Aalen's Additive Regression Model

Introduction

The dominating regression model in **survival analysis** is the **proportional hazards** model (or **Cox model**). Although very useful, it is clear that the Cox model cannot cover all relevant situations and that alternatives are needed. There are a number of other possible models, parametric and nonparametric ones. A member of the latter group is the additive regression model suggested by Aalen [2, 3]. One reason for seeking alternatives to the Cox model is that practitioners applying this model may not fully understand its complexities nor be able to check assumptions like proportionality. There is undoubtedly a somewhat uncritical use of the Cox model in the medical field. Clearly, other models may be no more easy to use, but by trying different approaches, one may get more insight into the data and develop a more critical attitude to the whole analysis. After all, there is no reason to assume that hazards will always be proportional.

In fact, experience tells one that effects are sometimes proportional, sometimes additive (*see* **Additive Hazard Models**; **Additive Model**), and often in between. Even when proportionality is a reasonable assumption, one often sees that the proportionality coefficient decreases over time. In fact, this is to be expected from **frailty** considerations; one basic consequence of frailty theory is that **relative risk** will often be expected to decrease over time. One advantage of the additive regression model presented here is that effects of **covariates** are allowed to vary freely over time. In contrast, when applying Cox analysis with standard packages, the normal approach will be to let the coefficients be constant over time and deviations from this may be difficult to incorporate. Thus, a standard Cox analysis gives no information about how the effects change over time and valuable information may be lost. It may also occur that significant effects may be masked. For instance, analyzing a set of survival data, it was found that a covariate indicating the extent of spread of the cancer ("N-stage") was not significant in the Cox analysis, while an additive model showed a clearly significant effect for the first year, but with the effect disappearing later [3, Table I and Figure 7(d)].

From a practical statistical point of view, it has been asserted that additive effects may be more informative than proportional effects. A hazard ratio of 2, say, may not be of much interest if the underlying basic hazard is very small. Then, the suggestion of a substantial effect may be misleading, and the real effect is better brought out by looking at the difference between hazards, which is by an additive approach.

The additive regression model generalizes the **Nelson–Aalen estimator**. For simplicity, assume that one wants to compare two groups. One way of doing this would be to make a Nelson–Aalen curve within each group and then plot the difference between the two curves. Now, this would be a fine procedure if the groups were defined by **randomization**, say. Otherwise, one will have to introduce the **covariates**, or **confounders**, which may explain the difference and adjust for them. The question then arises how to adjust the difference between two Nelson–Aalen curves, and the additive regression model is an answer to this.

A weakness of the additive approach is that the hazard rate is not naturally constrained to be positive. This may have odd effects occasionally, especially when predicting survival for individuals with extreme covariates where negative hazard may arise; see [9]. However, this does not prevent the additive model from being useful in most cases. As pointed out in [18], there is one important case for which the possibility of negative hazard rate is no problem, namely, when modeling excess hazard (*see* **Excess Risk**), for example, in cancer epidemiology. This subject is discussed further below.

The Additive Model

As indicated by the word "additive", the model has a linear structure. To be specific, assume that one observes the possibly censored life times of a number of individuals, the **censoring** times being assumed to be stopping times in the martingale sense [1] (*see* **Counting Process Methods in Survival Analysis**). Let $\lambda_i(t)$ denote the hazard rate of individual i, n the number of individuals, and r the number of covariates in the analysis. Consider the column vector $\boldsymbol{\lambda}(t)$ of hazard rates $\lambda_i(t), i = 1, \ldots, n$. The linear model is given as follows:

$$\boldsymbol{\lambda}(t) = \mathbf{Y}(t)\boldsymbol{\alpha}(t) \tag{1}$$

where the $n \times (r+1)$ matrix $\mathbf{Y}(t)$ is constructed as follows: If the ith individual is a member of the **risk set** at time t, then the i th row of $\mathbf{Y}(t)$ is the vector $\mathbf{Z}^i(t) = (1, Z_1^i(t), Z_2^i(t), \ldots, Z_r^i(t))'$, where $Z_j^i(t)$, $j = 1, \ldots, r$ are, possibly **time-dependent, covariate** values. If the ith individual is not in the risk set at time t, then the corresponding row of $\mathbf{Y}(t)$ contains only zeros. All sample paths of $\mathbf{Y}(t)$ are assumed to be left-continuous functions of t.

The vector $\boldsymbol{\alpha}(t) = (\alpha_0(t), \alpha_1(t), \ldots, \alpha_r(t))'$ contains the important regression information: The first element is a baseline function; while the remaining elements, called *regression functions*, measure the influence of the respective covariates. These functions are allowed to vary freely over time.

When turning to estimation, we concentrate on the cumulative regression functions defined by $A_j(t) = \int_0^t \alpha_j(s)\, ds$. Let $\mathbf{A}(t)$ be the column vector with elements $A_j(t)$, $j = 0, \ldots, r$. This is estimated by an approach that is similar to that for ordinary linear models [2, 3], resulting in the following estimator:

$$\mathbf{A}^*(t) = \sum_{T_k \le t} \mathbf{X}(T_k)\mathbf{I}_k \tag{2}$$

Here $T_1 < T_2 < \cdots$ are the ordered event times, while $\mathbf{I}_k$ is a column vector consisting of zeros except for a one in the place corresponding to the subject who experiences an event at time T_k. The estimator is only defined over the time interval, where $\mathbf{Y}(t)$ has full rank. The matrix $\mathbf{X}(t)$ is a generalized inverse of $\mathbf{Y}(t)$ (*see* **Matrix Algebra**) and will ordinarily be defined by the ordinary **least squares** inverse:

$$\mathbf{X}(t) = [\mathbf{Y}(t)'\mathbf{Y}(t)]^{-1}\mathbf{Y}(t)' \tag{3}$$

The components of $\mathbf{A}^*(t)$ are intended to be plotted against time and to give information about effects of covariates. Notice that the regression functions are the derivatives of the cumulative functions, and so it is the slopes of the plots that are informative. A decreasing slope means a decreasing additive effect (but this may not imply that the *relative* effect decreases).

The components of $\mathbf{A}^*(t)$ converge asymptotically, under appropriate conditions, to normal processes with independent increments. An estimator for the **covariance matrix** of $\mathbf{A}^*(t)$ is given by:

$$\boldsymbol{\Omega}^*(t) = \sum_{T_k \le t} \mathbf{X}(T_k)\, \mathbf{I}_k^D\, \mathbf{X}(T_k)', \tag{4}$$

where $\mathbf{I}_k^D$ is a diagonal matrix with the vector $\mathbf{I}_k$ as diagonal.

The model may also be formulated in terms of **counting processes** where the justification of the estimator and its properties is more easily seen [6]. The linear nature of the additive model fits very nicely with the counting process apparatus of stochastic integrals and so on. An extensive theory for the model has been derived, including asymptotic theory, test statistics, and martingale **residuals**. The latter ones and other checking procedures developed for the Cox model apply equally well to the additive approach [4]. Various issues concerning the model and its generalizations have been dealt with in [9, 10, 12, 14, 15, 17].

A practical advice concerning the analysis may be given: It is usually advantageous to center the covariates (subtracting the mean) before analysis. Then, the estimate of the cumulative baseline function, $A_0^*(t)$, will have a clear interpretation, namely, as the estimated cumulative hazard of an "average" individual.

The Semiparametric Additive Risk Model

One practical problem with the additive model in the above form is that all effects are nonparametric, thus making the description of some covariate effects unnecessarily complicated even when it is not needed. McKeague and Sasieni [15] suggested a very useful submodel of the additive model

$$\boldsymbol{\lambda}(t) = \mathbf{Y}(t)\boldsymbol{\alpha}(t) + \mathbf{W}(t)\boldsymbol{\gamma}, \tag{5}$$

where the first component of the model, $\mathbf{Y}(t)\boldsymbol{\alpha}(t)$, is defined as above and the second component, $\mathbf{W}(t)\boldsymbol{\gamma}$, is defined similarly ($\mathbf{W}(t)$ is an $n \times q$ dimensional matrix), and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$ is a regression parameter. The covariates of the model are thus partitioned in those whose effects depend on time and those whose effects are constant. Lin and Ying [13] considered a special case of this model in which the nonparametric part of the model only contains a baseline. Scheike [17] suggested a procedure for testing if effects in the **semiparametric** model depends significantly on time, thus making a stepwise model reduction strategy possible. The semiparametric model allows the data analyst to reduce effects that are not time varying to a parametric form thus giving a much simpler description to those effects.

McKeague and Sasieni [15] derived explicit estimators for $\mathbf{A}(t)$ and $\boldsymbol{\gamma}$ and their standard errors that are simple to compute.

Example

As an example of the additive analysis, we will use the (PBC) data on survival of 418 patients with primary biliary cirrhosis presented in [8]. The source of our data set is the survival package of **S-Plus/R**. The following covariates are included: age (in years), log(albumin), bilirubin (dichotomized as 0 when bilirubin is less than 3.25 mg/dl and 1 otherwise), edema (dichotomized as 0 for no edema and 1 for edema present now or before), and log(prothrombin time). Figures 1 and 2 present cumulative regression functions for the covariates bilirubin and edema. Pointwise 95% **confidence intervals** are also indicated. The null hypothesis: $\alpha_j(s) = 0$ over a suitable interval, may be tested by the supremum test of Scheike [17]. Here, it gives the values of 5.72

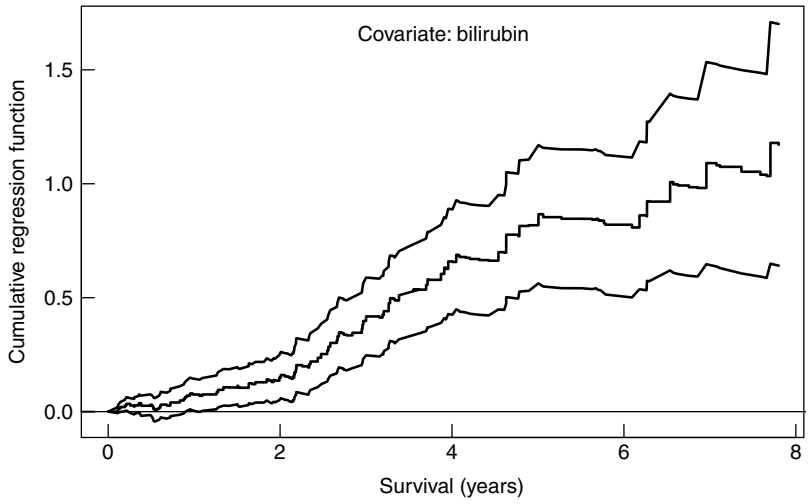

Figure 1 Estimated cumulative regression function for covariate bilirubin

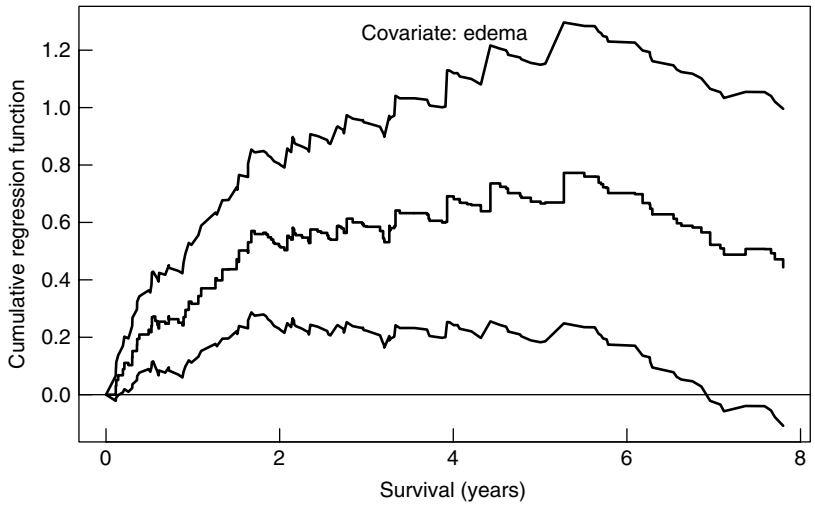

Figure 2 Estimated cumulative regression function for covariate edema

($p < 0.001$) and 4.00 ($p = 0.001$) for bilirubin and edema respectively. The plots may be interpreted as follows: For bilirubin, one sees a strongly positive slope, especially after 800 days, indicating a long-term effect on survival. For edema, the slope of the plot is largest to begin with. In fact, the plot soon levels off, and so it is clear that the effect of this covariate on survival is an initial effect that does not last. This is also found by Fleming and Harrington ([8], p. 191) by studying $\log(-\log(\text{survival}))$ plots, but the present procedure is a simpler way of discovering it, and it simultaneously adjusts for other covariates.

We now illustrate, the use of the semiparametric model and show that the data can be further summarized. In Aalen's additive model with all covariates (age, log(albumin), bilirubin, edema, and log(protime)), a test for constant covariate effects [17] gave the p value 0.88 for log(albumin). We, therefore, reduced the model to the semiparametric model in which log(albumin) had a constant effect, and other effects were time varying. In this model, it was found that age had a constant effect over time ($p = 0.89$). Further, stepwise model reduction lead to a model in which edema and log(protime) had time-varying effects (with **P-values** for constant effects at 0.01 and 0.04, respectively), and the remaining covariates were found to be well described by constant effects. Bilirubin had a constant effect of 0.143 (0.026), age 0.002 (0.001), and log(albumin) −0.263 (0.098). Note, that the cumulative regression effect of bilirubin shown in Figure 1 is well approximated by a line with slope 0.143.

Relative Survival Rate

There is an alternative useful representation of the results of the above analysis [19]. Consider a binary covariate with values 0 and 1, and let $A_j^*(t)$ be its cumulative regression function. Instead of plotting this, one could rather plot $R^*(t) = \exp(-A_j^*(t))$ versus t. This will be an estimate of the ratio between the survival curves, namely, the one with covariate value 1 divided by the one with covariate value 0, while the other covariates are kept constant. The quantity $R^*(t)$ is similar to what in epidemiology is termed a relative survival rate (*see* **Excess Mortality**).

The relative survival rates have been computed for the covariates bilirubin and edema in the example and presented in Figures 3 and 4. One sees that the relative survival declines to about 35% for bilirubin and a little above 60% for edema.

Excess Hazard Models

When studying the survival of cancer patients, one is interested in modeling the excess mortality, which is the mortality that remains when one subtracts the expected mortality (which is derived from ordinary

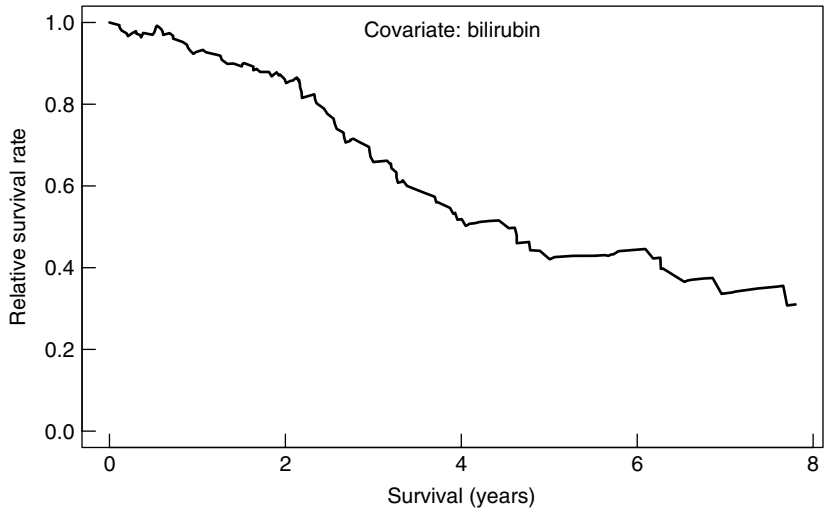

Figure 3 Estimated relative survival rate for covariate bilirubin

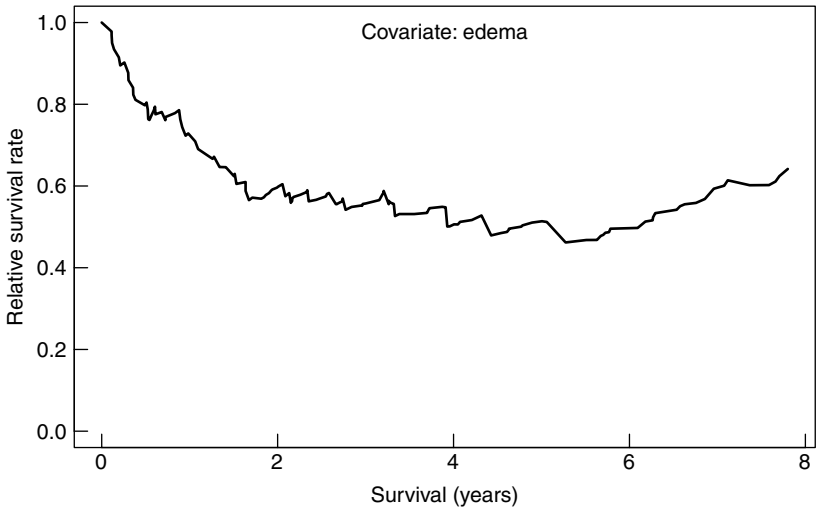

Figure 4 Estimated relative survival rate for covariate edema

life tables). It is this excess hazard that is supposed to be the cause-specific hazard related to the disease in question. Following up work of Andersen and Væth [7], Zahl [18] has extended the additive model to the excess hazards framework; see also [20]. Such excess hazards may well be negative and Zahl has shown that the additive model may give a better fit than the proportional hazards model.

Further Developments

The additive risk model, including the semiparametric version, has certain **robustness** properties that have been described along with robust **standard errors** in [17].

Estimating transition probabilities in **Markov chains** are of great interest in many practical applications. The additive model is suitable when one wants to adjust the transition probabilities for covariates [5]. An application of the additive model to adjusting for censoring in a more general multistage framework is given by Satten and Datta [16].

Software

An S-plus program, called Addreg, for making the cumulative regression plots is available on the web page `www.med.uio.no/imb/stat/addreg/`. Programs developed by T. Scheike are available on the web page: `www.biostat.ku.dk/~ts/`. Also, a program is available in Stata; see [11]. Finally, the survival package in S-plus contains a routine called aareg, which can make the plots described here.

References

[1] Aalen, O.O. (1978). Non-parametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[2] Aalen, O.O. (1980). *A Model for Non-Parametric Regression Analysis of Counting Processes, Lecture Notes in Statistics*, Vol. 2. Springer-Verlag, New York, pp. 1–25.

[3] Aalen, O.O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**, 907–925.

[4] Aalen, O.O. (1993). Further results on the non-parametric linear regression model in survival analysis, *Statistics in Medicine* **12**, 1569–1588.

[5] Aalen, O.O., Borgan, O. & Fekjær, H. (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach, *Biometrics* **57**, 993–1001.

[6] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[7] Andersen, P.K. & Væth, M. (1989). Simple parametric and non-parametric models for excess and relative mortality, *Biometrics* **45**, 523–535.

[8] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[9] Grønnesby, J.K. & Borgan, O. (1996). A method for checking regression models in survival analysis based on the risk score, *Lifetime Data Analysis* **2**, 315–328.

[10] Henderson, R. & Milner, A. (1991). Aalen plots under proportional hazards, *Applied Statistics* **40**, 401–409.
[11] Hosmer, D.W. & Royston, P. (2002). Using Aalen's linear hazard model to investigate time-varying effects in the proportional hazards model, *The Stata Journal* **2**, 331–350.
[12] Huffer, F.W. & McKeague, I.W. (1991). Weighted least squares estimation for Aalen's additive risk model, *Journal of American Statistical Association* **86**, 114–129.
[13] Lin, D.Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61–71.
[14] Martinussen, T. & Scheike, T. (2002). A flexible additive multiplicative hazard model, *Biometrika* **89**, 283–298.
[15] McKeague, I. & Sasieni, P. (1994). A partly parametric additive risk model, *Biometrika* **81**, 501–514.
[16] Satten, G.A. & Datta, S. (2002). Marginal estimation for multistage models: waiting time distributions and competing risks analyses, *Statistics in Medicine* **21**, 3–19.
[17] Scheike, T.H. (2002). The Additive Nonparametric and Semiparametric Aalen Model as the Rate Function for a Counting Process, *Lifetime Data Analysis* **8**, 247–262.
[18] Zahl, P.-H. (1996). A linear non-parametric regression model for the excess intensity, *Scandinavian Journal of Statistics* **23**, 353–364.
[19] Zahl, P.-H. & Aalen, O.O. (1998). Adjusting and comparing survival curves by means of an additive risk model, *Lifetime Data Analysis* **4**, 149–168.
[20] Zahl, P.-H. & Tretli, S. (1997). Long term survival of breast cancer in Norway by age and clinical stage, *Statistics in Medicine* **16**, 1435–1449.

ODD O. AALEN & THOMAS H. SCHEIKE

Aalen–Johansen Estimator

The survival data situation may be described by the **Markov process** with the two states "alive" and "dead". Splitting the state "dead" into two or more states, corresponding to different causes of death, a Markov model for **competing risks** is obtained. Another Markov model of importance for biostatistical research is the illness–death model with states "healthy", "diseased" and "dead". For survival data, the probability of a transition from state "alive" to state "dead" may be estimated as one minus the **Kaplan–Meier** estimator. The Kaplan–Meier estimator may be generalized to nonhomogeneous Markov processes with a finite number of states. Such a generalization was considered by Aalen [1] for the competing risks model and independently by Aalen & Johansen [2] and Fleming [5, 6] for the general case. In particular, the **product–integral** formulation of Aalen & Johansen [2] shows how the estimator, usually denoted the Aalen–Johansen estimator, can be seen as a matrix version of the Kaplan–Meier estimator.

Below, we first consider the competing risks model and the Markov illness–death model for a chronic disease. This gives illustrations of the Aalen–Johansen estimator in two simple situations where its elements take an explicit form. Then we present the Aalen–Johansen estimator in general, and show how it is obtained as the product–integral of the **Nelson–Aalen estimators** for the cumulative transition intensities. We also indicate briefly how this may be used to study its statistical properties. A detailed account is given in the monograph by Andersen et al. [3, Section IV.4].

Competing Risks

Assume that we want to study the time to death and cause of death in a homogeneous population. This situation with competing causes of death may be modeled by a Markov process with one transient state 0, corresponding to "alive", and k absorbing states corresponding to "dead by cause h", $h = 1, 2, \ldots, k$. The transition intensity from state 0 to state h is denoted $\alpha_{0h}(t)$ and describes the instantaneous risk of dying from cause h, i.e. $\alpha_{0h}(t)\,\mathrm{d}t$ is the probability that an individual will die of cause h in the small time interval $[t, t + \mathrm{d}t)$, given that it is still alive just prior to t. The $\alpha_{0h}(t)$ are also termed cause-specific **hazard rate** functions. For $h = 1, 2, \ldots, k$, we write $P_{0h}(s, t)$ for the probability that an individual in state 0 (i.e. alive) at time s will be in state h (i.e. dead from cause h) at a later time t. These transition probabilities are often termed cumulative incidence functions. Finally, let $P_{00}(s, t)$ denote the probability that an individual who is alive (i.e. in state 0) at time s will still be alive at a later time t. Then

$$P_{00}(s,t) = \exp\left[-\int_s^t \sum_{h=1}^{k} \alpha_{0h}(u)\,\mathrm{d}u\right], \qquad (1)$$

and

$$P_{0h}(s,t) = \int_s^t P_{00}(s,u)\alpha_{0h}(u)\,\mathrm{d}u, \qquad (2)$$

for $h = 1, 2, \ldots, k$.

Assume that we have a sample of n individuals from the population under study. Each individual is followed from an entry time to death or **censoring**, i.e. our observations may be subject to right censoring and/or left **truncation**. We denote by $t_1 < t_2 < \cdots$ the times when deaths (of any cause) are observed, and let d_{0hj} be the number of individuals who die from cause h (i.e. make a transition from state 0 to state h) at t_j. We also introduce $d_{0j} = \sum_{h=1}^{k} d_{0hj}$ for the number of deaths at t_j due to any cause, and let r_{0j} be the number of individuals at risk (i.e. in state 0) just prior to time t_j. Then the survival probability (1) may be estimated by the Kaplan–Meier estimator:

$$\hat{P}_{00}(s,t) = \prod_{s<t_j\le t}\left(1 - \frac{d_{0j}}{r_{0j}}\right), \qquad (3)$$

while the **cumulative incidence** function (2) may be estimated by

$$\hat{P}_{0h}(s,t) = \sum_{s<t_j\le t} \hat{P}_{00}(s,t_{j-1})\left(\frac{d_{0hj}}{r_{0j}}\right), \qquad (4)$$

for $h = 1, 2, \ldots, k$. Note that (4) is obtained from (2) by replacing $P_{00}(s,u) = P_{00}(s,u-)$ by $\hat{P}_{00}(s,u-)$ and $\alpha_{0h}(u)\,\mathrm{d}u$ by $\mathrm{d}\hat{A}_{0h}(u)$, the increment of the Nelson–Aalen estimator $\hat{A}_{0h}(t) = \sum_{t_j\le t} d_{0hj}/r_{0j}$ for the cumulative cause-specific **hazard rate** function $A_{0h}(t) = \int_0^t \alpha_{0h}(u)\,\mathrm{d}u$.

The variance of the Kaplan–Meier estimator (3) may in the usual way be estimated by Greenwood's formula (*see* **Kaplan–Meier Estimator**), while when there are no ties in the data,

$$\widehat{\mathrm{var}}\,\hat{P}_{0h}(s,t) = \sum_{s<t_j\le t} [\hat{P}_{00}(s,t_{j-1})\hat{P}_{0h}(t_j,t)]^2 \times (r_{0j}-1)r_{0j}^{-3}d_{0j} + \sum_{s<t_j\le t} \hat{P}_{00}(s,t_{j-1})^2[1-2\hat{P}_{0h}(t_j,t)] \times (r_{0j}-1)r_{0j}^{-3}d_{0hj}. \quad (5)$$

By breaking the ties at random, this variance estimator may also be used when there are a small number of tied observations (*see* **Tied Survival Times**). A more systematic treatment of variance estimation in the presence of ties is discussed below.

To illustrate the above results, we consider data on a cohort of uranium miners from the Colorado Plateau (see, for example, [7]). The cohort consisted of 3347 Caucasian male miners recruited between 1950 and 1960 and was traced for mortality outcomes to December 31, 1982, by which time there were 258 lung cancer deaths and 1000 deaths from other causes. Of these deaths, 145 and 442 occurred between 40 and 60 years of age. The data were collected to study the effects of radon exposure and smoking on mortality, but for our illustrative purposes we will study the (marginal) risk of death from lung cancer disregarding the information on these exposures.

We use the competing risks model with two competing causes of death, corresponding to "dead from lung cancer" (state 1) and "dead from other causes" (state 2), and with age as the time-scale. Figure 1 shows $\hat{P}_{01}(40,t)$ for $40 < t \le 60$, i.e. the estimated risk that a 40 years old miner will die from lung cancer between 40 and t years of age taking into account the risk of death from other causes. Pointwise 95% (log-transformed) confidence intervals based on the approximate normality of the Aalen–Johansen estimator (cf. below) are also shown. For comparison, Figure 1 also shows the estimated risk of lung cancer death disregarding the competing causes of death (computed as one minus the Kaplan–Meier estimator treating deaths from other causes as censorings). This estimate is sometimes interpreted as estimating the probability of death due to lung cancer, assuming this to be the only possible cause of death. Such an interpretation may be quite speculative, however; see the discussion in [9, Chapter 7]. The estimate disregarding competing risks is, of course, larger than the estimate that takes the competing causes of death into account; the difference between them increases with age as the risk of dying from other causes increases.

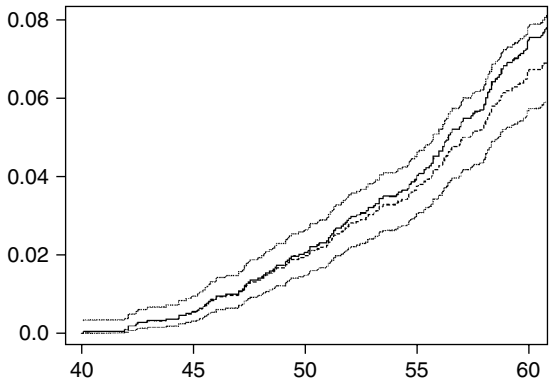

Figure 1 Aalen–Johansen estimate for the risk of dying from lung cancer taking into account the risk of death from other causes (- - - - - -) with 95% log-transformed confidence intervals (· · · · · · ·). Risk estimate disregarding other causes of death is also given (————)

An Illness–Death Model

To study the occurrence of a chronic disease as well as death in a homogeneous population, we may adopt the Markov illness–death model with states 0, 1 and 2 corresponding to "healthy", "diseased" and "dead", respectively, and where no recovery (i.e. transition from state 1 to state 0) is possible. The transition intensities of the model are denoted $\alpha_{01}(t), \alpha_{02}(t)$ and $\alpha_{12}(t)$ and describe the instantaneous risks of transitions between the states, i.e. $\alpha_{01}(t)\,\mathrm{d}t$ is the probability that an individual who is healthy just prior to time t will get diseased in the small time interval $[t, t+\mathrm{d}t)$, while $\alpha_{02}(t)\,\mathrm{d}t$ and $\alpha_{12}(t)\,\mathrm{d}t$ are the probabilities that an individual who is disease-free, respectively diseased, just before time t, will die in the small time interval $[t, t+\mathrm{d}t)$. For an individual who is healthy (i.e. in state 0) at time s, we write $P_{01}(s,t)$ for the probability that he is diseased (i.e. in state 1) at a later time t, while $P_{00}(s,t)$ is the probability that he is still healthy (i.e. in state 0) at

that time. Similarly, for an individual who is diseased (i.e. in state 1) at time s, we let $P_{11}(s,t)$ denote the probability that he is still alive (i.e. in state 1) at time t. Then we have

$$P_{00}(s,t) = \exp\left\{-\int_s^t [\alpha_{01}(u) + \alpha_{02}(u)]\,\mathrm{d}u\right\}, \quad (6)$$

$$P_{11}(s,t) = \exp\left[-\int_s^t \alpha_{12}(u)\,\mathrm{d}u\right], \quad (7)$$

$$P_{01}(s,t) = \int_s^t P_{00}(s,u)\alpha_{01}(u)P_{11}(u,t)\,\mathrm{d}u. \quad (8)$$

It is seen that (6) and (7) are of the same form as the survival probability in the survival data situation.

Assume, then, that we have a sample of n individuals from the population under study, and that each individual is followed from an entry time to death or censoring. Exact times of disease occurrences and deaths are recorded, and we denote by $t_1 < t_2 < \cdots$ the times of any observed event (disease occurrence or death). Furthermore, we let d_{01j} be the number of individuals who get diseased (i.e. make a transition from state 0 to state 1) at t_j, while d_{02j} and d_{12j} denote the numbers of disease-free, respectively diseased, individuals who die at that time. Finally, we introduce $d_{0j} = d_{01j} + d_{02j}$ for the total number of transitions out of state 0, and let r_{0j} and r_{1j} be the number of healthy (i.e. in state 0) and diseased (i.e. in state 1) individuals, respectively, just prior to time t_j. Then (6) and (7) may be estimated by the Kaplan–Meier estimators:

$$\hat{P}_{00}(s,t) = \prod_{s<t_j\le t}\left(\frac{1-d_{0j}}{r_{0j}}\right), \quad (9)$$

$$\hat{P}_{11}(s,t) = \prod_{s<t_j\le t}\left(\frac{1-d_{12j}}{r_{1j}}\right), \quad (10)$$

while an estimator for (8) is

$$\hat{P}_{01}(s,t) = \sum_{s<t_j\le t}\hat{P}_{00}(s,t_{j-1})\left(\frac{d_{01j}}{r_{0j}}\right)\hat{P}_{11}(t_j,t). \quad (11)$$

Note that (11) is obtained from (8) by replacing $P_{00}(s,u) = P_{00}(s,u-)$ by $\hat{P}_{00}(s,u-)$, $P_{11}(u,t)$ by $\hat{P}_{11}(u,t)$ and $\alpha_{01}(u)\,\mathrm{d}u$ by $\mathrm{d}\hat{A}_{01}(u)$, the increment of the Nelson–Aalen estimator $\hat{A}_{01}(t) = \sum_{t_j\le t} d_{01j}/r_{0j}$ for the cumulative disease intensity $A_{01}(t) = \int_0^t \alpha_{01}(u)\,\mathrm{d}u$. The variance of the Kaplan–Meier estimators (9) and (10) may be estimated by Greenwood's formula, while

$$\begin{aligned}\widehat{\mathrm{var}}\,\hat{P}_{01}(s,t) &= \sum_{s<t_j\le t}\hat{P}_{00}(s,t_{j-1})^2[\hat{P}_{11}(t_j,t) - \hat{P}_{01}(t_j,t)]^2 \\ &\quad\times (r_{0j}-1)r_{0j}^{-3}d_{01j} \\ &\quad+ \sum_{s<t_j\le t}[\hat{P}_{00}(s,t_{j-1})\hat{P}_{01}(t_j,t)]^2(r_{0j}-1)r_{0j}^{-3}d_{02j} \\ &\quad+ \sum_{s<t_j\le t}[\hat{P}_{01}(s,t_{j-1})\hat{P}_{11}(t_j,t)]^2(r_{1j}-1)r_{1j}^{-3}d_{12j},\end{aligned} \quad (12)$$

when there are no ties in the data, or when a few ties have been broken at random.

Before we illustrate these results, let us mention that other interpretations of the states are possible. In particular, in a study involving the treatment of cancer, state 0 could correspond to "no response to treatment", state 1 to "response to treatment" and state 2 to "relapse". The probability $P_{01}(s,t)$ is then the probability of being in response function suggested by Temkin [10] and sometimes used as an outcome measure when studying the efficacy of cancer chemotherapy. Another interpretation arises in the study of complications to a disease. Here, state 0 could correspond to "diseased with no complications", state 1 to "diseased with complications" and state 2 to "dead". This interpretation of the states is the one relevant for the following illustration.

The Steno Memorial Hospital in Greater Copenhagen has, since 1933, served as a diabetes specialist hospital for patients from the whole of Denmark. From the medical records at Steno we use for illustration data on the 374 female patients referred between 1933 and 1981 in whom the diagnosis insulin-dependent diabetes mellitus was established (usually by a general practitioner or another hospital) before the age of 10 years and between 1933 and 1972. The patients were followed from first contact with Steno to death, emigration, or December 31, 1984. One of the major complications of insulin-dependent diabetes is diabetic nephropathy, which is a sign of kidney failure. Seventeen patients had diabetic nephropathy at first admission to Steno, while 76 developed this complication during the observation period. The seriousness of diabetic nephropathy is reflected by the fact that among these 93 patients

54 were observed to die, whereas only 30 of the 281 patients who did not develop diabetic nephropathy died during the observation period.

We model the disease histories of the patients by the Markov illness–death model with the states 0 and 1 corresponding to "alive without diabetic nephropathy" and "alive with diabetic nephropathy", respectively, and with diabetes duration as time-scale. Figure 2 shows $\hat{P}_{01}(5, t)$, i.e. the estimated probability of being alive with diabetic nephropathy for patients without this complication five years after the onset of the disease. Pointwise 95% (log-transformed) **confidence intervals** based on the approximate normality of the Aalen–Johansen estimator (cf. below) are also shown. It is seen that the probability of being alive with diabetic nephropathy (among the group of patients we consider) first increases up to an estimated value of 17% after 23 years of diabetes duration, and then declines due to the high mortality among these patients.

It should be realized that Figure 2 is based on two crude assumptions. First, calendar time trends in mortality and incidence of diabetic nephropathy are not taken into account. Secondly, by using a Markov process to model the disease histories, the effect on mortality of the duration of diabetic nephropathy has been neglected. A point of less importance is that the exact times of onset of diabetic nephropathy were not known for nine of the 93 patients with this complication. For these nine patients, predicted times for the occurrence of diabetic nephropathy were used. A further discussion and analysis of the data are given, e.g. by Borch-Johnsen et al. [4]. The data were used for illustrative purposes by Andersen et al. [3] who also describe how the nine predicted times have been calculated.

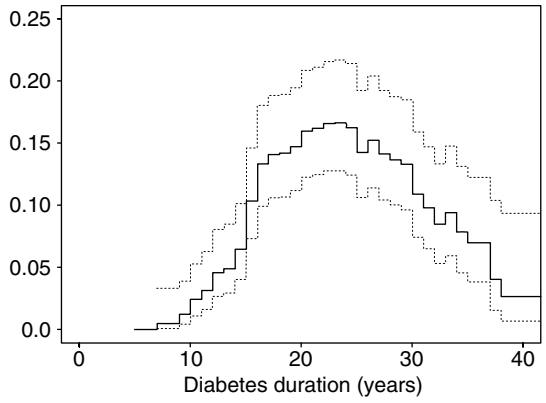

Figure 2 Aalen–Johansen estimate of the probability of being alive with diabetic nephropathy for female patients with diabetes onset before 10 years of age and with no sign of diabetic nephropathy five years after the onset of the disease (————). Pointwise 95% log-transformed confidence intervals are also shown (· · · · · · ·)

The General Case

We then consider a general Markov process with a finite number of states that may be used to model the life histories of individuals from a homogeneous population. Let $\mathcal{I} = \{0, 1, \ldots, k\}$ be the state space of the Markov process, and denote by $\alpha_{gh}(t)$ the transition intensity from state $g \in \mathcal{I}$ to state $h \in \mathcal{I}$, $g \neq h$. The transition intensities describe the instantaneous risks of transitions between the states, so $\alpha_{gh}(t)\,\mathrm{d}t$ is the probability that an individual who is in state g just before time t will make a transition to state h in the small time interval $[t, t + \mathrm{d}t)$. Furthermore, for all $g, h \in \mathcal{I}$, we let $P_{gh}(s, t)$ denote the probability that an individual who is in state g at time s will be in state h at a later time t, and we write $\mathbf{P}(s, t)$ for the $(k + 1) \times (k + 1)$ matrix of these transition probabilities. Only for simple Markov processes, like the competing risks and illness–death models considered earlier, is it possible to give explicit expressions for the $P_{gh}(s, t)$ in terms of the transition intensities, cf. (1), (2) and (6)–(8). We will see later, however, that the transition probability matrix $\mathbf{P}(s, t)$ itself can be expressed in terms of the $(k + 1) \times (k + 1)$ matrix $\boldsymbol{\theta}(t)$ of the transition intensities. First, we review the Aalen–Johansen estimator for $\mathbf{P}(s, t)$ and discuss the estimation of (co)variances.

Suppose that we have a sample of n individuals from the population under study. The individuals may be followed over different periods of time, so our observations of their life histories may be subject to left truncation and/or right censoring. A crucial assumption, however, is that truncation and censoring are independent so that the entry and censoring times do not carry any information on the risks of transitions between the states; cf. Andersen et al. [3, Sections III.2– 3] for a general discussion. We assume that exact times for transitions between the states are recorded, and denote by $t_1 < t_2 < \cdots$ the times when transitions between any two states are observed. Furthermore, for $g, h \in \mathcal{I}$, $g \neq h$, we

let d_{ghj} be the number of individuals who experience a transition from state g to state h at t_j, and introduce $d_{gj} = \sum_{h \neq g} d_{ghj}$ for the number of transitions out of state g at that time. Finally, we let r_{gj} be the number of individuals in state g just prior to time t_j. Then, the Aalen–Johansen estimator takes the form

$$\hat{\mathbf{P}}(s,t) = \prod_{s<t_j \leq t} (\mathbf{I} + \hat{\boldsymbol{\theta}}_j). \tag{13}$$

Here, $\mathbf{I}$ is the $(k+1) \times (k+1)$ identity matrix, $\hat{\boldsymbol{\theta}}_j$ is the $(k+1) \times (k+1)$ matrix with entry (g,h) equal to $\hat{\alpha}_{ghj} = d_{ghj}/r_{gj}$ for $g \neq h$ and entry (g,g) equal to $\hat{\alpha}_{ggj} = -d_{gj}/r_{gj}$, and the matrix product is taken in the order of increasing t_js. For simple models like the competing risks model and the illness–death model considered earlier, we are able to give explicit expressions for the elements of (13), cf. (3), (4), and (9)–(11). In general, however, this is not possible. But, in any case, a direct implementation of (13) is simple using software that can handle matrix multiplications (*see* **Matrix Computations**).

For any $g,h,m,r \in \mathcal{I}$, the covariance between $\hat{P}_{gh}(s,t)$ and $\hat{P}_{mr}(s,t)$ may be estimated by

$$\begin{aligned}
&\widehat{\mathrm{cov}}[\hat{P}_{gh}(s,t), \hat{P}_{mr}(s,t)] \\
&\quad = \sum_{i=0}^{k} \sum_{l \neq i} \sum_{s<t_j \leq t} \hat{P}_{gi}(s,t_{j-1}) \hat{P}_{mi}(s,t_{j-1}) [\hat{P}_{lh}(t_j,t) \\
&\qquad - \hat{P}_{ih}(t_j,t)][\hat{P}_{lr}(t_j,t) - \hat{P}_{ir}(t_j,t)] \\
&\qquad \times (r_{ij}-1) r_{ij}^{-3} d_{ilj},
\end{aligned} \tag{14}$$

provided that there are no ties in the data or that a small number of tied observations have been broken at random. Formulas (5) and (12) given earlier are special cases of (14). As an alternative to (14), or to handle ties in a systematic manner, one may use the recursion formula:

$$\begin{aligned}
&\widehat{\mathrm{cov}}[\hat{P}_{gh}(s,t_j), \hat{P}_{mr}(s,t_j)] \\
&\quad = \sum_{i=0}^{k} \sum_{l=0}^{k} \widehat{\mathrm{cov}}[\hat{P}_{gi}(s,t_{j-1}), \hat{P}_{ml}(s,t_{j-1})] \\
&\qquad \times (\delta_{ih} + \hat{\alpha}_{ihj})(\delta_{lr} + \hat{\alpha}_{lrj}) + \sum_{i=0}^{k} \hat{P}_{gi}(s,t_{j-1}) \\
&\qquad \times \hat{P}_{mi}(s,t_{j-1}) \widehat{\mathrm{cov}}(\hat{\alpha}_{ihj}, \hat{\alpha}_{irj}),
\end{aligned} \tag{15}$$

which describes how the estimated (co)variances are updated at the times of the observed transitions. (The estimates are constant between the t_js.) Here, δ_{ih} is a Kronecker delta, while $\widehat{\mathrm{cov}}(\hat{\alpha}_{ihj}, \hat{\alpha}_{irj})$ equals $(\delta_{hr} r_{ij} - d_{ihj}) r_{ij}^{-3} d_{irj}$ when $h \neq i, r \neq i$; it equals $-(r_{ij} - d_{ij}) r_{ij}^{-3} d_{irj}$ when $h = i \neq r$; and it equals $(r_{ij} - d_{ij}) r_{ij}^{-3} d_{ij}$ when $h = r = i$. When there are no ties in the data (14) and (15) give identical results.

Product–Integral Representation and Relation to the Nelson–Aalen Estimator

We now review how the transition probability matrix may be derived from the transition intensities $\alpha_{gh}(t)$ and describe how the Aalen–Johansen estimator is related to the Nelson–Aalen estimators for the cumulative transition intensities. To this end, we introduce $\alpha_{gg}(t) = -\sum_{h \neq g} \alpha_{gh}(t)$ and write $\boldsymbol{\theta}(t)$ for the $(k+1) \times (k+1)$ matrix with element (g,h) equal to $\alpha_{gh}(t)$. Then, the transition probability matrix $\mathbf{P}(s,t)$ is the unique solution to the Kolmogorov forward differential equation $(\partial/\partial t)\mathbf{P}(s,t) = \mathbf{P}(s,t)\boldsymbol{\theta}(t)$ with initial condition $\mathbf{P}(s,s) = \mathbf{I}$. By a general result for product–integrals (Volterra's equation), this solution takes the form $\mathbf{P}(s,t) = \prod_{(s,t]} [\mathbf{I} + \boldsymbol{\theta}(u)\,\mathrm{d}u]$. Alternatively, if we introduce the $(k+1) \times (k+1)$ matrix $\mathbf{A}(t)$ with elements $A_{gh}(t) = \int_0^t \alpha_{gh}(s)\,\mathrm{d}s$, we may write

$$\mathbf{P}(s,t) = \prod_{(s,t]} [\mathbf{I} + \mathrm{d}\mathbf{A}(u)]. \tag{16}$$

This product–integral representation of the transition probability matrix of a Markov process is not restricted to the situation where transition intensities exist. In fact (16) assumes only the existence of cumulative transition intensities $A_{gh}(t)$, which do not need to be absolutely continuous.

For $g \neq h$ we may estimate the cumulative transition intensity $A_{gh}(t)$ by the Nelson–Aalen estimator $\hat{A}_{gh}(t) = \sum_{t_j \leq t} \hat{\alpha}_{ghj}$, while $\hat{A}_{gg}(t) = -\sum_{h \neq g} \hat{A}_{gh}(t) = \sum_{t_j \leq t} \hat{\alpha}_{ggj}$. Let $\hat{\mathbf{A}}(t) = \sum_{t_j \leq t} \hat{\boldsymbol{\theta}}_j$ be the $(k+1) \times (k+1)$ matrix with these elements. By (16) it is reasonable to estimate the transition probability matrix by $\hat{\mathbf{P}}(s,t) = \prod_{(s,t]} [\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}(u)]$. But since $\hat{\mathbf{A}}(t)$ is a matrix of step functions with a finite number of jumps on $(s,t]$, this is nothing but the Aalen–Johansen estimator (13).

Thus, the Aalen–Johansen and Nelson–Aalen estimators are related in exactly the same way as are the transition probability matrix and the cumulative transition intensities themselves. This suggests that the Aalen–Johansen estimator is the canonical nonparametric estimator for the matrix of transition probabilities in a Markov process with a finite number of states. This statement is supported by the fact that it may also be given a **nonparametric maximum likelihood** interpretation [8].

Martingale Representation and Statistical Properties

The product–integral formulation of the Aalen–Johansen estimator is useful for the study of its statistical properties. We here indicate a few main steps and refer to Andersen et al. [3, Section IV.4] for a detailed account. For each $g \in \mathcal{I}$ we introduce an indicator $J_g(t)$, which is one if there is at least one individual in state g just before time t, and zero otherwise. Furthermore, for all $g, h \in \mathcal{I}$ define $A^*_{gh}(t) = \int_0^t J_g(u)\,\mathrm{d}A_{gh}(u)$, and let $\mathbf{A}^*(t)$ be the $(k+1) \times (k+1)$ matrix with these elements. Finally, we introduce $\mathbf{P}^*(s,t) = \prod_{(s,t]}[\mathbf{I} + \mathrm{d}\mathbf{A}^*(u)]$, and note that this is almost the same as $\mathbf{P}(s,t)$ (cf. (16)) when there is only a small probability that one or more states will be empty at times u between s and t. By a general result for product–integrals (Duhamel's equation), we may then write

$$\hat{\mathbf{P}}(s,t)\mathbf{P}^*(s,t)^{-1} - \mathbf{I} = \int_{(s,t]} \hat{\mathbf{P}}(s,u-)\,\mathrm{d}(\hat{\mathbf{A}} - \mathbf{A}^*)(u)\mathbf{P}^*(s,u)^{-1}. \quad (17)$$

Here, $\hat{\mathbf{A}} - \mathbf{A}^*$ is a $(k+1) \times (k+1)$ matrix of square integrable martingales (*see* **Nelson–Aalen Estimator**). It follows that the right-hand side of (17) is a matrix-valued stochastic integral, and therefore itself a $(k+1) \times (k+1)$ matrix of mean zero square integrable martingales. As a consequence of this

$$\mathrm{E}[\hat{\mathbf{P}}(s,t)\mathbf{P}^*(s,t)^{-1}] = \mathbf{I},$$

so the Aalen–Johansen estimator is almost **unbiased**. Furthermore, the predictable variation process of the matrix-valued martingale (17) suggests an estimator for the covariance matrix of $\hat{\mathbf{P}}(s,t)\mathbf{P}^*(s,t)^{-1}$, and based on this the (co)variance estimators (14) and (15) may be derived.

The martingale representation (17) is also key to the study of the large sample properties of the Aalen–Johansen estimator. For fixed s it may be shown that $\hat{\mathbf{P}}(s,\cdot)$, properly normalized, converges weakly to a matrix-valued Gaussian process. In particular, when also t is given, the Aalen–Johansen estimator (13) is asymptotically multinormally distributed, a fact that was used earlier in connection with the construction of confidence intervals.

References

[1] Aalen, O.O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models, *Annals of Statistics* **6**, 534–545.

[2] Aalen, O.O. & Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics* **5**, 141–150.

[3] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Borch-Johnsen, K., Andersen, P.K. & Deckert, T. (1985). The effect of proteinuria on relative mortality in Type 1 (insulin-dependent) diabetes mellitus, *Diabetologia* **28**, 590–596.

[5] Fleming, T.R. (1978). Nonparametric estimation for nonhomogeneous Markov processes in the problem of competing risks, *Annals of Statistics* **6**, 1057–1070.

[6] Fleming, T.R. (1978). Asymptotic distribution results in competing risks estimation, *Annals of Statistics* **6**, 1071–1079.

[7] Hornung, R. & Meinhardt, T. (1987). Quantitative risk assessment of lung cancer in U.S. uranium miners, *Health Physics* **52**, 417–430.

[8] Johansen, S. (1978). The product limit estimator as maximum likelihood estimator, *Scandinavian Journal of Statistics* **5**, 195–199.

[9] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[10] Temkin, N.R. (1978). An analysis for transient states with application to tumor shrinkage, *Biometrics* **34**, 571–580.

ØRNULF BORGAN

Absolute Risk

Absolute risk is defined as the probability that a disease-free individual will develop a given disease over a specified time interval given current age and individual risk factors, and in the presence of **competing risks**. In mathematical terms, the absolute risk of developing a disease of interest c_1 in the age interval $[a_1, a_2)$ in the presence of competing risks c_2 for a person of age a_1 and with initial **covariates** x is given by

$$\pi(a_1, a_2; x) = \frac{\displaystyle\int_{a_1}^{a_2} h_1(u;x) \exp\left\{-\int_0^u [h_1(v;x) + h_2(v;x)]\mathrm{d}v\right\} \mathrm{d}u}{\displaystyle\exp\left\{-\int_0^{a_1} [h_1(v;x) + h_2(v;x)]\mathrm{d}v\right\}}, \tag{1}$$

where $h_1(v;x)$ and $h_2(v;x)$ are, respectively, the cause-specific **hazards** of developing c_1 and c_2 for an individual with current age v and level x of covariates X. In this formula, the numerator represents the probability of developing the disease of interest c_1 between ages a_1 and a_2 in the presence of competing risks c_2 while the denominator represents the probability of being at risk at age a_1, namely free of c_1 and c_2. This formulation underscores the conditional nature of absolute risk. However, a simpler and equivalent formulation can be obtained as

$$\pi(a_1, a_2; x) = \int_{a_1}^{a_2} h_1(u;x) \times \exp\left\{-\int_{a_1}^{u} [h_1(v;x) + h_2(v;x)]\,\mathrm{d}v\right\} \mathrm{d}u. \tag{2}$$

The hazard $h_1(u;x)$ can be expressed as a function of both the baseline hazard $h_1(u)$ (i.e. the hazard in subjects at baseline level of covariates x) and the level x of covariates X. For instance, if the covariates X have a **multiplicative** effect on the hazard, then the multiplicative relationship $h_1(u;x) = h_1(u)rr(u;x)$ is obtained, where the multiplier $rr(u;x)$ is the relative rate, also termed the rate ratio, **incidence density ratio**, **hazard ratio** (the term which is used throughout this article), instantaneous relative risk or, loosely, **relative risk** (see the section "Related Quantities" below). If the covariates X have an **additive** effect on the hazard, then the additive relationship $h_1(u;x) = h_1(u) + d(u;x)$ is obtained, where the additive term $d(u;x)$ is the rate difference or hazard difference or incidence density difference. Upon considering such expressions, one can note that the value of absolute risk depends on both the incidence of disease in the population and the strength of the relationship between covariates and disease. One consequence is that, while the hazard ratio is often portable from one population to another (portability being more questionable for the rate difference), portability is not a property of absolute risk, as the baseline incidence rate of disease may vary widely among populations that are separated in time and location or even among subgroups of populations, possibly because of differing genetic patterns or differing exposure to unknown risk factors. Additionally, competing causes of death (competing risks) may also have different patterns among different populations which might also influence values of absolute risk.

An important consideration is that covariates X may be time-dependent (*see* **Time-dependent Covariate**), in which case one must rely on a more general formulation of (1) and (2) obtained by (i) replacing initial covariate value x in $\pi(a_1, a_2; x)$ by covariate history in interval $[a_1, a_2)$, namely $\{x(v), a_1 \leq v < a_2\}$, and (ii) by using generalized versions of cause-specific hazards, namely $h_1(v; x(v))$ and $h_2(v; x(v))$, in the right-hand terms of (1) and (2). Eqs. (1) and (2) correspond to the special case in which covariates X remain constant throughout the interval. However, unless it is possible to predict (in a probabilistic or deterministic manner) the future variation of covariates over time, estimation is based on (1) or (2) in their original form, and relies on the initial covariate value x and the assumption that it remains constant. This approach is likely to underestimate absolute risk if covariates the associated risks of which can only increase with time are considered. Such variables include, for instance, family history of breast cancer and number of previous breast biopsies for benign breast disease, which are used in estimating the absolute risk of breast cancer from the Breast Cancer Detection and Demonstration Project [47] (see the section "Estimation From Population-Based or Nested Case–Control Studies" below).

Range

Absolute risk is a probability and therefore lies between 0 and 1 and is dimensionless. A value of 0, while theoretically possible, would correspond to very special cases such as a purely genetic disease for an individual not carrying the disease gene. A value of 1 would be even more unusual and might again correspond to a genetic disease with a **penetrance** of 1 for a **gene** carrier (but, even in this case, the value should be less than 1 if competing risks cannot be ignored).

Synonyms

The term *absolute risk* or *absolute cause-specific risk* has been used by several authors, including Dupont [35], Benichou & Gail [13, 14], Benichou [11], and Langholz & Borgan [62]. However, it is not a universally accepted term. Alternative terms include risk [59], individualized risk [47], individual risk [94], crude probability [28], crude incidence [60], **cumulative incidence** [49], cumulative incidence risk [75] and absolute incidence risk [76]. It should be noted that the definition of the two latter terms [75, 76] ignores the concept of competing risks.

Interpretation and Usefulness

Absolute risk provides an individual measure of the probability of disease occurrence, and can therefore be useful in counselling. It is well suited to predicting **risk** for an individual, unlike the hazard ratio or the relative risk, which quantify the increase in the probability of disease occurrence relative to subjects at the baseline level of risk factors, but do not quantify that probability itself. Moreover, individualized absolute risk estimates over specific time intervals are often more useful than general statements about risk such as "one in nine women will develop breast cancer during her lifetime" [3].

Absolute risk has been used as a tool for individual counseling in breast cancer. Indeed, a woman's decision to embark on a program of intensive surveillance with mammography or even to undergo prophylactic mastectomy depends on her awareness of the medical options, on personal preferences, and on absolute risk. A woman may have several risk factors and an elevated hazard ratio, but if her absolute risk of developing breast cancer over the next 10 years is small, she may be reassured and she may be well advised simply to embark on a program of surveillance. Conversely, she may be very concerned about her absolute risk over a longer time period, such as 30 years, and she may decide to undergo prophylactic mastectomy if her absolute risk is very high [92]. An assessment of absolute risk (and its range of uncertainty) can help the woman understand the extent of the risk and can therefore be useful in helping the woman and her doctor define an acceptable medical plan [17, 44, 47].

Absolute risk is also useful in designing trials of interventions to prevent the occurrence of a disease (*see* **Prevention Trials**) because the sample sizes required for these studies (*see* **Sample Size Determination for Clinical Trials**) depend importantly on the absolute risk of developing the disease during the period of study [8]. Absolute risk has also been used to define **eligibility criteria** in such studies. For example, women were enrolled in a preventive trial to decide whether the drug Tamoxifen can reduce the risk of developing breast cancer. Because Tamoxifen is a potentially toxic drug and because it was to be administered to a healthy population, it was decided to restrict eligibility to women with somewhat elevated absolute risks of breast cancer. Only women over age 59 and younger women whose absolute risks were estimated to equal or exceed that of a typical 60-year-old woman were eligible to participate [8, 93].

Absolute risk can also be important in decisions affecting public health. For example, in order to estimate the absolute reduction in lung cancer incidence that might result from measures to reduce exposure to radon, one could categorize a general population into subgroups based on age, sex, smoking status, and current radon exposure levels, and then estimate the absolute reduction in lung cancer incidence, in the presence of competing risks, that would result from lowering radon levels in each subgroup [13, 42]. Such an analysis would complement estimation of population **attributable risk** and generalized impact fractions.

The concept of absolute risk is also useful in a clinical setting as a measure of the individualized probability of an adverse event, such as a recurrence or death in diseased subjects. In that context, absolute risk depends on **prognostic factors** of

recurrence or death, rather than on factors influencing the risk of incident disease, and the time-scale of interest is usually time from diagnosis or from surgery rather than age. Absolute risk is a useful tool to help define individual patient management and, for instance, the absolute risk of recurrence in the next three years might be an important element in deciding whether to prescribe an aggressive and potentially toxic treatment regimen. Such an application is discussed in Benichou & Gail [13], who consider the absolute risk of recurrence as a function of cell type and TN staging in patients with resected lung cancer. Korn & Dorey [60] provide other examples. Note that in such a setting, 1 minus the absolute risk of recurrence differs from the standard disease-free survival probability (obtained from the disease-free interval distribution or time to recurrence distribution) in that absolute risk takes into account competing risks (deaths from other causes than the disease under study). The difference is particularly large if competing death rates are high compared to the disease-related adverse event rate, as among older people.

Properties

Two main points need to be emphasized. First, as is evident from its definition, absolute risk can only be estimated in reference to a specified time interval. One might be interested in short time spans (e.g. five years), long time spans (e.g. 30 years), or even lifetime absolute risk. Of course, absolute risk increases as the time span increases. In the clinical setting, the time span might also vary with the context and the severity of the disease.

Absolute risk can be strongly influenced by the intensity of competing risks (typically competing causes of death). Absolute risk varies inversely as a function of death rates from other causes (denoted by $h_2(v;x)$ in (1) and (2)). The same result in the clinical setting may lead to differences between 1 minus the absolute risk and the disease-free survival probability (see the section "Interpretation and Usefulness" above). Indeed, disease-free survival applies best in the situation in which no competing causes (unrelated to the disease under study) are acting to kill the patient before the occurrence of the disease or adverse event of interest [13].

Estimability

It follows from its definition that absolute risk is estimable if and only if cause-specific hazard rates for the disease (or event) of interest c_1 as well as death rates from competing causes c_2 are estimable (*see* **Estimation**). Therefore, absolute risk is directly estimable from **cohort** and **case–cohort studies**, but **case–control** and **cross-sectional studies** have to be complemented with follow-up data. Absolute risk is estimable from nested **case–control studies** or **population-based case–control studies**, in which the cohort or the specified population from which cases and controls are selected provides the necessary complementary information on incidence rates. While the theoretic possibility exists to complement cross-sectional studies with follow-up data, such designs do not seem to have been implemented.

An important feature of absolute risk is that it takes into account competing risks, that is the possibility for an individual to die of an unrelated disease before developing the disease (or disease-related event) of interest. Absolute risk is identifiable without any unverifiable competing risk assumptions, such as the assumption that competing risks act independently of the cause of interest because, as Prentice et al. [86] emphasize, all functions of the cause-specific hazards in (1) and (2) are estimable. Chiang [28] used the term "crude" probability to describe absolute risk, the probability of experiencing c_1 *in the presence* of competing risks c_2. This quantity is relevant for individual predictions and other applications discussed above rather than the underlying (or "net" or "latent") probability of experiencing c_1 in the absence of competing risks. One minus the standard disease-free survival represents that underlying probability of experiencing c_1 in the absence of competing risks or under the (unverifiable) assumption of independence between time to c_1 and time to c_2 (see [13, 27, 28, 43, 55, 60] and [86] for more details). The only competing risk assumption needed to estimate absolute risk concerns subjects lost to follow-up, who are assumed to be randomly selected from those at risk at the time of loss (independent noninformative **censoring**) [13].

Sometimes, estimates of competing hazards h_2 are based on external sources such as **vital statistics**. For instance, Gail et al. [47] developed breast cancer absolute risk estimates and used mortality rates from year 1979 for all causes except breast cancer

(see also [11, 13, 14] and [60]). Although (1) and (2) allow for competing risk hazards h_2 to depend on covariate level x, it is frequently assumed that h_2 does not depend on x. It could also be assumed that h_2 depends on a set of covariates X' that are different from covariates X.

Estimation from Cohort Studies

Since all cause-specific hazards can be estimated from cohort studies, it follows that absolute risk can also be directly estimated from cohort studies. Estimation of cause-specific hazards from cohort data is a standard topic and details can be found in epidemiology or survival analysis textbooks (*see* **Survival Analysis, Overview**). However, the details of absolute risk estimation have been worked out under several models, and properties of absolute risk estimates have been studied and compared. A review is given here.

Covariate-free Estimates of Absolute Risk

The following methods are appropriate for a homogeneous study population. They are also used to provide estimates of composite absolute risk in populations; namely, overall estimates of absolute risk that do not distinguish among levels of covariates X. Parametric and **nonparametric** estimators are presented.

The "density method" [59, 76, 77] estimates absolute risk $\pi(a_1, a_2)$ by the cumulative (incidence) risk given by $1 - \exp\{-\Lambda(a_1, a_2)\}$, where $\Lambda(a_1, a_2)$ is the cumulative hazard for the event of interest, c_1. This formulation ignores competing risks. The term x is omitted in Λ because an overall rather than an exposure-specific absolute risk is considered. This approach is parametric, as it relies on a piecewise **exponential** distribution of time to c_1, which corresponds to a piecewise constant hazard of developing c_1. It ignores competing risks, and therefore applies only in the absence of competing risks, which constitutes an important limitation.

Benichou & Gail [13] developed direct parametric estimators of absolute risk. They derived direct estimators of $\pi(a_1, a_2)$ based on (1) or (2) (still ignoring covariates X) under exponential and piecewise exponential models. Under the exponential assumption, hazards h_1 and h_2 are constant, while under the piecewise exponential assumption, hazards h_{1i} and h_{2i} are piecewise constant. The expression for $\pi(a_1, a_2)$ under the piecewise exponential assumption is given by [13]

$$\pi(a_1, a_2) = \sum_i h_{1i}(h_{1i} + h_{2i})^{-1} \times [1 - \exp\{-(h_{1i} + h_{2i})\Delta_i\}]A(i), \quad (3)$$

with $A(i) = \prod_j \exp\{-(h_{1j} + h_{2j})\Delta_j\}$. In (3), the sum is taken over all time intervals included in $[a_1, a_2)$, i is the corresponding index, h_{1i} (respectively h_{2i}) denotes the (constant) hazard for cause c_1 (respectively c_2) in interval i, Δ_i is the width of interval i, and the product in $A(i)$ is taken over all time intervals in $[a_1, a_2)$, but the last one and indexed by j. For simplicity, a_1 and a_2 are taken to correspond to interval bounds.

Hazard rates h_{1i} can easily be estimated by d_{1i}/t_i, where d_{1i} and t_i, respectively, denote the observed number of events and **person-years** in interval i. Analogous estimates of competing hazards h_{2i} are given by d_{2i}/t_i, where d_{2i} denotes the observed number of events in interval i. Corresponding point estimates of $\pi(a_1, a_2)$ can be obtained by replacing hazards by their estimates in (3). Under the simple exponential assumption, no separate intervals are considered as the hazards h_1 and h_2 are considered constant throughout time. Eq. (3) simplifies, as the sum includes only one term and $A(i)$ equals 1. Estimates of hazards are obtained as for the piecewise exponential model with a single interval.

Unlike estimates with the density method, direct estimates of absolute risk with the exponential and piecewise exponential assumptions do not ignore competing risks, therefore providing estimates of the absolute risk of developing c_1 *in the presence* of competing risks. Moreover, as for the density method, absolute risk can be estimated for a much longer duration than the actual follow-up of individuals in the study if age is the time scale (open cohort), provided that there is no secular trend in age-specific disease incidence.

Variance estimates of the absolute risk estimate are obtained using the **delta method** [87], and corresponding **confidence intervals** follow. Details are given in Benichou & Gail [13] for the exponential and piecewise exponential models. Properties of point and variance estimators were studied by Benichou & Gail [13] for the case of a closed cohort. When the simple exponential model was correct,

simulations showed no or very little **bias** in point estimates of $\pi(a_1, a_2)$, and analytic and simulation results showed that substantial gains in efficiency could be achieved with a simple exponential analysis. **Simulations** showed that exponential and piecewise exponential analyses yielded nearly nominal coverage with better results under the log **transformation** of $\pi(a_1, a_2)$. When a **Weibull** model with a large shape parameter of 2 was correct, simulations showed that only the piecewise exponential analysis with a sufficient number of intervals achieved little or no bias as well as good coverage, while simpler models led to serious bias and consequent failure of coverage.

The **actuarial method** or **life table** method [23, 33, 39, 41, 59] is an approach that shares similarities with the piecewise exponential approach, although it was derived from a less parametric viewpoint. As with the piecewise exponential approach, time is split into intervals (indexed by i in this presentation). In each time interval i, the probability for an individual at risk at the beginning of the interval to survive the interval without developing c_1 is expressed as

$$S_i = \frac{\left(\frac{n_i - w_i}{2 - d_i}\right)}{\left(\frac{n_i - w_i}{2}\right)}, \tag{4}$$

where n_i denotes the number of subjects in the cohort at the beginning of interval i, d_i the number of events occurring in interval i, and w_i the number of subjects either lost to follow-up or developing c_2 (competing risks) in interval i. The actuarial approach is most appropriate when grouped data are available and the actual follow-up in each interval is not known. The person-years of follow-up for subjects lost to follow-up or developing c_2 in interval i is not used but, if one assumes that the **mean** withdrawal time occurs at the midpoint of the interval, then the denominator in (4) can be regarded as the effective number of persons at risk of developing the disease. That is, it represents the number of disease-free persons that would be expected to produce d_i events if all persons could be followed for the entire interval [38, 59, 66]. It can be regarded as a refinement of the simple cumulative method [59, 77] that ignores quantity w_i. Absolute risk is estimated by the cumulative (incidence) risk which, from the formulation in (4), is obtained as

$$1 - \prod_i S_i. \tag{5}$$

Since (5) ignores competing risks, the actuarial method applies in the absence of competing risks, which constitutes an important limitation, in an analogous manner as the density method. Moreover, as shown by several authors [33, 41], the actuarial method results in biased estimates of risks even in the unlikely and most favorable event (in terms of bias) of all withdrawals occurring at the interval midpoints. Alternative approaches based on different choices of the quantity to subtract from n_i (choices different from $w_i/2$) are not subject to less bias [38]. The problem can be improved best by using narrower intervals, but this is done at the expense of a larger **random error**.

Unlike the piecewise exponential models, the actuarial method does not require knowledge of follow-up time in each interval but only knowledge of the number at risk and the number of withdrawals. The piecewise exponential approach could, however, be used without knowledge of follow-up time by assigning a follow-up time of half the interval width to subjects who are lost to follow-up or who develop c_1 or c_2, in an analogous fashion as with the actuarial method [13]. The piecewise exponential approach has several advantages over the actuarial method. Bias is less of a problem with it, it takes competing risks into account, it applies naturally to open cohorts, and it extends easily to **regression**-based estimators (see below).

When individual follow-up times are all known, it is possible to estimate absolute risk nonparametrically as in Aalen [1], by substituting $\hat{\overline{G}}(t_1-)$, the right continuous **Kaplan–Meier estimate** [56] of surviving both c_1 and c_2 to time a_1 into the denominator of (1) and by replacing the numerator by $\sum \hat{\overline{G}}(t-)R^{-1}(t)$, where $R(t)$ is a left continuous process defining the number of subjects at risk just before t. The summation is over distinct times in $[a_1, a_2)$ at which events c_1 occur. The same estimator is discussed by Aalen & Johansen [2], Kay & Schumacher [57], Gray [49], Matthews [73], Keiding & Andersen [58], Benichou & Gail [13], and Korn & Dorey [60].

While nonparametric point estimates are easy to obtain, variance estimates are more complex and can be obtained in several ways. Results in Aalen [1, Theorem 2] can be used, as discussed in Benichou & Gail [13] and Korn & Dorey [60]. Alternatively, results in Aalen & Johansen [2, Theorem 4.3] can be used, as discussed by Keiding & Andersen [58]. Confidence intervals can then be obtained, based on

the log transformation, as suggested by Benichou & Gail [13] and Keiding & Andersen [58], or based on results of Dorey & Korn [34], who treat the lower and upper limit differently, a procedure that they claim is advantageous under heavy censoring.

Analytic and simulation results in Benichou & Gail [13] under exponential survival distributions show that the loss of efficiency of the nonparametric method is very small compared to a detailed piecewise exponential model and that nearly nominal coverage is obtained with the log transformation as for the piecewise exponential model. In simulations under a Weibull model with a large shape parameter of 2, very little bias and near nominal coverage was observed as with the piecewise exponential model [13]. These results suggest that properties of the piecewise exponential model and the nonparametric approach agree closely. The nonparametric approach does not make any assumption on the form of the hazards, but the piecewise constant assumption can be made less stringent by increasing the number of intervals. The piecewise exponential model has the advantage of simplicity of computation, in that it uses grouped data rather than individual data. Moreover, it is well suited to open cohorts.

These approaches yield an overall composite absolute risk and ignore covariates X. In order to obtain estimates that depend on the level of covariates, the cohort can be subdivided into subcohorts, and these approaches applied to resulting subcohorts defined by levels of X. This approach yields absolute risk estimates with low precision, however, if the subcohorts are small and have few events, as can happen if several risk factors have to be considered jointly (see [47] for further discussion and illustration, and [7] and [82] for further illustration with the actuarial method and breast cancer data). In order to remedy this problem, a natural approach is to model incidence rates h_1 and h_2 through regression models.

Covariate Models

Regression-based parametric methods are a direct extension of parametric methods for composite estimates. For instance, Benichou & Gail [13] studied exponential and piecewise exponential models. Under the piecewise exponential model, it is assumed that hazards for c_1 are products of a baseline hazard in interval i and a function of the covariates, usually (but not necessarily) expressed as $\exp(\beta^{\mathrm{T}} x)$. Baseline hazards as well as hazard ratio parameters β can be jointly estimated by maximizing the piecewise exponential **likelihood**. That likelihood is the same as that obtained by assuming that the number of events in each combination of time interval and level of X has a **Poisson distribution** with mean given by the product of the hazard times the corresponding number of person-years, that latter number being assumed constant [52, 61] (*see* **Poisson Regression in Epidemiology**). It is possible to include time by exposure **interactions** in covariates X so that the **proportional hazard** assumption is not required. Furthermore, hazards for cause c_2 are estimated separately. They are also assumed to be piecewise constant and can be assumed to depend on the set of covariates X, a different set X' if needed, or on no covariates. A point estimate of $\pi(a_1, a_2; x)$ is obtained by replacing quantities h_{1i} in (3) by quantities $h_{1i} \exp(\beta^{\mathrm{T}} x)$, where h_{1i} denotes the baseline hazard in the latter expression, and by plugging in **maximum likelihood** estimates of the parameters. Corresponding parameter estimates for competing hazards are estimated separately and also plugged in (3). A similar approach to point estimation can be taken for other parametric models such as a simple exponential model or a Weibull model [13].

As described in Benichou & Gail [13], variance estimates can be obtained by applying the delta method [87] and relying on the observed **information matrix** for all parametric models. Finite sample properties were studied by Benichou & Gail [13] through simulations based on a **clinical trial** of lung cancer [48]. Simulations used 392 patients, an accrual period of three years, and an additional follow-up of two years. Time to c_1 was assumed to be exponentially distributed and to depend on two covariates forming six joint levels, while time to c_2 was assumed to be exponential and not to depend on any covariates. Point estimates had little bias with piecewise exponential and exponential analyses. Variance estimates were also little biased and coverage was nearly nominal with all analyses except for the level of X with the fewest patients (12 patients) in which variance estimates and corresponding coverage were too small. Loss of efficiency could be appreciable when a detailed piecewise exponential was used compared to the simple exponential model.

Finally, a **semiparametric** estimator of absolute risk can be obtained, as outlined in Benichou & Gail [13]. The difference with the piecewise

exponential approach is that the hazard for c_1 is the product of an unspecified function of time (the baseline hazard) times a function of the covariates which is also usually of the form $\exp(\beta^T x)$ [31]. As for the piecewise exponential model, X may include time by exposure interaction and competing hazards can be assumed to depend on covariates X or X' or on no covariates (in the latter case, corresponding survival is estimated nonparametrically using the Kaplan–Meier product-limit estimator).

The expression for a semiparametric estimate of absolute risk is given in Benichou & Gail [13, formula (3.1)] and is a function of **partial likelihood** estimates [32] of hazard ratio parameters β and related **Nelson–Aalen estimates** of cumulative baseline hazards [6]. From results in Tsiatis [95] and Andersen & Gill [5] on the joint distribution of these parameter estimates, Benichou & Gail [13] derived an asymptotic variance estimator. No formal study of its finite sample properties has been undertaken.

These regression methods yield estimates of absolute risk with acceptable precision for several covariates. Regression-based methods are therefore well suited for individual prediction. Parametric or semiparametric approaches can be used. The piecewise exponential estimator seems to provide a good compromise between bias and precision, while being easy to implement both for open and closed cohorts.

Estimation from Population-based or Nested Case–Control Studies

Case–control studies provide data on the distributions of exposure respectively in diseased subjects (cases) and nondiseased subjects (controls) for the disease under study. These data are used to estimate hazard ratios or relative risks through the estimation of odds ratios, but are not sufficient to estimate exposure-specific incidence rates (the terms "hazard" and "incidence rate" are used indiscriminately in the remainder of the text) and absolute risks. In order to do so, case–control data have to be complemented by follow-up data. Either the cases and controls are selected from a follow-up study (*see* **Case–Control Study, Nested**) that provides either grouped data or individual data with survival-type information, or they are selected from a specified population in which an effort is made to identify all **incident cases** diagnosed during a fixed time interval (*see* **Case–Control Study, Population-based**) usually in a grouped form (number of cases and number of persons by age group). In both situations, full information on exposure is obtained only for cases and controls, but the complementary data provide information on composite incidence that can be combined with hazard ratio estimates to obtain exposure-specific incidence rates, as has long been recognized [29, 30, 68, 75, 76, 80].

The main estimation problem regards estimation of exposure- and age-specific hazards or incidence rates (age is the usual time scale in this context). Absolute risk estimates are then obtained from (1) or (2), and the delta-method [87] can be used to obtain the variance of absolute risk estimates based on the **covariance matrix** of incidence rate estimates. Parametric methods based on the piecewise exponential model (also termed the **Poisson regression** model) and the **logistic** model have been derived under a full likelihood approach, a **pseudo-likelihood** approach, and a hybrid approach. That latter approach will be described fully, because it has been used to obtain absolute risk estimates in practice. The former two approaches will be reviewed more briefly, because they have not yet been used to derive absolute risk estimates and fewer results are available. Finally, a semiparametric estimate of absolute risk based on partial likelihood has been proposed for nested case–control studies with time-matching of cases and controls, and will also be reviewed.

Parametric Approaches

The hybrid approach has been proposed by Gail et al. [47] as a multivariate extension of earlier work by Miettinen [75]. It relies on the possibility of estimating composite incidence rates h^*_{1i} from the population or follow-up data for each age group i or, in a more general fashion, for each stratum i defined by age and other factors observed in the follow-up or population data such as sex and region. Under a piecewise exponential assumption, the quantity h^*_{1i} is estimated by the ratio d_{1i}/t_i of the number of incident cases of disease c_1 to the number of person–years. Although information on exposure is obtained on cases and controls only, and not on the whole cohort or population, baseline incidence rates h_{1i} (for subjects at the baseline level of all exposure factors considered) can be obtained through the relationship [47, 75]:

$$h_{1i} = h^*_{1i}(1 - AR_i), \tag{6}$$

where AR_i is the attributable risk for disease c_1 in age group i or, more generally stratum i, for all exposure factors jointly, a quantity estimable from the case–control data. Gail et al. [47] suggested using the model-based approach of Bruzzi et al. [25], that incorporates **odds ratios** from **logistic regression**, for estimating attributable risk, and obtained a point estimate for h_{1i}. Upon multiplying that estimate by the corresponding odds ratio from logistic regression, they obtained an estimate of the incidence rate for each joint age and exposure level. Finally, incidence rates for competing risks can be obtained from the follow-up or population data, provided that those rates are assumed not to be influenced by the exposure factors for c_1. The latter assumption stems from the fact that it would be impossible to estimate hazard ratios for c_2 from case–control data for disease c_1. In fact, Gail et al. [47] used external data on national US mortality rates to estimate h_{2i} and obtained absolute risk estimates from the piecewise exponential model in formula (3).

Variance estimators are complex since incidence rate estimates involve odds ratio parameters obtained through logistic regression from the case–control data and counts of incident cases from the follow-up or population data. Estimators of variances and covariances of age- and exposure-specific incidence rates have been fully worked out by Benichou & Gail [14] for **simple random sampling**, **stratified random sampling**, **frequency matching** and individual **matching** in a simple setting. The approach relies on an extension of the delta-method to implicitly related **random variables** [12]. It takes into account all sources of variability; namely, the variance of hazard ratio estimates and of baseline incidence rate estimates, as well as the covariance between the two. Variance estimates of absolute risk estimates are then obtained through the delta-method [87] and take into account the variance of competing hazard estimates unless they are estimated from external sources and considered fixed, as in Gail et al. [47].

The hybrid approach can be regarded as relying on two models; namely, the piecewise exponential model and the logistic model (the **conditional logistic** model for individual matching and the unconditional logistic model for the three other ways of sampling controls). The baseline incidence rates are obtained by combining follow-up (or population) data and case–control data. Benichou & Gail [14] performed simulations based on the Breast Cancer Detection Demonstration Project (BCDDP) [9], a large follow-up study of 284 780 women, from which about 3000 cases and 3000 controls were selected (case–control study within a cohort or case–control study). They used a sample size of 100 000 women in each replication and generated piecewise exponentially distributed times to breast cancer occurrence by considering four age groups and two exposure factors forming six levels. A follow-up of five years was considered, and the possibility of dying from other causes (piecewise constant competing hazards not influenced by any covariates) was taken into account. Incident cases and frequency-matched controls were selected from the follow-up data. They found a small upward bias in absolute risk estimates due to the small upward bias incurred by using odds ratios to estimate hazard ratios when the rare disease assumption is violated in the context of such a study. Complete variance estimates had very little bias and yielded confidence intervals with near nominal coverage. Coverage was improved with the logit transform. Incomplete variance estimates that took into account only the variance of hazard ratio estimates from the case–control data were too small for small values of absolute risk, because they ignored the variances of baseline incidence rate estimates, and too large for larger values of absolute risk, because they ignored the negative covariances between hazard ratio estimates and baseline incidence rate estimates.

The hybrid approach was applied to the estimation of absolute risk of breast cancer from the BCDDP data as a function of age and four risk factors [47]. Details regarding variance estimation can be found in Benichou [11], who took into account special subsampling of cases and controls. Indeed, not all incident cases were used to estimate composite hazards and not all selected cases and controls were used to estimate hazard ratios. In order to implement these results and estimate absolute risk for new subjects, tables for point estimation were given by Gail et al. [47]. Practical implementation has been greatly facilitated by the development of the computer program RISK [10] and of graphs [17] that yield point estimates and confidence intervals of the absolute risk of developing breast cancer. Absolute risk is a widely used tool in individual counseling for breast cancer [17].

A pseudo-likelihood approach and a full likelihood approach have been proposed as alternatives to the hybrid approach [15]. They also rely on the piecewise exponential (or Poisson) model or logistic

model, although other parametric models could be used. They yield exposure-specific incidence rate estimates, but have not been fully developed to obtain absolute risk estimates, although this extension would be straightforward. Indeed, it would consist in (i) substituting in (1) or (2) age- and exposure-specific hazard estimates for disease c_1 and competing hazard estimates in order to obtain point estimates of absolute risk and (ii) using the delta-method [87] to derive variance estimates.

The pseudo-likelihood approach was presented by Benichou & Wacholder [15] in the context of a Poisson model (piecewise exponential model) and rests on the following principles. A full likelihood for the entire cohort or population could be written and maximized if information on exposure were available for all subjects in the population or cohort rather than just for the cases and controls. However, one can combine follow-up or population information in the form of number of events d_i and person-years t_i for each stratum i with the observed distribution of exposure in the case–control data to obtain estimates of number of events d_{1ij} and person-years t_{ij} for joint stratum level i and exposure level j. This is simply done by multiplying quantities d_{1i} (respectively t_i) by the observed proportion of cases (respectively controls) at exposure level j in stratum i. The rare disease assumption is used to obtain person-years from the conditional distribution of exposure in controls only. Substituting these estimated quantities, one obtains a Poisson pseudo-likelihood which is then maximized to obtain maximum pseudo-likelihood estimates of incidence rate parameters (baseline incidence rates and hazard ratios for a multiplicative model). Variance estimation relies on sandwich variance estimators [64] which allow for taking into account the additional component of variability incurred by the use of estimates of quantities d_{1ij} and t_{ij}.

The full likelihood approach differs from the pseudo-likelihood approach in that a full likelihood is written as a function of the incidence rate parameters to be estimated *and* a set of **nuisance parameters** for the conditional distribution of exposure given the stratum in the population. Rather than using the observed conditional distributions in cases and controls as with the pseudo-likelihood approach, the nuisance parameters are estimated jointly with the incidence rate parameters by maximization of the likelihood [15]. One obtains fully efficient maximum-likelihood estimates (rather than maximum pseudo-likelihood estimates) of all parameters (incidence rate and nuisance parameters), and variance estimates of the incidence rate parameters are obtained directly from the observed information matrix. In the context of a Poisson model, this approach is faced with the potential problem of a large number of parameters if several risk factors and stratum levels are considered. Even in the simple example of Benichou & Wacholder [15], with nine strata and eight exposure levels only, 60 nuisance parameters had to be estimated. This problem can be alleviated if one is willing to consider the logistic rather than the Poisson model, as pointed out by Greenland [50]. A prospective logistic model can be applied to the case–control data and yields maximum likelihood estimates of hazard ratio parameters. Furthermore, maximum likelihood estimates of baseline incidence rates are obtained by adding to the stratum parameter estimates from the logistic model a term corresponding to the logarithm of the ratio of sampling fractions among cases and controls in the stratum [50, 85]. The covariance matrix of estimates of baseline incidence rates and hazard ratios is obtained as described in Prentice & Pyke [85].

Upon comparing the pseudo-likelihood, full likelihood and hybrid approach on population-based case-control data of bladder cancer [51], Benichou & Wacholder [15] found that the hybrid approach seemed to be less efficient for incidence rate estimation than the other two approaches, which were themselves equally efficient. This efficiency loss might be due to the following conceptual difference regarding estimation of baseline incidence rates and hazard ratios among the three approaches. With the maximum likelihood and pseudo-likelihood approaches, these quantities are jointly estimated and their negative correlations fully accounted for in variance estimates. With the hybrid approach, crude incidence rates and hazard ratios are estimated separately and then combined to obtain stratum- and exposure-specific incidence rates and, as a consequence, negative correlations between estimates of baseline incidence rates and hazard ratios are not as strong, which results in larger variances [15]. Another potential advantage of the full likelihood and pseudo-likelihood approaches is that they directly estimate hazard ratios rather than odds ratios. Furthermore, if the Poisson (but not the logistic) model is used, they can be applied to more general models of risk; for example, models with an additive form using

rate difference parameters rather than hazard ratio parameters [15]. Finally, all three approaches require that cases and controls be selected at random [67] and that incident cases or at least a known proportion of them be fully identified [15].

Semiparametric Approach

The three parametric approaches described above apply to situations in which controls are not individually matched to cases. The hybrid approach can handle special cases of individual matching [14] but not time-matching, which characterizes **nested case–control studies** [24, 65, 71]. In that context, Langholz & Borgan [62] developed a semiparametric approach which can be regarded as an extension of the semiparametric approach for cohort studies described above (see the section "Estimation from Cohort Studies" above). The context is that of a nested case–control study (case–control within a cohort), in which cases develop from a cohort, and controls are selected from subjects still at risk. Therefore, individual follow-up times are needed and grouped data are not sufficient.

Incidence rates are expressed as the product of baseline incidence rates of an unspecified form times a function of the covariates representing the hazard ratio [31]. Hazard ratio parameter estimates are obtained from maximizing the partial likelihood of the **Cox regression model** for nested case–control data [81, 84]. Absolute risk estimates are obtained by combining partial likelihood hazard ratio parameter estimates and corresponding cumulative hazard estimates. Langholz & Borgan [62] showed that their proposed semiparametric estimate is asymptotically **normal** and provided a variance estimator based on results in Aalen & Johansen [2], Andersen et al. [6], and Borgan et al. [21]. Point estimates and corresponding variance estimates are based on simple sums or products of information from the case–control study, the estimated hazard ratio parameters, and the number at risk at the failure times. Competing risks can be taken into account provided that it is assumed that occurrence of c_2 is not influenced by the risk factors for occurrence of disease c_1. Finite sample properties of this approach have not been studied. A direct comparison with parametric approaches presented above is not possible because the semiparametric approach applies only to time-matched data, which the parametric approaches cannot handle. The semiparametric approach requires observation of individual follow-up time of each subject in the original cohort in order to form the risk sets for each failure time, and enable control selection. It is therefore potentially less widely applicable than the parametric approaches but makes no assumption on the baseline hazard. Finally, it has the advantage over the available parametric approaches of being able to handle continuous covariates.

Special Problems

Case–Cohort and Cross-sectional Designs

In the **case–cohort design**, information on exposure is gathered only in a subcohort of subjects randomly selected from the original cohort and among subjects who develop the disease [83]. It is therefore possible to estimate exposure-specific incidence rates and absolute risk directly from case–cohort data. However, the details of absolute risk estimation have not been worked out in the literature. **Cross-sectional studies** would need to be complemented by follow-up or population data in order to allow for incidence rate and absolute risk estimation, but such designs do not seem to have been implemented (*see* **Case–Control Study, Prevalent**).

Two-stage Case–Control Studies

In **two-phase case–control studies** [22, 98, 99], cases and controls are selected from a cohort or a population, as in a case–control study within a cohort or a population-based case–control study. Furthermore, a nested subsample of cases and controls is selected from original cases and controls on which information is gathered on exposure factors which are more difficult to obtain, such as X-ray data or **genetic markers**. Several parametric approaches have been developed to allow for hazard ratio and incidence rate estimation by an extension of the pseudo-likelihood approach for two-stage case–control data [16], pseudo-conditional likelihood methods [22, 90], and weighted likelihood methods [40, 54, 88, 89]. From incidence rate estimates from these various methods, it would be easy to obtain absolute risk estimates from (1) or (2).

Continuous Risk Factors

Absolute risk can be expressed as a function of both continuous and categorical risk factors. Model-based estimation methods presented above for cohort data (see the section "Estimation from Cohort Data" above) accommodate both types of variables. For case–control data however, the situation is different. Among parametric approaches, the hybrid approach yields point estimates that apply to both types of variables, but variance estimators have been developed only for categorical covariates. The full likelihood and pseudo-likelihood approach only apply to categorical covariates. The semiparametric approach is more flexible, in that it fully allows for continuous risk factors for point and variance estimation.

Time-dependent Risk Factors

Most estimation procedures presented above can be adapted to take into account **time-dependent covariates**. However, when absolute risk is used for individual prediction, estimation of absolute risk over time interval $[a_1, a_2)$ is based on the initial value of the covariates (i.e. the value at time a_1) and assumes that it stays constant over the whole interval, unless it is possible to predict (in a probabilistic or deterministic manner) the future variation of covariates over time (see the opening text).

Secular Trend

An important feature of the estimation methods described for cohort and case–control studies is that, by combining hazard estimates from different age intervals, absolute risk can be estimated for a much longer age interval than the actual follow-up of individuals in the study. To combine these hazard estimates into a single estimate of absolute risk, one must assume that there is no secular trend in disease incidence [59, Chapter 6].

Misclassification of Exposure

Misclassification of exposure could affect the validity of absolute risk estimates, but this problem, which has been studied for estimation of other measures (e.g. odds ratio, hazard ratio, and population attributable risk; *see* **Measurement Error in Epidemiologic Studies**) has not been studied for absolute risk estimation.

Use of Two Time Scales

In some applications, it may be important to consider two time scales, such as time from entry in the cohort (e.g. time from surgery, diagnosis, or first exposure) and age. Korn & Dorey [60] give guidelines and examples for that situation.

Selection of Risk Factors and Model Misspecification

Selection of risk factors on which to base absolute risk estimation is a difficult task. Complex multivariate models containing many risk factors will usually appear to describe the variation of risk in the data used to fit the model better than simpler models. Yet the simpler models often perform as well or better in predicting risk in other populations [37]. This is because complex models fit the statistical anomalies of the given sample as well as the reproducible features, whereas the simpler models tend to reflect the reproducible features only. It might therefore be preferable to choose factors for inclusion in the model that have been previously demonstrated to be important rather than to rely solely on the current data sample to select factors for inclusion [44].

A related problem is model **misspecification** which can lead to severe bias in absolute risk estimates and has to be considered carefully. Model misspecification can come from an inappropriate selection of risk factors, but also from incorrectly modeling the effect of included risk factors, from selecting the wrong model for time to event distribution, or from incorrectly assuming proportional hazards. Benichou & Gail [13] illustrate the potential severity of the problem in an example which suggests that using unsaturated rather than saturated models for covariate effects can lead to a **systematic error** that is potentially larger than **random error** (*see* **Generalized Linear Model**).

Validation

Given the potentially severe effects of model misspecification on absolute risk estimation, it is important to validate models used for absolute risk estimation. For instance, from internal validation results and two studies of external validation based on independent cohorts [20, 94] (*see* **Validation Study**), it appeared that the model developed by

Gail et al. [47] to estimate absolute risk of breast cancer from the BCDDP as a function of age and four risk factors produces valid estimates of absolute risk for women in regular screening as in the BCDDP, but yields estimates that tend to be too high when applied to unscreened or sporadically screened populations [20, 44, 45, 94], as had been cautioned in the initial paper [47].

Absolute Risk and Treatment Comparison

It might be useful to use absolute risk as a means of testing for treatment effect, especially given the availability of tests for comparing k treatment groups based on absolute risk [49]. However, use of absolute risk alone may be misleading. For example, if a cancer treatment increases h_2 but leaves h_1 unaffected, absolute risk will diminish in the treated group; yet overall survival is reduced and c_1-specific survival is unchanged. Instead, one should compare overall survival and estimates of the cause-specific survival curves in the treated and untreated groups, as is common practice. If h_2 is not affected by treatment, however, the change in absolute risk is a more realistic gauge of treatment benefit than a comparison of c_1-specific survival curves. If both h_1 and h_2 are affected, absolute risk gives useful descriptive information for summarizing the burden of recurrence in each group [13].

Overall Adjusted Absolute Risk

In order to obtain an overall measure of absolute risk at the population level, one might combine individualized estimates to obtain a direct adjusted value for the entire population by summing estimated values of absolute risk for a given level of the covariates over the distribution of the covariates in the reference population [13]. This procedure would yield a different estimate than that obtained by covariate-free estimation of absolute risk from the same population (see the section "Estimation from Cohort Studies" above). The adjusted procedure would be analogous to the methods for direct adjustment of survival curves described by Murphy & Haywood [79], Makuch [70], and Chang et al. [26], and the variance estimation methods of Gail & Byar [46] could be adapted.

Related Quantities

Attack Rate

In the investigation of a local outbreak of a **communicable disease**, a measure of interest is the absolute risk of developing the disease for the duration of the epidemic or the time during which primary cases occur. In this situation, absolute risk is often called an attack rate [59, 69].

Hazard Ratio and Relative Risk

As discussed above (see the section "Interpretation and Usefulness"), the hazard ratio, also called the relative rate, rate ratio, incidence density ratio, or instantaneous relative risk, is a useful measure in etiologic research that quantifies the strength of the relationship between exposure and disease, while absolute risk is more useful in individual prediction as a measure of the actual probability of disease for a given risk profile. Large hazard ratios may correspond to small absolute risks if the disease is rare and conversely.

Since incidence rates are a function of hazard ratios in multiplicative models, absolute risk is also a function of hazard ratios (and of baseline incidence rates) in those models. Alternatively, additive models can be used with the rate difference, also called hazard difference or incidence density difference, being the relevant parameter instead of the hazard ratio to measure the effect of covariates.

The term "relative risk" is frequently used to represent a hazard ratio or its estimator. Strictly speaking, however, relative risk refers to the ratio of absolute risks and not of incidence rates [59]. A synonym is "risk ratio" [76].

Incidence Rate

Absolute risk is a direct function of incidence rates, as is apparent from (1) and (2) that define absolute risk. As was mentioned above (see the sections "Estimability" and "Estimation" above), the problems of absolute risk estimability and estimation essentially reduce to those of incidence rate estimability and estimation.

Cumulative Risk and Cumulative Hazard

The relationships between absolute risk and cumulative risk and hazard have been defined in the section "Estimation from Cohort Studies" above.

Excess Risk

Excess risk [91], also called excess incidence [18, 69, 74], is defined as the difference between the incidence rates in the exposed and the unexposed. Like absolute risk, it takes into account the incidence of the disease in the unexposed and the strength of the association between exposure and disease. It can be expressed as the product of the baseline incidence rate times the hazard ratio minus 1, and it quantifies the difference in incidence that can be attributed to exposure at the individual level. Other terms have been used to denote this quantity; namely, "Berkson's simple difference" [96], "incidence density difference" [76], "excess prevalence" [96], and even "attributable risk" [72, 91].

Population Attributable Risk and Generalized Impact Fraction

Population **attributable risk** [63] and the generalized impact fraction [97] are measures that assess the public health consequences of an association between exposure and disease and the potential impact of prevention measures aimed at eliminating (population attributable risk) or reducing (generalized impact fraction) exposure in the population. As was mentioned above (see the section "Interpretation and Usefulness" above), absolute risk can be used to estimate the absolute reduction in incidence that would result from prevention measures in each subgroup of exposure, and can therefore be regarded as a useful complement to population attributable risk and the generalized impact fraction.

Floating Absolute Risk

The term "floating absolute risk", introduced by Easton et al. [36], refers to a concept unrelated to absolute risk, which may introduce some confusion. The purpose of those authors was to remedy the standard problem that hazard ratios are estimated in reference to a baseline group which in turn causes hazard ratio estimates for different levels of exposure to be correlated and may lead to lack of precision in hazard ratio estimates if the baseline group is small. The authors proposed a procedure to obtain hazard ratio estimates unaffected by these problems. They termed their proposed hazard ratio estimates "floating absolute risks" to indicate that their **standard errors** were not estimated in reference to an arbitrary baseline group.

Prospects and Conclusions

Despite the substantial development of methods for estimating absolute risk, there remain important research issues, including point and variance estimation for parametric case–control estimators when continuous risk factors are considered, the study of finite sample properties of nonparametric and semiparametric estimators and their comparison with parametric estimators, the comparison of the three main parametric approaches in case–control studies, the study of the effect of exposure misclassification on absolute risk estimation, and research issues regarding special problems (see the section "Special Problems" above).

An important issue is the development of tools to implement methods for absolute risk estimation. For instance, a graphic approach has been developed to convert relative to absolute risk [35]. Graphs [17] and a computer program [10] have been developed to estimate absolute risk of breast cancer as a function of age and four risk factors. More general programs would be worth developing.

Finally, an important challenge is to increase awareness of the proper interpretation and use of absolute risk in practice (e.g. in counseling, see [4, 17, 19, 53] and [78]), as well as of correct estimation techniques.

References

[1] Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models, *Annals of Statistics* **6**, 534–545.

[2] Aalen, O. & Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics* **5**, 141–150.

[3] American Cancer Society (1992). *Cancer Facts and Figures*. ACS, Atlanta.

[4] American Society of Human Genetics Ad Hoc Committee (1994). Statement of the American Society of Human

Genetics on Genetic Testing for Breast and Ovarian Cancer Predisposition, *American Journal of Human Genetics* **55**, i–iv.

[5] Andersen, P.K. & Gill, R.D. (1982). Cox's regression models for counting processes: a large-sample study, *Annals of Statistics* **4**, 1100–1120.

[6] Andersen, P.K., Borgan, Ø, Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[7] Anderson, D.E. & Badzioch, M.D. (1985). Risk of familial breast cancer, *Cancer* **56**, 383–387.

[8] Anderson, S.J., Ahnn, S. & Duff, K. (1992). *NSABP Breast Cancer Prevention Trial Risk Assessment Program, Version 2*, University of Pittsburgh Department of Biostatistics, Pittsburgh.

[9] Baker, L.H. (1982). Breast cancer detection demonstration project: five-year summary report, *CA Cancer Journal for Clinicians* **32**, 194–225.

[10] Benichou, J. (1993). A computer program for estimating individualized probabilities of breast cancer, *Computers and Biomedical Research* **26**, 373–382.

[11] Benichou, J. (1995). A complete analysis of variability of absolute risk from a population-based case–control study on breast cancer, *Biometrical Journal* **37**, 3–24.

[12] Benichou, J. & Gail, M.H. (1989). A delta-method for implicitly defined random variables, *American Statistician* **43**, 41–44.

[13] Benichou, J. & Gail, M.H. (1990). Estimates of absolute cause-specific risk in cohort studies, *Biometrics* **46**, 813–826.

[14] Benichou, J. & Gail, M.H. (1995). Methods of inference for estimates of absolute risk derived from population-based case–control studies, *Biometrics* **51**, 182–194.

[15] Benichou, J. & Wacholder, S. (1994). A comparison of three approaches to estimate exposure-specific incidence rates from population-based case–control data, *Statistics in Medicine* **13**, 651–661.

[16] Benichou, J., Byrne, C. & Gail, M.H. (1997). An approach to estimating exposure-specific rates of breast cancer from a two-stage case–control study within a cohort, *Statistics in Medicine* **16**, 133–151.

[17] Benichou, J., Gail, M.H. & Mulvihill, J.J. (1996). Graphs to estimate an individualized risk of breast cancer, *Journal of Clinical Oncology* **14**, 103–110.

[18] Berkson, J. (1958). Smoking and lung cancer: some observations on two recent reports, *Journal of the American Statistical Association* **53**, 28–38.

[19] Biesecker, B.B., Boehnke, M., Calzone, K., Markel, D.S., Garber, J.E., Collins, F.S. & Weber, B.L. (1993). Genetic counseling for families with inherited susceptibility to breast and ovarian cancer, *Journal of the American Medical Association* **269**, 1970–1974.

[20] Bondy, M.L., Lustbader, E.D., Halabi, S., Ross, E. & Vogel, V.G. (1994). Validation of a breast cancer risk assessment model in women with a positive family history, *Journal of the National Cancer Institute* **86**, 620–625.

[21] Borgan, Ø., Goldstein, L. & Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model, *Annals of Statistics* **23**, 1749–1778.

[22] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case–control data, *Biometrika* **75**, 11–20.

[23] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*. IARC Scientific Publications 82, Lyon.

[24] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[25] Bruzzi, P., Green, S.B., Byar, D.P., Brinton, L.A. & Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case–control data, *American Journal of Epidemiology* **122**, 904–914.

[26] Chang, I., Gelman, R. & Pagano, M. (1982). Corrected group prognostic curves and summary statistics, *Journal of Chronic Diseases* **35**, 669–674.

[27] Chiang, C.L. (1961). A stochastic study of the life table and its applications: III. The follow-up study with the consideration of competing risks, *Biometrics* **17**, 57–58.

[28] Chiang, C.L. (1968). *Introduction to Stochastic Processes in Biostatistics*. Wiley, New York.

[29] Cornfield, J. (1951). A method for estimating comparative rates from clinical data: applications to cancer of the lung, breast and cervix, *Journal of the National Cancer Institute* **11**, 1269–1275.

[30] Cornfield, J. (1956). A statistical problem arising from retrospective studies, in *Proceedings of the Third Berkeley Symposium*, Vol. IV, J. Neyman, ed. University of California Press, Monterey, pp. 133–148.

[31] Cox, D.R. (1972). Regression models and lifetables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[32] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[33] Cutler, S.J. & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival, *Journal of Chronic Diseases* **8**, 699–712.

[34] Dorey, F.J. & Korn, E.L. (1987). Effective sample sizes for confidence intervals for survival probabilities, *Statistics in Medicine* **6**, 679–687.

[35] Dupont, D.W. (1989). Converting relative risks to absolute risks: a graphical approach, *Statistics in Medicine* **8**, 641–651.

[36] Easton, D.F., Peto, J. & Babiker, A.G. (1991). Floating absolute risk: an alternative to relative risk in survival and case–control analysis avoiding an arbitrary reference group, *Statistics in Medicine* **10**, 1025–1035.

[37] Efron, B. (1986). How biased is the apparent error rate of a prediction rule?, *Journal of the American Statistical Association* **81**, 461–470.

[38] Elandt-Johnson, R.C. (1977). Various estimators of conditional probabilities of death in follow-up studies. Summary of results, *Journal of Chronic Diseases* **30**, 247–256.

[39] Elveback, L. (1958). Estimation of survivorship in chronic disease: the "actuarial" method, *Journal of the American Statistical Association* **53**, 420–440.

[40] Flanders, W.D. & Greenland, S. (1991). Analytic methods for two-stage case–control studies and other stratified designs, *Statistics in Medicine* **10**, 739–747.

[41] Fleiss, J.L., Dunner, D.L., Stallone, F. & Fieve, R.R. (1976). The life table: a method for analyzing longitudinal studies, *Archives of General Psychiatry* **33**, 107–112.

[42] Gail, M.H. (1975). Measuring the benefit of reduced exposure to environmental carcinogens, *Journal of Chronic Diseases* **28**, 135–147.

[43] Gail, M.H. (1975). A review and critique of some models used in competing risk analysis, *Biometrics* **31**, 209–222.

[44] Gail, M.H. & Benichou, J. (1992). Assessing the risk of breast cancer in individuals, in *Cancer Prevention*, V.T. De Vita & S.A. Rosenberg, eds. Lippincott, Philadelphia, pp. 1–15.

[45] Gail, M.H. & Benichou, J. (1994). Validation studies on a model for breast cancer risk (editorial), *Journal of the National Cancer Institute* **86**, 573–575.

[46] Gail, M.H. & Byar, D.P. (1986). Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect, *Biometrical Journal* **28**, 587–599.

[47] Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C. & Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, *Journal of the National Cancer Institute* **81**, 1879–1886.

[48] Gail, M.H., Eagan, R.T., Feld, R., Ginsberg, R., Godell, B., Hill, L., Holmes, E.C., Lubeman, J.M., Mountain, C.F., Oldham, R.K., Pearson, F.G., Wright, P.W., Lake, W.H., and the Lung Cancer Study Group (1984). Prognostic factors in patients with resected stage I non-small-cell lung cancer. A report from the Lung Cancer Study Group, *Cancer* **54**, 1802–1813.

[49] Gray, R.J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk, *Annals of Statistics* **16**, 1141–1151.

[50] Greenland, S. (1981). Multivariate estimation of exposure-specific incidence from case–control studies, *Journal of Chronic Diseases* **34**, 445–453.

[51] Hartge, P., Cahill, J.J., West, D., Hauck, M., Austin, D., Silverman, D. & Hoover, R. (1985). Design and methods in a multicenter case–control interview study, *American Journal of Public Health* **74**, 52–56.

[52] Holford, T.R. (1980). The analysis of rates and of survivorship using log-linear models, *Biometrics* **36**, 299–305.

[53] Hoskins, K.F., Stopfer, J.E., Calzone, K.A., Merajver, S.D., Rebbeck, T.R., Garber, J.E. & Weber, B.L. (1995). Assessment and counseling for women with a family history of breast cancer: a guide for clinicians, *Journal of the American Medical Association* **273**, 577–585.

[54] Kalbfleisch, J.D. & Lawless, J.F. (1988). Likelihood analysis of multi-stage models for disease incidence and mortality, *Statistics in Medicine* **7**, 149–160.

[55] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[56] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[57] Kay, R. & Schumacher, M. (1983). Unbiased assessment of treatment effects on disease recurrence and survival in clinical trials, *Statistics in Medicine* **2**, 41–58.

[58] Keiding, N. & Andersen, P.K. (1989). Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process, *Applied Statistics* **38**, 319–329.

[59] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont.

[60] Korn, E.L. & Dorey, F.J. (1992). Applications of crude incidence curves, *Statistics in Medicine* **11**, 813–829.

[61] Laird, N. & Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques, *Journal of the American Statistical Association* **76**, 231–240.

[62] Langholz, B. & Borgan, Ø. (1997). Estimation of absolute risk from nested case–control data, *Biometrics* **53**, 767–774.

[63] Levin, M.L. (1953). The occurrence of lung cancer in man, *Acta Unio Internationalis Contra Cancrum* **9**, 531–541.

[64] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[65] Liddell, J.C., McDonald, J.C. & Thomas, D.C. (1977). Methods of cohort analysis: appraisal by application to asbestos mining (with discussion), *Journal of the Royal Statistical Society, Series A* **140**, 469–491.

[66] Littell, A.S. (1952). Estimation of the t-year survival rate from follow-up studies over a limited period of time, *Human Biology* **24**, 87–116.

[67] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[68] MacMahon, B. (1962). Prenatal X-ray exposure and childhood cancer, *Journal of the National Cancer Institute* **28**, 1173–1191.

[69] MacMahon, B. & Pugh, T.F. (1970). *Epidemiology: Principles and Methods*. Little, Brown & Company, Boston.

[70] Makuch, R.W. (1982). Adjusted survival curve estimation using covariates, *Journal of Chronic Diseases* **35**, 437–443.

[71] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics* **29**, 479–486.

[72] Markush, R.E. (1977). Levin's attributable risk statistic for analytic studies and vital statistics, *American Journal of Epidemiology* **105**, 401–406.

[73] Matthews, D.E. (1988). Likelihood-based confidence intervals for functions of many parameters, *Biometrika* **75**, 139–144.

[74] Mausner, J.S. & Bahn, A.K. (1974). *Epidemiology: An Introductory Text*. W.B. Saunders, Philadelphia.

[75] Miettinen, O.S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention, *American Journal of Epidemiology* **99**, 325–332.

[76] Miettinen, O.S. (1976). Estimability and estimation in case–referent studies, *American Journal of Epidemiology* **103**, 226–235.

[77] Morgenstern, H., Kleinbaum, D.G. & Kupper, L.L. (1980). Measures of disease incidence used in epidemiologic research, *International Journal of Epidemiology* **9**, 97–104.

[78] Mulvihill, J.J., Safyer, A.W. & Bening, J.K. (1982). Prevention in familial breast cancer: counseling and prophylactic mastectomy, *Preventive Medicine* **11**, 500–511.

[79] Murphy, V.K. & Haywood, L.J. (1981). Survival analysis by sex, age group and hemotype in sickle cell disease, *Journal of Chronic Diseases* **34**, 313–319.

[80] Neutra, R.R. & Drolette, M.E. (1978). Estimating exposure-specific disease rates from case–control studies using Bayes' theorem, *American Journal of Epidemiology* **108**, 214–222.

[81] Oakes, D. (1981). Survival times: aspects of partial likelihood (with discussion), *International Statistical Review* **49**, 235–264.

[82] Ottman, R., King, M.C., Pike, M.C. & Henderson, B.E. (1983). Practical guide for estimating risk for familial breast cancer, *Lancet* **2**, 556–558.

[83] Prentice, R.L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[84] Prentice, R.L. & Breslow, N.E. (1978). Retrospective studies and failure time models, *Biometrika* **65**, 153–158.

[85] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case–control studies, *Biometrika* **66**, 403–411.

[86] Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Flournoy, N., Farewell, V.T. & Breslow, N.E. (1978). The analysis of failure times in the presence of competing risks, *Biometrics* **34**, 541–554.

[87] Rao, C.R. (1965). *Linear Statistical Inference and Its Application*. Wiley, New York, pp. 319–322.

[88] Reilly, M. & Pepe, M.S. (1995). A mean score method for missing and auxiliary covariate data in regression models, *Biometrika* **82**, 299–314.

[89] Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* **89**, 846–866.

[90] Schill, W., Jöckel, K-H., Drescher, K. & Timm, J. (1993). Logistic analysis in case–control studies under validation sampling, *Biometrika* **80**, 339–352.

[91] Schlesselman, J.J. (1982). *Case–Control Studies: Design, Conduct and Analysis*. Oxford University Press, New York.

[92] Schrag, D., Kuntz, K.M., Garber, J.E. & Weeks, J.C. (1997). Decision analysis – effects of prophylactic mastectomy and oophorectomy on life expectancy among women with *BRCA1* or *BRCA2* mutations, *New England Journal of Medicine* **336**, 1465–1471.

[93] Smigel, K. (1992). Breast cancer prevention trial takes off, *Journal of the National Cancer Institute* **84**, 669–670.

[94] Spiegelman, D., Colditz, G.A., Hunter, D. & Hertzmark, E. (1994). Validation of the Gail et al. model for predicting individual breast cancer risk, *Journal of the National Cancer Institute* **86**, 600–607.

[95] Tsiatis, A.A. (1981). A large-sample study of Cox's regression model, *Annals of Statistics* **9**, 93–108.

[96] Walter, S.D. (1976). The estimation and interpretation of attributable risk in health research, *Biometrics* **32**, 829–849.

[97] Walter, S.D. (1980). Prevention for multifactorial diseases, *American Journal of Epidemiology* **112**, 409–416.

[98] Weinberg, C.R. & Wacholder, S. (1990). The design and analysis of case–control studies with biased sampling, *Biometrics* **46**, 963–975.

[99] White, J.E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

JACQUES BENICHOU

Accelerated Failure-time Models

Accelerated failure-time models can be simply illustrated in the following way. Let T_0 be the survival time, under control conditions, from some origin to the occurrence of an event of interest, and suppose that application of a treatment, or exposure to a risk factor, modifies the survival time to $T = T_0/\theta$ for some fixed scaling parameter θ. Then the **median survival time** under the treatment or risk factor is $1/\theta$ times the median under the control, and indeed the time to reach *any* percentile of the treatment group will be $1/\theta$ times the time to reach the corresponding percentile of the controls. This proportional adjustment of the time-scale represents the simplest form of the accelerated failure-time assumption.

The term *accelerated failure time* derives from accelerated life testing, particularly in engineering and similar applications. In these, extrapolation is often required from high stress levels, designed to induce rapid failure under laboratory conditions, to lower stress levels which operate under normal conditions. The assumed link between the effects of the different levels is provided by the adjusted time-scale. Corresponding biostatistical applications include situations such as carcinogenicity or toxicity experiments, in which doses of a high level are applied under experimental conditions and the results extrapolated to lower doses via the accelerated failure-time assumption (*see* **Extrapolation, Low Dose**).

The Models

Let the proportion of cases in the control group surviving beyond time t be denoted by $S_0(t)$ (the *survival function*), and let the survival function under the treatment or exposure be $S(t)$. Then according to the accelerated failure-time model the two survival functions are related by

$$S(t) = S_0(\theta t),$$

and the *hazard functions* (*see* **Hazard Rate**) by

$$\lambda(t) = \theta\lambda_0(\theta t)$$

(*see* **Survival Distributions and Their Characteristics**). Under this assumption,

$$\Pr(\log T > t - \log\theta) = S_0(\mathrm{e}^t),$$

giving a location shift model on the log scale, namely

$$\log T = \beta_0 + \log\theta + \varepsilon,$$

where ε is a zero mean **residual**.

More general models are obtained by incorporating **covariates** or **explanatory variables** into θ. If $\mathbf{x}$ is a vector of covariates associated with an individual, then the survival function, given $\mathbf{x}$, is assumed to be of the form

$$S(t|\mathbf{x}) = S_0[\theta(\mathbf{x})t]$$

for an underlying survival function S_0 and function $\theta(\cdot)$, with hazard function

$$\lambda(t|\mathbf{x}) = \theta(\mathbf{x})\lambda_0[\theta(\mathbf{x})t].$$

Correspondingly, $\log T = \beta_0 + \log\theta(\mathbf{x}) + \varepsilon$. A particularly useful form is the **loglinear** regression model in which $\theta(\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x})$, which leads to the linear model $\log T = \beta_0 + \boldsymbol{\beta}'\mathbf{x} + \varepsilon$. Inferences concerning the regression parameters and the ways in which they influence survival can therefore be made using log survival times and **linear regression** methods, including methods that allow for **censored** survival times, where the survival time may be known only to exceed or be smaller than a given value. This may result from loss to follow-up, withdrawal for causes unrelated to the end point of interest (*see* **Competing Risks**), survival beyond the end of a trial, and so on.

Parametric Models

Parametric models under the accelerated failure-time assumption are obtained by specifying the underlying distribution S_0 and the form of dependence on $\mathbf{x}$ through $\theta(\cdot)$. Some important special cases of the underlying distribution include the *Weibull*, when $S_0(t) = \exp(-kt^\alpha)$, the **lognormal**, when $\log T$ has a normal distribution, and the *log-logistic*, when $S_0(t) = 1/(1 + kt^\alpha)$. The last model has received considerable attention, since one is often interested in the probability of survival beyond a fixed time (e.g.

five-year survival rates). If $\theta(\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x})$, then the logistic transform or log **odds ratio** is

$$\log\left\{\frac{S(t|\mathbf{x})}{[1-S(t|\mathbf{x})]}\right\} = -\log k - \alpha \log t - \alpha\boldsymbol{\beta}'\mathbf{x},$$

which is linear in the covariates $\mathbf{x}$, and for fixed t represents the familiar linear **logistic regression** model.

Estimation and Analysis

If we assume a fully parametric form for the accelerated failure-time model, then standard parametric inference procedures such as the use of **likelihood** methods are applicable, allowing for the possible presence of censored observations. With right-censored failure times and i indexing the individuals, provided the censoring and survival mechanisms are independent, the likelihood function is given by

$$l = \prod_i \lambda(y_i|\mathbf{x}_i)^{\delta_i} S(y_i|\mathbf{x}_i),$$

where y_i is the observed failure time or the time at which censoring occurs for the ith individual and δ_i is an indicator of censoring taking the value 1 if the failure time is observed and 0 if censored. Large-sample estimates of standard errors can be obtained from the observed Fisher **information**, since the presence of censoring will generally preclude the taking of expectations.

The Expectation-Maximization **(EM) algorithm** of Dempster et al. [6] often provides a convenient method of maximizing the likelihood with censored data (e.g. [5, Chapter 11]). This method is particularly useful if the distribution of $Z_i = \log T_i$ is a member of the regular **exponential family** in the mean parameter μ_i with variance V_i. Then the likelihood equations corresponding to derivatives of the log likelihood with respect to the parameters β_j in the mean take the simple form

$$\sum_i \frac{\tilde{z}_i - \mu_i}{V_i}\frac{\partial \mu_i}{\partial \beta_j} = 0,$$

where $\tilde{z}_i$ is z_i if the failure time is observed or $\mathrm{E}(Z_i|T_i > y_i)$ if the failure time T_i is censored at y_i. This is of the same form as the likelihood equations when all data are uncensored but with the censored values replaced by their conditional means. The E-step in the EM algorithm thus consists of replacing the censored responses by their estimated conditional means given the existing parameter estimates and the time at which censoring occurs, while the M-step corresponds to parameter updating by solving the likelihood equations treating the estimated values as if they were uncensored. The process is iterated until convergence. In general there will also be other parameters involved in the model which need to be estimated. For a discussion of the EM approach with (log)-normal responses see Aitkin [1].

Buckley & James [2] adopted a similar approach to deriving a semiparametric procedure in which the residual distribution remains unspecified. Suppose that the residuals ε_i are independent with common distribution and that $\mu_i = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i$. The likelihood equations then become

$$\sum_i (\tilde{z}_i - \beta_0 - \boldsymbol{\beta}'\mathbf{x}_i)x_{ij} = 0.$$

In the method of Buckley & James the conditional expectations for the censored responses are replaced in the equations by their estimates based on the **Kaplan–Meier** product-limit estimator of the residual distribution. An iterative estimation scheme analogous to the EM procedure therefore consists of starting with estimates of the β_j, obtaining the Kaplan–Meier residual distribution, replacing the censored responses by their estimated conditional means using the estimated residual distribution, and solving the normal equations, assuming these were the true responses, to update the parameter estimates. Some modifications are needed to account for the possibility of the Kaplan–Meier means being undefined when the largest residual is censored, and this will typically introduce some biases into the intercept estimates for small samples.

Unlike the fully parametric EM algorithm this iterative scheme need not converge, nor need the estimating equations have a unique nor exact solution due to discontinuities and nonmonotonicity. In these cases zero crossings or values closest to zero can be used. Extensions to **nonlinear regressions** or M-estimators [16] (*see* **Robustness**) are conceptually straightforward.

Whilst the method is simple to describe, obtaining theoretic properties has proved difficult, in part due to the issues of censored data in the right-hand tail of the distribution. Asymptotic properties have been

obtained under some conditions by Ritov [16] and Lai & Ying [10], who introduce a smooth weighting function to overcome instability due to censorship. Practical issues in the estimation of standard errors have been addressed by Weissfeld & Schneider [27], Smith [21], and Lin & Wei [11]. Approximations based on imputation via the data augmentation algorithm were proposed by Wei & Tanner [24] but, as noted by James [7], they do not appear to offer many advantages over the Buckley–James approach (*see* **Missing Data**).

Other semiparametric estimation methods have been proposed by Miller [13] and by Koul et al. [9]. The former is based on minimizing the weighted sum of squared residuals $\int \varepsilon^2 \, \mathrm{d}\hat{F}(\varepsilon)$, where $\hat{F}$ is the Kaplan–Meier estimator of the residual distribution function, while the latter is based on the observation that the quantities $\delta_i Z_i / [(1 - G(Z_i|\mathbf{x}_i)]$ have mean $\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i$, where G is the censoring distribution. Koul et al. use a **Bayesian** estimator of G, thus obtaining observable quantities which form the basis of **estimating functions**. Both the Miller and Koul et al. estimators appear to be sensitive to the relationship between the censoring times and the covariates – the former requiring that censorship relate linearly with the same slope parameters $\boldsymbol{\beta}$, the latter that there be no relationship (Miller & Halpern [14]). A comparison of semiparametric methods based on application to the Stanford heart transplant data is provided by Miller & Halpern.

Estimates of parameters and derivation of their properties can be based generally on appropriate test statistics. In the case of accelerated failure-rate models the **linear rank test** statistics with right-censored data introduced by Prentice [15] provide a basis for estimation and testing using **ranks** of the data. Similar rank procedures have been introduced by Louis [12], Tsiatis [23], and Wei et al. [26]. Ritov [16] discusses the asymptotic equivalence of the method of Tsiatis and the Buckley–James-type estimators.

Bayesian methods of analysis in the accelerated failure-time models are considered by Christensen & Johnson [3].

Comparison with the Proportional Hazards Model

It is instructive to compare the accelerated failure-time model with the **proportional hazards** model or **Cox model**. In the proportional hazards model the survival function is related to the underlying survival function S_0 by

$$S(t|\mathbf{x}) = S_0(t)^{\phi(\mathbf{x})},$$

and the hazards are related by

$$\lambda(t|\mathbf{x}) = \phi(\mathbf{x})\lambda_0(t)$$

for some function ϕ and covariates $\mathbf{x}$. The Cox model takes the loglinear form $\phi(\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x})$. In practice, whether it is the accelerated failure-time model or the proportional hazards model that is appropriate (if either) will depend on the mechanisms operating on the survival times through the covariates. The only distributions that satisfy both the accelerated failure-time and proportional hazards conditions are the Weibull distributions with underlying hazard functions of the form $\lambda_0(t) = \alpha k t^{\alpha-1}$, in which case $\phi(\mathbf{x}) = \theta(\mathbf{x})^{\alpha}$.

Ciampi & Etezadi-Amoli [4] suggested that both accelerated failure-time and proportional hazards models could be embedded into an extended model of the form

$$\lambda(t|\mathbf{x}) = h(\boldsymbol{\alpha}'\mathbf{x})\lambda_0[h(\boldsymbol{\beta}'\mathbf{x})t]$$

for some function h. Then, if $\boldsymbol{\alpha} = \boldsymbol{\beta}$ we have an accelerated failure-time model, while if $\boldsymbol{\beta} = \mathbf{0}$ the model is proportional hazards. Comparing the two thus reduces to testing the values of the parameters in this embedded model provided the underlying distribution is not Weibull.

Extensions and Further Reading

In many applications the covariates used for adjustment may also vary with time. Examples include calendar period effects, immunodeficiency status which changes over time, indicators of receipt of additional treatments at time t, and so on. Extensions of regression models to include **time-dependent covariates** have become relatively routine in many areas of application. Their incorporation into accelerated failure-time models leads to

$$S[t|\mathbf{x}(t)] = S_0\{\theta[\mathbf{x}(t)]t\},$$

where the notation $\mathbf{x}(t)$ now reflects the dependence of the covariates on the time under consideration. Fully parametric analyses in which both

the underlying distribution and the nature of the dependence of the covariates on time are completely specified may be carried out using, for example, likelihood methods. Robins & Tsiatis [18] and Robins [17] study an approach to the analysis of models with time-dependent covariates which is semiparametric in the sense that the dependence of the covariates on time is fully specified but where the underlying distribution S_0 remains unspecified.

Useful accounts of accelerated failure-time models can be found in Kalbfleisch & Prentice [8] and Cox & Oakes [5]. Wei [25] provides a comprehensive overview of nonparametric methods of estimation in accelerated failure-time models, and compares them with proportional hazards models.

In the econometric literature accelerated failure-time models are typically referred to as *tobit* models.

Software

Comprehensive parametric analyses of accelerated failure-time regression models are available in widely used packages such as **S-PLUS** [22] and *SAS* [19], as well as many other commercially available packages (*see* **Software, Biostatistical**). These incorporate response distributions such as the lognormal, Weibull, log-logistic, and Rayleigh (*see* **Parametric Models in Survival Analysis**) plus their transforms, and with various forms of censoring. More specialized survival analysis packages such as *Egret* [20] also accommodate censored regression models. Specific procedures for semi- and nonparametric analyses do not appear to be widely available.

References

[1] Aitkin, M. (1981). A note on the regression analysis of censored data, *Technometrics* **23**, 161–163.

[2] Buckley, J. & James, I. (1979). Linear regression with censored data, *Biometrika* **66**, 429–436.

[3] Christensen, R. & Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process, *Biometrika* **75**, 693–704.

[4] Ciampi, A. & Etezadi-Amoli, J. (1985). A general model for testing the proportional hazards and the accelerated failure time hypotheses in the analysis of censored survival data with covariates, *Communications in Statistics – Theory and Methods* **14**, 651–667.

[5] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[6] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[7] James, I.R. (1995). A note on the analysis of censored regression data by multiple imputation, *Biometrics* **51**, 358–362.

[8] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[9] Koul, H., Susarla, V. & Van Ryzin, J. (1981). Regression analysis with randomly right censored data, *Annals of Statistics* **9**, 1276–1288.

[10] Lai, T.L. & Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, *Annals of Statistics* **19**, 1370–1402.

[11] Lin, J.S. & Wei, L.J. (1992). Linear regression analysis based on Buckley-James estimating equations, *Biometrics* **48**, 679–681.

[12] Louis, T.A. (1981). Nonparametric analysis of an accelerated failure time model, *Biometrika* **68**, 381–390.

[13] Miller, R. (1976). Least squares regression with censored data, *Biometrika* **63**, 449–464.

[14] Miller, R. & Halpern, J. (1982). Regression with censored data, *Biometrika* **69**, 521–531.

[15] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika* **65**, 167–179.

[16] Ritov, Y. (1990). Estimation in a linear regression model with censored data, *Annals of Statistics* **18**, 303–328.

[17] Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors, *Biometrika* **79**, 321–334.

[18] Robins, J. & Tsiatis, A.A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates, *Biometrika* **79**, 311–320.

[19] SAS Institute Inc. (1995). *The SAS System for Windows, Release 6.11*. Cary.

[20] CYTEL Software Corporation (1996). *Egret for Windows*. Statistics and Epidemiology Research Corporation, Cambridge.

[21] Smith, P.J. (1988). Asymptotic properties of linear regression estimators under a fixed censorship model, *Australian Journal of Statistics* **30**, 52–66.

[22] Insightful Corporation (2001). *S-Plus 6 for Windows Users Guide*. StatSci, a division of Math-Soft, Inc., Seattle.

[23] Tsiatis, A.A. (1990). Estimating regression parameters using linear rank tests for censored data, *Annals of Statistics* **18**, 354–372.

[24] Wei, G.C.G. & Tanner, M.A. (1991). Application of multiple imputation to the analysis of censored regression data, *Biometrics* **47**, 1297–1309.

[25] Wei, L.J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis, *Statistics in Medicine* **11**, 1871–1879.

[26] Wei, L.J., Ying, Z. & Lin, D.Y. (1990). Linear regression analysis of censored survival data based on rank data, *Biometrika* **77**, 845–851.

[27] Weissfeld, L.A. & Schneider, H. (1987). Inferences based on the Buckley-James procedure, *Communications in Statistics – Theory and Methods* **16**, 177–187.

IAN JAMES

Accident and Emergency Medicine

Accident and emergency medicine is that specialty of medicine whose practitioners offer immediate medical care to people with major and minor injuries and illnesses presenting as emergencies to departments of Accident and Emergency (A&E) in general hospitals.

The specialty of A&E medicine and A&E departments in the UK have their counterparts in other countries. For example, in the United States of America (US), Canada, Sweden, Australia, New Zealand, and Spain, there are emergency rooms in general hospitals that provide the same service as A&E departments in the UK. The crucial functions of this service are the formulation of an early diagnosis, the institution of immediate therapies, and the timely referral if needed to the most appropriate specialty or agency to allow maximum chance of optimum recovery to be achieved. The conditions with which people present can vary widely. At one end lie true emergencies, where lifesaving treatment is needed within the first hour of onset. At the other end are a vast range of minor injuries and illnesses that could be managed in primary care or by individuals themselves.

Historical Development

In the UK, prior to the inception of the National Health Service (NHS) in 1948, free medical care for the poor had been varyingly available for several centuries from infirmaries run by local councils and from independent hospitals funded by charity. Most conditions seen would have been, as now, minor illnesses and injuries. This pattern of free care for the poor was followed in other countries. The debate about whether these minor conditions should be seen in A&E departments is not new. It was first described in the *Lancet* in 1849 [7]. Casualty departments primarily saw people with injuries caused by trauma. The report of the Medical Advisory Committee of the Central Health Services Council on Accident and Emergency Services in 1962, known as the Platt Report, recommended centralizing casualty services in general hospitals where all specialties were represented. They were to become receiving departments and be managed by orthopedic surgeons because of the predominance of trauma cases. They were to be called Accident and Emergency departments. There was an increasing realization that most of the real emergencies were medical cases with presenting conditions such as heart attacks and severe asthmatic or epileptic attacks. There was also much debate, as now, about whether the main role of such services should be to manage major emergency cases or anyone who presented. Parallel developments have occurred elsewhere. Early concern about the growth in the use of emergency departments was raised in the US in 1966 [22] and in the UK in the 1980s [13]. For example, the annual rate of first attendance at such departments per 1000 population in England rose from 105 in 1961 to 218 in 1984 [13]. This upward trend seems to have finally peaked in 1989 at a rate of 233 attendances per 1000 population [24]. A&E departments are now an integral part of general hospitals in the UK. But the debates about their true role (managing only emergency cases or offering an alternative to primary care) and their relation to trauma centers still rage in many countries [2, 24].

Different Types of Study

Descriptive and Analytical Epidemiology

Many researchers have tried to understand the determinants of the large geographical and temporal variation in first attendance. Some of these, and many others, have striven to show either the inappropriateness of much of the attendance at A&E departments or that much of it could be managed in general practice, or both. There have been a very large number of studies of single departments looking at these issues. Most of these have just used simple descriptive statistics with occasional use of the **chi-square test** and simple parametric tests (*see* **Hypothesis Testing**). One of the earliest was reported by Weinerman et al.in 1966 [22]. Many of these studies are referenced in [24] and in the report of the Anglia and Oxford emergency health care project steering group [2]. Fairley et al. [5] undertook one of the first studies of more than one department. They found that rates of use were highest for the age range 15–44 years, sex-specific rates were higher for males, and about 10% of attenders are admitted. These results have been replicated by many others. In general, the large majority of cases are due to trauma [5], although medical cases are relatively more common in inner city areas. Reilly [18] was one of the first to

show that general practitioners (*see* **General Practice**) could do much of the work undertaken by A&E departments. Holohan [8] was the first to describe the concept of social predicament as a key determinant in a large proportion of cases. The concept of inappropriateness has been much described. But the first systematic attempt to produce a classification scheme was in 1960 by the Nuffield Provincial Hospitals Trust.

Milner et al. [13] used **correlation** analysis, **multiple linear regression**, multiple **logistic regression** and a nonparametric test for the analyses of variance (*see* **Nonparametric Methods**) in their studies on temporal and geographical use. They found an eighteen-fold difference among health districts in the mean annual new attendance rates at A&E departments in England over the period 1974–1985. There was a rising trend in these rates which was statistically significant ($P < 0.05$) for 89% of districts. There was also a twenty-six-fold difference in the extent to which new attenders were reviewed [12]. The ratio of return attendances to first attendances (reattendance ratio) had declined significantly ($P < 0.05$) in 70% of districts. Investigation of the variation in the reattendance ratio among eight A&E departments showed that it was booked reattendance which largely determined sample reattendance ratios [12].

There is now an NHS common minimum data set for A&E departments in the UK [16]. This should facilitate comparative research among A&E departments.

Clinical Research

Discriminant analysis has been used very successfully for producing survival probabilities using logistic functions and regression weightings to allow the systematic audit of emergency care for cases of major trauma. This began in the US in the early 1980s with the Major Trauma Outcomes Study [3] and was later adopted in the UK [21]. The mortality rate in A&E departments in the UK is much less than 1% and trauma accounts for less than one-fifth [20]. Most of the deaths are due to medical emergencies such as myocardial infarction, stroke, or asthma [19]. **Randomization** has proven very difficult in care for life-threatening emergencies. There have been no randomized controlled trials (RCTs) (*see* **Clinical Trials, Overview**) of major trauma centers or emergency helicopter medical services. Major well-designed comparative studies of these have been undertaken in the UK without random allocation by the Medical Care Research Unit in Sheffield [17]. The major obstacle to randomization was the organization of care. The emergency nature of cases interacted with the ability to randomize responsively and quickly. The Medical Care Research Unit in Sheffield is currently running a randomized controlled trial in the UK of paramedical assistance as the first emergency contact which randomizes the paramedics rather than the patients. For the less urgent conditions, informed consent in randomized controlled trials has usually been sought (*see* **Ethics of Randomized Trials**).

A search of the nine emergency journals on Medline for 1995 found only 4% of articles described RCTs. These were usually studies of minor clinical developments.

Health Technology Assessment

Weinerman et al. [23] used descriptive statistics and χ^2 analysis to describe the possibilities of medical triage in 1963 in a pilot study. There have been many similar subsequent studies which claimed to have evaluated nurse triage and shown it to be beneficial. Apart from one, they have all either excluded a comparative arm, not used valid **outcome measures**, or been pilot studies. George et al. [6] in 1992 used a comparative design with triage being alternately on and off. They showed that triage patients waited on average longer than nontriaged. This was especially so for those most in need of urgent medical care.

Health technology assessment (*see* **Health Services Organization in the US**) came of age in accident and emergency medicine with the publication of an RCT with a cost-effectiveness analysis by Murphy et al. [15]. This group used valid intermediary outcome measures and found that general practitioners (GPs) were more cost effective than hospital doctors or nurses for managing primary care cases which presented to an A&E department.

The debate about trauma centers rages on. There is a shortage of good quality research evidence on the relative costs and benefits of the alternative forms of care for patients suffering major trauma. The UK Department of Health has funded a major comparative study of this [17] which shows, according to the

Department, that trauma centers are not cost effective in the shire counties of England.

Laboratory and Basic Sciences

There are many investigations undertaken in A&E departments in the UK. X-ray testing is the most common, followed by blood testing [12]. The opportunities for such tests are great. Head injuries, for example, are very common, as are twisted ankles. New technologies, such as MRI scanning or near-patient testing, are constantly being developed, which could have a major impact on A&E clinical practice. So far none of these technologies has been evaluated as rigorously as new drugs are. The lack of rigorous evaluation of tests is a general finding in health care. A recent **Cochrane Collaboration** has been established to try to rectify this. Details about its work and testing methodologies can be found on the World Wide Web at `http://wwwsom.fmc.flinders.edu.au FUSA/COCHRANE/sadtdoc.htm`.

In the nine emergency care journals found on Medline in 1995, the vast majority of original articles contained descriptive statistics. Correlation and predictive analysis were much less common.

Statistical Models

In the vast majority of studies only standard statistical methods have been used. **Time series** analysis was used recently to estimate the staffing requirements of A&E departments at various times depending on the **case-mix** presenting [14]. Milner had previously used the autoregressive integrative moving average (ARIMA) process (*see* **ARMA and ARIMA Models**) using the Box–Jenkins procedure to estimate future workloads in the Trent region of England [11]. Three time series were forecast. These were the first attendance rate, the ratio of return to first attendances, and the local resident population forecasts. These forecasts were then combined to produce forecasts for the district numbers of first, return, and total attendances. The theoretical ARIMA methods were applied without modification. There were two other examples of studies of emergency departments in the statistical literature on Medline in the period 1985–1996. The first was the 1992 National Ambulatory Medical Survey from the US **National Center for Health Statistics**. This was a descriptive survey of a random sample of attendances at hospital emergency and outpatient departments [10]. The second described the use of **correspondence analysis** as a screening method for indicants for clinical diagnosis through the application of the independent **Bayesian method** [4].

The proximity of the place of residence of an individual to a health care facility predicts its use by that individual. This general relationship has been found to hold for the use of A&E departments by Ingram et al. [9] and others.

Landmark Studies

Major Trauma Outcomes Studies

The Major Trauma Outcome Studies in the US [3] and UK [21] have allowed the quality of emergency care to be examined thoroughly by health care professionals as well as by purchasers and providers. These confidential studies allowed mortality rates for departments to be compared after adjusting for the nature and severity of the injury by means of the Revised Trauma Score and the Injury Severity Score and the patient's age.

Deaths in A&E Departments

The battle for the heart and soul of A&E medicine has long since been won. History and trauma favored orthopedic surgery. Technology and the diseases of affluence favored general medicine. The Platt Report started the revolution in the UK and various learned bodies continued it. But it was Shalley & Cross who stopped the debate with their study using descriptive statistics which showed in 1984 that most preventable deaths in A&E departments were due to medical conditions [19].

Inappropriate Attendance

Weinerman relaunched the debate in 1966 about inappropriate attendance in the US with a descriptive study of a case series of 2028 patients [22]. This followed the Nuffield Provincial Hospitals Study of casualty services in 1960 and the Platt Report in 1962. We have still not answered the question about appropriateness. We do however understand much better the policy and health service issues (*see* **Health Services Research, Overview**).

Reattendance

Reattendance of a patient at A&E departments was thought to be determined by the diagnosis and need for treatment. Milner et al. [12] showed using multiple logistic regression in 1992 that these were minor influences. The crucial factor was whether the doctor booked a patient to return. This propensity varied in an idiosyncratic manner among departments.

Particular Statistical Concepts, Problems, and Techniques

Accurate, population-based information on the **incidence rates** of minor injuries and illnesses is not available in the UK, unlike the US where the National Health Interview Survey reports this annually [1]. A similar regular survey from time to time in the UK would help to assess the appropriateness of the great geographical variation in the use of A&E departments.

There is a need to develop valid quantitative **health status instruments** for common A&E conditions such as twisted ankles, lacerations, head injuries, and strains and sprains, as well as the uncommon ones such as burns, and ear, nose, throat and eye disorders. They will have to be simple to administer. This will allow cost-effectiveness studies of the various alternative models of care to be undertaken.

Anticipated Developments

The central issue on the use of A&E departments in the UK is not discovering the determinants of such use. It is to secure an agreed policy on the basis of research evidence on the role of A&E departments. Currently they are providing a combination of services for hospital emergencies, minor injuries, alternative primary care, major trauma, and/or a fail-safe system for healthcare. There are two basic options for coping with the out-of-hours emergency problems and the overlap between general practice and A&E departments.

One model is to develop emergency primary health care centers for out-of-hours work or for a 24-hours-a-day service. The second model is to develop primary care within the A&E department. We need to know the cost effectiveness of these options. When there is agreement in a locality about the respective roles of hospital emergency services and general medical services, then there is an obligation to inform local people about using these health services appropriately.

The Cochrane Collaboration is systematically reviewing the literature by health problem through a world-wide collaboration based on Cochrane Centers and health problem collaborative groups (`http://cochrane@mcmaster.ca`). There will be a systematic attempt to bring together knowledge on emergency care from the collaborative groups.

References

[1] Adams, F.P. & Marano, M.A. (1996). *Current Estimates from the National Health Interview Survey 1994*. US Department of Health and Human Services. Vital and Health Statistics Series, no. 10. No. 193 DHSS Pub No. (PHS) 95–1521. US Government Printing Office, Washington.

[2] Anglia and Oxford Emergency Health Care Steering Group (1995). *Opportunities in Emergency Health Care*. Anglia and Oxford NHS-E Regional Office, Oxford.

[3] Champion, H.R., Copes, W.S., Sacco, W.J., Lawnwick, M.M., Bain, L.W., Gann, D.S., Gennarelli, T., Mackenzie, E. & Schwaitzberg, S. (1990). A new characterisation of injury severity, *Journal of Trauma* **30**, 1356–1365.

[4] Crichton, N.J. & Hinde, J.P. (1989). Correspondence analysis as a screening method for indicants for clinical diagnosis, *Statistics in Medicine* **8**, 1351–1362.

[5] Fairley, J. & Hewett, W.C. (1969). Survey of casualty departments in Greater London, *British Medical Journal* **2**, 375–377.

[6] George, S., Read, S., Westlake, L., Williams, B.T., Fraser-Moodie, A. & Pritty, P. (1992). Evaluation of nurse triage in a British accident and emergency department, *British Medical Journal* **304**, 876–878.

[7] Hodgson, J. (1849). The genteel out-patient abuse at the public charities, *Lancet* **ii**, 705.

[8] Holohan A.M. (1976). Accident and Emergency departments: illness and accident behaviour, in *The Sociology of the NHS*, N. Stacey, ed. Sociological Review Monographs 22, London, pp. 111–119.

[9] Ingram, D.R., Clarke, D.R., & Murdie, R.A. (1987). Distance and the decision to visit an emergency department, *Social Science and Medicine* **12**, 55–62.

[10] McCaig, L.F. & McLemore, T. (1994). Plan and operation of the National Hospital Ambulatory Medical Survey, *Vital and Health Statistics* **34**, 1–78.

[11] Milner, P.C. (1988). Forecasting the demand on accident and emergency departments in health districts in the Trent region, *Statistics in Medicine* **7**, 1061–1072.

[12] Milner, P.C., Beeby, N. & Nicholl, J. (1991). Who should review the walking wounded? Reattendance at

Accident and Emergency Departments, *Health Trends* **23**, 36–44.

[13] Milner, P.C., Nicholl, J.P. & Williams, B.T. (1988). Variation in demand for Accident and Emergency departments in England from 1974 to 1985, *Journal of Epidemiology and Community Health* **42**, 274–278.

[14] Morris, R.W., Leikin, J.B., Eckenrode, P. & Boston, D. (1990). The effects of time of trauma patient presentation on emergency department utilization, *Progress in Clinical Biological Research* **341A**, 201–211.

[15] Murphy, A.W., Bury, G., Plunkett, P.K., Gibney, D., Smith, M., Mullan, E. & Johnson, Z. (1996). Randomised controlled trial of general practitioners versus usual medical care in an urban accident and emergency department, *British Medical Journal* **312**, 1135–1142.

[16] NHS-E Information Management Group (1995). *Hospital Services Module, Version 3.0 of the NHS Data Manual: Chapter 4 Accident and Emergency Service*. National Health Service Executive, Leeds.

[17] Nicholl, J.P., Williams, B.T. & Brazier, J.E. (1993). Management of trauma, *British Medical Journal* **307**, 683–684.

[18] Reilly, P.M. (1981). Primary care and accident and emergency departments in an urban area, *Journal of the Royal College of General Practice* **31**, 223–230.

[19] Shalley, M.J. & Cross, A.B. (1984). Which patients die in an accident and emergency department?, *British Medical Journal* **289**, 419–421.

[20] Underhill, T.J. & Finlayson, B.J. (1989). A review of trauma deaths in an accident and emergency department, *Archives of Emergency Medicine* **6**, 90–96.

[21] Wardrope, J., Cross, S.F. & Fothergill, D.J. (1990). One year's experience of major trauma outcome study methodology, *British Medical Journal* **301**, 156–159.

[22] Weinerman, E.R., Ratner, R.S., Robbins, A. & Lavenhar, M.A. (1966). Determinants of the use of hospital emergency services, *American Journal of Public Health* **56**, 1037–1056.

[23] Weinerman, E.R., Rutzen, S.R. & Pearson, D.A. (1963). Effects of medical "Triage" in hospital emergency service, *Public Health Reports* **80**, 389–399.

[24] Williams, B., Nicholl, J. & Brazier, J. (1996). *Epidemiology-based needs assessment: Accident and Emergency Departments*. National Health Service Executive, Leeds.

P. MILNER

Accident Proneness

Research into the concept of accident proneness was motivated by the desire to find effective ways of reducing accidents. The initial hope was to identify the individuals most "prone" to have accidents and to nullify their problems in some way. Accident proneness was viewed as a personal psychological factor which affected the individual's probability of suffering an accident. The original context was industrial accidents during the 1914–1918 war. A very substantial research literature grew up over the succeeding half-century, especially as road accidents became a serious social and economic problem.

The concept of an accident as a purely random event had led Bortkiewicz [2, 3] to develop the **Poisson distribution** as a model for the number, X, of fatal accidents at work in a given time interval

$$\Pr(X = x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots, \quad 0 < \lambda. \tag{1}$$

His data sets included the well-known data on deaths from cavalry horse kicks. The Poisson model assumes that all individuals have the same probability (proportional to λ) of having an accident. The model implies that if you remove from the population under consideration those members who have had the highest number of accidents over a period of time, then this will have no effect whatsoever on the distribution of accidents in the population in subsequent periods.

Greenwood & Woods [5] and Greenwood & Yule [6] challenged the idea of pure randomness in their investigation into factory accidents. They put forward three competing hypotheses:

1. *Pure chance*, leading to the Poisson distribution, (1).
2. *True contagion*, i.e. the hypothesis that all individuals initially have the same probability of having an accident, but that this probability changes each time an accident is incurred. This led to their "biased distribution". If the probability of an accident remains unchanged after the occurrence of the first accident, then they described the outcome as the "burnt fingers distribution".
3. *Apparent contagion*, i.e. the hypothesis that individuals have constant but unequal probabilities of having an accident. This became known as accident proneness in the literature. It gives rise to a mixed Poisson distribution. Greenwood & Yule's well-known model for accident data assumes that the probability of an accident varies from individual to individual according to a gamma (c, k) distribution (*see* **Gamma Distribution**); the outcome is that the overall distribution of accidents in the population is a **negative binomial distribution** with

$$\begin{aligned}\Pr(X = x) &= \int_0^\infty \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}\frac{\mathrm{e}^{-c\lambda}c^k\lambda^{k-1}\,\mathrm{d}\lambda}{\Gamma(k)} \\ &= \binom{k+x-1}{x}\left(\frac{1}{c+1}\right)^x\left(\frac{c}{c+1}\right)^k, \\ &\qquad x = 0, 1, 2, \ldots, \ 0 < c, \ 0 < k\end{aligned} \tag{2}$$

(*see* **Contagious Distributions**).

However, it is easy to construct a *true contagion* model which also leads to the negative binomial distribution of (2) – a good empirical fit of the negative binomial distribution to population accident data cannot therefore distinguish between true contagion and accident proneness.

During the 1950s a number of authors (including Arbous & Kerrich [1]) tried to detect accident proneness by examining individuals' accident records in two consecutive periods. The general finding was that in practice this bivariate approach requires very large data sets. Arbous & Kerrich gave a good review of contemporary theories of accident occurrence.

Cresswell & Froggatt [4] in their study of bus driver accidents rejected the idea of accident proneness in favor of a fourth model:

4. *Spells*; here each driver is assumed to be susceptible to random spells (periods of time) during which accidents may befall him/her randomly with a probability that is the same for all drivers. They called the outcome distribution "long" or "short" according to whether further accidents might not or might occur randomly outside a spell. For their long model

$$\begin{aligned}\Pr(X = x) &= \sum_{j=0}^{\infty}\frac{\mathrm{e}^{-j\lambda}(j\lambda)^x}{x!}\frac{\mathrm{e}^{-\phi}\phi^j}{j}, \\ &\qquad x = 0, 1, 2, \ldots, \ 0 < \lambda, \ 0 < \phi.\end{aligned} \tag{3}$$

There is no simple expression for these probabilities. The distribution is better known in the statistical literature as the Neyman type A.

It soon became apparent that the problems of distinguishing between the various hypotheses are very severe. The Neyman type A distribution can easily be given a proneness interpretation [7] and, similarly, the negative binomial distribution can be given a spells interpretation.

A major problem which has bedevilled accident proneness as a concept is its exact definition – how is proneness to be distinguished from other aspects of personal risk, e.g. age or experience? This does not seem to have been resolved satisfactorily.

Prior to 1968, accident models assumed constant environmental risk as opposed to personal risk. Irwin [8] criticized this assumption and introduced a fifth type of model:

5. *Accident liability and accident proneness*; this incorporates the concept of accident liability resulting from varying environmental exposure. Irwin developed a three-parameter "Generalized Waring" distribution that assumes randomness while taking into account varying accident liability as well as varying accident proneness. He set $\theta = 1/(c+1)$ in (2) and assumed that θ has a **beta(a,b)** distribution, giving

$$\begin{aligned}\Pr(X = x) &= \int_0^1 \binom{k+x-1}{x} \theta^x (1-\theta)^k \\ &\quad \times \frac{\theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta}{B(a,b)} \\ &= \frac{(b+k-1)!(a+b-1)!}{(b-1)!(k-1)!(a-1)!} \\ &\quad \times \frac{(k+x-1)!(a+x-1)!}{(a+b+k+x-1)!x!}, \\ &\quad x = 0, 1, 2, \ldots, \end{aligned} \tag{4}$$

where $0 < k$, $0 < a$, $0 < b$. The theory underlying this model has been studied in depth by Xekalaki [11, 12] both for a single time period and for a subdivided time period. Discrimination between proneness and liability is theoretically possible but it is difficult to achieve this in practice.

Most of the work on proneness and related concepts has involved accident count data and hence discrete distributions. An alternative approach is to examine interaccident times (involving continuous distributions). This has received some attention but it runs into problems similar to those with count data – it is particularly difficult to get reliable large-scale data on interaccident times.

There are two major books on the statistical analysis of accident data. Both involve large data sets. The two books display strongly contrasting views on accident theory – Cresswell & Froggatt [4] favor the spells hypothesis while Shaw & Sichel [10] strongly endorse the accident-proneness approach. Kemp [9] gave a detailed review of work on proneness and related topics from 1920 to 1970. He concluded that "from a practical point of view (e.g. in terms of its contribution to accident prevention), the concept of accident proneness had proved singularly ineffectual". Nevertheless, the study of accident proneness was valuable in the development of statistical methodology.

By the early 1980s the golden age of accident proneness theorizing was over. Very little theoretical research appears to have taken place since then. Attention had moved towards **risk** evaluation and analysis. This may well reflect the view that whether or not proneness in a narrow sense does exist, in practice there are other very important factors that contribute to a particular individual's accident record.

References

[1] Arbous, A.G. & Kerrich, J.E. (1951). Accident statistics and the concept of accident proneness, *Biometrics* **7**, 340–432.

[2] Bortkiewicz, L.von (1898). *Das Gesetz der Kleinen Zahlen*. Teubner, Leipzig.

[3] Bortkiewicz, L.von (1915). Über die Zeitfolge zufälliger Ereignisse, *Bulletin de l'Institut International de Statistique* **20**, 30–111.

[4] Cresswell, W.L. & Froggatt, P. (1963). *The Causation of Bus Driver Accidents*. Oxford University Press, London.

[5] Greenwood, M. & Woods, H.M. (1919). A report on the incidence of industrial accidents upon individuals with special reference to multiple accidents, *Industrial Fatigue Research Board Report*, Vol. 4. HMSO, London.

[6] Greenwood, M. & Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society, Series A* **83**, 255–279.

[7] Irwin, J.O. (1964). The personal factor in accidents-a review article, *Journal of the Royal Statistical Society, Series A* **127**, 438–451.

[8] Irwin, J.O. (1968). The generalized Waring distribution applied to accident theory, *Journal of the Royal Statistical Society, Series A* **131**, 205–225.

[9] Kemp, C.D. (1970). "Accident proneness" and discrete distribution theory, in *Random Counts in Scientific Work*. Vol. 2, *Random Counts in Biomedical and Social Sciences*, G.P. Patil, ed. Pennsylvania State University Press, University Park, pp. 41–65.

[10] Shaw, L. & Sichel, H.S. (1971). *Accident Proneness*. Pergamon Press, Oxford.

[11] Xekalaki, E. (1983). The univariate generalized Waring distribution in relation to accident theory: proneness, spells, or contagion?, *Biometrics* **39**, 887–895.

[12] Xekalaki, E. (1984). The bivariate generalized Waring distribution and its application to accident theory, *Journal of the Royal Statistical Society, Series A* **147**, 488–498.

ADRIENNE W. KEMP & C.D. KEMP

Actuarial Methods

Historically, probability theory and statistical methods have played a central part in actuarial science, in both theory and practice. Indeed, the motto of the Institute of Actuaries, the professional body in England and the first to be established worldwide is *certum ex incertis*. Furthermore, the early development of these subjects was inextricably linked, with many of the principal contributors to actuarial theory also making notable contributions to probability and statistics – for example, **John Graunt**, Abraham de Moivre, Thomas Simpson, Daniel and Nicholas **Bernoulli**, and Erastus de Forest. Also, some modern statistical models have little-known actuarial antecedents, e.g. Böhmer's development in 1912 of the product limit estimator of **Kaplan–Meier**, and Du Pasquier's analysis in 1913 of multiple state and **competing risk** models; see Haberman & Sibbett [28] for further discussion.

We begin our review with a brief consideration of the nature of actuarial science. Actuarial science is concerned with the financial management of financial security systems – these can be defined as "mechanisms for reducing the adverse financial impact of random events that prevent the fulfillment of reasonable expectations" [3]. These systems have the important characteristics of **risk** transfer and risk pooling [12] but certain fundamental limitations. For example, they are restricted to reducing the consequences of random events that create losses that can be measured in monetary terms. Secondly, such systems do not directly reduce the probability of a loss occurring.

Examples of situations where random events may cause financial losses would include the following:

1. The destruction of property by fire or natural catastrophe (storm, hail, flood, landslide, earthquake, volcanic eruption) is usually considered a random event in which the loss can be measured in monetary terms.
2. A damage award imposed by a court as a result of a negligent event is often considered a random event with resulting monetary loss.
3. Prolonged illness may occur unexpectedly and result in financial losses in terms of reduced income and extra health care expenses.
4. Death of a young adult may occur while long-term family and business commitments remain unfulfilled.
5. Survival to an advanced age may deplete an individual's resources for meeting the cost of living, including long-term care.

One of the key tasks for an actuary advising financial security systems is the management of uncertainty. This process can be broken down into a number of distinct stages; for example, one classification would be: identification of information sources; collection of data; analysis; model construction (*see* **Model, Choice of**); **sensitivity analysis**; **prediction**; monitoring the model assumptions in the light of emerging experience (*see* **Model Checking**); updating the model.

Survival Model (or Life Table): Structure

The **survival** model is concerned with representing the mortality of individuals. Here, we consider single lives, although the extension to contingencies involving multiple lives is straightforward [3, 21]. The initial assumption is that the time from birth to death can be represented by a continuous **random variable** T_0. We define the distribution function of T_0 and the survival function of T_0 as follows:

$$F_0(t) = \Pr(T_0 \le t); \tag{1}$$

$$S_0(t) = \Pr(T_0 > t) = 1 - F_0(t). \tag{2}$$

If we consider an individual aged $x(> 0)$ currently, then we can define a random variable T_x to be his/her future lifetime, conditional on him/her having survived to age x. Then the distribution function of T_x is defined as

$$F_x(t) = \Pr(T_x \le t) = \Pr(T_0 \le x + t | T_0 > x), \tag{3}$$

which is written as ${}_tq_x$ in actuarial notation (and as q_x in the special case when $t = 1$), and the survival function is defined as

$$S_x(t) = \Pr(T_x > t) = \Pr(T_0 > x + t | T_0 > x), \tag{4}$$

which is written as ${}_tp_x$ in actuarial notation.

It is then straightforward to demonstrate the connection between F_x and F_0, and between S_x and S_0, namely

$$F_x(t) = \frac{F_0(x + t) - F_0(x)}{1 - F_0(x)} \tag{5}$$

and

$$S_x(t) = \frac{S_0(x+t)}{S_0(x)}. \tag{6}$$

The force of mortality at age x, μ_x, is defined as

$$\mu_x = \lim_{h \to 0} \left[\frac{\Pr(x < T_0 \leq x + h | T_0 > x)}{h} \right]. \tag{7}$$

The force of mortality is also described as the **hazard rate**.

The probability density function of T_x, $f_x(t)$, is defined by

$$f_x(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_x(t),$$

and is linked to the force of mortality through the following relationship which follows from this definition:

$$f_x(t) = {}_t p_x \mu_{x+t}, \tag{8}$$

so that

$$\mu_{x+t} = \frac{f_x(t)}{S_x(t)}.$$

This gives rise to a differential equation for ${}_t p_x$,

$$\frac{\mathrm{d}}{\mathrm{d}t} {}_t p_x = -{}_t p_x \mu_{x+t}, \tag{9}$$

which can be integrated with the boundary condition ${}_0 p_x = 1$ to give the following useful and important formula:

$${}_t p_x = \exp\left(- \int_0^t \mu_{x+s} \, \mathrm{d}s \right). \tag{10}$$

In numerical applications of the survival model it is common to impose simplifying assumptions on the distribution of T_x within a particular year of age. The two most commonly used such assumptions are:

1. a **uniform distribution** of deaths, i.e.

$$f_x(t) = \text{constant for } 0 \leq t \leq 1;$$

2. a constant force of mortality, i.e.

$$\mu_{x+t} = \text{constant for } 0 \leq t \leq 1.$$

An important modification to the survival model is the development of a select survival model, for use in many applications, in particular life insurance. The survival model is constructed from observations for certain population groups, differentiated by characteristics such as sex, geographical area, and type of insurance purchased. The age at entry to the group under consideration can have a significant influence on the resulting probabilities.

To focus the discussion, we consider an individual who has just purchased life insurance at age x. Since life insurance is carefully underwritten and only lives in good health are accepted (sometimes after a medical examination), it is reasonable to expect that a person who has just purchased insurance at, say, age x will be in better health than a person who bought insurance t years ago, say, at age $x - t$, and is now also aged x (*ceteris paribus*). This dependence of health status on t and x will have an impact on the probabilities of survival and is allowed for by a select survival model. Specifically, the probabilities of death are graded according to age at entry and duration of membership. The notation is to represent the one-year conditional probability of death for a person who entered at age x and who is now aged $x + t$ as $q_{(x)+t}$. Then the selection effect is represented by the sequence of inequalities:

$$q_{(x)} < q_{(x-1)+1} < q_{(x-2)+2} < \cdots, \tag{11}$$

where each probability refers to the conditional probability of death for a person aged x with different periods of membership. Empirically, we find that the selection effect becomes negligible some years, say r, after entry. We represent this feature by requiring that

$$q_{(x-r)+r} = q_{(x-r-1)+r+1} = \cdots = q_x.$$

r is then called the select period and q_x is called the ultimate probability of death at age x.

Selection arises in other practical circumstances, for example for persons purchasing life annuities and for those retiring after disablement. This latter case provides an example of negative selection for which the inequalities in (11) would be reversed.

The survival model can be traced back to Graunt's landmark contribution with the setting up of the first **life table** in 1662. The first authors to have used this life table were the Huygens brothers, who corresponded on the probabilistic interpretation of various life table indices. The first life table in the modern sense is widely attributed to **Halley** in 1693. Further historical details can be found [28, 29].

Survival Model and the Life Table: Actuarial Applications

In general, life insurance policies involve benefits payable by the insurer to the policyholder contingent on the policyholder's status and, in return, premiums are payable by the policyholder while he or she is alive. The benefits may consist of a single payment or a series of payments, and may be dependent on the policyholder having just died (as in life insurance) or being alive (as in an annuity). The financial management and control of life insurance depends critically on the survival model (and life table) which is used by actuaries in the calculations of premiums, reserves, surrender values, and other functions; see [3, 21] for further discussion.

Life insurance mathematics was developed by de Moivre (with his book of 1725, which was the first text on this subject) and Simpson (with his book of 1742), although their prime focus was on annuity rather than insurance contracts. Dodson was the first to demonstrate (in 1755) how modern life insurance could be operated with level annual premiums calculated on the basis of age at entry, and how this level charge for an increasing risk leads to the build-up of a reserve.

Survival Model: Estimation of Parameters

Estimation of $F_x(t)$, $S_x(t)$, $f_x(t)$, or μ_{x+t} will enable us to specify the distribution of T_x, given certain mildly restrictive conditions.

The simplest experiment would be to observe a large number of individuals, born in a particular time interval: then the proportion alive at age $t > 0$ would provide an estimate of $S_0(t)$. This is a nonparametric approach (*see* **Nonparametric Methods**), leading to a step function which would become more regular if the sample size were increased. Such an approach is not practicable because of the length of time it would take to specify fully the survival function and because it may not be possible to observe the deaths of all the lives in the study, because of censoring (*see* **Censored Data**). In medical statistics, however, this type of experiment is widely used, and estimators like the Kaplan–Meier (product limit) and the **Nelson–Aalen estimators** have been developed which allow for censored observations. The so-called *actuarial estimator* also enjoys wide use when the data are in grouped form [16]. We consider a partition of the survival period as follows:

$$0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = \infty,$$

and assume that the total population of lives at time t_0, N_0, is of the same exact age (and suppress age in the notation). Let d_i be the observed number of deaths and w_i the number of right-censored observations (or losses) during the interval (t_i, t_{i+1}). Let N_i be the number of lives at risk at the start of the interval (t_i, t_{i+1}), i.e. just after time t_i.

Then $N_{i+1} = N_i - d_i - w_i$, and the *actuarial estimate* of $F(t)$ is

$$\hat{F}(t) = 1 - \prod_{\substack{j\geq 0 \\ t_{j+1}\leq t}} \left(1 - \frac{d_j}{N_j - \frac{1}{2}w_j}\right). \qquad (12)$$

In actuarial practice, it is normal to use a different experimental plan and base estimation on data gathered within a short time interval – for example, four calendar years for the standard life tables prepared by the UK Continuous Mortality Investigation Bureau (CMIB). As a consequence, we observe several cohorts within a well-defined *window* rather than one cohort over its full life history (see Figure 1). As a result, we might not be sampling from the same distribution and it may be necessary to impose further assumptions on our model (for example, that survival probabilities are constant with respect to calendar time). Limiting the observation time to a rectangle defined by a specific period of time and an age interval, say x to $x + 1$, also introduces censoring: lives enter observation at a known time and survivors leave observation at a known time (when the investigation period ends or on attaining age $x + 1$), while

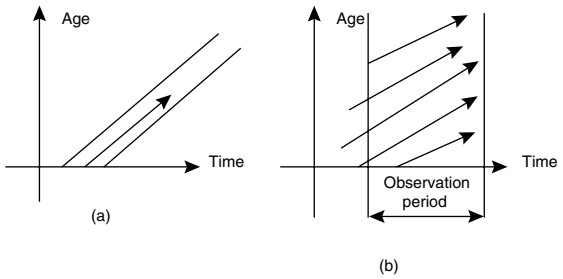

Figure 1 Lexis diagrams illustrating different experimental plans. (a) Cohort-based; (b) fixed period

for deaths and other types of exit (for example, surrenders of a policy), the time of exit will be random. Further discussion of censoring is provided by [1].

To take the actuarial approach further, we follow Broffit [5] and consider one particular form of censoring. We consider N lives to be observed between ages x and $x+1$. For the ith life we let $x+a_i$ be the age at which observation starts and $x+b_i$ the age at which observation must cease if the life survives to that age. Then $b_i = \min(1, a_i + c_i - e_i)$, where e_i and c_i are the dates of entry into the study and of the end of the study. The key point is that a_i, e_i, and c_i, and hence b_i, are known in advance.

We define two random variables

$$D_i = \begin{cases} 1, & \text{if the } i\text{th life is observed to die,} \\ 0, & \text{if the } i\text{th life is not observed to die,} \end{cases}$$

and T_i such that $x + T_i =$ age at which observation of the ith life ceases.

We also define $W_i = T_i - a_i$, the waiting time or time spent under observation for the ith life.

We note that

$$D_i = \begin{cases} 0, & \text{if and only if } W_i = b_i - a_i, \\ 1, & \text{if and only if } 0 < W_i < b_i - a_i. \end{cases}$$

The outcome of observing these random variables is a sample (d_i, w_i), and we define $w = \sum_{i=1}^{N} w_i$ and $d = \sum_{i=1}^{N} d_i$, the total waiting time and total number of deaths observed, respectively.

The **maximum likelihood** estimator for μ is then

$$\hat{\mu} = \frac{D}{W}, \tag{13}$$

where $D = \sum_{i=1}^{N} D_i$ and $W = \sum_{i=1}^{N} W_i$, and the corresponding estimate is $\tilde{\mu} = d/w$.

In many applications the randomness of W_i is ignored and it is usual to write the realized value w as E_x^c, the central *exposure* to risk. Assuming a constant force of mortality as before, the assumption that D has a **Poisson distribution** with parameter μE_x^c leads to the estimator

$$\hat{\mu} = \frac{D}{E_x^c}. \tag{14}$$

The Poisson model is not exact given the above experimental design, but it is a good approximation in many applications.

Given the estimated values of μ from (13) or (14), it is then possible to construct a survival model, using the standard results described earlier. By using estimates from successive ages and time periods, it is also possible to construct a cohort life table, as depicted schematically in Figure 1(a).

Many of the early life tables were based on the experience of individuals who purchased annuities (usually from the government) or who participated in tontines: for example, those constructed by Struyck in 1740, Kersseboom in 1742, and Deparcieux in 1746. These life tables were based on the cohort design of Figure 1(a). In 1749, the Swedish General Register Office was established and the first national set of population data started to be collected from that date. Wargentin combined death registration data for 1755–1763 and the triennial censuses of 1757, 1760, and 1763 (used to approximate exposed to risk figures) to estimate values of q_x (1766). This procedure was taken up and developed further by Price in 1783 to produce the first modern life tables based on the experimental design of Figure 1(b). The first published life table in 1828 based on the mortality experience of an insurance company was reported by Morgan in 1828.

Multiple State and Multiple Decrement Models: Structure

The survival model can be regarded as a two-state model (Figure 2), with two states "alive" and "dead" and transitions permitted in one direction only. This model can be extended to include any number of states, with transitions between them in either direction. Two examples with important applications in actuarial work, are the multiple decrement model (Figure 3, widely used in pensions applications) and the three-state disability model (Figure 4, widely used in disability insurance applications).

We consider initially the case where there are n possible states. Let $S(x)$, $0 \le x \le \infty$, be a continuous-time, time-inhomogeneous **Markov process** with a finite state space ($n < \infty$), and suppose that we interpret "$S(x) = 1$" to mean "the individual is in state 1 at age/time x".

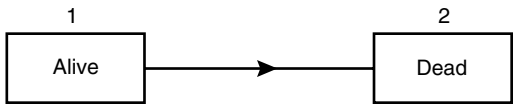

Figure 2 Two-state survival model

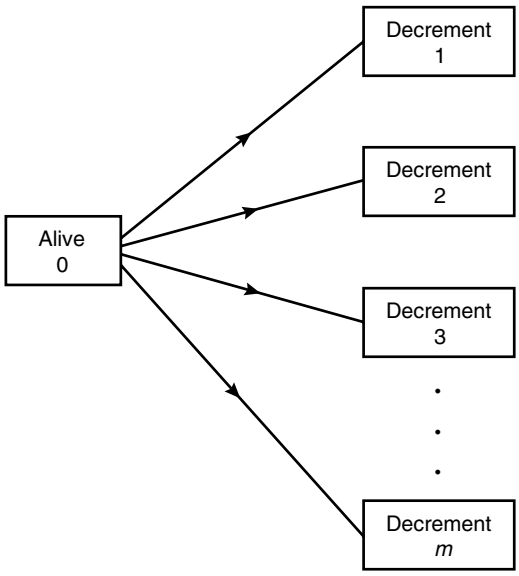

Figure 3 Multiple decrement model

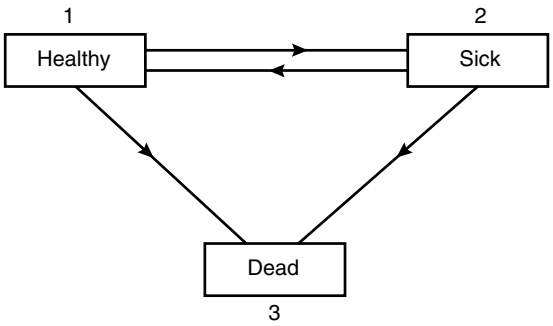

Figure 4 Three-state multiple state model

We define the conditional transition probability

$$ {}_tp_x^{ab} = \Pr[S(x+t) = b | S(x) = a], $$

and the occupancy probability

$$ {}_tp_x^{\overline{aa}} = \Pr[S(x+u) = a, \quad \text{for all } u \in (0,t) | S(x) = a]. $$

Corresponding to the definition of the force of mortality, (7), in the survival model, we define the transition intensity from state a to state b at age x by

$$ \mu_x^{ab} = \lim_{h \to 0+} \left(\frac{{}_hp_x^{ab}}{h} \right), \quad \text{for } a \neq b. $$

Then, we can derive the Chapman–Kolmogorov forward differential equations:

$$ \frac{\partial}{\partial t} {}_tp_x^{ab} = \sum_{j \neq b} {}_tp_x^{aj} \mu_{x+t}^{jb} - {}_tp_x^{ab} \mu_{x+t}^{bj} \tag{15} $$

and

$$ \frac{\partial}{\partial t} {}_tp_x^{\overline{aa}} = -{}_tp_x^{\overline{aa}} \sum_{j \neq a} \mu_{x+t}^{aj}, \tag{16} $$

which are generalizations of (9). Given estimates of the transition intensities, this set of equations (15) can be solved numerically (see, for example, [8, 31]) or analytically in certain special cases (see [42] for a discussion of piecewise constant transition intensities). Eq. (16) can be integrated directly, in a similar manner to (9), leading to

$$ {}_tp_x^{\overline{aa}} = \exp\left(-\int_0^t \sum_{j \neq a} \mu_{x+s}^{aj} \, ds \right), \tag{17} $$

which plays an important role in the estimation of the transition intensities from observed data.

Semi-Markov Model

In the above discussion, the Markov assumption has been made: that transition intensities (and probabilities) at time t depend (at least explicitly) on the current state at that time only. More realistic, and possibly more complex, models can be constructed considering, for example:

1. the dependence of some intensities (and probabilities) on the age x at time 0, corresponding, for example, to the issue of an insurance policy
2. the dependence of some intensities (and probabilities) on the time spent in the current state since the latest transition to that state
3. the dependence of some intensities (and probabilities) on the total time spent in some states since the policy issue.

The consideration of point 1 implies the use of issue-select intensities, corresponding to $\mu_{(x)+t}$ for the survival model and life table. This extension does not imply the use of more complex models since it is implicitly allowed for by the Markov assumption for the process $S(t)$.

With point 2, the Markovian property of the process $S(t)$ is lost. Nevertheless, there are practicable ways of dealing with this assumption, which is of practical importance in disability insurance where transitions from the disabled state would depend on the duration of the current disability. One general (and complex) approach leads to **semi-Markov** processes [8, 11], while a simpler approach requires the "splitting" of some states [11, 31].

The aim of point 3 is to stress the individual's life history. This assumption can lead to intractable models. (However, particular aspects of this assumption can be introduced without a dramatic increase in complexity.)

Multiple State Models: Actuarial Applications

As for the survival model, multiple state models are used for the determination of premiums and reserves for insurance policies, operating in a multiple state environment – for example, income protection insurance policies in the UK which provide an annuity while the individual is "sick or disabled" subject to certain qualifying conditions [40]. For a further discussion, see [25].

Correspondingly, the multiple decrement model is widely used in applications in defined pension schemes where the actuary's objective is to determine the contribution rate for current members and to calculate reserves at regular intervals and to monitor the financial health of the scheme.

In some practical applications, it is important to allow for the effects of selection arising from the effect of different transitions. For example, where withdrawals are associated with lower than average mortality rates, increased mortality in the continuing population results; where early retirements are associated with higher than average mortality, the result is decreased mortality in the continuing population.

Multiple State Models: Estimation of Parameters

For illustration we consider the three-state model of Figure 4; extensions to the more general case are straightforward. We consider an observation period of perhaps several calendar years and assume that each individual represents an independent realization of the underlying **stochastic process**, $S(y)$, where y is the individual's age. We assume that, while under observation, we can observe the time and type of each transition that an individual makes. We focus, for inference purposes, on the age interval $(x, x+1)$, over which we assume that the transition intensities are constants, μ_x^{12}, μ_x^{13}, μ_x^{21}, μ_x^{23}.

The observations in respect of a single life are now:

1. the times between successive transitions
2. the numbers of transitions of each type.

The form of the **likelihood** means that it suffices to record the total waiting time spent in each state. Following Sverdrup [51], we then define

C_j = waiting time of the jth life in the healthy state
W_j = waiting time of the jth life in the disabled state
S_j = number of transitions from healthy $\longrightarrow$ disabled by the jth life
R_j = number of transitions from disabled $\longrightarrow$ healthy by the jth life
D_j = number of transitions from healthy $\longrightarrow$ dead by the jth life
U_j = number of transitions from disabled $\longrightarrow$ dead by the jth life,

and totals $C = \sum_1^N C_j$ (and so on). It can then be shown that the maximum likelihood estimators are, respectively,

$$\hat{\mu}_x^{13} = \frac{D}{C}, \quad \hat{\mu}_x^{23} = \frac{U}{W},$$
$$\hat{\mu}_x^{12} = \frac{S}{C}, \quad \hat{\mu}_x^{21} = \frac{R}{W}. \tag{18}$$

We note that each estimator is the ratio of two random variables: number of transitions and waiting time (or central exposed to risk).

It may be important to be able to estimate the **moments** of these estimators, for example when comparing the results of two sets of observations or comparing one experience with a given standard experience. It is a well-known result of maximum likelihood theory that the asymptotic distribution of each $\hat{\mu}$ is **normal** with mean μ and variance $\mu/E(C)$ or $\mu/E(W)$, as appropriate.

The history of the development of multiple state models has been fully described by [13] and [46].

These models can be traced back to **Bernoulli's** memoir of 1766, in which he applied the methods of differential calculus to a problem involving a three-state decrement model and then solved the resulting differential equations under certain constraints. The problem was concerned with the incidence of smallpox in the population and measuring the efficacy of inoculation. Bernoulli's ideas were developed by a number of authors, in particular Lambert in 1772, Cournot in 1843, Makeham in 1867, Karup in 1875, and Du Pasquier in 1912.

Du Pasquier's work is very significant, presenting an early application of Markov processes and laying the foundations for modern actuarial applications to disability insurance, long-term care insurance, and critical illness policies, *inter alia* [40].

Projections

Almost all aspects of the actuarial management of financial security systems like insurance companies and pension funds require the projection forward of the financial status and the underlying cash flows using a survival model, multiple state model, or multiple decrement model, as appropriate.

The methods used are essentially those of the component method of population projection, and can be traced back to Webster's early calculations for the Scottish Ministers' Widows Fund set up in 1743. Techniques, which were originally deterministic, have now been extended to allow for stochastic projections, based on **simulations**, of portfolios of policies and ultimately of companies.

In life insurance, financial projections require, *inter alia*, estimates of mortality rates and withdrawal rates. For pension schemes, the projections require estimates of rates of mortality, withdrawal, disability, and retirement. For health insurance, a multiple-state model with rates of incidence of disability, recovery, withdrawal, and mortality would be used. In these cases, it would be normal to attempt to model the variation of the probabilities with secular time (as well as age, for example), so that **extrapolations** can be made. The forecasting methods receiving most attention consist of regression based methods (using generalized linear models: for example [48]) and methods based on the Lee-Carter method (for example, [34, 35, 44].

In the case of financial calculations associated with annuities and pensions, it is important to note that the improving **life expectancy** and the downward secular trend in mortality rates (observed in many countries) (*see* **Morbidity and Mortality, Changing Patterns in the Twentieth Century**) need to be allowed for explicitly in the calculation of premiums and reserves. Failure to make such an allowance can have serious financial consequences for an insurer because improving life expectancy would mean that benefits would have to be paid for longer than anticipated. An example of modeling the impact of mortality trends as insurance portfolios is [38].

Similarly, where an upward trend in mortality is suspected, it is important to recognize this in life insurance calculations. This has been an important feature of recent discussions in respect of the impact of **AIDS** (see [14]).

For health insurance (based on the model in Figure 4), we would note the likely relationship between the probability of recovery and the employment prospects for the individual and the economic environment [24]. An important area of recent development has centered on the modeling of dependence between demographic risks and between demographic and financial risks. This application has been based on copulas [18, 49].

Graduation

Graduation may be regarded as the principles and methods by which a set of observed probabilities is adjusted to provide a suitable basis for inferences and further practical calculations to be made.

We consider for the moment a set of age-specific crude probabilities of death, $\mathring{q}_x$, or forces of mortality (i.e. hazard rates), $\mathring{q}_x$, which have been calculated from a set of observations. These values can each be regarded as a sample from a larger population and thus contain some random fluctuations. If we believed that the true q_x (or μ_x) were independent, then the crude values would be our final estimates of the true, underlying rates. However, a common, prior opinion about the form of these true rates is that each is closely related to its neighbors. This relationship is expressed by the belief that the true rates progress smoothly from one age to the next. So the next step is to graduate the crude rates to produce smooth estimates, $\tilde{q}_x$ (or $\tilde{\mu}_x$) of the true rates. This is done by systematically revising the crude values to remove the random fluctuations. This can be considered as a

cheaper and more practicable alternative to increasing the size of the original investigation.

The graduation process is an essential step in the construction of a survival model in ensuring that the model displays the required degree of smoothness. Then, the functions of practical importance calculated from the model (and leading to insurance premiums, reserves, surrender values, etc.) have results that share this important property of smoothness.

Graduation methods tend to fall under the categories *parametric* (*see* **Parametric Models in Survival Analysis**) and *nonparametric*. For a full review, readers should consult [2] and [36].

Parametric methods involve the fitting of a mathematical function to $\mathring{q}_x$ or $\mathring{\mu}_x$, with the parameters being determined by a formal procedure such as maximum likelihood estimation. Although in the context of the assumed function such methods are efficient (*see* **Efficiency and Efficient Estimators**), they are always liable to some degree of **bias** since no preassigned function will represent exactly the true (and unknown) values of q_x or μ_x. Nonparametric methods aim to give more stable estimates than the crude values by combining data at different values of x, but without presupposing any particular mathematical form for q_x or μ_x. Like parametric methods, they are liable to give biased estimates, but in such a way that it is possible to balance explicitly an increase in bias with a decrease in sampling variation. With nonparametric methods, like kernel methods (*see* **Density Estimation**), the amount of smoothing of the crude data can be varied continuously over a continuous range (e.g. by the choice of bandwidth). In contrast, the smoothness of parametric methods can only be regulated in discrete steps, for example by increasing the degree of the polynomial or by increasing the number of knots in a cubic **spline**. The properties of such curves will also tend to change abruptly. However, parametric methods are able to achieve higher degrees of smoothness than nonparametric methods through their use of explicit mathematical formulae, and may be more useful for extrapolation beyond the data range available.

Parametric Methods

We consider initially the graduation of an index of mortality like q_x or μ_x with respect to age.

Forfar et al. [17] give a comprehensive description of the methodology used in the UK to graduate survival models. We reformulate the methodology using the framework of **generalized linear models** (GLMs); for a full review see [27].

A GLM is characterized by independent response variables (Y_u with $u = 1, 2, \ldots, n$) with distribution specified by

$$\mathrm{E}(Y_u) = m_u, \quad \mathrm{var}(Y_u) = \frac{\phi V(m_u)}{\omega_u} \tag{19}$$

comprising a variance function V, a scale parameter $\phi(> 0)$, and prior weights ω_u. **Covariates** enter via a linear predictor,

$$\eta_u = \sum_{j=1}^{p} x_{uj}\beta_j, \tag{20}$$

with specified structure (x_{uj}) and unknown parameters β_j linked to the mean response through a known, differential, monotonic link function g with

$$g(m_u) = \eta_u. \tag{21}$$

The suffices or units u have a structure which is either intrinsic or imposed. The data comprise realizations (y_u) of the independent response variables, matched to the structure of the units. Generally, in any one study, the detail of the distribution and link are fixed, while the predictor structure may be varied.

Model fitting is by maximizing the quasi log likelihood (*see* **Quasi-likelihood**), leading to a system of equations in the unknown β_js which need to be solved numerically. Full details can be found in [37], which also describes the calculation of the **standard errors** for the parametric estimators, based on standard statistical theory. For members of the exponential family of distributions (*see* **Parametric Models in Survival Analysis**), the quasi log likelihood equates to the log likelihood.

The raw data would normally comprise the number of recorded deaths a_x accruing from matching exposures (or **person-years at risk**) r_x over a range of ages x, in a specific calendar period.

The approach would then be to model the actual numbers of deaths A_x as Poisson variables when targeting μ_x and as **binomial** variables when targeting q_x. Thus, for μ_x graduations with responses (A_x),

$$m_x = \mathrm{E}(A_x) = r_x\mu_{x+1/2},$$
$$V(m_x) = m_x, \quad \omega_x = 1, \quad \phi = 1,$$

and for q_x graduations with responses (A_x),

$$m_x = r_x q_x, \quad V(m_x) = m_x\left(1 - \frac{m_x}{r_x}\right),$$
$$\omega_x = 1, \quad \phi = 1.$$

The formulas underpinning a (parametric) graduation can be presented as predictor–link relationships with age x as the sole covariate.

The most common choice for η_x is to use polynomial predictors, or a set of orthogonal polynomials (*see* **Orthogonality**) for reasons of convenience of computing and interpretation. Splines and break-point predictors have also been used.

Parametric models of mortality have a long history, dating from Gompertz's "law" of 1825,

$$\mu_x = \exp(b_0 + b_1 x),$$

Makeham's modification of 1860,

$$\mu_x = a_0 + \exp(b_0 + b_1 x),$$

and Thiele's proposal of 1872,

$$\mu_x = \exp(b_0 - b_1 x) + a_1 \exp\left[-\frac{(x-c)^2}{2b_2}\right] + \exp(b_3 + b_4 x).$$

Later suggestions include the use of a **logistic** family of curves [39] and Heligman & Pollard's model involving a combination of double-exponential and **lognormal** curves for representing the odds function $[q_x/(1-q_x)]$ [30]. These mark progress towards a parametric model for the full age range.

Nonparametric Methods: Moving Weighted Average Graduation

Moving weighted average graduation methods (*see* **Moving Average**) were among the first nonparametric methods to be developed. The adjusted average formulae were largely developed by de Forest in the 1870s in a series of rather obscure papers which were rescued from oblivion and the results extended by Wolfenden [55]. For comments on the importance of de Forest's contributions see [50]. In this approach a weighted average of consecutive crude values is taken, i.e.

$$\hat{q}_x = \sum_{s=-m}^{s=m} a_s \mathring{q}_{x+s}. \tag{22}$$

The most successful formulas have symmetric coefficients $a_s = a_{-s}$. When considering the optimality of the coefficients a_s, it is useful to consider the crude rate as a random variable and express it as

$$\mathring{q}_x = q_x + r_x,$$

where q_x is the true rate and r_x is the residual error. An essential feature of any graduation is that the graduated rates should be smooth in some sense. With moving weighted averages (MWA), one approach is to choose weights that give the smoothest graduations, *ceteris paribus*. London [36] provides a fuller discussion of this approach.

The problems caused by MWA methods failing to give smoothed values of the first and last m observations have recently been addressed by Greville [23], among others.

Nonparametric Methods: Kernel Methods

Kernel estimation methods are used for estimating a probability density function (*see* **Density Estimation**). Thus, if $x_1, x_2, \ldots, x_n$ are some observed values of the random variable X, then the kernel estimate of the density at x is

$$\hat{f}(x) = \frac{1}{nb}\sum_{i=1}^{n} k_b(x - x_i), \tag{23}$$

where $k_b(x) \equiv k(x/b)$ is a kernel function which satisfies

$$\int_{-\infty}^{\infty} k(x)\,\mathrm{d}x = 1.$$

The bandwidth b governs the amount of smoothing which is applied. The larger the value of b is, the more smooth is the resulting estimate. In effect, a kernel density estimate is formed by placing a kernel function at each data point and then summing these functions to form the estimate. A more complete discussion of kernel density estimation is given in [45, 47].

We assume that for a set of ages x_i, $i = 1, \ldots, n$, we are given a measure of the exposed to risk e_i and the observed number of deaths d_i.

Two kernel estimators have been suggested, both of which are closely related to MWA graduation,

namely

$$\hat{q}_x^{\mathrm{CH}} = \frac{\sum_{i=1}^{n} d_i k_b(x - x_i)}{\sum_{i=1}^{n} e_i k_b(x - x_i)} \tag{24}$$

and

$$\hat{q}_x^{\mathrm{NW}} = \frac{\sum_{i=1}^{n} \mathring{q}_{x_i} k_b(x - x_i)}{\sum_{i=1}^{n} k_b(x - x_i)}. \tag{25}$$

In the context of graduation, $\hat{q}^{\mathrm{CH}}$ was introduced by Copas & Haberman [9] and $\hat{q}^{\mathrm{NW}}$ by Ramlau–Hausen [41]. The latter is related to the Nadaraya–Watson estimator and can be viewed as a continuous analogue of MWA graduation [19]. The choice of bandwidth is discussed in some detail in [20].

Nonparametric Methods: Whittaker–Henderson Methods

The nonparametric methods described above can be regarded as local, in the sense that the graduated value at a given age depends only on the observed values for arguments within a stipulated distance from the given argument. Global methods allow each graduated value to depend on all the observed data. The principal such method is Whittaker–Henderson graduation, devised by Whittaker [54].

The approach is based on a minimization of

$$S = \sum_{i=1}^{n} w_{x_i}(u_{x_i} - y_{x_i})^2 + h \sum_{i=1}^{n-s} (\Delta^s u_{x_i})^2, \tag{26}$$

where y_{x_i} denotes the crude values and u_{x_i} denotes the resulting graduated values at age x_i, with $i = 1, 2, \ldots, n$. S combines a measure of goodness of fit of the graduation and a measure of the smoothness of the sequence of graduated values, moderated by a positive parameter h chosen by the user to reflect the relative importance that they wish to attach to these conflicting characteristics. It is common to choose $s = 2$ or 3.

When $h = 0$, S is minimized when $u_{x_i} = y_{x_i}$ so that no graduation is needed. As h tends to 0, fit is emphasized over smoothness. When h becomes large, the second term dominates and in the limit the graduating curve becomes the least squares fitted polynomial of degree $s - 1$.

(w_{x_i}) is a set of positive weights chosen by the user, although it is common to choose for w_{x_i} the reciprocal of an estimate of the variance of the observation y_{x_i}. Then, the graduated values u_{x_i} are constrained to be close to the more reliable observations (i.e. those with smaller variances) and to be approximately a polynomial of degree $s - 1$, where the observations are less reliable.

We can rewrite (26) in matrix notation,

$$\mathbf{S} = (\mathbf{u} - \mathbf{y})' \mathbf{W} (\mathbf{u} - \mathbf{y}) + \boldsymbol{h} (\mathbf{K}\mathbf{u})' \mathbf{K}\mathbf{u}, \tag{27}$$

where $\mathbf{y}$ is the vector of observed values, $\mathbf{u}$ is the vector of graduated values, and $\mathbf{W}$ is the $n \times n$ diagonal matrix with successive diagonal elements equal to w_{x_i}. $\mathbf{K}$ is an $(n - s) \times n$ matrix with entries k_{ij}, where

$$k_{ij} = (-1)^{s+i-j} \binom{s}{j-i}.$$

It is then straightforward to show that $\mathbf{S}$ is minimized by $\mathbf{u}$ satisfying

$$(\mathbf{W} + \boldsymbol{h}\mathbf{K}'\mathbf{K})\mathbf{u} = \mathbf{W}\mathbf{y}. \tag{28}$$

As an extension, the loss function in (26) has been adapted to the fitting of a continuous curve, namely the smoothing spline of DeBoor [15]. This is discussed in more detail in [22].

A **Bayesian** interpretation of Whittaker–Henderson graduation has been provided by Taylor [52], and Verrall [53] has shown that the approach is equivalent to a dynamic **regression** analysis in which one parameter of the fitted line is allowed to vary stochastically. A two-dimensional version of Whittaker–Henderson graduation has been introduced by Knorr [33].

Other Nonparametric Methods

Related global methods include Bayesian [32] and information theoretic methods [4].

Tests of a Satisfactory Graduation

Two characteristics of a graduation require examination: smoothness and **goodness of fit** to the

observed data. These qualities are in competition, as is formalized in the criterion S for Whittaker–Henderson graduation; see (26).

The degree of smoothness required of a graduation is subjective and depends on the use to be made of the results; for applications in life insurance, it is essential that the resulting functions (e.g. premiums) are smooth. A parametric function (not a piecewise function) may, of course, be assumed to be smooth. In other cases, it has become customary to examine low-order finite differences of the graduated values and to consider their size and progression with respect to age.

For measuring the goodness of fit, it is common practice to tabulate the residuals, defined as the difference between the graduated value and the observed value at the relevant ages. The **diagnostics** are augmented by a variety of residual plots (*see* **Residuals for Survival Analysis**), including the normal and **half-normal** plots, and a battery of tests (including the standardized deviations test, cumulative deviations test, **serial correlations** test, **sign test**, changes of sign tests, and grouping of sign tests; see [2] for a full discussion).

Risk Classification and Regression

An important feature of insurance systems is the classification of risks for the purposes of fixing premiums. The classic economic argument in favor of risk classification is to combat **adverse selection**, the tendency of high risks to be more likely to buy insurance or to buy larger amounts than low risks [12].

In this context, the **proportional hazards** model of Cox [10] has become widely used for the modeling of the dependence of the force of mortality on a range of covariates (e.g. sex, blood pressure, weight). Following the notation of statistics, we let λ denote the force of mortality (or hazard rate) and then propose

$$\lambda(t, z_i) = \lambda^*(t) \exp\left(\sum_{j=1}^{p} z_{ij}\beta_j\right), \qquad (29)$$

where $\lambda(t, \mathbf{z}_i)$ is the hazard rate at time t for a person with known covariates given by the vector $\mathbf{z}_i$ (with elements z_{ij}), (β_i) is a set of parameters to be estimated, and $\lambda^*(t)$ is a baseline hazard rate at time t. Then each factor $\mathbf{z}_i$ enters the hazard in a multiplicative fashion. In this formulation, only $\lambda^*(t)$ depends on time, but the model can be adapted to feature time-dependent covariates.

If we assume that the $\lambda^*(t)$ values are known, then we can formulate the model as a GLM and hence produce parameter estimates for the (β_j); examples are provided by [26, 43].

Results from such studies of insurance mortality have demonstrated that the total mortality risk can be represented by a statistical model involving a linear combination of a number of factors (possibly with **interactions**). The results have proved very useful for insurance purposes. Further, the major medico-actuarial studies of mortality and survival experience of insured lives characterized by a range of covariates have been of considerable importance in public health terms – for example, the link between build, blood pressure, and mortality demonstrated by the Build and Blood Pressure Studies in the US [6, 7].

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Benjamin, B. & Pollard, J.H. (1993). *The Analysis of Mortality and Other Actuarial Statistics*, 3rd Ed. Institute and Faculty of Actuaries, Oxford.

[3] Bowers, N.L., Gerber, H.U., Jones, D.A., Hickman, J.C. & Nesbitt, C.J. (1986). *Actuarial Mathematics*. Society of Actuaries, Chicago.

[4] Brockett, P.L. (1991). Information theoretic approach to actuarial science: a unification and extension of relevant theory and applications, *Transactions of the Society of Actuaries* **43**, 73–144.

[5] Broffit, J.D. (1984). Maximum likelihood alternatives to actuarial estimators of mortality rates, *Transactions of the Society of Actuaries* **36**, 77–142.

[6] *Build and Blood Pressure Study 1959* (1960). Society of Actuaries, Chicago.

[7] *Build and Blood Pressure Study 1979* (1980). Society of Actuaries, Chicago.

[8] Continuous Mortality Investigation Bureau (1991). Continuous Mortality Investigation Report No. 12. Institute and Faculty of Actuaries, Oxford.

[9] Copas, J.B. & Haberman, S. (1983). Non-parametric graduation using kernel methods, *Journal of the Institute of Actuaries* **110**, 135–156.

[10] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–200.

[11] Cox, D.R. & Miller, H.D. (1965). *The Theory of Stochastic Processes*. Chapman & Hall, London.

[12] Cummins, J.D., Smith, B.E., Vance, R.N. & Van der Hei, H.L. (1982). *Risk Classification in Life Insurance*. Kluwer, Boston.

[13] Daw, R.H. (1979). Smallpox and the double decrement table: A piece of actuarial pre-history, *Journal of the Institute of Actuaries* **106**, 299–318.

[14] Daykin, C.D., Clark, P.N.S., Eves, M.J., Haberman, S., Lockyer, J., LeGrys, D.J., Michaelson, R.W. & Wilkie, A.D. (1988). The impact of HIV infection and AIDS on insurance in the United Kingdom, *Journal of Institute of Actuaries* **115**, 727–837.

[15] DeBoor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

[16] Elandt-Johnson, R.G. & Johnson, N.L. (1980). *Survival Models and Data Analysis*. Wiley, New York.

[17] Forfar, D.O., Wilkie, A.D. & McCutcheon, J.J. (1988). On graduation by mathematical formula, *Journal of the Institute of Actuaries* **115**, 1–149.

[18] Frees, E.W. & Valdez, E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* **2**(1), 1–25.

[19] Gavin, J., Haberman, S. & Verrall, R.J. (1993). Moving weighted average graduation using kernel estimation, *Insurance: Mathematics and Statistics* **12**, 113–126.

[20] Gavin, J., Haberman, S. & Verrall, R.J. (1994). On the choice of bandwidth for kernel graduation, *Journal of the Institute of Actuaries* **121**, 119–134.

[21] Gerber, H.U. (1995). *Life Insurance Mathematics*, 2nd Ed. Springer-Verlag, Zurich.

[22] Green, P.J. & Silverman, B.W. (1994). *Non Parametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London.

[23] Greville, T.N.E. (1981). Moving weighted average smoothing extended to extremities of the data. I: Theory, *Scandinavian Actuarial Journal* **1981**, 38–55.

[24] Haberman S. (1987). Long-term sickness and invalidity benefits: Forecasting and other actuarial problems, *Journal of the Institute of Actuaries* **114**, 467–533.

[25] Haberman, S. & Pitacco, E. (1999). *Actuarial Models for Disability Insurance*. CRC Press, Bocce Raton.

[26] Haberman, S. & Renshaw, A.E. (1990). Generalized linear models and excess mortality from peptic ulcers, *Insurance: Mathematics and Economics* **9**, 21–32.

[27] Haberman, S. & Renshaw, A.E. (1996). Generalized linear-models and actuarial science, *Statistician* **45**, 407–436.

[28] Haberman, S. & Sibbett, T. (1995). *History of Actuarial Science*. Pickering & Chatto, London.

[29] Hald, A. (1990). *A History of Probability and Statistics and Their Applications Before 1750*. Wiley, New York.

[30] Heligman, L. & Pollard, J.H. (1980). The age pattern of mortality, *Journal of the Institute of Actuaries* **107**, 49–74.

[31] Jones, B.L. (1994). Actuarial calculations using a Markov model, *Transactions of the Society of Actuaries* **46**, 227–250.

[32] Kimeldorf, G.S. & Jones, D.A. (1967). Bayesian graduation, *Transactions of the Society of Actuaries* **19**, 66–112.

[33] Knorr, F.E. (1984). Multidimensional Whittaker-Henderson graduation, *Transactions of the Society of Actuaries* **36**, 213–240.

[34] Lee, R.D. (2000). The Lee-Carter Method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal* **4**(1), 80–93.

[35] Lee, R.D. & Carter, L. (1992). Modelling and Forecasting the time series of US mortality. *Journal of the American Statistical Association*, **87**, 659–671.

[36] London, D. (1985). *Graduation: The Revision of Estimates*. Actex, Winsted.

[37] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[38] Olivieri, A. (2001). Uncertainty in mortality projections: an actuarial perspective, *Insurance; Mathematics and Economics* **29**, 231–245.

[39] Perks, W. (1932). On some experiments in the graduation of mortality statistics, *Journal of the Institute of Actuaries* **63**, 12–40.

[40] Pitacco, E. (1995). Actuarial models for pricing disability benefits: Towards a unifying approach, *Insurance: Mathematics and Economics* **16**, 39–62.

[41] Ramlau-Hansen, H. (1983). The choice of a kernel function in the graduation of counting process intensities, *Scandinavian Actuarial Journal* **1983**, 165–182.

[42] Ramsay, C.M. (1989). AIDS and the calculation of life insurance functions, *Transactions of the Society of Actuaries* **41**, 393–422.

[43] Renshaw, A.E. (1988). Modelling excess mortality using GLIM, *Journal of the Institute of Actuaries* **115**, 299–315.

[44] Renshaw, A.E. & Haberman, S. (2003). Lee-Carter mortality forecasting: a parallel generalized linear modeling approach for England and Wales mortality projections. *Journal of the Royal Statistical Society Series C* **52**, 1–19.

[45] Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York.

[46] Seal, H.L. (1977). Studies in the history of probability and statistics XXXV. Multiple decrements or competing risks, *Biometrika* **64**, 429–439.

[47] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[48] Sithole, T.Z., Haberman, S. & Verrall, R.J. (2000). An investigation into parametric models for mortality projections, with applications to immediate annuitants' and life office pensioners' data. *Insurance: Mathematics and Economics* **27**, 285–312.

[49] Sklar A. (1973). Random variables, joint distribution functions and copulas. *Kybernetika*, **9**, 449–460.

[50] Stigler, S.M. (1978). Mathematical statistics in the early states, *Annals of Statistics* **6**, 239–265.

[51] Sverdrup, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries

and transfers between different states of health, *Skandinavisk Aktuaritidskrift* **48**, 184–211.

[52] Taylor, G.C. (1992). A Bayesian interpretation of Whittaker-Henderson graduation, *Insurance: Mathematics and Economics* **11**, 7–16.

[53] Verrall, R.J. (1993). A state space formulation of Whittaker graduation, with extensions, *Insurance: Mathematics and Economics* **13**, 7–14.

[54] Whittaker, E.T. (1923). On a new method of graduation, *Proceedings of the Edinburgh Mathematical Society* **41**, 63–75.

[55] Wolfenden, H.H. (1925). On the development of formulae for graduation by linear compounding, with special reference to the work of Erastus L. De Forest, *Transactions of the Actuarial Society of America* **26**, 81–121.

(*See also* **Demography**)

STEVEN HABERMAN

Adaptive and Dynamic Methods of Treatment Assignment

The simplest design of a randomized clinical trial is to enter a predetermined number of patients (i.e. use a fixed sample size) and to assign treatment by randomization with equal probability for each patient. In practice, trials are rarely conducted in this fashion. More commonly, both the manner in which patients are allocated a treatment, and the decision to terminate the study, are based on patient-specific information that accumulates during the progress of the trial. The terminology used to identify the different kinds of methods has not been consistent. However, in this article the following taxonomy will be used to classify the methods. A *dynamic* allocation method is one in which information on patient covariates that predict the clinical outcome is used to determine the treatment assignment. By contrast, an *adaptive* allocation method is one that uses accumulating outcome data to affect the treatment selection. In the broad context of adaptive designs, *sequential* designs (*see* **Sequential Analysis**) are prespecified analytic rules that guide the decision to terminate the trial on the grounds that the evidence favoring one of the treatments has become persuasive.

Dynamic Treatment Allocation

A completely randomized design (*see* **Experimental Design**) is relatively simple to implement and prevents selection bias. It also ensures that all hypothetical permutations of the treatment assignments are equiprobable, under the null hypothesis, and thus forms the basis for a conventional permutation test, if this is the analysis of choice. The disadvantages of complete randomization include inefficiency in small trials, due primarily to the risk of imbalanced treatment totals, and the possibility that important prognostic factors may also be imbalanced by chance, reducing the credibility of the results of the trial. The simplest way to avoid imbalance in treatment totals is to randomize groups of individuals in "blocks", with equal numbers of each treatment in each block (*see* **Randomized Treatment Assignment**). Imbalance in important prognostic factors can be reduced by allocating randomly permuted blocks within the strata (*see* **Stratification**) defined by the factors [26]. This method, randomly permuted blocks in strata, is probably the most widely used randomization method, and it is easily implemented, since all of the allocation sequences can be prescribed before the start of the trial, by creating sequences of blocks for each stratum. That is to say, the dynamic aspects of this method are embedded in the stratum-specific sequences of allocations, and so no dynamic calculations are necessary in the course of the trial to determine the next treatment allocation.

The method of randomly permuted blocks in strata rapidly degenerates as the number of strata increases. For example, a trial with five stratification factors, and three categories for each factor, would have $3^5 = 243$ distinct strata. Therefore, in the course of the trial many of the strata will accrue few, if any, patients (unless the sample size is very large), rendering the blocking ineffective. In effect, the method reduces to complete randomization as the number of strata increases. To offset this problem there are a number of methods that balance the factors individually, i.e. marginally, without requiring balance within all factor combinations.

Suppose that there are f factors, and l_f levels in factor f. At any given point in the trial the treatment allocations of the previous patients will have created some amount of imbalance among the factors. Let t_{ijk} be the total number of patients in the jth level of factor i that have been allocated to treatment $k, i = 1, \ldots, f, j = 1, \ldots, l_f, k = 1, \ldots, r$, where r is the number of treatments. The trial is balanced for factor i level j to the extent that $t_{ij1}, \ldots, t_{ijr}$ are similar. If the next patient to be randomized possesses factor i at the jth level, then one can consider the effect that each of the possible treatment allocations would have on this balance. Balance must be characterized by a mathematical function. Taves [21] proposed the popular minimization method, where balance is characterized by the range of treatment totals, and the treatment is selected by minimizing the sum of the ranges across all of the factors. Pocock & Simon [18] proposed a more general version of this method in which the treatment is selected by a biased coin randomization, with the biased coin probabilities determined by the balancing function. They suggested the use of either the range or the variance as balancing functions. Their overall balancing function

involves a weighted sum of the balancing functions of the individual factors, where the weights could be assigned on the basis of the relative importance of the prognostic factors. That is, if $t^*_{ijk}(k)$ represents the treatment totals if treatment k is allocated, and if $F_{ij}[t^*_{ijk}(k)]$ is the balance for the jth level of the ith factor under these circumstances, then the overall balance is

$$B_k = \sum_i \sum_j w_i F_{ij}[t^*_{ijk}(k)].$$

Note that only the unique levels of each factor for the new patient are affected by the choice of k. The values of B_k are especially easy to update and compute if the variance is used as the balancing function [11]. The biased coin probabilities are then determined on the basis of B_k. For example, if the r treatment assignments are ranked from the one that would lead to the least imbalance, $k = 1$, to the one that would lead to the greatest imbalance, $k = r$, then one of the formulas suggested by Pocock & Simon is to select $p_1 > r^{-1}$, and set $p_k = (1 - p_1)/(r - 1)$ for $k = 2, \ldots, r$. In this case the degree of randomization is inversely related to p_1, and the design is fully randomized if $p_1 = r^{-1}$. The use of biased coins in this context, rather than deterministic allocations, was originally proposed by Efron [10], in part to ensure that the trial is truly "randomized", enabling the calculation of an appropriate reference distribution for a permutation test (making use of the biasing probabilities), and in part to reduce the risk of **selection bias**.

Numerous other treatment allocation schemes have been proposed. Notably, it has been shown that balance is a characteristic of design optimality for the linear model [7], and efficient designs have been developed in the context of the theory of **optimal design** [3]. Various simulation studies and general empirical evidence demonstrate that all of these algorithms are effective at balancing stratification factors, even very early in the trial when there is a risk that the trial might have to be terminated unexpectedly. The numerous proposed methods have been reviewed in detail [14].

The validity of conventional statistical tests subsequent to the use of stratified or minimization-type schemes has been a topic of debate. In general, stratification has the effect of making the treatment groups more alike that would be expected by chance. This tends to make the unadjusted estimator of the treatment effect more precise, but the variance estimator is positively biased, and thus unadjusted statistical tests are conservative. Simon provides a review of historical discussion of this issue in the context of agricultural experiments in the 1930s [19]. To correct for this effect, it is necessary to perform a stratified analysis, stratified by the same factors employed in the design. This may be inconvenient if numerous factors were used in a minimization-type scheme. However, the distortion of the p values is only substantial for strong prognostic factors, and so it will typically be unnecessary to adjust for all factors in the analysis. Biased-coin designs affect the validity of standard permutation tests owing to the fact that different allocation sequences are not equiprobable, and it is possible in theory to correct this problem by simulating the correct reference distribution [19].

Adaptive Designs

Adaptive designs, i.e. designs which depend on the accumulating outcome data, have been researched and discussed extensively since the 1950s. Pioneering work in this area was accomplished by Armitage, who adapted the sequential probability ratio test for application to medical trials [2]. Such a scheme allows for a formal termination rule at any time based on a global significance level. That is to say, it accounts for the fact that multiple analyses of the data will increase the chances of a false positive finding, and so the stopping boundaries are adjusted to offset this multiplicity problem. For many years this methodology appears to have been well known but little used. However, a series of papers in the late 1970s and early 1980s succeeded in popularizing the concept, via the development of group sequential stopping rules, in which a relatively small number of pre-specified interim analyses are envisaged (*see* **Data and Safety Monitoring**). These new methods were developed in recognition of the fact that large multicenter trials are usually subject to regular analyses by data-monitoring committees (*see* **Data Monitoring Committees**). The simplest method involves setting a single significance level for each analysis [17]. However, it appears that methods with very strict criteria early in the trial, and a final criterion close to the nominal level (e.g. 5%), such as the O'Brien and Fleming rule, are more popular [15].

An entirely different formulation of this problem also led to much research and debate, stemming

from the ideas of Anscombe [1] and Colton [9]. This approach is designed to optimize the stopping rule on the basis of an appropriate **loss function**, in contrast to the arbitrariness inherent in using significance tests. To do this, it is necessary to construct a "patient horizon", i.e. the total number of patients either on the trial or affected by the trial results in the future via the choice of the best treatment. In this model the responses to each treatment are assumed normally distributed with equal variances, and patients are randomized until the boundary is crossed, after which all remaining patients are assigned to the superior treatment up to the patient horizon. The optimal boundary is evaluated by trading the losses incurred by randomizing half the patients to the inferior treatment, and the losses incurred by making the wrong decision and assigning all future patients, up to the horizon, to the inferior treatment. Tabulated boundaries for this problem are provided by Chernoff & Petkau [8]. Even greater optimization is theoretically possible by optimizing the proportions randomized to the treatments on the basis of the emerging data [5, 13]. A perceived problem with this kind of approach is that the formulation is considered by most experts to be too simplistic to be a credible approximation to the realities of clinical research [16]. The patient horizon is a spuriously precise expression of a vague concept. As a result, this approach is not used widely.

A closely related formulation is the two-armed bandit problem [20]. Zelen popularized this concept in the context of medical trials, calling it the play-the-winner rule [25]. Conceptually this rule involves randomly selecting treatments using urn-sampling, where the numbers of balls in the urn are changed as outcomes are recorded. If outcomes are recorded immediately, i.e. before the next allocation, then a modified play-the-winner rule assigns the subsequent patient to the same treatment following a successful outcome, and to the opposite treatment following a failure. Generalizations to this idea have been studied by numerous investigators, especially randomized versions that do not allow deterministic allocations [24]. The basic rationale presented for play-the-winner (or biased coin) adaptive designs is that it is preferable on ethical grounds to assign more patients to the treatment that appears to be generating the superior outcomes, and indeed it has been shown that such designs do allocate fewer patients to the "inferior" treatment compared with an equal allocation design, after fixing the probability of a correct selection [25]. Although this method has not been used frequently in practice, it was employed in a highly controversial study of extracorporeal membrane oxygenation therapy (ECMO) in newborn infants [4]. This trial was concluded after 12 patients were treated, only one of whom was allocated the control treatment (the only failure in the study). The subsequent permutation-based analysis, calculated on the basis of the biased-coin design, led to a marginally significant result [23]. However, the methodology received much criticism [6], and a subsequent confirmatory trial involving a randomized consent design (*see* **Ethics of Randomized Trials**) also led to great controversy [22].

The decision to continue or terminate a clinical trial on the basis of the available evidence is a highly charged issue that continues to engender debate among statisticians, clinical investigators, and ethicists, and even the role of randomization continues to be disputed. Frequently, in the course of a trial, relevant data from a related trial or a meta-analysis (*see* **Meta-analysis of Clinical Trials**) become available, and this may influence the decision to continue or terminate the study. The merits of formalizing the use of such information have been debated at length [12].

References

[1] Anscombe, R.F. (1963). Sequential medical trials, *Journal of the American Statistical Association* **58**, 365–383.

[2] Armitage, P. (1975). *Sequential Medical Trials*, 2nd Ed. Wiley, New York.

[3] Atkinson, A.C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors, *Biometrika* **69**, 61–67.

[4] Bartlett, R.H., Roloff, D.W., Cornell, R.G., Andrews, A.F., Dillon, P.W. & Zwishenberger, J.B. (1985). Extra corporeal circulation in neonatal respiratory failure: a prospective randomized study, *Pediatrics* **76**, 479–487.

[5] Bather, J.A. (1981). Randomized allocation of treatments in sequential experiments, *Journal of the Royal Statistical Society, Series B* **43**, 265–292.

[6] Begg, C.B. (1990). On inferences from Wei's biased coin design for clinical trials (with discussion), *Biometrika* **77**, 467–484.

[7] Begg, C.B. & Iglewicz, B. (1980). A treatment allocation procedure for sequential clinical trials, *Biometrics* **36**, 81–90.

[8] Chernoff, H. & Petkau, A.J. (1981). Sequential medical trials involving paired data, *Biometrika* **68**, 119–132.

[9] Colton, T. (1963). A model for selecting one of two medical treatments, *Journal of the American Statistical Association* **58**, 388–400.

[10] Efron, B. (1971). Forcing a sequential experiment to be balanced, *Biometrika* **58**, 403–417.

[11] Freedman, L.S. & White, S.J. (1976). On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial, *Biometrics* **32**, 691–694.

[12] Geller, N., Freedman, L.S., Lee, Y.J. & Dersimonian R., eds (1996). Conference on Meta-Analysis in the Design and Monitoring of Clinical Trials, *Statistics in Medicine* **15**, 1233–1323.

[13] Gittins, J.C. (1979). Bandit processes and dynamic allocation indices, *Journal of the Royal Statistical Society, Series B* **41**, 148–177.

[14] Kalish, L.A. & Begg, C.B. (1985). Treatment allocation methods in clinical trials: a review, *Statistics in Medicine* **4**, 129–144.

[15] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[16] Peto, R. (1985). Discussion of the papers by J.A. Bather and P. Armitage, *International Statistical Review* **53**, 31–34.

[17] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.

[18] Pocock, S.J. & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial, *Biometrics* **31**, 103–115.

[19] Simon, R. (1979). Restricted randomization designs in clinical trials, *Biometrics* **35**, 503–512.

[20] Smith, C.V. & Pyke, R. (1965). The Robbins-Isbell two-armed bandit problem with finite memory, *Annals of Mathematical Statistics* **36**, 1375–1386.

[21] Taves, D.R. (1974). Minimization: a new method of assigning patients to treatment and control groups, *Clinical Pharmacology and Therapeutics* **15**, 443–453.

[22] Ware, J.H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion), *Statistical Science* **4**, 298–340.

[23] Wei, L.J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule, *Biometrika* **75**, 603–606.

[24] Wei, L.J. & Durham, S. (1978). The randomized play-the-winner rule in medical trials, *Journal of the American Statistical Association* **73**, 840–843.

[25] Zelen, M. (1969). Play the winner rule and the controlled clinical trial, *Journal of the American Statistical Association* **64**, 131–146.

[26] Zelen, M. (1974). The randomization and stratification of patients to clinical trials, *Journal of Chronic Diseases* **27**, 365–375.

COLIN B. BEGG

Adaptive Designs for Clinical Trials

If investigators planning a clinical trial knew all that was necessary to design it, they would select an optimal statistical procedure to test the primary hypothesis of interest. In practice, however, one rarely has sufficient information available at the time one designs a trial – the variability of the primary outcome measure may be unknown, the effect size uncertain, and the expected compliance to therapy, a conjecture. Particularly for a long-term study, modifications to standard medical practice that may occur as the trial progresses may produce unanticipated changes that affect parameters used to design the trial. Lacking firm estimates for parameters integral to the design of a study, the investigators may choose to sacrifice statistical optimality in exchange for an adaptive design that provides flexibility. This section discusses a variety of such adaptive designs for clinical trials.

One can imagine many different types of adaptations. Some adaptive designs, by changing the allocation ratio during the course of the trial (play the winner trials or drop the loser trials) aim to increase the probability of assigning the best treatment to the participants in the trial (*see* **Adaptive and Dynamic Methods of Treatment Assignment**). Other adaptive designs incorporate aspects of both a *dose-finding* and *confirmatory study*; others may allow change in endpoint or modifications of entry criteria. This article discusses a class of adaptive designs that modify sample size during the course of the trial. Two types of such designs are available – those whose purpose is to end a study early if the answers are clear and those whose purpose is to increase the information in a trial, either by increasing sample size or length of follow-up, to maintain desired statistical power. We deal here primarily with two-stage adaptive designs. For a general description of the theoretical underpinnings of two-stage adaptive designs, see [16].

The types of designs considered in this article encompass the classes of design with an experimental and control arm (*see* **Clinical Trials, Overview**), a preselected primary endpoint, and a criterion specifying the requirement to preserve, or nearly preserve, the preselected *Type I error* rate. These designs aim to prevent bias, not only the technical bias defined by inflation of the Type I error rate, but also bias that may creep into the study by loosening the protective firewalls that separate the blinded data from those involved in the conduct of the study. In particular, the article addresses **sequential analysis**, *futility analysis*, conditional power (*see* **Cooperative Heart Disease Trials**), and designs that permit changes to sample size or follow-up time in response to internal estimates of either variability or effect size. Designs that allow changes to sample size on the basis of these internal estimates are called "internal pilot" designs [23]. Some authors reserve the word "adaptive" for the special case of internal pilot designs that use effect size to modify the sample size. This article does not discuss more general adaptive designs that allow such changes as dropping a study arm during the course of the trial, redefining the primary endpoints, selecting a different test statistic, modifying the study population, or changing the allocation ratio during the study.

The oldest type of adaptive design goes by the name "sequential analysis". These methods, now part of the standard tools of biostatistics, have been widely used for several decades and experienced clinical trialists understand their properties well. While sequential designs are less efficient than the optimal fixed sample design, most schemes in common use incur only small losses in **efficiency**. They allow a trial to stop early with the declaration of statistically significant benefit for the treated group. Many clinical investigators expect to see a sequential plan as part of a clinical trial, especially a long-term trial or a trial with a clinical outcome. In designing such trials, biostatisticians should think not only of the primary endpoint, but also about supportive endpoints and subgroups of potential importance. Stopping a trial early may allow declaration of success for the primary endpoint, but if the sample size is small at the time the study ends, the observed effect size may considerably overestimate the true effect. Moreover, the results may have ambiguous interpretations for other important questions.

A second type of adaptive approach that permits early termination with protection of Type I error rate is the so-called futility analysis. A trial may be declared futile if the experimental therapy is not so bad as to be unsafe, but if the probability of showing benefit is low. In this type of design, the group watching the trial, often the **Data Monitoring Committee**, may recommend ending the trial early if

it assesses that continuing is "futile". The method of assessing futility may be based on conditional power [13] and the B-value [14]; it may be based on a **confidence limit**, or a boundary for excluding a specific effect [8]. The criterion for defining futility often depends on the secondary objectives of the trial. A trial examining proof-of-concept may stop early for futility if the concept appears unfounded; similarly, a confirmatory trial, or a trial that follows an unsuccessful one, may have a low threshold for stopping early and declaring futility. Designers of a first Phase 3 trial (*see* **Clinical Trials, Overview**) however, may be reluctant to stop for futility because they intend to use data from the trial to learn a lot about the new therapy and early stopping produces too little information for rich exploration of the data. By the same token, designers of a trial of a therapy in common use may wish to continue the trial even if the chance of finding benefit is low because clear evidence of no efficacy is important for the public health. Futility analyses that are based on conditional power rely on the stochastic independence of nonoverlapping periods of the trial [14].

Internal pilot designs, unlike sequential analysis and futility analysis, incorporate the possibility of increased sample size. Two classes of such designs are available: those that use data internal to the trial to reestimate one or more **nuisance parameters** and those that reestimate the effect size. In both cases, the new estimate provides the basis for recalculating sample size. All these designs provide a hedge against having made poor estimates of parameters in designing the trial; however, this hedge can be costly. Midcourse estimates, which are based on a fraction of the total sample size originally projected, are often imprecise. This imprecision is more serious for designs that aim to estimate effect size than they are for those that estimate nuisance parameters.

The simplest approaches use data from the first, or internal pilot phase, to estimate **variance**, and then apply this new estimate to a **sample size** formula. A paper by Stein spawned these methods [22]. Many variants are available. Some use unblinded data [3, 22, 23], some **blinded** [11]. Some use formulas that do not correct for potential inflation of Type I error rate, and some address the inflation directly. Methods are available for **normal** [3–5, 7, 11, 23, 24, 26] and **binomial** [10, 12] tests as well as for repeated measures (*see* **Multiplicity in Clinical Trials**) [25].

The blinded and unblinded approaches have different properties. From the point of view of statistical operating characteristics, the blinded versions are generally preferable when the specified effect size is close to the true effect [26]. Especially for binomial and time-to-failure outcomes (*see* **Survival Analysis, Overview**), choosing between blinded and unblinded assessment requires balancing the risk of overestimating the sample size, which can occur in the blinded cases, with providing too much information to the investigators, which can occur in unblinded cases. For example, consider a trial designed to demonstrate a difference in proportion of failure from 0.4 to 0.3. In the blinded case, one would expect a pooled event rate of 0.35. Seeing a rate lower than that would prompt an increase in sample size. If, however, the observed rates were 0.4 and 0.1, that is, a 75% reduction at the first stage, the observed pooled rate of 0.25 would lead to a considerable, and unnecessary, increase in sample size. A method based on the estimated **placebo** rate alone would leave the sample size unchanged.

Both blinded and unblinded methods protect the Type I error rate quite well. In fact, even naïve estimates that simply calculate the sample size at the second stage using the estimated rate in the placebo without any correction for having made an interim look at the data incur only minimal inflation in Type I error rate [23]. Other methods, for example, [5] correct for the look and hence preserve Type I error rate more precisely. Thus, the criterion for selecting a method for increasing sample size on the basis of internal estimates of information (variance, proportion, or **hazard** should be based less on the operating characteristics (*see* **Animal Screening Systems**) of the procedures and more on practical contingencies. If the study team can separate the estimates from the operation of the study, then unblinded assessments based on the placebo rates may be appealing; if they cannot, then blinded methods are preferable.

Several very different methods are available for internal pilot designs that use effect size as the basis for changing sample size. Some approaches combine ***p*-values** from the two stages [1]; some base sample size on conditional power [19]; some rely on unequal weighing of data from the two phases of the study [6, 15, 20].

Procedures that allow increasing the sample size to control the conditional power implicitly permit increasing the sample size if the effect size at

the interim look is low. The procedure calculates the conditional power at the first stage and then increases the sample size to maintain desired conditional power. Consider a two-stage trial that tests a one-sided (*see* **Alternative Hypothesis**) **null hypothesis** H_0. At each stage one calculates p-values, p_1 and p_2. If $p_1 \leq \alpha_1 < \alpha$, the trial stops and rejects H_0. If $p_1 > \alpha_0 > \alpha$, the trial stops for futility. If $\alpha_1 \leq p_1 < \alpha_0$, the trial continues to the second stage and the final decision is based on a combination function $C(p_1, p_2)$ that rejects (does not reject) H_0 if $C(p_1, p_2) \leq c(> c)$. If $\alpha_0 = 1$, the trial will not stop for futility and if $\alpha_0 = 1$, the trial will not stop early [2]. The conditional error function [19], $CE(p_1) = \text{Prob}(\text{reject } H_0|p_1)$, is the probability of rejecting the null hypothesis conditional on observing a p-value of p_1 at the first stage. Let $A(p_1)$ be a monotonic function such that $\alpha_1 + \int_{\alpha_1}^{\alpha_0} A(p_1)\,\mathrm{d}p_1 = \alpha$. A rule that rejects H_0 when $p_1 \leq \alpha_1$ or when $\alpha_1 < p_1 \leq \alpha_0$ and $p_2 \leq A(p_1)$ controls the Type I error rate. If $\alpha_1 < p_1 \leq \alpha_0$ then $A(p_1)$ is the conditional error function [18]. For a description of various conditional error functions, see [19].

One simple, very flexible, two-stage method calculates the z-statistic (*see* **Standard Normal Deviate**), z_1 halfway though the trial [20]. The investigators decide on the basis of z_1 whether to increase the sample size (decreases are not allowed). If they decide not to change the sample size, the study continues to its planned end. If they change the sample size, they may use any method to calculate the second stage sample, n_2. The rejection region is $z^* = (z_1 + z_2)/2^{1/2} > z_\alpha$ and the p-value is $1 - \Phi\{(z_1 + z_{2)}/2^{1/2}\}$. Because Z_1 and Z_2 are independent **standard normal variates** for any sample size function $n_2(z_1)$, z^* also has a standard normal distribution. If the sample size remains unchanged, then the loss of efficiency is very small because z^* is only slightly larger than the usual fixed-sample z-score.

An option for one-sided tests uses a variance-spending sequential method [9], a technique that allows one to change sample size in response to an effect size different from expected. In these designs, one constructs a final test statistic using a weighted average of the sequentially collected data, where the observed data prior to each stage determines the weight function for that stage [21]. The goal is to terminate a trial early when the treatment effect is large or when the new therapy is harmful but to ensure an adequate sample size when the true effect is small. The final test statistic is a weighted average of the test statistics at each stage. One selects the weights to maintain the variance of the final test statistic in order to preserve the Type I error rate. Thus, in general, not all observations have equal weight.

Two other approaches combine sequential analysis with potential increases in sample size [6, 15]. These methods, like the variance-spending sequential methods above, maintain the Type I error rate by assigning different weights to different stages of the study. Consider a trial with K planned interim analyses. At interim analyses $1, 2, \ldots, L$, perform the prespecified group sequential test. At the Lth interim analysis, if the monitoring boundary is not crossed, then adjust the sample size, up or down, by the factor $N(\delta/\Delta_L)$ where N is the original sample size per group and Δ_L the observed treatment difference at the Lth analysis. Both approaches [6] and [15], though slightly different in theory and implementation, have similar properties.

While the ability to modify one's study on the basis of an observed effect has considerable appeal (you *can* have your cake and eat it), the methods are not problem free. The estimate of effect size can be quite biased, so special estimators must be employed. The estimated treatment effect at the end of the first stage may be imprecise, so the recalculated sample size may be either too large or too small and there is some question about the propriety of changing the effect size one wants to detect. Moreover, such designs can be extremely inefficient relative to comparable fixed sample size designs or classical sequential designs [17].

None of these methods should substitute for a careful design. Classical sequential analysis, futility analysis, and sample size recalculation on the basis of reestimated nuisance parameters incur little loss of efficiency; however, adaptive methods that use effect size to change sample size may be very inefficient, requiring much larger sample sizes than a well-designed sequential plan. Thus, one should be cautious in the use of this type of adaptation. During the design phase, one should identify the parameters projected with greatest uncertainty and select midcourse changes specifically to address those uncertainties. A strategy that deliberately chooses a low sample size in the hope that an adaptive design will bail one out courts serious inefficiency. On the other hand, failing to modify one's plans when the

data from a trial show important deviations from the expected may render the results of a trial ambiguous.

References

[1] Bauer, P. (1989). Multistage testing with adaptive designs, *Biometrie und Informatik in Medizin and Biologie* **20**, 130–148.

[2] Bauer, P. & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses, *Biometrics* **50**, 1029–1041 (Correction, *Biometrics* **52**, 380, 1996).

[3] Birkett, M. & Day, S. (1995). Internal pilot studies for estimating sample size, *Statistics in Medicine* **13**, 2455–2463.

[4] Bristol, D. (1993). Sample size determination using an interim analysis, *Journal of Biopharmaceutical Statistics* **3**, 159–166.

[5] Coffey, C. & Muller, K. (1999). Exact test size and power of a Gaussian error linear model for an internal pilot study, *Statistics in Medicine* **18**, 1199–1214.

[6] Cui, L., Hung, H.M.J. & Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials, *Biometrics* **55**, 853–857.

[7] Denne, J. & Jennison, C. (1999). Estimating the sample size for a t-test using an internal pilot, *Statistics in Medicine* **18**, 1575–1585.

[8] Emerson, S. & Fleming, T. (1989). Symmetric group sequential test designs, *Biometrics* **45**, 905–923.

[9] Fisher, L.D. (1998). Self-designing clinical trials, *Statistics in Medicine* **17**, 1551–1562.

[10] Gould, A. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate, *Statistics in Medicine* **11**, 55–66.

[11] Gould, A. & Shih, W. (1991). Sample size reestimation without unblinding for normally distributed outcomes with unknown variance, *Communications in Statistics* **21**, 2833–2853.

[12] Herson, J. & Wittes, J. (1993). The use of interim analysis for sample size adjustment, *Drug Information Journal* **27**, 753–760.

[13] Lan, K., Simon, R. & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials, *Communications in Statistics* **C1**, 207–219.

[14] Lan, K. & Wittes, J. (1988). The B-value: a tool for monitoring data, *Biometrics* **44**, 579–585.

[15] Lehmacher, W. & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials, *Biometrics* **55**, 1286–1290.

[16] Liu, Q., Proschan, M.A. & Pledger, G.W. (2002). A unified theory of two-stage adaptive designs, *Journal of the American Statistical Association* **97**, 1034–1041.

[17] Mehta, C.R. & Tsiatis, A. (2001). Flexible sample size considerations using information-based interim monitoring, *Drug Information Journal* **35**, 1095–1112.

[18] Posch, M. & Bauer, P. (2002). Promises and limitations of adaptive designs for clinical research, *IBC*. Freiburg, Germany.

[19] Proschan, M. & Hunsberger, S. (1995). Designed extension of studies based on conditional power, *Biometrics* **51**, 1315–1324.

[20] Proschan, M., Liu, Q. & Hunsberger, S. (2002). Practical mid-course sample size modification in clinical trials, *Controlled Clinical Trials* **24**, 4–15.

[21] Shen, Y. & Fisher, L. (1999). Statistical inference for self-designing clinical trials with one-sided hypothesis, *Biometrics* **55**, 190–197.

[22] Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance, *Annals of Mathematical Statistics* **16**, 243–258.

[23] Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials, *Statistics in Medicine* **9**, 65–72.

[24] Wittes, J.T., Schabenberger, O., Zucker, D.M., Brittain, E. & Proschan M. (1999). Internal pilot studies I: type I error rate of the naive t-test, *Statistics in Medicine* **18**, 3481–3491.

[25] Zucker, D. & Denne, J. (2002). Sample size redetermination for repeated measures studies, *Biometrics* **58**, 548–559.

[26] Zucker, D.M., Wittes, J.T., Schabenberger, O. & Brittain, E. (1999). Internal pilot studies II: comparison of various procedures, *Statistics in Medicine* **18**, 3493–3509.

JANET WITTES

Adaptive Sampling

Animal populations are often highly clustered (*see* **Clustering**). For example, fish can form large, widely scattered schools with few fish in between. Even rare species of animals may form small groups that are hard to find. Applying standard sampling methods such as **simple random sampling** of plots to such a population could yield little information, with most of the plots being empty. Adaptive cluster sampling, the most well-known form of adaptive sampling, is based on the simple idea that when some animals are located on a sample plot, the neighboring plots (and possibly their neighbors as well) are added to the sample. The hope is to find the whole cluster.

Methods of estimation were initially developed in the three pioneering papers of Thompson [27–29] and the sampling book by Thompson [30]. The methodology is described in detail by Seber & Thompson [38], and by Thompson & Seber [51].

Adaptive Methods

With adaptive sampling, the selection of sampling units (or plots) at any stage of the process depends on information from the units already selected. **Sequential** sampling could therefore be regarded as an adaptive method of sampling, but with the sample size rather than the method of selecting the units being adaptive. We note that the **network** (multiplicity) **sampling** introduced by Sirken and colleagues [39] (see [21] for references) is different from adaptive sampling, though they both use the idea of a network.

Adaptive cluster sampling, which we discuss in detail below, is the most common adaptive method. It is a form of biased sampling, technically known as unequal probability sampling, which arises when sampling clusters of different sizes. The probability of selecting a plot will depend on the size of the animal cluster in which the plot is embedded. We find, not surprisingly, that the standard **Horvitz–Thompson** (HT) and Hansen–Hurwitz (HH) estimators for unequal probability sampling (cf. [16] and [18]) can be modified to provide **unbiased** estimators.

Another adaptive method, which has been described as adaptive allocation, can be used when the population is divided up into strata or primary units, each consisting of secondary units. An initial sample of secondary units is taken in each primary unit. If some criterion is satisfied, for example the average number of animals per sampled unit in the primary unit is greater than some prechosen number, then a further sample of units is taken from the *same* primary unit. Kremers [22] developed an unbiased estimator for this situation. If the clumps tend to be big enough so that they are spread over several primary units, we could use what is found in a particular primary unit to determine the level of the sampling in the next unit. This is the basis for the theory developed by Thompson et al. [49]. Other forms of augmenting the initial sample which give biased estimates are described by Francis [14, 15] and Jolly & Hampton [19, 20].

Adaptive Cluster Sampling

As indicated briefly above, adaptive cluster sampling begins with an initial sample and, if individuals are detected on one of the selected units, then the neighboring units of that unit are sampled as well. If further individuals are encountered on a unit in the neighborhood, then the neighborhood of that unit is also added to the sample, and so on, thus building up a cluster of units. If the initial sample includes a unit from a clump, then the rest of the clump will generally be sampled. Such an approach will give us a greater number of individuals.

To set out the steps involved in adaptive cluster sampling, we begin with a finite population of N units (plots) indexed by their "labels" $(1, 2, \ldots, N)$. With unit i is associated a variable of interest y_i for $i = 1, 2, \ldots, N$. Up till now we have referred to y_i as the number of animals on the ith unit. However, as well as counting numbers of individuals, we may wish to measure some other characteristic of the unit, for example plant biomass or pollution level, or even just note the presence or absence of some characteristic using an indicator variable for y_i. In addition to rare species and pollution studies, we can envisage a wide range of populations that would benefit from adaptive sampling, for example populations which form large aggregations such as fish, marine mammals, and shrimp. It has also been used for sampling animal habitats [28], and we can add mineral deposits and rare infectious diseases in human populations (e.g. **AIDS**) to our list. Having defined y_i, our aim is to select a sample, observe the

y values for the units in the sample, and then estimate some function of the population y values such as the population total $\sum_{i=1}^{N} y_i = \tau$ or the population mean $\mu = \tau/N$. Before sampling begins we need to do three things.

The first is to define, for each unit i, a neighborhood consisting of that unit and a set of "neighboring" units. For example we could choose all the adjacent units with a common boundary which, together with unit i, form a "cross". Neighborhoods can be defined to have a variety of patterns; the units in a neighborhood do not have to be contiguous. However, they must have a "symmetry" property, that is if unit j is in the neighborhood of unit i, then unit i is in the neighborhood of unit j. We assume, for the moment, that these neighborhoods do not depend on y_i.

The next step is to specify a condition C (for instance, $y > c$ where c is a specified constant), which determines when we add a neighborhood or not, and the third step is to decide on the size of the initial sample size n_1.

We begin the sampling process by taking an initial **random sample** of n_1 units selected, usually without replacement (*see* **Sampling With and Without Replacement**), from the N units in the population. Whenever the y value of a unit i in the initial sample satisfies C, all units in the neighborhood of unit i are added to the sample. If, in turn, any of the added units satisfies the condition, still more units are added. The process is continued until a cluster of units is obtained which contains a "boundary" of units called *edge* units that do not satisfy C. If a unit selected in the initial sample does not satisfy C, then there is no augmentation and we have a cluster of size one. The process is demonstrated in Figure 1 where the units are plots and the neighborhood forms a cross. Here y_i is the number of animals on plot i and $c = 0$ so that a neighborhood is added every time animals are found. In Figure 1(a) we see one of the initial plots which happens to contain one animal. As it is on the edge of a "clump" we see that the adaptive process leads to the cluster of plots in Figure 1(b).

We note that even if the units in the initial sample are distinct, as in sampling without replacement, repeats can occur in the final sample as clusters may overlap on their edge units or even coincide. For example, if two nonedge units in the same cluster are selected in the initial sample, then that whole cluster occurs twice in the final sample. The final sample

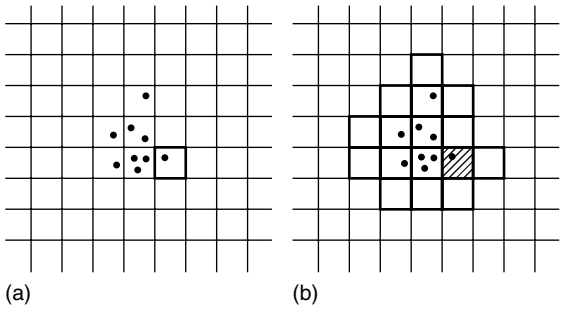

Figure 1 (a) Initial sample plot; (b) cluster obtained by adding adaptively

then consists of n_1 (not necessarily distinct) clusters, one for each unit selected in the initial sample.

Unbiased Estimation

Although the cluster is the natural sample group, it is not a convenient entity to use for theoretical developments because of the double role that edge units can play. If an edge unit is selected in the initial sample, then it forms a cluster of size 1. If it is not selected in the initial sample, then it can still be selected by being a member of any cluster for which it is an edge unit. We therefore introduce the idea of the network A_i for unit i which is defined to be the cluster generated by unit i but with its edge units removed. In Figure 1(b) we get the sampled network by omitting the empty units from the sampled cluster. Here the selection of *any* unit in the network leads to the selection of *all* of the network. If unit i is the only unit in a cluster satisfying C, then A_i consists of just unit i and forms a network of size 1. We also define any unit which does not satisfy C to be a network of size 1 as its selection does not lead to the inclusion of any other units. This means that all clusters of size 1 are also networks of size 1. Thus any cluster consisting of more than one unit can be split into a network and further networks of size 1 (one for each edge unit). In contrast to having clusters which may overlap on their edge units, the distinct networks are *disjoint* and form a *partition* of the N units.

Since the probability of selecting a unit will depend on the size of the network it is in, we are in the situation of unequal probability sampling and the usual estimates based on equal probability sampling will be biased. However, as already mentioned,

we can consider the Horvitz–Thompson (HT) and Hansen–Hurwitz (HH) estimators, the latter being used in sampling with replacement. These estimators, however, require that we know the probability of selection of each unit in the final sample. Unfortunately these probabilities are only known for units in networks selected by the initial sample, and not for the edge units attached to these networks. For example, the probability π_i that an initial sampling unit falls in the network containing unit i is

$$\pi_i = 1 - \binom{N - m_i}{n_1} \Big/ \binom{N}{n_1},$$

where m_i is the number of units in this network.

Therefore, in what follows, we ignore all edge units that are not in the initial sample and use only network information when it comes to computing the final estimators. Motivated by the HT estimator for the population **mean** μ, we consider

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} y_i \frac{I_i}{\mathrm{E}[I_i]},$$

where I_i takes the value 1 if the initial sample intersects network A_i, and 0 otherwise. It is clear that $\hat{\mu}$ is an unbiased estimator for sampling with or without replacement.

Another possible estimator (motivated by the HH estimator) that is also obviously unbiased for sampling, with or without replacement, is

$$\tilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} y_i \frac{f_i}{E[f_i]},$$

where f_i is the number of times that the ith unit in the final sample appears in the estimator, that is the number of units in the initial sample which fall in (intersect) A_i determined by unit i. We note that $f_i = 0$ if no units in the initial sample intersect A_i. It can be shown that

$$\tilde{\mu} = \frac{1}{n_1} \sum_{i=1}^{n_1} w_i = \overline{w},$$

say, where w_i is the mean of the observations in A_i, i.e. $\overline{w}$ is the mean of the n_1 (not necessarily distinct) network means. Di Consiglio & Scanu [12] studied the asymptotic behaviors of $\hat{\mu}$ and $\tilde{\mu}$. They proved that, under suitable conditions, Hajek's theorem [17] for asymptotic normality distribution can be applied to both estimators. However, confidence intervals based on asymptotic approximations may not be appropriate when the sample size is relatively small. Christman & Pontius [9] used several bootstrap percentile methods for constructing confidence intervals under adaptive cluster sampling. They showed, in a simulation study, that the coverage by the bootstrap method was closer to nominal coverage than the normal approximation.

In addition to the above two types of estimator, there is a third type of estimator that can be used. Since a network can be selected more than once, a more efficient design might be to "remove" a network from further consideration once it has been selected, i.e. select networks without replacement. We can than use an estimator due to Murthy [26]; details are given by Salehi & Seber [33]. Salehi & Seber [35] gave a direct proof of Murthy's estimator which extends the use of this estimator to sequential and some adaptive sampling schemes.

Rao–Blackwell Modification

In the above unbiased estimates that we introduced, we did not make use of the y values from the edge units. With this loss of information we would expect to be able to find more efficient estimates using all the sample data. We now show how we can do this.

An adaptive sample can be defined as one for which the probability of obtaining the sample depends only on the distinct unordered y observations in the sample, and not on the y values outside the sample. In this case d, the set of distinct unordered labels in the sample together with their associated y values, is minimal **sufficient** for μ. This is proved for "conventional designs" by Cassel et al. [7] and Chaudhuri & Stenger [8], and their proofs readily extend to the case of adaptive designs. (This extension is implicit in [2] and it is given in [51].) This means that an unbiased estimator that is not a function of d can be "improved" by taking the expectation of the estimator conditional on d to give an estimator with smaller **variance**. For example, consider three unbiased estimators of μ, namely $\overline{y}_1$ (the mean of the initial sample of n_1 units), $\hat{\mu}$, and $\tilde{\mu}$. Each of these depends on the order of selection as they depend on which n_1 units are in the initial sample; $\tilde{\mu}$ also depends on repeat selections; and when the

initial sample is selected with replacement, all three estimators depend on repeat selections. Since all three estimators are not functions of the minimal sufficient statistic d we can apply the **Rao–Blackwell theorem**. If T is any one of the three estimators, then $E[T|d]$ will give a better unbiased estimate, i.e. one with smaller variance. We find that this estimator now uses all the units including the edge units. Salehi [30], using an approach based on the inclusion–exclusion formula, has derived analytical expressions for the Rao–Blackwell version of the modified HH and HT estimators and their variance estimators. Felix-Medina [13], using a different approach, has also derived analytical expressions for their variances.

Applications and Extensions

In applications, other methods are sometimes used for obtaining the initial sample. For instance, in forestry the units are trees, and these are usually selected by a method of unequal probability sampling where the probability of selecting a tree is proportional to the basal area of a tree (the cross-sectional area of a tree at the basal height – usually 4.5 feet in the US). Roesch [29] described a number of estimators for this situation and derivations are given in [51].

In ecology, larger sample units other than single plots are often used. For example, a common sampling unit is the strip transect, which we might call the primary unit. In its adaptive modification, the strip would be divided up into smaller secondary units, and if we find animals in one of its secondary units we would sample units on either side of that unit, with still further searching if additional animals are sighted while on this search. Strips are widely used in both aerial and ship surveys of animals and marine mammals.

Here the aircraft or vessel travels down a line (called a line transect) and the area is surveyed on either side out to a given distance. Thompson [44] showed how the above theory can be applied to this sampling situation. He pointed out that a primary unit need not be a contiguous set of secondary units. For example, in some wildlife surveys the selection of sites chosen for observation is done systematically (with a random starting point) and a single systematic selection then forms the primary unit (*see* **Systematic Sampling Methods**). We can then select several such primary units without replacement and add adaptively as before. Such a selection of secondary units will tend to give better coverage of the population than a simple random sample. Acharya et al. [1] used sstematic adaptive cluster sampling (SACS) to sample three rare tree species in a forest area of about 40 ha in Nepal. They checked its applicability and showed that, for some cases, its efficiency of density estimation relative to conventional systematic sampling, increased by up to 500%.

Clearly other ways of choosing a primary unit to give better coverage are possible. Munholland & Borkowski [24, 25] and Borkowski [3] suggest using a **Latin square** $+1$ design selected from a square grid of secondary units. The Latin square gives a secondary unit in every row and column of the grid, and the extra (i.e. $+1$) unit ensures that any pair of units has a positive probability of being included in the initial sample. The latter requirement is needed for unbiased variance estimation. Salehi [31] suggested using a systematic Latin square sampling $+1$ design selected from a rectangular grid of secondary units.

In some situations it is hard to know what c should be for the condition $y > c$. If we choose c too low or too high we end up with a feast or famine of extra plots. Thompson [48] suggested using the data themselves, in fact the **order statistics**. For example, c could be the rth largest y value in the initial sample statistic so that the neighborhoods are now determined by the y values. This method would be particularly useful in pollution studies where the location of "hot spots" is important. In a study of contaminated sites, the advantages and disadvantages of this sampling scheme, when used along with composite sampling, have been discussed briefly by Correl [11].

Another problem, regularly encountered with animal population studies, is that not all animals are detected. Thompson & Seber [50] developed tools for handling incomplete detectability for a wide variety of designs including adaptive designs thus extending the work of Steinhorst & Samuel [42]. In the presence of incomplete detection, Pollard and Buckland [27] developed an adaptive sampling method in shipboard line transect survey. The survey effort is increased when the number of observation exceeds some limit. This increase is achieved by zigzagging for a period, after which the ship returns to the nominal (straight line) cruise track. They use distance sampling theory (see [6] to find the estimator.

Often we are in a **multivariate** situation where one needs to record several characteristics or measurements on each unit, e.g. the numbers of different species. Thompson [47] pointed out that any function of the variables can be used to define the criterion C, and obtained unbiased estimates of the mean vector and **covariance matrix** for these variables.

We can use any of the above methods in conjunction with **stratification**. If we do not allow the clusters to cross stratum boundaries, then individual stratum estimates are independent and can be combined in the usual fashion. However, Thompson [45] extended this theory to allow for the case where clusters do overlap. Such an approach makes more efficient use of sample information.

Finally we mention the "model-based" or **"superpopulation"** approach (cf. Särndal et al. [37], for example). Here the population vector $\mathbf{y}$ of y values is considered to be a realization of a random vector $\mathbf{Y}$ with some joint distribution F, which may depend on an unknown parameter ϕ. In a **Bayesian** framework ϕ will have a known **prior distribution**. For this model-based approach, Thompson & Seber [51] indicate which of the results for conventional designs carry over to adaptive designs and which of those do not. They also show in their Chapter 10 that **optimal designs** tend to be adaptive.

Relative Efficiency

An important question one might ask about adaptive sampling is "How does it compare with, say, simple random sampling?" This question is discussed by Thompson & Seber [51, Chapter 5] and some guidelines are given. Cost considerations are also important. Simple examples given by them throughout their book suggest that there are large gains in efficiency to be had with clustered populations. Clearly, it will depend on the degree of clustering in the population. Two **simulation** studies that shed some light on this are given by Smith et al. [40] and Brown [4]. These two studies suggested that the HT estimator is more efficient than the HH estimator. Salehi [32] found some support for this analytically, and recommended use of the HT estimator despite the HH estimator being easier to compute.

Adaptive cluster sampling is an efficient method for sampling rare and clustered populations when cluster sizes are large relative to unit sizes. Smith et al. [41] used adaptive cluster sampling for estimating the density of freshwater mussel populations. Since some of the populations were rare and clustered, but with small cluster sizes, adaptively added units were mainly edge units, with little or no gain in efficiency.

Designing an Adaptive Survey

There are several problems associated with adaptive sampling. First, the final sample size is random and therefore unknown. Furthermore, as we saw above, the unit selection probabilities depend on the initial sample size n_1. How then can we use a pilot survey, for example, to design an experiment with a given efficiency or expected cost – an approach which is used for conventional designs such as simple random sampling? Secondly, if an inappropriate criterion C for adding neighborhoods is used, then there may be a "feast or famine" of sampling units. If too many units are being added at each initially selected unit then we end up sampling too many units. Alternatively we might not get enough units. Thirdly, a lot of effort can be expended in locating initial units as we must travel to the site of each such unit.

Recently, a two-stage scheme has been developed by Salehi & Seber [34] which helps us to deal with all three problems in a reasonably optimal manner. To use this scheme, we divide the population of (secondary) units into, say, M primary sampling units (PSUs), each containing $N_0 = N/M$ secondary units. A simple random sample of m primary units is then taken and adaptive cluster sampling is carried out in each of the selected primary units using an initial sample of n_0 units. We again have two schemes, depending on whether networks are allowed to cross PSU boundaries or not, and two estimators (HT and HH) for each scheme. To design such an experiment, we use the HT estimator with nonoverlapping boundaries and choose a pilot sample of m_p PSUs but with the *same* initial sample size of n_0 units in each of the selected primary units. The theory based on the pilot survey now works, that is, we can now determine m to achieve a given efficiency or cost. The reason for this is that the network selection probabilities in a PSU are the same for both the pilot survey and the survey planned; both depend on n_0.

Another method of controlling the overall sample size is to use a method called restricted adaptive

cluster sampling, proposed by Brown & Manly [5]. Here the units are selected sequentially for the initial sample until a desired sample size is reached. The sampling therefore "restricts" the initial sample to one that produces a final sample size that is either at or just over the defined limit. The HT and HH estimators are now biased but under some circumstances the bias can be estimated well by **bootstrapping**. Lo et al. [23] used the restricted method to estimate Pacific hake larval abundance.

Salehi & Seber refno36 provided an unbiased estimator for the restricted method. Using a simulation study, they showed that the unbiased estimator has a smaller **mean square error** than the biased estimators. They also considered a restricted method when the networks are selected without replacement and obtained its unbiased estimator. Christman and Lan [10] introduced inverse adaptive cluster sampling, which is a especial case of restricted adaptive cluster sampling.

References

[1] Acharya, B., Bhattarai, A., De Gier, A., & Stein, A. (2000). Systematic adaptive cluster sampling for the assessment of rare tree species in Nepal, *Forest Ecology and Management*, **137**, 65–73.

[2] Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory, *Sankhyā, Series A* **31**, 441–454.

[3] Borkowski, J.J. (1999). Network inclusion probabilities and Horvitz-Thompson estimation for adaptive simple Latin square sampling, *Environmental and Ecological Statistics*, **6**, 291–311.

[4] Brown, J.A. (1994). The application of adaptive cluster sampling to ecological studies, in *Statistics in Ecology and Environmental Monitoring*, D.J. Fletcher & B.F.J. Manly, eds. University of Otago Press, Dunedin, New Zealand, pp. 86–97.

[5] Brown, J.A. & Manly, B.F.J. (1998). Restricted adaptive cluster sampling, *Journal of Environmental and Ecological Statistics*, **5**, 49–63.

[6] Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laak, J.L. (1993). *Distance Sampling: Estimating Abundance of Biological Populations*. Chapman and Hall, New York and London.

[7] Cassel, C.M., Särndal, C.E. & Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York.

[8] Chaudhuri, A. & Stenger, H. (1992). *Survey Sampling: Theory and Methods*. Marcel Dekker, New York.

[9] Christman, M.C. & Pontius, J.S. (2000). Bootstrap confidence intervals for adaptive cluster sampling, *Biometrics*, **56**, 503–510.

[10] Christman, M.C. & Lan F. (2001). Inverse adaptive cluster sampling, *Biometrics*, **57**, 1096–1105.

[11] Correll, R.L. (2001). The use of composite sampling in contaminated sites- a case study, *Environmental and Ecological Statistics*, **8**, 185–200.

[12] Di Consiglio, L. & Scanu, M. (2001). Some results on asymptotic in adaptive cluster sampling, *Statistics & Probability Letters*, **52**, 189–197.

[13] Félix-Medina, M.H. (2000). Analytical expressions for Rao-Blackwell estimators in adaptive cluster sampling, *Journal of Statistical Planing and Inference*, **84**, 221–236.

[14] Francis, R.I.C.C. (1984). An adaptive strategy for stratified random trawl surveys, *New Zealand Journal of Marine and Freshwater Research* **18**, 59–71.

[15] Francis, R.I.C.C. (1991). Statistical properties of two-phase surveys: comment, *Canadian Journal of Fisheries and Aquatic Science*, **48**, 1128.

[16] Hansen, M.M. & Hurwitz, W.N. (1943). On the theory of sampling from finite populations, *Annals of Mathematical Statistics* **14**, 333–362.

[17] Hájek, J. (1960). *Limiting Distributions in Simple Random Sampling from a Finite Population*, Publication of the Mathematical Institute of the Hungarian Academy of Sciences, pp. 361–374.

[18] Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

[19] Jolly, G.M. & Hampton, I. (1990). A stratified random transect design for acoustic surveys of fish stocks, *Canadian Journal of Fisheries and Aquatic Science*, **47**, 1282–1291.

[20] Jolly, G.M. & Hampton, I. (1991). Reply to comment by R.I.C.C. Francis, *Canadian Journal of Fisheries and Aquatic Science* **48**, 1228–1129.

[21] Kalton, G. & Anderson, D.W. (1986). Sampling rare populations, *Journal of the Royal Statistical Association, Series A* **147**, 65–82.

[22] Kremers, W.K. (1987). Adaptive Sampling to Account for Unknown Variability Among Strata, *Preprint No. 128*. Institut für Mathematik, Universität Augsburg, Germany.

[23] Lo, C.H., Griffith, D. and J.R. Hunter (1997). Using a restricted adaptive sampling to estimate pacific hake larval abundance. *CalCOFI*, Rep. **38**, 103 –113.

[24] Munholland, P.L. & Borkowski, J.J. (1993). Adaptive Latin Square Sampling +1 Designs, *Technical Report No. 3-23-93*. Department of Mathematical Sciences, Montana State University, Bozeman.

[25] Munholland, P.L. & Borkowski, J.J. (1996). Latin square sampling +1 designs, *Biometrics* **52**, 125–136.

[26] Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement, *Sankhyā* **18**, 379–390.

[27] Pollard, J.H. & S.T. Buckland (1997). A strategy for adaptive sampling in shipboard line transect surveys, *Rep. International Whaling Commission*, **47**, 921–931.

[28] Ramsey, F.L. & Sjamsoe'oed, R. (1994). Habitat association studies in conjunction with adaptive cluster samples, *Journal of Environmental and Ecological Statistics* **1**, 121–132.

[29] Roesch, F.A., Jr (1993). Adaptive cluster sampling for forest inventories, *Forest Science* **39**, 655–669.

[30] Salehi M.M. (1999). Rao-Blackwell versions of the Horvitz-Thompson and Hansen-Hurwitz in adaptive cluster sampling. *Journal of Environmental and Ecological Statistics*. **6**, 183–195.

[31] Salehi M.M., (2002). Systematic simple Latin Square sampling (+1) design and its optimality. *Journal of propagations on probability and statistics*, **2**, 191–200.

[32] Salehi M.M. (2002). Comparison between Hansen-Hurwitz and Horvitz-Thompson for adaptive cluster sampling, *Journal of Environmental and Ecological Statistics*. In press.

[33] Salehi, M.M. & Seber, G.A.F. (1997). Adaptive cluster sampling with networks selected without replacement, *Biometrika*, **84**, 209–219.

[34] Salehi, M.M. & Seber, G.A.F. (1997). Two stage adaptive cluster sampling, *Biometrics*, **53**, 959–970.

[35] Salehi M.M. & Seber, G.A.F. (2001). A new proof of Murthy's estimator which applies to sequential sampling. *Australian & New Zealand Journal of Statistics*, **43**, 901–906.

[36] Salehi M.M. & Seber, G.A.F. (2002). Unbiased estimators for restricted adaptive cluster sampling. *Australian & New Zealand Journal of Statistics*, **44**, 63–74.

[37] Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[38] Seber, G.A.F. & Thompson, S.K. (1994). Environmental adaptive sampling, in *Handbook of Statistics*, Vol. 12: *Environmental Sampling*, G.P. Patil & C.R. Rao, eds. North-Holland/Elsevier Science, New York, pp. 201–220.

[39] Sirken, M.G. (1970). Household surveys with multiplicity, *Journal of the American Statistical Association* **63**, 257–266.

[40] Smith, D.R., Conroy, M.J. & Brakhage, D.H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl, *Biometrics*, **51**, 777–788.

[41] Smith, D.R., Villella, R.F. & Lemari, D.P. (2002). Application of adaptive cluster sampling to low-density populations of freshwater mussels. *Environmental and Ecological Statistics*, in press.

[42] Steinhorst, R.K. & Samuel, M.D. (1989). Sightability adjustment methods for aerial surveys of wildlife populations, *Biometrics* **45**, 415–425.

[43] Thompson, S.K. (1990). Adaptive cluster sampling, *Journal of the American Statistical Association* **85**, 1050–1059.

[44] Thompson, S.K. (1991). Adaptive cluster sampling: Designs with primary and secondary units, *Biometrics* **47**, 1103–1115.

[45] Thompson, S.K. (1991). Stratified adaptive cluster sampling, *Biometrika* **78**, 389–397.

[46] Thompson, S.K. (1992). *Sampling*. Wiley, New York.

[47] Thompson, S.K. (1993). Multivariate aspects of adaptive cluster sampling, in *Multivariate Environmental Statistics*, G.P. Patil & C.R. Rao, eds. North-Holland/Elsevier Science, New York, pp. 561–572.

[48] Thompson, S.K. (1996). Adaptive cluster sampling based on order statistics, *Environmetrics* **7**, 123–133.

[49] Thompson, S.K., Ramsey, F.L. & Seber, G.A.F. (1992). An adaptive procedure for sampling animal populations, *Biometrics* **48**, 1195–1199.

[50] Thompson, S.K. & Seber, G.A.F. (1994). Detectability in conventional and adaptive sampling, *Biometrics* **50**, 712–724.

[51] Thompson, S.K. & Seber, G.A.F. (1996). *Adaptive Sampling*. Wiley, New York.

GEORGE A.F. SEBER & SALEHI M. MOHAMMAD

Additive Hazard Models

While most modern analyses of survival data focus on **multiplicative models** for **relative risk** using **proportional hazards** models, some work has been done on the development of additive hazard models. **Aalen's additive regression model** [1, 2, 6, 8] is a general nonparametric additive hazard model of the form:

$$\sum \lambda_i(t)x_i, \tag{1}$$

where the λ_i are nonparametric hazard functions associated with **covariate** x_i, which may be **time-dependent**. Andersen & Væth [3] discuss a special case of this model of the form

$$\beta(t)\mu^*(t) + \gamma(t), \tag{2}$$

where $\gamma(t)$ and $\beta(t)$ are functions to be estimated and $\mu^*(t)$ is a known function describing rates in some reference population. McKeague & Sasieni [8] discuss a partly parametric version of (1). For inference and other issues related to models (1) and (2) see **Aalen's additive regression model**. In what follows we discuss a class of parametric additive hazard models that are of special interest in studies in which we want to describe data on survival in terms of how the excess risk (or rate) depends on one or more "exposures", and how these exposure-specific risks depend on other factors, such as sex, age at exposure, or time since exposure. These models are used extensively in studies of **radiation** effects on cancer [10, 11] and are applicable in a wide variety of **occupational** and other studies.

In many applications it is useful to consider additive hazard models of the form:

$$\lambda_0(t, \beta_0, z_0) + \lambda_1(t, \beta_1, z_1), \tag{3}$$

$$\lambda_0(t, \beta_0, z_0)[1 + \lambda_1(t, \beta_1, z_1)]. \tag{4}$$

In these models, $\lambda_0(\cdot)$ represents a background **hazard (rate)** function that depends on time (t), and other covariates, z_0, with parameters β_0, while $\lambda_1(\cdot)$ describes the excess hazard (excess absolute rate) (3) or excess relative risk (4) as a function of time and covariates, z_1, with parameters β_1. In general, covariates affecting the excess risk will include some measure of exposure and may be time-dependent. It is also common for some covariates, e.g. sex, to affect both the background and excess risks. That is, some covariates may appear in both z_0 and z_1.

When working with parametric additive hazard models one must specify functional forms for the background and excess risks. In many problems **log-linear models** provide an adequate description of the background rates. Commonly used models for the logarithm of the background rate are linear or polynomial functions of t or $\log(t)$, though linear or quadratic **splines** in t or $\log(t)$ can also be useful. Other covariates, such as sex or **birth cohort**, may affect the intercepts or slopes in such models.

For an exposure, d, it is often useful to consider **excess risk** models of the form

$$\lambda_1(\cdot) = \rho(d, \beta_d)\gamma(t, z, \beta_1),$$

where we assume that other factors act multiplicatively on the shape of the **dose–response** function $\rho(\cdot)$. Dose–response functions may be described using linear, quadratic, linear spline, categorical, or other functions of dose, while **effect modification** is often modeled using loglinear functions of other covariates.

For example, in an analysis of mortality from cancers other than leukemia in Japanese atomic bomb survivors over a 40-year period it was found that the effect of radiation could be described quite well using an additive **excess relative risk** model in which the linear dose effect depends on sex and decreases loglinearly with increasing age at exposure ($agex$) with no significant effects of age. One way to write this model for the excess relative risk is

$$\lambda_1(\cdot) = \beta_1 dose \times \exp(\beta_2\, female + \beta_3\, agex),$$

where *female* is an indicator variable that is 1 for women and 0 for men. An alternative model describing excess absolute cancer death rates for atomic bomb survivors cancer data in terms of age at death (age) is

$$\begin{aligned}\lambda_1(\cdot) &= \beta_1\, dose \times \exp[\beta_2 \ln(age)] \\ &= \beta_1\, dose \times (age)^{\beta_2}.\end{aligned}$$

These two models were found to describe the excess cancer risks equally well.

Generalizing the Models

There are a number of useful generalizations and extensions of models (3) and (4). An important extension, which is closely related to model (1), generalizes the simple standardized mortality ratio (SMR) (*see* **Standardization Methods**) used in many epidemiologic studies. In particular, external data on background rates can be incorporated into either model by inclusion of these rates as a covariate in the background term, which may also include additional parameters to describe the ratio of the external background rates and the rates in the study population. We can write this background rate model as

$$\lambda_0(t, z_0, \beta_0) = \mu^*(t, z_0)\gamma(z_0, \beta_0).$$

Breslow et al. [5] discuss this multiplicative SMR model in some detail; extension to additive hazard models is straightforward.

Models (3) and (4) can also be extended to investigate the joint effects (*see* **Synergy of Exposure Effects**) of multiple exposures by the inclusion of additional excess hazard terms (for independent additive effects) or by allowing **interactions** between different exposure effects in a single excess hazard term.

Several authors (e.g. [4] and [9]) have proposed hybrid parametric families with a continuous index parameter γ such that models like (3) and (4) correspond to specific values of γ. The method of Aranda-Ordaz [4] involves the use of a hybrid family that includes complementary log–log and negative complementary log models to analyze survival data grouped into equal length intervals. This family includes models in which the excess hazard or the log relative risk are modeled as linear functions of the covariates. Muirhead & Darby [9] proposed a hybrid model of the form

$$\{\lambda_0(t, \beta_0, z_0)^\gamma + [1 + \lambda_1(t, \beta_1, z_1)]^\gamma - 1\}^{1/\gamma}.$$

When γ equals 1 this corresponds to the excess hazard model (4) and, in the limit as γ approaches 0, it corresponds to the excess relative model (3).

The primary use of these hybrid models is to compare how well the data of interest are described by models in which either the relative or absolute excess risk is constant over time. It is generally easier and more informative to address this question through a simple comparison of the fits of time-constant excess hazard and excess relative risk models, together with analyses of the effect of allowing the excess hazards or relative risks to depend on time. When excess risks are allowed to depend on time there is usually little difference in the fit of models (3) and (4). In this case, the hybrid models contain little information about the index parameter (i.e. the **profile likelihood** function for γ is quite flat).

Parameter Estimation and Inference

It is generally impractical to develop likelihood-based **estimating function** equations for the use of fully specified parametric models based on (3) and (4) with ungrouped survival data. However, parameter estimation for both of these classes of models is relatively straightforward when done using **Poisson regression** methods for grouped survival data. With suitable Poisson regression **software** it is possible to fit a version of model (4) in which the background hazard is modeled using separate multiplicative parameters for each time period, possibly, with stratification on other factors. This model, which is closely related to the stratified semiparametric proportional hazards model, can be written

$$\lambda_0(t, z, \beta) = \eta_{st}\lambda(z, \beta), \tag{5}$$

where the η_{st} is a parameter describing the hazard for the tth time period in a stratum, s, defined by other factors.

Since model (4) is a proportional hazards model, parameter estimation can also, in principle, be carried out using **partial likelihood** or **counting process methods**; in which case, the background rate would be replaced by, or include, a nonparametric baseline hazard function. Lin & Ying [7] outline a method for fitting simple semiparametric additive excess rate models (3) (*see* **Semiparametric Regression**) to ungrouped data using counting process methods.

Unfortunately, since the additive models of interest are almost always not simple linear or loglinear functions of the covariates, standard Poisson regression and proportional hazards modeling software is of little use in fitting these models. However, the Epicure software package, which was designed for working with general parametric and semiparametric additive risk models, can compute **maximum likelihood** estimates for a broad range of additive hazard models.

Since parameter estimation for additive hazard models is generally carried out using **likelihood** or partial likelihood methods, inference about parameters of interest can be carried out using the standard asymptotic methods, including Wald, score, and **likelihood ratio tests**. However, because of the nonlinear nature of the models and, in many applications, the limited information on excess risks, asymptotic standard errors and, hence, hypothesis tests and confidence intervals based on Wald tests can be quite misleading. Likelihood ratio tests and profile likelihood-based confidence intervals are the preferred methods of inference when working with additive hazard models.

Hazard functions are, by definition, nonnegative. This constraint is addressed implicitly by multiplicative hazard models. However, for additive hazard models it is possible that one of the components of the hazard (usually the "excess") can be negative. The implicit constraint in model (3) is that $\lambda_1(\cdot) > -\lambda_0(\cdot)$, while that for model (4) is $\lambda_1(\cdot) > -1$. These implicit constraints can make it difficult to fit additive hazard models for some data sets. In simple excess risk models, these constraints can be addressed by the choice of the parameterization (e.g. modeling the log of the linear dose–response slope) or by restricting the range of some parameters (e.g. restricting a linear dose–response slope in a simple linear excess relative risk model to be greater than minus one over the maximum dose). However, these approaches are inadequate for every problem.

Summary

Parametric additive hazard functions such as those described in (3) and (4) are useful and, in some settings, natural, alternatives to the semiparametric multiplicative hazards that have come to dominate the analysis of survival data in recent years. These models are especially useful for dose–response analyses in which one is primarily interested in the characterization of excess risks and how the excess depends on other factors. Parameter estimation and inference for additive hazard models is most easily carried out using nonlinear models and Poisson regression methods for **grouped survival data**.

References

[1] Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes, *Springer Lecture Notes in Statistics*, Vol. 2. Springer-Verlag, New York, pp. 1–25.

[2] Aalen, O.O. (1989). A linear regression model for the analysis of lifetimes, *Statistics in Medicine* **8**, 907–925.

[3] Andersen, P.K. & Væth, M. (1989). Simple parametric and non-parametric models for excess and relative mortality, *Biometrics* **45**, 523–535.

[4] Aranda-Ordaz, F.J. (1983). An extension of the proportional-hazards model for grouped data, *Biometrics* **39**, 109–117.

[5] Breslow, N.E. Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[6] Huffer, F.W. & McKeague, I.W. (1991). Weighted least square estimation for Aalen's additive risk model, *Journal of the American Statistical Association* **86**, 114–129.

[7] Lin, D.Y. & Ying, Z. (1994). Semi-parametric analysis of the additive risk model, *Biometrika* **81**, 61–72.

[8] McKeague, I.W. & Sasieni, P.D. (1994). A partly parametric additive risk model, *Biometrika* **81**, 501–514.

[9] Muirhead, D.R. & Darby, S.C. (1987). Modeling the relative and absolute risks of radiation-induced cancers (with discussion), *Journal of the Royal Statistical Society, Series A* **150**, Part 2, 83–118.

[10] Pierce, D.A., Shimizu, Y., Preston, D.L., Væth, M. & Mabuchi, K. (1996). Studies of the mortality of atomic bomb survivors. Report 12, Part I. Cancer Mortality 1950–1990, *Radiation Research* **146**, 1–27.

[11] Preston, D.L. (1990). Modeling radiation effects on disease incidence, *Radiation Research* **124**, 343–344.

DALE L. PRESTON

Additive Model

It is common, though potentially confusing, in discussions of risks and rates to make a distinction between additive and **multiplicative risk models** (e.g. [1, Chapter 4]). Under the additive or **excess risk** (rate) model the risk is described as

$$R = R_0 + E(z), \tag{1}$$

where R_0 is the background risk and $E(z)$ is an excess risk function associated with "exposure", z. Under the multiplicative or **relative risk model**, "exposure" is assumed to have a multiplicative effect on the rates:

$$R = R_0 \times RR(z), \tag{2}$$

where $RR(z)$ is the **relative risk** function.

The confusion in referring to (1) and (2) as additive and multiplicative models arises because the functions used to describe the excess risk in (1) or the relative risk in (2) can include both additive and multiplicative components. In particular, the simple **excess relative risk** model $RR(z) = 1 + \beta z$ is often called an additive model. To make a clear distinction between the form of the risk function and the nature of the functions used to model the components of risk, it is best to describe (1) and (2) as excess risk and relative risk models, respectively. If this is done, then the term additive model can be used to refer to excess or relative risk models that involve additive components. With this definition of additive models, excess risk models are intrinsically additive because they always include the sum of background and excess risks, while relative risk models may be either multiplicative, e.g. $RR(z) = \exp(\beta z)$, or additive, e.g. $RR(z) = 1 + \beta z$. Thomas [4] and Breslow & Storer [2] describe general relative risk functions that include both additive and multiplicative models. Realistic excess risk models often involve sums of multiplicative models for the background and excess risk functions. For example, in a **dose–response** analysis it might be appropriate to allow the excess risk associated with a given dose (d) to depend on sex (s) or time since exposure (t) by considering a multiplicative model for the excess risk of the form

$$E(d, s, t) = \beta_{1s} d \times t^{\theta}.$$

Preston et al. [3] describe a general class of additive models that are useful in working with either excess or relative risks.

The articles on **Parametric Models in Survival Analysis** and **Poisson Regression in Epidemiology** describe some specific additive models and discuss methods for parameter estimation and inference with such models.

References

[1] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research, Vol. II. The Design and Analysis of Cohort Studies*, IARC Scientific Publication No. 82. Oxford University Press, New York.

[2] Breslow, N.E. & Storer, B.E. (1985). General relative risk functions for case-control studies, *American Journal of Epidemiology* **122**, 149–162.

[3] Preston, D.L., Lubin, J., Pierce, D.A. & McConney, M.E. (1993). *Epicure User's Guide*. Hirosoft International Corp., Seattle.

[4] Thomas, D.C. (1981). General relative risk functions for survival time and matched case-control studies, *Biometrics* **37**, 673–686.

DALE L. PRESTON

Additive–Multiplicative Intensity Models

Introduction

The **proportional hazards** model introduced by Cox (*see* **Cox Regression Model**) has been so dominant in **survival analysis** that other models have had a hard time getting the attention they deserve. The Cox model is extremely useful and has many desirable properties, but obviously also has some shortcomings. One important shortcoming is that the model has a hard time describing **time-varying** effects (non-proportional effects), and although some work has been aimed at extending the model to overcome this problem [7, 9, 15], there is still some way to a fully satisfactory apparatus to deal with time-varying effects. Also, some **covariates** will have effects that are not well described as being multiplicative, and some of the available tests may reveal this [13].

One important alternative to the proportional hazards model is the **additive hazard model** and, in particular, **Aalen's additive hazard regression model**. Aalen's additive hazard regression model is completely **nonparametric** and includes covariates additively in the model thus leading to an **excess risk** interpretation. The Aalen model is very flexible and will estimate all covariate effects as time-varying effects. A **semiparametric** version of the model was suggested by McKeague and Sasieni [8] and in a special case by Lin and Ying [4] (*see* **Aalen's Additive Regression Model**). One advantage of the additive models is that time-varying effects are easy to estimate (with explicit estimators) and that inferential procedures for making conclusions about the time-varying effects exist. The semiparametric version of the model has not received much attention in practical work but has the advantage that explicit estimators are given and that the flexibility of the Aalen model can be used only for those covariate effects where it is needed, whereas other covariate effects can be described by parameters.

The additive and multiplicative models postulate different relationships between the hazard and covariates, and it is seldom clear which of the models should be preferred. The models may often be used to complement each other and to provide different **summary measures**. Sometimes, however, covariate effects are best modeled as multiplicative and other covariate effects are best modeled as being additive, and then one must combine the additive and multiplicative models.

Sometimes the data will only give little guidance on whether a covariate effect should be described as multiplicative or additive, but then the choice of additive or multiplicative effects will not be critical for the interpretation of the data. We illustrate in the example how certain tests can be used to decide if a covariate has an additive or multiplicative effect.

The additive and proportional hazard models may be combined in various ways to achieve flexible and useful models. We shall here consider two models that are based on either adding or multiplying the multiplicative Cox model and the additive Aalen model. This leads to two quite different models that are both quite flexible and useful. When the models are added, it leads to the proportional excess hazard models. Several parametric versions of such models exist (*see* **Additive Hazard Models**). For the version of the proportional excess hazard model considered here, the additive part can be thought of as modeling the baseline mortality, whereas the multiplicative part describes the excess risk due to different exposure levels. When the models are multiplied, it leads to a flexible model termed the Cox–Aalen model below. For this model, some covariate effects are believed to result in multiplicative effects, whereas other effects are better described as additive. In the article on **additive hazard models**, an example from cancer mortality is used to illustrate structures similar to those in the Cox–Aalen model.

Some notation is needed. We here use the **counting process** formulation. Assume that i.i.d. subjects are observed over some observation period $[0, \tau]$ and give rise to counting process data $N_i(s)$ with at risk indicator $Y_i(s)$, excess risk indicator $\rho_i(t)$, and $(p+q)$-dimensional covariates $(X_i^{\mathrm{T}}(s), Z_i^{\mathrm{T}}(s))$. Let $\boldsymbol{N}(t) = (N_1(t), \ldots, N_n(t))^{\mathrm{T}}$ be an n-dimensional counting process and define matrices $\boldsymbol{X}(t) = (Y_1(t)X_1(t), \ldots, Y_1(t)X_n(t))^{\mathrm{T}}$, $\boldsymbol{X}_\rho(t) = (\rho_1(t)X_1(t), \ldots, \rho_1(t)X_n(t))^{\mathrm{T}}$, and $\boldsymbol{Z}(t) = (Y_1(t)Z_1(t), \ldots, Y_1(t)Z_n(t))^{\mathrm{T}}$. Finally, let $\mathrm{diag}(w_i)$ denote an $n \times n$ diagonal matrix with elements $w_1, \ldots, w_n$, $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_n(t))^{\mathrm{T}}$, where $\phi_i(t) = \rho_i(t) \exp(Z_i^{\mathrm{T}}(t)\beta)$ and define $\tilde{\boldsymbol{X}}(\beta, t) = (\boldsymbol{X}_\rho(t), \boldsymbol{\phi}(t))$, an $(p+1) \times n$ matrix (*see* **Matrix Algebra**).

Proportional Excess Hazard Models

Lin and Ying [5] considered the following additive–multiplicative intensity model

$$\lambda(t) = Y(t)\left[g(Z^{\mathrm{T}}(t)\beta) + \lambda_0(t)h(X^{\mathrm{T}}(t)\gamma)\right], \quad (1)$$

where $Y(t)$ is an at risk indicator, $(X(t), Z(t))$ is a $p+q$ dimensional covariate vector, $(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})$ is a $p+q$ dimensional vector of regression coefficients and $\lambda_0(t)$ is an unspecified baseline hazard. Both h and g are assumed known. One problem with this model is that only the baseline is time-varying and therefore data with time-varying effects will often not be well described by the model. When additional time-varying effects are included in the model, the model will get added flexibility and it turns out that it is relatively simple to extend the model to deal with time-varying effects such as in the flexible additive–multiplicate intensity model [6], where the intensity is of the form

$$\lambda(t) = Y(t)\left[X^{\mathrm{T}}(t)\alpha(t)) + \rho(t)(\lambda_0(t)\exp\{Z^{\mathrm{T}}(t)\beta\}\right], \quad (2)$$

where both $Y(t)$ and $\rho(t)$ are at risk (excess risk) indicators, $\alpha(\cdot)$ is a q-vector of time-varying regression functions, $\lambda_0(t)$ is the baseline hazard of the excess risk term, and β is a p-dimensional vector of **relative risk** regression coefficients. The at risk indicators $Y(t)$ and $\rho(t)$ may be equivalent as in the Lin and Ying model, but sometimes one will have a baseline group where there is no excess risk. It should be verified that the model is **identifiable**. The model is an extension of the Lin and Ying model when $g(x) = x$ and $h(x) = \exp(x)$ and the model is the sum of an additive Aalen model and a Cox model. Sasieni [10] considered the special case of this model where $\alpha^{\mathrm{T}}(t)X_i(t)$ is replaced by a known function of $X_i(t)$. We shall consider estimation of the unknown parameters β, $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ and $\Lambda(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$; see [6] for additional theoretical details. Essentially, the model reduces to Aalen's additive risk model for known β and this may be utilized to obtain a score equation (*see* **Likelihood**) for β that only depends on observed quantities. Zahl [14] illustrated the use of the model with examples from breast and colon cancer.

Using the matrix notation introduced above, we can write an unweighted version of the score equation for β as

$$U(\beta) = \int_0^\tau \boldsymbol{Z}^{\mathrm{T}}(t)\mathrm{diag}(\phi_i(t))[I - \tilde{\boldsymbol{X}}(\beta, t) \times \{\tilde{\boldsymbol{X}}^{\mathrm{T}}(\beta, t)\tilde{\boldsymbol{X}}(\beta, t)\}^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}(\beta, t)]\,\mathrm{d}\boldsymbol{N}(t) = 0, \quad (3)$$

where $\phi_i(t) = \rho_i(t)\exp(Z_i(t)^{\mathrm{T}}\beta)$.

Now, denoting the solution to the score equation as $\hat{\beta}$, we estimate $B = (A(t), \Lambda(t))$ by

$$\widehat{B}(t) = \int_0^t \left\{\tilde{\boldsymbol{X}}^{\mathrm{T}}(\hat{\beta}, t)\tilde{\boldsymbol{X}}(\hat{\beta}, t)\right\}^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}(\hat{\beta}, t)\,\mathrm{d}\boldsymbol{N}(t). \quad (4)$$

An alternative estimation strategy is to iterate between fitting the model with β or $A(t)$ known [14].

The model extends both the Cox and the Aalen model and may have potential use to investigate **goodness of fit** for these models.

The Multiplicative Cox–Aalen Model

A different way of combining additive and multiplicative models are given by the Cox–Aalen model [11, 12]

$$\lambda_i(t) = Y_i(t)\left[X_i^{\mathrm{T}}(t)\alpha(t)\right]\exp(Z_i^{\mathrm{T}}(t)\beta). \quad (5)$$

The Cox–Aalen allows a very flexible (additive) description of covariate effects of $X_i(t)$ while allowing other covariate effects to act multiplicatively on the hazard. One alternative way of thinking about the model is to consider it as an approximation to the general stratified hazard model $\lambda(t, X_i(t))\exp(Z_i^{\mathrm{T}}(t)\beta)$ suggested by Dabrowska [1]. Compared to the Dabrowska model, some structure is introduced to make the estimation easier and to help facilitate the interpretation of the covariate effects.

To estimate the parameters of the model, an approximate **maximum likelihood** score equation is suggested and the estimators are studied in [11, 12]. The key to solving the model is to notice that for known β we have a usual Aalen model where the nonparametric terms can be estimated.

To estimate the regression parameter β, we solve the score equation

$$U(\beta) = \int_0^\tau \boldsymbol{Z}^{\mathrm{T}}(t)[I - \boldsymbol{W}(t, \beta)\boldsymbol{X}(t) \times \{\boldsymbol{X}^{\mathrm{T}}(t)\boldsymbol{W}(t, \beta)\boldsymbol{X}(t)\}^{-1}\boldsymbol{X}^{\mathrm{T}}(t)]\,\mathrm{d}\boldsymbol{N}(t) = 0, \quad (6)$$

where $\boldsymbol{W}(t, \beta) = \text{diag}(\exp(Z_i(t)^{\mathrm{T}}\beta))$. This score equation reduces to Cox's partial score equation when $\boldsymbol{X}(t) \equiv (Y_1(t), \ldots, Y_n(t))^{\mathrm{T}}$. An estimator of the cumulative baseline $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ is

$$\hat{A}(t) = \int_0^t \left\{ \boldsymbol{X}^{\mathrm{T}}(t)\boldsymbol{W}(t, \hat{\beta})\boldsymbol{X}(t) \right\}^{-1} \boldsymbol{X}^{\mathrm{T}}(t)\,\mathrm{d}\boldsymbol{N}(t). \tag{7}$$

Note the strong resemblance with the Aalen estimator. These estimators may be improved by the use of weight matrices.

Effect-modification

For models where effects are modeled solely as either multiplicative or additive, effect-modification is just another word for **interaction** (*see* **Effect Modification**) on the chosen scale. For the additive–multiplicative models considered in the sections "Proportional Excess Hazard Models" and "The Multiplicative Cox–Aalen Model", however, this is no longer valid.

Generally, multiplicative effects will lead to some interaction on the hazard, and additive effects will lead to interaction on the log-hazard. For the proportional excess model, both the multiplicative and additive effects will lead to interaction on the overall log-hazard. The model is constructed to give multiplicative effects on the excess risk. The multiplicative effects of the Cox–Aalen model are linear on the log-hazard and lead to interaction on the hazard, and vice-versa for the additive effects of the model.

Assume, for example, that gender has a multiplicative effect on some hazard, with males having relative risk θ compared to that of females, and that some treatment has an additive (or multiplicative) effect, with treated having intensity $\lambda_1(t)$ and untreated having intensity $\lambda_0(t)$. Then the treated females will have excess risk $e(t) = \lambda_1(t) - \lambda_0(t)$, whereas the excess risk for males will be modified by the effect of gender to $\theta e(t)$. Note that these properties refer specifically to the chosen scale at which one considers covariate effects.

The Cox–Aalen model has the useful property that it allows effect-modification of the covariates included in the proportional part of the model; some examples of the use of such a model is given under **additive hazard models**. Note, that the effect-modification, however, must be the same for all effects in the additive part of the model (just as in the Cox model). The model may be extended to allow different effect-modification for the different effects in the additive model.

Example

To illustrate the use of the Cox–Aalen model, we consider the data that was also used to illustrate the use of the Aalen model (*see* **Aalen's Additive Regression Model**). The data are given in [2] and gives the survival on 418 patients with primary biliary cirrhosis. The source of our data set is the survival package of S-Plus/R (*see* **S-PLUS and S**; **R**). The following covariates were used for the modeling: age, log(albumin), bilirubin (dichotomized as 0 when bilirubin is less than 3.25 mg/dL and 1 otherwise), edema (present/not present), and log(prothrombin). To resemble the analysis for the additive risk model as closely as possible, we only consider the data for the first 3000 days.

First, considering the Cox model to fit the data, a modified version of the cumulative score test [3] (see [12]) showed that log(prothrombin) had a nonproportional effect with a ***P* value** at 0.001. We therefore included log(prothrombin) in the additive part of the model. With log(prothrombin) in the additive part of the model, a similar score test revealed that edema had nonproportional effects ($p = 0.04$). So even though the test statistic is not dramatic, it seems preferable with a more flexible description of the effect of edema.

We therefore consider the Cox–Aalen model with baseline, edema and log(prothrombin) as additive components and age, log(albumin), bilirubin as multiplicative effects. The score test for proportional effects gave P values 0.12 for bilirubin, 0.91 for age and 0.81 for log(albumin). The covariates age and log(albumin) were centered to give a meaningful interpretation of the additive part of the model. The log-relative risk estimates (standard error) were 1.46 (0.19) for bilirubin, 0.033 (0.0073) for age, and −2.58 (0.57) for log(albumin).

Figure 1 gives the cumulative effect of edema, whose shape resembles that of the estimate from Aalen's additive risk model. Denote the cumulative effect of edema as $A_e(t)$. Now, for a subject with mean age and mean log(albumin) and bilirubin (less

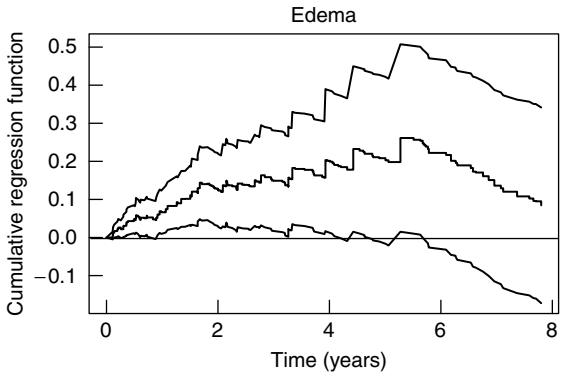

Figure 1 Cumulative regression function for time-varying excess risk effect of edema with 95 % pointwise confidence intervals

than 3.25), the presence of edema leads to a survival that lowered by $\exp(-A_e(t)$ (relative survival). For a subject with proportional effects leading to a total relative risk at R, however, the relative survival is $\exp(-R \cdot A_e(t))$.

Software

Software to fit the models using R (S-plus) is available from `www.biostat.ku.dk/~ts`

References

[1] Dabrowska, D.M. (1997). Smoothed Cox regression, *Annals of Statistics* **25**, 1510–1540.

[2] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[3] Lin, D.Y., Wei, L.J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals, *Biometrika* **80**, 557–572.

[4] Lin, D.Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61–71.

[5] Lin, D.Y. & Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes, *Annals of Statistics* **23**, 1712–1734.

[6] Martinussen, T. & Scheike, T.H. (2002). A flexible additive multiplicative hazard model, *Biometrika* **89**, 283–298.

[7] Martinussen, T., Scheike, T.H. & Skovgaard, I.M. (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models, *Scandinavian Journal of Statistics* **28**, 57–74.

[8] McKeague, I.W. & Sasieni, P.D. (1994). A partly parametric additive risk model, *Biometrika* **81**, 501–514.

[9] Murphy, S.A. & Sen, P.K. (1991). Time-dependent coefficients in a Cox-type regression model, *Stochastic Processes and their Applications* **39**, 153–180.

[10] Sasieni, P.D. (1996). Proportional excess hazards, *Biometrika* **83**, 127–141.

[11] Scheike, T.H. & Zhang, M.-J. (2002a). An additive-multiplicative Cox-Aalen model, *Scandinavian Journal of Statistics* **28**, 75–88.

[12] Scheike, T.H. & Zhang, M.-J. (2003). Extensions and applications of the Cox-Aalen Survival Model, *Biometrics*; **59**, 1033–1045.

[13] Therneau, T.M. & Grambsch, P.M. (2000). *Modelling Survival Data*. Springer-Verlag, New York.

[14] Zahl, P.H. (2003). Regression analysis with multiplicative and time-varying additive regression coefficients with examples from breast and colon cancer, *Statistics in Medicine* **22**, 1113–1127.

[15] Zucker, D.M. & Karr, A.F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach, *Annals of Statistics* **18**, 329–353.

THOMAS H. SCHEIKE

Administrative Databases

Administrative databases are derived from information routinely and systematically collected for purposes of managing a health care system [8]. Over the last few years, hospitals and insurers have often used such data to examine admissions, procedures, and lengths of stay. Because of new technologies allowing linkages between databases (*see* **Record Linkage**) and the increasing availability of comprehensive population information, studies using administrative data no longer need focus just on the amount and type of care. Instead, the accessibility and breadth of administrative data have made them a more general resource for the study of both health and health care (*see* **Health Services Research, Overview**) [56].

Questions such as the following can be approached using population-based administrative data:

1. How does the use of procedures, medications, and other health services vary with personal characteristics, such as age, gender, race, income, and health status (*see* **Descriptive Epidemiology**; **Health Care Utilization Data**)?
2. How does the use of these services vary with the source or mechanism of payment (*see* **Health Care Financing**)?
3. How does the use of these services vary across hospitals, communities, and regions (*see* **Small Area Variation Analysis**)?
4. How do the short-term and long-term outcomes of health care vary with personal, payer, and geographic or system characteristics (*see* **Outcomes Research**)?
5. How do total health care costs, and the distribution of component costs, vary with personal, payer, and geographic or system characteristics?
6. Is high use of specific health services associated with better outcomes? Are unhealthy populations "underserved" by the health care system?
7. How have the use of health services and the outcomes of care changed over time?
8. What is the appropriate level of health care resources for a population or region?
9. In what areas is the health care system consuming excess resources and, therefore, deserving of regulatory or market constraints?
10. How are the outcomes and processes of health care related? Which physicians, hospitals, nursing facilities, and health plans have the best outcomes and processes of care? How can the quality of care provided elsewhere be improved?
11. How does the natural history of disease vary with personal and geographic characteristics? How has it changed over time?
12. Are diagnostic or therapeutic methods as effective in the community as they are in randomized controlled trials (*see* **Clinical Trials, Overview**)?

Figure 1 presents a view of an ideal administrative database with a research registry playing a central role. Such a registry, with its ability to generate meaningful information on each individual's life course, helps multiply the number of health and health care studies that can be performed. Nonetheless, each of the associated files – alone or in conjunction with others – may permit important research.

This article provides an overview of the use of administrative databases for research on **clinical epidemiology**, health services, and population health. We wish to present a framework for understanding how administrative data can accurately and cost-effectively generate health and health care information for communities and populations.

Common Types of Administrative Data

The inclusiveness of administrative databases is strongly related to the requirements of health insurance plans and regulatory agencies. In Canada, where the population of each province is covered by a single insurance plan, health care data are comprehensive. The more complicated situation in the US has resulted in a loss of important utilization data associated with the Medicare and Medicaid programs [73]. In other developed countries, many administrative data files hold information about eligibility or enrollment, life events, claims and services, special programs, and providers. We present a brief overview of each type of file in what follows.

Registries

A population registry incorporates information on birth, death, mobility within a catchment area (such as

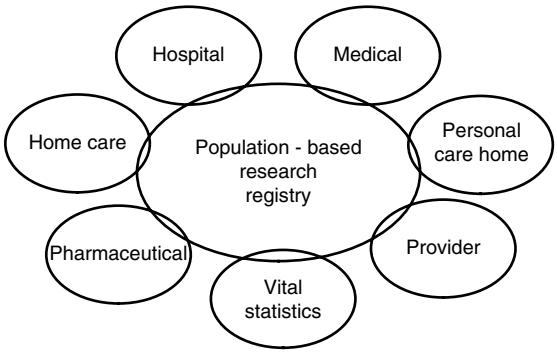

Figure 1 An ideal administrative database

a province), and in- and out-migration for all people enrolled with an insurance or benefit plan. A registry is essential for following a study cohort (*see* **Cohort Study**), providing denominators for analyses of **rates**, and updating data for each enrolled individual. When administrative registries are compared and combined into a research registry which accurately defines the health insurance status for each resident over many years the value of the tool is enhanced [64].

A research registry permits extensive checks on "subject misidentification". If an individual's identifiers are incorrect, utilization, loss to follow-up, and mortality may be misassigned. In one study, a lack of adjustment for women who were not appropriately identified, because they either left the province or changed their health care number, led to a serious underestimate of the number of women who may have developed cancer following breast augmentation [4].

Standardized systems for updating and reviewing the quality of registry data are critical. Typical checks on the accuracy of a population registry rely on other sources of information (**disease registers, vital statistics** files, **census** data, etc.). Such checks include comparisons between the number and characteristics of individuals in particular categories (such as age/gender/place of residence) with aggregate statistics from organizations such as the census [28, 64]. Often, there are opportunities to compare registry information on individual mortality and loss to follow-up with primary data collection; these comparisons are particularly useful in uncovering errors affecting a small percentage of the population.

Life Events

Acquiring and maintaining up-to-date demographic information generally requires integrating files from different sources. Life events such as birth (date and place), marriage (date and place), and death (date, place, and cause) are typically recorded in vital statistics files. In Canada, vital statistics files need to be better coordinated with utilization data maintained by provincial health departments to provide a standard health registration number on all death records.

In the US, linkages between vital statistics and health care utilization files (*see* **Record Linkage**) are rather unusual. However, the Medicare Provider Analysis and Review dataset from the Health Care Financing Agency is linked to an enrollment file indicating the date of death. Some states, such as California, have also linked patient discharge data to birth and death records. Both Canadian (through the Statistics Canada Mortality File) and American statistical agencies (through the **National Center for Health Statistics'** National Death Index) provide an additional route for funded investigators to access mortality data.

Claims and Services

Physician services recorded in an administrative database may include information regarding physician visits, surgical procedures, immunizations, prescription drugs, and diagnostic tests such as Papanicolaou tests. Descriptive fields identify the patient, the physician, and, where relevant, the institution. Typically, each claim describes a single service or event for a specific patient by a specific provider on a single date. If a single visit results in several billable services, more than one computer record may be generated; attention to detail is critical so that a single test or service is not counted twice (*see* **Drug Utilization Patterns**).

Documentation requirements affect the extent to which any recording system captures a population's ambulatory care patterns. In Manitoba, neither visits to the provincial cancer foundation (after the first) nor visits to health care providers other than physicians are recorded. However, because of the frequency of fee-for-service care and the requirement that salaried physicians submit "dummy" claims, from 90% to 98% of all physician-provided ambulatory care is captured in the existing system.

Administrative files describing hospital or nursing home stays are often maintained by regulatory agencies or provider associations. Each hospital discharge abstract describes services provided to a patient during a specified period in that facility. Most hospital discharge files are similar, including patient identifiers, demographic information such as gender and age or birth date, dates of admission and separation (discharge), diagnostic codes, payer source and charges, and procedure codes for all surgery (*see* **Classifications of Medical and Surgical Procedures**). Files for nursing home stays are similar but usually include little, if any, diagnostic and surgical information.

Although many private and government programs maintain files that describe individuals' use of their services, the agency managing the data must be willing to permit linkage (*see* **Record Linkage**). Service files may be maintained by nutrition programs (e.g. the Women, Infants, Children program in the US), immunization programs, local public health programs, education and rehabilitation programs, special programs for children with developmental disorders, and various voluntary agencies.

Provider Data

Physician information will obviously depend on the requirements of the relevant insurer or registrar. Items such as age, sex, education, specialty, and experience are likely to be available. Such data are particularly useful for physician manpower planning (*see* **Health Workforce Modeling**) and for comparisons of practice patterns among different types of physicians.

Health care organizations often submit data about their organizational, structural, and financial characteristics to trade associations and regulatory agencies. This information can be linked to individual-level data, so that differences in the utilization, costs, and outcomes of health care can be correlated with institutional characteristics. This permits testing hypotheses related to the impact of hospital size, ownership, teaching status, financial health, intensive care availability, and nurse staffing levels. Provider information collected by trade associations may be relatively difficult to obtain or may be incomplete because participation in the data system is typically voluntary.

Uses of Administrative Data

As fiscal restraint and organizational change continue, administrative databases can help to answer questions regarding the complex interplay among population characteristics, health status (*see* **Quality of Life and Health Status**), and health care utilization patterns. Such analyses can target health reform efforts, highlight the correlates of apparent overuse or underuse, and identify low-variation and high-variation services for which discretion plays a greater role (*see* **Health Care Utilization Data**; **Drug Utilization Patterns**). Differences in utilization can be related to:

1. Individual characteristics, such as age, gender, race, income, education, medical history, and comorbidity.
2. Payer characteristics, such as the source and method of payment [15, 42].
3. Characteristics of health care organizations, such as teaching status, size, ownership, location, and staffing levels [20].
4. Characteristics of small areas, states and provinces, and countries [22, 74].

Studies comparing the use of specific health services may provide important information for managing the delivery of preventive health services to an entire population (e.g. opthalmologic examination of diabetic patients), for directing additional resources to generally underserved individuals, and for redesigning health care organizations to deliver higher priority services [72]. Administrative data are particularly important for evaluating the impact of changes in the health care system, such as capitating physician payment, restricting pharmaceutical formularies, and closing hospital beds.

Profiling physician use of health services is a popular application in the 1990s [3]. Hospitals profile their physicians' length of stay pattern and their prescribing of high-cost medications. Health plans describe physician choice of screening tests, as well as their subspecialty referral rates and hospitalization rates. Such studies have revealed, for example, that physicians vary considerably in their use of Papanicolaou testing, with some overtesting and others markedly undertesting relative to current guidelines [65]. Physicians also vary in their management of breast cancer and in their proclivity to hospitalize, even after controlling for patient characteristics [21, 61]. Disease-specific utilization measures that assess

ambulatory care quality and access have several uses: to reward physicians who use preventive services appropriately, to counsel physicians who do not, and to channel patients away from primary care physicians who overuse technology-intensive and subspecialty services [46].

In the US, administrative data are also being used to profile health care delivery systems, such as health maintenance organizations (HMOs) and independent practice associations (IPAs). Large employers and employer coalitions have pressured health insurance plans to produce information on quality as well as on price. In response, the National Committee for Quality Assurance [48] developed a Health Plan Employer Data Information System (HEDIS) which includes a set of quality indicators representing use of various preventive procedures. To promote more comprehensive measures, several major purchasers and managed care advocates established the Foundation for Accountability [76]. Many of these measures may be based on administrative data, although operational definitions were not released by early 1997.

Costs

Because of global budgeting, Canadian hospital databases typically do not include direct cost data. Most administrative databases in the US, however, include data on total charges and/or charges for selected components of care (e.g. pharmacy, laboratory, supplies). Databases maintained by payers, such as state Medicaid programs and insurance companies, also include information on allowable charges or actual payments to providers. Although these payments sometimes represent only a fraction of the billed charges, payments by government health insurance programs represent a better measure of public investment than billed charges. These financial data can be used for five general types of research:

1. Cost-profiling studies to show the mean costs incurred by specific physicians, hospitals, and other health care providers. Similar to the utilization profiles discussed above, cost profiles can be used by health plans to "delist" individual providers and offer incentives for improved financial performance.
2. Cost-of-illness studies to estimate the aggregate cost of medical and nonmedical care for a specific condition. This aggregate cost may be stratified by payer or by demographic characteristics to show how the condition's economic impact is distributed.
3. Cost-containment studies to evaluate the effects of various strategies, such as pharmacy benefit caps, mandatory second-opinion programs, and preauthorization for hospital care.
4. Cost-effectiveness studies to compare the costs of various diagnostic or therapeutic strategies to achieve a given health benefit (e.g. quality-adjusted life year gained or cancer death prevented).
5. Benefit–cost studies to compare the total economic costs and benefits of a health-related intervention (*see* **Health Economics**).

Outcomes

Administrative data can unobtrusively provide useful information about selected health outcomes. Death is perhaps the best example, being well defined, clearly documented, and easily verified. However, some administrative data sets only capture in-hospital deaths, while others suffer from delayed or inexact reporting of death. In-hospital or 30-day death has been used as a measure of quality of care for such high-risk conditions as myocardial infarction, stroke, pneumonia, and congestive heart failure.

Recently, researchers have refined **risk-adjustment** models that use administrative data to estimate expected and risk-adjusted mortality rates for physicians, hospitals, and health systems. These risk-adjustment models have been developed by American government agencies such as the Health Care Financing Administration [39] and the California Office of Statewide Health Planning and Development [41], provider organizations such as the Hospital Research and Educational Trust, proprietary organizations such as 3M Health Information Systems and MediQual Systems [30], and independent investigators [10, 11]. For most conditions, these risk-adjustment models do not discriminate between decedents and survivors as well as those that include detailed clinical data [31, 39]. This has been most clearly demonstrated for stroke, pneumonia [33, 34] and coronary artery bypass surgery [25]. Administrative and clinical data allow

comparable discrimination only when post-admission complications are used in estimating the probability of death [32, 41].

When administrative data are used to estimate indirectly standardized mortality ratios (*see* **Standardization Methods**) or mortality z scores for hospitals, the results correlate reasonably well with those of more sophisticated models based on either clinical data or more carefully abstracted data (e.g. $r = 0.75$ to 0.87) [25, 32, 39, 59]. In practice, however, hospitals frequently shift above or below any **outlier** threshold (e.g. $P < 0.05$, $P < 0.01$) when clinical data are added to, or substituted for, administrative data. The **predictive value** of being labeled as a mortality outlier using administrative data may be as low as 50%–64% [25, 39].

The validity of risk-adjusted mortality estimates based on administrative data is supported by some, but not all, studies linking outcomes and processes of care. Two studies found significant hospital-level correlations between risk-adjusted mortality measures and quality problem rates determined by peer review organizations, although these correlations varied across states and across conditions [26, 71]. Physician-rated preventable deaths for pneumonia and stroke (but not myocardial infarction) occurred more often at high-mortality than at low-mortality hospitals, although explicit process scores did not differ [12]. Hospitals with low risk-adjusted mortality after myocardial infarction (based on administrative data) administered aspirin more quickly and were more aggressive with catheterization and revascularization than high-mortality hospitals [41]. These studies suggest that administrative data can be used to help identify process deficiencies.

For conditions that cause significant morbidity but are rarely fatal, death has limited utility as a measure of quality of care. Other measures developed from administrative data focus on adverse events and event-free survival, but their validity has not been well established. These are:

1. Post-operative complications can be identified using **International Classification of Diseases (ICD)**-9-CM diagnoses and procedures. Early measures suffered from **misclassification** bias because they were limited to diagnoses explicitly labeled as complications; such diagnoses are likely to be poorly documented by physicians and undercoded by hospitals, [10, 40]. More recent measures, developed both for specific conditions and for broader categories of patients, demonstrate better face and content validity, but still lack construct validity (*see* **Health Status Instruments, Measurement Properties of**) [30, 58, 67]. One promising "comorbidity-adjusted complication risk" measure incorporates physicians' consensus estimates of the probability that any secondary diagnosis is a complication of a given admitting diagnosis, but this has only been validated to a limited extent [2].
2. Post-operative length of stay may be an indicator of post-operative complications, especially if one identifies long-stay outliers instead of just comparing mean lengths of stay. This measure is free from coding bias and may be more predictive of true complications than measures based on ICD-9-CM diagnoses [40, 60]. However, length of stay may be heavily influenced by social factors, such as marital status and homelessness, and by the local availability of skilled nursing beds.
3. Post-discharge readmissions can be identified using linked hospital discharge or claims data. Being more common than death for most conditions, readmission is a promising measure of quality [77]. Panels of specialists convened by the US Health Care Financing Administration have developed lists of readmission diagnoses that indicate adverse events after orthopedic, cardiac, and general surgery [53]. Some investigators have found that "unplanned" readmissions are related to unresolved concerns at discharge, but others found no association with physician quality-of-care ratings (at one hospital) or quality problems determined by peer review organizations [1, 27, 70]. Readmission rates may well be associated with hospital-bed availability, complicating efforts to link readmissions with quality measures [17].
4. After certain procedures, such as hip fracture repair and elective discectomy, reoperations may represent treatment failures. However, reoperations tend to be low-frequency events and, for many conditions, may reflect disease progression more than quality of care. The validity of reoperation rates as a quality measure is unknown.
5. Emergency room visits and unscheduled physician visits after hospital discharge can be ascertained using claims data, but the reliability and validity of such measures are unknown. Given

the poor quality of outpatient diagnostic coding, separating routine visits from complications may be impossible.

In general, administrative data may be a less biased source of outcomes data than **case series** reported from prominent institutions or randomized controlled trials with rigid **eligibility and exclusion criteria**. Such information about mortality and other patient outcomes in the "real world" may help patients choose among therapeutic options. Given the universality of administrative data, differences in outcomes can be related to individual characteristics, such as age, gender, race, income, education, medical history, and comorbidity. These analyses can identify unusual but powerful risk factors for adverse outcomes and sort out the independent effects of multiple variables to facilitate clinical decision making.

Studies of characteristics of health care organizations (teaching status, size, ownership, location, and staffing levels) can highlight the general features of high-quality organizations and promote the regionalization of treatments having a clear association between hospital/surgeon volume and outcomes (e.g. coronary bypass surgery, angioplasty [37]). Relatively little research has studied the influence of payer characteristics, such as the source and method of payment. Analyses of small areas, states and provinces, and countries (*see* **Mortality, International Comparisons**) can elucidate the separate impact of physiologic and sociocultural factors. For example, American–Canadian studies suggest that a one- to four-day preoperative delay after hip fracture is associated with mortality simply by being a marker of preoperative morbidity [62].

Studies using administrative data to evaluate the comparative effectiveness of specific treatments suffer from the limitations of **nonrandomized** designs (*see* **Observational Study**), such as **confounding** because of self-selection into treatment groups (*see* **Selection Bias**), and potential misclassification of risk factors and outcomes. However, they appear useful when a randomized controlled trial is too costly, impractical, or unethical (because a therapy has already been adopted as the "standard of care"). Through careful subset analyses, for example, incidental appendectomy during open cholecystectomy was found to increase the risk of death [78].

New techniques from economics and epidemiology may also help control for differences in pretreatment health status. By using instrumental variables to account for selection bias, McLellan and colleagues found care within the first 24 hours to have more effect on myocardial infarction mortality in the elderly than later invasive procedures [43]. The **rate** ratio approach takes an epidemiologic perspective, using temporal profiles of death rates in the year after surgery to adjust for preoperative death risk [66].

Administrative data can support both randomized and nonrandomized designs. For example, administrative data external to, and independent of, a randomized trial were used to construct health histories of breast disease prior to entry in the controversial Canadian National Breast Screening Study; no definitive evidence supporting a nonrandom allocation of women was found [7]. A nonrandomized, but controlled, approach to introducing competing vaccines (e.g. acellular pertussis) (*see* **Vaccine Studies**) could generate product-specific estimates of clinical effectiveness and serious adverse reactions. Administrative data could be incorporated into a system to provide early warning of adverse events following administration of vaccines, antibiotics, antiarrhythmics, and other medications (*see* **Surveillance of Diseases**; **Postmarketing Surveillance of New Drugs and Assessment of Risk**). Many hospitals already use such systems internally, but they could be expanded on a regional or national scale.

Interventions

Administrative data can actually be used to carry out health interventions (if **confidentiality** constraints permit contacting individual patients) (*see* **Confidentiality in Epidemiology**). Client-oriented interventions (e.g. reminder postcards) and provider-oriented interventions (e.g. targeted feedback) both appear promising [69]. Although trials of reminder systems have apparently never been performed for an entire population, they would not be difficult to implement from guidelines published by various task forces on preventive services [52]. Some American health plans are already using enrollment lists to generate reminders. There are many unanswered questions about the optimal number, timing, and type of reminders for preventive care (*see* **Screening Benefit, Evaluation of**).

Other interventions might focus on identifying candidates for a randomized controlled trial or patients who seem to have "fallen between the

cracks" in the health care system. For example, claims data could be used to identify **AIDS** patients who have experienced an episode of Pneumocystis pneumonia but are not receiving prophylaxis to prevent recurrence. Administrative data could be used to identify persons with possible drug–drug interactions and those who might benefit from home care services because of their high risk of institutionalization (*see* **Pharmacoepidemiology, Adverse and Beneficial Effects**).

Characteristics of Administrative Data

Strengths

The unique advantages of administrative data derive largely from their accessibility, inclusiveness, and flexibility [19]:

1. Using data that have already been collected instead of primary data provides an opportunity to conduct research relatively quickly and cost-effectively.
2. The natural, real-world setting maximizes external validity. Observed outcomes of care using an experimental design may be less representative of "real-world" care than measures derived from administrative data. The inclusiveness of administrative data facilitates selection of representative samples of people, health care providers, and institutions for study.
3. Large sample sizes make small effect sizes detectable.
4. Standards or benchmarks (e.g. for defining an adequate level of care) can be established empirically.
5. Long-term demographic trends (e.g. aging population), changes in utilization, and cost increases can be disentangled as factors influencing system-wide costs [9].
6. Unique person-specific identifiers facilitate both **cohort** and **case–control studies** by the capacity to create comparison groups, to track individuals over time, and to construct histories of prior utilization and events [51, 64].
7. Utilization and mortality rates can be compared across communities to understand better the relationships among such variables [75].
8. Linkage with other databases creates opportunities for cost-effectively adding information collected using other methodologies or by other investigators [49].
9. The absence of the **recall bias** often associated with **surveys** increases the validity of the constructs or measures generated from administrative data.
10. The reliability and validity of many data sets are known and monitored; additional studies can be carried out by comparing records from several independent sources.

Limitations

Some of the limitations of administrative databases, and the extent to which they can be overcome, are presented in the following discussion.

1. *Reliability:* Maintaining data quality in medical, hospital, and nursing home files is an ongoing challenge. Variation in training and supervision leads to differences in coding styles. Information on intensive care and emergency room visits is often not standardized across hospitals. Definitions may be vague or difficult to apply. States and provinces differ in their definitions of institutions (particularly rehabilitation, long-term care hospitals, and extended care units). Ethnicity (*see* **Ethnic Groups**) presents particular problems. In Canada, a definition of Indians with rights specified by treaty has not been coordinated between different levels of government. Defining racial categories has also been problematic in the US, where patients with the same ethnic heritage may self-identify differently.
2. *Validity:* Reimbursement using **diagnosis-related groups (DRGs)** and public comparisons of **risk-adjusted** outcomes may result in bias because of pressures on reporting entities and fear of bad publicity [68].
3. *ICD-9-CM coding:* The lack of operational criteria for defining each disease compromises the quality of data generated by the International Classification of Diseases (ICD) system [16]. Variability among clinicians in diagnosing the same condition differently can produce unreliable statistics [38]. The extent to which shifts in disease rates over time are due to changes in occurrence, or to technological advances that generate improved diagnostic methods, can be difficult to discern. Illness severity and disease

progression affect both prognosis and therapy, but these factors can rarely be recorded using ICD-9-CM.

4. *Personal identifiers:* Many databases do not have personal identifiers that permit reliable record linkage or tracking of individuals over time.
5. *Timeliness:* Delays in submission and processing hinder time-dependent studies, leading to criticisms that one-or two-year-old data do not reflect current practices.
6. *Scope:* Populations of interest, such as the uninsured in the US, may not be included. Generally, only certain categories of billable services can be ascertained. Information on conditions (e.g. the common cold) that do not usually require contact with the health care system may be incomplete. Information on activities (e.g. immunizations) performed by nonmedical health care professionals may not be collected [54]. Similarly, hospital outpatient care was once relatively unimportant and some information systems still do not capture these data. Finally, as insurers move away from fee-for-service payment and toward capitated contracts with medical groups, they should establish incentives to ensure accurate recording of contact data.
7. *Content:* Many important data elements are not collected. For example, a lack of physiologic data (e.g. blood pressure, laboratory values) limits clinical risk-adjustment. Less specific and less standardized cost data are available in Canada, where hospitals receive global budgets, than in the US.

Database Issues: Problems and Opportunities

Reliability and Validity

Concerns about accuracy can be addressed by assessing the agreement among several data sources. In Manitoba, agreement among patient surveys, physician claims, and clinical measures is relatively high for diabetes and hypertension; this is reflected in both overall prevalence and case identification [47, 55]. The flexibility of administrative data easily permits such **sensitivity analysis**; varying the number of years in a patient history or using more stringent inclusion criteria (for example, requiring more than one diagnosis) is relatively easy. As histories of physician visits or hospital stays are incorporated into statistical models, the availability and quality of such data will become increasingly important [45, 51].

Discharge abstracts from the index hospitalizations typically fail to identify many patients with such conditions as angina, congestive heart failure, cerebrovascular disease, and prior acute myocardial infarction, although high agreement between abstracts and hospital records has been observed for diabetes [36, 64]. The number of co-morbid conditions (*see* **Co-morbidity**) identified can be increased by adding more diagnostic fields to the hospital discharge abstract, reviewing abstracts from earlier hospital stays, and using physician claim histories. In practice, however, increasing the number of co-morbidities available for each patient has only moderately improved the prediction of outcomes [63].

ICD-9-CM Coding

Using multiple ICD-9-CM codes to capture all cases of interest addresses problems that arise from use of a single code [16]. Enumerating asthma diagnoses on physician claims, along with possibly associated diagnoses (e.g. bronchitis, chronic obstructive pulmonary disease), helped estimate the extent to which an increase in the apparent prevalence of asthma was real or due to changed diagnostic coding [13]. This is especially important when a nonspecific ICD-9-CM code (e.g. unspecified intracranial hemorrhage) may substitute for a more specific code. Associated codes may clarify the presentation or severity of the condition of interest. For example, the severity of a liver injury may be estimated from the associated surgical repair codes [56].

Researchers should be familiar with techniques for assessing the quality of ICD-9-CM coding and with some of the important findings. Diagnoses are coded more reliably at the three-digit level than at the four- or five-digit level, while some diagnoses and procedures are markedly undercoded [35]. In general, coding quality in the US has improved over the past two decades [18], major procedures are coded more reliably than minor procedures [57], diagnoses that affect DRG assignment are coded more reliably than those that do not [23], and any diagnosis (e.g. mitral regurgitation, congestive heart failure) is more likely to be coded when clinically severe than when mild [36]. Truncation of the data at five or fewer diagnosis fields leads to systematic underreporting of co-morbidities

among patients who die [57]. American hospitals appear to resequence physicians' diagnoses to optimize reimbursement, so the principal diagnosis may not always be the cause of admission [29]. Because of substantial variation in coding quality across facilities [44], hospitals with implausibly low prevalences of major co-morbidities may legitimately be excluded from comparative studies [41].

New Directions

Several promising approaches to increasing the scope and content of research using administrative data are under way. Record linkage, performed with stringent controls to protect confidentiality, is critical both for expanding available data content and facilitating studies of reliability and validity. Because having identifiers to permit following individuals and linking files is so important for research that can benefit the public, efforts to persuade the jurisdictions that do not allow such tracking to do so are clearly in order [50].

Primary data collection should be targeted to fill in gaps in administrative data; for example, public health nurses can enter childhood immunization data to supplement physician claims and provide a picture of population coverage. Since routinely collected patient information about symptoms and functional ability might facilitate outcomes research, several American medical centers have created systems that include such functional measures as well as traditional administrative variables.

As noted in the earlier section on reliability and validity, health status information on individuals and areas from physician and hospital data can be compared fruitfully with that produced using other methodologies [6]. Registry, vital statistics, and census data are all being used to construct small area measures of premature mortality, **life expectancy**, and cause-specific mortality [5, 14]. Census data provide community-level socioeconomic information; in special circumstances, Canadian individual utilization data may be linked to the census [28].

Having a single payer with universal coverage facilitates developing integrated systems with population-wide scope. Comparing a Canadian population health information system to what might be done in the US, Greenfield [24] noted: "Just beginning to collect similar information on those who have health care coverage or can provide it for themselves in our disorganized multipayer system would require a significant expenditure of funds." Although lack of data on the American uninsured will remain a major obstacle to generating a true population health information system, at least one state (Minnesota) is building a state-wide data system on all aspects of health care (*see* **Health Services Data Sources in the US**).

This ability to use multiple files suggests new ways to look at problems. Thus, the integrated system in Manitoba has permitted the examination of physician supply and hospital bed supply within a population health context [62]. Ongoing research is directed toward examining rural hospital performance using two types of indicators: **population-based** (to describe the characteristics of the people living in the area served by each hospital) and hospital-based (to "describe the activities at each rural hospital"). Such contextually sensitive work represents just one new approach to using administrative data to further our knowledge of health and health care.

Acknowledgments

This work was supported by HEALNet (the Canadian Networks of Centres of Excellence Program), by a National Health Research and Development Program Career Scientist Award to Dr. L.L. Roos, and by the Manitoba Centre for Health Policy and Evaluation.

References

[1] Ashton, C.M., Kuykendall, D.H., Johnson, M.L., Wray, N.P. & Wu, L. (1995). The association between the quality of inpatient care and early readmission, *Annals of Internal Medicine* **122**, 415–421.

[2] Brailer, D.J., Kroch, E., Pauly, M.V. & Huang, J. (1996). Comorbidity-adjusted complication risk: a new outcome quality measure, *Medical Care* **34**, 490–505.

[3] Brand, D.A., Quam, L. & Leatherman, S. (1995). Medical practice profiling: concepts and caveats, *Medical Care Research and Review* **52**, 223–251.

[4] Bryant, H. & Brasher, P. (1995). Breast implants and breast cancer: re-analysis of a linkage study, *New England Journal of Medicine* **332**, 1535–1539.

[5] Carstairs, V. & Morris, R. (1991). *Deprivation and Health in Scotland.* Aberdeen University Press, Aberdeen.

[6] Cohen, M.M. & Macwilliam, L. (1995). Measuring the health of the population, *Medical Care* **33**, Supplement, DS21–DS42.

[7] Cohen, M.M., Kaufert, P.A., Macwilliam, L. & Tate, R.B. (1996). Using an alternative data source to

examine randomization in the Canadian National Breast Screening Study, *Journal of Clinical Epidemiology* **49**, 1039–1044.

[8] Cohen, M.M., Roos, N.P., DeCoster, C., Black, C. & Decker, K.M. (1995). Manitoba's population-based databases and long-term planning: beyond the hospital databases, *Healthcare Management Forum* **8**, 5–13.

[9] Demers, M. (1996). Factors explaining the increase in cost for physician care in Quebec's elderly population, *Canadian Medical Association Journal* **155**, 1555–1560.

[10] DesHarnais, S.I., McMahon, L.F., Wroblewski, R.I. & Hogan, A.J. (1990). Measuring hospital performance: the development and validation of risk-adjusted indexes of mortality, readmissions, and complications, *Medical Care* **28**, 1127–1141.

[11] Deyo, R.A., Cherkin, D.C. & Ciol, M.A. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases, *Journal of Clinical Epidemiology* **45**, 613–619.

[12] Dubois, R.W., Rogers, W.H., Moxley, J.H., Draper, D. & Brook, R.H. (1987). Hospital inpatient mortality: is it a predictor of quality?, *New England Journal of Medicine* **317**, 1674–1680.

[13] Erzen, D., Roos, L.L., Manfreda, J. & Anthonisen, J. (1995). Changes in asthma severity in Manitoba, *Chest* **108**, 16–23.

[14] Eyles, J., Birch, S., Chambers, J., Hurley, J. & Hutchinson, B. (1991). A needs-based methodology for allocating health care resources in Ontario, Canada: development and an application, *Social Science and Medicine* **33**, 489–500.

[15] Feinglass, J. & Holloway, J. (1991). The initial impact of the Medicare prospective payment system on U.S. health care: a review of the literature, *Medical Care Review* **48**, 91–115.

[16] Feinstein, A.R. (1988). Scientific standards in epidemiologic studies of the menace of daily life, *Science* **242**, 1257–1263.

[17] Fisher, E.S., Wennberg, J.E., Stukel, T.A. & Sharp, S.M. (1994). Hospital readmission rates for cohorts of Medicare beneficiaries in Boston and New Haven, *New England Journal of Medicine* **331**, 989–995.

[18] Fisher, E.S., Whaley, F.S., Krushat, W.M., Malenka, D.J., Fleming, C., Baron, J.A. & Hsia, D.C. (1992). The accuracy of Medicare's hospital claims data: progress has been made, but problems remain, *American Journal of Public Health* **82**, 243–248.

[19] Flood, A.B. (1990). Peaks and pits of using large data bases to measure quality of care, *International Journal of Technology Assessment in Health Care* **6**, 253–262.

[20] Fuchs, V.R. & Garber, A.M. (1990). The new technology assessment, *New England Journal of Medicine* **323**, 673–677.

[21] Goel, V., Olivotto, I., Hislop, T.G., Sawka, C., Coldman, A., Holowaty, E.J., & the British Columbia/ Ontario Working Group (1997). Patterns of initial management of node-negative breast cancer in two Canadian provinces, *Canadian Medical Association Journal* **156**, 25–35.

[22] Goel, V., Williams, J.I., Anderson, G.M., Blackstien-Hirsch, P., Fooks, C. & Naylor, C.D., eds (1996). *Patterns of Health Care in Ontario. The ICES Practice Atlas*. Canadian Medical Association, Institute for Clinical Evaluative Sciences, Ottawa.

[23] Green, J. & Wintfeld, N. (1993). How accurate are hospital discharge data for evaluating effectiveness of care?, *Medical Care* **31**, 719–731.

[24] Greenfield, L. (1996). Without universal coverage, health care use data do not provide population health, *Milbank Quarterly* **74**, 33–36.

[25] Hannan, E.L., Kilburn, H., Lindsey, M.L. & Lewis, R. (1992). Clinical versus administrative data bases for CABG surgery: does it matter? *Medical Care* **30**, 892–907.

[26] Hartz, A.J., Gottlieb, M.S., Kuhn, E.M. & Rimm, A. (1993). The relationship between adjusted hospital mortality and the results of peer review, *Health Services Research* **27**, 765–777.

[27] Hayward, R.A., Bernard, A.M., Rosevear, J.S., Anderson, J.E. & McMahon, L.F. (1993). An evaluation of generic screens for poor quality of hospital care on a general medicine service, *Medical Care* **31**, 394–402.

[28] Houle, C., Berthelot, J-M., David, P., Mustard, C.A., Roos, L.L. & Wolfson, M.C. (1996). *Project on Matching Census 1986 Database and Manitoba Health Care Files: Private Households Component (Analytical Studies Branch Research Paper Series, No. 91)*. Statistics Canada, Ottawa.

[29] Hsia, D.C., Ahern, C.A., Ritchie, B.P., Moscoe, L.M. & Krushat, W.M. (1992). Medicare reimbursement accuracy under the prospective payment system, 1985 to 1988, *Journal of the American Medical Association* **268**, 896–899.

[30] Iezzoni, L.I. (1994). Risk and outcomes, in *Risk Adjustment for Measuring Health Care Outcomes*, L.I. Iezzoni, ed. Health Administration Press, Ann Arbor.

[31] Iezzoni, L.I. (1995). Risk adjustment for medical effectiveness research: an overview of conceptual and methodological considerations, *Journal of Investigative Medicine* **43**, 136–150.

[32] Iezzoni, L.I., Ash, A.S., Shwartz, M., Daley, J., Hughes, J.S. & Mackiernan, Y.D. (1996). Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method, *American Journal of Public Health* **86**, 1379–1387.

[33] Iezzoni, L.I., Shwartz, M., Ash, A.S., Hughes, J.S., Daley, J. & Mackiernan, Y.D. (1995). Using severity-adjusted stroke mortality rates to judge hospitals, *International Journal of Quality of Health Care* **7**, 81–94.

[34] Iezzoni, L.I., Shwartz, M., Ash, A.S., Hughes, J.S., Daley, J. & Mackiernan, Y.D. (1996). Severity measurement methods and judging death rates for pneumonia, *Medical Care* **34**, 11–28.

[35] Institute of Medicine (1977). *Reliability of Hospital Discharge Abstracts*, National Academy of Sciences, Washington.

[36] Jollis, J.G., Ancukiewicz, M., DeLong, E.R., Pryor, D.B., Muhlbaier, L.H. & Mark, D.B. (1993). Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research, *Annals of Internal Medicine* **119**, 844–850.

[37] Jollis, J.G., Peterson, E.D., DeLong, E.R., Mark, D.B., Collins, S.R., Muhlbaier, L.H. & Pryor, D.B. (1994). The relation between the volume of coronary angioplasty procedures at hospitals treating Medicare beneficiaries and short-term mortality, *New England Journal of Medicine* **331**, 1625–1629.

[38] Koran, L.M. (1975). The reliability of clinical methods, data and judgements (Part 2), *New England Journal of Medicine* **293**, 695–701.

[39] Krakauer, H., Bailey, R.C., Skellan, K.J., Stewart, J.D., Hartz, A.J., Kuhn, E.M. & Rimm, A.A. (1992). Evaluation of the HCFA model for the analysis of mortality following hospitalization, *Health Services Research* **27**, 317–335.

[40] Kuykendall, D.H., Ashton, C.M., Johnson, M.L. & Geraci, J.M. (1995). Identifying complications and low provider adherence to normative practices using administrative data, *Health Services Research* **30**, 531–554.

[41] Legnini, M.W., Zach, A., Richards, T., Romano, P.S., Remy, L.L. & Luft, H.S. (1996). *Second Report of the California Hospital Outcomes Project. Acute Myocardial Infarction*. Vol. 2: *Technical Appendix*. Office of Statewide Health Planning and Development, Sacramento.

[42] Leibson, C.L., Naessens, J.M., Campion, M.E., Krishan, I. & Ballard, D.J. (1991). Trends in elderly hospitalization and readmission rates for a geographically defined population: pre- and post-prospective payment, *Journal of the American Geriatrics Society* **39**, 895–904.

[43] McClellan, M.B., McNeil, B.J. & Newhouse, J.P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables, *Journal of the American Medical Association* **272**, 859–866.

[44] Meux, E.F., Stith, S.A. & Zach, A. (1988). *Report of Results from the OSHPD Reabstracting Project: An Evaluation of the Reliability of Selected Patient Discharge Data, July through December 1988*. Office of Statewide Health Planning and Development, Sacramento.

[45] Mitchell, J.B., Ballard, D.J., Whisnant, J.P., Ammering, C.J., Matchar, D.B. & Samsa, G.P. (1996). Using physician claims to identify postoperative complications of carotid endarterectomy, *Health Services Research* **31**, 141–152.

[46] Moy, E. & Hogan, C. (1993). Access to needed follow-up services. Variations among different Medicare populations, *Archives of Internal Medicine* **153**, 1815–1823.

[47] Muhajarine, N., Mustard, C.A., Roos, L.L., Young, T.K. & Gelskey, D.E. (1997). A comparison of survey data and physician claims data for detecting hypertension, *Journal of Clinical Epidemiology* **50**, 711–718.

[48] National Committee for Quality Assurance (1996). *HEDIS 2.5* Washington.

[49] Newcombe, H.B. (1987). Record linking: the design of efficient systems for linking records into individual and family histories, in *Textbook of Medical Record Linkage*, J.A. Baldwin, E.D. Acheson & W.J. Graham, eds. Oxford University Press, Oxford, pp. 15–38.

[50] Newcombe, H.B. (1994). Cohorts and privacy, *Cancer Causes and Control* **5**, 287–291.

[51] Nichol, K.L., Margolis, K.L., Wuorenma, J. & Von Sternberg, T. (1994). The efficacy and cost effectiveness of vaccination against influenza among elderly persons living in the community, *New England Journal of Medicine* **331**, 778–784.

[52] Ornstein, S.M., Garr, D.R., Jenkins, R.G., Rust, P.F. & Arnon, A. (1991). Computer-generated physician and patient reminders: tools to improve population adherence to selected preventive services, *Journal of Family Practice* **32**, 82–90.

[53] Riley, G., Lubitz, J., Gornick, M., Mentnech, R., Eggers, P. & McBean, M. (1993). Medicare beneficiaries: adverse outcomes after hospitalization for eight procedures. *Medical Care* **31**, 921–949.

[54] Roberts, J.D., Poffenroth, L.A., Roos, L.L., Bebchuk, J.D. & Carter, A.O. (1994). Monitoring childhood immunizations: a Canadian approach, *American Journal of Public Health* **84**, 1666–1668.

[55] Robinson, J.R., Young, T.K., Roos, L.L. & Gelskey, D.E. (1997). Estimating the burden of disease: comparing administrative data and self-reports, *Medical Care* **35**, 932–947.

[56] Romano, P.S. & Luft, H.S. (1992). Getting the most out of messy data: problems and approaches for dealing with large administrative data sets, in *Medical Effectiveness Research Data Methods* (AHCPR Pub. No. 92-0056), M.L. Grady & H.A. Schwartz, eds. US Department of Health and Human Services, Rockville, pp. 57–75.

[57] Romano, P.S. & Mark, D.H. (1994). Bias in the coding of hospital discharge data and its implications for quality assessment, *Medical Care* **72**, 81–90.

[58] Romano, P.S., Campa, D.R. & Rainwater, J.A. (1997). Elective cervical discectomy in California: postoperative in-hospital complications and their risk factors, *Spine*, to appear.

[59] Romano, P.S., Roos, L.L., Luft, H.S., Jollis, J.G., Doliszny, K. & the Ischemic Heart Disease Patient Outcomes Research Team (1994). A comparison of administrative versus clinical data: coronary artery bypass surgery as an example, *Journal of Clinical Epidemiology* **47**, 249–260.

[60] Romano, P.S., Zach, A., Luft, H.S., Rainwater, J.A., Remy, L.L. & Campa, D. (1995). The California Hospital Outcomes Project: using administrative data to

compare hospital performance, *The Joint Commission Journal on Quality Improvement* **21**, 668–682.

[61] Roos, N.P. (1992). Hospitalization style of physicians in Manitoba: the disturbing lack of logic in medical practice, *Health Services Research* **27**, 361–384.

[62] Roos, N.P., Black, C., Wade, J. & Decker, K. (1996). How many general surgeons do you need in rural areas? Three approaches to physician resource planning in southern Manitoba, *Canadian Medical Association Journal* **155**, 395–401.

[63] Roos, L.L., Walld, R., Ramano, P.S. & Roberecki, S. (1996). Short-term mortality after repair of hip fracture: do Manitobans do worse?, *Medical Care* **34**, 310–326.

[64] Roos, L.L., Mustard, C.A., Nicol, J.P., McLerran, D.F., Malenka, D.J., Young, T.K. & Cohen, M.M. (1993). Registries and administrative data: organization and accuracy, *Medical Care* **31**, 201–212.

[65] Russell, L.B. (1994). *Educated Guesses: Making Policy About Medical Screening Tests*. University of California Press, Los Angeles.

[66] Seagroatt, V. & Goldacre, M. (1994). Measures of early postoperative mortality: beyond hospital fatality rates, *British Medical Journal* **309**, 361–365.

[67] Silber, J.H., Rosenbaum, P.R., Schwartz, J.S., Ross, R.N. & Williams, S.V. (1995). Evaluation of the complication rate as a measure of the quality of care in coronary artery bypass graft surgery, *Journal of the American Medical Association* **274**, 317–323.

[68] Steinwald, B. & Dummit, L.A. (1989). Hospital case-mix change: sicker patients or DRG creep?, *Health Affairs* **8**, (Summer), 35–47.

[69] Tannenbaum, T.N., Gyorkos, T.W., Abrahamowicz, M., Bedard, L., Carsley, J., Franco, E.D., Delage, G., Miller, M.A., Lamping, D.L. & Grover, S.A. (1994). Immunization delivery methods: practice recommendations, *Canadian Journal of Public Health* **85**, Supplement 1, S37–S40.

[70] Thomas, J.W. (1996). Does risk-adjusted readmission rate provide valid information on hospital quality? *Inquiry* **28**, 258–270.

[71] Thomas, J.W., Holloway, J.J. & Guire, K.E. (1993). Validating risk-adjusted mortality as an indicator for quality of care, *Inquiry* **30**, 6–22.

[72] Weiner, J.P., Parente, S.T., Garnick, D.W., Fowles, J., Lawthers, A.G. & Palmer, R.H. (1995). Variation in office-based quality: a claims-based profile of care provided to Medicare patients with diabetes, *Journal of the American Medical Association* **273**, 1503–1508.

[73] Welch, W.P. & Welch, H.G. (1995). Medicare analysis: fee-for-data: a strategy to open the HMO black box, *Health Affairs* **14**, (Winter), 104–116.

[74] Welch, W.P., Miller, M.E., Welch, H.G., Fisher, E.S. & Wennberg, J.E. (1993). Geographic variation in expenditures for physicians' services in the United States, *New England Journal of Medicine* **328**, 621–627.

[75] Wennberg, J.E. & Cooper, M.M. (1996). *The Dartmouth Atlas of Health Care in the United States*. American Hospital Publishing, Chicago.

[76] Wilson, E. (1995). Getting the FAccts (Foundation for Accountability) straight, *Health Systems Review* **28**, 12–14.

[77] Wray, N.P., Ashton, C.M., Kuykendall, D.H., Petersen, N.J., Souchek, J. & Hollingsworth, J.C. (1995). Selecting disease-outcome pairs for monitoring the quality of hospital care, *Medical Care* **33**, 75–89.

[78] Wu Wen, S., Hernandez, R. & Naylor, C.D. (1995). Pitfalls in nonrandomized outcomes studies: the case of incidental appendectomy with open cholecystectomy, *Journal of the American Medical Association* **274**, 1687–1691.

LESLIE L. ROOS, PATRICK S. ROMANO & PATRICIA FERGUSSON

Admixture in Human Populations

Human society is stratified primarily because of linguistic and cultural differences. This stratification has resulted in restrictions on free interbreeding. There is usually much more interbreeding within a linguistic/cultural stratum than between such strata. Breeding between members of different strata results in an exchange of **genes**. Sometimes, because of large-scale migration, there is a sudden infusion of genes from one population to another. For example, because of slave trading from Africa to North America in the eighteenth century, there has been an exchange of genes between the African Black and American White gene pools. Since children of marriages between Blacks and Whites were socially regarded as belonging to the Black population, this flow of genes was unidirectional – from Whites to Blacks. Exchange of genes has effects on disease profiles, especially in respect of those diseases that are primarily genetic. Human geneticists have long been interested in estimating extents of admixture between populations based on genetic data. This is an important problem, especially because statistics pertaining to migration and other demographic parameters are often unavailable.

Statistical Models and Estimation Procedures

One Biallelic Locus

Consider a specific allele at a codominant biallelic genetic locus. (Codominance is assumed for algebraic simplicity. If there is dominance, minor modifications in formulas are necessary.) Suppose $p_1, p_2, \ldots, p_p$ are the frequencies of this allele in P populations. Consider a hybrid population formed by admixture of these P parental populations in proportions μ_1, $\mu_2, \ldots, \mu_P (0 \leq \mu_i \leq 1; \sum_{i=1}^{P} \mu_i = 1)$. Then, in the hybrid population, the frequency of the allele will be [2]

$$p_{\mathrm{H}} = \sum_{i=1}^{P} \mu_i p_i. \tag{1}$$

The problem is to obtain estimates of the μ_is.

For simplicity, suppose $P = 2$. Then, (1) reduces to

$$p_{\mathrm{H}} = \mu_1 p_1 + (1 - \mu_1) p_2. \tag{2}$$

Assuming that p_1 and p_2 are known without error,

$$\mu_1 = \frac{p_{\mathrm{H}} - p_2}{p_1 - p_2}. \tag{3}$$

The above assumption is crucial and unrealistic since allele frequencies are, in practice, not known without error and are estimated from samples of individuals drawn from the parental and hybrid populations. If $x_1, x_2, \ldots, x_P$, and x_{H} denote the estimates of $p_1, p_2, \ldots, p_P$, and p_{H}, then an estimate m_1 of μ_1 is

$$m_1 = \frac{x_{\mathrm{H}} - x_2}{x_1 - x_2}, \tag{4}$$

$$\mathrm{var}(m_1) \approx \frac{\mathrm{var}(x_{\mathrm{H}}) + m_1^2 \mathrm{var}(x_1) + (1 - m_1)^2 \mathrm{var}(x_2)}{(x_1 - x_2)^2}, \tag{5}$$

where $\mathrm{var}(X_i) = x_i(1 - x_i)/n_i$ = sampling variance of x_i (*see* **Sampling Distributions**), and n_i = number of individuals sampled from the ith population; $i = H, 1, 2, \ldots, P$.

In the above we have made another crucial assumption: there are no other sources (e.g. genetic drift (*see* **Population Genetics**)) contributing to the variances of allele frequencies. In what follows, we continue to make this assumption. Some references to studies relaxing this assumption are, however, provided later.

Several Biallelic Loci

If data on $L(> 1)$ biallelic loci are available, then one can obtain estimates $m_1^{(l)}$ and $\mathrm{var}(m_1^{(l)})$ for the lth locus ($l = 1, 2, \ldots, L$) using (4) and (5). Cavalli-Sforza & Bodmer [3] suggest the following combined estimate:

$$\overline{m_1} = \frac{\sum_{l=1}^{L} \left[m_1^{(l)} \Big/ \mathrm{var}\left(m_1^{(l)}\right) \right]}{\sum_{l=1}^{L} \left[1 \Big/ \mathrm{var}\left(m_1^{(l)}\right) \right]}, \tag{6}$$

with

$$\mathrm{var}(\overline{m_1}) = \frac{1}{\sum_{i=1}^{L} \left[1 \Big/ \mathrm{var}\left(m_1^{(l)}\right) \right]}. \tag{7}$$

Multiallelic Loci

As in (1), a single allele provides a single equation. Thus, if $P > 2$, estimation of the μ_is is not possible without generalizing the model. Roberts & Hiorns [17] provide the generalization. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P) = ((x_{ij}))_{k\times P}$, where x_{ij} denotes the estimated frequency of the ith allele $(i = 1, 2, \ldots, k)$ at a codominant locus in the jth parental population $(j = 1, 2, \ldots, P)$. Let $\mathbf{y}_{k\times 1} = (y_1, y_2, \ldots, y_k)'$ denote the estimated allele frequencies in the hybrid population formed by admixture of the P parental populations in proportions $\boldsymbol{\mu}_{P\times 1} = (\mu_1, \mu_2, \ldots, \mu_P)'$. Then, if the x_{ij}s are known without error (that is, are actually known values of the corresponding population parameters), then

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\mu}. \tag{8}$$

Hence, the ordinary **least squares** estimate, $\mathbf{m}$, of $\boldsymbol{\mu}$ is:

$$\mathbf{m} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{9}$$

Unfortunately, because $\sum_{i=1}^{k} x_{ij} = 1$, for all $j = 1, 2, \ldots, P$, $\mathbf{X}'\mathbf{X}$ is a singular matrix. Roberts & Hiorns [17, 18] suggest eliminating data on one allele chosen arbitrarily, so that $\mathbf{X}'\mathbf{X}$ becomes nonsingular. However, in that case, the estimates of the admixture proportions become dependent on which allele is eliminated.

Elston [7] instead suggests a generalized least squares solution:

$$\mathbf{m} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{10}$$

where $\mathbf{V}$ is the dispersion matrix of $\mathbf{y}$, which is the dispersion matrix of the underlying **multinomial distribution**. Of course, because of the constraint that $\sum_{i=1}^{k} y_i = \sum_{i=1}^{k} x_{ij} = 1$, $\mathbf{V}$ is singular. Therefore, as suggested by Roberts & Hiorns [17, 18], in the present case too, data on one allele need to be eliminated arbitrarily to ensure nonsingularity of $\mathbf{V}$ and $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$. However, the advantage in using the generalized least squares approach is that the estimate $\mathbf{m}$ remains the same irrespective of the allele that is eliminated [7].

In addition to the problem of singularity, unconstrained least squares estimation (ordinary or generalized) does not guarantee $m_j \geq 0$, for all $j = 1, 2, \ldots, P$, and $\sum_{j=1}^{P} m_i = 1$. Elston [7] shows that the estimates

$$\mathbf{m}^* = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{y}^*, \qquad m_P = 1 - \sum_{j=1}^{P-1} m_j, \tag{11}$$

where $\mathbf{y}^*_{k\times 1} = \mathbf{y} - \mathbf{x}_P$, $\mathbf{X}^*_{k\times(P-1)}$ is the matrix whose jth column is $\mathbf{x}_j - \mathbf{x}_P$ and $\mathbf{m}^*_{(P-1)\times 1} = (m_1, m_2, \ldots, m_{P-1})'$, satisfy $\sum_{j=1}^{P} m_j = 1$. However, there is still no guarantee that $m_j > 0$, for all $j = 1, 2, \ldots, P$. Elston [7] suggests that, in practice, if any of the m_js is <0, the $\mathbf{m}$ should be recomputed with the smallest m_j set equal to 0.

The procedure for **maximum likelihood** estimation of $\boldsymbol{\mu}$ was suggested by Krieger et al. [10] and improved upon by Elston [7] to accommodate the constraint $\sum_{j=1}^{P} \mu_j = 1$. Under this model (Eq. (8)), the **likelihood** L that n_h alleles of type h are observed in a random sample of $2n$ alleles (that is, n individuals) from the hybrid population is

$$L \propto \prod_{h=1}^{k} (\mathbf{x}_h'\boldsymbol{\mu})^{n_h}, \tag{12}$$

since the n_hs follow a multinomial distribution with parameters $(\mathrm{E}(\mathbf{y});\ 2n)$. In practice, the likelihood is numerically maximized to obtain the mle of $\boldsymbol{\mu}$ (see [7] for details).

Generalizations of these procedures when allele frequency data on several loci are available are straightforward under the assumption that the allele frequencies are known without error. However, when we take into account fluctuations due to sampling a finite number of individuals to estimate allele frequencies in parental and hybrid populations, the procedure for estimating admixture proportions becomes statistically more complicated. Long & Smouse [13] have proposed a maximum likelihood estimation procedure taking sampling fluctuations into account. In addition to contemporary sampling fluctuations, there is another source of stochastic fluctuation of allele frequencies in each generation due to finiteness of population sizes (*genetic drift*). The effect of genetic drift is more difficult to take into account. Some attempts, albeit not completely satisfactory, have been made in this direction, notably by Thompson [19] and Wijsman [20].

Estimating Admixture from Genetic Similarities

Pollitzer [16] first formalized the intuitive notion that when there is differential gene flow from parental populations to a hybrid population, there will be a direct relationship between the amounts of gene flow and genetic similarities of the hybrid population from the parental populations. Thus, in principle, it should be possible to estimate admixture proportions from observed genetic similarity (or **genetic distance**) values. The initial statistical attempt made by Balakrishnan [1] has come under strong criticism [9, 6]. Chakraborty [4, 5] proposes an estimation procedure based on Nei's [14] gene identity coefficient. This coefficient, J_{ij}, is defined as the probability that two genes drawn at random, one from the ith population and the other from the jth population, are identical ($i, j = H, 1, 2, \ldots, P$). A **consistent estimator** of J_{ij} is [16]:

$$\hat{J}_{ij} = \frac{\sum_{l=1}^{L} \sum_{k=1}^{r_l} x_{ikl} x_{jkl}}{L}, \tag{13}$$

where x_{ikl} is the frequency of the kth allele at the lth locus in the ith population; r_l is the number of alleles at the lth locus; and L is the total number of randomly selected loci from the genome for which data are available. Chakraborty [5] shows that

$$\mathrm{E}(\mathbf{D}_{\mathrm{H}}) = \mathbf{D}\boldsymbol{\mu}^*, \tag{14}$$

where $\mathbf{D}_{\mathrm{H}} = (J_{1\mathrm{H}} - J_{1P}, J_{2\mathrm{H}} - J_{2P}, \ldots, J_{P\mathrm{H}} - J_{PP})'$, $\mathbf{D} = ((J_{ij} - J_{iP}))_{P\times(P-1)}$, $\boldsymbol{\mu}^* = (\mu_1, \mu_2, \ldots, \mu_{(P-1)})'$. Hence, the ordinary least squares estimator of $\boldsymbol{\mu}^*$ is:

$$\mathbf{m}^* = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{D}_{\mathrm{H}}. \tag{15}$$

While nonsingularity of $\mathbf{D}'\mathbf{D}$ is assured in this approach, there is no guarantee that the m_js will be nonnegative. When negative estimates are encountered, the strategy mentioned earlier may be adopted. By considering the dispersion matrix of $\mathbf{D}_{\mathrm{H}}$, evaluated empirically using formulas given in [15] and [12], generalized least squares estimators of $\boldsymbol{\mu}^*$ are obtained in a straightforward way.

Dynamics of Admixture

In the formulations above we have implicitly assumed that the admixture is a static, one-time phenomenon. In reality, there is usually a continued, long-term exchange of genes among populations. For example, as we mentioned in the beginning, migration of Africans to North America continued for more than a century. The total period of slave trading was about three centuries (from about 1650 to 1950), although the trading predominantly took place in the eighteenth century. Because of interbreeding with Whites, in each generation a fraction of the gene pool of African Blacks brought to North America was replaced by genes of Whites. Let b_0 denote the original frequency of an allele in Black Africans prior to any interbreeding with Whites, w denote the frequency of this allele among American Whites and μ denote the fraction of the gene pool of Blacks replaced per generation by genes of Whites. (Both w and μ are assumed to remain constant over generations.) Then, as in (2),

$$b_i = (1 - \mu)b_{i-1} + \mu w, \tag{16}$$

where b_i denotes the frequency of the allele among Blacks after i generations of interbreeding.

Thus, from (16), we obtain:

$$(1 - \mu)^i = \frac{b_i - w}{b_0 - w}. \tag{17}$$

Using (17), Glass & Li [8] estimated that the accumulated amount of White genes in the American Black gene pool was 30.6% ($\hat{\mu} = 0.0358$) based on the R^o allele frequency of the Rh **blood group** system and $i = 10$ generations (period of admixture $\approx$300 years; generation time $\approx$30 years). (This estimate has subsequently been questioned and revised. See [6] for a discussion, further references, examples and a detailed review.)

Data on frequencies of "private" alleles (i.e. unique alleles present in one population but not in others) are very helpful for admixture estimation. However, estimation procedures using such data need modifications from those presented above [11].

A Caveat

One of the major impediments to admixture estimation in humans has been the lack of accurate identities of parental populations of a hybrid population. All

the estimators presented above are quite sensitive to fluctuations of parental allele frequencies. Therefore, unless identities of, and allele frequencies in, parental populations are known accurately, estimates of admixture may be quite unreliable.

References

[1] Balakrishnan, V. (1973). Use of distance in hybrid analysis, in *Genetic Structure of Populations*, N.E. Morton, ed. University of Hawaii Press, Honolulu, pp. 268–273.

[2] Bernstein, F. (1931). Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung, in *Comitato Italiano per lo Studio dei Problemi della Populazione*. Instituto Poligrafico dello Stato, Roma, pp. 227–243.

[3] Cavalli-Sforza, L.L. & Bodmer, W.F. (1971). *The Genetics of Human Populations*. Freeman, San Francisco.

[4] Chakraborty, R. (1975). Estimation of race admixture-A new method, *American Journal of Physical Anthropology* **42**, 507–511.

[5] Chakraborty, R. (1985). Gene identity in racial hybrids and estimation of admixture rates, in *Genetic Microdifferentiation in Man and Other Animals*, J.V. Neel & Y.R. Ahuja, eds. Indian Anthropological Association, Delhi, pp. 171–180.

[6] Chakraborty, R. (1986). Gene admixture in human populations: Models and predictions, *Yearbook of Physical Anthropology* **29**, 1–43.

[7] Elston, R.C. (1971). The estimation of admixture in racial hybrids, *Annals of Human Genetics* **35**, 9–17.

[8] Glass, B. & Li, C.C. (1953). The dynamics of racial admixture-An analysis based on the American Negro, *American Journal of Human Genetics* **5**, 1–19.

[9] Korey, K.A. (1978). A critical appraisal of methods for measuring admixture, *Human Biology* **50**, 343–360.

[10] Krieger, H., Morton, N.E., Mi, M.P., Azevado, E.S., Freire-Maia, A. & Yasuda, N. (1965). Racial admixture in North-Eastern Brazil, *Annals of Human Genetics* **29**, 113–125.

[11] Li, Z. (1995). A multiplicative random effects model for meta-analysis with application to estimation of admixture component, *Biometrics* **51**, 864–873.

[12] Li, W.-H. & Nei, M. (1975). Drift variances of heterozygosity and genetic distance, *Genetical Research* **25**, 229–248.

[13] Long, J.C. & Smouse, P.E. (1983). Intertribal gene flow between the Ye'cuana and Yanomana: genetic analysis of an admixed village, *American Journal of Physical Anthropology* **61**, 411–422.

[14] Nei, M. (1972). Genetic distance between populations, *American Naturalist* **106**, 283–292.

[15] Nei, M. & Roychoudhury, A.K. (1974). Sampling variances of heterozygosity and genetic distance, *Genetics* **76**, 379–390.

[16] Pollitzer, W.S. (1964). Analysis of a triracial hybrid, *Human Biology* **36**, 362–373.

[17] Roberts, D.F. & Hiorns, R.W. (1962). The dynamics of racial admixture, *American Journal of Human Genetics* **14**, 261–277.

[18] Roberts, D.F. & Hiorns, R.W. (1965). Methods of analysis of the genetic composition of a hybrid population. *Human Biology* **37**, 38–43.

[19] Thompson, E.A. (1973). The Icelandic admixture problem, *Annals of Human Genetics* **37**, 69–80.

[20] Wijsman, E.M. (1984). Techniques for estimating genetic admixture and applications to the problems of the origin of the Icelanders and Ashkenazi Jews, *Human Genetics* **67**, 441–448.

PARTHA P. MAJUMDER

Admixture Mapping

The mapping of genes that control a trait or a disease is usually carried out by **linkage analysis** of family data on the trait/disease and **markers** [6]. However, there are other ways of mapping genes. One of these is by studying hybrid populations, that is populations that have arisen as a result of the admixture of two or more populations. The suggestion that hybrid populations can provide useful information about **linkage** was first pointed out by Rife [7]. Linkage between two loci results in the nonrandom **association** of alleles at the loci [8]. When two diallelic loci with alleles (A_1 and A_2) and (B_1 and B_2), respectively, are unlinked, then the frequency (f_{ij}) of the gamete A_iB_j ($i, j = 1, 2$) in the population will be

$$f_{ij} = p_i q_j,$$

where p_i and q_j denote, respectively, the frequencies of the alleles A_i and B_j in the population. If $D = f_{11}f_{22} - f_{12}f_{21}$ is used as a measure of the association of alleles at the two loci in gametes, which is known as **linkage disequilibrium**, then D is expected to be zero if the two loci are unlinked. However, if the two loci are linked and if θ denotes the recombination fraction between the two loci, then in a random-mating population,

$$D^{(t)} = (1-\theta)^t D^{(0)}, \tag{1}$$

where $D^{(0)}$ denotes the initial value of linkage disequilibrium in the population and $D^{(t)}$ denotes the value of linkage disequilibrium in the population after t generations of random mating [8]. It is clear from (1) that linkage disequilibrium will decay rapidly unless the two loci are closely linked, that is unless θ is close to zero.

The admixture between populations also leads to linkage disequilibrium between loci [1, 5]. Consider a population (Z) that is formed by the admixture of two populations X and Y, in proportions m ($\neq 0$) and $(1-m)$, respectively. Assume that no further admixture has taken place in the population Z. Then,

$$D_Z^{(t)} = (1-\theta)^t D_Z^{(0)}, \tag{2}$$

where $D_Z^{(t)}$ is the linkage disequilibrium in the admixed population Z in generation t and $D_Z^{(0)}$ is the linkage disequilibrium in population Z immediately after its formation by the admixture of populations X and Y, which is given by [2]

$$D_Z^{(0)} = mD_X^{(0)} + (1-m)D_Y^{(0)} + m(1-m) \times [p_1^{(x)} - p_1^{(Y)}][q_1^{(x)} - q_1^{(Y)}], \tag{3}$$

where $D_X^{(0)}$ and $D_Y^{(0)}$ are the initial linkage disequilibria in populations X and Y, respectively, $p_1^{(x)}$ and $q_1^{(x)}$ are the frequencies of alleles A_1 and A_2, respectively, in population X, and $p_2^{(Y)}$ and $q_2^{(Y)}$ are these frequencies in population Y. Equation (3) shows that even if the parental populations X and Y are in linkage equilibrium, if the allele frequencies in the parental populations X and Y are different, then there will be linkage disequilibrium in the admixed population. If the values of m, t, $[p_1^{(x)} - p_1^{(Y)}]$ and $[q_1^{(X)} - q_1^{(Y)}]$ are known, then θ can be estimated from (2) and (3). Appropriate likelihood ratio tests have been formulated [2] to test whether an observed value of linkage disequilibrium in the admixed population is due to linkage or due to admixture.

As was pointed out by Rife [7], in situations where the parental populations are of opposite homozygous **genotypes** ($A_1A_1B_1B_1 \times A_2A_2B_2B_2$ or $A_1A_1B_2B_2 \times A_2A_2B_1B_1$), then linkage from data from the admixed population may be detected in much the same way as from data on experimental crosses. This idea has been further extended. Since these extensions are similar to linkage analysis in experimental crosses, the term "admixture mapping" has been proposed [10].

With information about the ancestry of alleles at marker loci in individuals in the admixed population, one approach [3, 4] has been to test for association with states of ancestry on chromosomes drawn from the admixed population, conditioning on parental admixture. At each locus on each gamete there are two possible states of ancestry: X or Y. It has been shown [3] that from **Mendel's law** of independent assortment, conditional on parental admixture, there is no association between the ancestry of alleles at different loci if the loci are unlinked. Thus, the null hypothesis of no linkage is that, conditional on parental admixture, the odds ratio for the association between states of ancestry at any two loci on the same gamete is 1. Rejection of the **null hypothesis** is an indication of linkage between the loci under consideration. Using Bayesian statistical methodology, McKeigue et al. [4] have shown that the posterior distribution of parental admixture and ancestry at each marker locus, conditional on the observed

marker genotype data, can be generated by **Markov chain** simulation. They have also provided a statistical method of detecting linkage and of detecting misspecification of ancestry-specific allele frequencies within the admixed population under study.

A second approach [10] is based on an extension of the transmission-disequilibrium test (TDT) [9] (*see* **Disease-marker Association**). However, it should be noted that the usual TDT is not a special case of the multipoint TDT proposed [10] for this purpose. In this approach, a statistical method has been developed by conditioning on the ordered multilocus genotypes of parents and testing for association of marker **haplotypes** with the trait/disease under study. One major difference between this approach and the approach described in the previous paragraph is that the conditioning in this approach is on parental genotype, while in the other approach it is by parental admixture. Another major difference is the nature of data. While in McKeigue et al.'s [4] approach genotype data on parents are not essential, Zheng & Elston's [10] approach relies on nuclear family data. In both approaches, association arises only in the presence of linkage. Modeling in the second approach [10] is done through frequency distributions of haplotypes in the admixed population as a function of population history. Thus, probabilities of haplotypes on specific trait allele-carrying gametes that are derived from each of the two parental populations are calculated taking into account various modalities of multilocus recombination and the history of admixture. Using these probabilities, a "multipoint TDT" is derived. Unlike the usual TDT [9], in the present approach, what is scored is whether a *haplotype* in a parent is a transmitted one or a nontransmitted one to an offspring. The data therefore comprise counts of transmitted and nontransmitted haplotypes, the **likelihood** of which have been derived under various assumptions and scenarios [10]. A permutation test has been proposed as a test of significance [10].

While both approaches described above seem to perform well in a statistical sense, it remains unclear which of these two approaches is statistically more efficient.

References

[1] Cavalli-Sforza, L.L. & Bodmer, W.F. (1971). *Genetics of Human Populations*. Freeman, San Francisco.

[2] Chakraborty, R. & Weiss, K.M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci, *Proceedings of the National Academy of Sciences* **85**, 9119–9123.

[3] McKeigue, P.M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations by conditioning on parental admixture, *American Journal of Human Genetics* **63**, 241–251.

[4] McKeigue, P.M., Carpenter, J.R., Parra, E.J. & Shriver, M.D. (2000). Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations, *Annals of Human Genetics* **64**, 171–186.

[5] Nei, M. & Li, W.-H. (1973). Linkage disequilibrium in subdivided populations, *Genetics* **75**, 213–219.

[6] Ott, J. (1991). *Analysis of Human Genetic Linkage*. The Johns Hopkins University Press, Baltimore.

[7] Rife, D.C. (1954). Populations of hybrid origin as source material for the detection of linkage, *American Journal of Human Genetics* **6**, 26–33.

[8] Robbins, R.B. (1918). Some applications of mathematics to breeding problems. III, *Genetics* **3**, 375–389.

[9] Spielman, R.S. & Ewens, W.J. (1996). The TDT and other family-based tests for linkage disequilibrium and association, *American Journal of Human Genetics* **59**, 983–989.

[10] Zheng, C. & Elston, R.C. (1999). Multipoint linkage disequilibrium mapping with particular reference to the African-American population, *Genetic Epidemiology* **17**, 79–101.

(*See also* **Admixture in Human Populations**; **Population Genetics**)

PARTHA P. MAJUMDER

Adoption Studies

Adoption usually refers to the rearing of a nonbiological child in a family. This practice is commonplace after wars, which leave many children orphaned, and is moderately frequent in peacetime. Approximately 2% of US citizens are adoptees.

Historically, adoption studies have played a prominent role in the assessment of genetic variation in human and animal traits [10]. Most early studies focused on cognitive abilities [9], but there is now greater emphasis on psychopathology [5] and physical characteristics, such as body mass [20]. Adoption studies have made major substantive contributions to these areas, identifying the effects of genetic factors where they were previously thought to be absent [3, 12, 15].

In recent years the adoption study has been overshadowed by the much more popular twin study [17] (*see* **Twin Analysis**). Part of this shift may be due to the convenience of twin studies and the complex ethical and legal issues involved in the ascertainment and sampling of adoptees. Certain Scandinavian countries – especially Denmark, Sweden and Finland [8, 13, 14] – maintain centralized databases of adoptions and thus have been able to mount more representative and larger adoption studies than elsewhere.

The adoption study is a "natural experiment" that mirrors cross-fostering designs used in genetic studies of animals, and therefore has a high face validity as a method to resolve the effects of genes and environment on individual differences. Unfortunately, the adoption study also has many methodological difficulties. First, is the need to maintain **confidentiality**, which can be a problem even at initial ascertainment, as some adoptees do not know that they are adopted. Recent legal battles for custody fought between biological and adoptive parents make this a more critical issue than ever. Secondly, in many substantive areas, e.g. psychopathology, there are problems with sampling, in that neither the biological nor the adoptive parents can be assumed to be a random sample of parents in the population. For example, poverty and its sequelae may be more common among biological parents who have their children adopted into other families than among parents who rear their children themselves. Conversely, prospective adoptive parents are, on average, and through self-selection, older and less fertile than biological parents. In addition, they are often carefully screened by adoption agencies, and may be of higher socio-economic status than nonadoptive parents. Statistical methods (see below) may be used to control for these sampling biases if a random sample of parents is available. Some studies indicate that adoptive and biological parents are quite representative of the general population for demographic characteristics and cognitive abilities [19], so this potential source of bias may not have substantially affected study results.

Thirdly, selective placement is a common methodological difficulty. For statistical purposes, the ideal adoption study would have randomly selected adoptees placed at random into randomly selected families in the population. Often there is a partial **matching** of the characteristics of the adoptee (e.g. hair and eye color, religion and ethnicity) to those of the adopting family. This common practice may improve the chances of successful adoption. Statistically, it is necessary to control for the matching as far as possible. Ideally, the matching characteristics used should be recorded and modeled. Usually, such detailed information is not available, so matching is assumed to be based on the variables being studied and modeled accordingly (see below). In modern adoption studies, these methods are used often [18, 19].

Types of Adoption Study

Nuclear families in which at least one member is not biologically related to the others offer a number of potential comparisons that can be genetically informative (see Table 1). Of special note are monozygotic (MZ) twins reared apart (MZ_A) (*see* **Zygosity Determination**). Placed into uncorrelated environments, the correlation between MZ twins directly estimates the proportion of variation due to all genetic sources of variance ("broad **heritability**"). Estimation of heritability in this way is statistically much more powerful than, e.g. the classical twin study that compares MZ and dizygotic (DZ) twins reared together (MZ_T and DZ_T). With MZ_A twins the test for heritability is a test of the null hypothesis that the correlation is zero, whereas the comparison of MZ_T and DZ_T is a test of a difference between correlations. Environmental effects shared by members of a twin pair (known as "common", "shared" or "family" environment or "C") are excluded by design. If this source of variation is of interest, then additional groups of relatives, such as unrelated individuals reared together,

Table 1 Coefficients of genetic and environmental variance components quantifying resemblance between adopted and biological relatives, assuming random sampling, mating and placement

Relationship	Variance component						
	V_A	V_D	V_{AA}	V_{AD}	V_{DD}	E_S	E_P
BP–BC	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	0	1
BP–AC	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	0	0
AP–AC	0	0	0	0	0	0	1
AC–BC	0	0	0	0	0	1	0
BC–BC_T	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	1	0
BC–BC_A	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	0	0
MZ_T–MZ_T	1	1	1	1	1	1	0
MZ_A–MZ_A	1	1	1	1	1	0	0

V_A – additive genetic; V_D – dominance genetic; V_{AA} – additive × additive interaction; V_{AD} – additive × dominance interaction; V_{DD} – dominance × dominance interaction; E_S – environment shared by siblings; E_P – environment shared or transmitted between parent and child. Relationships are: MZ – monozygotic twin; DZ – dizygotic twin; BP – biological parent; BC – biological child; AP – adoptive parent; AC – adopted child. The subscripts T and A refer to reared together and reared apart, respectively.

are needed to estimate it. Similar arguments may be made about across-generational sources of resemblance. Heath & Eaves [11] compared the power to detect genetic and environmental transmission across several twin-family (twins and their parents or twins and their children) adoption designs.

Methods of Analysis

Most modern adoption study data are analyzed with Structural Equation Models (SEM) [2, 17]. SEM is an extension of **multiple linear regression** analysis that involves two types of variable: *observed variables* that have been measured, and *latent variables* that have not. Two variables may be specified as causally related or simply correlated from unspecified effects. It is common practice to represent the variables and their relationships in a path diagram (*see* **Path Analysis in Genetics**), where single-headed arrows indicate causal relationships, and double-headed arrows represent correlations. By convention, observed variables are shown as squares and latent variables are shown as circles.

Figure 1 shows the genetic and environmental transmission from biological and adoptive parents to three children. Two of the children are offspring of the biological parents (sibs reared together) while the third is adopted. This diagram may also be considered as multivariate, allowing for the joint analysis of multiple traits. Each box and circle then represents a vector of observed variables. Multivariate analyses (*see* **Multivariate Analysis, Overview**) are particularly important when studying the relationship between parental attributes and outcomes in their offspring. For example, harsh parenting may lead to psychiatric disorders. Both variables should be studied in a multivariate genetically informative design such as an adoption or twin study to distinguish between the possible direct and indirect genetic and environmental pathways.

From the rules of path analysis [22, 23] we can derive predicted covariances among the relatives, in terms of the parameters of the model in Figure 1. These expectations may, in turn, be used in a structural equation modeling program such as Mx [16] to estimate the parameters using **maximum likelihood** or some other **goodness-of-fit** function. Often, simpler models than the one shown will be adequate to account for a particular set of data.

A special feature of the diagram in Figure 1 is the dotted lines representing delta-paths [21]. These represent the effects of two possible types of selection: assortative mating, in which husband and wife correlate; and selective placement, in which the adoptive and biological parents are not paired at random. The effects of these processes may be deduced from the Pearson–Aitken selection formulas [1]. These formulas are derived from **linear regression** under the assumptions of multivariate linearity and homoscedasticity. If we partition the variables into selected variables, X_S, and unselected variables X_N, then it can be shown that changes in the covariance of X_S lead to changes in covariances among X_N and the cross-covariances (X_S with X_N). Let the original (preselection) covariance matrix of X_S be **A**, the original covariance matrix of X_N be **C**, and the covariance between X_N and X_S be **B**. The preselection matrix may be written

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{B}' & \mathbf{C} \end{array}\right).$$

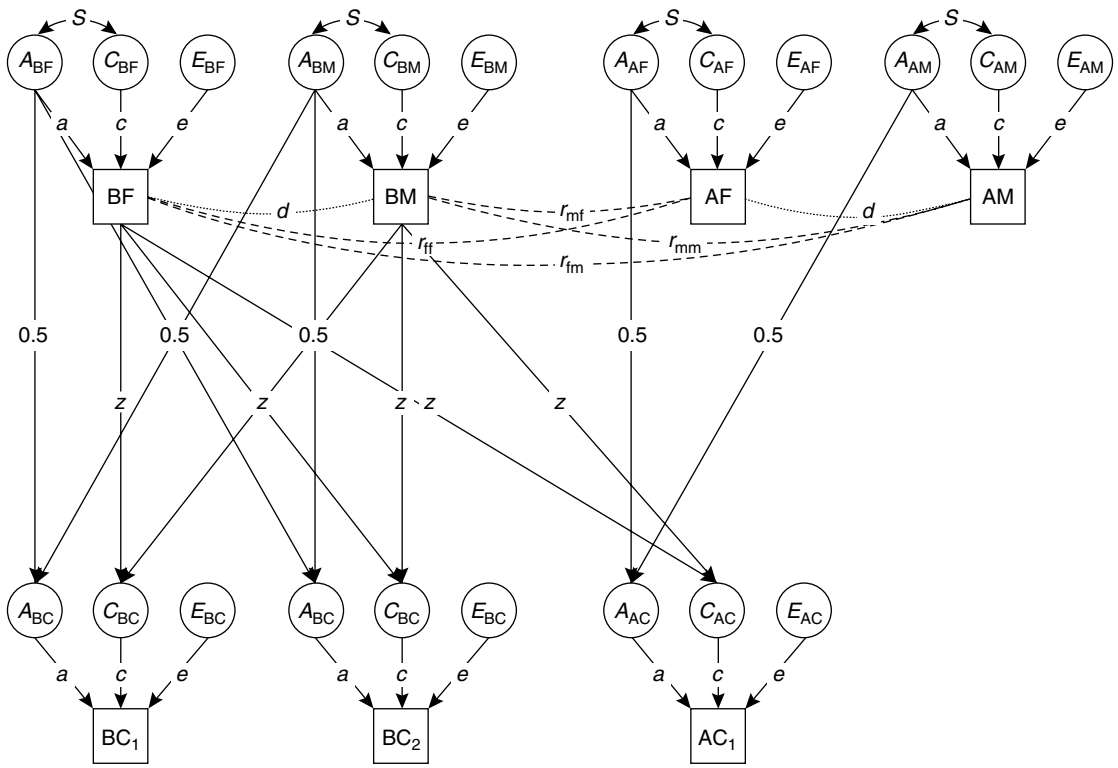

Figure 1 Path diagram showing sources of variation and covariation between: adoptive mother, AM; adoptive father, AF; their own biological children, BC_1 and BC_2; a child adopted into their family, AC_1; and the adopted child's biological parents, BF and BM

If selection transforms **A** to **D**, then the new covariance matrix is given by

$$\left(\begin{array}{c|c} \mathbf{D} & \mathbf{DA}^{-1}\mathbf{B} \\ \hline \mathbf{B}'\mathbf{A}^{-1}\mathbf{D} & \mathbf{C}-\mathbf{B}'(\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{DA}^{-1})\mathbf{B} \end{array}\right).$$

Similarly, if the original means are $(\mathbf{x}_s : \mathbf{x}_n)'$ and selection modifies $\mathbf{x}_s$ to $\tilde{\mathbf{x}}_s$, then the vector of means after selection is given by

$$[\mathbf{x}_s : \mathbf{x}_n + \mathbf{A}^{-1}\mathbf{B}(\mathbf{x}_s - \tilde{\mathbf{x}}_s)]'.$$

These formulas can be applied to the covariance structure of all the variables in Figure 1. First, the formulas are applied to derive the effects of assortative mating, and secondly, they are applied to derive the effects of selective placement. In both cases, only the covariances are affected, not the means. An interesting third possibility would be to control for the effects of nonrandom selection of the biological and adoptive relatives, which may well change both the means and the covariances.

Selected Samples

A common approach in adoption studies is to identify members of adoptive families who have a particular disorder, and then examine the rates of this disorder in their relatives (*see* **Ascertainment**). These rates are compared with those from control samples. Two common starting points for this type of study are (i) the adoptees (the adoptees' families method) and (ii) the biological parents (the adoptees study method). For rare disorders, this use of selected samples may be the only practical way to assess the impact of genetic and environmental factors.

One limitation of this type of method is that it focuses on one disorder, and is of limited use for examining **co-morbidity** between disorders. This

limitation is in contrast to the **population-based** sampling approach where many characteristics – and their covariances or co-morbidity – can be explored simultaneously.

A second methodological difficulty is that ascertained samples of the disordered adoptees or parents may not be representative of the population. For example, those attending a clinic may be more severe or have different risk factors than those in the general population who also meet criteria for diagnosis, but do not attend the clinic.

Genotype × Environment Interaction

The natural experiment of an adoption study provides a straightforward way to test for **gene–environment interaction**. In the case of a continuous phenotype, interaction may be detected with linear regression on

1. the mean of the biological parents' phenotypes (which directly estimates **heritability**)
2. the mean of the adoptive parents' phenotypes
3. the product of points 1 and 2.

Significance of the third term would indicate significant G × E interaction. With binary data such as psychiatric diagnoses, the rate in adoptees may be compared between subjects with biological or adoptive parents affected, vs. both affected. G × E interaction has been found for alcoholism [7] and substance abuse [6].

Logistic regression is a popular method to test for genetic and environmental effects and their interaction on **binary** outcomes such as psychiatric diagnoses. These analyses lack the precision that structural equation modeling can bring to testing and quantifying specific hypotheses, but offer a practical method of analysis for binary data. Analysis of binary data can be difficult within the framework of SEM, requiring either very large sample sizes for asymptotic weighted **least squares** [4] or integration of the **multivariate normal** distribution over as many dimensions as there are relatives in the pedigree, which is numerically intensive.

References

[1] Aitken, A.C. (1934). Note on selection from a multivariate normal population, *Proceedings of the Edinburgh Mathematical Society, Series B* **4**, 106–110.

[2] Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

[3] Bouchard, Jr., T.J. & McGue, M. (1981). Familial studies of intelligence: A review, *Science* **212**, 1055–1059.

[4] Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

[5] Cadoret, R.J. (1978). Psychopathology in adopted-away offspring of biologic parents with antisocial behavior, *Archives of General Psychiatry* **35**, 176–184.

[6] Cadoret, R.J., Troughton, E., O'Gorman, T.W. & Heywood, E. (1986). An adoption study of genetic and environmental factors in drug abuse, *Archives of General Psychiatry* **43**, 1131–1136.

[7] Cloninger, C.R., Bohman, M. & Sigvardsson, S. (1981). Inheritance of alcohol abuse: cross-fostering analysis of adopted men, *Archives of General Psychiatry* **38**, 861–868.

[8] Cloninger, C.R., Bohman, M., Sigvardsson, S. & von Knorring, A.L. (1985). Psychopathology in adopted-out children of alcoholics: the Stockholm adoption study, in *Recent Developments in Alcoholism*, Vol. 3, M. Galanter, ed. Plenum Press, New York, pp. 37–51.

[9] DeFries, J.C. & Plomin, R. (1978). Behavioral genetics, *Annual Review of Psychology* **29**, 473–515.

[10] Fuller, J.L. & Thompson, W.R. (1978). *Foundations of Behavior Genetics*. Mosby, St. Louis.

[11] Heath, A.C. & Eaves, L.J. (1985). Resolving the effects of phenotype and social background on mate selection, *Behavior Genetics* **15**, 15–30.

[12] Heston, L.L. (1966). Psychiatric disorders in foster home reared children of schizophrenic mothers, *British Journal of Psychiatry* **112**, 819–825.

[13] Kaprio, J., Koskenvuo, M. & Langinvainio, H. (1984). Finnish twins reared apart: smoking and drinking habits. Preliminary analysis of the effect of heredity and environment, *Acta Geneticae Medicae et Gemellologiae* **33**, 425–433.

[14] Kety, S.S. (1987). The significance of genetic factors in the etiology of schizophrenia: results from the national study of adoptees in Denmark, *Journal of Psychiatric Research* **21**, 423–429.

[15] Mendlewicz, J. & Rainer, J.D. (1977). Adoption study supporting genetic transmission in manic-depressive illness, *Nature* **268**, 327–329.

[16] Neale, M.C. (1995). *Mx: Statistical Modeling*, 3rd Ed. Box 980126 MCV, Richmond, VA 23298.

[17] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer, Boston.

[18] Phillips, K. & Fulker, D.W. (1989). Quantitative genetic analysis of longitudinal trends in adoption designs with application to IQ in the Colorado Adoption Project, *Behavior Genetics* **19**, 621–658.

[19] Plomin, R. & DeFries, J.C. (1990). *Behavioral Genetics: A Primer*, 2nd Ed. Freeman, New York.

[20] Sorensen, T.I. (1995). The genetics of obesity, *Metabolism*, **44**, 4–6.

[21] Van Eerdewegh, P. (1982). *Statistical Selection in Multivariate Systems with Applications in Quantitative Genetics*, PhD thesis. Washington University, St. Louis.

[22] Vogler, G.P. (1985). Multivariate path analysis of familial resemblance, *Genetic Epidemiology* **2**, 35–53.

[23] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.

M.C. NEALE

Adverse Selection

Some health care providers (e.g., clinicians, hospitals, or HMOs) and insurers serve populations that are substantially sicker or more difficult to care for than average. A provider or insurer with sicker than average patients experiences adverse selection; one with healthier than average patients has favorable selection. Action taken to achieve favorable selection is called "skimming" or "cream skimming".

Biased selection (either favorable or adverse is not necessarily a problem, if such differences are recognized and accounted for when paying for care or holding providers accountable for their patients' health outcomes. However, when providers are paid a fixed price for each patient, or they are penalized for expending more resources than their peers for patient care, or for not achieving as good outcomes, providers will compete to treat healthy patients and will be discouraged from caring for those with complex problems.

Further Reading

Cutler, D.M. & Zeckhauser, R.J. (2000). The Anatomy of Health Insurance Chapter 11 in *Newhouse Handbook of Health Economics*, A.J. Culyer and J.P., eds. Elsevier Science B.V. p. 563–643.

Mello, Michelle M., Stearns, Sally C., Norton Edward C. & Ricketts, III Thomas C. *Health Services Research*, June, 2003. Understanding biased selection in Medicare HMOs. (Brief Article).

ARLENE S. ASH

Age-of-onset Estimation

Age-of-onset estimation refers to the estimation of the distribution, as a function of age, of the time a trait first appears. Typically, the time of first occurrence is measured by the age at diagnosis. Many diseases exhibit variable age of onset where subjects carrying a susceptibility **gene** for a disease develop the disease at an earlier age. Consequently, interest is in estimating the age of disease onset as a function of **genotype** and characterizing how the distribution may vary across subpopulations, or in the presence of gene–gene or **gene–environment interactions**. Other names for the age-of-onset distribution include the age-specific **penetrance**, age-specific risk, cumulative risk, or cumulative incidence.

For many diseases, such as cancer, age is the primary risk factor; the risk of developing disease increases with age. However, not all subjects will develop a specific disease or trait in their lifetime. Therefore age of onset is studied among living subjects who are at risk for the trait of interest, in the presence of death from other causes. Observations are censored for participants on whom the event is not observed, either because the subjects are unaffected at time of observation or because they die without ever having developed the trait. In the literature this has been approached using two different modeling philosophies. One assumes the population is a mixture of susceptible and nonsusceptible individuals, nonsusceptible individuals being subjects who would never develop the trait no matter how long they lived [6, 7, 14]. This is the approach that we describe below. The other philosophy assumes all individuals are susceptible and would eventually become affected if they only lived long enough [1, 4, 13, 22, 24].

Let g denote genotype and x a vector of exposures. Then the cumulative risk at age a_1 for a carrier of genotype g can be expressed using the improper distribution function

$$F_g(a_1;x) = 1 - S_g(a_1;x) = \phi_g \left\{ 1 - \exp\left[- \int_0^{a_1} \lambda_g(a;x)\ \mathrm{d}a \right] \right\}, \quad (1)$$

where ϕ_g estimates the proportion of subjects who will develop the trait in their lifetime and $\lambda_g(a;x)$ is the hazard, the instantaneous probability of the trait developing among subjects at risk, given their genotype and exposures, as a function of age. For an individual who is unaffected at age a_0, the risk of developing the trait by age a_1 is

$$\pi(a_0, a_1; x) = \frac{\int_{a_0}^{a_1} h_g(a;x)\, S_g(a;x)\, S_c(a;x)\ \mathrm{d}a}{S_g(a_0;x)\, S_c(a_0;x)}, \quad (2)$$

where $h_g(a;x) = \partial/\partial a F_g(a;x)/S_g(a;x)$ and $S_c(a;x)$ is the probability of surviving up to age a due to causes of death unrelated to the trait of interest. For designing prevention studies or for the purpose of risk management in unaffected individuals, this latter quantity may be more relevant than the "pure" trait-specific cumulative risk given above.

The variability in age of onset as a function of other exposures, X, can be modeled through the hazard. The hazard is expressed in two parts: the hazard of disease onset in subjects with the baseline levels of exposure, $\lambda_g(t)$, and a term involving the level of exposure, x. Several models may be proposed. The most common is the proportional hazards model, $\lambda_g(t;x) = \lambda_g(t)RR(t;x)$, where the exposures have a multiplicative effect on the baseline hazard. The multiplicative term of the covariates is often called the hazard ratio or the **relative risk**. Alternatively, the effect on the hazard could be additive as in the additive hazard model, $\lambda_g(t;x) = \lambda_g(t) + RD(t;x)$, where $RD(t;x)$ is the risk difference. A third paradigm is an accelerated failure time model, $\lambda_g(t;x) = \lambda_g(t \cdot w(t;x))w(t;x)$, where the effect of the covariates is multiplicative on age.

Alternative formulations for the age-of-onset distribution use **regressive** logistic or regressive linear models. These approaches model the probability that an individual is affected at age a by $\phi_g w_g(a,x)$, and the probability that an individual is unaffected at age a by $1 - \phi_g W_g(a,x)$. The function w_g is the logistic or normal density and is modeled as a regressive function of age a and covariates x; W_g is its corresponding cumulative distribution function.

Uses

Estimating age of onset is important in studies of disease etiology. Allowing for variable age of onset in segregation analysis can help find evidence for new major genes and the use of age-of-onset estimates in **linkage analysis** can assist in the localization of

putative susceptibility genes. Additionally, estimates of the proportion of cases explained by a gene as a function of age of onset may guide the development of new studies for identifying other risk factors for disease. For example, the genes BRCA1 and BRCA2 explain a large proportion of breast cancer among early onset cases but do not explain the majority of cases among women with late onset disease. Therefore studies for identifying new risk factors may focus on cases with late onset breast cancer.

Accurate estimates of age of onset are also important for individual **genetic counseling** and for the planning of future prevention studies. A subject's age-specific cumulative risk for developing disease may influence the recommended age at which regular screening should begin or the length of time between visits. Also, it may influence preventive measures a subject may take. For a woman who has a low risk of developing breast cancer, regular surveillance may be her selected regimen. However, a woman having a high lifetime risk of breast cancer may opt for a prophylactic mastectomy. Accurate risk estimates and their variability will help a patient in determining her own course of action.

In the design of prevention trials, accurate estimates of age of disease onset are needed for estimating sample size. Another use, shown by a trial for studying the drug Tamoxifen as a chemopreventive agent for breast cancer, is for screening eligible study participants. Owing to the potential toxicities related to the administration of the drug to a healthy population, investigators enrolled subjects at high risk of breast cancer. Eligible women were either over the age of 59 or younger but having the same age-specific risk as an average 60-year-old woman [18].

Study Design

A careful study design is critical for obtaining good estimates of age-specific penetrance [7, 20]. First, the disease may be uncommon. For instance, among women in the US who live to age 85, one in nine (11%) will develop breast cancer [2]. Other cancers such as colon cancer occur less frequently. Four percent of the US population are expected to develop colon cancer in their lifetime. A second reason why a study design is important is that the genes that are strong risk factors for disease, BRCA1 and BRCA2 for breast cancer and MLH1 and MSH2 for colon cancer, are rare. Standard cohort and case–control studies would need prohibitively large sample sizes to study these genes. As a result, family studies have been implemented. Family studies have the advantage that they can be completed relatively quickly. Disease status on all family members can be obtained in a single interview and data for several diseases can be collected at the same time. However, a possible source of **bias** includes the differential participation rates of probands (*see* **Ascertainment**) based on **family history** of disease [7, 8]. A second source of bias is reporting error on the disease status of relatives. Although reporting error is not specific to family studies, it could occur at a higher frequency.

The following completed or ongoing family study designs are proposed for estimating age-specific cumulative risk.

Case–Control Family Study

In a **case–control** family study, affected study participants ("cases") are selected during a fixed time period from a population-based sample of newly identified diseased subjects. Unaffected study participants ("controls") are selected, generally by frequency matching to the cases based on age, sex, and geographic region of residence. Data are collected from the study participants on the history of disease in a specified group of relatives (e.g. first-degree relatives). Blood samples for genotyping may or may not be obtained from the participants or relatives. When genotypes on the participants are available, the design is named the genotyped–proband design [7].

Studies that have published estimates of age-specific penetrance of breast cancer from a case–control family study include a study of breast cancer in the Cancer and Steroid Hormone Study (CASH) [3] and a study of three US case–control studies with cases selected for ovarian cancer [23]. Additional estimates using data from breast cancer (case-only) families were reported from a population-based study in Australia [10].

Kin–Cohort Design

In the kin–cohort design, study participants are volunteers. They provide a blood sample and complete a short, self-administered questionnaire reporting on risk factors and family history of disease in first- and second-degree relatives. The relatives ("kin")

form a **cohort** from which age-specific cumulative risk is estimated. A **kin–cohort study** of Ashkenazi Jews living in Washington DC provided age-specific cumulative risk estimates for three BRCA1 **mutations** [19].

Multistage Sampling

Multistage sampling designs have been proposed for estimating the penetrance of rare genes in families. In order to increase the frequency of the gene variants in the sample, investigators have proposed to oversample probands with a positive family history of disease. In stage 1, participants are stratified based on their family history of disease in first-degree relatives. Subjects are then randomly sampled, conditional on their family history. The University of Southern California (USC) Consortium Colorectal Cancer Family Registry applied a multistage design in developing a registry of colorectal cancer cases and their families. They select all cases with a positive family history of disease and 16% of cases with a negative family history.

High-risk Families

Samples of high-risk families that were originally collected for the purpose of linkage analysis have also been used to characterize age of disease onset. Such families, generally collected in clinics, have a large number of affected relatives and variable family structure. In addition to the usual trait information on relatives, **marker** genotypes are measured on a large number of family members. The Breast Cancer Linkage Consortium has used a collection of families having multiple cases of breast and/or ovarian cancer to characterize age of disease onset. One analysis focused on families with four or more cases of either breast or ovarian cancer that reported linkage to marker D17S579, a genetic marker $\sim$2 cM distal to BRCA1 [5].

Estimation Methods

Many approaches have been developed for estimating age-specific penetrance from family studies. These include **likelihood** [1, 6, 7, 9, 14], **pseudo-likelihood** [13], **marginal likelihood** [4], weighted score [17, 24], and method of moments [22] approaches. We introduce the different methods for each of the above-mentioned designs. We begin by describing the likelihood for the data from a population-based case–control family study.

Case–Control Family Study. Let g_0 denote the proband's disease susceptibility genotype and $y_0 = (a_0, \delta_0)$ the phenotype, where a_0 is age at diagnosis if affected or current age if unaffected, and δ_0 is disease status (1 = affected, 0 = unaffected). The phenotypes and susceptibility genotypes of J relatives are given by $\mathbf{y}_1 = (y_{11}, \ldots, y_{1J})$ and $\mathbf{g}_1 = (g_{11}, \ldots, g_{1J})$. For relatives who died without ever developing the trait, a_{1j} is their age at death. The likelihood is conditioned on the disease status of the proband to adjust for their selection conditional on disease status. Assuming that phenotypes are conditionally independent within a family given genotypes, the likelihood contribution from a single family is

$$L(\boldsymbol{\Omega}, q; \mathbf{y}_1 | y_0) = \frac{\sum_{g_0, \mathbf{g}_1} p(g_0, \mathbf{g}_1; q)\, f(y_0 | g_0; \boldsymbol{\Omega}) \prod_{j=1}^{J} f\left(y_{1j} | g_{1j}; \boldsymbol{\Omega}\right)}{\sum_{g_0} p(g_0; q)\, f(y_0 | g_0; \boldsymbol{\Omega})}, \tag{3}$$

where the parameter q denotes the susceptibility allele frequency, $p(g_0, \mathbf{g}_1; q)$ is the probability of the unobserved family genotypes, and $f(y|g; \boldsymbol{\Omega})$ is the density function characterized by the age of onset parameters $\boldsymbol{\Omega}$. This density can be from a logistic function [6, 14]. Alternatively, and the case we describe in more detail below, it can be based on a hazard function. The genotype probabilities are computed under the usual assumptions of **Hardy–Weinberg equilibrium**, random mating, and a known mode of inheritance. To incorporate measured susceptibility genotypes on relatives into a joint likelihood, the observed genotypes are removed from the sum in the numerator of (3).

Both parametric and semiparametric models for the hazard function have been proposed. Gail et al. [7] parameterize the cumulative risk function using an improper Weibull function, $F_g(a) = \phi_g\{1 - \exp[-(\gamma_g a)^{\alpha_g}]\}$. Assuming that censoring is unrelated to genotype, they write the density as

$$f_g(a) = \lambda_g(a)^{\delta} S_g(a) G(a), \tag{4}$$

where $\lambda_g(a) = \phi_g \alpha_g \gamma_g^{\alpha_g} a^{\alpha_g - 1} \exp[-(\gamma_g a)^{\alpha_g}]/S_g(a)$ and $G(a)$ is the probability of surviving all causes of death unrelated to the disease of interest. Since $G(a)$ does not depend on genotype, it does not affect the penetrance estimates. For a semiparametric modeling approach, Moore et al. [13] propose a piecewise exponential survival model. A common set of age cut-points is selected and age intervals created for estimating separate hazard functions in gene carriers and noncarriers, e.g. less than 30 years old, 30–39 years, 40–49 years, 50–59 years, 60–69 years, 70–79 years, 80 or more years. Then they estimate a constant hazard function for each separate age interval.

For relatively simple family structures and parametric models, the loglikelihood can be directly maximized or an **EM algorithm** approach may be used. For more complex structures a **Monte Carlo** EM method may be preferred [12]. Moore et al. [13] present a pseudolikelihood alternative for estimating the age-specific risk estimates for their semiparametric model. All these approaches are analogous to estimating penetrance parameters in a **segregation analysis**. When measured genotypes are incorporated into the likelihood, it may be referred to as a modified segregation analysis. These methods assume conditional independence of disease given genotype which, if violated, can lead to biased penetrance estimates. Alternative methods that allow for residual familial correlation have been developed for randomly sampled families. These are described below for the kin–cohort design.

Kin–Cohort Design. Wacholder et al. [22] proposed a method of moments approach to estimate age of onset for the kin–cohort design. Using Kaplan–Meier methods they estimated the proportion of kin who develop disease separately for carrier participants and noncarrier participants. Noting that relatives form a mixture of carriers and noncarriers, they presented a method to decompose the risk estimates from kin into its genotype-specific parts. Under dominant gene action, the first-degree relatives of participants who carry a disease mutation are approximately a 50 : 50 mixture of carriers and noncarriers. For first-degree relatives of noncarrier participants the mixture is $\pi : (1 - \pi)$, where π is the population frequency of carrying the variant gene. Similar to the likelihood-based methods, this analysis assumes a known mode of inheritance of disease, a constant **gene frequency** across age (i.e. censoring due to the competing risk of death from an unrelated cause is not a function of the genotype under study), and homogeneity of risk given the genetic variant of interest. The primary critique of the analysis is that the estimates of cumulative risk from the use of Kaplan–Meier methods can be nonmonotonic. Whenever a noncarrier kin has an event at a time that the carrier kin does not, the cumulative risk will decrease. Also, it does not utilize the disease status of the proband.

These criticisms can be overcome by a likelihood-based analysis. Risk estimates derived from likelihood-based approaches are always monotonically increasing. Furthermore, the likelihood includes information on the proband's disease status. In the kin–cohort design, probands are volunteers from the population, so it is not necessary to condition (3) on the proband's phenotype. However, an added assumption is necessary, namely that there is no differential survival of cases by genotype before the time that they are selected into the study [7]. Gail et al. [7] showed that the kin–cohort design allows a modest reduction in necessary sample sizes compared with standard epidemiologic cohort or case–control designs. Larger reductions are possible if additional genotypes from relatives can be obtained [7].

Other strengths of a likelihood approach are that it allows for additional modeling of **covariates** and also allows the likelihood to be extended to include the disease status of more distant relatives. However, these estimates can suffer from their own **biases** under model misspecification. As mentioned earlier, **maximum likelihood** relies on a conditional independence assumption of disease given genotype and covariates [7] which, if violated, can result in biased risk estimates. To allow for the residual familial aggregation of disease arising from shared genetic and/or environmental effects, a **marginal likelihood** approach has been proposed [4]. Alternatively, one could add a family-specific random effect [12, 16]. Currently, these two approaches are limited to designs where the study participants are randomly sampled from a general population.

Multistage Sampling. A weighted score approach has been proposed for estimating risk from multistage sampled families [17, 24]. For a two-stage design, the analysis weights the score contribution of the data in stage 2 by the inverse of the proportion of

participants sampled in that stratum. Specifically, if S_k denotes stratum k $(k = 1, 2)$, then the weighted score equation is

$$U_W(\mathbf{\Omega}) = \sum_{k=1}^{2} \frac{1}{f_{S_k}} \sum_{y \in \text{stage } 2} \frac{\partial}{\partial \mathbf{\Omega}} \ln L(\mathbf{\Omega}, q; \mathbf{y}_1, g_0 | y_0), \tag{5}$$

where f_{S_k} denotes the fraction sampled in stratum S_k. In using this approach to design the USC Colorectal Cancer Family Registry, Siegmund et al. [17] found that over-sampling probands with a positive family history of disease could improve the efficiency of the penetrance estimates in gene carriers.

High-risk Families. The retrospective likelihood is used for estimating penetrance when families are sampled based on the occurrence of multiply affected individuals. This approach is also known as maximizing the lod score, or the mod score approach [9]. In this likelihood, marker genotypes (**m**) are measured on a large number of relatives, and their distance from the susceptibility locus is modeled using a recombination fraction (θ). The contribution to the retrospective likelihood from a single family is proportional to the probability of the family marker genotypes given all phenotypes:

$$L(\mathbf{\Omega}, q, \theta; \mathbf{m}|\mathbf{y}) \propto \frac{\sum_{\mathbf{g}} f(\mathbf{y}|\mathbf{g}; \mathbf{\Omega})\, p(\mathbf{g}|\mathbf{m}; q, \theta)}{\sum_{\mathbf{g}} f(\mathbf{y}|\mathbf{g}; \mathbf{\Omega})\, p(\mathbf{g}; q)}, \tag{6}$$

where $p(\mathbf{g}|\mathbf{m}; q, \theta)$ is the probability of unobserved susceptibility genotypes in the family given the observed marker genotypes as a function of the susceptibility allele frequency q and the distance between the marker and the susceptibility locus, θ. Random mating and Hardy–Weinberg equilibrium are assumed. Parameters are estimated using maximum likelihood and the same conditional independence of phenotype given genotype assumptions are needed. Vieland & Hodge [21] point out that for general pedigree structures the retrospective likelihood does not provide an exact ascertainment correction. However, the adjustment based on phenotype is believed to yield, in general, essentially unbiased penetrance estimates.

For a discussion of bias and efficiency based on different likelihood approaches, see Kraft & Thomas [11].

Populations

The experience of estimating the age-of-onset distribution for breast cancer shows the large variability that can exist across study populations. Early estimates from a large population-based case–control study in the US found a 67% risk of breast cancer by age 70 for women who carried a susceptibility gene [3]. Cases were young women, aged 20–54, with newly diagnosed breast cancer. After the cloning of BRCA1 identified that the gene was a risk factor in families carrying excess cases of breast and ovarian cancer, a similar study was undertaken using incident ovarian cancer cases. That study reported a 69% risk of breast cancer by age 70, supporting the earlier estimates [23]. Neither of these studies measured specific genetic variants in their samples.

In studies that have measured specific BRCA1 variants or a linked marker locus, a range of risk estimates for carriers has emerged. The analysis of high-risk families identified by the Breast Cancer Linkage Consortium reported an 85% risk of breast cancer (confidence interval 51%–91%) by age 70 [5]. The kin–cohort study of Ashkenazi Jews in Washington DC estimated a risk of 56% [19]; a population-based sample of Australian women selected for having onset prior to the age of 40 reported a risk of 40% [10].

The large variability accompanying many of these estimates (±15%–20%) does not on its own appear sufficient to explain the apparent inconsistencies of the estimates from the high-risk and population-based samples. One possible explanation for the higher risk estimates from the multiple-case families is ascertainment bias due to the selection of families showing linkage to BRCA1 (lod > 1.0). The general practice of selecting families based on lod scores depends on the marker distribution in the family and can lead to overestimates of penetrance in carriers from the retrospective likelihood [15]. The authors report that their estimates are not sensitive to this cut-point so they believe such bias is negligible in their data. Another possibility is that the increased risk estimates are capturing a **correlation** of disease due to unmeasured risk factors shared among the relatives. This is biologically plausible and can explain the experience

of genetic counselors that gene carriers in families having a strong history of disease appear to be at increased risk of disease over carriers from the general population. For this reason, genetic counselors prefer to counsel subjects based on estimates derived from families with a similar family history. Estimates derived from clinic-based families are used for counseling individuals with a strong family history of disease and estimates derived from population-based samples are preferred for counseling negative family history subjects.

Conclusions

Accurate estimates of age of onset are crucial to clinical management and designing disease prevention studies. They are also valuable in studies of disease etiology. More work is needed on understanding the subtleties of applying simple genetic models to complex traits and on extending current methods to fit more complex models. It will be very important to characterize how risks vary across different populations, and their possible modification by exposure, both environmental and genetic.

References

[1] Abel, L. & Bonney, G.E. (1990). A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases, *Genetic Epidemiology* **7**, 391–407.

[2] American Cancer Society (1992). *Cancer Facts and Figures*. ACS, Atlanta.

[3] Claus, E.B., Risch N. & Thompson, W.D. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study, *American Journal of Human Genetics* **48**, 232–242.

[4] Chatterjee, N. & Wacholder, S. (2001). A marginal likelihood approach for estimating penetrance from the kin-cohort design, *Biometrics* **57**, 245–252.

[5] Easton, D.F., Ford, D., Bishop, D.T. & the Breast Cancer Linkage Consortium (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers, *American Journal of Human Genetics* **56**, 265–271.

[6] Elston, R.C. & George, V.T. (1989). Age of onset, age at examination, and other covariates in the analysis of family data, *Genetic Epidemiology* **6**, 217–220.

[7] Gail, M.H., Pee, D., Benichou, J. & Carroll, R. (1999). Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotype-proband designs, *Genetic Epidemiology* **16**, 15–39.

[8] Gail, M.H., Pee, D. & Carroll, R. (1999). Kin-cohort designs for gene characterization, *Journal of the National Cancer Institute Monographs No. 26* **26**, 55–60.

[9] Hodge, S.E. & Elston, R.E. (1994). Lods, wrods, and mods: the interpretation of lod scores calculated under different models, *Genetic Epidemiology* **11**, 329–342.

[10] Hopper, J.L., Southey, M.C., Dite, G.S., Jolley, D.J., Giles, G.G., McCredie, M.R.E., Easton, D.F., Venter, D.J. & the Australian Breast Cancer Family Study (1999). Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2, *Cancer Epidemiology, Biomarkers and Prevention* **8**, 741–747.

[11] Kraft, P. & Thomas, D.C. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods, *American Journal of Human Genetics* **66**, 1119–1131.

[12] Li, H. & Thompson, E.A. (1997). Semiparametric estimation of major gene and family-specific random effects for age of onset, *Biometrics* **53**, 282–293.

[13] Moore, D.F., Chatterjee, N., Pee, D. & Gail, M.H. (2001). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study, *Genetic Epidemiology* **20**, 210–227.

[14] Schnell, A.H., Mahindra Karunaratne, P., Witte, J.S., Dawson, D.V. & Elston, R.C. (1997). Modeling age of onset and residual familial correlations for linkage analysis of bipolar disorder, *Genetic Epidemiology* **14**, 675–680.

[15] Siegmund, K.D., Gauderman, W.J. & Thomas, D.C. (1999). Gene characterization using high-risk families: a sensitivity analysis of the MOD score approach, *American Journal of Human Genetics* Supplement **65**, A398.

[16] Siegmund, K.D., Todorov, A.A. & Province, M.A. (1999). A frailty approach for modeling diseases with variable age of onset in families: the NHLBI Family Heart Study, *Statistics in Medicine* **18**, 1517–1528.

[17] Siegmund, K.D., Whittemore, A.S. & Thomas, D.C. (1999). Multistage sampling for disease family registries, *Journal of the National Cancer Institute Monographs No. 26* **26**, 43–48.

[18] Smigel, K. (1992). Breast cancer prevention trial takes off, *Journal of the National Cancer Institute* **84**, 669–670.

[19] Struewing, J.P., Hartge, P., Wacholder, S., Baker, S.M., Berlin, M., McAdams, M., Timmerman, M.M., Brody, L.C. & Tucker, M.A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews, *The New England Journal of Medicine* **336**, 1401–1408.

[20] Thomas, D.C. (1999). Design of gene characterization studies: an overview, *Journal of the National Cancer Institute Monographs No. 26* **26**, 17–23.

[21] Vieland, V.J. & Hodge, S.E. (1996). The problem of ascertainment for linkage analysis, *American Journal of Human Genetics* **58**, 1072–1084.

[22] Wacholder, S., Hartge, P., Struewing, J.P., Pee, D., McAdams, M., Brody, L. & Tucker, M. (1998). The kin-cohort study for estimating penetrance, *American Journal of Epidemiology* **148**, 623–630.

[23] Whittemore, A.S., Gong, G. & Itnyre, J. (1997). Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer: Results from three U.S. population-based case-control studies of ovarian cancer, *American Journal of Human Genetics* **60**, 496–504.

[24] Whittemore, A.S. & Halpern, J. (1997). Multi-stage sampling in genetic epidemiology, *Statistics in Medicine* **16**, 153–167.

K. SIEGMUND

Age–Period–Cohort Analysis

Age–period–cohort analysis refers to a family of statistical techniques for understanding temporal trends of an outcome, such as cancer **incidence**, in terms of three related time variables: the subject's age, the subject's date of birth (birth cohort), and calendar period. The fundamental ideas underlying these three perspectives of time have been understood by social scientists and public health researchers for many years. Early applications of these ideas employed innovative graphical presentations of data, but more recently investigators have also employed modeling and more formal **hypothesis testing** to understand better the separate contributions of each of these factors. Attempts to quantify the contributions of each factor have forced analysts to address the fact that age, period, and cohort are linearly dependent factors whose main effects cannot be uniquely and simultaneously estimated. This phenomenon is referred to as the **identifiability** *problem*. Available data do, however allow one to estimate the degree of curvature or departure from overall trends.

Suppose that we are interested in whether a **screening** program for breast cancer has had an impact on the **incidence rates** in a defined population. Such a program would identify cases at an earlier stage when the disease can be more effectively treated. However, shortening the time to detection would also be expected to result in a temporary rise in the calendar year (period) effect before returning to the long-term time trend. One approach we might try is to estimate the difference in the period effect before and immediately after the screening program, but we shall see that this is not an estimable quantity when we try to adjust for effects of age and year of birth. However, we can estimate a change in slope immediately before and after the program began, because this depends on curvature, and thus it is estimable. In fact, it may be reasonable to test the hypothesis that the screening program changes the slope by deflecting an ongoing trend.

Temporal Perspectives for Events

To understand the rationale for age–period–cohort analysis, as well as its inherent limitations, we first define the different time perspectives that give rise to the dynamic changes in a population, and then indicate the logical problems that arise when we try to consider all three factors simultaneously.

Definitions of Age, Period, and Cohort

Age refers to time since birth or, more generally, to time since a subject entered a study. *Period*, on the other hand, refers to the calendar date at which the outcome was determined. Finally, *cohort* identifies the calendar time when an individual was born, or entered a study; cohort thus provides an index for generational effects. The purpose of age–period–cohort analyses is to determine the separate contributions of age, period, and cohort to the outcome under consideration.

Vital statistics are often analyzed for age, period, and cohort effects. These data are readily available, and sometimes yield early hints on the etiology of a disease.

Age often influences risk of disease and socioeconomic outcomes. Hence, it would usually be essential to consider this factor in any analysis.

Period effects tend to be factors that impact all individuals under observation on a particular date, regardless of their age. For example, because everyone breathes essentially the same air, if disease incidence is affected by ambient air pollution in all age groups, and if levels of pollutants have changed over time, then we might expect to see period effects for each age group. However, not all period effects need be due to changes in causative agents. Artifacts, such as changes in medical diagnostic practice, or technology can introduce changes in disease incidence that would be manifested in the data as period effects.

Cohort effects can be attributed to factors related to the year of birth. A disease that is associated with poor nutrition in the mother might be expected to have higher incidence in cohorts born during a war or a famine. However, cohort effects may not be limited to events around the time of birth, because the cohort can also be thought of as a generational identifier. For example, cigarette smoking most commonly begins in late teens or early twenties, so that major changes in the marketing of cigarettes would affect primarily the generations who happened to be in the vulnerable age group on the date at which such marketing changes occurred. Hence, we might expect to see an effect due

to cohort for diseases that are strongly associated with the smoking of cigarettes in populations that have experienced major shifts in cigarette sales.

Early analyses of period and cohort relied mainly on descriptive plots of the data. More recently, investigators have tried to formalize the study of disease trends by fitting models that include time effects, or by considering **nonparametric** approaches to the analysis. These attempts have forced a recognition of the inherent limitation in these analyses, namely a nonidentifiability of some model parameters.

Collinearity of Age, Period, and Cohort

Figure 1 gives a **Lexis diagram**, showing the only possible diagonal paths that may be traversed by an individual under study, and it also demonstrates the relationship among the three temporal measures under consideration. The diagonal paths represent individual cohorts, c. If an event occurs to an individual of age a in year p, then a particular cohort $c = p - a$ must be involved. Hence, $a - p + c = 0$, and these time measures are linearly dependent. This dependence leads to aliasing of parameters, i.e. a fundamental inability to identify completely the separate contributions for each of the individual time factors. We usually think of an effect due to a particular factor as the contribution from that factor if other factors are held constant. However, this concept is clearly nonsensical when the factors are functionally related, as they are here.

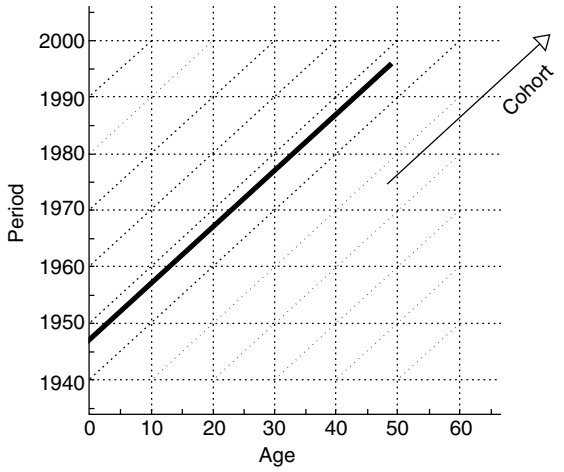

Figure 1 A Lexis diagram showing the relationship between age, period, and cohort

Interval Divisions

Population-based data are usually tabulated for the calculation of rates, by grouping age and period into categorical intervals, as seen in Table 1. Five- or 10-year intervals are most commonly used, although for large regions these rates are often reported annually, which implies one-year intervals for period. Because the grouping results in a somewhat crude measure for age and period, there remains some ambiguity when one tries to identify a corresponding cohort. For example, if a death occurred in someone aged 50–54 in 1990–94, then that individual could have been born as early as January 1, 1935, or as late as December 31, 1945, a span of 10 years, which is twice the width of the age and period interval. In addition, the intervals are overlapping, as we can see from the fact that for the next age group, 55–59, an individual could have been born between January 1, 1940, and December 31, 1950. From the Lexis diagram shown in Figure 1 we can see the pattern of age and period intervals traversed by different cohorts. While age and period uniquely define a cohort, we have lost that uniqueness in cohort definition when time has been categorized. Hence, the same cohort may pass through different age groups during a particular period interval. The same problem arises for the third time factor, when any of the other two time factors are categorized. Tarone & Chu [37] have suggested using a finer grid when possible. For example, in their analysis of breast cancer mortality they employ two-year intervals for age and period. Nevertheless, smaller overlaps remain.

If we know the cohort for each individual, then we can obtain nonoverlapping cohort categories, along with the other two time measures, by further grouping the data along the diagonal, as shown in Figure 1 [32]. When the age and period intervals are of equal width, identical width cohort intervals can be selected so that each square is divided into two triangles along the diagonal, thus achieving the same degree of precision as age and period, and avoiding the overlap at the same time.

Table 1 Observed number of cases, denominators and rates for lung cancer incidence in Connecticut males, 1935–1984

Age	1935–44	1945–54	1955–64	1965–74	1975–84
Number of cases					
20–29	1	3	4	6	7
30–39	10	20	28	31	40
40–49	70	115	195	289	281
50–59	247	543	885	1 300	1 418
60–69	395	1 057	1 992	2 780	3 769
70–79	209	790	2 001	3 017	4 354
80–89	60	231	673	1 453	2 270
Denominators					
20–29	1 537 781	1 380 360	1 555 934	2 322 128	2 769 374
30–39	1 406 807	1 615 355	1 632 000	1 863 489	2 343 684
40–49	1 258 708	1 493 910	1 800 315	1 727 315	1 800 233
50–59	1 143 763	1 232 189	1 514 848	1 789 483	1 660 060
60–69	770 224	980 496	1 095 932	1 336 181	1 561 113
70–79	437 017	567 892	723 242	756 609	941 025
80–89	145 147	207 148	273 417	345 493	389 562
Rate × 100 000					
20–29	0.07	0.22	0.26	0.26	0.25
30–39	0.71	1.24	1.72	1.66	1.71
40–49	5.56	7.70	10.83	16.73	15.61
50–59	21.60	44.07	58.42	72.65	85.42
60–69	51.28	107.80	181.76	208.06	241.43
70–79	47.82	139.11	276.67	398.75	462.69
80–89	41.34	111.51	246.14	420.56	582.71

Graphical Displays of Temporal Trends

Graphical displays offered the first approach for analyzing the effects of age, period, and cohort. One may plot the **response surface** against two time axes, such as the graph showing lung cancer incidence rates for women in Connecticut plotted on the age × period plane shown in Figure 2. The vertical axis shows the natural logarithm of the rate per 100 000 person-years experience, and we shall use this outcome to demonstrate each graphical method. While such graphs convey a broad picture of relationships, it is harder to extract some essential details. It is not easy, and sometimes impossible, to pick out the magnitude of the incidence in a particular surface plot. Other features are also unclear in two-dimensional representations of a three-dimensional figure, including whether the rates are changing on one axis, at a fixed value of the other axis. This is especially difficult for the missing time axis, i.e. cohort, in this figure. Obvious alternatives to this particular graph would entail the use of the age × cohort plane or even the period × cohort plane, although the latter is not used when, as usual, there is good reason to believe that age exerts a strong effect on the response.

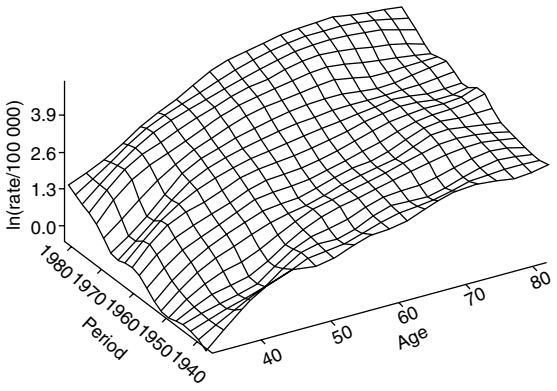

Figure 2 Natural log of the lung cancer incidence rates for Connecticut women plotted against age and period

An alternative display projects this response surface onto the age × response plane, as shown in Figure 3. Such figures were used by Korteweg [23] and others to recognize the effect of birth cohort on disease incidence. In this graph, solid lines connect

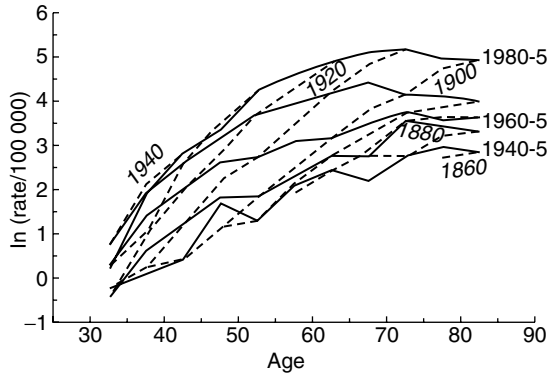

Figure 3 Natural log of the lung cancer incidence rates for Connecticut women by age (solid lines with regular font connect constant periods, broken lines with italic font connect constant cohorts)

the age-specific rates for the identical periods, and broken lines for specific cohorts. Note that the age-specific rates decline at older ages for the solid lines corresponding to fixed periods. The constant cohort (broken) lines increase monotonically with age, consistent with a belief that lung cancer risk increases with age. Because it is biologically implausible that rates should decline with age, we are led to reject the age–period model and, instead, to consider age and cohort as explanatory factors for disease trends. Similar reasoning was used by earlier investigators to suggest cohort as an important factor for some diseases. The cohort lines also tend to be more nearly parallel than the period lines, a feature that is especially relevant for the more formal models that can be fitted to these rates, as discussed below.

Projections of the response surface on the two remaining time axes can also be used (Figure 4). Each line shows either the period or the cohort trend for a particular age group. Because age so dominates these trends, these graphs better highlight some of the more subtle features for period and cohort trends. If these lines are more nearly parallel for either the period or the cohort axes, then that factor offers a more **parsimonious** description of the age-specific rates. We use the same scale for period and cohort so that the bend in a line will have the same visual impact for either period and cohort. Otherwise, the period axis would be more spread out, thus visually diluting some of the curvatures, as we can see in a period plot of the same data in a typically proportioned graph in Figure 5. By stretching the period axis, curves

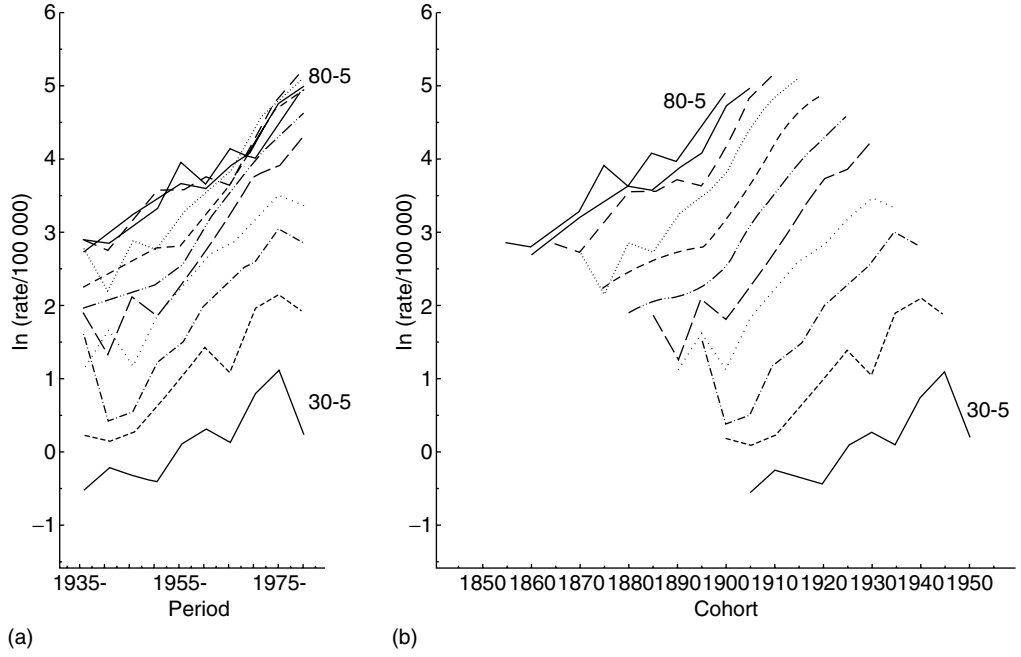

Figure 4 Natural log of the lung cancer incidence rates for Connecticut women by (a) period and (b) cohort

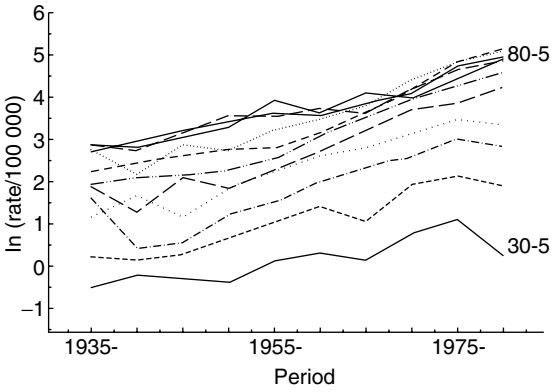

Figure 5 Natural log of the lung cancer incidence rates for Connecticut women by period

begin to appear more nearly straight and more nearly parallel. Because there are generally more cohorts than periods, the period trends will tend to look straighter unless we show the axes on the same scale. Hence, to facilitate the comparison of period and cohort effects, it is important to use the same abscissal scale.

Contour plots offer yet another approach for displaying features of the response surface by projecting lines producing the same response onto the age × period plane, as shown in Figure 6. The contours represent lines of constant lung cancer incidence, and are labeled according to ln(incidence/100 000). Similar graphs can be produced using the age × cohort or period × cohort planes [22]. Following the lines parallel to the age (period) axis, we can tell the rate at which the surface is increasing by how rapidly we cross the contour lines. Regions where the contours are parallel to the age (period) axis do not exhibit a change in incidence with age (period). Figure 6 also shows the constant cohorts as diagonal dotted lines, and we can make similar interpretations with respect to cohort by observing whether the contours are crossed or are parallel to this axis. These graphs can be especially useful when trying to understand complex patterns, such as the contour graph for Hodgkin's disease shown in Figure 7, in which there is more than one mode.

Another refinement that can assist in the interpretation of trends is to smooth the rates before preparing the contour plot. For example, Cislaghi et al. [7] give contour plots on the age–cohort plane for observed rates, fitted rates using a **polynomial regression** model, and residuals that show the adequacy of the model over the entire plane. Other variations include the use of models with period effects, or **spline functions**, in place of polynomials.

The cohort lines shown in Figure 6 have a different scale because they are diagonals on a rectangular age × period grid. Weinkam & Sterling [40] propose the use of a triangular lattice that represents the plane consisting of the locus of possible combinations of age, period, and cohort in three-dimensional age–period–cohort space. In this way they are able to present an identical scale for each time element, while at the same time using a two-dimensional time plane. This approach emphasizes that there are really only

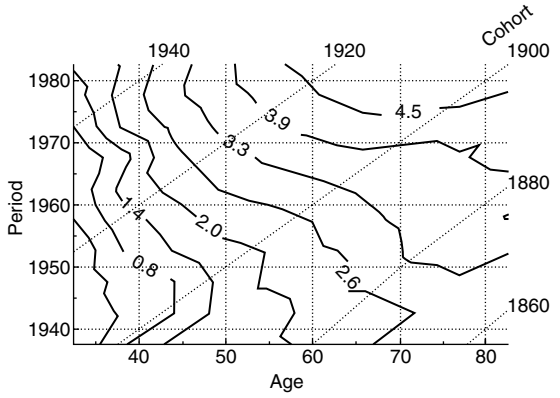

Figure 6 Contour plot for natural log of the lung cancer incidence rates for Connecticut women by age and period (cohorts are shown by the diagonal broken lines)

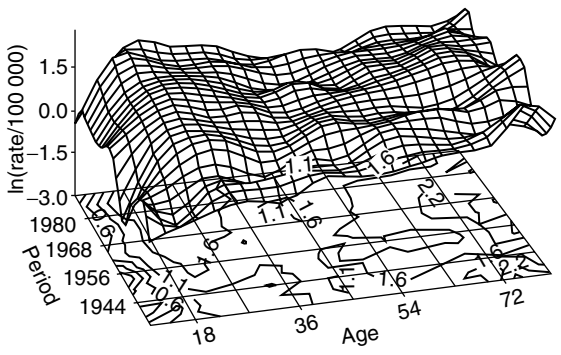

Figure 7 Surface and contour plot for the natural log of Hodgkin's disease incidence rates for Connecticut women by age and period

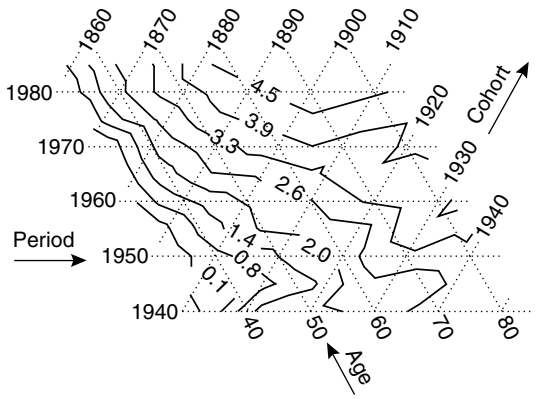

Figure 8 Contour plot for the natural log of the lung cancer incidence rates for Connecticut women on an age–period–cohort triangular lattice

two time dimensions that underlie an age–period-cohort analysis. Figure 8 shows the contours for the lung cancer example using this approach. Interpretation of the graph is similar to that of the contour plots discussed earlier.

Modeling Temporal Effects

In this section we consider age–period–cohort analyses that arise from fitting models to data. Because of the identifiability problem that arises from the **collinearity** among the time factors, it is impossible to determine parameters uniquely in models based on a linear combination of age, period, and cohort factors. We discuss several proposals to overcome this difficulty.

Vital statistics are often presented as **rates**, found by taking the ratio of the number of events divided by the total person-years experience. It is common to assume that the numerator has a **Poisson distribution**, and that the log rate is a linear function of specified regressor variables. Models of this form belong to the class of generalized linear models, which can be readily fitted using standard statistical software (*see* **Software, Biostatistical**).

Additive Effects

To formulate a linear model for the temporal effects, we first consider the case where data have been tabulated by dividing age and period into categories of equal width. This is the most common situation in practice, and instances where the interval widths are different for age and period actually give rise to still further complications [13, 15]. Let $i(= 1, \ldots, I)$ represent the age groups, $j(= 1, \ldots, J)$ the periods and $k(= 1, \ldots, K = I + J - 1)$ the cohorts. In a typical table, i represents the row index, j the column index, and k the upper-left to lower-right diagonals, beginning with the cell in the lower-left corner of the table. These indices are also linearly dependent, $k = j - i + I$, so the issue of collinearity remains. A typical **additive model** can be given by

$$Y_{ijk} = \phi_0 + \phi_{\mathrm{a}i} + \phi_{\mathrm{p}j} + \phi_{\mathrm{c}k} + \varepsilon_{ijk},$$

where Y_{ijk} represents the response (perhaps the log rate), ϕ_0 is an intercept, other parameters in the model ($\phi_{\mathrm{a}i}$, $\phi_{\mathrm{p}j}$ and $\phi_{\mathrm{c}k}$) represent age, period, and cohort effects, and ε_{ijk} is a **random error**. This equation has the same general form as **analysis of variance** models, and additional constraints must be made. One approach is to set the parameters arbitrarily at one level to zero, $\phi_{\mathrm{a}1} = \phi_{\mathrm{p}1} = \phi_{\mathrm{c}1} = 0$, say. Alternatively, we can adopt the usual constraints, $\sum_i \phi_{\mathrm{a}i} = \sum_j \phi_{\mathrm{p}j} = \sum_k \phi_{\mathrm{c}k} = 0$, which will be used in the remainder of this discussion. Unfortunately, forcing the parameters to satisfy these constraints does not entirely resolve the identifiability problem; a further constraint is necessary if one is to obtain a unique set of parameter estimates. Many regression packages allow for the possibility of a linear dependence among the **covariates** by employing a generalized inverse when fitting a model, which results in additional arbitrary constraints. The results can differ widely depending on the constraints used in the analytical software, and the order in which the factors are assigned to the model [15].

Linear Dependencies in the Design Matrix. Parameters under the usual constraints can be determined by setting up a **dummy variable** design matrix. Let the age columns of the design matrix be given by

$$\mathbf{A} = (\mathbf{A}_1 \quad \mathbf{A}_2 \ldots \mathbf{A}_{I-1}),$$

where the ith column is defined as

$$A_i = \begin{cases} 1, & \text{if } i\text{th age group,} \\ -1, & \text{if } I\text{th age group,} \\ 0, & \text{otherwise,} \end{cases}$$

thus yielding the parameters ϕ_{ai}, $i = 1, \ldots, I-1$, and $\phi_{aI} = -\sum_{i=1}^{I-1} \phi_{ai}$. The period, **P**, and cohort, **C**, components of the design matrix are similarly defined. Kupper et al. [24, 25] show that the columns of the overall design matrix formed by concatenating all three components satisfy

$$\sum_{i=1}^{I-1}\left[i - \frac{I+1}{2}\right]\mathbf{A}_i - \sum_{j=1}^{J-1}\left[j - \frac{J+1}{2}\right]\mathbf{P}_j + \sum_{k=1}^{K-1}\left[k - \frac{K+1}{2}\right]\mathbf{C}_k = 0.$$

Thus, these columns are linearly dependent. A condition for the existence of a unique set of parameter estimates in a regression model is that the design matrix be of full column rank. Hence, a model that simultaneously includes age, period, and cohort effects does not yield a unique set of estimates, which is referred to as the identifiability problem.

Partitions into Linear and Curvature Effects. One convenient way of representing trends for a particular factor is to include the overall trend or slope and the departure from that trend, namely the curvature. This approach represents the age effects obtained under the usual constraints as

$$\phi_{ai} = \left(i - \frac{I+1}{2}\right) \times \beta_a + \gamma_{ai},$$

where β_a is the overall slope, and γ_{ai} is the curvature. Period and cohort parameters can be represented in a similar way. Holford [17] proposed using the usual **least squares** estimate of the slopes, which can be expressed as a linear **contrast** among the age parameters $\beta_a = \mathbf{C} \times \phi_a$, where the contrast vector has elements

$$C_i = \left[i - \frac{I+1}{2}\right] \times \frac{12}{I(I-1)(I+1)}$$

for equally spaced intervals. This is the first-order **orthogonal** polynomial contrast (*see* **Polynomial Approximation**). We call β_a the "least squares linear component". Alternatively, Clayton & Schifflers [8] use the **mean** of the successive differences to represent the slope, which reduces to $(\phi_{aI} - \phi_{a1})/(I-1)$, and thus depends only on parameters in the first and last age groups. Both of these approaches for defining slopes can also be modified by restricting the range over which the slope is determined, either by defining the contrast appropriately in the case of using least squares, or by choosing groups other than the first and last when calculating the mean differences, i.e. using $(\phi_{ai} - \phi_{ai'})/(i - i')$.

Curvature terms can be determined by taking the difference between the estimated parameters and the fitted value from a **simple linear regression**, i.e. the residuals. If the least squares linear component is used, then these residuals are

$$\gamma_{ai} = \phi_{ai} - \left[i - \frac{I+1}{2}\right] \times \beta_a.$$

Identifiability Problem for Parameters. Because of the collinearity among the three temporal factors, a unique set of parameter estimates cannot be obtained without further constraints. Using different constraints can change not only the magnitude of the parameters, but the direction of trend for each time factor, thus profoundly influencing the conclusions from an analysis. The partitioning of the temporal effects into linear and curvature components provides one useful way of reducing the number of parameters involved in the collinearity, leading to a better understanding of its effect. It has been shown that the curvature parameters, such as γ_{ai}, are invariant, regardless of the parameterization or constraints on linear components [15, 33]. The same is not the case for the slopes (β_a, β_p, and β_c), which can arbitrarily take any value, $\beta_\cdot \in (-\infty, \infty)$. While each slope parameter may vary widely, all these parameters can only do so while maintaining a specific relationship among themselves. This constrained relationship suggests the use of estimable functions of the parameters, i.e. functions that do not depend on the constraints adopted to find a particular set of parameter estimates.

For an arbitrary pair of numbers, (r, s), the linear function, $r\beta_a + s\beta_p + (s-r)\beta_c$ is invariant to the particular set of parameters obtained, i.e. it is an estimable function of the slopes [15]. For example, by setting $r = s = 1$, we see that $\beta_a + \beta_p$ is estimable. Likewise, $r = 0$ and $s = 1$ demonstrate that $\beta_p + \beta_c$ is estimable, and in a similar fashion we can find other combinations of the slopes that are not affected by arbitrary constraints applied to obtain a particular set of parameters.

The completely unlimited range of values that can be arbitrarily assigned to an individual slope is certainly a serious drawback of these analyses. But

through the use of estimable functions we can see that if any one slope is determined, then the other two are immediately identified as well. With this in mind, any underlying quantity representing the individual slopes can be expressed as

$$\beta_a^* = \beta_a + \nu,$$
$$\beta_p^* = \beta_p - \nu,$$
$$\beta_c^* = \beta_c + \nu,$$

where β_a, β_p, and β_c are the true slopes and ν is an indeterminant parameter. For example, if we are particularly interested in period trends, β_p, then it is disconcerting that the estimated slope might be either increasing or decreasing depending on the unknown ν. However, β_a^* also depends on the same indeterminant constant, so that if it is implausible on substantive grounds for rates to decrease with age, then the values for ν that make the age slope negative must be implausible for β_p^* and β_c^* as well. If one can somehow show that ν lies within a particular range, then there is a corresponding range of values that must hold for the period and cohort effects as well.

We can observe the effect of nonidentifiability of the linear terms by considering a model in which we ignore the curvature components

$$Y = \mu + a \times \beta_a + p \times \beta_p + c \times \beta_c.$$

Because of the linear dependence between the time factors, we can add $0 = \nu \times (a - p + c)$ to the right-hand side, yielding

$$\begin{aligned} Y &= \mu + a(\beta_a + \nu) + p(\beta_p - \nu) + c(\beta_c + \nu) \\ &= \mu + a \times \beta_a^* + p \times \beta_p^* + c \times \beta_c^*, \end{aligned}$$

which is the model based on parameters obtained using a particular set of constraints.

Example. To illustrate the result from fitting age–period–cohort models, consider the data on lung cancer incidence in Connecticut men shown in Table 1. An analysis of deviance for a **loglinear model** fitted to these incidence rates is shown in Table 2, suggesting that the model does give a good fit to the data overall, and that each of the time components is statistically significant. The adequacy of the model can be further confirmed by an analysis of the residuals. Notice that the change in the scaled deviance (*see* **Model, Choice of**) attributable to age has $7 - 2 = 5$ **degrees of freedom**. The additional reduction in degrees of freedom arises because a model that includes period and cohort includes terms that are completely aliased with linear age. Hence, the test for the age effect when period and cohort are included in the model is only a test of age curvature. Likewise, the contribution for each of the factors is one less than the usual degrees of freedom that result from including a categorical factor in a model.

Table 2 Summary of analysis of deviance from fitting a loglinear model to the data in Table 1

Source	df	Scaled deviance	*P*
Goodness of fit	15	15.99	0.3825
Age|period, cohort	5	2907.06	<0.0001
Period|age, cohort	3	118.15	<0.0001
Cohort|age, period	9	464.87	<0.0001

We can observe the alternative sets of parameters from fitting the various models in Table 3. Despite the large discrepancies among these models, they each give identical fitted values. The second column was obtained by simply including age, period, and cohort in the regression model. In this instance the program set a parameter to zero when it discovered the first column of the design matrix that identified it as not being of full rank. Hence, the last two cohort effects are zero.

The third column includes linear age and cohort terms (period is not included because of the linear dependence), followed by dummy variables which constrain the first and last curvatures to be zero. The coefficients for the age and period terms correspond to net trends identified by considering the mean of successive differences, and the linear age effect estimates $\beta_a + \beta_p$, and cohort $\beta_c + \beta_p$. By dropping the curvatures we can readily see the source of the degrees of freedom for each effect, because we have applied two constraints on the components not accounted for by linear trend.

The final column shows the results of partitioning the effects into a least squares slope, and the corresponding residuals. These can be determined by: (i) simple linear regression on the parameters where the slope identifies the linear component and the residuals are the curvature; (ii) estimating a contrast using the approach described below; or (iii) forming a design matrix that is orthogonal to the linear component [21].

Table 3 Alternative parameter estimates obtained by fitting a loglinear model to the data in Table 1

	Default constraints	Mean change	Least squares
Intercept	−4.8338	−18.7566	−8.7477
$a + p$	–	1.5099	1.5433
$c + p$	–	0.3962	0.3738
Age			
20–29	−8.0544	0.0000	−0.7816
30–39	−6.1367	0.5754	−0.2449
40–49	−4.0035	1.3661	0.5073
50–59	−2.2868	1.7404	0.8429
60–69	−1.0821	1.6027	0.6665
70–79	−0.3444	0.9980	0.0233
80–89	0.0000	0.0000	−1.0134
Period			
35–44	−0.6699	0.0000	−0.1210
45–54	−0.3154	0.1871	0.0713
55–64	−0.1041	0.2309	0.1202
65–74	−0.0322	0.1352	0.0298
75–84	0.0000	0.0000	−0.1003
Cohort			
1855	−2.2875	0.0000	−0.6699
1865	−1.7225	0.3362	−0.3165
1875	−1.0593	0.7707	0.1352
1885	−0.6079	0.9933	0.3751
1895	−0.3056	1.0668	0.4658
1905	−0.2121	0.9316	0.3478
1915	−0.1045	0.8104	0.2439
1925	0.0750	0.7612	0.2118
1935	0.0654	0.5228	−0.0093
1945	0.0000	0.2287	−0.2862
1955	0.0000	0.0000	−0.4977
Scaled deviance (df = 15)	15.99	15.99	15.99

Approaches to Identifiability

By its very nature there cannot be a solution to the identifiability problem in the usual sense. We have already seen that alternative constraints can yield very different parameter trends, as we see from a graph of the age, period, and cohort effects in Figure 9. Notice that we can rotate the period slope 180° without affecting the fit of the model, but as we rotate period in a clockwise direction, there is a corresponding counterclockwise rotation for age and cohort. Proposals for ways to obtain a particular set of parameter estimates are necessarily arbitrary, and must be subject to critical evaluation when trying to interpret the results from an analysis. A variety of solutions have been proposed, but each has potentially serious limitations. Alternatively, we can limit our summaries to estimable functions of the parameters, thus avoiding the arbitrariness of any particular solution.

Parameter Constraints. A unique set of parameter estimates for a model of time trends is obtained by setting constraints on the parameters. Sometimes these are selected arbitrarily by a regression program that makes use of a particular generalized inverse when finding **maximum likelihood** estimates. However, it is better for the analyst to specify the constraint and to understand the implications of that constraint.

Drop a Factor. Perhaps the simplest approach to nonidentifiability is the attempt to avoid it by not considering all three factors simultaneously. When fitting such a two-factor model, the interpretation of the results seems quite straightforward, and it may in fact be a very reasonable approach if such a model gives a good fit to the data. However, implicit in any model that drops one of the factors is that it has no effect, i.e. there is neither curvature nor a linear effect due to the factor. As we have already seen, the latter cannot be addressed from the data so that there may still be a lingering source of **bias** in the parameters – the unidentifiable constant ν – which could have influence even if the model shows a good fit to the data.

Equate Two Effects. A second approach to finding a unique set of parameters is to equate just two of the effects for one of the model factors [3, 4, 13], rather than equating all effects for one factor to zero. For instance, two adjacent period effects may be set equal to each other because there is reason to believe that no changes occurred during that epoch, e.g. $\phi_{p1} = \phi_{p2}$. A variation on this approach is an assumption that the mean of the successive differences is zero, which reduces to $\phi_{p1} = \phi_{pJ}$ in the case of period. This is actually the constraint automatically specified by some regression programs for the last factor specified for a model. This constraint is very simple to apply, and it forces the parameters to return to their original level, which can yield parameters that are similar to those obtained by setting the period slope to zero, as discussed below.

The advantage of this approach to the nonidentifiability problem is that it is quite simple to understand

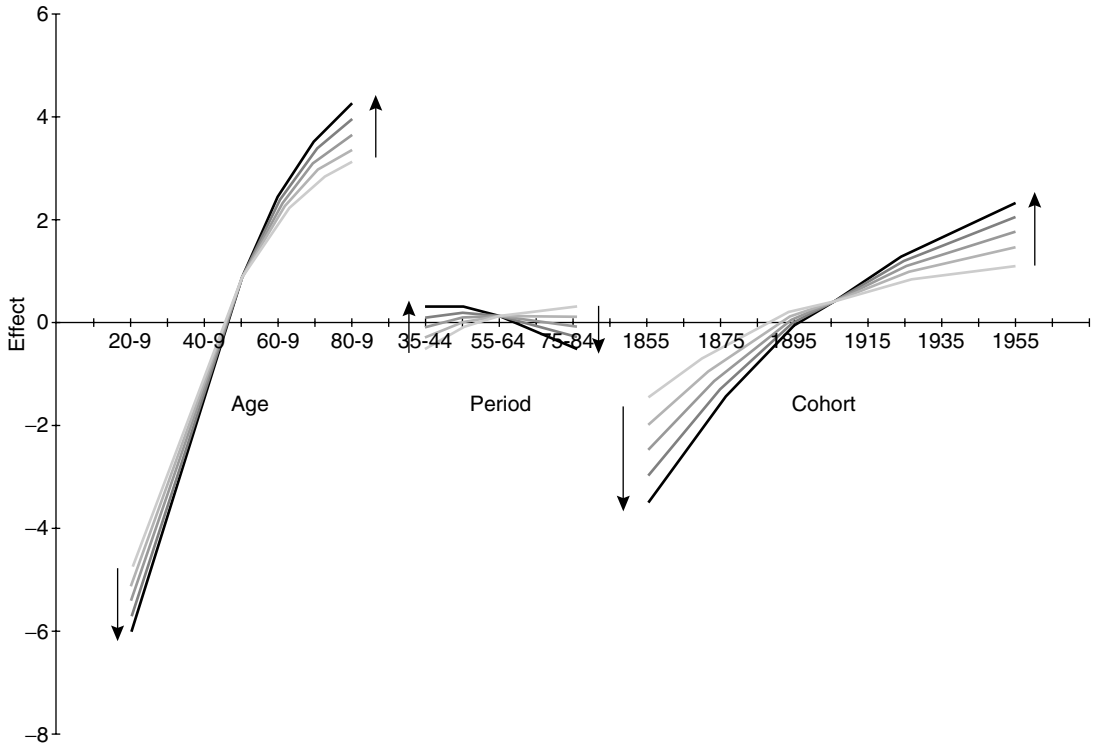

Figure 9 Age, period, and cohort effects in which the period slope takes values $-0.2(0.1)0.2$.

and apply. In addition, it does not force equality for all the effects, as is the case when one of the factors is dropped entirely from the model. Unfortunately, there is usually no more solid basis for equating two effects than reasoning such as: "There is no reason to expect a change during these years; therefore, they will be assumed to be equal." It is often true that equality of the fourth and fifth periods is just as logical as the first and second in a particular situation, and the resulting parameter estimates can vary considerably for these equally plausible assumptions.

Minimize Euclidean Distance to Two-factor Models. Osmond & Gardner [31] propose an approach that estimates the unidentifiable parameter, ν. Their criterion uses the Euclidean distance between the parameters from the age–period–cohort model, and a corresponding model that drops one of the factors, e.g. $\|\phi(\nu) - \phi_{(c)}\|$ in the case of a model that drops cohort. Because age plays a vital role in most responses, it is not eliminated entirely; but rather the fitted values from an age-only model are introduced as offset terms. The estimation criterion is to minimize

$$g(\nu) = \frac{\|\phi(\nu) - \phi_{(c)}\|}{\rho_c} + \frac{\|\phi(\nu) - \phi_{(p)}\|}{\rho_p} + \frac{\|\phi(\nu) - \phi_{(a)}\|}{\rho_a},$$

where ρ_c, ρ_p, and ρ_a are the residual mean squares from the respective models.

While this approach has the advantage of offering a unique set of model parameters, there is some question about whether the criterion is appropriate in general. At one level it seems sensible to give parameters from two-factor models with poor fit (high residual mean squares) less weight, until we recall that we can only estimate curvature in a three-factor model. A factor with a great deal of curvature will result in a relatively large residual mean square for the reduced model, thus receiving less weight according to the estimation criterion. But it is not clear why the parameter that identifies linear trend should be

related to curvature, or indeed to parameters from a two-factor model that may give a poor fit to the data. Alternative underlying models for the overall trends for the time factors can give rise to the same responses, as we have seen, and this approach does not guarantee that we will find the "correct" underlying trend.

Set a Slope to Zero. An alternative to excluding a factor altogether is to assume that its slope is zero. For example, we might specify that there is no period slope, $\beta_p = 0$ [15]. This approach is an immediate extension of the deletion method in that both assume that the overall slope for one factor is zero; however, this approach does not also require that all the curvature terms be set to zero. There may still be unidentifiable bias for all three slopes. Another variation on this theme is to fix the slope over a shorter span of time, rather than the entire span. For example, we might assume that there is no trend with period for the years 1940–1969, a span of three 10-year periods.

Roush et al. [34, 35] undertook a systematic study of cancer incidence using data from the Connecticut Tumor Registry. This analysis focused primarily on the curvature effects for each of the time factors, but in the summary graphs a period slope of zero was specified, $\beta_p = 0$. The rationale for this approach was that: (i) there is a strong biologic basis for an age effect on cancer, so that if only one factor is unimportant it is likely to be either period or cohort; (ii) empirical results strongly suggest that cohort has a stronger association with cancer incidence than period; and (iii) the assumption that $\beta_p = 0$ was less restrictive than ignoring the effect of period altogether.

Restrict the Range of the Slopes. Another way to select constraints on the parameters is to employ theoretical knowledge about the underlying process with respect to one of the time factors. Wickramaratne et al. [41] analyzed the effects of age, period, and cohort on risk of major depression in five US communities. Although a specific assumption about the overall trends with period and cohort was not imposed, it seemed reasonable to assume that there was not a decreasing trend with either period or cohort, i.e. $\beta_p \geq 0$ and $\beta_c \geq 0$. Adding β_c to both sides of the first inequality, and β_p to the second gives $\beta_p + \beta_c \geq \beta_c \geq 0$ and $\beta_p + \beta_c \geq \beta_p \geq 0$. Note that the upper bounds are estimable in each case. Similarly, the age slope must satisfy the inequality $\beta_a + \beta_p \geq \beta_a \geq \beta_a - \beta_c$, which also has estimable upper and lower bounds. Using these bounds, Wickramaratne et al. [41] were able to obtain the qualitative result that there was an increasing trend in the risk of major depression in the cohort born during the years 1935–1944, even though it was not possible to obtain a point estimate for the trend.

Estimable Functions. To avoid the adoption of arbitrary assumptions, estimable functions of the parameters offer summaries that are identical for any particular set of model parameters. In this section we discuss several estimable functions that have been found to be useful.

Forecasting Based on Age–Period–Cohort Models. The problem of **forecasting** trends is one that is difficult because one must necessarily make assumptions regarding trends beyond the range of existing data, which in general cannot be verified. For example, we might assume that the trends of the past will continue into the future, an albeit strong assumption [7, 25] which may well be unwarranted in a particular instance. Nevertheless, it is an assumption that is commonly made in other contexts, and it is one that seems reasonable in the absence of contradictory information. If we make a linear **extrapolation** for all three time parameters, then the resulting projected rates are identifiable [16, 30]. This property can be demonstrated by using a model that only includes linear terms, remembering that more complicated models that include curvature terms present no new problems, because the curvature parameters are estimable. The resulting model for the ith age, jth period, and kth cohort is

$$Y_{ijk} = \mu + i \times \beta_a + j \times \beta_p + k \times \beta_c.$$

Following the same cohort in time by increasing the age and the period index by one unit, gives

$$Y_{i+1,j+1,k} = \mu + (i+1)\beta_a + (j+1)\beta_p + k \times \beta_c,$$

and the difference between the two rates,

$$Y_{i+1,j+1,k} - Y_{i,j,k} = \beta_a + \beta_p,$$

which is an estimable function of the slopes, as we have already seen.

Drift Based on Mean of Successive Differences. We have already noted Clayton & Schifflers' [8, 9] suggestion of using the mean of successive differences as an estimate of overall trend for a particular factor. They also proposed the sum of the period and cohort slopes, $\beta_p + \beta_c$, as an indicator of the overall trend for the outcome during the span of time covered by the data, which they call the net drift. This is an estimable function of the model parameters, and hence it is unique. We can estimate the net drift for any set of parameter estimates by taking the contrast

$$\frac{\phi_{pJ} - \phi_{p1}}{J-1} + \frac{\phi_{cK} - \phi_{c1}}{K-1}.$$

Alternatively, we can estimate the drift for any range of periods and cohorts by using the contrast

$$\frac{\phi_{pj^*} - \phi_{pj}}{j^* - j} + \frac{\phi_{ck^*} - \phi_{ck}}{k^* - k}.$$

Drift Based on Slopes. As an alternative to determining drift on the basis of the mean of first differences, we can use the sum of the least squares estimates of the slopes. This can be expressed in terms of a linear contrast among the period and cohort parameters, $\beta_p + \beta_c = (\mathbf{C}_p'|\mathbf{C}_c') \cdot (\boldsymbol{\phi}_p'|\boldsymbol{\phi}_c')'$, where the elements of contrast vectors are the first-order orthogonal polynomial contrasts defined previously. We accomplish this by concatenating the slope contrasts for period and cohort. The variance for the contrast can be estimated from $\text{var}(\beta_p + \beta_c) = (\mathbf{C}_p'|\mathbf{C}_c') \times \text{var}(\boldsymbol{\phi}_p'|\boldsymbol{\phi}_c')(\mathbf{C}_p'|\mathbf{C}_c')'$.

Curvature Estimates. While nonidentifiability of the slope for a time factor implies that we cannot identify the overall trend, there remains useful information in the curvatures or departures from linear trend which are estimable. In fact, any contrast that is orthogonal to the first-order contrast for linear trend is estimable when the equal age and period intervals are used. Hence, we can determine any aspect of the shape of the curves, including information on whether the trends are concave upward or downward.

This approach can also be applied when looking for spikes in the overall trend lines by considering second differences, such as $D_k = \phi_{c,k} - 2\phi_{c,k+1} + \phi_{c,k+2}$ in the case of cohort [8, 36]. Tango & Kurashina [36] studied such effects by comparing mortality from diabetes, ischemic heart disease, liver cirrhosis, and suicide around the Showa Era, 1925–1940. They found that men born in this era had a higher than expected risk compared with the overall trend among men born in surrounding cohorts. This particular cohort experienced nutritional deprivation in adolescence during World War II, and they contributed extensively to the rapid economic expansion during the 1960s which introduced profound changes to Japanese society.

Because of the overlapping of cohort intervals, Tango & Kurashina [36] also suggested estimating the average of second differences, $(D_k + D_{k+1})/2$.

If we think of second differences as comparing slopes between adjacent points, then a natural extension is to consider changes in slopes over longer spans of time [38]. Suppose that we wish to consider slopes for two such cohort epochs, ϕ_{c1} and ϕ_{c2}, respectively. We already know that because of the identifiability problem, we can only estimate slopes that are aliased, $\phi_{c1}^* = \phi_{c1} + \nu$ and $\phi_{c2}^* = \phi_{c2} + \nu$. However, the difference is estimable because the indeterminant constant, ν, is canceled out.

The change in slope can be estimated using a contrast matrix formed by subtracting vectors that give slopes over the corresponding epochs. For example, if $\mathbf{C}_1$ is the cohort contrast for the slope during the first epoch, and $\mathbf{C}_2$ during the second, then the change in slopes is determined by $(\mathbf{C}_1 - \mathbf{C}_2)$. To illustrate this using the 11 cohorts represented in Table 1, suppose we wish to determine whether there is a significant change in cohort trend for men born from the turn of the century until 1925, and men born in the years following 1925. The resulting contrast would be

$$\begin{aligned}
&(0\ \ 0\ \ 0\ \ 0\ \ 0\ \ -0.5\ \ 0\ \ 0.5\ \ 0\ \ 0\ \ 0)\\
&-(0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ -0.3\ \ -0.1\ \ 0.1\ \ 0.3)\\
&\quad =(0\ \ 0\ \ 0\ \ 0\ \ 0\ \ -0.5\ \ 0\ \ 0.8\ \ 0.1\ \ -0.1\ \ -0.3),
\end{aligned}$$

which yields a Wald statistic of 1.39 on 1 df, and a contrast estimate of 0.172 (se = 0.149).

Autoregressive Models for Time Effects. We have already noted the identifiability of second differences. Berzuini & Clayton [6] propose an autoregressive model for the time effects (*see* **ARMA and ARIMA Models**) in which successive parameters are given by

$$\phi_{c,k} = 2\phi_{c,k-1} - \phi_{c,k-2} + \varepsilon_{ck}$$

in the case of cohort, with similar expressions for age and period. Each term is clearly related to the two previous parameters, along with an added random perturbation, $\varepsilon_{ck} \sim \mathrm{N}(0, \sigma_c^2)$. Berzuini & Clayton describe a **Bayesian method** for estimating the model parameters that employs a Markov chain Monte Carlo algorithm. One of the interesting extensions that this approach offers is Bayesian forecasting of rates, which is also an estimable function of the model parameters, as noted above. This is in contrast to the use of an autoregressive model for the cohort effect by Lee & Lin [26], who used this model to obtain a unique set of parameter estimates. As always, unique estimates for the nonestimable functions of the parameters depend on strong, unverifiable assumptions.

Design Matrices. An alternative to the use of contrasts for parameter estimates is to construct a design matrix with linearly independent columns which will yield a set of parameter estimates that are unique. Of course, the usual approach of constructing a design matrix using dummy variables for each level of age, period, and cohort will necessarily contain a linear dependence because of the dependence among the indices noted above.

A partitioned design matrix that will partition the effects into the linear and curvature components described above can be written as

$$\mathbf{X} = [\mathbf{1}|\mathbf{A}_\mathrm{L}|\mathbf{A}_\mathrm{C}|\mathbf{P}_\mathrm{L}|\mathbf{P}_\mathrm{C}|\mathbf{C}_\mathrm{L}|\mathbf{C}_\mathrm{C}],$$

where each row corresponds to the rate in a particular age and period group. The columns represented by $\mathbf{A}_\mathrm{L}$ and $\mathbf{A}_\mathrm{C}$ are the linear and curvature components for age, and the remaining elements of the design matrix are the corresponding terms for period and cohort. Regression parameters that correspond to the components of this design matrix are given by the vector

$$\boldsymbol{\Theta} = (\phi_0|\beta_\mathrm{a}|\gamma_\mathrm{a}|\beta_\mathrm{p}|\gamma_\mathrm{p}|\beta_\mathrm{c}|\gamma_\mathrm{c})'.$$

The column vector for the age slope, $\mathbf{A}_\mathrm{L}$, may be defined as having the elements $A_{\mathrm{L}i} = i - [I + 1]/2$ for the ith age group. For period and cohort, the slope columns, $\mathbf{P}_\mathrm{L}$ and $\mathbf{C}_\mathrm{L}$, are similarly defined. The linear dependence in this design matrix is readily apparent because $\mathbf{P}_\mathrm{L} = \mathbf{A}_\mathrm{L} + \mathbf{C}_\mathrm{L}$. Thus, a model that includes $\mathbf{A}_\mathrm{L}$ and $\mathbf{C}_\mathrm{L}$ would have already effectively included $\mathbf{P}_\mathrm{L}$. The resulting regression parameter associated with the remaining $\mathbf{A}_\mathrm{L}$ will actually estimate the sum of the age and period slopes, $\beta_\mathrm{a} + \beta_\mathrm{p}$. Likewise, the $\mathbf{C}_\mathrm{L}$ parameter would estimate $\beta_\mathrm{c} + \beta_\mathrm{p}$. These are two of the estimable functions of the slopes noted earlier.

Curvature elements of the design matrix are given by the remaining regressor variables that saturate the effect. If the first and last columns for age in a design matrix containing a 0–1 indicator variable for each category are dropped from the model, then we are effectively constraining $\phi_{\mathrm{a}1} = \phi_{\mathrm{a}I}$, which ultimately yields slopes that correspond to those obtained by considering means of successive differences. However, if we choose to represent the curvatures using variables that are orthogonal to the linear term, then we obtain slopes that correspond to the least squares slopes.

Summary of Estimable Effects. We now summarize the effect of the nonidentifiability problem on our ability to address questions of scientific interest. Without making strong assumptions, we cannot estimate overall changes for age, period, and cohort. This is a severe limitation because some of the most interesting scientific questions relate to whether trends are increasing or decreasing. However, there remain a number of questions that can still be addressed using quantities that are estimable, including:

1. predicted values;
2. slope changes or a deflection of an overall trend;
3. temporary spikes or departures from overall trend; and
4. net drift or combined period and cohort changes.

Interesting scientific questions can often be framed in terms of quantities that are estimable, although it may require us to conceptualize a problem in a different way. Ultimately, we want to produce valid estimates, and if we wish to do that without making strong assumptions about the model parameters, then we need to limit ourselves to estimable functions.

Nonlinear Effects. The difficulty caused by the nonidentifiability problem has led some to consider the use of intrinsically **nonlinear** models. One example is the model used by Moolgavkar et al. [29]:

$$Y_{ijk} = \mu + \phi_{\mathrm{p}j} + \phi_{\mathrm{c}k} + \phi_{\mathrm{a}i} \cdot \delta_j + \varepsilon_{ijk},$$

which was also considered by James & Segal [21]. In this model, $\phi_{\mathrm{p}j}$ and $\phi_{\mathrm{c}k}$ represent the effects due to

period and cohort, respectively. The effect of age, ϕ_{ai}, is included along with a multiplicative factor involving period, δ_j, which can modify the effect of age. Even though a unique set of parameters can result from this model, the parameters can be difficult to estimate and they can be inherently unstable. A special case occurs when

$$\delta_j = 1, \quad \text{for all } j,$$

which is identical to the usual model, and in instances where this gives a good fit to data the parameters are likely to be unstable [9, 36].

Another approach that sometimes leads to nonlinear models involves the theoretical introduction of external information for one of the parameters, thus specifying its functional form. For example, Holford et al. [19] discuss the use of various forms of the multistage carcinogenesis model described by Armitage & Doll [1] for use in the analysis of lung cancer incidence. In this case the effect of age has the same form as a **Weibull hazard**

$$A(a) \propto a^{\omega},$$

where the parameter ω represents the number of stages minus one. If the response represents the log hazard, then the functional form for the age effect becomes

$$\phi_{ai} = \omega \times \ln(a_i),$$

which is no longer linear in age. This and related nonlinear models for age do yield a unique set of model parameters, thus avoiding the adoption of a constraint. However, the unique set of parameters relies heavily on the success of a particular mathematical model in describing the effect of age. Even so, the parameters can be extremely unstable and difficult to estimate using the usual maximum likelihood approach. In addition, the estimates of overall trend can vary widely depending on the model chosen for the effect of age [27, 28].

Replace Temporal Variables. Underlying most studies of time trends is the idea that the effect of time is related to some factor that will affect the outcome. If this is correct, then a better analysis will include a more direct measure of the factor, rather than using time as a surrogate measure. A problem in using a model that employs information on causative agents is that we must have population data on exposure over time. Lung cancer is one instance where such a study is feasible, because we know that cigarette smoking is by far the leading cause of the disease [12, 39], and exposure information is available. Even so, there are several additional facts that must be kept in mind when developing a model, including: (i) smokers often begin in their late teens or early twenties, a fairly narrow age range, so that a change in cigarette consumption primarily affects individuals in these age groups; (ii) there can be a long lag (over 20 years) between the time one starts to smoke until cancer is diagnosed; (iii) not everyone has the same exposure, because some consume more cigarettes, and some cigarettes pose a greater risk; (iv) there is a cumulative effect of smoking, so that individuals who consume the same number of cigarettes per day at one point in time would have different cumulative exposures if they began at different times; and (v) individuals who quit smoking have a risk intermediate between current smokers and those who never smoked [11].

The contribution to risk from beginning to smoke is clearly identifiable with birth cohort [see (i) above]. However, the introduction of filters and other manufacturing changes in cigarettes would be associated with a period effect. Other factors could have components that are identified with both cohort and period. For instance, certain generations may be more health conscious and thus able to quit smoking more readily, a cohort effect, but the overall population might also be influenced by a report from the Surgeon General or an antismoking television advertising campaign, a period effect.

Brown & Kessler [5] fitted a model to US lung cancer mortality that used US data on cigarette composition over time, which would be expected to affect primarily the period parameters. The period effect was expressed as a linear function of a measure of tar, so that the log rate became

$$Y_{ijk} = \phi_0 + \phi_{ai} + \beta X_j + \phi_{ck} + \varepsilon_{ijk},$$

where X_j is a measure of the population's tar exposure for the jth period. While estimates of the prevalence of smoking were not included in the model, the pattern in the estimated cohort effects was similar to the temporal pattern of smoking prevalence, estimated from sample surveys in men and women. The successful use of information on changes in cigarette composition in this particular instance does not necessarily imply that the approach

will be uniformly successful. If there were a strictly linear trend in the mean tar content over time, then the X_j would be linearly dependent on I and k, resulting once again in nonidentifiability. Along similar lines, Holford et al. [20] used population information on the prevalence of smokers, ex-smokers, and mean years smoked to model incidence rates in Connecticut (*see* **Smoking and Health**).

Higher-order Models

Thus far we have only addressed models that include main effects for age, period, and cohort. We now consider work on higher-order models that allow for **interactions**, either with the time factors themselves, or with other groups. The identifiability problem continues to manifest itself in these more complex models; nevertheless, it is possible to address some substantive questions.

Interactions with Temporal Factors. Each of the models considered above assumes that the effect of a time factor is not modified by the level of the others, i.e. there are no interactions. Part of the problem with considering interactions between these temporal variables arises because in any two-factor model the incorporation of an interaction results in a saturated model. For example, if we only consider age and period the model becomes

$$Y_{ij} = \mu + \phi_{\mathrm{a}i} + \phi_{\mathrm{p}j} + \phi_{\mathrm{ap},ij} + \varepsilon_{ij},$$

where $\phi_{\mathrm{a}i}$ and $\phi_{\mathrm{p}j}$ are the main effects for age and period, and $\phi_{\mathrm{ap},ij}$ is the interaction. Comparing this model with the age–period–cohort model, we can see that the only difference between the two is that the first model includes a cohort effect, $\phi_{\mathrm{c}k}$, and the second an age–period interaction, $\phi_{\mathrm{ap},ij}$. Because the interactions saturate the model that only includes main effects due to age and period, we can think of the cohort effect as a particular type of age–period interaction [25]. In a similar way we can describe the period effect as a particular type of age–cohort interaction or the age effect as a particular type of period–cohort interaction, both of which are saturated models. Fienberg & Mason [14] have studied polynomial models for the three time factors, and have indicated which interactions can be identified. The interpretation of interactions in higher-order polynomial models is difficult under the best of circumstances, without complicating things still further with the nonidentifiability problem. Hence, considerable care is needed when introducing interactions into these models.

An alternative to polynomial interactions among temporal factors is simply to split times as, for example, in the comparison of cohort trends in breast cancer among women younger and older than 50 [18]. In one study this division was used because of a suggestion that breast cancer trends may differ between pre- and postmenopausal women [2]. In this instance the model may be written as

$$\log \lambda_{ijk} = \begin{cases} \mu + \phi_{\mathrm{a}i} + \phi_{\mathrm{p}j} + \phi_{\mathrm{c}k} + \phi_{\mathrm{ac},k} + \varepsilon_{ijk}, \\ \quad \text{if } i < i_0, \\ \mu + \phi_{\mathrm{a}i} + \phi_{\mathrm{p}j} + \phi_{\mathrm{c}k} - \phi_{\mathrm{ac},k} + \varepsilon_{ijk}, \\ \quad \text{if } i \geq i_0, \end{cases}$$

where $\phi_{\mathrm{ac},k}$ represents the difference from the mean log rate, and i_0 is the age category where the split is made. This type of interaction would cause no additional difficulties if we were only considering an age–period model, but another complication does arise when cohort is involved. If we follow the cohort diagonals in a typical table of rates, we see that by changing the row or age group at which the interaction occurs, we are at the same time changing the set of cohorts involved in making inferences about the interaction.

Interactions with Nontemporal Factors. Nonidentifiability also affects our ability to compare trends among groups by testing for interactions with the temporal variables. Once again it is convenient to partition the trends into linear and curvature components and, as before, only the linear terms are affected by the identifiability problem. We can express the differences between two groups by

$$(\beta^*_{\mathrm{a}1} - \beta^*_{\mathrm{a}2}) = (\beta_{\mathrm{a}1} - \beta_{\mathrm{a}2}) + (\nu_1 - \nu_2),$$
$$(\beta^*_{\mathrm{p}1} - \beta^*_{\mathrm{p}2}) = (\beta_{\mathrm{p}1} = \beta_{\mathrm{p}2}) - (\nu_1 - \nu_2),$$
$$(\beta^*_{\mathrm{c}1} - \beta^*_{\mathrm{c}2}) = (\beta_{\mathrm{c}1} - \beta_{\mathrm{c}2}) + (\nu_1 - \nu_2),$$

which are clearly aliased. In some circumstances we may be able to assume that the trends for one of the factors are identical for the groups. For example, we may be willing to assume that the age effect reflects an underlying biological process that is the same for all populations, i.e. $\beta_{\mathrm{a}1} - \beta_{\mathrm{a}2} = 0$. Hence, equating the two age linear trends identifies $\nu_1 - \nu_2$, and thus the remaining slope differences.

To illustrate how this result can be used in practice, consider the comparison of lung cancer trends for men and women in Connecticut. We can force the age trends to be equivalent for males and females by fitting a model that includes A, P, C, S, S·P and S·C, where S represents sex and where a dot indicates interaction terms. However, we might question whether it is reasonable to equate the age trends, because there is a variety of biological factors that might result in differences between men and women, including the possible effects of hormonal changes resulting from the menopause. Hence, we may have to settle for comparing the estimable interactions, such as those with curvature trends or drift. Equating age effects may be more reasonable when comparing rates for the same gender among different geographic regions [10]. However, Clayton & Schifflers [9] indicate that there is still the danger of regional differences in age-specific exposure to risk factors which can affect the age parameters.

Polynomials and Splines. In the analyses above, the time factors are treated as categoric, which allows for complete curvature flexibility for the trends. However, in some instances the variances for the individual responses may be large, such as the case when rates are based on small numbers of cases. In these circumstances it may be desirable to smooth the curvature, either by representing the effects by polynomials or by using **spline functions**. For age, these can represented by

$$\phi_{ai} = \left(i - \frac{I+1}{2}\right)\beta_a + X_{a2i}\beta_{a2} + \cdots + X_{api} \times \beta_{ap},$$

where X_{api} represents a regressor variable for a pth-order polynomial or a particular term in a spline function. The usual representation of these regressor variables is highly collinear with the first-order linear term, $i - (I+1)/2$. Thus, the identifiability problem can result in parameters that are difficult to interpret, unless an effort is made to identify the separate components of trend using the principles discussed above.

We can partition the effects into least squares linear components and the remaining curvature by defining regressor variables that are orthogonal to the linear components of trend. In the case of polynomials, this can be accomplished by employing orthogonal polynomials, which give rise to underlying slope parameters that can be interpreted in much the same way as in the models described earlier. Likewise, alternative representations of curvature, such as spline functions, can be constructed by defining variables that are orthogonal to the linear term. This can be accomplished for age by using

$$\mathbf{X}_a^* = \mathbf{X}_a - \mathbf{L}_a(\mathbf{L}_a' \times \mathbf{L}_a)^{-1} \times \mathbf{L}_a' \times \mathbf{X}_a,$$

where $\mathbf{L}_a$ is a vector of linear regressors, $\mathbf{X}_a$ is a matrix or regressor variable, and $\mathbf{X}_a^*$ is the matrix or regressor variable that is orthogonal to $\mathbf{L}_a$. A similar method can be applied to the period and cohort effects. An example of this method is shown in an analysis of thyroid cancer incidence in Connecticut [42].

Nonparametric Methods

A **nonparametric** approach to the analysis of period and cohort trends has been developed by Tarone & Chu [37]. To address the question of a cohort effect, age-specific rates are compared between adjacent cohorts and the total number of decreases is used to construct a permutation test. The **null hypothesis** is that there is no trend with cohort, and the mean and variance of the number of decreases expected out of n comparisons between successive cohorts in the same age group are $n/2$ and $(n+2)/12$ respectively. In a typical rectangular age–period matrix of rates, not all cohorts are represented by the same number of age groups. For example, in the data shown in Table 1 only one age group can be compared for the 1855 and 1955 cohorts, but four each can be made for the four cohorts from 1885 to 1915. The expected number of decreases is

$$\frac{1+2+3+4\times 4+3+2+1}{2} = 14,$$

and the sum of the corresponding variances is 4. Only three decreases are observed in the table, yielding the test statistic $z = (3-14)/4 = -2.75$, which may be compared with a **standard normal deviate**. In this case we can conclude that the rates are not constant across the cohorts.

A similar analysis can be conducted across periods, which results in the same observed and expected numbers of decreases; only the total variance has changed. In the example from Table 1, four comparisons are made in each of seven age groups, resulting

in a total variance of $7 \times (4 + 2)/12 = 3.5$, instead of 4.

Additional refinements offered by Tarone & Chu [37] include the consideration of blocks of cohorts for analysis to address the possibility that cohort effects may only be important during certain epochs and not others (*see* **Blocking**). This raises the issue of **multiple comparisons** that must be taken into account when trying to interpret the results. They also suggest comparing the results of analyses obtained by forming blocks of cohorts with blocks of periods to determine whether one factor predominates in the overall direction of the trends.

References

[1] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

[2] Avila, M.H. & Walker, A.M. (1987). Age dependence of cohort phenomena in breast cancer mortality in the United States, *American Journal of Epidemiology* **126**, 377–384.

[3] Barrett, J.C. (1973). Age, time, and cohort factors in mortality from cancer of the cervix, *Journal of Hygiene (Cambridge)* **71**, 253–259.

[4] Barrett, J.C. (1978). The redundancy factor method and bladder cancer mortality, *Journal of Epidemiology and Community Health* **32**, 314–316.

[5] Brown, C.C. & Kessler, L.G. (1988). Projections of lung cancer mortality in the United States: 1985–2025, *Journal of the National Cancer Institute* **80**, 43–51.

[6] Berzuini, C. & Clayton, D. (1994). Bayesian analysis of survival on multiple time scales, *Statistics in Medicine* **13**, 823–838.

[7] Cislaghi, C., Negri, E., La Vecchia, C. & Levi, F. (1988). The application of trend surface models to the analysis of time factors in Swiss cancer mortality, *Sozial-und Präventivmedizin* **33**, 259–373.

[8] Clayton, D. & Schifflers, E. (1987). Models for temporal variation in cancer rates. I: age-period and age-cohort models, *Statistics in Medicine* **6**, 449–467.

[9] Clayton, D. & Schifflers, E. (1987). Models for temporal variation in cancer rates. II: age–period–cohortAge–period–cohort models, *Statistics in Medicine* **6**, 469–481.

[10] Day, N.E. & Charnay, B. (1982). Time trends, cohort effects, and aging as influence on cancer incidence, in *Trends in Cancer Incidence*, K. Magnus, ed. Hemisphere, Washington, pp. 51–65.

[11] Doll, R. (1971). The age distribution of cancer: implications for models of carcinogenesis, *Journal of the Royal Society, Series A* **134**, 133–166.

[12] Doll, R. & Peto, R. (1981). The causes of cancer, *Journal of the National Cancer Institute* **66**, 1192–1308.

[13] Fienberg, S.E. & Mason, W.M. (1978). Identification and estimation of age–period–cohort models in the analysis of discrete archival data, in *Sociological Methodology 1979*, K.F. Schuessler, ed. Jossey-Bass, San Francisco, pp. 1–67.

[14] Fienberg, S.E. & Mason, W.M. (1985). Specification and implementation of age, period and cohort models, in *Cohort Analysis in Social Research*, W.M. Mason & S.E. Fienberg, eds. Springer-Verlag, New York, pp. 45–88.

[15] Holford, T.R. (1983). The estimation of age, period and cohort effects for vital rates, *Biometrics* **39**, 311–324.

[16] Holford, T.R. (1985). An alternative approach to statistical age–period–cohort analysis, *Journal of Clinical Epidemiology* **38**, 831–836.

[17] Holford, T.R. (1991). Understanding the effects of age, period, and cohort on incidence and mortality rates, *Annual Reviews of Public Health* **12**, 425–457.

[18] Holford, T.R., Roush, G.C. & McKay, L.A. (1991). Trends in female breast cancer in Connecticut and the United States, *Journal of Clinical Epidemiology* **44**, 29–39.

[19] Holford, T.R., Zhang, Z. & McKay, L.A. (1994). Estimating age, period and cohort effects using the multistage model for cancer, *Statistics in Medicine* **13**, 23–41.

[20] Holford, T.R., Zhang, Z., Zheng, T. & McKay, L.A. (1996). A model for the effect of cigarette smoking on lung cancer incidence in Connecticut, *Statistics in Medicine* **15**, 565–580.

[21] James, I.R. & Segal, M.R. (1982). On a method of mortality analysis incorporating age-year interaction, with application to prostate cancer mortality, *Biometrics* **38**, 433–443.

[22] Jolley, D. & Giles, G.G. (1992). Visualizing age–period–cohort trend surfaces: a synoptic approach, *International Journal of Epidemiology* **21**, 178–182.

[23] Korteweg, R. (1951). The age curve in lung cancer, *British Journal of Cancer* **5**, 21–27.

[24] Kupper, L.L., Janis, J.M., Karmous, A. & Greenberg, B.G. (1985). Statistical age–period–cohort analysis: a review and critique, *Journal of Chronic Diseases* **38**, 811–830.

[25] Kupper, L.L., Janis, J.M. Salama, I.A. Yoshizawa, C.N. & Greenberg, B.G. (1983). Age–period–cohort analysis: an illustration of the problems in assessing interaction in one observation per cell data, *Communications in Statistics – Theory and Methods* **12**, 2779–2807.

[26] Lee, W.C. & Lin, R.S. (1996). Autoregressive age–period–cohort models, *Statistics in Medicine* **15**, 273–281.

[27] Moolgavkar, S.H. & Venzon, D.J. (1979). A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor, *Mathematical Biosciences* **47**, 55–77.

[28] Moolgavkar, S.H. & Knudson, A.G. (1981). Mutation and cancer: a model for human carcinogenesis, *Journal of the National Cancer Institute* **66**, 1037–1052.

[29] Moolgavkar, S.H., Stevens, R.G. and Lee, J.A.H. (1979). Effect of age on incidence of breast cancer in females, *Journal of the National Cancer Institute* **62**, 493–501.

[30] Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates, *International Journal of Epidemiology* **14**, 124–129.

[31] Osmond, C. & Gardner, M.J. (1982). Age, period and cohort models applied to cancer mortality rates, *Statistics in Medicine* **1**, 245–259.

[32] Robertson, C. & Boyle, P. (1986). Age, period, and cohort models: the use of individual records, *Statistics in Medicine* **5**, 527–538.

[33] Rogers, W.L. (1982). Estimable functions of age, period, and cohort effects, *American Sociology Review* **47**, 774–796.

[34] Roush, G.C., Schymura, M.J., Holford, T.R., White, C. & Flannery, J.T. (1985). Time period compared to birth cohort in Connecticut incidence rates for twenty-five malignant neoplasms, *Journal of the National Cancer Institute* **74**, 779–788.

[35] Roush, G.C., Holford, T.R., Schymura, M.J. & White, C. (1987). *Cancer Risk and Incidence Trends, The Connecticut Perspective*. Hemisphere, New York.

[36] Tango, T. & Kurashina, S. (1987). Age, period and cohort analysis of trends in mortality from major diseases in Japan, 1955 to 1979: peculiarity of the cohort born in the early Showa Era, *Statistics in Medicine* **6**, 709–726.

[37] Tarone, R.E. & Chu, K.C. (1992). Implications of birth cohort patterns in interpreting trends in breast cancer rates, *Journal of the National Cancer Institute* **84**, 1402–1410.

[38] Tarone, R.E. & Chu, K.C. (1996). Evaluation of birth cohort patterns in population disease rates, *American Journal of Epidemiology* **143**, 85–91.

[39] US Public Health Service (1979). *Smoking and Health: A Report of the Surgeon General. US Department of Health*, Education and Welfare, Public Health Service, Washington.

[40] Weinkam, J.J. & Sterling, T.D. (1991). A graphical approach to the interpretation of age–period–cohort data, *Epidemiology* **2**, 133–137.

[41] Wickramaratne, P.J., Weissman, M.M. Leaf, P.J. & Holford, T.R. (1989). Age, period and cohort effects on the risk of major depression: results from five United States communities, *Journal of Clinical Epidemiology* **42**, 333–343.

[42] Zheng, T., Holford, T.R. Chen, Y., Ma, J.Z., Flannery, J., Liu, W., Russi, M. & Boyle, P. (1996). Time trend and age–period–cohort effects on incidence of thyroid cancer in Connecticut, *International Journal of Cancer* **67**, 504–509.

THEODORE R. HOLFORD

Aging Models

Models for aging, senescence, and biologic lifespan have come under intensive scrutiny in recent years due to increasing general scientific and medical interest in aging [7]. Projections of "oldest-old mortality", i.e. the mortality of those aged 85 or more in human populations, are important to plan for the future of society. The future of health care, pension, and social security systems critically depends on an assessment of oldest-old mortality and lifespan [9, 22, 25]. Currently, the most long-lived, well-documented person was Jeanne Calmont of France, who died at age 122 in 1997, providing an example of how far human lifespan can extend. Of particular interest recently has been the study and analysis of an observed slowing of mortality that occurs for the survivors into the ranks of the oldest-old (see [5, 15, 29, 32]).

Characteristic features of lifetime data in aging and **demographic** research [14], which distinguish such data from typical survival data such as obtained in cancer clinical trials, include the following. (i) All individuals in a **cohort** enter the study simultaneously (no staggered entry). Entry into the study occurs at a fixed age, often at birth. (ii) A cohort is observed until the last member is dead, and lifetime normally refers to the entire lifespan of each individual. (iii) **Censoring** and **truncation** are only seldom encountered. (iv) The data are usually aggregated in a **life table**. The exact time of death is rarely recorded. The aggregation intervals vary between days (as in biodemographic studies) to five year intervals (as in some human studies). The aggregated nature of the data must be taken into account for model fitting and inference. (v) The behavior of mortality in the right tail, including the study of extreme lifetimes, is of particular interest (oldest-old mortality). (vi) Large initial total cohort sizes are necessary to assure a sufficiently large group of oldest-old. In some studies, this is achieved by simultaneously observing many smaller cohorts, a fact that needs to be taken into account in the analysis. (vii) Individual-level Covariates and events may be recorded at irregular observation intervals before death.

In spite of these particular features, the quantification of aging and mortality is based on the same concepts as are used in survival analysis (*see* **Survival Distributions and Their Characteristics**). The dynamics of mortality in dependence on age are usually measured in terms of the force of mortality, also referred to as instantaneous death rate, **hazard rate**, or hazard function. The hazard rate or force of mortality at age t is defined as

$$\lambda(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t + \Delta > T \geq t | T \geq t),$$

where T denotes lifetime (assumed to be a continuous random variable). The force of mortality $\lambda(t)$ describes the instantaneous risk of dying at age t.

Other functions which equivalently characterize the lifetime distribution (see Cox [6]) are:

1. The survival function

$$\overline{F}(t) = P(T \geq t), \quad t \geq 0,$$

which is related to $\lambda(t)$ via

$$\overline{F}(t) = \exp\left[- \int_0^t \lambda(u)\, \mathrm{d}u \right],$$

$$\lambda(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log[\overline{F}(t)].$$

2. The probability density function

$$f(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T < t + \Delta), \quad t \geq 0,$$

which is related to $\lambda(t)$ via

$$f(t) = \lambda(t) \exp\left[- \int_0^t \lambda(u)\, \mathrm{d}u \right],$$

$$\lambda(t) = \frac{f(t)}{\displaystyle\int_0^t f(u)\, \mathrm{d}u}.$$

3. The remaining **life expectancy** function, also referred to as expected residual life function,

$$r(t) = \mathrm{E}(T - t | T \geq t), \quad t \geq 0,$$

where $\mathrm{E}(\cdot|\cdot)$ denotes conditional expectation. The remaining life expectancy function is related to $\lambda(t)$ via

$$r(t) = \int_t^\infty \exp\left[- \int_t^u \lambda(v)\, \mathrm{d}v \right] \mathrm{d}u,$$

$$\lambda(t) = \left\{ \frac{\mathrm{d}}{\mathrm{d}t}[r(t)] + 1 \right\} \Big/ r(t).$$

Note that $r(0)$ is the expected lifetime from birth for individuals. The relative merits of hazard function estimates and remaining life expectancy function estimates are discussed below in the context of a data example.

Mortality data typically are in the form of a life table $(n_i, d_i), i \geq 1$, where for time intervals $[\Delta(i-1), \Delta i]$ of length Δ, n_i is the number of subjects alive and under observation at the beginning of the interval, and d_i is the number of observed deaths in the interval. For data exploration, one usually computes and plots the central death rate or actuarial estimate for such data. Evaluated at $t_i = \Delta(i - 1/2)$, the central death rate is

$$\tilde{q}_c(t_i) = \frac{2d_i}{\Delta(n_i + n_{i+1})}.$$

The central death rate has a rapidly rising variance as the numbers of subjects at risk n_i declines.

A common approach for further analysis is to fit a parametric model to the data, using the maximum likelihood method (*see* **Parametric Models in Survival Analysis**). Given a parametric model with hazard rate (force of mortality) $\lambda(t, \boldsymbol{\theta})$, where $t \geq 0$ denotes age and $\boldsymbol{\theta} \in \mathcal{R}^p$, $p \geq 1$, is a parameter vector, the likelihood function is found to be

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{\infty} \{F(\Delta i, \boldsymbol{\theta}) - F[\Delta(i-1), \boldsymbol{\theta}]\}^{d_i},$$

where $F(t, \boldsymbol{\theta}) = 1 - \exp[-\int_0^t \lambda(u, \boldsymbol{\theta})\, \mathrm{d}u]$ is the distribution function. The maximum likelihood estimator is then, as usual, $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathcal{R}^p} L(\boldsymbol{\theta})$. Modified versions of L are used for censored data or special sampling schemes, such as occur in the nematode study of Brooks et al. [3].

The standard parametric model for aging and mortality data is the Gompertz model [10], which stipulates that the force of mortality rises exponentially,

$$\lambda(t) = \beta_0 \exp(\beta_1 t), \quad \beta_0, \beta_1 > 0, t > 0,$$

or linearly on the log scale, $\log[\lambda(t)] = \log \beta_0 + \beta_1 t$. This is the hazard rate of an extreme value distribution. The Gompertz model can be motivated in various ways. It is, for instance, obtained as a special case in random walk models of aging, considering physiological age states in a state space [33].

The Gompertz model can also be derived via the "disposable soma theory" of aging, which assumes that reproduction takes away resources for repair and thus leads to faster senescence [2]. Biological consequences and comparisons of Gompertz and Weibull models were studied in [23].

Fitting the Gompertz model to mortality data from various sources, some authors found that, when considering a sample of fitted parameters, the linear relationship $\log \hat{\beta}_0 = c_0 - c_1 \hat{\beta}_1$ with positive constants c_0, and c_1 [24] appears to hold. While this indicates simply a negative correlation between the two fitted parameters, far-reaching consequences have been claimed, including the existence of an upper limit to human lifespan [8].

Other fairly flexible parametric aging models which can be derived from various assumptions are the two-parameter Weibull model [31]

$$\lambda(t) = \frac{\beta_0}{\beta_1} \left(\frac{t}{\beta_1}\right)^{\beta_0 - 1}, \quad \beta_0 > 1, \beta_1 > 0, t > 0,$$

with

$$\log(\lambda(t)) = \log \beta_0 - \beta_0 \log \beta_1 + (\beta_0 - 1) \log t,$$

and the Gompertz–Makeham model [16]

$$\lambda(t) = \beta_0 + \beta_1 \exp(\beta_2 t), \quad \beta_0, \beta_1, \beta_2 > 0.$$

The latter is an extension of the Gompertz model, adding a baseline mortality.

Many other parametric distributions may reasonably be fitted to lifetime data; for instance, a shock model in which the force of mortality is composed of a sum of parameterized Gaussian-type peaks [35]. Examples of Gompertz and Gompertz–Makeham hazard rates (force of mortality) are shown in Figure 1.

Extensions of Gompertz, Gompertz–Makeham, and other models have been proposed [27] by assuming that the individuals vary randomly in their **frailty**, a variation that could be rooted in genetics or environment. Such random effects models are also referred to as heterogeneous or compositional models. An example is the Gompertz model with frailty,

$$\lambda(t|Z) = Z \exp(\beta_1 t), \quad \beta_1 > 0,$$

where Z is a random variable with a **gamma** or **inverse Gaussian distribution**. This model has been discussed by various authors [1, 13].

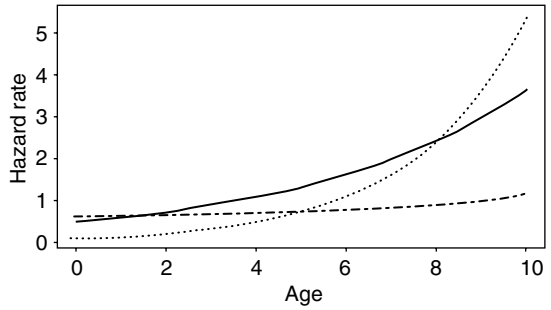

Figure 1 Force of mortality shown for two Gompertz models with parameters $(\beta_0, \beta_1) = (0.5, 0.2)$(———) $(\beta_0, \beta_1) = (0.1, 0.4)$(· · · ·), and a Gompertz–Makeham model with parameters $(\beta_0, \beta_1, \beta_2) = (0.5, 0.1, 0.2)$ (– · – · –)

The idea underlying the frailty approach is that each individual is following its own frailty-determined trajectory of mortality. The ensemble of these random forces of mortality determines the population force of mortality. It is important to note that one cannot conclude that an individual force of mortality is the same as the population force of mortality. This mistake has been made many times in the literature, starting with Gompertz [10]. Indeed, it was demonstrated by Vaupel & Yashin [26] that even if all individuals in a cohort have monotone increasing force of mortality, the population force of mortality may nevertheless be decreasing. The distinction between population and individual force of mortality is of great importance for an understanding of the biologic determinants of aging.

The problems of overfitting and lack of interpretability associated with parametric modeling of aging and mortality have led to recent renewed interest in nonparametric modeling. This requires primarily the nonparametric estimation of the force of mortality from life table data. This problem has a history of more than a century, starting with Gram [11] and important contributions by Hoem [12]. Denote a generic smoother S, like a smoothing **spline** or local linear fit or kernel smoother, based on weight functions W_i and smoothing scatterplot data (t_i, Y_i), by

$$S(t, (t_i, y_i)_{i=1,\ldots,n}) = \sum_{i=1}^{n} W_i(t) Y_i.$$

Letting $\hat{q}(t) = S[t, (t_i, \tilde{q}_i)_{i=1,\ldots,n}]$, which is a smoothed version of the central death rate, a reasonably smoothed force of mortality can be obtained as ([19, 30])

$$\hat{\lambda}(t) = -\log[1 - \hat{q}(t)].$$

An analogous estimate for the remaining life expectancy function is obtained by

$$\hat{r}(t) = S\left\{t, \left(t_i, \frac{1}{n_i}\sum_{j \in N_i}(T_j - t_i)\right)_{i=1,\ldots,p}\right\},$$

where T_j is the lifetime of the jth subject, N_i is the index set of those subjects still at risk at t_i, $N_i = \{j : T_j \geq t_i\}$, and n_i is the number of elements of N_i. The implementation of such smoothers requires choice of a smoothing parameter [18].

As an example for the application of these nonparametric procedures, and to demonstrate how force of mortality and remaining life expectancy function complement each other, consider data on cohorts of female medflies. For details of these data and their analysis, see Müller et al. [20]. Force of mortality and remaining life expectancy function estimates are shown in Figure 2. They have been computed for two groups of medflies: a protein-deprived group (solid lines) and a full-diet group (dashed lines).

A striking finding is a prominent peak in the force of mortality at around day ten, which appears for the protein-deprived group only. This peak corresponds

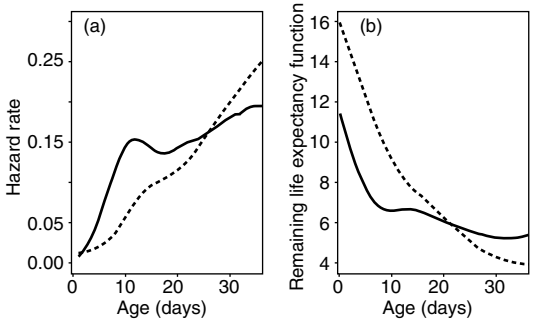

Figure 2 Nonparametric function estimates for force of mortality and remaining life expectancy, based on data of two cohorts of protein-deprived (———) and full diet (- - - -) female medflies of more than 100 000 medflies each. Estimates of hazard rate (force of mortality) (a) and of remaining life expectancy function (b): (a) is a modified version of Figure 1 of Müller et al. [20]

to a sharp drop in the remaining life expectancy function at around the same time. It can be interpreted as the signature of a vulnerable period for female medflies at reproduction. Since such a phenomenon would be hard to anticipate with parametric aging models, this example demonstrates that nonparametric modeling, properly implemented, is capable of establishing novel features by letting the data speak for themselves.

Future work on modeling of aging data is needed to address a multitude of open questions. This is an area of research with impact for society, medicine, and life sciences, in which statisticians can and should make important contributions. Open problems concern the further development of dynamic stochastic models [17, 34], and the incorporation of life history and **covariate** information for lifetime data [4] Addressing these problems requires the development of innovative lifetime regression models, that extend the classical tools of models for the deceleration of aging of the oldest–old [28] and models that include and predict secular trends in mortality, in particular the continuing increases in life expectancy [21, 28].

References

[1] Aalen, O.O. (1988). Heterogeneity in survival analysis, *Statistics in Medicine* **7**, 1121–1137.

[2] Abrams, P.A. & Ludwig, D. (1995). Optimality theory, Gompertz' law and the disposable soma theory of senescence, *Evolution* **49**, 1055–1066.

[3] Brooks, A., Lithgow, G. & Johnson, T. (1994). Rates of mortality in populations of *Caenorhabditis elegans*, *Science* **263**, 668–670.

[4] Capra, W.B. & Müller, H.G. (1997). Time accelerated models for response curves, *Journal of the American Statistical Association* **92**, 72–83.

[5] Carey, J.R., Liedo, P., Orozco, D. & Vaupel, J.W. (1992). Slowing of mortality rates at older ages in large medfly cohorts, *Science* **258**, 457–461.

[6] Cox, D.R. & Oakes, D.R. (1990). *Analysis of Survival Data*. Chapman & Hall, London.

[7] Finch, C.E. (1990). *Longevity, Senescence and the Genome*. University of Chicago Press, Chicago.

[8] Fries, J.F. (1980). Aging, natural death and the compression of morbidity, *New England Journal of Medicine* **303**, 130–135.

[9] Gavrilov, L.A. & Gavrilova, N.S. (1991). *Biology of Life Span: A Quantitative Approach*. Harwood Associates, Chur, Switzerland; Harwood Academic Publishers, New York.

[10] Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, *Philosophical Transactions* **27**, 510–519.

[11] Gram, J.P. (1883). Ueber Entwickelung reeller Functionen in Reihen mittelst der Methode der Kleinsten Quadrate, *Journal of Mathematics* **94**, 41–73.

[12] Hoem, J. (1976). On the optimality of modified minimum chi-square analytic graduation, *Scandinavian Journal of Statistics* **3**, 89–92.

[13] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer.

[14] Juckett, D.A. & Rosenberg, B. (1993). Comparison of the Gompertz and Weibull functions as descriptors for human mortality distributions and their intersection, *Mechanisms of Aging and Development* **69**, 1–31.

[15] Koenker, R. & Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis, *Journal of the American Statistical Association* **96**(454), 458–468.

[16] Makeham, W.M. (1860). On the law of mortality and the construction of annuity tables, *Journal of the Institute of Actuaries* **8**, 301–310.

[17] Manton, K.G. (1999). Dynamic paradigms for human mortality and aging, *Journals of Gerontology Series A-Biological Sciences and Medical Sciences* **54**(6), B247–B254.

[18] Müller, H.-G. & Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths, *Biometrics* **50**, 61–76.

[19] Müller, H.-G., Wang, J.-L. & Capra, B. (1997). From lifetables to hazard rates: the transformation approach, *Biometrika*, **84**, 881–892.

[20] Müller, H.-G., Wang, J.-L., Capra, B., Liedo, P. & Carey, J.R. (1997). Early mortality surge in protein-deprived medflies causes reversal of sex differential of life expectancy in Mediterranean fruit flies, *Proceedings of the National Academy of Sciences* **94**, 2762–2765.

[21] Oeppen, J. & Vaupel, J.W. (2002). Demography–Broken limits to life expectancy, *Science* **296**(5570), 1029–1031.

[22] Perls, T.T. (1995). The oldest-old, *Scientific American*, **1**, 70–75.

[23] Ricklefs, R.E. & Scheuerlein, A. (2002). Biological implications of the Weibull and Gompertz models of aging, *Journals of Gerontology Series A-Biological Sciences and Medical Sciences* **57**(2), B69–B76.

[24] Riggs, J.E. & Mileccia, R.J. (1992). Using the Gompertz-Strehler model of aging and mortality to explain mortality trends in industrialized countries, *Mechanisms of Aging and Development* **65**, 217–228.

[25] Suzman, R.M., Willis, D. & Manton, K. (1992). *The Oldest Old*. Oxford University Press, New York.

[26] Vaupel, J.W. & Yashin, A.I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics, *American Statistician* **39**, 176–185.

[27] Vaupel, J.W., Manton, K.G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **16**, 439–454.

[28] Vaupel, J.W., Carey, J.R., Christensen, K., Johnson, T.E., Yashin, A.I., Holm, N.V., Iachine, I.A., Kannisto, V., Khazaeli, A.A., Liedo, P., Longo, V.D., Zeng, Y., Manton, K.G. & Curtsinger, J.W. (1998). Biodemographic trajectories of longevity, *Science* **280**(5365), 855–860.

[29] Wachter, K.W. (1999). Evolutionary demographic models for mortality plateaus, *Proceedings of the National Academy of Sciences of the United States of America* **96**(18), 10544–10547.

[30] Wang, J.L., Müller, H.G., Capra, W.B. & Carey, J.R. (1994). Rates of mortality in populations of *Caenorhabditis elegans*, *Science* **266**, 827–828.

[31] Weibull, W.A. (1951). A statistical distribution function of wide applicability, *Journal of Applied Mechanics* **18**, 293–297.

[32] Weitz, J.S. & Fraser, H.B. (2001). Explaining mortality rate plateaus, *Proceedings of the National Academy of Sciences of the United States of America* **98**(26), 15383–15386, 2001.

[33] Woodbury, M.A. & Manton, K. (1977). A random walk model of human mortality and aging, *Theoretical Population Biology* **11**, 37–48.

[34] Yashin, A.I. & Manton, K.G. (1997). Effects of unobserved and partially observed covariate processes on system failure: A review of models and estimation strategies, *Statistical Science* **12**(1), 20–34.

[35] Zelterman, D. & Curtsinger, J.W. (1995). Survival curves subjected to occasional insults, *Biometrics* **51**, 1140–1146.

(*See also* **Gerontology and Geriatric Medicine**)

HANS-GEORG MÜLLER

Agreement, Measurement of

Reliability of measurements taken by clinicians or diagnostic devices is fundamental to ensure efficient delivery of health care. Consequently, clinicians and health professionals are becoming more aware of the need for evaluating the extent to which measurements are error-free and the degree to which clinical scores might deviate from the truth. Specifically, the recorded ratings or findings made during clinical appraisal need to be consistent, whether recorded by the same clinician on different occasions or by different clinicians within a short period of time. The consistency of the ratings reflect agreement, which is a distinct type of association. A clearly defined measure of agreement describes how consistent one clinician's rating of a patient is with what other clinicians have reported (interclinician reliability), or how consistently a clinician rates a patient over a number of occasions (intraclinician reliability). High agreement is indicative of how reproducible the results might be at different times or at other laboratories [9].

Investigators often have some latitude on the choice of how to measure the characteristics of interest in assessing agreement between raters. One practical aspect of this decision may relate to the implications of measuring the characteristic on a continuous or categorical scale. For categorical measurements, or when the levels of a continuous characteristic are categorized, the **kappa** coefficient and its variants seem to be appropriate tools to measure agreement among raters. The kappa coefficient gives an estimate of the proportion of agreement above chance [12]. For interval or continuous scale measurements, we estimate interclinician reliability with the "intraclass correlation coefficient" (ICC) (*see* **Correlation**).

In this paper we review some of the well-known indices of agreement, the conceptual and statistical issues related to their estimation, and interpretation for both categorical and interval scale measurements.

Cohen's Kappa and Darroch's Measure of Category Distinguishability

Let n subjects be classified into c **nominal** scale categories $1, \ldots, c$ by two clinicians using a single rating protocol, and let π_{jk} be the joint probability that the first clinician classifies a subject as j and the second clinician classifies the subject as k. Let $\pi_{j\cdot} = \sum_k \pi_{jk}$, and $\pi_{\cdot k} = \sum_j \pi_{jk}$. There are two questions that need to be addressed; the first is related to the interclinician **bias**, or the difference between two sets of **marginal probabilities** $\pi_{j\cdot}$ and $\pi_{\cdot j}$, while the second is related to the magnitude of $\sum \pi_{jj}$, or the extent of agreement of the two clinicians about individual subjects or objects.

Cohen [12] proposed that a coefficient of agreement be defined by

$$\kappa = \frac{\sum_{j=1}^{c} (\pi_{jj} - \pi_{j\cdot}\pi_{\cdot j})}{1 - \sum_{j=1}^{c} \pi_{j\cdot}\pi_{\cdot j}} \tag{1}$$

as a measure of agreement between two raters or clinicians. Cohen's justification was that the sum of the diagonal probabilities, $\pi_0 = \sum \pi_{jj}$, is the percentage of agreement between the two raters. Since $\pi_e = \sum \pi_{j\cdot}\pi_{\cdot j}$ is the probability of random or chance agreement, it should be subtracted from π_0. The division by $1 - \pi_e$ results in a coefficient whose maximum value is 1, which is attained when $\pi_{jk} = 0$, $j \neq k$. An estimate of κ is obtained by substituting n_{jk}/n for π_{jk}, where n_{jk} is the observed frequency for the j, kth cell.

The definition of κ given in (1) is suitable for $c \times c$ tables with nominal response categories. For ordinal response, Cohen [13] introduced the weighted kappa, κ_w, to allow each cell j, k to be weighted according to the degree of agreement between the jth and kth categories. Assigning weights $0 \leq d_{jk} \leq 1$ to the j, k cell with $d_{jj} = 1$, Cohen's weighted kappa is

$$\kappa_w = \frac{\sum_{j=1}^{c}\sum_{k=1}^{c} d_{jk}(\pi_{jk} - \pi_{j\cdot}\pi_{\cdot k})}{1 - \sum_{j=1}^{c}\sum_{k=1}^{c} d_{jk}\pi_{j\cdot}\pi_{\cdot k}}. \tag{2}$$

The large sampling distribution of the estimated κ_w has been investigated by Everitt [24] and Fleiss et al. [35]. The equivalence of κ_w to the ICC was shown by Fleiss & Cicchetti [32], Fleiss & Cohen [33], Krippendorff [46], and Schouten [61, 62].

In many circumstances the categories into which subjects are classified do not have clear objective definitions. As a result, clinicians may interpret the category definitions differently and the categories may not be completely distinguishable from each other, even by the same clinician. Darroch & McCloud [16] defined the degree of distinguishability from the joint classification probabilities for two clinicians. They derived an average measure of degree of distinguishability, δ, as

$$\delta = 2\sum_{j<k} \frac{\pi_{jj}\pi_{kk} - \pi_{jk}\pi_{kj}}{\pi_{jj}\pi_{kk}} \Big/ c(c-1). \qquad (3)$$

We estimate δ by substituting n_{jk}/n for π_{jk}.

Aickin's Alpha for Nominal Responses

It is evident from the definition of κ that it represents a fraction of subjects not classified in some category by chance; that is, they are classified for reasons other than chance. Aickin [2] attempted to make the notion of "agreement for cause" concrete by introducing another measure of agreement termed "α-measure", later referred to as Aickin's α. He based his argument on the idea that subjects to be classified are drawn from a population which is a mixture of two subpopulations. The first subpopulation consists of subjects which are difficult to classify, so that agreement between the two raters will be by chance alone. The second subpopulation consists of subjects that are easy to classify, so the raters will always agree (agreement for cause). The proposed parameter α is defined as the fraction of the entire population that consists of items that are classified identically for cause rather than by chance.

Interestingly, a case for Aickin's α can be made from reviewing the literature on the reliability of clinical methods. Koran [43] reported a study by Conn et al. [14] on physicians' agreement in diagnosing varices by esophagoscopy. In that study, two "experienced endoscopists" examined 39 male cirrhotic patients for esophageal varices during the same esophagoscopic examination. When the physicians disagreed, the one not reporting varices usually reported prominent mucosal folds, with which varices may be confused. The authors noted that "most diagnostic difficulties occur in the patients in whom esophageal varices are small". Clearly, the more prominent a sign, the easier it should be to recognize. One may argue, then, that in the population of male cirrhotic patients, the fraction with prominent signs is α (those which are easy to classify), while $1-\alpha$ have less prominent signs and therefore are difficult to diagnose.

According to Aickin's setup, let $\pi_{\mathrm{r}}(j)$ and $\pi_{\mathrm{c}}(j)$, $j = 1, 2, \ldots, c$, be any two probability distributions on the classification categories. The joint distribution π_{jk}, governing the classification of a subject by the first clinician in category j, and the second clinician in category k, is defined by

$$\pi_{jk} = (1-\alpha)\pi_{\mathrm{r}}(j)\pi_{\mathrm{c}}(k) + \alpha s^{-1} d_{jk}\pi_{\mathrm{r}}(j)\pi_{\mathrm{c}}(k), \qquad (4)$$

where $d_{jk} = 1$ if a row classification of j and column classification of k are considered to be in agreement; $d_{jk} = 0$ otherwise, and $s = \sum d_{jk}\pi_{\mathrm{r}}(j)\pi_{\mathrm{c}}(k)$.

This can be seen as a mixture of two discrete distributions. The first occurs with probability $1-\alpha$, and is a distribution under which the two classifications are independent with marginal probabilities $\pi_{\mathrm{r}}(j)$ and $\pi_{\mathrm{c}}(k)$ for the two raters. The second which occurs with probability α is a distribution under which there can only be perfect agreement. In this manner the parameter α acquires its meaning as the fraction of the population that produces "agreement for cause" between the two clinicians. A consequence of (4) is

$$\alpha = \frac{\sum_{jk} d_{jk}\pi_{jk} - \sum_{jk} d_{jk}\pi_{\mathrm{r}}(j)\pi_{\mathrm{c}}(k)}{1 - \sum_{jk} d_{jk}\pi_{\mathrm{r}}(j)\pi_{\mathrm{c}}(k)}. \qquad (5)$$

This shows that the parameter α follows the pattern of kappa-like statistics given in (1). The fundamental difference lies in the fact that $\pi_{\mathrm{r}}(j)$ and $\pi_{\mathrm{c}}(k)$ are not marginal table probabilities, but rather the marginal probabilities of the subpopulation of difficult-to-classify patients.

The above model contains $2(c-1)$ marginal probabilities and the α parameter (total of $2c-1$ parameters). Since the saturated model (*see* **Generalized Linear Model**) contains $c^2 - 1$ parameters, the number of **degrees of freedom** are $(c^2-1) - (2c-1) = c(c-2)$. For model fitting by **maximum likelihood** with application to cancer registry data (*see* **Disease Registers**), see Aickin [2].

Monotonic Agreement

Other measures of agreement for ordinal data, which do not involve any assumptions concerning the exact size of the interval between pairs of ordinal classes, are Kendall's τ [42] and **Goodman and Kruskal's** γ [37]. These two statistics measure monotonic agreement. The basic building blocks of most ordinal measures are the concepts of concordant and discordant pairs of observations. For example, select two subjects at random from the $c \times c$ table and let X_l and Y_l represent the lth subject's score by the first and the second rater; X_m and Y_m stand for the corresponding score for the mth subject. A pair is said to be concordant if one of the two subjects is higher (or lower) on both X and Y than the other person. Specifically, if $X_l > X_m$ and $Y_l > Y_m$, or $X_l < X_m$ and $Y_l < Y_m$, then the pair $(X_l, Y_l)(X_m, Y_m)$ is concordant. The simplest way to calculate the total number of concordant pairs, N_C, is to multiply each cell frequency, n_{jk}, by the total number of subjects falling in cells lying to the right and below it and then summing the results. If, on the other hand, $X_l > X_m$ and $Y_l < Y_m$, or $X_l < X_m$ and $Y_l > Y_m$, then the pair is discordant. The total number of discordant pairs in a table, N_D, is obtained by multiplying each cell frequency, n_{jk}, by the total number of subjects in the cells lying to the left and below it, and then summing the results. If there are tied observations, they are given the average of the ranks they would have received if there had been no ties. The formula for τ is

$$\tau = \frac{N_C - N_D}{\left\{\left[\binom{n}{2} - T_1\right]\left[\binom{n}{2} - T_2\right]\right\}^{1/2}}, \qquad (6)$$

where $T_1 = \frac{1}{2}\sum_{j=1}^{c} n_{j\cdot}(n_{j\cdot} - 1)$, $n_{j\cdot}$ being the number of tied observations on the jth group of ties of rater 1, and $T_2 = \frac{1}{2}\sum_{j=1}^{c} n_{\cdot j}(n_{\cdot j} - 1)$, $n_{\cdot j}$ being the number of tied observations in the jth group of ties of rater 2.

Binary Responses: Agreement in the 2 × 2 Table

One of the most familiar and extensively studied types of cross-classification in medical research is the **2 × 2 table**, as shown in Table 1.

Table 1 Classification probabilities into two categories by two raters

		Clinician 1		
		Disease	No disease	
Clinician 2	Disease	π_{11}	π_{12}	$\pi_{1\cdot}$
	No disease	π_{21}	π_{22}	$\pi_{2\cdot}$
		$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	

In addition to the simplicity of computing measures of agreement in such tables, many such measures reduce to functions of the cross-product ratio, $\Phi = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$, which is the most widely known measure of association in epidemiologic studies.

Recall that a crude measure of agreement is $\pi_0 = \pi_{11} + \pi_{22}$, which is estimated by $\hat{\pi}_0 = (n_{11} + n_{22})/n$. This measure is equivalent to Dunn & Everitt's [22] "matching coefficient of numerical taxonomy". If the clinicians are diagnosing a rare condition, the fact that they agree on the absence of the condition (the frequency, n_{22}) may be considered uninformative. A better measure of agreement in this case is estimated by

$$s = \frac{n_{11}}{n_{11} + n_{12} + n_{21}}. \qquad (7)$$

This is the Jaccard coefficient of numerical taxonomy [20, 22]. Before we discuss other indices of agreement for **binary data**, we show the relationship between **association** and agreement. Such a relationship is harder to demonstrate when the number of categories is larger than two.

First, the Pearson product moment **correlation** in a 2×2 table is

$$\rho = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{(\pi_{1\cdot}\pi_{2\cdot}\pi_{\cdot 1}\pi_{\cdot 2})^{1/2}}, \qquad (8)$$

and its sample estimate is

$$\hat{\rho} = \frac{n_{11}n_{22} - n_{12}n_{21}}{(n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2})^{1/2}}. \qquad (9)$$

The value of ρ varies between -1.0 and 1.0. It equals zero if the two sets of ratings are independent. From Eq. (8), $\rho = 1.0$ if $\pi_{12} = \pi_{21} = 0$, and $\rho = -1.0$ if $\pi_{11} = \pi_{22} = 0$. In this sense, the correlation coefficient gives both the direction and strength of association.

If we standardize the 2×2 table so that both row and column marginal totals are (1/2, 1/2) while the cross-product ratio ϕ remains unchanged, the adjusted cell probabilities are

$$\pi_{11}^* = \pi_{22}^* = \frac{1}{2}\left(\frac{\phi^{1/2}}{\phi^{1/2}+1}\right)$$

and

$$\pi_{12}^* = \pi_{21}^* = \frac{1}{2}\left(\frac{1}{\phi^{1/2}+1}\right)$$

[7, p. 379]. It can be shown that

$$\rho = \frac{\phi - 1}{(\phi^{1/2}+1)^2}.$$

Another well-known measure of association is Yule's Q. It is defined as

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}} = \frac{\phi - 1}{\phi + 1}. \tag{10}$$

Clearly, the cell probabilities in Table 1 can be reparameterized and rewritten as functions of any of the above measures of association. In terms of ρ we have

$$\begin{aligned}
\pi_{11} &= \pi_{1\cdot}\pi_{\cdot 1} + \rho\omega, \\
\pi_{22} &= \pi_{2\cdot}\pi_{\cdot 2} + \rho\omega, \\
\pi_{12} &= \pi_{1\cdot}\pi_{\cdot 2} - \rho\omega, \qquad (11)
\end{aligned}$$

and

$$\pi_{21} = \pi_{\cdot 1}\pi_{2\cdot} - \rho\omega,$$

where $\omega = (\pi_{1\cdot}\pi_{2\cdot}\pi_{\cdot 1}\pi_{\cdot 2})^{1/2}$. Shoukri et al. [66] showed that Cohen's kappa can be written as

$$\kappa = \frac{2\rho\omega}{\pi_{1\cdot}\pi_{\cdot 2} + \pi_{\cdot 1}\pi_{2\cdot}} \tag{12}$$

$$= \frac{2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})}{\pi_{1\cdot}\pi_{\cdot 2} + \pi_{\cdot 1}\pi_{2\cdot}}. \tag{13}$$

Note that, when the two raters are **unbiased** relative to each other, i.e. $\pi_{1\cdot} = \pi_{\cdot 1}$, then $\kappa = \rho$. It is also noted that perfect association ($\rho = 1$) does not generally imply perfect agreement (unless $\pi_{1\cdot} = \pi_{\cdot 1}$).

Rogot & Goldberg [59] proposed another index of agreement based on the **conditional probabilities** $\pi_{11}/\pi_{1\cdot}$, $\pi_{11}/\pi_{\cdot 1}$, $\pi_{22}/\pi_{2\cdot}$, $\pi_{22}/\pi_{\cdot 2}$. Their proposed index is

$$A_1 = \frac{1}{4}\left[\frac{\pi_{11}}{\pi_{1\cdot}} + \frac{\pi_{11}}{\pi_{\cdot 1}} + \frac{\pi_{22}}{\pi_{2\cdot}} + \frac{\pi_{22}}{\pi_{\cdot 2}}\right]. \tag{14}$$

The chance expected value of A_1 was shown by Fleiss [28] to be 1/2. Hence, their chance corrected measure is

$$M(A_1) = \frac{A_1 - \frac{1}{2}}{1 - \frac{1}{2}} = \frac{2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})}{\left(\frac{1}{\pi_{1\cdot}\pi_{2\cdot}} + \frac{1}{\pi_{\cdot 1}\pi_{\cdot 2}}\right)}.$$

Recently, Hirji & Rosove [40] argued that an ideal measure of agreement should have the following characteristics:

1. In the case of perfect agreement, it should yield a standard value, usually 1.
2. In the case of perfect disagreement, it should also yield a standard value, of -1.
3. When the two raters are independent, it should return a value of 0.

They proposed an index of agreement that satisfies the above characteristics. They defined λ_i such that

$$1 + \lambda_i = \frac{\pi_{ii}}{\pi_{i\cdot}} + \frac{\pi_{ii}}{\pi_{\cdot i}}. \tag{15}$$

Clearly, $1 + \lambda_1$ is the sum of the conditional probabilities of agreement given that the first rater classifies a patient as diseased and the conditional probability of agreement given that the second rater classifies the patient as diseased, and λ_2 has a complementary interpretation. Note that $-1 \le \lambda_i \le 1$. Hirji & Rosove [40] defined an overall measure of agreement, λ, as

$$\begin{aligned}
\lambda &= \frac{\lambda_1 + \lambda_2}{2} \\
&= 2A_1 - 1. \qquad (16)
\end{aligned}$$

It is easy to see that the chance corrected value of λ is 0, and that it satisfies the above three characteristics. The maximum likelihood estimate of λ is obtained by replacing π_{ii} by n_{ii}/n. Hirji & Rosove [40] extended their index of agreement to the case of multiple categories.

Armitage et al. [3] proposed, as another index of agreement, the **standard deviation** of the subject's total scores, where a subject scores 2 if both raters judged them positive, 1 if one observer judged a subject positive and the other negative, and 0 if both observers judged a subject negative. Their index of agreement is easily shown to be

$$\mathrm{SD}^2 = \pi_{11} + \pi_{22} - (\pi_{11} - \pi_{22})^2. \tag{17}$$

Fleiss [28] noted that the above measure is inadequate since it does not have the range of values required by the traditional index. He suggested rescaling SD^2 to become

$$RSD^2 = \frac{\pi_{11} + \pi_{22} - (\pi_{11} - \pi_{22})^2}{1 - \left(\frac{\pi_{1\cdot} + \pi_{\cdot 1}}{2} - \frac{\pi_{2\cdot} + \pi_{\cdot 2}}{2}\right)^2}, \quad (18)$$

which will have the desired range of variation. In fact, $RSD^2 = 1$ if $\pi_{12} = \pi_{21} = 0$ and $RSD^2 = 0$ if $\pi_{11} = \pi_{22} = 0$. As before, a **consistent estimator** of RSD can be obtained by replacing π_{ij} by n_{ij}/n.

Under marginal homogeneity ($\pi_{1\cdot} = \pi_{\cdot 1} = \pi$), or when the raters are deemed unbiased relative to each other (as in test–retest reliability studies), Table 1 can be rewritten as Table 2.

Thus, (11) becomes

$$\pi_{11}(\kappa) = \pi^2 + \kappa\pi(1-\pi),$$
$$\pi_{22}(\kappa) = (1-\pi^2) + \kappa\pi(1-\pi), \quad (19)$$

and

$$\pi_{12}(\kappa) = \pi_{21}(\kappa) = \pi(1-\pi)(1-\kappa).$$

The maximum likelihood estimates of π and κ are given, respectively, as

$$\hat{\pi} = \frac{2n_{11} + n_{12} + n_{21}}{n} \quad (20)$$

and

$$\hat{\kappa} = \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})}. \quad (21)$$

This estimator of κ is identical to the estimator of an intraclass correlation coefficient for 0–1 data [72, pp. 294–296; [9]], and was proposed by Scott [63] as a measure of agreement between two clinicians when their underlying base rates are the same (i.e. marginal homogeneity). Bloch & Kraemer [9] derived a $\sin^{-1}$ transformation (*see* **Delta Method**) to stabilize the variance of $\hat{\kappa}$. Calculations of confidence intervals are eased using such a transformation.

Table 2 2×2 table with marginal homogeneity

		Clinician 1		
		Disease	No disease	
Clinician 2	Disease	π_{11}	π_{12}	π
	No disease	π_{21}	π_{22}	$1-\pi$
		π	$1-\pi$	

The observed frequencies n_{11}, n_{12}, and n_{22} follow a **multinomial distribution** conditional on $n = n_{11} + n_{12} + n_{21} + n_{22}$, with estimated probabilities $\hat{\pi}_{11}(\kappa)$, $\hat{\pi}_{12}(\kappa)$, and $\hat{\pi}_{22}(\kappa)$, where we obtain $\hat{\pi}_{ij}(\kappa)$ by replacing π by $\hat{\pi}$ in (20). It follows that

$$\chi^2_G = \frac{[n_{11} - n\hat{\pi}_{11}(\kappa)]^2}{n\hat{\pi}_{11}(\kappa)} + \frac{[n_{12} - n\hat{\pi}_{12}(\kappa)]^2}{n\hat{\pi}_{12}(\kappa)} + \frac{[n_{22} - n\hat{\pi}_{22}(\kappa)]^2}{n\hat{\pi}_{22}(\kappa)} \quad (22)$$

has a limiting **chi-square distribution** with one **degree of freedom**.

Donner & Eliasziw [19] obtained corresponding two-sided **confidence limits** on κ by finding the admissible roots $(\hat{\kappa}_L, \hat{\kappa}_U)$ to the equation $\chi^2_G = \chi^2_{(1,1-\alpha)}$, which is cubic in $\hat{\kappa}$, where $\chi^2_{(1,1-\alpha)}$ is the $100(1-\alpha)$ percentile point of the chi-square distribution with one degree of freedom. They provided explicit expressions for $\hat{\kappa}_L$ and $\hat{\kappa}_U$; this method of estimation was referred to as the **goodness-of-fit** (GOF) method.

The **simulation** study conducted by Donner & Eliasziw [18] showed that the coverage levels associated with the GOF procedure are close to nominal over a wide range of parameter values (π, κ) in samples having as few as 25 subjects.

Some Remarks on the Use of Kappa

The purpose of this Section is to bring to the reader's attention some of the conceptual issues that arise when the kappa coefficient is used as an index of quality of measurements for a binary variable. Some of these issues have received attention; we mention, among others, Carey & Gottesman [11], Spitznagel & Helzer [69], Feinstein & Cicchetti [25, 26], and Thompson & Walter [70, 71].

Since the device by which subjects can be correctly classified may not be available, then neither of the two raters is a valid indicator of the true state of the subject to be classified. However, the magnitude of the simple index of chance corrected agreement between the two raters may provide a valid

interpretation of the true state of a subject. Thompson & Walter [70] showed that the kappa coefficient depends not only on the **sensitivity** and **specificity** of the two raters, but also on the true **prevalence** of the condition. They showed that, under the assumption that the classification errors are conditionally independent (an assumption that may hold if the two raters have a different biological basis for classifying subjects), kappa is given by

$$\text{kappa} = \frac{2\pi(1-\pi)(1-\theta_1-\eta_1)(1-\theta_2-\eta_2)}{p_1(1-p_2)+p_2(1-p_1)}, \tag{23}$$

where π is the true proportion having the condition, $\theta_i = 1 - \textit{specificity}$ for the ith rater, $\eta_i = 1 - \textit{sensitivity}$ for the ith rater, and $p_i = \pi(1-\eta_i) + (1-\pi)\theta_i$ is the proportion classified as having the condition according to the ith rater ($i = 1, 2$). The strong dependence of kappa on the true prevalence π complicates its interpretation as an index of agreement. Thompson & Walter [70] stated that it is not appropriate to compare two or more kappa values when the true prevalences of the conditions compared may differ. For further discussion on misinterpretation and misuse of kappa, we refer the reader to Bloch & Kraemer [9], Thompson & Walter [71], and Maclure & Willett [52]. Other issues related to modeling of kappa can be found in recent reviews by Kraemer [45] and Agresti [1].

Agreement of Multiple Raters Per Subject

In the previous section we discussed indices of agreement for present/absent characteristics as measured by two raters. Here we discuss the issue of agreement when more than two raters classify groups of subjects for dichotomous data. We distinguish between two situations: (i) when the subjects are evaluated by the same group of clinicians. This situation occurs in practice when a group of clinicians are presented with samples of slides, X-rays, or radiograms, and based on some clearly identified protocol each item is classified as having/not having the characteristic; (ii) when subjects are classified by different (possibly unequal) numbers of raters. For example [34], the subjects may be hospitalized mental patients, the studied condition may be the presence or absence of some psychological disorder, and the raters may be those psychiatry residents, out of a much larger pool, who happen to be on call when a patient is newly admitted. Not only may the particular residents responsible for one patient be different from those responsible for another, but different numbers of residents may provide diagnoses on different patients.

Let Y_{ij} represent the assessment of the ith subject by the jth rater, $(i = 1, \ldots, n; j = 1, \ldots, k)$, with $Y_{ij} = 1$ if the ith subject is judged by the jth rater to have the condition, and 0 otherwise. Let $Y_{i\cdot}$ represent the total number of raters who judged the ith subject to have the condition, and let $y_{\cdot j}$ represent the total number of subjects the jth rater judges to have the condition. Finally, let $Y_{\cdot\cdot}$ represent the total number of subjects for which the condition is judged to be present.

Since the raters differ in their sensitivities and specificities, it may be of interest to test whether these differences are statistically significant. This test is equivalent to testing the equality of the observed marginal probabilities. The appropriate test of marginal homogeneity for binary data is the use of Cochran's Q statistic [27; 20, pp. 141–142].

If we make the a priori assumption of no rater bias, then an estimate of the reliability kappa can be obtained from the **analysis of variance** (ANOVA) just as if the results were interval scores. From the ANOVA table a **variance components** estimate of reliability is

$$\hat{\rho}_{1\omega} = \frac{\text{MSBS} - \text{MSE}}{\text{MSBS} + (k-1)\text{MSE}}, \tag{24}$$

where MSBS is the between-subject mean square, MSE is the mean square error, and k is the number of raters. For a reasonably large number of subjects, $\hat{\rho}_{1\omega}$ is approximately equivalent to

$$R_1 = \frac{\text{SSBS} - (\text{SSBR} + \text{SSE})}{\text{SSBS} + \text{SSBR} + \text{SSE}}, \tag{25}$$

where SSBR is the sum of squares due to raters, SSBS is the sum of squares due to subjects, and SSE is the error sum of squares. Fleiss [28] demonstrated that, for $k = 2$, R_1 in (25) is mathematically identical to $\hat{\kappa}$ in (21). Therefore the assumption of marginal homogeneity in the calculation of chance-corrected kappa is equivalent to that of ignoring rater bias using the one-way ANOVA model for derivation of the variance component estimate of the intraclass reliability coefficient. If one were not prepared to

assume the lack of rater bias, then the appropriate model would either be a two-way **random effects** ANOVA, or a two-way mixed model [20, p. 146] (*see* **Experimental Design**). For the random effects model, the appropriate intraclass correlation can be estimated by

$$\hat{\rho}_{2\omega\mathrm{R}} = \frac{\hat{\sigma}_{\mathrm{g}}^2}{\hat{\sigma}_{\mathrm{g}}^2 + \hat{\sigma}_{\mathrm{c}}^2 + \hat{\sigma}_{\mathrm{e}}^2}, \tag{26}$$

where

$$\hat{\sigma}_{\mathrm{g}}^2 = \frac{\mathrm{MSBS} - \mathrm{MSE}}{k},$$
$$\hat{\sigma}_{\mathrm{c}}^2 = \frac{\mathrm{MSBR} - \mathrm{MSE}}{n},$$
$$\hat{\sigma}_{\mathrm{e}}^2 = \mathrm{MSE},$$

and MSBR is the "between raters" mean square. $\hat{\sigma}_{\mathrm{g}}^2, \hat{\sigma}_{\mathrm{c}}^2$, and $\hat{\sigma}_{\mathrm{e}}^2$, are, respectively, the variance component estimates for subjects, raters, and error. Fleiss [28] demonstrated that (26) is approximately the same as

$$R_2 = \frac{\mathrm{SSBS} - \mathrm{SSE}}{\mathrm{SSBS} + \mathrm{SSE} + 2\mathrm{SSBR}}. \tag{27}$$

For $k = 2$, R_2 is mathematically equivalent to the estimated value of kappa in (13), i.e.

$$R_2 = \hat{\kappa} = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{1\cdot}n_{\cdot 2} + n_{\cdot 1}n_{2\cdot}}. \tag{28}$$

In the second situation, when subjects are assigned different numbers of patients, Fleiss & Cuzick [34] extended the definition of the estimate of kappa to

$$\hat{\kappa} = 1 - \frac{1}{n(\overline{k} - 1)\hat{\pi}(1 - \hat{\pi})} \sum_{i=1}^{n} \frac{R_i(k_i - R_i)}{k_i}, \tag{29}$$

where

$$\overline{k} = \frac{1}{n}\sum_{i=1}^{n} k_i, \quad R_i = \sum_{j=1}^{k_i} y_{ij} \quad \text{and} \quad \hat{\pi} = \frac{1}{n\overline{k}} \sum_{i=1}^{n} R_i.$$

They also showed that $\hat{\kappa}$ is asymptotically (as $n \to \infty$) equivalent to the estimated intraclass coefficient

$$\hat{\kappa} = \frac{\mathrm{MSBS} - \mathrm{MSE}}{\mathrm{MSBS} + (k_0 - 1)\mathrm{MSE}}, \tag{30}$$

where

$$k_0 = \frac{1}{n-1}\left[\sum k_i - \frac{\sum k_i^2}{\sum k_i}\right]$$

(see Fleiss [30]).

Another estimate of the reliability kappa was constructed by Mak [53], and was given as

$$\tilde{\kappa} = 1 - \frac{2}{n(n-1)} \times \left[\sum_i \frac{R_i}{k_i} \sum_i \frac{k_i - R_i}{k_i} - \sum \frac{R_i(k_i - R_i)}{k_i^2}\right]. \tag{31}$$

When all the k_i are equal, the estimators $\hat{\kappa}$ of (30) and $\tilde{\kappa}$ are asymptotically equivalent as $n \to \infty$.

Agreement of Multiple Readings with Unanimous and Majority Rules

In this Section we discuss agreement from a different direction. If a single observation per subject does not produce a satisfactory value for kappa, then a sufficient number of repeated observations on each subject may produce a score close to the consensus score. Kraemer [44] has shown that, if there are k independent raters, then the reliability of the proportion of positive ratings is

$$R \simeq \frac{k\hat{\kappa}}{1 + k\hat{\kappa}}, \tag{32}$$

where $\hat{\kappa}$ is the estimated kappa when each subject is measured once by each rater.

Lachenbruch [47] introduced a **sequential** strategy for the use of the k tests. The tests are assumed to be given in a fixed order: the second test is applied after the results of the first are known, the third test is applied after the results of the second are known, and so on. We may consider one of two rules for the combination of the individual test results. To declare a subject as positive, the unanimity rule requires that all of the individual tests yield positive results. The majority rule requires that the majority of the individual tests yield positive results (which means an odd number of diagnostic tests being administered).

For example, if we have three diagnostic tests which are given in a fixed order, the unanimity rule implies that the negative individuals give the results

$(-)$, $(+-)$, $(++-)$, while the positives are $(+++)$. Assuming, for simplicity, that each of the three tests has the same specificity η and also the same sensitivity θ, then the specificity of the unanimous rule is given by

$$\text{SPEC} = 1 - (1 - \eta)^3,$$

and its sensitivity is

$$\text{SENS} = \theta^3.$$

Clearly, SENS $< \theta$, while SPEC $> \eta$.

The above derivations are based on the assumption that θ and η are constant across subjects. Lachenbruch [47] considers cases when this assumption is relaxed. Note also that the above derivations are dependent on the assumption of independence of the diagnostic tests. Lui [51] assessed the effect of the intraclass correlation ρ under the unanimity rule. When $\rho = 1$, the multiple reading procedure will yield the same sensitivity and specificity as those of a single reading. For small values of ρ, SENS $< \theta$ and SPEC $> \eta$, as stated above.

Interval Scale Agreements

In the first part of this review we were concerned with measures of agreement for categorical responses. Categories may be nominal, ordinal, or the result of categorizing a continuous variable. The advantage of such categorization makes the index of agreement easier to comprehend and interpret; a disadvantage is the dependence of the value of the index of agreement on the number of categories. Hermann & Kliebsch [39] demonstrated that quadratically weighted kappa coefficients tend to increase with the number of categories. Their findings contrast with findings by MacLure & Willett [52] for unweighted kappa coefficients, which decrease with the number of categories.

In the second part of this article, we discuss agreement for inherently continuous measurement, whereby categorization may not be advantageous. For example, if two trained nurses measure the weight of an infant to the nearest milligram, experience shows us that the two measurements will usually not be identical, and that differences of 10–20 g are not uncommon. Differences may, in part, be due to the effect of the rater and in part due to **measurement error**. Dunn [21] reported that it is not uncommon among clinical psychologists, for example, to use **linear regression** to associate the two sets of measurements, which is not appropriate when the two rating devices commit measurement error. Moreover, product–moment correlation coefficients are measures of association and should not be used as indices of agreement.

In his recent review, Dunn [21] classified reliability studies into two types. The first involves the comparison of two or more raters (or measuring instruments) and the second explicitly examines the sources of variability in measurements. The distinction between the two is not always clear-cut.

The simplest design used to assess the reliability or the agreement between sets of scores is the one-way random effects model. Suppose that we have n patients and we would like to take several measurements by a single device. How can we assess the consistency of the set of measurements taken from each patient? The one-way model stipulates that

$$Y_{ij} = \mu + s_i + e_{ij}, \tag{33}$$

where Y_{ij} is the jth measurement taken on the ith subject, μ is the bias, s_i is the subject effect, and e_{ij} is a random measurement error, assumed independent of s_i, where $s_i \sim \text{N}(0, \sigma_\text{s}^2)$ and $e_{ij} \sim \text{N}(0, \sigma_\text{e}^2)$.

The reliability estimate of R is defined as

$$R = \frac{\hat{\sigma}_\text{s}^2}{\hat{\sigma}_\text{s}^2 + \hat{\sigma}_\text{e}^2} \tag{34}$$

(see [23, 54, 31], and [4]). Here, $\hat{\sigma}_\text{s}^2$ and $\hat{\sigma}_\text{e}^2$ are the estimates of the corresponding variance components and are obtained from the one-way random effects ANOVA. This reliability estimate is the familiar estimate of the intraclass correlation coefficient (ICC) (Snedecor & Cochran [68]). It is clear, then, that a precise estimate of R depends on the precision of estimating σ_s^2 and σ_e^2. It is noted by Dunn [21] that R is not a fixed characteristic of a measuring device–it changes with the population of subjects being sampled. This is analogous to the effect of prevalence on the kappa statistic. However, the estimate is useful as an indicator of how a particular device will perform in a particular clinical setting. As can be seen from the definition of the ICC estimate in (34), low ICC occurs when the variation between subjects is low relative to that within subjects. This means that, in a typical **reliability study** it is desirable to have low differences between readings

for a given subject and large differences between subjects. Large between-subject differences usually reflect the condition where the raters are given the chance to test their skills with a full range of a measurement scale. Normal subjects, or subjects with predefined severity, are examples of a narrow range study where raters' reliability is a test only of agreement on the absence of the condition or a subsection of the disease scale, and not a test of the instrument on the full range (see [5]).

Rater Comparison

Two Raters

Bland & Altman [8] described the following clinical experiment. Data on cardiac stroke volume or blood pressure using direct measurement without adverse effects are difficult to obtain. The true values remain unknown. Instead, indirect methods are used, and a new method has to be evaluated by comparison with an established technique. If the new method agrees sufficiently well with the old, then the old may be replaced. When the two methods are compared, neither provides an unequivocally correct measurement. We need to assess the degree of agreement.

Bland & Altman [8] recommended plotting the difference between the two measurements $(Y_{i1} - Y_{i2})$ against their mean $(Y_{i1} + Y_{i2})/2$. This plot can be useful in detecting systematic bias, **outliers**, and whether the variance of the measurements is related to the mean.

Alternatively, we can simply plot Y_{i1} against Y_{i2}. We would like to see, within tolerable error, that the measurements fall on a 45° line through the origin. Lin [49] provided several graphs demonstrating how the Pearson correlation coefficient fails to detect any departure from the 45° line. For example, if the measurement taken by rater $2(Y_{i2})$ has systematic bias relative to rater 1, i.e. $Y_{i1} = Y_{i2} - c$, where c is a fixed constant, then Pearson's correlation will attain its maximum value of 1 while there is little or no agreement between Y_{i1} and Y_{i2}. The **least squares** approach may fail to detect departure from the intercept equal to 0 and slope equal to 1 (see [48; 65, p. 37].

Based on a **random sample** of n subjects, where the ith subject provides the pairs of measurements (Y_{i1}, Y_{i2}) taken by the two raters, Lin [49] constructed a measure of agreement between readings which he called "concordance correlation" (CC). Assuming that $(Y_{i1}, Y_{i2}), i = 1, 2, \ldots, n$, have a **bivariate normal distribution** with means μ_1, μ_2 and **covariance matrix**

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{35}$$

where ρ is Pearson's correlation, then the concordance correlation is defined as

$$\rho_c = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}. \tag{36}$$

Let $\beta_1 = \rho\sigma_1/\sigma_2$ and $\beta_0 = \mu_1 - \beta_1\mu_2$. Then,

$$\rho_c = \frac{2\beta_1\sigma_2^2}{(\sigma_1^2 + \sigma_2^2) + [(\beta_0 - 0) + (\beta_1 - 1)\mu_2]^2}, \tag{37}$$

and thus the sample estimate of ρ_c is

$$\hat{\rho}_c = \frac{2s_{12}}{s_1^2 + s_2^2 + (\overline{y}_1 - \overline{y}_2)^2}. \tag{38}$$

The concordance correlation has the following properties:

1. $-1 \le -|\rho| \le \rho_c \le |\rho| < 1$
2. $\rho_c = 0$ if and only if $\rho = 0$
3. $\rho_c = \rho$ if and only if $\sigma_1 = \sigma_2, \mu_1 = \mu_2$
4. $\rho_c = \pm 1$ if and only if $\rho = \pm 1, \sigma_1 = \sigma_2, \mu_1 = \mu_2$.

It is also clear from (36) that the magnitude of ρ_c is inversely related to the bias $= |\mu_1 - \mu_2|$.

As an alternative procedure for assessing agreement between the two raters, Bradley & Blackwood [10] suggested regressing $Y_i = (Y_{i1} - Y_{i2})$ on $X_i = (Y_{i1} + Y_{i2})/2$. A simultaneous test of $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ is conducted using the F statistic,

$$F(2, n-2) = \frac{\sum Y_i^2 - \text{SSReg}}{2\text{MSReg}},$$

where SSReg and MSReg are the residual sum of squares and the mean square with $n - 2$ degrees of freedom, respectively, from the regression of Y on X.

Suppose now, as in Bland & Altman [8], that each of the two raters or methods provides two replicates, as in Table 3.

Let

$$x_{ij} = \mu + s_i + \xi_{ij} \tag{39}$$

Table 3 Comparing two raters with two replicates per subject

Rater	Subject 1	2	...	n
1	x_{11}	x_{21}		x_{n1}
	x_{12}	x_{22}		x_{n2}
2	y_{11}	y_{21}		y_{n1}
	y_{12}	y_{22}		y_{n2}

and

$$y_{ij} = \mu + s_i + \eta_{ij}, \quad i = 1, 2, \ldots, n; j = 1, 2. \tag{40}$$

It is assumed that $s_i \sim \mathrm{N}(0, \sigma_s^2)$, $\xi_{ij} \sim \mathrm{N}(0, \sigma_\xi^2)$, and $\eta_{ij} \sim \mathrm{N}(0, \sigma_\eta^2)$, and that s_i, ξ_{ij}, and η_{ij} are mutually independent. As can be seen, the relative bias between the two raters, μ, is assumed constant. The above equations represent regression models where both variables are measured with error.

Dunn [21] suggested that the analysis starts with estimating the within-subjects mean squares by fitting two separate one-way ANOVAs – one for each rater. The estimated reliabilities are

$$R = \frac{\mathrm{BSMS} - \mathrm{WSMS}}{\mathrm{BSMS} + \mathrm{WSMS}} \tag{41}$$

(see [31]), where BSMS is the between subjects mean square and WSMS the corresponding within subjects mean square.

Let $x_i^* = (x_{i1} + x_{i2})/2$ and $y_i^* = (y_{i1} + y_{i2})/2$. Grubbs [38] showed that the maximum likelihood estimates of σ_s^2, σ_ξ^2, and σ_η^2 are given, respectively, as $\hat{\sigma}_s^2 = s_{xy}, \hat{\sigma}_\xi^2 = 2(s_{xx} - s_{xy})$, and $\hat{\sigma}_\eta^2 = 2(s_{yy} - s_{xy})$, where $(n-1)s_{\mathrm{ab}} = \sum_{i=1}^n (a_i - \overline{a})(b_i - \overline{b})$ and $\overline{a} = 1/n \sum_{i=1}^n a_i$.

The **null hypothesis** that the two raters are equally precise ($\mathrm{H}_0 : \sigma_\xi^2 = \sigma_\eta^2$) can be tested using a result due to Shukla [67], who showed that H_0 is rejected whenever $t_0 = r[(n-2)/(1-r^2)]^{1/2}$ exceeds $|t_{n-2,\alpha/2}|$, where $t_{\alpha/2}$ is the cutoff point in the t-table at $100(1-\alpha/2)\%$ confidence and $n-2$ degrees of freedom, and r is Pearson's correlation between u_i and v_i, $u_i = x_i^* + y_i^*$, $v_i = x_i^* - y_i^*$. Approximate $100(1-\alpha)\%$ confidence limits on the relative precision $q = \sigma_\xi^2/\sigma_\eta^2$ are

$$q_{\mathrm{U}} = \frac{b + \sqrt{c}}{a - \sqrt{c}}, \qquad q_{\mathrm{L}} = \frac{b - \sqrt{c}}{a + \sqrt{c}},$$

with $a = \hat{\sigma}_\eta^2/2$, $b = \hat{\sigma}_\xi^2/2$, and $c = [t_{\alpha/2}^2(s_{ss}s_{yy} - s_{xy}^2)]/(n-2)$.

Regression models with **errors in variables**, or **structural equations** to assess reliability of two raters, have received considerable attention from many researchers. For example, we refer the reader to Kelly [41], Linnett [50], Nix & Dunston [56], and the earlier work of Deming [17].

Example

Table 4 provides measurements derived from an experiment in microbiology. The primary aim was to determine the number of colonies of the *E. coli* 0157:H7 pathogen in contaminated fecal samples collected from 12 beef carcasses. For a given faecal sample, the number of colonies was determined by a new test (Petrifilm HEC) and by a "standard test" in two subsamples; results are recorded as the logarithm of the number of colonies (Table 4). The first two rows correspond to the repeated determinations based on the use of the standard test, and the second two rows correspond to the repeated determinations of the new test. (These data were kindly provided by Dr Christine Power from the Ontario Veterinary College, Guelph, Ontario.)

We begin the analysis by first investigating the repeatability of each of the two tests separately. Following the recommendations of Bland & Altman [8], we plotted the difference between the two against their sum (Figure 1).

Table 4 Logarithm of the number of colonies of *E. coli* 157:H7 in samples taken from 12 beef carcasses

Test (subsample)		1	2	3	4	5	6	7	8	9	10	11	12
Standard	1	2.356	2.149	2.452	2.255	2.694	2.43	2.322	2.322	2.491	2.322	2.322	2.491
	2	2.384	2.263	2.417	2.299	2.684	2.44	2.491	2.041	2.322	2.322	2.491	2.785
New test	1	2.283	2.061	2.322	2.162	2.068	2.322	2.491	2.041	2.322	2.491	2.041	2.785
	2	2.265	1.987	2.316	2.127	2.111	2.28	2.491	2.041	2.041	2.71	2.322	2.322

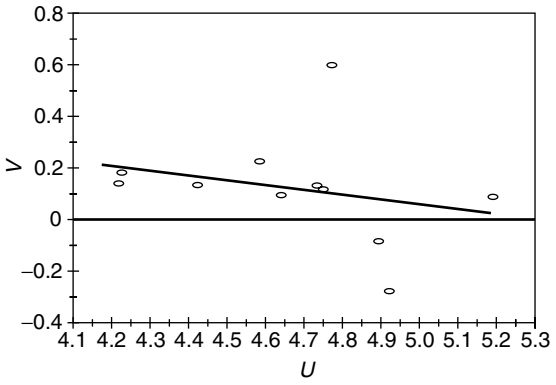

Figure 1 Plot of the difference, V, vs. the sum, U, with regression line imposed

Dunn [21] pointed out that the graph can be extremely useful in (i) allowing us to detect systematic bias and (ii) looking for outliers. The regression of the difference on the sum shows that observation #5 has a large standardized **residual**. We subsequently produced a one-way ANOVA to obtain estimates of the corresponding test–retest reliabilities using the intraclass correlation (41). The results are given in Table 5.

As can be seen, the two tests have equivalent test–retest reliability estimates. The models (39) and (40) assume that the relative biases of the two estimates are constant. This allows us to estimate the error variances using the Grubbs [38] method. If X represents the mean of the standard test log counts, and Y the mean of the new test log counts, then $\hat{\sigma}_s^2 = 0.010$, $\hat{\sigma}_\xi^2 = 0.026$, and $\hat{\sigma}_\eta^2 = 0.057$. It appears that the standard test is almost twice as precise as the new test.

Table 5 ANOVA of the logarithm of the number of colonies for the standard test and the new test

(a) Standard test (intraclass correlation = 0.61)

Source of variation	Sum of squares	df	Mean square
Cow	0.500	11	0.045
Residual	0.134	12	0.011
Total	0.634	23	

(b) New test (intraclass correlation = 0.62)

Source of variation	Sum of squares	df	Mean square
Cow	0.837	11	0.076
Residual	0.215	12	0.018
Total	1.052	23	

We now proceed to test the significance of the difference $(\sigma_\xi^2 - \sigma_\eta^2)$. Using the results of Shukla [67], we have $s_{xx} = 0.023$, $s_{yy} = 0.038$, and $s_{xy} = 0.010$. The Pearson correlation between $U = X + Y$ and $V = X - Y$ is $r = -0.27$ and $t_0 = -0.87$. We are unable to detect a significant difference between σ_ξ^2 and σ_η^2. This may be because the sample size was insufficient.

Multiple Raters

The simplest reliability study involves having each member of a sample of n subjects rated once by each member of a sample of k raters. Raters might be considered as fixed, or as a random sample drawn from a potentially larger population of raters. If raters are assumed fixed, the estimate of reliability can be obtained from the two-way mixed effects ANOVA model,

$$Y_{ij} = \mu + s_i + r_j + e_{ij}, \tag{42}$$

where Y_{ij} is the score of the jth rater on the ith patient, μ is the bias, $s_i \sim \mathrm{N}(0, \sigma_s^2)$, $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$, and the $r_i, \ldots, r_k$ are the raters effects such that $\sum r_j = 0$. Both s_i and e_{ij} are independent. From Fleiss [31], the appropriate reliability estimate is

$$R_f = \frac{n(\mathrm{BSMS} - \mathrm{WSMS})}{n(\mathrm{BSMS}) + (k-1)\mathrm{BRMS} + (k-1)(n-1)\mathrm{WSMS}}. \tag{43}$$

The components of (43) are as defined in (41), with BRMS as the between-raters mean square. If raters are assumed random, then the added assumptions that $r_j \sim \mathrm{N}(0, \sigma_r^2)$ and that r_j, s_i, and e_{ij} are mutually independent give an estimate of reliability identical to $\hat{\rho}_{2\omega R}$ in (26); simplified, this becomes

$$R_r = \frac{n(\mathrm{BSMS} - \mathrm{WSMS})}{n(\mathrm{BSMS}) + k(\mathrm{BRMS}) + (nk - n - k)\mathrm{WSMS}}. \tag{44}$$

Remarks

It is evident that **estimation** of indices of agreement for interval scale measurements is tied to estimation of variance components. The total variance is decomposed into subjects effect, raters effect (if raters are considered random), and the error component. The

traditional ANOVA, either one-way or two-way, is used when the data are balanced and/or complete (no missing data). For unbalanced data, one may use the maximum likelihood (ML) and **restricted maximum likelihood** (REML) [57, 64]. The advantage of ML or REML methods is that we obtain estimates of the **standard errors** of the estimates. Depending on the nature of the study, more complex designs other than the ANOVA may be used. Bassin et al. [6] used **crossover designs** for method comparisons. Other designs such as **balanced incomplete block designs** [27, 29] and hierarchical designs [36] can be used for nested reliability studies (*see* **Hierarchical Models**).

We emphasize that there is no single procedure which can be used to assess raters' reliability agreement. Bartko [5] pointed out that procedures for exploring agreement should be based upon the nature of the study and the purposes of various agreement measures. Table 6 summarizes the basic formulas mentioned in this article.

Computer Programs

Cyr & Francis [15] provided a menu-driven PASCAL program to compute Cohen's kappa and weighted kappa using the two types of weights, d_{jk}.

$$\text{(i) } d_{jk} = 1 - \frac{(j-k)^2}{(c-1)^2}, \quad \text{(ii) } d_{jk} = 1 - \frac{|j-k|}{c-1},$$

and other user-defined weights.

The SAS software (*see* **Software, Biostatistical**) [60] has a number of procedures referred to as "PROC" statements. PROC FREQ provides estimates of a variety of measures of association in $c \times c$ contingency tables, including an unweighted estimate of kappa. As an illustration, using Dunn's [20, p. 24] data, PROC FREQ in SAS provides the following estimates of kappa: unweighted Cohen's kappa ($= 0.21$) and weighted kappa ($= 0.58$), using the weights as described in (ii) above. Alternately, Cyr & Francis's, [15] program which uses the quadratic weights in (i), gives 0.80 as an estimate of weighted kappa. For multiple raters and the 0–1 category, estimates of kappa, in (24), (26), and (30), are obtainable using PROC ANOVA or PROC GLM in SAS. The SAS procedures provide the appropriate sum of squares and the corresponding mean squares to calculate reliability kappa estimates. It is also possible to use **StatXact** [55] to calculate Cohen's kappa or a weighted kappa statistic.

Estimates of variance components from the balanced one-way random effects, two-way random effects, and mixed effects models can be obtained using either PROC ANOVA with the RANDOM statement or PROC VARCOMP in SAS. For more complex designs with multiple levels of nesting and crossing, PROC MIXED in SAS may be used to estimate the components of variance. For unbalanced data or when some data points are missing, Robinson's [58] REML program can be used to obtain estimates of variance components and hence the intraclass correlations. With little programming experience in SAS, PROC UNIVARIATE, PROC CORR, and PROC REG are quite easy to implement and may be used to calculate the concordance correlation and Bradley & Blackwood [10] statistics.

Table 6 Summary of basic formulas

Parameter	Interpretation
Eq. (1)	Cohen's kappa – nominal-scale measure of agreement between two raters
Eq. (2)	Cohen's weighted kappa – used as an ordinal scale measure of agreement between two raters
Eq. (3)	Darroch & McCloud measure of category distinguishability
Eq. (14)	Rogot & Goldberg index of agreement
Eq. (18)	Armitage et al.index of disagreement
Eq. (24)	Intraclass kappa – obtained from the one-way random effects model under the assumption of no interrater bias
Eq. (26)	Reliability kappa – obtained from the two-way random effects model accounting for rater's effect
Eq. (36)	Concordance correlation – measures the departure from the 45° line

References

[1] Agresti, A. (1992). Modelling patterns of agreement and disagreement, *Statistical Methods in Medical Research* **1**, 201–218.

[2] Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability

model and its relation to Cohen's kappa, *Biometrics* **46**, 293–302.

[3] Armitage, P., Blendis, L.M. & Smyllie, H.C. (1966). The measurement of observer disagreement in the recording of signs, *Journal of the Royal Statistical Society, Series A* **129**, 98–109.

[4] Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability, *Psychological Reports* **19**, 3–11.

[5] Bartko, J.J. (1994). General Methodology II - Measures of agreement: a single procedure, *Statistics in Medicine* **13**, 737–745.

[6] Bassin, L., Borghi, C., Costa, F.V., Strocchi, E., Mussi, A., & Ambrossioni, E. (1985). Comparison of three devices for measuring blood pressure, *Statistics in Medicine* **4**, 361–368.

[7] Bishop, Y.M.M., Feinberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[8] Bland, M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **1**, 307–310.

[9] Bloch, D.A. & Kraemer, H. (1989). 2X2 kappa coefficient: measure of agreement or association, *Biometrics* **45**, 269–287.

[10] Bradley, E.L. & Blackwood, L.G. (1989). Comparing paired data: a simultaneous test of means and variances, *American Statistician* **43**, 234–235.

[11] Carey, G. & Gottesman, H. (1978). Reliability and validity in binary ratings: areas of common understanding in diagnosis and symptom ratings, *Archives of General Psychiatry* **35**, 1454–1459.

[12] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurements* **20**, 37–46.

[13] Cohen, J. (1968). Weighted kappa: nominal scale agreement with provisions for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213–220.

[14] Conn, H.O., Smith, H.W. & Brodoff, M. (1965). Observer variation in the endoscopic diagnosis of esophageal varices: a prospective investigation of the diagnostic validity of esophagoscopy, *New England Journal of Medicine* **272**, 830–834.

[15] Cyr, L. & Francis, K. (1992). Measures of clinical agreement for nominal and categorical data: the kappa coefficient, *Comparative Biological Medicine* **22**, 239–246.

[16] Darroch, J.N. & McCloud, P. (1986). Category distinguishability and observer agreement, *Australian Journal of Statistics* **28**, 371–388.

[17] Deming, W.E. (1943). *Statistical Adjustment of Data*. Wiley, New York.

[18] Donner, A. & Eliasziw, M. (1987). Sample size requirements for reliability studies, *Statistics in Medicine* **6**, 441–448.

[19] Donner, A. & Eliasziw, M. (1992). A goodness of fit approach to inference procedures for the kappa statistics: confidence interval construction, significance testing and sample size estimation, *Statistics in Medicine* **11**, 1511–1519.

[20] Dunn, G. (1989). *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors*. Oxford University Press, New York.

[21] Dunn, G. (1992). Design and analysis of reliability studies, *Statistical Methods in Medical Research* **1**, 123–157.

[22] Dunn, G. & Everitt, B.S. (1982). *An Introduction to Mathematical Taxonomy*. Cambridge University Press, Cambridge.

[23] Ebel, R. (1951). Estimation of the reliability of ratings, *Psychometrika* **16**, 407–423.

[24] Everitt, B.S. (1968). Moments of the statistics kappa and weighted kappa, *British Journal of Mathematical and Statistical Psychology* **21**, 97–103.

[25] Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes, *Journal of Clinical Epidemiology* **43**, 543–549.

[26] Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low kappa: II. Resolving the paradoxes, *Journal of Clinical Epidemiology* **43**, 551–558.

[27] Fleiss, J.L. (1965). Estimating the accuracy of dichotomous judgements, *Psychometrika* **30**, 469–479.

[28] Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait, *Biometrics* **31**, 651–659.

[29] Fleiss, J.L. (1981). Balanced incomplete blocks designs for interrater reliability studies, *Applied Psychological Measurements* **5**, 105–112.

[30] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[31] Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

[32] Fleiss, J.L. & Cicchetti, D.V. (1978). Inference about weighted kappa in the non-null case, *Applied Psychological Measurements* **2**, 113–117.

[33] Fleiss, J.L. & Cohen, J. (1973). The equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurements* **33**, 613–619.

[34] Fleiss, J.L. & Cuzick, J. (1979). The reliability of dichotomous judgement: unequal number of judges per subject, *Applied Psychological Measurement* **3**, 537–542.

[35] Fleiss, J.L., Cohen, J. & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa, *Psychological Bulletin* **72**, 323–327.

[36] Goldsmith, C.H. & Gaylor, D.W. (1970). Three stage nested designs for estimating variance components, *Technometrics* **12**, 487–498.

[37] Goodman, L.A. & Kruskal, W.H. (1954). Measures of association for cross classifications, *Journal of the American Statistical Association* **49**, 732–764.

[38] Grubbs, F.E. (1948). On estimating precision of measuring instruments and product variability, *Journal of the American Statistical Association* **43**, 243–264.

[39] Hermann, B. & Kliebsch, U. (1996). Dependence of weighted kappa coefficient on the number of categories, *Epidemiology* **7**, 199–202.
[40] Hirji, K. & Rosove, M. (1990). A note on interrater agreement, *Statistics in Medicine* **9**, 835–839.
[41] Kelly, G.E. (1985). Use of the structural equations model in assessing the reliability of a new measurement technique, *Applied Statistics* **34**, 258–263.
[42] Kendall, M.G. (1938). A new measure of rank correlation, *Biometrika* **30**, 81.
[43] Koran, L.M. (1975). The reliability of clinical methods, data and judgement, *New England Journal of Medicine* **293**, 642–646, 695–701.
[44] Kraemer, H.C. (1979). Ramification of a population model for κ as a coefficient of reliability, *Psychometrika* **44**, 461–472.
[45] Kraemer, H.C. (1992). Measurement of reliability for categorical data in medical research, *Statistical Methods in Medical Research* **1**, 183–199.
[46] Krippendorf, K. (1970). Bivariate agreement coefficients for reliability of data, in *Sociological Methodology*, F. Borgatta & G.W. Bohrnstedt, eds. Jossey-Bass, San Francisco, pp. 139–150.
[47] Lachenbruch, P.A. (1988). Multiple reading procedures: the performance of diagnostic tests, *Statistics in Medicine* **7**, 549–557.
[48] Leugrans, S. (1980). Evaluating laboratory measurement techniques, in *Biostatistics Case-Book*, R. Miller, B. Efron, B. Brown & L. Moses, eds. Wiley, New York, pp. 190–219.
[49] Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility, *Biometrics* **45**, 255–268.
[50] Linnet, K. (1990). Estimation of the linear relationship between the measurements of two methods with proportional errors, *Statistics in Medicine* **9**, 1463–1473.
[51] Lui, K.J. (1992). A note on the effect of the intraclass correlation in the multiple reading procedure with a unanimity rule, *Statistics in Medicine* **11**, 209–218.
[52] Maclure, M. & Willett, W.C. (1987). Misinterpretation and misuse of the kappa coefficient, *American Journal of Epidemiology* **126**, 161–169.
[53] Mak, T.K. (1988). Analyzing intraclass correlation for dichotomous variables, *Applied Statistics* **37**, 344–352.
[54] Maxwell, A.E. & Pilliner, A.E.G. (1968). Deriving coefficients of reliability and agreement for ratings, *British Journal of Mathematical and Statistical Psychology* **21**, 105–116.
[55] Mehta, C. & Patel, N. (1995). *StatXact 3 for Windows; User Manual*. Cytel Software Corporation, Cambridge, Mass., Chapter 23.
[56] Nix, A.B.J. & Dunston, F.D.J. (1991). Maximum likelihood techniques applied to method comparison studies, *Statistics in Medicine* **10**, 981–988.
[57] Robinson, D.L. (1987). Estimation and use of variance components, *Statistician* **36**, 3–14.
[58] Robinson, D.L. (1987). *REML User Manual*. Scottish Agricultural Statistical Service, Edinburgh.
[59] Rogot, E. & Goldberg, I.D. (1966). A proposed index for measuring agreement in test-retest studies, *Journal of Chronic Diseases* **19**, 991–1006.
[60] SAS Institute, Inc. (1992). *SAS Release 6.07*. SAS Institute Inc., Cary.
[61] Schouten, H.J.A. (1985). Statistical Measurement of Interrater Agreement, Ph.D. Doctoral Dissertation. Erasmus University, Rotterdam.
[62] Schouten, H.J.A. (1986). Nominal scale agreement among observers, *Psychometrika* **51**, 453–466.
[63] Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding, *Public Opinion Quarterly* **19**, 321–325.
[64] Searle, S.R. (1987). *Linear Models for Unbalanced Data*. Wiley, New York.
[65] Shoukri, M.M. & Edge, V.L. (1996). *Statistical Methods for Health Sciences*. CRC Press, Boca Raton.
[66] Shoukri, M.M., Martin, S.W. & Mian, I.U.H. (1995). Maximum likelihood estimation of the kappa coefficient from models of matched binary responses, *Statistics in Medicine* **14**, 83–99.
[67] Shukla, G.K. (1973). Some exact tests on hypothesis about Grubbs estimators, *Biometrics* **29**, 373–377.
[68] Snedecor, G.W. & Cochran, W.G. (1980). *Statistical Methods*, 7th Ed. Iowa State University, Ames.
[69] Spitznagel, E.L. & Helzer, J.E. (1985). A proposed solution to the base rate problem in the kappa statistics, *Archives of General Psychiatry* **42**, 725–728.
[70] Thompson, W.D. & Walter, S.D. (1988). A reappraisal of the kappa coefficient, *Journal of Clinical Epidemiology* **41**, 949–958.
[71] Thompson, W.D. & Walter, S.D. (1988). Kappa and the concept of independent errors, *Journal of Clinical Epidemiology* **41**, 969–970.
[72] Winer, B.J. (1971). *Statistical Principles in Experimental Design*. McGraw-Hill, New York.

(*See also* **Agreement, Modeling of Categorical**; **Kappa and its Dependence on Marginal Rates**; **Observer Reliability and Agreement**)

M.M. SHOUKRI

Agreement, Modeling of Categorical

In many situations two observers evaluate and classify each of n subjects into one of K mutually exclusive categories. For $K = 2$ a common measure of chance-corrected agreement between the two raters is Cohen's **kappa** (κ) statistic [6]. The measure has been extended to $K > 2$, where the K categories may be ordered [7] (*see* **Kappa and its Dependence on Marginal Rates**). Some authors, e.g. Agresti [1], argue that a single measure of agreement may be too simple for some situations. For example, one may want to model agreement after adjustment for chance agreement and ordinal association [1, 2] (*see* **Ordered Categorical Data**). Additionally, **covariates** may influence the probabilities of classification and may need to be considered prior to assessment of agreement [3]. Below we discuss models that address these issues.

Assumption of Marginal Homogeneity

Let p_{ij} be the probability of a subject being classified in category i by the first rater and category j by the second rater. Then the marginal probability of classification into category i for the first rater is $p_{i\cdot} = \sum_j p_{ij}$ and similarly $p_{j\cdot} = \sum_i p_{ij}$ for classification into category j by the second rater. Some models of agreement assume that $p_{i\cdot} = p_{\cdot i}$ for $i = 1, \ldots, K$. This assumption of marginal homogeneity is stringent, but is consistent with the definition of intraclass **correlation** for continuous data that assumes equal means and variances for the two variables [14]. Note that marginal homogeneity may be a prerequisite for high agreement [2]. Quite simply, one would not expect agreement to be high when the marginal totals are discrepant because the raters are probably not using the same underlying classification scheme. In the **loglinear model** described below, the assumption of marginal homogeneity is testable.

2 × 2 Tables

For the usual kappa statistic, the row and column marginal totals in **2 × 2 tables** are considered fixed. Since sample sizes may be small in studies of agreement, the number of discordant observations may be small. In this case computing a ***P* value** based on the exact permutation distribution (with marginals fixed) may be preferable (e.g. as implemented in **StatXact** 3 for Windows, Cambridge, MA; (*see* **Exact Inference for Categorical Data**). Alternatively, one can assume marginal homogeneity and compute an intraclass correlation [5, 9] as discussed below. The corresponding permutation distribution would be conditional on the overall number of positive and negative responses, but allow the cell frequencies in Table 1 to vary.

The marginal probability of classification for a particular subject may depend on one or more subject-specific covariates as well. For example, a radiologist may be more likely to classify a mammographic abnormality as breast cancer in the presence of known breast cancer risk factors, such as a family history of breast cancer or advanced age. A stratified kappa (*see* **Stratification**) can be used with weights determined by the sample size for each covariate combination comprising the strata [4]. Failure to account for these **confounders** (i.e. collapsing across strata) may lead to inflated estimates of agreement [4].

As the number of confounders becomes large, the stratified kappa would be based on sparse tables and would exhibit poor asymptotic behavior. Therefore, it is desirable to adjust directly for subject-specific covariates that may influence the raters in their classification of the subject. For the $K = 2$ case one can use a **logistic regression** model for linking subject-specific covariates to the marginal probability of classification and include a term for the intraclass correlation θ [3]. The model does assume marginal homogeneity, with the subject-specific covariates influencing each rater's marginal probability in the same way.

Let n_{ij} be the frequency in the ith row (rater 1) and the jth column (rater 2). Assuming marginal homogeneity leads to the trinomial distribution given in Table 1. This model (without covariates) is discussed elsewhere [5, 11]. Note that the model gives θ as the marginal probability of a positive response for each rater. **Maximum likelihood** estimates of ρ and θ are

$$\hat{\rho} = \frac{4n_{11}n_{22} - (n_{12} + n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})}$$

and

$$\hat{\theta} = \frac{2n_{11} + n_{12} + n_{21}}{2n},$$

Table 1 Trinomial model

Classification	Both positive	Discordant	Both negative
Frequency	n_{11}	$n_{12} + n_{21}$	n_{22}
Probability	$\theta^2 + \rho\,\theta(1-\theta)$	$2(1-\rho)\,\theta\,(1-\theta)$	$(1-\theta)^2 + \rho\,\theta(1-\theta)$

with estimated variance estimate for $\hat{\rho}$ given by [5, 9]:

$$\text{var}(\hat{\rho}) = \frac{1-\hat{\rho}}{n}\left[(1-\hat{\rho})(1-2\hat{\rho}) + \frac{\hat{\rho}(2-\hat{\rho})}{2\hat{\theta}(1-\hat{\theta})}\right].$$

Using the modeling approach derived below, a score test of $H_0 : \rho = 0$ can be derived. Without covariates, the score test is simply $n\hat{\rho}^2$, assumed to be distributed as χ^2 (1 df) under the **null hypothesis**. The estimate $\hat{\rho}$ is the intraclass correlation for dichotomous outcomes. If the formula for the intraclass correlation for continuous data is applied to the dichotomous data, then the estimate $\hat{\rho}$ is obtained [14]. In practice, $\hat{\rho}$ will be quite close to the kappa statistic and equal to it when $n_{12} = n_{21}$.

Inclusion of Covariates

To include subject-specific covariates, let $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ be the $p+1$ vector of covariates for subject i. Assume a logit link function (*see* **Generalized Linear Model**) between the **mean** θ_i and the covariate vector $\mathbf{x}_i$, i.e. logit $(\theta_i) = \mathbf{x}_i\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ a parameter vector to be estimated. This **multinomial** model may be fitted as a conditional logistic regression model with a generalized **relative risk** function [3].

$$r_i = \exp(z_i\beta) + w_i\rho - \frac{w_i - 1}{3},$$

where $\mathbf{x}_i$ and w_i are functions of the covariates and the observed outcome for person i. This additive risk function decomposes into a part that incorporates the covariates, a part that depends on the intraclass correlation, and an "offset". **Confidence limits** for θ may be derived using either a standard Wald interval or a **profile likelihood** interval. Alternatives to the asymptotic bounds are desirable for establishing a confidence interval for ρ since Wald-based confidence intervals have poor coverage probabilities for estimates of agreement [4].

The model also allows for easy derivation of a score test of $H_0 : \rho = 0$. Let Y_{ij} indicate whether the ith person has been classified into the jth cell of Table 1, $i = 1, \ldots, n$; $j = 1, 2, 3$. Let $\tilde{\boldsymbol{\beta}}$ be an estimate of $\boldsymbol{\beta}$ assuming $\rho = 0$, i.e. breaking the pairing and estimating $\boldsymbol{\beta}$ using a standard logistic regression model for the $2n$ observations. If the risk (denoted $\tilde{r}$) is based on $\tilde{\boldsymbol{\beta}}$, then a one-step estimate of ρ is

$$\hat{\rho}_{1-\text{step}} = \frac{1}{\text{n}}\sum_i\left(\frac{Y_{i1}}{\tilde{r}_{i1}} - \frac{2Y_{i2}}{\tilde{r}_{i2}} + \frac{Y_{i3}}{\tilde{r}_{i3}}\right)$$

and the score test (χ^2 with 1 df; *see* **Likelihood**) of $H_0 : \rho = 0$ is $n\hat{\rho}^2_{1-\text{step}}$. Limitations of the model include the marginal homogeneity assumption and restriction to $K = 2$.

Extension to $K \times K$ Tables

In many situations ratings are made on a **nominal** or ordinal scale with $K > 2$. Tanner & Young [15] proposed a loglinear model for nominal scales. Suppose that Y_{ij} is the number classified into row i and column j by raters 1 and 2, respectively. Assume a **Poisson distribution** for Y_{ij} with the loglink function $\log \mu_{ij} = z_{ij}^{\text{T}}\boldsymbol{\beta}$. A possible model would be $\log \mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij}$, where the $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ parameters correspond to the marginal distributions for raters 1 and 2, respectively, and δ_{ij} is zero for $i \neq j$ and equal to δ_i otherwise. A test of statistically significant agreement is given by the difference in deviances between this model and the independence model, $\log \mu_{ij} = \mu + \alpha_i + \beta_j$. The **likelihood ratio test** is assumed to have a **chi-square distribution** with K degrees of freedom for $K > 2$. This particular model is called the "**quasi-independence** model" because raters 1 and 2 are assumed to be independent given they disagree. The model forces the fitted values on the diagonal to be equal to the observed values. An intermediate model is given by constraining $\delta_1 = \delta_2 = \ldots \delta_K = \delta$. This model does not constrain the diagonals to equal the observed values

and can be tested against both the independence and quasi-independence models by likelihood ratio tests.

The above model does not assume marginal homogeneity, but it is not difficult to make this assumption and test its validity. The difference in deviances between models $\log \mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij}$ and $\log \mu_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij}$ has a chi-square distribution on $K - 1$ degrees of freedom. The α parameters for the latter model are estimated using covariates $X_l = I(l = i) + I(l = j)$ for $l = 2, \ldots, K$, where I is an indicator function. When $K = 2$ the likelihood ratio test of $H_0 : \delta = 0$ under marginal homogeneity is identical to that for $H_0 : \rho = 0$ in the first section.

For ordinal categories it is likely that if there is disagreement the two raters choose ratings that are more similar rather than more distant. A weighted kappa may be used with weights determined by linear or squared differences between the scores corresponding to the categories [7]. However, these models can accommodate covariates only by stratification and may be too simple for most situations. Accordingly, Agresti [1] proposed a generalization of the loglinear model intended to accommodate the association between the two ordinal scales. The discussion below is based largely on Agresti's generalization of loglinear models of nominal association [10, 15].

Suppose that u_i is the score associated with the ith category of the ordinal scale with $u_1 < u_2 < \ldots < u_K$. In most cases $u_i = i$, but the scores could depend on midpoints of categorized values or on subjective a priori experience. A possible model is

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j + u_i u_j \theta + \delta_{ij},$$

where θ indexes the linear-by-linear association and δ_{ij} indexes residual agreement between the raters. A comparison of deviances for this model vs. $\log \mu_{ij} = \mu + \alpha_i + \beta_j + u_i u_j \theta$ gives a test of agreement on K degrees of freedom. Unlike the weighted kappa statistic, "partial credit" for being similar rather than identical is attributed primarily to association rather than agreement. Agresti [1] recommends a slightly simpler model with $\delta_1 = \delta_2 = \cdots = \delta_K = \delta$, since the diagonal elements are not constrained to be equal to the observed values. This model has $(K - 1)^2 - 2$ df for assessing **goodness of fit** and allows a direct test of agreement $H_0 : \delta = 0$ and/or linear association $H_0 : \theta = 0$.

This model may also be modified to assume marginal homogeneity:

$$\log \mu_{ij} = \mu + \alpha_i + \alpha_j + u_i u_j \theta + \delta_{ij},$$

with δ_{ij} either a scalar (δ) or a **K**-vector on the diagonal. The latter model again imposes a perfect fit on the diagonal. Finally, one can include categorical subject-level covariates that may influence the marginal probabilities. For example, the subjects may be grouped into m age groups, and the $m(K \times K)$ tables fit by a common model with age entering as a main effect. Furthermore, age may be tested as an **effect modifier** of δ using **interaction** terms. Controlling for several **confounders** simultaneously will often not be possible due to sparseness of the tables, however.

Multiple Raters

In many cases several raters will rate all or some of the subjects and classify them into the K categories. In this situation extensions to kappa have been proposed [8, 13]. If all pairwise tables are considered, the tables are correlated since a rater may appear in several tables. Schouten [13] proposes using a **jackknife** variance estimator to assess significance in this setting. A unified loglinear model of agreement across multiple raters is not yet available. Tanner & Young [15] discuss a loglinear model for comparison of several raters to a standard.

If marginal probabilities are allowed to vary with each rater, the number of parameters tends to infinity when the number of raters does. In this case marginal homogeneity may be required. One might consider all pairwise tables in a loglinear model under marginal homogeneity. It would be necessary to correct the standard errors using a **generalized estimating equations** approach [12]. Alternatively, confidence limits could be found by **bootstrapping** the contributions from each rater.

Example

An agreement and accuracy study was performed to study the ability of radiologists to classify screening mammograms. The research was supported by the National Cancer Institute (CA63731, PI Stephen Taplin, MD). We consider

only a subset of the data here. Ten radiologists classified 113 screening mammograms into five categories: (i) normal; (ii) normal with benign findings; (iii) probably benign; (iv) suspiciously abnormal; and (v) highly suspicious of malignancy. The 113 mammograms included a mix of negative and positive mammograms. Table 2 shows the distribution for a single pair of raters and Table 3 shows the classification for all possible pairing of raters (45) classifying each of the 113 subjects, yielding a total of 5085 paired ratings.

Data for the two raters in Table 2 were analyzed using a loglinear model, though sparseness in the Table may be problematic. The Agresti [1] linear association model fits well (deviance $\chi^2 = 15.71$ on 14 df), but is not significantly different from a model that assumes marginal homogeneity that also provides a reasonable fit (deviance $\chi^2 = 21.64$ on 18 df). Under the two models, the estimates of the scalar δ are 1.065 (se 0.466) and 0.862 (se 0.448), respectively. For the marginal homogeneity model, both the Wald test ($z = 1.923$) and likelihood ratio test ($\chi^2 = 3.38$ on 1 df) show weak (one-tailed) significance for agreement beyond linear association. By comparison the unweighted and linear weighted kappas are quite high – 0.620 and 0.774, respectively.

Table 2 Classification by two raters of 113 screening mammograms

	Rater 1					
Rater 2	1	2	3	4	5	Total
1	75	1	3	1	0	80
2	1	1	0	0	1	3
3	5	2	4	0	1	12
4	0	0	2	1	3	6
5	0	0	0	0	12	12
Total	81	4	9	2	17	113

Table 3 Pairwise classification by 10 raters of 113 screening mammograms

Lower rating	Higher rating				
	1	2	3	4	5
1	3073	178	631	32	15
2		24	109	13	3
3			217	127	76
4				68	178
5					341

The loglinear model dissects the apparent interrater agreement into a strong linear association and weak residual agreement.

If all 45 pairwise tables are considered, then the median estimate of δ is 0.244 when not assuming marginal homogeneity and 0.093 under this assumption. In 43 of 45 tables, marginal homogeneity resulted in a smaller estimate of agreement. Finally, if one models all 45 tables simultaneously (assuming marginal homogeneity), then the overall δ estimate is 0.103, but the standard error needs to be corrected for the induced correlations among pairings.

Conclusions

Agreement between two fixed raters may be modeled using either a conditional logistic regression model ($K = 2$) or a loglinear model ($K \geq 2$). The former allows for continuous or categorical covariates, while the latter can incorporate only categorical covariates. More complex models are needed for more than two raters. Marginal homogeneity may be a prerequisite for strong agreement to be observed.

References

[1] Agresti, A. (1988). A model for agreement between ratings on an ordinal scale, *Biometrics* **44**, 539–548.

[2] Agresti, A. (1992). Modelling patterns of agreement and disagreement, *Statistical Methods in Medical Research* **1**, 201–218.

[3] Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates, *Biometrics* **52**, 695–702.

[4] Barlow, W., Lai, M.Y. & Azen, S.P. (1991). A comparison of methods for calculating a stratified kappa, *Statistics in Medicine* **10**, 1465–1472.

[5] Bloch, D.A. & Kraemer, H.C. (1989). 2×2 kappa coefficients: measures of agreement or association, *Biometrics* **45**, 269–287.

[6] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[7] Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213–220.

[8] Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin* **76**, 378–382.

[9] Fleiss, J.L. & Davies, M. (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance, *American Journal of Epidemiology* **115**, 841–845.

[10] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having

ordered categories, *Journal of the American Statistical Association* **74**, 537–552.

[11] Hale, C.A. & Fleiss, J.L. (1993). Interval estimation under two study designs for kappa with binary classifications, *Biometrics* **49**, 523–534.

[12] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[13] Schouten, H.J.A. (1986). Nominal scale agreement among observers, *Psychometrika* **51**, 453–466.

[14] Snedecor, G.W. & Cochran, W.G. (1967). *Statistical Methods*, 6th Ed. Iowa State University Press, Ames.

[15] Tanner, M.A. & Young, M.A. (1985). Modeling agreement among raters, *Journal of the American Statistical Association* **80**, 175–180.

(*See also* **Categorical Data Analysis**; **Confounding**; **Observer Reliability and Agreement**)

WILLIAM BARLOW

AIDS and HIV

The human immunodeficiency virus (HIV) can be transmitted from person to person. The immune system of individuals infected with HIV deteriorates, leaving them susceptible to infection and the development of various diseases. Certain conditions of poor health in HIV infected individuals qualify for a diagnosis of acquired immune deficiency syndrome (AIDS). Statistical issues associated with data from the HIV epidemic and HIV/AIDS related studies emerged soon after AIDS was formally defined in 1982. At that time the number of cases began to increase alarmingly and early concerns were with **forecasting** the incidence of AIDS diagnosis, estimating the distribution of the survival time with AIDS and estimating the distribution of the time from infection until diagnosis with AIDS. The latter time period is often called the **incubation period**.

The area of AIDS and HIV data has been a rich source of statistical challenges, which are reported in an extensive literature. Fusaro et al. [8] give an annotated bibliography of the early literature. Jewell [11] reviews some statistical issues associated with the study of HIV and AIDS data, and some of these are described in detail in the book by Brookmeyer & Gail [6].

Extrapolating AIDS Incidence

The rising number of AIDS diagnoses and the devastating consequence for its victims naturally generated an interest in the current trend of AIDS incidence and the likely number of cases in the near future. Owing to the lack of knowledge about the way AIDS cases arose, early attempts at forecasting AIDS incidence were based on fitting curves of a simple algebraic form to the observed AIDS incidences and then extrapolating these curves into the future. More specifically, **polynomial regression** models in time were fitted to the AIDS counts, or logarithms of the counts, and then **extrapolated**. The basis for making projections of AIDS incidences by extrapolating fitted polynomials is weak. Furthermore, the approach does not give an estimate of the rate of development of new infections over time, i.e. the HIV infection curve, which is of considerable interest. For these reasons, there soon developed an interest in using models that contain at least some aspects of the way AIDS cases arise.

Transmission Models for HIV

Only models that describe the way HIV is transmitted from person to person contain all aspects of the way AIDS cases are generated. There is a substantial body of work on infectious disease models (*see* **Communicable Diseases**), and attempts were soon made to adapt these models to HIV transmission.

There are several features that distinguish the HIV epidemic from epidemics of other communicable diseases. First, several different modes of transmission are identified for HIV, such as sexual transmission, injecting drugs with contaminated needles, and receiving contaminated blood or blood products during medical therapy. This means that comprehensive models contain many parameters that need to be estimated, from data that are personal, highly sensitive, and sparse. The models of Hethcote & Van Ark [9] illustrate this. To keep models simple it is common to restrict attention to a single risk category with the assumption that transmission within that group occurs without significant transmission from infective individuals in other risk categories.

Another distinguishing feature is that the incubation period usually has a duration of several years, with the consequence that knowledge about the disease, therapy, and behavior change significantly during the course of the epidemic. In statistical terms this means that model parameters do not remain constant.

These difficulties give rise to considerable uncertainty about conclusions reached by the use of transmission models. As a consequence the preferred methods for assessing the size of the HIV epidemic and making projections of AIDS incidence are based on the method of **back-calculation**, which is now described.

Reconstructing the HIV Infection Curve

This topic has been a major focus of statistical studies in the HIV/AIDS area. Reconstruction of the realized, but unobserved, HIV incidences is of interest because they indicate the size of the epidemic and are useful for predicting future AIDS incidences and the health care costs resulting from AIDS cases.

Let H_t and A_t denote the number of HIV infections and AIDS diagnoses in the discrete time unit t, typically a quarter, and let τ be the last time unit for which reliable data are available. We can relate $\mu_t = \mathrm{E}(A_t)$, the mean AIDS incidence in time unit t, to the HIV incidences at earlier times by

$$\mu_t = \sum_{s=1}^{t} \lambda_s f_{t-s}, \quad t = 1, 2, \ldots, \tag{1}$$

where $\lambda_s = \mathrm{E}(H_s)$ and f_r is the probability that the duration of the incubation period is r time units. This convolution equation holds for any transmission model, relying only on subjects having independent incubation periods.

The method of back-calculation, or back-projection, uses (1) to reconstruct the HIV infection curve, by a method that ignores the fact that HIV cases arise by infection. Most versions of the method of back-projection reconstruct the $H_1, H_2, \ldots, H_\tau$ by estimates of $\lambda_1, \lambda_2, \ldots, \lambda_\tau$, assuming that a precise estimate of the incubation period distribution $f_0, f_1, \ldots$ is available from data of large **cohort studies**.

Typically, the $H_1, H_2, \ldots, H_\tau$ are assumed independent Poisson variates (*see* **Poisson Processes**). This produces a very convenient working model, because with independent incubation periods it makes the AIDS incidences independent observations of Poisson variates with means given by (1), making **maximum likelihood** estimation an option for $\lambda_1, \lambda_2, \ldots, \lambda_\tau$. Rosenberg & Gail [24] show that for large epidemics this model assumption gives similar results to some other model assumptions. Simulation results show that this model gives sensible reconstructed HIV incidences even when data are generated by a transmission model [3].

Leaving the λ_s as separate parameters, in the spirit of functional estimation, makes their estimates overly sensitive to minor perturbations in the AIDS incidence data, so some form of smoothing (*see* **Nonparametric Regression**) needs to be imposed. When they first proposed back-projection in the AIDS context, Brookmeyer & Gail [6] assumed a piecewise constant form for λ_s, $s = 1, 2, \ldots, \tau$; that is, the λ_s take only a small number of distinct values. However, a smooth parametric form has also been used for the λ_s [27], while Isham [10] uses a smooth parametric form for the μ_t and deconvolutes (1).

A preference has evolved for leaving the λ_s as separate parameters but constraining them to a smooth form, because no simple natural parametric form is apparent for the HIV infection curve. Another, possibly more important, reason for this preference is a concern that with the use of parametric models we may not get an appropriate reflection of how the precision of the reconstructed HIV infection curve varies over time.

A variety of methods have been explored for reconstructing the HIV infection curve by smoothing a nonparametric estimate, including use of the **EM algorithm** with a smoothing step added to each iteration of the usual E- and M-steps for maximum likelihood estimation [5]. This method is very easy to program because the iterations involve only simple explicit expressions. Another approach is to estimate the λ_s by **penalized maximum likelihood**:

$$L(\boldsymbol{\lambda}; \mathbf{a}) - \gamma \Psi(\boldsymbol{\lambda}).$$

Here $L(\boldsymbol{\lambda}; \mathbf{a})$ is the **likelihood** function corresponding to the observed AIDS cases and $\Psi(\boldsymbol{\lambda})$ is a function that penalizes any estimate that is not sufficiently smooth. For example, $\Psi(\boldsymbol{\lambda}) = \sum_j (\lambda_{j+1} - 2\lambda_j + \lambda_{j-1})^2$ is used by Bacchetti et al. [2].

Bayesian methods seem natural for reconstructing the HIV infection curve, since posterior probabilities are appropriate descriptions of likely realizations of unobserved incidences, and these have been used to get smooth back-projections [18]. **Ridge regression** [21] and **splines** [24] have also been applied to obtain smooth reconstructions.

AIDS counts alone contain very little information about HIV infections in the recent past so that, naturally, reconstruction of the HIV incidences for recent time units is less precise than reconstruction for the distant past.

The large number of variations in the methods of back-projection are matched by an equally wide range of ways in which changes in the incubation distribution, over time, have been incorporated. These changes are needed because the incubation period has changed over time as a result of both therapy effects and changes in the definition of AIDS. Early methods for incorporating the dependence of the f_r on the time of infection were simple and crude, but there were soon a variety of ways in which these effects were incorporated in a more descriptive way [6].

Methods for incorporating auxiliary information into back-projections have been developed to

improve the precision of reconstructed HIV infection curves. For example, age at AIDS diagnosis is usually observed and its inclusion as a **covariate** improves precision and enables the estimation of age-specific **relative risks** of infection [4].

Estimating the Incubation Distribution from Retrospective Ascertainment Data

Estimation of the probability distribution for the incubation period has also been a major focus of statistical studies in the HIV/AIDS area. It is of interest because this distribution is needed to advise patients, to monitor disease progression, and for back-projection. Its estimation presents difficulties because the time of infection is usually unknown.

For individuals infected with HIV by a single blood transfusion it is possible to determine the time of infection. However, even for data on these subjects, estimation requires new methodology, because subjects usually enter the study sample only when they have been diagnosed with AIDS. At this point their time of infection can only be ascertained retrospectively. Shorter incubation periods are overrepresented when subjects are sampled in this way, so it is important that the method of estimation takes account of the way the data arise. Lui et al. [19] estimated the incubation distribution by ignoring the infection process that led to the data, considering each retrospectively ascertained incubation time as an observation from a truncated Weibull distribution. Medley et al. [20] allow for HIV infections to occur according to a Poisson process with the rate having a parametric form $\lambda(t;\boldsymbol{\beta})$ at time t and taking a parametric form $f(x;\boldsymbol{\theta})$ for the incubation distribution. They consider the forms

$$\lambda(t;\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 t) \quad \text{and} \quad \lambda(t;\boldsymbol{\beta}) = \beta_0 + \beta_1 t$$

for the infection rate and consider both the gamma and the Weibull distributions for f.

As pointed out by Kalbfleisch & Lawless [13], there is an **identifiability** problem when we try to estimate both the Poisson rate, λ, and the incubation distribution, f, nonparametrically, and this reflects itself even in the parametric setting by making the estimates of certain parameters imprecise. Nonparametric estimation of the shape of the incubation distribution is facilitated by a reparameterization to reverse-time **hazard rates** [17].

If it were possible to follow up a random sample of subjects infected by a blood transfusion, then the analysis of data on incubation periods would be a standard problem of **survival analysis**. Retrospective ascertainment data is the extreme case when information on the incubation period is available only for AIDS cases. Between these two extremes is the situation where the incubation periods are known for those who have developed AIDS and we also know the time since infection for some of those infected but still without AIDS. Analysis of such data requires modeling of the way the HIV infected individuals are detected.

The effect of covariates on the incubation distribution can be studied by focusing on the reverse-time hazard $h(x) = f(x)/F(x)$, where f is the density and F the distribution function. The analysis resembles survival analysis [17], and a **proportional hazard** regression analysis can be applied in reverse time [14].

Estimating the Incubation Distribution from Other Data

Estimation of the incubation distribution from prospective cohort studies has contributed to the development of a methodology for **interval censored** data. The time of infection is usually unknown, but tests for antibodies to HIV on subjects over time produce a time of last seronegative and a time of first seropositive. This has motivated extensions of known self-consistency algorithms to deal with doubly **censored data** [7], as well as truncation (*see* **Truncated Survival Times**) [26].

With a method analogous to the method of back-projection, Bacchetti & Jewell [1] use the convolution equation, (1), to estimate the incubation distribution by using the λ_s as estimated from data on a large cohort study. They obtain a smooth nonparametric estimate by using a penalized likelihood approach.

Monitoring Markers of the Immune System in Subjects

Numerous markers of the state of the immune system have been considered for monitoring disease progression. Of these the $CD4^+$ count seems the best indicator of disease progression. These counts are

highly variable, and the count declines at a rate that differs between individuals. The typical analysis fits a model that is linear in time, using **random effects**, to some transformation of the $CD4^+$ counts. The five papers in Jewell et al. [12, Section 3] give a good overview of the relevant statistical methodology.

Reporting Delays

Surveillance systems are usually prone to delays in the registration of cases, and AIDS surveillance reports are no exception. The number of AIDS cases in a surveillance report is usually less than the number of AIDS cases that have actually been diagnosed by the time of its publication, because of random delays until dates of AIDS diagnoses are entered onto the surveillance register (*see* **Surveillance of Diseases**).

AIDS surveillance data must be appropriately corrected for reporting delays before they can be used either for reconstruction of the HIV infection curve or for forecasting of AIDS incidence. This was recognized quite early, and the public access data on AIDS incidence in the US contain an adjustment for reporting delays.

Such adjustments require knowledge of the probability distribution of reporting delays. This distribution needs to be estimated from data on dates of diagnosis and subsequent dates of entry onto the surveillance register. A feature of these data is that both the date of diagnosis and the date of entry onto the register usually become available only at the time of entry onto the register. In other words, the available data have the same form as retrospective ascertainment data on the incubation periods of transfusion-acquired HIV infection, with the consequence that the methods of analysis carry over to reporting delay data. In particular, the methods of nonparametric estimation described in Kalbfleisch & Lawless [13] and Lagakos et al. [17] can be used.

A comprehensive analysis of reporting delay data, involving covariates, is a useful way of identifying weaknesses in the reporting system and of detecting changes in the reporting delay distribution over time. Methods for testing the **stationarity** of reporting delays over time, using a proportional reverse-time hazards model, are described by Kalbfleisch & Lawless [13] and Pagano et al. [21]. The use of **generalized linear models** for identifying significant factors, such as geographic region for example, is described by Zeger et al. [29].

Forecasting AIDS Incidence

Before using AIDS surveillance data to predict future AIDS incidence the available data must be corrected for reporting delays. Rosenberg [23] describes a simple way to do this.

All methods of forecasting extrapolate a curve like the convolution equation, (1), beyond the period for which data are available. For example, a prediction of the AIDS incidence for time unit τ', where $\tau' > \tau$, is given by

$$\mu_{\tau'} = \sum_{s=1}^{\tau'} \lambda_s f_{\tau'-s}. \qquad (2)$$

If a parametric form is assumed for the λ_s, then its smaller number of parameters have been estimated from the available data and extrapolating the time argument presents no difficulty, but relies on the assumed parametric form remaining appropriate for time points in the future. When the λ_s have been estimated as separate parameters, i.e. nonparametrically, then (2) contains $\lambda_{\tau+1}, \ldots, \lambda_{\tau'}$, which are not known. Some assumptions need to be made about those values. These assumptions are not crucial for short-term predictions since their multipliers $f_{\tau'-\tau-1}, \ldots, f_0$ will then be very small, because short incubation periods are rare.

The f_r contained in (2) are assumed known but are in fact estimated from available data. This means that knowledge about probabilities f_r is very imprecise for large r, adding to the uncertainty of predictions.

Infectivity

Rates of transmission between individuals are a central concern for infectious diseases. HIV can be transmitted between sexual partners. Estimating the risk of transmission between partners and determining the factors that affect this risk has been a focus of some studies.

To gain an understanding of sexual transmission of HIV, fundamental questions have had to be addressed with new statistical methodologies. Kaplan [15] used simple Bernoulli process models to investigate whether the number of partners alone can explain sexual transmission of HIV or whether the number of sex acts must be counted. Shiboski & Jewell [25] extend the investigation to whether the infectivity changes during the course of a partnership and

identify factors which influence the infectiousness of the initial infective partner and the susceptibility of uninfected partners. Analysis is complicated by limitations in the data, which include uncertainty in the times of infection of HIV positive individuals. Their methodology is akin to those of survival analysis.

When there is uncertainty about which partner is the source of infection, the EM-algorithm can be used to estimate the probabilities of transmission between partners.

Methods for Studying Intervention

Treatments intended to delay disease progression in HIV infected subjects are often therapies for just one or two of the many AIDS-defining illnesses. **Competing risk** models are therefore a natural tool for the assessment of the effectiveness of therapy in **clinical trials**. Another issue requiring statistical consideration is the assessment of drug **compliance** during clinical trials [16], because drugs tend to be self-administered and have undesirable side-effects.

Trials for determining the efficacy of vaccines (*see* **Vaccine Studies**), often aimed at preventing the onset of disease rather than preventing infection with HIV, present special challenges in the HIV/AIDS context [22]. One major issue of concern, which also has relevance to trials that assess other therapies of disease progression, is that the long incubation period means that estimation of the vaccine efficacy, with adequate precision, requires a very long trial. To avoid such delays there is interest in the use of **surrogate endpoints**. In particular, a reduction in the loss of $CD4^+$ cells during HIV infection is considered a promising sign of vaccine impact.

A second issue of concern is that subjects are able to determine, by being tested, whether they are receiving a vaccine rather than a placebo. This may induce a feeling of security in individuals who know they are receiving the vaccine, with a possible change in behavior to less safe sexual practices. This may obscure a modest protection offered by the vaccine.

Statistical methods have also been used to evaluate needle exchange programs.

Screening for HIV Infection

The seroprevalence of HIV infection is low in many parts of the population. In low **prevalence** regions there will tend to be more **false positives** than real positives. Furthermore, large samples are needed to obtain a nonzero estimate of the prevalence. This prompts the pooling of sera samples and performing tests on the pooled samples. Statistical studies have shown that not only does this result in cost saving, but the estimate of the seroprevalence can be improved significantly [28].

References

[1] Bacchetti, P. & Jewell, N.P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times, *Biometrics* **47**, 947–960.

[2] Bacchetti, P., Segal, M.R. & Jewell, N.P. (1993). Back-calculation of HIV infection rates, *Statistical Science* **8**, 82–119.

[3] Becker, N.G. & Chao, X. (1994). Dependent HIV incidences in back-projection of AIDS incidence data, *Statistics in Medicine* **13**, 1945–1958.

[4] Becker, N.G. & Marschner, I.C. (1993). A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data, *Biometrika* **80**, 165–178.

[5] Becker, N.G., Watson, L.F. & Carlin, J.B. (1991). A method of non-parametric back-projection and its application to AIDS data, *Statistics in Medicine* **10**, 1527–1542.

[6] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, New York.

[7] DeGruttola, V. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics* **45**, 1–11.

[8] Fusaro, R.E., Jewell, N.P., Hauck, W.W., Heilbron, D.C., Kalbfleisch, J.D., Neuhaus, J.M. & Ashby, M.A. (1989). An annotated bibliography of quantitative methodology relating to the AIDS epidemic, *Statistical Science* **4**, 264–281.

[9] Hethcote, H.W. & Van Ark, J.W. (1992). *Modelling HIV transmission and AIDS in the United States*. Springer-Verlag, Berlin.

[10] Isham, V. (1989). Estimation of incidence of HIV infection, *Philosophical Transactions of the Royal Society of London, Series B* **325**, 113–121.

[11] Jewell, N.P. (1990). Some statistical issues in studies of the epidemiology of AIDS, *Statistics in Medicine* **9**, 1387–1416.

[12] Jewell, N.P., Dietz, K. & Farewell, V.T., eds (1992). *AIDS Epidemiology: Methodological Issues*. Birkhäuser, Boston.

[13] Kalbfleisch, J.D. & Lawless, J.F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS, *Journal of the American Statistical Association* **84**, 360–372.

[14] Kalbfleisch, J.D. & Lawless, J.F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags, *Statistica Sinica* **1**, 19–32.

[15] Kaplan, E.H. (1990). Modeling HIV infectivity: must sex acts be counted?, *Journal of AIDS* **3**, 55–61.

[16] Kim, H.M. & Lagakos, S.W. (1993). Assessing drug compliance using longitudinal marker data, with application to AIDS, *Statistics in Medicine* **13**, 2141–2153.

[17] Lagakos, S.W., Barraj, L.M. & DeGruttola, V. (1988). Nonparametric analysis of truncated survival data with applications to AIDS, *Biometrika* **75**, 515–523.

[18] Liao, J. & Brookmeyer, R. (1995). An empirical Bayes approach to smoothing in backcalculation of HIV infection rates, *Biometrics* **51**, 579–588.

[19] Lui, K.-J., Peterman, T.A., Lawrence, D.N. & Allen, J.R. (1988). A model-based approach to characterize the incubation period of paediatric transfusion-associated Acquired Immunodeficiency Syndrome, *Statistics in Medicine* **7**, 395–401.

[20] Medley, G.F., Billard, L., Cox, D.R. & Anderson, R.M. (1988). The distribution of the incubation period for the acquired immunodeficiency syndrome (AIDS), *Proceedings of the Royal Society of London, Series B* **233**, 267–277.

[21] Pagano, M., DeGruttola, V., MaWhinney, S. & Tu, X.M. (1992). The HIV epidemic in New York City: projecting AIDS incidence and prevalence, in *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz & V.T. Farewell, eds. Birkhäuser, Boston, pp. 123–142.

[22] Rida, W.N. & Lawrence, D.N. (1994). Some statistical issues in HIV vaccine trials, *Statistics in Medicine* **13**, 2155–2177.

[23] Rosenberg, P.S. (1990). A simple correction of AIDS surveillance data for reporting delays, *Journal of AIDS* **3**, 49–54.

[24] Rosenberg, P.S. & Gail, M.H. (1991). Backcalculation of flexible linear models of the human immunodeficiency virus infection curve, *Journal of the Royal Statistical Society, Series C* **40**, 269–282.

[25] Shiboski, S. & Jewell, N.P. (1990). Statistical analysis of HIV infectivity based on partner studies, *Biometrics* **46**, 1133–1150.

[26] Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies, *Biometrics* **51**, 1096–1104.

[27] Taylor, J.M.G. (1989). Models for the HIV infection and AIDS epidemic in the United States, *Statistics in Medicine* **8**, 45–58.

[28] Tu, X.M., Litvak, E. & Pagano, M. (1994). Screening tests: can we get more by doing less?, *Statistics in Medicine* **13**, 1905–1919.

[29] Zeger, S.L., See, L.-C. & Diggle, P.J. (1989). Statistical methods for monitoring the AIDS epidemic, *Statistics in Medicine* **8**, 3–21.

NIELS G. BECKER & JISHENG CUI

Akaike's Criteria

Two of the criteria that have been widely used for model-choice are associated with H. Akaike.

FPE

If a (not necessarily true) p-parameter linear model is fitted by unweighted **least squares** to n observations $y_1, \ldots, y_n$, giving fitted values $\hat{y}_1, \ldots, \hat{y}_n$ and Residual Sum of Squares (*see* **Analysis of Variance**) $(n-p)s_p^2$, say, then the Final Prediction Error criterion (FPE) may be defined by

$$\text{FPE} = \left(1 + \frac{p}{n}\right) s_p^2.$$

This formula was derived by Jones [6] as the "full model" value of a model-choice criterion equivalent to C_p, for observations uncorrelated and homoscedastic (*see* **Scedasticity**) with variance σ^2 (*see* **Mallows' C_p Statistic** Statistic). Jones showed that

$$\text{E(FPE)} \geq \text{E}\left\{\frac{\sum[\hat{y}_i - y_i']^2}{n}\right\},$$

where y_i' is an independent replication of y_i, with equality if $\text{E}(s_p^2) = \sigma^2$ (which almost requires the model to be true after all). The FPE formula was independently derived by Akaike [1] in the context of autoregression (*see* **ARMA and ARIMA Models**), as a model-choice criterion in its own right. Provided p/n and C_p/n are small, FPE and C_p give the same model ranking: for then, denoting equivalence or approximate equivalence by $\cong$,

$$\text{FPE} \cong \left(1 + \frac{2p}{n}\right) \text{RSS}_p \cong 1 + \frac{C_p}{n}.$$

AIC

For a (not necessarily true) model $\mathcal{M}$ with p parameters $\theta = (\theta_1, \ldots, \theta_p)$ and **likelihood** function $L(\theta|y)$ (taken as a rewritten probability density function (pdf)), the Akaike Information Criterion [2] is

$$\text{AIC} = -2\log L[\hat{\theta}(y)|y] + 2p,$$

where $\hat{\theta}$ maximizes $L(\theta|y)$. Under conditions on the likelihood functions in $\mathcal{M}$ that involve their regularity and high informativeness, AIC is an approximately unbiased estimate of

$$\text{E}^{\mathcal{M}} = \text{E}_y\{\text{E}_Y[-2\log L[\hat{\theta}(y)|Y]\},$$

in which Y has, independently, the same unknown true distribution as the data y. Since the inner expectation over Y is a respected **information** theory index of the discrepancy between the pdf of Y (in $\mathcal{M}$) for $\theta = \hat{\theta}(y)$ and its true pdf, AIC has been well received as a model-choice criterion.

When $\mathcal{M}$ is a normal linear model with independent observations with *known* variance σ^2 and we take $L(\theta|y) = \exp\{-\sum[y_i - \text{E}(y_i)]^2/2\sigma^2\}$,

$$\text{AIC} = \frac{(n-p)s_p^2}{\sigma^2} + 2p = C_p + n$$

so that AIC and C_p are then equivalent. If σ^2 is one of the estimated parameters in $\mathcal{M}$ and we take $L(\theta|y) = \sigma^{-n}\exp\{-\sum[y_i - \text{E}(y_i)]^2/2\sigma^2\}$, then

$$\text{AIC} = n\log(\hat{\sigma}_p^2) + 2p + n + 2,$$

where $\hat{\sigma}_p^2 = (1 - p/n)s_p^2$ is the maximum likelihood estimate (mle) of σ^2 in $\mathcal{M}$. When p/n is small,

$$\text{FPE} \cong \log \text{FPE} \cong \log(\hat{\sigma}_p^2) + \frac{2p}{n}$$

and FPE and AIC are then approximately equivalent (which therefore extends to C_p when C_p/n is small too).

The asserted equivalences or approximate equivalences are purely functional and therefore *operationally* effective. "An asymptotic equivalence" of a different sort was proved by Stone [9] between AIC and a particular (not necessarily attractive) choice A of the **cross-validatory** criterion. The connection is of limited interest, since it is an operational, approximate equivalence (for small p/n) only when the model $\mathcal{M}$ used to generate both AIC and A is actually true.

Applications of AIC select models with the smallest value of AIC, a practice labeled MAICE. This practice has been widely criticized for its asymptotic inconsistency in selecting true models, meaning that it has been *theoretically* established that, *for large n and small* p/n, MAICE may have a high probability of picking an "over-fitting" model in which there is

a sharply defined, simpler submodel truly generating the data. This criticism is misplaced, given that AIC was designed to optimize, for ordinarily sized data sets, the approximation of an essentially unidentifiable true distribution, by an almost inevitably untrue model selected by MAICE from some merely reasonable family of distributions.

This is not to say that there is no room for improvement in the performance of MAICE in fulfilling its objective. Most sets of data are not big enough for the complex inferences that scientists wish to draw from them. So the practical effectiveness of model selection methods often depends on their capacity to handle cases where p/n is not small. Since the penalty term $2p$ in AIC was derived for small p/n, there must be doubts whether AIC can handle such cases (see p. 83 of the application-oriented book of Sakamoto et al. [8]). A modification of AIC to deal with moderate to large p/n has been proposed by Hurvich & Tsai [5] which has been generalized to multivariate **y** by Bedric & Tsai [3] and Fujikoshi & Satoh [4].

The book by Linhart & Zucchini [7] gives a useful, nonpartisan overview of the whole area of model choice, including applications.

References

[1] Akaike, H. (1970). Statistical predictor identification, *Annals of the Institute of Statistical Mathematics* **22**, 203–217.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Proceedings of the Second International Symposium on Information Theory*, B.N. Petrov & F. Csaki, eds. Akademiai Kiado, Budapest, pp. 267–281.

[3] Bedrick, E.J. & Tsai, C-L. (1994).. Model selection for multivariate regression in small samples, *Biometrics* **50**, 226–231.

[4] Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and c_p in multivariate linear regression, *Biometrika*, **84**, 707–716.

[5] Hurvich, C.M. & Tsai, C.L. (1989). Regression and time series model selection in small samples, *Biometrika* **76**, 297–307.

[6] Jones, H.L. (1946). Linear regression functions with neglected variables, *Journal of the American Statistical Association* **41**, 356–369.

[7] Linhart, H. & Zucchini, W. (1986). *Model Selection*. Wiley, New York.

[8] Sakamoto, Y., Ishiguro, M. & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. Reidel, Dordrecht.

[9] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B* **39**, 44–47.

(*See also* **Model, Choice of**)

M. STONE

Algorithm

An algorithm is a sequence of instructions for carrying out a well-defined task. The term is generally used to describe a series of logical steps that, when implemented within a computer program, will perform a desired computational task. This article will discuss issues that are generally important in finding or choosing such an algorithm.

The task must be specified precisely. The algorithm should then do the task correctly and do it economically (i.e. with acceptably small computational and memory cost). These, and several related points, will be illustrated with reference to some common tasks. We observe that there are tasks whose computational demands are so severe that they are not tractable, on current computers or on any currently imaginable digital computer.

The sorting of a set of numbers into numerical order is an unambiguous task. But suppose that the permutation (rearrangement) that places the numbers in order is at the same time applied to a second list of numbers. For example, the permutation of the numbers in the first row is at the same time applied to the numbers in the second row.

$$\begin{array}{ccc} 311 & 311 & 231 \\ 7 & 6 & 9 \end{array}$$

There are then two solutions. In one of these, the columns are placed in the order 3, 1, 2. In the second solution, the columns are placed in the order 3, 2, 1. There are several possibilities:

- Either result may be acceptable.
- Whenever there is a choice, the original order is preserved. Thus column 1 would precede column 2.
- Whenever there is a choice, the numbers in the second column appear in numerical order.

In file compression, different algorithms will lead to different compressed files. What is important is not to obtain the same compressed files, but to make a substantial reduction in the size of the file, with modest computational effort. The basis for comparing algorithms that perform reliably is the trade-off that they offer between computational requirements for compression and decompression, and the amount of compression achieved. For information on widely used file compression programs and formats, see the web page `http://www.programmersheaven.com/zone22/cat208/`.

Because different **internet** search engines use different algorithms, they will not always find the same sites and they will rank them differently. Some engines are better for some types of search, for example, searching for papers that have been published on the web, and others for other types of search, for example, searching for travel information.

The calculation of a variance and linear least squares are examples of **floating point** calculations, that is, calculations with decimal numbers that are stored using the floating point format that is standard on modern digital computers. Since these calculations will be carried out on computers that have limited floating point precision, two distinct algorithms that ostensibly solve the same numerical problem will almost inevitably differ in the precision that they achieve on an actual computer.

Finally, note the idea of computational complexity. This has to do, not with the complexity of the code that will carry out the computation, but with the number of operations that are required. Two algorithms that handle the same task can have very different demands; additionally, there are tasks that, irrespective of the algorithm used, make heavy or perhaps impossible demands on computer resources.

For example, one way to assess the significance of a one-sample t-statistic (*see* **Student's *t* Statistics**) is to refer it to its permutation distribution, that is, to the distribution of the statistic that arises from all the 2^n possible assignments of sign ($+$ or $-$) to the n values (*see* **Randomization Tests**). With $n = 20$, there are 1 048 576 possible assignments of sign, and the calculation is tractable. With $n = 50$, however, and assuming a computer that can handle the calculation for 10^6 of the 2^{50} t-statistics in each second, it will take more than 35 years to complete the calculations. Thus, this calculation is not tractable, for any except quite small values of n. Additionally, storage of all the 2^{50} t-statistics would require a little more than a million gigabytes, and might seem an issue for this calculation. Storage problems can however be sidestepped by building up information on the distribution as calculations proceed. Fortunately, calculation of all 2^{50} t-statistics is for practical purposes unnecessary. By taking a sufficient number of random samples from the distribution, the distribution can be estimated with high accuracy.

Algorithm Design

Where several alternative algorithms are available for a task, there are various criteria that may be important in comparing them. Typically, some algorithms will be preferred on one criterion (e.g. high numerical accuracy), and others on other criteria (e.g. computational time and/or storage requirements).

Strategies for algorithm design vary greatly from one problem area to another. The issues involved in designing algorithms for **matrix** computations are different from those for the design of sorting algorithms. Hence the differences in style and content found between Cormen et al. [7] who examine algorithms that are widely important in computing, Higham [11] whose interest is numerical algorithms, and Monahan [15] who examines algorithms that are important for statistical calculations.

A widely applicable general strategy is, in essence, that of top-down programming – complex problems are broken down into simpler subproblems, which are then further broken down. It is sometimes possible to break the problem down into smaller instances of the same problem, in what is called *recursion*. Sorting algorithms are among those that are suited to the use of recursion [7]. Another widely used device is *iteration*, which involves repeated execution of the same set of statements. Recursive algorithms can be attractive because of the initial simplicity and elegance of the code. An iterative version is often more efficient for the final implementation.

The design of a good algorithm will typically involve the use of theoretical insights, heuristic arguments that suggest (but do not establish beyond doubt) effective ways to proceed, trial and error, and the use of insights from the scrutiny of other algorithms that have a similar purpose.

Specific criteria will now be noted that are important for algorithm design. Note that assessment with respect to such criteria as optimality and accuracy can depend strongly on the particular **computer architecture** and organization. Tuning, that is, adaptation of implementation details, is often required to get the best performance on a particular computing system.

Optimality

Optimality has to do with computational cost. Often, the cost is dominated by one particular operation, for example, multiplication, or number of comparisons made. Or it may be dominated by sets of operations that are closely linked, for example, comparisons made and exchange of elements. Analysis of optimality aims to get results that, as far as possible, are independent of the different relative costs, on different computers, of operations that may be of interest. Often it is possible to identify operations or sets of operations that increasingly dominate the cost as the size of the problem (e.g. the number of elements that are to be sorted) increases. These are then the focus of attention. The relative costs of different sorting algorithms, for the sorting of 10 numbers, are of no consequence. For sorting a million numbers, the difference does matter.

Computer Memory or Storage Requirements

Not only do algorithms differ in computational complexity, but also in the amount of storage space that is required. For example, the usual implementations of the *Mergesort* algorithm that I discuss below require the use of two auxiliary arrays, of total length equal to the length of the vector that is to be sorted. By contrast, the *Quicksort* algorithm sorts in place, that is, no additional storage is required. This may be a consideration in choosing between the algorithms.

Simplicity

Simple forms of description make it easier to understand an algorithm, to verify correctness, and to ensure that the algorithm is robust against unusual or extreme inputs. In the choice between computational efficiency and simplicity, the more efficient algorithm will usually be preferred.

Accuracy and Precision

Questions of accuracy and precision arise because calculations with real numbers are carried out using **floating point arithmetic**. A key issue is to ensure that inaccuracies do not accumulate unreasonably; (*see* **Matrix Computations**).

Languages for Describing Algorithms

Computer languages, because they trade off ease of human comprehension against computational efficiency and demands that arise from the tradition of

the language, may not be ideal for the communication of algorithms to other humans. This has been a reason for the use of various forms of pseudocode, where simplicity, ease of description, and communication are the major considerations. The syntax may be modeled on a widely used computer language.

Benefits from using a computer language for describing as well as for implementing algorithms include:

- Implementation on actual machines forces resolution of language ambiguities.
- The description is immediately accessible, without explanation of language conventions, to the community of individuals with skills in the language or in a related language.
- The algorithm can be directly exposed to a range of checks on a computer.
- Changes to the implemented code can be immediately reflected in the description of the algorithm. Such changes may be required to correct bugs, to reduce execution time, to allow the program to handle a wider range of inputs, or to handle illegal inputs.
- Once carefully tested, the implementation is immediately available for incorporation into computer programs.

If a computer language is used, it is important to choose a language whose primitives, that is, the set of abilities that are immediately available in the language, are appropriate for algorithms of this general type. The Matlab language and associated syntax has been popular with numerical analysts. For statistical algorithms, the S language, whether as implemented in **S-PLUS** or as in **R**, may be a good choice. Functions that the algorithm writer may add to the language extend the set of "primitives" that are available for the coding of algorithms. The Perl and Python languages are widely regarded as versatile tools for text manipulation, string, and general-purpose programming.

Practical implementations of computer languages encode a large number of algorithms, which users of those languages take for granted. These include algorithms for parsing source language statements and turning them into more immediately executable code, algorithms for handling addition, subtraction, multiplication, and division, algorithms for extracting square roots, for calculating logarithms, and so on. Already, in the implementation of the statements of the computer language that will be used, algorithms are pervasive. Algorithms for handling such "primitive" operations are primarily a matter for compiler writers and other specialists, and will not be discussed further in this article.

Visual Devices

Visual devices can be helpful in the development and description of algorithms. The best known of these are flowcharts and structure diagrams. Flowcharts emphasize the sequence of operations within the algorithm or, more generally, within a computer program. Structure diagrams emphasize the command structure of the code. The master routine delegates tasks to a succession of lower level routines that, in their turn, delegate tasks to routines that are further down the hierarchy. There is a preference for algorithmic descriptions that are associated with simple forms of structure diagram.

Computational Complexity

Optimality, that is, keeping computational time to a minimum, is a serious issue for operations that must be carried out a large number of times. Hence the emphasis on *computational complexity*, that is, the investigation of how the number of operations of the predominate type increases with the size of the problem. This has led to attention to the asymptotic behavior of algorithms, that is, to the behavior as the size of the problem increases. Sorting algorithms will be used for illustration.

Bubblesort is a simple comparison-based algorithm. While it is not recommended for practical use, it is a useful point of reference, in making comparisons with more satisfactory algorithms. A first pass though the data uses the comparison and perhaps the exchange of adjacent elements to bring the largest element into the final position. The next pass then brings the next largest element into the second last position, and so on. The algorithm requires $N(N-1)/2$ comparisons and, on average, half that number of exchanges. In order to simplify the discussion, exchanges are commonly treated as a tax on comparisons, with the tax rate varying from one algorithm to another and from one computer implementation to another.

It follows from the discussion above that the computational cost of the *Bubblesort* algorithm is

$cN(N-1)$, where c is a suitably chosen constant, so that the algorithm is $O(N^2)$. For certain special inputs, there are sorting algorithms, not based on comparisons, that are $O(N)$. The relevant constant c_0 will be different, and, indeed, the operations that dominate the computational cost are not comparisons. What is important is that for sufficiently large N, in this example for $N > c_0/c$, the inequality $c_0 N < cN^2$ is satisfied. Recall the earlier remark that the interest is in computational cost when N is "large". Algorithms that are $O(N)$ or better are said to be *scalable* [5].

Two efficient and widely used sorting algorithms are *Mergesort* and *Quicksort*. *Mergesort* has best-case, average-case, and worst-case performance that is $O(N \log N)$. Most implementations require the use of an auxiliary array for storage of intermediate results, which can be a disadvantage relative to *Bubblesort* and *Quicksort*. As usually described, *Quicksort* has best-case and average-case performance that is $O(N \log N)$, with $O(N^2)$ worst-case performance. The algorithm can however be modified to be $O(N \log N)$, even in the worst case [7]. Depending on the computing implementation, the factor by which $N \log(N)$ must be multiplied to give a realistic time can be smaller than for *Mergesort*.

More than Polynomial Complexity

An algorithm that is $O(N^p)$ for some p has polynomial complexity. If an algorithm is not $O(N^p)$ for any p, then the computational cost will make the calculation intractable for any except relatively small N. An algorithm that is $O(N^{500})$ would be just about as intractable as an algorithm that is exponential in N, that is, an algorithm that is $O(e^{aN})$ for some $a > 0$. In practice, however, when polynomial complexity can be achieved, p is typically quite small, no more than 3.

In order to illustrate the reach of computational complexity, consider the naive use of **maximum likelihood** methods such as are described in [8], for the estimation of evolutionary trees from nucleic acid sequence data, in what is known as phylogenetic reconstruction. With n taxa, the number of unrooted bifurcating trees with n labelled tips is $(2n-5)!/[(n-3)!2^{n-3}]$. It can be shown that this makes the problem, without accounting for the increase in computation per tree as n increases, at least $O(n^{n-2})$. Examination of all likelihoods is possible only for n less than about 12; for larger n, it is necessary to resort to one of a number of probabilistic and/or heuristic methods that seem likely to give a good approximation to the maximum likelihood solution.

A problem with approximate methods is that it is often impossible to be sure that the problem has been "solved", and there has been a continuing search for "better" methods. The website `http://evolution.genetics.washington.edu/phylip/software.html` is an interesting source of information on the large variety of software and methods that have been developed for the estimation of evolutionary trees.

There is an important special class of problems that are said to be NP-complete. For details of the definition, see [6, 7]. Several are problems of considerable practical importance. For these problems, no polynomial time algorithm is known. All are suspected to require more than polynomial time; however, this has not been proved. Several of them have polynomial time approximate solutions that are adequate for most practical purposes (Cormen et al.[7]). One of the best-known examples is the traveling salesman problem: Given the array of $n(n-1)/2$ distances between n cities that a salesman must visit, find the minimum distance route that passes through all n cities. Cook[6] presents an overview of computational complexity.

Algorithms and Models

Many algorithms that are in practical use lack a totally convincing theoretical basis. They may be developed using a *heuristic* approach, effectively a trial and error approach that tries what seems to make sense, then testing it to see whether it works. When such approaches are used in statistical data analysis, they lead to the use of models that may be fairly described, following Breiman [4], as algorithmic. Although Breiman does not explain how algorithmic models arise, some of the motivations are:

- Extensions of algorithms, for example, normal theory statistical models, into contexts where the theory that motivated the models is no longer plausible.
- Algorithms for models of a learning process, leading to the use of the term *machine learning*. **Neural networks** have this character.

- Algorithms that mimic processes that have been found useful in taxonomic identification or medical diagnosis (*see* **Computer-aided Diagnosis**. **Tree-structured Statistical Methods**, often called decision tree methods (*see* **Decision Analysis in Diagnosis and Treatment Choice**), mimic the creation of a botanical or other taxonomic key.

If the purpose of the model is prediction, then the key issue is the ability to make accurate predictions for data that are different from those used for the development of the model. It is important to devise realistic tests. Further discussion would take us outside of the scope of this present article. See, in particular, the contributions of Cox and Efron to the discussion that followed [4].

Parallel Algorithms

Clustered computing systems, with multiple processors that can be used in parallel, are now relatively cheap and straightforward to build; *see* **Computer Architecture and Organization**. The structuring of algorithms to take advantage of such multiple processor systems and their associated software is, in general, much less straightforward. Problems where the major part of the computation splits cleanly into distinct subproblems, so that the structuring is simple, are said to be *embarrassingly parallel*. Thus, for a parallel sort, the data are split into parts, the parts are sorted separately, and then, as in *Mergesort*, the separate parts are merged. Modern computational power, and such developments as parallel processors, encourage the contemplation of methods for which the computations were formerly prohibitive; for example, **computer-intensive** statistical methods, virtual reality systems, three-dimensional medical imaging (*see* **Image Analysis and Tomography**), global climate models, and so on. Growth in computational power continually extends the range of computations and their associated algorithms that are of practical interest.

Future Reading and References

The books by Knuth [12–14] are classics. Harel [10] has extensive bibliographic notes, useful in indicating where to look for additional information on published algorithms or associated literature. Cormen et al. [7] is encyclopedic in its coverage. As with Harel, the focus is on nonnumerical algorithms. For numerical and matrix algorithms, see [2, 11, 20]. Thisted [21] and Monahan [15] are useful sources of information on statistical algorithms. For algorithm design and analysis, see [1, 3, 17, 19]. Algorithm design is a creative problem-solving activity, to which the discussion in Pólya [16] is relevant. On algorithms for parallel computing, see [9, 18, 22, 23]; (*see also* **Computer Architecture and Organization**).

Sources of information on algorithms and computation include *ACM Transactions on Mathematical Software, Communications of the ACM, Computational Statistics, Computer and Mathematics with Applications, IEEE Transactions on Computers, Journal of Computational and Graphical Statistics, Journal of the ACM, SIAM Journal on Computing*, and *Statistics and Computing*, where research articles appear regularly.

References

[1] Aho, A.V., Hopcroft, J.E. & Ullman, J.D. (1975). *Design and Analysis of Computer Programs*. Addison-Wesley, Reading.

[2] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. & Sorensen, D. (1999). *LAPACK Users' Guide*, 3rd Edn. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[3] Bentley, J. (2000). *Programming Pearls*. 2nd Ed. Addison-Wesley, Reading, MA.

[4] Breiman, L. (2001). Statistical modeling: the two cultures. With discussion, *Statistical Science* **16**, 199–231.

[5] Christen, P., Hegland, M., Nielsen, O., Roberts, S., Strazdins, P.E. & Altas, I. (2001). Scalable parallel algorithms for surface fitting and data mining, *Parallel Computing* **27**, 941–961.

[6] Cook, S.A. (1983). An overview of computational complexity, *Communications of the ACM* **26**, 400–408.

[7] Cormen, T.H., Leiserson, C.E. & Rivest, R.L. (1989). *Introduction to Algorithms*. McGraw-Hill, New York.

[8] Felsentein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach, *Journal of Molecular Evolution* **17**, 368–376.

[9] Gentleman, W.M. (1973). On the relevance of various cost models of complexity, in *Complexity of Sequential and Parallel Numerical Algorithms*, J.F. Traub, ed. Academic Press, London.

[10] Harel, D. (1992). *Algorithms: The Spirit of Computing*, 2nd Ed. Addison-Wesley, Reading.

[11] Higham, N.J. (1996). *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[12] Knuth, D.E. (1968). *The Art of Computer Programming*, *Volume 1: Fundamental Algorithms*. Addison-Wesley, Reading.

[13] Knuth, D.E. (1969). *The Art of Computer Programming*, *Volume 2: Semi-numerical Algorithms*. Addison-Wesley, Reading.

[14] Knuth, D.E. (1973). *The Art of Computer Programming*, *Volume 3: Sorting and Searching*. Addison-Wesley, Reading.

[15] Monahan, J.F. (2001). *Numerical Methods of Statistics*. Cambridge University Press, Cambridge.

[16] Pólya, G. (1945). *How to Solve It: A New Aspect of Mathematical Method*. Doubleday, New York.

[17] Purdom, P.W. Jr. & Brown, C.A. (1985). *The Analysis of Algorithms*. Holt, Rinehart and Winston, New York.

[18] Quinn, M.J. (1987). *Designing Efficient Algorithms for Parallel Computers*. McGraw-Hill, New York.

[19] Sedgewick, R. & Flajolet, P. (1998). *An Introduction to the Analysis of Algorithms*. 2nd Ed. Addison-Wesley, Reading, MA.

[20] Stewart, G.W. (1998). *Matrix Algorithms*, *Volume 1: Basic Decompositions*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[21] Thisted, R.A. (1988). Elements of Statistical Computing, *Numerical Computation*. Chapman and Hall, New York, pp. 134–135.

[22] van de Velde, E.F. (1994). *Concurrent Scientific Computing*. Springer-Verlag, New York.

[23] Wilkinson, D.J. & Yeung, S.K.H. (2004). A sparse matrix approach to Bayesian computation in large linear models, *Computational Statistics and Data Analysis*, **44**, 493–516.

JOHN H. MAINDONALD

Allometry

Allometry is the study of shape differences associated with size. Shape changes in growing organs or whole organisms may be triggered by either biological or physical needs. For example, a simple spherical organism may use its entire surface area for nutrient intake and respiration. If the organism doubles in diameter, then there is a fourfold increase in surface area, A, and an eightfold increase in volume, V. Food requirements are likely to be roughly proportional to volume. Spherical shape can only be maintained if additional nutrient requirements are met by increasing the efficiency of intake or by the inclusion of an increasing amount of inactive organic matter (akin to the woody structure in trees). To both of these there must be an upper limit at which the organism must either stop growing or increase its *effective* A/V ratio either by (i) convolution or branching, or (ii) alteration in shape, implying a different growth rate in some directions or in some parts of the organism. Animals and plants must also adapt their structure to meet physical demands. For example, the strength of a bone is proportional to its cross-sectional area.

The term *allometry* is used by some writers with more specific meanings usually relating to differences in proportions correlated with changes in absolute magnitude of the total organism, or of specific parts under consideration [2]. Variables measured may be morphological, physiological, or chemical. Gould also proposed, and this is now widely accepted, that the term be used regardless of the mathematical expression used to characterize the relationship between variables. We emphasize this point because many bivariate studies have been concerned with the so-called equation of *simple allometry* where two size variables, x and y, satisfy approximately a relationship of the form

$$y = \alpha x^{\beta}, \tag{1}$$

or equivalently

$$\ln\, y = \ln\, \alpha + \beta\, \ln\, x. \tag{2}$$

Relationships of the form (1) or (2) hold between many pairs of size measurements, e.g. between $x =$ *head height* and $y =$ *total height* for individuals in man and many other species from a few weeks after conception to maturity. The ratio x/y decreases with age because head height decreases as a proportion of total height. This is allied to the brain being a larger proportion of total body mass at birth than it is at maturity, because a relatively large brain is needed at birth to ensure that essential bodily functions to maintain life and stimulate growth are possible, but as the individual grows, other parts of the body (e.g. arms and legs) grow at a faster rate than the brain so as to make possible new activities by the developing individual.

Intuitively we think of size measurements x and y like lengths and masses as measures of absolute magnitude, and proportions such as x/y as shape indicators. If the latter remain constant in time, then this implies no shape change with respect to the particular size measurements and this situation is often described as *isometry* and corresponds to $\beta = 1$ in (1). Eq. (1) was proposed by Huxley [5] and has been used in a variety of contexts since.

Extensions from a **bivariate distribution** ($p = 2$) to a **multivariate distribution** ($p > 2$), where p size measurements are made on organs or parts of an organism, have received considerable attention. A diversity of approaches to suitable generalizations from the case $p = 2$ have been advocated by Teissier [9], Jolicoeur [6], Hopkins [4], and others, who recommended analyses of size and shape on the basis of either **factor analysis** or **principal components analysis**. Their approaches were basically empirical and often required assumptions which, although intuitively reasonable, were not readily amenable to statistical verification.

Mosimann [8] made a major breakthrough in the p-variate case by defining *size variables* and *shape vectors* associated, for example, with a set of distances between specified points. He defines sameness of shape in terms of vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ of distances by a relationship $\mathbf{x}_2 = c\mathbf{x}_1$, where c is a constant. Subject to certain axioms, he proved several theorems about size and shape. However, in this context, definitions of equality of shape of two organisms depend critically on how many and what measurements are made. For example, if an animal is 100 cm long from snout to tail, stands 80 cm high at its back legs, and the top of its head is 160 cm above ground, then we may argue on this evidence that it has the same shape as an animal with corresponding measurements 125 cm, 100 cm, and 200 cm, where $c = 1.25$. However, if we know also that the torso length of the

animals are, respectively, 50 cm, and 40 cm, then we would no longer consider the animals to be the same shape.

Bookstein [1] recognized the fundamental weakness in the above approaches to be the linearity implied by distance vectors and the constraints imposed by Euclidean geometry. He proposed two approaches to make the analysis of shape and shape changes more realistic. The first was by introducing concepts of curvature and tangent directions at a series of what he called landmark points. The second was a formalization of the concept of coordinate transformations first proposed by Thompson [10] using biorthogonal grids, and this is relevant also to more general aspects of morphometrics such as comparisons of shapes of corresponding parts of different species. His analyses are essentially mathematical rather than statistical in nature although he recognizes the statistical element in the interpretation of his results. A summary of the recent work on shape analysis using Euclidean distance matrix analysis can be found in [7].

For a review of conceptual and statistical difficulties associated with classic bivariate and multivariate allometry, see [3].

References

[1] Bookstein, F.L. (1978). *The Measurement of Biological Shape and Shape Change*. Springer-Verlag, Berlin.
[2] Gould, S.J. (1966). Allometry and size in ontogeny and phylogeny, *Biological Reviews* **41**, 587–640.
[3] Hills, M. (1982). Allometry, in *Encyclopedia of Statistical Sciences*, Vol. 1, S. Kotz & N.L. Johnson, eds. Wiley, New York.
[4] Hopkins, J.W. (1966). Some considerations in multivariate allometry, *Biometrics* **22**, 747–760.
[5] Huxley, J.S. (1924). Constant differential growth ratios and their significance, *Nature* **114**, 895–896.
[6] Jolicoeur, P. (1963). The multivariate generalization of the allometry equation, *Biometrics* **19**, 497–499.
[7] Lele, S.H. & Richtsmeier, J.T. (2001). *An Invariant Approach to the Statistical Analysis of Shapes*. Chapman & Hall/CRC, Boca Raton, Florida.
[8] Mosimann, J.E. (1970). Size allometry; size and shape variables with characterizations of the lognormal and gamma distributions, *Journal of the American Statistical Association* **65**, 930–945.
[9] Teissier, G. (1955). Sur la détermination de l'axe d'un nuage recitligne de points, *Biometrics* **11**, 344–357.
[10] Thompson, D.W. (1917). *On Growth and Form*. Cambridge University Press, Cambridge.

P. SPRENT

Alternative Hypothesis

An *alternative hypothesis* (usually symbolically represented as H_a: or H_1:) is a statement about a population or set of populations. It stands in contradistinction to the **null hypothesis**, stating what the conclusion of the experiment would be if the null hypothesis were rejected. While the null hypothesis usually expresses what the result of the experiment would indicate if nothing statistically significant results (a negative conclusion), the alternative hypothesis is stated in positive terms, i.e. something of significance is noted. The alternative hypothesis may be *two-sided* or *one-sided*. A *two-sided* alternative hypothesis would conjecture that if the null hypothesis is not acceptable, then the findings might indicate either a larger *or* smaller value for a parameter (for one population), or a difference (direction unspecified) among values for the parameter across the several populations. Notationally, a *two-sided* alternative hypothesis would take the form H_a: $\theta \neq \theta_0$ for a one-population setting, or H_a: not all θ_j are equal, for a setting with several populations. To illustrate, if μ is the population mean for one population and μ_0 is the value specified by the null hypothesis, then H_a: $\mu \neq \mu_0$; and if μ_1 and μ_2 are the means for two populations, then the alternative hypothesis might be H_a: $\mu_1 \neq \mu_2$.

The alternative hypothesis is *one-sided* when a directional relationship is indicated; that is, the value of a parameter (for one population) is smaller or larger (but only one direction is hypothesized), e.g. H_a: $\theta < \theta_0$. If several populations are being studied, a one-sided alternative would indicate specific directional relationships among the values of the parameter for the several populations, e.g. H_a: $\theta_1 < \theta_2$, if $k = 2$.

To illustrate, if μ_1 is the mean length of hospitalization for a population of males following coronary artery bypass surgery (CABG), and μ_2 is the comparable parameter for a population of females, then, if H_0: $\mu_1 = \mu_2$, the one-sided alternative hypothesis might be H_a: $\mu_1 < \mu_2$, i.e. that females have on the average a shorter postoperative hospitalization than males. A specific alternative might state how much shorter, say H_a: $\mu_2 = \mu_1 - 2$; i.e. females average 2 days less than males.

Additional terminology about the null and alternative hypothesis is sometimes used, to indicate the number of possible choices for the parameter under hypothesis. For example, for one population H_0: $\theta = \theta_0$ would be called a *simple* hypothesis because it specifies only a single choice for θ. On the other hand, H_a: $\theta \neq \theta_0$, would be termed a *composite* alternative hypothesis because there are many possible choices for θ (i.e. all values $\neq \theta_0$).

M.A. SCHORK

Alternative Medicine

Alternative medicine is not an entity easy to define. The **World Health Organization (WHO)** has defined alternative medicine as all forms of health-care provision which "usually lie outside the official health sector" [2], and the US **National Institutes of Health** Office of alternative medicine defines it as "therapies that are unproven" [3]. This last definition has the advantage that a proper evaluation of a therapy demonstrating its efficacy would move this therapy from alternative to orthodox. The problem with the word "alternative" is that most so-called alternative procedures are not offered as an alternative but rather as a complement to orthodox medicine; complementary medicine is therefore an expression which is often used and has been recommended. However, restricting the problem to complementary therapies ignores the fact that some proponents of alternative methods do reject orthodox medicine, and claim to offer truly alternative medical systems. The list of procedures that are considered as complementary medicine is very long and the most prevalent forms of therapy are acupuncture, homeopathy, and manipulation, i.e. osteopathy and chiropractic [4].

The proportion of the population reporting use of complementary medicine varies in Europe between 49% in France and 20% in the Netherlands [5], and is equal to 34% in the US [3]. Complementary therapies are generally used because they have fewer side effects than conventional therapies and with the aim to control symptoms. Despite this wide use, these practices remain outside of the mainstream of medicine without being completely accepted nor completely rejected, and their efficacy has not been properly evaluated.

Most alternative therapies have been in use for centuries. Ayur Veda, a traditional healing system from India, currently popular in the US is 5000 years old. The first Chinese text on acupuncture is more than 2000 years old, Hahnemann defined the principles of homeopathy 200 years ago, and Still developed a system of osteopathy in 1874 [14].

Types of Studies

A substantial number of randomized controlled trials (*see* **Clinical Trials, Overview**) evaluating alternative medical practice have been conducted, but few trials used an adequate methodology. The Groupe de Recherches et d'Essais Cliniques en Homoeopathie (GRECHO) trial of homeopathy in post-operative ileus [12] gives a good illustration of a well conducted trial. Six hundred patients undergoing planned abdominal surgery were randomized into four groups (*see* **Randomization**), one being left untreated and the other three receiving, under double blind conditions (*see* **Blinding or Masking**), either opium plus raphanus or opium plus raphanus placebo or a double placebo. Opium was at a dilution of 10^{30} and raphanus at a dilution of 10^{10}. The outcome (*see* **Outcome Measures in Clinical Trials**) was the time to recovery of bowel movements after surgery. The sample size was 600, to have a 95% chance to demonstrate a reduction in the time to recovery of bowel movements from 100 hours to 80 hours assuming a standard deviation equal to 40 hours for this measure (type I error of 5%) (*see* **Sample Size Determination for Clinical Trials**). There were no significant differences between any of the groups ($P > 0.30$). The conclusion of the trial was that the resumption of intestinal transit is not affected by placebo or by opium, either alone or associated with raphanus, in the concentrations studied.

Most studies are not that carefully designed. It is not often that the placebo effect is evaluated by comparing a placebo-treated group with a group left untreated. Some trials are described as randomized but show an obvious misconception of randomization [1]. The sample size is rarely based on statistical considerations and is much too small. The analysis is often biased by exclusion of patients. The main conclusion is often based on results observed on a subgroup of patients.

Proper quantitative **meta-analyses of clinical trials** are very difficult if not impossible to perform because of the heterogeneity between studies both in terms of patient selection and, more importantly, in terms of outcome. Nevertheless, qualitative literature reviews are feasible although they do not provide a synthetic measure of the efficacy of the therapy under study. Meta-analyses of trials evaluating alternative therapies are also difficult since many trial results are published in journals not listed in common databases, and in journals having limited resources to review clinical trial reports.

Landmark Studies

Trials of homeopathy have been reviewed by Hill & Doyon [6] and by Kleijnen et al. [7]. Hill & Doyon's review is restricted to 40 randomized trials, and excludes trials which are described as randomized but show obvious misconception of randomization, whereas Kleijnen et al.'s review includes 107 trials of which only 68 are said to be randomized. The two reviews reach different conclusions: Hill & Doyon concluding that their review failed to provide acceptable evidence that homeopathic treatments are effective, and Kleijnen et al. concluding that the evidence of clinical trials is positive, but not sufficient.

Shekelle et al. [16] reviewed studies of the safety and efficacy of spinal manipulation. Twenty-five controlled trials were identified. The conclusion is that spinal manipulation is of short-term benefit in the subgroup of patients with uncomplicated, acute low-back pain. Data are insufficient concerning the efficacy of spinal manipulation for chronic low-back pain. This review failed to report the results on the overall 25 trials.

Richardson & Vincent [15] and Vincent & Richardson [17] reviewed studies of acupuncture in the relief of pain including both randomized and nonrandomized studies. Their conclusion, based on a selection of studies including those described by the authors themselves as seriously flawed, is that there is good evidence for short-term effectiveness but weaker evidence for longer-term effectiveness.

Kleijnen et al. [8] identified 13 trials described as randomized or double-blind evaluating the efficacy of acupuncture in asthma. They conclude that no studies of high quality seem to have been published, and that claims of the efficacy of acupuncture in the treatment of asthma are not supported by the results of well-performed clinical trials.

Law & Tang [9], in a review of randomized controlled trials of smoking cessation interventions including eight trials of acupuncture on 2759 patients, concluded that acupuncture is ineffective. Li et al. [10] reviewed studies evaluating the effect of acupuncture on gastrointestinal function and disorders, and identified five trials published in China, two being described as randomized. The conclusion is that "more systematic, carefully designed and properly controlled studies are needed".

In conclusion, there does not seem to be any sound evidence of the efficacy of the alternative therapies considered here.

Problems and Solutions

Numerous arguments have been brought forward by practitioners and proponents of alternative therapies for *not* evaluating rigorously alternative therapies, and these therapies have been considered for a long time as very difficult, if not impossible, to evaluate. In 1986, the British Medical Association published a report [14] which concluded that an assessment of the value of alternative therapies

> would be *feasible* in the sense of not being totally impossible. For many therapies a formal trial would be quite inappropriate. In some cases... because the treatment was alleged to be necessarily different for each individual patient, it clearly would rule out any trial based on comparisons between patients.

We think that, in most instances, placebo-controlled trials are feasible and constitute a necessary first step since their aim is to establish the efficacy of the treatment. Placebo-controlled trials are feasible, and indeed have been conducted to study homeopathy, acupuncture, and chiropractic, using placebo drugs for homeopathy and sham acupuncture or a sham manipulation.

For acupuncture and chiropractic, double-blind trials are not recommended since a truly blind procedure would have to be carried out by a naive inexperienced practitioner who may not produce an adequate standard of treatment. Single-blind trials with independent outcome assessment are recommended. For homeopathy, double-blind trials are recommended. The argument that the treatment has to be adapted to each case is not a problem if one uses a whole placebo pharmacopoeia. Contrary to what is stated by Long & Mercer [11], blinding of the therapist does not prevent the monitoring of progress and the alteration of treatment; it is perfectly possible to adapt a placebo treatment.

Practitioners of alternative medicine have argued that trials were not possible because they thought that the methodology of trials imposed a strict selection of the patient population, this selection being based on a classification of diseases that was not relevant to them. We think that **eligibility and exclusion criteria** may be broadly defined, should correspond

to the practice of the therapists, and should respect the uncertainty principle [18].

In the same spirit, alternative therapies usually imply a flexible and individualized treatment. In principle this is not an obstacle to a placebo-controlled randomized trial, but varied sham manipulations may be difficult to organize, and the lack of conviction of the manipulator performing a sham treatment may be associated with a poorer outcome, leading to biased results.

According to Mercer et al. [13], complementary therapists insist on including both subjective and objective measures. We argue that the evaluation of the efficacy of alternative therapies must be based on a main outcome measure, defined uniquely and clinically meaningful. The analysis must be unbiased, based on the **intention to treat** principle, and include all randomized patients. The results should not be based on a subgroup analysis.

In conclusion, alternative therapies must be evaluated as rigorously as conventional therapies: trials should be designed, conducted, and analyzed with adequate methods, and should also comply with the principles of the complementary therapy studied.

Meta-analyses, having the quality of the trials they include, should be avoided because of the poor quality of the trials performed so far. Overviews are preferable. They imply a critical review of the methodology of each trial. They should include nothing but properly randomized trials. Studies where the treatment assigned to the next patient is predictable are not considered as adequately randomized. If possible, the efficacy of the therapy must be evaluated in an objective manner. It is not adequate to summarize the result of each trial by the conclusion it claims to have reached, without any discussion of the validity of this conclusion.

Anticipated Developments and Unresolved Problems

It is amazing that these therapies have been in use for such a long time without ever being properly evaluated. A meta-analysis of 12 surveys studying how physicians perceive complementary medicine in six countries concluded that it may be useful but that randomized controlled trials were urgently needed [4]. Despite this conclusion there does not seem to be any pressure to perform these evaluations.

Because alternative medicine has a cost, the only incentive for a rigorous evaluation will come, if it ever comes, from the regulatory authorities or from health insurance systems, unless the strong lobbies of practitioners and of manufacturers of homoeopathic treatments succeed in maintaining the status quo of wide use without any evidence of efficacy.

References

[1] Aulagnier, G. (1985). Action d'un traitement homéopathique sur la reprise du transit, *Homéopathie* **6**, 42–45.

[2] British Medical Association (1993). *Complementary medicine*. Oxford University Press, Oxford.

[3] Cassileth, B.R. & Chapman, C.C. (1996). Alternative cancer medicine: a ten-year update, *Cancer Investigation* **14**, 396–404.

[4] Ernst, E., Resch, K.L. & White, A.R. (1995). Complementary medicine. What physicians think of it: a meta-analysis, *Archives of Internal Medicine* **155**, 2405–2408.

[5] Fisher, P. & Ward, A. (1994). Complementary medicine in Europe, *British Medical Journal* **309**, 107–111.

[6] Hill, C. & Doyon, F. (1990). Review of randomized trials of homoeopathy, *Revue d'Épidémiologie et de Santé Publique* **38**, 139–147.

[7] Kleijnen, J., Knipschild, P. & ter Riet, G. (1991). Clinical trials of homoeopathy, *British Medical Journal* **302**, 316–323.

[8] Kleijnen, J., ter Riet, G. & Knipschild, P. (1991). Acupuncture and asthma: a review of controlled trials, *Thorax* **46**, 799–802.

[9] Law, M. & Tang, J.L. (1995). An analysis of the effectiveness of interventions intended to help people stop smoking, *Archives of Internal Medicine* **155**, 1933–1941.

[10] Li, Y., Tougas, G., Chiverton, S.G. & Hunt, R.H. (1992). The effect of acupuncture on gastrointestinal function and disorders, *American Journal of Gastroenterology* **87**, 1372–1381.

[11] Long, A.F. & Mercer, G. (1995). *Reviewing the State of the Evidence on Efficacy and Effectiveness of Complementary Therapies*. Nuffield Institute for Health, University of Leeds.

[12] Mayaux, M.J., Guihard-Moscato, M.L., Schwartz, D., Benveniste, J., Coquin, Y., Crapanne, J.B., Poitevin, B., Rodary, M., Chevrel, J.P. & Mollet, M. (1988). Controlled clinical trial of homoeopathy in postoperative ileus, *Lancet* **i**, 528–529.

[13] Mercer, G., Long, A.F. & Smith, I.J. (1995). *Researching and Evaluating Complementary Therapies: The State of the Debate*. Nuffield Institute for Health, University of Leeds.

[14] Payne, J.P., Black, D., Brownlee, G., Cundy, J.M., Mitchell, G.M., Quilliam, J.P., Rees, L. & Dornhorst, A.C. (1986). *Alternative Therapy*. British Medical Association, London.

[15] Richardson, P.H. & Vincent, C.A. (1986). Acupuncture for the treatment of pain: a review of evaluative research, *Pain* **24**, 15–40.

[16] Shekelle, P.G., Adams, A.H., Chassin, M.R., Hurwitz, E.L. & Brook, R.H. (1992). Spinal manipulation for low-back pain, *Annals of Internal Medicine* **117**, 590–598.

[17] Vincent, C.A. & Richardson, P.H. (1986). The evaluation of therapeutic acupuncture: concepts and methods, *Pain* **24**, 1–13.

[18] Yusuf, S., Held, P., Teo, K.K. & Toretsky, E.R. (1990). Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria, *Statistics in Medicine* **9**, 73–86.

CATHERINE HILL & FRANÇOISE DOYON

American Public Health Association

The American Public Health Association (APHA) is the largest organization of public health professionals in the world, numbering over 31 000 members in 1997 organized into 31 sections and forums representing the major scientific disciplines and programmatic concerns of public health. Headquartered in Washington, DC, the main objective of APHA is the protection and improvement of public health by exercising leadership in the development of health policy and action.

APHA programs include publications, conferences, and advocacy and action on public health issues. The *American Journal of Public Health* is a monthly peer-reviewed scientific journal and *The Nation's Health* is a monthly newsletter reporting on legislation and policy issues. Professional publications include *Control of Communicable Diseases in Man* (15th Ed.), *Chronic Disease Epidemiology and Control, Standard Methods for the Examination of Water and Wastewater* (18th Ed.), *Compendium of Methods for the Microbiological Examination of Foods* (3rd Ed.), and *Standard Methods for the Examination of Dairy Products* (16th Ed.). The Annual Meeting and Exhibition, featuring more than 500 scientific and special theme sessions, has been attended by 10 000 to 15 000 public health professionals in recent years. Through its Governing Council and an array of standing and *ad hoc* committees, the Association makes recommendations on public health policy, establishes standards in a variety of public health areas, provides educational material for professional and lay use, collaborates in research projects, and enhances the professional stature of public health workers.

Since its founding in 1872, APHA has been a vigorous advocate for improving public health programs, strengthening the legislative and organizational infrastructure of community programs, and providing adequate resources to improve the health of disadvantaged groups. Much of its initial impact devolved around the control of yellow fever at the end of the nineteenth century, but it was also a leading force for international health activities, promoting a national health board (which finally resulted in the establishment of the US Department of Health, Education, and Welfare in 1953) and advocating sanitation and disease surveillance programs (*see* **Surveillance of Diseases**). Over the years it helped establish appropriate standards for the investigation and control of **communicable diseases**, especially tuberculosis and venereal diseases, nutritional disorders, and maternal and child health programs. The Association was at the forefront in advocating fluoridation of water supplies, limiting tobacco use, reduction of occupational and environmental hazards, control of drug abuse, vaccination programs, and wider choices in reproductive health and family planning. Recent issues addressed by the Association include national health care reform, state and federal funding for health programs, model community health standards, air pollution control, injury and violence, and HIV/**AIDS** programs.

The Statistics Section of the APHA

The Vital Statistics Section was founded in 1908 and renamed the Statistics Section in 1948. When established, the Section accounted for about 13% of the Association's membership, but with the growth of the size of APHA and the scope of public health disciplines and activities, the Section constituted only 1.7% of primary section affiliations in 1997. Following a period of steady growth from about 1911 with 93 members until 1971, when it reached 698 members, the size of the Statistics Section had declined to 539 in 1997. The membership in the early twentieth century was dominated by physicians who had responsibilities for state vital registration systems (*see* **Vital Statistics, Overview**), and by staff of the Bureau of the Census and of insurance companies (*see* **Actuarial Methods**). By the 1930s, academicians became a prominent portion of the membership, and this group has tended to dominate in recent years as biostatistical issues and health research have come to the forefront of statistical activities. In keeping with the evolving issues in public health and the interests of its membership, the Section was first primarily concerned with establishing a national vital statistics program (which was first considered complete for births and deaths in 1932) and in the classification of causes of death (*see* **Cause of Death, Underlying and Multiple**). Subsequently the Section's activities have involved setting standards in statistical practice including definitions used in vital registration records,

tabular presentation, contents of statistical reports of city and state health departments, and use of age adjustment (*see* **Standardization Methods**). In the 1960s, the Section embarked on a project for the APHA to produce a set of monographs on mortality and morbidity for a variety of conditions. The 16 Vital and Health Statistics Monographs were a milestone in health statistics reporting. The Section has maintained interest in, and support for, the National Health Surveys conducted by the **National Center for Health Statistics** and was an advocate for the formation of the National Death Index to aid epidemiologic research (*see* **Health Services Data Sources in the US**). In recent years the Section has dealt with issues at the cutting edge of health statistics including increased use of computers and large databases (*see* **Administrative Databases**), linking of data files (*see* **Record Linkage**), and statistical aspects of **clinical trials** and **health services research**.

The Statistics Section sponsors an annual award in honor of **Mortimer Spiegelman** to recognize outstanding young statisticians working in the public health arena.

MANNING FEINLEIB

American Statistical Association

The American Statistical Association (ASA) is a unique organization among professional associations. ASA's strength comes from the diversity of its membership, which includes different educational levels, disciplines, and types of employers. Members include professionally trained B.S., M.S., and Ph.D. level statisticians, users of statistics, government and industry statisticians, economists, psychologists, chemists, sociologists, and other scientists who use statistical techniques, as well as policy makers who have an interest in data. This diversity creates interaction among members that fosters ASA's continuous evolution, including the creation of new Chapters, Sections, Committees, publications, meetings, and events.

Diversity has been a consistent theme of the ASA since its beginning in 1839, when five men in Boston founded the group whose interest was in gathering and reporting data. Statistics, as a discipline in itself, was not known at that time. Yet these founders had an interest in seeing and analyzing data reported by medical societies, census takers, occupational groups, and others. Much of their interest was in getting facts in front of the public for policy debates and research.

The five founders had different backgrounds – ministry, law, medicine, journalism, and politics. They were interested in **vital statistics**, mortality data, accurate **censuses**, and analyzing data for guidance in diagnosing problems. In fact, an early ASA focus was in the accuracy of the US census. Edward Jarvis, ASA president from 1852 through 1882, worked with the census to improve classifications. Early ASA members helped develop the standards for classification and data presentation.

Much more active advocacy by the ASA took place then than occurs now. One of the first things the ASA did after its creation was to launch a critique of the 1840 census. An interesting debate on how questions are asked and answered was an ongoing part of the critique. The ASA was also active in promoting the creation of a permanent Bureau of the Census in 1902.

Many other professional societies were begun between 1860 and 1900, but today most of them have disbanded. The ASA continued its activities, expanding its growth and influence throughout the twentieth century. ASA moved to New York in the 1920s, then to Washington, DC in 1934. The office is now located in Alexandria, Virginia.

In the beginning, ASA held quarterly meetings with an invited membership. Those small meetings grew into an annual meeting that attracts large numbers of members and guests; in fact there were over 5500 participants in 2003. Initially, there was no ASA national office, as the President and Secretary handled most ASA business in rented office space in Boston. The ASA now has a staff of 37 led by an appointed Executive Director.

As ASA grew to include national and international members in every field of statistical practice, ASA organized into Sections, Chapters, and Committees. Chapters are arranged geographically, representing 77 areas across the United States and Canada. Sections are subject-area and industry-area interest groups covering 22 subdisciplines. Over 60 Committees coordinate meetings, publications, education, careers, and special-interest topics involving statisticians.

To bring together geographically diverse members, ASA created local Chapters. Though there are many Chapters that vie for the honor of being the first, Los Angeles was the first chartered ASA Chapter. Other groups called themselves Chapters and had local meetings, but neglected the official chartering. Chapters provide various services for their local members, depending on their size, constituency, and needs. Each has at least one meeting a year, but many have monthly meetings and the Washington Statistical Society, the Washington DC Chapter, has meetings almost on a weekly basis. Many of the Chapters offer short courses, some have social gatherings, and some publish newsletters posting job openings and updating members on news and upcoming events. Chapter members appreciate the opportunities to hear about new techniques and job opportunities and keep up with the world of statistics. A Council of Chapters, which includes a representative from each of the Chapters, was established in 1984 to govern the Chapters. Nearly half of ASA's membership belongs to at least one Chapter.

ASA membership also has vast subject-area diversity – statisticians work in medicine, education, biometrics, business and economics, government statistics, engineering, quality control, computing, consulting, marketing, sports, the environment, and many

other areas. To be sure that the ASA was meeting the needs of working statisticians, it began establishing Sections. The first was the Biometrics Section, established in 1938. Though the **International Biometric Society** began shortly thereafter, the activity of the Biometrics Section increased. Indeed, the relationship between the Biometrics Section and the Society was strong and continues today. The Biometrics Section is one of the ASA Sections involved in the program for the Eastern North American Region's (ENAR) Spring Meeting of the Biometric Society. In the mid-1940s, a Section on the Training of Statisticians, later renamed the Section on Statistical Education, was formed. Sections did not proliferate at the rate that Chapters did. One reason may have been that Sections were represented on the ASA Board of Directors, and a second may have been that there was little encouragement to form new Sections. However, by the mid-1980s, there was a demand by members to form more Sections. One vital service of Sections is their role in formulating the program for the annual meeting. Each Section has a representative on the Program Committee and organizes a certain number of sessions for the annual meeting. In addition, Sections hold short courses, maintain electronic bulletin boards and email lists for their members, write newsletters, and keep their members informed of new developments in their fields of interest. There are now 22 active Sections in the ASA, with two-thirds of the ASA membership belonging to at least one Section. A Council of Sections, consisting of one representative from each group, governs the Sections.

Unlike Chapters and Sections, which are responses to members' professional interests, Committees are working groups for ASA internal policies and functions. There are many different kinds of Committees, some of which take care of internal business such as ethics, fellows, nominations, and planning. Others are concerned with minority groups, international participation, budgeting, ASA's relations to other professional organizations, or management of the ASA's growing array of journals and magazines. Committees are usually appointed by the President and report to the Board of Directors. There are now over 60 active Committees in the ASA.

Another important service the ASA offers to its members is its publications. In 1888, ASA established the ***Journal of the American Statistical Association*** (*JASA*), which has long been considered the premier journal of statistical science and is now the most widely cited journal in all mathematical sciences. Though the types of articles in it have changed over time, it is a strong link to members' interests, presenting articles describing new theories and applications. *JASA* is primarily seen as an outlet for the publication of academic members' articles, though members who work in government and industry also publish there. *The American Statistician* is aimed at a broader audience. It has articles on theory as well, but also contains commentary on statistical issues of the day, book reviews, software reviews, and a teaching corner. *Amstat News* is a monthly magazine with news about the profession, the activities of the ASA, and its members.

ASA has expanded its publications over the years to include eight professional journals, three magazines, and various brochures and information kits. Many ASA professional journals are sponsored jointly with other associations. *Technometrics* was developed in the l950s with the now named American Society for Quality. The *Current Index to Statistics* is published with the Institute of Mathematical Statistics. The *Journal of Computational and Graphical Statistics* is produced with the Institute of Mathematical Statistics and the Interface Foundation of North America. The *Journal of Agricultural, Biological, and Environmental Statistics* is published jointly with the International Biometric Society. The *Journal of Educational and Behavioral Statistics* is produced with the American Educational Research Association. ASA also offers a book series in conjunction with the Society for Industrial and Applied Mathematics.

As the ASA began working more with people who used data and were interested in statistics but were not trained statisticians, the idea for a publication more of interest to the general public was advanced. *Chance* came into being as a publication of Springer-Verlag, but is now jointly published by Springer-Verlag and ASA, with all editorial content determined by ASA. This publication is popular with a large group of people, but especially among those who teach statistics at the community college and undergraduate level. Another publication of more general interest is *Stats: The Magazine for Students of Statistics*.

ASA offers many of its journals and magazines on the Internet. Members have access to the JSTOR online database of the *Journal of the American Statistical Association* and *The American Statistician* as a benefit of ASA membership. Access to JSTOR

allows members to search all ASA journals, including over 100 years of *JASA* in full-text format. *JASA* is available through JSTOR from 1888, the inaugural year of the journal, through the volume published five years prior to the current year. Other journals and magazines publish abstracts or selected whole articles on the ASA web page or publication-specific sites.

The ASA has always been interested in education. The second Section formed was on the training of statisticians. For many years, the emphasis was on educational opportunities for members, or for statisticians. In the last 20 years, the ASA has expanded its view to include education for nonstatisticians, with a special emphasis on teaching statistics in grade schools and high schools. The ASA has developed workshops, with funding from the National Science Foundation, for teachers who help transfer knowledge about the collection, analysis, and presentation of data. Those workshops for teachers of both mathematics and science have been very effective. ASA also offers educational opportunities for members to include short courses beyond those offered at the annual meeting.

Similarly, the ASA has expanded its meetings. The Joint Statistical Meetings (JSM) is the largest gathering of statisticians held. It is jointly sponsored by the American Statistical Association, International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada. Attended by more than 5500 people in 2003, activities of the meeting include oral presentations, panel sessions, poster presentations, continuing education courses, an exhibit hall with state-of-the-art statistical products and opportunities, a career placement service, society and Section business meetings, Committee meetings, social activities, and networking opportunities. The location of JSM changes every year to ensure all members have an equal opportunity to attend.

Though an annual meeting for the entire membership is a cornerstone, the ASA has other regular meetings. About every other year there is a conference on **radiation** and health, sponsored by one or more federal agencies, depending on the exact topic. There is also a workshop cosponsored by ASA and the **Food and Drug Administration** that takes place each year. Certain Sections, including the Survey Methods Research Section, Health Policy Statistics Section, and Biopharmaceuticals Section, have meetings or conferences every other year on topics of interest to their members. Books that have been highly useful to practitioners have followed these conferences. In recent years, the ASA has cosponsored many different kinds of meetings, ranging from those focused on teaching statistics in business schools to those focused on undergraduate research. The ASA works with its Sections and Chapters to help in setting up meetings.

Over the years, the ASA has set up a series of awards. The most widely known is its selection of Fellows. Each year, members are elected for this honor, recognizing their achievements in statistics and service to the ASA. An annual Founders Award was established to recognize a few members for their exceptional service to the ASA. In total, there are 10 awards given by the national office, plus those given by individual Chapters, Sections, and Committees. Most awards are given on an annual or biannual basis.

The changing needs of the ASA membership keep pushing the Association forward, developing new services and products for its members. The diversity that was present at the beginning of ASA has been responsible for its growth and success as a professional association.

For those interested in more information about the ASA, please contact ASA, 1429 Duke Street, Alexandria, VA 22314-3402, USA; `www.amstat.org`, or email *asainfo@amstat.org*.

BARBARA BAILAR, WILLIAM B. SMITH &
MEGAN R. KRUSE

Analysis of Covariance

Analysis of covariance (ANCOVA), in modern usage, describes statistical models in which a model to compare groups (which usually involves indicator or **dummy variables** such as discussed under **analysis of variance**), incorporates a continuous **covariate** or covariates obtained at the level of the basic measurement unit. The primary focus is usually assumed to be on potential group differences, not on the continuous covariate(s). Snedecor & Cochran [3] describe four basic uses for analysis of covariance models. These include the increase of precision in designed experiments, the adjustment for sources of **bias in observational studies**, to throw light on treatment effects in randomized experiments (*see* **Randomized Treatment Assignment**), and the study of **regression** in multiple classifications. In what follows, we discuss these basic uses and the statistical modeling involved. We also discuss the estimation of adjusted treatment means, and the interpretation and pitfalls associated with the technique.

Increasing Precision in Designed Experiments

As discussed in the article on **Analysis of Variance**, an experimenter can often group experimental units into homogeneous sets by **blocking** on factors related to the response. The levels of the treatment can then be assigned, randomly, to the units within a block. This blocking reduces experimental error, and increases the precision with which we can make statements about the treatment effects. However, blocking on all factors related to the response is not feasible, and even after blocking on major factors, there may be other variables related to the response, which can be measured on the experimental units. The analysis of covariance is a statistical technique developed originally to allow experimenters to incorporate information on a factor or factors which varied across experimental units, and which thus contributed to the experimental error term in the analysis of variance models. In agricultural experiments, for example, by using fields as blocks, a number of climate variables could be controlled. However, even within a field, there could be variation in fertility from plot to plot. If the effect of that fertility variable could be captured in an appropriate statistical model, then the unexplained variability would be reduced and, thus, more precise statements about the effects of the treatments could be made.

In a study of the effects of drug treatment on blood pressure, we may determine that age, gender and treatment center may be important blocking variables because of their potential relationship to blood pressure, and the feasibility in a controlled trial of assigning patients to blocks defined by levels of these factors. However, a subject's blood pressure at pretest will very likely be related to posttest blood pressure. If we can incorporate information on pretest blood pressure into the analysis of variance model, then we should be able to reduce the experimental error term and make more precise comparisons of the effects of the drug (*see* **Baseline Adjustment in Longitudinal Studies**). Cox [1] calls such supplementary observations, which may be used to improve precision, *concomitant variables*.

It is important to consider the scientific requirements, or assumptions, on concomitant observations [1] in order that comparisons between treatments "adjusted" for levels of the covariate are meaningful. These requirements are given below, along with the consequences of a violation of the assumptions. In general, if there is a constant difference in the response between the treatment groups for all values of the concomitant variable(s), then the adjusted treatment effect is this constant difference. However, because we have not had control of the values of the concomitant variable, we must rely on a statistical model to make the adjustment. The following are the requirements for these concomitant observations.

1. The concomitant observation is assumed to represent a factor (or a constellation of factors, e.g. fertility) at the level of the experimental unit which is *unaffected by the treatment*. It can be measured *before* the treatments are assigned or applied (e.g. pretest blood pressure), *before* the effect of the treatment has had time to develop, or on some part of the process which is *not* related to the assignment of the treatment (e.g. the time of day that the measurement is taken). Randomization should ensure that the covariate does not differ significantly between treatment groups; however, this can be checked. If the groups differ with respect to the covariate, they

will be compared at a value of the covariate which is not typical of either group. Further, the statistical model will need to **extrapolate** beyond the region where there is the most data for both groups, and this makes an assumption that the statistical model is correct in that region, for both groups. If the treatment *does* affect the concomitant variable, then adjusting for differences in the concomitant variable may well "adjust away" the actual treatment differences.

2. The relationship between the response and the covariate must be the same for all groups. That is to say, there is no **interaction** between the treatments and the covariate (*see* **Treatment-covariate Interaction**). If there is an interaction, then it does not make sense to compare the groups at a single value of the covariate, since any difference noted will not apply for other values of the covariate. This assumption is equivalent to saying that the relationship between the response and the concomitant variable(s) should appear as parallel curves, one for each treatment group [1]. Again, the assumption of parallel curves can be checked at the time of modeling.

Adjustment for Bias in Observational Studies

In **observational studies**, groups may differ on factors related to the response because of naturally occurring phenomena that lead to self-selection into groups, or historical differences in how groups have developed. Snedecor & Cochran [3] give the example of a study examining the relationship between obesity and physical activity on the job. Since obesity is related to age, and there may be differences between the age structures for different occupations, adjusting for age differences between the workers in the study by analysis of covariance will help model the response and adjust for the age bias. However, for the reasons noted above, considerable care needs to be taken in the interpretation of such models for the following reasons.

1. If the occupations differ significantly with respect to age, then we will obtain a comparison of the relationship between high and low activity occupations for an individual of a fixed age. There may be few individuals of that age in either occupation, so the comparison may be meaningless, and will necessarily involve an extrapolation of the statistical model.
2. The observational nature of the data make it impossible to determine if it is the physical inactivity on the job which leads to the obesity, or whether individuals who are obese select certain occupations. While the analysis of covariance may allow for a comparison adjusting for age, it cannot answer this very basic question.

The Nature of Treatment Effects in Designed Experiments

In many investigations there will be variables that can be measured on the experimental units which can help elucidate the mechanisms responsible for the treatment effects. For example, in the blood pressure study mentioned earlier, we could also measure a subject's pretest and posttest weight. A natural question is whether the posttest blood pressure can be explained by the change in weight which might have occurred over the course of treatment. Analysis of covariance provides a modeling approach to examine the differences between groups in posttest blood pressure using pretest blood pressure and weight change as covariates. It is important to note that adjusting for weight change may, in fact, remove the difference between groups; however, to claim on this basis that the drug had no effect would be misleading if the effect of the drug was to increase weight loss. In this latter case we might report that there was a difference between the two groups in posttest blood pressure (adjusting for pretest differences in blood pressure), and that, at the same time, there was a difference in weight loss; the loss of weight was associated with a reduction in blood pressure.

Again, this example points out the need to be very careful in interpreting the differences (or lack thereof) between adjusted means. Cox [1] presents a thorough discussion of this type of situation, in which he interprets a series of possible data patterns.

Regression in Multiple Classification

Suppose we wished to model and examine the relationship between blood pressure and physical activity across several socioeconomic strata, and age groups. We could develop a regression model with indicator variables which coded for socioeconomic

status and age group, and a continuous variable which provides the measure of physical activity for each individual. By adding regression variables which code for the interaction between socioeconomic status and physical activity, say, we could assess whether the relationship between blood pressure and physical activity was the same for all socioeconomic strata. If there was no interaction, the significance of the indicator variables coding for socioeconomic status would assess whether there are significant differences in blood pressure between levels of socioeconomic status, after adjusting for physical activity differences through the statistical model (i.e. when we compare two different individuals with the same level of activity).

Again, the caveats discussed above apply. We should determine whether physical activity differs significantly between groups and whether the response curves are parallel, before proceeding to discuss these adjusted means.

The Model

To illustrate the technique, we consider a randomized block experiment (*see* **Randomized Complete Block Designs**) in which t treatments are randomly assigned to experimental units in each of b blocks. Furthermore, assume that for each experimental unit, we have a measurement on an explanatory variable (e.g. pretest blood pressure). If we define the covariate measurement from the plot in the jth block which received the ith treatment as x_{ij}, we can write the model as

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma(x_{ij} - \overline{x}_{..}) + \varepsilon_{ij},$$

$$\text{with constraints } \sum_{i=1}^{t} \alpha_i = 0, \sum_{j=1}^{b} \beta_j = 0.$$

(Note that the subscript dot indicates that averaging has been taken over the relevant subscript(s), so $\overline{x}_{..}$ is the average of the covariate over all tb experimental units.)

In the model, μ represents the overall mean effect, α_i and β_j represent the effect of the ith treatment and jth block, respectively, and ε_{ij} represents the unexplained variation in Y_{ij}. We usually assume $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$, independently.

In addition, we add the term $x_{ij} - \overline{x}_{..}$ to reflect the variation in the covariate about its average; γ is the regression coefficient associated with that covariate. If $\gamma = 0$, then the covariate information is not related to the response after adjusting for treatments and blocks, i.e. there is no increase in precision resulting from adjusting for the covariate. The transformation of the covariate from x_{ij} to $x_{ij} - \overline{x}_{..}$ is useful to produce a regression variable which is **orthogonal** to the mean. Furthermore, note that if $x_{ij} = \overline{x}_{..}$, then the expected value of the ith treatment mean is $\mathrm{E}\left(\sum_{j=1}^{b} Y_{ij}/b\right) = \mu + \alpha_i$. That is, "adjusting for the covariate" provides an estimate, via the statistical model, for the treatment mean for a hypothetical plot with $\overline{x}_{..}$ as covariate value (see subsequent discussion).

Estimates of Parameters

From this model, the **least squares** estimates (**maximum likelihood** estimates under the above model for the ε_{ij}s) are

$$\hat{\mu} = \overline{y}_{..};$$

$$\hat{\alpha}_i = (\overline{y}_{i.} - \overline{y}_{..}) - \hat{\gamma}(\overline{x}_{i.} - \overline{x}_{..});$$

$$\hat{\beta}_j = (\overline{y}_{.j} - \overline{y}_{..}) - \hat{\gamma}(\overline{x}_{.j} - \overline{x}_{..});$$

and

$$\hat{\gamma} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{b}(Y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})(x_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}_{..})}{\sum_{i=1}^{t}\sum_{j=1}^{b}(x_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}_{..})^2}.$$

If we define E_{yy}, and E_{xx} to be the Residual Sum of Squares from a randomized block model for the Y_{ij}s and x_{ij}s, respectively, and E_{xy} to be the Residual Sum of Cross-Products, then we have

$$\hat{\gamma} = \frac{E_{xy}}{E_{xx}}.$$

Note the similarity between the above expression for $\hat{\gamma}$ and the expression for the least squares estimate of the slope in a simple **linear regression** model. Thus, the estimate of γ is equivalent to that obtained from a simple linear regression of the **residuals** from a randomized block model for the Y_{ij}s on the residuals from a randomized block model for the x_{ij}s. The estimate will be different from zero when

there is residual variation in the Y_{ij}s explained by the x_{ij}s.

The estimate of σ^2 can be obtained as usual from the $\hat{\varepsilon}_{ij}$s; that is,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{b}(Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}(x_{ij} - \overline{x}_{..}))^2}{(t-1)(b-1)-1} = \frac{(E_{yy} - E_{xy}^2/E_{xx})}{((t-1)(b-1)-1)}.$$

We note that the first term in the numerator is the Residual Sum of Squares from the randomized block model for the Y_{ij}s. The second term, which must be positive, will significantly reduce the randomized block Residual Sum of Squares for the Y_{ij}s whenever $\hat{\gamma}$ is significantly different from zero.

The extension to additional covariates, or to **transformations** of the covariates [e.g. adding $(x_{ij} - \overline{x}_{..})^2$ to the above model] requires no new theory. It will become more difficult to compute the required quantities as the model becomes more complex, but with existing **software** these computations are easily performed.

Adjusted Treatment Means: Estimation and Testing

As discussed above, the analysis of covariance allows us to estimate treatment means which are adjusted for the covariate via the statistical model (i.e. estimated at $x_{ij} = \overline{x}_{..}$). For example, for the randomized block design above, the adjusted treatment mean for treatment i is

$$\sum_{j=1}^{b}\frac{\hat{Y}_{ij}}{b} = \hat{\mu} + \hat{\alpha}_i = \overline{y}_{i.} - \hat{\gamma}(\overline{x}_{i.} - \overline{x}_{..}).$$

Thus, if $\hat{\gamma}$ is positive (i.e. in general, large values of Y_{ij} are associated with large values of x_{ij}), and $\overline{x}_{i.} > \overline{x}_{..}$ (reflecting the fact that, on average, treatment i was applied to units with larger values of the covariate), then the observed mean for treatment i will be adjusted downwards, as required. The ith adjusted treatment mean has variance

$$\text{Variance}_{\text{Adjusted Treatment Mean}} = \sigma^2\left[\frac{1}{b} + \frac{(\overline{x}_{i.} - \overline{x}_{..})^2}{\sum_{i=1}^{t}\sum_{j=1}^{b}(x_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}_{..})^2}\right],$$

and, using the estimate for σ^2 given earlier, we can obtain **confidence intervals** for any adjusted mean; the relevant reference distribution being $t_{(t-1)(b-1)-1}$, i.e. **Student's *t*** with $(t-1)(b-1)-1$ **degrees of freedom**.

Likewise, we can test an hypothesis based on a **contrast** in the adjusted treatment means. If $\boldsymbol{c}_k$ is a $t \times 1$ vector of constants with $\sum_{i=1}^{t} c_{ik} = 0$, then, similarly to the development given in the analysis of variance entry, the ratio

$$F = \frac{\left(\sum_{i=1}^{t} c_{ik}\,\text{Adjusted Treatment Mean}_i\right)^2 \Big/ \sum_{i=1}^{t} c_{ik}^2}{\text{Residual Mean Square}_{\text{Full Model}}},$$

will follow the $F_{1,(t-1)(b-1)-1}$ distribution (i.e. ***F* distribution** with 1 and $(t-1)(b-1)-1$ degrees of freedom) under the hypothesis that the contrast in the (adjusted) treatment means is zero.

To test the global hypothesis that the adjusted treatment effects are zero, the extra sum of squares principle F test (*see* **Analysis of Variance**) can be employed. That is, under the hypothesis H: $\alpha_i = 0, i = 1, \ldots, t$, the model becomes

$$Y_{ij} = \mu + \beta_j + \gamma(x_{ij} - \overline{x}_{..}) + \varepsilon_{ij},$$

which is the model for a completely randomized design (one-way classification) with a covariate observed on each plot.

If T_{yy} represents the "Treatments Sum of Squares" for the response variable under the randomized block design, and T_{xx} and T_{xy} are defined analogously for the x_{ij}s and the cross-products of the response and covariate, then, following the discussion of the randomized block design, we can estimate γ in the

hypothesized model as

$$\hat{\hat{\gamma}} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{b}(Y_{ij} - \overline{y}_{.j})(x_{ij} - \overline{x}_{.j})}{\sum_{i=1}^{t}\sum_{j=1}^{b}(x_{ij} - \overline{x}_{.j})^2} = \frac{(E_{xy} + T_{xy})}{(E_{xx} + T_{xx})}.$$

The Residual Sum of Squares from the hypothesized model is

$$\text{Residual Sum of Squares}_{\text{H: }\alpha_i=0} = \sum_{i=1}^{t}\sum_{j=1}^{b}((Y_{ij} - \overline{y}_{.j}) - \hat{\hat{\gamma}}(x_{ij} - \overline{x}_{..}))^2 = E_{yy} + T_{yy} - \frac{(E_{xy} + T_{xy})^2}{(E_{xx} + T_{xx})}.$$

Then, the difference in the Residual Sum of Squares between the hypothesized model and the full model forms the basis for the extra sum of squares principle F test. Thus, if

$$\Delta = \text{Residual SS}_{\text{H: }\alpha_i=0} - \text{Residual SS}_{\text{Full Model}},$$

then the extra sum of squares principle F test statistic is

$$F = \frac{\Delta/(t-1)}{\text{Residual SS}_{\text{Full Model}}/((t-1)(b-1)-1)}.$$

The resulting value can be compared with tables of the $F_{(t-1),(t-1)(b-1)-1}$ distribution.

The difference between the two Residual Sums of Squares in the numerator is

$$\Delta = T_{yy} - \left(\frac{(E_{xy} + T_{xy})^2}{(E_{xx} + T_{xx})} - \frac{E_{xy}^2}{E_{xx}}\right),$$

which we note is the Treatment Sum of Squares for the response, T_{yy}, adjusted for the covariate.

With certain software packages (e.g. SAS [4]), the programs that fit normal linear models will compute type III sums of squares and mean squares (*see* **Analysis of Variance**). The type III sums of squares provide, for each source, the sum of squares after adjusting for all other terms in the model. Thus, if the covariate is included in a model, the type III mean square for treatments will be the relevant numerator for the F test, and the F test will be provided in the table. For interactive modeling, such as that provided by GLIM [2] (*see* **Software, Biostatistical**), the change in deviance when the "factor" coding for treatments is dropped from the full model, divided by $(t-1)$, will give the numerator for the test, the denominator being the deviance for the full model divided by its degrees of freedom.

Checking Assumptions on the Covariate

In this section we briefly discuss how the important issues surrounding the meaning of "adjusting for the covariate" can be assessed. For a discussion of methods for checking other aspects of the fit of the model, see **Model Checking, Diagnostics,** and **Residuals**.

The first concern described above is that if the covariate is significantly different for different treatments, then the adjustment may require extrapolation beyond the data on which the model was developed to covariate/treatment combinations which did not (or may not) occur. This raises both scientific and statistical concerns. Related to this, if the treatment affects the covariate, then we may adjust away the effect of the treatment on the response by adjusting the treatment means to the same value of the covariate. An analysis of variance conducted on the x_{ij}s will help address this concern. If there are significant treatment effects for the x_{ij}s, we must be very careful in interpreting the results of the analysis of covariance.

The second issue was that the relationship between the response and the covariate should be a set of parallel curves, one for each treatment, in order that we get the same treatment difference at all values of the covariate. In the linear relationship case, we can address this issue by adding $t-1$ columns to the X matrix, representing the product of the covariate and each of the $t-1$ columns coding for the treatments. If adding (deleting) these $t-1$ interaction columns to (from) the model reduces (increases) the Residual Sum of Squares significantly, we have evidence that the curves are not parallel. If the model contains m columns in X to represent the effect(s) of the covariate(s), we require $m \times (t-1)$ additional columns to assess fully the interactions. These should be added (deleted) in sets of $t-1$.

Thus, we should:

1. Check the covariate(s), via analysis of variance, for significant treatment effects.
2. Fit the analysis of covariance model, and the model with interactions between treatments and the covariate(s). Check the significance of the interactions of the covariate and treatments.
3. If there is no difference between treatment groups in the covariate, and no interaction between treatments and the covariate in the augmented analysis of covariance model, proceed to the analysis of covariance as described earlier. If problems are identified, refer to scatter plots (*see* **Graphical Displays**) of the relationship between the response and the covariate to determine the type of statement which can be made safely.

Comparison with other Modeling Strategies

Consider the experiment described earlier. Subjects with high blood pressure are randomly assigned (m per group) to a drug group or a placebo group. A pretest measure of blood pressure is taken and then blood pressure is reassessed at posttest. We have (at least) three choices for the analysis.

1. Take the difference in blood pressure for each subject (posttest−pretest), and analyze these differences via a **paired *t* test**. This analysis investigates whether the mean blood pressure change is the same for the two groups.
2. Treat the two measures of blood pressure as the response, and analyze the data as for a repeated measures design (*see* **Longitudinal Data Analysis, Overview**). This is an example of the class of designs described in the analysis of variance entry where there are two sources of experimental error to consider: *between* subjects and *within* subjects. The relevant question is whether there is an interaction between treatment and test. That is, is the relationship between pre- and posttest measures the same for both groups?
3. Use the pretest measure as a covariate, and then conduct an analysis of covariance with the posttest measure as response. This analysis compares the mean posttest blood pressure at the average pretest blood pressure for the two groups.

To compare these analyses, define Y_{ijk} to be the measured value of blood pressure for the kth test on the jth subject in the ith group. A model for the first analysis can be written as

$$Y_{ij2} - Y_{ij1} = \mu^* + \alpha_i + \varepsilon^*_{ij},$$

or

$$Y_{ij2} = \mu^* + \alpha_i + Y_{ij1} + \varepsilon^*_{ij},$$

where μ^* is the overall mean, α_i is the effect of the ith group, and ε^*_{ij}, the experimental error term, reflects the effects of uncontrolled factors which differ from pretest to posttest for the same subject. We note that this model is just a special case of the model developed above for the analysis of covariance. Specifically, if $\gamma = 1$, and $\mu = \mu^* + \overline{y}_{..1}$, we get the above model from the analysis of covariance model. Thus, the analysis of covariance provides a more general analysis of these data.

A model for the second analysis can be written

$$Y_{ijk} = \mu + \alpha_i + \delta_{ij} + \tau_k + (\alpha\tau)_{ik} + \varepsilon_{ijk},$$

where δ_{ij} and ε_{ijk} are the between-subjects and within-subjects experimental error terms, respectively, τ_k represents the effect of the kth test time, and $(\alpha\tau)_{ik}$ represents the effect of the interaction of the ith level of treatment group and the kth test time.

In this model, the relevant hypothesis is $\mathrm{H}: (\alpha\tau)_{ik} = 0, i = 1, 2; j = 1, 2,$ which states that the effect of test time is the same for both groups. This test is algebraically equivalent to the test of no difference between the adjusted treatment means as provided by the analysis of covariance. Consequently, we could model this situation using either approach. However, the analysis of covariance may be easier to interpret and explain, and by considering the assumptions of no differences in pretest measures, and parallel curves for the different treatments, we are more naturally led to assess whether, for instance, the difference in posttest blood pressure for each treatment is the same at all levels of pretest blood pressure. Furthermore, it provides the additional information concerning the relationship between the pretest and posttest blood pressures.

Concluding Remarks

The above discussion has concentrated on the analysis of covariance in designs in which there is estimation of a single source of experimental error. More complicated designs, such as those discussed in the analysis of variance entry, allow for the estimation of two or more sources of error, such as between-subjects variation due to differences in the subjects, and within-subjects variation due to differences between the test times for multiple measurements on the same subject. In such designs, covariate information could be available on the subjects (e.g. age) and/or on the test times (e.g. temperature). The analysis and interpretation of treatment differences will need to consider the relevant covariates. Again, this requires very careful modeling to ensure that the appropriate estimates of error are used in assessing treatment differences.

This same modeling strategy can be applied beyond the normal linear model described above to **generalized linear models**, allowing for the adjustment for covariates in a wide range of models.

Since the analysis of covariance models are just regression models in which there are both indicator variables and continuous covariates, the theory of model fitting, and inference about regression parameters is not different than that discussed elsewhere for **multiple linear regression** models. However, as with any statistical modeling, it is essential that the interpretation of the model be correct and clearly presented. The discussion of these models in Cox [1] provides an excellent overview of the issues involved in interpretation.

References

[1] Cox, D.R. (1958). *Planning of Experiments*. Wiley, New York.
[2] NAG (Numerical Algorithms Group) (1985). *The GLIM System Release 3.77 Manual*. NAG, Oxford.
[3] Snedecor, G.W. & Cochran, W.G. (1967). *Statistical Methods*, 6th Ed. Iowa State University Press, Ames.
[4] SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6*, 4th Ed., Vol. 2. SAS Institute Inc., Cary.

K.S. BROWN

Analysis of Variance for Longitudinal Data

Analysis of variance (ANOVA) methods are some of the most commonly used techniques for the analysis of sources of variation in both experimental and observational studies. It therefore seems quite natural to ask how they might be adapted for use on **longitudinal data**. Typically, one might recognize data coming from a longitudinal study as coming from a nested design (a **split-plot** experiment, for example), with the serial measurements (subplots) being nested within each of the subjects (plots). Designs involving repeated or serial measurements differ from the typical nested experiment, however, in that the times of observation must by their very nature follow a strict temporal sequence and that, in general, measurements made on occasions close together will be more highly correlated than those further apart. Be warned: *an analysis that ignores* **serial correlation** *will almost certainly be invalid.*

This article briefly introduces the use of ANOVA methods for longitudinal data – typically, repeated measures experiments – and then discusses assumptions necessary for their validity, methods of testing for departures from these assumptions and, finally, methods for the adjustment of F tests to compensate for these departures. In case the reader feels that, after reading about the pitfalls of the use of ANOVA, it would be better to use other methods of analysis, these other methods (particularly **multivariate analysis of variance** (MANOVA) and random effects models for longitudinal data) are briefly mentioned. We will illustrate the methods for two simple repeated measures designs. In the first, each member of a single group of subjects provides measurements on each of k separate occasions. The second design involves the use of two or more groups of subjects, and again each subject within these groups provides measurements on k occasions. For both designs it will be assumed that the timing of the measurements is the same for all subjects, but the spacing between successive measurements does not, necessarily, remain constant. The more complex repeated measures designs involving changing treatment conditions over time (**crossover designs**) will not be discussed here.

Growth Curves in a Single Group of Subjects

In this design each member of a group of n subjects is observed on each of k occasions. The structure of the data is analogous to that arising from a randomized blocks experiment, with the times corresponding to treatments and the subjects being the blocks. Beware of the differences, however. The order of the successive measurements in the longitudinal data is fixed, but the allocation of treatments within blocks is randomized. There will be serial correlation in the longitudinal data that will not be present in the data arising from randomized blocks. The aim of the analysis, however, is similar. We wish to test for differences in the means for the different occasions (treatments), having first allowed for overall subject (block) differences.

The Assumed Model

The statistical model for the repeated measures data is given by

$$Y_{ij} = \mu + \alpha_j + \omega_i + \varepsilon_{ij},$$

where Y_{ij} is the measurement for the ith subject on the jth occasion, μ is the grand mean, α_j is the effect associated with the jth occasion, ω_i the effect associated with the ith subject and, finally, ε_{ij} is the error term for the ith subject on the jth occasion. Note that as there is no replication of measurements on each occasion we have to assume that there is no time-by-subject interaction (analogous to the assumption of no treatment-by-blocks interaction in the randomized blocks experiment) – this interaction being completely confounded with the error term in the following ANOVA model. It is very straightforward to carry out a routine two-way analysis of variance and construct an F test for the null hypothesis that $\alpha_j = 0$ for all j. The resulting F statistic for a test of time trends will have $(k-1)$ and $(n-1)(k-1)$ degrees of freedom. Note in passing that if $k = 2$, the above analysis is equivalent to carrying out the much simpler **paired *t* test** on the changes between the first and second occasions. The F test from one is the square of the t statistic from the other, and the resulting P values are therefore identical.

Necessary Assumptions for a Test of Time Trends

As in any analysis of variance, in order for the above F test to be valid we need to make a series of assumptions. We assume that the repeated measures are measured on an interval or ratio scale of measurement (*see* **Measurement Scale**). We also make the usual assumptions concerning random sampling from the population, independence of subjects, and normality. There is also a homogeneity assumption similar to that required for between-subjects designs. This is the assumption that for any two occasions – say r and s – the difference $Y_r - Y_s$ must have the same population variance for every pair of occasions. The variance of the difference $Y_r - Y_s$ can be written as

$$\begin{aligned}\sigma^2_{Yr-Ys} &= \sigma^2_{Yr} + \sigma^2_{Ys} - 2\text{cov}(Y_r, Y_s)\\ &= \sigma^2_{Yr} + \sigma^2_{Ys} - 2\rho_{rs}\sigma_{Yr}\sigma_{Ys},\end{aligned}$$

where ρ_{rs} is the population correlation for measurements taken on occasions r and s. Huynh & Feldt [4] and Rouanet & Lépine [8] showed that this homogeneity of the time-difference variance assumption is equivalent to the assumption that the population covariance has a certain form, referred to as **sphericity** or *circularity*. A special case of sphericity is *compound symmetry*. A covariance matrix possesses compound symmetry if and only if all the variances are equal to each other and all the covariances are equal to each other. An equivalent statement is that all the repeated measures have the same variance ($\sigma^2_r = \sigma^2_s = \sigma^2$, for all r and s) and that the correlations between measures for all pairs of occasions are equal ($\rho_{rs} = \rho$, for all r and $s, r \neq s$). Under the assumption of compound symmetry, the variance of the difference between Y_r and Y_s becomes $2\sigma^2(1-\rho)$ for any r and $s, r \neq s$. In summary, compound symmetry implies sphericity, but not vice versa. In practice, however, one is unlikely to come across data demonstrating sphericity but not compound symmetry. One exception is in a simple follow-up study with $k = 2$. In this situation, the sphericity assumption always holds (there being only two variances and a unique correlation) but compound symmetry very often will not (the variation at follow-up being higher than at the start of the experiment, for example). When $k = 2$, the sphericity assumption is not needed and the F test is always valid, provided, of course, that the other assumptions are true.

Several tests of sphericity are available, the most commonly used being that according to Mauchly [6]. These sphericity tests are of limited value, however, because of their sensitivity to nonnormality. The Mauchly test seems to be poor for detecting small departures from sphericity despite the fact that these small departures can produce substantial bias in the standard F tests. Several authors, including Hand & Crowder [3] and Maxwell & Delaney [7], for example, recommend making the more realistic assumption that sphericity does *not* hold and automatically adjust the F tests accordingly (see below). The reader can find a brief discussion of the Mauchly test in [3]. For tests of departure from compound symmetry, see [8].

Adjustments to the F tests

When the assumption of sphericity is false, it is possible to perform an adjusted F test of the equality of means over time. Box [1] derived a measure, usually denoted by ε, which measures the departure of the covariance matrix from sphericity. A matrix that displays sphericity always has an ε-value of one and departure from sphericity leads to a lowering of ε, with an absolute minimum of $1/(k-1)$. Box [1] showed that if one calculates the required F statistic and compares it with the critical values from an $\boldsymbol{F}$ **distribution** with numerator and denominator degrees of freedom given by $\varepsilon(k-1)$ and $\varepsilon(n-1)(k-1)$, respectively, then the results will be approximately correct. On its own, however, this information is of limited value since we do not know the true value of ε. One solution to the problem is to use an estimate of ε, based on the observed covariance matrix – see [3] for further details. This is the so-called Greenhouse–Geisser estimate provided by many software packages. A more refined estimate of ε, due to Huynh & Feldt [5], is also produced by these software packages, although Maxwell & Delaney [7] recommend the routine use of the Greenhouse–Geisser estimate because the former can occasionally fail to control properly the Type 1 error rate. A final option is to use the known lower bound for ε in the adjustments rather than its estimated value – see [2]. Here, the adjustment factor is simply $1/(k-1)$. This is a conservative procedure, but if the F test is still significant after making this adjustment we can at least feel fairly safe in the validity of the result. Note that the use of any of these three adjustments does not need any change in

the analysis of variance – the observed F statistics remain unchanged – but simply a change in the degrees of freedom of the theoretical F variate in order to obtain a valid ***P* value**.

Comparison of Two or More Groups

Typically, we have the results of a randomized controlled trial in which n patients have each been randomly allocated to one of m treatments. Let us assume that each treatment group has exactly r patients allocated to it, and that there are no drop-outs. Each patient then provides regular follow-up measurements, yielding a total of k repeated measures per patient. Here, the analogy is with a traditional **split plot** design where the patients are equivalent to plots and the measures repeated over time are the subplots. This analogy implies the allocation of the degrees of freedom in an analysis of variance as in Table 1. The test of real interest in this context is that for the treatments by occasions interaction: are the time trends constant across the m groups? The F statistic for this interaction is compared to an F distribution with $(m-1)(k-1)$ and $m(r-1)(k-1)$ degrees of freedom. Despite the analogy, however, the data from this repeated measures design should *not* be treated as if it were from a split-plot experiment. The degrees of freedom should be adjusted using the appropriate Greenhouse–Geisser ε estimate in order to take into account the likely serial correlation in the data (although, as before, when $k = 2$ we do not have a problem). Note again that the adjustments are made to the theoretical F variate, with no adjustment whatsoever to the observed value of the test statistic. This is done automatically in many readily available software packages. Full details are provided in [7]. It is also important to remember that even these adjusted F tests are dependent on the additional assumption necessary for the analysis of data from this design: the covariance matrices are the same (apart from sampling fluctuations) across the m treatment groups.

Table 1

Between patients	$mr-1$
Treatments	$m-1$
Residual	$m(r-1)$
Within patients	$mr(k-1)$
Occasions	$k-1$
Treatments by occasions	$(m-1)(k-1)$
Residual	$m(r-1)(k-1)$

Advantages and Problems with the Use of ANOVA

The main advantage of the traditional ANOVA approach to the analysis seems to be both familiarity (amongst the data analysts and the readers of the resulting reports) and availability of easy-to-use software. It is difficult to think of any others! An alternative, which is quite popular in the social and behavioral sciences (including psychiatry), is to drop the assumptions concerning the correlational structure of the repeated measures and move on to use explicitly multivariate tests through the use of MANOVA. Both the adjusted ANOVA and the MANOVA approaches depend on the assumption of homogeneity of covariance matrices across groups, however. The multivariate approach involves the construction and testing of a set of transformed variables representing the within-subject differences for the within-subject factor (here, occasions) – typically a set of $k-1$ orthogonal polynomial **contrasts**. When $k = 2$ the univariate and multivariate methods are identical. One can carry out tests on the contrasts separately (is there a significant linear trend, for example, or is this trend the same in the m groups?) or as part of a multivariate test (using **Hotelling's T^2** statistic, for example).

Both the ANOVA and MANOVA approaches are dependent on complete data. If there are gaps or drop-outs, the patients with missing data have to be dropped from the analysis. This might be a source of considerable bias in the results. Another, although less common, problem is the collection of haphazardly spaced data instead of the measurements being made at fixed times for everyone in the study. One possible approach is to use random effects models with either **maximum likelihood** (ML) or **restricted maximum likelihood** (REML) estimation.

One obvious alternative to the use of ANOVA is to calculate a measure of change for each subject in the study – the summary measures approach to the analysis of repeated measures. In the context of the designs discussed above, this might be an estimate of linear trend. For the first design (k observations on each of n subjects) one can simply use a single-sample t test to assess whether this trend

is statistically significant. It is also very straightforward to produce confidence intervals for this trend. For the m-group design we can enter the summary statistics themselves into a simple one-way ANOVA to test for group differences. The advantage of the latter approach is that it is extremely simple to carry out and the results even easier to interpret than the output from split-plot design (with or without the almost obligatory adjustments). Although we are still using ANOVA, we are avoiding the pitfalls arising from serial correlation by replacing the k repeated measures by a simple, single, response feature of primary interest.

References

[1] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance. II. Effects of inequality of variance and of correlation between errors in the two-way classification, *Annals of Mathematical Statistics* **25**, 484–498.

[2] Greenhouse, S.W. & Geisser, S. (1959). On the methods in the analysis of profile data, *Psychometrika* **24**, 95–112.

[3] Hand, D. & Crowder, M. (1996). *Practical Longitudinal Data Analysis*. Chapman & Hall, London.

[4] Huynh, H. & Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements design have exact F-distributions, *Journal of the American Statistical Association* **65**, 1582–1589.

[5] Huynh, H. & Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom for sample data in randomized and split-plot designs, *Journal of Educational Statistics* **1**, 69–82.

[6] Mauchly, J.W. (1940). Significance test for sphericity of a normal n-variate distribution, *Annals of Mathematical Statistics* **29**, 204–209.

[7] Maxwell, S.E. & Delaney, H.D. (1990). *Designing Experiments and Analyzing Data*. Wadsworth, Belmont.

[8] Rouanet, H. & Lépine, D. (1970). Comparison between treatments in repeated-measures design: ANOVA and multivariate methods, *British Journal of Mathematical and Statistical Psychology* **23**, 147–163.

GRAHAM DUNN

Analysis of Variance

Analysis of variance (ANOVA) is one of the most commonly used statistical techniques, with applications across the full spectrum of biostatistics. The first reference to the technique appeared in the work of **R.A. Fisher** [5] in which he discussed the analysis of causes of human variability under a Mendelian scheme of inheritance (*see* **Mendel's Laws**). The first reference in Fisher's published work to the analysis of variance table was in 1923 [7], in a paper on the response of 12 different varieties of potato to the application of six manure treatments. The technique was fully discussed in Fisher's 1925 book [6].

Overview

One of the principal uses of statistical models is to attempt to explain variation in measurements. This variation may be due to a variety of factors, including variation from the measurement system, variation due to environmental conditions which change over the course of a study, variation from individual to individual (or experimental unit to experimental unit), etc. Factors which are not controlled from observation to observation can introduce variation in measured values. In designed experiments, the experimenter deliberately changes the levels of experimental factors to induce variation in the measured quantities, to lead to a better understanding of the relationship between experimental factors and the response (*see* **Experimental Design**). Other factors related to the response, called **blocking** factors, can be held fixed at one level to create a block of homogeneous experimental units which, in the absence of the effects of other factors, might be expected to produce measured responses with small variability. The experimental factors can then be manipulated on the units within the block. In a second block, the blocking factors can be held fixed at other levels and the experimental factors manipulated on the units within the block, etc. Effective blocking on factors related to the response can produce more precise estimates of the differences between the levels of experimental factors, while, at the same time, allowing more generalizable conclusions. To ensure that unnecessary variation in the measured responses is not introduced, other factors may be deliberately held fixed throughout the experiment (e.g. use of a standardized measurement system). Finally, **randomization** of the experimental factors to the experimental units serves to balance the effects of uncontrolled factors across the levels of the experimental factors to avoid the conscious or subconscious **confounding** of uncontrolled factors with those the experimenter is manipulating (*see* **Randomized Treatment Assignment**).

In **observational studies**, factors are typically not controlled – the data are obtained the way nature provides them. However, modeling and understanding the relationship between the observed values of the response and the observed values of explanatory variables collected with the response remains an important aim. The lack of control of extraneous factors either by blocking on levels of factors related to the response, or through randomization, makes the interpretation of models for observational data difficult, even if the basic analysis techniques are the same.

Analysis of variance is a commonly used technique for analyzing the relative contributions of identifiable sources of variation to the total variation in measured responses. Understanding the potential sources of variation prior to the development of a statistical model is very important. To develop an effective experimental design and/or to aid in the development and understanding of a statistical model, factors related to the response can be listed under categories such as measurement, environment, individual, method, etc. The cause and effect diagram, from the statistical process control literature (e.g. [8]), is an effective way to summarize potential sources of variation.

In a designed experiment, those variables which are to be experimentally manipulated, those which are used to define homogeneous sets of units (blocks), and those which are carefully controlled at fixed levels (e.g. measurement system variables) can be identified. Variables which have not been identified as experimental, blocking, or controlled factors may be contributing to variation in the response; however randomization will offer some insurance that their effects on the response are not systematically linked to the levels of the experimental factors.

For data from an observational study, the identification of such sources of variation can lead to the development of a statistical model in which variation in the response variable, associated with available explanatory variables, is "explained" through

inclusion of appropriate terms in the model. Again, variation from observational unit to observational unit in variables which are not included in the model can contribute to the "unexplained" (and possibly systematic) variation in the response. Finally, as described above, because the analyst cannot exert control over the process which gave rise to the data, unequivocal interpretations of the findings are very difficult, if not impossible.

The Model

In the description of the model, it will be assumed that there are n measurements on a response variable, Y, (i.e. $Y_1, Y_2, \ldots, Y_n$), and that associated with each measured response, there is a (row) vector of **explanatory variables** measured on, or related to the same unit as the response. Denote the elements of the vector associated with the ith response, Y_i, as $x_{i1}, x_{i2}, \ldots, x_{ip}$.

The elements of the vector of explanatory variables related to the response can be of several types. These include variables (assumed interval or ratio scaled; *see* **Measurement Scale**) such as age, height, or fitness score, and possibly higher powers of such variables (e.g. age^2), or other transformations (e.g. ln age). They can also include indicator or **dummy variables** which identify the levels of a nominal variable such as city or gender, or allow the distinction between the levels of an ordinal scale variable (e.g. satisfaction coded from "highly satisfied" to "completely dissatisfied"). Finally, an intercept term is usually required in the model. A variable, say x_{i1}, with $x_{i1} = 1$, for $i = 1, \ldots, n$ allows for the intercept term.

The **general linear model** links the response variable for the ith unit, Y_i, to the vector of variables related to the response. The form of this model is

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$
$$i = 1, \ldots n(p < n).$$

In matrix form, the model can be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is an $n \times p$ matrix which has the (row) vectors of variables as rows, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of residuals.

The role of the ε_is is to represent all sources of variation in the response which have not been accounted for by the variables included in the model. Thus, this vector represents the residual, or unexplained, variation in the measured response Y_i not accounted for by the model. This variation can be due to the effects of variables not included in the model, or to the misspecification of the form of their effects.

In the design of experiments literature, the ε_i term is often called the "experimental error" term to represent the combined effects of all those factors not controlled by the experimenter. For example, variation introduced by the measurement system, variation due to factors such as temperature or diet which the experimenter has not controlled, or variation in how the same treatment was applied to different subjects or units will contribute to this term. Randomization and **blinding** are designed to ensure, as far as possible, that the uncontrolled effects are not systematically related to terms in the model.

In observational studies, similar factors contribute to the residual term. However, unlike experiments in which randomization has been used to assign treatments to experimental units, it is often not reasonable to assume that such factors are unrelated to terms in the model.

In what follows we make the usual general linear model assumptions on the distribution of the ε_is. That is, it will be assumed that ε_i is a random variable with

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2; \quad i = 1, \ldots, n,$$

and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ independent.

These assumptions state that, on average, the residual terms are neither positive or negative so that the model does not systematically under- or overpredict; that the variation in these terms is the same for all experimental units and does not depend on the size of Y_i, and that the value of the ith residual term is not predictive of the value of any other residual term. These model assumptions can be checked using techniques discussed elsewhere (e.g. [2], and [4]) (*see* **Residuals**).

For tests of hypotheses (*see* **Hypothesis Testing**) about terms in the model, or for the calculation of **confidence intervals** for specific model parameters $(\beta_1, \beta_2, \ldots, \beta_p; \sigma^2)$, an additional assumption which describes the probability distribution of the ε_is is required. By far the most commonly used assumption

is that

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2).$$

Taken together, these assumptions imply that the residual (error) terms behave like an independent sample from a **normal distribution**, with mean 0 and common variance σ^2.

Estimation of Parameters

If the columns of the $\mathbf{X}$ matrix are linearly independent (i.e. $\mathbf{X}$ is of rank p), then the **least squares** estimate of the $p \times 1$ vector $\boldsymbol{\beta}$ which is equivalent to the maximum likelihood estimate of $\boldsymbol{\beta}$ under the normal distribution model described above is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The values of the response variable *predicted* by the model can then be obtained as

$$\hat{Y}_i = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip},$$

or, in matrix terms,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

The difference between the measured value of the response variable, Y_i, and the predicted value of the same variable, $\hat{Y}_i$, is called the ith estimated residual $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$. If the model describes the data well, then we would expect $Y_i - \hat{Y}_i$ to be "small". However, if the model is not providing a good description of the data, then $Y_i - \hat{Y}_i$ will be "large". It is these estimated residuals which form the basis for the model **diagnostics**, discussed elsewhere, used to assess the form and fit of the model. Furthermore, the quantity $\sum \hat{\varepsilon}_i^2/(n-p)$, which provides a measure of the departures from the model, is the usual estimate of σ^2, the variance of the residuals.

The Analysis of Variance Table

For the analysis of variance, we start with a measure of the variation in the measured response variables *before* any model is established for the Y_is. A convenient measure of the total variation in the measured response variables is given by the Total Sum of Squares

$$\text{Total SS} = \sum (Y_i - \overline{y})^2, \quad \text{where } \overline{y} = \sum \frac{Y_i}{n},$$

which is just $(n-1)$ times the estimated sample variance of the Y_is. If we have chosen an appropriate form for the model and have included the important predictors of the response, we expect $\hat{Y}_i$ to be close to Y_i. Thus, we will have "explained" a "large" portion of this variation. In fact, the variation left *unexplained* by the model can be summarized by the quantity

$$\text{Residual SS} = \sum (Y_i - \hat{Y}_i)^2,$$

which is just $n - p$ times the estimate of the variance of ε_i.

Thus, the variation in the response "explained" by the model will be the difference between the measure of the total variation in the response and the variation unexplained by the model. We have

total variation in the response
= variation explained by the model
+ variation unexplained by the model.

The "explained variation" is called the Model Sum of Squares (or Regression Sum of Squares). Algebraically, we can show that

$$\begin{aligned}\text{Model SS} &= \text{Total SS} - \text{Residual SS} \\ &= \sum (Y_i - \overline{y})^2 - \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (\hat{Y}_i - \overline{y})^2.\end{aligned}$$

In matrix terms, the split of the Total Sum of Squares into the Model Sum of Squares and Residual Sum of Squares can be written as

$$\mathbf{Y}'\mathbf{Y} - n\overline{y}^2 = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - n\overline{y}^2 + (\mathbf{Y} - \mathbf{X}'\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}'\hat{\boldsymbol{\beta}}).$$

It is this decomposition of the total sum of squares (as a measure of the total variation in the data prior to model fitting) into that portion "explained" by the model and that portion left "unexplained" that is the basis of the analysis of variance.

It is common practice to summarize the results of the analysis of variance in an *analysis of variance table*. The standard format of the table is given in Table 1. There are columns identifying the source of the variation, the degrees of freedom associated with each source, the corresponding sum of squares, and the mean square (sum of squares divided by degrees of freedom). Another column, headed "F", which gives the ratio of the Model Mean Square and the

Table 1 The analysis of variance table

Source	df	Sum of squares	Mean square	F
Model	$p-1$	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - n\overline{y}^2$	$(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - n\overline{y}^2)/(p-1)$	$\frac{\text{Mean Square}_{\text{Model}}}{\text{Mean Square}_{\text{Residual}}}$
Residual	$n-p$	$(\mathbf{Y}-\mathbf{X}'\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}'\hat{\boldsymbol{\beta}})$	$(\mathbf{Y}-\mathbf{X}'\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}'\hat{\boldsymbol{\beta}})/(n-p)$	
Total	$n-1$	$\mathbf{Y}'\mathbf{Y} - n\overline{y}^2$		

Residual Mean Square is usually added. This column is used for tests of hypotheses associated with sources of variation which will be discussed later.

As discussed, the Residual Mean Square in the analysis of variance table provides an estimate of σ^2, the (assumed constant) variance of the distribution of the ε_is.

Tests of Hypotheses – The Extra Sum of Squares Principle *F* Test

In most applications of the analysis of variance, interest will focus on the testing of hypotheses concerning the regression parameters (i.e. the β_is in the above model). In particular, tests of hypotheses of the form

$$\text{H: } \beta_k = 0,$$

or, more generally,

$$\text{H: } \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0,$$

are the most common. Each of the above is equivalent to asking whether the variable or variables associated with the β parameter(s) in the statement of the hypothesis are related to the response, *after adjusting for the effects of other variables included in the model*.

Intuitively, if we fit the full model and compute the Residual Sum of Squares, and then refit a reduced model without the variables corresponding to the β_is specified in the hypothesis, we should be able to judge the appropriateness of the hypothesis by examining the increase in the Residual Sum of Squares. If the increase is "large", then the variables we removed were important in explaining the variation in the response, after considering the effect of the variables remaining in the model. The assessment of whether this increase in the Residual Sum of Squares is "large" can be conducted by a statistical test based on the size of the increase in the Residual Sum of Squares, standardized by the Residual Sum of Squares from the full model.

Consider the hypothesis

$$\text{H: } \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0,$$

and define

$$\Delta = \text{Residual SS}_{\text{Reduced Model}} - \text{Residual SS}_{\text{Full Model}}.$$

Then the test statistic

$$F = \frac{\Delta/(p-q)}{\text{Residual SS}_{\text{Full Model}}/(n-p)}$$

assesses the change in the Residual Sum of Squares per variable removed from the model, relative to the residual mean square from the full model. That is, large values of the test statistic occur whenever the increase in the Residual Sum of Squares is "large" with respect to what might be expected by chance alone.

If the full model is assumed to be an adequate description of the variation in the response, and if the residuals are assumed to follow the normal distribution as described above, then under the hypothesis, H:$\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$, the quantity F behaves like a random variable from the $F_{(p-q),(n-p)}$ distribution (i.e. ***F* distribution** with $p-q$ and $n-p$ **degrees of freedom**). Consequently, a test of significance can be performed in which the observed value of F, F_{observed}, is compared to the tables of the $F_{(p-q),(n-p)}$ distribution. The significance level of the data with respect to the hypothesis is, then, $\Pr(F_{(p-q),(n-p)} > F_{\text{observed}})$ (*see* **Level of a Test**). *Small* values of the significance level are associated with *large* values of the observed test statistic, and cast doubt on the hypothesis.

One important application of this principle involves assessing whether the independent variables, collectively, explain a significant portion of the

variability in the response. In a model with an intercept term (i.e. $x_{i1} = 1, i = 1, \ldots, n$) this involves testing

$$\text{H: } \beta_2 = \beta_3 = \cdots = \beta_p = 0.$$

The reduced model under the hypothesis is

$$Y_i = \beta_1 + \varepsilon_i, \quad i = 1, \ldots, n.$$

In this reduced model, $\hat{\beta}_1 = \overline{y}$, and the Residual Sum of Squares is

$$\text{Residual SS}_{\text{Model With Intercept Only}} = \sum (Y_i - \overline{y})^2,$$

which is just the Total Sum of Squares for the Y_is.

Then if

$$\begin{aligned} \Delta &= \text{Residual SS}_{\text{Model With Intercept Only}} \\ &\quad - \text{Residual SS}_{\text{Full Model}} \\ &= \text{Total SS} - \text{Residual SS}_{\text{Full Model}} \\ &= \text{Model SS}, \end{aligned}$$

then the extra sum of squares principle F test statistic is

$$\begin{aligned} F &= \frac{\Delta/(p-1)}{\text{Residual SS}_{\text{Full Model}}/(n-p)} \\ &= \frac{\text{Model Mean Square}}{\text{Residual Mean Square}}. \end{aligned}$$

The column headed "F" in the analysis of variance table contains this ratio. When compared with the $F_{(p-1),(n-p)}$ tables, this statistic provides a test of whether the combined effect of all independent variables, and hence the fitted model, explains a significant portion of the variability in the response.

Partitioning the Model Sum of Squares

In many applications, interest will focus on examining various submodels of the fitted model. For example, in a **polynomial regression** model

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i,$$

an obvious question is whether the model involving only the linear term would describe the data almost as well as the more complicated quadratic model. Since the linear model is a special case of the quadratic model (i.e. $\beta_3 = 0$), the residual sum of squares for the best fitting linear model cannot be smaller than that for the best fitting quadratic model. The difference,

$$\begin{aligned} &\text{Residual SS}_{(\text{linear})} - \text{Residual SS}_{(\text{linear, quadratic})} \\ &\quad = \text{Model SS}_{(\text{linear, quadratic})} - \text{Model SS}_{(\text{linear})} \\ &\quad = \text{Model SS}_{(\text{quadratic}|\text{linear})}, \end{aligned}$$

if large, indicates that the quadratic model is "explaining" much more of the variation in the measured response variable than is the linear model. The extra sum of squares principle F test provides one means of testing whether the variation explained by the quadratic term is significant, over and above that which is explained by the linear term.

Similarly, if the measured response is systolic blood pressure, $x_{i1} = 1$ codes for the intercept, x_{i2} represents *weight* and x_{i3} represents *age* for the ith subject, then the difference

$$\begin{aligned} &\text{Residual SS}_{(\text{Model with weight})} \\ &\quad - \text{Residual SS}_{(\text{Model with weight and age})}, \end{aligned}$$

represents the additional portion of the variation "explained" when age is added to a model involving weight, over that explained by a model involving weight alone. If this difference is large, it suggests that age is an important variable in the explanation of the variability in systolic blood pressure even when the effect of weight has been modeled in this fashion.

If we define Model $\text{SS}_{(\text{weight, age})}$ to be the Model Sum of Squares for a model containing both weight and age, and Model $\text{SS}_{(\text{weight})}$ as the model sum of squares for a model containing only weight, then we can define

$$\begin{aligned} \text{Model SS}_{(\text{age}|\text{weight})} &= \text{Model SS}_{(\text{weight, age})} \\ &\quad - \text{Model SS}_{(\text{weight})}, \end{aligned}$$

as the "additional" sum of squares explained by age when added to a model containing weight.

Then, we can modify the Model Sum of Squares entries in the analysis of variance table to allow us to look at this partitioning of the Model Sum of Squares. Since,

$$\begin{aligned} \text{Model SS}_{(\text{weight, age})} &= \text{Model SS}_{(\text{weight})} \\ &\quad + \text{Model SS}_{(\text{age}|\text{weight})}, \end{aligned}$$

Table 2 Analysis of variance table for the blood pressure example

Source	df	Sum of squares	Mean square	F
Model (weight, age)	2	Model $SS_{(\text{weight, age})}$	Model $SS_{(\text{weight, age})}/2$	$\frac{\text{Model } MS_{(\text{weight, age})}}{\text{Residual MS}}$
Model (weight)	1	Model $SS_{(\text{weight})}$	Model $SS_{(\text{weight})}/1$	$\frac{\text{Model } MS_{(\text{weight})}}{\text{Residual MS}}$
Model (age|weight)	1	Model $SS_{(\text{age}\mid\text{weight})}$	Model $SS_{(\text{age}\mid\text{weight})}/1$	$\frac{\text{Model } MS_{(\text{age}\mid\text{weight})}}{\text{Residual MS}}$
Residual	$n-2$	Residual $SS_{(\text{weight, age})}$	Residual $SS_{(\text{weight, age})}/(n-2)$	
Total	$n-1$	$\mathbf{Y}'\mathbf{Y} - n\overline{y}^2$		

we have the analysis of variance table shown in the Table 2.

Then F ratios, reading down the column, test respectively, the hypotheses:

1. The model involving weight and age does not explain a significant portion of the variation in blood pressure.
2. The model involving weight alone does not explain a significant portion of the variation in blood pressure.
3. After adjusting for weight, age does not explain a significant portion of the variation in blood pressure. That is, for individuals of the same weight, there is no significant linear relationship between blood pressure and age.

It is important to note that we could also have decomposed the Model Sum of Squares as:

$$\text{Model } SS_{(\text{weight, age})} = \text{Model } SS_{(\text{age})} + \text{Model } SS_{(\text{weight}\mid\text{age})}.$$

The interpretation of the F tests is analogous to that described above, with the role of weight and age reversed. Because of the observational nature of many data sets, the explanatory variables may be highly correlated. In such cases, it will often be very difficult to obtain an unequivocal interpretation of the results of these F tests. For example, the model involving both age and weight may be highly significant, but both Model $SS_{(\text{age}\mid\text{weight})}$ and Model $SS_{(\text{weight}\mid\text{age})}$ may be small. Conversely, neither age nor weight on its own may explain a significant portion of the variation in blood pressure, but the model involving both may be highly significant. A class of techniques known as **variable selection** methods (e.g. [3, 4]) is often used to attempt to untangle the relationships between the response variable and correlated predictor variables.

In some software packages (e.g. SAS [12]), the analysis of variance table output will contain reference to type I and type III sums of squares. (Types II and IV are also available but will not be discussed here.) In brief, with type I sums of squares, the Model Sum of Squares is decomposed in the order that the terms are specified in the model. Thus, if the model is specified as

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$
$$(\text{i.e. } Y = \text{Weight} + \text{Age}),$$

we get

$$\text{Model } SS_{(\text{weight, age})} = \text{Model } SS_{(\text{weight})} + \text{Model } SS_{(\text{age}\mid\text{weight})},$$

and the two sums of squares terms on the right of the equation are provided as the type I sums of squares.

A type III sum of squares is the contribution to the Model Sum of Squares due to a term in the model after adjusting for *all other* terms specified in the model. Thus, in the example above, the type III sums of squares would be Model $SS_{(\text{age}\mid\text{weight})}$ and Model $SS_{(\text{weight}\mid\text{age})}$. Note that type I sums of squares will total the Model Sum of Squares, whereas the type III sums of squares will not unless the columns of the $\mathbf{X}$ matrix are orthogonal, as described in the next section.

Orthogonality

A very special case of partitioning the Model Sum of Squares occurs when the columns of the $\mathbf{X}$ matrix

are orthogonal (perpendicular) (*see* **Orthogonality**). In general, if two $n \times 1$ (column) vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal, then

$$\mathbf{x}_1'\mathbf{x}_2 = 0.$$

If the columns of the matrix $\mathbf{X}$ are mutually orthogonal (i.e. each pair of columns is orthogonal), then $\mathbf{X}'\mathbf{X}$ will be a diagonal matrix. In particular, the entry in position (j, j) is $\mathbf{x}_j'\mathbf{x}_j$, where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. The least squares estimate of $\boldsymbol{\beta}_j$ (equivalent to the maximum likelihood estimate under the assumption of a common normal distribution for the $\varepsilon_i s$), will be

$$\hat{\beta}_j = \sum_{i=1}^{n} x_{ij} Y_i \Bigg/ \sum_{i=1}^{n} x_{ij}^2,$$

and this estimate will not change as orthogonal columns are added or deleted from the $\mathbf{X}$ matrix.

Further, when the columns of $\mathbf{X}$ are orthogonal, the Model Sum of Squares decomposes into $p - 1$ components, with the jth component given by

$$\left(\sum_{i=1}^{n} x_{ij} Y_i\right)^2 \Bigg/ \sum_{i=1}^{n} x_{ij}^2.$$

(Note that the term $n\overline{y}^2$ which has been used to correct the Total and Model Sums of Squares for the intercept term, is just

$$n\overline{y}^2 = \left(\sum_{i=1}^{n} 1 \times Y_i\right)^2 \Bigg/ \sum_{i=1}^{n} 1^2,$$

consistent with the above.)

Thus, the Model Sum of Squares can be written

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - n\overline{y}^2 \\
&\quad = \sum_{j=1}^{p} \left(\left(\sum_{i=1}^{n} x_{ij} Y_i\right)^2 \Bigg/ \sum_{i=1}^{n} x_{ij}^2 \right) - n\overline{y}^2 \\
&\quad = \sum_{j=2}^{p} \left(\left(\sum_{i=1}^{n} x_{ij} Y_i\right)^2 \Bigg/ \sum_{i=1}^{n} x_{ij}^2 \right),
\end{aligned}$$

and we have a complete decomposition of the Model Sum of Squares into $p - 1$ one-degree-of-freedom sums of squares.

The ratio of any of these one-degree-of-freedom sums of squares to the Residual Mean Square provides an F test of the hypothesis that the relevant β_i term is zero. Because of the complete orthogonality, the sum of squares associated with any column in the $\mathbf{X}$ matrix will not change as columns orthogonal to it are added to or deleted from the model. Note that in such cases the type I and type III Sums of Squares (e.g. SAS [12]) will be equal.

This result is very important to the design of experiments. As discussed, in the designed experiment, the investigator has freedom to choose the levels of factors to be manipulated, the combinations of factors to be used in the experiment, and the number and construction of blocks which will be used to produce homogeneous units on which to experiment. With the proper choice of levels of experimental factors, and the proper attention to balance of the experimental factors across blocks, it is possible to design orthogonality into the experiment. This makes the analysis of the results relatively straightforward. More importantly, the interpretation of the results is clear since the orthogonality of the design ensures protection against the confounding of the experimental factors. For example, by randomly assigning predetermined levels of exercise to the same number of younger and older subjects, and measuring blood pressure at a later time, it would be possible to estimate the effects of exercise on blood pressure independently of age (and vice versa).

With observational studies, no such balance is guaranteed. Consequently, it is common to find explanatory variables which are highly correlated. The interpretation of the results is difficult. If younger subjects have lower blood pressure and exercise more than do older subjects, it may be difficult to separate the roles of exercise and age on blood pressure.

Analysis of Variance in Designed Experiments

In this section, we briefly describe the analysis of variance in designed experiments. The design of experiments is a very broad topic, and specific designs are described elsewhere (e.g. [1] and [11]) (*see* **Randomized Complete Block Designs**; **Latin Square Designs**; **Factorial Experiments**). However, we will discuss the use of the analysis of variance technique with reference to some simple designs to illustrate the basic features of the technique.

Consider a randomized block experiment in which t treatments are to be studied in an experiment involving b blocks. That is, the experimenter has isolated b sets of t units each on which to experiment. The units within a block are similar with respect to factors thought to be related to variability in the response. Units in different blocks may well be quite different with respect to factors thought to be related to the response. The t treatments are randomly assigned to the units within a block, and one measurement is taken from each unit. Thus, we have $n = t \times b$ measurements on the response variable to analyze.

Note that by assigning different treatments to relatively homogeneous units, the experimenter has deliberately introduced variation into the experiment. However, the sources of this variation are known and, thus, can be dealt with at the time of analysis.

In what follows, we will assume a **fixed effects** model. That is, the inference about the effects of the treatments, and about the effects of the blocks is relevant only for the treatments and blocks used in the particular experiment. Later, we will briefly describe **random effects** models in which the treatments (and/or blocks) are assumed to be a random sample from a population of treatments (blocks). In such cases, the inference will apply to the population of treatments (blocks) from which the actual treatments (blocks) were a sample.

To better describe the model, we change notation slightly. Define Y_{ij} to be the measured response of treatment i in block j. Then we describe the measured value as composed of an effect common to all observations, the effect due to the treatment applied, and the effect due to the block which contains the unit as follows:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, \ldots, t; \quad j = 1, \ldots, b.$$

In this model, μ represents an effect common to all observations, α_i represents the effect of the ith treatment relative to μ, β_j represents the effect of the jth block relative to μ, and ε_{ij} represents the uncontrolled variation (experimental error) from the unit in the jth block receiving the ith treatment.

While this model may seem different from the model defined earlier, there is a correspondence between them which is useful to keep in mind. If we write the measured values in a $tb \times 1$ column vector such that the observation on treatment i in block j is in location $i^* = (j-1)t + i$, then we have "strung out" the observed values so that the first observation is on treatment 1 in block 1, the second is on treatment 2 in block 1, ..., the tth observation is on treatment t in block 1. Then observation $t+1$ is the observation on treatment 1 in block 2, etc. We can then define a matrix $\mathbf{X}$ which will contain indicator variables to code for the different treatments and the different blocks in the experiment.

For example, define a $tb \times 1$ column vector $\mathbf{x}_1$ such that $x_{i1} = 1$, for $i = 1, \ldots, tb$, to provide for an intercept term (μ).

Define t $tb \times 1$ column vectors $\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{t+1}$ such that the kth entry of $\mathbf{x}_{1+i}$ is 1 if the kth observation is on treatment i; otherwise the kth entry of $\mathbf{x}_{1+i}$ is 0.

Likewise, define b $tb \times 1$ column vectors $\mathbf{x}_{t+2}, \ldots, \mathbf{x}_{t+b+1}$ such that the kth entry of x_{t+1+j} is 1, if the kth observation is from block j; otherwise the kth entry of $\mathbf{x}_{t+1+j}$ is 0.

Then we form the matrix $\mathbf{X}$ with columns given by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t+b+1}$. For example, if $t = 2$ and $b = 3$, then we get the $\mathbf{X}$ matrix shown in Table 3.

With this definition for $\mathbf{X}$, we could write the model in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \varepsilon,$$

where $\mathbf{Y}$ is the $tb \times 1$ vector of observations "strung out" as described above, $\mathbf{X}$ is given in Table 3, and $\boldsymbol{\gamma}$ is a $t + b + 1$ vector of parameters with

$$\boldsymbol{\gamma}' = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3).$$

However, in this model, $\mathbf{X}$ is not of full rank; the columns of $\mathbf{X}$ are not linearly independent. For example, column 1 is the sum of columns 2 and 3, and also the sum of columns 4, 5, and 6. Thus, we cannot obtain unique estimates of the parameters.

Table 3 $\mathbf{X}$ matrix for the randomized block design for $t = 2$, $b = 3$ with no constraints

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

However, if we place some restrictions or constraints on the parameters, we can arrive at estimates which are useful for describing the variation in the measured response. There are two commonly used sets of constraints for models of this type; these will be briefly described. It is important to note that these different systems of constraints will lead to different estimates of the parameters in the model, and the interpretation of the estimates will depend on the type of constraint applied. However, tests of hypotheses about the overall effect of treatments or blocks will be the same under both sets of constraints.

Indicator Variable Constraints

The simplest way to resolve the issue of linear dependence in the columns of **X** is to remove columns to produce a set which is linearly independent but which still allow for the estimation of meaningful quantities. In the example in Table 3, dropping the second and fourth columns of **X** gives a matrix which has linearly independent columns. Dropping these columns is equivalent to the set of constraints

$$\alpha_1 = 0; \quad \beta_1 = 0.$$

Under this model, the parameter μ is an estimate of the mean response from the unit in block 1 which received treatment 1. The parameter α_2 measures the difference in mean response between units *in the same block* which received treatment 2 and those which received treatment 1. Similarly, β_2 represents the difference in mean response between units in block 2 and units in block 1 *which received the same treatment*. A test of the hypothesis that $\alpha_2 = 0$ is then a test of no difference between treatments, adjusting for the differences in blocks, as required.

These constraints are used in many computing packages (e.g. GLIM [9]). Their major disadvantage is that, since the columns of **X** are not orthogonal, when the parameters related to treatments (blocks) are dropped from the model, the estimated values of the other parameters in the model will change, requiring a refit of the model. However, this disadvantage is only minor with modern computing packages.

Of course the choice of which columns to drop is arbitrary. In general, setting up the model so that natural reference categories are formed has some advantage for interpreting the results of the analysis.

Analysis of Variance Constraints

An alternative set of constraints which allows for hand calculation of all quantities required is

$$\sum_{i=1}^{t} \alpha_i = 0; \quad \sum_{j=1}^{b} \beta_j = 0.$$

Under these constraints, the parameter μ represents the overall mean response across all units used in the experiment. The parameter α_i represents the difference in mean response for units treated with treatment i, relative to the overall mean response. Similarly, the parameter β_j represents the difference in mean response for units in block j, relative to the overall mean response.

If we write the model in terms of α_2, β_2, and β_3 using these constraints (i.e. $\alpha_1 = -\alpha_2; \beta_1 = -\beta_2 - \beta_3$), we have the new **X** matrix, $\mathbf{X}^*$, given in Table 4, and the model can be written as

$$\mathbf{Y} = \mathbf{X}^* \boldsymbol{\gamma}^*,$$

where $\boldsymbol{\gamma}^{*\prime} = (\mu, \alpha_2, \beta_2, \beta_3)$.

In this model the column coding for the mean effect is orthogonal to those coding for treatment effects and to those coding for block effects. Similarly, the columns coding for treatment effects are orthogonal to those coding for the block effects. This orthogonality is a natural result of the balance in the design (each treatment appears the same number of times in each block), and the parameterization of the model. It follows that if this model is fitted and then the columns coding for the treatment effects are deleted and the model refitted, the estimates of the parameters for the mean and for the blocks will be unchanged. It further follows that we can get an easy decomposition of the Model Sum of Squares into a source due to treatments and a source due to blocks.

Table 4 $\mathbf{X}^*$ matrix for the randomized block design for $t = 2$, $b = 3$ with analysis of variance constraints

$$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

These sources will have $t-1$ and $b-1$ degrees of freedom respectively due to the single constraint on the $t(b)$ treatment (block) parameters.

Further it is quite easy to show that the least squares estimates (maximum likelihood estimates under the common normal errors model) of the parameters in the model are

$$\hat{\mu} = \overline{y}; \quad \hat{\alpha}_i = \overline{y}_{i.} - \overline{y}. \quad \text{and} \quad \hat{\beta}_j = \overline{y}_{.j} - \overline{y}_{..};$$
$$i = 1, \ldots, t; \quad j = 1, \ldots, b,$$

where $\overline{y}_{i.} = \sum_{j=1}^{b} Y_{ij}/b$, and $\overline{y}_{.j} = \sum_{i=1}^{t} Y_{ij}/t$ are the average responses for all units which received treatment i, and the average response for all units in block j, respectively. The Residual Sum of Squares for this model is then

$$\begin{aligned}
&\text{Residual SS}_{\text{Full Model}} \\
&= \sum_{i=1}^{t}\sum_{j=1}^{b} \hat{\varepsilon}_{ij}^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{b} (Y_{ij} - \hat{\alpha}_i - \hat{\beta}_j)^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{b} (Y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{b} (Y_{ij} - \overline{y}_{..})^2 - b\sum_{i=1}^{t} (\overline{y}_{i.} - \overline{y}_{..})^2 \\
&\quad - t\sum_{j=1}^{b} (\overline{y}_{.j} - \overline{y}_{..})^2.
\end{aligned}$$

The quantities $b\sum_{i=1}^{t}(\overline{y}_{i.} - \overline{y}_{..})^2$ and $t\sum_{j=1}^{b}(\overline{y}_{.j} - \overline{y}_{..})^2$ are called the Treatments and Blocks Sum of Squares, respectively. For example, if there is considerable variability of the $\overline{y}_{i.}$s about their mean $(\overline{y}_{..})$, there is evidence that not all the treatments are producing the same average response and the Treatments Sum of Squares will be large. Since each treatment appears in each block, differences between blocks cannot be accounting for the differences between the treatment averages. Similar conclusions will obtain in examining the Blocks Sum of Squares.

The Analysis of Variance Table for the Randomized Block Design

To develop the analysis of variance table for the randomized block design, we adopt the analysis of variance constraints parameterization, and note that the hypotheses of interest are usually that there is no difference between the treatments, or no difference between the blocks; i.e.

$$\text{H: } \alpha_1 = \alpha_2 = \cdots = \alpha_t = 0 \quad \text{or}$$
$$\text{H: } \beta_1 = \beta_2 = \cdots = \beta_b = 0.$$

Under the first hypothesis above, the model becomes

$$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}; \quad i = 1, \ldots, t; \quad j = 1, \ldots, b,$$

and we again have, due to the orthogonality of the columns representing the block parameters and the column representing the mean parameter, that

$$\hat{\mu} = \overline{y}_{..}; \quad \hat{\beta}_j = \overline{y}_{.j} - \overline{y}_{..}, \quad j = 1, \ldots, b;$$

and

$$\begin{aligned}
\text{Residual SS}_{\text{H:}\alpha_i=0} &= \sum_{i=1}^{t}\sum_{j=1}^{b} (Y_{ij} - \overline{y}_{..})^2 \\
&\quad - t\sum_{j=1}^{b} (\overline{y}_{.j} - \overline{y}_{..})^2.
\end{aligned}$$

Then, to test H: $\alpha_1 = \alpha_2 = \cdots = \alpha_t = 0$, we compute the Extra Sum of Squares Principle F test statistic.

Let

$$\begin{aligned}
\Delta &= \text{Residual SS}_{\text{H:}\alpha_i=0} - \text{Residual SS}_{\text{Full Model}} \\
&= \text{Treatments SS}.
\end{aligned}$$

Then

$$\begin{aligned}
F &= \frac{\Delta/(t-1)}{\text{Residual SS}_{\text{Full Model}}/(t-1)(b-1)} \\
&= \frac{\text{Treatments MS}}{\text{Residual MS}_{\text{Full Model}}}.
\end{aligned}$$

We compare the observed value of the test statistic to the tables of the $F_{(t-1),(t-1)(b-1)}$ distribution.

The above can be summarized in the analysis of variance Table given in Table 5.

Table 5 Analysis of variance table for a randomized block design

Source	df	Sum of squares	Mean square	F
Model	$t+b-2$			
Treatments	$t-1$	$b\sum_{i=1}^{t}(\overline{y}_{i.}-\overline{y}_{..})^2$	Treatments SS/$(t-1)$	$\dfrac{\text{Treatments MS}}{\text{Residual MS}}$
Blocks	$b-1$	$t\sum_{j=1}^{b}(\overline{y}_{.j}-\overline{y}_{..})^2$	Blocks SS/$(b-1)$	$\dfrac{\text{Blocks SS}}{\text{Residual MS}}$
Residual	$(t-1)(b-1)$	Residual $SS_{\text{Full Model}}$	Residual $SS_{\text{Full Model}}/(t-1)(b-1)$	
Total	$n-1$	$\mathbf{Y}'\mathbf{Y}-n\overline{y}_{..}^2$		

Orthogonal Contrasts for Planned Comparisons

In many experiments in which t treatments are to be compared, there may be some a priori (i.e. planned) comparisons which the investigator has in mind. For example, suppose that $t = 3$, and that the three treatments represent three doses (0, 10, and 20 units) of a drug. Rather than whether there are differences between the doses, the question of interest might be whether the response to the drug changes linearly with dose.

Factorial designs (*see* **Factorial Experiments**; **Factorial Designs in Clinical Trials**) provide another important example. Suppose $t = 8$ and the eight treatments represent all combinations of three drugs (Q, R, and S), each of which has two levels (0 and 20 units). Questions involving comparing the high and low doses of each drug, and determining whether the effect of one drug depends on the level of another drug (i.e. whether there is an interaction between the drugs), are likely to be more relevant than whether there is a difference between the eight treatments.

Consider the t means, $\overline{y}_{1.}, \overline{y}_{2.}, \ldots, \overline{y}_{t.}$, as a vector of t responses, and define a $t \times (t-1)$ matrix $\mathbf{C}$ such that the columns of $\mathbf{C}$ are mutually orthogonal and such that the sum of the t entries in each column is zero. The columns of $\mathbf{C}$ thus defined are called **contrasts**. While there may be many ways to define a set of orthogonal contrasts, generally there will be a main set which provides the answers to the relevant analysis questions.

For example, for the question of linearity described in the first experiment above, we order the means so that they correspond to the averages on units receiving 0, 10, and 20 units, respectively. If we define the contrast, $\mathbf{c}_1' = (-1, 0, 1)$, then $\sum_{i=1}^{t} c_{i1}\overline{y}_{i.}$ will be close to zero whenever the first mean and third mean are of the same size. If this were the case, there would be no suggestion that the means were increasing linearly. The contrast $\mathbf{c}_2' = (-1, 2, -1)$ is orthogonal to $\mathbf{c}_1$, and if $\sum_{i=1}^{t} c_{i2}\overline{y}_{i.}$ is close to zero, it suggests that there is no parabolic (quadratic) relationship between the means and dose.

For the 2^3 factorial design described above, let $\overline{y}_{qrs}$ represent the mean of all observations when the qth level of drug Q, the rth level of drug R, and the sth level of drug S are administered ($q, r, s = 1, 2$). Consider the matrix $\mathbf{C}$ shown in Table 6. If the treatment means are arranged in the vector $(\overline{y}_{111}, \overline{y}_{211}, \overline{y}_{121}, \overline{y}_{221}, \overline{y}_{112}, \overline{y}_{212}, \overline{y}_{122}, \overline{y}_{222})'$, then the first column of $\mathbf{C}$ is a contrast which, when applied to the vector of means, assesses whether the observations on the high and low levels of Q are equal. Similarly, columns 2 and 4 assess the *main effects* of R and S; i.e. whether observations on the high and low levels of R and S, respectively, are equal. Column 3 examines whether the difference between high and low levels of drug Q are different for the levels of drug R. Similarly, columns 5 and 6 assess the *interactions* of drugs Q and S, and R and S, respectively. The final column assesses the three-way interaction, i.e. whether the interaction of

Table 6 Matrix of contrasts, $\mathbf{C}$, for the 2^3 factorial design

$$\begin{pmatrix} -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Q and R depends on the levels of S. In each case, if $\sum_{i=1}^{t} c_{ik}\overline{y}_{i.}$ is close to zero, we can conclude that the particular effect is not important.

Because of the orthogonality, the Treatment Sum of Squares can be decomposed into $t-1$ one-degree-of-freedom sum of squares, each of which can be used to test whether the corresponding sum is significantly different from zero. We have

$$\sum_{i=1}^{t}(\overline{y}_{i.}-\overline{y}_{..})^2=\sum_{k=1}^{t-1}\left(\left(\sum_{i=1}^{t}c_{ik}\overline{y}_{i.}\right)^2\Big/\sum_{i=1}^{t}c_{ik}^2\right),$$

where c_{ik} is the ith entry in the kth column of $\mathbf{C}$. This result has exactly the same form and follows directly from the result given earlier on the decomposition of the Model Sum of Squares when the columns of $\mathbf{X}$ are orthogonal. Here, the "Model Sum of Squares" from the regression of the vector of means on the columns of $\mathbf{C}$ will be equivalent to the "Total Sum of Squares" of the means since we have t means and have defined $t-1$ columns in $\mathbf{C}$. Again, there is a one-degree-of-freedom sum of squares for the "intercept" term. Indeed, the requirement that the columns of $\mathbf{C}$ sum to zero is to ensure orthogonality with the overall mean.

Thus, the Treatment Sum of Squares (which for the randomized block design is just b times the expression on the left above) can be decomposed into $t-1$ one-degree-of-freedom sums of squares. (For the randomized block design, the kth such sum of squares would be $b(\sum_{i=1}^{t} c_{ik}\overline{y}_{i.})^2/\sum_{i=1}^{t} c_{ik}^2$.) Each of these sums of squares can be used to test certain comparisons among the treatment means, by forming the F statistic with the sum of squares as numerator and residual mean square as denominator.

Note that the development given here assumes that there are the same number of observations on each treatment, as would be the case in most designs, including the randomized block design described above. If the design is unbalanced, this breakup of the Treatment Sum of Squares may still be possible, but the $\mathbf{C}$ matrix needs to be constructed based on the original vector of responses, taking the different numbers of observations into account.

Designs Involving More Than One Source of Error

In the discussion above, we have considered models in which, because of the design, it was only possible to estimate a single measure of experimental error. As discussed initially, there are many potential sources of variation in any statistical investigation. Often through the design used, we may be able to separate sources of variation (*see* **Variance Components**) and arrive at different error terms for different comparisons. We illustrate this briefly with a simple example. Suppose we are interested in the effect of a drug on heart rate during exercise. We randomly assign m individual subjects to one of three levels of the drug (0, 10, or 20 units). The experimental procedure involves measuring heart rate, for each subject, under two exercise conditions (moderate and extreme), using a standard protocol. Thus, we have $6m$ measurements in total.

If we consider the factors associated with variation in heart rate, some of these are factors which vary from individual to individual (e.g. age, fitness level, weight, etc.). Others are factors which would vary between test times on the same individual (e.g. temperature, fatigue, etc.) In fact, we could measure the *within-subject* variation by taking repeat measurements on the same subject under similar conditions, and measure the *between-subject* variation by taking measurements on different subjects under similar conditions. It is reasonable to think that under similar conditions there will be more variation from individual to individual than from test time to test time within a single subject (*see* **Longitudinal Data Analysis, Overview**; **Split Plot Designs**).

The design described above allows us to estimate the between-subject and within-subject contributions to the overall experimental error. The between-subject contributions provide the benchmark against which to test treatments applied at the level of the subject (i.e. the levels of the drug). The within subjects contributions provide the baseline variability against which to compare treatments applied within subjects (i.e. the differing exercise regimes, and their interaction with the dose of the drug).

A model for this experiment includes two sources of error, which, for this design, are separable. A model is

$$Y_{ijk}=\mu+\alpha_i+\varepsilon_{ij}+\beta_k+(\alpha\beta)_{ik}+\delta_{ijk},$$

where Y_{ijk} represents the observation on exercise level k for the jth subject in the ith drug group; μ is the overall mean; α_i represents the effect of drug level $i\,(i=1,2,3)$; β_k represents the effect of exercise level $k(k=1,2)$; and $(\alpha\beta)_{ik}$ represents the effect of the interaction of drug level i and exercise

level k. Furthermore, ε_{ij} gives the between-subjects experimental error term for subject j in the ith drug group, and δ_{ijk} represents the within-subjects experimental error term for the measurement on exercise regimen k, for the jth subject in group i.

For the purposes of estimation, we make the usual "fixed" effects assumptions, i.e.

$$\sum_{i=1}^{3}\alpha_i = 0; \quad \sum_{k=1}^{2}\beta_k = 0;$$

$$\sum_{i=1}^{3}(\alpha\beta)_{ik} = \sum_{k=1}^{2}(\alpha\beta)_{ik} = 0.$$

Furthermore, we assume

$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_{\mathrm{I}}^2), \quad \text{independently;}$$

and

$$\delta_{ijk} \sim \mathrm{N}(0, \sigma_{\mathrm{II}}^2),$$

independently, and independent of ε_{ij}.

Under this model, we obtain the following estimates:

$$\hat{\mu} = \overline{y}_{...}; \quad \hat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{...}; \quad \hat{\beta} = \overline{y}_{.j.} - \overline{y}_{...};$$

and

$$(\alpha\hat{\beta})_{ik} = \overline{y}_{i.k} - \overline{y}_{i..} - \overline{y}_{..k} + \overline{y}_{...},$$

where the notation again implies averages are taken over the subscript denoted by a subscript dot. The sums of squares associated with each of the fixed effects are obtained as before. That is, the Drug Groups Sum of Squares is just

$$\text{Drugs SS} = \sum_{k=1}^{2}\sum_{j=1}^{m}\sum_{i=1}^{3}(\overline{y}_{i..} - \overline{y}_{...})^2$$

$$= 2m\sum_{i=1}^{3}(\overline{y}_{i..} - \overline{y}_{...})^2.$$

There are two sources of experimental error to estimate here. The between-subjects residual sum of squares is obtained by combining both measurements on the exercise regimen, and measuring the variation between subjects within drug groups. Thus we obtain,

$$\text{Residual SS}_{\text{Between Subjects}}$$

$$= \sum_{k=1}^{2}\sum_{i=1}^{3}\sum_{j=1}^{m}(\overline{y}_{ij.} - \overline{y}_{i..})^2,$$

which is based on $3(m-1)$ degrees of freedom (i.e. $m-1$ degrees of freedom in each of three drug groups).

The within-subjects residual sum of squares is the overall residual sum of squares from the model. It can be thought of as the measure of the total variation between the measurements made on the same individual (the total within-individual variation which has $3m$ degrees of freedom) less the variation which the experimenter deliberately induced within individuals; that is, less the exercise sum of squares (one degree of freedom) and the drug–exercise interaction sum of squares (two degrees of freedom). Thus, it could be calculated as:

$$\text{Residual SS}_{\text{Within Subjects}}$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{m}\sum_{k=1}^{2}(Y_{ijk} - \overline{y}_{ij.})^2$$

$$- \sum_{i=1}^{3}\sum_{j=1}^{m}\sum_{k=1}^{2}(\overline{y}_{..k} - \overline{y}_{...})^2$$

$$- \sum_{i=1}^{3}\sum_{j=1}^{m}\sum_{k=1}^{2}(\overline{y}_{i.k} - \overline{y}_{i..} - \overline{y}_{..k} + \overline{y}_{...})^2,$$

and it will have $3m - 1 - 2$ degrees of freedom.

Then, to test the between-subject effects (i.e. drug levels) or contrasts involving the means for the drug groups, we use the residual mean square from the between-subjects portion of the analysis in the F test. For the within-subjects sources of variation (i.e. exercise and the drug–exercise interaction), we use the residual mean square from the within-subjects part of the analysis in the F test. An abbreviated analysis of variance table is shown in Table 7.

Random Effects and Expected Mean Squares

In our discussion to this point we have assumed we were interested only in making statements about the treatments and blocks actually used in the experiment, and not some larger population of treatments or blocks from which those actually used were a random sample. The theory behind the analysis of variance allows us to deal with this latter situation. Note that in designed experiments it is very rare that

Table 7 Analysis of variance table for the drug–exercise example

Source	df	Sum of squares	F
Between Subjects	$3m-1$		
Drugs	2	$2m\sum_i(\overline{y}_{i..}-\overline{y}_{...})^2$	$\dfrac{\text{Drugs MS}}{\text{Residual MS}_{\text{Between Subjects}}}$
$\text{Residual}_{\text{Between Subjects}}$	$3(m-1)$	$2\sum_i\sum_j(\overline{y}_{ij\cdot}-\overline{y}_{i..})^2$	
Within Subjects	$3m$		
Exercise	1	$3m\sum_k(\overline{y}_{..k}-\overline{y}_{...})^2$	$\dfrac{\text{Exercise MS}}{\text{Residual MS}_{\text{Within Subjects}}}$
Drug × Exercise	2	$m\sum_i\sum_k(\overline{y}_{i.k}-\overline{y}_{i..}-\overline{y}_{..k}+\overline{y}_{...})^2$	$\dfrac{\text{Drug} \times \text{Exercise MS}}{\text{Residual MS}_{\text{Within Subjects}}}$
$\text{Residual}_{\text{Within Subjects}}$	$3m-3$	Subtraction from total	
Total	$6m-1$	$\mathbf{Y}'\mathbf{Y}-n\overline{y}_{..}^2$	

either treatments or blocks are chosen randomly from a larger population. Blocks used in an experiment may be thought to be representative of the larger population of blocks; however, it is common for blocks to be deliberately chosen to span the range of possible situations in which the treatments may be applied, and the assumption, at least, that the blocks are a simple random sample of all blocks is unlikely to be true.

In sample survey applications it may be more realistic to think of random samples of blocks (e.g. towns), and then further random samples within towns. In such cases we would like the inference to be made across the population of towns, and not just those actually used in the data collection.

A full treatment of this topic is not possible here. However, we illustrate the ideas in a simple example. For a thorough treatment of the subject of fixed, random and mixed effects models, see [1] and [10].

Consider a model for a randomized block experiment

$$Y_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij},$$
$$\text{with } \varepsilon_{ij}\sim \text{N}(0,\sigma^2), \text{ independently,}$$

and assume the blocks used in the experiment represent a random sample of blocks from a larger population. If this population is large relative to the number of blocks used in the experiment, it may be reasonable to assume that the block effects have a distribution which can be described by a normal probability model. That is, the β_js are randomly distributed according to the model

$$\beta_j\sim \text{N}(0,\sigma_{\text{B}}^2),\quad j=1,\ldots,b,\quad \text{independently,}$$
$$\text{and independently of the } \varepsilon_{ij}\text{s.}$$

Under this model, $Y_{ij}\sim \text{N}(\mu+\alpha_i,\sigma^2+\sigma_{\text{B}}^2)$, and the model induces a **correlation**, termed the intra-class correlation $(\sigma_{\text{B}}^2/(\sigma^2+\sigma_{\text{B}}^2))$, between observations in the same block, while maintaining the independence of observations in different blocks. The relevant hypothesis for the lack of block effects is, then, H: $\sigma_{\text{B}}^2=0$. Interestingly, this hypothesis can be tested in the same manner as in the fixed effects case, although the interpretation is different.

For example, consider the Block Mean Square, $t\sum_{j=1}^b(\overline{y}_{.j}-\overline{y}_{..})^2/(b-1)$. Then with the fixed effects assumptions on the α_is (i.e. $\sum_{i=1}^t\alpha_i=0$), and using the fact that $\text{var}(\sum_{i=1}^t Y_{ij})=t(\sigma^2+(t-1)\sigma_{\text{B}}^2)$, due to the correlation between observations in the same block, we can derive the expected value of the Blocks Mean Square as follows:

$$\begin{aligned}
&\text{E}(\overline{y}_{.j}-\overline{y}_{..})^2\\
&=\text{var}(\overline{y}_{.j}-\overline{y}_{..})+(\text{E}(\overline{y}_{.j}-\overline{y}_{..}))^2\\
&=\text{var}(\overline{y}_{.j})+\text{var}(\overline{y}_{..})-2\text{cov}(\overline{y}_{.j},\overline{y}_{..})+0\\
&=\frac{\sigma^2+t\sigma_{\text{B}}^2}{t}+\frac{\sigma^2+t\sigma_{\text{B}}^2}{tb}-\frac{2t(\sigma^2+t\sigma_{\text{B}}^2)}{t^2b}\\
&=\frac{(b-1)(\sigma^2+t\sigma_{\text{B}}^2)}{tb}
\end{aligned}$$

So,

$$\mathrm{E}\left(t\sum_{j=1}^{b}\frac{(\overline{y}_{\cdot j}-\overline{y}_{\cdot\cdot})^2}{b-1}\right)=\sigma^2+t\sigma_{\mathrm{B}}^2.$$

In this manner, we can construct expected mean squares for other sources in the analysis of variance table. In many designs, these can be useful to indicate the appropriate mean squares for tests of hypotheses about effects in the model. For example, for the randomized block design discussed earlier, Table 8 gives the expected mean squares for the design in which treatment is a fixed effect and blocks are assumed to be a random effect.

The table suggests that if $\sigma_{\mathrm{B}}^2=0$, the ratio of the Blocks Mean Square to the Residual Mean Square should be close to one, and that large values of this ratio will cast doubt on H: $\sigma_{\mathrm{B}}^2=0$. In fact, under this hypothesis, the ratio will have the $F_{(b-1),(t-1)(b-1)}$ distribution, as before. Similarly, the ratio of the Treatment Mean Square to the Residual Mean Square will be large when $\sum\alpha_i^2$ is large; that is, when there is evidence against H: $\alpha_1=\alpha_2=\cdots=\alpha_t=0$. The F test applies in this case, as well.

There has been much written on the estimation of the individual variance terms in models involving random effects. It is clear that the residual mean square is an unbiased estimate of σ^2. Further, in the model discussed above, an unbiased estimate of σ_{B}^2 is given by

$$\widehat{\sigma_{\mathrm{B}}^2}=\frac{\text{Block MS}-\text{Residual MS}}{t},$$

although it is possible for the estimate to be negative when calculated in this manner. The intra-class correlation discussed above can likewise be estimated from these quantities.

While there is no difference in the test statistics for hypotheses about treatment or block effects between this simple mixed model and the fixed effects model discussed above, there are differences when one wishes to make confidence interval statements about, say, the expected value of an observation or treatment mean taken at a future time. For the fixed effects model, this will involve specifying both the treatment and the block since both α_i and β_j are involved in the expression for the expected value. The variance term, $\mathrm{var}(Y_{ij})$, involves just σ^2 and constants; that is

$$\mathrm{var}(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j)=\sigma^2\left(\frac{1}{b}+\frac{1}{t}-\frac{1}{tb}\right).$$

For the mixed model described above, only specification of the treatment is required; the assumption is that the observation will be taken from a randomly selected block. Thus, the block effect appears in the variance expression. So

$$\mathrm{var}(\hat{\mu}+\hat{\alpha}_i)=(\sigma^2+\sigma_{\mathrm{B}}^2)\left(\frac{1}{b}\right)$$

is the variance of a future observation on the ith treatment.

The development of the expected mean squares for complicated designs, including mixed effects designs and designs with more than one source of error, is beyond the scope of this article. Discussions and algorithms for determining these expected mean squares can be found in [1, 10], and [11]. The SAS computing package [12] will also generate expected mean squares for a given model specification.

Table 8 Expected mean squares for the randomized block design with treatment effects fixed and block effects random

Source	df	Expected mean square
Model	$t+b-2$	
Treatments	$t-1$	$\sigma^2+b\sum\alpha_i^2/(t-1)$
Blocks	$b-1$	$\sigma^2+t\sigma_{\mathrm{B}}^2$
Residual	$(t-1)(b-1)$	σ^2
Total	$n-1$	

References

[1] Anderson, V.L. & McLean, R.A. (1974). *Design of Experiments: A Realistic Approach*. Marcel Dekker, New York.

[2] Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London.

[3] Daniel, C. & Wood, F.S. (1980). *Fitting Equations to Data*, 2nd Ed. Wiley, New York.

[4] Draper, N. & Smith, H. (1981). *Applied Regression Analysis*, 2nd Ed. Wiley, New York.

[5] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[6] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[7] Fisher, R.A. & MacKenzie, W.A. (1923). Studies in crop variation. II: The manurial response of different potato varieties, *Journal of Agricultural Science* **13**, 311–320.

[8] Kane, V.E. (1989). *Defect Prevention: Use of Simple Statistical Tools*. Marcel Dekker, New York.

[9] NAG (Numerical Algorithms Group) (1985). *The GLIM System Release 3.77 Manual*. NAG, Oxford.

[10] Scheffé, H. (1961). *The Analysis of Variance*. Wiley, New York.

[11] Snedecor, G.W. & Cochran, W.G. (1967). *Statistical Methods*, 6th Ed. Iowa State University Press, Ames.

[12] SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6*, 4th Ed., Vol .2. SAS Institute Inc., Cary.

(*See also* **Analysis of Covariance**; **Goodness of Fit**; **Multiple Linear Regression**)

K.S. BROWN

Analytic Hierarchy Process

Among the most important decisions made in society today are those relating to the life and health of people. How does one make a decision when faced with many conflicting factors that determine the best choice among the available alternatives of that decision? Most of these factors may be intangibles and the choice is not simply a matter of making financial trade-offs. How does one pool the judgments of doctors and other experts with their varying degrees of expertise to obtain the best decision? Who should bear the costs, and how much control should there be, what should be legal and what should not, who should receive an organ in transplant operations, and what is the best treatment for a certain type of disease? If we have standard measurements for how average people score on medical tests and if the readings of an individual on several tests do not meet these standards, how far off do they have to be before that individual is suspected of having a disease? These are examples of decision making that occur in the health and medical professions (*see* **Decision Analysis in Diagnosis and Treatment Choice**; **Decision Theory**).

Procedures for finding the answers to such questions are found in the new and rapidly spreading field of multicriteria decision making. An application of this decision-making theory is the analytic hierarchy process (AHP), which is defined below and also illustrated with an example of choosing a hospice.

The Analytic Hierarchy Process (AHP)

The analytic hierarchy process [3, 4, 6, 8] subdivides a complex decision-making problem or planning issue into its components or levels, and arranges these levels into an ascending hierarchic order. At each level of the hierarchy, the components are compared relative to each other using a pairwise comparison scheme. The components of a given level are related to an adjacent upper level and thereby generate an integration across the levels of the hierarchy. The result of this systematic process is a set of priorities or relative importance, or method of scaling between the various actions or alternatives. The relative priority weights can provide guidelines for the allocation of resources among the entities at the lower level.

Structuring any decision problem hierarchically is an efficient way to deal with and identify the major components of the problem. There is no single hierarchic structure to use in every problem. When hierarchies are designed to reflect likely environmental scenarios, corporate objectives, current and proposed product/market alternatives, and various medical strategy options, the AHP can provide a framework and methodology for the determination of a number of key decisions.

The AHP allows its users flexibility in constructing a hierarchy to fit their needs. The AHP also provides an effective structure for group decision making by imposing a discipline on the group's thought processes. The necessity of assigning a numerical value to each variable of the problem helps decision makers to maintain cohesive thought patterns by deriving the relative weight of each component of the hierarchy: criteria and alternatives. In this manner, one determines the optimum alternative. The AHP has been applied successfully to a variety of problems in planning [10], prioritization [6], resource allocation [10], conflict resolution [9], decision making, and forecasting or prediction [11], as well as in health care [1, 2, 7]. The AHP is a special case or subset of the analytic network process (ANP), which uses a network structure that allows dependence and feedback instead of a hierarchy.

The AHP focuses on dominance matrices and their corresponding measurement in contrast with the proximity, profile, and conjoint measurement approaches [12]. It goes beyond the Thurston [13] comparative judgment approach by relaxing the assumption of **normality** on the parameters, e.g. equal **variance**, zero covariance, and restriction of the type of comparisons. It is based on a trade-off concept whereby one develops the trade-off in the course of structuring and analyzing a series of simple reciprocal pairwise comparison matrices.

Some Detail

A measurement methodology is used to establish priorities among the elements within each level or stratum of the hierarchy. This is accomplished by asking the decision maker to evaluate each set of elements in a pairwise fashion with respect to their

parent element in the adjacent higher stratum. This measurement methodology provides the framework for data collection and analysis and constitutes the heart of the analytic hierarchy process. The degree of importance of the elements at a particular level over those in the succeeding level is measured by the paired comparisons. To ensure meaningful comparisons the elements are placed in homogeneous groups of a few elements in each to ensure consistency, with a pivot element from one group to the next. Each paired comparison made by a decision maker providing the judgments requires estimating how many times more one element has the property than the other element. The judgments are expressed verbally as equal, moderate, strong, very strong, and extreme. With these judgments are associated the absolute numbers (how many times more): 1, 3, 5, 7, 9. The numbers 2, 4, 6, 8 are used for compromise between the verbal judgments. In addition, reciprocals are used to represent the inverse comparison. This scale is used to compare any homogeneous set. When the elements are not homogeneous, a pivot element is used to link one cluster to an adjacent cluster. The choice of scale values to correspond to feelings and judgments is founded in both theory and in considerable experimental work, and is not to be taken lightly. When applied to comparisons of things whose measurements are already known, using this scale gives very close values when the judgments are given by an expert. For the mathematics behind the AHP and numerous applications, see [3], [5], and [6].

Unlike traditional ways of measurement that begin with a scale and apply it to measure things, in the AHP, we begin with things, measure them in pairs by using the lesser object as the unit, and then derive a scale of relative values from the pairwise measurements. The scale comes after and not before the objects. In this case, we do not need a unit, we only want relative values and it turns out that both the comparisons and the derived scale belong to the strongest possible kind of scale like the real numbers. It is known as an absolute scale, invariant under multiplication by the identity.

At each level, a set of priorities is obtained which numerically corresponds to the relative importance of the elements of that level relative to an element at an upper level. For a given level in the hierarchy if there are n elements, the solution technique will result in an n-element **eigenvector** of local priorities by solving the principal **eigenvalue** problem $(A - \lambda I)X = 0$. The components of the principal eigenvector correspond to the relative importance of each element. These priorities are now used as weighting factors for the eigenvectors generated at the next lower level in the hierarchy until all levels are completed. Applying this procedure at each level and weighting the next level and so on to the lowest level will result in a composite priority vector for the alternatives at the bottom level of the hierarchy.

One needs to use the eigenvector to derive the ratio scale priorities. Other techniques such as the method of **least squares** can minimize error but do not capture the dominance expressed numerically by the judgments. The eigenvalue process can determine the true order of dominance despite any inconsistencies or intransitivities that may occur in the judgments [5].

Example: the Hospice Problem

The following application is explained in greater detail in [6]. A hospital is concerned with the costs of the facilities and staffing involved in taking care of terminally ill patients. Often these patients do not need as much medical attention as do other patients. Those who best utilize the limited resources in a hospital are patients who require the medical attention of its specialists and advanced technology equipment–whose utilization depends on the demand of patients admitted into the hospital. The terminally ill need medical attention only episodically. Most of the time such patients need psychological support. For the mental health of the patient, home therapy may be most beneficial. From the medical standpoint, especially during a crisis, the hospital provides a greater benefit. Costs include economic costs as well as intangibles, such as inconvenience and pain. The planning association of the hospital wanted to develop alternatives for caring for terminally ill patients and to choose the best one from the standpoint of the patient, the hospital, the community, and society at large. To study the problem, one needs to deal with the benefits and costs of the decision separately (*see* **Health Economics**).

Approaching the Problem

The problem was which hospice to choose. There were three possible models under consideration: in Model I, the hospital provides full care to the

patients; in Model II, the family cares for the patient at home, and the hospital provides only emergency treatment (no nurses go to the home); and in Model III, the hospital and the home share patient care (with visiting nurses going to the home).

Two hierarchies were created by the decision makers: one for benefits and one for costs (Figures 1 and 2). For both hierarchies, the goal was to choose the best hospice. That goal is placed at the top of each hierarchy. The two hierarchies descend from the more general criteria in the second level to secondary subcriteria in the third level, then to tertiary subcriteria in the fourth level, and on to the alternatives at the bottom or fifth level.

For the benefits hierarchy (Figure 1) we decided that the decision should benefit the recipient, the institution, and society as a whole. We located these three elements on the second level of the benefits hierarchy. As the decision would benefit each party differently, it was thought important to specify the types of benefits for the recipient and the institution. Recipients want physical, psychosocial, and economic benefits. We located these benefits in the third level of the hierarchy. Each of these in turn needed further decomposition into specific items in terms of which of the decision alternatives could be evaluated. For example, while the recipient measures economic benefits in terms of reduced costs and improved productivity, the institution needed the more specific measurements of reduced length of stay, better utilization of resources, and increased financial support from the community. There was no reason to decompose the societal benefits into third-level subcriteria, and hence societal benefits

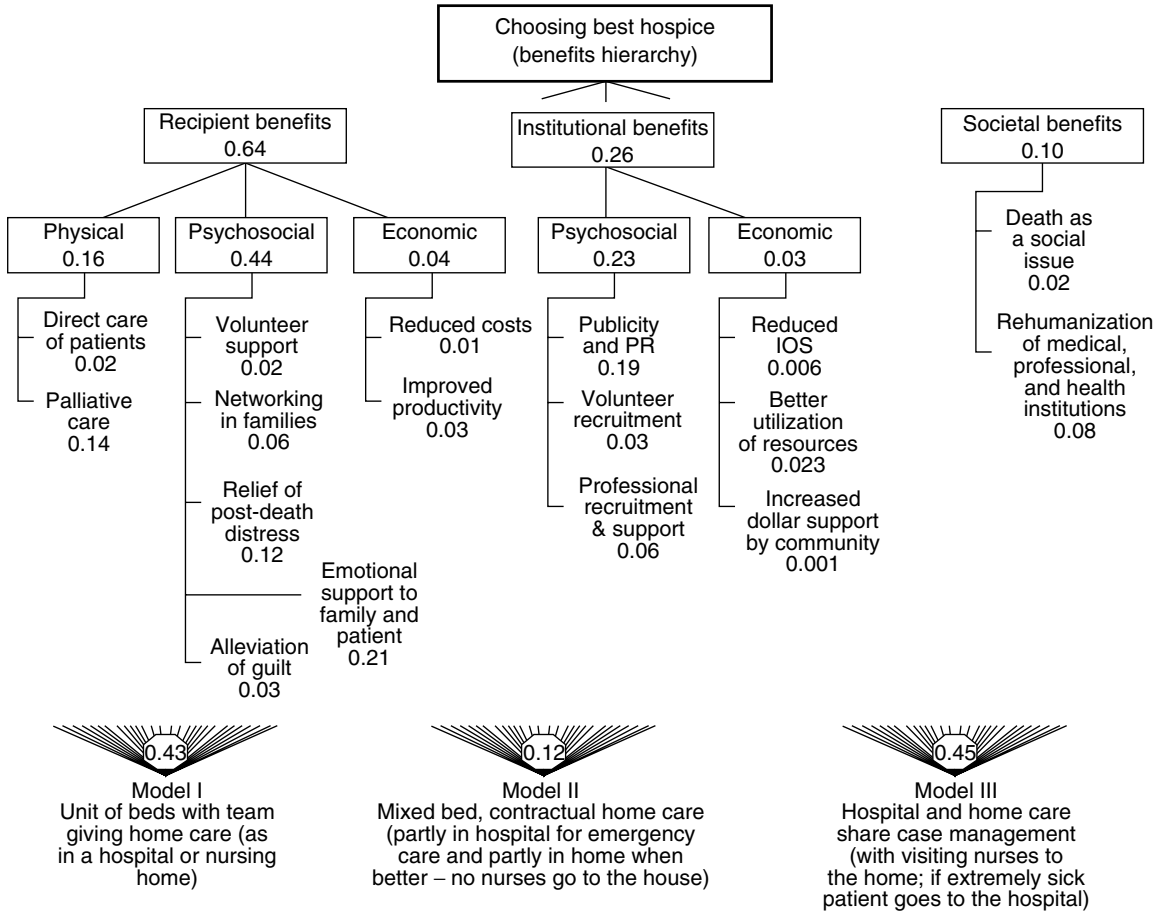

Figure 1 Benefits of choosing best hospice

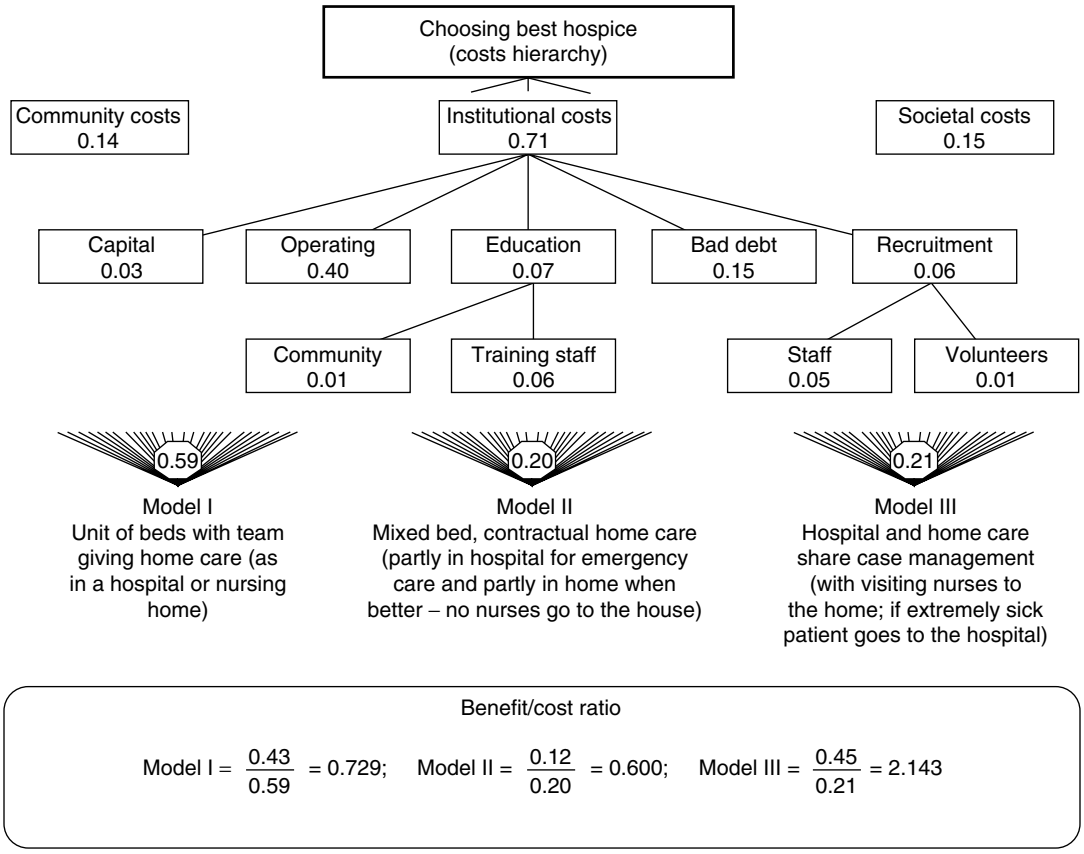

Figure 2 Costs of choosing best hospice

connect directly to the fourth level. The three hospice models are located on the bottom or fifth level of the hierarchy.

In the costs hierarchy there were also three major interests that would incur costs or burdens: community, institution, and society. In this decision the costs incurred by the patient were not included as a separate factor. Patient and family could be thought of as part of the community. We thought decomposition was necessary only for institutional costs. We included five such costs in the third level: capital costs, operating costs, education costs, bad debt costs, and recruitment costs. Educational costs apply to staff and volunteers. Since both the costs hierarchy and the benefits hierarchy concern the same decision, they both have the same alternatives in their bottom levels, even though the costs hierarchy has fewer levels.

Note that, even before preference judgments are introduced, this method of structuring the problem has changed an unformed problem into a structured problem.

Judgments and Comparisons

For both the cost and the benefit models, we compared the criteria and subcriteria according to their relative importance with respect to the parent element in the adjacent upper level. For example, one judges the importance of the three benefits criteria, as shown in Table 1. Recipient benefits were determined to be moderately more important than institutional benefits, and are assigned the absolute number 3 in the (1, 2) or first-row second-column position. The reciprocal value is automatically entered in the (2, 1) position, where institutional benefits on the left are compared with recipient benefits at the top. Similarly, a 5, corresponding to strong dominance or importance, is assigned to recipient benefits over social

Table 1 Comparing benefits of the hospice major benefit criteria

Choosing best hospice	Recipient	Institutional benefits	Social benefits	Priorities
Recipient benefits	1	3	5	0.64
Institutional benefits	1/3	1	3	0.26
Societal benefits	1/5	1/3	1	0.10

benefits in the (1, 3) position, and a 3, corresponding to moderate dominance, is assigned to institutional benefits over social benefits in the (2, 3) position, with corresponding reciprocals in the transpose positions of the matrix.

The priorities column at the right in Table 1 is the normalized eigenvector of the matrix of priorities. A matrix $A = (a_{ij})$ is consistent if and only if $a_{ij}a_{jk} = a_{ik}$ for all i, j, and k. If all the judgments were consistent, the priorities would be equivalent to the normalized sum of the elements in each row. In this case they were not. For example, institutional benefits should dominate social benefits by 5/3 and not 3. The eigenvalue solution, which handles inconsistency, must be used because, in general, people, though knowledgeable, are somewhat inconsistent. The software based on the AHP can show which judgment is the most inconsistent, and suggest the value that best improves consistency. The program determined that $a_{13} = 5$ is the most inconsistent judgment in this matrix, and suggested that instead of 5, 9 should be used. However, this recommendation may not lead to priorities in harmony with one's understanding of the real world. One might prefer to leave it at 5 and focus on the second most inconsistent judgment instead, or change the 5 to a 6 or a 7.

Similar tables were constructed for each of the other parent nodes in the model. The results of these tables are weighted and added to derive the overall priorities of the subcriteria just above the alternatives – the three models being evaluated. The first column of Table 2 shows these synthesized priorities, which sum to one for the benefits and for the costs.

The decision makers also rated each of the three hospice models with respect to each covering benefit and each covering cost. Those ratings are shown in columns 2, 3, and 4 of Table 2. For example, the first row shows that Model I provided the most direct care to the patient and Model II the least. Each of these rows sums to 1. The synthesis numbers are the sum of the product of the priorities of the subcriteria and the model weights.

The benefit/cost ratio in the last row of Table 2 is simply the quotient of the corresponding synthesis for each model. Model III would be chosen because it has the highest expected ratio of benefit to cost. This model – shared management between the hospital and home care with visiting nurses on a day-to-day basis, with hospital care for emergencies – was adopted for the community. Rounding off decimals gives slightly different numbers from the Benefit to cost ratios shown in Figure 2.

Finally, we performed a **sensitivity analysis** by varying the priorities of the criteria to determine the stability of the best alternative. We asked, for example, what would happen if the priority for institutional costs was less than the 0.71 value it received (Figure 2). The resulting ranking of the three models was fairly stable to small or realistic perturbations in the relative weights of the criteria.

Absolute Measurement–Rating Alternatives One at a Time

Cognitive psychologists have recognized for some time that people are able to make two kinds of comparisons–absolute and relative. In absolute comparisons, people compare alternatives with a standard in their memory that they have developed through experience. In relative comparisons, they compared alternatives in pairs according to a common attribute, as we did throughout the hospice example.

People use absolute measurement (sometimes also called rating) to rank independent alternatives one at a time in terms of rating intensities for each of the criteria. An intensity is a range of variation of a criterion that enables one to distinguish the quality of an alternative for that criterion. An intensity may be expressed as a numerical range of values if the criterion is measurable or in qualitative terms.

For example, if ranking students is the objective and one of the criteria on which they are to be ranked

Table 2 Synthesis

		Distributive mode		
Benefits	Priorities	Model I	Model II	Model III
Direct care of patient	0.02	0.64	0.10	0.26
Palliative care	0.14	0.64	0.10	0.26
Volunteer support	0.02	0.09	0.17	0.74
Networking in families	0.06	0.46	0.22	0.32
Relief of postdeath stress	0.12	0.30	0.08	0.62
Emotional support of family and patient	0.21	0.30	0.08	0.62
Alleviation of guilt	0.03	0.30	0.08	0.62
Recipient economic reduced costs	0.01	0.12	0.65	0.23
Improved productivity	0.03	0.12	0.27	0.61
Publicity and PR	0.19	0.63	0.08	0.29
Volunteer recruitment	0.03	0.64	0.10	0.26
Professional recruitment and support	0.06	0.65	0.23	0.12
Reduced length of stay	0.006	0.26	0.10	0.64
Better utilization of resources	0.023	0.09	0.22	0.69
Increased financial support	0.001	0.73	0.08	0.19
Death as a social issue	0.02	0.20	0.20	0.60
Rehumanization of institutions	0.08	0.24	0.14	0.62
Synthesis		0.428	0.121	0.451
Costs				
Community costs	0.14	0.33	0.33	0.33
Institutional capital costs	0.03	0.76	0.09	0.15
Institutional operating costs	0.40	0.73	0.08	0.19
Institutional costs educating community	0.01	0.65	0.24	0.11
Institutional costs training staff	0.06	0.56	0.32	0.12
Institutional bad debt	0.15	0.60	0.20	0.20
Institutional costs recruiting staff	0.05	0.66	0.17	0.17
Institutional costs recruiting volunteers	0.01	0.60	0.20	0.20
Societal costs	0.15	0.33	0.33	0.33
Synthesis		0.583	0.192	0.224
Benefit/cost ratio		0.734	0.630	2.013

is performance in mathematics, the mathematics ratings might be: excellent, good, average, below average, poor; or, using the usual school terminology, A, B, C, D, and F. Relative comparisons are first used to set priorities on the ratings themselves. If desired, one can fit a continuous curve through the derived intensities. This concept may go against our socialization. However, it is perfectly reasonable to ask how much an A is preferred to a B or to a C. The judgment of how much an A is preferred to a B might be different under different criteria. Perhaps for mathematics an A is very strongly preferred to a B, while for physical education an A is only moderately preferred to a B. So the end result might be that the ratings are scaled differently. For example, one could have the following scale values for the ratings:

	Math	Physical education
A	0.50	0.30
B	0.30	0.30
C	0.15	0.20
D	0.04	0.10
E	0.01	0.10

The alternatives are then rated or ticked off one at a time on the intensities.

I will illustrate absolute measurement with an example. A firm evaluates its employees for raises. The criteria are dependability, education, experience, and quality. Each criterion is subdivided into intensities, standards, or subcriteria (Figure 3). The managers set priorities for the criteria by comparing them in pairs. They then pairwise compare the intensities according to priority with respect to their parent criterion (as in Table 3) or with respect to a subcriterion if they are using a deeper hierarchy. The priorities of the intensities are divided by the largest intensity for each criterion to put them in ideal form (second column of priorities in Figure 3).

Table 3 shows a paired comparison matrix of intensities with respect to dependability. The managers answer the question: which intensity is more important and by how much with respect to dependability. Finally, the managers rate each individual (Table 4) by assigning the intensity rating that applies to him or her under each criterion. The scores of these intensities are each weighted by the priority of its criterion and summed to derive a total ratio scale score for the individual (shown on the right of Table 4). These numbers belong to a ratio scale, and the managers can give salary increases precisely in proportion to the ratios of these numbers. Adams gets the highest score and Kessel the lowest. This approach can be used whenever it is possible to set priorities for intensities of criteria; people can usually do this when they have sufficient experience with a given operation. This normative mode requires that alternatives be rated one by one without regard to how many there may be and how high or low any of them rates on prior standards. Some corporations have insisted that they no longer trust the normative standards of their experts and that they prefer to

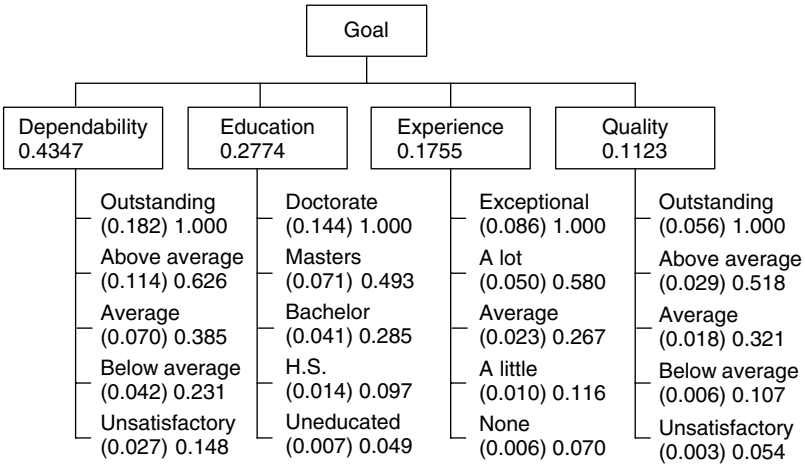

Figure 3 Criteria with priorities and their intensities both prioritized and idealized

Table 3 Ranking intensities: Which intensity is preferred most with respect to dependability and how strongly?

Intensities	Outstanding	Above average	Average	Below average	Unsatisfactory	Priorities
Outstanding	1.0	2.0	3.0	4.0	5.0	0.419
Above av	1/2	1.0	2.0	3.0	4.0	0.263
Average	1/3	1/2	1.0	2.0	3.0	0.160
Below av	1/4	1/3	1/2	1.0	2.0	0.097
Unsatisfact	1/5	1/4	1/3	1/2	1.0	0.062
						C.R. = 0.015

Table 4 Ranking alternatives. The priorities of the intensities for each criterion are divided by the largest one and multiplied by the priority of the criterion. Each alternative is rated on each criterion by assigning the appropriate intensity. The weighted intensities are added to yield the total on the right

Alternatives	Dependability 0.4347	Education 0.2774	Experience 0.1775	Quality 0.1123	Total
1. Adams, V.	Outstanding	Bachelor	A little	Outstanding	0.646
2. Becker, L.	Average	Bachelor	A little	Outstanding	0.379
3. Hayat, F.	Average	Masters	A lot	Below average	0.418
4. Kessel, S.	Above av	H.S.	None	Above average	0.369
5. O'Shea, K.	Average	Doctorate	A lot	Above average	0.605
6. Peters, T.	Average	Doctorate	A lot	Average	0.583
7. Tobias, K.	Above av	Bachelor	Average	Above average	0.456

make paired comparisons of their alternatives. Still, when there is wide agreement on standards, the absolute mode saves time in rating a large number of alternatives.

Other applications for AHP in the health care environment are as follows:

1. Choosing the most efficient hospital management information supply order system. The main criteria in this application were cost, speed, simplicity, flexibility, quality, and security. Most of these criteria had two or more subcriteria which were used to choose the best of the three alternatives: manual, computer, and Wand system, with the highest priority going to the last alternative.
2. Choosing a corporate health plan from among five options: fee for services (FFS), and Health Maintenance Organization (HMO), which has the three options: Physician Provider Organization (PPO), Individual Practice Association (IPA), and Staff and Group.
3. Selecting the best way to provide health care for everyone.
4. Performing a benefit/cost analysis to decide on an infant formula policy: sell to industrial countries, sell to Third World countries, stop selling.
5. Conducting a benefit/cost analysis as to whether drugs should be legalized in the US (No was nearly double Yes).
6. Doing a benefits (B), opportunities (O), costs, (C) and risks (R) called BOCR analysis to determine the best of the following five alternatives for solving a physicians heart problem: bypass operation, medicinal treatment, angioplasty, transplant, and doing nothing. When one does costs and risks, one needs to ask which is more costly and which is more risky because the smaller element is used as the unit and the larger one is estimated as a multiple of that unit. It cannot be done the other way by estimating the smaller as a fraction of the larger without using the smaller first as the unity. The importance of each of the four BOCR merits is determined by finding the best of the alternatives with respect to each one. Because there may be several criteria for each merit and because the alternatives are put into ideal form for each criterion, the winning or ideal alternative may not receive an overall value of one for that merit. That alternative is then rated as a representative of its merit using a set of strategic criteria and

their priorities used to evaluate all the decisions made for this type of problem. Here, the strategic criteria or subcriteria are prioritized as in a hierarchy and are then each assigned different intensities like high medium and low, or excellent, very good, average, and poor that are prioritized, then idealized by dividing by the largest value thus making the priority 1 as the standard and then assigning an alternative one intensity for each criterion and finally, multiplying the idealized priorities of the intensities by the priorities of their corresponding criteria and adding to obtain the overall importance rating of that alternative and hence, also of the merit it represents. In this manner, one obtains a rating for each of the four BOCR merits. These ratings are then normalized by dividing each value by their sum to obtain their priorities that are used to weight the priorities of the alternatives under each and then take the sum of the benefits and opportunities and subtract from it the sum of the costs and the risks for each alternative. We note that the outcome for some or all the alternatives may be negative and we have negative priorities.

7. Choosing a treatment for breast cancer, based on both physician and patient values.

In conclusion, the AHP and its generalization to feedback networks the Analytic network Process (ANP) (there is a book on the ANP and its applications by this author, 2001, RWS Publications) are finding increasing uses in the medical field. The software package Super Decisions, enables users to implement easily the AHP/ANP in decision-making. It can be downloaded from creativedecisions.net. The ANP has powerful predictive content as has been demonstrated in many examples involving market share and in predicting turn around in the US economy, and the presidential elections since 1976.

References

[1] Cook, D.R., Staschak, S. & Green, W.T. (1990). Equitable allocation of livers for orthotopic transplantation: an application of the analytic hierarchy process, *European Journal of Operational Research* **48**, 49–56.

[2] Odynocki, B. (1979). Planning the National Health Insurance Policy: An Application of the Analytic Hierarchy Process in Health Policy Evaluation and Planning, *Dissertation*. University of Pennsylvania.

[3] Saaty, T.L. (1990). *The Analytic Hierarchy Process*. RWS Publications, 4922 Ellsworth Avenue, Pittsburgh.

[4] Saaty, T.L. (1991). A natural way to make momentous decisions, *Journal of Scientific and Industrial Research* **51**, 561–571.

[5] Saaty, T.L. (1994). *Fundamentals of Decision Making with the Analytic Hierarchy Process*. RWS Publications, 5001 Baum Boulevard, Pittsburgh.

[6] Saaty, T.L. (1995). *Decision Making for Leaders*. RWS Publications, 4922 Ellsworth Avenue, Pittsburgh.

[7] Saaty, T.L. (1995). Decision making in the health care system – seven cases, *International Journal of Management and Systems* **10**, 219–258.

[8] Saaty, T.L. (1996). *Decision Making With Dependence and Feedback: The Analytic Network Process*. RWS Publications, 5001 Baum Boulevard, Pittsburgh.

[9] Saaty, T.L. & Alexander, J.M. (1989). *Conflict Resolution: The Analytic Hierarchy Approach*. Praeger, New York.

[10] Saaty, T.L. & Kearns, K.P. (1985). *Analytical Planning: The Organization of Systems*. Pergamon Press, New York (now through RWS Publications).

[11] Saaty, T.L. & Vargas, L.G. (1991). *Prediction, Projection and Forecasting*. Kluwer, Boston (now through RWS Publications).

[12] Shepard, R.N. (1972). taxonomy of some principal types of data and of multidimensional methods for their analysis, in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. 1, R.N. Shepard, ed. Seminar Press, New York.

[13] Thurston, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 273–286.

THOMAS L. SAATY

Analytic Epidemiology

Analytic epidemiology denotes epidemiologic studies, such as **cohort studies**, **cross-sectional studies**, and **case–control studies**, that obtain individual-level information on the **association** between disease status and exposures of interest. Such analytic studies often include individual-level information on potential **confounders** and **effect modifiers**. Usually such studies are designed to evaluate predetermined hypotheses concerning possible causal relationships between exposure and risk of disease (*see* **Causation**). Analytic epidemiology is distinguished from **descriptive epidemiology**, which focuses on quantifying trends and rates of disease in populations and on **ecologic studies** that attempt to correlate rates of disease in populations with average levels of exposure in such populations. Descriptive studies are often used to generate etiologic hypotheses that are tested in subsequent analytic studies.

MITCHELL H. GAIL

Ancillary Statistics

In a parametric model $f(\mathbf{y};\boldsymbol{\theta})$ for a **random variable** or vector $\mathbf{Y}$, a statistic $\mathbf{A} = a(\mathbf{Y})$ is ancillary for $\boldsymbol{\theta}$ if the distribution of $\mathbf{A}$ does not depend on $\boldsymbol{\theta}$. As a very simple example, if $\mathbf{Y}$ is a vector of independent, identically distributed random variables each with mean $\boldsymbol{\theta}$, and the sample size is determined randomly, rather than being fixed in advance, then $\mathbf{A}$ = number of observations in $\mathbf{Y}$ is an ancillary statistic. This example could be generalized to more complex structure for the observations $\mathbf{Y}$, and to examples in which the sample size depends on some further parameters that are unrelated to $\boldsymbol{\theta}$. Such models might well be appropriate for certain types of sequentially collected data arising, for example, in clinical trials.

Fisher [5] introduced the concept of an ancillary statistic, with particular emphasis on the usefulness of an ancillary statistic in recovering **information** that is lost by reduction of the sample to the **maximum likelihood** estimate $\hat{\boldsymbol{\theta}}$, when the maximum likelihood estimate is not minimal **sufficient**.

An illustrative, if somewhat artificial, example is a sample $(Y_1, \ldots, Y_n)$, where now n is fixed, from the uniform distribution on $(\theta, \theta + 1)$. The largest and smallest observations, $(Y_{(1)}, Y_{(n)})$, say, form a minimal sufficient statistic for θ. The maximum likelihood estimator of θ is any value in the interval $(Y_{(n)} - 1, Y_{(1)})$, and the **range** $Y_{(n)} - Y_{(1)}$ is an ancillary statistic. In this example, while the range does not provide any information about the value of θ that generated the data, it does provide information on the precision of $\hat{\theta}$. In a sample for which the range is 1, $\hat{\theta}$ is exactly equal to θ, whereas a sample with a range of 0 is the least informative about θ.

A theoretically important example discussed in Fisher [5] is the location model (*see* **Location–Scale Family**), in which $\mathbf{Y} = (Y_1, \ldots, Y_n)$, and each Y_i follows the model $f(y - \theta)$, with $f(\cdot)$ known but θ unknown. The vector of **residuals** $\mathbf{A} = (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$, where $\overline{Y} = n^{-1} \sum Y_i$, has a distribution free of θ, as is intuitively obvious, since both Y_i and $\overline{Y}$ are centered at θ. The uniform example discussed above is a special case of the location model, and the range $Y_{(n)} - Y_{(1)}$ is also an ancillary statistic for the present example. In fact, the vector $\mathbf{B} = (Y_{(2)} - Y_{(1)}, \ldots, Y_{(n)} - Y_{(1)})$ is also ancillary, as is $\mathbf{C} = (Y_1 - \hat{\theta}, \ldots, Y_n - \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimate of θ. A *maximal* ancillary provides the largest possible conditioning set, or the largest possible reduction in dimension, and is analogous to a minimal sufficient statistic. **A**, **B**, and **C** are maximal ancillary statistics for the location model, but the range is only a maximal ancillary in the location uniform.

An important property of the location model is that the exact conditional distribution of the maximum likelihood estimator $\hat{\theta}$, given the maximal ancillary $\mathbf{C}$, can be easily obtained simply by renormalizing the **likelihood** function:

$$p(\hat{\theta}|\mathbf{c};\theta) = \frac{L(\theta;\mathbf{y})}{\int L(\theta;\mathbf{y})\,\mathrm{d}\theta}, \tag{1}$$

where $L(\theta;\mathbf{y}) = \prod f(y_i;\theta)$ is the likelihood function for the sample $\mathbf{y} = (y_1, \ldots, y_n)$, and the right-hand side is to be interpreted as depending on $\hat{\theta}$ and $\mathbf{c}$, using the equations $\sum \partial\{\log f(y_i;\theta)\}/\partial\theta|_{\hat{\theta}} = 0$ and $\mathbf{c} = \mathbf{y} - \hat{\theta}$.

The location model example is readily generalized to a **linear regression** model with nonnormal errors. Suppose that, for $i = 1, \ldots, n$, we have independent observations from the model $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i$, where the distribution of ε_i is known. The vector of standardized residuals $(Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/\hat{\sigma}$ is ancillary for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ and there is a formula similar to (1) for the distribution of $\hat{\boldsymbol{\theta}}$, given the residuals.

It is possible, and has been argued, that Fisher's meaning of ancillarity included more than the requirement of a distribution free of $\boldsymbol{\theta}$: that it included a notion of a physical mechanism for generating the data in which some elements of this mechanism are "clearly" not relevant for assessing the value of $\boldsymbol{\theta}$, but possibly relevant for assessing the accuracy of the inference about $\boldsymbol{\theta}$. Thus, Kalbfleisch [7] makes a distinction between an experimental and a mathematical ancillary statistic. Fraser [6] developed the notion of a structural model as a physically generated extension of the location model. Efron & Hinkley [4] gave particular attention to the role of an ancillary statistic in estimating the variance of the maximum likelihood estimator.

Two generalizations of the concept of ancillary statistic have become important in recent work in the theory of **inference**. The first is the notion of approximate ancillarity, in which the distribution of $\mathbf{A}$ is not required to be entirely free of $\boldsymbol{\theta}$, but free of $\boldsymbol{\theta}$ to some order of approximation. For example, we might require that the first few

moments of $\mathbf{A}$ be constant (in $\boldsymbol{\theta}$), or that the distribution of $\mathbf{A}$ be free of $\boldsymbol{\theta}$ in a neighborhood of the true value $\boldsymbol{\theta}_0$, say. The definition used by Barndorff-Nielsen & Cox [1] is that $\mathbf{A}$ is qth order locally ancillary for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$ if $f(\mathbf{a};\boldsymbol{\theta}_0 + \delta/\sqrt{n}) = f(\mathbf{a};\boldsymbol{\theta}_0) + O(n^{-q/2})$. Approximate ancillary statistics are also discussed in McCullagh [9] and Reid [11]. The notion of an approximate ancillary statistic has turned out to be rather important for the asymptotic theory of statistical inference, because the location family model result given in (1) can be generalized, to give the result

$$p(\hat{\boldsymbol{\theta}}|\mathbf{a};\boldsymbol{\theta}) \doteq c(\boldsymbol{\theta},\mathbf{a})|j(\hat{\boldsymbol{\theta}})|^{1/2}\frac{L(\boldsymbol{\theta};\mathbf{y})}{L(\hat{\boldsymbol{\theta}};\mathbf{y})}, \qquad (2)$$

where c is a normalizing constant, $j(\boldsymbol{\theta}) = -\partial^2 \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ is the observed Fisher information function, $\mathbf{a}$ is an approximately ancillary statistic, and in the right-hand side $\mathbf{y}$ is a function of $\hat{\boldsymbol{\theta}}, \mathbf{a}$. This approximation, which is typically much more accurate than the normal approximation to the distribution of $\hat{\boldsymbol{\theta}}$, is known as Barndorff-Nielsen's approximation, or the p^* approximation, and is reviewed in Reid [10] and considered in detail in Barndorff-Nielsen & Cox [1]. In (2) the likelihood function is normalized by a slightly more elaborate looking formula than the simple integral in (1), but the principle of renormalizing the likelihood function has still been applied. A distribution function approximation analogous to (2) is also available: see Barndorff-Nielsen & Cox [1] and Reid [11].

Suppose that the parameter $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, where $\boldsymbol{\psi}$ is the parameter of interest and $\boldsymbol{\lambda}$ is a **nuisance parameter**. For example, $\boldsymbol{\psi}$ might parameterize a regression model for survival time as a function of several covariates, and $\boldsymbol{\lambda}$ might parameterize the baseline **hazard** function. If we can partition the minimal sufficient statistic for $\boldsymbol{\theta}$ as $(\mathbf{S}, \mathbf{T})$, such that

$$f(\mathbf{s},\mathbf{t};\boldsymbol{\theta}) = f(\mathbf{s}|\mathbf{t};\boldsymbol{\psi})f(\mathbf{t};\boldsymbol{\lambda}), \qquad (3)$$

then $\mathbf{T}$ is an ancillary statistic for $\boldsymbol{\psi}$ in the sense of the above definition. Factorizations of the form given in (3) are the exception, though, and we more often have a factorization of the type

$$f(\mathbf{s},\mathbf{t};\boldsymbol{\theta}) = f(\mathbf{s}|\mathbf{t};\boldsymbol{\psi})f(\mathbf{t};\boldsymbol{\psi},\boldsymbol{\lambda}) \qquad (4)$$

or

$$f(\mathbf{s},\mathbf{t};\boldsymbol{\theta}) = f(\mathbf{s}|\mathbf{t};\boldsymbol{\psi},\boldsymbol{\lambda})f(\mathbf{t};\boldsymbol{\lambda}). \qquad (5)$$

An example of (4) is the **two-by-two table**, with ψ the log **odds ratio**. The conditional distribution of a single cell entry, given the row and column totals, depends only on ψ: this is the basis for **Fisher's exact test** (*see* **Conditionality Principle**). Although it is sometimes claimed that the row total is an ancillary statistic for the parameter of interest, this is in fact not the case, at least according to the definition of ancillarity discussed here. Some more general notions of ancillarity have been proposed in the literature, but have not proved to be widely useful in theoretical developments. Further discussion of ancillarity and conditional inference in the presence of nuisance parameters can be found in Liang & Zeger [8] and Reid [10].

Ancillary statistics are defined for parametric models, so would not be defined, for example, in Cox's **proportional hazards** regression model (*see* **Cox Regression Model**). Cox [2] did originally argue, though, that the full likelihood function could be partitioned into a factor that provided information on the regression parameters $\boldsymbol{\beta}$ and a factor that provided no information about $\boldsymbol{\beta}$ in the absence of knowledge of the baseline hazard: the situation is analogous to (4) but, as was pointed out by several discussants of [2], the likelihood factor that is used in the analysis is not in fact the conditional likelihood for any observable random variables. Cox [3] developed the notion of **partial likelihood** to justify the now standard estimates of $\boldsymbol{\beta}$.

References

[1] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London. (This is the only book currently available that gives a survey of many of the main ideas in statistical theory along with a detailed discussion of the asymptotic theory for likelihood inference that has been developed since 1980 (see **Large-sample Theory**). Chapter 2.5 discusses ancillary statistics, and Chapters 6 and 7 consider the p^* approximation and the related distribution function approximation.).

[2] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[3] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[4] Efron, B. & Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion), *Biometrika* **65**, 457–487.

[5] Fisher, R.A. (1934). Two new properties of mathematical likelihood, *Proceedings of the Royal Society, Series A* **144**, 285–307.
[6] Fraser, D.A.S. (1968). *The Structure of Inference*. Wiley, New York.
[7] Kalbfleisch, J.D. (1975). Sufficiency and conditionality, *Biometrika* **62**, 251–259.
[8] Liang, K.-Y. & Zeger, S.L. (1995). Inference based on estimating functions in the presence of nuisance parameters, *Statistical Science* **10**, 158–172.
[9] McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall, London.
[10] Reid, N. (1988). Saddlepoint approximations in statistical inference, *Statistical Science* **3**, 213–238. (A review of the p^* approximation and related developments, up to 1987.).
[11] Reid, N. (1995). The roles of conditioning in inference, *Statistical Science* **10**, 138–157.

Further Reading

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator, *Biometrika* **70**, 343–365. (Among the early papers on the p^* approximation, this is one of the easier ones.)
Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio, *Biometrika* **73**, 307–322. (Introduces the distribution function approximation based on the p^* approximation.)
Fisher, R.A. (1973). *Statistical Methods and Scientific Inference*, 3rd Ed. Oliver & Boyd, Edinburgh. (Includes a discussion of ancillary statistics with many examples.)
Jorgensen, B. (1994). The rules of conditional inference: is there a universal definition of nonformation?, *Journal of the Italian Statistical Society* **3**, 355–384. (This considers several definitions of ancillarity in the presence of nuisance parameters.)
Kalbfleisch, J.G. (1985). *Probability and Statistical Inference*, Vol. II, 2nd Ed. Springer-Verlag, New York. (This is one of the few undergraduate textbooks that treats ancillarity in any depth, in Chapter 15.)

N. REID

Animal Screening Systems

Animal screening systems are experimental protocols designed primarily to screen substances in laboratory animals for possible adverse effects in humans. It is necessary to use animals as surrogates for humans because of ethical concerns associated with the intentional exposure of human subjects to potentially harmful substances in a controlled experiment and the frequent lack of comprehensive epidemiology data associated with unintentional exposures. The most important of the animal screening systems are discussed below.

Long-term Rodent Toxicity/Carcinogenicity Studies

The objective of long-term rodent carcinogenicity experiments is to determine if the administration of a test substance to laboratory animals will alter the normal pattern of tumor development (*see* **Tumor Incidence Experiments**). The test substance may be given in the diet or administered by other routes, such as inhalation, skin paint, or oral gavage. A typical experiment uses male and female rats and mice, with three dosed groups and a control, each group containing 50 animals. These animals are given the test substance for most of their natural lifespan (generally two years). Following necropsy, tissues taken from a number of different organ sites are examined microscopically for evidence of carcinogenic effects.

Important statistical issues that arise in the design, analysis, and interpretation of long-term toxicity/carcinogenicity studies include: (i) selecting the number, magnitude, and spacing of doses [3, 20] (ii) use of survival-adjusted methods in the evaluation of tumor data [13, 17] (iii) **multiple comparison** adjustments for the large number of tumor sites and types evaluated [11, 15, 36] (iv) use of **historical control** data [16, 17] and (v) adjustment for potentially **confounding** factors such as body weight that may be correlated with tumor incidence [31]. For further discussion of these issues, see **Tumor Incidence Experiments**. Studies of this type are often called **bioassays** although this term is more properly applied to experiments for the estimation of relative potency (*see* **Biological Assay, Overview**).

The carcinogenicity screening assay is the first step (often denoted the hazard identification phase) of a broader risk assessment effort to determine the magnitude of human risk that may be associated with exposure to the test substance. Thus, another major statistical issue in this area is the development of appropriate methodology to permit an extrapolation of experimental results (i) from high to low doses and (ii) from species to species [26] (*see* **Extrapolation, Low Dose**; **Extrapolation**). Critical in this evaluation is the proposed mechanism of action for the development of tumors and the effective dose of the chemical reaching the target site [27].

Short-term Toxicity Studies

The long-term rodent bioassay is generally preceded by short-term (generally 90 day) toxicity studies, which evaluate variables such as body and organ weights, histopathology, clinical pathology (hematology, clinical chemistry, and urinalysis data), sperm morphology and vaginal cytology, and immunotoxicology and neurobehavioral effects [4]. In related studies, the metabolism and distribution of the test substance are also evaluated. The objectives of these experiments are: (i) to assess chemical toxicity to permit an appropriate dose selection for the long-term study; (ii) to identify possible target organs for adverse effects; and (iii) to understand better the mechanism which may be responsible for the adverse effects observed. Statistical issues that arise in the evaluation of these studies include the multiplicity of parameters evaluated (*see* **Simultaneous Inference**), adjustment for **covariates** that may be affecting the observed responses, and choice of an appropriate model relating adverse effect to dose.

Screens for Anti-carcinogenesis

While chemically related decreases in tumor incidence frequently occur in long-term rodent carcinogenicity studies [18], animal studies can also be designed specifically to screen for anti-carcinogenic effects. Typically, a known carcinogen (such as dimethylbenzanthracene) is used to induce site-specific tumors in a short period of time, and the test compound is then administered to determine if it can counteract the development of these induced tumors [2]. Tamoxifen, now commonly used in

humans in adjuvant therapy for breast cancer [5], was first discovered to suppress the growth of chemically induced mammary tumors in rodents [21]. The knowledge gained from anti-carcinogenesis studies may also be useful in the design of human **clinical trials**. The statistical issues that arise in the design, analysis, and interpretation of anti-carcinogenesis screens are similar to those discussed above for toxicity/carcinogenicity studies.

A procedure for the mass screening of compounds for anti-tumor activity should aim to select the small proportion of active materials, and to reject the higher proportion of inactive materials, as economically and reliably as possible. Studies of the operating characteristic curve [22, 29] (*see* **Power**) and application of **decision theory** [6, 9], support the use of a multistage screen, in which unpromising materials can be rejected quickly, while more promising ones are tested further before acceptance.

Reproductive and Developmental Toxicology Experiments

These studies are designed to evaluate the impact of chemical exposure or other test agents on the developing fetus [24] (*see* **Teratology**). The objectives of these experiments are: (i) to assess reproductive performance; (ii) to identify developmental defects due to exposure to chemicals; (iii) to study the biologic mechanisms of developmental toxicants; and (iv) to establish "safe" conditions for their use or consumption by humans [14]. Typically, pregnant dams are administered the agent of interest during the period of gestation, and the animals are later sacrificed and examined to determine whether or not they exhibit chemically related effects on fetal implantation, mortality, and malformations [19].

The most important statistical issue associated with the evaluation of developmental toxicity data is how to take litter effects into account [12, 19]. That is, fetuses sampled from the same female represent multiple observations on a single experimental unit (the litter), and it is likely that the individual fetal responses will be correlated. If this correlation is not taken into account, any calculated test statistics or confidence intervals could be adversely affected (*see* **Correlated Binary Data**; **Multilevel Models**). Other important statistical issues include the development of tests for multiple binary factors such as different types of malformations [23], and the development of mathematical models for low dose risk estimation [14] (*see* **Dose–Response Models in Risk Analysis**).

Skin Corrosion Tests

The potential for chemicals to cause skin effects such as corrosion is a concern of industrial toxicologists in their assessments of possible worker and consumer safety issues, and animal models have often been used to screen substances for corrosive effects [25]. One protocol used for this purpose is the Draize test [8], in which the test substance is tested on the skin of albino rabbits. A material is considered corrosive if the structure of the tissue at the site of contact is destroyed or changed irreversibly after an exposure period typically ranging from four hours to 48 hours.

Because of concerns regarding the humane treatment of laboratory animals, one important statistical issue in this area is to devise a testing strategy that minimizes animal use. In addition, a high priority research activity is to develop *in vitro* methods for assessing corrosivity that do not require direct exposure of laboratory animals. Several promising new *in vitro* methods have been proposed [25, 37].

New Alternative Methods for Screening

Laboratory animal researchers are currently seeking new screening systems that are as (or more) predictive of human health hazard as the current methods, but can be conducted in a shorter time frame, with less expense, and requiring fewer animals. The 1993 NIH Revitalization Act in the US directed the various NIH Institutes to develop and validate alternate methods for acute and chronic safety testing that will reduce animal use, replace animals with nonanimal methods or lower species, or refine animal use to decrease or eliminate pain or distress. Certain of these alternative methods are discussed below.

Transgenic Animals

One of the potentially most important new alternative screening systems is the use of transgenic mouse lines, which provides the opportunity to develop relatively short-term *in vivo* models to identify carcinogens and other toxic agents. Such models include

transgenic mice carrying reporter genes that may serve as targets for mutagenic events, or mice carrying specific oncogenes or inactivated tumor-suppressor genes that are important factors contributing to the **multistage process of carcinogenesis** [7]. Mouse lines with defined genetic alterations that result in overexpression or inactivation of a gene intrinsic to carcinogenesis, but that are insufficient alone for neoplastic conversion, are promising models for chemical carcinogen identification and evaluation. Such studies may provide advantages in shortening the time required for bioassays and improving the accuracy of carcinogen identification [33]. One important statistical issue in this area is to develop appropriate model validation procedures.

Fish Models for Carcinogenicity Studies

The US National Toxicology Program (NTP) recently initiated studies to evaluate the feasibility of using two small fish models – Medaka and guppy – to determine the toxic and carcinogenic potential of selected chemicals. This project will include a comparative evaluation of molecular lesions in both the rodent and fish tumors to identify similarities and differences in activated oncogenes and/or mutations in tumor-suppressor genes. The greatest screening potential for fish at this time appears to be as a mid-tier carcinogenesis screen that could be conducted in a shorter time frame and less expensively than the two-year rodent bioassay [32].

Frog Embryo Teratogenesis Assay (FETAX)

This assay is a 96-hour whole-embryo developmental toxicity test that utilizes embryos of the South African clawed frog. Embryos are exposed to the test chemical continuously from the early blastula to free-swimming larvae stages. FETAX endpoints can potentially detect all four manifestations of mammalian developmental toxicity: growth retardation, structural malformations, death and functional deficits. An NTP multilaboratory validation study with 20 coded chemicals (including strong, weak, and nonteratogens) has recently been completed to determine further the repeatability and reliability of the FETAX system [32].

Computer-based Prediction Systems

Another method of reducing the number of animals used in screening tests is to derive computer-based systems to predict the outcome of such tests. In the area of carcinogenicity, a number of different methods have been proposed, some based on chemical structure [10, 30] and others based on the activity of the test compound in surrogate biological test systems such as **mutagenicity** and short-term toxicity studies [1, 34]. These approaches typically employ **multiple linear regression** techniques, **discriminant analysis**, and/or artificial **neural network**/decision tree strategies to identify important predictor variables for the endpoint of interest (*see* **Tree-structured Statistical Methods**). For an overview of these various methods, see Richards [28]. For the results of an application of these prediction systems to actual carcinogenicity data, see [35].

References

[1] Bahler, D. & Bristol, D.W. (1993). The induction of rules for predicting chemical carcinogenesis in rodents, in *Intelligent Systems for Molecular Biology*, L. Hunter, J. Shavlik & D. Searls, eds. AAAI/MIT Press, Menlo Park, pp. 29–37.

[2] Boone, C.W., Steele, V.E. & Kelloff, G.L. (1992). Screening for chemopreventive (anticarcinogenic) agents in rodents, *Mutation Research* **267**, 251–255.

[3] Bucher, J.R., Portier, C.J., Goodman, J.I., Faustman, E.M. & Lucier, G.W. (1996). National Toxicology Program studies: principles of dose selection and applications to mechanistic based risk assessment, *Fundamental and Applied Toxicology* **31**, 1–8.

[4] Chhabra, R.S., Huff, J.E., Schwetz, B.S. & Selkirk, J. (1990). An overview of prechronic and chronic toxicity/carcinogenicity experimental designs and criteria used by the National Toxicology Program, *Environmental Health Perspectives* **86**, 313–321.

[5] Cuzick, J. & Baum, M. (1986). Tamoxifen and contralateral breast cancer, *Lancet* **2**, 282.

[6] Davies, O.L. (1963). The design of screening tests, *Technometrics* **5**, 481–489.

[7] Donehower, L.A., Harvey, M., Slagle, B.L., McArthur, M.J., Montgomery, C.J., Butel, J.S. & Bradley, A. (1992). Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumors, *Nature* **356**, 215–221.

[8] Draize, J.H., Woodward, G. & Calvery, H.O. (1944). Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes, *Journal of Pharmacology and Experimental Therapeutics* **82**, 377–390.

[9] Dunnett, C.W. (1961). Statistical theory of drug screening, in *Quantitative Methods in Pharmacology*, H. de Jonge, ed. North-Holland, Amsterdam, pp. 212–231.

[10] Enslein, K. & Craig, P.N. (1982). Carcinogenesis: a predictive structure-activity model, *Journal of Toxicology and Environmental Health* **10**, 521–530.

[11] Farrar, D.B. & Crump, K.S. (1988). Exact statistical tests for any carcinogenic effect in animal bioassays, *Fundamental and Applied Toxicology* **11**, 652–663.

[12] Gad, S. & Weil, C.S. (1986). *Statistics and Experimental Design for Toxicologists*. Telford Press, Caldwell.

[13] Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research*, Vol. III: The Design and Analysis of Long-term Animal Experiments. International Agency for Research on Cancer, Lyon.

[14] Gaylor, D.W. (1994). Dose-response modeling, in *Developmental Toxicology*, 2nd Ed. C.A. Kimmel & J. Buelke-Sam, eds. Raven Press, New York, pp. 363–375.

[15] Haseman, J.K. (1990). Use of statistical decision rules for evaluating laboratory animal carcinogenicity studies, *Fundamental and Applied Toxicology* **14**, 637–648.

[16] Haseman, J.K. (1992). Value of historical controls in the interpretation of rodent tumor data, *Drug Information Journal* **26**, 191–200.

[17] Haseman, J.K. (1995). Data analysis-statistical analysis and use of historical control data, *Regulatory Toxicology and Pharmacology* **21**, 52–59.

[18] Haseman, J.K. & Johnson, F.M. (1996). Analysis of National Toxicology Program rodent bioassay data for anticarcinogenic effects, *Mutation Research* **350**, 131–141.

[19] Haseman, J.K. & Piegorsch, W.W. (1994). Statistical analysis of developmental toxicity data, in *Developmental Toxicology*, 2nd Ed. C.A. Kimmel & J. Buelke-Sam, eds. Raven Press, New York, pp. 349–361.

[20] International Life Sciences Institute (ILSI) (1984). The selection of doses in chronic toxicity/carcinogenicity studies. in *Current Issues in Toxicology*, H.C. Grice, ed. Springer-Verlag, New York, pp. 9–49.

[21] Jordan, V.C. (1976). Effect of tamoxifen (ICI 46474) on initiation and growth of DMBA-induced rat mammary carcinomata, *European Journal of Cancer* **12**, 419–424.

[22] King, E.P. (1963). A statistical design for drug screening, *Biometrics* **19**, 429–440.

[23] Legler, J.M., Lefkopoulou, M. & Ryan, L.M. (1995). Efficiency and power of tests for multiple binary outcomes, *Journal of the American Statistical Association* **90**, 680–693.

[24] Manson, J.M. & Kang, Y.J. (1989). Test methods for assessing female reproductive and developmental toxicology, in *Principles and Methods of Toxicology*, A.W. Hayes, ed. Raven Press, New York, pp. 311–360.

[25] Perkins, M.A., Osborne, R. & Johnson, G.R. (1996). Development of an in vitro method for skin corrosion testing, *Fundamental and Applied Toxicology* **31**, 9–18.

[26] Portier, C. & Hoel, D. (1983). Low-dose rate extrapolation using the multistage model, *Biometrics* **39**, 897–906.

[27] Portier, C.J. & Kopp-Schneider, A. (1991). A multistage model of carcinogenesis incorporating DNA damage and repair, *Risk Analysis* **11**, 535–543.

[28] Richards, A.M. (1994). Application of SAR methods to non-congeneric data bases associated with carcinogenicity and mutagenicity: issues and approaches, *Mutation Research* **305**, 73–97.

[29] Roseberry, T.D. & Gehan, E.A. (1964). Operating characteristic curves and accept-reject rules for two and three stage screening procedures, *Biometrics* **20**, 73–84.

[30] Rosenkranz, H.S. & Klopman, G. (1990). New structural concepts for predicting carcinogenicity in rodents: an artificial intelligence approach, *Teratogenesis, Carcinogenesis, and Mutagenesis* **10**, 73–88.

[31] Seilkop, S.K. (1995). The effect of body weight on tumor incidence and carcinogenicity testing in B6C3F1 mice and F344 rats, *Fundamental and Applied Toxicology* **24**, 247–259.

[32] Stokes, W.S. (1994). Alternative test method development at the National Toxicology Program, in *Proceedings of the Toxicology Forum Winter Meeting*. Toxicology Forum, Washington, pp. 302–313.

[33] Tennant, R.W., French, J.E. & Spalding, J.W. (1995). Identifying chemical carcinogens and assessing potential risk in short-term bioassays using transgenic mouse models, *Environmental Health Perspectives* **103**, 942–950.

[34] Tennant, R.W., Spalding, J., Stasiewicz, S. & Ashby, J. (1990). Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by the National Toxicology Program, *Mutagenesis* **5**, 3–14.

[35] Wachsman, J.T., Bristol, D.W., Spalding, J., Shelby, M. & Tennant, R.W. (1993). Predicting chemical carcinogenesis in rodents, *Environmental Health Perspectives* **101**, 444–445.

[36] Westfall, P.H. & Soper, K.A. (1998). Weighted multiplicity adjustments for animal carcinogenicity test, *Journal of Biopharmaceutical Statistics* **8**(1), 23–44.

[37] Whittle, E. & Basketter, D.A. (1993). The in vitro skin corrosivity test: comparison of in vitro human skin with in vivo data, *Toxicology In Vitro* **7**, 265–268.

(*See also* **Serial-sacrifice Experiments**)

JOSEPH K. HASEMAN

Anthropometry

That the ancient world was involved in measuring the human body is not in doubt, although little has survived in the form of data. The Egyptians measured the stature of their kings, the Chinese measured height and head circumference, and in Mesopotamia they could measure weight on a balance. Sculptors have always necessarily been aware of body proportions; Greek sculptors worked to a ratio of one full male height to $7\frac{1}{2}$ head heights in general and to eight head heights on occasion, the justification being that the former reflects the true physical proportion but the latter the proportion as it appears from eye level. Aristotle noted that the human body gains in length between birth and five years about the same amount as it gains during the rest of its life, an approximation as valid today as it was in 300 BC.

The Renaissance rediscovered the glories of ancient Greece, so it is perhaps not surprising that although Leon Battista Alberti constructed an instrument for measuring the human body around 1450, its sole purpose was the making of better statues.

The term Anthropometry can be ascribed to Johann Sigismund Elsholtz (1623–1688), a German physician, who entitled his graduation thesis *Anthropometria*. He described an anthropometer for measuring heights and parts of heights of the human body. A 72 in. man should ideally measure 28 in. down to the navel, 36 in. to the pubic bone, etc. and Elsholtz's interest lay in discerning whether divergences from these proportions could be ascribed to various diseases.

The eighteenth century witnessed a surge of interest in human measurement, led mainly by the military, who preferred tall soldiers to short ones. But scientific interest also flourished, and among those active was George LeClerc, Count of Buffon (1707–1788), of Buffon's needle (*see* **Stereology**), who measured the length and weight of a number of fetuses and newborns. J.G. Roederer (1726–1763) also measured the length and weight of newborn infants but with a view to detecting immaturity in the baby, an early example of the use of measurement as a diagnostic tool. A particular high-water mark was provided by Buffon's friend Count Philibert Guéneau de Montbeillard (1720–1785) who measured his son's height some 38 times between birth and age 19 years and so provided the first known longitudinal study of human height. Graphs, based on these data, of height and height velocity – the latter clearly showing the pubertal growth spurt – can be seen in Tanner [20], who provides a comprehensive history of human growth.

Lambert Adolphe Jacques Quetelet (1796–1874) also made longitudinal measurements of his son and daughter and two daughters of a friend. Furthermore, by analyzing very large samples of measurements made on the military, he concluded that height, as well as other dimensions such as chest circumference, were **normally distributed** (Gaussian) in a population. An unfortunate consequence of this was that he concocted the idea of "l'homme moyen" – the average man – who represented the *type* or standard in the population although, as others have pointed out, such a person was thereby rather dull. Quetelet constructed tables of **mean** heights and weights for children based on **cross-sectional** data although the sample size at each age was rather small. Moreover, his interest in shape led to his definition of the Quetelet index, now the body mass index (see below).

The *cephalic index* (head width to length ratio) was an early nineteenth-century invention for use by anthropologists in characterizing ethnic and racial groups.

Sir Francis Galton (1822–1911) collected measurements of height, weight, and the circumferences of chest, upper arm, and head in his study of family likeness and inherited characteristics that led to the concept of **regression**. However, an important impetus to measuring children in the nineteenth century was social concern about the effects that dismal living and working conditions were having on the health of the poor, and much of the work, particularly around the 1830s, was done in association with the various Factory Acts. Similar investigations at this time were carried out in France. Later in the century the British Association for the Advancement of Science set up a number of Anthropometric Committees to collect data on heights, weights, etc. and height surveys, particularly in schools and colleges, were carried out in Boston, USA, and Italy.

At the beginning of the twentieth century, longitudinal studies, often measuring many physical variables, were set up in several centers in Europe and America. The twentieth century later saw the clinical measuring of several dimensions of the human body for diagnostic purposes while, in quite different

areas, measurements were required by the clothing industry for the making of garments and in the field of ergonomics for the design of tools and equipment in the workplace.

Direct Measures

Height (Length)

Height is measured by a stadiometer, on which a horizontal board is brought into contact with the top of the head and its distance then measured from the footplate. The Harpenden stadiometer contains a dial that displays the height and needs calibrating at regular intervals. A number of inexpensive plastic instruments are now available with built-in scales. Some of these need careful positioning and calibration. During measurement, the subject's head should be placed in the Frankfurt plane; that is, with an imaginary line drawn from the center of the ear hole to the lower border of the eye socket set horizontally. A number of experiments have been carried out with these instruments, the conclusions being that they are about equally reliable. In particular Voss et al. [24], after measuring children and wooden poles, concluded that a child contributes at least 90% of the error **variance** when using one of these instruments, the remainder originating with the measurer and the instrument. They also concluded, in agreement with other workers, that the **standard deviation** of a single height measurement is about 0.25 cm. This is an average value over children, the standard deviation varying from child to child, with some children being almost static in their posture while others are rather springy. Subsequent experiments suggest that this value also obtains for adults.

Up to age 2 years, it is usual to measure supine length, rather than height, on a calibrated measuring mat. The standard deviation of a single measurement is estimated to be about 0.3 cm in newborn infants and about 0.4 cm at 6 weeks and 8 months [7].

There are problems in trying to measure a subject more than once and then averaging the observations to reduce the effect of **measurement error**. Several minutes of activity need to elapse between measurements for them to be independent, and the scale on the instrument needs to be read by an independent observer to avoid the very strong **bias** that can result when a measurer remembers a previous reading.

In 1724 the Royal Society of London was told of the Reverend Joseph Wasse's experiment with a nail in the wall that he could touch in the morning but not later in the day owing to diurnal shrinkage of the human body (*see* **Circadian Variation**). This has been investigated since by many researchers [17], who have found that daily shrinkage can be greater than 2 cm, and has led to a stretching method of measuring height that involves gentle upward pressure on the mastoid processes. This reduces only partly the effects of diurnal shrinkage and the variable posture of the subject. However, as shown in [24], different measurers stretch by different amounts so that, in clinical practice, subjects should be measured on repeated visits to the clinic by the same measurer and at the same time of day. Seasonal variation in growth and catch-up growth after illness are also well known. Hall [12] discusses clinical considerations in measuring height (as well as weight and head circumference).

The end of the nineteenth century saw the first attempts at producing centile charts (*see* **Quantiles**) for height based on cross-sectional data consisting of a number of heights made at each age. Their purpose was merely to describe the population, but later such charts were used for clinical diagnostic purposes or for **screening**. Any child with height below some threshold level, say the 3rd centile, can be referred and investigated further for possible pathology. To this end the charts of Tanner et al. [23], modified by Tanner & Whitehouse [22], were published and served as the standard in the UK for the next 30 years. They give smoothed 3rd, 10th, 25th, 50th, 75th, 90th, and 97th height (supine length to age 2) centiles from birth to age about 19 years for boys and girls separately. The corresponding charts for the US standards for height were given by Tanner & Davies [21]. Standards and charts also exist for several other countries, while much of the developing world uses charts prepared by the **World Health Organization** (**WHO**), although these are based on US data.

Charts of height velocity are also given in [21–23]. If H_1 is the height of a subject at time 1 and H_2 the height at a later time 2, then $D = H_2 - H_1$ is the incremental change in height which, when measured over the interval of a year, is known as the *height velocity*. This can be used for diagnostic purposes, in particular to detect a child who, although at an acceptable height, has a growth rate that starts

to fall. It is, however, subject to a larger measurement error than height.

Height, but not height velocity, standards for the UK have now been updated and the corresponding charts published by the Child Growth Foundation (CGF), but see Freeman et al. [10] for a full description. As explained by Cole [4], nine centiles are shown, against the original seven, these being spaced two-thirds of a standard deviation score (see below) apart. Data from seven sources were used in the construction of the charts, which were plotted by means of the LMS method and **penalized maximum likelihood** of Cole & Green [5]. In this, if $H(t)$ is the height, say, of a male at t years, $M(t)$ is its **median**, and $S(t)$ its coefficient of variation (*see* **Standard Deviation**), then $L(t)$ can be found from a Box–Cox **power transformation** of $H(t)$ so that

$$Z = \frac{\left[\frac{H(t)}{M(t)}\right]^{L(t)} - 1}{L(t)S(t)}$$

has approximately a standard normal distribution. Once estimates of $L(t)$, $M(t)$, and $S(t)$ have been obtained, the $C_{100\alpha}(t)$ centile of the height distribution at age t is given by

$$C_{100\alpha}(t) = M(t)[1 + L(t)S(t)z_\alpha]^{1/L(t)},$$

where z_α is the **standard normal deviate** corresponding to the centile ($z_\alpha = 1.341$ for the 91st centile, etc.). In fact, $L(t)$ was not found to differ significantly from 1 at any age and so was set uniformly to 1.

It has been suggested that when a child is measured on two separate occasions, the height observed on the second occasion should be assessed conditionally on the height obtained on the first. If height H, at a particular age, has mean μ and standard deviation σ, and h is an observed value of H, then $h^{\mathrm{SDS}} = (h - \mu)/\sigma$ is known as the *standard deviation score* (SDS) of h. If H is also normally distributed, then the centile of h^{SDS} can be determined from the standard normal distribution. Moreover, the conditional SDS of h_2 given h_1 can be found from the relation

$$(h_2|h_1)^{\mathrm{SDS}} = \frac{(h_2^{\mathrm{SDS}} - \rho h_1^{\mathrm{SDS}})}{(1 - \rho^2)^{1/2}},$$

where ρ is the coefficient of **correlation** between h_1 and h_2. This is well over 0.9 in the years approaching puberty [1].

Note that if $D = H_2 - H_1$ is a height velocity, then the observed velocity conditional on h_1 is given by $d|h_1 = h_2|h_1 - h_1$, so that **inferences** based on conditional height at time 2 and conditional velocity up to time 2 are equivalent.

Weight

Much of what has been written about height also applies to weight. Modern measuring instruments are highly reproducible with repeated observations differing by only a few grams. The main source of variability lies in the subject but is not so short term, being dependent on the contents of the stomach, bowel, and bladder. Moreover, the weight distribution in a population at any age is not generally normally distributed, although log (weight) is often sufficiently nearly normal for practical purposes (*see* **Lognormal Distribution**). The charts of Tanner & Whitehouse [22] have provided the weight and weight velocity standards for the UK since the 1960s although the weight standards have now been updated by Freeman et al. [10] and the corresponding charts published by the CGF. The LMS method was again used, but the values of $L(t)$ were found significantly different from 1. The charts of the US weight standards can be found in Hamill et al. [14].

The human body often is regarded as made up of two components, lean body mass and fat. In view of the latter, body weight can decrease as well as increase with time. The most commonly used method for estimating the weights of these two components involves taking four skinfold measurements with calipers at four recognized sites, namely the triceps, biceps, subscapular, and supra-iliac. Body density can then be estimated by a regression equation of the form

$$\text{density} = a + b \log_{10} S,$$

where S is the sum of the four skinfolds in millimeters and a and b are constants that depend on the sex and age of the subject. Body fat percentage can next be estimated by Siri's formula,

$$\text{body fat percentage of total weight} = 100\left(\frac{4.95}{\text{density}} - 4.5\right),$$

which is based on the assumption that the density of fat is 0.9 kg/l and of fat-free mass 1.1 kg/l.

Finally, body fat weight and, by subtraction from total body weight, the weight of lean body mass, can be calculated. Values of the regression coefficients for adults can be found in Durnin & Wormersley [9].

The situation with children is, however, more complicated, for a number of reasons, including the change in the density of fat-free mass with age. In a **validation study** of several suggested formulas, Reilly et al. [18] consider them all sufficiently unreliable as to state, "For the time being skinfolds might best be regarded as indices (rather than measures) of body fatness in individuals, or means of estimating body fatness of groups."

In hospitals and laboratories with suitable equipment, body density can be evaluated more directly by measuring the weight in air followed by the weight in water to estimate the volume; or the volume may be estimated from the amount of water displaced. In either case, since the aim is to measure the volume of body tissues alone, a correction needs to be made to remove residual lung volume from the body volume.

More direct methods of measuring lean body mass, such as potassium emission counting and dual energy X-ray absorptiometry (DXA), are available and have good reproducibility. See Jebb & Elia [16] for a review of methods.

Other Dimensions

Head circumference is measured in the neonatal period to detect diseases associated with a large head, such as hydrocephalus, or a small head. It is measured with a plastic or metal tape along a line midway between the eyebrows and the hairline at the front of the head and the occipital prominence at the back. The main source of error lies therefore in determining this line. Height (length), weight, and head circumference are the only dimensions measured in every child in the UK, and Hall [12] recommends that this last should be measured before discharge from hospital following birth and again at approximately 6–8 weeks of age. Charts of population standards are usually available for the early years of life only; the charts for the UK, published by the CGF, cover just the first year while those for the US, given in [14], cover the first three years.

The *sitting height* (SH) of a subject is measured on a table with a stadiometer attached, the height being the distance from the crown of the head to the table. Unpublished research suggests that the measurement error involved is larger than that for height in view of the added difficulty in placing the subject correctly. Sitting height subtracted from standing height leads to the *subischial leg length* (SLL).

A comparison of SH with SLL can aid the diagnosis of skeletal dysplasias. It is also known that SH may be disproportionately long relative to SLL in precocious puberty. UK charts date from 1978 and are available. They are, however, based on small numbers of children from several years ago and so should be treated with caution when used today.

In view of the inherent variability of a height measurement, clinical practitioners have sought other, less variable, dimensions that might prove useful in monitoring growth, particularly over a very short time period. These include the lengths of individual bones made from radiographs and the *lower leg length* measured on a knemometer, an instrument not unlike a stadiometer but with the headboard placed on the knee with the subject in the sitting position. Since this variable is essentially the sum of a few bone dimensions, it does indeed display a small variance. Insufficient correlation with height makes it unsuitable for making judgments about height. However, it can be used as an early indicator that growth hormone is being successful in the treatment of short stature [25].

Standards and charts have been compiled for a large number of other measures that are useful in diagnosing various syndromes. Many of these can be found in Hall et al. [13].

Derived Measures

Surface Area

Although some earlier attempts at measuring the surface area (SA) of the human body are recorded, the main work began in about 1850. Several methods were devised, including coating the body with paper or adhesive type, surface integration using inked discs, and triangulation. Interest lay initially in determining the number of pores on the skin or the total force exerted on the body by the atmosphere. Over 1000 cases were reported, and these are comprehensively discussed in [3].

The need for such estimation today arises from the fact that both renal function and dosage of cancer chemotherapy are expressed in relation to surface area rather than weight. Clearly the measurement of

SA by any direct method is impossible in routine clinical work. However, even from an early period, estimating formulas were devised based on height H, and weight W. Many of these take the form

$$\mathrm{SA} = cH^{a_1}W^{a_2},$$

where a_1, a_2, and c are constants. One of the first examples is due to DuBois & DuBois [8] and sets $a_1 = 0.725$ and $a_2 = 0.425$ if H is measured in centimeters, W in kilograms, and SA in square meters. This was based on a sample of only nine subjects, including a male cretin and a female cadaver, and is still in common use today!

Gehan & George [11] carried out a bivariate regression analysis on the log of the above model using as data the 401 postnatal cases given in [3] for which measured values are listed for all of H, W, and SA. This finds $a_1 = 0.422, a_2 = 0.515$, and $c = 0.0235$, values recommended by Bailey & Briars [2], who extend their analysis to provide **standard errors** and explain the equivalence of a number of suggested models of the above form.

Body Mass Index

A number of weight for height indices have been suggested for measuring the fatness of an individual, but the one that has established itself as the standard in adults is Quetelet's *body mass index* (BMI) defined by W/H^2, where weight W is measured in kilograms and height H in metres. An individual with a value less than 20 is generally regarded as underweight, while one with a value over 25 is overweight. This latter range is sometimes subdivided into three (25–30, 30–40, and >40) that define three grades of obesity.

Despite its routine clinical use in adults, the application of BMI to children has always been suspect since its distribution in the population changes quickly with age up to, and even beyond, adolescence and maturity. To overcome this problem, Rosenthal et al. [19] have shown that, for healthy London school children between the ages of about $4\frac{1}{2}$ years and 19, the index $W/H^{2.88}$ is independent of height or age, and they have provided centile charts and nomograms for its use. On the other hand, Cole et al. [6] have given centile charts of BMI for UK children from birth to age 23 years on the basis of data used in updating standards for height and weight. Centile charts for white US children aged 1–19 years are to be found in Hammer et al. [15].

References

[1] Bailey, B.J.R. (1994). Monitoring the heights of prepubertal children, *Annals of Human Biology* **21**, 1–11.

[2] Bailey, B.J.R. & Briars, G.L. (1996). Estimating the surface area of the human body, *Statistics in Medicine* **15**, 1325–1332.

[3] Boyd, E. (1935). *The Growth of the Surface Area of the Human Body*. The University of Minnesota Press, Minneapolis.

[4] Cole, T.J. (1994). Do growth chart centiles need a face lift?, *British Medical Journal* **308**, 641–642.

[5] Cole, T.J. & Green, P.J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood, *Statistics in Medicine* **11**, 1305–1319.

[6] Cole, T.J., Freeman, J.V. & Preece, M.A. (1995). Body mass index reference curves for the UK, 1990, *Archives of Disease in Childhood* **73**, 25–29.

[7] Doull, I.J.M., McCaughey, E.S., Bailey, B.J.R. & Betts, P.R. (1995). Reliability of infant length measurement, *Archives of Disease in Childhood* **72**, 520–521.

[8] DuBois, D. & DuBois, E.F. (1916). A formula to estimate the approximate surface area if height and weight be known, *Archives of Internal Medicine* **17**, 863–871.

[9] Durnin, J.V.G.A. & Wormersley, J. (1974). Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged 16–72 years, *British Journal of Nutrition* **32**, 77–97.

[10] Freeman, J.V., Cole, T.J., Chinn, S., Jones, P.R.M., White, E.M. & Preece, M.A. (1995). Cross sectional stature and weight reference curves for the UK, 1990, *Archives of Disease in Childhood* **73**, 17–24.

[11] Gehan, E.A. & George, S.L. (1970). Estimation of human body surface area from height and weight, *Cancer Chemotherapy Reports* **54**, 225–235.

[12] Hall, D.M.B., ed. (1996). *Health for All Children*, 3rd Ed. Oxford University Press, Oxford.

[13] Hall, J.G., Froster-Iskenius, U.G. & Allanson, J.E. (1989). *Handbook of Normal Physical Measurements*. Oxford University Press, Oxford.

[14] Hamill, P.V.V., Drizd, T.A., Johnson, C.L., Reed, R.B., Roche, A.F. & Moore, W.M. (1979). Physical growth: National Center for Health Statistics percentiles, *American Journal of Clinical Nutrition* **32**, 607–629.

[15] Hammer, L.D., Kraemer, H.C., Wilson, D.M., Ritter, P.L. & Dornbusch, S.M. (1991). Standardized percentile curves of body mass index for children and adolescents, *American Journal of Diseases of Children* **145**, 259–263.

[16] Jebb, S.A. & Elia, M. (1993). Techniques for the measurement of body composition: a practical guide, *International Journal of Obesity* **17**, 611–621.

[17] Lampl, M. (1992). Further observations on diurnal variation in standing height, *Annals of Human Biology* **19**, 87–90.

[18] Reilly, J.J., Wilson, J. & Durnin, J.V.G.A. (1995). Determination of body composition from skinfold thickness: a validation study, *Archives of Disease in Childhood* **73**, 305–310.

[19] Rosenthal, M., Bain, S.H., Bush, A. & Warner, J.O. (1994). Weight/height$^{2.88}$ as a screening test for obesity or thinness in schoolage children, *European Journal of Pediatrics* **153**, 876–883.

[20] Tanner, J.M. (1981). *A History of the Study of Human Growth*. Cambridge University Press, Cambridge.

[21] Tanner, J.M. & Davies, P.S.W. (1985). Clinical longitudinal standards for height and height velocity for North American children, *Journal of Pediatrics* **107**, 317–329.

[22] Tanner, J.M. & Whitehouse, R.H. (1976). Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty, *Archives of Disease in Childhood* **51**, 170–179.

[23] Tanner, J.M., Whitehouse, R.H. & Takaisha, M. (1966). Standards from birth for height, weight, height velocity, and weight velocity: British children, 1965, *Archives of Disease in Childhood* **41**, 454–471, 613–635.

[24] Voss, L.D., Bailey, B.J.R., Cumming, K., Wilkin, T.J. & Betts, P.R. (1990). The reliability of height measurement, *Archives of Disease in Childhood* **65**, 1340–1344.

[25] Wales, J.K.H. & Milner, R.D.G. (1987). Knemometry in assessment of linear growth, *Archives of Disease in Childhood* **62**, 166–171.

(*See also* **Growth and Development**)

B.J.R. BAILEY

Anticipation

Anticipation denotes a phenomenon characterized by increasing severity and/or earlier onset of a disease in successive generations; some definitions also include greater recurrence risk (*see* **Genetic Counseling**) among the possible factors delineating a progressive degeneration from parent to child across the generations. Observations of apparent anticipation in myotonic dystrophy date from the early twentieth century, and there are indications that the first observations of the phenomenon may be traceable to the nineteenth century [4].

In 1948, Penrose [6] noted that the clinical impression of anticipation could be an artifact induced by **ascertainment** bias, and cited ways in which preferential ascertainment of certain types of pedigrees could result in the appearance of anticipation in hereditary diseases. These included preferential ascertainment of late onset parents, who may have had an enhanced likelihood of reproduction relative to persons with earlier onset, and of children with very early onset, who might be more likely to attract medical attention. Both would tend to create the appearance of earlier onset in the offspring generation. Another possible source of **bias** is the simultaneous identification of parent–offspring affected pairs, in which the parent would have had a notably longer time of exposure to risk of onset, resulting again in a tendency toward the observation of later onset ages in the parental generation. Such rationales cast serious doubt on the reality of anticipation as a biologically based phenomenon in the minds of many geneticists.

However, in the late twentieth century a new type of dynamic mutation was linked to the clinical observation of anticipation in at least some disorders. In this type of mutation, called trinucleotide repeat expansion, the number of triplet repeats can increase with the transmission of genetic material from one generation to the next (*see* **DNA Sequences**; **Genetic Markers**). An increased number of such repeats at a particular genomic site within an individual is associated with disease for a set of disorders including myotonic dystrophy, Friedrich's ataxia, Huntington disease, and Fragile X syndrome. Moreover, the number of repeats has been found to correlate with severity and onset age, providing a molecular mechanism explaining the phenomenon of anticipation in these and other hereditary diseases [7, 8]. A variety of models have been put forward to explain how such expansions disrupt gene expression, and there is evidence that the specific mechanism may vary with the particular syndrome [4]. Because the likelihood of such expansion can differ with parental gender, this type of dynamic mutation also provides an explanation for the phenomenon of parent-of-origin effect often associated with disorders displaying anticipation (*see* **Parental Effects**).

Anticipation has been reported for a growing number of disorders not found to be associated with trinucleotide repeat expansion, and certainly the existence of unstable deoxyribonucleic acid (DNA) disorders does not preclude the possibility of other mechanisms leading to anticipation [5, 7]. Furthermore, the potential for ascertainment-related artifact should not be discounted. Since large, population-based prospective studies are rare, there is persistent interest in statistical methods for the detection of anticipation that are able to appropriately adjust for ascertainment bias. A number of such approaches have focused on correction for bias induced by differential age at interview for pedigree members of different generations. Such circumstances result in a tendency for parents to have longer periods of exposure to risk than their offspring [1, 3, 9]. However, even methods that successfully adjust for such truncation effects may be adversely affected by other commonly encountered sources of ascertainment bias, including problems arising from decreased fertility among affected persons and preferential selection of pedigrees with multiple affected individuals [3, 9]. **Simulation** studies have identified difficulties associated with a number of available methods, including inflation of type I error rates, and deficiencies of power [1, 3, 9], but also point toward promising new designs, such as those utilizing collateral relatives [2], for the statistical evaluation of anticipation.

References

[1] Heiman, G.A., Hodge, S.E., Wickramaratne, P. & Hsu, H. (1996). Age-at-interview bias in anticipation studies: computer simulations and an example with panic disorder, *Psychiatric Genetics* **6**, 61–66.

[2] Hoh, J., Heitjan, D.F., Merette, C. & Ott, J. (2001). Ascertainment and anticipation in family studies, *Human Heredity* **51**, 23–26.

[3] Huang, J. & Vieland, V. (1997). A new statistical test for age-of-onset anticipation: application to bipolar disorder, *Genetic Epidemiology* **14**, 1091–1096.

[4] McGinnis, M.G. (1996). Anticipation: an old idea in new genes, *American Journal of Human Genetics* **59**, 973–979.

[5] Paterson, A.D., Naimark, D.M.J., Vincent, J.B., Kennedy, J.L. & Petronis, A. (1998). netic anticipation in neurological and other disorders, in *Genetic Instabilities and Hereditary Neurological Diseases*, R.D. Wells, S.T. Warren & M. Sarmiento, eds. Academic Press, New York, pp. 413–427.

[6] Penrose, L.S. (1948). The problem of anticipation in pedigrees of dystrophia myotonica, *Annals of Eugenics* **14**, 125–132.

[7] Petronis, A. & Kennedy, J.L. (1995). Unstable genes – unstable mind?, *American Journal of Psychiatry* **152**, 164–172.

[8] Sherman, S.L. (1997). Evolving methods in genetic epidemiology. IV. Approaches to non-Mendelian inheritance, *Epidemiologic Reviews* **19**, 44–51.

[9] Vieland, V.J. & Huang, J. (1998). Statistical evaluation of age-at-onset anticipation: a new test and evaluation of its behavior in realistic applications, *American Journal of Human Genetics* **62**, 1212–1227.

DEBORAH V. DAWSON

Antidependence Models

One important approach to the analysis of continuous repeated measurements data uses a *linear regression model* to relate the *expected value*, or mean, of each subject's sequence of response measurements to covariates such as treatment group and time of measurement. The sequence of observations from a subject will deviate in some random way from the postulated regression model. With repeated measurements it cannot typically be assumed that these deviations, or residuals, are independent and equally variable – the assumptions that underly conventional multiple regression analysis. The pattern of variances and correlations among the residuals from a subject is called the *covariance structure* of the repeated measurements and a full analysis requires this to be specified as part of the model. The ante-dependence model defines a family of covariance structures for repeated measurements.

We assume that we have a repeated measurements trial or experiment in which all subjects are observed at the same times relative to the start. There may be missing values. Many models of covariance structure have been proposed for this setting, some, such as the *compound-symmetry* (or uniform) and *stationary first-order autoregressive* models, depending on a small number of parameters (*see* **Analysis of Variance for Longitudinal Data**; **ARMA and ARIMA Models**; **Time Series**). Such simple models, particularly those with variances that are constant across time, are often too restrictive to provide an adequate representation of the variances and correlations observed in practice with repeated measurements. At the other extreme we have the most general model: the so-called *unstructured* covariance matrix. This allows a different variance for each time of measurement and a different correlation for every pair of time points. With T times of measurement, the unstructured matrix has $T(T+1)/2$ parameters and this figure grows at the rate of the *square* of the number of times. While, by definition, the unstructured covariance matrix must provide an adequate model for the observed structure, the large number of parameters means that, when sample sizes are not large, the subsequent analysis can be inefficient. The class of ante-dependence models provides structures that are intermediate in complexity between the very simple models with a fixed number of parameters and the unstructured model. Typically, the number of parameters in an ante-dependence model grows at the same rate as the number of time points. The class includes the unstructured matrix as a least efficient special case.

The class of ante-dependence covariance structures can be defined in several ways. The definition, in terms of *conditional independence*, illustrates well the rationale behind the structure and provides a clear link with the important topic of graphical models [7]. Following Gabriel [1, 2] and Kenward [4] we say that the sequence of random variables $Y_1, \ldots, Y_T$ has an ante-dependence structure of order r (AD(r)) if, for every $t > r$,

$$Y_t | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-r}$$

is conditionally independent of $Y_{t-r-1}, \ldots, Y_1$. In other words, once we have taken account of the r observations preceding Y_t, the remaining preceding observations carry no additional information about Y_t. Typically, we would expect r to be small. This is plausible in many settings, but there are important exceptions; for example, where changes over time are nonlinear and times of rapid change vary at random among subjects.

If the $\{Y_t\}$ follow a Gaussian distribution, then the assumption of an AD(r) structure is equivalent to the assumption that the *inverse* of the covariance matrix has a band form of order $r+1$; that is, all elements of the inverse are zero except for the leading diagonal and the r diagonals immediately above and below it. This establishes an important distinction between the ante-dependence structure and the majority of models for repeated measurements covariances structures: the latter typically impose structure on the covariance matrix itself rather than the inverse. An important special case is the AD(1) structure which corresponds to a tri-diagonal inverse: this is a well-known property of processes with a first-order *Markov* structure, to which the AD(1) structure corresponds.

The AD(r) structure imposes no constraints on the constancy of variance or covariance with respect to time; that is, in terms of second-order moments, it is not *stationary*. The familiar stationary autoregressive [AR(r)] covariance structures can be expressed as special cases of the AD(r) structures. This lack of stationarity means that the structure is particularly well suited to situations where the variability fluctuates during the course of a trial and where intervals between times of measurement are not

equal. Additional smoothness can be introduced by imposing a polynomial form on the variances and auto-regressions as in the so-called "structed antedependence" models [5].

The AD(r) structure can be combined with any appropriate linear model for the mean profiles and fitted to data using conventional likelihood and restricted likelihood methods as described, for example, in [3]. The resulting inferences are based on asymptotic results.

The AD structures have an additional advantage when used in the special (and restricted) setting in which the saturated means model is used; that is, in which a different parameter is used for each combination of between-subject covariate and time of measurement. Such models are implicit in repeated measurements analysis of variance. Under the AD structure, pivotal **likelihood ratio test** statistics can be constructed for overall profile comparisons for any between-subject comparison and, when applied to successive differences among the repeated measurements, for interactions of such comparisons with time. The test statistics can be calculated in a simple way from univariate analyses of covariance and have known finite sample distributions. They can be regarded as generalizations of Wilks' lambda statistic from repeated measurements **multivariate analysis of variance**. Further, the same construction produces pivotal likelihood ratio statistics with known distributions even when there are drop-outs; that is, when some sequences terminate early. Full details are given in [4]. In this way, the AD structure leads to analyses that provide a practical likelihood-based alternative to simple analysis-of-variance-based methods for analyzing repeated measurements with drop-out. The validity of the AD-based analyses rests on the assumption that the drop-out process is random [6], while *ad hoc* methods such as modified repeated measurements analysis of variance require the stricter assumption of completely random drop-out.

References

[1] Gabriel, K.R. (1961). The model of ante-dependence for data of biological growth, *Bulletin de l'Institut International Statistique* **39**, 253–264.

[2] Gabriel, K.R. (1962). Ante-dependence analysis of an ordered set of variables, *Annals of Mathematical Statistics* **33**, 201–212.

[3] Jennrich, R.I. & Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics* **42**, 805–820.

[4] Kenward, M.G. (1987). A method for comparing profiles of repeated measurements, *Applied Statistics* **36**, 296–308.

[5] Nunez-Anton, V. & Zimmerman, D.L. (2000). Modelling non-stationary longitudinal data, *Biometrics*, **56**, 699–705.

[6] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

[7] Whittaker, J.C. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

(*See also* **Longitudinal Data Analysis, Overview**)

M.G. KENWARD

Antithetic Variable

The efficiency of a **simulation** can be increased by the judicious use of variance-reduction methods, of which antithetic variables (or variates) is one example, due originally to Hammersley & Morton [2].

The basic idea is simply illustrated in the context of integral estimation (*see* **Numerical Integration**). Consider the following example, taken from Morgan [4].

The integral,

$$I = \int_0^1 (1 - x^2)^{1/2} \, \mathrm{d}x, \tag{1}$$

is known to take the value $\pi/4$.

If a random sample, $U_1, \ldots, U_n$, is taken from a **uniform distribution** on (0,1), then a crude **Monte Carlo** approach estimates I by

$$I_c = \frac{1}{n} \sum_{i=1}^{n} (1 - U_i^2)^{1/2}. \tag{2}$$

This is because the integral in (1) can be interpreted as the expectation, $\mathrm{E}[(1 - U^2)^{1/2}]$, where U has the uniform, U(0, 1) distribution. The estimate, I_c, is then seen as the sample average, providing an unbiased estimate of I. It is straightforward to show that

$$\mathrm{var}(I_c) \approx \frac{0.0498}{n}.$$

The method of antithetic variates estimates I by

$$I_A = \frac{1}{2n} \sum_{i=1}^{n} \left\{ (1 - U_i^2)^{1/2} + [1 - (1 - U_i)^2]^{1/2} \right\}. \tag{3}$$

Clearly, since $1 - U$ also has a U(0, 1) distribution, $\mathrm{E}[I_A] = I$, but the negative correlation between U and $1 - U$ has a variance-reducing effect. If we are trying to estimate π by these approaches, then the method of crude Monte Carlo requires a random sample from the U(0, 1) distribution that is slightly more than nine times larger than that required by the approach based on (3) to achieve the same variance.

An attractive practical demonstration of the value of using antithetic variates is again in terms of estimation π, but through Buffon's cross replacing Buffon's needle (*see* **Stereology**); see Hammersley & Morton [2] and Morgan [4]. In simple computer simulations, antithetic variates can be readily obtained if the inversion method is employed to transform **pseudo-random** U(0, 1) variates to provide realizations of random variables from other desired distributions; see Ripley [6]. For examples and variations on this approach in the context of simulating queueing systems, see Page [5] and Mitchell [3]. The paper by Schruben & Margolin [7] provides an application in a medical context, in which patients with heart problems queue for beds in a hospital unit. The antithetic variate approach has wide application; see, for example, Green & Han [1].

References

[1] Green, P.J. & Han, X.-L. (1992). Metropolis methods, Gaussian proposals, and antithetic variables, in *Stochastic Models, Statistical Methods and Algorithms in Image Analysis*, P. Barone, A. Frigessi & M. Piccioni, eds, *Lecture Notes in Statistics*, Vol. 74. Springer-Verlag, Berlin, pp. 142–164.

[2] Hammersley, J.M. & Morton, K.W. (1956). A new Monte Carlo technique: antithetic variates, *Proceedings of the Cambridge Philosophical Society* **52**, 449–475.

[3] Mitchell, B. (1973). Variance reduction by antithetic variates in GI/G/1 queueing simulations, *Operations Research* **21**, 988–997.

[4] Morgan, B.J.T. (1984). *Elements of Simulation*. Chapman & Hall, London.

[5] Page, E.S. (1965). On Monte Carlo methods in congestion problems II – simulation of queueing systems, *Operations Research* **13**, 300–305.

[6] Ripley, B.D. (1987). *Stochastic Simulation*. Wiley, New York.

[7] Schruben, L.W. & Margolin, B.H. (1978). Pseudo-random number assignment in statistically designed simulation and distribution sampling experiments, *Journal of the American Statistical Association* **73**, 504–525.

BYRON J.T. MORGAN

ARMA and ARIMA Models

Public health institutions frequently collect notifications of diseases, entries into a hospital, injuries due to accidents, etc. at weekly or monthly intervals. Consecutive observations of such "time series data" are likely to be dependent. In environmental medicine, where series such as daily concentrations of pollutants are analyzed, it is evident that stochastic dependence of consecutive measurements may be important: a high concentration of a pollutant today, for example, has a certain inertia, that is, a tendency to be high tomorrow as well (positive autocorrelation). Dependence of consecutive observations may be equally important when data such as blood glucose are recorded within an single patient.

An important class of models having the flexibility to represent the stochastic dependence of consecutive data are autoregressive integrated moving average (ARIMA) models. Box & Jenkins [3] presented a detailed and influential account of these models and "ARIMA modeling" has become well established in such fields as economics and industry. The method of model identification, estimation, and checking is now often referred to as "the Box–Jenkins approach" and ARIMA models are also often called *Box–Jenkins models*.

ARIMA models may be particularly useful for **forecasting**. Forecasts of epidemiologic time series, for example, are often needed by public health organizations, since it is clearly of interest to know what frequencies of diseases might be expected in the future in order to better plan the distribution of resources. It may also be of interest to assess relations between two ARIMA time series, a "response" or "output" series, such as the daily number of patients coming to a clinic, and "explanatory" or "input" series, such as daily concentrations of a pollutant, daily mean temperature or other climatic series. Such situations may be represented adequately by the so-called **transfer function models**. Analogous questions arise when studying time series data recorded in an individual subject. Studies on individual subjects may have great potential for the investigation of biological mechanisms. Other questions are concerned with "changes" of time series: how efficient was a preventive program to decrease the monthly number of accidents? How did the pattern of morbidity in a population change after an environmental accident? These and other related questions may be investigated by an extension of ARIMA modeling [4] called *intervention analysis*.

The ARMA Model

Denote the observations at equally spaced times $t, t-1, t-2, \ldots$ by $z_t, z_{t-1}, z_{t-2}, \ldots$. For simplicity assume that $\mathrm{E}(z_t) = 0$ (otherwise the z_t may be considered as deviations from their mean). Let a_t, $a_{t-1}, a_{t-2}, \ldots$ be a white noise (*see* **Noise and White Noise**) series consisting of independent identically distributed random variables whose distribution is normal with mean zero and variance σ_a^2. It is helpful to think of the a_t as a series of "random shocks".

To begin, assume that the present observation, z_t, is linearly dependent on the previous observation, z_{t-1}, and on the random shock, a_t:

$$z_t = \phi z_{t-1} + a_t, \quad \text{where } \phi \text{ is a parameter.} \qquad (1)$$

Since z_t is regressed on z_{t-1} it is called an autoregressive model of first order [abbreviated AR(1)] model.

Alternatively, one may express z_t as a linear combination of the present and the previous random shock:

$$z_t = a_t - \theta a_{t-1}, \quad \text{where } \theta \text{ is a parameter.} \qquad (2)$$

This expression is called a moving average model of first order [abbreviated MA(1)] model.

The above two basic models are special cases of two more general models:

$$z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t, \qquad (3)$$

called an autoregressive model of order p [AR(p) model] and

$$z_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \qquad (4)$$

called a moving average model of order q [MA(q) model].

By combining these two equations one obtains what is called the autoregressive moving average model of order p and q [ARMA(p, q) model]:

$$z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \qquad (5)$$

This representation is relatively cumbersome to read and the situation becomes worse when considering generalizations, for example seasonality. The use of the backshift operator notation considerably improves the situation. The *backward shift operator* B is such that

$$Bz_t = z_{t-1}, \qquad B^k z_t = z_{t-k}. \tag{6}$$

The AR(1) model may then be written:

$$\begin{aligned} z_t &= \phi z_{t-1} + a_t, \\ z_t - \phi B z_t &= a_t, \\ (1 - \phi B) z_t &= a_t. \end{aligned} \tag{7}$$

Analogously, the AR(p) model may be written:

$$(1 - \phi_1 B - \cdots - \phi_p B^p) z_t = a_t, \tag{8}$$

and the MA(q) model may be written:

$$z_t = (1 - \theta_1 B - \cdots - \theta_q B^q) a_t. \tag{9}$$

Combining the AR(p) model and the MA(q) model, we obtain the ARMA(p, q) model:

$$\begin{aligned} &(1 - \phi_1 B - \cdots - \phi_p B^p) z_t \\ &\quad = (1 - \theta_1 B - \cdots - \theta_q B^q) a_t \end{aligned}$$

or

$$\phi(B) z_t = \theta(B) a_t, \tag{10}$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$.

The two polynomials in B, $\phi(B)$ and $\theta(B)$, are called the autoregressive and moving average operators, respectively. If the polynomial $\phi(B)$ has complex roots corresponding to $\phi(B) = 0$ (where B is viewed as a complex variable), then this indicates that the series contains random or quasi-periodic components [3].

To obtain more insight into the structure of the models, we use the backshift operator notation and write the AR(1) model in a different form:

$$\begin{aligned} (1 - \phi B) z_t &= a_t, \\ z_t &= (1 - \phi B)^{-1} a_t, \end{aligned}$$

or

$$z_t = (1 + \phi B + \phi^2 B^2 + \cdots) a_t,$$

or

$$z_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \cdots. \tag{11}$$

Thus, the current observation, z_t, is given by an (exponentially) weighted sum of random shocks. The relation shows that the AR(1) model can also be represented by a MA(∞) model. This duality holds, in general, between AR and MA models. In particular, the AR(1) and the MA(1) models are both generated by the white noise series, but they differ strongly in absorbing the random shocks. This difference is reflected in the different dependence structure and in the different forecasting properties of the models.

It has been shown that ARMA(p, q) models may be represented by a weighted sum of random shocks:

$$z_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \tag{12}$$

or

$$z_t = \psi(B) a_t,$$

where $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$. Comparing with (10) we see that $\psi(B) = \theta(B)/\phi(B)$. z_t is the output from a "linear filter" whose input is white noise and the "transfer function" $\psi(B)$ is a rational function of B. Representation (12) is not used for estimation since, in general, it contains an infinite number of parameters. However, the ARMA representation $\phi(B) z_t = \theta(B) a_t$ contains only $p + q$ parameters. Often, it is possible to find a parsimonious and adequate ARMA(p, q) representation with $p \leq 2$ and $q \leq 2$.

The Nonseasonal ARIMA Model

If the roots of the polynomial $\phi(B)$ lie outside the unit circle, it may be shown than an ARMA(p, q) process is stationary. Stationarity signifies that the probability structure of the series does not change with time. In particular, a stationary series has a constant mean and variance and a covariance structure that depends only on the difference between two time points.

Experience in industry, economics and, more recently, in medicine have shown that many time series are not stationary. However, it has been found that the series of first differences,

$$w_t = z_t - z_{t-1} = \nabla z_t, \tag{13}$$

is often stationary. The symbol $\nabla = 1 - B$ is called the (ordinary) differencing operator. If a series has to be differenced one time to obtain stationarity, then the model corresponding to the original series

is called an integrated ARMA model of order $p, 1, q$ or an ARIMA$(p, 1, q)$ model. If differencing has to be performed d times to obtain stationarity the model is written

$$\phi(B)\nabla^d z_t = \theta(B)a_t \tag{14}$$

and called an ARIMA(p, d, q) model.

To stabilize the variance it may be useful to consider transformations of the raw data; in particular, the logarithmic or the square root transformation [3, 11] (*see* **Power Transformations**).

The Seasonal ARIMA Model

Box & Jenkins have extended the above concepts to cope with **seasonal time series**. The model is obtained in two steps. Consider the case of monthly data.

1. An observation for a particular month is related to the observation for 12 months, 24 months, etc. previously by

$$z_t - \Phi_1 z_{t-12} - \Phi_2 z_{t-24} - \cdots - \Phi_P z_{t-12P} = \alpha_t - \Theta_1 \alpha_{t-12} - \Theta_2 \alpha_{t-24} - \cdots - \Theta_Q \alpha_{t-12Q},$$

or

$$\Phi(B^{12})z_t = \Theta(B^{12})\alpha_t, \tag{15}$$

where the AR and MA operators are now polynomials in B^{12},

$$\Phi(B^{12}) = 1 - \Phi_1 B^{12} - \Phi_2 B^{24} - \cdots - \Phi_P B^{12P},$$
$$\Theta(B^{12}) = 1 - \Theta_1 B^{12} - \Theta_2 B^{24} - \cdots - \Theta_Q B^{12Q},$$

and

$$B^{12}\alpha_t = \alpha_{t-12}.$$

Capital letters are used to distinguish this from the nonseasonal ARIMA model (14).

2. The error component, α_t, for a particular month is related to that for previous months by the usual ARMA model:

$$\phi(B)\alpha_t = \theta(B)a_t. \tag{16}$$

Joining the seasonal and the nonseasonal parts gives

$$\phi(B)\Phi(B^{12})z_t = \theta(B)\Theta(B^{12})a_t. \tag{17}$$

Extending the concept of ordinary differencing to seasonal differencing by forming seasonal differences,

$$w_t = z_t - z_{t-s} = \nabla_s z_t, \tag{18}$$

where $\nabla_s = 1 - B^s$ is the seasonal differencing operator and $s = 12$, for, for example, monthly data, one obtains the seasonal ARIMA model:

$$\phi(B)\Phi(B^s)\nabla^d \nabla_s^D z_t = \theta(B)\Theta(B^s)a_t, \tag{19}$$

abbreviated to the ARIMA$(p, d, q)(P, D, Q)_s$ model.

The Autocorrelation Function and Model Identification

The dependence structure of a stationary time series is characterized by the **autocorrelation function** (ACF). The ACF is defined as the correlation between z_t and z_{t+k}:$\rho_k = \text{cor}(z_t, z_{t+k})$. k is called the time lag.

The ACF is estimated by the empirical ACF:

$$r_k = \frac{c_k}{c_0}, \quad k = 0, 1, 2, \ldots,$$

where

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z}) \quad \text{and} \quad \bar{z} = \frac{1}{n}\sum_{t=1}^{n} z_t. \tag{20}$$

c_k are the empirical autocovariances.

The empirical ACF is the main tool for the identification of the ARIMA model. ACFs of the basic processes have a typical shape. The ACF of the AR(1) process decays exponentially. The ACF of the MA(1) process has only ρ_1 nonzero and of the MA(2) process only ρ_1 and ρ_2 nonzero. Valuable complementary tools for model identification are the partial autocorrelation function [3] and the inverse autocorrelation function [5].

To obtain an adequate ARIMA model, Box & Jenkins have suggested the following procedure (the "Box–Jenkins method"):

0. "Make the series stationary", that is, consider transformations to stabilize the variance, consider ordinary and seasonal differencing.
1. Choose a provisional model; in particular, by looking at the empirical ACF.

2. Estimate the model parameters (standard software such as SAS or BMDP allow maximum likelihood estimation of Box–Jenkins models).
3. Check the adequacy of the model. In particular, check the ACF of the residuals for white noise.

If the model does not fit the data adequately, then go back to point 1 and choose an improved model. Among different models that represent the data equally well choose the simplest one, that is, the model with the fewest parameters. If this concept of parsimony is ignored, poor forecasts may result.

Forecasting

To obtain forecasts $\hat{z}_{t+h}$ for h time units (days, months, etc.) ahead from an ARIMA model, one writes the corresponding model equation by replacing (i) future values of the random shocks a by zero and past values by observed residuals; (ii) future values of z by the corresponding forecasts; (iii) past values of z by their observed values.

The following example illustrates how to obtain forecasts for an AR(1) model:

$$\begin{aligned} h=1:\quad & z_{t+1} = \phi z_t + a_{t+1}, \\ & \hat{z}_{t+1} = \phi z_t, \\ h=2:\quad & z_{t+2} = \phi z_{t+1} + a_{t+2}, \\ & \hat{z}_{t+2} = \phi \hat{z}_{t+1} = \phi(\phi z_t) \\ & \qquad = \phi^2 z_t, \text{ etc.} \end{aligned} \tag{21}$$

By continuing this procedure one may see that the forecasts corresponding to the AR(1) model follow an exponential curve.

The Transfer Function Model

Box & Jenkins [3] have developed an important extension allowing us to analyze relations between an "output" series (e.g. the daily number of patients with a specified disease coming to a clinic) and one or several "input" series (e.g. the daily concentrations of pollutants, daily mean temperature, etc.). In the **transfer function models** the output series, y_t, is considered to be composed of two parts:

$$y_t = u_t + n_t. \tag{22}$$

u_t is the part that may be explained in terms of one (or several) input series x_t (concentration of a pollutant, etc.). n_t is an ARIMA process as described above. It represents the unexplained part of y_t (noise process).

It is assumed that the explained part, u_t, is given by a weighted sum of the present and past values of the input, x_t:

$$u_t = v_0 x_t + v_1 x_{t-1} + \cdots \tag{23}$$

or

$$u_t = v(B) x_t,$$

where

$$v(B) = v_0 + v_1 B^1 + v_2 B^2 + \cdots.$$

$v(B)$ is called the transfer function or the impulse response function; $v_0, v_1, \ldots$ are called transfer function weights. Eq. (23) is not a parsimonious representation of the transfer function model, but it is useful for model identification via the "prewhitened cross-correlation function" (see below).

A parsimonious "rational lag representation" of u_t can be obtained by writing $v(B)$ as the quotient of two polynomials in B, $v(B) = \omega(B)/\delta(B)$ [3]. Thus, $\omega(B)u_t = \delta(B)x_t$ and the transfer function model is given by

$$y_t = \left[\frac{\omega(B)}{\delta(B)}\right] x_t + \left[\frac{\phi(B)}{\theta(B)}\right] a_t. \tag{24}$$

Identification of a Transfer Function Model

Identification of Univariate Models for Input and Output Series

The univariate models of the *input* series are necessary in order to obtain a guess of the transfer function (via the prewhitened cross-correlation function).

The univariate model of the *output* series has two purposes:

1. It provides an initial guess of the noise process, n_t.
2. The residual variance of the univariate model may be used as a "yardstick" when comparing different transfer function models.

The Cross-correlation Function and Prewhitening

The relation between two time series, x_t and y_t, is determined by the cross-correlation function (CCF): $\rho_{xy}(k) = \text{cor}(x_t, y_{t+k}), k = 0, \pm 1, \pm 2, \ldots$. This function determines the correlation between the two series as a function of the time lag, k.

The main tool to identify a transfer function model is the empirical CCF $r_{xy}(k)$. However, a basic difficulty arises in the interpretation of the empirical CCF. As discussed by Box & Jenkins [3] the empirical CCF between two completely unrelated time series, which are themselves autocorrelated, can be very large due to chance alone. In addition, the cross-correlation estimates at different lags may be correlated. This is due to the autocorrelation within each individual series.

Box & Jenkins [3] proposed a way out of this difficulty called "prewhitening": the ARIMA model for the input series converts the correlated series x_t into an approximately independent series α_t. Applying the identical operation to the output series, y_t, produces a new series, β_t. It may be shown that the CCF between α_t and β_t (called the prewhitened CCF) is proportional to the transfer function. Thus, the empirical prewhitened CCF allows one to obtain a guess of the transfer function. An analogous iterative procedure as described above for ordinary ARIMA models leads to an adequate parsimonious transfer function model.

Example

The purpose of this study was the assessment of the relation between environmental time series and respiratory symptoms in preschool children. During about one year daily concentrations of SO_2, NO_2 and other environmental time series were collected in Basle. Simultaneously, the daily number of respiratory symptoms per child in a randomly selected group of preschool children was recorded. This series is termed "SYMPTOMS". Since January and February are the months with the strongest winter heating (SO_2), a separate model was identified for this "winter period". In a first step, ARIMA models for the individual series SO_2, NO_2, and SYMPTOMS had to be identified.

For the input series, SO_2, the Box–Jenkins method of model identification was straightforward. The mean-range plot [11] of SO_2 showed a tendency for the range to increase with the mean, whereas for $\ln(SO_2)$ the range was approximately independent of the mean, indicating that the logarithmic transformation stabilizes the variance. Figure 1 (lower curve) shows the plot of $\ln(SO_2)$. Figure 2 shows the corresponding ACF and PACF. The observed pattern (slow

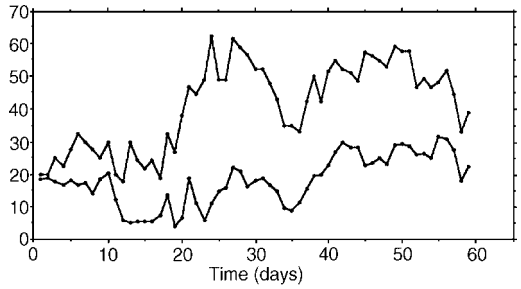

Figure 1 Upper curve: series SYMPTOMS × 100. Lower curve: rescaled series of $\ln(SO_2)$; Day 1 corresponds to 1 January 1986

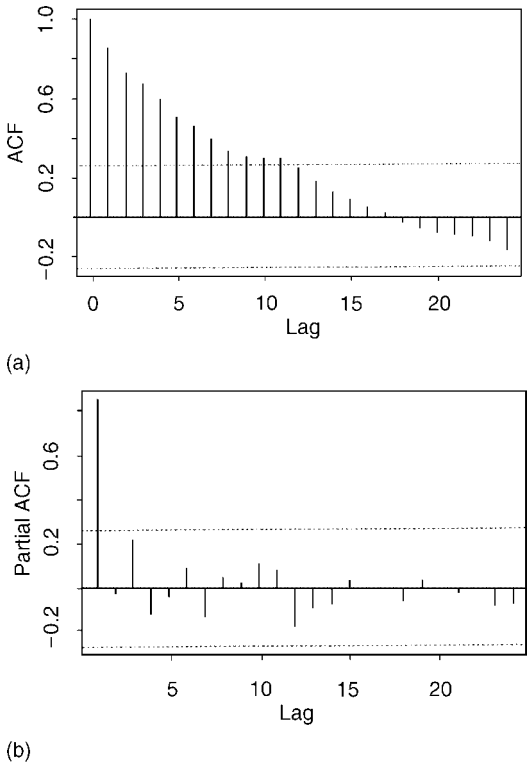

Figure 2 (a) Autocorrelation function (ACF) of series $\ln(SO_2)$; (b) partial autocorrelation function of series $\ln(SO_2)$

decay of ACF and marked peak at lag 1 in the PACF) suggests tentative fitting of an AR(1) model. The ACF and PACF of the residuals showed no marked peaks. The goodness-of-fit test [1] showed no sign of model inadequacy. A similar AR(1) model was found for the input series NO_2.

The output series SYMPTOMS is plotted in Figure 1 (upper curve). The ACF of this series showed a similar pattern. Tentative fitting of an AR(1) model gave an acceptable fit. However, the estimated autoregressive coefficient was close to one ($\phi = 0.96$), indicating possible nonstationarity. Thus, for the series SYMPTOMS one had to choose between two competing ARIMA models. Fitting a MA(1) model to the series of differences, that is, fitting an ARIMA (0,1,1) model, showed no sign of model inadequacy. Akaike's information criterion (AIC) was 388 for the AR(1) model and 379 for the ARIMA (0,1,1) model. In addition, the residual variance was somewhat smaller in the latter model. Both signs indicated superiority of the (nonstationary) ARIMA (0,1,1) model over the (stationary) AR(1) model. The corresponding estimated univariate model is shown in the first line of Table 1.

The CCF between the series $\ln(SO_2)$ and the series SYMPTOMS before prewhitening was not interpretable: "significant" coefficients are "smeared" over a large range of positive and negative time lags and there are typical nonsense coefficients suggesting that a high number of respiratory symptoms today are expected to be followed by high pollution during the next 10 days! After prewhitening, one marked positive peak is found at time lag zero, while all other coefficients are not significantly different from zero.

These results suggested the following transfer function model for y_t (SYMPTOMS) and $x_t(\ln(SO_2))$:

$$y_t = v_0 x_t + n_t, \quad \text{where } \nabla n_t = (1 - \theta B)a_t, \quad (25)$$

or, written compactly in standard notation,

$$\nabla y_t = v_0 \nabla x_t + (1 - \theta B)a_t. \quad (26)$$

The diagnostic checks of the residuals of the transfer function models showed no sign of model inadequacy. The same type of model was identified for SYMPTOMS and $\ln(NO_2)$ as the input series. In addition, a two-input model was fitted with $\ln(SO_2)$ and $\ln(NO_2)$ as the input series.

The summary of the models is presented in Table 1. One sees from the residual variances of the corresponding one-input models that the series SO_2 contributes more to the explanation of the series SYMPTOMS than the series NO_2. The two-input model shows no stronger reduction of the residual variance of SYMPTOMS than the one-input model with SO_2. Thus, the transfer function model revealed that input is related "instantaneously" with output; in particular, there is no "delayed" effect of the pollutant (no transfer function weight different from zero at nonzero time lags). A more detailed discussion of this example can be found in [10].

Literature

The following suggestions for further reading may be helpful. A thorough introduction to ARIMA models and transfer function models may be found in [1].

Table 1 Comparison of univariate and transfer function models fitted for output series y_t (SYMPTOMS) and input series x_{1t} [$\ln(SO_2)$] and x_{2t} [$\ln(NO_2)$]

Model type	Estimated model	Residual variance ($\times 10^4$)
Univariate:	$\nabla y_t = (1 - 0.24B)a_t$ ± 0.13	0.00387
One-input:		
x_{1t} : $\ln(SO_2)$	$\nabla y_t = 0.078\nabla x_{1t} + (1 - 0.26B)a_t$ $\pm 0.17 \qquad \pm 0.13$	0.00288
x_{2t} : $\ln(NO_2)$	$\nabla y_t = 0.068\nabla x_{2t} + (1 - 0.19B)a_t$ $\pm 0.022 \qquad + 0.13$	0.00339
Two-input:		
x_{1t} : $\ln(SO_2)$, x_{2t} : $\ln(NO_2)$	$\nabla y_t = 0.067\nabla x_{1t} + 0.025\nabla x_{2t} + (1 - 0.26B)a_t$ $\pm 0.20 \qquad \pm 0.24 \qquad \pm 0.13$	0.00288

An introduction to ARIMA models using biological and medical datasets is given by Diggle [7]. The classical reference to ARIMA models is Box & Jenkins [3]. Jenkins [11] provides instructive case studies in the fields of business, industry and economics. For studying specific medical time series problems the following articles are thought to be of interest. Examples of pitfalls in the analysis of relations between seasonal series in epidemiology are presented in [2]. Identification of seasonal ARIMA models representing infectious diseases is presented in some detail in [8]. A review, examples and references of studies concerned with ARIMA modeling in medicine are given in [9]. Applications and references of studies concerned with ARIMA modeling of single patient data may be found in [6].

References

[1] Abraham, B. & Ledolter, J. (1983). *Statistical Methods for Forecasting*. Wiley, New York.

[2] Bowie, C. & Prothero, D. (1981). Finding causes of seasonal diseases using time series analysis, *International Journal of Epidemiology* **10**, 87–92.

[3] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, Revised Ed. Holden-Day, San Francisco.

[4] Box, G.E.P. & Tiao, G.C. (1975). Intervention analysis with applications to economic and environmental problems, *Journal of the American Statistical Association* **70**, 70–79.

[5] Chatfield, C. (1979). Inverse autocorrelations, *Journal of the Royal Statistical Society, Series A* **142**, 363–377.

[6] Crabtree, B.F., Ray, S.C., Schmidt, P.M., O'Connor, P.J. & Schmidt, D.D. (1990). The individual over time: time series applications in health care research, *Journal of Clinical Epidemiology* **43**, 241–260.

[7] Diggle, P.J. (1990). *Time Series. A Biostatistical Introduction*. Clarendon Press, Oxford.

[8] Helfenstein, U. (1986). Box-Jenkins modelling of some viral infectious diseases, *Statistics in Medicine* **5**, 37–47.

[9] Helfenstein, U. (1996). Box-Jenkins modelling in medical research, *Statistical Methods in Medical Research* **5**, 3–22.

[10] Helfenstein, U., Ackermann-Liebrich, U., Braun-Fahrländer, Ch., Wanner, H.U. (1991). Air pollution and diseases of the respiratory tracts in pre-school children: a transfer function model, *Journal of Environmental Monitoring and Assessment* **17**, 147–156.

[11] Jenkins, G.M. (1979). *Practical Experiences with Modelling and Forecasting Time Series*. Gwilym Jenkins, St. Helier.

(*See also* **Multiple Time Series**; **Spectral Analysis**)

ULRICH HELFENSTEIN

Artificial Intelligence

Artificial intelligence (AI) has been defined as the study of artificial systems which exhibit intelligent behavior. This definition neatly sidesteps the need to define precisely what "intelligence" itself is!

Research in artificial intelligence splits broadly into two camps. The first seeks to yield greater understanding of how naturally intelligent systems (human brains, for example) function. This area has developed according to the principle that we can only be sure we understand the natural system being investigated if we can build a model which behaves in the same way as that system. This perspective views artificial intelligence research as a subdomain of cognitive psychology. The second camp seeks to build systems (computer programs, robots, etc.) which behave in apparently intelligent ways, regardless of whether the way in which this behavior is achieved emulates naturally intelligent systems. An analogy for the motivation underlying the second camp is with flight: we could fly by building ornithopters, emulating the flapping wings of birds; but we need not – we could, instead, build a helicopter or a fixed wing aeroplane. Instead of a subdomain of psychology, then, the second camp might be viewed as a subdomain of engineering.

Examples of systems of the second kind that can already be encountered in regular use are expert systems, natural language understanding systems, software verification systems, symbolic algebra systems (*see* **Computer Algebra**), and game-playing machines. The last example here provides a nice illustration of how apparently intelligent behavior can be achieved without necessarily emulating humans: chess-playing programs have now been developed which can compete at grand master level. However, they achieve their successes through strategies quite different from those of human grand masters. In particular, they adopt massive searches of the state–space of the chess game, whereas humans conduct focused searches of a much smaller space.

Artificial intelligence research is fundamentally interdisciplinary. It overlaps with computer science, cognitive psychology, statistics, mathematics, engineering, biology, linguistics, and other disciplines, and has developed through several stages. Early work was characterized by inflated claims of imminent achievement. In the context of the time, these were quite understandable. Such early work picked, as its problem domains, areas such as logical problems, puzzles, and the kind of matching problems found in IQ tests. From a human perspective, problems such as these clearly require intelligence. The early success of computer programs on such problems was taken to imply that mere scaling up would enable domains such as natural language understanding and visual object recognition to be readily tackled. However, this turned out not to be the case. More than a question of scale was involved. In retrospect, it can be seen that the early problem domains were defined in terms of a small and precise dictionary of concepts. (The extreme example is arithmetic, with a dictionary of 14 symbols – the digits and four arithmetic operators. Computers can perform arithmetic effectively instantaneously, making a human feel rather stupid by comparison.) This is in contrast to the huge ill-defined dictionaries of concepts of more "realistic" domains. (However, there are real practical problems which involve well-defined concept dictionaries, and where modern AI systems, notably expert systems, have been effectively applied.) As a consequence, the early expectations of rapid practical applications were not fulfilled. Instead, a long hard haul has been required, in which developments in the theory and methods of AI have been matched and supported by dramatic developments in computer hardware. The progress that has been made in AI technology over the past two decades owes a great deal to the latter.

The bulk of the early work in AI was based on the *symbol manipulation* paradigm, mentioned in the preceding paragraph. In this, a basic dictionary of symbols is defined, along with relationships between them. These symbols are combined into more complex structures, with relationships between these structures; and this is repeated, each time moving up a level of complexity until extremely complex structures are created. Achieving structural complexity in this way, by means of series of levels, emulates the situation in other domains. Examples are using letters to build words, words to build sentences, sentences to build paragraphs, and paragraphs to build books, or the natural example of atoms, molecules, cells, and multicellular organisms. Special purpose languages (*see* **Computer Languages and Programs**) were developed for AI programming, notably LISP, which followed this paradigm.

The symbol manipulation approach toward constructing intelligent machines has continued, but

over the past decade a new paradigm has also attracted considerable interest. This is the *connectionist* paradigm. Digital computers are essentially *serial* machines. They read a command, act on it as appropriate, and then move on to the next one. Biological brains, however, are anything but serial. They are *parallel*. They consist of huge numbers of cells (neurons), each connected to vast numbers of other cells, not merely connected to a "preceding" and "following" cell. Thus, the fact that the processing speed of a single neuron is tiny compared to that of a single electronic switch becomes irrelevant. With 1000 electronic switches connected in a line, switching at a rate of one on/off per millisecond, serial computation means that a second will elapse before a signal can propagate from one end to the other and switch them all on. In contrast, if the switches are in parallel, they can all switch on simultaneously. This parallelism explains how the brain can carry out certain kinds of operations much faster than can a digital computer. Recognition of this fact was an early stimulus behind connectionist or parallel approaches to artificial intelligence.

In the sections that follow we examine some important subdomains of AI research.

Knowledge Representation

One key to effective problem solving is finding a good way to think about the problem. In general, the key to effective knowledge manipulation is finding a way to represent the knowledge which permits ready search and restructuring to match the objectives. Several representations are particularly important in AI.

Semantic networks have nodes representing objects, with labeled links connecting these nodes. The links represent relationships between the objects and the labels on the links specify attributes and types of relationships. This is a powerful general knowledge representation. For example, geometric diagrams or visual scenes can be represented: objects might be tables, plates, and cutlery, with relationships such as "on top of" and "beside" being represented by links. Perhaps at the other extreme, stories and verbal discourse can be represented: nodes might represent individuals and objects within the story, while changes in the way individuals feel about each other can be represented by changing values of labels on links of the net.

Production systems represent knowledge in terms of *antecedent–consequent rules*. The left-hand sides, or antecedents, of such rules consist of a set of conditions which must be satisfied by the items in a database. When they are satisfied the rule is said to *fire* – it carries out some operation (the consequent), such as altering the database. The system cycles through a set of such rules: each time a rule fires the database is changed, until some terminating state is reached. This is the basic representation underlying *expert systems* (see, for example, the summary of the MYCIN project described in [2]). In the case of medical diagnosis, for example, an initial set of signs and symptoms is fed into the database and the system cycles through its rule base, updating and modifying the database until a diagnosis is reached (*see* **Computer-aided Diagnosis**; **Decision Analysis in Diagnosis and Treatment Choice**).

Logic has the advantages that it is well-understood, rigorous, and (by definition) completely formalized. Different kinds of logic – propositional calculus, first-order predicate calculus, second-order predicate calculus – permit increasingly complex situations to be represented. Logical structures involve predicates, variables, constants, logical connectives, and quantifiers. In particular, *predicates* are building blocks which can take particular values. Thus *blue(book_6)* might be used to indicate that the colour of "book_6" is blue and *lift(John, book_6)* might indicate that John lifted book_6. Considerable effort has gone into developing automatic proof procedures in logic, so that powerful systems can be developed. The important AI language Prolog is based on a subset of predicate calculus.

Search

If knowledge representation is one key to problem solving, then effective search strategies are the other. Many problems reduce to finding the best, or at least a good, solution to a particular question in a space consisting of a large number of potential solutions. "Large" here can often mean astronomically vast, so that no exhaustive search will ever be conceivably feasible, even by the fastest of imaginable computers ever. (In chess for example, there are around 10^{120} possible games. Put this in the context of there being around 3×10^7 seconds in a year.) Examples of problems requiring efficient searches are graph

matching problems (such as matching a parsed input spoken sentence to a dictionary of sentences, or matching a segment of a semantic network to the right part of a larger net), identifying eligible production rules in a rule base, or finding a proof of a theorem (in logic, perhaps). This last example requires finding a path from the premises of the theorem to the conclusion of the theorem.

Sophisticated search methods, such as branch and bound or mathematical programming, guarantee finding the global optimum (*see* **Optimization and Nonlinear Equations**), and extend the range of problems from those that can be tackled by exhaustive search, but even they fail in the face of really large search spaces.

Of course, efficient search strategies will be familiar to statisticians, in the form of forward and backward stepwise methods in regression (*see* **Variable Selection**), discriminant analysis (*see* **Discriminant Analysis, Linear**), and other model building situations. Such methods achieve their aim of making the search feasible by restricting the search to a subspace of the complete space – and so risk missing the global optimum. Classical search methods, such as steepest descent and other mathematical *optimization* methods, may be stymied by the nature of the search space: it often involves categorical variables, so that the function for which an optimum is sought is not differentiable (or even continuous). Another general strategy which can facilitate efficient search is that of breaking the problem down into components. For some situations, we might be able to show that "if A is true" and "if B is true" then the original problem is solved. If we are lucky, A and B separately may be much easier to prove than the original problem.

Often progress can be made by utilizing problem specific general guidelines about what might assist in finding a good solution. Such guidelines are known as *heuristics* [7]. Examples of heuristics are summary values of the estimated strengths of positions in chess games and the use of samples to infer (with the risk of error) some general property of a population. We see from the latter that statisticians are masters at certain kinds of heuristic reasoning!

Constraints on solutions can be extremely effective in narrowing down the potential search space. In vision and speech recognition, for example, we can utilize information about possible global structures to restrict the possible components (parts of an image, words or phonemes in speech) which need to be tested.

Search strategies can often be described as *tree structures*. Suppose, for example, the problem involves matching the description (of a semantic net or a logical structure, say) of one structure (A) to that of another (B). The nodes of a tree might represent different descriptions of B, with higher level nodes being general descriptions and lower level nodes being more specific (filling in the values or restricting the ranges of variables, for example). We start at the highest (most general) level – which A necessarily matches – and work our way down, looking at more specific descriptions. Our aim is to find that completely specified description – that leaf node – of the B tree which matches A. We can undertake this search *depth first* or *breadth first*. The former involves running right down one branch until the end is reached and (assuming a perfect match is not obtained) backing up and following a neighbouring branch. The latter involves looking a little bit down all the branches, hoping that some will be eliminated early on.

Connectionist Approaches

The earliest parallel system explored in depth was the *perceptron*. In statistical terms this is a simple linear model used to predict class membership – as in linear discriminant analysis. It takes as inputs the values of several variables and forms their weighted sum, which is then compared with a threshold to determine predicted class membership. The variables thus contribute in a parallel way to the decision. The difference between discriminant analysis and the perceptron involves mainly the parameter estimation methods and the criterion being optimized. The parameter estimation methods of the perceptron involve sequential presentation of the data points, with iterative updating, rather than the algebraic solution of linear discriminant analysis; and the perceptron minimizes misclassification rate rather than the ratio of between-to-within-group distances to determine the best set of weights.

Unfortunately, as was shown by Minsky & Papert [5] in an influential book, the capabilities of the perceptron are severely limited. Minsky & Papert did describe how to overcome this problem – by using extra stages of linear combinations and

applying nonlinear transformations prior to the combination – but the parameter estimation problems seemed intractable at the time. It was not until several years later, with progress in hardware capabilities, and after several authors had presented estimation techniques, that interest again picked up. Since then, the interest that the idea of the **neural network** or *connectionist* approaches have attracted has been remarkable – especially from a statistical perspective, since from that perspective such models can be seen to be mere generalizations of **logistic regression** models (and, indeed, alternative, similarly flexible, statistical models, have also been developed).

One way of looking at neural networks is to regard them as searching the space of possible transformations of the vector of input variables, so as to find a set which permits a highly discriminatory linear combination. This lies in sharp contrast to earlier methods of **pattern recognition** (and, indeed, discriminant analysis), which required the system developer to identify good transformations of the input vectors. It means that little expert knowledge of the problem domain is required. This is obviously attractive to many potential users, since it implies a saving of time and effort.

Conclusions

Although the early promise of quick solutions did not materialize, steady development, coupled with the huge advances in computing hardware (*see* **Computer Architecture and Organization**), has led to significant progress in AI. The widespread use of computers (often concealed within devices and machines) means that the impact of this progress is likely to be substantial over the next few years.

An illustration of one area of development in which this is likely to be the case is language processing. It was realized, some time ago, that perfect translation between natural languages needed a deep representation of the ideas being expressed. It is insufficient, for example, merely to use a huge dictionary of words and expressions. However, effective language processing systems can be produced if the aims are less grandiose; in particular, if a restricted set of expressions are involved or if the system drives the interaction, as in computer interviewing to collect data for medical diagnosis (*see* **Computer-assisted Interviewing**). Sophisticated versions of such systems develop internal models of the patient as they proceed with the interview. Similar systems are being developed for computer-aided instruction, where more focused teaching is possible if the system refines a model of the student's abilities and likely responses as the instruction session proceeds.

One of the difficulties with which AI research has to cope is the perception that once a problem has been analyzed and a system built to tackle it, then "intelligence" is no longer required. In this sense, AI research is a doomed enterprise. In spite of that, the results of AI research are beginning to be felt in everyday life, as computers become more and more ubiquitous. In many areas, the problems tackled by AI are identical to those tackled by statisticians. Each of the two disciplines brings their own strengths to tackling the problems, and a rich synergistic interaction can result.

Recommended books in the area include [6] and [10]. Some medical applications are described in [9], and its potential role in psychiatry is outlined in [3]. Neural networks and their relationship to statistical methods for tackling the same problems are described in [1, 4], and [8].

References

[1] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

[2] Buchanan, B.G. & Shortliffe, E.H. (1984). *Rule-Based Expert Systems*. Addison-Wesley, Reading.

[3] Hand, D.J. (1985). *Artificial Intelligence and Psychiatry*. Cambridge University Press, Cambridge.

[4] Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. Wiley, Chichester.

[5] Minsky, M. & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, Mass.

[6] Nilsson, N.J. (1980). *Principles of Artificial Intelligence*. Springer-Verlag, Berlin.

[7] Pearl, J. (1984). *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, Reading.

[8] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

[9] Szolovits. P. (1982). *Artificial Intelligence in Medicine*. Westview, Boulder.

[10] Winston, P.H. (1992). *Artificial Intelligence*, 3rd Ed. Addison-Wesley, Reading.

(*See also* **Neural Network**)

DAVID J. HAND

Ascertainment

What is the ascertainment problem? In brief, this problem arises from the fact that for most genetic disorders of interest, families are not selected for study at random, but come to the attention of investigators (i.e. are "ascertained") via some process which may or may not be well understood. These deviations from random sampling can introduce **bias** into genetic analysis if they are not correctly understood and allowed for.

In particular, for a **segregation analysis** one needs to incorporate the ascertainment model into the analysis (for example, by including the probability model for ascertainment in the likelihood of the genetic model); otherwise, serious distortion may result. Ascertainment has been believed not to be an issue for **linkage analysis** as long as individuals were ascertained based only on a single trait (e.g. [4] and [13]), i.e. based on either the trait of interest or the marker, but not both. However, see also the section "Sequential Sampling" below.

A very simple example will illustrate the general principles. Imagine we are studying a genetic disease, thought to be inherited as a rare recessive (*see* **Genotype**), and we wish to determine whether the **segregation ratio** within nuclear families is $\approx 25\%$ (as predicted for a rare recessive disease). Imagine further that we are able to locate every family in our catchment area with at least one affected child, and that every such family is willing and able to participate in our study. The actual proportion of affected children that we will observe in this hypothetical example will be not 25%, but rather a higher proportion. The apparent distortion arises because those families which, by chance alone, failed to produce any affected children will not enter the sample. For example, say we had an ideal population of 10 000 two-child families who were at risk to have an affected child (i.e. both parents are carriers). With each birth, such a mating type (pair of parents) has a probability of 3/4 of having an *un*affected child. Thus, on average, $(0.75)^2$ of these two-child families, i.e. 5625 families, would have *no* affected children and would be unable, a priori, to enter our sample. Of the remaining 4375 families that would enter our sample, 3750 would have one affected and one unaffected child, and 625 would have both children affected. If we naively counted up the total number of affected children ($3750 + [625 \times 2] = 5000$) and divided by the total number of children in our sample ($4375 \times 2 = 8750$ children), without awareness of the 5625 families that we did not "see", we would estimate the segregation ratio as 57.1%, which is seriously distorted from the expected segregation ratio of 25%.

Moreover, there is no guarantee that we will even find every family with at least one affected child. In many situations families with more affected children are more likely to come to the attention of an investigator than those with fewer affected children, and this fact can introduce additional distortion into segregation analysis.

If all details of the ascertainment model are known and can be modeled probabilistically, then they can in most cases (except sequential sampling; see below) be incorporated into the likelihood required for any type of statistical–genetic analysis. The difficulty arises when the ascertainment model is not known, or is not known accurately.

In the next section we explore in some detail the "classical" ascertainment model of Weinberg [34] and Morton [23], based on the concept of "probands". In the third section we consider some alternative models, along with their criticisms of the proband concept. The fourth section considers some other methods of dealing with ascertainment, other than incorporating the ascertainment model into the likelihood. The fifth section examines two additional complications: sequential sampling and stoppage. The final section summarizes the conclusions of this article.

"Classical" Ascertainment Model

The classical model, also called the π-model, was formulated by Weinberg [34] and specifically applied to genetics by Morton [23]. This model requires the concept of a "proband", defined by Morton as "an affected person who at any time was detected independently of the other members of the family, and who would therefore be sufficient to assure selection of the family in the absence of other probands" [23]. Then the "ascertainment probability" π for an individual is defined as the probability that any affected individual becomes a proband, i.e. is ascertained. The ascertainment probability for a *family* is defined to be the probability that a family has at least one

"proband". A family that has at least one proband is then assumed to be ascertained. Note that the event "to be ascertained" is defined differently for an individual and a family. An individual is ascertained if s/he becomes a proband, whereas a family is ascertained if it contains at least one proband.

The fundamental quantity that we need to formulate is the probability distribution of r affected children within s-child families, *within our ascertained dataset*, i.e.

$$\Pr_s(r \text{ affected children}|\text{sibship is ascertained}). \quad (1)$$

We will refer to the quantity in (1) as the "fundamental ascertainment probability", i.e. the fundamental segregation probability in the presence of ascertainment.

These probabilities were originally developed for nuclear families, i.e. families consisting of two parents and their children, under the assumption that only children could be probands; however, the formulas can also be applied to extended pedigrees, in which any class of members may potentially be probands.

The ascertainment probability π can assume values in the interval $0 < \pi \leq 1$. We consider first two special cases, (a) and (b), then the general formulation, (c).

(a) If every affected individual is sure to become a proband, then $\pi = 1$. This is called "truncate selection" by Morton [23], or "complete ascertainment" by many other authors. It follows that every family with at least one affected member will be ascertained; thus families with at least one affected member appear in the dataset in the same relative proportions as they appear in the general population. Let p represent the segregation probability (e.g. 0.25 in the example above). When $\pi = 1$, the segregation analysis likelihood for a sibship of s children takes on the following relatively simple form, since the fundamental ascertainment probability in (1) can be rewritten:

$$\begin{aligned}
&\Pr_s(r \text{ affected children}|\text{sibship is ascertained}) \\
&\quad = \Pr(r \text{ affected children}|\geq 1 \text{ child is affected}) \\
&\quad = \frac{\Pr(r \text{ affected children})}{\Pr(\geq 1 \text{ child is affected})} \\
&\quad = \frac{\Pr(r \text{ affected children})}{1 - \Pr(0 \text{ children are affected})} \qquad (2)
\end{aligned}$$

for $r \geq 1$. Therefore

$$\begin{aligned}
&\Pr_s(r \text{ affected children}|\text{sibship ascertained}) \\
&\quad = \frac{\binom{s}{r} p^r (1-p)^{s-r}}{1 - (1-p)^s}. \qquad (3)
\end{aligned}$$

The numerator of (3) is a standard binomial probability, for an s-child family with r affected children, and the denominator gives the probability of at least one affected child, as in the denominator of (2). The complete formula in (3) represents a "truncate binomial" probability distribution.

Complete ascertainment can arise when one is able to ascertain every affected person in the population under study – for example, if all population members belong to a centralized health care system or a national health registry. Complete ascertainment may also occur if a disease is severe and rare, and there exists just one medical center in a certain geographic area that specializes in that disease. In such a situation it may be reasonable to assume that every family in that area with at least one affected child will come to that center and will be studied.

Table 1 (right column) summarizes the properties of truncate selection. Note that the segregation ratio is distorted because of the at-risk families who happen not to have an affected child, but that there is no *additional* distortion to that.

(b) A completely different kind of scenario occurs when the probability for any one affected child to become a proband is very small, and, concomitantly, the probability that any given family will have *more than one* proband is essentially zero. This can happen, for example, if the investigator ascertains only children attending second grade within one's geographic area of study: the probability that any given affected child will happen to be in second grade at this time is small; and essentially no families will have *two* affected children in second grade at the same time. Scenarios such as this correspond to letting π approach zero. This ascertainment model is known as "single selection" or "single ascertainment" and represents our second special case. It can be shown that under single ascertainment, the likelihood in (1) is

Table 1 Features of the classical π-model, applied to s-child sibships, ascertained through the children

$\pi \to 0$	π in between	$\pi = 1$
Name		
"Single selection" (single ascertainment)	"Multiple incomplete selection" (multiple ascertainment)	"Truncate selection" (complete ascertainment)
Number of probands		
One proband/sibship	>1 proband/sibship, but not every affected child is a proband	Every affected child is a proband
"Distortion" in the segregation ratio occurs because not every at-risk sibship actually has an affected child		
Additional distortion		
Sibships with r affected children occur r times more often than in their population proportions	In-between additional distortion	No additional distortion; sibships with $\geq$1 affected child occur in their population proportions
Probability distribution		
Binomial with $s-1$, $r-1$. (Exact analytic maximum likelihood estimate (MLE) of p)	No special probability distribution. (No exact analytic MLE of p)	Truncate binomial probability distribution (No exact analytic MLE of p)
Application		
"Every second-grader..." (see text)	"Real life..." (other ascertainment schemes)	"National Health Registry" (see text)
Equation in text:		
(4)	(7)	(3)

given by

$$\Pr_s(r \text{ affected children} | \text{sibship ascertained}) = \binom{s-1}{r-1} p^{r-1}(1-p)^{s-r}. \quad (4)$$

This represents a simple binomial probability in p, but applied to a sibship of $s-1$ children, of whom $r-1$ are affected.

Since (4) corresponds to simply removing the proband from the family, the correction for single ascertainment is sometimes referred to as "discarding the proband". However, the reader should note that "discarding the proband" works only for single ascertainment of nuclear families with a single mating type. More generally, one can think of "conditioning on the proband" to allow for single ascertainment [20].

It can also be shown that the probability that a sibship will be ascertained is then *proportional* to the number of affected sibs. For example, a family with two affected children is twice as likely to be ascertained as a family with one affected child. Thus, sibships with multiple affected children will be *over*-represented in the sample, compared with their proportions in the general population. See Table 1, left column.

This proportionality property is a powerful result and can be taken as the fundamental *defining* characteristic of single ascertainment [31]. It can be shown that in a number of circumstances single ascertainment is "special" in ways that do not hold for other ascertainment schemes [20]. For example, if extended pedigrees have been selected under single ascertainment, segregation analysis can be performed on them without bias simply by "conditioning on the proband", whereas under other ascertainment schemes, correcting for ascertainment becomes difficult or even impossible. This remarkable result holds even if the pedigrees were collected "sequentially"; see below. In addition, Haghighi & Hodge [18] have shown that in a simple model of differential parent-of-origin effects, single ascertainment is the only situation in which the parent-of-origin effect can be ignored. However, note that in the case of "stoppage" (see below), even single ascertainment does not allow for simple solutions.

(c) If the probability that an affected individual will become a proband is neither vanishingly small nor equal to unity, then some but not all affected individuals will be probands, and some but not all families will have multiple probands. This is known as "(incomplete) multiple selection" or "multiple ascertainment", and represents the most complicated situation under the classical model. The full likelihood is derived as follows:

$$\Pr(r \text{ affected children}|\text{sibship ascertained}) = \frac{\Pr(\text{sibship asc'd}|r \text{ aff. ch.})\Pr(r \text{ aff. ch.})}{\Pr(\text{sibship asc'd})}. \quad (5)$$

The first factor in the numerator of (5), Pr(sibship ascertained|r affected children), is derived by recognizing that under the assumptions of the proband model, the probability of an affected child becoming a proband is binomial. Thus, among r affected children, the probability of at least one proband equals one minus the binomial probability of *no* probands, i.e. $1-(1-\pi)^r$. The second factor in the numerator, Pr(r affected children), is the binomial probability for r affected children in an s-child family, $\binom{s}{r}p^r(1-p)^{s-r}$. The denominator of (5) represents the sum of possible numerator terms, summed over $r = 1$ to s, i.e.

$$\sum_{r=1}^{s}[1-(1-\pi)^r]\binom{s}{r}p^r(1-p)^{s-r} = 1-(1-p\pi)^s. \quad (6)$$

Putting together the numerator and denominator of (5) yields the general likelihood for the π-model of ascertainment for a nuclear family, with a single mating type:

$$\Pr_s(r \text{ affected children}|\text{sibship ascertained}) = \frac{\binom{s}{r}p^r(1-p)^{s-r}[1-(1-\pi)^r]}{1-(1-p\pi)^s}. \quad (7)$$

Equations (3) and (4) are special cases of (7). The reader can confirm that setting $\pi = 1$ in (7) will yield (3), whereas letting $\pi \to 0$ in (7) yields (4). The middle column of Table 1 summarizes the features of this more general case.

Equations (3), (4) and (7) give the probabilities for a single s-child family. However, typically a dataset consists of a number of sibships of different sizes, so the likelihoods of the individual families are multiplied across all families. Let R represent the total number of affected children in all the sibships, and let C represent the total number of unaffected children; also, let n_s denote the number of s-child families in the dataset. Then the total log-likelihood over the whole dataset can be written:

$$\log L(p) = R\log p + C\log(1-p) - \sum_s n_s \log[1-(1-p\pi)^s]. \quad (8)$$

The quantity in (8) is then maximized with respect to p to yield the desired maximum likelihood estimates (MLEs).

The π-model can be incorporated into the likelihood for any kind of genetic analysis. Equations (3), (4), and (7) show the likelihoods for the relatively simple situation of nuclear families and a single mating type. Beyond that, numerous extensions are possible. One we mention briefly is simultaneous estimation of p and π. So far we have assumed that the ascertainment probability π has a known value, but in other situations one might need to incorporate π as an unknown parameter (either as another parameter of interest to be estimated, or as a "nuisance parameter"). One approach is to record the actual number of probands a in each family. In (7), replace $[1-(1-\pi^r)]$, i.e. the binomial probability of "at least one proband", with the binomial probability of "exactly a probands": $\binom{r}{a}\pi^a(1-\pi)^{r-a}$. Thus, the appropriate probability is now:

$$\Pr_s(r \text{ affected children}, a \text{ probands}|\text{sibship asc'd}) = \frac{\binom{s}{r}p^r(1-p)^{s-r}\binom{r}{a}\pi^a(1-\pi)^{r-a}}{1-(1-p\pi)^s}. \quad (9)$$

The likelihood is now $L(p,\pi)$ and can be maximized to estimate p and π.

For further details, as well as other extensions, such as multiple parental mating types, multiple loci, complex pedigrees, etc. the interested reader is referred to, for example, [5, 6, 9, 12, 17, 23, 27], and [30].

All these situations involve ascertainment through the children, or what Morton [23] denoted "incomplete selection". If nuclear families are ascertained

through the parents ("complete selection"), then the situation is simpler, and no ascertainment correction is needed. There is no longer any bias, and the segregation ratio can simply be estimated from the standard binomial distribution. For example, if there were 25 families, of 3 children each, and of those 75 children, 35 were affected, then the MLE of segregation ratio p would be simply $\hat{p} = 35/75 = 0.47$. An approximate 95% confidence interval about this estimate, 0.47 ± 1.96 (SEP), where SEP = standard error of the proportion $\approx \sqrt{[(0.47)(0.53)/75]} \approx 0.058$, would not rule out a fully penetrant autosomal dominant disease, for which $p = 0.50$.

Alternative Ascertainment Models

It has been recognized from the beginning (e.g. [23]) that the concept of "proband" is not always applicable in real-life situations and that therefore the π-model does not necessarily hold. Greenberg [15] described realistic ascertainment scenarios in which whole families are ascertained by processes that do not permit one to define "probands" meaningfully. For example, consider the following "screening" scenario. A certain proportion of pregnant women come to a certain clinic to have the fetus screened for a genetic condition. But once a woman has had an affected fetus identified, she is much more likely to return for screening of her subsequent pregnancies. Which affected fetuses are the "probands"? Whether one labels only the first identified one as a proband, or all affected fetuses as probands, one violates the assumption of independence, which is critical to the definition of a proband. Greenberg [15] documented the magnitude of bias and the probability of wrong conclusions that can be introduced into a segregation analysis if the investigator mistakenly models ascertainment as a proband-based process when it is not.

Note also that it is not legitimate simply to designate the *first* member of the family to come to one's attention (that is, the index case) as the "proband". Index cases do not necessarily satisfy the independence requirement. One common and dangerous error is to designate the index case as the proband and then reason that since one has only one index case per family, therefore one has satisfied the conditions of single ascertainment. This is false reasoning [22]. One's ascertainment scheme represents single ascertainment only if the proportionality property described above is satisfied, i.e. only if families with two affected children are twice as likely to be ascertained as those with one affected child, families with three affected children are three times as likely to be ascertained as those with one affected child, etc.

Stene [31] and Ewens & Shute [10] discussed models of family-based ascertainment, for example such that Pr(family is ascertained|r affected members) $\propto r^k$, where k can be any real number. Thus, $k = 0$ corresponds to "complete ascertainment" and $k = 1$ to "single ascertainment" above. However, other values of k, such as $k = 2$ ("quadratic ascertainment"), do not correspond to any values of π in the π-model. This implies that the cases $\pi \to 0$ and $\pi = 1$ do *not* provide "limits" on ascertainment models; rather, there exist numerous possible ascertainment models that do not fit into the "π" framework at all. These family-based models do not designate any individuals as probands, and thus they circumvent the difficulties in the proband concept.

Let us examine the "quadratic ascertainment" model more closely, as an example. In this model a family with, for example, two affected children is four times more likely to be ascertained than one with one affected child (as opposed to being twice as likely, as under single ascertainment). To derive the fundamental ascertainment probability we return to (5). In the numerator, the probability Pr(sibship ascertained|r affected children), which formerly equaled $1 - (1 - \pi)^r$, now equals βr^2, where β is the constant of proportionality. Thus, the desired probability becomes

$$\Pr(r \text{ affected children}|\text{sibship ascertained}) = \frac{\binom{s}{r} p^r (1-p)^{s-r} \beta r^2}{\sum_{r=1}^{s} \binom{s}{r} p^r (1-p)^{s-r} \beta r^2}, \tag{10}$$

instead of (7). Simplifying the denominator and canceling β yields

$$\Pr(r \text{ affected children}|\text{sibship ascertained}) = \frac{\binom{s}{r} p^r (1-p)^{s-r} r^2}{sp[1 + (s-1)p]}. \tag{11}$$

Equation (11) provides an example of how ascertainment models other than the "classical" one can still

be formulated precisely and mathematically. It also illustrates the fact that this "quadratic" model cannot be viewed as a special case of (7), nor can any child in these families be identified as the "proband".

Ginsburg & Axenovich [14] suggested a "cooperative binomial ascertainment" model which allows the ascertainment probability for an individual to depend on the number of potential probands per pedigree.

Even though these alternative ascertainment models do not fit the mold of the classical π-model, they can be incorporated into the likelihood if they are known, as we have seen with quadratic ascertainment above.

Other Methods of Dealing with Ascertainment

As mentioned, if one knows the true ascertainment model, then in theory one can incorporate that probability model into genetic analyses (except in some cases of sequential sampling; see below). The major problem in genetic epidemiology arises when the ascertainment model is not known. Often human families come to the attention of investigators via pathways that are haphazard, ill-defined, or completely unknown. Thus, it is worthwhile to consider alternative ways to deal with ascertainment when the mode of ascertainment is not known or is known only vaguely.

One way to circumvent the whole problem of ascertainment is to condition the likelihood on all information that could possibly have influenced the ascertainment of the family. For example, if one is studying nuclear families and does not know how these families were ascertained, but does know that ascertainment occurred only via the *children*, then one can condition the likelihood on the children's phenotypic information [11]. One then has a *conditional* likelihood, which yields valid maximum likelihood estimators of genetic parameters. Ewens & Shute [11] call this an "ascertainment assumption free" (AAF) approach.

The disadvantage of the AAF approach is that the resultant estimates of genetic parameters may have very large variances. In the above example, once one has conditioned the likelihood on the phenotypes of the children, there is usually little genetic information "left" in the parents. In the extreme, if one could not even be sure through whom ascertainment may have occurred, one would have to condition on the phenotypes of *all* family members, and the variances of the estimators would be infinite.

Thus, in practice this method is not recommended for pure segregation analysis. Where this approach does become practical is when the investigator also has a **marker** locus (*see* **Linkage Analysis, Model-based**) that is reasonably tightly linked to the disease. The resultant analysis can be viewed as a form of combined segregation-and-linkage analysis. If one has a linked marker locus, then even though one has "conditioned out" much of the pure trait information, the information remaining from the *cosegregation* of the marker and the disease can help to determine the mode of inheritance of the disease. It has been shown that conditioning on all trait data in a combined segregation–linkage analysis in this way is equivalent to maximizing the maximum lod score (*see* **Linkage Analysis, Model-based**) over genetic models. For more details, see [3, 7, 16, 19], and [25].

Another way to deal with the ascertainment problem would be to develop good approximations. For example, Sawyer [26] has suggested that treating ascertainment as if it is single, even when it is not single, may provide a reasonable approximation over a broad range of ascertainment schemes. In another approach, Rabinowitz [24] proposes using a pseudo-likelihood, which yields asymptotically unbiased estimates (although variance is inflated) of genetic parameters under a class of ascertainment models broader than those allowed by the classical π-model. More work needs to be done on these and other approximation-based approaches.

Additional Complications

We conclude with two additional complications, whose effects are not yet fully understood: sequential sampling of pedigrees, and stoppage.

Sequential Sampling

We call it sequential sampling when, after a pedigree is ascertained, decisions about who within the pedigree will be included in the study are made in a "proband-dependent" or sequential manner. For example, one possible sequential-sampling rule is: "Include all available first-degree relatives of the

proband. For any of these relatives who are affected, include all *their* available first-degree relatives. Continue until no new person included in this way is affected, i.e. continue until no sampled pedigree member has any additional affected first-degree relatives". In an influential paper, Cannings & Thompson [2] considered pedigrees sampled following this kind of sequential scheme. They showed that under single ascertainment, conditioning the likelihood on observed pedigree structure, and dividing by the population probability of a proband, would yield the correct likelihood (as long as the sequential sampling rules follow certain reasonable commonsense restrictions). Subsequently, Vieland & Hodge [32] demonstrated that the result in [2] holds *only* for single ascertainment and, moreover, that under other modes of ascertainment, the correct likelihood for these kinds of sampling situations inherently cannot be formulated – even when the ascertainment model is known [32]; see also [8]. Rabinowitz's pseudo-likelihood [24] still yields asymptotically unbiased estimates in this situation, but the variance is inflated and must be approximated (see above).

Later, Vieland & Hodge [33] showed that *linkage analysis* is also affected by ascertainment issues when pedigrees are sampled sequentially, though it is not clear whether the effect is ever large enough to be of practical concern [28]. This finding violates the commonly accepted wisdom that linkage analysis is immune to ascertainment issues.

Stoppage

Standard segregation analysis assumes that the observed distribution of sibship sizes in the dataset (FSD = family-size distribution) is independent of the segregation ratio p. However, for certain serious diseases with early onset and diagnosis, e.g. autism, parents may *change* their original desired family size after having one or more affected children, thus violating that assumption. If parents reduce their family size, then the phenomenon is called "stoppage". Thus, stoppage also represents a type of biased ascertainment, in that families will display a smaller number of affected children than they "should", based on the value of the genetic parameter p. This situation has been investigated by Jones & Szatmari [21] and Brookfield et al. [1]. More recently, Slager et al. [29] showed that stoppage can be considered as a special case of sequential within-family sampling [2]. They demonstrated that if families are ascertained completely at "random",[1] the presence of stoppage does not bias estimates of p, but under any other ascertainment schemes, including those in (3), (4), and (7) above, stoppage introduces an additional bias, which can be quite large. Slager et al. [29] derived the full correct likelihood for a stoppage model in which after the birth of each affected child, there is a stoppage probability d that the parents will have no more children, even if they had originally intended a larger family. They showed that unless one already knows the FSD of the population from which the data are drawn, correcting for stoppage is difficult. Even when there is single ascertainment, the likelihood does not simplify, unlike in so many other situations involving single ascertainment.

Conclusions

In this article we have defined the ascertainment problem and tried to convey some idea of its nature and importance. We have shown how, starting with basic probability principles, one can incorporate ascertainment assumptions into genetic analysis, starting with simple ascertainment models and proceeding to more complex ones. However, some situations remain difficult or even intractable, and research still remains to be done in this area.

References

[1] Brookfield, J.F.Y., Pollitt, R.J. & Young, I.D. (1988). Family size limitation: a method for demonstrating recessive inheritance, *Journal of Medical Genetics* **25**, 181–185.

[2] Cannings, C. & Thompson, E.A. (1977). Ascertainment in the sequential sampling of pedigrees, *Clinical Genetics* **12**, 208–212.

[3] Clerget-Darpoux, F. & Bonaïti-Pellié, C. (1992). Strategies on marker information for the study of human diseases, *Annals of Human Genetics* **46**, 145–153.

[4] Edwards, J.H. (1971). The analysis of X-linkage, *Annals of Human Genetics* **34**, 229–250.

[1] Random ascertainment is the situation in which every family – even those with *no* affected children – has an equal probability of being ascertained, independent of the number of affected children. This situation could arise during epidemiological survey sampling; or if families are ascertained through affected parents, rather than through affected children.

[5] Elandt-Johnson, R.C. (1971). *Probability Models and Statistical Methods in Genetics*. Wiley, New York, Chapter 18.

[6] Elston, R.C. (1980). gregation analysis, in *Current Developments in Anthropological Genetics*, Vol. 1, J.H. Mielke & M.H. Crawford, eds. Plenum, New York.

[7] Elston, R.C. (1989). Man bites dog? The validity of maximizing lod scores to determine mode of inheritance, *American Journal of Medical Genetics* **34**, 487–488.

[8] Elston, R.C. (1995). Twixt cup and lip: how intractable is the ascertainment problem?, *American Journal of Human Genetics* **56**, 15–17.

[9] Elston, R.C. & Sobel, E. (1979). Sampling considerations in the gathering and analysis of pedigree data, *American Journal of Human Genetics* **31**, 62–69.

[10] Ewens, W.J. & Shute, N.C.E. (1986). The limits of ascertainment, *Annals of Human Genetics* **50**, 399–402.

[11] Ewens, W.J. & Shute, N.C.E. (1986). A resolution of the ascertainment sampling problem. I. Theory, *Theoretical Population Biology* **30**, 388–412.

[12] George, V.T. & Elston, R.C. (1991). Ascertainment: an overview of the classical segregation analysis model for independent sibships, *Biometrics Journal* **33**, 741–753.

[13] Gershon, E.S. & Matthysse, S. (1977). X-linkage: ascertainment through doubly ill probands, *Journal of Psychiatric Research* **13**, 161–168.

[14] Ginsburg, E.Kh. & Axenovich, T.I. (1992). A cooperative binomial ascertainment model, *American Journal of Human Genetics* **51**, 1156–1160.

[15] Greenberg, D.A. (1986). The effect of proband designation on segregation analysis, *American Journal of Human Genetics* **39**, 329–339.

[16] Greenberg, D.A. (1989). Inferring mode of inheritance by comparison of lod scores, *American Journal of Medical Genetics* **35**, 480–486.

[17] Greenberg, D.A. & Lange, K.L. (1982). A maximum likelihood test of the two locus model for coeliac disease, *American Journal of Medical Genetics* **12**, 75–82.

[18] Haghighi, F. & Hodge, S.E. (2001). Likelihood formulation of parent-of-origin effect on segregation analysis, including ascertainment *American Journal of Human Genetics* **70**, 142–156.

[19] Hodge, S.E. & Elston, R.C. (1994). Lods, wrods, and mods: the interpretation of lod scores calculated under different models, *Genetic Epidemiology* **11**, 329–342.

[20] Hodge, S.E. & Vieland, V.J. (1996). The essence of single ascertainment, *Genetics* **144**, 1215–1223.

[21] Jones, M.B. & Szatmari, P. (1988). Stoppage rules and genetic studies of autism, *Journal of Autism and Developmental Disorders* **18**, 31–40.

[22] Marazita, M.L. (1995). Defining "proband", *American Journal of Human Genetics* **57**, 981–982.

[23] Morton, N.E. (1959). Genetic tests under incomplete ascertainment, *American Journal of Human Genetics* **11**, 1–16.

[24] Rabinowitz, D. (1996). A pseudolikelihood approach to correcting for ascertainment bias in family studies, *American Journal of Human Genetics* **59**, 726–730.

[25] Risch, N. (1984). Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type 1 diabetes, *American Journal of Human Genetics* **36**, 363–386.

[26] Sawyer, S. (1990). Maximum likelihood estimators for incorrect models, with an application to ascertainment bias for continuous characters, *Theoretical Population Biology* **38**, 351–366.

[27] Sham, P. (1998). *Statistics in Human Genetics* Arnold, London.

[28] Slager, S.L. & Vieland, V.J. (1997). Investigating the numerical effects of ascertainment bias in linkage analysis: development of methods and preliminary results, *Genetic Epidemiology* **14**, 1119–1124.

[29] Slager, S.L., Foroud, T., Haghighi, F., Spence, M.A. & Hodge, S.E. (2001). Stoppage: an issue for segregation analysis, *Genetic Epidemiology* **20**, 328–339.

[30] Spence, M.A. & Hodge, S.E. (1996). gregation analysis, in *Emery and Rimoin's Principles and Practice of Medical Genetics*, 3rd Ed., D.L. Rimoin, J.M. Connor, R.E. Pyeritz & A.E.H. Emery, eds. Churchill Livingstone, New York, Chapter 7, pp. 103–109.

[31] Stene, J. (1978). Choice of ascertainment model I. Discrimination between single-proband models by means of birth order data, *Annals of Human Genetics* **42**, 219–229.

[32] Vieland, V.J. & Hodge, S.E. (1995). Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework, *American Journal of Human Genetics* **56**, 33–43.

[33] Vieland, V.J. & Hodge, S.E. (1996). The problem of ascertainment for linkage analysis, *American Journal of Human Genetics* **58**, 1072–1084.

[34] Weinberg, W. (1928). Mathematische Grundlagen der Probandenmethode, *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* **48**, 179–228.

S.E. HODGE

Aspin–Welch Test

The assumptions used in deriving the independent samples (or two-sample) t test are: (i) within each of two groups the observations are independently, identically, and normally distributed; (ii) observations are independently distributed across groups, and (iii) the populations from which the observations in the two groups are drawn have equal variances (*see* **Student's *t* Distribution**; **Behrens–Fisher Problem**). If (iii) is violated, then the actual type I error rate, τ, is near the nominal type I error rate, α, provided the group sizes are equal and sufficiently large (*see* **Hypothesis Testing**). Results in [8] suggest that the required sample size is between 8 and 15, depending on how large a discrepancy between τ and α one will tolerate. If (iii) is violated and the group sizes are unequal, then $\tau < \alpha$ if the larger sample comes from the population with the larger variance and $\tau > \alpha$ if the smaller sample comes from this population.

An alternative to the test statistic used in the independent samples t test is

$$t' = \frac{\overline{x}_1 - \overline{x}_2}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{1/2}}.$$

Welch [9] proposed using t' with the approximate **degrees of freedom** (APDF) critical value $t_{1-\alpha}(\nu)$, where

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}.$$

In practice ν is estimated by substituting sample variances for population variances.

Welch [10] proposed a method for approximating the critical value for t' by a power series in $1/f_i = 1/(n_i - 1)$ and presented the solution to order $(1/f_i)^2$:

$$z\left[1 + \frac{(1+z^2)V_{21}}{4} - \frac{(1+z^2)V_{22}}{2} + \frac{(3+5z^2+z^4)V_{32}}{3} - \frac{(15+32z^2+9z^4)V_{21}^2}{32}\right],$$

where

$$V_{ru} = \frac{\left(\sum_{i=1}^{2} \frac{(s_i^2/n_i)^r}{f_i^u}\right)}{\left(\sum_{i=1}^{2} s_i^2/n_i\right)^r}$$

and $z = z_{1-\alpha}$. Aspin [2] presented the solution to order $(1/f_i)^4$ and Aspin [3] presented an abbreviated table of fourth-order critical values.

Results in Lee & Gurland [7] indicate that with small sample sizes (≤ 9) τ is nearer to α when the fourth-order critical value is used than it is when the second-order or the APDF critical values are used. However, all three critical values controlled τ fairly well. Algina et al. [1] investigated larger group sizes and reported that the APDF critical value and the second-order critical value result in estimates of τ that typically agree to the third or fourth decimal place.

Using t' to test equality of means has two shortcomings. Yuen [14] demonstrated lower **power** for t' when the data are long tailed. Both Yuen's tests on trimmed means and Wilcox's [13] test on one-step M-estimates of location (*see* **Robustness**) can be used to address this problem. Wilcox [12] demonstrated that when the data in either or both groups are drawn from skewed populations, the numerator and denominator of t' are not independent. The **simulation** in Algina et al.indicates that as a result of the lack of independence large discrepancies can occur between τ and α. When the larger group is drawn from the population with the smaller variance, τ can be much larger than α. If one is willing to use trimmed means or one-step M-estimates when the data are skewed, then the second problem can be addressed by using Yuen's and Wilcox's tests, respectively.

Welch [11] generalized the APDF approach to the one-way layout with $G \geq 3$ groups and Johansen [6] further generalized the approach to multi-way layouts and to test multivariate hypotheses. James [4] generalized the series approach to the one-way layout and James [5] generalized the approach to test multivariate hypotheses for the one-way layout.

References

[1] Algina, J., Oshima, T.C. & Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second-order

test under nonnormality and inequality of variance when there are two groups, *Journal of Educational and Behavioral Statistics* **19**, 275–292.

[2] Aspin, A.A. (1947). An examination and further development of a formula arising in the problem of comparing two mean values, *Biometrika* **35**, 88–96.

[3] Aspin, A.A. (1949). Tables for use in comparison whose accuracy involves two variances, separately estimated, *Biometrika* **36**, 290–293.

[4] James, G.S. (1951). The comparison of several groups of observations when the ratios of population variances are unknown, *Biometrika* **38**, 324–329.

[5] James, G.S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown, *Biometrika* **41**, 19–43.

[6] Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression, *Biometrika* **67**, 85–92.

[7] Lee, A.F.S. & Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances, *Journal of the American Statistical Association* **70**, 933–941.

[8] Ramsey, P.H. (1980). Exact Type I error rates for robustness of Student's t test with unequal variances, *Journal of Educational Statistics* **5**, 337–349.

[9] Welch, B.L. (1938). The significance of the difference between two means when the population variances are unequal, *Biometrika* **29**, 350–362.

[10] Welch, B.L. (1947). The generalization of "Students" problem when several different population variances are involved, *Biometrika* **34**, 23–35.

[11] Welch, B.L. (1951). On the comparison of several mean values: an alternative approach, *Biometrika* **38**, 330–336.

[12] Wilcox, R.R. (1990). Comparing the means of two independent groups, *Biometrical Journal* **32**, 771–780.

[13] Wilcox, R.R. (1992). Comparing one-step m-estimators of location corresponding to two independent groups, *Psychometrika* **58**, 71–78.

[14] Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances, *Biometrika* **61**, 165–176.

(*See also* **Scedasticity**)

JAMES ALGINA

Association, Measures of

For **categorical data**, measures of **association** are used to quantify the degree of relationship between variables. In particular, a high degree of association between two variables indicates that knowledge about the level of one variable increases the ability to predict accurately the level of the other variable; a low level of association would indicate that the two variables tend to be independent of one another.

Interest in measures of association arose as early as the late 1800s, instigated by the study of meteorological phenomena, smallpox vaccinations, and anthropology. Discussion of such measures began in earnest in the early 1900s. Pearson [28, 30] believed that the measure of association should be based on the correlation for a presumed bivariate continuous distribution that underlies the contingency table; his *tetrachoric correlation* for **two-by-two contingency tables** and *contingency coefficient* for $I \times J$ tables are such measures. Relying on the inherent discreteness of the categorical variables involved, Yule [37, 38] developed measures of association that assumed nothing about underlying continuous distributions, such as Q (now called *Yule's Q*), named in honor of the Belgian statistician **Quetelet**. Goodman & Kruskal [17] wrote a series of four landmark papers on measures of association for cross classifications emphasizing *interpretable* measures, including a thorough history and bibliography on the subject (*see* **Goodman–Kruskal Measures of Association**). We will begin our discussion with 2×2 tables and then cover the more complicated case of $I \times J$ tables.

2 × 2 Tables

Several measures of association for 2×2 tables are based on the value of the Pearson **chi-square** (χ^2) statistic. Note that it is inappropriate to use the value of χ^2 itself as a measure of association for two dichotomous variables because it is a function of the sample size (see [10, Section 21]); Fleiss [11, p. 59] provides an example of this point.

In what follows, n_{ij} represents the frequency for the ith level of the row variable and the jth level of the column variable, $i, j, = 1, 2$, and a "+" in the subscript represents summation over the subscript replaced (e.g. $n_{1+} = n_{11} + n_{12}$). Then, one of the χ^2-based measures of association is the *phi coefficient*:

$$\phi = \left(\frac{\chi^2}{n_{++}}\right)^{1/2},$$

where

$$\chi^2 = \frac{n_{++}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}.$$

Two additional measures based on ϕ are the *mean square contingency*, ϕ^2 [8, 28], and Pearson's [28] *coefficient of contingency*:

$$C = \left[\frac{\phi^2}{(1+\phi^2)}\right]^{1/2}.$$

Values of ϕ close to zero indicate very little association, while values of ϕ close to one imply close to perfect predictability. The maximum possible value of ϕ is 1 if the marginal distributions are equal, but less than 1 otherwise. Fleiss [11] provides, as a rule of thumb, that any value of ϕ less than 0.30 or 0.35 may be taken to indicate no more than trivial association. There are some disadvantages of the phi coefficient as a measure of association. For instance, the value of ϕ in **case–control studies** is not comparable to the value of ϕ in **cohort studies** [11, Chapter 6]. Also, if one or both of the characteristics being studied is obtained by dichotomizing a continuous random variable, then the value of ϕ is strongly dependent on where the continuous variable is divided [6] (*see* **Categorizing Continuous Variables**). This lack of invariance of the phi coefficient, among other reasons, led Goodman and Kruskal to recommend against the use of χ^2-like statistics as measures of association, except perhaps as they relate to **loss functions** and proportional prediction [17, pp. 10, 26, and 29–30].

A basic measure of association in 2×2 tables is the *cross-product ratio* or **odds ratio** (see [37] and [27]). The odds ratio is defined as

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}},$$

where p_{ij} represents the probability associated with the ith row and jth column of the contingency table, $i, j = 1, 2$. The odds ratio is useful as a measure of association because (i) it is appropriate for a number of different sampling models, (ii) it serves as the

building block for loglinear model theory, and (iii) it has a number of important properties, as follows:

1. invariance (except perhaps for direction) under interchange of rows and/or columns,
2. invariance (except perhaps for direction) under row and/or column multiplication,
3. clear interpretation: p_{11}/p_{12} is the odds of an item being in the first column given that it is in the first row, and p_{21}/p_{22} is the corresponding odds for the second row; then α is the ratio of these two odds, just as the name implies, and
4. usefulness in $I \times J$ tables by considering several 2×2 tables.

The value of α falls in $[0, \infty)$ and is symmetric in the sense that two odds ratios, α_1 and α_2, with $\log \alpha_1 = -\log \alpha_2$, represent the same degree of association, but in opposite directions. If $\alpha = 1$, then the row variable and column variable are independent; if $\alpha \neq 1$, then they are associated or dependent. The observed odds ratio,

$$a = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

is the maximum likelihood estimate of α. To construct confidence intervals, note that $\ln a$ is normally distributed with mean $\ln \alpha$ for large samples, and

$$\text{se}(\ln a) = (n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1})^{1/2}.$$

For reviews of alternative ways of obtaining confidence intervals for α, see [13] and [11]. Some authors prefer to add a continuity correction of 0.5 to each of the cell frequencies in the expressions above [11] (*see* **Yates's Continuity Correction**). Other suggested improved estimates are given in [4], [12], and [14].

A large number of the measures of association for 2×2 tables discussed in [17] are monotone functions of the odds ratio. For example, Yule's Q can be written

$$Q = \frac{(a-1)}{(a+1)}.$$

Edwards [9] showed that the odds ratio and functions of it are the only statistics that are invariant to both row/column interchange and to multiplication within rows/columns by a constant, and recommends that they be used to measure association in 2×2 tables. See [3] for related results concerning $I \times J$ tables.

Simple measures, such as the difference of proportions, $(p_{11}/p_{1+}) - (p_{21}/p_{2+})$, and **relative risk**, $(p_{11}/p_{1+})/(p_{21}/p_{2+})$, have a long history of use. Goodman & Kruskal [17] noted that the relative risk was used by Quetelet in 1849. The magnitude of the relative risk is similar to that of the odds ratio whenever the probability of response level one is close to zero for both groups. For further discussion of these measures, see [2] and [11].

While single numbers such as the odds ratio can summarize the association in 2×2 tables, it is difficult to summarize the association in $I \times J$ tables for $I > 2$ and/or $J > 2$ by a single number without some loss of information.

$I \times J$ Tables

The above measures can be generalized to $I \times J$ tables, with $I > 2$ and/or $J > 2$. For instance, the association in an $I \times J$ table can be described with a set of $(I-1)(J-1)$ odds ratios. Such a set is not unique; one basic set is

$$\alpha_{ij} = \frac{p_{ij}p_{IJ}}{p_{Ij}p_{iJ}},$$

$i = 1, 2, \ldots, I-1$ and $j = 1, 2, \ldots, J-1$. Or, the set of *local* odds ratios is

$$\alpha_{ij} = \frac{p_{ij}p_{i+1,j+1}}{p_{i,j+1}p_{i+1,j}},$$

$i = 1, 2, \ldots, I-1$ and $j = 1, 2, \ldots, J-1$. Such sets of odds ratios can be used to describe features of the association in the table.

The usefulness and interpretation of measures of association in $I \times J$ tables depend upon the nature of the variables involved (*see* **Measurement Scale**). Generally, a categorical variable is *ordinal* (*see* **Ordered Categorical Data**), in which case there is a natural order associated with its levels, or **nominal**, in which case there is no natural order associated with the levels. Examples of ordinal variables are *severity of disease* (mild, medium, severe) and *opinion* (strongly disagree, disagree, neutral, agree, strongly agree). For nominal variables, we might have *political party affiliation* or *diagnosis*.

Ordinal Variables

For ordinal variables, measures of association describe the degree to which the relationship is

monotone, i.e. whether Y tends to increase as X does. More specifically, a pair of subjects is *concordant* if the subject ranking higher on variable X also ranks higher on variable Y; they are *discordant* if the subject ranking higher on X ranks lower on Y (*see* **Ranks**; **Rank Correlation**). The pair is *tied* if the subjects have the same classification on X and/or Y. For a pair of observations, the probability of concordance is

$$\Pi_c = 2\sum_i \sum_j p_{ij} \left(\sum_{h>i} \sum_{k>j} p_{hk} \right),$$

and the probability of discordance is

$$\Pi_d = 2\sum_i \sum_j p_{ij} \left(\sum_{h>i} \sum_{k<j} p_{hk} \right).$$

The difference, $\Pi_c - \Pi_d$, is used in several measures of association for ordinal variables; the association is positive if $\Pi_c - \Pi_d > 0$ and negative if $\Pi_c - \Pi_d < 0$.

For those pairs that are untied on both variables, the probability of concordance is $\Pi_c/(\Pi_c + \Pi_d)$ and the probability of discordance is $\Pi_d/(\Pi_c + \Pi_d)$. Then, a measure proposed by Goodman & Kruskal [17] is *gamma*, where

$$\gamma = \frac{(\Pi_c - \Pi_d)}{(\Pi_c + \Pi_d)}.$$

This is simply the difference in the above two probabilities. In fact, γ tells us how much more probable it is to get like than unlike orders in the two classifications when two individuals are chosen at random from the population.

If C is the observed number of concordant pairs and D is the observed number of discordant pairs, then the sample version of gamma is $g = (C-D)/(C+D)$. Note that $-1 \le \gamma \le 1; \gamma = 1$ if $\Pi_d = 0$ and $\gamma = -1$ if $\Pi_c = 0$. That is, $|\gamma| = 1$ under monotonicity. Independence implies $\gamma = 0$; however, the converse is not true except in the 2×2 case. And, γ is invariant (except for sign) to the reversal in the category orderings of one variable. For 2×2 tables, γ is the same as Yule's Q, and hence is related to the odds ratio, α. In fact, γ is a strictly monotone transformation of α from the $[0, \infty)$ scale onto the $[-1, +1]$ scale.

There are a number of other ordinal measures of association that may be useful in a given application, such as Kendall's [19, 20] τ_b and Stuart's [35] τ_c. For a survey of such measures, see [24], [20], [17], and [1, Chapters 9 and 10].

Nominal Variables

For nominal variables, the concepts of positive/negative association and monotonicity are no longer appropriate for measuring the relationship. Instead, the most interpretable indices are those that describe the proportional reduction in variance from the marginal distribution to the conditional distributions of the response. One such measure is *Goodman & Kruskal's tau* [17] (also called the *concentration coefficient*):

$$\tau = \frac{\sum_i \sum_j p_{ij}^2/p_{i+} - \sum_j p_{+j}^2}{1 - \sum_j p_{+j}^2}.$$

Goodman & Kruskal provided the following interpretation: τ is the proportional reduction in the probability of an incorrect guess obtained by making predictions on Y using the classification on X. A large value of τ corresponds to a strong association because it indicates that we can guess Y much better when we know X than when we do not know X.

An alternative index for proportional reduction in variation, called the *uncertainty coefficient*, is proposed by Theil [36]:

$$U = -\frac{\sum_i \sum_j p_{ij} \ln(p_{ij}/p_{i+}p_{+j})}{\sum_j p_{+j} \ln(p_{+j})}.$$

Both τ and U take values on [0, 1], and $\tau = U = 0$ is equivalent to the independence of X and Y. The condition that, for each i, $p_{ij}/p_{i+} = 1$ for some j (no conditional variation) is equivalent to $\tau = U = 1$.

The larger τ or U is, the stronger is the association between the variables. However, note that these measures tend to decrease as the number of categories of the response variable increases.

$I \times J$ Tables, Other Cases

The measures of ordinal association mentioned above are appropriate if one variable is ordinal and the other is nominal with two categories. If one variable is ordinal and the other is nominal with more than two categories, then *ridits* may be used to measure the association; see [5, 11, Section 9.4, 25, 32, and 1, Sections 9.3 and 10.2].

To determine the association between, say, the ith level of A (row variable) and the B (column) polytomy, simply collapse all levels of A other than the ith to form a variable with two levels: the ith level and all other levels. Then the resulting $2 \times J$ table can be analyzed. This suggestion was made by Pearson [29]. This can be generalized to the study of the association between a particular set of A categories and a particular set of B categories; see [17, p. 38].

When two polytomies, X and Y, are nominal and *asymmetric* (i.e. X precedes Y chronologically, causally, or otherwise), Goodman & Kruskal [17] propose as a measure of association,

$$\lambda_y = \frac{\sum_i \max_j(p_{ij}) - \max_j(p_{+j})}{1 - \max_j(p_{+j})}.$$

This measure, first suggested by Guttman [18], gives the proportion of errors that can be eliminated by taking account of knowledge of the X classifications of individuals. Similarly, λ_x may be defined, as well as a measure for the symmetrical case, λ. See [17, pp. 10–15] for further discussion of these. Another measure, Somers' D [34], is an asymmetric modification of Kendall's τ_b (see [17]). Note that the measures τ and U also treat the variables asymmetrically.

Because λ_x, λ_y, and λ depend on marginal frequencies, one may wish to weight columns or rows (**standardization** of marginal distributions); see [38]. For example, it may be reasonable to base the association measure on a table for which all X-levels are equiprobable; an example of this is given in [17, pp. 15–16]. Other forms of standardization are given in [27].

More than Two Factors

When there are more than two polytomies, the *multiple association* between A and all other variables can be assessed by forming a two-way table with rows representing the A-polytomy and columns representing all possible combinations of levels of the remaining polytomies. The *partial association* in this case is the association between two of the variables with the effect of the others averaged out in some sense. Goodman & Kruskal [17, pp. 30–31] and Kendall & Stuart [21, pp. 571–575] discuss such measures.

When combining evidence from several four-fold tables, it is often of interest to know if the degree of association is consistent from one group to another, and if so, what the best estimate of the common value for the measure of association is. Mantel & Haenszel [26] proposed a measure that estimates a common odds ratio for several 2×2 tables (*see* **Mantel–Haenszel Methods**). Fleiss [11, Chapter 10] and Sokal & Rohlf [33] discuss these issues further.

Miscellaneous Topics

The traditional χ^2-like measures of association, unlike the λ and γ measures discussed by Goodman & Kruskal [17], assume the value zero if and only if there is independence in the cross classification. Note that independence is sufficient for $\lambda = 0$ and $\gamma = 0$, but not necessary. Goodman & Kruskal [17, p. 52] argue that λ and γ measure one dimension or aspect of association, and hence may be zero even when there is association along some other dimension.

A **misclassification error** can alter the degree or even direction of an association. A number of researchers have analyzed the effects of misclassification on measures of association; see [31], [22], [23], [15], and [7].

Measures of **agreement** are used in cases where the classes are the same for the two polytomies ($I \times I$ table) but differ in that assignment to level depends on which of two methods of assignment is used. For example, two psychiatrists may independently classify each of n_{++} subjects according to four diagnoses (4×4 table), and a measure of the agreement between their diagnoses is of interest. Measures of *association* differ from measures of *agreement* because there can be strong association without strong agreement in a table.

The asymptotic sampling theory for association measures is based on the sampling scheme used

to generate the contingency table and the sample estimator for the measure. The most common sampling schemes for contingency tables are based on the **multinomial distribution** or the product-multinomial distribution (i.e. independent multinomial distributions for each row or column). The estimator used for a measure of association is simply the sample analogue of the population measure; in almost every instance this is the maximum likelihood estimator. Goodman & Kruskal [17, pp. 76–146] develop formulas for the standard errors of some of their coefficients under various sampling models; also, see [16]. While most of the formulas are complex, major statistical software packages routinely provide standard errors for many of the measures of association (*see* **Software, Biostatistical**).

References

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

[2] Agresti, A. (2002). *Categorical Data Analysis* 2nd Ed. Wiley, New York.

[3] Altham, P.M.E. (1970). The measurement of association of rows and columns for an $r \times s$ contingency table, *Journal of the Royal Statistical Society, Series B* **32**, 63–73.

[4] Anscombe, F.J. (1956). On estimating binomial response relations, *Biometrika* **43**, 461–464.

[5] Bross, I.D.J. (1958). How to use ridit analysis, *Biometrics* **14**, 18–38.

[6] Carroll, J.B. (1961). The nature of the data, or how to choose a correlation coefficient, *Psychometrika* **26**, 347–372.

[7] Copeland, K.T., Checkoway, H., McMichael, A.J. & Holbrook, R.H. (1977). Bias due to misclassification in the estimation of relative risk, *American Journal of Epidemiology* **105**, 488–495.

[8] Doolittle, M.H. (1885). The verification of predictions (abstract), *Bulletin of the Philosophical Society of Washington* **7**, 122–127.

[9] Edwards, A.W.F. (1963). The measure of association in a 2×2 table, *Journal of the Royal Statistical Society, Series A* **126**, 109–114.

[10] Fisher, R.A. (1948). *Statistical Methods for Research Workers*, 10th Ed. Hafner, New York.

[11] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[12] Gart, J.J. (1966). Alternative analyses of contingency tables, *Journal of the Royal Statistical Society, Series B* **28**, 164–179.

[13] Gart, J.J. (1971). The comparison of proportions: a review of significance tests, confidence intervals and adjustments for stratification, *Review of the International Statistical Institute* **39**, 148–169.

[14] Gart, J.J. & Zweifel, J.R. (1967). On the bias of various estimators of the logit and its variance, with application to quantal bioassay, *Biometrika* **54**, 181–187.

[15] Goldberg, J.D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table, *Journal of the American Statistical Association* **70**, 561–567.

[16] Goodman, L.A. (1964). Simultaneous confidence intervals for contrasts among multinomial populations, *Annals of Mathematical Statistics* **35**, 716–725.

[17] Goodman, L.A. & Kruskal, W.H. (1979). *Measures of Association for Cross Classifications*. Springer-Verlag, New York.

[18] Guttman, L. (1941). An outline of the statistical theory of prediction. Supplementary Study B-1, in *The Prediction of Personal Adjustment*, Horst, Paul and others, eds. Bulletin 48, Social Science Research Council, New York, pp. 253–318.

[19] Kendall, M.G. (1945). The treatment of ties in rank problems, *Biometrika* **33**, 239–251.

[20] Kendall, M.G. (1970). *Rank Correlation Methods*, 4th Ed. Hafner, New York.

[21] Kendall, M.G. & Stuart, A. (1979). *The Advanced Theory of Statistics*, Vol. 2, 4th Ed. Griffin, London.

[22] Keyes, A. & Kihlberg, J.K. (1963). The effect of misclassification on estimated relative prevalence of a characteristic, *American Journal of Public Health* **53**, 1656–1665.

[23] Koch, G.G. (1969). The effect of non-sampling errors on measures of association in 2×2 contingency tables, *Journal of the American Statistical Association* **64**, 852–863.

[24] Kruskal, W.H. (1958). Ordinal measures of association, *Journal of the American Statistical Association* **53**, 814–861.

[25] Landis, J.R. Heyman, E.R. & Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests, *International Statistical Review* **46**, 237–254.

[26] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[27] Mosteller, F. (1968). Association and estimation in contingency tables, *Journal of the American Statistical Association* **63**, 1–28.

[28] Pearson, K. 1904. Mathematical contributions to the theory of evolution XIII: on the theory of contingency and its relation to association and normal correlation, *Draper's Co. Research Memoirs, Biometric Series*, no. 1.

[29] Pearson, K. (1906). On a coefficient of class heterogeneity or divergence, *Biometrika* **5**, 198–203.

[30] Pearson, K. (1913). On the probable error of a correlation coefficient as found from a fourfold table, *Biometrika* **9**, 22–27.

[31] Rogot, E. (1961). A note on measurement errors and detecting real differences, *Journal of the American Statistical Association* **56**, 314–319.

[32] Semenya, K., Koch, G.G., Stokes, M.E. & Forthofer, R.N. (1983). Linear models methods for some rank function analyses of ordinal categorical data, *Communications in Statistics – Theory and Methods* **12**, 1277–1298.

[33] Sokal, R.R. & Rohlf, F.J. (1995). *Biometry*, 3rd Ed. Freeman, New York.

[34] Somers, R.H. (1962). A new asymmetric measure of association for ordinal variables, *American Sociological Review* **27**, 799–811.

[35] Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables, *Biometrika* **40**, 105–110.

[36] Theil, H. (1970). On the estimation of relationships involving qualitative variables, *American Journal of Sociology* **76**, 103–154.

[37] Yule, G.U. (1900). On the association of attributes in statistics, *Philosophical Transactions of the Royal Society, Series A* **194**, 257–319.

[38] Yule, G.U. (1912). On the methods of measuring association between two attributes (with discussion), *Journal of the Royal Statistical Society* **75**, 579–642.

(*See also* **Contingency Table**; **Loglinear Model**)

HARRY J. KHAMIS

Association

Two variables may be said to be *mutually dependent* if the distribution of values of one variable depends on the value taken by the other. *Association* is a common form of dependence affecting changes in the mean values, or some other measures of level of response; association thus implies that the general level of one variable changes according to the value of the other variable. In informal usage it often carries the further implication that the relationship is monotone; for instance, in statements such as "Exposure to factor *A* is associated with an increased risk of disease *B*".

A well-known and very important precept is that association does not imply **causation**: two variables may be associated merely because each is in turn associated with one or more other variables, although the two in question are not causally related to each other (*see* **Correlation**). The term "association" is closely related to, and almost synonymous with, "correlation". The distinction is that correlation is a measure of closeness to a linear relationship, between either the original variables or some transformation of them, whereas a close association between quantitative variables may be markedly nonlinear. Association need not be monotone: one variable may tend to rise and then fall as the other increases (*see* **Correlation**, Figure 2).

The term "association" is also used to describe relationships between categorical variables (*see* **Association, Measures of**). With ordinal variables, the interest may lie in the degree of concordance; that is, whether movements along the scale of categories of one variable tend to be accompanied by similar movements in the other variable. When one variable is nominal the order of its categories is undefined, and association here would imply merely that the distribution of the other variable varied with the category of the nominal variable. An example is the association between blood groups (a nominal variable) and a particular disease, where the prevalence or incidence of the disease may vary from one blood group to another.

Association between two or more random variables implies departure from *independence* (*see* **Statistical Dependence and Independence**). For a description of concepts and measures of dependence between random variables, see [2] and [3]. Most measures of dependence, defined over the range (0, 1), are such that nonzero values imply positive or negative association. The term "association" has been used [1] for a specific form of dependence between a set of random variables, requiring that any nondecreasing functions of any pair of variables in the set should be nonnegatively correlated.

References

[1] Esery, J., Proschan, F. & Walkup, W. (1967). Association of random variables with applications, *Annals of Mathematical Statistics* **38**, 1466–1474.

[2] Jogdeo, K. (1982). Dependence, concepts of, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 324–334.

[3] Lancaster, H.O. (1982). Dependence, measures and indices of, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 334–339.

PETER ARMITAGE

Assortative Mating

Assortative mating for a characteristic is a process whereby biological parents are more similar (or occasionally more dissimilar) for a phenotypic trait than they would be if mating occurred at random in the population. The two characteristics usually given as examples in human populations are height and intelligence. There are many others; see, for example, [5, 8] and [11]. The phenotypic similarity can induce changes in genotypic relationships. Evaluating these changes is challenging, and usually involves complex mathematical modeling of genetic inheritance and environmental effects (*see* **Gene-environment Interaction**). Furthermore, genetic and environmental factors that affect those characteristics for which there is assortative mating may also affect other characteristics, producing an observed correlation between parents for these other characteristics.

Much of the early literature in **population genetics** theory considered the effects of assortative mating on population characteristics, and can be traced back to Jennings [7] for a phenotypic trait being determined by a single locus, to Fisher's notoriously difficult paper [4] for a continuous characteristic arising from a multifactorial model, and to Wright's [15] formulation of path analysis (which underlies the approach taken today by some behavioral geneticists). References to some of the more extensive literature of the 1960s and 1970s can be found in population genetics texts, such as Crow & Kimura [1] and Ewens [2]. Recent published scientific literature appears to be more concerned with assortative mating in animal populations. For human behavioral traits, assortative mating is still being incorporated in models based essentially on a path analysis approach.

The problem in accommodating assortative mating is that the choice of a mathematical model is not very clear-cut. Certainly, from the geneticist's viewpoint the choice is not as noncontroversial as it would be if, say, the nonrandom process had been the result of **inbreeding**. The essential difficulty in modeling assortative mating is that different social and behavioral patterns in the population can produce different outcomes. There is no agreement concerning these patterns. Furthermore, many assumptions must be made, either implicitly or explicitly. Then, if one takes the traditional, population genetic, approach of following the process to equilibrium, the consequences may lack biological (and perhaps social) relevance or realism. Some of the difficulties will be overviewed briefly in the context of assortative mating for a multifactorial trait, based on Fisher's classical multifactorial model.

First, a general mathematical representation is given that distinguishes between two-sided assortative mating, where each sex is involved in the choice of mates, and one-sided assortative mating, where only one sex selects its mates. Let m and m_{p} represent analogous phenotypic trait values in the male and male-parent populations, and, similarly, f and $f_{\uparrow \mathrm{p}}$ for the female and female-parent values. Then for two-sided assortative mating we can write

$$p(m_{\mathrm{p}}, f_{\uparrow \mathrm{p}}) = q(m)q(f)a(m, f), \tag{1}$$

where p and q represent appropriate probability functions, and a the assortation function. Under the hypothesis of random mating, $a = 1$. Otherwise, a could be quite complex; for instance, it might represent a sum over a proportion of the population mating assortatively with respect to this characteristic and the remainder at random. For one-sided assortative mating, we can write

$$p(m_{\mathrm{p}}, f_{\uparrow \mathrm{p}}) = q(m)Q(f|m), \tag{2}$$

where Q is a conditional probability or density. Such a model is usually not considered particularly realistic for human populations.

Now consider Fisher's multifactor model for the mode of inheritance of a continuous trait [9]. In this model, it is supposed that the measurements x and y of males and females, respectively, are the result of a large number of independently segregating factors, and are thus normally distributed. Taking the means as zero and variances as σ^2, the probability that a measurement lies in the range $(x, x + \mathrm{d}x)$ is given by $(2\pi\sigma^2)^{-1/2} \exp(-x^2/2\sigma^2)\,\mathrm{d}x$ (giving q), and analogously for y. Furthermore, it is supposed that the joint probability distribution of male and female parental values is bivariate normal with correlation ρ, so the joint probability for parents of having values in the ranges $(x, x + \mathrm{d}x)$, $(y, y + \mathrm{d}y)$ is

$$[2\pi\sigma^2(1-\rho^2)^{1/2}]^{-1} \times \exp - \left[\frac{(x^2 - 2\rho xy + y^2)}{2\sigma^2(1-\rho^2)}\right] \mathrm{d}x\,\mathrm{d}y$$

(giving p). Substituting into (1) above and solving we obtain a to be

$$(1-\rho^2)^{-1/2}\exp-\left[\frac{(\rho^2x^2-2\rho xy+\rho^2y^2)}{2\sigma^2(1-\rho^2)}\right].$$

Although this expression was regarded as the "relative probability" of a mating between two individuals with values x and y [4], in theory it can be unbounded. An interpretation can be given as one-sided assortative mating, based on the formulation in (2). To accommodate two-sided assortative mating and maintain the normality assumptions, selection against individuals having extreme values needs to be introduced, and details have been explored by Wilson [12]. For a single characteristic, the consequences of introducing a selection function that maintains the normality assumptions may not be particularly realistic. For example, the model may require too large a proportion of the population to remain biologically celibate. These modeling problems may be overcome by inbedding the assortation process into a more complex (and possibly more realistic) framework [14], so that assortative mating for a single characteristic is regarded as a consequence of mixing, simultaneously, over more than just one variable.

The classical multifactorial model also assumes that the observed measurement of an individual in the population can be split into two independent parts, the genetic value (given by the sum of contributions from many loci which are assumed initially to be independent) and the environmental value. Furthermore, it is assumed that there is no **interaction** between genotype and environment, and no epistatic effects (*see* **Genotype**) between loci. Moreover, Fisher [4] shows that the genetic value of an individual in the general population also can be split into two independent parts, the representative value and the dominance deviation value, where the representative value is the sum taken over all loci of the value of the genotype at each locus fitted by least squares, and the dominance deviation value is the sum over all loci of the deviation of the genotypic value at each locus from this value. If the population is such that initially each locus is in **Hardy–Weinberg equilibrium** and there is linkage equilibrium between loci, then it can be shown that the effect of assortative mating is to lose the Hardy–Weinberg equilibrium, and to produce **linkage disequilibrium** [12, 13]. Of most interest historically is the correlation, ρ, between relatives, and some important examples are

$$\frac{\rho(\text{parent–child}) = \frac{1}{4}c_1c_2\sigma^{-2}(\varepsilon_1^2+2\rho\varepsilon_1\varepsilon_2+\varepsilon_2^2)}{\left[\frac{1}{2}\sigma^{-2}(\varepsilon_1^2+\varepsilon_2^2)^{1/2}\right]=\rho_c},$$

$$\frac{\rho(\text{grandparent–grandchild})=\frac{1}{2}\rho_c(1+\rho c_1c_2\varepsilon_1\varepsilon_2\sigma^{-2})}{\left[\frac{1}{2}\sigma^{-2}(\varepsilon_1^2+\varepsilon_2^2)^{1/2}\right]},$$

$$\rho(\text{sibs}) = \tfrac{1}{4}c_1[1+c_2-2c_1c_2^2 + c_1c_2^2\sigma^{-2}(\varepsilon_1^2+\varepsilon_2^2+2\rho\varepsilon_1\varepsilon_2)], \quad (3)$$

where ε_1^2 and ε_2^2 are the variances in the populations of male parents and female parents, and

$$c_1 = \frac{\text{var(genetic values)}}{\text{var(observed values)}}, \quad (4)$$

and

$$c_2 = \frac{\text{var(representative values)}}{\text{var(genetic values)}}. \quad (5)$$

If $\varepsilon_1^2=\varepsilon_2^2=\sigma^2$, then the correlations simplify to those found by Fisher [4].

These formulas assume no epistasis, no **gene–environment interaction** or covariation, as well as no environmental covariation between relatives. Although these effects can be incorporated into statistical analyses (see [6]), there has been no systematic study of the effects of incorporating all realistic factors and processes. Moreover, whether a set of assumptions is reasonable or not depends on the particular trait being studied, and often will be controversial, especially as many of the assumptions will not be verifiable in practice. The effect of assortative mating on analyses has been investigated in only a few situations. It has been shown that assortative mating can affect the estimation and interpretability of **heritability** [10], and lowers the power of **linkage** studies [3].

References

[1] Crow, J.F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.

[2] Ewens, W.J. (1979). *Mathematical Population Genetics*. Springer-Verlag, Berlin.

[3] Falk, C.T. (1997). Effect of genetic heterogeneity and assortative mating on linkage analysis: A simulation study, *American Journal of Human Genetics* **61**, 1169–1178.

[4] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[5] Hebebrand, J., Wulftange, H., Goerg, T., Ziegler, A., Hinney, A., Barth, N., Mayer, H. & Remschmidt, H. (2000). Epidemic obesity: are genetic factors involved via increased rates of assortative mating? *International Journal of Obesity* **24**, 345–353.

[6] Hopper, J.L. (1993). Variance components for statistical genetics: Applications in medical research to characteristics related to human diseases and health, *Statistical Methods in Medical Research* **2**, 199–224.

[7] Jennings, H.S. (1916). The numerical results of diverse systems of breeding, *Genetics* **1**, 53–89.

[8] Maes, H.H., Neale, M.C., Kendler, K.S., Hewitt, J.K., Silberg, J.L., Foley, D.L., Meyer, J.M., Rutter, M., Simonoff, E., Pickles, A. & Eaves, L.J. (1998). Assortative mating for major psychiatric diagnoses in two population-based samples, *Psychological Medicine* **28**, 1389–1401.

[9] Moran, P.A.P. & Smith, C.A.B. (1966). Commentary on R.A. Fisher's paper on "The correlation between relatives on the supposition of Mendelian inheritance", *Eugenics Laboratory Memoirs*, Vol. XLI. Cambridge University Press, Cambridge.

[10] Rice, T.K. & Borecki, I.B. (2001). Familial resemblance and heritability, *Advances in Genetics* **42**, 35–44.

[11] Spuhler, J.N. (1968). Assortative mating with respect to physical characteristics, *Eugenics Quarterly* **15**, 128–40.

[12] Wilson, S.R. (1973). The correlation between relatives under the multifactorial model with assortative mating. I & II, *Annals of Human Genetics* **37**, 189–215.

[13] Wilson, S.R. (1978). A note on assortative mating, linkage and genotypic frequencies, *Annals of Human Genetics* **42**, 129–130.

[14] Wilson, S.R. (1981). An alternative approach to modelling the assortative mating process, *Biometrical Journal* **23**, 581–589.

[15] Wright, S. (1921). Systems of mating. III. Assortative mating based on somatic resemblance, *Genetics* **6**, 144–161.

(*See also* **Adoption Studies**; **Path Analysis**)

SUSAN R. WILSON

Asymptotic Relative Efficiency (ARE)

One of the most important problems of statistical practice is point **estimation** of an unknown parameter, say θ. In most cases, there are many apparently reasonable estimators of θ. For example, if one wants to estimate the **mean** of a normally distributed characteristic, it seems reasonable to estimate it by the mean of the characteristic from a sample. Since for normal random variables, the mean and **median** are the same, it also seems reasonable to use the median of the sample values as an estimate. Indeed, each estimate is a consistent estimate in this case (*see* **Consistent Estimator**); i.e. each estimate $\hat{\theta}$ satisfies $\Pr(|\hat{\theta} - \theta| > \varepsilon) \to 0$ as $n \to \infty$, for any given $\varepsilon > 0$. Asymptotic efficiency is a common method to discriminate between two reasonable estimates when there is nothing to discriminate between them from the viewpoint of consistency.

Typically, two consistent estimates, say $\hat{\theta}_1$ and $\hat{\theta}_2$, will also have limiting normal distributions, i.e. $\sqrt{n}(\hat{\theta}_i - \theta) \to \mathrm{N}(0, \sigma_i^2(\theta))$, $i = 1, 2$. In such a case, it is common to approximate the variance of $\hat{\theta}_i$ by $\sigma_i^2(\theta)/n$. The exact variance of $\hat{\theta}_i$ may be hard to calculate for a fixed sample size n, and thus the approximation really does become important. Since variance is a natural measure of accuracy of an estimate, it seems natural to define the efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ as the ratio $\sigma_2^2(\theta)/\sigma_1^2(\theta)$. In principle, the quantities $\sigma_1^2(\theta)$ and $\sigma_2^2(\theta)$ may depend on the unknown parameter θ. However, fortunately, in many important problems of statistics, they are just fixed positive constants not depending on θ, and therefore the asymptotic relative efficiency (ARE) σ_2^2/σ_1^2 has the very appealing interpretation of being one number summarizing the performance of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$. The values of ARE are between 0 and ∞, and ARE > 1 corresponds to $\hat{\theta}_1$ being more efficient than $\hat{\theta}_2$.

Example 1 Suppose $X_1, X_2, \ldots, X_n$ are independent observations from a $\mathrm{N}(\theta, 1)$ population. Then, $\sqrt{n}(\overline{X} - \theta) \to \mathrm{N}(0, 1)$ in distribution, and $\sqrt{n}(M - \theta) \to \mathrm{N}(0, \pi/2)$ in distribution, where M is the median of the sample data $X_1, X_2, \ldots, X_n$. Thus, according to our definition, the asymptotic relative efficiency of the sample median with respect to the sample mean for a normally distributed population is $2/\pi \approx 0.63$.

Interestingly, the situation reverses and the sample median becomes a more efficient estimate if the observations $X_1, X_2, \ldots, X_n$ are instead obtained from a population with a double exponential density, $1/2 \exp(-|x - \theta|)$. In this case, $\sqrt{n}(\overline{X} - \theta) \to \mathrm{N}(0, 2)$ and $\sqrt{n}(M - \theta) \to \mathrm{N}(0, 1)$, and the asymptotic relative efficiency of the sample median with respect to the sample mean is 2.

Example 2 Sir Ronald Fisher, one of the founding fathers of much of statistics as we know it today, once had a communication with A. Eddington, a noted physicist, about how to estimate the standard deviation of a normal distribution. Thus, if $X_1, X_2, \ldots, X_n$ are independent samples from the $\mathrm{N}(\theta, \sigma^2)$ distribution where both parameters are unknown, the specific question was a comparison of the two estimates

$$\hat{\sigma}_1 = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{n}{2}\right)} \left[\sum_{i=1}^{n} (x_i - \overline{x})^2\right]^{1/2}$$

$$\text{and} \quad \hat{\sigma}_2 = \frac{\sqrt{\pi}}{[2(n-1)n]^{1/2}} \sum_{i=1}^{n} |x_i - \overline{x}|,$$

based on the **mean deviation**, where $\Gamma(\cdot)$ denotes the Euler gamma function. Each estimate is unbiased for estimating σ. It is known that $\hat{\sigma}_1$ is the uniformly **minimum variance unbiased (MVU) estimator** of σ for each fixed sample size. So in fixed samples, $\hat{\sigma}_1$ is more efficient than $\hat{\sigma}_2$. An interesting question would be if, even asymptotically, it has an ARE > 1. Using standard methods of large sample theory, it is seen that $\sqrt{n}(\hat{\sigma}_1 - \sigma) \to \mathrm{N}(0, \sigma^2/2)$ and $\sqrt{n}(\hat{\sigma}_2 - \sigma) \to \mathrm{N}(0, [(\pi - 2)/2]\sigma^2)$ in distribution as $n \to \infty$. Thus, applying the definition of the ARE, the ARE of $\hat{\sigma}_1$ with respect to $\hat{\sigma}_2$ is $\pi - 2$, which is indeed larger than 1.

Efficient Estimates

A question of natural interest is the following: is there such a thing as a "most efficient" estimate (*see* **Efficiency and Efficient Estimators**), and how do we formulate such a concept? It turns out that in parametric estimation problems, it is

indeed possible to easily formulate such a concept. Thus, suppose $X_1, X_2, \ldots, X_n$ are independent observations from a population with density $f(x|\theta)$. The quantity $I(\theta) = -E_\theta[\,\mathrm{d}^2 f(x|\theta)/\,\mathrm{d}\theta^2]$, whenever the definition makes sense, is called the Fisher **information** function. Let $\hat{\theta}$ be any estimate of θ that is consistent and asymptotically normal, i.e. $\sqrt{n}(\hat{\theta} - \theta) \to \mathrm{N}(0, \sigma^2(\theta))$. Then, under some (frequently satisfied) regularity conditions on the density $f(x|\theta)$, it is true that $\sigma^2(\theta) \geq 1/I(\theta)$ (exceptions may occur at a "few" values of θ; this phenomenon is known as superefficiency, but we will not worry about this). Thus, any estimate $\hat{\theta}$ of θ which actually attains the bound $\sigma^2(\theta) \equiv 1/I(\theta)$ can be legitimately called an *efficient* estimate of θ.

In a given problem there are usually many efficient estimates of the unknown parameter θ. Standard methods of estimation typically result in efficient estimates, although in finite samples they may have different variances, biases, etc. This reinforces the fundamental issue that efficiency and ARE are intrinsically asymptotic indices in nature, but one hopes that if one estimate is more efficient than another according to the definition of ARE, in moderate samples it outperforms the other estimate as well. Among the standard methods of point estimation, **maximum likelihood** estimates and Bayes estimates typically are all efficient estimates, although exceptions to these general phenomena can and do occur. For instance, if the number of nuisance parameters grows with an increasing sample size, then maximum likelihood estimates of the most important parameter will usually not be efficient. Also, **method of moments** estimates may not be efficient even in very simple problems.

Other Measures of Efficiency

Besides point estimation, another very important problem of statistics is that of **hypothesis testing**. As in point estimation, there are usually many reasonable tests of a specified hypothesis and it is useful to have a concept of efficiency of one test with respect to another. Various efficiency measures have been proposed here too, primarily among them **Pitman efficiency** and Bahadur efficiency.

The Pitman efficiency is defined in the following way: fix a type I error probability or level α, fix an alternative θ, and specify a desired power $1 - \beta$ at this alternative. Let $n_i(\alpha, \beta, \theta)$ denote the minimum sample size required by the ith test, $i = 1, 2$, to achieve this goal. The Pitman efficiency of the first test with respect to the second is taken as the limit of the ratio $n_2(\alpha, \beta, \theta)/n_1(\alpha, \beta, \theta)$ as the alternative $\theta \to \theta_0$ at a suitable rate, where θ_0 is exactly (the) boundary between the null and the alternative hypothesis.

Of course, there are a number of subtle issues involved here. Dependence of this limit on α and β would make universal interpretation of the efficiency value difficult. Also, the limit itself should exist for the definition to make any sense. Finally, the boundary value θ_0 may not be unique. In almost all problems that commonly occur, fortunately these subtleties do not cause any problems and one has a quite good efficiency measure.

Bahadur efficiency proceeds along the same lines, except that one lets α tend to zero, keeping the alternative θ fixed. Thus the Bahadur efficiency can depend on both θ and the desired particular power $1 - \beta$. Fortunately, again, usually dependence on β does not occur, although dependence on θ does. Thus, in contrast to the Pitman measure of efficiency, which is usually one single number, the Bahadur efficiency measure is a curve or a function – a function of the specified alternative θ. This is actually good in some sense, as one has an efficiency measure that discriminates between two competing tests based on which alternative values are really important in the given context. Bahadur's original approach [3] was to compare the rates at which the P values corresponding to the two tests converge to zero at the specified θ. However, the two descriptions are equivalent.

Example 3 Suppose $X_1, X_2, \ldots, X_n$ are independent observations from the double exponential density (1/2) $\exp(-|x - \theta|)$ and we wish to test H_0: $\theta \leq 0$ vs. H_1: $\theta > 0$. The following two tests appear to be reasonable (*see* **Sign Tests**):

1. Sign test.
   Count N = number of sample values > 0.
   Reject H_0 if N is large.
2. Median test.
   Find the median M of the sample values.
   Reject H_0 if M is large (large positive).

The exact critical values for each test (i.e. what is to be regarded as a "large" value) can be found

by large-sample considerations (or even exactly, although it may involve numerical computing and randomization).

Now, it turns out that the Pitman efficiency of the sign test with respect to the median test is 1. So the Pitman measure does not discriminate between the two tests. Interestingly enough, the Bahadur efficiency does, and indeed, Sievers [27] shows that the Bahadur efficiency of the sign test with respect to the median test equals

$$e_{\mathrm{B}}(\theta) = \log \frac{1}{\{4g(\theta)[1-g(\theta)]\}^{1/2}} \{\log 2 + g(\theta)\log g(\theta) + [1-g(\theta)]\log[1-g(\theta)]\}^{-1},$$

where $g(\theta) = (1/2)\mathrm{e}^{-\theta}$.

$e_B(\theta)$ is seen to be > 1 for any $\theta > 0$, establishing for double exponential data that the sign test could be regarded as a better choice than the median test. Table 1 gives the Bahadur efficiency at selected values of θ.

Relevance for Finite Samples

An important practical question is how closely the ARE approximates the ratio of the variances of two estimators in finite samples. It is difficult to give a very general answer to this, but in many examples the fixed sample relative efficiency seems to converge monotonically to the asymptotic relative efficiency, and the approximation becomes quite close for sample sizes $\geq$ 25. For example, for estimating the mean of a normal distribution, the ARE of the median with respect to the mean differs from the asymptotic value $2/\pi$ by at most 6.8% for sample sizes $\geq$ 20. Trimmed means are also common alternatives to usual sample averages as estimates of population means (*see* **Trimming and Winsorization**). A certain amount of trimming of the smallest and the largest observations causes the effect of potential outliers to be decreased and has other nice advantages. Usually 5 or 10% trimming from each side is recommended; see [6]. For the 10% trimmed mean estimate for estimating a normal mean, the fixed sample relative efficiency with respect to the regular mean differs by at most 2.75% from the asymptotic value 0.975 for sample sizes $\geq$ 20. Thus, there is some empirical evidence that the ARE reasonably approximates the fixed-sample efficiency in moderate sample sizes. Expansions of the fixed sample quantity in which the asymptotic quantity is the leading term have also been attempted, frequently on a case-by-case basis. (See [2, 12, 18], and [22] for such developments.)

Table 1

θ	0	0.1	0.25	0.5	1	2	5
$e_{\mathrm{B}}(\theta)$	1	1.003	1.017	1.057	1.182	1.545	3.214

Sensitivity with Respect to Underlying Distribution

It is entirely possible that one estimate or test is more efficient than another if samples are obtained from one distribution, but loses these advantages, maybe drastically, for a fairly similar distribution. The comparison of the sample median and the sample mean is a good illustration. The mean has an efficiency of 1.57 with respect to the median if samples are known to come from a normal distribution, but this efficiency drops to 0.5 if data instead come from double exponential density, described before. Yet it is not easy to distinguish between the two distributions from moderate samples by using common methods, graphical or otherwise. Bickel & Lehmann [6] provide some concrete results in this direction. For example, they show that if samples are obtained from any density that is symmetric and unimodal about the mean, then the 5% trimmed mean has an ARE of at least 0.83 with respect to the mean for estimating the population mean, and, of course, for many such densities the efficiency is substantially larger than 1. This may be used as an argument for using the 5% trimmed mean if concerns about the exact density from which one is sampling exist. More information on this can be found in [23] and [28].

Concepts of Higher-order Efficiency

As stated before, in parametric estimation problems, it is customary to have many estimates which are fully efficient. It then becomes necessary, at least from a theoretical standpoint, to have a criterion to distinguish among them. The concept of second-order efficiency (now usually referred to as third-order efficiency) was introduced to address this issue.

(See [25, 26, 1, 14], and [16] for later developments.) The idea is to derive an expansion for n times the variance of a statistic, in which the leading term is the Fisher information function and subsequent terms decrease in reciprocals of powers of $1/n$. The second term is used as a comparison among different estimators, or simply for selecting an estimator which is first- as well as second-order efficient. In a peculiar result, Pfanzagl [24] showed that often first-order efficiency automatically implies second-order efficiency as well, making it necessary to consider higher-order efficiencies as a basis for comparison and selection. Bickel et al. [9] and Ghosh [15] expand on these results and ideas.

Complex Models

Parametric models using a given functional form for the density are often convenient choices, and perhaps restrictive. Similarly, the assumption that the sample observations are independent also often does not meet the criteria of realism. Real data often have a positive **serial correlation** or have a **time series** character. Models broader than parametric can be of various types; **nonparametric models** have been the popular alternative. In standard nonparametric modeling, very little is assumed about the density besides some minimal features, mostly to do with shape and symmetry, unimodality, etc. Intermediate between fully parametric and fully nonparametric models are the recent **semiparametric models**. It should be mentioned that complexity may arise not just from more complex models, but also because the quantity to be estimated is more complex than a simple thing like a mean or variance. For example, Bickel & Ritov [8], and Hall & Marron [20] consider estimation of $\int [f'(x)]^2 \, dx$, the integrated squared derivative of a density.

Efficient estimation in complex models has a large literature, of a substantially more difficult nature, as expected. The literature includes [4], [7], [10], [13], [19], [21], and [29]. Efficiency for dependent samples also has a substantial literature, but is more scattered. Grenander & Rosenblatt [17] is a classic reference which established efficiency of the sample mean for estimating the mean of a **stationary** process under quite mild conditions. Brockwell & Davis [11] give more information and discuss more problems. Efficient estimation in a relatively recent class of time series models known as long memory processes appears to be of a totally different qualitative nature. This can be seen in [5].

References

[1] Akahira, M. & Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*. Springer-Verlag, New York.

[2] Albers, W. (1974). *Asymptotic Expansions and the Deficiency Concept in Statistics, Mathematical Centre Tract*, Vol. 58, Amsterdam.

[3] Bahadur, R.R. (1960). Stochastic comparison of tests, *Annals of Mathematical Statistics* **31**, 276–295.

[4] Begun, J.M., Hall, W.J., Huang, W.M. & Wellner, J.A. (1983). Information and asymptotic efficiency in parametric–semiparametric models, *Annals of Statistics* **11**, 432–452.

[5] Beran, J. (1995). *Statistics for Long Memory Processes*. Chapman & Hall, New York.

[6] Bickel, P.J. & Lehmann, E.L. (1976). Descriptive statistics for nonparametric models, *Annals of Statistics* **4**, 1139–1158.

[7] Bickel, P.J. & Ritov, Y. (1987). Efficient estimation in the errors in variables model, *Annals of Statistics* **15**, 513–540.

[8] Bickel, P.J. & Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates, *Sankhyā, Series A* **50**, 381–393.

[9] Bickel, P.J., Chibisov, D.M. & van Zwet, W.R. (1981). On efficiency of first and second order, *International Statistical Review* **49**, 169–175.

[10] Bickel, P.J., Klassen, C.A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[11] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*. Springer-Verlag, New York.

[12] Chandra, T. & Ghosh, J.K. (1978). Comparison of tests with same Bahadur efficiency, *Sankhyā, Series A* **40**, 253–277.

[13] Chen, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models, *Annals of Statistics* **23**, 1102–1129.

[14] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency), *Annals of Statistics* **3**, 1189–1242.

[15] Ghosh, J.K. (1994). *Higher Order Asymptotics*. Institute of Mathematical Statistics, Hayward.

[16] Ghosh, J.K., Sinha, B.K. & Wieand, H.S. (1980). Second order efficiency of mle with respect to a bounded bowl-shaped loss function, *Annals of Statistics* **8**, 506–521.

[17] Grenander, U. & Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series*. Wiley, New York.

[18] Groenboom, P. & Oosterhoff, J. (1980). Bahadur Efficiency and Small Sample Efficiency: A Numerical

Study, *Report SW 68–80*. Mathematische Centrum, Amsterdam.

[19] Groenboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser, Boston.

[20] Hall, P. & Marron, J.S. (1987). Estimation of integrated squared density derivatives, *Statistics & Probability Letters* **6**, 109–115.

[21] Hasminskii, R.Z. & Ibragimov, I.A. (1983). On asymptotic efficiency in the presence of an infinite dimensional nuisance parameter, in *USSR – Japan Symposium*. Springer-Verlag, New York, pp. 195–229.

[22] Hodges, J.L. Jr & Lehmann, E.L. (1976). Deficiency, *Annals of Mathematical Statistics* **41**, 783–801.

[23] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

[24] Pfanzagl, J. (1979). First order efficiency implies second order efficiency, in *Contributions to Statistics*, J. Hajek Memorial Volume, J. Jureckova, ed. Academia, Prague, pp. 167–196.

[25] Rao, C.R. (1961). Asymptotic efficiency and limiting information, in *Proceedings of Fourth Berkeley Symposium*, Vol. I, pp. 531–545.

[26] Rao, C.R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion), *Journal of the Royal Statistical Society, Series B* **24**, 46–72.

[27] Sievers, G.L. (1969). On the probability of large deviations and exact slopes, *Annals of Mathematical Statistics* **40**, 1908–1921.

[28] Staudte, R.G. & Sheather, S.S. (1990). *Robust Estimation and Testing*. Wiley, New York.

[29] van der Vaart, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models, *Annals of Statistics* **24**, 862–878.

ANIRBAN DASGUPTA

Attributable Fraction in Exposed

The attributable fraction in the exposed (AF_{E}) is defined as the proportion of disease cases that can be attributed to an exposure factor among the exposed subjects only [2–5]. It can be formally written as

$$AF_{\mathrm{E}} = \frac{[\Pr(D|E) - \Pr(D|\overline{E})]}{\Pr(D|E)}, \qquad (1)$$

where $\Pr(D|E)$ is the probability of disease in the exposed individuals, E, and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same subjects but with all exposure eliminated. It is also called the **attributable risk** among the exposed [1, 3] and can be rewritten as

$$AF_{\mathrm{E}} = \frac{(RR - 1)}{RR}, \qquad (2)$$

a one-to-one increasing function of the **relative risk** (RR). It can be seen to equal the attributable risk in the special case of an exposure present in all subjects in the population (exposure **prevalence** of 1).

When the exposure factor under study is a risk factor ($RR > 1$), it follows from the above definition that AF_{E} lies between 0 and 1, and it is usually expressed as a percentage. AF_{E} increases with the strength of the association between exposure and disease measured by the relative risk and tends to 1 for an infinitely high relative risk. AF_{E} is equal to zero when there is no **association** between exposure and disease ($RR = 1$). Negative values of AF_{E} correspond to a protective exposure ($RR < 1$), in which case AF_{E} is not a meaningful measure.

While AF_{E} has some usefulness in measuring the disease-producing impact of an association between exposure and disease, it is much less useful than the attributable risk and does not share the same public health interpretation. This is because it is only a one-to-one transformation of relative risk. It does not take the prevalence of exposure into account, and, for instance, it can be high even if the prevalence of exposure is low in the population. Moreover, while AR estimates for several risk factors can be compared meaningfully to assess the relative importance of these risk factors at the population level, such is not the case with AF_{E} estimates. Indeed, each AF_{E} estimate refers to a different group which is specific to the risk factor under consideration (subjects exposed to that risk factor). However, one advantage of AF_{E} over attributable risk is that, since it does not depend on the prevalence of exposure, portability from one population to another is less problematic than with attributable risk and depends only on the portability of relative risk.

Issues of estimability and **estimation** of AF_{E} are the same as those for relative risk. AF_{E} can be estimated from the main types of epidemiologic studies (**cohort**, **case–control**, **cross-sectional**, **case–cohort**). Point estimates are obtained from point estimates of relative risk (**odds ratio** in case–control studies). **Variance** estimates can be obtained from variance estimates of relative risk (odds ratio in case–control studies) through the **delta method** [6], which yields:

$$\mathrm{var}(\widehat{AF}_{\mathrm{E}}) = \frac{\mathrm{var}(\widehat{RR})}{RR^4}, \qquad (3)$$

where $\widehat{RR}$ denotes a point estimate of relative risk (odds ratio in case–control studies), and $\widehat{AF}_{\mathrm{E}}$ denotes a point estimate of AF_{E}.

References

[1] Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case–control studies: a review, *Statistics in Medicine* **10**, 1753–1773.

[2] Cole, P. & MacMahon, B. (1971). Attributable risk percent in case–control studies, *British Journal of Preventive and Social Medicine* **25**, 242–244.

[3] Levin, M.L. (1953). The occurrence of lung cancer in man, *Acta Unio Internationalis Contra Cancrum* **9**, 531–541.

[4] MacMahon, B. & Pugh, T.F. (1970). *Epidemiology: Principles and Methods*. Little, Brown & Company, Boston.

[5] Micttinen, O.S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention, *American Journal of Epidemiology* **99**, 325–332.

[6] Rao, C.R. (1965). *Linear Statistical Inference and Its Application*. Wiley, New York, pp. 319–322.

JACQUES BENICHOU

Attributable Risk

The attributable risk (AR), first introduced by Levin [45], is a widely used measure to assess the public health consequences of an association between an exposure factor and a disease. It is defined as the proportion of disease cases that can be attributed to exposure and can be formally written as:

$$AR = \frac{[\Pr(D) - \Pr(D|\overline{E})]}{\Pr(D)}, \qquad (1)$$

where $\Pr(D)$ is the probability of disease in the population, which may have some exposed (E) and some unexposed ($\overline{E}$) individuals, and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same population but with all exposure eliminated.

The AR takes into account both the strength of the association between exposure and disease and the **prevalence** of exposure in the population. This can be seen, for instance, through rewriting AR from (1), using **Bayes' theorem**, as [15, 53]:

$$AR = \frac{[P(E)(RR - 1)]}{[1 + P(E)(RR - 1)]}, \qquad (2)$$

a function of the prevalence of exposure, $P(E)$, and the **relative risk** RR. Therefore, while the relative risk is mainly used to establish an association in etiologic research, AR has a public health interpretation as a measure of preventable disease. A high relative risk can correspond to a low or high AR depending on the prevalence of exposure, which leads to widely different public health consequences. One implication is that, while the relative risk is often portable from one population to another, as the strength of the association between disease and exposure might vary little among populations, portability is not a property of AR, as the prevalence of exposure may vary widely among populations that are separated in time or location.

Range

When the exposure factor under study is a risk factor (relative risk > 1), it follows from the above definition that AR lies between 0 and 1, and is, therefore, very often expressed as a percentage. AR increases both with the strength of the **association** between exposure and disease measured by the relative risk, and with the prevalence of exposure in the population. A prevalence of 1 (or 100%) yields a value of AR equal to $(RR - 1)/RR$, and AR tends to 1 for an infinitely high relative risk provided the prevalence is greater than 0.

AR is equal to zero when either there is no association between exposure and disease ($RR = 1$) or no subject is exposed in the population.

Finally, negative values of AR are obtained for a protective exposure (relative risk < 1). In this case, AR varies between 0 and $-\infty$ and AR is not a meaningful measure. Either one must consider reversing the coding of exposure to go back to the situation of a positive AR or one must consider a different parameter; namely, the prevented fraction (*see* **Preventable Fraction** and the section "Related Quantities" below).

Synonyms

Numerous terms have been used in the literature instead of attributable risk. Attributable risk was the term originally introduced by Levin [45], but it is not a universally accepted term because (i) the word "risk" may be misleading as AR does not represent a risk in the usual sense and (ii) it may not allow a clear enough distinction from the more restrictive concept of attributable risk (or fraction) in the exposed (*see* **Attributable Fraction in Exposed** and the section "Related Quantities" below). Most common alternative terms are population attributable risk [47] and population attributable risk percent [15], etiologic fraction and fraction of etiology [53], and attributable fraction [33, 43, 57]. Up to 16 terms have been used to denote attributable risk in the literature [26].

Interpretation and Usefulness

AR is used to assess the potential impact of prevention programs aimed at eliminating exposure from the population. It is often thought of as the fraction of disease that could be eliminated if exposure could be totally removed from the population.

However, this interpretation can be misleading because, for it to be strictly correct, the three following conditions have to be met. First, estimation

of AR has to be **unbiased** (see the section "Estimation" below). Secondly, the exposure factor has to be causal rather than merely associated with the disease (*see* **Causation**). Thirdly, elimination of the risk factor has to be without having any effect on the distribution of other risk factors. Indeed, as it might be difficult to alter the level of exposure to one factor independently of other risk factors, the resulting change in disease load might be different from the AR estimate [74]. For these reasons, various authors elect to use weaker definitions of AR, such as the proportion of disease that can be related or linked, rather than attributed, to exposure [53].

Despite these limitations, AR can serve as a useful guide in assessing and comparing various prevention strategies. It should be noted that authors have estimated AR in situations where causality was far from being established, and the association between exposure and disease still tentative and controversial. For instance, Alavanja et al. [2] estimated the risk of lung cancer attributable to elevated saturated fat intake in a population of nonsmoking women in the state of Missouri. They interpreted their estimate as quantifying the potential impact of eliminating elevated saturated fat exposure were it later proven to be causally related to lung cancer. This use of AR is somewhat controversial, the more so when the association is not well established, and it needs, therefore, to be presented with proper qualification.

Estimation of AR can usefully be complemented by applying the AR estimate to the **incidence rate** of the disease in the population to see not just what proportion of disease, but how many cases per unit of time are attributable to exposure. Moreover, multiplying an estimate of $1 - AR$ times an estimate of the incidence rate in the population yields an estimate of the incidence rate in the unexposed (baseline incidence rate), which can be useful in a perspective of etiologic research. For instance, Silverman et al. [66], estimated the risk of pancreatic cancer attributable to alcohol in white and black men separately in the United States. They found that the substantial difference in incidence rates between black and white men (16.0 vs. 12.8 per 100 000 person-years, a 25% higher rate in black men) could be explained in part by the higher AR for alcohol among black men, since the race difference among the unexposed (i.e. having removed the contribution of alcohol) was reduced by almost half (14.2 vs. 12.5 per 100 000 person-years, a 14% higher rate in black men). The higher AR estimate among black men was itself related to both a higher relative risk estimate for elevated alcohol consumption and a higher prevalence of that exposure among black men.

Finally, AR can be considered for not just one, but several, risk factors in combination. One can be interested in the potential effect on disease load of removing these risk factors from the population. Alternatively, one might interpret an AR estimate for all known risk factors as a gauge of what is known about the disease etiology, and its complement to 1 as a gauge of what remains unexplained by known risk factors. For instance, Madigan et al. [48] estimated at 41% the AR of breast cancer for well-established risk factors; namely, later age at first birth, nulliparity, family history of breast cancer in a first-degree relative and higher socioeconomic status. They argued in favor of more etiologic research to find new risk factors, whether genetic, hormonal, or biological, to account for the remaining 59% of unexplained breast cancer cases. In fact, most authors in that field have come to similar conclusions and the AR figure of 50% or less is a useful indicator and reminder of the need for new research directions in breast cancer etiology.

Properties

Two basic properties of AR need to be emphasized. First, AR values greatly depend on the definition of the reference level for exposure (unexposed or baseline level). A larger proportion of subjects exposed corresponds to a more stringent definition of the reference level and, as one keeps depleting the reference category from subjects with higher levels of risk, AR values and estimates keep rising. This property has a major impact on AR estimates as was illustrated by Benichou [5] and Wacholder et al. [71]. For instance, Benichou [5] found that the AR estimate of esophageal cancer for an alcohol consumption greater or equal to 80 g/day (reference level of 0–79 g/day) was 38% in the Ille-et-Vilaine part of France [70], and increased dramatically to 70% for an alcohol consumption greater or equal to 40 g/day (more restrictive reference level of 0–39 g/day). This property plays a role whenever studying a continuous exposure with a continuous gradient of risk and when there is no obvious choice of threshold. Therefore, AR estimates must be reported with reference to a

clearly defined baseline level in order to be interpreted validly. In the previous example, one notes that the interpretation in preventive terms would differ for the two AR estimates. One would conclude that 70% (respectively 38%) of all esophageal cancers in Ille-et-Vilaine can be attributed to an alcohol consumption of at least 40 g/day (respectively 80 g/day) and could potentially be prevented by reducing alcohol consumption to less than 40 g/day (respectively 80 g/day).

The second main property is distributivity. If several categories of exposure are considered instead of just one, then the sum of the category-specific ARs (see the section "Special Problems" below) equals the AR calculated from combining those categories into a single exposed category, regardless of the number and the divisions of the categories that are formed [5, 71, 74], provided the reference category remains the same. This property applies strictly to unadjusted AR estimates and to adjusted AR estimates calculated on the basis of a saturated model (see below) [5]. In other situations, it applies approximately [71]. For instance, Wacholder et al. [71] used data on malignant mesothelioma [67], and obtained an unadjusted AR estimate equal to 82% for a nontrivial (moderately low, medium, or high) likelihood of exposure to asbestos, identical to the sum of the respective category-specific AR estimates of 13%, 6%, and 64% for moderately low, medium, and high likelihoods of exposure. Thus, if an overall AR estimate for exposure is the focus of interest, there is no need to break the exposed category into several mutually exclusive categories, even in the presence of a gradient of risk with increasing exposure.

Estimability

AR can be estimated from the main types of epidemiologic studies, namely **cohort**, **case–control**, **cross-sectional**, and **case–cohort studies**. It can be seen immediately that all quantities in (1) are estimable from all four types of studies except case–control studies. For case–control studies, one has to consider (2) and estimate $P(E)$ from the proportion exposed in the controls, making the rare-disease assumption also involved in estimating odds ratios rather than relative risks. Alternatively, one can rewrite (1) using Bayes' theorem in yet another manner as

$$AR = \frac{\Pr(E|D)(RR - 1)}{RR}. \tag{3}$$

In (3), the quantity $\Pr(E|D)$ can be directly estimated from the diseased individuals (cases) and RR can be estimated from the **odds ratio**. Therefore, AR is estimable from case–control studies as well.

Often, cohort studies are based on groups with a different prevalence of exposure than the general population. This renders AR estimates obtained from cohort studies less applicable to the general population and might explain why AR is seldom estimated from cohort studies.

Estimation

Since Levin [45] first introduced AR, there has been a very active research in AR estimation and numerous developments have appeared, particularly in recent years. Case–control studies have been the most explored. The outline given here applies to cohort, case–control and cross-sectional studies, unless stated otherwise. While AR is estimable from case–cohort studies, no study of AR estimation methods seems to have been published, and case–cohort studies will be considered separately (see the section "Special Problems" below).

Unadjusted Estimation

From the three types of studies considered, it is easy to obtain unadjusted (crude) AR estimates, either from (2) or (3). No other factor than the exposure of interest is considered, and the data are limited to exposure and disease state. For instance, one obtains the following estimate both from (2) and (3) in case–control studies:

$$AR = \frac{(n_1 m_0 - m_1 n_0)}{m_0 n}, \tag{4}$$

where n_0 and n_1 respectively denote the numbers of unexposed and exposed cases ($n_0 + n_1 = n$) and m_0 and m_1 the numbers of unexposed and exposed controls ($m_0 + m_1 = m$).

Variance estimates can be obtained from the **delta-method** [60] by considering the following distributions. In case–control studies, the quantities n_1 and m_1 have independent **binomial distributions**

with respective indexes n and m considered as fixed (exposure is random conditional on disease status). In cohort studies, the quantities n_0 and n_1 have binomial distributions with respective fixed indexes $n_0 + m_0$ and $n_1 + m_1$ considered as fixed (disease status is random conditional on exposure). In cross-sectional studies, one has to consider the full (unrestricted) **multinomial** model in which all four quantities n_0, n_1, m_0, and m_1 come from a common multinomial distribution with index $n + m$ considered as fixed (exposure and disease status are random).

Once a variance estimate is obtained, a standard **confidence interval** for AR can be constructed based on the asymptotic **normal distribution** of AR. Alternatively, Walter [73] suggested using the interval based on the log transformed variable $\log(1 - AR)$, and Leung & Kupper [44] based the interval on the logit-transformed variable $\log[AR/(1 - AR)]$ (*see* **Transformations**). Whittemore [77] noted that the log-transformation yields a wider interval than the standard interval for $AR > 0$. Leung & Kupper [44] showed that the interval based on the logit transform is narrower than the standard interval for values of AR strictly between 0.21 and 0.79, whereas the reverse holds outside this range for positive values of AR. While the coverage probabilities of these intervals have been studied in some specific situations and partial comparisons have been made, no general study has been performed to determine their relative merits in terms of coverage probability.

Unadjusted estimates of AR are, in general, biased, because they fail to take into account other risk factors that **confound** the association between exposure and disease. The problem is analogous to estimation of relative risks or odds ratios, and has been studied by several authors [53, 74–78]. It is one of inconsistency rather than small-sample **bias**. Walter [75] showed that, if X_1 and X_2 are two binary exposure factors and if one is interested in estimating an AR for X_1, then the following applies. The crude AR estimate is unbiased if and only if at least one of the following two conditions is true:

1. X_1 and X_2 are independently distributed in the population, that is:

$$\Pr(X_1 = 0, X_2 = 0)\Pr(X_1 = 1, X_2 = 1) = \Pr(X_1 = 0, X_2 = 1)\Pr(X_1 = 1, X_2 = 0),$$

   where level 0 denotes the absence of exposure and 1 the exposed category.

2. Exposure to X_2 alone does not increase disease risk; that is:

$$\Pr(D|X_1 = 0, X_2 = 1) = \Pr(D|X_1 = 0, X_2 = 0).$$

Therefore, if X_2 acts as a true confounder of the association between exposure X_1 and the disease, then the crude estimate of AR is inconsistent, as is a crude estimate of relative risk or odds ratio. When neither condition 1 nor 2 is true, the direction of the bias can be determined. If X_2 alone increases risk, then the bias is positive (AR is overestimated) if X_1 and X_2 are positively correlated, and negative if they are negatively correlated [75]. When considering several factors $X_j (j = 2, \ldots, J)$, conditions 1 and 2 can be extended to a set of $2(J - 1)$ analogous **sufficient** conditions concerning factors X_1 and $X_j (j = 2, \ldots, J)$ as shown by Walter [75].

Adjusted Estimation – Inconsistent Approaches

Let us first note that two simple adjusted estimation approaches discussed in the literature are inconsistent. The first approach ever proposed to obtain adjusted AR estimates, based on decomposing AR into exposure and confounding effects [74], was shown to be inconsistent [27] and, accordingly, bias was exhibited in **simulations** for the **crossover design** [26, 27]. The approach based on using (2) and plugging in an adjusted relative risk estimate (odds ratio estimate in case–control studies), along with an estimate of $P(E)$, has also been advocated [15, 55], but it too has been shown to yield inconsistent estimates [28, 32] of AR, and, accordingly, bias was exhibited in simulations for the crossover design (i.e. under the unrestricted multinomial model) [26, 27].

Two adjusted approaches based on **stratification**, the **Mantel–Haenszel** approach and the weighted-sum approach, yield valid estimates.

Adjusted Estimation – The Mantel–Haenszel Approach

The Mantel–Haenszel (MH) approach has been developed by Greenland [29] and Kuritz & Landis [38, 39]. It allows adjustment for one or more polychotomous factors forming J joint levels or strata. It is based on the formulation of AR as a function of the relative risk (odds ratio in case–control studies) and

the prevalence of exposure in diseased individuals, as given by (3). One plugs in an estimate of $\Pr(E|D)$ (given by the observed proportion of cases exposed) and an estimate of the common adjusted relative risk (odds ratio in case–control studies). The MH estimate of the common odds ratio [49] can be used in case–control studies, while MH-type estimates of the common relative risk [36, 69] can be used in cohort or cross-sectional studies.

Other choices than the MH estimator of odds ratio or MH-type estimators of relative risk are possible, and this approach could be more generally termed the common relative risk (odds ratio in case–control studies) approach. Other choices have been suggested, such as an internally standardized mortality ratio [53] (*see* **Standardization Methods**) or the **maximum likelihood** estimator from **logistic regression** [29]. MH-type estimators combine properties of lack of (or very small) bias even for sparse data (e.g. individually matched case–control data), good efficiency except in extreme circumstances [10–12, 42], and the existence of consistent variance estimators even for sparse data ("dually-consistent" variance estimators) [29, 62].

While point estimation with the MH approach is simple, variance estimation is more complex. Variance estimators can be obtained either through applications of the delta method [38, 39] or by relying on asymptotic properties of first derivatives of log likelihood functions [29]. Finite sample properties were studied by simulations under the assumption of a common odds ratio or relative risk. It was found that bias in estimating AR was negligible in case–control studies with **simple random sampling** [38], **stratified random sampling** [29] and individual **matching** [39], as well as in cross-sectional studies [26, 27]. Variance estimates were also unbiased and coverage probabilities close to nominal for those various designs.

The crucial assumption in the MH approach is that of a common or homogeneous relative risk or odds ratio, which amounts to a lack of **interaction** between the adjustment factor(s) and the exposure factor (no **effect modification**). If interaction is present, the MH estimator of AR is inconsistent, which was illustrated in simulations for the crossover design [26, 27]. Greenland [29] proposed a modification of the MH approach, consisting in defining H levels out of the original J levels formed by adjustment factors. The H levels are defined so that, within each of them, the odds ratio or relative risk can be considered as homogeneous and is estimated separately. This constitutes a possible solution although (i) the definition of the H levels, which is critical to this modified approach, is somewhat arbitrary in view of the low **power** of tests to detect interaction, and (ii) finite sample properties of this modified approach might not be as favorable as the original MH approach and bias might arise as with the weighted-sum approach (see below). Indeed, this modified approach is a hybrid approach, being intermediate between the MH and weighted-sum approaches.

Adjusted Estimation – The Weighted-sum Approach

This approach allows adjustment for one or more polychotomous factors forming J levels or strata. AR is written as a weighted sum of the ARs over strata, namely [74, 77, 78]:

$$AR = \sum_j w_j AR_j, \tag{5}$$

where AR_j and w_j are respectively the AR specific to level j and the corresponding weight. Setting w_j as the proportion of diseased individuals (cases) in level j yields an asymptotically unbiased estimator of AR, which can be seen to be a maximum-likelihood estimator [73, 77]. This choice of weights defines the "case-load method". An alternative choice of weights, called "precision-weighting" is given by setting w_j as the inverse variance of the AR estimate in level j over the sum of inverse variances over all levels [25]. It can be shown to be an inconsistent estimator of AR except in special circumstances [27].

The weighted-sum approach does not require the assumption of a common relative risk or odds ratio. The odds ratios or relative risks are estimated separately for each level j. No restrictions are placed on them, which corresponds to a saturated model (*see* **Generalized Linear Model**). Thus, the weighted-sum approach not only accounts for confounding but also for interaction. It is interesting to note that, under the assumption of a common relative risk or odds ratio, the weighted-sum approach yields the same expression for AR as the MH approach [5].

Point estimates are easy to obtain for the various types of designs. Variance estimates can be obtained from specializing the distributions described above

(see the subsection "Unadjusted Estimation" above) for each level j and applying the delta method [60]. They have been worked out for case–control designs [77], cross-sectional designs [27] and cohort designs [48].

Unlike the MH approach, small-sample bias is an issue with the weighted-sum approach, at least for case–control designs. Negative bias was found in simulations of case–control studies for **frequency matching** and simple random sampling of the controls, under the assumption of a common odds ratio and with case-load weighting [38, 77]. This bias was substantial for sparse data and a high prevalence of exposure in controls. Precision-weighting yielded a positive bias of similar magnitude [38]. The strong negative (or positive) bias renders the approach inappropriate for individual matching [38, 77]. A tendency towards conservative variance estimates and confidence intervals was also observed [38, 77]. For crossover designs, however, the results were much more favorable, as no severe small-sample bias was found in simulations, whether or not a common relative risk was considered [26, 27].

Model-based Adjusted Estimation

The MH approach rests on the assumption of a common relative risk or odds ratio and yields biased estimates in the case of interaction between exposure and adjustment factors. The weighted-sum approach does not impose any structure on the relative risk or odds ratio and its variation with levels of adjustment factors, but is plagued by problems of small sample bias, particularly for case–control designs. A natural alternative has been to develop adjustment procedures based on **regression** models in order to take advantage of their flexible and unified approach to efficient parameter estimation and **hypothesis testing**.

Walter [74] first suggested this route, and others have followed [22, 68]. Greenland [29] proposed a modification of the MH approach for case–control studies, consisting in substituting a maximum-likelihood estimate of the odds ratio from conditional logistic regression rather than the MH estimate of odds ratio in (3), and he worked out the corresponding variance estimate for AR. This modification could be applied to other designs but retains the constraint of a homogeneous odds ratio.

The full generality and flexibility of the regression approach was first exploited by Bruzzi et al. [14] who expressed AR as:

$$AR = 1 - \sum_j \sum_i \frac{\rho_{ij}}{RR_{i|j}}. \tag{6}$$

In this formula, the first sum is taken over all J levels formed by polychotomous adjustment factors, and the second sum is taken over all exposure levels (usually one unexposed level and one exposed level). The quantity ρ_{ij} represents the proportion of diseased individuals (cases) with respective levels i and j of exposure and adjustment factors, while $RR_{i|j}$ represents the relative risk for level i of exposure given level j of adjustment factors. An informal proof of (6) can be found in Bruzzi et al. [14] and a more formal one in Benichou [5].

The model-based approach based on (6) is very general in several respects. First, while it was derived by Bruzzi et al. [14] for case–control studies, it can be used as well for cohort and cross-sectional studies. For all three designs, an estimate is obtained by replacing ρ_{ij} by the observed proportion among diseased individuals, and by replacing $RR_{i|j}$ by a maximum likelihood estimate obtained from a regression model. In case–control studies, an estimate of the odds ratio from unconditional or conditional logistic regression can be used; in cross-sectional studies, an estimate of the relative risk from unconditional logistic regression can be used; in cohort studies, an estimate of the relative risk from unconditional logistic regression or from **Poisson regression** can be used. Models with additive forms have also been proposed [17].

Secondly, since estimates of odds ratio and relative risk are obtained from regression models, this approach provides a unified framework for testing hypotheses and selecting models. In particular, interaction terms can be introduced in the model, tested and retained or not, depending on the result of the test. This approach allows control for confounding and interaction and essentially parallels the estimation of the relative risk or odds ratio. **Parsimony** can be balanced against bias and the "best" model selected. More elaborate models (e.g. models with interaction terms) protect against inconsistency but can lead to small-sample bias and larger **random error**, while more parsimonious models have the reverse properties.

Thirdly, the model-based approach is general in that it includes the crude and other adjusted

approaches as special cases [5]. The unadjusted approach corresponds to models with exposure only. The MH approach corresponds to models with exposure and confounding, but no interaction terms between exposure and confounding factors. The weighted-sum approach corresponds to fully saturated models with all interaction terms. Intermediate models are possible; for instance, models allowing for interaction between exposure and one confounder only, or models in which the main effects of some confounders are not modeled in a saturated way.

While point estimates are easy to obtain, variance estimates are complex because they involve **covariances** between quantities ρ_{ij} and $RR_{i'|j'}$ that are related implicitly (rather than explicitly) through score equations. Benichou & Gail [8] worked out a variance estimator for all types of case–control studies (with simple random sampling, stratified random sampling, frequency-matching and individual matching of the controls), using an extension of the delta method to implicitly related random variables [7]. Basu & Landis [3] used a similar approach for cohort and cross-sectional designs. In case–control studies, simulations showed little or no bias in most situations [8]. However, as the data became sparse, negative bias was observed with the unconditional logistic model [8]. This could be remedied by the use of more parsimonious models when appropriate or the use of conditional logistic regression. Use of the latter approach, however, remains a research issue, as variance estimates have been derived for conditional logistic regression only for the situation of individual matching. Finally, variance estimates were unbiased and coverage probabilities close to nominal for all types of case–control studies in the aforementioned simulations [8].

Greenland & Drescher [31] have made the point that Bruzzi et al.'s estimator of AR is not exactly a maximum likelihood estimator, and have proposed a modified approach in order to obtain a maximum likelihood estimator. The proposed modification consists in using a model-based estimate of quantities ρ_{ij} rather than estimating these quantities from the corresponding observed quantities. They developed point and variance estimators for case–control designs under the unconditional logistic model, and for cohort designs under the unconditional logistic model and the **Poisson** model. In case–control studies, their approach can be seen as a generalization of Drescher & Schill's approach [24]. Variance estimators rely on the delta method [60] rather than on the implicit delta method [7] as for Bruzzi et al.'s estimator.

The two model-based approaches are identical for fully saturated models, in which case they also coincide with the weighted-sum approach. More generally, provided that the model is not misspecified, the two approaches are practically equivalent, as was illustrated by simulations for the case–control design [31]. Point and variance estimators differed only trivially between the two approaches, with mean differences equal to less than 0.001 and correlations in excess of 0.999. In simulations of the cohort design for Greenland & Drescher's modified model-based approach, some downward bias was found, in a similar way to what had been observed for Bruzzi et al.'s model-based approach for case–control designs [8], and variance estimates appeared to be without substantial bias [31].

In practice, the two model-based approaches seem, therefore, to differ very little. The maximum likelihood approach might be more efficient for small samples, although no difference was observed in simulations of the case–control design even for samples of 100 cases and 100 controls. The maximum likelihood approach might be less robust to model **misspecification**, however, as it relies more heavily on the model for the relative risk or odds ratio. In one circumstance, the distinction between the two approaches is unequivocal. The modified approach does not apply to the conditional logistic model, and if that model is to be used (notably, in case–control studies with individual matching), Bruzzi et al.'s original approach is the only possible choice.

Special Problems

Case–Cohort Design

In the case–cohort design, information on exposure is gathered only in a subcohort of subjects randomly selected from the original cohort and among subjects who develop the disease [59]. Case–cohort data contain information on the prevalence of exposure and allow estimation of the relative risk. Therefore, AR is estimable from case–cohort data, and all estimation methods presented above could, in principle, be used to estimate AR. However, the details have not been worked out in the literature and variance estimators may prove complex to derive.

Risk Factor with Multiple Levels of Exposure

It has been seen above that, because of the distributive property (see the section "Properties" above), it is sufficient to consider one overall exposed level to estimate AR. However, several levels of exposure are worth considering when estimates of AR at specified levels are of interest.

The concept of partial or level-specific AR corresponds to the proportion of disease cases that can be attributed to a specified level of exposure, and may have important policy implications for screening groups at highest risk of disease, for instance [21, 53, 74]. All estimation methods described above can be extended to produce such AR estimates. In particular, Denman & Schlesselman [21] have developed unadjusted estimates for case–control designs. The model-based approach lends itself naturally to this problem, as (6) need only be slightly modified (all RR are set equal to 1 in it, except those for the exposure level of interest). Use of the model-based approach is illustrated in Coughlin et al. [16] who considered the esophageal cancer case–control data mentioned above (see the section "Properties" above) and showed that the AR for "moderate" drinkers (40–79 g/day) was higher than that for heavy drinkers (120+ g/day) (27% vs. 22%), suggesting that prevention policies targeting "moderate" drinkers might be potentially more effective than those aimed at heavy drinkers in that population.

Finally, AR estimates have been developed for a continuous exposure [8], but their main interest is not for AR estimation but, rather, for the estimation of a related quantity; namely, the generalized impact fraction (see below).

Multiple Risk Factors

When there are several risk factors at play, it is useful to estimate AR for each risk factor separately as well as an overall AR for all risk factors jointly. Contrary to some investigators' intuition, the sum of AR estimates for each risk factor does not equal the overall AR except in special circumstances. Walter [76] showed that the equality holds for two risk factors if and only if either no subject is exposed to both risk factors or the effect of the two risk factors on disease incidence is additive. This generalizes into a set of J sufficient conditions when more than two factors forming J levels are taken into account [76]. Another important result is that, if the risk factors are statistically independent and their joint effect on disease incidence is **multiplicative** (i.e. no interaction on a multiplicative scale is present), then the complement to 1 of the overall AR is equal to the product of the complements to 1 of the separate ARs [14, 53, 74].

Finally, it has been recommended to consider a single exposed level defined by exposure to at least one risk factor, and a reference level defined by exposure to no risk factor, in order to estimate the overall AR for several risk factors [5, 71]. However, this procedure, while appealingly simple, can lead to a very small reference level and thus a very unstable AR estimate. For this reason, some authors prefer to use the model-based approach (6) and retain one parameter for each exposure factor in an overall relative risk or odds ratio model to obtain a more stable AR estimate [20, 40].

Misclassification of Exposure

The effects of **misclassification** of exposure have been studied by several authors [35, 71, 76]. AR has a "canceling feature" [76] in that misclassification may result in compensatory effects. For example, if misclassification of exposure is **nondifferential**, reduced **specificity** of exposure classification (marked by the presence of **false positive** subjects in terms of exposure) biases the odds ratio or relative risk towards the null (*see* **Bias Toward the Null**), but exposure prevalence increases, so that the net result is an absence of bias. However, still for nondifferential misclassification, reduced **sensitivity** (marked by the presence of **false negative** subjects in terms of exposure) biases AR estimates towards the null. Hsieh & Walter [35] gave a formal proof of this result and Wacholder et al. [71] a heuristic proof based on the distributive property of AR. Moreover, this downward bias increases with the prevalence of exposure.

Thus, in order to minimize bias in estimating AR, a sensitive classification scheme is an appropriate strategy, even when specificity is exceedingly low. In other words, the estimate of AR is unbiased as long as all exposed individuals are classified as exposed, regardless of the proportion of unexposed subjects who are misclassified nondifferentially as exposed. This is illustrated by Wacholder et al. [71]

with case–control data on mesothelioma and asbestos exposure (see the section "Properties" above). As exposure to asbestos is hard to prove, there are categories with "moderately low" or "medium" probability of exposure in their example. They recommend to consider these subjects as exposed in order to obtain a perfectly sensitive classification. However, there is a price to pay when high sensitivity is obtained at the expense of reduced specificity. Precision decreases (the variance of the AR estimate increases) when the definition of exposure encompasses levels of exposure that have the same risk of disease as the unexposed [71]. Therefore, there might be a tradeoff between bias and precision.

If misclassification is differential with diseased subjects (cases) being falsely classified as exposed more often that nondiseased subjects (controls), then the estimate of AR, as well as of relative risk or odds ratio, will be biased upward [71].

Finally, it should be noted that Hsieh studied the effect of disease status misclassification ("outcome misclassification") on AR estimation, and defined conditions under which bias occurs in this less common situation [34].

Use of AR to Determine Sample Size

Browner & Newman [13] derived formulas for **sample size determination** in case–control studies that are based on AR instead of odds ratio. Upon comparing sample size and power estimates based on the detection of a given AR with conventional estimates based on the detection of a given odds ratio, they found the following results. For a rare dichotomous exposure, case–control studies having little power to detect a small odds ratio may still have adequate power to detect a small AR. However, even relatively large case–control studies may have inadequate power to detect a small AR when the exposure is common. Such sample size calculations may be useful when the public health importance of an association is of primary interest, as further discussed by Coughlin et al. [16] and Adams et al. [1].

Ordinal Data

Basu & Landis [4] considered the situation where the disease classification is not dichotomous (diseased, nondiseased) but includes more than two **ordered categories** (e.g. none, mild, moderate, severe) and the exposure factor has at least two ordinal levels (e.g. none, low, medium, high exposure). They extended the concept of AR to that special case in order to quantify the potential extent of disease reduction in the target population relative to each increasing level of the ordinal disease classification, which could be realized if the exposure factor were eliminated. They developed model-based estimates based on a cumulative logit model, assuming a proportional odds structure for cohort, case–control and cross-sectional designs, and obtained corresponding variance estimates based on the delta method for implicitly related random variables [7].

Recurrent Disease Events

Pichlmeier & Gefeller [58] extended the concept of AR to diseases that may recur, such as some skin diseases (e.g. urticaria, psoriasis) or chronic diseases like asthma, epilepsy, or multiple sclerosis. They defined the "recurrent attributable risk" as the proportion of disease events (first occurrence plus recurrences) that can be attributed to an exposure factor. This concept is of interest for risk factors that also act as **prognostic factors** of recurrences. Point estimators were derived for cohort, case–control and cross-sectional designs based on the unadjusted, weighted-sum and MH approaches. Corresponding approximate variance estimators were developed using the delta method [60].

Conceptual Problems

The public health interpretation of AR refers to the proportion of cases that are excess cases, i.e. that would not have occurred if exposure had not occurred. Greenland & Robins [33, 61] identified a different concept. From a biologic or legal perspective, one might refer to cases for which exposure played an etiologic role, i.e. cases for which exposure was a contributory cause of the outcome. They argued for the distinction between "excess fraction" and "etiologic fraction" to refer to the standard and new concept, respectively. While the "excess fraction" can be estimated under the usual conditions for validity of an epidemiologic study (e.g. lack of biases), estimation of the etiologic fraction requires nonidentifiable biologic assumptions about exposure action and interactions [61]. It is true, however, that the interpretation of the excess fraction also depends

on considerations of causality, as discussed in the section "Interpretation and Usefulness" above.

Related Quantities

AR in Exposed

The attributable risk in the exposed or **attributable fraction in the exposed** (AF_E) is defined as the proportion of disease cases that can be attributed to an exposure factor among the exposed subjects only [15, 45, 47, 53]. It can be formally written as:

$$AF_E = \frac{[\Pr(D|E) - \Pr(D|\overline{E})]}{\Pr(D|E)}, \qquad (7)$$

where $\Pr(D|E)$ is the probability of disease in the exposed individuals (E) and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same subjects but with all exposure eliminated.

Excess Incidence

Excess incidence δ is defined as the difference between the incidence rate in the exposed and the incidence rate in the unexposed, or $\delta = \lambda_1 - \lambda_0$ [9, 47, 51]. It takes into account the incidence of the disease in the unexposed and the strength of the association between exposure and disease, as it can be rewritten as $\delta = \lambda_0(RR - 1)$. It can be seen to equal the numerator of the AR in the exposed (7) if the latter is expressed in terms of incidence rates rather than probabilities. Its main interest lies, however, at the individual, rather than the population, level. It quantifies the difference in incidence that can be attributed to exposure for an individual. Other terms have been used to denote this quantity; namely, "**excess risk**" [63], "Berkson's simple difference" [74], "incidence density difference" [54], "excess prevalence" [74], or even "attributable risk" [50, 63], which may have introduced some confusion. Moreover, it should not be confused with the concept of "excess fraction" [33, 61] (see above). Estimation of δ pertains to estimation of incidence rates.

Prevented Fraction

When considering a protective exposure or intervention, an intuitively appealing alternative to attributable risk (AR) is the prevented fraction (PF). The prevented fraction measures the impact of an association between a protective exposure and disease at the population level. It is sometimes called the **preventable fraction**, although this term may have a different meaning. It is defined as the proportion of disease cases averted by a protective exposure or intervention [53]. It can be written formally as:

$$PF = \frac{[\Pr(D|\overline{E}) - \Pr(D)]}{\Pr(D|\overline{E})}, \qquad (8)$$

where $\Pr(D)$ is the probability of disease in the population, which may have some exposed (E) and some unexposed ($\overline{E}$) individuals, and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same population but with all (protective) exposure eliminated. Another formulation of PF is the proportion of cases prevented by the (protective) factor or intervention among the totality of cases that would have developed in the absence of the factor or intervention [53], which is why the denominator in (8) is the hypothetical probability of disease in the population *in the absence of the protective factor*.

Generalized Impact Fraction

The generalized impact fraction (or generalized attributable fraction) was introduced by Walter [75] and Morgenstern & Bursic [56] as a measure that generalizes AR. It is defined as the fractional reduction of disease that would result from changing the current distribution of exposure in the population to some modified distribution; namely, $[\Pr(D) - \Pr^*(D)]/\Pr(D)$, where $\Pr(D)$ and $\Pr^*(D)$, respectively, denote the probability of disease under the current distribution of exposure and under the modified distribution of exposure. AR corresponds to the special case in which the modified distribution puts unit mass on the lowest risk configuration and can be used to assess interventions aimed at eliminating exposure. A level-specific AR corresponds to the special case where the modified distribution of exposure differs from the current distribution in that subjects at the specified level of exposure are brought to the lowest risk configuration and can be used to assess interventions aimed at eliminating exposure in that specified group only. The generalized impact fraction is a general measure that can be used to assess various interventions,

targeting all subjects or subjects at specified levels, and aimed at modifying the exposure distribution (reducing exposure), but not necessarily eliminating exposure.

It has been used, for instance, by Lubin & Boice [46] who considered the impact on lung cancer of a modification in the distribution of radon exposure consisting in truncating the current distribution at various thresholds. Wahrendorf [72] used this concept to examine the impact of various changes in dietary habits on colorectal and stomach cancers.

Methods to estimate the generalized impact fraction are similar to methods for estimating AR. However, unlike for AR, it might be useful to retain the continuous nature of risk factors to define the modification of the distribution considered (for instance, a shift in the distribution), and extensions of methods for estimating AR for continuous factors [8] are useful. Drescher & Becher [23] proposed extending the model-based approaches of Bruzzi et al. [14] and Greenland & Drescher [31] to estimate the generalized impact fraction in case–control studies and considered categorical as well as continuous exposure factors.

Probability of Causation – Assigned Share

Cox [18, 19] proposed a method of partitioning the increase in disease risk among several risk factors for subjects jointly exposed to them. The part corresponding to each factor is called the assigned share or probability of causation [6, 18, 19, 41, 64, 65]. It is useful in a legal context to assign shares of responsibility to risk factors in tort liability cases but does not have a population interpretation, unlike AR [6]. The assigned share enjoys the additive property that the sum of separate assigned shares for two (or more) factors is equal to the joint assigned share for these factors [19].

Prospects and Conclusion

Although most important issues about AR estimation have been settled, research is still needed on specific points, such as the use of resampling methods (*see* **Bootstrap Method**) to estimate variance and confidence intervals [30, 37], the development of model-based estimates based on conditional logistic regression for stratified or frequency-matched, case–control studies [8], improvements of the weighted-sum approach in case–control studies, or the development of AR estimators for case–cohort designs and complex survey designs (e.g. **cluster sampling**). Research is also needed for issues regarding quantities related to AR (see above), special problems (see above), and for software development [52] (*see* **Software, Biostatistical**).

The biggest challenge at this point, however, might be the need to encourage the proper use and interpretation of AR in practice and to make investigators aware of correct estimation techniques.

References

[1] Adams, M.J., Khoury, M.J. & James, L.M. (1989). The use of attributable fractions in the design and interpretation of epidemiologic studies, *Journal of Clinical Epidemiology* **42**, 659–662.

[2] Alavanja, M.C.R., Brownson R.C., Benichou, J., Swanson, C. & Boice, J.D. (1995). Attributable risk of lung cancer in lifetime nonsmokers and long-term exsmokers (Missouri, USA), *Cancer Causes and Control* **6**, 209–216.

[3] Basu, S. & Landis, J.R. (1995). Model-based estimation of population attributable risk under cross-sectional sampling, *American Journal of Epidemiology* **142**, 1338–1343.

[4] Basu, S. & Landis, J.R. (1993). Model-based estimates of population attributable risk for ordinal data, Personal communication.

[5] Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case-control studies: a review, *Statistics in Medicine* **10**, 1753–1773.

[6] Benichou, J. (1993). Re: "Methods of adjustment for estimating the attributable risk in case-control studies: a review" (letter), *Statistics in Medicine* **12**, 94–96.

[7] Benichou, J. & Gail, M.H. (1989). A delta-method for implicitly defined random variables, *American Statistician* **43**, 41–44.

[8] Benichou, J. & Gail, M.H. (1990). Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models, *Biometrics* **46**, 991–1003.

[9] Berkson, J. (1958). Smoking and lung cancer. Some observations on two recent reports, *Journal of the American Statistical Association* **53**, 28–38.

[10] Birch, M.W. (1964). The detection of partial associations, I: the 2×2 case, *Journal of the Royal Statistical Society, Series B* **27**, 313–324.

[11] Breslow, N.E. (1981). Odds ratio estimators when the data are sparse, *Biometrika* **68**, 73–84.

[12] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. 1: *The Analysis of Case-Control Studies*. Scientific Publications No. 32, International Agency for Research on Cancer, Lyon.

[13] Browner, W.S. & Newman, T.B. (1989). Sample size and power based on the population attributable fraction, *American Journal of Public Health* **79**, 1289–1294.

[14] Bruzzi, P., Green S.B., Byar, D.P., Brinton, L.A. & Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data, *American Journal of Epidemiology* **122**, 904–914.

[15] Cole, P. & MacMahon, B. (1971). Attributable risk percent in case-control studies, *British Journal of Preventive and Social Medicine* **25**, 242–244.

[16] Coughlin, S.S., Benichou, J. & Weed, D.L. (1994). Attributable risk estimation in case-control studies, *Epidemiologic Reviews* **16**, 51–64.

[17] Coughlin, S.S., Nass, C.C., Pickle, L.W., Trock, B. & Bunin, G. (1991). Regression methods for estimating attributable risk in population-based case-control studies: a comparison of additive and multiplicative models, *American Journal of Epidemiology* **133**, 305–313.

[18] Cox, L.A. (1984). Probability of causation and the attributable proportion of risk, *Risk Analysis* **4**, 221–230.

[19] Cox, L.A. (1985). A new measure of attributable risk for public health applications, *Management Science* **7**, 800–813.

[20] D'Avanzo, B., La Vecchia C., Negri, E., Decarli, A. & Benichou, J. (1995). Attributable risks for bladder cancer in Northern Italy, *Annals of Epidemiology* **5**, 427–431.

[21] Denman, D.W. & Schlesselman, J.J. (1983). Interval estimation of the attributable risk for multiple exposure levels in case-control studies, *Biometrics* **39**, 185–192.

[22] Deubner, D.C., Wilkinson, W.E., Helms, M.J., Tyroler, H.A. & Hames, C.G. (1980). Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia, *American Journal of Epidemiology* **112**, 135–143.

[23] Drescher, K. & Becher, H. (1997). Estimating the generalized attributable fraction from case-control data, *Biometrics* **53**, 1170–1176.

[24] Drescher, K. & Schill, W. (1991). Attributable risk estimation from case-control data via logistic regression, *Biometrics* **47**, 1247–1256.

[25] Ejigou, A. (1979). Estimation of attributable risk in the presence of confounding, *Biometrical Journal* **21**, 155–165.

[26] Gefeller, O. (1990). A simulation study on adjusted attributable risk estimators, *Statistica Applicata* **2**, 323–331.

[27] Gefeller, O. (1992). Comparison of adjusted attributable risk estimators, *Statistics in Medicine* **11**, 2083–2091.

[28] Greenland, S. (1984). Bias in methods for deriving standardized mortality ratio and attributable fraction estimates, *Statistics in Medicine* **3**, 131–141.

[29] Greenland, S. (1987). Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data, *Statistics in Medicine* **6**, 701–708.

[30] Greenland, S. (1992). The bootstrap method for standard errors and confidence intervals of the adjusted attributable risk (letter), *Epidemiology* **3**, 271.

[31] Greenland, S. & Drescher, K. (1993). Maximum-likelihood estimation of the attributable fraction from logistic models, *Biometrics* **49**, 865–872.

[32] Greenland, S. & Morgenstern, H. (1983). Morgenstern corrects a conceptual error (letter), *American Journal of Public Health* **73**, 703–704.

[33] Greenland, S. & Robins, J.M. (1988). Conceptual problems in the definition and interpretation of attributable fractions, *American Journal of Epidemiology* **128**, 1185–1197.

[34] Hsieh, C.C. (1991). The effect of non-differential outcome misclassification on estimates of the attributable and prevented fraction, *Statistics in Medicine* **10**, 361–373.

[35] Hsieh, C.C. & Walter, S.D. (1988). The effect of non-differential misclassification on estimates of the attributable and prevented fraction, *Statistics in Medicine* **7**, 1073–1085.

[36] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont.

[37] Kooperberg, C. & Petitti, D.B. (1991). Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study, *Epidemiology* **2**, 363–366.

[38] Kuritz, S.J. & Landis, J.R. (1988). Summary attributable risk estimation from unmatched case-control data, *Statistics in Medicine* **7**, 507–517.

[39] Kuritz, S.J. & Landis, J.R. (1988). Attributable risk estimation from matched case-control data, *Biometrics* **44**, 355–367.

[40] La Vecchia C., D'Avanzo, B., Negri, E., Decarli, A. & Benichou, J. (1995). Attributable risks for stomach cancer in Northern Italy, *International Journal of Cancer* **60**, 748–752.

[41] Lagakos, S.W. & Mosteller, F. (1986). Assigned shares in compensation for radiation-related cancers (with discussion), *Risk Analysis* **6**, 345–380.

[42] Landis, J.R., Heyman, E.R. & Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests, *International Statistical Review* **46**, 237–254.

[43] Last, J.M. (1983). *A Dictionary of Epidemiology*. Oxford University Press, New York.

[44] Leung, H.K. & Kupper, L.L. (1981). Comparison of confidence intervals for attributable risk, *Biometrics* **37**, 293–302.

[45] Levin, M.L. (1953). The occurrence of lung cancer in man, *Acta Unio Internationalis contra Cancrum* **9**, 531–541.

[46] Lubin, J.H. & Boice, J.D. Jr (1989). Estimating Rn-induced lung cancer in the United States, *Health Physics* **57**, 417–427.

[47] MacMahon, B. & Pugh, T.F. (1970). *Epidemiology: Principles and Methods*. Little, Brown, & Company, Boston.

[48] Madigan, M.P., Ziegler, R.G., Benichou, J., Byrne, C. & Hoover, R.N. (1995). Proportion of breast cancer cases in the United States explained by well-established risk factors, *Journal of the National Cancer Institute* **87**, 1681–1685.

[49] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[50] Markush, R.E. (1977). Levin's attributable risk statistic for analytic studies and vital statistics, *American Journal of Epidemiology* **105**, 401–406.

[51] Mausner, J.S. & Bahn, A.K. (1974). *Epidemiology: An Introductory Text*. W.B. Saunders, Philadelphia.

[52] Mezzetti, M., Ferraroni, M., Decarli, A., La Vecchia, C. & Benichou, J. (1996). Software for attributable risk and confidence interval estimation in case-control studies, *Computers and Biomedical Research* **29**, 63–75.

[53] Miettinen, O.S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention, *American Journal of Epidemiology* **99**, 325–332.

[54] Miettinen, O.S. (1976). Estimability and estimation in case-referent studies, *American Journal of Epidemiology* **103**, 226–235.

[55] Morgenstern, H. (1982). Uses of ecologic analysis in epidemiological research, *American Journal of Public Health* **72**, 1336–1344.

[56] Morgenstern, H. & Bursic, E.S. (1982). A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population, *Journal of Community Health* **7**, 292–309.

[57] Ouellet, B.L., Romeder, J.M. & Lance, J.M. (1979). Premature mortality attributable to smoking and hazardous drinking in Canada, *American Journal of Epidemiology* **109**, 451–463.

[58] Pichlmeier, U. & Gefeller, O. (1997). Conceptual aspects of attributable risk in the case of recurrent disease events, *Statistics in Medicine* **16**, 1107–1120.

[59] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[60] Rao, C.R. (1965). *Linear Statistical Inference and Its Application*. Wiley, New York, pp. 319–322.

[61] Robins, J.M. & Greenland, S. (1989). Estimability and estimation of excess and etiologic fractions, *Statistics in Medicine* **8**, 845–859.

[62] Robins, J.M., Breslow, N.E. & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse-data and large-strata limiting models, *Biometrics* **42**, 311–323.

[63] Schlesselman, J.J. (1982). *Case-Control Studies. Design, Conduct and Analysis*. Oxford University Press, New York.

[64] Seiler, F.A. (1986). Attributable risk, probability of causation, assigned shares, and uncertainty, *Environment International* **12**, 635–641.

[65] Seiler, F.A. & Scott, B.R. (1987). Mixture of toxic agents and attributable risk calculations, *Risk Analysis* **7**, 81–90.

[66] Silverman, D.T., Brown, L.M., Hoover, R.N., Schiffman, M., Lillemoe, K.D., Schoenberg, J.B., Swanson, G.M., Hayes, R.B., Greenberg, R.S., Benichou, J., Schwartz, A.G., Liff, J.F. & Pottern, L.M. (1995). Alcohol and pancreatic cancer in Blacks and Whites in the United States, *Cancer Research* **55**, 4809–4905.

[67] Spirtas, R., Heineman, E.F., Bernstein, L., Beebe, G.W., Keehn, R.J., Stark, A.S., Harlow, B.L. & Benichou, J. (1994). Malignant melanoma: attributable risk of asbestos exposure, *Occupational and Environmental Medicine* **51**, 804–811.

[68] Sturmans, F., Mulder, P.G.H. & Walkenburg, H.A. (1977). Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage, *American Journal of Epidemiology* **105**, 281–289.

[69] Tarone, R.E. (1981). On summary estimators of relative risk, *Journal of Chronic Diseases* **34**, 463–468.

[70] Tuyns, A.J., Pequignot, J. & Jensen, O.M. (1977). Le cancer de l'oesophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac, *Bulletin of Cancer* **64**, 45–60.

[71] Wacholder, S., Benichou, J., Heineman, E.F., Hartge, P. & Hoover, R.N. (1994). Attributable risk: advantages of a broad definition of exposure, *American Journal of Epidemiology* **140**, 303–309.

[72] Wahrendorf, J. (1987). An estimate of the proportion of colo-rectal and stomach cancers which might be prevented by certain changes in dietary habits, *International Journal of Cancer* **40**, 625–628.

[73] Walter, S.D. (1975). The distribution of Levin's measure of attributable risk, *Biometrika* **62**, 371–374.

[74] Walter, S.D. (1976). The estimation and interpretation of attributable risk in health research, *Biometrics* **32**, 829–849.

[75] Walter, S.D. (1980). Prevention for multifactorial diseases, *American Journal of Epidemiology* **112**, 409–416.

[76] Walter, S.D. (1983). Effects of interaction, confounding and observational error on attributable risk estimation, *American Journal of Epidemiology* **117**, 598–604.

[77] Whittemore, A.S. (1982). Statistical methods for estimating attributable risk from retrospective data, *Statistics in Medicine* **1**, 229–243.

[78] Whittemore, A.S. (1983). Estimating attributable risk from case-control studies, *American Journal of Epidemiology* **117**, 76–85.

JACQUES BENICHOU

Autocorrelation Function

An autocorrelation is simply a correlation between two random variables X_t and X_s that are both part of the same **time series** (or other stochastic process, e.g. a spatial random field). From a single realization of the process it is possible to estimate the autocorrelations only if the process may be assumed to be second-order stationary, so that the autocorrelation between X_t and X_s depends only on the difference or *lag* between t and s, $\tau = t - s$. (In a spatial process, the autocorrelation function may also depend on the direction of the displacement between t and s unless the process is isotropic). For second-order stationary processes, the autocorrelation function (ACF), denoted ρ_τ, is defined as the correlation between variables separated by a lag τ.

If the process has been observed at equally spaced points, $t = 1, \ldots, N$, then the autocorrelation function ρ_τ at lag τ may be estimated by

$$r_\tau = \frac{c_\tau}{c_0},$$

where c_τ, the sample autocovariance function, is given by

$$c_\tau = \sum_t^{N-k} \frac{(x_t - \overline{x})(x_{t+\tau} - \overline{x})}{N}. \qquad (1)$$

The sampling properties of this estimator are discussed in [2, Chapter 48] where it is shown that the bias in c_τ is of the order $1/N$. To reduce this bias for small N, an alternative estimator is sometimes used in which the denominator N in (1) is replaced by $N - \tau$. However, Jenkins & Watts [1] show that the alternative estimator generally has a larger mean square error than (1). Also, the advantage of (1) is that it yields positive semidefinite autocovariances, a useful property for estimating the spectrum. For a purely random process with an autocorrelation function equal to zero for $\tau > 0$, the mean and variance of the sample autocorrelations are approximately given by

$$\mathrm{E}(r_\tau) \approx \frac{-1}{N}$$

and

$$\mathrm{var}(r_\tau) \approx \frac{-1}{N},$$

respectively.

A plot of the sample autocorrelation function, called a *correlogram*, may be used to explore the time series. For example, a slowly decreasing correlogram may be due to nonstationarity (i.e. a trend in the time series) and an oscillating correlogram indicates seasonal fluctuations. The correlogram is also helpful in identifying the order q of a moving average process [MA(q)], since the autocorrelations of an MA(q) are nonzero only for lags $\leq q$. Confidence limits of $-1/N \pm 2/\sqrt{N}$ may be drawn on a correlogram to help assess which autocorrelations differ significantly from zero. For autoregressive processes [AR(p)], the autocorrelation function decreases only slowly and the *partial* autocorrelation function is more useful for identifying the order of the process. An example of a time series and its correlogram are given in Figure 1 for the lutenizing hormone levels in blood samples from a healthy woman.

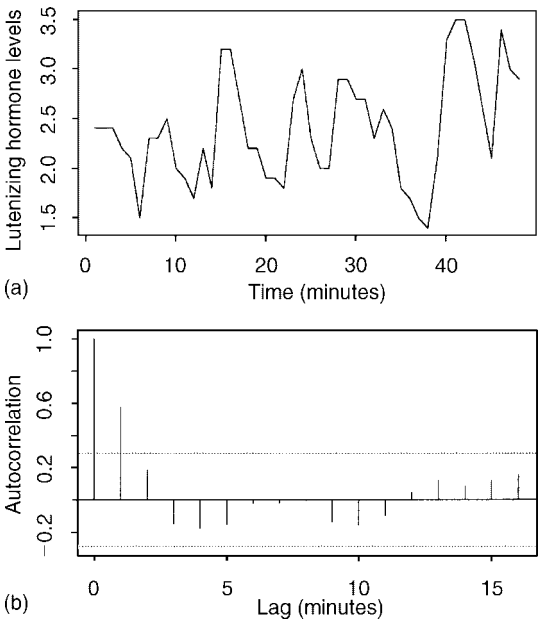

Figure 1 Levels of lutenizing hormone in blood samples taken from a healthy woman every 10 minutes (a) and the autocorrelation function with approximate 95% confidence limit for zero autocorrelation (b)

References

[1] Jenkins, G.M. & Watts, D.G. (1968). *Spectral Analysis and its Applications*. Holden-Day, San Francisco.

[2] Kendall, M.G., Stuart, A. & Ord, J.K. (1983). *The Advanced Theory of Statistics*, Vol. 3. Griffin, London.

(*See also* **ARMA and ARIMA Models**; **Coherence Between Time Series**; **Spectral Analysis**)

SOPHIA RABE-HESKETH

Average Age at Death

The fallacy of using average age at death as an "alias" to summarize **life expectancy** and other aspects of mortality plays a prominent role in the history of statistics [1, p. 23]. While its use for this purpose is tempting because of its availability, it is nonetheless incorrect [2]. Proper analysis of mortality involves the determination of age-specific mortality rates, which requires denominator data on the age distribution of the population [2], and failure to account for the age distributions of the underlying populations is the principal determinant of this fallacy (*see* **Denominator Difficulties**). Age-distribution data come from the total population being studied, including those who are still alive as well as those who have died.

Although summarizing mortality using the average age of death may be a convenient measurement, it (alone) is often neither a useful nor helpful measure [2, 3]. Information regarding frequency distributions and variability around the average also should be considered. Whereas, for example, the average age at death from a particular disease may be 55 years, all deaths may have occurred in persons younger than 50 years and older than 60 years – information that is not conveyed by the **mean** age alone.

Furthermore, average age of death can be a misleading statistic for other reasons [2, 3]. For instance, in the comparison of longevity among persons in various occupations (*see* **Occupational Mortality**), the average age at death depends, at least in part, on the age at entry into an occupation as well as age at exit, if exit occurs for reasons other than death. **Death certificates** provide age at death data; age at job entry and exit are less readily available. Also, average age at death may be determined by the intensity of the exposure to risk as well as by the duration of the exposure.

Andersen [1, p. 13] remarked on a study comparing the average length of life for male symphony orchestra conductors and for the entire US male population.

On average, the conductors lived about 4 years longer. The methodological flaw was that because age at entry was birth, those in the US male population who died in infancy and childhood were included in the calculation of the average life span, whereas only men who survived to become conductors could enter the conductor cohort. The apparent difference in longevity disappeared after accounting for **infant and perinatal mortality** in the US male population.

In comparing two groups with respect to mortality experience, both Andersen [1] and Colton [2] emphasized the importance of considering the age distribution of the groups. Arguments about life span based on the average age at death ignore those who are still alive. If among those still living there are more who are elderly in one group compared with the other, then differential mortality experience does not necessarily explain group differences in average age at death. Even if the groups had identical age-specific death rates, the group with the larger number of elderly individuals will have the higher average age of death.

Rather than looking at average age at death, Rothman [4] suggests that we should compare the *risk of death* among orchestra conductors (or whatever group is being looked at) with the risk of death among other people who have attained the same ages as the conductors. Average age at death is only a characteristic of those who die and does not reflect the risk of death.

References

[1] Andersen, B. (1990). Miscellaneous examples, in *Methodological Errors in Medical Research. An Incomplete Catalogue*. Blackwell Scientific, Oxford, pp. 12–24.

[2] Colton, T. (1974). Fallacies in numerical reasoning, in *Statistics in Medicine*. Little, Brown & Company, Boston, pp. 289–313.

[3] Hill, A.B. & Hill, I.D. (1991). Fallacies and difficulties: incidence and causes of mortality, in *Bradford Hill's Principles of Medical Statistics*, 12th Ed. Edward Arnold, London, pp. 258–263.

[4] Rothman, K.J. (2002). Introduction to epidemiologic thinking, in *Epidemiology: An Introduction*. Oxford University Press, New York, pp. 5–6.

HOWARD M. KRAVITZ

Axes in Multivariate Analysis

Visual presentation of data is valuable throughout any statistical analysis, right from the preliminary stages (where graphical examination of sample values can often reveal features that are difficult to detect in a table of numbers) through to the conclusion (where graphs and charts often provide the most effective way of presenting the results; *see* **Graphical Displays**). **Multivariate analysis** particularly benefits from pictorial support, as a typical multivariate sample contains too many numbers to be readily assimilable in tabular form. Indeed, many multivariate techniques can be regarded primarily as mechanisms for systematic exploration of multidimensional sample spaces in which sample individuals are represented as points. In this article we highlight just one important aspect of such representations.

Suppose that p variables $X_1, X_2, \ldots, X_p$ have been measured on each of n sample individuals, and write $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ for the vector of p values observed on the ith individual, $i = 1, 2, \ldots, n$. Assume for the present that all variables are *quantitative*, i.e. on a numerical scale (*see* **Measurement Scale**). Then the sample can be modeled geometrically as a swarm of n points in p-dimensional space, by associating each variable X_j with an **orthogonal** axis in this space and assigning the observed value $\mathbf{x}_i$ to the point with coordinates $(x_{i1}, x_{i2}, \ldots, x_{ip})$ on these axes. Thus the original variables constitute a fundamental system of axes in multidimensional space. Furthermore, if the data matrix has been mean-centered, then the origin of these axes is at the centroid of the swarm of points.

However, this geometrical model is of little immediate practical utility because we can graphically depict only a few dimensions (usually just two), but multivariate data sets generally involve more than two variables. Some approximation is therefore necessary, and many techniques of multivariate analysis are concerned with identifying either single *directions* in the sample space along which something "interesting" happens, or (low-dimensional) *subspaces* into which the points should be *projected* to highlight some relevant sample features. The two objectives are in essence the same, as any k-dimensional subspace can be defined simply by specifying k mutually orthogonal directions in the original space to act as coordinate axes for the subspace. Also, since most multivariate analyses operate on mean-centered data, any such subspace axes are essentially lines through the origin of the sample space.

Now any line $\mathcal{L}$ through the origin can be specified by a unit vector starting at the origin and having an end point on the line. If the end point of this vector has coordinates $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$, then unit length implies the condition $\sum_i a_i^2 = 1$, so the a_i can be interpreted as direction **cosines** of the vector with each of the original axes. The coordinates of any point $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ in the sample space can likewise be thought of as a vector from the origin to this point, so by elementary vector theory it follows that the projection of the point $\mathbf{x}$ onto the line $\mathcal{L}$ is a distance $\mathbf{a}'\mathbf{x} = \sum_i a_i x_i$ from the origin. We can therefore view $\mathcal{L}$ as an *axis* in the original space, and the point $\mathbf{x}$ has coordinate value $\mathbf{a}'\mathbf{x}$ on this axis. For convenience, the defining unit vector $\mathbf{a}$ is often referred to as the line $\mathcal{L}$.

We can specify any number of axes $\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}, \ldots$ in this way. Two such axes $\mathbf{a}_i, \mathbf{a}_j$ are *orthogonal* (i.e. at right angles to each other) if $\mathbf{a}_i'\mathbf{a}_j = 0$, and if the sample space is of dimension p, then *any* set of p mutually orthogonal lines can be chosen as reference axes for it. A particular set of p mutually orthogonal $\mathbf{a}_i$ constitutes a **rotation** of the original coordinate axes. Taking a subset of k of these axes defines a k-dimensional subspace of the original space into which the data swarm can be projected.

Given the association between the measured variables $X_1, X_2, \ldots, X_p$ and the original axes in the sample space, we thus see that all those multivariate techniques that obtain linear combinations of the X_i are in fact identifying new axes in the sample space. Moreover, if the linear combinations are orthogonal, then so are the corresponding axes, in which case k such linear combinations define a k-dimensional subspace of the sample space. Such is the case, for example, with successive components in **principal component analysis** and with the linear combinations derived in various forms of **projection pursuit**. In these cases, projecting the data points into several such orthogonal dimensions enables us to obtain a low-dimensional approximation to the full data representation; this is the objective, for example, of principal component score plots.

Other multivariate techniques result in linear combinations that are either derived implicitly from a

statistical model, or ones that are nonorthogonal. In the former case, for example in **factor analysis**, we can still view the combinations as defining axes and subspaces of the sample space, but direct projection of points into these subspaces may not necessarily correspond to derived scores. In the latter case, for example in canonical **discriminant analysis**, the axes will be **oblique**, so care must be taken with projection or representation of points.

A final point worth noting about such data representations is that it is often instructive to project the original axes into any k-dimensional subspace generated from orthogonal linear combinations, as these projections show the inclination of the derived subspace to the original axes. Such projections are obtained as biplots in principal component analysis, for example (*see* **Graphical Displays**).

The above ideas are based on the assumption that all the variables are quantitative, allowing the formulation of an underlying model of points in space with coordinates given by variable values. Many multivariate data sets, however, contain **categorical data**, either, **nominal** or ordinal variables (*see* **Ordered Categorical Data**), which do not permit the direct formulation of such a model. It is nevertheless possible to *construct* a model by means of **multidimensional scaling**. This construction requires a matrix of dissimilarities between every pair of sample members to be calculated (*see* **Similarity, Dissimilarity, and Distance Measure**), whereupon the scaling technique will find a k-dimensional configuration of points representing sample individuals in which between-point distances approximate between-individual dissimilarities as closely as possible. New axes and subspaces can be sought in this representation in the same way as above. However, since the multidimensional scaling method does not associate variables with coordinate axes in the constructed representation, there is no longer any association between new axes and linear combinations of variables.

(*See also* **Battery Reduction**; **Matrix Algebra**)

W.J. KRZANOWSKI

Axioms of Probability

Probability theory, like many other branches of mathematics such as geometry, for example, is a subject whose development arose out of an attempt to provide a rigorous mathematical model for observable real-world phenomena (*see* **Foundations of Probability**). The real-world phenomenon in the case of probability theory is chance or random behavior involving a physical system or a biological process. While some probabilistic ideas date back to India as early as the fifth century B.C. [2, p. 1], the formal study of probability is generally attributed to have its origin in the correspondence between Fermat and Pascal in the 1650s concerning various gambling issues. Today, probability theory has numerous applications in diverse fields such as statistical inference, number theory, the physics of particle movement, economics, the social sciences, the biological sciences, genetics, epidemiology, and demography.

The probability of an event is the abstract counterpart to the real-world notion of the long-run relative frequency of the occurrence of the event through replicating the experiment over and over again. For example, if a medical researcher asserts that the probability that a particular medical procedure results in a cure for those inflicted with a certain disease is (say) 0.85, then the researcher is asserting that in the long run 85% of those inflicted with the disease who receive the medical procedure will be cured. The phrase "in the long run" suggests that the theoretical underpinnings involve the notion of *limit* as the sample size $n \to \infty$. While the long-run relative frequency approach for defining the probability of an event may seem natural and intuitive, it raises serious mathematical questions. Does the limit of the relative frequency always exist as $n \to \infty$ and is the limit always the same irrespective of the experimental outcome? It is easy to see that the answers are in the negative. Indeed, it is within the realm of possibility that in the example above the proportion cured fluctuates repeatedly from near 0 to near 1 as $n \to \infty$. So in what sense can it be asserted that the limit exists and equals 0.85? The answer to this question is provided using the axiomatic (or measure-theoretic) approach.

The problems arising from the relative frequency approach are eliminated by the axiomatic approach which was developed by Kolmogorov in [1]. Kolmogorov's probability model is defined in terms of a triplet $(\Omega, \mathcal{F}, P)$ (called a *probability space*) whose components will now be described via the following four axioms:

Axiom 1. Ω is a nonempty reference set.

Axiom 2. $\mathcal{F}$ is a *σ-algebra* of subsets of Ω, that is, $\mathcal{F}$ is a nonempty collection of subsets of Ω satisfying
- (i) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ and
- (ii) $\{A_n, n \geq 1\} \subseteq \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Axiom 3. P is a *measure* on $\mathcal{F}$, that is, P is a real valued function defined on $\mathcal{F}$ satisfying
- (i) $P(\emptyset) = 0$,
- (ii) $P(A) \geq 0$ for each $A \in \mathcal{F}$, and
- (iii) if $\{A_n, n \geq 1\}$ is a sequence of disjoint sets in $\mathcal{F}$, then $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

Axiom 4. $P(\Omega) = 1$.

The set Ω is the abstract counterpart of the collection of primitive outcomes of a not completely determined real-world experiment. The objects $\omega \in \Omega$ are called *sample points* and Ω is called the *sample space* of the experiment. A member of $\mathcal{F}$ is referred to as an *event*. For $A \in \mathcal{F}$, its *probability* $P(A)$ is the abstract counterpart to the long-run or limiting relative frequency of the occurrence of A when the experiment is indefinitely repeated.

It is natural to take $\mathcal{F}$ to be the power set of Ω if Ω is countable. (The power set of Ω is the set of all subsets of Ω.) However, if Ω is uncountable, then profound measure-theoretic considerations would force such a choice of $\mathcal{F}$ to preclude the existence of a probability measure P on $\mathcal{F}$. Consequently, $\mathcal{F}$ can be smaller than the power set of Ω but should be large enough so as to contain all subsets of Ω whose probability would be of practical or theoretical interest.

It should be apparent that probability theory as a subject has two sides. On one side is the mathematical use of measure theory, whereas the other side concerns a random experiment arising in connection with a physical system or biological process. The measure-theoretic side gives probability theory its mathematical rigor; the experimental side gives it its application and often its inspiration.

Nothing in the axioms of probability (except Axiom 4) indicates the value of $P(A)$ for a particular event A. The axioms only stipulate that however P is defined on $\mathcal{F}$, it must satisfy Axioms 1–4. The actual experiment under consideration determines the

way in which probabilities should be defined. Various paradoxes in probability theory can arise from different interpretations of the actual experiment, all of which can be reasonable (see [5]). Such paradoxes do not indicate an inconsistency in the axioms of probability.

Moreover, it is interesting to note that there is absolutely nothing in the probability axioms 1–4 which refers to the notion of a limiting relative frequency when an experiment is indefinitely repeated. Shafer [3] advocates actually incorporating in some manner the notions of repetition and long-run relative frequency directly into the axiomatic framework of probability in order to emphasize the unity of the field. The notion of limiting relative frequency is a nontrivial *consequence* of Axioms 1–4 and is made precise by the Borel *strong law of large numbers* (SLLN). The Borel SLLN asserts that if $\{A_n, n \geq 1\}$ is a sequence of independent events all with the same probability p, then

$$P\left(\lim_{n\to\infty} \hat{p}_n = p\right) = 1, \tag{1}$$

where $\hat{p}_n = \sum_{j=1}^{n} I(A_j)/n$ is the proportion of $\{A_1, \ldots, A_n\}$ to occur, $n \geq 1$. [Here, $I(A_j)$ denotes the indicator function of the event A_j.] Hence, with probability 1, the "sample proportion", $\hat{p}_n$, approaches the "population proportion", p, as the "sample size" $n \to \infty$. It is this version of the SLLN which thus provides the theoretical justification for the long-run relative frequency approach to probability theory and so the SLLN lies at the very foundation of statistical science. It should be noted, however, that the convergence in (1) is not pointwise on Ω but, rather, is pointwise on some subset of Ω *having probability 1*. Thus, any practical interpretation of p via (1) would require that one has a priori an intuitive understanding of the notion of an event having probability 1.

The following example of Stout [4, p. 9] illustrates an application of (1) to the field of biostatistics. Consider a new drug for which the proportion p of patients who will be cured by the drug is unknown. A medical researcher continuously estimates p by using the proportion $\hat{p}_n$ of the first n patients treated with the drug who get cured. The medical researcher is interested in knowing if there will ever be a point in the sequence of patients such that, with high probability, $\hat{p}_n$ will be within ε of p *and stay* within ε of p (where $\varepsilon > 0$ is a prescribed tolerance). The answer is affirmative since (1) is equivalent to the assertion that, for given $\varepsilon > 0$ and $\delta > 0$, there exists a positive integer $N = N_{\varepsilon,\delta}$ such that

$$P\left(\bigcap_{n=N}^{\infty} [|\hat{p}_n - p| \leq \varepsilon]\right) \geq 1 - \delta.$$

That is to say, the probability is arbitrarily close to 1 that $\hat{p}_n$ will be arbitrarily close to p simultaneously for all n beyond some point. Consequently, the SLLN (1) is not only of theoretical significance, but also is of practical significance.

References

[1] Kolmogorov, A.N. (1933). *Foundations of the Theory of Probability*. Springer-Verlag, Berlin (in German). English translation: Chelsea, New York, 1950; 2nd English Ed. 1956.

[2] Rao, M.M. (1984). *Probability Theory with Applications*. Academic Press, Orlando.

[3] Shafer, G. (1990). The unity and diversity of probability, *Statistical Science* **5**, 435–462.

[4] Stout, W.F. (1974). *Almost Sure Convergence*. Academic Press, New York.

[5] Székely, G.J. (1986). *Paradoxes in Probability Theory and Mathematical Statistics*. Akadémiai Kiadó, Budapest.

(*See also* **Law of Large Numbers**)

ANDREW ROSALSKY & RICHARD SCHEAFFER

Back-calculation

Back-calculation – also called back-projection – estimates past infection rates of an epidemic infectious disease by working backward from observed disease **incidence** using knowledge of the **incubation period** between infection and disease. Although potentially applicable to various diseases, it was first proposed [15, 16] to study the acquired immune deficiency syndrome (**AIDS**) epidemic and has mainly been applied in this area. Performance of back-calculation requires a technical framework for defining and maximizing a likelihood to obtain estimated infection rates, detailed information about key inputs such as incubation and reporting completeness assumptions, and a strategy for assessing uncertainty in the key inputs and resulting uncertainty in back-calculated estimates.

The Basic Method

Suppose that we have n nonoverlapping intervals, (T_{j-1}, T_j), $j = 1, \ldots, n$; let Y_j be the number of persons developing disease in the jth interval, and assume that no infections occurred before time T_0. In practice, these intervals will often be calendar months or quarters. For example, AIDS incidence in the US is reported as the number of new diagnoses each month. Thus, a discrete-time formulation is realistic for practical applications, and we will employ such notation here, assuming a monthly time scale. Back-calculation is based on the following convolution equation:

$$\mathrm{E}(Y_j) = \sum_{i=1}^{j} \theta_i, D_{ij}, \quad j = 1, \ldots, n, \tag{1}$$

where θ_i is the expected number of new infections in month i and D_{ij} is the probability of developing disease in month j given infection in month i; that is, the probability that the incubation time is equal to $j - i$ given infection in month i. Back-calculation is thus a deconvolution method. Given a set of observed values $\mathbf{y} = (y_1, \ldots, y_n)$, it uses known D_{ij} to find a θ likely to have produced $\mathbf{y}$ via (1).

Eq. (1) only specifies the first moment of the Y_j, so implementation of the strategy requires additional specifics. A simple approach is to assume that the Y_j are independent with **Poisson** error structure. This follows from an assumption that infections arise according to a nonhomogeneous **Poisson process**. In our discrete-time framework, this means that the number of infections in month i is Poisson with expectation θ_i and the numbers of infections in different months are independent. This assumption produces a log **likelihood** (up to a constant) of

$$l(\mathbf{y}|\boldsymbol{\theta}) = \sum_{j=1}^{n} \left[y_j \log \left(\sum_{i=1}^{j} \theta_i D_{ij} \right) - \sum_{i=1}^{j} \theta_i D_{ij} \right], \tag{2}$$

which can be maximized to obtain an estimate of $\boldsymbol{\theta}$. To avoid ill-posedness [48], some structure must be imposed on $\boldsymbol{\theta}$. For example, a parametric model might specify that $\theta_i = f_{\boldsymbol{\beta}}(i)$, where $\boldsymbol{\beta}$ is a (small) vector of parameters and f is a family of functions indexed by $\boldsymbol{\beta}$. Projected values of Y_k for $k > n$ (yet to be observed) can be obtained from (1) using estimates $\hat{\theta}_j$, with $\hat{\theta}_j$ obtained by extrapolation for $n < j \leq k$.

Finding parameters that maximize the likelihood (2) may be possible using general numerical approaches such as the Newton–Raphson method, depending on the complexity of the structure imposed on θ (*see* **Optimization and Nonlinear Equations**). Using an expectation-maximization (**EM**) algorithm, however, can greatly simplify the computations. Consider the complete data $\{x_{ij}\}$, where x_{ij} is the number of persons infected in month i and diagnosed in month j. Under the assumptions leading to (2), these counts are independent Poisson with means $\theta_i D_{ij}$. The complete-data log likelihood is therefore (up to a constant)

$$\sum_{i=1}^{n} \left[x_i \log(\theta_i) - \theta_i \sum_{j=i}^{n} D_{ij} \right], \tag{3}$$

where $x_i = \sum_{j=i}^{n} x_{ij}$. This is a linear function of x_i, so its expected value can be calculated using the formula [30]

$$\mathrm{E}(x_i|\mathbf{y}, \boldsymbol{\theta}) = \sum_{j=i}^{n} y_j \frac{\theta_i D_{ij}}{\sum_{k=1}^{j} \theta_k D_{kj}}. \tag{4}$$

The EM algorithm begins with an initial guess for the parameters that determine $\boldsymbol{\theta}$, calculates the expected value of (3) using (4) (the E-step), and finds the

new values of the parameters that maximize (3) (the M-step). The E- and M-steps are iterated until the parameter estimates converge. The simple forms of (3) and (4) make this approach computationally easy.

Inputs

Back-calculation requires a known incubation distribution (the D_{ij}) and accurate data on incidence of disease. In addition, a realistic model for infection patterns must be specified.

Incubation

The estimate of the infection pattern $\boldsymbol{\theta}$ depends crucially on the assumed incubation distribution [6, 8, 58]. Accurate estimates, however, may be difficult to obtain. Estimation of distributions of incubation times from human immunodeficiency virus (HIV) infection to AIDS diagnosis illustrates many potential problems in obtaining accurate D_{ij}. These include inherent limitations in available data, heterogeneity of distributions in different populations, and changes over time (nonstationarity) (*see* **Stationarity**).

Data Sources. Inherent limitations arise because HIV infection is usually not immediately detected, and observed incubation times are therefore only available for special groups whose times of HIV infection can be determined retrospectively. These include persons whose HIV infection can be traced to a particular blood transfusion and those whose time of infection can be bracketed by antibody testing of stored specimens from various times in the past. These special groups may not be representative of the wider population for which back-calculation is to be performed. In addition, the data from these sources may suffer from right-truncation [38, 39, 42] or double-censoring [8, 27]. Such data require specialized statistical analysis and convey less information than fully observed data. An extreme form of double-censoring is present for **prevalent** cases – those already infected at the time that they were recruited into a cohort study and followed for development of AIDS. Such subjects are known to have been infected at some time between the start of the epidemic and their time of recruitment (an interval-censored starting time), and their time of AIDS may be right-censored. An additional difficulty is that persons who had already developed AIDS may have been excluded from recruitment. Thus, methods for left-truncated data must be used. Prevalent cohort participants are much more numerous than those with infection times that are more narrowly bracketed by a positive HIV antibody test preceded by a negative one, so investigators have attempted to utilize data from prevalent subjects by imputing infection times using laboratory markers that change with length of infection [36, 47] and by other methods [5, 41, 59].

Heterogeneity. Back-calculation is usually applied to entire populations defined by region of residence and possibly by risk behaviors, but data on incubation times come from highly selected, small groups of persons with known infection times. This would not be a problem if all populations and all the groups shared the same incubation distribution, or if all differences could be accurately explained and quantified in terms of readily measured characteristics such as age. This, however, does not appear to be the case. Direct comparison of data from different sources has shown statistically significant differences between different groups [12], including groups of gay men of similar ages who differ on other characteristics [7, 8, 60, 65]. This heterogeneity adds considerable uncertainty about what incubation distribution to use for a particular population, especially understudied populations such as women or intravenous drug users.

Nonstationarity. Estimation would be simplified if the D_{ij} all depended only on the elapsed time, $j - i$; that is, if the incubation distribution were stationary. In general, however, the chance of developing disease may be nonstationary and also depend on the time of infection, i, or on the current time, j. For example, diagnosis of AIDS may depend on several factors that change over time, including availability and effectiveness of preventive treatments, changes in the official case definition of AIDS [21, 22, 24], changes in care-seeking patterns, and possible evolution of the virus toward more or less virulence. Such phenomena place even greater demands on limited data. A widely used approach has been to assume that no factor other than treatment has caused nonstationarity, and to use data on effectiveness of treatment and usage rates of treatment over time to modify stationary incubation estimates [13]. This results in models in which persons infected more recently have

longer incubation times. A more direct approach is to examine the special groups discussed above who have approximately known infection dates, and to use standard **survival analysis** methods to estimate the effect of infection date or calendar time (a **time-dependent covariate**) on development of AIDS. This estimates the net effect of all factors that may be changing over time. Several such studies have found that incubation times remained constant or actually shortened in the late 1980s and early 1990s compared with earlier in the epidemic [12, 29, 35, 62]. This suggests that other factors, such as more aggressive care seeking or evolution of the virus, accelerated diagnosis of AIDS more than it was slowed by beneficial treatment effects. A dramatic source of nonstationarity in the US is the expansion in 1993 of conditions officially qualifying as an AIDS diagnosis [24]. This had such a large and sudden impact that accurate estimates of the affected D_{ij} may not be obtainable. In addition, introduction of potent protease inhibitors and increased use of combination antiretroviral therapy starting in the mid-1990s will also influence incubation times.

Incidence

In practice, true disease incidence is not observed exactly because of imperfections in the surveillance system. Incompleteness of the observed incidence arises from reporting delays and from underreporting; that is, cases who are never reported. This incompleteness must be corrected before back-calculation is applied. In addition, the incidence series may contain short-term perturbations, such as seasonal patterns, that can be adjusted out to make long-term trends clearer and to improve the accuracy of back-calculation.

Reporting Delay. Because there is often a lag between diagnosis of disease and the time that it is recorded and tabulated, recent incidence is incomplete. This typically causes a downturn in recent incidence that would severely distort back-calculation results if left uncorrected. The usual strategy is to estimate for each month j a completeness factor, R_j, the proportion of true incidence that has been reported. If y_j^* denotes observed incidence, one can apply back-calculation to corrected counts $y_j = y_j^*/R_j$. (Alternatively, the R_j can be incorporated directly into the back-calculation procedure [6]). A common practice is to exclude counts that are so recent that they are estimated to be less than 50% complete. If surveillance provides both date of disease and date of report, incompleteness due to reporting delay can be estimated if one assumes a maximum possible length of delay [25]. This requires specialized methods [14, 39, 50] because of the severe right-truncation caused by the fact that the only cases available for analysis are those with short enough delays to have been already reported. In addition, dependence of delays on case characteristics and changes in delay patterns over time can be estimated. Estimates of changes over time can be strongly influenced by irrelevant shifts in the patterns of very short delays, so modifications to avoid this lead to better estimates [1]. In the US, the 1993 change in the AIDS case definition apparently had a strong impact on reporting, even of cases meeting the earlier definition [4].

Underreporting. In addition to delays in reporting, there also may be cases that are never reported. Such underreporting of AIDS cases has been investigated to a limited extent by cross matching reported AIDS cases to cases identified by other means, notably **death certificates** that list HIV under cause of death [19, 31]. The proportion of cases found by other means that are not also in the **surveillance** system provides an estimate of the underreporting rate. (More sophisticated **capture–recapture** methods have not been widely used.) Studies of underreporting require use of personal identifiers, and so are usually carried out at a local level. They also usually apply to a specific time period. Consequently, extensive systematic data on underreporting is typically not available for the population under consideration, and assuming constant underreporting of between 10% and 20% is a common practice. Such assumptions must be combined with estimated incompleteness due to reporting delay to obtain R_j that reflect both sources of incompleteness.

Short-term Patterns. Season can influence the incidence of some infectious diseases, notably AIDS, and incidence of AIDS in a particular month is also influenced by how many workdays it includes [2]. The lengths of calendar months also vary by 10%. These short-term influences increase month-to-month variability and can degrade back-calculation results. Performance can be improved by estimating these

effects and adjusting them out of the incidence series to be used [2, 3].

Infection Model

As noted above, some structure must be imposed on $\boldsymbol{\theta}$ to allow stable estimation. Both parametric and nonparametric approaches have been used. Parametric approaches include smooth families indexed by two or three parameters [26, 58, 63], as well as step functions with four or five steps, within which infection rates are assumed to be constant [16, 55]. Although these step models are not plausible, they are flexible, and have been made more so by modifications to allow adaptive selection of cutpoints between steps [56]. Nonparametric approaches do not directly parameterize $\boldsymbol{\theta}$ but obtain smooth estimates by either adding a smoothing step after the M-step of the EM algorithm [11], or by using **penalized maximum likelihood** [6] or **ridge regression** [49]. Some methods combine aspects of the parametric and nonparametric approaches [13, 32].

Sensitivity Analyses

Asymptotic standard errors for estimated infection rates and future incidence projections can be obtained from the observed **information matrix** for $\boldsymbol{\theta}$ or by **bootstrap** methods, but this captures only a small part of the real uncertainty. Possible errors in the inputs to back-calculation cause much greater uncertainty in the results. **Sensitivity analyses** that employ a wide variety of inputs (consistent with available data on incubation and incidence) can serve to more realistically illustrate the plausible range of possibilities. (**Bayesian methods** that incorporate **priors** for the various inputs could offer a more formal assessment of uncertainty [20, 58], but these have not been widely used.) If the range of plausible inputs is large, as in the case of HIV and AIDS, exhaustive exploration of possible uncertainty may be difficult. Sensitivity analyses for back-calculation from AIDS incidence have generally considered from two to five possible incubation distributions [6, 54, 63], and often not considered uncertainty in the incidence series. An additional difficulty when uncertainties in the inputs are wide is that back-calculated results from some inputs may contradict what is known (at least qualitatively) from other sources, such as cross-sectional prevalence surveys or cohort studies. Simply dropping the offending inputs from the sensitivity analyses, however, is not adequate, because the remaining set of possibilities will be too narrow, even if the original set was adequate. This is because there is a continuum of plausible possibilities between the eliminated and retained possibilities, some of which are consistent with the outside information. When inputs that are a priori plausible produce implausible results, formal methods to combine back-calculation with the outside data [17] should be considered, as should the possibility that back-calculation cannot meaningfully improve on what is known directly from the outside data.

Incubation

The assumed incubation distribution strongly influences estimated infection rates. This is apparent from the forms of (1) and (2), where the incubation terms D_{ij} and the θ_i always appear multiplied together. Different plausible AIDS incubation distributions can lead to estimates of cumulative HIV infections in the US that differ by factors of two or more, while providing nearly identical fits to the observed AIDS incidence data [6, 8]. This implies that errors in the D_{ij} will not be detectable in the back-calculation process itself, because their influence will be masked by compensating errors in the estimate of $\boldsymbol{\theta}$ and no lack of fit will be apparent. This and the sources of uncertainty noted above underscore the need for careful sensitivity analysis of incubation assumptions. Nonstationarity in the incubation distribution can also influence back-calculated estimates. For example, a slowdown in incidence will be attributed to an earlier decline in infections if a stationary incubation is assumed, but could also be explained by recent lengthening of incubation times.

Disease Incidence

Assumptions about reporting delay and underreporting can strongly influence projections. Differing reasonable assumptions about underreporting and late reporting of AIDS cases diagnosed through 1991 in the US resulted in two-year projections that differed by 20% or more [1], and additional uncertainty about very late reporting increases this difference to at least 30% [3, 25]. In addition, one can see from (1) that underreporting has a direct impact on the estimate

of $\boldsymbol{\theta}$. For example, assuming constant 80% reporting instead of 90% reporting would increase all of the imputed y_j by about 13% (0.9/0.8), resulting in a corresponding proportional increase in all of the estimated θ_i.

Refinements

A wide variety of technical refinements, extensions, and modifications of back-calculation have been studied. Notable among these are methods for: incorporating results of HIV **prevalence** surveys [17]; allowing for dependence in the HIV infection process [9]; estimating **overdispersion** in a **quasi-likelihood** approach [17, 43]; utilizing data on HIV tests [45, 52]; using age at time of AIDS to back-calculate age-specific HIV incidence [10, 53]; nonparametric modeling of infection rates, including data-driven choices of smoothness parameters [6, 32, 44]; incorporating knowledge of the size of the susceptible population [61]. Because of the considerable uncertainty about crucial inputs, however, these refinements may not be able meaningfully to improve the accuracy of back-calculation.

Limitations

A key limitation of back-calculation is the need for accurate inputs, as noted above. Because uncertainty in the results comes mainly from uncertainty about these inputs, estimates of pure statistical uncertainty are misleading. Two additional limitations of the method are that it provides little information about recent infection rates and that projections can be overly sensitive to recent incidence.

Back-calculation is primarily useful with epidemic infectious diseases for which there is a substantial lag between infection and disease. If a disease is in a steady state or if disease rapidly follows infection, then infection rates can be adequately ascertained directly from disease incidence. Because of this focus, there will be little direct information about recent infection rates and back-calculated estimates of θ_j for j close to n will be determined mainly by implicit extrapolation, from either the parametric model of $\boldsymbol{\theta}$ or the form of the smoothness assumption. For example, because few persons develop AIDS within two years following HIV infection, back-calculation from AIDS incidence provides little information about infection rates in the last two years.

Projections from back-calculation can be overly sensitive to counts near the end of the incidence series. This is particularly true for AIDS if seasonal patterns are not adjusted out of the incidence series [3]. For example, anomalously high AIDS incidence in the US in the first half of 1987 caused projections from back-calculations to be too high, which was interpreted as evidence for a treatment-induced downturn in incidence [28]. The projections would have been more accurate if deseasonalized incidence had been used, and would have been much better if a more robust projection method had been used. Two- or three-year projections based on incidence through the end of 1986 also would have been fairly accurate [3].

Alternatives

A simple alternative for projecting future incidence is empirical **extrapolation** [34, 40, 46, 64]. This can be reasonably accurate [3], but provides no information about infection rates and has no ability to anticipate changes in trajectory. Measurement of infections in cross-sectional surveys and cohorts followed over time provides direct information on prevalence and incidence of infections. Such studies are most useful when performed anonymously on specimens collected for other purposes [33, 51], because this can eliminate the potentially serious problem of **nonresponse bias** [23]. In **cohort studies** of incidence, serious dropout bias can result from the fact that higher-risk subjects may be more likely to fail to return for follow-up testing (*see* **Nonignorable Dropout in Longitudinal Studies**). Markers of recent infection can be used to estimate current incidence without relying on follow-up and to correct dropout bias [18], provided that the initial sample is representative and that the average duration of the marker is known. One can deduce the shape of the infection density from the mix of laboratory markers, such as CD4 counts, in one or more cross-sectional surveys of infected individuals [57]. This requires a representative sample of infected persons and detailed knowledge of how the marker evolves over time since infection, which may be more difficult to obtain than the incubation information required by back-calculation [37]. Mathematical epidemic modeling is used mainly to further qualitative understanding, and typically requires too detailed input to provide useful

quantitative results (*see* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**).

References

[1] Bacchetti, P. (1994). The impact of lengthening AIDS reporting delays and uncertainty about underreporting on incidence trends and projections, *Journal of Acquired Immune Deficiency Syndromes* **7**, 860–865.

[2] Bacchetti, P. (1994). Seasonal and other short-term influences on United States AIDS incidence, *Statistics in Medicine* **13**, 1921–1931.

[3] Bacchetti, P. (1995). Historical assessment of some specific methods for projecting the AIDS epidemic, *American Journal of Epidemiology* **141**, 776–781.

[4] Bacchetti P. (1996). Reporting delay of deaths with AIDS in the United States, *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **13**, 363–367.

[5] Bacchetti, P. & Jewell, N.P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times, *Biometrics* **47**, 947–960.

[6] Bacchetti, P., Segal, M.R. & Jewell, N.P. (1993). Back-calculation of HIV infection rates, *Statistical Science* **8**, 82–119.

[7] Bacchetti, P., Koblin, B.A., van Griensven, G.J.P. & Hessol, N.A. (1996). Determinants of HIV disease progression among homosexual men, *American Journal of Epidemiology* **143**, 526.

[8] Bacchetti, P., Segal, M.R., Hessol, N.A. & Jewell, N.P. (1993). Differing AIDS incubation periods and their impacts on reconstructing HIV epidemics and projecting AIDS incidence, *Proceedings of the National Academy of Sciences* **90**, 2194–2196.

[9] Becker, N.G. & Chao, X. (1994). Dependent HIV incidences in back-projection of AIDS incidence data, *Statistics in Medicine* **13**, 1945–1958.

[10] Becker, N.G. & Marschner, I.C. (1993). A method for estimating the age-specific relative risk of HIV infection for AIDS incidence data, *Biometrika* **80**, 165–178.

[11] Becker, N.G., Watson, L.F. & Carlin, J.B. (1991). A method of nonparametric back-projection and its application to AIDS data, *Statistics in Medicine* **10**, 1527–1542.

[12] Biggar, J. (1990). AIDS incubation in 1891 HIV seroconverters from different exposure groups, *AIDS* **4**, 1059–1066.

[13] Brookmeyer, R. (1991). Reconstruction and future trends of the AIDS epidemic in the United States, *Science* **253**, 37–42.

[14] Brookmeyer, R. & Damiano, A. (1989). Statistical methods for short-term projections of AIDS incidence, *Statistics in Medicine* **8**, 23–34.

[15] Brookmeyer, R. & Gail, M.H. (1986). Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States, *Lancet* **2**, 1320–1322.

[16] Brookmeyer, R. & Gail, M.H. (1988). A method for obtaining short term predictions and lower bounds on the size of the AIDS epidemic, *Journal of the American Statistical Association* **83**, 301–308.

[17] Brookmeyer, R. & Liao, J. (1990). Statistical modelling of the AIDS epidemic for forecasting health care needs, *Biometrics* **46**, 1151–1163.

[18] Brookmeyer, R., Quinn, T., Shepherd, M., Mehendale, S., Rodrigues, J. & Bollinger, R. (1995). The AIDS epidemic in India: a new method for estimating current human immunodeficiency virus (HIV) incidence rates, *American Journal of Epidemiology* **142**, 709–713.

[19] Buehler, J.W., Berkelman, R.L. & Stehr-Green, J.K. (1992). The completeness of AIDS surveillance, *Journal of Acquired Immune Deficiency Syndromes* **5**, 257–264.

[20] Carlin, J.B. & Gelman, A. (1993). Comment: assessing uncertainty in backprojection, *Statistical Science* **8**, 104–106.

[21] Centers for Disease Control (1985). Revision of the case definition of acquired immune deficiency syndrome for national reporting-United States, *Morbidity and Mortality Weekly Reports* **34**, 373–375.

[22] Centers for Disease Control (1987). Revision of the CDC surveillance case definition for acquired immunodeficiency syndrome, *Morbidity and Mortality Weekly Reports* **36**, 3S–15S.

[23] Centers for Disease Control (1991). Pilot study of a household survey to determine HIV seroprevalence, *Morbidity and Mortality Weekly Reports* **40**, 1–5.

[24] Centers for Disease Control and Prevention (1992). 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults, *Morbidity and Mortality Weekly Reports* **41**, (No. RR-17), 1–18.

[25] Cooley, P.C., Hamill, D.N., Meyers, L.E. & Liner, E.C. (1993). The assumption of no long reporting delays may result in underestimates of US AIDS incidence, *AIDS* **7**, 1379–1381.

[26] Day, N.E., Gore, S.M., McGee, M.A. & South, M. (1989). Predictions of the AIDS epidemic in the UK: the use of the back projection method, *Philosophical Transactions of the Royal Society of London, Series B* **325**, 123–134.

[27] DeGruttola, V. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics* **45**, 1–11.

[28] Gail, M.H., Rosenberg, P.S. & Goedert, J.J. (1990). Therapy may explain recent deficits in AIDS incidence, *Journal of Acquired Immune Deficiency Syndromes* **3**, 296–306.

[29] Gauvreau, K., DeGruttola, V. & Pagano, M. (1994). The effect of covariates on the induction time of AIDS using improved imputation of exact seroconversion times, *Statistics in Medicine* **13**, 2021–2030.

[30] Green, P.J. (1990). On use of the EM algorithm for penalized likelihood estimation, *Journal of the Royal Statistical Society, Series B* **52**, 443–452.

[31] Greenberg, A.E., Hindin, R., Nicholas, A.G., Bryan, E.L. & Thomas, P.A. (1993). The completeness of AIDS case reporting in New York City, *Journal of the American Medical Association* **269**, 2995–3001.

[32] Greenland, S. (1996). Historical HIV incidence modelling in regional subgroups: use of flexible discrete models with penalized splines based on prior curves, *Statistics in Medicine* **15**, 513–525.

[33] Gwinn, M., Pappaioanou, M., George, J.R., Hannon, W.H., Wasser, S.C., Redus, M.A., Hoff, R., Grady, G.F., Willoughby, A., Novello, A.C., Petersen, L.R., Dondero, T.J., Jr & Curran, J.W. (1991). Prevalence of HIV infection in childbearing women in the United States. Surveillance using newborn blood samples, *Journal of the American Medical Association* **265**, 1704–1708.

[34] Healy, M.J.R. & Tillet, H.E. (1988). Short-term extrapolation of the AIDS epidemic, *Journal of the Royal Statistical Society, Series A* **151**, 50–65.

[35] Hessol, N.A., Koblin, B.A., van Griensven, G.J.P., Bacchetti, P., Liu, J.Y., Stevens, C.E., Coutinho, R.A., Buchbinder, S.P. & Katz, M.H. (1994). Progression of human immunodeficiency virus type 1 (HIV-1) infection among homosexual men in hepatitis B vaccine trial cohorts in Amsterdam, New York City, and San Francisco 1978–1991, *American Journal of Epidemiology* **139**, 1077–1087.

[36] Hoover, D.R., Taylor, J.M.G., Kingsley, L., Chmiel, J.S., Munoz, A., He, Y. & Saah, A. (1994). The effectiveness of interventions on incubation of AIDS as measured by secular increases within a population, *Statistics in Medicine* **13**, 2127–2139.

[37] Jewell, N.P. & Kalbfleisch, J.D. (1992). Marker models in survival analysis and applications to issues associated with AIDS, in *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz & V.T. Farewell, eds. Birkhauser, Boston.

[38] Kalbfleisch, J.D. & Lawless, J.F. (1988). Estimating the incubation period for AIDS patients, *Nature* **333**, 504–505.

[39] Kalbfleisch, J.D. & Lawless, J.F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS, *Journal of the American Statistical Association* **84**, 360–372.

[40] Karon, J.M., Devine, O.J. & Morgan, W.M. (1989). Predicting AIDS incidence by extrapolating from recent trends, in *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez, ed. Springer-Verlag, New York, pp. 58–88.

[41] Kuo, J.-M., Taylor, J.M.G. & Detels, R. (1991). Estimating the AIDS incubation period from a prevalent cohort, *American Journal of Epidemiology* **133**, 1050–1057.

[42] Lagakos, S.W., Barraj, L.M. & DeGruttola, V. (1988). Nonparametric analysis of truncated survival data with applications to AIDS, *Biometrika* **75**, 515–523.

[43] Lawless, J.F. & Sun, J. (1992). A comprehensive back-calculation framework for the estimation and prediction of AIDS cases, in *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz & V.T. Farewell, eds. Birkhauser, Boston.

[44] Liao, J. & Brookmeyer, R. (1995). An empirical Bayes approach to smoothing in backcalculation of HIV infection rates, *Biometrics* **51**, 579–588.

[45] Marschner, I.C. (1994). Using time of first positive HIV test and other auxiliary data in back-projection of AIDS incidence, *Statistics in Medicine* **13**, 1959–1974.

[46] Morgan, W.M. & Curran, J.W. (1986). Acquired immunodeficiency syndrome: current and future trends, *Public Health Reports* **101**, 459–465.

[47] Munoz, A., Wang, M.-C., Bass, S., Taylor, J.M.G., Kingsley, L.A., Chmiel, J.S. & Polk, B.F. (1989). Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men, *American Journal of Epidemiology* **130**, 530–539.

[48] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems, *Statistical Science* **1**, 502–527.

[49] Pagano, M., DeGruttola, V., MaWhinney, S. & Tu, X.M. (1992). The HIV epidemic in New York City: statistical methods for projecting AIDS incidence and prevalence, in *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz & V.T. Farewell, eds. Birkhauser, Boston.

[50] Pagano, M., Tu, X.M., DeGruttola, V. & MaWhinney, S. (1994). Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data, *Biometrics* **50**, 1203–1214.

[51] Pappaioanou, M., Dondero, T.J., Peterson, L.R., Onorato, I.M., Sanchez, C.D. & Curran, J.W. (1990). The family of HIV seroprevalence surveys: objectives, methods, and uses of sentinel surveillance for HIV in the United States, *Public Health Reports* **105**, 113–119.

[52] Raab, G.M., Fielding, K.L. & Allardice, G. (1994). Incorporating HIV test data into forecasts of the AIDS epidemic in Scotland, *Statistics in Medicine* **13**, 2009–2020.

[53] Rosenberg, P.S. (1994). Backcalculation models of age-specific HIV incidence rates, *Statistics in Medicine* **13**, 1975–1990.

[54] Rosenberg, P.S. (1995). Scope of the AIDS epidemic in the United States, *Science* **270**, 1372–1375.

[55] Rosenberg, P.S. & Gail, M.H. (1990). Uncertainty in estimates of HIV prevalence derived by backcalculation, *Annals of Epidemiology* **1**, 105–115.

[56] Rosenberg, P.S. & Gail, M.H. (1991). Backcalculation of flexible linear models of the human immunodeficiency virus infection curve, *Applied Statistics* **40**, 269–282.

[57] Satten, G.A. & Longini, I.M. (1994). Estimation of incidence of HIV infection using cross-sectional marker surveys, *Biometrics* **50**, 675–688.

[58] Taylor, J.M.G. (1989). Models for the HIV infection and AIDS epidemic in the United States, *Statistics in Medicine* **8**, 45–58.

[59] Taylor, J.M.G., Kuo, J.-M. & Detels, R. (1991). Is the incubation period of AIDS lengthening?, *Journal of Acquired Immune Deficiency Syndromes* **4**, 69–75.

[60] van Griensven, G.J.P., Veugelers, P.J., Page-Shafer, K.A., Kaldor, J.M. & Schechter, M.T. (1996). Determinants of HIV disease progression among homosexual men, *American Journal of Epidemiology* **143**, 525.

[61] Verdecchia, A. & Mariotto, A.B. (1995). A back-calculation method to estimate the age and period HIV infection intensity, considering the susceptible population, *Statistics in Medicine* **14**, 1513–1530.

[62] Veugelers, P.J., Page, K.A., Tindall, B., Schechter, M.T., Moss, A.R., Winkelstein, W.W., Jr., Cooper, D.A., Craib, K.J.P., Charlebois, E., Coutinho, R.A. & van Griensven, G.J.P. (1994). Determinants of HIV disease progression among homosexual men registered in the Tricontinental Seroconverter Study, *American Journal of Epidemiology* **140**, 747–758.

[63] Wilson, S.R., Fazekas de St. Groth, C. & Solomon, P.J. (1992). Sensitivity analyses for the backcalculation method of AIDS projections, *Journal of Acquired Immune Deficiency Syndromes* **5**, 523–527.

[64] Zeger, S.L., See, L.-C. & Diggle, P.J. (1989). Statistical methods for monitoring the AIDS epidemic, *Statistics in Medicine* **8**, 3–21.

[65] Zwahlen, M., Vlahov, D. & Hoover, D.R. (1996). Determinants of HIV disease progression among homosexual men, *American Journal of Epidemiology* **143**, 523–525.

Bibliography

In addition to the above references, see the following for readable discussions of many topics related to back-calculation as applied to the AIDS epidemic:

Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, New York.

PETER BACCHETTI

Backward and Forward Shift Operators

The backward shift operator B is defined for time series X_t by

$$B^k X_t = X_{t-k}.$$

Similarly, the forward shift operator F is defined by

$$F^k X_t = X_{t+k}.$$

These operators provide a convenient notation for defining times-series models. For example, the first-order autoregressive process can be written as

$$X_t = \alpha B X_t + Z_t.$$

Treating B as if it were a constant, models may be manipulated using simple algebra. For example, the first-order autoregressive model can be rewritten as

$$X_t = \frac{Z_t}{1 - \alpha B} = (1 + \alpha B + \alpha^2 B^2 + \cdots) Z_t.$$

The general ARMA model of order p, q is given by

$$\Phi(B) X_t = \Theta(B) Z_t,$$

where

$$\Phi(B) = 1 - \alpha_1 B - \cdots - \alpha_p B^p$$

and

$$\Theta(B) = 1 + \beta_1 B + \cdots + \beta_q B^q.$$

Another useful aspect of the backward shift operator here is that it allows the conditions for **stationarity** and invertibility to be stated simply. The ARMA process, for example, is stationary if the roots of $\Phi(B) = 0$ lie outside the unit circle and it is invertible if the roots of $\Theta(B) = 0$ lie outside the unit circle.

(*See also* **ARMA and ARIMA Models**)

SOPHIA RABE-HESKITH

Bacterial Growth, Division, and Mutation

> The problems of growth and division of cells lie in the very heart of biology [9].

Bacteria normally reproduce asexually, by binary fission. When organisms are introduced to a new environment, for instance by inoculation into a new growth medium, they undergo a *lag phase*, during which no growth takes place. When they have adapted to the new medium, they enter a so-called *logarithmic* phase, during which they reproduce with an approximately constant mean generation time, which may be as short as 20 minutes, leading to approximately exponential growth in the population size. In due course, the population reaches saturation level, with the depletion of nutrients and the accumulation of waste products, and there is an increasing proportion of nondividing and dying cells.

Models of cell growth and division in bacteria were considered more than 60 years ago (see [10] and [12]; reviewed by Harvey [7], Cooper [5], and Koch [16]). The evolution of models of bacterial growth was closely followed by development of models of the cell cycle in budding and fission yeast, cell cultures, and mammalian cells (*see* **Cell Cycle Models**). Many important biological hypotheses concerning the organization of the cell cycle and its effect on cell population dynamics were first tested in bacterial models, and were modified later for description of the eukaryotic cell-cycle. In some of these theoretical models new mechanisms have been hypothesized, which later have been confirmed in experiments on spontaneous and induced mutations, signal transduction pathways, cell cycle regulation, and programmed cell death. Thus, for many years, modeling of the bacterial cell cycle has been a "proving ground" for refinement of theoretical models in cell biology.

The major questions addressed in bacterial growth models have concerned:

1. stochastic and deterministic models of cell population dynamics (see reviews in [6] and [18]);
2. different models of the cell cycle in individual cells, such as growth control (review in [16]), random transition (review in [31]), and mitotic clocks [45];
3. generational dependence models such as continuum [5], supramitotic control [5, 34], multiple transitions [4, 31], and clonal inheritance [6];
4. cell-cycle regulation models [43, 44];
5. unequal cell-division models [13];
6. spontaneous mutation models and fluctuation tests (reviews in [24] and [25]).

Some of these models were derived by generalization of earlier bacterial models, though the majority of the models were designed for description of experiments with cell lineages and colony-forming assays and did not discriminate between bacterial and eukaryotic cells. Some of the bacterial models were later applied to the analysis of cancer cells, on the assumption of uncontrolled division of cancer cells which was widely accepted at the time (*see* **Tumor Growth**).

From the biological viewpoint such a generalization of bacterial models might not be justified, since there are significant evolutionary differences in organization of the cell cycle in prokaryotes and eukaryotes. However, as Nurse noted [27], many experiments suggest that cell-cycle control may be qualitatively similar in microbial cells and eukaryotic cells, and that a quantitative difference is due to a difference in the rates of cell progression through the deterministic and stochastic stages of the cell cycle.

Over five decades, bacterial models have evolved in diverse directions. We briefly review major directions of modeling which have provided important progress in the understanding of biological and mathematical aspects of cell growth, division and mutations.

Cell-Cycle Control Models

Several classes of empirical models of cycle control in individual cells have provided different answers to the problem of estimation of the generational time distributions (τ-distributions) in populations of dividing cells.

Early observations on cultured bacterial cell pedigrees [10] demonstrated significant variability of many observable parameters such as interdivisional times of individual cells, and growth rates of individual cells and clones. However, many models of exponentially growing bacterial colonies approximated this process in deterministic fashion. These models postulated that for large numbers of cells growing

with stable constant supply of nutrients, the variability of growth rates is not significant.

The *growth control model* (also referred to as a *size control model*) postulates that bacterial cells divide after a permissible size is reached [17]. Thus, the rate of cell division in this model is determined by the rate of cell growth. Variation of interdivisional times was assumed to be due to normally distributed fluctuations of duration of actual division. Direct implications of such postulates were that the size of bacteria at division should be constant, mother–daughter **correlation** of size and generation time should be equal to -0.5, and the correlation of sizes and generation times between sisters should be positive.

Deterministic models of growth control failed to reproduce many observations on exponentially growing cell cultures with constant concentration of substrate: namely, high variation of growth rates and colony sizes, and positive mother–daughter correlations. Later modifications of growth control models incorporated stochastic components, assuming variation of cell sizes as a result of unequal division. Such models were in better agreement with the observations [16].

Transition probability models were developed in the 1970s and 1980s as a result of pulse-labeling experiments in cell cultures [3, 4, 36]. Later modifications of the transition probability model included multiple random transitions and were also applied to bacterial and other cell cultures (see reviews by Rigney [31] and Staudte et al. [37]).

The transition probability model, originally proposed by Smith & Martin [36], postulates that the cell cycle is composed of two stages: A state and B phase. In the A state a cell does not progress to cell division, and it can transit to the B phase with constant probability. In the B phase a cell requires a constant time to complete division. The rate of division is determined by the random time period a cell spent in the A state and the constant time period in the B phase. The experimental support of this model came from findings of an exponential component in the distribution of cell-cycle times for many types of cells. The model accommodated a genetically predetermined constant duration of cell cycle, and variation of cell-cycle time as an effect of changes in the environment.

The mitotic oscillator model (also referred to as *internal clock* or *spontaneous oscillator*) was introduced in 1990 (reviews in [28] and [42]). It reflects new experimental data on the interaction of newly discovered protein complexes, MPF and cyclin [8]. Interactions between these complexes and associated protein kinase cdc2 drive cell-cycle progression by periodic changes of their enzymatic activity, operating as a spontaneous minimum oscillator [45]. The mitotic oscillator model incorporated these new data and reproduced oscillations which effectively defined duration of the cell cycle. This model demonstrated that experimentally identifiable molecular mechanisms can provide control of division in a deterministic fashion. This result was in support of the classical growth control model.

During recent years, more members of the cyclin protein family have been identified in yeast and mammalian cells [29], thus providing solid biological evidence in support of Nurse's idea of quantitative but not qualitative differences in cell-cycle organization of bacteria and eukaryotes [27]. These newly discovered cyclins are activated only during specific cell-cycle phases, and are deactivated at (or near) restriction points previously identified in transition probability models. This new biological evidence provides a solid ground for reconciliation of two major views, deterministic and stochastic, of the cell-division cycle in bacteria and mammalian cells [23, 26].

Spontaneous Mutation Models and Fluctuation Tests

Bacteria experience mutations of various types, which occur apparently randomly, with constant probability per cell division. For instance, a population growing from a single organism sensitive to a drug may, on reaching a population size of, say, 10^8, contain a number of resistant organisms, having descended from one or more cells that mutated during growth. In some instances, back-mutation may occur, from resistant to sensitive type. Mutation is, thus, a more random phenomenon than growth. Deterministic theories of mutating populations may be useful as broad descriptions of population dynamics, especially when population sizes are large. Many situations, though, are dominated by random variation, and stochastic theories are needed.

Deterministic Population Dynamics

Consider a mixed population of two forms of intermutating organisms, with x members of the wild-type X, and y members of the mutant type Y. Suppose the two strains have exponential growth rates a and b, respectively, and that the mutation rates (expressed as the proportion of new organisms which differ in type from the parent) are λ from X to Y, and ν from Y to X. The differential equations are easily solved. If the two growth rates are equal, the proportion of mutants in the population will approach asymptotically the value $\lambda/(\lambda + \nu)$.

Experiments to study the long-term development of a mixed population are hampered by the restrictions on the duration of the logarithmic phase of growth. These restrictions are overcome by experiments in continuous cultures, in which a liquid medium is continuously changed so as to provide the conditions for constant growth [41]. Long-term experiments, interpreted by the deterministic theory, provide rough estimates of the growth rate (and a check on the near-equality of growth rates for the two types), and of the two mutation rates [2, 35].

Stochastic Models

As noted earlier, the case for a stochastic treatment arises mainly because of the effects of random mutation, rather than to improve the description of bacterial growth. These effects are especially important in short-term experiments, such as *fluctuation tests* in which replicate cultures are grown from small inocula. In such an experiment, with replicate cultures grown to the same final population size, n, the proportion of mutant organisms is likely to be relatively small, and back-mutation can be ignored. The expected number of *mutations* (as distinct from *mutant organisms*) will be approximately $\mu = \lambda n$, and the numbers in different cultures should follow a **Poisson distribution** with mean μ. These numbers are not directly observable, because of the reproduction of mutant progeny, but what can be observed is the proportion of cultures with no mutants (and, hence, no mutations). This should be the zero term of the Poisson distribution, $e^{-\mu}$. This provides a simple way of estimating the mutation rate λ, since the total population size, n, can be estimated by standard methods.

Further information is provided by the distribution, between replicate cultures, of the number of mutant organisms per culture. This will clearly be extremely variable and highly skewed. A culture in which, by a rare chance, an early mutation occurred, will accumulate a large number of progeny descended from the first mutant. Most cultures will experience their first mutation much later, and the mutant progeny will be relatively few. Some features of the distribution were described by Luria & Delbrück [20], who conducted experiments on the resistance of *Escherichia coli*, strain B, to bacteriophage. They found highly skew distributions, of the type expected on mutation theory, and considered that their findings strongly supported the view that resistance was caused by mutation, rather than by some form of Lamarckian-like adaptation.

Luria & Delbrück assumed deterministic growth, at the same rate a for both strains, and obtained asymptotic expressions for the mean and variance of the number of mutants, Y, at time t, in replicates with initial size x_0:

$$\mathrm{E}(Y) = x_0 \lambda\, at \exp(at) = \lambda n \ln\left(\frac{n}{x_0}\right) \qquad (1)$$

and

$$\begin{aligned}\mathrm{var}(Y) &= x_0 \lambda \left[\exp(2at) - \exp(at)\right] \\ &= \frac{\lambda n (n - x_0)}{x_0}. \end{aligned} \qquad (2)$$

The dependence of (1) and (2) on x_0 is due to its effect only on the extreme upper tail of the distribution: the general shape of the distribution is unaffected. The **method of moments** is therefore a poor basis for inference about λ from the distribution of Y.

Lea & Coulson [19] derived a probability **generating function** for Y, now regarded as a discrete random variable:

$$G(\mu, z) = (1 - z)^{\mu(1-z)/z}. \qquad (3)$$

Eq. (3) can be derived under various assumptions, asymptotically for large n and small λ, and independently of x_0 [1].

Expansion of (3) gives the individual probabilities. For instance, the probabilities of 0 and 1 mutant are, respectively, $p_0 = e^{-\mu}$ (in agreement with the Poisson derivation), and $p_1 = (1/2)\mu e^{-\mu}$. Lea & Coulson provided tables of the distribution for selected values of μ up to 15. They also showed that a transformed

variable,

$$\chi = \frac{11.6}{y/\mu - \ln \mu + 4.5} - 2.02,$$

is approximately normally distributed with zero mean and unit variance.

Various methods of estimation from Lea & Coulson's distribution are described and illustrated by the authors and Armitage [2]. The use of the distribution is complicated by various possible deviations from the model.

Other models of mutations and several alternative methods of mutation rate estimation were introduced by Armitage [1], Mandelbrot [22], and Koch [15]. These models generalized the Luria & Delbrück model and the results of Lea & Coulson, and included the effects of phenotypic lag (that is, delay in detection of a mutant phenotype in clones), and differential fitness (difference in growth rates of mutant and wild-type cells) [1, 15].

Fluctuation analysis, as introduced by Luria & Delbrück, has been used in biological laboratories worldwide, and was adopted for the analysis of experiments with bacteria, yeast, and mammalian cells. Once again, bacterial models were applied to experimental conditions drastically different from the originally specified postulates of the Luria & Delbrück model (as had happened before with models of bacterial growth).

During the last decade a number of publications have provided a thorough revaluation of earlier mutation models and methods of mutational rate estimation [11, 30, 32, 33, 38, 40]. New models of mutation have taken into account reversible mutations and two-stage mutations. Several new computational methods were introduced for mutation rate estimation in mammalian cells [14, 32, 33, 39].

Mathematically interesting discussions of the Luria & Delbrück model have been published recently by Pakes [30], Stewart [39], and Nadas et al. [24, 25]. Many recent publications have pointed out that "certain mathematical artifacts had been appended by others to the Luria–Delbrück model" [24, 25], thus providing a source of confusion to the measurement of spontaneous mutation rates. Specifically Ma et al. [21] have pointed out, as noted by Armitage [1], that the Lea & Coulson approximation (3) has infinite mean and variance. As a mathematically sensible alternative several models have been proposed that avoid the infinite mean [33]. Some of them have adopted the **Galton–Watson** branching process [32]. A modification of fluctuation tests described by Nadas et al. [24, 25] allows one to estimate mutation rates using sufficiently large initial cell populations.

References

[1] Armitage, P. 1952. The statistical theory of bacterial populations subject to mutation, *Journal of the Royal Statistical Society, Series B* **14**, 1–40.

[2] Armitage, P. 1953. Statistical concepts in the theory of bacterial mutation, *Journal of Hygiene* **51**, 162–184.

[3] Boyd, A.W. 1983. Cell cycle kinetics data can be simulated by a simple chemical kinetic model, *Journal of Theoretical Biology* **101**, 355–372.

[4] Brooks, R.F., Bennett, D.C. & Smith, J.A. 1980. Mammalian cell cycles need two random transitions, *Cell* **19**, 493–504.

[5] Cooper, C. 1991. Bacterial Growth and Cell Division. *Biochemistry and Regulation of Prokaryotic and Eukaryotic Division Cycles*. Academic Press, New York.

[6] Gusev, Y. & Axelrod, D.E. 1995. Evaluation of models of inheritance of cell cycle times: computer simulation and recloning experiments, in *Mathematical Population Dynamics: Analysis of Heterogeneity*, O. Arino, D. Axelrod & M. Kimmel, eds. Wuerz Publishing, Winnipeg, pp. 97–116.

[7] Harvey, J.D. 1983. Mathematics of microbial age and size distributions, in *Mathematics in Microbiology*, M. Bazin, ed. Academic Press, New York, pp. 1–35.

[8] Hyver, C. & Guyader, H.L. 1990. MPF and cyclin: modeling of the cell cycle minimum oscillator, *BioSystems* **24**, 85–90.

[9] Jagers, P. 1975. *Branching Processes with Biological Applications*. Wiley, New York.

[10] Kelly, C.D. & Rahn, O. 1932. The growth rate of individual bacterial cells, *Journal of Bacteriology* **23**, 147–153.

[11] Kendal, W.S. & Frost, P. 1988. Pitfalls and practice of Luria-Delbrück fluctuation analysis: a review, *Cancer Research* **48**, 1060–1065.

[12] Kendall, D.G. 1952. On the choice of mathematical model to represent normal bacterial growth, *Journal of the Royal Statistical Society, Series B* **14**, 41–44.

[13] Kimmel, M. & Axelrod, D.E. 1991. Unequal cell division, growth regulation and colony size of mammalian cells: a mathematical model and analysis of experimental data, *Journal of Theoretical Biology* **153**, 157–180.

[14] Kimmel, M. & Axelrod, D.E. 1994. Fluctuation test for two-stage mutations: application to gene amplification, *Mutation Research* **306**, 45–60.

[15] Koch, A.L. 1982. Mutation and growth rates from Luria-Delbrück fluctuation tests, *Mutation Research*, **14**, 365–374.

[16] Koch, A.L. 1991. Evolution of ideas about bacterial growth and their pertinence to higher cells, in: *Mathematical Population Dynamics*, O. Arino, D.E. Axelrod & M. Kimmel, eds. Lecture Notes in Pure and Applied Mathematics, Vol. 131. Marcel Dekker, New York, pp. 561–575.

[17] Koch, A.L. & Schaecter, M. 1962. A model for the statistics of the cell division process, *Journal of General Microbiology* **29**, 435–454.

[18] Kuczek, T. 1984. Stochastic modeling for the bacterial life cycle, *Mathematical Biosciences* **69**, 159–169.

[19] Lea, D.E. & Coulson, C.A. 1949. The distributions of the number of mutants in bacterial populations, *Genetics* **49**, 264–289.

[20] Luria, S.E. & Delbrück, M. 1943. Mutations of bacteria from virus sensitivity to virus resistance, *Genetics* **28**, 491–511.

[21] Ma, W.T., Sandri, G.v.H. & Sarkar, S. 1992. Analysis of the Luria-Delbrück distribution using discrete convolution powers, *Journal of Applied Probability* **29**, 255–267.

[22] Mandelbrot, B. 1974. A population birth-and-mutation process, *Journal of Applied Probability* **11**, 437–444.

[23] Murray, A.W. & Kirschner, M.W. 1989. Dominoes and clocks: the union of two views of the cell cycle, *Science* **246**, 614–621.

[24] Nadas, A., Goncharova, E.I. & Rossman, T.G. 1996. Maximum likelihood estimation of spontaneous mutation rates from large initial populations, *Mutation Research* **351**, 9–17.

[25] Nadas, A., Goncharova, E.I. & Rossman, T.G. 1996. Mutation and infinity: improved statistical methods for estimating spontaneous rates, *Environmental and Molecular Mutagenesis* **28**, 90–99.

[26] Novak, B. & Tyson, J.J. 1995. Mathematical modeling of the cell division cycle, in *Mathematical Population Dynamics: Analysis of Heterogeneity*, O. Arino, D. Axelrod & M. Kimmel, eds. Wuerz Publishing, Winnipeg, pp. 155–170.

[27] Nurse, P. 1980. Cell cycle control – both deterministic and probabilistic?, *Nature* **286**, 9–10.

[28] Nurse, P. 1990. Universal control mechanism regulating onset of M-phase, *Nature* **344**, 503–507.

[29] Obeyesekere, M.N., Tucker, S.L. & Zimmerman, S.O. 1995. Mathematical models for regulation of the cell cycle via the concentrations of cellular proteins, in *Mathematical Population Dynamics: Analysis of Heterogeneity*, O. Arino, D. Axelrod & M. Kimmel, eds. Wuerz Publishing, Winnipeg, pp. 171–182.

[30] Pakes, A.G. 1993. Remarks on the Luria-Delbrück distribution, *Journal of Applied Probability* **30**, 991–994.

[31] Rigney, D.R. 1986. Multiple-transition cell cycle models that exhibit transition probability kinetics, *Cell Tissue Kinetics* **19**, 23–37.

[32] Rossman, T.G., Goncharova, E.I. & Nadas, A. 1995. Modeling and measurement of the spontaneous mutation rate in mammalian cells, *Mutation Research* **328**, 21–30.

[33] Sarkar, S., Ma, W.T. & Sandri, G.v.H. 1992. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants, *Genetica* **85**, 173–179.

[34] Sennerstam, R. & Stromberg, J.O. 1995. Contradictory conclusions from subcompartments of G1 phase are resolved by the two-subcycle cell cycle model, in *Mathematical Population Dynamics: Analysis of Heterogeneity*, O. Arino, D. Axelrod & M. Kimmel, eds. Wuerz Publishing, Winnipeg, pp. 201–216.

[35] Shapiro, A. 1946. The kinetics of growth and mutations in bacteria, *Cold Spring Harbor Symposium on Quantitative Biology* **11**, 228–234.

[36] Smith, J.A. & Martin, L. 1973. Do cells cycle?, *Proceedings of the National Academy of Sciences* **76**, 1279.

[37] Staudte, R.G., Guiguet, M. & Collyn d'Hooghe, M. 1984. Additive models for dependent cell populations, *Journal of Theoretical Biology* **109**, 127–146.

[38] Stewart, F.M. 1991. Fluctuation analysis: the effect of plating efficiency, *Genetica* **84**, 51–55.

[39] Stewart, F.M. 1994. Fluctuation tests: how reliable are the estimates of mutation rates?, *Genetics* **137**, 1139–1146.

[40] Stewart, F.M., Gordon, D.M. & Levin, B.R. 1990. Fluctuation analysis: the probability distribution of the number of mutants under different conditions, *Genetics* **124**, 175–185.

[41] Stocker, B.A.D. 1949. Measurements of rate of mutation of flagellar antigenic phase in Salmonella typhi-murium, *Journal of Hygiene* **47**, 398–413.

[42] Thron, C.D. 1991. Mathematical analysis of model of the mitotic clock, *Science* **254**, 122–123.

[43] Traganos, F., Kimmel, M., Bueti, C. & Dazynkiewicz, Z. 1987. Effects of inhibition of RNA and protein synthesis on CHO cell cycle progression, *Journal of Cell Physiology* **133**, 277–287.

[44] Tyson, J.J. 1987. Size control of cell division, *Journal of Theoretical Biology* **126**, 381–391.

[45] Tyson, J.J. 1991. Modelling the cell division cycle: cdc2 and cyclin interactions, *Proceedings of the National Academy of Sciences* **88**, 7328–7332.

(*See also* **Branching Processes**; **Stochastic Processes**)

Y. GUSEV & PETER ARMITAGE

Bagging and Boosting

Introduction

The past decade has witnessed an explosion of machine learning papers and research. Given the large number of **algorithms** proposed and explored for classification and regression problems (*see* **Tree-structured Statistical Methods**), it is not surprising that methods that combine these algorithms have been proposed as well. Statisticians have employed prediction averaging in many settings. In **forecasting**, averaging different models is a common way to reduce the **variance** and the **bias** inherent in choosing a single model form. In the machine learning context, two ensemble methods called *bagging and boosting* have become popular for combining models.

Bagging

Bagging is an acronym for "*bootstrap aggregating*". First introduced by Leo Breiman in 1996 [2], bagging is simple to describe and to implement. Given a certain model form (like a classification tree, or a **linear regression**), take repeated **bootstrap** samples of the original data set, refitting the model each time. Each of these models will be slightly different, and the predictions at each point will vary from model to model. The "bagged" prediction for each point is an "average" over the predictions of all the models for that point.

More precisely, let the data set on which we will build the model be described by $\{(y_i, x_i), i = 1, \ldots, N\}$. This is the so-called *training set*. Now, suppose we have a model, f, whose prediction for the ith point is $f(x_i) = \hat{y}_i$. Call the set of predictions of all the data points $\{\mathcal{L}\}$. Take repeated bootstrap samples from the training set, each consisting of N cases as well, drawn with replacement from the original $\{(y_i, x_i)\}$. For each of these, refit the model $f^{(\mathrm{B})}$ and get a new set of predictions $\{\mathcal{L}^{(\mathrm{B})}\}$. Repeat this many times.

For a **classification** problem, we may want to predict the class of x_i. If there are j classes, each prediction will be a class $j \in \{1, \ldots, J\}$. For each case x_i, we will have many different predictions given by the successive bootstrapped data sets. The usual method for obtaining the "bagged" prediction is to let these different models "vote". Take the most frequent **modal** predicted class as the bagged prediction. For a regression problem, where the output of f is numeric, we use the same procedure, but take an average (or **median** or other averaging procedure) of the predictions.

The most common class of models used for bagging are decision trees (using classification or regression trees, depending on the form of the response). Leo Breiman's implementation of bagged trees are called "Random Forests" [3].

Because of the fact that it averages predictions across many bootstrapped samples, it seems logical that the bagged prediction reduces the prediction variance. This fact is demonstrated by Breiman [2]. Bagging works best on models that have inherently high variance. Of course, the amount of variance reduction and the resulting improvement in error rate are dependent on the data set and the model employed, but empirical studies have shown that variance reductions of 10 or even 20% are not uncommon (see [1]). Unfortunately, bagging appears to do little to reduce the bias, which is often a problem with small decision trees and other so-called weak learners.

Boosting

Boosting is also a method for combining an ensemble of predictions, but it does so in quite a different way from bagging. Again we start with a model f from which we get a prediction on the training data $\{(y_i, x_i), i = 1, \ldots, N\}$. In boosting, however, we reweight the data set depending on the performance of the model. The observations that are misclassified are given *higher* weights and the model is refit. This process is repeated many times. At the end, we have a succession of predictions from these various reweighted fits. The boosted prediction is a weighted "vote" of the outputs from these models, this time reweighted by the performance of the model. Boosting puts successively more importance on the misclassified data by increasing their weights, thus making the problems harder as the number of models increases.

Boosting has received a lot of attention, especially in the machine learning literature where it first appeared [4, 5, 10]. In its simplest form, it is used on classification problems, as originally described by Freund and Shapire [5]. To make things even simpler, we will describe the case of only two classes:

-1 or $+1$. Here is the original "Adaboost" algorithm from [5] as described by Friedman, Hastie, and Tibshirani in [8]:

1. Initialize training set using weights $w_i = 1/N$ for all i.
2. Repeat a, b, c for $m = 1, 2, \ldots .M$:
   a. Fit the model $f_m(x)$ using weights w_i. The output for each point is either -1 or $+1$.
   b. Compute the overall weighted error rate $e_m = \left(\sum w_i (f_m(x_i) \neq y_i)\right)/\sum w_i$ and the adjusted overall weighted error rate $c_m = \log((1 - e_m)/e_m)$
   c. Set $w_i = w_i \exp(c_m \times 1_{y_i \neq f_m(x_i)})$ and renormalize so that $\sum w_i = 1$.
3. The resulting output $f_b(x_i)$ for each point is a weighted average of the predictions of the each $f_m : f_b(x_i) = \text{sign}\left(\sum_m c_m f_m(x_i)\right)$.

(The function $1_{\{A\}}$ gives the value 1 if A is true; 0 otherwise.) The sign function is used at the final stage because each prediction is either -1 or $+1$. Thus, if the (weighted) majority of votes is positive, the boosted prediction is $+1$; otherwise it is -1. Other schemes can be used to combine the models when the number of classes is greater than 2, or the output is continuous. Notice that boosting used weighting twice. The model is refit each time with cases by whether they were correctly predicted or not, with misclassified cases receiving *higher* weights. For the final predictions, the results of all the models are averaged, but models with lower error rates are given higher weight.

In spite of the simplicity of the algorithm, the reason for its success in reducing both the bias and variance of the individual models remained mysterious for some time. A great deal of light was shed by Friedman, Hastie, and Tibshirani [8, 9] who showed that the Adaboost algorithm fits an additive model using a loss function that is well suited for classification. Armed with this insight, they suggest several variants of boosting for both classification and regression. A version using boosted decision trees called TreeBoost (and later Multiple Additive Regression Trees or MART) was proposed by Friedman [6, 7].

The reason why boosting reduces both the bias and variance of simple learning algorithms is still not completely understood. Nor is it clear exactly why boosting seems to avoid the problem of overfitting even though hundreds, or even thousands of successive models are fit. However, the empirical evidence is strong that in a great many data sets, boosting can substantially lower the overall error rate (e.g. see [1]). Like bagging, the final "model" obtained by boosting is not interpretable, because the predictions are obtained by averaging over many different models. However, in the case of boosted trees, some progress has been made for help in interpreting variable importance and the influence of the individual predictors [6].

Further Comments

Bagging and boosting share several properties. Both are used to improve the prediction accuracy of a class of models by creating a "committee" of models that then vote to combine their predictions. Bagging does so by resampling from the original data set while boosting successively focuses on the part of the data set that was not fit well. Because they combine predictions, both are useful for reducing variance of predictions when compared with individual model fits. Both appear to be resistant to overfitting in spite of the number of models that are created. Boosting appears to be able to reduce bias as well. They also share a disadvantage in that the final combined model is not interpretable. While an individual tree is rule-based and "explains" why it makes the predictions it does, the combined model simply takes a majority vote of a committee.

References

[1] Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants, *Machine Learning* **36**(1–2), 105–139.

[2] Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**(2), 123–140.

[3] Breiman, L. (1999). *Random Forests – Random Features*, Technical Report 567, Statistics Department, University of California, Berkeley.

[4] Freund, Y. (1995). Boosting a weak learning algorithm by majority, *Information and Computation* **121**(2), 256–285.

[5] Freund, Y. & Shapire, R.E. (1996). Experiments with a new boosting algorithm, *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann, San Francisco, pp. 148–156.

[6] Friedman, J.H. (2002). *Tutorial – Getting Started with MART in R*, Technical Report, Department of Statistics, Stanford University, Stanford, CA, `http://www-stat.stanford.edu/~jhf/r-mart/tutorial/tutorial.pdf`.

[7] Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29**(5), 1189–1232.

[8] Friedman, J.H., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), *Annals of Statistics* **28**(2), 337–407.

[9] Hastie, T., Tibshirani, R. & Friedman, J. (2002). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York.

[10] Shapire, R.E. (1990). The strength of weak learnability, *Machine Learning* **5**(2), 197–227.

RICHARD DE VEAUX

Balanced Incomplete Block Designs

When evaluating treatment effects, it is often desirable to assign treatments randomly within homogeneous blocks of experimental units, thus eliminating the effect of differences between blocks when evaluating the differences between treatments (*see* **Blocking**). A **randomized complete blocks design** includes all treatments of interest within each block. In some circumstances, however, there are more treatments of interest than units available per block, so such a design is not possible. There are many naturally occurring blocks that contain limited numbers of experimental units. In some studies of the effect of innoculation on lesions in plants, each leaf may be a block and the two halves of each leaf are experimental units. Studies using twin pairs in humans as blocks are obviously limited to two units per block (*see* **Twin Analysis**). Studies of growth rates in animals often use litters as blocks, and there is a limited number of animals per litter. If two or more factors are crossed to form many blocks, there may be inadequate resources for measuring enough units for each block or combination of factors. A randomized block design with less than the full number of treatments in each block is an **incomplete block design**. If all pairs of treatments occur equally often, it is said to be a balanced incomplete block design (bibd).

The data from a bibd can be analyzed using standard **analysis of variance**. In a bibd, all pairwise treatment contrasts have equal efficiency, so this design is ideal when all pairwise contrasts are of equal interest. The effect of blocks, or interblock information, can also be analyzed from a bibd, although this is usually not of interest.

The concept of the bibd was originally proposed by **Yates** [18]. If there are a total of t treatments to be tested, then the total number of pairs of treatments is $\frac{1}{2}t(t-1)$. If only two units per block are available, then a bibd would assign each pair of treatments to an equal number of blocks; the number of blocks must then be some multiple of $\frac{1}{2}t(t-1)$. If there are $t-2$ units available in each block, then a complementary bibd arrangement is to delete each pair of treatments one at a time and assign the remaining $t-2$ treatments to each block; again, this would require the number of blocks to be some multiple of $\frac{1}{2}t(t-1)$. Yates provides bibd arrangements for different numbers of treatments, up to 10, and different numbers of units per block. He also outlines the standard **least squares** analysis and addresses issues of efficiency.

The standard notation for the parameters of a bibd design is (t, b, r, k, λ), where t is the number of treatments, b is the number of blocks, r is the number of replications of each pair of treatments, k is the number of experimental units per block, and λ is the number of replications of each pair of treatments. In order for the design to be balanced, the relationship $\lambda = r(k-1)/(t-1)$ must hold. A bibd is not possible, then, unless $r(k-1)/(t-1)$ is an integer. Following the initial work by Yates, there has been much attention to constructing more bibds. The problem is to find a set of parameters that satisfy the above condition, and to specify an appropriate arrangement of treatments within blocks given a set of values for the parameters. In the study of combinatorics, this problem is a special case of construction of tactical configurations. Bose [3] originally proposed the method of symmetric differences, which is an application of combinatoric methods to the construction of bibds. Further work on construction of bibds includes tables of specific arrangements or methods of constructing arrangements for different feasible combinations of parameter values [1, 5–11].

A nested balanced incomplete block design is a bibd with two systems of blocks, one nested within the other, such that ignoring either system of blocks leaves a bibd based on the remaining system of blocks. Nested bibds were originally proposed by Preece [12]. Preece provides several arrangements for such designs and gives the least squares equations for analysis. Singh & Dey [13] introduced the balanced incomplete block design with nested rows and columns, which adds yet another dimension to the blocking scheme, organized as rows and columns nested within the initial blocking system. Singh & Dey describe analysis of such designs and propose several arrangements. Further work has been done on construction of bibds with nested rows and columns [2, 4, 14–17].

References

[1] Abel, R.J.R. (1994). Forty-three balanced incomplete block designs, *Journal of Combinatorial Theory, Series A* **65**, 252–267.

[2] Agarwal, H.L. & Prasad, J. (1982). Some methods of construction of balanced incomplete block designs with nested rows and columns, *Biometrika* **69**, 481–483.

[3] Bose, R.C. (1939). On the construction of balanced incomplete block designs, *Annals of Eugenics* **9**, 353–399.

[4] Cheng, C. (1986). A method of constructing balanced incomplete-block designs with nested rows and columns, *Biometrika* **73**, 695–700.

[5] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York.

[6] Davies, O.L. (1956). *Design and Analysis of Industrial Experiments*, 2nd Ed. Hafner, New York.

[7] Fisher, R.A. & Yates, F. (1953). *Statistical Tables for Biological, Agricultural, and Medical Research*, 4th Ed. Oliver & Boyd, Edinburgh.

[8] Hanani, H. (1975). Balanced incomplete block designs and related designs, *Discrete Mathematics* **11**, 255–369.

[9] Hanani, H. (1989). BIBDs with block-size seven, *Discrete Mathematics* **77**, 89–96.

[10] Mathon, R. & Rosa, A. (1989). On the (15, 5, λ)-family of BIBDs, *Discrete Mathematics* **77**, 205–216.

[11] Mathon, R. & Rosa, A. (1990). Tables of parameters of BIBDs with $r \leq 41$ including existence, enumeration and resolvability results: an update, *Archives of Combinatorics* **30**, 65–96.

[12] Preece, D.A. (1967). Nested balanced incomplete block designs, *Biometrika* **54**, 479–486.

[13] Singh, M. & Dey, A. (1979). Block designs with nested rows and columns, *Biometrika* **66**, 321–326.

[14] Sreenath, P.R. (1989). Construction of some balanced incomplete block designs with nested rows and columns, *Biometrika* **76**, 399–402.

[15] Sreenath, P.R. (1991). Construction of balanced incomplete block designs with nested rows and columns through the method of differences, *Sankhyā, Series B* **53**, 352–358.

[16] Uddin, N. & Morgan, J.P. (1990). Some constructions for balanced incomplete block designs with nested rows and columns, *Biometrika* **77**, 193–202.

[17] Uddin, N. & Morgan, J.P. (1997). Further constructions for orthogonal sets of balanced incomplete block design with nested rows and columns, *SankyaB* **59**, 156–163.

[18] Yates, F. (1936). Incomplete randomized blocks, *Annals of Eugenics* **7**, 121–140.

SALLY FREELS

Ban Estimates

The BAN (best asymptotically normal) estimator is optimal in large samples in a similar way that the **Minimum Variance Unbiased (MVU) estimator** is in small samples. The concept was originally developed by Neyman [8] for parameter **estimation** in models involving the **multinomial distribution**. Suppose each of n randomly selected individuals can be assigned to $s+1$ mutually exclusive categories $C_0, C_1, \ldots, C_s$ with probabilities $\mu_0, \mu_1, \ldots, \mu_s$, where $\mu_0 + \mu_1 + \cdots + \mu_s = 1$. Let $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_s]'$ denote the s-dimensional vector of parameters where μ_0 is determined by the constraint to sum to 1. Let $n_0, n_1, \ldots, n_s$ be the random variables representing the counts of individuals in the $s+1$ categories where $n_0 + n_1 + \cdots + n_s = n$. Let $\mathbf{y}_n = [y_1, y_2, \ldots, y_s]' = [(n_1/n), (n_2/n), \ldots, (n_s/n)]'$ denote the s-dimensional vector of sample proportions where $y_0 = n_0/n$ is determined by the constraint to sum to 1. The crucial feature for asymptotic results is the following:

$$\sqrt{n}(\mathbf{y}_n - \boldsymbol{\mu}) \to N_s[\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\mu})], \tag{1}$$

where $\text{var}(\mathbf{y}_n) = \boldsymbol{\Sigma}(\boldsymbol{\mu})/n$. These count data can be arranged in two-way, $\ldots$, k-way tables and are the subject of **categorical data analysis**. The arrangement of the counts into tables suggests various hypotheses, e.g. independence of row and column effects in a two-way **contingency table**. A hypothesis H can often be described by a set of *constraint equations*

$$\text{H: } \mathbf{f}(\boldsymbol{\mu}) = 0, \tag{2}$$

where f is a vector-valued function of dimension $r < s$ having differential $\mathbf{F}(\boldsymbol{\mu}) = \partial \mathbf{f}(\boldsymbol{\mu})/\partial \boldsymbol{\mu}$ satisfying certain regularity conditions.

From this point on we assume sufficient regularity conditions for the different mathematical operations to be defined and convergence results to hold. Except where noted, all results, regularity conditions, and technical details may be found in [1–3]. A standard approach of estimating μ given H is by **maximum likelihood**. Let $\boldsymbol{\mu}_{\text{E}} = \boldsymbol{\mu}_{\text{E}}(\mathbf{y}_n)$ denote the maximum likelihood estimator of $\boldsymbol{\mu}$ given H. Then $\boldsymbol{\mu}_{\text{E}}$ has the property of being consistent asymptotically normal (CAN)

$$\sqrt{n}(\boldsymbol{\mu}_{\text{E}} - \boldsymbol{\mu}) \to N_s(\mathbf{0}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{F}'(\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}')^{-1}\mathbf{F}\boldsymbol{\Sigma}), \tag{3}$$

with *minimal* asymptotic covariance matrix:

$$\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{F}'(\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}')^{-1}\mathbf{F}\boldsymbol{\Sigma}, \tag{4}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\mu})$ and $\mathbf{F} = \mathbf{F}(\boldsymbol{\mu})$.

A CAN estimator with minimal asymptotic covariance matrix is *best*, and is referred to as BAN. The notion of *minimal* is based on the following order relationship among symmetric matrices: $A \leq B$ if $B - A$ is nonnegative definite. In practice, BAN estimators only exist subject to *regularity* conditions and are denoted RBAN. Therefore RBAN estimators are *regular* CAN estimators with *minimal* asymptotic covariance matrix. Maximum likelihood estimators are often difficult to compute and Neyman [8] showed that minimum Pearson χ^2, Neyman χ^2, and linearized Neyman χ^2 estimators are also RBAN [6]. RBAN estimators can also be obtained as the roots of certain linear forms [5].

General Multivariate Distributions

We now drop the assumption of the multinomial distribution. Let $x_1, x_2, \ldots, x_n$ be a random sample from an arbitrary s-dimensional multivariate distribution where $\text{E}(\mathbf{x}_i) = \boldsymbol{\mu}$ and $\text{var}(\mathbf{x}_i) = \boldsymbol{\Sigma}(\boldsymbol{\mu})$. Let $\mathbf{y}_n = (\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)/n$ be the s-dimensional average of the random sample. Then $\mathbf{y}_n$ satisfies (1) and any estimator $\boldsymbol{\mu}_{\text{E}} = \boldsymbol{\mu}_{\text{E}}(\mathbf{y}_n)$ of $\boldsymbol{\mu}$ given (2) is RBAN if it satisfies (3). If $\boldsymbol{\mu}_{\text{E}}$ is any estimator of $\boldsymbol{\mu}$ and $\mathbf{f}(\boldsymbol{\mu}_{\text{E}}) = 0$, then $\boldsymbol{\mu}_{\text{E}}$ is said to be *admissible* (this is a term in the BAN literature not to be confused with the decision-theoretic meaning; *see* **Decision Theory**). An *admissible* RBAN estimator $\boldsymbol{\mu}_{\text{P}}$ may be obtained as the solution to the following equation:

$$\boldsymbol{\mu} = \mathbf{y}_n - \boldsymbol{\Sigma}\mathbf{F}'(\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}')^{-1}\mathbf{F}(\mathbf{y}_n - \boldsymbol{\mu}), \tag{5}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\mu})$ and $\mathbf{F} = \mathbf{F}(\boldsymbol{\mu})$.

Another *admissible* RBAN estimator that may be easier to compute is μ_{N}, the solution to the following equation:

$$\boldsymbol{\mu} = \mathbf{y}_n - \boldsymbol{\Sigma}_n\mathbf{F}'(\mathbf{F}\boldsymbol{\Sigma}_n\mathbf{F}')^{-1}\mathbf{F}(\mathbf{y}_n - \boldsymbol{\mu}), \tag{6}$$

where $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}(\mathbf{y}_n)$ and $\mathbf{F} = \mathbf{F}(\boldsymbol{\mu})$.

If one is willing to sacrifice admissibility, then the following *closed-form solution* is RBAN (but not guaranteed to be admissible):

$$\boldsymbol{\mu}_{\text{N}^*} = \mathbf{y}_n - \boldsymbol{\Sigma}_n\mathbf{F}_n'(\mathbf{F}_n\boldsymbol{\Sigma}_n\mathbf{F}_n')^{-1}\mathbf{f}(\mathbf{y}_n), \tag{7}$$

where $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}(\mathbf{y}_n)$ and $\mathbf{F}_n = \mathbf{F}(\mathbf{y}_n)$.

Eqs. (5) and (6) are in the form

$$\boldsymbol{\mu} = h(\boldsymbol{\mu}), \tag{8}$$

which together with (7) suggest an iterative solution for $\boldsymbol{\mu}_{\mathrm{P}}$ and $\boldsymbol{\mu}_{\mathrm{N}}$:

$$\boldsymbol{\mu}_{\mathrm{P}(i+1)} = h(\boldsymbol{\mu}_{\mathrm{P}(i)}), \tag{9}$$

with $\boldsymbol{\mu}_{\mathrm{P}(0)} = \boldsymbol{\mu}_{\mathrm{N}^*}$, where $\boldsymbol{\mu}_{\mathrm{P}(i)} \to \boldsymbol{\mu}_{\mathrm{P}}$, and

$$\boldsymbol{\mu}_{\mathrm{N}(i+1)} = h(\boldsymbol{\mu}_{\mathrm{N}(i)}), \tag{10}$$

with $\boldsymbol{\mu}_{\mathrm{N}(0)} = \boldsymbol{\mu}_{\mathrm{N}^*}$, where $\boldsymbol{\mu}_{\mathrm{N}(i)} \to \boldsymbol{\mu}_{\mathrm{N}}$.

RBAN estimators usually have associated χ^2 test statistics for the hypothesis H. With this in mind, define the following Pearson and Neyman χ^2 functions:

$$\chi^2_{\mathrm{P}}(\boldsymbol{\mu}) = n(\mathbf{y}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_n - \boldsymbol{\mu}), \tag{11}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\mu})$,

$$\chi^2_{\mathrm{N}}(\boldsymbol{\mu}) = n(\mathbf{y}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}_n^{-1}(\mathbf{y}_n - \boldsymbol{\mu}), \tag{12}$$

where $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}(\mathbf{y}_n)$.

If $\boldsymbol{\mu}_B$ is any RBAN estimator, then the following hold:

$$\chi^2_{\mathrm{P}}(\boldsymbol{\mu}_B) \to \chi^2(r) \tag{13}$$

and

$$\chi^2_{\mathrm{N}}(\boldsymbol{\mu}_{\mathrm{B}}) \to \chi^2(r), \tag{14}$$

where r is the dimension of $\mathbf{f}$.

Multivariate Exponential Family Distributions

Let $x_1, x_2, \ldots, x_n$ be a random sample from the s-dimensional multivariate **exponential family** [4] with probability density function (pdf) of the form:

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp[\mathbf{x}'\boldsymbol{\theta} - q(\boldsymbol{\theta})]. \tag{15}$$

This family includes the s-variate multinomial, s-variate negative multinomial, s-variate Poisson, and s-variate logarithmic series among others. Let $\mathbf{y}_n = (\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)/n$ be the s-dimensional average of the random sample. Then $\mathbf{y}_n$ satisfies (1), where $\mathrm{E}(\mathbf{x}_i) = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{x}_i) = \boldsymbol{\Sigma}(\boldsymbol{\mu})$. Therefore all the results of the previous section apply. In addition, we have the following equivalence:

$$\boldsymbol{\mu}_{\mathrm{P}} = \text{maximum likelihood estimator}, \tag{16}$$

$$\chi^2_{\mathrm{P}}(\boldsymbol{\mu}_{\mathrm{P}}) = \text{Rao's efficient score statistic}, \tag{17}$$

$$\chi^2_{\mathrm{N}}(\boldsymbol{\mu}_{\mathrm{N}^*}) = \text{Wald statistic} \tag{18}$$

(*see* **Likelihood**). The above RBAN estimation and test criteria therefore reduce to familiar methods for the multivariate exponential family.

Sum Symmetric Power Series (SSPS) Distributions

A special case of the s-variate exponential family is the SSPS family of distributions [7] for count data which include the s-variate multinomial, s-variate negative multinomial, and s-variate Poisson. For this family the s-variate vector $\mathbf{x}$ from (15) has the structure $\mathbf{x} = [n_1, n_2, \ldots, n_s]'$, where the n_i represent count data (i.e. nonnegative integers). For the SSPS family the functions (11) and (12) have the familiar form of the Pearson χ^2 and Neyman χ^2 for multinomial data:

$$\begin{aligned}\chi^2_{\mathrm{P}}(\boldsymbol{\mu}) &= n(\mathbf{y}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_n - \boldsymbol{\mu}) \\ &= \sum_i \frac{(n_i - n\mu_i)^2}{(n\mu_i)},\end{aligned} \tag{19}$$

$$\begin{aligned}\chi^2_{\mathrm{N}}(\boldsymbol{\mu}) &= n(\mathbf{y}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}_n^{-1}(\mathbf{y}_n - \boldsymbol{\mu}) \\ &= \sum_i \frac{(n_i - n\mu_i)^2}{(n_i)},\end{aligned} \tag{20}$$

where $\sum_i$ represents the summation for $i = 0, 1, \ldots, s$, and μ_0 and n_0 have special definitions based on constraints. The constraints for the multinomial were $\mu_0 + \mu_1 + \cdots + \mu_s = 1$ and $n_0 + n_1 + \cdots + n_s = n$. Constraints in the other cases are described in [1].

Return to the Multinomial Distribution

Because the multinomial distribution is a special case of the SSPS family, the general RBAN estimation and test criteria that we developed reduce to familiar

forms for the multinomial distribution:

$$\boldsymbol{\mu}_{\mathrm{P}} = \text{maximum likelihood estimator}, \quad (21)$$

$$\boldsymbol{\mu}_{\mathrm{N}} = \text{minimum Neyman } \chi^2 \text{ estimator}, \quad (22)$$

$$\boldsymbol{\mu}_{\mathrm{N}^*} = \text{linearized minimum Neyman } \chi^2 \text{ estimator}, \quad (23)$$

$$\chi^2_{\mathrm{P}}(\boldsymbol{\mu}_{\mathrm{P}}) = \text{Pearson } \chi^2 \text{ test statistic}, \quad (24)$$

$$\chi^2_{\mathrm{N}}(\boldsymbol{\mu}_{\mathrm{N}}) = \text{Neyman } \chi^2 \text{ test statistic}, \quad (25)$$

$$\chi^2_{\mathrm{N}}(\boldsymbol{\mu}_{\mathrm{N}^*}) = \text{linearized Neyman } \chi^2 \text{ test statistic}. \quad (26)$$

Other Directions

The results in the section on general multivariate distributions hold if $\mathbf{y}_n$ is composed of averages of multiple independent random samples. They also hold if $\mathbf{y}_n$ is an arbitrary random vector (not necessarily an average) that satisfies (1). Similar results hold if the hypothesis H in (2) is expressed in other ways. *Constraint equations* are the most general form, but it may be more natural to express H in the form of *freedom equations* $\boldsymbol{\mu} = g(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is of dimension r. Another way to model H is $d(\boldsymbol{\mu}) = e(\boldsymbol{\lambda})$ or as a special case the **general linear model** $d(\boldsymbol{\mu}) = X\lambda$ for design matrix X. These situations and their RBAN estimates and test criteria are covered in [2].

References

[1] Bemis, K.G. & Bhapkar, V.P. On the equivalence of some test criteria based on BAN estimators for the multivariate exponential family, *Journal of Statistical Planning and Inference* **6**, 277–286.

[2] Bemis, K.G. & Bhapkar, V.P. On BAN estimators for chi squared test criteria, *Annals of Statistics* **11**, 183–196.

[3] Bemis, K.G. & Bhapkar, V.P. BAN estimation for chi-square test criteria in categorical data, *Communications in Statistics – Theory and Methods* **12**(11), 1211–1223.

[4] Berk, R.H. Consistency and asymptotic normality of mle's for exponential models, *Annals of Mathematical Statistics* **43**, 193–204.

[5] Ferguson, T.S. A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities, *Annals of Mathematical Statistics* **29**, 1046–1062.

[6] Hsiao, C. *Minimum chi-square, in Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley-Interscience, New York, pp. 518–522.

[7] Joshi, S.W. & Patil, G.P. Certain structural properties of the sum-symmetric power series distributions, *Sankhyā, Series A* **33**, 175–184.

[8] Neyman, J. Contribution to the theory of the χ^2 test, in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 239–273.

(*See also* **Large-sample Theory**)

KERRY G. BEMIS

Barahona–Poon Test

The Barahona–Poon test is used to detect nonlinear dynamics in time series (*see* **Nonlinear Time Series Analysis**). Nonlinear dynamical systems may generate deterministic signals, which can be easily mistaken for random noise. Distinguishing nonlinear deterministic dynamics, or "**chaos**", from random noise helps understand the physical process generating the data, and may improve short-term prediction (*see* **Forecasting**). In practice, however, identifying nonlinear dynamics is difficult because methods to detect chaos are often degraded by measurement noise, and need long data sets. The test proposed by C. S. Poon and M. Barahona is particularly suitable to detect chaos in short and noisy time series [1], typical of biological studies where nonstationarities limit the length of the observations [3]. The method compares linear and nonlinear models fitted to the data, and rejects the **null hypothesis** (i.e. that the time series is stochastic with linear dynamics) if at least one nonlinear model is significantly more predictive than all the linear models considered.

The technique is based on modeling the time series $y(n)$ $(n = 1, \ldots, N)$ as the output of a discrete Volterra series of nonlinear degree d and memory k. The predicted value of the series at time n, $y_{d,k}(n)$, is calculated as:

$$\begin{aligned} y_{d,k}(n) = {} & a_0 + a_1 y(n-1) + a_2 y(n-2) \\ & + \cdots + a_k y(n-k) + a_{k+1} y(n-1)^2 \\ & + a_{k+2} y(n-1) y(n-2) \\ & + \cdots + a_{M-1} y(n-k)^d \end{aligned} \tag{1}$$

When $d = 1$, $y(n)$ is simply modeled by an autoregressive linear equation of order k (*see* **ARMA and ARIMA Models**). The $M = (k+d)!/(k!d!)$ coefficients a_m, corresponding to all the combinations of $y(n-i)$ $(i = 1, \ldots, k)$ up to degree d, can be estimated from the data by a fast **algorithm** [2]. The one-step prediction power of each model, $\varepsilon^2_{d,k}$, is:

$$\varepsilon^2_{d,k} = \frac{\displaystyle\sum_{n=1}^{N} (y_{d,k}(n) - y(n))^2}{\displaystyle\sum_{n=1}^{N} (y(n) - \overline{y})^2} \tag{2}$$

where $\overline{y} = \sum_{n=1}^{N} y(n)/N$. Thus $\varepsilon^2_{d,k}$ is the variance of the prediction errors normalized by the variance of the time series.

First, one fixes the total dimension M. Then the best linear model is found by setting $d = 1$ in (1), and by iteratively increasing k from 1 to M. At each step k, a cost function is defined as

$$\mathrm{C}(r) = \log(\varepsilon_{d,k}) + \frac{r}{\mathrm{N}} \tag{3}$$

where r is the number of polynomials in the truncated Volterra series. The best linear model is the one minimizing $\mathrm{C}(r)$. To find the best nonlinear model, $\mathrm{C}(r)$ is computed again for $d > 1$ and for increasing values of k up to $r = M$. Similarly, the best nonlinear model is the one with the lowest $\mathrm{C}(r)$. The presence of "chaos" is indicated by a cost function lower for the best nonlinear model than for the best linear one. In this case, an objective statistical criteria is applied to reject the null hypothesis that nonlinear models are no better than linear models. For Gaussian prediction errors, the F-test is used to reject, with a certain level of confidence, the hypothesis that ε^2 is the same for the best linear and nonlinear model. Alternatively, the nonparametric **Wilcoxon–Mann–Whitney** rank-sum statistic is used.

References

[1] Barahona, M. & Poon, C.S. (1996). Detection of nonlinear dynamics in short, noisy time series, *Nature* **381**, 215–217.

[2] Korenberg, M.J. (1988). Identifying nonlinear difference equation and functional expansion representations: the fast orthogonal algorithm, *Annals of Biomedical Engineering* **16**, 123–142.

[3] Poon, C.S. & Merril, C.K. (1997). Decrease of cardiac chaos in congestive heart failure, *Nature* **389**, 492–495.

(*See also* **Time Series**)

PAOLO CASTIGLIONI

Barnard, George Alfred

Born: Walthamstow, UK; 23 September, 1915
Died: Brightlingsea, UK; 30 July, 2002

Barnard was born of working-class parents in suburban London and from the local school gained a scholarship to St. John's College, Cambridge, graduating in mathematics. From 1937 to 1939, he held a postgraduate award at Princeton University, USA, mainly studying mathematical logic. The war saw a return to England and a post with an engineering firm, displaying in the move a versatility that embraced theory and practice, a feature that remained with him throughout his life. In 1942, he joined the Ministry of Supply in a group dealing with production problems. There his serious interest in statistics began, to be continued at Imperial College London from 1945 until 1966, when he moved to the University of Essex, retiring in 1975. His retirement years were partly spent at other universities, especially that at Waterloo, Canada, but he kept up a lively correspondence with colleagues throughout the world. His political activities at Princeton resulted in his being refused a US visa for many years. He had the most engaging personality, perhaps shown at its best in debate where the author had the feeling that Barnard wanted to get at the truth, not just to defend his own position.

Barnard made major contributions to the basic ideas of statistics. He supported the **Bayesian** position in **inference**, restricting its use to cases where there existed prior knowledge that could be expressed in terms of probability, and in decision making where losses could be quantified (*see* **Decision Theory**), for example, in **quality control**. In cases where prior knowledge of this type was not available, he advocated methods based on **likelihood**; but he also felt there were cases where even a likelihood was not available and he would use tail-area significance tests (*see* **Hypothesis Testing**). He was an eclectic, but one whose reasons for using a tool were always clear. He was a great admirer of **R.A. Fisher** and was one of the very few people who could dispute with the great man without a touch of sycophancy.

He may have been the first to introduce the concept of conjugate distributions into the Bayesian canon, referring to them as distributions closed under sampling (*see* **Exponential Family**). In the field of likelihood, he was a pioneer, being the first to recognize the **likelihood principle**, this in 1947. He put likelihood inference onto an axiomatic basis [1] and produced, with Jenkins and Winsten, a seminal paper [2], which provided further support for the likelihood concept, applying it to many important statistical models. Both papers are still worth reading. As a consequence of understanding gained there, he was an advocate of the idea that, in most circumstances, the stopping rule was irrelevant to inference; an idea that has still not found general acceptance (*see* **Sequential Analysis**). He also advocated the method of pivots, where a function of parameter and data can be found having a known distribution, so that the data density can be transferred to provide one for the parameter.

It was not just the foundations that interested him: he was also concerned with the practical uses of statistics, for example, the British standard for condoms was largely written by him. A conversation with him could extend from the relevance of Gödel's theorem in inference to the design of schemes for industrial inspection, where he introduced the cumulative-sum (or cusum) chart (*see* **Quality Control in Laboratory Medicine**). He wrote on a great variety of practical problems acting, in his own words, as a midwife between the data and the decision. He was awarded the Deming medal of the American Association of Quality Control in 1991.

He was president of three societies, Operations Research, Institute of Mathematics and its Applications, and the **Royal Statistical Society**. Each of the presidential addresses is marked by the great breadth of subject matter, coming both from his mathematical interests and his experiences in serving on committees. He was a member of the University Grants Committee from 1967 to 1972, concerned with government funds for British universities and with the problem of allocating students to universities, taking into account the preferences of both applicant and faculty. He chaired the Computer Board and was, in 1970, an early advocate of extensive computer facilities in universities for which he was criticized by both academics and the press. He had the satisfaction of living to see his ideas accepted.

Barnard had a delightful personality and was very popular. His politics were to the left and he had a reasoned dislike of religions, so that he was never an establishment type of person. The book edited

by S. Geisser et al. [3] is a collection of essays in his honor and contains a complete bibliography of his works up to then. Unfortunately, he never wrote a book.

References

[1] Barnard, G.A. (1949). Statistical Inference, *Journal of the Royal Statistical Society Series B* **11**, 115–149, (with discussion).

[2] Barnard, G.A., Jenkins, G.M. & Winsten, C.B. (1962). Likelihood inference and time series, *Journal of the Royal Statistical Society Series A* **125**, 321–372, (with discussion).

[3] Geisser, S., Hodges, J.S., Press, S.J. & Zellner, A. eds. (1990). *Bayesian and Likelihood Methods in Statistics and Econometrics*. North Holland, Amsterdam.

DENNIS V. LINDLEY

Bartlett, Maurice Stevenson

Born: June 18, 1910, in London, UK.
Died: January 8, 2002, in Exmouth, UK.

Maurice Bartlett was the leading figure amongst British statisticians starting their careers in the 1930s. He published important work at an early age and remained active throughout his life. He made important contributions to statistical inference and methodology and was a pioneer in the development of the theory of stochastic processes. His interest in biological and medical applications led to important work in factor analysis, mathematical epidemiology, and the spatial analysis of field experiments.

Maurice Stevenson Bartlett was born in Chiswick, London. He said that his lifelong interest in probability began in school with the chapter in Hall and Knight's *Algebra*. In 1929, he was awarded a scholarship to Queens' College, Cambridge where he read mathematics. Whilst an undergraduate, he published a paper with John Wishart on **moment** statistics, and he went on to achieve first-class honors with distinction. He enrolled as a graduate student under Wishart, publishing with him a second paper, and received the Rayleigh Prize in 1933. As an aside from statistics, he attended lectures on physics by Eddington and Dirac, thus starting a lifelong interest in theoretical physics.

In 1933, he was appointed Assistant Lecturer in statistics at University College London, under **E.S. Pearson**. This was the center of academic statistics in England at the time: also at University College were **R.A. Fisher**, J.B.S. Haldane, and **J. Neyman**, with all of whom Bartlett would have made contact. But he stayed there only one year, for in 1934, he was appointed to the post of statistician at an agricultural research station of the Imperial Chemical Industries (ICI). This brought him more forcibly into the world of applied statistics, which he relished. His publications during the next four years ranged widely over areas of science well beyond the confines of agriculture and chemistry: the theory of inbreeding, the estimation of intelligence, **factor analysis**, field and laboratory sampling errors, cotton production, and nutrition. Methodological topics included the effect of nonnormality on **Student's *t* distribution**, important results on **sufficiency**, **marginal** and **partial likelihood**, the eponymous **Bartlett's Test** for homogeneity of variance, multiple regression, and **interactions** in **contingency tables**.

In 1938, Bartlett was appointed to a lectureship at Cambridge, in the Faculty of Mathematics rather than that of Agriculture to which statistics had hitherto been attached. In 1940, after the onset of war, he was assigned to an establishment of the Ministry of Supply concerned with the development of rocket batteries, and he divided his time between London and a testing station in Wales. During this period, he worked with other statisticians including F.J. Anscombe, D.G. Kendall, and P.A.P. Moran, and a later meeting with J. Moyal helped to stimulate his growing interest (shared with Kendall) in **stochastic processes**.

Bartlett returned to Cambridge in 1946, concentrating his research on **time series**, **Brownian motion, and diffusion processes**. An important by-product of the latter topic was a 1946 paper on the large-sample theory of **sequential** tests, reproducing many of Wald's results by a different route. He was invited to visit the University of North Carolina at Chapel Hill for four months, where he lectured

on stochastic processes, producing a set of notes, which he used for a repeat of the course on his return to Cambridge and as the basis for his later book. He joined D.G. Kendall and Moyal in a historic three-paper symposium on stochastic processes at the **Royal Statistical Society** in 1949, and in 1955, he published his important book *An Introduction to Stochastic Processes.* This was highly influential in expounding the subject to a wider audience and in making available operational methods for pursuing further research in stochastic processes.

In 1947, Bartlett moved to Manchester to occupy the Chair in Mathematical Statistics, where he enjoyed the company of an outstanding group of mathematicians and developed new courses in the undergraduate and graduate statistics programmes. He became interested in the mathematical modeling of epidemics, especially in the study of measles where the periodicity of measles epidemics could be explained in part by a stochastic model, whereas the corresponding deterministic model predicted a steady endemic state (*see* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**). His book *Stochastic Population Models in Ecology and Epidemiology* appeared in 1960.

In 1960, Bartlett was appointed to the Chair in Statistics at University College London in succession to Egon Pearson. He continued to publish widely on topics including stochastic processes, time series, **multivariate analysis**, spatial analysis, and statistical physics. A book of *Essays in Probability and Statistics* appeared in 1962.

Bartlett's final academic move, in 1967, was to Oxford, where a new Chair of Biomathematics, with a Fellowship at St. Peter's College, had been created as a successor to posts previously held by D.J. Finney and N.T.J. Bailey. Bartlett regarded this somewhat surprising move as a challenge, hoping to stimulate and contribute to the development of mathematical and statistical modeling in biology and medicine. Although the department was in the Biology faculty, Bartlett found the biologists less enthusiastic about his plans than the mathematicians. Nevertheless, he continued his researches in population biology and revived an interest in the spatial models for field experiments studied by Papadakis in 1937. He published *Probability, Statistics and Time* in 1975 and *The Statistical Analysis of Spatial Pattern* in 1976.

During the academic year 1973–1974, and on two occasions after retirement, he visited the Australian National University. Retirement in no way diminished his interest and productivity in a wide range of topics, now including new ones such as **chaos theory**. The volume of selected papers [1] lists 167 papers and 12 letters to *Nature* published by 1989.

As Peter Whittle remarked in an obituary in the *Independent* newspaper, "Bartlett was no self-publicist; both his written and his spoken exposition verged on the terse." Nevertheless, the authority with which he wrote and spoke was self-evident, as was his generosity and goodwill. His younger colleagues and research students received many acts of kindness, and regarded him with great affection.

Bartlett became a Fellow of the Royal Society in 1961 and an Honorary Member of the **International Statistical Institute** in 1980. He served as President of the Manchester Statistical Society (1959–1960), the British Region of the **International Biometric Society** (1964–1966), the Association of Statistics in the Physical Sciences (1965–1967) and the **Royal Statistical Society** (RSS) (1966–1967), and was a Silver and Gold **Guy** Medallist of the RSS. He received the Weldon Prize and Medal of the University of Oxford in 1971 and was elected a Foreign Associate of the US National Academy of Sciences in 1993.

A fuller account of Bartlett's career and a sensitive appreciation of his personality are contained in the obituary notice by Joe Gani [2], who summarizes as follows: "He was profoundly dedicated to research and scholarship, achieving his life's work with exemplary modesty, integrity and humanity."

References

[1] Bartlett, M.S. (1989). *Selected Papers of M.S. Bartlett*. Charles Babbage Research Centre, Winnipeg.

[2] Gani, J. (2002). Professor M.S. Bartlett FRS, 1910–2002, *Statistician* **51**, 399–405.

PETER ARMITAGE

Bartlett's Test

For testing the equality of **variances** of k **normal** populations, Bartlett's modified **likelihood ratio test** statistic is given by T/C, where

$$T = \sum_{i=1}^{k} (n_i - 1) \ln \left[\frac{s_{\mathrm{p}}^2}{s_i^2} \right],$$

$$C = 1 + \frac{1}{3(k-1)} \left\{ \left[\sum_{i=1}^{k} \left(\frac{1}{n_i - 1} \right) \right] - \frac{1}{N-k} \right\},$$

$n_1, \ldots, n_k$ are the sample sizes of the k independent samples, $N = \sum n_i, s_1^2, \ldots, s_k^2$ are the respective unbiased sample variances with denominators $n_i - 1$, and $s_{\mathrm{p}}^2 = (N-k)^{-1} \sum_{i=1}^{k} (n_i - 1) s_i^2$ is the pooled sample variance.

For normal data and even quite small sample sizes (say, $n_i \geq 5$), T/C may be compared with critical values of χ_{k-1}^2, the **chi-square distribution** with $k-1$ **degrees of freedom**. If desired, exact critical values for T/C are available for certain sample sizes (e.g. [7]), or they may be obtained simply by **Monte Carlo** methods. It hardly seems worth the effort, though, since even a slight departure from the normal data assumption will make type I error probabilities (*see* **Level of a Test**) deviate from the nominal level by much more than that caused by using the asymptotic χ_{k-1}^2 critical values in place of exact ones.

Bartlett [1] proposes T/C as a modification of the usual asymptotic form of the normal likelihood ratio statistic $T^* = -2 \ln L$. First he replaces n_i by $n_i - 1$ and puts the **unbiased** s_i^2 in place of the biased form of the sample variances (which arise naturally in normal maximum likelihood estimation). This modification converts T^* to T. Equipped with exact critical values, the resulting test based on T for normal data is unbiased, whereas the test based on T^* is biased. Bartlett then notes that, under the null hypothesis of equal variances, $\mathrm{E}(T) = (k-1)C$ to order $O(n_i^{-3})$ and thus that T/C converges more rapidly to a χ_{k-1}^2 random variable than does T by itself. This type of correction is generally called "Bartlett's correction" and has been the focus of much research.

Bartlett's test has been extended to the multivariate case (see [10, Chapter 8] and [11, Chapter 7]) and to designed experiments [12].

Unfortunately, Bartlett's test is very sensitive to nonnormality. This can be seen by noting that under the null hypothesis of equal variances T/C converges to $(1/2)(\beta_2 - 1)$ times a χ_{k-1}^2 random variable as sample sizes grow large, where β_2 is the fourth moment **kurtosis** coefficient which is equal to 3 for the normal distribution. For example, in [3, Table 1] we find that if $\beta_2 = 5$ and $k = 2$, then the approximate level of a nominal $\alpha = 0.05$ test is $\Pr(2\chi_1^2 > 3.84) = 0.166$. If $k = 5$, then the level jumps to $\Pr(2\chi_4^2 > 9.49) = 0.315$ and gets worse as k increases.

Many investigators feel that this extreme sensitivity to nonnormality is unacceptable. Thus, more robust methods have been proposed including adjustment of critical values for T/C by estimating β_2 [4, 8], **analysis of variance** (ANOVA) on absolute deviations from means [9] and from medians [5], and **bootstrap** estimation of critical values for T/C [2]. Monte Carlo comparisons of procedures are given in Conover et al. [6] and Boos & Brownie [2]. In terms of simplicity and robustness, the ANOVA on absolute deviations from medians is a good alternative to Bartlett's test when normality is in doubt.

References

[1] Bartlett, M.S. (1937). Properties of sufficiency and statistical tests, *Proceedings of the Royal Society of London, Series A* **160**, 268–282.

[2] Boos, D.D. & Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances, *Technometrics* **31**, 69–81.

[3] Box, G.E.P. (1953). Non-normality and tests on variances, *Biometrika* **40**, 318–335.

[4] Box, G.E.P. & Andersen, S.L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption, *Journal of the Royal Statistical Society, Series B* **17**, 1–26.

[5] Brown, M.B. & Forsythe, A.B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances, *Journal of the American Statistical Association* **69**, 364–367.

[6] Conover, W.J., Johnson, M.E. & Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data, *Technometrics* **23**, 351–361.

[7] Glaser, R.E. (1976). Exact critical values for Bartlett's test for homogeneity of variances, *Journal of the American Statistical Association* **69**, 364–367.

[8] Layard, M.W.J. (1973). Robust large-sample tests for homogeneity of variances, *Journal of the American Statistical Association* **71**, 488–490.

[9] Levene, H. (1960). Robust tests for equality of variances, in *Contributions to Probability and Statistics*, I. Olkin, ed. Stanford University Press, Palo Alto.

[10] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

[11] Rencher, A.C. (1995). *Methods of Multivariate Analysis*. Wiley, New York.

[12] Zelen, M. (1959). Factorial experiments in life testing, *Technometrics* **1**, 269–288.

DENNIS D. BOOS

Bartlett's Test

DENNIS D. BOOS

Volume 1, pp. 292–293

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Baseline Adjustment in Longitudinal Studies

The word *baselines*, as used in connection with longitudinal studies, may mean one of three things. First, it can be applied to *demographic characteristics* of the patient, which are either unchanging (such as sex), or which change slowly (such as height) or at the same rate (such as age), for all subjects during the course of the study. Secondly, it can mean *true baselines*: measurements taken at some earlier moment (perhaps prior to treatment) of the same variable which is to be used as a measure of outcome. Thirdly, it can indicate *baseline correlates*: measurements taken at an earlier moment than that which will be used to judge the outcome of the study, but not on the same variable (although correlated with it), but which may vary strongly during the course of the study.

At first sight, these three types of baseline seem to be quite different, but from the point of view of many powerful approaches to analyzing data (for example **analysis of covariance**), there is no essential distinction between them, and the common tendency to regard the second sort as being capable of a special use for which the others are not available, i.e. to judge the ability of treatment to effect a change, is false [17]. Nevertheless, except where otherwise specified, it will be implicitly assumed below that the second sort of baseline is being referred to.

The Randomized Trial

Analysis of Covariance

Consider the case of a randomized parallel group **clinical trial** in which outcome measures at the end of the trial are available for both treatment and control groups and baseline measurements prior to treatment have been taken also. Suppose that measurements on patients in the treatment group are X_{ti} (baselines) and Y_{ti} (outcomes), $i = 1$ to n_t. Similarly, let the corresponding measurements in the control group be X_{ci} and Y_{ci}, $i = 1$ to n_c, and define

$$q = \left(\frac{1}{n_t} + \frac{1}{n_c}\right).$$

Suppose that the variance–**covariance matrix** in the two groups is identical and equal to

$$\mathbf{\Sigma}_{x,y} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

Sometimes the further assumption is made that $\sigma_x = \sigma_y$. Although this may frequently be approximately true, there is no particular reason why this assumption should be valid in general, especially since patients are often selected for entry into clinical trials on the basis of baseline measurements.

If the baselines are ignored altogether, then the effect of treatment may be estimated using raw outcomes only as $\hat{\tau}_{raw} = \overline{Y}_t - \overline{Y}_c$ with variance $\text{var}(\hat{\tau}_{raw}) = q\sigma_y^2$. A popular alternative way to analyze such a trial is to construct so-called *change scores*, $Z_{ti} = Y_{ti} - X_{ti}$ and $Z_{ci} = Y_{ci} - X_{ci}$ and define the treatment estimator $\hat{\tau}_{change} = \overline{Z}_t - \overline{Z}_c$ with $\text{var}(\hat{\tau}_{change}) = q\sigma_z^2 = q(\sigma_y^2 + \sigma_x^2 - 2\rho\sigma_x\sigma_y)$. An alternative, but equivalent representation of the change score estimator is as $\hat{\tau}_{change} = (\overline{Y}_t - \overline{Y}_c) - (\overline{X}_t - \overline{X}_c)$. A more general estimator of the treatment effect is given by $\hat{\tau}_\beta = (\overline{Y}_t - \overline{Y}_c) - \beta(\overline{X}_t - \overline{X}_c)$, with variance which may be written as

$$\text{var}(\hat{\tau}_\beta) = q[(\beta\sigma_x - \rho\sigma_y)^2 + (1 - \rho^2)\sigma_y^2]. \qquad (1)$$

The covariance of this estimator with the difference at baseline is

$$\text{cov}_{\hat{\tau},\text{base}} = q(\rho\sigma_x\sigma_y - \beta\sigma_x^2). \qquad (2)$$

The raw outcomes and change-score estimator can be regarded as special forms of the general estimator with β equal to 0 and 1, respectively.

Thus, these three treatment estimators can be seen merely as ways of adjusting the observed difference at outcome using the baseline difference. The first simply relies on the fact that in a randomized trial, in the absence of a difference between treatments, the expected value of the mean difference at outcome is zero. Hence, the factor by which the observed outcome needs correcting, in order to judge the treatment effect, is also zero. The second corresponds to the assumption that the difference at outcome, in the absence of a treatment effect, is expected to be the difference in the means of the baselines. This naive assumption is, in fact, false [2, 6, 15, 18]. The third allows for a more general system of predicting what the outcome difference would have been in the

absence of any treatment effect as a function of the mean difference at baseline.

Analysis of covariance corresponds to choosing a suitable value for β. There are two standard motivations that lead to an equivalent result. The first finds the value of β that minimizes (1) and the second finds the value of β that sets (2) to zero. By inspection of (1) and (2), it is obvious that the value of β that does this is $\beta' = \rho\sigma_y/\sigma_x$ and that the resulting variance is then

$$(1 - \rho^2)\sigma_y^2. \tag{3}$$

β' is the regression of Y on X and this adjustment is known as *analysis of covariance*. It is quite general and can be used for baseline variables of the first and third type discussed above and not just of the second. In practice the **nuisance parameters** ρ, σ_x, and σ_y are unknown and the regression coefficient, β', has to be estimated and the (slight) nonorthogonality observed in nearly all trials means that the expected variance of the treatment estimate is slightly higher than given by (3).

These considerations generalize readily to the case where we have multiple outcomes and baselines. A simple and attractive computational solution is to use appropriate summary measures based on the outcomes (*see* **Summary Measures Analysis of Longitudinal Data**) and adjust these using either multiple baselines or some summary of them in an analysis of covariance. Some good practical advice has been given in [10] and also in [8].

Red Herrings

It is important to note that since analysis of covariance produces both the minimum variance estimator and the estimator which is uncorrelated with the baseline difference, other approaches are deficient in this respect. For example, using the change score alone does not, in fact, deal with the problem of chance imbalance in the baselines, since the change score is correlated with the baselines. Indeed, if the correlation is very low, then it may even produce a higher variance than using the raw outcomes alone.

It is also sometimes argued that clinical relevance may decide the toss between these approaches, but this is just nonsense. All these approaches measure the same thing, and indeed, for a trial in which baseline values are perfectly balanced, give exactly the same answer. Furthermore, because it has the smaller variance, a covariance adjusted estimator from a given trial would actually be expected to predict the change score estimator in a subsequent trial better than the change score estimator itself! Note also, as has been nicely shown by Laird, that adjusting the change scores using analysis of covariance gives exactly the same answer as adjusting the raw scores [9].

Errors in variables [4] also do not affect the validity of analysis of covariance, which can simply be seen as a means of incorporating what is known into the analysis [14, 15]. For example, where no baselines have been measured, using raw outcomes is valid because there is no baseline information to condition on, a fact which will then properly be reflected in the high variance of the treatment estimate. Where baselines have been observed, these observations can be used to decrease the variance of the treatment estimate and to adjust for any chance imbalance. But any chance imbalance observed is a chance imbalance in the observations, and this is what must be adjusted for.

A final point of some confusion is whether one should be interested in estimating a trend or an average. Where there is a single outcome measure, there is essentially no distinction, because, as has already been explained, adjusting the trend measure $Y - X$ for X gives exactly the same result as adjusting the outcome measure Y for X [9]. Where a series of outcome measures is made at different time points, however, a different treatment estimate is possible at each of these time points. To the extent that such separate estimates are made, there is again no distinction between the two approaches. However, a series of such independent estimates may miss the opportunity to make a powerful summary, and if it is decided therefore to do this, inevitably the choice of a suitable summary arises. Appropriate choice depends on the way in which the treatment effect grows over time. For example, regular therapy with a beta-agonist in asthma is likely to produce a near constant bronchodilation over time and hence a mean outcome measure is suitable. However, hormone replacement therapy in osteoporosis may have an effect on bone mineral density which increases with time so that some sort of slope estimate is appropriate. If the choice can be made on the basis of suitable prior information, then it avoids the difficulties to which data-dependent choices are liable. Contrary to what

is sometimes claimed, however, the choice does not depend directly on the way in which outcomes change for individual subjects, since these may be affected by strong (but irrelevant for causal purposes) secular trends.

Testing for Baseline Balance

A rather foolish but common use which is made of baselines is to check that the groups in a randomized clinical trial are "balanced" [1, 3, 13, 16]. A significant result, however, can only mean either that a type I error has been committed or that the **randomization** mechanism itself is flawed. However, where a significant result is found, trialists are generally most reluctant to impugn the conduct of the trial itself and instead treat the result as being the result of a randomly bad allocation. However, **hypothesis tests**, if they are to be used at all, should be used to test hypotheses, not to describe samples. For these and other reasons, whereas such tests of baseline balance might form a legitimate part of data monitoring and quality control, they should not form part of a general strategy of analyzing clinical trials and, in particular, should not be used to decide whether or not to use analysis of covariance [16].

Adjusting for Baselines Taken After the Start of Treatment

Adjusting for late baselines, whether by analysis of covariance or by simple change scores, can be extremely misleading since the treatment effect will be adjusted by a "baseline" difference which may itself reflect the effect of treatment, thus attenuating the final estimate. In the context of survival analysis this controversial topic goes by the name of **time-dependent covariates**. Wherever such adjustments are made, one should be extremely careful in interpreting the result.

Cut-off Designs

There has been some interest in designs that use observed baseline measurements to allocate patients to treatment [7, 19]. For example, the more severely ill get the experimental treatment, whereas those who are less ill get the standard treatment. In such trials some form of adjustment for the baselines is then mandatory. Quite apart from the stronger assumptions required to analyze such trials, the strong degree of imbalance leads to a considerable inflation of variance compared with the randomized design. Such trials have been described in [19] and [7] and some criticisms are given in [17]. The fact that the baselines used for assignment may be measured with error is also not a problem for such trials [11].

Cohort Studies

Baseline measurements can be extremely important in the context of epidemiological cohort studies which, however, have several problems not shared by the controlled clinical trial. From one point of view it may be supposed that analysis of covariance is even more important, since the simple comparison of mean outcome scores between an exposure and a control group is not likely to have an expected value equal to the exposure effect of interest. Therefore some sort of baseline adjustment is required. However, this can cause problems.

First, the errors in variable problem is potentially serious here in the sense that the assignment mechanism of subjects to exposure is not ignorable in Rubin's sense [12, 4] (*see* **Nonignorable Dropout in Longitudinal Studies**). However, if assignment probabilities are related to true covariate values, then adjusting for observed covariate values will not deal with the problem adequately. (This is not to say that it may not be preferable to some of the alternative forms of adjustment used. However, the point remains controversial [5, 14].) Secondly, the definition of exposure itself may be unclear in a way that makes adjustment for baselines problematic. Consider the case of salt and hypertension and a cohort study that measures salt consumption (presumably with some difficulty) and blood pressure in a cohort of individuals from a given time point. Whether and how we should adjust for baseline blood pressure is at least partly bound up with what we are trying to measure: lifetime exposure in terms of salt consumption, which will already have had an effect on the baseline measurements, or "downstream" consumption. Perhaps for these reasons, but no doubt also because computational progress has made this a much more feasible option, there is an increasing interest in complex modeling in this field.

Uncontrolled Studies

Simply comparing outcome with baseline by forming the mean difference for a single treatment group in an uncontrolled study is an inadequate way to assess the effect of treatment. Quite apart from the usual trend biases to which such studies are liable, if the subjects have been selected on the basis of extreme baseline measurements, then the study will be subject to **regression to the mean** [2, 6]. The price one then has to pay to get a reasonable estimate of the effect of treatment is complex statistical modeling with attendant doubt and skepticism as to the result [18].

References

[1] Altman, D.G. (1985). Comparability of randomised groups, *Statistician* **34**, 125–136.

[2] Altman, D. & Bland, M. (1994). Regression towards the mean, *British Medical Journal* **308**, 1499.

[3] Canner, P.L. (1991). Covariate adjustment of treatment effects in clinical trials, *Controlled Clinical Trials* **12**, 359–366.

[4] Carroll, R.J. (1989). Covariance analysis in generalized linear measurement error models, *Statistics in Medicine* **8**, 1075–1093.

[5] Carroll, R.J. (1990). Author's reply, *Statistics in Medicine* **9**, 585–586.

[6] Chuang-Stein, C. (1993). The regression fallacy, *Drug Information Journal* **27**, 1213–1220.

[7] Finkelstein, M.O., Levin, B. & Robbins, H. (1996). Clinical and prophylactic trials with assured new treatment for those at greater risk. I and II, *American Journal of Public Health* **86**, 691–705.

[8] Frison, L. & Pocock, S. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design, *Statistics in Medicine* **11**, 1685–1704.

[9] Laird, N. (1983). Further comparative analyses of pretest post-test research design, *American Statistician* **37**, 329–330.

[10] Laird, N. & Wang, F. (1990). Estimating rates of change in randomized clinical trials, *Controlled Clinical Trials* **11**, 405–419.

[11] Reichardt, C.S., Trochim, W.M.K. & Cappelleri, J.C. (1995). Reports of the death of regression discontinuity analysis are greatly exaggerated, *Evaluation Review* **19**, 39–63.

[12] Rubin, D. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.

[13] Senn, S.J. (1989). Covariate imbalance and random allocation in clinical trials, *Statistics in Medicine* **8**, 467–475.

[14] Senn, S.J. (1990). Covariance analysis in generalized linear measurement error models, *Statistics in Medicine* **9**, 583–585.

[15] Senn, S.J. (1994). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design, *Statistics in Medicine* **13**, 197–198.

[16] Senn, S.J. (1994). Testing for baseline balance in clinical trials, *Statistics in Medicine* **13**, 1715–1726.

[17] Senn, S.J. (1995). A personal view of some controversies in allocating treatment to patients in clinical trials, *Statistics in Medicine* **14**, 2661–2674.

[18] Senn, S.J. & Brown, R.A. (1989). Maximum likelihood estimation of treatment effects for samples subject to regression to the mean. *Communications in Statistics – Theory and Methods* **18**, 3389–3406.

[19] Trochim, W.M.K. & Cappelleri, J.C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials, *Controlled Clinical Trials* **13**, 190–212.

(*See also* **Longitudinal Data Analysis, Overview**; **Generalized Linear Models for Longitudinal Data**)

STEPHEN SENN

Battery Reduction

Often a researcher is in the position where he has n variables of interest under investigation, but desires to reduce the number for analysis or later data collection. Specifically, a researcher may desire to select a subset of m variables from the original n variables that reproduce as much as possible of the information contained in the original n variables. In other words, he may desire to find the subset of m variables which accounts for a large proportion of the variance of the original n variables. For example, if he has a long **questionnaire** measuring the effect of a given treatment on the day-to-day activities of a certain population of patients, there may be concern about the burden such a questionnaire places upon the patient. So there is a need to try to reduce the size of the questionnaire (or reduce the battery of questions) without substantially reducing the information obtained from the full questionnaire. To accomplish this he can perform battery reduction using the data collected from patients who completed the full battery of questions at some time in the past.

There are a number of procedures for performing battery reduction. In the following, we illustrate the concept using **Gram–Schmidt** transformations. Cureton & D'Agostino [1, Chapter 12] contains complete details of this procedure. Also, D'Agostino et al. [2] have developed a macro in SAS (*see* **Software, Biostatistical**) which carries out this procedure which is available from `ralph@math.bu.edu`.

Assume that the n variables on which we would like to perform battery reduction are denoted $X_1, \ldots, X_n$. Assume also that these n variables are standardized with mean zero and variance unity. Then the total variance explained by $X_1, \ldots, X_n$, is n, the number of variables. To find the subset of m variables which will explain as much as possible the variance of $X_1, \ldots, X_n$, we first perform a **principal components analysis** and decide upon the m components to be retained. These are the components that account for the salient variance in the original data set. The SAS [3] procedure PRINCOMP can be used to perform principal components analysis. The SAS [3] procedure FACTOR can also be employed. Both procedures automatically standardize the variables before employing principal components. Note also that the above-mentioned battery reduction macro created by D'Agostino et al. [2] automatically standardizes the variables and creates these components as part of its battery reduction.

Once m is determined, let $\mathbf{A}$ denote the $n \times m$ matrix in which the columns contain the **correlations** of $X_i, i = 1, \ldots, n$, to the m principal components. Symbolically $\mathbf{A}$ is

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix}.$$

The jth column contains the correlations of the original variables X_i to the jth component, and the sum of the squares of all the a_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, of $\mathbf{A}$ equals the amount of the total variance of the original n variables that is explained by the m retained components. We refer to this as *salient variance*. In principal components analysis, $\mathbf{A}$ is referred to as the *initial component matrix*. It is also often referred to as the *initial factor matrix*. The elements of $\mathbf{A}$ are called the *loadings*. The sum of the squares of the loadings of the ith row of $\mathbf{A}$ equals the proportion of variance of $X_i, i = 1, \ldots, n$, explained by the m principal components. This is called the **communality** of X_i, symbolized as h_i^2.

Now, to find the subset of m variables which explains, as much as possible, the salient variance of the original n variables, we can employ the Gram–Schmidt **orthogonal rotations** to the $n \times m$ initial component matrix $\mathbf{A}$. The goal of the Gram–Schmidt rotation in battery reduction is to rotate $\mathbf{A}$ into a new $n \times m$ component matrix, where the variable accounting for the largest proportion of the salient variance (call this "variable 1") has a nonzero loading on the first component, but zero loadings on the remaining $m - 1$ components; the variable accounting for the largest proportion of residual variance ("variable 2"), where residual variance is the portion of the salient variance which is not accounted for by the variable 1, has a nonzero loading on the first two components, but zero loadings on the remaining $m - 2$ components; the variable accounting for the largest proportion of second-residual variance ("variable 3") has a nonzero loading on the first three components, but zero loadings on the remaining $m - 3$ components, etc. until the variable accounting for the largest proportion of the $(m - 1)$th residual

variance ("variable m") is found. Variables 1 through m are then the variables which reproduce, as much as possible, the variance retained by the m principal components, and so also the salient variance contained in the original n variables. In the vocabulary of principal components analysis, variable 1 is the first *transformed component*, variable 2 is the second, etc. To determine how much of the original variance of all n variables is explained by the m transformed components, we simply compute the sum of squares of all the loadings in the final $n \times m$ Gram–Schmidt rotated matrix (this should be close to the sum of squares of the elements of the $n \times m$ initial component matrix **A**). The following example will illustrate the use of the Gram–Schmidt process in battery reduction.

In the **Framingham** Heart Study, a 10-question depression scale was administered (so $n = 10$), where the responses were No or Yes to the following (the corresponding name to which each question will hereafter be referred is enclosed in parentheses):

1. I felt everything I did was an effort (EFFORT).
2. My sleep was restless (RESTLESS).
3. I felt depressed (DEPRESS).
4. I was happy (HAPPY).
5. I felt lonely (LONELY).
6. People were unfriendly (UNFRIEND).
7. I enjoyed life (ENJOYLIF).
8. I felt sad (FELTSAD).
9. I felt that people disliked me (DISLIKED).
10. I could not get going (GETGOING).

A Yes was scored as 1 and No as 0 except for questions 4 and 7, where this scoring was reversed so that a score of 1 would indicate depression for all questions.

After performing a principal components analysis on this data, there were three components with variances greater than unity. The variances of these three components were 3.357, 1.290, and 1.022 for a percentage variance explained equal to $100 \times (3.357 + 1.290 + 1.022)/10 = 56.69\%$. Thus, using the Kaiser rule for selecting the number of retained components [1], we set m equal to 3 for this example. The 10×3 initial component matrix **A** is in Table 1.

Now, to use Gram–Schmidt transformations to determine the three *variables* which explain the largest portion of the salient variance from the original 10 variables, we do the following:

Table 1 Initial component matrix **A** for Framingham Heart Study depression questionnaire

	a_1	a_2	a_3	h^2
EFFORT	0.60	0.15	0.41	0.55
RESTLESS	0.39	0.07	0.55	0.46
DEPRESS	0.77	−0.13	−0.10	0.62
HAPPY	0.70	−0.23	−0.06	0.55
LONELY	0.64	−0.23	−0.21	0.51
UNFRIEND	0.35	0.68	−0.33	0.69
ENJOYLIF	0.52	−0.27	−0.27	0.42
FELTSAD	0.71	−0.22	−0.20	0.59
DISLIKED	0.34	0.72	−0.22	0.68
GETGOING	0.58	0.20	0.47	0.60

Note: $h^2 = a_1^2 + a_2^2 + a_3^2$ is the communality.

1. Find, from **A** in Table 1, the variable which explains the largest proportion of salient variance from the original 10 variables. This is the variable UNFRIEND, with a sum of squares of loadings (communality) across the three components equal to $0.35^2 + 0.68^2 + (-0.33)^2 = 0.69$.
2. Take the loadings of UNFRIEND from Table 1 (0.35, 0.68, −0.33) and normalize them (i.e. divide each element by the square root of the sum of the squares of all three elements). This yields the normalized loadings: 0.42, 0.82, −0.40.
3. Create a 3×3 ($m \times m$) matrix $\mathbf{Y}_1$, which, in the Gram–Schmidt process, is given by

$$\mathbf{Y}_1 = \begin{bmatrix} a & b & c \\ k_2 & -ab/k_2 & -ac/k_2 \\ 0 & c/k_2 & -b/k_2 \end{bmatrix},$$

where $a = 0.42$, $b = 0.82$, $c = -0.40$ (the normalized row of UNFRIEND from **A**), and $k_2 = (1 - a^2)^{1/2}$. Thus,

$$\mathbf{Y}_1 = \begin{bmatrix} 0.42 & 0.82 & -0.40 \\ 0.91 & -0.38 & 0.18 \\ 0 & -0.44 & -0.90 \end{bmatrix}.$$

4. Calculate $\mathbf{A}\mathbf{Y}_1'$, which is shown in Table 2. Note that, for UNFRIEND, the only nonzero loading is on the first component (or first column). This loading is equal to the square root of the sum of squares of the original loadings of UNFRIEND in matrix **A** (thus, no

Table 2 $\mathbf{B} = \mathbf{A}\mathbf{Y}_1'$

	b_1	b_2	b_3	res. h^2
EFFORT	0.21	0.56	−0.44	0.51
RESTLESS	0.00	0.43	−0.53	0.47
DEPRESS	0.26	0.73	0.15	0.56
HAPPY	0.13	0.71	0.15	0.52
LONELY	0.16	0.63	0.29	0.48
UNFRIEND	0.84	0.00	0.00	0.00
ENJOYLIF	0.11	0.53	0.36	0.41
FELTSAD	0.20	0.69	0.28	0.55
DISLIKED	0.82	0.00	−0.12	0.01
GETGOING	0.22	0.54	−0.51	0.55

Note: res. h^2 = residual communality $= b_2^2 + b_3^2$.

Table 3 Final rotated reduced component matrix, **C**

	c_1	c_2	c_3	h^2
EFFORT	0.21	0.46	0.54	0.55
RESTLESS	0.00	0.31	0.61	0.46
DEPRESS	0.26	0.75	0.00	0.63
HAPPY	0.13	0.73	0.00	0.55
LONELY	0.16	0.68	−0.16	0.51
UNFRIEND	0.84	0.00	0.00	0.70
ENJOYLIF	0.11	0.59	−0.25	0.42
FELTSAD	0.20	0.73	−0.14	0.59
DISLIKED	0.82	−0.02	0.12	0.67
GETGOING	0.22	0.43	0.61	0.60

Note: $h^2 = c_1^2 + c_2^2 + c_3^2$ is the final communality.

"information" explained by UNFRIEND is lost during the rotation process). For each of the remaining variables in Table 2, we have the following: (i) the squares of the elements in the first column are the portions of the variances of these variables which are accounted for by UNFRIEND; and (ii) the sum of the squares of the elements in the second and third columns is the residual variance (i.e. the variance of the variables not accounted for by UNFRIEND).

5. Find the variable which explains the largest proportion of residual variance (i.e. has the largest residual communality). This is the variable DEPRESS, with a sum of squares of loadings across the last two columns of Table 2 which is equal to $0.73^2 + 0.15^2 = 0.56$.
6. Take the loadings of DEPRESS from Table 2 (0.73, 0.15) and normalize them. This yields the normalized loadings: 0.98, 0.20.
7. Create a 2×2 matrix $\mathbf{Y}_2$, which, in the Gram–Schmidt process, is given by

$$\mathbf{Y}_2 = \begin{bmatrix} b & c \\ c & -b \end{bmatrix},$$

where $b = 0.98$, $c = 0.20$ (the normalized row of DEPRESS from the last two columns of Table 2). Thus,

$$\mathbf{Y}_2 = \begin{bmatrix} 0.98 & 0.20 \\ 0.20 & -0.98 \end{bmatrix}.$$

8. Postmultiply the last two columns of $\mathbf{A}\mathbf{Y}_1'$ by $\mathbf{Y}_2'$; the result is shown in the last two columns of Table 3. The first column of Table 3 is the first column of $\mathbf{A}\mathbf{Y}_1'$. Together, the three columns are called the rotated reduced component matrix (matrix **C** of Table 3).
   Note that, for DEPRESS, the loading on the last component (or last column) is zero. The sum of squares of the loadings (the final communality) of DEPRESS in Table 3 is, within rounding error, equal to the square root of the sum of squares of the loadings of DEPRESS in the initial component matrix **A** (0.63 vs. 0.62; thus, no "information" explained by DEPRESS is lost during the rotation process). For the remaining variables in the second column of Table 3, the elements are the portions of the variances of these variables which are accounted for by DEPRESS.
9. The last of the three variables which explains the largest portion of variance in the original 10 variables is GETGOING, since its loading is largest in the last column of Table 3.
10. The sum of squares of all the loadings in Table 3 is approximately equal, within rounding error, to the sum of squares of loadings in **A**.

Thus the three variables UNFRIEND, DEPRESS, and GETGOING alone retain approximately the same variance that was retained by the first three principal components (which involved all 10 original variables). We have reduced the original battery of ten questions to three.

The above is presented only as an illustration. It is unlikely that a researcher would need to perform a battery reduction on 10 simple items such as in the example. However, there could be a tremendous gain if the original n was, say, 100 and the number of retained components m was only 10.

Also, the above example focused on finding the m variables that reproduce the variance retained by the principal components. There may be variables with low communalities (thus not related to the other variables). The researcher may want to retain these also. For a discussion of this and presentations of other battery reduction methods, see Cureton & D'Agostino [1, Chapter 12].

References

[1] Cureton, O. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[2] D'Agostino, R.B., Dukes, K.A., Massaro, J.M. & Zhang, Z. (1992). in *Proceedings of the Fifth Annual Northeast SAS Users Group Conference*, pp. 464–474.

[3] SAS Institute, Inc. (1990). *SAS/STAT User's Guide, Release 6.04*, 4th Ed. SAS Institute, Inc., Cary.

(*See also* **Cluster Analysis, Variables**; **Psychometrics, Overview**)

JOSEPH M. MASSARO

Bayes Factors

The Bayes factor provides a number for quantifying the evidence in favor of a scientific theory, and hence provides a **Bayesian** approach to **hypothesis testing** or ascertainment. Its terminology is apparently due to Good [12], but the underlying philosophy and proposed usage has been described by Jeffreys [13, 14]. For the comparison of two competing hypotheses, the Bayes factor is the posterior **odds** in favor of one of the hypotheses when the prior odds of the hypotheses are equal. The Bayes factor is more precisely defined as follows.

Definition

Let H_0 and H_1 denote two competing hypotheses under which data D are thought to have arisen. The **prior** probabilities for H_0 and H_1 are given by $P(H_0)$ and $P(H_1) = 1 - P(H_0)$, and the posterior probabilities by $P(H_0|D)$ and $P(H_1|D) = 1 - P(H_0|D)$, respectively. By **Bayes's theorem**, the latter probabilities can be expressed as

$$P(H_k|D) = \frac{P(D|H_k)P(H_k)}{P(D|H_0)P(H_0) + P(D|H_1)P(H_1)}, \tag{1}$$

for $k = 0, 1$. Using (1), the posterior odds ratio in favor of H_0 can be expressed as

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)} \times \frac{P(H_0)}{P(H_1)}. \tag{2}$$

From (2), one can see that the prior odds get transformed into the posterior odds via multiplication by the factor

$$BF_{01} = \frac{P(D|H_0)}{P(D|H_1)}, \tag{3}$$

which is termed the Bayes factor of H_0 to H_1. Note that by (2), the Bayes factor is the ratio of posterior to prior odds, regardless of the value of prior odds, and also the posterior odds when the prior odds of H_0 and H_1 are equal. Kass and Raftery [15] summarize Jeffreys' guidelines for interpreting the Bayes factor, which suggest interpretation based on half-units on the log base 10 scale [14]. On this scale, a log base 10 BF_{01} value greater than 2, or raw BF_{01} value greater than 100, represents decisive evidence in favor of H_0, a value between 1 and 2, strong evidence, a value between 1/2 and 1, substantial evidence, and a value between 0 and 1/2, evidence not worth more than a bare mention.

In the simple case, H_0 and H_1 entail no free parameters to be estimated, and B_{01} is the **likelihood ratio** statistic. In the more common scenario, H_0 and H_1 hypothesize a model $p(D|\theta_k, H_k)$ for the data D with unknown parameters θ_k, and prior distributions $\pi(\theta_k|H_k)$ for θ_k, for $k = 0, 1$, respectively. In this case, the densities $P(D|H_k)$ appearing in the numerator and denominator of BF_{01} are the integrated **likelihoods** against the prior densities:

$$P(D|H_k) = \int_{\theta_k} p(D|\theta_k, H_k)\pi(\theta_k|H_k)\,\mathrm{d}\theta_k, \tag{4}$$

for $k = 0, 1$. As seen by (4) two challenges for implementation of Bayes factors are the calculation of the integrals, which may involve highly peaked integrands over high dimensional spaces, and the specification of prior distributions, for which the Bayes factor may be sensitive.

Calculation

There has been considerable research in the calculation of integrals (4) of the type required by the Bayes factor. Exact analytic evaluation of the integrals is possible for a restricted set of scenarios, including **exponential family** distributions with conjugate priors; see for example, [6, Chapter 9]. Otherwise, some type of numerical approximation is needed. Many standard numerical methods, such as numerical quadrature (*see* **Numerical Integration**), however, can be very inefficient for these types of integrals since the dimension may be high and/or the integrand highly peaked near the **maximum likelihood** estimate. Evans and Swartz [8, 9] provide a review of numerical **algorithms** for integrals appearing in the Bayes factor.

Laplace's asymptotic method provides a surprisingly accurate approximation to integrals of the form (4) [16, 22, 23]. One version of the approximation, termed fully exponential, is obtained by assuming that the posterior density, which is proportional to the integrand in (4), becomes highly peaked at the posterior mode $\tilde{\theta}$ as the sample size n increases. Here the subscript k has been dropped for brevity. Denote by $\tilde{\ell}(\theta)$, the log of the integrand in (4).

Laplace's approximation works by expanding $\tilde{\ell}(\theta)$ in a Taylor series about $\tilde{\theta}$. Exponentiating this Taylor series expansion and ignoring some of the latter terms yields an integrand that resembles a **multivariate normal distribution** with mean $\tilde{\theta}$ and variance–**covariance matrix** $\tilde{\Sigma} = [-D^2\tilde{\ell}(\tilde{\theta})]^{-1}$, where $D^2\tilde{\ell}(\theta)$ denotes the Hessian matrix of second derivatives of $\tilde{\ell}(\theta)$. Integrating this integrand yields the following approximation:

$$P(\hat{D}|H) = (2\pi)^{d/2}|\tilde{\Sigma}|^{1/2}P(D|\tilde{\theta}, H)\pi(\tilde{\theta}|H), \quad (5)$$

where d is the dimension of θ. Under the Fisher regularity conditions, given in Kass et al. [16], the relative error of the approximation is $O(n^{-1})$. Thus, when the approximation is applied to both the numerator and denominator of BF_{01}, the relative error of the resulting approximation is also $O(n^{-1})$. A variation of the approximation replaces $\tilde{\theta}$ with the maximum likelihood estimator $\hat{\theta}$ of $P(D|\theta, H)$, and obtains relative error of order $O(n^{-1/2})$. Using the latter approximation, Schwarz [21] derived a rough approximation to the Bayes factor, with error term on the log scale of order $O(1)$, as

$$S = \log P(D|\hat{\theta}_0, H_0) - \log P(D|\hat{\theta}_1, H_1) - \tfrac{1}{2}(d_0 - d_1)\log(n), \quad (6)$$

where d_0 and d_1 are the dimensions of the parameter spaces under H_0 and H_1, respectively. The expression in (6) is a modified log likelihood ratio statistic, which is interpretable even for nonnested models (*see* **Separate Families of Hypotheses**). Kass and Wasserman [17] show that for the case where H_0 is nested in H_1 and with an interesting choice of unit-information prior for the parameters tested, S is an approximation to the log of the Bayes factor to order $O(n^{-1/2})$. Minus twice the Schwarz criterion is often called the Bayesian Information Criterion (BIC), which is similar to **Akaike's Information Criterion** (AIC) [1]. When applied to each model separately, the BIC penalizes models of higher dimensionality more than the AIC; the penalty factor subtracted from the maximized log likelihood for the AIC is proportional to the number of parameters d, whereas for the BIC, it is proportional to $d \log n$.

The increased feasibility of **Markov Chain Monte Carlo** (MCMC) methods [11] for simulating samples from the posterior distribution of parameters has spawned a large variety of new **simulation** and simulation/asymptotic approximation hybrid approaches for calculating the Bayes factor; see, for example, [4, 5, 7, 10, 18, 20, 24]. The choice of method that obtains the highest accuracy depends on features of the individual problem.

Prior Distributions

Prior distributions play a key role in the Bayes factor. First, unlike in Bayesian estimation problems, where as the sample size approaches infinity, the influence of the prior distribution diminishes, the Bayes factor remains sensitive to the choice of prior, even asymptotically. This can be inferred somewhat from approximation (5). Therefore, careful consideration must be given to the selection of the prior distribution and **sensitivity analyses** to assess robustness of conclusions to a range of plausible priors must be performed. In the case of testing nested hypotheses, say where a parameter θ_0 is fixed under H_0 but unknown under H_1, the Bayes factor in favor of H_0 often increases as the prior variance for θ_0 under H_1 increases. This phenomenon, called *Bartlett's paradox* [2], emphasizes that proper informative priors should be used for the parameters under test.

Improper priors pose problems for Bayes factors in that either the integrals comprising the numerator and denominator may not converge, or that the arbitrary proportionality constant of improper priors leads to arbitrariness in definition of the Bayes factor. In the case of testing nested hypotheses, however, many authors are not bothered by the use of improper reference priors for the **nuisance parameters** [14, 17]. The intrinsic Bayes factor of Berger and Pericchi [3] and the fractional Bayes factor of O'Hagan [19] are two proposals for modifying Bayes factors to accept reference, possibly improper, prior distributions, by utilizing part of the data as a training sample for the prior.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, B.N. Petrox & F. Caski, eds. Akademiai Kiado, Budapest, p. 267.

[2] Bartlett, M.S. (1957). Comment on "A Statistical Paradox" by D.V. Lindley, *Biometrika* **44**, 533–534.

[3] Berger, J.O. & Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association* **91**, 109–122.

[4] Carlin, B. & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo, *Journal of the Royal Statistical Society, Series B* **57**, 473–484.

[5] Chib, S. (1995). Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association* **90**, 1313–1321.

[6] DeGroot, M.H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

[7] DiCiccio, T.J., Kass, R.E., Raftery, A. & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations, *Journal of the American Statistical Association* **92**, 903–915.

[8] Evans, M. & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems (Disc:V11 P54-64), *Statistical Science* **10**, 254–272.

[9] Evans, M. & Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, Oxford.

[10] Gelfand, A.E. & Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

[11] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[12] Good, I.J. (1958). Significance tests in parallel and in series, *Journal of the American Statistical Association* **53**, 799–813.

[13] Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability, *Proceedings of the Cambridge Philosophical Society* **31**, 203–222.

[14] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Oxford University Press, Oxford.

[15] Kass, R.E. & Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**, 773–795.

[16] Kass, R.E., Tierney, L. & Kadane, J.B. (1990). The validity of posterior asymptotic expansions based on Laplace's method, in *Bayesian and Likelihood Methods in Statistics and Econometrics*, S. Geisser, J.S. Hodges, S.J. Press & A. Zellner, eds. North-Holland, New York, pp. 473–488.

[17] Kass, R.E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *Journal of the American Statistical Association* **90**, 928–934.

[18] Newton, M.A. & Raftery, A.E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion), *Journal of the Royal Statistical Society Series B* **56**, 3–48.

[19] O'Hagan, A. (1995). Fractional Bayes factors for model comparison (Disc:P118-138), *Journal of the Royal Statistical Society, Series B* **57**, 99–118.

[20] Raftery, A.E. (1995). Hypothesis testing and model selection via posterior simulation, in *Practical Markov Chain Monte Carlo*, W. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 163–188.

[21] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**, 461–464.

[22] Tierney, L. & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**, 82–86.

[23] Tierney, L., Kass, R.E. & Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *Journal of the American Statistical Association* **81**, 82–86.

[24] Verdinelli, I. & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio, *Journal of the American Statistical Association* **90**, 614–618.

DONNA K. PAULER

Bayes' Theorem

The uncertainty, expressed through probability, that you feel about something will depend on your knowledge at the time you state that probability and could change as additional information becomes available. For example, the probability that a woman has breast cancer will depend on your knowing that she is an apparently healthy woman of 40 years of age with three children and no family history of breast cancer. It would change if she tested positive on a screening test for the condition. Bayes' theorem describes how this change, as a result of extra information, should be evaluated.

Let C be the uncertain event under consideration – in the example, the event that the woman has breast cancer. Let $\Pr(C)$ be the probability that C obtains, based on your knowledge at the time the probability is evaluated. Let $+$ denote the additional knowledge – in the example, the positive test result. The revised probability is written $\Pr(C|+)$, which reads "the probability of C given (expressed by the vertical line) the result $+$". Bayes' theorem relates $\Pr(C|+)$ to $\Pr(C)$. It is most easily understood if probability is replaced by **odds**. The odds on C is $\Pr(C)/\Pr(\sim C)$, written $\mathrm{O}(C)$, where $\sim C$ denotes the complementary event – in the example, not having breast cancer. So $\Pr(\sim C)$ is the probability that C does not obtain, namely $1 - \Pr(C)$. In this notation, Bayes' Theorem says

$$\mathrm{O}(C|+) = \frac{\Pr(+|C)}{\Pr(+|\sim C)} \times \mathrm{O}(C).$$

In words, to change the odds as a result of the extra information $+$, the original odds are multiplied by $\Pr(+|C)/\Pr(+|\sim C)$. The multiplier is called the **likelihood ratio** for C, on evidence $+$. In the example, $\Pr(+|C)$ is the probability that a woman with breast cancer will test positive – a true positive; $\Pr(+|\sim C)$ is the similar probability for a woman without breast cancer – a false positive.

Note the distinction between $\Pr(C|+)$ occurring in the odds and $\Pr(+|C)$ occurring in the likelihood ratio. The first expresses your uncertainty about the cancer when a positive result is available; the second concerns your uncertainty about whether a woman with breast cancer will test positive. The reversal of C and $+$ is very important. Bayes' theorem shows how this reversal occurs. The confusion between the two probabilities is called the "prosecutor's fallacy", perhaps because of its frequent appearance in legal cases (*see* **Medico–Legal Cases and Statistics**).

An alternative expression of the theorem is

$$\Pr(C|+) = \frac{\Pr(+|C)\Pr(C)}{\Pr(+)},$$

passing from $\Pr(C)$ to $\Pr(C|+)$. Here a new probability appears,

$$\Pr(+) = \Pr(+|C)\Pr(C) + \Pr(+|\sim C)\Pr(\sim C).$$

This is the probability of a positive result when the breast cancer state is unknown: in the example, for a healthy 40-year-old woman with three children and no family history of breast cancer.

The theorem is usually ascribed to the Rev Thomas Bayes in 1763 (*see* **Bayes, Thomas**). A convenient, modern reprint is [1]. It plays a fundamental role in one approach to statistical **inference** called *Bayesian* statistics. No one disputes the mathematics; the interpretation of $\Pr(C|+)$ is, however, controversial.

Reference

[1] Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances, *Biometrika* **45**, 293–315. Reprint of the original with a biographical note by G.A. Barnard.

(*See* **Bayesian Methods**; **Likelihood**)

DENNIS V. LINDLEY

Bayes' Theorem

DENNIS V. LINDLEY

Volume 1, pp. 304–304

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Bayes, Thomas

Born: 1701 (?) in Hertfordshire, UK.
Died: April 7, 1761, in Tunbridge Wells, UK.

Thomas Bayes, son of a nonconformist minister, spent most of his adult life in a similar position in Tunbridge Wells, England. He was educated at Edinburgh University and was a fellow of the Royal Society. He is today remembered for a paper that his friend Richard Price claimed to have found amongst his possessions after death. It appeared in the Society's Proceedings in 1763 and has often been republished [1].

By the middle of the eighteenth century it was well understood that if, to use modern terminology, in each of n independent trials the chance of success had the same value, θ say, then the probability of exactly r successes was given by the **binomial distribution**

$$\Pr(r|\theta, n) = {}^{n}C_{r}\theta^{r}(1-\theta)^{n-r}.$$

James **Bernoulli** had established the weak **law of large numbers** and **de Moivre** had found the normal approximation (*see* **Normal Distribution**) to the binomial. The passage from a known value of θ to the empirical observation of r was therefore extensively appreciated. Bayes studied the inverse problem: What did the data (r, n) say about the chance θ? There already existed partial answers in the form of significance tests (*see* **Hypothesis Testing**).

Bayes proceeded differently using the theorem that nowadays always bears his name, though it does not appear explicitly in the 1763 paper,

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

for events A and B with $\Pr(B) \neq 0$ (*see* **Bayes' Theorem**). The theorem permits the inversion of the events in $\Pr(B|A)$ into $\Pr(A|B)$. Applied when A refers to θ and B to the empirical r, we have

$$\Pr(\theta|r, n) \propto \Pr(r|\theta, n)\Pr(\theta|n).$$

The result effects the passage from the binomial, on the right, to a probability statement about the chance, on the left. It therefore becomes possible to pass from the data to a statement about what are probable, and what are improbable, values of the chance. This elegantly and simply solves the problem, except for one difficulty. It requires a value for $\Pr(\theta|n)$, a probability distribution for the chance before the result of the trials has been observed. It is usual to describe this as the **prior distribution** (prior, that is, to r) and the final result as the posterior distribution. Thus, the theorem describes how your views of θ change, from prior to posterior, as a result of data r. Bayes discussed the choice of prior but his approach is ambiguous. He is usually supposed to have taken $\Pr(\theta|n)$ **uniform** in (0,1) – the so-called Bayes' postulate – but an alternative reading suggests he took $\Pr(r|n)$ to be uniform. Mathematically these lead to the same result.

The theorem is of basic importance because it provides a solution to the general problem of **inference** or induction. Let H be a universal hypothesis and E empirical evidence bearing on H. Bayes' theorem says

$$\Pr(H|E) \propto \Pr(E|H)\Pr(H),$$

expressing a view about the hypothesis, given the evidence, in terms of the known probability of the evidence, given the hypothesis, and the prior view about H. As more evidence supporting H accrues, having large probability on H, so even the skeptic, with low $\Pr(H)$, will become convinced, $\Pr(H|E)$ will approach one and the hypothesis accepted. Many people, following **Jeffreys**, who extensively developed these ideas into a practicable scientific tool, hold that this provides a description of the scientific method.

These ideas have been extensively developed into a systematic treatment of statistics and decision making, termed *Bayesian*. The ideas therein differ from those adopted in the classical school of statistics. All this is a long way from Bayes' original problem and its resolution. He would doubtless be astonished were he to realize how his wonderful idea has been extended and his name used (*see* **Bayesian Methods**).

Reference

[1] Bayes, T. (1763). Essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society* **53**, 370–418. Reprinted in *Biometrika* **45**, 1958 293–315, with a note by G.A. Barnard.

DENNIS V. LINDLEY

Bayes, Thomas

DENNIS V. LINDLEY

Volume 1, pp. 304–305

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Bayesian Approaches to Cure Rate Models

Introduction

Survival models incorporating a cure fraction, often referred to as **cure rate models**, are becoming increasingly popular in analyzing data from cancer **clinical trials** (*see* **Clinical Trials, Early Cancer and Heart Disease**). The cure rate model has been used for modeling time-to-event data for various types of cancers, including breast cancer, non-Hodgkin's lymphoma, leukemia, prostate cancer, melanoma, and head and neck cancer, where for these diseases, a significant proportion of patients are "cured". Perhaps the most popular type of cure rate model is the mixture model discussed by Berkson and Gage [2]. In this model, we assume a certain fraction π of the population is "cured", and the remaining $1-\pi$ are not cured. The survivor function for the entire population, denoted by $S_1(y)$, for this model is given by

$$S_1(y) = \pi + (1-\pi)S^*(y), \tag{1}$$

where $S^*(y)$ denotes the survivor function for the noncured group in the population. Common choices for $S^*(y)$ are the exponential and Weibull distributions. We shall refer to the model in (1) as the *standard cure rate model*. The standard cure rate model has been extensively discussed in the statistical literature by several authors, including Farewell [13, 14], Goldman [15], Greenhouse and Wolfe [17], Halpern and Brown [18, 19], Gray and Tsiatis [16], Sposto, Sather, and Baker [31], Laska and Meisner [25], Kuk and Chen [24], Yamaguchi [39], Taylor [34], Ewell and Ibrahim [12], Stangl and Greenhouse [32], and Sy and Taylor [33]. The book by Maller and Zhou [28] gives an extensive discussion of frequentist methods of **inference** for the standard cure rate model. Although the standard cure rate model is attractive and widely used, it has some drawbacks. In the presence of **covariates**, it cannot have a **proportional hazards** structure if the covariates are modeled through π via a binomial regression model (*see* **Generalized Linear Model**). Proportional hazards are a desirable property in survival models when doing covariate analyses. Also, when including covariates through the parameter π via a standard binomial regression model, (1) yields improper posterior distributions for many types of noninformative improper **priors**, including an improper **uniform** prior for the regression coefficients. (see Chen, Ibrahim & Sinha [7]) This is a crucial drawback of (1), since it implies that **Bayesian** inference with (1) essentially requires a proper prior. These drawbacks can be overcome with an alternative definition of a cure rate model, which we discuss in the next section.

Univariate Cure Rate Models

We present a formulation of the parametric cure rate model discussed by Yakovlev et al. [37], Yakovlev [36], and Yakovlev and Tsodikov [38]. A Bayesian formulation of this model is given in [7]. The alternative model can be derived as follows. Suppose that for an individual in the population, let N denote the number of *metastasis-competent* tumor cells for that individual left active after the initial treatment. A metastasis-competent tumor cell is a tumor cell that has the potential of metastasizing. Further, assume that N has a **Poisson distribution** with mean θ. Let Z_i denote the random time for the ith metastasis-competent tumor cell to produce detectable metastatic disease. That is, Z_i can be viewed as a promotion time for the ith tumor cell. Given N, the random variables Z_i, $i = 1, 2, \ldots$, are assumed to be independent and identically distributed with a common distribution function $F(y) = 1 - S(y)$ that does not depend on N. The time to relapse of cancer can be defined by the random variable $Y = \min\{Z_i, 0 \le i \le N\}$, where $P(Z_0 = \infty) = 1$. The survival function for Y, and hence the survival function for the population, is given by

$$\begin{aligned} S_{\text{pop}}(y) &= P(\text{no metastatic cancer by time } y) \\ &= P(N=0) + P(Z_1 > y, \ldots, Z_N > y, \\ &\qquad N \ge 1). \end{aligned} \tag{2}$$

After some algebra, we obtain

$$\begin{aligned} S_{\text{pop}}(y) &= \exp(-\theta) + \sum_{k=1}^{\infty} S(y)^k \frac{\theta^k}{k!} \exp(-\theta) \\ &= \exp(-\theta + \theta S(y)) = \exp(-\theta F(y)). \end{aligned} \tag{3}$$

Since $S_{\text{pop}}(\infty) = \exp(-\theta) > 0$, (3) is not a proper survival function. As Yakovlev and Tsodikov [38]

point out, (3) shows explicitly the contribution to the failure time of two distinct characteristics of tumor growth: the initial number of metastasis-competent tumor cells and the rate of their progression. Thus, the model incorporates parameters bearing clear biological meaning. Aside from the biological motivation, the model in (3) is suitable for ' any type of survival data that has a surviving fraction. Thus, survival data which do not "fit" the biological definition given above can still certainly be modeled by (3) as long as the data has a surviving fraction and can be thought of as being generated by an unknown number N of latent competing risks (Z_i's). Thus, the model can be useful for modeling various types of survival data, including time to relapse, time to death, time to first infection, and so forth.

We also see from (3) that the cure fraction (i.e. cure rate) is given by

$$S_{\text{pop}}(\infty) \equiv P(N = 0) = \exp(-\theta). \tag{4}$$

As $\theta \to \infty$, the cure fraction tends to 0, whereas as $\theta \to 0$, the cure fraction tends to 1. The subdensity corresponding to (3) is given by

$$f_{\text{pop}}(y) = \theta f(y) \exp(-\theta F(y)), \tag{5}$$

where $f(y) = \mathrm{d}/\,\mathrm{d}yF(y)$. We note here that $f_{\text{pop}}(y)$ is not a proper probability density since $S_{\text{pop}}(y)$ is not a proper survival function. However, $f(y)$ appearing on the right side of (5) is a proper probability density function. The **hazard** function is given by

$$h_{\text{pop}}(y) = \theta f(y). \tag{6}$$

Following Chen, Ibrahim, and Sinha [7], we can now construct the **likelihood** function as follows. Suppose we have n subjects, and let N_i denote the number of metastasis-competent tumor cells for the ith subject. Further, we assume that the N_i's are independently and identically distributed (i.i.d.) Poisson random variables with mean θ, $i = 1, 2, \ldots, n$. We emphasize here that the N_i's are not observed, and can be viewed as latent variables in the model formulation. Further, suppose Z_{i1}, Z_{i2}, …, Z_{i,N_i} are the i.i.d. promotion times for the N_i metastasis-competent cells for the ith subject, which are unobserved, and all have proper cumulative distribution function $F(.)$, $i = 1, 2, \ldots, n$. In this subsection, we will specify a **parametric** form for $F(.)$, such as a **Weibull** or **gamma** distribution. We denote the indexing parameter (possibly vector valued) by $\boldsymbol{\psi}$, and thus write $F(.|\boldsymbol{\psi})$ and $S(.|\boldsymbol{\psi})$. For example, if $F(.|\boldsymbol{\psi})$ corresponds to a Weibull distribution, then $\boldsymbol{\psi} = (\alpha, \lambda)'$, where α is the shape parameter and λ is the scale parameter. Let y_i denote the survival time for subject i, which may be right censored, and let ν_i denote the censoring indicator, which equals 1 if y_i is a failure time and 0 if it is right censored. The observed data is $D_{\text{obs}} = (n, \boldsymbol{y}, \boldsymbol{\nu})$, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, and $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$. Also, let $\boldsymbol{N} = (N_1, N_2, \ldots, N_n)'$. The complete data is given by $D = (n, \boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{N})$, where $\boldsymbol{N}$ is an unobserved vector of latent variables. The complete data likelihood function of the parameters $(\boldsymbol{\psi}, \theta)$ can then be written as

$$L(\theta, \boldsymbol{\psi}|D) = \left(\prod_{i=1}^{n} S(y_i|\boldsymbol{\psi})^{N_i - \nu_i} \left(N_i f(y_i|\boldsymbol{\psi})\right)^{\nu_i}\right) \times \exp\left\{\sum_{i=1}^{n}(N_i \log(\theta) - \log(N_i!)) - n\theta\right\}. \tag{7}$$

Throughout the remainder of this section, we will assume a Weibull density for $f(y_i|\boldsymbol{\psi})$, so that

$$f(y|\boldsymbol{\psi}) = \alpha y^{\alpha-1} \exp\{\lambda - y^{\alpha} \exp(\lambda)\}. \tag{8}$$

We incorporate covariates for the parametric cure rate model (3) through the cure rate parameter θ. When covariates are included, we have a different cure rate parameter, θ_i, for each subject, $i = 1, 2, \ldots, n$. Let $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$ denote the $p \times 1$ vector of covariates for the ith subject, and let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ denote the corresponding vector of regression coefficients. We relate θ to the covariates by $\theta_i \equiv \theta(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, so that the cure rate for subject i is $\exp(-\theta_i) = \exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$, $i = 1, 2, \ldots, n$ This relationship between θ_i and $\boldsymbol{\beta}$ is equivalent to a canonical link for θ_i in the setting of generalized linear models. With this relation, we can write the complete data likelihood of $(\boldsymbol{\beta}, \boldsymbol{\psi})$ as

$$L(\boldsymbol{\beta}, \boldsymbol{\psi}|D) = \left(\prod_{i=1}^{n} S(y_i|\boldsymbol{\psi})^{N_i - \nu_i} \left(N_i f(y_i|\boldsymbol{\psi})\right)^{\nu_i}\right) \times \exp\left\{\sum_{i=1}^{n}\left[N_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log(N_i!) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right]\right\}, \tag{9}$$

where $D = (n, \boldsymbol{y}, X, \boldsymbol{v}, \boldsymbol{N})$, X is the $n \times p$ matrix of covariates, $f(y_i|\boldsymbol{\psi})$ is the Weibull density given above, and $S(y_i|\boldsymbol{\psi}) = \exp(-y_i^{\alpha}\exp(\lambda))$. If we assume independent priors for $(\boldsymbol{\beta}, \boldsymbol{\psi})$, then the posterior distributions of $(\boldsymbol{\beta}, \boldsymbol{\psi})$ are conditionally independent given $\boldsymbol{N}$. We mention that the part of the complete data likelihood in (9) involving $\boldsymbol{\beta}$ looks exactly like a Poisson generalized linear model with a canonical link, with the N_i's being the observables.

Various **semiparametric** alternatives to (9) have been proposed. Ibrahim, Chen, and Sinha [22, 23] specify a special type of prior process for $F(y)$, and Chen, Harrington, and Ibrahim [4] assume a piecewise **exponential** distribution for $F(y)$ (*see* **Grouped Survival Times**). Chen and Ibrahim [5] and Chen, Ibrahim, and Lipsitz [6] consider semiparametric cure rate models with missing covariate data, and Chen, Ibrahim, and Sinha [8] present multivariate extensions to (9), allowing for multivariate cure rate models. A recent review paper on cure rate models is given by Tsodikov, Ibrahim and Yakovlev [35]. **Joint** cure rate models for longitudinal and survival data have been considered by Law, Taylor, and Sandler [26], Brown and Ibrahim [3], and Chen, Ibrahim, and Sinha [9].

Multivariate Cure Rate Models

There does not appear to be a natural multivariate extension of the standard cure rate model in (1) (*see* **Multivariate Survival Analysis**). Even if such an extension were available, it appears that a multivariate mixture model would be extremely cumbersome to work with from a theoretical and computational perspective. As an alternative to a direct multivariate extension of (1), we examine the model discussed in Chen, Ibrahim, and Sinha [8], called the *multivariate cure rate model*. This model proves to be quite useful for modeling multivariate data in which the joint failure time random variables have a surviving fraction and each marginal failure time random variable also has a surviving fraction. The model is related to the univariate cure rate model discussed by Yakovlev et al. [37] and Asselain et al. [1]. To induce the correlation structure between the failure times, we introduce a **frailty** term [10, 20, 29], which is assumed to have a positive stable distribution. A positive frailty assumes that we have Cox's [11] proportional hazards structure conditionally (i.e. given the unobserved frailty). Thus, the marginal and conditional hazards of each component have a proportional hazards structure, and thus remain in the same class of univariate cure rate models.

For clarity and ease of exposition, we will focus our discussion on the bivariate cure rate model, as extensions to the general multivariate case are quite straightforward. The bivariate cure rate model of Chen, Ibrahim, and Sinha [8] can be derived as follows. Let $\boldsymbol{Y} = (Y_1, Y_2)'$ be a bivariate failure time, such as Y_1 = time to cancer relapse and Y_2 = time to death, or Y_1 = time to first infection, and Y_2 = time to second infection, and so forth. We assume that (Y_1, Y_2) are not ordered and have support on the upper orthant of the plane. For an arbitrary patient in the population, let $\boldsymbol{N} = (N_1, N_2)'$ denote latent (unobserved) variables for (Y_1, Y_2), respectively. We assume throughout that N_k has a Poisson distribution with mean $\theta_k w$, $k = 1, 2$, and (N_1, N_2) are independent. The quantity w is a frailty component in the model, which induces a correlation between the latent variables (N_1, N_2). Here we take w to have a positive stable distribution indexed by the parameter α, denoted by $w \sim S_{\alpha}(1, 1, 0)$, where $0 < \alpha < 1$. Although several choices can be made for the distribution of w, the positive stable distribution is quite attractive, common, and flexible in the multivariate survival setting.

Let $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i})'$ denote the bivariate promotion time for the ith metastasis-competent tumor cell. The random vectors $\boldsymbol{Z}_i$, $i = 1, 2, \ldots$ are assumed to be independent and identically distributed. The cumulative distribution function of Z_{ki} is denoted by $F_k(t) = 1 - S_k(t)$, $k = 1, 2$, and F_k is independent of (N_1, N_2). The observed survival time can be defined by the random variable $Y_k = \min\{Z_{ki}, 0 \le i \le N_k\}$, where $P(Z_{k0} = \infty) = 1$ and N_k is independent of the sequence $Z_{k1}, Z_{k2}, \ldots$, for $k = 1, 2$. The survival function for $\boldsymbol{Y} = (Y_1, Y_2)'$ given w, and hence the survival function for the population given w, is given by

$$
\begin{aligned}
S_{\text{pop}}(y_1, y_2|w) &= \prod_{k=1}^{2} [P(N_k = 0) \\
&\quad + P(Z_{k1} > t_k, \ldots, Z_{kN_k} > t_k, N_k \ge 1)] \\
&= \prod_{k=1}^{2} \left[\exp(-w\theta_k) + \left(\sum_{r=1}^{\infty} S_k(y_k)^r \right.\right. \\
&\quad \left.\left. \times \frac{(w\theta_k)^r}{r!} \exp(-w\theta_k) \right) \right]
\end{aligned}
$$

$$= \prod_{k=1}^{2} \exp\{-w\theta_k + \theta_k w S_k(y_k)\}$$

$$= \exp\left\{-w\left[\theta_1 F_1(y_1) + \theta_2 F_2(y_2)\right]\right\}, \quad (10)$$

where $P(N_k = 0) = P(Y_k = \infty) = \exp(-\theta_k)$, $k = 1, 2$. We emphasize here that the primary roles of $\mathbf{N_k}$ and $\boldsymbol{Z}_i$ are that they only facilitate the construction of the model and need not have any physical or biological interpretation at all for the model to be valid. They are quite useful for the computational implementation of the model via the Gibbs sampler as discussed below and thus are defined primarily for this purpose. The model in (10) is valid for *any* time-to-event data with a cure rate structure as implied by (10) and the subsequent development. Thus, the model can be useful for modeling various types of failure time data, including time to relapse, time to death, time to infection, time to complication, time to rejection, and so forth. In addition, the frailty variable w serves a dual purpose in the model – it induces the correlation between Y_1 and Y_2 and at the same time relaxes the Poisson assumption of N_1 and N_2 by adding the same extra Poisson variation through their respective means $\theta_1 w$ and $\theta_2 w$.

Following Ibragimov and Chernin [21], the $S_\alpha(1, 1, 0)$ density for w $(0 < \alpha < 1)$ can be expressed in the form

$$f_s(w|\alpha) = a w^{-(a+1)} \int_0^1 s(u) \times \exp\left\{-\frac{s(u)}{w^a}\right\} du, \quad w > 0, \quad (11)$$

where

$$a = \frac{\alpha}{1-\alpha} \text{ and } s(u) = \left(\frac{\sin(\alpha\pi u)}{\sin(\pi u)}\right)^a \times \left(\frac{\sin[(1-\alpha)\pi u]}{\sin(\pi u)}\right),$$

and the Laplace transform of w is given by $E(\exp(-sw)) = \exp(-s^\alpha)$. A useful reference on stable distributions is Samorodnitsky & Taqqu [30]. Using the Laplace transform of w, a straightforward derivation yields the unconditional survival function

$$S_{\text{pop}}(y_1, y_2) = \exp\{-[\theta_1 F_1(y_1) + \theta_2 F_2(y_2)]^\alpha\}. \quad (12)$$

The joint cure fraction implied by (12) is $S_{\text{pop}}(\infty, \infty) = \exp(-[\theta_1 + \theta_2]^\alpha)$. From (12), the marginal survival functions are

$$S_k(y) = \exp(-\theta_k^\alpha (F_k(y))^\alpha), \quad k = 1, 2. \quad (13)$$

Equation (13) indicates that the marginal survival functions have a cure rate structure with probability of cure $\exp(-\theta_k^\alpha)$ for Y_k, $k = 1, 2$. It is important to note in (13) that each marginal survival function has a proportional hazards structure as long as the covariates, $\boldsymbol{x}$, only enter through θ_k. The marginal hazard function is given by $\alpha\theta_k^\alpha f_k(y)(F_k(y))^{\alpha-1}$, with attenuated covariate effect $(\theta_k(x))^\alpha$, and $f_k(y)$ is the survival density corresponding to $F_k(y)$. This property is similar to the earlier observations made by Oakes [29] for the ordinary bivariate stable frailty survival model.

In addition, we can express the marginal survival functions in (13) in terms of standard cure rate models. We can write

$$\begin{aligned} S_k(y) &= \exp(-\theta_k^\alpha (F_k(y))^\alpha) \\ &= \exp(-\theta_k^\alpha) + (1 - \exp(-\theta_k^\alpha)) \\ &\quad \times \left(\frac{\exp(-\theta_k^\alpha (F_k(y))^\alpha) - \exp(-\theta_k^\alpha)}{1 - \exp(-\theta_k^\alpha)}\right) \\ &= \exp(-\theta_k^\alpha) + (1 - \exp(-\theta_k^\alpha)) S_k^*(y), \quad (14) \end{aligned}$$

where

$$S_k^*(y) = \frac{\exp(-\theta_k^\alpha (F_k(y))^\alpha) - \exp(-\theta_k^\alpha)}{1 - \exp(-\theta_k^\alpha)}, \quad k = 1, 2.$$

It is easily shown that $S_k^*(y)$ defines a proper survivor function. Thus, (14) is a standard cure rate model with cure rate given by $\pi_k = \exp(-\theta_k^\alpha)$ and survivor function for the noncured population given by $S_k^*(y)$ for $k = 1, 2$.

The parameter α $(0 < \alpha < 1)$ is a scalar parameter that is a measure of association between (Y_1, Y_2). Small values of α indicate high association between (Y_1, Y_2). As $\alpha \to 1$, this implies less association between (Y_1, Y_2), which can be seen from (12). Following Clayton [10] and Oakes [29], we can compute a local measure of dependence, denoted by $\theta^*(y_1, y_2)$, as a function of α. For the multivariate cure rate model in (12), $\theta^*(y_1, y_2)$ is well defined, and is given by

$$\theta^*(y_1, y_2) = \alpha^{-1}(1-\alpha)(\theta_1 F_1(y_1) + \theta_2 F_2(y_2))^{-\alpha} + 1. \quad (15)$$

We see that $\theta^*(y_1, y_2)$ in (15) decreases in (y_1, y_2). That is, the association between (Y_1, Y_2) is greater when (Y_1, Y_2) are small and the association decreases over time. Such a property is desirable, for example, when Y_1 denotes time to relapse and Y_2 denotes time to death. Finally, we mention that a global measure of dependence such as Kendall's τ (*see* **Rank Correlation**) or the Pearson **correlation** coefficient is not well defined for the multivariate cure rate model (12), since no moments for cure rate models exist due to the improper survival function.

The multivariate cure rate model presented here is attractive in several respects. First, the model has a proportional hazards structure for the population hazard, conditionally as well as marginally, when covariates are entered through the cure rate parameter, and thus has an appealing interpretation. Also, the model is computationally feasible. In particular, by introducing latent variables, efficient **Markov chain Monte Carlo algorithms** can be developed that enable us to sample from the joint posterior distribution of the parameters. Chen, Ibrahim, and Sinha [8] discuss a modified version of the collapsed Gibbs technique of Liu [27] for efficient Gibbs sampling from the posterior distribution.

The likelihood function for this model can be obtained as follows. Suppose we have n subjects, and let N_{ki} denote the number of latent risks for the ith subject, $i = 1, 2, \ldots, n$, $k = 1, 2$. Further, we assume that the N_{ki}'s are independent Poisson random variables with mean $w_i\theta_k$, $i = 1, 2, \ldots, n$ for $k = 1, 2$. We also assume the $w_i \sim S_\alpha(1, 1, 0)$, and the w_i's are i.i.d. We emphasize here that the N_{ki}'s are not observed, and can be viewed as latent variables in the model formulation. Further, suppose $Z_{ki1}, Z_{ki2}, \ldots, Z_{ki,N_{ki}}$ are the independent latent times for the N_{ki} latent risks for the ith subject, which are unobserved, and all have cumulative distribution function $F_k(.)$, $i = 1, 2, \ldots, n$, $k = 1, 2$. Chen, Ibrahim, and Sinha [8] specify a parametric form for $F_k(.)$, such as a Weibull or gamma distribution. We denote the indexing parameter (possibly vector valued) by $\boldsymbol{\psi}_k$, and thus write $F_k(.|\boldsymbol{\psi}_k)$ and $S_k(.|\boldsymbol{\psi}_k)$. For example, if $F_k(.|\boldsymbol{\psi}_k)$ corresponds to a Weibull distribution, then $\boldsymbol{\psi}_k = (\xi_k, \lambda_k)'$, where ξ_k is the shape parameter and λ_k is the scale parameter. Let y_{ki} denote the failure time or censoring time for subject i for the kth component, and let indicator $\nu_{ki} = 1$, if y_{ki} is an observed failure time and 0 if it is a censoring time. Let $\boldsymbol{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kn})$, $\boldsymbol{\nu}_k = (\nu_{k1}, \nu_{k2}, \ldots, \nu_{kn})$, $\boldsymbol{N}_k = (N_{k1}, N_{k2}, \ldots N_{kn})$, $k = 1, 2$, and $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)'$. The complete data is given by $D = (n, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{N}_1, \boldsymbol{N}_2, \boldsymbol{w})$, where $\boldsymbol{N}_1$, $\boldsymbol{N}_2$, and $\boldsymbol{w}$ are unobserved random vectors, and the observed data is given by $D_{\text{obs}} = (n, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$. Further, let $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \boldsymbol{\psi}_2')'$. The likelihood function of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ based on the complete data D is given by

$$
\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\psi}|D) &= \left(\prod_{k=1}^{2} \prod_{i=1}^{n} S_k(y_{ki}|\boldsymbol{\psi}_k)^{N_{ki}-\nu_{ki}} \right. \\
&\quad \left. \times \left(N_{ki} f_k(y_{ki}|\boldsymbol{\psi}_k) \right)^{\nu_{ki}} \right) \\
&\quad \times \exp\left\{ \sum_{i=1}^{n} (N_{ki} \log(w_i\theta_k) - \log(N_{ki}!) - w_i\theta_k) \right\},
\end{aligned} \tag{16}
$$

where $f_k(y_{ki}|\boldsymbol{\psi}_k)$ is the density corresponding to $F_k(y_{ki}|\boldsymbol{\psi}_k)$. We assume a Weibull density for $f_k(y_{ki}|\boldsymbol{\psi}_k)$, so that

$$
f_k(y|\boldsymbol{\psi}_k) = \xi_k y^{\xi_k - 1} \exp\left\{ \lambda_k - y^{\xi_k} \exp(\lambda_k) \right\}. \tag{17}
$$

To construct the likelihood function of the observed data, $L(\boldsymbol{\theta}, \boldsymbol{\psi}, \alpha|D_{\text{obs}})$, we integrate (16) with respect to $(\mathbf{N}, \boldsymbol{w})$ assuming a $S_\alpha(1, 1, 0)$ density for each w_i, denoted by $f_s(w_i|\alpha)$, leading to

$$
\begin{aligned}
&L(\boldsymbol{\theta}, \boldsymbol{\psi}, \alpha|D_{\text{obs}}) \\
&\quad \equiv \int_{R^{+n}} \sum_{(\mathbf{N_1}, \mathbf{N_2})} L(\boldsymbol{\theta}, \boldsymbol{\psi}|D) \times \left[\prod_{i=1}^{n} f_s(w_i|\alpha) \right] \mathrm{d}\boldsymbol{w} \\
&\quad = \theta_1^{d_1} \theta_2^{d_2} \alpha^{d_1+d_2} \left[\prod_{k=1}^{2} \prod_{i=1}^{n} f_k(y_{ki}|\boldsymbol{\psi}_k)^{\nu_{ki}} \right] \\
&\quad \times \prod_{i=1}^{n} \left\{ [\theta_1 F_1(y_{1i}|\boldsymbol{\psi}_1) + \theta_2 F_2(y_{2i}|\boldsymbol{\psi}_2)]^{(\alpha-1)(\nu_{1i}+\nu_{2i})} \right\} \\
&\quad \times \prod_{i=1}^{n} \left[\alpha^{-1}(1-\alpha)(\theta_1 F_1(y_{1i}|\boldsymbol{\psi}_1) \right. \\
&\quad \left. + \theta_2 F_2(y_{2i}|\boldsymbol{\psi}_2))^{-\alpha} + 1 \right]^{\nu_{1i}\nu_{2i}} \\
&\quad \times \prod_{i=1}^{n} \exp\left\{ -(\theta_1 F_1(y_{1i}|\boldsymbol{\psi}_1) + \theta_2 F_2(y_{2i}|\boldsymbol{\psi}_2))^{\alpha} \right\},
\end{aligned} \tag{18}
$$

where $f_s(w_i|\alpha)$ denotes the probability density function of w_i defined by (11), $d_k = \sum_{i=1}^{n} \nu_{ki}$ for $k = 1, 2$, $R^{+n} = R^+ \times R^+ \times \ldots \times R^+$, and $R^+ = (0, \infty)$. As before, we incorporate covariates for the cure rate model (12) through the cure rate parameter $\boldsymbol{\theta}$. Let $\boldsymbol{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the $p \times 1$ vector of covariates for the ith subject, and let $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})'$ denote the corresponding vector of regression coefficients for failure time random variable Y_k, $k = 1, 2$. We relate $\boldsymbol{\theta}$ to the covariates by

$$\theta_{ki} \equiv \theta(\boldsymbol{x}_i'\boldsymbol{\beta}_k) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k), \tag{19}$$

so that the cure rate for subject i is

$$\exp(-\theta_{ki}) = \exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)), \tag{20}$$

for $i = 1, 2, \ldots, n$ and $k = 1, 2$. Letting $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, we can write the observed data likelihood of $(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha)$ as

$$\begin{aligned}
&L(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha|D_{\text{obs}}) \\
&= \left(\alpha^{d_1+d_2} \prod_{k=1}^{2} \prod_{i \in \mathcal{D}_k} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)\right) \\
&\times \left[\prod_{k=1}^{2} \prod_{i=1}^{n} f_k(y_{ki}|\boldsymbol{\psi}_k)^{\nu_{ki}}\right] \\
&\times \prod_{i=1}^{n} \left\{[\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1) F_1(y_{1i}|\boldsymbol{\psi}_1) \right. \\
&\left. + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2) F_2(y_{2i}|\boldsymbol{\psi}_2)]^{(\alpha-1)(\nu_{1i}+\nu_{2i})}\right\} \\
&\times \prod_{i=1}^{n} \left\{\frac{1-\alpha}{\alpha}[\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1) F_1(y_{1i}|\boldsymbol{\psi}_1) \right. \\
&\left. + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2) F_2(y_{2i}|\boldsymbol{\psi}_2)]^{-\alpha} + 1\right\}^{\nu_{1i}\nu_{2i}} \\
&\times \prod_{i=1}^{n} \exp\left\{-(\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1) F_1(y_{1i}|\boldsymbol{\psi}_1) \right. \\
&\left. + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2) F_2(y_{2i}|\boldsymbol{\psi}_2))^{\alpha}\right\},
\end{aligned} \tag{21}$$

where $\mathcal{D}_k$ consists of those patients who failed according to Y_k, $k = 1, 2$, $D_{\text{obs}} = (n, \boldsymbol{y}_1, \boldsymbol{y}_2, X, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$, X is the $n \times p$ matrix of covariates, $f_k(y_{ki}|\boldsymbol{\psi}_k)$ is given by (17), and

$$f_k(y_{ki}|\boldsymbol{\psi}_k) = 1 - \exp\{-y_{ki}^{\xi_k} \exp(\lambda_k)\}. \tag{22}$$

Chen, Ibrahim, and Sinha [8] consider a joint improper prior for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \alpha)$ of the form

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha) &= \pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \alpha) \\
&\propto \pi(\boldsymbol{\psi}_1)\pi(\boldsymbol{\psi}_2) I(0 < \alpha < 1) \\
&= \prod_{k=1}^{2} \pi(\xi_k, \lambda_k) I(0 < \alpha < 1),
\end{aligned} \tag{23}$$

where $I(0 < \alpha < 1) = 1$ if $0 < \alpha < 1$, and 0 otherwise. Thus, (23) implies that $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and α are independent *a* priori, $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ are independent *a* priori with an improper uniform prior, α has a proper uniform prior over the interval $(0, 1)$, and $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ are independent and identically distributed as $\pi(\boldsymbol{\psi}_k)$ *a* priori. They also assume that

$$\pi(\xi_k, \lambda_k) = \pi(\xi_k|\nu_0, \tau_0)\pi(\lambda_k), \tag{24}$$

where

$$\pi(\xi_k|\delta_0, \tau_0) \propto \xi_k^{\delta_0 - 1} \exp\{-\tau_0 \xi_k\}, \text{ and}$$
$$\pi(\lambda_k) \propto \exp\{-c_0 \lambda_k^2\},$$

and δ_0, τ_0, and c_0 are specified hyperparameters. With these specifications, the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha)$ based on the observed data $D_{\text{obs}} = (n, \boldsymbol{y}_1, \boldsymbol{y}_2, X, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$ is given by

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha|D_{\text{obs}}) &\propto L(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha|D_{\text{obs}}) \\
&\times \prod_{k=1}^{2} \pi(\xi_k|\delta_0, \tau_0)\pi(\lambda_k),
\end{aligned} \tag{25}$$

where $L(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha|D_{\text{obs}})$ is given by (21). Chen, Ibrahim, and Sinha [8] show that the posterior distribution in (25) using the noninformative improper prior (23) is proper under some very general conditions.

References

[1] Asselain, B., Fourquet, A., Hoang, T., Tsodikov, A.D. & Yakovlev, A.Y. (1996). A parametric regression model of tumor recurrence: An application to the analysis of clinical data on breast cancer, *Statistics and Probability Letters* **29**, 271–278.

[2] Berkson, J. & Gage, R.P. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association* **47**, 501–515.

[3] Brown, E.R. & Ibrahim, J.G. (2003). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials, *Biometrics* **59**, 686–693.

[4] Chen, M-H., Harrington, D.P. & Ibrahim, J.G. (2002). Bayesian models for high-risk melanoma: A case study of ECOG trial E1690, *Applied Statistics* **51**, 135–150.

[5] Chen, M.H. & Ibrahim, J.G. (2001). Maximum likelihood methods for cure rate models with missing covariates, *Biometrics* **57**, 43–52.

[6] Chen, M-H., Ibrahim, J.G. & Lipsitz, S.R. (2002). Bayesian methods for missing covariates in cure rate models, *Lifetime Data Analysis* **8**, 117–146.

[7] Chen, M-H., Ibrahim, J.G. & Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association* **94**, 909–919.

[8] Chen, M-H., Ibrahim, J.G. & Sinha, D. (2002). Bayesian inference for multivariate survival data with a surviving fraction, *Journal of Multivariate Analysis* **80**, 101–126.

[9] Chen, M-H., Ibrahim, J.G. & Sinha, D. (2004). A new joint model for longitudinal and survival data with a cure fraction, *Journal of Multivariate Analysis* **91**, 18–34.

[10] Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familiar tendency in chronic disease incidence, *Biometrika* **65**, 141–151.

[11] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[12] Ewell, M. & Ibrahim, J.G. (1997). The large sample distribution of the weighted log rank statistic under general local alternatives, *Lifetime Data Analysis* **3**, 5–12.

[13] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* **38**, 1041–1046.

[14] Farewell, V.T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics* **14**, 257–262.

[15] Goldman, A.I. (1984). Survivorship analysis when cure is a possibility: A Monte Carlo study, *Statistics in Medicine* **3**, 153–163.

[16] Gray, R.J. & Tsiatis, A.A. (1989). A linear rank test for use when the main interest is in differences in cure rates, *Biometrics* **45**, 899–904.

[17] Greenhouse, J.B. & Wolfe, R.A. (1984). A competing risks derivation of a mixture model for the analysis of survival, *Communications in Statistics - Theory and Methods* **13**, 3133–3154.

[18] Halpern, J. & Brown, B.W. Jr. (1987a). Cure rate models: Power of the log rank and generalized Wilcoxon tests, *Statistics in Medicine* **6**, 483–489.

[19] Halpern, J. & Brown, B.W. Jr. (1987b). Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test, *Controlled Clinical Trials* **8**, 177–189.

[20] Hougaard, P. (1986). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678.

[21] Ibragimov, I.A. & Chernin, K.E. (1959). On the unimodality of stable laws, *Theory of Probability and its Applications* **4**, 417–419.

[22] Ibrahim, J.G., Chen, M-H. & Sinha, D. (2001a). *Bayesian Survival Analysis*. Springer-Verlag, New York.

[23] Ibrahim, J.G., Chen, M-H. & Sinha, D. (2001b). Bayesian semi-parametric models for survival data with a cure fraction, *Biometrics* **57**, 383–388.

[24] Kuk, A.Y.C. & Chen, C-H. (1992). A mixture model combining logistic regression with proportional hazards regression, *Biometrika* **79**, 531–541.

[25] Laska, E.M. & Meisner, M.J. (1992). Nonparametric estimation and testing in a cure rate model, *Biometrics* **48**, 1223–1234.

[26] Law, N.J., Taylor, J.M.G. & Sandler, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure, *Biostatistics* **3**, 547–563.

[27] Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem, *Journal of the American Statistical Association* **89**, 958–966.

[28] Maller, R. & Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

[29] Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**, 487–493.

[30] Samorodnitsky, G. & Taqqu, M.S. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, London.

[31] Sposto, R., Sather, H.N. & Baker, S.A. (1992). A comparison of tests of the difference in the proportion of patients who are cured, *Biometrics* **48**, 87–99.

[32] Stangl, D.K. & Greenhouse, J.B. (1998). Assessing placebo response using Bayesian hierarchical survival models, *Lifetime Data Analysis* **4**, 5–28.

[33] Sy, J.P. & Taylor, J.M.G. (2000). Estimation in a proportional hazards cure model, *Biometrics* **56**, 227–336.

[34] Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture models, *Biometrics* **51**, 899–907.

[35] Tsodikov, A.D., Ibrahim, J.G. & Yakovlev, A.Y. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association* **98**, 1063–1078.

[36] Yakovlev, A.Y. (1994). Letter to the Editor, *Statistics in Medicine* **13**, 983–986.

[37] Yakovlev, A.Y., Asselain, B., Bardou, V.J., Fourquet, A., Hoang, T., Rochefediere, A. & Tsodikov, A.D. (1993). A simple stochastic model of tumor recurrence and its applications to data on premenopausal breast cancer, in *Biometrie et Analyse de Dormees Spatio-Temporelles*, Vol. 12, B. Asselain, M. Boniface, C. Duby, C. Lopez, J.P. Masson & J. Tranchefort eds. Société Francaise de Biométrie, ENSA Rennes, France, pp. 66–82.

[38] Yakovlev, A.Y. & Tsodikov, A.D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, New Jersey.

[39] Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in Japan, *Journal of the American Statistical Association* **87**, 284–292.

Further Reading

Ibrahim, J.G. & Chen, M-H. (2000). Power prior distributions for regression models, *Statistical Science* **15**, 46–60.

(*See also* **Bayesian Survival Analysis**; **Bayesian Model Selection in Survival Analysis**)

JOSEPH G. IBRAHIM, MING-HUI CHEN & DEBAJYOTI SINHA

Bayesian Decision Models in Health Care

In many situations, analysts are required to assess the probability of a unique event, where there are no relevant historical patterns. Sometimes, the environment has changed so radically that the past trends for a familiar event are no longer relevant. Still other times, it is theoretically possible to gather historical data, but time or money limitations prevent data collection. In these circumstances, a "Bayesian subjective probability" model is appropriate (*see* **Bayesian Methods**). The process for creating a Bayesian probability model is explained elsewhere [4].

Bayesian probability models have been used to model complex health care issues such as predicting who will sue a hospital [2], assessing probability of mortality from myocardial infarctions [3], and predicting which health planning project is most likely to succeed [5]. Here, we give an example of the way the model was applied to create an index for predicting preventable hospitalization [1].

A health insurance company wished to adjust premiums on the basis of the preventable and modifiable health risks of individual members. Although many risk factors for hospitalization had been identified (e.g. smoking), no studies had been done that combined the various risks into one aggregate scale. To accomplish this task, leading researchers met in a consensus panel. During this meeting a model was constructed that summarized their opinions about the overall risk of an individual engaged in different lifestyles.

The goal of the exercise was to predict the probability of a major hospitalization during the next three years. The **explanatory variables** of interest were alcohol and tobacco use, weight, blood pressure, dyslipidemia, risk of trauma, and depression. For each factor and its potential levels, experts were asked to estimate two probabilities. For example, the experts individually and then as a group answered the following two questions related to smoking:

1. Of 100 people who *have* been hospitalized, how many smoked a pack a day in the last three years?
2. Of 100 people who *have not* been hospitalized, how many smoked a pack a day in the last three years?

The ratio of the answers to these two questions provided the **likelihood ratio** associated with smoking more than two packs a day. Experts also estimated the prior odds for different population groups, and a predictive model was created [5]. Finally, experts rated 64 hypothetical cases and the average of these ratings was compared with the Bayesian model predictions. The model, which was found to agree with the expert judgments, took only two days to construct, which is considerably less time than would have been required to collect and analyze a large data set. Not only are subjective models quick to construct, but they may also be more valid and generalizable than models based on a specific data set that may have significant biases.

References

[1] Alemi, F., Gustafson, D.H. & Johnson, M. (1986). How to construct a subjective index, *Evaluation and Health Profession* **9**, 42–52.

[2] Driver, J.F. & Alemi, F. (1995). Forecasting without historical data: Bayesian probability models utilizing expert opinions, *Journal of Medical Systems* **19**, 359–374.

[3] Eisenstein, E.L. & Alemi, F. (1994). An evaluation of factors influencing Bayesian learning systems, *Journal of the American Medical Informatics Association* **1**, 274–284.

[4] Gustafson, D.H., Cats-Baril, W.L. & Alemi, F. (1992). *Systems to Support Health Policy Analysis*. Health Administration Press, Ann Arbor.

[5] Gustafson, D.H., Cats-Baril, W.L. & Alemi, F. (1992). Using Bayesian and MAV models to analyze implementation, in *Systems to Support Health Policy Analysis*. Health Administration Press, Ann Arbor.

F. ALEMI

Bayesian Decision Models in Health Care

F. ALEMI

Volume 1, pp. 313–313

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Bayesian Measures of Goodness of Fit

Introduction

With the relatively recent advent of **Markov Chain Monte Carlo** techniques, and their flexibility and ease of application, we might be tempted to believe that we can now fit almost any practical **Bayesian** model, and report inference based on the MCMC's sampled approximation of the posterior **likelihood**. However, this suggested analysis does not at any stage involve checking that the model is indeed "sensible" and, following the model construction and criticism **algorithm** laid down by Carota et al. [6], simply concluding the analysis after completion of the MCMC computation may be unwise and potentially misleading. Exactly as with classical analyses, we should pursue some indication that the fitted model actually fits well. That is, we should critically assess whether the data appears to violate some or all of the model assumptions.

In classical approaches, **goodness of fit** statistics measure the fit, "distance", or "discrepancy" between the specified model and the data. Under a Bayesian approach, this description still holds, but the definition of the "model" now includes a **prior distribution**, which to some extent complicates this assessment.

In the summary that follows, we shall describe Bayesian methods for assessing goodness of fit, which use only the model (and prior) *specified*, and do not refer to any particular alternative formulations. From a strict Bayesian point of view, the reader should note that we may be on thin ice here; as for the prior, if we truly believe the model, then we have no reason to question it, and if there are alternative models, then these should be described and the models compared via their posterior probabilities, as one might when using **Bayes Factors**, or when conducting Bayesian model averaging (*see* **Bayesian Methods for Model Comparison**). This is not the same as the common situation considered here, where we have no specific alternative models, but do have some measures that (we believe) realistically reflect how well the specified model is performing. In summary then, this article will deal with discrepancies between model and data of types, which are somehow prespecified by the user, and therefore a more rigorous title might cite "measures of *practical* goodness of fit".

In this review, we deal first with the general ideas used and their application to relatively simple nonhierarchical models, and then cover the more complicated issues that **hierarchical** structures bring. The field is relatively new, and builds on the classical goodness-of-fit theory with which the reader should compare these methods.

There are a number of related issues of which the reader should be aware but which we will not cover here. *Overfitting* occurs when the model has too many parameters for the size of dataset considered, and it will generally fit suspiciously well. The methods described here are generally derived for testing the opposite problem of poor fit, but most methods could be adapted to look for data at the other extreme.

If we are not seeking an overall measure of fit, but rather methods for choosing the best of a selection of well-defined models, then this is *model comparison* (*see* **Model, Choice of**). A number of authors have considered Bayesian approaches to this issue; for a general review and new contributions, *see* **Bayesian Methods for Model Comparison**.

Finally, if our measure of goodness of fit is not only part of an inferential analysis as considered here, but forms some section of a decision-making process with associated costs, decisions to reject any model because of poor fit must be driven by the processes that determine how poor fit will be ascertained (*see* **Decision Theory**). We must specify the purpose for which the model is unacceptable, and hence a **utility** function. For a short discussion and references, see Draper's comment in [13].

Methods for Assessing Goodness of fit in Nonhierarchical Bayesian Models

***P* values.** Define $T(X)$ to be some measure of discrepancy between the data x, and the model, where T is chosen by the practitioner. Evidence against the model is indicated by large values of T, and we wish to examine what the chances are of seeing a value of T worse than the one we get from our observed data x_{obs}. The probability of observing any of these worse values is given by the ***P* value**

$$p = \mathbb{P}_H(T(X) \geq T(x_{\text{obs}})). \tag{1}$$

To obtain P values, the practitioner therefore has to choose both an appropriate discrepancy measure T

and the distribution H (or density h) for X under the assumed model, which automatically define the distribution of T.

The "natural" [3] Bayesian choice for h is the *prior predictive distribution*, so that

$$h_{\text{prior pred}} = \int f(x|\theta)\pi(\theta)\,\mathrm{d}\theta. \tag{2}$$

However, this cannot be used if our prior π is improper, which precludes the use of some reference distributions, and is therefore rather unsatisfactory. A suggested alternative is to use the *posterior predictive distribution*;

$$h_{\text{post pred}} = \int f(x|\theta)\pi(\theta|x_{\text{obs}})\,\mathrm{d}\theta, \tag{3}$$

where $\pi(\theta|x_{\text{obs}})$ of course is proportional to $f(x|\theta)\pi(\theta)$ by **Bayes' Theorem**.

Defined as the P value obtained using $H_{\text{post pred}}$ from (3) in (1), $p_{\text{post pred}}$ is asymptotically distributed as $U(0, 1)$, and so is interpretable in the same way as a classical P value [9, 10]. However, examining (3), we see that this approach leads to us comparing our observed statistic $T(X_{\text{obs}})$ with a distribution derived using x_{obs}. This uses the data twice, and can lead to overly conservative inferences [1].

Bayarri and Berger [2, 1], have proposed two alternatives, using principles similar to those used in (frequentist) **conditional** and **partial likelihood** arguments. To define these, we denote the density of $T(X)$ by $f(t|\theta)$, its observed value by $t_{\text{obs}} = T(x_{\text{obs}})$, and similarly for other functions of the data.

The *conditional predictive P value*, for some discrepancy function T, requires the user to choose another function U, which is preferably approximately independent of T, does not include T, and contains as much information as possible about θ. (The reader may be interested to compare this with the problem of choosing statistics to condition on for conditional inference.) A simple example, splitting the data into $T = x_1, \ldots, x_k$ and $U = x_{k+1}, \ldots, x_n$ was put forward by Evans [11]. In general, using

$$h_{\text{cond}}(t|u) = \int f(t|u, \theta)\pi(\theta|u)\,\mathrm{d}\theta, \tag{4}$$

where $\pi(\theta|u) \propto f(u|\theta)\pi(\theta)$, we get

$$p_{\text{cond pred}}(t) = \mathbb{P}_{H_{\text{cond}}(\cdot|u_{\text{obs}})}(T \geq t_{\text{obs}}). \tag{5}$$

This P value effectively conditions out almost all of the information about θ, leaving a distribution for T, which reflects the fit of the model f rather than the prior θ. Because the data is (preferably) partitioned into U and T, we should not use any of it twice, and improper priors can be used as long as $\pi(\theta|u)$ is proper. However, the choice of U may not be obvious, and even if some U can be found $p_{\text{cond pred}}$ may be hard to compute.

An alternative to finding an appropriate U is to adjust the posterior predictive distribution $h_{\text{post pred}}$ from (3) to remove the contribution from t_{obs}. This is done by conditioning on T, so we get

$$h_{\text{part pred}}(t) = \int f(t|\theta)\pi_{\text{part}}(\theta)\,\mathrm{d}\theta, \tag{6}$$

where

$$\pi_{\text{part}}(\theta) \propto \frac{f(x_{\text{obs}}|\theta)\pi(\theta)}{f(t_{\text{obs}}|\theta)} \tag{7}$$

Using $H_{\text{part pred}}$ in place of H_{cond} in (5) gives the **partial predictive *P* value**; it is easier to compute while, due to its derivation, we can expect it to give similar inferences to $p_{\text{cond pred}}(t)$.

The work of Gelman et al. [12, 13] defines P values similar to those derived from (1) but explicitly includes the idea of *replication*. Here we imagine that we have x^{rep}, a replicated dataset that might be observed if the experiment were run again, using exactly the same model M, parameters θ and, importantly, the same ancillary statistics, written as $A(X)$. We move from considering $f(x|\theta)\pi(\theta)$ to $f(x|\theta)\pi(\theta)f(x^{\text{rep}}|\theta)$. The classical goodness-of-fit statistic can be written as

$$p_{\text{class}}(x, \theta) = \mathbb{P}(T(x^{\text{rep}}) \geq T(x)|A, M, \theta), \tag{8}$$

for some fixed but unknown θ. The difficulty in the classical application of p_{class} lies in finding some pivotal T, which has a distribution at least approximately free of θ. The Bayesian analogue [12, 13] avoids this by averaging p_{class} over the posterior distribution for θ; hence,

$$\begin{aligned} p_{\text{Bayes}}(x) &= \mathbb{P}(T(x^{\text{rep}}) \geq T(x)|A, M) \\ &= \int p_{\text{class}}(x, \theta) f(\theta|A, M, x)\,\mathrm{d}\theta, \end{aligned} \tag{9}$$

and given a prior for θ, we can compute p_{Bayes} for any dataset. This approach is in turn generalizable in two

ways: firstly, we could take a fully Bayesian approach and allow for θ^{rep}, parameters replicated in the same way as for x^{rep}. Second, we can generalize the discrepancy function T to include some dependence on θ, and consider the model as ill-fitting when

$$p'_{\text{Bayes}} = \mathbb{P}(T(X, \theta) \geq T(x_{\text{obs}}, \theta) | x_{\text{obs}}, A(x_{\text{obs}})) \tag{10}$$

is small. This can in turn be generalized if we change from the symmetric use of discrepancy function T in (10) to the more general $D(x, x^{\text{rep}}, \theta, \theta^{\text{rep}})$, where we also allow for parameter replication as mentioned above. Then (9) becomes

$$p''_{\text{Bayes}}(x) = \mathbb{P}(D(x, x^{\text{rep}}, \theta, \theta^{\text{rep}}) > 0). \tag{11}$$

As a special case of (11), putting $D(x) = \min_\theta D(x, \theta)$, we remove the dependence of p_{Bayes} on θ, and for linear models obtain the classical tests of goodness of fit, comparing a sum of standardized **residuals** from the data with the best fit of the model, a **χ^2 test**. For many models, such a construction also gives some idea of which data points are *causing* the poor fit of the data, clearly a useful tool in many situations, for example, the exclusion of outliers. The analysis of residuals in this way forms the basis of much of frequentist testing, even outside of linear models. For a Bayesian version of these case-influence diagnostics, see the work of Chaloner [7], later generalized by Weiss [27–28]. This method of constructing goodness-of-fit measures is explored further in the section on hierarchical models.

The connection with classical residuals is attractive, but not straightforward. The classical residual has several forms, but they are fundamentally a measure of distance between a data point and its fitted value under the model. Summing these gives the classical overall measure of fit mentioned above. The Bayesian approach gives a posterior distribution for parameters, not fitted values, and so each residual also has a distribution, as does the sum of all residuals. We could of course use "plug-in" values like the posterior mean or median, but this contradicts the Bayesian practice of properly allowing for all sources of uncertainty. Calibrating the size of residuals in a meaningful way is also difficult and often sensitive to parameterization changes, unlike the P value techniques discussed here.

Bayes Factors. All the methods of the previous section result in some form of tail probability, a value between 0 and 1, where this scale indicates poor to good measure of goodness of fit. Care should be taken that this value is distributed (at least approximately) as $U(0, 1)$ if we want to interpret these as P values [9, 10]. This potential incoherence is noted by Conigliani et al. [8], who suggest the alternative use of *Bayes factors*. To use these, we *must* exactly specify all the alternative priors, models and associated parameters, giving a series of $f_i(x|\theta_i)$ and $\pi_i(\theta_i)$ for each model M_i.

(To keep to our remit of discussing models where no alternative is readily available, we shall here only mention models where a "general alternative" can be reasonably assumed; typically, these are models for discrete data (*see* **Categorical Data Analysis**), for example, **binomial** or **Poisson**, where the alternative is that the data comes from a **multinomial** distribution where the cell probabilities have a Dirichlet distribution. Strictly speaking, this is model comparison, but as we have no intention of accepting the **alternative hypothesis**, it can be viewed as constructing a diagnostic measure sensitive to deviations from the **null hypothesis**, or a measure of goodness of fit.)

The Bayes factor in favor of model 1 against the alternative model 2 is

$$B(x) = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\,\mathrm{d}\theta}{\int f_2(x|\theta_2)\pi_2(\theta_2)\,\mathrm{d}\theta}, \tag{12}$$

which can be interpreted as the ratio of posterior **odds** (of model 2 to model 1) to the prior odds. A large value of B indicates that model 1 fits better than the alternative, and scales exist for converting numerical Bayes factors into strengths of evidence against the model, for example, [16]. However, similarly to the prior predictive P value, the standard Bayes factor cannot be defined when either $\pi_1(\theta_1)$ or $\pi_2(\theta_2)$ is improper [21]. As discussed above, where we use a "general alternative" for π_2, this choice may well be improper. It should at least be rather diffuse, in which case the Bayes factor is sensitive to the degree of flatness in π_2 [8, 13].

Various solutions to these problems have been proposed. O'Hagan [20] developed the *partial Bayes factor*, where part of the data is used to update the prior so that it is proper, and the rest used to calculate the Bayes factor. If we split x into training sample y

and remainder data z, this gives

$$B(z|y) = \frac{\int f_1(z|\theta_1, y)\pi_1(\theta_1|y)\,\mathrm{d}\theta}{\int f_2(z|\theta_2, y)\pi_2(\theta_2|y)\,\mathrm{d}\theta}, \quad (13)$$

$$= \frac{B(x)}{B(y)}. \quad (14)$$

However, the problematic choice remains of which training sample y to use, and therefore how much of $B(x)$ to use. Various options are available, among them, a simple fixed proportion of the data, or Berger and Pericchi's [4] *intrinsic Bayes factor*, which uses all possible training samples, averaging the results. A further refinement is available through the *fractional Bayes factor* [8, 21]. Here, if the training sample y is assumed to be a given proportion b of the data, it is noted that the likelihood associated with *any* training sample is approximately $f(x|\theta)^b$, and so we get

$$B_{\text{frac}}(x) = \frac{q_1(b, x)}{q_2(b, x)}, \quad (15)$$

where

$$q_i(b, x) = \frac{\int f_i(x|\theta_i)\pi_i(\theta_i)\,\mathrm{d}\theta}{\int f_i(x|\theta_i)^b\pi_i(\theta_i)\,\mathrm{d}\theta}. \quad (16)$$

O'Hagan [21] suggests several ways of choosing b under different circumstances.

For continuous data, the problem of choosing a "general alternative" to the model being checked is more difficult, and techniques are less well advanced. Verdinelli and Wasserman [25] use as an alternative, a likelihood made up of a series of Gaussian packets, spread along the whole real line, choosing them appropriately so that the Bayes factor is *consistent*; that is, tends to ∞ or 0 depending on whether the model is true or false. Robert and Rousseau [23] restrict their attention to the interval [0, 1], postulating that if a model F_θ is not true for data X, then $F_\theta(X)$ does not have a **uniform** $U(0, 1)$ distribution, but is instead distributed as a mixture of **beta** likelihoods.

Methods for Assessing Goodness of fit in Hierarchical Bayesian Models

Assessing goodness of fit becomes more difficult where more than one level of random behavior takes place. For example, we might model pupil performance within school performance within county performance – postulating that each stage will contribute its own uncertainty. We can label the pupil performance as level I, the schools as level II, and so on. Such a model is called *hierarchical*, where ultimately, we place priors on all the parameters left unspecified in the hierarchy. (This is very similar to a **multilevel model**.) Hierarchical models have been found to have the desirable property of **robustness** to the choice of prior [14], and so we might expect that they perform well even if the model is slightly **misspecified**.

The difficulty in testing goodness of fit in hierarchical models comes because we have no direct observations beyond the first level; any data about school performance will contain uncertainty due to pupils. We do not have any data that depends only on parameters from levels II and above (known as intermediate parameters, and labelled ϕ). It is therefore difficult to define a discrepancy statistic T, which assesses how well the model fits at these levels.

One approach to this is motivated by a remark in [5]; if the likelihood and prior give conflicting information about intermediate parameters ϕ, this suggests faults in the model, in other words, lack of fit. This idea is developed by O'Hagan [22], who suggests several ways of measuring this conflict.

Another approach that gives measures of fit for individual data points is **cross-validation**, or "leave-one-out" techniques. Here the model is re-analyzed without a particular data point or points, and the fitted value for some (intermediate) parameter compared with the original estimate. Each data point (or group of points) is left out in turn; if fitted values behave erratically when some data is missing, this indicates poor model fit. However, as these methods will require rerunning the MCMC chain every time we leave a point out to assess its impact, they will almost always be too computationally burdensome to be practical. For details on cross-validation, and more tractable approximations to it, see the work of Stern and Cressie (2000) [24] its development by Marshall and Spiegelhalter (2003) [19].

A final approach used for assessing goodness of fit in Bayesian hierarchical or classical multilevel models involves *embedding* the model in some more general framework [17, 15]. If the model can be considered as a special, large case of, for example, a linear model, then goodness-of-fit **diagnostics** appropriate to the larger linear model can be used.

Embedding the model in this way may require the addition of artificial data points to reflect the hierarchical structure or prior distributions used, which is rather unsatisfactory, leading to accusations that we are fitting the problem to the diagnostics, not the diagnostics to the problem.

An early, but very useful and specifically Bayesian embedding procedure is given by Wakefield et al. [26], used to check the assumption of a normal **random effects** distribution. Here, instead of the **normal distribution** assumed to be "correct", an extension to the **multivariate *t* distribution** is used, represented as a scaled mixture of normals. If the scaling parameter corresponding to individual random effects is too small, we have evidence of poor fit of the model for that part of the data.

References

[1] Bayarri, M.J. & Berger, J. (1997). Measures of Surprise in Bayesian Analysis, Technical report number 97-46, ISDS, Duke University.

[2] Bayarri, M.J. & Berger, J. (1999). Quantifying surprise in the data and model verification (with discussion), in *Bayesian Statistics*, Vol. 6 J.M. Bernardo, ed. Oxford University Press, Oxford, pp. 53–82.

[3] Bayarri, M.J. & Castellanos, M.E. (2000). A comparison between p-values for goodness-of-fit checking, in *Bayesian Methods with Applications to Science, Policy, and Official Statistics: Selected Papers from ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis*, E.I. George, ed. pp. 1–10, unknown Conference held in Hersonissos, (Greece) as yet.

[4] Berger, J.O. & Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction, *The Journal of the American Statistical Association* **91**, 109–122.

[5] Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion), *Journal of the Royal Statistical Society, Series A, General* **143**, 383–430.

[6] Carota, C., Parmigiani, G. & Polson, N.G. (1996). Diagnostic measures for model criticism, *Journal of the American Statistical Association* **91**(434), 753–762.

[7] Chaloner, K. (1994). Chapter residual analysis and outliers in Bayesian hierarchical models, *Aspects of UNcertainty: A Tribute to D. V. Lindley*, Wiley, Chichester, pp. 149–157.

[8] Conigliani, C., Castro, J.I. & O'Hagan, A. (2000). Bayesian assessment of goodness of fit against nonparametric alternatives, *The Canadian Journal of Statistics* **28**(2), 327–342.

[9] De la Horra, J. & Rodriguez-Bernal, M.T. (1999). The posterior predictive p-value for the problem of goodness of fit, *Test* **8**, 117–128.

[10] de la Horra, J. & Rodriguez-Bernal, M.T. (2001). Posterior predictive p-values: what they are and what they are not, *Test* **10**, 75–86.

[11] Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise, *Communications in Statistics-theory and Methods* **26**(5), 1125–1143.

[12] Gelman, A. (2003). A Bayesian formulation of exploratory analysis and goodness-of-fit testing, *International Statistical Review* **71**(2), 369–382.

[13] Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion), *Statistica Sinica* **6**, 733–807.

[14] Gustafson, P. (1996). Robustness considerations in Bayesian analysis, *Statistical Methods in Medical Research* **5**(4), 357–373.

[15] Hodges, J.S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **60**(3), 497–536.

[16] Kass, R.E. & Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**(430), 773–795.

[17] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **58**(4), 619–678.

[18] Marshall, E.C. & Spiegelhalter, D.J. (2003a). Simulation-based Tests for Divergent Behaviours in Hierarchical Models, Technical report, Biostatistics group, Imperial College, London.

[19] Marshall, E.C. & Spiegelhalter, D.J. (2003b). Approximate cross-validatory predictive checks in disease mapping models, *Statistics in Medicine* **22**(10), 1649–1660.

[20] O'Hagan, A. (1991). Contribution to the discussion of 'posterior Bayes factors', *Journal of the Royal Statistical Society, Series B* **53**, 136.

[21] O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **57**, 99–138.

[22] O'Hagan, A. (2003). HSSS model criticism, in *Highly Structured Stochastic Systems*, P.J. Green, N.L. Hjort & S.T. Richardson eds. Oxford University Press; Oxford 423–453.

[23] Robert, C.P. & Rousseau, J. (2002). *A Mixture Approach to Bayesian Goodness of Fit*, Technical report number 0209, Université de Paris IX-Dauphine, Paris; CEREMADE.

[24] Stern, H. & Cressie, N. (2000). Posterior predictive model checks for disease mapping models, *Statistics in Medicine* **19**, 2377–2397.

[25] Verdinelli, I. & Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families, *Annals of Statistics* **26**(4), 1215–1241.

[26] Wakefield, J.C., Smith, A.F.M., Racine-Poon, A. & Gelfand, A.E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler, *Applied Statistics* **43**, 201–221.

[27] Weiss, R.E. (1996a). *Bayesian Model Checking with Applications to Hierarchical Models*, Technical report number 200, Department of Statistics, UCLA, Los Angeles, CA.

[28] Weiss, R.E. (1996b). An approach to Bayesian sensitivity analysis, *Journal of the Royal Statistical Society, Series B* **58**, 593–607.

[29] Weiss, R.E. & Cho, M. (1998). Bayesian marginal influence assessment, *Journal of Statistical Planning and Inference* **71**, 163–177.

KENNETH RICE

Bayesian Methods for Contingency Tables

Motivation for Bayesian Methods

There is an extensive literature on the classical analysis of **contingency tables**. Bishop, Fienberg, and Holland [7] and Agresti [1] illustrate the use of **loglinear models** to describe the **association** structure in a multidimensional contingency table. Chi-square statistics are used to examine independence and to compare nested loglinear models (*see* **Chi-square Tests**; **Chi-square Distribution**); ***P* values** are used to assess statistical significance. **Estimation** and **hypothesis testing** procedures rest on asymptotic results for multinomial random variables (*see* **Multinomial Distribution**).

Several problems with classical **categorical data analyses** motivate consideration of **Bayesian methods**. First, there is the problem of estimating cell probabilities and corresponding expected cell counts from a sparse multiway contingency table with some empty cells. It is desirable to obtain positive smoothed estimates of the expected cell counts, reflecting the knowledge that the cell probabilities all exceed zero. Second, in comparing models, there is difficulty in interpreting the evidence communicated by a test statistic's P value, as noted for instance by Diaconis and Efron [11] in the simple case of a P value for detecting association in a two-way table with large counts. This problem motivates a Bayesian approach to measuring evidence. Lastly, there is the potential bias in estimating association measures from models arrived at by classical model selection strategies, for example, for choosing the best loglinear model (*see* **Model, Choice of**). One typically uses a fitted model to estimate both an association parameter and the variability of the estimated parameter, while ignoring uncertainty in the process of arriving at the model on which estimation is based. Bayesian methods allow a user to explicitly model the uncertainty among a class of possible models by means of a **prior distribution** on the class of models, so that the posterior estimates of association parameters explicitly account for uncertainty about the "true" model on which estimates should ideally be based.

Early Bayesian Analyses of Categorical Data

Early Bayesian analyses for categorical data focused on tractable approximations for posterior distributions and measures of evidence for two-way tables. For a multinomial random variable $\{y_{ij}\}$ with cell probabilities $\{\theta_{ij}\}$, Lindley [27] considered the posterior distribution of the log contrast $\lambda = \Sigma\Sigma a_{ij} \log \theta_{ij}$, where $\Sigma\Sigma a_{ij} = 0$. In a **2 × 2 table** with cell probabilities $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$, one example of a log contrast is the log **odds ratio**

$$\lambda = \log \theta_{11} - \log \theta_{12} - \log \theta_{21} + \log \theta_{22}. \quad (1)$$

If $\{\theta_{ij}\}$ is assumed to have a Dirichlet $(\{\alpha_{ij}\})$ distribution of the form $p(\{\theta_{ij}\}) \propto \prod \theta_{ij}^{\alpha_{ij}-1}$, Lindley showed that the posterior distribution for λ is approximately normal with mean and variance given respectively by

$$\lambda^* = \sum_i \sum_j a_{ij} \log(\alpha_{ij} + y_{ij}),$$

$$v^* = \sum_i \sum_j a_{ij} (\alpha_{ij} + y_{ij})^{-1}. \quad (2)$$

Lindley used this approximation to obtain the posterior density of the log odds ratio and to develop a Bayesian statistic for testing independence in a 2×2 table (*see* **Independence of a Set of Variables, Tests of**).

The formal way of comparing models from a Bayesian perspective is by the use of **Bayes factors**. If y denotes the data and θ denotes a parameter, a Bayesian model is described by a sampling density for the data, $f(y|\theta)$, and a prior density for the parameter, $g(\theta)$. If one has two Bayesian models $M_1 : \{f_1(y|\theta_1), g_1(\theta_1)\}$ and $M_2 : \{f_2(y|\theta_2), g_2(\theta_2)\}$, then the Bayes factor in support of model 1 over model 2 is given by

$$BF_{12} = \frac{m_1(y)}{m_2(y)}, \quad (3)$$

where $m_i(y)$ is the marginal or predictive density of the data for model M_i:

$$m_i(y) = \int f_i(y|\theta_i) g_i(\theta_i) \, d\theta_i. \quad (4)$$

The Bayes factor may also be interpreted as the ratio of posterior odds of model 1 to model 2, given the data, to the corresponding prior odds.

Jeffreys [20] was one of the first to develop Bayes factors in testing for independence in a 2×2 table. Under the independence model H, the cell probabilities can be expressed as $\{\alpha\beta, \alpha(1-\beta), (1-\alpha)\beta, (1-\alpha)(1-\beta)\}$, where α and β are marginal probabilities of the table. Under the dependence hypothesis, Jeffreys expressed the probabilities as $\{\alpha\beta + \gamma, \alpha(1-\beta) - \gamma, (1-\alpha)\beta - \gamma, (1-\alpha)(1-\beta) + \gamma\}$. By assuming independent uniform $(0, 1)$ prior distributions on α, β, and γ, Jeffreys developed an approximate Bayes factor in support of the dependence hypothesis.

Suppose one observes multinomial $\{y_1, \ldots, y_t\}$ with cell probabilities $\{\theta_1, \ldots, \theta_t\}$ and total count $n = \sum_{i=1}^{t} y_i$. Good [14] noted that simple relative frequency estimates of the probabilities θ_i can be poor when data are sparse, and studied the alternative estimates $(y_i + k)/(n + kt)$, where n is the fixed total count and k is a "flattening constant". This estimate is the mean of the posterior density assuming that the $\{\theta_i\}$ have a symmetric Dirichlet distribution of form

$$p(\theta_i) \propto \prod_{i=1}^{t} \theta_i^{k-1}.$$

In practice, it may be difficult for a user to specify the Dirichlet parameter k. Good then advocated use of a prior distribution $g(k)$ for k, resulting in a prior for the $\{\theta_i\}$ of hierarchical form

$$g(\{\theta_i\}) = \int_0^\infty \frac{\Gamma(tk)}{(\Gamma(k))^t} \prod_{i=1}^{t} \theta_i^{k-1} g(k)\, \mathrm{d}k. \qquad (5)$$

As will be seen later, Good also used this form of a mixture of symmetric Dirichlet distributions to develop Bayes factors for testing independence in contingency tables.

Bayesian Smoothing of Contingency Tables

One difficulty with large contingency tables is that observed sampling zeros in some cells may lead to poor estimates of the underlying cell probabilities. One *ad hoc* adjustment is to add 1/2 to each observed count, as in Good's [14] approach with prespecified flattening constant $k = 1/2$. Fienberg and Holland [13] were interested in developing better estimates for cell probabilities in these tables with sparse counts. They first considered the conjugate Dirichlet model $g(\{\theta_i\} \propto \prod \theta_i^{K\lambda_i - 1}$ as a prior for the cell probabilities, where λ_i is the prior mean of θ_i and K is a precision parameter. The use of this conjugate prior results in the posterior mean estimate

$$\hat{\theta}_i = \left(\frac{n}{n+K}\right)\frac{y_i}{n} + \left(\frac{K}{K+n}\right)\lambda_i. \qquad (6)$$

Since the hyperparameter K is unknown, Fienberg and Holland [13] developed an **empirical Bayes** estimator. For fixed $\{\lambda_i\}$, they showed that the risk of $\hat{\theta}_i$, under squared error loss, is equal to

$$R(\hat{\theta}, \theta) = \left(\frac{n}{(n+K)}\right)^2 (1 - ||\theta||^2) + \left(\frac{K}{(n+K)}\right)^2 n||\theta - \lambda||^2. \qquad (7)$$

The value of K that minimizes this risk is $\hat{K} = (1 - ||\theta||^2)/(||\theta - \lambda||^2)$. If one replaces K with the estimate $\hat{K}$ in the expression for $\hat{\theta}_i$, one obtains the empirical Bayes estimate

$$\theta_i^* = \left(\frac{n}{n+\hat{K}}\right)\frac{y_i}{n} + \left(\frac{\hat{K}}{\hat{K}+n}\right)\lambda_i. \qquad (8)$$

Fienberg and Holland [13] showed that the estimates $\{\theta_i^*\}$ had good risk properties relative to the **maximum likelihood** estimates $\{y_i/n\}$. In practice, one can choose the prior means $\{\lambda_i\}$ to reflect one's prior beliefs, or choose data-dependent values for $\{\lambda_i\}$ based on the estimated expected counts from a log-linear model. For example, one might in this way shrink estimates towards the independence model for two-way contingency tables, or towards conditional independence or the no-three-way **interaction model** for three-way tables (*see* **Shrinkage Estimation**).

A number of alternative fully Bayesian methods have been proposed for smoothing contingency table counts. One approach is based on normal prior distributions placed on components of a logit representation of a cell probability (*see* **Binary Data**). For a two-way table with counts $\{y_{ij}\}$ and cell probabilities $\{\theta_{ij}\}$, Leonard [24] defines the multivariate logit

$$\gamma_{ij} = \text{logit}\ \theta_{ij} = \log \theta_{ij} + D(\theta), \qquad (9)$$

where $D(\theta)$ is chosen to ensure that the probabilities sum to one. The logit is decomposed as

$$\gamma_{ij} = \alpha_i + \beta_j + \lambda_{ij}, \qquad (10)$$

where the terms correspond respectively to a row effect, a column effect, and an **interaction** effect in the two-way table. To model the belief that the set of row effects $\{\alpha_i\}$ is **exchangeable**, Leonard [24] uses the two-stage prior

1. $\alpha_1, \ldots, \alpha_I$ are independent $N(\mu_\alpha, \sigma_\alpha^2)$;
2. $\mu_\alpha, \sigma_\alpha^2$ independent, with μ_α having a vague flat prior and $\nu_\alpha \tau_\alpha \sigma_\alpha^{-2}$ distributed chi-squared with ν_α degrees of freedom (*see* **Exchangeability**). The hyperparameter τ_α represents a prior estimate at σ_α^2 and ν_α measures the sureness of this prior guess.

Similar exchangeable prior distributions are placed on the sets of column effects $\{\beta_j\}$ and interaction effects $\{\lambda_{ij}\}$. Leonard [24] used this model to find posterior modal estimates of the probabilities. When the interaction effects are set equal to zero, these Bayesian estimates smooth the table towards an independence structure. Nazaret [29] extended this multivariate logit representation to three-way tables.

Albert and Gupta [5] and Epstein and Fienberg [12] perform similar smoothing using mixtures of conjugate priors. To model the belief that the cell probabilities satisfy an independence structure, Albert and Gupta [5] assign the $\{\theta_{ij}\}$, a Dirichlet distribution with precision parameter K and prior cell means $\{\lambda_{ij}\}$ satisfying the independence structure $\lambda_{ij} = \lambda_{i+}\lambda_{+j}$, with $\lambda_{i+}, \lambda_{+j}$ respectively the prior row and column marginal means: $\lambda_{i+} = \sum_j \lambda_{ij}, \lambda_{+j} = \sum_i \lambda_{ij}$. At the second-stage of the prior, these marginal prior means $\{\lambda_{i+}\}$ and $\{\lambda_{+j}\}$ are assigned vague **uniform distributions**. The posterior mean of the cell probability θ_{ij} can be expressed as

$$\left(\frac{n}{n+K}\right)\frac{y_{ij}}{n} + \left(\frac{K}{K+n}\right)\mathrm{E}(\lambda_{i+}\lambda_{+j}), \qquad (11)$$

where $\mathrm{E}(\cdot)$ denotes the expectation over the posterior distribution of $\{\lambda_{i+}\,\lambda_{+j}\}$. This estimate is a compromise between the usual unconditional maximum likelihood estimate y_{ij}/n and the estimate under an independence model. Epstein and Fienberg [12] generalized this conjugate approach by modeling the prior mean of θ_{ij} by a logit model. Laird [23] and Knuiman and Speed [22] perform Bayesian smoothing by applying a **multivariate normal** prior to the vector of logarithms of expected cell counts. (King and Brooks [21] show that a multivariate normal prior on the loglinear model parameters induces a multivariate lognormal prior on the expected cell counts of the contingency table.)

Bayesian Interaction Analysis

Leonard and Novick [26] describe the use of a Bayesian **hierarchical model** to explore the interaction structure of a two-way contingency table. Given cell counts $\{y_{ij}\}$ from independent **Poisson distributions** with respective means $\{\theta_{ij}\}$ they assume, at the first stage, that the θ_{ij} have independent **Gamma distributions** with respective means ξ_{ij} and precision parameter α. The means ξ_{ij} are presumed to satisfy the independence structure $\log \xi_{ij} = \mu + \lambda_i^A + \lambda_j^B$. At the second stage of the prior, all unknown parameters are given vague distributions. The posterior distribution of the precision parameter α is informative about the **goodness of fit** of the independence model. In addition, Leonard and Novick [26] consider the posterior distributions of the "parametric **residuals**" $\{\log \theta_{ij} - \log \xi_{ij}\}$ to explore the dependence pattern in the table. Leonard, Hsu, and Tsui [25] consider an alternative interaction analysis based on a nonhierarchical prior. They obtain approximations to the joint posterior distribution of $\{\theta_{ij}\}$, and dependence in the table is studied by considering the posterior distribution of the interaction parameters $\{\theta_{ij} - \theta_{i+} - \theta_{+j} + \theta_{++}\}$, with $\theta_{i+} = \sum_j \theta_{ij}$, $\theta_{+j} = \sum_i \theta_{ij}$, and $\theta_{++} = \sum_i \sum_j \theta_{ij}$.

Bayesian Tests of Equiprobability and Independence

I. J. Good, in a large series of papers, developed Bayes tests for contingency tables under a variety of sampling models. We illustrate the general approach by considering Good's construction of a significance test for equiprobability of a multinomial probability vector [15]. As usual, one observes multinomial $\{y_1, \ldots, y_t\}$ with cell probabilities $\{\theta_1, \ldots, \theta_t\}$ and total count $n = \sum_{i=1}^t y_i$. The hypothesis of interest is $H : \theta_i = 1/t, i = 1, \ldots, t$, and the usual classical test statistic is Pearson's chi-square,

$$X^2 = \sum_{i=1}^{t} \frac{(y_i - \mathrm{E}(y_i))^2}{\mathrm{E}(y_i)} = \frac{t}{n}\Sigma\left(y_i - \frac{n}{t}\right)^2, \qquad (12)$$

which is asymptotically distributed as chi-square with $t - 1$ degrees of freedom.

To develop a Bayes factor to test H against the alternative hypothesis $\overline{H}: \theta_i \neq 1/t$ for some $i = 1, \ldots, t$, note that the density of the data $\{y_1, \ldots, y_t\}$ is fully specified under the hypothesis H. Let H_k denote the hypothesis that the $\{\theta_i\}$ have a symmetric Dirichlet (k) distribution with density $g(\theta) \propto \prod \theta_i^{k-1}$. The hyperparameter k, since it is difficult to specify, is given a log Cauchy prior

$$g(k) = \left(\frac{1}{\pi k}\right)\left(\frac{\lambda}{\lambda^2 + \left(\frac{\log k}{\mu}\right)^2}\right) \tag{13}$$

(*see* **Cauchy Distribution**). Then the Bayes factor BF of $\overline{H}$ against H is given by $BF = \mathrm{E}(BF(k)|\lambda, \mu) = \int BF(k)g(k)\,\mathrm{d}k$, where $BF(k)$ is the Bayes factor of H_k against equiprobability,

$$BF(k) = \frac{t^n \Gamma(tk) \prod_{i=1}^{t} \Gamma(y_i + k)}{\Gamma(k)^t \Gamma(n + tk)}. \tag{14}$$

To illustrate and compare Bayesian and classical measures of evidence, we consider 150 voters of whom 61, 53, and 36 expressed preferences for candidates 1, 2, and 3, respectively. The chi-square statistic for equiprobability is $X^2 = 6.52$ with an associated P value of 0.0384. The following table gives values of the Bayes factor of H_k against H for a range of values of the Dirichlet parameter k. A Dirichlet prior with

k	0.1	0.5	1	2	10	100	1000
$BF(k)$	0	0.2	0.5	0.9	2.8	2.0	1.1

a large value of k, say $k = 1000$, concentrates virtually all of its probability very near the equiprobability hypothesis. The marginal density under $\overline{H}$ is then the data density averaged over models very close to H, and hence only slightly exceeds the density under H. Thus, the Bayes factor only slightly exceeds one. In contrast, Dirichlet priors with very small values of k become increasingly vague. The marginal density from such a vague prior gives roughly equal weights to data densities calculated throughout the parameter space, including parameter regions with which the data are much less compatible than with H. For such values of k, the marginal density under H is thus much higher than under H_k, and the resulting Bayes factor is low; equivalently, the odds of H_k relative to H given the data are much reduced. In fact, the Bayes factor is not defined by the use of an improper prior with $k = 0$.

Note that for the observed data the maximum value of $BF(k)$ in the table is 2.8; for a wide range of log Cauchy priors that might be used in practice, BF will be smaller than 1. Consequently, in contrast with the classical result, the Bayesian measures indicate that there is little evidence against equiprobability in these data. Generally, when testing a point **null hypothesis**, a classical P value overstates the evidence against the hypothesis compared to a Bayes factor test statistic. [6].

Good [16], Crook and Good [8], and Good and Crook [17] extended the above methodology in developing Bayes tests for two-way contingency tables. Tests were constructed on the basis of mixtures of conjugate priors for the three sampling models (multinomial, product-multinomial, multivariate hypergeometric) corresponding respectively to fixed overall table total n or to fixed totals along one or both marginal dimensions (*see* **Hypergeometric Distribution**). An objective of this analysis was to assess whether the marginal totals convey any evidence for or against independence of rows and columns. In the multinomial sampling situation (model 1) where only the total count is fixed, the Bayes factor in support of the dependence hypothesis $\overline{H}$ over the independence hypothesis H is given by

$$BF_1 = \frac{P(\{y_{ij}\}|\overline{H})}{P(\{y_{ij}\}|H)}, \tag{15}$$

where $P(\{y_{ij}\}|\overline{H})$ and $P(\{y_{ij}\}|H)$ are the marginal probabilities of the data under the hypotheses $\overline{H}$ and H, respectively. The marginal probability under $\overline{H}$ is computed using a prior on the vector of cell probabilities $\{\theta_{ij}\}$ that is a mixture of symmetric Dirichlet distributions, and the probability under H is computed by placing similar Dirichlet mixtures as priors on the vectors of marginal cell probabilities $\{\theta_i\}$ and $\{\theta_j\}$. Good and Crook also developed Bayes factors against independence in the situations where either the row or column totals were fixed (model 2), and in the situation where both row and column totals were fixed (model 3). The evidence provided by the row and column totals alone is defined to be the ratio $FRACT = BF_1/BF_3$ of the Bayes factors under model 1, in which information about rows and columns is observed, and model 3, in which

row and column margins are arbitrarily fixed. One conclusion from their studies was that $FRACT$ is usually between 0.5 and 2.5, indicating that the row and column totals typically contain a modest amount of evidence against independence. These Bayesian measures have the advantage that they explicitly allow for the sampling model and do not depend on asymptotic theory. In addition, they can be used as classical test statistics against independence, and tests based upon them can be shown to possess good **power**. Gunel and Dickey [19], and Albert [2] also develop Bayes factors for two-way tables based on conjugate priors and mixtures of conjugate priors, respectively.

Bayes Factors for GLM's with Application to Loglinear Models

Raftery [32] presents a general approach for testing within **generalized linear models**, with direct applications to comparisons of loglinear models for multiway contingency tables. Recall that if one observes data D and has two possible models M_1 and M_2, then the evidence in support of the model M_1 is given by the Bayes factor $BF_{12} = (P(D|M_1)/P(D|M_2))$, where $P(D|M_k)$ is the marginal probability

$$P(D|M_k) = \int P(D|\theta_k, M_k)\, p(\theta_k|M_k)\, \mathrm{d}\theta_k. \quad (16)$$

Raftery presents several methods for approximating $P(D|M_k)$. By the Laplace method for integrals, one has the approximation

$$P(D|M_k) \approx (2\pi)^{p_k/2} |\Psi_k|^{1/2} P(D|\tilde{\theta}_k, M_k)\, p(\tilde{\theta}_k|M_k),$$

where $\tilde{\theta}_k$ is the posterior mode, Ψ_k is the inverse of the negative of the Hessian matrix of $\log P(D|\theta_k, M_k)\, p(\theta_k|M_k)$ evaluated at the mode, p_k is the number of parameters of the model, and $|A|$ indicates the determinant of the matrix A (*see* **Matrix Algebra**).

The following approximation was developed as an alternative that capitalizes on quantities available from standard generalized linear model software:

$$2 \log BF_{12} \approx \chi^2 + (E_1^* - E_2^*),$$
$$E_k^* = -\log |I_k| + 2 \log P(\hat{\theta}_k|M_k) + p_k \log(2\pi), \quad (17)$$

where $\hat{\theta}_k$ and I_k are respectively the maximum likelihood estimate and observed **information matrix** from fitting the model M_k, and X^2 is the classical χ^2 "drop in deviance"test statistic for comparing the two models M_1 and M_2. One important issue is the choice of prior on the **regression** coefficients. Raftery [32] discusses suitable "vague" choices of hyperparameters to use in a testing situation.

To illustrate the use of Bayes factors in model choice, Raftery [32] considers the data shown below from a **case–control study** in which oral contraceptive histories were compared between groups of women having suffered myocardial infarction and control women who had not [34]. The table shows a cross-classification of case (Myocardial infarction) or control status (M), Age category (A), and history of oral Contraceptive use (C).

Suppose we wish to compare the two loglinear models M_1 and M_2, where M_1 denotes the no three-way interaction model indicating that the **relative risk** relating disease and oral contraceptive history (estimated in this context by the odds ratio) is constant across age groups, and the more complicated model M_2 indicating that the relative risk is constant from ages 25 to 34 but may shift to a different constant during ages 35 to 49. Using a classical loglinear analysis, the difference in deviances is 4.7 on one degree of freedom and the P value is 0.03, indicating different relative risks for the age groups 25 to 34 and 35 to 49. Computation of a Bayes factor, in contrast, slightly favors the simpler model M_1.

One advantage of this Bayesian approach is that it can explicitly allow for model uncertainty in the

	Age group (A)									
	25–29		30–34		35–39		40–44		45–49	
	Myocardial infarction (M)									
Oral contraceptives (C)	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Not used	224	2	390	12	330	33	362	65	301	93
Used	62	4	33	9	26	4	9	6	5	6

estimation of parameters of interest. Where there are many possible models $\{M_1, \ldots, M_K\}$, the posterior distribution of the parameter θ may be expressed as the mixture of posteriors:

$$p(\theta|D) = \sum_{i=1}^{K} p(\theta|M_k, D)p(M_k|D), \qquad (18)$$

where $p(\theta|M_k, D)$ is the posterior of θ under model M_k and $p(M_k|D) \propto P(M_k)P(D|M_k)$ is the posterior probability of model M_k. Using the above example, Raftery [32] illustrates this "model averaging" approach in estimating a relative risk parameter when there are two plausible models; see [33] for additional applications of this method.)

Use of BIC in Sociological Applications

If one ignores terms of order $O(1)$ or smaller, one gets a further approximation to the log marginal density of the data under model M_k:

$$\log P(D|M_k) \approx \log P(D|\hat{\theta}) - \frac{p_k}{2}\log n \qquad (19)$$

where, as above, p_k is the number of parameters in model M_k [31]. Twice the difference between values of this approximation for a reduced versus a saturated model is the BIC (Bayesian Information Criterion) measure for assessing the overall fit of a model M_k:

$$BIC_k = L_k^2 - \mathrm{d}f_k \log n, \qquad (20)$$

where L_k^2 is the usual deviance statistic and df_k is the associated degrees of freedom of this statistic. Two models M_j and M_k can be compared by the difference $BIC_k - BIC_j$. Raftery [31] gives tables helpful for interpreting the significance of a computed BIC value and comparing it with traditional P values. Specifically, a BIC measure gives precise guidelines on how one should adjust a significance level pertaining to model comparisons as the sample size increases, in order to avoid including trivial complexity in a final model.

Bayesian Model Search for Loglinear Models

Recent advances in Bayesian computing have increased interest in using Bayesian models to search for the "best" loglinear model for a multidimensional contingency table. Madigan and Raftery [28] define some general principles that should be expressed in the behavior of any model selection strategy. One principle is that models that predict the data far less well than the best model should be discarded; that is, models M_k such that $(\max_l P(M_l|D))/P(M_k|D) \geq C$ should be removed from consideration. A second principle, "Occam's Razor", states that if two models predict the data equally well, the simpler should be preferred (*see* **Parsimony**). Madigan and Raftery [28] describe how one can search through the space of models by use of Bayesian posterior model probabilities. In this approach, a model is represented by a directed graph with a node for each variable, and the dependence structure is represented by edges connecting pairs of nodes. Hyper-Dirichlet priors are used to represent prior opinion about the model parameters [9].

Albert [3, 4] describes Bayesian model selection based on priors placed directly on terms of the loglinear model. For a three-way table with a saturated loglinear model represented by

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \qquad (21)$$

Albert [3] places a multivariate normal prior directly on the vector $(u, \boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_{12}, \boldsymbol{u}_{13}, \boldsymbol{u}_{23}, \boldsymbol{u}_{123})'$ formed by stringing out the sets of u-terms, and models are defined by means of priors that constrain sets of u-terms to zero. The Laplace method [35] is used to compute the model probabilities.

To illustrate the use of a Bayesian model selection strategy, we use the table below, which classifies test rats in an experiment with respect to dose of a possible carcinogen (D), time of death (sacrificed at 132 weeks or age, or prematurely) (T), and presence or absence of cancer at necropsy (C) [13].

	Time of death (T)			
	Premature		At sacrifice	
	Dose (D)			
Cancer (C)	Low	High	Low	High
Present	4	7	0	7
Absent	26	16	14	14

A classical stepwise model search leads to the choice of the model [T], [CD], which indicates that cancer and dose are independent of survival to the end of the experiment. In [3], a Bayesian model search is performed over $2^4 = 16$ models consisting of combinations of the presence or absence of each of the four sets of interaction terms $u_{DT}, u_{DC}, u_{CT}, u_{DTC}$. The table below gives the posterior model probabilities for the six models with the largest values. This Bayesian analysis is similar to that of the classical analysis in the sense that the model with the highest posterior probability includes the interaction u_{CD} and no other interactions, but the probability of this model is only 0.39 and the remaining models have relatively small posterior probabilities. Bayesian estimates of the association between dose and cancer in the table will account for the uncertainty of the "best" loglinear model.

Dellaportas and Forster [10]) propose a **simulation**-based approach for finding the best loglinear model. In this approach, they assume that the vector of logarithms of the expected cell counts has a multivariate normal prior distribution. They work with a parameterization under which all parameters are identifiable and linearly independent, and choose vague priors for these parameters. For searching through the model space, they propose a strategy based on the reversible jump **Markov chain Monte Carlo** (MCMC) algorithm [18]. An attractive feature of this algorithm is that one can move between models of different dimension. These authors apply their model selection strategy, and those of [32] and [28], to finding the best model for a three-way contingency table. Although there are differences in the computed posterior model probabilities, all of these approaches select the same loglinear model. In the normal regression context, George and McCulloch [17] propose an alternative Bayesian algorithm, Stochastic Search Variable Selection (SSVS), for searching through the space of all possible models. Ntzoufras et al. [30] extend the SSVS approach to loglinear modeling.

Interactions included	Posterior probability
None	0.05
u_{CD}	0.39
u_{CA}, u_{CD}	0.07
u_{CD}, u_{ACD}	0.09
u_{CD}, u_{AD}	0.18
u_{CA}, u_{CD}, u_{AD}	0.04

Much of the early Bayesian methodology for contingency tables was devoted to issues regarding computation due to the difficulties in computing integrals of several variables. However, by virtue of great advances in computing posterior distributions by simulation, it is now possible to fit sophisticated Bayesian models for high-dimensional contingency tables. Further, Bayesian advances may be expected, especially with respect to criticism of single loglinear models, and model selection among large classes of hierarchical and graphical models (*see* **Hierarchical Models**).

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, New York.

[2] Albert, J.H. (1989). A Bayesian test for a two-way contingency table using independence priors, *Canadian Journal of Statistics* **18**, 347–363.

[3] Albert, J.H. (1996). Bayesian selection of log-linear models, *Canadian Journal of Statistics* **24**, 327–347.

[4] Albert, J.H. (1997). Bayesian testing and estimation of association in a two-way contingency table, *Journal of the American Statistical Association* **92**, 685–693.

[5] Albert, J.H. & Gupta, A.K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables, *Annals of Statistics* **10**, 1261–1268.

[6] Berger, J.O. & Delampady, M. (1987). Testing precise hypotheses, *Statistical Science* **2**, 317–335.

[7] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

[8] Crook, J.F. & Good, I.J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables: Part II., *Annals of Statistics* **8**, 1198–1218.

[9] Dawid, A.P. & Lauritzen, S.L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models, *Annals of Statistics* **21**, 1272–1317.

[10] Dellaportas, P. & Forster, J.J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models, *Biometrika* **86**, 615–633.

[11] Diaconis, P. & Efron, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square Statistic, *Annals of Statistics* **13**, 845–874.

[12] Epstein, L.D. & Fienberg, S.E. (1992). Bayesian estimation in multidimensional contingency tables, in *Bayesian Analysis in Statistics and Economics*, P.K. Goel & N.S. Iyengar, eds. Springer-Verlag, New York, 27–42.

[13] Fienberg, S.E. (1980). Using loglinear models to analyze cross-classified categorical data, *The Mathematical Scientist* **5**, 13–30.

[14] Good, I.J. (1965). *The Estimation of Probabilities*. M.I.T. Press, Cambridge, MA.

[15] Good, I.J. (1967). A Bayesian significance test for multinomial distributions, *Journal of the Royal Statistical Society, B* **29**, 399–431.

[16] Good, I.J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, *Annals of Statistics* **4**, 1159–1189.

[17] Good, I.J. & Crook, J.F. (1987). The robustness and sensitivity of the mixed-dirichlet Bayesian test for "independence" in contingency tables, *Annals of Statistics* **15**, 670–693.

[18] Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Journal of the Royal Statistical Society B* **82**, 711–732.

[19] Gunel, E. & Dickey, J.M. (1974). Bayes factors for independence in contingency tables, *Biometrika* **61**, 545–557.

[20] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Oxford University Press, Oxford.

[21] King, R. & Brooks, S.P. (2001). Prior induction in log-linear models for general contingency table analysis, *The Annals of Statistics* **29**, 715–747.

[22] Knuiman, M.W. & Speed, T.P. (1988). Incorporating prior information into the analysis of contingency tables, *Biometrics* **44**, 1061–1071.

[23] Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables, *Biometrika* **65**, 581–590.

[24] Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables, *Journal of the Royal Statistical Society B* **37**, 23–37.

[25] Leonard, T., Hsu, J.S.J. & Tsui, K.W. (1989). Bayesian marginal inference, *Journal of the American Statistical Association* **84**, 1051–1058.

[26] Leonard, T. & Novick, M.R. (1986). Bayesian full rank marginalization for two-way contingency tables, *Journal of Educational Statistics* **11**, 33–56.

[27] Lindley, D.V. (1964). The Bayesian analysis of contingency tables, *Annals of Mathematical Statistics* **35**, 1622–1643.

[28] Madigan, D. & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of the American Statistical Association* **89**, 1535–1546.

[29] Nazaret, W.A. (1987). Bayesian log-linear estimates for three-way contingency tables, *Biometrika* **74**, 401–410.

[30] Ntzoufras, I., Forster, J.J. & Dellaportas, P. (2000). Stochastic search variable selection for log-linear models, *Journal of Statistical Computation and Simulation* **68**, 23–38.

[31] Raftery, A.E. (1994). Bayesian model selection in social research, in *Sociological Methodology 1995*, P.V. Marsden, ed. Blackwell Publishing, Cambridge, MA, pp. 111–163.

[32] Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models, *Biometrika* **83**, 251–266.

[33] Raftery, A.E. & Richardson, S. (1996). Model selection for generalized linear models via GLIB, with application to nutrition and breast cancer, in *Bayesian Biostatistics*, D.A. Berry & D.K. Stangl, eds. Marcel Dekker, New York, pp. 321–353.

[34] Shapiro, S., Slone, D., Rosenberg, L., Kaufman, D.W., Stolley, P.D. & Miettinen, O.S. (1979). Oral-contraceptive use in relation to myocardial infarction, *Lancet* **313**, 743–747.

[35] Tierney, L. & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**, 82–86.

Further Reading

Dellaportas, P., Forster, J.J. & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC, *Statistics and Computing* **12**, 27–36.

Evans, M., Gilula, Z. & Guttman, I. (1993). Computational issues in the Bayesian analysis of categorical data: log-linear and Goodman's RC model, *Statistica Sinica* **3**, 391–406.

Fienberg, S.E. & Holland, P.W. (1973). Simultaneous estimation of multinomial cell probabilities, *Journal of the American Statistical Association* **68**, 683–689.

George, E.I. & McCulloch, R.E. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association* **88**, 881–889.

Giudici, P. (1998). Smooth sparse contingency tables: a graphical Bayesian approach, *Metron* **56**, 171–187.

Leonard, T. (1993). The Bayesian analysis of categorical data – a selective review, *Aspects of Uncertainty (A Tribute to D. V. Lindley)*, John Wiley and Sons, New York, pp. 283–310.

Madigan, D. & York, J.C. (1995). Bayesian graphical models for discrete data, *International Statistical Review* **63**, 215–232.

Raftery, A.E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information, *Journal of the Royal Statistical Society B* **48**, 249–250.

Spiegelhalter, D.J. & Smith, A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information, *Journal of the Royal Statistical Society B* **44**, 377–387.

(*see also* **Bayesian Measures of Goodness of Fit**; **Bayesian Methods for Model Comparison**)

JAMES H. ALBERT

Bayesian Methods for Model Comparison

This article aims to give a brief overview of methods for comparing competing Bayesian models. This is a large topic of ongoing research interest; for a more thorough review see Key, Pericchi, and Smith or Bernardo and Smith [2, Chapter 6]. Pertinent discussion can also be found in the entries on **Bayes Factors**, **Bayesian measures of goodness of fit**, **Lindley's paradox**, and **choice of model**.

Bayes Factors and Problems Associated with Them

The basic tools of Bayesian model comparison are easily derived from a **decision-theoretic** viewpoint. Suppose we have a series of models M_i, to which we assign **prior** probabilities $P(M_i)$. (For the purposes of this article, "model" includes the distribution of the data and all prior structures associated with parameters.) If we assume the simple **utility** function that scores one for a correct choice of model and zero otherwise, then for data x, it follows simply that the optimal decision is to choose the model with highest posterior probability $p(M_i|x)$. When comparing two models, it is therefore natural to consider

$$\frac{p(M_i|x)}{p(M_j|x)} = \frac{p(x|M_i)}{p(x|M_j)} \times \frac{P(M_i)}{P(M_j)}, \qquad (1)$$

and so given our choice of prior P, the Bayes factor $p(x|M_i)/p(x|M_j)$ gives a straightforward measure for model comparison. See the **Bayes Factor** entry for more details on how to calibrate and interpret Bayes factors.

For a prespecified finite collection of competing models and proper priors, Bayes factors are simple to interpret and calculate. They also follow the **Likelihood Principle** (see e.g. Bernardo and Smith [2, p. 454]) and are unaffected by integrating out nuisance levels of a hierarchical model. But they cannot be defined if we use improper priors and are unstable when using diffuse priors; these properties are described and various modifications proposed in the entry on **Bayesian measures of goodness of fit** or Berger and Pericchi [1]. For further discussion, and a review of techniques used for numerical computation of Bayes factors, see Carlin and Louis [3, pp. 206–219] and Han and Carlin [5].

Lindley's Paradox

Lindley's paradox is a problem related to the stability of Bayes factors for diffuse priors [9]. If the observed data are not "close" to the alternative models, the use of Bayes factors may lead us to accept a hypothesis/model that is rejected by classical **hypothesis tests**. The most famous example of this comes when we have a sample of **normal** data $x_1, x_2, \ldots$. We assume that all X_i are independently distributed $N(\mu, \sigma_0^2)$ for some known σ_0 and wish to compare the models

$$H_0\colon\ \mu = \mu_0 \quad vs \quad H_1\colon\ \mu \neq \mu_0, \qquad (2)$$

where μ_0 is known explicitly. Whenever the sample mean $\overline{x}$ is too far from μ_0, the classical framework rejects H_0. Let us now compare a similar pair of Bayesian models, using the same assumptions that $X_i \sim N(\mu, \sigma_0^2)$, with

$$M_0\colon\ \mu = \mu_0 \quad or \quad M_1\colon\ \mu \sim N(\mu_1, \sigma_1^2), \qquad (3)$$

for some known μ_1, σ_1. Lindley [9] showed that if we let σ_1 become sufficiently large, so that our prior information on μ is extremely diffuse, then use of Bayes factors will lead us to accept M_0 *whatever* the value of $\overline{x}$, even if it is far from μ_0, therefore contradicting the classical approach.

The potentially *huge* diffuseness in the prior is the key to unlocking the "paradox" here. The Bayes factor chooses between two models: one centered around μ_0, which may be a poor reflection of the data, and the extremely diffuse alternative that may give *even less* support to the data, owing to its being so thinly spread. This is a different choice to that posed in the classical analysis, and so different conclusions may be expected. The effect should be taken into account whenever sharp, point values for parameters are compared with diffuse alternative formulations.

Avoiding the Lindley Paradox

Several authors have questioned the value of testing "point" models or hypotheses, as reviewed in Kass and Raftery [6]. However, given the vast frequentist literature on this subject, there is a clear

motivation for Bayesians to seek priors, which are "reasonable" for the alternative model M_1 above. We want to specify a diffuse, weakly informative prior without assuming so little information that we run into Lindley's paradox. A neat solution is given by Kass and Wasserman [10], who advocate setting $\sigma_1 = \sigma_0$ above; this gives a prior that contains the same amount of information as we get from a single observation. They generalize this for multivariate parameters and show that for nested hypotheses (*see* **Hierarchical Models**) the resulting Bayes factors are asymptotically equivalent to the BIC (Bayesian Information Criterion) [11] (or Schwarz Criterion), written as

$$BIC = 2 \log f(x|\hat{\theta}) - 2p \log n \tag{4}$$

for likelihood f, maximum likelihood parameter estimate $\hat{\theta}$, a total of p parameters, and sample size n. This is similar to the work of Smith and Spiegelhalter [12], who derive priors for parameters in nested linear models that are shown to be equivalent to the BIC and the related AIC (Akaike Information Criterion).

The Bayes factor's clear definition makes it desirable to interpret any (sensible) model comparison method as a Bayes Factor for some particular set of prior beliefs, akin to the decision-theoretic results that all admissible decision rules are Bayes rules for some prior. This attractive unifying principle is, however, only helpful for finite numbers of competing models, where we additionally assume that one of the models is correct. Ordinarily, we are only looking for a "best" model, or perhaps just an adequate one, not a strictly "true" one, and this increases the subjectivity in defining prior probabilities for the different models under comparison. See the $\mathcal{M}$-open $\mathcal{M}$-closed discussion by Bernardo and Smith [2, pp. 384–385] for further discussion. In reality, we might additionally expect some models to be motivated by the data itself, further complicating the reasonably simple Bayes factor framework.

Model Averaging, RJMCMC, Utility Discussion

Before exploring Bayesian model comparison techniques beyond the Bayes factor, it is important to ask whether model comparison is to be just one analytical step toward some other ultimate goal, like parameter **estimation** or **prediction** of future behavior. If so, then perhaps the most "fully" Bayesian method for model choice is *not* to choose one model "as if true". Then, following Key et al. [7], we modify all the competing models, and for a decision about some parameter with true value ω, choose the action a that maximizes

$$\int u(a, \omega) \sum_i p_i(\omega|x, M_i)\, p_i(M_i|x)\, \mathrm{d}\omega, \tag{5}$$

where $u(a, \omega)$ is the utility obtained by action a for ω. The summation term here indicates that we are model averaging. In principle, as the amount of data present increases, the choice a should get arbitrarily close to the true parameter value, which is of course equivalent to choosing the correct model. However, this approach neatly produces the analysis we are really interested in without having to make a strict model choice at any stage. An extremely similar approach is used when the choice of model is considered a parameter value, where we subsume all competing models into one large hierarchical model; see the section on Reversible Jump in the **Markov Chain Monte Carlo** entry.

The dependence on utility functions in the model averaging above is clear and essential, and this carries over to other forms of model comparison; if we are to compare models rigorously there must be some specification of the *purpose* for which we are comparing them. For further references, see Draper's discussion of Spiegelhalter et al. [13]. If a utility function can be established, the decision-theoretic framework simply dictates that we use the model that maximizes expected utility. Nonetheless, while clearly desirable in terms of clarifying and simplifying the analysis, deciding on a utility function is difficult and extremely subjective, just like choosing a prior. We therefore consider methods that do not require specific utility functions and instead are based around some measure of comparative goodness of fit to the data.

Some Alternatives to Bayes Factors

Deviance Information Criterion (DIC)

Classical techniques for comparing (nested) models are based around the models' deviances and their numbers of parameters, or degrees of freedom. Bayesian analogues of these quantities are combined in the Deviance Information Criterion (DIC) of

Spiegelhalter et al. [13]. DIC is defined as

$$DIC = 2p_D + D(\overline{\theta}), \tag{6}$$

where

$$p_D = \mathbb{E}_{\theta|y}[-2\log\{p(y|\theta)\}] + 2\log[p\{y|\tilde{\theta}(y)\}] \tag{7}$$

evaluated at $\tilde{\theta}(y) = \mathbb{E}(\theta|y)$ is a measure of the "effective number of parameters" and $D(\overline{\theta})$ is a measure of the "Bayesian Deviance"

$$D(\theta) = -2\log\{p(y|\theta)\} + 2\log\{f(y)\}, \tag{8}$$

again evaluated at the posterior mean $\mathbb{E}(\theta|y)$. We seek **parsimonious** models, which combine low deviance with the minimum number of parameters, and so models that minimize DIC are preferred over alternatives.

DIC has been used in many analyses but further innovation seems possible; negative values of p_D are possible, different parameterizations lead to different values of DIC, and for **multilevel models**, it is not clear which levels of the model should be integrated out (e.g. **random effects**) and which left "in focus". Trevisani and Gelfand [14] show that integrating out different sets of **nuisance parameters** can lead to varying support for the same underlying hierarchical model. However, this applies to other **likelihood**-based criteria, like AIC or BIC, and is not a specifically Bayesian problem.

Although in most practical applications it will be unrealistically optimistic to assume that one of the competing models is "true", in this hypothetical situation, we might prefer methods that, asymptotically, select the correct model. As with AIC, it is possible to find situations in which DIC fails to do this.

Expected Posterior Deviance (EPD)

Another criterion-based approach is the use of Expected Posterior Deviance (EPD), developed initially by Laud and Ibrahim [8] and developed more formally by Gelfand and Ghosh [4]. Here we work with the posterior predictive distribution for new data;

$$f(x_{\text{new}}|x) = \int f(x_{\text{new}}|\theta)\, p(\theta|x)\, \mathrm{d}\theta. \tag{9}$$

The user must choose a discrepancy function $d(x_{\text{new}}, x)$, generally taken to be the (classical) deviance measure, although any reasonable measure of distance between the data sets is acceptable. Then the model that minimizes

$$EPD = \mathbb{E}(d(x_{\text{new}}, x)|x, M_i) \tag{10}$$

is selected. The interpretation of EPD is simple and does not require asymptotic arguments. Furthermore, its calculation, for any particular model, is simple and easily added on to standard MCMC calculations.

Cross-validation

Another alternative to Bayes factors in model comparison is the use of **cross-validation** as an extension of Bayesian measures of goodness of fit. Using the predictive likelihood as a measure of utility, Vehtari and Lampinen [15] suggest comparing models by their expected utility estimates; as with other cross-validation procedures, when data points are removed by more than "one at a time", the computational burden rises quickly and is likely to be unrealistic for large models evaluated using MCMC.

References

[1] Berger, J. & Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison, in *Model Selection*, volume 38 of *Lecture Notes Monograph Series*, P. Lahiri, ed. Institute of Mathematical Statistics, Beachwood, pp. 135–207.

[2] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester, UK.

[3] Carlin, B.P. & Louis, T.A. (2000). *Bayes and empirical Bayes methods for data analysis*, 2nd Ed. Chapman & Hall, New York.

[4] Gelfand, A.E. & Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach, *Biometrika* **85**(1), 1–11.

[5] Han, C. & Carlin, B.P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review, *Journal of the American Statistical Association* **96**(455), 1122–1132.

[6] Kass, R.E. & Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**(430), 773–795.

[7] Key, J.T., Pericchi, L.R. & Smith, A.F.M. (1999). Bayesian model choice: What and why? (with discussion), in *Bayesian Statistics 6*, J.M. Bernardo, ed. Oxford University Press, Oxford, pp. 343–370.

[8] Laud, P.W. & Ibrahim, J.G. (1995). Predictive model selection, *Journal of the Royal Statistical Society, Series B, Methodological* **57**, 247–262.

[9] Lindley, D.V. (1957). A statistical paradox, *Biometrika* **44**, 187–192.

[10] Robert, Kass.E. & Wasserman, Larry. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *Journal of the American Statistical Association* **90**, 928–934.

[11] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**(2), 461–464.

[12] Smith, A.F.M. & Spiegelhalter, D.J. (1980). Bayes factors and choice criteria for linear models, *Journal of the Royal Statistical Society, Series B, Methodological* **42**, 213–220.

[13] Spiegelhalter, D.J., Best, N.G., Carlin B.P. & van der Linde, A. (2003). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **64**(4), 583–616.

[14] Trevisani, M. & Gelfand, Alan E. (2003). Inequalities between expected marginal log-likelihoods, with implications for likelihood-based model complexity and comparison measures. *Canadian Journal of Statistics*, **31**(3), 239–250.

[15] Vehtari, A. & Lampinen, J. (2003). Expected utility estimation via cross-validation, in *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith & M. West, eds. Oxford University Press, Oxford, pp. 701–710.

KENNETH RICE

Bayesian Methods in Clinical Trials

A clinical trial is an experiment carried out to gain knowledge about the relative benefits of two or more treatments. Typically, this is part of a gradual accrual of knowledge: a trial to confirm benefits in a large population may follow much careful work on smaller scale studies, or a study may be asking essentially the same question as several other studies. Conventionally, clinical trials are analyzed formally as an individual trial, and their contribution to accruing knowledge then assessed informally. However, increasingly the technique of **meta-analysis** is used to combine the information from similar trials into a formal summary.

More generally, researchers may wish to frame the following questions: "What do we think about the relative benefits of the treatments before knowing the results from this trial?" "What information can be gained from the results of this trial?" "Considering the results of this trial in the light of previous understanding, what do we now think about the relative benefits of the treatments?"

If this seems too subjective, an alternative way of casting this framework is to ask: "What is the previous evidence on the relative benefits of treatment?" "What is the current evidence from this trial?" "What is the updated evidence, once we combine the previous with the new evidence?"

The concept of updating of beliefs or evidence is the essence of Bayesian statistics. This article explains the essential concepts through a simple example, and then discusses some of the issues raised, namely the legitimate sources of previous beliefs or evidence, including the question of subjectivity, and implications for the design of trials and Bayesian reporting of clinical trials. A particular area of application is data monitoring (*see* **Data and Safety Monitoring**).

Most of the discussion is in the context of two-group parallel trials, partly for simplicity of exposition, but mainly because of the preeminence of this design in practice. The framework is, however, completely general, and applies to more complex designs. For the combination of results of several trials, possibly with other evidence, Bayesian meta-analysis is outlined. Clinical trials are often used as part of wider decision-making processes. Bayesian statistics is sometimes set in the context of decision-making, and the implications of this are discussed. Finally, there is a note on computational issues.

An Example

Consider the following example: after a heart attack, thrombolysis is often indicated. There is a tension between whether this is done at home once the ambulance arrives, which confers the advantage of speed, or in hospital, which is a more optimal environment, but necessitates a delay in treatment. The GREAT trial was run to compare these two strategies [8]. When the trial reported, there were 13 deaths out of 163 patients in the home group, and 23 out of 148 in the hospital group. The authors estimated a reduction in mortality of 49%. Some commentators were skeptical that a halving of mortality was really possible. Pocock & Spiegelhalter [14] carried out a Bayesian analysis. They judged, ignoring the trial results, that home treatment probably conferred some benefit, say a 15%–20% reduction, but that a 40% reduction, let alone a halving of mortality, was fairly unlikely. These beliefs are termed the *prior distribution*. The evidence from the trial is described through the *likelihood function*. Combining these two gives the *posterior distribution* of beliefs. This gives an estimate of the reduction of mortality of about 25%, but still says that the extremes of no effect or of a halving of mortality are unlikely.

Differences from classical analyses include the incorporation of prior beliefs, the absence of P values, and the absence of any idea of hypothetical repetitions of such studies. The posterior estimate of effect and its surrounding uncertainty via a *credibility interval* is analogous to a classical point estimate and its associated confidence interval, but has a direct interpretation in terms of belief. As many people interpret a confidence interval as the region in which the effect probably lies, they are essentially acting as Bayesians.

Mathematical Formalization

We can express the preceding analysis formally as

$$P(\theta|\text{data}) \propto P(\text{data}|\theta) \times P(\theta).$$

$P(\theta)$ is the prior distribution expressing initial beliefs about the parameter of interest. In the example, this would be the difference in mortality rates, described using the Normal distribution. $P(\text{data}|\theta)$ is the likelihood function, expressing the statistical model of variability for the data given the parameters. In the example, a Normal distribution is assumed for the difference in binomial proportions. $P(\theta|\text{data})$ is the posterior distribution of beliefs. Its shape depends on the previous two distributions, but where the prior distribution and likelihood are assumed to be Normal, the resulting posterior distribution is Normal.

The equation is usually expressed and worked with in its proportional form: if needed, the constant of proportionality is obtained by integrating the right-hand side with respect to θ, which ensures that the posterior distribution is properly defined, integrating to one.

Clinicians will have views on how different these treatments need to be before it becomes unethical to randomize patients between them. For example, some may think that if the home thrombolysis is no worse than hospital and no more than 20% better than hospital treatment, that **randomization** is acceptable, whereas once there is reasonable evidence that the difference is outside this range, a decision can be reached as to which is preferable. This range is often termed the *region of equivalence*. One end is essentially the same as the "clinically important difference" used conventionally when deciding how large studies should be, and the other end may be the point of no difference. An alternative is to have the range symmetrical about the point of no difference: the range of equivalence in the GREAT trial might be that home is no more than 20% better or worse than hospital. (This is often used in bio-**equivalence studies**.)

Sources of Prior Distributions

The Bayesian approach just outlined gives a framework for updating beliefs or evidence. There are several possible sources of prior distributions. Spiegelhalter et al. [18] recommend that there is no need to select just one, and outline a *community of priors* that can be used for interpretation.

The *reference prior* represents minimal prior information. This is the least subjective, and analyses based on this act as a useful baseline against which to compare analyses using other priors. The *clinical prior* formalizes the opinion of well-informed specific individuals. The *skeptical prior* formalizes the belief that large treatment differences are unlikely. This can be set up, for example, as having a mean of no treatment effect, and only a small probability of the effect achieving a clinically relevant value. By contrast, the *enthusiastic prior* can be specified, for example, with a mean equivalent to a clinically relevant effect, and only a small probability of no effect, or worse.

The reference prior, skeptical prior, and enthusiastic prior are essentially mathematical constructs, calibrated using points such as that of no effect, and the clinically relevant effect. By contrast, a clinical prior is intended to represent the current state of knowledge. Where possible, it should be based on good evidence, such as a meta-analysis of relevant randomized controlled trials. Where this is not possible, evidence from nonrandomized studies may be needed. Alternatively, subjective clinical opinion may form the basis of a prior distribution. Elicitation of opinion can be carried out using techniques such as interviews, questionnaires or interactive computer packages with feedback [5, 13, 21]. These are not mutually exclusive: for example, subjective judgment about relevance or changed circumstances may be needed to modify results from an objective meta-analysis.

For example, in a Bayesian analysis of a cancer clinical trial comparing high-energy neutron therapy versus the standard of photon therapy [18], priors from two sources were used. The first was from a survey of interested clinicians, which showed beliefs favoring neutron therapy; the second was from a meta-analysis of related studies of low-energy neutron therapy, which showed a detrimental effect compared with placebo (*see* **Blinding or Masking**). The data from an interim analysis (*see* **Data and Safety Monitoring**) were against neutron therapy, and, starting from either prior, the posterior belief in a worthwhile benefit was small, with the weight of posterior evidence on a harmful effect. The data monitoring committee (*see* **Data Monitoring Committees**) had actually stopped the trial at that stage (on classical analyses), and the Bayesian analyses express explicitly the wisdom of that decision.

Region of Equivalence

The region of equivalence is the area in which *equipoise* exists: a patient or his/her doctor is indifferent to which of the two treatments is used. Whilst

there is a reasonable probability that the treatments are in equipoise, randomization is ethical.

There are close parallels with the specification of the alternative hypothesis in the design of clinical trials based on the classical statistical paradigm. For a Bayesian analysis of a classically designed trial, an obvious choice for a region of equivalence is to take the points associated with the null and alternative hypotheses. When Bayesian thinking is informing the design, the range of equivalence is often elicited from clinicians using similar techniques to those used for elicitation of prior beliefs. In the neutron therapy trial described above, clinicians had also been asked about how good neutron therapy would need to be before it should be routinely used. They said (on average) that one-year survival of 50% would need to be increased to 61.5%. The range of equivalence was then taken as being between no improvement and an improvement of this magnitude.

The region of equivalence provides a useful benchmark for the design of trials, for reporting of results and for data-monitoring. These are discussed in more detail below. The region of equivalence is often determined in a relatively informal fashion by clinicians. Wider questions, about whose equipoise is really relevant, and what considerations should inform this, point towards a decision-making perspective. These are considered at the end of the article.

Bayesian Power

Classical power calculations for clinical trials are carried out by specifying a null hypothesis that two treatments do not have different effects on the outcome of interest, and an alternative hypothesis that the difference in outcome is equal to some prespecified value. The risks of wrong decisions under these two hypotheses are then fixed at chosen levels, which then determine the necessary sample size. These calculations are essentially conditional on the choice of the alternative hypothesis. There is as yet no consensus on a Bayesian approach to **sample size determination** for clinical trials. Some advocate focusing on a reasonable probability of getting a posterior interval less than a certain width, while others take an explicit decision-making perspective, with utilities either essentially "information", or some trading-off of health benefits and cost. A wide-ranging discussion of Bayesian sample size calculation can be found in a special issue of *The Statistician* [16]. See also [18] and [20].

Data Monitoring

In many trials, results accrue fast relative to patient recruitment. In this situation, data-monitoring committees are often set up to review the data to ensure that equipoise still exists, and it is still ethical to enter patients into the trial. Statistically, the challenge is to guard against stopping a trial too early, as an overreaction to early dramatic results, whilst protecting new trial patients from inappropriate randomization. From a classical statistical perspective, this is often formalized in terms of adjusting significance levels.

One Bayesian approach [18] formalizes it differently. At the start of the trial, a skeptical prior (see above) is used to represent the view that there is not too much difference between treatments. Only when the data dominate this sufficiently would early stopping be considered. The effect of such a prior is to put a brake on early results. For a trial of esophageal cancer comparing surgery with preoperative chemotherapy and surgery, this approach was used [6]. It has been shown that there is a close tie-up between this approach and classical group sequential designs (*see* **Sequential Methods for Clinical Trials**), in that a particular design, say with five interim analyses and a Pocock boundary, corresponds to a Bayesian procedure with a particular choice of prior distribution [7].

An alternative Bayesian approach to monitoring takes a much more decision-theoretic perspective. For a trial of influenza vaccination of Navajo children, monitoring included explicit consideration of future children and their risk of influenza [4].

Complex Trial Designs

The two-group parallel trial described so far is important, but not the only trial design. For more complex designs, the same framework of prior distribution/likelihood/posterior distribution outlined above still holds, although because there are more parameters, careful specification is needed, for example in parameterization. Bayesian methods have been developed for other designs, including **crossover** trials [9] and **factorial** trials [1, 15].

Bayesian Reporting of Clinical Trials

A good report of a trial specifies the question being addressed, describes the design and conduct of the trial, gives results, makes formal statistical inference from these, discusses sensitivity to assumptions, and then interprets the trial in the context of other relevant research. Many of these do not differ from usual good practice, but some aspects can benefit from formalization using Bayesian procedures [11, 19].

The results of the trial should be described clearly, and in enough detail that another reader could carry out alternative analyses if desired. Formal statistical inference follows the prior/likelihood/posterior analysis outlined above. The posterior distribution then represents a summary of beliefs/evidence about the parameters of interest. This is most fully represented graphically, but can be further summarized by giving a *95% credibility interval*. In addition, it is often useful to give the probability that an effect is in a particular region, for example the probability that the parameter lies above the region of equipoise, or, for a bio-equivalence study, the probability that the parameter lies inside it.

The results section of the report should certainly include an analysis that starts from an uninformative prior distribution. If there are other well-specified prior distributions, then an analysis using these can also be presented in full. For example, if Bayesian data-monitoring has been used, then analyses with relevant prior distributions are appropriate. Sensitivity analyses should also be carried out. These may be for sensitivity to the specification of the prior distribution, but also to the specification of other parameters in the model. For example, Grieve [9] presents plots for a bio-equivalence study looking at sensitivity to prior beliefs on the treatment effect. Sensitivity to other choices of the region of equipoise may be needed.

The discussion section of the report often contains more speculative interpretation. This can usefully be formalized through Bayesian analysis. If other opinions can be captured, for example by a skeptical or enthusiastic prior distribution, then the appropriate posterior distributions can be presented here. If there are other similar studies, then a Bayesian meta-analysis (see below) can be used to provide a combined estimate of the effects of interest.

In all Bayesian reporting the separate elements of the prior distributions and likelihood should be clearly specified and appropriately justified, so that the posterior distribution may be clearly interpreted. The likelihood should not be controversial, since it comes from the data, but specification of prior distributions is more difficult. A good rule of thumb is that if the prior distribution is based on belief, then the posterior distribution should be interpreted as an updated statement of belief, but if the prior distribution represents a summary of hard evidence, then the posterior distribution represents an updated summary of hard evidence.

Bayesian Meta-Analyses of Clinical Trials

Bayesian statistics is essentially about the updating of evidence. So far in this article the focus has been on individual trials, but where several trials address essentially the same question, a combined analysis is desirable. Bayesian meta-analysis extends Bayesian ideas used for a single trial to multiple trials. Previous evidence is expressed through *prior distributions* about quantities of interest: in a meta-analysis of binary outcomes, this will include, for example, the log odds ratio. Current data are expressed through the *likelihood*, based on an appropriate model. The *posterior distribution* for quantities of interest can then be obtained. The Bayesian framework also allows calculation of the probability that the odds ratio is at least say 1, or at least 3, which cannot be done in the classical framework.

After a careful search for all relevant trials, it seems strange to combine objective trial data with subjective opinion. In the meta-analysis context, it may be reasonable to use noninformative priors, which give intuitively interpretable results. This is particularly true for the main comparison. However, it may be useful to bring in judgment on some of the other parameters, on which the trials are less helpful, such as the size of the random effects. It is important to carry out sensitivity analyses on assumptions made. Examples of Bayesian meta-analyses include modeling random effects in a meta-analysis of urinary tract infections [17], incorporating external evidence on heterogeneity in a trial of cirrhosis [10] and modeling heterogeneity in relation to underlying risk [22].

Decision-Making with Clinical Trials

The focus in this article has been on the estimation of effects of interest using the accrued evidence. The

purpose of accruing evidence is to make decisions. Bayesian statistics leads naturally towards explicit decision-making.

There is some debate as to whether clinical trials are, in themselves, decision-making contexts. Some (Lindley and others in discussion of [18]) argue they are, whereas Spiegelhalter et al. [19] argue that an individual clinical trial can be put to a variety of purposes, and so it is better not to construe the trial as a decision in itself.

Ashby & Smith [2] argue more generally that evidence-based medicine is about making decisions and the Bayesian approach is a natural one to adopt. When a decision is to be made, the following should be identified: the decision-maker, the possible actions, the uncertain consequences, the possible sources of evidence, and the utility assessments required. For example, a patient is diagnosed with esophageal cancer. He is advised that until recently routine treatment has been surgery, but a new suggestion is to precede the surgery by a course of several weeks of chemotherapy. The *decision-maker* is the patient, who may effectively delegate to his doctor. The *possible actions* are to undergo surgery, or to opt for the combination treatment. The *uncertain consequences* are the length of his survival, the side-effects (such as severe nausea), and the delay in completion of treatment. The possible *sources of evidence* relating to his expected survival come from routine data such as cancer registries, and relating to the additional benefit of combined treatments from clinical trials. The *utility assessment* required is the patient's trade-offs between extra survival, side-effects and time spent undergoing treatment. Within this framework, evidence from clinical trials plays a very important role.

Computation

Some of the simple analyses in this article can be done analytically, using nothing more than a hand calculator. BUGS is a general-purpose package written to facilitate the fitting of complex Bayesian models [21]. It is available from `http://www.mrc-bsu.cam.ac.uk/bugs/`, and can handle the kinds of analyses referred to in this article.

Bayesian Clinical Trials in Practice

For many years the principles of Bayesian statistics have been well understood. Implementation in practical areas such as clinical trials has been hampered, until recently, by computational complexity. However, with the growth in modern computing power, the situation is changing rapidly. Analyses of real complex studies are relatively recent, and their use as the first or primary approach even newer. A Bayesian analysis now offers an intuitive approach, combined with the power to deal with complexity when necessary. Bayesian clinical trials, and integrated summaries of them using Bayesian analyses, are finding their place in practice.

A comprehensive discussion of Bayesian clinical trials with excellent references based on systematic review, can be found in Spiegelhalter et al. [19] and several case studies of Bayesian clinical trials in Berry & Stangl [3] and Kadane [12].

References

[1] Abrams, K.R., Ashby, D., Houghton, J. & Riley, D. (1996). Tamoxifen and cyclosphosphamide – synergists or antagonists?, in *Bayesian Biostatistics*, D. Berry & D. Stangl, eds. Marcel Dekker, New York.

[2] Ashby, D. & Smith, A.F.M. (2000). Evidence-based medicine as Bayesian decision-making, *Statistics in Medicine* **19**, 3291–3305.

[3] Berry, D.A. & Stangl, D. (1996). *Bayesian Biostatistics.* Marcel Dekker, New York.

[4] Berry, D.A., Wolff, B.C. & Sack, D. (1992). Public health decision making: a sequential vaccine trial, in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 79–96.

[5] Chaloner, K. & Verdinelli, I. (1995). Bayesian experimental design: a review, *Statistical Science* **10**, 273–304.

[6] Fayers, P.M., Ashby, D. & Parmar, M.K.B. (1997). Bayesian data monitoring in clinical trials, *Statistics in Medicine* **16**, 1413–1430.

[7] Freedman, L.S. & Spiegelhalter, D.J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials, *Controlled Clinical Trials* **10**, 357–367.

[8] GREAT Group (1992). Feasibility, safety, and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial, *British Medical Journal* **305**, 548–583.

[9] Grieve, A.P. (1985). A Bayesian analysis of the two-period crossover design for clinical trials, *Biometrics* **41**, 979–990.

[10] Higgins, J.P.T. & Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis, *Statistics in Medicine* **15**, 2733–2749.

[11] Hughes, M.D. (1991). Practical reporting of Bayesian analyses of clinical trials, *Drug Information Journal* **25**, 381–393.

[12] Kadane, J.B. (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*. Wiley, New York.

[13] Parmar, M.K.B., Spiegelhalter, D.J. & Freedman, L.S. (1994). The chart trials: Bayesian design and monitoring in practice, *Statistics in Medicine* **13**, 1297–1312.

[14] Pocock, S.J. & Spiegelhalter, D.J. (1992). Grampian region early anistreplase trial, *British Medical Journal* **305**, 1015.

[15] Simon, R. & Freedman, L.S. (1997). Bayesian design and analysis of two × two factorial clinical trials, *Biometrics* **53**, 456–64.

[16] Smeeton, N.C. & Adcock, C.J., eds. (1997). Sample size determination, *Statistician* **46**, 129–291 (special issue).

[17] Smith, T.C., Spiegelhalter, D.J. & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study, *Statistics in Medicine* **14**, 2685–2699.

[18] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society* **157**, 357–416.

[19] Spiegelhalter, D.J., Myles, J.P., Jones, D.R. & Abrams, K.R. (2000). Bayesian methods in health technology assessment: a review, *Health Technology Assessment* **4**(38).

[20] Tan, S.B. & Smith, A.F.M. (1998). Exploratory thoughts on clinical trials with utilities, *Statistics in Medicine* **17**, 2771–2791.

[21] Thomas, A., Spiegelhalter, D.J. & Gilks, W.R. (1992). BUGS: a program to perform Bayesian inference using Gibbs sampling, in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 837–842.

[22] Thompson, S.G., Smith, T.C. & Sharp, S.J. (1997). Investigating underlying risk as a source of heterogeneity in meta-analysis, *Statistics in Medicine* **16**, 2741–2758.

DEBORAH ASHBY

Bayesian Methods

Bayesian formulations appear in various guises in statistics, most fundamentally for assessing and interpreting posterior distributions for unknown quantities, such as population means or predictions of future observations. Considered more broadly, posterior distributions can be used to calculate posterior expectations that can be compared for decision-making purposes. And modern frequentist **decision theorists**, who are mostly anti-Bayesian in outlook, nevertheless value Bayesian procedures because the class of Bayesian decision rules is typically complete in the sense that for every non-Bayesian procedure one can find a Bayesian procedure that is at least as good [33]. This article stresses the original use for computations whose inputs are data and a probability model, and whose outputs are posterior distributions of unknown quantities of interest, a canonical example being uncertain assessment of a population mean from data on a random sample.

The simplest case of sampling a dichotomy was given a non-Bayesian treatment by Jacob Bernoulli [3] in a famous posthumous publication (*see* **Bernoulli Family**). Bernoulli was interested in estimating from data the chance of an individual of known current age surviving to a specified later age, and with this goal in mind he derived the **binomial** sampling distribution and concluded that when samples are large enough one can with high *prior* probability be nearly certain that the sample relative frequency will be close to the population relative frequency. Fifty years later the same estimation problem was reformulated by **Bayes** [1] in another famous posthumous paper showing how to associate a posterior probability distribution with the population relative frequency given the sample relative frequency.

Bayes' method was soon taken up by Laplace [23] as a fundamental principle of inference, and applied both to sampling and measurement error models. Gauss [17] claimed that he proposed the method of **least squares** and applied it to tracking an asteroid, all this in 1795 while still in his teens. Although many years later Gauss also proposed the sampling theory justification that Neyman dubbed the Gauss–Markov theorem, Gauss's original derivation applied Bayesian reasoning to a model that used the normal law of error. By 1800, Bayesian inference was well on its way to becoming the primary technical approach for probabilistic evaluation of statistical estimates. Although not without detractors, it remained in this position for more than a century, the key technique being to use as a point estimate (*see* **Estimation**) the **mode** of a posterior density of an unknown quantity assuming a **uniform** prior density.

The underlying technical concepts of Bayes' paradigm are joint, conditional, and marginal **probabilities**, where joint probability equals **conditional probability** times **marginal probability** or, in symbols,

$$\Pr(A \text{ and } B) = \Pr(A|B) \times \Pr(B).$$

Not only did Bayes [1] explicitly define this relation, he cleverly went on to make twofold use of it, first as written, and then with the roles of A and B interchanged. Thus, if A denotes the information in an observed sample, and B the unknown properties of the population from which the sample is randomly drawn, and if the two factors on the right side of the formula are assumed given, with $\Pr(A|B)$ being what is now called the sampling model, and $\Pr(B)$ what is now called the **prior distribution** of B, then the formula yields on the left-hand side the joint uncertainty of A and B prior to observation of A. Interchanging A and B in the formula, and performing minor algebraic rewriting, one obtains:

$$\Pr(B|A) = \frac{\Pr(A \text{ and } B)}{\Pr(A)},$$

which yields on the left-hand side Bayes' proposed posterior distribution of B given the observation A. In these formulas, $\Pr(A)$ can be calculated from $\Pr(A \text{ and } B)$ by summing over the possible values of B, so that the essential inputs to the computation of $\Pr(B|A)$ are the right-hand-side terms in the first equation, namely the sampling model $\Pr(A|B)$, which in Bayes' example is Bernoulli's binomial distribution, and the prior distribution $\Pr(B)$ of the target quantity B, which Bayes chose with some trepidation to be a uniform prior distribution for the unknown value of the binomial parameter (*see* **Bayes' Theorem**).

From the mid-nineteenth century onward, Bayesian inference has had persistent critics balanced by steadfast defenders. The criticisms are basically twofold. The more fundamental of the pair is a stern objectivist principle denying that probability conceived as subjective degree of belief has any acceptable place in science. The more pragmatic position

is that the logic embodied in Bayes' concept of conditional probability makes sense, but in practice the method is often unusable, or at least compromised, because unlike the sampling probabilities $\Pr(A|B)$ the prior probabilities $\Pr(B)$ are typically hard to specify in a way grounded in scientific experience.

A pragmatic position was presented long ago in a remarkably modern way by Edgeworth [11]. The critics who "heaped ridicule upon Bayes' theorem and the inverse method" were justified only under "the pretence, here deprecated, of eliciting knowledge out of ignorance, something out of nothing". Edgeworth averred that "the so-called intellectual probability is not essentially different from the probability which is founded upon special statistics," and that the change from objective to **subjective probability**, or in critic Boole's terms "material" to "intellectual" probability, "is not from experience into dreamland, but from a particular to a more general sort of experience", and if this is somewhat imprecise it is not unlike other imprecisions routinely encountered in science. Edgeworth refers to Cournot [9, Section 95, p. 69] for the argument that the arbitrariness apparent in the widespread Bayesian use of uniform priors may have little effect a posteriori. Another passage notes Cournot's introduction of the term "subjective probability, as he calls it" and argues that Cournot's suggestion that subjective probabilities only be used "to regulate the conditions of a bet" is too narrow, and should be extended to allow that such probabilities may "afford an hypothesis which may serve as a starting point for further observation".

In the twentieth century, practical statistical inference via sampling distributions has overshadowed Bayesian methods, at least until relatively recently. **R.A. Fisher** was largely responsible for initiating the shift of statistical practice toward Bernoulli's *direct* use of sampling distributions, as contrasted with Bayes' *inverse* use through combination with a prior distribution. Fisher [13] had evidently learned while a student of the criticisms of Boole, Venn, and others, and he consciously sought to construct a theory of **estimation** not dependent on Bayesian prior distributions. Fisher held a pragmatic attitude that understood and approved Bayesian logic, but felt it was limited to situations where a "superpopulation" of parameter values was available to support a choice of prior distribution. **Jerzy Neyman** also developed a theory of estimation based on **sampling distributions** that followed Fisher's attempts by about 10 years, and Neyman likewise recognized the need to break free from the Bayesian formulation [28] before beginning to develop the strongly frequentist and behaviorist theories that went hand in glove with Neyman's strong rejection of subjective probability. Fisher had a different view of probability [15], much more in tune with the pragmatism of Edgeworth than with Neyman's hard line rejection of subjective probability, and this may well have been the underlying cause of Fisher's aggressive and lifelong attacks on Neyman and company [27] for what he saw as excessive preoccupation with mathematical theories insufficiently connected with the practice of making uncertain inferences.

By 1950, despite Fisher's continuing efforts, most mathematical statisticians had adopted Neyman's positions on the use of sampling distributions, specifically, on the importance of evaluating long-run properties of statistical procedures under hypothetical repetitions, to be used in turn for comparing and choosing among procedures. Not only was Fisher's star in decline, but the earlier Bayesian tradition was lost in the shrouds of time, and regarded as a historical relic. It is interesting that when a Bayesian revival began, especially in the US in the mid-1950s, it remained for a considerable time in the behavioral or decision-theoretic mold of the school that Neyman founded. Perhaps the most influential member of the neo-Bayesian resurgence of the 1950s and 1960s was the mathematical statistician **Jimmie Savage**. Savage's axioms [30] were directed at the construction of formal models that simultaneously specified probabilities and utilities, and even his more informal forays into applied statistics [31] emphasized decision-oriented thinking more than direct assessment of uncertainty through subjective probability. The movement also paid tribute to unreconstructed Bayesians such as the mathematician **Bruno de Finetti** [10] and the geophysical scientist **Harold Jeffreys** [22] who were more interested in inferential statistics than in decision making, as was Dennis Lindley (e.g. [25]) who, especially through students, was pivotal in developing the strong contemporary school of Bayesian applied statistics in the UK. In the US, an early supporter was the highly regarded biostatistician **Jerome Cornfield**; see Cornfield [7, 8] and Zelen [34]. It was not until the mid-1980s, however, under the impetus of rapidly developing computer technologies, that Bayesian applied statistics started to exhibit capabilities of handling

complex scientific phenomena (e.g. [5] and [20]) in ways that older sampling theories appear ill-designed to address. Another important feature of modern Bayesian statistics is its straightforward adaptability to **prediction** [18].

Contemporary Bayesian Analysis: The Problem of many Parameters, and Applications in Complex Circumstances

Inference methods based on sampling distributions made big strides in the area of small sample theory in the first half of this century. Many exact **hypothesis testing** and **confidence** procedures became available for practical application to small data sets whose analysis was comfortably within the capability of the limited computers of the time. It turned out, however, that Bayesian analogs of commonplace non-Bayesian small sample procedures were easily derived and were dependent on the same analytically tractable mathematical forms, as illustrated by the conjugate Bayes theory of Raiffa & Schlaifer [29]. Furthermore, there was typically little practical difference in the interpretations associated with competing Bayesian and sampling theory methods, albeit with some interesting exceptions such as the Lindley [24] paradox. The practical congruence is especially evident for sampling models with limited parameter sets and enough data to guarantee quite accurate estimates, since, as can be demonstrated by large-sample theory, estimation errors, whether evaluated through sampling distributions or through posterior distributions, then have variances and covariances approximately given by the inverse of the Fisher **information matrix**.

The approximate similarity of Bayesian and non-Bayesian inferences is a fortuitous mathematical result that hides very different logical arguments, as was already obvious in the original papers of Bernoulli and Bayes. The logical differences are highlighted by the differing treatments of **nuisance parameters**, defined as parameters in a stochastic model whose unknown values confound the uncertainty in a primary estimate. When using sampling theory methods, it is sometimes possible to finesse nuisance parameters through a mathematical trick, as illustrated by the famous device of **Student's *t* distribution** that permits exact small sample theory concerning the mean of a normal population despite the unknown population variance nuisance parameter. Fisher and others were quick to discern mathematically straightforward extensions of **studentization** to **general linear models** with many parameters and to **multivariate normal** models. But in truth, such tricks are mainly limited to relatively few and often unrealistic sampling models, and otherwise the elimination of nuisance parameters can only be achieved approximately through the cruder device of substituting point estimates in place of unknown nuisance parameter values. By contrast, the Bayesian argument has a unified approach to elimination of nuisance parameters, namely, the reduction of a joint posterior distribution to a marginal distribution, technically describable as integrating the full joint posterior density over the nuisance parameters.

Associated with these different methods of dealing with nuisance parameters are different nonuniquenesses of inferences derived Bayesianly and non-Bayesianly from the same parametric sampling model. With samples of moderate size, deviations from the congruent asymptotic theories of **efficient** estimation depend in ways that are complex and varied, involving specific choices of estimators in the sampling theory case, and on choices of prior distributions in the Bayesian case. Thus, estimates depend on more than the data and sampling model under either theory. The Bayesian argument transparently specifies that a joint prior distribution for all the unknowns is both logically and practically necessary when sample sizes are small, a principle that has no analog in Neyman-Pearson theory and is explicitly rejected by frequentists.

Just as the task of estimating a single mean was an important test case for the early development of basic inference methodologies, the more complex task of simultaneously estimating several or many means becomes an important test case for contemporary methods of statistical inference. Biostatistical examples are abundant. For example, experimental treatment effects may vary from center to center in a **multicenter** randomized clinical trial, or vary from study to study as in a **meta-analysis**. The concept of variation between and within groups goes back to R.A. Fisher's contributions to **experimental design** and the **analysis of variance** (ANOVA) [13, 14] examples being variation among and within varieties of a crop, or among blocks and among plots within a block, arising in **randomized block** experimentation. A currently popular term for structures

with variation within and variation among levels of statistical units is hierarchical models (*see* **Multilevel Models**).

A basic question posed by hierarchical structure is whether the mean of a group should be estimated using only sample data from within the group, or whether one can borrow strength from data on other groups. One seminal idea is that of **shrinking** a mean estimated from data within one group part way to a grand mean of the estimated means of several groups, typically by using a weighting of the individual mean and the grand mean with weights summing to unity. Under frequentist theory, it is typical to evaluate such shrinking or smoothing procedures under a hypothesis of random sampling within groups, while group mean parameters are assumed fixed and nonrandom. Under Bayesian theory the distinction between fixed and random becomes nonoperational since parameters are assigned a prior distribution, which might be that the group mean parameters are viewed as independently drawn at random from a superpopulation, as in the models called **random effects** models, or originally the Model II of Eisenhart [12], in the ANOVA literature. The Bayes vs. non-Bayes distinction is further blurred by the use of the term **empirical Bayes** for shrinking methods advocated by the frequentist school, the rationale being that the sources of the methods lie in corresponding Bayesian models, so that it is Bayesian interpretation that is rejected in favor of frequentist evaluation. The natural Bayesian position is of course that a genuine prior distribution should be specified and combined with the likelihood from within groups sampling, so that the Bayesian empirical Bayes methods, as they are confusingly called, are interpreted Bayesianly as well as motivated Bayesianly. As is often the case, totally Bayesian interpretations provide simpler and more direct explanations of inferential issues.

Simultaneous estimation of many means is a convenient vehicle for illustrating several features of the Bayesian paradigm. One such is the remarkable flexibility of the method. Modern computers can be easily programmed to simulate joint posterior distributions, the marginal posterior distribution of any function of the set of means, not only their mean, or their standard deviation, but any complex quantity, such as, perhaps, the subset of means exceeding some threshold. There are competing sampling theory methods, such as **simultaneous confidence region** procedures, but they are more limited and difficult to use and interpret accurately. Another Bayesian advantage is automatic differentiation among parameters such that the data are more and less informative. For example, if some means are accurately estimated because the samples are large, while others are poorly estimated because samples are small, Bayesian procedures will automatically smooth less toward the grand mean for the larger samples and more for the smaller samples, without the need for special and often difficult derivations of estimators and their sampling distributions. Of course, there is a price to pay in terms of specification of believable prior distributions. A characteristic of a fully Bayesian treatment of hierarchical models is the necessity of parametric prior distribution modeling of lower level parameters, the parameters at the second level being called hyperparameters with priors called hyperpriors. In practice, however, these may not be as abstract as they first appear. For example, there is likely to be "more general experience", to repeat a quote above from Edgeworth, about the variation of group means, not much different from empirical knowledge of variation within groups, and sufficiently formalizable to be usable in practice.

A case can be made that Bayesian inference methods have the greatest advantage over competing sampling theoretic procedures when the phenomena are complex. By contrast, many statisticians preach keeping it simple as a fundamental principle of practice, and point to Occam's razor in support of the principle (*see* **Parsimony**). Thus, for example, it is not uncommon that a carefully designed and executed clinical trial producing large volumes of data at high cost comes down to a trivially computable single ***P*-value** from a combination of multicenter **two-by-two tables**. But Occam actually advocates "no complexity without necessity", and one necessity is to extract from complex and expensive data sets information that may exist at different levels of aggregation, at different times, at different locations, exhibiting dependence on many different covariates, and so forth. A variety of statistical technologies ought to be applied in sequence, including exploratory fitting of empirical models, construction of stochastic models, and Bayesian evaluations of associated uncertainties in estimates and predictions from models. All of these analyses contribute to constructing and refining the complex models appropriate for a thorough assessment of real experiments. The simplicity and clarity

of Bayesian principles, together with recently developed computational means of implementing these principles, allows use of Bayesian methods in complex situations that available competing inferential technologies are too cumbersome to deal with.

For biostatistics in particular, applications can already be found in a wide variety of serious investigations. For example, the proceedings volume [16] of a recent conference on case studies featuring Bayesian methods contains ten articles of which seven involve biological or biomedical studies. The titles of these articles include the key phrases, "organ blood flow measurement with colored microspheres", "elicitation, monitoring, and analysis for an AIDS clinical trial", "reconstruction of medical images", "multiple sources in the analysis of a nonequivalent control group design", "optimal design for heart defibrillators", "longitudinal care patterns for disabled elders ... missing data". Zelen [35] makes persuasive arguments for the use of Bayesian methods for case–control studies in medical statistics. Berry & Stangl [4] present further case studies. These analyses support the contention that Bayesian analysis is coming of age as a pillar of applied statistics. The sympathetic review of Breslow [6] is also informative.

Modeling for Bayesian Analysis

A statistical or probabilistic model constructed for use in Bayesian analysis is conventionally represented as having the two parts denoted above by $\Pr(A|B)$ and $\Pr(B)$, or data model and prior distribution. If one accepts the pragmatic view illustrated above by quotes from Edgeworth, so that objective and subjective probability are two sides of the same coin, the first term $\Pr(A|B)$ is indistinguishable in origin and interpretation from stochastic models that applied probabilists and frequentist statisticians develop when analyzing a specific real-world situation.

There are important inputs to stochastic models that precede any consideration of probabilistic uncertainty, and also precede data analysis carried out in support of model choice. First, it is essential to reflect knowledge and understanding of scientific context. Typical realistic models rest on a large system of variables constructed so as to formally represent placeholders for both observable and unobservable facts and quantities, to an extent judged to be appropriate and necessary for the purposes at hand. The necessities include the ability to represent factors and variables that capture causal mechanisms operating in the underlying science, and mechanisms of sample selection and experimental manipulation introduced by the statistician.

Once a framework for knowledge representation is in place, the statistician can typically recognize repeated instances of similar entities such as plots within blocks, and blocks themselves. These are the raw materials from which frequency counts can be made, but before such counts, whether observed sample units or unobserved population units, can be taken as estimating or representing probabilities it is necessary to impose an intellectual judgment of symmetry, usually called **exchangeability** in the Bayesian literature [2], meaning that uncertain expectations of any single unit are defined by an unweighted distribution across all units. Under pragmatism, neither symmetry nor frequency assumptions dominate the construction of probabilities, but instead they work in tandem.

In a specific application, a probability model is hypothesized initially through processes that are partly art and partly science, partly subjective and partly objective, partly reliant on informal recollection of what survived critical analysis in similar situations in the past, and partly on being roughly in accord with regularities and empirical models found through data exploration of old and new studies. Given an initial tentative model choice, including any prior distributions required by Bayesian methods, there begins an open-ended process of model criticism and model revision and refinement. A model selected for purposes of implementing and reporting a Bayesian analysis is inevitably a compromise among competing needs to reflect background knowledge and understanding, to render manageable the amount of detail represented and yet to maintain fidelity to the full implications of the data from the study under analysis.

The statistical research literature and texts such as [19] emphasize the use of data in the revision and ultimate choice of models. Traditional non-Bayesian significance tests of **goodness of fit** are well established tools of statistical practice that many pragmatic Bayesians can accept as helpful for model assessment. The Bayesian literature also has its own approaches to significance testing, such as model comparison through the use of Bayes factors or relative weights of evidence. Significance tests can only be part of the story, however, because what can

be detected from data by these methods depends on how adequate and informative the available data happen to be, so that important model failures can easily be missed by significance testing methods. Consequently, a different strategy for model evaluation called **sensitivity analysis** is also indicated in most applications. Sensitivity analysis means the comparison of Bayesian inferences computed from alternative model choices. If several analyses, each based on models that are judged plausible, yield inferences about the same unknown quantities that are sufficiently different to have substantively significant practical consequences, it may be wise to conclude that the study cannot support the demands made on it.

Computation

Once a two-part probability model has been established, Bayesian analysis of a given data set is automatic in principle. The remaining difficulties are largely computational. Specification of the model needs to be in a form allowing numerical representation of the joint probability density of both observables and unobserved quantities of interest, this is because repeated computation of numerical values of the joint density are needed, where the observed variables are held fixed at their values in the data, while the unknowns of interest are varied across plausible values. The basic task is to pass from the joint posterior density of a generally lengthy vector of unknowns to practically interpretable and useful representations of marginal posterior distributions of singletons or small subsets of the unknowns. These marginal representations are precisely defined in a mathematical sense, and may have subcomponents that can be numerically evaluated from analytically tractable mathematical representations. But for a core set of parameters they are typically amenable only to methods of **numerical integration**. When Bayesian methods were attempted mainly for models with small parameter sets, it was often possible to proceed by approximating intractable densities with analytically integrable forms, sometimes obtained from asymptotic theory. By the mid-1980s the limitations of these algorithms had become a roadblock preventing the widespread use of Bayesian methods. Then it was realized that Monte Carlo approaches, such as the methods of Hastings [21], that had originally been pioneered by physicists, were adaptable to marginalizing the high dimensional densities that were appearing more and more in statistics [5, 20, 32] (*see* **Markov Chain Monte Carlo**).

The method of Monte Carlo integration rests on repeated computation of possible values of a set of unknown quantities, varying these values in a way that mimics their posterior distribution. It follows that restriction of these repeated **simulations** to subset of the unknowns varies in a way that similarly mimics the marginal posterior distribution of the subset. Three technical features are characteristic of the Monte Carlo revolution of Bayesian computation. One is that the successive draws from the posterior are generally done, not to be independent from step to step, but rather each draw is dependent on the preceding draw as in a Markov chain, and hence the label MCMC for these methods. The essential feature that makes the simulations directly interpretable is that, despite the dependence, the marginal distribution associated with each draw properly mimics the desired posterior distribution. The dependence does of course complicate the task of understanding the accuracy of the Monte Carlo integration, which is itself a statistical inference problem (that paradoxically is being studied at present mainly by frequentist methods). A second key idea is that of the Gibbs sampler, being a particular way to define an MCMC algorithm that cycles through the variables sampling small subsets holding the remaining variables fixed, the reason for this being that these small individual simulations are achievable by much simpler and faster algorithms than are available for joint simulation of larger sets of variables. The third feature that goes by the name Metropolis–Hastings involves a mathematical trick that facilitates simulation from densities that are known only up to an unknown scale factor, a situation that is more common than not in Gibbs steps. Textbook expositions of MCMC methods are only recently coming on stream (e.g. [19]).

Arguments For and Against Bayesian Methods

As noted above, hard line opponents of Bayesian inference methods in statistics reject the method because it depends on a concept of subjective probability that has no place in science. Defenders, such as Edgeworth quoted above, in their turn reject this position as extreme, noting that scientific practice

has many soft aspects. Analysis of uncertainty is indeed mostly treated in science using soft informal language, while mathematical models are most prominently used for representing objective real-world phenomena. It can be argued, however, that mathematization can also be beneficial for formal representation of uncertainty, not to replace soft informal descriptions but to underpin and support them, much as formal representations of physical systems and laws bring order and validity to informal explanations of the physical world. If it is accepted that long-run frequency cannot be logically linked to specific uncertainties without an accompanying judgment about equally likely drawing of cases, then it appears that formal subjective probability is a sine qua non of the mathematization of uncertainty.

Proponents of Bayesian statistics, in attempting to overturn the establishment position of hard line objectivists, have sought to legitimize the mathematics of uncertainty through axiomatic systems that formally represent beliefs and actions (e.g. [2]). Such proponents themselves often take an extreme position that, as Edgeworth noted, is easy to ridicule. In their enthusiasm for the beauty and precision of formal systems created by axioms, they argue that the system must always be used in full. But practical use of standard Bayesian logic requires complete specification of both $\Pr(A|B)$ and $\Pr(B)$, which for complicated A and B can be difficult, especially if one is required to convince one's fellow scientists, and through them a skeptical public. Bayesian advocates may have damaged their cause by themselves taking the extreme position of insisting that probabilistic frameworks be supplied, if necessary by questioning nominated experts and forcing them to state positions that they may be unable to support from their bases of expert knowledge and experience. It is not the fact of expert opinion that is in doubt, but only the forcing of it to produce a specific form of mathematized uncertainty, or uncertainty plus **utility** for action-based systems.

Most statisticians understand the concept of conditional probability and recognize its appealing qualities as a mechanism that supports informal judgments of uncertainty given data, and they similarly understand the validity of using conditional probabilities for computation of expectations that legitimately guide actions. For such statisticians, axioms have become superfluous, and the difficulty with Bayesian methods is mainly one of understanding applied circumstances and doing the hard and not always successful work of constructing a judgmentally sound and convincing probability model for Bayesian inference and decision making. To this end, Bayesian theorists have developed useful supporting ideas such as those derived from de Finetti's theory of exchangeability [10] and concepts for checking and rebuilding models. The pragmatic view required in applied statistics needs to avoid both extreme positions and focus on scientific modeling. Bayesian thinking may then reassert in the twenty-first century the prominent role that it had in nineteenth century scientific thinking, not only as an attractive ideal, but also as a central element of practice.

References

[1] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances *Philosophical Transactions of the Royal Society of London* **53**, 370–418 and **54**, 296–325. Reprinted in *Biometrika* **45** (1958) 293–315.

[2] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.

[3] Bernoulli, J. (1713). *Ars Conjectandi*. Basel.

[4] Berry, D.A. & Stangl, D.K. (1996). *Bayesian Biostatistics*. Marcel Dekker, New York.

[5] Besag, J., Green, P., Higdon, D. & Mengerson, K. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science* **10**, 3–66.

[6] Breslow, N. (1990). Biostatistics and Bayes (with discussion), *Statistical Science* **5**, 269–298.

[7] Cornfield, J. (1966). A Bayesian test of some classical hypotheses with applications to sequential clinical trials, *Journal of the American Statistical Association* **61**, 577–594.

[8] Cornfield, J. (1969). The Bayesian outlook and its applications, *Biometrics* **25**, 617–657.

[9] Cournot, A.A. (1843). *Exposition de la Theorie de Chances et des Probabilites*. Paris.

[10] de Finetti, B. (1970). *Teoria della Probabilita*, 1 and 2. Einaudi, Turin. English translation: *Theory of Probability*, Vol. 1 (1974) and Vol. 2 (1975). Wiley, Chichester.

[11] Edgeworth, F.Y. (1884). The philosophy of chance, *Mind* **9**, 223–235.

[12] Eisenhart, C. (1947). The assumptions underlying the analysis of variance, *Biometrics* **3**, 1–38.

[13] Fisher, R.A. (1925/1973). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[14] Fisher, R.A. (1935/1960). *Design of Experiments*. Oliver & Boyd, Edinburgh.

[15] Fisher, R.A. (1958). The nature of probability, *Centennial Review* **2**, 261–274. Reprinted in *Papers of R.A. Fisher*, J.H. Bennett, ed. **5**, 384–397.

[16] Gatsonis, C.A., Hodges, J.S. Kass, R.E. & Singpurwalla, N.D., eds (1993). *Case Studies in Bayesian Statistics*. Springer-Verlag, Berlin.

[17] Gauss, C.F. (1809). *Teoria Motus Corporium Coelestum*. Hamburg. English translation by Charles Henry Davis (1857).

[18] Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall, London.

[19] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

[20] Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. & Kirby, A.J. (1993). Modelling complexity: applications of Gibbs sampling in medicine (with discussion), *Journal of the Royal Statistical Society, Series B* **55**, 39–52.

[21] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.

[22] Jeffreys, H. (1939/1961). *Theory of Probability*. Oxford University Press, Oxford.

[23] Laplace, P.S. (1774). Mémoire sur la probabilité des causes par les évènemens, *Mémoires de mathématique et de physique presentés à l'Académie royale des sciences, par divers savans, & lûs dans ses assemblées* **6**, 621–656. English translation in *Statistical Science* **1** (1986) 359–378.

[24] Lindley, D.V. (1957). A statistical paradox, *Biometrika* **44**, 187–192.

[25] Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, Cambridge.

[26] Neyman, J. (1957). Inductive behavior as a basic concept of the philosophy of science, *Review of the International Statistical Institute* **25**, 22–35.

[27] Neyman, J. (1961). The silver jubilee of my dispute with Fisher, *Journal of the Operations Research Society of Japan* **3**, 145–154.

[28] Pearson, E.S. (1962). Some thoughts on statistical inference, *Annals of Mathematical Statistics* **33**, 394–403.

[29] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard University Press, Boston.

[30] Savage, L.J. (1954). *The Foundations of Statistics*. Wiley, New York. 2nd Ed. Dover, New York, 1972.

[31] Savage, L.J., Bartlett, M.S., Barnard, G.A., Cox, D.R., Pearson, E.S. & Smith, C.A.B. (1962). *The Foundations of Statistical Inference: A Discussion*. Methuen, London.

[32] Smith, A.F.M. & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion), *Journal of the Royal Statistical Society, Series B* **55**, 3–23.

[33] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

[34] Zelen, M. (1982). The contributions of Jerome Cornfield to the theory of statistics, *Biometrics* **28**, Supplement, 11–15.

[35] Zelen, M. (1986). Case-control studies and Bayesian inference, *Statistics in Medicine* **5**, 261–269.

(*See also* **Foundations of Probability**; **Inference**)

A.P. DEMPSTER

Bayesian Model Selection in Survival Analysis

Introduction

Model comparison is a crucial part of any statistical analysis (*see* **Model, Choice of**). Owing to recent computational advances, sophisticated techniques for **Bayesian** model comparison in **survival analysis** are becoming increasingly popular. There has been a recent surge in the statistical literature on Bayesian methods for model comparison, including articles by George and McCulloch [19], Madigan and Raftery [31], Ibrahim and Laud [27], Laud and Ibrahim [30], Kass and Raftery [28], Chib [9], Raftery, Madigan, and Volinsky [36], George, McCulloch, and Tsay [20], Raftery, Madigan, and Hoeting [35], Gelfand and Ghosh [17], Clyde [11], Chen, Ibrahim, and Yiannoutsos [7], Chib and Jeliazkov [10], and Spiegelhalter et al. [42]. Articles focusing on Bayesian approaches to model comparison in the context of survival analysis include Madigan and Raftery [31], Raftery, Madigan, and Volinsky [36], Sahu, Dey, Aslanidou, and Sinha [37], Ibrahim Chen [21], Aslanidou, Dey, and Sinha [2], Chen, Harrington, and Ibrahim (2002) and [5], Sinha, Chen, and Ghosh [39], Ibrahim, Chen, and MacEachern [23], Chen and Ibrahim [6], Ibrahim, Chen, and Sinha [24], and Ibrahim, Chen, and Sinha [25]).

The scope of Bayesian model comparison is quite broad, and can be investigated via **Bayes factors**, model **diagnostics**, and **goodness-of-fit** measures (*see* **Goodness of Fit in Survival Analysis**; **Bayesian Measures of Goodness of Fit**). In many situations, one may want to compare several models that are not nested. Such comparisons are common in survival analysis, since, for example, we may want to compare a fully **parametric model** versus a **semiparametric** model, or a **cure rate model** versus a **Cox model**, and so forth. In this article, we discuss several methods for Bayesian model comparison, including Bayes factors and posterior model probabilities, the Bayesian Information Criterion (BIC), the Conditional Predictive Ordinate (CPO), and the L measure.

Posterior Model Probabilities

Perhaps the most common method of Bayesian model assessment is the computation of posterior model probabilities. The Bayesian approach to model selection is straightforward in principle. One quantifies the prior uncertainties via probabilities for each model under consideration, specifies a **prior distribution** for each of the parameters in each model, and then uses **Bayes theorem** to calculate posterior model probabilities. Let m denote a specific model in the model space $\mathcal{M}$, and let $\boldsymbol{\theta}^{(m)}$ denote the parameter vector associated with model m. Then, by Bayes theorem, the posterior probability of model m is given by

$$p(m|D) = \frac{p(D|m)p(m)}{\sum_{m \in \mathcal{M}} p(D|m)p(m)}, \qquad (1)$$

where D denotes the data,

$$p(D|m) = \int L(\boldsymbol{\theta}^{(m)}|D)\pi(\boldsymbol{\theta}^{(m)})\,\mathrm{d}\boldsymbol{\theta}^{(m)}, \qquad (2)$$

$L(\boldsymbol{\theta}^{(m)}|D)$ is the likelihood, and $p(m)$ denotes the prior probability of model m.

In Bayesian model selection, specifying meaningful prior distributions for the parameters in each model is a difficult task requiring contextual interpretations of a large number of parameters. A need to look for some useful automated specifications then arises. Reference priors can be used in many situations to address this. In some cases, however, they lead to ambiguous posterior probabilities, and require problem-specific modifications such as those in Smith and Spiegelhalter [41]. Berger and Pericchi [3] have proposed the intrinsic Bayes factor, which provides a generic solution to the ambiguity problem. However, reference priors exclude the use of any real prior information one may have. Even if one overcomes the problem of specifying priors for the parameters in the various models, there remains the question of choosing prior probabilities $p(m)$ for the models themselves. A **uniform** prior on the model space $\mathcal{M}$ may not be desirable in situations where the investigator has prior information on each subset model. To overcome difficulties in prior specification, power priors [22] can be used to specify priors for $\boldsymbol{\theta}^{(m)}$ as well as in specifying $p(m)$ for all $m \in \mathcal{M}$. We now describe this in the context of Bayesian variable subset selection.

Variable Selection in the Cox Model

Variable selection is one of the most frequently encountered problems in statistical data analysis. In

cancer or **AIDS clinical trials**, for example, one often wishes to assess the importance of certain **prognostic factors** such as treatment, age, gender, or race in predicting survival outcome. Most of the existing literature addresses variable selection using criterion-based methods such as the **Akaike Information Criterion** (AIC) [1] or Bayesian Information Criterion (BIC) [38]. As is well known, Bayesian variable selection is often difficult to carry out because of the challenge in

1. specifying prior distributions for the regression parameters for all possible models in $\mathcal{M}$;
2. specifying a prior distribution on the model space; and
3. computations.

Let p denote the number of covariates for the full model and let $\mathcal{M}$ denote the model space. We enumerate the models in $\mathcal{M}$ by $m = 1, 2, \ldots, \mathcal{K}$, where $\mathcal{K}$ is the dimension of $\mathcal{M}$ and model $\mathcal{K}$ denotes the full model. Also, let $\boldsymbol{\beta}^{(\mathcal{K})} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ denote the regression coefficients for the full model including an intercept, and let $\boldsymbol{\beta}^{(m)}$ denote a $p_m \times 1$ vector of regression coefficients for model m with an intercept, and a specific choice of $p_m - 1$ covariates. We write $\boldsymbol{\beta}^{(\mathcal{K})} = (\boldsymbol{\beta}^{(m)'}, \boldsymbol{\beta}^{(-m)'})'$, where $\boldsymbol{\beta}^{(-m)}$ is $\boldsymbol{\beta}^{(\mathcal{K})}$ with $\boldsymbol{\beta}^{(m)}$ deleted. We now consider Bayesian variable selection for the Cox model based on a discretized gamma process on the baseline hazard function with independent increments (*see* **Bayesian Survival Analysis**). Let $0 = s_0 < s_1 < \cdots < s_J$ be a finite partition of the time axis and let

$$\delta_j = h_0(s_j) - h_0(s_{j-1}) \tag{3}$$

denote the increment in the baseline hazard in the interval $(s_{j-1}, s_j]$, $j = 1, 2, \ldots, J$, and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_J)'$. For $j = 1, 2, \ldots, J$, let d_j be the number of failures, $\mathcal{D}_j$ be the set of subjects failing, c_j be the number of right **censored** observations, and $\mathcal{C}_j$ the set of subjects that are censored. Under model m, the **likelihood** can be written as

$$L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta} | D^{(m)}) = \prod_{j=1}^{J} \left\{ \exp\{-\delta_j (a_j + b_j)\} \times \prod_{k \in \mathcal{D}_j} [1 - \exp\{-\eta_k^{(m)} T_j\}] \right\}, \tag{4}$$

where $\eta_k^{(m)} = \exp(\boldsymbol{x}_k^{(m)'} \boldsymbol{\beta}^{(m)})$, $x_k^{(m)}$ is a $p_m \times 1$ vector of **covariates** for the ith individual under model m, $X^{(m)}$ denotes the $n \times p_m$ covariate matrix of rank p_m, and $D^{(m)} = (n, \boldsymbol{y}, X^{(m)}, \boldsymbol{v})$ denotes the data under model m. The rest of the terms in (4) are defined as follows:

$$a_j = \sum_{l=j+1}^{J} \sum_{k \in \mathcal{D}_l} \eta_k^{(m)} (s_{l-1} - s_{j-1}),$$

$$b_j = \sum_{l=j}^{J} \sum_{k \in \mathcal{C}_l} \eta_k^{(m)} (s_l - s_{j-1}), \tag{5}$$

and $T_j = (s_j - s_{j-1}) \sum_{l=1}^{j} \delta_l$. We have written the model here assuming that $\boldsymbol{\delta}$ does not depend on m. This is reasonable here, since our primary goal is variable selection, that is, to determine the dimension of $\boldsymbol{\beta}^{(m)}$. In this light, $\boldsymbol{\delta}$ can be viewed as a nuisance parameter in the variable selection problem. A more general version of the model can be constructed by letting $\boldsymbol{\delta}$ depend on m.

To construct a class of informative priors for $\beta^{(m)}$, one can consider the class of power priors, as discussed in [22, 24, 26]. Following Ibrahim, Chen, and Sinha [24], the power prior under model m can be written as

$$\pi(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}, a_0 | D_0^{(m)}) \propto L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta} | D_0^{(m)})^{a_0} \times \pi_0(\boldsymbol{\beta}^{(m)} | c_0) \pi_0(\boldsymbol{\delta} | \boldsymbol{\theta}_0) \pi(a_0 | \alpha_0, \lambda_0), \tag{6}$$

where $D_0^{(m)} = (n_0, \boldsymbol{y}_0, X_0^{(m)}, \boldsymbol{v}_0)$ is the historical data under model m, $\pi_0(\boldsymbol{\delta} | \boldsymbol{\theta}_0) \propto \prod_{j=1}^{J} \delta_j^{f_{0j}-1} \exp\{-\delta_j g_{0j}\}$, $\pi(a_0 | \alpha_0, \lambda_0) \propto a_0^{\alpha_0 - 1} (1 - a_0)^{\lambda_0 - 1}$, and $\boldsymbol{\theta}_0 = (f_{01}, g_{01}, \ldots, f_{0J}, g_{0J})'$ and (α_0, λ_0) are prespecified hyperparameters.

An attractive feature of the power prior for $\boldsymbol{\beta}^{(m)}$ in variable selection problems is that it is semiautomatic in the sense that one only needs a one-time input of $(D_0^{(m)}, c_0, \boldsymbol{\theta}_0, \alpha_0, \lambda_0)$ to generate the prior distributions for all $m \in \mathcal{M}$.

Choices of prior parameters for $\boldsymbol{\delta}$ can be made in several ways. One may take vague choices of prior parameters for the δ_j's such as $f_{0j} \propto g_{0j}(s_j - s_{j-1})$ and take g_{0j} small. This choice may be suitable if there is little prior information available on the baseline hazard rate. More informative choices for $\boldsymbol{\theta}_0$ can be made by incorporating the historical data $D_0^{(m)}$

into the elicitation process. A suitable choice of f_{0i} would be an increasing estimate of the baseline **hazard rate**. To construct such an estimate under model m, we can fit a **Weibull** model via **maximum likelihood** using $D_0^{(m)} = (n_0, \boldsymbol{y}_0, X_0^{(m)}, \nu_0)$ as the data. Often, the fit will result in a strictly increasing hazard. We denote such a hazard by $h^*(s|D_0^{(m)})$. Thus, we can take $f_{0j} = b_{0j} h^*(s_j|D_0^{(m)})$, where $b_{0j} = s_j - s_{j-1}$. In the event that the fitted Weibull model results in a constant or decreasing hazard, doubt is cast on the appropriateness of the gamma process as a model for the hazard, and we do not recommend this elicitation method. There are numerous other approaches to selecting this baseline hazard. Alternative classes of parametric models may be fit to $D_0^{(m)}$ or a nonparametric method such as that of Padgett and Wei [33] may be used to construct an increasing hazard.

Let the initial prior for the model space be denoted by $p_0(m)$. Given the historical data $D_0^{(m)}$, the prior probability of model m for the current study based on an update of y_0 via Bayes theorem is given by

$$p(m) \equiv p(m|D_0^{(m)}) = \frac{p(D_0^{(m)}|m)\, p_0(m)}{\sum_{m \in \mathcal{M}} p(D_0^{(m)}|m)\, p_0(m)}, \quad (7)$$

where

$$p(D_0|m) = \int L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}|D_0^{(m)}) \pi_0(\boldsymbol{\beta}^{(m)}|\, \mathrm{d}_0) \times \pi_0(\boldsymbol{\delta}|\boldsymbol{\kappa}_0)\, \mathrm{d}\beta^{(m)}\, \mathrm{d}\boldsymbol{\delta}, \quad (8)$$

$L(\boldsymbol{\delta}, \boldsymbol{\beta}^{(m)}|D_0^{(m)})$ is the likelihood function of the parameters based on $D_0^{(m)}$, $\pi_0(\boldsymbol{\beta}^{(m)}|d_0)$ is the initial prior for $\boldsymbol{\beta}^{(m)}$ given in (6) with d_0 replacing c_0, and $\pi_0(\boldsymbol{\delta}|\kappa_0)$ is the initial prior for $\boldsymbol{\delta}$ with $\boldsymbol{\kappa}_0$ replacing $\boldsymbol{\theta}_0$. We take $\pi_0(\boldsymbol{\beta}^{(m)}|d_0)$ to be a $N_{p_m}(0, d_0 W_0^{(m)})$ distribution, where $W_0^{(m)}$ is the submatrix of the diagonal matrix $W_0^{(\mathcal{K})}$ corresponding to model m. Large values of d_0 will tend to increase the prior probability for model m. Thus, the prior probability of model m for the current study is precisely the posterior probability of m given the historical data $D_0^{(m)}$, that is, $p(m) \equiv p(m|D_0^{(m)})$. This choice for $p(m)$ has several additional nice interpretations. First, $p(m)$ corresponds to the usual Bayesian update of $p_0(m)$ using $D_0^{(m)}$ as the data. Second, as $d_0 \to 0$, $p(m)$ reduces to $p_0(m)$. Therefore, as $d_0 \to 0$, the historical data $D_0^{(m)}$ have a minimal impact in determining $p(m)$. However, as $d_0 \to \infty$, $\pi_0(\boldsymbol{\beta}^{(m)}|d_0)$ plays a minimal role in determining $p(m)$, and in this case, the historical data plays a larger role in determining $p(m)$. The parameter d_0 thus serves as a tuning parameter to control the impact of $D_0^{(m)}$ on the prior model probability $p(m)$. It is important to note that we use a scalar parameter c_0 in constructing the power prior $\pi(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}, a_0|D_0^{(m)})$ in (6), while we use a *different* scalar parameter d_0 in determining $p(m)$. This development provides us with great flexibility in specifying the prior distribution for $\boldsymbol{\beta}^{(m)}$ as well as the prior model probabilities $p(m)$. Finally, we note that when there is little information about the relative plausibility of the models at the initial stage, taking $p_0(m) = 1/\mathcal{K}$, $m = 1, 2, \ldots, \mathcal{K}$, *a priori* is a reasonable "neutral" choice.

To compute $p(m)$ in (7), we follow the Monte Carlo approach of Ibrahim and Chen [21] to estimate all of the prior model probabilities using a single Gibbs sample from the full model (*see* **Markov Chain Monte Carlo**). In the context of Bayesian **variable selection** for **logistic regression**, Chen, Shao, and Ibrahim [8] use a similar idea to compute the prior model probabilities. This method involves computing the marginal distribution of the data via ratios of normalizing constants and it requires posterior samples *only* from the *full model* for computing the prior probabilities for all possible models. The method is thus very efficient for variable selection. The technical details of this method are given in Ibrahim, Chen, and Sinha [24], Ibrahim, Chen, and MacEachern [23] and Chen, Shao, and Ibrahim [8].

Criterion-based Methods

Bayesian methods for model comparison usually rely on posterior model probabilities or Bayes factors, and it is well known that to use these methods, proper prior distributions are needed when the number of parameters in the two competing models are different. In addition, posterior model probabilities are generally sensitive to the choices of prior parameters, and thus one cannot simply select vague proper priors to get around the elicitation issue. Alternatively, criterion-based methods can be attractive in the sense that they do not require proper prior distributions in general, and thus have an advantage over posterior model probabilities in this sense. However, posterior model probabilities are intrinsically well **calibrated** since probabilities are relatively easy to interpret,

whereas criterion-based methods are generally not easy to calibrate or interpret. Thus, one potential criticism of criterion-based methods for model comparison is that they generally do not have well-defined calibrations.

Recently, Ibrahim, Chen, and Sinha [25] proposed a Bayesian criterion called the *L measure* [27, 30], for model assessment and model comparison, and proposed a calibration for it. The L measure can be used as a general model assessment tool for comparing models and assessing goodness of fit for a particular model, and thus in this sense, the criterion is potentially quite versatile. A recent extension of the L measure, called the weighted L measure, can be found in [4].

Consider an experiment that yields the data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. Denote the joint sampling density of the y_i's by $f(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of indexing parameters. We allow the y_i's to be fully observed, right censored, or interval censored. In the right-censored case, y_i may be a failure time or a censoring time. In the interval-censored case, we only observe the interval $[a_{l_i}, a_{r_i}]$ in which y_i occurred. Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)'$ denote future values of a replicate experiment. That is, $\boldsymbol{z}$ is a future response vector with the same sampling density as $\boldsymbol{y}|\boldsymbol{\theta}$. The idea of using a future response vector $\boldsymbol{z}$ in developing a criterion for assessing a model or comparing several models has been well motivated in the literature by Geisser [14] and the many references therein [17, 27, 30].

Let $\eta(\cdot)$ be a known function, and let $y_i^* = \eta(y_i)$, $z_i^* = \eta(z_i)$, $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)'$, and $\boldsymbol{z}^* = (z_1^*, z_2^*, \ldots, z_n^*)'$. For example, in survival analysis, it is common to take the logarithms of the survival times, and thus in this case $\eta(y_i) = \log(y_i) = y_i^*$. Also, $\eta(y_i) = \log(y_i)$ is a common transformation in **Poisson regression**. It is also common to take $\eta(\cdot)$ to be the identity function (i.e. $\eta(y_i) = y_i$), as in normal **linear regression** or logistic regression, so that in this case, $y_i^* = y_i$ and $z_i^* = z_i$.

For a given model, we first define the statistic

$$L_1(\boldsymbol{y}^*, \boldsymbol{b}) = \mathrm{E}[(\boldsymbol{z}^* - \boldsymbol{b})'(\boldsymbol{z}^* - \boldsymbol{b})] + \delta(\boldsymbol{y}^* - \boldsymbol{b})'(\boldsymbol{y}^* - \boldsymbol{b}), \quad (9)$$

where the **expectation** is taken with respect to the posterior predictive distribution of $\boldsymbol{z}^*|\boldsymbol{y}^*$. The posterior predictive density of $\boldsymbol{z}^*|\boldsymbol{y}^*$ is given by

$$\pi(\boldsymbol{z}^*|\boldsymbol{y}^*) = \int f(\boldsymbol{z}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}^*)\,\mathrm{d}\boldsymbol{\theta}, \quad (10)$$

where $\boldsymbol{\theta}$ denotes the vector of indexing parameters, $f(\boldsymbol{z}^*|\boldsymbol{\theta})$ is the sampling distribution of the future vector $\boldsymbol{z}^*$, and $\pi(\boldsymbol{\theta}|\boldsymbol{y}^*)$ denotes the posterior distribution of $\boldsymbol{\theta}$. The statistic in (9) takes the form of a weighted discrepancy measure. The vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)'$ is an arbitrary location vector to be chosen and δ is a nonnegative scalar that weights the discrepancy based on the future values relative to the observed data. The general criterion in (9) is a special case of a class considered by Gelfand and Ghosh [17], which are motivated from a Bayesian **decision theoretic** viewpoint.

In scalar notation, (9) can be written as

$$L_1(\boldsymbol{y}^*, \boldsymbol{b}) = \sum_{i=1}^{n}\{\mathrm{Var}(z_i^*|\boldsymbol{y}^*) + (\mu_i - b_i)^2 + \delta(y_i^* - b_i)^2\}, \quad (11)$$

where $\mu_i = \mathrm{E}(z_i^*|\boldsymbol{y}^*)$. Thus, we see that (11) has the appealing decomposition as a sum involving the predictive variances plus two squared "bias" terms, $(\mu_i - b_i)^2$ and $\delta(y_i^* - b_i)^2$, where δ is a weight for the second bias component.

The $\boldsymbol{b}$ that minimizes (11) is

$$\hat{\boldsymbol{b}} = (1 - \nu)\boldsymbol{\mu} + \nu\ \boldsymbol{y}^*, \quad (12)$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$, $\nu = \delta/(\delta + 1)$, which upon substitution in (11) leads to the criterion

$$L_2(y^*) = \sum_{i=1}^{n}\mathrm{Var}(z_i^*|y_i^*) + \nu\sum_{i=1}^{n}(\mu_i - y_i^*)^2. \quad (13)$$

Clearly, $0 \le \nu < 1$, where $\nu = 0$ if $\delta = 0$, and $\nu \to 1$ as $\delta \to \infty$. The quantity ν plays a major role in (13). It can be interpreted as a weight term in the squared bias component of (13), and appears to have a lot of potential impact on the ordering of the models, as well as characterizing the properties of the L measure and calibration distribution. Ibrahim, Chen, and Sinha [25] theoretically show that certain values of ν yield highly desirable properties of the L measure and the calibration distribution compared to other values of ν. They demonstrate that the choice of ν has much potential influence on the properties

of the L measure, calibration distribution, and model choice in general. On the basis of their theoretical exploration, $\nu = 1/2$ is a desirable and justifiable choice for model selection. When $\nu = 1$, 13 reduces to the criterion of Ibrahim and Laud [27] and Laud and Ibrahim [30].

If $\boldsymbol{y}^*$ is fully observed, then (13) is straightforward to compute. However, if $\boldsymbol{y}^*$ contains right-censored or interval-censored observations, then (13) is computed by taking the expectation of these censored observations with respect to the posterior predictive distribution of the censored observations. Let $\boldsymbol{y}^* = (\boldsymbol{y}^*_{\text{obs}}, \boldsymbol{y}^*_{\text{cens}})$, where $\boldsymbol{y}^*_{\text{obs}}$ denotes the completely observed components of $\boldsymbol{y}^*$, and $\boldsymbol{y}^*_{\text{cens}}$ denotes the censored components. Here, we assume that $\boldsymbol{y}^*_{\text{cens}}$ is a random quantity and $\boldsymbol{a}_l < \boldsymbol{y}^*_{\text{cens}} < \boldsymbol{a}_r$, where $\boldsymbol{a}_l$ and $\boldsymbol{a}_r$ are known. For ease of exposition, we let $D = (n, \boldsymbol{y}^*_{\text{obs}}, \boldsymbol{a}_l, \boldsymbol{a}_r)$ denote the observed data. Then (13) is modified as

$$L(\boldsymbol{y}^*_{\text{obs}}) = \mathrm{E}_{\boldsymbol{y}^*_{\text{cens}}|D}[1\{\boldsymbol{a}_l < \boldsymbol{y}^*_{\text{cens}} < \boldsymbol{a}_r\} L_2(\boldsymbol{y}^*)], \tag{14}$$

where $1\{\boldsymbol{a}_l < \boldsymbol{y}^*_{\text{cens}} < \boldsymbol{a}_r\}$ is a generic indicator function taking the value 1 if $\boldsymbol{a}_l < \boldsymbol{y}^*_{\text{cens}} < \boldsymbol{a}_r$ and 0 otherwise, and the expectation $\mathrm{E}_{\boldsymbol{y}^*_{\text{cens}}|D}$ is taken with respect to the posterior predictive distribution $f(\boldsymbol{y}^*_{\text{cens}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)$. Note that $\boldsymbol{a}_l < \boldsymbol{y}^*_{\text{cens}} < \boldsymbol{a}_r$ means that the double inequalities hold for each component of these vectors. If, for example, all n observations are censored, then the above notation means $a_{l_i} < y^*_{\text{cens},i} < a_{r_i}, \quad i = 1, \ldots, n$, where $\boldsymbol{a}_l = (a_{l_1}, \ldots, a_{l_n})'$, $\boldsymbol{a}_r = (a_{r_1}, \ldots, a_{r_n})'$, and $\boldsymbol{y}^*_{\text{cens}} = (y^*_{\text{cens},1}, \ldots, y^*_{\text{cens},n})'$. Small values of the L measure imply a good model. Specifically, we can write (14) as

$$L(\boldsymbol{y}^*_{\text{obs}}) = \int\!\!\int_{\boldsymbol{a}_l}^{\boldsymbol{a}_r} L_2(\boldsymbol{y}^*) f(\boldsymbol{y}^*_{\text{cens}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)\, \mathrm{d}\boldsymbol{y}^*_{\text{cens}}\, \mathrm{d}\boldsymbol{\theta}, \tag{15}$$

where $f(\boldsymbol{y}^*_{\text{cens}}|\boldsymbol{\theta})$ is the sampling density of $\boldsymbol{y}^*_{\text{cens}}$ and $\pi(\boldsymbol{\theta}|D)$ is the posterior density of $\boldsymbol{\theta}$ given the observed data D. If $\boldsymbol{y}^*$ has right-censored observations, then $\boldsymbol{a}_r = \infty$, and $\boldsymbol{a}_l$ is a vector of censoring times. If $\boldsymbol{y}^*$ has interval-censored observations, then $(\boldsymbol{a}_l, \boldsymbol{a}_r)$ is a sequence of finite interval censoring times. If $\boldsymbol{y}^*$ is fully observed, that is, $\boldsymbol{y}^*_{\text{obs}} = \boldsymbol{y}^*$, then (14) reduces to (13), and therefore, $L(\boldsymbol{y}^*_{\text{obs}}) \equiv L_2(\boldsymbol{y}^*)$ in this case.

Conditional Predictive Ordinate

The CPO statistic is a very useful model assessment tool, which has been widely used in the statistical literature under various contexts. For a detailed discussion of the CPO statistic and its applications to model assessment, see [12, 14, 16, 40]. For the ith observation, the CPO statistic is defined as

$$\begin{aligned}\mathrm{CPO}_i &= f(y_i|D^{(-i)}) \\ &= \int f(y_i|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{x}_i)\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D^{(-i)})\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\boldsymbol{\lambda}, \end{aligned} \tag{16}$$

where y_i denotes the response variable and $\boldsymbol{x}_i$ is the vector of covariates for case i, $D^{(-i)}$ denotes the data with the ith case deleted, and $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D^{(-i)})$ is the posterior density of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ based on the data $D^{(-i)}$. From (16), we see that CPO_i is the marginal posterior predictive density of y_i given $D^{(-i)}$, and can be interpreted as the height of this marginal density at y_i. Thus, large values of CPO_i imply a better fit of the model.

For most models for survival data, a closed form of CPO_i is not available. However, a Monte Carlo estimator of CPO_i can be obtained using a single MCMC sample from the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)$, where D denotes the data including all cases. The implementational details for computing CPO_i can be found in [8, Chapter 10].

For comparing two competing models, we examine the CPO_i's under both models. The observation with a larger CPO value under one model will support that model over the other. Therefore, a plot of CPO_i's under both models against observation number should reveal that the better model has the majority of its CPO_i's above those of the poorer fitting model. In comparing several competing models, the CPO_i values under all models can be plotted against the observation number in a single graph.

An alternative to CPO plots is the summary statistic called the logarithm of the Pseudo-**marginal likelihood** (LPML) (see [15]), defined as

$$\mathrm{LPML} = \sum_{i=1}^{n} \log(\mathrm{CPO}_i). \tag{17}$$

In the context of survival data, the statistic LPML has been discussed by Gelfand and Mallick [18] and Sinha and Dey [40]. To compare LPML's from two different studies for a given model, we propose to use

a modification of (17), which is the average LPML given by

$$\text{ALPML} = \frac{\text{LPML}}{n}, \tag{18}$$

where n is the sample size. The statistic ALPML can be interpreted as the relative pseudo-marginal likelihood.

We see from (16) that LPML is always well defined as long the posterior predictive density is proper. Thus, LPML is well defined under improper priors, and in addition, it is very computationally stable. Therefore, LPML has a clear advantage over the Bayes factor as a model assessment tool, since it is well known that the Bayes factor is not well defined with improper priors, and is generally quite sensitive to vague proper priors. In addition, the LPML statistic also has clear advantages over other model selection criteria, such as the *L measure*. The L measure is a Bayesian criterion requiring finite second moments of the sampling distribution of y_i, whereas the LPML statistic does not require existence of any moments. Since for example, cure rate models have improper survival functions, no moments of the sampling distribution exist, and therefore the L measure is not well defined for these models (*see* **Bayesian Approaches to Cure Rate Models**).

Bayesian Model Averaging

A popular approach to model selection is Bayesian model averaging (BMA). In this approach, one base's inference on an average of all possible models in the model space $\mathcal{M}$, instead of a single "best" model. Suppose $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K\}$, and let Δ denote the quantity of interest such as a future observation, a set of regression coefficients, or the utility of a course of action. Then, the posterior distribution of Δ is given by

$$\pi(\Delta|D) = \sum_{k=1}^{K} \pi(\Delta|D, \mathcal{M}_k) p(\mathcal{M}_k|D), \tag{19}$$

where D denotes the data, $\pi(\Delta|D, \mathcal{M}_k)$ is the posterior distribution of Δ under model $\mathcal{M}_k$, and $p(\mathcal{M}_k|D)$ is the posterior model probability. Equation (19), called BMA, consists of an average of the posterior distributions under each model weighted by the corresponding posterior model probabilities. The motivation behind BMA is based on the notion that a single "best" model ignores uncertainty about the model itself, which can result in underestimated uncertainties about quantities of interest, whereas BMA in (19) incorporates model uncertainty.

The implementation of BMA is difficult for two reasons. First, $p(\mathcal{M}_k|D)$ can be difficult to compute. Second, the number of terms in (19) can be enormous. One solution to reduce the number of possible models in (19) involves applying the Occam's window algorithm of Madigan and Raftery [31]. Two basic principles underlie this ad hoc approach. First, if a model predicts the data far less well than the model that provides the best predictions, then it has effectively been discredited and should no longer be considered. Thus, models not belonging to

$$\mathcal{A}' = \left\{ \mathcal{M}_k : \frac{\max_l \{p(\mathcal{M}_l|D)\}}{p(\mathcal{M}_k|D)} \leq C \right\} \tag{20}$$

are excluded from (19), where C is chosen by the data analyst and $\max_l\{p(\mathcal{M}_l|D)\}$ denotes the model with the highest posterior probability. A common choice of C is $C = 20$. The number of models in Occam's window increases as C decreases. Second, appealing to Occam's razor, models that receive less support from the data than any other simpler models are excluded. That is, models from (19) are excluded if they belong to

$$\mathcal{B} = \left\{ \mathcal{M}_k : \exists \mathcal{M}_l \in \mathcal{M}, \mathcal{M}_l \subset \mathcal{M}_k, \frac{p(\mathcal{M}_l|D)}{p(\mathcal{M}_k|D)} > 1 \right\}. \tag{21}$$

Thus, (19) is replaced by

$$\pi(\Delta|D) = \frac{\sum_{\mathcal{M}_k \in \mathcal{A}} \pi(\Delta|D, \mathcal{M}_k) p(D|\mathcal{M}_k) p(\mathcal{M}_k)}{\sum_{\mathcal{M}_k \in \mathcal{A}} p(D|\mathcal{M}_k) p(\mathcal{M}_k)}, \tag{22}$$

where $\mathcal{A} = \mathcal{A}' \backslash \mathcal{B} \in \mathcal{M}$, $p(D|\mathcal{M}_k)$ is the marginal likelihood of the data D under model $\mathcal{M}_k$, and $p(\mathcal{M}_k)$ denotes the prior model probability.

This strategy greatly reduces the number of possible models in (19), and now all that is required is a search strategy to identify the models in $\mathcal{A}$. Two further principles underlie the search strategy. The first principle – Occam's window – concerns interpreting the ratio of posterior model probabilities

$p(\mathcal{M}_1|D)/p(\mathcal{M}_0|D)$, where $\mathcal{M}_0$ is a model with one less predictor than $\mathcal{M}_1$. If there is evidence for $\mathcal{M}_0$, then $\mathcal{M}_1$ is rejected, but to reject $\mathcal{M}_0$, stronger evidence for the larger model $\mathcal{M}_1$ is required. These principles fully define the strategy. Madigan and Raftery [31] provide a detailed description of the algorithm and mention that the number of terms in (19) is often reduced to fewer than 25.

The second approach for reducing the number of terms in (19) is to approximate (19) using an MCMC approach. Madigan and York [32] propose the MCMC model composition (MC^3) methodology, which generates a **stochastic process** that moves through the model space. A **Markov chain** $\{\mathcal{M}(l), l = 1, 2, \ldots\}$ is constructed with state space $\mathcal{M}$ and equilibrium distribution $p(\mathcal{M}_k|D)$. If this Markov chain is simulated for $l = 1, 2, \ldots, L$, then under certain regularity conditions, for any function $g(\mathcal{M}_k)$ defined on $\mathcal{M}$, the average

$$\hat{G} = \frac{1}{L}\sum_{l=1}^{L} g(\mathcal{M}(l)) \tag{23}$$

converges almost surely to $\mathrm{E}(g(\mathcal{M}_k))$ as $L \to \infty$. To compute (19) in this fashion, set $g(\mathcal{M}_k) = \pi(\Delta|D, \mathcal{M}_k)$. To construct the Markov chain, define a neighborhood $\mathrm{nbd}(\mathcal{M}_*)$ for each $\mathcal{M}_* \in \mathcal{M}$ that consists of the model $\mathcal{M}_*$ itself and the set of models with either one variable more or one variable fewer than $\mathcal{M}_*$. Define a transition matrix q by setting $q(\mathcal{M}_* \to \mathcal{M}'_*) = 0$ for all $\mathcal{M}'_* \notin \mathrm{nbd}(\mathcal{M}_*)$ and $q(\mathcal{M}_* \to \mathcal{M}'_*)$ constant for all $\mathcal{M}'_* \in \mathrm{nbd}(\mathcal{M}_*)$. If the chain is currently in state $\mathcal{M}_*$, then we proceed by drawing $\mathcal{M}'_*$ from $q(\mathcal{M}_* \to \mathcal{M}'_*)$. It is then accepted with probability

$$\min\left\{1, \frac{p(\mathcal{M}'_*|D)}{p(\mathcal{M}_*|D)}\right\}. \tag{24}$$

Otherwise, the chain stays in state $\mathcal{M}_*$.

To compute $p(D|\mathcal{M}_k)$, Raftery [34] suggests the use of the Laplace approximation, leading to

$$\log(p(D|\mathcal{M}_k)) = \log(L(\hat{\boldsymbol{\theta}}_k|D, \mathcal{M}_k)) - p_k \log(n) + O(1), \tag{25}$$

where n is the sample size, $L(\hat{\boldsymbol{\theta}}_k|D, \mathcal{M}_k)$ is the likelihood function, $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}_k$ under model $\mathcal{M}_k$, and p_k is the number of parameters in model $\mathcal{M}_k$. This is the BIC approximation derived by Schwarz [38]. In fact, (25) is much more accurate for many practical purposes than its $O(1)$ error term suggests. Kass and Wasserman [29] show that when $\mathcal{M}_j$ and $\mathcal{M}_k$ are nested and the amount of information in the prior distribution is equal to that in one observation, then the error in (25) is $O(n^{-1/2})$, under certain assumptions, rather than $O(1)$. Raftery [34] gives further empirical evidence for the accuracy of this approximation.

BMA for Variable Selection in the Cox Model

Volinsky, Madigan, Raftery, and Kronmal [45] discuss how to carry out variable selection in the Cox model using BMA. Equation (19) has three components, each posing its own computational difficulties. The predictive distribution $\pi(\Delta|D, \mathcal{M}_k)$ requires integrating out the model parameter $\boldsymbol{\theta}_k$. The posterior model probabilities $p(\mathcal{M}_k|D)$ similarly involve the calculation of an integrated likelihood. Finally, the models that fall into $\mathcal{A}$ must be located and evaluated efficiently.

In (19), the predictive distribution of Δ, given a particular model $\mathcal{M}_k$, is found by integrating out the model parameter $\boldsymbol{\theta}_k$:

$$\pi(\Delta|D, \mathcal{M}_k) = \int \pi(\Delta|\boldsymbol{\theta}_k, D, \mathcal{M}_k)\pi(\boldsymbol{\theta}_k|D, \mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}_k. \tag{26}$$

This integral does not have a closed-form solution for the Cox model. Volinsky, Madigan, Raftery, and Kronmal [45] use the MLE approximation:

$$\pi(\Delta|D, \mathcal{M}_k) \approx \pi(\Delta|\hat{\boldsymbol{\theta}}_k, D, \mathcal{M}_k). \tag{27}$$

In the context of model uncertainty, this approximation was used by Taplin [43] and found it to give an excellent approximation in his **time series regression** problem; it was subsequently used by Taplin and Raftery [44] and Draper [13].

In regression models for survival analysis, analytic evaluation of $p(D|\mathcal{M}_k)$ is not possible in general, and an analytic or computational approximation is needed. In regular statistical models (roughly speaking, those in which the MLE is consistent and asymptotically normal), $p(D|\mathcal{M}_k)$ can be approximated by (25) via the Laplace method [34].

Equation (19) requires the specification of a prior on the model space. When there is little prior information about the relative plausibility of the models considered, taking them all to be equally likely *a priori* is a reasonable "neutral" choice. When prior information about the importance of a variable is available, a prior probability on model $\mathcal{M}_k$ can be specified as

$$p(\mathcal{M}_k) = \prod_{j=1}^{p} \pi_j{}^{\delta_{kj}} \, (1 - \pi_j)^{1-\delta_{kj}}, \qquad (28)$$

where $\pi_j \in [0, 1]$ is the prior probability that $\boldsymbol{\theta}_j \neq 0$, and δ_{kj} is an indicator of whether or not variable j is included in model $\mathcal{M}_k$. Assigning $\pi_j = 0.5$ for all j corresponds to a uniform prior across the model space, while $\pi_j < 0.5$ for all j imposes a penalty for large models. Using $\pi_j = 1$ ensures that variable j is included in all models. Using this framework, elicitation of prior probabilities for models is straightforward and avoids the need to elicit priors for a large number of models. For an alternative approach, when expert information is available, see [32].

Model Selection Using BIC

Model selection criteria such as BIC are often used to select variables in regression problems. Following Volinsky and Raftery [46], we use BIC to determine the best models (where models are variable subsets) in a class of censored survival models.

When censoring is present, it is unclear whether the penalty in BIC should use n, the number of observations, or d, the number of events. When using the partial likelihood, there are only as many terms in the partial likelihood as there are events d. Kass and Wasserman [29] indicate that the term used in the penalty should be the rate at which the Hessian matrix of the log-likelihood function grows, which suggests that d is the correct quantity to use. However, if we are to use a revised version of BIC, it is important that the new criterion continue to have the asymptotic properties that Kass and Wasserman derived. In fact, the revised BIC does have these properties, with a slightly modified outcome. Suppose that

$$-\frac{1}{d} D^2 l(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) - I_u(\boldsymbol{\theta}, \boldsymbol{\psi}) = O_p(n^{-1/2}), \qquad (29)$$

where $I_u(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the expected Fisher **information** for one uncensored observation (the *uncensored unit information*) and $D^2 l(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}})$ denotes the second derivative of the log-likelihood evaluated at $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}})$. If (29) holds, then the new BIC (with d in the penalty) is an $O_p(n^{-1/2})$ approximation to twice the Bayes factor where the prior variance on $\boldsymbol{\theta}$ is now equal to the inverse of the uncensored unit information. By using d in the penalty instead of n, it can be shown that this asymptotic result holds, the only difference being in the implicit prior on the parameter. More details regarding BIC for the Cox model can be found in Volinsky and Raftery [45].

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *International Symposium on Information Theory*, B.N. Petrov & F. Csaki, eds. Akademia Kiado, Budapest, pp. 267–281.

[2] Aslanidou, H., Dey, D.K. & Sinha, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods, *Canadian Journal of Statistics* **26**, 33–48.

[3] Berger, J.O. & Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association* **91**, 109–122.

[4] Chen, M.-H., Dey, D.K. & Ibrahim, J.G. (2003). Bayesian criterion based model assessment for categorical data, *Biometrika* **91**, 45–63.

[5] Chen, M.-H., Harrington, D.P. & Ibrahim, J.G. (2002). Bayesian cure rate models for malignant melanoma: a case study of ECOG trial E1690, *Applied Statistics* **51**, 135–150.

[6] Chen, M.-H. & Ibrahim, J.G. (2001). Bayesian model comparisons for survival data with a cure fraction, *Bayesian Methods with Applications to Science, Policy and Official Statistics*, Office for Official Publications of the European Communities, Luxembourg, pp. 81–90.

[7] Chen, M.-H., Ibrahim, J.G. & Yiannoutsos, C. (1999). Prior elicitation and Bayesian computation for logistic regression models with applications to variable selection, *Journal of the Royal Statistical Society, Series B* **61**, 223–242.

[8] Chen, M.-H., Shao, Q.-M. & Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

[9] Chib, S. (1995). Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association* **90**, 1313–1321.

[10] Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association* **96**, 270–281.

[11] Clyde, M.A. (1999). Bayesian model averaging and model search strategies, in *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 157–185.

[12] Dey, D.K., Chen, M.-H. & Chang, H. (1997). Bayesian approach for nonlinear random effects models, *Biometrics* **53**, 1239–1252.

[13] Draper, D. (1995). Assessment and propagation of model uncertainty, *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

[14] Geisser, S. (1993). *Predictive Inference: an Introduction*. Chapman & Hall, London.

[15] Geisser, S. & Eddy, W. (1979). A predictive approach to model selection, *Journal of the American Statistical Association* **74**, 153–160.

[16] Gelfand, A.E., Dey, D.K. & Chang, H. (1992). Model determinating using predictive distributions with implementation via sampling-based methods (with discussion), in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 147–167.

[17] Gelfand, A.E. & Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach, *Biometrika* **85**, 1–13.

[18] Gelfand, A.E. & Mallick, B.K. (1995). Bayesian analysis of proportional hazards models built from monotone functions, *Biometrics* **51**, 843–852.

[19] George, E.I. & McCulloch, R.E. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association* **88**, 881–889.

[20] George, E.I., McCulloch, R.E. & Tsay, R.S. (1996). Two approaches to Bayesian model selections with applications, in *Bayesian Analysis in Econometrics and Statistics – Essays in Honor of Arnold Zellner*, D.A. Berry, K.A. Chaloner & J.K. Geweke, eds. Wiley, New York, pp. 339–348.

[21] Ibrahim, J.G. & Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazards models, *Sankhyā, Series B* **60**, 48–64.

[22] Ibrahim, J.G. & Chen, M.-H. (2000). Power prior distributions for regression models, *Statistical Science* **15**, 46–60.

[23] Ibrahim, J.G., Chen, M.-H. & MacEachern, S.N. (1999). Bayesian variable selection for proportional hazards models, *The Canadian Journal of Statistics* **27**, 701–717.

[24] Ibrahim, J.G., Chen, M.-H. & Sinha, D. (2001a). *Bayesian Survival Analysis*. Springer-Verlag, New York.

[25] Ibrahim, J.G., Chen, M.-H. & Sinha, D. (2001b). Criterion based methods for Bayesian model assessment, *Statistica Sinica* **11**, 419–443.

[26] Ibrahim, J.G., Chen, M.-H. & Sinha, D. (2003). On optimality properties of the power prior, *Journal of the American Statistical Association* **98**, 204–213.

[27] Ibrahim, J.G. & Laud, P.W. (1994). A predictive approach to the analysis of designed experiments, *Journal of the American Statistical Association* **89**, 309–319.

[28] Kass, R.E. & Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**, 773–795.

[29] Kass, R.E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *Journal of the American Statistical Association* **90**, 928–934.

[30] Laud, P.W. & Ibrahim, J.G. (1995). Predictive model selection, *Journal of the Royal Statistical Society, Series B* **57**, 247–262.

[31] Madigan, D. & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of the American Statistical Association* **89**, 1535–1546.

[32] Madigan, D. & York, J. (1995). Bayesian graphical models for discrete data, *International Statistical Review* **63**, 215–232.

[33] Padgett, W.J. & Wei, L.J. (1980). Maximum likelihood estimation of a distribution function with increasing failure rate based on censored observations, *Biometrika* **67**, 470–474.

[34] Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models, *Biometrika* **83**, 251–266.

[35] Raftery, A.E., Madigan, D. & Hoeting, J.A. (1997). Bayesian model averaging for linear regression models, *Journal of the American Statistical Association* **92**, 179–191.

[36] Raftery, A.E., Madigan, D. & Volinsky, C.T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance, in *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 323–350.

[37] Sahu, S.K., Dey, D.K., Aslanidou, H. & Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data, *Lifetime Data Analysis* **3**, 123–137.

[38] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**, 461–464.

[39] Sinha, D., Chen, M.-H. & Ghosh, S.K. (1999). Bayesian analysis and model selection for interval-censored survival data, *Biometrics* **55**, 585–590.

[40] Sinha, D. & Dey, D.K. (1997). Semiparametric Bayesian analysis of survival data, *Journal of the American Statistical Association* **92**, 1195–1212.

[41] Smith, A.F.M. & Spiegelhalter, D.J. (1980). Bayes factors and choice criteria for linear models, *Journal of the Royal Statistical Society, Series B* **43**, 213–220.

[42] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society, Series B* **64**, 583–639.

[43] Taplin, R.H. (1993). Robust likelihood calculation for time series, *Journal of the Royal Statistical Society, Series B* **55**, 829–836.

[44] Taplin, R.H. & Raftery, A.E. (1994). Analysis of agricultural field trials in the presence of outliers and fertility jumps, *Biometrics* **50**, 764–781.

[45] Volinsky, C.T., Madigan, D., Raftery, A.E. & Kronmal, R.A. (1997). Bayesian model averaging in proportional hazards models: assessing the risk of a stroke, *Applied Statistics* **46**, 433–448.

[46] Volinsky, C.T. & Raftery, A.E. (2000). Bayesian information criterion for censored survival models, *Biometrics* **56**, 256–262.

Further Reading

JOSEPH G. IBRAHIM & MING-HUI CHEN

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**, 711–732.

Bayesian Survival Analysis

Introduction

Nonparametric and semiparametric **Bayesian methods** in **survival analysis** have recently become quite popular due to recent advances in computing technology and the development of efficient computational **algorithms** for implementing these methods. Such methods have now become quite common and well accepted in practice, since they offer a more general modeling strategy that contains fewer assumptions. The literature on nonparametric Bayesian methods has been recently surging, and all of the references are far too enormous to list here. In this chapter, we discuss several types of Bayesian survival models, including **parametric models** as well as models involving nonparametric prior processes for the baseline **hazard** or **cumulative hazard**. Specifically, we examine piecewise constant hazard models, the gamma process, the beta process, correlated prior processes, and the Dirichlet process, with much of the focus being on the **Cox model**. In each case, we give a development of the prior process, construct the **likelihood** function, derive the posterior distributions, and discuss **Markov Chain Monte Carlo** (MCMC) sampling techniques for inference. We also give references to other types of Bayesian models, including **frailty** models, **joint models for longitudinal and survival data** flexible classes of **hierarchical models**, **accelerated failure time models**, **multivariate survival** models, **spatial** survival models, and Bayesian model **diagnostics**.

There are two fundamental approaches to **semiparametric** Bayesian survival analysis, one based on continuous time and the other based on discrete time prior processes. The discrete time approach is an approximation to the continuous time approach. The continuous time approach can be viewed as a limiting case of the discrete time approach. In practice, the model development and implementation of the continuous time approach is much more complicated than that of discrete time models. Moreover, there is the general perception that not much is gained in the continuous time approach since its discrete approximation can be made arbitrarily accurate to approximate the continuous time version. Thus, in practice, discrete time models are most often used over their continuous time versions. Following the book by Ibrahim, Chen, and Sinha [57], we primarily focus on discrete time approaches to semiparametric Bayesian survival analysis in this chapter. Some key references for continuous time Bayesian survival analysis include [29, 38, 55, 60, 67, 68, 89, 90]. References discussing computational implementation of continuous time models include [29, 67, 68, 90] and the references therein.

The rest of this chapter is organized as follows. In the section, "Fully Parametric Models", we review Bayesian parametric survival models. In the next section, we discuss semiparametric Bayesian methods for survival analysis and focus on the **proportional hazards** model of Cox [27]. We examine the piecewise constant, gamma, beta, and Dirichlet process models. In the final section, we give several references to other types of models and applications in Bayesian survival analysis.

Fully Parametric Models

Bayesian approaches to fully parametric survival analysis has been considered by many in the literature. The statistical literature in Bayesian parametric survival analysis and life-testing is too enormous to list here, but some references dealing with applications to medicine or public health include [1, 2, 19, 30, 52, 62].

The most common types of parametric models used are the **exponential**, **Weibull**, and **lognormal** models.

The exponential model is the most fundamental parametric model in survival analysis. Suppose we have independent and identically distributed (i.i.d.) survival times $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, each having an exponential distribution with parameter λ, denoted by $\mathcal{E}(\lambda)$. Denote the **censoring** indicators by $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$, where $\nu_i = 0$ if y_i is right censored and $\nu_i = 1$ if y_i is a failure time. Let $f(y_i|\lambda) = \lambda \exp(-\lambda y_i)$ denote the density for y_i, $S(y_i|\lambda) = \exp(-\lambda y_i)$ denotes the survival function and $D = (n, \boldsymbol{y}, \boldsymbol{\nu})$ denotes the observed data. We can write the likelihood function of λ as

$$L(\lambda|D) = \prod_{i=1}^{n} f(y_i|\lambda)^{\nu_i} S(y_i|\lambda)^{1-\nu_i}$$
$$= \lambda^d \exp\left(-\lambda \sum_{i=1}^{n} y_i\right), \qquad (1)$$

where $d = \sum_{i=1}^{n} \nu_i$. The conjugate **prior** for λ is the **gamma** prior. Let $\mathcal{G}(\alpha_0, \lambda_0)$ denote the gamma distribution with parameters (α_0, λ_0), with density given by

$$\pi(\lambda|\alpha_0, \lambda_0) \propto \lambda^{\alpha_0 - 1} \exp(-\lambda_0 \lambda).$$

Then, taking a $\mathcal{G}(\alpha_0, \lambda_0)$ prior for λ, the posterior distribution of λ is given by

$$\begin{aligned}
\pi(\lambda|D) &\propto L(\lambda|D)\pi(\lambda|\alpha_0, \lambda_0) \\
&\propto \left(\lambda^{\sum_{i=1}^{n} \nu_i} \exp\left\{-\lambda \sum_{i=1}^{n} y_i\right\}\right)\left(\lambda^{\alpha_0 - 1} \exp(-\lambda_0 \lambda)\right) \\
&= \lambda^{\alpha_0 + d - 1} \exp\left\{-\lambda\left(\lambda_0 + \sum_{i=1}^{n} y_i\right)\right\}. \qquad (2)
\end{aligned}$$

Thus, we recognize the kernel of the posterior distribution in (2) as a $\mathcal{G}(\alpha_0 + d, \lambda_0 + \sum_{i=1}^{n} y_i)$ distribution. The posterior mean and variance of λ are thus given by

$$\begin{aligned}
\mathrm{E}(\lambda|D) &= \frac{\alpha_0 + d}{\lambda_0 + \sum_{i=1}^{n} y_i} \qquad \text{and} \\
\mathrm{Var}(\lambda|D) &= \frac{\alpha_0 + d}{\left(\lambda_0 + \sum_{i=1}^{n} y_i\right)^2}. \qquad (3)
\end{aligned}$$

The posterior predictive distribution of a future failure time y_f is given by

$$\begin{aligned}
\pi(y_f|D) &= \int_0^\infty \pi(y_f|\lambda)\pi(\lambda|D)\,\mathrm{d}\lambda \\
&\propto \int_0^\infty \lambda^{\alpha_0 + d + 1 - 1} \exp\left\{-\lambda(y_f + \lambda_0 + \sum_{i=1}^{n} y_i)\right\} \mathrm{d}\lambda \\
&= \Gamma(\alpha_0 + d + 1)\left(\lambda_0 + \sum_{i=1}^{n} y_i + y_f\right)^{-(d + \alpha_0 + 1)} \\
&\propto \left(\lambda_0 + \sum_{i=1}^{n} y_i + y_f\right)^{-(d + \alpha_0 + 1)}. \qquad (4)
\end{aligned}$$

The normalized posterior predictive distribution is thus given by

$$\begin{aligned}
&\pi(y_f|D) \\
&= \begin{cases} \dfrac{(d + \alpha_0)\left(\lambda_0 + \sum_{i=1}^{n} y_i\right)^{(\alpha_0 + d)}}{\left(\lambda_0 + \sum_{i=1}^{n} y_i + y_f\right)^{(\alpha_0 + d + 1)}} & \text{if } y_f > 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (5)
\end{aligned}$$

In the derivation of (4) above, we need to evaluate a gamma integral, which thus led to the posterior predictive distribution in (5). The predictive distribution in (5) is known as an *inverse beta* distribution and is discussed in detail in [3].

To build a **regression** model, we introduce **covariates** through λ, and write $\lambda_i = \varphi(\boldsymbol{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\varphi(.)$ is a known function. A common form of φ is to take $\varphi(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. Another form of φ is $\varphi(\boldsymbol{x}_i'\boldsymbol{\beta}) = (\boldsymbol{x}_i'\boldsymbol{\beta})^{-1}$. Feigl and Zelen [42] also discuss this regression model. Using $\varphi(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, we are led to the likelihood function

$$\begin{aligned}
L(\boldsymbol{\beta}|D) &= \prod_{i=1}^{n} f(y_i|\lambda_i)^{\nu_i} S(y_i|\lambda_i)^{1 - \nu_i} \\
&= \exp\left\{\sum_{i=1}^{n} \nu_i \boldsymbol{x}_i'\boldsymbol{\beta}\right\} \exp\left\{-\sum_{i=1}^{n} y_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right\}, \qquad (6)
\end{aligned}$$

$D = (n, \boldsymbol{y}, X, \boldsymbol{\nu})$ and X is the $n \times p$ matrix of covariates with ith row $\boldsymbol{x}_i'$. Common prior distributions for $\boldsymbol{\beta}$ include an improper **uniform** prior, that is, $\pi(\boldsymbol{\beta}) \propto 1$, and a **normal** prior. In the regression setting, closed forms for the posterior distribution of $\boldsymbol{\beta}$ are generally not available, and therefore one needs to use **numerical integration** or Markov chain Monte Carlo (MCMC) methods. Before the advent of MCMC, numerical integration techniques were employed by Grieve [52]. However, due to the availability of statistical packages such as BUGS, the regression model in (6) can easily be fitted using MCMC techniques. Suppose we specify a p-dimensional normal prior for $\boldsymbol{\beta}$, denoted by

$N_p(\boldsymbol{\mu}_0, \Sigma_0)$, where $\boldsymbol{\mu}_0$ denotes the prior mean and Σ_0 denotes the prior covariance matrix. Then the posterior distribution of $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|D) \propto L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta}|\boldsymbol{\mu}_0, \Sigma_0), \tag{7}$$

where $\pi(\boldsymbol{\beta}|\boldsymbol{\mu}_0, \Sigma_0)$ is the multivariate normal density with mean $\boldsymbol{\mu}_0$ and covariance matrix Σ_0. The posterior in (7) does not have a closed form in general, and thus MCMC methods are needed to sample from the posterior distribution of $\boldsymbol{\beta}$. The statistical package BUGS can be readily used for this model to do the Gibbs sampling (*see* **Software, Biostatistical**).

The Weibull model is perhaps the most widely used parametric survival model. Suppose we have independent identically distributed survival times $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, each having a Weibull distribution with shape parameter α and scale parameter γ. It is often more convenient to write the model in terms of the parameterization $\lambda = \log(\gamma)$, leading to

$$f(y|\alpha, \lambda) = \alpha y^{\alpha-1} \exp(\lambda - \exp(\lambda)y^{\alpha}). \tag{8}$$

Let $S(y|\alpha, \lambda) = \exp(-\exp(\lambda)y^{\alpha})$ denote the survival function. We can write the likelihood function of (α, λ) as

$$\begin{aligned} L(\alpha, \lambda|D) &= \prod_{i=1}^{n} f(y_i|\alpha, \lambda)^{v_i} S(y_i|\alpha, \lambda)^{1-v_i} \\ &= \alpha^d \exp\left\{d\lambda + \sum_{i=1}^{n}(v_i(\alpha-1)\log(y_i) - \exp(\lambda)y_i^{\alpha})\right\}. \end{aligned} \tag{9}$$

When α is assumed known, the conjugate prior for $\exp(\lambda)$ is the gamma prior. No joint conjugate prior is available when (α, λ) are both assumed unknown. In this case, a typical joint prior specification is to take α and λ to be independent, where α has a gamma distribution and λ has a normal distribution. Letting $\mathcal{G}(\alpha_0, \kappa_0)$ denote a gamma prior for α, and $N(\mu_0, \sigma_0^2)$ denote the normal prior for λ, the joint posterior distribution of (α, λ) is given by

$$\begin{aligned} \pi(\alpha, \lambda|D) &\propto L(\alpha, \lambda|D)\pi(\alpha|\alpha_0, \kappa_0)\pi(\lambda|\mu_0, \sigma_0^2) \\ &\propto \prod_{i=1}^{n} f(y_i|\alpha, \lambda)^{v_i} S(y_i|\alpha, \lambda)^{(1-v_i)} \\ &\quad \times \pi(\alpha|\alpha_0, \kappa_0)\pi(\lambda|\mu_0, \sigma_0^2) \\ &= \alpha^{\alpha_0+d-1} \exp\left\{d\lambda + \sum_{i=1}^{n}(v_i(\alpha-1)\log(y_i) \right. \\ &\quad \left. - \exp(\lambda)y_i^{\alpha}) - \kappa_0\alpha - \frac{1}{2\sigma_0^2}(\lambda-\mu_0)^2\right\}. \end{aligned} \tag{10}$$

The joint posterior distribution of (α, λ) does not have a closed form, but it can be shown that the conditional posterior distributions $[\alpha|\lambda, D]$ and $[\lambda|\alpha, D]$ are log-concave, and thus Gibbs sampling is straightforward for this model.

To build the Weibull regression model, we introduce covariates through λ, and write $\lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. Common prior distributions for $\boldsymbol{\beta}$ include the uniform improper prior, that is, $\pi(\boldsymbol{\beta}) \propto 1$, and a normal prior. Assuming a $N_p(\boldsymbol{\mu}_0, \Sigma_0)$ prior for $\boldsymbol{\beta}$ and a gamma prior for α, we are led to the joint posterior

$$\begin{aligned} \pi(\boldsymbol{\beta}, \alpha|D) &\propto \alpha^{\alpha_0+d-1} \exp \\ &\quad \times \left\{\sum_{i=1}^{n}\left(v_i\boldsymbol{x}_i'\boldsymbol{\beta} + v_i(\alpha-1)\log(y_i) - y_i^{\alpha}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right)\right. \\ &\quad \left. - \kappa_0\alpha - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\}. \end{aligned} \tag{11}$$

Closed forms for the posterior distribution of $\boldsymbol{\beta}$ are generally not available, and therefore one needs to use numerical integration or MCMC methods. Owing to the availability of statistical packages such as BUGS, the Weibull regression model can easily be fitted using MCMC techniques. The development for the lognormal model, gamma models, **extreme value** model, an other parametric models is similar to that of the Weibull model. A multivariate extension of the Weibull model (*see* **Multivariate Weibull Distribution**) includes the Poly-Weibull model of Berger and Sun [8].

Semiparametric Models

Piecewise Constant Hazard Model

One of the most convenient and popular discrete time models for semiparametric survival analysis is the piecewise constant hazard model (*see* **Grouped Survival Times**). To construct this model, we first construct a finite partition of the time axis, $0 < s_1 < s_2 < \cdots < s_J$, with $s_J > y_i$ for all $i = 1, 2, \ldots, n$.

Thus, we have the J intervals $(0, s_1]$, $(s_1, s_2]$, ..., $(s_{J-1}, s_J]$. In the jth interval, we assume a constant baseline hazard $h_0(y) = \lambda_j$ for $y \in I_j = (s_{j-1}, s_j]$. Let $D = (n, \boldsymbol{y}, X, \nu)$ denote the observed data, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$ with $\nu_i = 1$ if the ith subject failed and 0 otherwise, and X is the $n \times p$ matrix of covariates with ith row $\boldsymbol{x}_i'$. Letting $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)'$, we can write the likelihood function of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ for the n subjects as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) = \prod_{i=1}^{n}\prod_{j=1}^{J}(\lambda_j \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^{\delta_{ij}\nu_i} \times \exp\left\{-\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right\}, \quad (12)$$

where $\delta_{ij} = 1$ if the ith subject failed or was censored in the jth interval, and 0 otherwise, $\boldsymbol{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denotes the $p \times 1$ vector of covariates for the ith subject, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the corresponding vector of regression coefficients. The indicator δ_{ij} is needed to properly define the likelihood over the J intervals. The semiparametric model in (12), sometimes referred to as a piecewise exponential model is quite general and can accommodate various shapes of the baseline hazard over the intervals. Moreover, we note that if $J = 1$, the model reduces to a parametric exponential model with failure rate parameter $\lambda \equiv \lambda_1$. The piecewise exponential model is a useful and simple model for modeling survival data. It serves as the benchmark for comparisons with other semiparametric or fully parametric models for survival data.

A common prior of the baseline hazard $\boldsymbol{\lambda}$ is the independent gamma prior $\lambda_j \sim \mathcal{G}(\alpha_{0j}, \lambda_{0j})$ for $j = 1, 2, \ldots, J$. Here α_{0j} and λ_{0j} are prior parameters that can be elicited through the prior mean and variance of λ_j. Another approach is to build a prior correlation among the λ_j's [70, 81] using a correlated prior $\boldsymbol{\psi} \sim N(\boldsymbol{\psi}_0, \Sigma_\psi)$, where $\psi_j = \log(\lambda_j)$ for $j = 1, 2, \ldots, J$ (*see* **Multivariate Normal Distribution**).

The likelihood in (12) is based on continuous survival data. The likelihood function based on grouped or discretized survival data is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) \propto \prod_{j=1}^{J} G_j^*,$$

where

$$G_j^* = \exp\left\{-\lambda_j \Delta_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\boldsymbol{x}_k'\boldsymbol{\beta})\right\} \times \prod_{l \in \mathcal{D}_j}[1 - \exp\{-\lambda_j \Delta_j \exp(\boldsymbol{x}_l'\boldsymbol{\beta})\}], \quad (13)$$

$\Delta_j = s_j - s_{j-1}$, $\mathcal{R}_j$ is the set of patients at risk, and $\mathcal{D}_j$ is the set of patients having failures in the jth interval.

Models Using a Gamma Process

The gamma process is perhaps the most commonly used nonparametric prior process for the Cox model. The seminal paper by Kalbfleisch [60] describes the gamma process prior for the baseline cumulative hazard function (see also [13]). The gamma process can be described as follows: Let $\mathcal{G}(\alpha, \lambda)$ denote the gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$. Let $\alpha(t), t \geq 0$, be an increasing left continuous function such that $\alpha(0) = 0$, and let $Z(t), t \geq 0$, be a **stochastic process** with the properties:

(i) $Z(0) = 0$;
(ii) $Z(t)$ has independent increments in disjoint intervals; and
(iii) for $t > s$, $Z(t) - Z(s) \sim \mathcal{G}(c(\alpha(t) - \alpha(s)), c)$.

Then the process $\{Z(t) : t \geq 0\}$ is called a *gamma process* and is denoted by $Z(t) \sim \mathcal{GP}(c\alpha(t), c)$. We note here that $\alpha(t)$ is the mean of the process and c is a weight or confidence parameter about the mean. The sample paths of the gamma process are almost surely increasing functions. It is a special case of a Levy process whose characteristic function is given by

$$\mathrm{E}[\exp\{iy(Z(t) - Z(s)\}] = (\phi(y))^{c(\alpha(t)-\alpha(s))}, \quad (14)$$

where ϕ is the **characteristic function** of an infinitely divisible distribution function with unit mean. The gamma process is the special case $\phi(y) = \{c/(c - iy)\}^c$.

Gamma Process on Cumulative Hazard

Under the Cox model, the joint probability of survival of n subjects given the covariate matrix X is

given by

$$P(\mathbf{Y} > \boldsymbol{y}|\boldsymbol{\beta}, X, H_0) = \exp\left\{-\sum_{j=1}^{n} \exp(\boldsymbol{x}_j'\boldsymbol{\beta}) H_0(y_j)\right\}. \quad (15)$$

The gamma process is often used as a prior for the cumulative baseline hazard function $H_0(y)$. In this case, we take

$$H_0 \sim \mathcal{GP}(c_0 H^*, c_0), \quad (16)$$

where $H^*(y)$ is an increasing function with $H^*(0) = 0$. H^* is often assumed to be a known parametric function with hyperparameter vector $\boldsymbol{\gamma}_0$. For example, if H^* corresponds to the exponential distribution, then $H^*(y) = \gamma_0 y$, where γ_0 is a specified hyperparameter. If $H^*(y)$ is taken as Weibull, then $H^*(y) = \eta_0 y^{\kappa_0}$, where $\boldsymbol{\gamma}_0 = (\eta_0, \kappa_0)'$ is a specified vector of hyperparameters. The marginal survival function is given by

$$P(\mathbf{Y} > \boldsymbol{y}|\boldsymbol{\beta}, X, \boldsymbol{\gamma}_0, c_0) = \prod_{j=1}^{n} [\phi(i V_j)]^{c_0(H^*(y_{(j)}) - H^*(y_{(j-1)}))}, \quad (17)$$

where $V_j = \sum_{l \in \mathcal{R}_j} \exp(x_l'\boldsymbol{\beta})$, $\mathcal{R}_j$ is the risk set at time $y_{(j)}$ and $y_{(1)} < y_{(2)} < \cdots, < y_{(n)}$ are distinct ordered times. For continuous data, when the ordered survival times are all distinct, the likelihood of $(\boldsymbol{\beta}, \gamma_0, c_0)$ can be obtained by differentiating (17). Note that this likelihood, used by Kalbfleisch [60], Clayton [24], and among others, is defined only when the observed survival times are distinct. In the next subsection, we present the likelihood and prior associated with grouped survival data using a gamma process prior for the baseline hazard.

Gamma Process with Grouped-data Likelihood

Again, we construct a finite partition of the time axis, $0 < s_1 < s_2 < \cdots < s_J$, with $s_J > y_i$ for all $i = 1, \ldots, n$. Thus, we have the J disjoint intervals $(0, s_1]$, $(s_1, s_2]$, …, $(s_{J-1}, s_J]$, and let $I_j = (s_{j-1}, s_j]$. The observed data D is assumed to be available as grouped within these intervals, such that $D = (X, \mathcal{R}_j, \mathcal{D}_j : j = 1, 2, \ldots, J)$, where $\mathcal{R}_j$ is the risk set and $\mathcal{D}_j$ is the failure set of the jth interval I_j. Let h_j denote the increment in the cumulative baseline hazard in the jth interval, that is,

$$h_j = H_0(s_j) - H_0(s_{j-1}), \quad j = 1, 2, \ldots, J. \quad (18)$$

The gamma process prior in (16) implies that the h_j's are independent and

$$h_j \sim \mathcal{G}(\alpha_{0j} - \alpha_{0,j-1}, c_0), \quad (19)$$

where $\alpha_{0j} = c_0 H^*(s_j)$, and H^* and c_0 are defined in the previous subsection. Thus, the hyperparameters (H^*, c_0) for h_j consist of a specified parametric cumulative hazard function $H^*(y)$ evaluated at the endpoints of the time intervals, and a positive scalar c_0 quantifying the degree of prior confidence in $H^*(y)$. Now writing $H_0 \sim \mathcal{GP}(c_0 H^*, c_0)$ implies that every disjoint increment in H_0 has the prior given by (19). Thus, the grouped-data representation can be obtained as

$$P(y_i \in I_j|\boldsymbol{h}) = \exp\left\{-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \sum_{k=1}^{j-1} h_k\right\} \times [1 - \exp\{-h_j \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}], \quad (20)$$

where $\boldsymbol{h} = (h_1, h_2, \ldots, h_J)'$. This leads to the grouped-data likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{h}|D) \propto \prod_{j=1}^{J} G_j, \quad (21)$$

where

$$G_j = \exp\left\{-h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\boldsymbol{x}_k'\boldsymbol{\beta})\right\} \times \prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp(\boldsymbol{x}_l'\boldsymbol{\beta})\}]. \quad (22)$$

Note that the grouped-data likelihood expression in (22) is very general and not limited to the case when the h_j's are realizations of a gamma process on H_0. Since the cumulative baseline hazard function H_0 enters the likelihood in (22) only through the h_j's, our parameters in the likelihood are $(\boldsymbol{\beta}, \boldsymbol{h})$ and thus we only need a joint prior distribution for $(\boldsymbol{\beta}, \boldsymbol{h})$. One important case is that when one considers the piecewise constant baseline hazard of the previous section with $h_j = \Delta_j \lambda_j$ and $\Delta_j = s_j - s_{j-1}$. In this case, we observe a great similarity

between the likelihoods (17) and (22). In the absence of covariates, (22) reduces to

$$G_j = \exp\{-h_j(r_j - d_j)\}\{1 - \exp(-h_j)\}^{d_j}, \quad (23)$$

where r_j and d_j are the numbers of subjects in the sets $\mathcal{R}_j$ and $\mathcal{D}_j$, respectively.

A typical prior for $\boldsymbol{\beta}$ is a $N_p(\boldsymbol{\mu}_0, \Sigma_0)$ distribution. Thus, the joint posterior of $(\boldsymbol{\beta}, \boldsymbol{h})$ can be written as

$$\pi(\boldsymbol{\beta}, \boldsymbol{h}|D) \propto \prod_{j=1}^{J}\left[G_j h_j^{(\alpha_{0j}-\alpha_{0,j-1})-1}\exp(-c_0 h_j)\right] \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\}.$$

Relationship to Partial Likelihood

Kalbfleisch [60] and more recently, Sinha, Ibrahim, and Chen [86] show that the **partial likelihood** defined by Cox [28] can be obtained as a limiting case of the marginal posterior of $\boldsymbol{\beta}$ in the Cox model under a gamma process prior for the cumulative baseline hazard. Towards this goal, discretize the time axis as $(0, s_1]$, $(s_1, s_2]$, ..., $(s_{J-1}, s_J]$, and suppose $H_0 \sim \mathcal{GP}(c_0 H^*, c_0)$. Let $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ denote the ordered failure or censoring times. Therefore, if $h_j = H_0(y_{(j)}) - H_0(y_{(j-1)})$, then

$$h_j \sim \mathcal{G}(c_0 h_{0j}, c_0), \quad (24)$$

where $h_{0j} = H^*(y_{(j)}) - H^*(y_{(j-1)})$. Let $A_j = \sum_{l\in\mathcal{R}_j}\exp(\boldsymbol{x}_l'\boldsymbol{\beta})$ and $\mathrm{E_{GP}}$ denote expectation with respect to the gamma process prior. Then, we have

$$P(\mathbf{Y} > \boldsymbol{y}|X, \boldsymbol{\beta}, H_0) = \exp\left\{-\sum_{j=1}^{n}\exp(\boldsymbol{x}_j'\boldsymbol{\beta})H_0(y_{(j)})\right\} = \exp\left\{-\sum_{j=1}^{n}h_j\sum_{l\in\mathcal{R}_j}\exp(\boldsymbol{x}_l'\boldsymbol{\beta})\right\} \quad (25)$$

and

$$\mathrm{E_{GP}}\left[P(\mathbf{Y} > \boldsymbol{y}|X, \boldsymbol{\beta}, H_0)|H^*\right] = \prod_{j=1}^{n}\left(\frac{c_0}{c_0 + A_j}\right)^{c_0 h_{0j}} = \prod_{j=1}^{n}\exp\left\{c_0 H^*(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)\right\}. \quad (26)$$

Now let $\boldsymbol{\theta} = (\boldsymbol{\beta}', h_0, c_0)'$, where $h_0(y) = (\mathrm{d}/\mathrm{d}y)H^*(y)$. We can write the likelihood function as

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= \prod_{j=1}^{n}\exp\left\{c_0 H^*(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)\right\} \\ &\quad \times\left\{-c_0\frac{\mathrm{d}H^*(y_{(j)})}{\mathrm{d}y_{(j)}}\left(\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)\right)\right\}^{\nu_i} \\ &= \prod_{j=1}^{n}\exp\left\{H^*(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)^{c_0}\right\} \\ &\quad \times\left\{-c_0 h_0(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)\right\}^{\nu_i}. \end{aligned} \quad (27)$$

Let $d = \sum_{i=1}^{n}\nu_i$ and $h^* = \prod_{j=1}^{n}[h_0(y_{(j)})]^{\nu_i}$. Now we have

$$\lim_{c_0\to 0}\exp\left\{H_0^*(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)^{c_0}\right\} = 1$$

for $j = 1, 2, \ldots, n$, and

$$\lim_{c_0\to 0}\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right) = \log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{A_j}\right) \approx -\frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{A_j}$$

for $j = 1, 2, \ldots, n-1$. Thus, we have

$$\lim_{c_0\to 0}\frac{L(\boldsymbol{\theta}|D)}{c_0^d\{-\log(c_0)\}^{\nu_n}h^*} \approx \prod_{j=1}^{n}\left[\frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{A_j}\right]^{\nu_i}. \quad (28)$$

We see that the right-hand side of (28) is precisely Cox's partial likelihood.

Now if we let $c_0 \to \infty$, we get the likelihood function based on $(\boldsymbol{\beta}, h_0)$. To see this, note that

$$\begin{aligned} &\lim_{c_0\to\infty}\left[\exp\left\{H^*(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)^{c_0}\right\}\right. \\ &\quad \left.\times\left\{-c_0 h_0(y_{(j)})\log\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{c_0 + A_j}\right)\right\}^{\nu_j}\right] \\ &= \exp\left\{-H^*(y_{(j)})\exp(\boldsymbol{x}_j'\boldsymbol{\beta})\right\}\left\{h_0(y_{(j)})\exp(\boldsymbol{x}_j'\boldsymbol{\beta})\right\}^{\nu_j}, \end{aligned} \quad (29)$$

and therefore,

$$\lim_{c_0\to\infty} L(\boldsymbol{\beta}, c_0, h_0|D) = \prod_{j=1}^{n} \left(\exp\left\{-H^*(y_{(j)})\exp(\boldsymbol{x}'_j\boldsymbol{\beta})\right\}\right) \times \left\{h_0(y_{(j)})\exp(\boldsymbol{x}'_j\boldsymbol{\beta})\right\}^{v_j}. \quad (30)$$

Thus, we see that (30) is the likelihood function of $(\boldsymbol{\beta}, h_0)$ based on the proportional hazards model.

Gamma Process on Baseline Hazard

An alternative specification of the semiparametric Cox model is to specify a gamma process prior on the hazard rate itself. Such a formulation is considered by Dykstra and Laud [38] in their development of the extended gamma process. Here, we consider a discrete approximation of the extended gamma process. Specifically, we construct the likelihood by using a piecewise constant baseline hazard model and use only information about which interval the failure times fall into. Let $0 = s_0 < s_1 < \cdots < s_J$ be a finite partition of the time axis and let

$$\delta_j = h_0(s_j) - h_0(s_{j-1}) \quad (31)$$

denote the increment in the baseline hazard in the interval $(s_{j-1}, s_j]$, $j = 1, 2, \ldots, J$, and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_J)'$. We follow Ibrahim, Chen, and MacEachern [56] for constructing the approximate likelihood function of $(\boldsymbol{\beta}, \boldsymbol{\delta})$. For an arbitrary individual in the population, the survival function for the Cox model at time y is given by

$$S(y|\boldsymbol{x}) = \exp\left\{-\eta\int_0^y h_0(u)\,\mathrm{d}u\right\} \approx \exp\left\{-\eta\left(\sum_{i=1}^{J}\delta_i(y - s_{i-1})^+\right)\right\}, \quad (32)$$

where $h_0(0) = 0$, $(u)^+ = u$ if $u > 0$, 0 otherwise, and $\eta = \exp(\boldsymbol{x}'\boldsymbol{\beta})$. This first approximation arises since the specification of $\boldsymbol{\delta}$ does not specify the entire hazard rate, but only the δ_j. For purposes of approximation, we take the increment in the hazard rate, δ_j, to occur immediately after s_{j-1}. Let p_j denote the probability of a failure in the interval $(s_{j-1}, s_j]$, $j = 1, 2, \ldots, J$. Using (32), we have

$$p_j = S(s_{j-1}) - S(s_j) \approx \exp\left\{-\eta\sum_{l=1}^{j-1}\delta_l(s_{j-1} - s_{l-1})\right\} \times \left[1 - \exp\left\{-\eta(s_j - s_{j-1})\sum_{l=1}^{j}\delta_l\right\}\right]. \quad (33)$$

Thus, in the jth interval $(s_{j-1}, s_j]$, the contribution to the likelihood function for a failure is p_j, and $S(s_j)$ for a right-censored observation. For $j = 1, 2, \ldots, J$, let d_j be the number of failures, $\mathcal{D}_j$ be the set of subjects failing, c_j be the number of right-censored observations and $\mathcal{C}_j$ is the set of subjects that are censored. Also, let $D = (n, \boldsymbol{y}, X, \boldsymbol{v})$ denote the data. The grouped-data likelihood function is thus given by

$$L(\boldsymbol{\beta}, \boldsymbol{\delta}|D) = \prod_{j=1}^{J}\left\{\exp\{-\delta_j(a_j + b_j)\} \times \prod_{k\in\mathcal{D}_j}[1 - \exp\{-\eta_k T_j\}]\right\}, \quad (34)$$

where $\eta_k = \exp(\boldsymbol{x}'_k\boldsymbol{\beta})$,

$$a_j = \sum_{l=j+1}^{J}\sum_{k\in\mathcal{D}_l}\eta_k(s_{l-1} - s_{j-1}),$$

$$b_j = \sum_{l=j}^{J}\sum_{k\in\mathcal{C}_l}\eta_k(s_l - s_{j-1}), \quad (35)$$

and

$$T_j = (s_j - s_{j-1})\sum_{l=1}^{j}\delta_l. \quad (36)$$

We note that this likelihood involves a second approximation. Instead of conditioning on exact event times, we condition on the set of failures and set of right-censored events in each interval, and thus we approximate continuous right-censored data by grouped data. Prior elicitation and Gibbs sampling for this model has been discussed in [57] in detail.

Beta Process Models

We first discuss time-continuous, right-censored survival data without covariates. Kalbfleisch [60] and Ferguson and Phadia [45] used the definition of the cumulative hazard $H(t)$ as

$$H(t) = -\log(S(t)), \tag{37}$$

where $S(t)$ is the survival function. The gamma process can be defined on $H(t)$ when this definition of the cumulative hazard is appropriate. A more general way of defining the hazard function, which is valid even when the survival time distribution is not continuous, is to use the definition of Hjort [55]. General formulae for the cumulative hazard function $H(t)$ are

$$H(t) = \int_{[0,t]} \frac{\mathrm{d}F(u)}{S(u)}, \tag{38}$$

where

$$F(t) = 1 - S(t) = 1 - \prod_{[0,t]} \{1 - \mathrm{d}H(t)\}. \tag{39}$$

The cumulative hazard function $H(t)$ defined here is equal to (37) when the survival distribution is absolutely continuous. Hjort [55] presents what he calls a beta process with independent increments as a prior for $H(.)$. A beta process generates a proper cdf $F(t)$, as defined in (38), and has independent increments of the form

$$\mathrm{d}H(s) \sim \mathcal{B}(c(s)\,\mathrm{d}H^*(s), c(s)(1 - \mathrm{d}H^*(s))), \tag{40}$$

where $\mathcal{B}(a, b)$ denotes the **beta distribution** with parameters (a, b). Owing to the complicated convolution property of independent beta distributions, the exact distribution of the increment $H(s)$ is only approximately beta over any finite interval, regardless of how small the length of the interval might be. See [55] for formal definitions of the beta process prior and for properties of the posterior with right-censored, time-continuous data. It is possible to deal with the beta process for the baseline cumulative hazard appropriately defined under a Cox model with time-continuous data, but survival data in practice is commonly grouped within some grid intervals, where the grid size is determined by the data and trial design. So for practical purposes, it is more convenient and often sufficient to use a discretized version of the beta process [55, 82] along with grouped survival data. The beta process prior for the cumulative baseline hazard in (40) has been discussed by many authors, including Hjort [55], Damien, Laud, and Smith [29], Laud, Smith, and Damien [68], Sinha [82], and Florens, Mouchart, and Rolin [46]. Here we will focus only on the discretized beta process prior with a grouped-data likelihood.

Within the spirit of the definition of the cumulative hazard function $H(t)$ defined in (38), a discretized version of the Cox model can be defined as

$$S(s_j|\boldsymbol{x}) = P(T > s_j|\boldsymbol{x}) = \prod_{k=1}^{j} (1 - h_k)^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}, \tag{41}$$

where h_k is the discretized baseline hazard rate in the interval $I_k = (s_{k-1}, s_k]$. The likelihood can thus be written as

$$L(\boldsymbol{\beta}, \boldsymbol{h}) = \prod_{j=1}^{J} \left((1 - h_j)^{\sum_{i \in \mathcal{R}_j - \mathcal{D}_j} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \right) \times \prod_{l \in D_j} (1 - (1 - h_j)^{\exp(x_l'\boldsymbol{\beta})}), \tag{42}$$

where $\boldsymbol{h} = (h_1, h_2, \ldots, h_J)'$. To complete the discretized beta process model, we specify independent beta priors for the h_k's. Specifically, we take $h_k \sim \mathcal{B}(c_{0k}\alpha_{0k}, c_{0k}(1 - \alpha_{0k}))$, and independent for $k = 1, 2, \ldots, J$. Though it is reasonable to assume that the h_k's are independent from each other *a* priori, the assumption of an exact beta distribution of the h_k's is only due to an approximation to the true time-continuous beta process. Thus, according to the time-continuous beta process, the distribution of the h_k's is not exactly beta, but it can be well approximated by a beta distribution only when the width of I_k is small.

Under the discretized beta process defined here, the joint prior density of $\boldsymbol{h}$ is thus given by

$$\pi(\boldsymbol{h}) \propto \prod_{j=1}^{J} h_j^{c_{0j}\alpha_{0j}-1} (1 - h_j)^{c_{0j}(1-\alpha_{0j})-1}.$$

A typical prior for $\boldsymbol{\beta}$ is a $N_p(\boldsymbol{\mu}_0, \Sigma_0)$ prior, which is independent of $\boldsymbol{h}$. Assuming an arbitrary

prior for $\boldsymbol{\beta}$, the joint posterior of $(\boldsymbol{\beta}, h)$ can be written as

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{h}|D) &\propto L(\boldsymbol{\beta}, \boldsymbol{h}|D)\pi(\boldsymbol{h})\pi(\boldsymbol{\beta}) \\ &= \prod_{j=1}^{J}\left((1-h_j)^{\sum_{i\in\mathcal{R}_j-\mathcal{D}_j}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right) \\ &\quad\times\prod_{l\in\mathcal{D}_j}\left(1-(1-h_j)^{\exp(\boldsymbol{x}_l'\boldsymbol{\beta})}\right) \\ &\quad\times\prod_{j=1}^{J}h_j^{c_{0j}\alpha_{0j}-1}(1-h_j)^{c_{0j}(1-\alpha_{0j})-1}\pi(\boldsymbol{\beta}). \quad (43)\end{aligned}$$

Given the prior structure of (43), with no covariates, the posterior distribution of the h_j's given grouped survival data is also independent beta with

$$h_j|D \sim \mathcal{B}(c_{0j}\alpha_{0j} + \mathrm{d}_j, c_{0j}(1-\alpha_{0j}) + r_j - \mathrm{d}_j), \quad (44)$$

where $D = \{(\mathrm{d}_j, r_j), j = 1, 2, \ldots, J\}$ denotes the complete grouped data. The joint posterior of the h_j's given interval-censored data is not as straightforward as (44), and is discussed in [57].

Correlated Prior Processes

The gamma process prior of Kalbfleisch [60] assumes independent cumulative hazard increments. This is unrealistic in most applied settings, and does not allow for borrowing of strength between adjacent intervals. A correlated gamma process for the cumulative hazard yields a natural smoothing of the survival curve. Although the idea of smoothing is not new [5, 7, 9, 48, 80], its potential has not been totally explored in the presence of covariates. Modeling dependence between hazard increments has been discussed by Gamerman [48] and Arjas and Gasbarra [5]. Gamerman [48] proposes a **Markov** prior process for the $\{\log(\lambda_k)\}$, by modeling

$$\log(\lambda_k) = \log(\lambda_{k-1}) + \varepsilon_k, \qquad \mathrm{E}(\varepsilon_k) = 0, \quad \text{and} \quad \mathrm{Var}(\varepsilon_k) = \sigma_k^2. \quad (45)$$

Arjas and Gasbarra [5] introduced a first-order autoregressive structure on the increment of the hazards by taking

$$\lambda_k|\lambda_{k-1} \sim \mathcal{G}\left(\alpha_k, \frac{\alpha}{\lambda_{k-1}}\right) \quad (46)$$

for $k > 1$ (*see* **ARMA and ARIMA Models**). Nieto-Barajas and Walker [74] propose dependent hazard rates with a Markovian relation, given by

$$\lambda_1 \sim \mathcal{G}(\alpha_1, \gamma_1), \; u_k|\lambda_k, v_k \sim \mathcal{P}(v_k\lambda_k), \quad v_k|\xi_k \sim \mathcal{E}\left(\frac{1}{\xi_k}\right), \quad (47)$$

$$\lambda_{k+1}|u_k, v_k \sim \mathcal{G}(\alpha_{k+1} + u_k, \gamma_{k+1} + v_k), \quad (48)$$

and

$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}),$$

for $k \geq 1$, where $\pi(\boldsymbol{\beta})$ denotes the prior for $\boldsymbol{\beta}$, which can be taken to be a normal distribution, for example, and $\mathcal{P}(v_k\lambda_k)$ denotes the Poisson distribution with mean $v_k\lambda_k$.

Dirichlet Process Models

The Dirichlet process is perhaps the most celebrated and popular prior process in nonparametric Bayesian inference. The introduction of the Dirichlet process by Ferguson [43, 44] initiated modern day Bayesian **nonparametric methods**, and today this prior process is perhaps the most important and widely used nonparametric prior. Notable articles using the Dirichlet process in various applications include [14, 31, 33, 39, 40, 49, 50, 63, 64, 66, 71, 73, 76, 89, 90].

The Dirichlet process provides the Bayesian data analyst with a nonparametric prior specification over the class of possible distribution functions $F(y)$ for a random variable Y, where $F(y) = P(Y \leq y)$. In Bayesian nonparametric inference with the Dirichlet process prior, the typical approach is to specify a prior distribution over the space all possible cumulative distribution functions, $F(t) = 1 - S(t)$. To define the Dirichlet process formally, let the sample space be denoted by Ω, and suppose $\Omega = B_1 \cup B_2 \cup \cdots \cup B_k$, where the B_j's are disjoint. The B_j's, for example, can be disjoint intervals. Then a stochastic process P indexed by elements of a particular partition $B = \{B_1, B_2, \ldots, B_k\}$ is said to be a Dirichlet process on (Ω, B) with parameter vector α, if for any partition of Ω, the random vector $(P(B_1), P(B_2), \ldots, P(B_k))$ has a Dirichlet distribution with parameter $(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k))$.

The parameter vector α is a probability measure, that is, a distribution function itself so that we can

write $\alpha = F_0(\cdot)$, where $F_0(\cdot)$ is the prior hyperparameter for $F(\cdot)$, and thus $\alpha(B_j) = F_0(b_{2j}) - F_0(b_{1j})$. The hyperparameter $F_0(.)$ is often called the *base measure* of the Dirichlet process prior.

Finally, we can define a weight parameter c_0 $(c_0 > 0)$ that gives prior weight to $F_0(\cdot)$, so that $(F(B_1), F(B_2), \ldots, F(B_k))$ has a Dirichlet distribution with parameters $(c_0F_0(B_1), c_0F_0(B_2), \ldots, c_0F_0(B_k))$. Finally, we say that F has a Dirichlet process prior with parameter c_0F_0 if $(F(B_1), F(B_2), \ldots, F(B_k))$ has a Dirichlet distribution with parameters $(c_0F_0(B_1), c_0F_0(B_2), \ldots, c_0F_0(B_k))$ for every possible partition of the sample space $\Omega = B_1 \cup B_2 \cup \cdots \cup B_k$. Some of the earliest work on Dirichlet processes in the context of survival analysis is based on the work of Ferguson and Phadia [45] and Susarla and Van Ryzin [89]. Susarla and Van Ryzin derive the Bayes estimator of the survival function under the Dirichlet process prior and also derive the posterior distribution of the cumulative distribution function with right-censored data. In this section, we summarize the fundamental results of Ferguson and Phadia [45] and Susarla and Van Ryzin [89]. Letting $S(t)$ denote the survival function, Susarla and Van Ryzin [89] derive the Bayes estimator of $S(t)$ under the squared error **loss**

$$L(\hat{S}, S) = \int_0^\infty (\hat{S}(t) - S(t))^2 \, \mathrm{d}w(t), \tag{49}$$

where w is a weight function, that is, a nonnegative decreasing function on $(0, \infty)$ and $\hat{S}(t)$ is an estimator of $S(t)$. Susarla and Van Ryzin [89] show that the Bayes estimator $\hat{S}(u)$ under squared error loss is given by

$$\hat{S}(u) = \frac{c_0(1 - F_0(u)) + N^+(u)}{c_0 + n} \times \prod_{j=k+1}^{l} \left(\frac{(c_0(1 - F_0(y_{(j)}) + N(y_{(j)})}{c_0(1 - F_0(y_{(j)}) + N(y_{(j)}) - \lambda_j} \right) \tag{50}$$

in the interval $y_{(j)} \le u \le y_{(l+1)}$, $l = k, k+1, \ldots, m$, with $y_{(k)} = 0$, $y_{(m+1)} = \infty$. The **Kaplan–Meier estimator** of $S(u)$ [61] is a limiting case of (50) and is obtained when $F_0 \to 1$, as shown by Susarla and Van Ryzin [89]. Other work on the Dirichlet process in survival data includes Kuo and Smith [66], where they used the Dirichlet process for doubly censored survival data. Generalizations of the Dirichlet process have also been used in survival analysis. Mixture of Dirichlet Process (MDP) models have been considered by [14, 39, 63, 64, 71]. The MDP model [39, 71] removes the assumption of a parametric prior at the second stage, and replaces it with a general distribution G. The distribution G then in turn has a Dirichlet process prior [43], leading to

$$\begin{aligned} &\text{Stage 1:} \quad [\boldsymbol{y}_i|\theta_i] \sim \Pi_{n_i}(h_1(\boldsymbol{\theta}_i)), \\ &\text{Stage 2:} \quad \boldsymbol{\theta}_i|G \overset{i.i.d.}{\sim} G, \\ &\text{Stage 3:} \quad [G|c_0, \boldsymbol{\psi}_0] \sim \mathcal{DP}(c_0 \cdot G_0(h_2(\boldsymbol{\psi}_0))), \end{aligned} \tag{51}$$

where G_0 is a w-dimensional parametric distribution, often called the *base measure*, and c_0 is a positive scalar. The parameters of a Dirichlet process are $G_0(\cdot)$, a probability measure, and c_0, a positive scalar. The parameter $c_0G_0(\cdot)$ contains a distribution, $G_0(\cdot)$, which approximates the true nonparametric shape of G, and the scalar c_0, which reflects our prior belief about how similar the nonparametric distribution G is to the base measure $G_0(\cdot)$.

There are two special cases in which the MDP model leads to the fully parametric case. As $c_0 \to \infty$, $G \to G_0(\cdot)$, so that the base measure is the prior distribution for $\boldsymbol{\theta}_i$. Also, if $\boldsymbol{\theta}_i \equiv \boldsymbol{\theta}$ for all i, the same is true. For a more hierarchical modeling approach, it is possible to place prior distributions on $(c_0, \boldsymbol{\psi}_0)$. The specification in (52) results in a semiparametric specification in that a fully parametric distribution is given in Stage 1 and a nonparametric distribution is given in Stages 2 and 3. References [31–33], discuss the implementation of MDP priors for $F(t) = 1 - S(t)$ in the presence of right-censored data using the Gibbs sampler. A Bayesian nonparametric approach based on mixtures of Dirichlet priors [4, 43, 44] offers a reasonable compromise between purely parametric and purely nonparametric models. The MDP prior for F can be defined as follows. If ν is some prior distribution for $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \Theta$, and $c_0 > 0$ for each $\boldsymbol{\theta}$, then if $F|\boldsymbol{\theta} \sim \mathcal{DP}(c_0F_{0\boldsymbol{\theta}})$, then F, unconditional on $\boldsymbol{\theta}$ is a mixture of Dirichlet's. The weight parameter may depend on $\boldsymbol{\theta}$, but in most applications it does not. For the case where the data are not censored, there is a closed-form expression for the posterior distribution of F. From this result, it can be easily seen that in the uncensored case, estimators based on mixtures of Dirichlet priors continuously interpolate between those based on the purely parametric and nonparametric models. For large values of c_0,

the estimators are close to the Bayes estimator based on the parametric model. On the other hand, for small values of c_0, the estimators are essentially equal to the **nonparametric maximum likelihood** estimator. One therefore expects that the same will be true for the case where the data are censored.

For the censored case, there is no closed-form expression for the posterior distribution of F given the data, and one has to use **Monte Carlo methods**. A Gibbs sampling scheme described by Doss and Huffer [32] and Doss and Narasimhan [33] will enable us to estimate the posterior distributions of interest. Other generalizations of the Cox model have been examined by Sinha, Chen, and Ghosh [83], and problems investigating interval-censored data have been investigated by Sinha [82]. A nice review paper on Bayesian survival analysis is given in [84]. Further details on Bayesian semiparametric methods can be found in [57].

Other Topics

Fully parametric Bayesian approaches to frailty models are examined in [82], where they consider a frailty model with a Weibull baseline hazard. Semiparametric approaches have also been examined. Clayton [24] and Sinha [81, 82] consider a gamma process prior on the cumulative baseline hazard in the proportional hazards frailty model. Gray [51], Sahu, Dey, Aslanidou, and Sinha [82], Sinha and Dey [84], Aslanidou, Dey, and Sinha [80], and Sinha [80] discuss frailty models with piecewise exponential baseline hazards. Qiou, Ravishanker, and Dey [75] examine a positive stable frailty distribution, and Gustafson [53] and Sargent [78] examine frailty models using Cox's partial likelihood [28]. Posterior likelihood methods for frailty models include those of Sinha [80]. Gustafson [53] discusses Bayesian hierarchical frailty models for multivariate survival data, in which the hierarchical model has elements common with the work of Clayton [24], Gray [51], Sinha [81], Stangl [87], and Stangl and Greenhouse [88]. For detailed summaries of these models, see [57].

Bayesian approaches to joint models for longitudinal and survival data have been considered by Faucett and Thomas [41], Wang and Taylor [93], Brown and Ibrahim [10, 11], Law, Taylor, and Sandler [69], Brown, Ibrahim, and DeGruttola [12], and Ibrahim, Chen, and Sinha [58], Other topics in Bayesian methods in survival analysis include proportional hazards models built from monotone functions, [50] and accelerated failure time models [23, 59, 65, 91, 92]. Survival models using Multivariate adaptive regression **splines** (MARS) have been considered by Mallick, Denison, and Smith [72]. Changepoint models have been considered by Sinha, Ibrahim, and Chen [85], and flexible classes of hierarchical survival models have been considered by Gustafson [54] and Carlin and Hodges [16]. Bayesian methods for model diagnostics in survival analysis have been considered in [50, 77, 79], Bayesian latent residual methods given in [7], and the prequential methods of Arjas and Gasbarra in [6]. Bayesian spatial survival models have been considered by Carlin and Banerjee [15]. Bayesian methods for **missing** covariate data in survival analysis include [21, 22]. Other work on Bayesian survival analysis with specific applications in epidemiology and related areas include [18, 35], applications in fertility include [34, 36], and the references therein. Bayesian survival methods in carcinogenicity studies include [37, 47]. Books discussing Bayesian survival analysis include [17, 20, 25, 26, 57].

References

[1] Achcar, J.A., Bolfarine, H. & Pericchi, L.R. (1987). Transformation of survival data to an extreme value distribution, *The Statistician* **36**, 229–234.

[2] Achcar, J.A., Brookmeyer, R. & Hunter, W.G. (1985). An application of Bayesian analysis to medical follow-up data, *Statistics in Medicine* **4**, 509–520.

[3] Aitchison, J. & Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, New York.

[4] Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics* **2**, 1152–1174.

[5] Arjas, E. & Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler, *Statistica Sinica* **4**, 505–524.

[6] Arjas, E. & Gasbarra, D. (1997). On prequential model assessment in life history analysis, *Biometrika* **84**, 505–522.

[7] Aslanidou, H., Dey, D.K. & Sinha, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods, *Canadian Journal of Statistics* **26**, 33–48.

[8] Berger, J.O. & Sun, D. (1993). Bayesian analysis for the poly-Weibull distribution, *Journal of the American Statistical Association* **88**, 1412–1418.

[9] Berzuini, C. & Clayton, D.G. (1994). Bayesian analysis of survival on multiple time scales, *Statistics in Medicine* **13**, 823–838.

[10] Brown, E.R. & Ibrahim, J.G. (2003a). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data, *Biometrics* **59**, 221–228.

[11] Brown, E.R. & Ibrahim, J.G. (2003b). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials, *Biometrics* **59**, 686–693.

[12] Brown, E.R., Ibrahim, J.G. & DeGruttola, V. (2005). A flexible joint Bayesian model of multiple longitudinal biomarkers and survival, *Biometrics* **61**.

[13] Burridge, J. (1981). Empirical Bayes analysis for survival time data, *Journal of the Royal Statistical Society, Series B* **43**, 65–75.

[14] Bush, C.A. & MacEachern, S.N. (1996). A semiparametric Bayesian model for randomized block designs, *Biometrika* **33**, 275–285.

[15] Carlin, B.P. & Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data, (with discussion), in *Bayesian Statistics 7*, J. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith & M. West, eds. Clarendon Press, Oxford, pp. 45–63.

[16] Carlin, B.P. & Hodges, J.S. (1999). Hierarchical proportional hazards regression models for highly stratified data, *Biometrics* **55**, 1162–1170.

[17] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

[18] Chen, M.-H., Dey, D.K. & Sinha, D. (2000). Bayesian analysis of multivariate mortality data with large families, *Applied Statistics* **49**, 129–144.

[19] Chen, W.C., Hill, B.M., Greenhouse, J.B. & Fayos, J.V. (1985). Bayesian analysis of survival curves for cancer patients following treatment, in *Bayesian Statistics 2* J.O. Berger, J. Bernardo & A.F.M. Smith, eds. North-Holland, Amsterdam, pp. 299–328.

[20] Chen, M.-H., Shao, Q.-M. & Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

[21] Chen, M.H., Ibrahim, J.G. & Lipsitz, S.R. (2002). Bayesian methods for missing covariates in cure rate models, *Lifetime Data Analysis* **8**, 117–146.

[22] Chen, M.-H., Ibrahim, J.G. & Shao, Q.M. (2004). On propriety of the posterior distribution and existence of the maximum likelihood estimator for regression models with covariates missing at random, *Journal of the American Statistical Association* **99**, 421–438.

[23] Christensen, R. & Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process, *Biometrika* **75**, 693–704.

[24] Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models, *Biometrics* **47**, 467–485.

[25] Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley and Sons, New York.

[26] Congdon, P. (2003). *Applied Bayesian Modelling*. John Wiley and Sons, New York.

[27] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[28] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[29] Damien, P., Laud, P.W. & Smith, A.F.M. (1996). Implementation of Bayesian non-parametric inference based on beta processes, *Scandinavian Journal of Statistics* **23**, 27–36.

[30] Dellaportas, P. & Smith, A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling, *Applied Statistics* **42**, 443–459.

[31] Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling, *The Annals of Statistics* **22**, 1763–1786.

[32] Doss, H. & Huffer, F. (1998). *Monte Carlo Methods for Bayesian Analysis of Survival Data Using Mixtures of Dirichlet Priors*, Technical Report, Department of Statistics, Ohio State University.

[33] Doss, H. & Narasimhan, B. (1998). Dynamic display of changing posterior in Bayesian survival analysis, in *Practical Nonparametric and Semiparametric Bayesian Statistics*, D. Dey, P. Müller & D. Sinha. eds. Springer-Verlag, New York, pp. 63–84.

[34] Dunson, D.B. (2001). Bayesian modeling of the level and duration of fertility in the menstrual cycle, *Biometrics* **57**, 1067–1073.

[35] Dunson, D.B., Chulada, P. & Arbes, S.J. (2003). Bayesian modeling of time-varying and waning exposure effects, *Biometrics* **59**, 83–91.

[36] Dunson, D.B. & Dinse, G.E. (2000). Distinguishing effects on tumor multiplicity and growth rate in chemoprevention experiments, *Biometrics* **56**, 1068–1075.

[37] Dunson, D.B. & Dinse, G.E. (2001). Bayesian incidence analysis of animal tumorigenicity data, *Applied Statistics* **50**, 125–141.

[38] Dykstra, R.L. & Laud, P.W. (1981). A Bayesian nonparametric approach to reliability, *The Annals of Statistics* **9**, 356–367.

[39] Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior, *Journal of the American Statistical Association* **89**, 268–277.

[40] Escobar, M.D. & West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**, 578–588.

[41] Faucett, C.J. & Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach, *Statistics in Medicine* **15**, 1663–1685.

[42] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[43] Ferguson, T.S. (1973). A Bayesian analysis of some non-parametric problems, *The Annals of Statistics* **1**, 209–230.

[44] Ferguson, T.S. (1974). Prior distributions on spaces of probability measures, *The Annals of Statistics* **2**, 615–629.

[45] Ferguson, T.S. & Phadia, E.G. (1979). Bayesian nonparametric estimation based on censored data, *The Annals of Statistics* **7**, 163–186.

[46] Florens, J.P., Mouchart, M. & Rolin, J.M. (1999). Semi-and non-parametric Bayesian analysis of duration models with Dirichlet priors: A survey, *International Statistical Review* **67**, 187–210.

[47] French, J.L. & Ibrahim, J.G. (2002). Bayesian methods for a three-state model for rodent carcinogenicity studies, *Biometrics* **58**, 906–916.

[48] Gamerman, D. (1991). Dynamic Bayesian models for survival data, *Applied Statistics* **40**, 63–79.

[49] Gelfand, A.E. & Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response, *Biometrika* **78**, 657–666.

[50] Gelfand, A.E. & Mallick, B.K. (1995). Bayesian analysis of proportional hazards models built from monotone functions, *Biometrics* **51**, 843–852.

[51] Gray, R.J. (1994). A Bayesian analysis of institutional effects in a multicenter cancer clinical trial, *Biometrics* **50**, 244–253.

[52] Grieve, A.P. (1987). Applications of Bayesian software: two examples, *The Statistician* **36**, 283–288.

[53] Gustafson, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data, *Biometrics* **53**, 230–242.

[54] Gustafson, P. (1998). Flexible Bayesian modelling for survival data, *Lifetime Data Analysis* **4**, 281–299.

[55] Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models of life history data, *The Annals of Statistics* **18**, 1259–1294.

[56] Ibrahim, J.G., Chen, M.-H. & MacEachern, S.N. (1999). Bayesian variable selection for proportional hazards models, *The Canadian Journal of Statistics* **27**, 701–717.

[57] Ibrahim, J.G., Chen, M.-H. & Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.

[58] Ibrahim, J.G., Chen, M.-H. & Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine studies, *Statistica Sinica* **14**, 863–883.

[59] Johnson, W. & Christensen, R. (1989). Nonparametric Bayesian analysis of the accelerated failure time model, *Statistics and Probability Letters* **7**, 179–184.

[60] Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data, *Journal of the Royal Statistical Society, Series B* **40**, 214–221.

[61] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[62] Kim, S.W. & Ibrahim, J.G. (2000). On Bayesian inference for parametric proportional hazards models using noninformative priors, *Lifetime Data Analysis* **6**, 331–341.

[63] Kleinman, K.P. & Ibrahim, J.G. (1998a). A semiparametric Bayesian approach to the random effects model, *Biometrics* **54**, 921–938.

[64] Kleinman, K.P. & Ibrahim, J.G. (1998b). A semiparametric Bayesian approach to generalized linear mixed models, *Statistics in Medicine* **17**, 2579–2596.

[65] Kuo, L. & Mallick, B.K. (1997). Bayesian semiparametric inference for the accelerated failure-time model, *The Canadian Journal of Statistics* **25**, 457–472.

[66] Kuo, L. & Smith, A.F.M. (1992). Bayesian computations in survival models via the Gibbs sampler, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer Academic, Boston, pp. 11–24.

[67] Laud, P.W., Damien, P. & Smith, A.F.M. (1998). Bayesian nonparametric and covariate analysis of failure time data, in *Practical Nonparametric and Semiparametric Bayesian Statistics*, D. Dey, P. Muller & D. Sinha, eds. Springer-Verlag, New York, pp. 213–225.

[68] Laud, P.W., Smith, A.F.M. & Damien, P. (1996). Monte Carlo methods for approximating a posterior hazard rate process, *Statistics and Computing* **6**, 77–83.

[69] Law, N.J., Taylor, J.M.G. & Sandler, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure, *Biostatistics* **3**, 547–563.

[70] Leonard, T. (1978). Density estimation, stochastic processes and prior information, *Journal of the Royal Statistical Society, Series B* **40**, 113–146.

[71] MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior, *Communications in Statistics – Theory and Methods* **23**, 727–741.

[72] Mallick, B.K., Denison, D.G.T. & Smith, A.F.M. (1999). Bayesian survival analysis using a MARS model, *Biometrics* **55**, 1071–1077.

[73] Newton, M.A., Czado, C. & Chappell, R. (1996). Bayesian inference for semiparametric binary regression, *Journal of the American Statistical Association* **91**, 142–153.

[74] Nieto-Barajas, L.E. & Walker, S.G. (2000). *Markov Beta and Gamma Processes for Modeling Hazard Rates*, Technical Report, Imperial College, London.

[75] Qiou, Z., Ravishanker, N. & Dey, D.K. (1999). Multivariate survival analysis with positive frailties, *Biometrics* **55**, 637–644.

[76] Ramgopal, P., Laud, P.W. & Smith, A.F.M. (1993). Nonparametric Bayesian bioassay with prior constraints on the shape of the potency curve, *Biometrika* **80**, 489–498.

[77] Sahu, S.K., Dey, D.K., Aslanidou, H. & Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data, *Lifetime Data Analysis* **3**, 123–137.

[78] Sargent, D.J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting, *Biometrics* **54**, 1486–1497.

[79] Shih, J.A. & Louis, T.A. (1995). Assessing gamma frailty models for clustered failure time data, *Lifetime Data Analysis* **1**, 205–220.

[80] Sinha, D. (1998). Posterior likelihood methods for multivariate survival data, *Biometrics* **54**, 1463–1474.

[81] Sinha, D. (1993). Semiparametric Bayesian analysis of multiple event time data, *Journal of the American Statistical Association* **88**, 979–983.

[82] Sinha, D. (1997). Time-discrete beta process model for interval-censored survival data, *Canadian Journal of Statistics* **25**, 445–456.

[83] Sinha, D., Chen, M.-H. & Ghosh, S.K. (1999). Bayesian analysis and model selection for interval-censored survival data, *Biometrics* **55**, 585–590.

[84] Sinha, D. & Dey, D.K. (1997). Semiparametric Bayesian analysis of survival data, *Journal of the American Statistical Association* **92**, 1195–1212.

[85] Sinha, D., Ibrahim, J.G. & Chen, M.-H. (2002). Bayesian models for survival data from cancer prevention studies, *Journal of the Royal Statistical Society, Series B* **63**, 467–477.

[86] Sinha, D., Ibrahim, J.G. & Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood, *Biometrika* **90**, 629–641.

[87] Stangl, D.K. (1995). Prediction and decision making using Bayesian hierarchical models, *Statistics in Medicine* **14**, 2173–2190.

[88] Stangl, D.K. & Greenhouse, J.B. (1998). Assessing placebo response using Bayesian hierarchical survival models, *Lifetime Data Analysis* **4**, 5–28.

[89] Susarla, V. & Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations, *Journal of the American Statistical Association* **71**, 897–902.

[90] Walker, S.G., Damien, P., Laud, P.W. & Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion), *Journal of the Royal Statistical Society, Series B* **61**, 485–528.

[91] Walker, S.G. & Mallick, B.K. (1996). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing, *Journal of the Royal Statistical Society, Series B* **59**, 845–860.

[92] Walker, S.G. & Mallick, B.K. (1999). A Bayesian semiparametric accelerated failure time model, *Biometrics* **55**, 477–483.

[93] Wang, Y. & Taylor, J.M.G. (2001). Jointly modelling longitudinal and event time data, with applications to AIDS studies, *Journal of the American Statistical Association* **96**, 895–905.

(*See also* **Bayesian Approaches to Cure Rate Models**; **Bayesian Model Selection in Survival Analysis**)

JOSEPH G. IBRAHIM, MING-HUI CHEN &
DEBAJYOTI SINHA

Behrens–Fisher Problem

Consider the problem of testing for a difference between the means of two normal distributions (*see* **Hypothesis Testing**). Let $y_{j1}, \ldots, y_{jn_j}$ denote a random sample of size n_j from a normal distribution with mean μ_j and variance $\sum_j^2$, $j = 1, 2$. Furthermore, let the mean and variance of sample j be $\overline{y}_j = \sum_{i=1}^{n_j} y_{ji}/n_j$ and $s_j^2 = \sum_{i=1}^{n_j}(y_{ji} - \overline{y}_j)^2/(n_j - 1)$, respectively, $j = 1, 2$. Attention is focused on the parameter of interest, $\delta = \mu_2 - \mu_1$. If the two population variances are assumed to be equal, then the usual two-sample t test may be adopted to test $H_0: \delta = 0$ vs. $H_a: \delta \neq 0$, using the standard pooled variance estimate.

When no assumptions about the variances of the two populations are made, this is known as the Behrens–Fisher problem. The standard advice given in introductory textbooks is first to carry out a test of the hypothesis $\sigma_1^2 = \sigma_2^2$ based on the quantity $F = s_2^2/(s_1^2)$. Under the hypothesis of common variances, this statistic has an ***F* distribution** with $(n_2 - 1, n_1 - 1)$ degrees of freedom. If the hypothesis is not rejected, then it is often recommended to test the equality of the means using the usual two-sample t test (*see* **Student's *t* Distribution**). Since a failure to reject the hypothesis of equal variances does not imply that the variances are indeed equal, the validity of this procedure has been questioned.

A well-defined procedure for testing the difference between the means of two normal distributions can, in fact, be defined provided the variance ratio $\sigma_2^2/\sigma_1^2 = \rho^2$ is specified. The relevant distributional results are

$$t(\delta; \rho) = \frac{\overline{y}_2 - \overline{y}_1 - \delta}{\left\{\left(\frac{1}{n_1} + \frac{\rho^2}{n_2}\right)\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2/\rho^2}{n_1+n_2-2}\right]\right\}^{1/2}} \sim t_{n_1+n_2-2} \quad (1)$$

and

$$F(\rho) = \frac{s_2^2}{s_1^2\rho^2} \sim F_{n_2-1, n_1-1}. \quad (2)$$

Calculation of $t(0; \rho)$ leads to a significance level for a test of $\delta = 0$ as a function of ρ. Confidence intervals for δ may be similarly derived for specified ρ [1].

When ρ is not specified most inference procedures are based on the Behrens–Fisher statistic,

$$t(\delta; \hat{\rho}) = \frac{\overline{y}_2 - \overline{y}_1 - \delta}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}}, \quad (3)$$

which may be obtained by replacing ρ^2 with $\hat{\rho}^2 = s_2^2/s_1^2$ in (1). Behrens's solution [2] was shown by Fisher [5] to arise from a **fiducial** probability calculation. This is equivalent to obtaining the distribution of $t(\delta, \hat{\rho})$ by averaging out ρ over its fiducial distribution. The use of a fiducial distribution is controversial and this procedure is not generally accepted. It can also be obtained, however, by a **Bayesian** calculation which involves a uniform **prior** on $\log \rho$.

Similar **hypothesis testing** is that for which, under the null hypothesis, the probability of exceeding a specified critical value is equal to the required size, α, and does not depend on values for parameters other than the one of interest. Welch [10] derived an approximately similar test for the Behrens–Fisher problem which involved the statistic $t(\delta, \hat{\rho})$, but used different critical values than those advocated by Behrens and Fisher (*see* **Aspin–Welch Test**). Although Linnik [7] has shown that no similar tests which use the **sufficient statistics** in a reasonably smooth way exist, it has been shown numerically that Welch's test is very nearly similar [11].

The requirement of similarity is viewed as unnecessary by some and was the basis of Fisher's criticism of Welch's test. In particular, in the case $n_1 = n_2 = 7$ and under the null hypothesis of common means, the probability of the test statistic exceeding Welch's 0.1 level critical value, conditional on $\hat{\rho}^2 = 1$, is greater than or equal to 0.108 [6]. Thus, if $\hat{\rho}^2 = 1$, the probability of a type I error with Welch's test is known to be greater than the nominal value, and hence the test is anticonservative. Standard arguments of **conditional** inference suggest that the fact that the nominal error rate is valid on average, and upon repeated applications, is not relevant for the interpretation of a particular data set for which it is known that $\hat{\rho}^2 = 1$. Robinson [8] has shown that this situation does not arise with Behrens's test, but that the test may be somewhat conservative. In general, the Behrens–Fisher problem continues to provide a focus for discussions on the foundations of statistical inference.

From a practical point of view, the debate surrounding the Behrens–Fisher problem is most relevant for the analysis of very small samples. Indeed, for larger samples the approximate solution based on normal distributions and neglecting the errors in estimating variances should be satisfactory [3, p. 155]. However, in some settings, the relevance of testing for a difference in means when variances differ dramatically warrants serious consideration.

Another consideration from the data analytic point of view is that $t(0, \rho)$ can be calculated, and significance levels determined, for a range of plausible values of ρ. This range could be determined from a **likelihood ratio** based **confidence interval** for ρ through the use of (2) [9]. The alternative of using a posterior density for ρ based on **empirical Bayes** methodology is suggested in [4]. This would also facilitate a summary inference, although if qualitative conclusions vary widely as a function of ρ, a summary significance level may not provide all of the relevant information.

For illustration, consider the examples in [9]. Summary data from the two examples are given in Table 1.

There is negligible evidence against the hypothesis of equal variances in either example. In Example A, the Behrens–Fisher significance level is 0.005 and the value from Welch's test is 0.0015. A 90% confidence interval for ρ is (0.65, 1.91), and significance levels based on $t(0, \rho)$ for ρ values of 0.65, 1.00 (corresponding to the usual two-sample t test), $\hat{\rho} =$ 1.16, and 1.91 are 0.009, 0.002, 0.001, and 0.002, respectively. Thus inferences about δ are not much influenced by the value of ρ and the choice of the summary procedure does not appear to be critical.

Table 1

	Sample 1			Sample 2		
Example	n_1	$\overline{y}_1$	s_1	n_2	$\overline{y}_2$	s_2
A	9	22.20	0.6498	15	21.12	0.7541
B	8	0.8081	0.1369	4	0.4940	0.1629

In Example B, the Behrens–Fisher significance level is 0.037 and the corresponding value from Welch's statistic is 0.021. The 90% confidence interval for ρ is considerably wider at (0.53, 3.30). The significance levels for ρ values of 0.53, 1.00, $\hat{\rho} =$ 1.19, and 3.30 are 0.006, 0.005, 0.008, and 0.145, respectively. In this case, different plausible values for ρ lead to different qualitative conclusions, and this should be considered in interpreting any summary significance level.

References

[1] Barnard, G.A. (1984). Comparing the means of two independent samples, *Applied Statistics* **33**, 266–271.

[2] Behrens, B.V. (1929). Ein betrag zur fehlenberechnung bei wenigen beobachtungen, *Landwistschaftlich Jahrbuch* **68**, 807–836.

[3] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[4] Duong, Q.P. & Shorrock, R.W. (1996). On Behrens-Fisher solutions, *Statistician* **45**, 57–63.

[5] Fisher, R.A. (1935). The fiducial argument in statistical inference, *Annals of Eugenics* **6**, 391–398.

[6] Fisher, R.A. (1956). On a test of significance in Pearson's Biometrika tables (no. 11), *Journal of the Royal Statistical Society, Series B* **18**, 56–60.

[7] Linnik, Y.V. (1968). *Statistical Problems with Nuisance Parameters*. American Mathematical Society. New York. Translations of mathematical monographs, no. 20 (from the 1966 Russian edition).

[8] Robinson, G.K. (1976). Properties of Student's t and of the Behrens-Fisher solution to the two means problem, *Annals of Statistics* **4**, 963–971.

[9] Sprott, D.A. & Farewell, V.T. (1993). The difference between two normal means, *American Statistician* **47**, 126–128.

[10] Welch, B.L. (1947). The generalization of "Student's" problem when several different population variances are involved, *Biometrika* **34**, 28–35.

[11] Welch, B.L. (1956). Note on some criticisms made by Sir Ronald Fisher, *Journal of the Royal Statistical Society, Series B* **18**, 297–302.

RICHARD J. COOK & VERN T. FAREWELL

Benefit/Risk Assessment in Prevention Trials

Benefit/risk assessment (B/RA) is a mathematical procedure to estimate the probability of detrimental outcomes, beneficial outcomes and the net-effect anticipated from exposure to a given agent (*see* **Multiple Endpoints in Clinical Trials**). B/RAs of health-related outcomes are used for public health planning, decision-making regarding health care financing (*see* **Cost-effectiveness in Clinical Trials**) and therapeutic decision-making in clinical practice [12, 24]. Information obtained from B/RAs based on findings from controlled clinical trials, particularly those with double-masking of treatment (*see* **Cost-effectiveness in Clinical Trials**), are most informative for health care planning and decision-making because such information is less likely to be biased than is information obtained from observational studies [20, 23, 26]. Thus, B/RAs in prevention trials are an excellent source of information to use as the basis for the types of health care planning and decision-making mentioned above. However, in a **prevention trial** a B/RA is primarily performed as a supplement for planning, monitoring and analyzing the trial. It is designed to provide a global assessment of all potential beneficial and harmful effects that may occur as a result of a treatment that is being evaluated as a means to reduce the incidence of some particular disease or condition. The Women's Health Initiative (WHI), the Breast Cancer Prevention Trial (BCPT) and the Study of Tamoxifen and Raloxifene (STAR) are examples of large-scale, multicenter prevention trials (*see* **Multicenter Trials**) which included B/RA as part of the trial methodology [7, 8, 22, 28, 29].

Compared with treatment trials, the need for the type of information provided by a B/RA may be greater in prevention trials. This situation exists because prevention trials usually involve healthy persons among whom only a small proportion may develop the disease of primary interest during the course of the trial [5–7, 13, 22, 29, 30]. As such, all participants are subjected to the risks of therapy during the course of the trial, but relatively few will receive a preventive benefit from the therapy. In this setting, the use of B/RAs provides an additional mechanism to ensure that all participants comprehend the full extent of potential benefits and risks, and that they make a well-informed decision about the interchange of benefits and risks they are willing to accept by participating in the trial (*see* **Ethics of Randomized Trials** and **Medical Ethics and Statistics**). The use of B/RA in prevention trials also provides a method to evaluate the global effect of the therapy as a safeguard against subjecting trial participants to an unforeseen harmful net effect of treatment. Once the results of the trial are known and the true levels of benefits and risks of the therapy have been established, the individualized B/RA employed in the trial can become the basis for the development of a B/RA methodology that could be used in the clinical setting to facilitate the decision-making process for individuals and their health care providers who may be considering the use of preventive therapy. The trial results can also be used to develop a population-based B/RA to identify changes in patient loads for the outcomes affected by the preventive therapy that would be anticipated as health care professionals incorporate the use of the preventive therapy into their clinical practice. This information could in turn be used for decision-making regarding the planning for and use of health care resources.

Types of B/RAs Performed in Prevention Trials

In a prevention trial a B/RA can take one of three forms which can be classified according to the nature of the population that constitutes the basis for the assessment. These include assessments based on the general population, those based on the trial cohort and those based on an individual trial participant. Each of these forms of B/RA is performed for specific purposes, namely to support various aspects of the conduct of the trial.

A B/RA based on the general population is often performed pre-trial as part of the justification for initiating the trial. The purpose of this form of assessment is to demonstrate the potential net health benefit to society that could be obtained if the therapy being evaluated in the trial actually exhibits the efficacy that is anticipated. This type of assessment is the most generalized form. It is usually accomplished by estimating effects on a national basis assuming the therapy is administered to all susceptible individuals or to a subset of high-risk individuals and demonstrating that there is a significant net benefit when

comparing the number of cases prevented with the estimates for the number of additional cases of detrimental outcomes that may be caused as a side-effect of the therapy.

A B/RA based on the trial cohort is performed during the course of trial implementation as part of the safety monitoring effort (*see* **Data and Safety Monitoring**). It can be accomplished in a regimented fashion as part of the formal plan for the interim monitoring of the trial or as an informal tool used by the data monitoring committee (*see* **Data and Safety Monitoring**) to assess the overall safety of the therapy being evaluated. This type of assessment is not usually necessary during a trial if the anticipated effects from the therapy involve only a few outcomes or if the anticipated beneficial effects substantially outweigh the anticipated detrimental effects. However, in complex situations where the anticipated outcomes affected by the therapy involve multiple diseases or conditions and/or the magnitude of the anticipated net benefit may not be large, a B/RA based on the trial cohort can be a very useful supplement for trial surveillance as a method of monitoring the global effect of all beneficial and detrimental outcomes combined (*see* **Multiplicity in Clinical Trials**). A notable difference between a B/RA based on the general population and one based on the study cohort is in the nature of the measures that are provided by these two forms of assessment. A risk assessment based on a general population provides a measure of the theoretical net effect of the therapy from estimates of anticipated beneficial and detrimental outcomes. In contrast, a risk assessment based on the trial cohort determines the observed net effect of therapy based on outcomes actually experienced by the cohort during the course of the trial.

A B/RA based on an individual trial participant is similar to that of the population-based assessment in that it is also a theoretical estimate. In this case the assessment is not made for the general population, but instead for a specific subpopulation of persons who have the same risk factor profile (age, sex, race, medical history, family history, etc.) for the anticipated beneficial and detrimental outcomes as that of a particular individual participating in the trial. Information from this type of assessment is used to facilitate the communication to each potential trial participant of the nature of the benefits and risks that are anticipated for them as a result of taking therapy during trial participation. This type of individualized B/RA is used in prevention trials when the nature of anticipated effects is complex and benefit/risk communication is a more difficult task due to the interplay of multiple beneficial and detrimental outcomes. When it is used in this manner, it becomes an integral part of the process to obtain informed consent for each individual's participation in the trial.

Alternative Structures of the Benefit/Risk Algorithm Used in Prevention Trials

The core components of a B/RA are the measures of the treatment effect for each of the health outcomes (*see* **Outcome Measures in Clinical Trials**) that may be affected by the therapy being assessed. In this instance the treatment effect is defined as the difference between the probability that the outcome will occur among individuals who do not receive the therapy being evaluated (p_0) and the probability that the outcome will occur among those who do receive the therapy (p_1). For outcomes beneficially affected by therapy, the treatment effect ($p_0 - p_1$) will have a positive sign, representing cases prevented by therapy. For outcomes detrimentally affected by therapy, the treatment effect will have a negative sign, representing cases caused by therapy.

In its simplest structure, the benefit/risk analysis is summarized by an index of net effect (Δ) as the summation of treatment effects for all outcomes affected. If there are I number of outcomes affected by therapy, then the basic algorithm for the B/RA is defined as:

$$\Delta_1 = \sum_{i=1}^{I} (p_{0,i} - p_{1,i}). \qquad (1)$$

When the sign of the index of net effect is positive, the therapy exhibits an overall beneficial health effect. When the sign is negative, the therapy has an overall detrimental effect.

When dealing with a B/RA based on the trial cohort, the probabilities of (1) are obtained directly from the observations in the trial. When dealing with assessments based on the general population or the individual trial participant, the probabilities utilized as anticipated values among those who do not receive therapy (p_0) are usually taken from some type of national database or from prospective studies

of large populations that included measurements of the outcomes of interest. The probabilities used in these latter types of assessments as anticipated values among those who receive therapy are determined by multiplying the anticipated probability among those not treated by the relative risk (untreated to treated) anticipated as the treatment effect. For example, if we anticipate that treatment will reduce the incidence of a particular outcome by 35% then the anticipated relative risk would be 0.65 and the value used for p_1 would be 0.65 p_0. If we anticipate that treatment will increase the incidence of an outcome by 30%, then the anticipated relative risk would be 1.30 and the value used for p_1 would be 1.30 p_0. Estimates of the anticipated treatment effects for each outcome are taken from the literature dealing with pharmacokinetics, animal studies (*see* **Preclinical Treatment Evaluation**) and studies in humans undertaken as preliminary investigations of the therapy as an agent to prevent the disease, or from human studies in which the therapy was being used as an agent for the treatment of disease (*see* **Phase I Trials** and **Phase II Trials**).

In the prevention trial setting it is often advantageous to utilize structures of the benefit/risk algorithm other than that defined in (1). Since a B/RA based on the trial cohort is meant to be performed as part of the effort to monitor safety during the trial, an alternative structure of the benefit/risk algorithm can be used to facilitate this effort. This structure incorporates a standardization of the differences between the probabilities among those receiving and not receiving the therapy being evaluated. In this situation the index of net effect is defined as:

$$\Delta_2 = \frac{\sum_{i=1}^{I}(p_{0,i} - p_{1,i})}{\text{s.e.}\left[\sum_{i=1}^{I}(p_{0,i} - p_{1,i})\right]}. \qquad (2)$$

In this structure, the index of net-effect (Δ_2) becomes a standardized value with an N(0,1) distribution. As such, the standardized values are Z-scores. Critical values of this index of net effect in the form of Z and $-Z$ can then be used as cut-points for global monitoring indicating that there is a significant net effect that is beneficial or detrimental, respectively.

In addition to that for the standardized score, there are other structures of the algorithm used in the prevention trial setting. Instead of expressing the differences between those treated and not treated in terms of the probabilities of the anticipated outcomes, an alternative structure of the algorithm is that based on differences between treatment groups in terms of the number of cases of the outcomes. The structure of the algorithm based on the difference in terms of the number of cases is defined as:

$$\Delta_3 = \sum_{i=1}^{I}(n_{0,i} - n_{1,i}), \qquad (3)$$

where n_0 is the number of cases occurring among those who do not receive the therapy being evaluated and n_1 is the number of cases among those who do receive the therapy. This structure of the algorithm is that which is utilized to perform B/RAs based on the general population. This type of assessment is meant to justify the need for a trial by demonstrating the potential health benefit to society. The net effect to society is more effectively communicated to a greater proportion of individuals when it is expressed as the number of cases prevented from (3) than when it is expressed as the probability from (1). This facilitation of risk communication is also the reason that (3) is preferred over (1) for B/RAs based on individual trial participants where the specific goal of the assessment is to enhance the individual's comprehension of benefits and risks that may be experienced as a result of trial participation.

For a population-based assessment, the numbers of cases in (3) are determined by multiplying the anticipated probabilities $p_{0,i}$ and $p_{1,i}$ by the number of persons in the general population, frequently that of the total US, to obtain an estimate of the number of cases that may be prevented or caused by treatment for each outcome on an annual basis. For an individual participant-based assessment, the numbers of cases in (3) are determined by multiplying the anticipated probabilities by a fixed sample size (N) of theoretical individuals who all have a risk factor profile similar to that of the individual being assessed. A fixed period of follow-up time (t) is assumed to obtain the number of cases prevented or caused by treatment in t years among N individuals. In scenarios where the length of follow-up is long and/or the population is of older age, the estimation of $n_{0,i}$ and $n_{1,i}$ should incorporate the competing risk of mortality that would be anticipated. If d is the probability of dying and RR is the relative risk

anticipated for the outcome of interest, the adjusted expected number of cases among those not treated can be calculated as:

$$n_{0,i} = N\left\{\frac{p_{0,i}}{(p_{0,i}+d_i)}\right\}[1-\exp\{-t(p_{0,i}+d_i)\}], \tag{4}$$

and the adjusted expected number of cases among those treated can be calculated as:

$$n_{1,i} = N\left\{\frac{RRp_{0,i}}{(RRp_{0,i}+d_i)}\right\} \times [1-\exp\{-t(RRp_{0,i}+d_i)\}]. \tag{5}$$

In most prevention trials the outcomes that are potentially affected by the therapy being evaluated encompass a wide range of severity. A simple adding together of the risks of beneficial and detrimental effects without including a consideration of the relative severity of the outcomes may not be appropriate or desirable. For example, suppose a therapy is anticipated to prevent breast cancer and hip fractures, but may cause an increase in uterine cancer and cataracts. Is it appropriate to equate one case of breast cancer prevented to one case of cataracts caused or equate one case of hip fracture prevented to one case of uterine cancer caused? In situations where it is important to include a consideration of the relative severity of the outcomes affected by the therapy, the equations described above for determining the index of net effect can be modified to incorporate a weighting of the outcomes. If w_i is used to represent the weight for each of the I outcomes, then the modification to (3) to incorporate weighting of the outcomes is:

$$\Delta_4 = \sum_{i=1}^{I} w_i(n_{0,i} - n_{1,i}). \tag{6}$$

Equations (1) and (2) can be modified in a similar fashion by including w_i as a multiplier of the quantity of difference in the probabilities.

Methodological and Practical Issues with B/RA in Prevention Trials

There are several issues to be faced when performing a B/RA in a prevention trial. These issues concern the variability of the index of net effect, weighting the outcomes by severity, estimating the index of net effect for individuals with specific profiles of risk factors and communicating the findings of a B/RA to individual participants. Some discussion of each of these issues is presented below.

The estimates of $p_{0,i}$ and $p_{1,i}$ used in a B/RA have a variability associated with them in terms of the strength of evidence supporting the treatment effect and in terms of the precision of the treatment effect. If this variability is substantial, then it may be necessary to incorporate consideration of the variability into the B/RA. Freedman et al. [8, 25] have described a Bayesian approach to incorporating a measure of variability into the index of net effect when it is measured in the form of weighted, standardized probabilities. They assume a skeptical prior distribution based on the strength of the preliminary evidence used as the anticipated treatment effect for each outcome potentially affected by therapy. Gail et al. [10] have described a method to incorporate a measure of variability into the estimate of the index of net effect measured in the form of a weighted number of cases. Their method involves bootstrapping, based on the 95% confidence intervals of the anticipated relative risk associated with treatment for each outcome, to determine the probability that the net number of cases is greater than zero.

The values used for weighting the differences between those treated and not treated can be based on a utility function related to the severity of the outcome, preferences in terms of levels of risk acceptability or other considerations. However, the best choice of a utility function is not always obvious. A measure of mortality such as the case-fatality ratio is one possible utility. If this type of weighting is used, then the choice of the one-year, five-year or ten-year case-fatality ratios would be an issue because the relative weighting of the outcomes could likely be very different depending on which time period for case-fatality is used. Also, weights based on case-fatality would eliminate the consideration of any nonfatal outcome, which would not be preferable if there were several nonfatal outcomes of interest or if a nonfatal outcome has a significant impact on morbidity. Issues also arise with the use of rankings based on the impact on quality of life or preferences regarding the acceptability of risk [1, 11]. The issues with these utilities arise because the rankings are often subjective in nature, based on the opinions of a relatively small panel of individuals, and it

is possible that the rankings of outcomes could differ substantially depending on the population from whom the opinions are ascertained [2, 15, 16]. In light of these issues, attempting to identify a basis for weighting a B/RA is a practical problem that can be difficult to resolve. The preferred choice for any particular trial could differ from one group of individuals to another. As such, if a B/RA is planned as part of trial monitoring, it is essential that the data monitoring committee reviews and reaches a consensus regarding the proposed weighting before it initiates review of the outcome data.

To accomplish the individualization desired for B/RAs based on individual trial participants, it is necessary to provide estimates of effect specific to the individual's full spectrum of risk factors for each of the outcomes expected to be affected by the therapy of interest. A problem likely to be faced when performing individualized assessments is the unavailability of probability estimates specific to the individual's full set of risk factors. For outcomes with several relevant risk factors to be considered or for outcomes that have not been studied in diverse populations, estimates of the outcome probabilities for a specific category of risk factor profiles may not exist. In some cases, multivariate regression models are available that can be used to predict probabilities of outcomes for specific risk factor profiles from data based on the general population. Examples of such models include those for heart disease, stroke and for breast cancer [4, 9, 14, 17–19]. However, the models currently available are primarily limited to those for the more common diseases and are not generally applicable to all race and sex populations. Also, relatively few of these models have been well validated. Thus, in practice it is often necessary to use estimates of outcome probabilities for individualized B/RAs that are taken from populations that are more representative of the general population than of the population specific to the risk factor profile of the individual being assessed. When this is the case, the limitations of the methodology need to be recognized and used in this light. Nonetheless, a B/RA that has been individualized to the extent possible is more informative to a trial participant than one based on the general population. Additional discussions of the limitations of individualized B/RAs can be found in presentations concerning individualized B/RAs for the use of tamoxifen to reduce breast cancer risk [3, 4, 27].

Communicating the results of a B/RA to an individual is a skilled task. An effort must be made to provide information in a manner that facilitates the individual's comprehension [21]. Tools are needed to facilitate this effort. These tools must be developed before the initiation of the trial and included as part of the protocol approved by the Institutional Review Board. Relatively little work has been done in the area of developing tools for communicating the benefits and risks of participation in a prevention trial. However, some tools have been developed that serve as examples for future development.

Tools to enhance the communication of B/RA information to women screened for participation were developed for use in the BCPT [7, 22]. Since the conclusion of this trial, the tools were refined for use in the STAR trial [28]. Table 1 provides an example of the type of tool used in the STAR trial to inform potential participants regarding their individualized B/RA. This tool was developed based on the principles put forth by the participants of the National Cancer Institute's workshop convened to develop information to assist in counseling women about the benefits and risks of tamoxifen when used to reduce the risk of breast cancer. This workshop and the specific methodology used for the B/RA are described by Gail et al. [10]. There were several key working premises that guided the development of the STAR trial tool displayed in Table 1. The premises were considerations of form and format to facilitate the participant's comprehension of their individualized B/RA. These working premises were to: (1) avoid the use of probabilities and relative risk as these concepts are not readily understood by the nonstatistician; (2) provide information for each outcome anticipated to be affected by therapy; (3) group the information presented by severity of the outcomes; (4) provide detailed information for the outcomes with more severe consequences and provide an estimate of effects among those not treated so the individual can understand the context in which to place the expected treatment effects; and (5) limit the tool to one page of data presentation to reduce the amount of data overload perceived by the individual. The precise considerations involved in any prevention trial may differ; however, working premises of this nature designed to enhance comprehension should always be employed when developing tools to communicate B/RA information to potential trial participants.

Table 1 Example of data presentation tool for communicating the benefits and risks of tamoxifen therapy

The information below provides the number of certain events that would be expected during the next five years among 10 000 untreated women of your age (age_X), race ($race_Y$) and five-year breast cancer risk ($risk_Z$). To help you understand the potential benefits and risks of treatment, these numbers can be compared with the numbers of expected cases that would be prevented or caused by five years of tamoxifen use

Severity of event	Type of event	Expected number of cases among 10 000 untreated women	Expected effect among 10 000 women if they all take tamoxifen for five years
			Potential benefits
Life-threatening events	Invasive breast cancer	$N_{0,1}$ cases expected	$N_{1,1}$ of these cases may be prevented
	Hip fracture	$N_{0,2}$ cases expected	$N_{1,2}$ of these cases may be prevented
			Potential risks
	Endometrial cancer	$N_{0,3}$ cases expected	$N_{1,3}$ more cases may be caused
	Stroke	$N_{0,4}$ cases expected	$N_{1,4}$ more cases may be caused
	Pulmonary embolism	$N_{0,5}$ cases expected	$N_{1,5}$ more cases may be caused
			Potential benefit
Other severe events	*In situ* breast cancer	$N_{0,6}$ cases expected	$N_{1,6}$ of these cases may be prevented
			Potential risk
	Deep vein thrombosis	$N_{0,7}$ cases expected	$N_{2,7}$ more cases may be caused
Other events	*Potential benefits*:	Tamoxifen use may reduce the risk of a certain type of wrist fracture called Colles' fracture by about 39%, and also reduce the risk from fractures of the spine by about 26%.	
	Potential risk:	Tamoxifen use may increase the occurrence of cataracts by about 14%.	

References

[1] Bennett, K.J. & Torrance, G.W. (1996). Measuring health state preferences and utilities: ratings scale, time trade-offs and standard gamble techniques, *in Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 253–265.

[2] Boyd, N.F., Sutherland, H.J., Heasman, K.Z., Tritchler D.L. & Cummings B.J. (1990). Whose utilities for decision analysis?, *Medical Decision Making* **1**, 58–67.

[3] Costantino, J.P. (1999). Evaluating women for breast cancer risk-reduction therapy, in *ASCO Fall Education Book*. American Society of Clinical Oncology, pp. 208–214.

[4] Costantino, J.P., Gail, M.H., Pee, D., Anderson, S., Redmond, C.K. & Benichou, J. (1999). Validation studies for models to project the risk of invasive and total breast cancer incidence, *Journal of the National Cancer Institute* **91**, 1541–1548.

[5] Cummings, S.R., Echert, S., Krueger, K.A., Grady, D., Powles, T.J., Cauley, J.A., Norton, L., Nickelsen, T., Bjarnason, N.H., Morrow M., Lippman M.E., Black, D., Glusman, J.E. & Jordan, V.C. (1999). The effect of raloxifene on risk of breast cancer in postmenopausal women: results from the MORE randomized trial, *Journal of the American Medical Association* **281**, 2189–2197.

[6] Ettinger, B., Black, D.M., Mitlak B.H., Knickerbocker, R.K., Nickelsen, T., Genant, H.K., Christiansen, C., Delmas, P.D., Zanchetta, J.R., Stakkestad, J., Gluer, C.C., Krueger, K., Cohen, F.J., Eckert, S., Ensrud, K.E., Avioli, L.V., Lips, P. & Cummings, S.R. (1999). Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial, *Journal of the American Medical Association* **282**, 637–645.

[7] Fisher, B., Costantino, J.P., Wickerham, D.L., Redmond, C.K., Kavanah, M., Cronin, W.M., Vogel, V., Robidoux, A., Dimitrov, N., Atkins, J., Daly, M., Wieand, S., Tan-Chiu, E., Ford, L. & Wolmark, N. (1998). Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 study, *Journal of the National Cancer Institute* **90**, 1371–1388.

[8] Freedman, L., Anderson, G., Kipnis, V., Prentice, R., Wang, C.Y., Rousouw, J., Wittes, J. & DeMets, D. (1996). Approaches to monitoring the results of long-term disease prevention trials: examples from the Women's Health Initiative, *Controlled Clinical Trials* **17**, 509–525.

[9] Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C. & Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, *Journal of the National Cancer Institute* **81**, 1879–1886.

[10] Gail, M.H., Costantino, J.P., Bryant, J., Croyle, R., Freedman, L., Helzlsouer, K. & Vogel V. (1999). Weighing the risks and benefits of tamoxifen for preventing breast cancer, *Journal of the National Cancer Institute* **91**, 1829–1846.

[11] Guyatt, G., Feeny, D. & Patrick, D. (1993). Measuring health-related quality of life, *Annuals of Internal Medicine* **118**, 622–629.

[12] Haynes, R.B., Sackett, D.L., Gray, J.A.M., Cook, D.J. & Guyatt, G.H. (1996). Transferring evidence from research to practice: 1. The role of clinical care research evidence in clinical decisions, *APC Journal Club* **125**, A14–A15.

[13] Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B. & Vittinghoff, E. (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/Progestin Replacement Study (HER) Research Group, *Journal of the American Medical Association* **280**, 605–613.

[14] Liao, Y., McGee, D.L., Cooper, R.S. & Sutkowski, M.B. (1999). How generalizable are coronary risk prediction models? Comparison of Framingham and two other national cohorts, *American Heart Journal* **137**, 837–845.

[15] Llewellyn-Thomas, H.A. (1995). Patients' health care decision making: a framework for descriptive and experimental investigations, *Medical Decision Making* **15**, 101–106.

[16] Llewellyn-Thomas, H.A., Naylor, C.D., Cohen, M.N., Baskiniski, A.S., Ferris, L.E. & Williams, J.E. (1992). Studying patients' preferences in health care decision making, *Canadian Medical Association Journal* **147**, 859–864.

[17] Lloyd-Jones, D.M., Larson, M.G., Beiser, A. & Levy, D. (1999). Lifetime risk of developing coronary heart disease, *Lancet* **353**, 89–92.

[18] Manolio, T.A., Kronmal, R.A., Burke, G.L., O'Leary, D.H. & Price, T.R. (1996). Short-term predictors of incidence stroke in older adults. The Cardiovascular Health Study, *Stroke* **27**, 1479–1486.

[19] Menotti, A., Jacobs, D.R., Blackburn, H., Krombout, D., Nissinen, A., Nedeljkovic, S., Buzina, R., Mohacek, I., Seccareccia, F., Giampaoli, S., Dontas, A., Aravanis, C. & Toshima, H. (1996). Twenty-five year prediction of stroke deaths in the seven countries study: the role of blood pressure and its changes, *Stroke* **27**, 381–387.

[20] Pocock, S.J. & Elbourne, D.R. (2000). Randomized trials or observational tribulations?, *The New England Journal of Medicine* **342**, 1907–1909.

[21] Redelmeier, D.A., Rozin, P. & Kahneman D. (1993). Understanding patients' decision–cognitive and emotional perspectives, *Journal of the American Medical Association* **270**, 72–76.

[22] Redmond, C.K. & Costantino, J.P. (1996). Design and current status of the NSABP Breast Cancer Prevention Trial, *Recent Results in Cancer Research* **140**, 309–317.

[23] Sackett, D.L. (1997). Bias in analytical research, *Journal of Chronic Diseases* **32**, 51–63.

[24] Simon, G., Wagner, E. & VonKorff, M. (1995). Cost-effectiveness comparisons using "real world" randomized trials: the case of the new antidepressant drugs, *Journal of Clinical Epidemiology* **48**, 363–373.

[25] Spiegelhalter, D.J., Freedman, L. & Parmar, M.K.B. (1994). Bayesian approaches to randomization clinical trials, *Journal of the Royal Statistical Society, Series A* **157**, 357–416.

[26] Steineck, G. & Ahlbom, A. (1992). A definition of bias founded on the concept of the study base, *Epidemiology* **3**, 477–482.

[27] Taylor, A.L., Adams-Cambell, L. & Wright, J.T. (1999). Risk/benefit assessment of tamoxifen to prevent breast cancer – still a work in progress, *Journal of the National Cancer Institute* **19**, 1792–1973.

[28] Wolmark, N., Wickerham, D.L., Costantino, J.P. & Cronin, W. (1999). *NSABP Protocol P2: Study of Tamoxifen and Raloxifene (STAR) for the Prevention of Breast Cancer*. National Surgical Breast and Bowel Project, Pittsburgh, Pennsylvania.

[29] Women's Health Initiative Study Group (1998). Design of the Women's Health Initiative clinical trial and observational study, *Controlled Clinical Trials* **19**, 61–109.

[30] Writing Group for the PEPI Trial (1995). Effects of estrogen/progestin regimens on heart disease risk factors in postmenopausal women: the Post-menopausal Estrogen/Progestin Intervention (PEPI) Trial, *Journal of the American Medical Association* **273**, 199–208.

(*See also* **Adaptive and Dynamic Methods of Treatment Assignment**; **Noncompliance, Adjustment for**)

JOSEPH P. COSTANTINO

Benjamin, Bernard

Born: 8 March 1910, London.
Died: 15 May 2002, London.

Bernard Benjamin had a distinguished career as a statistician, actuary, and demographer, achieving distinction and the highest levels of recognition and honor in each of these fields. Among his particular interests were the effective use of routinely collected data and the application of statistical modeling for making decisions. For example, he was one of United Kingdom's pioneers on applications of statistical methods to nonlife insurance (*see* **Actuarial Methods**).

Bernard Benjamin was born in 1910, the youngest of eight children. He went to school in SE London and went on to study physics part-time at Sir John Cass College (then affiliated to the University of London), graduating in 1933. Meanwhile, he had started work in 1928 as an actuarial assistant to the London County Council (LCC) pension fund, and qualified as an actuary in 1941. From 1936, he worked as a statistician in the public health sections of the LCC until 1943, when military service took him to the Royal Air Force, where he continued his work as a statistician. After the war, he returned to the LCC and public health, and undertook his PhD (also on a part-time basis) on the analysis of tuberculosis mortality.

In 1952, Benjamin was appointed Chief Statistician at the General Register Office (GRO), later to be absorbed into the Office of Population Censuses and Surveys and then into the **Office for National Statistics**, marking his move into **demography**, and promotion to management and the leadership of a major public-sector department. After 11 years at the GRO, he was appointed Director of Statistics at the Ministry of Health in 1963, and then in 1966, became the first Director of the Intelligence Unit of the Greater London Council (GLC). The unit was set up as part of local government reorganization in London, and its task was to make sure that the GLC had "economic and other information at the right time in the right way", to quote Benjamin. In this role, he brought together the entire planning and transportation research staff into a cohesive and effective unit. But this reorganization did not survive without Benjamin's leadership skills and foresight after his retirement in 1970.

A strong theme of Benjamin's later working life was a series of retirements followed by new beginnings. Thus, in 1970, he became Director of Statistical Studies at the newly established Civil Service College. Then in 1973, he retired and joined City University as the Foundation Professor of actuarial science, establishing and designing the first BSc programme in actuarial science in the country. Although Benjamin enjoyed teaching undergraduates, he took particular pleasure in the supervision of PhD students and, over the next decade, he taught a steady stream of research students who were working on statistical methods applied to demography and actuarial science.

In his roles at the GRO, GLC, and Civil Service College, Benjamin was concerned with the collection and analysis of statistics and the presentation of results for the solving of practical problems in public health or demography. He was particularly adept at conveying statistical ideas in a clear manner, without recourse to jargon or indeed mathematical notation.

Benjamin retired in 1975 from City University (because of his wife's failing health), but continued as a visiting professor. On his final departure from City University, he was appointed emeritus professor and awarded an honorary DSc for his contributions to education and research in statistics and actuarial science.

Benjamin's scientific work was extensive. He published over 100 papers in leading statistical, actuarial, and demographic journals over almost 40 years. A notable achievement was his 1954 report on the growth of pension rights and their impact on the national economy, which became the actuarial profession's principal evidence to the 'Phillips Committee' on the economic and financial problems for the provision for old age. When the actuarial profession sought to update this landmark report 30 years later (at a time of government questioning of the role of public pension provision), it was to Benjamin that they turned to lead the research team that produced the monograph entitled *Pensions: The Problems of Today and Tomorrow,* in 1987.

Benjamin was able to write concisely and interestingly and his first drafts were almost of final draft quality. This talent contributed to a series of successful textbooks and monographs: *Elements of Vital Statistics* (1959), *Social and Economic Differentials*

in Fertility (1965), *Social and Economic Factors Affecting Mortality* (1965), *Health and Vital Statistics* (1968), *Demographic Analysis* (1968), *Medical Records* (1977) and *Population Statistics* (1989). The textbooks *Analysis of Mortality and other Actuarial Statistics* (1970, with new editions in 1980 and 1993) and *General Insurance* (1977) have become seminal works of international standing in the actuarial field. His last book, *Mortality on the Move* (1993), appeared at a time of accelerating mortality decline in many industrialized countries, and is now widely cited.

Among his honors and achievements, Benjamin was the UK representative on the UN Population Commission from 1955 to 1963, the honorary consultant in Medical Statistics to the Army, a member of the statistics committee of the Social Science Research Council, and secretary-general of the International Union for the Scientific Study of Population from 1962 to 1963. He was vice president (1963–1966) and president of the Institute of Actuaries (1966–1968), honorary secretary (1956–1963) and president of the **Royal Statistical Society** (1970–1972), and was uniquely awarded the highest honors of both bodies – the Gold Medal (1975) and the **Guy** Medal in Gold (1986), respectively.

At a personal level, Benjamin was modest and self-effacing, yet he was a determined and clear-sighted manager and leader who inspired respect and loyalty. As a colleague, he was both encouraging and supportive, qualities that made him both an excellent PhD supervisor and a research collaborator. He had a number of active hobbies before his sight failed, describing himself both as an 'amateur' pianist and painter. The latter activity gave him much pleasure and his watercolors were of a higher standard than he would admit to.

STEVEN HABERMAN

Berkson, Joseph

Born: May 15, 1899, in Brooklyn, New York.
Died: September 12, 1982, in Rochester, Minnesota.

Joseph Berkson, the sixth child of Russian immigrants, Henry and Jennie (Berkman) Berkson, was educated in the New York public schools and Townsend Harris Hall before attending the College of the City of New York from which he received a B.S. degree in 1920. He obtained an M.A. degree (Physics) from Columbia University in 1922 and two doctoral degrees from Johns Hopkins, an M.D. in 1927 and a Dr.Sc. in statistics in 1928.

Upon completion of these degrees, Dr Berkson accepted positions at Johns Hopkins as an assistant in the School of Hygiene and Public Health and as an Associate in the Institute of Biologic Research. He remained at Johns Hopkins for three years before accepting a position as a Macy Foundation Fellow in Physiology at the Mayo Clinic, which he began in September 1931. Berkson remained at Mayo until his retirement in 1964, first as Acting Director of the Statistics Division and then Head of the Division of Biometry and Medical Statistics, a position he held from January 1, 1934, to July 1, 1964. He was Associate Professor and later Professor (1949) of Biometry in the Mayo Graduate School of Medicine.

Subsequent to the completion of his degrees but prior to coming to Mayo, Berkson published ten manuscripts [1–5, 15, 19, 20, 24, 26]. Several merit comment because they are informative about his lifelong interests. Eight of these first papers [2, 3, 5, 15, 19, 20, 24, 26] are reports of physiological studies (two with the famous Louis Flexner [15, 24]) reflecting Berkson's training and an interest which he retained throughout his career. However, there were harbingers of statistical interests which would blossom in later years. His fourth paper [26] (with the famous epidemiologist Lowell Reed), "The Application of the Logistic Function to Experimental Data", provides the first evidence of his interest in the **logistic functions**, an interest that would reemerge much later in his professional life.

His fifth paper [1] involved a probability nomogram (another lifelong interest of Berkson's). It is difficult to imagine today with the plethora of computers that calculational shortcuts, nomograms, tables, special graph paper, etc. were all essential aids to an applied statistician throughout much of Berkson's professional career. In addition to publishing several papers including nomograms which he designed, Berkson also designed graph paper with a variety of different scales which was useful for many data analysis problems.

Berkson's first purely statistical publication [4] in the *Annals of Mathematical Statistics* was simply titled **"Bayes' Theorem"**. Symbolically, one of his last papers [14] – published 47 years later in the *International Statistical Review* – was entitled, "My Encounter with Neo-Bayesianism". He was certainly no Neo-Bayesian.

During World War II he served as a colonel in the office of the Air Surgeon General in Washington, DC, and in 1946 he was awarded the Legion of Merit. The accompanying Presidential citation reads:

> Colonel Joseph Berkson, Medical Corps, as Chief of the Statistics Division, Office of the Surgeon, Headquarters Army Airforces, contributed immeasurably to the advancement of medical statistics in developing new methods of presenting and interpreting statistical data as applied to Army Airforce matters. His display of professional skill and high standards of efficiency in disease control contributed to a great extent to the success of the Army Airforce health programs.

At the Mayo Clinic Dr Berkson was known as a maverick who tolerated fools reluctantly and "marched to the beat of his own drummer". He was probably

best known to the Mayo staff for three contributions, the most significant of which was the introduction of the then relatively new Hollerith card and the development of diagnostic and procedure coding systems that were used for over 40 years at Mayo [7, 8]. Berkson rejected the existing, widely touted, classification scheme with the quote, "There are hundreds of diseases in that book no one can get and an equal number of diagnoses no one can find". His ingenious coding system, using the new punch card capability, combined into a single scheme two related but not identical functions. First, cross-indexing of medical conditions was made possible so that specific groups of patient histories could easily be identified and retrieved for research or patient care and, secondly, the designation of main numbers facilitated administrative tabulations. Thousands of research papers by Mayo physicians, surgeons, epidemiologists, and statisticians were made possible because of this system.

From the perspective of his Mayo colleagues, Berkson's second major contribution came in the area of **survival analysis**. His first manuscript on the topic [6] was published in 1934 and involved the appropriate construction of a **life-table** to describe the survival experience of a group of patients following an operation. Several manuscripts were subsequently published using his methodology [18, 21, 22, 25] and his own work culminated in two papers with Robert Gage [16, 17], the first published in 1950 (*Mayo Clinic Proceedings*) and the second in 1952 (***Journal of the American Statistical Association***) on the estimation of survival rates and the construction of survival curves for cancer patients. Berkson insisted that research by Mayo investigators involving patient survival use his methods, and as late as 1980 there was a rule in place at the editorial office of the *Mayo Clinic Proceedings* that such papers had to have the approval of a statistician. He made perhaps an even more important contribution to the field by introducing **Jerzy Neyman** to the problem and by working with Lila R. Elveback on her dissertation. Thus, when Elveback and William F. Taylor both came to Mayo, a nucleus was formed which produced fruitful research in this area and which established a precedent continued by Mayo statisticians to this day.

Finally, everyone at Mayo knew about Berkson's attitude regarding the studies purporting to have established a relationship between smoking and lung cancer (*see* **Smoking and Health**). Interestingly, long before this controversy, Berkson had published a manuscript on tobacco and coronary disease [23]. Being a well trained physiologist and also a statistician, Berkson found the statistical studies associating smoking and lung cancer less than persuasive of causality (*see* **Causation**). Thus, all who were interested, and many who were not, were aware of Berkson's disdain for the studies upon which the smoking/lung cancer connection were based. Berkson's plea for more direct physiological evidence of causality and his observations that arguments associating tobacco use and lung cancer lacked specificity because they could be applied equally to tobacco use and other diseases may seem a little strained now, but they were legitimate at the time. In one of his major papers on the topic [13] in the *Mayo Clinic Proceedings*, Berkson discusses some of his concerns with the smoking/lung cancer data. Here he calls upon his "Berkson Bias" paper (*see* **Berkson's Fallacy**) [10] published some years earlier to argue that studies of hospitalized patients, whether retrospective or prospective, have a likely **selection bias** which may invalidate results. Then, turning to the question of **specificity**, he argues in his eloquent style that the question raised by the findings in the American Cancer Society study of higher death rates among cigarette smokers is not, "Does cigarette smoking cause cancer of the lungs?" so much as "What disease does cigarette smoking not cause?"

Finally, and also eloquently, he offers his argument for biological considerations,

> A disquieting element in the array of observations which have been assembled pointing the finger of accusation at smoking as a cause of lung cancer is that it is so ample, yet it is so exclusively statistical. There are lacking observations of the pathologic process of which the statistics are only the supposed reflection. Actually the American Cancer Society study does not point specifically to association of smoking and cancer, for all specific diseases for which the number of cases permits examination show association, exhibiting a larger death rate among smokers than among nonsmokers. Therefore, if the association found is not statistically spurious and is to be explained as a biologically causative effect, it is not on these findings specifically a carcinogenic effect but something which influences broadly whatever may increase the susceptibility of the organism to fatal disease. Now, the most important known cause of cancer and some other diseases, notably those of the cardiovascular system, is age. We might say speculatively that smoking

> accelerates the rate of living and advances age and age causes cancer. The supposed effect of smoking, if it exists, may be to stimulate those trophic processes, of which little is known, that constitute the biology of aging. The idea is not entirely implausible or without support in existing literature. It is in keeping with Pearl's idea that duration of life is inversely related to the rate of living and with Pearl's own study of the effect of heavy smoking on longevity.

Although his Mayo medical colleagues were unlikely to be aware of it, Berkson made major contributions to theoretical statistics. Nothing seems to have held Berkson's attention throughout his career as much as the logistic function, providing him with an opportunity to delve into some of the more esoteric aspects of statistical inference. Among the 118 papers in his bibliography at least 28 contain some, usually substantial, references to the logistic function. In three of his first 10 manuscripts he used and discussed at length the logistic function as a model of the rate of a chemical reaction. Then, in 1944, he published what he may have assumed would be a little blip in the statistical literature, a manuscript entitled, "Application of the Logistic Function to Bio-assay" [9]. In the ensuing 13 years he published 14 more papers on this and related topics and embarked on a sometimes humorous, sometimes vitriolic, exchange of views with **Fisher**, Finney, and **Bliss**. This exchange may have begun over the relative merits of the probit and logit (*see* **Quantal Response Models**) (apparently Berkson coined this term) in bioassay problems (*see* **Biological Assay, Overview**), but soon escalated into a crusade on Berkson's part to persuade the mathematical statistical community to pay attention to **estimation** principles other than the **maximum likelihood (ML)** and to acknowledge the advantage of calculational simplicity. In this first paper Berkson pointed out that probit analysis, the bioassay analytic method in vogue, assumed the Gaussian distribution (*see* **Normal Distribution**) for certain relationships of dose and susceptibility. If true, this certainly suggests the use of the integrated normal for such analyses. He goes on to point out, however, that the logistic function may have a better theoretic basis because it applies "to a wide range of physicochemical processes".

In applying the integrated normal approach, parameter estimation based on the principle of ML needed iterative and time-consuming calculations. Berkson applied the principle of weighted **least squares** to the estimation of the logistic parameters, which by a special simplification could be obtained directly without iteration. This simplification involved obtaining the logit corresponding to a given mortality rate, and he provided in the 1944 manuscript [9] a nomogram to do this for mortality rates ranging from 0.7% to 99.3%.

It is difficult today to conceive of a world in which such giants of the statistical profession would engage in the rhetoric that followed this publication based on questions of iterations and calculational simplicity. Berkson concluded this first paper with the simple statement that he believed that the work of fitting the logit to be considerably simpler and less time consuming than using probits, as advocated by Bliss and Fisher. In a later exchange [12], Fisher apparently accused Berkson of saying that the probit calculations took 30 times as along as the logit and indicated that in his (Fisher's) experience the two were about equivalent. To this Berkson responded by describing an experiment in which two sets of bioassay data were presented to a "computer" (presumably an individual using a mechanical calculator) with instructions to apply both methods of estimation to the data and to record faithfully the amount of time necessary to complete each. He reported that for one set the probit method required 295 minutes while the logit method required 12 minutes. Berkson admitted that this was less than a 30-fold difference but suggested that it was still substantial and deserving of attention.

Of course the argument between Fisher and Berkson did not hinge on how much time a calculation took. Berkson's minimum logit chi-square approach did not yield maximum likelihood estimates (MLEs), a point he did not deny, but he railed against the notion that there was something sacred about MLEs which precluded use or discussion of alternatives. He was particularly offended by the constant defense of the probit method as producing MLEs when it depended on choice of starting value and number of iterations actually carried out. In one hilarious passage [11, p. 591] he defined a hierarchy of estimators depending on the number of iterations conducted.

> The procedure of "probit analysis" as widely advanced and practiced, consisting of a single cycle of iteration based on a provisional graphical estimate, actually is not a maximum likelihood estimate, but only a somewhat modified graphical solution. Since

> it is a step in the right direction toward the maximum likelihood estimate, perhaps it is entitled to the designation "likelihood estimate." If one or two more iterations are performed, it could be called a "very likelihood estimate"; if as many as 9 iterations are accomplished, as in the example from Irwin and Cheeseman, an "exceedingly likely likelihood estimate", and so forth. A really mathematical maximum likelihood estimate in the present circumstance is rarely attainable, but this estimate appears to be held so noble an objective that perhaps we should be contented only to aspire to achieve it. However, it must be remembered that it is solely to the actual maximum likelihood estimate that the optimum properties pertain, which Finney insistently claims for "probit analysis". These optimum properties do not refer to a "likelihood estimate," nor even to a "practically good enough" maximum likelihood estimate.

On Christmas Eve 1935 Berkson was married to Susanna G. Cacioli, a Mayo translator (Italian) from 1927 to 1948. Although they had no children, the family of Dr Frank Falsetti – Susanna's son by an earlier marriage – has fond memories of times with the Berksons and the role Joe played in their lives.

In 1978 his colleagues and successors in statistics at the Mayo Clinic honored Berkson by naming the main departmental conference room after him. They presented him with two bound volumes of his published works which included solicited comments from some of his illustrious colleagues in the statistical world. These included **W.G. Cochran, Jerzy Neyman**, John Tukey, and others. They all reminisced with Joe about their interactions during the previous decades of their professional lives. Neyman acknowledged that Berkson had introduced him to some of the intriguing statistical problems found in medical research, mentioning specifically work in survival analysis and **competing risks** that Neyman and Evelyn Fix worked on. Cochran said, "An important reason for honoring Joe is that he was about the best writer in the business. I was never in doubt as to what Joe meant in a sentence. Moreover, he was a delight to read, ... Joe was continually reminding professional statisticians of their responsibility for keeping their heads clear." Tukey recalled that, "In the late 1940s there was a round-robin letter group consisting of (Berkson and Tukey), George Brown, Churchill Eisenhart, and Charlie Windsor." Tukey suggests that the correspondence was lively and enjoyable and apparently often heated as Berkson's wife, Susie, reported to him that Joe, upon coming home to find letters would read them and often shout, "They can't do this to me", and off he would go to his upstairs office to begin typing his reply.

Berkson was an active and aggressive statistician. He was also an active member of several statistical societies and was honored by many of them. One of the original group who founded the Biometric Society (*see* **International Biometric Society (IBS)**), he held two early regional offices. He was a fellow of the American Association for the Advancement of Science, the **American Public Health Association**, the **American Statistical Association**, and the **Royal Statistical Society**. In the twilight of his career, he was elected to the National Academy of Sciences (April 1979).

Although Berkson was obviously a statistician, he liked to play the role of the humble medical doctor simply trying to get along in this harsh mathematical world. This image was a facade, as his work illustrates, but it was a facade he maintained throughout his career. There appear to be no publications or written references, where degrees were quoted, in which any other than his M.D. degree was mentioned. His final paper was published in 1978 and he had remained in active contact with his statistical colleagues at Mayo and throughout the world until the last few years of his life.

At the August 1983 American Statistical Association meeting, a session organized by Joe's colleague, William F. Taylor, was presented in his honor. James Grizzle and Lucien LeCam made presentations and Fredrick Mosteller, Churchill Eisenhart, and John Tukey participated with the others in a panel discussion regarding interactions with their friend and colleague Joseph Berkson. Grizzle said the following:

> He (Berkson) had a serious concern about the fundamental properties of the statistics he was using. He did not feel comfortable with them until he understood their finite sample size properties and he was dismayed when others advocated statistical methods that did not, in his opinion, have a completely coherent philosophical (i.e., mathematical) base. There is no doubt that Berkson was a serious productive scientist. He was absolutely tenacious in his investigations. He persisted in martialing mathematical and empirical evidence for his views over approximately a 40-year period. He related his views to new developments in methodology in interesting and useful ways to the end of his career.

References

[1] Berkson, J. (1929). A probability nomogram for estimating the significance of rate differences, *American Journal of Hygiene* **9**, 695–699.

[2] Berkson, J. (1929). The mechanics of teleology, *Quarterly Review of Biology* **4**, 415–419.

[3] Berkson, J. (1929). Growth changes in physical correlation – height, weight, and chest-circumference, males, *Human Biology* **1**, 462–502.

[4] Berkson, J. (1930). Bayes' theorem, *Annals of Mathematical Statistics*, 42–56.

[5] Berkson, J. (1930). Evidence of a seasonal cycle in human growth, *Human Biology* **2**, 523–538.

[6] Berkson, J. (1934). The construction of life tables and the use of the method of calculating survivals after operation, *Proceedings of the Staff Meetings of the Mayo Clinic* **9**, 380–385.

[7] Berkson, J. (1936). A system of codification of medical diagnoses for application to punch cards with a plan of operation, *American Journal of Public Health* **26**, 606–612.

[8] Berkson, J. (1941). A punch card designed to contain written data and coding, *Journal of the American Statistical Association* **36**, 535–538.

[9] Berkson, J. (1944). Application of the logistic function to bio-assay, *Journal of the American Statistical Association* **39**, 357–365.

[10] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data, *Biometric Bulletin* **2**, 47–53.

[11] Berkson, J. (1953). A statistically precise and relatively simple method of estimating bio-assay with quantal response, based on the logistic function, *Journal of the American Statistical Association* **48**, 565–599.

[12] Berkson, J. (1954). Comments on R.A. Fisher, "The analysis of variance with various binomial transformations", *Biometrics* **10**, 130–151.

[13] Berkson, J. (1955). The statistical study of association between smoking and lung cancer, *Proceedings of the Staff Meetings of the Mayo Clinic* **30**, 319–348.

[14] Berkson, J. (1977). My encounter with neo-Bayesianism, *International Statistical Review* **45**, 1–8.

[15] Berkson, J. & Flexner, L.B. (1928). On the rate of reaction between enzyme and substrate, *Journal of General Physiology* **11**, 433–457.

[16] Berkson, J. & Gage, R. (1950). Calculation of survival rates for cancer, *Proceedings of the Staff Meetings of the Mayo Clinic* **25**, 270–286.

[17] Berkson, J. & Gage, R. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association* **47**, 501–515.

[18] Berkson, J., Gage, R. & Wilder, R.M. (1947). Mortality and longevity among patients with diabetes mellitus, *Proceedings of the American Diabetic Association* **133**, 144.

[19] Berkson, J. & Hollander, F. (1930). On the equation for the reaction between invertase and sucrose, *Journal of the Washington Academy of Sciences* **20**, 157–171.

[20] Berkson, J. & Schultz, G. (1929). The question of compensating variability, *American Journal of Physical Anthropology* **13**, 131–137.

[21] Berkson, J., Walters, W., Gray, H.K. & Priestley, J.T. (1952). Mortality and survival in cancer of the stomach: a statistical summary of the experience of the Mayo Clinic, *Proceedings of the Staff Meetings of the Mayo Clinic* **27**, 137–151.

[22] Cabot, H. & Berkson, J. (1939). Neoplasms of the testis: a study of the results of orchidectomy, with and without irradiation, *New England Journal of Medicine* **220**, 192–195.

[23] English, J.P., Willius, F.A. & Berkson, J. (1940). Tobacco and coronary disease, *Journal of the American Medical Association* **35**, 556–558.

[24] Flexner, L.B., Berkson, J., Winters, H. & Wolman, I. (1929). Antitrypic titre in pregnancy and in hyperthyroidism, *Proceedings of the Society of Experimental Biology and Medicine* **26**, 592–595.

[25] MacLean, A.R. & Berkson, J. (1951). Mortality and disability in multiple sclerosis, *Journal of the American Medical Association* **146**, 1367–1369.

[26] Reed, L.J. & Berkson, J. (1929). The application of the logistic function to experimental data, *Journal of Physical Chemistry* **33**, 760–779.

W. MICHAEL O'FALLON

Berkson's Fallacy

Berkson's fallacy, also referred to as Berkson's bias or Berkson's paradox, was first described in 1946 by **Joseph Berkson**, a physician in the Division of Biometry and Medical Statistics at Mayo Clinic. Berkson demonstrated mathematically that an **association** reported from a **hospital-based case–control study** can be distorted if cases and controls experience differential hospital admission rates with respect to the suspected causal factor [1]. His hypothetical example involved the association between two medical conditions – cholecystitis (the suspected causal factor) and diabetes (the outcome of interest). Assuming a hospital-based study, he defined controls as persons with a third condition, refractive errors, not thought to be correlated with cholecystitis. Calculations were based on the following assumptions: (i) the incidence of cholecystitis does not vary between diabetics and persons with refractive errors in the general population (i.e. **relative risk** and **odds ratio** close to 1.0); (ii) hospital admission rates *do* vary between diabetics and persons with refractive errors (5% and 20%, respectively); (iii) persons with cholecystitis experience a 15% probability of hospitalization; and (iv) the probabilities of hospitalization for the three conditions – diabetes, refractive errors, cholecystitis – behave independently and combine together according to the laws of probability. Using these conditions, a hospitalized subset was defined from Berkson's fabricated general population (Tables 1 and 2). Comparison of these two populations reveals that the association between cholecystitis and diabetes apparent in the hospitalized data (odds ratio of 1.89 calculated from data in Table 2) is not indicative of the "true" association (or lack of it) in the general population (odds ratio of 0.90 calculated from data in Table 1).

Berkson's bias remained theoretical and was largely disregarded by epidemiologists [3] until 1978 when Roberts et al. provided the first empirical support using data from household surveys of health care utilization [2]. They examined associations between several medical conditions and documented

Table 1 Cholecystitis and diabetes, hypothetical general population[a]

	Cholecystitis	Not cholecystitis	Total
Diabetes[b]	3 000	97 000	100 000
Refractive errors (not diabetic)	29 700	960 300	990 000
Total	32 700	1 057 300	1 090 000
Cholecystitis in diabetic group			3%
Cholecystitis in control group (refractive errors)			3%
Difference			0%

[a]Adapted from [1].
[b]10 000 of the 100 000 cases of diabetes also have refractive errors (300 cases with cholecystitis and 9700 cases without cholecystitis); the refractive errors control group contains no known cases of diabetes.

Table 2 Cholecystitis and diabetes, hypothetical hospital population[a]

	Cholecystitis	Not cholecystitis	Total
Diabetes	626	6 693	7 319
Refractive errors (not diabetic)	9 504	192 060	201 564
Total	10 130	198 753	208 883
Cholecystitis in diabetic group			8.55%
Cholecystitis in control group (refractive errors)			4.72%
Difference			+3.83%

Hospital admission rates for cholecystitis = 0.15, diabetes = 0.05, refractive error = 0.20.
[a]Adapted from [1].

significant differences between community- and hospital-based risk estimates. These **biases** occurred in both directions.

Berkson's original representation of the admission rate bias was based on a conservative assumption of independence for disease-specific admission rates [1]. In practice, however, a given medical condition may exacerbate a second condition, increasing the probability of differential hospitalization rates. Additionally, other circumstances such as the manifestation and severity of symptoms, treatment regimen of choice, and specialization of certain hospitals (or physicians practicing within certain hospitals) in treating given medical conditions, may further increase the disparity between case and control admission rates.

Berkson's fallacy has been primarily described for a certain subset of analyses in which the association of interest is between two medical conditions. It is conceivable that a similar bias might impact **case–control studies** considering a nonmedical **explanatory variable**, if: (i) the explanatory variable is represented disproportionately in a hospital setting; and (ii) cases and controls experience differential hospital admission rates. Berkson gives a hypothetical example of a study of occupation as an explanatory variable for heart disease in which one occupation group is more likely to present to a hospital for heart disease treatment than another.

Finally, Berkson's bias may have applications beyond the hospital setting. For example, a study of drug use and violent crime in a prison population, using nonviolent criminals as the control group, might result in a different risk estimate than the same study performed in a community-based population.

It is not possible to correct for admission rate bias during analysis. Berkson's bias, like other biases, is a design issue that needs consideration prior to initiating a case–control study drawing participants from a select segment of the general population.

References

[1] Berkson, J. (1946). Limitation of the application of fourfold tables to hospital data, *Biometrics Bulletin* **2**, 47–53.

[2] Roberts, R.S., Spitzer, W.O., Delmore, T. & Sackett, D.L. (1978). An empirical demonstration of Berkson's bias, *Journal of Chronic Diseases* **31**, 119–128.

[3] Sartwell, P.E. (1974). Retrospective studies: a review for the clinician, *Annals of Internal Medicine* **81**, 381–386.

LAURA A. SCHIEVE

Bernard, Claude

Born: July 12, 1813, in St-Julien, France.
Died: February 10, 1878, in Paris, France.

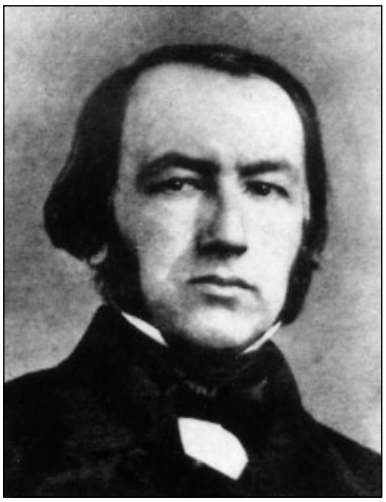

A young man of a very modest extraction (his parents were vineyard workers), Claude Bernard received a pious and humanistic education, and started working as an assistant pharmacist. At 21 he ventured to Paris with the ambition of making a career in literature. He was advised rather to learn a job that would enable him to earn a living. He then turned to medicine, which he studied conscientiously, although clinical work did not much appeal to him. As a resident at the Hotel-Dieu, under François Magendie, he realized what he really wanted to do: experimental work in physiology. For four years he assisted Magendie in his laboratory at the Collège de France, and became an expert at using animal vivisection to trace physiologic facts.

After graduating as an MD with a thesis on gastric juice (1843), he failed to get a teaching position in Paris and reluctantly considered settling as a country doctor in his native province of Beaujolais. In the meantime, he spent his wife's dowry trying to run a private research laboratory of his own, in the latin quarter in Paris. Finally, at the end of 1847, he was appointed as substitute of Magendie at the Collège de France. When in 1852 Magendie retired he was entrusted with his teaching post and his laboratory. He officially became professor of medicine at the Collège after Magendie's death in 1855. The years 1843–1855 were a very creative period, fertile in discoveries: on the nerve control of gastric digestion (1843–1845); on the digestive role of bile and on the functions of cranial nerves (1844–1845); on the mechanism of carbon monoxide intoxication and on the inhibitory action of the vagus nerve on the heart (1846); on the glycogenic function of the liver (1848); on the role of the pancreas and on the metabolism of carbohydrates (1849); on curare poisoning, on vasomotor nerves (1852); and so on. His results on the release of sugar by the liver were presented as a thesis for the doctorate in science (1853). A chair of general physiology was then created for him at the Faculty of Sciences of the University of Paris (1854), and later transferred to the Museum d'Histoire Naturelle (1868), while he went on teaching experimental medicine at the Collège de France. In the later part of his life, Bernard was elected to the Académie des Sciences (1854), to the Académie de Médecine (1861), and to the Académie Française (1869). He was the first French scientist to be honored with a state funeral.

Bernard's teaching at the Collège (ten volumes of *Lessons* were published) attracted large audiences of physicians and physiologists from all over the world. Although not a brilliant lecturer, he communicated a vivid sense of laboratory work in the making: describing techniques such as the experimental section of nerves, showing how experiments are guided by preconceptions (hypotheses), and how preconceptions are dismissed (refuted) by experimental arguments. His teaching at the Sorbonne and at the Museum was more theoretic. It reflects the philosophical turn that Bernard took in his later years, especially when, from 1860, several episodes of illness occasioned prolonged stays for recovery in his native village. He then attempted to conceptualize both his research methodology and his project for the development of the science of general physiology that he had helped to shape.

The *Introduction to the Study of Experimental Medicine* (1865) was meant as a preparatory step toward a comprehensive treatise on the *Principles of Experimental Medicine*, which was never completed (a posthumous putative reconstruction of the *Principles*, based on draft chapters, was published in 1947). The *Lessons on the Properties of Living Tissues* (taught at the Sorbonne), published in 1866, stressed that general physiology aims at identifying traits which are identical in all living animals,

thus revealing the essence of vital phenomena. The *Lessons on the Vital Phenomena Common to Animals and Plants* (taught at the Museum), published in 1878, right after Bernard's death, are an exacting meditation on the unity of living processes and on the intricacy of organic construction and destruction in such processes.

Bernard's views on biostatistics are stated in the *Introduction* (see Part II, Chapter 2, section "De l'emploi du calcul dans l'étude des phénomènes des êtres vivants; des moyennes et de la statistique"), and more explicitly in the draft versions of the *Principles* (especially Chapter VI and VII): "It is easier to statistically count cases pro and con than to conduct proper experimental reasoning" (*Principles*, VII). Bernard wanted to take medicine from the state of "a conjectural science based on statistics" to the state of "an exact science based on experimental determinism". It is not sufficient, he said, to observe (for instance) that inoculation protects against smallpox in 95% of cases. You want to know the exact mechanism by which inoculation is effective, and why it fails to work in particular cases (exceptions do not occur just by chance – they must be explained). Statistical knowledge is descriptive; it has no explanatory power. So-called statistical "laws" are "true in general and false in particular". Should medicine be practiced as an "active science" rather than as an art of guessing, physicians would track physiologic processes in the laboratory up to the point at which there is no uncertainty left.

Bibliography

Books of Claude Bernard (a selection, including works on scientific methodology):

Bernard, C. (1853). *Recherches sur une Nouvelle Fonction du Foie, Considéré comme Organe Producteur de Matière Sucrée chez l'Homme et les Animaux*. Thèse présentée à la Faculté des sciences de Paris pour obtenir le grade de docteur ès sciences naturelles. Reprinted together with Communications by Claude Bernard to the French Académie des Sciences and to the Société de biologie, by M.D. Grmek, under the title *Notes, Mémoires et Leçons sur la Glycogenèse Animale et le Diabète*. Tchou, Paris, 1965.

Bernard, C. (1865). *Introduction à l'Étude de la Médecine Expérimentale*. Paris (and numerous later editions).

Bernard, C. (1878–1879). *Leçons sur les Phénomènes de la vie Communs aux Animaux et aux Végétaux*, by A. Dastre, ed. 2 vols. Paris. First volume reprinted with a Preface by G. Canguilhem, Vrin, Paris, 1966.

Bernard, C. (1937, posth.). *Pensées, Notes Détachées*, L. Delhoume, ed. Baillière, Paris.

Bernard, C. (1942, posth.). *Le Cahier Rouge*, L. Delhoume, ed. (fragmentary). Gallimard, Paris.

Bernard, C. (1947, posth.). *Principes de Médecine Expérimentale*, Avant propos by L. Binet, Introduction and notes by L. Delhoume. PUF, Paris.

Bernard, C. (1965, posth.). *Cahier de Notes 1850–1860*, M.D. Grmek, ed. (complete edition). Gallimard, Paris.

Studies on Bernard's conception of the scientific method:

Canguilhem, G. (1970). "L'idée de médecine expérimentale selon Claude Bernard", and "Théorie et technique de l'expérimentation chez Claude Bernard", in *Etudes d'Histoire et de Philosophie des Sciences*, 2nd Ed. Vrin, Paris.

Grmek, M.D. (1973). *Raisonnement Expérimental et Recherche Toxicologique chez Claude Bernard*. Droz, Geneva.

Grande, F. & Visscher, M.B., eds (1967). *Claude Bernard and Experimental Medicine*. Cambridge, Mass.

Halpern, B., ed. (1967). *Philosophie et Méthodologie Scientifique de Claude Bernard*. Fondation Singer-Polignac Paris.

Olmsted, J.M.D. & Harris Olmsted, E. (1952). *Claude Bernard and the Experimental Method in Medicine*. Schuman, New York.

For a comprehensive bibliography of Bernard's works and of works on Bernard, see: Grmek, M.D. (1967). *Catalogue des Manuscrits de Claude Bernard, avec la Bibliographie de ses Travaux Imprimés et des Études sur son Oeuvre*. Masson, Paris.

A. FAGOT-LARGEAULT

Bernoulli Family

Of the many distinguished members of the Bernoulli family of Basle, Switzerland, three in particular made important contributions to the development of probability, statistics, and epidemiology: James (1654–1705) and his nephews Nicholas (1687–1759) and Daniel (1700–1782), sons of two of his brothers. The family had originally come to Basle from Antwerp as Protestant refugees, fleeing the government of the Duke of Alba.

James's reputation rests on his posthumous *Ars conjectandi (The Art of Conjecture)* brought out in 1713 by Nicholas. The first of the four parts is a commentary, with text, on Christian Huygens' book *De ratiociniis in aleae ludo (On Reckoning in Games of Chance)*, which had appeared in 1657; part II, on permutations and combinations, is a generally inferior version of **Pascal's** similar *Traité du triangle arithmétique* (published in 1665), with which James was unfamiliar, but it does contain the **binomial distribution** which Pascal had only given for the case of equal chances (hence "Bernoulli trials"); and part III applies the methods discussed earlier to games of chance. Part IV is the seminal part of the work, in which James inaugurates statistical **estimation** theory by stating and proving the first **limit theorem** in probability, the **law of large numbers**.

Nicholas, besides seeing his uncle's book through the press, became interested in 1712 in the biostatistical problem of estimating the probability of a birth being male from some data, and then comparing the distribution of the observations with the binomial model. To do this, he needed an approximation to the binomial distribution, which led him to improve James's limit theorem, sharpening two of the inequalities involved. Nicholas was also the originator, in 1713, of the St Petersburg problem, famous in probability theory as a game with an infinite mathematical **expectation** – so what should be a fair stake?

To Daniel, we owe a novel derivation of the **normal distribution** from the binomial (also in connection with sex-ratio data), the first explicit commendation of the method of **maximum likelihood** for estimation (1778), and a suggested resolution of the St Petersburg paradox (1738). For this, he applied the Pascalian notion of **utility** to the value of money, arguing that the utility of large sums fell away so that the expected utility was not infinite, and that therefore a finite stake was in order. This established him as one of the founders of mathematical economics. He made many contributions to applied mathematics, including the "Bernoulli equation" for fluid flow (in *Hydrodynamica*, 1738).

In 1760, Daniel discussed a model for the mortality from smallpox in various age groups, deriving a differential equation relating the number of survivors and the number of those at age x who had not yet had smallpox. In the same year, he advocated inoculation against smallpox as a means of increasing average survival times (*see* **Life Expectancy**).

A.W.F. EDWARDS

Bernoulli Family

A.W.F. EDWARDS

Volume 1, pp. 385–385

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Bertillon Family

A dynasty of French natural scientists and physicians interested in vital and social statistics [1].

Louis-Adolphe Bertillon (1821–1883), a physician, was a son-in-law of Achille Guillard (1789–1876), the "founder" of **demography**. Bertillon studied the use of **means** and distributions (e.g. of heights), following **Quetelet**. He was familiar with the work of **William Farr**, and he corrected some of Guillard's methods.

Jacques Bertillon (1851–1922), son of Louis-Adolphe and also a physician, was a demographer and crusaded for a higher French birth rate. In 1893, he chaired a committee of the **International Statistical Institute** to prepare a new classification of causes of death, synthesizing methods used in different countries (*see* **International Classification of Diseases (ICD)**).

Alphonse Bertillon (1853–1914), another son of Louis-Adolphe, produced a system of identification of criminals and other individuals ("bertillonage"), by bodily measurements, photographs, and so on. Fingerprints had been used earlier, and were espoused by **Galton**.

Reference

[1] Lécuyer, B.-P. (1987). Probability in vital and social statistics: Quetelet, Farr, and the Bertillons, in *The Probabilistic Revolution*, Vol. 1, *Ideas in History*, L. Krüger, L.J. Daston & M. Heidelberger, eds. MIT Press, Cambridge, Mass., pp. 317–335.

Beta Distribution

The beta distribution is a probability distribution of a continuous random variable taking value in the interval [0, 1]. Its probability density function is

$$f(x; a, b) = B(a, b)^{-1} x^{a-1} (1 - x)^{b-1}, \quad 0 \leq x \leq 1,$$

where a and b are parameters satisfying $a > 0$ and $b > 0$, and where $B(a, b) = \int_0^1 t^{a-1}(1 - t)^{b-1}\, \mathrm{d}t$ is called the beta function. The rth moment of the distribution is

$$\mathrm{E}(X^r) = \prod_{i=0}^{r-1} \frac{a + i}{a + b + i}.$$

The mean and variance are

$$\mathrm{E}(X) = \frac{a}{a + b},$$

$$\mathrm{var}(X) = \frac{ab}{(a + b)^2(a + b + 1)}.$$

The beta density function has a variety of possible shapes, including mound-shaped when a and b exceed 1, **U-shaped** when they are both less than 1, strictly increasing when $a > 1$ and $b = 1$, strictly decreasing when $a = 1$ and $b > 1$, and the **uniform distribution** over [0, 1] when $a = b = 1$. It is symmetric about 1/2 when $a = b$, and otherwise the direction of **skewness** is determined by $a - b$.

If X has a beta distribution with parameters a and b, then $(b/a)X/(1 - X)$ has an ***F* distribution** with $2a$ and $2b$ degrees of freedom. If Y_1 and Y_2 are independent **gamma** random variables with unit scale parameter and shape parameters a and b, then $Y_1/(Y_1 + Y_2)$ has a beta distribution with parameters a and b.

The beta distribution is often used to model proportions. For instance, in **Bayesian** inference it is the conjugate **prior distribution** when the observations have a **binomial distribution**, given the proportion. In that case, the marginal distribution of the observations is the **beta-binomial**. The Dirichlet distribution is a multivariate form of the beta.

For further details about the beta distribution, see [2]. An alternative two-parameter family over the unit interval, called the *logistic-normal* [1], assumes that the logit of the variable has a normal distribution.

References

[1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

[2] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Vol. 2, 2nd Ed. Wiley, New York.

Alan AGRESTI

Beta Distribution

Alan AGRESTI

Volume 1, pp. 389–390

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Beta-binomial Distribution

It is often the case that an individual provides repeated binary outcomes. For example, in ophthalmology, a subject almost always has information available on the right and left eyes. In obstetrics, one sometimes has outcomes from multiple pregnancies of the same woman. The primary outcome variable is often binary (e.g. reduced visual acuity = visual acuity 20/50 or worse in an individual eye). Such data are referred to as *clustered*. Clusters may be defined not only by replicates for a single individual, but also by outcomes for different related cluster members (e.g. members of a family). The use of standard methods for analyzing categorical data (e.g. the chi-square test for 2×2 tables) which are usually based on either the **binomial distribution** or the **hypergeometric distribution** using the cluster member as the unit of analysis is invalid because the assumption of independence of binary outcomes for different cluster members is often not correct.

An attractive alternative model in this setting is the beta-binomial model.

Definition

Under the beta-binomial model, we assume that there are t_i cluster members in the ith cluster, $i = 1, \ldots, k$. For the ith cluster (or unit) we assume that any cluster member (or subunit) has probability p_i of being affected, where p_i follows a **beta distribution** with parameters a and b. A beta distribution with parameters a and b (both > 0) is a probability distribution over the interval (0,1) with density $\{\Gamma(a+b)/[\Gamma(a)\Gamma(b)]\}p^{a-1}(1-p)^{b-1}, 0 < p < 1$. It has mean $= a/(a+b)$ and variance $= ab/[(a+b)^2(a+b+1)]$. Conditional on p_i, the outcomes for different cluster members are independent. The resulting marginal distribution of $Y_i =$ number of affected cluster members in the ith cluster is referred to as the beta-binomial distribution, which is given by

$$\Pr(Y_i = k) = \binom{t_i}{k} \frac{a_k b_{t_i - k}}{(a+b)_{t_i}}, \quad k = 0, 1, \ldots, t_i, \tag{1}$$

where $a_k =$ the rising factorial $\prod_{j=0}^{k-1}(a+j)$ and b_{t_i-k} and $(a+b)_{t_i}$ are defined similarly.

The expected value of Y_i is $t_i a/(a+b)$ and the variance is $t_i ab/(a+b)^2 + t_i(t_i-1)ab/[(a+b)^2(a+b+1)]$. Standard maximum likelihood methods can be used for parameter estimation [2, 3, 10]. Notice that the variance of the beta-binomial distribution is always greater than the corresponding variance of a binomial distribution with the same expected value, which is given by $t_i ab/(a+b)^2$. This property is referred to as **overdispersion** or extrabinomial variation and provides a rationale for why the beta-binomial distribution can be used more generally to model overdispersed binary data.

Measures of Dependence Among Subunits within a Cluster

The beta-binomial model can be used to quantify dependence between outcomes for two subunits within the same cluster. Two measures that are useful for this purpose are the intraclass **correlation** and the pairwise **odds ratio**. These are given by

$$\begin{aligned} &\text{intraclass correlation} \\ &\quad = \operatorname{corr}(y_{ij_1}, y_{ij_2}) \\ &\quad = 1/(a+b+1), \end{aligned} \tag{2}$$

$$\begin{aligned} &\text{pairwise odds ratio} \\ &\quad = \frac{\Pr(y_{ij_1}=1, y_{ij_2}=1)\Pr(y_{ij_1}=0, y_{ij_2}=0)}{\Pr(y_{ij_1}=1, y_{ij_2}=0)\Pr(y_{ij_1}=0, y_{ij_2}=1)} \\ &\quad = \frac{(a+1)(b+1)}{ab}, \end{aligned} \tag{3}$$

where y_{ij} is the outcome for the jth subunit within the ith cluster. In words, the pairwise odds ratio is a measure of the odds in favor of disease for the j_1th subunit if the j_2th subunit is affected divided by the odds in favor of disease for the j_1th subunit if the j_2th subunit is not affected. For either measure, as $a+b$ decreases, the aggregation between subunits in a cluster increases. If $a+b$ approaches ∞, then the correlation between subunits approaches 0 and the beta-binomial distribution converges to the binomial distribution. If $a+b=0$, then there is perfect correlation between subunits within a cluster. Note that negative correlation is not allowed under the beta-binomial distribution.

The Treatment of Covariates

An important issue is how to use the beta-binomial distribution in the presence of **covariates**. If there is only a single binary covariate (e.g. a treatment indicator variable in a clinical trial), then a simple approach is to fit separate beta-binomial models for a treated and control group and compare the parameters for the two groups. In particular, disease prevalence can be estimated by $a/(a+b)$ and compared between the two groups [11]. More generally, several authors have considered generalizations of the beta-binomial model for correlated binary data in the presence of covariates. Prentice [4] has proposed the following model for the joint distribution of t correlated binary variables in the presence of unit-specific covariates x:

$$\Pr(y_{i+}|x) = \frac{\prod_{j=0}^{y_{i+}-1}[p(x)+\lambda(x)j]\prod_{j=0}^{t-y_{i+}-1}[1-p(x)+\lambda(x)j]}{\prod_{j=0}^{t-1}[1+\lambda(x)j]}, \quad (4)$$

where $p(x)$ is a parametric model relating the marginal probability of disease to x, $\lambda(x) = \delta(x)/[1-\delta(x)]$, and $\delta(x)$ is the intraclass correlation for disease probabilities between cluster members as defined above. This model has the advantage that (i) the marginal probability is explicitly specified as a function of covariate values, and (ii) it reduces to a beta-binomial model if no covariates are present. The disadvantages are (i) only unit-specific covariates are possible, and (ii) the dependence among cluster members is parameterized by a correlation rather than an odds ratio; in general, constraints need to be placed on the correlations so that the probabilities in (4) are ≤ 1.

Rosner [6] proposed a polytomous logistic regression model (*see* **Polytomous Data**) which allows for both unit- and subunit-specific covariates for clustered binary data of the form

$$\Pr(y_i|x_i) = \frac{a_{s_i}b_{t_i-s_i}\exp\left[\sum_{j=1}^{t_i} y_{ij}(\beta x_i^{(0)}+\gamma x_i^{(j)})\right]}{\sum_{z_i} a_{z_{i+}}b_{t_i-z_{i+}}\exp\left[\sum_{j=1}^{t_i} z_{ij}(\beta x_i^{(0)}+\gamma x_i^{(j)})\right]}, \quad (5)$$

where $x_i^{(0)}$ is an $(N_p \times 1)$ vector of unit-specific covariates, $x_i^{(j)}$ is an $(N_e \times 1)$ vector of subunit-specific covariates, $s_i = y_{i+}$, and the summation in the denominator is over all possible permutations $z_i = (z_{i1}, \ldots, z_{it_i})$ of zeros and ones. If no covariates are present, then this model reduces to a beta-binomial distribution with parameters a and b. The measure of dependence in this model is the pairwise odds ratio as defined in (3), which can be shown to be independent of covariate values x. The most natural interpretation of the regression parameters is in conditional form as follows:

$$\ln\left[\frac{p_{ij}}{(1-p_{ij})}\right] = \ln\left[\frac{(a+s_{-j})}{(b+t_i-1-s_{-j})}\right] + \beta x^{(0)} + \gamma x^{(j)}, \quad (6)$$

where s_{-j} is the number of successes among the $t_i - 1$ subunits excluding subunit j. In this context, if $x_p^{(0)}$ is the pth unit-specific variable, then $\beta_p = \ln$ (odds in favor of disease for a 1 unit increase in $x_p^{(0)}$ holding all other variables constant including the disease status of the other cluster members). Thus, β (and likewise γ) has a conditional rather than a marginal interpretation. Qu et al. [5] and Connolly & Liang [1] have considered an extension of (5) of the form

$$\Pr(y_i|x_i) = c(\theta,\beta,\gamma)\exp\left[\sum_{k=0}^{s_i-1} f_t(k,\theta) + \sum_{j=1}^{t_i} y_{ij}(\beta x_i^{(0)}+\gamma x_i^{(j)})\right], \quad (7)$$

where θ is a vector of correlation parameters and $f_t(k,\gamma)$ is an arbitrary function representing the log odds in favor of being affected at a specific visit conditional on there being k successes among the remaining $t_i - 1$ visits and all covariate values being 0. It can be shown that (5) is a special case of (7).

More Complex Patterns of Correlation

An implicit assumption of the beta-binomial model is that subunits are exchangeable within a cluster. A natural generalization is to allow for data structures with multiple levels of nesting. For example, in the ophthalmologic setting one may have data from

several members of the same family where each person provides data on two eyes. In this setting, it would be expected that the correlation between two eyes of the same person is different from the correlation between two eyes from different family members, although both correlations may be greater than zero. To accommodate this data structure [7] let p be the probability that an eye is affected for a person from an average family across all families in the population, let p_i be the probability that an eye is affected for an average person from the ith family, and p_{ij} be the probability that an eye is affected for the jth person in the ith family. We model

$$f_1(p_i|p) \sim \text{beta}\,[\lambda_0 p, \lambda_0(1-p)] \equiv \text{beta}(a, b),$$
$$f_2(p_{ij}|p_i) \sim \text{beta}\,[\lambda_I p_i, \lambda_I(1-p_i)]. \tag{8}$$

The resulting marginal distribution of the p_{ij} is referred to as a *compound beta-binomial distribution* and is a function of p, λ_0 and λ_1. Under this model, the odds ratio between eyes from two different family members is

$$OR_1 = \frac{(a+1)(b+1)}{ab},$$

while the odds ratio between two different eyes of the same person is

$$OR_2 = \frac{(\lambda_1 a + \lambda_1 + \lambda_0 + 1)(\lambda_1 b + \lambda_1 + \lambda_0 + 1)}{(\lambda_1 a)(\lambda_1 b)}.$$

Newton–Raphson methods can be used to obtain **maximum likelihood** estimates of the parameters of this model (*see* **Optimization and Nonlinear Equations**).

Another variant of the beta-binomial model is obtained when one can subdivide a cluster into multiple subclasses, for example when one has both parents and children in the same cluster. In the case of c classes, one wants to generalize the beta-binomial model to allow for $c(c-1)/2$ interclass **correlations** between outcomes for members of different classes, and c intraclass correlation parameters between outcomes for members of the same class. To accomplish this goal, let

$$\lambda_i = \sum_{k=1}^{c} w_{ik}\theta_k, \quad i = 1, \ldots, c, \tag{9}$$

where $\theta_1, \ldots, \theta_c$ are independent beta (a, b) random variables and $\sum_{k=1}^{c} w_{ik} = 1, i = 1, \ldots, c$. It follows that the intraclass correlation between outcomes for subunits in the ith class of the same cluster is given by

$$\rho(y_{ij_1}, y_{ij_2}) = \sum_{k=1}^{c} \frac{w_{ik}^2}{(a+b+1)}, \quad i = 1, \ldots, c, \tag{10}$$

while the interclass correlation between outcomes for pairs of subunits in the i_1th and i_2th class of the same cluster is given by

$$\rho(y_{i_1 j_1}, y_{i_2 j_2}) = \sum_{k=1}^{c} \frac{w_{i_1 k} w_{i_2 k}}{(a+b+1)},$$
$$i_1, i_2 = 1, \ldots, c, i_1 \neq i_2. \tag{11}$$

Note that if $w_{ii} = 1, i = 1, \ldots, c$, then we have c independent subclasses with all intraclass correlations $= 1/(a+b+1)$ and all interclass correlations $= 0$. If $w_{i1} = 1, i = 1, \ldots, c$, then the model in (9) reduces to the ordinary beta-binomial model in (1). Based on (9) one can obtain an explicit function for the joint likelihood of $y = (y_l, \ldots, y_c)$, where y_i is a $1 \times t_i$ vector of outcomes in the ith class. The resulting model is referred to as a beta-binomial mixture model [8, 9].

References

[1] Connolly, M.A. & Liang, K.Y. (1988). Conditional logistic regression models for correlated binary data, *Biometrika* **75**, 501–506.

[2] Crowder, M.J. (1978). Beta-binomial ANOVA for proportions, *Applied Statistics* **27**, 34–37.

[3] Griffiths, D.A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease, *Biometrics* **29**, 637–648.

[4] Prentice, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, *Journal of the American Statistical Association* **394**, 321–327.

[5] Qu, Y.S., Williams, G.W., Beck, G.J. & Goormastic, M. (1987). A generalized model of logistic regression for correlated data, *Communications in Statistics* **16**, 3447–3476.

[6] Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired data situations, *Biometrics* **40**, 1025–1035.

[7] Rosner, B. (1989). Multivariate methods for clustered binary data with more than one level of nesting, *Journal of the American Statistical Association* **84**, 373–380.

[8] Rosner, B. (1992). Multivariate methods for binary longitudinal data with heterogeneous correlation over time, *Statistics in Medicine* **11**, 1915–1928.

[9] Rosner, B. (1992). Multivariate methods for clustered binary data with multiple subclasses, with application to binary longitudinal data, *Biometrics* **48**, 721–731.

[10] Smith, D.M. (1983). Maximum likelihood estimation of the parameters of the beta-binomial distribution (algorithm AS189), *Applied Statistics* **32**, 196–204.

[11] Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics* **31**, 949–952.

(*See also* **Logistic Regression**; **Rasch Models**)

BERNARD ROSNER

Bias From Diagnostic Suspicion in Case–Control Studies

Diagnostic suspicion **bias** occurs when there is **systematic error** in case ascertainment. For instance, a knowledge about the subject's prior exposure to a putative cause (family history, being exposed to an epidemic, taking certain drugs, and certain occupational exposure) may increase diagnostic search for the disease, and as a result, exposed subjects are more likely to have the disease diagnosed than the nonexposed [10]. It can occur in both a **case control** and a **cohort study**. In case–control studies, diagnostic suspicion bias is a **selection bias**, whereas in cohort studies, it is considered as a measurement or information bias since it occurs in subjects already included in the study [9].

Diagnostic suspicion bias can result in overestimation of an exposure's effect on the risk of disease [13]. For example, in patients with a suspected risk factor, physicians may perform the diagnosis more carefully. So, they are more likely to make the correct diagnosis in patients with the disease of interest and also more likely to make a **false positive** diagnosis. Conversely, in patients without the suspected risk factors, presence of the disease is less aggressively sought, and consequently, physicians are more likely to make a false negative diagnosis. A false positive diagnosis, however, is unlikely. Figure 1 illustrates the described scenario. Suppose that among the exposed, 90% of subjects who truly have the disease and 20% of subjects who truly do not have the disease are diagnosed to have the disease, while among the nonexposed, 70% who truly have the disease and 0% who truly do not have the disease are diagnosed positive. The unbiased **odds ratio** is AD/BC, while the putative odds ratio becomes $[(0.9A + 0.2B) \times (D + 0.3C)]/[(0.8B + 0.1A) \times 0.7C]$ when diagnostic suspicion bias occurs. Clearly, this odds ratio would in general overestimate the true odds ratio.

Diagnostic suspicion bias is more likely to occur when objective criteria for reliable diagnosis are difficult to establish [12]. For some diseases, such as rheumatoid arthritis, a whole host of symptoms and signs may be present. Different clinicians may interpret these symptoms, signs, and various laboratory test results differently, which may lead to very different diagnoses. Sometimes, diagnoses are classified as "definite", "probable", and "possible". In this situation, with the same borderline symptoms and signs, subjects with the suspected risk factor may be diagnosed "possibly" to have the disease, whereas those without the risk factor may be diagnosed negative.

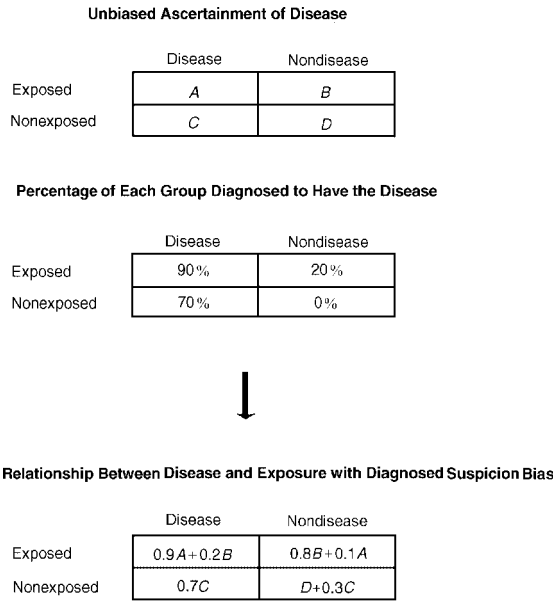

Figure 1 Illustration of diagnostic suspicion bias in case–control studies

Diagnostic suspicion bias is closely related to surveillance bias and **detection bias**. Surveillance bias occurs when individuals under frequent or close surveillance are more likely to have disease diagnosed [11]. For example, postmenopausal women taking estrogen may be more likely to have breast cancer diagnosed because of frequent physician visits. Detection (unmasking) bias occurs when an exposure, rather than causing disease, causes symptoms that precipitate a search for the disease [10]. An early report of postmenopausal estrogens and endometrial cancer was criticized on this ground [5, 6]. It was suggested that subclinical cancers were being diagnosed more frequently in exposed women because estrogen use could cause symptomless patients to bleed, which, in turn, led to more thorough diagnoses. As with diagnostic suspicion bias, surveillance and detection biases can result in a spurious increase in the odds ratios.

Examples of Studies on Checking Diagnostic Suspicion Bias

In a study of oral contraceptives and deep-vein thrombosis and pulmonary thrombosis, Vessey & Doll [14] investigated the possibility that a history of oral contraceptives use might influence doctors' diagnoses. They asked an independent investigator who had no knowledge about the patients' exposure to classify the diagnoses as "possible", "probable", or "certain". It was hypothesized that if the diagnoses were influenced by knowledge of a patient's contraceptive history, the association would be strongest in patients for whom the diagnosis was least certain, because patients in this group were most likely to be diagnosed only because of their oral contraceptives use history. Results indicated, however, that the proportions of subjects using oral contraceptives increased with increasing certainty of diagnosis, suggesting that the clinical diagnoses were not biased by knowledge about the exposure.

Fox & White [3] examined whether diagnostic suspicion bias affected the observed increase in mortality rate of bladder cancer among workers in the rubber industry. The bias might arise from the publicity that followed the initial discovery of excess bladder cancer among these workers. Awareness of the association between work in the rubber industry and bladder cancer among doctors might increase the chance of recording the bladder cancer as the underlying cause on the death certificates of workers in the rubber industry than on the death certificates of other people. The researchers compared death certificates for bladder cancer cases from the rubber industry with those from other groups identified from the National Cancer Registry. They found that proportions of death certificates with mention of bladder cancer as the underlying cause on the certificates were similar between workers in the rubber industry and patients from other occupations, indicating that doctors' awareness did not explain the rise of bladder cancer death among these workers.

Foreman et al. [1] found that intrauterine device (IUD) usage at conception significantly increased septic second-trimester fetal loss. They considered the possibility that, with borderline evidence of infection, patients with an IUD in place might be more frequently diagnosed as septic than those without an IUD in place. To check this bias, they restricted the analyses to only blatant cases (cases with temperature of 39.4°C or higher). The results remained unchanged, indicating diagnosis suspicion bias was not an explanation for the observed association.

Several case–control studies reported a positive association between aspirin use and Reye's syndrome [4, 7, 15]. However, diagnostic suspicion bias may partially explain this association because the diagnosis procedure could be affected by the extensive previous publicity about a relationship between aspirin and Reye's syndrome. Forsyth et al. [2] carried out a further case–control study in which "diagnostic-suspicion" patients (Reye's syndrome was initially suspected but definitely ruled out) were used as an alternative control group. They found that aspirin use in this control group was very low and equal to the rate in the control group identified from communities, suggesting that the diagnosis of Reye's syndrome was not affected by the knowledge of aspirin use and publicity about the association.

Prevention and Minimization

To prevent diagnostic suspicion bias, one must make sure that the disease of interest is sought with equal vigor in exposed and nonexposed subjects [13]. In other words, exposed and nonexposed people should have the same chance to be detected as cases. This may be implausible in circumstances where prevalence cases are selected because diagnoses are made prior to the study. However, multiple data sources can be used to verify diagnoses and minimize **misclassification** [8]. For example, one may review hospital records, death certificates, and pathology reports, in addition to the data from a disease registry, to identify cases.

When cases are gathered prospectively (incidence cases), one important strategy to minimize bias is to ensure that all activities associated with case ascertainment follow the same, standard protocol so that diagnoses are made in the same way in the exposed and nonexposed [8]. In randomized **clinical trials**, outcome assessment is conducted in a blinded manner with regard to treatment conditions. In case-control studies, a blinded assessment requires that neither clinicians nor patients be aware of the study hypothesis. In some cases, independent investigators may be needed to evaluate all available evidence

without having knowledge about the exposure [14] (*see* **Blinding or Masking**).

One way to control diagnostic suspicion bias in data analysis is to stratify cases (*see* **Stratification**) on the basis of certainty of diagnosis or severity of disease [11]. The investigator can also simply restrict the analyses to blatant cases, but this may reduce the **power** of the study. As with other biases in case-control studies, diagnostic suspicion bias should be taken into consideration when a study is conceived.

References

[1] Foreman, H., Stadel, B.V. & Schlesselman, S. (1981). Intrauterine device usage and fetal loss, *Obstetrics and Gynecology* **58**, 669–677.

[2] Forsyth, B.W., Horwitz, R.I., Acampora, D., Shapiro, E.D., Viscoli, C.M., Feinstein, A.R., Henner, R., Holabird, N.B., Jones, B.A., Karabelas, A.D.E., Kramer, M.S., Miclette, M. & Wells, J. (1989). New epidemiologic evidence confirming that bias does not explain the aspirin/Reye's syndrome association, *Journal of the American Medical Association* **262**, 2517–2524.

[3] Fox, A.J. & White, G.C. (1976). Bladder cancer in rubber workers, *Lancet* **1**, 1009–1011.

[4] Halpin, T.J., Holtzhauer, F.J., Campbell, R.J., Hall, L.J., Correa-Villasenor, A., Lanese, R., Rice, J. & Hurwitz, E.S. (1982). Reye's syndrome and medication use, *Journal of the American Medical Association* **248**, 687–691.

[5] Horwitz, R.I. & Feinstein, A.R. (1978). Alternative analytic methods for case-control studies of estrogens and endometrial cancer, *New England Journal of Medicine* **299**, 1089–1094.

[6] Horwitz, R.I. & Feinstein, A.R. (1979). Analysis of clinical susceptibility bias in case-control studies: Analysis as illustrated by the menopausal syndrome and the risk of endometrial cancer, *Archives of Internal Medicine* **139**, 1111–1113.

[7] Hurwitz, E.S., Barrett, M.J., Bregman, D., Gunn, W.J., Pinsky, P., Schonberger, L.B., Drager, J.S., Kaslow, R.A. & Burlington, D.B. (1987). Public Health Service study on Reye's syndrome and medications: Report of the main study, *Journal of the American Medical Association* **257**, 1905–1911.

[8] Kopec, J.A. & Esdaile, J.M. (1990). Bias in case-control studies: A review, *Journal of Epidemiology and Community Health* **44**, 179–186.

[9] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown & Company, New York.

[10] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[11] Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, and Analysis*. Oxford University Press, New York.

[12] Smith, M.W. (1981). The case control or retrospective study in retrospect, *Journal of Clinical Pharmacology*, **21**, 269–274.

[13] Sutton-Tyrrell, K. (1991). Assessing bias in case-control studies: Proper selection of cases and controls, *Stroke* **22**, 938–943.

[14] Vessey, M.P. & Doll, R. (1968). Investigation of relation between use of oral contraceptives and thromboembolic disease, *British Medical Journal* **2**, 199–205.

[15] Waldman, R.J., Hall, W.N., McGee, H. & Van Amburg, G. (1982). Aspirin as a risk factor in Reye's syndrome, *Journal of the American Medical Association* **247**, 3089–3094.

(*See also* **Bias in Case–Control Studies**; **Bias in Observational Studies**; **Bias, Overview**)

F.B. HU

Bias from Exposure Effects on Controls

In population-based case–control studies, controls are selected at random from the source (or "base") population (*see* **Case–Control Study, Population-based**). In hospital-based case–control studies, cases in the hospital with a disease of interest are compared with controls in the hospital who have other diseases (*see* **Case–Control Study, Hospital-based**). In hospital-based case–control studies, the exposure **odds ratio** comparing cases with controls may be a **biased** estimate of the **relative risk** of disease in the underlying source population if the risks of the control diseases are themselves associated with the exposure under study. For example, in a pioneering hospital-based case–control study of the risk of lung cancer from smoking, Doll & Hill [2] included subjects with bronchitis among the controls. Because the risk of bronchitis is now known to be increased by smoking (*see* **Smoking and Health**), we can infer that estimates of the relative risk of lung cancer from smoking were biased downward by the inclusion of such controls. A quantitative treatment of such bias is given by Breslow & Day [1, pp. 153–154].

References

[1] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2: *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[2] Doll, R. & Hill, A.B. (1952). A study of the aetiology of carcinoma of the lung, *British Medical Journal* **2**, 1271–1286.

(*See also* **Bias in Case–Control Studies**)

MITCHELL H. GAIL

Bias from Exposure Suspicion in Case–Control Studies

Exposure suspicion **bias** is the mirror image of *diagnostic suspicion* bias (*see* **Bias From Diagnostic Suspicion in Case–Control Studies**). It occurs when there is **systematic** error in the ascertainment of exposure. For example, a knowledge of the subject's disease status may increase the intensity of a search for exposure to the putative cause, and as a result, exposure information is gathered more thoroughly in cases than in controls [9]. Exposure suspicion bias is specific to a **case-control studies**, in which exposure is usually ascertained after the outcome event has occurred.

As with diagnostic suspicion bias, exposure suspicion bias can spuriously increase the effect of the exposure on the risk of disease. Consider a study of the presence of a carotid bruit and the occurrence of transient ischemic attack (TIA) [13]. In patients with symptoms of cerebral ischemia, a physician listens very carefully and is able to detect a bruit most of the time if it is truly present. Meanwhile, a **false positive** diagnosis of a bruit is likely to occur owing to awareness of the suspected association. On the contrary, in patients without cerebral ischemia, the presence of a bruit is less aggressively sought, so a **false negative** diagnosis of a bruit is likely to occur. A false positive diagnosis, however, is unlikely. The described scenario is illustrated in Figure 1. Suppose that among the diseased, 90% of subjects who are truly exposed and 20% who are truly nonexposed are detected to have the exposure; while, among the nondiseased, 70% who are truly exposed and 0% who are truly nonexposed are detected to have the exposure. The unbiased **odds ratio** is AD/BC. The putative odds ratio becomes $[(0.9A + 0.2C) \times (D + 0.3B)/(0.7B) \times (0.8C + 0.1A)]$ when exposure suspicion bias occurs. This putative odds ratio would in general overestimate the true odds ratio.

Exposure suspicion bias can arise from several sources. One source relates to clinicians who examine the patients, record the exposure information, and perform histologic assessments and laboratory tests. The second source is the interviewer who inquires about whether or not the patients are exposed to the suspected risk factor. For example, an interviewer

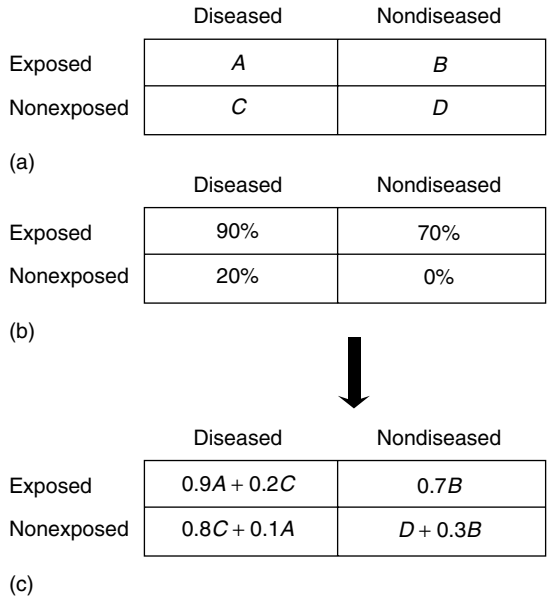

Figure 1 Exposure suspicion bias in case–control studies. (a) Unbiased ascertainment of exposure; (b) the percentage of each group detected to have the exposure; (c) the relationship between exposure and disease with exposure suspicion bias

with knowledge of the study hypothesis may tend to probe the diseased subjects more intensely for histories of exposure and may encourage certain responses among either cases or controls through language, tone, or "body language". A third source is the interviewed subjects or proxies who report or recall exposure information. For example, individuals with specific diseases may be more likely to recall exposures if they suspect that the exposure is related to their disease. Mothers of malformed infants may recall certain exposure more thoroughly than mothers of healthy infants if they believe the exposure to be the cause of the adverse outcome [8]. Differential **recall bias** often tends to overestimate the **risk**, but the opposite may also occur [8].

Examples of Studies on Checking Exposure Suspicion Bias

In Doll & Hill's case–control study of smoking and lung cancer [1], potential bias on reporting of smoking history might arise due to interviewers' knowledge of the diagnosis (*see* **Smoking and Health**).

The researchers checked this possibility by comparing reported smoking habits of patients of confirmed lung cancer with those whose diagnoses were later not confirmed. They found that smoking habits of patients with erroneous diagnoses resembled those of the controls and differed significantly from those of confirmed cases. This indicates that reporting of smoking history was not affected by interviewer's knowledge about the disease status.

Forsyth et al. [4] investigated the possibility of biased recall of medications of the child by the parent or guardian in a case-control study of the association between aspirin use and Reye's syndrome. The bias might arise from physicians' consideration about Reye's syndrome, their repeated questioning about aspirin use, and intensive publicity about the association. It was discounted since the reported aspirin use among those who were suspected initially of having Reye's syndrome, but later were confirmed not to be cases, was found to be as low as in the control group.

MacKenzie & Lippman [7] examined the effect of knowledge about pregnancy outcomes on the reporting of exposure histories by comparing reports from early pregnancy and postdelivery. It was argued that if exposure suspicion bias did occur, cases would add more new information about the exposure and delete less of the previously reported exposure. Results showed that changes in reporting were similar among cases (mothers of died, stillborn, and malformed infants), controls, and mothers whose infants had less serious problems.

Werler et al. [16] compared interview data with exposure information documented during pregnancy in obstetric records and found that serious recall biases existed for many exposure variables. These biases, however, were **nondifferential** in regard to pregnancy outcomes: that is, mothers of severely malformed infants did not recall better or worse than mothers of less severely malformed infants.

Prevention and Minimization

To some extent, the strategies for preventing and minimizing exposure suspicion bias mirror those for preventing and minimizing diagnostic suspicion bias. The key is that the exposure of interest is sought with equal vigor in diseased and nondiseased subjects [13]. In other words, cases and controls should have the same chance to be classified as "exposed" or "nonexposed". This may reduce the degree of differential exposure **misclassification** (*see* **Differential Error**). But nondifferential misclassification caused by the general inability of individuals to report accurately about the histories of exposure may still exist, which often biases the results toward a weak or null association (*see* **Bias Toward the Null**).

In principle, exposure suspicion bias can be diminished by **blinding** the clinicians who perform the physical examinations, histologic assessment, and laboratory tests, the interviewers, and even the subjects themselves to ensure comparability of measurement of the exposure [6]. Blinding implies keeping clinicians, interviewers, and patients ignorant of both the study hypothesis and the classification of the subject as a case or a control. This can be easily achieved for clinicians who perform histologic assessment and laboratory tests. In most actual field operations, blinding interviewers and patients, however, may not be feasible. Nevertheless, **interviewer bias** and differential recall can be reduced by using standardized, uniform data collection procedures and by training the interviewers for unbiased probing [3, 10, 18]. Objective or independent sources of exposure history can also be used.

To address the potential influence of knowledge of disease status on the report of exposure, several authors have advocated the use of "affected", "restricted", or "pseudo" controls [7, 12, 16]. It is argued that if equally sick controls are chosen, interest and/or preoccupation with one's medical state would be similar for the two groups and, as a result, exposure history would be examined or probed in the same way in the two groups. However, use of affected controls may introduce **selection bias** [2, 14] or a subtle **confounder** [5]. To circumvent these problems, one can select controls with a variety of admission diagnoses.

Some authors have suggested that respondents who are aware of the etiologic hypothesis should be excluded from the analyses to control exposure suspicion bias [17]. However, others cautioned that exclusion of "knowledgeable" subjects might introduce selection bias and should not be advocated as general practice [15]. Nevertheless, side-by-side comparison of analyses with and without consideration of subjects' knowledge about the study hypothesis would provide an excellent opportunity to examine whether the knowledge does in fact influence ascertainment of exposure.

Use of biologic markers has received increasing interest in epidemiologic studies [11]. Bio-markers have been developed to assess environmental exposures, nutritional factors, and genetic susceptibility. Although survey methods will continue to play a dominant role in most case-control studies, integration of laboratory methods would in general increase the quality of exposure ascertainment [11].

References

[1] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung: a preliminary report, *British Medical Journal* **2**, 739–748.

[2] Drews, C., Greenland, S. & Flanders, W.D. (1993). The use of restricted controls to prevent recall bias in case-control studies of reproductive outcomes, *Annals of Epidemiology* **3**, 86–92.

[3] Feinstein, A.L. (1979). Methodologic problems and standards in case-control research, *Journal of Chronic Diseases* **32**, 35–41.

[4] Forsyth, B.W., Horwitz, R.I., Acampora, D., Shapiro, E.D., Viscoli, C.M., Feinstein, A.R., Henner, R., Holabird, N.B., Jones, B.A., Karabelas, A.D.E., Kramer, M.S., Miclette, M. & Wells, J. (1989). New epidemiologic evidence confirming that bias does not explain the aspirin/Reye's syndrome association, *Journal of the American Medical Association* **262**, 2517–2524.

[5] Khoury, M.J., James, L.M. & Erickson, J.D. (1994). On the use of affected controls to address recall bias in case-control studies of birth defects, *Teratology* **49**, 273–281.

[6] Kopec, J.A. & Esdaile, J.M. (1990). Bias in case-control studies: a review, *Journal of Epidemiology and Community Health* **44**, 179–186.

[7] MacKenzie, S.G. & Lippman, A. (1989). An investigation of report bias in a case-control study of pregnancy outcome, *American Journal of Epidemiology* **129**, 65–75.

[8] Martinez-Frias, M.L. (1993). Interviewer bias and maternal bias, *Teratology* **47**, 531–532.

[9] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[10] Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, and Analysis*. Oxford University Press, New York.

[11] Schulte, P.A. & Perera, F.P., eds (1993). *Molecular Epidemiology: Principles and Practices*. Oxford University Press, New York.

[12] Smith, M.W. (1981). The case control or retrospective study in retrospect, *Journal of Clinical Pharmacology* **21**, 269–274.

[13] Sutton-Tyrrell, K. (1991). Assessing bias in case-control studies: proper selection of cases and controls, *Stroke* **22**, 938–943.

[14] Swan, S.H., Shaw, G.M. & Schulman, J. (1992). Reporting and selection bias in case-control studies of congenital malformations, *Epidemiology* **3**, 356–363.

[15] Weiss, N.S. (1994). Should we consider a subject's knowledge of the etiologic hypothesis in the analysis of case-control studies?, *American Journal of Epidemiology* **139**, 247–249.

[16] Werler, M.M., Pober, B.R., Nelson, K. & Holmes, L.B. (1989). Reporting accuracy among mothers of malformed and nonmalformed infants, *American Journal of Epidemiology* **129**, 415–421.

[17] Werler, M.M., Shapiro, S. & Mitchell, A.A. (1993). Periconceptional folic acid exposure and risk of occurrent neural tube defects. *Journal of the American Medical Association* **269**, 1257–1261.

[18] Wynder, E.L. (1994). Investigator bias and interviewer bias: the problem of reporting systematic error in epidemiology, *Journal of Clinical Epidemiology* **47**, 825–827.

(*See also* **Bias in Case–Control Studies**; **Bias in Observational Studies**; **Bias, Overview**)

F.B. HU

Bias from Historical Controls

An historical control trial (HCT) has been defined [8] as a trial that compares the experience of a prospectively treated group with that of either a previously published series or with previously treated patients at the same institution (the historical controls). HCTs are often compared with both randomized **clinical trials** (RCT) and with **observational studies**. In an RCT, treatment assignment is randomly assigned to all subjects (*see* **Randomization**), and **covariate** information and outcomes are collected prospectively. In an **observational study**, treatment assignment is not determined by the investigator. Because RCTs have become the standard by which medical therapies are judged, the focus here will be on the comparison between an HCT and RCT in terms of feasibility, cost, and **bias**.

Advantages of Historical Control Trials

There are both **sample size** considerations and ethical arguments (*see* **Ethics of Randomized Trials**) in support of HCTs. These have been described in [1, 4], and [5].

Sample size considerations center on the reduced number of new patients needed to conduct an HCT with the same **power** as an RCT. As an example, assume that the response, x, in the control group, and, y, in the treatment group, have the same **variance** σ^2 and that there are n_c subjects in the control group and n_T subjects in the treatment group. Then $\text{var}(\overline{x} - \overline{y}) = \sigma^2(1/n_c + 1/n_T)$. A typical clinical trial sets $n_c = n_T = n$. In an HCT in which data for the n_c patients in the control group are already available, only n_T patients – one-half that in the clinical trial – would be needed. If the response in the control group is assumed known with certainty or if n_c is much larger than n_T, only $n_T/2$ – one-quarter that in the clinical trial – would be needed. For rare diseases or for expensive trials the advantage is clear. In addition, a smaller sample size reduces the time required for recruitment and so shortens the trial duration.

Clinical trials are most often undertaken to prove a new treatment superior to standardized ones based on previously accumulated data. As a consequence, a randomized trial allocates to some patients a treatment considered by the investigators to be inferior, while the use of historical controls allows all patients to be given the new, possibly better, treatment. The ethical dilemma of giving a treatment thought to be inferior is avoided. Physicians and patients might then be more willing to participate in a trial when only a single (better) treatment is involved. HCTs are also easier to organize.

Disadvantages

HCTs are actually observational studies, in that the investigator does not prospectively assign a treatment in the control group. In HCTs, the estimated treatment effect may be biased if controls differ systematically from the treatment group in a way that affects prognosis. Differences may occur either in the selection of patients (*see* **Selection Bias**) or in their subsequent evaluation and treatment. The analysis requires either a demonstration that baseline covariates thought to be potentially **confounding** are similarly distributed in the two groups, or a model-based approach to adjust for the baseline differences (see [6] for modeling in the context of **survival analysis**) (*see* **Baseline Adjustment in Longitudinal Studies**).

Sacks et al. [9] note that criteria for inclusion in the treatment group are usually more stringent than for inclusion in the control group (*see* **Eligibility and Exclusion Criteria**). Poor risk patients may not be offered the new treatment, which as a consequence may appear superior. It is also possible that patients with a worse prognosis may be more likely to enter a trial as "last chance" therapy, making the new treatment appear worse. Advances in techniques for cancer staging may also distort comparisons [3]. The less sensitive staging techniques in earlier years make historical controls appear to have less advanced disease, so they will appear to do worse when compared with the treatment group. Ancillary care may also be different for randomized controls, either because of a difference in surveillance or because of medical advancement in other fields.

Several authors have empirically compared HCTs and RCTs designed to answer similar questions. Diehl & Perry [2] matched historical control groups and randomized control groups for six different types of cancers. In 18 of 43 comparisons, survival or relapse-free survival differed by more than 10 percentage points, being worse in the HCT group in

17 of the 18 comparisons. Sacks et al. [9] reached a similar conclusion in a survey of 106 papers on therapeutic questions including cirrhosis with varices, coronary artery disease, acute myocardial infarction, colon cancer, melanoma, and habitual abortion. While only 20% of RCTs found benefit, 79% of HCTs did, despite similar outcomes for the treated patients in the two types of study. The differences were in the outcomes for the control groups, which tended to be poorer in the HCTs. Miller et al. [7] analyzed the results for 221 comparisons from 188 articles in six surgery journals in 1983. For primary outcomes, 79% of 19 HCTs showed the innovation to be better compared with only 50% of 20 randomized controlled trials. For secondary outcomes, 75% of eight HCTs concluded the intervention was better compared with only 57% of 61 randomized controlled trials.

In an RCT, randomization provides a theoretical foundation by which a treatment effect can be estimated and an hypothesis tested without the use of covariate information (*see* **Randomization Tests**). Randomization alone eliminates **bias**. Such is not the case for HCTs, in which the burden is to remove or reduce bias. As a consequence, HCTs may suffer from an inability to convince colleagues who require a RCT for confirmation. Funding for an HCT may then be difficult.

Requirements for a Valid Historical Control Study

Pocock [8] gives four requirements for a valid HCT: (i) the control group has received the precisely defined treatment in a recent previous study; (ii) the criteria for eligibility, workup, and evaluation must be the same (*see* **Outcome Measures in Clinical Trials**); (iii) **prognostic factors** should be completely known and be the same for both treatment groups; (iv) no unexplained indications lead one to expect different results. Gehan [4] adds an additional requirement: (v) if differences in prognostic features exist between the treatment and control groups, these should not be sufficient to explain any observed differences in outcome.

Summary

The potential for bias in historical control trials often requires complex design and analysis. For some therapies, accepted treatments have more commonly come from HCTs than from randomized controlled clinical trials (see Gehan [4] in the context of acute leukemia). In practice, HCTs tend to give results that favor the intervention. HCTs and RCTs are best viewed as complementary techniques. A large-scale randomized trial can be used to confirm the results of HCTs, and HCTs can be used to support the results of RCTs when there is difficulty in repeating trials that have shown a benefit for one of the treatment groups.

References

[1] Cranberg, L. (1979). Do retrospective controls make clinical trials "inherently fallacious?", *British Medical Journal* **2**, 1265–1266.

[2] Diehl, L.F. & Perry, D.J. (1986). A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid?, *Journal of Clinical Oncology* **4**, 1114–1120.

[3] Dupont, W.D. (1985). Randomized vs. historical clinical trials: are the benefits worth the cost?, *American Journal of Epidemiology* **122**, 940–946.

[4] Gehan, E.A. (1984). The evaluation of therapies: historical control studies, *Statistics in Medicine* **3**, 315–324.

[5] Gehan, E.A. & Freireich, E.J. (1974). Non-randomized controls in cancer clinical trials, *New England Journal of Medicine* **290**, 198–203.

[6] Keiding, N. (1995). Historical controls and modern survival analysis, *Lifetime Data Analysis* **1**, 19–25.

[7] Miller, J.N., Colditz, G.A. & Mosteller, F. (1989). How study design affects outcomes in comparisons of therapy. II: surgical, *Statistics in Medicine* **8**, 455–466.

[8] Pocock, S.J. (1976). The combination of randomized and historical controls in clinical trials, *Journal of Chronic Diseases* **29**, 175–188.

[9] Sacks, H. Chalmers, T.C. & Smith, H. (1982). Randomized versus historical controls for clinical trials, *American Journal of Medicine* **72**, 233–240.

MICHAEL L. BEACH & JOHN BARON

Bias from Loss to Follow-up

Loss to follow-up **bias** results when subjects lost from a cohort (*see* **Cohort Study**) have different health response distributions from subjects who remain in follow-up. For example, if sicker patients are lost from a cohort during follow-up, the estimated survival distribution (*see* **Survival Analysis, Overview**) will be biased upward. As another example, if a cohort of subjects with the human immunodeficiency virus (HIV) are being followed in a natural history study to track decreases in T-helper lymphocyte (CD4+ lymphocyte) levels, and if the subjects with low CD4+ lymphocyte levels are dropped from the study in order to begin treatment, then the CD4+ lymphocyte levels in those remaining on study will be upwardly biased. If loss to follow-up bias is greater in an exposed cohort than in an unexposed cohort, the estimates (*see* **Estimation**) of exposure effects will be biased, but if the same degree of loss to follow-up bias operates in both cohorts, **nondifferential error** will result, and estimated exposure effects may be **unbiased** or nearly unbiased.

(*See also* **Bias from Nonresponse**; **Bias in Cohort Studies**; **Bias in Observational Studies**; **Bias, Overview**; **Missing Data in Epidemiologic Studies**)

MITCHELL H. GAIL

Bias from Nonresponse

Nonresponse **bias** results when some members of the intended **study population** fail to provide required data (the nonresponders), and when those who respond are not representative of the entire study population (*see* **Nonresponse**). In comparative studies, such as studies comparing exposed and unexposed **cohorts**, nonresponse bias may severely distort estimates of **exposure effect** if the degree of nonresponse bias differs in the exposed and unexposed groups, resulting in differential nonresponse bias. If the degree of nonresponse bias is the same in the exposed and unexposed groups, then the nonresponse bias is said to be nondifferential, and the bias in the estimate of exposure effect may be minimal, or even zero, depending on which measure of exposure effect is used.

(*See also* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**; **Bias, Overview**; **Validity and Generalizability in Epidemiologic Studies**)

MITCHELL H. GAIL

Bias from Stage Migration in Cancer Survival

> When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states (attributed to Will Rogers, American humorist, 1879–1935).

Stage, the indicator of a tumor's anatomic dissemination, is a key predictor of cancer survival. In its simplest form, a tumor may be classified as "localized" or "stage I" if there is no evidence of spread of tumor beyond the organ of origin, "regional" or "stage II" if there is evidence of tumor spread to adjacent tissues or lymph nodes, and "distant" or "stage III" if there is evidence that the tumor is disseminated to other organs [1, 7]. "Localized" or "stage I" tumors generally have the best survival **prognosis**, while tumors with distant metastases ("stage III") have the worst. Tumors are "staged" based on morphologic and/or clinical evidence of the tumor's anatomic dissemination. Technological advances in the latter half of the twentieth century have increased the use of and reliance on sophisticated diagnostic imaging procedures for identifying disseminated disease (*see* **Image Analysis and Tomography**).

Feinstein et al. [5] demonstrated that a cancer **cohort** that has undergone use of sophisticated diagnostic imaging will have a different stage distribution than a cohort the members of which were staged with less use of imaging technology. The use of diagnostic imaging uncovers "silent" tumor dissemination to regional and distant sites that previously would have escaped clinical detection. This results in a shift of patients from less advanced to more advanced stages of disease. The effect of this stage migration is to improve cancer survival rates artifactually for both the early stage and late stage patients in the cohort subjected to imaging procedures.

Stage migration and its effect on survival rates were demonstrated in two cohorts of lung cancer patients, one diagnosed between 1953 and 1964 and the other diagnosed in 1977. The latter cohort contained staging data similar to that used in the first cohort; additionally, the latter cohort contained data from diagnostic imaging procedures not available to the first cohort. When six-month survival rates for the two cohorts were compared, the latter cohort had higher survival rates for each stage and for the cohort as a whole (Table 1). When the stage distribution of the two cohorts was compared, the latter cohort had a lower proportion of patients with stage I disease and a higher proportion of patients with stage III disease (Table 2, columns 1 and 2).

Table 1 Six month survival rates for 1953–1964 and 1977 cohorts (data from [4])

	Cohort	
	1953–1964, $n = 1266$	1977, $n = 131$
Stage I	0.75	0.92
Stage II	0.57	0.72
Stage III	0.30	0.42
Total	0.44	0.55

Table 2 Stage distributions of 1953–1964 cohort and 1977 cohort with and without diagnostic imaging data (data from [4])

	1953–1964, $n = 1266$	1977 with imaging data, $n = 131$	1977 without imaging data, $n = 131$
Stage I	0.22	0.18	0.32
Stage II	0.14	0.14	0.19
Stage III	0.64	0.68	0.49
Total	1.00	1.00	1.00

The 1977 cohort was then staged using only those data points available on the first cohort (Table 2, column 3). Without 1977 imaging data, 32% of the 1977 cohort were classified as stage I, compared with 18% when imaging data were used (Table 2, column 2). This reflects a "migration" of cases out of stage I to more advanced stages with the use of diagnostic imaging. While 49% of the 1977 cohort were classified as stage III without imaging data, 68% of the 1977 cohort were classified as stage III when imaging data were used. The stage migration from less to more advanced stages in this cohort with the use of diagnostic imaging data is detailed in Table 3.

The effect of stage migration on cohort survival rates was demonstrated by applying a standardized clinical staging system to both cohorts. There were no clinically or statistically significant differences in survival rates between the two cohorts when members were staged using standardized clinical criteria. The authors concluded that the increase in survival rates (shown in Table 1) between the two cohorts was a

Table 3 Stage migration of 1977 cohort cases when diagnostic imaging data were applied (data from [4])

Number of cases without imaging data		Number of cases with imaging data		
		Stage I	Stage II	Stage III
Stage I	42	24	1	17
Stage II	25	0	17	8
Stage III	64	0	0	64
Total	131	24	18	89

statistical artifact, rather than a real improvement in lung cancer survival [5].

The artifactual increase in survival rates over time is due to the shift of previously undetectable poor-prognosis patients (those with clinically "silent" metastases) from stage I to more advanced stages. When these poorer prognosis patients are removed from the best prognosis group, the survival rate of that group increases. Since their advanced disease is asymptomatic ("silent"), their prognosis is somewhat better than that of the original members of the poorest prognosis group (stage III), whose advanced disease is symptomatic. This raises the survival rate of the poor prognosis group, and of the cohort as a whole, without any changes in individual survival times. To paraphrase the quote attributed to Will Rogers, "When the poor prognosis patients left the best prognosis group and moved to the poorest prognosis group, they raised the survival rates in both groups".

Although the Will Rogers Phenomenon was identified and described in the comparison of cancer cohorts from different eras in time, it can occur at any time when staging methods vary between the cohorts compared. For example, the comparison of survival rates in concurrent cohorts from different geographic regions having diversity in access to imaging technology may be subject to this type of **bias**.

A slightly different form of stage migration may occur with advances in cancer treatment. This was demonstrated by Bosl et al. [2] in patients who received platinum-based chemotherapy for advanced stage germ cell tumors. As the success of this chemotherapy regimen for treating advanced disease became known, clinicians began using it for early stage germ cell tumors. Previously, these localized tumors would have been treated with surgery or radiation alone. This shifted better-prognosis patients into chemotherapy treatment (formerly reserved for poorest prognosis patients), thus improving survival rates for chemotherapy. Additionally, use of diagnostic imaging procedures increased for germ cell tumors, resulting in the classic stage migration bias described above.

Stage migration resulting in artificially inflated cancer survival rates can be avoided by ensuring that reproducible staging methods are used for study cohorts. Standardized clinical staging systems that consider data available from every patient, regardless of access to diagnostic imaging technology, have been advanced for lung and prostate cancer [3, 4, 6].

References

[1] Beahrs, O.H., Carr, D.T. & Rubin, P. eds. (1978). *Manual for Staging Cancer*, American Joint Committee for Cancer Staging and End Results Reporting. Whiting Press, Chicago.

[2] Bosl, G.J., Geller, N.L. & Chan, E.Y.W. (1988).. Stage migration and the increasing proportion of complete responders in patients with advanced germ cell tumors, *Cancer Research* **48**, 3524–3527.

[3] Clemens, J.D., Feinstein, A.R., Holabird, N. & Cartwright, C. (1986). A new clinical-anatomic tagging system for evaluating prognosis and treatment of prostate cancer, *Journal of Chronic Diseases* **39**, 913–928.

[4] Feinstein, A.R. & Wells, C.K. (1990). A clinical severity staging system for patients with lung cancer, *Medicine* **69**, 1–33.

[5] Feinstein, A.R., Sosin, D.M. & Wells, C.K. (1985). The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer, *New England Journal of Medicine* **312**, 1604–1608.

[6] Pfister, D.G., Wells, C.K., Chan, C.K. & Feinstein, A.R. (1990). Classifying clinical severity to help solve problems of stage migration in nonconcurrent comparisons of lung cancer therapy, *Cancer Research* **50**, 4664–4669.

[7] National Cancer Institute (1994). *SEER Cancer Statistics, Review, 1973–1991: Tables and Graphs*. *NIH Publication No. 94-2789*. National Cancer Institute, Bethesda.

KAREN SMITH BLESCH

Bias from Survival in Prevalent Case–Control Studies

In a prevalent case–control study, the exposures of prevalent cases sampled from among living cases are compared with the exposures of living noncases (*see* **Case–Control Study, Prevalent**). Because an exposure that causes disease may also influence the probability that an incident case will survive long enough to be sampled from the population of prevalent cases, exposure **odds ratios** from prevalent case–control studies may yield **biased** estimates of the odds ratio of etiologic interest that relates exposure to the risk of incident disease.

(*See also* **Bias, Overview**; **Biased Sampling of Cohorts**)

MITCHELL H. GAIL

Bias in Case–Control Studies

In recent years, the concept of a *study base* [5, 10, 26, 28] as the source from which any analytical epidemiologic study is derived has gained widespread acceptance [25]. Under this concept, **case–control** and **cohort studies** represent alternative approaches to sampling and information gathering from a definable population/time experience, and the **biases** that arise are a consequence of doing so inappropriately. A common earlier view [11] was that case–control studies are uniquely susceptible to bias because they "look back" from the outcome to the exposure, whereas cohort studies "look forward". For that reason, it has sometimes been claimed that case–control studies are intrinsically more susceptible to bias than cohort studies. Today, it is better recognized that while certain biases occur more commonly when using one or the other approach, others affect them equally, and the problems are not fundamentally different.

The unifying concept of a single study base might be expected in its turn to lead to unified definitions of bias, applicable both to case–control and cohort studies, and attempts to create such definitions have been made [18, 42]. But the matter is complex, and thus far none has gained wide currency. In this article, we use the existing terminology as applied to case–control studies [3, 25].

Bias is present in a case–control study if there is systematic distortion in the data that leads to an **odds ratio** estimate that is different from the true odds ratio in the study base. Because the bias is systematic, large sample sizes do not eliminate it; indeed, the only effect of enlarging sample sizes is to produce biased estimates that are more precise. Bias may arise as a consequence of **systematic errors** in the selection of cases or controls, or errors in the recording of exposure data, or because of **confounding**. When there is **nondifferential** misclassification of exposure data among cases and controls, the usual effect is to bias odds ratio estimates towards unity (*see* **Bias Toward the Null**); if such misclassification is substantial, so may be the bias. Failure to adjust for confounding may distort odds ratio estimates towards or away from unity and, again, the bias may be substantial. The reader interested in a discussion of nondifferential misclassification is referred to the articles **Misclassification Error** and **Measurement Error in Epidemiologic Studies**. Here, we focus on two remaining sources of distortion, **selection bias** and **information** bias. The term information bias is sometimes used to denote both nondifferential and differential misclassification of exposure [36]; here, we use it to denote only differential misclassification (*see* **Differential Error**).

Selection Bias

Selection bias exists when cases or controls are selected in a way that is not representative of the respective exposure distributions in the study base.

Specification of the Study Base

A fundamental step in the avoidance of selection bias is to ensure that the cases and controls are drawn from the same study base. Otherwise, if the **prevalence** of exposure is different in the different bases, bias is unavoidable. A primary study base is one in which the population/time experience, including the cases that occur, can be specified (e.g. new cases of acute myeloid leukemia (AML) occurring in the population of Massachusetts from 1990 to 1994). In a **population-based case–control study**, all cases are identified and selected; alternatively, a representative sample (e.g. a **random sample**) is selected. In either instance, a properly specified **control** series consists of noncases sampled from the same study base. When the base is well defined, and all cases are identifiable, it is possible in principle to sample them using methods that are **unbiased**. In practice, however, problems such as **nonresponse** may nonetheless lead to bias, as discussed below.

Sometimes it may not be possible to specify a primary base, as happens when a series of cases is selected without full insight into the population/time experience from which they are drawn (e.g. new cases of AML diagnosed in one hematology laboratory from 1990 to 1994). The secondary study base may then be conceived of as that population/time experience from which any person would have been selected as a case had he (or she) developed the disease under study. The proper selection of **controls** requires that they be sampled from that hypothetical secondary study base. Operationally, such controls

are often sampled from the same source as the cases (in this example, the control series could perhaps comprise a sample of persons with normal blood counts, recorded in the same laboratory). In other words, certain selection characteristics of the cases determine the secondary study base, which cannot otherwise be specified. Since it is not possible to identify members of a secondary base explicitly, it is necessary to rely on judgment and experience in order to select controls likely to be representative of the exposure in the hypothetical base.

It is worth illustrating how incorrect specification of a secondary study base may give rise to selection bias. Consider a hypothetical study of radon exposure and lung cancer, carried out in one hospital. The hospital has a large thoracic surgery department and selectively admits lung cancer cases from an entire city, in some areas of which household radon levels are high, while in other areas they are low. The hospital admits patients with conditions other than lung cancer only from an immediately adjacent area, in which household radon levels are high. In this example, the cases and controls have not been selected from the same study base, the exposure rates in the different bases are different, and the data are biased. The bias could only be overcome if the case series were to be restricted to those resident in the same area as the controls – in which case the comparison groups are now drawn from the same study base.

Selection Bias due to Nonresponse

Assuming the study base is correctly specified, selection bias may nevertheless arise if there is differential sampling or identification of cases and controls. A common way in which differential sampling may occur is if there are substantial losses in the enrollment of cases or controls originally deemed eligible for inclusion (nonresponse). In a population-based case–control study, all cases, or a representative sample of those that occur in the study base, are included. Alternatively, in a study with a secondary base, the cases should be representative of those occurring in that base (e.g. a single hospital). In each instance, potentially eligible controls should constitute a representative sample of all noncases in the same base. In practice, however, it is virtually inevitable that some of the cases or controls initially specified as eligible will not be enrolled (e.g. because of failure to trace subjects, refusal to participate, severe illness, or death). If, either among the cases or the controls, the exposure rate is systematically different among those who are and are not enrolled, there is selection bias due to nonresponse.

Bias due to nonresponse is negligible if close to 100% of the cases and controls scheduled for sampling are successfully enrolled. Some studies come close to meeting that objective. Other studies do not, and the greater the proportion unenrolled, the greater must be the concern about possible selection bias. Hospital-based case–control studies tend to have higher response rates than population-based studies, especially if it is necessary to collect biological samples.

In the face of high nonresponse rates, some limited reassurance about the absence of material selection bias may be gained when it can be shown that distributions of known characteristics (such as age, sex, or residence) are similar among enrolled and unenrolled subjects. That reassurance may be unjustified, however, if the compared variables are not themselves correlates of the exposure.

Sensitivity analyses are sometimes used in an attempt to cope with high nonresponse rates [15]. Varying assumptions are made about possible exposure rates among unenrolled subjects, and their effects upon the magnitude of any given association are then assessed. Clearly, however, the assumptions may be incorrect, and that possibility limits the interpretability of the data. In general, the more the attrition, the greater is the potential for bias. To limit this source of bias, there is simply no substitute for high enrollment rates.

Selection Bias in the Identification of Study Subjects

Selection bias may also arise if cases or controls are identified in a way that is not independent of the exposure (*see* **Detection Bias**). To illustrate how the biased identification of cases may occur, consider a hypothetical example in which each of two women has a tender and swollen leg due to deep vein thrombosis (DVT); one woman takes oral contraceptives (OCs – a known cause [40]), the other does not. The OC taker is aware that she is at risk of DVT, and so consults her physician, who correctly makes the diagnosis and admits her to a hospital; the nonuser stays home, undiagnosed; both women recover. Even though the diagnosis, when made, is

correct, the case identification is incomplete, and exposure-dependent: a case–control study that enrolls cases of DVT, regardless of whether it is population-based or derived from a secondary study base, would overestimate the association with OC use, because knowledge of the exposure increases the likelihood that exposed cases would be included in the study.

This type of selection bias may take many forms, as can be illustrated further by the following examples. In a study of breast cancer risk in relation to female hormone use, "screening" or "detection" bias may arise if hormone users are more commonly subjected to mammography than nonusers (e.g. because of concern about possible breast cancer risk). As a result, users are more commonly diagnosed than nonusers as having breast cancer that might not otherwise have become clinically apparent for many years [39].

There are instances in which the biases brought about by **screening** may not be at all subtle. It has been estimated, for example, that over 60% of men over the age of 60 years have asymptomatic prostatic cancer [12]. It is now possible to detect cases that would otherwise have remained asymptomatic for many years, and perhaps for life, by means of a new test (the prostate-specific antigen test [41]). Any **association** of prostatic cancer with a correlate of the likelihood of undergoing such a test (e.g. high socioeconomic status) would likely be biased. That bias may be reduced, or perhaps avoided, in a study restricted to cases that must inevitably come to diagnosis, regardless of screening, because of symptoms that oblige them to seek medical care, such as hematuria or bone pain due to metastases. As a general rule, studies that enroll cases (or controls) from screening programs run a substantial risk of selection bias [4, 29].

A similar bias arises when registries that selectively record exposed cases are used as sources for case enrollment (*see* **Disease Registers**). Perhaps the best-known example of this type was the American Registry of Blood Dyscrasias [48], which was initiated in the mid-1950s and maintained for over a decade following reports of an association of aplastic anemia with the use of the antimicrobial drug, chloramphenicol [44]. Exposed cases were far more likely than nonexposed cases to be reported. There is little doubt that chloramphenicol does indeed increase the risk of aplastic anemia, but it is now clear that the association was overestimated [19]. This example illustrates how important it is to ensure that *all* cases within any specified study base, whether primary or secondary, should have the same chance of being identified and included in a case–control study, regardless of exposure status. It also illustrates the limited interpretability of **case series** reported to regulatory agencies, or to medical journals, without reference to the background occurrence of nonexposed cases – or, indeed, without reference to the exposure prevalence among suitably selected controls.

Biased identification of cases may also occur when a cluster of exposed cases gives rise to a hypothesis, and then the same cluster is included in an independent study mounted to confirm the hypothesis. Thus, if a cluster of cases of leukemia occurs in the vicinity of a nuclear power plant [1], it would be inappropriate to include that cluster in an independent study designed to test the hypothesis that leukemia is associated with proximity to a nuclear power plant.

Bias due to the Selection of Nonrepresentative Controls from a Secondary Study Base

Hospitalized patients continue to constitute the most commonly selected controls in case–control research [22, 25, 45–47], and the potential problems posed by their selection serve well to illustrate the biases that may arise when controls are selected from a hypothetical secondary study base.

Particular attention must be paid to ensure that hospitalized patients selected for inclusion as controls have been admitted for diseases that are independent of the exposure under study. An illustration of how bias may occur if this is not done is the classical study by Doll & Hill of smoking and lung cancer [7], in which the control series included patients with chronic bronchitis, a disease not appreciated at the time to be tobacco-related [8, 9]. The magnitude of the association with lung cancer was somewhat underestimated for that reason. The association was nevertheless identified, because most of the control diagnoses were independent of smoking status, and because smoking was more strongly associated with lung cancer than with chronic bronchitis (*see* **Smoking and Health**).

Despite the risk of biases of this type, hospital-based studies have remained a mainstay of case–control research. When well conducted, such studies have continued to document important and valid associations (e.g. OCs and myocardial

infarction [34]). There are several reasons. Interview data are usually less biased among hospital controls (see the section "Information Bias" below), and hospital-based studies are usually easier to conduct than studies that enroll community controls: response rates are usually higher, and when needed, high success rates in obtaining blood or tissue samples can be achieved. Perhaps the most important reason, however, is that there is today a better appreciation of the steps that should be taken to ensure that the selection of hospital controls is unbiased.

In formal terms, the selection of hospital controls is unbiased if the control diagnoses that are selected for inclusion are representative of the exposure distribution in the hypothetical secondary base. In practice, there is usually no reason why such controls cannot be identified if only those persons are selected whose reason for hospital admission (the *primary* diagnosis) is independent of the exposure under study. Those admitted for conditions that are not independent of the exposure should be excluded. When in doubt, one should opt for *exclusion*: what matters is that the selection of those subjects that are *included* should be valid. Clearly, the valid selection of hospital controls calls for experience and judgment. If that judgment is called into question, the interpretation of hospital-based case–control data can sometimes be controversial.

Hospitalized patients, whether cases or controls, commonly have more than one diagnosis, and it is important to note that *secondary* diagnoses are irrelevant to the selection of controls, unless the *secondary* diagnoses have also influenced the selection of the cases (which is unusual). For example, in a study of the risk of myocardial infarction (MI) in relation to OC use [34], cases admitted for a *primary* diagnosis of MI who had a *secondary* diagnosis of diabetes mellitus (a condition that is inversely associated with OC use) were not excluded; correspondingly, controls admitted with *primary* diagnoses unrelated to OC use, such as trauma, but with a *secondary* diagnosis of diabetes, were also not excluded. Instead, potential confounding due to diabetes was controlled in the analysis. If, however, we conceive of a hypothetical study in which patients admitted for MI are excluded if they are also diabetic, then controls admitted for trauma who also happen to be diabetic should also be excluded.

As a general rule, persons whose primary diagnoses are acute conditions for which admission is obligatory (e.g. trauma; appendicitis) meet the requirement of independence, as may persons with other conditions (e.g. elective admission for cataract surgery). However, it is always necessary to consider the particular hypothesis under study, and to use informed judgment. For example, consider a study of the risk of ovarian cancer in relation to OC use [35]: among women hospitalized for trauma, the reason for admission is likely to be independent of the exposure; such women would be eligible as controls. However, if the example is changed to a study of breast cancer risk in relation to alcohol intake [33]), trauma would not be a suitable control diagnosis because its occurrence may not be independent of the exposure.

Reassurance that the identification of hospital controls is unbiased may be gained if the exposure rates among major diagnostic categories (e.g. trauma, acute infections, orthopedic conditions) are uniform: in that circumstance, bias is only possible if the selection of an entire control series is biased, and biased to the same degree for each diagnostic category. However, the confident demonstration of uniformity requires that the categories be large enough to ensure that the rates in each of them are reasonably precise.

As an alternative, it has been suggested [25] that a hospital control series should include as wide a range of diagnoses as possible. If it can be assumed that most of them will be independent of the exposure, then any bias, if present, will be diluted. This latter option is seldom acceptable unless there are good grounds to be reasonably sure that by far the overwhelming majority of the individual diagnoses that led to admission are, indeed, independent of the exposure. This is rarely the case. For this reason, the selection of a random sample of an entire hospital population, without any regard for diagnostic eligibility, can seldom be defended.

Despite the generally distinguished record of hospital-based case–control studies, some epidemiologists have argued that hospital controls are almost always unrepresentative of exposure in the population at large [43, 49]. That argument ignores the premise that when the cases represent a secondary study base (as is usual in hospital-based studies), the only valid control series may be patients admitted to the same hospitals as the cases.

One theoretical drawback to the sampling of hospital controls is that the judgment that the included conditions are independent of the exposure is an assumption, and one that is not needed when selecting

controls in a **population-based study**. In addition, there is evidence to suggest that certain exposures differ for in-hospital and out-of-hospital populations [25, 45–47]. That evidence, however, has been derived from studies that did not take into account the specific eligibility of each control diagnosis in the context of the specific hypothesis under study – an essential step in the proper selection of hospital controls. Nevertheless, there may be circumstances when the exposure under study (e.g. alcohol [33]) influences admission across such a wide range of diagnoses that it may be difficult or impossible to select a valid series of hospital controls.

Bias due to the Nonrepresentative Selection of Controls from a Primary Study Base

Selection bias may arise in analogous ways when population-based controls are chosen [45–47]. For example, if the sampling scheme is based on incomplete coverage of the base population (e.g. a motor vehicle owners' registry, or **random digit dialing**), it may underrepresent people of low socioeconomic status (because they do not have cars or telephones). Similarly, the selection of other controls, such as friends of the cases, or classmates, may give rise to other problems. For example, nonexposed friends of an exposed case may tend selectively to participate in a study because they would like to help. As with hospital controls, it remains important to use judgment in ensuring that population-based controls, however selected, are representative of the study base.

In an idealized example of a **population-based case–control study** with a 100% response rate among the sampled cases and among controls, confidence in the validity of the findings would be greater than for an otherwise identical study in which hospital controls are selected because no unverifiable assumptions about representativeness are required. In practice, however, that theoretical advantage is commonly not achieved because response rates among population controls tend to be considerably lower than among hospital controls, and lower still when it is necessary to obtain biological samples.

Partly in order to circumvent the problem of low response rates in the selection of population controls, random digit dialing [16, 25, 46] has been advocated as one way to obtain high participation rates, at least in societies with almost universal telephone coverage. This method had its origins in market research and opinion polls, and its application in epidemiologic research enjoyed some early success. However, with the passage of time, answering machines, voice mail, call forwarding, and an increasingly hostile attitude in society to what are perceived to be invasions of privacy, have lowered response rates, and sometimes even rendered such rates unmeasurable (because the presence or absence in the household of a potentially eligible control could not be determined) [13]. Despite these difficulties, however, adequate participation rates can sometimes be achieved if the interviewers are carefully trained and care is taken with the wording of invitations to participate (*see* **Interviewing Techniques**).

Selection Bias in Nested Case–Control Studies

In recent years **nested case–control studies** have come to play an increasingly important role in case–control methodology. In a nested case–control study, the cases are members of a cohort who develop a given condition, and the controls are a sample of noncases selected from the same cohort, and followed for the same length of time. There are several advantages to this approach: the cases and controls are unambiguously representative of the same study base; if the follow-up has been successful nonresponse rates are low; and information bias (see below) is avoided, since exposure status is usually determined before the subject qualifies as a case. A further advantage is that it may be easier to obtain biological specimens from people who are already collaborating in a study. All of these advantages were demonstrated in a study [31] that documented an increased risk of stomach cancer in relation to antecedent *Helicobacter pylori* infection, as determined from immunological assays of frozen serum specimens that had been collected and stored an average of 14 years earlier. As a general rule, however, a major disadvantage to the conduct of nested case–control studies is that it may not be possible to assemble sufficient cases, unless the follow-up study is massive.

Information Bias

Information bias exists when cases or controls report their exposures differently, or when the information is solicited differently (as noted above, in this article, bias due to nondifferential misclassification of exposure is excluded from the definition). The likelihood

that information bias will occur is greatest when the study subject, or those responsible for collecting the data, know the hypothesis.

Differential reporting (**recall bias**) may occur if cases aware of the hypothesis tend to report their exposures more fully than controls, with resultant overestimation of the odds ratio. For this bias to occur, it is not necessary to assume that cases may report exposures that did not actually take place (although they may overestimate duration or dosage). Even if the controls share knowledge of the hypothesis with the cases, if they are healthier they may have less reason to probe their memories. For example, one study of breast cancer risk in relation to OC use [24] specifically informed participants of the hypothesis, thus rendering recall bias all but unavoidable (this is also an example of biased solicitation of information – see below). Cases may also be prompted to recall their exposures more completely if their memories (but not those of the controls) have already been "primed" by repeated questioning from their medical attendants about the putative cause before they are interviewed by the study personnel.

A lack of awareness of the hypothesis reduces the likelihood of information bias, but it does not necessarily eliminate it. Hospitalized cases, for example, because of the setting in which the questions are asked, may remember their exposures better than population controls interviewed at home. For this reason, the interviewing of controls in a hospital setting may reduce the likelihood of information bias. Similarly, without any specific hypothesis in mind, patients with cancer, or mothers who have given birth to children with birth defects, may be more inclined than controls to probe their memories for possible "causes", even if such "causes" have not specifically been hypothesized.

These examples illustrate how cases might report their exposures more fully than controls. Sometimes, however, the reverse may occur. In a hypothetical study of trauma in relation to alcohol intake, for example, the cases might understate their consumption relative to controls if they are embarrassed at having contributed to their own illness.

Much the same considerations that apply to recall bias on the part of the study subjects may also apply when those responsible for the data collection are aware of the hypothesis, and it is not uncommon for such awareness to coexist both among the subjects and the study personnel, as in the OC/breast cancer example [24] mentioned above. Or, to give another example, in a further study of the same question [30], women with breast cancer were interviewed face-to-face by a single male physician, while the controls were subjected to telephone interviews, conducted by two female interviewers. The biased solicitation of exposure information may be quite subtle: the inflection of an interviewer's voice, the "body language", the use of open-ended questions, or the way in which they are worded may all influence the respondent's answers.

It is sometimes argued that the presence of **dose–response** or duration–response effects constitutes evidence against information bias. This argument may have merit inasmuch as long-duration exposures, and perhaps high doses, are less likely to be misremembered than short-duration exposures or low doses. The countervailing argument, however, is that cases may tend systematically to overreport, and controls to underreport, duration or dosage. Thus, apparent duration or dosage gradients cannot necessarily be taken as evidence against information bias.

Occasionally, it is possible to avoid or minimize information bias. To give some examples: in a study of breast cancer risk in relation to use of the antihypertensive drug reserpine [20], women were questioned before their breast lumps were biopsied: the cases were those with breast cancer, and the controls were women given a diagnosis of benign breast disease. (But it should be noted that selection bias may have been present if reserpine increased the risk of benign breast disease – an instance in which the control diagnosis would not be independent of the exposure.) In a study of spermicide use at the time of conception in relation to Down's syndrome [23], pregnant women were questioned about exposure before they underwent an amniocentesis: the cases were fetuses with trisomy 21, the controls were fetuses with normal chromosome counts. And in a study of uterine cancer in relation to conjugated estrogen use [50], medical records were examined for prescriptions after all information on case or control status was masked (*see* **Blinding or Masking**): the biased recording of the exposure information was thus avoided.

Information bias can sometimes be assessed by the independent evaluation of exposure, using information from other sources. For example, some interview-based case–control studies have suggested

that induced abortion increases the risk of breast cancer [6]. Information bias could account for the association if women with breast cancer more fully report such a sensitive exposure than do control women [32]. Evidence that this may be so is suggested by the results of a recent Danish cohort study [27] based on national registry data. An increased risk was ruled out, and in that study there was no information bias [17], since the data on abortion status were recorded before the breast cancer outcomes were observed. Indeed, this example serves to illustrate how cohort studies can avoid information bias.

Unfortunately, as illustrated by the abortion/breast cancer example, the circumstances in which information bias can confidently be ruled out in case–control studies tend to be the exception rather than the rule. Usually, even if the investigator judges information bias to be minimal, it may not be possible to demonstrate that this is so. It is necessary to resort to the next best alternative, which is to design studies in which the potential for information bias is reduced as much as possible: for example, by concealment of the hypothesis from the study subjects, and the interviewers – or, if that is not possible, by avoiding mention of the hypothesis; by the use of highly structured and unambiguous questions; memory prompts (e.g. photographs of OCs) to maximize recall; the administration of questionnaires as soon as possible, before there is a substantial opportunity for memory loss; and the rigorous training of interviewers (*see* **Interviewing Techniques**; **Questionnaire Design**).

Even with optimal study design, the question of whether information bias is, or is not, sufficient to invalidate an association is ultimately a matter of judgment. For example, in a study documenting an increased risk of sinonasal cancer among workers exposed to wood dust [21], we may judge that occupational exposure is likely to be equally well remembered by cases and controls. Alternatively, we may judge that information bias is likely, as with the example of breast cancer risk in relation to a history of induced abortion [6, 32].

Conclusions

In this article, we have described two types of systematic bias (confounding, of course, is a third). Yet, in the past, the view has sometimes been taken that there are many more types of bias, each of them sufficiently different to require separate classification, that may affect observational studies: Sackett [37] described more than 35 (*see* **Bias, Overview**). However, all the specific biases that have been reported can readily be classified as instances of selection bias and information bias. For example, **Berkson's fallacy** [2], the proposition that the selection of hospitalized cases may be biased if admission for the condition under study is dependent on the coexistence of another condition, is a form of selection bias.

Systematic bias due to confounding has not been considered in this article. But it is important to mention that selection bias or information bias may affect not only the recording of exposures, but also the recording of confounders. Indeed, both the differential and the nondifferential recording of a confounder can lead to residual confounding, with a bias that can act in either direction [14, 38].

Information bias is sometimes mentioned as the Achilles' heel of case–control methodology. One of the major advantages of follow-up studies, relative to case–control studies, is that exposures are usually measured before the health outcomes occur, thus reducing or eliminating the likelihood of information bias. However, that advantage may be offset by biases that sometimes affect cohort studies, such as high nonresponse rates on follow-up, with differential losses according to exposure status. Another potential disadvantage is that changes in exposure status over time may be missed in cohort studies, unless the recording of the variables at issue is updated frequently (*see* **Bias in Cohort Studies**). And to complete the picture, certain biases (e.g. confounding, selection bias due to knowledge of exposure, or due to selective screening according to exposure status) may affect both approaches. In short, neither the case–control nor the cohort approach can circumvent all sources of bias, and they should, instead, be thought of as complementary strategies, each with certain strengths and certain weaknesses.

Since bias cannot be entirely eliminated in **observational studies**, concern about whether its existence is sufficient to invalidate any given association may be reduced if, in any study, the magnitude of the effect, relative to the magnitude of the plausible biases that may exist, is considerable. By contrast, the possibility of bias limits the interpretability of small associations. Concern about validity may also be reduced when a

variety of well-conducted studies based on different epidemiologic methods, and some of them based on nonepidemiologic methods, converge on the same large and relatively invariant association – an obvious example being lung cancer and smoking (*see* **Smoking and Health**), for which the validity of a causal connection (*see* **Causation**; **Hill's Criteria for Causality**) has long been beyond dispute.

Finally, one strength of the case–control approach can also be considered a major weakness: the ease with which case–control studies can sometimes be done, relative to cohort studies, means that the method can also more easily be abused. Case–control studies should be carried out by experienced investigators who are aware of their limitations, and they should be designed to anticipate and cope with potential sources of bias. When this has been done, there can be no doubt that they have made a major contribution to medical knowledge and to public health.

References

[1] Beral, V. (1990). Childhood leukemia near nuclear plants in the United Kingdom: the evolution of a systematic approach to studying rare disease in small geographic areas, *American Journal of Epidemiology* **132**, Supplement, S63–S68.

[2] Berkson, J. (1976). Limitations of the application of fourfold table analysis to hospital data, *Biometrics Bulletin* **2**, 47–53.

[3] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. I. *The Analysis of Case–Control Studies*, *IARC Scientific Publication* No. 32. International Agency for Research on Cancer (IARC), Lyon.

[4] Cole, P. & Morrison, A.S. (1980). Basic issues in population screening for cancer, *Journal of the National Cancer Institute* **64**, 1263–1272.

[5] Cornfield, J. & Haenszel, W. (1960). Some aspects of retrospective studies, *Journal of Chronic Diseases* **11**, 523–524.

[6] Daling, J.R., Malone, K.E., Voigt, L.F., White, E. & Weiss, N.S. (1994). Risk of breast cancer among young women: relationship to induced abortion, *Journal of the National Cancer Institute* **86**, 1584–1592.

[7] Doll, R. & Hill, A.B. (1952). A study of the aetiology of carcinoma of the lung, *British Medical Journal* **2**, 1271–1286.

[8] Doll, R. & Hill, A.B. (1964). Mortality in relation to smoking: ten years' observation of British doctors, *British Medical Journal* **1**, 1399–1410, 1460–1467.

[9] Doll, R. & Peto, R. (1976). Mortality in relation to smoking: 20 years' observations on male British doctors, *British Medical Journal* **ii**, 1525–1536.

[10] Dorn, H.F. (1959). Some problems arising in prospective and retrospective studies of the etiology of disease, *New England Journal of Medicine* **261**, 571–579.

[11] Feinstein, A.R. (1975). The epidemiologic trohoc, the ablative risk ratio, and retrospective research, *Journal of Clinical Pharmacology and Therapy* **14**, 291–306.

[12] Gittes, R.F. (1991). Carcinoma of the prostate, *New England Journal of Medicine* **324**, 236–245.

[13] Greenberg, E.R. (1990). Random digit dialing for control selection. A review and a caution on its use in studies of childhood cancer, *American Journal of Epidemiology* **131**, 1–5.

[14] Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564–569.

[15] Greenland, S. (1996). Basic methods for sensitivity analysis of biases, *International Journal of Epidemiology* **25**, 1107–1116.

[16] Hartge, E.R., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., Hoover, R.N. & Waksberg, J. (1984). Random digit dialing in selecting a population-based control group, *American Journal of Epidemiology* **120**, 825–833.

[17] Hartge, P. (1997). Abortion, breast cancer, and epidemiology (Editorial), *New England Journal of Medicine* **336**, 127–128.

[18] Kass, P.H. (1992). Converging toward a "Unified Field Theory" of epidemiology (Editorial), *Epidemiology* **3**, 473–474.

[19] Kaufman, D.W., Kelly, J.P., Levy, M. & Shapiro, S. (1991). *The Drug Etiology of Agranulocytosis and Aplastic Anemia*. Oxford University Press, Oxford.

[20] Kewitz, H.J., Jesdinsky, H., Schröter, P. & Lindtner, E. (1977). Reserpine and breast cancer in women in Germany, *European Journal of Clinical Pharmacology* **2**, 79–83.

[21] Leclerc, A., Martinez Cortes, M., Gerin, G., Luce, D. & Brugere, J. (1994). Sinonasal cancer and wood dust exposure: results from a case–control study, *American Journal of Epidemiology* **140**, 340–349.

[22] Linet, M.S. & Brookmeyer, R. (1987). Use of cancer controls in case–control cancer studies, *American Journal of Epidemiology* **125**, 1–11.

[23] Louik, C., Mitchell, A., Werler, M., Hanson, J. & Shapiro, S. (1987). Maternal exposure to spermicides in relation to certain birth defects, *New England Journal of Medicine* **317**, 474–478.

[24] Lund, E., Meirik, O., Adami, H.O., Bergstrom, R., Christoffersen, T. & Bergsjo, P. (1989). Oral contraceptive use and premenopausal breast cancer in Sweden and Norway: possible effects of a different pattern of use, *International Journal of Epidemiology* **18**, 527–532.

[25] MacMahon, B. & Trichopoulos, D. (1996). *Epidemiology. Principles and Methods*. Little, Brown & Company, Boston.

[26] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[27] Melbye, M., Wohlfahrt, J., Olsen, J.H., Frisch, M., Westergaard, T., Helweg-Larsen, K. & Andersen, P.K. (1997). Induced abortion and the risk of breast cancer, *New England Journal of Medicine* **336**, 81–85.

[28] Miettinen, O.S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. Wiley, New York.

[29] Morrison, A.S. (1985). *Screening in Chronic Disease*, 2nd Ed. Oxford University Press, Oxford.

[30] Olsson, H., Moller, T.R. & Ranstam, J. (1989). Early oral contraceptive use and breast cancer among premenopausal women: final report from a study in southern Sweden, *Journal of the National Cancer Institute* **81**, 1000–1004.

[31] Parsonnet, J., Friedman, G.D., Vandersteen, D.P., Chang, Y., Vogelman, J.H., Orentreich, N. & Sibley, R.K. (1991). *Helicobacter pylori* infection and the risk of gastric carcinoma, *New England Journal of Medicine* **325**, 1127–1131.

[32] Rosenberg, L. (1994). Induced abortion and breast cancer: more scientific data are needed (Editorial), *Journal of the National Cancer Institute* **86**, 1569–1570.

[33] Rosenberg, L., Metzger, L.S. & Palmer, J.R. (1993). Epidemiology of breast cancer. Alcohol consumption and the risk of breast cancer: a review of the evidence, *Epidemiology Review* **15**, 133–144.

[34] Rosenberg, L., Palmer, J.R., Lesko, S.M. & Shapiro, S. (1990). Oral contraceptive use and the risk of myocardial infarction, *American Journal of Epidemiology* **131**, 1009–1016.

[35] Rosenberg, L., Shapiro, S., Slone, D., Kaufman, D.W., Helmrich, S., Miettinen, O.S., Stolley, P., Rosenshein, N.B., Schottenfeld, D. & Engle, R.L. (1982). Epithelial ovarian cancer and combination oral contraceptives, *Journal of the American Medical Association* **247**, 3210–3212.

[36] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown & Company, Boston.

[37] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[38] Shapiro, S., Castellana, J.V. & Sprafka, J.M. (1996). Alcohol-containing mouthwashes and oropharyngeal cancer: a spurious association due to underascertainment of confounders?, *American Journal of Epidemiology* **144**, 1091–1095.

[39] Skegg, D.C.G. (1988). Potential for bias in case–control studies of oral contraceptives and breast cancer, *American Journal of Epidemiology* **127**, 205–212.

[40] Stadel, B.V. (1981). Oral contraceptives and cardiovascular disease, *New England Journal of Medicine* **305**, 612–618.

[41] Stamey, T.A., Yang, N., Hay, A.R., McNeal, J., Freiha, F.S. & Redwine, E. (1987). Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate, *New England Journal of Medicine* **317**, 909–916.

[42] Steineck, G. & Ahlbom, A. (1992). A definition of bias founded on the concept of the study base, *Epidemiology* **3**, 477–482.

[43] Swan, S.H., Shaw, G.M. & Schulman, J. (1992). Reporting and selection bias in case–control studies of congenital malformations, *Epidemiology* **3**, 356–363.

[44] Volini, I.F., Greenspan, I., Ehrlich, I., Gonner, J.A., Felfenfeld, O. & Schwartz, S.R. (1950). Hemopoietic changes during administration of chloramphenicol, *Journal of the American Medical Association* **42**, 1333–1335.

[45] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case–control studies. I. Principles, *American Journal of Epidemiology* **135**, 1019–1028.

[46] Wacholder, S., Silverman, D.T., McLaughlin, J.K. & Mandel, J.S. (1992). Selection of controls in case–control studies. II. Types of controls, *American Journal of Epidemiology* **135**, 1029–1041.

[47] Wacholder, S., Silverman, D.T., McLaughlin, J.K. & Mandel, J.S. (1992). Selection of controls in case–control studies. III. Design options *American Journal of Epidemiology* **135**, 1042–1050.

[48] Welch, H., Lewis, C.N. & Kerlan, I. (1954). Blood dyscrasias, a nationwide survey, *Antibiotics and Chemotherapy* **4**, 607.

[49] West, D.W., Sehaman, K.L., Lyon, J.L., Robison, L.M. & Allred, R. (1984). Differences in risk estimation from a hospital and a population-based case–control study, *International Journal of Epidemiology* **13**, 235–239.

[50] Ziel, H.K. & Finkle, W.D. (1975). Increased risk of endometrial carcinoma among users of conjugated estrogens, *New England Journal of Medicine* **293**, 1167–1170.

SAMUEL SHAPIRO & LYNN ROSENBERG

Bias in Cohort Studies

An epidemiologic **cohort** (or follow-up) study is typically performed by: (i) identifying a group of subjects who are at risk for a disease or condition of interest; (ii) determining the exposure status of each individual; and (iii) observing the subjects over time for the occurrence of the health outcome(s) under investigation. While this approach is advantageous in that it ensures that the temporal relationship between exposure and outcome is unambiguous, cohort studies are susceptible to the same kinds of **bias** (i.e. **selection**, **misclassification**, and **confounding**) as are other types of study design (*see* **Bias in Observational Studies**; **Bias, Overview**).

Selection bias occurs when the study population available for analysis is not representative of the (theoretical) cohort of all eligible participants. This may result from biased sampling of the eligible cohort and/or selective losses from the study population during follow-up (*see* **Biased Sampling of Cohorts**). The common attribute of all sources of selection bias is that the effect estimated from the available study population is meaningfully different from the one that would have been obtained had all subjects theoretically eligible to participate been included in the analysis. A potential source of nonrepresentative sampling is self-selection, whereby subjects who become aware of a study volunteer themselves for participation. If such volunteers have a different probability of developing the outcome of interest compared with the group of all eligible subjects [1, 4], then the result may be a biased estimate of effect. A related concept is the "healthy worker effect," based on the observation that people in the workforce have lower mortality rates than members of the general population [2, 4]. Studies that utilize workers must take this situation into account in order to avoid biased results (*see* **Occupational Epidemiology**).

Other important potential sources of selection bias in cohort studies include losses to follow-up and **nonresponse** during data collection (*see* **Cohort Study**; **Missing Data in Clinical Trials**; **Missing Data in Epidemiologic Studies**). However, bias is created only if data are missing disproportionately from one or more cells of the **2 × 2 table** that relates a dichotomous exposure to a dichotomous outcome. Follow-up on a given subject may be incomplete for a variety of reasons. The subject may choose to withdraw his or her consent and no longer participate in the study. More commonly, the investigator simply loses contact with the subject, and thus cannot know with confidence whether he or she experienced the outcome of interest during the relevant follow-up period. When the outcome of interest is time to an event (e.g. death, diagnosis of disease, relapse, etc.), those individuals who do not experience the event during the study period are said to be **censored**. An assumption that is generally made with regard to such studies is that subjects who are censored at a given time have similar risks compared with those not censored at that time, i.e. that the censoring is "noninformative". If this assumption is violated, then bias results. Biased sampling may also occur if the probability that one is selected into the study sample depends upon whether he or she experiences the event of interest during a prescribed time window.

A final example of selection bias is illustrated by the "prevalent" cohort study. In such a study, subjects who already have a disease or other health condition are enrolled and then followed over time for events such as disease progression, relapse, or death. The goal is to obtain information about the natural history of the disease; however, problems of interpretation arise if the time since disease diagnosis remains unknown and is not uniform across subjects. Such difficulties must be weighed against the effort and costs required to assemble a cohort of newly diagnosed subjects, as would be done with the analogous "incident" cohort study.

Another type of bias to which cohort studies are susceptible is misclassification bias, a distortion in effect estimation that occurs when measurement errors result in incorrect classification of the exposure and/or disease status of study participants (*see* **Measurement Error in Epidemiologic Studies**). Such misclassification errors may be differential or nondifferential. When errors made in classifying subjects along one axis (i.e. exposure or disease) are independent of the subject's status on the other axis, the misclassification is said to be **nondifferential**. If the magnitude of the error along one axis varies according to the category of the other axis (e.g. disease status is misclassified more frequently among the unexposed), then differential misclassification has occurred [4]. This distinction is of value, since, for dichotomous exposure and disease variables, nondifferential misclassification leads consistently to an underestimation of the magnitude of

the association (**bias toward the null**). In contrast, bias from differential misclassification can be in any direction.

Classification errors arise from a variety of sources, including imprecise measurement tools, mistaken or missed diagnoses, and conscious or unconscious inaccuracies in self-reported disease and/or exposure information. An inaccurate diagnostic tool could lead to either over- or underascertainment of cases, causing disease misclassification that may be either differential or nondifferential with respect to exposure status. A potential source of differential disease misclassification is **detection bias**, whereby exposed subjects are followed more closely than their unexposed counterparts and are thus less likely to have unrecognized subclinical disease. The behavior of study personnel can also affect the accuracy of the data being collected. For example, an interviewer who is aware of both the study hypothesis and the exposure status of study participants may be more thorough in his or her questioning of exposed subjects regarding signs and symptoms indicative of the outcome of interest. Such **interviewer bias** is another potential source of differential misclassification and could thus lead to invalid study results.

The final category of bias which may affect cohort studies is **confounding**. Confounding operates similarly in all types of study designs, occurring when the effect of the exposure of interest is mixed up with that of one or more "extraneous" variables. The result can be over- or underestimation of the true effect of the exposure, the magnitude of the bias depending upon the nature of the relationships among the confounder(s), the exposure, and the disease. Confounding can be addressed in the design stage of a study – using **randomization**, restriction or matching – or in the analysis stage by employing stratified analysis (*see* **Stratification**) or applying a mathematical modeling technique [3] (*see* **Matching**; **Matched Analysis**).

References

[1] Criqui, M.H., Austin, M. & Barrett-Connor, E. (1979). The effect of non-response on risk rations in a cardiovascular disease study, *Journal of Chronic Diseases* **32**, 633–638.

[2] Fox, A.J. & Collier, P.F. (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry, *British Journal of Preventive and Social Medicine* **30**, 225–230.

[3] Kleinbaum, D.G., Kupper, L.L. & Morganstern, H. (1982). *Epidemiologic Research*. Van Nostrand Reinhold, New York.

[4] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown & Company, Boston.

HOLLY A. HILL & DAVID G. KLEINBAUM

Bias in Observational Studies

Bias can be defined as any **systematic error** (in contrast to sampling error) that results in inaccurate **estimation** of the effect of an exposure on an outcome. Such errors may occur in the design and/or analysis phase of an epidemiologic study and may result in either over- or underestimation of the true effect. Since bias is due to systematic rather than **random error**, the magnitude of the bias is not affected by sample size. Studies that produce effect estimates free of bias are said to be internally valid. An estimate which is internally valid may or may not be considered externally valid as well; the contrast between internal and external validity is discussed below. Although many sources of bias have been identified [14, 20, 21, 23], biases can generally be classified into one of three categories: **selection bias**, information bias, and **confounding** [14].

Selection Bias

Selection bias refers to a distortion in the estimate of effect resulting from (i) the manner in which subjects are selected for the study population and/or (ii) selective losses from the study population prior to data analysis. There are many sources of selection bias, and more than one source can contribute to bias in a given study. The common attribute of all sources of selection bias is that the effect estimated from the available study population is meaningfully different from the one that would have been obtained had all subjects theoretically eligible to participate been included in the analysis. Selection bias can occur under any type of study design; however, it is of special concern in the design and conduct of **case–control studies** because the outcome has already occurred prior to selection of study subjects.

Sources of Selection Bias

In **cohort** (follow-up) **studies**, two important potential sources of selection bias include losses to follow-up and **nonresponse** during data collection (*see* **Bias in Cohort Studies**; **Missing Data in Clinical Trials**; **Missing Data in Epidemiologic Studies**). Both situations result in missing exposure and/or outcome information at the time of analysis for some eligible subjects. This creates the potential for selection bias, depending upon whether information is missing disproportionately from one or more cells of the **2 × 2 table** that relates a dichotomous exposure to a dichotomous outcome (Table 1). Because the degree of bias relates to the amount of missing data in a cell relative to each of the other cells, selection bias may occur even with a fairly high overall response rate and/or very little loss to follow-up. For example, a cohort study might be conducted in which 90% of all subjects originally assembled into the cohort remain available for analysis at the end of the study (i.e. 10% of subjects were lost to follow-up). If the losses to follow-up are concentrated among the exposed subjects who ultimately developed the disease, the true **relative risk** could be underestimated by a substantial amount. Conversely, there may be no selection bias despite small response rates and/or large follow-up losses in each exposure category. If only 20% of all eligible subjects choose to participate in a study, but this 20% represents a true **random sample** of all potential participants, the resulting estimate of relative risk will be **unbiased**. In fact, even an assumption this stringent is not required. As long as, within each exposure category, the likelihood of being selected into the study (and available for analysis) is the same for subjects who develop the disease and subjects who do not, the risk for each exposure group (and thus the relative risk) can be estimated without bias.

The validity of any effect estimated from case–control data depends in part upon the appropriate choice of a comparison (**control**) group. The purpose of the controls is to estimate the **prevalence** of the exposure(s) of interest in the population from which the cases emerged. Any control group that yields over- or underestimates of this prevalence produces biased study results, unless there are compensating biases in the case sample. In terms of the 2×2 table (Table 1), selection bias results from an imbalance in the probability of being selected into the study (or remaining for the analysis) across the four

Table 1 2×2 table

	Exposed	Unexposed
Diseased	*a*	*b*
Nondiseased	*c*	*d*

cells. For example, **detection** (also called diagnostic or unmasking) **bias** [12, 21] can result from closer follow-up or more intense scrutiny of exposed vs. unexposed subjects. The detection of a higher proportion of subclinical outcomes among the exposed leads to an overrepresentation of exposed cases in a case–control study. Hospital-based case–control studies are subject to unique types of selection bias as a result of factors associated with admission to the hospital (*see* **Case–Control Study, Hospital-based**). For example, **Berkson's fallacy** (bias) [3] results when an individual with two or more medical conditions is more likely to be hospitalized than someone with only one of the conditions. Thus, in a study utilizing hospital cases, there could be an apparent **association** between two conditions that does not exist in the general population. More generally, exposed cases may have a different chance of entering the hospital than nonexposed cases. Likewise, exposed and unexposed participants with control diseases may have different chances of hospital admission. These selection effects can bias the estimates of exposure effect [23]. Therefore, the choice of an appropriate control group can be difficult in a hospital-based case–control study. If the conditions for which controls have been hospitalized are associated with the exposure under study, then bias will result (*see* **Bias in Case–Control Studies**).

Certain types of selection bias may operate in either cohort or case–control studies. Among these is self-selection (or volunteer) bias, whereby subjects self-refer for participation in a study. Especially if the study hypothesis has been publicized, people who volunteer to become involved may differ in important ways from the group of all potentially eligible participants [5, 20]. Self-selection can also occur prior to the initiation of a study. For example, it has been observed that active workers experience lower mortality than the general population, presumably because one must maintain a certain degree of health in order to remain a part of the workforce [8, 20]. Studies utilizing workers as subjects must therefore account for this "healthy worker effect" by choosing an appropriate comparison group to avoid the risk of drawing invalid conclusions (*see* **Occupational Epidemiology**).

A final illustration of selection bias involves the use of prevalence data to draw conclusions about incidence, for example in a **cross-sectional study** or a case–control study employing prevalent cases (*see* **Case–Control Study, Prevalent**). The problem, commonly referred to as selective survival, arises when persons with the disease of interest are unavailable to participate in a study because they have died prior to the study's initiation. If exposure status happens to be over- or underrepresented in the survivors, then the use of prevalence data to estimate incidence-based effect estimates can lead to biased results (*see* **Bias from Survival in Prevalent Case–Control Studies**).

Addressing Selection Bias

Efforts to avoid or minimize selection bias should be emphasized over attempts to correct for it in the analysis stage. As implied above, one of the most important ways of achieving this goal is through the careful choice of an appropriate comparison group, i.e. the controls should be representative of the population from which the cases emerged. It has been recommended that researchers conducting case–control studies use two or more control groups, as a means of drawing some conclusions about the likelihood of selection bias [16, 17]. If the effect estimate remains the same regardless of which control group is utilized, then this offers some degree of reassurance that selection bias has been avoided (although the possibility remains that all estimates are equally biased). If the estimated effects differ, then one is left with the decision of which control group is the most suitable. Other strategies for minimizing the potential for selection bias include efforts to achieve high response and follow-up rates and to assure equal opportunity for disease detection among exposed and unexposed subjects. Case–control studies using incident cases (including **nested case–control studies**) are preferable to those using prevalent cases or to hospital-based studies.

The degree to which selection bias can be corrected after the collection of data depends on whether reliable estimates of the underlying selection or loss probabilities can be determined [14]. Since these probabilities are rarely known with accuracy, a suggested strategy is to consider a range of values of these parameters to assess the magnitude and direction of the bias that may be operating in a given study.

Information Bias

Information bias refers to a distortion in effect estimation that occurs when measurement of either the exposure or the disease is systematically inaccurate. This presentation focuses on "**misclassification** bias", the term used when such **measurement errors** result in incorrect classification of the exposure and/or disease status of study participants. It is useful to distinguish two types of misclassification – nondifferential and differential – since the distinction has implications for the direction and overall impact of the bias. If the misclassification is **nondifferential**, the errors made in classifying subjects along one axis (i.e. exposure or disease) are independent of their status with regard to the other axis. Differential misclassification occurs when such classification errors along one axis are not independent of the other axis [20]. For example, if a certain proportion of exposed subjects are mistakenly designated as unexposed, but the probability of misclassifying exposure is the same among diseased and nondiseased, then the result is nondifferential misclassification of exposure. If a certain proportion of diseased subjects are mistakenly designated as nondiseased, and the proportion misclassified varies by exposure status, then this represents differential misclassification of disease. In the case of the simple 2×2 table, nondifferential misclassification leads consistently to an underestimation of the magnitude of association between exposure and disease. Since the biased estimate is in the direction of no exposure–disease association, this phenomenon is termed **bias toward the null**. The situation becomes more complex when polytomous rather than dichotomous exposure variables are employed. In this circumstance, it is possible for nondifferential misclassification to result in bias away from the null [6]. Bias from differential misclassification can be in any direction. Therefore, depending upon the situation, differential misclassification can result in an underestimation (bias toward the null) or an overestimation (bias away from the null) of the magnitude of an association. The biased and unbiased effect estimates can even be on opposite sides of the null value ("crossover bias").

It should be borne in mind that misclassification of both exposure and disease can occur in the same study and that errors can be made simultaneously in both directions (e.g. some truly diseased subjects are mistakenly classified as nondiseased while some truly nondiseased are mistakenly classified as diseased). Misclassification probabilities are often expressed in terms of **sensitivity** and **specificity**, terms that are more frequently used in discussions of **screening** or **diagnostic test accuracy** [9, 14].

Sources of Information (Misclassification) Bias

Classification errors can occur in any type of study and may be due to imprecise measurement (of exposure and/or disease), mistaken or missed diagnoses, conscious or unconscious inaccuracies in self-reported information, or any other factor that causes a subject to be placed into the wrong cell of the 2×2 table (Table 1). For example, if subjects are followed over time for the occurrence of a disease, some may develop subclinical disease which goes unrecognized by the investigators. Such subjects would be misclassified as nondiseased. If exposed subjects are under greater scrutiny than the unexposed, then they may be less likely to have undiagnosed subclinical disease. This implies that **detection bias**, described above as a type of selection bias for case–control studies, can lead to differential misclassification in a follow-up study. With an inaccurate diagnostic tool, overascertainment of cases is also possible, and could be either differential or nondifferential with respect to exposure status.

Another potential source of misclassification has to do with the quality of information provided by study subjects. It can probably be assumed that some degree of misclassification is inevitable when subjects are asked to report exposures or to provide other aspects of their medical histories. **Recall bias**, which induces differential misclassification, occurs when the accuracy of self-reported information varies across comparison groups [1]. For example, people who have recently been diagnosed with an illness may be seeking an explanation and therefore could be more motivated to recall past exposures than are those unaffected by the illness. This could lead to an underestimate of exposure prevalence among controls compared to cases, causing the **odds ratio** to be artificially inflated. Recall bias is a potential problem in any case–control study in which subjects are asked to recall previous exposures. Recall bias would be unlikely to affect cohort studies because exposure information is usually based on exposure status at baseline. Misclassification is also

likely if it becomes necessary to gather data from surrogate respondents, depending upon the nature of the information requested and the level of detail required. Another potential source of inaccuracy has been termed "social desirability bias", which results from subjects' natural reluctance to report exposures and/or behaviors that are deemed socially unacceptable [24].

Those who are conducting the study can also have an impact on the degree and type of misclassification that occur during data collection. Knowledge of the study hypothesis and the comparison group to which a subject belongs could influence the behavior of personnel responsible for conducting interviews. For example, in a case–control study of the potential association between oral contraceptive use and development of venous thrombosis, subjects known to have experienced a venous thrombosis might be probed more deeply than controls for a history of oral contraceptive use. A similar situation might arise in a cohort study if an interviewer were more thorough in his or her questioning of exposed vs. unexposed subjects regarding signs and symptoms indicative of the outcome of interest. Such **interviewer bias** is a potential source of differential misclassification and could thus lead to invalid study results.

Addressing Information (Misclassification) Bias

Although there is always likely to be some degree of inaccuracy in measuring both exposures and outcomes, steps can be taken to minimize classification errors and reduce the probability that such errors will be differential (see below). Since nondifferential misclassification is more predictable in its impact and tends to result in underestimation of effects, it has generally been considered to be less of a threat than differential misclassification [20]. It should be noted that if a study finds a significant relationship between an exposure and an outcome, it is illogical to dismiss the study results on the grounds that nondifferential misclassification is present, since the effect estimate obtained in the absence of such misclassification would only be stronger.

One strategy for addressing misclassification bias is to ensure that all study participants are subject to the same follow-up procedures and standardized diagnostic criteria. To the extent possible, both the study personnel responsible for data collection and the study subjects should be blinded as to the main hypothesis under investigation (*see* **Blinding or Masking**). Interviewers must follow standardized protocols and, if practical, should remain unaware of the comparison group (i.e. case/control or exposed/unexposed) to which study participants from whom they gather information belong (*see* **Interviewing Techniques**). Acceptance of surrogate responses may not be appropriate for information that is subjective, highly personal, time-specific, or otherwise difficult to obtain from someone other than the subject him- or herself.

After data collection, correction for misclassification depends on the availability of information on probabilities of misclassification (e.g. sensitivity and specificity estimates) for variables that have been misclassified. Several authors have offered correction procedures: Barron [2] and Copeland et al. [4] provide correction formulae for 2×2 tables that assume nondifferential misclassification. These formulas were extended to allow for differential misclassification by Kleinbaum et al. [14] and to arbitrary multiway cross-classifications (*see* **Contingency Table**) by Korn [15]. Greenland & Kleinbaum [10] provide correction formulae for matched data. Espeland & Hui [7] use **loglinear models** and **maximum likelihood** estimation to incorporate estimates of nondifferential misclassification probabilities gathered either by resampling the study population, sampling a separate population, or a priori assumption. Reade-Christopher & Kupper [19] use **logistic regression** to correct for nondifferential misclassification of exposure with a priori assumptions about misclassification probabilities. Also, recent work by Satten & Kupper [22] provides odds ratio regression methods when we have available only the probability of exposure (POE) for each study subject, and where these POE values are assumed to be known without error.

Correction for misclassification bias often requires information from a **validation study**, which may not be available. If such data are not available, then the best approach to evaluating the results of a study is to assess the probable magnitude and direction of suspected misclassification errors and discuss the likely impact of such errors on the estimated effect. Formulas used for correction for misclassification are useful for this purpose.

Confounding

Confounding is a type of bias that occurs when the effect of the exposure of interest is mixed up with that of one or more "extraneous" variables. This can result in an observed exposure–disease relationship being attributed exclusively to the exposure of interest, when in reality the relationship is due, either wholly or in part, to the effect of another variable or variables (i.e. **confounders**). Confounding can also create the appearance of an exposure–disease relationship when in fact none exists. The amount of bias introduced can be large or small, depending upon the nature of the relationships among the confounder(s), the exposure, and the disease. Confounding can lead to an over- or underestimation of the true effect and can also result in an estimated effect that is on the opposite side of the null from its true value.

Sources of Confounding

There is essentially only one source of confounding – the presence of certain key relationships between an extraneous variable and both the exposure and the disease. The first requirement is that the extraneous variable be a risk factor for the disease, i.e. that it is either causally related to the disease or is a correlate of a causal factor [14, 20] (*see* **Causation**). More specifically, the status of a confounder as a risk factor for the disease must be independent of its association with the exposure of interest; therefore, it must be a risk factor among the unexposed. The second criterion is that the confounder be associated with the exposure of interest. Theoretically, this relationship should hold in the source population that produces the cases of disease [18]. Practically speaking, it is generally assessed among all subjects in a follow-up study and among the controls in a case–control study. A final criterion is that the confounder should not be an "intervening variable" in the causal pathway between exposure and disease. In other words, if an extraneous variable were actually a measure of some type of biological alteration caused by the exposure, which in turn went on to cause the disease, then such a variable would not fulfill the criteria for a confounder but would simply be a mediator between exposure and disease.

In addition to the theoretical considerations described above, there is also a data-based criterion for assessing confounding. The crude measure is considered to be confounded "in the data" if it differs meaningfully in value from the "adjusted" estimate of effect that removes the influence of the extraneous variables being assessed as possible confounders. This "data-based" criterion is also referred to as the "**collapsibility**" criterion [11]. The absence of data-based confounding implies that the strata considered when controlling for a potential confounder can be collapsed (or pooled) without introducing bias.

Addressing Confounding

Although applicable only to experimental rather than observational studies, one technique for decreasing the likelihood of confounding is **randomization**. If the exposure of interest is allocated randomly to study subjects, then the probability that the exposure will be associated with a potential confounder is greatly reduced (*see* **Randomized Treatment Assignment**). Importantly, this benefit is gained for previously unidentified (and unobserved) potential confounders as well as for those suspected potential confounders that are measured. Restriction is another means of avoiding an unwanted relationship between the exposure and an extraneous variable. For example, if gender is a risk factor for the outcome of interest, and there is concern that gender will be unequally distributed between exposure groups, then restricting the study to either males or females will eliminate the possibility of confounding by this variable. However, the advantage of this approach must be weighed against the potential threat to generalizability (external validity, see below) that may result. Another strategy for addressing confounding is the practice of **matching**, whereby the comparison group is selected to be similar to the index group with regard to key variables that are suspected confounders. Although confounding can be controlled in unmatched designs, the primary statistical advantage of matching is that it can make control of confounding more efficient (i.e. increase the precision of an adjusted estimate) [14, 20].

In contrast to the other types of bias, viable options also exist for addressing confounding in the analysis phase of a study. If only one potential confounder is identified, then the simplest approach is to perform a stratified analysis (*see* **Stratification**), calculating a separate effect estimate for subjects in each category of the potential confounder (e.g. separate estimates for males and females, smokers and

nonsmokers, etc.). If the estimates are similar across categories (i.e. there is no **interaction**), then they can be combined into a single adjusted estimate which removes the influence of the potential confounder (*see* **Mantel–Haenszel Methods**; **Matched Analysis**). When several potential confounders are identified, the assessment of confounding is complicated, requiring simultaneous control of several variables as well as determination of which subset of potential confounders is most appropriate for simultaneous control [14]. The "gold standard" (i.e. most valid subset) to which all subsets should be compared contains the entire set of potential confounders. The use of stratified analysis is not a viable option when there is even a moderate number of potential confounders, since the number of strata becomes too large and individual strata may contain few or no subjects. An alternative approach uses a mathematical modeling procedure (e.g. **logistic regression**), and requires determining whether the estimated measure of effect changes meaningfully when potential confounders are deleted from the model [13, 14]. The term "meaningfully" implies a decision that does not involve statistical testing, but rather the consideration of biologic and/or clinical experience about the importance of a change in the effect measure. Variables identified as nonconfounders from this approach may be dropped from the model provided their deletion leads to a gain in precision (i.e. narrower **confidence interval**). In the absence of interaction, the assessment of confounding simplifies to monitoring changes in the estimated coefficient of the exposure variable. However, if there is interaction, then the assessment is more subjective because the collective change in several coefficients must be monitored. Consequently, when interaction is present, we recommend keeping all potential confounders in the model. Note, furthermore, that whether stratified analysis or mathematical modeling is used, misclassification of confounders can lead to incomplete or incorrect control of confounding [9]. Also, the use of regression modeling for control of confounding may yield misleading results if incorrect assumptions are made about the form and characteristics of the model being fit. Important characteristics to consider include the specific variables (e.g. "risk factors") chosen for control, the quantitative form that such variables should take in the model, the interaction effects to be evaluated, and the measurement scale (e.g. **additive** or **multiplicative**) used to assess interaction effects (*see* **Effect Modification**; **Relative Risk Modeling**).

Internal vs. External Validity

The discussion to this point has centered upon the issue of internal validity, or lack of bias. One may also be interested in assessing the external validity of an estimated effect (*see* **Validity and Generalizability in Epidemiologic Studies**). The distinction between internal and external validity has to do with the population about which inferences are to be made [14]. An internally valid effect estimate is one that correctly describes the association between exposure and outcome in the **target population**, i.e. the collection of individuals upon which the study was designed to focus and about whom one is able to draw direct conclusions. This is the group that has been sampled, though not necessarily in a random fashion. By contrast, external validity, or generalizability, refers to the making of inferences to an external population beyond the restricted interest of the particular study from which the effect is estimated. Assessing external validity involves making a judgment regarding whether the results of a study can be extended to apply to individuals who are dissimilar in some aspects (e.g. age, race, occupation) compared with those upon whom the study was focused.

References

[1] Austin, H., Hill, H.A., Flanders, W.D. & Greenberg, R.S. (1994). Limitations in the application of case-control methodology, *Epidemiologic Reviews* **16**, 65–76.

[2] Barron, B.A. (1977). The effects of misclassification on the estimation of relative risk, *Biometrics* **33**, 414–418.

[3] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data, *Biometrics Bulletin* **2**, 47–53.

[4] Copeland, K.T., Checkoway, H., McMichael, A.J., & Holbrook, R.H. (1977). Bias due to misclassification in the estimation of relative risk, *American Journal of Epidemiology* **105**, 488–495.

[5] Criqui, M.H., Austin, M. & Barrett-Connor, E. (1979). The effect of non-response on risk ratios in a cardiovascular disease study, *Journal of Chronic Diseases* **32**, 633–638.

[6] Dosemeci, M., Wacholder, S. & Lubin, J.H. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value?, *American Journal of Epidemiology* **132**, 746–748.

[7] Espeland, M.A. & Hui, S.L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors, *Biometrics* **43**, 1001–1012.

[8] Fox, A.J. & Collier, P.F. (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry, *British Journal of Preventive and Social Medicine* **30**, 225–230.

[9] Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564–569.

[10] Greenland, S. & Kleinbaum, D.G. (1983). Correcting for misclassification in two-way tables and matched-pair studies, *International Journal of Epidemiology* **12**, 93–97.

[11] Greenland, S. & Robins, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding, *International Journal of Epidemiology* **15**, 412–418.

[12] Horwitz, R.I. & Feinstein, A.R. (1978). Alternate analytic methods for case-control studies of estrogens and endometrial cancer, *New England Journal of Medicine* **299**, 1089–1094.

[13] Kleinbaum, D.G. (1994). *Logistic Regression: A Self-Learning Text*. Springer-Verlag, New York.

[14] Kleinbaum, D.G., Kupper, L.L. & Morganstern, H. (1982). *Epidemiologic Research*. Van Nostrand Reinhold, New York.

[15] Korn, E.L. (1981). Hierarchical log-linear models not preserved by classification error, *Journal of the American Statistical Association* **76**, 110–112.

[16] Lilienfeld, A.M. & Lilienfeld, D.E. (1980). *Foundations of Epidemiology*. Oxford University Press, New York.

[17] MacMahon B. & Pugh, T.F. (1970). *Epidemiology – Principles and Methods*. Little, Brown & Company, Boston.

[18] Miettinen, O.S. & Cook, E.F. (1981). Confounding: Essence and detection, *American Journal of Epidemiology* **114**, 593–603.

[19] Reade-Christopher, S.J. & Kupper, L.L. (1991). Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data, *Biometrics* **47**, 535–548.

[20] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown & Company, Boston.

[21] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[22] Satten, G.A. & Kupper, L.L. (1993). Inferences about exposure-disease associations using probability-of-exposure information, *Journal of the American Statistical Association* **88**, 200–208.

[23] Schlesselman, J.J. (1982). *Case-Control Studies*. Oxford University Press, New York.

[24] Wynder, E.L. (1994). Investigator bias and interviewer bias: The problem of reporting systematic error in epidemiology, *Journal of Clinical Epidemiology* **47**, 825–827.

HOLLY A. HILL & DAVID G. KLEINBAUM

Bias Toward the Null

Bias toward the null usually refers to an effect of **nondifferential errors** in exposure measurements that reduces the apparent effect of the exposure on the dependent variable, which might be a health outcome. For **linear regressions**, bias toward the null is called attenuation (*see* **Measurement Error in Epidemiologic Studies**). Attenuation (or bias toward the null) does not change the sign of the coefficient of exposure in the regression, but it reduces the absolute magnitude toward zero. Similar effects are found for estimates of log relative odds in **logistic regression** and for **odds ratios** in **2 × 2 tables** (*see* **Bias in Observational Studies**; **Misclassification Error**). The odds ratios are biased toward, but not beyond, unity. There are, however, more complex situations in which nondifferential error can induce a bias away from the null and thus exaggerate an apparent exposure effect or induce a bias that reverses the direction of an apparent exposure effect, as mentioned in the articles cited above.

Bias toward the null can also result from other sources of bias, such as **confounding** and **selection bias**.

MITCHELL H. GAIL

Bias, Nondifferential

When comparing exposed with unexposed groups, **unbiased** estimates of **exposure effect** may result if the same (nondifferential) **biases** affect each exposure group. Nondifferential bias is bias that affects each exposure (or treatment) group in such a way that the resulting exposure effect measure remains unbiased. For example, suppose that 10% of the exposed group and 5% of the unexposed group develop cancer in a given time period, corresponding to a true **relative risk** of $10\%/5\% = 2.0$. Suppose, however, that follow-up procedures fail to detect 20% of incident cancers in each group, resulting in apparent cancer risks of 8% and 4%, respectively. Despite the fact that each of these **risks** is biased, the relative risk, $8\%/4\% = 2.0$, is unbiased. Thus, with respect to relative risk, these biases are nondifferential. If, instead, the chosen measure of exposure effect was the risk difference, $10\% - 5\% = 5\%$, these same errors would yield a biased estimate of $8\% - 4\% = 4\%$. Thus, an error process may induce nondifferential bias with respect to one effect measure but differential bias with respect to another measure.

Nondifferential bias results from a **nondifferential error** process. In the previous example, the underestimates, 8% and 4%, depended only on the corresponding true values, 10% and 5%, respectively, and not on the exposure group, because the error process missed 20% of incident cancers, regardless of exposure group. Thus, the error was nondifferential.

(*See also* **Misclassification Error**)

MITCHELL H. GAIL

Bias, Overview

Bias is defined as the "deviation of results or inferences from the truth, or processes leading to such deviation" [12]. In other words, it is the extent to which the expected value of an estimator differs from a population parameter. Bias refers to **systematic errors** that decrease the validity of estimates, and does not refer to **random errors** that decrease the precision of estimates. Unlike random error, bias cannot be eliminated or reduced by an increase in sample size.

Bias can occur as a result of flaws in the following stages of research [17]:

1. literature review,
2. study design,
3. study execution,
4. data collection,
5. analysis,
6. interpretation of results, and
7. publication.

Literature Review Bias

Literature review bias (syn. reading-up bias) refers to errors in reading-up on the field [17]. Examples include:

Foreign language exclusion bias: literature reviews and **meta-analyses** that ignore publications in foreign languages [5].

Literature search bias: caused by lack of a computerized literature search, incomplete search due to poor choice of keywords and search strategies, or failure to include unpublished reports or hard-to-reach journals through interlibrary loans.

One-sided reference bias: investigators may restrict their references to only those studies that support their position [17].

Rhetoric bias: authors may use the art of writing to convince the reader without appealing to scientific fact or reason [17].

Design Bias

Design bias refers to errors occurring as a result of faulty design of a study [12]. This can arise from faulty selection of subjects, noncomparable groups chosen for comparison, or inappropriate sample size.

Selection Bias

Selection bias is a distortion in the estimate of effect resulting from the manner in which subjects are selected for the study population. Bias in selection can arise: (i) if the **sampling frame** is defective, (ii) if the sampling process is nonrandom, or (iii) if some sections of the **target population** are excluded (noncoverage bias) [14].

Sampling frame bias. This type of bias arises when the sampling frame that serves as the basis for selection does not cover the population adequately, completely, or accurately [14]. Examples include:

Ascertainment bias: arising from the kind of patients (e.g. slightly ill, moderately ill, acutely ill) that the individual observer is seeing, or from the diagnostic process which may be determined by the culture, customs, or individual disposition of the health care provider [12]. (See also diagnostic access bias.)

Berkson bias (*see* **Berkson's Fallacy**) (syn. admission rate bias, hospital admission bias): caused by selective factors that lead hospital cases and controls in a case–control study to be systematically different from one another [1, 6].

Centripetal bias: the reputations of certain clinicians and institutions cause individuals with specific disorders or exposures to gravitate toward them [17].

Diagnostic access bias: patients may not be identified because they have no access to diagnostic process due to culture or other reasons. (See also ascertainment bias, hospital access bias.)

Diagnostic purity bias: when "pure" diagnostic groups exclude comorbidity, they may become nonrepresentative [17].

Hospital access bias: patients may not be identified because they are not sick enough to require hospital care, or because they are excluded from hospitals as a result of distance or cost considerations. (See also ascertainment bias, diagnostic access bias, referral filter bias.)

Migrator bias: migrants may differ systematically from those who stay home [17].

Neyman bias (syn. attrition bias, prevalence–incidence bias, selective survival bias; *see* **Bias from Survival in Prevalent Case–Control Studies**):

caused by excluding those who die before the study starts because the exposure increases mortality risk [4, 6].

Telephone sampling bias: if **telephone sampling** is used to select a sample of individuals, then persons living in households without telephones would be systematically excluded from the study population, although they would be included in the target population.

Nonrandom sampling bias. This type of bias arises if the sampling is done by a nonrandom method, so that the selection is consciously or unconsciously influenced by human choice [14]. Examples include:

Autopsy series bias: resulting from the fact that autopsies represent a nonrandom sample of all deaths [12].

Detection bias (syn. selective surveillance bias, verification bias): caused by errors in methods of ascertainment, diagnosis, or verification of cases in an epidemiologic investigation, for example verification of diagnosis by laboratory tests in hospital cases, but not in cases outside the hospital [6, 12]. (See also diagnostic work-up bias, unmasking bias.)

Diagnostic work-up bias (syn. sequential-ordering bias): arises if the results of a diagnostic or screening test affect the decision to order the "**gold standard**" procedure that provides the most definitive result about the disease [16], for example those who have a negative screening test are systematically excluded from the gold standard procedure [3]. (See also detection bias, unmasking bias.)

Door-to-door solicitation bias: subjects obtained by door knocking are more likely to be the elderly, unemployed, and less active individuals who tend to stay at home.

Previous opinion bias: the tactics and results of a previous diagnostic process on a patient, if known, may affect the tactics and results of a subsequent diagnostic process on the same patient [17]. (See also diagnostic work-up bias.)

Referral filter bias: as a group of patients are referred from primary to secondary to tertiary care, the concentration of rare causes, multiple diagnoses, and severe cases may increase [17]. (See also hospital access bias.)

Sampling bias: caused by the use of nonprobability sampling methods that do not ensure that all members of the population have a known chance of selection in the sample [12] (*see* **Quota, Representative, and Other Methods of Purposive Sampling**).

Self-selection bias (syn. self-referral bias): subjects contact the investigators on their own initiative in response to publicity about the investigation.

Unmasking bias (syn. signal detection bias): an innocent exposure may become suspect if, rather than causing a disease, it causes a sign or symptom which leads to a search for the disease [17] (*see* **Bias From Diagnostic Suspicion in Case–Control Studies**). (See also detection bias, diagnostic work-up bias.)

Noncoverage bias. This type of bias arises if some sections of the population are impossible to find or refuse to cooperate [14]. Examples include:

Early-comer bias (syn. latecomer bias): "early-comers" from a specified sample may exhibit exposures or outcomes which differ from those of "latecomers" [6], for example early-comers in a study tend to be healthier, and less likely to smoke [17]. (See also response bias.)

Illegal immigrant bias: when census data are used to calculate death rates, bias is caused by illegal immigrants who appear in the numerator (based on death records) but not in the denominator (based on census data).

Loss to follow-up bias: caused by differences in characteristics between those subjects who remain in a cohort study and those who are lost to follow-up [6] (*see* **Bias from Loss to Follow-up**).

Response bias (syn. nonrespondent bias, volunteer bias): caused by differences in characteristics between those who choose or volunteer to participate in a study and those who do not [7, 12] (*see* **Bias from Nonresponse**). An example is the forecast of the US presidential election in a 1936 survey of 10 million individuals that went wrong because the response rate was only 20%, and the respondents presumably came from a higher social class than the general electorate [14]. (See also early-comer bias.)

Withdrawal bias: caused by differences in the characteristics of those subjects who choose to withdraw and those who choose to remain [6, 12].

Noncomparability Bias

Noncomparability bias occurs if the groups chosen for comparison are not comparable. Examples include:

Ecological bias (syn. **ecologic fallacy**): the associations observed between variables at the group level on the basis of ecological data may not be the same as the associations that exist at the individual level.

Healthy Worker Effect (HWE): an observed decrease in mortality in workers when compared with the general population [4] (*see* **Occupational Epidemiology**). This is a type of membership bias [6].

Lead-time bias (syn. zero time shift bias): occurs when follow-up of two groups does not begin at strictly comparable times, for example when one group has been diagnosed earlier in the natural history of the disease than the other group owing to the use of a screening procedure [12] (*see* **Screening Benefit, Evaluation of**).

Length bias: caused by the selection of a disproportionate number of long-duration cases (cases who survive longest) in one group and not in the other. An example is when prevalent cases, rather than incident cases, are included in a case–control study [12].

Membership bias: membership in a group (e.g. workers, joggers) may imply a degree of health which differs systematically from that of the general population because the general population is composed of both healthy and ill individuals [6, 17].

Mimicry bias: an innocent exposure may become suspect if, rather than causing a disease, it causes a benign disorder which resembles the disease [17].

Nonsimultaneous comparison bias (syn. noncontemporaneous control bias): secular changes in definitions, exposures, diagnoses, diseases, and treatments may render noncontemporaneous controls noncomparable [17], for example use of historical controls [12] (*see* **Bias from Historical Controls**).

Sample Size Bias

Samples that are too small may not show effects even when they are present; samples that are too large may show tiny effects of little or no practical significance [17]. Another name for sample size bias is wrong sample size bias.

Study Execution Bias

Study execution bias refers to errors in executing the experimental maneuver (or exposure) [17]. Examples include:

Bogus control bias: when patients who are allocated to an experimental maneuver die or sicken before or during its administration and are omitted or reallocated to the control group, the experimental maneuver will appear spuriously superior [17].

Contamination bias: when members of the control group in an experiment inadvertently receive the experimental maneuver, the differences in outcomes between experimental and control patients may be systematically reduced [17] (*see* **Bias Toward the Null**).

Compliance bias: in experiments requiring patient adherence to therapy, issues of efficacy become **confounded** with those of compliance, for example when high-risk coronary patients quit exercise programs [17] (*see* **Noncompliance, Adjustment for**).

Data Collection Bias

Data collection bias (syn. information bias, **measurement error**, **misclassification** bias, observational bias) refers to a flaw in measuring exposure or outcome that results in differential quality or accuracy of information between compared groups [12] (*see* **Bias, Nondifferential**). Bias in data collection can arise from (i) defective measuring instruments, (ii) wrong data source, (iii) errors of the observer, (iv) errors of the subjects, and (v) errors during data handling.

Instrument Bias

Instrument bias (syn: instrument error) refers to defects in the measuring instruments [17]. This may be due to faulty calibration, inaccurate measuring instruments, contaminated reagents, incorrect dilution or mixing of reagents, etc. [12]. Examples include:

Case definition bias: definition of cases, for example based on different versions of **International Classification of Diseases** (ICD) codes, or first-ever cases vs. recurrent cases, may change over time or across regions, resulting in inaccurate trends and geographic comparisons [13]. (See also diagnostic vogue bias.)

Diagnostic vogue bias: the same illness may receive different diagnostic labels at different points in space or time, for example the British term "bronchitis" vs. North American "emphysema") [17]. (See also case definition bias.)

Forced choice bias: questions that provide inadequate choices, for example only "yes" and "no", and without other choices like "do not know" or "yes but

do not know type", may force respondents to choose from the limited choices. (See also scale format bias.)

Framing bias: preference depends on the manner in which the choices are presented, for example telling a prospective candidate for surgery that an operation has a 5% mortality, vs. 95% survival rate.

Insensitive measure bias: when **outcome measures** are incapable of detecting clinically significant changes or differences, type II errors occur [17].

Juxtaposed scale bias (syn. questionnaire format bias): juxtaposed scales, a type of self-report response scale which asks respondents to give multiple responses to one item, may elicit different responses than when separate scales are used [10].

Laboratory data bias: data based on laboratory test results are subject to errors of the laboratory test including faulty calibration of the instruments, contaminated or incorrect amounts of reagents, etc.

Questionnaire bias: leading questions or other flaws in the questionnaire may result in a differential quality of information between compared groups [6] (*see* **Questionnaire Design**).

Scale format bias: even vs. odd number of categories in the scale for the respondents to choose from can produce different results, for example (Agree) 1–2–3 (Disagree) tends to obtain neutral answers, i.e. 2, while (Agree) 1–2–3–4 (Disagree) tends to force respondents to take sides. (See also forced choice bias.)

Sensitive question bias: sensitive questions such as personal or household incomes, sexual orientation, or marital status, may induce inaccurate answers.

Stage bias: method for determining stage of disease of patients may vary across the groups being compared, across geographic areas, or through time, leading to spurious comparison of stage-adjusted survival rates (*see* **Bias from Stage Migration in Cancer Survival**) [9].

Unacceptability bias: measurements which hurt, embarrass or invade privacy may be systematically refused or evaded [17].

Underlying/contributing cause of death bias: results of data analysis will be different depending on whether the underlying or the contributing cause of death as recorded on the death certificates is used (*see* **Cause of Death, Underlying and Multiple**; **Death Certification**).

Voluntary reporting bias: voluntary reporting system vs. mandatory reporting system can generate differences in the quality and completeness of routine data.

Data Source Bias

Data source bias refers to wrong, inadequate, or impossible source or type of data. Examples include:

Competing death bias: some causes of death (e.g. cancers) are associated with older age, while others (e.g. infectious diseases) are associated with younger age. Therefore in places where infectious diseases are prevalent, the cancer rates will be underestimated owing to competing causes of death from infectious diseases (*see* **Competing Risks**).

Family history bias: positive family history is not an accurate indicator of familial aggregation of a disease and the influence of genetic factors, because it is a function of the number of relatives and the age distribution of relatives [11].

Hospital discharge bias: hospital discharge data do not reflect hospital admission data since they are affected by length of hospital stay, and therefore do not provide accurate information for disease incidence.

Spatial bias: many environmental data used for health applications, for example geographic information systems (GIS), derive from point measurements at monitoring or survey stations. Unfortunately, many environmental monitoring networks are too sparse spatially and biased towards high pollution sites, generating an inaccurate pollution surface [2].

Observer Bias

Observer bias is due to differences among observers (interobserver variation) or to variations in readings by the same observer on separate occasions (intraobserver variation) [12] (*see* **Observer Reliability and Agreement**). Examples include:

Diagnostic suspicion bias (syn. diagnostic bias): a knowledge of the subject's prior exposure to a putative cause (e.g. ethnicity, drug use, cigarette smoking) may influence both the intensity and the outcome of the diagnostic process [6, 17] (*see* **Bias From Diagnostic Suspicion in Case–Control Studies**).

Exposure suspicion bias: a knowledge of the subject's disease status may influence both the intensity and outcome of a search for exposure to the putative cause [6, 17] (*see* **Bias from Exposure Suspicion in Case–Control Studies**).

Expectation bias: observers may systematically err in measuring and recording observations so that they concur with prior expectations, for example house officers tend to report "normal" fetal heart rates [17].

Interviewer bias: caused by interviewers' subconscious or even conscious gathering of selective data [12], for example questions about specific exposures may be asked several times of cases but only once of controls [17]. Can result from interinterviewer or intrainterviewer errors [6].

Therapeutic personality bias: when treatment is not blind, the therapist's convictions about efficacy may systematically influence both outcomes and their measurement (e.g. desire for positive results) [17] (*see* **Blinding or Masking**).

Subject Bias

Subject bias (syn. "observee" bias) refers to the inaccuracy of the data provided by the subjects (respondents, "observees") at the time of data collection. Examples include:

Apprehension bias: certain measures (e.g. pulse, blood pressure) may alter systematically from their usual levels when the subject is apprehensive (e.g. blood pressure may change during medical interviews) [17].

Attention bias (syn. Hawthorne effect): study subjects may systematically alter their behavior when they know they are being observed [17].

Culture bias: subjects' responses may differ because of culture differences, for example some **ethnic groups**, because of their cultural background, do not want to share publicly their pain or problems such as unemployment, marital troubles, youth crime, and parental difficulties.

End aversion bias: subjects usually avoid end of scales in their answers, try to be conservative, and wish to be in the middle.

Faking bad bias (syn. hello–goodbye effect): subjects try to appear sick in order to qualify for support. Also, subjects try to seem sick before, and very well after, the treatment.

Faking good bias (syn. social desirability bias): socially undesirable answers tend to be underreported. (See also unacceptable disease bias, unacceptable exposure bias.)

Family information bias: the family history and other historical information may vary markedly depending upon whether the individual in the family providing the information is a case or a control, for example different family histories of arthritis may be obtained from affected and unaffected siblings [17].

Interview setting bias: whether interviews are conducted at home, in a hospital, the respondent's workplace, or the researcher's office may affect subjects' responses.

Obsequiousness bias: subjects may systematically alter questionnaire responses in the direction they perceive desired by the investigator [17].

Positive satisfaction bias (syn. positive skew bias): subjects tend to give positive answers, typically when answering satisfaction questions.

Proxy respondent bias (syn. surrogate data bias): for deceased cases or surviving cases (e.g. brain tumors) whose ability to recall details is defective, soliciting information from proxies (e.g. spouse or family members) may result in differential data accuracy.

Recall bias: caused by differences in accuracy or completeness of recall to memory of prior events or experiences [6], for example mothers whose children have had leukemia are more likely than mothers of healthy children to remember details of diagnostic X-ray examinations to which these children were exposed *in utero* [12].

Reporting bias (syn. self-report response bias): selective suppression or revealing of information such as past history of sexually transmitted disease [12]. (See also unacceptable disease bias, unacceptable exposure bias, sensitive question bias.)

Response fatigue bias: questionnaires that are too long can induce fatigue among respondents and result in uniform and inaccurate answers.

Unacceptable disease bias: socially unacceptable disorders (e.g. sexually transmitted diseases, suicide, mental illness) tend to be underreported [12]. (See also reporting bias, faking good bias.)

Unacceptable exposure bias: socially unacceptable exposures (e.g. smoking, drug abuse) tend to be underreported. (See also reporting bias, faking good bias.)

Underlying cause bias (syn. rumination bias): cases may ruminate about possible causes for their illness and thus exhibit different recall or prior exposures than controls [17]. (See also recall bias.)

Yes-saying bias: some subjects tend to say "yes" to all questions.

Data Handling Bias

Data handling bias refers to the manner in which data are handled. Examples include:

Data capture error: errors in the acquisition of the data in digital form, normally by manual encoding (coding error), digitizing (data entry error), scanning, or electronic transfer from pre-existing data bases [2]. (See also data entry bias.)

Data entry bias: difference in data entry practices may cause unreal observed differences in geographic variations in incidence rates [18]. (See also data capture error.)

Data merging error: incorrect merging of data from different databases, for example erroneous merging and failure to merge as a result of illegible handwriting on the routine forms, different dates of service recorded in different databases, etc. (See also record linkage bias.)

Digit preference bias (syn. end-digit preference bias): in converting analog to digital data, observers may record some terminal digits with an unusual frequency [17], for example rounding off may be to the nearest whole number, even number, multiple of 5 or 10, or, when time units like a week are involved, [8, 15], etc. [12].

Record linkage bias: computerized **record linkage** is based on a probabilistic process based on identifiers (*see* **Matching, Probabilistic**). Some identifiers, e.g. some surnames, may have a poor record linkage weight, causing linkage problems, and therefore tend to exclude subjects having those identifiers.

Analysis Bias

Analysis bias results from errors in analyzing the data. It can arise from (i) lack of adequate control of **confounding** factors, (ii) inappropriate analysis strategies, and (iii) *post hoc* analysis of the data set.

Confounding Bias

Confounding bias occurs when the estimate of the effect of the exposure of interest is distorted because it is mixed with the effect of a confounding (extraneous) factor. A confounding factor must be a risk factor for the disease, be associated with the exposure under study, and not be an intermediate step in the causal path between the exposure and the disease [6]. Examples include:

Latency bias: failure to adjust for the **latent period** in the analysis of cancer or other chronic disease data.

Multiple exposure bias: failure to adjust for multiple exposures.

Nonrandom sampling bias: when a study sample is selected by nonrandom (nonprobability) sampling, failure to account for variable sampling fractions in the analysis may introduce a bias, for example weighting by the strata population sizes is needed for a disproportionate stratified sample (*see* **Stratified Sampling**).

Standard population bias: choice of standard population will affect estimation of standardized rates (a weighted average of the category-specific rates) [7] (*see* **Standardization Methods**).

Spectrum bias (syn. case mix bias): heterogeneous groups of patients with different proportions of mild and severe cases can lead to different estimates of **screening** performance indicators [16].

Analysis Strategy Bias

Analysis strategy bias (syn. analysis method bias) refers to problems in the analysis strategies. Examples include:

Distribution assumption bias: wrong assumption of **sampling distribution** in the analysis, for example time variables follow **lognormal distribution** rather than **normal distribution**, and therefore geometric mean time rather than **mean** time should be used [8].

Enquiry unit bias: choice of unit of enquiry may affect analysis results, for example with the school as the unit of enquiry, half the high schools offered no physics, but when the student becomes the unit of enquiry, only 2% of all high school students attended schools that offered no physics, since the small schools do not teach physics (*see* **Unit of Analysis**).

Estimator bias: the difference between the expected value of an estimator of a parameter and the true value of this parameter [12], for example **odds ratio** always overestimates **relative risk**).

Missing data handling bias: how **missing data** are handled, for example treated as a missing case vs. interpreted as a "no" answer, will lead to different results.

Outlier handling bias: arising from a failure to discard an unusual value occurring in a small sample, or due to exclusion of unusual values that should be

included [12]. The latter is also called tidying-up bias (the exclusion of outliers or other untidy results which cannot be justified on statistical grounds) [17].

Overmatching bias: matching on a nonconfounding variable that is associated with the exposure but not the disease can lead to conservative estimates in a matched case–control study [6].

Scale degradation bias: the degradation and collapsing of measurement scales tend to obscure differences between groups under comparison [17].

Post Hoc *Analysis Bias*

Post hoc analysis bias refers to the misleading results caused by *post hoc* questions, data dredging, and subgroup analysis (*see* **Treatment-covariate Interaction**). Examples include:

Data dredging bias: when data are reviewed for all possible associations without prior hypothesis, the results are suitable for hypothesis-generating activities only [17].

Post hoc *significance bias*: when decision levels or "tails" for type I and type II errors are selected after the data have been examined, conclusions may be biased [17].

Repeated peeks bias: repeated peeks at accumulating data in a randomized trial are not independent, and may lead to inappropriate termination [17] (*see* **Sequential Analysis**).

Interpretation Bias

Interpretation bias arises from inference and speculation, for example failure of the investigator to consider every interpretation consistent with the facts and to assess the credentials of each, and mishandling of cases that constitute exceptions to some general conclusion [12]. Examples include:

Assumption bias (syn. conceptual bias): arising from faulty logic or premises or mistaken beliefs on the part of the investigator, for example having correctly deduced the mode of transmission of cholera, John Snow falsely concluded that yellow fever was transmitted by similar means [12].

Cognitive dissonance bias: the belief in a given mechanism may increase rather than decrease in the face of contradictory evidence [17].

Correlation bias: equating **correlation** with causation leads to errors of both kinds [17].

Generalization bias (syn. lack of external validity): generalizing study results to people outside the study population may produce bias, for example generalizing findings in men to women (*see* **Validity and Generalizability in Epidemiologic Studies**).

Magnitude bias: when interpreting a finding, the selection of a scale of measurement may markedly affect the interpretation, for example $1 000 000 may also be 0.0003% of the national budget [17].

Significance bias: the confusion of statistical significance, on the one hand, with biologic or clinical or health care significance, on the other hand, may lead to fruitless studies and useless conclusions [17] (*see* **Clinical Significance Versus Statistical Significance**).

Underexhaustion bias: the failure to exhaust the hypothesis space may lead to erroneous interpretations [17].

Publication Bias

Publication bias refers to an editorial predilection for publishing particular findings, e.g. positive results, which can distort the general belief about what has been demonstrated in a particular situation [12] (*see* **Meta-analysis of Clinical Trials**). Examples include:

All's well literature bias: scientific or professional societies may publish reports or editorials which omit or play down controversies or disparate results [17].

Positive results bias: authors are more likely to submit, and editors accept, positive than null results [17].

Hot topic bias (syn. hot stuff bias): when a topic is hot, investigators and editors are tempted to publish additional results, no matter how preliminary or shaky [17].

References

[1] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data, *Biometrics Bulletin* **2**, 47–53.

[2] Briggs, D.J. & Elliott, P. (1995). The use of geographical information systems in studies on environment and health, *World Health Statistics Quarterly* **48**, 85–94.

[3] Choi, B.C.K. (1992). Sensitivity and specificity of a single diagnostic test in the presence of work-up bias, *Journal of Clinical Epidemiology* **45**, 581–586.

[4] Choi, B.C.K. (2000). A technique to re-assess epidemiologic evidence in light of the healthy worker effect: the

case of firefighting and heart disease, *Journal of Occupational and Environmental Medicine* **42**, 1021–1034.

[5] Choi, B.C.K. (1996). Occupational cancer in developing countries, *American Journal of Epidemiology* **144**, 1089.

[6] Choi, B.C.K. & Noseworthy, A.L. (1992). Classification, direction, and prevention of bias in epidemiologic research, *Journal of Occupational Medicine* **34**, 265–271.

[7] Choi, B.C.K., de Guia, N.A. & Walsh, P. (1999).. Look before you leap: stratify before you standardize. *American Journal of Epidemiology* **149**, 1087–1096.

[8] Choi, B.C.K., Pak, A.W.P. & Purdham, J.T. (1990). Effects of mailing strategies on response rate, response time, and cost in a questionnaire study among nurses, *Epidemiology* **1**, 72–74.

[9] Farrow, D.C., Hunt, W.C. & Samet, J.M. (1995). Biased comparisons of lung cancer survival across geographic areas: effects of stage bias, *Epidemiology* **6**, 558–560.

[10] Hunt, D.M., Magruder, S. & Bolon, D.S. (1995). Questionnaire format bias: when are juxtaposed scales appropriate. A call for further research, *Psychological Reports* **77**, 931–941.

[11] Khoury, M.J. & Flanders, W.D. (1995). Bias in using family history as a risk factor in case-control studies of disease, *Epidemiology* **6**, 511–519.

[12] Last, J.M. (2000). *A Dictionary of Epidemiology*, 4th Ed. Oxford University Press, New York.

[13] May, D.S. & Kittner, S.J. (1994). Use of medicare claims data to estimate national trends in stroke incidence, 1985–1991, *Stroke* **25**, 2343–2347.

[14] Moser, C.A. & Kalton, G. (1971). *Survey Methods in Social Investigation*. Gower, Brookfield.

[15] Neyman, J. (1955). Statistics – servant of all sciences, *Science* **122**, 401.

[16] Ransohoff, D.F. & Feinstein, A.R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests, *New England Journal of Medicine* **299**, 926–930.

[17] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[18] Sarti, C. (1993). Geographic variation in the incidence of nonfatal stroke in Finland: are the observed differences real?, *Stroke* **24**, 787–791.

(*See also* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**)

BERNARD C.K. CHOI & ANITA W.P. PAK

Bias, Protopathic

Protopathic (or reverse-causality) **bias** is a consequence of a differential **misclassification** of exposure related to its timing of occurrence. It is observed when a change in exposure taking place in the time period following disease occurrence is incorrectly thought to precede disease occurrence. It can be observed for exposures that may change with time and for diseases for which the date of occurrence is difficult to determine accurately because of an insidious development of symptoms over a prolonged period of time. The term protopathic bias was coined by Horwitz and Feinstein [1]. Suppose that a long delay between early (or protopathic) symptoms and disease suspicion or diagnosis is commonly observed. Suppose also that many patients decrease or stop their exposure to a risk factor as a result of early disease symptoms. This will create a protopathic downward bias upon assessing the association between exposure and disease. For instance, early symptoms of chronic obstructive pulmonary disease (*e.g.*, dyspnea) can start a long time before disease occurrence is suspected or diagnosed. If smoking patients spontaneously reduce their smoking as a result of early disease symptoms, then the magnitude of the association between current smoking and risk of chronic obstructive pulmonary disease will appear smaller than it is (downward bias). Conversely, protopathic bias can correspond to an upward bias if exposure is started or increased during the time period ranging from the start of symptoms to disease diagnosis. For instance, patients may be prescribed a drug to alleviate early disease symptoms before the disease is actually diagnosed. This will yield an apparent association between that drug and disease risk (upward bias). Protopathic bias is a frequent concern in pharmacoepidemiology studies (*see* **Pharmacoepidemiology, Overview**; **Pharmacoepidemiology, Adverse and Beneficial Effects**) because drug prescription and consumption often vary over time and may change in response to early disease-related symptoms. However, it can be observed potentially in all fields of epidemiology. It is a more serious concern **in case-control** studies than in **cohort** studies because timings of exposure and actual disease start need to be ascertained retrospectively, which increases the likelihood of protopathic bias.

Reference

[1] Horwitz, R.I. & Feinstein, A.R. (1979). Methodologic standards and contradictory results in case-control research, *The American Journal of Medicine* **66**, 556–564.

(*See also* **Bias in Case–Control Studies**; **Bias, Overview**; **Bias in Cohort Studies**; **Bias in Observational Studies**)

JACQUES BENICHOU

Bias

Bias is the expected deviation of an estimate (*see* **Estimation**) from the true quantity to be estimated. If an estimator $\hat{\theta}$ of a parameter θ has **expectation** $\theta + b$, the quantity b is called the bias. If $\hat{\theta}$ converges to $\theta + b$ as the sample size increases, then $\hat{\theta}$ is said to have asymptotic bias b. Some biases result from the small sample properties of the estimator used and vanish asymptotically. Most biases that result from **systematic error**, however, such as **selection biases** or biases in measuring outcomes, persist as the sample size increases. Indeed, increasing the sample size does not eliminate such biases but only leads to more precise biased estimates.

(*See also* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**; **Bias, Overview**)

MITCHELL H. GAIL

Biased Sampling of Cohorts

The epidemiologic **cohort study** involves a sample of individuals followed over time. Individuals are monitored to ascertain the incidence of various endpoints such as the incidence of disease, infection, or death. The goal is to estimate the absolute **incidence rates** of the event or to identify **covariates** called risk factors that modify this risk. Individuals may also be monitored for changes in various markers of health status such as blood pressure measurements in prospective studies of cardiovascular disease, or CD4+ T cell counts or viral load measurements in natural history studies of **AIDS**. Usually cohort studies are prospective because subjects are monitored following establishment of the cohort. In a **historical cohort study**, however, earlier records are used to define membership in the cohort and to determine subsequent changes in health status.

Traditional approaches to the analysis of epidemiologic cohort studies include **person-years** analyses for rare events [1], **survival analyses** for more general time-to-event data [9], and **longitudinal data analysis** for repeated marker data [14]. While classical statistical methodologies routinely address sampling variation, other more systematic sources of error and **bias** that can overwhelm sampling variation are sometimes ignored. For example, **selection bias** resulting from biased sampling of the cohort either at enrollment or during the course of follow-up can seriously distort incidence and **relative risk** estimates. The objective of this article is to review how these different forms of sampling bias arise and how they can affect the results of studies. Other biases that result from **confounding** and measurement errors are discussed in other articles (*see* **Measurement Error in Epidemiologic Studies**; **Bias in Observational Studies**; **Bias in Cohort Studies**).

Self-Selection into the Study

Cohort studies may involve a sample of self-selected volunteers. In some circumstances, the self-selection into cohorts may be an important source of bias. Individuals may be solicited to participate in a cohort study through questionnaires and invitations, and those who respond and choose to participate may differ from those who do not participate with respect to known and unknown disease risk factors. For example, the **Framingham Study**, a long-term cohort study of heart disease initiated about 1950, issued an invitation to every town resident in the age range 20–70 to join the study. A lower death rate was subsequently observed among those individuals who chose to participate compared with nonparticipants [20]. One explanation for the mortality difference was that nonparticipants may be selectively frailer because study participation required a clinic visit. Although one might expect the differences in mortality rates to diminish over time, the Framingham Study found higher mortality rates among nonparticipants at both two and five years after study invitation. In another example involving a cohort study of British physicians, lower mortality rates were observed among physicians who replied to an initial questionnaire compared with those physicians who did not respond [1, 15]. In a population-based **cross-sectional study** of cardiovascular disease (*see* **Cardiology and Cardiovascular Disease**), less cardiovascular disease was found among respondents compared with nonrespondents [10]. In the above examples, less disease was observed among the study participants or responders. However, it is also possible to observe more disease among respondents. For example, the **Centers for Disease Control** investigated leukemia incidence among individuals near the Smoky Atomic Test in Nevada [7, 35]. Among the 18% of subjects who were self-referred and contacted the investigators, because of publicity about the study, there were four leukemia cases. Among the remaining 82% of subjects who were traced by investigators, there were also four leukemia cases. These data suggest that those individuals with disease were more likely to respond voluntarily and to participate in the study.

Self-selection is a potentially important source of bias particularly if the main comparison is with an external **control** group. For example, if the mortality rates of a self-selected group of smokers were compared with general population mortality rates, then the effect of smoking could be masked if the study participants were healthier than the general population. However, if the main comparison is with an internal control group, then self-selection may not be a source of bias. For example, among those who self-selected into the study, the mortality rates of smokers could be compared with those of nonsmokers. If the

assumption is that the effect of self-selection was comparable for both smokers and nonsmokers, then the relative risk of smoking based on the self-selected group may be unbiased. However, the absolute mortality rates of the self-selected groups of smokers and nonsmokers may be a biased estimate of the rates in the general population. There is, however, no guarantee that the relative risks are unbiased because self-selection may act differentially on the exposure groups. Both empirical and theoretical investigations of the impact of self-selection on relative risks have been performed [11, 21, 29].

A related bias is the healthy worker effect (*see* **Occupational Epidemiology**). The healthy worker effect occurs in cohort studies of occupational risks [1, 34]. For example, a cohort study may be conducted to evaluate the health effects associated with working in a particular occupational setting. The mortality rates among those individuals who are employed in the occupation might be compared with an external control group. However, employed individuals may have a lower mortality rate than the general population that is also made up of unemployed. Indeed, individuals may leave employment upon the onset of severe life-threatening diseases. Internal control groups help to correct the bias from the healthy worker effect, where an exposed group of employees is compared with an unexposed group of employees. For example, workers exposed to carbon disulfide were compared with workers in the paper industry as well as with the general Finnish population [23, 28]. Although the exposed workers had the highest coronary heart disease mortality rates, the rates among both exposed and unexposed workers in the paper industry were considerably lower than the rates in the general population.

Follow-Up Bias

Cohort studies typically follow individuals until an event occurs. However, there is often **incomplete follow-up**; that is, an individual may have follow-up data only up to time t and there is no information on the status of the individual beyond t. In that case, all that is known is that the individual did not have an event prior to time t. Such observations are right censored (*see* **Censored Data**). There are different reasons for censoring. One reason may be because the cohort study is ending and the investigators wish to analyze the data. In this case we say the individual is administratively censored. We say the individual is lost to follow-up if the individual is right censored because of all other reasons including the patient has moved away or the individual no longer wishes to participate in the study.

Most statistical methods for the analysis of follow-up data from cohort studies assume that individuals censored at some time t have similar risks as individuals not censored at t. This is noninformative censoring [30]. Generally, the assumption of noninformative censoring is plausible for administrative censoring because the reason for the censoring was external from the individual. If the assumption of noninformative censoring is violated, then we say there is follow-up bias. If the individuals lost to follow-up are at lower risk than those who remain under follow-up, then the event rates will be overestimated. For example, those individuals who are particularly healthy may move away from the study area. Alternatively, if the individuals lost to follow-up are at higher risk of an event than those who remain under follow-up, then the event rates will be underestimated. For example, the frailer individuals may be too weak to attend clinic visits, and are thus the ones lost to follow-up. Of particular concern are studies where the event of interest is a particular cause of death, and individuals who die of other causes are considered censored [17] (*see* **Competing Risks**). The assumption of noninformative censoring may be violated if there is dependence between the cause of death of interest and another cause of death. This can occur if an unknown risk factor is associated with both causes of death. The dependence can be either positive or negative. For example, positive dependence arises between coronary heart disease and some types of cancer if smoking, which is a risk factor for both, is ignored. Alternatively, negative dependence may arise if alcohol consumption is ignored because alcohol may increase risk for some cancers but decrease risk for coronary heart disease [36].

Unfortunately, the assumption that censoring is noninformative cannot usually be verified from observable data [38]. One way to evaluate the sensitivity of the **Kaplan–Meier** estimate to the assumption is to create bounds assuming perfect positive and negative dependence between the censoring and survival times [33]. The lower bound of the estimated survival curve is obtained by assuming that censored observations experience the

event immediately after being censored. The upper bound is obtained by assuming censored observations never experience the event. Dependent censoring cannot only bias survival curve estimation but can also reverse the effect of a true risk factor and make it appear protective [36].

One approach to control the bias introduced by informative censoring is to attempt to identify a variable (covariate) such that in a given level or stratum of the variable, censoring is noninformative. Such a variable would be associated both with the risks of disease and the risk of loss to follow-up. The statistical analysis must then account for this covariate using either **regression** or stratified analyses (*see* **Stratification**). Similar problems arise in longitudinal data analysis where dropouts are informative as opposed to completely random [14]. Some approaches for modeling the dropout process in longitudinal data analysis have been proposed that allow the probability of dropout at time t to depend on the history of measurements up to time t [13].

Truncation

Truncation is a potentially important source of bias in cohort studies with nonstandard sampling. Generally, truncation occurs if the probability that an individual is sampled for follow-up depends on the individual's event time. There are different forms of truncation. Left truncation arises if the individual comes under observation only if the individual's event (survival) time exceeds a known time, which we call the truncation time. Right truncation arises if the individual is included in the cohort only if the individual's event time is less than a known (truncation) time. Left truncation is discussed in more detail in this section, and an example of right truncation is discussed in detail in a later section of this article.

As a simple example of left truncation, consider a study whose objective is to identify **prognostic factors for survival** among patients with a particular disease. Now suppose only individuals with the disease who are *alive* at calendar time C are eligible for sampling into the study. Under this sampling design only individuals with survival times $t_i \geq C - u_i$, where u_i is the calendar time of diagnosis of disease have the opportunity to be included in the cohort. Individuals with shorter survival times are selectively excluded. Unless special adjustments are made for the left truncated data, standard statistical analyses can be seriously biased. For example, suppose the standard **nonparametric** Kaplan–Meier (product–limit) estimate is calculated based on the data (t_i, δ_i), where t_i is the event time for uncensored individuals or the last follow-up time for censored individuals and δ_i is a right censoring indicator that indicates if the individual had an event at time t_i (in which case, $\delta_1 = 1$) or was censored at time t_i (in which case $\delta_i = 0$). The Kaplan–Meier estimate of the survival function, $S(t)$ (probability of surviving beyond time t) is

$$\hat{S}(t) = \prod_{\{t_i \leq t\}} \left(\frac{n_i - d_i}{n_i} \right),$$

where n_i is the number at risk at t_i, i.e. the number of individuals with survival times greater or equal to t_i who have not been previously censored, and d_i are the numbers of (uncensored) events that occurred at t_i. Under usual sampling (i.e. a **random sample** of individuals are chosen for the cohort), the standard Kaplan–Meier estimate is a **consistent estimate** of the survival curve [2]. However, if the data are left truncated, then the standard Kaplan–Meier estimate of the survival curve will overestimate the true survival curve. The intuition for this bias is as follows. If we look at all individuals diagnosed with disease at time u_i, only those with long survival ($t_i \geq C - u_i$) are included in the data set, and this selection results in an overestimation of survival probabilities.

The correct Kaplan–Meier analysis that accounts for left truncation requires a different definition of the risk set. The correct definition of the risk set at time t to account for left truncation includes only those individuals who have neither had an event nor been censored prior to t and *who are under active follow-up at time t*. That is to say we require for an individual to be included in the risk set at time t that (i) the individual has not had an event or been censored prior to time t and (ii) $u_i + t \geq C$. The Kaplan–Meier estimator based on this definition of risk sets has been called the truncation product–limit estimator, and its theoretical properties have been studied by Tsai et al. [37]. Implicit in the truncation product–limit estimator is the assumption that risks (hazards) of the event depend only on follow-up time and depend neither on calendar time of disease diagnosis nor calendar time of study enrollment. These **stationarity** assumptions are discussed in [39].

Parametric estimation of survival curves (*see* **Parametric Models in Survival Analysis**) from left truncated data also requires modification of traditional methods. The key idea is that the contributions from each individual to the **likelihood** function must be *conditional* on the sampling criteria. Specifically, because individuals are required to be alive at calendar time C, we can account for left truncation by conditioning on the event that an individual's survival time is at least $C - u_i$. Then the likelihood function for estimating the parameters in a survival distribution $S(t)$ with density $f(t)$ from data (t_i, δ_i, u_i) for N individuals is

$$\prod_{i=1}^{N} \left[\frac{f(t_i)}{S(C-u_i)} \right]^{\delta_i} \left[\frac{S(t_i)}{S(C-u_i)} \right]^{1-\delta_i}.$$

The usual naive likelihood function that did not account for left truncation would not include the denominator terms $S(C - u_i)$ in the above likelihood function.

Special care is also needed in computing incidence rates when the data are left truncated. The incidence rate is usually calculated as the ratio of the number of events divided by the total person time of follow-up. This estimator of incidence is justified if the hazard of the event is constant over time. A source of confusion in the calculation of total person time is the amount of person time contributed by left truncated individuals. A naive analysis might allocate person time equal to t_i for all individuals, even if left truncated. However, that would lead to an underestimation of event incidence rates because events that occurred before C are not included in the calculation, but person time of the left truncated individuals is included. The correct analysis that accounts for left truncation has individuals contribute person time only during the period that they are actually under observation [1, 3]. Thus, in our example, a left truncated individual contributes person time accrued only after enrollment at calendar time C, that is, the contribution is $t_i - (C - u_i)$.

Special care is also needed in applying the **proportional hazards model** to left truncated data. The model formalizes the relative risk concept for studies of time to response and generalizes it to the regression setting. The model assumes that the time-specific incidence **(hazard) rate** in an exposed population is proportional to the incidence rate in an unexposed population. For a covariate with two levels the model is

$$\lambda_1(t) = \theta \lambda_0(t), \tag{1}$$

where $\lambda_1(t)$ and $\lambda_0(t)$ are the hazard rates at time t among those with and without the risk factor, respectively, and the parameter θ is the hazard ratio (*see* **Hazard Ratio Estimator**) (or relative risk). An important question is: What is the time scale t on which the proportional hazards model (1) is defined? The most appropriate time scale is the one that needs the most careful control to obtain valid comparisons between the treatment groups. Often the time scale is follow-up time, i.e. t is the time the individual has been under active follow-up. Sometimes, however, there is a more natural time scale whose origin is defined by some initiating event [39]. For example, in studies of chronic disease incidence, the most appropriate time scale might be chronological age. In a prognostic factor study for survival among patients with a disease, the origin of the time scale might be the time of disease diagnosis. In this last example, if the time scale refers to time since diagnosis (as opposed to follow-up time) special care is needed in constructing the risk sets of the **partial likelihood** analysis when the data are left truncated. A naive proportional hazards analysis that defined risk sets in the standard way (i.e. all individuals with event times greater or equal to t_i and not previously censored) can yield biased estimates of the relative risk. For example, suppose the cohort study recruits individuals alive with disease at calendar time $C = 1995$. The covariate of interest is disease diagnosed before 1990 ($X = 0$) or after 1990 ($X = 1$). Even if there were no differences in the hazards of death by calendar year of diagnosis ($\theta = 1$), the naive proportional hazards analysis would give $\theta > 1$ (aside from sampling variation). This is because all individuals diagnosed before 1990 ($X = 0$) who have survival times that are less than five years would not have an opportunity to be sampled and included in the cohort. However, among all individuals diagnosed after 1990 ($X = 1$), at least some of the individuals with survival times less than five years could have an opportunity to be sampled. The correct proportional hazards analysis of left truncated data requires an adjusted definition of risk sets similar to the adjustments described previously for Kaplan–Meier estimation. Basically, individuals should be included in risk sets only if they are under active follow-up at that time.

Truncation in Retrospective Studies

Some studies have selected individuals for enrollment based on the occurrence of the primary event of interest, and then studied these cases retrospectively to ascertain their survival time in order to estimate survival curves. The key feature of this type of design is that only individuals who have an event are eligible for inclusion in the study. For example, the first study of the **incubation period** of AIDS involved only a sample of AIDS cases who developed AIDS from a blood transfusion [32]. These patients were identified and selected for study *because they developed AIDS*. These patients were then studied retrospectively to determine the dates of transfusion with infected blood. Thus, the incubation period was determined as the time between blood transfusion and AIDS diagnosis. Another example of a retrospective study with truncation involved the selection of pediatric AIDS patients whose only known risk was maternal transmission; the dates of infection were assumed to be the dates of birth. In this case the incubation period is the time between birth and AIDS diagnosis. These are examples of retrospective studies with *right* truncated data in which individuals with long incubation periods are selectively excluded because they may not yet have had an event at the time of sampling (i.e. the time that individuals are selected for inclusion in the study).

There are also examples of **retrospective studies** with *left* truncated data. For example, a study was conducted to estimate the interval between first exposure to the human immunodeficiency virus (HIV) and the development of detectable HIV antibodies (seroconversion); this interval is called the preantibody phase [41]. This study selected only those individuals who seroconverted late in calendar time (in the late 1980s). Sera samples from these individuals had been stored and were tested by PCR to ascertain the time of first exposure to HIV. However, these individuals are a biased sample and over-represent the longer preantibody durations because individuals with short preantibody duration would have been more likely to seroconvert earlier in calendar time and thus not be included in the study [40, 42].

The remainder of this section considers the analysis of retrospective studies with right truncated data. The methods and issues are illustrated with the study of incubation periods of transfusion associated AIDS described previously. Figure 1 illustrates

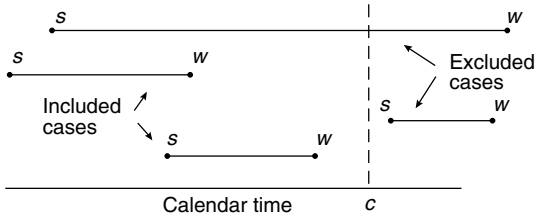

Figure 1 Schematic illustration of a retrospective study of cases (transfusion-associated AIDS) with right truncated data. Only cases diagnosed before calendar time of sampling (C) are included (i.e. $w_i < C$): s = infection date; w = AIDS diagnosis date; c = case ascertainment date

the sampling scheme: all (AIDS) cases diagnosed before some calendar time C are sampled. The figure illustrates the main problems with the analysis and interpretation of this type of data. First, since the data involve only individuals who have experienced events (AIDS diagnoses), without strong parametric assumptions, they can provide no information about the prospective probability that an infected individual eventually develops the disease. Secondly, the sampling scheme tends to over-represent individuals with shorter incubation periods. The data are right truncated because individuals with very long incubation periods are selectively excluded. The data consist of the calendar dates of infection s_i and the calendar dates of diagnosis w_i. The time to event (incubation period) is $u_i = w_i - s_i$. The criterion for inclusion in the data set is that $u_i \leq T_i$, where $T_i = C - s_i$ is called the truncation time. As illustrated in Figure 1, individuals with longer incubation periods tend to be selectively excluded because such individuals are less likely to have developed the endpoint (AIDS) at the time of ascertainment. Failure to account for such **length-biased** sampling will cause the incubation time to be underestimated.

Both nonparametric and parametric statistical procedures have been proposed for estimating the distribution function of the times to event from such data. Nonparametrically, the best that can be done is to estimate the distribution function conditional on the time to event being less than the maximum truncation time. The maximum truncation time T^* refers to the longest event time that could possibly be observed under this sampling procedure. There are simple computational approaches for calculating the nonparametric estimate of $F^*(t)$, which is the cumulative probability that an incubation period is less than t conditional on it being

less than T^*. This is based on expressing F^* as the product of conditional probabilities as follows [26, 31]. Let $t_1, t_2, \ldots, t_n$ be the ordered observed event times. In the AIDS example, these are the incubation times. The nonparametric estimate $\hat{F}^*(t)$ of $F^*(t)$ is

$$\hat{F}^*(t_s) = \prod_{j=s+1}^{n} \left(1 - \frac{d_j}{m_j}\right), \quad s = 1, \ldots, n-1, \tag{2}$$

and $\hat{F}^*(t_n) = 1.0$, where d_j is the number of individuals with event times exactly equal to t_j, and m_j are the numbers of individuals with truncation times greater than or equal to t_j (i.e. $T_i = C - s_i \geq t_i$) and whose event times are less than or equal to t_j. The estimate (2) is a step function with jumps at observed incubation times; $\hat{F}^*$ reaches the value 1.0 at the largest observed event time. The estimate $\hat{F}^*$ accounts for the length-biased sampling that arises from right truncation. A naive estimate that was based simply on the proportion of event times less than t would grossly overestimate the true distribution function $F(t)$ and would suggest that event times are shorter than they really are.

Parametric approaches can also be used to analyze retrospective studies with right truncated data. While some parametric assumptions may permit estimation not only of the conditional distribution F^* but also the unconditional distribution F, the resulting estimates of F are extremely imprecise and depend strongly on parametric assumptions. This is because parametric approaches do not circumvent the main weakness in the data; it is not possible to observe event times greater than the maximum truncation time. Several likelihood functions have been proposed for the parametric analysis of retrospective data on cases with right truncation [5, 26]. The differences in the various likelihood functions that have been proposed arise from using different conditioning events. At a minimum, the likelihood function must condition on having an event prior to the case ascertainment time C. In addition, some of the proposed likelihood functions condition on the time origins (dates of infection s_i).

Prevalent Cohort Studies

Prevalent cohort studies are used to study the **natural history** of disease [6]. The prevalent cohort study consists of a sample of individuals who have a condition or disease at the time of enrollment in the study. These individuals are then followed over time to monitor endpoints such as disease progression or death. In some situations the durations of time the individuals have been prevalent with the disease or condition prior to enrollment are known. For example, Cnaan & Ryan [8] considered survival analysis of a registry of sarcoma patients seen at certain institutions that included some patients who were initially diagnosed elsewhere. These data are left truncated because the patients diagnosed at other institutions must have survived long enough to be included in the sample. The methods for left truncated data outlined in the previous section are required to account adequately for the sampling scheme [8].

A more serious complexity arises in the analysis and interpretation of prevalent cohort studies if the durations of time the individuals have been prevalent with the disease or condition prior to enrollment is unknown. This section is concerned with the issues in the analysis and interpretation of prevalent cohort data when the prior durations are unknown.

An example of a prevalent cohort concerns a study of HIV infected individuals with the objective to estimate rates of progression to AIDS and to identify covariates that modify these rates. In this example, individuals who are alive and previously infected with HIV are eligible for enrollment. These individuals are then followed for the onset of disease progression (AIDS). The main complexity is that the previous calendar times of infection are unknown. Figure 2 gives a schematic illustration of the prevalent cohort study. There are three time scales: calendar time (s), the time from infection (u),

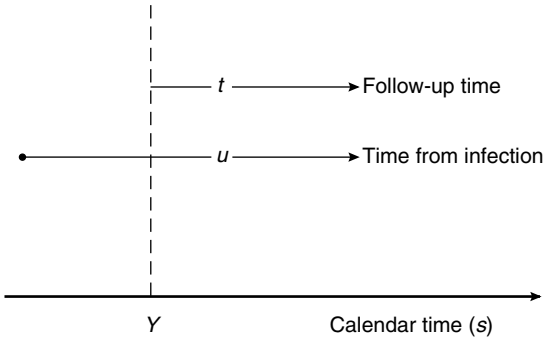

Figure 2 Schematic illustration of the prevalent cohort study

and follow-up time (t). A prevalent sample of individuals who are HIV infected and alive is taken at calendar time Y. The main advantage of the prevalent cohort study is that it can be performed more rapidly than can traditional cohort studies of individuals with incident (new) HIV infection. The traditional cohort study requires a sample of newly infected individuals and there could be considerable expense and effort entailed to identify such a sample. However, a number of important problems arise in the interpretation and analysis of a prevalent cohort that do not arise with a series of newly infected (incident) individuals because the duration of time a person has been infected prior to the beginning of follow-up is not known.

This section outlines the biases and problems of interpretation in prevalent cohorts. Although the prevalent cohort is discussed using the HIV/AIDS example described above, the conclusions, of course, apply to many other settings: for example, a study among prevalent carriers of hepatitis B surface antigen in order to identify modifiers of risk for hepatocellular carcinoma. The unifying feature is that there is an initiating event of a disease (e.g. infection) that defines the natural biological time scale, and individuals who are prevalent with the condition are then enrolled. This results in left truncated data but, unfortunately, one cannot analyze prevalent cohorts using methods for left truncated data because the truncation times are unknown.

Suppose analyses are performed on the time scale of the observed follow-up time (t) instead of the desired, but unobservable, natural time scale (u) such as time from infection. Specifically, how do estimates derived from prevalent cohorts of the probability of an event within t years of follow-up, $F_p(t) = 1 - S_p(t)$, relate to $F(t)$ which is the probability of an event within t years of infection? The proportion of persons in a prevalent cohort who develop disease within t years of follow-up, $F_p(t)$, does not in general approximate $F(t)$. Only if the hazard function on the natural time scale $\lambda(u)$ is constant (an **exponential distribution** for F) do the two coincide. This follows from the lack of memory property of an exponential distribution, which implies that a newly infected individual will progress to the event at the same rate as an individual who has been alive for some time. However, if the hazard $\lambda(u)$ is monotonically increasing, then individuals in the prevalent cohort will be at greater risk of an event than are newly infected individuals. That is to say, the cumulative probability of an event within t years of follow-up of a prevalent cohort is larger than that based on t years of follow-up of an incident (newly infected) cohort [$F_p(t) > F(t)$]. The direction of the bias is reversed for a decreasing hazard. No general statements can be made about the direction of the bias for nonmonotonic hazard functions. Regardless of the shape of the hazard, $F_p(t)$ is a lower bound on the ultimate proportion of individuals who will have an event, $F(\infty)$. Brookmeyer & Gail [4] derived exact expressions for the distribution function on the observed follow-up time scale, $F_p(t)$, in terms of the probability density of infection times among cohort members and the true distribution function F. The magnitude of the biases depend both on the hazard function, $\lambda(t)$, and the density of calendar times of prior infection (the initiating event) among those individuals in the prevalent cohort. For example, the bias would be small if the prevalent cohort is assembled near the beginning of the epidemic, in which case the backward recurrence times (or the times from infection to the onset of follow-up) would be short (*see* **Back-calculation**).

The prevalent cohort study is a rapid and convenient approach to identify cofactors and markers of disease progression. However, because the onset date is unknown, there are biases that result from using follow-up time instead of time from infection. The most important bias associated with identifying cofactors from prevalent cohorts is called *onset confounding* [4, 6]. This occurs when the unknown calendar date of infection is associated both with the risk of disease and the cofactor under study. A subgroup may appear at higher risk of progression to disease simply because they were infected earlier than another subgroup. For example, individuals in one geographic region may exhibit a higher progression rate to disease than other individuals. This finding could be an artifact if individuals in one city were infected earlier in calendar time, and the hazard function $\lambda(t)$ is increasing. The requirement to insure no onset confounding is that the probability densities of infection times among individuals infected before calendar time Y is the same in the two subgroups. Onset confounding can be controlled by stratification on factors such as geographic region. Stratification on a covariate is useful provided we are not interested in determining whether the covariate itself is a cofactor of disease progression.

Unfortunately, even if a covariate has no direct effect on the probability density of infection times among members of the prevalent cohort so that there is no onset confounding, relative risk estimates obtained from prevalent cohorts may still be biased. To see this, assume that a cofactor with two levels obeys the following simple proportional hazards model:

$$\lambda_1(u) = \theta\lambda_0(u),$$

where $\lambda_1(u)$ and $\lambda_0(u)$ are the disease incidence rates at time u among those with and without the factor, respectively, and θ is the ratio of the hazards. In this model, the underlying primary time scale is the natural but unobservable scale u, time since first infection. If this model holds, but a proportional hazards analysis is performed based on follow-up time, then tests of the **null hypothesis** $H_0 : \theta = 1$ will be valid provided there is no confounding. However, estimates of θ based on the incorrect assumption of proportional hazards on the observed follow-up time scale will usually be biased for θ. The term *differential length biased sampling* is used to refer to this bias, which results from differences in the distributions of prior durations of infection (backward recurrence times) between the two prevalent subgroups [4]. Differential length-biased sampling may bias the relative risk from a prevalent cohort, and the direction of the bias depends on whether the hazard function is increasing or decreasing, as discussed below.

If the hazard function is increasing, then relative risk estimates obtained from follow-up on a prevalent cohort will be biased toward unity. A theoretical proof is given in [4] but an intuitive justification is as follows. Infected persons with a risk factor are at higher risk of disease than infected persons without the risk factor. Persons sampled for the prevalent cohort who are in the low-risk group will tend to have longer prior durations of infection than persons in the high-risk group. This is because high-risk persons infected many years earlier are more likely to have developed disease and thus be excluded from the prevalent cohort. Since low-risk persons tend to have been infected for a longer time, their disease is further advanced [with an increasing hazard $\lambda_0(u)$] and therefore the disparity in risk of disease between two groups is reduced, biasing the relative risk toward 1. Analogously, if the hazard function $\lambda_0(u)$ is decreasing, then the relative risk will be biased away from 1. Fortunately, the magnitude of differential length-biased sampling phenomena is never enough to reverse the conclusion, that is to push the relative risk to the other side of 1. Furthermore, there are two situations when the effect of differential length bias could be expected to be negligible: (i) if there is little dispersion in the infection dates (in which case all backward recurrence times are nearly identical) and (ii) if the hazard $\lambda_0(u)$ is small so that only a small proportion of those infected before the initiation of the prevalent cohort study develop disease before the end of follow-up and are selectively excluded (see [4] for a more formal statement of the conditions when the bias from differential length biased sampling is small).

Other types of biases in prevalent cohorts arise in the analysis of the **time-dependent covariate**, that is the variable whose value changes over time u. There are two types of time-dependent covariates, "external" and "internal" [27]. The main distinction is that internal time-dependent covariates reflect the health of the individual (markers) while external time-dependent covariates are applied externally, such as a random assignment to a treatment group. For example, consider an external time-dependent variable which takes effect at some point after the onset of follow-up. An example might be a treatment given to some members of a prevalent cohort. If the treatment is assigned randomly (*see* **Randomization**), then the condition for no onset confounding is satisfied. Unfortunately, even without onset confounding, relative risk estimates of the treatment effects will still be biased [4]. Specifically, assume that the effect of the treatment is to multiply the hazard by θ, that is consider a proportional hazards model with a time-dependent covariate $x(u)$, where $x(u)$ is 0 before initiation of treatment and 1 thereafter:

$$\lambda_1(u) = \theta^{x(u)}\lambda_0(u). \tag{3}$$

Then, the relative risk estimate of θ based on the proportional hazards analysis with a time-dependent covariate on the follow-up time scale t will yield an estimate of θ that is biased toward unity, regardless of whether the hazard function is strictly increasing or decreasing. This bias results because the analysis controls for follow-up time t when in fact the analysis should control for the unknown time from infection u. The result that the risk estimates are biased toward unity is seen intuitively from the following argument. Suppose, without loss of generality, that $\theta > 1$. Then the effect of the variable $x(u)$ is to

accelerate disease, especially among frail individuals. If the hazard is increasing, then the frail individuals are those that have been infected for a longer time. Therefore the individuals with $x(u) = 1$ tend to be selectively depleted from the frail individuals, and the net effect is to decrease the disparity in risk between those with $x(u) = 1$ and those with $x(u) = 0$. This bias has been called *frailty selection*. A similar argument holds if the hazard is decreasing. In either case, the effect of frailty selection is to bias the relative risk toward unity.

Internal time-dependent covariates, also known as *markers*, present different issues. Markers track the progression of disease and change value over the course of follow-up. Markers change in response to disease progression and may convey information about the duration of infection. For example, persons with abnormal marker levels are likely to have been infected longer than persons with normal levels. Suppose the proportional hazards model (3) holds. In model (3) the parameter θ reflects the disease–marker association controlling for duration of infection; that is θ quantifies the prognostic information in the marker over and above the prognostic information in the duration of infection. Unfortunately, in prevalent cohort studies of markers that are performed on the follow-up time scale, estimates of θ in model (3) as well as hypothesis tests of $\theta = 1$ may not be valid and are not comparable with results obtained from an incident cohort. Furthermore, no general statements can be made about the direction of the biases because both frailty selection and onset confounding are operating. For example, a high relative risk associated with an elevated marker may reflect the fact that individuals with the elevated marker have been infected longer.

The various biases associated with prevalent cohorts are summarized in Table 1. Because of these biases, prevalent cohort studies pose serious limitations for studying the disease–marker association of model (3). Prospective studies of an incident cohort are required to disentangle the role of markers and duration of infection on disease risk. Nevertheless, prevalent cohort studies of markers may serve other important purposes. For example, baseline values of markers measured at enrollment ($t = 0$) are useful for prognostic purposes. Survival analyses and proportional hazards analyses on the scale of time since enrollment address the prognostic information in the baseline marker over and above time since enrollment. This is useful for counseling individuals, and for deciding on a course of treatment. These analyses answer the question: "What prognostic information does the baseline marker value provide in addition to *time since enrollment*?" The question "What prognostic information does the baseline marker value provide in addition to *time since infection*?" cannot be answered from such studies and analyses. Nonetheless, such analyses are useful, because the dates of infection are usually unknown in clinical practice; such analyses thus provide important prognostic information for advising patients from similar prevalent cohorts about risk. Such studies may also identify important variables for stratification and adjustment in controlled **clinical trials** of individuals with prevalent infection.

Selection and Regression Towards the Mean

Cohort studies are sometimes performed among the individuals at highest risk of disease. For example, a double blind trial of clofibrate in the primary prevention of ischemic heart disease randomized men who were in the upper third of the distribution of serum cholesterol values [22]. In the Hypertension Detection and Follow-up Program Cooperative Group Study [24], individuals with elevated blood pressure at initial screening were enrolled for follow-up. In these examples, the individuals selected were sampled for inclusion in the study because measurements on a variable at initial screening were extreme. In some instances this selection process can be a source for bias because of **regression toward the mean**.

Regression towards the mean refers to the phenomenon that if a variable is extreme on the first measurement, then later measurements may tend to be closer to the center of the distribution [12]. Regression towards the mean was first described by **Sir Francis Galton** who found that offspring of tall parents tended to be shorter than their parents while offspring of short parents tended to be taller. Galton called this regression toward mediocrity [18].

As a simple example, consider a study to evaluate a treatment to lower blood pressure. Individuals are screened for blood pressure and those in the highest decile are enrolled in the cohort study and given the treatment. Some of those extremely high

Table 1 Bias of relative risk estimate, θ^*, from a prevalent cohort study compared with relative risk estimate, θ, obtained from an incident cohort study[a] (adapted from [6])

		Effect of bias			
		Increasing hazard		Decreasing hazard	
Type of factor	Source of bias	$\theta > 1$[b]	$\theta = 1$	$\theta > 1$[b]	$\theta = 1$
Fixed cofactor affects risk of infection nonmultiplicatively	Onset confounding	No reliable inference from prevalent cohort			
Fixed cofactor unrelated to or acts multiplicatively on risk of infection	Differential length-biased sampling	Biased toward 1 ($1 < \theta^* \leq \theta$)	Unbiased ($\theta^* = 1$)	Biased away from 1 ($\theta^* \geq \theta$)	Unbiased ($\theta^* = 1$)
Time-dependent cofactor which takes effect after enrollment	Frailty selection	Biased toward 1 ($1 \leq \theta^* \leq \theta$)	Unbiased ($\theta^* = 1$)	Biased toward 1 ($1 \leq \theta^* \leq \theta$)	Unbiased ($\theta^* = 1$)
Marker	Onset confounding and frailty selection	No general statements about direction of bias			

[a] θ^* and θ are the large sample expected values of the relative risks obtained from prevalent and incident cohorts, respectively.
[b] Analogous results hold for $\theta < 1$.

measurements at initial screening will decline at the follow-up measurement not because of the efficacy of the treatment but because the initial extreme high measurements were statistical flukes. A naive analysis could lead one to conclude incorrectly that the decline in mean blood pressure is evidence that the treatment is effective. In this example it was especially important to account for regression towards the mean because the study design was a before–after comparison without a control group. Regression towards the mean can lead one to conclude incorrectly that not only are there treatment effects, but there are also **treatment–covariate interactions**. For example, several studies of serum cholesterol lowering diets, such as the National Diet Heart Study, have reported that individuals with initially high serum cholesterol levels experience greater reductions than individuals with initially low cholesterol levels [16].

The effect of regression towards the mean can be formalized [12, 19]. Suppose the only individuals enrolled in a cohort study are those whose initial screening measurement, y_1, are greater than a prespecified value, k. A second measurement, y_2, is taken on follow-up. Suppose y_1 and y_2 are each **normally distributed** with **mean** μ and **variance** σ^2. Then

$$\mathrm{E}(y_1|y_1 > k) - \mathrm{E}(y_2)|y_1 > k) = c\sigma(1-\rho), \quad (4)$$

where c is a positive constant that depends on k, μ and σ, and ρ is the **correlation** coefficient between y_1 and y_2. Thus, even though there is no difference in expected measurements at baseline and follow-up in the entire population $[\mathrm{E}(y_1) = \mathrm{E}(y_2) = \mu]$, there is an expected decline in the two measurements in the *sampled cohort* because of the selection criterion $(y_1 > k)$. The effect of regression towards the mean becomes greater as ρ approaches 0. There is no regression towards the mean if $\rho = 1$.

One approach for accounting for regression towards the mean is to have a suitable control group. For example, if individuals are randomized to either a treated or control group, then both groups could be expected to have the same amount of regression towards the mean, and any significant differences between groups could be attributed to real treatment effects. If a control group is not available, then corrections can be made for the regression towards the mean [12, 25]. The basic idea of these corrections is based on (4) and uses either external or internal estimates of the parameters of the equation.

Various suggestions have been made for improvements in study design to minimize regression towards the mean. For example, the initial selection criteria could be based on the mean $\overline{y}_1$ of n measurements rather than on only a single measurement y_1. Then, only individuals with $\overline{y}_1 > k$ are sampled for inclusion in the cohort. It can be shown that $\mathrm{E}(\overline{y}_1 - y_2|\overline{y}_1 > k)$ goes to 0 as n gets large [12]. Thus, the effect of regression towards the mean can be reduced by using the average of a number of initial measurements as the basis of the selection criteria. Another proposed approach is to use an initial measurement as the basis for selecting individuals into the study. However, then a second initial measurement is taken on the selected sample, and it is this second measurement that is used as the baseline measure to calculate change from the subsequent follow-up measurement. Under this scheme there will be no regression towards the mean if the observations are equicorrelated [12, 16].

References

[1] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*. Vol. 2: The Design and Analysis of Cohort Studies. International Agency for Research on Cancer, Lyon.

[2] Breslow, N.E. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics* **2**, 437–453.

[3] Brookmeyer, R. (1987). Time and latency considerations in the quantitative assessment of risk, in *Epidemiology and Health Risk Assessment*, L. Gordis, ed. Oxford University Press, Oxford, pp. 178–188.

[4] Brookmeyer, R. & Gail, M.H. (1987). Biases in prevalent cohorts, *Biometrics* **43**, 739–749.

[5] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, Oxford.

[6] Brookmeyer, R., Gail, M.H. & Polk, B.F. (1987). The prevalent cohort study and the acquired immunodeficiency syndrome, *American Journal of Epidemiology* **126**, 14–24.

[7] Caldwell, G.G., Kelley, D.B. & Heath, C.W., Jr (1980). Leukemia among participants in military maneuvers at a nuclear bomb test: a preliminary report, *Journal of the American Medical Association* **244**, 1575–1578.

[8] Cnaan, A. & Ryan, L. (1989). Survival analysis in natural history studies of disease, *Statistics in Medicine* **8**, 1255–1268.

[9] Cox, D.R. & Oakes, S.D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[10] Criqui, M.H., Barrett-Connor, E. & Austin, M. (1978). Differences between respondents and nonrespondents

in a population based cardiovascular disease study, *American Journal of Epidemiology* **108**, 367–372.

[11] Criqui, M.H., Austin, M. & Barrett-Connor, E. (1979). The effect of nonresponse on risk ratios in a cardiovascular disease study, *Journal of Chronic Diseases* **32**, 633–638.

[12] Davis, C.E. (1976). The effect of regression to the mean in epidemiological and clinical studies, *American Journal of Epidemiology* **104**, 493–498.

[13] Diggle, P.J. & Kenward, M.G. (1994). Informative dropouts in longitudinal data analysis (with discussion), *Applied Statistics* **43**, 49–93.

[14] Diggle, P.J. et al. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

[15] Doll, R. & Hill, A.B. (1954). The mortality of British doctors in relation to their smoking habits. A preliminary report, *British Medical Journal* **ii**, 1451–1455.

[16] Ederer, F. (1972). Serum cholesterol: effects of diet and regression toward the mean, *Journal of Chronic Disease* **25**, 277–289.

[17] Gail, M.H. (1975). A review and critiques of some models in competing risk analysis, *Biometrics* **35**, 209–222.

[18] Galton, F. (1985). Regression towards mediocrity in hereditary stature, *Journal of the Anthropology Institute* **15**, 246–263.

[19] Gardner, M.J. & Heady, J.A. (1973). Some effects of within person variability in epidemiological studies, *Journal of Chronic Diseases* **26**, 781–795.

[20] Gordon, T.F.E., Moore, F.E., Shurtleff, D. & Dawber, T.R. (1959). Some epidemiologic problems in the long-term study of cardiovascular disease. Observations on the Framingham Study, *Journal of Chronic Diseases* **10**, 186–206.

[21] Greenland, S. (1977). Response and follow-up bias in cohort studies, *American Journal of Epidemiology* **106**, 184–187.

[22] Heady, J.A. (1973). A cooperative trial in the primary prevention of ischemic heart disease using clofibrate, design methods and progress, *Bulletin of the World Health Organization* **48**, 243–256.

[23] Hernberg, S.M., Nurminen, M. & Tolonen, N. (1973). Excess mortality from coronary heart disease in viscose rayon workers, *Work Environmental Health* **10**, 93–98.

[24] Hypertension Detection and Follow-up Program Cooperative Group (1977). Blood pressure studies in 14 communities, *Journal of the American Medical Association* **237**, 2385–2391.

[25] James, K.E. (1973). Regression toward the mean in uncontrolled clinical studies, *Biometrics* **29**, 121–130.

[26] Kalbfleisch, J.D. & Lawless, J.F. (1989). Inference based on retrospective ascertainment: an analysis of data on transfusion related AIDS, *Journal of the American Statistical Association* **84**, 360–372.

[27] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[28] Kelsey, J.L. & Thompson, W.D. (1986). *Methods in Observational Epidemiology*. Oxford University Press, Oxford.

[29] Kleinbaum, D.G., Morgenstern, H. & Kupper, L.L. (1981). Selection in epidemiologic studies, *American Journal of Epidemiology* **113**, 452–463.

[30] Lagakos, S.W. (1979). General right censoring and its impact on the analysis of survival data, *Biometrics* **35**, 139–156.

[31] Lagakos, S.W., Barraj, L.M. & DeGruttola, V. (1988). Nonparametric analysis of truncated survival data with application to AIDS, *Biometrika* **75**, 515–523.

[32] Lui, K.-J. Lawrence, D.N., Morgan, W.M., Peterman, T.A., Haverkos, H.W. & Bregman, D.J. (1986). A model based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome, *Proceedings of the National Academy of Sciences* **83**, 3051–3055.

[33] Peterson, A. (1976). Bounds for a joint distribution function with fixed subdistribution functions. Applications to competing risks, *Proceedings of the National Academy of Sciences* **73**, 11–13.

[34] Robbins, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-applications to control of the healthy workers effect, *Mathematical Modelling* **7**, 1393–1512.

[35] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company, Boston.

[36] Slud, E. & Byar, D. (1988). How dependent causes of death can make risk factors appear protective, *Biometrics* **44**, 265–269.

[37] Tsai, W.Y., Jewell, N.P. & Wang, M.C. (1987). A note on the product-limit estimator under right censoring and left truncation, *Biometrika* **74**, 883–886.

[38] Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences* **72**, 20–22.

[39] Wang, M.C., Brookmeyer, R. & Jewell, N.P. (1993). Statistical models for prevalent cohort data, *Biometrics* **49**, 1–11.

[40] Winkelstein, W., Royce, R.A. & Sheppard, H.W. (1990). Median incubation time for human immunodeficiency virus (HIV) (letter), *Annals of Internal Medicine* **112**, 797.

[41] Wolinsky, S.M., Rinaldo, C.R. & Phair, J. (1990). Response to letter, *Annals of Internal Medicine* **112**, 797–798.

[42] Wolinksy, S.M., Rinaldo, C.R. & Kwok, S. (1989). Human immunodeficiency virus type 1 (HIV-1) infection a median of 18 months before a diagnostic Western Blot, *Annals of Internal Medicine* **111**, 961–972.

(*See also* **Cross-sectional Study**; **Incidence–Prevalence Relationships**)

RON BROOKMEYER

Binary Data

Data are said to be binary when the observed value of a response variable falls into one of two possible categories. For example, a patient in a clinical trial to compare alternative therapies may or may not experience relief from symptoms. Similarly, an insect exposed to a particular concentration of an insecticide may either be alive or dead after a certain period of time. The two possible values of the binary response variable for each individual are usually coded as 0 and 1. Expressed in this way, the observations are referred to as *ungrouped* binary data.

In some circumstances, interest centers on a set of individuals that have all been treated in the same manner. The binary responses for each individual in a set are then combined to give a proportion. Thus a batch of insects may be exposed to an insecticide, and the number of insects that respond is expressed as a proportion of the number exposed. Data in this form are referred to as *grouped* binary data.

Usually, grouped binary data will be obtained for a number of sets of individuals. Thus, in the insecticide example we may wish to explore how different batches of insects respond to a range of concentrations of the insecticide. The total number that are killed in each batch is then recorded for each concentration. Interest then centers on how the probability that an insect dies is related to the concentration of the insecticide. Data in this form are commonly encountered in bioassay (*see* **Biological Assay, Overview**).

Probability Models for Binary Data

The random variable associated with the binary response of the ith of n individuals in a study, Y_i, say, will take a value y_i, where y_i is either 0 or 1. These two possible values of the response variable are often referred to as failure and success, respectively. We will write p_i for the probability that the ith individual experiences a success, so that $p_i = \Pr(Y_i = 1)$. The random variable Y_i then has a *Bernoulli distribution*, and

$$\Pr(Y_i = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i},$$
$$y_i = 0, 1; \quad i = 1, 2, \ldots, n.$$

Now suppose that we have a set of m binary observations, $y_1, y_2, \ldots, y_m$, which are mutually independent. If each binary response has the same success probability p, then the total number of successes in the m binary observations is $y = \sum_{i=1}^{m} y_i$, and the corresponding random variable $Y = \sum_{i=1}^{m} Y_i$ has a **binomial distribution** with parameters m and p. We then have that

$$\Pr(Y = y) = \binom{m}{y} p^y (1 - p)^{m-y},$$
$$y = 0, 1, \ldots, m,$$

and this distribution is written as $B(m, p)$. More generally, consider the situation where we have n sets of binary data, such that in the jth set, $j = 1, 2, \ldots, n$, there are y_j successes amongst m_j binary observations. If we write p_j for the true probability of a success for an individual in the jth set, then the values y_j are observations on $B(m_j, p_j)$ random variables. Ungrouped binary data can be regarded as a special case in which $m_j = 1$ for each set, and we would then be back to n binary observations.

Modeling Binary Data

In studies that lead to ungrouped binary data there will usually be other variables recorded for each individual. For example, consider a study to compare two alternative treatments for prostatic cancer, where the response variable concerns whether or not an individual dies within the three-year period following entry to the trial. The age of the patient, and the values of variables such as tumor size, tumor grade, and the level of serum acid phosphatase, may all affect the prognosis of an individual. In analyzing the data from the study we would examine how the three-year survival probability depends on these variables, as well as on the treatment group.

For grouped binary data the values of any explanatory variables will perforce be the same for each of the binary observations in a given set. For example, in the insecticide study, suppose that batches of 20 insects are exposed to one of two chemicals applied at one of four concentrations. The resulting data will be the eight proportions of insects that die, out of the 20 exposed, for each of the eight combinations of chemical and concentration. Of course,

these grouped binary data could be ungrouped to give the $20 \times 8 = 160$ binary observations, but there are advantages in working with the grouped data.

When the structure of a data set is particularly simple, i.e. when there are just one or two explanatory variables recorded for each individual, methods of analysis based on a **contingency table** may be sufficient. However, it is usually much more informative to describe the relationship between the response probabilities and explanatory variables using a statistical model. This approach is now described and illustrated. Full details can be found in [2], [6], [12], and in the research monograph of McCullagh & Nelder [9] on **generalized linear models**.

Example: Incidence of Sore Throat

It will be convenient to illustrate the modeling process using a specific example, and for this we will use data based on a study to compare two devices used in securing the airway in patients undergoing surgery with general anesthesia. The two devices were the laryngeal mask airway (LMA) and the tracheal tube (TT); in the data set the device is denoted by a variable, 0 for LMA and 1 for TT. In addition, information on each patient's age (in years), sex (0 = female, 1 = male), duration of surgery (in minutes) and on whether or not a lubricant was used by the anesthetist (0 = not used, 1 = used) was recorded. The response variable of interest is whether or not a patient experienced a sore throat on waking (0 = no sore throat, 1 = sore throat), and so is binary. Data for 35 patients are given in Table 1.

In analyzing these data we investigate whether the probability of a sore throat depends on the variables age, sex, duration, and lubricant, and the extent of any differences between the two types of airway device.

Table 1 Data on the incidence of sore throat following anesthesia

Patient	Age	Sex	Lubricant use	Duration	Type	Response
1	48	1	0	45	0	0
2	48	1	0	15	0	0
3	39	0	1	40	0	1
4	59	1	0	83	1	1
5	24	1	1	90	1	1
6	55	1	1	25	1	1
7	35	0	1	35	0	1
8	23	1	1	65	0	1
9	57	0	1	95	0	1
10	34	1	1	35	0	1
11	56	0	1	75	0	1
12	35	0	0	45	1	1
13	37	0	1	50	1	0
14	30	1	1	75	1	1
15	45	1	1	30	0	0
16	60	1	0	25	0	1
17	35	1	1	20	1	0
18	41	1	0	60	1	1
19	67	0	1	70	1	1
20	25	0	0	30	0	1
21	63	0	1	60	0	1
22	26	0	1	61	0	0
23	47	0	0	65	0	1
24	27	0	0	15	1	0
25	18	0	1	20	1	0
26	64	0	0	45	0	1
27	48	0	0	15	1	0
28	28	1	0	25	0	1
29	54	1	0	15	1	0
30	58	1	1	30	0	1
31	59	1	1	40	0	1
32	67	1	0	15	1	0
33	43	1	1	135	1	1
34	63	1	0	20	1	0
35	41	0	0	40	1	0

Models for Binary Data

Methods used in fitting **multiple regression** models to continuous data, based on ordinary **least squares**, cannot be applied directly to binary or binomial response data. This is because these methods do not take proper account of the fact that the data have a binomial distribution, and second they may well lead to fitted probabilities outside the range (0, 1). Instead, the probability scale is transformed from the range (0, 1) to $(-\infty, \infty)$, and a linear model is adopted for the transformed value of the response probability.

Of the possible transformations, the *logistic transform* is the one most commonly used. The logistic transform or logit of p is $\log\{p/(1-p)\}$, written $\text{logit}(p)$. If $x_{1i}, x_{2i}, \ldots, x_{ki}$ are the values of k explanatory variables $X_1, X_2, \ldots, X_k$ for the ith individual, $i = 1, 2, \ldots, n$, then the **logistic regression** model for p_i, the response probability for that individual, is given by

$$\text{logit}(p_i) = \log\left\{\frac{p_i}{1-p_i}\right\} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}.$$

Other transformations of p_i that might be used are the *probit* and *complementary log–log transformations*. The probit of a probability p is that value of ξ for which $\Phi(\xi) = p$, where $\Phi(\cdot)$ is the standard normal distribution function, so that probit $(p) = \Phi^{-1}(p)$. The complementary log–log transform of p is $\log\{-\log(1-p)\}$. For practical purposes the logistic and probit transformations will often give similar results, but the logistic transformation is computationally more convenient and leads directly to **odds ratios**, which help in the interpretation of fitted models. Unlike the logistic and probit transformations, the complementary log–log transformation is not symmetric about $p = 0.5$, but it does arise in particular areas of application, such as a **serial dilution assay** and the analysis of interval-censored survival data.

In what follows the transformation of p is denoted by the function $g(\cdot)$, so that a general model for binary data, or a **quantal response model**, can be expressed as

$$g(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, \quad i = 1, 2, \ldots, n. \tag{1}$$

This model is, in fact, a member of the class of models known as *generalized linear models*, and the function $g(\cdot)$ is known as the *link function*.

Fitting the Model

The model for a binary response probability can be fitted using the method of **maximum likelihood**. The **likelihood** of n binomial observations is

$$L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}, \tag{2}$$

where the p_i are related to the β coefficients through (1). Note that for ungrouped binary data the n_i in (2) are all equal to unity.

The values of the βs that maximize this likelihood function, denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, can only be obtained numerically, and either the Newton–Raphson method or Fisher scoring are generally used (*see* **Optimization and Nonlinear Equations**). These methods lead to the **information matrix**, from which the standard errors of the $\hat{\beta}$s can be found.

Once estimates of the βs in (1) have been obtained, the corresponding fitted probability for the ith individual, $\hat{p}_i$, is found from

$$g(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}, \quad i = 1, 2, \ldots, n. \tag{3}$$

Fortunately, computer software for obtaining the maximum likelihood estimates of the β coefficients in (1) is widely available. The statistical packages MINITAB, GLIM, Genstat, SAS, S-PLUS, SPSS, STATA, and many others all have facilities for modeling binary data (*see* **Software, Biostatistical**). Naturally, there are differences in the numerical methods used to estimate the βs, in the types of variable that can be included in the model (some software does not allow factors to be fitted directly), and in the format of the resulting output.

Goodness of Fit of a Model

Suppose that a model containing certain explanatory variables, known as the *current model*, is fitted to grouped binary data. The agreement between n observed proportions y_i/n_i, $i = 1, 2, \ldots n$, and the corresponding fitted values under the model of interest, $\hat{p}_i$, can be assessed using a quantity known as the *deviance*. This is defined to be $-2\{\log \hat{L}_c - \log \hat{L}_f\}$, where $\hat{L}_c$ is the maximized likelihood under the current model, and $\hat{L}_f$ is the maximized likelihood under the *full* or *saturated model*. The latter model is one that is a perfect fit to the data, and so the fitted probabilities under the full model are simply the observed proportions y_i/n_i. If the current model is satisfactory, then the fit of this model will not be too different from that of the full model. In this case $\hat{L}_c$ will be similar to $\hat{L}_f$, and the deviance will be close to zero. However, if the current model is a poor fit, then $\hat{L}_c$ may be very much smaller than $\hat{L}_f$, and the deviance will be large.

We can calculate a deviance for ungrouped binary data, where $n_i = 1$ for each observation, but then the deviance turns out to be uninformative about the goodness of fit of a model. For example, in the particular case of the linear logistic model, the deviance is a function of $\hat{p}_i$ alone, and so can tell us nothing about the agreement between the binary observations y_i and the corresponding fitted probabilities. Consequently, the deviance for ungrouped binary data cannot be used as a summary measure of goodness of fit. It is

therefore desirable to group binary data when possible, since this leads to an overall measure of the goodness of fit of a model.

Distribution of the Deviance

From results concerning the asymptotic distribution of the maximized likelihood ratio statistic (see, for example, [5]), it follows that the deviance for binomial data has an asymptotic χ^2 distribution on $n - \nu$df, where ν is the number of unknown βs in the current model. Thus a well fitting model should have a deviance that is not significantly large relative to the percentage points of a $\chi^2_{n-\nu}$ distribution. Since $E(\chi^2_{n-\nu}) = n - \nu$, a useful rule of thumb is that the deviance should be close to its corresponding number of degrees of freedom in a satisfactory model. However, caution must be exercised in using such general measures of model adequacy, and the fit of a model should also be assessed critically using the model-checking diagnostics described in a later section.

For binary data the deviance does not have an asymptotic χ^2 distribution, and so even when an appropriate model has been fitted, the deviance will not necessarily be close to its number of degrees of freedom. Nevertheless, we shall see in the next section that this quantity is valuable in comparing models fitted to either grouped or ungrouped binary data.

Comparing Alternative Models

The main use of the deviance is in comparing alternative models for a binary or binomial response variable. Suppose that one model contains terms that are additional to those in another. The difference in deviance between the two models then reflects the extent to which the additional terms improve the fit of the model. For example, suppose that the following two nested models, labeled Model (1) and Model (2), are to be compared:

Model (1):

$$g(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_h x_{hi};$$

Model (2):

$$g(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_h x_{hi} + \beta_{h+1} x_{h+1,i} + \cdots + \beta_k x_{ki}.$$

Denoting the deviance under each model by D_1 and D_2, so that these deviances have $n - h - 1$ df and $n - k - 1$df, respectively, D_2 will be smaller that D_1, since Model (2) contains more terms than Model (1). The difference in deviances, $D_1 - D_2$, will reflect the additional effect of the corresponding variables $X_{h+1}, X_{h+2}, \ldots, X_k$, after $X_1, X_2, \ldots, X_h$ have been included in the model. This difference in deviance has an approximate χ^2_{k-h} distribution on the hypothesis that the additional $k - h$ variables are not needed in the model. In general, changes in deviance can be compared with appropriate percentage points of a χ^2 distribution to determine whether or not terms need to be included in or excluded from a model. In fact, the difference in deviance between two nested models is asymptotically χ^2 for both binary and binomial data, and so procedures for comparing models apply equally to grouped and ungrouped binary data.

Using the deviance to compare models fitted to binary or binomial response data, the most appropriate combination of variables for describing observed variation in such data can be determined. Methods for identifying subsets of explanatory variables are described in texts on linear regression modeling, e.g. Draper & Smith [7] and Montgomery & Peck [11]. See also [10] and Appendix 2 of [6].

The method for comparing models described above is based on the asymptotic distribution of changes in deviance. For the proper application of this method, the data must not be too sparse, i.e. proportions must not be based on small numbers of individuals, or the number of binary observations needs to be large relative to the number of parameters being fitted. If this is not the case, then there will be difficulties associated with the convergence of the algorithm for fitting the models, and certain parameter estimates may appear to have unusually large standard errors. In such cases one cannot rely on the asymptotic properties of test statistics. Recent computational advances have led to the development of exact methods for analyzing contingency tables and for logistic regression. These techniques are described elsewhere, and are implemented in the software packages StatXact and LogXact.

Example: Incidence of Sore Throat

We first identify which of the variables relating to the age of the patient (A), the sex of the patient (S), the duration of surgery (D) and the use of a lubricant (L)

Table 2 Deviances on fitting logistic regression models

Variables in model	Deviance	df
Constant only	46.18	34
A	45.88	33
S	46.18	33
L	44.08	33
D	33.65	33
T	42.58	33
$D + T$	30.14	32

need to be included in a logistic regression model for the probability of a sore throat before the effect of type of airway (T) is considered. The deviances on fitting certain logistic regression models to the data in Table 1 are given in Table 2. These show that there are no significant reductions in deviance when either A, S, or L are added to the model that contains a constant only. We also find that these variables do not become relevant in the presence of others and so need not be considered further. However, the decrease in deviance on adding D to a model that contains a constant term alone is 12.53 on 1 df, which is highly significant ($P < 0.001$).

The main focus of the study is to compare the two airway devices. When the variable T is added to the model that contains D, the reduction in deviance is 3.51 on 1 df, which is significant at the 10% level ($P = 0.061$). There is therefore evidence of a difference in the two types of airway device, after allowing for the effect of the duration of surgery. Once D and T are included in the model, no further variables lead to a significant reduction in deviance. There is also no need to include interactions between any variables.

The equation of the fitted logistic regression model for the probability of a sore throat is

$$\text{logit}(\hat{p}_i) = -1.417 + 0.069D_i - 1.659T_i, \quad i = 1, 2, \ldots, 35, \tag{4}$$

where D_i and T_i are the values of D and T for the ith individual. This analysis shows that the probability of a sore throat is dependent upon the duration of surgery and the type of airway device used. The positive coefficient of D in the model shows that the probability of a sore throat increases with the duration of surgery, while the negative coefficient of T indicates that there is a higher probability of a sore throat when $T = 0$, i.e. when the laryngeal mask airway is used.

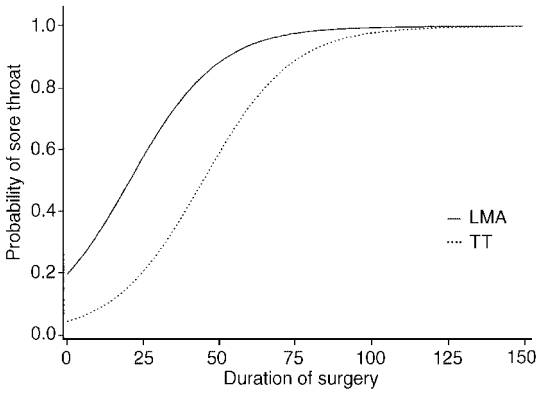

Figure 1 The fitted probability of a sore throat plotted against duration of surgery for laryngeal mask airway (LMA) and tracheal tube (TT)

It is informative to plot the fitted response curves, and such a graph is shown in Figure 1 where we have plotted the observed and fitted probability of a sore throat, $\hat{p}_i$, on the vertical axis. The logistic transform of $\hat{p}_i$ could also have been plotted, in which case the graph would show straight-line relationships between the transformed fitted probability and duration of surgery. This graph shows clearly the relationship between the probability of a sore throat and duration of surgery, and the extent of the difference due to the type of airway.

Model Checking

Once a model has been fitted it is essential to check that the fitted model is appropriate (*see* **Model Checking**). After all, inferences drawn on the basis of an incorrect model will simply be wrong. There are a number of ways in which a model may be unsatisfactory. The linear component of the model may be incorrectly specified in that it may not include explanatory variables that really should be in the model, or variables that are included may need to be transformed. An incorrect choice of link function may have been made, or there may be observations not well fitted by the model, termed outliers. In binary data, outliers may correspond to cases where a failure has been misclassified as a success. There may be values that unduly influence quantities such as parameter estimates. The assumption of a binomial distribution for grouped binary data may also be invalid, possibly

because of the nonindependence of the constituent binary observations.

Techniques for examining the adequacy of a fitted model, known collectively as model-checking diagnostics, are described and illustrated in Chapter 5 of Collett [2]. Some key papers in this area are [13, 8, 14], and [4].

Analysis of Residuals

Much information about model adequacy can be obtained from residuals. Suppose that a model is fitted to n binomial observations of the form y_i/n_i and, as usual, let $\hat{p}_i$ be the fitted probability for the ith observation, $i = 1, 2, \ldots, n$. There are a number of possible residuals for use in binary data analysis, including *Pearson residuals*, given by

$$r_i = \frac{y_i - n_i\hat{p}_i}{[n_i\hat{p}_i(1-\hat{p}_i)]^{1/2}},$$

and the *deviance residuals*

$$d_i = \text{sgn}(y_i - n_i\hat{p}_i) \times (\text{deviance component for } i\text{th observation})^{1/2}.$$

The squares of these residuals sum respectively to Pearson's χ^2 statistic and the deviance for the fitted model. In the particular case of logistic regression, the deviance residuals are

$$\text{sgn}(y_i - n_i\hat{p}_i)\left[2y_i\log\left(\frac{y_i}{n_i\hat{p}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{p}_i}\right)\right]^{1/2}.$$

Both types of residuals can be standardized by division by $(1-h_i)^{1/2}$, where h_i is a quantity known as the *leverage*. This is the ith diagonal element of the matrix $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$, known as the hat matrix, in which $\mathbf{X}$ is the matrix of explanatory variables known as the design matrix, and $\mathbf{W}$ is a diagonal matrix of weights with elements $n_i/\{\hat{p}_i(1-\hat{p}_i)[g'(\hat{p}_i)]^2\}$, where $g'(\cdot)$ is the derivative of $g(.)$ with respect to p.

The standardized deviance residuals, $r_{Di} = d_i/(1-h_i)^{1/2}$, are recommended for general use and can be plotted against the observation number or index to give an index plot, explanatory variables in or out of the model, or the linear predictor, $\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$. Observations that are outliers will be shown in an index plot as having unusually large residuals. The pattern in a plot of residuals against explanatory variables in the model may indicate the need for a transformation of that variable, and the pattern in a plot of residuals against the linear predictor may also suggest that the linear component of the model is not correct.

A **half-normal** plot, possibly supplemented by simulation envelopes (Atkinson [1]), can be helpful in revealing model inadequacy. These plots are based on the absolute values of the residuals arranged in ascending order, denoted $|r|_{(j)}$, $j = 1, 2, \ldots, n$. The values $\Phi^{-1}\{(j + n - \frac{1}{8})/(n + \frac{1}{4})\}$ are then plotted against the $|r|_{(j)}$. Outliers will correspond to points in the top right-hand corner of the plot. A simulated envelope indicates the region of the plot where the points should lie, if the model is satisfactory. Consequently, the occurrence of points outside such an envelope indicates that the fitted model is inappropriate.

Most of these plots of residuals are designed for use with grouped binary data, and corresponding plots of residuals obtained from binary response data may not be informative. This is because the plots can have a pattern even when the fitted model is appropriate. For example, plots of residuals obtained from binary data against the linear predictor will show two hyperbolas, corresponding to the observations of 0 and 1. However, index plots and half-normal plots of standardized deviance residuals can be useful in assessing model adequacy.

Influential Observations

An observation is said to be *influential* if its omission from the data set has a substantial effect on model-based inferences. In the assessment of influence, it turns out that the leverage, h_i, is an important quantity. Observations that are distant from the others in terms of the explanatory variables alone have unusually large values of the leverage and so an index plot of the leverage will reveal such observations. They may be influential, but need not necessarily be. The values of the leverage can be obtained directly from some software packages. In situations where a package does not provide the leverage, but gives the variance or standard error of the linear predictor, $\hat{\eta}_i$, the values of h_i can be found using the result $h_i = \text{var}(\hat{\eta}_i)w_i$, where w_i are the weights used

in the model-fitting process, often called the *iterative weights*. For the linear logistic model, $w_i = n_i\hat{p}_i(1-\hat{p}_i)$.

Observations that influence the set of parameter estimates in a model can be detected using the analog of a statistic proposed by Cook [3]. This is an approximation to the change in the maximized log likelihood when the ith observation is omitted from the data base, given by

$$D_i = \frac{h_i r_{Pi}^2}{\nu(1-h_i)}, \quad i = 1, 2, \ldots, n,$$

where $r_{Pi} = r_i/(1-h_i)^{1/2}$ is the standardized Pearson residual, h_i is the leverage and ν is the number of β coefficients in the fitted model. Approximations such as this only involve quantities that can be obtained from fitting the model to the complete data set and so do not require that the model be actually fitted to each reduced data set.

Unusually large values of the D statistic will indicate observations that unduly affect the set of parameter estimates. If such observations were omitted, then the parameter estimates may change quite markedly, and as a result so would conclusions drawn from the fitted model.

It is often of interest to examine the impact of each observation on a particular parameter estimate, $\hat{\beta}_j$, say. For this, we use an approximation to the change in $\hat{\beta}_j$ brought about by excluding each observation in turn from the data base. This is the quantity denoted $\Delta_i\hat{\beta}_j$ and given by

$$\Delta_i\hat{\beta}_j = \frac{(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}_{j+1}\mathbf{x}_i(y_i - n_i\hat{p}_i)}{(1-h_i)\mathrm{se}(\hat{\beta}_j)},$$

where $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}_{j+1}$ is the $(j+1)$th row of the variance–covariance matrix of the parameter estimates and $\mathbf{x}_i$ is the vector of explanatory variables for the ith observation. This statistic is widely referred to as a delta–beta.

The Binomial Assumption

When an appropriate model is fitted to grouped binary data the deviance is expected to be close to its corresponding number of degrees of freedom. If we find that this deviance is too large, and this cannot be explained by an incorrect model, the presence of outliers and so on, then it may be that there is a positive correlation between the binary responses that form the binomial data. If so, the number of successes will exhibit more variability than the binomial distribution allows. This phenomenon is known as **overdispersion**. It is important to note that the deviance can only indicate overdispersion when the data are grouped. In ungrouped data, the deviance does not necessarily have a χ^2 distribution and so the deviance need not be close to the number of degrees of freedom for a satisfactory model.

Example: Incidence of Sore Throat

On fitting the model for the probability of a sore throat given in (4), index plots of the residuals, leverage, and Cook's D statistic were obtained. These plots are shown in Figures 2–4. The index plot of the residuals shows that the model does not fit particularly well, in that several observations have relatively

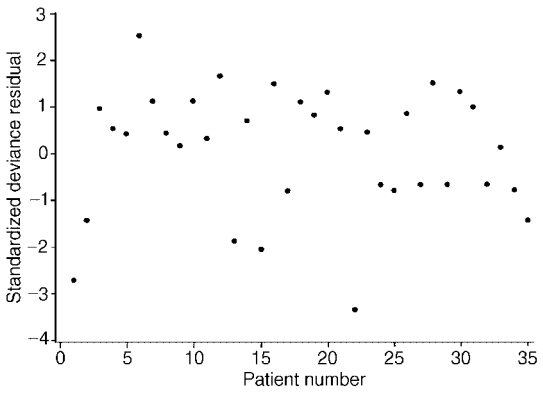

Figure 2 Index plot of the residuals

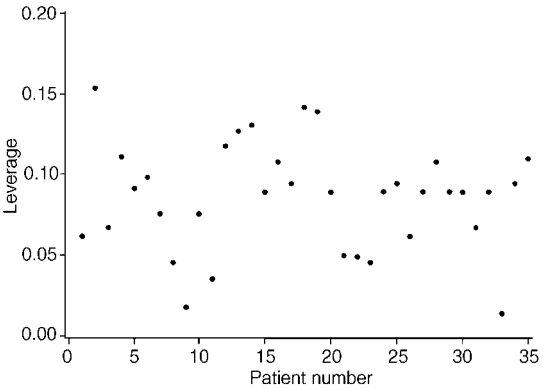

Figure 3 Index plot of leverage

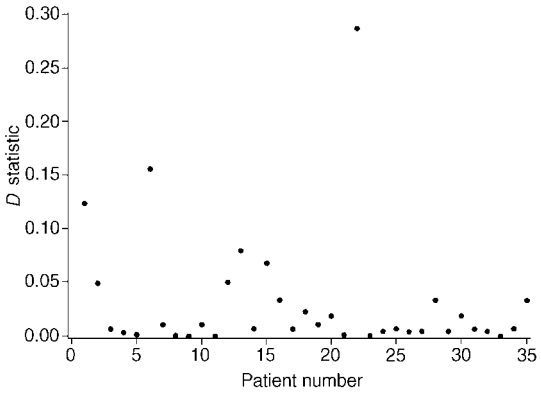

Figure 4 Index plot of the D statistic

large residuals, particularly for patient 22. Under the model, the fitted probability of a sore throat for this individual is 0.94, and yet this patient did not report a sore throat on waking.

The index plot of the leverage shows that there are no patients with an unusual combination of the values of the explanatory variables, but the index plot of the D statistic clearly shows that the observation from patient 22 is influential. The next step would be to investigate whether there have been any errors in recording the data for this particular patient. If not, the actual effect that the data for this patient has on the form of the model will need to be studied.

References

[1] Atkinson, A.C. (1981). Two graphical displays for outlying and influential observations in regression, *Biometrika* **68**, 13–20.

[2] Collett, D. (2002). *Modelling Binary Data* 2nd Ed. Chapman & Hall/CRC, Boca Raton.

[3] Cook, R.D. (1977). Detection of influential observations in linear regression, *Technometrics* **19**, 15–18.

[4] Copas, J.B. (1988). Binary regression models for contaminated data (with discussion), *Journal of the Royal Statistical Society, Series B* **50**, 225–265.

[5] Cox, D.R., & Hinkley, D.V. (1977). *Theoretical Statistics*. Chapman & Hall, London.

[6] Cox, D.R., & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London.

[7] Draper, N.R., & Smith H. (1981). *Applied Regression Analysis*, 2nd Ed. Wiley, New York.

[8] Landwehr, J.M., Pregibon, D.A., & Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models (with discussion), *Journal of the American Statistical Association* **79**, 61–83.

[9] McCullagh, P.J., & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall/CRC, London.

[10] Miller, A.J. (2002). *Subset Selection in Regression* 2nd Ed. Chapman & Hall/CRC, Boca Raton.

[11] Montgomery, D.C., & Peck, E.L. (1982). *Introduction to Regression Analysis*. Wiley, New York.

[12] Morgan, B.J. (1984). *Analysis of Quantal Response Data*. Chapman & Hall/CRC, London.

[13] Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics* **9**, 705–724.

[14] Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions, *Applied Statistics* **36**, 181–191.

(*See also* **Logistic Regression**; **Rasch Models**)

D. COLLETT

Binomial Distribution

The binomial distribution is an important discrete distribution arising in many biostatistical applications. To fix ideas, consider a **binary** (or Bernoulli) response from individual subjects, denoted by "success" and "failure". For example, these responses could denote smoking status (smoker vs. nonsmoker), whether or not a patient develops complications following surgery, or other outcomes having two possible states. Assuming a sample of n independent responses or trials, each with a common probability of success, p, the total number of successes, Y, follows a binomial distribution with probability mass function:

$$\Pr(Y=k)=\binom{n}{k}p^k(1-p)^{n-k}, \quad k=0,1,\ldots,n,$$

where

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

and $0 < p < 1$. It is often convenient to denote the probability of failure by $q = 1 - p$. To express that Y follows a binomial distribution with parameters n and p, we write $Y \sim \text{bin}(n, p)$, where "$\sim$" is read as "is distributed as". In practice, the assumptions of independence and common success probability may not be strictly accurate, but the binomial distribution may still give a reasonable representation. Figure 1 provides a histogram of the binomial probabilities for the case where $n = 12$ and $p = 0.4$.

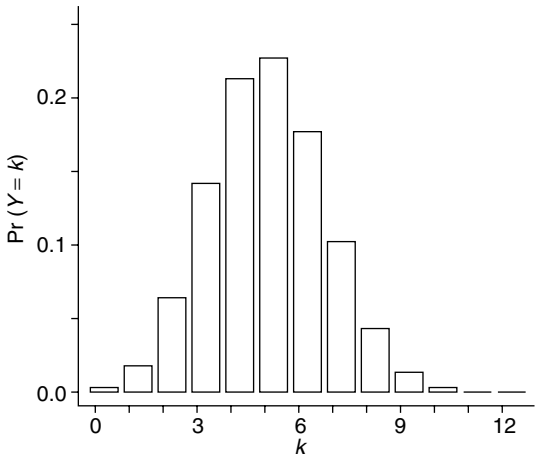

Figure 1 Histogram of the binomial distribution when $n = 12$ and $p = 0.4$

This distribution is one of the oldest to have been studied, dating back to James Bernoulli's *Ars Conjectandi* of 1713 (*see* **Bernoulli Family**). However, binomial coefficients are found in the earlier work of **Pascal**. In fact, the name of the distribution arises from fact that $\Pr(Y = k)$ is the $(k + 1)$st term in the binomial expansion of $(q + p)^n$.

Properties of the Binomial Distribution

The mean and variance are given by

$$\text{E}(Y)=\sum_{k=0}^{n}k\times\Pr(Y=k)=np$$

and

$$\text{var}(Y)=\sum_{k=0}^{n}(k-np)^2\times\Pr(Y=k)=npq.$$

The standard deviation is given by $\text{SD}(Y) = (npq)^{1/2}$. Thus if $Y \sim \text{bin}(12, 0.4)$, then $\text{E}(Y) = 4.8$ and $SD(Y) = 1.70$. The distribution is symmetric if $p = 1/2$, with the **skewness** increasing as p moves away from 1/2 in either direction. The variance and the standard deviation of the binomial distribution decrease as p deviates from 1/2, with the smallest variability near $p = 0$ or 1. As n increases, the binomial distribution gets more symmetric and more closely approximated by a **normal distribution**. Numerical values for $\Pr(Y = k)$ for selected values of n and p are available in published tables (see, for example, Rosner [14] or the references given by Johnson & Kotz [10]). When evaluating many binomial probabilities for the same n and p, it is convenient to use the recursion formula

$$\Pr(Y=k+1)=\left[\frac{n-k}{k+1}\right]\times\frac{p}{q}\times\Pr(Y=k),$$
$$k=0,1,\ldots,n-1.$$

If $Y_1 \sim \text{bin}(n_1, p)$ is independent of $Y_2 \sim \text{bin}(n_2, p)$, then $Y_1 + Y_2 \sim \text{bin}(n_1 + n_2, p)$ and

$$\Pr(Y_1=j|Y_1+Y_2=k)=\frac{\binom{n_1}{j}\binom{n_2}{k-j}}{\binom{n_1+n_2}{k}} \quad (1)$$

for $\max(0, k - n_2) \le j \le \min(n_1, k)$. The **conditional** distribution (1) is a (central) **hypergeometric distribution**.

Calculation of sums of binomial probabilities can be cumbersome, so approximations have been sought. Several approximations to the binomial distribution for large values of n are available. For $Y \sim \text{bin}(n, p)$, application of the **central limit theorem** gives

$$\Pr(Y \le k) \approx \Phi\left[\frac{k - np}{npq}^{1/2}\right], \qquad (2)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf). However, this approximation does not work well in practice and is not recommended. Comparing the histogram of the binomial probability mass function with the approximating normal density having the same mean np and variance npq suggests that a continuity correction of 1/2 may be appropriate, giving

$$\Pr(Y \le k) \approx \Phi\left[\frac{k + \frac{1}{2} - np}{npq}^{1/2}\right]. \qquad (3)$$

This approximation should work well when n is large and p is "central" (not too far from 1/2). For example, as a "rule of thumb" Rosner [14] suggests this approximation can be used when $npq > 5$. Approximating $\Pr(Y \le 8)$ for $n = 24$ and $p = 0.4$ gives 0.252 from (2) and 0.323 from (3), compared with the true value of 0.328. Other more accurate normal approximations, some based on the arc sine square root (or angular) transformation (*see* **Delta Method**), are available [8, 10–12]. Peizer & Pratt [13] developed an extremely accurate, though more complicated, normal approximation.

When n is large and p is small, the binomial distribution can be approximated by a **Poisson distribution**

$$\Pr(Y \le k) \approx \exp(-np) \sum_{j=0}^{k} \frac{(np)^j}{j!},$$

since in this case the mean and variance of the binomial distribution are similar to that of a Poisson distribution with mean np. As a rule of thumb, Rosner [14] suggests that this approximation may be adequate for $n \ge 100$ and $p \le 0.01$. Other more accurate approximations for small p are available [11].

Point Estimation of *p*

In most applications, the number of trials, n, is known but the success probability, p, is not. Then the natural estimate of the success probability, p, is the observed proportion of successes, y/n, which will be denoted by $\hat{p}$. This estimator is **unbiased**, since $\text{E}(\hat{p}) = p$, and is both the **method of moments** and the **maximum likelihood** estimator of p. Application of the **Cramér–Rao inequality** shows that $\hat{p}$ is also the uniformly **minimum variance unbiased estimator** of p.

Since $\text{var}(Y) = npq$ and $\hat{p} = y/n$, it follows that $\text{var}(\hat{p}) = pq/n$, which can be estimated by $\hat{p}\hat{q}/n$. A measure of the precision of $\hat{p}$ is given by its standard error, $(\hat{p}\hat{q}/n)^{1/2}$. Thus the precision increases with sample size.

It can be shown that $\hat{p}$ is admissible under both squared error loss and relative squared error loss (see, for example, [2] and [15]). Thus, there can be no uniformly better estimator than $\hat{p}$ under these loss functions, although other estimators may perform better over wide ranges of p. Santner & Duffy [15] review various **Bayesian** and related methods which incorporate prior knowledge about the unknown p for developing alternative estimators to $\hat{p}$.

When n is unknown, the maximum likelihood and method-of-moments estimators of n can be extremely sensitive to minor fluctuations in the data. Consideration of this problem has been given by Aitkin & Stanisopoulos [1], Casella [5], Hall [9], and the references therein.

Hypothesis Testing

On some occasions, it is of interest to test H_0: $p = p_0$ vs. the alternative H_a: $p \ne p_0$ (a two-sided alternative; *see* **Hypothesis Testing**). Whether or not to reject the null hypothesis, H_0, will depend on how far the sample proportion of successes, $\hat{p}$, is from p_0, as well as on the precision of $\hat{p}$. Using the central limit theorem and assuming that a normal assumption is reasonable (say when $np_0q_0 \ge 5$), under H_0 it follows that

$$\hat{p} \sim \text{N}\left(\frac{p_0,\, p_0q_0}{n}\right). \qquad (4)$$

It is more convenient to standardize $\hat{p}$ by subtracting the expected value under H_0 and dividing by the

standard error under H_0, thus creating the test statistic

$$z = \frac{\hat{p} - p_0}{\left(\frac{p_0 q_0}{n}\right)^{1/2}},$$

which has an approximately unit normal distribution (mean 0, variance 1) under H_0. For a two-sided α level test, we would reject H_0 when either $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$, where $\Phi(z_{1-\alpha}) = 1 - \alpha$. Alternatively, the p value for this hypothesis test is given by $2\Phi(z)$ when $\hat{p} < p_0$ or $2[1 - \Phi(z)]$ when $\hat{p} \geq p_0$.

When n is not large or p is "extreme" (close to 0 or 1), the normal approximation (4) will not perform well, and a more accurate normal approximation needs to be used. Alternatively, we could base the test on the exact binomial probabilities (*see* **Exact Inference for Categorical Data**). In this case, the p value is given by $\min(2\Pr(Y \leq y|p = p_0), 1)$ when $\hat{p} < p_0$ or $\min(2\Pr(Y \geq y|p = p_0), 1)$ when $\hat{p} \geq p_0$, where y is the observed number of successes in the sample.

These hypothesis tests can be modified to accommodate one-sided alternatives. Rosner [14] reviews **power** and **sample size estimation** issues for the one-sample binomial test.

Confidence Interval Estimation of *p*

In most applications it is more informative to calculate a **confidence interval** for p than to perform a hypothesis test. The most commonly used form of confidence interval is based on a normal approximation. Using $\hat{p} \sim N(p, pq/n)$, it follows that

$$\Pr\left(p - z_{1-\alpha/2}\left(\frac{pq}{n}\right)^{1/2} < \hat{p} < p + z_{1-\alpha/2}\left(\frac{pq}{n}\right)^{1/2}\right) = 1 - \alpha.$$

Approximating pq/n by $\hat{p}\hat{q}/n$ and rearranging the inequalities shows that an approximate two-sided $100(1 - \alpha)\%$ confidence interval for p is given by

$$\left(\hat{p} - z_{1-\alpha/2}\left(\frac{\hat{p}\hat{q}}{n}\right)^{1/2}, \hat{p} + z_{1-\alpha/2}\left(\frac{\hat{p}\hat{q}}{n}\right)^{1/2}\right). \quad (5)$$

This method of interval estimation, commonly called the Wald method, should only be used for n large and p "central", e.g. when $n\hat{p}\hat{q} > 5$ (see, for example, [14]). However, Vollset [16] maintains that this rule of thumb does not ensure adequate accuracy. Improved confidence intervals based on more accurate normal approximations are presented and reviewed by Blyth & Still [4] and Vollset [16]. Vollset suggests that score test-based confidence intervals for p, based on inversion of the test without estimating the standard error, have much better (closer to nominal) coverage probabilities than many other methods, particularly the intervals (5), and are only slightly more difficult to calculate.

When use of a normal approximation is not valid, one can resort to the binomial distribution to construct exact confidence intervals. Clopper & Pearson [6] proposed an exact two-sided $100(1 - \alpha)\%$ confidence interval for p of the form (p_L, p_U), where p_L and p_U satisfy $\Pr(Y \geq y|p = p_L) = \alpha/2$ and $\Pr(Y \leq y|p = p_U) = \alpha/2$. These intervals, often called *tail intervals*, have attractive symmetry and monotonicity properties, but require iterative calculations. Using the relationships between binomial, beta, and ***F* distributions** gives an alternative form for these limits, namely

$$p_L = \frac{y}{[y + (n - y + 1) \times F_{1-\alpha/2}(2(n - y + 1), 2y)]}, \quad \text{for } 1 \leq y \leq n$$

(with $p_L = 0$ for $y = 0$) and

$$p_U = \frac{y + 1}{\left\{\frac{y + 1 + (n - y)}{F_{1-\alpha/2}[2(y + 1), 2(n - y)]}\right\}}, \quad \text{for } 0 \leq y \leq n - 1$$

(with $p_U = 1$ for $y = n$). These forms may be attractive if one has access to tables of the F distribution.

Although the tail intervals have intuitive appeal, they are extremely conservative. That is, although the intervals are guaranteed to have at least $100(1 - \alpha)\%$ coverage for all values of p, the actual coverage probabilities are often much greater than the nominal level. Blyth & Still [4] review various exact confidence intervals for p and present tables of less conservative exact 95% and 99% confidence intervals when $n \leq 30$.

Construction of one-sided confidence intervals for p proceeds in a similar fashion as above. Exact one-sided confidence intervals can be derived from the tail intervals. Blyth [3] examines approximate one-sided confidence bounds.

Vollset [16] provides a detailed comparison of the operating characteristics of various two-sided confidence intervals for p, including the Wald intervals, exact intervals, and other approximate confidence intervals. He suggests that only the exact intervals and the score test-based confidence intervals approximately achieve the desired coverage levels. In particular, Vollset recommends the continuity corrected score test-based confidence intervals for p, as they are less conservative and less tedious to calculate than exact intervals, but have coverage close to the nominal levels.

Extensions

In certain applications it is convenient to think in terms of the **odds** of success, $p/(1-p)$, or the log odds, rather than the probability of success, p. Thus, an odds of 2 means that the probability of success is twice the probability of failure, or that $p = 2/3$. The odds are particularly useful when comparing two binomial proportions via the **odds ratio**, or when modeling the log odds of success as a linear function of covariates in **logistic regression**.

The **multinomial distribution** is a generalization of the binomial distribution which allows responses with more than two outcomes. The **negative binomial distribution** arises as the distribution of the number of failures of independent Bernoulli trials until a prespecified number of successes is obtained. Correlated binary responses could be modeled with parametric assumptions, e.g. using the **beta-binomial distribution**, or via methods for **overdispersion**.

The books by Collett [7] and Rosner [14] review basic properties of the binomial distribution. Johnson & Kotz [10] and Santner & Duffy [15] give many more details.

References

[1] Aitkin, M. & Stanisopoulos, M. (1989). Likelihood analysis of a binomial sample size problem, in *Contributions to Probability and Statistics*, L.J. Gleser, M.D. Perlman, S.J. Press & A.R. Sampson, eds. Springer-Verlag, New York, pp. 399–411.

[2] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed. Springer-Verlag, New York, Chapter 4.

[3] Blyth, C.R. (1986). Approximate binomial confidence limits, *Journal of the American Statistical Association* **81**, 843–855.

[4] Blyth, C.R. & Still, H.A. (1983). Binomial confidence intervals, *Journal of the American Statistical Association* **78**, 108–116.

[5] Casella, G. (1986). Stabilizing binomial n estimators, *Journal of the American Statistical Association* **81**, 172–175.

[6] Clopper, C.J. & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial, *Biometrika* **26**, 404–413.

[7] Collett, D. (1991). *Modeling Binary Data*. Chapman & Hall, London, Chapter 2.

[8] Freeman, M.F. & Tukey, J.W. (1950). Transformations related to the angular and the square root, *Annals of Mathematical Statistics* **21**, 607–611.

[9] Hall, P. (1994). On the erratic behavior of estimators of N in the binomial N, p distribution, *Journal of the American Statistical Association* **89**, 344–352.

[10] Johnson, N.L. & Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Wiley, New York, Chapter 3.

[11] Molenaar, W. (1970). *Approximations to the Poisson, Binomial, and Hypergeometric Distribution Functions*. Mathematisch Centrum, Amsterdam.

[12] Molenaar, W. (1973). Simple approximations to the Poisson, binomial, and hypergeometric distributions, *Biometrics* **29**, 403–407.

[13] Peizer, D.B. & Pratt, F. (1968). A normal approximation for binomial, F, beta, and other common related tail probabilities, I, *Journal of the American Statistical Association* **63**, 1416–1456.

[14] Rosner, B. (1995). *Fundamentals of Biostatistics*, 4th Ed. Duxbury Press, Belmont, Chapters 4–7.

[15] Santner, T.J. & Duffy, D.E. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York, Chapters 1 and 2.

[16] Vollset, S.E. (1993). Confidence intervals for a binomial proportion, *Statistics in Medicine* **12**, 809–824.

(*See also* **Categorical Data Analysis**; **Two-by-Two Table**)

DAVID WYPIJ

Bioassay

This term is an abbreviation of, and is effectively synonymous with, **biological assay**. In the classical form of biological assay, an experiment is conducted on biological material, to determine the relative potency of test and standard preparations. In recent decades, the term *bioassay* has been used in a more general sense, to denote any experiment in which responses to various doses of externally applied agents are observed in animals or some other biological system. The emphasis here is to measure the effects of various agents on the response variable, and no attempt is made to estimate a relative potency. This usage is common, for instance, in programmes for screening substances for possible carcinogenic effects. Such experiments may include "negative controls" (inert materials) and "positive" controls (known carcinogens), but these controls are used to validate the experimental system rather than as a basis for the estimation of relative potency.

(*See also* **Animal Screening Systems**; **Extrapolation, Low Dose**; **Extrapolation**)

PETER ARMITAGE

Bioavailability and Bioequivalence

The US Code of Federal Regulations (CFR) defines the *bioavailability* of a drug product as the rate and extent to which the active drug ingredient or therapeutic moiety of the drug product is absorbed and becomes available at the site of drug action. Bioavailability studies are usually conducted to assess the pharmacological characteristics of a new drug product during phase I clinical development and to serve as a surrogate for the clinical evaluation of generic drug products. A comparative bioavailability study refers to the comparison of bioavailabilities of different formulations of the same drug (e.g. tablets vs. capsules) or different drug products (e.g. a generic drug vs. a brand-name drug). For the approval of generic drugs, a **bioequivalence** assessment, as a surrogate for the clinical evaluation of the generic drug products, is based on the following *fundamental bioequivalence assumption*. That is, when two formulations of the same drug, or different drug products, are claimed bioequivalent it is assumed that they are therapeutically equivalent [5]. Note that the US Food and Drug Administration (FDA) does not require a complete new drug application (NDA) submission for the approval of a generic drug if the sponsor can provide evidence of bioequivalence in average bioavailability between the generic drug and the brand-name drug through an abbreviated new drug application (ANDA) from bioequivalence studies.

The concepts of bioavailability and bioequivalence did not become public issues until the late 1960s when concern was raised that a generic drug product might not be as bioavailable as that manufactured by the innovator. In 1970, the FDA began to request evidence of biological availability in applications submitted for the approval of certain new drugs. In 1974, a Drug Bioequivalence Study Panel was formed by the US Office of Technology Assessment (OTA) to examine the relationship between the chemical and therapeutic equivalence of drug products. On the basis of the recommendations in the OTA report, the FDA published a set of regulations for the submission of bioavailability data in certain new drug applications. These regulations became effective on July 1, 1977, and are currently codified in 21 CFR Part 320.

In 1984, the FDA was authorized to approve generic drug products under the Drug Price Competition and Patent Term Restoration Act. In recent years, as more generic drug products have become available, there has been concern that generic drug products may not be comparable in identity, strength, quality or purity to the innovator drug product. To address this concern, the FDA conducted a hearing on Bioequivalence of Solid Oral Dosage Forms in Washington, DC in 1986. As a consequence of the hearing, a Bioequivalence Task Force (BTF) was formed to evaluate the current procedures adopted by the FDA for the assessment of bioequivalence between immediate solid oral dosage forms. The BTP report was issued in January 1988. Based on the recommendations of the report by the BTF, guidance on statistical procedures for bioequivalence studies was issued by the FDA Division of Bioequivalence Office of Generic Drugs in 1992 [12].

Pharmacokinetic Parameters

In bioavailability studies, the rate and extent of drug absorption are usually characterized by some **pharmacokinetic** parameters, such as the area under the blood or plasma concentration–time curve (AUC), maximum concentration ($C_{\max}$), time to reach maximum concentration ($t_{\max}$), elimination half-life ($t_{1/2}$), and rate constant (k_e). Gibaldi & Perrier [16] provided a comprehensive overview of these pharmacokinetic parameters. Among these pharmacokinetic parameters, AUC is considered the primary measure for the extent of absorption, which provides information regarding the total amount of the drug absorbed in the body. For comparative bioavailability studies, bioequivalence is usually assessed by bioequivalence measures, such as the difference in means or the ratio of means.

Between 1977 and 1980, the FDA proposed a number of decision rules for assessing bioequivalence in average bioavailability [25]. These decision rules include the 75/75 rule, the 80/20 rule, and the ± 20 rule. The 75/75 rule claims bioequivalence if at least 75% of individual subject ratios (i.e. relative individual bioavailability of the generic (test) product to the innovator (reference) product) are within (75%,

125%) limits. The 80/20 rule concludes bioequivalence if the test average is not statistically significantly different from the reference average and if there is at least 80% power for detection of a 20% difference of the reference average. The BTF does not recommend these two decision rules because of their undesirable statistical properties. The ±20 rule suggests that two drug products are bioequivalent if the average bioavailability of the test product is within ±20% of that of the reference product with a certain assurance (say, 90%). Recently, the FDA guidance recommended an 80/125 rule for log-transformed data. The 80/125 rule claims bioequivalence if the ratio of the averages between the test product and the reference product falls within (80%, 125%) with a 90% assurance.

The ±20 rule and the 80/125 rule are currently acceptable to the FDA for the assessment of bioequivalence in average bioavailability. Current FDA guidance suggests that the 80/125 rule be used for pharmacokinetic parameters such as AUC and $C_{\max}$ after log-transformation. The guidance, however, does not indicate which criterion should be used for other pharmacokinetic parameters. It should be noted that, based on current practice of bioequivalence assessment, either the ±20 rule or the 80/125 rule can be applied to all pharmacokinetic parameters and all drug products across all therapeutic areas.

Designs of Bioavailability Studies

The *Federal Register* [15] indicated that a bioavailability study (single-dose or multiple-dose) should be crossover in design. A **crossover design** is a modified randomized block design in which each block (i.e. subject) receives more than one formulation of a drug at different time periods. The most commonly used study design for the assessment of bioequivalence between reference (R) and test (T) formulations is a two-sequence, two-period crossover design, denoted by (RT, TR), which is also known as the standard crossover design. For the standard crossover design, each subject is randomly assigned to either sequence 1 (R–T) or sequence 2 (T–R). In other words, subjects within sequence R–T (T–R) receive formulation R(T) during the first dosing period and formulation T(R) during the second dosing period. Usually, the dosing periods are separated by a washout period of sufficient length for the drug received in the first period to be completely metabolized and/or excreted by the body.

In practice, when differential carry-over effects are present, the standard crossover design may not be useful because the formulation effect is confounded with the carry-over effect. In addition, the standard crossover design does not provide independent estimates of intrasubject variability for each formulation, because each subject only receives each formulation once. To overcome these drawbacks, Chow & Liu [6] recommend a higher-order crossover design be used. A higher-order design is defined as a crossover design in which either the number of periods or the number of sequences is greater than the number of formulations to be compared. The commonly used higher-order crossover designs include Balaam's design (TT, RR, RT, TR), the two-sequence dual design (TRR, RTT), and the optimal four-sequence design [(TRRT, RTTR) or (TTRR, RRTT, TRRT, RTTR)]. For comparing more than two formulations, Jones & Kenward [19] recommend that a Williams design be used. For example, for comparing three formulations (R, T_1, and T_2), the design (RT_2T_1, T_1RT_2, T_2T_1R, T_1T_2R, T_2RT_1, RT_1T_2) is useful.

For bioequivalence trials, a traditional approach for **sample size determination** is to conduct a **power** analysis based on the 80/20 decision rule. This approach, however, is based on point hypotheses rather than interval hypotheses and, therefore, may not be statistically valid [5]. Phillips [24] provides a table of sample sizes that are based on power calculation of Schuirmann's two one-sided tests procedure using the bivariate **noncentral *t* distribution**. However, no formulas are provided. An approximate formula for sample size calculations is provided in [20].

Statistical Methods

For a standard two-sequence, two-period crossover design, let Y_{ijk} denote the response (e.g. logarithm of AUC) of the ith subject in the kth sequence at the jth period, where $i = 1, \ldots, n_k$, $k = 1, 2$, and $j = 1, 2$. Under the assumption that there are no carry-over effects, Y_{ijk} can be described by the following statistical model:

$$Y_{ijk} = \mu + S_{ik} + F_{(j,k)} + P_j + e_{ijk},$$

where μ is the overall mean; S_{ik} is the random effect of the ith subject in the kth sequence; P_j is the fixed

effect of the jth period; $F_{(j,k)}$ is the direct fixed effect of the formulation in the kth sequence, which is administered at the jth period; and e_{ijk} is the within-subject random error in observing Y_{ijk}.

The commonly used approach for assessing bioequivalence in average bioavailability is the method of the classical (shortest) **confidence interval**. Let μ_T and μ_R be the mean of test and reference formulation, respectively. Then, under a normality assumption, the classical $(1-2\alpha)\times 100\%$ confidence interval for $\mu_T - \mu_R$ can be obtained as follows:

$$L = (\overline{Y}_T - \overline{Y}_R) - t(\alpha, n_1+n_2-2)\hat{\sigma}_d \times \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2},$$

$$U = (\overline{Y}_T - \overline{Y}_R) + t(\alpha, n_1+n_2-2)\hat{\sigma}_d \times \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2},$$

where $\overline{Y}_T$ and $\overline{Y}_R$ are least-squares means for the test and reference formulations, $t(\alpha, n_1+n_2-2)$ is the upper αth critical value of a t distribution with n_1+n_2-2 degrees of freedom, and $\hat{\sigma}_d^2$ is given by

$$\hat{\sigma}_d^2 = \frac{1}{n_1+n_2-2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(d_{ik}-\overline{d}_{\cdot k})^2,$$

where

$$d_{ik} = \tfrac{1}{2}(Y_{i2k}-Y_{i1k}) \quad \text{and} \quad \overline{d}_{\cdot k} = \frac{1}{n_k}\sum_{i=1}^{n_k} d_{ik}.$$

According to the 80/125 rule, if the exponentiations of L and U are within (80%, 125%), then the two formulations are bioequivalent.

On the basis of the interval hypotheses, Schuirmann [27] proposed a procedure consisting of two one-sided tests to evaluate whether the bioavailability of the test formulation is too high (safety) for one side and is too low (efficacy) for the other side. Thus, we conclude that the two formulations are bioequivalent if

$$T_L = \frac{(\overline{Y}_T - \overline{Y}_R) - \theta_L}{\hat{\sigma}_d\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2}} > t(\alpha, n_1+n_2-2)$$

and

$$T_U = \frac{(\overline{Y}_T - \overline{Y}_R) - \theta_U}{\hat{\sigma}_d\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2}} < -t(\alpha, n_1+n_2-2),$$

where $\theta_L = \ln(0.8) = -0.2231$ and $\theta_U = \ln(1.25) = 0.2231$ are the limits for bioequivalence. Note that the confidence interval approach is operationally equivalent to Schuirmann's two one-sided tests procedure [5].

Several methods have been proposed for the assessment of average bioequivalence. These methods include the Westlake symmetric confidence interval [29], Chow & Shao's joint confidence region approach [9], Anderson & Hauck's test for interval hypotheses [1], the Bayesian approach for the highest posterior density (HPD) interval, and nonparametric methods [5, 10, 17]. Note that some of these methods are operationally equivalent in the sense that they will reach the same decision on bioequivalence. More details can be found in [5].

Drug Interchangeability

In recent years, as more generic drug products have become available, the efficacy, safety, and quality of generic drug products have become issues of public concern in health care. However, for the approval of a generic drug, regulatory agencies such as the FDA require only that a bioequivalence trial be conducted to provide evidence of bioequivalence in average bioavailability. An approved generic drug can be used as a substitute for the innovator drug product. The regulatory agencies, however, do not require that bioequivalence among generic drugs be provided. Therefore, whether the brand-name drug with its many generic copies can be used interchangeably is an issue of great regulatory and scientific concern.

Basically, drug interchangeability can be classified as drug prescribability or drug switchability. Drug prescribability is referred to as the physician's choice for a new patient, when prescribing an appropriate drug product, between a brand-name drug and a number of generic copies shown to be bioequivalent. Under current regulation for average bioequivalence, Chow & Liu [8] suggest that one should perform a meta-analysis for post-approval bioequivalence review to ensure drug prescribability. The idea is to assess bioequivalence among generic drugs

based on individual bioequivalence submissions. It is suggested that a warning be issued if a significant bioinequivalence is observed between any two generic drug products. To ensure drug prescribability prospectively, many researchers have recommended that, in addition to bioequivalence in average bioavailability, bioequivalence be established in the variability of bioavailability between generic drug products and the brand-name drug product [21, 23]. This concept is known as *population bioequivalence*.

Drug switchability is related to the switch from a drug product (e.g. a brand-name drug) to an alternative product (e.g. a generic copy of the brand-name drug) within the same subject, whose concentration of the drug has been titrated to a steady, efficacious and safe level. To ensure drug switchability, it is necessary to establish bioequivalence within each individual. This concept is known as *individual bioequivalence* [2]. In the past few years, several methods with different criteria for bioequivalence within each subject have been proposed. These methods and criteria can be classified as either probability-based [1, 11] or moment-based [18, 26, 28]. These methods, however, fail to evaluate adequately the equivalence between distributions within the same individual. Recently, although the use of individual bioequivalence as an alternative regulatory requirement for assessment of bioequivalence has attracted much attention [13, 14], many regulatory, scientific, and practical issues remain unresolved [3, 7].

Other Issues

To account for the variability of bioavailability and assess drug interchangeability, it is recommended that a replicated crossover design be used [4, 22]. In addition, the FDA is seeking alternative pharmacokinetic parameters, decision rules and statistical methods for population and individual bioequivalence. Some unresolved scientific issues of particular interest include the impact of add-on subjects for drop-outs, the use of female subjects in bioequivalence trials, *in vitro* dissolution as a surrogate for *in vivo* bioequivalence, post-approval bioequivalence, and international harmonization for bioequivalence requirements among the European Community, Japan, and the US. A comprehensive overview of these issues can be found in [5].

References

[1] Anderson, S. & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials, *Communication in Statistics – Theory and Methods* **12**, 2663–2692.

[2] Anderson, S. & Hauck, W.W. (1990). Consideration of individual bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics* **18**, 259–273.

[3] Chen, M.L. (1997). Individual bioequivalence – a regulatory update, *Journal of Biopharmaceutical Statistics* **7**, 5–11.

[4] Chow, S.C. (1996). Statistical considerations for replicated design, *Proceedings of the FIP BIO International 1996*, Business Center for Academic Societies, Tokyo, Japan, pp. 107–112.

[5] Chow, S.C. & Liu, J.P. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies.* Second Edition, Revised and Expanded, Marcel Dekker, New York.

[6] Chow, S.C. & Liu, J.P. (1992). On assessment of bioequivalence under a higher-order crossover design, *Journal of Biopharmaceutical Statistics* **2**, 239–256.

[7] Chow, S.C. & Liu, J.P. (1995). Current issues in bioequivalence trials, *Drug Information Journal* **29**, 795–804.

[8] Chow, S.C. & Liu, J.P. (1997). Meta-analysis for bioequivalence review, *Journal of Biopharmaceutical Statistics* **7**, 97–111.

[9] Chow, S.C. & Shao, J. (1990). An alternative approach for the assessment of bioequivalence between two formulations of a drug, *Biometrical Journal* **32**, 969–976.

[10] Cornell, R.G. (1990). The Evaluation of Bioequivalence Using Nonparametric Statistics in *Drug Absorption and Disposition: Statistical Considerations*, K.S. Albert, ed. American Pharmaceutical Association, Academy of Pharmaceutical Sciences, Washington, pp. 51–57.

[11] Esinhart, J.D. & Chinchilli, V.M. (1994). Extension to the use of tolerance intervals for assessment of individual bioequivalence, *Journal of Biopharmaceutical Statistics* **4**, 39–52.

[12] FDA (1992). *Guidance on Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design. Division of Bioequivalence*, Office of Generic Drugs, Food and Drug Administration, Rockville.

[13] FDA (2001). *Guidance on Statistical Approaches to Establishing Bioequivalence*, Food and Drug Administration, Rockville.

[14] FDA (2003). *Guidance on Bioavailability and Bioequivalence Studies for Orally Administrated Drug Products-General Considerations*, Food and Drug Administration, Rockville.

[15] *Federal Register* (1977). Vol. 42 No. 5, Sections 320. 26(b). Marcel Dekker, New York.

[16] Gibaldi, M. & Perrier, D. (1982). *Pharmacokinetics.* Marcel Dekker, New York.

[17] Hauschke, D., Steinijans, V.W. & Diletti, E. (1990). A distribution-free procedure for the statistical analysis of bioequivalence studies, *International Journal of Clinical Pharmacology, Therapy and Toxicology* **28**, 72–78.

[18] Hyslop, T.F., Hsuan, F., & Holder, D.J. (2000). A small-sample confidence interval approach to assess individual bioequivalence, *Statistics in Medicine*, **19**, 2885–2897.

[19] Jones, B. & Kenward, M.G. (1989). *Design and Analysis of Crossover Trials*. Chapman & Hall, London.

[20] Liu, J.P. & Chow, S.C. (1992). On power calculation of Schuirmann's two one-sided tests procedure in bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics* **20**, 101–104.

[21] Liu, J.P. & Chow, S.C. (1992). On assessment of bioequivalence in variability of bioavailability, *Communications in Statistics – Theory and Methods* **21**, 2591–2608.

[22] Liu, J.P. (1995). Use of the repeated cross-over designs in assessing bioequivalence, *Statistics in Medicine* **14**, 1067–1078.

[23] Metzler, C.M. & Huang, D.C. (1983). Statistical methods for bioavailability and bioequivalence, *Clinical Research Practices and Drug Regulation Affairs* **1**, 109–132.

[24] Phillips, K.F. (1990). Power of the two one-sided tests procedure in bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics* **18**, 137–144.

[25] Purick, E. (1980). Bioavailability/bioequivalency regulations: an FDA perspective, in *Drug Absorption and Disposition: Statistical Considerations*, K.S. Albert, ed. American Pharmaceutical Association, Academy of Pharmaceutical Sciences, Washington, pp. 115–137.

[26] Schall, R. & Luus, H.G. (1993). On population and individual bioequivalence, *Statistics in Medicine* **12**, 1109–1124.

[27] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.

[28] Sheiner, L.B. (1992). Bioequivalence revisited, *Statistics in Medicine* **11**, 1777–1788.

[29] Westlake, W.J. (1976). Symmetrical confidence intervals for bioequivalence trials, *Biometrics* **32**, 741–744.

SHEIN-CHUNG CHOW & JEN-PEI LIU

Bioequivalence

Two drug formulations are bioequivalent if their absorption characteristics are closely similar. The most important characteristics are the extent and rate of absorption, which together define the **bioavailability** of a drug formulation. Thus, *bioequivalence* is the comparable bioavailability of drug formulations.

Kinetic Measures of Bioavailability

Measures After a Single Drug Administration

When drugs are not administered directly into the systemic circulation, as in the case of oral intake, at least a fraction of the dose should first be absorbed. As a result, the concentration in blood and plasma initially rises (Figure 1). The drug is then distributed to various tissues and also starts being eliminated from the body. Consequently, after reaching a peak, the concentration declines (Figure 1).

The principal measures (metrics) characterizing bioavailability, and therefore bioequivalence, are the area under the curve (AUC) contrasting concentration with time, and the maximum concentration, $C_{\max}$, which is observed at the time of $T_{\max}$ (Figure 1). AUC is a measure of the extent of absorption. AUC rises as the absorbed fraction of drug dose, i.e. the amount reaching the circulation, increases.

$C_{\max}$ is the most widely used index for the evaluation of absorption rates, especially in comparative studies. $C_{\max}$ actually reflects several processes of drug disposition, including the extent of absorption. However, if, in two drug formulations, all processes except the absorption rate have the same magnitudes (often a good assumption, at least approximately), then differences of $C_{\max}$ values indeed indicate deviations between absorption rates. A higher $C_{\max}$ signals (everything else being equal) a faster absorption rate.

$T_{\max}$ is also determined by various processes of drug disposition. Consequently, contrasts of $T_{\max}$ reflect deviations between absorption rates only if magnitudes characterizing other processes of drug disposition remain the same. Under this condition, a smaller $T_{\max}$ indicates a higher absorption rate.

AUC, $C_{\max}$, and $T_{\max}$ are often referred to as model-free measures of bioavailability because they are determined independently of assumed **pharmacokinetic** models.

Measures After Repeated Drug Administrations

Following repeated administrations of a given drug dose, D, the concentration in plasma and blood eventually reaches a quasi-steady state. In this condition, after each administration of the drug, the concentration first rises, passes through a maximum, $C_{\max}$, and then declines toward a minimum or so-called trough value, $C_{\min}$. The time profile of concentrations is illustrated in Figure 2.

The kinetic measures of bioavailability parallel in the steady state the metrics noted for single administration. AUC is again an index of the extent of absorption. However, it is evaluated after repeated administrations by measuring the area recorded during the time interval between two dosings (the dosing or maintenance interval, T). This AUC is numerically identical to the value obtained following a single drug administration, provided that the drug exhibits first-order kinetics (i.e. the rate of change in the concentration is proportional to the concentration) and that this is identical under the two conditions.

Absorption rates are usually less important when considering repeated drug administrations than after a single drug administration. Nevertheless, $C_{\max}$ is still important, especially as an index of drug safety; unusually high concentrations could indicate a danger of toxicity.

Additional measures of bioavailability are considered later.

Assessment of Bioequivalence

Several statisticians have made distinguished contributions since the early 1970s to developing the methodology for the assessment of bioequivalence. They notably include Carl M. Metzler and Wilfred J. Westlake and also, among others, Sharon Anderson, Walter W. Hauck, Jochen Mau, & Bruce E. Rodda. Chow & Liu [5] reviewed the statistical aspects of the evaluation of bioequivalence.

Procedures will be briefly described which are applied most widely at present. Sauter et al. [10] illustrated examples for detailed calculations evaluating bioequivalence following single drug administrations and in the steady state.

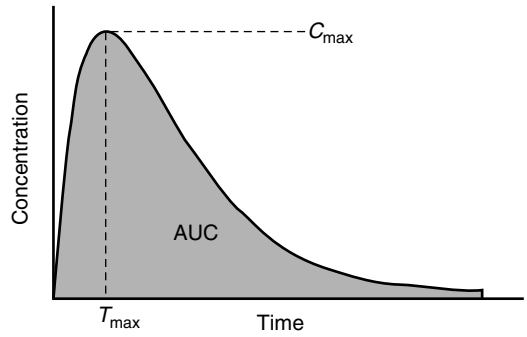

Figure 1 Time course of plasma concentration following a single oral administration of a drug

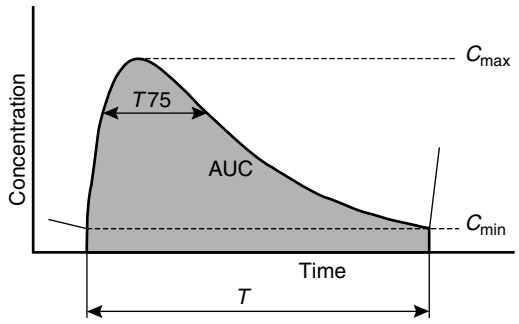

Figure 2 Time course of plasma concentration following repeated oral administrations of a drug

Two One-Sided Tests Procedure

The two one-sided tests (*see* **Alternative Hypothesis**) procedure is nowadays widely applied for the assessment of bioequivalence. Yee [20] presented some features of the approach and Schuirmann [14] elaborated them. The goal is to compare the **mean** kinetic responses of a reference (R) formulation and an investigated test (T) drug product. The **null hypothesis** to be tested is that of bioinequivalence. The test is subdivided into two one-sided problems. They assume bioinequivalence if the difference between the means is either less than or equal to a regulatory criterion θ_1 or (/and) larger than or equal to another regulatory value θ_2:

$$H_{01}: \mu_T - \mu_R \leq \theta_1; \qquad H_{02}: \mu_T - \mu_R \geq \theta_2. \quad (1)$$

The null hypothesis could be rejected in favor of an alternative hypothesis indicating bioequivalence:

$$H_a: \theta_1 < \mu_T - \mu_R < \theta_2. \quad (2)$$

If both null hypotheses are true, their evaluation at the $\alpha/2$ significance level indicates a consumer risk on both sides with this probability.

The two null hypotheses are in practice assessed by the use of **confidence intervals**. Various approaches have been proposed. The application of the shortest interval at the α level [14, 18] has been widely adopted. It yields, at a given significance level (and consumer risk), the highest **power**, i.e. the smallest risk for producers when the two formulations are in fact bioequivalent [16].

Implementation of the Two One-Sided Tests Procedure

Most kinetic quantities are considered to have multiplicative character (their multiplication and division – e.g. whether their magnitudes should be raised or lowered by a factor of 2 or 10 – appear to be relevant). Correspondingly, their errors are also thought to be multiplicative (*see* **Multiplicative Model**) and not additive [16, 19]. Therefore, they are typically evaluated and compared in their logarithmic form. Consequently, μ_T and μ_R are estimated after the logarithmic **transformation** of the investigated quantity, e.g. from log AUC or log C_{max}. The regulatory limits are usually considered to be symmetrical. Consequently, $\theta_2 = -\theta_1$ in the logarithmic scale. Times of the observations are not regarded to have multiplicative character. Moreover, they are recorded only at discrete sampling points. Therefore, T_{max} is generally not transformed and not assessed with regulatory limits. $\alpha = 0.10$ is widely applied.

The bioequivalence of two drug formulations is evaluated generally in two-period, two-sequence **crossover** trials. The kinetic responses (e.g. AUC) estimated for both formulations are contrasted in each subject. From the difference of individual logarithmic responses, their average and its 90% confidence interval are calculated. Bioequivalence is declared if the limits are in the regulatory range, between θ_1 and θ_2. The values of θ_1 and θ_2 are considered below.

Distribution-Free Procedure

Hauschke et al. [8] described a **nonparametric** procedure which, in its implementation of the two-sided hypotheses, took into account the structure of two-period, two-sequence, crossover studies. The kinetic

parameters recorded with the test and reference formulations (X_T and X_R) are contrasted within each individual:

$$X_i = X_{Ti} - X_{Ri}, \quad i = 1, \ldots, n.$$

Let the numbers of subjects in the two sequences of drug administration be n_1 and n_2, with $n_1 + n_2 = n$. A total of $n_1 n_2$ differences in the two sequences,

$$X_{j1} - X_{j^*2}, \quad j = 1, \ldots, n_1, \, j^* = 1, \ldots, n_2,$$

are formed and sorted by magnitude. The median of the pairwide differences is a Hodges–Lehmann point estimator (*see* **Estimation**) of $2(\mu_T - \mu_R)$. Ninety percent confidence limits are also obtained from the ranked differences.

Rapidly Evolving Issues

Various issues remain unresolved about the evaluation of bioequivalence. Two important topics are discussed which have particular relevance in biostatistics. They are developing rapidly and, therefore, their resolution in the near future is anticipated.

Individual Bioequivalence

The criteria discussed so far for the acceptance of bioequivalence involve the comparison of average kinetic parameters. Thus, the resulting similarity of two drug formulations is referred to as *average bioequivalence*. It ensures that the efficacy and safety of the new drug product is, on average, similar to that of the reference formulation. It is recognized that the efficacy and safety of the reference product was thoroughly evaluated during its development.

Average bioequivalence ensures that an individual who had not been exposed to either drug formulation, would have generally similar responses to both products. Thereby the *prescribability* of the test formulation is demonstrated. When, as often happens, patients are already receiving the reference product, then its substitution by the test formulation is contemplated. The issue is therefore the *switchability* from one formulation to another [2]. Thus, the similarity of responses and kinetic parameters *within* individuals, and therefore *individual bioequivalence*, becomes important.

To assess individual bioequivalence, the intrasubject variances of the reference and test products (σ^2_{WR} and σ^2_{WT}) need to be estimated. This can be accomplished if three- or four-period crossover trials are conducted. This design also enables the estimation of the subject × formulation **interaction** (σ^2_{SF}), which can be a major source of lack of individual bioequivalence.

The procedures proposed by several authors for the evaluation of individual bioequivalence can be separated into two principal categories [1, 11]. Probability-based approaches assume that deviations between individual kinetic parameters for the two formulations (X_R and X_T) would remain within a specified range and have a probability $\Pr(|X_T - X_R| \leq r)$, which should be sufficiently high. Moment-based procedures consider the second moment of $X_T - X_R$. A measure would extend that given for average bioequivalence, (2), by including terms for σ^2_{SF}, σ^2_{WR}, and σ^2_{WT}. A one-sided criterion is generally

$$(\mu_T - \mu_R)^2 + \delta\sigma^2_{SF} + \phi\sigma^2_{WR} + \gamma\sigma^2_{WT} < \theta_I. \qquad (3)$$

The regulatory criterion θ_I as well as the coefficients δ, ϕ, and η will have to be determined. For instance, it has been suggested that $\delta = \phi = 1$ and $\gamma = -1$ [13]. Sheiner [15] recommended $\delta = \gamma = 1$ and $\phi = 0$. Also other values have been proposed for the coefficients.

An unscaled measure such as that given above can be divided by a combination of σ^2_{WR} and σ^2_{WT}. The resulting scaled bioequivalence criterion differs intrinsically from its unscaled counterpart. It is anticipated that scaled comparisons will be particularly useful for assessing the bioequivalence of drugs which exhibit high intraindividual variability; the analysis of their equivalence is very difficult and at times even impossible by applying the methodology of average bioequivalence.

Similar conclusions can be drawn for probability-based measures of individual bioequivalence since they have close relationships to the corresponding moment-based measures [9, 12].

Kinetic Measures for the Evaluation of Bioequivalence

There is general agreement that AUCs measure well the extent of absorption, and that the application of

the two one-sided tests procedure to relative AUCs appropriately evaluates the equivalence of extents of absorption of two drug formulations. Consequently, two drug products are considered to be bioequivalent by this criterion if the 90% confidence interval around the geometric average of individual AUC ratios is between 0.80 and 1.25 [3].

A similar consensus has not been reached about metrics evaluating the equivalence of absorption rates. C_{max} is used most frequently. The various regulatory agencies apply differing criteria to indicate equivalence. For example, the **Food and Drug Administration** in the US has requirements for C_{max}s which parallel those for AUCs: the 90% confidence interval around the geometric average of individual C_{max} ratios should be between 0.80 and 1.25. The European Union recognizes that C_{max} is determined generally with larger variation than AUC. Therefore, its condition for the equivalence of C_{max}s is less demanding: the stated 90% confidence limits should be between 0.70 and 1.43. Canadian regulatory requirements do not invoke confidence limits and are, therefore, even less demanding: it is sufficient for the declaration of bioequivalence if the geometric average of individual C_{max} ratios is between 0.80 and 1.25.

In addition to the diversity of regulatory criteria, questions have been raised about the usefulness of C_{max} for determining the equivalence of absorption rates. C_{max} notably reflects various kinetic quantities and processes including the extent of absorption. Moreover, C_{max} responds to changes in absorption rates very insensitively, particularly in the steady state. The statistical properties of the metric are also unfavorable. Interestingly, its variation in the steady state can be higher or lower than that observed after a single drug administration, depending on the contributions of various sources of variation [21].

Therefore, reasonably, alternative measures have been suggested for assessing the equivalence of absorption rates after a single drug administration. They include, following a single drug administration, C_{max}/AUC, AUC measured until the peak of the reference formulation (partial AUC), and the intercept obtained by linear extrapolation from the ratios of concentrations of the two formulations measured in the early stage of a study [4, 6, 7].

After repeated drug administrations, in the steady state, the most frequently applied metric is, in addition to C_{max}, the peak–trough fluctuation, PTF; $\mathrm{PTF} = (C_{max} - C_{min})/C_{ave}$, where C_{ave} and C_{min} are the average and minimum concentrations, respectively, during a dosing interval [17]. Other measures include the Swing $[(C_{max} - C_{min})/(C_{min})]$, the AUC above C_{ave} normalized by the total AUC within a dosing interval (AUCF), and the duration (T75) of the concentration peak at the level of 3/4 of its adjusted height, i.e. at $0.75C_{max} + 0.25C_{min}$ (see Figure 2).

Little is known at present about the properties of these metrics, which are being explored extensively by several investigators.

References

[1] Anderson, S. (1993). Individual bioequivalence: a problem of switchability (with discussion), *Biopharmaceutics Report* **2**(2), 1–11.

[2] Anderson, S. & Hauck, W.W. (1990). Considerations of individual bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics* **18**, 259–273.

[3] Cartwright, A.C., Gundert-Remy, U., Rauws, G., McGilveray, I., Salmonson, T. & Walters, S. (1991). International harmonization and consensus DIA meeting on bioavailability and bioequivalence testing requirements and standards, *Drug Information Journal* **25**, 471–482.

[4] Chen, M.L. (1992). An alternative approach for assessment of rate of absorption in bioequivalence studies, *Pharmaceutical Research* **9**, 1380–1385.

[5] Chow, S.C. & Liu, J.P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker, New York.

[6] Endrenyi, L. & Al-Shaikh, P. (1995). Sensitive and specific determination of the equivalence of absorption rates, *Pharmaceutical Research* **12**, 1856–1864.

[7] Endrenyi, L., Fritsch, S. & Yan, W. (1991). C_{max}/AUC is a clearer measure than C_{max} for absorption rates in investigations of bioequivalence, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **29**, 394–399.

[8] Hauschke, D., Steinijans, V.W. & Diletti, E. (1990). A distribution-free procedure for the statistical analysis of bioequivalence studies, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **28**, 72–78.

[9] Holder, D.J. & Hsuan, F. (1993). Moment-based criteria for determining bioequivalence, *Biometrika* **80**, 835–846.

[10] Sauter, R., Steinijans, V.W., Diletti, E., Böhm, A. & Schulz, H.-U. (1992). Presentation of results from bioequivalence studies, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **30**, 233–256.

[11] Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar, *Biometrics* **51**, 615–626.

[12] Schall, R. (1995). Unified view of individual, population and average bioequivalence, in *BioInternational 2: Bioavailability, Bioequivalence and Pharmacokinetic Studies*, H.H. Blume & K.K. Midha, eds. Medpharm, Stuttgart, pp. 91–106.

[13] Schall, R. & Luus, H.E. (1993). On population and individual bioequivalence, *Statistics in Medicine* **12**, 1109–1124.

[14] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.

[15] Sheiner, L.B. (1992). Bioequivalence revisited, *Statistics in Medicine* **11**, 1777–1788.

[16] Steinijans, V.W. & Hauschke, D. (1990). Update on the statistical analysis of bioequivalence studies, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **28**, 105–110.

[17] Steinijans, V.W., Sauter, R., Jonkman, J.H.G., Schulz, H.U., Stricker, H. & Blume, H. (1989). Bioequivalence studies: single vs multiple doses, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **27**, 261–266.

[18] Westlake, W.J. (1981). Bioequivalence testing – a need to rethink, *Biometrics* **37**, 589–594.

[19] Westlake, W.J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations, in *Biopharmaceutical Statistics for Drug Development*, K.E. Peace, ed. Marcel Dekker, New York, pp. 329–352.

[20] Yee, K.F. (1986). The calculation of probabilities in rejecting bioequivalence, *Biometrics* **42**, 961–965.

[21] Zha, J. & Endrenyi, L. (1997). Variation of the peak concentration following single and repeated drug administrations in investigations of bioavailability and bioequivalence, *Journal of Biopharmaceutical Statistics* **7**, 191–204.

(*See also* **Dose-response in Pharmacoepidemiology**; **Drug Approval and Regulation**; **Drug Interactions**)

LASZLO ENDRENYI

Bioinformatics in Functional Genomics

With the sequencing of the human genome (or more accurately, a handful of human genomes), we are now said to be in a post-genomic era (*see* **Human Genome Project**). However, this term is confusing, since it is only now, with the availability of at least a draft outline of the genome of multiple organisms, that we can even begin systematically to deconstruct the relationship of the genetically programmed, physiologic behavior of an organism to the constituent **genes** comprising its individual version of the genome of its species. In this deconstruction, several kinds of biological information are available: **DNA sequences**, physical maps, genetic maps, gene **polymorphisms**, protein structure, gene expression (*see* **Gene Expression Analysis**), and protein interaction effects. The collection of these diverse data in large, internationally curated databases has produced an urgent need for systematic quantitative analysis, which in this domain often goes by the name of *bioinformatics*. The information in Figure 1 provides perhaps the best motivation for applying information sciences to the functional genomics enterprise. Since the invention of deoxyribonucleic acid (DNA) sequencing 25 years ago, the number of gene sequences deposited in international repositories, such as GenBank, has grown exponentially, culminating in the sequencing of the entire human genome in 2001. The *knowledge* about these genes (as measured by the number of papers published in biomedicine, a proxy measurement) has also been growing exponentially but at a much slower rate. As shown, the number of GenBank entries has fast outstripped the growth of MEDLINE entries. This difference serves as a proxy for the large gap between our knowledge of the functioning of the genome and the generation of raw genomic data. Yet GenBank entries represent only a fraction of the various kinds of data (listed above) generated from our investigations of the human genome. This exponentially expanding volume of data must somehow be sifted and linked to the biological phenomena of interest. Accomplishing this exhaustively, reliably, and reproducibly – credibly – is possible only with the application of algorithmic implementations on computers. This has led to an unprecedented demand for investigators who have

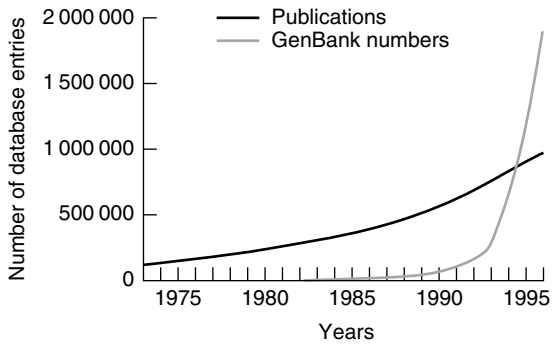

Figure 1 Cumulative growth of molecular biology and genetics literature (black) compared with DNA sequences (grey). Articles in the "G5" (molecular biology and genetics) subset of MEDLINE are plotted alongside DNA sequence records in GenBank over the same time period. The former data wave obtained with the help of R.M. Woodsmall of NCBI and the latter data are available (`ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt`). No attempt has been made to eliminate data redundancy among either the DNA sequence records or information contained in the literature

the required knowledge to manipulate large data sets. These skills may come from investigators in fields as diverse as computational physics, chemical engineering, operations research, and financial modeling. However, once these skills are applied to the domain of functional genomics, they can be collectively described as bioinformatic techniques. Although there is a wide overlap between the methodologies of bioinformatics and those of biostatistics, both the nature of the data and the computer-science orientation of many early practitioners of bioinformatics color much of the current research in and applications of bioinformatics. The breadth of applications of information science to biomedical research and practice far exceeds the scope of a brief article. Consequently, the focus here will be on the bioinformatic efforts that appear to be the most challenging and in the greatest demand: the elucidation of the function of genes – otherwise known as *functional genomics*.

Much of functional genomics has been and will continue to be the hypothesis-driven biological research that has been pursued for the past decades. Addressed here is a computationally intensive branch of functional genomics that has emerged as a result of the practical implementation of technologies for

assessing thousands of genes at a time.[1] The incredible confluence over the past five years of disparate technologies, such as robotics, fluorescence detection, photolithography, and the human genome project, has made it possible for present-day biologists to use ribonucleic acid (RNA) expression microarray detection technologies to greatly increase data about cells in various states. With the commercial tools currently available, a single experiment using RNA expression–detection microarrays can now provide systematic quantitative information on the expression of up to 60 000 unique RNAs within cells in any given physiological state (i.e. all expressed genes can be measured in any cell type under all conditions under which that cell will function).

cDNA and oligonucleotide microarray technology can be used not only to determine various cell functions but also to analyze more complex systems, such as traits with multigenic origins or those linked to the environment [16]. Microarrays can be used in time series to measure how a particular intervention [32, 51] may start a transcriptional program – that is, change the expression of large numbers of genes in a reproducible pattern determined by inherent genetic regulatory networks – and to measure gene expression in the appropriate tissue in groups of patients with and without a particular disease [3] or with two different diseases [28].

The ability to measure such RNA expression affords an opportunity to reduce our dependence on a priori knowledge (or biases) and to allow the biology of organisms to point us in potentially fruitful directions in our investigations. That is, much of the current mission of bioinformatics and functional genomics is a *hypothesis-generating* effort, which, if carefully crafted, can lead to a highly productive set of investigations using more conventional hypothesis-driven research.

Gene expression–detection microarrays are notable not because their ability to measure gene expression is unique, since many technologies have permitted quantitative or semiquantitative measurement of gene expression for well over two decades. What distinguishes gene expression–detection microarrays (and other genome-scale technologies) from these older technologies is the ability to measure tens of thousands of genes at a time, a *quantitative* change of the scale of gene measurement that has led to a *qualitative* change in our ability to understand regulatory processes at the cellular level.

Several approaches have been developed during the past four years to analyze basic RNA expression data sets. The central hypothesis of these techniques is that improved techniques in bioinformatics will enable us to analyze larger data sets of measurements from RNA expression–detection microarrays and thereby to discover the true biological functional pathways in gene regulation.

The related discipline of *clinical informatics* refers to the application of information science to various aspects of clinical care. Although clinical informatics is not addressed here, many of the problems that have dogged clinical informaticians (and, for that matter, biostatisticians) will confront bioinformaticians as they attempt to apply their basic science findings to clinical problems [36].

Why Do We Need New Techniques?

A scientist trained in quantitative techniques or even a biologically trained scientist taking a first look at a typical genomic study might ask the following, quite legitimate, question: Why isn't this field amenable to standard biostatistical techniques? After all, we are trying to understand the relationship between multiple variables and the mechanisms the relationships reveal, and the development of biostatistical techniques to analyze large studies that have large numbers of cases with many variables has a long history.

The following are the types of questions asked by conventional epidemiologic studies: What risk factors are associated with heart disease? Does smoking cause disease? On the surface, these questions seem similar to many of those posed about genetic risk factors for acute and chronic disease. Yet a review of the bioinformatics/functional genomics literature from the past three years reveals that most analyses in this field have used techniques borrowed from the computational sciences and machine-learning communities in particular. There are good reasons for this bias towards the computational sciences that have little to do with disciplinary parochialism.

Figure 2 sketches out a fundamental difference between a typical epidemiologic/clinical study and

[1] Expression microarrays, because they have been the most impressive recent examples of massive parallel acquisition of genomic data, are the canonical example used here. However, this discussion is equally applicable to other genomic technologies.

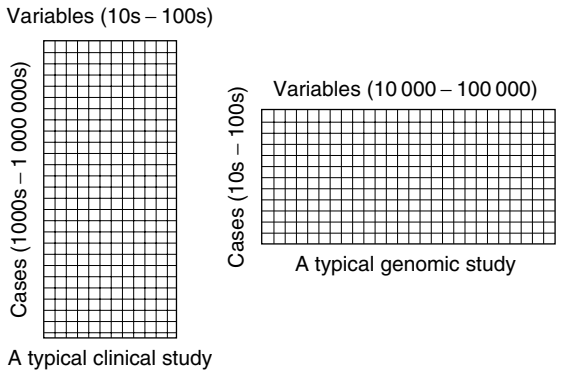

Figure 2 A major difference between classical clinical and epidemiologic studies and microarray analyses

a typical genomic study. A comprehensive epidemiologic study will often involve thousands to tens of thousands of subjects, such as in the Nurses' Health Study [8] or the Framingham Heart Study [19], and the measurement of tens or even hundreds of variables (often longitudinally). In contrast, a typical genomic study involves only tens or, exceptionally, hundreds of cases, each with tens of thousands of measured variables.

Initially the low number of cases in a genomic study may have been due to the high cost of the microarrays (in 1999 on the order of several thousand US dollars per microarray and in 2001 in the low hundreds) but the scarcity of cases in a typical functional genomic study will increasingly relate to the scarcity of appropriate biologic samples. Because these experiments involve measuring gene expression, a particular tissue (e.g. brain, muscle, fat) must be obtained under the right conditions, in contrast to studies using genomic DNA, for which more easily obtained blood samples will suffice. Samples of nonblood tissues may be very difficult to obtain in human populations. Yet although only tens of cases are involved, each case requires the measurement of tens of thousands of variables corresponding to the expression of tens of thousands of genes measurable with microarray technology. The result of the large number of variables as compared with the number of cases is a highly underdetermined system, i.e. these measurements are of very high dimensionality (on the order of tens of thousands) but with the provision of only a small number of cases to explore this high-dimensional space. Stated differently, there are a great many ways the variables being measured could be interrelated mechanistically, which may be difficult to model with the relatively small number of observations. Many of the assumptions underlying standard biostatistical techniques do not hold up well in these systems because of their high dimensionality and underdetermined nature. While statisticians have done quite a lot of research on the analysis of underdetermined systems of high dimensionality, only relatively recently has this work found its way into mainstream functional genomic studies.

The Functional Genomics Dogma

In the first two years of the publication of significant articles regarding the large-scale application of microarray technologies, numerous special-purpose or adapted machine-learning algorithms were described in the literature. Self-organizing maps [53], dendrograms [3, 21, 32, 51], K-means clusters, support vector machines [12], neural networks [13, 40, 55, 58], and several other methodologies (borrowed largely from the machine-learning community of computer science) have been employed. Most of these have worked reasonably well for the purposes described in the papers.

There is a central underlying assumption, or dogma, of all these techniques for expression analysis. Simply put, it is assumed that genes that appear to be expressed in similar patterns are in fact related mechanistically. Furthermore, the corollary to this assumption is that although genes may distantly affect the function of other gene products, they fall into groups of more tightly regulated mechanisms. For instance, the genes that govern chromosome function or meiosis may be more tightly linked to each other than to the genes involved with another function, such as apoptosis. This has been the basis of our collective experience in biologic investigations over the last century: that some groups of proteins have closer interactions than others. Often such groups have been organized into pathways such as glycolysis, the Kreb's cycle, and other metabolic pathways in which the gene products, called enzymes, have to work in concert. Other, more obvious functional clusters are those of structural proteins that have to come together in a conserved and reproducible fashion to serve their purpose, whether they are the components of the ribosomal unit or the histoproteins essential for the maintenance of chromatin structure.

On this basis, it is possible to impute functional clustering of genes whose expression patterns approximate one another, i.e. that they have related functions. Several important caveats are worth noting here. First, it remains unclear just how discrete the functional groupings of gene function is in the cellular apparatus. Individual gene products may have so many different roles under different circumstances that several of them partake in essential roles in significantly different functions.[2] The second caveat is that the term *functionally related* is not in itself well specified. Similar patterns of expression of more than one gene could signify the following possible relationships (the list is not exhaustive):

1. two genes having gene products that physically interact;
2. one gene encoding a transcriptional factor for the other gene;
3. two genes having different functions but similar promoter sequences; or
4. two genes both with promoter sequences bound by repressors that are knocked off when a nuclear receptor is activated even though the two genes have widely disparate functions.

Of course there is a level of abstraction at which *all genes* are functionally related by their roles in keeping the cell alive and producing whatever components are needed for the rest of the organism. But below this level of abstraction are many alternative and, by their nature, "sloppy" definitions of clustering. We should therefore be somewhat wary of the claim that similarity in expression corresponds to similarity in function. Nevertheless, this is a useful starting point for many analyses of a genome whose function remains, by and large, unknown at this time.

The question of what constitutes a similar expression pattern is also poorly defined, or at least has multiple alternative definitions. For example, similarity could mean that patterns of change over time are similar, that absolute levels of expression are similar at any given point in time, or that patterns of expression are perfectly opposite but well choreographed. Just which dissimilarity or similarity measure is chosen for examining patterns of expression will influence the kind of functional clusters that we expect.

[2] One example of this is the transcriptional factor Sonic Hedgehog, which in some tissues at some times is involved in cell proliferation and in others is involved in cell differentiation processes.

Supervised vs. Unsupervised Learning

The preceding sections have motivated the need for computational techniques for the analysis of gene expression that are qualitatively different from those of traditional epidemiologic biostatistics. Because the data sets are of high dimensionality but the number of cases is relatively small, the number of solutions that could explain the observed behavior is quite large. For this reason, the machine-learning community has recognized the potential role for techniques now used to explore high-dimensional spaces (such as those of voice or face recognition) for the exploration of genomic data sets.

Two useful broad categorizations of the techniques used by the machine-learning community are *supervised* learning techniques and *unsupervised* learning techniques, also commonly known as *classification* techniques and *clustering* techniques, respectively. The two techniques are easily distinguished by the presence of external labels of cases. For example, it is necessary to label a tissue as obtained from a patient with acute myelogenous leukemia (AML) or one with acute lymphocytic leukemia (ALL) [28] before applying a supervised learning technique to create a method of learning those labels. In an unsupervised learning technique, such as finding those genes that are co-regulated across all the samples, this type of organization of the data operates independently of any external labels. The kinds of variables (also known as *features* in the language of the machine-learning community) that characterize each *case* in a data set can be quite varied. Each case can include measures of clinical outcome, gene expression, gene sequence, drug exposure, proteomic measurements, or any other discrete or continuous variable believed to be of relevance to the case.

The two types of machine learning are generally used to answer different types of questions. In supervised learning, the goal is typically to obtain a set of variables (e.g. expressed genes as measured on a microarray) that can be used reliably to make a diagnosis, predict future outcome, predict future response to pharmacologic intervention, or categorize that patient or tissue or animal as part of a class of interest. In unsupervised learning, the typical application is to find either a completely novel cluster of genes with putative common (but previously unknown) function or, more commonly, to obtain a

cluster or group of genes that appear to have patterns of expression similar to those of a gene (i.e. that fall into the same cluster) already known to have an important function. The goal of unsupervised learning in this context is to find more details about the mechanism by which the known gene works and to find other genes involved in that same mechanism, either to obtain a more complete view of a particular cellular physiology or, in the case of pharmacologically oriented research, to identify other possible therapeutic targets.

Although the distinct goals of supervised versus unsupervised machine learning techniques may appear rather obvious, it is important to be aware of the implications of these differences for study design. For example, an analyst may be asked to find classifiers between two types of malignancy, as was done in the investigation by Golub et al. of AML and ALL [28]. However, the lists of genes that reliably divide the two malignancies may have little to do with the actual pathophysiologic causes of the two diseases and may not represent any particular close relationship of those genes and function. Why is this? One possibility is that the small amounts of change of some gene products, such as transcriptional activators and genes, e.g. *p53*, cause large downstream changes in gene expression. That is, with only a subtle change, an important upstream gene may cause dramatic changes in the expression in several pathways that are functionally only distantly related but are highly influenced by the same upstream gene. When a classification algorithm is applied directly to the gene expression levels, the algorithm will naturally identify those genes that undergo the most change between the two or more states being classified. A study design geared towards the application of a supervised learning technique may thereby generate a useful artifact for classification, diagnosis, or even prognosis, but will not necessarily lead to valuable insights into the biology underlying the classes obtained. To obtain such insights, unsupervised or clustering methodologies are more likely to be rewarding, as they reveal how genes will affect each other's function. More generally, let us consider the other cases for which gene expression values are not the only data type. A given case may include several thousand gene-expression measurements but also several hundred phenotypic measurements such as blood pressure, laboratory values, or the response to a chemotherapeutic agent. Here again a clustering algorithm can be used to find the features that are most tightly coupled in the observed data. In a generalization of the functional genomics dogma, these tight associations (in space and/or time) can thereby lead to the development of hypotheses that may be tested by standard techniques.

The list of bioinformatic techniques is growing rapidly. Table 1 includes only a brief subset, with references to published biomedic works incorporating each technique and starting with the taxonomy of unsupervised and supervised machine learning. For the experienced biostatistician, many of these techniques will be familiar tools called by a new label or applying a new jargon.

The Immediate Future of Bioinformatics and Functional Genomics

In the last five years of the genomic revolution and the development of concomitant bioinformatic methodologies, the principal weakness of the field has been the poor quality and irreproducibility of many of the measurements made [17, 54]. This is particularly true of microarray-expression experiments[3] that make broad and unsubstantiated claims about gene function on the basis of one, two, or three measurements. As the science and engineering of functional genomics develop (and become more cost-effective), many of the tried and true techniques of biostatistics are being applied to genomic data (often by biostatisticians) [39, 47]. These efforts are still in their infancy but represent the beginnings of the integration of genomic data into the armamentarium of the biostatistician as yet another type of data for clinical trials and basic biologic investigation.

This brief article has only touched on one of the main themes of bioinformatics. Because the bioinformatic and genomic endeavor is so large, much mundane groundwork remains to be covered, and this too is an important component of the enterprise. This includes the development of standardized nomenclatures for describing genes, gene variants, and gene products [6, 37, 41, 57]; the development of standardized data models and databases for storing and sharing genomic data acquired in thousands of laboratories [22, 26, 50];

[3] This is also true of all the massively parallel genomic measurement modalities, whether using proteomic arrays, tissue arrays, or resequencing arrays.

Table 1 Bioinformatic techniques

Unsupervised
Analysis looking for characteristics in the data set, without a priori input on cases or genes
- Feature determination: Determine genes with interesting properties, without specifically looking for a specific pattern to be determined
- Principal component analysis and singular value decomposition: Determine genes explaining the majority of the mathematical **variance** in the data set [5, 23, 31, 46, 56]
- Cluster determination: Determine groups of genes or samples with similar patterns of gene expression
  - Nearest neighbor clustering: The number of clusters is decided first, the boundaries of the clusters are calculated, then each gene is assigned to a single cluster [9]
  - Agglomerative clustering: Bottom-up method, where clusters start as empty, then genes are successively added to the existing clusters
    Dendrogram algorithm: Groups are defined as subtrees in a phylogenetic-type tree, created by comprehensively measuring a pairwise metric, such as the correlation coefficient [56]
    Two-dimensional dendrograms: Both genes and samples are clustered separately
  - Divisive or partitional clustering: Top-down method, where large clusters are successively broken into smaller ones until each subcluster contains only one gene
    Matrix incision tree [35]
    Two-way clustering binary tree [4]
    Coupled two-way clustering [27]
    Cluster affinity search technique [11]
    Gene shaving [30]
- Network determination: Determine networks of gene–gene or gene–phenotype interactions
  Bayesian networks [24]
  Hybrid petri networks [44]
  Boolean regulatory networks [1, 2, 40, 52, 59]
  Relevance networks: Determines associations between features (genes, phenotypic measures, or samples) [14, 15, 17]

Supervised
Analysis to determine ways to accurately split into or predict groups of samples or disease based on external (typically expert-provided) labels
- Single feature or sample determination: Find genes or samples that match a particular a priori pattern
  - Naive Bayes classifier [10, 18]
  - Naive Bayes global relevance [45]
- Multiple-feature determination: Find combinations of genes that match a particular a priori pattern
  - Decision trees: Use the training set of genes or samples to construct a decision tree to help classify test samples or test genes. Typically uses entropy as the classification measure [20]
  - Support vector machines: First take the set of measured genes, then create a richer feature set with combinations of genes, then find a hyperplane that linearly separates groups of samples in this larger multidimensional space [12, 18, 25]
  - Tree harvesting [29]
  - Boosting [9]

distributed annotation efforts – allowing laboratories around the world to contribute in a controlled manner to databases documenting the function of each gene [7, 34, 38, 48, 49]; and leverage of the existing biomedical literature for genomic analysis [33, 42, 43].

References

[1] Akutsu, T., Miyano, S. & Kuhara, S. (2000). Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function, *Journal of Computational Biology* **7**, 331–343.

[2] Akutsu, T., Miyano, S. & Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics* **16**, 727–734.

[3] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**, 503–511.

[4] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. & Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed

by oligonucleotide arrays, *Proceedings of the National Academy of Sciences* **96**, 6745–6750.

[5] Alter, O., Brown, P.O. & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy of Sciences* **97**, 10101–10106.

[6] Antonarakis, S.E. (1998). Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group, *Human Mutation* **11**, 1–3.

[7] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genetics* **25**, 25–29.

[8] Belanger, C., Hennekens, C., Rosner, B. & Speizer, F. (1978). The Nurses' Health Study, *American Journal of Nursing* **78**, 1039–1040.

[9] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z. (2000). Tissue classification with gene expression profiles, *Journal of Computational Biology* **7**, 559–583.

[10] Ben-Dor, A., Friedman, N. & Yakhini, Z. (1999). In *International Conference on Computational Biology (RECOMB)*, ACM, Tokyo, pp. 31–38.

[11] Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999). Clustering gene expression patterns, *Journal of Computational Biology* **6**, 281–297.

[12] Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proceedings of the National Academy of Sciences* **97**, 262–267.

[13] Brunak, S., Engelbrecht, J. & Knudsen, S. (1990). Neural network detects errors in the assignment of mRNA splice sites, *Nucleic Acids Research* **18**, 4797–4801.

[14] Butte, A. & Kohane, I.S. (1999). In *Fall Symposium, American Medical Informatics Association*, N. Lorenzi, ed., Hanley & Belfus, Washington, pp. 711–715.

[15] Butte, A. & Kohane, I. (2000). In *Pacific Symposium on Biocomputing 2000*, R. Altman, K. Dunker, L. Hunter, K. Lauderdale & T. Klein, eds, World Scientific, Hawaii, pp. 418–429.

[16] Butte, A., Tamayo, P., Slonim, D., Golub, T.R. & Kohane, I.S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, *Proceedings of the National Academy of Sciences* **97**, 12182–12186.

[17] Butte, A.J., Ye, J., Niederfellner, G., Rett, K., Häring H.U., White, M.F. & Kohane, I.S. (20001). In *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 6–17.

[18] Chow, M.L., Moler, E.J. & Mian, I.S. (2001). Identifying marker genes in transcription profiling data using a mixture of feature relevance experts, *Physiological Genomics* **5**, 99–111.

[19] Dawber, T., Meadors, G. & Moore, F. (1951). The Framingham Study: epidemiological approaches to heart disease, *American Journal of Public Health* **41**, 279–286.

[20] Dietterich, T.G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Computation* **10**, 1895–1923.

[21] Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences* **95**, 14863–14868.

[22] Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M. & Boguski, M.S. (1998). Data management and analysis for gene expression arrays, *Nature Genetics* **20**, 19–23.

[23] Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics, *Nature Biotechnology* **18**, 1157–1161.

[24] Friedman, N., Linial, M., Nachman, I. & Péer, D. (2000). Using Bayesian networks to analyze expression data, *Journal of Computational Biology* **7**, 601–620.

[25] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics* **16**, 906–914.

[26] Gardiner-Garden, M. & Littlejohn, T.G. (2001). A comparison of microarray databases, *Briefings in Bioinformatics* **2**, 143–158.

[27] Getz, G., Levine, E. & Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data, *Proceedings of the National Academy of Sciences* **97**, 12079–12084.

[28] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. & Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**, 531–537.

[29] Hastie, T., Tibshirani, R., Botstein, D. & Brown, P. (2001). Supervised harvesting of expression trees, *Genome Biology* **2**, 3.1–3.12.

[30] Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. & Brown, P. (2000). "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology* **1**, 3.1–3.21.

[31] Hilsenbeck, S.G., Friedrichs, W.E., Schiff, R., O'Connell, P., Hansen, R.K., Osborne, C.K. & Fuqua, S.A. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance, *Journal of the National Cancer Institute* **91**, 453–459.

[32] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Jr, Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. & Brown, P.O. (1999). The transcriptional program in the response of human fibroblasts to serum, *Science* **283**, 83–87.

[33] Jenssen, T.K., Laegreid, A., Komorowski, J. & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression, *Nature Genetics* **28**, 21–28.

[34] Karp, P.D. (2000). An ontology for biological function based on molecular interactions, *Bioinformatics* **16**, 269–285.

[35] Kim, J.H., Ohno-Machado, L. & Kohane, I.S. (2001). In *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 30–41.

[36] Kohane, I.S. (2000). Bioinformatics and clinical informatics: the imperative to collaborate (editorial), *Journal of the American Medical Informatics Association* **7**, 512–516.

[37] Kuska, B. (1997). Scientists reach a turning point with gene nomenclature, *Journal of the National Cancer Institute* **89**, 1332–1334.

[38] Lewis, S., Ashburner, M. & Reese, M.G. (2000). Annotating eukaryote genomes, *Current Opinion in Structural Biology* **10**, 349–354.

[39] Li, C. & Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proceedings of the National Academy of Sciences* **98**, 31–36.

[40] Liang, S., Fuhrman, S. & Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures, in *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 18–29.

[41] Maltais, L.J., Blake, J.A., Eppig, J.T. & Davisson, M.T. (1997). Rules and guidelines for mouse gene nomenclature: a condensed version. International Committee on Standardized Genetic Nomenclature for Mice, *Genomics* **45**, 471–476.

[42] Masys, D.R. (2001). Linking microarray data to the literature, *Nature Genetics* **28**, 9–10.

[43] Masys, D.R., Welsh, J.B., Lynn Fink, J., Gribskov, M., Klacansky, I. & Corbeil, J. (2001). Use of keyword hierarchies to interpret gene expression patterns, *Bioinformatics* **17**, 319–326.

[44] Matsuno, H., Doi, A., Nagasaki, M. & Miyano, S. (2000). Hybrid Petri net representation of gene regulatory network, in *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 341–352.

[45] Moler, E.J., Radisky, D.C. & Mian, I.S. (2000). Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in *S. cerevisiae*, *Physiological Genomics* **4**, 127–135.

[46] Raychaudhuri, S., Stuart, J.M. & Altman, R.B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series, in *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 455–466.

[47] Schadt, E.E., Li, C., Su, C. & Wong, W.H. (2000). Analyzing high-density oligonucleotide gene expression array data, *Journal of Cellular Biochemistry* **80**, 192–202.

[48] Schulze-Kremer, S. (1997). Adding semantics to genome databases: towards an ontology for molecular biology, *Proceedings of the International Conference on Intelligent Systems in Molecular Biology* Vol. 5, pp. 272–275.

[49] Schulze-Kremer, S. (1998). Ontologies for molecular biology, in *Pacific Symposium Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 695–706.

[50] Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S. et al. (2001). The Stanford Microarray Database, *Nucleic Acids Research* **29**, 152–155.

[51] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell* **9**, 3273–3297.

[52] Szallasi, Z. & Liang, S. (1998). Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies, in *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 66–76.

[53] Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. (1999). Analysis of gene expression data using self-organizing maps, *FEBS Letters* **451**, 142–146.

[54] Tsien, C.L., Libermann, T.A., Gu, X. & Kohane, I.S. (2001). In *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed., World Scientific, Hawaii, pp. 496–507.

[55] Weinstein, J.N., Kohn, K.W., Grever, M.R., Viswanadhan, V.N., Rubinstein, L.V., Monks, A.P. et al. (1992). Neural computing in cancer drug development: predicting mechanism of action, *Science* **258**, 447–451.

[56] Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. & Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development, *Proceedings of the National Academy of Sciences* **95**, 334–339.

[57] White, J.A., McAlpine, P.J., Antonarakis, S., Cann, H., Eppig, J.T., Frazer, K., Frazal, J., Lancet, D., Nahmias, J., Pearson, P., Peters, J., Scott, A., Scott, H., Spurr, N., Talbot, C., Jr & Povey, S. (1997). Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee, *Genomics* **45**, 468–471.

[58] Wu, C., Whitson, G., McLarty, J., Ermongkonchai, A. & Chang, T. (1992). Protein classification artificial neural system, *Protein Science* **1**, 667–677.

[59] Wuensche, A. (1998). Genomic regulation modeled as a network with basins of attraction, in *Pacific Symposium on Biocomputing*, Vol. 6, R. Altman, ed. World Scientific, Hawaii, pp. 89–102.

(*See also* **Gene Expression Analysis**)

ISAAC S. KOHANE & ATUL BUTTE

Bioinformatics

Bioinformatics is an emerging field that was once considered to be the part of computational biology that explicitly dealt with the management of the increasing number of large **databases**, including methods for data retrieval and analyses, and **algorithms** for sequence similarity searches, structural predictions, functional predictions and comparisons, and so forth. Very recently, the field has been rapidly evolving, not only because of the impact of the various genome projects (*see* **Human Genome Project**) but also because of the development of experimental technologies such as microarrays for gene expression analyses and mass spectrometry for detection of protein–protein interactions. Currently, and increasingly, bioinformatics is being widely viewed as a more fundamental discipline that also encompasses mathematics, statistics, physics, and chemistry. Further, the field is already looking forward to what is currently termed a "systems biology" approach and to simulations of whole cells with incorporation of more levels of complexity; see, for example, [2].

The stated goal for many researchers is for developments in bioinformatics to be focused at finding the fundamental laws that govern biological systems, as in physics. However, if such laws exist, they are a long way from being determined for biological systems. Instead, the current aim is to find insightful ways to model limited components of biological systems and to create tools that biologists can use to analyze data. Examples include tools for statistical assessment of the similarity between two or more DNA sequences or protein sequences (*see* **Sequence Analysis**), for finding genes in genomic DNA (*see* **DNA Sequences**), for quantitative analysis of functional genomics data (*see* **Genetic Markers**), for estimating differences in how genes are expressed in, say, different tissues (*see* **Gene Expression Analysis**), for analysis and comparison of genomes from different species, for phylogenetic analysis (*see* **DNA Sequences**), for DNA sequence alignment and assembly (*see* **Hidden Markov Models**; **EM Algorithm**), and so on. Such tools involve statistical modeling of biological systems; see, for example, [1].

Much biological data arise from mechanisms that have a substantial probabilistic component, the most significant being the many random processes inherent in biological evolution, and also from randomness in the sampling process used to collect the data. Another source of variability or randomness is introduced by the biotechnological procedures and experiments used to generate the data. So, the basic goal is to distinguish the biological "signal" from the "noise". Today, as experimental techniques are being developed for studying genomewide patterns, such as expression arrays, the need to appropriately deal with the inherent variability has multiplied astronomically. For example, we have progressed from studying one or a few genes in comparative isolation to being able to evaluate thousands of genes (or expressed sequence tags) simultaneously. Not only must methodologies be developed that scale up to handle the enormous data sets generated in the postgenomic era but these also need to become more sensitive to the underlying biological knowledge and better understanding of the mechanisms that generate the data. For biostatisticians, research has reached an exciting and challenging stage at the interface of computational statistics and biology. The need for novel approaches to handle the new genomewide data (including that generated by microarrays) has coincided with a period of dramatic change in approaches to statistical methods and thinking. This "quantum" change has been brought about, or even has been driven by, the potential of ever more increasing computing power. What was thought to be intractable in the past is now feasible, and so new methodologies need to be developed and applied.

Unfortunately, too many of the current practices in the biological sciences rely on methods developed when computational resources were very limiting and are often either (a) simple extensions of methods for working with one or a few outcome measures, and do not work well when there can be thousands of outcome measures, or (b) ad hoc methods (that are commonly referred to being "statistical" or "computational") that make many assumptions for which there are no (biological) justifications. The challenge now is to creatively combine the power of the computer with relevant biological and **stochastic process** knowledge to derive novel approaches and models, using minimal assumptions, that can be applied at genomic wide scales. Such techniques comprise the foundation of bioinformatic methods in the future.

Useful web resources are starting to appear. For example, "Functional and Comparative Genomics" has been developed by the US Department of Energy Office of Science, Office of Biological and

Environmental Research, Human Genome Program; see, for example, `http://www.ornl.gov/TechResources/Human_Genome/faq/compgen.html` and links therein.

Federated databases and bioGrids are at the cutting edge of modern biological technology. To date, Grid development has focused on the basic issues of storage, computation, and resource management needed to make a global life-science community's information and tools accessible in a high-performance environment. In the longer term, the purpose of the Grids is to deliver a collaborative and supportive environment that will enable geographically distributed scientists to achieve research goals more effectively, while enabling their results to be used in developments elsewhere. The *in silico* biological experimental process will use efficient tools that allow the e-life scientists to seamlessly link together databases and analytical tools, extract relevant information from free texts, and harness available computational resources for CPU-intensive tasks. Also, there will be an increasing overlap and seamlessness between areas of bioinformatics and Public Health Informatics. For the biostatistician, in the future, both data analysis and simulation will be done increasingly at high speed achieved by parallel processing and heterogeneous distributed processing.

References

[1] Ewens, W.J. & Grant, G.R. (2001). *Statistical Methods in Bioinformatics*. Springer-Verlag, New York.

[2] Kanehisa, M. & Bork, P. (2003). Bioinformatics in the post-sequence era, *Nature Genetics Supplement* **33**, 305–310.

SUSAN R. WILSON

Biological Assay, Overview

This article mainly emphasizes the classical aim of biological assay (or *bioassay*), to estimate *relative potency*, arising out of a need for **biological standardization** of drugs and other products for biological usage. There is a basic difference between biological and chemical endpoints or responses: the former exhibits greater (bio)variability and thereby requires *in vivo* or *in vitro* biological assays wherein a standard preparation (or a reference material) is often used to have a meaningful interpretation of relative potency. However, the term *bioassay*, has also been used in a wider sense, to denote an experiment, with biological units, to detect possible adverse effects such as carcinogenicity or mutagenicity (*see* **Software for Clinical Trials**; **Mutagenicity Study**). In the context of environmental impact on biosystems, *toxicodynamic* and *toxicokinetic* (TDTK) models as well as *physiologically based pharmacokinetic* (PBPK) models have been incorporated to expand the domain of bioassays; *structure–activity relationship information* (SARI) is often used to consolidate the adoption of bioassays in a more general setup; the genesis of dosimetry (or *animal studies*) lies in this complex. The use of *biomarkers* in studying environmental toxic effects on biological systems, as well as in carcinogenicity studies, has enhanced the scope of bioassays to a greater interdisciplinary field; we need to appraise, as well, bioassays in this broader sense. Further, recent advances in **bioinformatics** have added new frontiers to the study of biological systems; bioassay models are gaining more popularity in the developing area of *computational biology*. Our appraisal of bioassay would remain somewhat incomplete without an assessment of the role of *Pharmacogenomics* as well as *Toxicogenomics* in establishing a knowledge base of the chemical effects in biological systems. The developments in genomics during the past eight years, have opened the doors for a far more penetrating level of research focusing on the **gene-environment interaction** in conventional experiments with biological units, and thereby calling for drastically different statistical resolutions for bioassays. We include also a brief synopsis of these recent developments.

Traditionally, in a bioassay, a test (new) and a standard preparation are compared by means of reactions that follow their applications to some biological units (or subjects), such as subhuman primates (or human) living tissues or organs; the general objective being to draw interpretable statistical conclusions on the *relative potency* of the test preparation with respect to the standard one. Usually, when a drug or a *stimulus* is applied to a subject, it induces a change in some measurable characteristic that is designated as the *response* variable. In this setup, the dose may have several chemically or therapeutically different ingredients while the response may also be multivariable. Thus the *stimulus–response* or **dose–response** relationship for the two preparations, both subject to inherent stochastic variability, are to be compared in a sound statistical manner (with adherence to biological standardization) so as to cast light on their relative performance with respect to the set objectives. Naturally, such statistical procedures may depend on the nature of the stimulus and response, as well as on other extraneous experimental (biological or therapeutical) considerations. As may be the case with some competing drugs for the treatment of a common disease or disorder, the two (i.e. test and standard) preparations may not have the same chemical or pharmacological constitution, and hence, statistical modeling may be somewhat different than in common laboratory experimentation. Nevertheless, in many situations, the test preparation may behave (in terms of the response/tolerance distribution) as if it is a dilution or concentration of the standard one. For this reason, often, such bioassays are designated to compare the relative performance of two drugs under the *dilution–concentration* postulation, and are thereby termed *dilution assays*.

Dilution assays are classified into two broad categories: *Direct dilution* and *indirect dilution* assays. In a direct assay, for each preparation, the exact amount of dose needed to produce a specified response is recorded, so that the response is certain while the dose is a nonnegative random variable that defines the *tolerance distribution*. Statistical modeling of these tolerance distributions enables us to interpret the relative potency in a statistically analyzable manner, often in terms of the parameters associated with the tolerance distributions. By contrast, in an indirect assay, the dose is generally administered at some prefixed (usually nonstochastic) levels, and at each level, the response is observed for subjects included in the

study. Thus, the dose is generally nonstochastic and the stochastic response at each level leads to the tolerance distributions that may well depend on the level of the dose as well as the preparation. If the response is a quantitative variable, we have an indirect *quantitative* assay, while if the response is *quantal* in nature (i.e. all or nothing), we have a *quantal* assay (*see* **Binary Data**). Both of these indirect assays are more commonly addressed in statistical formulations.

Within this framework, the nature of the dose–response regression may call for suitable **transformations** on the dose variable (called the *dosage* or *dose-metameter*) and/or the response variable, called the *response-metameter*. The basic objective of such transformations is to achieve a linear dosage-response regression (*see* **Linear Regression, Simple**), which may induce simplifications in statistical modeling and analysis schemes. In view of the original dilution structure, such transformations may lead to different designs for such assays, and the two most popular ones are (i) **parallel-line assays** and (ii) **slope-ratio assays**. Within each class, there is also some variation depending on the (assumed) nature of tolerance distributions, and within this setup, the *probit* (or normit) and *logit* transformations, based on normal and logistic distributions respectively, are quite popular in statistical modeling and analysis of bioassays. Bliss [2] contains an excellent account of the early developments in this area, while the various editions of Finney [6] capture more up-to-date developments, albeit with a predominantly parametric flavor. We refer to these basic sources for extensive bibliography of research articles, particularly in the early phase of developments where biological considerations often dominated statistical perspectives.

In this framework, it is also possible to include bioassays that may be considered for **bioavailability and bioequivalence** studies, though basically there are some differences in the two setups: Bioassays for assessing relative potency relate to *clinical therapeutic* **equivalence trials**, while in **bioequivalence** trials, usually, the relative bioavailability of different formulations of a drug are compared. Thus, in bioequivalence studies, the *pharmacologic* results of administrating essentially a common drug in alternative forms, such as a capsule versus a tablet, or a liquid dose of certain amount, capsules (tablets) or liquid forms of larger dose versus smaller dose with increased frequency of prescription, or even the administration of a drug at different time of the day, such as before breakfast or sometime after a meal, and so on, are to be assessed in a valid statistical manner. In this sense, the active ingredients in the drug in such alternative forms may be essentially the same, and differences in bioavailability reflect the form and manner of administration. We shall see later on that these basic differences in the two setups call for somewhat different statistical formulations and analysis schemes.

Direct Dilution Assays

As an illustrative example, consider two toxic preparations (say, S and T), such that a preparation is continuously injected into the blood stream of an animal (say, cat) until its heart stops beating. Thus, the response (death) is certain, while the exact amount of the dose (X) required to produce the response is stochastic. Let X_S and X_T stand for the dose (variable) for the standard and test preparation, and let $F_S(x)$ and $F_T(x)$, $x \geq 0$, be the two tolerance distributions. The *fundamental assumption* of a direct dilution assay is the following:

$$F_T(x) = F_S(\rho x), \quad \text{for all } x \geq 0, \tag{1}$$

where $\rho(>0)$ is termed the relative potency of the test preparation with respect to the standard one. Standard parametric procedures for drawing statistical conclusions on ρ are discussed fully in the classical text of Finney [6], where other references are also cited in detail. If $F_S(.)$ is assumed to be a **normal distribution** function, then ρ is characterized as the ratio of the two means, as well as the ratio of the two standard deviations. Such simultaneous constraints on means and variances vitiate the simplicity of achieving optimality of parametric procedures (in the sense of **maximum likelihood** estimators and related **likelihood ratio tests**). On the other hand, if we use the log-dose transformation on the two sets of doses, and the resulting dosage distributions, denoted by $F_S^*(.)$ and $F_T^*(,)$ respectively, are taken as normal, then they have the same variance, while the difference of their means define $\log \rho$. Interestingly enough, in the first case, the estimator of ρ is the ratio of the sample arithmetic means, while in the other case, it turns out as the ratio of the sample geometric means. A different estimator emerges when one uses a *power-dosage* (as is common in slope-ratio assays). Thus,

in general, these estimators are not the same, and they depend sensibly on the choice of a dosage. This explains the lack of invariance property of such parametric estimates (as well as associated test statistics) under monotone dosage transformations.

From an operational point of view, an experimenter may not have the knowledge of the precise dosage, and hence, it may not be very prudent to assume the normality, lognormality or some other specific form of the tolerance distribution. Therefore, it may be reasonable to expect that an estimator of the relative potency should not depend on the chosen dosage as long as the latter is strictly monotone. For example, if the true tolerance distribution is *logistic* while we assume it to be (log)normal, the sample estimator may not be **unbiased** and fully **efficient**. Even when the two tolerance distributions are taken as normal, the ratio of the sample means is not unbiased for ρ. In this respect, such parametric procedures for the estimation of the relative potency (or allied tests for the fundamental assumption) are not so **robust**, and any particular choice of a dosage may not remain highly efficient over a class of such chosen tolerance distributions. **Nonparametric** procedures initiated by Sen [17, 18, 19] and followed further by Shorack [26], and Rao and Littell [14], among others, eliminate this arbitrariness of dosage selection and render robustness to a far greater extent. Basically, we may note that **ranks** are invariant under strictly monotone (not necessarily linear) transformations on the sample observations. As such, a test for the fundamental assumption in (1) based on appropriate rank statistic remains invariant under such transformations. Similarly, if an estimator of the relative potency is based on suitable rank statistics, it remains invariant under such strictly monotone dosage transformations. Both the **Wilcoxon–Mann–Whitney** two-sample rank-sum test and the (Brown–Mood) **median** test statistics were incorporated by Sen [17] for deriving nonparametric estimators of relative potency, and they also provide distribution-free **confidence intervals** for the same parameter. If there are m observations $X_{S1}, \ldots, X_{Sm}$ for the standard preparation and n observations $X_{T1}, \ldots, X_{Tn}$, for the test preparation, we define the differences

$$Y_{ij} = X_{Si} - X_{Tj}, \text{ for } i = 1, \ldots, m; \; j = 1, \ldots, n. \tag{2}$$

We arrange the $N(= mn)$ observations Y_{ij} in ascending order of magnitude, and let $\tilde{Y}_N$ be the median of these N observations. If N is even, we take the average of the two central **order statistics**. Then $\tilde{Y}_N$ is the Wilcoxon score estimator of $\log \rho$, and it is a robust and efficient estimator of $\log \rho$. The estimator is invariant under any strictly monotone transformation on the dose. Similarly, the confidence interval for $\log \rho$ can be obtained in terms of two specified order statistics of the Y_{ij}, and this is a distribution-free and robust procedure. A similar procedure works out for the median procedure; for general rank statistics, generally, an iterative procedure is needed to solve for such robust R-estimators (*see* **Robust Regression**). Rao and Littell [14] incorporated the two-sample **Kolmogorov–Smirnov test** statistics in the formulation of their estimator. For computational convenience, because of the invariance property, it is simpler to work with the log-dose dosage, and in that way, the estimators of the log-relative potency correspond to the classical rank estimators in the two-sample location model.

These direct dilution assays require the measurement of the exact doses needed to produce the response; this may not be the case if there are some *latent effects*. For example, the time taken by the toxic preparation to traverse from the point of infusion to the heart multiplied by the infusion rate may account for such a latent effect. In general, the situation may be much more complex. This naturally affects the fundamental assumption in (1), and variations in the modeling and statistical analysis to accommodate such effects have been discussed in [6] and [17] in the parametric and nonparametric cases respectively.

Indirect Dilution Assays

As an example, consider two drugs, A and B, each administered at $k(\geq 2)$ prefixed levels (doses) $d_1, \ldots, d_k$. Let X_{Si} and Y_{Ti} be the response variable for the standard and test preparation respectively. These drugs may not have the same chemical ingredients, and may not have the same dose levels. It is not necessary to have the same doses for both the preparations, but the modifications are rather straightforward, and hence we assume this congruence. We assume first that both X_{Si} and Y_{Ti} are continuous (and possibly nonnegative) **random variables**. Suppose further that there exist some dosage $x_i = \xi(d_i), i =$

$1, \ldots, k$ and response-metameter $X^* = g(X)$, $Y^* = g(Y)$, for some strictly monotone $g(.)$, such that the two dosage–response **regressions** may be taken as linear, namely, that

$$Y^*_{Ti} = \alpha_T + \beta_T x_i + e_{Ti}, \quad X^*_{Si} = \alpha_S + \beta_S x_i + e_{Si}, \tag{3}$$

for $i = 1, \ldots, k$, where for statistical inferential purposes, certain distributional assumptions are needed for the error components e_{Ti} and e_{Si}, $i = 1, \ldots, k$. Generally, in the context of log-dose transformations, we have a parallel-line assay, while slope-ratio assays arise typically for **power transformations**. Thus, in a parallel-line assay, the two dose–response regression lines are taken to be parallel, and further that the errors e_{Ti} and e_{Si} have the same distribution (often taken as normal). In this setup, we have then $\beta_S = \beta_T = \beta$ (unknown), while $\alpha_T = \alpha_S + \beta \log \rho$, where ρ is the relative potency of the test preparation with respect to the standard one. This leads to the basic estimating function

$$\log \rho = \frac{\{\alpha_T - \alpha_S\}}{\beta}, \tag{4}$$

so that if the natural parameters β, α_S and α_T are estimated from the acquired bioassay dataset, statistical inference on $\log \rho$ (and hence ρ) can be drawn in a standard fashion. For normally distributed errors, the whole set of observations pertains to a conventional linear model with a constraint on the two slopes β_S, β_T, so that the classical maximum likelihood estimators and allied likelihood ratio tests can be incorporated for drawing statistical conclusions on the relative potency or the fundamental assumption of parallelism of the two regression lines. However, the estimator of $\log \rho$ involves the ratio of two normally distributed statistics, and hence, it may not be unbiased; moreover, generally, the classical **Fieller's theorem** [6] is incorporated for constructing a confidence interval for $\log \rho$ (and hence, ρ), and it is known that this may result in an inexact coverage probability (*see* **Confidence Intervals and Sets**). Because of this difference in setups (with that of the classical linear model), design aspects for such parallel-line assays need a more careful appraisal. For equispaced (log −)doses, a symmetric $2k$-point design has optimal information contents, and are more popularly used in practice. We refer to [6] for a detailed study of such bioassay designs in a conventional normally distributed errors model. Two main sources of nonrobustness of such conventional inference procedures are the following:

(i) Possible nonlinearity of the two regression lines (they may be parallel but yet curvilinear);
(ii) Possible nonnormality of the error distributions.

On either count, the classical normal theory procedures may perform quite nonrobustly, and their (asymptotic) optimality properties may not hold even for minor departures from either postulation. However, if the two dose–response regressions (linear or not) are not parallel, the fundamental assumption of parallel-line assays is vitiated, and hence, statistical conclusions based on the assumed model may not be very precise.

In a slope-ratio assay, the intercepts α_S and α_T are taken as the same, while the slopes β_S and β_T need not be the same and their ratio provides the specification of the relative potency ρ. In such slope-ratio assays, generally, a power transformation: dosage $=$ (dose)$^\lambda$, for some $\lambda > 0$ is used, and we have

$$\rho = \left\{ \frac{\beta_T}{\beta_S} \right\}^{1/\lambda}, \tag{5}$$

which is typically a nonlinear function of the two slopes β_T and β_S, and presumes the knowledge of λ. In such a case, the two error components may not have the same distribution even if they are normal. This results in a heteroscedastic linear model (unless $\rho = 1$) (*see* **Scedasticity**), where the conventional linear estimators or allied tests may no longer possess validity and efficiency properties. Moreover, as ρ^λ is a ratio of two slopes, its conventional estimator based on usual estimators of the two slopes is of the ratio-type. For such ratio-type estimators, again the well-known Fieller Theorem [6] is usually adopted to attach a confidence set to ρ or to test a suitable null hypothesis. Such statistical procedures may not have the exact properties for small to moderate sample sizes. Even for large sample sizes, they are usually highly nonrobust for departures from the model-based assumptions (i.e. linearity of regression, the fundamental assumption, and normality of the errors). Again the design aspects for such slope-ratio assays need a careful study, and [6] contains a detailed account of this study. Because of the

common intercept, usually a $2k + 1$ point design, for some nonnegative integer k is advocated here.

The articles on **Parallel-line Assay** and **Slope-ratio Assay** should be consulted for further details. The primary emphasis in these articles is on standard parametric methods, and hence we discuss briefly here, the complementary developments of nonparametric and robust procedures for such assays. These were initiated in [21, 22] and also systematically reviewed in [23]. First, we consider a nonparametric test for the validity of the fundamental assumption in a parallel-line assay. This is essentially a test for the equality of slopes of two regression lines, and as in [21], we consider an aligned test based on the Kendall τ statistic (*see* **Rank Correlation**). For each preparation with the set of dosages as independent variate and responses as dependent variable, one can define the Kendall tau statistic in the usual manner. We consider the aligned observations $Y_{Ti} - bx_i$ and x_i, and denote the corresponding Kendall τ (in the summation but not average form) as $K_T(b)$, and for the standard preparation, an aligned Kendall's τ statistic is defined by $K_S(b)$, where we allow b to vary over the entire real line. Let then

$$K^*(b) = K_T(b) + K_S(b), \quad -\infty < b < \infty. \tag{6}$$

Note then that $K_T(b)$, $K_S(b)$, and hence $K^*(b)$ are all nonincreasing in b and have finitely many step-down discontinuities. Equating $K^*(b)$ to 0 [20], we obtain the pooled estimator $\hat{\beta}$ of β. Let us then write

$$\mathcal{L} = \frac{\{[K_T(\hat{\beta})]^2 + [K_S(\hat{\beta})]^2\}}{V_n}, \tag{7}$$

where V_n is the variance of the Kendall τ statistic under the hypothesis of no regression (and is a known quantity). This statistic has, under the hypothesis of homogeneity of β_T and β_S, closely central **chi-square distribution** with one **degree of freedom**. $\mathcal{L}$ is used as a suitable test statistic for testing the validity of the fundamental assumption of a parallel-line assay where the normality of the error components is not that crucial. In that sense it is a robust test. Moreover, having obtained the pooled estimator $\hat{\beta}$ of β, under the hypothesis of homogeneity of the slopes, we consider the residuals

$$\hat{Y}_{Ti} = Y_{Ti} - \hat{\beta}x_i, \quad \hat{Y}_{Si} = Y_{Si} - \hat{\beta}x_i, \tag{8}$$

for different i, and treating them as two independent samples, as in the case of dilution direct assays, we use the Wilcoxon–Mann–Whitney rank-sum test statistic to estimate the difference of the intercepts $\alpha_T - \alpha_S$ in a robust manner. As in the direct dilution assay, this estimator is the median of the differences of all possible pairs of **residuals** from the test and standard preparation respectively. A robust, **consistent** and asymptotically normally distributed estimator of $\log \rho$ is then obtained by dividing this estimator by the pooled estimator $\hat{\beta}$.

For drawing a confidence interval for $\log \rho$ (and hence, ρ), we can then use the Fieller Theorem by an appeal to the asymptotic normality of the estimator, or as in [21], consider a rectangular confidence set for β and $\alpha_T - \alpha_S$ by computing a coordinate-wise confidence interval for each with coverage probability $1 - \gamma/2$, and as in Figure 1, draw a robust confidence set for $\log \rho$ with coverage probability $1 - \gamma$.

Though this does not have an exact coverage probability, it is quite robust and works out well even for quite nonnormal error distributions. In the above setup, instead of the Kendall τ and the two-sample rank-sum statistics, we may use a general linear rank statistic for regression and a two-sample linear rank statistic for difference of location parameters, and obtain similar robust estimation and testing procedures. It is also possible to use general (aligned) M-statistics for this purpose (*see* **Robust Regression**). In general, such solutions

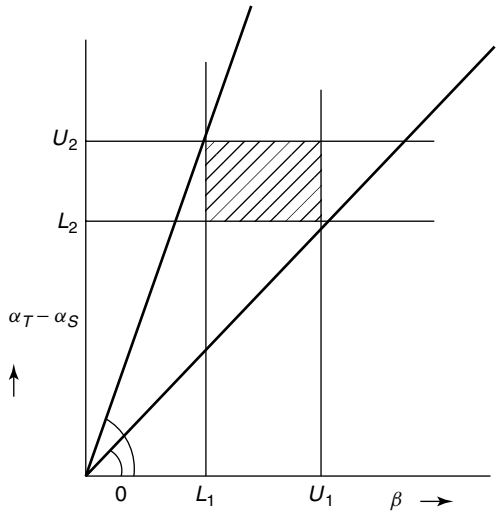

Figure 1 Graphical procedure for obtaining a nonparametric confidence interval for the log potency ratio in a parallel-line assay

are to be obtained by iterative methods, and hence, for simplicity and computational ease, we prescribe the use of the Kendall tau and two-sample rank- sum statistics for the desired statistical inference.

Next, we proceed to the case of slope-ratio assays, and consider first a nonparametric test for the validity of the fundamental assumption (of a common intercept but possibly different slopes). We define the Kendall tau statistics $K_T(b)$ and $K_S(b)$ as in the case of the parallel-line assay, and equating them to 0, we obtain the corresponding estimates of β_T and β_S, which are denoted by $\hat{\beta}_T$ and $\hat{\beta}_S$ respectively. Consider then the residuals

$$\tilde{Y}_{Ti} = Y_{Ti} - \hat{\beta}_T x_i, \quad \tilde{Y}_{Si} = Y_{Si} - \hat{\beta}_S x_i, \ \forall i. \tag{9}$$

We pool all these residuals into a combined set, and use the **Wilcoxon signed-rank** statistic to derive the corresponding rank estimator of the hypothesized common value of the intercept; this estimator, denoted by $\tilde{\alpha}$, is the median of all possible midranges of the set of residuals listed above. Let then $\hat{Y}_{Ti} = \tilde{Y}_{Ti} - \tilde{\alpha}$, $\hat{Y}_{Si} = \tilde{Y}_{Si} - \tilde{\alpha}$, $\forall i$, and for each preparation, based on these residuals, we consider the Wilcoxon signed-rank statistic. These are denoted by $\hat{W}_T$ and $\hat{W}_S$ respectively. As in the case of parallel-line assays, here we consider a test statistic for testing the validity of the fundamental assumption as

$$\mathcal{L} = \frac{\{\hat{W}_T^2 + \hat{W}_S^2\}}{V_n}, \tag{10}$$

where V_n is the variance of Wilcoxon signed-rank statistic under the hypothesis of symmetry of the distribution around 0 (and is a known quantity). When the fundamental assumption holds, the distribution of $\mathcal{L}$ is close to the central chi-square distribution with 1 degree of freedom, and hence a test can be carried out using the percentile point of this chi-square law. This test is quite robust and the underlying normality of the errors may not be that crucial in this context. Note that for the slope-ratio assay, granted the fundamental assumption of a common intercept, a natural plug-in estimator of ρ is given by

$$\hat{\rho} = \left\{ \frac{\hat{\beta}_T}{\hat{\beta}_S} \right\}^{1/\lambda}. \tag{11}$$

We may use the Fieller Theorem under an asymptotic setup to construct a confidence interval for ρ. Alternatively, as in the case of a parallel line assay, for a given $\gamma (0 < \gamma < 1)$, we may consider a distribution-free confidence interval of coverage probability $1 - \gamma/2$ for each of the two slopes β_T and β_S, and obtain a confidence interval for ρ^λ (and hence ρ). The situation is quite comparable to the Figure for the parallel-line assay, excepting that β_T and β_S are taken for the two axes. Here also, instead of the Kendall tau statistics and the Wilcoxon signed-rank statistics, general regression rank statistics and (aligned) signed-rank statistics (or even suitable M-statistics) can be used to retain robustness of the procedures without sacrificing much efficiency. However, the solutions are generally to be obtained by iterative procedures, and hence, we prefer to use the simpler procedures considered above.

Indirect Quantal Assays

In this type of (indirect) assays, the response is *quantal* (i.e. all or nothing) in nature. For each preparation (T or S) and at each level of administered dose, among the subjects, a certain number manifest the response while the others do not; these frequencies are stochastic in nature and their distribution depends on the dose level and the preparation (*see* **Quantal Response Models**). Thus, for a given dosage x, we denote by $F_T(x)$ and $F_S(x)$ the probability of the response for the test and standard preparation respectively. It is customary to assume that both $F_T(x)$ and $F_S(x)$ are monotone increasing in x, and for each $\alpha (0 < \alpha < 1)$, there exits unique solutions of the following

$$F_T(\xi_{T\alpha}) = \alpha, \quad \text{and} \quad F_S(\xi_{S\alpha}) = \alpha, \tag{12}$$

so that $\xi_{T\alpha}$ and $\xi_{S\alpha}$ are the α-**quantile** of the test and standard preparation; they are termed the $100\alpha\%$ effective dosage. In particular, for $\alpha = 1/2$, they are termed the **median effective dosage**. Whenever the response relates to death (as is usually the case with animal and toxicologic studies), the $\xi_{T\alpha}$, $\xi_{S\alpha}$ are also termed $100\alpha\%$-*lethal dosage*. In many studies, generally, low dosages are contemplated so that α is chosen to be small. This is particularly the case with **radioimmunoassays**, and we shall comment on that later on. Estimation of the $\xi_{T\alpha}$ and $\xi_{S\alpha}$ with due attention to their interrelations is the main task in a quantal assay. The concept of parallel-line and slope-ratio assays, as laid down for indirect quantitative assays, is also adoptable in quantal assays, and a

detailed account of the parametric theory based on normal, lognormal, logistic, and other notable forms of the distribution $F_T(x)$ is available with Finney [6, Chapter 17]. In this context, the *probit* and *logit* analyses are particularly notable, and we shall discuss them as well.

To set the ideas, we consider a single preparation at $k(\geq 2)$- specified dosage $d_1, \ldots, d_k$, where $d_1 < d_2 < \cdots < d_k$. Suppose that the dosage d_i has been administered to n_i subjects, out of which r_i respond positively while the remaining $n_i - r_i$ do not, for $i = 1, \ldots, k$. In this setup, the d_i, n_i are nonstochastic, while the r_i are random. The probability of a positive response at dosage d_i, denoted by $\pi(d_i)$, is then expressed as

$$\pi(d_i) = \pi(\theta + \beta d_i), \quad i = 1, \ldots, k, \tag{13}$$

where θ and β are unknown (intercept and regression) parameters, and $\pi(x)$, $-\infty < x < \infty$, is a suitable distribution function. In a parametric mold, the functional form of $\pi(.)$ is assumed to be given, while in nonparametrics, no such specific assumption is made. Note that the joint probability law of $r_1, \ldots, r_k$ is given by

$$\prod_{i=1}^{k} \binom{n_i}{r_i} \pi(\theta + \beta d_i)^{r_i} [1 - \pi(\theta + \beta d_i)]^{n_i - r_i}, \tag{14}$$

so that the **likelihood** function involves only two unknown parameters θ and β. The log-likelihood function or the corresponding estimating equations are not linear in the parameters, and this results in methodological as well as computational complications (*see* **Optimization and Nonlinear Equations**).

If $\pi(.)$ is taken as a **logistic distribution**, that is, $\pi(x) = \{1 + e^{-x}\}^{-1}$, then we have from the above discussion

$$\log \left\{ \frac{\pi(d_i)}{[1 - \pi(d_i)]} \right\} = \theta + \beta d_i, \quad i = 1, \ldots, k. \tag{15}$$

This transformation, known as the *logit* transformation, relates to a linear regression on the dosage, and simplifies related statistical analysis schemes. Thus, at least intuitively, we may consider the sample logits

$$Z_i = \log \left\{ \frac{r_i}{n_i} - r_i) \right\}, \quad i = 1, \ldots, k, \tag{16}$$

and attempt to fit a linear regression of the Z_i on d_i (*see* **Logistic Regression**). In passing, we may remark that technically r_i could be equal to zero or n_i (with a positive probability), so that Z_i would assume the values $-\infty$ and $+\infty$ with a positive probability, albeit for large n_i, this probability converges to zero very fast. As in practice, the n_i may not be all large; to eliminate this impasse, we consider the Anscombe correction to a **binomial** variable, and in (16), modify the Z_i as

$$Z_i = \log \left\{ \frac{(r_i + \frac{3}{8})}{(n_i - r_i + \frac{3}{8})} \right\}, \quad i = 1, \ldots, k. \tag{17}$$

Though the r_i have binomial distributions, the Z_i have more complex probability laws, and computation of their exact mean, variance, and so on, is generally highly involved. For large values of the n_i, we have the following

$$\sqrt{n_i}(Z_i - \theta - \beta d_i) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \{\pi(d_i)[1 - \pi(d_i)]\}^{-1}), \tag{18}$$

for each $i = 1, \ldots, k$, where the unknown $\pi(d_i)$ can be consistently estimated by the sample proportion $p_i = r_i/n_i$. Thus, using the classical *weighted* **least squares** *estimation* (WLSE) methodology (*see* **Categorical Data Analysis**), we may consider the quadratic norm

$$Q(\theta, \beta) = \sum_{i=1}^{k} n_i p_i (1 - p_i) \{Z_i - \theta - \beta d_i\}^2, \tag{19}$$

and minimize this with respect to θ, β to obtain the WLS estimators. Although the logit transformation brings the relevance of **generalized linear models** (GLM), the unknown nature of their variance functions makes the WLSE approach more appropriate for the suggested statistical analysis. In any case, the asymptotic flavor should not be overlooked.

If $\pi(x)$ is taken as the standard normal distribution function $\Phi(x)$, whose density function is denoted by $\phi(x)$, then we may consider the transformation

$$Z_i = \Phi^{-1}\left(\frac{r_i}{n_i}\right), \quad i = 1, \ldots, k, \tag{20}$$

known as the *probit* or *normit* transformation. Here also, it would be better to modify Z_i as

$$Z_i = \Phi^{-1}\left(\frac{(r_i + \frac{3}{8})}{(n_i + \frac{1}{2})}\right), \quad i = 1, \ldots, k. \tag{21}$$

Note that by assumption, $\Phi^{-1}(\pi(d_i)) = \theta + \beta d_i,\ i = 1, \ldots, k$, and this provides the intuitive appeal for a conventional linear regression analysis. However, the likelihood approach based on the product-binomial law encounters computational difficulties and loses its exactness of distribution theory to a greater extent. Here also, we would have complications in the computation of the exact mean, variance, or distribution of the Z_i, and hence, as in the logit model, we consider a WLSE approach in an asymptotic setup where the n_i are large. By virtue of the asymptotic normality of the $\sqrt{n_i}(p_i - \pi(d_i))$ (where again we take $p_i = (r_i + 3/8)/(n_i + 1/2)$), we obtain that for every $i \geq 1$,

$$\begin{aligned} &\sqrt{n_i}[Z_i - \theta - \beta d_i] \\ &\xrightarrow{\mathcal{D}} \mathcal{N}(0, \frac{\pi(d_i)[1 - \pi(d_i)]}{\phi^2(\Phi^{-1}(\pi(d_i)))}), \end{aligned} \tag{22}$$

so that we consider quadratic norm in a WLSE formulation

$$Q(\theta, \beta) = \sum_{i=1}^{k} \frac{n_i \phi^2(\Phi^{-1}(p_i))}{p_i(1 - p_i)}[Z_i - \theta - \beta d_i]^2, \tag{23}$$

and minimizing this with respect to θ, β, we arrive at the desired estimators.

For both the logit and probit models, the resulting estimators of θ, β are linear functions of the Z_i with coefficients depending on the n_i and the p_i. Therefore, the asymptotic normality and other properties follow by standard statistical methodology (*see* **Bartlett's Test**). Moreover the (asymptotic) dispersion matrix of these estimators, in either setup, can be consistently estimated from the observational data sets. Thus, we have the access to incorporate standard asymptotics to draw statistical conclusions based on these estimators.

Let us then consider the case of quantal bioassays involving two preparations (S and T), and for each preparation, we have a setup similar to the single preparation case treated above. The related parameters are denoted by θ_S, β_S and θ_T, β_T respectively, and for modeling the response distributions, we may consider either the logit or probit model, as has been discussed earlier. If we have a parallel-line assay, as in the case of an indirect assay, we have then

$$\beta_T = \beta_S = \beta \text{ unknown, and } \theta_T - \theta_s = \beta \log \rho, \tag{24}$$

so that based on the estimates $\hat{\theta}_S, \hat{\beta}_S, \hat{\theta}_T$ and $\hat{\beta}_T$, along with their estimated dispersion matrix, we can incorporate the WLSE to estimate the common slope β and the intercepts θ_S and θ_T. The rest of the statistical analysis is similar to the case of indirect assays. Moreover, this WLSE methodology is asymptotically equivalent to the classical likelihood-function–based methodology, so it can be regarded, computationally, as a simpler substitute for a comparatively complicated one. For a slope-ratio assay, we have similarly a common intercept while the ratio of the slopes provide the measure of the relative potency, and hence, the WLSE based on the individual preparation estimators can be adopted under this restriction to carryout the statistical analysis as in the case of an indirect assay.

Besides the logit or probit method, there are some other quasi-nonparametric methods, of rather an ad hoc nature, and among these, we may mention of the following estimators of the median effective dosage:

(i) The Spearman–Kärber estimator;
(ii) The Reed–Muench estimator, and
(iii) The Dragstedt–Behrens estimator.

These procedures are discussed in [7], p. 43. If the tolerance distribution is symmetric, the Spearman–Kärber estimator estimates the median effective dosage closely; otherwise, it may estimate some other characteristic of this distribution. Miller [13] studied the relative (asymptotic) performance of these three estimators, casting light on their bias terms as well. From a practical point of view, none of these estimators appears to be very suitable. Rather, if the $\pi(d_i)$ do not belong to the extreme tails (i.e. are not too small or close to 1), the logit transformation provides a robust and computationally simpler alternative, and is being used more and more in statistical applications. In passing, we may remark that

Finney [7, Chapter 10] contains some other techniques that incorporate modifications in the setup of usual quantal assays, such as the numbers n_i being unknown and possibly random, multiple (instead of binary) classifications, errors in the doses. In the following chapter, he also introduced the case of doses in mixtures that require a somewhat extended model and more complex statistical designs and analysis schemes. We shall comment on these below.

Stochastic Approximation in Bioassay

In the context of a quantal assay, we have the dosage-response model in terms of the tolerance distribution $\pi(d)$, and the median effective (lethal) dosage, LD50, is defined by the implicit equation $\pi(LD50) = 0.50$. In this context, for each preparation (standard or test), corresponding to initial dosage levels $d_1, \ldots, d_k$, we have estimates $p(d_1), \ldots, p(d_k)$ of the unknown $\pi(d_1), \ldots, \pi(d_k)$. We may set

$$p_i = \pi(d_i) + e(d_i), \quad i = 1, \ldots, k, \tag{25}$$

where the errors are (for large n_i, the number of subjects treated) closely normally distributed with zero mean and variance $n_i^{-1}\pi(d_i)[1 - \pi(d_i)]$. On the basis of this initial response data, we can choose an appropriate d_o for which the corresponding $p(d_o)$ is closest to $1/2$. Then, we let $d_{(1)} = d_o + a_o[p(d_o) - 1/2]$, for some $a_o > 0$, and recursively we set

$$d_{(j+1)} = d_{(j)} + a_j \left[p(d_{(j)}) - \frac{1}{2} \right], \quad \text{for some } \; a_j > 0; \; j \geq 0. \tag{26}$$

(*see* **Up-and-Down Method**) The aim of this **stochastic approximation** procedure, due to Robbins and Monro [15], is to estimate the LD50 without making an explicit assumption on the form of the tolerance distribution $\pi(d)$. But in this setup, the $p(d_{(j)})$ as well as the $d_{(j)}$ are stochastic elements, and for the convergence of this stochastic iteration procedure, naturally, some regularity conditions are needed on the $\{a_i; i \geq 0\}$ and $\pi(d)$ around the LD50. First of all, in order that the iteration scheme terminates with a consistent estimator of the LD50, it is necessary that the a_i converge to zero as i increases. More precisely, it is assumed in this context that

$$\sum_{n \geq 0} a_n \text{ diverges to } +\infty, \quad \text{but } \sum_{n \geq 0} a_n^2 \; < \; +\infty. \tag{27}$$

In addition, the continuity and positivity of the density function corresponding to the distribution function $\pi(x)$ at the population LD50 is also a part of the regularity assumptions. Further assumptions are needed to provide suitable (stochastic) rates of convergence of the estimator of the LD50 and its asymptotic normality and related large sample distributional properties. Once the LD50 values are estimated for each preparation, we may proceed as in the case of a quantal assay, and draw conclusions about the relative potency and other related characteristics. It is not necessary to confine attention specifically to the LD50, and any LD100α, for $\alpha \in (0, 1)$ can be treated in a similar fashion. In fact, Kiefer and Wolfowitz [12] considered an extension of the Robbins–Monro stochastic approximation procedure that is aimed to locate the maximum (or minimum) of a dose–response function that is not necessarily (piecewise or segmented) linear but is typically nonmonotone, admitting a unique extremum (maximum or minimum) of experimental importance. Such dose–response regressions arise in many toxicologic studies where a turn occurs at an unknown level. Often this is treated in a general **change-point** model framework. The main advantage of the stochastic approximation approach over the classical quantal assay approach is that no specific assumption is generally needed on the probability function $\pi(d)$, so that the derived statistical conclusions remain applicable in a much wider setup. On the other hand, the stochastic iteration scheme generally entails a larger number of subjects on which to administer the study, and often that may run contrary to the practicable experimental setups, especially with respect to cost considerations. In this general context, a significant amount of methodological research work has been carried out during the past 40 years, and an extensive review of the literature on stochastic approximation is made by Ruppert [16] where the relevant bibliography has also been cited. The scope of stochastic approximation schemes is by no means confined to quantal assays; they are also usable for quantitative bioassays, and even to other problems cropping up in far more general setups.

Radioimmunoassay

In *radioimmunoassays* antigens are labeled with radioisotopes, and in *immunoradiometric assays* antibodies are labeled. For a broad range of antigens, such radioligand assays enable the estimation of potency from very small quantities of materials and usually with high precision. Radioligand assays are based upon records of radiation counts in a fixed time at various doses, so that potency estimation involves the relation between counts of radioactivity and dose, generally both at low levels [8]. In many such studies, the regression function of the count of radioactivity on dose has been found to be satisfactorily represented by a logistic curve; however, the lower and upper asymptotes of such a curve are not necessarily equal to zero and one, but are themselves unknown parameters. This difference with the classical logistic distribution is reflected in a somewhat different form of the variance function of radiation counts. Unlike the **Poisson process**, the variance function may not be equal to the mean level of the radiation counts $U(d)$ (i.e. their expectation at a given dose level d); in many studies, it has been experimentally gathered that the variance function $V(d)$ behaves like $[U(d)]^{\lambda}$, where $\lambda(>0)$ typically lies between 1 and 2. For this reason, the usual **Poisson regression** model in *generalized linear models* (GLM) methodology may not be universally applicable in radioimmunoassays. Moreover, such radioligand assays may not be regarded as strictly bioassays, since they may not depend upon responses measured in living organisms or tissues. However, the advent of the use of biologic *markers* in mutagenesis studies and in molecular genetics, particularly during the past 20 years, has extended the domain of statistical perspectives in radioligand assays to a much wider setup of investigations, and strengthened the structural similarities between radioimmunoassays and the classical bioassays. They involve statistical modeling and analysis schemes of very similar nature, and in this sense, their relevance in a broader setup of bioassays is quite appropriate (*see* **Radioimmunoassay**).

Dosimetry and Bioassay

As has been noted earlier, a dose–response model exhibits the (mathematical) relationship between an amount of exposure or treatment and the degree of a biological or health effect, generally a measure of an adverse outcome. Bioassay and clinical trials are generally used in such dose–response studies. With the recent advances in **pharmacoepidemiology** (*see* **Dose-response in Pharmacoepidemiology**) as well as in risk analysis (*see* **Dose–Response Models in Risk Analysis**), bioassays have led to another broader domain of statistical appraisal of biological dose–response studies, known as *dosimetry* (or animal study). Pharmacoepidemiology rests on the basic incorporation of *pharmacodynamics* (PD) and *pharmacokinetics* (PK) in the development of the so called *structure–activity relationship information* (SARI) (*see* **Chemometrics**). Though a PD model directly relates to a dose–response model, the PK actions of the exposure or drug needs to be taken into account in the dose–response modeling. This is now done more in terms of SARI where the structure refers to the dose factors and activity refers to the biological reactions that follow the exposure (dose) to a specific species or organism. In a majority of cases, the target population is human, but owing to various ethical and other experimental constraints, human beings may not be usable to the full extent needed for such a dose–response modeling. As such, animal studies are often used to gather good background information, which is intended for incorporation in human studies in bioassay and clinical trials. Dosimetry pertains to this objective.

Dosimetry models intend to provide a general description of the uptake and distribution of inhaled (or ingested or absorbed) toxics (or compounds having adverse health effects) on the entire body system. For judgment on human population, such dosimetric models for animal studies need to be *extrapolated* with a good understanding of the *interspecies differences* (*see* **Dose–Response Models in Risk Analysis**). SARI is a vital component in enhancing such statistical validation of pooling the information from various animal studies and extrapolating to the human population. Most dose–response relationships are studied and through well-controlled animal bioassays with exposure or dose levels generally much higher than typically perceived in human risk analysis. In this respect, dosimetry is directly linked to bioassay, though in dosimetry, the SARI is more intensively pursued to facilitate extrapolation (*see* **Extrapolation, Low Dose**). PDPK aspects not only may vary considerably from subhuman primates to human beings, but also there is much less of control in

human exposure to such toxics. Also, metabolism in the human being is generally quite different from that in subhuman primates. An important element in this context is the *environmental burden of disease* (EBD) factor that exhibits considerable interspecies variation as well as geopolitical variation. Hence, ignoring the SARI part, a conventional dose–response model for a subhuman primate may not be of much help in depicting a similar model for human exposure. For the same reason, conventional statistical extrapolation tools may be of very limited utility in this interspecies extrapolation problems [25]. Finally, in many carcinogenicity studies, it has been observed that *xenobiotic* effects underlie such dose–response relations, and this is outlined in a later section (*see pharmacogenomics*).

Semiparametrics in Bioassays

The GLM methodology has been incorporated in a broad variety of statistical modeling and analysis schemes pertaining to a wide range of applications, and bioassays are no exceptions. Going back to the direct dilution assays, if we had taken both the distributions, F_S and F_T, as **exponentials** with respective means μ_S and μ_T, then the two distributions would have constant **hazard rates** $1/\mu_S$ and $1/\mu_T$ respectively, so that the relative potency ρ is equal to the ratio of the two hazard rates. Inspired by this observation, and by the evolution of the Cox [3] **proportional hazard model** (PHM) (*see* **Cox Regression Model**), research workers have attempted to relate the two survival functions $S_S(x) = P\{X_S > x\}$ and $S_T(x) = P\{X_T > x\}$ as

$$S_T(x) \;=\; [S_S(x)]^{\rho}, \quad x \geq 0, \tag{28}$$

and interpret ρ as the relative potency of the test preparation with respect to the standard one (*see* **Lehmann Alternatives**). Though this representation enables one to import the PHM-based statistical analysis tools for the estimation of the relative potency, for distributions other than the exponential ones, the interpretation of "dilution assays" may no longer be tenable under such a PHM. There is an alternative interpretation in terms of the parallelism of the two log-hazard functions, but that may not fit well with the fundamental assumption in dilution assays. For some related statistical analysis of bioassays based on GLM methodologies, we refer to [24], where indirect bioassays have also been treated in the same manner along with the classical parametrics.

Nonparametrics in Bioassays

The estimators of relative potency and tests for fundamental assumptions in dilution (direct as well as indirect) assays based on rank statistics, considered earlier, spark the first incorporation of nonparametrics in biological assays. However, these may be characterized more in terms of **semiparametrics**, in the sense that the assumed linearity of dose–response regressions was essentially parametric in nature, while the unknown form of the underlying tolerance distribution constitutes the nonparametric component. Thus, together they form the so-called semiparametric models. It is possible to incorporate more nonparametrics in bioassays mostly through the **nonparametric regression** approach. For direct dilution assays, such nonparametric procedures are quite simple in interpretation and actual formulation. We consider the log-dose transformation, so that the dosage for the test and standard preparations have the distributions $F_T^*(x)$ and $F_S^*(x)$, respectively, where $F_T^*(x) = F_S^*(x + \log \rho)$, for all x. If we denote the p-quantile of F_T^* and F_S^* by $Q_T(p)$ and $Q_S(p)$ respectively, then we have

$$Q_S(p) \;-\; Q_T(p) = \log \rho, \;\; \forall \; p \in (0, 1), \tag{29}$$

so that the well-known Q–Q *plot* for the two preparations results in a linear regression form, and this provides the statistical information to test for this fundamental assumption as well as to estimate the relative potency. A similar conclusion can also be drawn from a conventional P–P *plot* (*see* **Graphical Displays**). The classical Kolmogorov–Smirnov statistics (in the two-sample case) can be used for drawing statistical conclusions, and we may refer to Rao and Littell [14] for some related work. The situation is a bit more complex with indirect assays. In the classical parametric setup, we work with the expected response at different dosages, assuming of course a linear regression. In a semiparametric approach, this linearity of dosage-response regression is taken as a part of the basic assumption, but the distribution of the errors is allowed to be a member of a wider class, so that robust procedures based on rank or M-statistics are advocated instead of the classical WLSE. In a pure nonparametric setup, the linearity of the

dosage-response regression is not taken for granted. Therefore the two dosage-response regression functions may be of quite arbitrary nature, and yet parallel in an interpretable manner. The statistical task is therefore to assess this parallelism without imposing linearity or some other parametric forms. Here also, at a given dosage level, instead of the mean response level, we may consider median or a p-quantile, and based on such robust estimators, we draw statistical conclusions allowing the quantile functions to be of a rather arbitrary nature. Asymptotics play a dominant role in this context, and often this may require a relatively much larger sample size. On the other hand, in terms of robustness and validity, such pure nonparametric procedures have a greater scope than parametric or semiparametric ones.

Bioavailability and Bioequivalence Models

As has been explained earlier bioequivalence trials differ from conventional bioassays, as here, generally, the active substances in the drug are the same but the differences in bioavailability reflect the form and manner of administration. Such alternative modes may therefore call for additional restraints in the statistical formulation, and because of anticipated biological equivalence, there is less emphasis on relative potency and more on general equivalence patterns. For such reasons, regulatory requirements for establishing *average bioequivalence* of two preparations (that are variations of an essentially common drug) relate to a verification of the following:

A *confidence interval* for the relative potency, having the *confidence limits* ρ_L, ρ_U, lies between two specified endpoints, say $\rho_o < 1 < \rho^o$, with a high *coverage probability* (or *confidence coefficient*) γ. Generally, γ is chosen close to 1 (namely, 0.95), and also $\rho^0 = (\rho_o)^{-1}$ is very close to one.

These requirements in turn entail a relatively large sample size, and therefore, (group) sequential testing procedures (*see* **Sequential Analysis**) are sometimes advocated [9]. For general considerations underlying such bioequivalence trials, we refer to [1, 11, 29], where other pertinent references are cited. Generally, such statistical formulations are more complex than the ones referred to earlier.

As has been mentioned earlier, the term bioassay is used in a more general form, and this is equally true for bioequivalence and bioavailability models. Kinetic measures of bioavailability and pharmacokinetic parameters have been developed to meet the demand for such recent usage (*see* **Bioavailability and Bioequivalence**). We will illustrate this somewhat differently with *pharmacogenomics*, which is revolutionizing the field of *bioinformatics* and experiments with biological units, in general.

Pharmacogenomics in Modern Bioassays

Following Ewens and Grant [5], we take bioinformatics to mean the emerging field of science growing from the application of mathematics, statistics, and information technology, including computers and the theory surrounding them, to study and analysis of very large biological and, in particular, genetic data sets (*see* **Genetic Markers**). Having its genesis 50 years ago [28], the field has been fueled by the immense increase in the DNA data generation (*see* **DNA Sequences**). Earlier interpretation of bioinformatics with emphasis on *computational biology* by Waterman [27] also merits serious considerations, while Durbin et al. [4] had a view point geared by computer **algorithms** along with some heuristic usage of **hidden Markov models**.

At the current stage, gene scientists cannot scramble fast enough to keep up with the genomics, with developments emerging at a furious rate and in astounding detail. Bioinformatics, at least at this stage, as a discipline, does not aim to lay down some fundamental mathematical laws (which might not even exist in such a biological diversity). However, its utility is perceived in the creation of innumerable computer graphics and algorithms that can be used to analyze exceedingly large data sets arising in bioinformatics. In this context, naturally **data mining** and *statistical learning* tools (under the terminology *Knowledge Discovery and Data Mining* (KDDM)) are commonly used [10], though often in a heuristic rather than objective manner. There could be some serious drawbacks of statistical analysis based on such KDDM algorithms alone, and *model selection* has emerged as a challenging task in bioinformatics (*see* **Model, Choice of**).

Given the current status of bioinformatics as the information technology (advanced computing) based discipline of analyzing exceedingly high dimensional data with special emphasis on *genomics*, and that genomics looks at the vast network of genes, over

time, to determine how they interact, manipulate, and influence biological pathways, networks, as well as physiology, it is quite natural to heed to *genetic variation* (or **polymorphism**) in most studies involving biological units. Moreover, because of the drug-response relationship, basic in bioassay, it is natural to appraise the role of pharmacogenomics in this setup. Pharmacology is the science of drugs including materia medica, toxicology, and therapeutics, dealing with the properties and reactions of drugs, especially with relation to their therapeutic values. In the same vein, *pharmacodynamics*, a branch of pharmacology, deals with reactions between drugs and living structures; *pharmacokinetics* relates to the study of the bodily absorption, distribution, metabolism, and excretion of drugs. In bioequivalence trials, these tools have already been recognized as fundamental. *Pharmacogenetics* deals with genetic variation underlying differential response to drugs as well as drug metabolism. The whole complex constitutes the discipline: *Pharmacogenomics*. In the same way, *Toxicogenomics* relates to the study of gene-environmental interactions in disease and dysfunction to cast light on how genomes respond to environmental stress or toxics.

It is conceived that there are certain genes that are associated with disease phenotype, side effects, and drug efficacy. Also, because of inherent (genetic) variations and an enormously large number of genes as well as a very large pool of diseases and disorders, there is a genuine need of statistical methods to assess the *genetic mapping of disease genes*. Pharmacotoxicogenomics is therefore destined to play a fundamental role in biological assays, in the years to come.

Complexities in Bioassay Modeling and Analysis

There are generally other sources of variations, which may invalidate the use of standard statistical analysis schemes in bioassays to a certain extent. Among these factors, special mention may be made of the following:

(a) *Censoring of various types*,
(b) *Differentiable / Nondifferentiable measurement errors*,
(c) *Stochastic compliance of dose*,
(d) *Correlated multivariate responses*, and
(e) *Curse of dimensionality in genomics*.

It is generally assumed that **censoring** is usually of *Type I* (truncation of the experiment at a prefixed timepoint), *Type II* (truncation following a certain prefixed number or proportion of responses), and *random*, where the censoring time and response time are assumed to be stochastically independent, and moreover, the censoring is assumed to be *noninformative*, so that the censoring time distribution remains the same for both the preparations. In actual practice, this may not be generally true, and hence, effects of departures from such assumptions on the validity and efficacy of standard statistical procedures are therefore needed to be assessed. Measurement of the actual dose levels in quantal assays, or the response levels in an indirect assay may often be impaired to a certain extent by measurement errors. In statistical analysis, usually such measurement errors are assumed to be either *differentiable* or *nondifferentiable* type, and appropriate statistical models and related analysis schemes depend on such assumptions. In radioimmunoassays, dosimetric studies in pharmacokinetics, as well as in other types, not the full amount of a prescribed dose may go into the organ or experimental unit, and the actual consumption of the dose may be (often, highly) stochastic in nature. Therefore, the dose–response regression relation may be subject to *nonidentifiability* (*see* **Identifiability**)and **overdispersion** effects. This calls for more modifications of existing models and analysis schemes. Finally, when there are **multiple endpoints** with possibly binary or **polytomous** responses, a dimension reduction for the model-based parameters becomes necessary from statistical modeling and inference perspectives. Otherwise, an enormously large sample size may be needed to handle adequately, the full parameter model, and this may run contrary to the practical setup of an assay. The situation is worse when some of the responses are quantitative while the others are quantal or at best polychotomous. These naturally introduce more model complexities and call for more complicated statistical analysis tools.

References

[1] Anderson, S. & Hauck, W.W. (1990). Considerations of individual bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics* **18**, 259–273.
[2] Bliss, C.I. (1952). *The Statistics of Bioassay*. Academic Press, New York.

[3] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society B* **34**, 187–220.
[4] Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
[5] Ewens, W.J. & Grant, G.R. (2001). *Statistical Methods in Bioinformatics: An Introduction*. Springer-Verlag, New York.
[6] Finney, D.J. (1964). *Statistical Methods in Biological Assay*, 2nd Ed. Griffin, London.
[7] Finney, D.J. (1971). *Probit Analysis*, 3rd ed. University Press, Cambridge.
[8] Finney, D.J. (1976). Radioligand assay, *Biometrics* **32**, 721–730.
[9] Gould, A.L. (1995). Group sequential extensions of a standard bioequivalence testing procedure, *Journal of Pharmacokinetics and Biopharmaceutics* **23**, 57–86.
[10] Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.
[11] Hochberg, Y. (1955). On assessing multiple equivalences with reference to bioequivalence, in *Statistical Theory and Applications: Papers in Honor of H. A. David.*, H.N. Nagaraja, P.K. Sen & D.F. Morrison, eds. Springer-Verlag, New York, pp. 265–278.
[12] Kiefer, J. & Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function, *Annals of Mathematical Statistics* **23**, 462–466.
[13] Miller, R.G., Jr. (1973). Nonparametric estimators of the mean tolerance in bioassay, *Biometrika* **60**, 535–542.
[14] Rao, P.V. & Littell, R. (1976). An estimator of relative potency, *Communication of Statistics Series A* **5**, 183–189.
[15] Robbins, H. & Monro, S. (1951). A stochastic approximation method, *Annals of Mathematical Statistics* **22**, 400–407.
[16] Ruppert, D. (1991). Stochastic approximation, in *Handbook of Sequential Analysis*, B.K. Ghosh & P.K. Sen, eds. Marcel Dekker, New York, pp. 503–529.
[17] Sen, P.K. (1963). On the estimation of relative potency in dilution (-direct) assays by distribution-free methods, *Biometrics* **19**, 532–552.
[18] Sen, P.K. (1964). Tests for the validity of fundamental assumption in dilution (-direct) assays, *Biometrics* **20**, 770–784.
[19] Sen, P.K. (1965). Some further applications of nonparametric methods in dilution (-direct) assays, *Biometrics* **21**, 799–810.
[20] Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau, *Journal of the American Statistical Association* **63**, 1379–1389.
[21] Sen, P.K. (1971). Robust statistical procedures in problems of linear regression with special reference to quantitative bioassays, I, *International Statistical Review* **39**, 21–38.
[22] Sen, P.K. (1972). Robust statistical procedures in problems of linear regression with special reference to quantitative bioassays, II, *International Statistical Review* **40**, 161–172.
[23] Sen, P.K. (1984). Nonparametric procedures for some miscellaneous problems, in *Handbook of Statistics*, Vol. 4: Nonparametric Methods, P.R. Krishnaiah & P.K. Sen, eds. Elsevier, Holland, pp. 699–739.
[24] Sen, P.K. (1997). An appraisal of generalized linear models in biostatistical applications, *Journal of Applied Statistical Sciences* **5**, 69–85.
[25] Sen, P.K. (2003). Structure-activity relationship information in health related environmental risk assessment, *Environmetrics* **14**, 223–234.
[26] Shorack, G.R. (1966). Graphical procedures for using distribution-free methods in the estimation of relative potency in dilution (-direct) assays, *Biometrics* **22**, 610–619.
[27] Waterman, M.S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall, Cambridge.
[28] Watson, J.D. & Crick, F.H.C. (1953). Genetical implications of the structure of deoxyribonucleic acid, *Nature* **171**, 964–967.
[29] Westlake, W.J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations, in *Biopharmaceutical Statistics for Drug Developments*, K.E. Peace, ed. Marcel Dekker, New York, pp. 329–352.

Further Reading

Cox, C. (1992). A GLM approach to quantal response models for mixtures, *Biometrics* **48**, 911–928.
Moses, L.E. (1965). Confidence limits from rank tests, *Technometrics* **7**, 257–260.
Sen, P.K. (2002). Bioinformatics: statistical perspectives and controversies, in *Advances in Statistics, Combinatorics and Related Areas*, C. Gulati & S.N. Mishra, eds. World Science Press, London, pp. 275–293.

PRANAB K. SEN

Biological Standardization

Biological products are distinguished from chemical products by the biological nature, singly or in combination, of the source materials, of the production and purification procedures, and of the test methods needed to characterize such products or determine their potency. Biological products differ from chemical products in that they are not adequately characterized solely by their physical and chemical properties. Two biologicals may give the same results in chemical and physical tests but may have different activities when compared in biological tests. Measurement of the amount or concentration of any biological product thus requires an *in vivo* or *in vitro* **biological assay** system and a standard preparation or reference material. A biological standard for a product is a preparation such that the properties of a given amount of it do not change over time, and with which the properties of other samples of the product can be compared. Standards are essential if measurements are to be comparable from one assay to another. The potencies of the standard are customarily defined in arbitrary "units".

The use of biological activity as the basis for analytical, or assay, techniques developed rapidly in the late nineteenth and early twentieth centuries, with the development of vaccines and the discovery of vitamins and hormones. It was quickly recognized that biological test systems were variable, and the principles of biological standardization and quantitative approaches to biological test systems were formulated [6, 7, 10]. Central to these principles were the importance of comparison with a standard and the need to determine the variation of the biological system [2].

In some early uses of biological assays, attempts were made to define units in terms of the amount of product required to produce a specified effect. There are instances in which this continues today with, for example, attempts to define the amount of an antiviral agent in terms of the amount that will protect 50% of a population of cells, and to calibrate botulinum toxin for therapeutic use in terms of "mouse" units. However, it is virtually impossible to standardize a biological system so that whenever and wherever it is used the relation between the amount of a product or material and its response remains constant, and use of a standard thus leads to improved reproducibility [2, 9].

Any biological standard must fulfill certain conditions. All samples of the standard material must be uniform, so that the amount of material required to produce the observed effect or response is known. International standards are thus prepared in ampoules in such a way that the contents of any one ampoule are as nearly as possible identical to the contents of any other. The standard must be representative of the substance for which it is to serve as a standard, but does not necessarily have to be of high purity. A representative batch of product may be selected for the in-house standard by a manufacturer. For an international standard, suitability is usually shown by extensive characterization in an international collaborative study. The standard must be stable so that its effects do not change over time. Because of the nature of biological standards, direct tests for stability – that is, tests in real time – are not possible; predictions of stability may be based on the effects of storage at elevated temperatures on samples of the standard coupled with assumptions about the predictive nature of these effects. The quantity of standard available must be sufficient for the purposes for which it is intended. The **World Health Organization** has published guidelines for the preparation of standards detailing the way in which these conditions may be met, and setting specific requirements for international standards [13].

Biological standards serve a variety of specific purposes. Any laboratory routinely carrying out biological assays will include an in-house standard in these assays, and may also include "quality control samples" that might also be considered to be standards (*see* **Quality Control in Laboratory Medicine**). Although such standards provide comparability between assays within a laboratory, they may differ markedly between laboratories, as has been shown when a common sample is measured by several laboratories in terms of their individual standards (see, for example, [11]). Biological standards may be used in the development and validation of assay systems; for example, failure of an assay system to respond in a dose-related way to the standard would automatically invalidate the system for that material [12]. Standards are an essential part of the various quality assessment schemes which are operated for some types of assay, often by national authorities. International standards are indispensable not only for

measurement of biological materials but also for the controls necessary for modern medicines [8].

The World Health Organization (WHO) plays an important role in developing consensus guidelines on regulatory issues, and the WHO Constitution requires that WHO establish standards to assure the safety and efficacy of products used for the diagnosis, therapy, and prophylaxis of disease. This is a continuation of the development of international standards begun by the League of Nations [3–5]. Various international unions or federations representing individual scientific disciplines have also established commissions on issues related to measurement within their discipline, and may produce standards or recommend standards to WHO. Veterinary biologicals are also subject to regulation [1].

The World Health Assembly recommends that member states of WHO give official recognition to the International Standards and International Units, and publishes a list of these and their custodians [14], which is periodically updated. The majority of International Standards are held and distributed by the National Institute for Biological Standards (Blanche Lane, South Mimms, Potters Bar, Hertfordshire EN6 3QG, UK). The responsibility for the assurance of the quality and potency of biological products lies with the manufacturer, as regulated by the regional or national control authority and detailed in the national pharmacopoeias. The contribution of biological standardization to this assurance is widely recognized, and International Biological Standards are accepted as the primary standards for **calibration** of biologicals.

References

[1] Blancou, J. & Truszczynski, M. (1995). The role of international and regional organizations in the regulation of veterinary biologicals, *Revue Scientifique et Technique, Office International des Épizooties* **14**, 1193–1206.

[2] Burn, J.H. (1930). The errors of biological assay, *Physiological Reviews* **10**, 146–169.

[3] Cockburn, W.C. (1991). The international contribution to the standardization of biological substances. I. Biological standards and the League of Nations 1921–1946, *Biologicals* **19**, 161–169.

[4] Cockburn, W.C., Hobson, B., Lightbown, J.W., Lyng, J. & Magrath, D. (1991). The international contribution to the standardization of biological substances. II. Biological standards and the World Health Organization 1947–1990. General considerations, *Biologicals* **19**, 257–264.

[5] Cockburn, W.C., Hobson, B., Lightbown, J.W., Lyng, J. & Magrath, D. (1991). The international contribution to the standardization of biological substances. III. Biological standards and the World Health Organization 1947–1990. Specific activities and commentary, *Biologicals* **20**, 1–10.

[6] Dale, H. (1939). Biological standardization, *Analyst* **64**, 554–567.

[7] Irwin, J.O. (1950). Biological assays with special reference to biological standards, *Journal of Hygiene* **48**, 215–238.

[8] Jeffcoate, S.L., Corbel, M.J., Minor, P.D., Gaines Das, R.E. & Schild, G.C. (1993). The control and standardization of biological medicines, *Proceedings of the Royal Society of Edinburgh* **101B**, 207–226.

[9] McLellan, K., Gaines Das, R.E., Ekong, T.A.N. & Sesardic, D. (1996). Therapeutic Botulinum type A toxin: factors affecting potency, *Toxicon* **34**, 975–985.

[10] Miles, A.A. (1951). Biological standards and the measurement of therapeutic activity, *British Medical Bulletin* **7**, 283–291.

[11] Mire-Sluis, A.R., Gaines Das, R.E. & Padilla, A. (1997). Assisting the development of cytokines for research and as therapeutic agents: the WHO Cytokine Standardization Programme, *Journal of Immunological Methods*, in press.

[12] Storring, P.L. (1988). Biological assays of peptide and protein hormones, in *Improvement of Comparability and Compatibility of Laboratory Assay Results in Life Sciences*. A. Kallner, D. Bangham & D. Moss, eds. *Scandinavian Journal of Clinical and Laboratory Investigation*, Supplement 193.

[13] World Health Organization (1990). Annex 4: Guidelines for the preparation, characterization and establishment of international and other standards and reference reagents for biological substances, Technical Report Series, No. 800. WHO, Geneva.

[14] World Health Organization (1991). *Biological Substances, International Standards and Reference Reagents, 1990*. WHO, Geneva.

ROSE E. GAINES DAS

Biometrical Journal

The *Biometrical Journal* is an international journal for mathematical and statistical methods used in biological sciences in the widest sense: in biology, medicine, psychology, agriculture, forestry, ecology, and others. It is primarily addressed to mathematicians and statisticians working in these fields and to biologists and physicians having extensive contact with biometrical methods.

The *Biometrical Journal* was originally published by the Akademie Verlag, Berlin, and is now published by Wiley. The journal was founded in 1959 by Ottokar Heinisch and Maria Pia Geppert as the German publication *Biometrische Zeitschrift*. In the editorial to the first issue, the founding editors outlined the purpose of biometry as rendering objectivity to empirical results found in the biosciences by mathematical and statistical methods. This characterization of biometry as a joint methodologic concept for the various biological branches of the natural sciences has been well established by the development since 1959. The scope of biometry has been broadened since that time, as is apparent from the emergence of many new biometric methods in medicine, for example. Also, the number of researchers working in biometry has increased enormously, and the store of available biometric theories and practical methods has been extended systematically. The *Biometrical Journal* has participated in this flourishing of the biometric sciences in the past few decades. Consequently, it grew from four issues with a total of 290 pages in its first year to eight issues with 1024 pages by 1996, and a new format was introduced in 2004.

From 1967 to 1969, the journal was edited by Maria Pia Geppert and Erna Weber, from 1970 to 1988, by Erna Weber, and from 1988 to 1995, by Heinz Ahrens and Klaus Bellmann. In 1996, Jürgen Läuter assumed the post of the editor-in-chief, followed by Peter Bauer in 2000 and Edgar Brunner and Martin Schumacher in 2004.

In 1977, the journal's name was changed to *Biometrical Journal*. This change marked the continuous transition from German to English as the primary publication language and much increased the journal's international recognition. The relation with Germany and especially with the German and Austro-Swiss Regions of the **International Biometric Society (IBS)** remained strong without dominating the journal, and the journal is now truly international, as is immediately clear from a look at recent issues. Today, almost all published papers are in English, and contributions in the secondary publication languages, French and German, are a rare exception.

Original papers form the core of the *Biometrical Journal*'s contents, with the aim of trying to provide a link between theory and practice. Papers presenting new mathematical or statistical methods of potential benefit to the biosciences as well as interesting and original applications of existing theory to bioscientific problems are welcome, as are reviews and letters to the editors.

Special emphasis is laid on biomedical applications ranging from quantitative analysis in basic research to **clinical trials** methodology including **drug regulatory** aspects, and to methodological developments for research synthesis, which is of particular importance for the transfer of research results into practice (*see* **Meta-analysis of Clinical Trials**). An active discussion forum has been established.

With the support of a board of internationally well-renowned associate editors, who are chosen specialists in their respective areas of research, a competent and fast evaluation of manuscripts submitted online is guaranteed. Also, some initiatives have been introduced since January 1, 2004 as a part of the editors' intention to accelerate the submission and review process. Thus, electronic submission and handling has been made mandatory for authors and members of the editorial board.

The journal has its own website in order to provide broad accessibility to its online version for the readers to enable them to gain immediate and faster reach. An advanced subject-specific search facility is provided on the website. Authors can use this extensive, full-text search facility on a wide range of journals published by Wiley in order to learn about recent developments in their areas of research and interests, which helps them in preparing their articles.

The website, `www.biometrical-journal.de` provides more details on submission procedures, contact information, templates for manuscripts, and other relevant information for authors and readers.

JÜRGEN LÄUTER, EKKEHARD GLIMM,
EDGAR BRUNNER & MARTIN SCHUMACHER

Biometrics

Biometrics, a journal of the **International Biometric Society** (IBS), is published quarterly. The main objectives of the journal, which are listed at the IBS web site `http://www.tibs.org/`, are to promote and extend the use of mathematical and statistical methods in the biological sciences by describing and exemplifying developments in these methods and their application in a form accessible to statistical practitioners, experimenters, and others concerned with the analysis of data. The journal is intended to provide a medium for exchange of ideas by subject-matter specialists, those involved primarily with analysis of data and those concerned with the development of statistical methodology. Published papers may deal with statistical methodology applied to specific biological contexts, topics in mathematical biology, and statistical principles and methodology that are generally applicable to common challenges in the analysis of biological data.

Papers in the journal are currently organized into two main sections. The Regular Communications section includes statistical, authoritative, or review articles; and papers outlining development of novel statistical methods for the planning of experiments or interpretation of data, including demonstrations of utility and performance. Except for papers having to do with **experimental design**, which of necessity cannot refer to data that have not yet been collected, a centerpiece of most Regular Communications articles is a motivating and important data set exemplifying the scientific challenges on which methods are focused. The Consultant's Forum section presents papers illustrating the application of existing methods to new areas where they have not been previously used and permit new biological insights, papers clarifying or contrasting existing methods, or papers providing new guidance or tools for new or common data-analytic challenges.

The journal began in 1945 as the *Biometric Bulletin* of the **American Statistical Association**, and the name was changed in 1947 with Volume 3 onwards being called *Biometrics*. The journal continued to be published by the American Statistical Association until Volume 6, which was published by the Biometrics Society (now the IBS). Initially, the journal had an editor for Regular Communications and a Queries and Notes editor. In 1960, a book review section and dedicated editor were added. *Biometrics* continued in this form until 1974, with the Queries and Notes section becoming the Shorter Communications and Queries section, which was renamed as Shorter Communications in 1976. The Consultant's Forum section was added in 1977 with a view to include papers with novel data sets and novel use of existing methods. Some submissions offer specific comments regarding papers appearing in the journal previously; these are published as Reader Reaction articles. Authors of published articles to which these papers react are offered an opportunity to respond, and the reaction paper and response are published together.

The founding editor of *Biometrics* was **Gertrude M. Cox** (1945–1955), and the founding editor of the Queries and Notes/Shorter Communications section was **George W. Snedecor** (1945–1958). Table 1 lists all editors of the journal through 2004.

By 1999, the volume of submissions had increased to the point where it was no longer feasible for a single editor to handle the Regular Communications section; moreover, the distinction between this section and Shorter Communications had become blurred. Accordingly, on the basis of recommendations of an IBS committee charged with reviewing editorial structure, Shorter Communications was merged with the Regular Communications section; Consultant's Forum was maintained; and the journal adopted a three-editor system, where three "coeditors" handle submissions to both Regular Communications and Consultant's Forum. In addition, a Central Editorial Office was established to coordinate the editorial process among the three coeditors, who may reside in geographically diverse locations. The Office and the able journal Editorial Assistant, Ms Ann Hanhart, are located in Dallas, Texas. The position of "coordinating editor" was also created; this coeditor is responsible, in addition to usual editorial duties, for all administrative tasks (e.g. answering queries, compiling statistics on times to review, monitoring backlog) and serves as the point of contact for the Editorial Assistant regarding administrative matters. Coeditors are appointed to three-year terms according to a staggered entry scheme in which one completes his/her term and is replaced each year.

In accordance with recent technological advances, in 1999 the journal began accepting submissions electronically, and, by 2002, greater than 97% of all

Table 1 Editors of *Biometrics*. From 1999 on, coeditors under the three-editor system are listed by term

Editor	Shorter communications editor	Book review editor
G.M. Cox (1945–1955)	G.W. Snedecor (1945–1958)	
J.W. Hopkins (1956–1957)		
R.A. Bradley (1957–1962)	D.J. Finney (1959–1961)	J.G. Skellam (1960–1963)
M.R. Sampford (1962–1966)	J.A. Nelder (1962–1966)	W.T. Federer(1964–1972)
H.A. David (1966–1971)	P. Sprent (1967–1971)	
F.A. Graybill (1971–1975)	C.D. Kemp (1972–1975)	F.N. David (1972–1977)
F.B. Cady (1975–1979)	J.S. Williams (1976–1979)	R.M. Cormack (1978–1984)
P. Armitage (1979–1984)	J.J. Gart (1980–1984)	
D.L. Solomon (1984–1989)	R. Thompson (1984–1988)	C.D. & A.W. Kemp (1984–1999)
K. Hinkelmann (1989–1993)	N. Keiding (1989–1992)	
C.A. McGilchrist (1993–1997)	B.J.T. Morgan (1992–1996)	
R.J. Carroll (1997–2000)	L. Ryan (1996–1999)	Martin Ridout (2000–2002)
T. Pettitt (1999–2001)		
M. Davidian (2000–2002)		
D. Commenges (2000–2003)		
B. Cullis (2002–2004)		
X. Lin (2003–2005)		Iris Pigeot–Kuebler (2003–)
M. Kenward (2004–2006)		

submissions were electronic; virtually all are today. Moreover, starting with the January 2003 issue, *Biometrics* is available in electronic as well as print form to IBS members and authorized subscribers. In 2001, all issues of *Biometrics* from inception to a five-year lag from the current year were made available on the journal archival web resource JSTOR (`http://www.jstor.org`).

From modest beginnings, *Biometrics* has become a highly regarded international journal, whose papers are among the most referenced in biostatistical research. In 2001, the journal received 452 submissions, a number that continues to increase, and published 157 papers in all sections of the journal together with 66 book reviews and 17 brief book reports.

In 2003, there were approximately 90 associate editors of *Biometrics* drawn from all over the world, reflecting not only the truly international character of IBS but also the diverse range of areas of application and methodological development of published papers. The size and breadth of expertise represented on the editorial board frequently allows submissions to be handled by an associate editor whose interests are closely aligned with those of authors, which facilitates times to review that are among the swiftest in the statistical profession. In 2002, the median time to review for initial submissions was less than 10 weeks, with 98% of all first submissions receiving a review within 6 months.

The type of paper published in *Biometrics* has varied somewhat over the years while staying consistent with the aims of the journal. This variation has reflected the development of different areas of application. In early issues, biometry in agricultural research was the dominant theme. The journal continues to publish papers in this area related to both traditional design and analysis as well as newer topics such as the analysis of spatial data (*see* **Geographic Epidemiology**; **Spatial Models for Categorical Data**). The journal is also a forum for papers in ecology and wildlife statistics. In recent years, papers targeting applications in biomedical research have made up a large proportion of submissions and published papers. Environmental research applications (*see* **Environmental Epidemiology**) and research in genetics (*see* **Genetic Epidemiology**) are also reflected. Important, emerging new areas, such as molecular genetics (*see* **Molecular Epidemiology**; **Bioinformatics**), and medical imaging (*see* **Image Analysis and Tomography**), are frequently covered in the journal. Having methods developed in such a diversity of application areas published in the same journal promotes cross-fertilization of ideas. The journal properly reflects current interests of members of the IBS and at the same time, remains true to the emphasis on practical application as envisaged by its founders.

M. DAVIDIAN & C.A. MCGILCHRIST

Biometrika

Biometrika was founded by **Karl Pearson** and W.F.R. Weldon in consultation with **Francis Galton**. The first issue appeared in October 1901. From 1906 Pearson assumed entire editorial responsibility, until he was succeeded by his son, **E.S. Pearson**, in 1936. D.R. Cox took over in 1966 and acted until 1991. Since then D.V. Hinkley (1991–1992), A.P. Dawid (1992–1996), and D.M. Titterington (1996–) have edited the Journal. The Publication Editor, Ms. B.J. Sowan, has been involved with the Journal since 1971.

The Journal's origin was due in part to the Royal Society's request that, in papers submitted for publication, mathematics be kept apart from biological applications. The early volumes contained many diagrams, photographs, and tables of measurements of parts of the human body, animals, and plants. So far as the founding editors were concerned, they envisaged the Journal as an organ of "a spearhead of enthusiastic workers ... to lead a fight for the recognition of the place of mathematics in the biological field" [1], and this objective dominated the early decades of the Journal. There was a major change of emphasis when E.S. Pearson succeeded K. Pearson as Editor, reflecting the former's strong interest in and involvement with current theoretical and applied interests of the time. Since then the broad character of the Journal has gradually evolved in line with the increasing specialization of the field. Biological applications no longer receive special emphasis, although inevitably a substantial proportion of papers do involve directly or indirectly biomedical applications. Indeed, in his announcement of the changeover in editorship from E.S. Pearson to D.R. Cox, Tippett [2] noted that *Biometrika* "is now a general statistics journal which retains, however, a bias towards papers with a practical application and which are thus connected with the usefulness of statistics in some particular field, biometric or otherwise." With almost no change in this basic philosophy, Biometrika has evolved into a leading international journal for theory and methods across a wide spectrum of statistical topics, although certain specialisms are given emphasis from time to time, according to their current importance in statistical research. The Journal's centenary in 2001 was marked by the publication by a series of special articles that reflected on the contribution of Biometrika papers to the development of statistical science during the twentieth Century. These papers were reprinted in book form [3], together with a selection of particularly seminal articles from past volumes.

The number of papers published over the years is indicated in Table 1. There were in general two issues per year up to 1967, then three per year up to 1986 and four per year thereafter. The years of World Wars I and II, however, led to deviations from the general pattern.

Biometrika currently contains about 75–90 papers per year, in a volume of about 950 pages annually. The Journal is published by the Biometrika Trust, and the four parts of each volume appear in March, June, September, and December. The circulation is about 2300. Subscribers can now obtain online access and there is further rapidly increasing electronic usage of the journal, through institutional subscriptions.

Table 1 Variation in the size of *Biometrika* and the length of papers published, 1905–2003

Year	Pages	Issues	Papers		Main paper (ave. length)
			Main	Miscellanea	
1905	384	2	13	3	28
1920	132	1	7	0	19
1935	471	2	18	6	25
1945	85	1	8	4	9
1950	454	2	41	15	10
1965	675	2	45	30	13
1980	728	3	67	44	8
1995	892	4	55	23	14
2003	994	4	63	18	14

From 1997 the distribution, but not the publication, has been handled by Oxford University Press, who also look after the Journal's webpage, at `http://www.biomet.oupjournals.org`.

In general the Journal aims to publish new statistical theory and methodology that are capable of application in practical problems. As a comparison, the ***Journal of the Royal Statistical Society, Series B*** and the Theory and Methods Section of the ***Journal of the American Statistical Association*** contain material very similar to that in *Biometrika*. Papers that are essentially of mathematical interest only or that are applications of existing theory and methods are excluded; more explicit stress on biological applications is typically laid by and **Biometrics** and Biostatistics. A few review papers have been published, as well as a series on the history of probability and statistics (*see* **Biostatistics, History of**). There is a Miscellanea section for shorter articles. For some years now the submission rate has been at the level of close to 400 papers per year, so that the acceptance rate is about 20%. Nowadays electronic submission of papers, by email to the Editor, is the norm. There are about 15 Associate Editors, who advise the Editor over the bulk of the papers, although the Editor deals directly with some. All correspondence is channeled through the office of the Editor, and referees and Associate Editors act anonymously. The Publication Editor deals with the process of taking scientifically acceptable papers to their appearance on the printed page.

The Editor aims to minimize the time authors have to wait for feedback about their material, and publication takes 5–10 months from the date of final submission of an acceptable typescript.

References

[1] Pearson, E.S. (1936). Karl Pearson: an appreciation of some aspects of his life and work, Part I, *Biometrika* **28**, 193–257.

[2] Tippett, L.H.C. (1965). Editorial arrangements, *Biometrika* **52**, 1.

[3] Titterington, D.M. & Cox, D.R. (2003). *Biometrika: One Hundred Years*. Oxford University Press.

B.J. SOWAN & D.M. TITTERINGTON

Biostatistics, History of

From the seventeenth century to the present day, basic biologic phenomena (most notably mortality and morbidity) have been a central concern of those who collected and analyzed statistical data. For this reason, **John Graunt's** pioneering 1662 work *Observations upon the Bills of Mortality* sounds notes that are very similar in kind to biostatistical reports issued today – even though present-day statistical works make use of more sophisticated mathematical methods than their predecessors in earlier centuries. Consequently, the history of biostatistics can best be understood in terms of methodologic developments within statistical thinking. For analytical purposes, these methodologic developments can be divided into four phases: (i) the work of nineteenth-century statisticians who pioneered the concept that social patterns (including incidence of disease) could be shown to have "lawlike" characteristics [13, 15]; (ii) the mathematical work of **Karl Pearson** and his biometric associates at University College London in the early twentieth century; (iii) the interwar years when the methods of **hypothesis testing** (associated initially with the agricultural research of **R.A. Fisher**) were extended into the field of biomedicine; and (iv) the postwar rise of epidemiologic studies focusing on such celebrated discoveries as the association between cigarette smoking and lung cancer (*see* **Smoking and Health**) and the **Framingham study** of heart disease. Each of these four phases will be discussed in turn.

Nineteenth-Century Developments

As its name implies, statistics developed as a science concerned with information important to the state. With the rise of industrialization and democratic reforms in the early nineteenth century, Western governments became overwhelmed with what Hacking [7] has called an "avalanche of printed numbers" about their citizens, thereby leading writers to associate the term "statistics" specifically with information expressed in numeric form. One of the earliest examples from a biomedic context was **Bisset Hawkins'** [8] 1829 work *Elements of Medical Statistics*, which was concerned with "the application of numbers to illustrate the natural history of man in health and disease".

Throughout the course of the nineteenth century this numeric conception of statistics became institutionalized through the founding of statistical societies and the holding of international statistical conferences. In the English-speaking world, two of the most prominent societies were Section F of the British Association for the Advancement of Science (founded in 1833) and the Statistical Society of London (founded in 1834 (*see* **Royal Statistical Society**)). In both of these societies (and in many of the international conferences) one of the leading figures was the Belgian astronomer turned social statistician **Adolphe Quetelet** who helped to pioneer the concept that society had distinctly lawlike characteristics which could be revealed through the amassing of statistical evidence.

Most of the members of these statistical societies were not trained in the physical sciences like Quetelet; rather, they often came from the medical profession. As Lécuyer [10] has argued, several factors may account for the high proportion of physicians within these societies, namely the emergence of public health as a goal through the method of improved hygiene, the medical tradition of local investigation to improve the conditions of the poor, and the physicians' obvious professional association with the phenomena of death and sickness. All of these concerns lent themselves naturally to the amassing of numeric evidence.

In this period, two of the most prominent physicians to study public health problems statistically were the French physician, Louis René Villermé (1782–1863), and the English physician, **William Farr** (1807–1883). As William Coleman [4] has shown, Villermé corresponded with Quetelet as a result of their shared interest in describing society in numeric terms. In 1828, Villermé published a memoir positing a relationship between mortality and economic status, thereby firmly establishing his reputation as an advocate for a statistically based approach to public health problems. Similar concerns informed the work of William Farr who studied medicine in Paris under **P.-C.-A. Louis**, the advocate of the "numerical method". In 1839, Farr was appointed compiler of abstracts to the newly established General Register Office which had been founded to record all births and deaths within Great Britain. During Farr's long 41-year tenure at that institution, he made considerable contributions to the field of **vital statistics**

and developed multiple disease and occupational taxonomies to be used in the collecting of statistical evidence (*see* **Health Statistics, History of**).

The statistical record-keeping of individuals like Farr and his associates at the General Register Office proved to be indispensable for one of the major epidemiologic discoveries of the mid-nineteenth century, namely John Snow's [14] demonstration that cholera was a water-borne disease. After an outbreak of cholera near the Broad Street pump in London in 1854, Snow used data collected by the General Register Office to determine that there had been 83 deaths attributed to the disease during a three-day period. On closer examination Snow determined that in all but 10 of these cases the individuals had lived in households that were closer to the Broad Street pump than any other water source. Furthermore, in five of the households that could get water elsewhere, Snow interviewed the family members and discovered that they did indeed use the water from the Broad Street pump. Of the remaining five cases, three were children who probably also drank water emanating from this source since they attended a nearby school; the remaining two cases were dismissed by Snow as representing "only the amount of mortality from cholera that was occurring before the irruption took place". Although Snow's idea regarding the water-borne nature of cholera received little support during his lifetime (with the notable exception of William Farr), his paper has subsequently attained classic status as one of the most important epidemiologic discoveries prior to the discovery of the germ theory of disease.

The Biometrical School and the Mathematization of Statistics

Throughout the nineteenth century, statisticians conceived of their work as largely descriptive in nature with comparatively little emphasis placed on mathematical reasoning. This orientation was fundamentally changed by the creation of the biometrical school at University College London under the direction of the applied mathematician Karl Pearson (1857–1936). Pearson developed this school with the blessing (and financial backing) of the scientist **Francis Galton**, an English scientist who espoused the view that heredity played a decisive role in individual development. Thus, the primary *raison d'être* of Pearson's research was to provide scientific warrant to Galton's views through statistical analysis. Although Pearson clearly shared the views of his main financial backer, he also developed a full-blown philosophy of statistical reasoning arguing that, since all **inference** is based on the association of antecedents and consequents, all scientific reasoning is at its core fundamentally statistical. As a result, Pearson argued for the extension of statistical methods into potentially all domains of scientific endeavor, actively engaged in debates with other researchers over the proper interpretation of statistical data, and trained students in the biometric techniques that he pioneered.

In the field of biostatistics specifically, Pearson is remembered for engaging in a dispute with Almoth Wright (1861–1947) over the meaning of the statistics Wright had collected to demonstrate that antityphoid inoculation reduced the chance of infection for soldiers in the British Army. In critiquing Wright's conclusions, Pearson made use of one of the statistical constructs for which he is remembered today, namely the **correlation** coefficient, which was designed to measure the degree of association between two phenomena. Specifically, Pearson [12] found that the average correlation between immunity and inoculation was about 0.23, with individual results as high as 0.445 and as low as 0.021 (a one-to-one positive association would have generated a value of 1 and no relationship would have generated a value of 0). Since this was a very low correlation coefficient relative to other common therapeutic interventions (the protective character of vaccination at preventing mortality from smallpox was found to have a correlation coefficient of approximately 0.6), Pearson argued against the introduction of antityphoid inoculation as a standard practice. Although Pearson's criticisms did not convince the leaders of the British Army, who continued to test and use Wright's inoculation procedure, the episode is illustrative of Pearson's desire to show the applicability of his biometric techniques in the biomedical arena.

In addition to his theoretical innovations, Pearson was also important for training the physician **Major Greenwood** (1880–1949) in biometric methods; Greenwood launched his career by criticizing Wright's use of the so-called "opsonic index" for diagnostic purposes. Wright believed that there was a substance in blood serum (opsonin) which prepared bacteria to be ingested by the white blood corpuscles. Wright was able to measure the amount of

opsonin that was present in the blood by comparing the average number of microbes per leucocyte in a blood sample from a normal individual with the average number of microbes per leucocyte from an individual suspected of having a bacterial infection; subsequently, Wright computed the ratio of these two mean values which he called the "opsonic index". In general, Wright believed that if the opsonic index were higher than 1.2 or lower than 0.8, then this would indicate bacterial infection. In his critique of Wright, Greenwood [6] plotted the **frequency distribution** of the number of microbes per leucocyte and found the distribution to be markedly asymmetric or skew. Thus, Greenwood advocated that the **mode**, or most frequently occurring value, would be a better constant with which to measure the opsonic index than the mean. Although Greenwood did not convince Wright, his work did succeed in impressing Charles James Martin who was then director of the Lister Institute of Preventive Medicine; late in 1909 Martin offered Greenwood a position as medical statistician at the Lister Institute thereby helping to legitimate the use of biometric techniques in the analysis of medical statistical results.

Greenwood was not the only individual to draw on Pearsonian methods to study biomedical phenomena; another prominent follower was **John Brownlee**. Brownlee [3] utilized Pearson's insight that the Gauss–Laplace or **normal distribution** curve was, in fact, just a particular case of an entire family of frequency distribution systems. By attempting to fit Pearson's various frequency distributions to medical statistics of disease incidence during an epidemic, Brownlee hoped to classify epidemics according to the type of frequency distributions they approximated; he often found that the type of distribution produced was nearly symmetric (*see* **Epidemic Curve**).

The Interwar Years and the Birth of Experimental Epidemiology

On both sides of the Atlantic the interwar years saw an attempt to forge a new and experimental approach to epidemiology. Rather than continuing to rely solely on vital statistics of human populations, systematic attempts were made to study the rise and fall of epidemic diseases within populations of laboratory animals – most notably mice. In the UK, this work was based at the **Medical Research Council** and consisted of a collaborative endeavor between the bacteriologist W.W.C. Topley and Major Greenwood; Greenwood had become head of the statistical research unit of the Medical Research Council in 1927. In the US, the principal investigator was L.T. Webster and his associates at the Rockefeller Institute. In both countries, researchers focused on mouse typhoid in an attempt to understand the relative importance of environment, host, and agent factors in disease occurrence.

In addition to these experiments on populations of laboratory animals, the interwar years also saw important theoretic developments in the methods of statistical **inference** as pioneered initially by Ronald A. Fisher. Whereas Pearson and the early biometricians had been associated primarily with classifying **observational** data, Fisher was more directly concerned with **experimental** data and **hypothesis testing**. Fisher developed his statistical ideas after being appointed to the Rothamsted Experimental Station where he studied the differing productivity of various types of grain in agricultural field experiments. Drawing on these scientific findings, Fisher published a series of books on statistical methodology. In his 1935 work *The Design of Experiments* [5], he outlined the key importance of **randomization** in assigning different grain types to various tracts of land in order to remove subjective experimenter **bias**.

Fisher's focus on randomization proved to have profound implications for research medicine – especially for the development of the modern **clinical trial** (*see* **Clinical Trials, History of**). When faced with a shortage of the drug streptomycin during World War II, **Austin Bradford Hill** (who had succeeded Greenwood as head of the statistical division of the Medical Research Council) chose to design a rigorous clinical trial of the effect of this drug on bilateral pulmonary tuberculosis. Bradford Hill used random numbers to determine which patients received the experimental drug and which patients were to be controls. Although the trial contained a relatively small number of patients (107 overall with 55 allocated to the streptomycin group and the remaining 52 allocated to the control group), Hill's attention to methodologic detail meant that the results were decisive: 7% of the streptomycin patients died and 27% in the control group died. As Hill and his associates [11] observed in the 1948 report on the trial, "The difference between the two series is

statistically significant; the probability of it occurring by chance is less than one in a hundred" (*see* **Medical Research Council Streptomycin Trial**). Hill's streptomycin trial has often been seen as the standard against which all subsequent clinical trials have been judged.

The Rise of Postwar Epidemiology

Even though the study of epidemiology (or disease within populations) has a long history, the postwar era is significant in at least two respects: the use of more sophisticated statistical techniques in analyzing the etiology of disease; and the increasing shift in focus from infectious to chronic conditions. Although these twin facets of postwar epidemiology are distinguishable for analytical purposes, they actually were historically interwoven. With the transition from infectious to chronic disease as the principal reason for mortality and morbidity, research became centered less on the search for a specific agent (or germ) and more on the analysis of (multiple) environmental factors. Since multiple factors presented the problem of **confounding** causal relationships (*see* **Causation**), epidemiologists increasingly turned to the statistically trained who had dealt with similar issues in the context of social surveys. Of the many epidemiologic claims established through these methods, the two most famous examples were the researches establishing a link between cigarette smoking and lung cancer and the study of cardiovascular disease.

As discussed in Brandt [2], epidemiologic studies began to appear in the late 1940s and early 1950s indicating that cigarette smokers were at a higher risk of lung cancer than nonsmokers. Most of these studies were **retrospective** in nature – meaning that individuals who had already developed lung cancer were interviewed about their smoking habits after the fact; their responses were then compared with a control group of individuals who did not smoke (*see* **Case–Control Study**). However, two pioneering prospective studies were also launched at this time (*see* **Cohort Study**). In 1951, Richard Doll and Bradford Hill sent questionnaires to all British physicians inquiring about their smoking habits. When individuals who responded to their survey died, Doll and Hill obtained data about their cause of death. At about the same time, a similar study was being conducted in the US by E. Cuyler Hammond with the support of the American Cancer Society. Both studies implied conclusions consistent with retrospective studies: cigarette smoking increased one's risk of contracting cancer.

In the US, the most famous postwar epidemiologic investigation has been the Framingham Study, which was initiated in October 1947. As Susser [16] has argued, the Framingham Study has often been cited as the paradigmatic example of a prospective or "cohort study" which follows a specific group (or cohort) of individuals over their life courses to see what factors influence disease development. As its name implies, the researchers selected their study population from the residents of the town of Framingham, Massachusetts. By examining a sample of 30- to 59-year-old persons biennially, the researchers were able to test the role of such factors as cholesterol level, physical activity, diet, and life stress on the development of heart disease. In addition to specific empirical findings, the Framingham Study also generated important methodologic insights of how to deal with the variability of repeated measurements over time (*see* **Longitudinal Data Analysis, Overview**); however, the researchers could not solve the problem of when to terminate the study. As a result, more recent findings have centered on diseases related to aging (e.g. stroke) and follow-up studies of the offspring of the original participants.

Since the 1960s, several factors gave increasing prominence to statistically based ways of studying biomedic phenomena. After the Thalidomide scandal raised the specter of infant deformity, the clinical trial became a standard requirement before experimental drugs could be administered to the general public. In 1965, the *American Journal of Hygiene* changed its name to the *American Journal of Epidemiology* [1] reflecting the "greatly increased importance" of the "epidemiologic approach to disease". Finally, the discipline of epidemiology was put on a secure conceptual foundation by researchers on both sides of the Atlantic who articulated causality criteria for epidemiologic studies; these criteria were designed to serve as the chronic disease analog to Robert Koch's famous postulates for establishing causality for infectious disease. In the US, the most famous list of epidemiologic causality criteria was published as part of the Surgeon-General's 1964 [17] report positing a link between cigarette smoking and lung cancer; in the UK, the most famous list of causality

criteria was published by Austin Bradford Hill [9] early in 1965 (*see* **Hill's Criteria for Causality**).

References

[1] "Change in Name" (1965). *American Journal of Epidemiology* **81**, 1.

[2] Brandt, A.M. (1990). The cigarette, risk, and American culture, *Daedalus: Journal of the American Academy of Arts and Sciences* **119**, 155–176.

[3] Brownlee, J. (1918). Certain aspects of the theory of epidemiology in special relation to plague, *Proceedings of the Royal Society of Medicine (Section Epidemiology and State Medicine)* **11**, 85–132.

[4] Coleman, W. (1982). *Death is a Social Disease: Public Health and Political Economy in Early Industrial France*. University of Wisconsin Press, Madison.

[5] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[6] Greenwood, M. (1909). A statistical view of the opsonic index, *Proceedings of the Royal Society of Medicine* **2**, 145–155.

[7] Hacking, I. (1987). Was there a probabilistic revolution, 1800–1930?, in *The Probabilistic Revolution: Ideas in History*, Vol. 1, L. Krüger, L.J. Daston & M. Heidelberger, eds. MIT Press, Cambridge, Mass., pp. 45–55.

[8] Hawkins, B. (1829). *Elements of Medical Statistics*. Longman, Rees, Orme, Brown & Green, London.

[9] Hill, A.B. (1965). The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[10] Lécuyer, B.-P. (1987). Probability in vital and social statistics: Quetelet, Farr, and the Bertillons, in *The Probabilistic Revolution: Ideas in History*, Vol. 1, L. Krüger, L.J. Daston & M. Heidelberger, eds. MIT Press, Cambridge, Mass., pp. 317–335.

[11] Medical Research Council Streptomycin in Tuberculosis Trials Committee (1948). Streptomycin treatment of pulmonary tuberculosis, *British Medical Journal* **2**, 769–782.

[12] Pearson, K. (1904). Report on certain enteric fever inoculation statistics, *British Medical Journal* **2**, 1243–1246.

[13] Porter, T.M. (1986). *The Rise of Statistical Thinking, 1820–1900*. Princeton University Press, Princeton.

[14] Snow, J. (1855). *On the Mode of Communication of Cholera*, 2nd Ed. London.

[15] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge, Mass.

[16] Susser, M. (1985). Epidemiology in the United States after World War II: the evolution of technique, *Epidemiologic Reviews* **7**, 147–177.

[17] US Department of Health, Education, and Welfare (1964). *Surgeon General's Report, Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*, PHS Publication No. 1103. Government Printing Office, Washington.

J. ROSSER MATTHEWS

Biostatistics, Overview

This article presents a personal view of the general field of biostatistics. I describe the overall sense of the field, a few of its roots and principal originators over the past century and more, some of the methods used in biostatistics and types of problems that yield to a biostatistical approach (using as illustrations areas where biostatistics has had noteworthy impact). I also discuss the discipline in general as it exists today and mention some of the more important problem areas in biomedicine where biostatistics is now challenged to provide new methodology. Finally, I describe the profession of biostatistics in terms of its history, the present composition of professionals, its societies and journals, its entry points for new professionals and the current areas of activity of professionals in the field.

Biostatistical Roots, Development, and Examples of the Discipline

It is difficult to trace the roots of a field that has come, only recently, in the early part of this century, to be recognized as a distinct discipline. This difficulty is especially marked if the field is of hybrid character. Biostatistics focuses on the development and use of statistical methods to solve problems and to answer questions that arise in human biology and medicine. Thus it expands statistical theory and adapts it to bring specific methods to bear on questions of importance to the community of scientists, practitioners, and policy makers who have interest in health and in all health aspects of the human community. A few examples best illustrate the roots and development of the discipline.

There has always been interest in the general length of human life and its numerical description. Consideration of human longevity (*see* **Life Expectancy**) quickly prompts careful distinctions between, for example, recorded lengths of usual life (barring "early" death) and expected or average length of life, i.e. age at death, among all persons born in a given epoch and circumstance. Sometimes the interest can become sufficiently intense to prompt careful numerical thought, e.g. when insurance premiums are exacted on condition of further life length, or a plague has afflicted a nation. The biblical three score and ten and rough estimates of variation may not be sufficiently precise or specific to the need. Roman documents contain tables of the expected length of life for persons of various ages, for actuarial use in fixing prices of insurance (*see* **Actuarial Methods**). A millennium later, various Western European countries began to collect information systematically on births and deaths (*see* **Vital Statistics, Overview**). Curiosity prompted some individuals to examine this information and to note interesting characteristics of the collective information and striking regularities in what seemed on the surface as simply random events. **John Graunt's** [11] analysis of the Bills of Mortality, lists of dates of burials, births, and marriages, is an oft-cited landmark in the beginnings of such work, along with the work of **Quetelet**, Huygens, **Halley** and others. Their work led to the development of calculations of birth rates and death rates for various ages and to **life tables** that yielded calculation of the expected lengths of life for populations in well-defined geographic areas and subgroups of these populations, based on increasingly reliable and detailed counts of births and deaths. This early work produced useful information for use in detecting sources of unusual and preventable causes of death, due, for example, to contamination of water supply and sanitation systems.

These statistical methods for the collection and analysis of birth and death data led to many other uses of such data in government and in social sciences, for example in forecasting population size (*see* **Population Growth Models**) and economic conditions of populations and population subgroups. The mathematical and statistical methods of data collection, description, and **forecasting**, with the attendant concern for statistical precision, remain in many fields of application a major area for research by biostatisticians in government, universities, insurance firms, pharmaceutical firms, medical centers, and independent research centers.

Another root of biostatistics as a discipline stemmed from interest in the obvious resemblance of offspring to their parents. The work of Mendel with his pea plants (*see* **Mendel's Laws**), in the mid nineteenth century, is a critical and familiar landmark for all of us, in the beginnings of the use of systematically collected data, scientific speculation, and the application of numerical methods to the description of regularities in heritability amidst striking variation. Mendel's work focused on plants, but rediscovery of his work at the turn of the century

quickly drew the attention of statisticians as well as biologists and scientists to gain an understanding of the variation in human heritability (*see* **Human Genetics, Overview**). **Francis Galton** and **Karl Pearson** were two of the major figures in this area of science at the turn of the century. They were foremost contributors to the development of new concepts and statistical methods for describing and drawing inferences from data on the resemblance of parents to offspring. They were interested in resemblances in physical, psychological, and behavioral features of offspring to their parents. They developed concepts and statistical theory still basic to understanding this form of data. Central among these contributions were the concepts of **correlations** among variables and **regression** methods for predicting the characteristics of offspring from those of their parents. These concepts continue to be basic in the study of such relationships among variables in general. The concepts and methods continue to find new application in old areas of scientific research and in the newest areas of science, and are central in much of the work of biostatisticians today.

A still more recent, but most important area of biostatistical activity stems again first from agriculture, where **R.A. Fisher**, at Rothamsted Agricultural Station, in the 1920s, worked with agricultural scientists on their experimental studies. Even the most fastidiously controlled experiments involving animals (or plants) manifest responses that vary widely from one experimental animal to the next. Fisher proposed the method of randomizing the animals to the several treatments under study in the experiment, using specified probabilities of assignments (*see* **Randomization**). Of course, this would tend to balance treatment groups in terms of the characteristics, known or unknown, that cannot be controlled precisely. At the same time, through the probabilities used in the random assignments, a firm experimental basis in probability theory is provided for measuring the reliability of **inferences** drawn in comparing the effects of the treatments under study. The concept became firmly imbedded in experimental agriculture. A decade and more later, the method was introduced into experimental science in biomedicine, where the same problem of experimental control is encountered in laboratories and in experiments with patients. Where the experimental unit is the test tube, unwanted variation from tube to tube may be small and tolerable, though this is not always the case. Variation from animal to animal may not. Variation from one patient to the next, treated with the same experimental maneuver, even with the greatest care in patient selection and treatment, is rarely negligible (*see* **Experimental Design**).

Randomization of treatments to patients as a means of securing a controlled comparison of treatments, with statistical evaluation of the results came gradually into use (*see* **Clinical Trials, Overview**), beginning in the 1930s, following a few years after Fisher's introduction of the idea in the 1920s for use in studies in agricultural science. **Bradford Hill** [12] was among the foremost leaders in advocating the method for general use in clinical science. Randomization is now a hallmark of reliable experimental method in evaluating new and experimental medical treatments and in other medical maneuvers for prevention, for therapy, and for study of medical policies in general. Biostatisticians continue to work on the development and adaptation of these ideas for special applications in medicine, working to develop new experimental designs for the collection and use of clinical trial data for valid inference, and at the same time deriving safeguards that assure validity of the inferences and protection of the safety of the patients in these experiments (*see* **Ethics of Randomized Trials**). This is an area of intense development for theory, and for use in current clinical research and in much of medical research ranging from laboratory experiments to community interventions of public health maintenance. Meinert [17] and Pocock [21] provide some interesting notes on early clinical trials and broad discussion of statistical aspects of the organization, design, implementation, and analysis of the results of randomized clinical trials. Many books have been published over the last several decades on these topics.

Another area where biostatistics has been key to the advancement of a scientific area of inquiry, bringing statistical concepts, principles, and methods to bear on a specific question, is the indirect measurement of the strengths of compounds by administration of the compound to living organisms. Vitamins and hormones provide specific examples of the general problem. If a vitamin can be produced, for example, as an extract of some natural product, but the chemical composition of the ingredient(s) is unknown, then the resulting efficacy might be ascertainable by testing in animals. For example, many early vitamins were produced in this way and the strength of the

product was measured by the rate of growth among animals fed the product. It then remains to develop a systematic experimental method to measure the strength of a given "batch" of the material, so that uniform batches can be produced for research and clinical use. Otherwise, doses cannot be defined and efficacy and safety cannot be controlled. Variation from animal to animal, even in the closest of controlled experiments, makes the problem of measuring strength in this way problematic. Here biostatisticians [2, 7, 8] were able to bring statistical concepts and methods to the problem, developing definitions of strength and designing experiments and methods of statistical analysis to estimate the strength of a compound with a measure of the reliability of the estimate. The methods allowed calculation of the sizes of studies necessary for reliable use in the manufacture of various vitamins, hormones, and other pharmaceuticals that would allow comparisons of strength of compounds across manufacturers and time. These methods of bioassay (*see* **Biological Assay, Overview**) are essential, and in current use whenever the strength of the target compound cannot easily and specifically be characterized and measured by standard chemical or physical instrumentation. The methods have been adopted in many other areas where analogous questions have arisen, e.g. in measuring the strengths of toxic substances in the environment (*see* **Risk Assessment for Environmental Chemicals**), and in psychophysics (where, indeed, some of the basic concepts and methods had parallel early development). The methods have been adapted for use in psychological and educational testing theory (*see* **Psychometrics, Overview**) [14], where it is important to have tests of measurable performance and precision across time and across varied populations of subjects to be tested.

Nature of the Discipline

As is clear from the above examples, biostatistics is problem oriented. It is specifically directed to questions that arise in biomedical science. The methods of biostatistics are the methods of statistics (*see* **Statistics, Overview**) – concepts directed at variation in observations and methods for extracting information from observations in the face of variation from various sources, but notably from variation in the responses of living organisms and particularly human beings under study. Biostatistical activity spans a broad range of scientific inquiry, from the basic structure and functions of human beings, through the interactions of human beings with their environment, including problems of environmental toxicities and sanitation, health enhancement and education, disease prevention (*see* **Preventive Medicine**) and therapy, the organization of health care systems (*see* **Health Services Organization in the US**) and **health care financing**.

The details, depth, and breadth of modern biomedical and social science and the span of knowledge of mathematical, statistical, and calculational theory compel team approaches to the solution of modern scientific problems. Biostatisticians are members of many of these teams. The role of the biostatistician requires a special combination of tastes: a taste for quantitative methods, an understanding of and tolerance for variation in the data of scientific investigation, enjoyment in communication and in collaborative work with applied scientists, and a knowledge of statistical theory and methods.

The concepts and principles of mathematical statistics are the methods for describing regularity in the presence of variation, methods for prediction in the face of uncertainty, and methods for efficient study and experimentation when the results for the individual case or observation are uncertain. The biostatistician brings the characteristics of scientific curiosity to the specific question, a collegial bent, pleasure in dealing with scientific problems and an ability to communicate with the applied scientist, along with the skills and background in statistics necessary to invent, develop, or recognize special statistical methods that can aid in the solution and a knack for implementation of the new method to the problem.

As an illustration of this role of biostatistics in biomedical science I have cited above four areas where biostatistics has played a major, even a critical developmental role: life tables as a tool for adapting probability theory and statistical inference to the generation and use of birth and death data, as a method for measuring and comparing the health status (*see* **Health Status Instruments, Measurement Properties of**) of populations and for identifying health factors affecting populations and subgroups of populations (*see* **Analytic Epidemiology**); **correlation** and **regression** as measures of the strength of heritability, tools that have now been extended

throughout science to the study of cause and effect (*see* **Causation**); the clinical trial as a method for valid assessment of the efficacy of medical treatments of patients, again a tool that now plays a pervasive role in all of clinical science and beyond; and bioassay as an indirect method for precise measurement of the strength of biologically active compounds, such as hormones and vitamins.

There are many more areas where biostatistics has contributed greatly to the development of an area of scientific investigation, to the point where the special methods developed have become a large and essential part of the scientific armamentarium of the investigators in a specific scientific area and in the training of new investigators in that field. Genetics (*see* **Genetic Epidemiology**) is a special discipline in itself. **Demography**, actuarial science, and experimental methods in clinical trials, as well as bioassay are others.

One area of special importance is the area of design of randomized clinical trials. Much of the current interest in the comparison of medical therapies involves large numbers of patients, randomized to two or more treatments. These treatment groups or cohorts are followed for extended lengths of time to compare survival rates (or some other outcome such as disease recurrence) (*see* **Survival Analysis, Overview**). At the end of a prescribed length of study time the data present a length of time for each patient, the time to the event of interest, say death, or successful survival to the end of the study. The study data then present the statistical problem of summarization, comparison of the survival experience for each group, and inference regarding the meaning of the results. Such data appear much like those gathered centuries ago in John Graunt's Bills of Mortality [11] and first formed into life tables by Halley and others, as described previously. These early methods have been adapted and elaborated in detail for the design and analysis of clinical trials within the last half century. This body of statistical theory and practical methodology forms the basis for much of the statistical planning and analysis in both clinical therapeutics and prevention (*see* **Prevention Trials**) today.

Another area standing as a discipline in itself, but where biostatistics continues to play a central role, is epidemiology, the study of epidemics and the causes of infectious diseases (*see* **Communicable Diseases**), and more recently, the study of the causes and control of chronic diseases, such as cancers and heart disease [13]. The field of epidemiology has grown rapidly and investigators now focus and specialize on such diverse areas as injuries and deaths as a result of accidents – due to automobiles and traffic in general, sports, accidents in the home, and in special occupational activities. Environmental risk factors and their consequences is another area of study of great importance and interest to epidemiologists. We are all familiar with the public health importance of risk factors associated with life styles and behavior modes, e.g. in diet, exercise, sexual, and drug behaviors. These sources of health risk, and more, attract the epidemiologist. They present challenges to develop new methodologies that will allow efficient and valid inferences in identifying risks, estimating the magnitudes of the risks to the population, identifying subgroups of the population who are at extreme risk, and providing useful information to avoid such risks. The challenges are especially sharp and often subtle because randomized experiments are rarely possible. Instead, innovations in study design in the gathering of observations that will assure unbiased inference within reasonable limits of cost and time become all-important [13, 24] (*see* **Observational Study**).

Psychometrics [26] is yet another area, stemming, in part, from several strong biostatistical roots, and reaching back more than a century. One root goes back to the German psychophysicists who puzzled about the measurement of the relationship between physical stimuli, such as light rays and sound waves, and the reactions reported by human subjects. These scientists developed methods for describing the strength of the stimulus and the response of the subject. Their experimental approaches and the statistical methods they developed to describe their results yielded some of the concepts and techniques for bioassay as described earlier. Another root goes back to the work of Galton in the description of "correlations" (a word coined by Galton and Pearson in reference to this work) between human behavior of parent and child. These roots have led to much fundamental work in the broad field of psychology and to the development of the subfield called psychometrics, which concentrates on the use of statistical methods in research in human behavior, and the use of statistical techniques in areas such as educational testing.

Today there is much interest in the applications of psychology to the behavior of populations

and techniques for changing behavior toward more healthful life styles, e.g. with regard to drug and eating habits, exercise, sexual practices, systematic resort to **screening** for early detection of diseases such as breast cancer and prostate cancer, vaccinations (*see* **Vaccine Studies**) for disease prevention, and early maternal care in pregnancy. The techniques used in research in this area of behavioral medicine derive from many areas of biostatistics, particularly sample survey methods and clinical trial methods, as well as related fields that are heavily biostatistical, such as demography, vital statistics, epidemiology, and psychometrics.

Another area of renewed scientific research and application that draws heavily on biostatisticians and the use and development of new biostatistical methods is environmental toxicology. The statistical methods of epidemiology originated in questions of epidemics of contagious diseases spread to human from human, and from insects, rodents, and other sources in the environment. Epidemiologic questions today involve more subtle mechanisms of contamination and dysfunction, the detection and effects of noxious chemicals in the air and the soil, and the causes and effects of chronic disease in populations. The methods again are a blend of tried and of new statistical tools of epidemiology, but also new statistical tools adapted and developed from diverse areas of theory and scientific application for answering specific scientific questions in this area of investigation; for example, the concomitant variations of air pollutants and disease rates as air pollutants and disease rates vary through time on daily, weekly, and seasonal bases.

Yet another old but very new area of research that has developed with help from biostatistics is **health services research**. The questions concern just how medicine handles the health needs of the population it serves, what impact the medical system has, how to measure the impact, what factors are important in the effectiveness of a system for financing and caring for a population, and how the system might be altered for greater cost effectiveness. An excellent introduction to this field is the volume prepared under the direction of Kerr L. White, *Health Services Research: An Anthology* [27]. The work on the National Halothane Study [3] resulted in statistical methods for measuring and comparing the outcome of health care (for example, surgical mortality rates) among health care institutions (e.g. hospitals). Flood & Scott [10] described a study that exemplifies the methods by measuring factors in the organization of a hospital that affect outcomes of surgery and medical care. Health Services Research is central now in studying and informing health care finance and delivery. It draws on statistical methods in fields as diverse as medical economics, design and analysis of clinical trials, medical informatics and the design of medical data banks [25], the use of data banks (*see* **Data Archives**) for measurement and comparisons of **quality of care** among medical care systems, measurements of the **quality of life**, and the measurement of the health status of populations.

Current Focuses and Challenges in Biostatistics

The discipline of biostatistics is broad, its borders vague. This is necessarily so for any discipline that links a more theoretical or basic discipline to a spectrum of applied sciences. This is clearly true for other derivative disciplines, such as biochemistry and biophysics. The greatest contributors to the development of biostatistics have been, almost by definition, those leaders with a strength and interest in statistical theory but also with a clear vision of the methods and needs of scientific investigators in the pursuit of science, either generally or in a specific area of study. These central figures in the development of biostatistics mastered the art of compromise, developed methods that were and are the right blend of general theory but scientific specificity to problems common within and across areas of medical investigation and application. They developed methods that have the appeal of scientific and practical utility, bringing simplicity to the design and description of scientific experimentation and observation that are easily communicated and amenable to replication and testing.

For the reader interested in the flavor of current biostatistics, the first chapter of the book by Cox & Hinkley [5] gives a very nice description of how the statistician who has an eye on scientific applications thinks about the work. The book by Miller [18] is a personal view of the way an applied statistician goes about the work of bridging the gap between theory and practice.

What are the current research problems faced by the biostatistician? What are the needs of the biomedical investigator and applied scientist and practitioner? Where are the current challenges? The most important and lasting contributions of the future will come from biostatisticians who combine statistical power with scientific insight and curiosity about the important and general problems of medical research. Predicting new and fundamental breakthroughs in any field is problematic indeed. Who could have predicted the seminal work of K. Pearson, Galton, and Fisher?

Seminal contributions spring from genius, but there is much work yet to be done also at a lower and less innovative level, in further testing, adapting, extending, and exploring the new methods already proposed over the last few decades.

A further source of new biostatistical progress lies in new technologies in other fields that offer increments in explanatory power and insight when incorporated into biostatistical methodology. The most important and obvious example is the tremendous growth in computing power (*see* **Computer-intensive Methods**) and in medical informatics [25].

In addition, there is the important business of reviving methods of the past, ignored in their time or left relatively dormant, but attracting attention as possibly holding answers to new questions and new needs in scientific investigation. Here, for example, Bayes procedures (*see* **Bayesian Methods**) and **decision theory** assume a new luster in the eyes of the applied medical scientist and biostatistician. There is a growing interest and, indeed, a modestly growing use of Bayes' method and decision theory for quantitatively formulating knowledge, belief and cost estimates available at the planning stage of a clinical trial into detailed plans for the trial and the analysis of the results of the trial. The aim is an efficient wedding of what is known at the start of the trial with the results gathered, the whole then to be summarized and the detailed implications reported. Issues in the development of these methods for modern scientific experimentation, communication, and needs, in clinical trials and other areas, are interesting and will certainly continue [1]. It is clear, however, that if these methods are to be incorporated into routine and popular practice there will be a demand for more complex and flexible computer **software** and computational methods.

Focusing on the explosion in computer power over the past two decades as an example for exploitation, one can trace the roots of resampling methods back to the 1930s and earlier, with Fisher's description of the lady tasting tea [9] and the permutation tests of Pitman [20] (*see* **Exact Inference for Categorical Data**), the Quenouille [22] **jackknife method** for correcting bias through systematic resampling of the given data set, and the later **bootstrap methods** [6].

These ideas for testing, for estimation and for measuring the statistical precision of an estimate through resampling the data under analysis, were limited half a century ago by the computational work required in setting them to practical use, even though many of the examples were drawn from medical research settings. With the surge in computer power in the last two decades, measurement of the statistical precision of an estimate based on a sample, by repeated sampling of the sample itself, is indeed feasible for complex data sets, and in many cases these methods finesse the need for complex mathematical theory and approximations. It remains to further the theoretical work needed to draw the guidelines for the use of these methods, to set them on firm theoretical footings, to explore and define the areas of medical science where the methods can be used, and to incorporate the methods in software packages for routine use by biostatistical practitioners and medical scientists. With regard to the latter, some suggestions and currently available help appear in the appendix to Efron & Tibshirani [6].

The controlled randomized clinical trial was mentioned above as a central contribution of statistics to modern experimentation. It was gradually incorporated into the clinical sciences in medicine, and firmly adopted in clinical research and in regulatory affairs, setting the standards for approval of medical drugs (*see* **Drug Approval and Regulation**), devices and other medical accoutrements, for marketing and for medical practice. Currently, a drug proposed for marketing and practice may be subjected to study in thousands of patients, comparing groups treated with the experimental drug with groups treated with a standard drug, to reveal, if true, superiority in effect of the new drug over the standard drug (e.g. in preventing death following a heart attack). This approach to medical experimentation has seen much statistical work in developing designs of these experiments, particular to the medical clinic and to the kinds of problems in execution, analysis and decision making

that arise. The costs of the experiments can be huge; the future costs of erroneous decision even greater. The problems, ethical and otherwise, of working with human study subjects and preserving the validity of the study and the statistical inferences based on the results have presented, and continue to present, challenges to the biostatistician. Deep and useful further work is needed.

The preceding discussion of the clinical trial leads to another area of important focus for the biostatistician. Clinical trials are very expensive. Yet clinical knowledge becomes all important in these times, for several reasons. New drugs, new methods of surgery, the expansion of organ transplantation techniques and practice all lead to greater demand for more expensive medical care. It becomes essential that treatments be evaluated carefully for efficacy and safety, and that treatments be tailored to needs and to the balancing of cost with benefit. The randomized clinical trial plays a central role in obtaining reliable answers; but answers can also be obtained from information gathered in medical practice. Carefully gathered data on medical practice in both hospital and clinical settings, accessible by computer, can be a valuable basis for statistical comparisons of treatment outcomes, for patients with various degrees of illness and other concomitant characteristics (e.g. age, ancillary diseases, history of disease). The problems of statistical inference from such data banks are complex and have to do with the definitions of the variables, the number of variables, the collection of the data, computer storage and access, and prudent and valid statistical analysis of the data. Byar [4] cautions that in light of the difficulties in these complex nonrandomized studies, inferences that are drawn can be far from convincing. The statistical methods available to the investigator are not adequate to deal with the questions asked and the precision required of the answers. The classical methods of **multivariate analysis** are helpful and will continue to grow in their power and use, but new methods are needed. Biostatisticians draw ideas from the past, using the ideas of Bayes to correct hospital estimates from each of a group of hospitals for **regression to the mean** for the group, before drawing inferences about differences in practice effects from one hospital to another. Here, methods of more than a century past can be blended with the power of the modern computer to resample the huge masses of data in the data banks, to merge data from different data banks (*see* **Record Linkage**) to obtain estimates for comparison, and to evaluate the precision of estimates by resampling methods such as the bootstrap approach. Biostatistical work in this area, directed at the collection of data, merging data bank information, and making reliable inferences and decisions from such data is an area that demands the biostatistician's attention.

Chronic diseases themselves pose experimental problems that in turn lead to statistical questions of method. To answer a question regarding the relative efficacy of an experimental treatment to a control treatment, when the event of interest is mortality, patients must be followed for extended periods of time, years or even decades. When the measurement of interest is blood pressure, or physical strength, or kidney function, repeated measures on each individual must be made. The data gathered across subjects are necessarily asynchronous, even if the intention is to measure every patient according to a fixed schedule, say at monthly or annual intervals. Patients miss appointments or appear for extra appointments for various reasons related or unrelated to study goals (*see* **Missing Data**). Some are lost to follow-up, or die of one putative cause or another. Obviously, such data present problems for comparison of the groups with regard to efficacy of a therapy or mode of medical care. The consequent statistical problems in interpreting such longitudinally gathered data have been a concentrated focus of biostatistical research and application over recent years (*see* **Longitudinal Data Analysis, Overview**). The result has been a number of new proposals for analysis. The work has been extremely useful and the work will go on, again it being a blend of new theoretical proposals, much dependent on computer power for both method and for the examination and comparison of methods [14–16].

One final current area of challenge to biostatistics merits mention. It is an area that occupies teams of geneticists, probabilists, biostatisticians, and computer scientists. The new methods of isolating and mapping the human genetic structure (and that of other organisms as well), mapping the genetic structures of inheritance, and grasping the fine relationships between genetic structure and human function (*see* **Genetic Map Functions**) has opened up a huge potential for medical study in the prevention of disease, enhancement of health, and the treatment of disease. But, here again, the questions are new and call for new statistical concepts. The questions start

with the uncertainties in mapping the human genome, a vast undertaking with basic questions of just how to design the mapping efficiently and how to measure the uncertainties in the measurements and the implications of error with regard to the work in progress and to the "final" outcomes. There are questions having to do with the choice of persons to study, the kinds of familial structures to study, and how to weave together the genetic information and the health data to make a predictive model with measurable statistical precision. These results are only now being used in the laboratory and clinic to design vaccines to prevent disease, for example graft-vs.-host disease in bone marrow transplantation, or the conversion of HIV infection to frank **AIDS**, or to tailor a vaccine to a specific cancer in an individual patient as an aid in his/her fight to prevent further development of his/her personal cancer. The questions call for modification in the design and analysis of methods long used, but also the use of new methods aimed at new questions and newer cost and time-efficient approaches to the goals. Jurg Ott gives a comprehensive and thorough introduction to one of the large areas of very active further methodologic research, namely the linkage of genes in humans [19].

In summary, the times present Biostatistics with the triple challenge of completely new areas of science with new kinds of questions, e.g. in genetics, tremendously improved tools for use (notably in both theoretical statistics and computing), and new questions arising from the escalating costs of both medical care and medical research (demanding a closer focus on efficient research and parsimony in monitoring the costs and effectiveness of medical care itself).

Organization of Biostatistics as a Professional Discipline

It was felt by the editors that the focus and breadth of the body of specialized statistical knowledge and the size of the body of professionals in the field of medical and biological statistics justified the organization of an encyclopedia for this now well-established discipline. The general field of statistics itself has been clearly identified for nearly two centuries and the medical and medically related applications – the roots of biostatistics – in genetics, epidemiology, the basic medical sciences, and applications in demography, psychology, and government, have been present for more than a century. The **American Statistical Association (ASA)** was established in 1839, with a regularly published Journal. In England, the **Royal Statistical Society** was established in 1834 as the Statistical Society, with the aim of publishing "facts calculated to illustrate the condition and prospects of society". It publishes three series of journals, dealing with theory and with statistical methods and applications in all areas of science. The **International Biometric Society**, with world organization and membership, was established in 1947 and now has approximately 6200 members. The Biometrics Section of the ASA started a special professional bulletin in 1945, which was adopted by the Biometric Society in 1950 as the official journal of the Society and called ***Biometrics***. There are now journals in many countries devoted at least in part to biostatistics or special aspects of it, e.g. ***Statistics in Medicine***, ***Biometrical Journal*** (Germany), ***Controlled Clinical Trials***, ***Journal of Biopharmaceutical Statistics***, and ***Statistical Methods in Medical Research***.

Academic departments offering specialized statistical courses in the methods and applications of statistics in medicine have existed from the early nineteenth century. In the US, these offerings were generally in schools of public health, where the methods and applications were focused on the areas of sanitation, epidemiology, demography, and vital statistics. Today, most medical schools across the world offer formal instruction in biostatistics to the medical degree candidate (*see* **Teaching Statistics to Medical Students**) and to candidates for degrees in most of the allied medical fields, e.g. dentistry, nursing, health administration, sanitary engineering, where such degrees are offered by the university.

Degrees in biostatistics are offered at a number of universities. These are generally advanced degrees and most are offered in schools of public health. Course work for the master and doctorate degrees is directed to theory and applications of statistics to applications in the broad range of the medical sciences, with an emphasis on clinical and public health areas. Those interested in careers in biostatistics and best suited for satisfying and productive careers will be interested in the mathematical and theoretical aspects of science, will have a curiosity and inclination for the sciences, and will enjoy teamwork and scientific collaboration. Teaching is an important part of the biostatistician's role, whether it is the one-to-one teaching of the consultant to

biomedical colleagues or the more formal teaching of the classroom, to students of biostatistics or students of the biomedical sciences. Of course, specific roles for the biostatistician span the range of applications from government to academia to industry, and involve applications from the most basic of the medical sciences to applications in the clinical practice of medicine and to applications in the organization of health care, as well as to the activities of the government in the planning, financing, and distribution of health care, and the measurement of health care status and health progress in the population.

The reader interested in the field of biostatistics as a career that is located on the bridge between statistical theory and scientific investigation in biomedicine might start with reading selected papers in those professional journals that include occasional expository articles, interviews with professionals, and presidential addresses. *Statistical Science*, the *American Statistician* and the ***Journal of the American Statistical Association*** are several. The professional biostatistics and statistics societies also circulate publications that contain news of the societies and provide a view of activities, the accomplishments of its members, and programs of approaching professional meetings. It will be seen that biostatisticians can be found throughout all biomedical research and application – academia, research institutes, industry, and government.

References

[1] Ashby, D. (1993). Papers from the Conference on Methodological and Ethical Issues in Clinical Trials, *Statistics in Medicine* **12**, 1373–1534.

[2] Bliss, C.I. (1952). *The Statistics of Bioassay*. Academic Press, New York.

[3] Bunker, J.D., Forrest, W.N., Jr, Mosteller, F. & Vandam, L.D. (1969). *The National Halothane Study*. National Institutes of Health, Maryland.

[4] Byar, D. (1980). Why data bases should not replace randomized clinical trials, *Biometrics* **36**, 337–342.

[5] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[6] Efron, B. & Tibshirani, R.J. (1992). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[7] Emmens, C.W. (1948). *Principles of Biological Assay*. Chapman & Hall, London.

[8] Finney, D.J. (1978). *Statistical Method in Biological Assay*, 3rd Ed. Griffin, London.

[9] Fisher, R.A. (1966). *The Design of Experiments*, 8th Ed. Hafner, New York.

[10] Flood, A.B. & Scott, W.R. (1987). *Hospital Structure and Performance*. Johns Hopkins University Press, Baltimore.

[11] Graunt, J. (1962). Natural and Political Observations, mentioned in a following INDEX, and made upon the Bills of Mortality. William Hall, for John Martyn and James Allestry, for the Royal Society, Oxford.

[12] Hill, A.B. (1961). *Principles of Medical Statistics*. 7th Ed. Oxford University Press, New York.

[13] Kelsey, J.L., Whittemore, A.S., Evans, A.S. & Thompsen, W.D. (1996). *Methods in Observational Epidemiology*. Oxford University Press, New York.

[14] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[15] Liang, K.-Y. & Zeger, S.-L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[16] McCullagh, P. & Nelder, J.S. (1983). *Generalized Linear Models*. Chapman & Hall, London.

[17] Meinert, C.L. & Tonascia, S. (1986). *Clinical Trials: Design, Conduct and Analysis*. Oxford University Press, New York.

[18] Miller, R.G., Jr (1986). *Beyond ANOVA, Basics of Applied Statistics*. Wiley, New York.

[19] Ott, J. (1991). *Analysis of Human Genetic Linkage*, rev. Ed. The Johns Hopkins University Press, Baltimore.

[20] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population, *Journal of the Royal Statistical Society, Series B* **4**, 119–130.

[21] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, New York.

[22] Quenouille, M.H. (1956). Notes on bias in estimation, *Biometrika* **43**, 353–360.

[23] Redman, R., Nader, S., Wong, R., Brown, B.W., Jr & Arvin, A. (1997). Early reconstitution of immunity and decreased severity of herpes Zoster in bone marrow transplant recipients immunized with inactivated varicella vaccine, *Journal of Infectious Diseases*, in press.

[24] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown & Company, Boston.

[25] Shortliffe, E.H. & Perreault, L.E., eds (1990). *Medical Informatics*. Addison-Wesley, New York.

[26] Wain, H. & Messick, S., eds (1983). *Principles of Modern Psychological Measurement: A Festschrift for Frederic Lord*. Lawrence Erlbaum, Hillsdale.

[27] White, K.L., editor-in-chief (1992). *Health Services Research: An Anthology*, no. 534 Pan American Health Organization.

W. BYRON BROWN, JR

Biostatistics

Biostatistics, established in 2000 by Oxford University Press, publishes papers that advance statistical reasoning and methods relevant to studies of human health and disease. Authors and readers include academic and professional statisticians as well as other quantitatively oriented biomedical or public health researchers.

The Journal publishes papers of methodologic and substantive consequence. One of the aims of the journal is to achieve review-times that are significantly shorter than are typical of statistics journals, and more competitive with review-times for biomedical journals. The more rapid response to submitted papers is made possible by a small editorial board comprising outstanding scholars who themselves provide critiques of submitted papers and substantially determine what is published. In addition, accepted papers are accessible from the journal website prior to publication in the journal. The website also provides access to supplementary material including data, programs, and text that support the paper publication.

Although *Biostatistics* is a young journal, it is selective, publishing roughly one in four submitted papers. Electronic submissions are encouraged. The journal website `http://biostatistics.oupjournals.org/` provides further details.

PETER J. DIGGLE & SCOTT L. ZEGER

Birth Cohort Studies

Birth cohort studies are those which begin at or before the birth of their subjects, and continue to study the same individuals at later ages, on more than one occasion. They are a type of observational study in which "there is no randomization to exposure classes nor is there any attempt to manipulate the exposure" [10]. They vary in population size from large studies that aim to be nationally representative [12, 24, 33], to those that are area based and with populations of 1000 or more subjects [1, 8, 11, 14, 16, 19, 21, 26, 27, 29, 36]. Currently new, nationally representative birth cohort studies are being started in Denmark and Canada.

Although historical birth cohorts have been imaginatively used in epidemiology [3, 17] they are not discussed here.

Study Population

Selection

Prospective and historical birth cohort studies usually select their populations using time and/or geographical sampling frames. Three British birth cohort studies, which began at the birth of their subjects, each used a sampling frame of all births occurring in one week [12, 24, 33]. The oldest study followed up a class-stratified sample of all the single and legitimate births that occurred during a week, in 1946, and the two later studies followed up all births from the chosen week in 1958 and in 1970. The Avon longitudinal study of all births occurring in one English county used a year's births as a sampling frame, and recruited at antenatal clinics during that period, in order to collect data on risk exposure during pregnancy [14].

Population Size

In a birth cohort study the size of population must be selected at the outset, which may be far in time from some intended outcome measures. Definition of the outcome measures, their age-related incidence, and expected sample attrition at different future times can be used to calculate sample size. Large samples offer the opportunity to study relatively rare occurrences, but not without penalty. Inevitably, sample size is associated with frequency of data collection and with data quality. Large samples are more costly in terms of data collection and subjects can be so costly to contact that time intervals between data collections become long, and undue reliance has then to be given to subjects' recall of events and experiences occurring since the previous data collection.

Contact Maintenance

Contact maintenance is a constant task in a large birth cohort study. Annual contact is the ideal, so that changes of name and address can be kept up to date. In return, information on the study's work helps to maintain the subjects' interest. The two older British studies achieve these ends by means of a birthday card, which is an advantage of the time-based sampling frame. Each also sends an annual description of current work, and an annual request for information on address and name changes.

The tracing of lost contacts is relatively easy during the preschool and school years, when health and education systems may offer assistance, but in the subjects' adult lives investigators have to rely on information given by parents (rapport developed with parents during the school years is of value in these later years), and on agreements with others to forward letters to subjects.

Attrition and Representativeness

None of the British national studies found attrition and its effects on representativeness a serious problem during the preschool and school years, because of the help received from health and education authorities. But, whereas during the childhood and school years of the 1946 national birth cohort study – for instance, data collection was achieved from between 85% (lowest) and 96% (highest) of the live study population resident in Britain – the comparable adult range of response was from 67% to 85% (highest) [33].

Loss of sample members through failure to maintain contact is usually higher in those with the lowest educational attainment or interest in education, those living in the poorest socioeconomic circumstances, those who are single in adult life, and in the mentally ill, but not in those with serious physical illness [35].

Loss of sample members through death can, in Britain, be checked with the National Health Service register, on which study members in the two earlier birth cohorts are "flagged" so that death certification is notified to the study. In other countries similar methods of checking against files of the deceased may be possible (*see* **Death Indexes**).

Attrition in birth cohort studies through emigration, refusal, and loss of contact distort the representation of the study population. Distortion is caused also by inward migration. The 1946 birth cohort, for example, represented at age 43 years the native born, legitimate population, but not those who had been illegitimately born (4% of those born in the chosen week), nor the 5% of the British population of the same age as cohort members at 43 years who (in 1989) were not native born. The 1958 cohort augmented the population selected at birth by including in data collections at ages 7, 11, and 16 years all the children born in the sampled week, even if they had not been included in the original study of births.

Topics of Study

Unique Assets of Prospective Studies from Birth

Two groups of assets of birth cohort studies are conferred by their design. The first is an advantage of having information on individual developmental time passing, and the second is associated with the individual's experience of historical time.

First are assets conferred by prospective data collection. This method provides information on the sequence of events, which is essential to the understanding of causation, and of risk and protective factors. It also provides information which cannot be gained on all cohort members, or even at all in retrospect, or from records – for example, cognitive scores, and information on attitudes, hopes and aspirations, growth, and behavior. This includes information on individuals' physical and mental change over time, and exposure to illness risk. Birth cohorts that are general population samples can provide viable denominator as well as numerator information, and therefore the relative and absolute risks, the effects of differential mortality, and the heterogeneity of outcomes of the hypothesized risk can all be calculated with some accuracy.

The second asset of birth cohort studies conferred by their design is that their populations have passed through known historical times. Thus, for example, the earliest British birth cohort was born at a time of high likelihood of parental smoking, lived the first eight years of life in circumstances of wartime food rationing and the first two years without a national health service, experienced selective entry to secondary schools, and lived all of the infant and childhood years before the Clean Air Act greatly reduced atmospheric pollution. The later-born cohorts, by contrast, lived in less austere times, with increasing awareness of the risks of smoking, less atmospheric pollution from coal burning, and comprehensive education. Members of each cohort came to the historical high period of unemployment, which began in the 1980s, at different career stages. Members of the 1946 cohort lived through the early postwar polio epidemics, and mothers of the 1958 cohort were pregnant during the influenza epidemic in the winter of 1957–1958. These differences of experience can be used to investigate the effects of different kinds of risk and exposure to risk. For example, a study of schizophrenia in the 1958 cohort confirmed that exposure to influenza in pregnancy was associated with raised risk of schizophrenia in offspring [6]. Comparative studies of perinatal mortality in the three British cohorts have been used to examine the effects of changes in obstetric care over the 24 year period [5], and these studies of maternity and childbirth were why the national cohort studies began.

Effects of Historical Time on Topics of Study

Although birth cohort studies are necessarily science-led, they are also inevitably products of their time. This is seen in the population size, which has been conditioned in the past by available information technology and in the selection of data collected, as well as in the initial decision to begin the studies, as already described.

In retrospect, in a long running study it is easy to see what appear later to be omissions in data collection. For example, in the 1946 birth cohort no information was collected during the early years on parental smoking, because the recognition of its damaging effects came almost a decade after the study began. Similarly, in the same study the data on child health collected by school nurses followed the pattern of the current school medical examinations; with hindsight, information on biological function, such as

blood pressure and respiratory function, would have been invaluable.

Some kinds of information were perceived as important in the early years of the 1946 cohort, but there were no suitable research instruments for their collection. The importance of postweaning nutrition in childhood, for instance, although recognized, was very little studied in any of the British birth cohorts. None of the British studies has information on the parents' relationships with one another during the child's early years, because it was thought to be impossible to assess at earlier times, and later because available instruments were too time-consuming. In consequence, whilst each has been able to study the associations of parental divorce and separation with the child's health and well-being, it has not been possible to compare the apparent effects of divorce with those of living in harmonious and disharmonious family circumstances.

Not only is the choice of variables and measurement instruments a product of the time, so also is the general direction of the studies [27]. Current scientific and political expediency are vital forces that shape long-running studies. The 1946 cohort continued to be viable in the children's first five years because of current questioning of the value of home visiting undertaken by nurses involved in maternal and child welfare: the study was able to show the good effects of such a service [33]. During the cohort's school years, the effectiveness of the selective process for entry to secondary school (at age 11 years) was questioned. By having measures of cognitive attainment taken at age 8 years, three years before the selection process, the 1946 study was able to show that the method of selection was, as had been feared, biased in favor of the middle-class child. Furthermore, the study showed that the origins of this problem lay not simply with the selection process itself, but also in child–parent–teacher relationships from the earliest times at school, and in parent attitudes to education [7, 33]. Similarly, the 1958 cohort was well-placed to undertake studies of primary school education for the contemporary government enquiry into education in primary schools [13, 23]. In more recent times, the 1958 study has been ideally placed to investigate the effects of unemployment and preexisting circumstances on mental and physical health and on the acquisition of social capital in early adulthood, since cohort members were in their early years of work when the national rise in unemployment began [20]. The 1946 study is now contributing to understanding the processes of aging, and is well-placed to do so, not only because of its lifetime data, but also because this population represents the beginning of the boom years in the population of the elderly.

Previous and Later Generations

The two older British birth cohorts have some information on the parents of study members, particularly their educational attainment, occupation, marital status and stability, concern for the study child's education, and cause of death, as well as some information about health. This has made it possible to study social mobility, the effects of parental ill health on the life of the study child, and intergenerational differences in religious and political adherence, as well as the effect of some important aspects of family of origin, and cohort members' educational attainment [25, 34].

The first two British cohorts also have information on the offspring of cohort members. The 1946 study interviewed mothers of all first children born to male and female cohort members between ages 19 and 25 years. Information on these second-generation children was collected at the end of the preschool period, at age 4 years, and again at 8 years, because the purpose of the study was to explore further the finding in the previous generation that parental concern for education was strongly associated with attainment scores [32, 33]. The 1958 study, on the other hand, interviewed a randomly selected one in three of the cohort members who were parents at age 33 years about all their children ($n \simeq 4000$), to compare differences in a wide range of aspects of childhood in two generations.

Now that the 1946 cohort members are in middle life, questions have been included at the interview at 43 years about their relationships with, and (where necessary) care of, their elderly parents, who were by that time aged 71 years, on average [34].

Although these studies of the two generations adjacent to cohort members are useful for the study of intergenerational relationships, their populations cannot be regarded as representative of those generations.

Forward Planning

Forward planning is difficult in a long-term prospective study because the variables selected for current

data collection will be used for two purposes. They will be used as outcome measures in relation to data collected at earlier times. They will also become precursor variables used to test hypotheses about risk. It is therefore necessary, in planning data collections, to consider the needs of future hypothesis testing. For example, planning midlife data collection in the 1946 cohort has had to include consideration of requirements for testing hypotheses about health in later life. For instance, current hypotheses about skeletal fractures and repair in old age implicate smoking, diet, and exercise in midlife, and so information on these health-related habits will be collected. The risk is that by the time the cohort reaches old age, hypotheses will change, and new ideas will develop. Although that is an inevitable problem, in practice, for biological questions, measures of current function and its change over time, and the preservation of blood samples, provide a range of data possibilities for future use. In general, once a future biological, psychological, or social outcome has been defined, then, the demands of current hypotheses having been taken into account, the most detailed possible measures available should be taken to make future accurate assessments of change, and not to constrain future analyses.

In retrospect, it is clear that the 1946 study has been fortunate. Investigations of midlife cognitive function, for example, have been able to use data on this topic collected in childhood, adolescence, and early adulthood for the study of educational attainment. Similarly, information on birth weight and infant growth and development, originally collected to study social variation in physical development, has proved to be of unanticipated value in the study of midlife blood pressure and respiratory function, and of schizophrenia.

Other Sources of Data

In addition to cohort members themselves, and in childhood their parents, birth cohort studies have found it invaluable to collect information from other sources, for four reasons.

First, confirmatory information can usefully be collected, with permission, for example, from hospital records and educational certification bodies. Secondly, information can be collected to provide another view of the circumstance. For instance, in the 1946 study, teachers' comments added considerably to the information given by cohort members and by parents. Thirdly, information from other sources, including census data, can provide information on an area basis about exposure to such things as atmospheric pollution and the nature of water supplies, as well as local rates of unemployment, educational attainment, and socioeconomic structure. The fourth kind of information concerns the current social and scientific context. In a long-running prospective study it is difficult to know in retrospect the political and scientific pressures and social concerns of the day that affected the direction that the study took. They are difficult to ascertain years later, because histories of the period are usually concerned with political pressure and events rather than social and scientific histories, although there are notable exceptions. In trying to account for earlier choice of subjects, measures, and direction of analyses, a study logbook is needed to supplement such other sources as reports of funding bodies, government enquiries, and commissions.

Analysis

Coding

The effects of historical time in a long-running study can also be problematic in terms of how information is coded and stored. Classifications of illnesses and socioeconomic position change over time, and it is often necessary to recode information already coded, and to retain two or more sets of classifications both to study change over time and to be comparable with current work in other studies.

Concepts of Analysis

Now that birth cohorts, historical cohorts, and other forms of longitudinal studies have been collecting data for 50 years and more, the long-term nature of these investigations has brought a life-span perspective to analysis [15, 22, 31]. This has generated new ideas about the nature of psychological and biological aging, and raised the question of the role and measurement of adaptation with age [2]. So far, this has been largely a matter of psychological study, but questions about variation in biological vulnerability in relation to age are now being discussed in view of the data available from prospective and historical studies that encompass many years of life [3].

The most commonly used approach to analysis in birth cohort and other long-term investigations is concerned with variables, and the most commonly used method is linear regression. More recently, **structural equations models** have also been used to handle interactions between variables that measure function or environment [9]. Alternatively, a person-oriented, as compared with a variable-oriented, approach has been advocated [4, 18]. This method compares individuals' profiles of characteristics determined by analysis of clusters or patterns (*see* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**). Magnusson and Bergman [18] illustrate this approach in comparison with variable-based analysis, which in a study of precursors of crime and aggression would show how

> each of a number of single aspects of individual functioning – aggressiveness, hyperactivity, low school motivation, poor peer relations – is significantly related to various aspects of adult maladjustment. . . .Applying the pattern approach to this research area, the focus instead becomes: what typical patterns or configurations of adjustment problems of this kind actually exist in childhood; how are major adjustment problem areas in adult age interconnected; and what are the relationships between typical problem patterns in childhood and typical problem patterns in adulthood?

This account of the nature of birth cohort studies cannot describe all aspects, nor refer to every study. Broader summaries and reviews are given by Sontag [30], Mednick et al. [19], Schneider & Edelstein [28], and Young et al. [37].

References

[1] Alison, L.H., Counsell, A.M., Geddis, D.C. & Sanders, D.M. (1993). First report from the Plunket National Child Health Study: smoking during pregnancy in New Zealand, *Paediatric and Perinatal Epidemiology* **7**, 318–333.

[2] Baltes, P.B. & Baltes, M.M. eds (1990). *Successful Aging*. Cambridge University Press, Cambridge.

[3] Barker, D.J.P. (1991). *Fetal and Infant Origins of Adult Disease*. British Medical Journal, London.

[4] Block, J. (1971). *Lives Through Time*. Bancroft, Berkeley.

[5] Chamberlain, R., Chamberlain, G., Howlett, B. & Claireaux, A. (1975). *British Births 1970*, Vols 1 and 2. Heinemann, London.

[6] Done, D.J., Crow, T.J., Johnstone, E.C. & Sacker, A. (1994). Childhood antecedents of schizophrenia and affective illness: social adjustment at ages 7 and 11 years, *British Medical Journal* **309**, 699–703.

[7] Douglas, J.W.B. (1964). *The Home and the School*. McGibbon and Kee, London, pp. 119–128.

[8] Dragonas, T., Golding, J., Ignatyeva, R. & Prokhorskas, R., eds (1996). *Pregnancy in the 90s*. Sansom, Bristol.

[9] Dunn, G., Everitt, B. & Pickles, A. (1993). *Modelling Covariances and Latent Variables Using EQS*. Chapman & Hall, London.

[10] Feinleib, M., Breslow, N.E. & Detels, R. (1991). Cohort studies, in *The Oxford Textbook of Public Health*, W.W. Holland, R. Detels & R.G. Knox, eds. Oxford University Press, Oxford, pp. 145–159.

[11] Fergusson, D.M., Horwood, L.J., Shannon, F.T. & Lawton, J.M. (1989). The Christchurch Child Development Study: a review of epidemiological findings, *Paediatric and Perinatal Epidemiology* **3**, 302–325.

[12] Ferri, E., ed. (1993). *Life at 33*. National Children's Bureau, London.

[13] Fogelman, K., ed. (1983). *Growing Up in Great Britain*. Macmillan, London, pp. 231–326.

[14] Golding, J. (1989). European longitudinal study of pregnancy and childhood, *Paediatric and Perinatal Epidemiology* **3**, 460–469.

[15] Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies*. Academic Press, London.

[16] Kolvin, I., Miller, F.J.W., Scott, D.M., Gatzanis, S.R.M. & Fleeting, M. (1990). *Continuities of Deprivation? The Newcastle 1000 Family Study*. Avebury, Aldershot.

[17] Lumey, L.H., Ravelli, A.C.J., Wiessing, L.G., Koppe, J.G., Treffers, P.E. & Stein, Z.A. (1993). The Dutch famine birth cohort study, *Paediatric and Perinatal Epidemiology* **7**, 354–367.

[18] Magnusson, D. & Bergman, L.R. (1990). In *Straight and Devious Pathways from Childhood to Adulthood*, L.N. Robins & M. Rutter, eds. Cambridge University Press, Cambridge, pp. 101–115.

[19] Mednick, S.A., Baert, A.E. & Bachmann, B.P., eds. (1981). *Prospective Longitudinal Research*. Oxford University Press, Oxford.

[20] Montgomery, S., Bartley, M., Cook, D. & Wadsworth, M.E.J. (1996). Health and social precursors of unemployment in young men, *Journal of Epidemiology and Community Health* **50**, 415–422.

[21] Olsen, P., Laara, E., Rantakallio, P., Jarvellin, M.-R., Sarpola, A. & Hartikainen, A.-L. (1995). Epidemiology of preterm delivery in two birth cohorts with an interval of 20 years, *American Journal of Epidemiology* **142**, 1184–1193.

[22] Plewis, I. (1985). *Analysing Change*. Wiley, Chichester.

[23] Plowden Report (1967). *Children and Their Primary Schools*, Vol. 2. HMSO, London, pp. 401–543.

[24] Power, C. (1992). A review of child health in the 1958 birth cohort, *Paediatric and Perinatal Epidemiology* **6**, 81–110.

[25] Power, C., Manor, O. & Fox, J. (1991). *Health and Class: The Early Years*. Chapman & Hall, London.

[26] Rantakallio, P. (1988). The longitudinal study of the North Finland birth cohort of 1966, *Paediatric and Perinatal Epidemiology* **2**, 59–88.

[27] Roche, A. (1992). *Growth, Maturation and Body Composition: the Fels longitudinal study, 1929–1991*. Cambridge University Press, Cambridge, p. 199.

[28] Schneider, W. & Edelstein, W. eds (1990). *Inventory of European Longitudinal Studies in the Behavioural and Medical Sciences*. Max-Planck Institute, Berlin.

[29] Silva, P. (1990). The Dunedin Multidisciplinary Health and Development Study: a 15 year longitudinal study, *Paediatric and Perinatal Epidemiology* **4**, 76–107.

[30] Sontag, L.W. (1971). The history of longitudinal research, *Child Development* **42**, 987–1002.

[31] Sugarman, L. (1993). *Life-Span Development*. Routledge, London.

[32] Wadsworth, M.E.J. (1986). Effects of parenting style and preschool experience on children's verbal attainment, *Early Childhood Research Quarterly* **1**, 237–248.

[33] Wadsworth, M.E.J. (1991). *The Imprint of Time*. Oxford University Press, Oxford.

[34] Wadsworth, M.E.J. (1996). Social and historical influences on parent-child relations in midlife, in *The Parental Experience in Midlife*, C. Ryff & M.M. Seltzer, eds. Chicago University Press, Chicago, pp. 169–212.

[35] Wadsworth, M.E.J., Mann, S.L., Rodgers, B., Kuh, D.L., Hilder, W.S. & Yusuf, E.J. (1992). Loss and representativeness in a 43 year follow up of a national birth cohort, *Journal of Epidemiology and Public Health* **46**, 300–304.

[36] Yach, D., Cameron, N., Padayadhee, N., Wagstaff, L., Richter, L. & Fonn, S. (1991). Birth to ten: child health in South Africa in the 1990s. Rationale and methods of a birth cohort study, *Paediatric and Perinatal Epidemiology* **5**, 211–233.

[37] Young, C.H., Savola, L.K. & Phelps, E. (1991). *Inventory of Longitudinal Studies in the Social Sciences*. Sage, London.

M.E.J. WADSWORTH

Birth Defect Registries

Birth defects registries can be used to determine the newborn prevalence of birth defects and they can provide cases for studies [3]. Both of these functions are important foundations for the prevention of birth defects.

A birth defects registry is an organized database containing information on individuals born with specified congenital disorders, ascertained from a defined source. Birth defects are a heterogeneous group of conditions, present at birth, and not generally considered to include injuries suffered in and around the birth process. In the US, the March of Dimes Birth Defects Foundation defines a birth defect as "an abnormality of structure, function or body metabolism (inborn error of body chemistry) present at birth that results in physical or mental disability, or is fatal" [9].

There are more than 4000 known types of birth defects. These include conditions ranging from congenital malformations of unknown etiology, to genetic conditions due to abnormal genes or chromosomes, to conditions due to damage to the developing fetus related to prenatal maternal exposure to hazardous environmental agents. Thus, this term includes a range of disorders such as cleft lip and palate, heart malformations, clubfoot, Down syndrome, phenylketonuria (PKU), fragile X-linked mental retardation, fetal alcohol syndrome and the thalidomide syndrome. Not all of these conditions are necessarily medically or cosmetically significant. Some "minor" defects, such as small birthmarks or skin tags, are of no significant consequence.

Causes of Birth Defects

Birth defects may be caused by genetic or environmental factors (including drugs and chemicals, infectious agents, physical agents, and maternal health or nutritional factors) acting singly or in combination. The causes of more than half of all birth defects are unknown at present. Collectively, birth defects are a major cause of disability and morbidity, and in the US, represent the leading cause of infant mortality.

Information in Birth Defects Registries

Birth defects registries systematically collect information such as identifiers and demographic information, description of the defects present, family history, history of the pregnancy (including prenatal exposures), and other information potentially related to risk factors. This information may be gathered from vital records, hospital or clinic records, or other sources of case ascertainment. Registries that rely upon records collected for other purposes (e.g. birth certificates) are sometimes referred to as using "passive" methods of ascertainment, while those that use multiple sources and collect information from patient records are said to use "active" ascertainment. Information stored in birth defects registries is typically stored as both paper files and in electronic systems, often using a relational database. Considerable effort is expended to ensure data security, as well as the confidentiality and privacy of the individuals and families who are participants in the registry.

Usually, information collected is restricted to a defined age range such as the neonatal period or the first year of life. Indeed, since many defects are not identified before neonatal hospital discharge, the longer the ascertainment period used, the higher the frequency of children identified with birth defects in a particular birth cohort. Systems actively ascertaining cases throughout the first year of life typically find that 3.5%–5% of infants have such conditions.

Birth defects registries generally collect cases from a defined population, such as that in a specified geographic area. When they do, the registry can use birth certificate data collected for governmental purposes on all babies as the denominator. Thus, it becomes possible to determine the newborn prevalence of birth defects in this geographic area. Such population-based registries also permit case–control studies. Many countries have national registries. These programs share information through an International Clearinghouse for Birth Defects Monitoring Programs [4]. This organization has contributed to standardized terminology and methods and has helped in the organization of international collaborative research efforts. Non population-based registries can provide cases for studies that do not need a population base. Such collections of patients have been particularly useful for studies seeking to identify the gene responsible for a single gene cause of a birth defect.

Uses of Birth Defects Registries

Registries that are population-based permit the rates of birth defects to be determined. Such rates can be

monitored to look for changes that would suggest a newly introduced environmental/drug agent. It has been argued that had there been registries established prior to the thalidomide epidemic, the cause may have been determined sooner than it was. On a more positive note, when there is the possibility to prevent birth defects like rubella and folic-acid-preventable spina bifida and anencephaly, monitoring the trends can determine the effectiveness of prevention programs. There is currently great interest to determine whether or not the folic acid fortification plan in the US will provide full protection from folic-acid-preventable birth defects.

Given that we do not know the causes of the birth defects in the majority of infants born with them, a major function of birth defects registries has been to supply cases for studies seeking to find the causes of birth defects. The most common kind of epidemiologic study seeking to find causes is the **case–control study**. In these studies the exposures and family history of cases are compared with those of controls selected from the same population from which the case is drawn.

One notable example of such a study is the one from France that found that cases of spina bifida were much more likely to have been exposed to valproic acid than controls. This study led rapidly to a health warning in the US [6]. Another useful example is provided by the Atlanta Case–Control Study conducted by the US Centers for Disease Control and Prevention (CDC). This study provided the opportunity to examine possible associations between many kinds of birth defects and many drug and environmental exposures. The protective association between regular multivitamin consumption and the reduction in spina bifida and anencephaly shown in Atlanta and other observational studies, provided critical data for public health policy [8]. These studies led to a recommendation by the US Public Health Service (PHS) that women of reproductive age consume 400 μg of synthetic folic acid daily, rather than the 4000 μg that was used in the randomized controlled trial [2, 7]. Because data on so many exposures were collected, there remain data yet to be analyzed that may provide insight to other etiologies of these and other birth defects [5].

Recently, investigators have sought to improve the search for causes of birth defects by using both molecular markers of exposure and outcomes [10]. With the genome project soon to provide data on all genes, investigators will have even more powerful tools to try to understand the interplay between environmental and genetic risk factors in the cause of birth defects. Thus, registries are a rich resource for clinical and family studies, for health outcomes and services research, and for public health planning and programming [1].

Authority for Operation of Birth Defects Registries

Birth defects registries are generally established and operated under the authority of public health agencies. In the US, authority for such public health surveillance programs is generally the responsibility of the states. As a consequence, there are specific statutes authorizing such a registry or surveillance system, and there is considerable variation in specific goals, methods of financing, operational policies, and scope of disorders covered. Furthermore, other state and federal policies (e.g. health information and privacy policies) affect the type and circumstances under which research can proceed.

References

[1] Adams, M.M., Greenberg, F., Khoury, M.J., Marks, J.S. & Oakley, G.P. Jr (1985). Survival of infants with spina bifida – Atlanta, 1972–1979, *American Journal of Diseases of Childhood* **139**, 518–523.

[2] Centers for Disease Control and Prevention (1992). Recommendations for the use of folic acid to reduce the number of cases of spina bifida and other neural tube defects, *Morbidity and Mortality Weekly Reports* **41**, 1–7.

[3] Edmonds, L.D., Layde, P.M., James, L.M., Flynt, J.W. Jr, Erickson, J.D. & Oakley, G.P. Jr (1981). Congenital malformations surveillance: two American systems, *International Journal of Epidemiology* **10**, 247–252.

[4] Erickson, J.D. (1991). The International Clearinghouse for Birth Defects Monitoring Systems: past, present, and future, *International Journal of Risk and Safety Medicine* **2**, 239–248.

[5] Erickson, J.D. (1991). Risk factors for birth defects: data from the Atlanta birth defects case–control study, *Teratology* **43**, 41–51.

[6] Lammer, E.J., Sever, L.E. & Oakley, G.P. Jr (1987). Teratogen update: valproic acid, *Teratology* **35**, 465–473.

[7] MRC Vitamin Study Research Group (1991). Prevention of neural tube defects: results of the Medical Research Council Vitamin Study, *Lancet* **338**, 131–137.

[8] Mulinare, J., Cordero, J.F., Erickson, J.D. & Berry, R.J. (1988). Periconceptional use of multivitamins and the occurrence of neural tube defects, *Journal of the American Medical Association* **260**, 3141–3145.

[9] Petrini, J., ed. (1997). *March of Dimes Birth Defects Foundation StatBook*. March of Dimes Birth Defects Foundation, White Plains, p. 259.

[10] Shaw, G.M., Wasserman, C.R., Murray, J.C. & Lammer, E.J. (1998). Infant TGF-alpha genotype, orofacial clefts, and maternal periconceptional multivitamin use, *Cleft Palate-Craniofacial Journal* **35**, 366–370.

(*See also* **Cancer Registries**; **Infant and Perinatal Mortality**; **Teratology**)

JAMES W. HANSON &
GODFREY P. OAKLEY, JR

Birthweight

The **World Health Organization (WHO)** defines birthweight as the first weight of the fetus or baby after birth. Although parents commonly include the baby's birthweight on the cards they send to friends and relatives to tell them of the new arrival, there have been times when people considered it unlucky to weigh babies [1].

Throughout the nineteenth century, "prematurity" was often cited as a cause of death among babies, but attempts at definition do not appear to have been made before the twentieth century. In 1906, George Newman made a distinction between prematurity and immaturity, but quoted views that prematurity should be defined as having a birthweight under 2500 g or perhaps 3000 g [3].

Arvo Ylppo, a Finnish doctor working in Germany, suggested in 1919, in his review "On the physiology, care and fate of newborn babies", that "premature birth" should be defined as having a birthweight of 2500 g or less. He acknowledged, however that the term *frühgeburt*, or "premature birth", was inappropriate and suggested instead the term *unreifes kind*, or "immature child". Despite the fact that he acknowledged that the cutoff point of 2500 g was arbitrary and not necessarily related to other indicators of immaturity, it was adopted internationally as a definition of "prematurity" [6]. Countries using imperial weights substituted the corresponding weight of $5\frac{1}{2}$ lb.

By the 1970s the limitations of this definition of "prematurity" were becoming increasingly apparent. To distinguish between short gestation and slow fetal growth, separate definitions of "low birthweight" and "preterm" birth (*see* **Gestational Age**) were published in 1977 in the ninth revision of the **International Classification of Diseases (ICD)** [4] and repeated in the tenth revision [5]. At the same time the definition was changed from 2500 g or less to under 2500 g. Because of "digit preference"; that is, the tendency to choose round numbers, the change affected the continuity of time series [2]. WHO recommends that birthweight should preferably be measured within the first hour of life and the actual weight should be recorded to the degree of accuracy to which it is measured [5]. The classifications are as follows.

1. Low birthweight: less than 2500 g; that is, up to and including 2499 g.
2. Very low birthweight: less than 1500 g; that is, up to and including 1499 g.
3. Extremely low birthweight: less than 1000 g; that is, up to 999 g.

The mortality of babies varies considerably according to birthweight, with very high mortality rates among very small babies (*see* **Infant and Perinatal Mortality**). On average, babies from multiple births are lighter than singletons and a higher proportion of them have low birthweights. Average birthweight and the proportion of low-weight births also varies between socioeconomic groups. Among babies born to less favored sections of populations there is a tendency for lower mean birthweights and higher proportions of low-weight births than among babies born in more favored circumstances. There are also differences between **ethnic groups**. Studies using **record linkage** have shown successive babies born to the same women tend to have similar birthweights.

Research done by people concerned with evaluation of maternity care tends to center on the survival rates of low and very low birthweight babies and the health status (*see* **Quality of Life and Health Status**) of the surviving children. Among epidemiologists there is currently a keen interest in associations between the health of adults and their birthweight. The relative importance of circumstances at birth and in later life is hotly debated.

References

[1] Chambers, R. (1866). *The Book of Days. A Miscellany of Popular Antiquities*, Vol. II. W. and R. Chambers, London and Edinburgh.

[2] Macfarlane, A. & Mugford, M. (2000). *Birth Counts: Statistics of Pregnancy and Child-birth*, Vol. 1. 2nd Ed The Stationery Office, London.

[3] Newman, G. (1906). *Infant Mortality*. Methuen, London.

[4] World Health Organization (1977). *International Classification of Diseases. Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*, 9th revision, Vol. 1. WHO, Geneva.

[5] World Health Organization (1992). *International Classification of Diseases and Related Health Problems*, 10th revision, Vol. 1. WHO, Geneva.

[6] Ylppo, A. (1919). Physiologie und zum Schicksal der Frühgeborenen, *Zeitschrift für Kinderheilkunde* **24**, 1–110.

ALISON MACFARLANE

(*See also* **Vital Statistics, Overview**)

Biserial Correlation

Biserial correlation coefficients are measures of bivariate association that arise when one of the observed variables is on a measurement scale and the other variable takes on two values. There are several biserial coefficients, with appropriate choices based on the nature of the underlying bivariate population. Two common forms are the Pearson biserial correlation (hereafter referred to as the biserial correlation) and the point biserial correlation.

Pearson [9] developed the biserial correlation to estimate the product moment **correlation** ρ_{YZ} between two measurements Y and Z using data where Z is not directly observed. Instead of Z, data are collected on a categorical variable X which takes on the values $X = 1$ if Z exceeds a threshold level, and $X = 0$ otherwise. In many applications, the latent variable Z is conceptual rather than observable. The actual values used to code X do not matter, provided the larger value of X is obtained when Z exceeds the threshold. The point biserial correlation is the product moment correlation ρ_{YX} between Y and X.

We use data adapted from the study by Karelitz et al. [7] of 38 infants to illustrate ideas. Table 1 gives a listing of the data. The categorical variable X corresponds to whether the child's speech developmental level at age three is high ($X = 1$) or low ($X = 0$). The child's IQ score at age three is Y. The biserial correlation ρ_{YZ} is a reasonable measure of association when X can be viewed as a surrogate for an underlying continuum Z of speech levels. The point biserial correlation ρ_{YX} might be considered when the scientist is uninterested in the relationship between IQ and the underlying Z scale, or cannot justify the existence of such a scale.

The remainder of this article discusses methods for estimating the point biserial and the biserial correlation. Other forms of biserial correlation are briefly mentioned.

The Point Biserial Correlation

Suppose that a sample $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ is selected from the (Y, X) population. Let s_{YX} be the sample covariance between the $y_i s$ and the $x_i s$, and let s_Y^2 and s_X^2 be the sample variances of the $y_i s$ and the $x_i s$, respectively. The population correlation ρ_{YX} is estimated consistently by the sample point biserial correlation

$$r_{YX} = \frac{s_{YX}}{s_Y s_X} = \frac{(\overline{y}_1 - \overline{y}_0)}{s_Y}[\hat{p}(1-\hat{p})]^{1/2}, \quad (1)$$

where $\overline{y}_1$ and $\overline{y}_0$ are the average y values from sampled pairs having $x_i = 1$ and $x_i = 0$, respectively, and $\hat{p}$ is the observed proportion of pairs with $x_i = 1$.

The sampling distribution of r_{YX} is known only for certain models. Tate [12] derived the distribution of $T = (n-2)^{1/2} r_{YX}/(1 - r_{YX}^2)^{1/2}$ under the assumption that the conditional distributions of Y given $X = 1$ and given $X = 0$ are normal with identical variances. Tate [12] noted that T is equal to the usual two-sample Student's t statistic for comparing the means of the y samples having $x_i = 1$ and $x_i = 0$. The hypothesis $\rho_{YX} = 0$ is usually tested using this standard t test. Tate's [12] results are more complex for testing nonzero values of ρ_{YX}. In large samples, hypothesis tests and confidence intervals for ρ_{YX} can be based on a normal approximation to r_{YX} with mean ρ_{YX} and estimated variance

$$\widehat{\text{var}}(r_{YX}) = \frac{(1 - r_{YX}^2)^2}{n}\left\{1 - 1.5 r_{YX}^2 + \frac{r_{YX}^2}{4\hat{p}(1-\hat{p})}\right\}. \quad (2)$$

Das Gupta [5] generalized Tate's [12] results to nonnormal populations.

For the IQ data, the distributions of the IQ scores for the samples with $X = 0$ and $X = 1$ are slightly skewed to the right and have similar spreads. The mean IQ scores in the two samples are $\overline{y}_1 = 2779/22 = 126.318$ and $\overline{y}_0 = 1676/16 = 104.750$. With $\hat{p} = 22/38 = 0.579$ and $s_Y = 19.383$, we obtain

Table 1 IQ data for a sample of 38 children: X = speech developmental level (0 = low; 1 = high) and Y = IQ score

$X = 0$	Y:	87	90	94	94	97	103	103	104	106	108	109	109
		109	112	119	132								
$X = 1$	Y:	100	103	103	106	112	113	114	114	118	119	120	120
		124	133	135	135	136	141	155	157	159	162		

$r_{YX} = 0.557$, $\widehat{\text{sd}}(r_{YX}) = 0.103$, and $T = 4.024$ under Tate's assumptions.

A limiting feature of the point biserial correlation is that the range of ρ_{YX} is smaller than the usual reference range of -1.0 to 1.0. For example, the magnitude of ρ_{YX} cannot exceed 0.798 when Y is normally distributed. This restriction can lead to misinterpreting the strength of the sample correlation. Shih & Huang [10] examined this problem in a general setting, and offer a useful method to calibrate point biserial correlations.

Pearson's Biserial Correlation and Generalizations

Suppose that X is obtained by **categorizing a continuous variable** Z with $X = 1$ if $Z > \omega$ and $X = 0$ otherwise, where ω is a fixed but possibly unknown threshold. Let $f(t)$ and $F(t)$ be the pdf for Z and the cdf for Z, respectively. We assume without loss of generality that $\text{E}(Z) = 0$ and $\text{var}(Z) = 1$. The threshold ω is the upper pth percentile of Z, i.e. $\omega = F^{-1}(1 - p)$, where

$$p = \Pr(X = 1) = \Pr(Z > \omega) = 1 - F(\omega). \quad (3)$$

If the regression of Y on Z is linear, then the biserial correlation and the point biserial correlation are related by

$$\rho_{YZ} = \rho_{YX} \frac{[p(1-p)]^{1/2}}{\lambda(\omega, F)}, \quad (4)$$

where

$$\lambda(\omega, F) = \text{E}(XZ) = \int_{\omega}^{\infty} t f(t)\, \text{d}t = \int_{\omega}^{\infty} t\, \text{d}F(t). \quad (5)$$

The linear regression assumption is satisfied when (Y, Z) has a bivariate normal distribution, a common assumption, but holds for other elliptically symmetrical bivariate distributions as well.

Eq. (4) provides a way to estimate the biserial correlation from a sample of $(y_i, x_i)s$ when the cdf of Z is known. Bedrick [2] proposed a simple method-of-moments estimator,

$$\tilde{r}_{YZ} = r_{YX} \frac{[\hat{p}(1-\hat{p})]^{1/2}}{\lambda(\hat{\omega}, F)}, \quad (6)$$

where $\hat{\omega} = F^{-1}(1 - \hat{p})$ is the estimated threshold based on the proportion $\hat{p}$ of sampled pairs with $x_i = 1$. If Z is normally distributed, then (6) is Pearson's biserial estimator

$$r_{\text{Pb}} = \frac{r_{YX}}{\phi(\hat{\omega})}[\hat{p}(1-\hat{p})]^{1/2} = \frac{(\overline{y}_1 - \overline{y}_0)}{s_Y \phi(\hat{\omega})}\hat{p}(1-\hat{p}), \quad (7)$$

where $\phi(t)$ is the standard normal pdf.

Bedrick [2] showed that the asymptotic distribution of $\tilde{r}_{YZ}$ is normal with mean ρ_{YZ} and gave an expression for the large sample $\text{var}(\tilde{r}_{YZ})$. In earlier work, Soper [11] gave an estimator for $\text{var}(r_{\text{Pb}})$ when (Y, Z) is normal:

$$\widehat{\text{var}}(r_{\text{Pb}}) = \frac{1}{n}\left\{ r_{\text{Pb}}^4 + \frac{r_{\text{Pb}}^2}{\phi^2(\hat{\omega})}[\hat{p}(1-\hat{p})\hat{\omega}^2 + (2\hat{p}-1)\hat{\omega}\phi(\hat{\omega}) - 2.5\phi^2(\hat{\omega})] + \frac{\hat{p}(1-\hat{p})}{\phi^2(\hat{\omega})} \right\}. \quad (8)$$

Unlike the point biserial estimator, the magnitudes of r_{Pb} and $\tilde{r}_{YZ}$ can exceed 1.0. For the IQ data, $r_{\text{Pb}} = 0.694$ and $\widehat{\text{sd}}(r_{\text{Pb}}) = 0.135$.

Brogden [3] and Lord [8] generalized Pearson's estimator by relaxing the assumption that the distribution of Z is known. Bedrick [1, 2] gave a detailed study of Brogden and Lord's estimators. Cureton [4] and Glass [6] proposed versions of Brogden's estimator that are based on the ranks of the y sample.

As a final point, note that a **maximum likelihood estimator** (MLE) of ρ_{YZ} can be computed iteratively whenever a joint distribution for (Y, Z) can be specified. Tate [13] proposed the MLE of ρ_{YZ} as an alternative to Pearson's biserial estimator with bivariate normal populations. Although MLEs are fully efficient, Bedrick's [1, 2] results show that the asymptotic variances of Lord's estimator and the MLE are often close in normal and nonnormal populations.

References

[1] Bedrick, E.J. (1990). On the large sample distributions of modified sample biserial correlation coefficients, *Psychometrika* **55**, 217–228.

[2] Bedrick, E.J. (1992). A comparison of modified and generalized sample biserial correlation estimators, *Psychometrika* **57**, 183–201.

[3] Brogden, H.E. (1949). A new coefficient: application to biserial correlation and to estimation of selective inefficiency, *Psychometrika* **14**, 169–182.
[4] Cureton, E.E. (1956). Rank-biserial correlation, *Psychometrika* **21**, 287–290.
[5] Das Gupta, S. (1960). Point biserial correlation and its generalization, *Psychometrika* **25**, 393–408.
[6] Glass, G.W. (1966). Note on rank biserial correlation, *Educational and Psychological Measurement* **26**, 623–631.
[7] Karelitz, S., Fisichelli, V.R., Costa, J., Karelitz, R. & Rosenfeld, L. (1964). Relation of crying activity in early infancy to speech and intellectual development at age three years, *Child Development* **35**, 769–777.
[8] Lord, F.M. (1963). Biserial estimates of correlation, *Psychometrika* **28**, 81–85.
[9] Pearson, K. (1909). On a new method of determining the correlation between a measured character A and a character B, *Biometrika* **7**, 96–105.
[10] Shih, W.J. & Huang, W.-H. (1992). Evaluating correlation with proper bounds, *Biometrics* **48**, 1207–1213.
[11] Soper, H.E. (1914). On the probable error for the biserial expression for the correlation coefficient, *Biometrika* **10**, 384–390.
[12] Tate, R.F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation, *Annals of Mathematical Statistics* **25**, 603–607.
[13] Tate, R.F. (1955). The theory of correlation between two continuous variables when one is dichotomized, *Biometrika* **42**, 205–216.

(*See also* **Association, Measures of**; **Pearson, Karl**)

EDWARD J. BEDRICK

Bivariate Distributions

The study of the joint statistical behavior of pairs of random variables gives rise to bivariate distribution theory. For instance, Halperin et al. [1] examine the effect of systolic blood pressure and the number of cigarettes smoked per day on the probability of death. In this case the random variables of interest are the systolic blood pressure and the number of cigarettes smoked per day. While the ideas involved in the study of bivariate distributions follow along the lines of univariate distributions, the mathematical development of the results is more complicated. Three distinct types of bivariate distributions that arise in practical situations are identified and presented in this article.

Continuous Distributions

Random variables (X_1, X_2) are said to have a continuous distribution if and only if $\Pr\{X_i$ lies in the infinitesimal interval $(x_i, x_i + \mathrm{d}x_i)$ for $i = 1, 2\} = f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$, where $f(x_1, x_2)$ is called the probability density function (pdf) of (X_1, X_2) at the point (x_1, x_2) lying in the two-dimensional Euclidean space. Alternatively, it is possible to represent the probability that (X_1, X_2) lies in the two-dimensional region I as $\int\int f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$, with the integration running over the region I. The **moment generating function** (mgf) is defined by $M(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t_1^{x_1} t_2^{x_2} f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$, with its existence requiring that $\{-h_1 < t_1 < h_1, -h_2 < t_2 < h_2\}$ (see Hogg & Craig [2, p. 97]). As in univariate distributions, the mgf has a one-to-one relationship with the pdf. The marginal distributions of X_1 and X_2 have the mgfs given by $M(t_1, 0)$ and $M(0, t_2)$, respectively. A necessary and sufficient condition for the independence of X_1 and X_2 is that $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$. Under the assumption of the existence of the mgf, the **moments** of (X_1, X_2) can be determined as $\mathrm{E}[X_1^r X_2^s] = \mathrm{the}(r, s)\mathrm{th}$ mixed partial derivative of the mgf evaluated with each of the arguments set equal to zero. The inversion of the mgf is mathematically complicated, involving, in most instances, the use of transform theory. However, in many practical problems it is of a recognizable form leading to a simple way for its inversion. The mgf can be used to find the distributions of functions of (X_1, X_2). Thus, for example, for a random sample of size n the mgf of $\left(\sum_{i=1}^{i=n} X_{1i}, \sum_{i=1}^{i=n} X_{2i}\right)$ is $[M(t_1, t_2)]^n$, while that of $\{\overline{X}_1, \overline{X}_2\}$ is $[M(t_1/n, t_2/n)]^n$. At this point we introduce a few special types of continuous distributions:

1. The **bivariate normal distribution**. The standard form of the bivariate normal distribution is defined by the pdf $f(z_1, z_2) = c\exp\{-[z_1^2 - 2\rho z_1 z_2 + z_2^2]/2(1-\rho^2)]\}$, where the constant $c = 1/[2\pi(1-\rho^2)]^{1/2}$. For this form of distribution, the mean of each of the random variables is zero, the variances are each equal to one, with the coefficient of **correlation** ρ. A more general form of the distribution arises when we replace z_i by $(x_i - \mu_i)/\sigma_i$, $i = 1, 2$, and c by $c/\sigma_1\sigma_2$. It can shown that the means in this case are μ_i and the standard deviations are σ_i for $i = 1, 2$. The coefficient of correlation, being location and scale-free, remains at ρ. The distribution enjoys the regenerative property in that if (X_{1i}, X_{2i}) are n independent random variables having a standard bivariate normal distribution, then $\left[U = \sum_{i=1}^{n} X_{1i}/\sqrt{n}, V = \sum_{i=1}^{n} X_{2i}/\sqrt{n}\right]$ also have the standard bivariate normal distribution. Similar results can be established for the general bivariate normal distribution. In each case the marginal distributions are univariate normals. The bivariate normal distribution finds extensive applicability in studies involving two characteristics. Hutchinson & Lai [3, Chapter 19] have listed a wide variety of applications to medical research of the bivariate normal distribution and its modifications.
2. The **exponential distribution**. While a variety of forms of this distribution have been suggested in the literature, their applicability is, in all cases, directed toward a study of the reliability of competing systems. A widely used form is that of Marshall & Olkin [5]. Unfortunately, the pdf is quite complicated and as such will not be reproduced here. The cumulative distribution function is given by $F(x_1, x_2) = \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_3 \max(x_1, x_2)]$. An interesting application of the distribution in the medical context is given by Rai & Van Ryzin [6]. They consider in a **quantal response** context, the tolerance distribution for the occurrence of bladder and liver tumors as a consequence of exposure to various

levels of a carcinogen (see Hutchinson & Lai [3, Chapter 9]).

Discrete Distributions

Random variables (X_1, X_2) are said to have a discrete distribution if and only if the probability of the event $\{X_1 = x_1 \text{ and } X_2 = x_2\}$ is nonzero and is equal to $f(x_1, x_2)$, the probability function (pf) at (x_1, x_2), where x_1 and x_2 each assumes values over a finite set or a countable infinity of points in the Euclidean space. However, usually, the pf is taken to be positive over the non-negative integer values of x_1 and x_2. In most discrete distributions the probability **generating function** (pgf) is preferable because it has a simpler form than the mgf. The pgf is given by $\prod(t_1, t_2) = \sum_{x_2=0} \sum_{x_1=0} t_1^{x_1} t_2^{x_2} f(x_1, x_2)$, which can be readily seen to exist for all values of each of t_1 and t_2 in the interval $[-1, 1]$. There is a one-to-one relationship between the pgf and the pf. Thus, the pf at (r, s) can be determined from the pgf either as the ratio $\{[(r, s)\text{th mixed partial derivative of the pgf evaluated at } t_1 = t_2 = 0]/r!s!\}$ or as the coefficient of $t_1^{x_1} t_2^{x_2}$ in an expansion of the pgf in powers of t_1 and t_2. In addition, it is possible to determine, $\mathrm{E}[x_1^{(r)} x_2^{(s)}]$, the factorial moment (*see* **Moments**) of order (r, s), from the same mixed partial by setting $t_1 = t_2 = 1$. The marginal pgfs are given by $\prod(t_1, 1)$ and $\prod(1, t_2)$, respectively. A necessary and sufficient condition for independence is $\prod(t_1, t_2) = \prod(t_1, 1) \prod(1, t_2)$. Two of the most commonly occurring discrete distributions are the bivariate **Poisson** and the bivariate Neyman type A (*see* **Accident Proneness**). Reference may be made to Kocherlakota & Kocherlakota [4] for a discussion of the bivariate discrete distributions and their various forms.

Mixed Distributions

Although in most situations both of the random variables have distributions that are similar in form, in some instances one of the random variables could have a continuous distribution while the other is of a discrete form. The example cited at the beginning of this article typifies this situation. In this case the blood pressure has a continuous distribution while the number of cigarettes smoked per day is a discrete random variable. In situations of this type we may be particularly concerned with the **regression** of the continuous variable on the discrete variable.

References

[1] Halperin, M., Wu, M. & Gordon, T. (1979). Genesis and interpretation of differences in distribution of baseline characteristics between cases and non-cases in cohort studies, *Journal of Chronic Diseases* **32**, 483–491.

[2] Hogg, R.V. & Craig, A.T. (1995). *Introduction to Mathematical Statistics*, 5th Ed. Prentice-Hall, Princeton.

[3] Hutchinson, T.P. & Lai, C.D. (1990). *Continuous Bivariate Distributions, Emphasizing Applications*. Rumsby Scientific Press, Adelaide.

[4] Kocherlakota, S. & Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Marcel Dekker, New York.

[5] Marshall, A.W. & Olkin, I. (1967). A multivariate exponential distribution, *Journal of the American Statistical Association* **62**, 30–44.

[6] Rai, K. & Van Ryzin, J. (1984). Multihit models for bivariate quantal responses, *Statistics and Decisions* **2**, 111–129.

SUBRAHMANIAM KOCHERLAKOTA

Bivariate Normal Distribution

The bivariate normal distribution of the random variables X and Y has the joint density function

$$\begin{aligned}\phi(x, y) = {} & [2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}]^{-1} \\ & \times \exp\{-1/2[(x-\mu_1)^2/\sigma_1^2 \\ & - 2\rho[(x-\mu_1)/\sigma_1][(y-\mu_2)/\sigma_2] \\ & + (y-\mu_2)^2/\sigma_2^2]/(1-\rho^2)\},\end{aligned}$$

for x and y defined over the entire plane $-\infty < x < \infty$ and $-\infty < y < \infty$. The five parameters determining the location, dispersion, and orientation of the bivariate normal probability surface are the means $\mathrm{E}(X) = \mu_1$ and $\mathrm{E}(Y) = \mu_2$, the variances $\mathrm{var}(X) = \sigma_1^2$ and $\mathrm{var}(Y) = \sigma_2^2$, and the **correlation** $\mathrm{corr}(X, Y) = \rho$. For the density to be defined it is necessary that $-1 < \rho < 1$. Otherwise, when $\rho = \pm 1$ the distribution is *singular*, and the density function does not exist. The density is unimodal, with its peak at $x = \mu_1$ and $y = \mu_2$. The contours of planes through the density at constant elevations are elliptical, since each has the form

$$\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} = \text{constant}.$$

The random variable defined by the quadratic form has the **chi-squared distribution** with two degrees of freedom. When $\rho > 0$ the major axes of the ellipses have positive slopes in the (X, Y) plane, and negative orientation if $\rho < 0$. If $\rho = 0$ the axes of the ellipses are parallel with the coordinate axes of X and Y. When $\rho = 0$ and $\sigma_1^2 = \sigma_2^2$, the concentration ellipses are circular.

The transformations $Z_1 = (X - \mu_1)/\sigma_1$ and $Z_2 = (Y - \mu_2)/\sigma_2$ give the standardized bivariate normal distributions with zero means, unit variances, and the single parameter ρ. The **orthogonal** transformation of the standardized variates,

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix},$$

leads to the independent random variables U_1 and U_2, regardless of the value of ρ. If $\rho > 0$, U_1 corresponds to the major axis of the concentration ellipse of the density of Z_1 and Z_2, while U_2 is the minor axis variate.

The bivariate normal distribution arises from the **central limit theory** for a sequence of independent pairs of correlated random variables. It was first proposed in 1808 by Adrain [1] in the "circular" case of equal variances and zero correlation. **Laplace** [5] published a general normal density expression in 1812, based on his earlier work on the central limit theorem. Some 70 years later, **Galton** [3] noticed a pattern of concentric ellipses in tables of bivariate data. His request of the mathematician J.D.H. Dickson [2] for a distribution with that property led to the bivariate normal density.

Properties

The random variables X and Y are independent if and only if $\rho = 0$. The marginal distributions of X and Y are univariate normal with means and variances as in the bivariate distribution. The *conditional* distribution of Y for $X = x$ is also normal, with mean $\mathrm{E}(Y|X) = \mu_2 + (\rho\sigma_2/\sigma_1)(x - \mu_1)$ and variance $\mathrm{var}(Y|x) = \sigma_2^2(1 - \rho^2)$. The conditional mean, or *regression function* (*see* **Regression**), is linear in the values x of the fixed variable. Furthermore, the conditional variance is constant for all x. Those properties are the basis for regression analysis when both variables are random and are jointly bivariate normally distributed.

Johnson & Kotz [4] have given an extensive treatment of the history, properties, and inferential aspects of the bivariate normal distribution. They include several plots of the bivariate normal density surface for different values of ρ.

Probability Calculations

Tables were published by the National Bureau of Standards [7] of the probabilities

$$\begin{aligned}L(h, k, \rho) &= \int_h^\infty \int_k^\infty \phi(x, y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \Pr[(X > h) \cap (Y > k)]\end{aligned}$$

for bivariate normal (X, Y). Another related function $V(h, k)$ for computing probabilities over triangular and polygonal regions is also tabulated. Owen [8] published tables of a slightly different version of $V(h, k)$. Zelen & Severo [9, 10] expressed $L(h, k, \rho)$ in terms of $L(h, 0, \rho)$, and provided charts of the latter function. Mehta & Patel [6] have developed statistical computer software that will calculate the probability function $L(h, k, \rho)$.

References

[1] Adrain, R. (1808). Research concerning the probabilities of errors which happen in making observations, etc., *The Analyst; or Mathematical Museum* **1**, 93–109.

[2] Dickson, J.D.H. (1886). Appendix to "Family likeness in stature", by F. Galton, *Proceedings of the Royal Society of London* **40**, 63–73.

[3] Galton, F. (1889). *Natural Inheritance*. Macmillan, London.

[4] Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

[5] Laplace, P.S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

[6] Mehta, C. & Patel, N.R. (1994). *StaTable: Electronic Tables for Statisticians and Engineers*. Cytel Software Corporation, Cambridge, Mass.

[7] National Bureau of Standards (1959). *Tables of the Bivariate Normal Distribution Function and Related Functions*. US Government Printing Office, Washington.

[8] Owen, D.B. (1956). Tables for computing bivariate normal probabilities, *Annals of Mathematical Statistics* **27**, 1075–1090.

[9] Zelen, M. & Severo, N.C. (1960). Graphs for bivariate normal probabilities, *Annals of Mathematical Statistics* **31**, 619–624.

[10] Zelen, M. & Severo, N.C. (1964). Probability functions, in *Handbook of Mathematical Functions*, M.Abramowitz & I.Stegun, eds. National Bureau of Standards, Washington, pp. 936–940.

(*See also* **Multivariate Normal Distribution**; **Normal Distribution**)

DONALD F. MORRISON

Blinding or Masking

The term *blinding* (sometimes *masking*) applies primarily to **clinical trials** but is also commonly and increasingly used with reference to **analytic epidemiologic** studies. Most explanations of the term are in the context of clinical trials, and the majority of this article is focused towards trials, although various sections also describe blinding epidemiologic studies.

The fundamental idea in blinding is that the study patients, the people involved with their management, and those collecting the clinical data from studies should not be influenced by knowledge of the assigned treatment or, in an epidemiologic study, by knowledge of the main risk factors or outcomes. For example, if the factor being investigated is a treatment in a prospective trial, then neither patients, their physicians, nor those assessing their medical condition should know which treatment any particular patient is or has been receiving. It is not sufficient to argue that any individual patient cannot know the treatment identity because they only see their own medication and cannot compare it with medication given to other patients. If this were the case, a study comparing red tablets with green capsules could be described as patient blind. It is true that a patient given red tablets would not know that other patients may be given green capsules but this would not be sufficient to describe a study as blinded. As another example, if the factor being investigated is exposure to an environmental toxin in a **case–control study**, then whether a subject is a case or a control should be unknown to those who are collecting the data on exposure.

Reasons for Blinding

Patient Bias

Blinding patients to which treatment they have received in a clinical trial is particularly important when many of the parameters are subjective both in patient response and with more formal clinical assessment of response. One reason for this is described as the "placebo effect". It is a rather broad and ill-defined term encompassing a variety of responses that occur when patients are being "treated" with inactive placebo medication that, theoretically, should have no therapeutic impact. Use of the term is so broad that it covers psychological responses that should be expected as well as physical ones that should not. Various studies have described how placebos can give both positive (therapeutic benefit) and negative (adverse events) effects. Moscucci et al. [13] describe how, in an obesity trial of active medication (phenylpropanolamine) vs. placebo, some of the patients randomized to placebo indicated that they had improved control of their appetite. One area prone to such effects is antiemetic (control of nausea and vomiting) studies where the main outcome can be strongly influenced by psychological processes. Another obvious example is trials in **psychiatry**. Even in therapeutic areas where assessments are more objective, Schulz et al. [18] have reported that trials that are not double-blinded are more likely than blinded studies to show benefit (falsely) for the active intervention group.

Physician Bias in Clinical Management

Where possible, the managing physician should be blinded to prevent any possible bias in patient management. For example, the decision to withdraw patients from a study could be influenced by knowledge of which treatment they are receiving. If there has been a poor response and it is known that the patient has been assigned to placebo, then there could be a greater tendency to withdraw the patient. If adverse events are present and it is known that the patient has been assigned to active therapy, then a similar increased tendency to withdraw the patient may exist.

Another aspect of patient management that could be influenced by knowledge of which treatment a patient is receiving is the decision regarding dose adjustment. Similarly, a general problem exists in the comparison of complementary (or **"alternative") medicine** with other complementary medicines or with "conventional" medicine. Anthony [2] highlights a fundamental problem in this area of research, that often complementary medicine, properly administered, is designed individually for each patient, so that blinding the managing physician is impossible. In such cases it is best to ensure that the person assessing the response is blinded.

Bias in Evaluation

Whenever there is subjective judgment in evaluating clinical response, it is preferable to blind the person making the evaluation. In the field of periodontal trials, Imrey & Chilton [9] describe the need for blinding, because blinded evaluation greatly increases credibility and usually only marginally increases costs.

In guidelines for trials in scleroderma, White et al. [22] briefly but firmly state the need for blinding of clinical evaluations. They state that "assessments of global functioning and functional disability cannot be assured to be unbiased". Their terminology (the words *"cannot be assured"*) is pertinent since unblinded assessments do not necessarily lead to biased assessments, but the lack of blinding means that a lack of **bias** cannot be assumed: credibility and reliability are both compromised.

Bias in Data Management

Even away from the patient and clinic environment, bias is still possible and is a potentially important source of error within the various stages of **data management**. Stages include coding of adverse events, interpretation of ambiguous handwriting, and decisions on whether or not to query unlikely (but possibly correct) data values. Beyond the initial aspects of data management, bias can be introduced in the choice of statistical methods that are used or presented (or even in the order in which they are presented) and, indeed, in the choice of style and content of presentation of the data, regardless of any formal statistical methodology that may be applied. For these reasons, some statisticians consider that every detail of data presentation and analysis should be specified before the database is unblinded. Others disagree and argue that the most effective presentation and analysis requires complete knowledge and understanding of the data. This, necessarily, includes knowledge of the treatment groups.

Bias in Decisions Regarding Stopping a Trial

If interim analyses are planned and a data monitoring committee (*see* **Data and Safety Monitoring**) is to review accumulating trial results with a view to recommending continuing or stopping a study, Rockhold & Enas [15] suggest that the committee be presented with results in the form of "treatment A" and "treatment B" without revealing which is which. If more than one set of results is presented to the monitoring committee, some might suggest that the labels A and B are not necessarily kept the same across all sets of results so as to try to prevent accumulating evidence across a variety of efficacy and safety parameters from giving clues as to the treatment identity. A completely opposite view is taken by others, who argue that the welfare of the patients is more important, and that the data monitoring committee should be unblinded so as to make properly informed decisions.

Whichever way the data monitoring committee is presented with the data, it is important that other staff (particularly investigators) are still kept blind. In particular, the blinding should be maintained for each patient until all patients have completed the study: then the blind can be broken for all patients. This is often not liked in cases in which the first patient may finish the study several years before the last patient: still, the first patient's treatment allocation should not be revealed until all patients have completed the study. There are many reasons to justify this stance. If the blind were broken as each patient completed the study, then the response to treatment (the size of the treatment effect) could begin to emerge; many informal interim analyses might be carried out; some investigators might decide not to continue in the study based on their opinions of early results; some investigators may (consciously or subconsciously) change the type of patient they recruit to the study, and so on.

Bias in Treatment Allocation

Blinding is important in allocating treatments to trial subjects (*see* **Randomized Treatment Assignment**) since **randomization** alone will not necessarily ensure that different treatment groups are balanced at baseline. Spriet & Dupin-Spriet [20] illustrate this, showing how randomization may fail if blinding is not in place, and give suggestions for how to overcome the problem. Schemes such as alternate allocation have been suggested as adequate alternatives to true randomization, but the nonblinded (and therefore open to bias) mechanism of such a method argues against its use. A distinction has sometimes been drawn between *concealment* up to the point of allocation of treatment and *blinding* (or masking) for

steps taken to conceal group identity after allocation [17]. This is a helpful distinction, particularly in single-blind studies: total concealment may be possible, even if total blinding is not.

Bias in Reviewing Studies

In reviewing studies (particularly if the objective is to carry out a formal **meta-analysis**), the procedures for abstracting data should be highly standardized to help to eliminate bias. If the abstractor is to be blinded to, for example, the journal of publication or to the results when abstracting data on the methods, the method for ensuring blinding should be agreed beforehand and described in a written protocol. In a different context, Jadad et al. [10] conducted a randomized controlled study to investigate the impact of blinding on peer-review. They found that blinded assessments produced consistently lower (worse) and less variable ratings than unblinded assessments (*see* **Statistical Review for Medical Journals**).

Bias in Epidemiology

In data collection, Rose et al. [16] were concerned about the potential for bias in assessing blood pressure in case–control studies, where the assessor may know if a patient is a "case" or a "control". Blood pressure has a high measurement error due to factors including systematic observer bias, terminal digit preference and observer preference. To overcome these, they describe a modification of the standard sphygmomanometer that has a "random zero". The random zero sphygmomanometer is now standard in epidemiologic studies involving blood pressure measurements. Generally speaking, when exposure assessments have some subjective element, it seems sensible to use blinded assessment in case–control studies (*see* **Bias in Case–Control Studies**).

Levels of Blinding

The term "double-blinding" is widely accepted and understood to mean that both the patients and the treating physicians are unaware of which treatment is being used. Here we identify four levels of blinding and describe each. The terminology for all but double-blind is not standard, and in these cases, additional detail should be provided to explain what is meant in any particular setting.

Single-Blind

Many authors describe this as meaning that the patient is unaware of which treatment he or she is receiving. However, we may also use the term single-blind when the treating physician is unaware of the treatment assignment. This can happen if the study medications are different in appearance and it is arranged that the patient collects the medication from a pharmacy rather than from the treating physician. Neither interpretation of single-blind is wrong and for clarity the term ought, generally, to be stated as "single (patient) blind" or "single (investigator) blind".

Double-Blind

This is a very common term and is widely accepted to mean that the patient and the investigator are each blinded to the treatment allocation. The investigator is generally assumed to be both the provider (or at least the prescriber) of the medication and the assessor of its effect. When this is not the case, further levels of blinding as described below may be appropriate.

Triple-Blind

This is a less common term but its use is increasing. It is generally accepted to mean that, in addition to the patient and investigator each being blinded, those handling the data are also kept blinded until all decisions about data validity and classification have been made. Marginal decisions regarding patient eligibility or assessment are then made (and are seen to be made), independently of knowing the treatment assignment.

It is, of course, possible that the clinical aspects of a study could be carried out single (patient or investigator)-blind or even unblinded but that the **data management** is performed blinded. Strictly speaking, the study should then be called double-blind or single-blind, but not in the context of the interpretation of those terms given above. This highlights the importance of specifying exactly who was blinded and at what stages in the study.

Quadruple-Blind

This is a very unusual term but Chalmers et al. [4] use it to describe the situation where the patient, the

treating physician, the physician (or other person) assessing the response, and the data handlers are all kept blinded. The additional blinding is relevant when the person assessing the clinical outcome is different from the one who administers the treatment. For example, an oncologist may request a radiologist to assess an X-ray, or a general physician may request a swab sample to be assessed by a microbiologist.

As suggested above, the large number of different parties who may be blinded in a study necessitates one clearly explaining what is meant in the description of blinding, rather than relying on a simple phrase such as "single-blind study".

Methods of Blinding

Placebos

Placebos are commonly used in comparative controlled clinical trials. They are chemically inert compounds that closely resemble the active compound in all physical characteristics such as taste, smell, and appearance. The term "placebo" is also used more broadly to include medical procedures that are still "inert" but that also resemble the true procedure. In such cases, the term "sham" is often used instead of placebo (see later in this Section). One instance where a compound is used that is not a true placebo (that is, it does have some active ingredient) but takes the place of a placebo is in studies of topical skin preparations. In these studies, the term "vehicle" is sometimes used. This vehicle is the base compound used as a delivery mechanism. The base compound might have therapeutic benefits such as soothing and moisturizing or it may have adverse effects such as stinging or irritation. By using the vehicle as a comparator, rather than a pure placebo, the true therapeutic effect of the pharmaceutic compound can be measured over and above that of the base (delivery) compound.

Placebos for Comparison of Active Drugs

Placebos are not usually used when comparing two active compounds. However, it may be difficult to manufacture two active compounds so that they appear and taste identical. Prozac® (fluoxetine hydrochloride) and Haldol® (haloperidol) may, for example, be compared for their relative efficacy in treating obsessive compulsive disorder. The former is normally presented in a green and white enteric coated capsule; the latter is a pale blue tablet. The difference in appearance is the first and perhaps most frequent problem faced in blinding. There are two common approaches to solving the problem.

The first is to disguise the presentation of one or both medications. It may be possible to fit the Haldol tablets (either whole or dissected) into an enteric coating so that they resemble capsules of Prozac. An immediate problem is that the efficacy of "disguised" Haldol may have to be compared with that of "true" Haldol to demonstrate that no harmful (or beneficial) effect has been introduced by changing its presentation. Studies to confirm that the changed presentation of a medication has not affected its potency are known as **bioequivalence** studies. These have their own special blinding problems and are discussed later.

The second approach, often easier, is to use a "double-dummy" method whereby each patient takes both a tablet and a capsule. Patients assigned Haldol tablets also receive a placebo capsule that looks and tastes like a Prozac capsule; patients assigned Prozac capsules also receive a placebo tablet that looks and tastes like a Haldol tablet. Every patient, therefore, receives both capsules and tablets but no patient knows whether they are receiving active tablets or active capsules.

Sham Procedures

The blinding of physical treatments is often difficult. Deyo et al. [6] give an example of treating chronic low back pain using transcutaneous electrical nerve stimulation (TENS), and using "dummy" TENS units. This appeared to have been less successful for blinding the patient than for blinding the investigators. The study evaluated 125 randomized patients. Clinicians guessed the allocation correctly in 61% of cases (only just better than the chance value of 50%); but patients randomized to TENS all guessed that their units were functioning, while most (84%) patients randomized to dummy TENS guessed that their units were not functioning.

Also difficult is the blinding of surgical procedures, particularly the blinded comparison of surgical vs. nonsurgical procedures. The ethics of sham operations are clearly questionable and could rarely, if ever, be justified. It is worth noting, however, that

sham surgery has been used as a placebo and was reported by Cobb et al. [5]. The procedure was for internal mammary artery ligation to treat angina pectoris, and patients were randomized to the full surgical procedure or to a sham surgical procedure that ended at the point immediately after making the initial incision. Patients were informed that they were taking part in an evaluation of the surgical procedure but were not informed of its double-blind nature. Randomization was after the incision had been made (that is, as late as possible), and both groups were (therefore!) subjected to anesthesia (although it was a local – not general – anesthetic). In all respects, therefore, patients were blinded to the randomization scheme. The study was small (only 17 patients) and concluded no effect of ligation over that of the sham operation in terms of exercise tolerance.

It may be less difficult to blind patients in trials of alternative surgical procedures, and blinded assessment of the patient could also be arranged. Blinding surgeons is clearly impossible, and conflicting results of similar trials comparing early vs. delayed surgery for acute cholecystitis, reported by van der Linden & Sunzel [21] and Lahtinen et al. [12], have been attributed to the lack of blinding of the surgeons.

Blinding of Evaluators

We have already mentioned under "Levels of Blinding" that, when it is not possible to blind the treating physician, a blinded evaluator may be used to record patients' responses to treatment. Such procedures are particularly applicable when the assessment has a subjective element and when the investigator is likely to recall the treatment given to the patient. Examples include: comparing alternative surgical procedures, since a surgeon often recalls which patients received which procedure; comparing alternative counseling procedures for patients suffering from post-traumatic shock; and comparing alternative instructional programs for patient self-care (in dental hygiene, for example).

Difficulties of Blinding

Drawbacks of Blinding

The advantages of blinding (generally relating to the "fairness" or "lack of bias" of patient assessments) are important, but blinding is not without disadvantages. Drawbacks include: the practical difficulty of formulating placebos (see the example in the section on "Placebos for Comparison of Active Drugs"); the dangers associated with emergency situations in which someone, possibly having no connection with the trial, may need to know which treatment a patient is taking; and the ethical problems of assigning (some) patients placebo while they remain ignorant of whether they are receiving the experimental or the inert compound. Allen [1] argues that "double-blinding of drug trials to prevent bias deprives physicians of information they need in order to comply with their duty to treat patients and do them no harm", and claims that a physician cannot properly treat a patient without knowing what treatment the patient is already receiving (*see* **Ethics of Randomized Trials**). However, see below the Section entitled "Other Issues" regarding arrangements for breaking the blinding when clinically necessary.

Ineffectiveness

Reference has already been made to some assessment of how well blinding actually worked in specific studies. Ney [14] strongly challenges the effectiveness of blinding. He reports that in a review of clinical trial reports published over a 10-year period, in fewer than 5% of studies described as double-blind was the blinding actually checked. He states that, "In most instances where they were checked they were found not to be blind, and in many instances when they were not checked, they could not have been blind".

Other authors have also challenged the effectiveness of blinding. Greenberg & Fisher [7], for example, discuss trials of antidepressants. Many antidepressants induce dry mouth or constipation and when compared to placebo, the question of which patients are on which treatment arm may not be difficult to answer. They draw the extreme (and perhaps debatable) conclusion that "in the main, all past studies of antidepressant effectiveness are open to question". Known side-effect profiles of other drugs in other therapeutic areas may create similar problems.

The difficulty of genuinely matching treatments in trials intended to be double-blind is described in a study by Hill et al. [8], who used a panel of four observers, each to assess 22 pairs of agents that had been used in double-blind trials. In only five pairs was the match described as "excellent", and in seven there were obvious differences that were detectable

by all four observers. Color and taste were the most frequent causes of mismatching.

Ironically, it may be most difficult to maintain blinding in comparative trials that show the largest clinical effects. Even if that is true, the possibility of the blind being broken during the study does not necessarily justify abandoning plans to blind the study at the outset.

Other types of ineffective blinding have also been described. Senn [19] considers comparing two transdermal patches, A and B. The patches are different in appearance so that placebo patches for each (PA for placebo A and PB for placebo B, respectively) need to be prepared. A three-arm double-dummy study may be envisaged with randomization to A + PB or B + PA or PA + PB. However, to circumvent the practical difficulty of patients wearing two patches, an alternative four-arm randomization could be to A or B or PA or PB. Now, although a patient (or observer) will not know if the patch A or PA is active or placebo, they will know that they are not receiving medication B. Similarly, patients using patches B or PB will know they have not been randomized to A. Thus, this scheme will not necessarily allow an unbiased comparison of A with B.

James et al. [11] have approached the assessment of the effectiveness of blinding from a methodologic point of view and arrive at an index of success of blinding ranging from 0 (complete lack of blinding) to 1 (complete blinding). Subjects are asked to guess whether the treatment was active or placebo. Unlike the common **kappa** statistic, their method incorporates the "don't knows" as well as the correct and incorrect guesses. The "don't knows" are, of course, the ideal response, implying complete success of the blinding. Like the kappa statistic, their method allows assessment of whether the knowledge of the treatment in any given study is significantly greater than chance. Even if "statistical significance" is demonstrated, it is not always easy to interpret the meaning of the index's value; for example, it is unclear if a value of 0.7, say, represents a good or a poor level of blinding.

Other Issues

Arrangements for Breaking Blinding

It is necessary to make arrangements for breaking the blind for any particular patient. When a patient experiences an adverse event that may be medication-related, or a patient requires medication for a concurrent illness where drug interaction is possible, the blind may need to be broken. Occasionally the situation may constitute an emergency. Trials in hospitalized patients generally pose fewer problems since the hospital's own pharmacy is likely to be dispensing all medication and would itself hold a master copy of the randomization codes. In an outpatient setting there is a greater danger that immediate access to the randomization codes might not be possible. In such cases, all patients participating in the trial might carry an identifying card with them that gives details of the trial they are in, the group organizing it (whether that be a hospital, academic institution, or a pharmaceutical company), and an emergency 24-hour telephone number. An alternative arrangement is to provide a tear-off label on the patient's medication that reveals details of the true identity of the medication (and further contact details). This ensures that immediate identification of medication is possible but also allows patients (or others) to break the code for nonessential reasons.

Regulatory Issues

Guidelines within the regulated pharmaceutical industry clearly describe the requirement for blinding. The US **Food and Drug Administration (FDA)** requires that the specific procedures for blinding should be included in clinical study reports, including a description of the labeling on bottles, the appearance, shape, smell, and taste of different medications, and the circumstances in which the blind may be broken (and by whom). They comment on what special precautions should be taken if blinding cannot be achieved and require justification from the sponsor if blinding is deemed unnecessary. The International Conference on Harmonization (ICH) mirrors quite closely the comments of the FDA. In Europe, the Committee on Proprietary Medicinal Products (CPMP) makes similar comments but also addresses some different aspects of the problem. They refer to the possible influence of the "attitudes of patients to the treatments". By this they mean that, although double-blinding is the ideal approach, if this involves double dummy methods, then the administration scheme may be sufficiently different from clinical practice to influence patient motivation and compliance. In such a case, double-blinding

may not necessarily be optimal and some compromise between the level of blinding and similarity of the dosage regimen to be used in routine practice may be preferable (*see* **Drug Approval and Regulation**).

Reporting Blinding

Bailar & Mosteller [3] suggest that a "statement that a study was 'blind' or 'double-blind' is rarely enough". Details of the methods of blinding should be given along with an assessment of the effectiveness (or otherwise) of the blinding. Chalmers et al. [4] have reported on how to assess the quality of clinical trials and consider that blinding ("quadruple-blinding", as described earlier) is the most important feature. "It is not sufficient", they state, "to assume that a double-blind procedure is effective." They propose that the effectiveness of the blinding should always be investigated and reported.

Equivalence Trials

Blinding fails to protect against bias in trials aiming to test whether two treatments are equivalent (*see* **Equivalence Trials**). The same problem applies to a study that fails to demonstrate a treatment difference and, retrospectively, consideration moves to whether the treatments could be considered (reliably) to have a similar therapeutic effect. The problem is that blinding is intended to avoid a false conclusion that one treatment is superior to the other when in fact they are equivalent. However, it does not adequately protect against falsely concluding that the two treatments are equivalent when in fact one is superior. This is because such bias can be introduced by factors such as treatment **noncompliance** and **measurement error** even when they apply equally to the two treatment groups.

Psychological Resistance

The objective of blinding is to contribute to the elimination of bias that may be introduced intentionally or unintentionally by the many individuals involved in a research project. Some take the view that they need not be blinded. Physicians treating a patient may claim that their assessment of the disease is not influenced by knowing which treatment a patient has received; patients may feel that their reactions to the treatment are not colored by knowledge of which treatment they have been given. Insistence on blinding may be seen as casting doubt on the integrity of these individuals. This attitude sometimes presents an obstacle. Blinding makes it more difficult to bias results and helps to ensure the credibility of the results of a study. For this reason, efforts to overcome psychological resistance to blinding are worthwhile.

References

[1] Allen, A.D. (1989). Making moral decisions on a double-blinded drug trial in progress without breaking the code: a primer on posterior analysis, *Medical Decision Making* **9**, 207–216.

[2] Anthony, H.M. (1987). Some methodological problems in the assessment of complementary therapy, *Statistics in Medicine* **6**, 761–771.

[3] Bailar, J.C. & Mosteller, F. (1992). *Medical Uses of Statistics*, 2nd Ed. New England Journal of Medicine Books, Boston.

[4] Chalmers, T.C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial, *Controlled Clinical Trials* **2**, 31–49.

[5] Cobb, L.A., Thomas, G.I., Dillard, D.H., Merendino, K.A. & Bruce, R.A. (1959). An evaluation of internal-mammary-artery ligation by a double-blind technic, *New England Journal of Medicine* **260**, 1115–1118.

[6] Deyo, R.A., Walsh, N.E., Schoenfeld, L.S. & Ramamurthy, S. (1990). Can trials of physical treatments be blinded? The example of transcutaneous electrical nerve stimulation for chronic pain, *American Journal of Physical Medicine and Rehabilitation* **69**, 6–10.

[7] Greenberg, R.P. & Fisher, S. (1994). Suspended judgement-Seeing through the double-masked design: a commentary, *Controlled Clinical Trials* **15**, 244–246.

[8] Hill, L.E., Nunn, A.J. & Fox, W. (1976). Matching quality of agents employed in "double-blind" controlled clinical trials, *Lancet* **1**, 352–356.

[9] Imrey, P.B. & Chilton, N.W. (1992). Design and analytic concepts for periodontal clinical trials, *Journal of Periodontology* **63**, 1124–1140.

[10] Jadad, A.R., Moore, R.A., Carroll, D., Jenkinson, C., Reynolds, D.J.M., Gavaghan, D.J. & McQuay, H.J. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary?, *Controlled Clinical Trials* **17**, 1–12.

[11] James, K.E., Bloch, D.A., Lee, K.K., Kraemer, H.C. & Fuller, R.K. (1996). An index for assessing blindness in a multi-centre clinical trial: disulfiram for alcohol cessation – a VA cooperative study, *Statistics in Medicine* **15**, 1421–1434.

[12] Lahtinen, J., Alhava, E.M. & Aukee, S. (1978). Acute cholecystitis treated by early and delayed surgery. A

controlled clinical trial, *Scandinavian Journal of Gastroenterology* **13**, 673–678.

[13] Moscucci, M., Byrne, L., Weintraub, M. & Cox, C. (1987). Blinding, unblinding, and the placebo effect: an analysis of patients' guesses of treatment assignment in a double-blind clinical trial, *Clinical Pharmacology and Therapeutics* **41**, 259–265.

[14] Ney, P.G. (1989). Double-blinding in clinical trials [letter], *Canadian Medical Association Journal* **140**, 15.

[15] Rockhold, F.W. & Enas, G.G. (1992). Practical approaches to the design and conduct of interim analyses, in *Biopharmaceutical Sequential Statistical Applications*, K.E. Peace, ed. Marcel Dekker, New York, pp. 19–28.

[16] Rose, G.A., Holland, W.W. & Crowley, E.A. (1964). A sphygmomanometer for epidemiologists, *Lancet* **1**, 296–300.

[17] Schulz, K.F., Chalmers, I., Grimes, D.A. & Altman, D.G. (1994). Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals, *Journal of the American Medical Association* **272**, 125–128.

[18] Schulz, K.F., Chalmers, I., Hayes, R.J. & Altman, D.G. (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association* **273**, 408–412.

[19] Senn, S. (1995). A personal view of some controversies in allocating treatment to patients in clinical trials, *Statistics in Medicine* **14**, 2661–2674.

[20] Spriet, A. & Dupin-Spriet, T. (1994). Allocation to treatment groups in randomized open trials, *Journal of Pharmaceutical Medicine* **4**, 1–5.

[21] van der Linden, W. & Sunzel, H. (1970). Early versus delayed operation for acute cholecystitis, *American Journal of Surgery* **120**, 7–13.

[22] White, B., Bauer, E.A., Goldsmith, L.A., Hochberg, M.C., Katz, L.M., Korn, J.H., Lachenbruch, P.A., LeRoy, E.C., Mitrane, M.P., Paulus, H.E., Postlethwaite, A.E. & Steen, V.D. (1995). Guidelines for clinical trials in systemic sclerosis (scleroderma): I. Disease modifying interventions, *Arthritis and Rheumatism* **38**, 351–360.

SIMON J. DAY

Bliss, Chester Ittner

Born: 1899, in Springfield, Ohio.
Died: 1979.

Photograph supplied by Yale University Archives

Chester Bliss studied entomology at Ohio State University, earning a B.A. in 1921. He continued these studies at Columbia University where he was awarded an M.A. in 1922 and a Ph.D. in 1926. Upon his graduation, he took a position as an entomologist in the Department of Agriculture, which he held until 1933 when his work was cut short by the depression. Moving to London, he attended lectures given by **R.A. Fisher** and collaborated with him on research over the next two years [6]. It was during this period that he did much of his work on probit analysis (*see* **Quantal Response Models**).

From London, he took a position at the Institute of Plant Protection in Leningrad in 1936. In 1938, he returned to the US, where he became a biometrician at the Connecticut Agricultural Experiment Station in New Haven, Connecticut until he retired in 1971. In 1942, Dr Bliss was appointed Lecturer at Yale University, where he taught and collaborated on research projects throughout his career [6].

Dr Bliss was convinced of the value of sound statistical methods in biology, and he devoted his career to developing practical approaches to data analysis, and making these methods available to scientists in the field. His research focused on estimating the potency of biological agents (*see* **Biological Assay, Overview**), his best known being contributions to the development of probit analysis. He played a major role in founding the **International Biometric Society**, and was appointed as the principal statistical contributor to the *U.S. Pharmacopeia*.

Chester Bliss' work on the use of the probability integral **transformation**, or probit, was done in collaboration with R.A. Fisher. The response at a given dose is the proportion of observations in which the specified outcome was observed. To linearize the sigmoid curve that is often observed for such data, the proportion was transformed using the inverse of the standard normal distribution function. As this transformation is undefined for 0 or 1, their first approach was to eliminate these observations from the analysis. This troubled Bliss a great deal, because these were, after all, valid observations, so one would then be ignoring some of the data. Bliss' persistence lead Fisher to derive the full **maximum likelihood** solution for probit analysis. A hallmark of Bliss' work was in making these methods accessible to biologists, and, in his books, examples of these techniques were described in painstaking detail, so that they could be readily followed by those in the field.

While the **American Statistical Association** had formed a Biometrics Section in 1938, Dr Bliss was appointed to a committee in 1945 to report on the merits of a separate American Biometric Society. However, following the committee's report in 1946, it was decided to postpone a decision on the formation of a separate society. Bliss was once again galvanized to press for a separate society devoted to statistics in biology when the program for the fall 1947 meeting of the **International Statistical Institute** in Washington, DC, appeared and was found to contain virtually no biological applications. On short notice, organizational and financial arrangements were made for the first International Biometric Conference at the Marine Biological Laboratory in Woods Hole, Massachusetts, on September 5–6, 1947. Dr Bliss served from 1948 to 1955 as the first Secretary for the Biometric Society, and was its Treasurer from 1951 to 1956. Together with **Gertrude Cox**, the editor of ***Biometrics***, the society was nurtured and grew into a vital organization with a diverse worldwide membership.

The editors of the *U.S. Pharmacopeia* recognized that accurate estimates of potency for biological agents required not only careful laboratory work, but also sound methods for analyzing the data. For many years, Dr Bliss provided advice on the methods of data analysis that should be used for the various agents, bringing his training in biology, as well as his insights into the application of statistical methods, to this standard reference.

During his career, he wrote over 130 articles, most of which dealt with various aspects of bioassay. These included work on methods for analyzing vitamins, analgesics, insulin, digitalis, penicillin, thiamin, radiation, parathyroid extract, adrenal cortex extract, cardiac glucosides, insecticides, and anthelmintics [1, 2]. In 1952 he published a book entitled *The Statistics of Bioassay, With Special Reference to the Vitamins* [3], which was to be followed by a three-volume work on *Statistics in Biology* [4, 5]. The first two volumes appeared in 1967 and 1970, respectively, and he was working on the third volume at the time of his death. His style was always meticulous and clear so that it could be followed readily by nonmathematicians.

In recognition of his contributions to biostatistics, he was elected an honorary life member of the Biometric Society, an Honorary Fellow of the **Royal Statistical Society**, and a Fellow of the Institute of Mathematical Statistics and of the American Statistical Association.

References

[1] Bliss, C.I. (1935). Comparison of dosage mortality data, *Annals of Applied Biology* **22**, 307–335.
[2] Bliss, C.I. (1945). Confidence limits for biological assays, *Biometric Bulletin* **1**, 57.
[3] Bliss, C.I. (1952). *The Statistics of Bioassay with Special Reference to the Vitamins*. Academic Press, New York.
[4] Bliss, C.I. (1967). *Statistics in Biology*, Vol. 1. McGraw-Hill, New York.
[5] Bliss, C.I. (1970). *Statistics in Biology*, Vol. 2. McGraw-Hill, New York.
[6] Cochran, W.G. (1979). Obituary for Chester Ittner Bliss, 1899–1979, *Biometrics* **35**, 715–717.

THEODORE R. HOLFORD & C. WHITE

Blocking

Blocking is a term describing strategies that are used in the design of experiments (*see* **Experimental Design**) and of **sample surveys**. It is also used to describe certain statistical methods that are used in the statistical analysis of data that can arise from **observational studies** as well as experimental studies (*see* **Clinical Trials, Overview**).

In the planning of experimental studies, the term *blocking* refers to the strategy of taking into consideration characteristics of the experimental units in the assignment of experimental treatments to units. The overall objectives of this strategy are to ensure that the assignment would result in minimization of **biases** caused by relationships between the outcome or dependent variable and the properties of the experimental units. For example, in a randomized trial of three treatments for carcinoma of the breast, it is possible that premenopausal women might differ from postmenopausal women in their response to these therapies. To ensure that each treatment has the same (or close to the same) proportion of premenopausal and postmenopausal women, one might perform **randomization** separately by menopausal status (i.e. *block* by menopausal status and randomize separately within each block). The various specific methods of blocking are discussed elsewhere (*see* **Balanced Incomplete Block Designs**; **Lattice Designs**; **Randomized Complete Block Designs**).

In the design of sample surveys, the blocks are generally called *strata*, and the objective of *blocking* in this scenario is to allocate the sample to strata in such a way that the resulting estimates have low **standard errors** under the cost constraints imposed on the survey (*see* **Stratified Sampling**; **Stratified Sampling, Allocation in**).

In analysis of data from observational studies, *blocking* is an analysis strategy used to control for **confounding** by variables that are measured on the **nominal** or ordinal scale, or by variables grouped on the basis of some variable (*see* **Stratification**). Examples of techniques of stratified analysis for quantitative dependent variables include two or higher way **analysis of variance**, and for categorical dependent variables include **contingency table** methods such as the Cochran–**Mantel–Haenszel** class of tests.

PAUL S. LEVY

Blood Groups

The term *blood groups* encompasses the products of **genes** expressed as molecules on the surface of erythrocytes in humans. Beginning in 1901 with Landsteiner's discovery of the ABO blood group system, these markers have been the subject of intense interest, primarily because of their clinical importance in the transfusion of blood, but secondarily because of their usefulness as **genetic markers**. The major impetus for study, of course, was the need to match for some of these groups in the transfusion of blood products. Included in the classification of major blood groups are the ABO, Rhesus(Rh), MNSs, P, Kell, Duffy, Kidd, Lutheran, and Lewis groups. All blood units transfused must be matched for the A and B antigens of the ABO system because of naturally occurring antibodies against them. All units are also matched for the D antigen of the Rh system because of the extreme immunogenicity of that antigen.

The chemical nature of these molecules has been worked out during the last 20 years with the Rh antigens being characterized only with the tools of molecular biology. Since the A and B substances of the ABO group appear in body secretions, such as saliva in some persons, as well as on the surface of the red cell, these antigens were among the first to be well characterized chemically. The two genes in the system (A and B) control the production of enzymes that place carbohydrates on specific sites of a precursor molecule, which itself is under genetic control (the H gene). The O gene produces no enzyme. Inheritance in the ABO system follows a simple co-dominant pattern for A and B with O being a true recessive. The Rh system, originally thought to be controlled by one gene producing three factors (or alternatively by three closely linked genes) is now known to be made up of two distinct proteins controlled by two genes. One gene controls the presence or absence of factor D and the second controls a molecule that at the serological level displays two factors C/c and E/e. The presence of factor D produces the familiar Rh positive when tested serologically. The problem with characterizing the Rh molecules was the absence of any secreted form and the intimate relationship between the Rh molecules and the cell membrane (see [1] for more details on systems, methods, and biochemistry).

Before the advent of molecular biology, the ease of the techniques used to detect blood group antigens (simple agglutination or agglutination augmented with an antiglobulin reagent) made the red cell blood groups the most important tools in the study of human **population genetics**. Blood groups could be studied both in families and populations and their genetics (in most cases) followed **Mendel's laws**. Population studies involving blood groups allowed the characterization of both historical movements of populations and present-day population isolates. Examples of the former are the pattern of blood group B in populations of Eastern Europe reminiscent of the Mongol invasions of the early post-Roman world. Examples of the latter might be the characterization of pockets of Rh negative populations such as the Basques of northern Spain. (For more information see [2].)

Although their use in population genetics is being partially supplanted by molecular techniques, blood groups remain of paramount importance in the provision of blood products for transfusion.

References

[1] Issitt, P.D. (1985). *Applied Blood Group Serology*, 3rd Ed. Montgomery, Miami. (Preferably, *see* Issitt, P.D. & Anstee, D.J. *Applied Blood Group Serology* 4th Ed. Expected from Montgomery, Miami, 1998).

[2] Mourant, A.E., Kopec, A.C. & Domaniewska-Sobczak, K. (1976). *The Distribution of the Human Blood Groups and Other Polymorphisms*, 2nd Ed. Oxford University Press, London.

E. REISNER

Bonferroni Inequalities and Intervals

Let $A_1, A_2, \ldots, A_n$ be any random events. Boole's inequality, or the first-order Bonferroni inequality, states that the probability that at least one of these events occurs is less than or equal to the sum of the probabilities of the individual events:

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i). \tag{1}$$

The majority of statistical applications of (1), commonly referred to as Bonferroni's inequality, deal with testing of multiple hypotheses (*see* **Hypothesis Testing**) and related **simultaneous confidence intervals** (*see* **Multiple Comparisons**; **Simultaneous Inference**). Suppose that n hypotheses are tested and that all of the hypothesis are true: the *complete* null hypothesis. If the ith hypothesis is tested at the α_i level of significance (*see* **Level of a Test**), where

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = \alpha, \tag{2}$$

then it follows from (1) that the probability of falsely rejecting at least one of the null hypotheses is at most α. The most common choice of α_i are α/n.

Similarly, the Bonferroni technique yields simultaneous confidence intervals for a set of n parameters by adding and subtracting the required estimated standard error multiplied by the $\alpha_i/2$ percentile point of an appropriate distribution from the point estimates of interest, where (2) holds. In a related encyclopedia article, Alt [2] presents formulas for the most commonly used Bonferroni confidence intervals.

In addition to always yielding simultaneous confidence intervals, it is also particularly easy to obtain adjusted simultaneous ***P* values** when we apply the easily understood Bonferroni technique [76].

In the multiple comparison setting where the **analysis of variance** (ANOVA) is used, methods superior to the Bonferroni method are often available. For example, Tukey's multiple comparison procedure is superior to the Bonferroni procedure for comparing all pairwise means in a one way ANOVA [40, 47]. The Bonferroni method of multiple comparisons is often the method of choice when a small subset of all possible comparisons is of interest [20], in settings where **covariate** adjustments are employed [20, 49], in situations in which distributions other than the **multivariate normal** or **multivariate *t*** are employed [59], or in discrete data settings such as obtaining simultaneous confidence intervals for **multinomial** proportions [4].

The major advantage of the Bonferroni procedure is its generality and flexibility. In the multiple endpoint setting it can be employed when some outcomes are quantitative and others are qualitative, unlike most multivariate techniques. In the multiple comparison with **multiple endpoints** setting, the Bonferroni technique can be applied to control for the multiple endpoints in conjunction with any technique to control for the multiple comparisons for each endpoint [49]. Consequently, corrections are made for the multiplicity of endpoints without making assumptions about the joint distributions of the outcomes.

The Bonferroni technique is also employed in a variety of research areas, such as spatial correlations [28], **nonparametric regression** [19], simultaneous confidence intervals for survival probabilities [1] (*see* **Survival Analysis, Overview**), obtaining optimal cutpoints [29], and detecting outliers in growth curve modeling [9] (*see* **Nonlinear Growth Curve**).

Textbooks by Neter et al. [49] and Fleiss [20] present an overview of the uses of Bonferroni's inequality in basic statistical applications. More detailed applications in multiple comparison settings are provided in texts by Hochberg & Tamhane [32], Hsu [40], and Miller [47]. The first two references are good sources for stepwise multiple testing procedures based on Bonferroni's inequality. Books by Morrison [48] and Srivastava & Carter [66] present applications of Bonferroni's inequality in **multivariate analysis**.

Multiple Comparisons for a Single Outcome

A competitor to the Bonferroni technique in the general linear model is the F projection method of Scheffé, the S-method. The S-method, which can be used to control the family-wise error rate for all possible contrasts, is useful in a hypothesis generating framework. Alt [2] presents more detail of the S-method. Both the Bonferroni and S-method

are applicable regardless of the correlation structure present. Applications in which both methods are applicable, in which more powerful methods do not exist, and in which it is valid to calculate both and use the least conservative include the **analysis of covariance** (ANCOVA) [20, 49], inverse prediction [49] (*see* **Calibration**), simultaneous **prediction** intervals [15, 47, 49], and simultaneous **tolerance intervals** [43, 47]. The S-method should be employed if some comparisons are not specified in advance. The Bonferroni method is generally less conservative than the S-method, unless a large number of comparisons is made [3, 16, 21]. However, when the error degrees of freedom is very low, the S-method is superior to the Bonferroni method [46].

Whenever the product inequality

$$\Pr\left(\bigcap_{i=1}^{n} A_i^c\right) \geq \prod_{i=1}^{n} \Pr(A_i^c) \tag{3}$$

is valid, then it follows from De Morgan's law that

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq 1 - \prod_{i=1}^{n} \Pr(A_i^c). \tag{4}$$

Whenever (4) is satisfied it yields a less conservative correction for multiple testing problems than (1), although the improvement is slight for small values of α [16]. Sidak [63] showed that this is the case for two-sided significance tests based on the multivariate normal and multivariate t distribution, including ANOVA applications. Sidak [64] showed that the Studentized maximum modulus yields a slightly less conservative procedure than the product inequality which controls the **experiment-wise error rate** for two-sided testing problems in the ANOVA setting. This procedure, known as the GT2 method, is available in SAS.

Additional applications in which the Bonferroni technique of simultaneous inference is the method of choice include three-decision problems [11] and **randomization tests** [51].

Sverdrup [68, 69] and Hjort [30] show how the S-method and the Bonferroni method are applicable in simultaneous inference settings whenever **generalized maximum likelihood** estimators, which are asymptotically normally distributed with a **covariance matrix** which can be consistently estimated, are employed. They show that the Bonferroni method is usually superior to the S-method in such applications, which include categorical data.

The Bonferroni bound is accurate when the number of comparisons is not large, when the $\Pr(A_i)$ are less than 0.1, and when the positive dependence among the A_i is not large. The technique also works better for continuous data than for discrete data. Tarone [71] shows how the Bonferroni inequality can be modified to obtain less conservative corrections for multiple testing of **categorical data**.

Multiple Outcomes for Two Groups

A sequentially rejective procedure of Holm [34] controls the experiment-wise error rate and is less conservative than the classical Bonferroni procedure. With the Holm procedure, one first tests the outcome with the largest observed difference at the α/n significance level, the usual Bonferroni adjustment. If one fails to reject the largest observed difference at the α/n significance level, then one fails to reject each of the null hypotheses. Otherwise, the endpoint with the second largest observed difference is tested at the $\alpha/(n-1)$ significance level. Testing progresses from the strongest to the weakest observed difference and stops with failure to reject the remaining hypotheses as soon as one fails to reject a null hypothesis. At each stage the significance level is increased, with the ith significance test being performed at the $\alpha/(n-i+1)$ significance level. For example, if five outcomes are tested, then significance levels of $\alpha/5, \alpha/4, \alpha/3, \alpha/2$, and α are employed. Wright [76] and Troendle [72] argue for more widespread use of sequentially rejective procedures instead of the classical Bonferroni correction. An extensive literature is available on modifications of Holm's sequentially rejective procedure and applications of the procedure in new situations [6, 7, 14, 17, 18, 31, 33, 35, 56–58, 65].

When many hypotheses are tested, even sequentially rejective procedures are overly conservative. In such applications one may wish to control the false discovery rate [8], the expected proportion of falsely rejected hypotheses, rather than the experiment-wise error rate.

A competitor to Bonferroni adjustments, which takes advantage of the correlation between outcomes and accommodates distributional characteristics of the individual outcomes, is based on resampling and **Monte Carlo** simulation [74].

A single global test statistic based on all of the individual outcomes may also be employed [50, 53, 70] in place of the Bonferroni procedure (*see* **Multiple Endpoints, Multivariate Global Tests**).

Higher Order Bonferroni Inequalities and Extensions

We define the first- and second-order Bonferroni sums as

$$S_1 = \sum_{i=1}^{n} \Pr(A_i) \quad \text{and} \quad S_2 = \sum_{j=1}^{i-1} \sum_{i=2}^{n} \Pr(A_i \cap A_j).$$

The second- (first-)order Bonferroni bound is a lower (upper) bound to the probability of a union:

$$b_2 = S_1 - S_2 \leq \Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq S_1 = b_1. \quad (5)$$

Higher-order Bonferroni bounds are obtained by alternately adding odd order Bonferroni sums and subtracting even order Bonferroni sums from lower order bounds. Even (odd) order Bonferroni bounds are lower (upper) bounds to the probability of a union. A related class of bounds, sometimes known as Galambos bounds [22], are obtained by taking linear combinations of Bonferroni sums. Unlike Bonferroni bounds, odd (even) order Galambos bounds are lower (upper) bounds.

Second or lower order bounds are often preferred in applications due to computational considerations. A useful second order lower bound [22], referred to as an extended Bonferroni-type bound by Galambos (23), is

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \geq \frac{2S_1}{k} - \frac{2S_2}{[k(k-1)]} = b_G, \quad (6)$$

where $k = 2 + [2S_2/S_1]$ and $[y]$ indicates the largest integer less than y. When $k = 2$, $b_G = b_2$, the second-order Bonferroni bound, and otherwise $b_G > b_2$.

The most widely used second-order upper bound among a class of bounds based on Hunter's inequality [41] is

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq S_1 - \sum_{i=2}^{n} \Pr(A_i \cap A_{i-1}). \quad (7)$$

Stoline [67] advocates the use of Hunter's inequality in ANOVA applications where there is a strong positive dependence structure. In most situations in which ANOVA is applied, the improvement of (7) over (1) is moderate [32, 40]. However, Bauer & Hackl [5] and Worsley [75] show that (7) is a substantial improvement over (1), which is very accurate, in many other situations with a strong degree of positive dependence.

When both upper and lower bounds of order 2 or less are required, then one should use (6) and (7) rather than (5), as in Bjornstad & Butler [10].

The bound given in (7) has been extended to higher-order bounds by Bolviken [12], Schwager [61], and Hoover [36]. This class of bounds exhibit nesting: the accuracy of the bounds increases as the order increases. Higher-order Bonferroni bounds are not nested [61]. When high-dimension multivariate probabilities need to be accurately estimated, these higher-order bounds are required. One such application involves scan statistics [24, 25, 42, 73]. If precision is required and computational considerations are not limiting, then higher order bounds produce substantially narrower simultaneous confidence intervals and simultaneous prediction intervals [26, 27, 55].

The bound in (6) has been extended to higher order bounds [2, 13, 22, 23, 54].

Improvements to higher-order bounds have been developed [37, 39, 44, 45, 62].

Bonferroni-type bounds are also available for the probability of exactly r or at least r of n events occurring [2, 13, 23, 38, 52, 60].

A comprehensive treatment of Bonferroni inequalities is presented in the book by Galambos [23].

References

[1] Afifi, A.A., Elashoff, R.M. & Lee, J.J. (1986). Simultaneous non-parametric confidence intervals for survival probabilities from censored data, *Statistics in Medicine* **5**, 653–662.

[2] Alt, F. (1982). Bonferroni inequalities and intervals, in *The Encyclopedia of Statistical Sciences, Vol. 1* S. Kotz & S. Johnson, eds. Wiley, New York, pp. 294–300.

[3] Alt, F. & Spruill, C. (1977). A comparison of confidence intervals generated by the Scheffé and Bonferroni method, *Communications in Statistics: Theory and Methods* **A6**, 1503–1510.

[4] Bailey, B.J.R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based

on transformations of cell frequencies, *Technometrics* **22**, 583–589.

[5] Bauer, P. & Hackl, P. (1985). The application of Hunter's inequality in simultaneous testing, *Biometrical Journal* **83**, 25–38.

[6] Bauer, P. & Hackl, P. (1987). Multiple testing in a set of nested hypotheses, *Statistics* **18**, 345–349.

[7] Bauer, P., Hackl, P., Hommel, G. & Sonnemann, E. (1986). Multiple testing of pairs of one-sided hypotheses, *Metrika* **33**, 121–127.

[8] Benjamamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **83**, 289–300.

[9] Bhandary, M. (1995). Detection of outliers in growth curve models, *Communications in Statistics: Theory and Methods* **A24**, 1923–1940.

[10] Bjornstad, J.F. & Butler, R.W. (1988). The equivalence of backward elimination and multiple comparisons, *Journal of the American Statistical Association* **83**, 136–144.

[11] Bohrer, R., Chow, W., Faith, R., Joshi, V.M. & Wu, C.-F. (1981). Multiple decision rules for factorial designs: Bonferroni wins again!, *Journal of the American Statistical Association* **76**, 119–124.

[12] Bolviken, E. (1988). Some probability bounds relevant for Bonferroni significance levels under Gaussian models, *Scandinavian Journal of Statistics* **15**, 281–297.

[13] Boros, E. & Prekopa, A. (1989). Closed form two-sided bounds for probabilities that at least r and exactly r of n events occur, *Mathematics of Operations Research* **2**, 317–342.

[14] Cheung, S.H. & Holland, B. (1994). A step down procedure for multiple tests of treatment versus control in each of several groups, *Statistics in Medicine* **13**, 1261–1267.

[15] Chew, V. (1968). Simultaneous prediction intervals, *Technometrics* **10**, 323–330.

[16] Dunn, O.J. (1959). Confidence intervals for means of dependent normally distributed variables, *Journal of the American Statistical Association* **54**, 613–621.

[17] Dunnett, C.W. & Tamhane, A.C. (1993). Power comparisons of some step up multiple test procedures, *Statistics and Probability Letters* **51**, 55–58.

[18] Dunnett, C.W. & Tamhane, A.C. (1995). Step up multiple testing of parameters with unequally correlated estimates, *Biometrics* **51**, 217–227.

[19] Eubank, R.L. & Speckman, P.L. (1993). Confidence bands in nonparametric regression, *Journal of the American Statistical Association* **88**, 1287–1301.

[20] Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

[21] Fuchs, C. & Sampson, A.R. (1987). Simultaneous confidence intervals for the general linear model, *Biometrics* **43**, 457–469.

[22] Galambos, J. (1977). Bonferroni inequalities, *Annals of Probability* **5**, 577–581.

[23] Galambos, J. (1996). *Bonferroni-type Inequalities with Applications*. Springer-Verlag, New York.

[24] Glaz, J. (1992). Approximations for tail probabilities and moments of the scan statistic, *Computational Statistics and Data Analysis* **14**, 213–227.

[25] Glaz, J. (1993). Approximations for the probabilities and moments of the scan statistic, *Statistics in Medicine* **12**, 1845–1852.

[26] Glaz, J. (1993). Approximate simultaneous confidence intervals, in *Multiple Comparisons, Selection, and Applications in Biometry*, F.M. Hoppe, ed. Marcel Dekker, New York, pp. 149–166.

[27] Glaz, J. & Ravishanker, N. (1991). Simultaneous prediction intervals for multiple forecasts based on Bonferroni and product-type inequalities, *Statistics and Probability Letters* **12**, 57–63.

[28] Hall, P. (1988). On confidence intervals for spatial parameters estimated from nonreplicated data, *Biometrics* **44**, 271–277.

[29] Hilsenbeck, S.G. & Clark, G.M. (1996). Practical p-value adjustments for optimally selected cutpoints, *Statistics in Medicine* **15**, 103–112.

[30] Hjort, N. (1988). On large-sample multiple comparison methods, *Scandinavian Journal of Statistics* **15**, 259–271.

[31] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

[32] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[33] Holland, B.S. & Copenhaver, M.D. (1987). An improved sequentially rejective Bonferroni test procedure, *Biometrics* **43**, 417–423.

[34] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.

[35] Hommell, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test, *Biometrika* **75**, 383–386.

[36] Hoover, D.R. (1990). Subset complement addition upper bounds – an improvement of inclusion-exclusion with applications, *Journal of Statistical Planning and Inference* **12**, 195–202.

[37] Hoppe, F.M. (1985). Iterating Bonferroni bounds, *Statistics and Probability Letters* **3**, 121–125.

[38] Hoppe, F.M. (1993). Beyond inclusion-and-exclusion: natural identities for P[exactly t events] and P[at least t events] and resulting inequalities, *International Statistical Review* **61**, 435–446.

[39] Hoppe, F.M. & Seneta, E. (1990). A Bonferroni-type identity and permutation bounds, *International Statistical Review* **58**, 253–262.

[40] Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, New York.

[41] Hunter, D. (1976). An upper bound for the probability of a union, *Journal of Applied Probability* **13**, 597–603.

[42] Krauth, J. (1992). Bounds for the upper-tail probability of the circular ratchet scan statistic, *Biometrics* **48**, 1177–1185.

[43] Lieberman, G.J. & Miller, R.G. (1963). Simultaneous tolerance intervals for regression, *Biometrika* **50**, 155–168.

[44] Margaritescu, E. (1986). A note on Bonferroni inequalities, *Biometrical Journal* **28**, 937–943.

[45] Margolin, B.H. & Mauer, W. (1976). Tests of the Kolmogorov-Smirnov type for exponential data with unknown scale, and related problems, *Biometrika* **63**, 149–160.

[46] Mi, J. & Sampson, A.R. (1993). A comparison of Bonferroni and Scheffé bounds, *Journal of Statistical Planning and Inference* **36**, 101–105.

[47] Miller, R.G. (1881). *Simultaneous Statistical Inference*. Springer-Verlag, New York.

[48] Morrison, D.F. (1990). *Multivariate Statistical Methods*. McGraw-Hill, New York.

[49] Neter, J., Wasserman, W. & Kunter, M.H. (1990). *Applied Linear Statistical Models*. Irwin, Boston.

[50] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1087.

[51] Petrondas, E.A. & Gabriel, K.R. (1983). Multiple comparisons by rerandomization tests, *Journal of the American Statistical Association* **78**, 949–957.

[52] Platz, O. (1985). A sharp upper bound for the occurrence of at least m out of n events, *Journal of Applied Probability* **22**, 978–981.

[53] Pocock, S.J., Geller, N.L. & Tsiatis, A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics* **43**, 487–498.

[54] Prekopa, A. (1990). Sharp bounds on probabilities using linear programming, *Operations Research* **38**, 227–239.

[55] Ravishanker, N., Hochberg, Y. & Melnick, E.L. (1987). Approximate simultaneous prediction intervals for multiple forecasts, *Technometrics* **29**, 371–376.

[56] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality, *Biometrika* **77**, 363–365.

[57] Rom, D.M., Costello, J. & Connell, L.T. (1994). On closed test procedures for dose–response analysis, *Statistics in Medicine* **13**, 1583–1596.

[58] Rom, D.M. & Holland, B. (1995). A new closed multiple testing procedure for hierarchical families of hypotheses, *Journal of Statistical Planning and Inference* **46**, 265–275.

[59] Royen, T. (1991). Multivariate Gamma distributions with one-factorial accompanying correlation matrices and applications to the distribution of the multiple range, *Metrika* **38**, 299–315.

[60] Sathe, Y.S., Pradhan, M. & Shah, S.P. (1980). Inequalities for the probability of the occurrence of at least m of n events, *Journal of Applied Probability* **17**, 1127–1132.

[61] Schwager, S.J. (1984). Bonferroni sometimes loses, *American Statistician* **38**, 192–197.

[62] Seneta, E. (1988). Degree, iteration, and permutation in improving Bonferroni-type bounds, *Australian Journal of Statistics* **30**, 27–38.

[63] Sidak, Z. (1967). Rectangular confidence regions for means of multivariate normal distributions, *Journal of the American Statistical Association* **62**, 626–633.

[64] Sidak, Z. (1971). On probabilities of rectangles in multivariate student distributions: their dependence on correlations, *Annals of Mathematical Statistics* **42**, 169–175.

[65] Simes, S.M. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **30**, 507–512.

[66] Srivastava, M.S. & Carter, E.M. (1983). *An Introduction to Applied Multivariate Statistics*. North-Holland, New York.

[67] Stoline, M.R. (1983). The Hunter method of simultaneous inference and its recommended use for applications having large known correlation structure, *Journal of the American Statistical Association* **78**, 366–370.

[68] Sverdrup, E. (1986). Multiple comparison and the likelihood ratio testing: general theory and applications to categorical data, *Scandinavian Actuarial Journal* **13**, 13–63.

[69] Sverdrup, E. (1990). The delta multiple comparison method: performance and usefulness, *Scandinavian Journal of Statistics* **17**, 115–134.

[70] Tang, D., Geller, N.L. & Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints, *Biometrics* **49**, 23–30.

[71] Tarone, R.E. (1990). A modified Bonferroni method for discrete data, *Biometrics* **46**, 515–522.

[72] Troendle, J.F. (1996). A permutational step-up method of testing multiple outcomes, *Biometrics* **52**, 846–859.

[73] Wallenstein, S., Naus, J. & Glaz, J. (1993). Power of the scan statistic for detection of clustering, *Statistics in Medicine* **12**, 1829–1843.

[74] Westfall, P.H. & Young, S.S. (1993). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustments*. Wiley, New York.

[75] Worsley, K.J. (1982). An improved Bonferroni inequality and applications, *Biometrika* **69**, 297–302.

[76] Wright, S.P. (1992). Adjusted p-values for simultaneous inference, *Biometrics* **48**, 1005–1013.

T. COSTIGAN

Bonferroni, Carlo Emilio

Born: January 28, 1892, in Bergamo, Italy.
Died: August 18, 1960, in Firenze, Italy.

Carlo Emilio Bonferroni studied for the degree of *laurea* in Torino (Turin) under Peano and Segre, became *incaricato* (assistant professor) at the Turin Polytechnic, and then in 1923 took up the chair of financial mathematics at the Economics Institute in Bari. In 1933, he transferred to Firenze (Florence) where he held his chair until his death.

The obituary of him by Pagni [5] lists his works under three main headings: **actuarial** mathematics (16 articles, 1 book); **probability** and statistical mathematics (30, 1); and analysis, geometry, and rational mechanics (13, 0). His name is known in the statistical world for the contents of just two of these papers.

The two articles cover similar ground, but the 1935 article [3] is directed to a specific application, that is, life assurance, whereas the 1936 article [4] is more abstract. In the latter he developed formulas for the probability that of n events, exactly r, at least r, at least 1, at most r occur. He finally arrived at the sets of inequalities [his formulas (27) and (28)] which bear his name (*see* **Bonferroni Inequalities and Intervals**). As he noted, his formula (28) is a generalization of the inequality of Boole.

Apart from these, he also had interests in the foundations of probability. Two relevant articles are an inaugural lecture [2] and a more formal article published earlier, but written about the same time [1]. He developed a strongly frequentist view of probability (*see* **Inference**), denying that subjectivist views can even be the subject of mathematical probability. A quote from the lecture perhaps gives the flavor.

> A weight is determined directly by a balance. And a probability, how is that determined? What is, so to say, the probability balance? It is the study of frequencies which gives rise to a specific probability [2, p. 32].

After this point his statistical work moved on to work on relationship. He does not seem to have returned to probability again.

References

[1] Bonferroni, C.E. (1924). Intorno al concetto di probabilità, *Giornale di Matematica Finanziara* **6**, 105–133. The article itself is dated December 1925.

[2] Bonferroni, C.E. (1927). Teoria e probabilità, in *Annuario del R Istituto Superiore di Scienze Economiche e Commerciali di Bari per L'anno Accademico 1925–1926*. Bari, pp. 15–46. Given as the inaugural lecture for the academic year 1925–1926 on November 22, 1925.

[3] Bonferroni, C.E. (1935). Il calcolo delle assicurazioni su gruppi di teste, in *Studi in Onore Del Professore Salvatore Ortu Carboni*. Rome, pp. 13–60.

[4] Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3–62.

[5] Pagni, P. (1960). Carlo Emilio Bonferroni, *Bollettino del Unione Matematica Italiana* **15**, 570–574.

MICHAEL E. DEWEY

Bonferroni, Carlo Emilio

MICHAEL E. DEWEY

Volume 1, pp. 528–529

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Bootstrap Method

Bootstrap methods are procedures for the empirical estimation or approximation of **sampling distributions** and their characteristics. Their primary use lies in the estimation of accuracy measures, such as bias and variance, for parameter estimators, and in construction of **confidence sets** or **hypothesis tests** for population parameters. They are applied in circumstances in which the form of the population from which the observed data have been drawn is unknown. They prove particularly useful where very limited sample data are available and traditional parametric modeling and analysis are difficult or unreliable.

Bootstrap methods are closely related to other data resampling methods of error assessment, such as the **jackknife**, but are more widely applicable and can provide more accurate inference, although they generally require more computation. An introduction to bootstrap methodology which stresses applications is given by Efron & Tibshirani [6], while detailed accounts of theory are given by Hall [9] and Shao & Tu [11]. A critical evaluation of the importance of bootstrap methods in different contexts is given by Young [13]. A concise summary of the methods is given by Efron & Tibshirani [5], while the revolution that they offer for statistical practice is discussed in very accessible terms by Diaconis & Efron [3].

The bootstrap principle was formalized by Efron [4]. It may be summarized for a general situation as follows. We have data $Y = (Y_1, \ldots, Y_n)$ (not necessarily independent and identically distributed) and a statistical model P under which the data are obtained. Usually, P can be described by the joint distribution of Y, or by some quantities that uniquely determine this joint distribution. Suppose that we wish to estimate the distribution of a **random variable** or "pivot" $R_n(Y; P)$, or some characteristic of that distribution. Then the data Y are used to estimate P by $\hat{P}$. Letting Y^* be a bootstrap data set generated from $\hat{P}$, then the bootstrap estimator of the distribution of $R_n(Y; P)$ is the conditional distribution of $R_n(Y^*; \hat{P})$, given Y. Where this conditional distribution is not expressible as an explicit function of Y, **simulation** by **Monte Carlo Methods** can be used to construct an approximation to the bootstrap estimator. The bootstrap can therefore be applied to any situation in which an underlying model P can be postulated and estimated and where one can sample from the estimated model $\hat{P}$.

Bootstrap methods are most fully developed for the case in which $Y_1, \ldots, Y_n$ are an independent and identically distributed (iid) sample. Extensions to independent, but not iid, data problems, such as **regression** problems, are often straightforward. Extensions to the dependent data setting are less well developed. In that context, care must be taken to account for the dependence structure in the data: some remarks are made below. We concentrate here mainly on the case of an iid sample.

The bootstrap may be applied parametrically or nonparametrically. With the former, we assume some parametric form for P, estimate any unknown parameters in its specification, typically by **maximum likelihood**, and so obtain $\hat{P}$. With the latter, we assume nothing about the form of P, and $\hat{P}$ is taken as the "empirical distribution function", usually denoted by F_n, of the given sample data. This distribution puts an equal point mass n^{-1} on each observed data point $Y_i, i = 1, \ldots, n$. Then a "bootstrap sample" $Y^* = (Y_1^*, \ldots, Y_n^*)$ is generated by independently sampling, with replacement, from the given data points (*see* **Sampling With and Without Replacement**).

We illustrate the *bootstrap* for the problem of variance estimation, adopting the notation of Shao & Tu [11]. Let $Y_1, \ldots, Y_n$ be iid from an (unknown) distribution F and let $T_n \equiv T_n(Y_1, \ldots, Y_n)$ be a given statistic, such as a parameter estimator. Then the variance of T_n is $\mathrm{var}_F(T_n) = \beta(F)$, say, a function of the unknown F.

If T_n is simple, then we can obtain an explicit expression for $\beta(F)$ as a function of unknown quantities, such as population **moments**, and then estimate $\mathrm{var}_F(T_n)$ by substituting estimates, constructed from the sample, for these unknowns. But, usually, this standard approach to variance estimation is too complicated to be useful. Bootstrap methods enable $\beta(F)$ to be estimated quite generally.

The bootstrap estimator of $\mathrm{var}_F(T_n)$ is

$$v_{\mathrm{BOOT}} = \mathrm{var}_*[T_n(Y_1^*, \ldots, Y_n^*)|Y_1, \ldots, Y_n],$$

where $\{Y_1^*, \ldots, Y_n^*\}$ is an iid sample from $\hat{F}$, an estimator of F, and $\mathrm{var}_*(\cdot|Y_1, \ldots, Y_n)$ denotes the conditional variance, given $Y_1, \ldots, Y_n$. An appealingly simple choice is to take $\hat{F} = F_n$.

In circumstances in which $\beta(F)$ is available explicitly, as a known function of F, the bootstrap

estimator v_{BOOT} is just a substitution estimator $\beta(\hat{F})$, and may be computed exactly and analytically. Usually $\beta(F)$ is not known explicitly, so we cannot evaluate v_{BOOT} exactly. However, in this case, Monte Carlo simulation can be used to approximate v_{BOOT} numerically. We repeatedly draw data samples from $\hat{F}$ and then use the sample variance of the values of T_n computed from these bootstrap samples as an approximation to v_{BOOT}.

We draw $\{Y_{1b}^*, \ldots, Y_{nb}^*\}, b = 1, \ldots, B$, for suitable B, independently from $\hat{F}$, conditional on $Y_1, \ldots, Y_n$; compute $T_{n,b}^* = T_n(Y_{1b}^*, \ldots, Y_{nb}^*)$ and approximate v_{BOOT} by the Monte Carlo approximation

$$v_{\mathrm{BOOT}}^{(B)} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{l=1}^{B}T_{n,l}^*\right)^2. \qquad (1)$$

Then $v_{\mathrm{BOOT}} = \lim_{B\to\infty} v_{\mathrm{BOOT}}^{(B)}$.

The traditional approach to variance estimation when $\beta(F)$ is unavailable explicitly is to first obtain an explicit asymptotic approximation to $\beta(F)$ and then estimate the unknown quantities in this asymptotic formula. Bootstrap methods therefore (i) avoid the need for analytic calculation and approximation, and (ii) avoid the errors associated with asymptotic approximation. This may lead to greater accuracy than is obtained from classical approaches to error assessment.

The standard error of T_n is $[\mathrm{var}_F(T_n)]^{1/2}$. Its bootstrap estimator is $\sqrt{v_{\mathrm{BOOT}}}$, approximated by $\sqrt{v_{\mathrm{BOOT}}^{(B)}}$ in circumstances in which Monte Carlo approximation is necessary.

Example 1

Efron & Tibshirani [6] present the following small data set, representing the survival times in days of seven mice receiving a new medical treatment after a test surgery: 94, 197, 16, 38, 99, 141, 23. Denote the observations by $Y_1, \ldots, Y_n$, where $n = 7$, so that $Y_1 = 94$, $Y_2 = 197$, and so on. Let the ordered observations be $Y_{(1)} < \cdots < Y_{(n)}$, so that $Y_{(1)} = 16$, $Y_{(2)} = 23$, and so on. We consider use of the bootstrap to estimate the standard error of three statistics: $T^{(1)} = n^{-1}\sum_{i=1}^{n} Y_i$, the mean survival time, the value of which is 86.9 for these data; $T^{(2)} = Y_{(4)}$, the median survival time, here equal to 94; and $T^{(3)} = (n-2)^{-1}\sum_{i=2}^{n-1} Y_{(i)}$, a trimmed mean (*see* **Trimming and Winsorization**), which takes the value 79.0. For the statistics $T^{(1)}$ and $T^{(2)}$ no Monte Carlo simulation is necessary, as the bootstrap variance estimator v_{BOOT} has an explicit expression in terms of $Y_1, \ldots, Y_n$. In the case of the mean $T^{(1)}$, for example, $v_{\mathrm{BOOT}} = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2/n^2$, where $\overline{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i \equiv T^{(1)}$. The bootstrap variance estimator for $T^{(2)}$ is given by a more complicated formula: see Efron & Tibshirani [6, Chapter 2]. The bootstrap standard error estimates are 23.36 and 37.83 for $T^{(1)}$ and $T^{(2)}$ respectively. In the case of $T^{(3)}$ the Monte Carlo approach must be used to obtain an approximation to v_{BOOT}. A series of B bootstrap samples are drawn from the given data and an approximation $v_{\mathrm{BOOT}}^{(B)}$ to v_{BOOT} computed from (1). A bootstrap sample is drawn by randomly sampling, with replacement, from the original datapoints $Y_1, \ldots, Y_n$. A typical bootstrap sample might be, for instance, $Y^* = (38,\ 99,\ 16,\ 99,\ 99,\ 16,\ 94)$. A bootstrap estimate of the standard error of $T^{(3)}$ computed from $B = 200$ randomly drawn bootstrap samples was 31.45. The statistic $T^{(3)}$ is therefore estimated to be more variable than the mean, but less variable than the median, for this situation.

Our description of the bootstrap variance and standard error estimators is easily generalized to other more complicated problems. Of particular importance is bootstrap distribution estimation. Often, an accuracy measure of a particular statistic T_n is a characteristic of the sampling distribution of T_n. If the bootstrap is used to estimate this sampling distribution, an estimator of the accuracy measure is provided by the corresponding characteristic of the estimated sampling distribution. Viewed in these terms, the bootstrap variance estimator v_{BOOT}, for example, is just the variance of a bootstrap estimator of the sampling distribution of T_n.

Consider the problem of estimating the sampling distribution of a pivot $R_n(Y_1, \ldots, Y_n; F)$:

$$H_F(x) = \Pr[R_n(Y_1, \ldots, Y_n; F) \le x], \qquad (2)$$

where $Y_1, \ldots, Y_n$ are iid from F.

The bootstrap estimator is

$$\begin{aligned} H_{\mathrm{BOOT}}(x) &\equiv H_{\hat{F}}(x) \\ &= \Pr_*[R_n(Y_1^*, \ldots, Y_n^*; \hat{F}) \le x | Y_1, \ldots, Y_n], \end{aligned}$$

where $Y_1^*, \ldots, Y_n^*$ are iid from $\hat{F}$ and $\Pr_*(\cdot|Y_1, \ldots, Y_n)$ denotes the probability under $\hat{F}$, conditional on $Y_1, \ldots, Y_n$. If $H_{\mathrm{BOOT}}(x)$ is not available explicitly as a function of $Y_1, \ldots, Y_n$, we may use a Monte Carlo approximation to $H_{\mathrm{BOOT}}(x)$:

$$H_{\mathrm{BOOT}}^{(B)}(x) = \frac{1}{B}\sum_{b=1}^{B} I[R_n(Y_{1b}^*, \ldots, Y_{nb}^*; \hat{F}) \le x],$$

where $\{Y_{1b}^*, \ldots, Y_{nb}^*\}, b = 1, \ldots, B$, are independent bootstrap samples from $\hat{F}$, and I is the indicator function.

For estimating the sampling distribution of T_n, we simply set $R_n(Y_1, \ldots, Y_n; F) = T_n$. When T_n is used to construct a confidence set for a parameter θ related to F, we might use $\sqrt{n}(T_n - \theta)$ or the studentized pivot $(T_n - \theta)/S_n$, where S_n is an estimator of the standard deviation of T_n. The confidence set may be derived from the sampling distribution of $R_n(Y_1, \ldots, Y_n; F) = \sqrt{n}(T_n - \theta)$ or $(T_n - \theta)/S_n$, as indicated below.

The classical approach to distribution estimation obtains a simple theoretical formula for $H_F(x)$, exact or approximate, and substitutes estimators for unknown quantities in the theoretical formula. When $R_n(Y_1, \ldots, Y_n; F) = \sqrt{n}(T_n - \theta)$, usually $H_F(x)$ can be approximated by $\Phi(x/\sigma)$, where Φ is the distribution function of N(0,1) and σ is an unknown parameter related to F. If $\hat{\sigma}$ is an estimator of σ, then we estimate $H_F(x)$ by $\Phi(x/\hat{\sigma})$. When $R_n(Y_1, \ldots, Y_n; F) = (T_n - \theta)/S_n$, $H_F(x)$ can often be approximated by $\Phi(x)$. But use of the bootstrap can provide a better (more accurate) approximation to $H_F(x)$ simply, without analytic calculation.

Example 2

As an example, consider the case in which $R_n(Y_1, \ldots, Y_n; F) = \overline{Y}_n - \mu$, where $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and $\mu = \mathrm{E}(Y_1)$. Then, with $\hat{F} = F_n$, the empirical distribution function of $Y_1, \ldots, Y_n$,

$$H_{\mathrm{BOOT}}(x) = \Pr_*(\overline{Y}_n^* - \overline{Y}_n \le x|Y_1, \ldots, Y_n),$$

where $\overline{Y}_n^* = n^{-1}\sum_{i=1}^n Y_i^*$ denotes the mean of an independent sample of size n drawn from F_n.

As an illustration, consider the case in which $n = 10$ and the given data $Y_1, Y_2, \ldots, Y_{10}$ are a random sample from the normal distribution N(0, 1). The objective is to use the bootstrap to estimate the sampling distribution of $R_n(Y_1, \ldots, Y_n; F) = \overline{Y}_n - \mu$, without assuming anything about the population from which the sample has been drawn. The true sampling distribution is actually N(0, 0.1), so we may check how the bootstrap performs. We consider the data set in which the Y_is are: -2.03, -0.58, 0.60, 0.45, 1.22, -0.69, 0.33, -1.69, 0.57, and -0.62. The bootstrap estimate of the sampling distribution was constructed from $B = 1000$ bootstrap samples, and is shown in Figure 1. Superimposing the (true) N(0, 0.1) normal curve on the bootstrap histogram shows that the bootstrap provides a very accurate estimate of the sampling distribution of $\overline{Y}_n - \mu$.

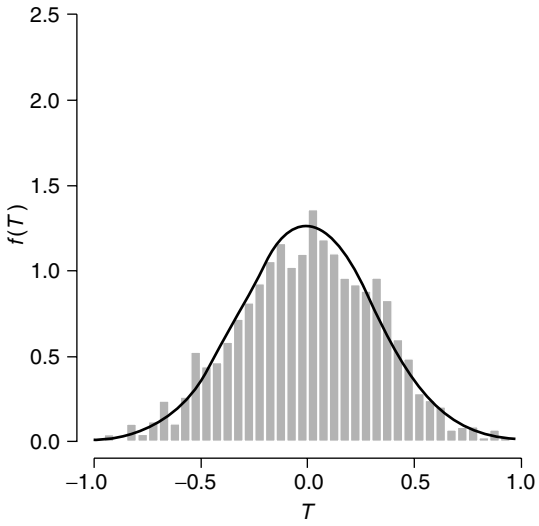

Figure 1 Bootstrap simulations, mean example

We now sketch other applications of the bootstrap. Further details are given by Shao & Tu [11, Chapter 1].

Bias Estimation

The bias of T_n as an estimator of θ is

$$\mathrm{bias}_F(T_n) = \int x \, \mathrm{d}H_F(x) - \theta,$$

with $H_F(x)$ given by (2), with $R_n = T_n$ (*see* **Unbiasedness**). The bootstrap bias estimator is obtained by substituting $\hat{F}$ and T_n for the unknown F and θ,

and is

$$b_{\mathrm{BOOT}} = \int x \, \mathrm{d}H_{\mathrm{BOOT}}(x) - T_n.$$

In situations in which the estimator b_{BOOT} has no explicit analytic form, it is approximated, using the Monte Carlo technique, by

$$b_{\mathrm{BOOT}}^{(B)} = \int x \, \mathrm{d}H_{\mathrm{BOOT}}^{(B)}(x) - T_n$$

$$= \frac{1}{B} \sum_{b=1}^{B} T_n(Y_{1b}^*, \ldots, Y_{nb}^*) - T_n.$$

Confidence Sets

Bootstrap confidence sets for an unknown parameter θ can be obtained using the percentiles of H_{BOOT} (or $H_{\mathrm{BOOT}}^{(B)}$).

As an illustration, let H_F be as given by (2), with $R_n(Y_1, \ldots, Y_n; F) = \hat{\theta}_n - \theta$, where $\hat{\theta}_n \equiv \hat{\theta}_n(Y_1, \ldots, Y_n)$ is an estimator of θ. An exact $1 - 2\alpha$ confidence interval for θ is

$$[\hat{\theta}_n - H_F^{-1}(1-\alpha), \hat{\theta}_n - H_F^{-1}(\alpha)].$$

Under repeated sampling of $Y_1, \ldots, Y_n$ from F, this interval contains the true value of θ with probability $1 - 2\alpha$. Approximating H_F by H_{BOOT} gives the bootstrap confidence interval for θ

$$[\hat{\theta}_n - H_{\mathrm{BOOT}}^{-1}(1-\alpha), \hat{\theta}_n - H_{\mathrm{BOOT}}^{-1}(\alpha)],$$

of *approximate* coverage $1 - 2\alpha$, under repeated sampling of $Y_1, \ldots, Y_n$ from F. This confidence interval is often called a "hybrid bootstrap" confidence interval [8]. A commonly used alternative is the "percentile bootstrap" confidence interval

$$[K_{\mathrm{BOOT}}^{-1}(\alpha), K_{\mathrm{BOOT}}^{-1}(1-\alpha)],$$

where $K_{\mathrm{BOOT}}(x)$ is the bootstrap estimator of the sampling distribution of the estimator $\hat{\theta}_n$.

Various alternative techniques for constructing bootstrap confidence sets, and their asymptotic properties, are described by Shao & Tu [11, Chapter 4].

Theoretical results on the performance of bootstrap methods are summarized by Shao & Tu [11]. In particular, it is known that for the iid case and bootstrap sampling from F_n, bootstrap estimators of the distributions of many commonly used regular statistics (such as means, functions of means, ***U*-statistics**, and sample **quantiles**) are consistent for the true sampling distributions and therefore give estimates that approach the correct values as the sample size n increases. Consistency of the bootstrap distribution estimator requires, roughly speaking, smoothness conditions that are almost the same as those required for asymptotic normality of the statistic, and certain further moment conditions. How good the bootstrap approximation is depends on the statistic to which it is applied. For nonstudentized statistics, the convergence rate of bootstrap estimators is the same as that for normal approximations. For a studentized statistic (*see* **Studentization**), the bootstrap estimator is better than a normal approximation. For the nonparametric bootstrap, consistency of bootstrap variance estimators requires stronger moment conditions than required for consistency of the bootstrap distribution estimator. Inconsistency of the bootstrap estimator in nonregular cases can often be rectified by the device of drawing bootstrap samples not of size n, but of size $m(n)$, which diverges to infinity more slowly than n. See [11, Chapter 3] for more detail of these asymptotic properties.

For confidence sets, of crucial importance is coverage accuracy. The coverage is the probability, under repeated sampling of Y from the underlying population, that the set contains the true value of the parameter of interest, and the confidence set is accurate if the actual coverage is close to the nominal desired coverage. Compared with confidence sets obtained by using the classical normal approximation, some sophisticated bootstrap confidence sets have greater theoretic accuracy, while simpler bootstrap confidence sets, such as the percentile and hybrid procedures, have the same accuracy, but are observed to work well in practice, and have the advantage of simplicity. See [8] and [11, Chapter 4].

The choice of the number of bootstrap samples B to be drawn in the Monte Carlo approach to the construction of bootstrap estimators is a delicate one. Shao & Tu [11, Section 5.4.1] consider the question in detail. As a rule of thumb, for moment estimators B should be between 50 and 200, while B should be considerable larger, of the order of at least 1000, for bootstrap distribution estimation and construction of confidence intervals. Other simulation approaches which reduce the computational costs

in bootstrap estimation are summarized by Shao & Tu [11, Chapter 5]: see also [9, Appendix I].

Procedures for bootstrap hypothesis testing are less well-developed than those for estimation and confidence set construction. The main point to be considered in the use of bootstrap procedures in this context is determination of the distribution from which bootstrap samples are drawn. The key principle is that bootstrap data should be generated from a distribution that satisfies the restrictions specified by the null hypothesis under test. Some of the methods used for determining this distribution are summarized by Shao & Tu [11, Section 4.5].

While our discussion so far has focused on the use of bootstrap methods with iid data, their implementation with non-iid data problems is often easy.

A key example concerns the simple **linear regression** model. Suppose that the data $Y_1, \ldots, Y_n$ are independent, of the form $Y_i = (Z_i, x_i)$, with the Z_i and x_i scalar. There are two representations of the model, and corresponding bootstrap procedures, depending on whether the x_i are considered random or fixed.

If the x_i are random, then the Y_i are iid from an unknown **bivariate distribution** P, and $\mathrm{E}(Z_i|x_i) = \beta x_i$, say. In this case P is estimated by the empirical distribution function of the Y_i. Bootstrap samples, used, for example, to estimate the sampling distribution of the **least squares** estimator $\hat{\beta}$, are drawn from this empirical distribution.

If the x_i are considered fixed, the model is that in which $Z_i = \beta x_i + \varepsilon_i$, with the ε_i iid from an unknown distribution F_ε with mean zero. Then F_ε can be estimated by the empirical distribution $\hat{F}_\varepsilon$ of the centered **residuals** $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - n^{-1}\sum_{j=1}^{n} \hat{\varepsilon}_j$, where $\hat{\varepsilon}_j = Z_j - \hat{\beta} x_j$. Bootstrap data $Z_1^*, \ldots Z_n^*$ are generated from iid data $\varepsilon_1^*, \ldots, \varepsilon_n^*$ with distribution $\hat{F}_\varepsilon$ by setting $Z_i^* = \hat{\beta} x_i + \varepsilon_i^*$.

Shao & Tu [11, Chapters 7 and 8] consider application of bootstrap methods to linear models, including **generalized linear models, nonlinear regression**, and **Cox regression models**. The same authors detail use of bootstrap methods in estimation problems arising in nonparametric and multivariate models, such as **nonparametric regression** models. A further important application of bootstrap methods lies in the analysis of **sample surveys**; see [11, Chapter 6].

To illustrate some of the key points relevant to bootstrapping dependent data, consider again Example 2 above, but now suppose the simplest dependence structure in statistical applications, that the Y_i are m-dependent; see [11, Section 9.1] for a formal definition. In these circumstances $\sqrt{n}(\overline{Y}_n - \mu)$ converges in distribution to $\mathrm{N}(0, \sigma_\infty^2)$, where $\sigma_\infty^2 \neq \mathrm{var}(Y_1)$ in general. This asymptotic distribution may be used to construct confidence intervals for μ, but only provided that a consistent estimate of σ_∞^2 is available. This may be far from straightforward to obtain: bootstrap methods are an attractive alternative.

While the bootstrap can bypass difficult problems associated with use of asymptotics (*see* **Large-sample Theory**), use of a bootstrap resampling scheme appropriate to independent data will fail to provide consistent approximation even in the case of weakly dependent processes. In our example, for instance, the simple bootstrap estimator of the variance of $\sqrt{n}\overline{Y}_n$ converges in probability to $\mathrm{var}(Y_1)$ and is therefore inconsistent for σ_∞^2, in general. Identification of a valid resampling scheme requires knowledge of the dependence structure of the observations.

As in the independent setting, the bootstrap can be applied parametrically to structured dependent data models, often with improvement over standard asymptotic procedures. Most developments to dependent data problems have considered such structured models; see, for example, [11, Chapter 9].

Considerable interest, however, lies in nonparametric resampling schemes. Künsch [10] proposed a "moving blocks" resampling scheme for stationary time series data. The basic idea here is to break the observed data series Y up into a collection of overlapping blocks of observations. Bootstrapped data series are obtained by independent sampling, with replacement, from among these blocks.

We illustrate this procedure in the context of the example above. Let b be a given block size. Define $\xi_i = (Y_i, \ldots, Y_{i+b-1})$ to be the block of b consecutive observations starting from Y_i, $i = 1, \ldots, n-b+1$. The moving blocks bootstrap is based on sampling with replacement from the collection $\{\xi_1, \ldots, \xi_{n-b+1}\}$. Suppose that k is an integer such that kb is approximately n, and let $\xi_1^*, \ldots, \xi_k^*$ be sampled independently and with replacement from $\{\xi_1, \ldots, \xi_{n-b+1}\}$. Let the $l = kb$ elements of $\xi_1^*, \ldots, \xi_k^*$ be concatenated into a single vector $(Z_1, \ldots, Z_l) \equiv (\xi_1^*, \ldots, \xi_k^*)$. Then $(Z_1, \ldots, Z_l)$ is

the bootstrap sample under the moving blocks bootstrap scheme and, for example, a bootstrap estimate of $\Pr[\sqrt{n}(\overline{Y}_n - \mu) \le z]$ is $\Pr[\sqrt{l}(\overline{Z}_l - \overline{Y}_n) \le z]$, where the probability is computed under the moving blocks resampling scheme, and where $\overline{Z}_l = l^{-1}\sum_{i=1}^{l} Z_i$. Consistency under the model of m-dependence is now achieved if b is allowed to grow to infinity with n.

The rate of approximation by the moving blocks method may be worse than the rate of normal approximation: the normal approximation yields better estimates of the sampling distribution of interest in this context. However, with suitable modification in the definition of the bootstrapped statistic, an improved approximation may be obtained. In the above example, such modification amounts to estimating $\Pr[\sqrt{n}(\overline{Y}_n - \mu) \le z]$ by $\Pr[\sqrt{l}(\overline{Z}_l - \mathrm{E}^*\overline{Z}_l) \le z]$, where $\mathrm{E}^*\overline{Z}_l$ denotes the expectation of $\overline{Z}_l$ under the moving blocks resampling scheme.

We conclude by reiterating the major considerations in use of bootstrap methods and with a data example.

1. Crucial to the practical effectiveness of bootstrap methods is how the model P is defined and estimated. Incorrect assumptions in postulating the model P, such as assuming independence for data which are actually correlated, may lead to incorrect conclusions.
2. Even if the model P is postulated correctly, the performance of the bootstrap relies on how well we can estimate it. In the iid case for example, it may be more effective to use a smoothed version of the empirical distribution function F_n in constructing the bootstrap estimator, rather than F_n itself; see [12] and [2]. The bootstrap itself may be applied to choose between different bootstrap estimators that with the smallest (bootstrap estimated) error. This *iterated bootstrap*, while in general demanding great computational expense, can also be used for the fine tuning of basic bootstrap procedures, by quantitative adjustment of the bootstrap estimator to account for estimated error; see [9, Chapter 1]. An important application of this method is in the refinement of bootstrap confidence sets, by bootstrap estimation of coverage error, and adjustment of the nominal coverage of the confidence set; see [7] and [1].
3. In cases in which parametric assumptions are justified, the bootstrap will do no better in general than the correct parametric technique. Bootstrap methods are also crucially conditioned by the available sample data. They are therefore only as good as the data with which they are provided. If the data set has **outliers** or influential points (*see* **Diagnostics**), and if the statistic being bootstrapped is nonrobust (*see* **Robustness**), bootstrap procedures can produce bad results, in the same way as classical approaches.

Example 3 (Bootstrap Bioequivalence)

The following application of the bootstrap is described by Efron & Tibshirani [6]. A drug company has applied each of three hormone supplement medicinal patches to eight patients suffering from a hormone deficiency. One of the three is "Approved", having received approval from the US **Food and Drug Administration (FDA)**. Another is a "Placebo" containing no hormone. The third is "New", being manufactured at a new facility, but otherwise intended to be identical to "Approved". The three wearings occur in random order for each patient and the blood level of the hormone is measured after each patch wearing: results are given in Table 1. The FDA requires proof of **bioequivalence** before approving sale of the product manufactured at the new facility.

Technically, let x be the difference between Approved and Placebo measurements on the same patient and let y be the difference between New and Approved:

$$x = \text{Approved} - \text{Placebo}, \qquad y = \text{New} - \text{Approved}.$$

Let μ and ν be the expectations of x and y and let

$$\rho = \frac{\nu}{\mu}.$$

Table 1 Bioequivalence data

Placebo	Approved	New	x	y
9 243	17 649	16 449	8 406	−1200
9 671	12 013	14 614	2 342	2601
11 792	19 979	17 274	8 187	−2705
13 357	21 816	23 798	8 459	1982
9 055	13 850	12 560	4 795	−1290
6 290	9 806	10 157	3 516	351
12 412	17 208	16 570	4 796	−638
18 806	29 044	26 325	10 238	−2719

The FDA criterion for bioequivalence is that the new facility matches the old facility within 20% of the amount of hormone that the old drug adds to the placebo blood levels, $|\rho| \leq 0.2$.

For the given data

$$(\overline{x}, \overline{y}) = (6342, -452) = (\hat{\mu}, \hat{\nu}),$$

giving an estimate of the ratio ρ as

$$\hat{\rho} = \frac{\hat{\nu}}{\hat{\mu}} = -0.071.$$

In formal terms, the FDA bioequivalence requirement is that a 90% central confidence interval for ρ lies within the range $(-0.2, 0.2)$. We use the percentile method to construct a nominal 90% confidence interval for ρ.

Let $\hat{F}$ be the distribution putting point mass 1/8 on each original data point (x_i, y_i), $i = 1, \ldots, 8$. Let $\{(x_1^*, y_1^*), \ldots, (x_8^*, y_8^*)\}$ be a sample drawn from $\hat{F}$. Such a sample gives a bootstrap replication of $\hat{\rho}$:

$$\hat{\rho}^* = \left(\sum_{i=1}^{8} y_i^*/8\right) \Big/ \left(\sum_{i=1}^{8} x_i^*/8\right).$$

A nominal 90% confidence interval for ρ is $(\hat{\rho}^*[0.05], \hat{\rho}^*[0.95])$, in terms of the lower 0.05 limit and upper 0.95 limit of $\hat{\rho}^*$ under the drawing of such bootstrap samples. An interval based on the drawing of 5000 random bootstrap samples (nearly exhaustive over all possible bootstrap samples) is $(-0.209, 0.123)$; see the bootstrap histogram in Figure 2, where the limits of the interval are shown by the broken lines.

The bioequivalence criterion is (just) violated. But is this interval to be trusted? The accuracy, under repeated sampling from the underlying population, of the chosen confidence interval procedure is an important part of the way the FDA decision making operates, and we must therefore be concerned at the accuracy of our percentile method interval. The iterated bootstrap can be used to refine the interval, by estimating the coverage error of the percentile method. In effect, we adjust the nominal level of the confidence interval, through the bootstrap, to deliver an interval that we believe will have coverage nearer to the required level 90%.

In Figure 3 is shown the graph of a bootstrap estimate of the coverage of the percentile method confidence interval, as a function of nominal coverage,

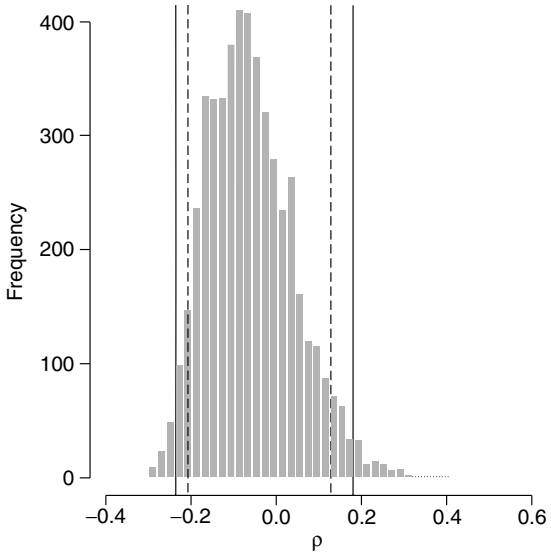

Figure 2 Bootstrap distribution, hormone data

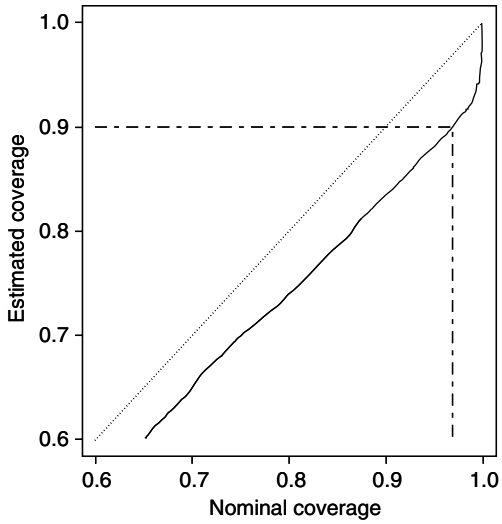

Figure 3 Calibration, hormone data

constructed from a Monte Carlo simulation. This simulation generated $B = 5000$ bootstrap samples from the given data. From each, the percentile method confidence interval of any given nominal coverage may be constructed, by drawing $C = 5000$ (second level) bootstrap samples. The proportion containing the observed value $\hat{\rho}$ estimates the coverage of the

percentile method interval, for that nominal coverage. This calibration curve shows that the percentile method confidence interval of nominal coverage 0.9 has estimated coverage 0.833: the percentile interval of nominal coverage 0.970 has estimated coverage equal to the required coverage 0.9. The iterated bootstrap confidence interval is therefore the percentile method interval of nominal coverage 97%. This interval, shown by the solid lines in Figure 2, is wider: the interval for ρ is now $(-0.237, 0.177)$. There is rather good evidence that the bioequivalence criterion is violated.

References

[1] Beran, R. (1987). Prepivoting to reduce level error of confidence sets, *Biometrika* **74**, 457–468.

[2] De Angelis, D. & Young, G.A. (1992). Smoothing the bootstrap, *International Statistical Review* **60**, 45–56.

[3] Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics, *Scientific American* **248**, 116–130.

[4] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**, 1–26.

[5] Efron, B. & Tibshirani, R.J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science* **1**, 54–77.

[6] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[7] Hall, P. (1986). On the bootstrap and confidence intervals, *Annals of Statistics* **14**, 1431–1452.

[8] Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *Annals of Statistics* **16**, 927–985.

[9] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

[10] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations, *Annals of Statistics* **17**, 1217–1241.

[11] Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

[12] Silverman, B.W. & Young, G.A. (1987). The bootstrap: to smooth or not to smooth?, *Biometrika* **74**, 469–479.

[13] Young, G.A. (1994). Bootstrap: more than a stab in the dark?, *Statistical Science* **9**, 382–415.

(*See also* **Bootstrapping in Survival Analysis**)

D. DE ANGELIS & G.A. YOUNG

Bootstrapping in Survival Analysis

Right Random Censoring

Let $Y_1, \ldots, Y_n$ be independent and identically distributed (i.i.d.) **random variables** (r.v.), called lifetimes. Such data in **survival analysis** have the typical feature that some of the Y_i are not fully observable due to various types of **censoring** and/or **truncation**. We restrict here to the model of random right censorship, which is described by considering another sequence $C_1, \ldots, C_n$ of i.i.d. r.v., called *censoring times*. The observations are the pairs $(T_1, \delta_1), \ldots, (T_n, \delta_n)$, where for $i = 1, \ldots, n$, $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \le C_i)$. As is mostly the case in survival analysis, we will assume that the Y's and C's are nonnegative, although this is not essential. The distribution functions of Y and C are denoted by F and G respectively. The model of random right censorship assumes that Y_i and C_i are independent for each i. This assumption entails that the T_i are i.i.d. with distribution function $H = 1 - (1 - F)(1 - G)$ and the δ_i are Bernoulli distributed with $\gamma = \mathrm{E}(\delta_1) = P(\delta_1 = 1) = \int_0^\infty (1 - G(s-))\,\mathrm{d}F(s)$. The **nonparametric maximum likelihood** estimator for the lifetime distribution F is the **Kaplan–Meier** [29] estimator $F_n(t)$ defined by

$$1 - F_n(t) = \prod_{T_i \le t} \left(1 - \frac{m_i}{M_i}\right)^{\delta_i} \tag{1}$$

where $m_i = \sum_{j=1}^n I(T_j = T_i)$ is the number of failures observed at T_i, and $M_i = \sum_{j=1}^n I(T_j \ge T_i)$ is the number at risk at T_i. If all T_i are different, then each $m_i = 1$ and $M_i = n - rank(T_i) + 1$, and in this case

$$1 - F_n(t) = \prod_{T_{(i)} \le t} \left(\frac{n - i}{n - i + 1}\right)^{\delta_{(i)}} \tag{2}$$

where $T_{(1)} \le T_{(2)} \le \cdots \le T_{(n)}$ are the **order statistics** of $T_1, \ldots, T_n$ and $\delta_{(1)}, \ldots, \delta_{(n)}$ are the corresponding δ's.

It is easy to see that F_n reduces to the usual empirical distribution function (*see* **Goodness of Fit**) if there is no censoring (all $\delta_i = 1$).

Efron's Bootstrap in the Right Random Censorship Model

It was Efron [17] who first proposed **bootstrap** procedures for right randomly censored observations. His first proposal is to take independent resamples from the Kaplan–Meier estimators of the lifetimes and the censoring times, and to combine them by taking minima and indicators.

Efron's bootstrap procedure I:

(1) Resample independently

$Y_1^*, \ldots, Y_n^* \overset{\text{i.i.d.}}{\sim} F_n$ (Kaplan–Meier estimator for F)

$C_1^*, \ldots, C_n^* \overset{\text{i.i.d.}}{\sim} G_n$ (Kaplan–Meier estimator for G)

(2) Form $(T_1^*, \delta_1^*), \ldots, (T_n^*, \delta_n^*)$, where $T_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = I(Y_i^* \le C_i^*)$.
(the definition of $G_n(t)$ is obtained from that of $F_n(t)$ by replacing δ_i by $1 - \delta_i$).

Efron's second proposal is to resample from the empirical distribution function of the observed minima and indicators.

Efron's bootstrap procedure II:
Resample $(T_1^*, \delta_1^*), \ldots, (T_n^*, \delta_n^*)$ as a random sample with replacement from the pairs $\{(T_1, \delta_1), \ldots, (T_n, \delta_n)\}$.

The first procedure could be called "model-based" since it uses the specific structure of random right censorship. The second procedure is more naive and could be called "model-free". An interesting result is that both procedures are identical if we assume that there are no ties between censored and uncensored values in the original sample. This means that both procedures give the same bootstrap values with the same probabilities. This equivalence no longer holds for left-truncated and right-censored data (see [6, 20,40]).

The Kaplan–Meier estimator based on the bootstrapped observations $(T_1^*, \delta_1^*), \ldots, (T_n^*, \delta_n^*)$ is then given by

$$1 - F_n^*(t) = \prod_{T_i^* \le t} \left(1 - \frac{m_i^*}{M_i^*}\right)^{\delta_i^*} \tag{3}$$

where $m_i^* = \sum_{j=1}^n I(T_j^* = T_i^*)$ and $M_i^* = \sum_{j=1}^n I(T_j^* \geq T_i^*)$.

Throughout, we will use the notations P^*, E^*, Var^*, ... for probability, **expectation**, **variance**, ... under the proposed resampling procedure (i.e. conditional on the original observations).

Weak Convergence of the Bootstrapped Kaplan–Meier Process

We consider the Kaplan–Meier process $n^{1/2}(F_n(t) - F(t))$, $0 \leq t \leq T_0$, where $T_0 < T_H = \min(T_F, T_G)$.

(Throughout, we use the following notation: for any distribution function L, we write $T_L = \inf\{t : L(t) = 1\}$ for the right endpoint of support). Let $D[0, T_0]$ denote the space of right continuous functions with left hand limits. The strong **consistency** of Efron's bootstrap for $n^{1/2}(F_n - F)$ has been established by different techniques ([33, 1]). The method in [1] uses **point processes** and martingale theory (*see* **Counting Process Methods in Survival Analysis**) to show the weak convergence result: if $T_0 < T_H$, then almost surely (a.s.) as $n \to \infty$,

$$n^{1/2}(F_n^*(\cdot) - F_n(\cdot)) \Rightarrow W(\cdot) \text{ in } D[0, T_0] \quad (4)$$

where $W(\cdot)$ is the same Gaussian process as for the original Kaplan–Meier process (obtained in [7]) (*see* **Large-sample Theory**). In [1] also, bootstrap versions of the **confidence** bands in [22] are obtained. The basic idea is the following: from the weak convergence result $n^{1/2}(F_n(\cdot) - F(\cdot)) \Rightarrow W(\cdot)$, we have that $\sup_{0\leq t\leq T_0} n^{1/2}|F_n(t) - F(t)| \xrightarrow{d} \sup_{0\leq t\leq T_0} |W(t)|$. Hence, an approximate $100(1-\alpha)\%$ confidence band can be obtained as $F_n(t) \pm c_\alpha n^{-1/2}$ $(0 \leq t \leq T_0)$, where c_α is such that $P(\sup_{0\leq t\leq T_0} |W(t)| \leq c_\alpha) = 1 - \alpha$. By the above result, c_α can be approximated by the $1-\alpha$ **quantile** of the bootstrap distribution of $\sup_{0\leq t\leq T_0} n^{1/2}|F_n^*(t) - F_n(t)|$. For related results, we also refer to the paper [24] on strong approximations and to [41] on the bootstrap for **multivariate survival** data.

Other Resampling Plans

There exist other resampling methods for censored data, different from the one of Efron [17] used above. Reid [35] proposed (in a context of estimation of the **median**) to take an i.i.d. resample from the Kaplan–Meier estimator of F and to work with the corresponding empirical distribution function.

Reid's bootstrap procedure

Resample $\widetilde{T}_1, \ldots, \widetilde{T}_n \overset{\text{i.i.d.}}{\sim} F_n$.

It is clear that such a resample always consists entirely of originally uncensored T_i. If $\widetilde{F}_n$ is the empirical distribution function of $\widetilde{T}_1, \ldots, \widetilde{T}_n$, then it is shown in [1] that for $T_0 < T_H$, a.s. as $n \to \infty$,

$$n^{1/2}(\widetilde{F}_n(\cdot) - F_n(\cdot)) \Rightarrow \widetilde{W}(\cdot) \quad \text{in } D[0, T_0] \quad (5)$$

but the Gaussian process $\widetilde{W}$ does not agree with the limiting process W of the original Kaplan–Meier estimator. It follows that Reid's bootstrap cannot be used to approximate the quantile that is needed in the construction of the confidence band for F.

Hjort [23] (in the context of **Cox's regression model**) proposed a resampling method as in Efron's Procedure I but with different construction of the censoring variables: if $\delta_i = 0$, then we observe the exact value of C_i and take $C_i^* = T_i$; if $\delta_i = 1$, then we know only that $C_i > T_i$ and we generate C_i^* from the Kaplan–Meier estimator of G, conditional on $C_i > T_i$.

Hjort's bootstrap procedure:

(1) Resample $Y_1^*, \ldots, Y_n^* \overset{\text{i.i.d}}{\sim} F_n$.
(2) Independently, resample $C_1^*, \ldots C_n^*$ as follows
- if $\delta_i = 0$, let $C_i^* = T_i$
- if $\delta_i = 1$, generate C_i^* from G_{T_i}, where

$$G_{T_i}(t) = \frac{G_n(t) - G_n(T_i)}{1 - G_n(T_i)} (t \geq T_i) \quad (6)$$

(3) Form $(T_1^*, \delta_1^*), \ldots, (T_n^*, \delta_n^*)$, where $T_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = I(Y_i^* \leq C_i^*)$.

If F_n^* is the Kaplan–Meier estimator formed with the above (T_i^*, δ_i^*), then in [30] (see also [16]) it is shown that for $T_0 < T_H$, a.s. as $n \to \infty$,

$$n^{1/2}(F_n^*(\cdot) - F_n(\cdot)) \Rightarrow W(\cdot) \text{ in } D[0, T_0] \quad (7)$$

where W is the same Gaussian process as for the original Kaplan–Meier process. Also, **Bayesian** bootstrapping and weighted bootstrapping have been suggested and studied; see, for example, [32] and [26]. It should also be mentioned that bootstrap methods

for arbitrary **counting process** models are discussed in [2].

Bootstrapping Kaplan–Meier Quantiles

For $0 < p < 1$, we denote the pth quantile of the lifetime distribution F by $\xi_p = F^{-1}(p) = \inf\{t : F(t) \geq p\}$. A simple estimator for ξ_p is the p-th quantile of the Kaplan–Meier estimator $F_n : \xi_{pn} = F_n^{-1}(p) = \inf\{t : F_n(t) \geq p\}$. Many large sample properties for quantile estimators can be derived from the asymptotic representations (Bahadur representations). They represent ξ_{pn} as follows:

$$\xi_{pn} = \xi_p + \frac{p - F_n(\xi_p)}{f(\xi_p)} + R_n(p) \qquad (8)$$

where $f = F'$ and $R_n(p)$ is a remainder term.

Weak convergence of the quantile process $n^{1/2}(\xi_{pn} - \xi_p)$ $(0 < p \leq p_0)$, where $0 < p_0 < \min(1, T_{G(F^{-1})})$ can be obtained. The limiting process is given by $-W(\xi_p)/f(\xi_p)$ $(0 < p \leq p_0)$, where W is the limiting Gaussian process of the Kaplan–Meier process.

Efron's bootstrap for Kaplan–Meier quantiles was studied in [33]. Let F_n^* be the Kaplan–Meier estimator based on the bootstrapped observations $(T_1^*, \delta_1^*), \ldots, (T_n^*, \delta_n^*)$ and let $\xi_{pn}^* = F_n^{*-1}(p) = \inf\{t : F_n^*(t) \geq p\}$. Their key result is a bootstrap version of the above representation theorem, from which weak convergence of $n^{1/2}(\xi_{pn}^* - \xi_p)$ follows. This, and the application to confidence bands for the quantile function is also discussed in [16].

A Practical Example

As a simple illustration of the use of the bootstrap in survival analysis, we consider the calculation of a confidence interval for the median survival time. We perform this on the Channing House data (see [25]) using the *R* software. Figure 1 shows the Kaplan–Meier survival estimator $1 - F_n(t)$ for the 97 men who lived in this retirement center. Of the 97 lifetimes, 46 were observed completely ($\delta = 1$) and 51 were censored ($\delta = 0$).

It follows that the median survival time is 1044 months. From each of 2000 resamples from the data $\{(z_1, \delta_1), \ldots, (z_{97}, \delta_{97})\}$, we calculated the Kaplan–Meier median. For the percentiles of these 2000 replicas of the median, we obtained

2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
1025	1029	1031	1036	1044	1055	1060	1080	1080

These values can be used be obtain simple confidence intervals for the median survival time. For example, a 90% confidence interval is given by [1029, 1080]. This method is called the percentile method. This and other results can also be found in [17].

Accuracy of the Bootstrap with Censored Data

Second-order asymptotics for the bootstrap have been studied to find the rate at which the bootstrap error tends to zero a.s. For uncensored data,

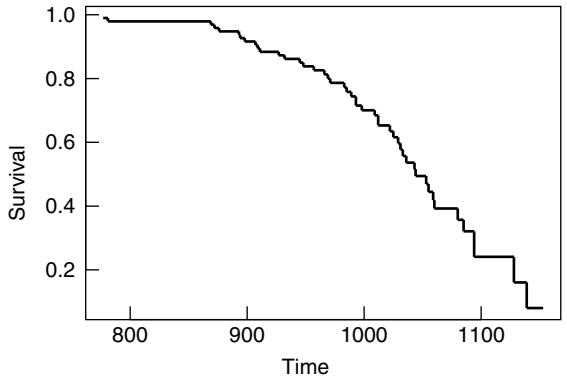

Figure 1 Kaplan–Meier survival curve $1 - F_n(t)$ for the Channing House data (time in months)

it is well known that for several classes of statistics, the bootstrap approximation is more accurate than the normal approximation. While the error of the normal approximation is typically $o(n^{-1/2})$ (by a Berry–Esseen theorem), the rate of convergence of the bootstrap estimator can, in some situations, be shown to be $o(n^{-1/2})$ a.s. (which is typically that of a one-term **Edgeworth expansion**). Edgeworth expansions for the **Studentized** Kaplan–Meier estimator $n^{1/2}(F_n(t) - F(t))/\widehat{\sigma}_G(t)$ and for the bootstrapped version $n^{1/2}(F_n^*(t) - F_n(t))/\widehat{\sigma}_G^*(t)$ have been derived in [10]. Here, $\widehat{\sigma}_G^2(t)$ is the Greenwood estimator of the variance given by

$$\widehat{\sigma}_G^2(t) = (1 - F_n(t))^2 n \sum_{T_i \le t} \frac{\delta_i}{M_i(M_i - m_i)} \tag{9}$$

and $\widehat{\sigma}_G^{*2}$ is the bootstrapped version (*see* **Kaplan–Meier Estimator**).

An important corollary is that (under the conditions: F and G continuous, $G(t) > 0$, and $1 - H(t) > 0$), we have a.s. as $n \to \infty$,

$$\sup_x \left| P^* \left(\frac{n^{1/2}(F_n^*(t) - F_n(t))}{\widehat{\sigma}_G^*(t)} \le x \right) - P\left(\frac{n^{1/2}(F_n(t) - F(t))}{\widehat{\sigma}_g(t)} \le x \right) \right| = o(n^{-1/2}). \tag{10}$$

This shows that bootstrapping the Studentized Kaplan–Meier estimator offers a better alternative than the normal approximation.(Compare with the $o(n^{-1/2})$ rate of the Berry–Esseen theorem in [9] and [28]).

In [27], it has been shown that modified bootstrapping (i.e. choosing a resample size m, which is possibly different from n) may lead to improved consistency rates for Kaplan–Meier quantiles. Modified bootstrapping for doubly censored data has also been considered in [5].

Bootstrapping in the Proportional Hazards Regression Model of Cox

Very often it happens that together with the observation of the ith individuals lifetime Y_i or censoring time C_i, one has also information on other characteristics Z_i (**covariates**, **explanatory variables**, design variables). Regression models study the effect of these covariates on the conditional distribution function of the true lifetimes $F_z(t) = P(Y \le t \mid Z = z)$.

In the **proportional hazards** model of Cox [11], the relation between a, possibly right censored, lifetime Y and the covariate Z is modeled via the **hazard rate** function

$$\lambda_z(t) = \lim_{\substack{h \to 0 \\ > }} \frac{1}{h} P(Y \le t + h \mid Y > t; Z = z). \tag{11}$$

(For simplicity, we assume here that the covariate Z is one-dimensional, but generalizations are possible). Cox's proportional hazards model specifies that $\lambda_z(t)$ is given by

$$\lambda_z(t) = \lambda_0(t) \mathrm{e}^{\beta z} \tag{12}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function (the hazard for an individual with $z = 0$) and β is an unknown regression parameter.

Suppose that there are n individuals in the study and that the observations are given as $(Z_1, T_1, \delta_1), \ldots, (Z_n, T_n, \delta_n)$, where for individual i, Z_i denotes the covariate and where, as before, $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \le C_i)$.

Cox's maximum **partial likelihood** estimator for β ([12]) is the value for β maximizing the partial likelihood function

$$L_n(\beta) = \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\beta Z_i}}{\sum_{j=1}^{n} \mathrm{e}^{\beta Z_j} I(T_j \ge T_i)} \right)^{\delta_i} \tag{13}$$

(or the log partial likelihood function $\log L_n(\beta)$).

The large-sample properties of Cox's maximum partial likelihood estimator were established in [36] (see also [3] for a counting process approach). The paper [36] shows strong consistency of the estimator $\widehat{\beta}_n$, and also asymptotic normality

$$n^{1/2}(\widehat{\beta}_n - \beta) \xrightarrow{d} N(0; \sigma^2(\beta))$$

where $\sigma^2(\beta) = \frac{1}{I(\beta)}$ with

$$I(\beta) = \int \left[\frac{\alpha_2(t, \beta)}{\alpha_0(t, \beta)} - \left(\frac{\alpha_1(t, \beta)}{\alpha_0(t, \beta)} \right)^2 \right] \times \alpha_0(t) \lambda_0(t) \, \mathrm{d}t$$

$$\alpha_k(t, \beta) = \int z^k e^{\beta z} P(T \geq t | Z = z) f(z)\, dz$$
$$k = 0, 1, 2 \qquad (14)$$

and f is the density of the covariate Z.

The asymptotic variance $\sigma^2(\beta) = 1/I(\beta)$ can be consistently estimated by $n/I_n(\widehat{\beta}_n)$, where $I_n(\beta) = -\partial/\partial\beta U_n(\beta)$ is the **information** function. Hence, the Studentized estimator is $(I_n(\widehat{\beta}_n))^{1/2}(\widehat{\beta}_n - \beta)$.

As suggested by Efron and Tibshirani [18], an obvious bootstrap procedure for Cox's proportional hazards regression model is to resample from the triples, then calculate the partial likelihood and then maximize it.

Efron and Tibshirani's bootstrap procedure

(1) Resample $(X_1^*, T_1^*, \delta_1^*), \ldots, (X_n^*, T_n^*, \delta_n^*)$ i.i.d. from the empirical distribution function of $(X_1, T_1, \delta_1), \ldots, (X_n, T_n, \delta_n)$.
(2) Calculate $\widehat{\beta}_n^*$: the value of β such that $L_n^*(\beta)$ is maximized.

A one-term Edgeworth expansion for $n^{1/2}(\widehat{\beta}_n - \beta)$ and related quantities is established in [21]. It is also shown that the bootstrap approximation is second-order correct: a.s. as $n \to \infty$,

$$\sup_{x \in \mathbb{R}} \left| P^* \left(\sqrt{I_n^*(\widehat{\beta}_n^*)}(\widehat{\beta}_n^* - \widehat{\beta}_n) \leq x \right) - P\left(\sqrt{I_n(\widehat{\beta}_n)}(\widehat{\beta}_n - \beta) \leq x \right) \right| = o(n^{-1/2}). \qquad (15)$$

Bootstrap confidence intervals for the regression parameter β, the distribution function $F_x(t)$, and the quantile function $F_x^{-1}(t)$ were studied by Burr [8]. The author performs a **Monte Carlo** study to compare different types of confidence intervals.

Bootstrapping in Nonparametric Regression Models

The relation between the distribution of the lifetime Y and the value of the covariate Z has also been analyzed in a completely **nonparametric** way, that is, without assuming any condition on the regression function. Beran [4] extended the definition of the Kaplan–Meier estimator to the regression context by proposing an entirely nonparametric estimator for $F_z(t) = P(Y \leq t \mid Z = z)$ based on a **random sample** of observations $(Z_1, T_1, \delta_1), \ldots, (Z_n, T_n, \delta_n)$. It is assumed that the r.v. Y_i and C_i are conditionally independent, given Z_i. This ensures **identifiability** of $F_z(t)$.

This estimator has been studied by several people. Uniform consistency has been shown in [4]. Weak convergence results and quantile functions were studied in [13, 14] and [15]. The paper [34] provides the counting process treatment and an extension of the regression model. Almost sure asymptotic representations in the fixed design case (i.e. nonrandom covariates Z_i) were obtained in [19] and [39]. In [37, 38] and [39], a further study was made of this Beran extension of the Kaplan–Meier estimator, the corresponding quantile estimator and the bootstrap versions.

The Beran estimator $F_{zh}(t)$ for $F_z(t)$ is given by

$$1 - F_{zh}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{w_{n(i)}(z; h_n)}{1 - \sum_{j=1}^{i-1} w_{n(j)}(z; h_n)} \right)^{\delta_{(i)}}. \qquad (16)$$

where $\{w_{ni}(z; h_n)\}$ is a sequence of **smoothing** weights depending on a kernel density and a positive bandwidth sequence $\{h_n\}$. We use the notation $w_{n(i)}(z; h_n)$ for the weight corresponding to $T_{(i)}$. The bootstrap procedure suggested in [39] consists of drawing the pairs $(T_1^*, \delta_1^*), \ldots, (T_n^*, \delta_n^*)$ with replacement from $(T_1, \delta_1), \ldots, (T_n, \delta_n)$, giving probability $w_{nj}(z_i; g_n)$ to (T_j, δ_j) for $j = 1, \ldots, n$. The bandwidth sequence $\{g_n\}$ is typically such that $g_n/h_n \to \infty$ as $n \to \infty$ ("oversmoothing"). In [38], strong consistency of this bootstrap is shown; see also [31] for a martingale approach.

References

[1] Akritas, M. (1986). Bootstrapping the Kaplan-Meier estimator, *Journal of the American Statistical Association* **81**, 1032–1038.
[2] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer Verlag, New York.
[3] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes, *Annals of Statistics* **10**, 1100–1120.

[4] Beran, R. (1981). Nonparametric Regression with Randomly Censored Survival Data, Technical Report, University of California, Berkeley.

[5] Bickel, P.J. & Ren, J.-J. (1996). The *m* out of *n* bootstrap and goodness of fit tests with double censored data, in *Robust Statistics, Data Analysis and Computer Intensive Methods*, H. Rieder, ed. Lecture Notes in Statistics 109, Springer, New York, p. 35–47.

[6] Bilker, W.B. & Wang, M.-C. (1997). Bootstrapping left truncated and right censored data, *Communications in Statistics-Simulation and Computation* **26**, 141–171.

[7] Breslow, N. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics* **5**, 437–453.

[8] Burr, D. (1994). A comparison of certain bootstrap confidence intervals in the Cox model, *Journal of the American Statistical Association* **89**, 1290–1302.

[9] Chang, M.N. & Rao, P.V. (1989). Berry-Esseen bound for the Kaplan-Meier estimator, *Communications in Statistics – Theory and Methods* **18**, 4647–4664.

[10] Chen, K. & Lo, S.-H. (1996). On bootstrap accuracy with censored data, *Annals of Statistics* **24**, 569–595.

[11] Cox, D.R. (1972). Regression model and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[12] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[13] Dabrowska, D.M. (1987). Non-parametric regression with censored survival time data, *Scandinavian Journal of Statistics* **14**, 181–197.

[14] Dabrowska, D.M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate, *Annals of Statistics* **17**, 1157–1167.

[15] Dabrowska, D.M. (1992). Variable bandwidth conditional Kaplan-Meier estimate, *Scandinavian Journal of Statistics* **19**, 351–361.

[16] Doss, H. & Gill, R. (1992). An elementary approach to weak convergence for quantile processes, with applications to censored survival data, *Journal of the American Statistical Association* **87**, 869–877.

[17] Efron, B. (1981). Censored data and the bootstrap, *Journal of the American Statistical Association* **76**, 312–319.

[18] Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, *Statistical Science* **1**, 54–77.

[19] González-Manteiga, W. & Cadarso-Suarez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications, *Journal of Nonparametric Statistics* **4**, 65–78.

[20] Gross, S.T. & Lai, T.L. (1996). Bootstrap methods for truncated and censored data, *Statistica Sinica* **6**, 509–530.

[21] Gu, M. (1992). On the Edgeworth expansion and bootstrap approximation for the Cox regression model under random censorship, *The Canadian Journal of Statistics* **20**, 399–414.

[22] Hall, W.J. & Wellner, J.A. (1980). Confidence bands for a survival curve from censored data, *Biometrika* **67**, 113–143.

[23] Hjort, N.L. (1985). Bootstrapping Cox's Regression Model, Technical Report 241 Department of Statistics, Stanford University, Stanford.

[24] Horváth, L. & Yandell, B.S. (1987). Convergence rates for the bootstrapped product-limit process, *Annals of Statistics* **15**, 1155–1173.

[25] Hyde, J. (1980). Testing survival with incomplete observations, in *Biostatistics Casebook*, R.G. Miller Jr., B. Efron, B.Wm. Brown Jr. & L.E. Moses, eds. Wiley, New York, p. 31–46.

[26] James, L.F. (1997). A study of a class of weighted bootstraps for censored data, *Annals of Statistics* **25**, 1595–1621.

[27] Janssen, P., Swanepoel, J. & Veraverbeke, N. (2002). The modified bootstrap error process for Kaplan-Meier quantiles, *Statistics and Probability Letters* **58**, 31–39.

[28] Janssen, P. & Veraverbeke, N. (1992). The accuracy of normal approximations in censoring models, *Journal of Nonparametric Statistics* **1**, 205–217.

[29] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[30] Kim, J. (1990). Conditional Bootstrap Methods for Censored Data. Unpublished Ph.D. dissertation, Florida State University, Department of Statistics.

[31] Li, G. & Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data, *Annals of the Institute of Statistical Mathematics* **53**, 708–729.

[32] Lo, A.Y. (1993). A Bayesian bootstrap for censored data, *Annals of Statistics* **21**, 100–123.

[33] Lo, S.-H. & Singh, K. (1986). The product-limit estimator and the bootstrap: some asymptotic representations, *Probability Theory and Related Fields* **71**, 455–465.

[34] McKeague, I.W. & Utikal, K.J. (1990). Inference for a nonlinear counting process regression model, *Annals of Statistics* **18**, 1172–1187.

[35] Reid, N. (1981). Estimating the median survival time, *Biometrika* **68**, 601–608.

[36] Tsiatis, A.A. (1981). A large sample study of Cox's regression model, *Annals of Statistics* **9**, 93–108.

[37] Van Keilegom, I. & Veraverbeke, N. (1996). Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles, *Communications in Statistics – Theory and Methods* **25**, 2251–2265.

[38] Van Keilegom, I. & Veraverbeke, N. (1997a). Weak convergence of the bootstrapped conditional Kaplan-Meier process and its quantile process, *Communications in Statistics – Theory and Methods* **26**, 853–869.

[39] Van Keilegom, I. & Veraverbeke, N. (1997b). Estimation and bootstrap with censored data in fixed design nonparametric regression, *Annals of the Institute of Statistical Mathematics* **49**, 467–491.

[40] Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data, *Journal of the American Statistical Association* **86**, 130–143.

[41] Yandell, B.S. & Horváth, L. (1988). Bootstrapped multi-dimensional product limit process, *Australian Journal of Statistics* **30**, 342–358.

Further Reading

Van Keilegom, I. & Veraverbeke, N. (1997c). Bootstrapping quantiles in a fixed design regression model with censored data, *Journal of Statistical Planning and Inference* **69**, 115–139.

Veraverbeke, N. (1997). Bootstrapping in survival analysis, *South African Statistical Journal* **31**, 217–258.

NOËL VERAVERBEKE

Bortkiewicz, Ladislaus von

Born: August 7, 1868, in St Petersburg, Russia.
Died: July 15, 1931, in Berlin, Germany.

Bortkiewicz studied initially in Russia, then in Gottingen under Wilhelm Lexis (1837–1914). From 1901, he was a professor at the University of Berlin, teaching statistics and economics. He contributed widely in theoretic, social, and economic statistics, developing especially the ideas of Lexis on dispersion in heterogeneous **binary data**.

He is best known in biostatistics for his presentation [1] of data on deaths from horse-kicks in 14 corps of the Prussian Army over a 20-year period. The variation in number of deaths is closely represented by a **Poisson distribution** (Table 1).

Table 1

Deaths	0	1	2	3	4	5-	Total
Frequency	–	–	–	–	–	–	–
Observed	144	91	32	11	2	0	280
Expected	139.0	97.3	34.1	8.0	1.4	0.2	280.0

The "Law of Small Numbers" [1] implied that for low **expectations** in **binomial** or Poisson data, considerable heterogeneity in the expectation could remain undetected from the marginal distribution. In this example, heterogeneity must exist, if only because the corps differed greatly in size. Further analysis of a reduced set of two-way data categorized by year and by corps [2, 4] reveals clear evidence of variation in expectation between both years and corps.

See [3] for further biographical details. The spelling of his name is variable.

References

[1] Bortkiewicz, L. von (1898). *Das Gesetz der kleinen Zahlen*. Teubner, Leipzig.

[2] Preece, D.A., Ross, G.J.S. & Kirby, S.P.J. (1988). Bortkewitsch's horse-kicks and the generalised linear model, *Statistician* **37**, 313–318.

[3] Sheynin, O.B. (1970). Bortkiewicz (or Bortkewitsch), Ladislaus (or Vladislav) Josephowitsch, in *Dictionary of Scientific Biography*, Vol. 1, C.C. Gillespie, ed. Scribner, New York, pp. 318–319.

[4] Winsor, C.P. (1947). Quotations: Das Gesetz der kleinen Zahlen, *Human Biology* **19**, 154–161.

PETER ARMITAGE

Bortkiewicz, Ladislaus von

PETER ARMITAGE

Volume 1, pp. 548–548

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Bradford Hill Lectures

Sir Austin Bradford **Hill** has been described as "the greatest medical statistician in the twentieth century" [1]. He pioneered the randomized controlled trial as *the* method of evaluating new treatments (*see* **Clinical Trials, Overview**) [3, 4], established criteria for evaluating evidence of **causation** in **epidemiology** [5], and collaborated on many key medical and public health research studies, both experimental and observational. Notably, the cohort study of British doctors [2] led by Sir Austin and Sir Richard Doll provided the first definitive evidence linking cigarette **smoking** with risks of lung cancer and coronary heart disease, and the British **Medical Research Council Streptomycin trial** in tuberculosis [6] is generally considered to be the first well-executed randomized clinical trial.

Sir Austin's genius lay in creating a collaborative spirit whereby statistical skills became truly appreciated by medical researchers. His key text "Principles of Medical Statistics", first published in 1937 as a series of *Lancet* articles, is still today considered essential reading; its enlarged 12th edition [7] with his son David Hill as joint author appeared shortly after Sir Austin's death in 1991.

Much of Sir Austin's career was at the London School of Hygiene and Tropical Medicine, where he was Professor of Medical Statistics from 1946 to 1961 and Dean of the School from 1955 to 1957. His legacy has led to the School being one of the world's leading academic centers for medical statistics and epidemiology both in research and teaching. To honor his memory, the School established the Bradford Hill Memorial Lecture series, which each year provides an opportunity for a leader in the field to give a perspective on our discipline and its relevance to key public health issues. The Bradford Hill Memorial lecturers have been as follows:

1992 Richard Doll
Sir Austin Bradford Hill and the progress of medical science

1993 Peter Armitage
Before and after Bradford Hill: Some trends in Medical Statistics

1994 David Cox
Causality

1995 Richard Peto
Clinical trials: where Bradford Hill went wrong

1996 Nick Day
Quantification and causality in epidemiology: beyond Bradford Hill

1997 Iain Chalmers
The Cochrane Collaboration: Problems, achievements and prospects

1998 Sheila Gore
Drugs, illegal addiction; High time for Bradford Hill's scientific method

1999 Richard Horton
Common sense and figures: the rhetoric of validity in medicine

2000 David Clayton
Biostatistics, Epidemiology and the post-Genomic Challenge

2001 David Strachan
The environment and disease: association & causation across three centuries

2002 David Spiegelhalter
Monitoring and comparing clinical performance – do we need 'clever' statistical methods?

2003 Valerie Beral
The causes of breast cancer

2004 Janet Darbyshire
Challenging Trials

References

[1] Doll, R. (1992). Sir Austin Bradford Hill and the progress of medical science, *British Medical Journals* **305**, 1521–1526.

[2] Doll, R. & Hill, A.B. (1954). The mortality of doctors in relation to their smoking habits, *British Medical Journals* **I**, 1451–1455.

[3] Hill, A.B. (1952). The clinical trial, *New England Journal of Medicine* **247**, 113–119.

[4] Hill, A.B. (1963). Medical ethics and controlled trials, *British Medical Journals* **I**, 1043.

[5] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[6] Hill, A.B. (1990). Memories of the British streptomycin trial in tuberculosis, *Controlled Clinical Trials* **11**, 77–79.

[7] Hill, A.B. & Hill, I.D. (1991). *Bradford Hill's Principles of Medical Statistics*, 12th Ed. Edward Arnold, London.

STUART J POCOCK

Bradley, Ralph A.

Born: November 28, 1923 in Smith Falls, Ontario, Canada.
Died: October 30, 2001, in Athens, Georgia, USA.

Photo provided by Marion Bradley

Ralph Bradley's contributions to the world of statistics fall under two headings: his statistical research (especially the Bradley–Terry test used extensively in taste-testing experiments) and his professional leadership role in statistical science, as evidenced by his development of statistical programs, by his presidency (1981) of the American Statistical Association (ASA) and by his editorial efforts. The conversation in Statistical Science [5] provides more details of his views and life.

Ralph Bradley grew up in Wellington, Ontario, and received his B.A. degree with honors in mathematics and physics from Queens University, Canada, in 1944. After serving in the Canadian army (1944–1945), he completed an M.A. in mathematics and statistics in 1946, also at Queens University. He completed his Ph.D. degree in statistics in 1949 at the University of North Carolina at Chapel Hill under the direction of Harold **Hotelling** as his major professor. After a year at McGill University in Montreal, Canada, Ralph returned to the United States and joined the new department of statistics at Virginia Polytechnic Institute (VPI) under Dr Boyd Harshbarger. Ralph stayed in the department for 10 years (1950–1959), became Boyd's right-hand man and with him laid the foundations and developed the department. Building upon this experience, Ralph traveled further south from his beloved Canada to Tallahassee, Florida, to become the head of a new statistics department at Florida State University (FSU). In the ensuing 19 years (1959–1978) as head, he developed the department and made it first-rate. He stayed another four years at Florida State before moving to the University of Georgia in 1982, where he remained until his death, first as a research professor (1982–1992) and then as research professor emeritus after his retirement.

While at VPI, Ralph Bradley collaborated with Milton Terry on a statistical test for paired comparisons, which became known as the **Bradley–Terry** test. The test involves ranking a series of two (of a total of t) treatments in **incomplete blocks** of size two. The test was motivated by a problem encountered when judges were asked to assess the relative merits of t treatments, making the comparisons two treatments at a time and expressing a preference for one treatment in each pair of two treatments (e.g. taste testing). Bradley and Terry [4] developed a model and the related **hypothesis testing** procedures, including developing rating estimates, and also conquered the not inconsiderable computational difficulties (in the days long before modern computing capabilities were freely available). This work was widely cited and led to several awards, including the 1957 J. Shelton Horsley Research Award. An elegant exposition of the universality of the application of the test was given in [6], which is an excellent illustration of Bradley's commitment to the importance of both theoretical and applied statistics in the advancement of statistical science. This balancing of theory and application permeated his research throughout his career, culminating with his work in meteorology in the late 1970s to the early 1980s [2] and his work on trend-free block designs and **blocking** criteria in later years (see, e.g. [3, 7]). Ralph Bradley had broad interests in statistics throughout his research career, including **nonparametric** statistics, **sequential analysis**, computational methods, design of experiments (*see* **Experimental Design**), and **multivariate** methods.

Though Ralph made important contributions to statistical research, his leadership in the development of statistics departments distinguished him from his peers. He played a key role in the growth of the department at VPI; he was the driving force of the department of statistics at FSU, as the founding head; and he provided senior leadership at the University

of Georgia. He wrote extensively on how to achieve and develop viable degree programs (see, e.g. [1]). His career was also distinguished by his editorial efforts, highlighted by his 6-year term (1957–1962) as an editor of **Biometrics** and his 44-year term (1954–1998) as a consulting editor of the Wiley Series in probability and statistics. He contributed mightily to these high-quality statistical publications.

Bradley's leadership capabilities were recognized by his election as President (1981) of the **American Statistical Association** (ASA), a singular honor recognizing his leadership skills and insights in guiding the future development of the profession. A major contribution was his leadership role in ASA's purchasing and moving to its own building in Alexandria, Virginia. Earlier, Bradley served the ASA in a variety of roles, including as vice president, as a member of the board of directors, and as a member of the publication committee. His overall contributions were recognized in 1992 with a Founders Award. Bradley's other contribution to the profession included being president (1965) of the Eastern North American Region of the **International Biometric Society** (IBS), with three terms (over three decades) on the IBS council, and a term as chair and member (1982–1988) of the National Research Council's Committee on Applied and Theoretical Statistics.

A mark of a true leader includes a willingness to seek advice and learn from the experiences of other leaders, and Ralph Bradley was such a leader. He often looked to Gertrude **Cox**, Boyd Harshbarger, Harold Hotelling, and Frank **Wilcoxon**, among others, for inspiration and advice. He understood and practiced the art of collaboration, especially as it pertained to professional leadership. The marks of his leadership are visible in the establishments of departments and degree programs not only where he held rank but also where he played related advisory roles. Further, he sought to instill in others the same ideals of cross-communication between colleagues and aided the development of leadership abilities when the potential excited him. His insistence on excellence from himself and those around him was a thread throughout his work that enhanced his contributions and defined his accomplishments.

Ralph's dedication to perfectionism enabled him to be a craftsman par excellence. He built a swivel chair from discarded parts of broken chairs, and he also built a roll-top desk. This involved the craftsmanship and patience that was typical of Ralph. He loved to fish; he loved to play bridge, even maintaining a noon-time game (maximum playing time 70 minutes!) with departmental faculty at FSU; and he loved sports, especially tennis, never allowing age to deter his enthusiasm for playing as hard as he could. His family supported him fully; his wife Marion was tireless, serving, for example, as his editorial assistant during his *Biometrics* editorship. He has two children Allan and Linda and four grandchildren. Ralph Bradley was an all-around man! He has left behind a rich and enduring legacy to the statistical profession, a legacy to be embraced and enjoyed by those who follow him.

References

[1] Bradley, R.A. (1993). The statistics profession – Some questions for the future, *Bulletin International Statistical Institute* **49**, 49–66.

[2] Bradley, R.A., Srivastava, S. & Lanzdorf, A. (1979). Some approaches to statistical analysis of a weather modification experiment, *Communications in Statistics – Theory and Methods* **A8**, 1049–1081.

[3] Bradley, R.A. & Stewart, F. (1991). Intrablock analysis of designs with multiple blocking criteria, *Journal of the American Statistical Association* **86**, 792–797.

[4] Bradley, R.A. & Terry, M.E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons, *Biometrika* **39**, 324–345.

[5] Hollander, M. (2000). A conversation with Ralph A. Bradley, *Statistical Science* **16**, 75–100.

[6] Terry, M.E., Bradley, R.A. & Davis, L.L. (1952). New designs and techniques for organoleptic testing, *Food Technology* **6**, 250–254.

[7] Yeh, C.-M., Bradley, R.A. & Notz, W.I. (1985). Nearly trend-free block designs, *Journal of the American Statistical Association* **80**, 985–992.

LYNNE BILLARD

Bradley–Terry Model

The Bradley–Terry model is an elegant unidimensional scaling method for summarizing purely ordinal data on paired "objects" (*see* **Paired Comparisons**). Such data are ubiquitous in the match-by-match records of individual and team sports involving paired competition. In biometry they are intrinsic to analyses of dominance behaviors in some animal species, and occur more generally in the study of subjective or objective phenomena that are not directly measurable, therefore requiring that objects be scaled by comparative ranking. Paired comparisons are used when it is psychologically impracticable or burdensome to rank triads or larger groups, or when operational difficulties may compromise validity or reliability of such rankings.

The Basic Model

Consider a set of objects $\mathcal{O} = \{O_i, i = 1, \ldots, I\}$, from among which certain pairs (O_j, O_k) are ordered, either by nature or by one or more observers ("judges"), as either $(O_j > O_k)$ or $(O_j < O_k)$. The judgments are presumed to be guided by the same criterion, which may be frankly subjective (e.g. judge's personal preference) or invoke a clearly objective quality. Pairs may be judged more than once, with the same or different outcomes. The Bradley–Terry model associates with each O_i a parameter $\lambda_i > 0$. Any single comparison between O_j and O_k is modeled as a Bernoulli trial with probabilities of the two outcomes $(O_j > O_k)$ and $(O_j < O_k)$ proportional to λ_j and λ_k (*see* **Binary Data**). Thus, the **odds** of O_j being "preferred" to O_k are λ_j/λ_k, and

$$\pi_{jk} = \Pr\{O_j > O_k\} = \frac{\lambda_j}{(\lambda_j + \lambda_k)}. \qquad (1)$$

For **identifiability** the λ_i are conventionally scaled to sum to 1, so that they satisfy the most basic formal properties of probabilities. However, only their ratios are consequential and interpretation as probabilities is generally inappropriate.

The model was proposed by Zermelo [41] with a view to resolving incomplete round-robin tournaments, and independently developed by Bradley & Terry in the early 1950s while studying sensory difference testing methods for evaluating food quality [12, 39]. It has since been widely applied to biometric and psychometric problems. Ford [27], who also rediscovered the model, clarified aspects of the asymptotic theory and convergence properties of algorithms for **maximum likelihood** estimation (*see* **Large-sample Theory**).

Guided by the principle of "independence of irrelevant alternatives", Luce [33] proposed a general model for choices among groups of objects that reduces to the Bradley–Terry model when the objects are judged only in pairs. Luce's model posits that the probabilities of ranking O_j or O_k first have a constant ratio λ_j/λ_k, whether O_j and O_k are being compared as a pair, or are included in any larger subset of $\mathcal{O}$. Since $\mathcal{O}$ itself is such a subset, in the Luce model λ_i actually is the probability that O_i is ranked first amongst the I objects. Thus, although it is applicable to complex data structures beyond paired comparisons, the Luce model incorporates very restrictive assumptions about the choice process.

Table 1 relates the Bradley–Terry model to antecedents in (i) the tradition of **psychometric** paired-comparison (or paired-choice) research dating back to psychophysical studies of Fechner [36], and (ii) the **quantal response bioassay** literature of the 1930s and 1940s that proved seminal to modern **categorical data** modeling [26].

To see the relationships, consider the general linear paired-comparison model [15, pp. 7–9]. This represents the π_{jk} as differences between intrinsic "true merits" M_j and M_k of O_j and O_k on an underlying continuum, transformed by a continuous and symmetric cumulative distribution function (cdf) F centered at zero. Thus,

$$\pi_{jk} = F(M_j - M_k). \qquad (2)$$

The merits constitute a latent unidimensional interval scaling of the objects, and form the structural portion of the linear model. The stochastic component

Table 1 Relationship of the Bradley–Terry model to analogous paired comparison and quantal bioassay methods

	Latent distribution	
	Gaussian	Logistic
Paired comparison	Thurstone–Mosteller	Bradley–Terry
Quantal bioassay	Probit analysis	Logit analysis

F follows from assuming that a paired-comparison judge observes M_j and M_k only after perturbation by additive random errors ε_j and ε_k. These errors vary across judges, observation times, or both, as appropriate to the situation, following a single continuous **bivariate distribution** for all pairs. If judges order pairs using the resulting observed merits $M_j + \varepsilon_j$ and $M_k + \varepsilon_k$, then the linear model follows by taking F as the marginal distribution of $\varepsilon_j - \varepsilon_k$.

The Gaussian (**normal**) and **logistic distributions** are natural candidates for F, the former yielding the classical Thurstone–Mosteller psychometric model [38], and the latter the Bradley–Terry model (for which one may take $M_i = \ln \lambda_i$). These are related, respectively, to probit and logit models for quantal bioassay, wherein the minimal drug dose required to produce a well-defined response in a test animal, when measured on an appropriate scale, is presumed to follow a normal (probit model) or logistic (logit model) distribution within the animal population. Both conceptually and formally, the difference in true merits in a Bradley–Terry model plays the same role as does the log dose in logit modeling of a quantal bioassay (*see* **Quantal Response Models**).

The Thurstone–Mosteller model arises simply from a **bivariate normal distribution** of merit perturbations, and so is a comfortable choice if a linear paired-comparison model is to be selected on the basis of first principles. In contrast, the Bradley–Terry model has the unappealing property that independent merit perturbations within pairs are type I **extreme value** (double-exponential) variates. Thompson & Singh [37] generate this distribution from an underlying psychophysical model in which perturbations are maxima rather than means of many independent identically distributed components. Despite this apparently substantial difference, however, the Gaussian and logistic distributions of perturbation differences are virtually indistinguishable in practice. Hence, the choice between them is rarely consequential for substantive inference.

In another respect, the Bradley–Terry model is most appealing. It is common in a tournament or other paired-comparison setting to rank the objects by their total "victory" counts, v_i. In a balanced round-robin tournament, $\mathrm{E}(v_i) = r\pi_i. = \sum_{j \neq i} \pi_{ij}$, where r is the number of replications. In any linear paired-comparison model, H_0: equality of these $\pi_{i\cdot}$ implies H_0': equality of the M_i. Hence, indifference may be tested using the v_i. But when some M_i are unequal, formal modeling may not support the naive ordering that the v_i suggest, especially when data arise from incomplete tournaments. Nevertheless, inference about the M_i is considerably simplified if one confines attention to these marginal scores. Among all linear paired-comparison models, only the Bradley–Terry model allows this. The v_i are **sufficient** for the M_i under this model, so one can always find the maximum likelihood ordering from the v_i, even when they do not themselves display it. Daniels [14] shows that the Bradley–Terry model also arises directly from some intuitively appealing approaches to "fair" scoring of round-robin tournaments.

Inference

Independent Choices

Inference is greatly simplified when all choices are viewed as statistically independent. This occurs (i) in the rare situation of randomly chosen judges who each separately view a single pair, or (ii) when no comparisons are influenced by common random effects, or by the results of any other comparisons. It is sometimes reasonable to assume independence under other circumstances, at least to a first approximation.

In this situation, data may be summarized without loss of information using either of two **contingency table** formats. We illustrate these using data from Davidson & Bradley [20] on overall quality of three chocolate puddings. In the "population-response" format (Table 2), each comparison is counted in the cell of the row labeled by the pair and the column labeled by the winner.

In the "repeated measures" format (Table 3), each comparison is counted in the cell of the winner's

Table 2 Population–response representation of paired-comparison results on overall quality of chocolate puddings A, C, and D

	Better			
Pair	A	C	D	Total
AC	10	10		20
AD	7		15	22
CD		14	9	23
Total	17	24	24	65

Table 3 Repeated-measures representation of paired-comparison results on overall quality of chocolate puddings A, C, and D

		Worse			
		A	C	D	Total
Better	A		10	7	17
	C	10		14	24
	D	15	9		24
	Total	25	19	21	65

row and the loser's column. Tables 2 and 3 are both "incomplete" contingency tables, inasmuch as a diagonal of each must be empty, by nature of the paired-comparison design. The Bradley–Terry model is equivalent to both the **quasi-independence** model for the nonzero cell counts of Table 2 and the **quasi-symmetry** model for the nonzero cell counts of Table 3 [25, 30]. These are **loglinear models**, equivalently expressible as **logistic regression** models or, in a broader context, as **generalized linear models** (GLMs) with a logit link function.

Under either the quasi-independence or the quasi-symmetry model, the row and column marginal totals of Table 1 (equivalently, the $n_{jk} + n_{kj}$ and the row sums from Table 2) form a complete set of sufficient statistics for the λ_i. If the design is connected, in that there is no proper subset of $\mathcal{O}$ whose members are compared only to each other, then maximum likelihood estimates of the λ_i, M_i, π_{jk}, and all contrasts among them may be obtained by algorithms employing **iterative proportional fitting** (IPF), the Newton–Raphson method, or Fisher's method of scoring (*see* **Optimization and Nonlinear Equations**). Essentially any **software** for loglinear contingency table analysis, multiple logistic regression, or generalized linear models may be used.

When the fitted cell counts for a counterpart of Table 2 and Table 3 are sufficiently large to validate asymptotic inference, the usual maximum likelihood machinery may be applied to obtain confidence intervals for linear contrasts among the π_{jk}, λ_i, or M_i, and for associated **hypothesis testing**. Other best asymptotically normal **(BAN) estimators** and asymptotically efficient test statistics (e.g. score and Wald statistics; *see* **Likelihood**) may also be used, and a locally asymptotically most stringent test of equivalence is available [4]. For example, the maximum likelihood estimates of λ_A, λ_C and λ_D from the chocolate pudding data in Tables 2 and 3, with estimated asymptotic standard errors, are $\hat{\lambda}_A = 0.253 \pm 0.060$, $\hat{\lambda}_C = 0.387 \pm 0.073$, and $\hat{\lambda}_D = 0.360 \pm 0.070$.

When the likelihood approach is taken and the number of choices between each pair in the design is not small, the deviance **chi-square statistic** (*see* **Generalized Linear Model**) based on the fitted counts may be used to test the fit of the Bradley–Terry model. The deviance for the pudding data is 2.50 with one df ($P = 0.11$), consistent with adequate fit. Parameters or parameter sets may be compared across groups of judges by incorporating appropriate nested or interaction terms into the corresponding logistic regression model.

Bradley & Terry [39] suggest that 15 judgments per pair are needed in a balanced design before the large-sample χ^2 approximation is satisfactory for the distribution of the **likelihood ratio** statistic for testing equivalence, $H_0 : \lambda_i = I^{-1}, i = 1, \ldots, I$. With smaller samples the approximation is unacceptably liberal, yielding higher than nominal type I error (*see* **Level of a Test**). While the small-sample distributions of this and related test statistics are generally analytically intractable and tedious to compute, tables have been prepared for common hypotheses and balanced designs [8, 12]. Regardless of balance, fast computational algorithms for exact small-sample logistic regression analysis may now also be applied to the multiple logistic regression formulation of a Bradley–Terry model, provided that the number of objects is not so large as to exceed available computational resources [29] (*see* **Exact Inference for Categorical Data**).

Correlated Choices

Inference is more challenging in the context of correlated choices, which generally occur due to heterogeneity of judges or of conditions under which comparisons occur. Evidence for such correlation may appear in the form of **overdispersion**, as reflected by an estimated scale parameter substantially greater than unity in a GLM fit. Such overdispersion produces liberality in hypothesis tests, spuriously narrowing confidence intervals and exaggerating the lack of fit. Corrections for overdispersion based on the scaled deviance do not

account for the underlying correlation structure. Several of the available approaches when correlations arise only from random judge effects are described briefly below.

For the case when each judge rates all pairs, the judges may be classified by their response patterns into an $I^{I(I-1)/2}$ sparse incomplete contingency table. Iterative algorithms for constrained optimization under **Poisson** sampling may then be used to obtain maximum likelihood estimates for a Bradley–Terry model of the marginal distributions of this table [3]. Happily, there is no need to model the nuisance correlation structure induced by the judge effects. Or, suppose each judge ranks a subset of all possible pairs, with the number of such subsets in the design small relative to the number of judges. Then the response pattern of each judge may be similarly classified into a contingency table whose margins are the pairs ranked by that judge. We may thus use joint weighted **least-squares** analysis of the sets of correlated marginal logits from these tables [30]. Covariances of choices are estimated by unrestricted maximum likelihood.

Generalized estimating equation (GEE) methods may be employed conveniently when all judges rank the same set of pairs. This approach can accommodate **covariates** associated with the judges and objects. Simple GEE (GEE1) with an independence "working" covariance structure may be used, with inference from a robust covariance estimator to account for the true association structure. Extended GEE (GEE2), which seeks increased efficiency in estimators of the M_i by modeling the association structure, may also be employed. Caution is indicated, however, as resulting estimators $\hat{M}_i$ may be biased if the association model is misspecified.

Prior distributions may also be introduced to model heterogeneity. The most natural approach, though computationally burdensome, is through the generalized linear mixed model (GLMM) [13], which introduces additive $\mathrm{MN}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ **random effects** on the M_i. Just as the random perturbations in the linear paired comparison model, these random effects produce multipliers of the choice odds λ_j/λ_k, yielding in essence a **variance components** model for the perturbations on the latent merit scale. The above use of multivariate Gaussian random effects on the M_i is traceable to the full **Bayesian** treatment of Leonard [32].

Design

Several workers, beginning with Abelson & Bradley [1], have considered **experimental design** aspects of paired comparison studies. The merits are usually presumed linear in the **explanatory variables**. Where the O_i embody factorial combinations of classification variables, factorial modeling is obviously desirable (*see* **Factorial Experiments**). More generally, **response surface** fitting and associated efficient designs are of interest when the objects are distinguished by interval-scaled variables (*see* **Measurement Scale**). El-Helbawy [24] reviews much of this body of work. Dillon et al. [23] consider the simultaneous estimation of Bradley–Terry parameters and classification of judges into latent groups, in the presence of covariates. David & Andrews [16] and Bhandari et al. [6] give sequential procedures for selecting the best treatment under a Bradley–Terry model. Advances in estimation for correlated choices and mixed models, mentioned in passing above, will likely stimulate more extensive work on efficient design.

Extensions

Ties and Preference Strengths

The Bradley–Terry model has been generalized to allow for a third "tied" outcome category, to extend further to multicategory ordinal judgments, to incorporate effects of the order in which objects are presented, and to include correlated judgments on multiple criteria.

Following the work of Glenn & David [28] on the Thurstone–Mosteller model, Rao & Kupper [35] extended the Bradley–Terry model by introducing a threshold of perception below which objects appear indistinguishable, and must be declared ties. Suppose objects differing in perceived merits by no more than τ are judged indistinguishable. Writing $\theta = \exp(\tau)$ and $\pi_{j=k}$ for the probability of a tie between O_j and O_k yields

$$\pi_{jk} = F(M_j - M_k - \tau) \tag{3}$$

for the general linear paired-comparison model, and

$$\pi_{jk} = \frac{\lambda_j}{(\lambda_j + \theta\lambda_k)} \tag{4}$$

for the Bradley–Terry model, with $\pi_{j=k} = 1 - \pi_{jk} - \pi_{kj}$ in each case.

In this model the ratio π_{jk}/π_{kj} no longer equals λ_j/λ_k, and depends on the discrimination threshold τ. Davidson [17] gives an alternate single-parameter model for ties that preserves the straightforward connection of π_{jk}/π_{kj} to the merits, at the price of assuming that $\pi_{j=k}$ is proportional to the geometric **mean** of π_{jk} and π_{kj}. The threshold parameter, τ, is replaced by a proportionality constant, ν, which must also be estimated.

The Bradley–Terry model assumes that O_j is preferred to O_k when the perceived difference in merits exceeds a "cut-point" of zero. Agresti [3] generalizes the Bradley–Terry and Rao–Kupper models to an arbitrary number of ordered response categories (e.g. much worse, worse, equal, better, much better) by assuming multiple cut-points for the perceived difference, producing an indifference zone and increasing degrees of preference (*see* **Ordered Categorical Data**). This leads in essence to simultaneous Bradley–Terry models for odds-ratios (i) comparing the perception that O_j is much better than O_k with the perception that O_k is much better than O_j, (ii) comparing the perception that O_j is at least better than O_k with the perception that O_k is at least better than O_j, and so on. These are proportional cumulative odds models in the spirit of McCullagh [34] and Walker & Duncan [40] (*see* **Polytomous Data**). The same structural models may be applied to "local" odds ratios formed after conditioning on a pair of adjacent response categories, e.g. the odds of considering O_j better vs. equal to O_k, relative to the odds of considering O_k better vs. equal to O_j. Maximum likelihood analysis of such models is straightforward [2].

Presentation Order

Models involving both additive and multiplicative adjustments have been proposed to incorporate a presentation order effect into the Bradley–Terry model. The additive approach assumes that a fixed proportion of choices, specific to each pair, is swayed by the order of presentation. If π^*_{jk} is the probability of choosing j over k given a specific order of presentation, then under this model

$$\pi^*_{jk} = \pi_{jk} \pm \delta_{jk}, \qquad \delta_{jk} \leq \min(\pi_{jk}, \pi_{kj}), \qquad (5)$$

depending on whether O_j is presented respectively first or second [5]. Thus, the range of admissible δ_{jk} depends on λ_j and λ_k.

The **multiplicative model** [18] applies the same concept to the log(odds) rather than the choice probabilities. In this case,

$$\ln\left(\frac{\pi_{jk}}{\pi_{kj}}\right) = \ln\lambda_j - \ln\lambda_k \pm \gamma_{jk}, \quad \gamma_{jk} > 0, \qquad (6)$$

depending, respectively, on whether O_j is presented second or first. In contrast to the additive model, simplifications such as $\gamma_{jk} = \gamma$, or intermediate modeling, are straightforward. In a particularly interesting earlier paper, Kousgaard [31] discusses models incorporating both ties and order effects, where the discrimination threshold and order effect may vary across judges. Conditional maximum likelihood inference is proposed (*see* **Logistic Regression, Conditional**).

Multiple Criteria

The Bradley–Terry model may be applied separately to choices using c different criteria, or under different conditions. This requires accounting for the correlation structure among multiple binary responses, for which direct modeling is difficult. Davidson & Bradley [19] propose a direct model in which the probability of a response pattern under independent Bradley–Terry selections is perturbed by a function of the underlying marginal preference parameters and $c(c-1)/2$ correlation parameters generalizing Pearson's ϕ coefficient for **two-by-two contingency tables** (*see* **Association, Measures of**). However, since the association structure is generally a nuisance rather than an object of inference itself, much success has been gained by large-sample approaches such as weighted least squares that adjust for the covariance structure without modeling it [30], or that incorporate covariance modeling into **robust** estimation of marginal parameters, such as GEE. In general, the large-sample techniques discussed above for handling correlated responses in univariate paired-comparison studies may be similarly applied to handle multivariate responses, sample size permitting.

The literature on Bradley–Terry models is large, as suggested by the comprehensive bibliographies of Davidson & Farquhar [21] and David [15]. The

reader is referred to the reviews by Bradley [9–11] and the classic monographs of David [15] and Bock & Jones [7]. Diggle et al.'s [22] discussion of recent work on **correlated binary** regression, though written in the context of **longitudinal data analysis** rather than paired comparisons, is useful for the reader interested in modeling correlated and/or multivariate choices.

References

[1] Abelson, R.M. & Bradley, R.A. (1954). A 2 (2 factorial with paired comparisons, *Biometrics* **10**, 487–502.

[2] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York, pp. 261–305.

[3] Agresti, A. (1992). Analysis of ordinal paired comparison data, *Applied Statistics* **41**, 287–297.

[4] Beaver, R.J. (1974). Locally asymptotically most stringent tests for paired comparison experiments, *Journal of the American Statistical Association* **69**, 423–427.

[5] Beaver, R.J. & Gokhale, D.V. (1975). A model to incorporate within-pair order effects in paired comparisons, *Communications in Statistics* **4**, 923–939.

[6] Bhandari, S.K., Hande, S.N. & Ali, M.M. (1993). An optimal sequential procedure for ranking pairwise compared treatments, *Calcutta Statistical Association Bulletin* **27**, 191–197.

[7] Bock, R.D. & Jones, L.V. (1968). *The Measurement and Prediction of Judgment and Choice*. Holden-Day, San Francisco.

[8] Bradley, R.A. (1954). The rank analysis of incomplete block designs. II. Additional tables for the method of paired comparisons, *Biometrika* **41**, 502–537.

[9] Bradley, R.A. (1976). Science, statistics, and paired comparisons (with discussion), *Biometrics* **32**, 213–232.

[10] Bradley, R.A. (1984). Paired comparisons: some basic procedures and examples, in *Handbook of Statistics*, Vol. 4, P.R. Krishnaiah & P.K. Sen, eds. North-Holland, Amsterdam, pp. 299–326.

[11] Bradley, R.A. (1985). Paired comparisons, in *Encylopedia of Statistical Sciences*, Vol. 6, S. Kotz, N.L. Johnson & C. Read, eds. Wiley, New York, pp. 555–560.

[12] Bradley, R.A. & Terry, M.B. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons, *Biometrika* **39**, 324–345.

[13] Breslow, N. & Clayton, D.G. (1993). Approximate inference in generalized linear models, *Journal of the American Statistical Association* **88**, 9–25.

[14] Daniels, H.E. (1969). Round-robin tournament scores, *Biometrika* **56**, 295–299.

[15] David, H.A. (1988). *The Method of Paired Comparisons*, 2nd Ed. Wiley, New York.

[16] David, H.A. & Andrews, D.M. (1987). Closed adaptive sequential paired-comparison selection procedures, *Journal of Statistical Computation and Simulation* **27**, 127–141.

[17] Davidson, R.R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments, *Journal of the American Statistical Association* **65**, 317–328.

[18] Davidson, R.R. & Beaver, R.J. (1977). On extending the Bradley-Terry model to incorporate within-pair order effects, *Biometrics* **33**, 693–702.

[19] Davidson, R.R. & Bradley, R.A. (1969). Multivariate paired comparisons: the extension of a univariate model and associated estimation and test procedures, *Biometrika* **56**, 81–95.

[20] Davidson, R.R. & Bradley, R.A. (1971). A regression relationship for multivariate paired comparisons, *Biometrika* **58**, 555–560.

[21] Davidson, R.R. & Farquhar, P.H. (1976). A bibliography on the method of paired comparisons, *Biometrics* **32**, 233–240.

[22] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon, Oxford, pp. 146–189.

[23] Dillon, W.R., Kumar, A. & de Borrero, M.S. (1993). Capturing individual differences in paired comparisons.: an extended BTL model incorporating descriptor variables, *Journal of Marketing Research* **30**, 42–51.

[24] El-Helbawy, A.T. (1992). Optimal paired comparison designs, in *Order Statistics and Nonparametrics: Theory and Applications*, P.K. Sen & I.A. Salama, eds. North-Holland, Amsterdam, pp. 349–361.

[25] Fienberg, S.E. & Larntz, K. (1976). Log linear representation for paired and multiple comparisons models, *Biometrika* **63**, 245–254.

[26] Finney, D.J. (1978). *Statistical Method in Biological Assay*, 3rd Ed. Griffin, London, pp. 349–403.

[27] Ford, L.R., Jr (1957). Solution of a ranking problem from binary comparisons, *American Mathematical Monthly* **64**, 28–33.

[28] Glenn, W.A. & David, H.A. (1960). Ties in paired-comparison experiments using a modified Thurstone-Mosteller model, *Biometrics* **16**, 86–109.

[29] Hirji, K.F., Mehta, C.R. & Patel, N.R. (1987). Computing distributions for exact logistic regression, *Journal of the American Statistical Association* **82**, 1110–1117.

[30] Imrey, P.B., Johnson, W.D. & Koch, G.G. (1976). An incomplete contingency table approach to paired-comparison experiments, *Journal of the American Statistical Association* **71**, 614–623.

[31] Kousgaard, N. (1976). Models for paired comparisons with ties, *Scandinavian Journal of Statistics* **3**, 1–14.

[32] Leonard, T. (1977). An alternative Bayesian approach to the Bradley-Terry model for paired comparisons, *Biometrics* **33**, 121–132.

[33] Luce, R.D. (1959). *Individual Choice Behavior*. Wiley, New York, pp. 1–37.

[34] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[35] Rao, P.V. & Kupper, L.L. (1967). Ties in paired-comparison experiments: a generalization of the Bradley-Terry model, *Journal of the American Statistical Association* **62**, 194–204.

[36] Stigler, S.S. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, Mass, pp. 239–254.

[37] Thompson, W.A., Jr & Singh, J. (1967). The use of limit theorems in paired comparison model building, *Psychometrika* **32**, 255–264.

[38] Thurstone, L.L. (1927). Psychophysical analysis, *American Journal of Psychology* **38**, 368–389.

[39] Virginia Agricultural Experiment Station (1951). *Statistical Methods for Sensory Difference Tests of Food Quality: Bi-Annual Report No. 2*. Virginia Agricultural Experiment Station, Blacksburg.

[40] Walker, S.H. & Duncan, D.B. (1967). Estimation of the probability of an event as a function of several independent variables, *Biometrika* **54**, 167–179.

[41] Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift* **29**, 436–460.

PETER B. IMREY

Branching Processes

A branching process is a description of the evolution of populations of individuals, which reproduce independently. In the most general cases, individuals inherit a *type* from some type space at birth; this type determines a probability measure on a set of possible life careers, which in their turn determine a **point process** giving the ages of bearing and the types of the children.

There is a whole array of different branching processes:

1. In classical (*Bienaymé-)* **Galton–Watson**, or *simple*, processes, there is only one type of individual and all individuals live for one time unit (season), giving birth to independent, identically distributed (iid) numbers of children living through the next season.
2. In Bellman–Harris, or *age-dependent*, processes, this pattern is generalized so that individuals can have iid lifespans according to an arbitrary distribution. At death they split into iid numbers of offspring, and so forth.
3. A special case of these are the *Markov branching* processes, obtained by choosing the lifespan distributions to be **exponential**.
4. *Sevastyanov* processes generalize Bellman–Harris processes by allowing dependence between lifespan and offspring numbers.
5. In *general* processes, births need no longer occur only at the end of a lifetime, but could be spread out randomly according to any point process.
6. All of these populations exist in single and multitype versions. In much literature the latter term indicates that there is a finite number of different types of individuals, but generally this need not be the case. In particular, much mathematically oriented literature deals with populations in which "type" is position in space. Biologically, "type" could be **genotype**; but it could also be a property such as mass or DNA content at birth.

Historically, branching processes originate from biologically or **demographically** motivated problems. There are, however, many other natural phenomena exhibiting a branching character, such as cascades of splitting particles, or polymerization processes. Lately, branching processes have been used in the study of **algorithm** performance. They occur in the formation of random fractal sets (*see* **Chaos Theory**). Limits of branching processes form so-called superprocesses, measure-valued random processes that are used to interpret nonlinear differential equations. The area thus also contains many aspects of pure mathematics, both analytic and more probabilistic.

The first problem of branching processes proper was that of determining extinction probabilities. In the single-type Galton–Watson case, this can be formulated and solved as follows.

If q denotes the probability that a population that started from one ancestor dies out, then, by independence (and since all individuals are the same type), the probability that a population with k ancestors dies out must be q^k. Hence if p_k is the probability of begetting k children, then

$$q = \sum_k p_k q^k.$$

In the interval [0, 1] this equation has exactly one solution, namely $q = 1$, if $m := \sum_k k p_k < 1$, or $m = 1$ and $p_1 < 1$, the *subcritical* and *critical* cases. It has one more solution if $m > 1$, and this is the real extinction probability in that *supercritical* case. The extinction problem of more general branching processes can be reduced to the Galton–Watson case through counting the numbers of individuals in successive generations.

In the subcritical and critical cases, it is of interest to study the conditional distribution of the population size, given that the population has not yet died out. For both of these cases limit theorems, as time passes, exist. In the subcritical case, it is the population size itself that has a limiting distribution; in the critical case, it behaves like the time since the population started multiplied by an exponentially distributed random variable.

In the cases in which extinction is not necessary, one can prove that nonextinction implies growth beyond all limits, at the exponential rate already argued by **Malthus**. The exponent in this exponential growth is called the *Malthusian parameter*, and it is determined by the reproduction laws of individuals of the various types. In branching processes it is usually denoted by α, and in deterministic population dynamics by r.

To see how the parameter is determined, consider one-type general processes and let $\mu(t)$ denote the expected number of children obtained by a t-aged

mother. Then the Malthusian parameter is defined by the equation

$$\int_0^\infty \exp(-\alpha t)\mu(\,\mathrm{d}t) = 1,$$

which usually has one (and only one) solution. The sub- and supercritical cases correspond to $\alpha < 0$ and $\alpha > 0$, and the critical to $\alpha = 0$. In multitype populations it is a more complicated object that should equal one, the so-called Perron root of the kernel, describing the expected numbers of children of various types from mothers of various types.

If z_t denotes population size at time t in a supercritical process, Malthus' law of growth then reads

$$z_t \sim cw \exp(\alpha t), \quad t \to \infty,$$

where c is a constant determined by the way in which population size is measured and w is a random variable which is not identically zero, if the individual offspring distributions have a finite first and logarithmic moment; that is, a condition of the form $\mathrm{E}[X \log^+ X] < \infty$ holds, X being basically the number of children per individual.

During the exponential growth the age distribution will stabilize to a distribution also determined by the individual reproduction distributions (a fact already known to Euler). Similarly, other aspects of population composition will stabilize, so that the properties of a typical individual (i.e. one sampled at random) converge, while time passes and the population grows. A particular aspect of this is the emergence of a typical history of individuals; for example, a typical mutation history. Such results can be used to infer properties of individuals from the composition of the whole population.

The applications of branching processes to biology predominantly concern either systems of rather simple organisms such as cells (*see* **Cell Cycle Models**) or bacteria (*see* **Bacterial Growth, Division, and Mutation**), reproducing at a rapid rate so that the asymptotic assertions of theory become relevant, or they concern conceptual matters, say in evolution. Branching processes can also be applied to approximate genetic and epidemic processes (*see* **Epidemic Models, Stochastic**).

Bibliography

Asmussen, S. & Hering, H. (1983). *Branching Processes*. Birkhäuser, Boston.

Athreya, K. & Jagers, P., eds (1997). *Classical and Modern Branching Processes*. Springer-Verlag, Berlin.

Athreya, K. & Ney, P. (1972). *Branching Processes*. Springer-Verlag, Berlin.

Guttorp, P. (1991). *Statistical Inference for Branching Processes*. Wiley, New York.

Harris, T.E. (1963). *The Theory of Branching Processes*. Springer-Verlag, Berlin.

Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley, Chichester.

Jagers, P. (1991). The growth and stabilization of populations (with discussion), *Statistical Science* **6**, 269–283.

Mode, C.J. (1971). *Multitype Branching Processes*. Elsevier, New York.

Sankaranarayan, G. (1989). *Branching Processes and its Estimation Theory*. Wiley Eastern, New Delhi.

Vatutin, V.A. & Zubkov A.M. (1987). Branching processes I, *Journal of Soviet Mathematics* 39.

Vatutin, V.A. & Zubkov A.M. (1993). Branching processes II, *Journal of Soviet Mathematics* 67.

Kimmel, M. & Axelrod, D. (2002). *Branching Processes in Biology*. Springer, New York.

Haccou, P., Jagers, P. & Vatutin, V. (2004). *Branching Processes: Variation, Growth and Extinction of Populations*. Cambridge University Press, Cambridge.

PETER JAGERS

Breslow–Day Test

The case–control method is a popular way to study the association between exposure and disease in epidemiology. Under the presence of nuisance factors, one could stratify the data into a series of 2×2 tables to control confounding. For example, one might study the relationship between lung cancer and smokers, controlling for age. If the association between exposure and disease is constant from stratum to stratum, then the **Mantel–Haenszel** (MH) summary odds ratio estimator is usually used to estimate the relative risk. Breslow & Day [2] provided a test for assessing the homogeneity of the odds ratios across tables.

Table 1 shows the data in the kth of a series of 2×2 tables, where $k = 1, \ldots, K$. The MH estimator is defined by $\hat{\psi} = \sum_k R_k / \sum_k S_k$, where $R_k = a_k d_k / N_k$ and $S_k = b_k c_k / N_k$. Breslow & Day [2, p. 142] proposed a statistic for testing the null hypothesis of homogeneity of the K true odds ratios. It sums up the squared deviations of observed and fitted values, each standardized by its variance:

$$\sum_{k=1}^{K} \frac{(a_k - A_k(\hat{\psi}))^2}{\mathrm{var}(a_k; \hat{\psi})}, \tag{1}$$

where $A_k(\hat{\psi})$ and $\mathrm{var}(a_k; \hat{\psi})$, denote the expected number and the asymptotic variance of exposed cases based on the MH fitted odds ratio $\hat{\psi}$, respectively.

Tarone [4] noted that by replacing the MH estimator $\hat{\psi}$ by the conditional maximum likelihood estimator, the Breslow–Day test statistic (1) becomes the conditional likelihood score test. Since the MH estimator is inefficient, Tarone [4] and Breslow [1] noted that the test statistic (1) is stochastically larger than a χ^2 random variable (*see* **Chi-square Distribution**) under the homogeneity hypothesis. The correct form for the test was derived by Tarone as

$$\sum_{k=1}^{K} \frac{[a_k - A_k(\hat{\psi})]^2}{\mathrm{var}(a_k; \hat{\psi})} - \frac{[\Sigma_k a_k - \Sigma_k A_k(\hat{\psi})]^2}{\Sigma_k \mathrm{var}(a_k; \hat{\psi})}, \tag{2}$$

where $A_k(\hat{\psi})$ is obtained from the quadratic equation

$$\frac{A_k(\hat{\psi})[N_{0k} - t_k + A_k(\hat{\psi})]}{[N_{1k} - A_k(\hat{\psi})][t_k - A_k(\hat{\psi})]} = \hat{\psi}, \tag{3}$$

and the asymptotic variance is given by

$$\mathrm{var}(a_k; \hat{\psi}) = \left[\frac{1}{A_k(\hat{\psi})} + \frac{1}{N_{1k} - A_k(\hat{\psi})} + \frac{1}{t_k - A_k(\hat{\psi})} + \frac{1}{N_{0k} - t_k + A_k(\hat{\psi})} \right]^{-1}. \tag{4}$$

When the number of strata is small and each table has large frequencies, Tarone's test statistic (2) follows an approximate χ^2 distribution on $K - 1$ degrees of freedom, under the homogeneity hypothesis.

The computer package StatXact [3] provides the Breslow–Day test, as does SAS (PROC FREQ) (*see* **Software, Biostatistical**). Unfortunately, they are both currently based on the incorrect test statistic (1).

Table 1 Notation for the kth of a series of 2×2 tables

	Observed frequencies		
Exposure	Case	Control	Total
Exposed	a_k	b_k	t_k
Unexposed	c_k	d_k	$N_k - t_k$
Total	N_{1k}	N_{0k}	N_k

References

[1] Breslow, N.E. (1996). Statistics in epidemiology: the case-control study, *Journal of the American Statistical Association* **91**, 14–28.

[2] Breslow, N.E. & Day N.E. (1980). *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies*. IARC, Lyon.

[3] Cytel Software Corporation (1995). *StatXact-3 for Windows. User Manual*. Cytel Software Corporation, Cambridge, Mass.

[4] Tarone, R.E. (1985). On heterogeneity tests based on efficient scores, *Biometrika* **72**, 91–95.

I-MING LIU

Breslow–Day Test

I-Ming Liu

Volume 1, pp. 560–560

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Broadband Smoothing

Broadband smoothing is a procedure to reduce the estimation variance of **fast Fourier transform (FFT)** power spectra (*see* **Spectral Analysis**), which results in estimation variance and frequency resolution not constant over the frequency axis, as in conventionally smoothed spectra.

Because of their stochastic nature, spectra of most biological processes are affected by large estimation variance. Smoothing decreases this variance. Typically, a smoothed spectrum is estimated by splitting a **time series** of length T into L shorter data segments of length $T_L < T$, by computing a spectrum over each segment, and by averaging the L spectra thus obtained. If $P_i(f_k)$ is the FFT spectrum in the ith segment, where $f_k = k/T_L$ is the frequency of the kth spectral line, the smoothed spectrum $P^S(f_k)$ is

$$P^S(f_k) = \frac{1}{L}\sum_{i=1}^{L} P_i(f_k). \tag{1}$$

If the number of independent spectra to be averaged increases, then the reduction of estimation variance is greater. On the other hand, if L increases, T_L decreases with a loss of frequency resolution. Thus, a compromise should be found between estimation variance and frequency resolution. When the spectrum is estimated over a broad band of frequencies, however, it is difficult to find a satisfactory compromise because the optimum trade-off in a frequency band might be unacceptable at other frequencies. In this case, broadband smoothing may considerably improve spectral analysis because resolution and estimation variance actually change with the frequency.

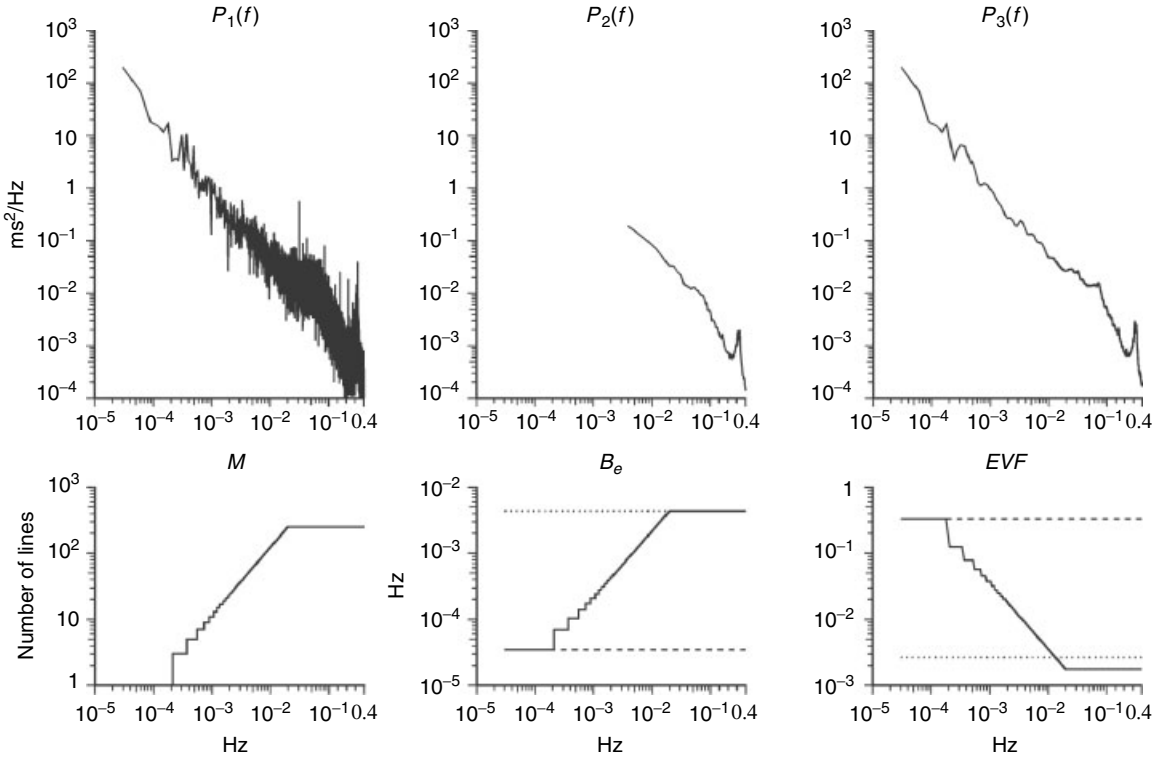

Figure 1 Comparison of traditional and broadband smoothing: data are the RR intervals from a 24 h Holter recording. *Upper panels*: $P_1(f)$ and $P_2(f)$ smoothed as in (1) with $L = 3$, $T_L = 8$ h (*left*) and $L = 378$, $T_L = 228.5$ s (*mid*); $P_3(f)$ smoothed by broadband smoothing of $P_1(f)$ (*right*). *Lower panels*: features of the broadband smoothing used for $P_3(f)$ plotted as functions of f: M is the number of averaged spectral lines; B_e is the equivalent bandwidth; EVF is the estimation variance factor; straight lines also show B_e and EVF for $P_1(f)$ (dashed) and $P_2(f)$ (dotted)

Broadband smoothing consists of a **moving average** of the raw spectrum $P(f_k)$:

$$P^S(f_k) = \sum_{i=-N(f_k)}^{N(f_k)} w_i(f_k) P(f_{k+i}), \qquad (2)$$

where the number of averaged spectral lines $M(f) = 2N(f) + 1$ increases with f to get the highest resolution at the lower frequencies and the more consistent variance reduction at the higher frequencies. A typical choice for $M(f)$ is

$$M(f) = \begin{cases} m_1 & \text{for } f < f^1 \\ int(af^b) & \text{for } f^1 \le f \le f^2 \\ m_2 & \text{for } f > f^2 \end{cases} \qquad (3)$$

with $m_1 \ll m_2$, $b = \log(m_1/m_2)/\log(f^1/f^2)$ and $a = m_1/(f^1)^b$.

The weights w_i are usually selected as:

$$w_i(f) = \frac{N(f) + 1 - |i|}{(N(f)+1)^2} \quad \text{for } -N(f) \le i \le N(f). \qquad (4)$$

An approximation of the variance of the smoothed estimate is [1, 3]

$$\text{Var}[P^S(f)] \cong \text{Var}[P(f)] \sum_{i=-N(f)}^{N(f)} w_i^2. \qquad (5)$$

The Estimation Variance Factor, $EVF = \sum_{i=-N(f_k)}^{N(f_k)} w_i^2$, quantifies the reduction in estimation variance. The frequency resolution can be measured by the equivalent bandwidth, B_e [2]. With the choices (3) and (4), EVF is lower than 1 and decreases with f, while B_e increases with f and is equal to $(N(f) + 1)\Delta f$, with Δf the resolution of the raw spectrum. Figure 1 illustrates the performances of broadband smoothing by comparing three RR-interval spectra from the same 24-hour Holter recording. The first two spectra are estimated by applying traditional smoothing schemes, one preserving high frequency resolution, the other providing consistent variance reduction, while the third spectrum is obtained by broadband smoothing.

References

[1] Di Rienzo, M., Castiglioni, P., Parati, G., Mancia, G. & Pedotti, A. (1996). Effects of sino-aortic denervation on spectral characteristics of blood pressure and pulse interval variability: a wide-band approach, *Medical and Biological Engineering and Computing* **34**, 133–141.

[2] Marple, S.L., Jr. (1987). *Digital Spectral Analysis*. Prentice-Hall, Englewood Cliffs.

[3] Oppenheim, A.V. & Schafer, R.W. (1975). *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs.

PAOLO CASTIGLIONI

Brownian Motion and Diffusion Processes

Brownian motion, or Brownian movement, was named after the English botanist Robert Brown who, in 1827, reported an experiment with pollen of a certain herb in an aqueous suspension. The particles contained in the pollen performed a continuous, haphazard zigzag movement. This movement, Brown pointed out, could not be attributed to life in the particles. During the remainder of the nineteenth century, various experiments showed that the movement depends on the size and mass of the particles. The smaller and lighter the particles, the faster the movement. The movement also depends on the medium. The movement of the particles in gases is faster than the movement in liquids, and the movement in smoke is faster still. The velocity of the movement increases as the temperature increases. (These points formed the basis of the **Ornstein–Uhlenbeck** processes described later in this article.)

Another aspect of the Brownian movement is *diffusion*. In physical science, diffusion is the random molecular movement by which matter is transported. Diffusion occurs in all forms of matter. The diffusion process is slower in liquids than in gases. It takes place also in solids. Generally, there are two parties to every diffusion process. When smoke bursts into the air, diffusion follows; smoke and air are the two parties to the diffusion process. If two bodies of liquids of different colors are placed in two adjacent compartments of a tank, separated by a center divider, then diffusion occurs as soon as the center divider is removed. The two bodies of liquids are the two parties to the diffusion process. If two disks of lead and gold are placed in such a position that their edges meet, the migration of lead into gold and of gold into lead will eventually take place. These are good examples of diffusion processes. A diffusion process will end, but the Brownian movement will continue. In reality, the diffusion processes are a form of the Brownian movement when two or more distinguishable groups of particles are involved. If there were no Brownian movement, then there would be no diffusion processes.

While the various physical characteristics of Brownian motion were discovered in the late nineteenth century, basic questions about Brownian motion remained unanswered until much later. What makes the minute particles engage in a perpetual movement was not determined for the next three-quarters of a century.

It was not until 1905 that Einstein [6] showed that the Brownian motion could be explained by assuming that the particles are subject to continual bombardment of the molecules in the surrounding medium. His pioneering work was generalized, extended, and experimentally verified by various physicists [13]. Brownian motion has since become a popular research topic in theoretical physics. It is now well explained by statistical mechanics and kinetic theory. Wiener [18] also developed the theoretical foundation of Brownian motion and explored its applications. For a history of the theory of Brownian motion, the reader is referred to Einstein [6] and the collected papers edited by Wax [17].

In the theory of stochastic processes, Feller [7] was among the first to study Brownian motion, and he extensively discussed diffusion processes as a major topic in the field. Feller also considered Brownian motion as a limiting case of a random walk as presented in the following section (see also [3]). One may find discussions on Brownian motion and diffusion processes in [1, 2, 4, 5, 9–14]. It should be noted, however, that in stochastic processes we often consider the movement of only *one particle*, in *one dimension*, and the net displacement in *one time period*. This is hardly a fair description of the Brownian movement, much less of diffusion processes. But such a simplification is necessary to make the basic concept clearer and the mathematics simpler. A good understanding of a simple case often makes it easier to understand the complicated real picture.

The purpose of this article is to give a brief review of the one-dimensional Brownian motion and diffusion processes. In the following sections we describe the processes in terms of net displacement from three different perspectives.

1. As a limiting case of a random walk.
2. As a mathematical model.
3. As a stochastic process.

Limiting Case of a Random Walk

Consider a random walk on the real line during the time interval $(0, t]$. At the initial time, $t = 0$, the particle is at the origin. A move of length Δx occurs with every time element Δt.

Let δ be the corresponding **random variable** with $\Pr\{\delta = +\Delta x\} = p$ and $\Pr\{\delta = -\Delta x\} = q$, where $q = 1 - p$. The expectation and the variance of δ are, respectively,

$$\mathrm{E}(\delta) = (p - q)\Delta x \quad \text{and} \quad \mathrm{var}(\delta) = 4pq(\Delta x)^2.$$

There are $t/\Delta t$ moves during the interval $(0, t)$ and $t/\Delta t$ independent and identically distributed random variables δ_i. The sum $\sum_i \delta_i = Z_t$ is the total net displacement on the real line during the interval $(0, t]$, or the position of the particle at time t. According to the **central limit theorem**, when $t/\Delta t$ becomes infinitely large, Z_t has a **normal** distribution with expectation and variance:

$$\mathrm{E}(Z_t) = \frac{(p - q)t\Delta x}{\Delta t}$$

and

$$\mathrm{var}(Z_t) = \frac{4pqt(\Delta x)^2}{\Delta t}.$$

In Brownian motion, we have $\Delta x \to 0$ and $\Delta t \to 0$, and both p and q close to 1/2. For Z_t to have a finite expectation and a bounded variance, we have $\Delta x \to 0$, $\Delta t \to 0$, and $p \to 1/2$ in such a way that

$$\frac{(p - q)\Delta x}{\Delta t} = c \quad \text{and} \quad \frac{(\Delta x)^2}{\Delta t} = D, \qquad (1)$$

where both c and D are positive constants. It follows that

$$\frac{(p - q)}{\Delta x} = \frac{c}{D},\ p = \frac{1}{2} + c\frac{\Delta x}{2D},$$
$$q = \frac{1}{2} - \frac{c\Delta x}{2D},$$

and

$$\mathrm{E}(Z_t) = ct \quad \text{and} \quad \mathrm{var}(Z_t) = Dt + o(\Delta x).$$

Consequently, the limiting probability density function of Z_t is

$$f_{Z_t}(x) = \frac{1}{(2\pi Dt)^{1/2}} \exp\left[-\frac{(x - ct)^2}{2Dt}\right]. \qquad (2)$$

Mathematical Model

For the total net displacement Z_t, we now write the probability

$$\Pr(Z_t = x) = v(x, t),$$

which is a function of both the value x and the length of time t. The purpose is to derive a formula for $v(x, t)$.

Consider $v(x, t + \Delta t)$, the probability of the particle being in position x at time $t + \Delta t$. For the particle to be in position x at $t + \Delta t$, it must be either at $x - \Delta x$ at time t followed by a shift of Δx to the right in $(t, t + \Delta t)$, or at $x + \Delta x$ at time t followed by a shift of Δx to the left in $(t, t + \Delta t)$. As a result,

$$v(x, t + \Delta t) = pv(x - \Delta x, t) + qv(x + \Delta x, t). \qquad (3)$$

Using Taylor's expansion for $v(x, t + \Delta t)$, $v(x - \Delta x, t)$, and $v(x + \Delta x, t)$ at (x, t), defining c and D from (1), and letting $\Delta t \to 0$ and $\Delta x \to 0$, we obtain the following partial differential equation:

$$\frac{\partial}{\partial t}v(x, t) = -c\frac{\partial}{\partial x}v(x, t) + \left(\frac{1}{2}\right)D\frac{\partial^2}{\partial x^2}v(x, t). \qquad (4)$$

This is a partial differential equation of second order; see [1] and [2] for its solution. It is easy to check that the following normal density function satisfies the partial differential equation (4):

$$v(x, t) = \frac{1}{(2\pi Dt)^{1/2}} \exp\left[-\frac{(x - ct)^2}{2Dt}\right]. \qquad (5)$$

Therefore the normal density (5) is the solution of the differential equation (4). We recognize, in passing, that (5) is the same as (2).

The differential equation (4) is the Fokker–Planck diffusion equation, well known in physics. The constant c is the drift coefficient and D is the diffusion coefficient.

Stochastic Process

In **Markov processes** the random variable $X(t)$ assumes discrete values. For $s < t$, the transition (conditional) probability is $P_{ij}(s, t) = \Pr[X(t) = j | X(s) = i]$, where $X(s) = i$ is the condition. In

Brownian motion and diffusion processes, both the time t and the random variable $X(t)$ are continuous. Instead of transition probabilities, we speak of transition density functions. Given $X(s) = x$, the transition density function of $X(t)$ at $X(t) = y$ is denoted by $f(x, s; y, t)$.

A process is time-homogeneous if the transition density, $f(x, s; y, t)$, depends only the difference, $t - s$, and not on s and t separately. A process is additive if $f(x, s; y, t)$ is a function of $y - x$ and not of x and y separately. Generally $(t - s) > 0$, but the difference, $y - x$, is not always positive. The Brownian motion diffusion processes discussed here are assumed to be both time-homogeneous and additive. For nonoverlapping time intervals, the corresponding transitions are assumed independent. For simplicity, we let the process start from the origin, with $X(0) = 0$, and write $f(0, 0; x, t)$ for the transition density function of $X(t)$ at $X(t) = x$.

Following the general practice in Markov processes, we derive the formula for $f(0, 0; x, t)$ in three steps:

1. establish a Chapman–Kolmogorov equation;
2. derive Kolmogorov differential equations; and
3. solve the differential equations for the transition density function $f(0, 0; x, t)$.

Chapman–Kolmogorov Equation

Consider three points on the time axis: $0 < s < t$, and the two adjacent intervals $(0, s)$ and (s, t). In addition to the transition density function $f(0, 0; y, t)$, there are two additional transition density functions: $f(0, 0; x, s)$ and $f(x, s; y, t)$. According to the conditions underlying Brownian motion, the transition during the time interval (s, t) is independent of the transition in the interval $(0, s)$. This essentially is the Markovian property. It follows that the product $f(0, 0; x, s) f(x, s; y, t)$ is the transition density of $X(t)$ at $X(t) = y$ given $X(0) = 0$, by way of $X(s) = x$ at time s. Since the random variable $X(s)$ must assume some value at s,

$$f(0, 0; y, t) = \int_{-\infty}^{\infty} f(0, 0; x, s) f(x, s; y, t) \, \mathrm{d}x, \tag{6}$$

which is the Chapman–Kolmogorov equation for the continuous process $X(t)$.

The Differential Equation

For a detailed account, see [1] and [2].

Consider the time interval $(0, t + \Delta t)$ and two adjacent intervals $(0, t)$ and $(t, t + \Delta t)$. Write two equations,

$$f(0, 0; x, t + \Delta t) = \int_{\infty}^{\infty} f(0, 0; z, t) f(z, t; x, t + \Delta t) \, \mathrm{d}z$$

and

$$f(0, 0; x, t) = f(0, 0; x, t) \int_{\infty}^{\infty} f(z, t; x, t + \Delta t) \, \mathrm{d}z,$$

and hence,

$$f(0, 0; x, t + \Delta t) - f(0, 0; x, t) = \int_{\infty}^{\infty} f(z, t; x, t + \Delta t)[f(0, 0; z, t) - f(0, 0; x, t)] \, \mathrm{d}z.$$

Using Taylor's expansion for the difference, $f(0, 0; z, t) - f(0, 0; x, t)$, we write

$$c = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{\infty}^{\infty} (x - z) f(z, t; x, t + \Delta t) \, \mathrm{d}z$$

and

$$D = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{\infty}^{\infty} (x - z)^2 f(z, t; x, t + \Delta t) \, \mathrm{d}z,$$

where c and D are the infinitesimal mean and variance of $X(t)$, respectively. In the present case, the process is homogeneous and additive: c and D are constants.

Letting $\Delta t \to 0$, we obtain the partial differential equation:

$$\frac{\partial}{\partial t} f(0, 0; x, t) = -c \frac{\partial}{\partial x} f(0, 0; x, t) + \frac{1}{2} D \frac{\partial^2}{\partial x^2} f(0, 0; x, t). \tag{7}$$

Formula (7) is the (forward) Kolmogorov differential equation for the diffusion process.

Solution for the Transition Density

The final step is to solve the differential equation (7) to find the formula for the transition density function

$f(0, 0; x, t)$. Once again we refer to [2] for details of the solution of the differential equation (7).

Eq. (7) is identical to formula (4) in the preceding section when $f(0, 0; x, t)$ is identified with $v(x, t)$. Therefore, the solution of the partial differential equation is

$$f(0, 0; x, t) = \frac{1}{(2\pi Dt)^{1/2}} \exp\left[-\frac{(x - ct)^2}{2Dt}\right]. \quad (8)$$

Identical formulas, in (2), (5) and (8), have thus been found by three different approaches.

Wiener–Levy Process

If the instantaneous distribution of $X(t)$ is "standardized" so that the mean $c = 0$ and the variance $D = 1$, then the differential equation (7) reduces to

$$\frac{\partial f(0, 0; x, t)}{\partial t} = \frac{1}{2} \frac{\partial^2 f(0, 0; x, t)}{\partial x^2},$$

and the solution is

$$f(0, 0; x, t) = \frac{1}{(2\pi t)^{1/2}} \exp\left(-\frac{x^2}{2t}\right),$$

which is the Wiener–Levy process [2].

Ornstein–Uhlenbeck Process

The Ornstein–Uhlenbeck process describes another aspect of Brownian motion, namely the velocity of the movement of a particle. As we noted in the introduction, there are many factors affecting the velocity of the movement, namely the particle's physical characteristics and the surrounding environment. The process is much more complicated than the simple displacement on the real line discussed above. Specifically, in the Ornstein–Uhlenbeck process we wish to determine the velocity of the movement between x and $x + \mathrm{d}x$ after time t, when at $t = 0$ the velocity was $x = x_0$. Let $X(t)$ denote the velocity of the particle at time t and let

$$f(x_0; x, t) = \Pr[X(t) = x | X(0) = x_0].$$

Uhlenbeck & Ornstein [15] established a partial differential equation for $f(x_0; x, t)$ and provided the solution:

$$f(x_0; x, t) = \left\{\frac{m}{2kT[1 - \exp(-2\beta t)]}\right\}^{1/2} \times \exp\left\{-\frac{m}{2kT} \frac{[x - x_0 \exp(-\beta t)]^2}{1 - \exp(-2\beta t)}\right\},$$

where $\beta = \overline{f} m^{-1}$. Here, m is the mass of the particle, $\overline{f}$ is the coefficient of friction, k is the coefficient of viscosity, and T is the absolute temperature.

In addition to the original publication [15], there was a second later publication [16]. A convenient reference for this process is [2].

A Final Remark

The above description covers a small but fundamental part of the Brownian motion and diffusion processes. A more extensive study of the processes requires advanced mathematics. However, even at this level of description, the Brownian motion and diffusion processes have many practical applications. For example, they have been used in studies of **population growth** [2] and in **population genetics** [8].

References

[1] Bailey, N.T.J. (1964). *The Elements of Stochastic Processes*. Wiley, New York.

[2] Bharucha-Reid, A.T. (1960). *Elements of the Theory of Markov Processes and Their Applications*. McGraw-Hill, New York.

[3] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Krieger, New York.

[4] Chung, K.L. (1960). *Markov Processes and Stationary Transition Probabilities*. Springer-Verlag, Heidelberg.

[5] Cox, D.R. & Miller, H.D. (1965). *The Theory of Stochastic Processes*. Wiley, New York.

[6] Einstein, A. (1956). *Investigations on the Theory of the Brownian Motion*. Dover, New York. (Contains translation of Einstein's 1905 papers.).

[7] Feller, W. (1950). Some recent trends in the mathematical theory of diffusion, Proceedings of International Congress of Mathematicians, Vol. 2. Cambridge, Mass., pp. 322–339.

[8] Feller, W. (1951). Diffusion processes in genetics, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 227–246.

[9] Feller, W. (1951). Two singular diffusion problems, *Annals of Mathematics* **54**, 173–182.

[10] Feller, W. (1954). Diffusion processes in one dimension, *Transactions of the American Mathematical Society* **77**, 1–31.

[11] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

[12] Karlin, S. (1966). *A First Course in Stochastic Processes*. Academic Press, New York.

[13] Parzen, E. (1962). *Stochastic Processes*. Holden-Day, San Francisco.

[14] Prabhu, N.U. (1965). *Stochastic Processes – Basic Theory and Its Applications*. Macmillan, New York.

[15] Uhlenbeck, G.E. & Ornstein, L.S. (1930). On the theory of' Brownian motion, *Physics Review* **36**, 823–841 (also in Wax 17).

[16] Wang, M.C. & Uhlenbeck, G.E. (1945). On the theory of Brownian motion II, *Review of Modern Physics* **17**, 323–342.

[17] Wax, N., ed. (1954). *Selected Papers on Noise and Stochastic Processes*. Dover, New York.

[18] Wiener, N. (1930). Generalized harmonic analysis, *Acta Mathematica* **55**, 117.

CHIN LONG CHIANG

Brownlee, John

Born: 1868.
Died: March 20, 1927.

Brownlee was qualified in both mathematics and medicine. He became Director of the Statistical Department of the (British) **Medical Research Council** in 1914, housed in the National Institute for Medical Research in London. He wrote several papers on **vital statistics**, including an important publication in 1916 [1], in which he demonstrated the value of displaying age-specific death rates from tuberculosis by the cohort or generation method (*see* **Age–Period–Cohort Analysis**). He wrote also on mathematical epidemiology, especially in relation to the periodicity of epidemics.

Reference

[1] Brownlee, J. (1916). Certain considerations regarding the epidemiology of phthisis pulmonalis, *Public Health* **29**, 130–145.

PETER ARMITAGE

BSE and vCJD

The transmissible spongiform encephalopathies constitute a group of uniformly fatal neurological degenerative diseases with the abnormal isoform of the cellular prion protein present. They include Creutzfeldt–Jakob Disease (CJD) and kuru, among others, and now also a new variant CJD (vCJD, formerly nvCJD) in humans, scrapie in sheep, and "mad cow disease" or bovine spongiform encephalopathy (BSE) in cattle. A 10-year history introduces salient biostatistical issues in BSE and vCJD.

The first confirmed diagnosis of BSE [54] was made in the UK on November 26, 1986. BSE became a notifiable disease on June 21, 1988 (*see* **Surveillance of Diseases**; **Disease Registers**). The ruminant feed ban [56] came into force on July 18, 1988, and the Bovine Offal (Prohibition) Regulations in November 1989 (in England and Wales) and January 1990 (in Scotland and Northern Ireland).

The first case of BSE after the ruminant feed ban in the offspring of a BSE-infected dam was announced on March 27, 1991. A BSE maternal **cohort study** of over 300 calf pairs with birth dates from August 1987 to November 1989 had been under way on three study farms since July 1989 [59]. All calves of BSE-affected dams in this study were born within 13 months of clinical onset in their dam. Other investigations of risk factors for BSE [12, 37, 55] included **case–control studies**. Despite concentration on BSE cases born after the introduction of the feed ban, there was **confounding** because exposure to contaminated feed had continued well beyond July 1988; reported **confidence intervals** for the **odds ratios** on dam-to-calf BSE transmission overlapped 1 and had upper bounds which excluded high overall rates.

Meanwhile, as a precaution in respect of human health, prospective UK surveillance of CJD was reactivated in May 1990. The remit of the CJD Surveillance Unit was to alert to any changes in the presentation, age-specific incidence, occupational distribution, or dietary correlates of CJD cases that might suggest that humans were affected by exposure to BSE. By November 1995, Gore considered that cases of CJD in five UK farmers and three young adults since May 1990 were more than happenstance: they signalled an epidemiological alert (*see* **Case Series, Case Reports**) [25].

On March 20, 1996, 10 UK cases of a new variant of Creutzfeldt–Jakob disease (vCJD, formerly known as nvCJD) [59] with distinctive neuropathology, methionine homozygosity at codon 129, comparative youth, longer survival from clinical onset, and psychiatric presentation were announced in a statement by the Spongiform Encephalopathy Advisory Committee (SEAC). In a *British Medical Journal* editorial on March 30, 1996, actions necessary to safeguard the public health and properly to acquire data to quantify risks were suggested, and interim analysis of the BSE maternal cohort study called for [26], calls which were initially spurned [2].

On July 29, 1996, SEAC announced that an interim analysis of 273 calf pairs in the BSE maternal cohort study had revealed a significantly enhanced risk of BSE among the offspring of BSE-affected dams compared with matched (*see* **Matching**) **controls** (controls were born in the same herd and calving season to dams which had reached at least six years of age without developing clinical signs of BSE). The risk difference was 10% (95% confidence interval 5 to 14%) when the cohort study ended [19, 57], and **relative risk** 3.2 (95% confidence interval 1.8 to 5.9). In August 1996, UK donors with a family history of CJD were asked to abstain from blood and tissue donation.

By the end of August 1996, a comprehensive account of the transmission dynamics and epidemiology of BSE in British cattle had been published by Anderson et al. [1]; they estimated that approximately 1 million cattle had been infected in the UK, over 160 000 having survived to develop clinical signs of BSE. Underreporting of BSE was substantial before July 1988 (that is, before BSE became notifiable and was compensated for) with an estimated two cases of BSE not reported for every reported case. The epidemiological parameters which Anderson et al. took into account were: age-dependent exposure/susceptibility; five-year mean incubation period; age-specific survivorship of cattle in Great Britain; decline in total herd size from over 13 million in 1974 to under 10 million in 1995; maternal and horizontal transmission [37] with several – but not all – scenarios chosen to reflect that, for BSE in cattle, high infectivity may be restricted to the late or symptomatic stage of the incubation period. **Back-calculation** from BSE cases [1] – without knowledge of (dam–offspring) relatedness and having made allowance for substantial underreporting of BSE

cases before July 1988 – could not distinguish compellingly between no maternal transmission, 10% maternal transmission in the six months before BSE onset in dam, and 10% maternal transmission in the year before dam's onset; **goodness-of-fit chi-square** differed by 10 on 219 degrees of freedom across the foregoing three scenarios but the fit became very substantially worse if the period over which maternal transmission took place was assumed to be three years (change of 240 in goodness of fit).

Meanwhile, medical scientists were giving priority to diagnostic tests for CJD: in September 1996, Hsich et al. [39] reported that, in patients with dementia, a positive immunoassay for the 14-3-3 brain protein in cerebrospinal fluid strongly supported a diagnosis of CJD. October 1996 heralded an even more important breakthrough. Collinge et al. [6] published a molecular analysis of prion strain variation: vCJD has a specific pattern of protease-resistant prion protein (PrP) on Western blot analysis which is distinct from other types of CJD and which resembles those of BSE transmitted to mice, domestic cats, and macaques [43]. These results were consistent with BSE being the source of vCJD, see also [5, 6, 31].

Collinge's molecular analysis has since been used by Deslys et al. [16] to confirm that the first vCJD patient in France had an indistinguishable electrophoretic pattern to UK cases (type 4) while another French patient, a possible case of vCJD – a 52-year-old methionine homozygous female with florid plaques in a brain specimen but dura-mater graft 11 years previously – had a type 2 pattern similar to that described by Collinge in an iatrogenic dura-mater linked CJD case. Molecular analysis of brain tissue from now-six farmer CJD cases in the UK did not find type 4; all from whom tissue was analyzed had classical CJD neuropathologically [37]. Importantly, Hill et al. [33] showed that it was possible to make the specific diagnosis of vCJD by Western blot analysis from frozen tonsillar tissue, see also [34, 35, 41, 47, 49]. These scientific developments, together with Europe's highly assessed geographical BSE risks [50], heralded both Swiss [17] and the European Union's rapid TSE testing in adult cattle and sheep [3, 46, 51]. Postmortem BSE testing, since 2001, of apparently healthy cattle aged over 30 months and risk stock aged over 24 months throughout Europe has transformed our understanding of BSE epidemiology. Active surveillance in cattle showed that clinical BSE cases account for one-third only of all BSE-test positives *and* that BSE positivity is 10 to 15 times higher in risk stock. Assuming differential mortality in late-stage BSE led to Great Britain's having had nearer 4 millions than 1 million BSE-infected cattle [20]. Higher scrapie positivity was also confirmed in fallen sheep [3] *and* apparent positives discovered in ARR/ARR sheep previously considered as scrapie-resistant: their somewhat unusual TSE test profile is being further investigated.

Not only has active BSE surveillance led to a three-fold increase in previous estimates for the extent of Great Britain's BSE epidemic in cattle [20] but, in man too, retrospective [34] and unlinked, anonymous, and ungenotyped testing of stored operative tissue (appendix or tonsil from 1995 to 2000) has discovered an unusual positivity pattern for abnormal prion, PrP^{SC} [36]. Because, to date, all vCJD phenotyped patients have been methionine homozygotes at codon 129, we do not know whether humans' abnormal PrP^{SC} profile in tonsil or appendix tissue is altered by phenotype. There is about a three-fold discordance in prevalence between back-calculation from vCJD cases [22, 23] and testing of tissues for abnormal PrP^{SC}, if *all positives* detected are assumed methionine homozygous and exposed dietarily.

Degree of belief in the proposition "BSE causes vCJD", as elicited from audiences of scientists, legislators and public health specialists in June to August 1996, was highly variable [27]: **modal** score was 8 out of 10, **median** 6, but the **mean** score was 5.4 with **standard deviation** of 2.8. And this despite headlines in the UK press in the week of March 20, 1996, which ranged from the *Daily Mirror*'s scoop on March 20 ("Official: Mad Cow Can Kill You. Govt to admit it today") to the *Evening Standard*'s on March 21 ("French Ban British Beef. Germany calls for Europe boycott as Dorrell warns 11m cattle may have to die") and the *Independent*'s on March 23, 1996 ("Beefgate").

Information on dietary exposure to BSE is critical, but difficult to obtain. CJD patients being too unwell, a relative is asked to recall the patient's diet over a lifetime and since 1985; the CJD Surveillance Unit itself places little confidence in the data so generated. Four indirect sources of information on dietary exposure to BSE may therefore be important: UK nutrition surveys conducted in the 1980s and 1990s (for evidence on strongly age-related dietary consumption, of beefburgers, for example [30]); unannounced

inspections at abattoirs by the State Veterinary Service (for evidence on the frequency and nature of breaches of regulations concerning specified bovine materials [28, 45]); an audit commissioned in the spring of 1996 by the Ministry of Agriculture, Fisheries and Food (for information by quinquennia on which bovine and ovine tissues went into which foods when); and back-calculation from BSE cases to infer the BSE infection curve [1] and annual numbers of BSE-infected bovines which were within one year of clinical onset (say) when slaughtered for human consumption. Cooper and Bird's synthesis [7–9] of the foregoing data sources predated upward revision by Ferguson et al. [20] of annual numbers of BSE-infected bovines, which were within one year of clinical onset when slaughtered for human consumption. Changes in the estimated preclinical BSE incidence pattern (not just level) could have implications for Cooper and Bird's predicted incidence of vCJD from the United Kingdom dietary exposure to BSE for the 1940–1969 and post-1969 birth cohorts [10]. Their projections suggested: that about three-fifth of predicted dietary vCJD onsets would be in males, see also [11]; and, over and above dietary exposure, that an age-dependent susceptibility function – as first conjectured by Valleron et al. [53] – or other exposure was required to match the age distribution of vCJD patients in the 1940–1969 birth cohort. Although the risk of infection with prion diseases increases with repeated challenges, Gravenor et al. [24] have found that it does so to a lesser extent than is expected if challenges combined independently or in a cumulative manner. This finding could also be a part explanation for implied age-dependent susceptibility.

Cousens et al. [14] hazarded [42] projections of vCJD by assuming simply that, until 1989, the number of people newly infected with the BSE agent was proportional to the number of BSE cases with onset in that year, as may be reasonable if bovine material is infectious to humans for only a short period before cattle develop clinical BSE. These very preliminary projections do not explain the apparently age- and genetics-related susceptibility to vCJD. Dates of onset of vCJD (seven in 1994, six in 1995), referral to the CJD Surveillance Unit, death and confirmation of diagnosis were plotted by Cousens et al. for the first 14 cases of vCJD in the UK: given the typically long delay between onset and confirmation, the final number with onset in 1994 or 1995 could be about 23 [14]. Regular updating – at least annually – of these crucial date data is epidemiologically essential, and by a variety of methods as reflected, or referenced, in a special TSE issue (Volume 12, Number 3) of Statistical Methods in Medical Research in 2003.

Unlike in **AIDS** [15, 26, 48], an independent group of statisticians and subject-matter specialists was not convened prior to 1996 to work on, and publish, BSE projections, but this investigative format did apply for UK's review in 2002–2003 of its Over Thirty Month Rule, whereby from April 1996, no part of UK's cattle slaughtered at 30+ months of age could enter human food or any feed chain [20]. A reinforced feed ban came into force from August 1, 1996 in the United Kingdom (and from January 1, 2001 elsewhere in European Union), after which it was hoped that maternal transmission would be the only route of cattle's BSE exposure. Events proved otherwise with maternal transmission being effectively ruled out for about a third of BSE cases born after the reinforced feed ban (BARBs) [52] so that in 2004, the United Kingdom was designing a BARB-controls study to investigate a low-level, third wave of BSE exposure (prior to original ruminant feed was first, between initial and reinforced feed bans was the second).

Three statistical teams had been invited to undertake further analyses [29] of the BSE maternal cohort study to determine whether the elevated risk of BSE among offspring of BSE-affected dams was due to a genetically enhanced risk of food-borne infection, to vertical transmission of the etiologic agent from dam to calf, or some combination of the two. Evidence of an enhanced BSE risk in calves born closer to disease onset in their dams would provide support for vertical transmission; and corroboration of the "infectious interval" could be sought from the BSE database in respect of calves born well after the feed ban to dams which ultimately developed BSE [18, 29]. Analysis of the BSE database was complicated by misidentification of dam–calf pairs and by the lack of survival data, which pointed up the need for improved cattle-tracing systems, such as Northern Ireland had and have subsequently come into force in Great Britain. To date, separate analyses of the BSE maternal cohort study and BSE database have been undertaken. A **Bayesian** formulation would allow for full propagation of uncertainty [21] with consideration also of lateral transmission and of regional exposure to contaminated feed in 1989 or later: the

BSE maternal cohort study is seen as external data and **sensitivity analysis** of assumptions about calf survivorship according to BSE status of dam would be crucial.

Statistical plans [28] for unannounced inspections at abattoirs by the State Veterinary Service were an outstanding issue in the aftermath of vCJD. Contrary to reporting standards in 1995 [45], they would allow defendable estimates of the extent of compliance with BSE and other regulations intended to safeguard public or animal health. Statistical planning in determining which brain and other tissue to preserve or test from cows in the over-30-month slaughter program and selective cull was only fully addressed under the auspices of the European Union's Scientific Steering Committee [3, 51]. Such tissue, particularly from UK cows born either between feed bans or after July 31, 1996, could be extremely valuable in studying age-specific pathogenesis and in determining the performance of diagnostic tests in asymptomatic animals.

Three main statistical issues in vCJD are acquisition of key data to underpin projections; evolution of a risk score for highly suspect vCJD; and follow-up of children born to parents with vCJD, of healthcare workers who attended the delivery of a mother with vCJD, and of healthcare or other workers who have had percutaneous exposure to vCJD or to BSE. Probable blood-borne vCJD transmission [4, 44], announced to both Houses of Parliament in December 2003, has underlined rights and responsibilities for limiting human-to-human transmission of vCJD and for surveillance. Blood-borne vCJD transmission had been anticipated by BSE (and scrapie) transmission risks of 10 to 20% in sheep [38, 40] via blood transfusion: of 24 sheep-recipients of BSE-infected transfusions, two were confirmed BSE cases (8%) and two others suspect; and of 21 sheep-recipients of scrapie-infected transfusions, at least four had succumbed (21%) [40].

From 5 April 2004, the United Kingdom excluded recipients of blood or tissue from donation. Preventing operative transmission of vCJD *and* quantifying human-to-human vCJD transmission risks need additional measures. Quantification of human-to-human vCJD transmission risks requires *identification of index patients* (incubating vCJD, positive for abnormal prion protein PrP^{SC} or otherwise at risk) *and their recipients* (of vCJD-implicated blood or tissue or surgical instruments), a *recipients' database*, which records exposure risk classification and donor, together with *flagging of recipients for mortality* and *recipients' agreement in life* that they be tested postmortem for abnormal PrP^{SC} and phenotyped at codon 129 since all vCJD cases to date have been methionine homozygous at codon 129, but only 40% of UK population is.

Surveillance options include attributable tonsil biopsy for abnormal PrP^{SC} at all autopsies [4] under 50 years of age (with positives phenotyped at codon 129), which would facilitate identification, follow-up, and risk quantification for recipients of "at-PrP^{SC}-risk" blood or tissue or implicated in a surgical web spun out from deceased positives.

Key data to underpin projections have been marshaled, such as: those from patients with kuru or iatrogenic CJD or from repeated challenge experiments in rodents to estimate the incubation period, its age dependence (if any), and possible dose-responsiveness [24]; age-specific and temporal changes in the consumption of foods, which may have contained the BSE agent [7–11, 20] to estimate dietary BSE exposure; those, including costs, from feasibility study on estimating the prevalence of abnormal prion in 10–50 year olds by molecular analysis of stored tonsillar or appendix tissue as forerunner of a national tonsil archive to facilitate testing or postmortem testing in the course of autopsies routinely performed for other reasons ranging from road traffic accidents to drugs overdose.

A second issue was to evolve, and update, a risk score for highly suspect vCJD, including for determining a suitable, if not optimal, sequence of investigations that is likely to reach a correct, definitive diagnosis quickly and with minimal trauma for the patient. Tonsil biopsy, now accepted in differential diagnosis of vCJD, has resolved this issue to such an extent that the postmortem rate has fallen considerably in vCJD-diagnosed patients. Careful documentation of the sequence of neurological signs was considered important [58]. The setting-up of a patient-preference protocol under Medical Research Council auspices at least affords all patients the option of standardized follow-up whether they elect for randomization or not. (see under "other studies" on `http://www.ctu.mrc.ac.uk/browse.asp` for initial options such as: "choice = to receive quinacrine now", "choice = to reject quinacrine now" and "choice = randomization to receive/not receive quinacrine now").

Finally, to quantify the risk (if any) of mother-to-child or of percutaneous transmission of vCJD [26] we need outstandingly to register exposed individuals so that the relevant denominators are knowable.

References

[1] Anderson, R.A., Donnelly, C.A., Ferguson, N.M., Woolhouse, M.E.J., Watt, C.J., Udy, H.J., Mawhinney, S., Dunstan, S.P., Southwood, T.R.E., Wilesmith, J.W., Ryan, J.B.M., Hoinville, L.J., Hillerton, J.E., Austin, A.R. & Wells, G.A.H. (1996). Transmission dynamics and epidemiology of BSE in British cattle, *Nature* **382**, 779–788.

[2] Arthur, C. (1996). Ministry vets spurn calls to pass on data from calf experiments, *Independent* April 12, p. 2.

[3] Bird, S.M. (2003). European Union's rapid TSE testing in adult cattle and sheep: implementation and results in 2001 and 2002, *Statistical Methods in Medical Research* **12**, 261–278.

[4] Bird, S.M. (2004). Recipients of blood or blood products "at vCJD risk". We need to define their rights and responsibilities and those of others, *British Medical Journal* **328**, 118–119.

[5] Bruce, M.E., Will, R.G., Ironside, J.W., et al. (1997). Transmission to mice indicates that 'new variant' CJD is caused by the BSE agent, *Nature* **389**, 498–501.

[6] Collinge, J., Sidle, K.C.L., Meads, J., Ironside, J. & Hill, A.F. (1996). Molecular analysis of prion strain variation and the aetiology of "new variant" CJD, *Nature* **383**, 685–690.

[7] Cooper, J.D. & Bird, S.M. (2002). UK bovine carcass meat consumed as burgers, sausages and other meat products: by birth cohort and gender, *Journal of Cancer Epidemiology and Prevention* **7**, 49–57.

[8] Cooper, J.D. & Bird, S.M. (2002). UK dietary exposure to BSE in beef mechanically recovered meat: by birth cohort and gender, *Journal of Cancer Epidemiology and Prevention* **7**, 59–70.

[9] Cooper, J.D. & Bird, S.M. (2002). UK dietary exposure to BSE in head meat: by birth cohort and gender, *Journal of Cancer Epidemiology and Prevention* **7**, 71–83.

[10] Cooper, J.D. & Bird, S.M. (2003). Predicting incidence of variant Creutzfeldt–Jakob disease from UK dietary exposure to bovine spongiform encephalopathy for the 1940 to 1969 and post-1969 birth cohorts, *International Journal of Epidemiology* **32**, 784–791.

[11] Chadeau-Hyam, M., Tard, A., Bird, S., le Guennec, S., Bemrah, N., Volatier, J-L. & Alperovitch, A. (2003). Estimation of the exposure of the French population to the BSE agent: comparison of the 1980-95 consumption of beef products containing mechanically recovered meat in France and the UK, by birth cohort and gender, *Statistical Methods in Medical Research* **12**, 247–260.

[12] Curnow, R.N. & Hau, C.M. (1996). The incidence of bovine spongiform encephalopathy in the progeny of affected sires and dams, *Veterinary Record* **138**, 407–408.

[13] Cousens, S., Everington, D., Ward, H.J.T., Huillard, J., Will, R.G. & Smith, P.G. (2003). The geographical distribution of variant Creutzfeldt–Jakob disease cases in the UK: what can we learn from it?, *Statistical Methods in Medical Research* **12**, 235–246.

[14] Cousens, S.N., Vynnycky, E., Zeidler, M., Will, R.G. & Smith, P.G. (1997). Predicting the CJD epidemic in humans, *Nature* **385**, 197–198.

[15] Day, N.E., Gore, S.M. & de Angelis, D. (1995). Acquired immune deficiency syndrome predictions for England and Wales (1992–1997): sensitivity analysis, information, decision, *Journal of the Royal Statistical Society, Series A* **158**, 505–524.

[16] Deslys, J.-P., Lasmezas, C.I., Streichenberger, N., Hill, A., Collinge, J., Dormont, D. & Kopp, N. (1997). New variant Creutzfeldt–Jakob disease in France, *Lancet* **349**, 30–31.

[17] Doherr, M.G., Hett, A.R., Cohen, C.H., Fatzer, R., Ufenacht, J.F., Zurbriggen, A. & Heim, D. (2002). Trends in prevalence of BSE in Switzerland based on fallen stock and slaughter surveillance, *Veterinary Record* **150**, 347–348.

[18] Donnelly, C.A., Ferguson, N.M., Ghani, A.C., Wilesmith, J.W. & Anderson, R.M. (1997). Analysis of the dam–calf pairs of BSE cases: confirmation of maternal risk enhancement. *Proceedings of the Royal Society of London, Series B* **264**, 1647–1656.

[19] Donnelly, C.A., Gore, S.M., Curnow, R.N. & Wilesmith, J.W. (1997). Preamble. The BSE Maternal Cohort Study – its purpose and findings, *Applied Statistics* **46**, 299–304.

[20] Ferguson, N.M. & Donnelly, C.A. (2003). Assessment of the risk posed by bovine spongiform encephalopathy in cattle in Great Britain and the impact of potential changes to current control measures, *Proceedings of the Royal Society of London, Series B* **270**, 1579–1584.

[21] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[22] Ghani, A.C., Donnelly, C.A., Ferguson, N.M. & Anderson, R.M. (2003). Updated projections of future vCJD deaths in the UK. *Biomed Central Infectious Diseases* **3**, 4 (see also `http://www.biomedcentral.com/1471-2334/3/4`).

[23] Ghani, A.C., Ferguson, N.M., Donnelly, C.A., Hagenaars, T.J. & Anderson, R.M. (1998). Estimation of the number of people incubating variant CJD, *Lancet* **352**, 1353–1354.

[24] Gravenor, M.B., Stallard, N., Curnow, R. & McLean, A.R. (2003). Repeated challenge with prion disease: The risk of infection and impact on incubation period, *PNAS* **100**,(19), 10960–10965 (see also `www.pnas.org/cgi/doi/10.1073/pnas.1833677100`).

[25] Gore, S.M. (1995). More than happenstance: CJD in farmers and young adults, *British Medical Journal* **311**, 1416–1418.

[26] Gore, S.M. (1996). Bovine Creutzfeldt–Jakob disease? Failures of epidemiology must be remedied, *British Medical Journal* **312**, 791–793.

[27] Gore, S.M. (1996). Address to Parliamentary & Scientific Committee. More than happenstance: CJD in farmers and young adults, *Science in Parliament* **53**, 2–3.

[28] Gore, S.M. (1996). Bovine spongiform encephalopathy and "The term safe is not the same as zero risk", *Journal of the Royal Statistical Society, Series A* **159**, 363–365.

[29] Gore, S.M., Gilks W.R. & Wilesmith, J.W. (1997). Bovine spongiform encephalopathy: maternal cohort study – exploratory analysis, *Applied Statistics* **46**, 305–320.

[30] Gregory, J., Foster, K., Tyler, H. & Wiseman, M. (1990). *The Dietary and Nutritional Survey of British Adults*. Office of Population, Censuses and Surveys and Ministry of Agriculture, Fisheries and Food. HMSO, London.

[31] Hill, A.F., Desbruslais, M., Joiner, S., Sidle, K.C.L., Gowland, I. & Collinge, J. (1997). The same prion strain causes vCJD and BSE, *Nature* **389**, 448–450.

[32] Hill, A.F., Will, R.G., Ironside, J. & Collinge, J. (1997). Type of prion protein in UK farmers with Creutzfeldt–Jakob disease, *Lancet* **350**, 188.

[33] Hill, A.F., Zeidler, M., Ironside, J. & Collinge, J. (1997). Diagnosis of new variant Creutzfeldt–Jakob disease by tonsil biopsy, *Lancet* **349**, 99–100.

[34] Hilton, D.A., Fathers, E., Edwards, P., Ironside, J.W. & Zajicek, J. (1998). Prion immunoreactivity in appendix before clinical onset of variant Creutzfeldt–Jakob disease, *Lancet* **352**, 703–704.

[35] Hilton, D.A., Ghani, A.C., Conyers, L., Edwards, P., McCardle, L., Penney, M., Ritchie, D. & Ironside, J. (2002). Accumulation of prion protein in tonsil and appendix: review of tissue samples, *British Medical Journal* **325**, 633–634.

[36] Hilton, D.A., Ghani, A.C., Conyers, L., Edwards, P., McCardle, L., Ritchie, D., Penney, M., Hegazy, D. & Ironside, J.W. (2004). Prevalence of lymphoreticular prion protein accumulation in UK tissue samples, *Journal of Pathology* **202**, x–y. DOI: 10.1002/path.1580.

[37] Hoinville, L.J., Wilesmith, J.W. & Richards, M.S. (1995). An investigation of risk factors for cases of bovine spongiform encephalopathy born after the introduction of the "feed ban", *Veterinary Record* **136**, 312–318.

[38] Houston, F., Foster, J.D., Chong, A., Hunter, N. & Bostock, C.J. (2000). Transmission of BSE by blood transfusion in sheep, *Lancet* **356**, 999–1000.

[39] Hsich, G., Kenney, K., Gibbs, C.J., Lee, K.H. & Harrington, M.G. (1996). The 14-3-3 brain protein in cerebrospinal fluid as a marker for transmissible spongiform encephalopathies, *New England Journal of Medicine* **335**, 924–930.

[40] Hunter, N., Foster, J., Chong, A., et al. (2002). Transmission of prion diseases by blood transfusion. *Journal of Gen. Virology* **83**, 2897–2905.

[41] Ironside, J.W., Hilton, D.A., Ghani, A., et al. (2000). Retrospective study of prion-protein accumulation in tonsil and appendix tissues. *Lancet* **355**, 1693–1694.

[42] Lancet (1996). Betraying the public over vCJD risk, *Lancet* **348**, 1529.

[43] Lasmezas, C.I., Deslys, J.-P., Demalmay, R., Adjou, K.T., Lamoury, F., Dormont, D., Robain, O., Ironside, J. & Hauw, J.-J. (1996). BSE transmission to macaques, *Nature* **381**, 743–744.

[44] Llewelyn, C.A., Hewitt, P.E., Knight, R.S.G., Amar, K., Cousens, S., Mackenzie, J. & Will, R.G. (2004). Possible transmission of variant Creutzfeldt–Jakob disease by blood transfusion, *Lancet* **363**, 417–21.

[45] Meat Hygiene Service (1996). *Annual Report and Accounts 1995/96*. HMSO, London, pp. 30–31.

[46] Moynagh, J. & Schimmel, H. (1999). Tests for BSE evaluated, *Nature* **100**, 105.

[47] O'Rourke K.I., Baszler T.V., Besser T.E., et al. (2000). Preclinical diagnosis of scrapie by immunohistochemistry of third eyelid lymphoid tissue. *Journal of Clinical Microbiology* **38**, 3254–3259.

[48] Report of an Expert Group (chairman N.E. Day) (1996). The incidence and prevalence of AIDS and prevalence of other severe HIV disease in England and Wales for 1995 to 1999: projections using data to the end of 1994. *Communicable Disease Report 6* (Review 1), R1–R24.

[49] Schreuder, B.E., van Keulen, L.J., Vromans, M.E., Langeveld, J.P. & Smits, M.A. (1998). Tonsillar biopsy and PrPSc detection in the preclinical diagnosis of scrapie, *Veterinary Record* **142**, 564–568.

[50] Scientific Steering Committee. *Opinion on Geographical Risk of Bovine Spongiform Encephalopathy (GBR)*. Brussels: European Commission, Health and Consumer Protection Directorate-General 2000.

[51] Scientific Steering Committee. *Opinion on Requirements for Statistically Authoritative BSE/TSE Surveys.* Brussels: European Commission, Health and Consumer Protection Directorate-General, November 2001.

[52] Scientific Steering Committee. *Opinion and Report on BSE in Great Britain's cattle born after 31 July 1996 [BARBs]*. Brussels: European Commission, Health and Consumer Protection Directorate-General, March 2003.

[53] Valleron, A.J., Boelle, P.Y., Will, R. & Cresbon, J.Y. (2001). Estimation of epidemic size and incubation time based on age characteristics of vCJD in the United Kingdom, *Science* **294**, 1726–1728.

[54] Wells, G.A.H., Scott, A.C., Johnson, C.T., Gunning, R.F., Hancock, R.D., Jeffrey, M., Dawson, M. & Bradley, R. (1987). A novel progressive spongiform encephalopathy in cattle, *Veterinary Record* **121**, 419–420.

[55] Wilesmith, J.W. (1994). Bovine spongiform encephalopathy: epidemiological factors associated with the emergence of an important new animal pathogen in Great Britain, *Seminars in Virology* **5**, 179–187.

[56] Wilesmith, J.W., Wells, G.A.H., Cranwell, M.P. & Ryan, J.B.M. (1988). Bovine spongiform encephalopathy: epidemiological studies, *Veterinary Record* **123**, 638–644.

[57] Wilesmith, J.W., Wells, G.A.H., Ryan, J.B.M., Gavier-Widen, D. & Simmons, M.M. (1997). A cohort study to examine maternally associated risk factors for bovine spongiform encephalopathy, *Veterinary Record* **141**, 239–243.

[58] Will, R. & Zeidler, M. (1996). Diagnosing Creutzfeldt–Jakob disease. Case identification depends on neurological and neuropathological assessment, *British Medical Journal* **313**, 833–834.

[59] Will, R.G., Ironside, J.W., Zeidler, M., Cousens, S.N., Estibeiro, K., Alperovitch, A., Poser, S., Pocchiari, M., Hofman, A. & Smith, P.G. (1996). A new variant of Creutzfeldt–Jakob disease in the UK, *Lancet* **347**, 921–925.

SHEILA M. BIRD

Burden of Disease

"Burden of disease" is the name given to a concept dealing with a range of medical statistics. It aims to give a comprehensive picture of how different diseases impact on society. Some analysts use this concept to allocate health care and research resources. Because it covers a range of impacts of disease on society, the burden is often referred to in the plural, a practice followed in this article.

"Burdens of disease" aims to give an account of the dimensions of damage inflicted by ill health on society. The main burdens covered are mortality, morbidity, and resources costs of care and treatment. Applications were developed first in the US, but examples here are based on experience in the UK. The main interest is in how the burdens of different diseases compare. It aims to answer questions like: Is heart disease a bigger killer than cancer? Which diseases afflict women especially? Which diseases impose greatest cost on healthcare services? Table 1 is taken from the latest UK table of burdens of disease [4] and summarizes the main results. Answering such questions raises a host of problems. Some of them arise from one of the burdens and some arise from attempts to bring them all together.

Mortality

What do we mean when we say that one disease is a bigger killer than another? Looking simply at unadjusted death rates will not capture the flavor of the question. Since all must die, the simple causing of death does not impose a burden on society. It is premature death that is the burden. One commonly adopted solution is to look at deaths below a certain age. Choosing the age is not simple. If it is set very low, then it will focus attention on a very narrow range of causes of death such as sudden infant death syndrome (SIDS), accidents, etc. Set too high and it loses any focus on policy issues. Over the age of, say, one hundred, the precise cause of death is of less interest than the survival thus far. Recent presentations have looked not at simple death rates, but weighted them by the years below a certain age. Thus SIDS deaths are given a high weight because they are seen as destroying almost a whole lifespan. While the usual presentation focuses on life-years lost, there is no reason in principle why both crude and weighted death rates should not be used. They reflect different concerns. Even though inevitable, the event of death is always painful to survivors and fearful to the dying. The curtailing of life is an additional loss. A further difficulty is whether the life-years lost calculation should use different age standards for men and women, to reflect their different **life expectancy**.

Compared with other burdens, the data difficulties of mortality statistics are relatively few. In many developed countries the measurement has become simpler with the recording of several **causes of death** on death certificates. There remain problems with causes like **AIDS** and suicides, where there will be a *bias* against recording such conditions on death certificates.

Morbidity

Data on morbidity tend to come from three main sources: administrative records associated with welfare benefits, surveys of physicians, and household surveys (*see* **Surveys, Health and Morbidity**). In the UK these are represented by sickness benefit records, the morbidity survey of general practitioners, and surveys like the Disability Survey and the Health Survey. The first dataset records total days of certificated sickness absence. The second source records the patient consulting rate for a given condition. Household surveys will, in principle, obtain a direct measure of the number of people suffering from a given condition at a particular point in time, and so come closest to a measure of the extent to which the population at a given moment is suffering from ill health. While certificated sickness measures the same thing in principle, it suffers, besides the problems associated with administrative sources like policy change, from the problem that only those otherwise in work will be recorded. It therefore omits most of the population who are ill, the old, and many of the chronically sick. The number of people consulting primary physicians gives a useful indication of something between incidence and prevalence of disease, but requires additional weighting to give a useful picture of the burden of morbidity. Chronic conditions clearly impose a greater burden than short-lived conditions, so this measure of prevalence has to be adjusted by some estimates of average duration.

Table 1 Twenty leading causes of mortality, patients consulting in general practice, and expenditure by ICD-9 subchapter. England except MSGP = England and Wales, GHS = Great Britain [4]

Rank	Mortality (1991)	(%)	Patients consulting in general practice[a] by cause[b] (1991/92)	Rate per 1000	NHS hospital expenditure (1992/93)	(%)	NHS primary care expenditure[c] (1992/93)	(%)	Community health and social care for adults net expenditure (1992/93)	(%)
1	Ischemic heart disease	26.2	Acute respiratory infections	242.0	Injury and poisoning	5.8	Mouth disease[e]	26.0	Learning disability	12.5
2	Stroke	12.1	Skin diseases	159.6	Learning disability	5.2	Acute respiratory infection	7.7	Stroke	7.1
3	Lung cancer	6.0	Infectious diseases other than TB	156.2	Symptoms	4.3	Eye disorders[d]	6.3	Other arthropathies	6.5
4	Pneumonia	5.0	Injury and poisoning	139.0	Stroke	4.2	Symptoms	5.7	Eye disorders	5.2
5	Chronic obstructive pulmonary diseases (COPD)	4.8	Symptoms	138.7	Schizophrenia	4.0	Injury and poisoning	3.8	Dementla	5.2
6	Colorectal cancer	3.0	Preventive medicine	138.0	Normal delivery	3.1	Infections other than TB	3.0	Neuroses	3.8
7	Arteries	2.8	Screening	121.0	Dementia	2.7	Skin diseases	3.0	Ear disorders	3.7
8	Cancer of other sites	2.7	Ear disorders	101.2	Ischemic heart disease	2.4	Heart failure	2.7	Rheumatism	3.7
9	Female breast cancer	2.4	Family planning	72.8	Other pregnancy	2.3	Preventive medicine	2.2	Prevention	3.5
10	Other heart	2.0	Other neuroses, etc.	64.9	Procedures and aftercare	2.3	Osteoarthritis	2.1	Alcohol, drugs, etc.	3.1
11	Dementia	1.7	Eye disorders	63.7	Skin diseases	2.0	Screening	2.0	Multiple sclerosis	2.7
12	Prostatic cancer	1.5	Social and social marital	61.3	Arteries and veins	2.0	Ear disorders	2.0	Other CNS and disorders of the PNS	2.7
13	Stomach cancer	1.5	Other back diseases	59.1	Other nonorganic psychoses	1.9	COPD	2.0	Osteoarthritis	2.3
14	Diabetes	1.4	Dorsopathies	59.1	Other neuroses, etc	1.9	Hypertension	1.8	Rheumatoid arthritis	2.1
15	Urinary cancer	1.3	Female genital tract	57.4	Female genital tract	1.9	Ischemic heart disease	1.8	Pregnancy	1.8
16	Pancreatic cancer	1.1	Sprains and strains	55.0	Social and social marital, etc.	1.8	Other urinary	1.6	Old age	1.7
17	Suicide and unknown motive	1.0	Rheumatism	49.0	Heart failure	1.6	Stroke	1.5	Other heart disease	1.6
18	Genito-urinary cancers female	1.0	Upper respiratory tract diseases	43.0	Pneumonia	1.6	Intestines and peritoneum	1.4	Schizophrenia	1.6
19	Other systemic cancers	1.0	Asthma	42.5	Eye disorders	1.5	Asthma	1.4	Epilepsy	1.6
20	Heart failure	1.0	Hypertension	41.9	Other respiratory	1.4	Other back disorders	1.2	Symptoms	1.6
Total leading causes		79.5				53.7		79.1		73.8
All other causes		20.5				35.4[f]		20.9		26.2[g]

Sources: Mortality: OPCS 1991.
Patient consultations: Morbidity Statistics from General Practice (MSGP4) 1991–92.
Expenditure: MSGP4, HES and Programme Budget.
[a]Patient consulting rates are based on consultations with general practitioners and practice nurses.
[b]78% of patients consult their GP at least once per annum. If a patient consults for more than one cause, they will be recorded once for each cause. Therefore, rates should not be added as there will be an element of double counting.
[c]NHS Primary care expenditure excluding pharmaceutical expenditure.
[d]Primary care expenditure relating to eye disorders include all costs attributable to General Optical Services.
[e]Primary care expenditure relating to mouth disease include all costs attributable to General Dental Services.
[f]Includes expenditure on day patients in hospitals, Accident and Emergency, Teaching Hospital Uplift (SIFTR) and "other" hospital expenditure.
[g]Includes unallocated expenditure.

Another weighting that is required is an adjustment for severity. A frequently used measure is the QALY or quality-adjusted life-year (*see* **Quality of Life and Health Status**). The implied days of sickness are weighted by the degree of suffering caused. A variety of scales have been developed such as the Euroqol [5], and estimates of discomfort and disability impact have been attached to different conditions.

While statistics on certificated sickness days may be of limited value in obtaining an overall picture of such suffering, they provide valuable subsidiary information on the burden imposed on the economy from not having workers available, on those who depend on the income they might otherwise have earned, and on the agencies that pay benefit.

Resource Costs

Societies devote considerable resources both to curing disease and to remedying some of its consequences. In most industrial countries much of this burden falls on the state, while in some there is considerable private finance of health insurance. However financed, these represent a use of resources which could otherwise perform some other useful function. Simplest to measure in most **health services** are the number of beds devoted to different conditions, and the average cost of bed occupancy. Slightly more difficult is out-patient (ambulatory) care, where records of conditions seem to be less systematically maintained. Where morbidity data are based on surveys of primary carers, these sources can also be used to estimate the cost burden of primary care. Here, of course, it is the consultation rate rather than the number of patients consulting which is relevant. Separate estimates are often required for dental care.

The disease classification of pharmaceuticals requires a certain amount of judgment to align with conditions. Classification becomes progressively more difficult with care which requires support against disability rather than curative interventions where the recording of causes of such conditions may be limited.

As well as pecuniary burdens borne both by the public and private sectors, there are nonpecuniary burdens – particularly in terms of caring. While such burdens are less often recorded in official compilations of burdens of disease, there are sources of such information in general and in dedicated household surveys.

There is conceptual difficulty with measuring the resource burden. It reflects what is considered appropriate by current medical practice. There is an implied but incorrect assumption that the scale of resources devoted to different conditions is such as to bring every sufferer back to a similar kind of condition. This is clearly far from the case. One of the conditions that imposes the heaviest burden on in-patient care is **stroke**, although it is unclear that the scale of those resources reflect the effectiveness of the interventions, compared, with, say, interventions in heart disease. The cost burdens therefore reflect a combination of prevalence and medical practice. If it were decided that more care should be devoted to mental illness, say, then there would be a perceived rise in burden, which might be accompanied by an unrecorded abatement in the severity of morbidity.

Uses and Abuses

In a 1996 publication the UK Department of Health presents burdens in a range of different categories [4]. There is no attempt to combine these into one overall indicator of burden. This was not the practice adopted in the first attempt by Black & Pole [1]. They combined all burdens together using implicit weights reflecting both the severity of disease relative to full health and death, and an implied pecuniary value of death. While it is always open to users to make such combinations using their own assumptions, the attribution of a money value to life is seen as too controversial to form the basis of a statistical publication.

Black & Pole, and many of those who have followed them, have seen burdens of disease as of potential value in decisions on where to direct preventive or curative resources and research which might develop them. While the epidemiologic mapping contained in burdens of disease provides an essential component of any system for making such allocations, it must be combined with indicators of the effectiveness of interventions. There is no point in throwing a large amount of resources at a disease, however burdensome, if it makes no difference to that burden.

Indicators of cost effectiveness are now available for a range – albeit still a fairly limited range of conditions [4]. The technique for effective resource allocation depends on a combination of burdens of disease statistics and cost-effectiveness information.

In practice nearly all interventions vary in their effectiveness depending both on the condition of the sufferer, the general circumstances, and the quality of the medical practitioner. An effective allocation of interventions will depend on the scale of the condition and the extent to which further intervention is likely to be effective in reducing the burden. It turns out that making such an allocation is relatively complicated. One scheme for doing so is described in Neuburger & Fraser [2].

Another potential use of burdens is to indicate the burden of a particular risk factor. Thus it would be possible to show the burden of smoking or car exhaust fumes by weighting together appropriate fractions of the burden of those disease associated with such risk factors. In terms of policy these could then be compared with interventions which might be taken to abate them.

Conclusion

Burdens of disease provides a valuable framework for compiling a range of data on different diseases. Used with care it can provide an invaluable tool for the development of health policy.

References

[1] Black, D.A. & Pole, J.D. (1975). Priorities in biomedical research; indices of burden, *British Journal of Social and Preventative Medicine* **29**, 222–227.

[2] Neuburger, H. & Fraser, N. (1993). *Economic Policy Analysis; A Rights Based Approach*. Aldershot, UK.

[3] UK Department of Health (1994). *Register of Cost-Effectiveness Studies*. Department of Health, London.

[4] UK Department of Health (1996). *Burdens of Disease: A Discussion Document*. Department of Health, London.

[5] Williams, A. (1995). *Measurement and Valuation of Health: A Chronicle*. Centre for Health Economics Discussion Paper 136, University of York, York.

H. NEUBURGER

Burden of Disease

H. NEUBURGER

Volume 1, pp. 573–576

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Byar, David P.

Born: February 23, 1938, in Lockland, Ohio.
Died: August 8, 1991, in Washington, DC.

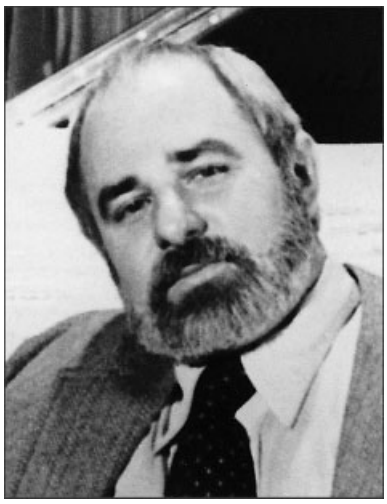

David Byar was a leading **clinical trials** methodologist and proponent of structured experiments for making treatment **inferences**. He played an important role in teaching and advocating the strengths of clinical trials through his positions at the National Cancer Institute. He also made substantial contributions to epidemiology and studies of disease prevention.

David Byar was born in Lockland, Ohio, and attended high school in Maryville, Tennessee, where he graduated as valedictorian. He received an AB degree from Emory University in 1960 and went on to graduate from Harvard Medical School in 1964. After a surgical internship in Denver, Colorado, Dr Byar worked for three years at the Armed Forces Institute of Pathology. It was at this time that he became interested in genitourinary tumors and learned the fundamentals of laboratory experimentation. He became increasingly interested in the sources of variation in laboratory experiments and the statistical methods for coping with them. He began studying statistics and, in 1968, he joined the National Cancer Institute at the invitation of his teacher, Dr John Bailar. Dr Byar assumed the responsibility as statistician for several clinical trials being conducted by the Veterans Administration Cooperative Urological Group. These trials were to be influential in both the clinical treatment of prostate cancer and the methodology of human experiments.

These studies indicated that diethylstilbestrol (DES) was an effective treatment for patients with advanced prostate cancer. The drug was associated with increased mortality from heart disease, so that the treatment was not appropriate for patients with early stage prostate cancer. These data illustrated to biostatisticians the necessity for studying **treatment–covariate interactions** in clinical trials.

In 1972, Dave Byar was appointed Head of the newly formed Clinical and Diagnostic Trials Section at the National Cancer Institute. He would remain in that position for 13 years. The section combined methodologic work in biostatistics and clinical trials with consultation on specific real-world biostatistical problems. Dr Byar built a strong program in methodology and recruited excellent colleagues to this Section. In keeping with one of Dr Byar's favorite quotes by Voltaire, "The price of freedom is eternal vigilance", he was repeatedly called upon to defend the merits of randomized (*see* **Randomization**) clinical trials at a time when emerging computer and database technology suggested that treatment inferences from such sources (*see* **Administrative Databases**) might replace designed experiments.

In 1981, Dr Byar was elected a fellow of the **American Statistical Association**. He was cited "for rare capacity, reflecting an unusual combination of medical and statistical expertise, to bring scientific rigor to clinical testing; for work in statistical theory; and for effectiveness as a communicator between statisticians and medical researchers". In 1984, the Clinical and Diagnostics Trial Section in the Biometry Branch moved to the National Cancer Institute's Division of Cancer Prevention and Control. Dave was subsequently named Branch Chief. During this time, he was very influential in the development of important cancer **prevention trials**, including dietary modification, **screening**, and smoking cessation studies (*see* **Smoking and Health**). He also became active in modifying clinical trial designs for the treatments of **AIDS**.

Dr Byar was elected to the **International Statistical Institute** in 1984. In 1991, he was named an Honorary Fellow of the **Royal Statistical Society** "in recognition of services to statistics".

Dr Byar's academic achievements were widely known and respected. However, he was equally well known as a teacher, friend, and person who enjoyed many aspects of life. He was a particularly accomplished pianist, and once considered a career in music. He loved good food, reading, theater, and engaging friends in stimulating conversation. Dr Byar traveled extensively internationally and had many friends all over the world.

David Byar died on August 8, 1991, following a long illness. He will be remembered always for his scientific contributions and as a source of inspiration to those who learned from and worked with him. On November 7–8, 1991, a scientific symposium titled "David Byar: An Accidental Career" was held at the **National Institute of Health**. David Byar's distinguished career and engaging personality were remembered by his colleagues and friends. The proceedings of this symposium were published in ***Controlled Clinical Trials*** (Vol. 16(4), 1995).

STEVEN PIANTADOSI

Calibration

Introduction

Statistical calibration, sometimes called inverse regression, is often used in biomedicine to estimate the value of one measurement (x) by some other measurement(s) (y) using a **regression** model. The need for calibration arises when the quantity to be calibrated is harder, or more expensive, to measure, or when the value was not recorded and cannot be retrieved. For example, instead of direct analysis through the "wet chemistry" methods, the **radioimmunoassay** (RIA), immunoradiometric assay (IRMA), or enzyme-linked immunosorbent assay (ELISA) can be used to estimate the minute concentration of hormones, enzymes, plasma tissue proteins, and monoclonal antibodies from the measurement of a radioactive count (RIA experiments) or an optical density (ELISA experiments) [29, 32, 77]. Near-infrared (NIR) spectroscopy is commonly used to assay the molecular contents of a sample through the absorbance spectra in a range of wavelengths [47, 57]. For example, in NIR spectroscopy, the spectral readings (x) are very precise but more expensive, while the chemical measurements (y) are less expensive to obtain but not as precise. Oman & Wax [61] describe how multivariate calibration can be applied to estimate fetal age by measuring the femur length and biparietal diameter of the fetus using ultrasound.

A calibration experiment is typically conducted in two stages, namely, the calibration stage and the inverse prediction stage. In the calibration stage, n pairs of training samples $(x_1, y_1), \ldots, (x_n, y_n)$ with known measurements are acquired to estimate the regression function $y = f(x)$, where x corresponds to the compositional variable to be calibrated (e.g. concentration) and y corresponds to the **instrumental variable** (e.g. absorbance of a certain wavelength in the NIR spectroscopy). In the inverse prediction stage, the objective is to estimate the unknown x_0 in a new sample by taking one or more measurements of y_0 from the same sample. The classical estimator of x_0 can be computed by taking $\hat{x}_0 = f^{-1}(y_0)$. Depending on the nature of the variables to be studied, calibration can be classified as the "absolute calibration", in which x is assumed to be measured without error, and the "relative or comparative calibration", in which both x and y are subject to measurement error (*see* **Errors in Variables**). Calibration experiments can also be characterized as the "controlled or designed calibration", where $x_1, \ldots, x_n$ in the training sample are prespecified to cover the range of all possible x_0 or the "nature or random calibration", where a sample of $(x_1, y_1), \ldots, (x_n, y_n)$ pairs is conveniently obtained.

A large collection of literature relating to the calibration problem addresses many areas, such as classical versus inverse estimators, point and **confidence interval** estimation, linear and nonlinear calibration, parametric and **nonparametric methods**, frequentist and **Bayesian** modeling, univariate and multivariate calibration, and many other issues such as measurement error, heteroscedasticity, and **optimal design**. Selected topics are discussed below, followed by a brief description of instrumental calibration and survey sampling. A comprehensive literature review of calibration can be found in [5, 62].

Classical Versus Inverse Estimators

Assume a **simple linear regression** model holds for the training sample: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_i are independent identically distributed (i.i.d.) N(0, σ^2), $i = 1, \ldots, n$. The **least square** estimates of β_0 and β_1 are $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, where $\overline{x}$ and $\overline{y}$ are sample means and

$$S_{xx} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n},$$

$$S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n}. \tag{1}$$

The classical estimator of the unknown x_0 is

$$x_C = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} = \overline{x} + (y_0 - \overline{y}) \frac{S_{xx}}{S_{xy}}. \tag{2}$$

Alternatively, Krutchkoff [40, 41] proposes estimating x_0 by regressing x on y: $x = \hat{\gamma}_0 + \hat{\gamma}_1 y$, where

$$\hat{\gamma}_1 = \frac{S_{xy}}{S_{yy}}, \hat{\gamma}_0 = \overline{x} - \hat{\gamma}_1 \overline{y}, \text{ and}$$

$$S_{yy} = \sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n}. \tag{3}$$

The resulting inverse regression estimator of x_0 is

$$x_I = \overline{x} + (y_0 - \overline{y})\frac{S_{xy}}{S_{yy}}. \quad (4)$$

It can be shown that except when $y_0 = \overline{y}$, $\hat{x}_I$ is always closer to $\overline{x}$ than $\hat{x}_C$ and can be considered as a **shrinkage estimator** shrinking toward $\overline{x}$. By taking a compound estimation approach, Lwin & Maritz [45] show that the classical and the inverse estimators can be derived with and without the asymptotic **unbiasedness** constraint, respectively. Therefore, the classical estimator is a **consistent estimator** but the inverse estimator is not. However, the inverse estimator has a smaller **mean square error** (MSE) than the classical estimator when the unknown x_0 is in the neighborhood of $\overline{x}$, defined as $(x_0 - \overline{x})^2 < S_{xx}[2 + (\sigma/\beta_1)^2/S_{xx}]$ [48]. The classical estimator is the **maximum likelihood** estimator under the appropriate model, while the inverse estimator corresponds to the Bayesian solution under a particular informative **prior distribution** [34]. Perng [64] shows that the inverse estimator can also be derived by **cross-validation** without any distributional assumptions. The difference between the classical and inverse estimators is quantified in [14]. Naszódi [58] proposes another estimator to correct the bias of the inverse estimator and claims that it is more efficient than the classical estimator. Srivastava [71] considers comparison of the inverse and classical estimators in the controlled linear calibration with a multivariate response and univariate **explanatory variable** when the **covariance matrix** is unknown. Later, Oman & Srivastava [60] derive exact expression for the MSE of the inverse estimator and compare with the ones previously derived for the MSE of the classical estimator in multi-univariate linear calibration.

Interval Estimation

The standard method of constructing a confidence interval for x_0 is to apply **Fieller's theorem** [26, 27] for estimating the ratio of two **normally distributed** random variables. The procedure leads to solving a quadratic inequality, and the resulting confidence interval can be a finite interval, a union of two semi-infinite intervals, or the whole real line. Intervals with infinite length occur when the slope of the calibration line is close to zero, where the validity of calibration is questionable. Graybill [31] advocates a two-stage conditional approach: Step 1: test the **null hypothesis** of zero slope at α level, H_0: $\beta_1 = 0$; and Step 2: if the test is not rejected, do not construct the confidence interval because x_0 cannot be estimated well when β_1 is close to zero. If the test is rejected, a $(1 - \alpha) \times 100\%$ interval can be constructed in the usual way and the resulting interval will have finite width. The coverage rate of the conditional confidence intervals is studied independently in [42, 72]. Approximate conditional inference with the angular **transformation** is discussed in [20] (*see* **Delta Method**). In multivariate calibration problem using a multivariate linear model, Mathew & Zha [52] construct some conservative confidence regions, which are nonempty and invariant under nonsingular transformations. Theoretic treatment of the interval estimation can be found in [69].

Cox [15] discusses the direct **likelihood** estimation of ratio parameters and gives the asymptotic variances of the maximum likelihood estimates. For confidence interval estimation, direct likelihood estimation offers a useful alternative to the standard method in linear models with large samples and can also be used for nonlinear models. Rosen & Cohen [67] propose a **bootstrap** confidence interval that is applicable to both parametric and nonparametric calibration curves. Müller & El-Shaarawi [55] investigate M-estimation (*see* **Robustness**) and bootstrapping techniques in the simple linear controlled calibration model and provide different types of confidence intervals for the calibration estimator. Simultaneous calibration intervals are described in [54, 69]. Methods for obtaining confidence bands for **polynomial regression** and **nonparametric regression** are given in [38]. Mathew & Sharma [51] consider the univariate calibration problem of constructing confidence regions for the unknown values of the explanatory variable. An exact confidence region for the multivariate calibration problem is proposed in [49]. Mathew & Zha [53] consider the calibration problem, that is, the calibration data will be used repeatedly in order to construct a sequence of confidence regions for a sequence of unknown values of the explanatory variables. Recently, Schechtman & Spiegelman [68] show that a nonlinear approach to single-use calibration curves gives confidence intervals centered at MLE. The nonlinear approach produces intervals even when the classical approach fails to do so. Mathew & Sharma [50] construct joint confidence regions for several unknown values of explanatory

variable in a normal multivariate linear model (*see* **Multiple Linear Regression**), when the variance covariance matrix is a scalar multiple of the identity matrix or a completely unknown positive definite matrix (*see* **Matrix Algebra**).

Multivariate Calibration

The advance in analytical methods and the wide use of computers have allowed investigators to collect enormous amounts of data easily and quickly; for example, in chromatography, infrared spectroscopy, or flow cytometry. Suppose a set of q instrument responses $\mathbf{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_q)$ are determined from a set of known p-dimensional compositions $\mathbf{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_p)$. The multivariate calibration is to estimate a single unknown $\mathbf{X}_0$ from the observed $\mathbf{Y}_0$. Brown [4] discusses the multivariate calibration in the context of random calibration, controlled calibration, forward and inverse regression, and Bayesian methods. Martens & Naes [47] give a comprehensive treatment on the subject in their book. Two introductory papers [3, 76] also provide useful overviews. Sundberg [74] reviews multivariate calibration in two approaches: the estimation approach (indirect regression – controlled calibration) and the prediction approach (direct regression – natural calibration). Bilinear and other less standard models are also briefly reviewed.

The central idea in multivariate calibration involves dimension reduction. Because the instrument responses $\mathbf{Y}$ are often measured in high dimension and are highly correlated, the standard multiple regression can result in unestimable or unstable models. Two approaches have been used to resolve the problem. One approach is through the principal component regression (PCR) (*see* **Reduced Rank Regression**), which performs the **principal component analysis** first on $\mathbf{Y}$ followed by regressing $\mathbf{X}$ on the principal component scores for calibration. The other approach is the partial least square regression (PLSR), which simultaneously estimates the underlying components in both $\mathbf{X}$ and $\mathbf{Y}$ and then performs the regression. Stone & Brooks [73] show that PCR and PLSR belong to a general class of "continuum regressions".

A generalization of the classical estimator in multivariate calibration can be found in [43]. When $q = 1$, the classical estimator has an infinite mean and mean square error. However, the classical estimator has a finite mean when $q > 2$ and a finite MSE when $q > 4$. The **profile likelihood** approach for multivariate calibration can be found in [6, 7]. Generalized least squares and covariance adjustment approaches are proposed by Naes [56]. Methods applying a stationary autoregressive process (*see* **ARMA and ARIMA Models**) to model serial dependence (*see* **Serial Correlation**), regression **splines**, and minimum length least squares are discussed in [18].

Nonlinear Calibration, Nonparametric Calibration, Robust Calibration, Measurement Errors, and Heteroscedasticity in Calibration

In immunoassay or **bioassay** applications, the **dose–response** curves are often nonlinear (*see* **Quantal Response Models**). For example, a fairly standard approach is to apply the four-parameter **logistic** model: $f(x) = \beta_1 + (\beta_2 - \beta_1)/\{1 + \exp[\beta_4(\log x - \beta_3)]\}$ [28]. Giltinan & Davidian [29] construct a general framework for nonlinear calibration by the nonlinear mixed effects model, which can account for the intra-assay variability. They also study the Bayesian approach and find that the **empirical Bayes** methods can gain considerable efficiency in a **simulation**. Schwenke & Milliken [70] compare interval estimation methods in nonlinear calibration derived from the distribution of $\hat{x}_0$ or the distribution of $\hat{\beta}$. They indicate that, although both methods attain the desired confidence coefficient, the method based on the distribution of $\hat{\beta}$ is more general because it does not depend on a closed-form solution of the inverse function.

Knafl et al. [39] apply the nonparametric regression technique to calibration when the functional form of $f(x)$ is not specified. Chambers et al. [12] study extensively the **robustness** and **efficiency** of various nonparametric estimators derived from the sample empirical distribution, kernel smoothing, and bias-calibrated predictors. They conclude that the nonparametric predictors are more robust and less biased against the model misspecification, but the loss of efficiency is unavoidable. Gruet [32] proposes a new approach leading to a direct statistical inference on the parameter of interest to solve calibration problems in a nonparametric setting. The method combines kernel and robust estimation techniques. Tiede

& Pagano [77] derive an **algorithm** for obtaining the M-estimates of nonlinear calibration curves occurring in radioimmunoassay. They recommend fitting the calibration curve by such a robust **nonlinear regression** procedure, especially when an **outlier** is present in the data. Kitsos & Müller [37] introduce estimators of robust linear calibration based on robust one-step M-estimators, which have a bounded asymptotic bias. Cheng & Van Ness [13] propose robust methods for the random calibration problem. Several approaches based on the standard regression model, the inverse regression model, and the measurement-error model are investigated to robustify calibration. When some assumptions for ordinary least squares are violated, regression techniques including nonparametric and robust approaches to regression analysis are reviewed in [2].

To account for the errors in working standards, Lwin & Spiegelman [46] develop an accurate calibration curve procedure as an extension of calibration intervals. Methods for correcting measurement errors can be found in [9, 10], while Thomas [75] derives a consistent maximum likelihood estimator of x_0 in multivariate calibration with measurement errors. For assays exhibiting variance heterogeneity, Davidian et al. [16] discuss the generalized least square approach of estimating the heteroscedasticity parameter and the calibration parameters. Liski & Nummi [44] consider the problem of prediction and inverse estimation in repeated measures models (*see* **Longitudinal Data Analysis, Overview**).

Bayesian Calibration

A Bayesian approach to calibration is discussed in Dunsmore [23]. He shows that the classical estimator is a special case of Bayesian estimators in which the error due to under- or overestimation is equally punishable in the **loss function**. He also points out that the width of the interval estimator is unaffected by y_0 in the Bayes' methods, while it increases as $|y_0 - \overline{y}|$ increases in the classical method. Hoadley [34] shows that in the simple linear calibration with standardized x_i's and one observed y_0, the inverse estimator is a Bayesian solution when the prior distribution of x_0 is a t distribution with $n - 3$ degrees of freedom, mean 0, and scale parameter $[(n + 1)/(n - 3)]^{1/2}$ (*see* **Student's *t* Distribution**). A thorough discussion of Bayesian methods can also be found in [35] and [17]. Eno & Ye [24] derive a reference prior and corresponding posterior inferences in calibration problem for polynomial regression models. Later, probability matching priors and a reference prior for the linear calibration problem are presented with the constant variance assumption relaxed in [25]. A Bayesian solution to nonlinear calibration is given in [66]. Multivariate Bayesian calibration, including the use of the Gibbs sampler approach to computing the posterior distribution in complex settings, is studied in [21, 22] (*see* **Computer-intensive Methods**; **Markov Chain Monte Carlo**).

Optimal Design and Residual Analysis

The optimal design for calibration is discussed in [63] to minimize the expected MSE $E(\hat{x}_0 - x)^2$. Buonaccorsi [8] gives the optimal design to minimize the asymptotic variance of $\hat{x}_0$. Barlow [1] describes the computation of the optimal design under the Bayesian framework. Oman [59] derives a statistic similar to the Cook's distance in the usual regression setup to measure the influence of a particular observation in calibration (*see* **Diagnostics**). They illustrate the use of **residual** analysis to assist the data analysis and improve the study design. Kitsos [36] considers the simple linear calibration problem through an optimal design theory to evaluate the approximate variance of the calibrating value and provide approximate confidence intervals.

Instrumental Calibration and Survey Sampling

The above sections describe statistical calibration for situations in which the calibration model is obtained to estimate inversely the compositional variable x_0 from the observed instrumental variable y_0. Another type of calibration, instrumental calibration, is also commonly applied in laboratories and industry to calibrate the accuracy and precision between instruments and to establish standards (*see* **Quality Control in Laboratory Medicine**). Cembroski et al. [11] outline approaches to assure optimal proficiency testing in hematology laboratory. Graves [30] gives procedures and criteria to standardize immunoassays for the prostate-specific antigen. Plummer et al. [65] discuss the application of calibration in multicenter cohort studies for correcting the bias at the cohort level and

the subject level. Sample size tables are provided to facilitate the design of multicenter cohort studies.

The application of instrumental calibration with categorical variables in survey sampling (*see* **Sample Surveys in the Health Sciences**) is studied in [33], where a measurement-error model for the data in registers is introduced (*see* **Disease Registers**). Deville & Särndal [19] discuss the use of the generalized raking (or **iterative proportional fitting**) in multiway tables and derive the general regression estimators to estimate the finite population totals in survey sampling on the basis of auxiliary information.

References

[1] Barlow, R.E. (1991). Computing the optimal design for a calibration experiment, *Journal of Statistical Planning and Inference* **29**, 5–19.

[2] Baumann, K. (1997). Regression and calibration for analytical separation techniques. Part II: Validation, weighted and robust regression, *Process Control and Quality* **10**, 75–112.

[3] Beebe, K.R. & Kowalski, B.R. (1987). An introduction to multivariate calibration and analysis, *Analytical Chemistry* **59**, 790–795.

[4] Brown, P.J. (1982). Multivariate calibration, *Journal of the Royal Statistical Society, Series B* **44**, 287–321.

[5] Brown, P.J. (1993). *Measurement, Regression, and Calibration*. Oxford University Press, Oxford.

[6] Brown, P.J. & Sundberg, R. (1987). Confidence and conflict in multivariate calibration, *Journal of the Royal Statistical Society, Series B* **49**, 46–57.

[7] Brown, P.J. & Sundberg, R. (1989). Prediction diagnostics and updating in multivariate calibration, *Biometrika* **76**, 349–361.

[8] Buonaccorsi, J.P. (1986). Design considerations for calibration, *Technometrics* **28**, 149–155.

[9] Buonaccorsi, J.P. (1991). Measurement errors, linear calibration and inferences for means, *Computational Statistics and Data Analysis* **11**, 239–257.

[10] Buonaccorsi, J.P. & Tosteson, T.D. (1993). Correcting for nonlinear measurement errors in the dependent variable in the general linear model, *Communications in Statistics–Theory and Methods* **22**, 2687–2702.

[11] Cembroski, G.S., Engebretson, M.J., Hackney, J.R. & Carey, R.N. (1993). A systems approach to assure optimal proficiency testing in the hematology laboratory, *Clinics in Laboratory Medicine* **13**, 973–985.

[12] Chambers, R.L., Dorfman, A.H. & Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration, *Journal of the American Statistical Association* **88**, 268–277.

[13] Cheng, C. & Van Ness, J.W. (1997). Robust calibration, *Technometrics* **39**, 401–411.

[14] Chow, S. & Shao, J. (1990). On the difference between the classical and inverse methods of calibration, *Applied Statistics* **39**, 219–228.

[15] Cox, C. (1990). Fieller's theorem, the likelihood and the delta method, *Biometrics* **46**, 709–718.

[16] Davidian, M., Carroll, R.J. & Smith, W. (1988). Variance functions and the minimum detectable concentration in assays, *Biometrika* **75**, 549–556.

[17] Davis, W.W. & DeGroot, M.H. (1982). A new look at Bayesian prediction and calibration, in *Statistical Decision Theory and Related Topics III* (in two volumes), Vol. 1, S.S. Gupta & J.O. Berger, eds. Academic Press, New York, pp. 271–289.

[18] Denham, M.C. & Brown, P.J. (1993). Calibration with many variables, *Applied Statistics* **42**, 515–528.

[19] Deville, J. & Särndal, C. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association* **87**, 376–382.

[20] Dobrigal, A., Fraser, D.A.S. & Gebotys, R. (1987). Linear calibration and conditional inference, *Communications in Statistics–Theory and Methods* **16**, 1037–1048.

[21] du Plessis, J.L. & van der Merwe, A.J. (1994). Inferences in multivariate Bayesian calibration, *Statistician* **43**, 45–60.

[22] du Plessis, J.L. & van der Merwe, A.J. (1995). A Bayesian approach to multivariate and conditional calibration, *Computational Statistics and Data Analysis* **19**, 539–522.

[23] Dunsmore, I.R. (1968). A Bayesian approach to calibration, *Journal of the Royal Statistical Society, Series B* **30**, 396–405.

[24] Eno, D.R. & Ye, K. (2000). Bayesian reference prior analysis for polynomial calibration models, *Test* **9**, 191–208.

[25] Eno, D.R. & Ye, K. (2001). Probability matching priors for an extended statistical calibration model, *Canadian Journal of Statistics* **29**, 19–35.

[26] Fieller, E.C. (1940). The biological standardization of insulin, *Journal of the Royal Statistical Society (Supplement)* **7**, 1–54.

[27] Fieller, E.C. (1954). Some problems in interval estimation, *Journal of the Royal Statistical Society, Series B* **16**, 175–185.

[28] Finney, D.J. (1976). Radioligand assay, *Biometrics* **32**, 721–740.

[29] Giltinan, D. & Davidian, M. (1994). Assays for recombinant proteins: a problem in non-linear calibration, *Statistics in Medicine* **13**, 1165–1179.

[30] Graves, H.C.B. (1993). Issues on standardization of immunoassays for prostatespecific-antigen: a review, *Clinical and Investigative Medicine–Medecine Clinique et Experimentale* **16**, 415–424.

[31] Graybill, F.A. (1976). Applications of the general linear model, in *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, pp. 267–340.

[32] Gruet, M.A. (1996). A nonparametric calibration analysis, *Annals of Statistics* **24**, 1474–1492.

[33] Heldal, J. & Spjetvoll, E. (1988). Combination of surveys and registers: a calibration approach with categorical variables, *International Statistical Review* **56**, 153–164.

[34] Hoadley, B. (1970). A Bayesian look at inverse linear regression, *Journal of the American Statistical Association* **65**, 356–369.

[35] Hunter, W.G. & Lamboy, W.F. (1981). A Bayesian analysis of the linear calibration problem, *Technometrics* **23**, 323–328.

[36] Kitsos, C.P. (2002). The simple linear calibration problem as an optimal experimental design, *Communications in Statistics–Theory and Methods* **31**, 1167–1177.

[37] Kitsos, C.P. & Müller, C.H. (1995). Robust linear calibration, *Statistics* **27**, 93–106.

[38] Knafl, G., Sacks, J. & Ylvisaker, D. (1985). Confidence bands for regression functions, *Journal of the American Statistical Association* **80**, 683–691.

[39] Knafl, G., Spiegelman, C., Sacks, J. & Ylvisaker, D. (1984). Nonparametric calibration, *Technometrics* **26**, 233–241.

[40] Krutchkoff, R.G. (1967). Classical and inverse regression methods of calibration, *Technometrics* **9**, 425–439.

[41] Krutchkoff, R.G. (1969). Classical and inverse regression methods of calibration in extrapolation, *Technometrics* **11**, 605–608.

[42] Lee, J.J. (1991). A note on the conditional approach to interval estimation in the calibration problem, *Biometrics* **47**, 1573–1580.

[43] Lieftinck-Koeijers, C.A.J. (1988). Multivariate calibration: a generalization of the classical estimator, *Journal of Multivariate Analysis* **25**, 31–44.

[44] Liski, E.P. & Nummi, T. (1995). Prediction and inverse estimation in repeated-measures models, *Journal of Statistical Planning and Inference* **47**, 141–151.

[45] Lwin, T. & Maritz, J.S. (1982). An analysis of the linear-calibration controversy from the perspective of compound estimation, *Technometrics* **24**, 235–242.

[46] Lwin, T. & Spiegelman, C.H. (1986). Calibration with working standards, *Applied Statistics* **35**, 256–261.

[47] Martens, H. & Naes, T. (1989). *Multivariate Calibration*. Wiley, New York.

[48] Martinelle, S. (1970). On the choice of regression in linear calibration. Comments on a paper by R.G. Krutchkoff, *Technometrics* **12**, 157–161.

[49] Mathew, T. & Kasala, S. (1994). An exact confidence region in multivariate calibration, *Annals of Statistics* **22**, 94–105.

[50] Mathew, T. & Sharma, M.K. (2002a). Joint confidence regions in the multivariate calibration problem, *Journal of Statistical Planning and Inference* **100**, 427–441.

[51] Mathew, T. & Sharma, M.K. (2002b). On the construction of multiple use confidence regions based on combined information in univariate calibration, *Journal of Statistical Planning and Inference* **103**, 151–172.

[52] Mathew, T. & Zha, W. (1996). Conservative confidence regions in multivariate calibration, *Annals of Statistics* **24**, 707–725.

[53] Mathew, T. & Zha, W. (1997). Multiple use confidence regions in multivariate calibration, *Journal of the American Statistical Association* **92**, 1141–1150.

[54] Mee, R.W., Eberhardt, K.R. & Reeve, C.P. (1991). Calibration and simultaneous tolerance intervals for regression, *Technometrics* **33**, 211–219.

[55] Müller, I. & El-Shaarawi, A.H. (2002). Confidence intervals for the calibraton estimator with environmental applications, *Environmetrics* **13**, 29–42.

[56] Naes, T. (1986). Multivariate calibration using covariance adjustment, *Biometrical Journal* **28**, 99–107.

[57] Naes, T., Irgens, C. & Martens, H. (1986). Comparison of linear statistical methods for calibration of NIR instruments, *Applied Statistics* **35**, 195–206.

[58] Naszódi, L.J. (1978). Elimination of the bias in the course of calibration, *Technometrics* **20**, 201–205.

[59] Oman, S.D. (1984). Analyzing residuals in calibration problems, *Technometrics* **26**, 347–353.

[60] Oman, S.D. & Srivastava, M.S. (1996). Exact mean squared error comparisons of the inverse and classical estimators in multi-univariate linear calibration, *Scandinavian Journal of Statistics* **23**, 473–488.

[61] Oman, S.D. & Wax, Y. (1984). Estimating fetal age by ultrasound measurements: an example of multivariate calibration, *Biometrics* **40**, 947–960.

[62] Osborne, C. (1991). Statistical calibration: a review, *International Statistical Review* **59**, 309–336.

[63] Ott, R.L. & Myers, R.H. (1968). Optimal experimental designs for estimating the independent variable in regression, *Technometrics* **10**, 811–823.

[64] Perng, S.K. (1987). A note on the inverse estimator for the linear calibration problem, *Communications in Statistics–Theory and Methods* **16**, 1743–1747.

[65] Plummer, M., Clayton, D. & Kaaks, R. (1994). Calibration in multi-centre cohort studies, *International Journal of Epidemiology* **23**, 419–426.

[66] Racine-Poon, A. (1988). A Bayesian approach to nonlinear calibration problems, *Journal of the American Statistical Association* **83**, 650–656.

[67] Rosen, O. & Cohen, A. (1995). Constructing a bootstrap confidence interval for the unknown concentration in radioimmunoassay, *Statistics in Medicine* **14**, 935–952.

[68] Schechtman, E. & Spiegelman, C. (2002). A nonlinear approach to linear calibration intervals, *Journal of Quality Technology* **34**, 71–79.

[69] Scheffé, H. (1973). A statistical theory of calibration, *Annals of Statistics* **1**, 1–37.

[70] Schwenke, J.R. & Milliken, G.A. (1991). On the calibration problem extended to nonlinear models, *Biometrics* **47**, 563–574.

[71] Srivastava, M.S. (1995). Comparison of the inverse and classical estimators in multi-univariate linear calibration, *Communications in Statistics–Theory and Methods* **24**, 2753–2767.

[72] Steffens, F.E. (1971). On confidence sets for the ratio of two normal means, *South African Statistical Journal* **5**, 105–113.

[73] Stone, M. & Brooks, R.J. (1990). Continuum regression; cross-validated sequentially contrasted prediction embracing ordinary least squares, partial least squares, and principal components regression, *Journal of the Royal Statistical Society, Series B* **52**, 237–269.

[74] Sundberg, R. (1999). Multivariate calibration–direct and indirect regression methodology, *Scandinavian Journal of Statistics* **26**, 161–191.

[75] Thomas, E.V. (1991). Errors-in-variables estimation in multivariate calibration, *Technometrics* **33**, 405–413.

[76] Thomas, E.V. (1994). A primer on multivariate calibration, *Analytical Chemistry* **66**, 795–804.

[77] Tiede, J.J. & Pagano, M. (1979). The application of robust calibration to radioimmunoassay, *Biometrics* **35**, 567–574.

J. JACK LEE & HOJIN MOON

Call-backs and Mail-backs in Sample Surveys

Call-backs in face-to-face and telephone surveys, and mail-backs in surveys conducted by mail (*see* **Surveys, Health and Morbidity**), are indispensable for achieving high response rates and lowering **nonresponse** error. Regardless of survey method, making only one attempt to contact a sampled household or individual will result in unacceptably low response rates and a very high likelihood that respondents will differ from nonrespondents [1, 7]. However, the role of additional contacts in reducing survey error differs somewhat by method.

For interview surveys the first contact is likely to result in substantial numbers of not-at-homes, an outcome that is related to multiple characteristics of the prospective respondent. Younger people, households with more occupants in the labor force, individuals who hold multiple jobs, and those whose life activities keep them away from home during the typical interviewing hours of late afternoon and early evening or for days at a time, are likely to be underrepresented when only one contact is made. For this reason protocols for well-designed telephone interviews often specify that 20 call-back attempts, or even more, be made before declaring that a sample unit is unavailable. In recent years the widespread use of answering machines and telephone number recognition devices that result in telephones not being answered even when people are at home, leaves additional call-backs, which may come at a time when calls are not being screened, as one of the few means with potential for reaching respondents. Furthermore, calls to a multiple-person household may reach a person who is not the desired respondent so that additional calls must be made to obtain that person (*see* **Telephone Sampling**).

Face-to-face household interview protocols are less likely to specify 20 or more call-backs because of the high costs associated with returning to a household that many times. In addition, face-to-face call-backs are more effective than telephone call-backs [8]. The appearance of a face-to-face interviewer at the door is more compelling than a telephone call. And, when no one is at home, clues from neighbors or household indicators (e.g. toys in the yard, newspapers not picked up) may help the interviewer plan the next attempted contact so that it will be more effective.

For mail surveys the situation is somewhat different. If addresses are correct, all sampled households or individuals will, in theory, receive the first contact. However, single mailings inevitably produce response rates that are unacceptably low. In addition, there is no feedback or clues, as is often the case for face-to-face interviews, as to why a questionnaire has not been returned.

Attempting to recontact sample units is only one of the factors that influence attempts to improve response rates and thereby reduce nonresponse error. For all types of surveys, attributes of the survey design, characteristics of the sampled individual, characteristics of the interviewer or mail-out materials, and the interaction between interviewer (or researcher) and prospective respondent are important [9]. For interviews it has been shown that longer questionnaires, survey topics that are uninteresting or threatening to respondents, and the interviewer's lack of experience can all reduce response rates. In addition some types of people, e.g. young adult males and people with poor heath status, are less likely to be interviewed successfully regardless of number of contacts [7] (*see* **Response Effects in Sample Surveys**).

The essential role of a survey design that emphasizes multiple contacts has been articulated by Groves et al. [9]. Their theory of survey participation posits that interviewers can increase response by tailoring their strategies to different individuals and maintaining interaction with them. This strategy suggests seeking another time to recontact the person rather than pressing the interviewee into a decision to be interviewed or not interviewed, thus using additional contacts as a means of avoiding refusals.

For mail surveys it has been shown that structural characteristics such as greater salience of the survey topic, sponsorship of the survey by government or university as opposed to market research company, and type of population (school or employee vs. general public) are likely to improve response rates [10]. Characteristics of the multiple contacts that have been shown to improve response rates include: prepaid token financial incentives [11], special postage (e.g. special delivery or certified),

stamped (vs. business reply) return envelopes, and personalization of correspondence [2], as well as respondent-friendly questionnaire layout [4] (*see* **Questionnaire Design**). Nonetheless, by far the most powerful influence of mail survey response rates is the number of attempts to contact respondents.

Research has also shown that contacts by a different survey mode can improve response rates to both interview and mail surveys. For example, sending a prior letter to prospective telephone respondents explaining the survey can improve response rates significantly [3, 7]. Similarly, a follow-up telephone call to mail nonrespondents can improve response rates by increasing the number of returned questionnaires and identifying ineligible sample units [2]. In addition, a strategy frequently used to increase response rates is to switch from one survey mode to another; for example, starting with the less expensive mail method and contacting nonrespondents in person to obtain the needed information as used for the US Decennial **Census**. However, it has been shown in the case of the Census that offering respondents the alternative of responding by mail or telephone to a mail contact will not improve response [5]. Rather, it is the fact that additional contacts are made by another mode that results in improved response rates.

The indispensable nature of call-backs and mail-backs across survey modes has been convincingly summarized by Goyder [6]. In a **meta-analysis** of nearly 500 different surveys, mostly conducted from the 1940s to the 1970s, he found on average that the mail surveys obtained response rates 7.5% lower than was the case for face-to-face interviews. He also found that the greater number of contacts used for face-to-face surveys partly accounted for this small difference. He models the determinants of response across methods, and shows that responses to different modes are influenced by similar factors, ranging from salience of topic to sponsorship. In this model, the number of call-backs or mail-backs is not only revealed to be an important determinant of response for all methods, but a unifying one that underlies the ability of each method to achieve high response rates.

References

[1] Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. Wiley, New York.

[2] Dillman, D.A. (1991). The design and administration of mail surveys, *Annual Review of Sociology* **17**, 225–249.

[3] Dillman, D.A., Gallegos, J.G. & Frey, J.H. (1976). Decreasing refusal rates for telephone interviews, *Public Opinion Quarterly* **50**, 66–78.

[4] Dillman, D.A., Sinclair, M.D. & Clark, J.R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys, *Public Opinion Quarterly* **57**, 289–304.

[5] Dillman, D.A., West, K.K. & Clark, J.R. (1994). Influence of an invitation to answer by telephone on response to census questionnaires, *Public Opinion Quarterly* **58**, 557–568.

[6] Goyder, J. (1987). *The Silent Minority: Nonrespondents on Sample Surveys*. Westview Press, Boulder.

[7] Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

[8] Groves, R.M. & Kahn, R.L. (1979). *Surveys by Telephone*. Academic Press, New York.

[9] Groves, R.M., Cialdini, R.B. & Couper, M.P. (1992). Understanding the decision to participate in a survey, *Public Opinion Quarterly* **56**, 475–495.

[10] Heberlein, T.A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature, *American Sociological Review* **43**, 447–462.

[11] James, J.M. & Bolstein, R. (1992). Response rates with large monetary incentives, *Public Opinion Quarterly* **56**, 442–453.

DON A. DILLMAN

Cancer Registries

History of Cancer Surveillance

Early forms of cancer surveillance involved registering cancers diagnosed in a population of interest for the purpose of providing accurate statistics on the morbidity and prevalence of cancer. The first attempts to do this were in the early 1700s in London, in Hamburg in 1900, and subsequently in the Netherlands, Spain, Portugal, Hungary, Sweden, Denmark, and Iceland during the period 1902–1908. These efforts were unsuccessful because doctors often refused to fill out the questionnaires needed to document each diagnosed cancer [27].

The first successful attempt at cancer registration took place in Mecklenberg in 1937. The data recorded on each cancer included the name of the patient, which facilitated the elimination of multiple records on the same case. Also, registration cards were sent to all medical practitioners, hospitals, and pathological institutes and there was telephone follow-up for the purpose of obtaining complete ascertainment as well as complete data on each case. Subsequently, similar surveys were conducted in Saxony-Anhalt, Saarland, and Vienna in 1939, but were soon discontinued because of political developments [27].

In the US, the first attempt to register cancers was initiated by the American College of Surgeons (ACOS) in 1921, and involved only malignancies of the bone [2]. Registration was expanded in the next decade to include other malignancies. The first cancer morbidity survey was conducted in 1937–38 in 10 metropolitan areas by the National Cancer Institute (NCI), and subsequent surveys were conducted in 1947–48 and 1969–71. In principle, all cancers diagnosed in residents of these areas during a one-year period were registered during each of the three time periods. A problem with surveys of this type was that the fate of cancer patients was not known. The lack of information on the survival of cancer patients indicated the need for alternative approaches to cancer registration.

In 1971, the National Cancer Act, announced as the "War on Cancer", called for the NCI to "collect, analyze, and disseminate all data useful in the prevention, diagnosis, and treatment of cancer. . .". This legislation led to the creation of the Surveillance, Epidemiology, and End Results (SEER) Program which was based at the NCI.

Case ascertainment for the SEER Program began on January 1, 1973, in several geographic areas of the US and its territories. Those areas that have participated in the program since 1975 include the states of Connecticut, Iowa, New Mexico, Utah, and Hawaii, and the metropolitan areas of Detroit, San Francisco/Oakland, Seattle, and Atlanta. Subsequent additions to the program included 10 predominantly black rural counties in Georgia in 1978, and American Indians residing in Arizona in 1980. In 1992, the program was further expanded to increase coverage of minority populations, especially Hispanics. The two new areas added were Los Angeles County, and four counties in the San Jose/Monterey area south of San Francisco. Alaskan natives in Alaska have been added to those populations covered by SEER. The SEER Program currently includes population-based data from about 14% of the US population and is reasonably representative of subsets of the different racial/ethnic groups residing in the US. Figure 1 provides a map of the SEER areas and Figure 2 gives the percentages and sizes of various populations included in geographic areas covered by SEER [13].

The SEER database contains records on more than two million cancers and is growing at the rate of more than 160 000 records per year. Other data resources used by the SEER Program include cancer mortality data by county for the total US, obtained from the National Center for Health Statistics (NCHS). To provide for the calculation of incidence and mortality rates, population estimates are obtained through an interagency agreement with the Census Bureau.

Other organizations are also involved in cancer surveillance activities in the US. The North American Association of Central Cancer Registries (NAACCR) was organized in 1987 as an umbrella organization for cancer registries, governmental agencies, professional organizations, and private groups in North America interested in enhancing the quality and use of cancer registry data. Most population-based cancer registries in the US and Canada are members. The mission of NAACCR is to support and coordinate the development, enhancement, and application of cancer registration techniques in population-based groups, so that data of high quality and completeness may be used for epidemiologic research, public health programs, and patient care to reduce the burden of cancer in North America.

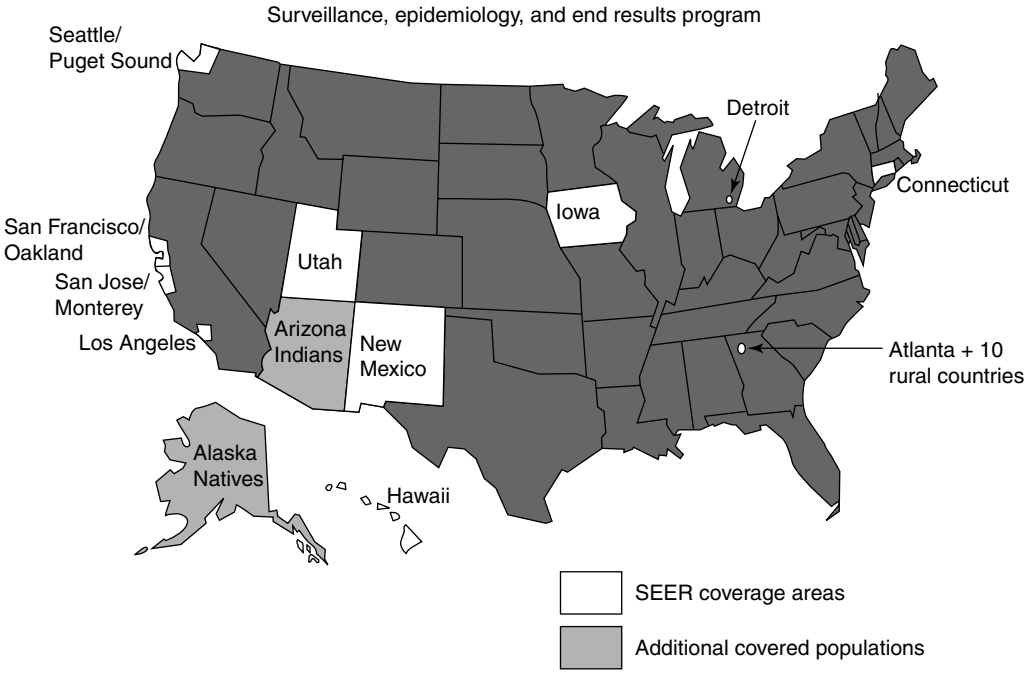

Figure 1 Map of US indicating areas and populations covered by the SEER Program

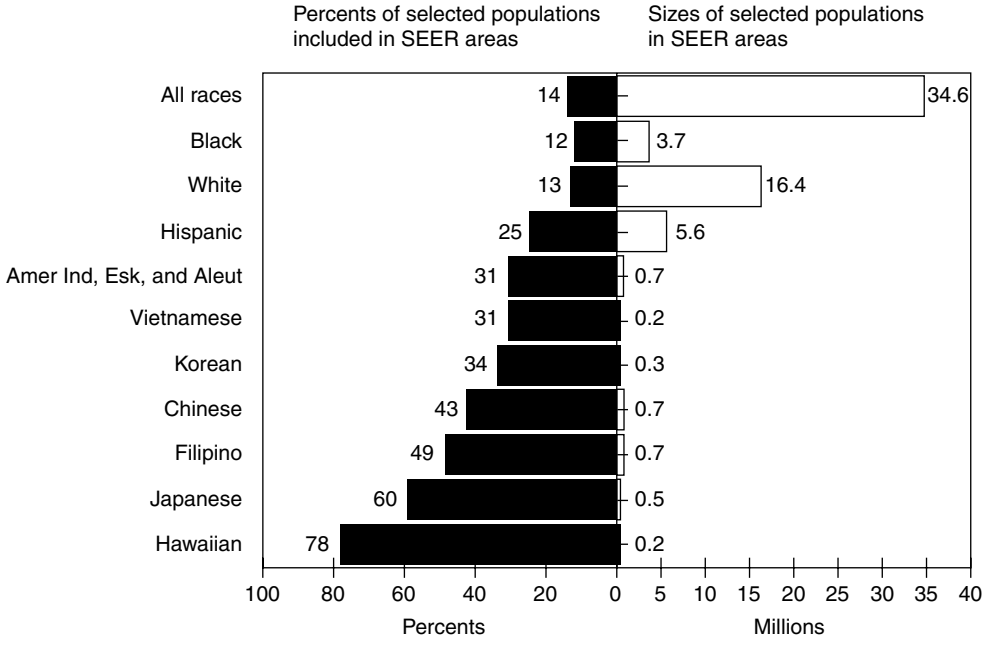

Figure 2 Percentages and sizes of various populations covered by the SEER Program

The American College of Surgeons Commission on Cancer and the American Cancer Society jointly founded the National Cancer Data Base (NCDB), which is a nationwide oncology outcomes database that includes data from over 1500 hospitals in 50 states, and is in its tenth year of operation. This database can be used to study patterns of care and factors associated with patient outcome, and patient care evaluation studies are periodically carried out in participating cancer registries [25]. Uses of this database focus on clinical surveillance of people with cancer, and cannot be used to calculate incidence rates.

Current Cancer Registration Practices

Cancer registration is the process of collecting data about patients with malignant diseases. The data collected identify the demographics of the patient with the disease, the type of cancer, how it is treated and the outcome of the patient. The data collected reside in a cancer registry, a term which can mean simply the database or data system that manages and analyzes the information or the data system and all of the associated systems and personnel who perform cancer surveillance and cancer control. Cancer registries serve several purposes.

A *hospital-based cancer registry* is a cancer data base maintained in a health care facility to collect pertinent information on all cancer patients who use the services of that facility for diagnosis, staging, and treatment. The service area for a hospital-based registry varies from facility to facility, depending on the types of specialty treatment it offers, the types of third-party payors it attracts, and a number of other factors. As a result, the number of potential patients in the facility's customer base can only be estimated. A hospital-based cancer registry can calculate frequency of cases and measure outcomes for the patients it monitors. A hospital-based cancer registry cannot calculate incidence rates because the denominator population is not known.

A *population-based cancer registry* is a centralized cancer database covering a known population, usually residents of a defined geographic area, such as a county or state. Because the population denominator can be counted or estimated by a census, a population-based registry can calculate incidence rates. Population-based registries are the principal source of cancer surveillance data.

Population-based registries must gather information on cancer patients from a variety of sources, including registries in hospital and other healthcare facilities, physician offices, pathology laboratories, and facilities outside the defined geographic area to which residents travel for cancer diagnosis and treatment. Population-based registries can be of two types: (1) those that report incidence only (the first report of a new cancer) or (2) multipurpose registries that collect data on incidence and subsequent outcomes.

A population-based registry is one type of *central registry*. A central registry collects data from a variety of sources but it may not be population-based. For example, a provider of cancer registry software may maintain a pooled database of all the cancer cases submitted by its customers, or a hospital corporation may pool the cases from all facilities it owns. In each of these cases, it is not possible to determine an appropriate denominator, so these central registries are not population-based.

Registry Operations

The four main aspects of registry operations are: case ascertainment, abstracting and coding, follow-up or mortality follow-back, and quality control. Data collection procedures will also be reviewed.

Case Ascertainment

Case ascertainment, also called casefinding, is the process of identifying patients with malignant disease who meet the criteria for inclusion in the registry. Because cancer surveillance requires monitoring of cancer incidence and mortality, case ascertainment must identify all cases of the disease in a defined population, regardless of where the cancer patient encounters the healthcare system; including hospitals, independent treatment centers, clinics, pathology laboratories, physician offices, and nursing homes. For practical purposes, the principal source of cancer information is the hospital health information or medical record, which includes all contacts with the hospital inpatient, outpatient and clinic. The medical record contains reports of diagnostic and staging procedures, physical examination, operations and other treatments. In addition, consultation reports from

outside pathology departments and physicians are usually retained as part of the medical record. Medical records are maintained as legal documents in most facilities where patients are treated; the exception is the pathology laboratory.

Most cancer patients come to a hospital at some point in their disease process, usually for a biopsy or treatment; thus, hospital medical records are an important source of casefinding. In hospitals, medical records are coded and indexed by disease and procedure so that patient records can be retrieved for analysis. The database containing these codes is one of the principal sources of case ascertainment in a healthcare facility. Specific codes for cancer diagnosis and treatment permit retrieval of records pertaining to reportable neoplasms that must be included in the registry.

A neoplasm is a "new growth" or tumor that develops somewhere in the body. The term neoplasm refers to either benign or malignant (having the potential to spread from the site of origin and ultimately kill the patient) tumors. A *reportable neoplasm* is a tumor that meets the inclusion criteria for a registry. Reportable neoplasms are well defined in the *International Classification of Diseases for Oncology*, a coded nomenclature published by the World Health Organization (WHO). This coding system defines each type of tumor and its behavior: benign, uncertain malignant potential, *in situ*, invasive, or metastatic. The reportable neoplasms collected by all general-purpose cancer registries are those that are malignant (*in situ* or invasive). Metastatic tumors (malignancy growing in a site at a distance from the organ in which it started) are not reported individually; rather, metastases are reported as progression of the tumor at the site of origin. Occasionally a central registry will require that another type of tumor be reported, such as benign brain tumors, which cannot spread but do have the potential to be lethal, and tumors of uncertain malignant potential, such as carcinoids of the appendix. On the other hand, a few cancers are very common and are associated with such a good prognosis that it is generally not necessary to monitor their outcomes, such as basal cell and squamous cell carcinomas of the skin and carcinoma *in situ* of the cervix.

All the inclusion and exclusion guidelines for case ascertainment are compiled into a *reportable list*, which the data collector uses to identify cases to be abstracted for the registry.

Abstracting and Coding

Abstracting is the process of deriving and recording pertinent data about each reportable case. The resulting document, the *abstract*, is an abridgment or summary of what happened to the patient, and may be in paper or electronic form. Data items include demographics of the patient, a description of the disease (site of origin, type of malignancy), stage at diagnosis (documentation of how far the cancer had spread when it was diagnosed), treatment, and the course of the disease from the time it was diagnosed. Parts of the abstract are encoded, such as site and type of cancer, stage, and treatment. In addition to the standard data items, some registries collect information on items of special interest, such as smoking history, family history of cancer, or co-morbid conditions.

The abstracting process is exacting and highly technical, requiring great attention to detail. The aim of abstracting is to collect the data about each cancer case as accurately as possible (high correlation between source document and abstract) and as consistently as possible for similar cases (all cases following the same rules). Abstracting rules and guidelines have been developed to cover nearly every situation, but human interpretation of both the facts of the case and the rules of abstracting can sometimes cause problems, and there is always the danger of incorrect data entry. As a result, a series of edits have been developed and included in most cancer registry database systems.

There are several types of edits, including range checks and logic checks. The simplest edits are range checks or allowable codes. Logic checks are a type of inter-item edit where the program looks at two or more data fields to ensure that they make sense together. For example, an error message should be generated when "sex" is male and "primary site" is cervix. Inter-item edits can be quite complex, such as looking at a morphologic diagnosis code as noninvasive, the corresponding stage at diagnosis coded as *in situ*, the method of diagnostic confirmation coded as histologic, and the sites of distant metastasis fields that should be left blank. Computer edits such as these are the first line of defense against inaccurate data. Other editing mechanisms and preventive measures such as training and standardization of procedures are described in the following sections on data collection and quality control.

An additional function of population-based central registries is case consolidation or case matching.

Because the registry receives reports from many sources, it is necessary to identify multiple reports on the same patient so that the case is not counted more than once. Case consolidation involves not only various computer algorithms but human review as well. For example, Hospital A might send in a report on Ric Smith with a birth date of 11-19-35 and a diagnosis of sigmoid colon cancer, and Hospital B might send in a report on Frederic Smith with a birthdate of 11-18-35 or 11-19-36 and a diagnosis of rectal cancer. The registry must decide whether these reports are about the same patient, and, furthermore, whether they are about the same cancer. Without a case consolidation operation in the registry, the numerator (newly diagnosed cases) of the incidence rate may be inflated.

Data Collection Procedures

When data collection for the SEER Program began in 1973, it was imperative that data be collected uniformly and systematically in all participating areas. As a result, the SEER Program published a series of manuals providing specific rules for case inclusion and staging. The most recent of these is the *SEER Program Code Manual, 3rd Ed.* [20]. Since its inception, the SEER Program has been a leader in documentation of data collection rules, training of data collectors, and quality assurance of the data collected. Many central registries in the US follow SEER rules even though they are not funded by the National Cancer Institute.

In addition to these coding guidelines developed in the US, an international body established definitions of what was considered to be cancer. The WHO has been publishing revisions of the International Classification of Diseases (ICD) on a decennial basis since 1893. Originally developed to code mortality, ICD has been modified to code all types of diseases and conditions, and the current edition, ICD-9-CM (Ninth Revision, Clinical Modification) is the coding standard for health care facilities and reimbursement through federal Medicare programs [26]. The next edition, ICD-10, is in use in vital statistics offices to code death certificates [30]. As the WHO began development of the ninth revision in the early 1970s, clinicians expressed a desire for a more complete coding system for neoplasms, one that would describe both where the tumor started (topography) and what the tumor was (morphology). ICD contained a coded list of anatomic sites for the topography, and another coding system, the *Systematized Nomenclature of Pathology* (SNOP), published by the College of American Pathologists, contained the codes for cell types or morphology [4]. SNOP was a functional descendant of the *Manual of Tumor Nomenclature and Coding* (MOTNAC), published in 1951 and revised in 1968 by the American Cancer Society. The WHO used the topography code structure from ICD-9 and selected the code structure from SNOP for the morphology codes. The first edition of the *International Classification of Diseases for Oncology* (ICD-O) was published in 1976 by the WHO [28]. A second edition using the alphanumeric topography codes from ICD-10 was published in 1990 and implemented in the US in 1992 [29]. A third edition of ICD-O is scheduled for publication in 2000. The College of American Pathologists maintains the descendant of SNOP, called the *Systematized Nomenclature of Medicine* (SNOMED) [5,6], now in its fourth generation as SNOMED RT (Reference Terminology) [24]. By international treaty, the neoplasm sections of SNOMED and ICD-O are identical, although the topography codes differ between the two coding systems.

US cancer organizations collected data for specific purposes; the American College of Surgeons for quality management of patient care, and the SEER Program for incidence, survival, and mortality statistics. For many years there was no effort on the part of these organizations to collaborate on the development of data-collection rules. An example of the resulting problems relates to the collection of data pertaining to stage of disease at diagnosis. There are currently four major staging systems in use in the US: Tumor–Node–Metastasis (TNM), a product of the International Union Against Cancer and the American Joint Committee on Cancer [1]; SEER Extent of Disease (EOD) [19]; Summary Staging [16] and SEER Historic Stage (local–regional–distant) [18]. These staging systems are not comparable for a number of cancers.

The American College of Surgeons Commission on Cancer (COC) uses TNM as the standard for coding stage of disease for hospitals participating in its approvals program. The SEER Program uses EOD as its data-collection standard and some versions of TNM can be derived from it [19], and SEER historic stage as its reporting standard which can be derived from EOD [18]. The National Program

of Cancer Registries (NPCR) uses Summary Stage as its standard [16]. As a consequence, a registry approved by the American College of Surgeons COC in a state receiving NPCR funds and an area where SEER data is collected must stage each case using three different systems, each having their own codes, timing rules, and inclusion/exclusion criteria. Efforts are in progress to define a single data set for staging and a single set of rules [10], but these efforts are far from fruition, much less implementation, data collection and analysis.

In the early 1980s, the SEER Program and the American College of Surgeons COC began meeting to resolve differences in data fields, such as field lengths, definitions, and code structures. The 1988 publications of the COC's *Data Acquisition Manual* [7] and *The SEER Program Code Manual*, 2nd Ed. [17] were in substantial agreement.

In 1987, the population-based central registries in the US and Canada formed an "organization of organizations" to share information on coding practices, registry operations, standards and other factors that affect the accuracy and reliability of published cancer information. One of the first activities of the NAACCR was to establish the Uniform Data Standards Committee (UDSC). The UDSC formalized the standardization efforts begun by SEER and the COC. This committee, consisting of representatives from all the standard-setting organizations, central registries, data collectors, registry software vendors, and other users of registry data, serves as a forum for identifying and resolving problems in data collection. The committee compiled all the rules regarding data collection and identified areas of discrepancy, publishing four volumes of standards in 1994: *I. Data Exchange Standards and Record Description*; *II. Data Standards and Data Dictionary*; *III. Standards for Completeness, Quality, Analysis and Management of Data*; and *IV. Standard Data Edits*.

Adherence to coding rules established by the UDSC and the NAACCR in general is voluntary. However, in the current practice of cancer registration in the US and Canada, all revisions to existing data fields, coding guidelines and data-collection rules, as well as proposed new data fields, data record layouts, and other enhancements, are discussed and voted upon by the members of the UDSC. An implementation date for approved changes is widely published so that software vendors can make changes in sufficient time to meet the needs of cases diagnosed after the implementation date, and the standard setters can publish necessary revisions to their data-collection manuals.

Outcome Measurements and Quality Control

As noted previously, a population-based cancer registry can be either incidence-only or multipurpose. If the registry is incidence-only, then the registry does no outcomes assessments. Outcomes measurement is the current vernacular for describing the results of treatment and the disease process in terms of survival rates and mortality. Outcomes processes include follow-up and mortality follow-back, two specific additional operations performed by multipurpose central registries. Follow-up is long-term surveillance of cancer patients. Once a patient is treated and rehabilitated, he or she resumes a relatively normal life, but monitoring for disease recurrence or sequelae of treatment must continue for the patient's lifetime. Follow-up is the process of contacting someone – either the patient directly or the patient's physician – to obtain current information on the status of the cancer. Ideal follow-up information includes a recent date of last contact, vital status (alive or dead), and disease status (free of disease, recurrent disease, a subsequent primary cancer, additional treatment, etc.). Most registries prefer to contact the patient's physician for this information as it will be more accurate, technical, and specific than that received from the patient. However, response rates are generally good when patients are contacted directly. Either type of direct contact is called *active follow-up*.

If a registry chooses not to contact the patient, follow-up information less complete than the ideal can be obtained by linking the cancer registry database with other governmental databases, such as voter registration, local tax rolls, and Department of Motor Vehicles (DMV) drivers' license renewal files. Little can be determined from these linkages other than the patient's vital status, and that might only be at the last point of contact with the agency. For example, DMV would only have a record of the last time the patient renewed his driver's license (possibly several years previously) or reported a change of address. This indirect method of obtaining the vital status of the patient is called *passive follow-up*. Another method of obtaining follow-up is linkage to Social Security Administration death lists and to the National Death

Index; this linkage will update only deceased patients, however.

Tracing a patient who no longer regularly visits a physician for his disease is both an art and a science. Confidentiality guidelines must be observed when information is requested about any patient, but the higher the percentage of complete follow-up, the more reliable are estimates of survival rates. The SEER standard for complete follow-up is to have current information (within the past 15 months) on at least 95% of all cases.

Occasionally a patient with cancer will be missed in the case ascertainment process and not be abstracted into the registry database. Missing a case lowers the incidence rate for that particular cancer; thus high standards of case completeness are important for accurate and reliable cancer data. If a previously unreported cancer patient dies, a cancer diagnosis on the death certificate may be the first and only report of the cancer case. It is good policy for a registry to follow-back a Death Certificate Only (DCO) case to see where it was missed in the casefinding process. Follow-back is the process of contacting physicians and facilities noted on the death certificate to review their medical records to determine any earlier diagnosis or treatment of the cancer. In many instances the case was simply missed, so the abstract is processed as a late report and reporting-source procedures are investigated. In other instances the death certificate diagnosis is the only identification of the case. These cases are tagged as DCO in the database, and usually very little information is known about them. A registry monitors its percentage of DCOs as part of its quality control processes.

Quality control encompasses all registry activities that monitor and resolve data problems. Quality control usually deals with facts and data items. On the other hand, quality improvement or quality management usually deals with procedures and processes. Quality control is performed on all aspects of registry operations. Standards have been established for case completeness, database completeness, accuracy and reliability of data, and timely reporting of cases to the registry. The purpose of quality control is to determine whether these standards are being met.

Quality has been defined as "fitness for use". Analysis of data which are not fit for use can result in incorrect conclusions and inappropriate cancer control and cancer surveillance activities. The principal components of data quality are accuracy, completeness and timeliness [9].

Completeness has at least two aspects: completeness of the database and completeness of the data in each record. Completeness of the database means that all cases in the population under investigation have been included for the specified time period. Without a complete database, incidence rates and relative frequencies may be inaccurate. Completeness of the database is a function of thorough case ascertainment, described above. Completeness is assessed by several techniques, including re-casefinding studies, projections of the number of cases reported in previous years, and the ratio of incidence to mortality for all cancers combined and for selected cancers. The SEER Program's target rate for database completeness is 98% complete reporting for a diagnosis year at the time the data are first submitted (14 months after the end of the diagnosis year).

Completeness of the data is a function of abstracting. This means that all data have been reported and there are no omissions, unnecessary blanks, or fields coded as unknown that should have been completed. It is possible to have a data field considered complete because there are no blanks, but unusable because the data are coded as unreportable or unknown. However, a data collector must find a balance between tracking down every last data item for 100% completeness, and coding unknown if the data are not easily obtained.

Timeliness is a corollary to completeness. It is presumed that every case may eventually be found, but the issue is how long to wait before using the data. There is a tradeoff between having potentially incomplete data available for use quickly, and having complete data available after so long a wait that the data are no longer current or useful. To meet the needs of most (if not all) data users, it is necessary to set a cut-off date and assess completeness at that time. Thus, timeliness is determined by setting a final date for data submission, and ensuring that all records have been submitted by that date.

Accuracy is the quality measure that establishes the reputation for a registry. Accuracy is necessary in all parts of registry operations. Accurate incidence rates neither overcount nor undercount the number of cases in the population. Accurate abstracting ensures that results are appropriate for research. Accurate follow-up permits survival rates and other outcomes to be measured correctly.

The SEER Program performs quality control studies annually that are designed to provide quantitative assessments of accuracy and completeness of the data collected by the various participating registries. Some findings from these studies have been published [31]. In the near future, findings from all quality control studies designed to provide estimates of completeness and accuracy both at the registry level and for all registries combined will be available on the Internet at `http://www-seer.ims.nci.nih.gov/`

Current Cancer Surveillance Activities

The discussion here will focus on the activities of the NCI, since it has been in the forefront of cancer surveillance activities for more than 60 years. As previously mentioned, other organizations are involved in cancer surveillance; however, an attempt will not be made to associate specific activities with specific organizations. At this time it is not clear how cancer surveillance responsibilities in the US will ultimately be divided, as that is currently being negotiated by the parties involved. However, it is reasonable to assume that the NCI will continue to play a major role in all aspects of cancer surveillance.

The fundamental tool of cancer surveillance is the population-based cancer registry. The establishment of the SEER Program in 1972 was the beginning of a new era in cancer surveillance in the US. Beginning in 1973, the continuous registration of all cancers diagnosed in residents of geographic areas initially covered by SEER allowed the calculation of incidence rates for calendar years beginning in 1973. Follow-up to determine vital status for all cancer patients in SEER areas, including the coding of cause of death, allowed the calculation of survival rates as the SEER database matured.

A question frequently raised is: How representative are the SEER areas of the total US population? To address this question, it is necessary to define "representativeness" in this context. It is certainly desirable to be able to derive cancer rates from SEER areas that approximate those for the total US by age, sex, and racial/ethnic group. But it is probably more important to be able to establish that the trends in cancer rates in SEER areas approximate those for the total US. SEER rates have been assumed to be reasonably representative of those from the total US. However, in the future, cancer rates in SEER areas will be modeled using ecologic data available from the census in order to refine national estimates.

Cancer mortality data for the total US have been used to examine the representativeness of trends in SEER areas versus those for the total US. This was done by systemically comparing trends in cancer mortality for selected cancer sites in SEER areas with those for the total US. It was concluded that, with few exceptions, mortality trends for selected cancers in SEER areas were representative of those for the total US [11]. Therefore, it seems reasonable to assume that trends in SEER incidence rates approximate those for the total US. Further information on the representativeness of SEER areas is given in Figure 3, which compares selected ecologic data from the 1990 census in SEER areas with that for the total US.

A variety of reports are available on cancer rates in SEER areas, the most recent of which includes data for the time period 1973–1996 [21]. There are also reports on cancer incidence rates published by the NAACCR which include data from SEER registries plus other US and Canadian registries that have been found to have data of high quality. The most recent NAACCR publication is for the time period 1992–1996 [3].

The scope of cancer surveillance has broadened considerably since the early attempts at cancer registration. Current cancer surveillance activities include: developing and reporting estimates of cancer

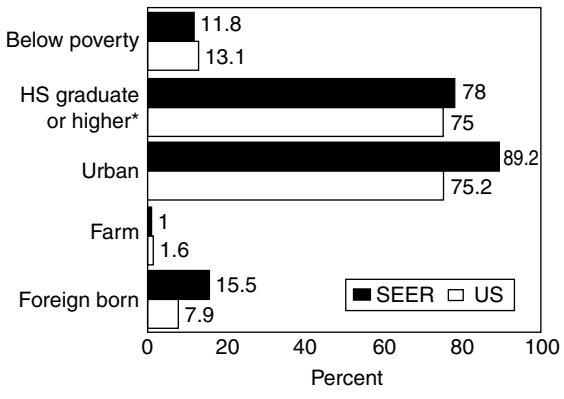

Figure 3 Comparison of selected ecologic variables from the 1990 census for the total population in SEER areas versus the total US population

incidence, prevalence, and mortality on a periodic basis for the total US; monitoring annual cancer incidence trends to identify unusual changes in specific forms of cancer occurring in population subgroups defined by geographic, demographic, and social characteristics and providing insight into their etiology; and providing continuing information on changes over time in the extent of disease at diagnosis, trends in therapy, and associated changes in patient survival. Of particular importance is the inclusion of sufficient numbers of various racial/ethnic populations, and other populations defined by a variety of measures including access to medical care, urban versus rural, and measures of poverty and socioeconomic status. Such research can benefit cancer prevention and control activities, and identify areas where improvements in treatment may be needed. Other uses of incidence, prevalence and survival data include identifying cancer sites showing unusual rates of increase or decrease that would warrant special etiologic investigation (e.g. non-Hodgkin's lymphoma increases associated with the acquired immune deficiency syndrome (AIDS) epidemic); helping health policy-makers set priorities for spending on research and for allocating resources among etiologic, prevention, diagnosis, treatment and control areas; and informing the general public and Congress on the extent and trends in the cancer burden. These data also have direct clinical relevance for advising individual patients. For example, age- and race-specific SEER data were used (together with other data on specific risk factors) to help develop a model for projecting the chance that a woman with particular risk factors would develop breast cancer in a given time period [8].

Other important research components to cancer surveillance include promoting studies designed to identify cancer risk factors amenable to cancer control interventions. These studies may pertain to the environment, occupation, socioeconomic status, tobacco, diet, screening practices, patterns of care, and determinants of the length and quality of patient survival. Other areas of investigation include planning, conducting, and supporting research related to evaluating patterns and trends in cancer rates and cancer-related risk factors. Also studied are health behaviors, cost of care, patient outcomes, health services as part of an attempt to determine the influence of such factors at the individual, societal, and systems level on patterns and trends in the various measures of cancer burden. Also included in this activity are identifying, improving and developing databases and methods for cancer-related surveillance research; maintaining, updating, and disseminating these databases and methods; and promoting and facilitating their use among investigators within the extramural research community and federal agencies. No attempt will be made to document all of these activities; however, information about them can be obtained from the following Internet sites: `http://www-dccps.ims.nci.nih.gov/arp/` and `http://www-dccps.ims.nci.nih.gov/srab/surveillance/survdesc.html`

The SEER contracts are primarily with cancer research organizations affiliated with universities. Thus, they provide an infrastructure for conducting analytic epidemiologic studies on a variety of emerging issues in cancer prevention and control which can be used by the NCI. The ability to do special studies was established in the early 1990s. The workscopes of SEER contracts were modified to include the capabilities of interviewing patients, conducting surveys of the covered populations, obtaining biological materials from patients and survey respondents, conducting methodologic research which utilizes cancer registry data, and establishing tissue banks. Standard competitive procurement procedures within the SEER framework have been used to plan and fund studies on a wide range of topics that have included identification of risk factors, quality of life, statistical modeling, etiology of trends in cancer rates, and operational issues pertaining to data collection and reporting. No attempt will be made to document the findings from these studies here, but they have resulted in a large number of peer-reviewed publications in scientific journals on a variety of issues pertaining to etiology, cancer control, quality of life, and registry operations [13].

There have also been significant efforts to involve the general research community in cancer surveillance activities. This has been done by distributing to cancer researchers public use files that include SEER data. The files are made available on CD-ROM and include more than two million individual records of cancers registered in SEER areas from 1973 to the most recent year for which complete data are available, population data, and documentation of all files. Also included is SEER*Stat which

is a free Windows-based computer program developed by the SEER Program to calculate incidence rates, frequencies, trends, and survival rates. Another program, called SEER*Prep, allows registries outside of the SEER Program to put their data into SEER*Stat, greatly facilitating analysis of their data. Currently, about 1500 SEER public use files are distributed annually, and a number of non-SEER registries are using the SEER*Stat software via SEER*Prep. More information about this software can be obtained on the Internet at: `http://www-seer.ims.nci.nih.gov/scientific systems/SEERStat/`

Recent developments in statistical methodology pertaining to the analysis of trends in age-adjusted rates deserve mention, since the analysis of trends is a fundamental activity of cancer surveillance. The first is the use of join point regression using log linear or linear models to describe trends in age-adjusted rates [14]. Models of this type assess the statistical significance of recent changes in trends as well as describe the trends over the period for which they are fit. Figure 4 presents a fit of a log linear join point regression model to the age-adjusted (1970 US Standard) mortality rates for all cancers combined for the total US for the total population and by sex. Annual percent changes and join points are given to describe the trends.

A second methodology of interest partitions a trend based on fitting linear regression models [12]. For example, it is possible to derive the relative contributions of various individual cancers or groups

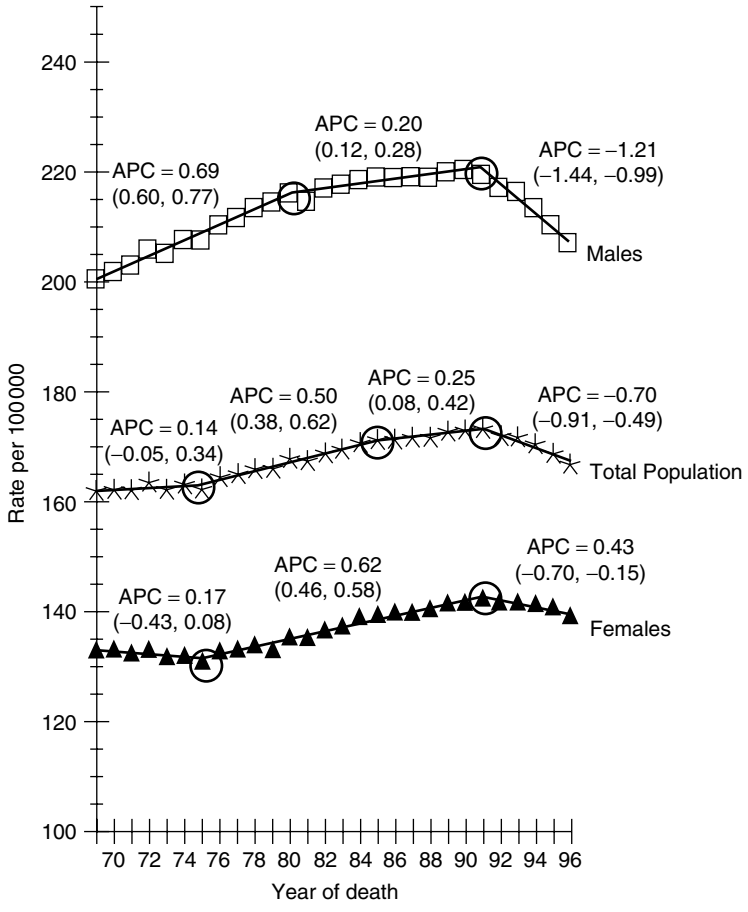

Figure 4 Fits of log linear join point regression models to cancer mortality trends based on age-adjusted rates (1970 US Standard) for the total population of the US and by sex. The linear segments are indicated by a solid line, and the join points are circled. The observed points are also given

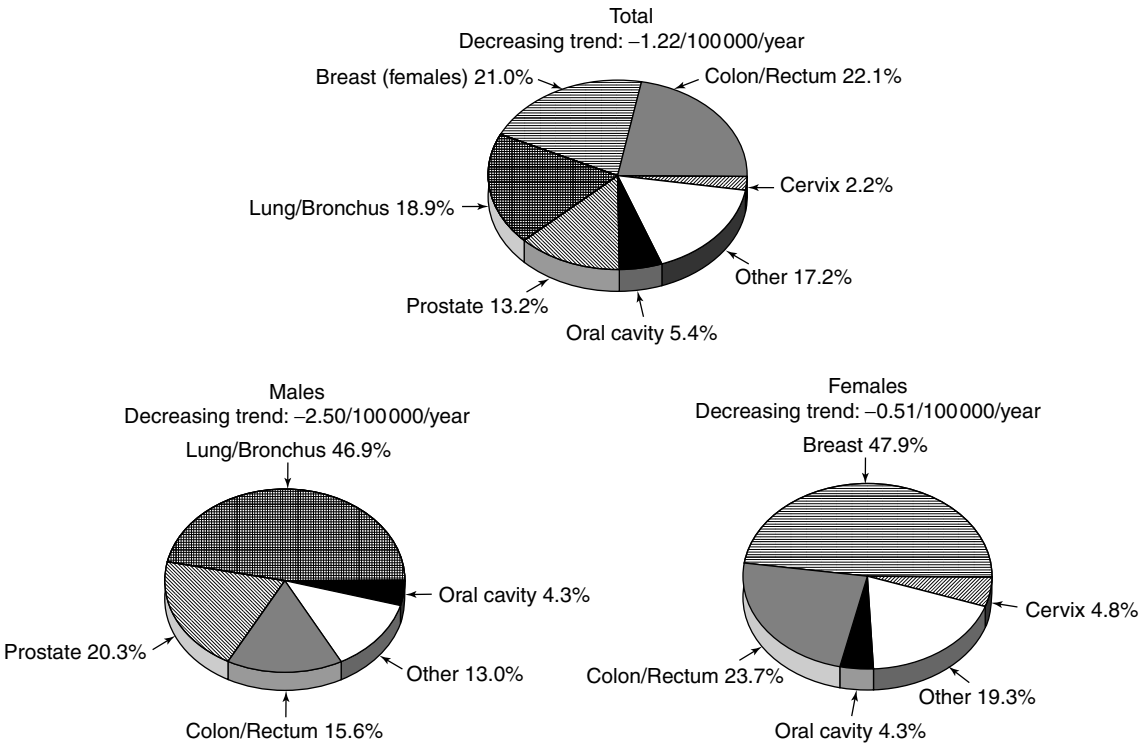

Figure 5 Partition of the cancer mortality trend from 1991 to 1996 based on fits of a linear regression model to the age-adjusted rates (1970 US Standard) for the total US population and by sex

of cancers to an increasing or decreasing trend in the age-adjusted rate for all cancers combined. This type of analysis provides useful information about the impact of targeted interventions on the overall trend in age-adjusted rates for a group of diseases of interest. Figure 5 presents a partition of the most recent cancer mortality decrease for the total population and by sex based on rates adjusted to the 1970 US Standard. The contributions of cancers for which interventions have been introduced into the general population are given. Thus, it is possible to quantify the contributions of cancers of the lung, oral cavity, colon and rectum, female breast, cervix, and prostate to the decreasing trend.

The Future of Cancer Surveillance

Medical practices have led to a new role for cancer surveillance. In theory, prevention, screening, and treatment interventions are tested for efficacy by randomized controlled trials. If such trials demonstrate that an intervention is efficacious, then it is introduced into the general population. An important role for cancer surveillance is to make an assessment of the impact of the intervention in the population by analyzing trends in incidence, survival, or mortality rates as appropriate. This paradigm has been violated in some cases in the past, particularly in the development of new screening tests where new tests have been introduced into the general population before establishing their efficacy in regard to reducing mortality. A prime example is the Prostate Specific Antigen (PSA) test for detecting prostate cancer [15]. The use of Spiral CT for diagnosing lung cancer may be a second example [23]. This practice has resulted in the use of surveillance data to not only establish some measure of the impact of the introduction of such a new test on population cancer rates, but to also make some assessment of its efficacy. If such practices continue, it will likely result in the establishment of more sophisticated surveillance systems directed toward accommodating this expanded role.

Cancer surveillance activities at the NCI have been reviewed by a committee of researchers from within and outside the Institute [22]. Recommendations for future directions have been made in a number of areas. The first priority is to expand the scope of surveillance research through additional data collection and methods development. Specific activities will include collection of data on patterns of care, health status, and quality of life, as well as cohort studies of newly diagnosed cancer patients for the purpose of documenting levels and trends in these parameters; collection of risk factor and screening data in defined populations, particularly those covered by high quality cancer registration; development of research methods to measure the dimensions of the cancer burden and factors affecting the burden, as well as methods to explain patterns and trends in cancer rates; and exploration of the feasibility and utility of employing geographic information systems for geocoding surveillance data and reporting geographic relationships among screening measures, risk factors (including environmental exposures), and improved cancer outcomes.

A second area of focus is to expand the scope of surveillance to improve the representativeness of cancer burden estimates. Specific activities will include expanding SEER population coverage to improve representation of ethnic minority and underserved populations including rural African Americans, Hispanics from Caribbean countries, American Indians, residents of Appalachia and other rural areas, especially those of lower socioeconomic classes; developing methods for improving national estimates of the cancer burden; and working with other organizations involved in cancer surveillance to develop a national cancer surveillance plan.

A third area to be addressed is the production and dissemination of a national report card on the cancer burden. Specific activities will include the collection, analysis, and dissemination of data on important cancer outcomes and trends in risk factors, screening, and treatment to be incorporated into a national cancer report card; and the development of improved methods for disseminating information via the report card and other NCI communications.

The fourth area to be addressed is the support of molecular and genetics research for surveillance. Specific activities will include the development of valid tools to assess family history of cancer, which will provide for the collection of data on the population prevalence of familial cancers; and the investigation of the feasibility of expanding population-based molecular and genetic biomarker studies within the Cancer Surveillance Research Program.

The final area to receive attention is the development of a training strategy for individuals interested in cancer surveillance research. Specifically, training pertaining to the needs of surveillance sciences will be developed along with a plan to incorporate surveillance training as a priority in mechanisms for training cancer prevention and control scientists. Much more detail is provided in the Cancer Surveillance Research Implementation Plan [22] which can be obtained at `http://camp.nci.nih.gov/dccps/`

Thus, in addition to their basic goals of supplying timely information on trends in site-specific cancer incidence, prevalence, and survival, cancer surveillance programs are evolving to provide improved quantitative benchmarks to document the impact of research advances in cancer prevention, detection, and treatment, and to identify problems that can be addressed through cancer prevention and control efforts.

References

[1] American Joint Committee on Cancer (1997). *AJCC Cancer Staging Manual*, 5th Ed. Lippincott-Raven, Philadelphia.

[2] Brennan, M.F., Clive, R.E. & Winchester D.P. (1994). The COC: its roots and destiny, *American College of Surgeons Bulletin* **79**, 14–21.

[3] Chen, V.W., Wu, X.C. & Andrews, P.A., eds. (1999). *Cancer Incidence in North America, 1991–1995; Volume One: Incidence*. North American Association of Central Cancer Registries, Sacramento.

[4] College of American Pathologists (1965). *Systematized Nomenclature of Pathology*. Chicago.

[5] College of American Pathologists (1977). *Systematized Nomenclature of Medicine*. Two Vols., R.A. Cote, ed., Skokie.

[6] College of American Pathologists (1993). *The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International*, R.A. Cote, D.J. Rothwell, J.L. Palotay, R.S. Beckett & L. Brocher, eds. Northfield.

[7] Commission on Cancer of the American College of Surgeons (1988). *Data Acquisition Manual*. American College of Surgeons, Chicago.

[8] Costantino, J.P., Gail, M.H., Pee, D., Anderson, S., Redmond, C.K., Benichou, J. & Wieand, H.S. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence, *Journal of the National Cancer Institute* **91**, 1541–1548.

[9] Department of Health and Human Services (1985). *Quality Control for Cancer Registries*. Public Health Service, National Institutes of Health, Bethesda. Out of print.

[10] Edge, S., Fritz, A., Clutter, G.G. Page, D.L., Watkins, S., Blankenship, C., Douglas, L. & Fleming, I. (1999). A unified cancer stage data collection system: preliminary report from the Collaborative Stage Task Force/American Joint Committee on Cancer, *Journal of Registry Management* **26**, 57–61.

[11] Frey, C.M., McMillen, M., Cowan, C.D., Horm, J.W. & Kessler, L.G. (1992). Representativeness of the Surveillance, Epidemiology, and End Results Program data: Recent trends in cancer mortality rates, *Journal of the National Cancer Institute* **84**,: 872–877.

[12] Hankey, B.F., Ries, L.A., Kosary, C.L., Feuer, E.J., Merrill, R.M. & Edwards, B.E. (1999). Partitioning linear trends in age-adjusted rates, *Cancer Causes and Control*.

[13] Hankey, B.F., Ries, L.A.G. & Edwards, B.K. (1999). The SEER Program: a national resource, *Cancer Epidemiology, Biomarkers, and Prevention*.

[14] Kim, H.J., Fay, M.P., Feuer, E.J. & Midthune, D.N. (1999). Permutation tests for joinpoint regression with applications to cancer rates, *Statistics in Medicine*.

[15] Mandelson, M.T., Wagner, E.H. & Thompson, R.S. (1995). PSA screening: a public health dilemma, *Annual Reviews Public Health* **16**, 283–306.

[16] National Cancer Institute (1986). *Summary Staging Guide, NIH Publication No. 86-2313*. National Institutes of Health, Baltimore. Hard copy available at: `http://www-seer.ims.nci.nih.gov/cgi-bin/pubs/order1.pl`.

[17] National Cancer Institute (1992). *The SEER Program Code Manual*, 2nd Ed., *NIH Publication No. 94–1999*. National Institutes of Health, Baltimore. Hard copy available at: `http://www-seer.ims.nci.nih.gov/cgi-bin/pubs/order1.pl`.

[18] National Cancer Institute (1993). *Comparative Staging Guide for Cancer: Major Cancer Sites*, Version 1.1., *NIH Publication No. 93-3640.* National Institutes of Health, Baltimore. Hard copy available at: `http://www-seer.ims.nci.nih.gov/cgi-bin/pubs/order1.pl`.

[19] National Cancer Institute (1998). *Extent of Disease 1998, Codes and Coding Instructions*, 3rd Ed., *NIH Publication No. 98-1999.* National Institutes of Health, Baltimore.

[20] National Cancer Institute (1998). *The SEER Program Code Manual*, 3rd Ed., *NIH Publication No. 98-2313*. National Institute of Health, Baltimore.

[21] National Cancer Institute (1999). *SEER Cancer Statistics Review, 1973–1996*. National Institutes of Health, Bethesda.

[22] National Cancer Institute (1999). *Cancer Surveillance Research Implementation Plan*. Surveillance Implementation Group, National Institutes of Health, Bethesda.

[23] Sone, S., Takishima, S., Li, F., Yang, Z., Honda, T., Maruama, Y., Hasegawa, M., Yamanda, T., Kubo, K. & Asakura, K. (1998). Mass screening for lung cancer with mobile spiral computed tomography scanner, *Lancet* **351**, 1242–1245.

[24] Spackman, K.A., Campbell, K.E. & Cote, R.A. (1997). SNOMED RT: a reference terminology for health care. AMIA Fall Symposium.

[25] Steele, G.D. Jr., Winchester, D.P. & Menck, H.R. (1994). The National Cancer Data Base. A mechanism for assessment of patient care, *Cancer* **73**, 499–504.

[26] US Public Health Service (1991). *International Classification of Diseases, Clinical Modification*. 9th rev., 4th Ed.

[27] Wagner, G. (1985). In *The Role of The Registry in Cancer Control*, D.M. Parkin, G. Wagner & C.S. Muir, eds., *IARC Scientific Publications*, No. 66. International Agency for Research on Cancer, Lyon.

[28] World Health Organization (1976). *International Classification of Diseases for Oncology*, 1st Ed. WHO, Geneva.

[29] World Health Organization (1990). *International Classification of Diseases for Oncology*, 2nd Ed., C. Percy, V. Van Holten & C. Muir, eds. WHO, Geneva.

[30] World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems*, 10th rev., 3 Volumes. WHO, Geneva.

[31] Zippin, C., Lum, D. & Hankey, B.F. (1995). Completeness of hospital cancer case reporting from the SEER Program of the National Cancer Institute, *Cancer* **76**, 2343–2350.

(*See also* **Vital Statistics, Overview**)

B. HANKEY & APRIL FRITZ

Candidate Gene

A candidate gene is a **gene** that is guessed to underlie a disease on the basis of a metabolic pathway thought to be involved in the pathogenesis of the disease. Each protein in such a pathway must have one or more genes involved in its synthesis, and these are thus candidates for the gene(s) underlying the disease. Thus the "candidate gene approach" to finding a disease gene presupposes that we have prior information that the disease under study is probably caused by a particular gene, which we try to confirm by studying allelic variation at a genetic marker locus. If the marker locus is in fact the disease locus, then a **disease–marker association** can be sought to determine the particular allele(s) involved. Similarly, there may be a disease–marker association if the candidate marker locus is close to the disease locus and the two sets of alleles are in **linkage disequilibrium**. However, it is possible for the polymorphisms of the candidate locus studied to be in gametic equilibrium with the disease locus, even if the two loci are in almost the same position, and in this case **linkage analysis** is necessary to detect the (intrafamilial) association between the two loci. Finding an association or tight linkage between a disease and a candidate marker locus supports, but does not prove, the belief that the candidate gene does in fact cause the disease being studied.

ROBERT C. ELSTON

Canonical Correlation

Multivariate observations are characterized by various characteristics of their marginal distributions (such as their location, dispersion, and functional forms), as well as by their stochastic interrelations. These **associations** can be assessed only from their joint distributions, often in terms of suitable parameters associated with them. In the same way that the location or scale measures of marginal distributions may not necessarily correspond to some algebraic constants appearing in their functional forms, measures of association also may not correspond to a finite dimensional vector of algebraic constants appearing in the joint distribution. The situation with the **multivariate normal distribution** is, of course, entirely different because here all the measures of location, dispersion, **skewness**, **kurtosis**, and (total, partial, multiple or subset) association of all the coordinate variables can be formulated explicitly in terms of the *mean* vector and *dispersion* matrix (*see* **Covariance Matrix**), which are the only natural parameters appearing in the functional forms of the distributions. To appreciate this picture thoroughly, consider a (finite) mixture of several multivariate normal distributions with possibly different mean vectors and/or dispersion matrices, where the nonnegative mixing coefficients add up to one (yielding a convex combination). The mixed–normal distribution in this setup is characterized by means of the mixing coefficients as well as the component mean vectors and dispersion matrices. Therefore the measures of location, dispersion, as well as association are all functions of a finite set of parameters, though some of the basic linearity properties of the conditional or joint distributions may no longer be tenable. A more complex situation arises with general nonnormal distributions even if we confine ourselves to the so-called *elliptically symmetric* distributions (such as the multivariate t distribution). Thus, it seems quite natural to introduce such measures in the most simple and classical cases of multinormal distributions, interpret them properly, and then proceed to more complex situations and examine how such interpretations are affected by these underlying complexities. This picture is more complex in the case of association measures than in location or scale measures.

The (Karl) Pearsonian *total* or *product-moment* **correlation** *coefficient* in the bivariate case provides the genesis of canonical correlations in the multivariate case, as well as its generalizations for more complex models. If $(X_1, X_2)'$ is a bivariate random vector (rv) with a bivariate cumulative distribution function (cdf) $F(x, y)$, $(x, y) \in \Re^2$, such that F admits finite moments of order one and two, then we may write the correlation coefficient ρ as

$$\rho = \frac{\mathrm{cov}(X_1, X_2)}{[\mathrm{var}(X_1)\mathrm{var}(X_2)]^{1/2}},$$

and this definition does not entail any particular form of the cdf F. In this setup, X_1 and X_2 are uncorrelated when $\rho = 0$, and they are positively or negatively correlated/associated when ρ is positive or negative, respectively. Note, furthermore, that in a general setup, uncorrelation may not imply independence, although the converse is always true. In the case of a normal F, however, uncorrelation and independence are equivalent. Furthermore, by the use of the classical *Cauchy–Schwartz* moment inequality, it follows that $-1 \leq \rho \leq 1$, where the upper bound is attained only when X_1 and X_2 are strictly linearly related with a positive slope (except perhaps on a set of null probability measures), and a similar case with a negative slope leads to the attainment of the lower bound. The *regression function* of X_2 on X_1 is $m_2(x) = \mathrm{E}[X_2|X_1 = x]$, $x \in \Re$, and X_2 is said to have a **linear regression** on X_1 if $m_2(x)$ is a linear function of x for all x (except on a set of null measures). Similar notations and definitions hold for $m_1(x) = \mathrm{E}[X_1|X_2 = x]$. If both these regression functions are linear with respective slopes β_{12} and β_{21}, then it is easy to verify that

$$\rho^2 = \beta_{12} \times \beta_{21}.$$

It is also easy to verify that $\mathrm{E}(X_2 - \mathrm{E}X_2)^2 = \mathrm{E}[X_2 - m_2(X_1)]^2 + \mathrm{E}[m_2(X_1) - \mathrm{E}X_2]^2$, so that for the linear regression case, we have

$$\rho^2 = \frac{\mathrm{E}[m_2(X_1) - \mathrm{E}X_2]^2}{\mathrm{E}(X_2 - \mathrm{E}X_2)^2}.$$

Thus, ρ^2 is interpreted as the *component of variation* of X_2, which can be ascribed due to the regression on X_1. This component is null only when $\rho = 0$. If the underlying distribution is **bivariate normal**, then the conditional distribution of X_2, given X_1, is also normal with the mean linearly dependent on X_1 and (conditional) variance equal to $[\mathrm{var}(X_2)](1 - \rho^2)$, which is independent of X_1. A similar property holds

for the conditional distribution of X_1, given X_2. These are referred to as the *linearity of regression* and *homoscedasticity* (*see* **Scedasticity**) properties of bivariate normal distributions, and a similar characterization holds in the general multivariate normal case as well. In this homoscedastic case, we have

$$1 - \rho^2 = \frac{\mathrm{E}[X_2 - m_2(X_1)]^2}{\mathrm{E}(X_2 - \mathrm{E}X_2)^2},$$

although such a conditional variance component interpretation may not generally hold in the absence of the homoscedasticity condition.

Let us now proceed to the case of two subsets of variates, one containing a single element X_1, and the other one having $p(\geq 1)$ elements $\mathbf{X}_2 = (X_{21}, \ldots, X_{2p})'$. Suppose that we want a single measure for the association between X_1 and $\mathbf{X}_2$. To motivate this measure, suppose that $\mathbf{X}' = (X_1, \mathbf{X}'_2)$ has a multivariate normal distribution. Then, the conditional distribution of X_1, given $\mathbf{X}_2$, is also univariate normal with the mean linear in $\mathbf{X}_2$ and a constant (conditional) variance γ^2 (independent of $\mathbf{X}_2$). If we denote the variance of the marginal distribution of X_1 by σ_{11}, then we can write

$$\gamma^2 = \sigma_{11}(1 - R^2),$$

where R^2 is nonnegative and is bounded from above by 1. It is equal to zero only when X_1 and $\mathbf{X}_2$ are stochastically independent, while it attains the upper bound 1 when $\gamma = 0$, i.e. X_1 is perfectly linearly dependent on $\mathbf{X}_2$ (except possibly on a set of null measures). We can partition σ_{11} into two **orthogonal** components $R^2\sigma_{11}$ and γ^2, representing respectively the variation due to the regression of X_1 on $\mathbf{X}_2$, and the residual unexplained by this regression. This R^2 is a natural extension of ρ^2 to the pseudo-multivariate situation, and is known as the squared *multiple correlation coefficient* of X_1 on $\mathbf{X}_2$. There is another interpretation of this multiple correlation coefficient that is particularly appealing to the development of our study of canonical correlations. Consider an arbitrary linear combination of the second set, namely, $\mathbf{a}'\mathbf{X}_2$, where $\mathbf{a} \in \Re^p$. Let $\boldsymbol{\sigma}_1$ denote the covariance vector between X_1 and $\mathbf{X}_2$, and $\boldsymbol{\Sigma}_{22}$ be the dispersion matrix of $\mathbf{X}_2$; then the Pearsonian correlation coefficient of X_1 and $\mathbf{a}'\mathbf{X}_2$ is given by

$$\rho(\mathbf{a}) = \frac{\mathbf{a}'\boldsymbol{\sigma}_1}{[\sigma_{11}(\mathbf{a}'\boldsymbol{\Sigma}_{22}\mathbf{a})]^{1/2}}.$$

Consider the problem of choosing $\mathbf{a}$ in such a way that $\rho(\mathbf{a})$ is a maximum. We may normalize $\mathbf{a}$ by setting $\mathbf{a}'\boldsymbol{\Sigma}_{22}\mathbf{a} = 1$. Thus, we need to maximize $\mathbf{a}'\boldsymbol{\sigma}_1$ with respect to $\mathbf{a}$, subject to the above normalizing constraint. The desired solution is given by

$$\mathbf{a}_0 \propto \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_1,$$

and as a result we obtain

$$\rho^2(\mathbf{a}_0) = \frac{\boldsymbol{\sigma}_1'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_1}{\sigma_{11}},$$

which can be shown to be equal to the R^2 introduced earlier. In this setup, $\mathbf{a}_0'\mathbf{X}_2$ is the best fit to X_1, in the sense that the mean square of the residuals from this fit is a minimum among all possible choices of linear combinations of $\mathbf{X}_2$. The variance component due to regression for this fit yields the multiple correlation between X_1 and $\mathbf{X}_2$. Although in the above discussion we were primarily motivated by the linearity of regression and homoscedasticity characterizations of multinormal distributions, the latter interpretation of the multiple correlation does not crucially depend on the multinormality assumption. The canonical correlations, introduced by Hotelling [30, 31], relate to the case of two or more subsets of variates, each having possibly more than one variate, and hence contain the total and multiple correlation measures as particular cases.

Foundation of Canonical Correlations

Hotelling's original idea was to extend the Pearsonian product–moment correlation for studying stochastic dependence or association between two groups of variables. In multivariate analysis, often we may have two (or more) sets each containing multiple variates. Their dependence picture, even when completely determined by second-order moments, rests on a complex of high dimensional covariance matrices. In many studies, particularly exploratory, there may be too many variates within each subset, and this may smudge the overall picture to a greater extent. To reduce the dimensionality of the data set by eliminating *redundant* or less important variates, Hotelling [29], led by Pearson's [46] idea of fitting planes by *orthogonal least squares*, introduced two basic concepts in multivariate analysis for analyzing correlation structures: (i) **principal component analysis**

and (ii) canonical correlation analysis (CCA). Though the principal component analysis relates to an *internal analysis*, i.e. within-group orthogonal decomposition for the study of dispersion, and the canonical correlations to an *external analysis*, i.e. between-group interrelations or correlations, conceptually they are interrelated, and hence we motivate the canonical correlation analysis through the concepts of principal component analysis as well.

Suppose that $\mathbf{X}$ is a stochastic $p(\geq 1)$ vector having mean vector $\boldsymbol{\theta}$ and dispersion matrix $\boldsymbol{\Sigma}$. Note that in this setup there are p unknown elements in the mean vector and $p(p+1)/2$ unknown elements in the dispersion matrix, so that there are in all $p(p+3)/2$ unknown parameters in the model (even if multinormality is imposed on the distribution of $\mathbf{X}$). If the underlying distribution is not multinormal, then we may have an even larger parameter space. The emphasis in both the cases of principal component and canonical correlation analysis is on linear combinations of the variables that capture essentially the entire statistical information and at the same time preserve some interpretable properties. The justification for such linear transformations of course stems from multivariate normal distributions which characteristically possess some *equivariance* properties under such transformations. Nevertheless, the results to follow may still be interpreted without the underlying multinormality assumption. The principal component analysis aims to reduce the dimension of a given distribution by linear orthogonal transformations which may summarize the dispersion picture without losing any significant information. Let $\mathbf{b}$ be a p-vector of unit length in L_2 norm (i.e. $\mathbf{b}'\mathbf{b} = 1$), and consider the linear compound $\mathbf{b}'\mathbf{X}$ which has mean $\mathbf{b}'\boldsymbol{\mu}$ and variance $\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}$. We choose the particular vector $\mathbf{b}_0$ which leads to the largest possible variance. For this, we maximize $\psi(\mathbf{b}) = \mathbf{b}'\boldsymbol{\Sigma}\mathbf{b} - \lambda(\mathbf{b}'\mathbf{b} - 1)$ with respect to $\mathbf{b}$, where λ is a Lagrangian multiplier. This leads to the following estimating equation:

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\mathbf{b} = \mathbf{0}, \tag{1}$$

so that λ satisfies the equation $|\boldsymbol{\Sigma} - \lambda\mathbf{I}| = 0$ (*see* **Eigenvalue**). If we denote the solutions (in λ) of (1) by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p (\geq 0)$ and identify them as the ordered *characteristic roots* of $\boldsymbol{\Sigma}$, then we may simultaneously introduce the *characteristic vectors* $\mathbf{b}_j$, $j = 1, \ldots, p$, by setting

$$\boldsymbol{\Sigma}\mathbf{b}_j = \lambda_j\mathbf{b}_j, \quad j = 1, \ldots, p$$

(*see* **Eigenvector**). Then $\mathbf{b}'_1\mathbf{X}$ is termed the *first principal component* and its variance is equal to λ_1. In general, for $j = 1, \ldots, p$, $\mathbf{b}'_j\mathbf{X}$ is the jth principal component with variance λ_j, which is the highest possible subject to the constraint(s) that the jth component is uncorrelated with all the previous $j-1$ components. If $\boldsymbol{\Sigma}$ is singular, of rank $q < p$, then $\lambda_j = 0$, for all $j > q$, so that q orthogonal components describe the structure completely. In many anthropometric, biometric and psychometric problems, there may be a large number of variables either having multi-**collinearity** or almost degeneracy. In such a case the prominent principal components convey a clear picture of the relevant compounds, obtainable by orthogonal transformations, that capture the essential statistical information. Of particular statistical importance are the so-called *biplots* which relate only to the first two principal components and sacrifice the others (*see* **Graphical Displays**). In that way a high-dimensional model is reduced virtually to a two-dimensional one, so that a statistical analysis can be performed with greater clarity.

Let us look into the canonical correlation problem in the same vein. We partition the p-vector $\mathbf{X}$ into two subvectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of dimension p_1 and $p_2 (p = p_1 + p_2)$ reflecting two sets of variables. Adopt a similar partition for $\boldsymbol{\mu}$, and the dispersion matrix $\boldsymbol{\Sigma}$ is then partitioned into p_1 and p_2 rows and columns, i.e. $\boldsymbol{\Sigma} = ((\boldsymbol{\Sigma}_{ij}))_{\{i,j=1,2\}}$. Although it is not necessary, for simplicity, we assume that both $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are nonsingular, while the rank of $\boldsymbol{\Sigma}_{12} = m[\leq \min(p_1, p_2)]$. If $m = 0$ (i.e. $\boldsymbol{\Sigma}_{12}$ is a null matrix), then the canonical correlations to be derived will be all equal to 0, and hence we assume that $m \geq 1$. Whenever m is small compared with p_1 and p_2, we shall see that the canonical correlations lead to substantial data reduction with respect to such between-group dependence studies. If p_1 and p_2 are both equal to 1, then their product moment correlation explains their interrelations. If either p_1 or p_2 is equal to 1, while the other is larger, then we can appeal to the multiple correlation measure to study their interdependence. If both p_1 and p_2 are greater than 1, then it may be more practical to consider only a few linear combinations from each set, chosen in such a way that they contain as much dependence structure as possible. This would then amount to a reduction in the dataset (dimension) for the study of between-group dependence structures with only insignificant loss of information. With this motivation, within each

subset we consider linear compounds chosen in such a way that they have unit variance, and then we desire to maximize their correlation. Thus, let $\mathbf{a}$ and $\mathbf{b}$ be two (p_1 and p_2) vectors such that

$$\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a} = 1 \quad \text{and} \quad \mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b} = 1.$$

Subject to these constraints, $\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}$ is a maximum. Using Lagrangian multipliers λ and ν, we consider the function

$$\psi(\mathbf{a}, \mathbf{b}) = \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} - \left(\frac{\lambda}{2}\right)(\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a} - 1) - \left(\frac{\nu}{2}\right)(\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b} - 1).$$

Taking the gradients with respect to $\mathbf{a}$ and $\mathbf{b}$, and equating them to zero, we arrive at the solution that $\lambda = \nu$ with

$$-\lambda\boldsymbol{\Sigma}_{11}\mathbf{a} + \boldsymbol{\Sigma}_{12}\mathbf{b} = 0,$$
$$\boldsymbol{\Sigma}_{21}\mathbf{a} - \lambda\boldsymbol{\Sigma}_{22}\mathbf{b} = 0,$$

so that the solution λ_1^2 is the largest characteristic root of either of the matrices

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \qquad \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{2}$$

or their other cyclic permutations. If we choose the characteristic vectors $\mathbf{a}_1$ and $\mathbf{b}_1$ corresponding to the largest characteristic root λ_1, then we obtain that the Pearsonian product–moment correlation of $U_1 = \mathbf{a}'_1\mathbf{X}^{(1)}$ and $V_1 = \mathbf{b}'_1\mathbf{X}^{(2)}$ is equal to λ_1, which is the largest possible correlation between a linear compound of $\mathbf{X}^{(1)}$ and another one of $\mathbf{X}^{(2)}$. That is why it is termed the *first canonical correlation* between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. We can consider then a second linear combination of $\mathbf{X}^{(1)}$ and a second one of $\mathbf{X}^{(2)}$, say U_2 and V_2, such that among all linear combinations uncorrelated with U_1 and V_1, λ_2, the correlation between U_2 and V_2, is the maximum. A little algebra shows that λ_2^2 is the second largest characteristic root of the same matrix in (2) for which λ_1^2 is the largest root, and the corresponding characteristic vectors are the coefficient vectors for U_2 and V_2. This process can continue until we have m of the characteristic roots $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_m^2 > 0$, and the corresponding sets of canonical variates U_j and V_j, $j = 1, \ldots, m$. Note that as the rank of $\boldsymbol{\Sigma}_{12} = m$, $\lambda_j^2 = 0$ for every $j > m$. Also note that like the product–moment, partial and multiple correlations, the canonical correlations are all scale-invariant with respect to each of the p coordinate variates.

At this stage we set, without loss of generality, $m \leq p_1 \leq p_2$, and denote by $\boldsymbol{\Lambda}$ the diagonal matrix of order $p_1 \times p_1$ with the leading elements $\lambda_1, \ldots, \lambda_m, 0, \ldots, 0$. Then the process of finding the canonical correlations leads to two sets of canonical combinations $\mathbf{U} = (U_1, \ldots, U_{p_1})'$ and $\mathbf{V} = (V_1, \ldots, V_{p_2})'$, such that the dispersion matrix of the p-vector with the two components $\mathbf{U}$ and $\mathbf{V}$ is given by

$$\begin{pmatrix} \mathbf{I} & \boldsymbol{\Lambda} & \mathbf{0} \\ \boldsymbol{\Lambda} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}, \tag{3}$$

where the last pivotal $\mathbf{I}$ is of order $(p_2 - p_1) \times (p_2 - p_1)$, and hence is a null matrix when $p_2 = p_1$, and $\boldsymbol{\Lambda}$ has $(p_1 - m)$ null diagonal elements when $p_1 \geq m$. This provides a clearly interpretable picture of CCA.

This latter representation of the canonical variates and their (canonical) correlations can also be interpreted in the light of *affine equivalence* studied in detail in [21, Chapter 10]. Recall that an affine transformation on a vector rv $\mathbf{X}$ is defined by $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{a}$, where $\mathbf{A}$ is nonsingular and $\mathbf{a}$ is some vector. The **Hotelling T^2** and other likelihood-based **multivariate analysis of variance** tests are affine invariant in the sense that affine transformations on the observable random vectors leave the statistic invariant. Eaton interpreted that if $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{a}$ with probability one, where $\mathbf{A}$ is nonsingular, then $\mathbf{Y}$ and $\mathbf{X}$ are affinely equivalent. Consider now measures of *affine dependence* between the two subvectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ which are functions of the covariance matrix $\boldsymbol{\Sigma}$ and are invariant with respect to affine transformations on $\mathbf{X}$. For example, if $\mathbf{Y}^{(j)}$ is affinely equivalent to $\mathbf{X}^{(j)}$, for $j = 1, 2$, then the measure of affine dependence between $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ should be the same as between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Since the elements of the covariance matrix $\boldsymbol{\Sigma}$ are translation invariant, one may as well take the shift vector $\mathbf{a} = \mathbf{0}$. Then, following Eaton [21], we may verify that the characteristic roots of (2) constitute the maximal invariant function under the compound group of affine transformations on the individual subset vectors. Since the canonical correlations are the squares of these characteristic roots, it follows that any invariant measure of affine dependence has to be a function of these canonical correlations. The canonical vectors $\mathbf{U}$ and $\mathbf{V}$ defined before (3) are affinely equivalent to $\mathbf{X}^{(1)}$

and $\mathbf{X}^{(2)}$, respectively, and (3) reveals their affine dependence through the diagonal matrix $\mathbf{\Lambda}$. Following Dempster [16] and Eaton [21], we note that the canonical correlations also have a natural geometrical interpretation as **cosines of the angles** between appropriate vector spaces. With biostatistical perspective in mind, we refrain from such abstractions. However, we refer to the two texts by Morrison [42] and Mardia et al. [39] for nice applications-oriented illustrations of canonical correlations and related topics. Below we consider some extension of canonical correlations to the case of more than two subsets of variates, as well as to some restricted type of dependence patterns.

Relationships with Other Multivariate Measures

One of the reasons for the broad appeal of canonical correlation analysis in practical applications is its ability to subsume many other multivariate measures, concepts, and methods. We have already commented on the relevance of the principal component analysis. Besides this, CCA brings together techniques like *multiple correlation* and *regression analysis, canonical discriminant analysis*, **correspondence analysis**, *analysis of* **contingency tables**, and **multivariate analysis of variance** and *covariance*. Some of these items are covered in greater detail in other articles, and hence we omit their definitions here. **Linear discriminant analysis** has also been covered as a special case under canonical correlation analysis by Takeuchi et al. [59]. In this particular case the canonical variables of one of the sets turn out to be the discriminators, while those of the other set provide the optimal scores to be assigned to the different populations. **Factor analysis**, another important area in multivariate analysis with special emphasis on **psychometry** and *mental testing*, also has affinity to canonical correlation analysis. To illustrate this relationship we consider the following problem, which uses canonical correlation analysis in a spirit somewhat similar to that in factor analysis. Suppose that $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are linearly dependent on a number m, of common, uncorrelated, unobservable factors $\mathbf{F} = (F_1, \ldots, F_m)'$ in the following way:

$$\mathbf{X}^{(j)} = \mathbf{A}_j \mathbf{F} + \mathbf{G}_j, \quad j = 1, 2,$$

where $\mathbf{G}_1$ and $\mathbf{G}_2$ are the specific factors, and we assume that $\mathbf{F}$, $\mathbf{G}_1$, and $\mathbf{G}_2$ are all uncorrelated (orthogonal). The goal is to determine the minimum m, the *effective number of common factors* for which such a representation is possible. It is known [49] that m is equal to the rank of $\mathbf{\Sigma}_{12}$, and hence canonical correlation analysis can be adapted to appraise this situation. As we consider the sample counterparts of the canonical correlations, we will observe that there are certain difficulties in attaching any reliability to the rank of the sample counterpart of $\mathbf{\Sigma}_{12}$, and the canonical correlation analysis has some advantages in this respect.

Let us examine the role of CCA in a related **prediction** problem, which is essentially allied to the **multivariate multiple regression** problem. We have the same partitioning, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, as before, and in the same way we find the partitioning $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ and the $\mathbf{\Sigma}_{ij}$, $i, j = 1, 2$, of the mean vector and covariance matrix, respectively. In a multivariate normal model the regression of $\mathbf{X}^{(1)}$ on $\mathbf{X}^{(2)}$ is given by

$$\mathrm{E}[\mathbf{X}^{(1)}|\mathbf{X}^{(2)}] = \boldsymbol{\mu}^{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}[\mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)}],$$

and the dispersion matrix of the residual vector $\mathbf{X}^{(1)} - \mathrm{E}[\mathbf{X}^{(1)}|\mathbf{X}^{(2)}]$ is given by

$$\mathbf{\Sigma}_{11.2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}.$$

It is easy to verify that the characteristic roots of $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{11.2}$ are nothing but the complements of the squared canonical correlations, namely $1 - \lambda_j^2$, $j = 1, \ldots, p_1$. Without this multinormality condition, one may still fit a linear regression of $\mathbf{X}^{(1)}$ on a linear compound $\mathbf{a} + \mathbf{B}\mathbf{X}^{(2)}$, and verify that the best fit, in the sense of having the minimum characteristic roots of the standardized dispersion matrix of the residual vector, corresponds to $\mathbf{a} = \boldsymbol{\mu}^{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}$ and $\mathbf{B} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$. Therefore, either way the canonical correlations tell us about the dispersion of the predicting vector, as well as the errors due to deviation from the predictor. Thus, we may use the canonical variates for the set $\mathbf{X}^{(2)}$ and come up with predictions under dimension reduction. We refer to Brillinger [8] for some allied studies. Note that in this setup the roles of the two sets are not the same, although in CCA we have a symmetric formulation.

The above development also leads us to two important concepts in multivariate analysis, namely the *canonical loadings* and *redundancy coefficient*. Canonical loadings can be introduced through the

canonical weights, which in optimum canonical correlations express the importance of a variable from one set with regard to the other set in obtaining a maximum correlation between the two sets. In a sense these canonical weights are related to the canonical variables introduced earlier [see (2) and the discussion following it]; we denoted these canonical variates for the two sets by $\mathbf{U}$ and $\mathbf{V}$, respectively. We write $\mathbf{U} = \mathbf{A}_1\mathbf{X}^{(1)}$ and $\mathbf{V} = \mathbf{A}_2\mathbf{X}^{(2)}$. To interpret CCA appropriately, one needs to look into these canonical coefficients along with the canonical loadings as defined below. A canonical loading, or structure, is the product–moment correlation between an original variable and its respective canonical variable, so that it reflects the degree to which a variable is represented by its canonical variable. Thus, if we denote the ith intragroup correlation matrix by $\mathbf{R}_{ii}$, $i = 1, 2$, then for the p_i canonical variables in the ith group, the canonical loadings are given by the rows of $\mathbf{R}_{ii}\mathbf{A}_i$, $i = 1, 2$. At this stage we treat the first set of variables as the *predicting* set, and the second one as the *criterion* set. While the proportion of variance in the criterion set explained by its canonical variates can be expressed in terms of the canonical loadings, we may like to study the proportion of variance in the criterion set which is explained by the predictor set. Here canonical correlations and canonical variables do not capture the full information, and we need redundancy analysis. The redundancy coefficient, proposed by Stewart & Love [57], represents the amount of variance in the criterion set that is redundant to the amount of variance in the predictor set. They showed that the redundancy coefficient is equivalent to regressing each variable in the criterion set in turn on all the variables in the predictor set, and then averaging the p_2 squared multiple correlation coefficients.

Canonical weights or patterns are used to assess the relationship between the original variables and the canonical variables so that they indicate the contribution of each variable to the respective within-set canonical variables. However, these weights do not necessarily accurately reflect the relative importance of the different variables. This may happen mainly due to the presence of *multicollinearity*, where some variable may obtain a small, even negative, weight because of the presence of some other variables yielding the degeneracy of the dispersion matrix. Also, because of multicollinearity, these coefficients may become very unstable. That is why some researchers have advocated the use of canonical loadings (structure) instead of canonical weights (pattern). But these measures can also suffer from similar drawbacks. The main difficulties surrounding CCA in such nonregular cases have not been well assessed, and hence one should exercise caution in interpreting canonical correlations in any particular setting. In passing, we may also remark that the canonical variables, for either set, are not observable, and hence may not always attach interpretable physical meanings. This drawback of CCA is of some concern in applied fields, particularly in biostatistics, because derived statistical results should be capable of being understood and interpreted by researchers from other fields who seek to make use of them. It is to be noted though that the canonical variables are of interest in their own right; they help to deepen the understanding of the original variables and in some cases may even suggest new measures. Kshirsagar [37] remarked that if canonical analysis had no other practical use, it could at least be used as a descriptive and exploratory tool. It summarizes the complex relationship and provides a useful method of reduction in the dimensionality problem.

Sample Canonical Correlation(s) and their Sampling Distributions

Suppose that $\mathbf{X}'_i = (\mathbf{X}_i^{(1)'}, \mathbf{X}_i^{(2)'})$, $i = 1, \ldots, n$ are n independent identically distributed (iid) random variables having the same partition of the mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$ as before. We define the sample mean vector $\overline{\mathbf{X}}_n$ and sample covariance matrix $\mathbf{S}_n$ as

$$\overline{\mathbf{X}}_n = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i,$$

$$\mathbf{S}_n = (n-1)^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)'.$$

Then $\mathbf{S}_n$ is translation-invariant and affine equivariant, and is unbiased for $\boldsymbol{\Sigma}$ whenever the latter exists. If we sample from a multivariate normal population, then the $\mathbf{X}_i$s have finite moments of all orders, and $\overline{\mathbf{X}}_n$ and $\mathbf{S}_n$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Hence the estimators of canonical correlations (which are functions of these natural parameters) can be based solely on such sufficient statistics. Such a motivation may not be rational if the underlying pdf is not

multinormal; these statistics then may not lead to the maximum likelihood estimators of the canonical correlations, although from the point of view of affine equivariance and unbiasedness, $\overline{\mathbf{X}}_n$ and $\mathbf{S}_n$ may still be considered appropriate for such data reduction.

With this motivation, and recalling the affine equivariance property of the sample covariance matrix, we partition $\mathbf{S}_n$ into $\mathbf{S}_{ij,n}$, $i, j = 1, 2$, in the same way as was done for $\mathbf{\Sigma}$, and consider the derived matrix

$$\mathbf{S}_n^* = \mathbf{S}_{12,n}\mathbf{S}_{22,n}^{-1}\mathbf{S}_{21,n}\mathbf{S}_{11,n}^{-1};$$

we also could have taken the permutation version of order $p_2 \times p_2$. Here also we take $p_1 \leq p_2$, and hence prefer to use the above definition of $\mathbf{S}_n^*$. It is easy to see that $\mathbf{S}_n^*$ satisfies the affinely equivalence property, and is translation invariant. Consider the characteristic roots $l_{1,n}^2 \geq l_{2,n}^2 \geq \ldots \geq l_{p1,n}^2 (\geq 0)$ of $\mathbf{S}_n^*$, and define the sample canonical correlations as their nonnegative roots:

$$l_{j,n} = j\text{th sample canonical correlation},$$
$$\text{for } j = 1, \ldots, p_1.$$

Note that the sample canonical correlations are all (ordered) stochastic (nonnegative) variables, and even if the population canonical correlations are zero for some $j (\geq 1)$ and onwards, the corresponding sample canonical correlations may not be identically equal to zero with probability one. Therefore for drawing statistical conclusions on the canonical correlations based on their sample counterparts, it is necessary to incorporate the **sampling distribution** of the latter in the statistical decision procedures.

For finite sample size n, the marginal (and certainly joint) distributions of the sample canonical correlations are, in general, extremely complicated, even if we confine ourselves to sampling from a multivariate normal distribution. However (in the case of a multinormal distribution), when the population canonical correlations are all null (so that the two subsets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent, implying $\mathbf{\Sigma}_{12} = \mathbf{0}$), the sampling distribution is much more manageable; an elegant derivation of this can be found in Kshirsagar [37]. This special case relates to the independence of the two subsets, and the classical **likelihood ratio test** statistic, $\mathfrak{L}_n$, can be easily shown to be given by

$$\mathfrak{L}_n = n[\ln|\mathbf{S}_{n,11}| + \ln|\mathbf{S}_{n,22}| - \ln|\mathbf{S}_n|]$$
$$= n \ln|\mathbf{I} - \mathbf{S}_n^*| = n\sum_{j=1}^{p_1} \ln(1 - l_{j,n}^2).$$

The null hypothesis distribution of this test statistic can be quite well approximated by the central **chi-square distribution** with $p_1 p_2$ degrees of freedom (df), and this approximation can be improved further, particularly for moderate values of n, by using the conventional Bartlett correction. For details we refer the reader to any standard multivariate statistical analysis text; Anderson [2] is an excellent source. A similar likelihood ratio test for the null hypothesis that all but the first $k (\leq p_1)$ characteristic roots are equal to 0, can be worked out in the same manner. The latter test of the hypothesis problem is of significant interest in deciding on the number of canonical correlations to be used in a given situation, or to estimate the rank of $\mathbf{\Sigma}_{12}$. The asymptotic distribution is again chi-square, this time with df $(p_1 - k)(p_2 - k)$; appropriate Bartlett correction works out along the same line. Constantine [10] derived the density of the sample canonical correlations in the general (multivariate normal) case when the population of canonical correlations are not necessarily null, and involves a hypergeometric function of two matrix arguments (and thereby is very difficult to implement in actual applications). From statistical considerations, therefore, there was a pressing need for suitable approximations to sampling distributions of sample canonical correlations, and, led by the pioneering work of T.W. Anderson in the late 1940s and early 1950s, a considerable amount of research work has been accomplished in this area. The first phase relates to the large-sample case and yet retaining the underlying multinormality assumption, and demonstrates the asymptotic normality of the $\sqrt{n}(l_{j,n} - \lambda_j)$, for different j, treating separately the cases of distinct population canonical correlations and their multiplicity. The parameters appearing in these asymptotic normal distributions themselves may depend in a rather complex manner on the unknown λ_j. Therefore in setting suitable confidence sets or testing composite null hypotheses on the canonical correlations, it may be necessary to estimate these functional parameters reliably from the sample data, and this may make it necessary to have an enormously large sample size – a postulation that is not always met in practice.

The second phase of development started with the observation that CCA is a meaningful tool for the study of between-group dependence, even when the underlying multinormality assumption may not be tenable. However, without the multinormality assumption the canonical correlations relate to measures of correlation vs. noncorrelation, but not necessarily independence. In such a general nonnormal case the asymptotic distribution of the sample canonical correlations has been studied by Muirhead & Waternaux [44] under the assumption that the population pdf admits finite fourth-order moments and that its canonical correlations are all distinct, i.e. the sample canonical correlations have a jointly asymptotically multinormal distribution. While the finiteness of the fourth-moment condition is not that stringent (needed even for the total correlation in a bivariate nonnormal distribution), without the distinctness of the population canonical correlations, the asymptotic distribution may not be normal without the latter condition. When the asymptotic normality holds, Muirhead & Waternaux were able to verify that the asymptotic variance of $\sqrt{n}(l_{j,n} - \lambda_j)$ is equal to

$$\begin{aligned}\gamma_j^2 = \{&4\lambda_j^2(1-\lambda_j^2)^2 + \lambda_j^4(\kappa_{j:4} + \kappa_{j+p_1:4}) \\ &+ 2\lambda_j^2(2+\lambda_j^2)\kappa_{j,p_1+j:2,2} \\ &- 4\lambda_j^3(\kappa_{j,p_1+j:3,1} + \kappa_{j,p_1+j:1,3})\},\end{aligned}$$

where the κ_j are different fourth-order cumulants of the population (*see* **Characteristic Function**). The formulas for the asymptotic covariances are even more cumbersome (see [44]). In practice, for setting **confidence intervals** or **hypothesis testing**, even for the canonical correlations in isolation, one would require a good estimator of the γ_j^2, and therefore the presence of the unknown cumulants signals further complications in the practical use of this asymptotic approximation. One possible way to eliminate this shortcoming is to estimate the unknown cumulants from the sample by the method of moments, and substitute these estimators in the above expression. Alternately, *resampling methods*, such as the **jackknife** or the **bootstrap**, can be used to estimate the γ_j nonparametrically with considerable computational ease. Das & Sen [14] have studied such resampling plans. Since characteristic roots typically relate to certain (implicit) equations involving polynomial functions, it is easy to verify that the canonical correlations, the λ_j, are all *smooth* functions of the population covariance matrix $\mathbf{\Sigma}$, and likewise, the sample canonical correlations are smooth functions of the sample covariance matrix $\mathbf{S}_n$. Furthermore, the elements of $\mathbf{S}_n$ are all $\boldsymbol{U}$**-statistics**, and hence are asymptotically jointly multinormal whenever the fourth moments are all finite. This provides an easy way to verify the asymptotic normality of the sample canonical correlations as well as to incorporate resampling plans to estimate their asymptotic dispersion matrix in a nonparametric manner.

It may also be remarked that the canonical correlations are invariant under multiplication of the original variables with nonsingular matrices. Hence, in most sampling theory works it is assumed without loss of generality that the population covariance matrix $\mathbf{\Sigma}$ is of the canonical form given in (3).

Generalizations and Excursions

Generalizations of the notion of canonical correlations to three or more subsets of variates were proposed by Roy [51]; he also developed the notion of partial canonical correlations between two subsets of variates when the others are held fixed, and proposed a test for the same. Anderson [2] incorporated the minimization of the Wilks *generalized variance* or determinant criterion to define the canonical correlations in the case of two or three subsets of variates. However, a more systematic approach to multigroup canonical correlation analysis was initiated by Horst [28] with further feedback from Kettenring [34, 35] and Sengupta [54]. Most of these methods call for the selection of canonical variables, one from each subset, such that some function of their correlation matrix is maximized. The different methods typically considered are

1. Maximization of the sum of the correlation coefficient (SUMCOR).
2. Maximization of the sum of squares of the elements of the correlation matrix (SSQCOR).
3. Maximization of the largest characteristic root of the correlation matrix (MAXVAR).
4. Minimization of the smallest characteristic root of the correlation matrix (MINVAR).
5. Minimization of the generalized variance of the correlation matrix (GENVAR).

Several computational **algorithms** that work satisfactorily in practice are available for these methods (see [54] and the references therein), although there is some need to justify their convergence properties theoretically.

The developments on CCA reported above all relate to the case where the dispersion matrix $\mathbf{\Sigma}$ is positive definite, and singular covariance matrices resulting from multicollinearity and/or redundancy are thereby excluded from this setup. Intuitively, at least, this study should not depend on the positive definiteness of the subset dispersion matrix, and incorporation of suitable *generalized inverses* (*see* **Matrix Algebra**) indeed validates the canonical correlation analysis in the case of a possibly singular dispersion matrix, without essentially any additional regularity assumptions over the conventional positive definite case. To illustrate this point, consider a $2p$-vector $\mathbf{X}$ partitioned into two p-vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, where the dispersion matrix of each subset is singular, say of rank $q(< p)$. Then we first write $\mathbf{X}^{(j)} = \mathbf{B}_j \mathbf{Y}^{(j)}$, $j = 1, 2$, where the $\mathbf{Y}_j$ are q-vectors having nonsingular dispersion matrices $\mathbf{\Gamma}_j$, $j = 1, 2$, while the $\mathbf{B}_j$ are $p \times q$ matrices of rank q. This way we have

$$\mathbf{\Sigma}_{jk} = \mathbf{B}'_j \mathbf{\Gamma}_{jk} \mathbf{B}_k, \quad j, k = 1, 2,$$

and the canonical correlation analysis can be performed on the $\mathbf{Y}^{(j)}$ which satisfy the nonsingularity condition on their dispersion matrix. Note that in this way the number of nonzero canonical correlations will be less than or equal to q, although $p_1 = p_2 = p > q$. Khatri [36], Rao [50], Jewell & Bloomfield [33], Sengupta [55], and Baksalary et al. [5] have all incorporated different generalized inverses for such singular cases in various frameworks. In a nutshell, they have shown that no changes are needed from the traditional approach. However, in the canonical reduction of the dispersion matrix in (3) the three block-diagonal matrices $\mathbf{I}$ are replaced by $\mathbf{I}$ of lower order and a null complementary part.

In the same way as the product–moment correlation has been extended to *part*, *partial* and *bipartial* correlations, canonical correlations also have been extended to such part, partial, and bipartial forms. For example, with (the usual extension of notations and) three groups of variates $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, and $\mathbf{X}^{(3)}$, the partial canonical correlations between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ after the linear effect of $\mathbf{X}^{(3)}$ is removed can be obtained from the conditional covariance matrix between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Given $\mathbf{X}^{(3)}$, in the same way the usual canonical correlations are obtained from the unconditional covariance matrix. Thus, in essence, $\mathbf{\Sigma}_{ij}$ gets replaced by $\mathbf{\Sigma}_{ij.3} = \mathbf{\Sigma}_{ij} - \mathbf{\Sigma}_{i3} \mathbf{\Sigma}_{33}^{-1} \mathbf{\Sigma}_{3j}$ for $i, j = 1, 2$. Similarly, an example of part canonical correlations would be when the effect of $\mathbf{X}^{(3)}$ is removed from $\mathbf{X}^{(2)}$, but not from $\mathbf{X}^{(1)}$. This would require working with

$$\begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12.3} \\ \mathbf{\Sigma}_{21.3} & \mathbf{\Sigma}_{22.3} \end{pmatrix}.$$

Bipartial canonical correlations are relevant when there are four groups of variates and the effects of $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ are removed respectively from $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ (but not from both). See Rao [48] and Timm & Carlson [61] for more details on these topics.

In many problems, some structural symmetry is reflected in the correlation matrix $\mathbf{R}$, and this can simplify the computational algorithm for the canonical correlations. For example, if $\mathbf{R}$, a $2p \times 2p$ matrix, is partitioned into four matrices $\mathbf{R}_{ij}$, $i, j = 1, 2$, where

$$\mathbf{R}_{ii} = (1 - \rho_1)\mathbf{I} + \rho_i \mathbf{J}, \quad i = 1, 2;$$
$$\mathbf{R}_{12} = \rho_3 \mathbf{J}; \qquad \mathbf{J} = \mathbf{1}\mathbf{1}',$$

and the ρ_i, $i = 1, 2, 3$, are real numbers assuming values on $(-1, 1)$, then $\mathbf{R}_{12}$ is of rank 1, and hence only the first canonical correlation is nonzero. This computation can be carried out much more conveniently than in the general case of an arbitrary correlation matrix. Das & Sen [14] give additional illustrations of this type.

It is important to study how significant the changes in canonical correlations are when the entries of the covariance matrix shift by a small amount. An account of such perturbation analysis in this context can be found in Golub & Zha [25]. Styan [58] and Bérubé et al. [6] have studied the CCA in three-way layouts exploring orthogonality and connectedness.

Among other excursions in the field of canonical correlation analysis, *constrained* and *restricted* variations deserve special mention. As discussed earlier, one of the most common reservations held against the traditional analysis is the lack of interpretability of the canonical coefficients. In the context of categorical data models, DeSarbo et al. [17] introduced the idea of constrained canonical correlations to address this problem. They proposed allowing the coefficients for each cell (in the linear compound/canonical variable) to be either of the three entries $\{-1, 0, 1\}$. Thus,

for p_1 and p_2 categorical variables in the two subsets, there are in all $(3^{p_1}-1)(3^{p_2}-1)$ choices of the two (nonnull) coefficient vectors, and optimization is to be achieved within this finite set. They also discussed relevant computational aspects, comparing the complete enumeration with two algorithms that reduce the computation further. This simple alternative may have been more appealing because of the discrete nature of the data set they studied. Yanai & Takane [66] discussed the algebra of restricted canonical correlations with additional linear constraints (on the coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ for the two subsets) in the form

$$\mathbf{Aa} = \mathbf{0} \quad \text{and} \quad \mathbf{Bb} = \mathbf{0},$$

where $\mathbf{A}$ and $\mathbf{B}$ are $r_1 \times p_1$ and $r_2 \times p_2$ given matrices with $r_1 \leq p_1$ and $r_2 \leq p_2$. In view of the linear nature of these subspaces, they considered the projection technique to solve the optimization problem in this setup. A more general restricted canonical correlation analysis arises in many psychometric and biological problems, wherein the coefficients appearing in the linear compounds are restricted by certain inequalities, such as that these are all nonnegative or ordered within each subset (*see* **Isotonic Inference**). The most common restriction of this type relates to the case where all the coefficients $\mathbf{a}$ and $\mathbf{b}$ are set to be nonnegative, so that the derived canonical variables are *convex combinations* of the original variables. In this context, suitable interpretations, perhaps being better *representatives* of their respective group of variates, can be made of these canonical variables. Das & Sen [13] formulated the algebraic derivations of such restricted canonical correlations. It was shown how some general restricted models (including cases where only *some* of the coefficients are constrained by inequality-type restrictions) can be reduced to such convex combination models by simple transformations. It follows from their discussion that the largest squared restricted canonical correlation between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ is equal to one of the squared canonical correlations between ${}_{\mathbf{a}}\mathbf{X}^{(1)}$ and ${}_{\mathbf{b}}\mathbf{X}^{(2)}$, for some $\mathbf{a} \in \mathbf{W}_{p_1}$ and $\mathbf{b} \in \mathbf{W}_{p_2}$, with $\mathbf{W}_p = \{\mathbf{a} : \emptyset \neq \mathbf{a} \subseteq \{1, \ldots, p\}\}$ and ${}_{\mathbf{a}}\mathbf{X}$ stands for the $|\mathbf{a}|$-component vector consisting of those components of $\mathbf{X}$ whose indices belong to $\mathbf{a}$. This result also holds for the case of sample canonical correlations in the restricted case, and is of pivotal importance in the study of the sampling properties of such statistics. This was handled in a follow-up study, Das & Sen [15], along with the effectiveness of resampling plans in such restricted canonical correlation schemes. In a recent book, Hastie et al. [23] dealt with nonlinear canonical variables in neural network and flexible discriminant analysis. Friman et al. [27] used the above constrained analysis technique in adaptively filtering the functional MRI data.

Issues Related to Applications in Biostatistics

At present, biostatistics covers a broad domain of basic as well as applied sciences, wherein biological, medical and clinical, pharmacologic, environmental, epidemiologic, neural, and socioeconomic aspects have all mingled with statistical concepts and perspectives in a harmonious way. The advent of modern computers has opened up a wide avenue of **computer-intensive**, application-oriented research in this fertile field (see, for example, Sen [53]). Typically, in view of the high dimensionality and high volume of acquired data sets in such studies, it is often necessary to implement effective dimension reduction techniques, such as **projection pursuits**, etc. In many such studies it may be possible to identify multiple subsets of variates on suitable experimental grounds, and interrelations of (two or more) subsets of variates constitute the prime objective of the study. For example, in a therapeutic study, certain body characteristic variables (such as blood sugar, cholesterol and other lipids) are to be recorded before and immediately after a therapeutic course is complete, and sometimes to judge its long-term efficacy, also after a certain lapse of time. Provided the measurements at each stage conform to a set of biologically or medically contiguous variables, the concept of CCA, or its various ramifications considered above, can be incorporated to draw meaningful statistical conclusions. In this respect, often restricted analysis appears to be more appropriate from an interpretational point of view. If the measurements have a nonnegativity dependence (association) structure (within each set), then it is quite natural to work with convex CCA through nonnegativity conditions on the coefficients in the canonical variables. In **neural networks** such a phenomenon occurs quite often, and Das & Sen [14] have pointed out the relevance of restricted CCA along with the related sampling theory (in an asymptotic setup with due emphasis on applicable

resampling plans). Although such formulations are more appropriate for (nearly) multinormal distributions, they remain asymptotically quite viable for a large class of continuous distributions (having finite moments up to the fourth order at least). To cope with plausible departures from multinormality, both nonlinear and semilinear (where only powers of the linear combinations are allowed), versions of canonical correlation analysis have also been studied in the literature; see, for example, Gambus et al. [24], and Tielemans et al. [60].

In the period from 1985 to 1995 alone, there have been more than 200 published works that use canonical correlations in various branches of biostatistics. Clearly, any summary of all of them is beyond the scope of the current study. Instead, in what follows we address several key issues related to using CCA in biostatistics applications in general.

1. Typically with measurements related to nonnegative variates, even the marginal distributions are (mostly, positively) skewed, and hence suitable **transformations** on the coordinate variables (which are usually highly nonlinear) are advocated to induce more symmetry (if not normality) in the distributions of the transformed variables. This does not, however, guarantee that the joint distribution of the transformed variables would be closely multinormal. Thus the conventional multinormality assumption based canonical correlation analysis may not always be appropriate in such applications. The picture with nonnormal canonical correlations is comparatively better.
2. The motivation behind the use of CCA rests heavily on the linearity structure, and transformations mentioned before may distort such relationships even if the original variables had such structural properties. However, the very rationality of choosing the canonical variables as linear compounds of within-subset variates may be questionable, and without this linearity the foundation of CCA may not be firm enough. Even if such a linearity structure is tenable, but a linear combination of (within-set) variates is not that physically interpretable, motivations for CCA would be diffused to a certain extent. Therefore, in practice, before a CCA is adopted, the suitability of linear compoundability of (transformed) variable has to be examined carefully.
3. In classical CCA, underlying the normality and linearity structure there is the basic assumption that all the variables are quantitative in nature, and are continuous. For discrete variables, generally the linearity of regression and/or normality approximations may not always be feasible. This calls for alternate canonical measures of such association patterns. We illustrate this point with two important examples from biostatistics. First, consider the case of a *multivariate Poisson distribution*, which we introduce as follows. Consider a stochastic $t(\geq p)$ vector $\mathbf{Z} = (Z_1, \ldots, Z_t)'$ of independent **Poisson** variates with positive parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_t)'$, not necessarily all equal. We then express the observable stochastic vector $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{B}\mathbf{Z},$$

where $\mathbf{B}$ is a suitable $p \times t$ matrix of real constants. In order that $\mathbf{X}$ has a multivariate Poisson distribution, all the marginal ones need to be univariate Poisson, so that the elements of $\mathbf{B}$ are all nonnegative. Moreover, for the increments of the marginal distributions to be concentrated to the set of all nonnegative integers, we need to assume that the elements of $\mathbf{B}$ are binary (i.e. 0 or 1). Finally, in order that the distribution of each X_j is nondegenerate, we need that each row of $\mathbf{Z}$ is nonnull. All these conditions imply that the elements of $\mathbf{X}$ can only be associated nonnegatively, and that even in a pairwise case, these association parameters may not always be in the entire interval $(-1, 1)$. Clearly, if we need to incorporate CCA in this context, we need to consider the restricted version, where the restrictions facilitate the verification of the regularity conditions needed for $\mathbf{B}$. In this case, the restricted CCA makes more sense than its classical counterpart based on the usual matrix $\mathbf{S}_n^*$. Even so, the adequacy of normality approximation remains to be checked thoroughly. In the univariate case, usually the *square root* transformation is used for Poisson-type variables to stabilize the asymptotic variance and accelerate the normality approximations, although the rationality of linear compoundability of these transformed variables may not always be clear in practice (*see* **Power Transformations**). As a second illustration, consider *multivariate counting processes* that arise in survival analysis,

neural networks, and also reliability networks. In this context, too, there are certain structural constraints that may block the direct adaptation of CCA.

4. Clinical and educational psychometry, mental testing, and psychiatry have all been identified as vital domains requiring (bio)statistical methodology to a considerable extent. Multivariate analysis is a vital discipline for modeling and analyzing of experiments or studies in this broad domain. Interestingly, some of the classical illustrations in canonical correlations, principal components, and factor analysis have a distinctly psychometric or educational testing flavor, and hence it is worthwhile to examine the relevance of CCA in this domain. Typically, we conceive of some underlying *traits*, and convert the count data set on a dichotomous or **polytomous** classification to suitable **scores**, such as the Z scores or **normal scores** in psychometry, and use standard CCA (or multivariate analysis, in general) on such score data. Even if we have an adequately large sample size, such a conversion may not usually validate the use of conventional canonical correlations. For example, in a multidimensional contingency table, say a **two-by-two table**, these normal scores leading to the *tetrachoric correlations* (*see* **Association, Measures of**) do not capture the linearity of the regression, and hence the linear compoundability of the scores from different characteristics or tests remains open to question – the nonlinear/semilinear ramifications of canonical correlations may be adopted in these cases. The picture is quite different for external analyses, such as the multivariate analysis of variance and covariance models. In this setup, more basic categorical data models with latent continuous variates could provide an alternate and valid approach to (restricted) CCA.
5. In psychometric research, as well as in other areas of biostatistics, it is not uncommon to encounter a purely ranked data set in one or more dimensions. **Rank correlations** and measures of association have all received due attention in the literature, and CCA remains pertinent to this domain as well. Puri & Sen [47, Chapter 8] contains a broad coverage of such rank measures of association along with their statistical properties and sampling distributions. The emphasis has been mainly on the testing of statistical hypotheses, although the matrix of rank-based correlations amends readily to such complex measures of between-group association. One of the basic drawbacks of rank methods is their possible lack of affine invariance, and hence the linear compoundability of the original variables again may not be suitable. Nevertheless, the asymptotic joint multinormality of these sample rank correlations can at least intuitively justify the adoption of CCA when the sample size is large.
6. Mixed models relating to partly quantitative (continuous) responses and partly qualitative (categorical) ones are often encountered in biostatistical analysis. The classical **biserial correlation** is the precursor of such measures of association in the bivariate case. Their generalizations to canonical correlations need a considerable amount of care in formulation so that mathematical manipulations may not preclude statistical interpretations.

At the present time, some of the applications made in **pharmacokinetics**, biopharmaceutics, ecology, epidemiology, and environmetrics deserve critical appraisal for their validity; nevertheless, these represent commendable attempts to gain insight into the phenomena under study (see [1, 3, 4, 7, 9, 11, 12, 18, 20, 22, 24, 26, 28, 38, 40, 41, 43, 45, 52, 56, 60, 62–66].

References

[1] Alterman, A.I., Kushner, H. & Holahan, J.M. (1990). Cognitive functioning and treatment outcome in alcoholics, *Journal of Nervous Diseases* **178**, 494–499.

[2] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[3] Armitage, P., Hoffmann, R., Fitch, T., Morel, C. & Bonato, R. (1995). A comparison of period amplitude and power spectral analysis of sleep EEG in normal adults and depressed outpatients, *Psychiatric Research* **56**, 245–256.

[4] Azais-Braesco, V., Moriniere, C., Guesne, B., Partier, A., Bellenand, P., Bagnelin, D., Grolier, P. & Alix, E. (1995). Vitamin A status in the institutionalized elderly. Critical analysis of four evaluation criteria: Vitamin A intake, Serum retinol, Relative dose-response test (RDR) and Impression cytology with transfer (ICT), *International Journal for Vitamin and Nutrition Research* **65**, 151–161.

[5] Baksalary, J.K., Puntanen, S. & Yanai, H. (1992). Canonical correlations associated with symmetric

reflexive generalized inverses of the dispersion matrix, *Linear Algebra and Applications* **176**, 61–74.

[6] Bérubé, J., Hartwig, R.E. & Styan, G.P.H. (1991). On canonical correlations and the degree of nonorthogonality in the three-way layout, in *Statistical Sciences and Data Analysis, Proceedings of the Third Pacific Area Statistical Conference*, K. Matusita, ed. VSP, Netherlands, pp. 253–263.

[7] Braun, C.M.J. & Richer, M. (1993). A comparison of functional indexes, derived from screening tests, of chronic alcoholic neurotoxicity in the cerebral cortex, retina and peripheral nervous system, *Journal of Studies of Alcoholism* **54**, 11–16.

[8] Brillinger, D.R. (1974). *Time Series: Data Analysis and Theory*. Holt, Rinehart & Winston, New York.

[9] Clarke, G., Hops, H., Lewinsohn, P.M., Andrews, J., Seeley, J.R. & Williams, J. (1992). Cognitive-behavioral group treatment of adolescent depression: prediction of outcome, *Behavior Therapy* **23**, 341–354.

[10] Constantine, A.G. (1963). Some non-central distribution problems in multivariate analysis, *Annals of Mathematical Statistics* **34**, 1270–1285.

[11] Cserhatu, T. & Forgacs, E. (1995). Use of canonical correlation analysis for the evaluation of chromatographic retention data, *Chemometric Intelligence Laboratory Systems* **28**, 305–313.

[12] Culasso, F., Lenzi, A., Gandini, L., Lombardo, F. & Dondero, F. (1993). Statistical andrology: standard semen analysis and computer-assisted sperm motility analysis, *Archives of Andrology* **30**, 105–110.

[13] Das, S. & Sen, P.K. (1994). Restricted canonical correlations, *Linear Algebra and Applications* **210**, 29–47.

[14] Das, S. & Sen, P.K. (1995). Simultaneous spike-trains and stochastic dependence, *Sankhyā, Series B* **57**, 32–47.

[15] Das, S. & Sen, P.K. (1996). Asymptotic distributions of restricted canonical correlations and relevant resampling methods, *Journal of Multivariate Analysis* **56**, 1–19.

[16] Dempster, A.P. (1969). *Continuous Multivariate Analysis*. Addison-Wesley, Boston.

[17] DeSarbo, W.S., Hausman, R.E., Lin, S. & Thompson, W. (1982). Constrained canonical correlation, *Psychometrika* **47**, 489–516.

[18] Dickson, K.L., Waller, W.T., Kennedy, J.H. & Ammann, L.P. (1992). Assessing the relationship between ambient toxicity and instream biological response, *Environmental Toxicology and Chemometry* **11**, 1307–1322.

[19] Dillion, W.R. & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. Wiley, New York.

[20] Dishman, R.K., Darrcott, C.R. & Lambert, L.T. (1992). Failure to generalize determinants of self-reported physical activity to a motor sensor, *Medicine and Science in Sports and Exercise* **24**, 904–910.

[21] Eaton, M.L. (1983). *Multivariate Statistics: A Vector Space Approach*. Wiley, New York.

[22] Fogle, L.L. & Glaros, A.G. (1995). Contributions of facial morphology, age, and gender to EMG activity under biting and resting conditions: a canonical analysis, *Journal of Dental Research* **74**, 1496–1500.

[23] Friman, O., Borga, P., Lundberg, P. & Knutsson, H. (2003). Adaptive analysis of fMRI data, *NeuroImage* **19**(3), 837–845.

[24] Gambus, P.L., Gregg, K.M. & Shafer, S.L. (1995). Validation of alfentanil canonical univariate parameter as a measure of opioid effect on the electroencephalogram, *Anesthesia* **83**, 747–756.

[25] Golub, G.H. & Zha, H. (1994). Perturbation analysis of the canonical correlations of matrix pairs, *Linear Algebra and Applications* **210**, 3–28.

[26] Gregg, K.M., Varvel, J.R. & Shafer, S.L. (1992). Application of semilinear canonical correlation to the measurement of opioid drug effect, *Journal of Pharmacology and Biopharmacology* **20**, 611–635.

[27] Hastie, T., Tibshirani, R. & Friedman, J.H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics.

[28] Horst, P. (1961). Generalized canonical correlations and their applications to experimental data, *Journal of Clinical Psychology* **14**, 331–347.

[29] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441, 498–520.

[30] Hotelling, H. (1935). The most predictable criterion, *Journal of Educational Psychology* **26**, 139–142.

[31] Hotelling, H. (1936). Relation between two sets of variates, *Biometrika* **28**, 321–377.

[32] Hsu, P.L. (1941). On the limiting distribution of canonical correlations, *Biometrika* **32**, 38–45.

[33] Jewell, N.P. & Bloomfield, P. (1983). Canonical correlations of past and future for time series: definitions and theory, *Annals of Statistics* **11**, 837–847.

[34] Kettenring, J.R. (1971). Canonical analysis of several sets of variables, *Biometrika* **58**, 433–450.

[35] Kettenring, J.R. (1983). Canonical analysis, in *Encyclopedia of Statistical Sciences*, Vol. 1, S. Kotz & N.L. Johnson eds. Wiley, New York, pp. 354–365.

[36] Khatri, C.G. (1976). A note on multiple and canonical correlation for a singular covariance matrix, *Psychometrika* **41**, 465–470.

[37] Kshirsagar, A. (1972). *Multivariate Analysis*. Marcel Dekker, New York.

[38] Lehmann, D., Grass, P. & Meier, B. (1995). Spontaneous conscious covert cognition states and brain electric spectral states in canonical correlation, *International Journal of Psychophysics* **19**, 41–52.

[39] Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, New York.

[40] Meagher, T.R. (1992). The quantitative genetics of sexual dimorphism in Silene latifolia, *Evolution* **46**, 445–457.

[41] Moeur, M. & Stage, A.R. (1995). Most similar neighbor: an improved sampling inference procedure for natural resource planning, *Forest Science* **41**, 337–359.

[42] Morrison, D.F. (1967). *Multivariate Statistical Methods*. McGraw-Hill, New York.

[43] Mueller, W.H., Marbella, A., Harrist, R.B., Kaplowitz, H.J., Grubaum, J.A. & Labarthe, D.R. (1990). Body circumferences as measures of body fat distribution in 10 to 14-year-old schoolchildren, *American Journal of Human Biology* **2**, 117–124.

[44] Muirhead, R.J. & Waternaux, C.M. (1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for non-normal populations, *Biometrika* **67**, 31–43.

[45] Nilsson, A.N. & Svensson, B.W. (1995). Assemblages of dytiscid predators and culicid prey in relation to environmental factors in natural and clear-cut boreal swamp forest pools, *Hydrobiologia* **308**, 183–196.

[46] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine* **2**(sixth series), 559–572.

[47] Puri, M.L. & Sen, P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[48] Rao, B.R. (1969). Partial canonical correlations, *Trabajos Estadistica y Investigationes Operationes* **20**, 211–219.

[49] Rao, C.R. (1973). *Linear Statistical Inference and its Application*, 2nd Ed. Wiley, New York.

[50] Rao, C.R. (1981). A lemma on g-inverse of a matrix and computation of correlation coefficients in the singular case, *Communications in Statistics – Theory and Methods* **10**, 1–10.

[51] Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

[52] Saama, P.M., Mao, I.L. & Holter, J.B. (1995). Nutrition, feeding, and calves, *Journal of Dairy Science* **78**, 1945–1953.

[53] Sen, P.K. (1993). Statistical perspectives in clinical and health sciences: the broad way to modern applied statistics, *Journal of Applied Statistical Science* **1**, 1–50.

[54] Sengupta, A. (1983). Generalized canonical variables, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 326–330.

[55] Sengupta, A. (1991). Generalized correlations in the singular case, *Journal of Statistical Planning and Inference* **28**, 241–245.

[56] Smith, L.W., Patterson, T.L. & Grant, I. (1990). Avoidant coping predicts psychological disturbance in the elderly, *Journal of Nervous Diseases* **178**, 525–530.

[57] Stewart, D.K. & Love, W.A. (1968). A general canonical correlation index, *Psychological Bulletin* **70**, 160–163.

[58] Styan, G.P.H. (1986). Canonical correlations in three way layout, in *Pacific Statistical Congress*, I.S. Francis, B.F.J. Manly & F.C. Lam, eds. North-Holland, Amsterdam, pp. 433–438.

[59] Takeuchi, K., Yanai, H. & Mukherjee, B.N. (1982). *The Foundations of Multivariate Analysis*. Halsted Press, New York.

[60] Tielemans, E., Heederik, D. & van Pelt, W. (1994). Changes in ventilatory function in grain processing and animal feed workers in relation to exposure to organic dust, *Scandinavian Journal of Work and Environmental Health* **20**, 435–443.

[61] Timm, N.H. & Carlson, J.E. (1976). Part and bipartial canonical correlation analysis, *Psychometrika* **41**, 159–176.

[62] Vicario, A., Mazon, L.I., Agurre, A., Estomba, A. & Lostao, C. (1989). Relationships between environmental factors and morph polymorphism in Cepaea nemoralis, using canonical correlation analysis, *Genome* **32**, 908–912.

[63] Wade, J.B., Dougherty, L.M., Hart, R.P., Rafii, A. & Price, D.D. (1992). A canonical correlation analysis of the influence of neuroticism and extraversion on chronic pain, suffering, and pain behavior, *Pain* **51**, 67–73.

[64] Whaley, M.H., Kaminsky, L.A., Dwyer, G.B., Getchell, L.H. & Nortin, J.A. (1992). Predictors of over and underachievement of age-predicted maximal heart rate, *Medicine and Science in Sports and Exercise* **24**, 1173–1179.

[65] Williams, J.G. & Kleinfekter, K.J. (1989). Perceived problem-solving skills and drinking patterns among college students, *Psychological Reports* **65**, 1235–1244.

[66] Yanai, H. & Takane, Y. (1992). Canonical correlation analysis in linear constraints, *Linear Algebra and Applications* **176**, 75–89.

(*See also* **Multivariate Analysis, Overview**)

SHUBHABRATA DAS & PRANAB K. SEN

Capture–Recapture

In a typical capture–recapture experiment in biological and ecologic sciences, we place traps or nets in the study area and sample the population several times. At the first trapping sample a number of animals are captured; the animals are uniquely tagged or marked and released into the population. Then at each subsequent trapping sample we record and attach a unique tag to every unmarked animal, record the capture of any animal that has been previously tagged, and return all animals to the population. At the end of the experiment the complete capture history for each animal is known. Such experiments are also called mark-recapture, tag-recapture, and multiple record systems in the literature. The simplest type only includes two samples; one is the capture sample and the other the recapture sample. This special two-sample case is often referred to as a "dual system" or a "dual-record system" in the context of census undercount estimation.

The capture–recapture technique has been used to estimate population sizes and related parameters such as survival rates, birth rates, and migration rates. Biologists and ecologists have long recognized that it would be unnecessary and almost impossible to count every animal in order to obtain an accurate estimate of population size. The recapture information (or the proportion of repeated captures) by marking or tagging plays an important role because it can be used to estimate the number missing in the samples under proper assumptions. Intuitively, when recaptures in subsequent samples are few, we know that the size is much higher than the number of distinct captures. However, if the recapture rate is quite high, then we are likely to have caught most of the animals.

According to Seber [15], the first use of the capture–recapture technique can be traced back to Laplace, who used it to estimate the population size of France in 1786. The earliest applications to ecology include Petersen's and Dahl's work on fish populations in 1896 and 1917, respectively, and Lincoln's use of band returns to estimate waterfowl in 1930. More sophisticated statistical theory and inference procedures have been proposed since the paper by Darroch [5], who founded the mathematical framework of this topic. See [15]–[17] and references therein for the historical developments, methodologies, and applications.

The models are generally classified as either closed or open. In a closed model the size of a population, which is the main interest, is assumed to be constant over the trapping times. The closure assumption is usually valid for data collected in a relatively short time during a nonbreeding season. In an open model, recruitment (birth or immigration) and losses (death or emigration) are allowed. It is usually used to model the data from long-term investigations of animals or migrating birds. In addition to the population size at each sampling time, the parameters of interest also include the survival rates and number of births between sampling times. Here we concentrate on closed models because of their applications to epidemiology and health science.

Applications to Epidemiology

The capture–recapture model originally developed for animal populations has been applied to human populations under the term "multiple-record systems". A pioneering paper is that of Sekar & Deming [18], who used two samples to estimate the birth and death rates in India. Wittes & Sidel [19] were the first to use three-sample records to estimate the number of hospital patients. Related subsequent applications were given in an earlier overview by El-Khorazaty et al. [7].

Epidemiologists recently have shown renewed and growing interest in the use of the capture–recapture technique. As LaPorte et al. [13] indicated, the traditional public-health approaches to counting the number of occurrences of diseases are too inaccurate (surveillance), too costly (population-based registries; *see* **Disease Registers**), or too late (death certificates) for broad monitoring. They felt that it was time to start counting the incidences of diseases in the same way as biologists count animals. Two recent review articles [11, 12] by the International Society for Disease Monitoring and Forecasting proposed that the capture–recapture method would provide a technique for enhancing our ability to monitor disease. Reference [12] also reviewed its applications to the following categories: birth defects, cancers, drug use, infectious diseases, injuries, and diabetes as well as other areas of epidemiology.

The purpose of most applications to epidemiology is to estimate the size of a certain target population by merging several existing but incomplete lists

of names of the target population. If each list is regarded as a trapping sample and identification numbers and/or names are used as "tags", then it is similar to a closed capture–recapture setup for wildlife estimation. Now the "capture in a sample" corresponds to "being recorded or identified in a list", and "capture probability" becomes "ascertainment probability". Two major differences between wildlife and human applications are (i) there are more trapping samples in wildlife studies, whereas in human studies only two to four lists are available; and (ii) in animal studies there is a natural temporal or sequential time order in the trapping samples, whereas for epidemiologic data such order does not exist in the lists, or the order may be different for some individuals. Researchers in wildlife and human applications have respectively developed models and methodologies along separate lines. Three of these approaches are discussed after the data structure and assumptions are explained.

Data Structure and Assumptions

Ascertainment data for all identified individuals are usually aggregated into a categorical data form. We give in Table 1 a three-list hepatitis A virus example for illustration. The purpose of this study was to estimate the number of people who were infected by hepatitis in an outbreak that occurred in and around a college in northern Taiwan from April to July 1995. Our data are restricted to those records from students of that college. A total of 271 cases were reported from the following three sources: (i) P-list (135 cases): records based on a serum test conducted by the Institute of Preventive Medicine of Taiwan. (ii) Q-list (122 cases): records reported by the National Quarantine Service based on cases reported by the doctors of local hospitals. (iii) E-list (126 cases): records based on questionnaires collected by epidemiologists.

In Table 1, for simplicity, the presence or absence in any list is denoted by 1 and 0, respectively. There are seven observed cells Z_{100}, Z_{010}, Z_{001}, Z_{110}, Z_{011}, Z_{101}, and Z_{111}. Here $Z_{111} = 28$ means that there were 28 people recorded on all three lists; $Z_{100} = 69$ means that 69 people were recorded on list P only. A similar interpretation pertains to other records. Let n_1, n_2, and n_3 be the number of cases in P, Q and E, respectively. Then $n_1 = Z_{111} + Z_{110} + Z_{101} + Z_{100} = 135$. Similar expressions hold for n_2 and n_3. There is one missing cell, Z_{000}, the number of uncounted. The purpose is to predict Z_{000} or to estimate the total population size.

Table 1 Data on hepatitis A virus

Hepatitis A virus list			
P	Q	E	Data
1	1	1	$Z_{111} = 28$
1	1	0	$Z_{110} = 21$
1	0	1	$Z_{101} = 17$
1	0	0	$Z_{100} = 69$
0	1	1	$Z_{011} = 18$
0	1	0	$Z_{010} = 55$
0	0	1	$Z_{001} = 63$
0	0	0	$Z_{000} = ??$

A crucial assumption in the traditional approach is that the samples are independent. Since individuals can be cross classified according to their presence or absence in each list, the independence for two samples is usually interpreted from a 2×2 categorical data analysis in human applications. This assumption in animal studies is expressed in terms of the "equal-catchability assumption": all animals have the same probability of capture in each sample. However, this assumption is rarely valid in most applications. Lack of independence among samples leads to a bias for the usual estimators derived under the independence assumption. The bias may be caused by the following two sources:

1. List dependence within each individual (or substratum): that is, inclusion in one sample has a direct causal effect on any individual's inclusion in other samples. For example, an individual with a positive for the serum test of hepatitis is more likely to go to the hospital for treatment and thus the probability of being identified in local hospital records is larger than that of the same individual given as negative by the serum test. Therefore, the "capture" of the serum test and the "capture" of hospital records become positively dependent. This type of dependence is usually referred to as "list dependence" in the literature.
2. Heterogeneity between individuals (or substrata): even if the two lists are independent within individuals, the ascertainment of the two lists may become dependent if the capture probabilities

are heterogeneous among individuals. This phenomenon is similar to **Simpson's paradox** in categorical data analysis. That is to say, aggregating two independent **2 × 2 tables** might result in a dependent table. Hook & Regal [10] provided an example.

The above two types of dependences are usually confounded and cannot be easily disentangled in a data analysis without further assumptions. We discuss three approaches (the ecologic model, the **loglinear model**, and the sample coverage approach) that allow for the above two types of dependences.

Ecologic Models

Pollock, in his 1974 Ph.D. thesis and subsequent papers (e.g. Pollock [14]), proposed a sequence of models mainly for wildlife studies to relax the equal-catchability assumption. This approach aims to model the dependences by specifying various forms of "capture" probability. The basic models include (i) model M_t, which allows capture probabilities to vary with time; (ii) model M_b, which allows the capture of behavioral responses; and (iii) model M_h, which allows heterogeneous animal capture probabilities. Various combinations of these three types of unequal capture probabilities (i.e. models M_{tb}, M_{th}, M_{bh}, and M_{tbh}) are also proposed.

Only for model M_t are the samples independent. List dependence is present for models M_b and M_{tb}; heterogeneity arises for model M_h; and both types of dependences exist for models M_{bh} and M_{tbh}. For any model involving behavioral response, the capture probability of any animal depends on its "previous" capture history. However, there is usually no sequential order in the lists, so those models have limited use in epidemiology. Models M_h and M_{th} might be useful for epidemiological studies. Various estimation procedures have been proposed. See [15]–[17] for reviews.

Loglinear Models

The loglinear model approach is a commonly used technique for epidemiological data. Loglinear models that incorporate list dependence were first proposed by Fienberg [8] for dealing with human populations. Cormack [4] proposed the use of this technique for several ecologic models.

In this approach the data are regarded as a form of an incomplete 2^t **contingency table** (t is the number of lists) for which the cell corresponding to those individuals uncounted by all lists is missing. A basic assumption is that there is no t-sample **interaction**. This assumption implies an extrapolation formula for the number of uncounted. For three lists, the most general model is a model with main effects and three two-sample interaction terms. Various loglinear models are fitted to the observed cells and a proper model is selected using deviance statistics and the **Akaike information criterion**. The chosen model is then projected onto the unobserved cell.

List dependences correspond to some specific interaction terms in the model. As for heterogeneity, **quasi-symmetric** and partial quasi-symmetric models of **loglinear models** can be used to model some types of heterogeneity, i.e. **Rasch models** and their generalizations; see [1] and [6]. Since the quasi-symmetric or partial quasi-symmetric models are equivalent to assuming that some two-factor interaction terms are identical, the heterogeneity corresponds to some common interaction effects in loglinear models. Details of the theory and development are fully discussed in [11] and [12].

Sample Coverage Approach

The idea of sample coverage, originally from I.J. Good and A.M. Turing (Good [9]), has been used in species and animal population size estimation; see Chao & Lee [2]. The same approach was also applied to epidemiologic data in [3].

This approach aims to model dependences by some parameters, which are called "coefficients of variation", defined for two or more samples. The magnitude of the parameters measures the degree of dependence of samples. The two types of dependences are confounded in these measures. In the independent case, all dependence measures are zero. This general model encompasses the Rasch model and the ecologic models as special cases.

A common definition for the sample coverage of a given sample is the probability-weighted fraction of the population that is discovered in that sample. For multiple-sample type of data the sample coverage is

modified to be the probability-weighted fraction of the population that is jointly covered by the available samples. See [3] for a formal definition. The basic motivation here is that sample coverage can be well estimated even in the presence of two sources of dependences. Thus an estimate of population size can be obtained via the relation between the population size and sample coverage. Chao et al. [3] have shown that an estimator of C is $\hat{C} = 1 - (Z_{100}/n_1 + Z_{010}/n_2 + Z_{001}/n_3)/3$, which is one minus the average of the proportion of individuals listed in only one sample (i.e. singletons). Let D be the average of the distinct cases for three pairs of samples. In this approach, when all three samples are independent, a valid estimator is $\hat{N}_0 = D/\hat{C}$. If any type of dependence arises, then Chao et al. [3] attempt to account for the dependences by adjusting $D/\hat{C}$ based on a function of the estimates of the coefficients of variation. In the same reference, estimators of population size are proposed separately for high sample coverages (e.g. if $\hat{C}$ is over 55%) and low sample coverages.

Analysis of the Hepatitis Example (Low Sample Coverage)

Several loglinear models were fitted to the hepatitis data given in Table 1. Except for the saturated model, the loglinear models that do not take heterogeneity into account (e.g. models with one or two interaction terms) do not fit the data well, whereas all other models that take heterogeneity into account (quasi-symmetric and partial quasi-symmetric models) fit well. All those adequate models yielded very similar estimates – 1300 with an approximate estimated standard error of 520.

The coverage estimate is $\hat{C} = 51.27\%$, which is considered to be low. The average of the distinct cases for three pairs of samples is $D = 208.667$. If the incorrect independence is assumed, then an estimate would be $\hat{N}_0 = D/\hat{C} = 407$. (The loglinear independent model yields a similar estimate of 388.) It follows from Chao et al. [3] that the estimates for dependence measures are relatively large, which indicates that the three samples are pairwise positively dependent and $\hat{N}_0$ would generally underestimate. However, one cannot distinguish which type of dependence is the main cause of the bias. Incorporating the bias due to dependences along the sample coverage approach results in an estimate of 508. An estimated standard error of 40 is calculated by using a **bootstrap method** based on 1000 replications. The resulting 95% **confidence interval** is (407, 591) based on the same bootstrap replications.

This example shows that the loglinear and sample coverage approaches may give widely different estimates. Moreover, the example in [11] further shows that several loglinear models that fit the data equally well might also result in quite different estimates. **Simulation** comparisons of the two approaches and other examples with high sample coverages are provided in Chao et al. [3].

References

[1] Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort, *Biometrics* **50**, 494–500.

[2] Chao, A. & Lee S.-M. (1992). Estimating the number of classes via sample coverage, *Journal of the American Statistical Association* **87**, 210–217.

[3] Chao, A., Tsay, P.K., Shau, W.-Y. & Chao, D.-Y. (1996). Population size estimation for capture–recapture models with applications to epidemiological data, *Proceedings of Biometrics Section, American Statistical Association*, pp. 108–117.

[4] Cormack, R.M. (1989). Loglinear models for capture-recapture, *Biometrics* **45**, 395–413.

[5] Darroch, J.M. (1958). The multiple recapture census I. Estimation of a closed population, *Biometrika* **45**, 343–359.

[6] Darroch, J.N., Fienberg, S.E., Glonek, G.F.V. & Junker, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability, *Journal of the American Statistical Association* **88**, 1137–1148.

[7] El-Khorazaty, M.N., Imery, P.B., Koch, G.G. & Wells, H.B. (1977). A review of methodological strategies for estimating the total number of events with data from multiple-record systems, *International Statistical Review* **45**, 129–157.

[8] Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables, *Biometrika* **59**, 591–603.

[9] Good, I.J. (1953). On the population frequencies of species and the estimation of population parameters, *Biometrika* **40**, 237–264.

[10] Hook, E.B. & Regal, R.R. (1993). Effects of variation in probability of ascertainment by sources ("variable catchability") upon capture–recapture estimates of prevalence, *American Journal of Epidemiology* **137**, 1148–1166.

[11] International Society for Disease Monitoring and Forecasting (1995). Capture–recapture and multiple-record systems estimation I: history and theoretical

development, *American Journal of Epidemiology* **142**, 1047–1058.

[12] International Society for Disease Monitoring and Forecasting (1995). Capture–recapture and multiple-record systems estimation II: Applications in human diseases, *American Journal of Epidemiology* **142**, 1059–1068.

[13] LaPorte, R.E., McCarty, D.J., Tull, E.S. & Tajima, N. (1992). Counting birds, bees, and NCDs, *Lancet* **339**, 494–495.

[14] Pollock, K.H. (1991). Modelling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future, *Journal of the American Statistical Association* **86**, 225–238.

[15] Seber, G.A.F. (1982). *The Estimation of Animal Abundance*, 2nd Ed. Griffin, London.

[16] Seber, G.A.F. (1986). A review of estimating animal abundance, *Biometrics* **42**, 267–292.

[17] Seber, G.A.F. (1992). A review of estimating animal abundance II, *International Statistical Review* **60**, 129–166.

[18] Sekar, C. & Deming W.E. (1949). On a method of estimating birth and death rates and the extent of registration, *Journal of the American Statistical Association* **44**, 101–115.

[19] Wittes, J.T. & Sidel, V.W. (1968). A generalization of the simple capture–recapture model with applications to epidemiological research, *Journal of Chronic Diseases* **21**, 287–301.

(*See also* **Structural and Sampling Zeros**)

ANNE CHAO

Cardiology and Cardiovascular Disease

Cardiology is the study of the heart and its diseases. Medical treatises give long lists of potential causes of heart diseases. However, as the functioning of the heart hinges on a constant and adequate blood supply, many heart diseases can be traced back to diseases of the vessels bringing blood to the heart. Thus, in biomedical textbooks and in epidemiology, heart diseases and vascular diseases are usually treated under the common heading of cardiovascular diseases (CVDs). Cardiovascular diseases comprise all diseases classified according to the **International Classification of Diseases**, 9th Revision.

Cardiovascular diseases are one of the major causes of morbidity and mortality worldwide with a death count of around 14 million per year. As reported by the **World Health Organization** [41], in spite of a very marked decline over the last few decades (for instance, male mortality from CVD dropped by 60% in Japan, and by 50% in Australia, Canada, France, and the US), CVDs are responsible for 20% of all deaths worldwide: 50% of all deaths in industrialized countries and 16% of all deaths in developing countries. Absolute estimates for 1990 are as follows: 10.9 million deaths occurred in developed countries, and 45 million deaths in developing countries; of these, the deaths attributed to CVDs were 5.3 million for developed countries and 9 million for developing countries. However, mortality is not sufficient to describe the impact of CVDs. It is also estimated that 25%–30% of the CVD burden arises from their disabling sequelae other than death. The American Heart Association translated these mortality and morbidity considerations into **health economics** and predicted that the cost of CVDs in the US for the year 1996 would be US$151.3 billion [1].

The most common serious heart disease is indeed a disease of the vessels: coronary heart disease (CHD), defined as cardiac ischemia (insufficient blood supply) due to atherosclerosis of the coronary arteries. The American Heart Association, in its 1996 statistical supplement, reports that in 1993 CHD caused 489 970 deaths in the US, or 1 of every 4.6 deaths. Angina (chronic chest pains due to cardiac ischemia) is often the first symptom of CHD. Angina may evolve as a distinct disease, or a heart attack may follow its onset. On the other hand, a heart attack may strike asymptomatic subjects: indeed, there are no previous signs of CHD in 48% of men and in 63% of women who die of a heart attack. The end stage of CHD may be Congestive Heart Failure (CHF), which manifests as a failure of the heart to pump blood as needed. CHF is also seen as the last stage of other heart diseases such as valvular diseases and cardiomyopathies (diseases of the heart muscle). Besides contributing to an estimated 250 000 deaths a year, CHF is also listed as a direct cause of death (36 387 US deaths in 1992).

Arrhythmia (irregularities of the heart rhythm, 40 843 US deaths in 1992) may appear as a consequence of CHD or may have a distinct cause. The most serious arrhythmia is ventricular fibrillation (quivering of the heart, replacing its regular pumping function, 1461 US deaths in 1992). Although it represents a relatively modest burden as an official cause of death, it is thought to cause the overwhelming proportion of sudden deaths, estimated at about 250 000 per year in the US.

Not all heart diseases are vascular diseases, for example: rheumatic heart disease, valvular diseases, infectious cardiomyopathies and idiopathic cardiomyopathies. Similarly, not all vascular diseases directly affect the heart. **Stroke** (brain attack) is the second most important CVD: it caused 149 740 deaths in 1993 in the US, or 1 in 15 deaths. Hypertension, an underlying condition of many CVDs and indeed a contributing factor to atherosclerosis, is also considered a disease in its own right: as such, it killed 37 520 in 1993 in the US.

CVDs, in particular CHD, are in principle both treatable and preventable. The problem with treatment is that often patients die before getting the needed intervention. Hence, the emphasis is on developing diagnostic tools for cardiovascular diseases with the aim of preventing sudden episodes like heart attacks and, more generally, of slowing down the aggravation of subclinical pathological processes. The evolution of cardiac diagnostic devices has proceeded at great speed. Starting in the nineteenth century with Laënnec's invention of the stethoscope (1816) and with the introduction by Riva & Rocci (1896) of the modern blood pressure measuring device (sphygmomanometer), techniques have become increasingly sophisticated in the

twentieth century. Electrocardiography, developed by Einthoven in 1903, introduced a relatively simple measure of the heart's rhythm from which the functioning of the heart could be studied, while radiography and angiocardiography (radiography of arteries by means of appropriate contrasting substances) allowed visualization of the heart (1919). Cardiac catheterization (1931) allowed direct access to diseased areas of the heart, and the combination of catheterization with angiography is the basis of contemporary invasive, but very accurate, diagnostic devices, such as selective angiography, coronary arteriography, and digital subtraction angiography. Concern for reducing the invasive character of these procedures without sacrificing accuracy has spurred the most recent developments: echocardiography or noninvasive ultrasonography in the mid 1960s and magnetic resonance imaging in the early 1980s are good examples. For other historical milestones in cardiology, see Table 1.

In parallel with diagnostic advances, surgical or invasive treatment of CHD has also made impressive progress in the twentieth century. The 1950s saw the introduction of artificial aortic valves, open heart surgery, and regulation of arrhythmias by implantation of pacemakers. Progress continued in the following three decades with the 1967 first heart transplant by Barnard, the 1977 development of angioplasty (opening of blocked arteries using balloon catheter) by Gruentzig, and the first implant of an artificial heart in 1982 by DeVries. At the same time, pharmacologic approaches aimed at impeding the progress of CHD and at controlling arrhythmias have become increasingly sophisticated and effective. Control of hypertension, one of the leading causes of CHD progress, may now be achieved by several classes of medications, involving very different modes of action (e.g. thiazide diuretics, alpha, beta, calcium channel blockers, ACE inhibitors). Several classes of medications have also been developed to control hyperlipidemias (e.g. nicotinic acid, resins, statins, Gemfibrozil, probucol). To treat arrhythmias there are also several classes of drugs available, with varying modes of action. As another example, aspirin has been proved effective in prevention of a second heart attack, stroke, and mortality in patients who had survived a first heart attack. The great variety of potentially successful therapeutic approaches has prompted, since the mid 1960s, the development of large-scale, **multicenter trials** in cardiology to test the efficacy of these drugs. As for other areas of medicine, the randomized **clinical trial**, double-blinded whenever possible (*see* **Blinding or Masking**), is now considered the norm for monitoring progress and comparing alternative strategies. Table 2A summarizes some of the most important clinical trials in cardiology.

The successes sketched above are considered at least partially responsible for the decline in CHD mortality, which has been reduced by 49% since 1970. This reduction is attributable mostly to secondary prevention, i.e. prevention in patients with clinical manifestations of CHD. However, costs, both human and monetary, remain enormous: for instance, Haltky et al. [15], estimated the 5-year total medical cost of bypass surgery at US$58 498 and that of angioplasty at US$56 225. Clearly, these costs could be greatly reduced if the development of CHD could be prevented altogether (primary prevention).

The need for a major effort in the area of primary prevention, i.e. prevention of the development of CHD in healthy subjects, has become increasingly apparent since the first decades of the twentieth century, with the first description by Herrick (1912, see Table 1) of the relationship between atherosclerosis and CHD, and with the work of Muller (1930, see Table 1), who reports on the relationship between CVD and elevated cholesterol serum level. This awareness has culminated in the historic **Framingham Heart Study** [17]. Framingham is a small community of about 65 000 people, situated 21 miles from Boston. Over 5000 subjects from this community, aged between 30 and 62, were followed for 30 years, starting in 1948, with the aim of addressing a complete picture of the epidemiologic aspects of CHD, its major risk factors, and its evolution. The Framingham study helped identify the risk factors that are now common knowledge, such as heredity, sex, age, cigarette smoking, high cholesterol blood levels, hypertension, sedentary life-style, obesity, diabetes [1]. The Framingham study has served as a model for several similar large-scale endeavors, increasingly including active interventions to modify risk factors or to enhance prevention: a brief summary is given in Table 2B.

As a result of these major studies, our knowledge of CHD has progressed to a point where effective prevention is possible. For example, the National Heart, Lung and Blood Institute (US) estimated in 1994 that

Table 1 Historical events in cardiology

Event	Attributed to	Year
First description of blood circulation	William Harvey	1628
First intravenous injection in a human	Major	1667
First description of the heart structure	Raymond de Vieussens	1706
First known cardiac catheterization	Stephen Hales	1711
First blood pressure measurement	Stephen Hales	1733
Stethoscope	René T.H. Laënnec	1816
First human to human blood transfusion published	James Blundell	1818
Electric current accompanies each heart beat	Carlo Matteucci	1842
Describes an "action potential" accompanying each muscular contraction	Emil Dubois-Reymond	1843
Publication of work on action potential of the heart	Eudolph Albert Von Kolliker	1855
Record the heart's electrical current and shows it consists of two phases	J.B. Sanderson & F. Page	1878
Description of myocardial infarction	Karl Weigert	1880
Description of angina pectoris	Frederick Winsor	1880
Discovery of X-rays	Wilhelm C. Röntgen	1895
Sphygmomanometer	Riva, Rocci	1896
First successful closure of a stab wound of the heart (birth of cardiac surgery)	Ludwig Rehn	1897
Discovery of the first three human blood groups: A, B, and O	Karl Landsteiner[a]	1900
Anastomosis of small blood vessels	Alexis Carrel	1902
Discovery of the fourth blood group, AB	A. Decastello & A. Sturli	1902
Invention of the electrocardiograph	Williem Einthoven[b]	1903
Discovery of the sino atrial node (the origin of heartbeat)	Keith & Flack	1907
Publication of "The mechanism of the heart beat"	Thomas Lewis	1911
First description of heart disease resulting from hardening of the arteries	James B. Herrick	1912
First angiogram in a living person using potassium iodide	Heuser	1919
First soundly established surgical technique for severe mitral stenosis	Souttar	1925
First use of radioactive tracers in human	Blumgart	1926
Relation between CVD and elevated cholesterol serum level was first described	Muller	1930
First blood bank established in London		1930
First human right heart catheterization	Frostmann[c]	1931
First heart surgery	Robert E. Gross	1938
Discovery of angiotensin	Irvine H. Page	1938
Identification of the Rh factor	Karl Landsteiner et al.	1939
Invention of the ambulatory ECG	John Holter	1940
Isolation of albumine	Edwin Cohn	1940
Demonstration of the changing oxygen saturation in blood	Cournand & Richards[c]	1945
Invention of a plastic valve to repair aortic valve	Charles Hufnagel	1951
First successful open heart surgery	F. John Lewis	1952
First report of transthoracic pacing	Paul Zoll	1952
First use of a mechanical heart and blood purifier ECC (Extra Corporal Circulation)	John H. Gibbon	1953
First entirely implantable rechargeable pacemaker	Senning & Elmqvist	1958
First attempt to use fibrinolytic therapy (streptokinase) of MI intravenously	Fletchet	1958
Development of selective coronary angiography	Sones	1959
First real-time instrument for two-dimensional echocardiography	Hertz	1960
First external cardiac massage to restart a heart	J.R. Jude	1961
First Percutaneous Transluminal Coronary Angioplasty (PTCA)	Dotter & Judkins	1964
First radionuclide technique for measuring human myocardial blood flow	Ross	1964
First heart transplant	Christiaan Barnard	1967
First experimental PTCA balloon	Andreas Gruentzig	1974
First human coronary angioplasty	Andreas Gruentzig	1977

(*continued overleaf*)

Table 1 (*continued*)

Event	Attributed to	Year
Introduction of new immunosuppressant: cyclosporine		1978
First cardiac image using magnetic resonance imaging	Lauterbur	1978
First implant of a cardioverter defibrillator	Watkins, Reid, Mirowski et al.	1980
First successful heart lung transplant	Norman Shumway	1981
Invention of the artificial heart	Robert Jarvik	1982
First artificial heart (Jarvik-7) recipient, Barney Clark	Williem De Vries	1982
First real-time two-dimensional color-flow Doppler	Bommer et al.	1982
First use of coronary stent in humans	Sigwart et al.	1987

[a]Won the 1930 Nobel Prize for physiology and medicine for this discovery.
[b]Won the 1924 Nobel Prize for physiology and medicine for this invention.
[c]With Prossmann won the 1956 Nobel Prize for physiology and medicine for their contribution to the advancement of catheterization.
References for these events can be found in: Brandenburg, R.O., Fuster, V., Giuliani, E.R. & McGoon, D.C. (1987). *Cardiology Fundamentals and Practice*. Year Book Medical Publisher, Chicago; Braunwald, E. (1996). *Heart Disease: A Textbook of Cardiovascular Medicine*. W.B. Saunders, Toronto; Colman, R.W., Hirsh, J., Marder, V.J., Salzman, E.W. (1994). *Hemostasis and Thrombosis: Basic Principles and Clinical Practice*, 3rd Ed. J.B. Lippincott Company, Philadelphia; Fye, W.B. (1994). A history of the origin, evolution, and impact of electrocardiography, *American Journal of Cardiology* **73**, 937–949; Grossman, W. (1986). *Cardiac Catheterization and Angiography*. Lea & Febiger, Philadelphia; Harbert, J. & Da Rocha, A.F.G. (1984). *Textbook of Nuclear Medicine*. Vol. II: *Clinical Applications*. Lea & Febiger, Philadelphia; Mollison, P.L. (1972). *Blood Transfusion in Clinical Medicine*, 5th Ed. Blackwell Scientific Publications, Oxford; Schapira, J.N. & Harold, J.G. eds. (1982). *Two Dimensional Echocardiography and Cardiac Doppler*, 2nd Ed. William & Wilkins, Baltimore; Marcus, M.L., Schelbert, H.R., Skorton, D.J. & Wolf, G.L. (1991). *Cardiac Imaging: A Companion to Braunwald's Heart Disease*. W.B. Saunders, Toronto; Schlant, R.C. & Alexander, R.W. (1994). *Hurst's: the Heart*, 8th Ed. McGraw-Hill, New York.

by modifying the major known risk factors for CVD, it would be possible to increase the rate of the decline in CVD mortality to around 6% per year for an overall reduction of 50% in 10 years [25]. However, at the present state of knowledge, the cost of controlling risk factors such as hypertension and hyperlipidemias is formidable, since effective control would require lifelong medical treatments for very large subpopulations with these risk factors. For other risk factors such as life-style changes, control implies behavior modification. This is notoriously difficult and, for certain target groups, ineffective. As research continues towards the development of more powerful and less costly preventive approaches, further study concerning established and new potential risk factors is necessary. A deeper knowledge in this area could lead to the identification of smaller subpopulations at very high risk, in which it would be cost–effective to concentrate preventive efforts.

Statistics and Cardiovascular Diseases

The level of statistical sophistication present in the cardiology literature has radically improved since the 1980 paper by Glantz [12], owing, at least in part, to its impact on the editorial policies of two of the most prestigious journals in the field, *Circulation* and *Circulation Research*. In practical terms, this has meant that all papers are now scrutinized for correctness of their data analytic sections by a statistician, and that tests and procedures beyond the elementary *t*- (*see* **Student's *t* Statistics**) and chi-square tests have become prevalent. In a rapid survey of the present status of statistical methodology in cardiology, we have randomly selected 150 articles from volume 91 of *Circulation* (1995); the results are summarized in Table 3.

The figures in Table 3, compared with what is considered state-of-the-art in clinical biostatistics, show that there is an important time lag between development of new statistical methodologies and their current use in cardiology, as is indeed the case in many other highly specialized areas. For example, we find limited use of techniques for the analysis of survival times (18/150) (*see* **Survival Analysis, Overview**) and repeated measurements (26/150) (*see* **Longitudinal Data Analysis, Overview**). This is rather surprising, in view of the fact that improving survival is a major aim of clinical studies in

Table 2 Some of the most important CVD studies

A Clinical trials	Landmark publications[a]
ACME (Angioplasty Compared to MEdicine)	*New England Journal of Medicine* **326** (1992) 10–16
AMIS (Aspirin Myocardial Infarction Study)	*Journal of the American Medical Association* **243** (1980) 661–667
ART (Anturane Reinfarction Trial)	*New England Journal of Medicine* **302** (1980) 250–256
BHAT (Beta Blocker Heart Attack Trial)	*Journal of the American Medical Association* **247** (1982) 1701–1714
CAMLAT (Canadian Amiodarone Myocardial Infarction Arrhythmia Trial)	*American Journal of Cardiology* **72** (1993) 87F–94F
CASS (Coronary Artery Surgery Study)	*Circulation* (Suppl. 1) **63** (1981) 1–81
CAST (Cardiac Arrhythmia Suppression Trial)	*New England Journal of Medicine* **321** (1989) 406–412
CAVEAT (Coronary Angioplasty vs. Excisional Atherectomy Trial)	*New England Journal of Medicine* **329** (1993) 221–227
CDP (Coronary Drug Project)	*New England Journal of Medicine* **303** (1980) 1038–1041
GISSI (Gruppo Italiano per lo Studio Della Stepptochinasi Nell'Infarto Miocardico)	*Lancet* **i** (1986) 397–401
GUSTO I (Global Utilization of Streptokinase and TPA for Occluded Arteries)	*New England Journal of Medicine* **329** (1993) 673–682
PARIS (Persantine Aspirin Re-Infarction Study)	*Circulation* **62** (1980) 449–461
SOLVD (Study of Left Ventricular Dysfunction)	*New England Journal of Medicine* **325** (1991) 293–302
TIMI (Thrombolysis In Myocardial Infarction)	*Circulation* **76** (1987) 142–154
B Community trials	
ARIC (The Atherosclerosis Risk In Communities)	*American Journal of Epidemiology* **129** (1989) 687–702
British Male Doctors Trial	*British Medical Journal* **296** (1988) 313–316
CABG Patch (Coronary Artery Bypass Graft surgery with/without simultaneous epicardial Patch for automatic implantable cardioverter defibrillator)	*Circulation* **94** (Suppl II) (1996) II-248–II-253
Framingham Heart Study	*Annals of Internal Medicine* **74** (1971) 1–12
HAPPHY (Heart Attack Primary Prevention in Hypertension)	*Journal of Hypertension* **5** (1987) 561–572
HDFP (Hypertension Detection and Follow-up Program Cooperative Group)	*Journal of the American Medical Association* **242** (1979) 2562–2571
Helsinki Heart Study	*New England Journal of Medicine* **317** (1987) 1237–1245
ISIS-2 (Second International Study of Infarct Survival)	*Lancet* **2** (1982) 349–360
LRC-CPPT (Lipid Research Clinics Coronary Primary Prevention Trial)	*Journal of the American Medical Association* **251** (1984) 351–374
MRFIT study (Multiple Risk Factor Intervention Trial)	*Journal of the American Medical Association* **248** (1982) 1465–1477
National Cooperative Pooling Project	*Journal of Chronic Diseases* **31** (1978) 201–306
SHEP (Systolic Hypertension in the Elderly Program)	*Journal of the American Medical Association* **265** (1991) 3255–3264
Stanford Five-City Project	*American Journal of Epidemiology* **132** (1991) 235–249
The Rochester Coronary Heart Disease Project	*Mayo Clinic Proceedings* **64** (1989) 1471–1480
US Physicians' Health Study Aspirin component	*New England Journal of Medicine* **321** (1989) 129–135
US Nurses' Health Study	*New England Journal of Medicine* **313** (1985) 1044–1049
VA Cooperative Study Group on antihypertensive agents	*Journal of the American Medical Association* **202** (1967) 1028–1034
VA Cooperative Study Group on coronary artery surgery	*American Journal of Cardiology* **59** (1987) 1017–1023
WHI (The Women's Health Initiative)	Scheduled to end in 2007
WHO Cooperative Trial on Primary Prevention of Ischaemic Heart Disease	*Lancet* 379–385 (1980)
WHO MONICA Project (MONItoring and CArdiovascular)	Scheduled to end in 2000

[a]This is far from being an exhaustive review of the major studies and their publications. One very good source of studies is the Cochrane Controlled Trials Register (CCTR), which is part of the Cochrane Library (available on CD-ROM), containing more than 70 000 controlled trials in every medical field.

Table 3 Review of the "statistical analysis" sections of 150 randomly selected articles from *Circulation* in 1995
Statistical analysis section (number of lines)[a]: mean = 13.27; standard deviation = 11.8; median = 9; 25th percentile = 6; 75th percentile = 18

Type of analysis[b]	Frequency (n)	Percentage (n/150)(%)
Mann–Whitney U tests	25	16.7
Kruskal–Wallis or Friedman nonparametric ANOVA	2	1.3
Fisher's exact tests	12	8
Linear regression	32	21.3
Pearson coefficient of correlation	13	8.6
Spearman correlation	2	1.3
One-[c] and multiple-way ANOVA	54	36
Repeated measure ANOVA	26	17.3
ANCOVA	6	4
Logistic regression	16	10.6
Conditional logistic regression	1	0.6
Probit regression	1	0.6
ROC curve	2	1.3
Kaplan–Meier and log rank or Gehan tests	18	12
Cox proportional hazard	10	6.7
Time-dependent Cox proportional hazard	1	0.6
Mixed and random effect models	2	1.3
Kappa, Kendall Tau b statistic and intraclass correlation	4	2.7
McNemar and Mantel–Haenszel tests	5	3.3
Others (Kolmogorov test, Holm's adjustment ...)	3	2

[a]For those without a "statistical analysis" section we looked at the end of the "methods" section.
[b]Other than t tests and chi-square tests.
[c]One-way ANOVA with more than two groups.

cardiology and that most studies entail repeated assessment of cardiac function and/or **quality of life**. Moreover, several general techniques of data analysis, considered important in biostatistics, are absent from Table 3. Among these are: **Generalized Estimating Equations** (GEE), for the analyses of continuous and/or discrete longitudinal outcomes; **point processes**, for the analysis of multiple failures; signal analyses, to study, for example, cardiac rhythms in a clinical context; and trees or **neural networks**, for the development of powerful predictors or decision rules. At an even more general level, the total absence of **Bayesian methods** of analysis should also be noted.

By contrast, even a cursory look at the historical development of biostatistics in the last 50 years shows that some important methodological advances have originated from the need to treat CVD data. The Framingham study [17] greatly contributed to the development of multivariate methods for binary longitudinal data; an example is the seminal paper by Truett et al. [38], where **logistic regression** was introduced and its relationship to **discriminant analysis** discussed. The association of **Cornfield** with Framingham also contributed to the development of methods for the analysis of cross-classified data (*see* **Contingency Table**) [19].

The introduction of heart transplantation as a current surgical procedure generated the need to account properly for waiting time as a predictor of survival for patients receiving a new heart. The need was met by several papers culminating in the work by Crowley & Hu [9], who modeled heart transplant as a **time-dependent covariate** and modified the **proportional hazards** model for survival data so that time-dependent covariates could be properly treated. Other advances in survival analysis have also been stimulated by CVD data: Beck [2] looked at survival models with **competing risks** to analyze heart transplant data: Senthilselvan [33] used **penalized likelihood** to obtain a **nonparametric, spline**-based estimator of the hazard function for heart transplant

data, and Wassell & Moeschberger [40] and Pickles & Crouchley [27] studied **frailty** models for the analysis of bivariate survival data.

More recently, an important application to cardiology was at the origin of the development of Classification and Regression Trees (CART) (*see* **Tree-structured Statistical Methods**). Two cardiovascular related examples were examined. In the first, a prognostic tree was developed to identify patients at high risk of short-term mortality following a heart attack (less than 30 days), using 19 noninvasive, initial 24-hour variables. In the second example, a diagnostic tree was constructed from 40 noninvasive variables to classify patients as suffering from ischemic heart disease or not.

The advances in signal and **image analysis** in cardiology have been providing new challenges to statisticians since the 1970s. Cornfield's name is also associated with early work on the statistical analysis of electrocardiograms (ECG) (see [8] and [18] for reviews). The introduction of 24-hour blood pressure monitoring has provided yet another potentially important "signal" to analyze [21]: several approaches have been proposed; see Turney et al. [39] (weighted **least square analysis of covariance**), Selwyn & Difranco [32] (Gaussian mixed model), Gaffney et al. [11] (harmonic analysis), Somes et al. [35] (Fourier analysis). As for image analysis in cardiology, Bozzini et al. [3] studied the problem of heart potential mapping and Puterman et al. [28] studied the prognostic value of echocardiography; see also [37].

Bayesian methods, despite their appeal, are not the most popular approach in biostatistics, due, in part, to lack of software. This is reflected in the area of CVDs, in spite of the central role of Cornfield, who was a Bayesian [5]. However, some recent works should be cited: Christensen & Johnson [4] analyzed heart transplant survival data by a fully Bayesian method; Sharples [34] used the Gibbs sampler (*see* **Markov Chain Monte Carlo**) to estimate the marginal posterior distribution of the transition rates between grades of coronary heart disease and from each grade to death; L'Italien et al. [20] developed and validated a Bayesian model for perioperative cardiac risk assessment; and Spiegelhalter et al. [36] discussed a Bayesian methodology for evaluating prior beliefs about frequencies, with application to a case study in congenital heart diseases. Finally, Palmas et al. [26] developed a Bayesian approach to the enhancement of scintigraphic images by integration of diagnostic information.

Beside influencing major developments in data analysis, the problems arising from CVD studies have also stimulated much of the methodological advances in the design and the conduct of clinical trials. Cornfield's name appears again in a central role: see [6] for a review or early contributions. More recently, **Halperin** et al. [14] reviewed the field again, identifying four areas: (i) organizational structure for **multicenter** studies; (ii) design considerations related to patient risk, **noncompliance**, lag in treatment effect, and changing risks (the "**intention to treat** principle" seems to have originated from these considerations); (iii) periodic reviews of accumulating data (this stimulated developments of **sequential analysis** methods); (iv) design and analysis of longitudinal studies. In each of these areas, current research continues to be very active. Examples of recent contributions to (ii) are Efron & Feldman [10] on compliance, McMahon et al. [23] on **sample size** calculations for count outcomes, and Yateman & Skene [43] on computational intensive sample size calculations in complex situations. For some recent papers in area (iii), see Hallstrom et al. [13] on sequential monitoring (*see* **Data and Safety Monitoring**). For area (iv), see Hathaway & D'Agostino [16].

The design of epidemiologic studies to assess risk factors and community intervention effectiveness has also spurred important methodological advances. A primary example is another influential work by Cornfield [7] on **group randomization**. More recent examples are: Rehm et al. [30] on omitted variable bias, Marshall & Jackson [22] on the case–crossover design, and Schouten et al. [31] on risk and rate ratios estimation in the **case–cohort** design.

Meta-analysis (the analysis of the results of several studies for the purpose of integrating them) was only recently introduced in the health field, and one of the earliest applications was in the area of CVDs (see, for example, [44]). In spite of its limitations, including the important one linked to publication bias, meta-analysis must be credited for having accelerated the creation of national and international registries of clinical trials and epidemiologic studies, published or not, such as those of the Cochrane Controlled Trials Register (*see* **Cochrane Collaboration**).

Future Perspectives

The increasing computerization of biomedical and health-related data is introducing new and unforeseen developments in the study of CVDs, as it is in other areas of human activity. Availability of data has generally driven innovation in data analytic methodology. Since CVDs constitute the most important health burden in countries where computerization is most advanced, they can be expected to provide the most important single source of problems around which new methodology will be developed. The type of data that are being collected in the course of clinical trials, epidemiologic studies and, indeed, in daily medical practice and health care delivery, will most likely determine the future directions of progress in biostatistics.

What are the general characteristics of CVD data? Perhaps the first is the abundance of subjects. We are increasingly faced with situations in which sample size is not an issue. Instead, the problem is to extract reliable information from large quantities of data, with only vague and unstructured questions as a guide. Whether studying the **prognosis** for patients with a first major cardiovascular episode, or trying to understand the role of multiple risk factors in determining the development of CVD, the analyst is faced with the problem of constructing from data a predictor that is not only accurate, but also interpretable in the light of biomedical concepts. Traditional statistical techniques, which excel in obtaining reasonable answers to extremely pointed questions, are not adequate for the task. More promising are recent techniques such as the already mentioned CART, **generalized additive modeling**, and MARS (multivariate adaptive regression splines) which go under the general term of adaptive model building. These techniques aim at constructing prediction models directly from the data by means of flexible, data-dependent strategies. They are similar to methods like induction trees and neural networks, independently developed by researchers in machine learning and other areas of **artificial intelligence** (AI). Indeed, the exchanges between statisticians and AI researchers seem to be very fruitful, as witnessed, for example, by Michie et al. [24].

A second characteristic of CVD data is the multiplicity of sources. It is increasingly common to pool data from several large studies or private practices to find common features or in an attempt to resolve controversies arising from discordant findings. To meet this challenge, the establishment of a Bayesian perspective in cardiology would be of great utility. For instance, meta-analysis would benefit from such a perspective, which would offer a general framework within which various approaches could be compared.

A third characteristic of CVD data is their increasing complexity. For example, as prognosis improves, both outcome and predictor variables will have a richer longitudinal structure, since they are collected from subjects on repeated occasions, e.g. in the course of clinical trials or prevention studies. Further development in the area of longitudinal modeling will be necessary to analyze repeated measurements of continuous and discrete variables. Similarly, the analysis of event history data will require increasingly sophisticated application of the theory of point processes. Also, as the instrumentation for monitoring heart rhythms and blood pressure over time becomes accessible at relatively low cost, data of the future will include more and more signals, i.e. continuous functions of time, one or more for each of a large number of subjects. To treat such data without oversimplification will require major developments in the new research area known as functional data analysis. Even more complex data are those that take the form of functions on two- or three-dimensional space, such as images resulting from echocardiography or magnetic resonance. The theory of random fields has been recently applied to the analysis of medical images [42]. This could be the opening of an extremely fruitful new direction for research in biostatistics with important applications in CVDs.

Clearly, in the future, we can expect to see an increasing cooperation between clinical and statistical sciences toward the solution of major problems in the area of CVDs. It is not unreasonable to hope that the result of this cooperation will be a substantial alleviation of one of the major health burdens afflicting human populations.

References

[1] American Heart Association (1996). *Heart and Strokes Facts*, and *1996 Statistical Supplement*. `http://www.armhrt.org/hs96/Hsfacts.html`.

[2] Beck, G.J. (1979). Stochastic survival models with competing risk and covariates, *Biometrics* **35**, 427–438.

[3] Bozzini, M., DeTisi, F. & Lenarduzzi, L. (1984). An approximation method of the local type. Application to

a problem of heart potential mapping, *Computing* **32**, 69–80.

[4] Christensen, R. & Johnson, W. (1988). Modeling accelerated failure time with a Dirichlet process, *Biometrika* **75**, 693–704.

[5] Cornfield, J. (1969). The Bayesian outlook and its application, *Biometrics* **25**, 617–657.

[6] Cornfield, J. (1976). Recent methodological contributions to clinical trials, *American Journal of Epidemiology* **104**, 408–421.

[7] Cornfield, J. (1978). Randomization by group: a formal analysis, *American Journal of Epidemiology* **108**, 100–102.

[8] Cornfield, J., Dunn, R.A., Batchlor, C.D. & Pipberger, H.V. (1973). Multigroup diagnosis of electrocardiograms, *Computers and Biomedical Research* **6**, 97–120.

[9] Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data, *Journal of the American Statistical Association* **72**, 27–36.

[10] Efron, B. & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials (C/R: pp. 18–26), *Journal of the American Statistical Association* **86**, 9–17.

[11] Gaffney, M., Taylor, C. & Cusenza, E. (1993). Harmonic regression analysis of the effect of drug treatment on the diurnal rhythm of blood pressure and angina, *Statistics in Medicine* **12**, 129–142.

[12] Glantz, S.A. (1980). Biostatistics: how to detect, correct and prevent errors in the medical literature, *Circulation* **61**, 1–7.

[13] Hallstrom, A., McBride, R. & Moore, R. (1995). Toward vital status sweeps: a case history in sequential monitoring, *Statistics in Medicine* **14**, 1927–1931.

[14] Halperin, M., DeMets, D.L. & Ware, J.H. (1990). Early methodological developments for clinical trials at the National Heart, Lung and Blood Institute, *Statistics in Medicine* **9**, 881–892; discussion 903–906.

[15] Haltky, M.A., Rogers, W.J., Johnstone, I., Boothroyd, D., Brooks, M.M., Pitt, B., Reeder, G., Ryan, T., Smith, H., Whitlow, P., Wiens, R. & Mark, D.B. (1997). Medical care cost and quality of life after randomization to coronary angioplasty or coronary bypass surgery, *New England Journal of Medicine* **336**, 92–99.

[16] Hathaway, D.K. & D'Agostino, R.B. (1993). A technique for summarizing longitudinal data [see comments], *Statistics in Medicine* **12**, 2169–2178.

[17] Higgins, M.W. (1984). The Framingham Heart Study: review of epidemiological design and data, limitations and prospects, in *Genetic Epidemiology of Coronary Heart Diseases; Past, Present, and Future*, Alan R. Liss, Inc, New York.

[18] Kors, J.A. & Van Bemmel, J.H. (1990). Classification methods for computerized interpretation of the electrocardiogram, *Methods of Information in Medicine* **29**, 330–336.

[19] Kullback, S. & Cornfield, J. (1976). An information theoric contingency table analysis of the Dorn study of smoking and mortality, *Computer & Biomedical Research* **9**, 409–437.

[20] L'Italien, G.J., Sumita, D.P., Robert, C.H., Jeffrey, A.L., Mylan, C.C., Lee, A.F., Kenneth, A.B., Stuart, W.Z., Richard, P.C., Bruce, S.C. & Kim, A.E. (1996). Development and validation of a Bayesian model for perioperative cardiac risk assessment in a cohort of 1081 vascular surgical candidates, *JACC* **27**, 779–786.

[21] Marler, M.R., Jacob, R.G., Rolf, G., Lehoczky, J.P. & Shapiro, A.P. (1988). The statistical analysis of treatment effect in 24-hour ambulatory blood pressure recordings, *Statistics in Medicine* **7**, 697–716.

[22] Marshall, R.J. & Jackson, R.T. (1993). Analysis of case-crossover designs, *Statistics in Medicine* **12**, 2333–2341.

[23] McMahon, R.P., Proschan, M., Geller, N.L., Stone, P.H. & Sopko, G. (1994). Sample size calculation for clinical trials in which entry criteria and outcomes are counts of events. ACIP Investigators, Asymptomatic Cardiac Ischemia Pilot, *Statistics in Medicine* **13**, 859–870.

[24] Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.

[25] National Heart, Lung and Blood Institute (1994). *Report of the Task Force on Research in Epidemiology and Prevention of Cardiovascular Diseases*. US Department of Health and Human Services, Washington.

[26] Palmas, W., Denton, T.A., Morise, A.P. & Diamond, G.A. (1994). Afterimages: integration of diagnostic information through Bayesian enhancement of scintigraphic images, *American Heart Journal* **128**, 281–287.

[27] Pickles, A. & Crouchley, R. (1995). A comparison of frailty models for multivariate survival data, *Statistics in Medicine* **14**, 1447–1461.

[28] Puterman, M.L., Schumacher, P. & Sandor, G.G.S. (1990). Optimal choice of prognostic variables with an application to cardiac monitoring using M-mode echocardiography, *Statistics in Medicine* **9**, 273–286.

[29] Ramsay, J.O. & Dalzell, C.J. (1991). Some tools for functional data analysis, *Journal of the Royal Statistical Society, Series B* **53**, 539–561.

[30] Rehm, J., Arminger, G. & Kohlmeier, L. (1992). Using follow-up data to avoid omitted variable bias: an application to cardiovascular epidemiology, *Statistics in Medicine* **11**, 1195–1208.

[31] Schouten, E.G., Dekker, J.M., Kok, F.J., LeCessie, S., VanHouwelingen, H.C., Pool, J. & Vanderbroucke, J.P. (1993). Risk ratio and rate estimation in case-cohort design: hypertension and cardiovascular mortality, *Statistics in Medicine* **12**, 1733–1745.

[32] Selwyn, M.R. & Difranco, D.M. (1993). The application of large Gaussian mixed models to the analysis of 24 hour ambulatory blood pressure monitoring data in clinical trials, *Statistics in Medicine* **12**, 1665–1682.

[33] Senthilselvan, A. (1987). Penalized likelihood estimation of hazard and intensity functions, *Journal of the Royal Statistical Society, Series B* **49**, 170–174.

[34] Sharples, L.D. (1993). Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation, *Statistics in Medicine* **12**, 1155–1169.

[35] Somes, G.W., Harshfield, G.A., Arheart, K.L. & Miller, S.T. (1994). A Fourier series approach for comparing groups of subjects on ambulatory blood pressure patterns, *Statistics in Medicine* **13**, 1201–1210.

[36] Spiegelhalter, D.J., Harris, N.L., Bull, K. & Franklin, R.C.G. (1994). Empirical evaluation of prior belief about frequencies: methodology and a case study in congenital heart disease, *Journal of the American Statistical Association* **89**, 435–443.

[37] Storvik, G. & Switzer, P. (1992). Space-time modeling of simple connected objects: an application to detection of left ventricular cardiac boundaries from ultrasound images, in *Computer Science Statistics: Proceedings of the Twenty-Fourth Symposium of Interface*, H. Joseph Newton, ed. Interface Foundation of North America, Fairfax Station.

[38] Truett, J., Cornfield, J. & Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham, *Journal of Chronic Diseases* **20**, 511–524.

[39] Turney, E.A., Amara, I.A., Koch, G.G. & Stewart, W.H. (1992). Evaluation of alternative statistical methods for linear model analysis to compare two treatments for 24-hour blood pressure response [see comments], *Statistics in Medicine* **11**, 1843–1860.

[40] Wassell, J.T. & Moeschberger, M.L. (1993). A bivariate model with modified gamma frailty for assessing the impact of intervention, *Statistics in Medicine* **12**, 241–248.

[41] WHO (1991, 1992, 1993, 1994). *World Health Statistics Annual*. WHO, Geneva.

[42] Worsley, K.J. (1995). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images, *Annals of Statistics* **23**, 640–669.

[43] Yateman, N.A. & Skene, A.M. (1993). The use of simulation in the design of two cardiovascular survival studies, *Statistics in Medicine* **12**, 1365–1372.

[44] Yusuf, S., Peto, R., Lewis, J., Collins, R. & Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials, *Progress in Cardiovascular Diseases* **27**, 335–371.

ANTONIO CIAMPI & ANDRE COUTURIER

Case Fatality

The concept of case fatality refers to patients with a common defined index disease or other medical problem, not to healthy people. Case fatality indicates how serious a disease condition is in causing death to the patients, usually within a defined period of time. It is common to hear about case fatality without reference to the time period of follow-up of the patients, but this should be avoided for reasons of ambiguity. There can, nevertheless, be applications in which the follow-up time may be virtually zero as, for example, with heart attacks or automobile accidents.

Technically, case fatality is expressed as the proportion of the number of patients dying in the follow-up interval out of all patients under observation. This concept is useful only under a fairly complete follow-up, where the proportion of persons lost to follow-up or otherwise withdrawn alive is small. Moreover, **competing risks** of death can, in addition to the index disease, cause deaths among the patients. If the follow-up period is short, deaths due to competing risks unrelated to the index disease may be uncommon, and the case fatality indeed reflects the seriousness of the disease in an adequate way.

Conceptually, case fatality may be seen as a complement to survival. Thus, the methods of survival analysis can be employed in assessing case fatality. For example, the proportion of survivors after a one-year follow-up among patients diagnosed in Finland in 1967–1974 with cancer of the tongue was 64%. The case fatality within the first year was thus 36%.

The models in survival analysis generally are based on assumptions concerning the risk of dying for the patients. Thus, it would also be natural to express case fatality in terms of fatality or lethality rate of the disease by defining any death or death due to index disease as the main outcome event in survival analysis. This rate is the incidence of death or death due to disease and as a rate is calculated as the number of outcome events in the follow-up period divided by the appropriate person-time denominator (*see* **Person-years at Risk**). Although this may sound theoretically appealing, a conversion to a proportion-type measure produces a quantity with an easier numerical interpretation for clinical medicine.

T. HAKULINEN

Case Mix

Case mix refers to the characteristics of the patients and/or the medical problems treated by a *provider*, a term we will use generically to refer to any of the following: an individual clinician, a health care delivery "team" (such as a hospital or clinic or group of hospitals or clinics), or an entire health care delivery system (such as an HMO or a statewide program of subsidized care).

Utilization rates reflect case mix. For example, hernias do not require the same resources as heart attacks, and even among heart attack patients, need may substantially depend upon patient age, other medical problems (comorbidities) present, and the severity of the heart attack itself (disease-specific severity).

An informal sense of case mix is conveyed via distributions of patient characteristics or summary statistics (such as percentage of cases with diabetes, mean and standard deviation of patient age).

Some case-mix classification systems assign each case (e.g. an individual hospital admission, or a person enrolled in a health care delivery system) to one and only one category, which is relatively homogeneous with respect to the expected level of health care need. To facilitate comparisons of case mix across groups of cases (e.g. admissions occurring at distinct hospitals, or the patient panels of different providers), each category may be assigned a "weight" indicating the expected utilization of these cases in comparison with an average case. Thus, the weight 1.00 is used for average cases, while a weight of 1.10 indicates 10% higher expected utilization. Such a classification of hospital admissions into **diagnosis related groups (DRGs)** is used to calculate the payment for an individual **Medicare** hospital admission, proportional to its DRG weight. (*see* **Health Care Financing**) Some authors equate a hospital's case mix with its average DRG weight.

Differences in case-mix can be large and are addressed using **risk adjustment**.

ARLENE S. ASH

Case Series, Case Reports

Case reports are used by clinicians to describe responses to treatment, among other things. Epidemiologists may rely on case reports to find clues to disease etiology. Case reports can be very informative for rare diseases. Such reports identified chimney sweeping as a risk factor for scrotal cancer [1] in 1775, and many modes of transmission of the human immunodeficiency virus were identified in the early 1980s from case reports of the Acquired Immune Deficiency Syndrome (**AIDS**). For more reliable **inferences**, however, it is usually necessary to compare rates of exposure in cases with rates of exposure in disease-free controls to develop etiologic evidence (*see* **Case–Control Study**).

Reference

[1] Potts, P. (1775). Cancer scroti, in *Chirurgical Observations*. Hawes, Clarke & Collins, London, pp. 63–68.

MITCHELL H. GAIL

Case–Cohort Study

The case–cohort design is a method of sampling from an assembled epidemiologic **cohort study** or **clinical trial** in which a **random sample** of the cohort, called the *subcohort*, is used as a comparison group for all cases that occur in the cohort. This design is generally used when such a cohort can be followed for disease outcomes but it is too expensive to collect and process **covariate** information on all study subjects. Though it may be used in other settings, it is especially advantageous for studies in which covariate information collected at entry to the study is "banked" for the entire cohort but is expensive to retrieve or process (see examples below) and multiple disease stages or outcomes are of interest. In such circumstances, the work of covariate processing for subcohort members can proceed at the beginning of the study. As time passes and cases of disease occur, information for these cases can be processed in batches. Since the subcohort data are prepared early on and are not dependent on the occurrence of cases, statistical analyses can proceed at regular intervals after the processing of the cases. Furthermore, staffing needs are quite predictable. Motivated by the case–base sampling method for simple **binary** outcome data [15, 23], Prentice described the design and a **pseudo-likelihood** method of analysis (see below) for the case–cohort design.

Design

The basic components of a case–cohort study are the *subcohort*, a sample of subjects in the cohort, and *nonsubcohort cases*, subjects that have had an event and are not included in the subcohort. The subcohort provides information on the **person-time** experience of a random sample of subjects from the cohort or random samples from within strata (*see* **Stratification**) of a **confounding** factor. In the latter situation, differing sampling fractions could be used to align better the person-time distribution of the subcohort with that of the cases. Methods for sampling the subcohort include sampling a fixed number without replacement [26] (*see* **Sampling With and Without Replacement**) and sampling based on independent Bernoulli "coin flips" [34] (*see* **Binomial Distribution**). The latter may be advantageous when subjects are entered into the study prospectively; the subcohort may then be formed concurrently rather than waiting until accrual into the cohort has ended [30, 34]. Simple case–cohort studies are the same as case–base studies for simple **binary** outcome data. But, in general, portions of a subject's time on study might be sampled. For example, the subcohort might be "refreshed" by sampling from those remaining on study after a period of time [26, 36]. These subjects would contribute person-time only from that time forward. While the subcohort may be selected based on covariates, a key feature of the case–cohort design is that the subcohort is chosen without regard to failure status; methods that rely on failure status in the sampling of the comparison group are **case–control studies**.

Examples

Study of Lung Cancer Mortality in Aluminum Production Workers in Quebec, Canada. Armstrong et al. [1] describe the results of a case–cohort study selected from among 16 297 men who had worked at least one year in manual jobs at a large aluminum production plant between 1950 and 1988. This study greatly expands on an earlier cohort mortality study of the plant, which found a suggestion of increased rates of lung cancer in jobs with high exposures to coal tar pitch [12]. Through a variety of methods, 338 lung cancer deaths were identified. To avoid the expense associated with tracing subjects and abstraction of work records for the entire cohort, a case–cohort study was undertaken. To improve study efficiency a subcohort of 1138 subjects was randomly sampled from within year-of-birth strata with sampling fractions varying to yield a similar distribution to that of cases. This was accommodated in the analysis by stratification by these year-of-birth categories. The random sampling of subcohort members resulted in the inclusion of 205 cases in the subcohort. Work and smoking histories were abstracted for the subcohort and the additional 133 nonsubcohort cases. Cumulative exposure to coal tar pitch volatiles was estimated by linking worker job histories to measurements of chemical levels made in the plant using a **"job-exposure matrix"**. The analyses confirmed the lung cancer–coal pitch **association** observed in the earlier study and effectively ruled out confounding by smoking.

Women's Health Trial. To assess the potential health benefits of a low fat diet, a randomized trial of women assigned to low fat intervention and control groups has been undertaken. Of particular interest is the effect of this intervention on the risk of breast cancer. The study, as described in Self et al. [30], includes a cohort of 32 000 women between ages 45 and 69 whose percent calories from fat is greater than the **median** and who have at least one of a list of known risk factors for breast cancer. The study will involve 20 clinics across the US for a period of 10 years of follow-up. At two-year intervals, each participant will fill out four-day food records and food frequency questionnaires and blood will be drawn and stored. While evaluation of the intervention will be based on the full cohort, questions that require abstraction and coding of the questionnaires and blood lipid analyses are being addressed in a case–cohort study with a 10% sample serving as the subcohort. It was calculated that, relative to the entire cohort, this sample avoids about 80% of the cost of the analyses requiring these data with only a modest reduction of efficiency. The subcohort can also be used for making other comparisons between intervention and control groups. For example, the case–cohort sample could be used to investigate the joint relationship of blood hormone and nutrient levels and dietary intakes to breast cancer risk. Also, questions relating to other outcomes, such as cardiovascular disease, could be explored using the same subcohort as the comparison group, although additional data processing would be required for cases that occur outside the subcohort.

Statistical Analysis

Several methods have been developed to analyze case–cohort samples. Essentially, each of the methods available for the analysis of complete cohort data has an analog for the case–cohort sample. For point **estimation** of rate ratio parameters, the **likelihood** for full cohort data applied to the case–cohort data yields a valid estimator. However, estimation of the **variance** of point estimates, or tests of hypotheses (*see* **Hypothesis Testing**), requires adjustment to the standard full cohort variance estimators, as these will be too small. For likelihood-based methods, case–cohort sampling induces a covariance between score terms so that the variance of the **score** is given by $\Sigma + \Delta$, where Σ is the full cohort score variance and Δ is the sum of the covariances between the score terms. Since the subgroup used to compute the score terms has less variability than the full cohort, this covariance is positive. This leads to a larger variance for the parameter estimates, taking into account the subcohort sampling variability [19, 26, 29, 36]. Estimation of **absolute rates or risk** requires incorporation of the subcohort sampling fraction (or fractions) into the estimator.

Pseudo-likelihoods for Proportional Hazards Models

Assume the underlying model for disease rates has a **multiplicative** form:

$$\lambda[t, z(t); \beta_0] = \lambda_0(t) r[z(t); \beta_0],$$

where $r[z(t); \beta_0]$ is the rate ratio of disease for an individual with covariates $z(t)$ at time t and $r(0; \beta) = 1$, so $\lambda_0(t)$ is the rate of disease in subjects with $z = 0$. The pseudo-likelihood approach described by Prentice [26] parallels the **partial likelihood** approach to the analysis of full cohort data. We start with the full cohort situation and then return to the analysis of the case–cohort sample. The partial likelihood approach is illustrated in Figure 1 for a small hypothetical **cohort study** of 15 subjects. Each horizontal line represents one subject. A subject enters the study at some *entry time*, is *at risk*, denoted by the horizontal line, over some time period, and exits the study at some *exit time*. A subject may contract or die from the disease of interest, and thus be a *failure* (represented by "•" in Figure 1 or be **censored**, i.e. be alive at the end of the study, died never having had the disease of interest, or be lost to follow-up. At each failure time a *risk set* is formed that includes the *case*; namely, the failure at that failure time, and all *controls*, namely, any other cohort members who are at risk at the failure time (these are denoted by a "|" in Figure 1). The partial likelihood for full cohort data is based on the **conditional probabilities** that the case failed given that one of the subjects in the risk set failed at that time. With r_k the rate ratio and Y_k the "at risk" indicator for subject k at the failure time, and r_{case} the rate ratio associated with the case, the full cohort partial likelihood is

$$\prod_{\text{failure times}} \frac{r_{\text{case}}}{\sum_{\text{case and all controls}} Y_k r_k}.$$

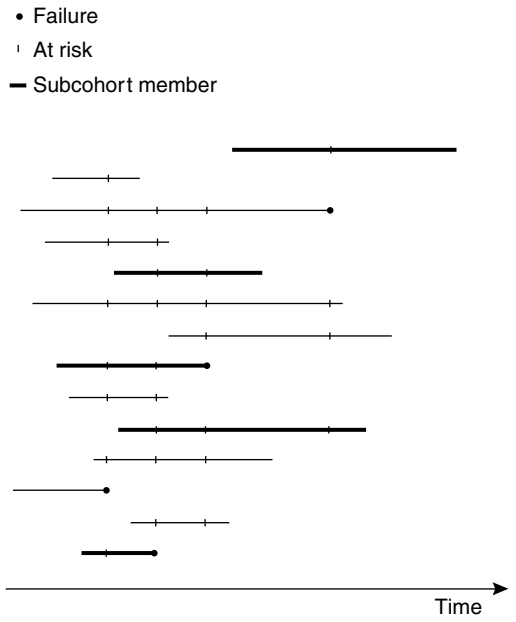

Figure 1 Prentice pseudo-likelihood approach to the analysis of case–cohort data. Pseudo-likelihood contributions are conditional probabilities based on the case and the subcohort members at risk at the failure time

Now in a case–cohort sample, covariate information is obtained for the subcohort and all nonsubcohort failures and only these subjects can contribute to the analysis. Prentice's pseudo-likelihood approach is illustrated in Figure 1 in which subcohort members are denoted by a thick horizontal line. For each failure, a *sampled risk set* is formed by the case and the controls who are in the subcohort (those with thick lines and a "|" at the failure time). As the figure indicates, subcohort members contribute to the analysis over their entire time on study, but the nonsubcohort failures contribute only at their failure times. Analogous to the full cohort partial likelihood, a pseudo-likelihood contribution is based on the conditional probability that the case fails given that someone fails among those in the sampled risk set. The pseudo-likelihood is then the product of such conditional probabilities over failure times:

$$\prod_{\text{failure times}} \frac{r_{\text{case}}}{\sum_{\text{case and subcohort controls}} Y_k r_k}, \qquad (1)$$

where the sum in the denominator is over the subcohort members when the case is in the subcohort and over the subcohort and nonsubcohort case when the case is not in the subcohort. This "likelihood" has the property that the expected value of the score is zero at β_0 but, as discussed above, the inverse information does not estimate the variance of the maximum pseudo-likelihood estimator. Prentice provided an estimator of the covariance Δ from the covariance between each pair of score terms, conditional on whether or not the failure occurring later in time was in the subcohort [17]. This is a rather complicated expression and only one software package has implemented it (Epicure, Hirosoft International Corp., Seattle, WA). Development of other methods of variance estimation has been an area of much research. These include "large sample" [29], **bootstrap** [37], "empirical" [9, 28], and influence function based [2, 21] methods. Simpler alternatives are the "asymptotic" [29] and the "robust" estimators [2, 21, 22]. Either may be computed by the simple manipulation of *delta beta* diagnostic statistics, which are an output option in many software packages [31]. The asymptotic estimator requires the sampling fractions, while the robust version estimates these from the data. Other methods of variance estimation have been proposed [9, 28, 37].

Absolute Risk Estimation

Estimation of the cumulative baseline hazard and related quantities parallel the nonparametric estimators based on the **Nelson–Aalen estimator** for full cohort data. Since the subcohort is a random sample from the full cohort, a natural estimator of the cumulative baseline hazard $\int_0^t \lambda_0(u)\,\mathrm{d}u$ is given by summing contributions for failure times up to t of the form:

$$\frac{1}{1/f \sum_{\text{subcohort}} r_k(\hat{\beta})},$$

where f is the proportion of the cohort in the subcohort [26, 29]. Again, adjustment of the cohort variance estimator is required. Cause-specific baseline hazard estimates for multiple outcomes have also been developed [25].

Other estimation methods and further developments

Alternative pseudo-likelihoods for the estimation of rate ratios of a similar form to (1) have been proposed. These involve differential weightings of the

r_k terms on the basis of the sampling fractions of those associated with subject k [2, 13, 29]. Iterative mean score methods have been proposed, which may yield more efficient estimators than (1) [8]. A method for analysis of generalized case–cohort sampling imputes rate ratio values for each cohort member using a "local averaging". Theoretically, this method is shown to have a superior efficiency to other methods, with methods for making the estimator optimal. Further research is needed to ascertain whether the increases are of practical importance. Methods for estimating standardized mortality ratios (*see* **Standardization Methods**) with a case–cohort sample have been described [35]. These involve "boosting up" the subcohort person-time in each age-year-exposure group "cell" by the inverse sampling fraction. Methods of variance adjustment are also discussed. When disease is rare and there is little censoring, methods of analyses for case-base studies with simple binary outcome data [23] will approximate the failure time analyses (e.g. [11, 17, 27, 33]). When exposure (or treatment) information is available on cohort members and additional information is to be collected for the case–cohort sample, an exposure-stratified subcohort may offer substantial efficiency advantages over random sampling. The analysis of this design uses a weighted variation of the pseudo-likelihood (1) and a generalization of the asymptotic variance estimator has been described [4].

Asymptotic Properties and Efficiency

Self & Prentice [29] give conditions for the **consistency** and asymptotic **normality** of the Prentice pseudo-likelihood for simple (stratified) case–cohort sampling. They show that the asymptotic variance of the maximum pseudo-likelihood estimator of relative risk parameters has the form $\Sigma^{-1} + \Sigma^{-1}\Delta\Sigma^{-1}$, where Σ is the full cohort variance of the score, and they provide a formula for the asymptotic sampling-induced covariance Δ. This covariance depends on the censoring distribution even when $\beta_0 = 0$, so that efficiencies relative to the full cohort analysis must take the censoring distribution into account. Assuming a cohort with complete follow-up over a fixed observation period, an **exponential** relative risk model for a single binary covariate, a subcohort that is a simple $100\alpha\%$ random sample of the cohort, and probability of failure during the observation period of d, they calculate the **asymptotic relative efficiency** as

$$\left\{1 + 2\frac{1-\alpha}{\alpha}\left[1 + \frac{1-d}{d}\log(1-d)\right]\right\}^{-1}.$$

A number of papers have derived the asymptotic variance and semiparametric efficiency bounds for the case–cohort design [7, 8, 37, 38]. These indicate that, although the pseudo-likelihood (1) is not generally semiparametric efficient, the potential loss of efficiency appears to be small, unless disease is common or the size of the subcohort is much smaller than the number of cases.

Comparison with Nested Case–Control Sampling

Nested case–control and case–cohort methods are the two main approaches to sampling from assembled cohort studies. The former takes a retrospective point of view by sampling time-matched controls after the outcome (failure) occurs. In contrast, case–cohort sampling is prospective and unmatched in the sense that the comparison group, the subcohort, is picked without regard to failure status. Considerations for choosing between the designs have been the subject of some interest [10, 18, 20, 24, 26, 32, 34]. We summarize some of these considerations below.

Prospective Studies

If the study is **retrospective** and has been assembled, the major consideration in choosing between sampling designs is the statistical efficiency for the proposed analyses and the information to be collected on the sample, as this will translate quite directly into cost. If the study is prospective in that the study group will be assembled as time passes and outcomes occur in the future, the decision about which design to choose will depend on whether it is advantageous to have a comparison group early on in the study or whether it is better to wait until near the end. If the sample is to be chosen at the beginning, or concurrent with accrual into a prospective study, the case–cohort study has a number of advantages. First, as discussed above, processing of covariate information for the subcohort may proceed early on in the study during the accrual period. During the follow-up period, data for cases arising outside the

subcohort could be processed in batches at various times. A nested case–control study requires waiting until cases occur and controls are selected for them, delaying the processing of covariate information until later in the study than would be required in the case–cohort design. Thus, the case–cohort study can potentially be completed sooner than the nested case–control study. Secondly, although subcohort members should not be treated differently from cases occurring outside the subcohort, the subcohort can serve as a sample for assessing compliance, or quality control, as the study proceeds. However, a nested case–control sample may be advantageous if it is important that processing of information be "blinded" (*see* **Blinding or Masking**) to case-comparison group status. Since case–control covariate information can be processed simultaneously, potential information bias can be avoided. This is not always possible with a case–cohort sample when subcohort data are processed early in the study.

Statistical Efficiency

Comparison of statistical efficiency for studying a single outcome has been a topic of much research. It has been conjectured that the case–cohort design should be more efficient than the nested case–control design. This belief has been based on a comparison of the contribution of a failure to the pseudolikelihood (1) with that of the corresponding nested case–control contribution. The former uses all subcohort members at risk at the failure time, whereas the latter uses only the controls selected for that case, usually resulting in the case–cohort having many more "controls per case". In fact, analytic and empirical efficiency comparisons indicate that in most situations encountered in practice, nested case–control sampling will be more efficient than the case–cohort, although often not by a large amount [18, 19, 34, 36]. The reason for the lower-than-anticipated relative efficiency is that the large number of controls per case in the case–cohort sample, which by itself increases efficiency, is offset by the sampling-induced positive **correlation** in score terms (see above), which lowers efficiency. The nested case–control design has relatively few controls per case, but there is no sampling-induced correlation between score terms [19].

Multiple Disease Outcomes

Since the subcohort is chosen without regard to failure status, it may serve as the comparison group for multiple disease outcomes. This would seem to be a great advantage over the nested case–control design, since controls are selected for specific cases. In fact, there are few published studies that exploit this feature of the design. Nevertheless, the most cost-effective use of the case–cohort design would seem to be to study a single set of explanatory factors and multiple outcomes. Thus, it seems likely that the case–cohort design may have application in clinical investigations in which researchers are often interested in multiple-event outcomes such as relapse, local and distant recurrence, and death as a function of a single set of treatment and **prognostic factors**. If, for instance, the prognostic factors involve expensive laboratory work, a case–cohort sample would be a natural way to reduce costs associated with the laboratory work, but still allow a full analysis of multiple endpoints. Using the same comparison group will result in correlation between estimates of the same parameter for different endpoints. Appropriate methods for the variance adjustment and hypothesis testing with multiple outcomes have been developed [25].

Matching

Often, it is desirable to match (*see* **Matching**) comparison subjects closely on certain factors, either to control for **confounding** or so that information of comparable quality may be obtained. For instance, it is common to compare a case with controls close in year of birth to adjust for secular trends in behavior. Fine matching, and matching based on time-dependent factors is accommodated in a natural way in a nested case–control sample. Matching may only be done crudely for case–cohort sampling and must be based on factors available at the time the subcohort is sampled.

Analysis Flexibility

The nested case–control design is inherently associated with methods for analysis of cohort data based on **semiparametric proportional hazards** models. Estimation of rate ratio parameters is based on partial likelihood methods and estimation of absolute-risk-related quantities is based on the Nelson–Aalen

estimator of the **cumulative hazard**. (One interesting exception to the restriction to proportional hazards models is estimation of excess risks using the **Aalen linear model** [3].) The case–cohort design is not associated with any particular model or method of analysis. Thus, in theory, "Poisson likelihood" or "grouped time" case–base analysis approaches, as well as the risk-set-based pseudo-likelihood (1), may be used for parameter estimation. Examples of estimation of parameters in nonproportional hazards models from case–cohort data include the additive hazards, proportional odds, and transformation regression models [5, 6, 14]. For a subcohort that is a simple random sample, changing time scales and analysis stratification variables poses no difficulties in the analysis. Since the nested case–control sample is bound to the risk set defined by the time scale and stratification variables used in matching controls to failures, these must be fixed in the analysis. However, **inference** from case–cohort samples is complicated by the need to adjust **standard errors** and test statistics for the sampling-induced covariance. Further, for testing, adjusted Wald and score tests are adapted in a natural way using the variance estimator [31], but a "pseudo-likelihood ratio test" is not available.

Computation

Standard conditional logistic regression software, for the analysis of matched case–control data, may be used to analyze rate ratio parameters from nested case–control studies (*see* **Software, Biostatistical**). Furthermore, if the numbers of subjects in the risk sets are known, absolute risk estimators and standard errors are relatively simple to compute [16]. Since the latter are based on standard nonparametric cumulative hazard and survival estimators, standard software for the analysis of full cohort data may be "tricked" into computing the nested case–control estimators. For case–cohort samples, standard **Cox regression** software may be used to estimate parameters but, as discussed above, special software is needed to estimate corresponding variances.

References

[1] Armstrong, B., Tremblay, C., Baris, D. & Gilles, T. (1994). Lung cancer mortality and polynuclear aromatic hydrocarbons: a case–cohort study of aluminum production workers in Arvida, Quebec, Canada, *American Journal of Epidemiology* **139**, 250–262.

[2] Barlow, W.E. (1994). Robust variance estimation for the case–cohort design, *Biometrics* **50**, 1064–1072.

[3] Borgan, O. & Langholz, B. (1997). Estimation of excess risk from case–control data using Aalen's linear regression model, *Biometrics* **53**, 690–697.

[4] Borgan, O., Langholz, B., Samuelsen, S.O., Goldstein, L. & Pogoda, J. 2000. Exposure stratified case-cohort designs, *Lifetime Data Analysis* **6**, 39–58.

[5] Chen, H.Y. 2001. Fitting semiparametric transformation regression models to data from a modified case-cohort design, *Biometrika* **88**(1), 255–268.

[6] Chen, H.Y. 2001. Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design, *Journal of the American Statistical Association* **96**(456), 1446–1457.

[7] Chen, Kani 2001. Generalized case-cohort sampling, *Journal of the Royal Statistical Society, Series B, Methodological* **63**(4), 791–809.

[8] Chen, K. & Lo, S-H. 1999. Case-cohort and case-control analysis with Cox's model, *Biometrika* **86**(4), 755–764.

[9] Edwardes, M.D. (1995). Re: Risk ratio and rate ratio estimation in case–cohort designs: hypertension and cardiovascular mortality, *Statistics in Medicine* **14**, 1609–1610.

[10] Ernster, V.L. (1994). Nested case–control studies, *Preventive Medicine* **23**, 587–590.

[11] Flanders, W.D., Dersimonian, R. & Rhodes, P. (1990). Estimation of risk ratios in case–base studies with competing risks, *Statistics in Medicine* **9**, 423–435.

[12] Gibbs, G.W. (1985). Mortality of aluminum reduction plant workers, 1950 through 1977, *Journal of Occupational Medicine* **27**, 761–770.

[13] Kalbfleisch, J.D. & Lawless, J.F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality, *Statistics in Medicine* **7**, 149–160.

[14] Kulich, M. & Lin,D.Y. 2000. Additive hazards regression for case-cohort studies, *Biometrika* **87**, 73–87.

[15] Kupper, L.L., McMichael, A.J. & Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk, *Journal of the American Statistical Association* **70**, 524–528.

[16] Langholz, B. & Borgan, Ø. (1997). Estimation of absolute risk from nested case–control data, *Biometrics* **53**, 767–774.

[17] Langholz, B. & Goldstein, L. 2001. Conditional logistic analysis of case-control studies with complex sampling, *Biostatistics*, **2**, 63–84.

[18] Langholz, B. & Thomas, D.C. (1990). Nested case–control and case–cohort methods of sampling from a cohort: a critical comparison, *American Journal of Epidemiology* **131**, 169–176.

[19] Langholz, B. & Thomas, D.C. (1991). Efficiency of cohort sampling designs: some surprising results, *Biometrics* **47**, 1563–1571.

[20] Langholz, B., Thomas, D.C., Witte, J.S. & Peters, R.K. (1995). Re: Thompson et al. a population based case–cohort evaluation of the efficacy of mammography screening for breast cancer, *American Journal of Epidemiology* **142**, 448–449.

[21] Lin, D.Y. & Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association* **88**, 1341–1349.

[22] Mark, S.D. & Katki, H. 2001. Influence function based variance estimation and missing data issues in case-cohort studies, *Lifetime Data Analysis* **7**(4), 331–344.

[23] Miettinen, O.S. (1982). Design options in epidemiology research: an update, *Scandinavian Journal of Work, Environment, and Health* **8**(Supplement 1), 1295–1311.

[24] Moulton, L.H., Wolff, M.C., Brenneman, G. & Santosham, M. (1995). Case–cohort analysis of case-coverage studies of vaccine effectiveness, *American Journal of Epidemiology* **142**, 1000–1006.

[25] Sorensen Per & Andersen, Per Kragh 2000. Competing risks analysis of the case-cohort design, *Biometrika* **87**(1), 49–59.

[26] Prentice, R.L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[27] Sato, T. (1992). Maximum likelihood estimation of the risk ratio in case–cohort studies, *Biometrics* **48**, 1215–1221.

[28] Schouten, E.G., Dekker, J.M., Kok, F.J., Le Cessie, S., van Houwelingen, H.C., Pool, J. & Vandenbrouke, J.P. (1993). Risk ratio and rate ratio estimation in case–cohort designs: hypertension and cardiovascular mortality, *Statistics in Medicine* **12**, 1733–1745.

[29] Self, S.G. & Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case–cohort studies, *Annals of Statistics* **16**, 64–81.

[30] Self, S., Prentice, R., Iverson, D., Henderson, M., Thompson, D., Byar, D., Insull, W., Gorbach, S.L., Clifford, C., Goldman, S., Urban, N., Sheppard, L. & Greenwald, P. (1988). Statistical design of the women's health trial, *Controlled Clinical Trials* **9**, 119–136.

[31] Therneau, T.M. & Li, H. 1999. Computing the Cox model for case cohort designs, *Lifetime Data Analysis* **5**, 99–112.

[32] Thompson, R.S., Barlow, W.E., Taplin, S.H., Grothaus, L., Immanuel, V., Salazar, A. & Wagner, E.H. (1994). A population-based case–cohort evaluation of the efficacy of mammographic screening for breast cancer, *American Journal of Epidemiology* **140**, 889–901.

[33] van den Brandt, P.A., Goldbohm, R.A. & van't Veer, P. (1995). Alcohol and breast cancer: results from the Netherlands cohort study, *American Journal of Epidemiology* **141**, 907–915.

[34] Wacholder, S. (1991). Practical considerations in choosing between the case–cohort and nested case–control designs, *Epidemiology* **2**, 155–158.

[35] Wacholder, S. & Boivin, J.-F. (1987). External comparisons with the case–cohort design, *American Journal of Epidemiology* **126**, 1198–1209.

[36] Wacholder, S., Gail, M.H. & Pee, D. (1991). Selecting an efficient design for assessing exposure–disease relationships in an assembled cohort, *Biometrics* **47**, 63–76.

[37] Wacholder, S., Gail, M.H., Pee, D. & Brookmeyer, R. (1989). Alternative variance and efficiency calculations for the case–cohort design, *Biometrika* **76**, 117–123.

[38] Zhang, H. & Goldstein, L. 2002. Information and asymptotic efficiency of the case-cohort sampling design in Cox's regression model, *Journal of Multivariate Analysis* in press.

BRYAN LANGHOLZ

Case–Control Study, Hospital-based

A hospital-based case–control study is a **case–control study** in which cases with a given disease are selected from persons with that disease in a given hospital or group of hospitals, and **controls** are patients with other diseases from those hospitals. Hospital-based case–control studies are relatively convenient to conduct and offer some advantages, compared with **population-based case–control studies**. First, a higher proportion of persons invited to join a hospital-based case–control study may agree to participate, especially when biologic samples such as blood specimens are required. This reduces the chance for one type of **selection bias** known as **nonresponse** bias. Secondly, cases and controls with other diseases may provide a similar quality of information when asked about previous exposures, thus reducing the chance of **recall bias** compared with **population-based studies** in which most controls are healthy.

There are two serious sources of **bias** that may affect hospital-based case–control studies and that are not present in population-based case–control studies. First, the pattern of referral of cases to a hospital may differ from the pattern of referral for persons with control diseases, resulting in selection bias because the controls are not representative of the source population from which the cases arise. Secondly, the exposure under study may affect the risk of the control conditions, causing a distortion of the **relative risk**.

(*See also* **Bias in Case–Control Studies**; **Bias in Observational Studies**; **Bias, Overview**)

MITCHELL H. GAIL

Case–Control Study, Nested

A nested **case–control** study is comprised of subjects sampled from an assembled epidemiological **cohort study** in which the sampling depends on disease status. Nested case–control studies are generally used when disease is rare and, at the minimum, disease outcome has been obtained for all cohort subjects, but it is too expensive to collect and/or process information on **covariates** of interest for the entire cohort. By sampling a small proportion of the nondiseased subjects, there is high cost efficiency for assessing associations between exposures and disease. "Standard" case–control studies, the most common study design in epidemiologic research, may often be viewed as nested case–control studies in which a portion of underlying cohort (usually among the nondiseased) has not been identified [15]. The distinction between standard and nested case–control studies is often ambiguous and, in fact, analysis methods appropriate to standard case–control studies are directly applicable to nested case–control studies. However, depending on the amount of information available in the assembled cohort, there may be a much wider range of design and analysis options for nested case–control studies than for a standard case–control study. So, **confounder** information available in the cohort data is often used to select controls that closely match cases. Also, unlike standard case–control studies, **absolute risk** may often be reliably estimated. Further, it is often possible to compare characteristics of participants to nonparticipants to assess the potential magnitude of selection or information bias (*see* **Bias in Case–Control Studies**). The advantages of nesting a case–control study in a cohort include convenience, cost-efficiency, high validity, and analytic flexibility, for example, [15, 16, 21, 30, 32, 35, 42, 59]. Methodologically, the paradigm of nested case–control sampling is *prospective*, with disease outcome random with probability dependent on covariates. In contrast, the paradigm for standard case–control studies is *retrospective*, with covariates random with distribution depending on disease status. To the extent that standard case–control studies can be viewed as having been sampled from a (perhaps poorly defined) cohort, nested case–control design and analysis developments apply to case–control studies generally.

Data Model for Nested Case–Control Studies Based on Risk Sets

Cohort data arises by observing a population for disease occurrence over some period of time. So, it is natural to represent nested case–control studies in relation to cohort generation. Figure 1 represents the basic features of a small hypothetical cohort study of 14 subjects. Each subject enters the study at some *entry time*, is *at risk*, denoted by the horizontal line, over some time period, and exits the study at some *exit time*. A subject may contract or die from the disease of interest, and thus be a *failure* (represented by "•" in Figure 1) or be *censored*, that is be alive at the end of the study, died never having had the disease of interest, or be lost to follow-up.

The link to nested case–control studies is in the organization of the cohort data into **risk sets** [19]. At any time, the *risk set* is defined to be all subjects under observation. Risk sets may be defined at single points in time, *continuous time risk sets* as in Figure 2 or in time intervals, *grouped time risk sets* as in Figure 3. Risk set members are identified by the "|" at the given time (or time interval). Continuous and grouped risk sets have the structure of individually matched (*see* **Matching**) or unmatched case–control sets, respectively. *Cases* in the risk set are failures at the failure time or time interval, while *controls* are the nonfailures. The nested case–control sample is drawn by sampling from the controls (and possibly from the cases) in the risk sets. Individually matched nested case–control studies arise by sampling from continuous time risk sets at the failure times, while unmatched nested case–control studies arise from sampling from the grouped time risk sets. These are illustrated in Figures 2 and 3, in which ○ represent sampled controls.

Examples

Occupational Cohort Study of TCDD Exposure and STS and NHL. **The International Agency for the Research of Cancer** (IARC) maintains an international register of 21 183 workers exposed to phenoxy herbicides, chlorophenols, and dioxins [52]. In

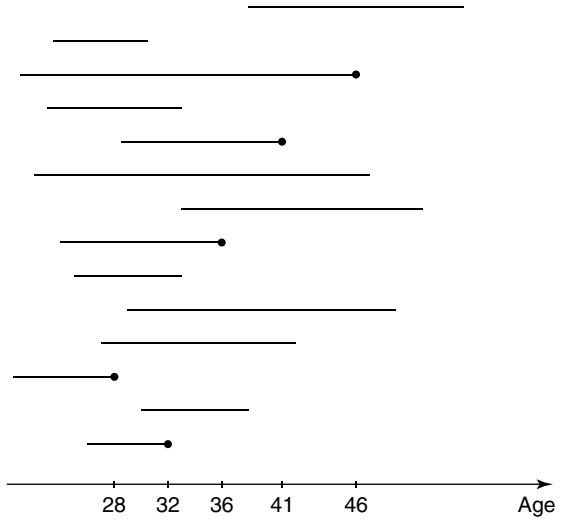

Figure 1 Cohort of 14 subjects. Each line represents the time on study for one subject. Subjects can either fail (represented by the •) or be censored (no •)

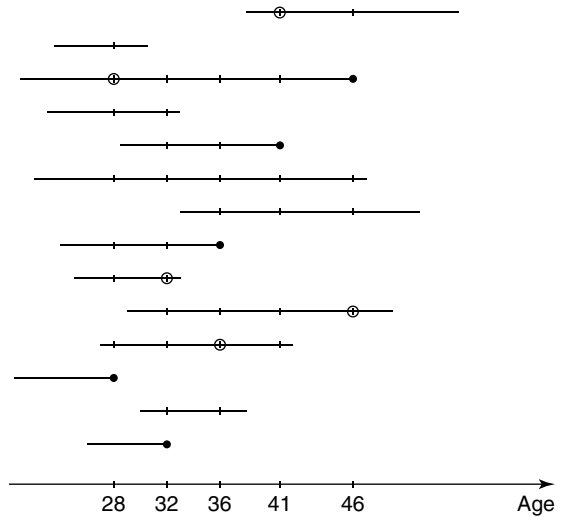

Figure 2 Continuous time risk sets at each failure time are represented by the "|" marks. The failure is the case in the risk set and the nonfailures are the controls. Single controls, sampled for each case are represented by the ○

a cohort mortality study analysis, standardized mortality ratios (SMRs) (*see*, **Standardization Methods**) of 1.96 and 1.29 were found for soft-tissue sarcomas (STS) and non-Hodgkin's lymphoma (NHL), respectively, comparing exposed to unexposed workers. In

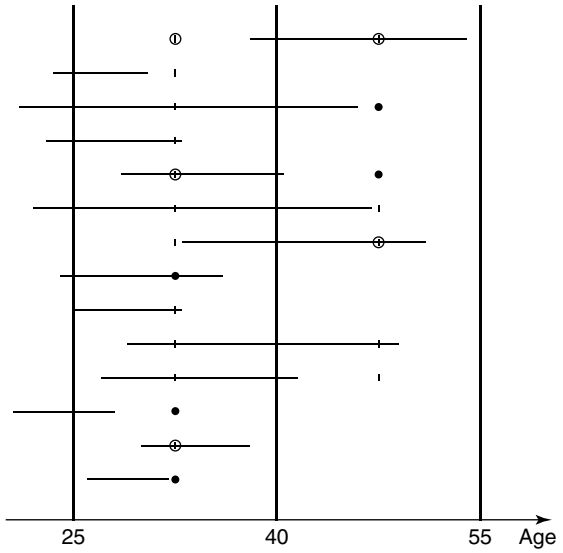

Figure 3 In this example, grouped time risk sets for 25 to 39 and 40 to 55 age groups are defined as subjects that are on study for any portion of the age interval and are indicated by the "|" marks. Nested case–control sampling in grouped time yields an unmatched case–control study structure with multiple cases per set (the •s) and sampled controls (indicated by ○s). Illustrated here, the number of sampled controls sampled to the number of cases

order to explore the effect of exposure to various agents more fully, a nested case–control study was undertaken in which for each of the 11 STS and 32 NHL cases, five controls were sampled from those from the same country, of the same gender, and same year of birth as the case [27]. For each subject in the nested case–control study, industrial hygienists assessed the degree of exposure to 21 chemicals or mixtures based on company records. Increasing trends of risk of STS and NHL were observed for a number of phenoxy herbicides including 2,4D, and TCDD. This study illustrates a number of potential advantages of nested case–control studies. First, having already assembled the workers cohort, the nested case–control study was a natural follow-up design in order to obtain more detailed exposure information. Second, the workers cohort has much higher **prevalence** of TCDD exposure than general population. Thus, the case–control study selected from this cohort will have much higher statistical **power** to investigate TCDD (and other chemical) associations with STS and NHL than a study of similar size from the general population. Third, on

the basis of the $(m-1)/m$ relative efficiency rule [16, 58], this nested case–control study of 258 subjects provides $5/6 = 83\%$ efficiency relative to an analysis of entire cohort of 21 183 for testing associations between single exposures and disease. Finally, because exposure assessment did not require contact with study subjects, recall and selection bias (*see* **Bias in Case–Control Studies**), common problems in standard case–control studies, were avoided.

Nested Case–control Study of Hypertensive Drugs and the Risk of Myocardio-infarction (MI) within a HMO Cohort. In this study, the cohort is defined to be patients within the Group Health Cooperative of Puget Sound who were prescribed hypertensive medication for some time during July 1989 through December 1993 [48]. Failures in the cohort were 623 MI cases. Grouped time risk sets were formed on the basis of calendar year and controls were randomly sampled (about 3 times the number of cases) within matching strata based on 10 year age group and gender. For each case–control study member, the types of antihypertensive drugs used were ascertained through computerized records, chart review, and interview. It was found that risk of MI was 60% higher among calcium channel blocker users compared to that among users of either diuretics alone or β-blockers, a finding that has resulted in a change in treatment strategy. Nesting this case–control study within the HMO cohort had similar advantages to the IARC study. First, the HMO computerized database allowed the identification of cohort members and MI outcome information in a fairly efficient way. There was a high participation rate and, since the type and period of use of drugs could be assessed using the pharmacy database, this information is not subject to information bias.

Residential Magnetic Field Exposure and Breast Cancer. The Multiethnic Cohort is a large population-based cohort from Los Angeles and Hawaii of men and women aged 45 to 74 at enrollment between 1993 and 1996. There were 52 112 female Los Angeles County residents who enrolled in the cohort and completed a self-administered questionnaire that included questions about menstrual and reproductive history, use of oral contraceptives and hormone replacement therapy, diet, and physical activity. For the nested case–control study, 751 breast cancer cases diagnosed by 1999 were ascertained through the National Cancer Institute's Surveillance and End Results (SEER) registry in Los Angeles (*see* **Cancer Registries**). Because the study duration was relatively short, the entire study period was considered as a single grouped time risk set. Controls were approximately **frequency matched**, according to the expected number of breast cancer cases, within self-reported ethnicity. Information on traditional breast cancer risk factors was obtained from the cohort baseline questionnaire (100% participation), and each case or control was invited to have an in-home interview about magnetic field exposures (75% participation). Using the baseline residence for questionnaire participants, wire code was obtained for 99% of all case–control subjects, but because permission was required, magnetic field measurements were obtained in homes of only 44% of subjects. No association between magnetic field measures or wire-code and breast cancer were found [36]. Although covariate information obtained through the interview would be subject to the same information biases as a standard case–control study, there was no selection or information bias for the baseline questionnaire and wire-code data. Further, potential bias, in particular, with regard to the missing patterns for magnetic field measurements, could be assessed using the other variables in the baseline questionnaire.

Nested Case–control Study of the Colorado Plateau Uranium Miners. The Colorado Plateau uranium miners cohort data were collected to assess the effect of occupational radon exposure on the mortality rates (e.g. [25, 39, 41]) (*see* **Radiation Epidemiology**). The cohort consists of 3347 Caucasian male miners who worked underground at least one month in the uranium mines of the four-state Colorado Plateau area and were examined at least once by Public Health Service physicians between 1950 and 1960. These miners were traced for mortality outcomes through December 31, 1982, by which time 258 lung cancer deaths had occurred. Miner radon exposure histories were estimated using job histories and mine radon levels. Although radon and smoking information are available on all cohort members, nested case–control samples with as many as 40 controls per case have been used to reduce the computational burden required to fit complex models exploring the timing of exposures and lung cancer mortality rates [24, 34, 57]. Each of the risk sets was formed by all those who were alive and had entered the study by the

age of death of the case and had attained that age in the same five-year calendar period as the case's date of death (matching by calendar time). The analyses based on the nested case–control data closely approximates the corresponding **Cox regression**, but fitting the models required a fraction of computing time; in the case of the latency models, this meant a reduction from a few days to less than an hour. Further, in the nested case–control data, radon and smoking summaries need only be computed at a fixed (failure) time rather than dynamically in a Cox regression. This makes it much easier to identify data errors and check exposure calculation routines. Absolute risk of lung cancer death, given radon and smoking histories, were estimated from nested case–control data from this cohort [29].

Statistical Analysis Based on the Proportional Hazards and Odds Models for Sampled Risk Set Data

Any of the analysis methods available for "standard" case–control studies, including **conditional** and unconditional **logistic regression** and **Mantel–Haenszel** methods may be used to estimate rate or **odds ratios** from a nested case–control study when the sampling is "simple". Here, we describe methods that are based on the risk set sampling data model that accommodates quite general sampling.

Proportional Hazards and Odds Models

The standard methods for analysis of nested case–control data correspond to and are generalizations of estimation methods for cohort data based on risk sets. Data analysis methods are derived from **semiparametric** models for disease occurrence, the **proportional hazards** or **proportional odds models** being appropriate to continuous or grouped time data, respectively [19]. Each is assumed to have **multiplicative** form

$$\lambda(t; z(t)) = \lambda_0(t) r(z(t); \beta_0) \tag{1}$$

where $r(z(t); \beta)$ is the rate (odds) ratio of disease for an individual with covariates $z(t)$ at time t and $r(0; \beta) = 1$, so $\lambda_0(t)$ is the rate (odds) of disease in subjects with $z = 0$. In continuous time, t refers to any time, while in grouped time the t is discrete and indexes the time intervals. The proportional hazards model may be obtained as the limit to the proportional odds model as the time interval lengths go to zero. As a consequence, the rate ratio parameter and odds ratio parameter in grouped time structure will be close when the probability of failure (rare disease) in each time interval is small.

Estimation of Rate Ratio Parameters from Continuous Time Data

The **partial likelihood** method for cohort data is based on the probability that a subject is a case given the risk set [19, 20]. Similarly, the partial likelihood for nested case–control data is based on the probability that a subject is a case given the case–control set [6, 35, 43, 55]. This will depend on the sampling method and leads to a **likelihood** of the form

$$\prod_{\text{failure times}} \frac{r_{\text{case}}(\beta)\pi_{\text{case}}}{\sum_{k \in \tilde{\mathcal{R}}} r_k(\beta)\pi_k} \tag{2}$$

where $\tilde{\mathcal{R}}$ is the case–control set, the r_k are computed at the failure time, and π_k is the probability of picking the particular case–control set if k was the case. These will generally be replaced by a convenient weights w_k that are proportional to the π_k.

For instance, for **simple random sampling** of $m-1$ controls from the $n-1$ in the risk set, $\pi_k = \binom{n-1}{m-1}^{-1}$. In this case, the π_k are the same for all case–control set members so we may take $w_k = 1$, which yields the "unweighted" conditional likelihood for standard matched case–control data (*see* **Matched Analysis**). Standard conditional logistic regression software may be used to estimate the rate ratio (β_0) parameters (*see* **Software, Biostatistical**).

Estimation of Odds Ratio Parameters from Grouped Time Data

Parallel to the continuous time situation, a partial likelihood is based on the probability that a set of subjects **D** are the cases given that the case–control set is $\tilde{\mathcal{R}}$ and is given by

$$\prod_{\text{grouped times}} \frac{\lambda^{|\mathbf{D}|}\, r_{\mathbf{D}}(\beta)\pi_{\mathbf{D}}}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}} \lambda^{|\mathbf{s}|} r_{\mathbf{s}}(\beta)\pi_{\mathbf{s}}}, \tag{3}$$

where $r_{\mathbf{s}}(\beta) = \prod_{j \in \mathbf{s}} r(Z_j; \beta)$, $|\mathbf{s}|$ is the number of elements in **s**, and $\pi_{\mathbf{s}}$ is the probability of picking the case–control set given that **s** is the set of

cases. For analysis, the $\pi_{\mathbf{s}}$ can be replaced by convenient weights $w_{\mathbf{s}}$ that are proportional to $\pi_{\mathbf{s}}$ [31]. For instance, in $1 : m-1$ frequency matching, the number of controls randomly sampled is $m-1$ times the number of cases. Then, the $\pi_{\mathbf{s}} = \binom{n-|\mathbf{D}|}{m-|\mathbf{D}|}^{-1}$ for all subsets $\mathbf{s}$ of the case–control set $\tilde{\mathcal{R}}$ that are of the same size as the case set. Cancellation of the common π from numerator and denominator leads to the standard (unweighted) conditional logistic likelihood for unmatched case–control data. On the other hand, case–based sampling (for example, [28]) in which a random sample of $m|\mathbf{D}|$ subjects (without regard to failure status) is drawn from the cohort and additional failures are included is "weighted" with $w_{\mathbf{s}} = \binom{|\mathbf{s}|}{m|\mathbf{D}|-(|\tilde{\mathcal{R}}|-|\mathbf{s}|)}$ [31].

Conditional Logistic Likelihood. A likelihood estimator that is closely related to the partial likelihood conditions on the number of cases so that the *conditional logistic likelihood* is given by

$$\prod_{\text{grouped times}} \frac{r_{\mathbf{D}}(\beta)\pi_{\mathbf{D}}}{\sum_{\mathbf{s}\subset\tilde{\mathcal{R}}:|\mathbf{s}|=|\mathbf{D}|} r_{\mathbf{s}}(\beta)\ \pi_{\mathbf{s}}}. \tag{4}$$

Unlike the partial likelihood (3) from which the baseline odds may often be estimated, the baseline odds parameter is conditioned out of (4). Also, unlike the partial likelihood, for each of the standard (simple) control selection methods, including frequency matching, Bernoulli trials (*see* **Binary Data**) and case-base, the conditional likelihood is the same, "unweighted" version [3]. The conditional likelihood is often used when the number of cases and/or controls in all or some case–control sets is small or when there are tied failure times in continuous time analyses (*see* **Tied Survival Times**).

Unconditional Logistic Regression. This is the most commonly used alternative for analysis of grouped time nested case–control studies with random sampling and is based on the product of "marginal" case/control probabilities within the case–control set

$$\prod_{\text{grouped times}} \prod_{j\in\tilde{\mathcal{R}}} \frac{[\lambda\ w_j\ r_j(\beta)]^{D_j}}{1+\lambda\ w_j\ r_j(\beta)}, \tag{5}$$

where D_j is a case–control status indicator, w_j is a marginal "inverse control sampling probability". The w_j depend both on β_0 and j, but for common situations, this dependence is small for "large samples". Further, because the probabilities are only marginal, the variance cannot generally be estimated as the "inverse information" (e.g., [11, 31]). The unconditional logistic likelihood also arises from two-phase studies, closely related to nested case–control studies, in which the cohort is taken as a fixed set of cases and controls from which a second stage sample is drawn from the first, for example, [11, 13, 53, 62, 67] (*see* **Case–Control Study, Two-phase**). In the case of simple random sampling (with $m-1$ controls per case), the w_j are approximately all equal to $(n-|\mathbf{D}|)/(m|\mathbf{D}|-|\mathbf{D}|)$ and a **nuisance parameter** θ_0 can be used in place of $\lambda_0\ w_j$ in the likelihood. In this special case, the variance of the odds ratio parameter may be estimated using the standard inverse information estimator [3].

Absolute Risk Estimation

Unlike a standard case–control study in which the cohort cannot be identified from a nested case–control study, it is possible to estimate the baseline **hazard rate** and, more generally, absolute risk quantities that are functions of the hazard.

Estimation of Risk from Continuous Time Data. If the number at risk n and the control selection probabilities π_k are known at each failure time, then **cumulative hazard** functions and survival functions may be estimated using a generalization of the Breslow estimator of the baseline hazard [1, 19, 29] (*see* **Hazard Ratio Estimator**). Let $z^0(t)$ be a covariate history and $r^0(t) = r(z^0(t); \beta_0)$ be the rate ratio (as a function of time) associated with z^0 according to the model. The basic components for estimators of risk are the jumps in the hazard at the failure times. With n the number at risk and $\hat{r}_k = r(z_k, \hat{\beta})$, the relative risk for individual k predicted using $\hat{\beta}$, the hazard jump from a case–control set is estimated by

$$\hat{r}^0 \Big/ n \sum_j \frac{\pi_j}{\sum_k \pi_k} \hat{r}_j, \tag{6}$$

where the sums are over case–control set members [29]. Note that setting $m = n$ and $z^0(t) \equiv 0$ yields the Breslow estimator of the baseline hazard for the full cohort [19]. For simple random sampling of $m-1$ controls, $\pi_j / \sum_k \pi_k = m$ so the denominator is given by $n/m \sum \hat{r}_k$. Cumulative hazard

and **Kaplan–Meier-type estimators** of risk and an **Aalen–Johansen-type estimator** of risk in the presence of **competing** causes of failure, as well as corresponding variance estimators have been described with application to the Colorado Plateau uranium miners study [9, 29].

Estimation of Risk from Grouped Time Data. Estimation of absolute risk from grouped time nested case–control data with general sampling is a topic of continuing research. When the overall risk of disease is known in the cohort, one approach uses the distribution of covariates in controls as representative of the cohort rates to infer risk within exposure subgroups [4]. When the w_j can be specified, then the baseline odds (and hence the risk) can evidently be estimated using the grouped time partial (3) or the unconditional (5) likelihoods.

Asymptotic Properties and Efficiency

The "likelihoods" for both continuous and grouped time are "partial" in the same sense as the Cox partial likelihood for full cohort data [19] in that the same subject may appear in multiple sets. In an extension to the **counting process** and martingale theory approach for full cohort data [2], nested case–control data is represented by a counting process $N_{i,\mathbf{r}}(t)$ for occurrences of both subject i becoming diseased and $\mathbf{r}$ the case–control set. Within this framework, the case–control set variability is constant, with sample size and the asymptotics driven by the increasing number of case–control sets. Conditions for the consistency and asymptotic normality of the partial likelihood rate ratio and baseline hazard estimators have been described for a wide range of sampling methods [6, 23]. Also provided are expressions for the asymptotic variance from which **efficiency**, statistical **power**, and **sample size** calculations can be made. For simple sampling, these are refinements of those for standard individually matched case–control studies that take into account the underlying failure time structure [23]. Performance of the partial likelihood under model **misspecification** under simple sampling has also been studied [63, 64]. For grouped time case–control data, the framework can be similarly defined with $N_{\mathbf{d},\mathbf{r}}(t)$ now indicating the set $\mathbf{d}$ of diseased subjects and $\mathbf{r}$ the case–control set. However, the asymptotic theory will depend on if and how the time intervals "shrink" as a function of sample size. For fixed time intervals, the number of case–control sets is fixed and the asymptotics are driven by increasing sample size within case–control sets and thus, the asymptotic theory is very different than in the continuous time situation. The theory for the "unweighted" conditional logistic (4) and unconditional logistic (5) likelihoods based on rejective sampling has been described [3]. However, a general theory has not been derived in the grouped time setting.

Other Approaches to Estimation and Other Models

Proportional Hazards Models. **Mantel–Haenszel** estimators for nested case–control studies with simple random sampling have been described and shown to be consistent [65, 66]. Methods for estimation of relative mortality have been described on the basis of an extension of the proportional hazards model [7, 14, 54].

Another class of estimators seeks to use case and control information at times other than when they were sampled, all with the goal of capturing more information from the case–control sample than the partial likelihood. There have been a number of methods that enlarge or restrict the controls used in the unweighted version of (2) in order to increase efficiency [33, 45, 49, 60]. Interestingly, even though these methods made "better" use of the sample, it was found that efficiency gains were modest at best, and often were worse in situations of practical importance [33]. Another method incorporates external rates and estimation is based on joint cohort Poisson and nested case–control partial likelihoods [56]. Methods have been proposed using an "inverse weighting" method [18, 51] as well as an estimator based on "local averaging". The latter was shown to be more efficient than earlier extensions and is more efficient than the partial likelihood in a range of situations [17]. All these methods show some improvement when disease is common and/or the rate ratio is large. Further work is necessary to establish the practical guidelines for when these methods offer significant benefits over the standard methods.

Nonproportional Hazards Models and Extensions. Methods for modeling **excess risk** and estimation of absolute risk based on the Aalen linear regression

model, from nested case–control studies have also been developed [8]. Nested case–control studies with appropriately sampled controls can be used to estimate transition rate ratio parameters for recurrences or multiple outcomes [37, 38] and for **Markov** transition probabilities [5]. A useful method for estimation of parameters in parametric models has been described [51].

Control Sampling Methods

General Guidelines. The likelihood methods are all valid (and most useful) if the risk sets are sampled independently over time so that subjects may serve as controls in multiple case–control sets and failures may be controls in "earlier" risk sets. Restricting controls to be used only once or using "pure" controls, those that never become cases, will, in theory, result in biased estimation [40] unless special analysis methods are used [45]. However, if disease is rare, this bias is generally negligible. Very general sampling methods can be accommodated by the risk set sampling likelihoods but a general guideline for useful designs is that the structure of the case–control set should not reveal the identity of the case [30].

Matching and Random Sampling. The most commonly used methods for selecting controls is to randomly sample from risk set members who "match" on a set of factors. For continuous time case–control studies, this means that the sampled controls will be similar to the individual case on these factors. For grouped time studies, the risk set will be partitioned into matching strata and controls (and cases) are sampled from within these strata. Although the choice of matching criteria will depend on the needs of the study, common matching factors include gender, race/ethnicity, calendar year, and/or year of birth. The latter is often desirable because, as in the hypertensive drug-MI example, a natural timescale is age but matching on year of birth aligns the cases and controls with respect to calendar time and thus assures comparable data quality and control for "secular trends" in diagnostic treatment practices [12].

Fine Matching. When a continuous matching factor is available on all cohort members, nearest neighbor and caliper matching are possible as continuous time control selection options [26]. For instance, in a study of occupational exposure to chemical agents and pancreatic cancer, for each case, the four controls in the risk set who most closely matched on date of birth were enrolled into the nested case–control study [22]. In this study, there was no random sampling at all and the nearest neighbor matching completely determined the control selection. The unweighted partial likelihood is not strictly correct under this sampling, but conditions have been described when it is asymptotically valid, as well as other analysis options when the matching factor is included as a covariate in the proportional hazard model [26].

Exposure Stratified Sampling Methods. Until recently, it was thought that control selection could not depend on exposure related variables [50]. In fact, unbiased estimation is possible if the appropriate control selection probabilities (weights) are specified in the likelihood. Such designs include **counter-matching**, variants of quota sampling, two-stage sampling, and exposure stratified case-base (**case–cohort**) sampling [6, 10, 11, 30, 31, 61].

Issues Related to Grouped Time Studies

Continuous Time as a Limiting Case of the Grouped Time Model. Parallel to the approach of Cox for cohort data organized into risk sets, individually matched case–control study designs and methods can be obtained from unmatched studies by "shrinking the time interval" to zero. Thus, the proportional odds model converges to the proportional hazards, $1 : m - 1$ frequency matching becomes $1 : m - 1$ individual matching and the grouped time partial (or conditional) likelihood converges to the continuous time partial likelihood.

Grouped Time as an Approximation to Reality. Because time is in reality "continuous", the grouped time approach necessarily involves a number of approximations, which may be critical when the grouped time intervals are large. One issue is the ambiguity in the definition of who is in the risk set, in particular, among those who are censored during a grouped time interval. The problems that can arise correspond to those associated with ignoring censoring in failure time data (but on the scale of the time interval). In Figure 3, we have defined subjects as being at risk in the interval if they are at risk during any part of the interval. Another problem with

grouping time is that there is not a single unambiguous "reference time" from which to compute time-dependent covariates. Various strategies have been to use the average time for the cases, as was done in the residential magnetic fields breast cancer example [36] or to randomly assign times to the controls based on the case failure time distribution within the interval. Unless there are large changes in at-risk status and/or covariate values over time intervals, strategies that reasonably approximate the continuous time risk sets will yield estimates that are close to the corresponding continuous time estimator.

Failure Time Analysis of Grouped Time Case–control Studies. A number of methods have been developed to estimate rate ratio parameters when the case–control study is sampled from grouped time risk sets. Most notable of these is the case–cohort method [44]. To see that this is a grouped time sampling, note that the sampling is the "case-base" method [28] in which a random sample of subjects (without regard to failure status) is drawn from the cohort and additional failures are included. The grouped time analysis methods based on estimation of *odds ratios* for exposures apply with the time interval taken as the entire study period; this analysis is subject to the pitfalls associated with grouping time described above. The case–cohort *analysis* method allows estimation of the *rate ratios* appropriately accounting for censoring and **time-dependent covariates** from the case-based sampled data. This idea was generalized to estimators of rate ratios from "simple" unmatched case–control data from the cohort [17]. A comparison of nested case–control and case–cohort approaches is given in the article **Case–Cohort Study**.

Nested and Standard Case–control Studies

Relevance of Nested Case–control Studies. Standard case–control studies may often be viewed as a nested case–control study within a nonassembled (and perhaps poorly defined) cohort. Thus, methods developed for nested case–control studies that do not require further knowledge of cohort information will apply to standard case–control designs. So, a number of study designs, including quota sampling, modified randomized recruitment, and individually matched two-stage studies have been proposed on the basis of the nested case–control study paradigm that do not require an assembled cohort [6, 30, 31]. Another example is a robust (*see* **Robustness**) variance estimator derived for continuous time $1 : m - 1$ nested case–control studies that may be used in standard individually matched case–control study analysis [64].

Difference in Methodological Approach for Nested and Standard Case–control Studies. Methodologically, perhaps the biggest difference between nested and "nonnested" case–control studies is the data model used to develop methods. Traditionally, case–control data is viewed as generated "retrospectively", with individual exposure independent random quantities, with distribution conditional on disease status (e.g. [46]). A key result is that, even under the retrospective model, odds ratio parameters are estimable using the unconditional logistic likelihood (5), (e.g. [47]). For simple sampling, either approach leads to similar methods and inference about odds ratio parameters in grouped time data [3]. However, methods developed under the risk set sampling model provide a connection to failure time cohort data and associated methods and, further, provide a natural framework for the development of methods for individually matched case–control data.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

[2] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[3] Arratia, R., Goldstein, L. & Langholz, B. (2005). Local central limit theorems, the high order correlations of rejective sampling, and logistic likelihood asymptotics, *Annals of Statistics*; to appear.

[4] Benichou, J. & Gail, M.H. (1995). Methods of inference for estimates of absolute risk derived from population-based case-control studies, *Biometrics* **51**, 182–194.

[5] Borgan, Ø (2002). Estimation of covariate-dependent markov transition probabilities from nested case-control data, *Statistical Methods in Medical Research* **11**, 183–202.

[6] Borgan, Ø, Goldstein, L. & Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model, *Annals of Statistics* **23**, 1749–1778.

[7] Borgan, Ø & Langholz, B. (1993). Non-parametric estimation of relative mortality from nested case-control studies, *Biometrics* **49**, 593–602.

[8] Borgan, Ø & Langholz, B. (1997). Estimation of excess risk from case-control data using Aalen's linear regression model, *Biometrics* **53**, 690–697.

[9] Borgan, Ø & Langholz, B. (1998). Risk set sampling designs for proportional hazards models, in *Statistical Analysis of Medical Data: New Developments*, B.S. Everitt & G. Dunn, eds. Arnold, London, pp. 75–100.

[10] Borgan, Ø, Langholz, B., Samuelsen, S.O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs, *Lifetime Data Analysis* **6**, 39–58.

[11] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two stage case-control data, *Biometrika* **75**, 11–20.

[12] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research. Volume II – The Design and Analysis of Cohort Studies*, Vol. 82, IARC Scientific Publications. International Agency for Research on Cancer, Lyon.

[13] Breslow, N.E. & Holubkov, R. (1997). Weighted likelihood, pseudo-likelihood, and maximum likelihood methods for logistic regression analysis of two-stage data, *Statistics in Medicine* **16**, 103–116.

[14] Breslow, N.E. & Langholz, B. (1987). Nonparametric estimation of relative mortality functions, *Journal of Chronic Diseases* **131**(Suppl. 2), 89S–99S.

[15] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[16] Breslow, N.E. & Patton, J. (1979). Case-control analysis of cohort studies, in *Energy and Health*, N.E. Breslow & A.S. Whittemore, eds. SIAM Institute for Mathematics and Society, SIAM, Philadelphia, pp. 226–242.

[17] Chen, K. (2001). Generalized case-cohort sampling, *Journal of the Royal Statistical Society, Series B, Methodological* **63**, 791–809.

[18] Chen, K. & Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model, *Biometrika* **86**, 755–764.

[19] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **34**, 187–220.

[20] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[21] Ernster, V.L. (1994). Nested case-control studies, *Preventive Medicine* **23**, 587–590.

[22] Garabrant, D.H., Held, J., Langholz, B., Peters, J.M. & Mack, T.M. (1992). DDT and related compounds and the risk of pancreatic cancer, *Journal of the National Cancer Institute* **84**, 764–771.

[23] Goldstein, L. & Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model, *Annals of Statistics* **20**, 1903–1928.

[24] Hauptmann, M., Behrane, K., Langholz, B. & Lubin, J.H. (2001). Using splines to analyze latency in the colorado plateau uranium miners cohort, *Journal of Epidemiology and Biostatistics* **6**, 417–424.

[25] Hornung, R.W. & Meinhardt, T.J. (1987). Quantitative risk assessment of lung cancer in U. S. uranium miners, *Health Physics* **52**, 417–430.

[26] Kim, S. & De Gruttola, V. (1999). Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial, *Lifetime Data Analysis* **5**(2), 149–72.

[27] Kogevinas, M., Kauppinen, T., Winkelmann, R., Becher, H., Bertazzi, P.A., Bueno de Mesquita, H.B., Coggon, D., Green, L., Johnson, E., Littorin, M., Lynge, E., Marlow, D.A., Mathews, J.D., Neuberger, M., Benn, T., Pannett, B., Pearce, N. & Saracci, R. (1995). Soft tissue sarcoma and non-Hodgkin's lymphoma in workers exposed to phenoxy herbicides, chlorophenols, and dioxins: two nested case-control studies, *Epidemiology* **6**, 396–402.

[28] Kupper, L.L., McMichael, A.J. & Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk, *Journal of the American Statistical Association* **70**, 524–528.

[29] Langholz, B. & Borgan, Ø (1997). Estimation of absolute risk from nested case-control data, *Biometrics* **53**, 767–774.

[30] Langholz, B. & Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies, *Statistical Science* **11**, 35–53.

[31] Langholz, B. & Goldstein, L. (2001). Conditional logistic analysis of case-control studies with complex sampling, *Biostatistics* **2**, 63–84.

[32] Langholz, B., Rothman, N., Wacholder, S. & Thomas, D.C. (1999). Cohort studies for characterizing measured genes, *Monographs Journal of the National Cancer Institute* **26**, 39–42.

[33] Langholz, B. & Thomas, D.C. (1991). Efficiency of cohort sampling designs: some surprising results, *Biometrics* **47**, 1563–1571.

[34] Langholz, B., Thomas, D.C., Xiang, A. & Stram, D. (1999). Latency analysis in epidemiologic studies of occupational exposures: application to the colorado plateau uranium miners cohort, *American Journal of Industrial Medicine* **35**, 246–256.

[35] Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977). Methods of cohort analysis: appraisal by application to asbestos miners, *Journal of the Royal Statistical Society A* **140**, 469–491.

[36] London, S.J., Pogoda, J.M., Hwang, K.L., Langholz, B., Monroe, K.R., Kolonel, L.N., Kaune, W.T., Peters, J.M. & Henderson, B.E. (2003). Residential magnetic field exposure and breast cancer risk in the multiethnic cohort study, *American Journal of Epidemiology* **158**, 969–980.

[37] Lubin, J.H. (1985). Case-control methods in the presence of multiple failure times and competing risks, *Biometrics* **41**, 49–54.

[38] Lubin, J.H. (1986). Extensions of analytic methods for nested and population-based incident case-control studies, *Journal of Chronic Diseases* **39**, 379–88.

[39] Lubin, J.H., Boice, J.D., Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Tirmarche, M., Woodward, A., Xiang, Y.S. & Pierce, D.A. (1994). *Radon and Lung*

Cancer Risk: A Joint Analysis of 11 Underground Miners Studies, NIH Publication 94–3644, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, Bethesda.

[40] Lubin, J.H. & Gail, M.H. (1984). Biased selection of controls for case-control analyses of cohort studies, *Biometrics* **40**, 63–75.

[41] Lundin, F.D., Wagoner, J.K. & Archer, V.E. (1971). *Radon daughter exposure and respiratory cancer, quantitative and temporal aspects*. Joint Monograph 1, U.S. Public Health Service, Washington.

[42] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics* **29**, 479–486.

[43] Oakes, D. (1981). Survival times: aspects of partial likelihood (with discussion), *International Statistical Review* **49**, 235–264.

[44] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[45] Prentice, R.L. (1986). On the design of synthetic case-control studies, *Biometrics* **42**, 301–310.

[46] Prentice, R.L. & Breslow, N.E. (1978). Retrospective studies and failure time models, *Biometrika* **65**, 153–158.

[47] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika* **66**, 403–411.

[48] Psaty, B.M., Heckbert, S.R., Koepsell, T.D., Siscovick, D.S., Raghunathan, T.E., Weiss, N.S., Rosendaal, F.R., Lemaitre, R.N., Smith, N.L., Wahl, P.W., et al. (1995). The risk of myocardial infarction associated with antihypertensive drug therapies, *Journal of the American Medical Association* **274**, 620–625.

[49] Robins, J.M., Prentice, R.L. & Blevins, D. (1989). Designs for synthetic case-control studies in open cohorts, *Biometrics* **45**, 1103–1116.

[50] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Ed. Lippincott-Raven Publishers, Philadelphia.

[51] Samuelsen, S.O. (1997). A pseudolikelihood approach to analysis of nested case-control data, *Biometrika* **84**, 379–394.

[52] Saracci, R., Kogevinas, M., Bertazzi, P.A., Bueno de Mesquita, B.H., Coggon, D., Green, L.M., Kauppinen, T., L'Abbé, K.A., Littorin, M., Lynge, E., Mathews, J.D., Neuberger, M., Osman, J., Pearce, N. & Winkelmann, R. (1991). Cancer morality in workers exposed to chlorophenoxy herbicides and chlorophenols, *Lancet* **338**, 1027–1032.

[53] Scott, A.J. & Wild, C.J. (1991). Fitting logistic models under case-control or choice based sampling, *Journal of the Royal Statistical Society Series B* **48**, 170–182.

[54] Suissa, S., Edwardes, M.D. & Boivin, J.F. (1998). External comparisons from nested case-control designs, *Epidemiology* **9**, 72–78.

[55] Thomas, D.C. (1981). General relative-risk models for survival time and matched case-control analysis, *Biometrics* **37**, 673–686.

[56] Thomas, D.C., Blettner, M. & Day, N.E. (1992). Use of external rates in nested case-control studies with application to the international radiation study of cervical cancer patients, *Biometrics* **48**, 781–794.

[57] Thomas, D.C., Pogoda, J., Langholz, B. & Mack, W. (1994). Temporal modifiers of the radon-smoking interaction, *Health Physics* **66**, 257–262.

[58] Ury, H.K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data, *Biometrics* **31**, 643–649.

[59] Wacholder, S. (1995). Design issues in case-control studies, *Statistical Methods in Medical Research* **4**, 293–309.

[60] Wacholder, S., Gail, M.H. & Pee, D. (1991). Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort, *Biometrics* **47**, 63–76.

[61] Weinberg, C.R. & Wacholder, S. (1990). The design and analysis of case-control studies with biased sampling, *Biometrics* **46**, 963–975.

[62] Whittemore, A. (1997). Multistage sampling designs and estimating equations, *Journal of the Royal Statistical Society B* **59**, 589–602.

[63] Xiang, A.H. & Langholz, B. (1999). Comparison of case-control to full cohort analyses under model misspecification, *Biometrika* **86**, 221–226.

[64] Xiang, A.H. & Langholz, B. (2003). Robust variance estimation for rate ratio parameter estimates from individually matched case-control data, *Biometrika* **90**, 741–746.

[65] Zhang, Z.-Z. (2000). On consistency of Mantel-Haenszel type estimators in nested case-control studies, *Journal of the Japan Statistical Society* **30**(2), 205–211.

[66] Zhang, Z.-Z., Fujii, Y. & Yanagawa, T. (2000). On Mantel-Haenszel type estimators in simple nested case–control studies, *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]* **29**, 2507–2521.

[67] Zhao, L.P. & Lipsitz, S. (1992). Designs and analysis of two-stage studies, *Statistics in Medicine* **11**, 769–782.

BRYAN LANGHOLZ

Case–Control Study, Population-based

A population-based **case–control study** is based on a well-defined source population from which all cases that arise in a given time period can be enumerated. **Controls** consist of **random samples** of persons without the disease of interest from the source population. Cases consist of all cases or a random sample of those cases. Because the total number of cases is known and the size of the source population can usually be estimated from **census** data or other sources, a population-based case–control study yields information not only on **relative risk**, but also, by combining information on relative risk with information on overall disease risk, on exposure-specific **absolute risk**.

Compared with **hospital-based case–control studies**, population-based case–control studies can, in principle, avoid **selection biases** that produce nonrepresentative samples of cases and controls. Moreover, there is no ambiguity about what constitutes an appropriate control. In practice, however, selection biases can arise if all cases are not identified or if persons selected for inclusion in the study refuse to participate. In addition, **recall bias** can distort the results of such a study if persons with disease provide a different quality of information on exposure from healthy controls selected from the general population.

(*See also* **Bias in Case–Control Studies**; **Bias in Observational Studies**; **Bias, Overview**)

MITCHELL H. GAIL

Case–Control Study, Prevalent

Incident cases represent the change from a nondiseased to a diseased state. For research purposes, a population at risk is defined as one whose members are, as of some arbitrary time point, disease free. Over time, incident cases emerge from that population. In a study with incident cases, the underlying assumption is that either all of the cases produced by the population are available for study, or that the study includes what may be taken as a **random sample** of all of the cases [18]. In this way the experience of the cases and the population at risk can, in principle, be used to investigate the etiology of the disease, conditional on adjustment for potential **confounding** [16, 18, 24]. When, for efficiency, a **case–control study** is done, the measure of effect estimated by the exposure **odds ratio** will be either the **incidence density ratio**, the **cumulative incidence ratio**, or the disease odds ratio, depending upon the sampling scheme that was used to select the **controls** [17, 20, 23]. However, whatever the estimated parameter, the potential for etiologic inference remains, conditional on adjustment for potential confounding.

The situation is more complex when prevalent cases are included and the intent of the research remains etiologic. Prevalent cases have made the transition from a nondiseased to a diseased state. However, they also *survived* to the time the study sample was obtained [6, 16, 29]. All prevalent cases were once incident cases; but not all incident cases survive long enough to become prevalent cases.

When a study includes prevalent cases, the question arises as to whether the prevalent case series can reasonably be taken to represent a random sample of all incident cases with respect to the distribution of etiologically relevant factors (and potential confounders). For the answer to be "yes", incident cases that did not survive long enough to become prevalent cases must be assumed to have been etiologically similar to the incident cases that survived and became prevalent cases [3]. Survival would thus be unrelated to any etiologically important factor [16, 17, 24]. For this special situation, the prevalent case–control study is interchangeable with the incident case–control study with respect to the validity of the relative measure of effect, though it may differ with respect to **power** [2].

In the more likely situation, the prevalent case series cannot reasonably be taken to represent a random sample of all incident cases [2, 16, 29]. In particular, cases with short survival times will be under represented; longer surviving cases will be over represented; and most significantly, the duration of survival may well be related to etiologically relevant factors. Risk factors for the disease therefore will be simultaneously related to etiology *and* prognosis [6, 17, 29]. The exposure odds ratio will not directly estimate the incidence ratios that are the target for etiologic research. Table 1 illustrates how estimates of relative effects from a case–control study with prevalent cases will or will not validly duplicate the estimate from a case–control study with incident cases.

Linking Prevalence- and Incidence-Based Studies

Freeman & Hutchison [7, 8] have demonstrated the interrelations between incidence and prevalence (*see* **Incidence–Prevalence Relationships**)–in particular, that

> prevalence, incidence, and duration of a condition or illness ... are interrelated in such a way that two of these quantities may be used to obtain the third. Data may be collected in the most expedient manner and the results expressed as both incidence and prevalence [7, p. 707].

At the core of these interrelationships is the assumption that the population is "steady" or "stationary" [1, 2, 7, 8, 16, 17, 21, 24]. For a population in a steady state, the immigration rate into the candidate pool equals the emigration rate from the candidate pool, so that the size of the candidate pool is constant over time. Similarly, the immigration rate into the prevalence pool equals the emigration rate from the prevalence pool, so that the size of the prevalence pool is constant over time. A corollary of these conditions is that the distribution of survival of incident cases remains constant. A practical consequence is that if the population is in a steady state, then estimates of disease incidence are not dependent upon the time period of the study. In a steady-state population, the prevalence odds (PO) equals the product of incidence density (ID) and the average duration ($\overline{D}$)

Table 1 Estimation of relative effects in incident and prevalent case–control studies

Frame A: Hypothetical data from an incident case–control study:

	Exposed	Unexposed	
Cases	50	50	
Controls	100	900	$OR = 9.0$

Frame B: Hypothetical data from a prevalent case–control study, where 50% of all cases do not survive long enough to be included in the study. Survival, however, is unrelated to exposure status:

	Exposed	Unexposed	
Cases	25	25	
Controls	100	900	$OR = 9.0$

Frame C: Hypothetical data from a prevalent case–control study, where 50% of exposed cases do not survive long enough to be included in the study. Survival is therefore related to exposure status:

	Exposed	Unexposed	
Cases	25	50	
Controls	100	900	$OR = 4.5$

Frame D: Hypothetical data from a prevalent case–control study, where 50% of unexposed cases do not survive long enough to be included in the study. Survival is therefore related to exposure status:

	Exposed	Unexposed	
Cases	50	25	
Controls	100	900	$OR = 18.0$

of the illness or condition (see Appendix A). If this relation is calculated for the exposed and unexposed segments of the population (with a subscript 1 indicating exposure and a subscript zero indicating nonexposure), then the prevalence odds ratio (POR) is

$$POR = \frac{PO_1}{PO_0} = \frac{ID_1 \times \overline{D}_1}{ID_0 \times \overline{D}_0} = IDR \times \frac{\overline{D}_1}{\overline{D}_0}. \quad (1)$$

In a case–control study with prevalent cases, the prevalence odds ratio (POR) estimates the incidence density ratio (IDR) if the population is in a steady state, and if the duration of disease among the exposed equals the duration of disease among the unexposed, that is, if survival is not related to an etiologically important factor (as in Table 1, Frame B).

Example 1: Nosocomial Infections

Freeman and his coauthors have used data from a "bed-to-bed" prevalence survey at a municipal hospital to illustrate the relations between prevalence and incidence [7, 8] and to study predictors of nosocomial infection [9–11]. They argue that it is reasonable to expect a hospital population to be in a steady state. Moreover, since admission, discharge, and occupancy data are collected daily, the conditions underlying a steady state can be empirically checked.

Data from their report on risk factors for nosocomial infection [9] are reproduced in Table 2, Frame A. Suppose the data represent a complete **cross-sectional** survey, and, for illustrative purposes, that the relation between use of an endotracheal tube and nosocomial infection is unconfounded. We can compute the prevalence ratio (PR) and the prevalence odds ratio (POR).

Table 2 Prevalence data on the relation between the need for an endotracheal tube and nosocomial infection

Frame A: Prevalence survey data:[a]

	Need for an endotracheal tube	
	Present	Absent
Nosocomial infection		
Present	7	90
Absent	4	544
$POR = 10.6$		

Frame B: Prevalent case–control data, with 50% sampling of controls:

	Need for an endotracheal tube	
	Present	Absent
Nosocomial infection		
present	7	90
absent	2	272
$POR = 10.6$		

[a]The data are from [9, Table 1, p 814].

The $PR = [7/11]/[90/634] = 4.5$; the $POR = [7/4]/[90/544] = 10.6$. While both point estimates are large, the POR is not a good estimate of the PR because the overall prevalence is not low (15%) and the exposure rate in the noncases is less than half the exposure rate in the total population surveyed (0.7% and 1.7%, respectively).

Table 2, Frame B shows data from a potential case–control study drawn from this prevalence survey. Suppose, for example, that determination of the exposure status of the 645 subjects was expensive. For efficiency, the exposure status could be determined for all cases and for a 50% sample of controls (again assuming no confounding). Since the parameter of interest is the *POR* (because of its relation to the *IDR*), the controls are drawn from the noncases. The *POR* of 10.6 will be a valid estimate of the incidence density ratio *if*:

1. the hospital population is in a steady state, i.e. if "the rate at which patients were admitted equaled the rate at which they were discharged, and the rate at which patients acquired active nosocomial infections and entered the prevalence pool equaled the rate at which they left the prevalence pool through recovery, death, or discharge ... [and] the rates at which new patients entered the hospital and new infections were acquired were ... constant ... [and the] distributions of durations of hospitalization and durations of infections were also ... constant" [11, p. 734]

and

2. the duration of nosocomial infection is equal for those who acquire an infection and have had an endotracheal tube and those who acquire an infection and have not had an endotracheal tube, i.e. that the exposure is not associated with the duration of the disease.

Freeman & McGowan [9, p. 815] report that the duration of nosocomial infection in the incident cases did not differ by the need for an endotracheal tube. The duration of infection was estimated from durations-to-date in the prevalent cases (the time from the onset of the infection to the time of the prevalence survey). Suppose, however, for illustrative purposes, that the durations of infection differed. The distribution of observed durations-to-date in the prevalent series can be converted into the distribution of durations of infection *from disease incidence* in the same steady-state population, as shown by Freeman & Hutchison [7]. In particular:

$$p_i(D) = \frac{[\overline{D}_i \times p_p(D)]}{D}, \qquad (2)$$

where

$p_i(D)$ = the proportion from the incidence series (designated by i) with duration D;
$\overline{D}_i$ = the average duration of the disease or condition from a series of incident cases;
$p_p(D)$ = the proportion from the prevalence series (designated by p) with duration D; and
D = a specific duration of the disease or condition from onset to termination of the disease or condition, or to removal from observation.

Suppose infections associated with a poor underlying condition have a longer duration than infections associated with a less serious underlying condition, and that the need for an endotracheal tube is a marker for a poor underlying condition. Table 3 (Frame A) shows hypothetical data from a prevalent case–control study, along with information on the distribution of durations-to-date for the prevalent cases, by exposure categories.

From (2) we obtain the distribution of durations from disease incidence in exposed and unexposed cases (Table 3, Frame B) up to a constant multiplier $\overline{D}_\mathrm{i}$. For example, $p_\mathrm{i}(5) = \overline{D}_\mathrm{i} \times 0.057/[\overline{D}_\mathrm{i} \times 0.057 + \overline{D}_\mathrm{i} \times 0.071] = 0.45$ for exposed cases. Using these distributions, we calculate $\overline{D}_\mathrm{i} = 7.8$ days for the group needing endotracheal tubes and 5.9 days for the group not needing endotracheal tubes (Table 3, Frame B).

A comparison of the observed distribution of durations-to-date from the prevalent series and the calculated distribution of the durations of disease in the incident series illustrates the length-biased sampling (*see* **Screening Benefit, Evaluation of**) that can occur when prevalent rather than incident cases are studied [16, 25, 29]. In particular, cases of long duration represent 71% of the prevalent series of exposed cases, but would constitute only 55% of the incident series. For the unexposed group, the cases of long duration represent 30% of the prevalent series, but would constitute only 18% of the incident series.

By using (1) and solving for the incidence density ratio, the prevalence odds ratio can be adjusted by the *inverse* average duration ratio to provide an estimate of the incidence density ratio. In particular:

$$IDR = POR \times \frac{\overline{D}_0}{\overline{D}_1}. \tag{3}$$

For the example in Table 3, the prevalence odds ratio of 10.6 can be adjusted by a factor of 0.76 to provide an estimate of the incidence density ratio of 8.0 (Table 3, Frame C).

Example 2: Neural Tube Defects

Length-biased sampling can be especially problematic in the study of risk factors for congenital malformations [12–15, 22, 28]. In principle, the goal

Table 3 Use of duration-to-date data to estimate duration of disease

Frame A: Hypothetical duration-to-date data:

	Need for an endotracheal tube	
	Present	Absent
Nosocomial infection		
present	7	90
absent	4	544
Duration		
10 days	5	27
5 days	2	63

Frame B: Calculation of distribution of duration of disease:

$p_\mathrm{i}(D) = [\overline{D}_\mathrm{i} \times p_\mathrm{p}(D)]/D$

For exposed cases:

$p_\mathrm{i}(5) = [\overline{D}_\mathrm{i} \times (2/7)]/5 = \overline{D}_\mathrm{i} \times 0.057$

$p_\mathrm{i}(10) = [\overline{D}_\mathrm{i} \times (5/7)]/10 = \overline{D}_\mathrm{i} \times 0.071$

For unexposed cases:

$p_\mathrm{i}(5) = [\overline{D}_\mathrm{i} \times (63/90)]/5 = \overline{D}_\mathrm{i} \times 0.14$

$p_\mathrm{i}(10) = [\overline{D}_\mathrm{i} \times (27/90)]/10 = \overline{D}_\mathrm{i} \times 0.03$

Distribution of duration for incidence series:

exposed cases: 5 days for 45%; 10 days for 55%

unexposed cases: 5 days for 82%; 10 days for 18%

Average duration of disease for incidence series:

exposed cases: $\overline{D}_{\mathrm{i}1} = 7.8$ days

unexposed cases: $\overline{D}_{\mathrm{i}0} = 5.9$ days

Frame C: Calculation of IDR:

$IDR = POR \times \overline{D}_0/\overline{D}_1$

$IDR = 10.6 \times 5.9/7.8 = 8.0$

would be to enroll women from the time of pregnancy and follow all of them through the termination of the pregnancy. All incident cases of malformations could be captured, and etiologic factors could be investigated. In practice, however, the situation is quite different. Women are often enrolled at the time of delivery of live or stillborn infants, so that only *prevalent cases* of malformations are available for study. Malformations at birth represent the *survivors* of early pregnancy, a time during which spontaneous and induced abortions occur, both of which are more likely in the presence of fetal malformations [13, 14, 22, 27, 31, 33]. If the duration of survival (early pregnancy loss versus survival to birth) is associated with an etiologically relevant factor, then the estimate of effect using only birth data will be biased because of either a toxic or a protective effect of the exposure on the fetus during the early prenatal period [13, 14, 22].

Consider the example of neural tube defects (NTDs), which include anencephaly, spina bifida, craniorrhachischisis, and iniencephaly. Leaving aside the issue of early spontaneous abortions [34], over the last two decades prenatal screening for these malformations has increased, and as a result, induced abortions post screening have also increased [30, 33]. Prevalent cases of NTDs at birth may thus be systematically different from incident cases with respect to any etiologic factor that is associated with obtaining prenatal screening and acting on the results of the screen. Velie & Shaw [32] demonstrate the magnitude of this bias by comparing the estimates of effect for a number of variables from a typical birth prevalence case–control study and from a case–control study that is more "incident-like" in that cases from induced abortions are included along with cases ascertained at birth. For illustrative purposes, we consider estimates of the effect of folic acid use on the occurrence of NTDs. Table 4 presents data adapted from Velie & Shaw [32].

Data from the prevalence case–control study show a strong protective effect of folic acid on the prevalence of NTDs, $POR = 0.37$. When the study is broadened to include early fetal deaths due to induced abortions following screening for NTDs that have been excluded from the birth prevalence study, the protective effect of folic acid remains, but it is considerably attenuated, $POR = 0.61$. The reason for this attenuation is that folic acid use is strongly and negatively associated with the odds of surviving early pregnancy and thereby being available for inclusion in the prevalence at birth study ($OR = 0.24$). Folic acid use may be associated with early prenatal care, which itself is associated with prenatal screening; and prenatal screening is associated with termination of affected fetuses.

The availability of data from "incident-like" and prevalence studies from the same population, in which the duration of survival is associated with a number of potential etiologic factors, allows for a demonstration of the way length-biased sampling can bias the estimate of effect. In particular:

1. From Eq. (1), $POR = IDR \times \overline{D}_1/\overline{D}_0$, the duration ratio can be estimated to be 0.61.
2. Alternatively, (4) below (from Freeman & Hutchison [8]) can be used to estimate the duration ratio:

$$\frac{\overline{D}_{i1}}{\overline{D}_{i0}} = \frac{[\pi_p(1)/\pi_i(1)]}{[\pi_p(0)/\pi_i(0)]}, \tag{4}$$

where

$\overline{D}_{i1}$ = the average duration of the disease or condition from a series of incident cases from the exposed group;
$\overline{D}_{i0}$ = the average duration of the disease or condition from a series of incident cases from the unexposed group;
$\pi_p(1)$ = the proportion of exposed cases in a prevalence series;
$\pi_p(0)$ = the proportion of unexposed cases in a prevalence series;
$\pi_i(1)$ = the proportion of exposed cases in an incidence series; and
$\pi_i(0)$ = the proportion of unexposed cases in an incidence series.

Again, the duration ratio would be estimated to be 0.61. From this calculation, the POR of 0.37 can be adjusted by the inverse duration ratio of 1.65 to obtain an estimate of the IDR of 0.61.

For this last sequence of calculations, information about the exposure frequency of the prevalent cases might be obtained from data on hand. A series of potential exposure frequencies of the *un*observed incident cases could be postulated, based on the best available extra-study information. These frequencies could then be used to estimate a range of plausible

Table 4 Prevalence and "incidence-type" data on the relation between folic acid and neural tube defects (NTDs)[a]

Frame A: Prevalent case–control data (cases from still and livebirths only):

	Folic acid use	
	Yes	No
Cases	151	164
Controls	374	149

$POR = 0.37$

Frame B: "Incidence-type" case–control data (cases from electively aborted fetuses as well as still and livebirths):

	Folic acid use	
	Yes	No
Cases	319	207
Controls	374	149

$POR = 0.61$

Frame C: The association of survival and exposure:

	Folic acid use	
	Yes	No
Survival to birth		
Yes	151	164
No	168	43

$OR = 0.24$

[a]Data are adapted from [32, Tables 2 and 3, pp. 476–477].

duration ratios. These ratios could then be used to produce a band of plausible incidence density ratios. The main purpose of this exercise would be to show the sensitivity of the observed POR to length-biased sampling when the potential etiologic factor under investigation could plausibly be related to survival.

Separating Prevalence- and Incidence-Based Studies

We have illustrated some of the relations between estimates based on prevalence data and those based on incidence data when the steady-state assumption holds [7, 8]. Provided exposures are not strongly related to survival, the calculations suggest that qualitative conclusions from prevalence studies will often agree with those from incidence-based data. Experience confirms this expectation. Yet, in "strict logic, ... there is no reason why [incidence based and prevalence based studies] should yield even qualitatively similar results" [5, p. 524].

Neyman illustrates this point with hypothetical data which he describes as "somewhat implausible" [19, p. 404]. Using only prevalent cases to study the relation between smoking and lung cancer, it is possible that a positive association in the prevalent series could mask a protective effect in the population as a whole, if the majority of cases in nonsmokers were highly lethal and therefore did not survive long enough to be included in the prevalence study. However implausible the substance of this example, the point is that in such situations the duration of survival will be strongly associated with the exposure and thereby bias the POR as an estimate of the IDR. In principle, the bias may be so large as to reverse the direction of the effect.

The plausibility of "balancing" or "off-setting" effects of a potential risk factor among a segment of the population that does not survive long enough to be eligible for prevalence studies has been examined in detail in the study of risk factors for congenital malformations [13, 14, 22]. Evidence for potential teratogens typically comes from studies of livebirths, which involve prevalent cases. However, the object of these studies is to discover the potential teratogenic effects on *all* the products of conception, in principle, through incident cases. The results from studies of livebirths may not agree *even qualitatively* with results applicable to all conceptuses. For example, regarding the relation between maternal cigarette smoking and the occurrence of Downs syndrome, Hook & Regal [14] have shown that an apparent protective relative effect as low as 0.3 among livebirths could be consistent with a null effect among all conceptuses if smoking increased the risk of embryonic and fetal deaths by a factor of only 1.1. Regal & Hook [22] provide general formulas for estimating the magnitude of effect on embryonic and fetal deaths that could account for the observed effect in livebirths. Hook & Regal [13] also describe the following variant of the "Yule–Simpson" paradox (*see* **Simpson's Paradox**). The relative exposure effects can be protective among embryonic and fetal deaths and protective among livebirths, yet among all conceptuses the effect is null. Alternately, the exposure can increase the risk both among embryos and fetuses who died and among livebirths, yet among all conceptuses the exposure effect is null. (This situation is statistically similar to confounding, but it is substantively distinct since the "confounding" variable is duration of survival.) Appendix B contains an example of data embodying the "Yule–Simpson" paradox. Hook & Regal [14] develop the concept of a "boundary value", namely a value such that if the observed risk from a suspected teratogen derived from a prevalent series of livebirths exceeds this value, then the explanation of the finding is unlikely to be due solely to selection factors related to the exclusion of cases that did not survive long enough to be included in the prevalence study. The boundary value is a function of the following: the proportion of unexposed subjects that do not survive to be included in the prevalence study; the proportion of unexposed cases that do not survive to be included in the prevalence study; and the proportion of exposed subjects that do not survive to be included in the prevalence study. Where estimates of these quantities are available, a boundary value can be calculated. Tables of boundary values are given for a range of determinants.

Conducting and Interpreting Prevalence-Based Studies

Bias from the inclusion of prevalent cases is a concern when the goal of the study is to discover potential etiologic factors. In this case, if the exposure is related to the duration of survival, then a series of prevalent cases will not mirror the targeted series of incident cases. However, there are situations where etiologic research is not the goal [18]. At times one may be mainly interested in factors that influence disease *prevalence*. The question, for example, may not be, "What factors *cause* infection in neonatal intensive care units (NICUs)?" but rather, "What factors *predict* the number of NICU beds needed to treat the infections that occur?" Concern with length-biased sampling is central to the etiologic question, but irrelevant for estimating factors that influence prevalence. For determining required medical services, the prevalent series is the direct research target; it is not intended to duplicate the incidence series.

The assumption of steady-state conditions is import ant for both descriptive and etiologic research with prevalent cases. If a population is in a steady state, then the size of the prevalence pool is constant. In this case a prevalence-based study of service needs at time t will be applicable to $T > t$, provided the population remains in a steady state. For etiologic research, the steady-state assumption underlies the conversion formulas presented by Freeman & Hutchison [7, 8]. While the assumption of steady-state conditions may well be "unrealistic in a literal sense, it [may well] be approximately true in many applications" [2, p. 194]. For example, in a steady-state population, there should be "a stable pattern in terms of how disease gets diagnosed," [6, p. 1110], and treatment should be reasonably stable. Medical advances that change the way a disease is diagnosed or treated will perturb, at least temporarily, the steady state of the population. Studies done with prevalent cases during the early introduction of AZT for the treatment of AIDS, or with livebirths following the introduction of prenatal screening, can

be misleading. For more detailed arguments regarding the difficulties of studying the risk factors for AIDS in a setting of changing criteria for diagnosis and methods of treatment, see Brookmeyer & Gail [4]. Hook & Regal [12–14, 22] offer cautions concerning the effects of changing prenatal practices on the study of risk factors for congenital malformations.

When the assumption of a steady-state population is plausible, the formulae from Freeman & Hutchison [7, 8] can be used to convert information obtained from prevalence studies to estimates that would result from incidence studies from the same population. The formulas are applicable where exposure–outcome relations are either unconfounded or can be conveniently stratified on confounders. Begg & Gray [2] have developed **multivariate** methods and proposed **quasi-likelihood** estimates for bias correction that can be obtained using GLIM with a log link function and a **gamma** error distribution. (When the exposure is dichotomous, and there is no confounding, the correction factor proposed by Begg & Gray reduces to the inverse duration ratio as described by Freeman & Hutchison [7, 8].)

When a population is in a steady state (or stationary), the size of the candidate pool and the prevalence pool are constant, and the age distribution of the population is stable [1, 2, 7, 8, 16, 17, 21, 24]. The population, or segment of the population considered, is not growing. Alho [1] considered the basic relation examined by Freeman & Hutchison [7, 8] and Begg & Gray [2] in the more general situation where the size of the population is either increasing or decreasing, that is, a situation where the population is "stable" but is not "stationary" or in a "steady state". Under this situation, when, for example, incidence is increasing with age and duration of disease is declining with age, the standard relation of the prevalence odds as the product of incidence density and average duration of disease will be an overestimate. Instead, the prevalence odds will equal "a weighted average of the age specific products of incidence and (discounted) expected duration" [1, p. 587] (*see* **Incidence–Prevalence Relationships**).

Suppose that a population (or a segment of a population) can be assumed to approximate a steady state, such that the relations described by Freeman & Hutchison [7, 8] and Begg & Gray [2] hold. In this case inferences about incidence densities can be made from either incidence- or prevalence-based studies. Begg & Gray [2] have compared the relative efficiency of prevalence studies to incidence studies, where the determinants of efficiency included the proportion of cases in the sample, the exposure frequency among controls, the true relative effect, and the duration of disease ratio. There was substantial variation in the relative efficiency of the prevalence-based study compared with the incidence-based study over a wide range of these determinant values. However, the authors conclude that

> for less extreme configurations [of the determinants of efficiency], the relative efficiency ranges from about 50% to 90%, so that an approximate rule of thumb is that on average, to achieve comparable precision, a prevalence study will require about three subjects for every two in an incidence study [2, p. 194].

While a prevalence-based study should be larger than the corresponding incidence-based study, it may be easier and less costly to enroll prevalent cases. However, the concerns about the precision of the study should always be subordinated to concerns about the validity of the study. Moreover, the validity of the study based on prevalent cases and aimed at etiologic investigation will depend on two factors: (i) the feasibility of obtaining accurate duration-to-date information on the duration from the onset of disease to the enrollment in the prevalence study, and (ii) the plausibility that the population (or population segment) approximates the steady-state requirements.

Conclusion

Examples from research on plausible teratogens highlight the importance of considering the possibility that prevalent studies can yield misleading estimates of exposure effects on incident disease. A comment from Sartwell & Merrell [26, p. 583] broadens this concern beyond studies that are traditionally described as "prevalence based":

> While the foregoing illustrations have dealt with limitations of prevalence and mortality rate, the interpretation of observed incidence rates also offers a challenge. What is actually obtained may more properly be termed a discovery rate, which may be quite different from the true incidence rate. Of the

cases with onset in any time period, some will be discovered early, others late, perhaps at death, and still others will escape recognition entirely. Thus, the discovery rate of any period will include cases with onsets covering a long time span, and moreover will be influenced by the clinical level at which cases are recognized.

Sartwell & Merrell [26] remind us that even a typical incidence series may well contain a trace of prevalence. To the extent that any survival (or persistence) criterion is required to allow the diagnosis of a disease, estimates of exposure effect may be distorted, as in the prevalent case–control study.

Appendix A: Derivation of $PO = ID \times \overline{D}$

Prevalence (P), incidence density (ID), and average duration ($\overline{D}$) are readily linked when the population is in a "steady state" [1, 2, 7, 8, 16, 17, 21, 24]. In a steady state the number of people entering the prevalence pool from the population at risk is equal to the number of people exiting the prevalence pool, through recovery or death.

At any time, the total population (N) can be divided into those individuals who have the disease (P) and those who are at risk for the disease ($N - P$). Entry into the prevalence pool is governed by the size of the population at risk ($N - P$), the rate of disease (ID), and the time period (δt) such that

$$\text{inflow} = ID \times \delta t \times (N - P).$$

Exit from the prevalence pool is governed by the size of the prevalence pool (P), the exit rate (ID_{exit}), and the time period (δt) such that

$$\text{outflow} = ID_{\text{exit}} \times \delta t \times P.$$

In a steady state, any rate is equal to the reciprocal of the average time to the event. Therefore, $ID_{\text{exit}} = 1/\overline{D}$, where $\overline{D}$ is the average duration of disease. In a steady state, inflow (to the prevalence pool) equals outflow (from the prevalence pool). Therefore

$$ID \times \delta t \times (N - P) = \frac{1}{\overline{D} \times \delta t \times P}.$$

By algebraic manipulation:

$$ID \times \overline{D} = \frac{P}{(N - P)} = PO.$$

Appendix B: Hypothetical Data Embodying the "Yule–Simpson" Paradox

The data in this appendix are from [13, Table 2, p. 56].

Stratum 1: Embryonic and fetal deaths

	Exposed	Unexposed	Total
Downs syndrome	732	1251	1983
Unaffected	58 359	89 658	148 017
All	59 091	90 909	150 000

Relative risk of Downs syndrome = 0.9

Stratum 2: Livebirths

	Exposed	Unexposed	Total
Downs syndrome	213	637	850
Unaffected	274 029	575 121	848 150
All	274 242	575 758	850 000

Relative risk of Downs syndrome = 0.7

Total population: All (recognized conceptuses)

	Exposed	Unexposed	Total
Downs syndrome	945	1888	2833
Unaffected	332 388	663 834	997 167
All	333 333	666 667	1 000 000

Relative risk of Downs syndrome = 1.0

References

[1] Alho, J.M. (1992). On prevalence, incidence and duration in general stable populations, *Biometrics* **48**, 587–592.

[2] Begg, C.B. & Gray, R.J. (1987). Methodology for case–control studies with prevalent cases, *Biometrika* **74**, 191–195.

[3] Borman, B. & Cryer, C. (1990). Fallacies of international and national comparisons of disease occurrence in the epidemiology of neural tube defects, *Teratology* **42**, 405–412.

[4] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, New York.

[5] Cornfield, J. & Haenszel, W. (1960). Some aspects of retrospective studies, *Journal of Chronic Diseases* **11**, 525–534.

[6] Dunn, J.E. (1962). The use of incidence and prevalence in the study of disease development in a population, *American Journal of Public Health* **52**, 1107–1118.

[7] Freeman, J. & Hutchison, G.B. (1980). Prevalence, incidence, and duration, *American Journal of Epidemiology* **112**, 707–723.

[8] Freeman, J. & Hutchison, G.B. (1986). Duration of disease, duration indicators, and estimation of the risk ratio, *American Journal of Epidemiology* **124**, 134–149.

[9] Freeman, J. & McGowan, J.E., Jr (1978). Risk factors for nosocomial infection, *Journal of Infectious Diseases* **138**, 811–819.

[10] Freeman, J. & McGowan, J.E., Jr (1981). Day-specific incidence of nosocomial infection estimated from a prevalence survey, *American Journal of Epidemiology* **114**, 888–901.

[11] Freeman, J., Rosner, B.A. & McGowan, J.E. Jr (1979). Adverse effects of nosocomial infection, *Journal of Infectious Diseases* **140**, 732–740.

[12] Hook, E.B. (1982). Incidence and prevalence as measures of the frequency of birth defects, *American Journal of Epidemiology* **116**, 743–747.

[13] Hook, E.B. & Regal, R.R. (1991). Conceptus viability, malformation, and suspect mutagens or teratogens in humans: the Yule–Simpson paradox and implications for inferences of causality in studies of mutagenicity or teratogenicity limited to human livebirths, *Teratology* **43**, 53–59.

[14] Hook, E.B. & Regal, R.R. (1993). Representative and misrepresentative associations of birth defects in livebirths. Conditions under which relative risks greater than unity in livebirths necessarily imply relative risks greater than unity in all conceptuses, *American Journal of Epidemiology* **137**, 660–675.

[15] Khoury, M.J., Flanders, W.D., James, L.M. & Erickson, J.D. (1989). Human teratogens, prenatal mortality, and selection bias, *American Journal of Epidemiology* **130**, 361–370.

[16] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont.

[17] Miettinen, O.S. (1976). Estimability and estimation in case-referent studies, *American Journal of Epidemiology* **103**, 226–235.

[18] Miettinen, O.S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. Wiley, New York.

[19] Neyman, J. (1955). Statistics–servant of all sciences, *Science* **122**, 401–406.

[20] Pearce, N. (1993). What does the odds ratio estimate in a case–control study?, *International Journal of Epidemiology* **22**, 1189–1192.

[21] Preston, S.H. (1987). Relations among standard epidemiologic measures in a population, *American Journal of Epidemiology* **126**, 336–345.

[22] Regal, R.R. & Hook, E.B. (1992). Interrelationships of relative risks of birth defects in embryonic and fetal deaths, in livebirths, and in all conceptuses, *Epidemiology* **3**, 247–252.

[23] Rodrigues, L. & Kirkwood, B.R. (1990). Case–control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls, *International Journal of Epidemiology* **19**, 205–213.

[24] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company, Boston.

[25] Rozencweig, M., Zelen, M., Von Hoff, D.D. & Muggia, F.M. (1978). Waiting for a bus: does it explain age-dependent differences in response to chemotherapy of early breast cancer?, *New England Journal of Medicine* **299**, 1363–1364.

[26] Sartwell, P.E. & Merrell, M. (1952). Influence of the dynamic character of chronic disease on the interpretation of morbidity rates, *American Journal of Public Health* **42**, 579–584.

[27] Seller, M.J. (1987). Unanswered questions on neural tube defects, *British Medical Journal* **294**, 1–2.

[28] Sever, L.E. (1983). Re: "Incidence and prevalence as measures of the frequency of birth defects", *American Journal of Epidemiology* **118**, 608–609.

[29] Simon, R. (1980). Length biased sampling in etiologic studies, *American Journal of Epidemiology* **111**, 444–452.

[30] Slattery, M.L. & Janerich, D.T. (1991). The epidemiology of neural tube defects: a review of dietary intake and related factors as etiologic agents, *American Journal of Epidemiology* **133**, 526–539.

[31] Stein, Z., Susser, M., Warburton, D., Wittes, J. & Kline, J. (1975). Spontaneous abortion as a screening device: the effect of fetal survival on the incidence of birth defects, *American Journal of Epidemiology* **102**, 275–290.

[32] Velie, E.M. & Shaw, G.M. (1996). Impact of prenatal diagnosis and elective termination on prevalence and risk estimates of neural tube defects in California, 1989–1991, *American Journal of Epidemiology* **144**, 473–479.

[33] Wald, N.J. (1984). Neural-tube defects and vitamins: the need for a randomized clinical trial, *British Journal of Obstetrics and Gynaecology* **91**, 516–523.

[34] Wilcox, A.J., Weinberg, C.R., O'Connor, J.F., Baird, D.D., Schlatterer, J.P., Canfield, R.E., Armstrong, E.G. & Nisula, B.C. (1988). Incidence of early loss of pregnancy, *New England Journal of Medicine* **319**, 189–194.

(*See also* **Bias in Case–Control Studies**; **Biased Sampling of Cohorts**; **Cross-sectional Study**)

JANET M. LANG

Case–Control Study, Sequential

Many **case–control studies** are based on previously identified cases and **controls**, together with their information on exposure and **confounders**. Some studies, however, require that data on newly **incident cases** and controls be collected and accumulated prospectively, not unlike the data collection for a controlled **clinical trial**. For example, when a new drug or exposure is introduced into a population, there may be no medical database (*see* **Administrative Databases**) or method to link records (*see* **Record Linkage**) to determine exposure and case status. When a concern exists about risk associated with these exposures and a case–control design is capable of addressing the issue, the conduct of a case–control study to evaluate and test the hypothesis requires the accumulation of new records and exposure information.

In these situations an investigator may wish to test the hypothesis (*see* **Hypothesis Testing**) repeatedly as the data accumulate in order to reduce the average sample size required by the study and obtain results sooner. Sequential case–control designs (*see* **Sequential Analysis**) are proposed first, to take account more efficiently of the accumulating collection of exposure data on cases and controls and, thus, decide when sufficient data are in hand to address the question and, secondly, to preserve the integrity of the **inferences** drawn from case–control studies when these studies are repeatedly analyzed as data accumulate.

The statistical methods developed for monitoring randomized clinical studies (*see* **Data and Safety Monitoring**) can be applied to the analysis of accumulating data in matched or unmatched case–control designs, and to the repeated analyses of measures of **association**, such as the **odds ratio** or **relative risk**. In a clinical trial, ethical concerns usually drive the need to terminate a study early in order to minimize exposure of study subjects to harmful treatments (*see* **Ethics of Randomized Trials**). The ethical concerns are different for case–control studies, but there may be ethical motivations for trying to terminate a case–control early. For example, a case–control study may be used to identify or confirm an important risk that requires expeditious public health policy decisions.

Sequential methods can also be used to continue accumulating data until the odds ratio has been estimated with acceptable precision. In this article, we concentrate on hypotheses testing, though we discuss applications to sequential **estimation**.

Implementing the Sequential Approach

The following discussion is based upon O'Neill & Anello [13]. Assume that an investigator is interested in studying whether or not there is an increased risk of an adverse outcome (cases) associated with exposure to a specified factor. Assume that there is a mechanism for uniformly identifying and collecting cases with the adverse outcome from a defined population and then, according to predetermined criteria, **matching** each case with respect to a matching variate to a control (person without the adverse outcome) at the time the case becomes available. Furthermore, assume that there is a mechanism for identifying in an unbiased manner the exposure to a risk factor for both case and matched controls. Thus, case and control matches become available over time, and at any point in time T a cumulative set of N_T cases and their matched controls is available. While it is not necessary that all cases be matched to the same number of controls, the example we provide assumes a constant matching ratio for ease of exposition. We will also assume that the relative risk is constant and independent of the matching variables and/or other covariables.

For one-to-one matching, the information is in the case–control pairs that are discordant with respect to the presence or absence of the risk factor. For multiple matching on **confounding** variables, the situation is somewhat more complex (e.g. [6]).

We illustrate two sequential designs with one-to-one matching. The first design monitors outcomes after each case–control pair is ascertained and uses the Wald [21] sequential probability ratio test (SPRT). The second design analyzes the data after successive groups of matched pairs have been accrued and is based on the theory of repeated significance testing for group sequential data. The group sequential approach can also be used for multiple-matched and unmatched case–control designs, as discussed in Pasternack & Shore [14, 15].

The Wald SPRT for a One-to-One Matched Design

For the ith case–control pair, let Y_{11} denote the exposure value (1 if exposed, 0 otherwise) assumed by the case and Y_{21} denote the value assumed by the control. For the ith case–control pair, let $P_{1i} = \Pr(\text{case is exposed to the factor under study})$ and $P_{2i} = \Pr(\text{control is exposed to the factor under study})$. The data from N matched case–control pairs may be summarized as in Table 1, where Z_{11} denotes the number of pairs in which both case and control are exposed, Z_{10} the number of pairs where only the case is exposed, and so forth.

In an individually matched retrospective case–control study, the **maximum likelihood** estimate of relative risk ρ is based upon the pairs with discordant exposure and is given by $\rho = Z_{10}/Z_{01}$. Thus, the effective sample size for **interval estimation** of ρ is $Z_{10} + Z_{01} = n$. In testing whether this estimate of the relative risk differs significantly from one, attention is restricted to the discordant pairs. In the ith such pair, the probability that the case is exposed and the control is not exposed is

$$\pi_i = \frac{P_{1i}(1 - P_{2i})}{P_{1i}(1 - P_{2i}) + (1 - P_{1i})P_{2i}}. \qquad (1)$$

The relative risk is $\rho = \pi/(1 - \pi)$ and is assumed constant for all pairs. The test of whether ρ differs significantly from one is equivalent to testing whether the conditionally **binomial** variate Z_{10}, based on n observations, has probability $\pi = \frac{1}{2}$ against an alternative that $\pi \neq \frac{1}{2}$.

To carry out the SPRT, it is essential that one specifies two simple hypotheses before the data are collected. That is, one specifies that the **null hypothesis** is $\rho_0 = 1$ and chooses a value of ρ for the alternative hypothesis H_1, call it ρ_1, which is of interest to detect. Furthermore, one must specify the type I and type II errors, α and β, respectively. Since we choose to examine the one-sided test in which the alternative of the form $\rho > 1$ is of interest (*see* **Alternative Hypothesis**), the SPRT tests the null hypothesis that $H_0 : \rho = 1$ vs. $H_1 : \rho = \rho_1$, which is equivalent to $H_0 : \pi = \frac{1}{2}$ vs. $H_1 : \pi_1 = \rho_1/(1 + \rho_1)$.

Table 1 Summary of data from N matched case–control pairs

		Control (Y_2)		
		1	0	Total
Case (Y_1)	1	Z_{11}	Z_{10}	Z_1
	0	Z_{01}	Z_{00}	$N - Z_1$
	Total	Z_2	$N - Z_2$	N

The SPRT is carried out as each case–control pair becomes available over time. The cumulative number of cases exposed to the factor, Z_{10}, among the subset of n exposure-discordant pairs is then compared with two parallel boundaries (see [13, Appendix A]). Crossing one boundary causes rejection of H_0 and crossing the other causes rejection of H_1.

The Group Sequential Design for a One-to-One Matched Design

Rather than analyze the data after each matched pair accrues, one can perform only K analyses after successive groups of M discordant pairs have accrued, giving rise to a maximum total sample size of $KM = N$ discordant pairs. Pasternack & Shore [16] described this group sequential approach for repeated tests of a hypothesis. The approach is based on the sequential use of the signed square root of the **McNemar test** statistic χ for accumulating discordant case–control pairs, namely

$$\chi_k^2 = \frac{(|Z_{10} - Z_{01}| - 1)^2}{Z_{10} + Z_{01}}. \qquad (2)$$

While we emphasize this application, one could also use the estimate of the log odds ratio from the conditional logistic model (*see* **Logistic Regression, Conditional**) divided by its **standard error**.

Letting χ_k denote the signed square root of (2) after k groups, the decision rule for the kth ($1 \leq k \leq K$) group sequential test of $\rho = 1$ vs. $\rho \neq 1$ is

1. If $\chi_k > Z_\alpha$ reject the null hypothesis that $\rho = 1$ in favor of the alternative, $\rho \neq 1$; then no additional accumulation of data is required.
2. (i) If $\chi_k < Z_\alpha$ and $k < K$, accumulate additional discordant pairs and then apply the $(k + 1)$th group sequential test.
   (ii) If $\chi_k < Z_\alpha$ and $k = K$, end the study and do not reject the null hypothesis that $\rho = 1$.

Note that Z_α is a constant in this "repeated testing design". To control the overall type I error at α, Z_α at each interim analysis will need to be adjusted.

For example, for planned three-group sequential analyses controlled at $\alpha = 0.05$, $Z_\alpha = 2.289$, corresponding to a nominal $\alpha' = 0.022$ for each analysis (see Pocock [17, Table 2]). Other group sequential designs, referred to in the discussion section, allow for varying the boundary with k.

A Comparison of Sample Sizes for Fixed and Sequential Designs

One of the main benefits of a sequential design is that fewer observations are needed on average than for a fixed sample size plan to reject the null hypothesis when the alternative hypothesis is true, thereby being a more efficient design by offering the chance of yielding a conclusion before the final fixed sample size is needed. The difference in sample size needed for the fixed and sequential designs for the one-to-one individually matched design method was described by O'Neill & Anello [13] for the SPRT, and by Pasternack & Shore [16] for the group sequential designs using repeated significance-testing methods.

For the SPRT, formulas exist [13, Appendix A] to calculate the average number of exposure discordant pairs needed to arrive at a decision when the null and alternative hypotheses are assumed true and for a value of ρ; namely, $\overline{\rho}$, midway between ρ_0 and ρ_1. These sample sizes can be compared with those needed in a fixed sample design. Table 2, adapted from O'Neill & Anello [13], presents the average number of exposure discordant pairs required in a SPRT and fixed sampling design for selected protocol parameters. The SPRT offers a substantial advantage in average required sample size, although there is no upper limit on the sample size.

For the group sequential designs based upon repeated significance testing, one can calculate the maximum sample size N needed and the average sample size $\overline{N}$ needed for a prespecified number, K, of interim analyses of accruing case–control exposure discordant pairs. This calculation depends on the number of interim analyses, K, or independent groups of exposure discordant pairs, planned in advance, and on the per-test significance level α' and its corresponding **standard normal deviate** Z for use in normal group sequential testing ([16, Table 2]). Table 3, extracted from Pasternack & Shore [16], who provide further discussion of the calculations, presents a comparison of the number of discordant pairs needed for the group sequential design relative to a fixed sample plan for a two-sided hypothesis test. For example, for $\rho = 2$, the fixed sample plan ($K = 1$) requires $N = 91$ pairs, compared with an average of only 64 pairs with $K = 4$. Note, however, with $K = 4$, the maximum trial size, 108, exceeds the fixed sample size, 91.

Table 2 Average number of exposure discordant pairs for SPRT and fixed sample designs, for selected relative risk $\rho : \alpha = 0.05, \beta = 0.10$ (one-sided)[a]

	$\rho = 2$			$\rho = 3$		
Fixed plan	73			30		
SPRT	H_0	H_1	H_ρ^-	H_0	H_1	H_ρ^-
	34	42	56	14	18	23

[a]Extracted from O'Neill & Anello [13]

Table 3 Comparison of the number of discordant pairs needed for the group sequential design relative to a fixed sample plan for a two-sided hypothesis test: $\alpha = 0.05$, $\beta = 0.10$ (two-sided). N = Maximum number of discordant pairs per group; $\overline{N}$ = Average number of discordant pairs per group[a]

Number of interim analyses (groups)	$\rho = 2$		$\rho = 3$	
K	N	$\overline{N}$	N	$\overline{N}$
1	91	91	38	38
2	100	71	42	30
3	105	66	45	28
4	108	64	44	27

[a]Extracted from Pasternack & Shore [16].

In Tables 2 and 3, the sample sizes presented for the average number of discordant pairs using the SPRT, the group sequential design, and the fixed sample design need to be adjusted to arrive at the expected total number of pairs required. To make the adjustment, one needs to divide the sample sizes in Tables 2 and 3 by the probability of obtaining a discordant pair, which is

$$Y = P_1(1 - P_2) + P_2(1 - P_1). \tag{3}$$

values for which are provided in [13, Table 3].

Discussion

To date, there are few examples of sequentially designed and analyzed case–control studies, though there are many examples of prospective collection

of exposure information on cases and controls that could easily adapt the sequential strategy. Several variants of the group sequential approach can be used for monitoring case–control studies. Rather than using the same boundary (i.e. the same nominal Type 1 error) at each interim analysis as proposed by Armitage et al. [3] and Pocock [17], we can use different critical levels for each k. For example, O'Brien & Fleming [11] proposed a monotonically decreasing set of nominal α levels, and Lan & DeMets [10] introduced a flexible method for allocating the Type 1 error over the number of testing times used. Rather than applying group sequential statistic methods to simple statistics such as the McNemar statistic, one can also use **score** statistics derived from logistic models in sequential tests, as described by Whitehead [22].

Despite the attractiveness of the group sequential design for reducing required sample sizes when performing hypothesis testing in case–control studies, we are unaware of completed case–control studies that formally use this methodology.

Only in unusual circumstances will there be a pressing need to terminate an **observational study** early, unlike a clinical trial. It may be important to obtain a large number of cases and controls to pursue analyses of the effects of confounding. Issues of data quality and case ascertainment may play a role also. For example, it may be necessary to validate exposure assessments and histopathologic diagnoses of disease. Such quality control activities may be logistically difficult to accomplish during the conduct of the sequential acquisition of cases and controls. These factors may make it more attractive to design and conduct a large fixed sample size study rather than to attempt a sequential design. However, one should be aware that if one initiates a fixed sample design but terminates the study early on the basis of interim analyses, one risks increasing the chance of type I error.

More often, an epidemiologist may be interested in obtaining a precise estimate of a risk parameter rather than in rejecting a null hypothesis expeditiously. To the extent that precise and valid estimation of exposure effects is the dominant goal of the observational study, sequential strategies may play a greater role in the future by requiring that data continue to be accrued until precise estimates of exposure risks are obtained.

Generally, sample sizes to test a hypothesis will be inadequate to provide a **confidence interval** for the odds ratio whose width is sufficiently narrow for precise estimation [12]. Many authors have considered sequential estimation procedures in a univariate and **multivariate** context, which may be relevant for the medical and public health applications for which case–control studies have found utility [1, 2, 4, 5, 7–9, 18–20].

References

[1] Anscombe, F.J. (1952). Large-sample theory of sequential estimation, *Proceedings of the Cambridge Philosophical Society* **48**, 600–607.

[2] Anscombe, F.J. (1953). Sequential estimation, *Journal of the Royal Statistical Society, Series B* **15**, 1–21.

[3] Armitage, P., MacPherson, C.K. & Rowe, B.C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

[4] Cabilio, P. (1977). Sequential estimation in Bernoulli trials, *Annals of Statistics* **5**, 342–356.

[5] Chow, Y.S. & Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean, *Annals of Mathematical Statistics* **36**, 457–462.

[6] Connett, J., Ejigou, A., McHugh, R. & Breslow, N. (1982). The precision of the Mantel–Haenszel estimator in case–control studies with multiple matching, *American Journal of Epidemiology* **116**, 875–877.

[7] Gleser, L.F. (1966). On the asymptotic theory of fixed size sequential confidence bounds for linear regression parameters, *Annals of Mathematical Statistics* **36**, 463–467. (See correction note: *Annals of Mathematical Statistics* **36** (1966) 1053–1055.)

[8] Grambsch, P. (1983). Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator, *Annals of Statistics* **11**, 68–77.

[9] Jennison, C. & Turnbull, B.W. (1993). Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses, *Biometrics* **49**, 31–44.

[10] Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–662.

[11] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[12] O'Neill, R.T. (1983). Sample size for estimation of the odds ratio in unmatched case–control studies, *American Journal of Epidemiology* **120**, 145–153.

[13] O'Neill, R.T. & Anello, C. (1978). Case–control studies: a sequential approach, *American Journal of Epidemiology* **108**, 415–424.

[14] Pasternack, B.S. & Shore, R.E. (1980). Sample sizes for group sequential cohort and case–control study designs, *American Journal of Epidemiology* **113**, 778–784.

[15] Pasternack, B.S. & Shore, R.E. (1980). Group sequential methods for cohort and case–control studies, *Journal of Chronic Diseases* **33**, 365–373.

[16] Pasternack, B.S. & Shore, R.E. (1981). Sample sizes for individually matched case–control studies, *American Journal of Epidemiology* **115**, 778–784.

[17] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.

[18] Robbins, H. & Sigmund, D.O. (1974). Sequential estimation of p in Bernoulli trials, in *Studies in Probability and Statistics*, E.J. Williams, ed. University of Melbourne.

[19] Srivastava, M.S. (1967). On fixed width confidence bounds for regression parameters and mean vector, *Journal of the Royal Statistical Society, Series B* **29**, 132–140.

[20] Srivastava, M.S. (1971). On fixed width confidence bounds for regression parameters, *Annals of Mathematical Statistics* **42**, 1403–1411.

[21] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

[22] Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials*. Ellis Horwood, Chichester.

ROBERT T. O'NEILL

Case–Control Study, Two-phase

Double sampling, also known as two-phase sampling, is a standard technique for **stratification** [23]. The investigator first draws a **random sample** from the population to measure the **covariates** needed for stratification. At phase two, random subsamples of varying size are drawn from within each stratum and the collection of data is completed for subjects selected at both phases. By using larger sampling ratios for the most informative strata, the efficiency of estimates of population parameters is enhanced.

The **case–control study** embodies a stratified sampling design where the strata depend on the outcome [5]. Case–control studies in epidemiology typically sample a large fraction of the **incident cases** of disease and a much smaller fraction of disease-free **controls** to evaluate the association between disease outcome and risk factors. This design is much more efficient than alternative **cohort** or **cross-sectional designs** for the study of a rare disease.

More complex double-sampling techniques offer the potential to enhance efficiency further and to reduce cost. White [37] proposed studying the **association** between a rare exposure and a rare disease by sampling at phase two on the basis of *both* disease and exposure status. She noted that the initial sample might itself be stratified by outcome, as in a case–control study, or not, as in a cohort or cross-sectional study. Another example of double sampling arises from studies of gene-environment interaction [1]. Efficiency is enhanced by limiting expensive genotyping to a sample stratified both by disease and by family history or a rare environmental exposure. A third example is the **validation study**. Here subsamples of cases and controls are drawn to make error-free measurements, so that parameter estimates may be adjusted for the attenuation caused by **measurement error** [9]. In all three examples, subjects not sampled for phase two have a portion of their data missing by design. There is a strong connection with the literature on **missing data** [21, 28].

This article reviews double sampling techniques for the study of **binary** outcomes, with particular emphasis on methods of fitting **logistic regression** models that appropriately utilize the data from both phases of sampling.

Stratified Sampling

Under the usual **superpopulation model**, the population from which subjects are sampled at phase one is regarded as infinite. Let S denote a **random variable** defined on this population whose values $S = j$ indicate the stratum and set $\pi_j = \Pr(S = j)$, $j = 1, \ldots, J$. Suppose for the moment that the object is to estimate the **mean** value, μ, of another random variable, U, and note that $\mu = \sum_j \pi_j \mu_j$, where $\mu_j = \mathrm{E}(U|S = j)$. Assuming that the π_j are known, appropriately **stratified sampling** yields a more informative estimate than does a **simple random sample** of like size [14].

Double Sampling for Stratification

Often the information needed for classification of population units into strata is not known in advance. Neyman [23] developed the theory of double sampling to handle this problem. At the first phase of sampling one draws a simple random sample of size N for stratum ascertainment and observes N_j, the number of sampled subjects, with $S = j$. At the second phase, subsamples of specified size n_j are drawn at random and without replacement from among the N_j in stratum j, for a total sample size of $n = \sum_j n_j$. Denote by U_{jk} the value of U for the kth subject in stratum j, $k = 1, \ldots, n_j$; by $f_j = n_j/N_j$ the known sampling fraction; and by $\overline{U}_j = n_j^{-1} \sum_k U_{jk}$ the stratum specific sample mean. With $\hat{\pi}_j = N_j/N$ and $\hat{\mu}_j = \overline{U}_j$, an obvious estimate of μ is

$$\hat{\mu} = \sum_{j=1}^{J} \hat{\pi}_j \hat{\mu}_j = N^{-1} \sum_{j=1}^{J} f_j^{-1} \sum_{k=1}^{n_j} U_{jk}. \qquad (1)$$

The Horvitz–Thompson Estimator

When regarded as an estimator of the (unknown) finite population mean of the N values of U for subjects sampled at phase one, $\hat{\mu}$ weights each of the n phase two observations by the inverse of its selection probability. This is the defining property of the famous **Horvitz–Thompson [17] estimator**. The **variance** as given in standard texts [11] is

$$\mathrm{var}(\hat{\mu}) = \frac{1}{N} \sum_{j=1}^{J} \pi_j (\mu_j - \mu)^2 + \sum_{j=1}^{J} \frac{\pi_j^2 \sigma_j^2}{n_j}$$

$$+ \frac{1}{N} \sum_{j=1}^{J} \frac{\pi_j (1 - \pi_j) \sigma_j^2}{n_j}, \qquad (2)$$

where $\sigma_j^2 = \text{var}(U|S = j)$. The first two terms dominate under an asymptotic scheme where each of the n_j increases proportionately with N. For fixed n, efficiency is enhanced by oversampling large strata or those with large σ_j^2 since this will reduce the value of the middle term in (2).

Weighted Likelihood and the Horvitz–Thompson Estimator

In case–control studies the parameters of interest are not the mean values of random variables but rather the **regression** coefficients in probability models for the association between a binary outcome variable Y and a vector $\mathbf{X}$ of explanatory variables. Suppose that

$$\begin{aligned} \Pr(Y = 1|\mathbf{X} = \mathbf{x}, S = j) &= \Pr(Y = 1|\mathbf{X} = \mathbf{x}) \\ &= F(\mathbf{x}'\beta), \end{aligned} \qquad (3)$$

where F denotes a known cumulative distribution function. Typical choices for F are the **logistic distribution** for logistic regression and the unit **normal distribution** for probit regression. The assumption of conditional independence between Y and S given $\mathbf{X}$ means simply that any dependence of the outcome probability on stratum is modeled in $\mathbf{X}$.

Under standard regularity conditions, **likelihood** theory tells us that β satisfies the expected score equation

$$\mu(\beta) = \text{E}[\mathbf{U}(\beta)] = \text{E}\left[\frac{\partial \log \Pr(Y|\mathbf{X}; \beta)}{\partial \beta}\right] = 0, \quad (4)$$

where $\mathbf{U}(\beta)$ is defined as the term in brackets. Even if (3) does not hold, the solution to (4) defines a parameter of interest since it identifies that member of a set of hypothesized models that best describes the population association. If observations on $(Y_t, \mathbf{X}_t)$ were available for all N subjects sampled at phase one, then β would be estimated by solving the score equations $\mathbf{U}(\beta) = \sum_{t=1}^{N} \mathbf{U}_t(\beta) = 0$. In a two-phase study, the unknown $\mathbf{U}$ is estimated using (1). Thus, following Whittemore [38], we define the Horvitz–Thompson estimator, $\hat{\beta}$, as the solution to the Horvitz–Thompson estimating equations

$$\hat{\mathbf{U}}(\beta) = \sum_{j=1}^{J} f_j^{-1} \sum_{k=1}^{n_j} \mathbf{U}_{jk}(\beta) = 0, \qquad (5)$$

where $\mathbf{U}_{jk}(\beta)$ denotes the **score** for the kth subject sampled from stratum j. The **consistency** and asymptotic normality of $\hat{\beta}$ follow from Huber's [19] theory of M-estimation. The asymptotic variance is given by the "**information** sandwich"

$$\begin{aligned} \text{var}_A(\hat{\beta}) = &\left[\frac{\partial \hat{\mathbf{U}}(\beta)}{\partial \beta}\right]^{-1} \\ &\times \text{var}_A[\hat{\mathbf{U}}(\beta)] \left\{\left[\frac{\partial \hat{\mathbf{U}}(\beta)}{\partial \beta}\right]'\right\}^{-1} \Bigg|_{\beta=\hat{\beta}}, \end{aligned} \qquad (6)$$

where the middle expression is obtained by extending (2) to vector-valued random variables.

The Horvitz–Thompson estimator and its analog have a long history. In econometrics it is known as the weighted exogenous sampling **maximum likelihood** estimator [22]. A finite population version was studied by Binder [4] and Chambless & Boyle [12]. Others have referred to the weighted likelihood method as **pseudo-likelihood** [15, 20], and mean score [25].

Two-phase Case–Control Studies

The distinguishing feature of a case–control study is that the variables used for stratification and sample selection include the outcome (case–control) status. Were this not the case, and the sample was stratified using explanatory variables alone, standard methods of binary regression analysis could be used to estimate the regression coefficients in the model (3). Such methods generally yield biased estimates, however, when applied to case–control data. Horvitz–Thompson **estimation** solves the **bias** problem.

Population-based Case–Control Samples

Suppose the population from which the cases and controls are sampled has been completely enumerated and that disease status Y and stratum S are known for everyone. This (finite) population may itself then be regarded as the preliminary random sample in the two phase design. Since the sampling fractions are known for both cases and controls, no restrictions need be placed on F and both **absolute** and **relative risk** parameters may be estimated [3].

To apply the Horvitz–Thompson estimator to **population-based case–control studies**, we extend the notation to allow the phase one strata to depend on both Y and S. Thus, denote by N_{1j} the number of cases ($Y = 1$) and by N_{0j} the number of controls ($Y = 0$) with $S = j$ at phase one, and let n_{ij} denote the numbers in the corresponding subsample at phase two. The Horvitz–Thompson estimator is the solution to

$$\hat{\mathbf{U}}(\beta) = \sum_{i=0}^{1} \sum_{j=1}^{J} f_{ij}^{-1} \sum_{k=1}^{n_{ij}} \mathbf{U}_{ijk}(\beta) = 0, \qquad (7)$$

where now $f_{ij} = n_{ij}/N_{ij}$ and $\mathbf{U}_{ijk}(\beta)$ denotes the likelihood score for the kth subject in stratum (i, j).

Separate Preliminary Samples of Cases and Controls

More typically, a complete enumeration of the population at risk is not available, and one simply samples controls from the same communities or hospital service areas in which the cases arose. Then the appropriate model treats the phase one subjects as separate random samples of N_1 cases and N_0 controls. Since the case and control sampling fractions are not known, the only quantities that may be estimated **consistently** are **odds ratio** (**relative risk**) parameters in "multiplicative intercept" models [18]. Attention is confined here to the logistic model $F(x) = 1/(1 + \mathrm{e}^{-x})$. Provided that the linear predictor $x'\beta$ includes a constant term β_0, the remaining βs represent log relative risks that are in principle estimable from the case–control sample. The logistic scores appearing in (7) are given by $\mathbf{U}_{ijk}(\beta) = \{Y_{ijk} - 1/[1 + \exp(-\mathbf{X}'_{ijk}\beta)\}\mathbf{X}_{ijk}$.

Assume momentarily that the marginal outcome probabilities are known and set $\alpha = \log[\Pr(Y = 1)/\Pr(Y = 0)]$. If the constant term $\log(N_1/N_0) - \alpha$ is added to β_0, then the Horvitz–Thompson estimating equations (7) are unbiased for all parameters, including the intercept, and the corresponding estimator is consistent [15]. The equations are easily solved using standard programs for logistic regression by treating the inverse sampling fractions f_{ij}^{-1} as prior weights (*see* **Software, Biostatistical**). The asymptotic variance is again given by (6), except that now the middle term is estimated by

$$\sum_{i=0}^{1} \sum_{j=1}^{J} f_{ij}^{-2} \left\{ \sum_{k=1}^{n_{ij}} \mathbf{U}_{ijk}^{\otimes 2} - \frac{1 - f_{ij}}{n_{ij}} \left(\sum_{k=1}^{n_{ij}} \mathbf{U}_{ijk} \right)^{\otimes 2} \right\} - \sum_{i=0}^{1} \frac{1}{N_i} \left(\sum_{j=1}^{J} f_{ij}^{-1} \sum_{k=1}^{n_{ij}} \mathbf{U}_{ijk} \right)^{\otimes 2},$$

where $\mathbf{u}^{\otimes 2}$ for any vector $\mathbf{u}$ denotes the matrix $\mathbf{u}\mathbf{u}'$. The last expression involving the N_i, which only affects the variance of $\hat{\beta}_0$, reflects the extra information obtained by assuming α to be known and has no counterpart in (2). In practice, α is not known. Since its value only affects the free parameter β_0, however, this does not matter. One simply ignores $\hat{\beta}_0$ and its **standard error**.

Maximum likelihood estimation

Scott and Wild [31, 32] derived the nonparametric maximum likelihood estimator of regression coefficients for the population based study. Breslow and Holubkov [8, 9] derived the nonparametric maximum likelihood estimator for logistic regression coefficients for the study with separate samples of cases and controls at phase one. The two estimators are identical, and semiparametric efficient, when the binary response model (3) is specified as logistic [28]. Although they can be substantially more efficient than Horvitz–Thompson under certain conditions, in many practical settings, the loss of information is slight [7, 28, 29]. Many researchers prefer the Horvitz–Thompson approach because, at least for the **population-based study**, it provides a consistent estimator of a meaningful population parameter even if the model (3) does not hold.

Pseudo-likelihood and pseudo-score estimators

Other methods of estimation of logistic regression coefficients have been developed that are typically of intermediate efficiency. Breslow & Cain [6] and Schill and colleagues [29, 30] developed closely related pseudo-likelihood estimators, deriving unbiased estimating equations by maximization of a product of conditional probabilities. For the Breslow–Cain version one fits the logistic regression model (3) to the phase two data, using as offsets in the

linear predictor the terms $\log(n_{1j}N_{0j}) - \log(n_{0j}N_{1j})$ for observations in stratum j to adjust for the biased sampling. A matrix formula is available that corrects the usual asymptotic variance matrix to account for the additional information coming from the phase one data. The Schill version involves fitting a logistic regression model jointly to the phase one and phase two data and also requires offsets. Whenever the linear predictor contains a separate term for the main effect of each stratum, the two pseudo-likelihood estimators and the maximum likelihood estimator are identical.

Chatterjee and colleagues [13] developed an estimator they termed pseudo-score. This uses the regression model to *estimate* the scores for subjects *not* sampled at phase two more efficiently than they are estimated by the mean score method [25], which as already noted is equivalent to Horvitz–Thompson estimation in the present setting. The pseudo-score estimator has the important practical feature that it may be implemented in situations where the phase two sample consists entirely of cases or entirely of controls. This may be particularly valuable when the covariates ascertained at phase two involve an invasive medial procedure. However, because of the near impossibility of checking model assumptions from the collected data, this feature should be exploited only if absolutely necessary and then with considerable caution.

Fixed vs. Random Phase Two Sample Sizes

So far we have assumed that the sample sizes, n_{ij}, at the second phase of sampling are fixed by the investigator after considering results of the phase one data collection. An alternative sampling strategy uses a random device to decide independently for each phase one subject, using selection probabilities p_{ij} that depend on (Y, S), whether to include the subject at phase two. This is known by econometricians as Manski–Lerman [22] sampling. In the context of case–control studies, Weinberg & Sandler [35] called it the randomized recruitment method and suggested it as an alternative to **frequency matching**. Others have referred to it as Bernoulli sampling. The associated sampling theory is simplified by the fact that the phase two observations are rendered statistically independent.

Both Horvitz–Thompson and pseudo-likelihood estimation procedures may be applied to data collected in randomized recruitment designs. Instead of dividing the log likelihood contributions by the observed sampling ratios, f_{ij}, one divides by the expected ones, p_{ij}. Use of the observed ratios is also justified; just condition on the n_{ij} and use the previous theory. Both empirical [36] and theoretical [28] studies show that use of the known selection probabilities results in less efficient estimates, however, so it is best to treat the phase two sample sizes as fixed even when they are not.

An Example

Table 1 shows sample sizes at phases one and two for a study of the association of operative mortality and gender in patients undergoing coronary bypass surgery. Two different designs were used at phase two: a "case–control" design in which the subsamples of 100 cases (deceased) and 100 controls (alive) were drawn without consideration of the stratum variable gender; and a "balanced" design in which equal numbers ($n_{ij} = 50$) were drawn from within each of the four cells defined by both outcome and gender. More generally a balanced design involving n phase two subjects would set $n_{ij} = \min(n/2J, N_{ij})$. Since the "case–control" design involves sampling only with regard to outcome, fitting ordinary logistic regression models yields valid estimates of relative risk parameters [24]. Fitting of ordinary logistic regression models to data from the "balanced" design, however, results in biased estimates of the gender effect since the association between outcome and gender at phase two is completely distorted by the nonproportional sampling ratios.

Table 1 Sample sizes for a study relating operative mortality to gender for patients undergoing coronary artery bypass surgery

	Male ($S = 1$)	Female ($S = 2$)
First phase sample (N_{ij})		
Alive ($Y = 0$)	6,666	1,228
Deceased ($Y = 1$)	144	58
Second phase sample: case–control (n_{ij})		
Alive	81	19
Deceased	67	33
Second phase sample: balanced (n_{ij})		
Alive	50	50
Deceased	50	50

Source: Cain & Breslow [10], reproduced by permission of the publisher.

Table 2 reports results of estimating logistic regression coefficients for gender and other covariates by two methods: ordinary logistic regression and "adjusted" logistic regression. Since gender is included in the model, simple adjustments may be applied directly to the results from ordinary logistic regression programs to calculate the maximum likelihood estimates and their standard errors [10]. Furthermore, adjusted and unadjusted results for the covariates other than gender are identical [6]. Note the severe distortion of the gender coefficient for the unadjusted analysis of the balanced data, and the substantial reduction in its standard error, when adjustment is made for the information available at phase one. Results of fitting the same model to the entire phase one data set, for which all the covariate values were in fact known, are shown for comparison.

Efficiency Gains from Stratification

The standard errors shown in Table 2 suggest there was little benefit with these data from use of the balanced design at phase two. In other situations one should expect the balanced design to yield more **efficient** estimates of coefficients that model stratum effects and their **interactions**, at the possible cost of some mild loss of efficiency for estimates of the other covariate effects. This conclusion is based on the asymptotic efficiencies shown in Table 3 for models with a binary stratum coefficient, β_1, and a binary covariate, β_2, with and without an interaction term, β_3. The table shows results for a study where all cases identified at phase one are also used at phase two. Here the principal efficiency gain is for the interaction effect. When subsamples of both cases and controls are taken at phase two, stratum and interaction effects are both estimated more efficiently using the balanced design.

While the balanced design seems reasonable on general grounds and has good efficiency properties over a large region of the parameter space, other designs offer greater efficiency in particular circumstances [10, 16]. For the measurement error problem with Horvitz–Thompson estimation, Reilly & Pepe [25] derived expressions for the optimal sampling fractions, f_{ij}, needed for estimation of a particular regression coefficient. They presented numerical results for a logistic regression model with a single explanatory variable, X, assumed to have a **standard normal distribution**.

Table 3 Large-sample efficiencies of the balanced design relative to the case–control design when the second phase sample contains all the cases but only a fraction of the controls from the first phase sample

		Relative efficiency		
e^{β_2}	θ[a]	β_1	β_2	β_3
0.2	0.2	1.02	0.83	1.43
0.2	1.0	1.18	0.88	1.45
0.2	5.0	1.34	0.93	1.65
1.0	0.2	0.99	0.74	2.30
1.0	1.0	1.00	0.81	2.09
1.0	5.0	0.98	0.82	2.06
5.0	0.2	1.14	0.76	2.94
5.0	1.0	1.22	0.81	2.12
5.0	5.0	1.00	0.76	1.74

Source: Cain & Breslow [10], reproduced by permission of the publisher.
[a]θ = odds ratio measure of association between S and X among controls.

Table 2 Logistic regression coefficients (and standard errors) for the data in Table 1

		Second phase sample: case–control		Second phase sample: balanced	
Model term	First phase sample	Unadjusted	Adjusted	Unadjusted	Adjusted
Constant	−3.271 (0.285)	−0.167 (0.615)	−3.812 (0.594)	0.990 (0.637)	−2.845 (0.606)
Female sex	0.634 (0.171)	0.650 (0.348)	0.690 (0.190)	−0.061 (0.301)	0.722 (0.189)
Diameter of arteries	−0.065 (0.016)	−0.030 (0.034)		−0.080 (0.033)	
CHF score	0.445 (0.072)	0.395 (0.160)		0.348 (0.165)	
Priority of surgery[a]					
Urgent	0.706 (0.181)	0.631 (0.365)		0.412 (0.350)	
Emergency	2.004 (0.232)	2.605 (1.072)		1.853 (0.800)	

Source: Cain & Breslow [8], reproduced by permission of the publisher.
[a]Relative to elective surgery.

Stratification was based on a binary, S, indicating whether or not X, when measured with normally distributed error, is positive. Although intended for cohort sampling at phase one, their results apply also to the two-phase case–control study provided that the intercept parameter, β_0, is interpreted as applying to the case–control sample rather than the source population.

Table 4 shows the optimal sampling fractions for a study involving $N = 500$ phase one subjects, 25% of whom are to be selected for the validation sample at phase two. Note that the balanced design here is optimal only when $\beta_0 = \beta_1 = 0$. As with all design problems, the optimal sampling fractions depend on the values of unknown parameters, so that prior information or a good guess is needed to achieve near optimality. Holcroft and Spiegelman [16] used numerical methods to determine optimal sampling fractions when regression parameters are estimated by maximum likelihood.

In addition to varying the sampling fractions used for the selection of the phase two sample, efficiency may be further increased by posthoc stratification in the analysis of the data [7]. The stratum factor S is constructed to represent combinations of levels of any covariate factors available at phase one, not just those used to stratify the sampling. The main limitation is the necessity for each stratum to contain both cases and controls at phase two, which means that continuous valued covariates available for all subjects cannot be utilized fully. Since the phase one data enter the Horvitz–Thompson estimation procedure only through the sampling ratios f_{ij}, post hoc stratification increases the amount of phase one information that is actually used in the analysis. This information is of particular value when stratification is based on surrogate explanatory variables that are highly correlated with those used in the regression model.

Table 4 Optimal sampling fractions f_{ij} when $\sigma = 0.25$[a]

(β_0, β_1):	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(2,2)
f_{01}	0.25	0.46	0.12	0.28	0.11	0.21
f_{02}	0.25	0.47	0.52	0.61	0.75	0.80
f_{11}	0.25	0.17	0.54	0.43	0.72	0.56
f_{12}	0.25	0.17	0.12	0.04	0.11	0.06

Source: Reilly & Pepe [21], reproduced by permission of the publisher.

[a]σ = standard deviation of measurement error distribution.

Other Complex Sampling Designs

Whittemore [38] studied Horvitz–Thompson estimation for designs with three or more sampling phases based on nested partitions of the sample space into increasingly fine strata. Noting that the selection probabilities for observations with complete covariate data were given by products of the inverse sampling fractions at each phase, she derived an explicit expression for the variance of the scores used in the information sandwich formula (6). Results were given for both fixed sample size and random recruitment at the second and each succeeding phase.

Whittemore & Halpern [39] gave an example of a three-phase study of the association between prostatic cancer and lifestyle factors. The first phase involved a case–control sample of men with and without a history of prostate cancer. Each was asked if he had a brother with the disease, and subsamples were drawn at phase two according to whether or not there was a positive reply. A complete family history was taken for those so selected. In a third phase of sampling all the subjects whose families had three or more cases of prostate cancer were asked to provide blood or tissue specimens for DNA analysis. Data of interest for statistical modeling included the complete family histories and the DNA results. Variance formulas were used to determine optimal sampling fractions at phases two and three for estimation of parameters in genetic models. Stratification of the phase two sample, using the reply to the simple question posed at phase one, remarkably improved efficiency.

Benichou et al. [2] provided another example of three-phase sampling involving women participating in a large-scale breast cancer demonstration project. Data on age, treatment center, and date of entry into the study were available initially for 280 000 women. The second phase involved selection of approximately 3000 breast cancer cases and 3000 controls with data on family history and clinical and reproductive risk factors. At the third phase, subsamples were selected on the basis of the availability of mammographic information. This example is best characterized as a problem with a complex pattern of **missing data**, because at phases two and three only those cases and controls with complete data on the relevant

covariates were retained. The authors developed their own pseudo-likelihood analysis, using the parametric **bootstrap** for calculation of standard errors.

The essential feature of the multiphase sampling design [38] that makes it amenable to simple Horvitz–Thompson estimation is the fact that the resulting data are subject to a *monotone* pattern of *missingness* [21]. The covariates may be ordered so that the groups of subjects missing each one of them are nested within each other. For other complex sampling designs the data analysis is inherently more difficult and may require additional parametric assumptions. Wacholder et al. [34], for example, proposed the *partial questionnaire* design for case–control studies as a means of reducing the length of the questionnaire administered to most subjects. Different subgroups of subjects are given distinct questionnaires that are missing different blocks of questions, so that the missingness pattern is *nonmonotone*. They developed a pseudo-likelihood estimation procedure that requires explicit estimation of the joint covariate distribution in the sample. Consequently, the technique is currently restricted to studies with a small number of discrete covariates.

Conclusions

Epidemiologists, especially those working in genetic epidemiology, have begun to recognize the value of two- and three-phase stratified sampling designs for case–control studies [1]. By incorporating available data into the design and analysis to the fullest possible extent, great savings in cost can be achieved with little or no sacrifice in statistical precision. Computational tools are now available that facilitate such designs and analyses. Horvitz–Thompson estimation of logistic regression coefficients has been implemented for data collected in complex sample surveys [33]. Macros for Horvitz–Thompson estimation in two-phase studies, and for the determination of optimal sampling fractions from pilot data, are now available for the popular package STATA [26, 27]. S-Plus functions are available for implementation of the more efficient pseudo-likelihood and maximum likelihood estimators [7].

This article has described two-phase stratified sampling designs and analyses for estimation of logistic regression parameters from binary outcome data. Similar methods have been developed for estimation of relative risk parameters in the Cox proportional hazards model for failure time data. Horvitz–Thompson estimation procedures are available, for example, for covariate stratified versions of the nested case–control study and of the case–cohort study. Further work is needed to implement more efficient estimation methods for these and other complex sampling designs.

Acknowledgment

This work was supported in part by a US Public Health Services Grant CA 40644.

References

[1] Andrieu, N. & Goldstein, A.M. (1998). Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods, *Epidemiologic Reviews* **20**, 137–147.

[2] Benichou, J., Byrne, C. & Gail, M. (1997). An approach to estimating exposure-specific rates of breast cancer from a two-stage case–control study within a cohort, *Statistics in Medicine* **16**, 133–151.

[3] Benichou, J. & Wacholder, S. (1994). Epidemiologic methods: a comparison of three approaches to estimate exposure-specific incidence rates from population-based case–control data, *Statistics in Medicine* **13**, 651–661.

[4] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review* **51**, 279–292.

[5] Breslow, N.E. (1996). Statistics in epidemiology: the case–control study, *Journal of the American Statistical Association* **91**, 14–28.

[6] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case–control data, *Biometrika* **75**, 11–20.

[7] Breslow, N.E. & Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms' tumor prognosis, *Applied Statistics*, **48**, 457–468.

[8] Breslow, N.E. & Holubkov, R. (1997a). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data, *Statistics in Medicine* **16**, 103–116.

[9] Breslow, N.E. & Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling, *Journal of the Royal Statistical Society, Series B* **59**, 447–461.

[10] Cain, K.C. & Breslow, N.E. (1988). Logistic regression analysis and efficient design for two-stage studies, *American Journal of Epidemiology* **128**, 1198–1206.

[11] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, New York.

[12] Chambless, L.E. & Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models, *Communications in Statistics – Theory and Methods* **14**, 1377–1392.

[13] Chatterjee, N., Chen, Y.-H. & Breslow, N.E. (2003). A pseudo-score estimator for regression problems with two-phase sampling, *Journal of the American Statistical Association 98*, in press.

[14] Cochran, W.G. (1963). *Sampling Techniques*, 2nd Ed. Wiley, New York.

[15] Flanders, W.D. & Greenland, S. (1991). Analytic methods for two-stage case–control studies and other stratified designs, *Statistics in Medicine* **10**, 739–747.

[16] Holcroft, C.A. & Spiegelman, D. (1999). Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified, *Biometrics* **55**, 1193–1201.

[17] Horvitz, D.G. & Thompson, D.J. (1951). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**, 663–685.

[18] Hsieh, D.A., Manski, C.F. & McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations, *Journal of the American Statistical Association* **80**, 651–662.

[19] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 221–233.

[20] Kalbfleisch, J.D. & Lawless, J.F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality, *Statistics in Medicine* **7**, 149–160.

[21] Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd Ed. Wiley, New York.

[22] Manski, C.F. & Lerman, S.R. (1977). The estimation of choice probabilities from choice based samples, *Econometrica* **45**, 1977–1988.

[23] Neyman, J. (1938). Contribution to the theory of sampling human populations, *Journal of the American Statistical Association* **33**, 101–116.

[24] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case–control studies, *Biometrika* **66**, 403–411.

[25] Reilly, M. & Pepe, M.S. (1995). A mean score method for missing and auxiliary covariate data in regression models, *Biometrika* **82**, 299–314.

[26] Reilly, M. & Salim, A. (2000). Mean score method for missing covariate data in logistic regression models, Stata Technical Bulletin 58, 25–27.

[27] Reilly, M. & Salim, A. (2000). Computing optimal sampling designs for two-stage studies, Stata Technical Bulletin 58, 37–41.

[28] Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* **89**, 846–866.

[29] Schill, W. & Drescher K. (1997). Logistic analysis of studies with two-stage sampling: a comparison of four approaches, *Statistics in Medicine* **16**, 117–132.

[30] Schill, W., Jöckel K.-H., Drescher, K. & Timm, J. (1993). Logistic analysis in case–control studies under validation sampling, *Biometrika* **80**, 339–352.

[31] Scott, A.J. & Wild, C.J. (1991). Fitting logistic regression models in stratified case–control studies, *Biometrics* **47**, 497–510.

[32] Scott, A.J. & Wild, C.J. (1997). Fitting regression models to case–control data by maximum likelihood, *Biometrika* **84**, 57–71.

[33] Shah, B.V., Folsom, R.E., LaVange, L.M., Wheeless, S.C., Boyle, K.E. & Williams, R.L. (1993). *Statistical Methods and Mathematical Algorithms Used in SUDAAN*. Research Triangle Institute, Research Triangle Park, North Carolina.

[34] Wacholder, S., Carroll, R.J., Pee, D. & Gail, M. (1994). The partial questionnaire design for case–control studies, *Statistics in Medicine* **13**, 623–634.

[35] Weinberg, C.R. & Sandler, D.P. (1991). Randomized recruitment in case–control studies, *American Journal of Epidemiology* **134**, 421–432.

[36] Weinberg, C.R. & Wacholder, S. (1990). The design and analysis of case–control studies with biased sampling, *Biometrics* **46**, 963–975.

[37] White, J.E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

[38] Whittemore, A.S. (1997). Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society, Series B* **59**, 589–602.

[39] Whittemore, A. & Halpern, J. (1997). Multistage sampling in genetic epidemiology, *Statistics in Medicine* **16**, 153–167.

NORMAN E. BRESLOW

Case–Control Study

The case–control design provides a framework for studying the relationship between possible risk factors and a disease by collecting information about exposure from those with disease but only from a fraction of the individuals under study who do not develop disease. When the disease is rare, this approach offers a major gain in efficiency relative to the full **cohort study**, in which an investigator seeks information on exposure for everyone. The savings compensate handsomely for the loss in the precision of estimates of parameters describing the relationship between exposure and disease that could have been obtained from studying everyone. In fact, the reduction in precision often is marginal. By collecting data on exposure about *cases*, the subjects who have developed disease, and **controls**, specially selected subjects without disease, the case–control design also compresses the time needed to complete the study. In a classic case–control study, Doll & Hill [10] recruited 649 male lung cancer cases and 649 male controls during an 18-month period in London. They were able to show a clear increase in **risk** with increasing daily cigarette consumption in this case–control study (*see* **Smoking and Health**). By contrast, in a cohort study of an equal number of men at the very highest risk – that is, very heavy smokers above age 70 – one would expect to find only a handful of lung cancer cases within 1.5 years, not nearly enough to draw convincing conclusions about the relationship between smoking and lung cancer.

A hypothetic example illustrates the extent of the savings. In Table 1 are displayed the results of a cohort study of 1 000 000 individuals who are followed for disease for one year; 10% of them are exposed.

The expected results from a case–control study in which all 56 of the cases from this cohort are studied are displayed in Table 2. Expected cell counts are also shown in Table 2. For example, the expected number exposed among the 56 studied controls is calculated as $56 \times (99\,984/999\,944) = 5.6$.

The estimate of the **odds ratio** for disease, (16/40)/(99 984/899 960) in Table 1, equals the estimate of the exposure odds ratio in Table 2, (16/40)/(5.6/50.4). Both odds ratios equal 3.6004, and approximate the risk ratio (*see* **Relative Risk**) $(16/100\,000)/(40/900\,000) = 3.60$ to four significant digits. Thus, the study of 112 individuals would give the same estimate as the study of 1 000 000, apart from random variation. While the 95% **confidence interval** (CI) for the odds ratio from the case–control study, (1.3–10.3), is substantially wider than the CI (2.0–6.4) from the full cohort study, using $5 \times 56 = 280$ controls instead of only 56 would narrow the CI for the case–control study to (1.8–7.2), which is notably closer to that of the full cohort study. This minor loss of precision is a small price to pay for the savings in exposure assessment costs and in time that may make feasible a study that would otherwise be too expensive.

In principle, although not always in practice, all case–control studies yield an **unbiased** estimate of the odds ratio and other functions of the odds. Most are designed so that the odds ratio directly estimates the relative risk or the incidence-rate ratio. However, only **population-based case–control studies** that yield estimates of overall disease risk or

Table 2 Expected values from case–control study in same setting as Table 1

	Diseased (cases)	Nondiseased (controls)	Relative odds
Exposed	16	5.6	3.60
Unexposed	40	50.4	1.00
Total	56	56	

Table 1 Hypothetical full cohort study

	Diseased	Nondiseased	Total	Relative risk[a]	Relative odds[b]
Exposed	16	99 984	100 000	3.60	3.60
Unexposed	40	899 960	900 000	1.00	1.00

[a]Relative risk = (16/100 000)/(40/900 000).
[b]Relative odds = (16/99 984)/(40/899 960).

rate in the population permit estimation of exposure-specific **incidence rates** and thus of all parameters that could be estimated from studying the entire cohort.

Along with these considerable design strengths, the case–control study has several weaknesses. Incomplete or inaccurate ascertainment of outcome and improper selection of controls can cause **selection bias**. Retrospective assessment of exposure history can lead to **nondifferential** and **differential** measurement error and **biased** estimates of exposure effects. As in any nonexperimental or **observational study**, **confounding** can distort the estimates of effect from a case–control study (*see* **Bias in Case–Control Studies**; **Bias in Observational Studies**; **Bias, Overview**; **Measurement Error in Epidemiologic Studies**; **Misclassification Error**).

The Range of Case–Control Studies

A MEDLINE search for papers published since 1992 found over 1500 entries per year mentioning case–control or one of its cognates, usually case–referent. The case–control study is a fundamental tool of epidemiology with broad application in areas as diverse as the etiology of cancer and birth defects, the effectiveness of vaccination and **screening** for disease, and the causes of automobile accidents.

Case–control studies vary greatly in scope, sources of data, and complexity. At one extreme are investigations of an outbreak, which may include fewer than ten cases (*see* **Communicable Diseases**). These studies often encompass a wide-ranging, open-ended examination of many exposures and host characteristics of the cases. Often, the selection of controls can precisely correspond to the source of cases because there is a roster of the source population (for example, in a hospital outbreak) or a convenient collection of willing participants. At the other extreme are multicenter, multiyear, highly focused studies of tens of thousands of cases and controls. These are not common, because of their high cost. More typical are studies of a few hundred cases and an equal number of controls selected without a roster, but with an algorithm intended to represent the population from which the cases arose. These intermediate-sized studies provide a practical approach when the relative risk is expected to be around 2 or greater and the exposure is reasonably common (10% or more).

Weaknesses of the Case–Control Approach

Case–control studies, like **cross-sectional** and observational cohort studies, suffer from the common drawbacks of all nonexperimental, or observational, research, stemming from the investigator's lack of control in assigning exposure. Foremost is the absence of **randomization** as a tool for reducing confounding. An observational study will not be as reliable as a **clinical trial** for investigating questions such as the effectiveness of a new treatment or screening program.

Even though a case–control study has no intrinsic shortcomings compared exposuto a nonexperimental full cohort study that collects information on everyone in the same setting, the case–control design has often been disparaged as fundamentally weaker than the full cohort study. Several conceptual, statistical, and practical reasons explain this negative attitude. Many early observers saw the case–control study as a "backward cohort study", with inference made from effect to cause. It was not obvious how to translate a difference in exposure between cases and controls into a parameter describing prospective risk until **Cornfield** in 1951 [7] showed theoretically that the exposure odds ratio from a case–control study approximates the disease risk ratio from a case–control study when the outcome is rare. Selection bias can arise from poor study design or poor implementation in choosing cases and controls. Retrospective ascertainment of information about exposure and confounders may yield inaccurate data leading to bias. These issues are discussed in detail later in this article.

Another apparent weakness of the case–control approach is that ordinarily it yields relative but not absolute measures of the effect of exposure on disease. It is possible, however, to estimate exposure-specific **absolute risk** and risk differences when the crude risk of disease is known in the study population [1, 7, 10, 30].

Case–Control Study as a Missing-Data Problem

A population-based case–control study can be regarded as a cohort study with many nondiseased subjects missing at random [30]. This view of the case–control study helps to resolve many conceptual issues.

It reveals when and how a broader class of parameters, including absolute risk and risk difference, can be estimated. It clarifies the requirements for proper control selection (*see* **Missing Data in Epidemiologic Studies**; **Missing Data**).

Consider a population-based case–control study to examine the effect of an exposure on the risk of developing disease. In the ideal study, the investigator is able to identify all cohort members newly diagnosed with disease during a specified follow-up period. These people with disease, or a random subset, become the cases in the study. Controls are a **random sample** of the noncases. The investigators obtain information on exposure that preceded the time of onset of disease from these cases and controls. Exposure information for those noncases who never develop disease during the study period will be *missing at random* if the investigator determines whose exposure will be collected, based only on disease status, which is known for individuals during the specified time. Thus, the case–control study is a missing-data problem, albeit with two unusual features: the "missingness" is a planned maneuver rather than an uncontrollable accident, and the ratio of missing to observed data can be extraordinarily high.

Under these assumptions, the cases and controls will have the same exposure distribution as the diseased and nondiseased, respectively, in the cohort, and the investigator can estimate from the case–control data all of the parameters estimable from the full cohort study. Indeed, under these assumptions, there are no intrinsic weaknesses to the case–control design. This outlook recognizes the prospective nature of the study, allows estimation of all parameters available from the full cohort, including absolute risk and risk difference, and demonstrates why the controls selected should have the same exposure distribution as other nondiseased individuals in the study population [30]. The inference from the missing data approach is identical to standard case–control **inference** in this setting [30].

Case–Control Studies to Estimate a Hazard Ratio

In the idealization described above, risk is described as the probability of developing disease during a fixed interval. If the study aims to estimate functions of **hazard rates** of disease, or numbers of new events per unit of person–time (*see* **Person-years at Risk**), the time element must be incorporated more precisely. For instance, in the standard **proportional hazards** analysis of the full cohort study designed to estimate the hazard ratio, the **partial likelihood** compares the exposure of a case to that of the members of the *risk set*; namely, all other members of the cohort who are at risk at the time of the event that defines when the cohort member became a case.

In the nested case–control study that would be undertaken in the same cohort, as first described by Thomas [27], exposure from only a few randomly selected members of each risk set is collected and used in a time-matched case–control analysis, an analog to partial likelihood. Again, except for the use of fewer individuals, there is no intrinsic difference between the full cohort and nested case–control analyses. All noncases in the risk set should be eligible and equally likely to be sampled as controls, even those who were previously selected as controls or who later develop disease [15]. Sampling at event times should be mutually independent in the nested study.

The **case–cohort design**, first described by Prentice [21], is a useful alternative with several practical advantages. The controls are selected as a single sample or *subcohort* from the entire cohort, including cases. While the sampling is not time-matched, in the analysis the **likelihood** at each event time uses the exposures of the case and of the subcohort members who are in the risk set at the event time. The fact that the subcohort is a random sample of the cohort leads to more flexibility in the analysis and allows the same controls to be used for analyses of several endpoints.

Design

There are three interlocking steps in planning the design of a case–control study:

1. Investigators must decide whether a cohort or case–control study is appropriate.
2. Investigators must determine who will be cases and controls in the study and how to assess exposure.
3. Investigators must decide on all the specific details to be included in the study protocol.

Full Cohort vs. Case–Control?

The first decision required in planning a case–control study is to determine whether the case–control design is more appropriate than a full cohort design [29]. The reasons for preferring a case–control study to the full cohort study are almost always practical, revolving around feasibility, economy, speed, and the need to study multiple exposures or their joint effects. On the other hand, a prospective cohort study sometimes affords an opportunity to collect more reliable exposure information, and can be used to study **multiple health outcomes** simultaneously. It can offer slightly more statistical precision. Finally, justifiably or not, the cohort study has more credibility.

Lower Cost vs. Higher Statistical Efficiency. Studying fewer subjects reduces the cost but also lowers the precision of the estimate of effect. When the disease is rare, the impact will be very modest, as the above example demonstrates. The **variance** estimate of the log-odds ratio estimate from **two-by-two tables** of the form in Table 1 or Table 2 is the sum of the reciprocals of the cell entries, so the size of the smallest cell in the two-by-two table is the factor limiting precision. When exposure is rare, this smallest cell almost always will be the number of exposed cases. This quantity is the same in the full cohort or in the case–control study performed in the same setting. Thus, the relative efficiency of the case-control study with k controls per case is $k/(k+1)$ compared to the full cohort [28]. The choice of design often boils down to whether to look for cases among the exposed (as in a cohort study) or exposed among cases (as in the case–control approach).

The clearest advantage for the case–control study occurs when the outcome of interest is rare and the exposure of interest is common. As the percentage of individuals experiencing the outcome during the follow-up period increases, the efficiency advantage of the case–control design diminishes. As the exposure of interest becomes rare, the ability of the case–control study to estimate an effect diminishes and a cohort design that ensures that individuals with the rare exposure will be followed for disease may become more advantageous.

Data Quality. Exposure assessment is the Achilles heel of the case–control study. If information collected retrospectively about exposure is of lower quality than concurrent data, more *nondifferential* **misclassification** or error, and consequently, attenuation of estimates of effect, almost inevitably ensue. Worse still, exposure information that is self-reported is susceptible to *differential* error or misclassification, namely different error patterns in cases and controls.

The resulting bias can work to exaggerate, attenuate, or reverse the direction of an effect. While differential error from interviews has been difficult to establish conclusively in particular situations, it seems realistic to assume that the accuracy and thoroughness of reports from cases, who are touched by the research question and whose lifestyle may be affected by the disease, will be greater than for controls. The effect of differential error is often called report or **recall bias**. Some nutritional epidemiologists are extremely skeptical of dietary data collected from cases and controls retrospectively, for fear of differential misclassification (*see* **Nutritional Exposure Measures**). By contrast, when previously written records are the source of exposure information, the errors are no different from those in a full cohort study. So a retrospective or even a prospective full cohort study would not automatically have higher data quality. Correspondingly, collection of reliable information on outcomes in all members of a cohort or a case–control study is also a challenge, especially for softer endpoints, such as infertility.

Other Scientific Issues. Apart from considerations of efficiency, reflecting the rarity of disease and exposure, other considerations come into play. When confounding poses a major problem for a study, accurate **confounder** assessment may dictate one design or the other. The need to study multiple exposures magnifies the advantage of the case–control design, while a cohort study allows additional outcomes to be included in the study with little increase in cost. Some well-established cohorts [37] have demonstrated that results on the relationships between multiple exposures and multiple exposures from a full cohort study can justify its substantially greater cost relative to a single case–control study.

Credibility. While most researchers and journals now appreciate the case–control design, some still consider case–control studies automatically suspect [11]. While this attitude is becoming less widespread, it may affect how one's work is accepted.

Choice of Setting

The specific setting for the study must be chosen within constraints imposed by logistics, convenience, and cost. Investigators must also consider the key factors that determine the quality of a case–control study in a particular location. How complete and accurate will the case ascertainment be? How rapidly will investigators receive reports of cases, thereby reducing the influence of the postdiagnosis period, such as effects of treatment, and the number of fatal or debilitated cases who might be excluded or whose exposure information may need to be collected from a proxy, such as a spouse or child? Is there a roster or **sampling frame**, possibly from electoral lists or a health insurance plan (*see* **Administrative Databases**), from which to select suitable controls? Are written records available to evaluate exposure, thereby reducing the possibility of differential misclassification? Are participants likely to give reliable information on exposure or confounders, including perhaps family medical history, prescription drug use, or highly personal questions about sexual history or a previous abortion? Are participation rates likely to be high? (*see* **Nonresponse**). Will participants be amenable to a procedure needed for the study, such as blood drawing for assessing a biomarker? What is the rate of occurrence of events and how will it affect the amount of time needed in the field? Is there enough heterogeneity of exposure to reduce the cost of a study and the number of subjects needed to achieve a specified precision?

Case–control studies can be oriented toward measuring the effect of exposure on disease **prevalence**, **cumulative hazard**, or **incidence rate**. Thus, the temporal perspective must be considered. Ought the study be limited to future cases, or can previously diagnosed individuals be used? Using only those cases that are newly diagnosed (**incident cases**) generally works to improve case ascertainment and participation, reduce reliance on proxy respondents for deceased or disabled cases and simplify control selection, but is slower and more costly. One subtly different definition of cases produces an estimate of cumulative risk rather than **incidence density ratio**; namely, when cases are all subjects who developed disease throughout the duration of follow-up of a population. Finally, diseases with poorly defined onset and long duration call for prevalence studies, with the definition of cases correspondingly changed to subjects who have the disease at the specified point in time, regardless of when they first developed it (*see* **Case–Control Study, Prevalent**).

Case and Control Selection

Case and control selection must be defined together because they are intrinsically linked. Miettinen's [20] concept of the *study base* helps to clarify this connection. The study base at a given time consists of those individuals who would become cases in the study if they developed disease at that time. When the study base is well-defined, the study is called a *primary-base* study or a *population-based case–control* study; cases are simply those members of the study base who experience the outcome and controls can be a random sample from the base. In this situation, it is possible to determine whether any individual is in or out of the study base at a given time and whether that individual is eligible to be a case or control in the study. The problem is making sure that all cases in the base come to the attention of the study investigators. The alternative starts with a set of cases, perhaps chosen for convenience, as in a *hospital-based case–control* study of lung cancer diagnosed at a single hospital during a single year. In these *secondary-base* studies, the study base is poorly defined because it is not always clear whether an individual who did not develop disease would have been a case in the hypothetic circumstance of development of disease. With no way to know whether a potential control would have come to the study hospital upon development of disease, random sampling for control selection is impossible. Thus, these secondary-base controls must be *assumed* to be an approximation to a hypothetic random sample that could characterize the study base. So in the primary base study, the difficulty is finding the cases, while, in the secondary base study, the difficulty is ensuring an appropriate set of controls.

Case Selection. In the idealized case–control study, all subjects with disease in the study base (or a random sample of them) become cases. In reality, some cases do not come to the attention of the investigators, some individuals are falsely called cases when in fact they do not meet the diagnostic criteria, and some eligible cases refuse to participate. In a study of male infertility, factors that lead to someone to regard lack of children as a problem might appear as

risk factors because of differential case ascertainment [31]. Inaccurate and incomplete case ascertainment can create selection bias as well as reduce precision. When there is ambiguity as to whether someone truly developed disease, as in the absence of a definitive pathology report, the standard practice of excluding the case is not harmless, if those lacking information have different exposure distributions than those with the information, perhaps because they are seen at a hospital in a poorer area [29].

Principles of Control Selection. There are three principles that underlie control selection: *study-base, comparable-accuracy*, and *deconfounding* [31]. The essence of the study-base principle is that controls can be used to characterize the distribution of exposure in the study base from which the cases arise. The comparable-accuracy principle calls for equal reliability in the information obtained from cases and controls so that there is no *differential* misclassification. Thus, a study of drug use during pregnancy as a risk factor for a specific type of birth defect might call for a control group of children who experienced a comparably serious outcome at birth so that the mothers of cases and controls would be equally likely to recall exposure during pregnancy accurately. The deconfounding principle allows elimination of confounding through control selection, such as through matching or **stratified sampling**, to be a consideration in control selection. These principles may conflict with one another and may have strong negative impacts on efficiency. They should not be regarded as absolute, but rather as points to consider in choosing a control group.

Controls for Studies with a Roster. In fortuitous situations, the investigator can use a roster listing all individuals and the period when they are in the study base. Investigators can then sample at random from the roster to satisfy the study-base criterion.

Controls for Primary-base Studies without a Roster. When a roster is not available and cannot be created from electoral or town residence lists, it is impossible to generate a random sample directly. A commonly used approach when there is no roster is **random digit dialing** (RDD) [34], an efficient way to generate a near-random sample often used in public opinion polling. RDD relies on dialing telephone numbers according to a strategy that yields representative samples. RDD suffers from several potential biases. RDD will not select individuals without phones, although it can compensate for households with multiple telephone lines. Furthermore, many people refuse to respond to telephone surveys, especially since the advent of answering machines (*see* **Telephone Sampling**). Empirically, controls chosen by RDD seem to be of higher socioeconomic class than a truly random sample would be. This violation of the study-base principle may be alleviated by adjustment for income or socioeconomic status.

Requirements for individual controls vary. In *incidence-density sampling* [14, 19, 22], used most commonly in primary-base studies, controls must be disease free at the time of diagnosis of the case to which they are matched. As in the nested case–control study, this design allows estimation of an incidence rate-ratio (and **relative hazard**) and eliminates the need for the rare-disease assumption [14, 19]. For *cumulative-incidence sampling*, controls are selected from among those who survive the study period without developing disease. Cumulative-incidence sampling of controls allows estimation of the risk ratio (**relative risk**), which approximates the relative hazard only when the rate of disease is low.

Secondary-Base Studies. Some diseases, including those not consistently detected in the general population, dictate an alternative to primary-base studies. For example, when case identification is incomplete, population controls may not be appropriate when completeness of case identification is differential by exposure and the selection bias cannot be corrected by adjustment for another variable. The most common secondary-base study is the *hospital-based case–control* study. Controls are patients seen at the same hospital as the cases, but for a different condition. This approach works well when two requirements are met. First, both cases and controls must be people who would have presented at the same hospital if they had *either* the case-defining illness or the control-defining condition. Secondly, the conditions used to select controls cannot be associated with the exposure. If these requirements are met, the distribution of the exposure in the controls reflects the distribution in the study base. The investigator seldom knows with certainty that both criteria are met, so compliance with the study-base criterion remains hard to verify convincingly.

A possible advantage of the hospital-based control group is more confidence that the equal accuracy criterion will be met. With equally serious illnesses, cases and controls ought to provide similarly complete and accurate reporting of past exposures. Thus, for the study of a specific birth defect, controls could be chosen from babies born with another birth defect of similar severity but known not to be related to the exposure of interest. Using controls with cancer at other sites for a study of a form of cancer may help with the equal-accuracy principle, but care must be taken so that cancer at the control site is not related to exposure.

Other Kinds of Control Groups. While population and hospital controls are the most commonly used kinds of control groups, investigators have used other options [32]. Use of patients from the *same primary care provider* as the case helps to insure that a control who developed the disease of interest would have become a case in the study. Use of *friends* of the cases can lead to bias in studies of factors related to sociability. Use of *relatives*, often siblings, as controls may reduce confounding by genetic factors. Each of these control groups requires a careful selection procedure to make sure that individuals are not being picked to be controls in a way that is related, directly or indirectly, to the factors under study.

Design Options. Matching on well-established confounders is a common practice in case–control studies. In case–control studies, matching serves to increase the precision of the estimated effect of exposure by making the distribution of the confounder identical in the cases and controls. Usually, the efficiency advantage from matching is small, and may not compensate for the extra cost and complexity, the exclusion of cases for whom no match is found, and the reduced flexibility of the analysis [33]. Other justifications of matching include control for nonquantitative variables such as neighborhood and the ability to control for confounding without making assumptions about the effect of the confounder in the risk model [33].

Only strong confounders should be considered as matching variables. Two-phase designs, discussed below, are more appropriate if one wants to estimate the effect of a variable considered for matching. Demographic variables such as race and sex and temporal variables such as age and calendar year (or decade) of first employment are the most suitable matching variables. Matching is always inappropriate on a factor that is a consequence of exposure.

Two-Phase Designs. These techniques [2, 36] (*see* **Case–Control Study, Two-phase**) are a more flexible generalization of matching, also used to increase efficiency or to reduce the cost of exposure assessment. In two-phase designs, detailed information on exposures and confounders is not ascertained for everyone, but only for subsets of cases and controls, with the selection probability depending on case status and on the value of another variable that is available for everyone. Instead of requiring, as in matching, that the distribution of the variable be the same in the control as in the cases, essentially arbitrary distributions in each group are specified. These two-stage designs allow the estimation of both main effects and **interactions**. For example, in a study designed to investigate the joint effects of domestic radon exposure, requiring expensive measurements, and smoking, which is easier to ascertain, on the risk of lung cancer, taking all cases and a random sample of controls would lead to a study with a preponderance of smoking cases and nonsmoking controls; matching controls to cases on smoking status would lead to small numbers of control nonsmokers as well. The assessment of interaction is much more efficient in the two-phase design where nonsmoking cases and smoking controls are oversampled [35].

Sample Size. There is an extensive literature on **sample size determination** for case–control studies [4, 24, 25]. As in the full cohort study, needed sample size is dependent on the variation in exposure in the study base. A key point is that increasing the ratio of controls to the harder-to-find cases increases the precision of the odds ratio estimate in an increasingly marginal way, especially for small effects. Ratios of controls to cases beyond four or five are usually not advisable because the successive gains in efficiency diminish. Indeed, the **asymptotic relative efficiency** for a study involving k controls per case is $k/(k+1)$, which takes on values of 0.5, 0.67, 0.75, 0.8, and 0.83 for k from 1 through 5 [28].

Fieldwork

The best-designed study will not be convincing unless the fieldwork is sound. In the field, case–control studies face the usual challenges of observational research: identifying all members of the study population, achieving an adequate response rate, collecting accurate data, and measuring potential confounders.

Most case–control studies include a questionnaire, because seldom have all of the exposure variables of interest been recorded in documents easily available to the investigator. Sometimes the study subject, or his surrogate, completes the questionnaire ("self-administered"); alternatively, an interviewer can pose the questions. A questionnaire can be computerized or on paper; an interview can be in person or by telephone (*see* **Interviewing Techniques**; **Questionnaire Design**). Depending on the hypotheses, investigators may also collect biologic specimens, samples of the study subject's present or past environment, and permission to contact agencies that have documented data about the exposures.

The case–control design poses some specific problems, as well. Since the cases have already developed the disease, it will not be possible to estimate the effects of exposure measures that are distorted by the disease, including weight and body biochemistry, unless the investigator has access to stored measures that were collected before disease onset. If it is not clear whether a measure is likely to be valid once the disease is clinically manifest, the investigator may conduct a specific methodologic pilot study. Sometimes, it is possible to examine the effects specific for stage of disease, in the expectation that post-onset distortions will be more pronounced with more advanced disease. In a similar fashion, the investigator will consider whether therapy influences the level of the exposure variable. If so, then cases need to be studied before therapy begins, or well after any of its influence has waned.

Just as diagnosis and treatment of a serious disease can cause biological changes in exposure variables, they also can cause changes in a patient's recollection or willingness to report various exposures. The resulting **recall bias** does not always go in a particular direction; the specific exposure needs to be considered, preferably with data on reporting bias from ancillary sources. Some exposures lend themselves to internal validation by studying a higher-quality exposure variable on a subset of subjects or by collection of validation data from other sources, such as medical records (*see* **Validation Study**). In that circumstance, some or all of the subjects reporting an illness or hospitalization will be asked to give permission for review of records; ideally, some of the reports of no hospitalization ought to be selected for review, too, although this is seldom practical. To minimize recall bias, the investigator also attends to the exact phrasing of questions, trying to leave very little room for interpretation or rumination. Sometimes investigators attempt to blind the interviewer to the case–control status of subject, but often the status of the subject becomes apparent anyway (*see* **Blinding or Masking**).

With access to prospectively collected data stored in records, the investigator can avoid the problem of differential misclassification stemming from the fact of diagnosis. Even with stored records, however, one source of differential misclassification could be present: minor abnormalities noted because of greater medical surveillance of the exposed may not have been detected in the unexposed (*see* **Bias From Diagnostic Suspicion in Case–Control Studies**; **Bias from Exposure Suspicion in Case–Control Studies**).

Analysis

The goal of the analysis of case–control studies is almost always to identify risk factors that are related to disease and to determine whether in fact the risk factors are causes of the disease (*see* **Causation**). As in other nonrandomized situations, the analysis must address the possibility of *confounding* and **effect-modification** by measured **covariates**.

The primary difficulty that is inherent to analysis of case–control studies is that the sampling is based on disease status while the parameters of interest relate to risk or rate of disease. Thus, it is not the difference in exposure frequency or **means** between cases and controls that is of direct interest, but estimates of the effect of determinants of disease on the rate of disease or on the probability of developing disease.

The analysis of case–control data can be exquisitely simple or tremendously complex. When the exposure and disease are each dichotomous (*see* **Binary Data**) and there are no other factors to

consider, the analysis reduces to a **two-by-two table** of exposure by disease status. Originally, Cornfield [7] proposed that the *odds ratio*, or cross product ratio in the two-by-two table could be used as an estimate of the risk ratio (or relative risk) when the disease was rare. **Mantel & Haenszel** [16] developed an estimator and a test statistic that could be used when combining tables over several strata, thereby controlling for confounding. Exact conditional approaches, not relying on asymptotic theory, are also available for obtaining inference on the common odds ratio, adjusted for confounders by **stratification** [3, 13, 18].

While **discriminant analysis** seems a natural tool to distinguish cases and controls, **logistic regression**, in which the dependent variable is the logarithm of the odds of disease, has two distinct advantages. It allows for exposures, confounders and effect-modifiers that are discrete or continuous, regardless of distribution [9] and yields valid estimates of relative-odds parameters from case–control data. Prentice & Pyke [23] proved that prospective logistic modeling – that is, of disease as a function of exposure – estimated relative risk parameters correctly and with full efficiency. If the sampling fractions of cases and controls are known, as in some population-based case–control studies, the intercept estimate from the case–control analysis can be combined with the ratio of the sampling fractions to yield a valid estimate of absolute risk. Logistic regression is now the most commonly used approach to analyze case–control data. Carroll et al. [6] extended the Prentice–Pyke result to show that many variations of case–control designs could be analyzed by logistic regression and given a prospective interpretation. Extensions to the logistic framework allow the handling of more complex sampling schemes, such as two-phase designs, of **nonlinear regression** effects of covariates, and of alternative models of joint effects of two risk factors, such as **additive** rather than **multiplicative** effects.

Control for a small number of categorical confounders can be achieved by the Mantel–Haenszel estimator of the odds ratio and corresponding **hypothesis test**. These simple procedures have excellent statistical properties. Nonetheless, logistic modeling is used routinely, because of its greater flexibility, for instance, in handling continuous variables [3]. In most modern studies, there will be more than two levels of exposure or one or more confounders and effect-modifiers to consider.

The analysis of matched pairs with a single dichotomous exposure variable uses only discordant pairs. It takes the ratio of pairs with the case exposed to those with the control exposed as the odds ratio estimate. The corresponding test of the **null hypothesis** that the odds ratio is one is equivalent to the hypothesis that the number of pairs with exposed cases among the discordant ones is **binomial** with probability 0.5 (*see* **Matched Analysis**; **McNemar Test**). While several extensions to more complex exposure variables and matching schemes were developed, the breakthrough in the analysis of matched data was the introduction by Breslow et al. [5] of **conditional logistic regression**, which allows general matching schemes, arbitrary exposures, continuous or discrete confounders (other than those used in the matching), and effect-modifiers.

An important variation of the analysis of case–control data allows for estimation of a hazard ratio rather than an odds ratio or risk ratio, as in nested case–control and case–cohort studies [21, 26, 27]. These designs are particularly useful when exposures vary with time, as does, for example, lifetime exposure to an environmental or occupational chemical. A conditional **likelihood** approach can be used here as well as where the matching is on time. In the contribution to the conditional likelihood at each event time, the exposure is accumulated only until the event time, exactly as if the analysis were prospective and no future data were available [21]. Furthermore, the same structure of the conditional likelihood as in the matched case–control study is used. As long as the controls are selected randomly from those at risk – that is, including future cases and independently of past use as a control or time of follow-up – the estimates of hazard ratio are valid, again reflecting the close relationship to the full cohort design. When the same controls are used for each case diagnosed during the control's follow-up, as in the case–cohort design, the estimates of the hazard ratio are also valid, but the variance estimate is more complex because the scores at each event time are not independent.

In most reports of case–control studies, no estimate of *absolute risk* or *absolute rate* is given. Methods are available for population-based case–control studies when the crude risk of disease is known [1, 7], and generally for nested case–control and case–

cohort studies [17, 37]. Furthermore, risk or rate difference and other nonlogistic models can be fit [30]. Methods for estimating the attributable risk and its variance in a general setting are also available [8].

Summary

The case–control study remains the most popular approach in **analytic epidemiology** because of its relatively low cost and high speed. Ascertainment of disease, selection of controls and measurement of exposure present substantial difficulties in almost every case–control study, but a large body of epidemiologic theory and experience provides guidance to meet these challenges.

References

[1] Benichou, J. & Wacholder, S. (1994). Epidemiologic methods: a comparison of the approaches to estimate exposure-specific incidence rates from population-based case–control data, *Statistics in Medicine* **13**, 651–661.

[2] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case–control data, *Biometrika* **75**, 11–20.

[3] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. I: it The Analysis of Case–Control Studies. International Agency for Research on Cancer, Lyon.

[4] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[5] Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.I. & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case–control studies, *American Journal of Epidemiology* **108**, 299–307.

[6] Carroll, R.J., Wang, S. & Wang, C.Y. (1995). Prospective analysis of logistic case–control studies, *Journal of the American Statistical Association* **90**, 157–169.

[7] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix, *Journal of the National Cancer Institute* **11**, 1269–1275.

[8] Coughlin, S.S., Benichou, J. & Weed, D.L. (1994). Attributable risk estimation in case–control studies, *Epidemiological Reviews* **16**, 51–64.

[9] Cox, D.R. (1966). Some procedures connected with the logistic qualitative response curve, in *Research Papers in Statistics: Festschrift for Neyman J.*, F.N. David, ed. Wiley, New York, pp. 55–71.

[10] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung: preliminary report, *British Medical Journal* **2**, 739–748.

[11] Feinstein, A.R. (1988). Scientific standards in epidemiologic studies of the menace of daily life, *Science* **242**, 1257–1263.

[12] Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C. & Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, *Journal of the National Cancer Institute* **81**, 1879–1886.

[13] Gart, J.M. (1970). Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals, *Biometrika* **57**, 471–475.

[14] Greenland, S. & Thomas, D.C. (1982). On the need for the rare disease assumption in case–control studies, *American Journal of Epidemiology* **116**, 547–553.

[15] Lubin, J. & Gail, M.H. (1984). Biased selection of controls for case–control analyses of cohort studies, *Biometrics* **40**, 63–75.

[16] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[17] McMahon, B. (1962). Prenatal X-ray exposure and childhood cancer, *Journal of the National Cancer Institute* **28**, 1173–1191.

[18] Mehta, C.R., Patel, N.R. & Gray, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables, *Journal of the American Statistical Association* **80**, 969–973.

[19] Miettinen, O.S. (1976). Estimability and estimation in case–referent studies, *American Journal of Epidemiology* **103**, 226–235.

[20] Miettinen, O.S. (1985). The "case–control" study: valid selection of subjects, *Journal of Chronic Diseases* **38**, 543–548.

[21] Prentice, R.L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[22] Prentice, R.L. & Breslow, N.E. (1978). Retrospective studies and failure time models, *Biometrika* **65**, 153–158.

[23] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case–control studies, *Biometrika* **66**, 403–411.

[24] Schlesselman, J.J. (1982). *Case–Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.

[25] Self, S.G. & Mauritsen, R.H. (1988). Power/sample size calculations for generalized linear models, *Biometrics* **44**, 79–86.

[26] Sheehe, R.R. (1962). Dynamic risk analysis in retrospective matched pair studies of disease, *Biometrics* **18**, 323–341.

[27] Thomas, D.C. (1977). Addendum to a paper by Liddell, F.D.K., McDonald, J.C. & Thomas, D.C., *Journal of the Royal Statistical Society, Series A* **140**, 483–485.

[28] Ury, H.K. (1975). Efficiency of case–control studies with multiple controls per case: continuous or dichotomous data, *Biometrics* **31**, 643–649.

[29] Wacholder, S. (1995). Design issues in case–control studies, *Statistical Methods in Medical Research* **4**, 293–309.

[30] Wacholder, S. (1996). The case–control study as data missing by design: estimating risk differences, *Epidemiology* **7**, 144–150.

[31] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case–control studies, I. Principles, *American Journal of Epidemiology* **135**, 1019–1028.

[32] Wacholder, S., Silverman, D.T., McLaughlin, J.K. & Mandel, J.S. (1992). Selection of controls in case–control studies, II. Types of controls, *American Journal of Epidemiology* **135**, 1029–1041.

[33] Wacholder, S., Silverman, D.T., McLaughlin, J.K. & Mandel, J.S. (1992). Selection of controls in case–control studies, III. Design options, *American Journal of Epidemiology* **135**, 1042–1051.

[34] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **73**, 40–46.

[35] Weinberg, C.R. & Sandler, D.P. (1991). Randomized recruitment in case–control studies, *American Journal of Epidemiology* **134**, 421–432.

[36] White, J.E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

[37] Willett, W.C., Stampfer, M.J., Colditz, G.A., Rosner, B.A. & Speizer, F.E. (1992). Relation of meat, fat and fiber intake to the risk of colon cancer in a prospective study among women, *New England Journal of Medicine* **323**, 73–77.

SHOLOM WACHOLDER & PATRICIA HARTGE

Case-only Gene Mapping

Genetic mapping of human disease **genes** refers to the identification of **markers** that are linked to the loci affecting disease status. When pedigree data are available, linkage studies are used to estimate the recombination fractions between marker and disease genes (**Linkage Analysis, Model-based**). This approach has proven to be very successful, although the relatively limited size of pedigrees does not allow very small recombination fractions to be estimated because of the lack of recombinants. Fine-scale mapping rests on the indirect approach of estimating population genetic parameters whose size depends on the recombination fraction, and these parameters are referred to collectively as "**association** parameters".

When individuals can be characterized as being affected or unaffected by a disease, marker–trait association can be addressed by comparing marker frequencies between these two categories (**Disease-marker Association**). The **case–control** approach compares frequencies between people with the disease and those chosen to be their matched controls. Alternatively, a comparison of the frequencies with which alternative marker alleles are transmitted to affected offspring forms the basis for the transmission disequilibrium class of tests.

It is also possible to infer linkage or association on the basis of marker data from affected individuals only. The extent to which individuals of a specified relationship share marker alleles can be predicted from classic **population genetics**, and linkage is inferred when there is more sharing than expected among affected relatives. This article describes procedures based on association among marker alleles within the affected population, and shows how these marker associations allow inferences to be drawn for marker–disease associations. Some of these ideas have been presented previously [1, 2].

Population Genetic Model

A disease is supposed to be affected by a locus **A**, with alleles A_r, in the sense that the probability of an individual of **genotype** A_rA_s being affected is $\phi_{rs} = \phi_{sr}$. If the population proportion of A_rA_s genotypes is P_{rs}, then the disease prevalence ϕ is $\sum_{r,s} P_{rs}\phi_{rs}$. When multiple loci contribute to disease susceptibility, the quantities P_{rs} and ϕ_{rs} would need to be extended. For a random-mating population, genotype proportions can be expressed as products of allele proportions, p_r for A_r, so that $\phi = \sum_{r,s} p_r p_s \phi_{rs}$. Unless a **candidate gene** is being considered, the disease genotypes and genotype-specific susceptibilities are unknown. This means that the number of alleles at the disease locus is also unknown.

Data can be collected on **marker** loci **M** with alleles M_i. If the proportion of individuals in the population formed from the union of A_rM_i and A_sM_j gametes is written as P_{sj}^{ri}, then the proportion of M_iM_j marker genotypes among affected or unaffected individuals is

$$\begin{aligned}\Pr(M_iM_j|\text{Aff.}) &= \sum_{r,s} P_{sj}^{ri}\frac{\phi_{rs}}{\phi},\\ \Pr(M_iM_j|\text{Unaff.}) &= \sum_{r,s} P_{sj}^{ri}\frac{1-\phi_{rs}}{1-\phi}.\end{aligned} \tag{1}$$

Under the **null hypothesis** of no marker–disease gene association, $P_{sj}^{ri} = P_{rs}P_{ij}$ and both these proportions reduce to the marker genotype proportion P_{ij}. The case–control test based on marker genotypes, therefore, is actually a test of the hypothesis $P_{sj}^{ri} = P_{rs}P_{ij}$. In the special case of a random union of gametes, one-locus genotype proportions are products of allele proportions and two-locus genotype proportions are products of gamete proportions: $P_{sj}^{ri} = p_{ri}p_{sj}$. Writing gamete frequencies in terms of allele frequencies and **linkage disequilibrium** coefficients, $p_{ri} = p_rq_i + D_{ri}$, leads to

$$\begin{aligned}\Pr(M_iM_j|\text{Aff.}) &= q_iq_j + \frac{1}{\phi}(q_i\delta_j + q_j\delta_i + \delta_{ij}),\\ \Pr(M_iM_j|\text{Unaff.}) &= q_iq_j - \frac{1}{1-\phi}(q_i\delta_j + q_j\delta_i + \delta_{ij}),\end{aligned} \tag{2}$$

where the marker allelic and genotypic association parameters are defined as

$$\begin{aligned}\delta_i &= \sum_{r,s} p_sD_{ri}\phi_{rs},\\ \delta_{ij} &= \sum_{r,s} D_{ri}D_{sj}\phi_{rs}.\end{aligned} \tag{3}$$

The marker-genotype case–control test is therefore a test of no marker–disease allele and genotype association, $\delta_i = \delta_{ij} = 0$. The random mating assumption has allowed three- and four-allele disequilibrium coefficients to be ignored. Only in the special case of two alleles at the disease locus can the test said to be one for linkage disequilibrium.

Adding more marker genotypes to recover marker allele frequencies provides, in the random-mating case,

$$\Pr(M_i|\text{Aff.}) = q_i + \frac{\delta_i}{\phi},$$
$$\Pr(M_i|\text{Unaff.}) = q_i - \frac{\delta_i}{1-\phi}, \qquad (4)$$

showing that the case–control test on marker allele frequencies is actually a test for no marker–allelic association, $\delta_i = 0$.

Case-only Tests

From the results in the previous section it is possible to express marker-genotype proportions in terms of marker allele proportions among affected or unaffected people. For the random-mating case,

$$\Pr(M_iM_i|\text{Aff.}) = \Pr(M_i|\text{Aff.})^2 + \frac{\phi\delta_{ii} - \delta_i^2}{\phi^2},$$
$$\Pr(M_iM_j|\text{Aff.}) = 2\Pr(M_i|\text{Aff.})\Pr(M_j|\text{Aff.}) + 2(\phi\delta_{ij} - \delta_i\delta_j)\phi^2, \qquad (5)$$

showing that a test for **Hardy–Weinberg** proportions among marker genotypes in the affected population is actually a test about marker allele and genotype associations with the disease locus. The test is not strictly a test that these associations are zero, as the marker locus will have Hardy–Weinberg proportions among affected people even if it is associated with the disease locus when the susceptibilities are multiplicative, $\phi_{rs} = \alpha_r\alpha_s$, for then $\phi\delta_{ij} = \delta_i\delta_j$.

The absence of a linkage disequilibrium between marker and disease loci results in Hardy–Weinberg equilibrium for the marker in the affected population. The converse does not hold unless there are only two alleles at the disease locus. Parallel results apply to the unaffected population, so that either population could be used. For rare diseases, however, $\phi < (1-\phi)$, and the departures from Hardy–Weinberg at the marker locus are expected to be greater among affected than unaffected people.

If the disequilibrium coefficients involving marker and disease alleles are all zero, whether these are for one or two alleles at each locus, then the amount of Hardy–Weinberg disequilibrium at the marker locus is the same for affected people, unaffected people, and the whole population. Rejecting Hardy–Weinberg at the marker locus in the affected population, therefore, may simply reflect nonrandom mating in the whole population. Preliminary tests for marker Hardy–Weinberg for a random sample taken without regard to disease status should therefore be conducted. If this test gives a significant result, then it is necessary to go back to case–control tests although even for those tests the power will be affected by nonrandom mating.

The analysis of marker-only associations can be extended to multiple marker loci. Tests of linkage disequilibrium for pairs of marker loci among affected people, for example, are actually tests for association of marker **haplotypes** with the disease locus in the whole population. These associations are functions of the linkage disequilibrium coefficients for each marker allele and each disease allele, and of the three-locus linkage disequilibrium coefficient for alleles from each of the three loci.

References

[1] Nielsen, D.M., Ehm, M.G. & Weir, B.S. (1999). Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus, *American Journal of Human Genetics* **63**, 1531–1540.

[2] Nielsen, D.M. & Weir, B.S. (1999). A classical setting for associations between markers and loci affecting quantitative traits, *Genetic Research* **74**, 271–277.

B.S. WEIR & D.M. NIELSEN

Categorical Data Analysis

Introduction

Categorical variables separate observations into groups, within which members share a common trait. This may be a nominal attribute, level of an ordinal scale, or numerical value or range derived from an interval or ratio scale (*see* **Measurement Scale**). In practice, the number of groups (categories) per variable is usually small to moderate, no more than 20. However, finer classifications are often defined by combinations of several variables. Procedures such as **Student's *t*-tests**, **analyses of variance** (ANOVAs), and their **nonparametric** analogues employ categorizations as independent or **explanatory variables** to study how continuous **random variables** vary between groups. *Categorical data analysis,* in contrast, involves categorical **response variables** and draws inferences to probability distributions of random category counts, or functions of them.

For instance, Table 1 reports the pretreatment percent labelling index (LI, continuous) and posttreatment remission status (RS, dichotomous) of acute myeloblastic leukemia patients. LI, which reflects the proportion of cells undergoing DNA synthesis, might be expected to predict RS. The relevant **null hypothesis** of independence implies mathematically that the conditional densities of LI, in the two groups differentiated by subsequent remission status ($RS = 1$ or $RS = 0$), are identical. However, testing for a difference in LI densities between these two *ex post facto*-differentiated populations with the usual Student's t or **Wilcoxon** rank sum test *is not* categorical data analysis, because these tests do not treat variation across the categories stochastically. It is biologically more natural, though, to view the RS outcomes for different patients as resulting from independent Bernoulli trials (*see* **Binary Data**) with fixed but unknown conditional probabilities $\pi_{LI} = \Pr\{RS = 1|LI\}$. One might then examine the **regression** of π_{LI} on LI. This latter approach incorporates the randomness of the RS dichotomy into analysis, and hence, falls within our scope.

When all variables are categorical, the cumbersome "list" or "case-record" format of Table 1 may be condensed by combining observations with identical values, that is, "levels", of all variables. Each observed pattern is listed once, with an added column for the count of such observations. When analysis is focused on one dependent variable, patterns identical for all independent variables share the same row, with added columns for counts in each dependent category (Table 2). A **contingency table** is a multiway cross-tabulation of counts for nonoverlapping groups. The dimensions correspond to categorical variables, the levels of which combine to define the groups. Such tables are usually more interpretable than condensed list formats, when most possible patterns of levels have been observed at least once. Table 3 shows the simplest example, a **two-by-two table** defined by combinations of levels of variables A and B. Rows and columns are labelled at tabular edges, or "margins", by levels of the corresponding variables. The physical intersection of a row and column is the "cell" corresponding to the combination of corresponding levels of each variable, and contains the count n_{ij} of observations with this pattern. Row and column sums are often placed at the right or bottom margins, respectively. Throughout, we replace a subscript with "+" to denote a sum of all terms with different values of the subscript. For certain study designs, specific combinations of levels may be unobservable or logically impossible. Corresponding cells are marked with a dash or other placeholder, or left empty, and the tables containing them are called "incomplete".

These concepts extend directly to tables of three or more variables. For convenience, counts from a contingency table of any dimension may be viewed in a two-dimensional rectangular array by nesting categories of some variables within those of others to form the rows (e.g. Table 2), columns, or both. Analogies between contingency tables and their counterpart tables of means in ANOVAs for **factorial experiments** have contributed greatly to the development of categorical data models.

Several features combine to give categorical data analysis a distinct flavor. Probability modeling is primarily confined to the **Poisson distribution**, for which the variance equals the mean, and its conditional descendants for which variability remains analytically inseparable from location. Hence, efficient analyses require weighting of observations. The dimension of the data space is tied to the number of observable category combinations rather than the number of categorical variables. Consequently, models are frequently of high dimension, requiring reliance on asymptotic inference (*see* **Large-sample**

Table 1 Pretreatment percent labelling index and posttreatment remission status (1 = in remission, 0 = relapsed) of acute myeloblastic leukemia patients; excerpted from [3]

Patient	Remission status	Percent labelling index
1	1	1.9
2	1	1.4
3	0	0.8
4	0	0.7
...	...	...
27	0	0.7

Table 2 Count list format for one dependent categorical variable: artificial data

Variable *A* level	Variable *B* level	Variable *C*		
		Level 1	Level 2	Level 3
1	1	5	7	4
1	2	6	3	2
1	3	11	2	14
2	1	1	12	9
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
17	2	8	13	1

Table 3 A 2×2 contingency table: cell and marginal counts

Variable *A*	Variable *B*		
	Level one	Level two	Marginal totals
Level one	n_{11}	n_{12}	n_{1+}
Level two	n_{21}	n_{22}	n_{2+}
Marginal totals	n_{+1}	n_{+2}	n_{++}

Theory) and/or complex computations. Most models for expected counts are inherently nonlinear, so **maximum likelihood estimates** (MLEs) for parameters are implicit nonlinear functions, rather than explicit linear functions, of observed marginal distributions or other summary statistics. This introduces subtle issues of model interpretation, and brings computational complexity into even low-dimensional settings. Lastly, complex structures of counts are modeled and smoothed by imposing symmetry constraints, analogous to those of factorial ANOVA or **polynomial regression**, on sets of either probabilities or model parameters. However, the inherent plausibility of constraints such as linearity in classical regression extends only infrequently to the categorical data setting, where a thoroughgoing empiricism flourishes by necessity.

Nevertheless, the same scientific questions about relationships among variables prompt continuous and categorical data analyses, which have been conceptually unified to a remarkable degree. For inference about a single proportion, *see* **Proportions, Inferences, and Comparisons**. Below, we treat the 2×2 table at length, since the most important concepts appear in simplest form at this level. Parallels with continuous data analysis are noted in discussions of general two- and three-dimensional tables. We then broadly survey categorical data modeling via unifying paradigms: **generalized linear models** (GLMs), weighted **least-squares** (WLS) functional modeling, **generalized estimating equations** (GEE), and **generalized linear mixed models** (GLMMs). Some attention is also given to **exact inference**, conditional logistic regression (*see* **Logistic Regression, Conditional**), and **Bayesian methods**. The need for brevity requires neglect of history. For sake of readability given the voluminous literature, we also omit detailed citation, providing instead a selected bibliography of texts, monographs, and a few papers, primarily reviews. For historical discussion and further references, see [2] and specific entries cross-referenced below. For additional remarks on comparison of two proportions, *see* **Proportions, Inferences, and Comparisons**. We confine our discussion to data that have been fully and accurately observed. Categorical data may also be missing or misclassified. (*See* **Missing Data**; **Misclassification Error**; **Misclassification Models**.)

The discussion of 2×2 tables is more detailed and in some portions more mathematical than the sections that follow. The reader desiring to learn the flavor of the subject with a bit less intensity may prefer to skim the sections entitled "Exact Inference for 2×2 Tables" and "Likelihood Ratio, Score and Wald Statistics," in favor of easier sledding beyond, and perhaps also refer to the entry on **Proportions, Inferences and Comparisons**.

Probability Models for 2 × 2 Contingency Tables

Four basic sampling models, the product-Poisson, multinomial, product-binomial, and noncentral hypergeometric, are commonly used to describe how data

in 2×2 tables such as Table 3 originate. We write m_{ij} for $E(n_{ij})$ under such a model. The general product-Poisson model, employed in spatial distribution and **incidence density** studies, assumes that four independent Poisson streams of events or individuals are counted over possibly different regions of space and/or time. Then

$$\Pr(\{n_{ij}\}) = \prod_{i=1}^{2}\prod_{j=1}^{2} \Pr(n_{ij}) = \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{(N_{ij}\lambda_{ij})^{n_{ij}} e^{-N_{ij}\lambda_{ij}}}{n_{ij}!}, \quad (1)$$

the product-Poisson distribution with rate parameters $\{\lambda_{ij}\}$ and known space-time "exposures" $\{N_{ij}\}$. For such data, $m_{ij} = N_{ij}\lambda_{ij}$. Typically, one is interested in how the λ_{ij} depend upon the levels of A and B, and particularly in how the ratios $\lambda_{i1}/\lambda_{i2}$ change with i, or how the ratios $\lambda_{1j}/\lambda_{2j}$ change with j. If these ratios do not change, then $\psi = \lambda_{11}\lambda_{22}/\lambda_{12}\lambda_{21} = 1$, and the $\ln \lambda_{ij}$ satisfy the additive model

$$\ln \lambda_{ij} = \mu + \gamma_{i*} + \gamma_{*j} \quad (2)$$

for some μ, γ_{i*}, and γ_{*j}. Such linear models for logarithms are also called **loglinear models**.

A simpler model results when the observational region is the same for all streams, with individuals or events classified jointly by variables A and B as they are observed. Then all $N_{ij} = N$ can be absorbed into the scale of the λ_{ij} and dropped from consideration. After rescaling, $m_{ij} = \lambda_{ij}$. Unless stated otherwise, "product-Poisson" below will refer to this simplified version.

The product-Poisson model for Table 3 generates other models through conditioning on marginal totals when these are fixed by design. For studies such as cross-sectional surveys that collect precisely n_{++} observations,

$$\Pr(\{n_{ij}\}) = n_{++} \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}, \quad (3)$$

a four-category **multinomial** distribution with the *unconditional* probabilities $\pi_{ij} = \Pr(A = i, B = j) = \lambda_{ij}/(\sum_{i=1}^{2}\sum_{j=1}^{2}\lambda_{ij})$. Here, $m_{ij} = n_{++}\pi_{ij}$. Central to such studies is whether the row and column categorizations A and B are informative about one another or independent. The answer lies with the ratio of the conditional **odds** π_{11}/π_{12} that $B = 1$ given $A = 1$ to the conditional odds π_{21}/π_{22} that $B = 1$ given $A = 2$,

$$\psi = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{m_{11}m_{22}}{m_{12}m_{21}} = \frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}}. \quad (4)$$

Since ψ is invariant when the roles of A and B are reversed, it is known simply as the **odds ratio**; $\psi = 1$ under independence and is higher or lower respectively when A and B have the same levels more or less often than independence would imply.

Two-group comparative designs, such as **case–control** or **cumulative incidence** cohort studies, have predetermined sums along one margin (e.g. n_{1+}, n_{2+}). Conditioning on these yields

$$\Pr(\{n_{ij}\}|\{n_{i+}\}) = \prod_{i=1}^{2} \binom{n_{i+}}{n_{i1}} \pi_i^{n_{i1}}(1-\pi_i)^{n_{i2}}, \quad (5)$$

the product of binomial distributions with sample sizes n_{i+} and **conditional probabilities** $\pi_i = \Pr(B = 1|A = i) = \pi_{i1}/(\sum_{j=1}^{2}\pi_{ij}) = \lambda_{i1}/(\sum_{j=1}^{2}\lambda_{ij})$. Under (5), $m_{i1} = n_{i+}\pi_i$ and $m_{i2} = n_{i+} - m_{i1}$. The object of these designs is comparison of π_1 and π_2 through inference about the difference $\Delta = \pi_1 - \pi_2$, the ratio $RR = \pi_1/\pi_2$, or the ratio $OR = \Omega_1/\Omega_2$ of the corresponding conditional odds $\Omega_i = \pi_i/(1-\pi_i)$. The first two measures of disparity are known to epidemiologists as the "risk difference" (*see* **Absolute Risk**) and "risk ratio" (*see* **Relative Risk**), respectively, while the latter is known as the risk odds ratio (ROR) for **cohort** designs, and as the exposure odds ratio (EOR) for case–control designs. Simple algebra shows that $OR = \Omega_1/\Omega_2$ is ψ, equal to unity under the homogeneity null hypothesis H_0 : $\Delta = \pi_1 - \pi_2 = 0$, and higher or lower depending on the sign of Δ. A fourth probability model is used when either "exact" or randomization analysis (*see* **Randomization Tests**) is desirable, as with small samples or some randomized **clinical trials**. In such a trial, one may hypothesize that a dichotomous outcome such as survival or death is preordained in the patient sample studied, in the sense of being unaffected by therapy. Once patients are selected, results then depend only on the outcome of **randomization**. Under this null assumption of no therapeutic impact, conditioning on *both* the designed treatment group sizes *and* the predetermined outcome counts for the selected patient group yields a null **hypergeometric distribution**. More generally, when the

row and column variables are dependent, the same conditioning gives the noncentral hypergeometric distribution

$$\Pr(\{n_{ij}\}|\{n_{i+}\},\{n_{+j}\},\psi) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}\psi^{n_{11}}}{\sum_u \binom{n_{1+}}{u}\binom{n_{2+}}{n_{+1}-u}\psi^u}, \quad (6)$$

where $\max(0, n_{1+}+n_{+1}-n_{++}) \le u \le \min(n_{1+}, n_{+1})$, and ψ is again the odds ratio $\pi_{11}\pi_{22}/\pi_{12}\pi_{21} = \lambda_{11}\lambda_{22}/\lambda_{12}\lambda_{21}$. Since ψ is the only unknown quantity, inferences based on (6) are free of **nuisance parameters**. When $\psi = 1$, (6) is the hypergeometric distribution, for which $m_{ij} = n_{i+}n_{+j}/n_{++}$.

Inference for 2 × 2 Tables

The odds ratio ψ is central to all probability models for 2×2 tables. Under a product-Poisson law, ψ reflects nonadditivity of the effects of A and B on the $\ln \lambda_{ij}$. The multinomial, product-binomial, and noncentral hypergeometric admit equivalent formulations of ψ as a measure of **association** between A and B. In each case, writing $\ln \psi = \ln m_{11} - \ln m_{12} - \ln m_{21} + \ln m_{22}$ expresses $\ln \psi$ as the usual **interaction** contrast for a two-way layout in the general linear model, applied here to the $\ln m_{ij}$ rather than to cell means of Gaussian variates. Thus, ψ is the interaction parameter of a loglinear model for variation among the $\ln m_{ij}$ that is structurally identical to the 2×2 factorial ANOVA model for expected cell means in the two-way layout. Conveniently, the MLE of ψ under Poisson, multinomial, and product-binomial laws is just the sample odds ratio $\widehat{\psi} = n_{11}n_{22}/n_{12}n_{21}$. It is thus attractive to characterize association in a 2×2 table by ψ, although the risk difference Δ or the risk ratio RR at times represent the association on a more practically relevant scale. We now describe methods of estimation and testing for these parameters, using **likelihood** inference or related asymptotically equivalent methods.

Exact Inference for 2 × 2 Tables

Under (6), $\Pr(\{n_{ij}\})$ is known for any specified value of ψ, the only unknown parameter. This allows exact probability calculations, and hence **exact inference**, either when both margins have been fixed by design so that (6) applies directly, or after conditioning on random margin(s) under (1), (3), or (5). The MLE under (6), $\widehat{\psi}_c (\neq \widehat{\psi})$, is known as the "conditional MLE" of ψ, and may be found iteratively as a root of the polynomial equation $n_{11} = E(n_{11}|\{n_{i+}\},\{n_{+j}\},\psi)$. The hypothesis $H_0 : \psi = \psi_0$ may be tested against $H_a : \psi \neq \psi_0$, using as ***P* value** the summed probabilities of the observed $\{n_{ij}\}$ and all tables $\{n^*_{ij}\}$ with identical margins $\{n_{i+}\}$ and $\{n_{+j}\}$ for which $\Pr(\{n^*_{ij}\}) \le \Pr(\{n_{ij}\})$ under H_0 (*see* **Hypothesis Testing**). For $\psi_0 = 1$, this is **Fisher's exact test**. For a one-tailed test, only tables in the direction of the alternative ($n^*_{11} \le$ or $\ge n_{11}$, as applies) contribute to the P value. Lower and upper $100(1-\alpha)\%$ confidence limits ψ_L and ψ_U for ψ may be found by inverting the two-tailed test. This requires iteratively solving $\alpha/2 = \sum_{u' \ge n_{11}} \Pr(u'|\{n_{i+}\},\{n_{+j}\},\psi_L)$ and $\alpha/2 = \sum_{u' \le n_{11}} \Pr(u'|\{n_{i+}\},\{n_{+j}\},\psi_U)$, where $\Pr(u'|\{n_{i+}\},\{n_{+j}\},\psi)$ is given by (6) with $n_{11} = u'$.

The above method may be extended to a broader class of procedures, and from 2×2 to $r \times c$ tables through the multivariate hypergeometric generalization of (6), as follows: (i) Order all possible tables with the observed $\{n_{i+}\}$ and $\{n_{+j}\}$ by a measure of discrepancy from H_0; (ii) test H_0 against H_a, using as P value the summed probabilities of all tables at least as compatible with H_a, under the selected ordering, as the observed table; (iii) form a $100(1-\alpha)\%$ two-sided **confidence interval** for the relevant unknown parameter θ (ψ in this instance), by including in the interval all θ_0 retained by such a two-sided α-level test of $H_0 : \theta = \theta_0$. Analogous methods can be used whenever a distribution giving exact probabilities under a simple hypothesis for all observable tables, and a reasonable method of ordering tables, are both available. For Fisher's exact test, the ordering criterion is the probability of the table under H_0 and (6). Other ordering criteria of interest include the asymptotic test statistics discussed in the following section and measures of row by column association. Computational barriers to such tests are falling rapidly.

Owing to the discreteness of (6), hypothesis tests that reject H_0 when an exact P value is $\le \alpha$ are conservative, in the sense that the actual type I error probability is typically below α, sometimes considerably so. For instance, for the famous tea tasting experiment Fisher used in presenting his exact test, the true type I error probability of the nominal $\alpha = 5\%$ level test

is 1.4%. Several approaches are available for gaining statistical power by more closely approaching the nominal level. One is to randomize Fisher's exact test. Under this approach, rejection or retention of H_0 for an outcome yielding the lowest nonsignificant P value under the exact test is determined randomly, with rejection probability set just sufficient to increase type I error to the desired level. This is rarely done in practice, however, because it allows researchers with identical data to reach discordant results based on a random process containing no information relevant to the scientific question.

A second approach, for data collected under (1), (3), or (5), is to enlarge the probability space, and thus, the set of possible P values, by testing under (5) rather than under (6). Such an "unconditional exact test" is accomplished by (i) exact testing under (5), conditional on each member of an interval $\pi\varepsilon[\pi_L, \pi_U]$ of presumed common probabilities $\pi_1 = \pi_2 = \pi$, and (ii) using the supremum of the conditional P values across the collection to determine the P value for inference. The fully unconditional approach takes $\pi_L = 0, \pi_U = 1$ and the supremum itself, $\sup_{\pi_0\varepsilon[0,1]}(P_{H_0:\pi=\pi_0})$, as the unconditional P value, while the formulation reduces to Fisher's exact test for the degenerate interval $\pi_L = \pi_U = n_{1+}/n_{++}$. Increased power over either of these extreme choices may often be obtained by maximizing only over a confidence interval for π, and adding the error probability of the interval to the supremum to obtain the unconditional P value; *see* **Proportions, Inferences, and Comparisons** for more detail.

A third approach bases inference on the "mid-P value", obtained through reducing the exact P value by half the probability of the observed table. Mid-P values mimic the null behavior of P values based on continuous sampling distributions more closely than do exact P values. A mid-P value has expectation 0.5 under H_0; mid-P values for one-sided tests in opposing directions sum to 1.0; and Pr(mid-P value $\leq \alpha$) is frequently much closer to α than Pr(exact P value $\leq \alpha$). However, type I error control for tests based on a mid-P value is less stringent than for conditional or unconditional exact tests, randomized tests, or for any hypothesis test properly constructed from a valid continuous sampling distribution, in the sense that Pr(mid-$P \leq \alpha$) can exceed α for some sample sizes. Similarly, the coverage probability of a confidence interval constructed by inverting mid-P-based hypothesis testing may fall short of its nominal confidence coefficient for some combinations of true probabilities and sample sizes. The mid-P value appears to be a satisfactory approximate remedy for the conservatism of Fisher's exact test in that the type I error from its repeated use in different situations can be expected to approximate the nominal level better than the error rate of Fisher's exact test. But the looser form of type I error control may compromise the suitability of mid-P values for use in settings where precise type I error control is essential to the integrity of a regulatory standard. *See* **Continuity Correction**; **Proportions, Inferences, and Comparisons**.

While inference should clearly be based on (6) when both margins of a 2×2 table are fixed by experimental design, there is no clear consensus on the best level of conditioning, and hence distributional model, to employ for inference from data generated by (1), (3), or (5). However, large-sample methods have been developed, based primarily though not exclusively on these less-conditional distributions, that are computationally simpler than exact methods, and more easily generalized to higher-dimensional contingency tables and statistical modeling of association structures involving several variables. Such large-sample methods dominate statistical practice. For moderate to large samples, these methods tend to give results close to those based on (6) and its multivariate extensions. They may not be reliable for use with sparse samples or in the context of highly eccentric distributions, but in such situations, often still provide a framework for exact analyses.

Approximate Large-sample Inference for 2 × 2 Tables

A hypergeometric-based large-sample test depends on the asymptotically Gaussian (**normal**) distribution of n_{11} as $n_{++} \uparrow \infty$ with $n_{1+}/n_{++} \to \pi_{1+}$ and $n_{+1}/n_{++} \to \pi_{+1}$. Then, by randomization **central limit theory**, the **Mantel–Haenszel** chi-square statistic

$$X^2_{\mathrm{MH}} = \frac{(n_{++}-1)(n_{++}n_{11} - n_{1+}n_{+1})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} = \frac{[n_{11} - (n_{1+}n_{+1}/n_{++})]^2}{\mathrm{Var}_{H_0}(n_{11})} = \frac{(n_{11} - m_{11})^2}{\mathrm{Var}_{H_0}(n_{11})} \tag{7}$$

has a limiting χ^2_1 distribution (**chi-square distribution** with one **degree of freedom**). Several other test statistics have the same null limiting distribution under product-Poisson, multinomial, or product-binomial models. Under these models, the MLE of m_{ij} remains $\widehat{m}_{ij} = n_{i+}n_{+j}/n_{++}$ under $H_0 : \psi = 1$, and is n_{ij} under the general alternative $H_a : \psi \neq 1$. The **chi-square likelihood ratio test** statistic for testing H_0 against H_a reduces in each case to $G^2 = -2 \ln \Lambda = 2 \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} \ln(n_{ij}/\widehat{m}_{ij})$. G^2 is sometimes abbreviated as $2 \sum \sum O \ln(O/\mathrm{E})$, where O and E respectively represent an observed cell count and its estimated expected value under H_0. However, the Pearson chi-square statistic $X^2_P = \sum_{i=1}^{2} \sum_{j=1}^{2} (n_{ij} - \widehat{m}_{ij})^2/\widehat{m}_{ij} = \sum \sum (O - \mathrm{E})^2/\mathrm{E}$, like X^2_{MH} a normalized, squared-error comparison of the n_{ij} with the $\widehat{m}_{ij}$, is more commonly used than either G^2 or X^2_{MH}. Since $X^2_{\mathrm{MH}} = [(n_{++} - 1)/n_{++}]X^2_P$, the distinction between these tests is minimal except when large-sample approximation is difficult to justify. It becomes important, however, in extensions to three-way tables (see Cochran–Mantel–Haenszel Tests below, and **Chi-square Distribution**). X^2_P and X^2_{MH} are sometimes modified to "continuity corrected" versions $X^2_{P_c}$ and $X^2_{\mathrm{MH}_c}$ by reducing each $|O - \mathrm{E}|$ in X^2_P, and equivalently $|n_{11} - m_{11}|$ in X^2_{MH}, by 1/2 to better approximate the behavior of Fisher's exact test (*see* **Yates's Continuity Correction**).

An alternate algebraic form of Pearson's statistic under product-binomial sampling is $X^2_P = z^2_P = (p_1 - p_2)^2/(n^{-1}_{1+} + n^{-1}_{2+})\overline{p}(1 - \overline{p})$, where $p_i = n_{i1}/n_{i+}$ and $\overline{p} = n_{+1}/n_{++}$. Here $p_1 - p_2$ is the MLE $\widehat{\Delta}$ of Δ, the denominator is the MLE $\widehat{\mathrm{Var}}_0(\widehat{\Delta})$ of $\mathrm{Var}(\widehat{\Delta})$ under $H_0 : \Delta = 0$, and z_P is a conventional large-sample **"standard normal deviate"** z-statistic. Since $\widehat{\mathrm{Var}}_0(\widehat{\Delta})$ is not consistent for $\mathrm{Var}(\widehat{\Delta})$ when $\Delta \neq 0$, z_P does not yield closed-form bounds of an asymptotic confidence interval for Δ. Instead, one must numerically invert the more general version of the Pearson criterion,

$$X^2_{P;\Delta_0} = \frac{((p_1 - p_2) - \Delta_0)^2}{\sum_{i=1}^{2} (\widetilde{\pi}_i(1 - \widetilde{\pi}_i)/n_{i+})}, \qquad (8)$$

where the $\widetilde{\pi}_i$ are MLEs of the π_i under the constraint $\widetilde{\pi}_1 - \widetilde{\pi}_2 = \Delta_0$, and the confidence interval is bounded by the lowest and highest values of Δ_0 retained by the hypothesis test.

A more convenient alternative to this computation is to instead estimate $\mathrm{Var}(\widehat{\Delta})$ by either its unrestricted MLE $\widehat{\mathrm{Var}}(\widehat{\Delta}) = \sum_{i=1}^{2} p_i(1 - p_i)/n_{i+}$ or **unbiased** estimator $\widetilde{\mathrm{Var}}(\widehat{\Delta}) = \sum_{i=1}^{2} p_i(1 - p_i)/(n_{i+} - 1)$, to form respective large-sample Gaussian confidence intervals $\widehat{\Delta} \pm z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{Var}}(\widehat{\Delta}))}$ or $\widehat{\Delta} \pm z_{1-\alpha/2}\sqrt{(\widetilde{\mathrm{Var}}(\widehat{\Delta}))}$. Indeed, $\widehat{\Delta} \pm z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{Var}}(\widehat{\Delta}))}$ is the "standard" interval in statistical pedagogy. Unfortunately, this interval is now known to be overly liberal, with slow convergence to its asymptotic coverage, below-nominal coverage in small to moderate samples, and performance notably inferior to that of the interval based on (8) and other alternatives. However, a practical and pedagogically useful replacement for this errant standard results from smoothing the 2×2 table slightly toward uniformity by adding one to each cell prior to calculating the interval. This correction is equivalent to basing the Gaussian interval not on the p_i, but rather on **Bayesian** point estimates of π_1 and π_2, chosen as posterior modes using independent, **uniform prior distributions** for the π_i. The resulting interval has the nominal behavior of its predecessor in large samples, and dramatically improved coverage in small samples, particularly when π_1 or π_2 approaches zero or one. As with any approximate interval, and intervals based on heuristic improvements to exact intervals such as mid-P, coverage can still be below nominal for specific combinations of n_{1+}, π_1, n_{2+}, and π_2.

Confidence intervals for ψ and RR may be obtained by an extension of the basic approach for Δ. The MLEs $\widehat{\psi}$ and $\widehat{RR}$ and their logarithms are smooth functions of the n_{ij}, and thus are asymptotically Gaussian. Convergence is faster on the symmetric logarithmic scale, where intervals including negative (hence inadmissible) values are also routinely avoided. From the **delta method**, the MLEs of the asymptotic standard errors (se's) of $\ln \widehat{\psi}$ and of $\ln \widehat{RR} = \ln[(n_{11}/n_{1+})/(n_{21}/n_{2+})]$ are respectively $\sqrt{(\sum \sum n^{-1}_{ij})}$ and $\sqrt{(n^{-1}_{11} - n^{-1}_{1+} + n^{-1}_{21} - n^{-1}_{2+})}$. Confidence intervals, generally symmetric, are determined on the ln scale using Gaussian critical values, and exponentiated to yield (asymmetric) intervals for ψ and RR. Although still approximate, these intervals behave more satisfactorily than does the Gaussian interval for Δ. Performance of the interval for ψ is improved by (i) extension of the interval to $\pm\infty$ respectively when $\min(n_{12}, n_{21}) = 0, \min(n_{11}, n_{22}) = 0$, and (ii) determining finite

boundaries as above after smoothing slightly toward independence, by distributing across the cells of the table a total of two additional observations in fractions proportionate to the $\widehat{m}_{ij}$. This is an **empirical Bayes** procedure in which the table is smoothed by applying a Dirichlet prior to the cell probabilities, with Dirichlet parameters summing to two and proportional to the estimated expected values under independence.

The confidence intervals for ψ, Δ, and RR, based on their respective MLEs and **consistent** asymptotic variance estimators, may each be used to generate alternative tests of $\psi = 1$. The statistic for Δ, $X_N^2 = \widehat{\Delta}^2/\widehat{\mathrm{Var}}(\widehat{\Delta})$, is of particular interest. X_N^2, known as Neyman's "minimum modified chi-square" statistic (*see* **Ban Estimates**), may be rewritten as $X_N^2 = \sum_{i=1}^2 \sum_{j=1}^2 (n_{ij} - \widetilde{m}_{ij})^2/n_{ij} = \sum\sum(O - \mathrm{E})^2/O$, where $\widetilde{m}_{i1} = n_{i+}\widetilde{p}$, $\widetilde{m}_{i2} = n_{i+} - \widetilde{m}_{i1}$, and $\widetilde{p}$ is the inverse-variance weighted average of the p_i. X_N^2 differs from X_P^2 in estimating expected counts slightly differently, and by weighting each squared deviation inversely to the corresponding observed rather than expected count. Though undefined when any cells are empty, X_N^2 converges in large samples to the same limiting distribution under $H_0 : \Delta = 0$ as do X_P^2 and G^2, as Neyman showed in conjunction with his development of best asymptotically normal (BAN) estimation.

Indeed, this limiting distribution is shared by a much broader class of test statistics, the "**power-divergence** (PD)" family of form

$$\mathcal{DIV}^{\varphi} = \left(\frac{2}{\varphi(\varphi+1)}\right) \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\left[\left(\frac{n_{ij}}{\overline{m}_{ij}}\right)^{\varphi} - 1\right], \tag{9}$$

where the $\overline{m}_{ij}$ are known or estimated expected counts under a null hypothesis. This is X_P^2 for $\varphi = 1$ and X_N^2 for $\varphi = -2$. The **likelihood ratio** G^2 is the limiting case as $\varphi \to 0$, and $\varphi \to -1$ yields a **Kullback–Liebler information** statistic. The minimum power-divergence estimator (mpe) $\{\widehat{m}_{ij}^{\varphi}\}$ minimizes $\mathcal{DIV}^{\varphi}$ under a null hypothesis H_0 for the m_{ij}. Thus, $\widehat{m}_{ij}$ and $\widetilde{m}_{ij}$ are respectively the MLE ($\varphi = 0$), and the "minimum Neyman chi-square" ($\varphi = -2$) estimators of m_{ij} under $H_0 : \Delta = 0$. Quite generally, mpes and PD statistics based on any mpe of m_{ij} share, respectively, common limiting null Gaussian and χ^2 distributions as $n_{++} \uparrow \infty$. (*See* **Power Divergence Methods**; **Proportions, Inferences, and Comparisons**; **Chi-square Tests**).

Likelihood Ratio, Score and Wald Statistics

Further progress requires more powerful and general tools. We therefore briefly sketch, heuristically, some likelihood results in quite general form. Consider a data vector $\boldsymbol{y}$ (e.g. $\boldsymbol{y} = (n_{11}, n_{12}, n_{21}, n_{22})'$) and associated loglikelihood function $l(\boldsymbol{\theta};\boldsymbol{y})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ is a parameter vector. One might wish to test $H_0: \boldsymbol{\theta}_2 = \mathbf{0}$. For instance, consider the saturated loglinear model for the 2×2 table,

$$\ln m_{ij} = \mu + \gamma_{i*} + \gamma_{*j} + \gamma_{ij} \tag{10}$$

under product-Poisson, multinomial, or product-binomial sampling. We assume that **identifiability** constraints are placed on the γ_{i*}, γ_{*j}, and γ_{ij}, so that μ, γ_{1*}, γ_{*1}, and $\gamma_{11} = \ln\psi$ determine the remaining parameters. Then $\boldsymbol{\theta}_1 = (\mu, \gamma_{1*}, \gamma_{*1})'$ and $\boldsymbol{\theta}_2 = (\gamma_{11})$ places testing of $\psi = 1$ or $\Delta = 0$ in this context.

When $l(\boldsymbol{\theta};\boldsymbol{y})$ is smooth in a neighborhood of $\boldsymbol{\theta}$, the MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1', \widehat{\boldsymbol{\theta}}_2')'$ is a solution of the likelihood equations $\partial l(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta} = \mathbf{0}$ and **converges in probability** to $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^{\circ} = (\boldsymbol{\theta}_1', \mathbf{0}')'$ and $\widehat{\boldsymbol{\theta}}^{\circ} = (\widehat{\boldsymbol{\theta}}_1^{\circ\prime}, \mathbf{0}')'$ be, respectively, $\boldsymbol{\theta}$ and its MLE under H_0. Under H_0, $\widehat{\boldsymbol{\theta}}_2$ from a large sample will likely be close to $\mathbf{0}$, a prediction that may be checked to evaluate the plausibility of H_0. Moreover, both $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^{\circ}$ will likely be close to $\boldsymbol{\theta}$, placing them both in the neighborhood of $\boldsymbol{\theta}$, where $l(\boldsymbol{\theta};\boldsymbol{y})$ is smooth. Thus, chances are that $l(\widehat{\boldsymbol{\theta}}^{\circ};\boldsymbol{y})$ will be close to $l(\widehat{\boldsymbol{\theta}};\boldsymbol{y})$, a second verifiable prediction of H_0. Finally, under H_0, the log-likelihood slope in any $\boldsymbol{\theta}_2$ direction has expectation $\mathbf{0}$ when evaluated either at $\boldsymbol{\theta}^{\circ}$ or $\widehat{\boldsymbol{\theta}}^{\circ}$, and should not often depart greatly from that. This may also be checked, evaluating at $\boldsymbol{\theta}^{\circ}$ if $\boldsymbol{\theta}_1$ is known and at $\widehat{\boldsymbol{\theta}}^{\circ}$ otherwise. We may thus test compatibility of the data with H_0 by checking these stochastically usual consequences for the parameter estimates, log-likelihood, and score.

The deviation of $l(\widehat{\boldsymbol{\theta}}^{\circ};\boldsymbol{y})$ from $l(\widehat{\boldsymbol{\theta}};\boldsymbol{y})$ is evaluated through the large-sample null χ_{ν}^2 distribution of $-2\ln\Lambda$, where $\nu = \mathrm{rank}\ \boldsymbol{\theta}_2$. The MLE and log-likelihood slope are evaluated in terms of their limiting **multivariate normal distributions**. The large-sample multivariate Gaussian distribution of $\widehat{\boldsymbol{\theta}}$ has mean $\boldsymbol{\theta}$ and $\mathrm{cov}_A(\widehat{\boldsymbol{\theta}}) = \mathcal{I}(\boldsymbol{\theta})^{-1}$, where $\mathcal{I}(\boldsymbol{\theta}) = -E([\partial^2 l(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'])$ is Fisher's **information matrix**. Consequently, the large-sample null marginal

distribution of the component $\widehat{\boldsymbol{\theta}}_2$ has mean $\mathbf{0}$ and $\text{cov}_A(\widehat{\boldsymbol{\theta}}_2) = \text{V}_{\widehat{\boldsymbol{\theta}}_2}(\boldsymbol{\theta})$, the lower right block of $\mathcal{I}(\boldsymbol{\theta})^{-1}$. Since this may be estimated consistently by $\text{V}_{\widehat{\boldsymbol{\theta}}_2}(\widehat{\boldsymbol{\theta}})$, the quadratic form $\text{Q}_W = \widehat{\boldsymbol{\theta}}_2' \, [\text{V}_{\widehat{\boldsymbol{\theta}}_2}(\widehat{\boldsymbol{\theta}})]^{-1}\widehat{\boldsymbol{\theta}}_2$ has the same null asymptotic χ^2_ν distribution as $-2\ln\Lambda$. Statistics of form Q_W are "Wald statistics", and corresponding chi-square tests are "Wald tests" (*see* **Likelihood**).

More generally, for estimation of a parameter vector $\boldsymbol{\theta}$, or a subvector $\boldsymbol{\theta}_2$ from a categorical data model, a "best asymptotically normal (BAN)" estimator $\boldsymbol{F}$ is a consistent, asymptotically efficient multivariate Gaussian estimator that has continuous partial derivatives with respect to the observed counts. BAN estimators share the same asymptotic null distribution, and mpes are BAN. Any quadratic form $\boldsymbol{F}'\widehat{\boldsymbol{V}}_{\boldsymbol{F}}^{-1}\boldsymbol{F}$, where $\widehat{\boldsymbol{V}}_{\boldsymbol{F}}$ is a consistent estimate of the asymptotic **covariance matrix** $\text{cov}_A(\boldsymbol{F})$ of $\boldsymbol{F}$, is an extended Wald statistic sharing the same asymptotic properties as Q_W under the null. Different extended Wald statistics for the same hypothesis share the same null asymptotic χ^2 distribution, though their non-null asymptotic cdfs may differ.

The log-likelihood slope $\boldsymbol{S}_2$, defined as $\partial l(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}_2$ evaluated at $\boldsymbol{\theta}_2 = \mathbf{0}$, is called the "likelihood score" or "Rao's efficient score" for $\boldsymbol{\theta}_2$. The large-sample null multivariate Gaussian distribution of $\boldsymbol{S}_2$ has $\text{cov}_A(\boldsymbol{S}_2) = \mathcal{I}_{22}(\boldsymbol{\theta}^\circ)$, the lower right block of $\mathcal{I}(\boldsymbol{\theta}^\circ)$. If $\boldsymbol{\theta}_1$ is known, then so is $\mathcal{I}_{22}(\boldsymbol{\theta}^\circ)$ under H_0. The "score statistic" for testing H_0 is then the quadratic form $Q_S(\boldsymbol{\theta}^\circ) = \boldsymbol{S}_2'(\boldsymbol{\theta}^\circ)[\mathcal{I}_{22}(\boldsymbol{\theta}^\circ)]^{-1}\boldsymbol{S}_2(\boldsymbol{\theta}^\circ)$ reflecting the magnitude of the score relative to its variability, which the "score test" evaluates using the χ^2_ν distribution as above. When $\boldsymbol{\theta}_1$ is unknown and the score is thus evaluated at $\widehat{\boldsymbol{\theta}}^\circ = (\widehat{\boldsymbol{\theta}}_1^{\circ\prime}, \mathbf{0}')'$, the use of the MLE $\widehat{\boldsymbol{\theta}}_1^\circ$ constrains $\boldsymbol{S}_1 = \partial l(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}_1$ to $\mathbf{0}$, and the observed score must be assessed against the conditional distribution under that restriction. The conditional covariance matrix is $\text{cov}_A[\boldsymbol{S}_2|(\boldsymbol{S}_1 = \mathbf{0})] = \mathcal{I}_{\theta_2|\theta_1} = \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}$, and the corresponding score statistic is $Q_S(\widehat{\boldsymbol{\theta}}^\circ) = \boldsymbol{S}_2'(\widehat{\boldsymbol{\theta}}^\circ)[\mathcal{I}_{\theta_2|\theta_1}(\widehat{\boldsymbol{\theta}}^\circ)]^{-1}\boldsymbol{S}_2(\widehat{\boldsymbol{\theta}}^\circ)$, also χ^2_ν under H_0. Equivalently, in terms of the overall score vector $\boldsymbol{S}' = (\boldsymbol{S}_1', \boldsymbol{S}_2')'$, this is $\boldsymbol{S}'(\widehat{\boldsymbol{\theta}}^\circ)\mathcal{I}(\widehat{\boldsymbol{\theta}}^\circ)^{-1}\boldsymbol{S}(\widehat{\boldsymbol{\theta}}^\circ)$.

Some score statistics are extended Wald statistics. For instance, in the **Poisson regression**, **logistic regression**, and loglinear models discussed below, a score statistic is a quadratic form in linear functions of residuals from the likelihood fit of H_0, with kernel the null-based MLE of their inverse asymptotic covariance matrix.

Most asymptotic test statistics in common use for categorical data take the form of either $-2\ln\Lambda$, Q_S, or Q_W, for appropriate distributions and parametric models. Wald and score tests are often substituted for likelihood ratio tests, because they are less computationally demanding when working with multivariable models. Maximum likelihood estimation frequently requires iterative numerical approximation. Each likelihood ratio test requires two MLEs, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_1^\circ$. In exploratory analyses, one may wish to examine the statistical significance of many individual parameters within a single model, using the results for model reduction. The MLE $\widehat{\boldsymbol{\theta}}$ for the baseline model is needed for every test, but $\widehat{\boldsymbol{\theta}}_1^\circ$ changes with each parameter or parameter set to be tested. In contrast, Wald tests depend only on $\widehat{\boldsymbol{\theta}}$. For instance, for testing 10 parameters individually, likelihood ratio testing requires 11 numerical optimizations as compared to 1 for the Wald test.

Similarly, when many candidate predictor variables are available for initiating a model or augmenting a baseline model, the score test is more convenient than the likelihood ratio test. Defining $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, as the parameter vectors of variables already in the model and of candidates for entry, only $\widehat{\boldsymbol{\theta}}_1^\circ$ from the baseline model is needed to construct a score statistic for testing candidate variables, individually or in groups, for incremental explanatory power. Although such computational economies have gradually become less important for individual models with the extraordinary increase in available computational power, computer software has tended to retain these practices to facilitate exploration of larger and more complex models.

The previous results for 2×2 tables can be better understood, and extended, after reinterpretation in the context of this section. Under product-binomial sampling, the hypothesis $H_0 : \Delta = 0$ in terms of probabilities is equivalent to $H_0 : \delta_i = 0$, $i = 1, 2$ in the model for expected cell counts $m_{i1} = n_{i+}(\mu + \delta_i)$, with a suitable identifiability constraint. In contrast to (10), this model may be reexpressed as linear constraints on the m_{ij}, since the n_{i+} are fixed in product-binomial sampling. Neyman chi-squares for such linear hypotheses are Wald statistics. Further, if nonlinear constraints on the m_{ij} in any hypothesis are replaced by their linear Taylor series approximations at $m_{ij} = n_{ij}$, then X^2_N for the "linearized" hypothesis

is a Wald statistic for its nonlinear parent. This gives a particularly simple recipe for asymptotically (first order) optimal inference, since minimum X^2_N estimators and the associated X^2_N may be obtained from one-step WLS computations.

Specifically, in the 2×2 context Neyman's X^2_N, obtained by inverting the large-sample Gaussian confidence interval for Δ, is the Wald test for $\Delta = 0$. Similarly, the **delta method**-based intervals for ψ and RR are inversions of Wald tests, which in this case are also linearized Neyman chi-square tests, for $\psi = \psi_0$ and $RR = RR_0$. G^2, and $X^2_P = X^2_{P;0}$ in (8), are respectively the likelihood ratio and score statistics for $\Delta = 0$, or $H_0 : \gamma_{11} = 0$, in (10).

More generally, Pearson's $X^2_{P;\Delta_0}$ (8) is the score statistic $Q_{S;\Delta_0}$ for testing $\Delta = \Delta_0$. Analogous score tests of $RR = RR_0$ and $\psi = \psi_0$ are obtained respectively from

$$Q_{\mathrm{S;RR}_0} = \frac{n_1(p_1 - \widetilde{\pi}_1)^2}{\widetilde{\pi}_1(1 - \widetilde{\pi}_1)} + \frac{n_2(p_2 - \widetilde{\pi}_2)^2}{\widetilde{\pi}_2(1 - \widetilde{\pi}_2)}, \qquad (11)$$

where $\widetilde{\pi}_i$ is the MLE of π_i under the constraint that $RR = RR_0$, and

$$Q_{\mathrm{S};\psi_0} = Cn_1(p_1 - \widetilde{\widetilde{\pi}}) \left[\frac{1}{n_1\widetilde{\widetilde{\pi}}_1(1 - \widetilde{\widetilde{\pi}}_1)} + \frac{1}{n_2\widetilde{\widetilde{\pi}}_2(1 - \widetilde{\widetilde{\pi}}_2)} \right], \qquad (12)$$

where $\widetilde{\widetilde{\pi}}_i$ is the MLE of π_i under the constraint that $\psi = \psi_0$ and C is a normalizing constant.

Comparative studies of score and Wald tests, and their associated confidence intervals, indicate that the score-based procedures generally approach nominal behavior more rapidly, and therefore, perform more acceptably in small to moderate samples, than the corresponding Wald procedures. As the associated computations have become less intimidating in the light of expanded computing power, the advantages of confidence intervals obtained by inverting two-sided tests based on the above score criteria have increasingly been recognized. Among asymptotically based intervals, these appear to maintain near-nominal coverage for a wide range of sample size and probability configurations.

For situations in which near is not enough, because nominal coverage in all conditions must be guaranteed, exact intervals are required. For this purpose as well, confidence intervals based on score statistics are increasingly used. In these situations, a score statistic is employed, as described earlier, as the criterion for ordering possible outcomes in computing an exact P value. Note that the standardization by estimated variance in the score statistics (8), (11), and (12) produces a finer partition of the sample space than the corresponding unstandardized measure of association, thus rendering exact analysis less discrete, and potentially reducing the conservatism of exact intervals. The difference between the true and nominal coverage of an exact interval can be further reduced, and the interval narrowed correspondingly, by unconditional rather than conditional exact testing, and by further restricting the range of unconditional analysis to values of the nuisance parameter that are not grossly implausible in light of the observed data.

Matched Pairs

The above discussion of 2×2 tables has omitted the simplest categorical analog to paired continuous data, 2×2 tables in which A and B represent a single dichotomy observed under different circumstances, or on members of n_{++} **matched pairs**. In such tables, association between the two dimensions is usually taken for granted, and ψ becomes a nuisance parameter. Interest shifts to discrepancies between the marginal probabilities π_{1+} and π_{+1}, expressed as the marginal risk difference $\Delta_M = \pi_{1+} - \pi_{+1}$, marginal risk ratio $RR_M = \pi_{1+}/\pi_{+1}$, or matched odds ratio $\psi_M = \pi_{12}/\pi_{21}$. Approaches introduced above may be used for inference about Δ_M, RR_M, and ψ_M. For Δ_M, see Marginal Homogeneity in Square Tables below (*see* **Square Contingency Table**). Matched categorical data are more easily treated generally, however, by viewing matched sets of observations as observational strata, with observations cross-classified in a three-way table with dimensions defined respectively by the strata, the condition of observation or other within-stratum factor, and the response variable. (*See* Cochran–Mantel–Haenszel tests below, and **McNemar Test**; **Matched Pairs With Categorical Data**; **Mantel–Haenszel Methods**).

Inference for $r \times c$ Tables

An $r \times c$ table may be viewed as a collection of 2×2 subtables, the challenge being to knit the subtables together in a reasonable structure for inference

that allows adequate error control. Ways of coping with the overparameterization in larger tables depend heavily on whether the variables A, B, or both are ordinal. Ordinality presents special opportunities for combining parameters across cells or incorporating scaling into a model for the m_{ij} (*see* **Ordered Categorical Data**).

The probability distributions defined above apply with $r \times c$ rather than $2 \times 2 = 4$ categories, or generalize directly: the noncentral hypergeometric to the multivariate noncentral hypergeometric, and the product-binomial to the product-multinomial. As with 2×2 tables, there are two distinct spheres of analysis. In the study of association between row and column variables, equivalently expressible as heterogeneity across rows of the conditional column distributions $\Pr(B = j|A = i) = m_{ij}/m_{i+}$, the null hypothesis is independence or conditional homogeneity, and the marginal distributions $\Pr(A = i) = m_{i+}/m_{++}$ and $\Pr(B = j) = m_{+j}/m_{++}$ are either known or of only secondary interest. Numerous indices have been developed to summarize, for particular purposes, the overall level of association in an $r \times c$ table. See **Association, Measures of** for enumeration and discussion of these. In contrast, in the study of repeated measurements or other matched data in square tables, the focus is on differences between these marginal distributions; the null hypothesis is marginal homogeneity and row by column association is of only secondary interest (*see* **Marginal Models**).

Marginal Homogeneity in Square Tables

We consider the simpler repeated measures case first. If the salient variable A is nominal, then H_0 : $m_{i+} - m_{+i} = 0, i = 1, \ldots, r$ is linear in the m_{ij}. In this case, $X_N^2 = Q_W$ is a quadratic form in an arbitrary $r - 1$ of the $(n_{i+} - n_{+i})$, with covariance matrix estimated at $m_{ij} = n_{ij}$, and is asymptotically χ^2_{r-1} under H_0. If A is ordinal with **scores** s_i, then comparison of the mean row score $\mu_r = \mathrm{E}(\overline{s}_r) = \mathrm{E}(n_{++}^{-1} \sum_{i=1}^{r} n_{i+}s_i)$ with the mean column score $\mu_c = \mathrm{E}(\overline{s}_c) = \mathrm{E}(n_{++}^{-1} \sum_{i=1}^{r} n_{+i}s_i)$ may be of primary interest. The corresponding Wald/Neyman χ^2_N for H_0 : $\mu_r - \mu_c = 0$ has null large-sample χ^2_1 distribution and is a contingency table counterpart of the **paired *t*-statistic**. Routine generation of covariance kernels for such tests is straightforward (see Weighted Least-Squares Analysis below). The tests may be inverted to obtain, in the former case, a confidence region for the vector of the $m_{i+} - m_{+i}$, and in the latter, a confidence interval for the scalar $\mu_r - \mu_c$.

When levels have no associated natural scores, it is often reasonable to apply the above methods with equally-spaced scores, especially for **questionnaire** data, with responses such as much improved, improved, the same, worse, and much worse. When linear scaling is unreasonable, data-derived scores may be used with appropriate conditioning. For this purpose, the $n_{ij}, i \neq j$, may be extracted and reformulated as an $(r - 1) \times 2$ table in which, for $k = 1, \ldots, r - 1$, $n^{\dagger}_{k1} = \sum \sum_{i-j=k} n_{ij}$ and $n^{\dagger}_{k2} = \sum \sum_{j-i=k} n_{ij}$. Row k of this table contains all elements of the parent table whose row and column categorizations differ by k levels. Rank scores may then be derived from the row sums $n^{\dagger}_{k+} = \sum_{l=1}^{2} n^{\dagger}_{kl} = \sum \sum_{|i-j|=k} n_{ij}$ of the reformulated table. Conditioning on the $n^{\dagger}_{k+}$, the $(r - 1) \times 2$ table may usually be treated as product-binomial with fixed row scores (see $r > c = 2$ below), yielding for example, categorical counterparts of the **signed-rank test**.

An alternative likelihood ratio approach for nominal data is based on the relationship of marginal homogeneity to two forms of interior symmetry. The model $m_{ij} = m_{ji}$ represents symmetric reflection off the main diagonal, or "total symmetry". The weaker **quasi-symmetry** model specifies that $\gamma_{ij} = \gamma_{ji}$ in (10). Total symmetry is the special case of quasi-symmetry with homogeneous margins. If quasi-symmetry can be justified, either *a priori* or by a nonsignificant **goodness-of-fit** test with adequate power, then the likelihood ratio statistic comparing total symmetry with quasi-symmetry is a valid test of marginal homogeneity. The df of the limiting χ^2_{r-1} distribution is the difference between the $(r^2 - r)/2 = r(r - 1)/2$ linearly independent constraints of total symmetry on the m_{ij} and the $[(r - 1)^2 - (r - 1)]/2$ constraints of quasi-symmetry on the $(r - 1)^2$ linearly independent γ_{ij} (*see* **Matched Pairs With Categorical Data**; **Square Contingency Table**).

$r > c = 2$

G^2, X_P^2, X_N^2, and exact hypergeometric methods extend directly from the 2×2 case, but neglect any ordering of A. Consequently, power functions of the resulting hypothesis tests fail to differentiate important and contextually plausible distributional

shifts or monotonic trend alternatives from irregular alternatives of equal magnitude but negligible plausibility. The appropriate method for incorporating ordinality depends on which margins are fixed by study design or conditioned upon in analysis. When scores s_i are available for the rows with the n_{i+} random and the n_{+j} fixed, it is natural to compare the column mean scores $\overline{s}_1 = n_{+1}^{-1} \sum_{i=1}^{r} m_{i1} s_i$ and $\overline{s}_2 = n_{+2}^{-1} \sum_{i=1}^{r} m_{i2} s_i$. This is readily done using $Q_W = X_N^2$ for the single linear contrast $\overline{s}_1 - \overline{s}_2 = 0$, yielding contingency table counterparts of Student's t-test and its associated confidence interval. When the n_{i+} and n_{+j} are fixed, an equivalent test may be obtained by substituting the null hypergeometric variance of this contrast for the estimated variance in Q_W.

When the n_{+j} are random and the n_{i+} are fixed, models relating the conditional probability $\pi_i = \pi_{i1} / \sum_{j=1}^{2} \pi_{ij}$ to s_i are natural analogs of univariate continuous regression analyses. The Cochran–Armitage trend test (*see* **Trend Test for Counts and Proportions**) is based on an unweighted least-squares fit of the simple linear regression of the p_i on the s_i. This amounts to a partitioning of the Pearson X_P^2, with df = r-1, into a multiple of the squared fitted regression coefficient and a weighted combination of squared deviations of observed from model-predicted proportions, respectively analogous to the regression and residual sums of squares for a continuous response. Under $H_0 : \pi_i = \pi$, both components have large-sample χ^2 distributions, with respectively 1 and $r - 2$ df. The first component is also valid as another t-test counterpart for detecting a difference in mean row scores between columns. Under the same model, the corresponding Wald/Neyman χ_N^2 uses a WLS-based slope estimate, while G^2 iterates the same WLS computation until convergence. These latter approaches are readily extended to polynomial regression models for the π_i, and can provide confidence intervals for estimated regression parameters. The residual X_N^2 and G^2 statistics retain their null χ_{r-2}^2 distribution under simple linear regression alternatives to H_0, and hence may be interpreted as lack-of-fit statistics.

The related partition of chi-square technique (*see* **Chi-square, Partition of**) is useful for isolating heterogeneity in tables of nominal variables where association between rows and columns has been established. If a nested sequence of hypotheses can be developed whose intersection implies independence, or homogeneity of rows or columns, then test statistics for corresponding collapsed versions of the original table are approximately additive, and have asymptotically independent null χ^2 distributions. This is best shown by example. Table 4 relates infant malformations to mother's alcohol consumption. For these data, $X_P^2 = 6.9494$ with df $= 2$, $P = 0.031$, indicating statistically significant variation in the incidence of malformation with alcohol intake during the first trimester of pregnancy. However, the effect is exclusively associated with consumption of at least three drinks daily, as shown by the components of X_P^2 comparing the first two categories of alcohol consumption ($X_P^2 = 0.0984$ with df $= 1$, $P = 0.754$), and comparing the last category of alcohol consumption with these two categories pooled ($X_P^2 = 6.8557$ with df $= 1$, $P = 0.009$). These components are nested because all categories contrasted by the first test are pooled together in the contrast examined by the second test. Although the test statistics for the component single degree of freedom (df) tests do not precisely sum to that for the overall two df test ($6.8557 + 0.0984 = 6.9541 \neq 6.9494$), this does not compromise the validity of the tests individually; exact partitioning may be obtained by analogous use of G^2 or X_N^2. Partitioning requires caution, however, under product-binomial sampling with disproportionately sampled population subgroups. Pooling of such sample subgroups may produce biased estimates of proportions in the pooled population categories, and invalid tests based upon them.

Table 4 Infant malformations, by mother's average number of alcoholic drinks/day in first trimester of pregnancy; adapted from [1]

	Infant malformations		
Drinks/Day	Absent	Present	Total deliveries
0	17 066	48	17 114
1–2	14 464	38	14 502
3+	952	7	959

When the n_{+j} are fixed, the n_{i+} are random, and natural scores are not at hand, equally-spaced (linear) scores may be used *a priori* where that seems reasonable, or rank scores may be derived from the marginal distribution $\{n_{i+}\}$. Binary (zero or one) scores differentiating groups of low and high categories, Wilcoxon **rank** scores, and scores that transform Wilcoxon scores to expected order statistics (*see* **Normal Scores**) or to the inverse of a continuous cdf, are commonly used. For fixed s_i, likelihood

ratio, score or Wald statistics may be obtained for the structural model $H_0 : \bar{s}_1 - \bar{s}_2 = \sum_{i=1}^{r} s_i(m_{i1}/n_{+1} - m_{i2}/n_{+2}) = 0$. When the s_i are rank scores, either the multivariate hypergeometric distribution should be employed or the scores should be treated as random and their variation incorporated into the test statistic. With the former approach, a quadratic form with covariance matrix from the multivariate (central) hypergeometric distribution may be used as test statistic. Under the latter approach, the Wald test, obtained by approximating the nonlinear $\bar{s}_1 - \bar{s}_2$ by a linear function of the n_{ij}, may be used. These statistics have limiting χ_1^2 distributions, and yield categorical data counterparts of the Wilcoxon rank sum and other rank tests for continuous data (*see* **Wilcoxon–Mann–Whitney Test**).

The tests above differ in their handling of nuisance parameters. Some, for example, the two-population likelihood ratio comparison of means of predetermined scores, employ test statistics whose asymptotic distributions are fully specified under the nominal null hypothesis. This is accomplished by collapsing over nuisance parameters, conditioning them out, or optimizing over their possible values. For others, for example, the Cochran–Armitage trend test and the hypergeometric-based comparison of means of data-derived scores, nuisance parameters still influence the distribution of the test statistic when the nominal null hypothesis is true, and are only removed by testing a more restrictive hypothesis. Such procedures are best understood as tests of the more restricted rather than the nominal null hypotheses, but with power concentrated against alternatives to the latter. They may not be unbiased tests of the nominal null, but this deficiency has minimal practical consequences.

Ordinality may also be handled by incorporating category distances, as determined by scores, directly into the structure of the m_{ij}. This allows **parsimonious** summarization and targeted testing of ordinal association. Roughly speaking, a set of statistics $\mathcal{S}$ is **sufficient** for a set of parameters if the likelihood equations for estimating the parameters involve the data only through the members of $\mathcal{S}$. Loglinear models may be constructed with parameters for which simple observed functions of scores, such as means and covariances, are sufficient statistics. In the $r \times 2$ table, writing $\gamma_{\text{loc}}^i = \ln m_{i1} - \ln m_{i2} - \ln m_{i+1,1} + \ln m_{i+1,2} = \ln \psi_i$, where γ_{loc}^i is also γ_{ij} as in (10), the model $\gamma_{\text{loc}}^i = (s_i - s_{i+1})\beta$ is a simple linear regression model for the $\text{logit}(\pi_i) = \ln[\pi_i/(1-\pi_i)]$ as a function of the s_i, with β the regression parameter. The mean score $n_{+1}^{-1} \sum_{i=1}^{r} n_{i1} s_i$ is sufficient for β in this model. When the model fits, the $r-1$ parameters γ_{loc}^i, $i = 1, \ldots, r-1$, reduce to a single β that, in conjunction with parameters for the marginal distributions, completely specifies the m_{ij}. An estimate $\widehat{\beta}$ can then be used to summarize the strength and direction of the change in conditional probability between rows, or the change in mean score between columns, in an observed table.

In biostatistics, this "logit model" was developed for the analysis of quantal **bioassay** data, where a test animal was presumed to respond or not respond to a drug or toxic agent depending on whether or not the animal's individual tolerance threshold was exceeded by the dose. The **logistic distribution** was used to model the distribution of tolerance thresholds across animals because of its similarity to the Gaussian distribution, coupled with its exceptional mathematical convenience. Animals were generally dosed in groups, with $s_i = \ln(\text{dose}_i)$. This analytic advance was seminal for loglinear, **logistic regression**, weighted least-squares, and **generalized linear modeling** (*see* **Quantal Response Models**; **Psychometrics, Overview**).

The odds ratio structure of the logit model may be applied outside the framework (10) of a loglinear model for cell probabilities. Suppose row margins are random. Rather than modeling the logarithms of the "local" odds ratios γ_{loc}^i, that is, the ln odds ratios of expected cell counts from adjacent rows in the $r \times 2$ table, one may model the logarithms of the corresponding cumulative odds ratios of the (r-1) 2×2 tables obtained by collapsing rows 1 through i and rows $(i+1)$ through r, for $i = 1, \ldots, r-1$. The ith of these overlapping, cumulative tables has expected entries $m_{1j}^i = \sum_{l=1}^{i} m_{lj}$, $m_{2j}^i = n_{+j} - m_{1j}^i$. Then the model $\gamma_{\text{cum}}^i = \ln(m_{11}^i m_{22}^i / m_{12}^i m_{21}^i) = (s_i - s_{i+1})\beta$ is a "cumulative" (or "global") odds ratio model. Multifactor linear models for logarithms of global rather than local odds are often called **proportional odds models**. Similarly, the ith of an alternative set of $(r-1)$ 2×2 derived tables may be constructed using the m_{ij} as first and the m_{2j}^i as second rows. The $(r-1)$ "continuation-ratio" odds ratios γ_{cr}^i from these tables are used primarily in analyses of grouped survival data, where $m_{ij} / \sum_{l=i+1}^{r} m_{lj}$ is the conditional odds and $m_{ij} / \sum_{l=i}^{r} m_{lj}$ is the conditional probability of an event such as death in the ith observation period, given survival through the preceding

periods without experiencing the event. Ratios of these probabilities and odds represent discrete data counterparts of **hazard ratios** for continuous survival data, and become essentially equivalent when rows correspond to sufficiently short time intervals. Thus, linear models for the corresponding $\ln \gamma_{\text{cr}}^i$ and $\ln RR_{\text{cr}}^i$ are discrete counterparts of **proportional hazards** models for continuous survival data (*see* **Survival Analysis, Overview**).

Many of the above ordinal models may be further extended by treating the scores as unknown parameters to be estimated rather than as fixed. (*See* **Ordered Categorical Data**).

$c > 2$

General $r \times c$ tables allow the possibilities of **polytomous** nominal and/or ordinal variables on either or both dimensions. The methods above extend readily. For instance, Student's t-statistic counterparts with large-sample χ_1^2 distributions extend directly to counterparts of analysis of variance $\boldsymbol{F}$ **tests** with asymptotic $\chi_{(c-1)}^2$ distributions. Conditional or unconditional exact tests, based either on ordering tables by their multivariate hypergeometric probabilities under independence (thereby generalizing Fisher's exact test), or on orderings according to pivot functions such as X_P^2, are available for tables with cell expectations too low to support inference based on large-sample chi-square approximations.

For tables with nominal columns and ordinal rows with associated fixed scores $\{s_i\}$, we may directly extend the linear logit model for the $r \times 2$ table by applying such models simultaneously within each pair of columns. The model for the full table parameterizes the local odds ratios $\psi_{ij} = \pi_{ij}\pi_{i+1,j+1}/\pi_{i,j+1}\pi_{i+1,j}$ as $\ln \psi_{ij} = (s_i - s_{i+1})(\xi_j - \xi_{j+1})$, where the $\{\xi_j\}$ are unknown column parameters. Within each pair of columns, odds ratios between rows are a multiple of the differences between the fixed row scores. The multiple, however, varies across column pairs according to differences among the ξ_j. Hence, $\beta_{j'j''} = \xi_{j'} - \xi_{j''}$ plays the role of β in the linear logit model of the $r \times 2$ table, for columns j' and j'', by representing the extent to which the row distribution in the column j' is stochastically higher or lower than the row distribution in the column j''.

When both rows and columns are ordinal with respective fixed scores $\{s_i^{(r)}\}$ and $\{s_j^{(c)}\}$, association may be examined through the **correlation** of row and column scores. If the true covariance or correlation is the primary function of the m_{ij} of interest, either a loglinear model may be constructed with an overall association parameter for which the sample covariance is a sufficient statistic, or inference may focus directly on the sample covariance and its distribution.

The model in the first case parameterizes the local odds ratios as $\ln \psi_i = (s_i^{(r)} - s_{(i+1)}^{(r)})(s_j^{(c)} - s_{(j+1)}^{(c)})\beta$. The sample covariance is sufficient for β. Within a given pair of rows, the log odds ratios between pairs of columns vary linearly with the difference in column scores, while within a given pair of columns, the log odds ratios between pairs of rows vary linearly with the difference in row scores. Hence, this "linear by linear association model" simultaneously applies the linear logit model for an $r \times 2$ table within all such tables formed by pairs of columns or pairs of transposed rows, and all these models share the same slope β. When both the $\{s_i^{(r)}\}$ and $\{s_j^{(c)}\}$ are equally spaced, the local log odds ratios for all 2×2 subtables formed from adjacent rows and columns throughout the $r \times c$ table are the same multiple of β, hence identical. In this "uniform association model", the single odds ratio e^β replaces the $(r-1)(c-1)$ parameters necessary to describe the association structure of an arbitrary $r \times c$ table. This representation approaches the simplicity and elegance of the correlation coefficient in the **bivariate normal** Gaussian distribution for continuous data. *A priori* justification for a uniform association model is rare, however, and this model is more useful for approximating a monotonic trend, or as a smoothing device for generating tests with power directed at linear correlation alternatives, than as a full representation of many data sets.

An alternative approach is direct inference from $\widehat{\text{cov}}(s^{(r)}, s^{(c)})$. The random portion of $\widehat{\text{cov}}(s^{(r)}, s^{(c)})$ is nonlinear in the n_{ij} under product-Poisson, multinomial, or product-multinomial sampling but linear, through $\sum\sum s_i^{(r)} s_j^{(c)} n_{ij}$, when conditioning on both row and column margins. Under such dual conditioning, the quadratic form statistic from $\widehat{\text{cov}}(s^{(r)}, s^{(c)})$ and its variance under the multivariate (central) hypergeometric distribution may be used as a test of independence with power function directed at linear correlation. Use of this test, which generalizes X_{MH}^2 in the 2×2 table, is analogous to testing $\rho = 0$ to detect association between continuous variables, when monotonic but not necessarily linear association is anticipated. Null covariance does not imply

independence for any of the relevant distributions, so optimizing the likelihood under $H_0 : \mathrm{cov}(s^{(r)}, s^{(c)}) = 0$ is difficult. Thus, likelihood ratio and score statistics are often obtained after imposing additional structure on the m_{ij}. However, the Wald statistic is straightforward to determine and, as it uses a consistent estimate of the variance of $\widehat{\mathrm{cov}}(s^{(r)}, s^{(c)})$ unrestricted by the null hypothesis, its null χ^2_1 distribution is valid without additional assumptions (*see* **Ordered Categorical Data**).

Three-dimensional Tables

Extensions of many methods for two-way tables are straightforward. Indeed, three-dimensional tables can be reduced to two-dimensional tables whenever two variables, say B and C, are conditionally independent within each level of the third variable A. In that case, the two-way marginal $A \times B$ and $A \times C$ tables are sufficient statistics for the cell parameters of the three-way table, and hence, contain all the information in the data about the joint distribution of A, B, and C. Moreover, extending (10), conditional association structures can be represented by loglinear models whose parameters correspond to the usual deviation contrasts of balanced three-way factorial ANOVA, applied to the $\ln m_{ij}$. Each **hierarchical model** (for which every interaction is accompanied by all lower order interactions and main effects of its members) represents a different association structure. All variables are independent under the main effects model. When a single pairwise interaction is present, the interacting variables are jointly independent of the third. When two pairwise interactions are present, the unpaired variables are conditionally independent of each other, given the third. When all pairwise interactions are present without three-way interaction, each conditional odds ratio relating two variables is constant over levels of the third variable.

Confounding and Effect Modification

However, the loglinear models corresponding to saturated and "no three-way interaction" ANOVA models introduce fundamental issues that occur only in contingency tables of three or more dimensions. These are most easily explained through a 2^3 table of expected counts. Thus, the layers of Table 5 represent two strata of a population, such as men and women (variable A), in which an exposure such as an environmental risk factor or a new clinical treatment regimen (variable B) has been related by observational study to a health outcome such as new disease or death (variable C). The subscript hij refers to the cell in stratum (layer) h, row i, and column j.

Table 5 A 2^3 Contingency table relating exposure to health outcome in two strata

Variable A: Stratum, Stratum 1

	Variable C: Outcome		
Variable B: Exposure	Present	Absent	Total
Exposed	n_{111}	n_{112}	n_{11+}
Not exposed	n_{121}	n_{122}	n_{12+}
Total	n_{1+1}	n_{1+2}	n_{1++}

Variable A: Stratum, Stratum 2

	Variable C: Outcome		
Variable B: Exposure	Present	Absent	Total
Exposed	n_{211}	n_{212}	n_{21+}
Not exposed	n_{221}	n_{222}	n_{22+}
Total	n_{2+1}	n_{2+2}	n_{2++}

The saturated model allows the possibility, as may occur in any observed table, that the odds ratios $\psi_{(h)} = m_{h11}m_{h22}/m_{h12}m_{h21}$ relating exposure to outcome in each stratum may vary. For example, the effect of a treatment on recovery from disease might be different for men than for women. In the presence of such three-way interaction, which epidemiologists term **effect modification**, stratum-specific odds ratios are required to properly summarize the exposure-disease relationship. Neither the marginal odds ratio nor any average of the $\psi_{(h)}$ adequately represents the data when the $\psi_{(h)}$ differ to a scientifically meaningful degree. The stratification variable is called an "effect modifier".

In the "no three-way interaction" model, which includes all two-way interactions, the stratum-specific odds ratios relating exposure to outcome are identical. Nevertheless, as in the saturated model, the relationship may not be reliably represented by the collapsed table formed from pooling the strata. This is garden-variety statistical **confounding** in a categorical data context, where the stratification variable is the "confounder". Owing to the separate relationships of B and C to the stratification variable A, the set of marginal odds ratios $\{\psi^M_{BC}\}$ relating pairs of categories of B to pairs of categories of C in the

collapsed table ignoring A may differ substantially from the set of common within-stratum odds ratios $\{\psi_{BC|A}\}$ relating the same pairs of categories in the three-way table. An extreme case, **Simpson's Paradox**, occurs when, for instance, in a general $h \times 2 \times 2$ table, the sign of $\ln \psi_{BC}^{M}$ differs from the common sign of the stratum specific $\ln \psi_{BC|A=h}$, $h = 1, \ldots, q$. Marginal and conditional odds ratios thereby indicate relationships in opposing directions. In an observational study, one might thus infer benefit to a harmful medical therapy by neglecting to control for such confounding.

When the no three-way interaction loglinear model holds, that is, when confounding occurs without effect modification, $\psi_{BC|A}$ properly adjusts for confounding by A, and its associated MLE may be used to summarize the exposure-disease relationship. When ordinal models apply within strata, a similar interpretation applies to within-stratum association parameters of the ordinal model.

Cochran-Mantel-Haenszel Tests

Cochran–Mantel–Haenszel (CMH) testing is an approach to demonstrate association between two variables after controlling for one or more confounders by stratification. The approach is useful in situations without effect modification such as just described, as well as in others where effect modification is present but within-stratum log odds ratios are predominantly of the same sign. From one perspective, CMH tests are simply score tests of the $\{\psi_{BC|A}\}$, or other indices of association, from conditional likelihood analyses of various loglinear models without three-way interaction. However, they are also simply derived as structurally model-free, multivariate hypergeometric-based tests for association in the presence of confounding. This latter perspective is more helpful in clarifying their broad utility.

We string out a $q \times r \times c$ table as the single vector $\boldsymbol{n} = (\boldsymbol{n}_1', \ldots, \boldsymbol{n}_q')'$ with $\boldsymbol{n}_h = (n_{h11}, \ldots, n_{h1c}, \ldots, n_{hr1}, \ldots, n_{hrc})'$. CMH test statistics take the quadratic form

$$Q_{\mathrm{CMH}} = \left(\sum_{h=1}^{q} \boldsymbol{A}_h(\boldsymbol{n}_h - \overset{\circ}{\boldsymbol{m}}_h)\right)' \left(\sum_{h=1}^{q} \boldsymbol{A}_h \mathrm{cov}(\boldsymbol{n}_h)\boldsymbol{A}_h'\right)^{-1} \times \left(\sum_{h=1}^{q} \boldsymbol{A}_h(\boldsymbol{n}_h - \overset{\circ}{\boldsymbol{m}}_h)\right), \quad (13)$$

where $\overset{\circ}{\boldsymbol{m}}_h = \mathrm{E}(\boldsymbol{n}_h)$ and $\mathrm{cov}(\boldsymbol{n}_h)$ are the mean and covariance matrix of $\boldsymbol{n}_h$ under a null multivariate hypergeometric distribution within-stratum h, and the $\boldsymbol{A}_h$ are typically of full-rank u. Under either (i) independent stratified randomization to the row or column dimension, or (ii) stratified random sampling in the absence of within-stratum row by column association, the distribution of $\boldsymbol{n}$ is product-multivariate hypergeometric and the kernel of Q_{CMH} is the exact covariance matrix of $\sum_{h=1}^{q} \boldsymbol{A}_h(\boldsymbol{n}_h - \overset{\circ}{\boldsymbol{m}}_h)$. Q_{CMH} is then asymptotically χ_u^2, by randomization central limit theory.

CMH tests address conformity with null expectations of across-strata linear combinations of within-stratum pivot functions. The noncentrality parameter of the asymptotic chi-square distribution, under an alternative with $E\boldsymbol{n}_h = \boldsymbol{m}_h$, is an across-strata linear combination of the $\boldsymbol{m}_h - \overset{\circ}{\boldsymbol{m}}_h$, proportional to a weighted average of functions of within-stratum association measures. Thus, CMH tests have substantial power only against alternatives in which the within-stratum associations are predominantly in the same direction, and for this reason are often called "average partial association tests". In addition to a general test against all forms of average partial association, the CMH approach yields more focused extensions of χ_{CMH}^2 to (i) a stratum-adjusted ANOVA F-statistic analog for testing differences among mean column scores between nominal rows; (ii) analogs of partial correlation tests based on *a priori* scores; and (iii) a variety of rank tests using scores based on overall or stratum-specific row and column margins. Since the null distribution is conditional on both margins of each stratum, such derived rank scores may be treated as fixed.

The null large-sample chi-square distribution of a CMH statistic is valid for increasingly large samples within a fixed set of strata, or under a "sparse asymptotic" situation in which the number of strata increases, while the sample sizes within them remain small. A prominent example of the latter is when each stratum consists of a matched case-control pair. This situation prompted development of the original Mantel–Haenszel statistic for q 2×2 tables, each with a single observation per row. The reformulation of a 2×2 table, in which a matched case–control pair is the observational unit and each pair is classified as [+, +], [+, −], [−, +], or [−, −] with respect to an exposure, into a $q \times 2 \times 2$ table in which the individual is the observational unit and classification is by a (i) pair identifier, (ii) exposure,

and (iii) disease, is a productive representation for inference about the matched-data analogues of Δ, RR, and ψ from independent samples. For instance, the Mantel–Haenszel summary odds ratio $\widehat{\psi}_{\text{MH}} = (\sum_{h=1}^{q} n_{h11}n_{h22}/n_{h++})/(\sum_{h=1}^{q} n_{h12}n_{h21}/n_{h++})$, arises naturally from treating matched sets as strata. For matched pairs, using the notation $n^{[+,+]}, n^{[+-]}, n^{[-,+]}, n^{[-,-]}$ for the respective counts in the 2×2 table classifying the matched pairs, $\widehat{\psi}_{\text{MH}} = (\sum_{h=1}^{q} n_{h11}n_{h22}/2)/(\sum_{h=1}^{q} n_{h12}n_{h21}/2) = n^{[+,-]}/n^{[-,+]}$ and, from (13), $Q_{\text{CMH}} = (n^{[+,-]} - n^{[-,+]})^2/(n^{[+,-]} + n^{[-,+]})$, the **McNemar test**. Stratified versions of relative risk and other epidemiologic measures of association are also used in conjunction with CMH statistics (*see* **Mantel–Haenszel Methods**).

Generalized Linear Models

We briefly review unifying paradigms for generalization and extension of methods above to data (i) of arbitrary dimension, and/or (ii) with continuous covariates, or other features incompatible with the assumption $m_{ij} \uparrow \infty$.

Generalized linear models (GLMs) provide a single framework for modeling expectations of Gaussian and other continuous data, Bernoulli indicator variables, Poisson, binomial, and (through a conditional Poisson argument) multinomial counts. This framework exploits the common structure of univariate **exponential families** of distributions, whose pdfs or pmfs may be written as Ke^L, where K and L each may be written as the product of a term depending exclusively on data and another depending exclusively on unknown parameters.

GLMs postulate independent observations $Y_1, \ldots, Y_n$ from a univariate exponential family

$$f_Y(y_i; \theta_i, \phi) = \exp\left\{\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)}\right] + c(y_i, \phi)\right\}, \tag{14}$$

for known functions $a_i(\cdot), b(\cdot)$, and $c(\cdot, \cdot)$ and fixed scale parameter ϕ. Let $b'(\cdot)$ and $b''(\cdot)$ be the first and second derivatives of $b(\cdot)$. In this case, $\text{E}(Y_i) = \mu_i = \mu_i(\theta_i) = b'(\theta_i)$ and $\text{var}(Y_i) = v_i = v_i(\theta_i, \phi) = b''(\theta_i)a_i(\phi)$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ be the vector of expectations. A GLM $\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta}$ restricts $\boldsymbol{g}(\boldsymbol{\mu}) = (g(\mu_1), g(\mu_2), \ldots, g(\mu_n))'$ to a $u < n$-dimensional linear subspace of $\Re^n$ spanned by columns of an $n \times u$ full-rank model matrix $\boldsymbol{X} = [x_{ik}]$. In the Gaussian case, $\boldsymbol{X}$ is the "design" or "model" matrix in a general linear model, and otherwise is totally analogous. Notationally suppressing the dependence of $\mu_i = \mu_i(\theta_i)$ on θ_i and of $v_i = v_i(\theta_i, \phi)$ on θ_i and ϕ, the likelihood equations for any GLM are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ik}}{v_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0, \quad k = 1, \ldots, u. \tag{15}$$

Note that the left-hand side of (15) is an inverse-variance weighted linear combination of the residuals $y_i - \mu_i$, and that these are the standard normal equations for a general linear model with rescaling for the transformation $g(\cdot)$ and inverse-variance weighting, as with weighted least-squares, to account for nonconstant variance. For fixed ϕ, (15) may be solved for $\boldsymbol{\beta}$, by iterated (equivalently, generalized) weighted least-squares (IWLS). When ϕ is known by the form of the distribution, this completes model estimation. When ϕ is an unknown scale parameter, as with Gaussian distributions for which $\phi = \sigma^2$, then ϕ may be estimated from the residuals to obtain a complete solution. The MLE of the covariance matrix $\text{cov}(\widehat{\boldsymbol{\beta}})$, $\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}})$, is a by-product of IWLS estimation. Likelihood ratio, Pearson chi-square, score and Wald statistics are often used to evaluate fit and assess significance of model parameters.

The "canonical link function" $g(\cdot) = b'^{-1}(\cdot)$ produces a linear model $\boldsymbol{\theta} = \boldsymbol{X\beta}$ for the "natural parameter" $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ of the exponential family (14). The u linear combinations $\sum_{i=1}^{n}(x_{ik}/a_i(\phi))y_i$ are then jointly sufficient for $\boldsymbol{\beta}$, and the likelihood equations set these sufficient statistics to their expectations. In the absence of differential weighting through the a_i, the sufficient statistics are simply inner products of the response vector with columns of $\boldsymbol{X}$. With convenient parameterizations, these may be means or mean scores, cross-product sums, marginal totals, or analogous functions of the observed data, which the MLEs are then constrained to reproduce.

This formulation is even more powerful than suggested by its exponential family underpinning. The first-order asymptotics of MLEs and likelihood ratios that justify most statistical applications depend upon the likelihood only through the moment-based

properties

$$\mathrm{E}\left(\frac{y_i-\mu_i}{v_i}\right)=0,$$
$$\mathrm{var}\left(\frac{y_i-\mu_i}{v_i}\right)=-\mathrm{E}\left[\left(\frac{\partial}{\partial\mu}\right)\left(\frac{y_i-\mu_i}{v_i}\right)\right]$$
$$=v_i^{-1}. \qquad (16)$$

Consequently, provided the first and second moments $\mathrm{E}(Y_i)=\mu_i(\theta_i)$ and $\mathrm{var}(Y_i)=v_i(\theta_i,\phi)$ are specified correctly, solutions to (15) largely behave like MLEs even if the data do not otherwise follow the distribution from which these moments were obtained. Thus, first-order optimal estimators may also be obtained from these estimating equations (15), when the moment functions are known and sufficiently regular, even in the absence of knowledge about the generating distribution. In such cases, the **quasi-likelihood** function

$$QL(\mu|y)=\int_y^\mu \frac{y-t}{v(\theta\phi)}\,\mathrm{d}t \qquad (17)$$

behaves like a likelihood and "quasi-likelihood analysis" therefore proceeds as if (17) is the true likelihood. Among their other uses, such quasi-likelihood analyses provide a convenient method for adjusting for observed overdispersion relative to what would ordinarily be expected from, for instance, binomial or Poisson variation. Quasi-likelihood thus provides a conceptual bridge between likelihood and weighted least-squares estimation (*see* Weighted Least-Squares (WLS) Functional Models below).

When f is Gaussian with $a_i(\phi)=a_i\phi$ for known a_i, the canonical link is the identity function, maximum likelihood reduces to weighted least-squares, and the IWLS algorithm terminates in one step. Categorical data GLMs include **Poisson regression**, multiple logistic regression, and general loglinear models.

Poisson Regression

For Poisson counts, the canonical link function is $\ln(\cdot)$. If exposure is uniform, the ith natural parameter is $\ln\mu_i=\ln\lambda_i$, and the corresponding GLM is a loglinear model for the Poisson intensity parameters. In epidemiologic studies, the λ_i usually are incidence densities, and an individual β_k often represents the $\ln RR$ associated either with comparison of the kth level of a categorical independent variable to a baseline level, or with a one unit difference of a continuous independent variable, other variables in the model held constant. The WLS iteration step is a regression on $\boldsymbol{X}$ of the residuals from the previous step, weighted inversely to the previous fitted values, which by the Poisson law are the current variance estimates.

In the more usual case of varying exposure measures N_i, the ith natural parameter is $\ln\mu_i=\ln N_i\lambda_i$, where the N_i are known. The linear predictor $\boldsymbol{X\beta}$ for $\ln\boldsymbol{\mu}$ is offset from the predictor for $\ln\boldsymbol{\lambda}$ by $\ln\boldsymbol{N}$, where $\boldsymbol{N}$ is the vector of known exposures. This is easily accommodated by estimating $\boldsymbol{\mu}$ and then rescaling to $\widehat{\boldsymbol{\lambda}}=\boldsymbol{D}_N^{-1}\widehat{\boldsymbol{\mu}}$, $\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\lambda}})=\boldsymbol{D}_N^{-1}(\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\mu}}))\boldsymbol{D}_N^{-1}$, where $\boldsymbol{D}_z$ is the diagonal matrix with main diagonal $\boldsymbol{z}$.

Multiple Logistic Regression

For binomial counts, the canonical link is the logit function, and the GLM generalizes the logit regression model previously described for $r\times 2$ contingency tables. As $\boldsymbol{X}$ may be an arbitrary full-rank design matrix, the model may include a combination of categorical and continuous predictors. Letting n_{i+} be the sample size and π_i be the "success" probability associated with the ith $(i=1,\ldots,r)$ pattern of predictor variables $\boldsymbol{x}_i$, the likelihood equations are simply $\sum_{i=1}^r n_{i1}\boldsymbol{x}_i=\sum_{i=1}^r n_{i+}\widehat{\pi}_i\boldsymbol{x}_i$. The sufficient statistics are means of the predictor variables in the respective outcome groups, and the MLE $\widehat{\boldsymbol{\beta}}$ generates predicted probabilities $\widehat{\pi}_i$ that reproduce these. The iteration step is $\widehat{\boldsymbol{\beta}}^{(t+1)}=\widehat{\boldsymbol{\beta}}^{(t)}+[\sum_{i=1}^r n_{i+}\widehat{\pi}_i^{(t)}(1-\widehat{\pi}_i^{(t)})\boldsymbol{x}_i\boldsymbol{x}_i']^{-1}[\sum_{i=1}^r(n_{i1}-n_{i+}\widehat{\pi}_i^{(t)})\boldsymbol{x}_i]$. The asymptotic covariance matrix for $\widehat{\boldsymbol{\beta}}$, $\mathrm{cov}(\widehat{\boldsymbol{\beta}})=[\sum_{i=1}^r n_{i+}\pi_i(1-\pi_i)\boldsymbol{x}_i\boldsymbol{x}_i']^{-1}$, may then be estimated by substitution.

The β_k are interpreted here similarly as in Poisson regression, but with odds ratios replacing rate ratios. Specifically, an individual β_k often represents the $\ln OR$ associated either with comparison of the kth level of a categorical independent variable to a baseline level, or with a one unit difference of a continuous independent variable, other variables in the model held constant. In an epidemiologic cumulative incidence cohort study or a clinical trial, π_i is usually the probability, in individuals with a known pattern of risk factors, of a disease outcome during a fixed observation period. In an epidemiologic incident case–control study, π_i is the probability of past

exposure to a risk factor in individuals with known current disease status and other sociodemographic factors and/or past exposures. Provided that cumulative incidence during the cohort study is low, in a stable population the odds ratios from these two types of studies should be similar and closely approximate the incidence density ratio, say λ_E/λ_U, between Exposed and Unexposed populations. For this reason, multiple logistic regression has become the most common tool for multivariable investigation in epidemiologic data analysis, particularly for control of confounding and effect modification. In randomized clinical medical trials, logistic regression is often used conventionally as both a primary analysis and, as a routine check for confounding, to simultaneously adjust for all other known predictors of a categorical outcome. This latter use is questionable since these adjusted models often contain dozens of variables, most of which are equitably randomized across treatment groups and unlikely to confound, and little attention is paid to the shapes of the corresponding simultaneous adjustments.

Related multinomial logit and loglinear models have also been developed, to apply the multiple logistic regression structure simultaneously either to dichotomies formed from category pairs from polytomous responses, or (through proportional odds) to cumulative probabilities obtained by dichotomizing ordered polytomies at multiple cut-points.

Loglinear Models

Loglinear models are a symmetric version of polytomous multiple logistic regression models, much as correlation analysis is a symmetric representation of multiple regression analyses for continuous data. When there is no clear "dependent variable" in a categorical data array, loglinear models simultaneously show how local odds relating pairs of levels of each single variable depend upon the levels or values of other variables. Loglinear models provide a basic paradigm for smoothing contingency tables, and for exploring causality through association structures in a spirit similar to that of **path analyses** of continuous data.

Let $\boldsymbol{\pi}' = (\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_r')'$, with $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ic_i})'$, be the single strung-out vector of probabilities for a product-multinomial distribution with r components, of which the ith has c_i categories. Also, let $\boldsymbol{K}$ and $\boldsymbol{J}$ be block diagonal with respective ith blocks $(\mathbf{1}_{c_i}\mathbf{1}_{c_i}')$ and $\mathbf{1}_{c_i}$. A loglinear model is defined by $\boldsymbol{\pi} = \boldsymbol{D}_\omega^{-1}\exp(\boldsymbol{X\beta})$, where $\boldsymbol{\omega} = \boldsymbol{K}\exp(\boldsymbol{X\beta})$ and $\boldsymbol{X}$ is a full-rank model matrix with columns linearly independent of the columns of $\boldsymbol{J}$. The latter requirement is an **identifiability** condition, since the model equation implies $\mathbf{1}_{c_i}'\boldsymbol{\pi}_i = 1$ for all i.

The multinomial populations represented by i, the response categories represented by j, or both may themselves be defined by combinations of levels of nominal and/or ordinal variables, the latter possibly with fixed scores. Such underlying structure can be represented in the model matrix $\boldsymbol{X}$. Though without columns for intercepts for the respective populations, $\boldsymbol{X}$ may otherwise be identical to model matrices for structures of cell means from continuous data. When all variables are nominal, common choices for $\boldsymbol{X}$ are matrices for a hierarchical factorial ANOVA model with interactions up to a given order.

Although the multinomial distribution does not directly fit the univariate exponential family formulation (14), the GLM machinery may nevertheless be used. For, letting $\boldsymbol{n}$ and $\boldsymbol{m}$ be respectively the strung-out vectors of observed counts n_{ij} and their expectations $m_{ij} = n_{i+}\pi_{ij}$ ordered as in $\boldsymbol{\pi}$, the likelihood equations $[\boldsymbol{J}, \boldsymbol{X}]'\boldsymbol{n} = [\boldsymbol{J}, \boldsymbol{X}]'\widehat{\boldsymbol{m}}$ for the loglinear model $\boldsymbol{X}$ are identical to those of the Poisson regression model $[\boldsymbol{J},\boldsymbol{X}]$. The MLEs reproduce the sufficient statistics $\boldsymbol{X}'\boldsymbol{n}$ while conforming to the population totals $\boldsymbol{J}'\boldsymbol{n}$. For instance, the sufficient statistics for the no three factor interaction hierarchical model of any nominal table are thus seen to be the observed two-way margins of the table, while the sufficient statistics for the linear by linear association model are the row and column marginal counts and the sum of products of row and column scores. This model thus reproduces the Pearson correlation coefficient between row and column scores.

Writing $\boldsymbol{X}_i$ for the submatrix of $\boldsymbol{X}$ corresponding to $\boldsymbol{\pi}_i$, the asymptotic covariance matrix of the MLE $\widehat{\boldsymbol{\beta}}$ is $\operatorname{cov}_A(\widehat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^r n_{i+}\boldsymbol{X}_i'(\boldsymbol{D}_{\boldsymbol{\pi}_i} - \boldsymbol{\pi}_i\boldsymbol{\pi}_i')\boldsymbol{X}_i\right]^{-1}$. Its maximum likelihood estimate, $\widehat{\operatorname{cov}}_A(\widehat{\boldsymbol{\beta}})$, is obtained by substitution of $\widehat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ and emerges as a by-product from the standard GLM Newton–Raphson (IWLS) fitting algorithm (*see* **Optimization and Nonlinear Equations**). For large hierarchical models, where repeated matrix inversion is problematic, the model may be fit by using iterative scaling to determine $\widehat{\boldsymbol{\pi}}$ directly (*see* **Iterative Proportional Fitting**). After substitution of $\widehat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$, the model equations

may then be solved for $\widehat{\boldsymbol{\beta}}$, and $\widehat{\mathrm{cov}}_A(\widehat{\boldsymbol{\beta}})$ obtained by direct substitution of $\widehat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ above. (*See* **Loglinear Model**).

Conditional and Exact Logistic Regression

GLM analyses are valid only under large-sample conditions in which MLEs are consistent (*see* **Large-sample Theory**). Classical likelihood theory demonstrates such consistency under asymptotics with indefinitely increasing information *per model parameter*. However, in certain situations involving matched data, the dimensionality of the parameter space increases with sample size, so that the information per parameter is bounded. These include cross-over experiments with parameters for each subject, pair-matched case–control studies with parameters for each matched pair, and clinical trials of rare diseases by large multicenter networks, with parameters for each center and few patients per center. In each of these cases, increasing sample size is accompanied by commensurate increase in parameters. Classical results on normality of the sufficient statistics for the model do not apply, and the MLE based on the full likelihood behaves poorly. Similarly, in small samples, the MLE may be substantially biased. The effects of nuisance parameters on estimation must somehow be controlled.

Conditional logistic regression applies standard likelihood theory after conditioning on the occurrence of precisely the patterns of predictor variables seen in the data from each matched set (*see* **Logistic Regression, Conditional**). The contribution to the conditional likelihood from each matched set is thus the ratio of the product-binomial probability of the observed data, expressed as a function of the linear predictor $\boldsymbol{x}_i'\boldsymbol{\beta}$, to the sum of such probabilities, over all permutations of the predictor variable patterns among individuals in that matched set. The full conditional likelihood is the product of these contributions across the matched sets. For 1:1 matched studies, the GLM computational algorithm may be formally applied by regressing a constant **dummy** outcome on the case–control differences of the predictors, highlighting the close relationship of this analysis, for a single dichotomous predictor, to the matched-pairs t-test. More generally, if each matched set in the logistic regression setting is identified with the **risk set** at an observed failure time in the survival data context, the conditional likelihood takes the same form as Cox's **partial likelihood** for **semiparametric** proportional hazards models for right-censored survival data (*see* **Cox Regression Model**). The linear predictor of the conditional likelihood lacks an intercept term, so that the model does not predict absolute risk. However, the likelihood is free of nuisance parameters, allowing consistent estimation of $\boldsymbol{\beta}$ as the number of matched sets increases, even if each matched set remains small. With the exception of examinations of **residuals**, most aspects of conditional logistic regression analysis are essentially identical to unconditional logistic regression via the GLM.

For quite small samples, however, the large-sample properties of even such conditional MLEs are of no help. Extensions are required of Fisher's exact test and other multivariate hypergeometric-based procedures for two-way tables. Estimators and hypothesis tests may be based on distributions that condition out not just nuisance parameters representing matched sets (as in conditional logistic regression), but *all parameters not being estimated or restricted by the null hypothesis.* These permutation distributions are obtained by conditioning on the observed values of sufficient statistics for unconstrained parameters. This computationally intensive method is now feasible and widely applied, although computations remain prohibitive for some problems. In a few cases at the other end of the spectrum, minimal computation is required because the conditional distribution employed is highly discrete, and can be so discrete as to provide little information about the parameters of interest. The technique is not a general cure for the "few subjects, many predictors" problem (*see* **Exact Inference for Categorical Data**; **Randomization Tests**).

Weighted Least-squares (WLS) Functional Models

Because they focus on modeling conditional expectations of individual observations on a single variable, as in logistic regression, or on cell counts determined by patterns of explanatory variables, as in loglinear models, GLMs provide an excellent framework for study of conditional probabilities and association structures. However, they are awkward for analyses of integrated data summaries such as marginal

distributions, survival rates, reliability measures, and association measures other than odds ratios, where the individual cell expectations are nuisance parameters for estimating the specific parametric functions of interest. WLS provides a convenient and powerful alternative approach to such analyses, through use of one-step WLS to obtain Neyman's minimum χ^2_N estimates and Wald Q_W statistics.

The WLS algorithm is based on the functional model $\boldsymbol{F}(\boldsymbol{m}) = \boldsymbol{X}\boldsymbol{\beta}$, with $\boldsymbol{F}(\boldsymbol{m}) = (F_1(m), \ldots, F_u(m))'$. The $F_k(\cdot)$ are second-order differentiable in the π_{ij} and $\boldsymbol{X}$ is a full-rank model matrix. The computation is performed on $\widehat{\boldsymbol{F}} = \boldsymbol{F}(\boldsymbol{n})$, an unrestricted consistent estimator of $\boldsymbol{F}(\boldsymbol{m})$, weighting using a consistent estimate of its asymptotic covariance matrix. The estimate is obtained as $\widehat{\boldsymbol{V}}_{\widehat{F}} = \widehat{\mathrm{cov}}_A(\boldsymbol{F}(\boldsymbol{n})) = \boldsymbol{H}(\boldsymbol{n})[\widehat{\mathrm{cov}}_A(\boldsymbol{n})]\boldsymbol{H}'(\boldsymbol{n})$, with $\boldsymbol{H}(\boldsymbol{n}) = [\partial F_k/\partial m_l]|_{\boldsymbol{m}=\boldsymbol{n}}$, by the delta method for "propagation of variances" introduced earlier in connection with confidence intervals for association measures in 2×2 tables. Here, $\widehat{\mathrm{cov}}_A(\boldsymbol{n})$ is the unrestricted MLE of the covariance matrix appropriate to the assumed distribution (e.g. $\boldsymbol{D}_{\boldsymbol{n}}$ for product-Poisson, and block diagonal with diagonal blocks $(\boldsymbol{D}_{\boldsymbol{n}_i} - (\boldsymbol{1}'\boldsymbol{n}_i)^{-1}\boldsymbol{n}_i\boldsymbol{n}_i')$ for product-multinomial data.

The estimator $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\widehat{\boldsymbol{V}}_{\widehat{F}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{V}}_{\widehat{F}}^{-1}\boldsymbol{n}$ minimizes the quadratic form $(\widehat{\boldsymbol{F}} - \boldsymbol{X}\boldsymbol{\beta})'\widehat{\boldsymbol{V}}_{\widehat{F}}^{-1}(\widehat{\boldsymbol{F}} - \boldsymbol{X}\boldsymbol{\beta})$, and is asymptotically multivariate Gaussian. Letting $\boldsymbol{V}_{\widehat{F}} = \boldsymbol{H}(\boldsymbol{m})[\mathrm{cov}_A(\boldsymbol{n})]\boldsymbol{H}'(\boldsymbol{m})$, the covariance matrix of the estimator is $\mathrm{cov}_A(\widetilde{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{V}_{\widehat{F}}^{-1}\boldsymbol{X})^{-1}$. Its estimate, $\widehat{\mathrm{cov}}_A(\widetilde{\boldsymbol{\beta}}) = (\boldsymbol{X}'\widehat{\boldsymbol{V}}_{\widehat{F}}^{-1}\boldsymbol{X})^{-1}$, may be used to form confidence regions and Wald test statistics for $\boldsymbol{\beta}$. When the F_k are explicit algebraic functions, $\boldsymbol{F}$ and $\boldsymbol{H}(\boldsymbol{n})$ can be expressed as compositions of matrix premultiplications and elementwise logarithmic and exponential transformations on n, simplifying the development of general software.

Validity of a WLS analysis depends on sample sizes large enough to ensure that $\boldsymbol{F}(\boldsymbol{n})$ is approximately multivariate Gaussian, so requirements vary with $\boldsymbol{F}$. WLS may be suitable for analysis of marginal mean scores of an ordinal variable in a panel survey from a high-dimensional and sparsely populated contingency table, but unsuitable for analysis of probabilities or log odds from a small table with few cells and larger sample size, if any cells are not observed or have very low expected values under a model. Slower convergence in distribution of Wald/Neyman chi-squares relative to score and likelihood ratio statistics has now been noted in a variety of categorical data situations through study of exact distributions and simulations. Consequently, greater caution is indicated than was once appreciated when using this method in moderate sample sizes, especially as the computational simplicity of closed-form parameter estimation becomes progressively less of an advantage over other methods.

However, for analysis of grouped data from sufficiently large samples, the flexibility of WLS allows simple handling of some problems that are quite awkward to address from other perspectives. WLS has been particularly important in the analysis of repeated measures contingency table data, in modeling of arbitrary measures of association from grouped data, and in modeling estimates from **sample surveys**, each of which may be a complex but explicit algebraic function of counts observed at the final level of a **multistage sampling** design. As WLS is a purely second-moment based approach, like quasi-likelihood, specification of the underlying probability model for counts is not required provided a consistent estimate of $\mathrm{cov}_A(\widehat{\boldsymbol{F}})$ is available. For survey data, such an estimate is sometimes obtained by some form of pseudoreplication rather than by the delta method.

Generalized Estimating Equations (GEE)

WLS methodology requires approximately Gaussian functions of counts prior to modeling, and so cannot accommodate covariates observed at the level of the individual observation. GLMs and quasi-likelihood analyses, which naturally accommodate individual level covariates, are overparameterized for many purposes, particularly when marginal rather than conditional distributions are of most interest. Both approaches attain asymptotic efficiency through consistent estimation of the optimal weight matrix at each stage of a WLS iteration. Suboptimal weighting in WLS or IWLS, nevertheless, yields a consistent estimator of β. **Generalized estimating equation** analysis is an IWLS approach that extends the quasi-likelihood estimating equations, seeking to combine advantages of GLM and WLS by accepting suboptimal weighting and some loss of asymptotic efficiency in order to gain flexibility and **robustness**. More specifically, if some efficiency loss can be tolerated, the information matrix may be modeled only approximately, without fully accounting for nuisance association parameters from the underlying probability

distribution. This is particularly advantageous for the analysis of correlated categorical responses, such as occur in repeated measures experiments or in repeated observations over time, in those instances when the association structure itself is largely a nuisance in relation to the objectives of data analysis (*see* **Correlated Binary Data**).

For $a_i(\phi) = \phi$, the GLM estimating equations may be written as

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta_k}\right)\left(\frac{y_i - \mu_i}{\mathrm{var}(Y_i)}\right) = 0, \quad k = 1, \ldots, u, \tag{18}$$

and interpreted as quasi-likelihood equations when the distribution is unknown but $\mathrm{var}(Y_i)$ can be written as $V(\mu_i)$. Let $\boldsymbol{Y}_i$ by the random vector and $\boldsymbol{y}_i$ be the corresponding vector of observations for the i^{th} individual or other observational unit. The generalized estimating equations are a straightforward multivariate extension of (18),

$$\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right) \boldsymbol{V}_i^{-1}(\boldsymbol{\mu}_i, \boldsymbol{\alpha})(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \boldsymbol{0}. \tag{19}$$

In this extension, $\boldsymbol{\mathcal{D}}_i$ is the matrix of derivatives of means of the elements of $\boldsymbol{Y}_i$ relative to those of $\boldsymbol{\beta}$ and $\boldsymbol{S}_i$ is the deviation vector $\boldsymbol{y}_i - \boldsymbol{\mu}_i$, directly generalizing $\partial \mu_i / \partial \beta_k$ and $(y_i - \mu_i)$. However, the true $\mathrm{Var}(Y_i)$ in the univariate case is replaced in the multivariate generalized estimating equations by a working model $\boldsymbol{V}_i(\boldsymbol{\mu}_i, \boldsymbol{\alpha}) = \boldsymbol{A}_i(\boldsymbol{\mu}_i)^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i(\boldsymbol{\mu}_i)^{1/2}$ for $\mathrm{cov}(\boldsymbol{Y}_i)$ with $\boldsymbol{R}_i(\boldsymbol{\alpha})$ a correlation matrix dependent on a vector $\boldsymbol{\alpha}$ of association parameters, and $\boldsymbol{A}_i(\boldsymbol{\mu}_i)$ a diagonal matrix of variances. If the dimension of $\boldsymbol{\alpha}$ is low, a parsimonious model thus replaces the more complex covariance structure of the underlying product-multinomial or other distribution. In the simplest implementation, known as "GEE1," (19) is solved by repeated cycles of IWLS estimation of $\boldsymbol{\beta}$ given $\boldsymbol{\alpha}$ from the previous step, and moment-based estimation of $\boldsymbol{\alpha}$ using the Pearson residuals $(y_{ij} - \mu_{ij})/\sqrt{([\boldsymbol{V}_i(\boldsymbol{\mu}_i, \boldsymbol{\alpha})]_{jj})}$ based on current values of the parameters. The resulting $\overline{\boldsymbol{\beta}}$ is asymptotically multivariate Gaussian with mean $\boldsymbol{\beta}$, whatever the working covariance structure employed. The asymptotic covariance matrix $\mathrm{cov}_A(\overline{\boldsymbol{\beta}})$ does depend on the working covariance structure, but a consistent "sandwich estimator" is

$$\left(\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} \boldsymbol{\mathcal{D}}_i\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} \mathrm{cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{\mathcal{D}}_i\right) \times \left(\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} \boldsymbol{\mathcal{D}}_i\right)^{-1} \tag{20}$$

evaluated at $\overline{\boldsymbol{\beta}}$, $\overline{\boldsymbol{\alpha}}$, and $\overline{\phi}$.

The general covariance estimator (20) is employed in two ways, with the appropriate choice depending on sample size and presumed adequacy of the covariance model. When the working covariance structure is based on a highly plausible underlying distributional model and/or reflects well the empirical covariance structure, then it is reasonable to substitute $\boldsymbol{V}_i$ for $\mathrm{cov}(\boldsymbol{Y}_i)$ in (20). This yields the model-based covariance estimator

$$\left(\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} \boldsymbol{\mathcal{D}}_i\right)^{-1}. \tag{21}$$

When the working covariance structure is correct, this estimator, evaluated at the fitted parameters $\overline{\boldsymbol{\beta}}$, $\overline{\boldsymbol{\alpha}}$, and $\overline{\phi}$, is consistent for the true covariance and gains efficiency from the information in the assumed covariance pattern. However, this covariance estimator is inconsistent if the working covariance structure is substantially incorrect, and invalid inferences can result. An alternative choice is to substitute the empirical estimator $\boldsymbol{S}_i \boldsymbol{S}_i'$ for $\mathrm{cov}(\boldsymbol{Y}_i)$. This yields the "robust" sandwich covariance estimator

$$\left(\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} \boldsymbol{\mathcal{D}}_i\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{S}_i \boldsymbol{S}_i') \boldsymbol{V}_i^{-1} \boldsymbol{\mathcal{D}}_i\right) \times \left(\sum_{i=1}^{n} \boldsymbol{\mathcal{D}}_i' \boldsymbol{V}_i^{-1} \boldsymbol{\mathcal{D}}_i\right)^{-1}, \tag{22}$$

also evaluated at the estimated parameters. "Robust" is used in this instance specifically to refer to consistency of this estimator for $\mathrm{cov}_A(\overline{\boldsymbol{\beta}})$ when the working covariance structure is misspecified, rather than in any more general sense of resistance to contamination. Large-sample confidence intervals and test statistics may be formed in the usual ways from (21) or (22). The robust sandwich estimator provides

great flexibility to the GEE method by providing reliable inferences in large samples with complex nuisance covariance structures. However, its convergence is not rapid even in standard Gaussian situations. Like other methods based on empirical variability estimates, such as Wald-based inference for some models, performance in small samples can be unacceptable in the sense that actual test levels can be noticeably higher, and confidence interval coverages noticeably lower, than their nominal specifications.

An alternative GEE implementation augments the estimating equations for the marginal mean structure with additional estimating equations for the second-moment structure. The estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are thus obtained from a simultaneous fit of the multivariate GLM and the residual variances and covariances or odds ratios. Alternative estimation methods for both GEE1 and this "GEE2" modification apply a Gauss–Newton algorithm to somewhat different versions of the estimating equations.

GEE methods have been successfully used for marginal modeling of **longitudinal** categorical data with covariates. In such analyses, net shifts in category distributions over time are of primary interest rather than the transition trajectories of individuals. If a dichotomous variable is studied at five time points, 26 association parameters may be needed to represent the association structure of the 2^5 table of response counts. A GEE analysis might crudely approximate this association structure using a patterned covariance matrix with $\boldsymbol{\alpha}$ of only one to three elements, say from a simple **time series** model. Accumulating evidence suggests that reasonable but simplified covariance models often sacrifice little efficiency. Modifications of GEE in the presence of **random effects**, and for **generalized additive models**, have also been developed (*see* **Marginal Models**; **Generalized Linear Models for Longitudinal Data**).

Generalized Linear Mixed Models (GLMMs)

Generalized linear mixed models (GLMMs) are GLMs in which the linear predictor $\boldsymbol{X\beta}$ is offset by additive contributions of random effects. The GLM $\boldsymbol{g(\mu) = \eta = X\beta}$ is thus generalized to $\boldsymbol{g(\mu_{(C)}) = \eta_{(C)} = X\beta_{(C)} + Z\zeta}$ with $\boldsymbol{\mu_{(C)}} = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{\zeta})$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ known model matrices, $\boldsymbol{\beta_{(C)}}$ fixed and unknown, and $\boldsymbol{\zeta}$ random with a specified, usually Gaussian, cdf. The subscript "(C)" is used to indicate that GLMM fixed effect parameters represent conditional effects given particular values of the random effects.

The random effects produce mixture distributions that account for extra-Poisson or extra-binomial/multinomial variability due to sample heterogeneity (*see* **Overdispersion**). They also allow description of associations among multiple categorical observations more parsimoniously than by a full multinomial representation or a loglinear model. With Gaussian random effects, association is modeled by the variances of underlying random effects distributions, that is, the usual **variance components** in the mixed model for Gaussian dependent variables. The variance components in a GLMM thus serve, in part, the same purpose as $\boldsymbol{\alpha}$ in GEE analyses. Beyond this, however, the random effects in $\boldsymbol{\zeta}$ can be estimated and the estimates used to characterize a correlated cluster of observations, such as sequential measurements on an individual over time, and to predict additional observations from the observational unit underlying the cluster. Two important special cases are the classical linear mixed model and the mixed logistic regression model, that is, the binomial GLM with canonical logit link and Gaussian random effects.

We emphasize that unless the GLMM link is the identity, which is rarely the case in categorical data analysis, the addition of random effects to the GLM *changes the meaning of the fixed effects* $\boldsymbol{\beta}$. While in the GLMM, as noted above, these are parameters $\boldsymbol{\beta_{(C)}}$ of the conditional distribution of the Y_i for given $\boldsymbol{\zeta}$, in the GLM they are parameters $\boldsymbol{\beta = \beta_{(U)}}$ of the unconditional distribution of the Y_i, that is, of the joint distribution of random $\boldsymbol{Y}$ and random $\boldsymbol{\zeta}$ integrated over $\boldsymbol{\zeta}$. Since this unconditional distribution is a marginal of the joint distribution, obtained by integrating out a mixing distribution that differentiates collections of correlated observations, the terms "marginal" and "population-averaged" are often substituted for unconditional in the GLMM literature; similarly, "subject-specific" is sometimes used in place of conditional.

The relationship of $\boldsymbol{\beta_{(U)}}$ to $\boldsymbol{\beta_{(C)}}$ can sometimes be approximated. For the mixed logistic regression model, the population-averaged parameters $\boldsymbol{\beta_{(U)}}$ are invariably attenuated in relation to the subject-specific parameters $\boldsymbol{\beta_{(C)}}$. An example of a practical interpretation of the attenuation is that odds ratios from analytic epidemiologic studies of smoking and lung cancer understate the biological impact

of an individual's decision to smoke or refrain from smoking.

Under a GLMM, estimation of either $\boldsymbol{\beta}_{(C)}$ or $\boldsymbol{\beta}_{(U)}$ is preferably performed by maximum likelihood, quasi-likelihood, or **restricted maximum likelihood** (REML) estimation from the marginal likelihood. This task is straightforward in principle but exceptionally challenging in practice. In addition to the nontrivial optimization problems posed by linear mixed models, for which the marginal likelihood is available analytically, estimation for the GLMM requires that the marginal likelihood be obtained numerically before it can be optimized. Methods currently available for doing this, by numerically integrating $\boldsymbol{\zeta}$ out of the likelihood or quasi-likelihood function using Gaussian quadrature or otherwise, or by generating observations from the marginal distribution using **Markov Chain Monte Carlo** (MCMC), have not been consistently successful for models with more than a few random effects or nested variance components. At the time of this writing, most analysis is performed using methods that maximize an approximate likelihood or that solve approximate quasi-likelihood or generalized estimating equations. Various approximations using Taylor series and Laplace expansions with differing simplifying assumptions have been implemented, sometimes using **profile likelihoods**. The approximate methods appear to work well in many circumstances, but not with small numbers of observations per random effect. Active research continues on the properties of each method and possible further improvements (*see* **Generalized Linear Models for Longitudinal Data**).

Bayes and Empirical Bayes (EB) Methods

GLMMs introduce additive random effects into the linear predictor of a GLM. This can be formally identical, for example, in a "random intercepts" model, to sampling subject-specific GLM fixed effects from a prior distribution. A substantial Bayes and EB literature employs prior distributions for contingency table probabilities, logits, or loglinear model parameters. Such methods can model heterogeneity, and smooth naive estimates from small samples or sparse tables towards more substantively reasonable values by "shrinking" them towards parsimonious model structures (*see* **Shrinkage**). Applications include removal of zero probability estimates from tables with random empty cells (*see* **Structural and Sampling Zeros**); the Agresti–Coull confidence interval for a single proportion π centered around a Bayesian point estimate using a **Beta**(2, 2) prior for π; adjustment of local standardized mortality rates by smoothing local geographical variation (*see* **Geographic Patterns of Disease**); and identification of suspect adverse drug effects from a large cross-classification, by drug and type of event, of spontaneous reports to the US Food and Drug Administration (*see* **Pharmacoepidemiology, Overview**).

The Beta distributions are the conjugate family for binomials, and the Dirichlet generalizations of Beta distributions form the conjugate family to multinomials. Both noninformative and Dirichlet priors have been applied directly to expected counts and cell probabilities. The Dirichlet priors may embody simple underlying structure, for example, a low-order loglinear model, as in the correction to the Wald confidence interval for a difference in two proportions that was mentioned above (*see* **Loglinear Model**).

A Bayesian precursor to the logit-Gaussian GLMM uses a Gaussian $N(\mu, \sigma^2)$ prior for logits from an $r \times 2$ table, a noninformative prior for μ, and a χ^2_ν prior for $\nu\lambda/\sigma^2$. Such independent prior formulations have also been used for row, column, and interaction effects in the saturated loglinear model for an $r \times c$ table. An EB approach to such tables uses flat priors for main effects and a Gaussian prior with estimated variance for interactions. Other EB applications employ Beta and **lognormal** priors for ratios of cell probabilities to expectations under independence, gamma and lognormal priors for relative risks in geographic epidemiology, and log-logistic priors for mortality proportions.

A particularly interesting area of Bayesian application, and a subject of intense current research, is the study of geographic variation and temporal-spatial modeling of counts, for instance, of animal populations and incident disease cases. Beyond recognition of its broad importance for ecological and epidemiological research, this field has received stimulus from the urgency associated with recent concerns pertaining to emerging infectious diseases and to bioterrorism. Poisson regression models may be employed in GLMMs with various spatial correlation structures, including local association implied by the reciprocal conditional distributions of a Markov random field. Empirical Bayes and full Bayes treatments of various models with priors on parameters of the random

effect distributions are under investigation (*See* **Geographic Epidemiology**).

Recent advances in computational algorithms for Bayesian inference are greatly assisting development and application of Bayesian methods, and of other methods when Bayes, frequentist, and likelihood approaches have common computational problems. For instance, a practical estimation algorithm for the probit-Gaussian GLMM embeds a Gibbs sampling step into the **EM algorithm** (*see* **Bayesian Methods for Contingency Tables**).

Software

We close with remarks on **software**, which abounds in widely available commercial packages and is available for newer methods in less polished shareware. Comments that follow reflect the authors' experience, are not product endorsements, are by no means exhaustive, and will inevitably be somewhat outdated when this is read. For the more advanced procedures that are implemented in multiple packages, the specifics of each implementation often vary substantially in computational approach, default approaches, and in the extent of analytic, algorithmic, and reporting customization available to the user.

Basic chi-square statistics, measures of association for contingency tables, and generalized linear models are available in virtually all major statistical packages, for example, $SAS^{®}$ (Version 9), $S-Plus^{®}$ (Version 6) and R (Version 1.8.1), $STATA^{®}$ (Version 8), and $SPSS^{®}$ (Version 12). General ordinal association models for $r \times c$ and higher-dimensional tables are not commonly preprogrammed, although specific models are often available depending upon the customer base of a particular product. However, cumulative logit (proportional odds) analysis is widely available, and ordinal loglinear models may readily be implemented using packages that allow arbitrary model matrices $\boldsymbol{X}$, which include all of the above and the GLM package GLIM (Version 4). (Note that caution is indicated in reading software brochures. Some use GLM to designate only the Gaussian generalized linear model with identity link function, that is, the "general" rather than "generalized" linear model.)

The most general prepackaged implementations of CMH and WLS methods are in *SAS*® PROC FREQ and PROC CATMOD, respectively, but these analyses involve simple quadratic forms and are easily incorporated into packages that allow matrix manipulation and user programming. Any unconditional logistic regression software may be used to perform conditional logistic regression analyses of paired data, while software for proportional hazards survival regression models may be used for such analyses of general matched data.

In two areas, (i) exact analyses and (ii) WLS or GEE-based methods for modelling data from sample surveys and other clustered research designs, there is increasing overlap between specialty software and general purpose statistical packages such as those above. *StatXact*® (Version 6) implements rapid algorithms for many exact analyses of two- and three-way contingency tables (*see* **StatXact**). *LogXact*® (Version 5) performs exact logistic regression analyses. Each of these packages has incorporated broader capabilities with recent releases. However, exact logistic regression is now available as an option of *SAS*® PROC LOGISTIC, a core of the most common analyses pioneered in *StatXact*® are similarly available as options in *SAS*® PROC FREQ, and the full *StatXact*® package with *SAS*® connectivity is available as a *SAS*® add-on. The *SPSS*® add-on *SPSS*® Exact Tests™ has similar capabilities.

Similarly, the *SUDAAN*® package has pioneered powerful adjustment capabilities for correlated data from stratified multistage cluster designs, with particular attention to the analyses of complex national surveys, but initially was restricted to descriptive analyses and a limited set of models. *SUDAAN*® 8.0, the release as of this writing, now implements an array of continuous, univariate and multinomial logistic, loglinear, and discrete and continuous semiparametric proportional hazards models, using GEE. The survey analysis package *WESVAR*® employs jackknife or balanced repeated replication variance estimation to fit a less extensive class of models, including logistic and multinomial logistic models. However, *STATA*® and *SAS*® now include core survey modeling capabilities that, particularly for *STATA*®, accommodate a substantial proportion – though by no means all – of data analyses for which *SUDAAN*®, *WESVAR*®, or similar specialty software would previously have been required. An *SPSS*® add-on, *SPSS*® Complex Samples™, offers similar capabilities for descriptive analyses and basic hypothesis tests. There is hope that the wider availability of convenient methods for proper variance estimation from clustered survey designs will overcome the reluctance of many

researchers to account for clustering in their analyses (*see* **Software for Sample Survey Data, Misuse of Standard Packages**).

GEE implementations are routinely available in *SAS*® (GENMOD AND NLMIXED procedures), *STATA*® (xtgee and other xt... commands), and *S − Plus*® (gee function in correlated data research library, and other functions in private libraries). *SAS*® (NLMIXED, %GLIMMIX macro) and *S − Plus*® (glme and xglm functions in correlated data research library) provide both approximate and exact approaches to fitting GLMMs. *STATA*® (xt... procedures) fits them using quadrature to obtain the true likelihood. BUGS (Bayesian analysis Using Gibbs Sampling, with Windows implementation WinBugs), a general platform for hierarchical Bayesian modeling, allows GLMM fitting based on the true likelihood through MCMC.

EPIINFO™ and *EGRET*® are packages particularly oriented to epidemiologists, the first relatively basic and the second more comprehensive and advanced but less user-friendly. *EGRET*® includes some capabilities of *StatXact*® for two-dimensional contingency tables, and for stratification-based adjustment of such tables for confounding, (*see* **Software, Epidemiological**).

References

[1] Graubard, B.I. & Korn, E.L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ contingency tables, *Biometrics* **43**, 471–476.

[2] Imrey, P.B., Koch, G.G. & Preisser, J.S. (1996). The evolution of categorical data modeling: a biometric perspective, in *Advances in Biometry*, H.A. David & P. Armitage, eds. Wiley, New York, pp. 89–114.

[3] Lee, E.T. (1974). A computer program for linear logistic regression analysis, *Computer Programs in Biomedicine* **4**, 80–92.

Further Reading

Aerts, M., Geys, H., Molenberghs, G. & Ryan, L.M. eds. (2002). *Topics in Modelling of Clustered Data*. Chapman & Hall/CRC, London.

Agresti, A. (1984). *Analysis of Ordered Categorical Data*. Wiley, New York.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. Wiley, New York.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT, Cambridge.

Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-control Studies*. International Agency for Research on Cancer, Lyon.

Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research. Volume 2: The Design and Analysis Cohort Studies*. International Agency for Research on Cancer, Lyon.

Cameron, A.C. & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York.

Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman and Hall, London.

Dey, D.K., Ghosh, S.J. & Mallick, B.K. eds. (2000). *Generalized Linear Models: A Bayesian Perspective*. Dekker, New York.

Diggle, P.J., Heagerty, P., Liang, K.-Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd Ed. Clarendon, Oxford.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd Ed. MIT, Cambridge.

Finney, D.J. (1978). *Statistical Method in Biological Assay*, 3rd Ed. Griffin, London.

Fleiss, J.L., Levin, B. & Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*, 3rd Ed. Wiley, New York.

Forthofer, R.N. & Lehnen, R.G. (1981). *Public Program Analysis: A New Categorical Data Approach*. Lifetime Learning Publications, Belmont.

Gokhale, D.V. & Kullback, S. (1978). *The Information in Contingency Tables*. Dekker, New York.

Good, I.J. (1965). *The Estimation of Probabilities*. MIT, Cambridge.

Goodman, L.A. (1978). *Analyzing Qualitative/Categorical Data: Log-linear Models and Latent Structure Analysis*, J. Magidson, ed. Abt, Cambridge.

Haberman, S. (1974). *The Analysis of Frequency Data*. University of Chicago, Chicago.

Haberman, S. (1978). *Analysis of Qualitative Data*. Volume 1: Introductory Topics. Academic Press, New York.

Haberman, S. (1979). *Analysis of Qualitative Data*. Volume 2: New Developments. Academic Press, New York.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd Ed. Wiley, New York.

Imrey, P.B., Koch, G.G. & Stokes, M.E. (1981). Categorical data analysis: some reflections on the log linear model and logistic regression. Part I: historical and methodological overview, *International Statistical Review* **49**, 265–283.

Imrey, P.B., Koch, G.G. & Stokes, M.E. (1982). Categorical data analysis: some reflections on the log linear model and logistic regression. Part II: data analysis, *International Statistical Review* **50**, 35–63.

Koch, G.G., Imrey, P.B., Singer, J.M., Atkinson, S.S. & Stokes, M.E. (1985). *Analysis of Categorical Data*. University of Montreal, Montreal.

Kuritz, S.J., Landis, J.R. & Koch, G.G. (1988). A general overview of Mantel-Haenszel methods: applications and recent developments, *Annual Review of Public Health* **9**, 123–160.

Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables, *Biometrika* **65**, 581–590.

Lawson, A., Biggeri, A., Böhning, D., LeSaffre, E., Viel, J.-F. & Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health*. Wiley, Baffins Lane.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables, *Journal of the Royal Statistical Society* **B37**, 23–37.

McCullagh, P. & Nelder, J.A. (1999). *Generalized Linear Models*, 3rd Ed. Chapman and Hall, London.

McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data, *Journal of the American Statistical Association* **89**, 330–335.

McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.

Stokes, M.E., Davis, C.S. & Koch, G.G. (2001). *Categorical Data Analysis Using the SAS® System*, 2nd Ed. SAS Institute, Cary; Wiley, New York.

Vonesh, E.F. & Chinchilli, V.M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, Dekker, New York, 381–444.

(*See also* **Matrix Algebra**)

PETER B. IMREY & GARY G. KOCH

Categorizing Continuous Variables

It is common, especially in medical research, for continuous variables to be converted into categorical variables by grouping values into two or more categories (*see* **Grouped Data**). Some reasons for this preference are statistical, notably the avoidance of certain assumptions about the nature of the data, but there are other less well-defined reasons. There seems to be a general preference in many clinical areas to categorize individuals, and this seems to carry over to the analysis and interpretation of continuous data. While there may be clinical value in such classifications, there is no statistical reason why all continuous variables should be treated this way. Furthermore, while creating categories may avoid certain statistical problems, it leads to new ones.

The use of grouping for descriptive purposes may allow a simpler presentation and is not especially problematic. When the approach is carried forward to data analysis, however, more serious problems may arise. Grouping may be seen as introducing an extreme form of measurement error, with an inevitable loss of power. It is questionable whether trading power for simplicity is a wise bargain. Nevertheless, there may be good reasons to categorize, so it is valuable to consider how this might be done. We consider the issues in relation to categorization of **explanatory variables** in **regression** models, but similar considerations apply more widely.

Effect of Categorizing

Converting a continuous variable into a categorical one will result in some loss of information; but it can be, and often is, argued that with three or more categories the loss is small and is offset by a gain in simplicity and the avoidance of assumptions. Connor [3] quantified the loss of information when categorizing a **normally distributed** variable which is linearly related to the outcome variable (also normally distributed) using the relative efficiency, which is based on the ratio of the expected variances of the estimated regression coefficients under the two models. For 2, 3, 4, or 5 groups, the efficiency relative to an ungrouped analysis is 65%, 81%, 88%, and 92% respectively (and is almost identical for an **exponentially distributed** variable).

Given the decision to categorize, it is not at all obvious how many groups to create. In practice the sample size should be one factor that influences the decision, as it is undesirable to have sparsely populated groups. The placing of the cutpoints may also not be obvious. Using optimally placed intervals, as derived by Connor [3] following Cox [5], is little different, in terms of efficiency, from using equally spaced intervals when the variable is normally distributed, but when the variable has an exponential distribution there is reduced efficiency with equiprobable intervals [9]. This result is important, as it is common to categorize in such situations, and equiprobable intervals are the norm. Morgan & Elashoff [11] examined the effect of categorizing a continuous covariate when comparing **survival** times. Here too, unequal groupings give increased efficiency.

Two Groups

Forcing all individuals into two groups simplifies the statistical analysis and may lead to easier interpretation of results. A **binary** split leads to a comparison of groups of individuals with high or low values of the measurement, or a comparison of proportions and a simple estimate of the difference between the groups (with its confidence interval). However, this simplicity is gained at the expense of throwing away a lot of information, as noted above. Cohen [4] observed that dichotomizing is equivalent to throwing a third of the data away. It thus leads to a loss of power to detect real relationships [15]. Patients are divided into just two categories, so that considerable variability may be subsumed. Using such binary variables in regression models may lead to **biased** estimates [15].

The choice of cutpoint to create the two groups may not be straightforward, and several approaches are possible. It is highly desirable for the choice of cutpoint not to be influenced by the actual data values. For a few variables there are recognized cutpoints which can be used (e.g. for body mass index). For some variables, such as age, it is usual to take a "round number" – an elusive concept, which in this context usually means a multiple of ten. Another possibility sometimes used is to use the upper limit of a normal range (reference interval). In the absence of any prior idea of a suitable cutoff, the most

common approach is to take the sample median. This gives two equal groups and is probably the best approach. However, using the median means that different studies will take different cutpoints, so that their results cannot easily be compared or combined. (Note that moving the cutpoint to a higher value leads to higher mean values in *both* groups.)

The arbitrariness of the choice of cutpoint may lead to the idea of trying more than one value and choosing that which, in some sense, gives the most satisfactory result. The temptation to perform multiple analyses should be strongly resisted. Taken to extremes, this approach leads to trying every possible cutpoint and choosing the value which minimizes the P value. Because of the multiple testing the overall false positive rate will be very high, being around 40% rather than the nominal 5% [2]. Also, the cutpoint chosen will have a wide confidence interval and will not be clinically meaningful and, crucially, the difference between the two groups will be overestimated. Although it is possible to correct the P value for multiple testing [2], the bias cannot currently be corrected for. Regrettably, this unacceptable strategy has become common in the cancer literature, especially in breast cancer [1, 2]. A similar proposal has appeared in the epidemiologic literature [17].

If the dependent variable is also a dichotomized continuous variable, the above effects are exaggerated. There is greater information loss [4], and there are two cutpoints to choose. The simple analysis of a **two-by-two table** can be via a **chi-square test** or an **odds ratio**. However, this is a drastic reduction of information. Maxwell & Delaney [10] studied the effect of dichotomizing two explanatory variables.

Several Groups

The simple approach of dichotomizing the explanatory variable may hide important complexities. The advantage of having several groups is that one can get a feel for the shape of a relation between the outcome variable and the explanatory variable without the need to specify the shape of that relation. Use of three or four categories is common, especially in epidemiology. It is rare for more than five groups to be used. With fewer groups there is an increased risk of pooling data for individuals with different risks. Occasionally, there may be a U-shaped relation between a variable and outcome. For example, Sather [13] showed such a relation between age and survival in children with acute lymphoblastic leukemia. With several groups such an effect can be seen clearly. With only two groups not only would this information be missed, but there could be a failure to detect any relation at all. When the true risk increases (or decreases) monotonically with the level of the variable of interest, the apparent spread of risk will increase with the number of groups used. It is thus possible to manipulate the message from the data, by choice of the number of groups.

It is most usual to divide the data into equal groups, for example using tertiles, quartiles or quintiles (*see* **Quantiles**) to divide the data values into three, four, or five groups respectively. Optimal groupings, as described above, are very rarely used. The only other reasonable approach is to split at round numbers, usually leading to groups of equal width rather than equal size, but not necessarily. The use of round numbers is esthetically pleasing, and will increase the likelihood of providing data comparable with those in other studies. It would be valuable to have standard groups for common variables, such as serum cholesterol or blood pressure. Data could still be analyzed in the authors' favored manner, but with results also presented using standard groups. This practice would considerably aid those carrying out meta-analyses.

After creation of groups, there are several options for the method of analysis. Because the groups are ordered it is desirable to use a method that specifically looks for a trend across the groups (*see* **Ordered Categorical Data**). An exception would be in the fairly rare case in which one had an a priori hypothesis of a "U-shaped" relation.

With k categories it is common to create $k - 1$ binary indicator variables (or **dummy variables**). These are entered into the **multiple regression** model. In order to test the effect of the explanatory variable X, all $k - 1$ dummy variables can be assessed at once using a single test with $k - 1$ degrees of freedom. This approach is lacking in power against a monotonic relation between the response variable Y and X. An alternative is to use $k - 1$ tests to examine each indicator variable separately, which may be done as part of a stepwise procedure. Without allowance for multiple comparisons, however, this method will lead to a increased risk of a false positive result. This method can lead to models which are hard to interpret. Walter et al. [16] recommend a different way

of creating dummy variables, where each represents whether or not an individual is in a particular group or a higher one. Here stepwise selection cannot lead to a nonsensical model, as each dummy variable represents a meaningful comparison. This approach is more sensible, but it does not seem to be much used. The problem of multiple testing remains, however.

A different strategy is to give each of the k groups a **score** (or rank) from 1 to k, and then to treat these scores as a continuous variable, say S. The regression coefficient for S then estimates the change in Y between adjacent groups. A closely similar approach involves using the mean of X within each category in place of group scores. This is a good general method, but it does impose some assumptions regarding the nature of the relation between the explanatory variable and outcome.

Creating Risk Groups

Rather than group according to the distribution of observed values of a variable, a preferable general approach is to base any grouping on outcome. This can be done either when there is a single variable of interest, or when several variables are considered simultaneously in a multiple regression model. The fitted values from the chosen model are obtained as a weighted sum of the variables in the model, where the regression coefficients are the weights. The weighted sum is often called a *prognostic index* (PI) (*see* **Prognosis**). If there are no continuous variables the PI will have as many possible values as there are distinct covariate patterns. The PI will be a continuous score if the model includes at least one continuous variable. Creating meaningful groups is often desirable for clinical reasons. As an example, the Nottingham breast cancer index [7], based on three variables, yields three risk groups. Such an approach is quite common in cancer. The advantage of keeping any grouping to the end of the analysis is that, while some categorization may be applied to some variables at an early stage, it is not required.

For grouping the PI, similar considerations apply as have been discussed. There are no easy answers, though. When creating two risk groups, the situation closely resembles the problem of finding an appropriate cutpoint for a continuous measurement or score for diagnostic purposes. Here too, a data-derived cutpoint can be obtained with a minimum P value – and again the P value but not the estimate can be adjusted for multiple comparisons [14].

Discussion

Categorization of continuous data is not necessary, and indeed is not a natural way of analyzing continuous data for most statisticians. Grouping is often used because of concern about the risks associated with mismodeling the relationship. Several strategies can be applied to assess whether the relation may in fact be curved and to model such relationships [6]. A second possibility is to use results obtained after categorization to guide the choice of an appropriate model [18] (*see* **Model, Choice of**). It is important to note that the use of categories does not remove assumptions about the nature of the relationship, as the above comments on analysis indicate.

Given the widespread practice of categorizing continuous variables, it is remarkable how little attention has been given to the way in which this should best be done. Several textbooks on statistics and epidemiology make no apparent mention of the topic, even though they present data that have been categorized. Among those that do mention the issue there is disagreement. Kramer [8, p. 179] says, with respect to creating categories, that "It is essential that these boundaries be decided a priori ... so that the investigator is not at liberty to pick a cutoff point that optimizes his chances for demonstrating statistical significance". (He considered only the case of a single cutpoint.) By contrast, Rothman [12] states that "it is often preferable to define the final categories after reviewing the data, notwithstanding the common advice that it is somehow more 'objective' to do so in ignorance of the distribution of observations in hand". These remarks indicate that, despite wide familiarity, this area of application would benefit from further research into desirable strategies.

References

[1] Altman, D.G. (1991). Categorizing continuous variables, *British Journal of Cancer* **64**, 975.

[2] Altman, D.G., Lausen, B., Sauerbrei, W. & Schumacher, S. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors, *Journal of the National Cancer Institute* **86**, 829–835.

[3] Cohen, J. (1983). The cost of dichotomization, *Applied Psychological Measurement* **7**, 249–253.

[4] Connor, R.J. (1972). Grouping for testing trends in categorical data, *Journal of the American Statistical Association* **67**, 601–604.

[5] Cox, D.R. (1957). Note on grouping, *Journal of the American Statistical Association* **52**, 543–547.

[6] Greenland, S. (1995). Dose–response and trend analysis in epidemiology: alternatives to categorical analysis, *Epidemiology* **6**, 356–365.

[7] Haybittle, J.L., Blamey, R.W., Elston, C.W., Johnson, J., Doyle, P.J., Campbell, F.C., Nicholson, R.I. & Griffiths, K. (1982). A prognostic index in primary breast cancer. *British Journal of Cancer* **45**, 361–366.

[8] Kramer, M.S. (1988). *Clinical Epidemiology and Biostatistics*. Springer-Verlag, Berlin.

[9] Lagakos, S.W. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable, *Statistics in Medicine* **7**, 257–274.

[10] Maxwell, S.E. & Delaney, H.D. (1993). Bivariate median splits and spurious statistical significance, *Psychological Bulletin* **113**, 181–190.

[11] Morgan, T.M. & Elashoff, R.M. (1986). Effect of categorizing a continuous covariate on the comparison of survival time, *Journal of the American Statistical Association* **81**, 917–921.

[12] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company, Boston, pp. 135–136.

[13] Sather, H.N. (1986). The use of prognostic factors in clinical trials, *Cancer* **58**, 461–467.

[14] Schulgen, G., Lausen, B., Olsen, J.H. & Schumacher, M. (1994). Outcome-oriented cutpoints in analysis of quantitative exposures, *American Journal of Epidemiology* **140**, 172–184.

[15] Selvin, S. (1987). Two issues concerning the analysis of grouped data, *European Journal of Epidemiology* **3**, 284–287.

[16] Walter, S.D., Feinstein, A.R. & Wells, C.K. (1987). Coding ordinal independent variables in multiple regression analyses, *American Journal of Epidemiology* **125**, 319–323.

[17] Wartenberg, D. & Northridge, M. (1991). Defining exposure in case-control studies: a new approach, *American Journal of Epidemiology* **133**, 1058–1071.

[18] Zhao, L.P. & Kolonel, L.N. (1992). Efficiency loss from categorizing quantitative exposures into qualitative exposures in case–control studies, *American Journal of Epidemiology* **136**, 464–474.

DOUGLAS G. ALTMAN

Cauchy Distribution

S.D. Poisson discovered the Cauchy distribution in 1824, long before its first mention by A.L. Cauchy [12]. Early interest in the distribution focused on its value as a counter-example which demonstrated the need for regularity conditions in order to prove important limit theorems [12] (*see* **Large-sample Theory**). The Cauchy distribution arises naturally as the ratio of two independent **standard normal** random variables (i.e. **Student's *t* distribution** with one **degree of freedom** is a Cauchy distribution). Also, if θ is **uniformly distributed** on $(-\pi/2, \pi/2)$, then $\tan\theta$ has a Cauchy distribution. The Cauchy distribution also arises as a mixture of normals: if Y follows the **chi-square distribution** with one degree of freedom, and, given $Y = y$, X is normal with mean 0 and variance y^{-1}, then X has a Cauchy distribution. This derivation of the Cauchy distribution as a mixture motivates its use for **robust regression** analysis of data sets in which the errors have longer than normal tails [9]. Another application of the Cauchy distribution is as a useful alternative to a normal **prior distribution** in the situation where a thicker tail than the normal distribution is reasonable [4].

The Cauchy distribution has density

$$f(x) = \frac{\beta}{\{\pi[\beta^2 + (x-\alpha)^2]\}}, \quad \text{where } \beta > 0,$$

and cumulative distribution function

$$F(x) = \frac{\left\{\arctan\left[\frac{(x-\alpha)}{\beta}\right] + \frac{\pi}{2}\right\}}{\pi}.$$

The distribution is symmetric about its median α, and has first and third quartiles $\alpha - \beta$ and $\alpha + \beta$, respectively (*see* **Quantiles**). The standard Cauchy distribution has $\alpha = 0$ and $\beta = 1$.

The **characteristic function** of the Cauchy distribution is

$$\phi(t) = \exp(it\alpha - \beta|t|).$$

From this it follows that the distribution does not have finite **moments** of any order. Linear combinations of independent Cauchy random variables also follow the Cauchy distribution. In particular, if X_1 and X_2 have independent Cauchy distributions with medians α_1 and α_2 and scale parameters β_1 and β_2, respectively, then $a_1X_1 + a_2X_2$ also has a Cauchy distribution with parameters $a_1\alpha_1 + a_2\alpha_2$ and $|a_1|\beta_1 + |a_2|\beta_2$, respectively.

The mean of n independent and identically distributed Cauchy random variables with median α and scale β also has a Cauchy distribution with median α and scale β. Hence, the sample mean is not a **consistent estimator** of the median α and, in fact, it offers no increase in accuracy compared with any single value [8]. Convenient estimators of α and β are available using the **order statistics** [3, 7, 11]. For example, the median is an **unbiased** estimator of α with 81% **asymptotic relative efficiency** [2]. For **maximum likelihood** estimates, care must be taken because the **likelihood** equations can have multiple roots. Haas et al. [5] have provided tables of critical values necessary to construct confidence intervals for estimates of α and β.

While the ratio of two independent normal random variables follows the Cauchy distribution, Cauchy distributions can also arise as the ratio of other, possibly dependent, random variables [1]. For additional characterizations and applications of the Cauchy distribution, see [6] and [10].

References

[1] Arnold, B.C. & Brockett, P.L. (1992). On distributions whose component ratios are Cauchy, *American Statistician* **46**, 25–26.

[2] Barnett, V.D. (1966). Order statistics estimators of the location of the Cauchy distribution, *Journal of the American Statistical Association* **61**, 1205–1218.

[3] Chernoff, H., Gastwirth, J.L. & Johns, M.V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation, *Annals of Mathematical Statistics* **38**, 52–72.

[4] Good, I.J. (1986). Some statistical applications of Poisson's work, *Statistical Science* **1**, 157–180.

[5] Haas, G., Bain, L. & Antle, C. (1970). Inference for the Cauchy distribution based on maximum likelihood estimators, *Biometrika* **57**, 403–408.

[6] Hall, P. (1994). On the erratic behavior of estimators of N in the binomial N, p distribution, *Journal of the American Statistical Association* **89**, 344–352.

[7] Johnson, N.L. & Kotz, S. (1970). *Continuous Distributions*, Vol. 1. Wiley, New York, pp. 154–165.

[8] Kendall, M.G. & Stuart, A. (1967). *The Advanced Theory of Statistics*, Vol. II, 2nd Ed. Griffin, London.

[9] Lange, K.L., Little, R.J.A. & Taylor, J.M.G. (1989). Robust statistical modeling using the *t* distribution, *Journal of the American Statistical Association* **84**, 881–896.

[10] McCullagh, P. (1992). Conditional inference and Cauchy models, *Biometrika* **79**, 247–259.

[11] Rothenberg, T.J., Fisher, F.M. & Tilanus, C.B. (1964). A note on estimation from a Cauchy sample, *Journal of the American Statistical Association* **59**, 460–463.

[12] Stigler, S.M. (1974). Cauchy and the witch of Agnesi: an historical note on the Cauchy distribution, *Biometrika* **61**, 375–380.

ROBERT J. GLYNN

Causal Direction, Determination

Causality analysis studies the cause–effect relationships among several variables (*see* **Causation**). There are several definitions of causality. Here we restate one that is in terms of probability [2]: C is a candidate cause of E if $\Pr(E|C) > \Pr(E|\overline{C})$. But the role of a potential **confounder**; B say, should be taken into account. It is possible that C leads to B and then B leads to E. Formally, C is a spurious cause if $\Pr(E|C \cap B) = \Pr(E|\overline{C} \cap B)$ which means that the real cause of E is B rather than C. Therefore, conditional on B, E is not affected by C. However, there may be a large number of potential confounders and some of them are unobserved, so that testing for $\Pr(E|C \cap B) = \Pr(E|\overline{C} \cap B)$ may be difficult. Causal relationship can be represented by a path diagram (*see* **Path Analysis**), on which the relationships between variables can be shown by arrows. For example $C \to E$ means that C is the cause of E; that is, B is the direct cause of E. Determining causal direction includes determining paths and their directions. A traditional causal analysis method is the **structural equation model**, in which the causes appear in the model as independent **explanatory variables** and the effects as dependent **response variables**. Bentler & Newcomb [1] proposed the following approach: (i) Form a path diagram. (ii) Form a **multivariate multiple regression** model with one equation for each effect, and all of its causes as regressors of this effect. The correlations between independent variables are represented by correlation matrices. (iii) Parameters are estimated and tested by model fitting. Special programs for this purpose are available (e.g. EQS) [4]. For an introduction to graphic models for several kinds of outcomes, see [3]. Although the above approaches are straightforward, the asymmetry between causes and effects may not be determined by the model and available data alone. Additional information such as temporal ordering or subjective knowledge may have to be used. Recently, acyclic graphic models have been used in causal analysis [6]. However, several assumptions are made which may not be realistic in practice and large sample sizes are needed to determine the causal relationship.

In **longitudinal** studies, repeated measurements are available. However, the structural equation models do not take the time order into account. An extension of these models for longitudinal data is the dynamic regression model which imposes cause–effect relationships on a multivariate time series [5] (*see* **Multiple Time Series**). For time series models, Granger's causality [8] is defined in terms of *predictability according to a law*. For example, suppose that X_t and Y_t, $t = 1, 2, \ldots, T$, are two time series. If adding the history of $Y_t(Y_1, \ldots, Y_{t-1})$ leads to a better prediction of X_t than using the history of X_t alone, then Y causes X.

A general form for time series causality models is [5]

$$x_t = \sum_{s=1}^{p} E_s x_{t-s} + \sum_{s=1}^{q} F_s y_{t-s} + u_{1t}, \qquad (1)$$

$$y_t = \sum_{s=1}^{r} G_s y_{t-s} + \sum_{s=1}^{v} H_s x_{t-s} + u_{2t}, \qquad (2)$$

where $\mathrm{var}(u_{1t}) = \mathbf{\Sigma}_1$ and $\mathrm{var}(u_{2t}) = \mathbf{\Sigma}_2$. Causal direction can be determined by testing $F_s = 0, s = 1, \ldots, q$, and $H_s = 0, s = 1, \ldots, v$. The computer package EQS [4] can also be used for the analysis of longitudinal data, particularly to determine causal directions. Recently, combined graphic and linear dynamic models [7] have been proposed for use in the analysis of time series models with complicated structures.

References

[1] Bentler, P.M. & Newcomb, M.D. (1992). Linear structure equation modelling with non-normal continuous variables, in *Statistical Models for Longitudinal Studies of Health*, J.H. Dwyer, M. Feinleib, P. Lippert & H. Hoffmeister, eds. Oxford University Press, New York.

[2] Cox, D.R. (1992). Causality: some statistical aspects, *Journal of the Royal Statistical Society, Series A* **155**, 291–302.

[3] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies*. Chapman & Hall, London.

[4] Dunn, G., Everitt, B. & Pickles, A. (1993). *Modelling Covariances and Latent Variables Using EQS*. Chapman & Hall, London.

[5] Geweke, J. (1984). Inference and causality in economic time series, in *Handbook of Econometrics*, Vol. 2, Z. Griliches & M.D. Intriligator, eds. North-Holland, Amsterdam, pp. 1101–1144.

[6] Pearl, J. (1995). Causal diagrams for empirical research (with discussion), *Biometrika* **82**, 669–710.
[7] Queen, C.M. & Smith, J.Q. (1993). Multiregression dynamic models, *Journal of the Royal Statistical Society, Series B* **55**, 849–870.
[8] Sobel, M.E. (1995). Causal inference in the social and behavioral sciences, in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg & M.E. Sobel, eds. Plenum, New York.

(*See also* **Hill's Criteria for Causality**; **Time Series**)

B. JONES & J. WANG

Causation

The concepts of cause and effect are central to most areas of scientific research, so it is not surprising that the literature on them could fill a small library. What may be surprising, given their importance, is that consensus about basic definitions and methods for causal **inference** is (at best) limited, despite some three centuries of debate. A brief review cannot do justice to the history and details of this debate, nor to all the schools of thought on causation. This article will, therefore, focus on a few major themes that have affected modern biostatistical practice. Of necessity, some aspects of the discussion are simplified relative to the literature, and the references should be consulted for more thorough descriptions. Entries on related topics are given in the final section.

Counterfactual Causation

At least as far back as the early eighteenth century, philosophers noted serious deficiencies in ordinary definitions of causation (e.g. see Hume [8]). For example, *Webster's New Twentieth Century Dictionary* [12] offers "that which produces an effect or result" as a definition of "cause". The circularity of this definition becomes apparent when one discovers "to cause" among the definitions of "produces". In early scientific treatises, an event (or set of events) A was said to cause a later event B if there was "constant conjunction" or "regularity" of the events, in that A (the cause) was inevitably followed by B ("the effect"). Mill [13] pointed out that such "constant conjunction" could always be the effect of a third event C preceding A and B; in other words, the regularity of B following A might only be due to **confounding**[4]. Informal definitions of "effect" suffer from the same problems, because "effect" as a verb is merely a synonym for "cause", while "effect" as a noun is defined as a "result", which is, in turn, defined as an "effect" in causal contexts.

Hume [7], however, offered in passing another view of causation that pointed a way out of circularity or confounding in the definition (even if confounding might be inevitable in the observation). In the present terminology, Hume proposed that A caused B if failure of A to occur would have been sufficient for failure of B to occur (see Lewis [10]). That is, by focusing on specific instances of causation, we could say that a specific event A caused a specific event B if occurrence of A was necessary for B under the observed background circumstances. Essentially, the same concept of causation can be found in [2] and [13] (both quoted in [22]).

Of course, the preceding definition falls short of the formalism necessary for rigorous logical analysis. Such analysis first appeared in the statistics literature in [14]. The basic idea is as follows: Suppose N units indexed by $i = 1, \ldots, N$ are to be observed in an experiment that will assign each unit to one of K treatments $x_1, \ldots, x_K$. For each unit, the outcome of interest is the value of a response variable Y_i. It is assumed that Y_i will equal y_{ik} if unit i is assigned treatment x_k. Suppose that one treatment level, say x_1, is designated the reference treatment against which other treatments are to be evaluated (typically, x_1 is "no treatment", placebo, or standard treatment). We may then define the causal *effect* of $x_k (k > 1)$ on Y_i relative to x_1 (the reference) to be $y_{ik} - y_{i1}$. Alternatively, if the response is restricted to positive values (such as blood pressure), we may define the causal effect as y_{ik}/y_{i1} or $\log y_{ik} - \log y_{i1}$.

This definition of effect leads naturally to a precise usage for the word "cause". Prior to treatment, we say y_k would cause a change of $y_{ik} - y_{i1}$ in Y_i; if $y_{ik} - y_{i1} = 0$, we say x_k would cause no change in Y_i. After the experiment, if unit i had received treatment k, then we say that x_k caused a change of $y_{ik} - y_{i1}$ in Y_i; otherwise, we say that x_k would have caused a change of $y_{ik} - y_{i1}$ in Y_i.

There are four crucial restrictions that the preceding formalism places on the notion of effect (and, hence, cause). First, effects are defined only within comparisons of treatment levels. To say that "drinking two glasses of wine a day lengthened Smith's life by four years" is meaningless by itself. A reference level must be at least implicit to make sense of the statement. Smith might have lived even longer had she consumed one rather than two glasses per day. As given, the statement could refer to no wine or four glasses per day or any other possibility.

Secondly, more subtly and profoundly, the formalism assumes that y_{ik}, the response of unit i under treatment k, remains well defined even if unit i is *not* given treatment k. In the philosophy literature, this assumption and the problems attendant with it

are recognized as problems of *counterfactual* analysis [10, 11, 24] (see also the discussion of Holland [6]). The statement "if x_k had been administered, then the response Y_i of unit i would have been y_{ik}" is called a *counterfactual conditional*: it asserts that Y_i would equal y_{ik} if, *contrary to fact*, x_k had been administered to unit i. Consider again Smith's drinking. Suppose she would contract cancer at age 70 if she drank two glasses of wine a day, but would instead die of a stroke at age 68 if she drank no wine. If Y_i is her time to cancer and she drank two glasses per day, how could we define her counterfactual time to cancer given no wine? Without this definition, the effect of two glasses of wine vs. none would be undefined.

The preceding problem is common in survival analysis when **competing risks** are present. The problem is not solved by attempting to condition on "absence of competing risks": such hypothetical absence is itself not a well-defined counterfactual state, even though standard probability calculations (as used in product-limit estimates) make it appear otherwise (see Kalbfleisch & Prentice [9, p. 166], and Prentice & Kalbfleisch [18], for further discussion). Rather, the definition of the response must be amended to include the competing risks if the counterfactual definition is to be applied. For example, Y_i could become the pair comprising time of cancer or competing risk and an indicator marking the event at that time.

Thirdly, the effects captured by the counterfactual definition are *net effects*, in that they include all indirect effects and **interactions** not specifically excluded by the treatment definition. For example, Smith's consumption of two glasses of wine per day rather than none may have given her four extra years of life solely because one night at a formal dinner it made her feel unsteady and she had a friend drive her home; had she not drunk, she would have driven herself and been hit and killed by a drunk driver. This sort of indirect effect is not one we would wish to capture when studying biological effects of wine use. It is, nonetheless, included in our measure of effect (as well as any estimate) unless we amend our treatment definition to include holding constant all "risky" activities that take place during Smith's life. Such amendment is sometimes (simplistically) subsumed under the clause of "all other things being equal (apart from treatment)", but can be a serious source of ambiguity when the intervention that enforces the amendment is not well defined.

A fourth restriction, which may be considered an aspect of the third, is that the formalism assumes that treatments not applied to a unit could have been applied. Suppose Smith would not, and could not, stop daily wine consumption unless forced physically to do so. The effect of her actual two-glass-a-day consumption vs. the counterfactual "no wine" would now be undefined without amending the treatment definition to include forcing Smith to drink no wine, e.g. by forcibly injecting Smith with antabuse each day of her life. Such amendment would be of little interest, not just because of its wild impracticality, but because of the side effects it would introduce.

The preceding restriction is sometimes accounted for by requiring that the counterfactual definition of "effect" applies only to "treatment variables". The latter are defined informally as variables subject to manipulation of their levels; an additional restriction is made that each possible level (treatment) for the treatment variable has nonzero probability of occurrence (e.g. see Holland [6]). One may sense here an echo of the circularity in ordinary definitions of cause, for this notion of manipulation embodies having an effect on treatment levels. Nonetheless, it has been argued that one strength of the counterfactual approach is its explication of the ambiguities inherent in defining cause and effect [10, 22].

One model of causation that has enjoyed some popularity in epidemiology is the sufficient-component cause (SCC) model introduced by Rothman [20, 21, Ch. 2]. This model presents causal mechanisms via schematic "pie charts" composed of slices representing necessary causal components of mechanisms. It can be shown that this model can be mapped into the general counterfactual framework, although it involves certain nonidentifiable elaborations [21, Ch. 18].

Probabilistic Causation

A number of authors have attempted to formalize causation through axioms governing the evolution of probabilities over time (e.g. see Suppes [25] and Eells [1]). Such systems have attracted little attention in biostatistics. Other approaches include probabilistic extensions of the counterfactual approach. One is based on the distribution of fixed potential

responses; that is, the joint distribution $F(y_1, \ldots, y_K)$ of $y_{i1}, \ldots, y_{iK}$ in a population of units. We may also consider conditional distributions in subpopulations defined by **covariates** such as age, sex, and received treatment. Population effects can be defined as averages of individual effects over populations; statistical procedures for inferences about these effects follow from assumptions about sampling and treatment assignment mechanisms. The basic ideas were present in Fisher [4] and Neyman [14], and have been elaborated more generally since [23].

Another extension considers potential responses that are distributional parameters specific to units. For example, we could consider the probability that a given atom emits a photon in the second following absorption of a photon ("treatment 2") minus the probability of emission in the same second if no photon had been absorbed ("treatment 1"). This probability difference is the effect of photon absorption on the atom relative to no absorption. In quantum mechanics, this difference (effect) is well defined *whether or not a photon is actually emitted*. In fact, under the standard quantum model, the emission indicator ($Y_i = 1$ if the atom emits a photon in the second, 0 if not) is *not* well defined under counterfactual alternatives to the actual history of the atom. Fortunately, the latter ambiguity appears to have no practical implications for the gross phenomena studied in biostatistics. It does, however, illustrate the possibility of considering probabilities and expectations (rather than events) as responses in the counterfactual definition, even for macrophenomena.

Causal Inference

Of causal inference there may be even more written and less agreed than for the basic definitions of cause and effect. A discussion of this literature is beyond the scope of the present article. Issues of **bias**, validity, and generalizability in causal inference are discussed elsewhere (*see* **Bias in Case–Control Studies**; **Bias in Observational Studies**; **Confounding**; **Validity and Generalizability in Epidemiologic Studies**). A discussion of criteria for causal inference [5] may be found in **Hill's Criteria for Causality** and in chapter 2 of Rothman and Greenland [21]; most of these criteria are informal, and there are many objections to their use [21, Ch. 2].

Formalisms for causal inference in biostatistics have thus far been restricted to counterfactual-based procedures, as in [1, 14, 23], and to methods based on path diagrams or directed graphs. Starting with Wright [26], path diagrams (*see* **Path Analysis**) (more recently called causal diagrams) have been used to display assumptions about the absence of particular effects and to provide a basis for algorithms useful in determining whether effects are identifiable from a given observational process (e.g. Robins [19], Pearl [16], and Pearl & Robins [17]; *see* **Identifiability**). Most of these approaches take the notion of cause as a primitive; a few define effects in a manner formally equivalent to the counterfactual definition extended to probabilistic domains [15]. The counterfactual approaches emphasize the importance of **randomization** in assuring identifiability of causal effects [3, 4, 19, 23].

References

[1] Eells, E. (1991). *Probabilistic Causality*. Cambridge University Press, New York.

[2] Fisher, R.A. (1918). The causes of human variability, *Eugenics Review* **10**, 213–220.

[3] Greenland, S. (1990). Randomization, statistics, and causal inference, *Epidemiology* **1**, 421–429.

[4] Greenland, S. Robins, J.M. & Pearl, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science* **14**, 29–46.

[5] Hill, A.B. (1965). The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[6] Holland, P.W. (1986). Statistics and causal inference (with discussion), *Journal of the American Statistical Association* **81**, 945–970.

[7] Hume, D. (1739;. reprinted 1888) *A Treatise of Human Nature*. Oxford University Press, Oxford.

[8] Hume, D. (1748;. reprinted 1988) *An Enquiry Concerning Human Understanding*. Open Court Press, LaSalle.

[9] Kalbfleish, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure-Time Data*. Wiley, New York.

[10] Lewis, D. (1973). Causation, *Journal of Philosophy* **70**, 596–567.

[11] Lewis, D. (1973). *Counterfactuals*. Blackwell, Oxford.

[12] McKechnie, J.L., ed. (1979). *Webster's New Twentieth Century Dictionary*. Simon & Schuster, New York.

[13] Mill, J.S. (1862). *A System of Logic, Ratiocinative and Inductive*, 5th Ed. Parker, Son & Bowin, London.

[14] Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: essai des principes. English translation by Dabrowska, D. & Speed, T. (1990), *Statistical Science* **5**, 463–472.

[15] Pearl, J. (1993). Comment: graphical models, causality, and intervention, *Statistical Science* **8**, 266–269.

[16] Pearl, J. (2000). *Causality*, Cambridge University Press, New York.

[17] Pearl, J. & Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables, in *Uncertainty in Artificial Intelligence*, R.L. Mantaras & D. Poole, eds. Morgan Kaufman, San Francisco, pp. 444–453.

[18] Prentice, R.L. & Kalbfleisch, J.D. (1988). Author's reply, *Biometrics* **44**, 1205.

[19] Robins, J.M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods, *Journal of Chronic Diseases* **40**, Supplement 2, 139S–161S.

[20] Rothman, K.J. (1976). Causes, *American Journal of Epidemiology* **104**, 587–592.

[21] Rothman, K.J. & Greenland, S (1998). *Modern Epidemiology*, 2nd Ed. Lippincott, Philadelphia.

[22] Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies, *Statistical Science* **5**, 472–480.

[23] Rubin, D.B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism, *Biometrics* **47**, 1213–1234.

[24] Stalnaker, R.C. (1968). A theory of conditionals, in *Studies in Logical Theory*, N. Rescher, ed. Blackwell, Oxford.

[25] Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland, Amsterdam.

[26] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.

SANDER GREENLAND

Cause of Death, Automatic Coding

Accurate selection and coding of the underlying cause of death based on death certificates using the **International Classification of Diseases (ICD)** is a labor-intensive task requiring special training and knowledge. Persons without a medical background must learn basic anatomy and physiology as a prerequisite to cause of death coding training, but even trained nurses and physicians must learn the detailed procedures and rules embodied in the ICD. **Vital statistics** offices have had a continuing burden to maintain a well-trained and experienced mortality coding staff. However, it was not until there was a heightened interest in coding not only the underlying cause of death but the other causes on the death certificate as well, that efforts to utilize computer technology to code death certificates began in earnest (*see* **Cause of Death, Underlying and Multiple**). The dual coding burden of the different procedures required to produce the two kinds of mortality data was prohibitive for most, if not all, countries interested in producing enhanced mortality statistics.

There were several independent approaches to automation of the international selection and modification rules, most notably in England and Wales and in the United States (US). Responsible government authorities in these two countries kept in close touch with each other's progress, but resource constraints in the former left the US **National Center for Health Statistics** as the single remaining major investor in research and development of an automated mortality coding system.

Originally designed for a large mainframe computer, the US system consisted of two subsystems, MICAR (Medical Information, Classification, and Retrieval) and ACME (Automated Classification of Medical Entities). MICAR required an ICD code or a standardized diagnostic abbreviation to be assigned by a coder to each condition reported, along with a location code indicating where the condition was written in relation to the other conditions entered on the death certificate (*see* **Death Certification**). This information was entered into the computer. The MICAR software searched its internal dictionary until the condition was found and then it assigned a dictionary reference number to each such term. The reference numbers and their death certificate location codes (i.e. MICAR output) formed the input to the ACME module which then, through a series of logical decision tables, applied the international selection and modification rules to arrive at the underlying cause of death. The MICAR output data could then also be used to produce multiple cause of death tabulations by applying additional computer procedures designed for that purpose.

The original versions of both MICAR and ACME have undergone many iterations since their early development in the late 1960s. The current versions are designed to run on a personal computer and are based on the latest revision of the ICD. The MICAR module, now known as SuperMICAR, accepts English language diagnostic text as well as standard disease abbreviations as input while earlier versions required a coder to assign code numbers to each diagnostic term. This allows the data to be entered by persons who can operate a keyboard but who have no need for familiarity with the ICD.

The main features of the MICAR/ACME approach are as follows:

1. The coding of death certificates can be done by coders with less training and knowledge than is required of underlying cause of death coders, although the total number of individual codes to be assigned is greater.
2. The ACME decision tables residing in the computer are logical reflections of the steps an underlying cause of death coder is trained to follow in the application of the international selection and modification rules and therefore result in underlying cause codes with a very high degree of agreement with those resulting from the work of highly experienced human coders.
3. The same original coding of death certificates can be used to produce both underlying cause of death statistics and multiple cause of death statistics.
4. There is a higher degree of consistency in the results, since variation due to differences in interpretation of the rules by coders is eliminated.
5. Changes in the ICD, its rules, or their interpretation can be implemented at any time by modifying the appropriate decision table(s) and reprocessing the records that have been processed prior to the change.

A number of countries plan to implement this automated system when they begin using the Tenth Revision of the ICD. Because of this widespread interest, there will be increased international involvement with the content and possible future modification of the decision tables and in the way multiple cause of death data are manipulated and presented. Countries planning to use this automated system have participated in planning meetings with the system designers and, in the case of the UK, have assisted in some of the research for the latest version. In addition, a few other automated systems in languages other than English have been developed using, in part, the ACME decision table logic. This broader involvement of many countries in the same or highly similar automated programs is expected to contribute to greater uniformity and comparability of international mortality data.

ROBERT A. ISRAEL

Cause of Death, Underlying and Multiple

The underlying cause of death is defined by the **World Health Organization (WHO)** as "(a) the disease or injury which initiated the train of morbid events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury" [5]. This definition recognizes the importance of the public health principle of prevention. By having information on the sequence of conditions leading to death, from the initial disease or condition, through diseases arising as consequences of the initial disease, on to the final or terminal condition, it is believed that interventions can be found to break the train of events and reduce mortality from selected causes of death.

The concept of attributing to each death a single cause for statistical tabulation and analytic purposes is rooted in the early development of disease classifications. From the outset and through the first part of the twentieth century, disease classifications were focused primarily on the causes of mortality, and it was the principal cause of death – not symptoms, other concurrent conditions, nor modes of dying – that was of primary interest. At the First International Revision Conference for the **International Classification of Diseases (ICD)**, convened in 1900, Bertillon proposed a set of rules for selecting the single cause of death to be used for statistical purposes when more than one condition was reported [4]. In subsequent years, however, there was little uniformity in practice among countries for the selection of a single cause of death and it was not until 1938, at the Fifth Decennial Revision Conference, that formal recognition at the international level was given to the statistical problem of selecting a single cause of death where more than one cause was given on the death certificate (joint causes of death). This question had been under study in the United States (US), and the Conference requested the US government to continue its investigations in this regard. Accordingly, the US government established the US Committee on Joint Causes of Death, comprised of members from the US, Canada, the UK, and representatives in an advisory capacity from the Interim Commission of the WHO [2]. In recognition of the work of this committee and the recommendations of the WHO Expert Committee for the Preparation of the Sixth Revision of the International Lists of Diseases and Causes of Death, the Sixth Decennial Revision Conference, at its meeting in 1948, not only adopted the Sixth Revision of the International Lists but also an International Form of Medical Certificate of Cause of Death, a formal definition of underlying cause of death, and a standardized set of selection rules for arriving at an underlying cause of death when more than one condition is reported (*see* **Death Certification**). The recommendations of the Conference were accepted by the World Health Assembly in 1948 and incorporated into Regulations No. 1 under Article 21(b) of the WHO Constitution. The regulations serve as guidelines to Member States for the compilation of morbidity and mortality statistics in accordance with the ICD [3]. This acceptance of a standard form for certifying causes of death, the definition of the underlying cause, and the rules for selecting it from more than one reported cause was a major step in the quest for international comparability of mortality data.

While the underlying cause of death continues to form the basis for mortality data for countries with medically certified deaths, some have observed that there is a loss of potentially useful information when several diseases or conditions are reported on a death certificate but only one is used for statistical reporting and analysis. With the decline in importance, especially in developed countries, of infectious diseases and a concomitant rise in **life expectancy**, the number of death certificates listing several conditions, particularly chronic illnesses, has risen noticeably. In those cases in which more than one disease is reported, all of the diagnostic entries on the death certificate are known collectively as "multiple causes of death". From the group of multiple causes on a death certificate, an underlying cause is chosen in accordance with the standard definition and procedures, but unless additional steps are taken, the remainder of the entries on the medical certification of cause of death contribute nothing to the statistical collection of mortality data. Recognizing this, several countries, particularly England and Wales, France, Sweden, and the US, began experimenting in the 1960s and subsequent years with ways to capture the additional information and analyze and present the findings [1]. At the international level, the WHO convened several meetings of interested countries, the first in London in 1969, to review multiple cause of death coding

procedures and to compare findings. At those international meetings it was generally agreed that there was not enough similarity of approach, nor a clearly superior methodology, to recommend a single international procedure. However, the growing number of interested countries were encouraged to continue to develop and use their own methodologies and to continue to keep each other informed of results. Furthermore, it was emphasized that some form of multiple cause of death analysis and data presentation was an important adjunct to the traditional underlying cause approach. Late in the 1990s, with a growing number of countries beginning to rely on a common automated computer coding scheme (*see* **Cause of Death, Automatic Coding**), there appeared to be more likelihood that several countries would mutually agree to a uniform procedure for multiple cause of death coding and analysis. This, when realized, would form an important step toward international comparability of multiple cause data to enhance the traditional underlying cause of death statistics.

References

[1] Israel, R.A., Rosenberg, H.M. & Curtin, L.R. (1986). Analytic potential for multiple cause-of-death data, *American Journal of Epidemiology* **124**, 161–179.

[2] United States Bureau of the Census (1939). *Manual of the International List of Causes of Death (Fifth Revision) and Joint Causes of Death (Fourth Revision)*. US Government Printing Office, Washington.

[3] World Health Organization (1967). *WHO Nomenclature Regulations*. World Health Organization, Geneva.

[4] World Health Organization (1977). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*, Vol. 1. World Health Organization, Geneva.

[5] World Health Organization (1992). *International Statistical Classification of Diseases and related health problems: 10th revision*, Vol. 2. World Health Organization, Geneva, p. 30.

ROBERT A. ISRAEL

Cell Cycle Models

The proliferation of cells is fundamental to the study of the growth of tissues and organs in development, as well as the growth of colonies of bacteria (*see* **Bacterial Growth, Division, and Mutation**), yeast, and tumors. Specifically, the cell cycle refers to the identifiable stages in the division cycle of a cell based on its DNA content which is duplicated (except in meiosis) before cells can divide to produce progeny. There are four basic phases of the cell cycle: G_1 (gap one), S (DNA synthesis), G_2 (gap two), and M (mitosis). The time spent in these four phases, from birth to the division of a cell is called the *cell-cycle time*.

Numerous mathematical models have been proposed for the study of the variation in the cell-cycle time and phase durations during the last few decades. Cell-cycle modeling encompasses three major areas. The oldest of these areas is the study of cell population growth. These studies focus on predicting the time-dependent dynamics of a cell population or the time-independent distribution of cells within various cell-cycle phases. Both deterministic and stochastic models based on age-dependent progression have been used to estimate the transit times of different cell-cycle phases. The second area, cell-cycle analysis, is related to cytometric studies, and is based on characterizing cells on the basis of their DNA content or other proliferative markers, indicating the position in the cell cycle. The third and most recent area is based on detailed kinetics of growth regulation within an individual cell. The models in this area are deterministic and use nonlinear multivariable ordinary differential equations. The models in the first two areas, considered as macroscopic models, have quantitative experimental results for validation. However, the models in the third area, the microscopic models, still can be compared only with observed qualitative behavior of the relevant variables. These three types of cell cycle models are discussed below.

Macroscopic Models

Cell Population Models

The experimental observations that cells could be labeled in the S-phase with radioactive thymidine, and that the labeled cells could be identified in mitosis by autoradiography, gave rise to cell-cycle models based on the fraction of labeled mitosis (FLM) curves. These models considered the variation in phase and cycle duration by introducing multivariable and/or time delay equations to express the temporal distribution of the population with the uncertainties expressed by stochastic terms. **Regression** analysis was used to estimate the coefficients of those models. Birth rate, death rate, cell-cycle time, and population doubling time are a few of the widely used parameters in these models. Typically, the assumptions such as (i) the variability of these parameters from one cell to another, (ii) the correlations of these parameters from one generation to another, or (iii) correlation of a cellular property such as the cell size to these variables, have been tested. Although many of these models could predict the population density for a short period of time (related to experiments) none can predict adequately the long-term behavior. This difficulty is due to the environmental changes, cell-to-cell communications, and intracellular signals of the cell population. Hence, the control of cell proliferation gives rise to differences in the parameters of these models. Since the number of contributions discussing such cell-cycle models is extremely large, we suggest to interested researchers the review papers by Swan [15], Rubinow [13], Eisen [6], White [24], Cooper [5] and Arino [1], and strictly for deterministic models the work by Webb [23], Tucker & Zimmerman [19], and Arino & Kimmel [2]. These mathematical models for cell population studies overlap with other types of mathematical modeling such as **branching** models, compartmental models, **Markov processes**, discrete models, and fluid flow models. Although they fall under different subcategories, they all incorporate the same basic physical principle: mass balancing. Related models are used in other biological studies, such as **tumor growth**, cell population analysis of tissues (e.g. regeneration of epithelial cells after irradiation), and hematopoiesis.

Cell-Cycle Analysis (Measurement of DNA Content)

While the original cell-cycle models owe their genesis to the FLM experiment, a second class of models has been developed based on the recognition that DNA content can be used to identify the position of cells within the cell cycle ranging from diploid DNA

content in G_1, increasing throughout the S-phase, and being at twice diploid in $G_2 + M$ preparatory to division. The development of appropriate DNA stains and equipment such as flow cytometers has made it possible to characterize populations of cells rapidly on the basis of differing subpopulations and to infer changes in the total population on the basis of the changes in the fractions of cells within each cell-cycle phase. While it is possible to define different DNA amounts and, hence, differing amounts of progression within the S-phase, in both G_1 and $G_2 + M$, cells appear indistinguishable on the basis of DNA content, so that finer structures of cell-cycle progression are obscured. Because a histogram of the DNA distribution of a population of cells typically appears with all G_1 cells appearing in a single region and $G_2 + M$ cells appearing in a second region with cells in the S-phase at all points in between, a wide variety of procedures have arisen for deconvolving the total distribution into separate parts to obtain estimates of the fractions of cells in each phase. A discussion of these procedures may be found in flow cytometry textbooks such as Shapiro [14] and Watson [22].

More recently it has become possible to measure other markers of cell proliferation such as Ki67, PCNA, various cyclins and markers of cell death within populations of cells, and to refine the location of a subpopulation within the division cycle. These methods generally use multi-parameter measurements leading to large data sets (e.g. 50 000 cells with six parameters on each) and are currently under study.

While these macroscopic models are still studied, especially for understanding bacterial growth or cell growth *in vitro*, modern research has revealed considerable details about the intracellular kinetics of the biochemical events that drive a cell through the cell cycle. This information, which is accumulating rapidly, has opened a new aspect of cell-cycle modeling: the development of microscopic models.

Microscopic Models

Macroscopic studies of cell populations have demonstrated the need for further studies to elucidate the cause of the variability in the cell-cycle time of two daughter cells. Technological advances in biochemistry have enabled scientists to identify the biochemical changes within a cell, which permits the simulation of mechanistic models. This section gives a more detailed account of the mathematical modeling of the microscopic view of the cell cycle.

A large body of experimental work focuses on understanding the regulation of the cell cycle in eukaryotes. Studies of cell-cycle progression have led to the elucidation of a variety of proteins which are required for appropriate entry and exit from the S-phase (DNA synthesizing), entry and exit from the M-phase (mitosis and cell division), cell death, cell rest (quiescence), and permanent cell arrest (differentiation). These proteins include cyclins, and their associated kinases, which together form complexes required for progression through the cell cycle. The activation and deactivation of these kinases are controlled by other proteins called activators and inhibitors, respectively. Biochemically, phosphorylation, dephosphorylation, and proteolysis are the main mechanisms of control of these kinases at specific phases of the cell cycle. Mathematically, the kinetics of these proteins are represented by a nonlinear multivariable, first-order dynamic system. It is assumed that the oscillatory solutions of this system will mimic continuous cell proliferation. While very early contributions by Hyver & Le Guyader [8], Norel & Agur [9], and Goldbeter [7] modeled the M-phase regulation by two to three variable systems, Tyson [20] proposed a model for the regulation of the same phase by a six-variable system. All these authors established necessary criteria for obtaining stable oscillatory solutions and discussed the bifurcational properties of reduced systems. The papers by Thron [16], Obeyesekere et al. [11], and Busenberg & Tang [4] discussed further mathematical aspects of these models. A very detailed model for the same phase kinetics, which introduces a ten-variable system, was published by Tyson & Novak [21]. Experimental work continued to elucidate proteins that regulate other phases, namely the S- and G_1-phases; models for these phases have been proposed by Obeyesekere et al. [10, 12]. The kinetics or the mechanisms necessary for a cell to enter the S-phase (i.e. pass a restriction point) has been an important biological question. Some of Thron's mathematical models address this issue [17, 18]. While many authors have contributed towards cell cycle regulation, some have validated their models with qualitative experimental results. For example, see [3] for such a contribution.

Presently, the search for proteins affecting the cell cycle is ongoing. Our understanding of the number

and types of proteins that control cell-cycle regulation is growing rapidly; however, knowledge of their interactions and the rate constants is still sketchy. Thus, the results of these models have been restricted to qualitative analysis. Future cell-cycle modeling should integrate experimental results with the models developed to date, or, in other words, incorporate the microscopic models into the macroscopic models via the parameters of the latter. Our ability to do this will result in significantly more realistic cell-cycle models.

References

[1] Arino, O. (1995). A survey of structured cell population dynamics, *Acta Biotheoretica* **43**, 3–25.

[2] Arino, O. & Kimmel, M. (1989). Asymptotic behavior of a nonlinear functional-integral equation of cell kinetics with unequal division, *Journal of Mathematical Biology* **27**, 341–354.

[3] Bar-Or, R.L. Maya, R. Segal, L.A. Alon, U. Levine, A.J. Oren, M. (2000). Generation of Oscillations by the p53-Mdm2 Feedback Loop: A Theoretical and Experimental Study, *Proceedings of the National Academy of Sciences* **97**, 11250–11255.

[4] Busenberg, S. & Tang, B. (1994). Mathematical models of the early embryonic cell cycle: the role of MPF activation and cyclin degradation, *Journal of Mathematical Biology* **32**, 573–596.

[5] Cooper, S. (1991). Conjectures on the mathematics of the cell cycle, in *Lecture Notes in Pure and Applied Mathematics, Mathematical Population Dynamics (Proceedings of the Second International Conference)*, Vol. 131, O. Arino, D.E. Axelrod & M. Kimmel, eds. Marcel Dekker, New York, pp. 539–546.

[6] Eisen, M. (1979). Mathematical models in cell biology and cancer chemotherapy, in *Lecture Notes in Biomathematics*, Vol. 30. Springer Verlag, New York, pp. 1–43.

[7] Goldbeter, A. (1991). A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase, *Proceedings of the National Academy of Sciences* **88**, 9107–9111.

[8] Hyver, C. & Le Guyader, H. (1990). MPF and cyclin: modelling of the cell cycle minimum oscillator, *Biosystems*, **24**, 85–90.

[9] Norel, R. & Agur, Z. (1991). A model for the adjustment of the mitotic clock by cyclin and MPF levels, *Science* **251**, 1076–1078.

[10] Obeyesekere, M.N., Hurbert, J.R. & Zimmerman, S.O. (1995). A model of the G1 phase of the cell cycle incorporating cyclin E/cdk2 complex and retinoblastoma protein, *Oncogene* **11**, 1199–1205.

[11] Obeyesekere, M.N., Tucker, S.L. & Zimmerman, S.O. (1992). Mathematical models for the cellular concentrations of cyclin and MPF, *Biochemical and Biophysical Research Communications* **184**, 782–789.

[12] Obeyesekere, M.N., Tucker, S.L. & Zimmerman, S.O. (1994). A model for regulation of the cell cycle incorporating cyclin A, cyclin B, and their complexes, *Cell Proliferation* **27**, 105–113.

[13] Rubinow, S.I. (1978). Age-structured equations in the theory of cell populations, in *Studies in Mathematics, Studies in Mathematical Biology, Part II. Populations and Communities*, Vol. 16, S.A. Levin, ed. The Mathematical Association of America, Washington, pp. 389–410.

[14] Shapiro, H. (1995). *Practical Flow Cytometry*. Wiley Liss, New York.

[15] Swan, G.W. (1977). *Some Current Mathematical Topics in Cancer Research*, University Microfilms International.

[16] Thron, C.D. (1991). Mathematical analysis of a model of the mitotic clock, *Science* **254**, 122–123.

[17] Thron, C.D. (1994). Theoretical dynamics of the cyclin B-MPF system: a possible role for p13suc1, *Biosystems* **32**, 97–109.

[18] Thron, C.D. (1996). A model for the bistable biochemical trigger of mitosis, *Biophysical Chemistry* **57**, 239–251.

[19] Tucker, S.L. & Zimmerman, S.O. (1988). A nonlinear model of population dynamics containing an arbitrary number of continuous structure variables, *SIAM Journal on Applied Mathematics* **48**, 549–591.

[20] Tyson, J.J. (1991). Modeling the cell division cycle: cdc2 and cyclin interactions, *Proceedings of the National Academy of Sciences* **88**, 7328–7332.

[21] Tyson, J.J. and Novak, B. (1993). Numerical analysis of a comprehensive model of M-phase control in Xenopus oocyte extracts and intact embryos, *Journal of Cell Science* **106**, 1153–1168.

[22] Watson, J.V. (1992). *Flow Cytometry Data Analysis: Basic Concepts and Statistics*. Cambridge University Press, New York.

[23] Webb, G.F. (1985). *Theory of Nonlinear Age-Dependent Population Dynamics*. Marcel Dekker, New York.

[24] White, R.A. (1981). A review of some mathematical models in cell kinetics, in *Developments in Cell Biology, Biomathematics and Cell Kinetics*, Vol. 8, M. Rotenberg, ed., Elsevier/North-Holland Biomedical Press, Amsterdam, pp. 243–261.

(*See also* **Mathematical Biology, Overview**)

MANDRI N. OBEYESEKERE

Censored Data

In classical statistics, the observations are frequently assumed to include *independent* random variables $X_1, \ldots, X_n$, with X_i having the density function

$$f_i^\theta(x) = \alpha_i^\theta(x) S_i^\theta(x),$$

where $\alpha_i^\theta(x)$ is the hazard function, $S_i^\theta(x)$ is the survival function, and θ is a vector of unknown parameters (*see* **Survival Distributions and Their Characteristics**). Then inference on θ may be based on the **likelihood** function,

$$L(\theta) = \prod_i f_i^\theta(X_i),$$

in the usual way. In survival analysis, however, one can rarely avoid various kinds of incomplete observation. The most common form of this is *right-censoring* where the observations are

$$(\tilde{X}_i, D_i), \quad i = 1, \ldots, n, \tag{1}$$

where D_i is the indicator $I\{\tilde{X}_i = X_i\}$, and $\tilde{X}_i = X_i$, the true survival time, if the observation of the lifetime of i is uncensored and $\tilde{X}_i = U_i$, the time of right-censoring, otherwise. Thus, $D_i = 1$ indicates an uncensored observation, $D_i = 0$ corresponds to a right-censored observation. Other kinds of incomplete observation will be discussed below.

Survival analysis, then, deals with ways in which inference on θ may be performed based on the censored sample (1). We would like to use the function

$$\begin{aligned} L^c(\theta) &= \prod_i \alpha_i^\theta(\tilde{X}_i)^{D_i} S_i^\theta(\tilde{X}_i) \\ &= \prod_i f_i^\theta(\tilde{X}_i)^{D_i} S_i^\theta(\tilde{X}_i)^{1-D_i} \end{aligned} \tag{2}$$

for inference, but there are two basic problems:

1. The presence of censoring may alter the hazard function of the lifetime X_i, i.e. the conditional distribution of X_i, given that i is alive at t ($X_i \geq t$) *and uncensored at* t ($U_i \geq t$), may be different from what it was in the uncensored case, i.e. just given $X_i \geq t$ (*dependent censoring*).
2. The observed right-censoring times, U_i, may contain information on θ (*informative censoring*).

An example of a dependent censoring scheme would be if, in a clinical trial with survival times as the outcome variables, one removed patients from the study while still alive and when they appeared to be particularly ill (or particularly well), so that patients remaining at risk are not representative of the group that would have been observed in the absence of censoring. In other words, dependent censoring represents a dynamic version of what in an epidemiologic context would be termed a **selection bias**. An example is provided below (Example 1). Mathematical formulations of independent censoring (conditions on the joint distribution of X_i and U_i) may be given, and it may be shown that several frequently used models for the generation of the times of right-censoring satisfy these conditions. The difficulty in a given practical context lies in the fact that the conditions may be impossible to verify, since they refer to quite hypothetical situations.

The second concept mentioned, noninformative censoring, is simpler and relates to the fact that if censoring is informative, then a more efficient inference on θ may be obtained than the one based on (2); see below.

Independent Censoring

The general definition of independent censoring given by Andersen et al. [2], Section III.2.2 for multivariate counting processes has the following interpretation for the special case of survival analysis with time-fixed **covariates**. The basic (uncensored) model is that conditional on covariates $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ the lifetimes $X_1, \ldots, X_n$ are independent, X_i having the hazard function

$$\alpha_i^\theta(t|\mathbf{Z}_i) \approx P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | X_i \geq t, \mathbf{Z}) / \mathrm{d}t. \tag{3}$$

Here, $I_{\mathrm{d}t}$ is the interval $[t, t + \mathrm{d}t)$ and $P^{\theta\phi}$ is the joint distribution of $X_1, \ldots, X_n, \mathbf{Z}$ *and* the censoring times. Note that the hazard function only depends on θ, i.e. ϕ is a nuisance parameter. Because of the conditional independence of X_i it follows that

$$P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | \mathcal{F}_{t-}) \approx \alpha_i^\theta(t|\mathbf{Z}_i) I\{X_i \geq t\} \, \mathrm{d}t,$$

where the *history* $\mathcal{F}_{t-}$ contains $\mathbf{Z}$ and all information on $X_1, \ldots, X_n$ from the interval $[0, t)$, i.e. values of X_i for i with $X_i < t$ and the information that $X_j \geq t$

for j with $X_j \geq t$. Let there now be given right-censoring times $U_1, \ldots, U_n$ and define the enlarged history $\mathcal{G}_t$ as the one containing $\mathcal{F}_t$ *and* all information on $U_1, \ldots, U_n$ from the interval $[0, t]$, i.e. values of $U_i \leq t$ and the information that $U_j \geq t$ for those j where $U_j \geq t$. The condition for independent censoring is then that

$$P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | \mathcal{F}_{t-}) = P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | \mathcal{G}_{t-}). \qquad (4)$$

It follows that *simple type I* censoring, where all U_i are equal to a *fixed* time, u_0, and *simple type II* censoring, where all U_i are equal to the kth smallest lifetime $X_{(k)}$ for some k between 1 and n, are both independent, since the right-censoring times in these cases give rise to no extra randomness in the model; that is, $\mathcal{F}_t = \mathcal{G}_t$.

In some models, $U_1, \ldots, U_n$ are assumed to be independent given $\mathbf{Z}$ and $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent identically distributed (iid). Then the assumption (4) reduces to

$$\alpha_i^{\theta}(t|\mathbf{Z}_i) \approx P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | X_i \geq t, U_i \geq t, \mathbf{Z})/\,\mathrm{d}t \qquad (5)$$

and it is fulfilled, e.g. if U_i and X_i are independent given Z_i. This is, for instance, the case in the *simple random* censorship model where $U_1, \ldots, U_n$ are iid and independent of $X_1, \ldots, X_n$.

Some authors take the condition (5) (which is less restrictive than (4)) as the definition of independent censoring; see, for example, [6], p. 128. However, (4) may be generalized to other models based on counting processes and both (4) and (5) cover the most frequently used mathematical models for the right-censoring mechanisms. These include both the models already mentioned, i.e. simple type I, type II and random censorship and various generalizations of these (e.g. progressive type I censorship (cf. Example 2, below), general random censorship, and randomized progressive type II censorship; see, [2, Section III.2.2]). Earlier contributions to the definition and discussion of independent censoring are the monographs by Kalbfleisch & Prentice [13], p. 120 and Gill [7], Theorem 3.1.1 and the papers by Cox [5], Williams & Lagakos [16], Kalbfleisch & MacKay [12] and Arjas & Haara [3], all of whom give definitions that are close or equivalent to (5). Another condition for independent censoring, stronger than (5) but different from (4), is discussed by Jacobsen [11].

From (4) and (5) it is seen that censoring is allowed to depend on covariates as long as these are included in the model for the hazard function of the lifetime distribution in (3). Thus, an example of a *dependent* censoring scheme is one where the distribution of U_i depends on some covariates that are not included there. This is illustrated in the following example.

Example 1: Censoring Depending on Covariates

Suppose that iid binary covariates, $Z_1, \ldots, Z_n$, have

$$P^{\theta\phi}(Z_i = 1) = 1 - P^{\theta\phi}(Z_i = 0) = \phi,$$

and that $X_1, \ldots, X_n$ are iid with survival function $S(t)$. The **Kaplan–Meier estimator** $\widehat{S(t)}$ based on the X_i then provides a consistent estimate of $\theta = S(\cdot)$, the marginal distribution of X_i. This may be written as

$$S(t) = \phi S_1(t) + (1 - \phi) S_0(t),$$

where $S_j(t)$, for $j = 0, 1$, is the conditional distribution given $Z_i = j$. Note that these may be different, e.g. $S_1(t) < S_0(t)$ if individuals with $Z_i = 1$ are at higher risk than those with $Z_i = 0$. Define now the right-censoring times U_i by

$$U_i = u_0, \text{ if } Z_i = 1, \qquad U_i = +\infty, \text{ if } Z_i = 0.$$

Then, for $t < u_0$ the Kaplan–Meier estimator will still consistently estimate $S(t)$, while for $t > u_0$, $\widehat{S(t)}/\widehat{S(u_0)}$ will estimate $S_0(t)/S_0(u_0)$. If, however, the covariate is included in the model for the distribution of X_i, i.e. $\theta = [S_0(\cdot), S_1(\cdot)]$, then $\widehat{S_j(t)}$, the Kaplan–Meier estimator based on individuals with $Z_i = j$, $j = 0, 1$, will consistently estimate the corresponding $S_j(t)$, also based on the right-censored sample (though, of course, no information will be provided about $S_1(t)$ for $t > u_0$).

It is seen that censoring is allowed to depend on the *past* and on external (in the sense of conditionally independent) random variation. This means that if, in a lifetime study, sex and age are included as covariates, then a right-censoring scheme, where, say, every year, one out of the two oldest women still alive and uncensored is randomly (e.g. by flipping a coin) chosen to be censored, is independent. However, a right-censoring scheme depending on the *future* is dependent. This is illustrated in the following example.

Example 2: Censoring Depending on the Future

Suppose that, in a clinical trial, patients are accrued at calendar times $T_1, \ldots, T_n$ and that they have iid lifetimes $X_1, \ldots, X_n$ (since entry) independent of the entry times. The study is terminated at calendar time t_0 and the entry times are included in the observed history, i.e. $Z_i = T_i$ in the above notation. If, at t_0, all patients are traced and those still alive are censored (at times $U_i = t_0 - T_i$) and, for those who have died, their respective lifetimes, X_i, are recorded, then this right-censoring is independent (being *deterministic*, given the entry times, so-called progressive type I censoring).

Consider now, instead, the situation where patients are only seen, for instance, every year, i.e. at times $T_i + 1, \ldots, T_i + k_i \leq t_0$ and suppose that if a patient does not show up at a scheduled follow-up time, then this is because he or she has died since last follow-up and the survival time is obtained. Suppose, further, that for the patients who are alive at the time, $T_i + k_i$, of their last scheduled follow-up, and who die before time t_0, there is a certain probability, ϕ, of obtaining information on the failure, whereas for those who survive past t_0 nothing new is learnt. If these extra survival times are included in the analysis and if everyone else is censored at k_i, then the right-censoring scheme is dependent. This is because the fact that patient i is censored at k_i tells the investigator that this patient is likely not to die before t_0 and the right-censoring, therefore, depends on the future. To be precise, if the average probability of surviving past t_0, given survival until the last scheduled follow-up time is $1 - \pi$, then the probability of surviving past t_0, given censoring at the time of the last scheduled follow-up, is $(1 - \pi)/[\pi(1 - \phi) + 1 - \pi]$, which is 1 if $\phi = 1, 1 - \pi$ if $\phi = 0$, and between $1 - \pi$ and 1, otherwise.

If, alternatively, everyone still alive at time $T_i + k_i$ were censored at k_i, then the censoring would be independent (again being deterministic given the entry times).

Another censoring scheme that may depend on the future relative to "time on study", but not relative to calendar time, occurs in connection with testing with replacement, see, for example, [8].

Let us finally in this section discuss the relation between independent right-censoring and **competing risks**. A competing risks model with two causes of failure, d and c, is an inhomogeneous **Markov process** $W(\cdot)$ with a transient state 0 ("alive"), two absorbing states d and c and two cause-specific hazard functions $\alpha_{0d}(t)$ and $\alpha_{0c}(t)$, e.g. Andersen et al. [1]. This generates two random variables:

$$X = \inf[t : W(t) = d]$$

and

$$U = \inf[t : W(t) = c],$$

which are incompletely observed since the observations consist of the transition time $\widetilde{X} = X \wedge U$ and the state $W(\widetilde{X}) = d$ or c reached at that time. The elusive concept of "independent competing risks" (e.g. [13, Section 7.2]) now states that in a population where the risk c is not operating, the hazard function for d is still given by $\alpha_{0d}(t)$. This condition is seen to be equivalent to censoring by U being independent. However, since the population where a given cause of failure is eliminated is usually completely hypothetical in a biological context, this formal equivalence between the two concepts is of little help in a practical situation and, as is well known from the competing risks literature (e.g. [4, 15], and [13, Chapter 7]), statistical independence of the random variables X and U cannot be tested from the incomplete observations $[\widetilde{X}, W(\widetilde{X})]$. What can be said about the inference on the parameter $\theta = \alpha_{0d}(\cdot)$ based on these data is that consistent estimation of θ may be obtained by formally treating failures from cause c as right-censorings, but that this parameter has no interpretation as the d failure rate one would have had in the hypothetical situation where the cause c did not operate.

For the concept of independent censoring to make sense, the "uncensored experiment" described in the beginning of this section should, therefore, be meaningful.

Likelihoods: Noninformative Censoring

The right-censored data will usually consist of

$$(\widetilde{X}_i, D_i, \mathbf{Z}_i; i = 1, \ldots, n)$$

and, under independent censoring, the likelihood can then be written using **product-integral** notation

$$L(\theta,\phi) = P^{\theta\phi}(\mathbf{Z}) \prod_i \prod_{t>0} \alpha_i^{\theta}(t)^{D_i(\mathrm{d}t)} [1 - \alpha_i^{\theta}(t)\mathrm{d}t]^{1 - D_i(\mathrm{d}t)} \times \gamma_i^{\theta\phi}(t)^{C_i(\mathrm{d}t)} [1 - \gamma_i^{\theta\phi}(t)\,\mathrm{d}t]^{1 - C_i(\mathrm{d}t)}. \quad (6)$$

Here, $D_i(\mathrm{d}t) = I\{X_i \in I_{\mathrm{d}t}\}$, $C_i(\mathrm{d}t) = I\{U_i \in I_{\mathrm{d}t}\}$, and $\alpha_i^\theta(t)$ and $\gamma_i^{\theta\phi}(t)$ are the conditional hazards of failure and censoring, respectively, given the past up until t- (including covariates). The likelihood (6) may be written as

$$L(\theta, \phi) = L^c(\theta)L^*(\theta, \phi),$$

with $L^c(\theta)$ given by (2) and where the contributions from censoring and covariates are collected in $L^*(\theta, \phi)$. Thus, the function (2), which is usually taken as the standard censored data likelihood, is, under independent censoring, a **partial likelihood** on which a valid inference on θ may be based. It is only the full likelihood for θ if $L^*(\theta, \phi)$ does not depend on θ, which is the case if censoring (and covariates) are *noninformative*. Thus, noninformative censoring is a statistical concept (while the concept of independent censoring is *probabilistic*) and means that the conditional hazard of censoring $\gamma_i^{\theta\phi}(t)$ does, in fact, not depend on θ, the parameter of interest.

An example of an informative right-censoring scheme could be in a study with two competing causes of failure and where only one of the two cause-specific failure rates is of interest; if the two cause-specific failure rates are *proportional* (as in the so-called Koziol–Green model for random censoring, [14]), then the failures from the second cause (the censorings) will carry information on the shape of the hazard function for the failure type of interest. It is, however, important to notice that even if the censoring is informative, then inference based on (2) will still be valid (though not fully efficient) and as it is usually preferable to make as few assumptions as possible about the distribution of the right-censoring times, the (partial) likelihood (2) is often the proper function to use for inference.

Other Kinds of Incomplete Observation

When observation of a survival time, X, is right-censored, then the value of X is only known to belong to an interval of the form $[U, +\infty)$. This is by far the most important kind of censoring for survival data, but not the only one. Thus, the observation of X is **interval-censored** if the value of X is only known to belong to an interval $[U, V)$ and it is said to be *left-censored* if $U = 0$.

It was seen above that under independent right-censoring a right-censored observation, U_i, contributed to the partial likelihood function with a factor $S^\theta(U_i)$, which was also the contribution to the *full* likelihood under noninformative censoring. Similarly, concepts of independent and noninformative interval-censoring may be defined as leading to a contribution of $S^\theta(U_i) - S^\theta(V_i)$ to, respectively, the partial and the full likelihood. These concepts have received relatively little attention in the literature; however, this way of viewing censoring is closely related to the concept of **coarsening at random**.

Formally, **grouped data**, where for each individual the lifetime is known only to belong to one of a fixed set of intervals $[u_{k-1}, u_k)$ with $0 = u_0 < u_1 < \cdots < u_m = +\infty$, are also interval-censored. However, the fact that the intervals are the same for everyone simplifies the likelihood to a binomial-type likelihood with parameters $p_k^\theta = S^\theta(u_{k-1}) - S^\theta(u_k)$, $k = 1, \ldots, m$.

Let us finally remark that while, following Hald [9; 10, p. 144], *censoring* occurs when we are able to sample a complete population but individual values of observations above (or below) a given value are not specified, truncation corresponds to sampling from an incomplete population, i.e. from a conditional distribution (*see* **Truncated Survival Times**). Left-truncated samples, where an individual is included only if his or her lifetime exceeds some given lower limit, also occur frequently in the analysis of survival data, especially in epidemiologic studies where hazard rates are often modeled as a function of age and where individuals are followed only from age at diagnosis of a given disease or from age at employment in a given factory.

References

[1] Andersen, P.K., Abildstrom, S. & Rosthøj, S. (2002). Competing risks as a multistate model. *Statistical Methods in Medical Research*. **11**, 203–215.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[3] Arjas, E. & Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates, *Scandinavian Journal of Statistics* **11**, 193–209.

[4] Cox, D.R. (1959). The analysis of exponentially distributed life-times with two types of failure, *Journal of the Royal Statistical Society, Series B* **21**, 411–421.

[5] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[6] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[7] Gill, R.D. (1980). Censoring and stochastic integrals, *Mathematical Centre Tracts* **124**, Mathematisch Centrum, Amsterdam.

[8] Gill, R.D. (1981). Testing with replacement and the product limit estimator, *Annals of Statistics* **9**, 853–860.

[9] Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Skandinavisk Aktuarietidsskrift* **32**, 119–134.

[10] Hald, A. (1952). *Statistical Theory with Engineering Applications*. Wiley, New York.

[11] Jacobsen, M. (1989). Right censoring and martingale methods for failure time data, *Annals of Statistics* **17**, 1133–1156.

[12] Kalbfleisch, J.D. & MacKay, R.J. (1979). On constant-sum models for censored survival data, *Biometrika* **66**, 87–90.

[13] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[14] Koziol, J.A. & Green, S.B. (1976). A Cramér-von Mises statistic for randomly censored data, *Biometrika* **63**, 465–474.

[15] Tsiatis, A.A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences* **72**, 20–22.

[16] Williams, J.A. & Lagakos, S.W. (1977). Models for censored survival analysis: constant sum and variable sum models, *Biometrika* **64**, 215–224.

(*See also* **Survival Analysis, Overview**)

PER KRAGH ANDERSEN

Censuses

According to a United Nations Population and Housing Census manual, "The fundamental purpose of the population census is to provide the facts essential to governmental policy-making, planning and administration" [3]. The manual goes on to state that "Population census results are also used in policy development and in management of national evaluation for programmes in such fields as ... maternal and child health, ... and welfare." Thus, the content of population censuses is designed primarily to serve official administrative functions of government, such as taxation, political representation, conscription, and revenue and resource allocation. While censuses may also provide invaluable information for planning and evaluation of social, health, and welfare programs, and description and monitoring of population composition and trends, crucial both to practical problem solving and for pure research, these applications are of secondary priority in determining census content.

To meet its basic objectives, national population censuses contain a "minimum data set" of demographic variables, which invariably includes age, sex, marital status, and geographic residence. While specific content varies widely from country to country, most national population censuses also include data on socioeconomic status, education, occupation, industry, economic activity, housing conditions, and, less frequently, health and/or disability status. Censuses in some countries may also include "cultural status" measures – for example, race, ethnicity, religion, language, or national origin.

The principal use of population censuses in epidemiologic research is in the construction of measures of **risk**. Risk is one of the most basic concepts in the study of epidemiology and is central to the study of the distribution, incidence, prevalence (*see* **Descriptive Epidemiology**), or transmission of disease, adverse health conditions, or outcomes. The risks of becoming ill, being injured, disabled, or of dying are typical examples. Formally, risk is defined as the probability that an event will occur within a specified period of time, as measured either by the calendar or by biological age [2]. Mathematically, risk is calculated as the number of events, or outcomes, occurring in the specified time period divided by the number of persons at risk at the *beginning* of that period. A widely used proxy measure of risk relates the same events to the number of **person-years at risk** among the risk group during the specified time interval. A typical example is the age-specific mortality rate, defined as the ratio of deaths to a specified age group in a specified time period, to the midyear population of that age. In turn, a midyear population of the time interval in question has been shown to be a rather accurate estimate of the person-years lived in the interval (*see* **Vital Statistics, Overview**). Various models have been developed by statisticians and demographers to link directly the true probability rate (risk) to the mortality rate. The approach is also applied to other health outcomes or status changes.

Not all members of a population are equally exposed to or are susceptible to disease, accident, death, or other changes in health status. To understand the epidemiology of adverse health conditions, or outcomes, or to design effective and efficient public health intervention programs, it is useful to identify and focus on those subsets of the population at elevated risk. For example, to study **maternal mortality**, the risk group (denominator) would not be the population at large, but would be narrowed to *females of child-bearing age*. Similarly, the study of prostate cancer mortality or morbidity would use *males at middle and old age*, perhaps separated into five-year age groups as the denominator in calculating risk. Studies of morbidity and mortality risk due to natural disasters – earthquakes, hurricanes, etc. – would limit the risk group to geographic localities in which such events are most likely to occur.

Data for risk numerators, i.e. the events, health outcomes, or health status changes, may be obtained from vital statistics (birth and death records), from special **disease registers** (cancer and congenital malformation registries), **surveillance** reports (notifications of "reportable diseases"), or from special surveys (national health surveys) (*see* **Surveys, Health and Morbidity**). In epidemiologic studies of national populations (as opposed to clinical studies), population censuses are most often the source of data for risk denominators.

As valuable as these variables from population censuses may be in differentiating risk categories in the population, decennial censuses by themselves are of limited value due to the relatively long intercensal periods. That is, any one census provides risk denominator data only every 10 years (or five years, as the case may be). Typically, censuses are

conducted every 10 years in years ending in "0", although nations of the British Commonwealth and some former colonies conduct their censuses in years ending in "1". Also, a small number of countries conduct their censuses every five years. Four methods have been developed to provide data between two adjacent censuses (already completed) or from the last census and until the next census results become available.

The first approach, to provide information between two completed censuses, involves some form of interpolation. In its simplest form the size of various population aggregates in intercensal years are assumed to change linearly between censuses, and annual estimates are calculated accordingly. More sophisticated patterns of change (logarithmic, for example) may be used in lieu of the linear assumption. However sophisticated the assumption of the pattern of intercensal change in the population risk groups, a serious limitation of this approach is that it cannot replicate irregular patterns, such as those occasioned by abrupt changes in natality or mortality patterns, in migration, or results of some sudden economic changes. Another limitation of interpolation methods is that estimates of the subpopulations generally do not add up to the totals in the overall population.

A second approach utilizing vital statistics, which are events recorded on an ongoing basis, may be used to refine intercensal interpolations and postcensal extrapolations, at least as far as irregular patterns of natality, mortality, or marriage and divorce are concerned. While vital statistics are useful in refining the denominator information, they cannot account for irregular population shifts due to changing migration patterns, or to social and economic conditions affecting the characteristics of the population.

Because of these limitations in the interpolation and extrapolation methods, some countries have instituted annual "mini" censuses – annual probability sample surveys of the total population. These surveys typically include basic demographic variables and a limited collection of socioeconomic variables. The "Current Population Survey" in the US is one such example [4].

An alternative to the intercensal sample population surveys is to use information from national population registers, such as those in use in Scandinavia, Japan, and elsewhere. These registers are, in fact, dynamic censuses in which changes in vitality, marital status, residence, and socioeconomic condition are recorded as they occur to each individual in the population. In practice, however, there are relatively few such systems of sufficient quality to warrant widespread use for epidemiologic research.

Population censuses provide an extremely important source of information identifying risk groups in a general population. While they may contain limited, or even no, specific health information, the basic demographic variables they do contain can be used very effectively to differentiate levels of risk. Age, sex, marital status, socioeconomic level, and place of residence are obvious examples generally strongly associated with risk levels. Census population data cross-classified by these variables provide direct estimates of the risk denominators, either the person-years lived or the bases for reconstructing the population initially at risk.

Other applications of censuses for epidemiologic research include censuses of housing and the preparation of special linkage studies. In many countries, a census of housing is conducted in conjunction with the census of population. This provides the opportunity to use housing, residential neighborhood, and related information to focus more precisely on populations at elevated health risks due to their living conditions and/or locations. The other approach is to link vital events, or other health-related events obtained from registers or special surveys, directly to individuals or households enumerated in the population census. While this approach may provide a particularly rich source of data for detailed epidemiologic study, it is expensive to carry out, technically difficult to accomplish in a rigorous and accurate manner, and may conflict with laws concerning privacy and **confidentiality** of personal information. Perhaps the classic example of this technique is the British linkage study utilizing the 1961 census [1] (*see* **Record Linkage**).

Thus, the population census is an invaluable tool for epidemiologists and biostatisticians. In spite of limitations in the coverage of variables, periodicity, and other problems cited above, census data are almost universally available and are easily adapted to identify a large number of highly differentiated population subgroups at elevated exposure and susceptibility to disease, injury, or death.

References

[1] Fox, A.J. & Goldblatt, P.O. (1982). *Socio-demographic Mortality Differentials*, OPCS Longitudinal Study, OPCS Series LS, No. 1. HMSO, London.

[2] Last, J.M. ed. (1983). *A Dictionary of Epidemiology*. Oxford University Press, Oxford.

[3] United Nations. (1996). *Principles and Recommendations for Population and Housing Census*, ST/ESA/STAT/Series/M/Rev.1 (in draft). United Nations, New York, p. 6.

[4] US Bureau of the Census. (1978). The Current Population Survey, Design and Methodology, Technical Paper 40, Government Printing Office, Washington.

R. HARTFORD

Centers for Disease Control and Prevention (CDC)

The Centers for Disease Control and Prevention (CDC) celebrated its 50th anniversary on July 1, 1996. The CDC evolved from a small mosquito-eradication effort in World War II to being the USA's primary agency in the promotion of health and the prevention of disease. This agency, the Malaria Control in War Areas (MCWA), was established in 1942 and was a national effort to keep military bases and essential war industry-related establishments in the southern US free from malaria. In 1946, MCWA became the Communicable Disease Center and was the agency of the US Public Health Service that directed efforts to prevent diseases such as malaria, polio, smallpox, toxic shock syndrome, Legionnaires' disease, and more recently, **AIDS**, Ebola virus, Hantavirus, monkeypox, and SARS. In the early days most efforts focused on prevention and control of unnecessary morbidity and mortality from infectious diseases of public health importance. Over the years these responsibilities have expanded to include contemporary threats to health, such as lead-paint poisoning, environmental and occupational hazards (e.g. pesticides, chemical warfare agents or hazards in the workplace), (*see* **Environmental Epidemiology**; **Occupational Epidemiology**); behavioral risks (smoking); the prevention of chronic diseases and injuries; and the promotion of healthy behavior, prenatal care, immunizations, and upgrading state and local public health agencies' readiness to infectious disease outbreaks and bioterrorism threats [41, 43]. As it expanded its responsibilities it has also changed its name. In 1967 it became the National Communicable Disease Center; in 1970, the Center for Disease Control; in 1980, the Centers of Disease Control; and in 1993, the Centers for Disease Control and Prevention. In 2003, CDC's sister agency, The Agency for Toxic Substances and Disease Registry (ATSDR) and CDC's National Center for Environmental Health merged and CDC's environmental health activities include ATSDR [41].

Today the CDC's mission is to promote health and **quality of life** by preventing and controlling disease, injury, and disability [40, 41, 43]. To accomplish this mission the CDC works with others throughout the US and world to monitor health; detect and investigate health problems; conduct research to enhance disease prevention; develop and advocate sound public health policies; implement prevention strategies; promote healthy behaviors; foster safe and healthy environments; and provide leadership and training. The CDC works in partnership with other agencies within the Department of Health and Human Services and other agencies in the US government, with state and local health departments, academic institutions, professional, voluntary, and community organizations, philanthropic foundations, school systems, churches, and other local institutions, industry, and labor [8, 40, 41]. In 2003, it had an annual budget of seven billion dollars and 9400 employees in 170 occupations and more that 5000 contractors in many locations, including field stations, states, and countries [41]. While the line between the responsibilities of the CDC and the **National Institutes of Health** (NIH) is not clearly defined, it was initially agreed that the NIH would focus more on basic research and the CDC would help the states recognize and control communicable diseases.

Statisticians have played an important role at the CDC. In the 1950s there were two groups of statisticians at the CDC, one in the Epidemiology Branch and the other in the Venereal Disease Branch. Later, in 1960, a third group in the Tuberculosis Branch located in Washington, D.C, moved to Atlanta [30]. Robert Serfling and Ida Sherman developed and used morbidity and mortality systems to monitor disease trends and to estimate excess mortality in influenza epidemics [30, 38]. These systems, and many other surveillance systems, have been developed and are being used to monitor diseases, behavioral risk factors, and traumatic occupational fatalities (*see* **Surveillance of Diseases**). Serfling and others modified and used **quota sampling** methods to study the distributions of polio cases [36, 37]. These studies showed that the highest proportion of paralytic cases were concentrated among young children living in depressed conditions. In some cities the central sections of cities with large, overcrowded black populations had the most cases. In some cases the populations were mostly white but the paralytic cases were concentrated among very poor young children. Based on these results of the surveys it was clear that the low levels of vaccination were a problem, providing the necessary justification and stimulus to pursue a much needed and vigorous vaccination program.

In the area of Venereal Diseases, Lida Usilton established a mechanical system for surveillance of venereal disease cases, their treatment progress, and follow-up and treatment of contacts [41]. **Life table** methodology was used to evaluate rapid antisyphilitic therapy and work was done to evaluate the effectiveness of syphilis contact investigations [19]. By the 1970s, mathematical modeling (*see* **Model, Choice of**), vaccine (*see* **Vaccine Studies**) and other **clinical trials**, **program evaluations**, cost–effectiveness and cost–benefit analyses (*see* **Health Economics**) and **time series** models were carried out in the Divisions of Sexually Transmitted Diseases and Tuberculosis [15, 22, 27, 29, 31–33]. Evaluations were done to compare the effectiveness of different gonorrhea control strategies. In the area of hospital-acquired infections, statisticians played a major role. The Study on the Efficacy of Nosocomial Infection Control (SENIC), a nationwide study of the effectiveness of programs to control hospital-acquired infections, was a large and sophisticated medical survey [14–17, 27, 28, 44]. Agent Orange and injury control were other areas where statisticians have worked to identify high-risk groups [3–7, 9, 23, 25]. Statisticians have played a major role in AIDS and sexually transmitted diseases (STDs) research [12, 13, 18, 20, 23, 24, 34]. In surveillance of diseases and in trend analysis, statisticians have made continuing contributions [2, 21, 26, 35, 39, 45]. In the Agency for Toxic Substances and Disease Registry, a National Exposure Registry (a listing of persons with documented environmental exposures) has been established [1, 10, 11] (*see* **Disease Registers**). The rate of reporting of adverse health outcomes in this registry is compared with national norms to assess impact of exposure.

The CDC now has statistics branches (or statisticians) in most areas of disease control and prevention working in the areas of environmental exposure and health, occupational safety and health, cancer, heart disease, sexually transmitted diseases, AIDS, tuberculosis, diabetes, hospital infections, birth defects, reproductive health, infectious diseases, bacterial and mycotic diseases, hospital infections, environmental exposures and hazards, immunizations, injury, genetics, prevention research, and so on. The statistical techniques used to analyze such data are varied, for example, **multivariate analysis**, **sequential analysis**, **categorical data analysis**, **stochastic processes**, **survival analysis**, time series analysis, **generalized linear models**, **decision analysis**, sample surveys, and estimation of **sensitivity** and **specificity** of **screening** and **diagnostic tests**.

Perhaps the biggest stimulus to increased use of data in monitoring morbidity, mortality, risk factors, costs of health care, quality of life, and other statistical information to guide policies to improve the health of the American people was when the **National Center for Health Statistics (NCHS)** became a part of the CDC in 1987. The NCHS is the principal health statistics agency of the US.

The CDC has developed and recommended many prevention and intervention strategies. Evaluation of the efficacy of these public health programs is essential before implementing prevention or intervention strategies on a broad basis. It is almost always more difficult to assess the long-term health effects of environmental and occupational hazards, or to determine the long-term effects of smoking (*see* **Smoking and Health**), lack of exercise, stress, or workplace hazards (often requiring years of study), than to design and implement programs testing drug or vaccine effectiveness. The demonstration of the efficacy of these prevention or intervention strategies is complex and often confounded by many uncontrollable, external factors, but this demonstration is essential before implementing them on a broad basis.

There are usually two levels of evaluation which require different study designs and analyses. The first level addresses control or prevention of disease in individuals, while the second concerns prevention or control in the general population. At the first level we may need a **clinical trial** to determine the efficacy of a drug, vaccine, or other intervention at the individual level. The second level is to show that we can control or prevent disease, alcohol and other drug misuse, or other risk behavior in a community.

This second stage of evaluation is both more complex and costly. It requires:

1. attention to sampling methodology and the possible **biases** introduced by the method of sampling, e.g. **telephone sampling**;
2. attention to the evaluation of diagnostic procedures and tests, i.e. that the tests are both sensitive and specific and that the evaluations are properly evaluated in different subgroups of the population;
3. that the intervention does not affect the surveillance system artifactually;

4. assurance that the surveillance system is impartial and **unbiased** in reaching different subgroups of the population;
5. knowledge of those intervention or prevention strategies appropriate for different subgroups at risk;
6. surveillance measures culturally appropriate to the subgroups of the population at risk;
7. that the community that receives the intervention must trust those who deliver the intervention; and
8. that the community must be involved in all phases of the intervention/prevention effort.

These phases include needs assessments of the community, setting objectives, developing, planning and implementing the intervention, including deciding what and how questions are asked (*see* **Questionnaire Design**) collecting and managing data (*see* **Data Management and Coordination**), and analysis and interpretation of results.

Thus, the different areas of disease control and prevention have provided statisticians with many opportunities for innovative statistical application and development to many different public health programs at CDC. For more information about CDC, visit the Web site [43].

References

[1] Burg, J.R. (1989). Policies and procedures for establishing the National Exposure Registry, *Journal of the American College of Toxicologists* **8**, 949–954.

[2] Caudill, S.P., Smith, S.J. & Cooper, G.R. (1989). Cholesterol-based personal risk assessment in coronary heart disease, *Statistics in Medicine* **12**, 295–309.

[3] Centers for Disease Control Vietnam Experience Study (1987). Post service mortality among Vietnam Veterans, *Journal of the American Medical Association* **257**, 790–795.

[4] Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: I. Psychosocial characteristics, *Journal of the American Medical Association* **259**, 2701–2707.

[5] Centers for Disease Control Vietnam Experience Study (1988). Health status of Vietnam Veterans: II. Physical health, *Journal of the American Medical Association* **259**, 2708–2714.

[6] Centers for Disease Control Vietnam Experience Study (1988). Health status of Vietnam Veterans: III. Reproductive outcomes and child health, *Journal of the American Medical Association* **259**, 2715–2719.

[7] Conn, J.M., Lui, K.-J. & McGee, D.L. (1989). A model-based approach to the imputation of missing data: home incidences, *Statistics in Medicine* **8**, 263–266.

[8] Etheridge, E.W. (1992). *Sentinel for Health, A History of the Centers for Disease Control*. University of California Press, Berkeley.

[9] Flanders, W.D. & Rhodes, P.H. (1987). Large sample confidence limits for regression standardized risks, risk ratios, and risk differences, *Journal of Chronic Diseases* **40**, 697–704.

[10] Gist, G.L. & Burg, J.R. (1995). Methodology for selecting substances for the National Exposure Registry, *Journal of Exposure Analysis and Environmental Epidemiology* **5**, 197–208.

[11] Gist, G.L., Burg, J.R. & Radtke, T.M. (1994). The site selection process for the National Exposure Registry, *Journal of Environmental Health* **56**, 7–12.

[12] Hadgu, A. (1996). Discrepant analysis and screening of *Chlamydia trachomatis*, *Lancet* **348**, 1309.

[13] Hadgu, A. (1996). The discrepancy in discrepant analysis, *Lancet*, **348**, 592–593.

[14] Haley, R.W., Hooton, T.M., Schoenfelder, J.R., Crossley, K.B., Quade, D., Stanley, R.C. & Culver, D.H. (1980). Effect of an infection surveillance and control program on the accuracy of retrospective chart review, *American Journal of Epidemiology* **111**, 543–555.

[15] Haley, R.W., Quade, D., Freeman, H.E., Bennett, J.V. & CDC SENIC Planning Committee (1980). Study in the efficacy of nosocomial infection control (SENIC Project): summary of study design, *American Journal of Epidemiology* **111**, 472–485.

[16] Haley, R.W., Schaberg, D.R., McClish, D.K., Quade, D., Crossley, K.B., Culver, D.H., Morgan, W.M., McGowan, J.E. Jr & Schachtman, R.H. (1980). The accuracy of retrospective chart review on measuring nosocomial infection rates: results of validation studies in pilot hospitals, *American Journal of Epidemiology* **111**, 516–533.

[17] Hooton, T.M., Haley, R.W. & Culver, D.H. (1980). A method of classifying patients according to the nosocomial infection risks associated with diagnoses and surgical procedures, *American Journal of Epidemiology* **111**, 556–573.

[18] Huiman, X.B., Caldwell, M.B., Thomas, P., Mascola, L., Ortiz, I., Hsu, H., Schulte, J., Parrott, R., Maldonado, Y., Byers, R. & the Pediatric Spectrum of Disease Clinical Consortium (1996). Natural history of the Human Immunodeficiency Virus disease in perinatally infected children: an analysis from the Pediatric Spectrum of Disease Project, *Pediatrics* **97**, 710–716.

[19] Iskrant, A.P., Bowman, R.W. & Donohue, J.F. (1948). Techniques in evaluation of rapid antisyphilitic therapy, *Public Health Reports* **63**, 965–977.

[20] Kafadar, K. & Karon, J.M. (1989). An analysis of AIDS incidence data by clustering trends, *Statistics in Medicine* **8**, 263–266.

[21] Katzoff, M. (1989). The application of time series forecasting methods to an estimation of problems using provisional mortality statistics, *Statistics in Medicine* **12**, 335–341.

[22] Kramer, M. & Reynolds, G.H. (1982). Evaluation of a gonorrhea vaccine other than simulation modeling,

in *Differential Equations and Applications in Ecology, Epidemics, and Population Problems*, S. Busenberg, & K.L. Cooke, eds. Academic Press, New York, pp. 97–113.

[23] Lui, K.-J., Lawrence, D.N., Morgan, W.M., Peterman, T.A., Haverkos, H.W. & Breg man, D.J. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome, *Proceedings of the National Academy of Sciences* **83**, 3051–3055.

[24] Lui, K.-J., Peterman, T.A., Lawrence, D.N. & Allen, J.R. (1988). A model based approach to characterize the incubation period of pediatric transfusion-associated acquired immunodeficiency syndrome, *Statistics in Medicine* **7**, 395–401.

[25] McGee, D.L. & Rhodes, P. (1989). Estimating trends in the effectiveness of seat belts in saving lives, *Statistics in Medicine* **8**, 379–385.

[26] Parker, R.A., (1989). Analysis of surveillance data with Poisson regression: a case study, *Statistics in Medicine* **12**, 285–294.

[27] Quade, D., Culver, D.H., Haley, R.W., Whaley, F.S., Kalsbeek, W.D., Hardison, C.D., Johnson, R.E., Stanley, R.C. & Shachtman, R.H. (1980). The SENIC sampling process: design for choosing hospitals and patients and results of sample selection, *American Journal of Epidemiology* **111**, 486–502.

[28] Quade, D., Lachenbruch, P.A., Whaley, F.S., McClish, D.K. & Haley, R.W. (1980). Effects of misclassifications on statistical inferences in epidemiology, *American Journal of Epidemiology* **111**, 503–515.

[29] Reynolds, G.H. (1973). A Control Model for Gonorrhea, *Ph.D. Dissertation* Emory University, University Microfilms.

[30] Reynolds, G.H. (1989). Closing remarks, *Statistics in Medicine* **8**, 397–400.

[31] Reynolds, G.H. & Chan, Y.K. (1975). A control model for gonorrhea, *Proceedings of the International Statistical Institute* **XLVI**, (Book 2), 256–279.

[32] Reynolds, G.H. & Colton, T. (1993). Editorial, *Statistics in Medicine* **12**, 191–192.

[33] Reynolds, G.H., Zaidi, A. & Kramer, M.A. (1983). Evaluation of gonorrhea control strategies using a computer simulation model, *Proceedings of the International Statistical Institute, Contributed Papers* **1**, 247–251.

[34] Sattan, G. & Longini, I. (1996). Markov Chains with measurement error: estimating the true course of a marker of the progression of Human Immunodeficiency Virus Disease, *Applied Statistics* **45**, 275–309.

[35] Schnell, D., Zaidi, A. & Reynolds, G. (1989). A time series analysis of gonorrhea surveillance data, *Statistics in Medicine* **12**, 343–352.

[36] Serfling, R.E. & Sherman, I.J. (1965). *Attribute Sampling Methods for Local Health Departments*. US Government Printing Office, Washington.

[37] Serfling, R.E., Cornell, R.G. & Sherman, I.J. (1960). The CDC Quota Sampling Technic with results of 1959 poliomyelitis vaccination surveys, *American Journal of Public Health* **50**, 1847–1857.

[38] Serfling, R.E., Sherman, I.J. & Houseworth, W.J. (1967). Excess pneumonia-influenza mortality by age and sex in three major influenza A2 epidemics, United States, 1957–58, 1960, and 1963, *American Journal of Epidemiology* **86**, 433–441.

[39] Stroup, D.F., Williamson, G.D., Herndon, J.L. & Karon, J.M. (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data, *Statistics in Medicine* **12**, 323–329.

[40] US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention (1996). *Fact Book; FY 1996*.

[41] US Department of Health and Human Services, Centers for Disease Control and Prevention. *The State of the CDC, Fiscal Year 2003*.

[42] Usilton, L.J. (1940). A mechanical system for record keeping of morbidity, treatment-progress, and control of venereal diseases, *American Journal of Public Health* **30**, 928–930.

[43] `www.cdc.gov`.

[44] Whaley, F.S., Quade, D. & Haley, R.W. (1980). Effects of method error on the power of a statistical test: implications of imperfect sensitivity and specificity in retrospective chart review, *American Journal of Epidemiology* **111**, 534–542.

[45] Zaidi, A., Schnell, D., Reynolds, G. (1989). Time series analysis of gonorrhea surveillance data, *Statistics in Medicine* **12**, 353–362.

GLADYS H. REYNOLDS

Central Limit Theory

Central limit theory asserts that the sum of a large number of none too large independent **random variables** is approximately normally distributed (*see* **Large-sample Theory**). Results such as these are important to statistical theory, because they provide general conditions under which the distribution of a **mean** is well approximated by the **normal distribution**. Central limit theory allows us to use the normal distribution in creating **confidence intervals, hypothesis testing**, and in many other statistical procedures.

The most basic central limit theorem (CLT) is for sums of independent, identically distributed (iid) random variables. See [2] and [5] for a general introduction to central limit theory.

IID CLT

Suppose $X_1, \ldots, X_n$ are iid random variables with finite mean μ and finite, nonzero variance σ^2. Then for $S_n = X_1 + \cdots + X_n$, the rescaled and centered sum,

$$\frac{S_n - n\mu}{n^{1/2}\sigma}, \tag{1}$$

converges in distribution to the normal law as n increases. That is, for all t,

$$\Pr\left[\frac{S_n - n\mu}{n^{1/2}\sigma} \leq t\right] \to \Phi(t), \tag{2}$$

where Φ is the cumulative distribution function for the standard normal.

There are many statistical applications that build on this CLT, but which are not themselves sums of iid random variables. These include **rank** statistics, ***U*-statistics**, and M-estimators (*see* **Robustness**). We present three examples here.

The Sample Median

Suppose $X_1, \ldots, X_n$ form an independent sample from a distribution with median 0 and positive density γ at 0. The sample **median**, M_n, is approximately normally distributed for large n. That is, the series $n^{1/2}2\gamma M_n$ converges in distribution to the standard normal. The sample median does not appear to be a sum of random variables, but notice that, for odd n,

$$\Pr(n^{1/2}M_n \leq t) = \Pr[\text{at least } (n+1)/2 \text{ observations are } \leq t/n^{1/2}], \tag{3}$$

and the problem reduces to one involving convergence in distribution of a sequence of **binomials**, which are sums of iid random variables.

The Sample Variance

In this example, let $X_1, \ldots, X_n$ be iid with mean μ, finite variance σ^2 and finite fourth central **moment** μ_4. The sample variance,

$$s_n^2 = (n-1)^{-1}\sum (X_i - \overline{X})^2, \tag{4}$$

is a U-statistic with kernel of degree 2. This can be more easily seen by rewriting s_n^2 as follows:

$$\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \tfrac{1}{2}(X_i - X_j)^2. \tag{5}$$

Using the Hoeffding decomposition, s_n^2 can be written as $n^{-1}\sum (X_i - \mu)^2$ plus negligible terms, and it can be shown that $n^{1/2}(s_n^2 - \sigma^2)$ is normally distributed, in the limit, with mean 0 and variance $\mu_4 - \sigma^4$.

Maximum Likelihood Estimator

Again we take $X_1, \ldots, X_n$ to be iid. Now we assume that the distribution of the Xs has density $f(x, \boldsymbol{\theta}_0)$, where $f(x, \boldsymbol{\theta}_0)$ belongs to a k-parameter **exponential family**, i.e. for $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^k$,

$$f(x, \boldsymbol{\theta}) = \exp[\boldsymbol{\theta}'\mathbf{T}(x) - B(\boldsymbol{\theta})]. \tag{6}$$

The **maximum likelihood** estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is that $\boldsymbol{\theta}$ which maximizes the log **likelihood**,

$$\frac{1}{n}\sum \boldsymbol{\theta}'\mathbf{T}(X_i) - B(\boldsymbol{\theta}). \tag{7}$$

A multivariate CLT for the iid sum of the $\mathbf{T}(X_i)$ leads us to a central limit theorem for $\hat{\boldsymbol{\theta}}_n$. Under suitable conditions [4], $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to a **multivariate normal** with variance–**covariance matrix** $\{\text{var}[\mathbf{T}(X)]\}^{-1}$.

Multivariate central limit theory for sums of iid random vectors follows from the CLT for iid random variables. Convergence in distribution of a sequence

of random variables is equivalent to pointwise convergence of the corresponding **characteristic functions**, providing the limit is continuous at the origin [2]. Characteristic functions are especially useful in establishing CLTs for sums of independent random variables because of their multiplicative properties. From this approach, we can show that convergence in the distribution of random vectors is equivalent to convergence in the distribution of all linear functions of the random vectors.

Multivariate CLT

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are iid random vectors in $\mathbf{R}^k$ with mean $\mathbf{0}$ and variance–covariance matrix $\mathbf{V}$, then $n^{-1/2}/\mathbf{S}_n$ converges in distribution to the multivariate normal with mean $\mathbf{0}$ and variance–covariance matrix $\mathbf{V}$.

History

A. De Moivre proved the first central limit theorem in 1733 [10]. It covers the classic example of counting the number of successes in a large number of independent Bernoulli trials (*see* **Binary Data**), each with probability p. Basically it says that, as the number of trials increases, the distribution of the number of successes, when properly standardized, approximately follows the normal distribution. **P.S. Laplace** established similar results in 1812 [18]. More formally, the De Moivre–Laplace central limit theorem is for a sequence of n independent Bernoulli trials that are 1 with probability p and 0 with probability $q = 1 - p$. The sum of the trials, S_n, follows the binomial distribution with parameters n and p. As n grows, the distribution of the sum approaches the normal distribution. By this we mean that, for large n, the chance that S_n is at most k is well approximated by the normal probability,

$$\Phi[(npq)^{-1/2}(k + \tfrac{1}{2} - np)]. \tag{8}$$

De Moivre–Laplace CLT

For β_n and α_n such that $(\beta_n - np)^3/n^2 \to 0$ and $(\alpha_n - np)^3/n^2 \to 0$, as $n \to \infty$,

$$\Pr(\alpha_n \le S_n \le \beta_n) \sim \Phi[(npq)^{-1/2}(\beta_n + \tfrac{1}{2} - np)] - \Phi[(npq)^{-1/2}(\alpha_n + \tfrac{1}{2} - np)], \tag{9}$$

where $\sim$ denotes asymptotic equivalence.

See [7] for a proof of this central limit theorem.

In the second half of the nineteenth century, P.L. Chebyshev developed limit theorems for sums of arbitrarily distributed random variables. His results are based on the **method of moments**. Liapounoff, in 1900 and 1901, derived more general central limit theorems using characteristic functions. The most classical central limit theorem is due to J.W. Lindeberg in 1922. Lindeberg's result holds for independent random variables that are not necessarily identically distributed. We consider the triangular array of random variables, where for each n, $X_{n1}, \ldots, X_{nk(n)}$ is a sequence of $k(n)$ independent random variables with mean 0 and variances $\sigma_{n1}^2, \ldots, \sigma_{nk(n)}^2$. If we let

$$b_n^2 = \sum_{j=1}^{k(n)} \sigma_{nj}^2, \tag{10}$$

then the Lindeberg condition is

$$\frac{1}{b_n^2} \sum_{j=1}^{k(n)} \mathrm{E} X_{nj}^2 \mathrm{I}(|X_{nj}| \ge \varepsilon b_n) \to 0, \quad \text{for every } \varepsilon > 0, \tag{11}$$

where $\mathrm{I}(\cdot)$ denotes an indicator function (*see* **Dummy Variables**).

Provided the Lindeberg condition is met, S_n/b_n converges in distribution to the normal law. The Lindeberg condition is almost both necessary and sufficient. Feller showed in 1935 that if each of the random variables in the sum is small, then the Lindeberg condition is both necessary and sufficient. This result is called the Lindeberg–Feller central limit theorem.

Lindeberg–Feller CLT

The rescaled sum S_n/b_n converges in distribution to the normal law and $\max_j \sigma_{nj}/b_n \to 0$ if and only if the Lindeberg condition holds.

Notice that the iid central limit theorem is a special case of the Lindeberg–Feller result. There, $k(n) = n$, $X_{nj} = (X_j - \mu)/n^{1/2}\sigma$, $b_n = 1$, and the Lindeberg condition holds by dominated convergence.

There are many generalizations of the Lindeberg–Feller result. They include bounds on the rates of convergence, convergence in the distribution of random functions, and convergence of sums of dependent random variables. In addition, limit laws can be

obtained for sums of random variables with infinite variance, and limit laws other than the normal have been studied. We briefly discuss these generalizations here.

Rates of Convergence

One well-known result on the rate of convergence to normality in the CLT was arrived at independently by Berry & Esseen.

Berry–Esseen

Suppose $X_1, \ldots, X_n$ are iid random variables with mean 0, variance σ^2, and $\mathrm{E}|X_i|^3 = \rho < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then, for all n and t,

$$|\Pr(S_n < tn^{1/2}\sigma) - \Phi(t)| \leq \frac{3\rho\sigma^{-3}}{\sqrt{n}}. \qquad (12)$$

For other rate results see Petrov [15].

Functional CLT

One of the simplest examples of a random function arises from an iid sample $X_1, \ldots, X_n$ from the **uniform distribution** on (0,1). The standardized empirical cumulative distribution function,

$$U_n(t) = n^{-1/2} \sum [\mathrm{I}(X_i \leq t) - t], \quad \text{for } 0 \leq t \leq 1, \qquad (13)$$

is a random function in the space $D[0, 1]$ of all real-valued functions that are right continuous at each point of [0,1) with left limits existing at each point of (0,1].

For each fixed t, $U_n(t)$ converges in distribution to the normal law with mean 0 and variance $t(1-t)$. The multivariate CLT says that, for each k-vector $(t_1, \ldots, t_k)'$, the random vector $[U_n(t_1), \ldots, U_n(t_k)]'$ converges in distribution to a multivariate normal with $\mathrm{cov}[U_n(t_i), U_n(t_j)] = \min(t_i, t_j) - t_i t_j$. A functional CLT says that U_n converges in distribution to the **Brownian** Bridge. Of particular interest is the limiting distribution of functionals of U_n such as the **Kolmogorov–Smirnov** statistic $\sup_t |U_n(t)|$ and the Cramér–Von Mises statistic $\int_0^1 U_n(t)^2 \mathrm{dt}$ (*see* **Kolmogorov–Smirnov and Cramer–Von Mises Tests in Survival Analysis**). See [1] for functional CLTs for these and related **stochastic processes**. See [8], [12], and [16] for functional CLTs for empirical processes on spaces of functions.

Dependence

There are many CLTs for sums of random variables that are not independent. One example is the CLT for martingale difference arrays [11].

Martingale CLT

Suppose $X_{n1}, \ldots, X_{nn}$ form a martingale difference array with respect to the sigma-fields $\mathcal{F}_{n0}, \mathcal{F}_{n1}, \ldots, \mathcal{F}_{nn}$. Then, provided

$$\sum_{j=1}^{n} \mathrm{E}(X_{nj}^2 | \mathcal{F}_{n,j-1}) \to \sigma^2 \qquad (14)$$

in probability, with σ^2 a positive constant, and

$$\sum_{j=1}^{n} \mathrm{E}[X_{nj} \mathrm{I}(|X_{nj}| > \varepsilon) | \mathcal{F}_{n,j-1}] \to 0 \qquad (15)$$

in probability for every ε, then the sum $\sigma^{-1} S_n$ converges in distribution to the standard normal law.

Other CLTs for martingales can be found in [14] and [17]. CLTs for sums of dependent random variables where the dependence satisfies a mixing condition can be found in [6].

More General CLTs

Central limit theorems can hold under very general conditions. The random variables need not have finite variance. For example, take $X_1, \ldots, X_n$ to be iid from a symmetric distribution where $\Pr(|X_1| > x) = x^{-2}$, for $x \geq 1$. Then, despite the fact that the X_is have infinite variance, $(n \log n)^{-1/2} S_n$ converges in distribution to the standard normal. This is established by a truncation argument, where new random variables Y_{nj} are created such that $Y_{nj} = X_j$ provided $|X_{nj}| \leq \sqrt{n}$ log log n, and $Y_{nj} = 0$ otherwise. These truncated variables have finite variances that are roughly of size log n, and their sum differs very little from the sum of the original random variables.

This result for random variables with infinite variance leads to the question of what conditions are

needed for a sum of iid random variables to have a normal limit (*see* **Limit Theorems**).

IID CLT

Suppose $X_1, \ldots, X_n$ are iid random variables. In order that there exist constants a_n and $b_n > 0$ such that $(S_n - a_n)/b_n$ converges in distribution to the normal law, it is necessary and sufficient that, as $x \to \infty$,

$$\frac{x^2 \Pr(|X_1| > x)}{\mathrm{E}\left[|X_1|^2 \mathrm{I}(|X_1| \le x)\right]} \to 0. \tag{16}$$

Sums of independent random variables can have limit laws other than the normal. For example, the sum of iid **Cauchy** random variables, when properly normalized, has a limiting Cauchy distribution. The set of possible limit laws for sums of iid random variables are the stable laws. See [3] and [9] for these and other results.

Sums based on triangular arrays of random variables can also have limit laws other than the normal. (The iid sequence is a special case of the triangular array, as noted above.) For example, suppose the X_{nj} are independent Bernoulli random variables with probability p_{nj}, $j = 1, \ldots, n$. If $p_{n1} + \cdots + p_{nn} \to \lambda$ and $\max p_{nj} \to 0$, then S_n converges in distribution to a Poisson law with parameter λ.

In general, we consider the triangular array of independent random variables $X_{n1}, \ldots, X_{nk(n)}$ for $k(n) \to \infty$. The summands are said to be uniformly asymptotically negligible (uan) if the X_{nj} converge uniformly to 0 in probability. For uan independent summands, the family of limit laws of the sequence $\sum X_{nj}$ coincides with the family of infinitely divisible distributions, which includes the stable laws. Examples of infinitely divisible distributions are the normal, Poisson, **gamma**, and **geometric**. See [2] and [13] for a more detailed treatment of this material.

References

[1] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[2] Billingsley, P. (1986). *Probability and Measure*, 2nd Ed. Wiley, New York.

[3] Breiman, L. (1968). *Probability*. Addison-Wesley, Reading.

[4] Chernoff, H. (1954). On the distribution of the likelihood ratio, *Annals of Mathematical Statistics* **25**, 573–578.

[5] Durrett, R. (1996). *Probability: Theory and Examples*, 2nd Ed. Duxbury, Belmont.

[6] Eberlein, E. & Taqqu, M.S. (1986). *Dependence in Probability and Statistics: A Survey of Recent Results*. Birkhauser-Verlag, New York.

[7] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd Ed. Wiley, New York.

[8] Giné, E. & Zinn, J. (1984). Some limit theorems for empirical processes, *Annals of Probability* **12**, 929–989.

[9] Gnedenko, B.V. & Kolmogorov, A.N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Reading.

[10] Hald, A. (1989). *A History of Probability and Statistics and Their Applications Before 1750*. Wiley, New York.

[11] Hall, P. & Heyde, C.C. (1980). *Martingale Limit Theory and Its Applications*. Academic Press, New York.

[12] Ledoux, M. & Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York.

[13] Loéve, M. (1963). *Probability Theory*, 3rd Ed. Springer-Verlag, New York.

[14] McLeish, D.L. (1974). Dependent central limit theorems and invariance principles, *Annals of Probability* **2**, 620–628.

[15] Petrov, V.V. (1974). *Sums of Independent Random Variables*. Springer-Verlag, Berlin.

[16] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

[17] Shiryayev, A.N. (1981). Martingales: recent developments, results, and applications, *International Statistical Review* **49**, 199–233.

[18] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, Mass.

DEBORAH NOLAN

Centroid Method

The centroid method is a factoring procedure in **factor analysis**. Before the advent of high-speed computers, this was a very popular factoring method. The initial work on the centroid method was presented by Burt [1] in 1917 and it was fully developed by Thurstone [3] in 1931. The centroid method relies on the idea that if the original variables are represented as a set of vectors, then the common factor can be interpreted as a vector which passes through the centroid of the terminal points for this set of vectors.

To compute the m centroid factors for a set of p variables ($m < p$), we first calculate the **correlation** matrix for the p variables. We denote this $p \times p$ matrix as $\mathbf{R}$ and the element at the ith row and jth column of $\mathbf{R}$ as r_{ij}. Next we sum the columns of the correlation matrix, including the diagonal element or **communality** estimate. If there are negative column sums, then we should reflect the negative sums by changing their signs. We then add all the column sums to yield a grand total. The first centroid **factor loadings** are now obtained by dividing each column sum by the square root of the grand total. In symbols, the first centroid factor loading $\mathbf{a} = (\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_p) = (\sum r_{i_1} / \sum\sum r_{ij}, \sum r_{i2} / \sum\sum r_{ij}, \ldots, \sum r_{ip} / \sum\sum r_{ij})$. These factor loadings are not final until we undo the earlier reflections that are applied to some of the negative column sums. There are two ways we can undo the change. One is to reverse the names of the variables that had the signs of their sums changed earlier and keep the loadings as they are. The alternative is to reflect the loadings of these variables and retain the original names of the variables. For the first factor loadings, we usually take the approach of reversing the names of the variables, because keeping all the loadings of this first factor positive will facilitate the rotation step that follows and also will simplify the subsequent interpretation of the rotated factor.

To determine the loadings on the second factor, we form the first residual correlation matrix. The elements of this residual matrix are denoted by $\{r_{ij \cdot a}\}$ which are equal to $r_{ij} - a_i a_j$. To obtain the second factor, **b**, we factor this residual matrix using the same computational procedure as was applied to the original correlation matrix, $\mathbf{R}$. To compute the third factor, **c**, we apply the same factoring procedure on the second residual matrix obtained from subtracting the second factor from the first residual correlation matrix. We obtain the elements of this second residual matrix $\{r_{ij \cdot b}\}$ as $r_{ij \cdot a} - b_i b_j$. Successive application of this procedure to the corresponding residual correlation matrix will give the complete centroid factor matrix.

The centroid method was originally derived as a mathematical approximation to the more difficult principal-axes procedure when computers were not generally available. Even though the centroid solution yields the same complexity of variables and factors, and also has the same variance contributions of the factors as the principal-axes procedure, it does not share the other important mathematical properties of the principal-axes solution, which include uniqueness and orthogonality. A treatment of the centroid method in factor analysis is given in Cureton & D'Agostino [2, Chapter 2].

References

[1] Burt, C. (1917). *The Distribution and Relations of Educational Abilities*. King & Son, London.

[2] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[3] Thurstone, L.L. (1931). Multiple factor analysis, *Psychological Review* **38**, 406–427.

(*See also* **Factor Analysis, Overview**)

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Chain Binomial Model

Let time be discrete and indexed $t = 0, 1, \ldots$. Let S_t be the number of individuals at risk for the event of interest (e.g. infection or death) at the beginning of time interval t, and let I_t be the number that experienced the event of interest at the beginning of time interval t. The event has a duration of at least one time interval. We let $p_t = 1 - q_t = f(t, \boldsymbol{\theta}, I_t)$ be the probability that an at-risk individual has a new event at the beginning of time interval time $t + 1$, with parameter $\boldsymbol{\theta}$. As shown, this probability can be a function, $f(\cdot)$, of t and I_t. We usually start with a closed population of $n = S_0 + I_0$ individuals. Then I_{t+1} is a **binomial** random variable that follows the conditional probability mass function

$$\begin{aligned}&\Pr(I_{t+1} = i_{t+1} | S_t = s_t, p_t)\\ &\quad = \binom{s_t}{i_{t+1}} p_t^{i_{t+1}} q_t^{s_t - i_{t+1}}, \quad s_t \geq i_{t+1}. \end{aligned} \tag{1}$$

In many cases, S_t is updated via the relationship

$$S_{t+1} = S_t - I_{t+1}, \tag{2}$$

although other relationships are possible (see below). The conditional expectation and variance of I_{t+1}, respectively, are

$$\mathrm{E}(I_{t+1} | s_t, p_t) = s_t p_t, \tag{3}$$

$$\mathrm{var}(I_{t+1} | s_t, p_t) = s_t p_t q_t. \tag{4}$$

Eqs (1) and (2) form the classical chain binomial model. Formal mathematical treatment of the model involves formulation of the discrete, two-dimensional **Markov chain** $\{S_t, I_t\}_{t=0,1,\ldots}$. I_t is the (binomial) random variable of interest, and S_t is updated using (2). The probability of a particular chain, $\{i_0, i_1, i_2, \ldots, i_r\}$, is given by the product of conditional binomial probabilities from (1) as

$$\begin{aligned}&\Pr(I_1 = i_1 | S_0 = s_0, p_0) \Pr(I_2 = i_2 | S_1 = s_1, p_1)\\ &\quad \times \cdots \Pr(I_r = i_r | S_{r-1} = s_{r-1}, p_{r-1})\\ &\quad = \prod_{t=0}^{r-1} \binom{s_t}{i_{t+1}} p_t^{i_{t+1}} q_t^{s_t - i_{t+1}}. \end{aligned} \tag{5}$$

The conditional expected value of I_{t+1} from (3) suggests the deterministic system of first-order difference equations

$$i_{t+1} = s_t p_t, \qquad s_{t+1} = s_t - i_{t+1}, \tag{6}$$

which can be analyzed as an approximation to the mean of the sample paths of the **stochastic process** $\{S_t, I_t\}_{t=0,1,\ldots}$. This system reduces to

$$s_t = s_{t-1} q_{t-1} = s_0 \prod_{l=0}^{t-1} q_l, \tag{7}$$

which is analyzed using methods from discrete mathematics (see, for example, [7] and [11]).

The Reed–Frost Model

History

The probabilistic form of the Reed–Frost epidemic model was introduced by the biostatistician Lowell J. Reed and the epidemiologist Wade Hampton Frost around 1930, as a teaching tool at Johns Hopkins University. It was developed as a mechanical model consisting of colored balls and wooden shoots. Although Reed and Frost never published their results, the work is described in articles and books by others (see [1, Chapters 14 and 18] and [2, Chapters 2 and 3]). An excellent description of the early Reed–Frost model is given by Fine [6]. The deterministic version of the Reed–Frost model has been traced back to the Russian epidemiologist P.D. En'ko, who used the model to analyze epidemic data in the 1880s (see [5]). The Reed–Frost version of the chain binomial and its extensions is used to study the dynamics of epidemics in small populations, such as families or day care centers, and to estimate transmission probabilities from epidemic data.

Formulation

In this case, S_t is the number of susceptible persons at the beginning of time interval t, and I_t is the number of persons who were newly infected at the beginning of time interval t. An infected person is infectious for exactly one time interval and then is removed; that is, becomes immune. Thus, a person infected at the beginning of time interval t will be infectious to others until the beginning of time interval $t + 1$. We let R_t be the number of removed persons at the beginning of time interval t, and then, by definition,

$$R_{t+1} = R_t + I_t = R_0 + \sum_{r=0}^{t} I_r. \tag{8}$$

Since the population is closed, we have $S_t + I_t + R_t = n$ for all t. We let $p = 1 - q$ be the probability that any two specified people make sufficient contact in order to transmit the infection, if one is susceptible and the other infected, during one time interval. We note that p is a form of the **secondary attack rate**. We assume **random mixing**. Then, if during time interval t there are I_t infectives, the probability that a susceptible will escape being infected over the time interval is q^{I_t}, and the probability that they will become a new case at the beginning of time interval $t + 1$ is $1 - q^{I_t}$. Thus $q_t = q^{I_t}$, and substituting into (1) yields

$$\Pr(I_{t+1} = i_{t+1} | S_t = s_t, I_t = i_t) = \binom{s_t}{i_{t+1}} (1 - q^{i_t})^{i_{t+1}} q^{i_t(s_t - i_{t+1})}, \quad s_t \geq i_{t+1}. \tag{9}$$

The epidemic process starts with $I_0 > 0$, and terminates at stopping time T, where

$$T = \inf_{t \geq 0} \{t : S_t I_t = 0\}. \tag{10}$$

The possible chains for a population of size 4 with one initial infective – that is, $S_0 = 3$, $I_0 = 1$ – are shown in Table 1.

The probability of no epidemic is defined as the probability that there will be no further cases beyond the initial cases. This probability is

$$\Pr(I_1 = 0 | S_0 = s_0, p_0) = q^{i_0 s_0}. \tag{11}$$

For example, if $S_0 = 10$, $I_0 = 1$, and $p = 0.05$, then the probability of no further cases beyond the initial case is 0.599. From (3), the conditional expected number of new cases in time interval t is $\mathrm{E}(I_{t+1} | s_t, p_t) = s_t(1 - q^{i_t})$. On the average, the epidemic process will not progress very far if the expected number of cases in the first generation is less than or equal to one; that is, $\mathrm{E}(I_1 | s_0, p_0) = s_0(1 - q^{i_0}) \leq 1$. In many cases, $i_0 = 1$, so that there will be few secondary cases if $s_0 p \leq 1$. Then, for example, if $S_0 = 10$, $I_0 = 1$, there will be few secondary cases if $p \leq 0.1$.

Table 1 Possible individual chains when $S_0 = 3$, $I_0 = 1$

Chain	Probability	Final size
$\{i_0, i_1, i_2, \ldots, i_T\}$		R_T
$\{1\}$	q^3	1
$\{1, 1\}$	$3pq^4$	2
$\{1, 1, 1\}$	$6p^2q^4$	3
$\{1, 2\}$	$3p^2q^3$	3
$\{1, 1, 1, 1\}$	$6p^3q^3$	4
$\{1, 1, 2\}$	$3p^3q^2$	4
$\{1, 2, 1\}$	$3p^3q(1 + q)$	4
$\{1, 3\}$	p^3	4

From (7), the deterministic counterpart of the Reed–Frost model is

$$s_t = s_0 q^{\sum_{l=0}^{t-1} i_l}, \tag{12}$$

which has been thoroughly analyzed (see, for example [7] and [11]).

In some cases, the distribution of the total number of cases, R_T, is the random variable of interest. We let J be the random variable for the total number of cases in addition to the initial cases, so that $R_T = J + I_0$. If we let $S_0 = k$ and $I_0 = i$, then the probability of interest is

$$\Pr(J = j | S_0 = k, I_0 = i) = m_{ijk}, \tag{13}$$

where $\sum_{j=0}^{k} m_{ijk} = 1$. Then, based on probability arguments (see, for example, [1]), we have the recursive expression

$$m_{ijk} = \binom{k}{j} m_{ijj} q^{(i+j)(k-j)}, \quad j < k, \tag{14}$$

and

$$m_{ikk} = 1 - \sum_{j=0}^{k-1} m_{ijk}. \tag{15}$$

The Reed–Frost model has several extensions and special cases. If it is hypothesized that the probability that a susceptible becomes infected does not depend on the number of infectives that he or she is exposed to, then

$$p_t = \begin{cases} p, & \text{if } I_t > 0, \\ 0, & \text{if } I_t = 0. \end{cases} \tag{16}$$

This model is known as the Greenwood model [8].

Longini & Koopman [12] modified the Reed–Frost model for the common case in which there is a constant source of infection from outside the population that does not depend on the number of infected persons in the population. We let $a_t = 1 - b_t$ be the probability that a susceptible person is infected during interval t due to contacts with infected persons outside the population, where

$$a_t > 0, \quad \text{if } t \leq T,$$

$$a_t = 0, \quad \text{if } t > T,$$

and T is a stopping time. Then $p_t = 1 - b_t q^{I_t}$. If we let $B = \prod_{t=0}^{T} b_t$, then B is the probability that a person escapes infection from sources outside of the population over the entire period $[0, T]$. We then define CPI $= 1 - B$ as the community probability of infection. Longini & Koopman derive the probability mass function

$$m_{ijk} = \binom{k}{j} m_{ijj} B^{(k-j)} q^{(i+j)(k-j)}, \quad j < k. \quad (17)$$

Usually, $i = 0$ for this model. This model reduces to (14) when $B = 1$.

Another extension of the Reed–Frost model is for infectious diseases that do not confer immunity following infection. In this case, there is no removed state, so that $S_t + I_t = n$. Then, since $S_{t+1} = n - I_{t+1}$, the model is a discrete, one-dimensional Markov chain $\{I_t\}_{t=0,1,\ldots}$. The transition probabilities for this process are

$$\begin{aligned} &\Pr(I_{t+1} = i_{t+1} | I_t = i_t) \\ &\quad = \binom{n - i_t}{i_{t+1}} (1 - q^{i_t})^{i_{t+1}} q^{i_t(n - i_t - i_{t+1})}, \\ &\quad i_t + i_{t+1} \le n. \end{aligned} \quad (18)$$

In this case, the disease in question can become "endemic". An interesting analytic question involves the study of the mean stopping time for the endemic process. From (6), the deterministic counterpart of this model is

$$i_{t+1} = (n - i_t)(1 - q^{i_t}), \quad (19)$$

which is a form of the discrete logistic function. The stochastic behavior of (18) has been analyzed by Longini [10], and the dynamics of (19) have been analyzed by Cooke et al. [4].

There are many other extensions of the Reed–Frost model depending on the particular infectious disease being analyzed, but a further key extension is to allow the infectious period to extend over several time intervals. In this case $p_t = f(t, \boldsymbol{\theta}, I_0, I_1, \ldots, I_t)$, and $\{S_t, I_t\}_{t=0,1,\ldots}$ is not a Markov chain. Special methods are used to analyze this model [14].

Inference

Data are usually in the form of observed chains, $\{i_0, i_1, \ldots, i_r\}$, for one or more populations, or final sizes, R_T, for more than one population. With respect to the former data form, suppose that we have N populations and let $\{i_{k0}, i_{k1}, \ldots, i_{kr}\}$ be the observed chain for the kth population. Then, from (5), the **likelihood** function for estimating $p = 1 - q$ is

$$L(p) = \prod_{k=1}^{N} \prod_{t=0}^{r-1} \binom{s_{kt}}{i_{kt+1}} (1 - q^{i_{kt}})^{i_{kt+1}} q^{i_{kt}(s_{kt} - i_{kt+1})}. \quad (20)$$

For final value data, let a_{ijk} be the observed frequencies of the m_{ijk}, from (17); $i = 1, \ldots, I$, $k = 1, \ldots, K$, and $j = 1, \ldots, k$. Then the likelihood function for estimating p and B is

$$L(p, B) = \prod_{i=1}^{I} \prod_{k=1}^{K} \prod_{j=0}^{k} m_{ijk}^{a_{ijk}}. \quad (21)$$

The logarithms of (20) and (21) are maximized using standard scoring routines (see, for example, Bailey [1], Becker [2], and Longini et al. [12, 13]) (*see* **Optimization and Nonlinear Equations**) or the corresponding **generalized linear model** (see Becker [2] and Haber et al. [9]). Extensions involve making both p and the CPI functions of covariates, such as age, level of susceptibility, or vaccination status (*see* **Vaccine Studies**). Bailey [1, Section 14.3] gives an example in which (20) is used to estimate $\hat{p} = 0.789 \pm 0.015$ (estimate ± 1 standard error) for the household spread of measles among children. In the case of the household spread of influenza, Longini et al. [13] use (21) to estimate $\hat{p} = 0.260 \pm 0.030$ for persons with no prior immunity and $\hat{p} = 0.021 \pm 0.026$ for persons with some prior immunity. In addition, they estimate $\widehat{CPI} = 0.164 \pm 0.015$ and $\widehat{CPI} = 0.092 \pm 0.013$ for persons with no and some prior immunity, respectively.

Life Tables

The chain binomial model forms the statistical underpinnings of the **life table** (see Chiang [3, Chapter 10]). In this case, p_t simply depends on the time interval. Then S_t is the random variable of interest, which is formulated in terms of the interval survival probabilities $q_t = 1 - p_t$. Many important life table indices are functions of q_t. For example, the probability that an individual who starts in the cohort at time zero, is still alive at the end of time interval r, denoted q_{0r}, is $q_{0r} = \prod_{t=0}^{r} q_t$. The expected

number alive at the beginning of time interval $r+1$ is $\mathrm{E}(S_{r+1}) = s_0 q_{0r}$. This model is a discrete, one-dimensional Markov chain $\{S_t\}_{t=0,1,\ldots}$. From (1) we see that the chain binomial model for S_t is simply

$$\Pr(S_{t+1} = s_{t+1} | S_t = s_t) = \binom{s_t}{s_{t+1}} q_t^{s_{t+1}} p_t^{s_t - s_{t+1}}, \quad s_t \geq s_{t+1}. \tag{22}$$

From (5), the probability of a particular chain $\{s_0, s_1, s_2, \ldots, s_r\}$ is

$$\begin{aligned} &\Pr(S_1 = s_1 | S_0 = s_0) \Pr(S_2 = s_2 | S_1 = s_1) \\ &\qquad \times \cdots \Pr(S_r = s_r | S_{r-1} = s_{r-1}) \\ &\quad = \prod_{t=0}^{r-1} \binom{s_t}{s_{t+1}} q_t^{s_{t+1}} p_t^{s_t - s_{t+1}}. \end{aligned} \tag{23}$$

For an observed chain $\{s_0, s_1, s_2, \ldots, s_r\}$, (23) is the likelihood function for estimating $\{q_0, q_1, \ldots, q_r\}$. The maximum likelihood estimators are

$$\hat{q}_t = s_{t+1}/s_t, \tag{24}$$

while the approximate variances, for large S_0, are

$$\mathrm{var}(\hat{q}_t) \approx p_t q_t / \mathrm{E}(S_t). \tag{25}$$

In addition, the $\hat{q}_t$ are unique, unbiased estimates of the q_t, and $\mathrm{cov}(\hat{q}_t, \hat{q}_l) = 0, t \neq l$. Estimators of most of the life table functions are based on the estimators $\hat{q}_t$.

References

[1] Bailey, N. (1975). *The Mathematical Theory of Infectious Diseases*, 2nd Ed. Griffin, London.

[2] Becker, N. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, New York.

[3] Chiang, C. (1984). *The Life Table and its Applications*. Krieger, Malabar.

[4] Cooke, K., Calef, D. & Level, E. (1977). Stability or chaos in discrete epidemic models, in *Nonlinear Systems and Applications – An International Conference*. Academic Press, New York.

[5] Dietz, K. (1988). The first epidemic model: a historical note on P.D. En'ko, *Australian Journal of Statistics* **30A**, 56–65.

[6] Fine, P. (1977). A commentary on the mechanical analogue to the Reed–Frost epidemic model, *American Journal of Epidemiology* **106**, 87–100.

[7] Frauenthal, J. (1980). *Mathematical Models in Epidemiology*. Springer-Verlag, Berlin.

[8] Greenwood, M. (1931). The statistical measure of infectiousness, *Journal of Hygiene* **31**, 336–351.

[9] Haber, M., Longini, I. & Cotsonis, G. (1988). Models for the statistical analysis of infectious disease data. *Biometrics* **44**, 163–173.

[10] Longini, I. (1980). A chain binomial model of endemicity, *Mathematical Biosciences* **50**, 85–93.

[11] Longini, I. (1986). The generalized discrete-time epidemic model with immunity: a synthesis, *Mathematical Biosciences* **82**, 19–41.

[12] Longini, I. & Koopman, J. (1982). Household and community transmission parameters from final distributions of infections in households, *Biometrics* **38**, 115–126.

[13] Longini, I., Koopman, J., Haber, M. & Cotsonis, G. (1988). Statistical inference for infectious diseases: risk-specified household and community transmission parameters, *American Journal of Epidemiology* **128**, 845–859.

[14] Saunders, I. (1980). An approximate maximum likelihood estimator for chain binomial models, *Australian Journal of Statistics* **22**, 307–316.

(*See also* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**; **Infectious Disease Models**)

IRA M. LONGINI, JR

Chalmers, Thomas Clark

Born: December 17, 1917, in Forest Hills, New York.
Died: December 27, 1995, in Hanover, New Hampshire.

Thomas C. Chalmers, M.D., a leader in the design, conduct, and evaluation of **clinical trials**, was born in Forest Hills, New York, where his father was a physician in private practice. Following a tradition set by his father and grandfather, he graduated in 1943 from Columbia University College of Physicians and Surgeons. After additional training in medical research in New York and at the Thorndike Memorial Laboratories of Boston City Hospital, he entered private practice in Cambridge, Massachusetts, in 1947. He soon became concerned over the lack of knowledge on the efficacy of accepted medical therapies. Having learned about **randomization** from **Sir Austin Bradford Hill**, he applied this principle to a study of the treatment of infectious hepatitis among American soldiers in Japan during the Korean War. This study, a 2×2 randomized factorial study of diet and bed rest (*see* **Factorial Designs in Clinical Trials**), designed in 1951, included estimates of the required number of patients and an evaluation of ineligible patients, withdrawals, and **compliance**.

Deciding to devote his career to research and education, he was Chief of Medical Services at the Lemuel Shattuck Hospital in Boston (1955–1968), Assistant Director for Research and Education for the Veterans' Administration in Washington (1968–1970), Director of the Clinical Center at the National Institutes of Health in Bethesda (1970–1973), and President and Dean of the Mount Sinai Medical Center and School of Medicine in New York City (1973–1983).

Dr Chalmers returned to Boston in 1983. Over the next 10 years he was on the faculty of the Harvard School of Public Health and Tufts University School of Medicine, was appointed a Distinguished Professor at the Boston Veterans' Administration Medical Center, and was a member and Chairman of the Board of Trustees of the Dartmouth Hitchcock Medical Center. In 1992, at age 75, he cofounded Meta-Works, Inc., a **meta-analysis** consulting company and moved to Lebanon, New Hampshire. He continued to teach in both Boston and New York and was actively involved in numerous meta-analytic studies.

Throughout his career he was a fervent advocate of randomized controlled trials in all areas of medicine and the education of students and physicians in the skills needed to evaluate these trials. His belief in the ethical need for randomization [6] (*see* **Ethics of Randomized Trials**) led to his recommendation to "begin randomization with the first patient" [1]. A corollary was the belief that developing trends should not be known by investigators during the conduct of the trial, but should be monitored by an independent policy advisory committee (*see* **Data Monitoring Committees**). In subsequent years he was a member (and frequently chairman) of Policy or Data Safety and Monitoring Boards for numerous multicenter clinical trials.

Dr Chalmers moved to Mount Sinai in 1973 because he wanted to influence the education of medical students, and to make both students and faculty aware of the need for properly conducted clinical trials. He became concerned that clinical trials were being conducted with insufficient sample sizes and in a review published in the *New England Journal of Medicine* [3] he sought to educate physicians concerning the importance of both type II errors (*see* **Hypothesis Testing**) and sample size calculations in the planning and evaluation of clinical trials (*see* **Sample Size Determination for Clinical Trials**). His method for assessing the quality of randomized controlled trials became a standard for the evaluation of published reports of clinical trials [2]. He was one of 12 founding members of the Society for Clinical Trials and established a student scholarship to encourage the involvement of students in clinical research.

Upon returning to Boston in 1983, Dr Chalmers continued to be interested in the combination of data from multiple clinical trials and introduced the readers of the *New England Journal of Medicine* to the meta-analysis of randomized controlled trials [5]. He educated numerous students in this method of analysis and continued to work with them on manuscripts up to within weeks of his death in 1995 from prostate cancer. An obituary and tributes to Chalmers have been given in a recent issue of **Controlled Clinical Trials** [4].

References

[1] Chalmers, T.C. (1972). Randomization and coronary artery surgery, *Annals of Thoracic Surgery* **14**, 323–327.

[2] Chalmers, T.C., Smith, H. Jr, Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. & Ambroz, A. (1981). A

method for assessing the quality of a randomized control trial, *Controlled Clinical Trials* **2**, 31–49.

[3] Freiman, J.A., Chalmers, T.C., Smith, H. & Kuebler, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials, *New England Journal of Medicine* **299**, 690–694.

[4] Knatterud, G.L. & Greenhouse, S.W. (1996). Tributes to Thomas C. Chalmers, MD, *Controlled Clinical Trials* **17**, 471–475.

[5] Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A. & Chalmers, T.C. (1987). Meta-analyses of randomized controlled trials, *New England Journal of Medicine* **316**, 450–455.

[6] Shaw, L.W. & Chalmers, T.C. (1970). Ethics in cooperative clinical trials, *Annals of the New York Academy of Sciences* **169**, 487–495.

E. WRIGHT

Change-point Problem

The general form of the change-point problem is to determine the unknown location τ, based on an ordered sequence of observations $x_1, \ldots, x_n$, such that the two groups of observations $x_1, \ldots, x_\tau$ and $x_{\tau+1}, \ldots, x_n$ follow distinct models. The index of ordering frequently refers to time, but in general it can be associated with any variable. The simplest example is a level-change model, where $x_1, \ldots, x_\tau$ are i.i.d $N(\mu_1, \sigma^2)$ and $x_{\tau+1}, \ldots, x_n$ are iid $N(\mu_2, \sigma^2)$ with $\mu_1 \neq \mu_2$. Another example is a two-phase **regression** or a linear regression switching model [6, 10], where $x_1, \ldots, x_\tau$ follow a model $\beta_{01} + \beta_{11}t$, for some predictor variable t, and $x_{\tau+1}, \ldots, x_n$ follow another model $\beta_{02} + \beta_{12}t$.

Figure 1 shows the breast cancer incidence rate in Sweden in 1990 for women between the age of 40 and 50. There is a dramatic change in slope around the age of 46, presumably because of hormonal changes that come with the onset of menopause. It is of interest to know the age when the change occurs.

The observed number of breast cancers are given in the following. Each age group is one-tenth of a year, and the age varies from 40.2 to 50, giving a total of 99 age groups.

6	1	4	6	2	2	4	3	6	5	1	3	2	5	5	4	6	5	4	5	2	5	2	6	5
9	8	7	6	7	6	3	5	10	11	4	4	4	10	6	4	7	7	7	6	10	11	8	10	8
11	3	12	8	13	9	5	7	11	10	12	8	11	11	6	11	13	7	9	12	12	7	11	10	8
10	8	8	10	10	8	7	14	6	8	11	6	5	7	14	6	8	5	9	7	10	11	8	4	

The associated number of person-years are the following. Rate is computed as the incidence divided by the person-years.

6389	6371	6352	6334	6315	6297	6278	6260	6241	6223	6204	6186	6167	6149	6130
6112	6093	6075	6056	6038	6019	6001	5982	5964	5945	5927	5908	5890	5871	5853
5834	5816	5797	5779	5760	5742	5723	5705	5686	5668	5649	5631	5612	5594	5575
5557	5538	5520	5501	5483	5465	5446	5428	5409	5391	5372	5354	5335	5317	5298
5280	5261	5243	5224	5206	5187	5169	5150	5132	5113	5095	5076	5058	5039	5021
5002	4984	4965	4947	4928	4910	4891	4873	4854	4836	4817	4799	4780	4762	4743
4725	4706	4688	4669	4651	4632	4614	4595	4577						

Page [16, 17] seems to be the earliest reference for this problem. Many authors, for example, [2, 5, 7, 11, 14], have considered extensions in various different directions and settings including **time series** data, multivariate observations, discrete observations, **Poisson process**, hazard or failure rate regressions (*see* **Survival Analysis, Overview**), **quality control** or surveillance problems, **sequential** applications and multiple change-point problems. The general techniques used are nonlinear **least-squares**, **maximum likelihood** [9, 19, 21], **nonparametric**/rank-based [3] and **Bayesian methods** [1]. Techniques may also be categorized as sequential or nonsequential. In sequential applications, the data x_i's are typically available one (or several) at a time and the objective is a quick detection of a change as soon as it occurs; the cumulative sum technique is the main tool in this area (*see* **Quality Control in Laboratory Medicine**). See [18] for a bibliography up to 1980 and a more recent review is given in [13].

A change-point problem may be viewed simply as a **nonlinear regression** problem, but from a theoretical point of view, the problem is rather nonstandard or nonregular. For example, even in the simplest example of a level-change model above, the likelihood function is not differentiable with respect to the parameter τ, so the standard theory does not apply. Strong simplifying assumptions are generally needed, for example, the existence of at most one change point, to derive theoretical results of the following types: consistency and asymptotic normality of the change-point estimate [4, 10, 13], and the asymptotic distribution of a maximum statistic [9, 19, 21] to decide if there is a change and, if so, where the

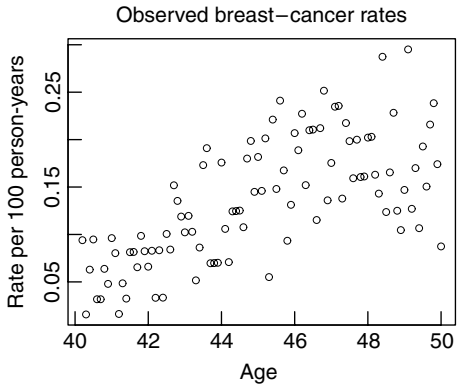

Figure 1 Breast cancer incidence for women between the age of 40 to 50 in Sweden in 1990

change is. Recent results are, for example, on the consistent estimation of the number and location of jumps [13].

From an application point of view, the nonlinear least-squares technique is the most straightforward for generating the estimate and this can be implemented easily using any statistical package that has a nonlinear least-squares routine. However, there does not seem to be a simple general result for the distribution theory of the change-point estimate; see [13]. Users may have to resort to **computer-intensive techniques** such as the **bootstrap method** to obtain a **confidence interval** [12].

We will review the nonlinear least-squares technique and express it in the language of quasi-likelihood so it naturally covers the **generalized linear model** setting [15]. It is common to start with the assumption of one change point and proceed by sequential splitting in the case of multiple change points [20]. Thus, let x_i, for $i = 1, \ldots, n$, be independent outcomes with mean and variance functions given by

$$Ex_i = \mu(t_i), \tag{1}$$

$$\text{Var } x_i = \phi v_i(\mu(t_i)), \tag{2}$$

where ϕ is a scale parameter; for example, for **normal** (Gaussian) outcomes, we might use $\phi = \sigma^2$ and $v_i(\mu) = 1$, and for **Poisson** outcomes $\phi = 1$ and $v_i(\mu) = \mu$. The mean function $\mu(t)$ is piecewise continuous function of the index variable t, which is not necessarily a time variable, with unknown regression parameters $\beta = (\beta_1, \beta_2)$ and the change-point τ according to

$$\mu(t) = \begin{cases} \mu_1(t, \beta_1) & \text{for } t < \tau \\ \mu_2(t, \beta_2) & \text{for } t \geq \tau. \end{cases} \tag{3}$$

Note that other **covariates** may also enter the regression function and the formulation would allow multivariate outcomes, but the change point is generally limited to one index variable. Denote by $\hat{\beta}$ and $\hat{\tau}$ the maximum **quasi-likelihood** estimates. Various quasi-likelihood models are now available for different types of outcome variables [15]. The standard estimation procedure alternates between τ and β in the following way. First, fix τ and estimate β using the standard (iterative-reweighted) least-squares [15] algorithm, which solves the estimating equation

$$\sum_i \frac{\partial \mu(t_i)}{\partial \beta} v_i^{-1}(\mu(t_i))\{x_i - \mu(t_i)\} = 0. \tag{4}$$

In cases where there is a jump discontinuity between the two functions μ_1 and μ_2, we can simply fit two separate regressions for $t < \tau$ and $t \geq \tau$, and estimate the scale parameter ϕ jointly. Denote by $\hat{\beta}(\tau)$ the estimate we obtain at this step.

To estimate τ, we compute the **profile** (quasi-)likelihood for τ by substituting $\hat{\beta}(\tau)$ in the quasi-likelihood function. Or, as a general approximation, we can use the Pearson's χ^2 statistic (*see* **Chi-square Tests**) as the objective function

$$Q(\tau) = \sum_i v_i^{-1}(\hat{\mu}(t_i))\{x_i - \hat{\mu}(t_i)\}^2, \tag{5}$$

where $\hat{\mu}(t)$ is evaluated at $\hat{\beta}(\tau)$. Thus, $\hat{\tau} \equiv \text{argmin}_\tau Q(\tau)$, which is a simple one-dimensional minimization problem, and $\hat{\beta} \equiv \hat{\beta}(\hat{\tau})$.

Several inferential issues are unresolved in general. One is the distribution of $Q(\hat{\tau})$, which is necessary to decide if the change at $\hat{\tau}$ is real. Theoretically, this is available under normal or **binomial** assumptions for a simple level-change model [9, 11, 13, 19, 21]. Secondly, there is no simple distribution theory for $\hat{\tau}$. Standard **large-sample theory** does not apply since the parameter is not regular. For both of these problems, we might use the permutation test (*see* **Randomization Tests**) or the bootstrap. Gibbs sampling has also been used for the Bayesian approach [1] (*see* **Markov Chain Monte Carlo**). With regards to the inference on $\hat{\beta}$, it is common to make it conditional on the observed value of $\hat{\tau}$,

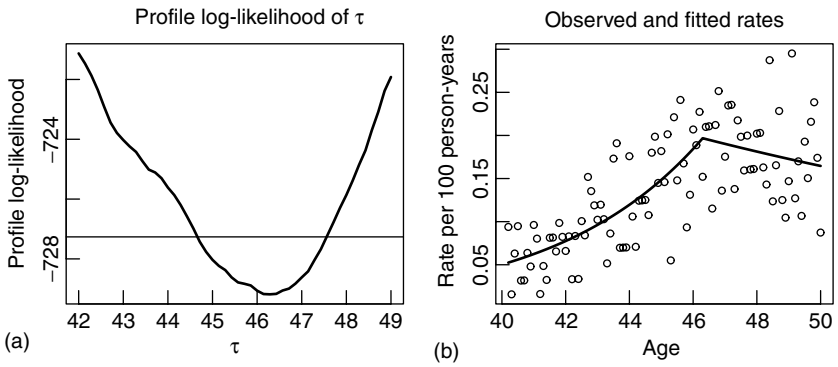

Figure 2 (a) Profile log-likelihood of the change parameter τ; (b) The observed and fitted values of breast cancer incidence

though it ignores the extra variability in the estimation of the change point.

In our example, we will assume that the observed number of breast cancer at each age group $y(t)$ is Poisson with mean $\mu(t) = N(t)\lambda(t)$, where $N(t)$ is the number of **person-years** at age t, and the rate parameter is given by

$$\log \lambda(t) = \begin{cases} \beta_{01} + \beta_{11}(t - 40) & \text{for } t < \tau \\ \beta_{02} + \beta_{12}(t - \tau) & \text{for } t \geq \tau. \end{cases} \quad (6)$$

Because we want $\lambda(t)$ to be continuous, we must have

$$\beta_{02} = \beta_{01} + \beta_{11}(\tau - 40). \quad (7)$$

Figure 2(a) shows the profile log-likelihood of τ, with a minimum at $\hat{\tau} = 46.3$ (the nominal likelihood-based 95% confidence interval is $44.7 < \tau < 47.6$.) The estimated parameters are

$$\begin{aligned} \hat{\beta}_{01} &= -7.60(\text{se} = 0.12) \\ \hat{\beta}_{11} &= 0.22(\text{se} = 0.025) \\ \hat{\beta}_{12} &= -0.05(\text{se} = 0.039). \end{aligned} \quad (8)$$

The estimated dispersion parameter is

$$\hat{\phi} = \frac{1}{99 - 4} \sum_t \frac{(y(t) - \hat{\mu}(t))^2}{\hat{\mu}(t)} = 0.76, \quad (9)$$

showing some underdispersion.

As a concluding remark, we note that there is a strong similarity between the most general change-point problems, involving, say, a continuous change and piecewise continuous functions, and general **nonparametric regression problems** for nonsmooth functions [8]. Recent advances in the latter, using, for example, the wavelet techniques, will produce natural competitors against the current nonsequential change-point analysis.

References

[1] Barry, D. & Hartigan, J.A. (1993). A Bayesian analysis for change point problems, *Journal of the American Statistical Association* **88**, 309–319.

[2] Box, G.E.P. & Tiao, G.C. (1965). A change in level of a non-stationary time series, *Biometrika* **52**, 181–192.

[3] Csörgo, M. & Horváth, L. (1988). Nonparametric methods for changepoint problems, *Handbook of Statistics*, Vol. 7, Oxford, North-Holland, pp. 403–425.

[4] Feder, P.I. (1975). On asymptotic distribution theory in segmented regression problem – identified case, *Annals of Statistics* **3**, 49–83.

[5] Fu, Y.-X. & Curnow, R.N. (1990). Maximum likelihood estimation of multiple change points, *Biometrika* **77**, 563–573.

[6] Gallant, A.R. & Fuller, W.A. (1973). Fitting segmented regression model whose joint points have to be estimated, *Journal of the American Statistical Association* **68**, 144–147.

[7] Gosh, J.K. & Joshi, S.N. (1992). On the asymptotic distribution of an estimate of the change point in a failure rate, *Communication in Statistics, Series A* **21**, 3571–3588.

[8] Hall, P. & Titterington, D.M. (1992). Edge-preserving and peak-preserving smoothing, *Technometrics* **34**, 429–440.

[9] Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location, *Journal of the American Statistical Association* **72**, 180–186.

[10] Hinkley, D.V. (1969). Inference about intersection in two-phase regression, *Biometrika* **56**, 495–504.

[11] Hinkley, D.V. & Hinkley, E.A. (1970). Inference about the change point in a sequence of binomial random variables, *Biometrika* **57**, 477–488.

[12] Hinkley, D.V. & Schechtman, E. (1987). Conditional bootstrap methods in the meanshift model, *Biometrika* **74**, 85–93.

[13] Krisnaiah, P.K. & Miao, B.Q. (1988). Review about estimation of change points, *Handbook of Statistics*, Vol. 7, Oxford, North-Holland, pp. 375–402.

[14] Liang, K.-Y., Self, S.G. & Liu, X. (1990). The Cox proportional hazards model with change point: an epidemiologic application, *Biometrics* **46**, 783–793.

[15] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Nodels*. Chapman & Hall, London.

[16] Page, E.S. (1954). Continuous inspection schemes, *Biometrika* **41**, 100–114.

[17] Page, E.S. (1955). A test for a change in a parameter occurring at an unknown point, *Biometrika* **42**, 523–527.

[18] Shaban, S.A. (1980). Change point problem and two-phase regression: an annotated bibliography, *International Statistical Review* **48**, 83–93.

[19] Srivastava, M.S. & Worsley, K.J. (1986). Likelihood ratio tests for a change in the multivariate normal mean, *Journal of the American Statistical Association* **81**, 199–205.

[20] Vostrikova, L.Ju. (1981). Detecting disorder in multidimensional random processes, *Soviet Mathematics Doklady* **24**, 55–59.

[21] Worsley, K.J. (1979). On the likelihood ratio test for a shift in location of normal populations, *Journal of the American Statistical Association* **74**, 365–367.

YUDI PAWITAN

Chaos Theory

Chaos derives from the Greek word $\chi\alpha o\varsigma$, from which one also obtains the English word "gas". Various probabilistic models have been proposed for the motion of gas molecules, such as the Ehrenfest **Markov chain** model for the diffusion of gas through a membrane [10]. However, in idealized situations – and even some practical ones – to which such models are applied, all motion is strictly deterministic. Hence, if one knew complete information about the momentum, energy, fields, etc. of the system in question, then, in principle, one could predict perfectly the state of the system at any future instant. However, in practice, this cannot be achieved. Now the motion of gas particles is not necessarily chaotic, in the strict sense of the word, but it provides a useful analogy. In loose terms, a system is chaotic if it is governed by a set of deterministic equations, but displays erratic, apparently random, behavior. Whilst one can attempt to provide a stochastic model for a chaotic system, one invariably obtains more substantive detail by considering the deterministic component and its contribution to the erratic nature of the observations.

Historically, the theory of chaos has been developed in the deterministic scenario, with Poincaré accredited with much of the foundation work [19], although recent statistical applications have endeavored to merge stochastic variation, or noise, into this setting. Consider first strictly deterministic chaos.

The phenomenon of chaos is produced by a function, possibly multivariable, either in discrete time, where the function comprises one or more linked difference equations, or in continuous time, where it comprises linked differential equations. Chaos is the erratic nature of the trajectories produced in the former case by iteration of the function, and in the latter by evolution of the differential equations. Chaos in continuous time must comprise at least three state variables, for otherwise trajectories would intersect and would thus not provide a unique solution at the point of intersection; however, discrete time chaos can exist in any Euclidean dimension.

There is not absolute agreement on a strict mathematical definition of chaos, although the following, reported fairly loosely, encapsulates most ideas. It follows the exposition given by Falconer [11]. Consider a function f defined on a domain D with an m-fold iterate (composition) denoted $f^{(m)}$. With this notation one can identify special kinds of points $x \in D$; for example, if $f(x) = x$, then x is a fixed point of f, and if $f^{(m)}(x) = x$ for the least $m > 1$, then x is a periodic point of order m. An attractor for f is the minimal compact set $A \subseteq D$ such that, for each $x \in D$, the set of all $f^{(m)}(x)$ which come arbitrarily close to A is nonnull. The chaotic nature, or otherwise, of f is characterized in terms of its attractor. In particular, the attractor must be such that (i) some point in the attractor produces a trajectory under f which is dense ("filling") in the attractor; (ii) the periodic points of the attractor are dense; and (iii) given any point on the attractor, one can find another arbitrarily close point on the attractor such that the two associated trajectories diverge sufficiently fast [22]. Respectively, these properties are (i) that the attractor is nondecomposable; (ii) that the attractor has some semblance of regularity; and (iii) that there is sensitive dependence upon initial conditions. In practical terms it is usually only possible to check condition (iii). Sensitive dependence upon initial conditions can be identified in many areas, from weather forecasting to tossing a coin: in the latter, the greater the angular spin and upward thrust on the coin the more difficult it is to predict the face it will show. Sensitive dependence upon initial conditions can be quantified in terms of an invariant of a dynamic system known as the Lyapunov exponent(s). Under an eigendecomposition, there is one such exponent for each Euclidean direction in which the map evolves, and, by invoking Taylor expansions, one can demonstrate that the Lyapunov exponent is the exponential rate of separation of initially nearby trajectories in the short term. (Since a chaotic map evolves in a bounded space, the separation of trajectories cannot continue indefinitely.)

If a map is chaotic, then often its attractor will be fractal-like: its fine, complex structure is not always "filling" in each Euclidean direction, and hence its dimension is not always an integer. Estimates of fractional dimension are obtained using box counting methods, amongst others. The analysis of fractal surfaces and the estimation of fractional dimension comprise research and applications which are closely allied with the study of chaos; see, for instance, Davies & Hall [9]. These topics are beyond the scope of this article.

Deterministic chaos possesses features that have stochastic interpretations. One can define a probability measure, ν, on the set D upon which the chaotic map f is defined by denoting the measure of a subset $C \subseteq D$ as the limiting proportion of time spent by an arbitrary trajectory in C (continuous time) or the proportion of its iterates therein (discrete time); that is, $\nu(C) = \lim_{m\to\infty} m^{-1}\#\{f^{(j)}(X_0) \in C; 1 \leq j \leq m\}$, where "#" denotes the cardinality of a set. Here X_0 is an arbitrarily chosen initial condition. For a wide class of dynamical systems (called "Axiom A" systems), ν does not depend on the choice of X_0; that is to say, the invariant measure is a "generic" feature of a dynamic system (which is the analog for the phrase "almost sure" in a probabilistic setting). The existence of an absolutely continuous invariant measure is known only for certain classes of maps (essentially, those that are one-dimensional and whose piecewise derivatives exceed unity in absolute value). If the dynamics of the system do not change as time evolves, then one has the equivalent of a strictly **stationary** time series. If one regarded the output of a chaotic dynamical system as a stochastic **time series**, then one could obtain its **autocorrelation function**, amongst other features. Whilst the deterministic interpretation of the variables would lead one to think that the variables are highly correlated, via the relationship induced by the map f, sensitive dependence upon initial conditions actually produces autocorrelation functions which can be zero at virtually all nonzero lags (within the stochastic interpretation of the series). Hence a chaotic series, produced by a strong deterministic relationship, can behave like an apparently purely random time series. It is for this reason, among other more technical ones, that random number generators in computers are based on chaotic maps (*see* **Pseudo-random Number Generator**).

One could point to papers in 1990, such as by Bartlett [3] and Wolff [27], which were among the first formal statistical treatments of chaos in the statistical literature, and to papers in 1992, e.g. [26], and [5], and [7], which were the first fruits of the initial activity. Of course, the scientific community more widely has known about chaos for much longer, and even statisticians, such as Tong & Lim [25], made observations of it before the 1990s. Such observations were in the context of **nonlinear time series** models, which progressively came to the fore from the 1970s [16], and which identified the relationship between modes of behavior of various models for different parameter values, charting the steady state, monotonic, periodic, limit cycle or chaotic behavior of realizations of the model; again see, for example, Tong & Lim [25]. Tong [24] draws together the theory of nonlinear time series and statistical aspects of dynamic systems, which also appears in an as yet unpublished book by K.S. Chan & H. Tong.

It is at this juncture that one is able to pose a more realistic setting for chaos by incorporating stochastic noise into the map (*see* **Noise and White Noise**). Just as time domain models for time series usually comprise a deterministic component and a stochastic component, one can formulate nonlinear autoregressions, for instance where the deterministic component is a chaotic map and noise is added into the system. $X \to f(X) + \varepsilon$, where ε represents an independent and identically distributed sequence of **random variables**. Many other formulations with noise are possible. Under this representation, the generic properties of the deterministic map f are unlikely to be related to the corresponding ones of the noisy chaotic map (for example, computations of Lyapunov exponents, fractal dimensions, and invariant measures). Moreover, interpretation of those quantities in the noisy case is difficult; for in the deterministic case there exists a geometric interpretation for them which does not make sense when noise infects the deterministic dynamics.

Parametric modeling of chaotic systems is almost impossible. Since a chaotic map cannot be linear, the family of possible maps for chaos is vast and unchartable. Unless one has considerable substantive information about the map, such as intimate knowledge of the physics governing the system, then parametric modeling usually is out of the question. However, in such cases where it is possible there has been some progress. Geweke [12] considers a realization of a deterministic chaotic map in which the series is "blurred" with an independent noise sequence. A method of parameter estimation using **maximum likelihood** in the presence of an extremely irregular **likelihood** function is demonstrated. The more realistic case of system noise is considered by Berliner [4], as described in the previous paragraph, who sets out a **Bayesian** approach to parametric model fitting. Assuming parametric knowledge of the map, one may adapt nonparametric kernel **density estimation** techniques to obtain the density

of the invariant measure of a given map, assuming that it exists and is absolutely continuous, as investigated by Hall & Wolff [15]. That technique further validates the block **bootstrap** for chaotic time series.

In fact, nonparametric modeling is where the best progress has been made. Yao & Tong [30, 31] adapt the idea of Lyapunov exponents to the stochastic setting by replacing it with a probability that trajectories will diverge exponentially fast. They call estimates of that probability a *sensitivity measure*. In an earlier paper they draw on the Taylor expansion formulation of the Lyapunov exponent to produce forecasts of a dynamic system and to bound those forecasts in terms of the Lyapunov exponent. Further in the endeavor of forecasting, they adapt a **Kullback–Leibler** statistic to identify local regions from which accurate k-step-ahead forecasts can be made and regions from which sufficient accuracy of forecasts does not exist. This is related to an adaptation of Lyapunov exponents to find local regions of relatively large or small exponential divergence of trajectories [28], which has also been studied in a local-time sense [18].

Estimates of dimension for noisy chaos [21] and interpretations of fractal properties for general time series models [8] have also been presented.

Adaptations have been made of devices used in the study of dynamic systems, such as the application of the correlation integral, used for estimating fractional dimension and, more generally, for determining the presence of chaos, as established by Grassberger & Procaccia [13], and also as a test statistic for independence in a time series, such as in Brock et al. [6] and Wolff [29]. Thus the epiphany of chaos in statistical realms has done much to complement the Box–Jenkins methodology.

Direct applications in human biology are few, although consideration of dynamic aspects have led to novel approaches in modeling. Three particular applications have received prominence in recent years. Of course, the classical discussion of chaos in biological systems *per se* is that of May [17], where simple one-dimensional maps governing biological populations were found mathematically and in reality to exhibit chaos. For various plain biological and social reasons, human population series would not be expected to display such behavior.

Earlier works on measles epidemics, such as those of Bartlett [3] and Schaffer [20], were scrutinized by Grenfell [14] and others of his co-workers subsequently in regard to chaotic dynamics in measles epidemics. Sugihara & May [23] exhibited a **cross-validation** device to distinguish between chaotic behavior (claimed to exist in measles epidemics and thought to be governed by a six-dimensional map) and purely stochastic behavior (claimed to exist in mumps and chicken-pox epidemics). It is initially plausible that childhood diseases might be chaotic. Consider a pendulum with a metallic weight: given a small displacement it will follow regular simple harmonic motion, but if it is placed near an object of the same magnetic polarity it can trace out a chaotic trajectory. This kind of forced oscillation, causing chaos, is present in childhood diseases, in that school terms and holidays "force" large numbers of potentially infectious children together and apart, respectively, and that forcing is superimposed on the natural oscillation of the epidemic. Subsequent scrutiny of epidemiologic time series raised issues about large amounts of noise present in the series, and the greater likelihood of limit cycles rather than chaotic behavior lying at the center of the dynamics (*see* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**; **Infectious Disease Models**).

Babloyantz [1, 2], among many others, has examined physiologic dynamics and found evidence of chaos. Along the lines of the above discussion, the forced oscillations of the heart were studied, and it was found that a healthy heart is chaotic! In its usual beating mode the differenced series of times between corresponding parts of the heart cycle shows evidence of low-dimensional chaos. What is more, at the onset of myocardial infarction (heart attack), chaos gives way to limit cycle patterns. There is a large research industry into this topic, as well as in infant respiration and analysis of electroencephalograms, to name but two others in the realm of physiological dynamics (*see* **Mathematical Biology, Overview**; **Clinical Signals**).

References

[1] Babloyantz, A. (1988). Is the normal heart a periodic oscillator?, *Biological Cybernetics* **58**, 203–211.

[2] Babloyantz, A. (1991). Predictability of human EEG – a dynamic approach, *Biological Cybernetics* **64**, 381–391.

[3] Bartlett, M.S. (1990). Chance or chaos? (with discussion), *Journal of the Royal Statistical Society, Series A* **153**, 321–347.

[4] Berliner, L.M. (1991). Likelihood and Bayesian prediction of chaotic systems, *Journal of the American Statistical Association* **86**, 938–952.

[5] Berliner, L.M. (1992). Statistics, probability and chaos (with discussion), *Statistical Science* **7**, 69–90.

[6] Brock, W.A., Dechert, W.D. & Scheinkman, J.A. (1986). A test for independence based on the correlation integral, *Technical Report*. University of Wisconsin, Madison.

[7] Chatterjee, S. & Yilmaz, M.R. (1992). Chaos, fractals and statistics (with discussion), *Statistical Science* **7**, 49–68.

[8] Cutler, C.D. (1994). A theory of correlation dimension for stationary time series (with discussion), *Philosophical Transactions of the Royal Society of London, Series A* **348**, 343–355.

[9] Davies, S. & Hall, P. (1998). Fractal analysis of surface roughness using spatial data, *Journal of the Royal Statistical Society, Series B*, to appear.

[10] Ehrenfest, P. (1912). The conceptual foundations of the statistical approach in mechanics, in *Encyclopedia of Mathematical Sciences* (in German). English translation by M.J. Moravcsik. Dover, New York (1959).

[11] Falconer, K. (1990). *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, Chichester.

[12] Geweke, J. (1993). Inference and forecasting for chaotic non-linear time series, in *Non-linear Dynamics and Evolutionary Economics*, P. Chen & R. Day, eds. Oxford University Press, Oxford.

[13] Grassberger, P. & Procaccia, I. (1983). Characterization of strange attractors, *Physics Review Letters* **50**, 346–349.

[14] Grenfell, B.T. (1992). Chance and chaos in measles dynamics (with discussion), *Journal of the Royal Statistics Society, Series B* **54**, 383–398.

[15] Hall, P. & Wolff, R.C.L. (1995). Properties of invariant distributions and Lyapunov exponents for chaotic logistic maps, *Journal of the Royal Statistical Society, Series B* **57**, 439–452.

[16] Jones, D.A. (1978). Nonlinear autoregressive processes, *Proceedings of the Royal Society of London, Series A* **360**, 71–95.

[17] May, R.M. (1975). Biological populations obeying difference equations: stable points, stable cycles and chaos, *Journal of Theoretical Biology* **51**, 511–524.

[18] Nychka, D., Ellner, S., McCaffrey, D. & Gallant, A.R. (1992). Finding chaos in noisy systems (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 399–426.

[19] Poincaré, H. (1885). Sur l'équilibre d'une masse fluide animée d'un mouvement de rotation, *Acta Mathematica* **7**, 259–380.

[20] Schaffer, W.M. (1985). Can non-linear dynamics elucidate mechanisms in ecology and epidemiology?, *IMA Journal of Mathematics with Applications in Medicine and Biology* **2**, 221–252.

[21] Smith, R.L. (1992). Estimating dimension in noisy chaotic time series (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 329–351.

[22] Sparrow, C. (1982). *The Lorenz Equations: Bifurcations, Chaos and Strange Attractors*. Springer-Verlag, New York.

[23] Sugihara, G. & May, R.M. (1990). Non-linear forecasting as a way of distinguishing chaos from measurement error in a time series, *Nature* **344**, 734–741.

[24] Tong, H. (1990). *Non-linear Time Series: A Dynamical Systems Approach*. Oxford University Press, Oxford.

[25] Tong, H. & Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 245–292.

[26] Tong, H. & Smith, R.L., eds (1992). Royal Statistical Society Meeting on Chaos, *Journal of the Royal Statistical Society, Series B* **54**, 301–474.

[27] Wolff, R.C.L. (1990). A note on the behavior of the correlation integral in the presence of a time series, *Biometrika* **77**, 689–697.

[28] Wolff, R.C.L. (1992). Local Lyapunov exponents: looking closely at chaos (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 353–371.

[29] Wolff, R.C.L. (1994). Independence in time series: another look at the BDS test (with discussion), *Philosophical Transactions of the Royal Society of London, Series A* **348**, 383–395.

[30] Yao, Q. & Tong, H. (1992). Quantifying the influence of initial values on non-linear prediction, *Journal of the Royal Statistical Society, Series B* **56**, 701–725.

[31] Yao, Q. & Tong, H. (1994). On prediction and chaos in stochastic systems (with discussion), *Philosophical Transactions of the Royal Society of London, Series A* **348**, 357–369.

(*See also* **Brownian Motion and Diffusion Processes**; **Markov Processes**; **Stochastic Processes**)

R.C.L. WOLFF

Characteristic Function

The characteristic function has several important theoretical applications. It is used to derive distributions for sums and other linear combinations of independent **random variables**. It is also used to determine limiting distributions of statistics as sample sizes become infinitely large (*see* **Large-sample Theory**). The resulting asymptotic distributions often provide useful approximations for making inferences from finite samples when the exact finite sample distributions are too complicated to be conveniently evaluated. The characteristic function can also be used to derive **moments** of distributions.

The characteristic function of a (real-valued) random variable X with distribution function $F(x)$ is a complex-valued function defined as

$$\phi(t) = \mathrm{E}[\exp(itX)] = \int_{-\infty}^{\infty} \exp(itx)\,\mathrm{d}F(x), \quad (1)$$

where $i \equiv \sqrt{-1}$ and t is any real number. For a continuous distribution with density function $f(x)$,

$$\phi(t) = \int_{-\infty}^{\infty} \exp(itx) f(x)\,\mathrm{d}x,$$

and for a discrete distribution on the nonnegative integers,

$$\phi(t) = \sum_{k=0}^{\infty} \exp(ikt)\Pr(x = k).$$

Introductory textbooks on statistical theory often use the **moment generating function** instead of the characteristic function to avoid the introduction of complex arithmetic. They have many of the same uses, but the characteristic function provides more general results because it exists for any distribution. It is absolutely continuous on the real line and satisfies (i) $\phi(0) = 1$, (ii) $|\phi(t)| \leq 1$, and (iii) $\phi(-t)$ is the complex conjugate of $\phi(t)$.

We may obtain the rth moment of X about zero from $\phi^{(r)}(0)$, the rth derivative of $\phi(t)$ evaluated at $t = 0$. In particular, if X has finite moments up to some order n, then $\phi(t)$ has continuous derivatives up to order n and

$$\mu_r = \mathrm{E}(X^r) = i^r \phi^{(r)}(0), \quad r = 1, \ldots, n. \quad (2)$$

Equivalently, μ_r is the coefficient of the rth term in the expansion

$$\phi(t) = 1 + \sum_{r=1}^{\infty} \mu_r \frac{(it)^r}{r!}. \quad (3)$$

The rth *cumulant* of X, denoted by κ_r, is the coefficient of the rth term in the corresponding expansion of the natural logarithm of the characteristic function, i.e.

$$\ln[\phi(t)] = \sum_{r=1}^{\infty} \kappa_r \frac{(it)^r}{r!}. \quad (4)$$

Moments of X about zero are obtained from cumulants as

$$\begin{aligned}
\mu_1 &= \kappa_1, \\
\mu_2 &= \kappa_2 + \kappa_1^2, \\
\mu_3 &= \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3, \\
\mu_4 &= \kappa_4 + 3\kappa_2^2 + 4\kappa_1\kappa_3 + 6\kappa_1^2\kappa_2 + \kappa_1^4.
\end{aligned}$$

Cumulants are also called *semi-invariants* because adding a constant to X does not affect the value of κ_r for $r \geq 2$. The term "cumulant" is motivated by the property that the rth cumulant of the sum of independent random variables is equal to the sum of the corresponding cumulants of the individual random variables for any $r \geq 1$ for which the cumulants exist. Stuart & Ord [7] provide more information on the computation and uses of cumulants and additional formulas relating cumulants to various types of moments.

Inversion formulas allow us to recover density functions and discrete probability distributions from characteristic functions. If $\phi(t)$ is absolutely integrable, i.e. $\int_{-\infty}^{\infty} |\phi(t)|\,\mathrm{d}t$ is finite, then $\phi(t)$ uniquely determines an absolutely continuous distribution with a bounded and uniformly continuous density function given by the formula

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx)\phi(t)\,\mathrm{d}t. \quad (5)$$

For discrete distributions on the nonnegative integers, we have

$$\Pr(X = k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itk)\phi(t)\,\mathrm{d}t. \quad (6)$$

Characteristic functions have important applications in the determination of distributions for linear combinations of independent random variables. For example, if $X_1, X_2, \ldots, X_n$ are independent random variables with respective characteristic functions $\phi_1(t), \phi_2(t), \ldots, \phi_n(t)$, then the characteristic function of $X_1 + X_2 + \cdots + X_n$ is the product

$$\phi(t) = \phi_1(t)\phi_2(t)\ldots\phi_n(t). \tag{7}$$

We complete the task by applying an inversion formula to $\phi(t)$ or simply recognizing the distribution function uniquely determined by $\phi(t)$. The characteristic function for $X_1 - X_2$ is

$$\phi(t) = \phi_1(t)\phi_2(-t). \tag{8}$$

Characteristic functions were originally introduced as a mechanism for deriving limiting distributions of sequences of random variables (*see* **Convergence in Distribution and in Probability**). Important applications include the derivation of **central limit theorems** for establishing the limiting **normal** (Gaussian) distributions of sample means or estimators of model parameters as sample sizes are increased, and the derivation of **chi-square distributions** for sums of squares and **goodness-of-fit** tests. The basic result is that if $\{X_n\}$ is a sequence of random variables such that X_n has distribution function $F_n(x)$ and characteristic function $\phi_n(t)$, then the pointwise convergence of $\phi_n(t)$ to a function $\phi(t)$ that is continuous at $t = 0$ implies that $F_n(x)$ converges to $F(x)$, the unique distribution function determined by $\phi(t)$, at all continuity points of $F(x)$. The converse of this result is also true, i.e. the convergence of the sequence $\{F_n(x)\}$ to a continuous distribution function $F(x)$ implies uniform convergence of $\{\phi_n(t)\}$ to $\phi(t)$ in any finite interval of t values as $n \to \infty$.

As an illustration, consider the limiting behavior of a sequence of Bernoulli random variables. Let X_n denote the number of successful outcomes in a series of n independent and identical trials where each trial has probability p of producing a successful outcome. Then, X_n has a **binomial distribution** with probability function

$$\Pr(X = k) = \binom{n}{k} p^r (1-p)^{n-r}, \quad \text{for } k = 0, 1, \ldots, n.$$

The characteristic function for the binomial distribution is

$$\begin{aligned}\phi(t) &= \sum_{k=0}^{n} \exp(itk) \binom{n}{k} p^k (1-p)^{n-k} \\ &= [1 - p + p\exp(it)]^n \\ &= \{1 + p[\exp(it) - 1]\}^n.\end{aligned}$$

Now consider a sequence of binomial random variables $\{X_n, n = 1, 2, \ldots\}$, where X_n is the number of successful outcomes in n independent trials with probability of success $p = \lambda/n$ on any single trial. In this scenario the probability of success on a single trial becomes smaller as the number of trials increases in such a way that the expected number of successes,

$$\mathrm{E}(X_n) = np = \lambda,$$

remains constant. Then, the characteristic function for X_n is

$$\phi_n(t) = \left[1 + \frac{\lambda[\exp(it) - 1]}{n}\right]^n,$$

and using the result that $(1 + a/n)^n \to \mathrm{e}^a$ as $n \to \infty$, it is easily seen that as $n \to \infty$,

$$\phi_n(t) \to \exp\{\lambda[\exp(it) - 1]\}$$

which is the characteristic function of a **Poisson** random variable with expectation λ. This shows that as the number of trials becomes large and the probability of success on any single trial becomes small, the binomial distribution for the number of successful outcomes approaches a Poisson distribution with the same mean.

When p remains constant as the number of trials is increased, X_n increases without bound and has no limiting distribution, but a central limit theorem can be used to show that the standardized count,

$$Z_n = \frac{X_n - np}{[np(1-p)]^{1/2}},$$

has a limiting standard normal (Gaussian) distribution. The characteristic function for the standard normal distribution is

$$\begin{aligned}\phi(t) &= \int_{-\infty}^{\infty} \exp(itx) \frac{1}{(2\pi)^{1/2}} \exp[-(1/2)x^2]\,\mathrm{d}x \\ &= \exp[-(1/2)t^2]. \end{aligned} \tag{9}$$

Consequently, the characteristic function for Z_n,

$$\phi_n(t) = \exp\{-itnp/[np(1-p)]^{1/2}\} \times [1 - p + p\exp\{it/[np(1-p)]^{1/2}\}]^n,$$

must converge to (9) for all t in any finite interval. Note that the effect on $\phi_n(t)$ of transforming X_n to $(X_n - \mu)/\sigma$ is to replace t by t/σ and multiply the result by $\exp(-it\mu/\sigma)$.

Stuart & Ord [7] provide a good introduction to the derivation and uses of characteristic functions. Additional properties of characteristic functions are reviewed by Laha [1]. For proofs and more precise mathematical statements of these results and additional developments, see Lukács [3, 4], Laha & Rohatgi [2], and Ramachandran [6]. Applications are discussed by Lukács & Laha [5] and Laha & Rohatgi [2].

References

[1] Laha, R.G. (1982). Characteristic functions, in *Encyclopedia of Statistical Sciences*, Vol. 1, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 415–422.

[2] Laha, R.G. & Rohatgi, V.K. (1979). *Probability Theory*. Wiley, New York.

[3] Lukács. E. (1970). *Characteristic Functions*, 2nd Ed. Griffin, London/Hafner, New York.

[4] Lukács, E. (1983). *Developments in Characteristic Function Theory*. Macmillan, New York.

[5] Lukács, E. & Laha, R.G. (1964). *Application of Characteristic Functions*. Griffin, London/Hafner, New York.

[6] Ramachandran, B. (1967). *Advanced Theory of Characteristic Functions*. Statistical Publishing Society, Calcutta.

[7] Stuart, A. & Ord, J.K. (1987). *Kendall's Advanced Theory of Statistics*, 5th Ed., Vol. 1. Griffin, London.

KENNETH KOEHLER

Chemometrics

Chemometrics is usually defined as the application of mathematical and statistical methods to problems in chemistry. Although most such applications involve simple statistical techniques, it is the use of multivariate statistical methods that is popularly associated with the term chemometrics (*see* **Multivariate Analysis, Overview**). These multivariate methods include some that will be familiar to many biostatisticians, such as **principal components analysis** and **discriminant analysis**, as well as some others that will not. The text by Massart et al. [5] gives a good coverage of chemometrics in its wider sense. Here we concentrate on the multivariate aspects.

One important area of application is to multivariate **calibration** problems [1, 4]. Here we wish to calibrate a rapid, usually indirect, measurement of some quantity against measurement by a reference method y. The basic measurement x produced by the indirect method is multivariate, possibly highly so. For example, the rapid method may be spectroscopic, in which case it is not untypical to have as x a spectrum measured at 100 or even 1000 wavelengths. Typically, one is faced with estimating a calibration equation $y = f(x)$ from around 50 training samples on which both x and y are measured, with **multiple regression** ruled out by the dimensionality of x and great care needed to avoid overfitting the limited data. Spectroscopic examples in clinical and pharmaceutical chemistry are common. A 1995 conference [2] included a dozen papers in this area, with the rapid spectroscopic measurement of blood parameters as one popular theme.

Methods that construct a prediction equation by linear regression of y on the scores of the training samples on a number of factors derived from the original predictor variables are widely used in such problems (*see* **Reduced Rank Regression**). In principal component regression we carry out a preliminary **principal components analysis** of the original predictor variables, retaining only a small number of components that explain most of the variance in x. The second step is to regress y on the principal component scores for these components, which are linear combinations of the original measurements. In partial least squares regression (PLSR, or just PLS) we also construct new predictor variables as linear combinations of the original ones, but now we choose the combinations to maximize the covariance between the constructed variables and y. **Cross-validation** is commonly used with both of these methods to select the appropriate number of factors to include in the equation. Frank & Friedman [3] discuss these techniques and compare them with other **shrinkage** methods such as **ridge regression**.

Another application with importance in the field of medicine is the study of quantitative structure–activity relationships, universally referred to by its acronym QSAR. The idea of QSAR is to try to relate quantitative physical and chemical descriptions of molecules to their biological activity. Examples of descriptors range from simple ones such as the solubility of the molecule in water, to complex and high-dimensional ones from computational chemistry, based for example on the molecule's electrostatic potential calculated on a lattice of points surrounding it. Given a database of molecules in which some desired biological activity – such as being an effective drug or pesticide – is present to a greater or lesser extent, the general problem is to relate this activity to descriptors of the molecules. Such a relationship would guide the drug designer in synthesizing new molecules or deciding, on the basis of relatively easily measured descriptors, which existing molecules to subject to expensive biological testing. When high-dimensional descriptors are used the statistical problems are similar to those in the multivariate calibration problem, and many of the same chemometric tools used in multivariate calibration have also been applied to QSAR. Stone & Jonathan [6, 7] review statistical aspects of QSAR and provide a thoughtful commentary on the use of these tools in general.

References

[1] Brown, P.J. (1993). *Measurement, Regression and Calibration*. Oxford University Press, Oxford.

[2] Davies, A.M.C. & Williams, P.C., eds. (1996). *Near Infrared Spectroscopy: the Future Waves. Proceedings of the Seventh International Conference on Near Infrared Spectroscopy, Montréal, Canada, August 6–11, 1995*. NIR Publications, Chichester.

[3] Frank, I.E. & Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics* **35**, 109–148.

[4] Martens, H. & Næs, T. (1989). *Multivariate Calibration*. Wiley, New York.

[5] Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y. & Kaufman, L. (1988). *Chemometrics: A Textbook*. Elsevier, Amsterdam.

[6] Stone, M. & Jonathan, P. (1993). Statistical thinking and technique for QSAR and related studies, part I: general theory, *Journal of Chemometrics* **7**, 455–475.

[7] Stone, M. & Jonathan, P. (1994). Statistical thinking and technique for QSAR and related studies, part II: specific methods, *Journal of Chemometrics* **8**, 1–20.

TOM FEARN

Chi-square Distribution; Properties

This article complements the introductory article on the **chi-square distribution** by presenting proofs of its derivation as the distribution of the sum of squares of standard normal deviates. Various properties of the distribution are derived. The second section of the article deals with the noncentral chi-square distribution, which arises when the component normal deviates, whose squares are summed, have nonzero means. This is an important result as it defines the **power** of standard chi-square tests.

The Central Chi-Square Distribution

Theorem 1. Let $X_1, X_2, \ldots, X_n$ be independent with the **standard normal** $\mathcal{N}(0, 1)$ density

$$f_{X_k}(x_k) = \frac{\mathrm{e}^{-x_k^2/2}}{\sqrt{2\pi}}, \quad -\infty < x_k < \infty,$$
$$k = 1, 2, \ldots, n.$$

Then

$$U_n = \sum_{k=1}^{n} X_k^2$$

has the central chi-square distribution $\chi_n^2(0)$ with density

$$f_{U_n}(s) = \frac{s^{(n/2)-1}\mathrm{e}^{-s/2}}{\Gamma(n/2)2^{n/2}}.$$

for $0 < s < \infty$, and 0 elsewhere.

Proof. For $\alpha > 0$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$, $\Gamma(1/2) = \sqrt{(\pi)}$, and $\Gamma(m + 1) = m!$ for integer $m \geq 0$.

We use complete induction on n and write $U_n = U_{(n-1)} + U_1$, where $U_1 = X_1^2$ and $U_{(n-1)} = \sum_{k=2}^{n} X_k^2$. Then forming the convolution using independence of U_1 and $U_{(n-1)}$

$$\begin{aligned} f_{U_n}(s) &= \int_0^s f_{U_{(n-1)}}(s - y) f_{U_1}(y)\,\mathrm{d}y \\ &= \int_0^s \frac{(s-y)^{(n-1)/2-1}\mathrm{e}^{-(s-y)/2}}{\Gamma((n-1)/2)2^{(n-1)/2}} \\ &\quad \times \frac{y^{1/2-1}\mathrm{e}^{-y/2}}{\Gamma(1/2)2^{1/2}}\,\mathrm{d}y \\ &= \frac{s^{n/2-1}\mathrm{e}^{-s/2}}{\Gamma(n/2)2^{n/2}} \\ &\quad \times \int_0^s \frac{\Gamma(n/2)(s-y)^{(n-1)/2-1}y^{1/2-1}}{\Gamma((n-1)/2)\Gamma(1/2)s^{n/2-1}}\,\mathrm{d}y \end{aligned}$$

Substituting $v = y/s$, we get

$$\frac{s^{n/2-1}\mathrm{e}^{-s/2}}{\Gamma(n/2)2^{n/2}} \int_0^1 \frac{\Gamma(n/2)v^{(n-1)/2-1}(1-v)^{1/2-1}}{\Gamma((n-1)/2)\Gamma(1/2)}\,\mathrm{d}v = \frac{s^{n/2-1}\mathrm{e}^{-s/2}}{\Gamma(n/2)2^{n/2}} \times 1$$

using the **beta**$((n-1)/2, 1/2)$ density integrates to 1. The parameter n of the $\chi_n^2(0)$ density is called the **degrees of freedom** parameter.

A proof using moment **generating functions** is as follows:

$$\begin{aligned} m_{X_k^2}(t) &= \mathrm{E}(\mathrm{e}^{tX_k^2}) = \int_{-\infty}^{\infty} \mathrm{e}^{tx^2}\frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x \\ &= (1-2t)^{-1/2}. \\ m_{U_n}(t) &= \prod_{k=1}^{n} m_{X_k}(t) = (1-2t)^{-n/2} \\ &= \int_0^{\infty} \mathrm{e}^{ts}\frac{s^{(n/2)-1}\mathrm{e}^{-s/2}}{\Gamma(n/2)2^{n/2}}\,\mathrm{d}s = \mathrm{E}(\mathrm{e}^{tU_n}) \end{aligned}$$

and we use the uniqueness of the moment generating function. The proof is similar using the uniqueness of the complex **characteristic function**,

$$\begin{aligned} \phi_{X_k^2}(t) &= \mathrm{E}(\mathrm{e}^{itX_k^2}) = \int_{-\infty}^{\infty} \mathrm{e}^{itx^2}\frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x \\ &= (1-2it)^{-1/2}. \\ \phi_{U_n}(t) &= \prod_{k=1}^{n} \phi_{X_k^2}(t) = (1-2it)^{-n/2} \\ &= \int_0^{\infty} \mathrm{e}^{its}\frac{s^{(n/2)-1}\mathrm{e}^{-s/2}}{\Gamma(n/2)2^{n/2}}\,\mathrm{d}s = \mathrm{E}(\mathrm{e}^{itU_n}), \end{aligned}$$

where $i = \sqrt{(-1)}$.

The chi-square distribution is a special case ($\alpha = n/2$, $\beta = 2$) of the **gamma** (α, β) distribution, which has density

$$f(x) = \frac{x^{\alpha-1}\mathrm{e}^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \quad \text{for } 0 < x < \infty,$$

and 0 elsewhere.

The incomplete gamma distribution function can be calculated from

$$G(x,a) = \int_0^\infty \frac{u^{a-1}\mathrm{e}^{-u}}{\Gamma(a)}\,\mathrm{d}u = \begin{cases} \dfrac{x^a\mathrm{e}^{-x}}{\Gamma(a)} \displaystyle\sum_{k=0}^{\infty} \frac{x^k}{(a+k)_{k+1}} & \text{for small } x \\ 1 - \dfrac{x^a\mathrm{e}^{-x}}{\Gamma(a)} c(x) & \text{for large } x \end{cases}$$

with $(a+k)_{k+1} = \prod_{j=0}^{k}(a+j)$ and the continued fraction

$$c(x) = 1/(x + a_0/(1 + a_1/(x + a_2/(1 + a_3/(x + \cdots/(x + a_{2k}/(1 + a_{2k+1}/(x + \cdots,$$

where $a_{2k} = k + 1 - a$, $a_{2k+1} = k + 1$ for $k = 0, 1, 2\ldots$ (see for example Abramowitz and Stegun [1, p. 263, 6.5.29, 6.5.31]). The gamma function $\Gamma(a)$ can be calculated accurately using Stirling's formula with the Binet function error term evaluated using the continued fraction of Jones and Thron [2, p. 350].

The cumulative distribution function (cdf) for the central $\chi_n^2(0)$ distribution is then

$$F_{U_n}(x) = P(U_n \le x) = G\left(\frac{x}{2}, \frac{n}{2}\right).$$

The cdf for the **Poisson**(λ) distribution can be determined from that of the chi-square cdf

$$\sum_{k=0}^{x} \frac{\lambda^k\mathrm{e}^{-\lambda}}{k!} = 1 - F_{U_m}(2\lambda),$$

where $m = 2(x+1)$ for $x = 0, 1, \ldots$.

The Noncentral Chi-square Distribution

Theorem 2. Let X_k be independent normal $\mathcal{N}(\mu_k, 1)$ for $k = 1, 2, \ldots, n$ with densities

$$f_{X_k}(x_k) = \frac{\mathrm{e}^{-(x_k-\mu_k)^2/2}}{\sqrt{2\pi}}, \quad -\infty < x_k < \infty.$$

The distribution of

$$U_n = \sum_{k=1}^{n} X_k^2$$

is the noncentral chi-square distribution $\chi_n^2(\delta_n^2)$, where $\delta_n^2 = \sum_{k=1}^{n} \mu_k^2$, with density

$$f_{U_n}(u) = \sum_{j=0}^{\infty} \mathrm{e}^{-\delta_n^2/2} \frac{(\delta_n^2/2)^j}{j!} \frac{u^{(2j+n)/2-1}\mathrm{e}^{-u/2}}{\Gamma((2j+n)/2)2^{(2j+n)/2}}$$

for $0 < u < \infty$, and 0 elsewhere.

Proof. Make the transformation

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} \frac{\mu_1}{\delta_n} & \frac{\mu_2}{\delta_n} & \frac{\mu_3}{\delta_n} & \cdots & \frac{\mu_n}{\delta_n} \\ \frac{\mu_2\mu_1}{\delta_2\delta_1} & \frac{-\delta_1^2}{\delta_2\delta_1} & 0 & \cdots & 0 \\ \frac{\mu_3\mu_1}{\delta_3\delta_2} & \frac{\mu_3\mu_2}{\delta_3\delta_2} & \frac{-\delta_2^2}{\delta_3\delta_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\mu_n\mu_1}{\delta_n\delta_{(n-1)}} & \frac{\mu_n\mu_2}{\delta_n\delta_{(n-1)}} & \frac{\mu_n\mu_3}{\delta_n\delta_{(n-1)}} & \cdots & \frac{-\delta_{(n-1)}^2}{\delta_n\delta_{(n-1)}} \end{pmatrix} \times \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix}.$$

In **matrix** notation $\mathbf{Z = AX}$, where $\mathbf{A}$ is the **orthogonal** $n \times n$ matrix above that satisfies $\mathbf{AA^T = A^TA = I_n}$, where $\mathbf{I_n}$ is the $n \times n$ identity matrix. It follows from the **multivariate normal distribution** for $\mathbf{X} : \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{I_n})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$ that $\mathbf{Z} : \mathcal{N}_n(\boldsymbol{\eta}, \mathbf{I_n})$ with $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\mu} = (\delta_n, 0, 0, \ldots, 0)^T$. Thus, $Z_1 : \mathcal{N}(\delta_n, 1)$, $Z_k : \mathcal{N}(0, 1)$ for $k = 2, 3, \ldots, n$ and they are independent. Also,

$$\sum_{i=1}^{n} Z_i^2 = \mathbf{Z^TZ} = \mathbf{X^TA^TAX} = \mathbf{X^TX} = \sum_{i=1}^{n} X_i^2 = U_n.$$

The density

$$\begin{aligned} f_{Z_1^2}(u_1) &= f_{Z_1}(\sqrt{u_1}) \left| \frac{\mathrm{d}\sqrt{u_1}}{\mathrm{d}u_1} \right| \\ &\quad + f_{Z_1}(-\sqrt{u_1}) \left| \frac{\mathrm{d}(-\sqrt{u_1})}{\mathrm{d}u_1} \right| \\ &= \frac{\frac{1}{2}u_1^{-1/2}\mathrm{e}^{(\sqrt{u_1}-\delta_n)^2/2}}{\sqrt{2\pi}} + \frac{\frac{1}{2}u_1^{-1/2}\mathrm{e}^{(-\sqrt{u_1}-\delta_n)^2/2}}{\sqrt{2\pi}} \\ &= \frac{u_1^{1/2-1}\mathrm{e}^{-u_1/2}}{\Gamma(1/2)2^{1/2}}\mathrm{e}^{-\delta_n^2/2}\left(\frac{\mathrm{e}^{\delta_n\sqrt{u_1}}+\mathrm{e}^{-\delta_n\sqrt{u_1}}}{2}\right) \\ &= \sum_{j=0}^{\infty} \mathrm{e}^{-\delta_n^2/2}\frac{(\delta_n^2/2)^j}{j!}\frac{u_1^{(2j+1)/2-1}\mathrm{e}^{-u_1/2}}{\Gamma((2j+1)/2)2^{(2j+1)/2}}. \end{aligned}$$

Forming the convolution $Z_1^2 + \sum_{k=2}^{n} Z_k^2$ term by term and using the convolution of $\chi^2_{(2j+1)}(0)$ and $\chi^2_{(n-1)}(0)$ is $\chi^2_{(n+2j)}(0)$ gives

$$f_{U_n}(u) = \sum_{j=0}^{\infty} \mathrm{e}^{-\frac{\delta_n^2}{2}} \frac{\left(\frac{\delta_n^2}{2}\right)^j}{j!} \frac{u^{(n+2j)/2-1}\mathrm{e}^{-u/2}}{\Gamma((n+2j)/2)2^{(n+2j)/2}}.$$

Thus, the noncentral $\chi_n^2(\delta_n^2)$ density is a Poisson$(\delta_n^2/2)$ probability mixture of central $\chi^2_{(n+2j)}(0)$ densities for $j = 0, 1, 2 \ldots, \infty$. If $\delta_n = 0$, then the distribution has the central chi-square $\chi_n^2(0)$ density.

For an alternate proof using moment generating functions, consider

$$\begin{aligned} m_{X_k^2}(t) = \mathrm{E}(\mathrm{e}^{tX_k^2}) &= \int_{-\infty}^{\infty} \mathrm{e}^{tx_k^2} \frac{\mathrm{e}^{-(x_k-\mu_k)^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x_k \\ &= \mathrm{e}^{-\mu_k^2/2}\frac{\mathrm{e}^{\mu_k^2/(2(1-2t))}}{\sqrt{1-2t}}. \end{aligned}$$

Then

$$\begin{aligned} m_{U_n}(t) &= \prod_{k=1}^{n} m_{X_k^2}(t) = \mathrm{e}^{-\delta_n^2/2}\frac{\mathrm{e}^{\delta_n^2/(2(1-2t))}}{(1-2t)^{n/2}} \\ &= \sum_{j=0}^{\infty} \frac{\mathrm{e}^{-\delta_n^2/2}(\delta_n^2/2)^j}{j!}(1-2t)^{-(n+2j)/2} \\ &= \sum_{j=0}^{\infty} \frac{\mathrm{e}^{-\delta_n^2/2}(\delta_n^2/2)^j}{j!} m_{\chi^2_{(n+2j)}(0)}(t). \end{aligned}$$

The parameter δ_n^2 is called the *noncentrality* parameter and n is the *degrees of freedom* parameter.

The characteristic function is

$$\phi_{U_n}(t) = \mathrm{e}^{-\delta_n^2/2}\frac{\mathrm{e}^{\delta_n^2/(2(1-2it))}}{(1-2it)^{n/2}},$$

where again $i = \sqrt{(-1)}$.

The cdf can be calculated from the incomplete gamma function as

$$\begin{aligned} F_{U_n}(x) = P(U_n \le x) &= \sum_{j=0}^{\infty} \frac{\mathrm{e}^{-\delta_n^2/2}(\delta_n^2/2)^j}{j!} \\ &\quad \times G\left(\frac{x}{2}, \frac{n+2j}{2}\right). \end{aligned}$$

We can calculate the moments of the noncentral chi-square random variable recursively. Let U_n be distributed as $\chi_n^2(\delta^2)$. Then for integer $r \ge 0$, we have

$$\begin{aligned} \mathrm{E}(U_n^r) &= \sum_{j=0}^{\infty} \frac{(\delta^2/2)^j\mathrm{e}^{-\delta^2/2}}{j!} \\ &\quad \times \int_0^{\infty} s^r \frac{s^{(n+2j)/2-1}\mathrm{e}^{-s/2}}{\Gamma((n+2j)/2)2^{(n+2j)/2}}\,\mathrm{d}s \\ &= \sum_{j=0}^{\infty} \frac{(\delta^2/2)^j\mathrm{e}^{-\delta^2/2}}{j!}\prod_{k=0}^{r-1}(n+2j+2k) \\ &= \sum_{j=0}^{\infty} \frac{(\delta^2/2)^j\mathrm{e}^{-\delta^2/2}}{j!}(n+2j)\prod_{k=0}^{r-2}(n+2+2j+2k) \\ &= n\mathrm{E}(U_{n+2}^{r-1}) + \delta^2\mathrm{E}(U_{n+4}^{r-1}), \end{aligned}$$

where U_m for $m = (n+2)$, $(n+4)$ have noncentral chi-square $\chi_m^2(\delta^2)$ distributions with the same parameter δ^2.

From this it follows by induction on integer $r = 1, 2, \ldots$ that

$$\mathrm{E}(U_n^r) = \sum_{k=0}^{r}\binom{r}{k}\delta^{2k}2^{r-k}\frac{\Gamma((n+2r)/2)}{\Gamma((n+2k)/2)}.$$

In particular, $\mathrm{E}(U_n) = n + \delta^2$, $\mathrm{Var}(U_n) = 2n + 4\delta^2$ and

$$\begin{aligned}
\mathrm{E}(U_n^2) &= n(n+2) + 2\delta^2(n+2) + \delta^4 \\
\mathrm{E}(U_n^3) &= n(n+2)(n+4) + 3\delta^2(n+2)(n+4) \\
&\quad + 3\delta^4(n+4) + \delta^6 \\
\mathrm{E}(U_n^4) &= n(n+2)(n+4)(n+6) + 4\delta^2(n+2) \\
&\quad \times (n+4)(n+6) + 6\delta^4(n+4)(n+6) \\
&\quad + 4\delta^6(n+6) + \delta^8.
\end{aligned}$$

For the central chi-square case with $\delta^2 = 0$, we have $\mathrm{E}(U_n) = n$, $\mathrm{Var}(U_n) = 2n$ and for integer $r \geq 0$,

$$\begin{aligned}
\mathrm{E}(U_n)^r &= 2^r \frac{\Gamma(r + (n/2))}{\Gamma(n/2)} \\
&= \begin{cases} 1 & \text{for } r = 0 \\ \prod_{k=0}^{r-1}(n+2k) & \text{for } r = 1, 2, \ldots \end{cases}.
\end{aligned}$$

If the $m \times 1$ random vector $\mathbf{X}$ has a multivariate normal distribution $\mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix},$$

$$\text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{pmatrix}$$

with $\boldsymbol{\Sigma}$ symmetric, positive semidefinite ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathrm{T}}$ and $\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v} \geq 0$ for all $\mathbf{v}$), then the quadratic form

$$\mathbf{X}^{\mathbf{T}}\boldsymbol{\Sigma}^{-}\mathbf{X} : \chi_r^2(\delta^2) \tag{1}$$

has the noncentral chi-square distribution where $\delta^2 = \boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-}\boldsymbol{\mu}$, the matrix $\boldsymbol{\Sigma}^{-}$ is a generalized inverse of $\boldsymbol{\Sigma}$ that satisfies $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$, and r is the rank of $\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}$ is positive definite ($\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v} > 0$ for all $\mathbf{v} \neq (0, 0, \ldots, 0)^{\mathrm{T}}$), then $\boldsymbol{\Sigma}^{-} = \boldsymbol{\Sigma}^{-1}$ is the inverse and we have full rank $r = m$.

The chi-square distribution lends its name to the chi-square statistic of Karl **Pearson** [4] (*see* **Chi-square Tests**). Let $X_1, X_2, \ldots, X_K$ have a multinomial joint distribution $\mathcal{M}(n, (p_1, p_2, \ldots, p_K))$ given by

$$\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K) \\
&\quad = \frac{n!}{x_1! x_2! \cdots x_K!} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K},
\end{aligned}$$

where $\sum_{k=1}^K p_k = 1$, $\sum_{k=1}^K x_k = n$, and $x_k \in \{0, 1, \ldots, n\}$. Then for testing the hypothesis

$$\begin{aligned}
H: \mathbf{p} &= (p_1, p_2, \ldots, p_K) = (p_{10}, p_{20}, \ldots, p_{K0}) \\
&= \mathbf{p_0},
\end{aligned}$$

where $\mathbf{p_0}$ is known, against the **alternative** $A: \mathbf{p} \neq \mathbf{p_0}$, we can reject for large values of the chi-square statistic

$$V_n = \sum_{k=1}^K \frac{(X_k - np_{k0})^2}{np_{k0}}.$$

(*see* **Hypothesis Testing**). For *near by* alternatives $p_k = p_{k0} + \eta_k/\sqrt{(n)}$, where $\sum_{k=1}^K \eta_k = 0$, it has a limiting noncentral chi-square distribution

$$V_n \xrightarrow{d} V : \chi_{K-1}^2(\delta^2)$$

as $n \to \infty$, where

$$\delta^2 = \lim_n \sum_{k=1}^K \frac{n(p_k - p_{k0})^2}{p_{k0}} = \sum_{k=1}^K \frac{\eta_k^2}{p_{k0}}.$$

This result uses the multivariate **central limit theorem** that the normalized random variables converge to a multivariate normal as $n \to \infty$:

$$\begin{aligned}
\mathbf{Z}_n^{\mathrm{T}} &= \left(\frac{X_1 - np_{10}}{\sqrt{n}}, \frac{X_2 - np_{20}}{\sqrt{n}}, \ldots, \frac{X_K - np_{K0}}{\sqrt{n}} \right) \\
&\xrightarrow{d} \mathbf{Z}^{\mathrm{T}} : \mathcal{N}_K((\eta_1, \eta_2, \ldots, \eta_K), \boldsymbol{\Sigma}_0),
\end{aligned}$$

where the variance–**covariance matrix** of rank $r = K - 1$ is

$$\begin{aligned}
\boldsymbol{\Sigma}_0 &= \begin{pmatrix} p_{10} & 0 & \ldots & 0 \\ 0 & p_{20} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_{K0} \end{pmatrix} - \begin{pmatrix} p_{10} \\ p_{20} \\ \vdots \\ p_{K0} \end{pmatrix} \\
&\quad \times (p_{10}, p_{20}, \ldots, p_{K0}).
\end{aligned}$$

Using the above quadratic form result (1) with $m = K$ and the generalized inverse

$$\boldsymbol{\Sigma}_0^- = \begin{pmatrix} 1/p_{10} & 0 & \dots & 0 \\ 0 & 1/p_{20} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/p_{K0} \end{pmatrix},$$

$$\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_0^- \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_0,$$

we have

$$V_n = \mathbf{Z}_n^{\mathrm{T}} \boldsymbol{\Sigma}_0^- \mathbf{Z}_n \xrightarrow{d} V = \mathbf{Z}^{\mathrm{T}} \boldsymbol{\Sigma}_0^- \mathbf{Z} : \chi^2_{K-1}(\delta^2),$$

where

$$\delta^2 = (\eta_1, \eta_2, \dots, \eta_K) \boldsymbol{\Sigma}_0^- \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_K \end{pmatrix} = \sum_{k=1}^{K} \frac{\eta_k^2}{p_{k0}}.$$

Setting $\eta_k = 0$ for the hypothesis, we have a limiting central $\chi^2_{K-1}(0)$ null distribution for V_n as $n \to \infty$.

Similar chi-square approximations hold for chi-square statistics in more complicated problems as well as for the generalized **likelihood ratio** statistic $-2\log_e(\lambda)$ of Wilks [6, 7].

More detailed references are [3, 5].

References

[1] Abramowitz, M. & Stegun, I. (1970). *Handbook of mathematical functions*, Dover Publications Inc., New York.

[2] Jones, W.B. & Thron, W.J. (1980). Continued fractions: analytic theory and applications, in *Encyclopedia of Mathematics and its Applications*, Vol. 11, Addison Wesley, Reading, MA, 350.

[3] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, 2nd Ed., Vol. 2, John Wiley, New York, 433–479, Chapter 29.

[4] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arizen from random sampling, *Philosophical Magazine, Series 5* **50**, 157–172.

[5] Tiku, M. (1985). Noncentral chi-square distribution, in *Encyclopedia of Statistical Sciences*, Vol. 6, N.L. Johnson, S. Kotz & C.B. Read eds. John Wiley, New York, pp. 276–280.

[6] Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses, *Annals of Mathematical Statistics* **9**, 60–62.

[7] Wilks, S.S. (1962). *Mathematical Statistics*. John Wiley, New York.

JEROME KLOTZ

Chi-square Distribution

If a random variable Z has a **standard normal** distribution [written $Z \sim \mathrm{N}(0, 1)$], then $W = Z^2$ is said to have a χ^2 distribution with one degree of freedom ($W \sim \chi^2(1)$). If $Z_1, Z_2, \ldots, Z_k$ are independent N (0, 1) random variables, then $W_k = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ is said to have a χ^2 distribution with k degrees of freedom ($W_k \sim \chi^2(k)$). This representation of a χ^2 random variable gives it an additive property important in applications: if U and V are independent random variables, and $U \sim \chi^2(r)$ and $V \sim \chi^2(s)$, then $U + V \sim \chi^2(r + s)$. This enables independent testing procedures of the same hypothesis to be combined.

The importance of the χ^2 distribution in statistical **hypothesis testing** initially derives from the following observation. If a large number n of binomial trials is performed with probability of success p in each trial, and X is the random variable recording the number of successes observed, then the normal approximation to the **binomial distribution** of the random variable X implies that if we write

$$W = \frac{(X - np)^2}{npq} \equiv \frac{(X - np)^2}{np} + \frac{(Y - nq)^2}{nq}, \quad (1)$$

then $W \sim \chi^2(1)$ approximately, where $q = 1 - p$ is the failure probability, and $Y = n - X$ is the number of failures. This generalizes to considering a situation in which on each trial k kinds of outcome can occur (rather than just success or failure), to read:

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

has approximately a $\chi^2(k - 1)$ distribution if n is large, with summation $\sum$ over the k classes. This generalization was introduced by Karl Pearson in 1900. It can be applied to test the hypothesis that the probabilities of outcomes corresponding to each of the k classes each have specified values. This is the celebrated **goodness-of-fit** test of Karl Pearson. It extends to testing hypotheses where the probabilities are assumed to have a general structure, as in testing for independence of attributes in a two-way **contingency table**.

Such tests are instances of a general theory of hypothesis tests based on the **likelihood ratio**, which uses chi-square as the large-sample distribution of a test statistic under a null hypothesis (*see* **Likelihood Ratio Tests**).

The probability distribution of a chi-square random variable is described by a **gamma** density, as was already shown in 1852 by the French statistician Irenée-Jules Bienaymé (1796–1878).

Historical detail and an extensive exposition of the theory and applications may be found in [1].

Reference

[1] Lancaster, H.O. (1969). *The Chi-Squared Distribution*. Wiley, New York.

(*See also* **Chi-square Distribution; Properties**; **Chi-square Tests**; **Multinomial Distribution**)

H.O. LANCASTER & E. SENETA

Chi-square Tests

Karl Pearson [34] originated the chi-square test as a goodness of fit test to determine if observed data are consistent with a proposed probability model. To use this test, we must partition the set of possible outcomes into a set of r mutually exclusive categories and count the number of observations falling into each category. The Pearson statistic

$$X^2 = \sum_{i=1}^{r} \frac{(N_i - m_i)^2}{m_i} \tag{1}$$

is an index of discrepancy between a set of observed counts $N_1, N_2, \ldots, N_r$ in the r categories and the corresponding set of expected counts $m_1, m_2, \ldots, m_r$ for the hypothesized probability model. Pearson showed that, as the expected counts increase, the distribution of X^2 approaches a central **chi-square distribution** with $r - 1$ **degrees of freedom** when the null hypothesis is true. Hence, the hypothesized model is rejected when the observed value of X^2 exceeds $\chi^2_{r-1,\alpha}$, the upper α percentile of the central chi-square distribution with $r - 1$ degrees of freedom, for some specified significance level α.

In many situations the expected counts $m_1, m_2, \ldots, m_r$ are functions of unknown parameters that must be estimated from the data. This would occur, for example, in testing the fit of the **normal distribution** with unspecified mean and variance, testing the fit of the **Poisson distribution** with an unspecified mean, or testing the independence hypothesis in a two-way **contingency table**. To achieve an asymptotic chi-square distribution for X^2, the unknown parameters must be estimated in an efficient manner from the observed category counts $N_1, N_2, \ldots, N_r$, and the degrees of freedom of the limiting chi-square distribution must be adjusted for the effective number of estimated parameters. This was pointed out by Fisher [10], who made adjustments to the degrees of freedom in applications to contingency tables. Later, Fisher [11] provided the first proof of the asymptotic chi-square distribution of X^2 in the general case where parameters are estimated from the data. We refer the reader to [6, 23], and [35] for more information on the early development of chi-square tests.

Tests with asymptotic chi-square distributions may also be derived from **likelihood ratio tests**. For testing a composite null hypothesis against a general alternative with data consisting of Poisson, binomial, or **multinomial** counts, the natural logarithm of the likelihood ratio multiplied by -2 has the form

$$G^2 = 2 \sum_{i=1}^{r} N_i \ln\left(\frac{N_i}{\hat{m}_i}\right), \tag{2}$$

where $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_r$ are **maximum likelihood** estimates of the expected counts for the model corresponding to the null hypothesis. This statistic is often referred to as the *deviance*. The Pearson statistic for a composite null hypothesis has the formula

$$X^2 = \sum_{i=1}^{r} \frac{(n_i - \hat{m}_i)^2}{\hat{m}_i}. \tag{3}$$

Both tests are members of the larger family of power divergence statistics identified by Cressie & Read [7]. Each member of this family of test statistics has the same asymptotic chi-square distribution when the null hypothesis is true, but different members have somewhat different **power** for detecting different types of departures from the null hypothesis in finite samples.

Applications to the analysis of contingency tables include chi-square tests of homogeneity for two or more **binomial distributions** (*see* **Two-by-Two Table**), or tests of homogeneity of two or more multinomial distributions, which are used, for example, to ascertain if different treatments are equally successful for treating a certain health problem. In higher-dimensional contingency tables chi-square tests are used to test hypotheses about various types of conditional independence and to assess the relative fit of members of nested sets of models.

Accurate use of the asymptotic chi-square distributions for these tests requires large values for the expected counts. This is an issue of practical concern as it limits the number of factor levels as well as the number of factors that can be used to cross-classify data in a contingency table. It also limits the number of categories that can be used to test the fit of a continuous distribution. Decisions must be made about both the number of intervals and either interval boundaries or the interval probabilities (*see* **Categorizing Continuous Variables**; **Grouped Data**). The power of the tests to detect deviations from the null hypothesis is also affected by these choices.

Standard formulas for chi-square tests such as (2) and (3) are derived from the assumption that the

observed counts are distributed as either independent Poisson counts, a single multinomial distribution, or a set of two or more independent multinomial or binomial distributions. The use of multinomial or binomial distributions for counts can often be justified through the use of **simple random sampling**. Large international, national, or multicenter health studies, however, often employ more complex sampling schemes which violate these standard assumptions and invalidate the use of some tests (*see* **Surveys, Health and Morbidity**). After reviewing standard applications of chi-square tests, we briefly consider some approaches to developing reliable chi-square tests for complex survey data.

Simple Goodness-of-Fit Tests

We first consider the use of a chi-square test as a goodness-of-fit test for a hypothesized distribution containing no unknown parameters. This is called a test of a simple null hypothesis. To test the simple null hypothesis that a set of observations $X_1, X_2, \ldots, X_n$ is a random sample from a population with continuous distribution function $F(x) = \Pr(X \le x)$, we partition the set of possible values that could be observed into r nonoverlapping intervals, say $(a_0, a_1], (a_1, a_2], \ldots, (a_{r-1}, a_r]$. When the null hypothesis is true, the probability that any single observation X_j falls into the ith interval is

$$\pi_i = \Pr(X_j \text{ falls in } (a_{i-1}, a_i]) = \int_{a_{i-1}}^{a_i} \mathrm{d}F(x), \quad (4)$$

and the expected count for the ith interval is $m_i = n\pi_i$. The null hypothesis is rejected if

$$X^2 = \sum_{i=1}^{r} \frac{(N_i - m_i)^2}{m_i} \ge \chi^2_{r-1,\alpha}.$$

The limiting chi-square distribution has $r - 1$ degrees of freedom because the r counts must satisfy a single constraint,

$$n = \sum_{i=1}^{r} N_i.$$

The interval probabilities and expected counts satisfy corresponding constraints,

$$1 = \sum_{i=1}^{r} \pi_i \quad \text{and} \quad n = \sum_{i=1}^{r} m_i.$$

We illustrate this application of the Pearson statistic with an analysis of the distribution of survival times for male residents of a total care facility.

Example 1

We want to test the null hypothesis that the number of months that male residents of a total care facility survive beyond age 65 has an **exponential** distribution with mean 180 months. Survival times beyond age 65 were recorded for a sample of 48 male residents, and the sorted times are shown in Table 1.

First, we partition the positive part of the real line into four adjacent intervals with probabilities $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$. The boundaries of the intervals are $a_0 = 0$, $a_1 = -180\ln(0.75) = 51.8$, $a_2 = -180\ln(0.5) = 125.8$, $a_3 = -180\ln(0.25) = 249.5$, and $a_4 = \infty$. The observed counts in these four intervals are (5, 8, 21, 14), the expected counts are $m_1 = m_2 = m_3 = m_4 = 12$, and the value of the Pearson statistic is

$$X^2 = \frac{(5-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(21-12)^2}{12} + \frac{(14-12)^2}{12} = 12.5.$$

The null hypothesis is rejected at the $\alpha = 0.01$ level of significance because $12.5 > 11.34 = \chi^2_{3,0.01}$. Consequently, the observed data are deemed to be inconsistent with a random sample of 48 survival times from an exponential distribution with mean 180 months.

Other Tests with Limiting Chi-Square Distributions

As previously noted, tests with limiting chi-square distributions are also obtained from likelihood ratio tests. The result that the natural logarithm of a

Table 1 Ordered survival times for male residents (months)

1	89	118	165	203	232	256	295
3	92	127	168	209	233	263	300
15	96	129	177	213	242	264	305
31	107	147	186	214	249	273	314
34	113	148	189	218	251	279	348
52	114	152	191	229	255	280	359

ratio of likelihoods multiplied by -2 has a limiting central chi-square distribution as the sample size becomes large, when the null hypothesis is true, was established under fairly general conditions by Wilks [54] and Wald [53]. For the simple goodness-of-fit test described in the previous section, the test statistic has the form

$$G^2 = 2\sum_{i=1}^{r} N_i \ln\left(\frac{N_i}{m_i}\right). \tag{5}$$

Two other statistics that have received some attention are the Freeman–Tukey [12] statistic

$$\mathrm{FT}^2 = 4\sum_{i=1}^{r}\left(\sqrt{N_i} - \sqrt{m_i}\right)^2 \tag{6}$$

and the Neyman [33] modified chi-square statistic

$$X_m^2 = \sum_{i=1}^{r}\frac{(N_i - m_i)^2}{N_i}. \tag{7}$$

These statistics are all members of the larger class of power divergence statistics identified by Cressie & Read [7]. Using

$$\mathbf{p} = (p_1, p_2, \ldots, p_r)' = \left(\frac{N_i}{n}, \frac{N_2}{n}, \ldots, \frac{N_r}{n}\right)'$$

to denote the vector of observed proportions, and

$$\pi = (\pi_1, \pi_2, \ldots, \pi_r)'$$

to denote the corresponding vector of true probabilities for the hypothesized distribution, the directed divergence of order λ of $\mathbf{p}$ from π is

$$I^{\lambda}(\mathbf{p}, \pi) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{r} p_i\left[\left(\frac{p_i}{\pi_i}\right)^{\lambda} - 1\right].$$

Although this quantity is a metric only for $\lambda = -1/2$, it provides a useful generalized **information** measure of the "distance" between $\mathbf{p}$ and π for any real number λ. For a test of a simple hypothesis, the divergence statistic

$$-2nI^{\lambda}(\mathbf{p}, \pi) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{r} N_i\left[\left(\frac{N_i}{m_i}\right)^{\lambda} - 1\right] \tag{8}$$

has an asymptotic chi-square distribution with $r-1$ degrees of freedom as $n \to \infty$. The Pearson statistic is obtained from (8) when $\lambda = 1$, the deviance is the limit as $\lambda \to 0$, the Freeman–Tukey statistic corresponds to $\lambda = -1/2$, and the Neyman statistic is the limit as $\lambda \to -1$.

For large samples, all members of the power divergence family have distributions that approach the same limiting chi-square distribution when the null hypothesis is correct, but different members may have different power for detecting different kinds of alternatives to the null hypothesis in small and moderate samples. No single member of this family dominates the other members with respect to power against all alternatives. Read & Cressie [43] suggest that values of λ between $\frac{1}{3}$ and $\frac{4}{5}$ provide a good compromise between relatively good power against most alternatives and the ability of the limiting chi-square distribution to approximate the distribution of the test statistic in small samples when the null hypothesis is true. This includes the Pearson statistic but excludes the deviance, Freeman–Tukey, and Neyman statistics. Read & Cressie recommend the power divergence test with $\lambda = \frac{2}{3}$, but we prefer the Pearson statistic because it exhibits very similar properties, it is well known and widely used, and it is commonly available in statistical software packages.

Wald Tests

To provide a broader view of the construction and uses of chi-square tests we consider a general approach for constructing chi-square tests from quadratic forms called Wald statistics (*see* **Likelihood**). The Pearson statistic can be derived as a Wald statistic, but this approach can be used to construct chi-square tests in many situations where the Pearson statistic does not have a limiting chi-square distribution.

To construct a Wald statistic, we must first select an r-dimensional random vector $\mathbf{Y}_n = (Y_{1n}, Y_{2n}, \ldots, Y_{rn})'$ that summarizes how results from a set of n observations deviate from what is expected to occur when the null hypothesis is true. A Wald statistic is a quadratic form,

$$\mathbf{Y}_n'\mathbf{C}\mathbf{Y}_n = \sum_{i=1}^{r}\sum_{j=1}^{r} c_{ij}Y_{in}Y_{jn}, \tag{9}$$

where $\mathbf{C}$ is an appropriate $r \times r$ symmetric matrix, and c_{ij} is the value in the ith row and jth column

of $\mathbf{C}$. To obtain a test statistic with a limiting central chi-square distribution, we must select $\mathbf{Y}_n$ so that it has a limiting multivariate normal distribution with mean vector $\mathbf{0} = (0, 0, \ldots, 0)'$ as $n \to \infty$ when the null hypothesis is true. Finally, the matrix $\mathbf{C}$ must satisfy the condition

$$\mathbf{\Sigma C \Sigma} = \mathbf{\Sigma}, \tag{10}$$

where $\mathbf{\Sigma}$ is the **covariance matrix** for the limiting distribution of $\mathbf{Y}_n$. The number of degrees of freedom for the limiting chi-square distribution of (9) is the trace of the matrix $\mathbf{C\Sigma}$.

If $\mathbf{\Sigma}^{-1}$ exists, then $\mathbf{C} = \mathbf{\Sigma}^{-1}$ is the unique solution to the condition of Eq. (10). It follows that the quadratic form

$$\mathbf{Y}_n' \mathbf{\Sigma}^{-1} \mathbf{Y}_n \tag{11}$$

provides a large sample chi-square test with r degrees of freedom as $n \to \infty$. On the other hand, if $\mathbf{\Sigma}$ does not have an inverse, then there will be an infinite number of choices for $\mathbf{C}$ that satisfy (10). Each of these choices will yield exactly the same value for (9), however, so it does not matter which one is used. In such cases, the degrees of freedom for the chi-square test, given by the trace of $\mathbf{C\Sigma}$, will be less than r. The inverse of $\mathbf{\Sigma}$ will not exist when the components of $\mathbf{Y}_n$ exactly satisfy one or more linear constraints. Then, individual components of $\mathbf{Y}_n$ have perfect **correlation** with linear combinations of other elements of $\mathbf{Y}_n$, so some components of $\mathbf{Y}_n$ are redundant. Another way to deal with this situation is to reduce the dimension of $\mathbf{Y}_n$ by deleting the minimum number of components required to break all of the linear constraints. There are many choices for the subset of components that can be deleted. Each choice corresponds to choosing a different $\mathbf{C}$ to satisfy (10) and results in the same value for the test statistic. Subtracting the number of deleted components from r gives the degrees of freedom for the chi-square test.

We will illustrate this recipe for constructing chi-square tests by using it to construct a simple goodness-of-fit test. In this application the set of possible outcomes is partitioned into a set of r nonoverlapping intervals. We use π_i to denote the probability that a random observation falls into the ith interval when the hypothesized distribution is correct. These probabilities are collected into the vector

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)'.$$

For a simple random sample of n observations, $X_1, X_2, \ldots, X_n$, the vector of observed counts for the r intervals, $\mathbf{N} = (N_1, N_2, \ldots, N_r)'$, has a multinomial distribution with sample size n and probability vector π when the hypothesized distribution is correct. Using

$$\mathbf{p} = (p_1, p_2, \ldots, p_r)' = \left(\frac{N_1}{n}, \frac{N_2}{n}, \ldots, \frac{N_r}{n}\right)'$$

to denote the vector of observed proportions, we consider the vector of scaled differences

$$\mathbf{Y}_n = n^{1/2}(\mathbf{p} - \boldsymbol{\pi}).$$

When the null hypothesis is true, $\mathbf{Y}_n$ has a limiting normal distribution where the means are all zero and the covariance matrix is $\mathbf{\Sigma} = (\mathbf{\Delta}_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}')$. Here $\mathbf{\Delta}_{\boldsymbol{\pi}}$ denotes a diagonal matrix with the elements of $\boldsymbol{\pi}$ on the main diagonal. It is easy to verify that $\mathbf{C} = \mathbf{\Delta}_{\boldsymbol{\pi}}^{-1}$ satisfies the condition of (10). Consequently, a large sample chi-square test is given by the quadratic form

$$\begin{aligned}\mathbf{Y}_n' \mathbf{\Delta}_{\boldsymbol{\pi}}^{-1} \mathbf{Y}_n &= n(\mathbf{p} - \boldsymbol{\pi})' \mathbf{\Delta}_{\boldsymbol{\pi}}^{-1} (\mathbf{p} - \boldsymbol{\pi}) \\ &= (\mathbf{N} - \mathbf{m})' \mathbf{\Delta}_{\mathbf{m}}^{-1} (\mathbf{N} - \mathbf{m}),\end{aligned}$$

which is a matrix expression for the Pearson statistic in (1). Here, $\mathbf{m} = n\boldsymbol{\pi}$.

In this case, the Wald statistic is equivalent to the Pearson statistic, but this is not always true. For a simple goodness-of-fit test, appropriate use of the Pearson statistic as a chi-square test depends on $\mathbf{\Sigma}$, the covariance matrix for the limiting normal distribution of $\mathbf{Y}_n = n^{1/2}(\mathbf{p} - \boldsymbol{\pi})$. If $\mathbf{\Delta}_{\boldsymbol{\pi}}$ does not satisfy the condition of (10) for the given $\mathbf{\Sigma}$, then the Pearson statistic is not equivalent to a Wald test and neither the Pearson statistic nor any other member of the family of power divergence statistics has a limiting chi-square distribution.

There are many other uses for Wald tests. For example, $\mathbf{Y}_n$ could be the difference between a vector of parameter estimates and a vector of hypothesized values for those parameters. Then the Wald test is used to simultaneously test that all the hypothesized parameter values are correct. Wald tests also provide a convenient way to obtain large sample chi-square tests for complex survey data.

Theoretical results for asymptotic chi-square distributions of quadratic forms are presented by Serfling [48, pp. 128–130] and Moore & Spruill [31]. Moore [29] provides straight forward extensions

to situations where $\boldsymbol{\Sigma}$ depends on the values of parameters that must be estimated from the data.

Goodness-of-Fit Tests for Composite Hypotheses

Suppose we want to test the composite null hypothesis that observations $X_1, X_2, \ldots, X_n$ were randomly sampled from a population distribution function in some family of distribution functions denoted by $F(x;\boldsymbol{\theta})$, where members of the family are distinguished by a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)'$. For example, we may wish to test the hypothesis that the observed data were randomly sampled from a normal distribution with unknown mean and variance. The construction of a chi-square test for a composite null hypothesis requires estimation of $\boldsymbol{\theta}$ as well as the selection of a partition of the real line into nonoverlapping intervals.

The formula for the test statistic and the related degrees of freedom will depend on the manner in which parameters are estimated and boundaries of the intervals are determined. There are two basic approaches. In the first approach the interval boundaries $-\infty = a_0 < a_1 < \cdots < a_r = \infty$ are selected before the data are observed. Then, the expected counts are functions of unknown parameters that must be estimated from the observed data. The form of the chi-square test and the resulting degrees of freedom will vary depending on whether the original observations, $X_1, X_2, \ldots, X_n$, or the interval counts, $N_1, N_2, \ldots, N_r$, are used to estimate $\boldsymbol{\theta}$. In the second approach, interval probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)'$ are specified. This fixes the values of the expected counts, but the interval boundaries depend on $\boldsymbol{\theta}$ and must be estimated from the observed data before the interval counts can be computed. The first approach is said to use fixed intervals and the second approach is said to use random intervals.

Chi-Square Tests with Fixed Intervals

Without considering the observed data, specify boundaries $-\infty = a_0 < a_2 < \cdots < a_r = \infty$ for r nonoverlapping intervals. Let $\mathbf{N} = (N_1, N_2, \ldots, N_r)'$ denote the corresponding vector of observed counts obtained by classifying the observations, $X_1, X_2, \ldots, X_n$, into the r intervals. The expected count for the jth interval is $m_j(\boldsymbol{\theta}) = n\pi_j(\boldsymbol{\theta})$, where

$$\pi_j(\boldsymbol{\theta}) = \Pr(a_{j-1} < X_i \le a_j) = \int_{a_{j-1}}^{a_j} \mathrm{d}F(x;\boldsymbol{\theta}). \tag{12}$$

The null hypothesis is tested by comparing the observed counts to the estimates of the expected counts obtained by substituting an appropriate estimate of $\boldsymbol{\theta}$ into (12).

The Pearson statistic, or any other member of the power divergence family, provides a test with an asymptotic central chi-square distribution if $\boldsymbol{\theta}$ is estimated from the observed counts $(N_1, N_2, \ldots, N_r)$ instead of the original observations $X_1, X_2, \ldots, X_n$. Typically, the maximum likelihood estimator (mle) of $\boldsymbol{\theta}$ for the multinomial distribution of $(N_1, N_2, \ldots, N_r)$ is used, but any other asymptotically equivalent estimator could also be used.

Example 2

As an illustration, we use the Pearson statistic to test the composite null hypothesis that the data in Table 1 were randomly sampled from an exponential distribution with unknown mean θ. The distribution function is $F(x;\theta) = 1 - \exp(-x/\theta)$ and the density function is $f(x;\theta) = \theta^{-1}\exp(-x/\theta)$ for $x > 0$ and $\theta > 0$. Using the $r = 4$ intervals, (0, 51.8], (51.8, 125.8], (125.8, 249.5], (249.5, ∞], considered in Example 1, the observed counts are $(N_1, N_2, N_3, N_4) = (5, 8, 21, 14)$.

From (12) the probability that a randomly selected observation from the hypothesized distribution falls into the jth interval is

$$\begin{aligned}\pi_j(\theta) &= \int_{a_{j-1}}^{a_j} \theta^{-1}\exp\left(\frac{-x}{\theta}\right)\mathrm{d}x \\ &= \exp\left(\frac{-a_{j-1}}{\theta}\right) - \exp\left(\frac{-a_j}{\theta}\right)\end{aligned} \tag{13}$$

and the expected counts are $m_j(\theta) = n\pi_j(\theta)$, for $j = 1, 2, 3, 4$. An estimate of θ is obtained by maximizing the log **likelihood** function

$$\begin{aligned}l(\theta; N_1, N_2, \ldots, N_r) &= \ln(n!) - \sum_{j=1}^{r}\ln(N_j!) \\ &+ \sum_{j=1}^{r} N_j \ln\left[\exp\left(\frac{-a_{j-1}}{\theta}\right) - \exp\left(\frac{-a_j}{\theta}\right)\right]\end{aligned}$$

for the multinomial distribution of the observed counts. The first derivative of the log likelihood function with respect to θ is

$$S_{\mathbf{N}}(\theta) = \sum_{j=1}^{r} \frac{N_j \partial \pi_j(\theta)}{\pi_j(\theta) \partial \theta} \tag{14}$$

$$= \frac{1}{\theta^2} \sum_{j=1}^{r} \frac{N_j}{\pi_j(\theta)} \left[a_{j-1} \exp\left(\frac{-a_{j-1}}{\theta}\right) - a_j \times \exp\left(\frac{-a_j}{\theta}\right) \right]. \tag{15}$$

This is called the score function. The mle for θ is obtained by setting the score function equal to zero and solving the resulting equation for θ. In this case the solution cannot be expressed as a simple function of the observed counts, but a numerical solution can be obtained. The resulting mle is $\hat{\theta}_{\mathbf{N}} = 239.48$. This estimator is given the subscript **N** to indicate that it is the mle for θ computed from the interval counts.

Substituting $\hat{\theta}_{\mathbf{N}}$ for θ in (13), we obtain estimates of expected counts

$$m_1(\hat{\theta}_{\mathbf{N}}) = 9.34, \qquad m_2(\hat{\theta}_{\mathbf{N}}) = 10.28,$$
$$m_3(\hat{\theta}_{\mathbf{N}}) = 11.45, \qquad n_4(\hat{\theta}_{\mathbf{N}}) = 16.93,$$

and the value of the Pearson statistic is

$$X^2 = \sum_{j=1}^{r} \frac{[N_j - m_j(\hat{\theta}_{\mathbf{N}})]^2}{m_j(\hat{\theta}_{\mathbf{N}})} = 10.99. \tag{16}$$

This test has $r - k - 1 = 4 - 1 - 1 = 2$ degrees of freedom. Since $X^2 = 10.99$ exceeds $\chi^2_{2,0.01} = 9.21$, the hypothesis that the data were sampled from an exponential distribution is rejected at the 0.01 level.

Alternatively, the original observations, $X_1, X_2, \ldots, X_n$, could be used to estimate unknown parameters in the hypothesized family of distributions. This avoids loss of information encountered in using interval counts. When an mle $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$ is computed from the original observations; however, neither the Pearson statistic, nor any other member of the power divergence family of tests, has a limiting chi-square distribution when the null hypothesis is correct. Chernoff & Lehmann [5] showed that the Pearson statistic has a limiting distribution, corresponding to a linear combination of independent chi-square random variables in this case.

Greenwood & Nikulin [14] show that a Wald statistic with a limiting chi-square distribution can be written as the sum of a Pearson statistic and a correction term. In this version of the Pearson statistic, estimates of the expected counts are evaluated using $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$ instead of $\hat{\boldsymbol{\theta}}_{\mathbf{N}}$. The formula for the test statistic is

$$X^2_{\mathrm{NR}} = \sum_{j=1}^{r} \frac{[N_j - m_j(\hat{\boldsymbol{\theta}}_{\mathbf{X}})]^2}{m_j(\hat{\boldsymbol{\theta}}_{\mathbf{X}})} + \mathbf{S}_{\mathbf{N}}(\hat{\boldsymbol{\theta}}_{\mathbf{X}})' [\mathbf{J}_{\mathbf{X}}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) - \mathbf{J}_{\mathbf{N}}(\hat{\boldsymbol{\theta}}_{\mathbf{X}})]^{-1} \mathbf{S}_{\mathbf{N}}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}), \tag{17}$$

where the $k \times 1$ vector $\mathbf{S}_{\mathbf{N}}(\hat{\boldsymbol{\theta}}_{\mathbf{X}})$ is the score function of the multinomial likelihood for the interval counts $(N_1, N_2, \ldots, N_r)$ evaluated at $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$, $\mathbf{J}_{\mathbf{N}}(\hat{\boldsymbol{\theta}}_{\mathbf{X}})$ is the Fisher **information matrix** for the multinomial likelihood function of the interval counts evaluated at $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$, and $\mathbf{J}_{\mathbf{X}}(\hat{\boldsymbol{\theta}}_X)$ is the Fisher information matrix for the likelihood function of the original observations evaluated at $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$. The large sample chi-square distribution for X^2_{NR} has $r - 1$ degrees of freedom when r intervals are used.

Example 3

As in Example 2, we test the composite null hypothesis that the data in Example 1 were randomly sampled from an exponential distribution with unknown mean θ. We use the same class intervals, but the mle for the hypothesized exponential distribution of $X_1, X_2, \ldots, X_n$ will be used to estimate θ in the evaluation of the expected counts.

The log likelihood function for the hypothesized exponential distribution is

$$l(\theta; \mathbf{X}) = -n \ln(\theta) - \theta^{-1} \sum_{i=1}^{n} X_i.$$

Setting $\partial l(\theta; \mathbf{X})/\partial\theta$ equal to zero and solving the resulting equation yields the sample mean,

$$\hat{\theta}_{\mathbf{X}} = n^{-1} \sum_{i=1}^{n} X_i = 186,$$

as the mle for θ. Substituting $\hat{\theta}_{\mathbf{X}}$ for θ in (13) yields estimates of expected counts

$$m_1(\hat{\theta}_{\mathbf{X}}) = 11.67, \qquad m_2(\hat{\theta}_{\mathbf{X}}) = 12.93,$$
$$m_3(\hat{\theta}_{\mathbf{X}}) = 11.86, \qquad m_4(\hat{\theta}_{\mathbf{X}}) = 12.55,$$

and the resulting value of the Pearson statistic is

$$X^2 = \sum_{j=1}^{r} \frac{[N_j - m_j(\hat{\theta}_{\mathbf{X}})]^2}{m_j(\hat{\theta}_{\mathbf{X}})} = 12.32.$$

To obtain the correction term, we evaluate the multinomial score function given by (15) at $\hat{\theta}_{\mathbf{X}} = 186$ to obtain $S_{\mathbf{N}}(\hat{\theta}_{\mathbf{X}}) = 0.0514814$. Since there is only one parameter, the Fisher information matrix for the multinomial distribution of the interval counts contains a single element

$$J_{\mathbf{N}}(\theta) = \mathrm{E}\left(-\frac{\delta^2 l(\theta;\mathbf{N})}{\delta\theta^2}\right) = \frac{n}{\theta^4}\sum_{j=1}^{r}\frac{1}{\pi_j(\theta)} \times\left[a_{j-1}\exp\left(\frac{-a_{j-1}}{\theta}\right) - a_j\exp\left(\frac{-a_j}{\theta}\right)\right]^2. \tag{18}$$

The Fisher information matrix for the hypothesized exponential distribution of the original observations is

$$J_{\mathbf{X}}(\theta) = \frac{n}{\theta^2}. \tag{19}$$

Evaluating (18) and (19) at $\hat{\theta}_{\mathbf{X}} = 186$ yields $J_{\mathbf{N}}(\hat{\theta}_{\mathbf{X}}) = 0.0010056$, $J_{\mathbf{X}}(\hat{\theta}_{\mathbf{X}}) = 0.0013874$ and a correction of 6.94 to the Pearson statistic. Then, the value of the test statistic is

$$X^2_{\mathrm{NR}} = 12.32 + 6.94 = 19.26.$$

This exceeds $\chi^2_{3,0.001} = 11.34$, and the null hypothesis is rejected at the 0.001 level.

Chi-Square Tests with Random Intervals

In this approach the interval probabilities $\pi_1, \pi_2, \ldots, \pi_k$ are specified and the interval boundaries must be estimated before the counts can be evaluated. For a hypothesized distribution function $F(x;\boldsymbol{\theta})$, the interval boundaries are defined as

$$a_j(\boldsymbol{\theta}) = F^{-1}(\pi_1 + \pi_2 + \cdots + \pi_j;\boldsymbol{\theta}), \tag{20}$$

for $j = 1, 2, \ldots, k$, where $F^{-1}(c;\boldsymbol{\theta}) \equiv \inf[x\colon F(x;\boldsymbol{\theta}) \geq c]$. The boundaries are estimated by replacing $\boldsymbol{\theta}$ in (20) with $\hat{\boldsymbol{\theta}}_X$, the maximum likelihood estimator (mle) computed from the original observations $X_1, X_2, \ldots, X_n$. The count for the jth interval, denoted by N_j^*, is the number of observations falling in the random interval $(a_{j-1}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}), a_j(\hat{\boldsymbol{\theta}}_{\mathbf{X}})]$.

A Wald test with a limiting chi-square distribution when $F(x;\boldsymbol{\theta})$ is the correct distribution function was developed by Rao & Robson [37]. This test statistic can also be written as a Pearson statistic plus a correction term,

$$X^2_{\mathrm{RR}} = \sum_{j=1}^{r}\frac{(N_j^* - n\pi_j)^2}{n\pi_j} + \mathbf{D}(\hat{\boldsymbol{\theta}}_{\mathbf{x}})'[\mathbf{J}_{\mathbf{x}}(\hat{\boldsymbol{\theta}}_{\mathbf{x}}) - n\mathbf{W}(\hat{\boldsymbol{\theta}}_{\mathbf{x}})\boldsymbol{\Delta}_{\boldsymbol{\pi}}^{-1}\mathbf{W}(\hat{\boldsymbol{\theta}}_{\mathbf{x}})]^{-1}\mathbf{D}(\hat{\boldsymbol{\theta}}_{\mathbf{x}}), \tag{21}$$

where $\boldsymbol{\Delta}_{\boldsymbol{\pi}}$ is a diagonal matrix with the elements of $(\pi_1, \pi_2, \ldots, \pi_k)$ on the main diagonal, $\mathbf{W}(\hat{\boldsymbol{\theta}}_{\mathbf{X}})$ is a $k \times r$ matrix with (i, j) element

$$w_{ij} = f[a_{j-1}(\theta);\theta]\frac{\partial a_{j-1}(\theta)}{\partial\theta_i} - f[a_j(\theta);\theta]\frac{\partial a_j(\theta)}{\partial\theta_i},$$

and $\mathbf{D}(\hat{\boldsymbol{\theta}}_{\mathbf{X}})$ is a $k \times 1$ vector where the ith element is

$$d_i = \sum_{j=1}^{r} w_{ij}\frac{N_j^*}{\pi_j}.$$

Here $f(x;\boldsymbol{\theta})$ is the density function for the hypothesized distribution.

Example 4

We test the composite null hypothesis that the data in Table 1 are a random sample from an exponential distribution with unknown mean θ. From Example 3, $\hat{\theta}_{\mathbf{X}} = 186$, and taking $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$, the interval boundaries are estimated as

$$a_j(\hat{\theta}_{\mathbf{X}}) = -\hat{\theta}_{\mathbf{X}}\ln\left(\frac{1-j}{r}\right).$$

Consequently, the intervals are (0, 53.51], (53.51, 128.93], (128.93, 257.85], (257.85, ∞], the observed counts are $(N_1^*, N_2^*, N_3^*, N_4^*)$ = (6,8,23,11), and

$$X^2 = \sum_{j=1}^{r}\frac{(N_j^* - n/r)^2}{n/r} = 14.5.$$

The correction term is evaluated with

$$\frac{\partial a_j(\theta)}{\partial\theta_i} = -\ln\left(\frac{1-j}{r}\right)$$

and

$$w_{ij} = \hat{\theta}_{\mathbf{X}}^{-1}\left\{\ln\left(\frac{1-j}{r}\right)\exp\left[-a_j\frac{\hat{\theta}_{\mathbf{X}}}{\hat{\theta}_{\mathbf{X}}}\right] - \ln\left[1 - \frac{j-1}{r}\right]\exp\left[-a_{j-1}\frac{\hat{\theta}_{\mathbf{X}}}{\hat{\theta}_{\mathbf{X}}}\right]\right\}.$$

Then,

$$\mathbf{W}(\hat{\theta}_{\mathbf{X}}) = [-0.001160, -0.000703, 0, 0.001863],$$

$\mathbf{D}(\hat{\theta}_{\mathbf{X}}) = 0.031639$, and

$$X^2_{\mathrm{RR}} = 14.5 + 2.72 = 17.22,$$

with $r - 1 = 3$ degrees of freedom. This test also rejects the null hypothesis that the observations were sampled from an exponential distribution at the 0.001 level.

Greenwood & Nikulin [14] provide a thorough review of chi-square tests with fixed and random intervals that includes proofs of the asymptotic chi-square distributions. Special formulas are presented for testing the fit of the normal distribution, **location-scale families** of continuous distributions, and both discrete and continuous members of the **exponential family** of distributions. They also review the literature on selection of intervals and the small sample behavior of these tests.

Selection of the number of intervals and either the interval boundaries or the interval probabilities affects both the power of a test to detect deviations from the hypothesized distribution and the accuracy of the chi-square distribution as an approximation of its finite sample distribution when the hypothesized distribution is correct. Generally, r equiprobable intervals are recommended when nothing is assumed about likely alternatives to the hypothesized distribution. Since accurate use of the asymptotic chi-square approximation requires large expected counts, r should not be taken to be too large. Choosing a value for r that is either too large or too small may reduce the power of the test against many alternatives. For a sequence of tests with $r = 2, 3, 4, \ldots$ equiprobable intervals, power against most alternatives initially increases, reaches a maximum at some value of r, and then declines as more intervals are used. Based on a large-sample approximation derived by Mann & Wald [26] and later refined by Schorr [47], Moore [30] recommends using roughly $r = 2n^{2/5}$ equiprobable intervals for a sample of n observations. Simulation results reported by Koehler & Gan [21] suggest that using fewer intervals, say $r = n^{2/5}$, provides better power against many alternatives. Greenwood & Nikulin [14] recommend $r \leq \min[\alpha^{-1}, \ln(n)]$, where α is the significance **level of the test**.

Simulation result reported by Kallenberg [18] indicate that appreciable gains in power can sometimes be achieved by using nonequiprobable intervals. Power for detecting heavy tailed alternatives, for example, may be improved by using more intervals with smaller probabilities in the tails of the distribution. A greater gain in power against a specific class of alternatives can often be achieved, however, by using a test that more directly focuses on the specific class of alternatives than the omnibus chi-square tests considered in (16), (17), and (21).

In testing the fit of a continuous distribution, the conversion of the original observations $X_1, X_2, \ldots, X_n$ into interval counts generally results in some loss of information. When the values of $X_1, X_2, \ldots, X_n$ are available, more powerful goodness-of-fit tests can be obtained from methods based on empirical distribution functions or probability plots. A variety of such tests are reviewed by Stephens [49, 50]. Unlike the chi-square tests considered in (16), (17), and (21), these tests typically do not have convenient asymptotic distributions. Limiting distributions generally depend on the hypothesized distribution and they are often intractable. Tables of finite sample percentiles (*see* **Quantiles**) exist for testing the fit of a few distributions, but use of these tests may require **Monte Carlo simulation** of finite sample critical values.

Finally, tests with asymptotic chi-square distributions that use the original observations and do not require the classification of observations into intervals can be derived from the asymptotic normality of the score function for the hypothesized distribution. Such tests are called score tests. Further power can be gained by restricting attention to classes of alternative distributions with density functions that deviate in a smooth manner from the density for the hypothesized distribution. These are called smooth tests of goodness of fit. Derivations and applications of score tests and smooth tests are reviewed by Rayner & Best [41], who recommend the use of smooth tests for testing the fit of continuous distributions and the use of the Pearson statistic for discrete distributions.

Applications to Categorical Data

Chi-square tests provide the primary method for making inferences from categorical response data. This

Table 2 A two-way contingency table

Row factor	Column factor $j=1$	$j=2$	...	$j=J$	Row totals
$i=1$	N_{11}	N_{12}	...	N_{15}	N_{1+}
$i=2$	N_{21}	N_{22}	...	N_{25}	N_{2+}
$i=I$	N_{I1}	N_{I2}	...	N_{IJ}	N_{I+}
Column totals	N_{+1}	N_{+2}	...	N_{+J}	N_{++}

includes tests of independence and homogeneity in contingency tables and tests of significance about parameters in **logistic regression** models and other **generalized linear models**. We restrict out attention to tests of independence or homogeneity in two-way contingency tables. Additional applications can be found in any book on logistic regression analysis, generalized linear models, or **categorical data analysis**. These include the pioneering work on **loglinear models** by Bishop et al. [3] and the more recent overview of categorical data analysis by Agresti [2]. Fienberg [9] provides a good introduction to both the asymptotic theory for chi-square tests and their use in categorical data analysis.

In the construction of a two-way contingency table, each observed response is classified with respect to a finite set of exhaustive and nonoverlapping categories for each of two factors. The form of the table is shown in Table 2, where N_{ij} is the number of responses achieving both the ith level of the row factor and the jth level of the column factor. The Pearson statistic for testing the null hypothesis of independence (or no association) between the row and column factors is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}, \tag{22}$$

where $\hat{m}_{ij}$ is the maximum likelihood estimate of the expected count given by the formula

$$\hat{m}_{ij} = \frac{(\text{row total}) \times (\text{column total})}{(\text{total for entire table})} = \frac{N_{i+}N_{+j}}{N_{++}}. \tag{23}$$

The deviance, or likelihood ratio test statistic, is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij} \ln\left(\frac{N_{ij}}{\hat{m}_{ij}}\right). \tag{24}$$

When the independence hypothesis is true, both X^2 and G^2, as well as any other member of the power divergence family, have asymptotic chi-square distributions with $(I-1)(J-1)$ degrees of freedom if the counts are independent Poisson counts, or the entire table of counts has a multinomial distribution, or the rows of the table correspond to independent multinomial or binomial distributions, or the columns of the table correspond to independent multinomial or binomial distributions.

Example 5

As part of a survey of attitudes toward primary health care education and practice, independent random samples of first-year medical students, fourth-year medical students and postgraduate residents were taken from national databases of the American Medical Association and the Association of American Medical Colleges [4]. Table 3 summarizes the specialty orientation of the respondents. Each column in this table corresponds to an independent multinomial distribution with four response categories.

In this case, the test of the hypothesis of independence of intended specialty and academic status is a test of homogeneity of the distributions across intended specialties for the three groups. Table 4 shows the estimates of the expected counts obtained from (23). The values of the Pearson statistic and the deviance are $X^2 = 14.50$ and $G^2 = 14.64$, respectively, both with $(4-1)(3-1) = 6$ degrees of freedom. The independence hypothesis is rejected at the 0.025 level of significance. Further inspection of data reveals a slightly lower percentage of first-year students orientated toward specialist careers, a lower percentage of fourth-year students interested in internal medicine, and a slightly lower percentage of residents orientated toward mixed and primary care practice.

Table 3 Specialty orientation of respondents

Specialty	Academic status Year 1	Year 4	Residents	Totals
Specialists	127	174	366	667
Internal medicine	29	19	72	120
Mixed	24	19	32	75
Primary care	71	95	175	341
Totals	251	307	645	1203

Table 4 Expected counts for Table 3

Specialty	Academic status			Totals
	Year 1	Year 4	Residents	
Specialists	139.17	170.21	357.62	667
Internal medicine	25.04	30.62	64.34	120
Mixed	15.65	19.14	40.21	75
Primary care	71.15	87.02	182.83	341
Totals	251	307	645	1203

This example illustrates the most basic application of a chi-square test to the analysis of a contingency table. For details on applications to more complex hypotheses, see **Categorical Data Analysis**, **Contingency Table**, and **Loglinear Model**.

It is important to realize that the X^2 and G^2 tests used in the previous example are not appropriate for all tables of counts. When entries in a contingency table do not correspond to either independent Poisson counts, or a single multinomial distribution, or several independent multinomial (binomial) distributions, then neither the Pearson statistic, nor the deviance, nor other members of the power divergence family will have limiting chi-square distributions. Such tables of counts arise from the use of complex sampling schemes involving various levels of **stratification** and **cluster sampling**. They can also arise in the analysis of **longitudinal** studies and repeated measures studies, where several responses are obtained from each respondent. Some care must be exercised in constructing the table of counts and selecting the test statistic.

As a simple illustration of a repeated measures study, we consider the much analyzed vision data for 7477 women, aged 30–39, employed in Royal Ordnance factories in Britain. These data were analyzed by Stuart [51], Grizzle et al. [15], and Bishop et al. [3], among others. Each woman received two classifications: one for quality of vision in the left eye and the other for quality of vision in the right eye. To test if the distributions across the four vision categories are the same for right and left eyes, one might consider computing the Pearson statistic for the counts in Table 5. The large sample chi-square approximation to the distribution of the Pearson statistic is inappropriate in this situation because it incorrectly treats the rows of Table 5 as two independent multinomial distributions. The rows of Table 5 are not independent multinomial distributions because vision quality in the right eye has a positive correlation with vision quality in the left eye of individual women.

Example 6

To obtain a Wald statistic with a limiting chi-square distribution, we must create a table of counts showing both the left and right eye results for each woman in the study. This is done in Table 6, where each woman appears in exactly one cell of the table. A multinomial distribution with 16 categories is appropriate for the counts in this table. We want to test the hypothesis that the distribution of the row totals is the same as the distribution of the column totals. This is called a test of marginal homogeneity.

To derive a Wald statistic, we use $\mathbf{N}_L = (1907, 2222, 2507)'$ to denote the column vector of counts in the three highest categories for left eye quality. Similarly, we use $\mathbf{N}_R = (1976, 2256, 2456)'$ to denote a column vector of counts in the three highest categories for right eye quality. Corresponding vectors of proportions are $\mathbf{P}_L = \mathbf{N}_L/n$ and $\mathbf{P}_R = \mathbf{N}_R/n$, and we can express the marginal homogeneity hypothesis as $E(\mathbf{P}_L - \mathbf{P}_R) = \mathbf{0}$. The proportions for the fourth vision category are not needed because proportions are constrained to sum to one across the four categories for both the right and left eyes.

Table 5 Vision quality for women employed in Royal Ordnance factories

Eye	Vision category				Totals
	Highest	Second	Third	Lowest	
Left	1907	2222	2507	841	7477
Right	1976	2256	2456	789	7477
Totals	3883	4478	4963	1630	14 954

Table 6 Vision quality for women employed in Royal Ordnance factories

Right eye category	Left eye vision category				Totals
	Highest	Second	Third	Lowest	
Highest	1520	266	124	66	1976
Second	234	1512	432	78	2256
Third	117	362	1772	205	2456
Lowest	36	82	179	492	789
Totals	1907	2222	2507	841	7477

Then, an appropriate Wald statistic is

$$X^2_{\mathrm{MH}} = n(\mathbf{P}_{\mathrm{L}} - \mathbf{P}_{\mathrm{R}})'\mathbf{V}^{-1}(\mathbf{P}_{\mathrm{L}} - \mathbf{P}_{\mathrm{R}}), \qquad (25)$$

where $\boldsymbol{V}$ is a consistent estimate of the covariance matrix for $\sqrt{n}(\mathbf{P}_{\mathrm{L}} - \mathbf{P}_{\mathrm{R}})$. A formula for $\boldsymbol{V}$ is derived from the multinomial distribution for the counts in Table 6 by noting that $\mathbf{P}_{\mathrm{L}} - \mathbf{P}_{\mathrm{R}} = \mathbf{AP}$, where $\mathbf{P} = n^{-1}(N_{11}N_{12}N_{13}N_{14}N_{21}\ldots N_{44})'$ denotes the sample proportions from Table 6 arranged as a 16×1 column vector and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix}$$

Since $\mathbf{P}$ is a consistent estimate of the probability vector for 16 joint vision categories, it follows that

$$\mathbf{V} = \mathbf{A}(\boldsymbol{\Delta}_{\mathbf{P}} - \mathbf{PP}')\mathbf{A}$$

is a consistent estimator for the covariance matrix of $\sqrt{n}(\mathbf{P}_{\mathrm{L}} - \mathbf{P}_{\mathrm{R}})$.

Evaluating $\boldsymbol{V}$, $\mathbf{P}_{\mathrm{L}}$, and $\mathbf{P}_{\mathrm{R}}$ from the data in Table 6, $X^2_{\mathrm{MH}} = 11.96$ with three degrees of freedom and the P value = 0.0075. The quality of vision is not the same for both eyes, and further inspection of the data reveals that quality of vision tends to be higher in the right eye for this population of women.

For additional details and illustrations, see **Square Contingency Table** and **Matched Pairs With Categorical Data**.

Randomization and Conditional Tests

The large-sample chi-square distribution for the Pearson statistic is also justified through randomly assigning subjects to treatment groups. When the treatments are equally effective, the randomization distribution of the possible values of X^2 corresponding to the possible random allocations of subjects to treatment groups will approximately have a chi-square distribution (*see* **Randomization Tests**). The chi-square approximation becomes more accurate as the number of subjects increases.

We could compute an exact P value for a randomization test searching through all possible tables of counts with the same row and column totals as the observed table of counts to determine the number of tables with larger X^2 values than the X^2 value for the observed table (*see* **Fisher's Exact Test**). Development of high-speed computers and efficient algorithms, like the Mehta & Patel [28] algorithm, which is able to avoid searching large blocks of tables with small X^2 values, have made it possible to apply this approach to moderately large tables. When the row and column totals are sufficiently large or the number of cells in the table is sufficiently large, even the most efficient search can overwhelm any computer. One alternative to searching through all the possible random allocations is to approximate the exact P value from the results of a sample from the possible random allocations. Agresti [1] provides an extensive review of randomization tests for contingency tables that includes a discussion of the controversy over the use of tests that condition on observed marginal counts.

Small Samples and Sparse Tables

The central chi-square distribution is a limiting distribution for X^2, G^2, and other members of the power divergence family as expected counts become infinitely large. It may not provide an accurate approximation to the distribution of these statistics if too many of the expected counts are too small. The point at which the chi-square approximation begins seriously to deteriorate depends on many factors, including the number of cells in the table, the total sample size, the relative sizes of the expected counts, and the formula for the test statistic.

A commonly used rule of thumb is that no expected count should be smaller than 1 and at least 80% of the expected counts should be no less than 5. This was part of the recommendations made in Cochran's extremely influential paper [6] on chi-square tests. While this rule of thumb guarantees accurate use of the chi-square distribution in computing P values for the Pearson statistic, simulation studies by Roscoe & Byars [45], Radlow & Alf [36], Larntz [24], Koehler & Larntz [22], and Read [42], among others, have shown that it can be relaxed for some members of the power divergence family. Larntz [24] found that the chi-square approximation for X^2 was reasonably accurate when expected counts were all larger than 1. Koehler & Larntz [22] found that the chi-square approximation for G^2 deteriorates sooner, leading to inflated type I error levels when many expected counts are between

1 and 5 and to deflated type I error levels and substantial loss of power when many expected counts were smaller than 1. Read [42] found that in sparse tables the chi-square distribution provided the most accurate approximation to exact P values for members of the power divergence family with λ in the interval [1/3, 3/2]. This includes X^2, but excludes G^2 and the Freeman–Tukey tests. This conclusion is supported by results reported by Kallenberg et al. [19], Rudas [46], Margolin & Light [27], and Hosmane [17]. Read & Cressie [43] recommend the power divergence statistic with $\lambda = 2/3$, but the Pearson statistic exhibits nearly the same behavior.

All members of the power divergence family of test statistics have discrete distributions, and finding a continuous distribution that provides a good general approximation for sparse tables is difficult. When a few, say k, of the r cells in a table of counts have very small expected counts, X^2 essentially behaves as if the table only had the $r - k$ cells with the large expected counts, with the exception of small probabilities at extreme values in the right tail of the distribution. The $C(m)$ distribution proposed by Cochran [6] and further studied by Yarnold [55] and Lawal & Upton [25] provides a good approximation to the distribution of X^2 in such situations. The $C(m)$ distribution is the distribution of a chi-square random variable with reduced degrees of freedom added to a weighted sum of squares involving independent Poisson random variables with small means. Although this approximation is useful for understanding the behavior of X^2 in the presence of small expected counts, it has limited practical value because percentiles of the $C(m)$ distribution depend on the number and values of the small expected counts in the table.

Given the current availability of high-speed computers, attractive alternatives to approximating the distribution of X^2, or some other member of the power divergence family, for sparse tables involve simulating the exact distribution of the statistic or computing P values from an exact conditional distribution of the test statistic. The randomization test of independence in a two-way contingency table discussed earlier is an example of an exact conditional test.

We have focused this discussion on tests of a null hypothesis against a completely general or unrestricted set of alternatives. Chi-square approximations for the distributions of X^2 and G^2 can provide reliable inferences in sparse tables of counts if the null hypothesis is tested against a suitably restricted alternative. For example, Haberman [16] showed that chi-square tests can be used to compare the fit of two nested loglinear models to sparse contingency tables (*see* **Hierarchical Models**). The basic criteria are that the number of parameters in the larger model and the difference in the number of parameters in the two models both must be small relative to the total sample size and number of cells in the table.

Correlated Responses and Complex Surveys

When a table of counts is obtained by adding across clusters and strata in a complex survey, neither the Pearson statistic nor other members of the power divergence family have asymptotic chi-square distributions when the null hypothesis is true. Rao & Scott [38] give a detailed account of the impact of survey design on the distribution of X^2 and G^2. Unless some adjustment is made to the standard chi-square approximation, positive correlations among responses within clusters lead to inflated type I error levels for both X^2 or G^2. Computing cell counts in a contingency table by combining results from different strata can have the opposite effect. If data from individual respondents are available, a Wald statistic with an asymptotic chi-square distribution can be constructed as described by Koch et al. [20]. Simulated comparisons of type I error levels and power for sparse tables, reported by Thomas & Rao [52], show that Wald tests can become unstable for cluster sampling when there are few clusters relative to the number of cells in the table, often resulting in inflated type I error levels. A **jackknife method** proposed by Fay [8] was shown to have better properties over a wide range of conditions.

Sometimes complete data for the individual respondents are not readily available and results from complex surveys are summarized as a table of counts with accompanying information of the survey **design effects**. In these situations there is not enough information to either evaluate a Wald statistic or use Fay's jackknife method. Rao & Scott [39] showed how to create approximate chi-square tests by using the information in the survey design effects to make simple adjustments to X^2 and G^2. The Thomas &

Rao [52] study also showed that these adjusted tests provide relatively accurate P values.

Roberts et al. [44] and Rao et al. [40] use similar adjustments for survey design effect to make inferences about parameters in logistic regression models fit to data from complex surveys. Morel [32] developed Wald tests for logistic regression. Gleser & Moore [13] review the literature regarding the application of goodness-of-fit tests to counts obtained from serially dependent observations and show that positive **serial correlation** also inflates type I error levels of chi-square tests.

References

[1] Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science* **7**, 131–177.

[2] Agresti, A. (1992). *Categorical Data Analysis*. Wiley, New York.

[3] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

[4] Block, S.D., Clark-Chiarelli, N., Peters, A.S. & Singer, J.D. (1996). Academia's Chilly Climate for Primary Care, *Journal of the American Medical Association* **276**, 677–682.

[5] Chernoff, H. & Lehmann, E.L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit, *Annals of Mathematical Statistics* **25**, 579–586.

[6] Cochran, W.G. (1952). The χ^2 test of goodness of fit, *Annals of Mathematical Statistics* **23**, 315–345.

[7] Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

[8] Fay, R.E. (1985). A jackknifed chi-squared test for complex samples, *Journal of the American Statistical Association* **80**, 148–157.

[9] Fienberg, S.E. (1979). The use of chi-squared statistics for categorical data problems, *Journal of the Royal Statistical Society Series B*, **41**, 54–64.

[10] Fisher, R.A. (1922). On the interpretation of χ^2 from contingency tables and the calculation of P, *Journal of the Royal Statistical Society* **85**, 87–94.

[11] Fisher, R.A. (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis, *Journal of the Royal Statistical Society* **87**, 442–450.

[12] Freeman, M.F. & Tukey, J.W. (1950). Transformations related to the angular and the square root, *Annals of Mathematical Statistics* **21**, 607–611.

[13] Gleser, L.J. & Moore, D.S. (1985). The effect of positive dependence on chi-squared tests for categorical data, *Journal of the Royal Statistical Society, Series B* **47**, 459–465.

[14] Greenwood, P.E. & Nikulin, M.S. (1996). *A Guide to Chi-Squared Testing*. Wiley, New York.

[15] Grizzle, J.E., Starmer, C.F. & Koch, G.G. (1969). Analysis of categorical data by linear models, *Biometrics* **25**, 489–504.

[16] Haberman, S.J. (1977). Log-linear models and frequency tables with small expected cell counts, *Annals of Statistics* **5**, 1148–1169.

[17] Hosmane, B. (1986). Improved likelihood ratio tests and Pearson chi-squared tests for independence in two dimensional contingency tables, *Communications in Statistics – Theory and Methods* **15**, 1875–1888.

[18] Kallenberg, W.C. (1985). On moderate and large deviations in multinomial distributions, *Annals of Mathematical Statistics* **13**, 1554–1580.

[19] Kallenberg, W.C.M., Oosterhoff, J. & Shriever, B.F. (1985). The number of classes in chi-squared goodness-of-fit tests, *Journal of the American Statistical Association* **80**, 959–968.

[20] Koch, G.G., Freeman, D.H. & Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys, *International Statistical Review* **43**, 59–78.

[21] Koehler, K.J. & Gan, F.F. (1990). Chi-square goodness-of-fit tests: cell selection and power, *Communications in Statistics – Simulation and Computation* **19**, 1265–1278.

[22] Koehler, K.J. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials, *Journal of the American Statistical Association* **75**, 336–344.

[23] Lancaster, H.O. (1969). *The Chi-Squared Distribution*. Wiley, New York.

[24] Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics, *Journal of the American Statistical Association* **73**, 253–263.

[25] Lawal, H.B. & Upton, G.J.G. (1980). An approximation to the distribution of the χ^2 goodness-of-fit statistic for use with small expectations, *Biometrika* **67**, 442–453.

[26] Mann, H.B. & Wald, A. (1942). On the choice of the number of intervals in the application of the chi-squared test, *Annals of Mathematical Statistics* **13**, 306–317.

[27] Margolin, B.H. & Light, R.L. (1974). An analysis of variance for categorical data II: small sample comparisons with chi square and other competitors, *Journal of the American Statistical Association* **69**, 755–764.

[28] Mehta, C.R. & Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $R \times C$ contingency tables, *Journal of the American Statistical Association* **78**, 427–434.

[29] Moore, D.S. (1977). Generalized inverses, Wald's method and the construction of chi-squared tests of fit, *Journal of the American Statistical Association* **72**, 131–137.

[30] Moore, D.S. (1986). Tests of chi-squared type, in *Goodness-of-Fit Techniques*, R.B. D'Agostino & M.A. Stephens, eds. Marcel Dekker, New York, pp. 63–95.

[31] Moore, D.S. & Spruill, M.C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit, *Annals of Statistics* **3**, 599–616.

[32] Morel, J. (1989). Logistic regression under complex survey designs, *Survey Methodology* **15**, 203–223.

[33] Neyman, J. (1949). Contribution to the theory of the χ^2 test. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp. 239–273.

[34] Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine* **5**(50), 157–175.

[35] Plackett, R.L. (1983). Karl Pearson and the chi-squared test, *International Statistical Review* **51**, 59–72.

[36] Radlow, R. & Alf, E.F. (1975). An alternate multinomial assessment of the accuracy of the χ^2 test of goodness of fit, *Journal of the American Statistical Association* **70**, 811–813.

[37] Rao, K.C. & Robson, D.S. (1974). A chi-squared statistic for goodness-of-fit tests within the exponential family, *Communications in Statistics – Theory and Methods* **3**, 1139–1153.

[38] Rao, J.N.K. & Scott, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data, *Annals of Statistics* **12**, 46–60.

[39] Rao, J.N.K. & Scott, A.J. (1987). On simple adjustments to chi-squared tests with sample survey data, *Annals of Statistics* **15**, 385–397.

[40] Rao, J.N.K., Kumar, S. & Roberts, G. (1989). Analysis of sample survey data involving categorical response variables: methods and software, *Survey Methodology* **15**, 161–186.

[41] Rayner, J.C.W. & Best D.J. (1989). *Smooth Tests of Goodness of Fit*. Oxford University Press, New York.

[42] Read, T.R.C. (1984). Small-sample comparisons for the power divergence goodness-of-fit statistics, *Journal of the American Statistical Association* **77**, 929–935.

[43] Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.

[44] Roberts, G., Rao, J.N.K. & Kumar, S. (1987). Logistic regression analysis of sample survey data, *Biometrika* **74**, 1–12.

[45] Roscoe, J.T. & Byars, J.A. (1971). An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic, *Journal of the American Statistical Association* **66**, 755–759.

[46] Rudas, T. (1986). A Monte Carlo comparison of the small sample behavior of the Pearson, the likelihood ratio and the Cressie–Read statistics, *Journal of Statistical Computation and Simulation* **24**, 107–120.

[47] Schorr, B. (1974). On the choice of the class intervals in the application of the chi-squared test, *Mathematische Operationsforschung und Statistik* **5**, 357–377.

[48] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[49] Stephens, M.A. (1986). Tests based on EDF statistics, in *Goodness-of-Fit Techniques*, R.B. D'Agostino & M.A. Stephens, eds. Marcel Dekker, New York, pp. 97–194.

[50] Stephens, M.A. (1986). Tests based on regression and correlation, in *Goodness-of-Fit Techniques*, R.B. D'Agostino & M.A. Stephens, eds. Marcel Dekker, New York, pp. 195–200.

[51] Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables, *Biometrika* **40**, 105–110.

[52] Thomas, D.R. & Rao, J.N.K. (1987). Small sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling, *Journal of the American Statistical Association* **82**, 630–636.

[53] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society* **54**, 426–482.

[54] Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses, *Annals of Mathematical Statistics* **9**, 60–62.

[55] Yarnold, J.K. (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution, *Journal of the American Statistical Association* **65**, 864–886.

KENNETH KOEHLER

Chi-square, Partition of

If ν is a positive integer, then the χ^2 random variable $\chi^2(\nu)$ with ν df is the sum of ν squared mutually independent standard normal random variables (*see* **Chi-square Distribution**). Thus the χ^2 variable has the additive property that, for positive integers ν and ν_j, $1 \le j \le k$, $k \ge 2$, a χ^2 variable with ν df decomposes into a sum of k mutually independent χ^2 variables, $\chi^2(\nu_j)$, with respective ν_j df, $1 \le j \le k$, if $\nu = \Sigma_{j=1}^{k} \nu_j$.

In the case of quadratic functions of normal variables, the Fisher–Cochran theorem (Cochran [1], Rao [30, p. 185]) provides a necessary and sufficient condition for a sum of squares to be decomposed into independent components with χ^2 distributions. Let Y_i, $1 \le i \le \nu$, be ν independent standard normal variables, and let $\mathbf{Y}$ be the n-dimensional column vector with coordinates Y_i for $1 \le i \le n$. For $1 \le j \le k$, let $\mathbf{A}_j$ be a symmetric $n \times n$ matrix with rank ν_j. Let

$$\mathbf{Y}'\mathbf{Y} = \sum_{j=1}^{k} \mathbf{Y}'\mathbf{A}_j\mathbf{Y}.$$

Then the $\mathbf{Y}'\mathbf{A}_j\mathbf{Y}$, $1 \le j \le k$, are independent $\chi^2(\nu_j)$ variables if and only if $\nu = \sum_{j=1}^{k} \nu_j$.

Such additive properties can be used in statistical inferential procedures. A χ^2 statistic suitable for testing a primary hypothesis of interest can be partitioned into components such that each component is a test statistic for a corresponding secondary hypothesis. Under certain conditions, such partitioning can lead to an insightful analysis of the problem under study.

Chi-Square Partitioning into Chi-Square Components

One of the earliest uses of a partition of a χ^2 statistic is found in Fisher [4, Chapter 9], who applies a χ^2 decomposition to study linkage in self-fertilized heterozygote corn plants. A general description of this approach is found in Cochran [2]. An additional early application appears in Haldane [20]. This approach is analogous to decompositions of treatment sums of squares in one-way **analysis of variance**. Let n_i, $1 \le i \le I$, be frequencies with a **multinomial distribution** with sample size N and with respective probabilities p_i. Consider the simple hypothesis that $p_i = q_i$, $1 \le i \le I$, where the probabilities q_i are positive and have sum 1. Let $m_i = Nq_i$ be the expectation of n_i under the null hypothesis. The classical Pearson χ^2 statistic is then

$$X^2 = \sum_{i=1}^{I} (n_i - m_i)^2/m_i$$

(*see* **Chi-square Tests**). As is well known, under the null hypothesis, X^2 has an asymptotic $\chi^2(I - 1)$ distribution. Fisher suggests use of fixed scores x_{ij}, $1 \le i \le I$, $1 \le j \le I - 1$, such that

$$\sum_{i=1}^{I} q_i x_{ij} = 0, \quad 1 \le j \le I - 1,$$

$$\sum_{i=1}^{I} q_i x_{ij} x_{ik} = 0, \quad 1 \le j < k \le I - 1,$$

and

$$d_j = \sum_{i=1}^{I} q_i x_{ij}^2 > 0, \quad 1 \le j \le I - 1.$$

If

$$Y_j = \left[\sum_{i=1}^{k} x_{ij} n_i\right]^2 \Big/ (nd_j^2), \quad 1 \le j \le I - 1,$$

then

$$X^2 = \sum_{j=1}^{I-1} Y_j.$$

Under the null hypothesis the Y_j, $1 \le j \le I - 1$, are asymptotically mutually independent random variables with χ^2 (1) distributions. This decomposition is particularly useful if the linear combinations $\sum_{i=1}^{I} p_i x_{ij}$, $1 \le j \le I - 1$, are meaningful in the problem under study.

Chi-square partitions have been considered for the χ^2 test of independence of two polytomous variables. Consider an $I \times J$ **contingency table** with frequencies n_{ij}, $1 \le i \le I$, $1 \le j \le J$. Let the frequencies have a multinomial distribution with sample size N and respective positive probabilities p_{ij}. Let

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$$

and

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij}.$$

Let $\hat{m}_{ij} = n_{i\cdot} n_{\cdot j}/N$. The customary Pearson χ^2 test statistic,

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - \hat{m}_{ij})^2/\hat{m}_{ij},$$

and the customary **likelihood ratio test** statistic,

$$L^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log(n_{ij}/\hat{m}_{ij}),$$

both have asymptotic $\chi^2_{(I-1)(J-1)}$ distributions under the null hypothesis of independence of row and column variables. An early suggestion of a partition appears in Pearson [28], who suggests the decomposition

$$X^2 = \sum_{i=1}^{I} X_i^2,$$

where

$$X_i^2 = \sum_{j=1}^{J} (n_{ij} - \hat{m}_{ij})^2/\hat{m}_{ij}.$$

This decomposition is exact, but the components do not have asymptotic χ^2 distributions. As in Lancaster [26] and Williams [31], an exact partition of X^2 may be based on the approach of Fisher [4]. Conditional on the observed marginal totals $n_{i\cdot}$ and $n_{\cdot j}$, select scores $x_{ii'}, 1 \le i \le I, 1 \le i' \le I-1$ and $y_{jj'}, 1 \le j \le J, 1 \le j' \le J-1$, so that

$$\sum_{i=1}^{I} n_{i\cdot} x_{ii'} = 0, \quad 1 \le i' \le I-1,$$

$$\sum_{i=1}^{I} n_{i\cdot} x_{ii'} x_{ii''} = 0, \quad 1 \le i' < i'' \le I-1,$$

$$c_i = \sum_{i=1}^{I} n_{i\cdot} x_{ii'}^2 > 0, \quad 1 \le i' \le I-1,$$

$$\sum_{j=1}^{J} n_{\cdot j} y_{jj'} = 0, \quad 1 \le j' \le J-1,$$

$$\sum_{j=1}^{J} n_{\cdot j} y_{jj'} y_{jj''} = 0, \quad 1 \le j' < j'' \le J-1,$$

$$d_j = \sum_{j=1}^{J} n_{\cdot j} y_{jj'} > 0, \quad 1 \le j' \le J-1.$$

Let

$$X^2_{i'j'} = \frac{N \left(\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} x_{ii'} y_{jj'} \right)^2}{c_{i'} d_{j'}},$$

$$1 \le i' \le I, 1 \le j' \le J.$$

Then

$$X^2 = \sum_{i'=1}^{I-1} \sum_{j'=1}^{J-1} X^2_{i'j'}. \tag{1}$$

Under the null hypothesis, the $X^2_{i'j'}$ are asymptotically mutually independent $\chi^2_{(1)}$ random variables. Cochran [3] provides a number of analyses which exploit this decomposition technique for applications such as testing for a linear trend in binomial proportions across levels of a quantitative covariate (*see* **Trend Test for Counts and Proportions**) and testing conditional independence between two binary variables while controlling for a third variable.

Lancaster [26] uses a special case of (1) to suggest an approximate partition of X^2 based on conventional χ^2 tests of independence for two-by-two contingency tables formed from the original table. This approximate partition is considered in Cochran [2], Kimball [24], and Kastenbaum [23], among others.

Kullback [25, pp. 173–174] and Gabriel [5] use L^2 to obtain decompositions based on tests of independence of subtables and tests of independence of collapsed tables. Let U be a partition of the integers 1 to I into p nonempty disjoint sets, and let V be a partition of the integers 1 to J in q nonempty disjoint sets. For a nonempty subset A of the integers 1 to I and a nonempty subset B of the integers 1 to J, let

$$n_{AB} = \sum_{i \in A} \sum_{j \in B} n_{ij}, \qquad n_{A\cdot} = \sum_{i \in A} n_{i\cdot},$$

$$n_{\cdot B} = \sum_{j \in B} n_{\cdot j}, \qquad n_{Aj} = \sum_{i \in A} n_{ij}, \quad 1 \le j \le J,$$

and

$$n_{iB} = \sum_{j \in B} n_{ij}, \quad 1 \le i \le I.$$

Let

$$L^2_{\cdot\cdot} = 2\sum_{A\in U}\sum_{B\in V} n_{AB}\log\left(\frac{Nn_{AB}}{n_{A\cdot}n_{\cdot B}}\right)$$

be the *generalized likelihood ratio* χ^2 statistic for testing the hypothesis $H_{\cdot\cdot}$ of independence for the collapsed $p \times q$ table with elements n_{AB} for A in U and B in V, let

$$L^2_{A\cdot} = 2\sum_{i\in A}\sum_{B\in V} n_{iB}\log\left(\frac{n_{iB}n_{A\cdot}}{n_{i\cdot}n_{\cdot B}}\right), \quad A \in U,$$

be the likelihood ratio χ^2 statistic for testing the hypothesis $H_{A\cdot}$ of independence for the column-collapsed subtable with elements n_{iB} for i in A and B in V, let

$$L^2_{\cdot B} = 2\sum_{A\in U}\sum_{j\in B} n_{Aj}\log\left(\frac{n_{Aj}n_{\cdot B}}{n_{A\cdot}n_{\cdot j}}\right), \quad B \in V,$$

be the likelihood ratio χ^2 statistic for testing the hypothesis $H_{\cdot B}$ of independence for the row-collapsed subtable with elements n_{Aj} for A in U and j in B, and let

$$L^2_{AB} = 2\sum_{i\in A}\sum_{B\in V} n_{iB}\log\left(\frac{n_{iB}n_{A\cdot}}{n_{i\cdot}n_{\cdot B}}\right), \quad A \in U, B \in V,$$

be the likelihood ratio χ^2 statistic for testing the hypothesis H_{AB} of independence for the subtable with elements n_{ij} for i in A and j in B.

Independence of the row and column variables of the original table holds if and only if $H_{\cdot\cdot}$ holds, $H_{A\cdot}$ holds for all A in U, $H_{\cdot B}$ holds for all B in V, and H_{AB} holds for all A in U and B in V. Thus

$$L^2 = L^2_{\cdot\cdot} + \sum_{A\in U} L^2_{A\cdot} + \sum_{B\in V} L^2_{\cdot B} + \sum_{A\in U}\sum_{B\in V} L^2_{AB}.$$

This decomposition may help to explain the dependence in the original table in terms of dependence in selected subtables and collapsed tables.

Let $f(A)$ denote the number of elements in a set A. Let

$$\nu_{\cdot\cdot} = (p-1)(q-1), \qquad \nu_{A\cdot} = [f(A)-1](q-1),$$

$$\nu_{\cdot B} = (p-1)[f(B)-1],$$

and

$$\nu_{AB} = [f(A)-1][f(B)-1],$$

so that

$$(I-1)(J-1) = \nu_{\cdot\cdot} + \sum_{A\in U}\nu_{A\cdot} + \sum_{B\in V}\nu_{\cdot B} + \sum_{A\in U}\sum_{B\in V}\nu_{AB}.$$

Adopt the convention that a χ^2 variable with zero degrees of freedom is a random variable which is always zero. The components in this decomposition are asymptotically independent under the null hypothesis and have asymptotic χ^2 distributions. Under the null hypothesis, the following asymptotic approximations apply:

$$L^2_{\cdot\cdot} \sim \chi^2(\nu_{\cdot\cdot}), \qquad L^2_{A\cdot} \sim \chi^2(\nu_{A\cdot}), \qquad L^2_{\cdot B} \sim \chi^2(\nu_{\cdot B}),$$

and

$$L^2_{AB} \sim \chi^2(\nu_{AB}).$$

Iverson [22] provides examples of repeated application of this decomposition. Replacement of likelihood ratio χ^2 statistics with Pearson χ^2 statistics leads to an approximate partition of χ^2 that includes Lancaster's decomposition as a special case. Gilula & Krieger [9, 10] derive a different yet analogous way of partitioning χ^2 which is exact for both L^2 and X^2.

Decompositions of the Pearson χ^2 have been much less successful in the case of multiway contingency tables, with the notable exception of Cochran's [3] decomposition of the Pearson χ^2 for conditional independence. An early attempt by Lancaster [27] was shown by Plackett [29] to be unsatisfactory. Useful χ^2 decompositions for multiway tables appear in Kullback [25, pp. 159–171], Goodman [11–14] and Haberman [18, Chapter 4]. These decompositions are based on likelihood ratio tests.

The likelihood ratio χ^2 provides a very general source of partitions of χ^2. In a typical example, a d-dimensional parameter $\boldsymbol{\theta}$ with coordinates θ_j, $1 \le j \le d$, is used to specify the distribution of some n-dimensional random vector $\mathbf{Y}$. The parameter $\boldsymbol{\theta}$ is in the interior of a d-dimensional parameter space Θ. Associated with the observation $\mathbf{Y}$ is a likelihood function L on Θ. Let $\boldsymbol{\gamma}$ be a member of the interior of Θ, and let γ_i be the ith coordinate of $\boldsymbol{\gamma}$. For each integer q from 0 to d, consider the hypothesis H_q that $\theta_i = \gamma_i$, $q < i \le d$. Let M_q be the maximum value of the likelihood $L(\hat{\boldsymbol{\theta}}_q)$ under the condition that $\hat{\boldsymbol{\theta}}_q$ is in Θ and has ith coordinate equal to γ_i

for $i > q$. Let $q(j), 1 \le j \le k, k > 1$, be integers such that $0 \le q(1), q(k) \le d$, and $q(j) < q(j+1)$ for $1 \le j < k$. Let

$$L^2(q, r) = 2\log(M_r/M_q)$$

be the likelihood ratio χ^2 for the null hypothesis that H_q and the alternative hypothesis H_r for $0 \le q < r \le d$. Integers q and r are given such that $r < q < d$. For any $q(j), 1 \le j \le k, k \ge 2$, such that $0 \le q(1), q(k) \le d$, and $q(j) < q(j+1), 1 \le j < k$, one obtains the decomposition

$$L^2[q(1), q(k)] = \sum_{j=1}^{k-1} L^2[q(j), q(j+1)].$$

Under common regularity conditions, $L^2[q(1), q(k)]$ and $L^2[q(j), q(j+1)]$ have asymptotic χ^2 distributions under $H_{q(1)}$, and the $L^2[q(j), q(j+1)]$ are asymptotically independent.

Of particular interest is the decomposition of χ^2 in Goodman [12] for a three-way contingency table. Let n_{ijk} be an $I \times J \times K$ contingency table which represents a cross classification of the polytomous variables A, B, and C, respectively, so that n_{ijk} is the number of observations with $A = i$, $B = j$, and $C = k$. In the proposed decomposition, H_1 is the hypothesis that A, B, and C are mutually independent, H_2 is the hypothesis that A and B are jointly independent of C, H_3 is the hypothesis that A and C are conditionally independent given B, H_4 is the hypothesis that A, B, and C are related by a model of no three-factor interaction, and H_5 is the saturated model that makes no assumptions about the relationships among A, B, and C. One obtains the decomposition

$$L^2(1,5) = L^2(1,2) + L^2(2,3) + L^2(3,4) + L^2(4,5).$$

In this decomposition, $L^2(1,2)$ is the conventional likelihood ratio χ^2 for a test of marginal independence of A and B, $L^2(2,3)$ is the conventional likelihood ratio χ^2 for a test of independence of B and C, $L^2(3,4)$ is the test of interaction of A and C given validity of the model of no three-factor interaction, and $L^2(4,5)$ is the test of validity of the model of no three-factor interaction. The statistic $L^2(1,5)$ is a test of validity for the model of mutual independence of A, B, and C. For generalizations to more complex tables, see [13] and [14].

Partitioning of χ^2 can also be used in contexts different from contingency tables and categorical data (e.g. [6]).

Chi-Square Partitioning into Non-Chi-Square Components

Hirschfeld [21] provides an alternative decomposition of the Pearson χ^2 based on canonical correlations. This decomposition forms the basis of **correspondence analysis** [17]. Consider the two-way table with frequencies $n_{ij}, 1 \le i \le I, 1 \le j \le J$, sample size N, and cell probabilities p_{ij}. Let

$$p_{i\cdot} = \sum_{j=1}^{J} p_{ij}, \quad 1 \le i \le I,$$

and

$$p_{\cdot j} = \sum_{i=1}^{I} p_{ij}, \quad 1 \le j \le J,$$

denote the row and column marginal probabilities, respectively. Let $K = \min(I-1, J-1)$. Then the canonical decomposition of p_{ij} is

$$p_{ij} = p_{i\cdot}p_{\cdot j}\left(1 + \sum_{k=1}^{K} \rho_k t_{ik} u_{jk}\right), \tag{2}$$

where ρ_k is nonincreasing in k,

$$\sum_{i=1}^{I} p_{i\cdot} t_{ik} = \sum_{j=1}^{J} p_{\cdot j} u_{jk} = 0, \quad 1 \le k \le K,$$

and

$$\sum_{i=1}^{I} p_{i\cdot} t_{ik} t_{ik'} = \sum_{j=1}^{J} p_{\cdot j} u_{jk} u_{jk'} = \begin{cases} 1, & k = k', \\ 0, & k \ne k'. \end{cases}$$

The canonical decomposition of n_{ij} is

$$n_{ij} = \hat{m}_{ij}\left(1 + \sum_{k=1}^{K} r_k \hat{t}_{ik} \hat{u}_{jk}\right), \tag{3}$$

where $|r_k|$ is nonincreasing in k,

$$\sum_{i=1}^{I} n_{i\cdot} \hat{t}_{ik} = \sum_{j=1}^{J} n_{\cdot j} \hat{u}_{jk} = 0, \quad 1 \le k \le K,$$

and

$$\sum_{i=1}^{I} n_{i\cdot}\hat{t}_{ik}\hat{t}_{ik'} = \sum_{j=1}^{J} n_{\cdot j}\hat{u}_{jk}\hat{u}_{jk'} = \begin{cases} N, & k = k', \\ 0, & k \neq k'. \end{cases}$$

With this decomposition, the Pearson χ^2 statistic X^2 for testing independence of row and column variables satisfies

$$X^2 = N\sum_{k=1}^{K} r_k^2 = \sum_{k=1}^{K} W_k^2,$$

where

$$W_k^2 = N^{-1}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\hat{t}_{ik}\hat{u}_{jk}\right)^2.$$

Unless $K = 1$, this decomposition does not lead to components with χ^2 distributions, as shown in Haberman [19]. Nonetheless, the size of the W_k^2 does provide an indication as to which ρ_k may be zero. Goodman [15, 16] and Gilula & Haberman [7, 8], among others, consider models in which, for some integer q from 0 to K, it is assumed that $\rho_k = 0$ for $k > q$.

References

[1] Cochran, W.G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance, *Proceedings of the Cambridge Philosophical Society* **30**, 178–191.

[2] Cochran, W.G. (1952). The χ^2 test of goodness of fit, *Annals of Mathematical Statistics* **23**, 315–345.

[3] Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.

[4] Fisher, R.A. (1925). *Statistical Methods for Research Workers*, 1st Ed. Oliver & Boyd, Edinburgh.

[5] Gabriel, K.R. (1966). Simultaneous test procedures for multiple comparisons on categorical data, *Journal of the American Statistical Association* **61**, 1081–1096.

[6] Gilula, Z. (1985). On the analysis of heterogeneity among populations, *Journal of the Royal Statistical Society, Series B* **47**, 15–23.

[7] Gilula, Z. & Haberman, S.J. (1986). Canonical analysis of contingency tables by maximum likelihood, *Journal of the American Statistical Association* **81**, 780–788.

[8] Gilula, Z. & Haberman, S.J. (1988). The analysis of multiway contingency tables by restricted canonical and restricted association models, *Journal of the American Statistical Association* **83**, 760–771.

[9] Gilula, Z. & Krieger, A.M. (1983). The collapsibility and monotonicity of Pearson's chi-square for collapsed contingency tables with applications, *Journal of the American Statistical Association* **78**, 176–180.

[10] Gilula, Z. & Krieger, A.M. (1989). Collapsed contingency tables and the chi-square reduction principle, *Journal of the Royal Statistical Society, Series B* **51**, 425–434.

[11] Goodman, L.A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with or without missing entries, *Journal of the American Statistical Association* **63**, 1091–1131.

[12] Goodman, L.A. (1969). On partitioning χ^2 and detecting partial association in three-way contingency tables, *Journal of the Royal Statistical Society, Series B* **31**, 486–498.

[13] Goodman, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications, *Journal of the American Statistical Association* **65**, 226–256.

[14] Goodman, L.A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables, *Journal of the American Statistical Association* **66**, 339–344.

[15] Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories, *Journal of the American Statistical Association* **76**, 320–334.

[16] Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *Annals of Statistics* **13**, 10–69.

[17] Greenacre, M.J. (1984). *Theory and Application of Correspondence Analysis*. Academic Press, New York.

[18] Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

[19] Haberman, S.J. (1981). Tests for independence in two-way contingency tables based on canonical correlation and on linear by linear interaction, *Annals of Statistics* **9**, 1178–1186.

[20] Haldane, J.B.S. (1939). Note on the preceding analysis of Mendelian segregations, *Biometrika* **31**, 67–71.

[21] Hirschfeld, H.O. (1935). A connection between correlation and contingency, *Proceedings of the Cambridge Philosophical Society* **31**, 520–524.

[22] Iverson, G.R. (1979). Decomposing chi-square: a forgotten technique, *Sociological Methods and Research* **8**, 143–157.

[23] Kastenbaum, M.A. (1960). A note on the additive partitioning of chi-square in contingency tables, *Biometrics* **16**, 416–422.

[24] Kimball, A.W. (1954). Short-cut formulas for the exact partition of χ^2 in contingency tables, *Biometrics* **10**, 452–458.

[25] Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[26] Lancaster, H.O. (1949). The derivation and partition of χ^2 in certain discrete distributions, *Biometrika* **36**, 117–129.

[27] Lancaster, H.O. (1951). Complex contingency tables treated by the partition of χ^2, *Journal of the Royal Statistical Society, Series B* **13**, 242–249.

[28] Pearson, K. (1906). On a coefficient of class heterogeneity or divergence, *Biometrika* **5**, 198–203.

[29] Plackett, R.L. (1962). A note on interactions in contingency tables, *Journal of the Royal Statistical Society, Series B* **24**, 162–166.

[30] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. Wiley, New York.

[31] Williams, E.J. (1952). Use of scores for the analysis of association in contingency tables, *Biometrika* **39**, 274–289.

(*See also* **Loglinear Model**)

Z. GILULA & SHELBY J. HABERMAN

Choi–Williams Distribution

The Choi–Williams distribution is a transform that represents the **spectral** content of nonstationary signal with a bidimensional time–frequency map. How well a time-frequency distribution represents a nonstationary signal depends on how fast the changes in the frequency content of the signal are. When they are relatively slow, the standard method of representation is the spectrogram, that is, the calculation of the Fourier (*see* **Fast Fourier Transform (FFT)**) spectrum over a short-time running window. When the spectral characteristics change more rapidly, a higher resolution in time and frequency is needed, and methods like the **Wigner–Ville Distribution** are preferred. A large class of time–frequency distributions of a signal $s(t)$ can be defined as

$$P(t,\omega) = \frac{1}{4\pi^2}\iiint s^*\left(t-\frac{1}{2}\tau\right)s\left(t+\frac{1}{2}\tau\right) \times e^{-j\theta t - j\omega\tau + j\theta u}\phi(\theta,\tau)\,\mathrm{d}u \times \mathrm{d}\tau \times \mathrm{d}\theta \quad (1)$$

where $s^*(t)$ is the complex conjugate of $s(t)$, ω is the angular frequency $2\pi f$, and $\phi(\theta,\tau)$ is a kernel that defines the type of distribution. For

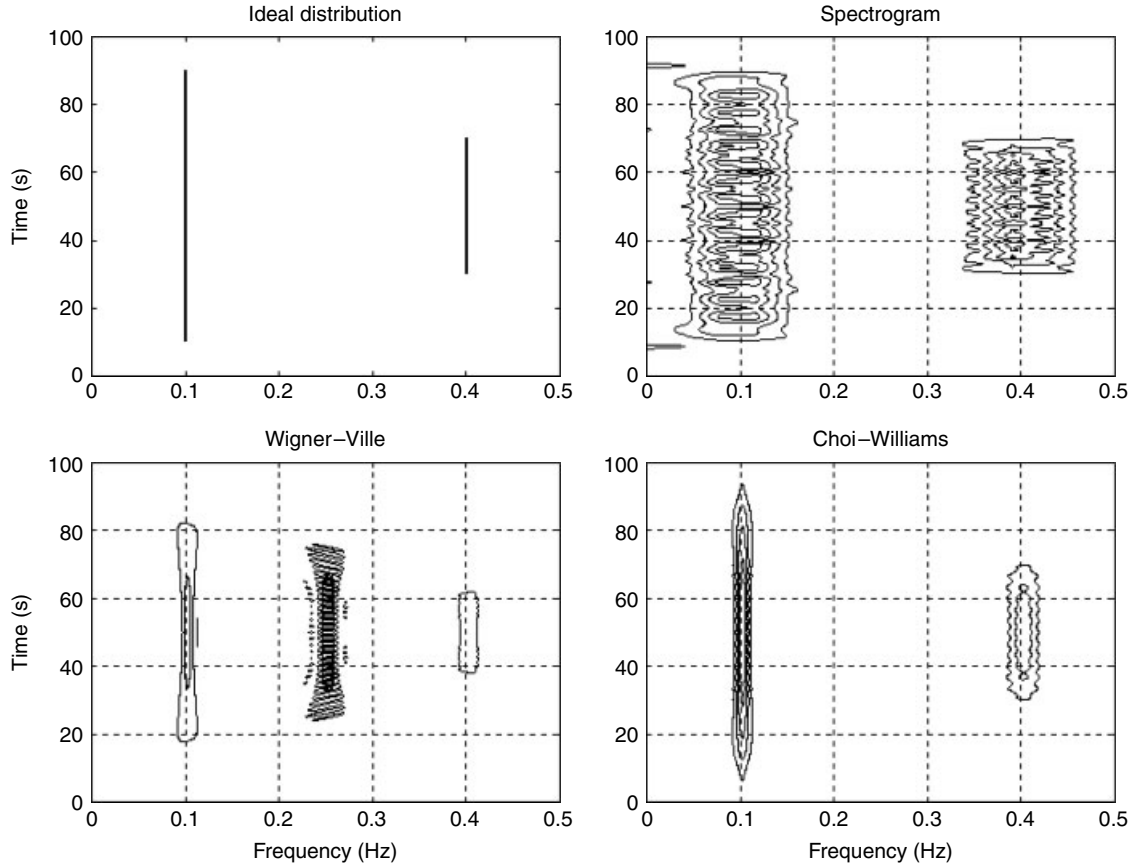

Figure 1 Comparison of time–frequency distributions for a multicomponent signal $s(t) = s_1(t) + s_2(t)$, where $s_1(t)$ is a 0.1-Hz sinusoid defined between 10 and 90 s and $s_2(t)$ is a 0.4-Hz sinusoid defined between 30 and 70 s. One expects an "ideal" distribution to show just two components, one at 0.1 Hz and one at 0.4 Hz, within the time periods 10–90 s and 30–70 s respectively. The spectrogram is a low-resolution approximation of the ideal distribution; the Wigner–Ville distribution shows a better resolution but also significant interferences; the Choi–Williams distribution provides a large suppression of interference terms with only a little loss in resolution

instance, (1) defines the Wigner–Ville distribution when $\phi(\theta, \tau) = 1$ and a spectrogram with running window $h(t)$ when $\phi(\theta, \tau) = \int h^*(u - \frac{1}{2}\tau)h(u + \frac{1}{2}\tau)e^{-j\theta u}\,\mathrm{d}u$ [4].

Choi and Williams started from (1) to address one of the main problems of the Wigner–Ville distribution: the presence of interference terms. These are spurious values indicating powers in regions of the time–frequency plane where one would expect zero values, and they are particularly prevalent in multicomponent signals. Their approach was to find the kernel $\phi(\theta, \tau)$ that minimizes the interference terms while still retaining the desirable properties of the Wigner–Ville distribution. They proposed the kernel

$$\phi(\theta, \tau) = e^{-\theta^2\tau^2/\sigma} \tag{2}$$

where σ is a parameter controlling the attenuation of interference terms [3]; the smaller σ is, the more the interferences are suppressed. Unfortunately, increased attenuation also leads to a loss of resolution in the time–frequency plane. By properly tuning σ, one can minimize the interferences and retain a sufficient time–frequency resolution.

Figure 1 compares the Choi–Williams distribution of a multicomponent signal with the spectrogram and the Wigner–Ville distribution. The Choi–Williams representation of the signal can largely suppress interference terms while still preserving a high resolution.

Choi–Williams distributions were used to describe the changes in the frequency content of several biological signals (*see* **Clinical Signals**), like the sounds produced by muscles [1] or the structure of heart-rate signals during changes in the autonomic tone [2].

References

[1] Barry, D.T. & Cole, N.M. (1990). Muscle sounds are emitted at the resonance frequency of the skeletal muscle, *IEEE Transactions on Biomedical Engineering* **37**, 525–531.

[2] Chan, H.L., Huang, H.H. & Lin, J.L. (2001). Time-frequency analysis of heart-rate variability during transient segments, *Annals of Biomedical Engineering* **29**, 983–996.

[3] Choi, H.I. & Williams, W.J. (1989). Improved time/ frequency representation of multicomponent signals using exponential kernels, *IEEE Transactions on Acoustic, Speech and Signal Processing* **37**, 862–971.

[4] Cohen, L. (1989). Time-frequency distributions–a review, *Proceedings of the IEEE* **77**, 941–981.

PAOLO CASTIGLIONI

Chronic Disease Models

Modeling chronic disease in human populations requires different substantive and technical principles than in modeling acute infectious and epidemic disease (e.g. [5, 54]). One is that the natural history of a chronic disease is a significant portion of the life span, e.g. atherosclerosis may start by ages 10–15 – autopsy studies of accident victims showed well developed atheromas in males aged 20–25 [74]. Lung tumors may initiate 10–50 years before clinical expression [55] (*see* **Latent Period**). Evidence is mounting that infectious agents can increase chronic disease risks. *Helicobacter pylori* (discovered in 1983) is a causative factor, not only in gastritis, but also in gastric ulcers [35], cancer [28, 73], lymphoma, and liver cancer [69]. Antibodies to *Chlamydia pneumoniae*, a pathogen causing pneumonia, bronchitis, pharyngitis, and sinusitis, have been isolated from atheromas [31]. Some models of long-term changes in circulatory disease risk (and certain cancers) implicate livestock viral infections as well as chronic infection with CMV or Epstein–Barr virus (e.g. [64]).

Because of the long natural history of chronic disease, there is more potential to interact with other diseases (e.g. diabetes and atherosclerosis), risk factor exposures (e.g. cigarette **smoking**; obesity), and age-related declines in physiology (e.g. senescent amyloid changes in the myocardium [38, 43], and temporally accumulated ischemic damage [46]). At late ages (e.g. 95+ years), distinguishing between chronic diseases and age-related losses of physiological function is difficult without biologically informed dynamic models (*see* **Mathematical Biology, Overview**) (e.g. [38, 47]), because multiple age-dependent interactions exist between chronic diseases processes – interactions perturbed by exogenous factors. Thus, it is necessary to model disease behavior as a multidimensional stochastic system evolving according to internally programmed age dynamics [91] and influenced by environmentally induced damage [98].

Chronic disease model applications encounter statistical problems such as (i) estimating parameters for high dimensional processes from partial, cross-temporal data measured with error, and (ii) the interdependence of mortality and state dynamic processes. **Longitudinal** studies of human chronic diseases do not provide sufficiently fine grained and lengthy **time series** data on physiological changes to construct models without extensive theory or ancillary data [79]. This is because the measurement burden on subjects in practice limits the number of time points sampled on chronic disease processes (e.g. every 1, 2, or 5 years) compared with the usual density of samples of processes in time series analyses (e.g. [2]). Additionally, missing data [71] problems arise as a result of left and right **censoring**.

Consequently, it is necessary to combine temporally sparse observations made of a disease with other empirical and theoretical information in a global **likelihood** function (e.g. [53]). This allows data sets with different observational strengths and measurement characteristics to be combined in a model of a complex system subject to chronic functional loss on multiple dimensions, and multiple functionally dependent modes of failure. Combined correctly, the complementary strengths of data sets can improve parameter estimates (e.g. [51]). Model-based data combining is different than in **meta-analysis of clinical trials** (e.g. [19, 20, 34]) in that instead of increasing statistical **power** about a hypothesis (e.g. the effects of chemotherapy on breast cancer survival) by "pooling" cases from observationally equivalent studies, the precision of the parameters in a highly structured process model are improved using data sets covering *heterogeneous* age/time and substantive domains. This is akin to the use of stochastic compartment models in **pharmacokinetics** (e.g. [39]), and complex biological systems (e.g. [61]), to make estimates of transitions through intermediate, unobserved states from the correlation of the temporal scheduling of multiple system inputs and outputs. In such models the focus is on "temporal" homogeneity, i.e. observations of individuals in different stages of a process but following similar trajectories. Thus, all individuals do not have identically the same parameter values, or exist in the same stage of the process; rather, processes are assumed to be described by the same systems of partial differential equations.

Because of these, and other, statistical issues modeling chronic disease requires specialized analytic and data-collection procedures. In what follows we discuss several data sets and the chronic disease features they describe. We then discuss data use in chronic disease modeling.

Observation Plans for Chronic Disease Processes

We review the properties of four observational plans.

Longitudinal Cohort Studies

The classic longitudinal **cohort study** of chronic disease and risk factors is the **Framingham Heart Study**, begun in 1949–1950 [16], of a cohort of 2336 males and 2873 females initially aged 29–62 years, where multiple risk factors (e.g. blood glucose, blood pressure, cholesterol, cigarette smoking, hematocrit, vital capacity, BMI) were assessed biennially. In addition, the times to health events (some clinically determined, e.g. by EKG) or death are recorded. The exams now constitute a long time series, covering a large portion of the adult life span, on changes of multiple risk factors. Biennial assessments allowed frequent updating of risk factor–disease relationships (e.g. [41]). Biennial data have also been used to update risk factor–disease relationships by using concurrent values at each wave and, assuming conditional (on measured risk factor values) independence of events in subsequent biennial periods, treating each interval as an observation on the process [96]. Though improving **logistic regression** estimates, this does not use data on risk factor changes – and their cross-temporal covariances – to estimate the age trajectories of measured risk factors, or latent state variables [14]. Feskens et al. [25, 26] showed how longitudinal data could be used to infer the parameters of the individual's physiological dynamics.

Over time, the Framingham study changed. Risk factors were added (e.g. assessments of cholesterol components such as high and low density lipoproteins) and some dropped (e.g. uric acid about wave 4). Ancillary samples (e.g. Framingham offspring) were generated. Other risk factors [e.g. Lp(a) [15, 83] and homocysteinemia [7]; fibrinogen and hyperhomocysteinemia interactions [90]] were less completely assessed so that hazard estimates may be affected by unobserved or incompletely observed risk factors.

Other longitudinal studies were started in the 1960s, often with similar risk factors (e.g. blood pressure, smoking, cholesterol, weight and height), but with longer intervals between measurements. Follow-up periods of five to 10 years were used in the Seven Country [45] and Charleston Heart Studies [44]. Longer intervals make models describing risk factor and disease dynamics for incompletely observed processes more important.

Additionally, focus shifted from examining circulatory disease risks in middle-aged (generally male) populations, to studying a range of diseases, and total mortality, in elderly male and female populations. This was because early goals were reached (i.e. significant associations of risk factors with circulatory diseases, such as cholesterol with CHD, were demonstrated) and because many studies were initiated during a period (1954–1968) when US male mortality *increased* 0.2% per year – and shortly after the 1930s and 1940s when male CHD risk increased in the US and the UK [42]. Indeed, US life expectancy was thought as possibly having reached biological limits in the 1960s [67]. Limits were built into the 1974 Social Security Trust Fund projections [66]. Social epidemiologists theorized that chronic disease risks were intrinsic to industrial societies (e.g. [18] and [70]).

In fact, female mortality declined 0.8% per annum from 1954 to 1968. Adult male CHD mortality began declining after 1968. From 1950 to 1991, age standardized US heart disease mortality declined 53%; stroke mortality 70% [68]. However, the causes of such declines could not be identified with existing data [88]. Early CVD risk factor interventions often did not reduce total mortality [65], sometimes due to adverse effects on other chronic diseases (e.g. diuretics initially used to lower blood pressure adversely affected mortality in diabetics, [92]), making it necessary to jointly examine multiple causes of death and multiple health outcomes (e.g., JAMA estrogen and program study, July 17, 2002). As **life expectancy** above age 65 (and 85) increased in the US and other developed countries (e.g. Japan, France, Sweden [60]), the need to extend existing cohort studies to describe risk factor–disease relationships to late ages became evident (*see* **Gerontology and Geriatric Medicine**). Rapid physiological change, the multiplicity of interacting chronic disease *and* disability processes, and the high levels of mortality at ages 80+ emphasized the need for biologically motivated, gender-specific models [56].

Sequential and Parallel Cross-Sectional Studies

If the nature of chronic disease–risk factor relationships were established in cohort studies, other studies

were needed to assess risk factor variation across populations. The WHO "MONICA" (Monitoring of trends and determinants in cardiovascular disease) studies assessed age and gender-specific trends in CVD risk factors in developing and developed countries (*see* **Surveillance of Diseases**). These studies, started 1985–1987, monitored mortality and risk factor changes in men and women aged 25–64 for 10 years. Forty-one centers in 27 countries used standardized protocols to monitor 118 populations containing 15 million persons aged 25–64. Two null hypotheses were examined. First that there was no relationship between 10-year trends in cholesterol, blood pressure, and smoking and CVD. Some centers examined HDL and LDL cholesterol and physical activity. The second was that there was no relationship between 10-year trends in 28 day case fatality rates and acute coronary care.

To evaluate the first hypothesis risk factor surveys were to be done at the beginning and end of the study period by drawing independent gender-specific **random samples** of at least 200 cases for each 10 years of age category from 25 to 64 from target populations. Survey data collection was standardized for demographic variables, smoking status, hypertension (and medications), blood pressure, cholesterol, serum thiocyanate, height, and weight. Ten-year risk factor trends were evaluated from distributional (means and variances) changes between the two independent samples.

Evaluation of the second hypothesis required sufficient events to have an 80% power for detecting a 2% per annum change in CVD incidence over 10 years, significant at a 5% level using a two-tailed test (*see* **Sample Size Determination**). Thus, each study had to generate a minimum of 240 morbid events per year for ages 25–64. Depending on local CVD rates this required populations of about 300 000. Standardized procedures were used to monitor disease events, and acute coronary care [89]. Morbid events were monitored using **death certificates** and by identifying heart attack and other CVD events from hospital records with either "hot" (i.e. screening admissions) or "cold" (i.e. abstracting discharge records) pursuit. Results on myocardial infarctions and coronary deaths reported for 38 populations in 21 countries refute the suggestion that high CHD rates are associated with high **case fatality** rates [89]. This study design has the problem that independent sampling of the continuous risk factor distribution at 0 and 10 years provides little power to distinguish risk factor changes due to period vs. cohort effects (*see* **Age–Period–Cohort Analysis**).

Measurement and Modeling. If models of risk factor and mortality processes are not used to analyze disease–risk dynamics, then additional measurement issues arise. Law et al. [49] examined the relationship between cholesterol and IHD mortality in 21 535 men followed from 1975 to 1982. Cholesterol was measured twice over three years on 5696 men with LDL cholesterol assessed in the second examination. Although cholesterol is "stable" (in contrast, say, to systolic blood pressure) it exhibited significant **regression to the mean**, i.e. the cholesterol–IHD relationship was underestimated by 41% (equivalent to the ratio of total, to between person, variance of cholesterol). Surrogate dilution bias arose by measuring total cholesterol instead of LDL (*see* **Measurement Error in Epidemiologic Studies**). Total cholesterol represents the net effect of trends in HDL ("good") and LDL ("bad") cholesterol. Surrogate dilution bias reduced the estimated effect of cholesterol on mortality by 14%. Bias was found at all ages examined. In a **stochastic process** model, regression to the mean would be represented as the balance of autoregressive (*see* **ARMA and ARIMA Models**) and diffusive forces and surrogate dilution bias by influential unmeasured variables [1]. The short study period suggests mortality selection bias would be small.

Law et al. [48] examined lags between cholesterol and IHD risk, reductions using incidence and cholesterol data from 10 prospective (cohort) studies, three international studies in multiple communities, and 28 randomized controlled trials. Estimates of regression to the mean, and surrogate dilution, bias were used to adjust observational study results. In cohort studies a 10% decrease in cholesterol was associated with a 27% decline in IHD (adjusted for blood pressure and smoking). This relationship held to at least age 80 (where the IHD reduction was 19%). In randomized trials, after five years, IHD risk reduction was 25%, or 92.6% of the total effect in cohort studies. Thus, not only was atherosclerosis reversible (e.g. [8] and [72]), but reversal could occur rapidly – even before the structural regression of atheromas [3, 6]. This suggests that the rate of progression of the physiological dynamics of IHD can be changed sufficiently in five years to prevent

pathological processes from passing a clinical threshold. Also, it suggested that atherosclerotic disease affected acute (e.g. within four months) physiological changes (e.g. vasodilation and vasoconstriction due to neuroendocrine factors [6]) interacting with chronic processes.

National Longitudinal Surveys

Risk factor–chronic disease relationships can be statically modeled using nationally representative surveys (*see* **Surveys, Health and Morbidity**). The US National Health and Nutrition Examination Surveys were done four times 1960–1962 to 1990–1994. The first, the National Health Examination Survey (NHESI; $N = 7710$ adults age 18–79; response rate 86.5%) measured chronic disease prevalence and risk factors including blood pressure and cholesterol. A nutrition component was added to create the 1971–1974 National Health and Nutrition Examination Survey (NHANESI; $N = 28\,048$ noninstitutionalized persons age 1–74; the response rate was 96%; 78% were examined). NHANES-II was conducted from 1976 to 1980 for a sample of 27 801 noninstitutionalized persons aged 6 months to 74 years, 73.1% of whom were examined. NHANES-III was done over six years, 1988–1994, in two phases. In Phase I, 20 277 persons were sampled; 77% were examined. Results from Phase I show that certain year 2000 goals of the 1988 National Cholesterol Education Program were achieved by 1991 [40, 84]. Comparisons of cholesterol, hypertension, and smoking across the four surveys show that they declined from 1960 to 1990 for persons aged 65–74 [68]. Phase II of NHANES-III (1991–1994) involved sampling another 20 000 persons. Risk factors were followed with new emphasis on studying the natural history of chronic diseases. In addition, longitudinal follow-ups of health outcomes for 10+ years were done for the second and third surveys (e.g., National Health and Examination Follow-Up Survey 1971–1984; NHEFS-I [24]).

Combined Select Population and Registry Data

A fourth observation plan combines "select" and "registered" population data (e.g. [22]). Studies were done of mesothelioma and cancer risks in asbestos exposed occupational groups (*see* **Occupational Epidemiology**) as well as disease risks in populations with healthy life styles (e.g. Seventh Day Adventists [50] and Mormon High Priests in California [21]). NCI's Surveillance of Epidemiology and End Results (SEER) program of population-based tumor registries does data base of occupational studies of ionizing radiation [11] (*see* **Disease Registers**), and geographic monitoring of cancer mortality rates using vital statistics (*see* **Geographic Epidemiology**), put select population results into a national context.

For example, the absolute risk of lung cancer mortality among former smokers was examined in the American Cancer Society, Cancer Prevention Study II (CPS-II), a prospective cohort study of 1.2 million voluntary participants, begun in 1982, and analyzed after six years of follow-up. Roughly 900 000 persons were analyzed after excluding persons below age 40, above 80, or with smoking cessation before age 30, or after 75.

A "person-time" logistic regression model, where each subject had a separate entry for each year in the study, was used. A quadratic function of age provided a good fit to the risk of lung cancer death with gender, education, and smoking as covariates. **Spline** terms were added for each five years of age cohorts of former smokers. To fit risks for quit smokers, quadratic and cubic functions of time elapsed after cessation were used. **Goodness of fit** was tested using a statistic due to Hosmer & Lemeshow [36].

The study showed that smoking cessation produced benefits at all ages, e.g. the risk of smokers quitting in their early 60s was 45% of continuing smokers. The analyses did not explicitly use a **multistage model of carcinogenesis** [33] so the polynominals used to describe the health effects of quitting smoking are difficult to interpret biologically.

Innovations in Measurements of Chronic Disease Processes

Measurement of Disease Risk at the Cellular Level. Refinements in measurement were made possible by new **assays** [e.g. polymerase chain replication (PCR) and enzyme-linked immunosorbent assay (ELISA)] of nuclear and mitochrondrial DNA. This stimulated the development of **molecular epidemiology** where biomarkers of risk factor exposures are examined [76]. Biomarkers may be genetic damage caused by carcinogens forming

nuclear DNA adducts, or changes in oncogene or tumor suppressor gene (e.g. P-53) expression that can trigger, or block, tumor initiation. Other biomarkers are acquired traits affecting carcinogen metabolism (e.g. metabolizing enzyme production stimulated by exposure), or DNA damage and repair (e.g. DNA repair methyltransferase [82]). Assays of cellular DNA have also proved useful in determining the effects of viruses on cancer and other chronic diseases (e.g. the role of viruses, or immunological response to viruses, in CVD [64]).

In molecular epidemiology, biomarkers are used in **cross-sectional**, **retrospective**, prospective, or **case–control studies**. By analyzing biomarkers, disease processes can be identified early, thereby increasing the likelihood of developing preventative strategies and reducing the time necessary to identify the emergence of disease in exposed organisms.

In a study of 40 persons having smoked at least one pack per day for one year, blood samples were drawn while smoking, and at 2.5, 8 and 14 months after cessation. Smoking abstinence was monitored by measuring plasma cotinine. PAH–(polycyclic aromatic hydrocarbons, e.g. benzo [α] pyrene [32]) DNA adducts in white blood cells was twofold higher while smoking than after eight months of abstinence. A regression model of log transformed biomarkers, adjusted for background, suggested a half-life of 23.4 weeks for PAH–DNA adducts. The coefficient of variation for PAH–DNA adducts was 106%, suggesting a high degree of intraindividual variability in response to exposure. These and other data suggest a higher degree of female physiological lung cancer susceptibility to smoking products [63, 78, 100].

In a second study, lung tissue was taken from persons undergoing surgery for lung cancer ($n = 54$) or noncancer pulmonary disease ($n = 20$). Phase-I cytochrome P-450 enzymes responsible for metabolizing smoking chemical constituents into carcinogens were increased by the stimulus of smoking varying from 11-fold to 440-fold between individuals. Certain Phase-II cytochrome P-450 enzymes responsible for detoxifying activated carcinogens [e.g. glutathione S-transferase (GST)] were decreased and varied 17-fold over individuals. Thus, smoking increased cancer risks by differentially affecting Phase I and II enzymes in the cytochrome P-450 system [30] with large interindividual variation in activity [77]. Another study suggested that vitamins C and E significantly reduced PAH–DNA adducts *only* in persons with a null GSTM1 (a **genotype** which may be 50% prevalent [32]). Thus, the antitumor effects of micronutrients may strongly interact with the genotype.

These studies directly examine the biochemistry of physiological processes leading to disease initiation. This detects pathological changes early, and perhaps better identifies genetic risk. It has the difficulty that, by isolating specific disease components, producing a model of the natural history of the disease in an individual, or of a disease's effect on a population (especially given high variability in individual enzyme levels), is made difficult and makes greater demands on theory, e.g. using a parametric **Weibull** hazard for the multistage model of carginogesis [99].

Modeling Shared and Correlated Frailty. To study chronic disease risk in related individuals, one needs to control (i) the **correlation** of phenotypic traits, (ii) differences in phenotypic correlations between biologically related individuals (*see* **Familial Correlations**), (iii) **censoring**, (iv) the effects of observed **covariates**, and (v) **gene–environment interactions**. These problems can be examined with "shared" and "correlated" **frailty** models [99].

For frailty "shared" between related individuals (e.g. twins), the **hazard**, where u_i are covariate values (with effect parameters β) for the ith member of a group with shared genetic traits (e.g. for twins; $i = 1, 2$), x_i is the individual's age (the same for twins) and Z_i, a latent shared frailty variable, is,

$$\mu(x_i, Z_i, u_i) = Z_i \lambda_0(x_i) \exp(\beta u_i) \qquad (1)$$

The univariate form of (1) is identifiable for shared frailty variables with a finite mean (e.g. the **gamma** distribution). Parameters of univariate and bivariate shared frailty models are consistent only if the correlation of Z_i for related individuals is 1.0.

In correlated frailty models it is assumed that, conditional on Z_i the life span T_i is independent, with the risk μ_i for each related individual modeled as a proportional hazard, i.e. $\mu(Z_i, x_i) = Z_i \mu(x_i)$. The marginal survival distribution, when the Z_i are gamma distributed (with, for two related individuals $i = 1, 2$, the variances and correlation $\rho_{z_1 z_2}$), is

$$S(x_1, x_2) = \left[S_1(x_1)^{1-(\sigma_1/\sigma_2)\rho_{z_1 z_2}} \times S_2(x_2)^{1-(\sigma_2/\sigma_1)\rho_{z_1 z_2}}\right] \left[S_1(x_1)^{-\sigma_1^2}\right.$$

$$+ S_2(x_2)^{-\sigma_2^2} - 1\Big]^{-\rho_{z_1z_2}/\sigma_1\sigma_2}, \qquad (2)$$

where $S_i(x_i)$ are univariate survival functions, with the correlation of Z_i in the range,

$$0 \le \rho_{z_1z_2} \le \min\left(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right). \qquad (3)$$

For same-sex monozygotic (MZ) and dyzygotic (DZ) twins it is assumed that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and that $S_1(x) = S_2(x)$. If covariate data, u_i, are available for a correlated frailty model, then assessing the individual hazard in (1) requires the following bivariate survival function:

$$\begin{aligned} S(x_1, x_2|u_1, u_2) &= S_1(x_1|u_1)^{1-(\sigma_2/\sigma_1)\rho_{z_1z_2}} \\ &\times S_2(x_2|u_2)^{1-(\sigma_1/\sigma_2)\rho_{z_1z_2}} \Big(S_1(x_1|u_1)^{-\sigma_1^2} \\ &+ S_2(x_2|u_2)^{-\sigma_2^2} - 1\Big)^{-\rho_{z_1z_2}/\sigma_1\sigma_2}, \end{aligned} \qquad (4)$$

where S_i is related to the integrated hazard, H:

$$S_i(x|u_i) = [1 + \sigma_1^2 \exp(\beta u_1) H(x)]^{-1/\sigma_i^2}. \qquad (5)$$

Estimation of the model with covariates is complicated because Z_i are no longer gamma distributed conditional on u_i. The **EM algorithm** may be modified for this case [37].

Longitudinal Surveys Linked to Administrative Record Systems. A third innovation is longitudinal surveys of the health of elderly populations linked to administrative records (*see* **Administrative Databases**). The 1982, 1984, 1989, 1994 and 1999 National Long Term Care Surveys (NLTCS [52]) used computerized Medicare data in two ways. First, the NLTCS samples were drawn from those records. Thus, 100% of the sample can be tracked to the end of the study, or time of death. Secondly, although NLTCS response rates were near 95%, nonrespondents were, in part, more seriously ill [58] (*see* **Bias from Nonresponse**). Health information in the Medicare files can be used to (i) adjust for health biases in nonresponse, and (ii) reduce right censoring by providing continuous time information on health changes up to the time of death.

In molecular epidemiologic studies diseases were decomposed into select physiological processes for small, intensively evaluated populations. The NLTCS examines the integrated effects of those processes on the dynamics of multiple dimensions of function in elderly persons in large, nationally representative, longitudinally followed samples. Studies show that functional changes reflect both the effects of chronic diseases (e.g. the effects of Alzheimer's disease on function) and those of risk states in elderly persons for subsequent chronic disease and decline (e.g. physical activity and stroke [85]). Thus, by following multiple functional measures over time, latent state variables for individuals, and their change with age/time, can be modeled as a multivariate stochastic process (e.g. [59]).

The NLTCS design is outlined in Figure 1. The US population is assessed five times (1982, 1984, 1989, 1994, 1999). Interviews at each date were conducted over four months, during which age-related functional loss is assumed in equilibrium with age and functional status-specific mortality.

Linked to interviews are Medicare records for persons age 65+ for 1982–2000 containing (i) continuous histories of health service use, and (ii) birth and death dates. For each NLTCS, a sample of 5000 persons passing age 65 between interviews is drawn to keep the sample representative of the US elderly population and to maintain a size of roughly 20 000 persons. Over five waves, 42 000 individuals were followed with 22 000 deaths observed from 1982 to 2000. Since functional loss is low below age 65, left-censoring bias is small. Over time, persons are right censored by mortality, though health service data are available to the time of death.

To model this multidimensional, cross-temporally sparse data set requires assumptions about the trajectory of functional changes for persons between surveys (e.g. A and B), persons dying between surveys (e.g. C) or persons newly entering the sample (E). Given the multiple functions assessed, multivariate procedures are used to identify the underlying state variables. Assumptions made about functional dynamics [e.g. does a person's state "jump" when assessed (trajectory $A_1 \to A_2$) or change linearly (trajectory $B_1 \to B_2$) between assessments?] and their interaction with the observational plan (i.e. the overlay of survey assessment intervals on illustrative trajectories A, B, C or E), can affect estimates of process stochasticity (e.g. the different propagation of uncertainty; dashed and continuous lines), and trajectories of the average level of function and mortality, at late ages.

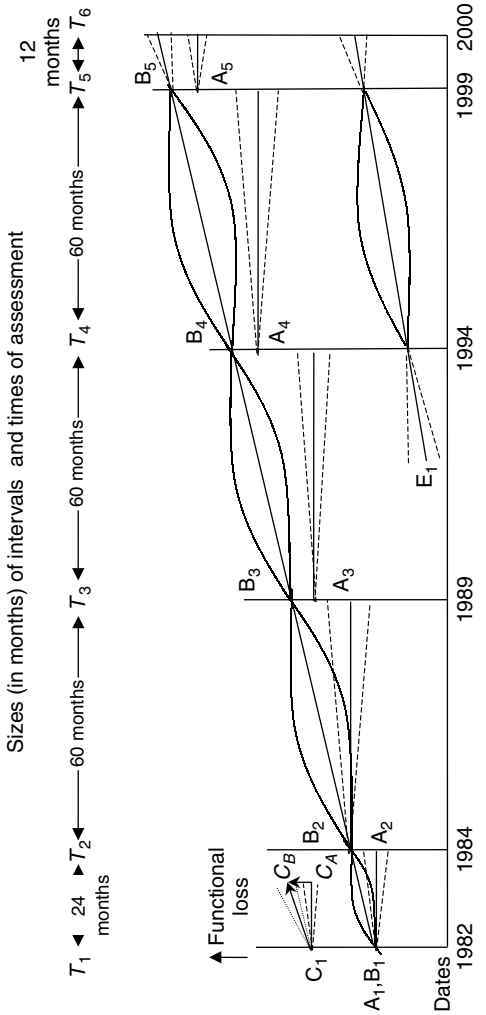

Figure 1 Temporal organization of the 1982, 1984, 1989, 1994 1999 NLTCS Longitudinal Observational Plan and illustrative trajectories of functional loss with time, and changes in the range of uncertainty about the level of functional loss relative to the time position between assessments

Hazard and Stochastic Process Models of Chronic Disease

In chronic disease models there are several types of time-related variation which can be exploited in different ways using ancillary data and theory.

Parametric Hazard Models Estimated from Failure Times

If only a few risk factors (e.g. age, sex, birth cohort/date, race, occupation, smoking) are measured, then a parametric function of age must be chosen to adjust for age increases in the physiological risk of disease (e.g. cancer) due to unobserved processes (*see* **Aging Models**).

Gompertz Hazard. A commonly used hazard is the Gompertz [29]:

$$\mu(a_{\text{it}}) = \alpha \exp(\theta a_{it}), \tag{6}$$

where θ represents the percent age increase in mortality risk μ, per unit time and a_{it} is person's i age at time t. This model describes the age dependence of mortality in many human [94] and animal [27] adult populations. Thus, θ has been interpreted as the intrinsic rate of aging (e.g. [80]). Risk factors may be represented by stratifying (6) into J discrete risk "groups" with scale parameters α_j. A stratified Gompertz, with the population in each strata known, describes the population age trajectory of disease risk as a weighted mixture of J Gompertz functions.

If membership in risk factor strata is not known, but the risk factor can be assumed continuously distributed according to a parametric (e.g. a gamma or **inverse Gaussian**) form with special properties relating to parameter changes under systematic mortality selection, a continuously mixed Gompertz hazard can be estimated by inferring the parameters of the mixing distribution from the deviation of observed and Gompertz predicted mortality rates (see below; Manton et al. [57]). The advantages of (6) are that the doubling time of risk is constant over age, and that it is an **extreme value** distribution, i.e. the progression of individual failure processes generates failure times with the same distribution in the population.

Weibull Hazard. The **Weibull** hazard [93] is used to analyze cancer risk [4],

$$\mu(a_{it}) = \beta(a_{it})^{m-1}, \tag{7}$$

where m is the number of genetic changes in a cell before a tumor initiates. In cancer, the changes are mutations in nuclear DNA that allow growth control to be lost. This function is consistent with observations of the age dependence of cancer mortality rates in populations (e.g. stomach cancer in England and Wales [87]). As carcinogenesis is studied using the techniques of molecular epidemiology, the biological validity of (7) as a model of cancer has been confirmed. For example, mutations in the p53 and p21 genes regulating cell growth and division – and possibly apoptosis in cells with DNA mutations – cause many cancers. Fearon & Vogelstein [23] found five mutations necessary for colon cancer initiation.

The Weibull is also an extreme value failure distribution where the doubling time for risk is *not* constant over age, but the probability of each mutation is. Thus, the mechanism of failure in the individual involves cells in a given tissue gradually accumulating DNA errors until the mth error occurs and a tumor initiates. This "multistage" model of carcinogenesis [4] requires that mutations be described by probabilities, i.e. each is sufficiently rare that the proportion of cells experiencing a mutation does not affect the risk of subsequent mutations. The time to the occurrence of the first cell, of N cells in a tissue, achieving m errors has the same distribution as the time to tumor initiation in a population.

Mixed Hazards. The Weibull does not fit increases in cancer risks after, say, age 75 [12]. This may be due to specific mutations occurring too rapidly to be modeled as a probability [62]. Modeling each by a hazard rate produces an overall hazard function (based on a series expansion) with additional terms slowing the age increase in risk as late events in the failure process become constrained by high transition rates for events earlier in the process. Such models produce estimates of m too large (e.g. 20–25) to be biologically plausible.

An alternative explanation for the slowing of the age increase of the Weibull is that the hazard applies to individuals. Unobserved risk factors are assumed to affect the probability of mutations, so individuals have different cancer risks. Thus, unobserved exogenous factors do not affect the number, m, of mutations needed for tumor initiation – only the m probabilities – whose effect is summarized by the individual's scale parameter, β_i. The effects of unobserved variables are represented by assuming that β_i

is distributed over individuals, according to a mixing distribution whose parameters have special properties under mortality selection. Mixed hazard functions can be examined using [17].

$$\overline{\mu}(a, \beta_i, \gamma, m) = \frac{\overline{\beta} a^{m-1}}{\left(1 + n\gamma \int_0^a \overline{\beta} u^{m-1}\, \mathrm{d}u\right)^{1/n}}, \quad (8)$$

where $\gamma = \mathrm{var}(\beta_i)/\mathrm{E}^2(\beta_i)$ is the squared coefficient of variation (CV) of the β_i, $\overline{\beta}$ is the mean of the distribution of β_i, and n determines the mixing distribution. For general n, the cumulative Weibull hazard is

$$H^{(n)}(a) = \int_0^a \frac{\overline{\beta} u^{m-1}}{\left(1 + (n\gamma\overline{\beta} u^m)/2\right)^{1/n}}\, \mathrm{d}u. \quad (9)$$

For $n = 1$ (the gamma distribution) this is

$$H^{(1)}(a) = \frac{1}{\gamma} \ln\left(1 + \frac{\gamma\overline{\beta} a^m}{m}\right) \quad (10)$$

and, for $n = 2$ (the inverse Gaussian distribution):

$$H^{(2)}(a) = \frac{1}{\gamma}\left[(1 + 2\gamma\overline{\beta} a^m)^{1/2} - 1\right]. \quad (11)$$

The gamma distribution implies a constant CV for β_i; the inverse Gaussian a decreasing CV. This model can be generalized to allow mixing across K observed risk factor strata, even if only the marginal risk factor distribution is observed [53]. The Gompertz is similarly generalized.

Semi-Parametric Hazard Models

For multiple risk factors measured over time, a more general model (*see* **Cox Regression Model**) is [13]

$$\mu(\mathbf{x}_i; t) = \lambda(t) \cdot \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}), \quad (12)$$

where $\lambda(t)$, the baseline hazard, is an unknown function of time, $\mathbf{x}_i$ is a J-element vector of risk factors, and $\boldsymbol{\beta}$ is a vector of coefficients. To make (12) depend on age it can be included in $\mathbf{x}_i$, and assuming $\lambda(t)$ is constant, λ_0, producing

$$\mu(\mathbf{x}_{it}; a_{it}) = \lambda_0 \exp(\theta a_{it} + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}), \quad (13)$$

where a_{it} is current age. Eq. (13) implicitly represents multiplicative interactions between risk factors because effects are additive in the logs of risk factors. Specifically, the second-order partial derivatives of (13) are [75]

$$\frac{\partial \mu(x_{it}, a_{it})}{\partial x_{ijt} \partial x_{ikt}} = \lambda_0 \beta_j \beta_k \exp(\theta a_{it}) \exp(\mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}). \quad (14)$$

The rate of change of $\mu(x_{it}, a_{it})$ in (14), relative to values of x_{ijt} and x_{ikt}, is a function of a proportionality factor, $\lambda_0 \beta_j \beta_k$, age, $\exp(\theta a_{it})$, and risk factors, $\exp(\mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta})$. Thus, age and risk factor values interact so that comparisons of risk factor effects across populations, where different risk factors are measured, or moments of the risk factor distribution differ, is difficult. A decomposition of (12) by cause has the problem that summing such functions produces a different distribution for total mortality (i.e. the sum of exponentials).

Combining Risk Factor Dynamics and Mortality

If the vector, $\mathbf{x}_{it}$ represents the physiological state of person i at t, then a model is needed to describe the joint evolution of $\mathbf{x}_{it}$, and disease risk. Generating a model requires assumptions about the temporal information in $\mathbf{x}_{it}$. Longitudinal studies provide only partial information on the individual's state, i.e. $\mathbf{x}_{it}$ is incomplete. For a model using incomplete data to have external and internal validity it needs to be:

1. consistent with biological theory,
2. decomposable by modes of failure, and
3. logically consistent in portraying risk factor dynamics and their interaction with mortality.

Many models do not have one or more of these properties. The exponential form of the Cox model has effects that depend on risk factor levels and is not decomposable by failure modes. Models appropriate to describe such data developed from modeling human aging as a multidimensional, stochastic process (e.g. [81] and [88]). Early models did not represent heterogeneity in individual risks. A model due to Woodbury & Manton [95] has the properties associated with conditional Gaussian distributions used to describe stochastic processes [97]. It describes chronic disease by two processes. One describes changes in $\mathbf{x}_{it}$ as J stochastic functions of their past values,

$$\mathbf{x}_{it} = \mathbf{u}_0 + \mathbf{u} a_{ijt-1} + R\mathbf{x}_{ijt-1} + \mathbf{e}_{it}, \quad (15)$$

where $\mathbf{x}_i$ are functions, possibly time dependent [i.e. R may be $R(t)$], of the prior state of the individual ($\mathbf{x}_{it-1}$), age (a_{ijt-1}), and stochastic errors $\mathbf{e}_{it-1}$, generated by "diffusion". If the $\mathbf{e}_{it-1}$ values are independent of the $\mathbf{x}_{it}$, then the diffusion process might be Gaussian (*see* **Brownian Motion and Diffusion Processes**). If the magnitude of $\mathbf{e}_{it-1}$ depends on values in $\mathbf{x}_{it}$, then the diffusion process is more complicated. Which coefficients can be estimated for (15) depends on the heterogeneity represented in the data.

The second process describes mortality as a quadratic function of risk factor values, i.e.

$$\mu(\mathbf{x}_{it}; a_{it}) = \left(\mu_0 + \mathbf{x}_{it}^{\mathrm{T}}\mathbf{b} + \tfrac{1}{2}\mathbf{x}_{it}^{\mathrm{T}}\mathbf{B}\mathbf{x}_{it}\right)\exp(\theta a_{it}), \tag{16}$$

where μ_0 is the constant force of mortality, $\mathbf{b}$ are linear coefficients adjusting for changes in the location of the hazard in the state space, $\mathbf{B}$ is a matrix of coefficients representing the quadratic and second-order interactions of the $\mathbf{x}_{it}$, and $\exp(\theta a_{it})$ represents the average effects of age related unobserved variables. Each coefficient in (16) can be made a function of age by multiplying it by $\exp(\theta a_{it})$. To relate (16) to (12) we add quadratic terms to (13) to produce, using a Taylor series expansion,

$$\mu(\mathbf{x}_{it}; a_{it}) = \lambda_0\left[1 + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \tfrac{1}{2}\mathbf{x}_{it}^{\mathrm{T}}(\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}})\mathbf{x}_{it} + O\left(\mathbf{x}_{it}^3\right)\right]\exp(\theta a_{it}), \tag{17a}$$

which can be expressed as

$$\mu(\mathbf{x}_{it}; a_{it}) = \left[\mu_0 + \mathbf{x}_{it}^{\mathrm{T}}\mathbf{b} + \tfrac{1}{2}\mathbf{x}_{it}^{\mathrm{T}}\mathbf{B}\mathbf{x}_{it} + O\left(\mathbf{x}_{it}^3\right)\right] \times \exp(\theta a_{it}). \tag{17b}$$

The difference between (17b) and (16) is the higher order interaction terms $O(\mathbf{x}_{it}^3)$. If these terms are negligible, then (17b) reduces to (16). The function in (16) has simpler second-order partial derivatives [75]:

$$\frac{\partial^2 \mu(\mathbf{x}_{it}; a_{it})}{\partial x_{ijt}\partial x_{ikt}} = \lambda_0\beta_j\beta_k\exp(\theta a_{it}), \tag{18}$$

which can be written

$$= \beta_{jk}\exp(\theta a_{it}).$$

Thus, quadratic hazard age and risk factor effects do not change over risk factor levels. The hazard in (16) has a number of useful properties. First, for L causes of death with the same θ (i.e. the same unobserved, age-related variables affect each cause) the quadratic forms are additive. Thus, total mortality can be consistently decomposed. Secondly, the failure mechanism represented by (16) is natural for multidimensional, biological systems where homeostatic forces keep trajectories of $\mathbf{x}_{it}$ close to a "central" region of the state space where mortality is a minimum. This is often empirically justifiable for chronic diseases (e.g. blood pressure at late ages). It may also be justified from assumptions about the dynamic response of complex biological systems to stress. Specifically, the principle of "hormesis" suggests that homeostatic (or homeorhetic) feedback mechanisms cause complex organisms to "over-control" in response to low-level environmental stress because there are time delays in responding to environmental insults due to the need for communication (e.g. by hormonal responses) between organ systems [86]. Thus, given a lag in responding, adjustments to environmental stress must be in "quanta" which, on average, overcompensate in responding to stress because an absolutely continuously graded response exactly matching the stress is not possible given the system's latency. Thus, low-level stresses "strengthen" the organism, i.e. the minimum mortality level occurs for a small but nonzero exposure. "Quanta" responses are common in humans. Without low-level exposure to pathogens, the immune system will not develop humoral responses. Alternately, the cytochrome P-450 enzyme system, although genetically programmed, may not produce a detoxifying enzyme except when stimulated by chemical stress. Thus, the quadratic hazard is the inverse of a biologically complex organism's fitness "response" to environmental stress.

Not only is the hormetic "dose–response" function reflective of a peak in an organism's "fitness" at a nonzero stress level [10], but the variability of responses is increased by structural heterogeneity in populations of responding organisms. This may be why "aged" organisms express "chronic" diseases. Models of genetic selection suggest that organisms evolve to use energy to maximize survival only to the end of the reproductive life span. This does not explain why organisms, like humans, survive long past the end of the reproductive life span and suffer chronic, degenerative diseases. Hormesis

suggests that this is a necessary response to environmental stresses, i.e. chronic diseases may be "over" responses to environmental insults. Neoplasia are tissue healing responses out of genetic (p53) control; similar arguments might be made for atheromas. Neurological diseases may be due to oxidative processes that have a function in meeting environmental stresses (e.g. the "glutamic" cascade of cell death in stroke) or in producing immunological responses.

Finally, the chronic diseases now recognized may not represent the physiological mechanisms that generate a particular pathology. Angiogenesis (i.e. the creation of new vasculature) may allow cancers to metastasize but may also permit the development of collateral circulation in a heart damaged by atherosclerosis. Many tumors develop because of the failure of the p53 gene to induce apoptosis due to mutation. Thus, is the "cancer" associated with the dysfunction of a specific organ the chronic disease, or mutations in the p53 gene that occur in many tissue types? Not only are there (i) intrinsic features of dynamic feedback systems in human organisms that dictate that the hazard is a U- or J-shaped function of exposure to environmental stress, but also (ii) a crucial problem in modeling chronic diseases is to redefine chronic disease as the understanding of disease dynamics on a molecular level improves.

Age changes in quadratic hazards with different θs are illustrated in Figure 2. Linear terms (b) describe the location of the minimum, x_{jt}^0, of the hazard relative to risk factors – a minimum that may be a function of age [i.e. $\mathbf{b}(t) = \mathbf{b}\exp(\theta a_{it})$]. The matrix of quadratic coefficients, $\mathbf{B}$, represents the age-dependent $\mathbf{B}(t) = \mathbf{B}\exp(\theta a_{it})$ curvature of the hazard relative to risk factor levels.

The **likelihood** for estimating the parameters of the two processes from a longitudinal study, where individual risk factor values are assessed at fixed time intervals, and the time of occurrence of specific lethal (or morbid) events are observed, can be written as [55] (i) the initial distribution, $\mathbf{x}_{i0}$, (ii) the regression (15) of $\mathbf{x}_{it}$ on past values in persons surviving a study interval, and (iii) the quadratic mortality function (16).

With estimates of (15) and (16) one can model the trajectory of mortality and risk factors in a cohort by using those parameters in systems of differential equations (e.g. [56]). Those differential equations describe not only the probability of survival to age a

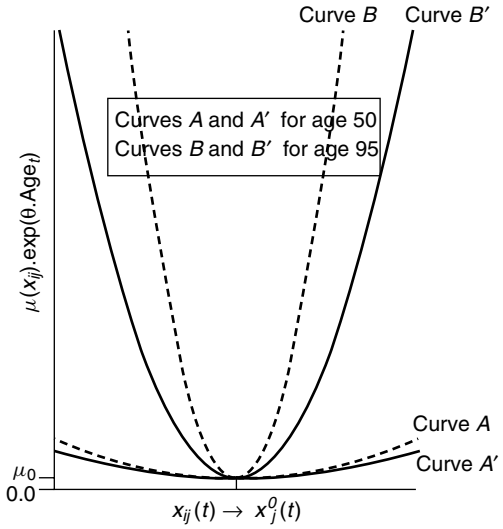

Figure 2 Changes in shape of the age-specific quadratic hazard functions at ages 50 and 95 due to the effects of unobserved variables represented by two values of the parameter θ describing the rate of aging (7.70 and 3.85%)

(conditional on changes in risk factors to a) but also the distribution of the $\mathbf{x}_{it}$ among survivors to age a.

A crucial model feature is how diffusion and risk factor heterogeneity interact with mortality over time. If the difference in the form of the partial differential equations is assumed to be updated (i) monthly or (ii) annually, then one obtains different survival patterns, or risk factor trajectories, at late age (Table 1).

In Table 1 three conditions are presented. The first assumes that, between assessments, the $\mathbf{x}_{it}$ are constant (i.e. person A in Figure 1). This produces, at age 95, a life expectancy of 5 years with an average of 46.2% of physical function preserved. Secondly, the $\mathbf{x}_{it}$ are assumed to change linearly between assessments (i.e. person B in Figure 1). Life expectancy at 95 increases to 5.4 years. The proportion of function maintained increases to 52.3%. Thirdly, state variables can be assumed to change linearly between assessments, but with mortality interacting with state variable changes every 12 months (instead of monthly). Life expectancy at age 95 is only 4.4 years and the proportion of function maintained is only 33.2%. Thus, conditions 1 and 2 represent different assumptions about $\mathbf{x}_{it}$ trajectories between assessments. Conditions 2 and 3 contrast the effect of the

Table 1 Life table models, constructed under three assumptions about state dynamics estimated from time varying covariates for females in the US NLTCS followed 1982 to 1991

	Condition 1. "Jump"processes with monthly mortality state variable interactions		Condition 2. Linear state variable dynamics with monthly mortality state variable interactions		Condition 3. Linear state variables dynamics with annual mortality state variable interaction	
Age (in years)	Life expectancy at age *a*	Average proportion of 27 functions maintained at age *a*	Life expectancy at age *a*	Average proportion of 27 functions maintained at age *a*	Life expectancy at age *a*	Average proportion of 27 functions maintained at age *a*
65	20.9	0.923	20.8	0.926	20.1	0.917
75	14.1	0.838	14.1	0.832	13.3	0.799
85	8.5	0.643	8.7	0.647	7.6	0.544
95	5.0	0.462	5.4	0.523	4.4	0.332
105	3.8	0.473	4.0	0.573	3.3	0.375

frequency at which mortality and dynamic processes interact in a model. In all three conditions the proportion of function maintained increases at late ages because, for elderly, impaired persons mortality rates eventually surpass the age rate of functional loss. Thus, the interactions of heterogeneity and mortality, and the dynamic balance of autoregression and diffusion, control the age trajectory of mortality and chronic disease processes at late ages. Thus, the model produced complex, nonlinear risk factor and mortality population trajectories at late ages. Multivariate stochastic process models are required to examine nonlinearities in risk factor dynamics and mortality at late ages.

Discussion

We have examined models of chronic disease. Techniques appropriate to modeling chronic disease change as information increases. At the least informed level we search for patterns of association between risk factors and disease risk. As information increases, measures of association may not fully exploit the available information. Thus, different models are needed for the different amounts of information generated by different observational plans.

One strategy uses improved measurement techniques to focus on increasingly detailed features of the chronic disease processes. Instead of modeling the relationship between cholesterol and CHD risk, one has to recognize that there are (i) different lipoproteins, some beneficial (e.g. HDL) and some not (LDL), and (ii) that there are multiple stages in atherogenesis, such as plaque initiation, plaque elaboration (involving lipid levels, oxidation of LDL, recruitment of macrophages having ingested oxidized LDL into plaques as "foam" cells; stimulation of arterial endothelial growth due to inflammatory responses and stimulation of local growth factor production [9]), plaque disruption, and thrombus formation. Each stage involves different risk factors and risk factor interactions. As measurement of the process improves, static models may describe process components operating over short time scales.

However, as longitudinal information accumulates, models that integrate stages of the disease process are needed to represent (i) the evolution of a multistage stochastic process, and (ii) the effects of those processes on the population distribution of outcomes. In these models the temporal organization of the stages of the chronic disease process is crucial, especially in developing interventions. Thus, the selection of models to analyze chronic disease has to be based on assessments of both the information generated by observational plans and the nature of the processes modeled.

Acknowledgment

This research was supported by grants from the National Institute on Aging.

References

[1] Akushevich, I., Akushevich, L., Manton, K., Yashin, A. (2002). Stochastic process model of mortality and aging: application to longitudinal data. Working paper, Center for Demographic Studies.

[2] Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[3] Anderson, T.J., Meredith, I.T., Yeung, A.C., Frei, B., Selwyn, A.P. & Ganz, P. (1995). The effect of cholesterol-lowering and antioxidant therapy on endothelium-dependent coronary vasomotion, *New England Journal of Medicine* **332**, 488–493.

[4] Armitage, P. & Doll, R. (1961). Stochastic models for carcinogenesis, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 19–38.

[5] Bailey, N.T. (1975). *The Mathematical Theory of Infectious Diseases*. Hafner, New York.

[6] Benzuly, K.H., Padgett, R.C., Kaul, S., Piegors, D.J., Armstrong, M.L. & Heistad, D.D. (1994). Functional improvement precedes structural regression of atherosclerosis, *Circulation* **89**, 1810–1818.

[7] Brattstrom, L., Israelsson, B., Norring, B., Bergquist, V., Thorne, J., Hultberg, B. & Hamfelt, A. (1990). Impaired homocysteine metabolism in early onset cerebral and peripheral occlusive arterial disease, *Atherosclerosis* **81**, 51–60.

[8] Brown, B.G., Zhao, X.Q., Sacco, D.E. & Albers, J.J. (1993). Lipid lowering and plaque disruption and clinical events in coronary disease, *Circulation* **87**, 1781–1791.

[9] Buja, L.M. & Willerson, J.T. (1994). Role of inflammation in coronary plaque disruption, *Circulation* **89**, 503–505.

[10] Calow, P. (1982). Homeostasis and fitness, *American Naturalist* **120**, 416–419.

[11] CEDR. (2002). Comprehensive Epidemiologic Data Resource. `http://cedr.lbl.gov`.

[12] Cook, N.R., Fellingham, S.A. & Doll, R. (1969). A mathematical model for the age distribution of cancer in man, *International Journal of Cancer* **4**, 93–112.

[13] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–202.

[14] Cupples, L.A., D'Agostino, R.B., Anderson, K. & Kannel, W.B. (1988). Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study, *Statistics in Medicine* **7**, 205–218.

[15] Dahlen, G.H., Guyton, J.R., Attar, M., Farmer, J.A., Kautz, J.A. & Gotto, A.M. (1986). Association of levels of lipoprotein Lp(a), plasma liquids, and other lipoproteins with coronary artery disease documented by angiography, *Circulation* **74**, 758–765.

[16] Dawber, T.R. (1980). *The Framingham Study: The Epidemiology of Arterosclerotic Disease*. Harvard University Press, Cambridge, Mass.

[17] Dubey, S.D. (1967). Some percentile estimators of Weibull parameters, *Technometrics* **9**, 119–129.

[18] Dubos, R. (1965). *Man Adapting*. Yale University Press, New Haven and London.

[19] Early Breast Cancer Trialists Collaborative Group (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy – Part I, *Lancet* **339**, 1–15.

[20] Early Breast Cancer Trialists Collaborative Group (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy – Part II, *Lancet* **339**, 71–85.

[21] Enstrom, J.E. (1989). Health practices and cancer mortality among active California Mormons, *Journal of the National Cancer Institute* **81**, 1807–1814.

[22] Enstrom, J.E. & Kanim, L.E. (1983). Populations at low risk, in G.R. Nowell, ed. *Cancer Prevention in Clinical Medicine*. Ravin, New York, pp. 49–78.

[23] Fearon, E.R. & Vogelstein, B. (1990). A genetic model for colo-rectal tumorigenesis, *Cell* **61**, 759–767.

[24] Feldman, J.J., Makuc, D.M., Kleinman, J.C. & Huntley, J.C. (1989). National trends in educational differentials in mortality, *American Journal of Epidemiology* **129**, 919–933.

[25] Feskens, E., Bowles, C. & Kromhout, D. (1991). *Journal of Clinical Epidemiology* **44**, 947–953.

[26] Feskens, E., Bowles, C. & Kromhout, D. (1992). Intra- and interindividual variability of glucose tolerance in an elderly population, *Journal of Clinical Epidemiology* **45**, 293–300.

[27] Finch, C.E. & Pike, M.C. (1996). Maximum life span predictions from the Gompertz mortality model, *Journal of Gerontology* **51**, B183–B194.

[28] Forman, D. (1991). *Helicobacter pylori* infection: a novel risk factor in the etiology of gastric cancer, *Journal of the National Cancer Institute* **83**, 1702–1703.

[29] Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, *Philosophical Transactions of the Royal Society of London* **114**, 513.

[30] Gonzalez, F.J. & Nebert, D.W. (1990). Evolution of the P450 gene superfamily: animal - plant "welfare" molecular drive, and human genetic differences in drug oxidation, *Trends in Genetics* **6**, 182–186.

[31] Grayston, J.T. (1993). Chlamydia in atherosclerosis, *Circulation* **87**, 1408–1409.

[32] Grinberg-Funes, R.A., Singh, V.N., Perera, F.P., Bell, D.A., Young, T.L., Dickey, C., Wang, L.W. & Santella, R.M. (1994). Polycyclic aromatic hydrocarbon – DNA adducts in smokers and their relationship to micronutrient levels and the glutathione-S-transferase M1 genotype, *Carcinogenesis* **15**, 2449–2454.

[33] Halpern, M.T., Gillespie, B.W. & Warner, K.E. (1993). Patterns of absolute risk of lung cancer mortality in former smokers, *Journal of the National Cancer Institute* **85**, 457–464.

[34] Hedges, L. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York.

[35] Hosking, S.W., Ling, T.K.W., Chung, S.C.S., Yung, M.Y., Cheng, A.F.B., Sung, J.J.Y. & Li, A.K.C. (1994). Duodenal ulcer healing by eradication of helicobacter pylori without anti-acid treatment: randomised controlled trial, *Lancet* **343**, 508–510.

[36] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[37] Iachine, I.A. (1995). Parameter estimation in the bivariate correlated frailty model with observed covariates via EM-algorithm. *Research Report of Population Studies of Aging*, No. 16. Odense University, Denmark, pp. 1–21.

[38] Jacobson, D.R., Pastore, R.D., Yaghoubian, R., Kane, I., Gallo, G., Buck, F.S. & Buxbaum, J.N. (1997). Variant-sequence transthyretin (isoleucine 122) in late-onset cardiac amyloidosis in black Americans, *New England Journal of Medicine* **336**, 466–473.

[39] Jacquez, J.A. (1972). *Compartmental Analysis in Biology and Medicine*. Elsevier, Amsterdam.

[40] Johnson, C.L., Rifkind, B.M., Sempos, C.T., Carroll, M.D., Bachorik, P.S., Briefel, R.R., Gordon, D.J., Burt, V.L., Brown, C.D., Lippel, K. & Cleeman, J.I. (1993). Declining serum total cholesterol levels among US adults: the National Health and Nutrition Examination Surveys, *Journal of the American Medical Association* **269**, 3002–3008.

[41] Kannel, W.B., Castelli, W.P., Gordon, T. & McNamara, P.M. (1971). Serum cholesterol, lipoproteins, and the risk of coronary heart disease. The Framingham study, *Annals of Internal Medicine* **74**, 1–12.

[42] Kaplan, G. & Keil, J. (1993). Socioeconomic factors and cardiovascular disease: a review of the literature, *Circulation* **88**, 1973–1998.

[43] Kasch, F.W., Boyer, J.L., Van Camp, S.P., Verity, L.S. & Wallace, J.P. (1993). Effect of exercise on cardiovascular ageing, *Age and Ageing* **22**, 5–10.

[44] Keil, J., Sutherland, S., Knapp, R., Lackland, D., Gazes, P. & Tyroler, H. (1993). Mortality rates and risk factors for coronary disease in black as compared

with white men and women, *New England Journal of Medicine* **329**, 73–78.

[45] Keys, A. (1980). Wine, garlic, and CHD in seven countries, *Lancet* **1**, 145–146.

[46] Kitzman, D.W. & Edwards, W.D. (1990). Minireview: age-related changes in the anatomy of the normal human heart, *Journal of Gerontology: Medical Sciences* **45**, M33–M39.

[47] Lakatta, E.G. (1985). Health, disease, and cardiovascular aging, in *America's Aging: Health in an Older Society*. National Academy Press, Washington, pp. 73–104.

[48] Law, M.R., Wald, N.J. & Thompson, S.G. (1994). By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease?, *British Medical Journal* **308**, 367–373.

[49] Law, M.R., Wald, N.J., Wu, T., Hackshaw, A. & Bailey, A. (1994). Systematic underestimation of association between serum cholesterol concentration and ischaemic heart disease in observational studies: data from the BUPA study, *British Medical Journal* **308**, 363–366.

[50] Lindsted, K.D., Tonstad, S. & Kuzma, J.W. (1991). Self-report of physical activity and patterns of mortality in Seventh-Day Adventist men, *Journal of Clinical Epidemiology* **44**, 355–364.

[51] Little, R.T.J. & Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika* **72**, 497–512.

[52] Manton, K.G., Corder, L. & Stallard, E. (1997). Chronic disability trends in the U.S. elderly population 1982 to 1994, *Proceedings of the National Academy of Sciences* **94**, 2593–2598.

[53] Manton, K.G., Lowrimore, G. & Yashin, A.I. (1993). Methods for combining ancillary data in stochastic compartment models of cancer mortality; generalization of heterogeneity models, *Mathematical Population Studies* **4**, 133–147.

[54] Manton, K.G. & Stallard, E. (1988). *Chronic Disease Modeling: Measurement and Evaluation of the Risks of Chronic Disease Processes*. Griffin, London.

[55] Manton, K.G. & Stallard, E. (1992). Compartment model of the temporal variation of population lung cancer risks, in *Biomedical Modeling and Simulation*, J. Eisenfled, D.S. Levine & M. Witten, eds. Elsevier/North-Holland, Amsterdam, pp. 75–81.

[56] Manton, K.G., Stallard, E. & Singer, B.H. (1994). Methods for projecting the future size and health status of the U.S. elderly population, in *Studies of the Economics of Aging*, D. Wise, ed. University of Chicago Press, Chicago, pp. 41–77.

[57] Manton, K.G., Stallard, E. & Vaupel, J.W. (1986). Alternative models for the heterogeneity of mortality risks among the aged, *Journal of the American Statistical Association* **81**, 635–644.

[58] Manton, K.G., Stallard, E. & Woodbury, M.A. (1991). A multivariate event history model based upon fuzzy states: estimation from longitudinal surveys with informative nonresponse, *Journal of Official Statistics* **7**, 261–293.

[59] Manton, K.G., Stallard, E., Woodbury, M.A. & Dowd, J.E. (1994). Time-varying covariates in models of human mortality and aging: multidimensional generalization of the Gompertz, *Journal of Gerontology: Biological Sciences* **49**, B169–B190.

[60] Manton, K.G. & Vaupel, J.W. (1995). Survival after the age of 80 in the United States, Sweden, France, England, and Japan, *New England Journal of Medicine* **333**, 1232–1235.

[61] Matis, J.H. & Wehrly, T.E. (1979). Stochastic models of compartmental systems, *Biometrics* **35**, 199–220.

[62] Moolgavkar, S.H. (1978). The multi-stage theory of carcinogenesis and the age distribution of cancer in man, *Journal of the National Cancer Institute* **61**, 49–52.

[63] Mooney, L.A., Santella, R.M., Covey, L., Jeffery, A.M., Bigbee, W. & Randall, M.C. (1995). Decline in DNA damage and other biomarkers in peripheral blood following smoking cessation, *Cancer Epidemiology Biomarkers Preventions* **4**, 627–634.

[64] Mozar, H.N., Bal, D.G. & Farag, S.A. (1990). The natural history of atherosclerosis: an ecologic perspective, *Atherosclerosis* **82**, 157–164.

[65] Multiple Risk Factor Intervention Trial Research Group (MRFIT) (1990). Mortality rates after 10.5 years for participants in the multiple risk factor intervention trial, *Journal of the American Medical Association* **263**, 1795–1801.

[66] Myers, G.C. (1981). Future age projections and society, in *Aging: A Challenge to Science and Society*, Vol. 2, Part II, W.M. Beattie, Jr, J. Piotrowski & M. Marois, eds. Oxford University Press, Oxford, pp. 248–260.

[67] National Center for Health Statistics (1964). *The Change in Mortality Trends in the United States*, Series 3, No. 1. Public Health Service, Washington.

[68] National Center for Health Statistics (1995). *Health, United States, 1994*. Public Health Service, Hyattsville.

[69] Nightingale, T.E. & Gruber, J. (1994). Helicobacter and human cancer, *Journal of the National Cancer Institute* **86**, 1505–1509.

[70] Omran, A.R. (1971). The epidemiologic transition: a theory of the epidemiology of population change, *Milbank Memorial Quarterly* **49**, 509–538.

[71] Orchard, G. & Woodbury, M.A. (1971). A missing information principle: theory and application, *Sixth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 697–715.

[72] Ornish, D., Brown, S.E., Scherwitz, L.W., Billings, J.H., Armstrong, W.T., Ports, T.A., McLanahan, S.M., Kirkeeide, R.L., Brand, R.J. & Gould, K.L. (1990). Can lifestyle changes reverse coronary heart disease? The Lifestyle Heart Trial, *Lancet* **336**, 129–133.

[73] Parsonnet, J. (1996). Helicobacter pylori in the stomach – a paradox unmasked, *New England Journal of Medicine* **335**, 278–280.

[74] Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Research Group (1990). Relationship of atherosclerosis in young men to serum lipoprotein cholesterol concentrations and smoking, *Journal of the American Medical Association* **264**, 3018–3024.

[75] Pekkanen, J., Manton., K.G., Stallard, E., Nissinen, A. & Karvonen, M.J. (1992). Risk factor dynamics, mortality and life expectancy differences between Eastern and Western Finland: the Finnish cohorts of the Seven Countries Studies, *International Journal of Epidemiology* **21**, 406–419.

[76] Perera, F. (1996). Insights into cancer susceptibility, risk assessment, and prevention, *Journal of the National Cancer Institute* **88**, 496–509.

[77] Petruzzelli, S., Camus, A.M., Carrozzi, L., Ghelarducci, L., Rindi, M. & Menconi, G. (1988). Long-lasting effects of tobacco smoking on pulmonary drug-metabolizing enzymes: a case - control study on lung cancer patients, *Cancer Research* **48**, 4695–4700.

[78] Risch, H.A., Howe, G.R., Jain, M., Burch, J.D., Holowaty, E.J. & Miller, A.B. (1993). Are female smokers at higher risk for lung cancer than male smokers? A case - control analysis by histologic type, *American Journal of Epidemiology* **138**, 281–293.

[79] Rudemo, M. (1973). State estimation for partially observed Markov chains, *Journal of Mathematical Analysis and Applications* **44**, 581–611.

[80] Sacher, G.A. (1977). *Life Table Modification and Life Prolongation*, J. Birren & C. Finch, eds. Van Nostrand Reinhold, New York.

[81] Sacher, G.A. & Trucco, E. (1962). The stochastic theory of mortality, *Annals of the New York Academy of Sciences* **96**, 985.

[82] Sekiguchi, M., Nakabeppu, Y., Sakumi, K. & Tuzuki, T. (1996). DNA-repair methyltransferase as a molecular device for preventing mutation and cancer, *Journal of Cancer Research and Clinical Oncology* **122**, 199–200.

[83] Selby, J.V., Austin, M.A., Sandholzer, C., Quesenberry, C.P., Zhang, D., Mayer, E. & Utermann, G. (1994). Environmental and behavioral influences on plasma lipoprotein(a) concentration in women twins, *Preventive Medicine* **23**, 345–353.

[84] Sempos, C.T., Cleeman, J.I., Carroll, M.D., Johnson, C.L., Bachorik, P.S., Gordon, D.J., Burt, V.L., Briefel, R.R., Brown, C.D., Lippel, K. & Rifkind, B.M. (1993). Prevalence of high blood cholesterol among US adults: an update based on guidelines from the second report of the national cholesterol education program adult treatment panel, *Journal of the American Medical Association* **269**, 3009–3014.

[85] Shinton, R. & Sagar, G. (1993). Lifelong exercise and stroke, *British Medical Journal* **307**, 231–234.

[86] Stebbing, A.R.D. (1987). Growth hormesis: a by-product of control, *Health Physics* **52**, 543–547.

[87] Stocks, P. (1953). A study of the age curve for cancer of the stomach in connection with a theory of the cancer producing mechanism, *British Journal of Cancer* **4**, 407–517.

[88] Strehler, B.L. & Mildvan, A.S. (1960). General theory of mortality and aging, *Science* **132**, 14–21.

[89] Tunstall-Pedoe, H., Kuulasmaa, K., Amouyel, P., Arveiler, D., Rajakangas, A. & Pajak, A. (1994). Myocardial infarction and coronary deaths in the World Health Organization MONICA Project. Registration procedures, event rates, and case-fatality in 38 populations from 21 countries in four continents, *Circulation* **90**, 583–612.

[90] von Eckardstein, A., Malinow, R., Upson, B., Heinrich, J., Schulte, H., Schonfeld, R., Kohler, E. & Assmann, G. (1994). Effects of age, lipoproteins, and hemostatic parameters on the role of homocyst(e)inemia as a cardiovascular risk factor in men, *Arteriosclerosis and Thrombosis* **14**, 460–464.

[91] Warner, H.R., Fernandes, G. & Wang, E. (1995). A unifying hypothesis to explain the retardation of aging and tumorigenesis by caloric restriction, *Journal of Gerontology Biological Science* **50**, B107–B109.

[92] Warram, J., Laffel, L., Valsania, P., Christlieb, A. & Krolewski, A. (1991). Excess mortality associated with diuretic therapy in diabetes mellitus, *Archives of Internal Medicine* **151**, 1350–1356.

[93] Weibull, W. (1939). A statistical theory of the strength of materials, *Ingeniors Vetenskaps Akademien Handlingar* **151**, 1–45.

[94] Wetterstrand, W.H. (1981). Parametric models for life insurance mortality data: Gompertz's law over time, *Transactions of the Society of Actuaries* **33**, 159–175.

[95] Woodbury, M.A. & Manton, K.G. (1977). A random walk model of human mortality and aging, *Theoretical Population Biology* **11**, 37–48.

[96] Wu, M. & Ware, J. (1979). On the use of repeated measurements of regression analysis with dichotomous responses, *Biometrics* **35**, 513–521.

[97] Yashin, A.I. (1985). *Statistic and Control of Stochastic Processes*. Springer-Verlag, New York.

[98] Yashin, A.I. & Manton, K.G. (1997). Effects of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies, *Statistical Science* **12**, 20–34.

[99] Yashin, A.I., Manton, K.G. & Iachine, I.A. (1996). Genetic and environmental factors in the etiology of chronic diseases: multivariate frailty models and estimation strategies, *Journal of Epidemiology and Biostatistics* **1**, 115–120.

[100] Zang, E.A. & Wynder, E.L. (1996). Differences in lung cancer risk between men and women: examination of evidence, *Journal of the National Cancer Institute* **88**, 183–192.

Further Reading

Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women, *Journal of the American medical Association* **288**(3), 321–333.

KENNETH G. MANTON

Chronomedicine

Definitions and Aims

Chronobiology (from *chronos*, time; *bios*, life; and *logos*, science) investigates the mechanisms underlying variability in the otherwise unassessed physiological range, including rhythms found in us, resonating with cycles around us. Broad time structures (chronomes) consisting of deterministic chaos and trends organized by rhythms are found in organisms and in their environments. They are mapped by chronomics as the reference values for both an applied chronomedicine and a basic chronobiology. Chronomics quantify health, identifying new disease risks, diagnosing predisease and overt illness, enabling timely and timed treatment (R_x), and validating the short- and long-term efficacy of a given R_x on an inferential statistical individualized (as well as population) basis. Chronomics-based mapping includes the cartography of rhythms in us and around us and of their associations, with **hypothesis testing** and parameter **estimation** yielding ***P*-values** and 95% **confidence intervals** for the everyday preventive as well as curative self- or professional care of a given patient, rather than only for research on groups.

Introduction

About-daily **circadian** and about-yearly (circannual) rhythms, popularly biological clocks and calendars are part of broader biological time structures, chronomes. So are many more features of intermodulating rhythms with widely differing frequencies including those of the action potentials in the human brain and heart at the high-frequency end. Also in the circulation, notably of neonates and the elderly, are oscillations with periods of about a week, month, half-year, and year, including cisyears and transyears with periods shorter or longer than one year. Near the other end of the rhythm spectrum are about 11-year and multidecadal cycles, not only outside us, but also within us, influencing other components such as circadians. Rhythms, **chaos**, and (e.g. age-related) trends are chronome components interacting as feed-sidewards, multifrequency time-specified intermodulations requiring inferential statistical quantification, replacing time-unqualified feedbacks and feedforwards.

Historical Development

Confusing variability in blood eosinophil cells was resolved by averaging and stacking data over the 24-hour day. Time plots (chronograms) revealed to the naked eye large amplitude rhythms dependent upon the adrenal cortex as a cyclic mechanism preparatory for daily activity. The temporal placement of this and other rhythms could be manipulated by shifts, among others, of lighting or feeding regimens and was altered by magnetic storms. Experiments in continuous light or darkness at constant temperature and humidity or studies of humans in isolation from society documented endogenous features of "circa" rhythms that persisted with a period slightly but statistically significantly different from their exact societal daily, weekly, yearly, or decadal counterpart. These studies led to the coinage of "circadian" and other "circa" rhythms. A circadian system was extended to hormones influencing these cells and other endocrines, to the nervous and other systems, and eventually to nucleic acid formation, as well as to the effects of drugs, magnetic storms, and other physical agents such as noise or radiation. A genetic basis of biological circadian rhythms is now documented, *inter alia*, by studies on human twins reared apart (*see* **Twin Analysis**) and by chemical mutagenesis (*see* **Mutagenicity Study**) and gene transfer in fruit flies. The prefix "circa" (about) conveys the desynchronized feature and the fact that rhythm characteristics can only be defined with some statistical uncertainty.

In isolation from society, nearly identical frequencies were found for cardiovascular and geomagnetic rhythms, the latter gauged by the planetary disturbance index, K_p. What is more conclusive, "subtraction" to the point of disappearance and reamplification of environmental cycles' amplitudes was associated with dampened and amplified biological rhythms with corresponding frequency, respectively. Without causal implications, such findings provide strong hints of associations, rendering it essential to examine and compare the frequencies and phases of biological and environmental rhythms. A desynchronization as a free-run of biological rhythms must be documented not only from societal and other artificial, for example, lighting schedules but also from magnetic or other terrestrial, atmospheric, solar and galactic, for example, cosmic ray and/or other near-matching environmental cycles. The latter may pull

and amplify a biological rhythm without necessarily synchronizing it.

Different Types of Study

For **longitudinal** nearly "womb-to-tomb" monitoring in the laboratory, sensors are available for the telemetry of many vital functions. Transverse or **cross-sectional** studies are often linked systematically into a hybrid (linked cross-sectional) design with repeated measures for spans of days, weeks, years, or decades on different variables of human subjects or groups being compared (*see* **Longitudinal Data Analysis, Overview**).

Landmark Studies

Nonrandom patterns of morbidity and mortality from different causes stem largely from the times (e.g. hours) of changing resistance. Ubiquitous rhythms account for the difference between life and death, as a function only of timing, when in the experimental laboratory the same stimulus – noise, *X*-irradiation, an endotoxin, or a drug – is applied with the same dose or intensity under the same conditions to similar groups of inbred animals at different rhythm stages (e.g. 4 hours apart covering 24 hours). Fundamental life processes, RNA and DNA synthesis, exhibit reproducible rhythms underlying a circadian cell cycle, with important applications in cancer chronotherapy with chemicals or radiation. When anticancer drugs act at a specific stage of the cell cycle, it is important to time their administration in such a way as to optimize tumor cell kill. The concurrent aim is to minimize the damage to target organs, when pertinent rhythms are in near antiphase. Alternatively, when the time of greatest effectiveness does not coincide with that of least toxicity, one can

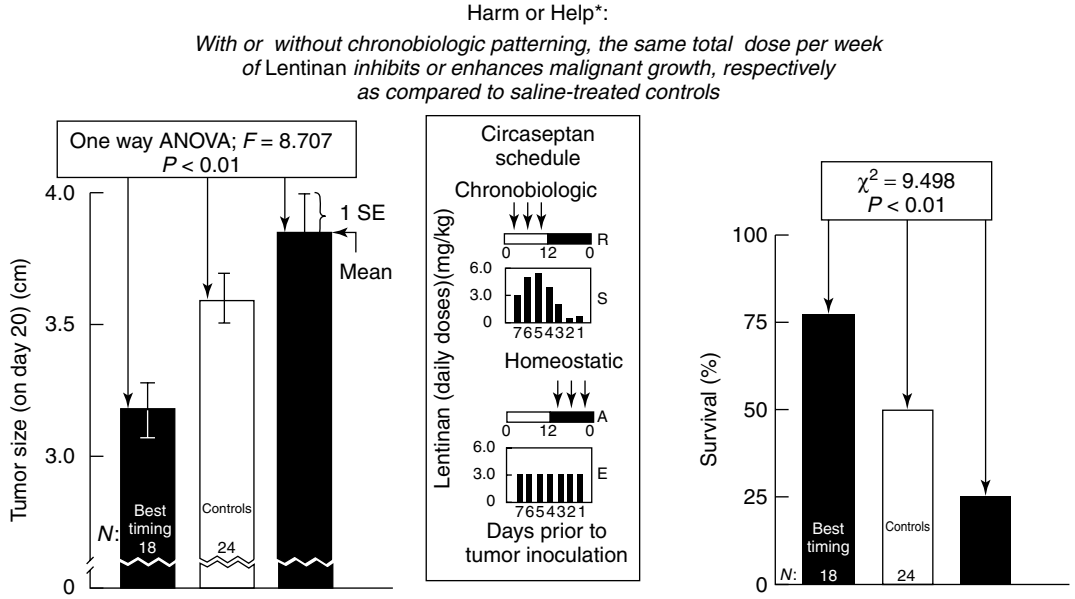

* Key: treatment during active (A) or rest (R) span with sinusoidal weekly pattern (S) or equal daily doses (E); $P < 0.01$. Difference in size corresponds to 50% difference in survival time; $N = N$ of animals. Therapeutic optimization by timing drug administration according to biologic week and day. Halberg E and Halberg F: Chronobiologia 7:95-120,1980.

Figure 1(a) The administration pattern of an immunomodulating drug accounts for the difference between the inhibition and stimulation of a subsequently implanted malignant growth. The conventional fixed daily dose pattern shortens survival time rather than lengthening it, as does a sinusoidally varying pattern adjusted to the body's rhythms, the *raison d'être* of chronotherapy. What remains to be proved in humans is that by resolving a time structure in both circadian and circaseptan aspects, clinical chronotherapy benefits from multifrequency timing, as it does from the use of circadian rhythmicity. See Figure 1b. Reproduced from Chronobiologia by permission of Franz Halberg

try to obtain this situation by manipulating before treatment the timing of host and cancer rhythms as much as possible, for example, by manipulating mealtimes. Findings on the critical importance of the circadian system have led to the fields of chronopharmacology and chronotherapy. Rhythms with other-than-circadian frequencies also matter: pretreatment with a sinusoidally patterned daily (and hourly) administration of the same total weekly dose of the immunomodulator lentinan can inhibit a subsequently implanted immunocytoma growth, when pretreatment with conventional equal daily doses enhances the same malignant growth in rats (Figure 1(a)). The use of tumor temperature as a marker rhythm to guide perioral cancer radiotherapy, as compared to the usual time-unspecified treatment, has doubled the two-year disease-free survival rate (Figure 1(b)). Further optimization involves about-yearly, about-weekly and circadian considerations, the latter two in keeping with Figure 1(a).

Other Clinical Uses of Chronomics

Chronome mapping leads to: (i) a positive definition of health in the light of reference standards for new endpoints (see Figure 2); (ii) a better understanding of mechanisms underlying changes of chronomes in healthy development (Figure 3) and against this reference standard, an earlier recognition of any disease process; (iii) the detection of chronome alterations before changes outside the physiological range occur, detecting predisease longitudinally in the stroke-prone, spontaneously hypertensive rat and in humans; and (iv) the opportunity to act preventively and rationally rather than after the fact of overt disease (Figure 4). A chronomic interpretation of serial data yields a location index (the MESOR (**m**idline **e**stimating **s**tatistic **o**f **r**hythm)) usually more accurate and more precise than the arithmetic **mean** (being associated with a smaller bias owing to the temporal placement of measurements and with a smaller **standard error** once other deterministic variation is accounted for) (Figure 5). This improved average, of great merit in itself, is only a dividend from the major merit of chronomics, namely, the provision of intuitively meaningful endpoints of dynamic changes (such as amplitudes, phases, waveforms and fundamental frequencies of rhythms), which convey useful information in their own right (Figure 2).

The rhythm characteristics (some shown in Figures 2 and 5) are confounded when dealing with

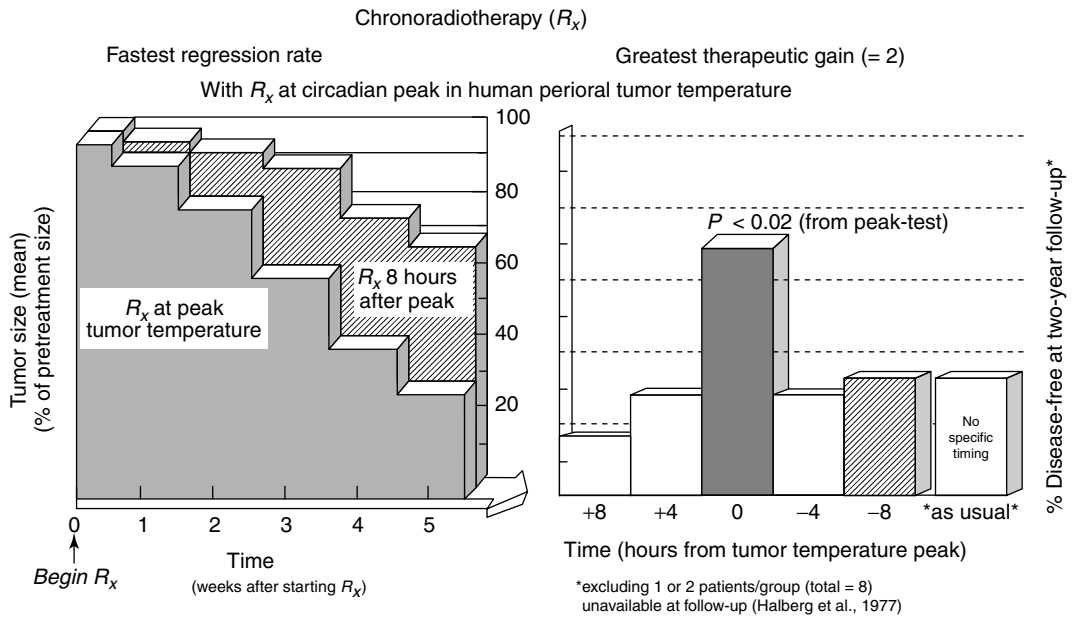

Figure 1(b) Doubling of two-year disease-free survival by radiotherapy administered at the circadian peak of tumor temperature. Reproduced from Chronobiologia by permission of Franz Halberg

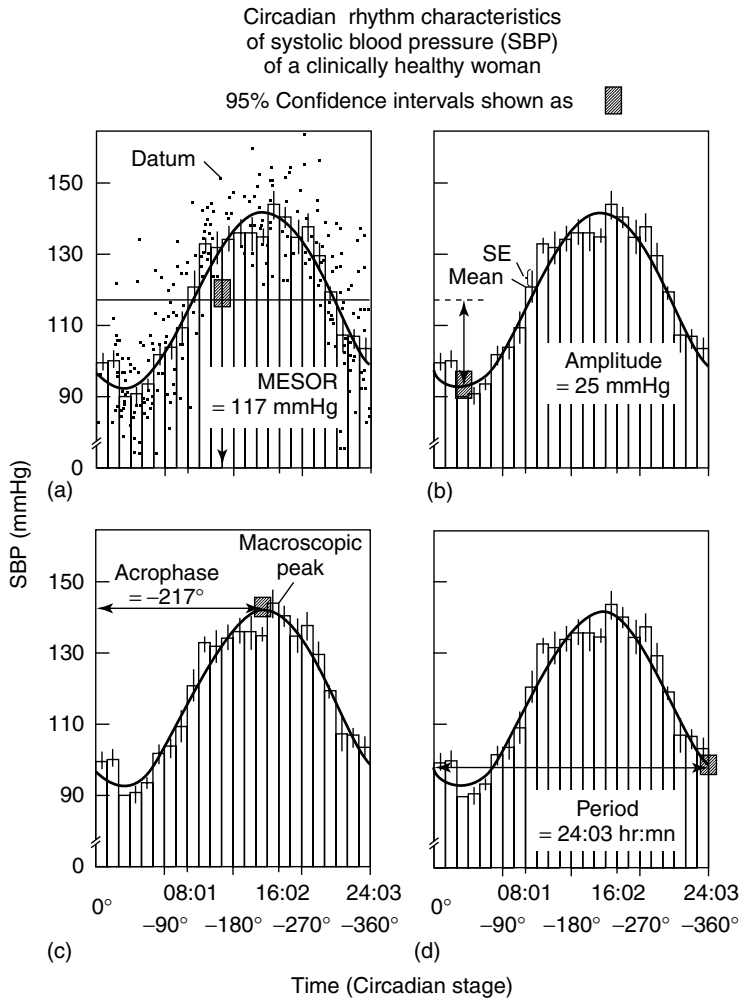

Figure 2 Illustration of circadian rhythm characteristics of systolic blood pressure of a clinically healthy woman estimated by linear–nonlinear least squares. Although the data (dots; (a)) were collected during a 9-day span and were analyzed as a longitudinal time series, the results are displayed after the data have been stacked over an idealized cycle with a period that was estimated before stacking to be 24.03 hours (d). The 95% confidence interval for the period estimate is much less than 1 hour, as can be seen from the rectangle at the tip of the arrow (d). Point-and-interval estimates are also shown for the MESOR, a rhythm-adjusted mean (a), for the amplitude, a measure of half the extent of predictable change within a cycle (b), and for the acrophase, a measure of the timing of overall high values recurring in each cycle (c). (see also abstract Figure 5.) The results also serve to indicate the large variability in systolic blood pressure, which is predictable to a large extent. These and other results, e.g., in Figure 4, also question the reliability, validity, and pertinence of single measurements used today for screening, diagnosis, and treatment of blood pressure disorders. Reproduced from Chronobiologia by permission of Franz Halberg

day–night ratios, for instance to classify patients as "dippers", "non-dippers", "reverse dippers", or "extreme dippers". A quantification of each parameter is then not available and consideration for the waveform is lost, as is all information on any extra-circadian components (there are many of them, and they can be important). Parameter estimation complemented by a nonparametric assessment of data stacked over an idealized day for comparison with time-specified reference values offers a readily

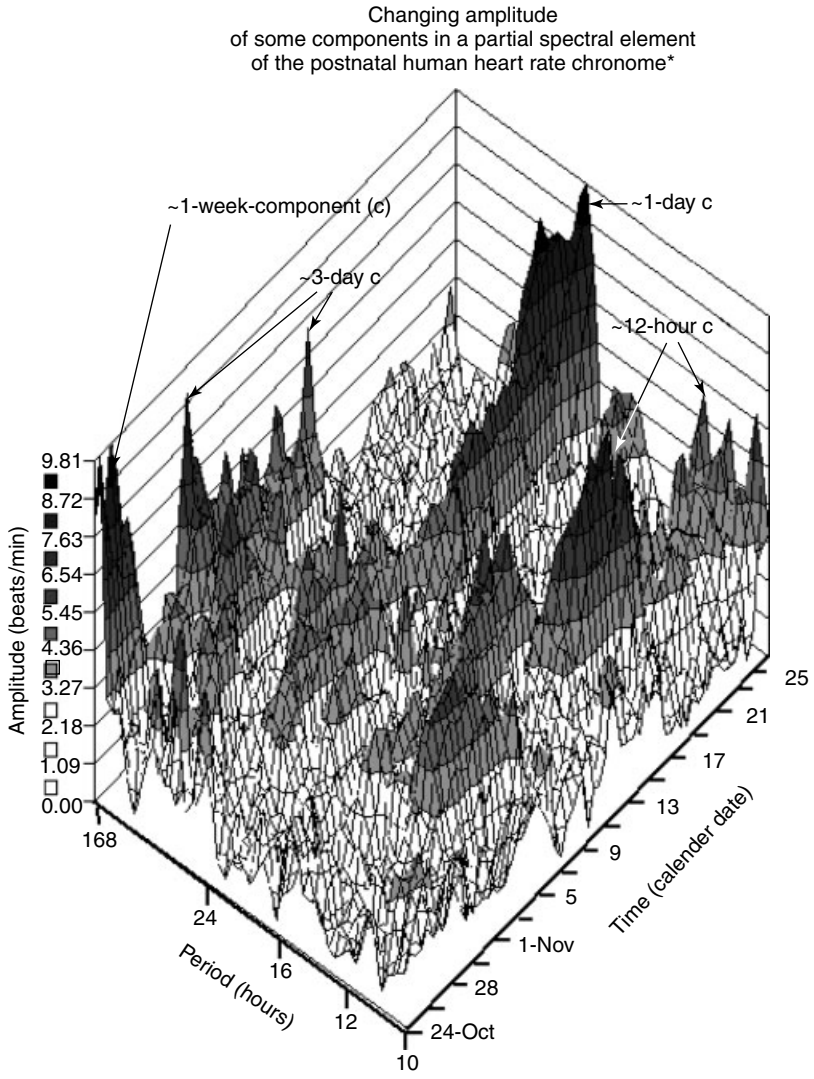

* in a healthy boy, born 19.10.1992, whose heart rate was measured at mostly 30-minute intervals from 20.10 for the ensuing 40 days, and analyzed as a moving spectrum in separate weekly intervals, displaced in 12-hour increments through the data set. An initially greater prominence of infradians (see ~1-week-component (c), left), shown by height and shading, corresponding to a larger amplitude, contrasts with the prominence of circadians and circasemidians in later weeks of life, while any ultradians with still higher frequencies and any trends and chaos, two other chronome elements, are unassessed in this gliding spectral window.

Figure 3 Early infradian over circadian prominence of human heart rate after birth. The oblique age scale ascends from the bottom middle to the right, giving the midpoints of 7-day intervals analyzed; trial periods are shown along a scale that ascends from the bottom middle to the left. Along the vertical scale of amplitudes initially no circadian peak, only an infradian (about-weekly) component is seen. The circadian and circasemidian components become noticeable by peaking several weeks later. Reproduced from Chronobiologia by permission of Franz Halberg

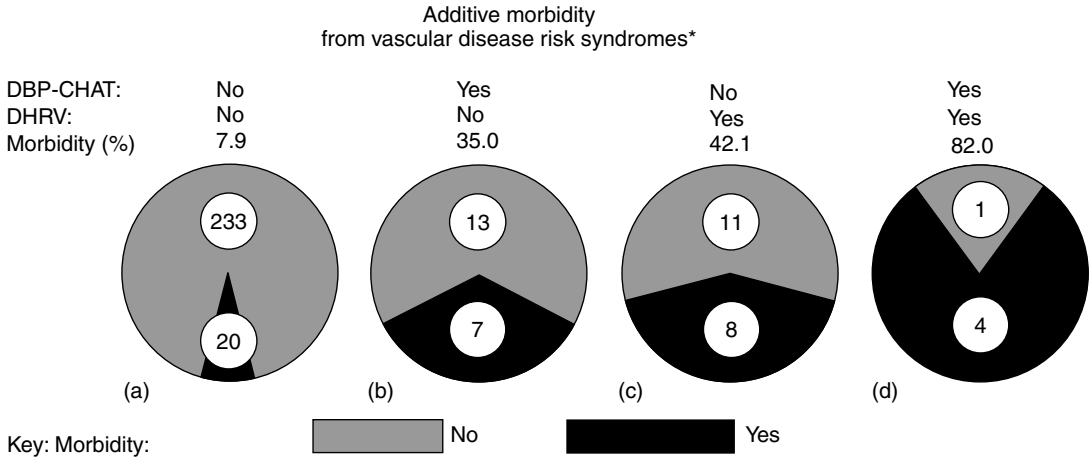

* DBP-CHAT: Diastolic circadian hyper-amplitude-tension.
DHRV: Deficient heart rate variability.
Morbidity: Cerebral ischemia, myocardial infarction, nephropathy and/or retinopathy during 6 years, without (a), with single (b, c) or both (d) disease risks.

7-day / 24-hour monitoring can detect in the neglected normal range abnormality in variabilities of blood pressure and heart rate that make the difference between <8 and 80% morbidity.

Figure 4 Pie charts compare the incidence of morbid vascular events among four groups of patients: (a) Those with an acceptable circadian blood pressure and heart rate variability; patients diagnosed with either (b). an excessive (above-threshold) circadian amplitude of diastolic blood pressure (DBP-CHAT), or (c) a decreased (below-threshold) circadian heart rate variability (DHRV), alone, and (d) patients diagnosed with both conditions. Results from a 6-year prospective study on 297 (121 MESOR-normotensive and 176 treated MESOR-hypertensive) patients, who each contributed a 48-hour record of blood pressure and heart rate measurements at 15-minute intervals at the start of study. The incidence of morbid events was checked at 6-month intervals for 6 years. Each patient's circadian characteristics of blood pressure and heart rate was interpreted in the light of reference standards obtained from independent studies of presumably clinically healthy subjects, matched by gender and age. CHAT (**c**ircadian **h**yper-**a**mplitude-**t**ension) was defined as a circadian amplitude of blood pressure above the upper (95% prediction) limit of acceptability and DHRV as a 48-hour standard deviation of heart rate below the lower limit of acceptability. Findings of Kuniaki Otsuka, in keeping with earlier studies of the spontaneously hypertensive stroke-prone Okamoto rat, and in keeping with human studies in Minnesota, Taiwan, Japan, Italy, and Germany, where outcomes are available with a 28-year perspective. Reproduced from Chronobiologia by permission of Franz Halberg

understood diagnosis, summarized on a form called a "sphygmochron" shown in Figure 6. Biostatistics should be made as simple as possible but not simpler, to paraphrase Einstein.

Examples of diagnoses are the disease risks: chronome alterations of heart rate variability, CAHRV, such as a **d**ecreased (under-threshold) **h**eart **r**ate **v**ariability (DHRV), and an excessive (over-threshold) variability of blood pressure (circadian hyper-amplitude-tension, CHAT). CHAT describes a blood pressure profile with a circadian amplitude above the upper 95% prediction limit of healthy peers matched by gender and age. CHAT can be a response during only a few days to stimuli such as conflict or grief (transient CHAT). CHAT beyond a week-long monitoring should prompt further monitoring and the initiation of non-drug treatment, e.g., by relaxation procedures. Patients with diastolic CHAT, whether MESOR-normotensive or MESOR-hypertensive, have a 720% increase in the risk of cerebral ischemic events. The diagnosis of CHAT requires both that data be collected around the clock and that the circadian amplitude be estimated and compared with available reference values. DHRV (decreased **h**eart **r**ate **v**ariability) is a deficient heart rate jitter (defined as a 24-hour standard deviation of heart rate below the lower 5% prediction limit of healthy peers), which carries a

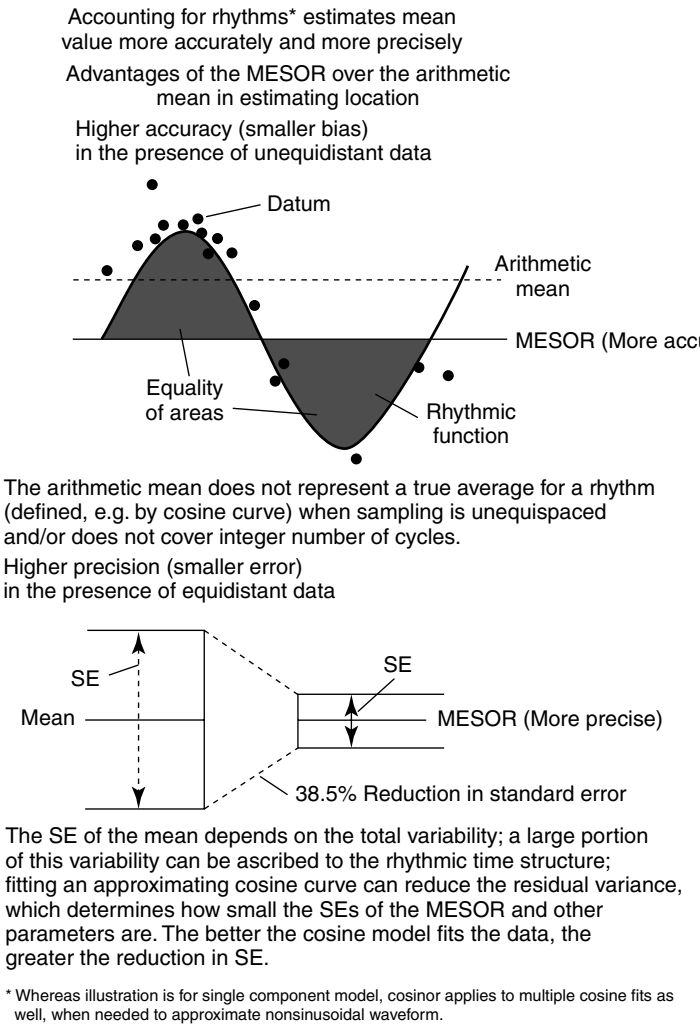

Figure 5 In the presence of periodicities, the use of statistics to resolve the time structure (chronome) usually yields a more accurate (top) and more precise (bottom) estimate of location (the MESOR, a rhythm-adjusted mean) than the arithmetic average. Reproduced from Chronobiologia by permission of Franz Halberg

550% increase in the risk of coronary artery disease and can be determined along with a check for CHAT by ambulatory monitoring without electrodes. When both CHAT and DHRV coexist, there is a doubling of the risk of vascular diseases (Figure 4). An above-threshold circadian pulse pressure (around-the-clock average difference between systolic and diastolic pressure, when the heart contracts and relaxes) further increases vascular disease risk, as may do an odd circadian timing, circadian ecphasia.

Statistical Concepts, Problems, and Solutions

Periodograms, Power Spectra, and Single (Usually Multiple-component) Linear-nonlinear Cosinors

In the precomputer era, periodograms (*see* **Spectral Analysis**) used at first necessitated equidistant data over several integer cycles of the components characterizing the data. The computations

SPHYGMOCHRON™ -S (short form)

Monitoring Profile Over Time;
Computer Comparison
with Peer Group Limits

Blood Pressure (BP) and Related Cardiovascular Summary

(Circadian Sphygmochron; from *sphygmo-*, of or relating to the circulation, notably blood pressure, as well as pulse and *chronos*, time)

Name _____ Patient # _____ No. of Profile: _____

Age _____ Sex M F Monitoring From _____ To _____, 20 _____

Time of: Awakening (A) _____ (_____) Falling Asleep (S) _____ (_____)
Day of profile (Habitually) Day of profile (Habitually)

R_x: _____ Comments[1,2] _____

Chronobiologic Characteristics

	Systolic BP (mmHg)		Diastolic BP (mmHg)		Heart Rate (bpm)	
	Patient Value	Peer Group Reference Limits	Patient Value	Peer Group Reference Limits	Patient Value	Peer Group Reference Limits
Adjusted 24-h Mean (MESOR)		Range		Range		Range
Predictable Change (Double Amplitude)		Range		Range		Range
Timing of Overall High Values (Acrophase) (hr:min)		Range		Range		Range

	STD (Min; Max)	STD (Min; Max)	STD (Min; Max)
Percent Time of Elevation			
Timing of Excess	(hr:min)	(hr:min)	(hr:min)
Extent of Excess During 24 Hours	(mmHg × hour)	(mmHg × hour)	(bpm × hour)
10-Year Cumulative Excess	(mmHg × hour) (In 1,000's units)	(mmHg × hour) (In 1,000's units)	(bpm × hour) (In 1,000's units)

Individualized bounded indices: (STD = Standard) (Min = Minimum) (Max = Maximum) (HBI = Hyperbaric index)

Intervention Needed

☐ No
☐ Yes ☐ Drug ☐ Non-Drug

More Monitoring Needed

☐ Annually
☐ As soon as possible
☐ Other specify _____

Prepared By _____ Date ____/____/____

1) Unusually long standing or lying-down during waking: unusual activity, such as exercise, emotional loads, or schedule changes, e.g., shiftwork, etc.: 2) Salt, calories, kind and amount, other, etc. Please enter this information into a separate diary along with daily times of getting up and retiring for sleep.

Figure 6 Sphygmochron, a computer-generated form used to summarize results from the combined parametric and nonparametric assessment of a blood pressure and heart rate profile. Results from both approaches are compared with reference values specified by gender and age, given in boxes next to the given subject's estimates of his/her rhythm characteristics. Reproduced by permission of Franz Halberg

were time-consuming and data could not always be obtained at regular intervals. Self-measurements of blood pressure or heart rate, for example, are not possible while sleeping. An alarm clock used to prompt a self-measurement leads to disturbance, which may affect the measurements and hence may prevent the rigorous assessment of spontaneous variation. Undue caution at the beginning of the computer era led to very conservative power spectra, with a great deal of smoothing. As the ubiquity of circadians was documented, **least-squares** procedures offered themselves for the test of anticipated rhythms in unequidistant data such as those collected in isolation from society in caves, or rooms without a clock. Thus, the single cosinor was developed. Here *single* refers to the analysis of a single series by the fit of one or, usually, of several components when the density and length of the data allow it. The addition of harmonic terms in the model quantifies the waveform when it is nonsinusoidal. The results are displayed along both rectangular and polar coordinates as point estimates and 95% confidence intervals. For time series spanning more than one or a few cycles, a chronobiologic serial section can be used, wherein the single cosinor is applied to successive consecutive or partly overlapping intervals to examine how the characteristics of a rhythm with a given frequency vary as a function of time. For long series involving components of several frequencies, chronobiologic serial sections of several orders can be applied to the original data or to the parameters obtained in a previous pass.

Least-squares procedures are well suited to the situation where anticipated rhythmic components have known approximate periods (τ_i). The least-squares fit of a model such as

$$Y(t) = \sum_{j=0}^{q} a_j t^j + \sum_{i=1}^{p} A_i \cos\left(\frac{2\pi t}{\tau_i} + \phi_i\right) + \varepsilon(t)$$

detects a rhythm with period τ_i by the zero-amplitude test ($H_0 : A_i = 0$). Confidence and/or prediction limits are derived for the parameters of all rhythmic components, whether they represent several physiologically different (multifrequency) rhythms and/or harmonics quantifying a nonsinusoidal waveform. In addition, any superimposed trend is detected by nonzero **polynomial** coefficients (a_j). Least-squares techniques to assess rhythms in short and sparse series led to several important developments.

1. Parameter comparisons: these can check for changes occurring on an individual basis, for example, to determine whether a given antihypertensive drug has lowered the circadian blood pressure amplitude of a patient with CHAT, or whether it is preferable to administer such a treatment at one versus another circadian stage.
2. Gliding spectral windows in combination with cumulative sums (CUSUM, Figure 7) (*see* **Quality Control in Laboratory Medicine**) ascertain that an effect of treatment occurs and persists with statistical significance in the given patient, when his/her blood pressure, in response to antihypertensive treatment, leaves the decision interval.
3. Phase zero trials: the parsimonious single cosinor method relying on prior information (such as the critical importance of circadian stage in relation to treatment efficacy) accounts for powerful chronobiologic pilots that are always useful but are named "phase zero trials" since usually they should precede the customary Phase I–III **clinical trials**, which then could be carried out at the "right time" determined in the phase zero trial.

For different signal-to-noise ratios and a relatively small number of subjects (≤ 20), in a hardly ever random (but rather somewhat periodic) world, the power of the single cosinor method exceeds that of a one-way analysis of variance (*see* **Experimental Design**), assuming that the (usually considered six) test times are equidistributed within one (e.g. circadian) cycle. The dangers of relying on a two-timepoint approach need to be stressed whenever the precise phase information is lacking and/or the individual's rhythm may be desynchronized in phase or period. This comment applies equally to the selection of only two time-spans, as discussed above in relation to blood pressure regarding the preference of assessing circadian characteristics over merely computing a day–night ratio. The merit of a six-timepoint design at the outset is its amenability to cost-effectively determine the optimal time and the likely gain to be derived from timing. For instance, in a six-subject, six-timepoint pilot study, the particular tested anticlotting properties of treatment with daily low doses of aspirin were optimal when the drug was administered shortly after awakening (Figure 8).

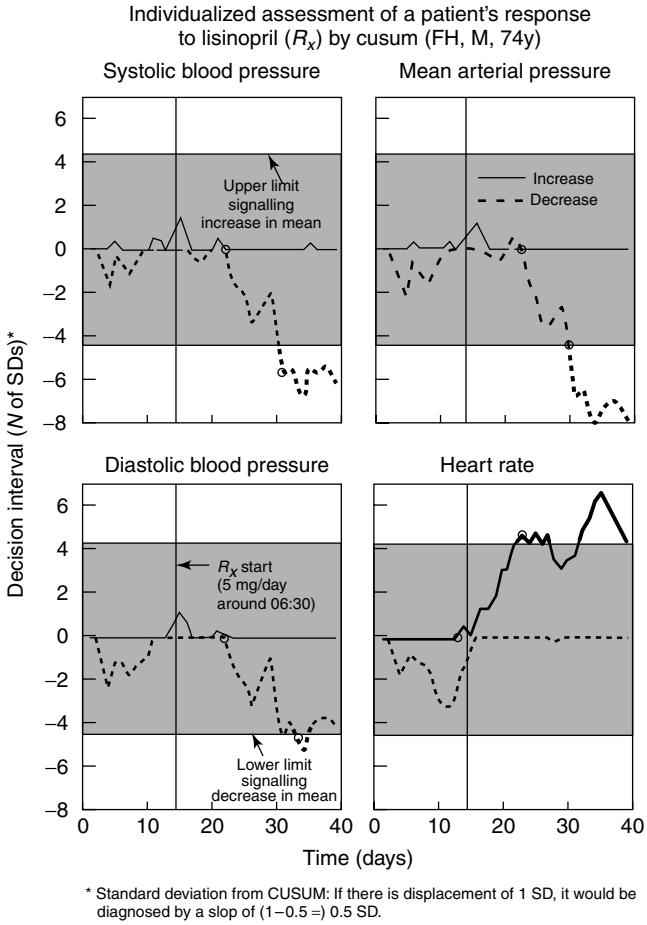

Figure 7 Control charts of daily mean values of blood pressure and heart rate data collected at 15-minute intervals around the clock. While the series of daily means is proceeding "in control" (i.e. at the pretreatment mean value), the CUSUM comprises two line graphs, signaling an increase or decrease in mean, respectively, that generally stay within the shaded "decision interval", plotted here as the horizontal lines at 4.4 and -4.4 standard deviations (SD). When the dashed curve breaks out downward of the decision interval boundary, it provides the validation of a decrease in daily blood pressure mean. The time at which the mean changed is estimated by tracking the line segment leading to the breakout back to the last occasion on which it lay on the horizontal axis. Thus, in the case of systolic blood pressure, the breakout occurs on day 30 (16 days after the start of treatment with the drug lisinopril) and the shift in pressure is estimated to have occurred on day 22 (8 days after lisinopril treatment started). An upward breakout for heart rate shows the desirability of continued monitoring to see whether a breakout is transient or sustained. In the case of blood pressure, a return into the decision interval can occur after several months of successful intervention, when an event led to the treatment's failure latter on (not shown). Reproduced from Chronobiologia by permission of Franz Halberg

Indications for Linear–Nonlinear Least-squares Cosinors, Spectra, and Cross-spectra

The time structure of a variable is usually synchronized by the environment. Transient changes are associated with the expression of partly endogenous variations when an organism is studied under conditions rendered as constant as possible in terms of illumination, temperature, humidity, access to food and water, and so on. Persisting rhythms assume periods, which usually remain close to their environmental match, yet differ with statistical significance, albeit slightly, from the period of the environmental cycle with which they had been synchronized earlier.

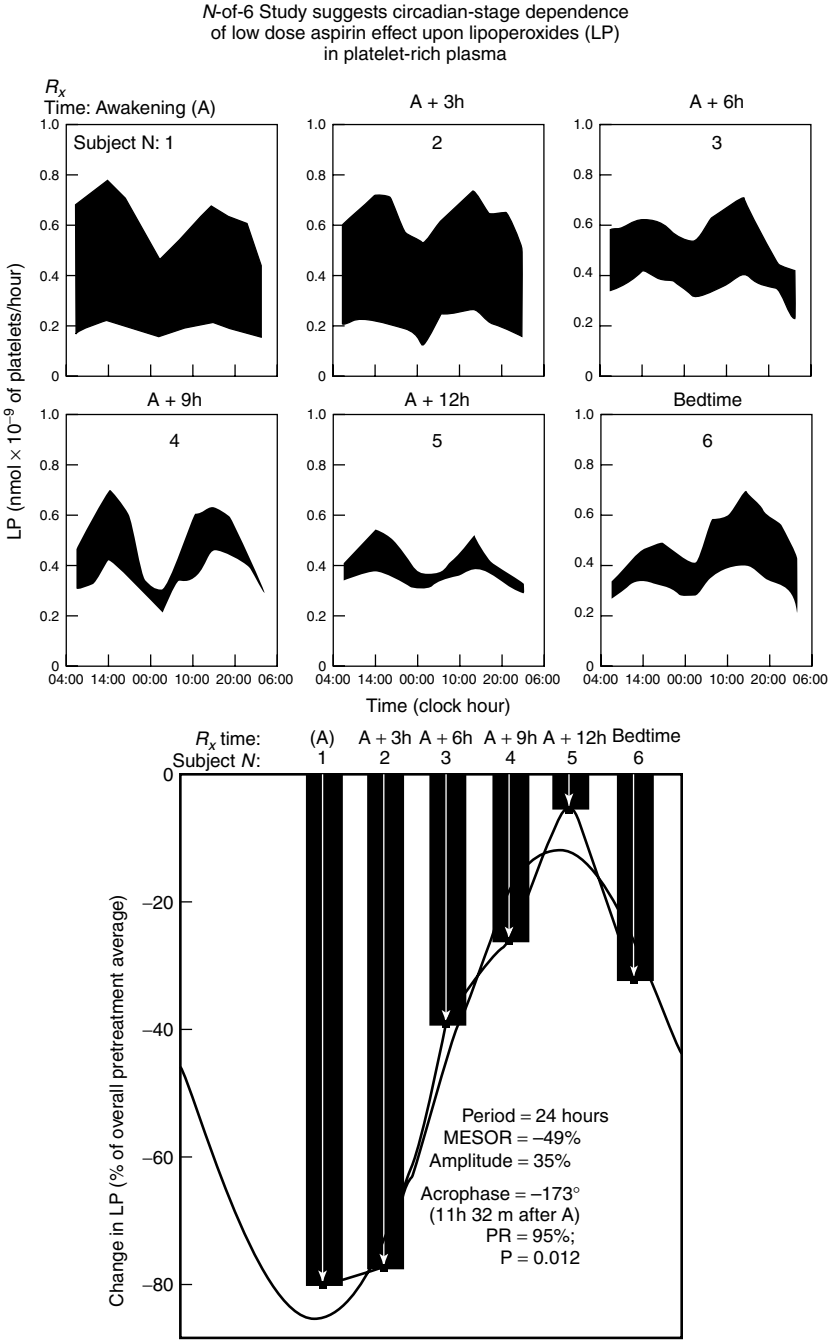

Figure 8 Power of "phase zero" chronobiologic pilots: by randomly assigning similar subjects to six different circadian stages, each to receive 100 mg/day of aspirin for one week, a large amplitude response rhythm can be assessed indicating that the lowering of lipoperoxides (LP) in platelet-rich plasma (a desired effect to prevent myocardial infarction) is maximal when aspirin is taken shortly after awakening and that aspirin does not have this effect when it is given 12 hours later. Reproduced from Chronobiologia by permission of Franz Halberg

Nonlinear least-squares techniques (*see* **Nonlinear Regression**) generally serve to estimate the period(s) with other rhythm characteristics. The combination of linear and nonlinear least squares relies on guess estimates from the former to assess by the latter the persisting circadian or other rhythms, for instance in the absence of time clues. Evidence accumulates for the desynchronization from societal schedules of about-weekly (circaseptan) rhythms *in vitro* and *in vivo*, and for their frequency multiplication to about-3.5-day (circasemiseptan) rhythms after enucleation and/or mutation in unicells.

Asynchronization from the calendar year is observed further for the case of an about 1.05-year component in the daily excretion of breakdown products of steroidal hormones recorded (with gaps) for 15 years. Figure 9 shows gliding (middle) and global (right) spectral windows of these data, complementing an earlier periodogram (left) obtained without an estimate of the period's uncertainty. Percentage rhythm (R^2) values and/or ordering P values (from the zero-amplitude test) can also be displayed along with amplitudes, each in separate gliding spectra, for biological and/or physical variables for the mapping of chronomes in and around us.

Methodological Challenge in Finding Out that Stormy Weather in Space Is a Health Hazard

The combined use on existing extensive databases of spectral coherence, superposed epochs, and other remove-and-replace approaches, as in **endocrinology** (allowing nature to ablate and replace certain frequencies, e.g. of the velocity changes in the solar

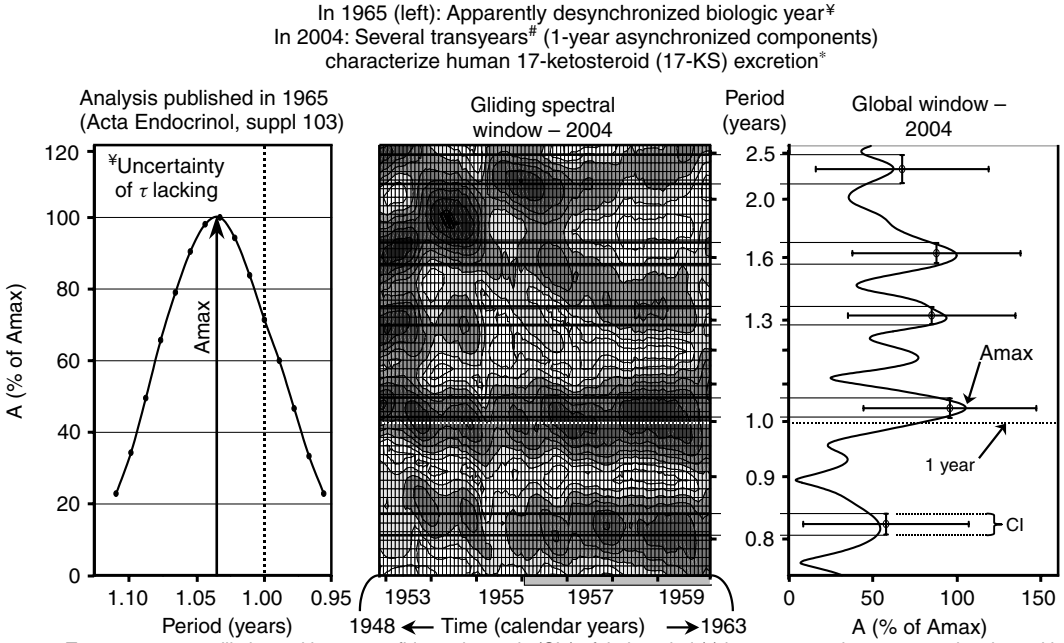

Transyears are oscillations with 95% confidence intervals (CIs) of their period (τ) between 1 and 2 years, overlapping neither of these values nor one another. Cisyears are oscillations with periods between 1 and 0.5 year with CIs overlapping neither 1-year nor 0.5-year, as seen for a period of about (but not equal to) 0.8 y. * about 15-year record, daily data (with gaps) from a clinically healthy man (CH, 43 – 58 y), N = 3719. Gliding interval = 8 years (y), increment = 1 month, longest trial τ = 2.265 y, shortest τ = 0.75 y, harmonic increment = 0.05. Darker shading corresponds to larger amplitude (A). Horizontal shaded bar starting in 1956 indicates span of intermittent androgen administration; data during treatment omitted from analysis.

Figure 9 Combination of gliding (middle) and global (right) spectral windows identifies several transyears, with periods of about 1.6, 1.3, and 1.05 years, and a cisyear with a period of about 0.8 year in a 15-year record of urinary excretion of 17-ketosteroids by a healthy man. All these components are validated nonlinearly with 95% confidence intervals of their periods non-overlapping precisely 1 year. The absence of an exact calendar year by cosinor (right) had been published in 1965. Reproduced from Chronobiologia by permission of Franz Halberg

wind), has suggested that magnetic storms are consistently associated with a decreased heart rate variability. This could constitute a physiological basis for an increased localized **incidence** of myocardial infarctions and strokes also observed in association with magnetic storms. Notably in the Arctic, around-the-clock electrocardiograms covering seven days allowed the comparison of data from days with high versus low geomagnetic activity. The long-term, eventually life-long concomitant systematic monitoring of physiological variables for alignment with ongoing physical monitoring is the aim of an international chronome initiative seeking information in different geographic/geomagnetic locations as reference values for chronomedicine, while also examining questions about mechanisms of external–internal chronome interactions.

The reciprocity of cycles in us and around us led to the recent discovery of components with periods slightly longer or shorter than one year that may coexist with the circannual variation. These are the far-transyear with a period of about 1.3 years characterizing the solar wind speed, the near-transyear and cisyear with periods of about 1.05 and about 0.95 year(s), respectively, some of which can also be found in some helio- and/or geomagnetic indices. Because such components can beat with the circannual variation, circannual studies are best conducted over long enough spans to avoid obtaining controversial results corresponding to spans when the two cycles are in or out of phase. Chronomics, the mapping of broad time structures, including these newly found cycles, can serve as useful reference for designing studies with appropriate sampling recommendations.

Population-mean Cosinor

This method, based on **multivariate** statistics, was developed for drawing inferences to be generalized by checking the extent of similarity of single-cosinor estimates among individuals selected at random from a homogeneous population.

Time-specified Reference Standards: Chronodesms

This chronobiologic alternative for usual value ranges collects serial data from clinically healthy subjects to derive reference limits (such as 90% prediction intervals) that account for multifrequency rhythms, age trends (from womb to tomb), and differences as a function of gender and ethnicity, considering both changes in mean value and **variance**. These limits are for the interpretation of single values and for that of rhythm parameters (parameterdesms) and noise characteristics. This approach to the monitoring of blood pressure and heart rate identifies patients with CHAT or DHRV, among others.

Measures of Excess and Deficit and Beyond with Signal Averaging

The recognition of chronomes provides more reliable answers to the question whether and when a time series is too high or too low and detects alterations in time structure in the absence of changes in operating overall average. Chronomics in addition to comparing endpoints of anticipated periodic components with a pertinent chronodesm to determine the extent, timing, and duration of any excess (or deficit) also numerically integrates the area (under and/or over the curve) delineated by the data when they are outside time-specified limits and the limits themselves. The time when most of the excess (deficit) occurs serves for diagnosis and for timing any intervention.

Control Charts

To assess an individual's trends in mean or in other chronome characteristics, cosinor methods are applied in intervals that are progressively displaced throughout the accumulating **time series**. This chronobiologic serial section can be combined with a self-starting cumulative sum (CUSUM) to detect chronome alterations or to assess the response to a given intervention (*see* **Quality Control in Laboratory Medicine**). Such an individualized approach, first used in chronomedicine for the monitoring of epileptic seizures and for adjusting treatment, is particularly indicated in assessing blood pressure and heart rate. Hawkins' self-starting CUSUM detects a shift in mean (MESOR, circadian amplitude, and/or any other pertinent chronome endpoint or chrone) and indicates when the change may have occurred (Figure 7). Moreover, the boundaries of the decision interval can be determined even from relatively few data prior to a given intervention. Gliding spectral

windows (Figures 3 and 9) provide a view, from above or from the side, of the change in both amplitude and period.

Anticipated Developments

Some nondrug treatments or an antihypertensive drug can lower an excessive circadian blood pressure amplitude when given at the right time (rhythm stage) or raise it when given at the wrong time. For instance, an α-adrenoceptor antagonist given for benign prostatic hypertrophy in the evening raises the circadian blood pressure amplitude, but does not have this effect when given in the morning, when the circadian blood pressure amplitude is restored within acceptable limits. Further clinical trials will have to examine the degree of generality of the finding already made, that the actual incidence of adverse vascular events can be reduced by antihypertensive agents capable of lowering an excessive circadian blood pressure amplitude when given at the right time(s).

Deterministic chaos theory of heart rate variability has associated complexity, if not irregularity, with health, while regularity (periodicity) has been regarded as an index of disease, with the focus primarily on spectral components with periods of seconds or a few minutes. Results along the 24-hour scale reveal that the circadian variation is better defined in health than in the presence of heart disease. Concomitant assessment of various chronome elements reveals rhythms in endpoints of "chaos". Trends in both rhythmic and chaotic endpoints are found as a function of age and in disease versus health. For example, the correlation dimension of fractal scaling separates healthy subjects from patients with coronary artery disease at 2 A.M., but not at 10 A.M. or 2 P.M. Sampling around the clock not only reveals a circadian rhythm in the correlation dimension of cardiac interbeat intervals in health, but also a variance transposition of an endpoint of chaos from the 24-hour to the 12-hour region of the rhythm spectrum, as a new feature of CAHRV in patients with heart disease.

Chronome-specified Interactions

Chronome-specified interactions among two or more different variables have been called feed-sidewards. Those among three rhythmic entities such as the pineal–pituitary–adrenocortical interactions have been modeled *in vitro*. A rhythmic sequence of stimulation, no-effect, and inhibition by a third entity, a modulator, such as the pineal, upon the interaction of two other entities, the actor and reactor, such as the pituitary and the adrenal, is gauged by the *in vitro* production of corticosterone (see Figure 10). Another feed-sideward applies to DNA labeling in bone. An ACTH analog leads to a circadian sequence of stimulation, no-effect, and inhibition. Studies of feed-sidewards will have to be extended to multiple interactions at several (e.g. circadian and about 7-day rhythmic) chronome components, since the results can be critically important (Figure 1), as different as the stimulation versus the inhibition of a malignant growth.

Automatic Closing of the Loop Between a Chronodiagnosis and Chronotherapy

Longitudinal monitoring of vital signs for surveillance has been advocated, at least for at-risk individuals. Rather than loosing all original data except those collected just prior to an event, as done in the black box of an airplane, the data steadily accumulating over years or decades can be analyzed in relatively short spans to extract the pertinent spectral, chaotic, and trend (chronome) characteristics. Endpoints extracted from such windows are stored, thus compacting the available information as a summary for each day and then for each week or for a longer span. The information is thus progressively updated as the window is displaced and enlarged in repeated passes as-one-goes, while components with progressively lower frequencies are thus gradually resolved. This continuous examining, compacting, and recycling of information based on progressively broader windows can detect the earliest chronome alterations at one or the other frequency, which may indicate an increased disease risk and can prompt the institution of countermeasures.

Automatic monitoring devices, miniaturized for long-term ambulatory use, some implanted under the skin or in the heart, are already available for research. The windowing, compacting, and recycling of telemetered data could provide a continuous medical examination, eventually available to everybody, thus contributing a thorough objective history of vital signs, preferably retrieved in response to the push of

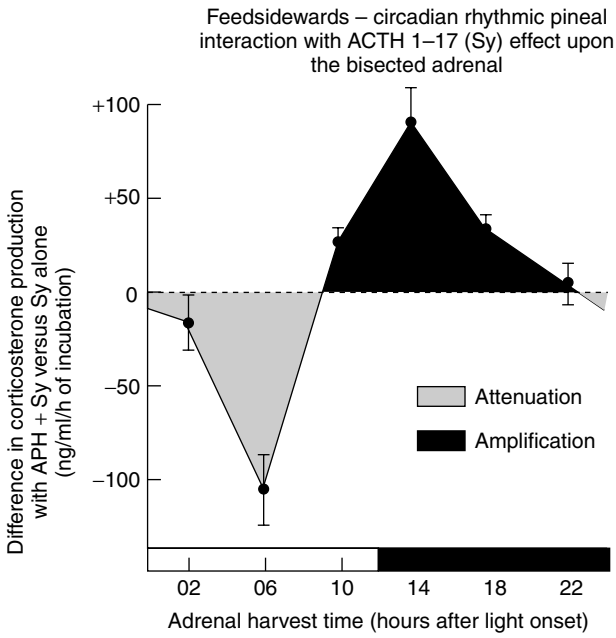

Figure 10 Much controversy can be resolved by studying the effect of the interaction by more than two entities at different rhythm stages: a third entity may modulate, in a predictable insofar as rhythmic fashion, the effect of a first entity upon a second. Predictable sequences of attenuation, no-effect and amplification can then be found. A case in point is corticosterone production by bisected adrenals stimulated by ACTH 1–17 (Sy) in the presence versus absence of pineal homogenate (APH). Such chronomodulations are part of (time-specified) feed-sidewards, for example, of rhythmic sequences of attenuation, no-effect and amplification by a modulator upon the interaction of an actor and a reactor. The figure summarizes five studies by Salvador Sanchez de la Peña with us. Reproduced from Chronobiologia by permission of Franz Halberg

a button. Eventually, the loop may be closed for automatic chronome-adjusted treatment with drug pumps and/or electrical treatment devices. Another merit of the chronome approach versus the airplane's black box may be the much more complete history of long-term antecedents to avoid the "crash" of catastrophic disease by timely and timed treatment.

Further Reading

Bingham, C., Arbogast, B., Cornélissen, G., Lee, J.K. & Halberg, F. (1982). Inferential statistical methods for estimating and comparing cosinor parameters, *Chronobiologia* **9**, 397–439.

Bingham, C., Cornélissen, G. & Halberg, F. (1993). Power of "Phase 0" chronobiologic trials at different signal-to-noise ratios and sample sizes, *Chronobiologia* **20**, 179–190.

Burioka, N., Cornélissen, G., Halberg, F., Kaplan, D.T., Suyama, H., Sako, T. & Shimizu, E. (2003). Approximate entropy of human respiratory movement during eye-closed waking and different sleep stages, *Chest* **123**, 80–86.

Cornélissen, G. & Halberg, F. (1994). *Introduction to Chronobiology*. Medtronic Chronobiology Seminar No. 7, April 1994, `http://www.msi.umn.edu/~halberg`.

Cornélissen, G. & Halberg, F. (1996). Impeachment of casual blood pressure measurements and the fixed limits for their interpretation and chronobiologic recommendations, in *Time-dependent Structure and Control of Arterial Blood Pressure*, F. Portaluppi & M.H. Smolensky, eds; *Annals of the New York Academy of Sciences* **783**, 24–46.

Cornélissen, G., Halberg, F., Breus, T., Syutkina, E.V., Baevsky, R., Weydahl, A., Watanabe, Y., Otsuka, K., Siegelova, J., Fiser, B. & Bakken, E.E. (2002). Non-photic solar associations of heart rate variability and myocardial infarction, *Journal of Atmospheric and Solar-Terrestrial Physics* **64**, 707–720.

Halberg, F. (1959). Physiologic 24-hour periodicity; general and procedural considerations with reference to the

adrenal cycle, *Zeitschrift für Vitamin-, Hormon und Fermentforschung* **10**, 225–296.

Halberg, F. (1969). Chronobiology, *Annual Review of Physiology* **31**, 675–725.

Halberg, F. (1980). Chronobiology: Methodological problems, *Acta Medica Romana* **18**, 399–440.

Halberg, F. (1983). Quo vadis basic and clinical chronobiology: promise for health maintenance, *American Journal of Anatomy* **168**, 543–594.

Halberg, F., Bakken, E., Cornélissen, G., Halberg, J., Halberg, E., Wu, J., Sánchez de la Peña, S., Delmore, P. & Tarquini, B. (1990). Chronobiologic blood pressure assessment with a cardiovascular summary, the sphygmochron, in *Blood Pressure Measurements*, W. Meyer-Sabellek, M. Anlauf, R. Gotzen & L. Steinfeld, eds. Steinkopff-Verlag, Darmstadt, pp. 297–326.

Halberg, F. & Bingham, C. (1987). The scope and promise of chronobiology and biostatistics: interpenetrating, inseparable disciplines, in *Proceedings of the Biopharmaceutical Section, American Statistical Association*. Chicago, August 15–18, pp. 11–32.

Halberg, F., Bingham, C. & Cornélissen, G. (1993). Clinical trials: the larger the better? *Chronobiologia* **20**, 193–212.

Halberg, F. & Cornélissen, G. (1995). International Womb-to-Tomb Chronome Initiative Group: Resolution from a meeting of the international society for research on civilization diseases and the environment (New SIRMCE confederation), in *Fairy Tale or Reality? Medtronic Chronobiology Seminar No. 8*. Brussels, March 17–18, 1995; `http://www.msi.umn.edu/~halberg/`, April 1995.

Halberg, F., Cornélissen, G., Katinas, G., Syutkina, E.V., Sothern, R.B., Zaslavskaya, R., Halberg, F., Watanabe, Y., Schwartzkopff, O., Otsuka, K., Tarquini, R., Perfetto, F. & Siegelova, J. (2003). Transdisciplinary unifying implications of circadian findings in the 1950s. *J Circadian Rhythms* **1**, 2. pp. 61 `www.JCircadianRhythms.com/content/pdf/1740-3391/1/2.pdf`.

Halberg, F., Cornélissen, G., Otsuka, K., Katinas, G. & Schwartzkopff, O. (2001). Essays on chronomics spawned by transdisciplinary chronobiology. Witness in time: Earl Elmer Bakken, *Neuroendocrinology Letters* **22**, 359–384.

Halberg, F., Cornélissen, G., Otsuka, K., Schwartzkopff, O., Halberg, J. & Bakken, E.E. (2001). Chronomics, *Biomedicine and Pharmacotherapy* **55**, (Suppl. 1), 153–190.

Halberg, F., Halberg, E., Barnum, C.P. & Bittner, J.J. (1959). Physiologic 24-hour periodicity in human beings and mice, the lighting regimen and daily routine, in *Photoperiodism and Related Phenomena in Plants and Animals*, Publ. No. 55, R.B. Withrow, ed. American Association for the Advancement of Science, Washington, DC, pp. 803–878.

Halberg, F., Lee, J.K. & Nelson, W.L. (1978). Time-qualified reference intervals – chronodesms, *Experientia (Basel)* **34**, 713–716.

Hawkins, D.M. (1987). Self-starting CUSUM charts for location and scale, *Statistician* **36**, 299–315.

Johnson, E.A., Haus, E., Halberg, F. & Wadsworth, G.L. (1959). Graphic monitoring of seizure incidence changes in epileptic patients, *Minnesota Medicine* **42**, 1250–1257.

Macey, S.L. ed. (1994). *Encyclopedia of Time*. Garland, New York.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters, *Journal of the Society of Industrial and Applied Mathematics* **11**, 431–441.

Nelson, W., Cornélissen, G., Hinkley, D., Bingham, C. & Halberg, F. (1983). Construction of rhythm-specified reference intervals and regions, with emphasis on "hybrid" data, illustrated for plasma cortisol, *Chronobiologia* **10**, 179–193.

Otsuka, K., Cornélissen, G. & Halberg, F. (1996). Predictive value of blood pressure dipping and swinging with regard to vascular disease risk. *Clinical Drug Investigation* **11**, 20–31.

Otsuka, K., Cornélissen, G. & Halberg, F. (1997). Circadian rhythmic fractal scaling of heart rate variability in health and coronary artery disease. *Clinical Cardiology* **20**, 631–638.

Otsuka, K., Cornélissen, G., Weydahl, A., Holmeslet, B., Hansen, T.L., Shinagawa, M., Kubo, Y., Nishimura, Y., Omori, K., Yano, S. & Halberg, F. (2001). Geomagnetic disturbance associated with decrease in heart rate variability in a subarctic area, *Biomedicine and Pharmacotherapy* **55**, (Suppl. 1), 51–56.

Reinberg, A. & Smolensky, M.H. (1983). Biological rhythms and medicine. *Cellular, metabolic, physiopathologic, and pharmacologic aspects*. Springer, New York.

Shinagawa, M., Kubo, Y., Otsuka, K., Ohkawa, S., Cornélissen, G. & Halberg, F. (2001). Impact of circadian amplitude and chronotherapy: relevance to prevention and treatment of stroke, *Biomedicine and Pharmacotherapy* **55**, (Suppl. 1), 125–132.

Smolensky, M., Halberg, F. & Sargent, F. II (1972). Chronobiology of the life sequence, in *Advances in Climatic Physiology*, S. Itoh, K. Ogata, H. Yoshimura, eds; Igaku Shoin Ltd., Tokyo, pp. 281–318.

Spector, N.H., Dolina, S., Cornélissen, G., Halberg, F., Markovic, B.M. & Jankovic, B.D. (1995). Neuroimmunomodulation: neuroimmune interactions with the environment, in *Handbook of Physiology, Section 4: Environmental Physiology*, M.J. Fregly & C.M. Blatteis, eds; American Physiological Society/Oxford University Press, New York, pp. 1537–1550.

Touitou, Y. & Haus, E. (1992). *Biological Rhythms in Clinical and Laboratory Medicine*. Springer-Verlag, Berlin, p. 730.

Zeman, M., Cornélissen, G., Balazova, K., Jozsa, R., Olah, A., Nagy, G., Csernus, V., Kaszaki, J., Pan, W.H., Bubenik, G. & Halberg, F. (2004). Circadian rhythm of melatonin in rat duodenum, in *Proceedings, Symposium: Chronobiology in Medicine. Dedicated to the 85th Anniversary of Professor Franz Halberg*, G. Cornélissen, T. Kenner, B. Fiser, J. Siegelova, eds; Masaryk University, Brno, pp. 95–97.

GERMAINE CORNÉLISSEN & FRANZ HALBERG

Circadian Variation

Most living organisms on earth experience an annual seasonal cycle caused by the earth's revolution around the sun, and a *circadian* (Latin *circa dies*: approximately one day) day/night cycle, caused by the earth's rotation around its own axis. Other rhythms, defined by cycle length, are as shown in Table 1.

Circadian variations may arise directly from the effects of the varying levels of electromagnetic radiation from the sun at different times of the day. In addition, many living organisms have evolved internally generated rhythms that do not depend entirely on external stimuli. The *endogenous circadian rhythms* in physiological functions such as body temperature, blood pressure, and mental alertness are responsible for the normal sleep/wake cycle and for jetlag. The eventual adaptation to local time by travelers indicates that the "body clock" is sensitive by external cues known as *synchronizers, entrainers* or *Zeitgeber*. The brain centre responsible for the generation and synchronization of circadian rhythms is believed to be the *suprachiasmatic nuclei* in the anterior hypothalamus, which receives input from retinal neurones, and which regulates the production of the hormone *melatonin* by the *pineal gland*. The endogenous periodic oscillations of the suprachiasmatic nuclei may be due to certain "clock genes" which exhibit rhythmic transcription by negative feedback control [3]. Understanding the "body clock" has important implications for air travel, shift work, as well as mental disorders such as mania and depression in which sleep disturbance is an important feature.

The rhythm of a variable is characterized by its *period* (time units per cycle) or *frequency* (the number of cycles per unit time), *acrophase* and *nadir* (the times at the maximum and minimum values of the variable, respectively), *amplitude* (half the difference between the maximum and the minimum values of the variable), and the *mesor* (the mean value of the variable over the cycle). The aim of statistical analysis in circadian rhythm research is usually to estimate these parameters in a population for the variables of interest, or to test for differences between these parameters in two or more populations. The data usually consist of measurements of subjects several times a day over one or more days.

Table 1 Rhythms

Rhythm	Length of period
Ultradian	Less than 20 h
Circadian	Between 20 and 28 h
Infradian	More than 28 h
Circaseptan	About seven days
Circatrigintan	About one month
Circaannual	About one year

In the simplest design, a group of subjects is randomized into two or more subgroups, and each subgroup is measured at a different time of the same day. For normally distributed variables, an **analysis of variance** (ANOVA) can be performed with "time of day" as a main effect, with or without other main effects (e.g. sex, disease, medication) or covariates (e.g. age), or interactions (e.g. time of day by medication). For discrete variables, **chi-square tests** or the **loglinear model** may be used. The advantage of this design is that it avoids practice effects, which may artifactually inflate the value of the variable at the second time point.

Another design that also avoids this problem is the *rolling Latin square* (*see* **Latin Square Designs**), which aims to subject the different time points to the same practice effect. An example of this design, for three time points, is shown in Table 2, where T represents a testing session.

The resulting data can be analyzed using simple ANOVA, with time of day and order (first, second or third) as main effects. This, however, ignores individual effects, and it is desirable to analyze such data with more sophisticated methods such as mixed effects models for longitudinal data, **multilevel models**, or *repeated measures models*. If the data are complicated by missing data or irregular time intervals, then the use of repeated measures models is even more desirable.

Table 2 Rolling Latin square design

	Day 1			Day 2		
	0600	0800	2200	0600	0800	2200
Group 1	T	T	T			
Group 2		T	T	T		
Group 3			T	T	T	

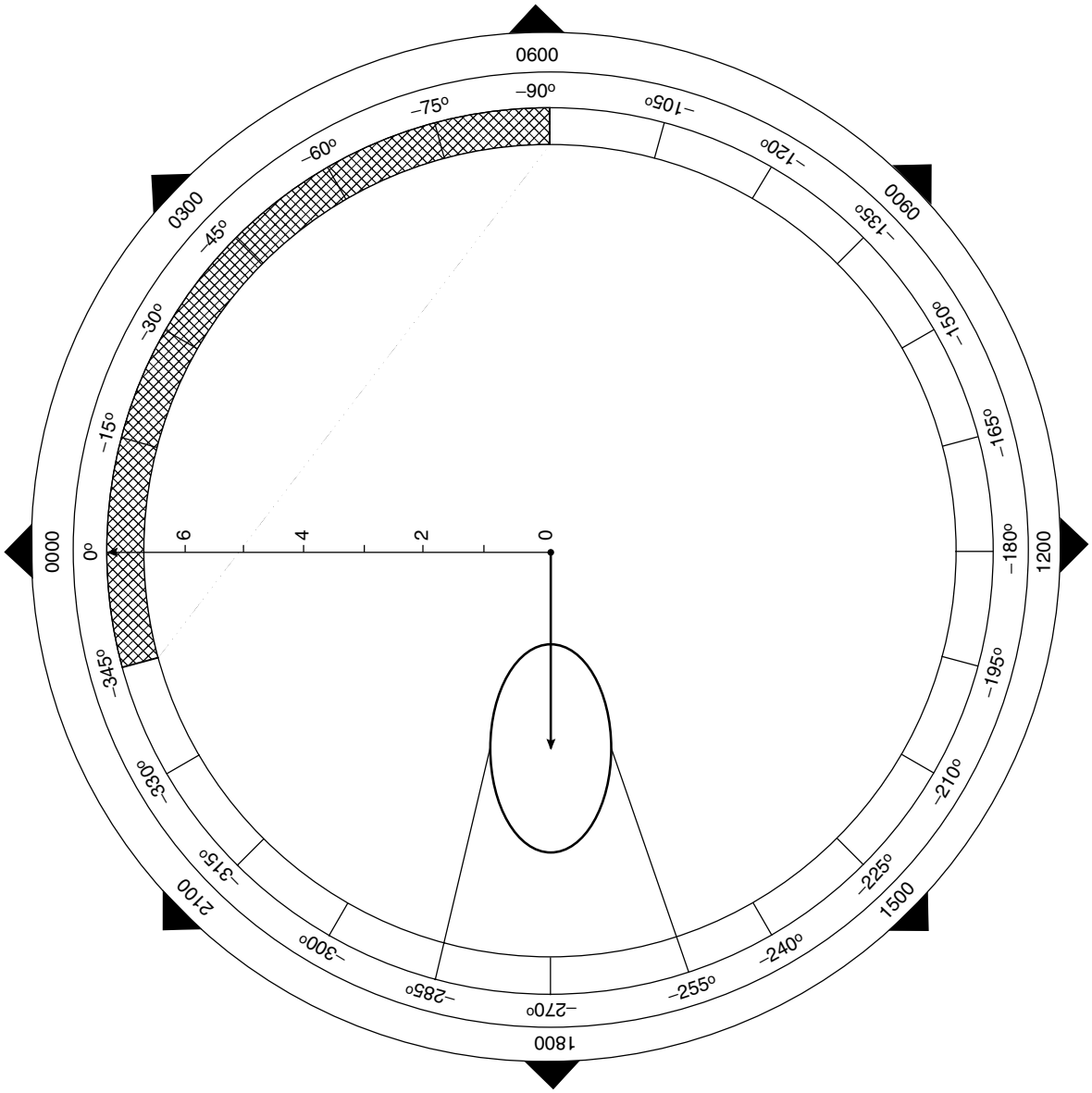

Figure 1 A cosinor display

In some studies a single subject is measured over several days and the resulting data constitute a single *time series*. Standard methods of time series analysis, such as the *correlogram* and the *spectrogram*, can then be applied. In addition, circadian rhythm researchers often use simple displays such as *Aschoff bars* and *Buys-Ballot Tables* [2].

A *sinusoid* with a period of 24 h is often fitted to data in order to estimate the amplitude and phase of the rhythm. As in *linear regression* (*see* **Linear Regression, Simple**), the significance of the fit can be tested by *F statistics*, and the *goodness of fit* between the sinusoid and the data measured by R^2, the *proportion of variance explained*. One popular extension of this approach is the *Minnesota cosinor technique* [1], which combines time series data from a number of subjects to give an overall summary of the circadian variation. The method involves estimating the phase (ϕ) and amplitude (α) of each subject, transforming these estimates to a Cartesian system ($x = \alpha \cos \phi$, $y = \alpha \sin \phi$), averaging x and y across subjects, and then back-transforming the averages to give estimates for ϕ and α of the sample. The circular confidence region for x and y gives rise to an elliptic confidence region for ϕ and α. A circadian rhythm is significant if the confidence region for ϕ and α does not include the origin (i.e. $\alpha = 0$). The result of such an analysis is usually presented in a *cosinor display*, which is a plot of ϕ and α and their confidence region in polar coordinates with time (i.e. phase) represented clockwise from 00 : 00 (north) through 06 : 00 (east), 12 : 00 (south), 18 : 00 (west) and back to 00 : 00. An illustrative example of a cosinor display is shown in Figure 1.

References

[1] Halberg, F., Tong, Y.L. & Johnson, E.A. (1967). Circadian system phase – an aspect of temporal morphology; procedures and illustrative examples, in *The Cellular Aspects of Biorhythms*, H. Von Mayerback, ed. Springer-Verlag, Berlin, pp. 20–48.

[2] Monk, T.H. (1984). Research methods of chronobiology, in *Biological Rhythms, Sleep, and Performance*, W.B. Webb, ed. Wiley, Chichester, pp. 27–57.

[3] Sassone-Corsi, P. (1994). Rhythmic transcription and autoregulatory loops: winding up the biological clock, *Cell* **78**, 261–364.

(*See also* **Chronomedicine**; **Spectral Analysis**; **Structural Time Series Models**; **Time Series**)

PAK SHAM

Circular Data Models

Circular data are data measured in the form of angles or two-dimensional orientations, and so can be represented as points on the circumference of a unit circle, or as (endpoints of) diameters of a unit circle.

Statistical models for circular data have found a particularly felicitous area of application in biology, not so much because of the range of models required (indeed, the converse is true) but because of the wealth of fascinating experiments the analysis of which requires the models and methods.

Typical of the problems of interest to biological scientists are those of bird navigation and of general orientations selected by particular creatures in response to experimental variation of their natural habitat (or of parts of themselves).

For the specific areas of medicine and the health sciences, there has been rather less use of circular models. (One of the more common applications has been to the study of **circadian variation**.) Nevertheless, there remains scope for their application. Subsequent sections of this article will review material which should be of most immediate value: modeling a single sample of circular data; **association** and **regression** involving a circular random variable; and **time series** models for circular data. The reader is referred to [4] for use of these models in statistical analysis.

Statistical Models for a Circular Population

In this section we introduce the most useful probability models for samples of circular data. However, some important points need to be made before this is done:

1. Probability distribution and density functions for *circular* data are defined differently from those for *real* (i.e. *linear*) data, with corresponding differences for the types of moments used. This issue is elaborated in the next subsection.
2. Circular data arise commonly in one of two forms:
   (i) *Vectorial* data, or data with a *sense of direction*, which can be represented as unit vectors, or as points on the circumference of a circle of unit radius; for example, arrival times (on a 24-hour clock) at out-patient clinics at a hospital.
   (ii) *Axial* data, or *undirected* data, which can be represented as oriented lines of unit length, or as (the two endpoints of) diameters of a unit circle; for example, the orientation of the principal axis of a blood cell on a plate being viewed under a microscope. If Θ is an axial random variable, then the values Θ and $\Theta + \pi$ are indistinguishable. The transformed variate $\Theta' = 2\Theta[\text{mod } 2\pi]$ is generally used for the purposes of modeling and analysis. Exceptionally, p-axial data are encountered; that is, data for which the values
   $$\Theta, \Theta + \frac{2\pi}{p}, \Theta + 2\frac{2\pi}{p}, \ldots, \Theta + (p-1)\frac{2\pi}{p}$$
   are indistinguishable. In this case the working variate becomes $\Theta' = p\Theta[\text{mod } 2\pi]$. Thus, special probability models are not required for these cases.
3. This discussion treats only univariate models: no tractable models for *multivariate* circular data have been developed, even for a case as simple as bivariate circular data. This is partly because, with the exception of the von Mises distribution, the most useful models for univariate unimodal circular data are not members of the **exponential family**, and in the case of the von Mises distribution, Mardia [7] has shown that the only bivariate distribution of exponential form with von Mises marginals is just the product of the marginal distributions.

Circular Probability Density Functions, Distribution Functions, and Trigonometric Moments

The *probability density function* (pdf) $f(\theta)$ of a continuous circular random variable Θ is a nonnegative continuous periodic function such that, for all θ,

$$f(\theta) = f(\theta + 2\pi)$$

and

$$\int_{\theta}^{\theta+2\pi} f(\phi)\, \mathrm{d}\phi = 1.$$

The *distribution function* $F(\theta)$ corresponding to $f(\theta)$ can be defined over any interval (θ_1, θ_2) by

$$F(\theta_2) - F(\theta_1) = \int_{\theta_1}^{\theta_2} f(\theta)\, \mathrm{d}\theta.$$

Corresponding to ordinary moments on the real line are trigonometric moments for Θ. The pth trigonometric moment is given by

$$\begin{aligned}\mu'_p &\equiv \rho_p \exp(\mathrm{i}\mu'_p)\\ &\equiv \alpha'_p + \mathrm{i}\beta'_p,\end{aligned}$$

where α'_p and β'_p are the pth cosine and sine moments, respectively. When $p = 1$, we simply write ρ for ρ_1, μ for μ_1; that is,

$$\mu'_1 = \rho \exp(\mathrm{i}\mu),$$

where μ is the *mean direction* and ρ the *mean resultant length*, $0 \le \rho \le 1$. If $\rho = 1$, the distribution is concentrated in a single direction (μ). However, $\rho = 0$ does *not* imply that the distribution is uniform: for example, a distribution with pdf such that $f(\theta) = f(\theta + \pi)$ will have $\rho = 0$.

The *central trigonometric moments* of Θ are obtained computed relative to the population mean direction:

$$\begin{aligned}\mu_p &\equiv \rho_p \exp(\mathrm{i}\mu_p)\\ &= \int_0^{2\pi} \exp[\mathrm{i}p(\theta - \mu)] f(\theta)\, \mathrm{d}\theta\\ &\equiv \alpha_p + \mathrm{i}\beta_p.\end{aligned}$$

When $p = 1$, we obtain $\alpha_1 = \rho$ and $\beta_1 = 0$.

Some functions of the first and second trigonometric moments are also of value. The *circular variance* of Θ is defined by

$$v = 1 - \rho, \quad 0 \le v \le 1.$$

The *circular standard deviation* is defined *not* as $\sqrt{v}$ but as

$$\sigma = \{-2\log(1 - v)\}^{1/2} \equiv \{-2\log\rho\}^{1/2}.$$

For v small (i.e. ρ near 1),

$$\sigma \simeq (2v)^{1/2} \qquad \text{or} \qquad \{2(1 - \rho)\}^{1/2},$$

the error in the approximation being less than 5% for $v < 0.18$: equivalently, $\rho > 0.82$. An associated measure of spread is the *circular dispersion*

$$\delta = \frac{(1 - \rho_2)}{2\rho^2},$$

the sample counterpart of which plays an important role in large-sample statistical inference for the mean direction.

Definitions of other quantities, such as measures of skewness and kurtosis, median and modal directions, and other measures of spread, are also available [4, Section 3.3].

Probability Distributions on the Circle: the Uniform Distribution

The **uniform distribution** plays a rather more important role in the analysis of circular data than for linear, as it provides the null model against which alternatives – unimodal or multimodal – are assessed. The pdf, df, and moments for this distribution and the other distributions described in this section are provided in Table 1.

Probability Distributions on the Circle: Unimodal Distributions

Here, we provide definitions and basic properties of the most useful probability models. The reader is referred to [6], [8], and [4] for more detailed discussion, methods of data analysis, and references to related work.

Wrapped Models

Let X be a random variable on the real line, with pdf $g(x)$ and df $G(x)$, and define $\Theta \equiv X[\mathrm{mod}\ 2\pi]$. Then Θ has a *wrapped* distribution on the circle, with pdf and df given, respectively, by

$$f(\theta) = \sum_{k=-\infty}^{\infty} g(\theta + 2k\pi)$$

and

$$F(\theta) = \sum_{k=-\infty}^{\infty} [G(\theta + 2k\pi) - G(2k\pi)].$$

Details about the properties of wrapped distributions can be found in [6, Section 3.4.8].

Two particular unimodal distributions obtained in this way, the wrapped Cauchy and the wrapped normal distributions, have found useful application in circular statistics. These are two-parameter symmetric unimodal distributions; their density and distribution functions and moments are shown in Table 1. Both have the uniform distribution as one limiting form, as the mean resultant length goes to zero. The wrapped

Table 1 Probability density functions, distribution functions, and central trigonometric moments for some standard circular distributions. $I_p(\kappa)$ denotes the modified Bessel function of order p. Note that $\beta_p = 0$, $p > 1$, as the distributions are symmetric

Distribution	Probability density function, $f(\theta)$	Distribution function, $F(\theta)$	μ	ρ	α_p, $p > 1$
Uniform	$1/2\pi$	$\theta/2\pi$	Un-defined	0	0
Wrapped Cauchy	$\frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}$, $0 \le \rho \le 1$	$\frac{1}{2\pi}\cos^{-1}\left[\frac{(1+\rho^2)\cos(\theta-\mu)-2\rho}{1+\rho^2-2\rho\cos(\theta-\mu)}\right]$	μ	ρ	ρ^p
Wrapped normal	$\frac{1}{2\pi}\left[1+\sum_{k=-\infty}^{\infty}\rho^{k^2}\cos k(\theta-\mu)\right]$, $0 \le \rho \le 1$	(Integral of $f(\theta)$)	μ	ρ	ρ^{ρ^2}
von Mises	$\frac{1}{2\pi I_0(\kappa)}\exp[\kappa\cos(\theta-\mu)]$, $\kappa \ge 0$	(Integral of $f(\theta)$)	μ	$\frac{I_1(\kappa)}{I_0(\kappa)}$	$\frac{I_p(\kappa)}{I_0(\kappa)}$

normal distribution behaves effectively as a normal distribution as ρ approaches unity (so that the effect of wrapping is negligible).

The wrapped Cauchy has found indirect use in algorithms for simulating data from von Mises distributions. The numerous desirable properties enjoyed by the normal distribution on the line are, on the circle, shared by the wrapped normal and von Mises distributions. For suitably chosen values of their dispersion parameters, all three distributions can be made very close to each other, so that samples from each are indistinguishable in practice, except with large data sets. As a consequence, one uses whatever is most convenient to the problem in hand.

The von Mises Distribution

Notwithstanding the preceding remarks about the closeness of the wrapped normal, wrapped Cauchy, and von Mises distributions, it is the von Mises distribution that enjoys many of the useful *inferential* properties possessed by the normal distribution on the line. As such, it is the most common model for a sample of unimodal circular data. As $\kappa \to 0$, the distribution tends to the uniform. The size of the *concentration* parameter κ is crucial for application of many parametric procedures. For $\kappa < 2$, corresponding to relatively dispersed distributions, there is significant density at the antimode (i.e. $f(\mu + \pi) >> 0$), and many of these procedures require (data-dependent) adjustment to work satisfactorily. As $\kappa \to \infty$, the von Mises distribution tends to the normal with variance $1/\kappa$.

Other Probability Models for Circular Data

Few other models for univariate circular data have found practical application. Batschelet [2] and Mardia [6] have described asymmetric models and models for discrete data, and some application has been found for mixture models, mainly mixtures of von Mises distributions.

Similarly, little is available by way of multivariate distributions involving circular random variables. Fisher [4] provides references to bivariate models in the literature.

Association and Regression Involving a Circular Random Variable

Whereas, with linear data, the topics of **association** and **regression** are sometimes presented independently of each other, this is not possible with circular data. At least for relationships involving both linear and circular variates, we need first to look one step ahead: Is the purpose of the proposed analysis to be able to predict the mean direction of Θ for a given value x of a linear variate X, or to predict the mean of X given that $\Theta = \theta$? Without answering this question it is not possible to assess the association or correlation between Θ and X. More details relating to this area can be found in [4, Chapter 6].

Some work has been done on the analysis of experiments with a circular response: see [1] for a survey of the literature.

Association Between Two Circular Random Variables

The joint distribution of two circular random variables Θ and Φ is concentrated on the surface of a torus. A simple analogue of complete linear association between real variates X and Y is so-called *T-linear association:*

$$\Theta = \Phi + \theta_0 [\text{mod } 2\pi] \qquad \text{(positive association)}$$

or

$$\Theta = -\Phi + \theta_0 [\text{mod } 2\pi] \qquad \text{(negative association).}$$

The extent to which Θ and Φ are so correlated can be measured by a simple analog of the linear correlation coefficient, namely

$$\rho_T = \frac{\mathrm{E}[\sin(\Theta_1 - \Theta_2)\sin(\Phi_1 - \Phi_2)]}{\{\mathrm{E}[\sin^2(\Theta_1 - \Theta_2)\sin^2(\Phi_1 - \Phi_2)]\}^{1/2}}$$

for variate pairs (Θ_1, Φ_1) and (Θ_2, Φ_2) distributed independently as (Θ_1, Φ_1): compare with the alternative representation of Pearson's product moment correlation coefficient as

$$\frac{\mathrm{E}[(X_1 - X_2)(Y_1 - Y_2)]}{\mathrm{E}[(X_1 - X_2)^2(Y_1 - Y_2)^2]^{1/2}}$$

(*see* **Correlation**).

Analogously to a monotone relation between X and Y, we can define a concept of complete *T-association* as one in which, when Θ moves clockwise (counterclockwise) so does Φ, with the opposite happening for negative association. Whereas, for linear variates, Kendall's τ provides a simple measure of monotone association based on two independent pairs (X_1, Y_1) and (X_2, Y_2) (*see* **Rank Correlation**), three independent pairs are required to define an analogous measure for circular variates.

Association between a Circular rv and a Linear rv

Let Θ and X be circular and linear random variables respectively, the association of which we seek to assess.

Linear–Circular Association. The simplest form of model for the conditional mean of X given Θ is provided by the periodic relationship

$$\begin{aligned} \mathrm{E}(X|\Theta = \theta) &= \alpha_0 + \beta_0 \cos(\theta - \theta_0) \\ &\equiv \alpha_0 + \alpha_1 \sin\theta + \alpha_2 \cos\theta. \end{aligned}$$

It is termed *C-linear* association. In this form, it is clear that an appropriate measure of association between X and Θ is a measure of multiple correlation between X and $(\sin\Theta, \cos\Theta)$.

By analogy with a monotone relationship between two linear random variables, C-linear association can be generalized to *C-association*:

$$\mathrm{E}(X|\Theta = \theta) = \alpha_0 + \beta_0 f[\cos(\theta - \theta_0)],$$

where $f(\cdot)$ is a monotone function of its argument. The extent of C-association can then be gauged by an analogue of Kendall's τ.

Circular–Linear Association. In the same way that modeling linear–circular association leads, at least for prediction purposes, to models relating linear variates, so the modeling of circular–linear association leads to models relating circular variates. The *circular–linear* association between Θ and X can be computed as the *circular–circular* association between Θ and $\Phi = 2\tan^{-1}(X)$.

Model for Linear–Circular Regression. From the previous discussion, it can be seen that this reduces to a standard multiple regression model.

Model for Regression with Circular Response. To model Θ as a function of a vector $\mathbf{x}$ of linear explanatory variables, suppose that the mean direction μ of Θ is related to $\mathbf{x}$ by the equation

$$\mu = \mu_0 + g(\boldsymbol{\beta}'\mathbf{x}),$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $g(\cdot)$ is a (monotone) *link function*, which maps the real line into the circle. An example of such a link function would be $g(u) = 2\tan^{-1}(u)$. The model can be fitted quite generally, using the directional analog of least squares; however, if Θ is modeled as a von Mises variate, likelihood methods become applicable.

For suggested approaches to modeling the mean direction of Θ conditional on a circular explanatory variable ϕ, see the discussion and references in [4, Section 6.4.5].

Time Series Models for Circular Data

As noted in the discussion of probability models, there are no satisfactory models for correlated circular data, as a consequence of which modeling time series of circular data poses rather more problems than modeling time series of linear data. It is helpful to distinguish series with moderate noise from very noisy series, so two approaches are described.

Each approach utilizes link functions (cf. the preceding discussion of angular regression). Let $g(x) : (-\infty, \infty) \to (-\pi, \pi)$ be an increasing function of x such that $g(0) = 0$ and $g(-x) = -g(x)$. (One such example is the arc tan link function suggested for regression.)

For such link functions, if X is a linear random variable, then $\Theta = g(X)$ is a circular random variable; and conversely, $g^{-1}(\cdot)$ transforms a circular variate to a linear one. Now define $\{\theta_t\}$ to be a *linked ARMA(p,q) process* (*see* **ARMA and ARIMA Models**), or a *LARMA(p,q) process,* if its linked linear process $\{g^{-1}(\theta_t)\}$ is an ARMA(p,q) process.

We need to differentiate dispersed from concentrated situations. Broadly speaking, a unimodal circular variate is dispersed if its density assigns non-negligible mass to all parts of the interval $(-\pi, \pi)$: for a von Mises variate, this corresponds to a distribution with $\kappa < 2$. A distribution concentrated on a subset of $(-\pi, \pi)$ provides the opportunity to use

approximate methods based on linear-variate theory. For concentrated time series, complete LARMA(p,q) processes can be fitted. For noisier series, the only methods available at present relate to fitting autoregressive models.

For more detailed discussion of these and other methods, see [3]–[5].

References

[1] Anderson, C.M. & Wu, C.F.J. (1995). Measuring location effects from factorial experiments with a directional response, *International Statistical Review* **63**, 345–363.

[2] Batschelet, E. (1981). *Circular Statistics in Biology*. Academic Press, London.

[3] Breckling, J. (1989). *The Analysis of Directional Time Series: Application to Wind Speed and Direction*. Springer-Verlag, Berlin.

[4] Fisher, N.I. (1995). *Statistical Analysis of Circular Data*, 1st Paperback Ed. Cambridge University Press, Cambridge.

[5] Fisher, N.I. & Lee, A.J. (1994). Time series analysis of circular data, *Journal of the Royal Statistical Society, Series B* **56**, 327–339.

[6] Mardia, K.V. (1972). *Statistics of Directional Data.* Academic Press, London.

[7] Mardia, K.V. (1975). Statistics of directional data, *Journal of the Royal Statistical Society, Series B* **37**, 349–393.

[8] Watson, G.S. (1983). *Statistics on Spheres*. Wiley, New York.

N.I. FISHER

Cladistic Analysis

In **association** studies, when a set of linked **markers** is available an analysis based on haplotypes is more powerful than one using only a single marker. **Haplotype analysis** does not require the discovery of every disease variant, but does assume that **linkage disequilibrium** exists between the markers and the disease variants. Moreover, the number of haplotypes can be very large when many markers are typed, thereby reducing the power to test for association. The critical challenge is to maintain precision without losing power by combining the information from different haplotypes. Templeton et al. [14–18] introduced cladistic analysis which incorporates information on the inferred evolutionary relationships of the sample haplotypes to identify disease variants. The central assumption behind this approach is that an unknown **mutation** causing a phenotypic effect occurred at some point in the evolutionary history of the population and became embedded within the historic structure represented by the cladogram. In other words, certain portions of the cladogram would display phenotypic effects different from the other portions, depending on the mutations that are shared. Thus, the cladogram defines a nested analysis that is efficient for detecting associations between measured genotypic variation and phenotypic variation at the population level. Cladistic analysis consists of two steps: constructing a cladogram of the haplotypes that reflects the evolutionary relationships; and conducting a nested analysis.

Constructing a Cladogram

Assuming we know the individual haplotypes, an evolutionary tree can be constructed by the method of maximum parsimony used for phylogeny reconstruction, as implemented in the computer program PAUP [11] or PHYLIP (software by J. Felsenstein, Version 3.57). The parsimony algorithm determines the unrooted tree that connects the observed haplotypes using the minimum number of mutations. The method of parsimony is robust and effective for reconstructing evolutionary relationships, although it does not use the frequencies of the haplotypes in the sample [12]. For example, Figure 1 shows a cladogram of 25 haplotypes found in the 13-kb alcohol dehydrogenase (ADH) (*see* **Gene**) locus in the fruit fly *Drosophila melanogaster* constructed using PHYLIP [15]. Each branch represents a single mutational change between

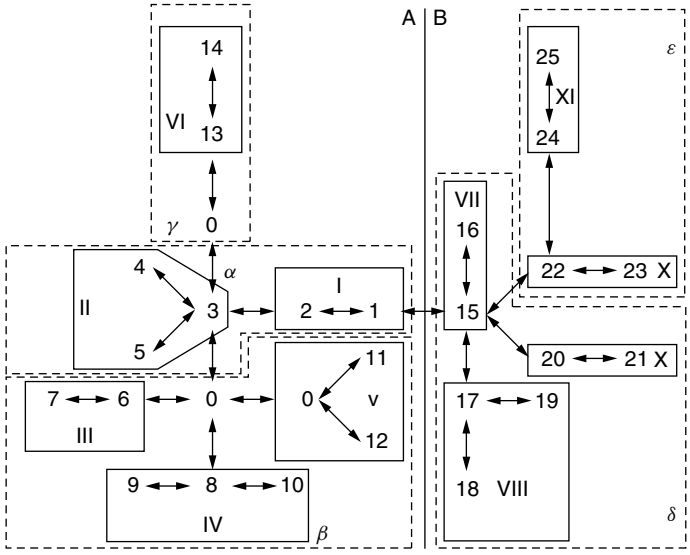

Figure 1 The cladogram defined by the nesting algorithm. The haplotypes enclosed by solid lines indicate the 1-step clades and are represented by Roman numerals. The dashed lines enclose the 2-step clades that are designated by Greek letters. A thick, solid line indicates the partitioning of the cladogram into two 3-step clades, designated by Roman letters. From Templeton et al. [1], reproduced with permission of the Genetics Society of America

haplotypes. Zeros refer to the inferred intermediate haplotypes that are not found in the sample but are necessary to connect the existing haplotypes. To summarize the evolutionary information Templeton et al. [15] developed a nesting algorithm to classify the haplotypes within clades. The nesting algorithm is as follows:

1. All the haplotypes in the cladogram are considered as 0-step clades.
2. Assume the n-step clades have been formed. Define the n-step clades with only one arrow pointing at them in the cladogram as "terminal" n-step clades, otherwise, as "internal" n-step clades.
3. Define the $n+1$ step clades as sets of all n-step clades that can be joined together by moving back one mutational step from the terminal n-step clades. After this operation, any remaining internal n-step clades that have not been incorporated into an $n+1$ step clade require further attention to be correctly classified. First, identify those remaining n-steps that are adjacent to (i.e. separated by one mutational step away from) an $n+1$-step clade that was defined in the initial operation. These n-step clades are then regarded as "terminal", and the operation is repeated. This procedure is iterated as needed until all n-step clades are members of an $n+1$-step clade.
4. Repeat step 3 until the entire cladogram can be united into a single category.

For example, haplotypes 4, 5, 7, 9, 10, 11, 12, 14, 16, 18, 19, 21, 23 and 25 in Figure 1 represent terminal 0-step clades and the remaining haplotypes are internal 0-step clades. Using the nesting algorithm, 1-step clades can be created as in Figure 1, represented by Roman numerals. To construct the 1-step clade II, for instance, the terminal 0-step clade haplotype 4 is found to be distinguished from haplotype 3 by one mutation. Similarly, haplotype 5 has one mutational difference with haplotype 3. Hence, haplotypes 3, 4 and 5 are joined into a 1-step clade. Repeating this procedure for all the terminal haplotypes leads to internal haplotypes 1 and 2 not being placed within a 1-step clade (haplotype 2 is one mutational step from the 1-step clade II; haplotype 1 is one step from the 1-step clade VII). Hence, according to step 3, both haplotypes 1 and 2 are now regarded as "terminal" and can be joined together because they are one mutational step from each other. Similarly, the 2-step clades can be defined from the 1-step clades as indicated by Greek letters, and the same procedure is applied to generate the 3-step clades, indicated as A and B in Figure 1. On the basis of these sequential operations, the algorithm thus gives a nested design.

Cladistic analysis has been recently applied in humans to study the association of angiotensin-1 converting enzyme (ACE) level with variation in the ACE gene [7, 20]. In the study of 159 randomly sampled Afro-Caribbeans by Zhu et al. [20], seven **polymorphisms** were genotyped within the ACE gene and 28 haplotypes were inferred using Clark's algorithm [1]. The positions of the seven polymorphisms are given in Figure 2. Eight common haplotypes, accounting for 83% of the total variability, were used for cladistic analysis, the remaining haplotypes being combined as R [20]. An unrooted evolutionary tree, shown in Figure 3, was first constructed from the eight haplotypes by the principle of maximum parsimony incorporated using PHYLIP. The nesting algorithm was then used to group the haplotypes. In Figure 3, haplotype 4 (note that numbers appear below the haplotypes) was grouped in clade C and haplotype 3 was grouped in clade B. Strictly, clades B and C should be considered as 2-step clades according to the above algorithm of Templeton et al. [15].

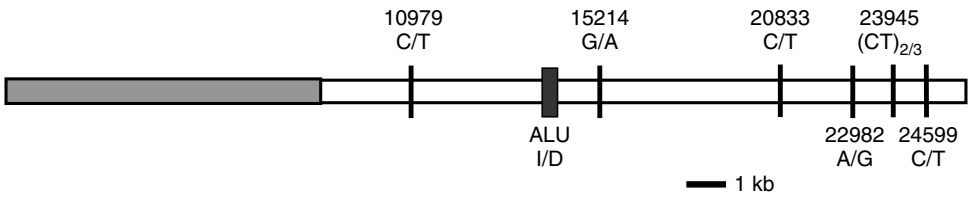

Figure 2 Polymorphic markers genotyped in ACE. From Zhu et al. [10], reproduced with permission of the University of Chicago Press

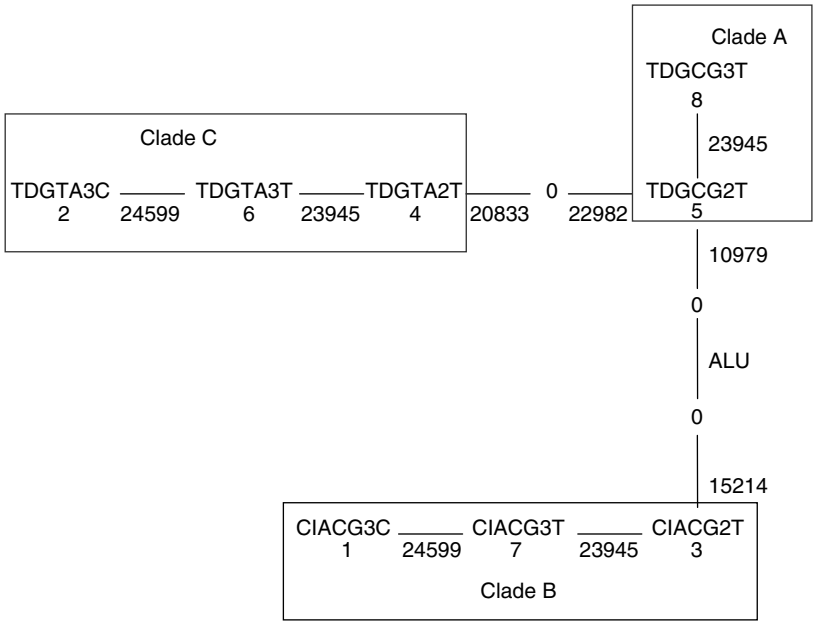

Figure 3 The eight most frequent haplotypes were used to infer a maximum parsimony tree. Haplotypes were then grouped with neighboring haplotypes into clades A, B and C. The maximum parsimony mutational connections among the haplotypes are indicated by solid lines, with the 0s representing all intermediate haplotypes that are missing from the sample. The number below the haplotypes corresponds to the haplotype, and the number below or beside the solid lines is the mutation site. The phenotypic effects attributed to these groups were tested under different model assumptions against plasma ACE activity using a measured haplotype analysis. From Zhu et al. [10], reproduced with permission of the University of Chicago Press

Nested Analysis

After defining a cladogram, there is the problem of how to carry out a nested analysis to estimate the phenotypic effects of specific haplotypes, realizing that we are dealing with the transmission of a phenotypic value from parent to offspring, and that parents pass on haplotypes, not **genotypes**, to their offspring. When homozygous genotypes are segregating in the population (i.e. each individual has two identical haplotypes), a NANOVA is an efficient method to detect and localize phenotypically important mutations for quantitative phenotypes. For example, Templeton et al. [15] performed a NANOVA of ADH level for the cladogram presented in Figure 1. The results of this NANOVA are given in Table 1. The most significant effect is associated with the transitional step between the 3-step clades A and B (Figure 1). There are also significant phenotypic effects found at the 1-step and 0-step levels. Since there are many 1-step and 0-step clades, further decomposition of variance is needed to localize the effects at these levels [15].

NANOVA will not work for populations whose members carry two different haplotypes, as in human populations. Consequently, assigning phenotypic effects to each haplotype is the core of the analysis. Fisher [5] developed two methods to measure the

Table 1 Nested analysis of variance of ADH activities in *D. melanogaster*. Reproduced from Templeton et al. [1] by permission of the Genetics Societies of America

Source	Sum of squares	Degrees of freedom	Mean square	F-statistic
3-Step clades	138.33	1	138.33	366.50**
2-Step clades	0.88	3	0.29	0.78
1-Step clades	12.93	6	2.16	5.71*
0-Step clades	19.14	14	1.37	3.62*
Error	6.04	16	0.38	

*Significant at the 1% level.
**Significant at the 0.1% level.

phenotypic effect of a haplotype: (a) the average excess of a haplotype, defined as the average phenotypic value of the haplotype minus the overall population mean; and (b) the average effect, defined as the least-squares regression coefficient from the linear relationship between the phenotype and the number of copies of each haplotype (0, 1 or 2) an individual possesses. Under the assumption of **Hardy–Weinberg equilibrium**, the average excess is equivalent to the average effect [13]. Templeton et al. [18] proposed a random permutation procedure using average excess to test the phenotypic associations for populations containing both heterozygous and homozygous genotypes. When this procedure was used to investigate the association between the ADH locus and ADH activity in *Drosophila melanogaster*, Templeton et al. found that the result of the random permutation procedure was consistent with that from NANOVA for this homozygous population. Zhu et al. [20] modeled the relationship between ACE plasma level, sex, age, and haplotype using a **linear regression** model for the cladogram in Figure 3. Thus, the least-squares estimates of the haplotype effects are their adjusted average effects. To localize the important mutation affecting ACE level variance, a series of model comparisons can be performed. On the basis of the cladogram in Figure 3, the hypothesis was first tested that the average effect on ACE level is the same for haplotypes within each clade. This comparison can be tested by the goodness-of-fit test between model [A, B, C, R] and the full model [1, 2, 3, 4, 5, 6, 7, 8, R], which assumes that each haplotype has a different average effect on the ACE level. Table 2 shows that this comparison is not statistically significant, indicating that the average effects are similar within each of the three clades A, B and C. Next the average effects on the ACE level of clades A, B and C were compared. Because of the evolutionary relationships revealed in Figure 3, only two such comparisons were necessary, namely the average ACE level between A and B and that between A and C. As shown in Table 2, no significantly different average effects were found between clades A and B, but a significant difference was found between clades A and C, after adjusting for **multiple comparisons**. One can see from the cladogram that there are two mutational differences between clades A and C, occurring at the connected positions 20 833 and 22 982, and three mutational differences between clades A and B. However, differences of this sort can also result from an ancestral recombination that occurred between sites 15 214 and 20 833. By calculating the linkage disequilibrium between each pair of polymorphisms and testing for recombination [2], Zhu et al. found that an ancestral recombination between 15 214 and 20 833 was in fact more likely than multiple mutations. Therefore the functional variant causing ACE variation is unlikely to be located in the segment between 10 979 and 15 214. Similar arguments also exclude the segment between 23 945 and 24 599 as the location of an ACE functional variant. The 9-kb interval between 15 214 and 23 945 therefore most likely bears the functional ACE variant.

Conclusion

Cladistic analysis is a potentially powerful method for analyzing association between a set of tightly linked markers and phenotypic variation, especially for **candidate gene** studies. Since cladistic analysis uses the historical relationships between haplotypes, the necessary multiple comparisons can be conducted objectively guided by the cladogram. Consequently, the number of multiple comparisons can be reduced and the power to detect association thus improved.

Table 2 Model comparisons

	P-value obtained against model [1, 2, 3, 4, 5, 6, 7, 8, R]			
Model	Overall	Male	Female	Overall by EM
A, D, C, R (2 = 6 = 4, 5 = 8, 1 = 7 = 3, R)	0.3809	0.1028	0.1028	0.2878
A = B, C, R	0.4344	0.1508	0.9770	0.3617
A = C, B, R	0.016	0.016	0.6244[a]	0.015

[a] *P* value is 0.0636 in the test against model [A, B, C, R]. From Zhu et al. [10], reproduced with permission of the University of Chicago Press.

The method has been successfully used to localize the deoxyribonucleic acid (DNA) region affecting the ACE level in different populations [7, 20]. In an analysis of simulated data at the Genetic Analysis Workshop 12, Zhu et al. [19] also localized the Q1 functional mutation to the correct region using cladistic analysis. Although cladistic analysis was developed for quantitative phenotypes, it can also be extended to **case–control studies** of qualitative traits. Templeton [14] studied the association between haplotypes and Alzheimer's disease by performing a series of nested 2 (case and control) $\times\, n(i)$ contingency table analyses, where $n(i)$ is the number of clades in the nested category i. If the sample is small, then a permutation chi-square test can be performed using the algorithm of Roff & Bentzen [8]. **Logistic regression** and a series of goodness-of-fit tests based on the **likelihood ratio** statistic can also be adapted for nested model comparisons when we are analyzing qualitative traits.

As with all statistical methods, cladistic analysis has its limitations. The power of the method is dependent on the fidelity with which the estimated cladogram reflects the true evolutionary relationships. Several factors, such as recombination and **gene conversion**, can affect the accuracy of this estimate. To keep these sources of error to a minimum, recombination and gene conversion should be relatively rare, and this is a reasonable assumption in a candidate gene association study. If this assumption is not valid, then analyses on subdivisions of the region may be necessary [17].

Cladistic analysis requires that each individual's haplotypes be known. When samples are random, the haplotypes can be inferred with a high degree of accuracy using Clark's [1] subtracting algorithm. For family data, SIMWALK2 developed by Sobel & Lange [9], which uses simulated annealing based on a Markov process, can be used. However, uncertainty of haplotype assignment has an impact on cladistic analysis. For random samples and quantitative phenotypes, Zhu et al. [21] proposed using a two-step approach to model the relationship between the phenotype and genetic markers: (a) estimate the haplotype frequencies using the **EM algorithm** – well-developed methods are available for this purpose [3,4,6,11]; and (b) use a mixture model approach to model the association between the trait and the haplotypes. Zhu et al. [21] applied this two-step approach to the Jamaican data set described above and obtained results consistent with that using Clark's algorithm (EM results are presented in the last column of Table 2).

In summary, cladistic analysis has emerged as an important tool for detecting associations between measured genetic and phenotypic variation at the population level, although some limitations must be anticipated when the assumptions are not met. Only further practical experience will define the full contribution this analytic method can make to **genetic epidemiology**.

References

[1] Clark, A.G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations, *Molecular Biology and Evolution* **7**, 111–122.

[2] Crandall, K.A. & Templeton, A.R. (1999). Statistical approaches to detecting recombination, in *The Evolution of HIV*, K.A. Crandall, ed. The Johns Hopkins University Press, Baltimore, pp. 153–176.

[3] Excoffier, L. & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, *Molecular Biology and Evolution* **12**, 921–927.

[4] Fallin, D. & Schork, N.J. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data, *American Journal of Human Genetics* **67**, 947–959.

[5] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[6] Hawley, M. & Kidd, K. (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes, *Journal of Heredity* **86**, 409–411.

[7] Keavney, B., McKenzie, C., Connell, J.M.C., Julier, C., Ratcliffe, P.J., Sobel, E., Lathrop, M. & Farrall, M. (1998). Measured haplotype analysis of the angiotensin-1 converting enzyme (ACE) gene, *Human Molecular Genetics* **7**, 1745–1751.

[8] Roff, D.A. & Bentzen, P. (1989). The statistical analysis of mitochondrial DNA polymorphisms: π^2 and the problem of small samples, *Molecular Biology and Evolution* **6**, 539–545.

[9] Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics, *American Journal of Human Genetics* **58**, 1323–1337.

[10] Stephens, M., Smith, N.J. & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from

population data, *American Journal of Human Genetics* **68**, 978–989.

[11] Swofford, D.L. (1998). *PAUP: Phylogenetic Analysis Using Parsimony, Release 4.01b1*. Sinauer Associates, Sunderland.

[12] Swofford, D.L., Olsen, G.J., Waddell, P.J. & Hillis, D.M. (1996). Phylogenetic inference, in *Molecular Systematics*, D.M. Hillis, C. Moritz & B.K. Mabel, eds, 2nd Ed. Sinauer Associates, Sunderland, pp. 407–514.

[13] Templeton, A.R. (1987). The general relationship between average effect and average excess, *Genetic Research* **49**, 69–70.

[14] Templeton, A.R. (1995). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus, *Genetics* **140**, 403–409.

[15] Templeton, A.R., Boerwinkle, E. & Sing, C.F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in drosophila, *Genetics* **117**, 343–351.

[16] Templeton, A.R., Crandall, K.A. & Sing, C.F. (1992). A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation, *Genetics* **132**, 619–633.

[17] Templeton, A.R. & Sing, C.F. (1993). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. IV. Nested analyses with cladogram uncertainty and recombination, *Genetics* **134**, 659–669.

[18] Templeton, A.R., Sing, C.F., Kessling, A. & Humphries, S. (1988). A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations, *Genetics* **120**, 1145–1154.

[19] Zhu, X., Cooper, R.S., Chen, G., Luke, A. & Elston, R.C. (2001). Localization of the Q1 mutation by cladistic analysis, *Genetic Epidemiology* **21**, S594–S599.

[20] Zhu, X., McKenzie, C., Forrester, T., Nickerson, D.A., Broeckel, U., Schunkert, H., Doering, A., Jacob, H.J., Cooper, R.S. & Rieder, R. (2000). Localization of a small genomic region associated with elevated ACE, *American Journal of Human Genetics* **67**, 1144–1153.

[21] Zhu, X., Zhao, H., Cooper, R.S. & Rieder, M.J. (2000). comparison of haplotype analysis using Clark's method and likelihood method, in *American Statistical Association, 2000, Proceedings of the Section on Biometrics*. American Statistical Association, Alexandria, pp. 163–166.

XIAOFENG ZHU, RICHARD S. COOPER &
ROBERT C. ELSTON

Classification, Overview

This article gives a general review of classification problems that occur in biometrics (*see* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**; **Discriminant Analysis, Linear**). Confusingly, the term *classification* has two complementary meanings in the statistical literature. Firstly, it is concerned with assigning a sample to one of a set of previously recognized classes, and secondly, it is concerned with the construction and description of the classes themselves.

In classical statistical writings, classification is concerned with assigning a name to a given sample based on the values it is observed to have on some set of variables; that is, it is concerned with identification. This usage is better described as discrimination; the theory of discriminant analysis is an important part of statistics [1, 8, 15, 22, 23, 27, 30]. Basically, discrimination is concerned with labeling the sample with some name: in botany, based on the observed features of a plant, the kind of flower is identified; in medicine, based on a set of symptoms and/or biochemical tests, the disease from which a patient is suffering is identified; in banking, based on a prospective customer's financial record, the bank needs to decide whether to issue a credit card. In these examples, the assignment classes are the species of plants, the diseases to which humans are susceptible, and the populations of credit-card holders and nonholders. In the latter case there are only two possibilities (to issue or not to issue a card), and commonly assignment is to one of two classes. In the medical example, except perhaps for general expert systems for medical diagnosis, physicians would normally already have narrowed things down to a very few possibilities – perhaps merely healthy and infected with some specific disease. With plants, we may already have decided on plant genus and merely wish to decide between the relatively few species within the genus; on the other hand, we may be totally at sea and first wish to decide between the genera. Thus, when discriminating, it is fundamental first to have recognized the set of classes of interest. The values of variables may or may not overlap the classes. An example of a variable which overlaps classes is given by blood pressure, which varies in the population and with two groups of patients – one healthy and one not. It is to be expected that some healthy patients will have as high blood pressures as some with heart disease. Nonoverlapping variables are usually **categorical**, such as color, which completely separates dandelions, which are yellow, from daisies, which are white. Broadly speaking, when classes are close, with *overlap*, probabilistic methods have to be used, but when classes are distinct, assignment can be satisfactorily accomplished by using nonprobabilistic methods. When probabilistic assignment is appropriate, then it is implied that there is some probability of getting things wrong, and this encourages the development of methodology that is concerned with minimizing the probability of incorrect classification. From the probabilistic point of view, nonprobabilistic assignment is trivial, for with no overlap there is no possibility of being wrong. Nevertheless, as shown below, interesting problems arise.

All the above is concerned with assignment to classes where, as we have seen, it is assumed that relevant classes are recognized at the outset. Often, these known classes will have been described from substantive research in the field of application. Thus, medical researchers describe diseases, botanists describe plant species, and bankers know that some customers are good risks and others are bad risks. In contrast to defining classes on the basis of substantive knowledge, classes may be formed by using statistical methods. This is the second kind of problem, referred to above, where one is given an unstructured collection of what we shall provisionally term *objects*, which are not differentiated in any way. The question arises as to whether these objects might with advantage be allocated to two or more classes. This is the problem of *constructing* classifications, and it occurs very widely. It has several variants, depending on (i) whether or not the variables used to describe the objects are **random variables**, and on (ii) whether the objects represent individual samples, as is usual in the statistical literature, or whether they themselves represent previously recognized classes which one wishes to group into fewer classes at a higher level. In the most simple cases, an object may represent a class with many identical members; this is typical of classes which represent well-separated biological populations; all buttercups are yellow and all cats have claws, so in these respects, describing one buttercup or cat describes them all. Alternatively, an object may represent an entire statistical population with well-defined distributional form, very

likely a formalization of a distinct biological population, in which case replication within the population is formally possible and replicate within-population individual samples may be available. In both these cases, we shall say that the objects are *structured*, that is to say that every given object has been assigned to one of K, say, population classes. In the first case each population is represented by a single object, it being assumed that there is no variation within any population with respect to the variables with which we are concerned, or at least that variation is unimportant. With this assumption we are nearly always concerned with qualitative/categorical variables because it is inconceivable that quantitative variables do not vary in populations. In the second case, there are several, n_k, sample replicates within the kth population, and we shall be concerned mainly with quantitative variables.

In another variant of the second kind of problem, that of constructing classifications, rather than differentiated objects we may have N *undifferentiated* objects. This would be the case if we had a random sample of size N drawn from a single population. However, in the context of classification, it is common to have what looks like a random sample of size N, but we may have a strong suspicion that these N cases might be classified with advantage into a few classes. That is, we are not given any form of classification at the outset, so assignment is not relevant, but nevertheless we may wish to construct a classification. We may hypothesize that the samples come from a mixture of K populations with some specified parametric forms, and wish to decide on the number of components in the mixture and to estimate the unknown parameters. Additionally, we may indeed assign each sample to one of the newly identified component populations.

Using the terminology developed above, Table 1 lists the kinds of classification problem discussed in the remainder of this article.

Discrimination – Assignment to Classes

We first discuss the problem of assigning samples, cases, and objects to preassigned classes. As Table 1 shows, all assignment problems are concerned with structured objects. These problems may be considered either in a probabilistic framework or not. We first consider nonprobabilistic assignment.

Nonprobabilistic Assignment

The most simple method of identification, that is of assigning a name to a specimen, is to compare the properties of the specimen with the entries in a table listing the properties of every named class of relevance. Such a table is called a *diagnostic table*. It is not necessary to list all properties – only some minimal subset that is sufficient for identification purposes. Perhaps more than one diagnostic table may be needed, each listing different subsets of properties. An example would be with the identification of plants where one table might be applicable to plants in their flowering state, another to their seeds and yet another to cover the vegetative state. Diagnostic tables are practicable only when deciding among a few putative classes, and more efficient methods, such as *diagnostic keys*, have to be sought for larger problems.

Diagnostic keys have been in use for several centuries to help identify plant specimens. Nearly every flora contains a key which helps botanists name a specimen plant. More recently, the idea of a diagnostic key has found wider applications either directly, as in diagnoses based on biochemical tests,

Table 1 The main types of classification problem discussed in this article

	Objects	Assignment	Construction
Nonprobabilistic	Structured	Matching	Maximal predictive classes
		Diagnostic keys and tables	Cluster analysis: k groups
			hierarchical
			others
Probabilistic	Unstructured	(Nothing to assign to)	Mixture problems
	Structured	Discriminant analysis	Not developed (but could be)

or conceptually, as in CART (*see* **Tree-structured Statistical Methods**), the more simple expert systems (*see* **Artificial Intelligence**), or as **neural networks**. The simple botanical key operates by requiring the botanist to decide whether the specimen they wish to identify has or does not have some feature. Table 2 shows the beginnings of a key to the genera of North American orchids.

This key shows several features of interest. First, we see that the key has a hierarchical structure, with the nodes numbered on the left-hand side. In its entirety, this particular key has 55 nodes but only the first four are shown; further nodes must be added to distinguish between the various species within each genus, thus allowing the specimen to be named. On the right-hand side, the response to each question shown leads either to another node, and another question, or to an identification given in italics with a page reference where further information may be found [31]. The tree is binary with two possible outcomes at each node. Nodes need not be **binary**; indeed, node 26 of this key has three outcomes:

26: Lip directed upward; spur absent	27
Lip directed downward; spur absent	29
Lip directed downward; spur present	30

Clearly, this node could be rearranged as two binary nodes: "Lip directed upward?" and "spur absent?". Most methodologic work on keys is concerned with binary keys. A "don't know" response can be accommodated by allowing the user to use both branches, and this possibility may be allowed for when constructing the key.

We have associated the term *questions* with nodes, but it is more usual to refer to *tests* rather than questions. **Algorithms** for key construction are heuristic and are aimed at minimizing (i) the number of nodes, or (ii) the number of different tests used, or (iii) the average number of steps to identification, or, when tests have a significant cost, (iv) the average cost of identification. In the latter two cases, the relative probabilities, or frequencies, of the various outcomes may become relevant. The cost of tests may be associated with the time it takes to do them. Clearly, in some applications, waiting for the outcome of one test before deciding what test to do next can be very inefficient. This is typical of biochemical applications where tests arise from laboratory work, often automated. Then, it may be desirable to do batches of tests in parallel, requiring methods for deciding which tests should be batched together. Other refinements allow for recovery from errors that have led one down the wrong branch, reticulation that takes one across branches (so the key no longer has the form of a simple tree), and check keys which allow one to verify the correctness of an identification by including further otherwise redundant tests. Payne & Preece [25] review the methodologic literature on keys.

In all of the above it is assumed that the outcomes of tests are error-free, so that probabilistic methods are not appropriate. This is often a reasonable assumption for taxonomic keys, particularly at the genus level where, for example, all *Hexalectris* have "Lateral sepals free at base". Uncertainty has not been neglected entirely because we have mentioned the possibility of accommodating "don't knows", of recovery from an error that has led one down an incorrect branch, and of verifying the correctness of an identification. Also the probabilities of possible outcomes may be taken into account in minimizing costs and average numbers of steps to identification.

Probabilistic Assignment

While distinct groups may be separated by nonprobabilistic methods such as the diagnostic keys discussed

Table 2 The beginnings of a key to the genera of North American orchids [31]

1. Orchids terrestrial in habit	2
Orchids epiphytic in habit	40
2. Orchids saprophytic, lacking chlorophyll or green leaves	3
Orchids not saprophytic, with green leaves or green bractlike scales on a green stem	5
3. Flowers, stem, and bracts white, leaves absent	*Cephalanthera* (p. 54)
Flowers, stem, and bracts brownish, purple, or yellowish; leaves absent	4
4. Lateral sepals free at base	*Hexalectris* (pp. 110–112)
Lateral sepals united at base, forming a mentum	*Corallorhiza* (pp. 112–114)
5. . . .	

above, finer distinctions will often require probabilistic methods of the kind described below. This occurs when tests or characteristics overlap classes. Blood pressure varies widely in the human population and what may be a high blood pressure for one person, indicating heart disease, may be normal in another person. Thus, blood pressure by itself cannot distinguish between the class of healthy people and the class of diseased people; the values of blood pressure *overlap* in the two classes, or groups. Methods for assigning samples to one or more groups each characterized by its own probability distribution are among the oldest multivariate problems studied by statisticians, and under the name *discriminant analysis* have a large statistical literature (see, for example, the books by Hand [15], Lachenbruch [22], and McLachlan [23]). To many statisticians *classification* and *discrimination* are synonymous.

Barnard [1], who was advised by Fisher, and Fisher [8] introduced discriminant analysis. Fisher was concerned with discriminating between three species of iris and Barnard with discriminating between skulls from four Egyptian dynastic periods. Thus, the classes are the three irises and the four dynasties; both studies had many samples from each class. Fisher's approach was to use **linear regression**, defining a **dummy variable** as the independent variable (or **explanatory variable**) for each class. Barnard, somewhat arbitrarily, used -3, -1, 1, and 3 to characterize the four dynastics. Fisher took the irises two classes at a time characterized by a dummy variable with values -1, 1; with two classes it is immaterial what values are given to the dummy variables. There were four measurements on the four iris species – sepal length and width, and petal length and width – so a linear regression of the dummy variable on the four dependent variables was performed and used for discrimination, assigning a sample to one group when the discriminant function is positive and to the other group when it is negative. It worked very well, with few misclassifications. The calculations are easy to do and the discriminant function, being linear, is easy to use. In addition, it turns out to be quite robust to a range of distributional assumptions about the dependent variables. Consequently, the method remains popular and is much used in biometric applications, including the discrimination between diseased and healthy groups. Regression methods are well-known to carry with them certain problems; they predict much better for the sample used to fit them than for subsequent samples, and this is especially true when an optimal subset is selected from all the variables available. Naturally, these problems transfer to linear discriminators. When there is sufficient data, they are often divided into two sets: one, *the training set*, is used to compute the discriminant function, and the other is used to validate it. The error rates of the validation set give a much truer assessment of performance.

A version of two-group discrimination that owes much to Fisher's linear discriminant arises as follows. Suppose p is the probability that a sample belongs to the first group and $q = 1 - p$ is the probability that it belongs to the second group. Then we may assume that $\log(p/q)$ is a linear function of the dependent variables (*see* **Logistic Regression**). This is *logistic discrimination*; the generalization is similar to that of discriminating between **loglinear models** and linear models (*see* **Generalized Linear Model**). Logistic discrimination requires the estimation only of the regression coefficients, which usually have far fewer parameters than are required for other methods, described briefly below. Also it has certain optimal properties for a range of distributional assumptions for the two populations (see [27] for a more detailed discussion and further references).

The discrimination methods so far described only have heuristic or intuitive appeal. Since Welch [30], there has been a continuous development of theoretical underpinning. This is based on assuming that the ith group has a known density function $f_i(\mathbf{x})$, and then examining the probabilities of **misclassification error** arising from various assignment rules R, say. There are two probabilities of misclassification $p(1|2)$ and $p(2|1)$, respectively: the probability of a sample from group 2 being classified as from group 1 and the probability of a sample from group 1 being classified as from group 2. An obvious optimal classification rule is to choose R to minimize $p(1|2) + p(2|1)$. Figure 1 illustrates the geometry of errors of misclassification. The locus $f_1(\mathbf{x}) = f_2(\mathbf{x})$ defines a rule that minimizes $p(1|2) + p(2|1)$, and any other locus will give a greater total error of misclassification. Thus, we are led to a discriminant rule to assign to groups 1 and 2 according to whether $f_1(\mathbf{x}) < f_2(\mathbf{x})$ or $f_1(\mathbf{x}) > f_2(\mathbf{x})$. If, as is very unlikely, $f_1(\mathbf{x}) = f_2(\mathbf{x})$, then it is arbitrary which group is chosen, and we may toss a coin to decide. This leads to discrimination on the basis of whether or

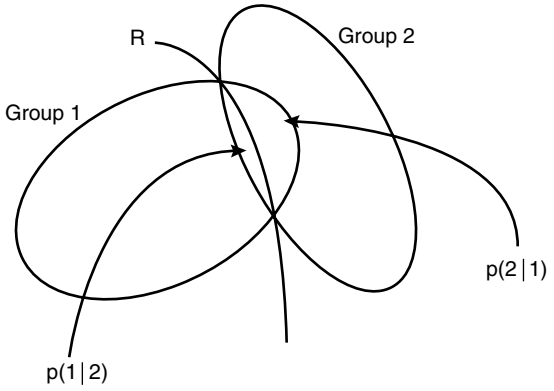

Figure 1 Illustration of the errors of misclassification for two groups. Although the elliptical shapes suggest normal distributions, they need not be so. The curve R separates space into two regions. Samples to the left of R are assigned to Group 1 and those to the right are assigned to group 2. R is supposed to show the locus of points with equal density and so passes through the intersections as shown

not the **likelihood ratio** criterion $f_1(\mathbf{x})/f_2(\mathbf{x})$ exceeds unity.

The likelihood ratio criterion is a good basis for discrimination and may be used whenever the functional forms $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are known. When both populations are characterized by **multivariate normal distributions**, with the same dispersion matrix $\mathbf{\Sigma}$, then the likelihood ratio criterion determines R to be linear, thus justifying Fisher's linear discriminant. However, when the dispersions differ, the best discriminator turns out to be quadratic. Although quadratic discriminators are optimal, under these assumption they are not robust to departures from normality, especially when the distributions show **skewness**. Errors of misclassification are not everything, and one may wish to take into account differing costs $c(1|2), c(2|1)$ for the different types of error, and also the **prior probabilities** q_1, q_2 of the groups. This replaces the likelihood ratio criterion by the *Bayes procedure* (*see* **Bayesian Methods**) with assignment based on

$$\frac{q_1 f_1(x)}{c(1|2)} \gtrless \frac{q_2 f_2(x)}{c(2|1)}.$$

Another criterion is the **minimax** rule, which minimizes the maximum error of misclassification.

The above has been concerned with discrimination among two groups, but the ideas are easily extended to k groups. The sample space is then divided into k regions, each corresponding to assignment to a different group. Assignment then is merely a matter of deciding in which region lies the sample to be identified. By far the most popular discrimination method of this kind is *canonical variate analysis* (*see* **Discriminant Analysis, Linear**), which applies to k groups each with a multinormal distribution with the same dispersion matrix $\mathbf{\Sigma}$. It is a generalization of linear discriminant analysis applied to k groups. The distance between the groups i and j is given by the **Mahalanobis distance** D_{ij} A sample may be assigned to the group with which it has the smallest Mahalanobis distance. In the **multidimensional scaling** version of canonical variate analysis, the distances may be approximated in a few dimensions and two-dimensional approximations may be viewed visually. With an appropriate scaling, the means of each group may be surrounded by circular confidence regions. When a sample falls within one of these circles, it may be assigned to the group to which it pertains. Samples not falling within any circle do not get assigned because they may be aberrant or they may more properly belong to a previously unrecognized group. For similar reasons, the possibility of not assigning a sample should be considered in all discrimination methods.

Errors of misclassification may be calculated from a knowledge of the distributions concerned. In all the above it is assumed that the parameters of all the distributions are known. In practice they have to be estimated from data, and these estimated values plug into the exact formulas. This introduces additional uncertainties into the accuracy of quoted errors of misclassification whose analysis is a highly technical matter [23].

Appropriate multivariate distributions underlying discrimination are not necessarily known, so **nonparametric methods** are valuable. This is one reason why the Fisher linear discriminant, in its regression interpretation, and logistic discrimination are so valuable. Another approach is to use kernel **density estimators** based on the data to form empirical density functions (see [27] for a review). Recently, neural networks have been used for discrimination. These work like diagnostic keys but, at each node of the tree, branching is decided on a single variable x according to whether $x \leq c$ or $x > c$, where c is a threshold to be determined [16].

Classification: the Construction of Classes

The above has been concerned with assigning to classes. Now we turn to constructing classes. As for assignment, class-construction may be considered either in a probabilistic framework or not. As Table 1 shows, sometimes we classify structured objects and sometimes unstructured objects, so the distinction between the two becomes important in the following.

Nonprobabilistic Classification

To a computer, n unstructured objects cannot be distinguished from n structured objects, so everything which is computable for the one is also computable for the other. Sometimes this makes sense (see the discussion of mixture problems which follows closely the K-group classification discussed next), but, for example, the hierarchical classification of unstructured objects makes little sense. Initially we shall be concerned with the classification of n structured objects, each representing a separate class as described above.

k-group classification classifies n structured objects into a specified number, k, of groups, sometime, termed a *partitioning* problem. Each object is described by p variables. The basis of the classification into k groups is the optimization of one of several possible criteria, C, which is discussed below. To optimize C we start with some initial classification. One strategy is to examine the effect of transferring an object from group i to group j, say. If this transfer improves C, then we let it stand, otherwise we replace the object in group i. Transfers are examined systematically until no further improvement is possible. Then we have attained a local optimum of C which gives a putative classification into k groups. Another strategy is to examine interchanges of objects between groups i and j, or a mixture of the two strategies may be used. A different starting configuration may give a better optimum, so usually it is recommended to try several starts; several methods have been suggested for finding a "good" start. It is rare for k to be known in advance, so the groupings given by different values of k may be found and the optimal values C_k examined to suggest an acceptable value of k. If the value of C_{h+1} is not much better than for C_h, then this indicates that we should choose $k = h$. This type of heuristic algorithm is much used and its efficiency is greatly increased if we can update C, following a transfer or interchange, rather than evaluate C *ab initio* every time. Fortunately, updating is straightforward for the three main choices of C, brief descriptions of which follow. For quantitative variables and a given grouping into k classes, we consider the between/within **multivariate analysis of variance** (MANOVA), which partitions the total sum of squares into a between-groups matrix $\mathbf{B}$ and a within-groups matrix $\mathbf{W}$. Then we may define C as follows: (i) k-means grouping (or minimum within-group sums-of-squares, trace $\mathbf{W}$); (ii) minimum $\det \mathbf{W}$; and (iii) maximal predictive classification. Because the total sum of squares $\mathbf{B} + \mathbf{W}$ is constant for all possible groupings, the grouping which minimizes trace $\mathbf{W}$ also maximizes trace $\mathbf{B}$. The latter may be interpreted as maximizing the sum of the squared distances between the group centroids (or means) weighted by the group sizes. If we ignore the group sizes, then we have the k-means algorithm which maximizes the sum of the squared distances between the group means or, equivalently, the sum of the squared distances between the means and the overall mean. If we operate on $\det \mathbf{W}$, then we have the same thing in terms of Mahalanobis D^2 and the canonical variate means (*see* **Mahalanobis Distance**); as with the k-means algorithm, we may operate in weighted or unweighted mode. We shall return to the minimization of $\det \mathbf{W}$ when discussing mixture problems, below.

Many classification problems of the type discussed above may be subsumed into a single algorithm which minimizes trace $(\mathbf{X} - \mathbf{GZ})\mathbf{W}^{-1}(\mathbf{X} - \mathbf{GZ})'$ over $\mathbf{G}$ and $\mathbf{Z}$, where $\mathbf{G}$ is an $n \times k$ indicator matrix of zeros and units giving class membership and $\mathbf{Z}$ is a $k \times p$ matrix giving the coordinates of cluster centers. A **least squares** algorithm achieves the minimization by alternating between estimating $\mathbf{Z}$ for fixed $\mathbf{G}$ and estimating $\mathbf{G}$ for fixed $\mathbf{Z}$. For the problems discussed above, every row of $\mathbf{G}$ will have unit sum, indicating that each object lies in one and only one class. By taking $\mathbf{W} = \mathbf{I}$ we arrive at the k-means problem and by taking $\mathbf{W}$ to be the pooled within-group sums-of-squares matrix we minimize the squared Mahalanobis distances between groups. By removing the unit sum constraint, overlapping clusters may be formed.

Maximal predictive classification operates on binary rather than quantitative data. Thus, each variable can take only values 1 and 0, which may represent

presence/absence or two equal-status categorical levels such as male/female. For each group we may predict the level that occurs most frequently for each variable and then count the number of correct predictions in the group; this will give more correct predictions than for any other choice of predictor. The choice of the criterion is now to choose the partition into k groups that maximizes over all groups the number, C_w, of correct predictions. Alternatively, we could choose C_b to minimize the number of correct predictions averaged over the $k = 1$ predictors for the wrong groups, C_w measures the homogeneity of groups, which we want to be good (i.e. big), while C_b gives a measure of the separation between groups, which we also want to be good (i.e. small). Unlike the situation in MANOVA, where $\mathbf{W} + \mathbf{B}$ is constant for all k, $C_w + C_b$ varies with k. The difference $C_w - C_h$, which balances homogeneity within classes with separation between classes, may be used to suggest suitable values of k [14].

If now we denote by C_k the optimal value given by the k-group criterion for k classes, and the groupings of samples for C_{k+1} happen to be nested within the groupings given by C_k for $k = 1, 2, \ldots, n$, then we have a natural hierarchical classification. Natural hierarchical classifications are very rare but when absent a *best* hierarchical classifications may be desired. We could define the best hierarchical classification as that which optimizes $\sum_{k=1}^{K} C_k$, where the groupings which generate the C_ks are constrained to be nested. This gives a very general way of defining hierarchical classification with respect to any C-criterion, but it has not been studied and almost certainly leads to formidable combinatoric computational problems. The heuristic algorithms, outlined above, for optimizing k-groups, partitioning, and hierarchic interio, usually deliver a local optimum but, until recently, were the only ones practicable. Recent advances in dynamic programming (see Hubert, Arabie and Meuluran, [18]) are beginning to offer the feasibility of finding the global optimum-atleast for n < 30, say.

The desire for seeking hierarchical classifications of structured objects seems to stem partly from taxonomy, where evolutionary considerations naturally lead to hierarchical relationships between living (and fossil) organisms, but also from the importance of hierarchical systems for the organization of many things, ranging from library books to government or management of any large organization (*see* **Numerical Taxonomy**). There is an enormous literature (e.g. reviews by Cormack [4] and Gordon [10, 11, 13] on hierarchical classification and related methods. Many of the original methods for determining hierarchical classifications are heuristic **algorithms** that operate on matrices giving the (dis)similarities between all pairs of the n objects (*see* **Similarity, Dissimilarity, and Distance Measure**). A typical algorithm is that for **single-linkage** cluster analysis, which at the ith step has i groups and joins the two groups which share the smallest dissimilarity between two objects – one chosen from each group. The algorithm starts with each object in a separate group and ends up with all objects in a single group; such algorithms are said to be *agglomerative*. *Divisive* algorithms work in the opposite way, starting with the n objects in a single group and subdividing existing groups at each stage until each object is in a separate group. It turns out that both agglomerative and divisive algorithms can be found for single linkage clusters, but this is not so for most other methods. The successive divisions of the groups in a hierarchical classification can be shown as in Figure 2.

This is an artificial example which shows that objects A and B join at a level of dissimilarity of 0.1 and are joined by object C at a level of dissimilarity of 0.2; objects D and E join at level 0.3 and combine with the other three objects at level 0.4. Such a diagram, with an associated scale of dissimilarities, is known as a dendrogram. A set of fitted dissimilarities may be calculated from any dendrogram and the fitted values for the dendrogram of Figure 2 are shown alongside. The fitted dissimilarity for objects i and j is obtained by finding the dissimilarity at which i and j first join. Such dissimilarities satisfy the *ultrametric* inequality $d_{ij} \leq \max(d_{ik}, d_{jk})$, which is a stronger form of the usual metric inequality satisfied by the

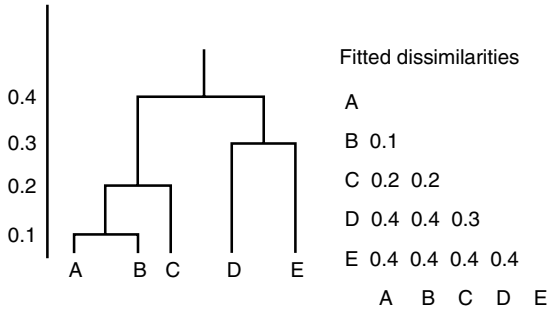

Figure 2 A dendrogram representation of a hierarchical classification with fitted ultrametric dissimilarities

sides of a triangle $d_{ij} \leq (d_{ik} + d_{jk})$. Not only does every dendrogram define dissimilarities that satisfy the ultrametric inequality, but also the converse is true [17, 19, 21]. It follows that the best hierarchical tree may be defined as the tree which minimizes

$$\sum_{i=1}^{n} (d_{ij} - \delta_{ij})^2, \tag{1}$$

where (d_{ij}) is the matrix of observed dissimilarities and (δ_{ij}) is the matrix of fitted ultrametrics. Unfortunately, such criteria are hard to minimize numerically, and existing algorithms are not very efficient. Therefore, heuristic algorithms remain popular (see, for example, [28]). Many of these algorithms can be unified by choosing different parameters in formulas that give the dissimilarity between any group and two merged groups as a linear combination of quantities that characterize the dissimilarities between the three groups concerned and their heights in the dendrogram. The most simple general-purpose formula giving the dissimilarity between group k and the amalgamation of groups i and j is

$$d_{k(i,j)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}|,$$

where the Greek letters are adjustable parameters which may be chosen to give different agglomerative algorithms. Unifications of these kinds, and there are several of them, facilitate programming general-purpose software for hierarchical classification encompassing many methods, but writing tailor-made programs for some of the special cases can be more efficient.

As well as characterizing trees by ultrametric distances, they may also be characterized by additive distances. Figure 3 illustrates the concepts concerned.

Now, rather than an accompanying scale of dissimilarities as in a dendrogram, the length of each branch is given. The resulting trees are termed *additive trees* because the fitted values are determined by the length of the shortest path between two endpoints, which is obtained by adding up the component parts of the path, as is shown in the table alongside the tree of Figure 3. Thus, the distance from A to C is $0.2 + 0.1 + 0.2 = 0.5$. For all quadruples i, j, k, l these fitted dissimilarities satisfy the four-point metric,

$$\delta_{ij} + \delta_{kl} \leq \delta_{il} + \delta_{jk} = \delta_{ik} + \delta_{jl},$$

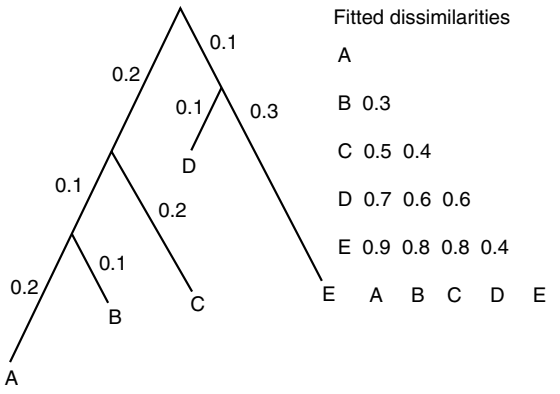

Figure 3 An additive tree representation of a hierarchical classification with fitted four-point metric dissimilarities

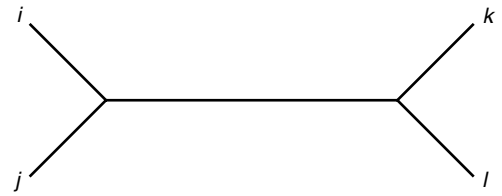

Figure 4

where the two longest sums are equal. This is often drawn as in Figure 4.

By setting $k = l$, so that $\delta_{kl} = 0$, it follows that the four-point inequality also satisfies the metric inequality for all triplets. Now (1) may be fitted by constraining the δ_{ij} to satisfy the four-point inequality, but efficient algorithms are available only for small problems. Instead, efficient heuristics are used which ensure iteratively that the four-point property is satisfied locally, e.g. by finding the best least-squares approximation to every set of four points (see [9] for an algorithm of this kind and a summary of related work). An alternative approach depends on a property of additive trees that they may be expressed as a sum of an ultrametric tree and a star tree, i.e. an additive tree with just one node. This may be written as

$$a_{ij} = u_{ij} + a_j + a_i, \quad i, j = 1, 2, \ldots, n,$$

where a_{ij} denotes the additive distance between classes i and j, a_i are additive constants, and u_{ij} denotes corresponding ultrametric distances. Thus, in the example of Figure 3, we may set $a_1 = 0.2$, $a_2 = 0.1$, $a_3 = 0.1$, $a_4 = 0.1$, and $a_5 = 0.1$, and verify that

the elements $u_{ij} = a_{ij} - a_j - a_i$ satisfy the ultrametric inequalities. The reader may check that the quantities a_i differ only by a constant, chosen to ensure that the u_{ij} are nonnegative, from the lengths of the branches of Figure 3 from the root node. Provided one has an algorithm for fitting ultrametric trees, this result allows one to construct a least squares algorithm for fitting additive trees by alternating between fitting the ultrametrics and the additive constants.

The previous result may be considered as part of the family of so-called *hybrid* models:

$$d_{ij} = u_{ij} + c_{ij}, \qquad d_{ij} = u_{ij} + a_i + a_j + c_{ij},$$

$$d_{ij} = \sum_{r=1}^{R} u_{ijr} + e_{ij}, \quad d_{ij} = u_{ij} + f_{ij} + e_{ij},$$

which place hierarchical classification problems in the wider context of statistical modeling and takes us rather far from the topic of classification under discussion.

The same objects may be given several different hierarchical classifications, either because they are based on different sets of variables or because they are based on different algorithms. The question therefore arises of deciding to what extent the classifications agree with one another. A first step is to determine an average or *consensus* classification. There are many ways to do this, the most simple of which is to fit a tree to the average (root mean square is better) of the ultrametric or additive distance matrices arising from each classification. A more popular approach is to devise combinatoric algorithms that search for objects that group together in all, or at least in many, of the separate classifications [5].

Sometimes, hierarchical classifications need to be constrained in some way. When classifying geographic, ecological and **image** data it is natural to constrain classes to be spatially contiguous so that the resulting clusters have spatial integrity. To impose constraints greatly complicates clustering algorithms, but considerable progress has been made [12].

As well as ultrametrics and additive metrics, other structures have been fitted to dissimilarity matrices. Critchley defines *ziggurats*, which are a special class of tree where the objects split off from the main tree one at a time. Diday defines pyramids which are not tree structures but are highly interlinked systematic rooted graphs. Indeed, the possibility of fitting interlinked, or overlapping, trees has been recognized for many years (see, for example, [20]). My impression is that these structures are currently mainly of research interest and have rarely been used in applications; perhaps they are more concerned with data description rather than with classification.

Probabilistic Classification

The main probabilistic method for forming classes from unstructured objects is known as the *multivariate mixture problem*. It has a long history in statistics and in its univariate form was first studied by Karl Pearson [26]. Pearson was concerned with data on several hundred shin bones from an archeological investigation. He knew that some were female and some were male, and wished to assign the bones to the two classes, after identifying and eliminating immature skeletal material. In this case, he knew that there were only two classes, but in problems where the objects are a mixture of samples from several populations, the number of classes is not necessarily known in advance. Thus, the mixture problem is to assign each of n given samples, on each of which the same p variables have been measured, to one of k classes and to determine the best value of k. Each class is assumed to be identified by a probability distribution function $f_i(x, \theta_i)$, $i = 1, 2, \ldots, k$, and a supplementary interest is to estimate the values of the parameters θ_i of these distributions. Being set up in a formal statistical context, a maximum likelihood approach can be taken. The likelihood is

$$L = \prod_{x \in C_1} f_1(x, \theta_1) \prod_{x \in C_2} f_2(x, \theta_2) \ldots \prod_{x \in C_k} f_k(x, \theta_k). \tag{2}$$

Thus, L is maximized for some permutation of the samples assigning them to the k classes, no class being empty, and the parameters θ_i, $i = 1, 2, \ldots, k$, being estimated from those samples assigned to C_i. Of course, there are an enormous number of possible permutations, so the computational problem is formidable and approximate heuristic algorithms are used. Transfer and interchange algorithms like those described above for nonprobabilistic classification are applicable, but whenever a sample is transferred from C_i to C_j the parameters θ_i and θ_j have to be reestimated which, unless a simple updating procedure is available, adds to the computational burden. An

important special case is when all the distributions have multinormal form with the same dispersion matrix. Then, the updating of the class-means and within-class dispersion matrix **W** is simple and it turns out that the permutation of the samples which maximizes the likelihood is when det **W** is minimized, thus leading back to one of the variants of the k-means problem and algorithm.

Mixture distributions may also be modeled as the distribution function

$$\sum_{i=1}^{k} p_i f_i(x, \theta_i),$$

where p_i gives the frequency of the ith population. The parameters, including the frequencies, may again be fitted by maximum likelihood. In the previous approach the frequencies may be estimated from the relative sizes of the classes C_i.

In both approaches it seems difficult to determine an optimal value of k, and I am not aware that anything better has been suggested than plotting the maximized likelihood against k and hoping to find a "dogleg" pattern in which the likelihood increases with k until a level is found where increasing k has little effect. The point where this occurs, if it does, gives a natural choice for a suitable value of k. A device which has some popularity is to add a penalty function to the likelihood (*see* **Penalized Maximum Likelihood**), so discouraging large values of k. Mixture problems have a large statistical literature, e.g. [6, 24].

Jardine & Sibson [20] suggest that classes be formed on the basis of what they term *information gain*. They use Renyi's measure of **information** to define the information in a classification and calculate the information gained from regarding the samples as belonging to a mixture of k classes rather than that all the samples belong to a single class. Similarly, information gain can be calculated when proceeding from any classification to one where the number of classes is decreased. Apart from using information gain rather than likelihood, the concepts and the computational problems are similar for the two approaches.

There is a hierarchical version of the mixture problem which seeks to classify n unstructured objects into k nested classes. Thus we seek a tree with k end points, each corresponding to a mixture component. It is rare for those who construct hierarchical classifications, even of structured objects, to be really interested in trees with n end points. Very often they will truncate the tree at some convenient level in the dendrogram. For example, in Figure 2 we might truncate at 2.5 leading to three nested classes (A, B, C), (D), and (E). Such ad hoc rules are frequently used, but I know of no direct modeling for this type of problem.

A problem that could be studied, but seems not to have been, is concerned with the classification of n objects structured into k groups. These objects may be parameterized as for discriminant analysis, and so may be represented by either n fully parameterized distribution functions or by samples from the n objects from which the parameters of the distributions may be determined. Some objects may be closer together than others, as measured, for example, by the Mahalanobis D^2, and it seems reasonable that these be classified together. In general we might wish to group the n objects into fewer homogeneous classes. In an extreme case, the objects may be characterized by point distributions so that there is no overlap and errors of misclassification are irrelevant. This suggests that some other basis for classification must be used to replace, or add to, probabilistic measures of distance between populations and between grouped populations.

Links Between Construction and Assignment to Classes

In this article I have stressed the distinction between constructing classes and assignment to classes. Yet the distinction is often blurred because one legitimate reason for constructing classes is so that future assignments to the classes should be optimal in some sense. We have seen that diagnostic keys are constructed so that identification is optimized either in terms of the number of tests that have to be done or in terms of costs. Of course, the classification is given at the outset, but if we were allowed to redefine our classes, perhaps we could reduce the number of tests required or the costs. Then we could say that the best classification is the one that minimizes these quantities. Unfortunately, there is a trivial solution with only one class, and then every specimen could immediately be assigned to that class without cost. The anomaly arises because we are not allowed to

redefine the basic classes. However, if we reallocated the species among the genera, then the key to the genera might be shorter or cheaper to use. In this way we could define a classification with optimal properties for assignment. Maximal predictive classification has such a property, as it can be shown that any object shares more properties with the class predictor for its own class than with the class predictors for all the other classes. Thus, an object may be assigned to its proper class merely by counting how many properties it shares with each class predictor. Mixtures estimated by maximum likelihood have a similar property, as can be seen from the following result, which derives from (2). Suppose (2) has been maximized and we move an object $\mathbf{x}$ from C_i to C_j. Then the likelihood becomes $L^* = Lf_j(\mathbf{x}, \theta_j)/f_i(\mathbf{x}, \theta_i) < L$. Thus, $f_j(\mathbf{x}, \theta_{ij}) \leq f_i(\mathbf{x}, \theta_i)$, showing that the boundary between objects assigned to classes C_i and C_j is precisely that given by the likelihood ratio criterion for discrimination regions. Thus, the classes obtained from the mixture model are optimal for discrimination.

Of special interest in the biometric context are problems of constructing classifications based on evolutionary concepts. Diagnostic keys and the hierarchical classifications discussed earlier are unashamedly utilitarian and, despite their hierarchical structure, make no evolutionary claims. There has been much controversy and confusion between utilitarian hierarchical classifications and classifications supposed to reflect evolutionary relationships (see, for example, [2]). This is a little surprising, because the Linnaean binomial classification of the natural world long predates Darwin and, in its original form, cannot be claimed to be based on evolutionary principles. For more than 100 years taxonomists have been extending and modifying Linnaean classifications and, for the most part, have attempted to mirror supposed evolutionary development. Similarity between organisms is likely to bear some relationship to shared genetic material but evolutionary convergence between, for example, some of the superficial features of fishes and whales shows that evolutionary classifications must be approached with care and that, although phenotypic similarity may be a satisfactory basis for constructing utilitarian classifications, it may not be for evolutionary classifications. Nevertheless, phenotypic classifications are likely to contain a substantial element of evolutionary classification, even though that is not their prime objective. Two approaches to evolutionary classification have been developed – one probabilistic and one not. The probabilistic approach is useful for constructing evolutionary trees for genetically close classes, such as different human populations, and is based on modeling genetic drift and mutation. Being a probabilistic model, in principle its parameters and the topology of the optimal tree may be estimated by maximum likelihood. The computational problems are formidable but some progress has been made (see, for example, [7]) that allow the method to be used for small problems. The nonprobabilistic approach, often termed *cladistics*, is based on the concept of *minimal evolution*, which is appropriate for categorical variables or characters. Minimal evolution requires the fitted tree to minimize the number of character changes as one moves from node to an adjacent node. There are difficulties, because with multistate characters the states have to be ordered from their presumed primitive forms to their most developed evolutionary forms. Further, the fossil record has considerable gaps and so one must allow for hypothetical missing organisms and a suitable root for the tree must be identified. When the length of a branch of a tree is defined as the number of character changes between its nodes, then the minimal evolution tree is the same as the minimal-length additive tree, in this context known as a Wagner tree. For further discussion, see [3, 29, 31].

References

[1] Barnard, M.M. (1935). The secular variation of skull characters in four series of Egyptian skulls, *Annals of Eugenics* **6**, 352–371.

[2] Calcraft, J. (1983). The significance of phylogenetic classifications for systematic and evolutionary biology, in *Numerical Taxonomy*, J. Felsenstein, ed., NATO Advanced Science Institutes, Series G, Ecological Sciences, Vol. 1, Springer-Verlag, Berlin, pp. 1–17.

[3] Colless, D.H. (1983). Wagner trees in theory and practice, in *Numerical Taxonomy*, J. Felsenstein, ed., NATO Advanced Science Institutes, Series G, Ecological Sciences, Vol. 1, Springer-Verlag, Berlin, pp. 259–278.

[4] Cormack, R.M. (1971). A review of classification (with discussion), *Journal of the Royal Statistical Society, Series A* **146**, 246–272.

[5] Day, W.H.E., ed. (1986). Consensus classification (special issue), *Journal of Classification* **3**.

[6] Everitt, B.S. & Hand, D.J. (1981). *Finite Mixture Distributions*. Chapman & Hall, London.

[7] Felsenstein, J. (1983). Statistical inference of phylogenies (with discussion), *Journal of the Royal Statistical Society, Series A* **134**, 321–367.

[8] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–188.

[9] Gascuel, O. & Levy, D. (1996). A reduction algorithm for approximating a (nonmetric) dissimilarity by a tree distance, *Journal of Classification* **13**, 129–155.

[10] Gordon, A.D. (1987). A review of hierarchical classification, *Journal of the Royal Statistical Society, Series A* **150**, 119–137.

[11] Gordon, A.D. (1996). Hierarchical classification, in *Clustering and Classification*, P. Arabie, L. Hubert & G.A. De Soete, eds., World Scientific, River Edge, pp. 65–121.

[12] Gordon, A.D. (1996). A survey of constrained classification, *Computational Statistics and Data Analysis* **26**, 17–29.

[13] Gordon, A.D. (1999). *Classification*, 2nd Ed. Chapman & Hall-CRC Press, London.

[14] Gower, J.C. (1975). Maximal predictive classification, *Biometrics* **30**, 643–654.

[15] Hand, D.J. (1981). *Discrimination and Classification*. Wiley, New York.

[16] Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. Wiley, Chichester.

[17] Hartigan, J.A. (1967). Representation of similarity matrices by trees, *Journal of the American Statistical Association* **62**, 1140–1158.

[18] Hubert, L., Arabie, P. & Meuluran, J. (2001). *Combinational Data Analysis: Optimization by Dynamic Programming*, SIAM monographs on discrete mathematics and its applications. Society for Industrial and Applied Mathematics, Philadelphia.

[19] Jardine, C.J., Jardine, M. & Sibson, R. (1967). The structure and construction of taxonomic hierarchies, *Mathematical Biosciences* **1**, 173–179.

[20] Jardine, N. & Sibson, R. (1971). *Mathematical Taxonomy*. Wiley, Chichester.

[21] Johnson, S.C. (1967). Hierarchical clustering schemes, *Psychometrika* **32**, 241–254.

[22] Lachenbruch, P.A. (1975). *Discriminant Analysis*. Hafner, New York.

[23] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

[24] McLachlan, G.J. & Basford, K.E. (1988). *Mixture Problems: Inference and Applications to Clustering*. Marcel Dekker, New York.

[25] Payne R.W. & Preece, D.A. (1980). Identification keys and diagnostic tables: a review (with discussion), *Journal of the Royal Statistical Society, Series A* **143**, 253–292.

[26] Pearson, K. (1894). Contributions to the mathematical theory of evolution. I. Dissection of frequency curves, *Philosophical Transactions of the Royal Society, Series A* **185**, 71–110.

[27] Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

[28] Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

[29] Wagner, W.H. (1981). Problems in the classification of ferns, in *Recent Advances in Botany*. University Press, Toronto, pp. 841–844.

[30] Welch, B.L. (1939). Note on discriminant functions, *Biometrika* **31**, 218–220.

[31] Wiley, E.O. (1981). *Phylogenetics. The Theory and Practice of Phylogenetic Classification*. Wiley, New York.

[32] Williams, J.G. & Williams, A.E. (1983). *Field Guide to North American Orchids*. Universe Books, New York.

JOHN C. GOWER

Classifications of Medical and Surgical Procedures

Many of the principles governing the construction of classifications of procedures parallel those relating to the **International Classification of Diseases (ICD)**. Like the ICD, classifications of procedures are designed to facilitate statistical analysis, with the structure and composition of categories reflecting their frequency of occurrence and surgical importance. Classifications of procedures are not intended to be surgical nomenclatures, although in common with other similar classifications they are designed to be accessed from the clinical terminology used to describe surgical and other operations.

An important additional consideration which often affects the construction of a procedure classification is an explicit definition of its scope. Although the earliest examples of such classifications were confined to those surgical operations normally carried out in operating theaters, gradually the scope of "surgery" has extended to procedures carried out in other environments. More recently there is increasing pressure for procedure classifications to include other forms of less invasive intervention, particularly those which entail the use of expensive resources.

The **World Health Organization (WHO)** considers that a Classification of Procedures should be a component of the "family" of disease and health-related classifications and indeed, for trial purposes in 1978, WHO published an International Classification of Procedures in Medicine (ICPM) [11], which was adopted by a few countries, and used as a basis for a national classification of surgical operations by a number of others. For example, a procedure classification developed in this way is an integral part of the United States International Classification of Diseases, 9th Revision Clinical Modification (ICD-9CM) [10].

Nevertheless, many countries have independently developed their own procedure classifications. Well-known examples include those from Canada [7] and the recently published Nordic Classification of Surgical Procedures [6]. This latter classification has been structured so that the listed codes form a tiny proportion (less than 0.5%) of the available space. This graphically illustrates a further particular feature of the structure of a procedure classification, in that it needs to be able to respond appropriately to the ever increasing developments in surgical techniques. The relatively poor acceptance of an international version to some extent reflects the fact that surgery is practiced differently in each country. More importantly, a procedure classification is frequently used as an important part of the billing process, or for other similar revenue purposes. In these respects, individual countries have widely disparate needs, and thus there is no uniform requirement or specification. A classification designed to meet the needs for groupings which easily aggregate on the basis of iso-resource operating theater costs would be constructed quite differently from one designed for epidemiologic purposes in order to add a proxy dimension of severity when combined with a suitable diagnostic classification such as the ICD.

In the UK, a classification of surgical operations has been available for use since 1944, when one which identified 443 categories of operation was published by the Medical Research Council [4]. The then General Register Office prepared and issued an updated version in 1950 [1], and revisions to this were subsequently issued in 1956 (first revision) [2], 1969 (second revision) [3] and 1975 (third revision) [8]. The current fourth revision (normally referred to as OPCS-4) was published by the then Office of Population Censuses and Surveys in 1990 [9].

Since 1990, the content of the tabular list categories in OPCS-4 has remained constant but the index has been regularly updated to give additional entries to these categories. Responsibility for maintenance of the classification now lies with the National Health Service Centre for Coding and Classification, but constant expansion of the frontiers of surgery presents its own particular problems. These are typified by the difficulties of fitting the recent growth in techniques of minimal access surgery within a basic classification structure not designed to accommodate them.

OPCS-4 has a wide range of national uses within the UK, including the presentation of statistics of hospital use, and as a central element in the construction of Health Care Resource Groups (HRGs) as defined by the National Casemix Office [5]. It has similar local uses, which also include a major role in the setting and monitoring of contracts. In some places it is also used in operating theater applications such as scheduling, although it was not designed for this latter purpose.

References

[1] General Register Office (1950). *Draft Code of Surgical Operations*. GRO, London.

[2] General Register Office (1956). *Code of Surgical Operations*. GRO, London.

[3] General Register Office (1969). *Classification of Surgical Operations*, 2nd rev. GRO, London.

[4] Medical Research Council (1944). *A Provisional Classification of Diseases and Injuries for Use in Compiling Morbidity Statistics*. HMSO, London.

[5] National Casemix Office (1995). *Turning Data into Information*. IMG G7059 NCMO, Winchester.

[6] Nordic Medico Statistical Committee (1996). *Classification of Surgical Procedures*, Vol. 46. NOMESCO, Copenhagen.

[7] Nosology Reference Center, Statistics Canada (1978). *Canadian Classification of Diagnostic Therapeutic, and Surgical Procedures*. Statistics Canada, Ottawa.

[8] Office of Population Censuses and Surveys (1975). *Classification of Surgical Operations*, 3rd rev. OPCS, London.

[9] Office of Population Censuses and Surveys (1990). *Tabular List of the Classification of Surgical Operations and Procedures*, 4th rev. HMSO. London.

[10] United States Department of Health and Human Services (1991). *The International Classification of Diseases, 9th Revision Clinical Modification*, Vol. 3: *Procedures*, 4th Ed. PHS (91)-1260. US Government Printing Office, Washington.

[11] World Health Organization (1978). *International Classification of Procedures in Medicine*. WHO, Geneva.

JOHN S.A. ASHLEY

Clinical Epidemiology

Clinical epidemiology involves the application of methods derived from epidemiology and other fields to the study of clinical phenomena, particularly diagnosis, treatment decisions, and outcomes. Clinical epidemiology has been characterized, somewhat immodestly but fairly accurately, as the basic science of clinical medicine. Clinical epidemiology is not a clearly delimited field. In its concern with the accuracy of diagnosis, the elements of treatment decisions, and the measurement of outcomes, clinical epidemiology overlaps substantially with clinical medicine. In its focus on physician and patient choices and the interactions between patients and physicians in clinical processes, it overlaps substantially with **health services research**. In its use of various quantitative methodologies to address questions of clinical relevance, clinical epidemiology overlaps substantially with epidemiology, economics, and other disciplines.

To facilitate an understanding of this broad field, it is helpful, if somewhat arbitrary, to consider clinical epidemiology in terms of its major areas. These include clinical measurement, diagnosis and screening, clinical decision making, measurement of treatment effects, clinical economics, and clinical study design.

Clinical Measurement

Clinical measurement, or clinimetrics, as it is sometimes called, is the foundation for all of clinical epidemiology. Sound measurements are the raw material for the various methodologies of the clinical epidemiologist. If this raw material is flawed, then it is very likely that the results and conclusions derived from it will also be flawed.

Clinical measurement involves some issues that are common to all forms of measurement and others that are almost uniquely applicable to measurement as a clinical process. The common issues for any measure include its reliability, validity, responsiveness, and generalizability. A measure is reliable if it provides consistent results when used to measure an unchanged phenomenon at different times or in different settings. A measure is valid if it truly measures the construct it is assumed to measure. A measure is responsive if it is sensitive to clinically meaningful change. A measure is generalizable if it can be usefully applied to subjects who differ by age, gender, race, diagnosis, or some other major characteristic (*see* **Validity and Generalizability in Epidemiologic Studies**).

Other issues in clinimetrics arise from the fact that clinical measurement typically involves an examiner, an examinee, and an examination technique, all of which are subject to **measurement error**. The clinical examiner may be inadequately trained or may be prone to **biased** measurement. The clinical examinee is subject to physiological variation and is prone to reporting bias. The examination methodology may be affected by a variety of factors, including calibration errors and changes in technique. The effects of the sources of measurement error can be cumulative, making clinical measurement a particularly challenging proposition.

Diagnosis and Screening

Diagnosis is a critical step in the clinical process, and the analysis of the diagnostic process is a major focus for clinical epidemiology. Diagnosis refers to the categorization of individuals who have come to a clinician with symptoms. **Screening** refers to the categorization of asymptomatic individuals in a clinical setting or in the general population.

The analysis of diagnosis in clinical epidemiology focuses on the performance of **diagnostic tests**. The key properties of a diagnostic test are its **sensitivity** and **specificity**. Sensitivity refers to the frequency with which a test is positive when it is applied to a group of individuals known to have a particular disease. Specificity refers to the frequency with which a test is negative when it is applied to a group of individuals known to be without a particular disease.

In the clinical setting, the sensitivity and specificity of a test, combined with the estimated **prevalence** of the expected diagnosis (sometimes referred to as the pretest probability of disease) determine the **predictive value** of a test. A positive predictive value estimates the probability that an individual with a positive diagnostic test result will actually have a particular disease. A negative predictive value estimates the probability that an individual with a negative test will actually be free of disease. Since pretest probability is substantially lower for screening than it is for diagnosis,

the predictive value of diagnostic tests may be substantially lower when they are used for screening.

Diagnosis in medicine is seldom based on the positive or negative result of a single test. In most diagnostic situations several diagnostic tests are available, with each providing a range of results rather than a simple positive or negative. Methods have been adopted to compare the diagnostic efficiency of different clinical tests across their range (**receiver operating characteristic (ROC) curves**) and to determine the probability of a diagnosis given the level of a diagnostic test result (**likelihood ratios**).

In many diagnostic or screening situations several tests are used in combination. In circumstances in which a diagnosis should not be missed, such as a highly treatable infectious disease, one or more diagnostic tests may be applied in parallel with any positive result leading to a diagnosis. This approach enhances the sensitivity of a diagnostic or screening strategy. In other settings, serial testing, in which a second test is done only if a first one is positive, minimizes the use of costly or dangerous second-stage tests. Serial testing lowers the sensitivity of the diagnostic or screening strategy, but it maximizes specificity and minimizes **false positive rates**.

Clinical epidemiologists also concern themselves with issues of intra- and interobserver agreement (*see* **Observer Reliability and Agreement**). In the realm of diagnosis, intraobserver agreement estimates the extent to which a clinician reproducibly categorizes subjects into diagnostic categories. Interobserver agreement estimates the extent to which two or more clinicians agree on the diagnostic categorization of subjects. Since agreement in either situation may occur by chance, **kappa** and other statistics are used as measures of agreement that adjust for chance. Clinical epidemiologists have repeatedly demonstrated that the adjusted diagnostic agreement between two trained clinicians may be surprisingly low.

Clinical Decision Making

The key methodology for the study of clinical decision making is formal **decision analysis**. Decision analysis involves the construction of detailed trees in which each decision point in a clinical treatment cascade is specified, probabilities are assigned to an exhaustive group of potential outcomes, and values are assigned to each outcome. To guide clinical practice, formal decision analysis can be quite helpful in clearly specifying the issues and potential outcomes in a given clinical decision (*see* **Decision Theory**). At the level of clinical epidemiology, formal decision analysis forms the basis for explicit cost and benefit estimates that serve to inform the choice between treatment alternatives (*see* **Health Economics**).

Measurement of Treatment Effects

A central topic in clinical epidemiology is the estimation of benefits and side-effect rates that result from clinical therapies. The estimation of benefits involves the specification and measurement of the major components of treatment outcome. Until recently, treatment outcome was usually estimated in terms of reduced mortality or improvements in the physiological manifestations of a particular disease, such as blood pressure reduction in hypertension treatment and blood sugar control in diabetes treatment.

In recent years, clinical epidemiologists, in concert with health services researchers, have expanded the measurement of treatment benefits to include the four major components of health status: physical function, psychological function, social function, and symptoms. These elements of health status are measured using a variety of methods, with an emphasis on recently developed questionnaire approaches. These questionnaires have proven to be as reliable, valid, responsive, and generalizable as more traditional clinical measures of benefit. In addition, these newer measures have more relevance for patients because they assess treatment benefits, such as functional capacity and symptom reduction, that are of particular concern to patients.

Treatment benefits may be measured using general methods that apply across diagnostic categories. The SF-36 health status questionnaire is the most widely used example of this approach. Other questionnaires are designed to measure the major elements of health status in particular categories of disease, such as arthritis or respiratory disease-specific measures (*see* **Questionnaire Design**).

Treatment benefits may also be estimated in terms of patient utilities. In this approach, such methods as **time tradeoff** and **standard gamble** are used to estimate patient-specific preferences for potential benefits. Although **utility** estimation may be methodologically difficult, it is conceptually appealing, and it provides a basis for comparing treatment benefits across different disease categories and patient groups.

There is no such thing as a treatment without costs. So in their efforts to develop an accurate and balanced assessment of medical treatments, clinical epidemiologists must measure the rate of adverse events and their severity as well as benefits. Drug toxicities, or adverse effects, may result from an overshoot in the intended effect (e.g. low blood sugar caused by a diabetes treatment), from undesirable but related physiological effects (e.g. stomach ulcers from arthritis drugs), and from apparently idiosyncratic effects (e.g. headache or skin rash from various drugs).

The study of the frequency and severity of adverse events falls into the realm of **pharmacoepidemiology**. For common treatment side-effects, follow-up of patients on medication or even data from **clinical trials** can be utilized to characterize the frequency and severity of side-effects. The rate of common side-effects can be characterized with the highest level of validity because these rates are often derived from close observations by practitioners who are actively monitoring clinical subjects for drug side-effects. For rarer adverse events, computerized clinical databases or claims databases (*see* **Administrative Databases**) are increasingly being used, for they are the best source of information on large numbers of people under treatment. Both approaches typically utilize the prescription as their measure of treatment exposure. Any side-effect of treatment that occurs outside the medical record or claims file may be difficult to capture and count. There is also an inherent problem in large-scale **observational studies** of drug use in that certain drugs may be given only to persons at high risk of certain side-effects, so that an association with those side-effects is to be expected. This **confounding** by indication has made it nearly impossible to perform pharmacoepidemiologic studies of drugs that are supposed to protect against other drug side-effects. Notwithstanding these difficulties, large-scale, computer-based pharmacoepidemiology studies offer promise in terms of identifying and quantifying the serious adverse effects of drugs.

Clinical Economics

Dollars are the other major cost of treatment, so clinical epidemiology must concern itself with the estimation of dollar costs. The first issue in doing any clinical cost study is to define the perspective from which costs are estimated. For example, a vaccination program will have different costs depending on whether costs are analyzed from the perspective of the State Health Department, the clinic providing the vaccinations, or the family whose child is vaccinated.

There are three major types of clinical cost study. The first is a descriptive study in which the costs of a treatment are measured in terms of direct medical costs (e.g. the cost of a clinic visit and the prescription), direct nonmedical costs (e.g. the cost of losing time from work to receive treatment), and indirect costs (e.g. the cost of reduced productivity caused by disease-related disability or death). Cost estimates may then be used in a cost-effectiveness study in which the costs of treatment are compared to benefits measured in terms of clinical or health status improvements. The third type of cost study is a cost–benefit study in which both the costs and benefits of treatment are measured in dollar terms. Although it is the most difficult form of cost study to carry out, the cost–benefit study has the advantage of allowing comparisons of very different treatments, such as a vaccination program for children vs. a hip replacement program for the elderly, because the benefits of each approach are measured using a single metric: dollars.

Clinical Study Design

The distinguishing characteristic of clinical studies is that they begin with subjects who have a particular diagnosis. Clinical study designs include **natural history studies**, randomized clinical trials, and N of 1 studies (*see* **Crossover Designs**). The natural history study focuses on analyzing prognosis in subjects who have a particular disease. Using the natural history design, one can evaluate the effects of treatments including important and common adverse events. One can also identify persons at high risk of experiencing a poor disease course who might be appropriate subjects for more aggressive treatment and other subjects who experience a benign disease outcome in whom aggressive treatment may not be indicated. In addition, data on prognosis can be utilized to develop "predictive models" identifying which patients might benefit from extensive, costly clinical evaluation (*see* **Predictive Modeling of Prognosis**). Thus natural

history studies of people with a particular disease provide valuable diagnostic and therapeutic information that can guide clinical decisions.

Clinical epidemiologists have identified several critical methodologic concerns in performing a natural history study. These concerns are similar to those of epidemiologic **cohort studies**, the main differences being that in natural history studies subjects already have disease. The **controls** in a natural history study are internal controls, and comparisons are made between different subsets of people with disease. Natural history studies are most accurate in so-called inception cohorts, in which patients are entered into the study just after diagnosis. If patients are entered long after diagnosis, patients who die or go into remission soon after diagnosis may be missed, thus biasing the study. Right censoring, or loss to follow-up of patients enrolled in the natural history study, must be avoided to give an accurate picture of the prognosis of disease. If dropouts tend to have poor prognosis, then the study would arrive at an inaccurately optimistic prediction of disease prognosis.

Prognostic information can be used to build prediction rules that will estimate patient prognosis based on the presence or absence of key clinical characteristics. To develop prediction rules, investigators generally study two different groups of subjects with disease. Using the first group, they develop the rule identifying those factors which affect prognosis. Using the second group, they test this prediction rule, attempting to confirm that the factors identified in the first sample generalize to the second. Such repeatability of a prediction role in two independent samples augurs well for the general applicability of this rule to yet other patient samples.

Summary

Clinical epidemiology is a heterogeneous and dynamic field in which methodologies drawn from epidemiology, economics, and psychometrics are applied to issues related to clinical measurement and clinical decision making. This applied science will undoubtedly continue to grow and become even more important as clinical scientists apply additional methodologies to the analysis of clinical problems.

Several general references on clinical epidemiology are listed in the Bibliography.

Further Reading

Fletcher, R.H., Fletcher, S.W. & Wagner, E.H. (1996). *Clinical Epidemiology: The Essentials*, 3rd Ed. Williams & Wilkins, Baltimore.

Friedland, D.L., Go, A.S., Davoren, J.B., Shlipak, M.G., Bent, S.W., Subak, L.L., Mendelson, T. (1998). *Evidence-Based Medicine: A Framework for Clinical Practice*. Appleton & Lange, Stamford.

Murphy, E.A. (1976). *The Logic of Medicine*. Johns Hopkins University Press, Baltimore.

Petitti, D.B. (2000). *Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. 2nd Ed. Oxford University Press, New York.

Sackett, D.L., Haynes, R.B., Guyatt, G.H. & Tugwell, P. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*, 2nd Ed. Little, Brown, & Company, Boston.

Spilker, B. (1996). *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed. Lippincott-Raven, Philadelphia.

Weiss, N.S. (1996). *Clinical Epidemiology: The Study of the Outcomes of Illness*, 2nd Ed. Oxford University Press, New York.

ROBERT F. MEENAN & DAVID T. FELSON

Clinical Signals

Clinical signals are obtained when human subjects are monitored, usually by measuring electrical activity in a particular part of the body. Typical signals are the electrical activity measured from the brain (EEG) and from the heart (ECG or EKG). Other signals are muscle activity (EMG), stomach (EGG), and blood pressure from an indwelling cannula. Signals not measured electronically include lung function (*see* **Pulmonary Medicine**), levels of electrolytes in the blood, and clinical symptoms measured at regular intervals such as hourly or daily.

Time series methods are applied routinely to analyze clinical signals in hospitals and they can often provide life-saving information. Many measurements vary considerably over short periods of time and it is in fact this variation that is indicative of health or sickness, and not the absolute level of the measurement. As examples, it is the variation in Peak Expiratory Flow Rate (a measure of lung function) which is diagnostic of asthma, not the absolute level and Heart Rate Variability (HRV) is used to diagnose fetal distress during labor.

The literature on the analysis of time-varying clinical signals has been dominated by engineers, who have developed their own methodology and literature in parallel with the statistical version, and sometimes the engineering methods lack an underlying statistical model, which can make dialog between the two disciplines difficult. A discussion of the different approaches taken by statisticians and engineers has been given recently [3].

Spectral Analysis

Perhaps the most common feature of the analysis of clinical signals is to look for regularly occurring features or rhythms. For humans to maintain stable bodily functions, clinical signals must be constrained to lie within certain limits and this tends to make patterns within signals recur regularly. A discussion of the role of rhythms in homeostasis has been given by Hyndman [15]. Rhythms often result from nonlinear feedback loops. A simple example will illustrate the point. To remain healthy, humans must maintain blood pressure to within certain narrow limits. Blood pressure is mediated through the baroreceptors, located in the wall of the aortic arch and in the wall of the carotid sinus. If blood pressure is too high, then signals from the baroreceptors result in vasodilation which drops the blood pressure. If pressure is too low, then vasoconstriction occurs to increase the blood pressure. The feedback mechanism is thought to be nonlinear, and incorporates a delay, and for these reasons, at rest these rhythms can occur spontaneously [16].

Spectral analysis involves decomposing a signal into individual frequency components where the amplitude of these components is proportional to the "energy" of the signal at that frequency [26]. It is a convenient method for summarizing a long time series and is a natural procedure if we believe there are rhythms in the data. Spectral analysis is the method of choice for the analysis of clinical signals.

Sampling

Some signals are essentially continuous, whereas others are discrete. For example, the heart rate is measured from surface electrodes on the chest from the ECG. Although the ECG is continuous, the heart rate is usually derived from the "R" wave in the ECG, which is a sudden spike just preceding the ventricular contraction. Thus, the heart beat signal is essentially a **point process**. Some authors have analyzed the interbeat intervals, thus arriving at a spectrum which estimates frequencies per beat, rather than per unit time. Others sample the heart rate (or RR interval) signal at regular intervals or filter the point process to produce a continuous signal which can be sampled [2, 12].

Spectral Analysis and Time-Dependent Spectra

The problem with spectral analysis is that it assumes that the signal is **stationary**. However, medical signals are not stationary in the usual sense. They contain rhythms that may come and go in the time interval, the frequencies may vary, or amplitudes of cycles at certain frequencies increase or decrease. Spectral analysis considers the entire time interval and so cycles that only occur in part of the interval will have their spectral peaks attenuated by the low power in other parts of the interval. One common solution is to assume that over a short interval the signal is stationary and so to split the time period

into nonoverlapping intervals and calculate the spectrum for each interval. These spectra can be displayed in a pseudo three-dimensional plot, with the time axis running into the page [30]. The difficulty here is that it is not realistic to think of a signal being stationary in sections. A better intuitive model is one in which the signal "evolves" slowly so that the nonstationary component is slow in comparison with the signal in which we are interested.

Priestley [25, 26] pointed out that a sine wave in which the amplitude changes over time will have Fourier components at *all* frequencies. However, it would seem sensible to consider it as having a single frequency, with a time-varying amplitude. This concept led him to define "semi-stationary" signals, which have an oscillatory form, and a method of estimating their time-dependent spectrum known as the "evolutionary spectrum". This method has not featured in applications, possibly because it is not widely known.

Many applications devise a joint function of time and frequency and obtain a distribution that will describe the intensity of a signal simultaneously in time and frequency. A method that originated in physics is the Wigner or Wigner–Ville distribution. Let $y(t)$ be a continuous signal dependent on time t. The Wigner estimate of the spectrum at frequency ω is

$$f_w(\omega) = \int_{-\infty}^{\infty} y(t - 0.5\tau)\, y(t + 0.5\tau) \exp(-i\omega\tau)\, \mathrm{d}\tau.$$

This distribution has been used extensively [7, 22]. However, in the raw form it has some undesirable properties. It lacks any physical interpretation in terms of energy and may take negative values for certain processes. If a signal contains two harmonic components, then the Wigner distribution will contain interference at frequencies between the two. Smoothing the distribution improves its properties and computationally it resembles Priestley's evolutionary spectrum, a relationship discussed by Hammond et al. [14].

A popular method of estimating time-dependent spectra is via a fitted autoregressive model (*see* **ARMA and ARIMA Models**). If we can assume that the data are generated by a finite autoregressive process of order p, then

$$y(t) + \alpha_1 y(t-1) + \cdots + \alpha_p y(t-p) = z(t),$$

where $\alpha_1, \alpha_2, \ldots, \alpha_p$ are constants and $z(t)$ is random noise with mean zero and variance σ^2.

Then the spectrum at frequency ω is given by

$$f_{\mathrm{AR}}(\omega) = \frac{\sigma^2}{\left|1 + \alpha_1 \exp(-i\omega) + \cdots + \alpha_p \exp(-i\omega p)\right|^2}.$$

Thus, a method of estimating the spectrum of a series would be to fit an AR(p) model to a section of data and insert the estimated parameters into the above equation. Contiguous sections will give successive spectra. In contrast to the usual window method of estimating the spectrum, which can be thought of as a "local smoothing" method, the AR method is a form of "global smoothing" over all frequencies [26].

One feature of sinusoidal functions is that they best model signals in which a cycle persists throughout the interval. Some clinical signals are characterized better by bursts of energy, and these can be modeled by the newly emerging technique of wavelets. A link between wavelets and evolutionary spectral analysis has been given recently by Priestley [27]. Another form of spectrum, based on regularly occurring zeros and ones rather than a sinusoidal function, is known as the *Walsh spectrum*. It has certain computational advantages over the conventional spectrum, but is difficult to interpret. It has been used as a complementary tool to conventional spectral analysis [29].

Heart Rate and Blood Pressure Variability

Three major components are to be found in a typical heart rate spectrum [2, 13, 28] and these are also present in the blood pressure spectrum. A region of activity occurs at around 0.25 Hz, which is attributable to respiration (respiratory sinus arrhythmia) and this is thought to be a marker of vagal (parasympathetic) activity. A second component at around 0.1 Hz arises from spontaneous vasomotor activity within the blood pressure control system and is mediated by vagal and sympathetic activity [13]. A third, low-frequency component at around 0.04 Hz is thought to arise from thermoregulatory activity.

Other cycles that have been detected in the heart rate include one with a wavelength of about 90 minutes [23], corresponding to rapid eye movement in

sleep, and one supposedly relating to a "perception" cycle with a wavelength of about 10 minutes [9].

Spectral analysis has a number of clinical applications. It has been suggested that heart rate variability at particular frequencies relates to mental workload [11]. It has been shown that the heart rate may become "entrained", i.e. fluctuate in phase with regularly occurring tasks and used to measure their difficulty [4].

In diabetic neuropathy, vagal denervation causes a loss of spontaneous heart rate variability. This also may cause a reduction in the frequency at which vasomotor oscillations occur, possibly due to a decrease in conduction velocity [18]. Thus spectral analysis may be used to measure the severity of denervation. For example, Bianchi et al. [1] showed that spectral analysis could discriminate 21 diabetics with neuropathy from 19 without neuropathy.

The Electroencephalogram (EEG)

The EEG is electrical activity of the brain measured by electrodes at the surface of the skull. There is an immense amount of literature devoted to the spectral analysis of EEGs. In particular, six spectral peaks can be identified [19, 29]. These peaks, with a typical range of frequencies are: delta 1 (0.5–2.0 Hz), delta 2 (2.0–4.0 Hz), theta (4.0–8.0 Hz), alpha (8.0–12.0 Hz), sigma (12.0–14.0 Hz), and beta (14.0–20.0 Hz). The peaks can be used, for example to classify different levels of sleep. Jervis et al. [17] used Walsh and Fourier transforms to show that the EEG can discriminate between subjects with Huntingdon's Chorea and normal subjects. Recently there has been interest in describing neural processes in the context of nonlinear dynamics, and in particular deterministic chaos. For example, Palus [24] has suggested that the EEG has true randomness and is not chaotic (*see* **Chaos Theory**). This is a new and fast-emerging field.

Other Surface Electrical Signals

Electrodes can be used to measure surface potentials at a number of sites, and rhythms can be detected. The electrogastrogram (EGG) refers to the surface measurement of electrical activity of the stomach. Two sorts of activity can be detected: slow waves with a frequency of about three cycles per minute and spikes with a frequency of between 0.5 Hz and 1 Hz [5, 6].

For monitoring pregnancy and parturition, the uterine electromyogram (EMG) has been proven to be an efficient tool and Duchêne et al. [10] discussed the use of an autoregressive method and a smoothed Wigner distribution for estimating the spectrum.

Hormone Levels

Hormone of various kinds, including luteinizing hormone and growth hormone, are known to be released in a pusatile fashion, and it is the pulsating levels of the hormone that cause its action rather than the absolute level [20]. Using samples of blood taken every five minutes, luteinizing hormone has been shown to have a cyclic component with a frequency of about one cycle per hour. Murdoch et al. [21] showed peaks in five out of six women ranging from 51 minutes cycle length to 120 minutes and more rapid oscillations at 20 minutes and 13.3 minutes cycle length. These were also seen, but less strongly, in four other women. This methodology has recently been extended to enable the analysis of replicated series, so that a more global analysis can be carried out [8].

Conclusion

There are a wide variety of methods in use to carry out the analysis of biomedical signals. To some extent this is inevitable; different signals have different characteristics and the methods have to adapt to these. In applications there appears little unaminity as to optimum methods, many authors deriving their own spectral smoothing windows or updating algorithms *ab initio*.

References

[1] Bianchi, A., Bontempi, B., Cerutti, S., Gianoglio, P., Comi, G. & Sora M.G.N. (1990). Spectral analysis of heart rate variability signal and respiration in diabetic subjects, *Medical and Biological Engineering and Computing* **28**, 205–211.

[2] Campbell, M.J. (1983). Spectral analysis applied to physiological signals in human subjects, *Bulletin of Applied Statistics* **10**, 175–193.

[3] Campbell, M.J. (1996). Spectral analysis of clinical signals: an interface between medical statisticians

and medical engineers, *Statistical Methods in Medical Research* **5**, 51–66.

[4] Charnock, D.M. & Manenica, A. (1978). Spectral analysis of R–R intervals under different work conditions, *Ergonomics* **21**, 103–108.

[5] Chen, J.D.Z., Stewart, W.R. & McCallum, R.W. (1993). Spectral analysis of episodic rhythmic variations in the cutaneous electrogastrogram, *IEEE Transactions on Biomedical Engineering* **40**, 128–135.

[6] Chen, J., Vandewalle, J., Sansen, W., Vantrappen, G. & Janssen, J. (1990). Adaptive spectral analysis of cutaneous electrogastric signals using autoregressive moving average modelling, *Medical and Biological Engineering and Computing* **28**, 531–536.

[7] Cohen, L. (1989). Time frequency distributions – a review, *8 Proceedings of the IEEE* **77**, 941–981.

[8] Diggle P.J. & al Wasel, I. (1997). Spectral analysis of replicated biomedical time series, *Applied Statistics* **46**, 31–71.

[9] Doust, J.S.L. (1978). A free running endogenous rhythm of the resting heart rate in man, *Canadian Journal of Physiology and Pharmacology* **56**, 83–86.

[10] Duchêne, J., Devedeux, D., Mansour, S. & Marque, C. (1995). Analyzing uterine EMG: tracking instantaneous burst frequency, *IEEE Engineering in Medicine and Biology* **March/April**, 124–140.

[11] Ettema, J.H. & Zielhuis, R.L. (1971). Physiological parameters of mental load, *Ergonomics* **14**, 137–144.

[12] French, A.S. & Holden, A.V. (1971). Alias free sampling of neuronal spike trains, *Kybernetik* **8**, 165–171.

[13] Furlan, R., Guzzetti, S., Crivellaro, W., Dassi, S., Tinelli, M., Baselli, G., Cerutti, S., Lombardi, F., Pagani, M. & Malliani, A. (1990). Continuous 24-hour assessment of the neural regulation of systemic arterial pressure and RR variabilities in ambulant subjects, *Circulation* **81**, 537–547.

[14] Hammond, J.K., Harrison, R.F., Tsao, Y.H. & Lee, J.S. (1993). The prediction of time-frequency spectra using covariance-equivalent models, in *Developments in Time Series Analysis. In Honour of Maurice B. Priestley*, T. Subba Rao, ed. Chapman & Hall, London.

[15] Hyndman, B.W. (1974). The role of rhythms in homeostasis, *Kybernetik* **15**, 227–236.

[16] Hyndman, B.W., Kitney, R.I. & Sayers, B.McA. (1971). Spontaneous rhythms in physiological control systems, *Nature* **233**, 339–341.

[17] Jervis, B.W., Coellio, M. & Morgan, G.W. (1989). Spectral analysis of EEG responses, *Medical and Biological Engineering and Computing* **27**, 230–238.

[18] Kitney, R.I. & Rompelman, O. (1987). New trends in the application of heart-rate variability analysis, in *The Beat-by-Beat Investigation of Cardiovascular Function*, R.I. Kitney & O. Rompelman, eds. Clarendon Press, Oxford.

[19] Larsen, H. & Lai, D.C. (1980). Walsh spectral estimates with applications to the classification of EEG signals, *IEEE Transactions on Biomedical Engineering* **27**, 485–492.

[20] Lincoln, D.W., Fraser, H.M., Lincoln, G.A., Martin, G.B. & McNeilly, A. (1985). Hypothalmic pulse generators, *Recent Progress in Hormone Research* **41**, 369–419.

[21] Murdoch, A.P., Diggle, P.J., Dunlop, W. & Kendall-Taylor, P. (1985). Determination of the frequency of pulsatile luteinizing hormone secretion by time series analysis, *Clinical Endocrinology* **22**, 341–346.

[22] Novak, P. & Novak, V. (1993). Time/frequency mapping of the heart rate, blood pressure and respiratory signals, *Medical and Biological Engineering and Computing* **31**, 103–110.

[23] Orr, W.E. & Hoffman, J.H. (1974). A 90 min biorhythm: methodology and data analysis using modified periodograms and complex demodulation, *Transactions on Biomedical Engineering* **21**, 130–143.

[24] Palus, M. (1996). Nonlinearity in normal human EEG: cycles, temporal asymmetry, nonstationarity and randomness, not chaos, *Biological Cybernetics* **75**, 389–396.

[25] Priestley, M.B. (1965). Evolutionary spectra and non-stationary processes, *Journal of the Royal Statistical Society, Series B* **27**, 204–237.

[26] Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

[27] Priestley, M.B. (1996). Wavelets and time-dependent spectral analysis, *Journal of Time Series Analysis* **17**, 85–105.

[28] Sayers, B.McA. (1973). Analysis of heart rate variability, *Ergonomics* **16**, 17–32.

[29] Thakar, N.V., Guo, X., Vaz, C.A., Laguna, P., June, R., Caminal, P., Rix, H. & Hanley, D.F. (1993). Orthonormal (Fourier and Walsh) methods of time-varying evoked potentials in neurological injury, *IEEE Transactions on Biomedical Engineering* **40**, 213–221.

[30] Van der Schee, E.J. & Grashuis, J.L. (1987). Running spectrum analysis as an aid in the representation and interpretation of electrogastric signals, *Medical and Biological Engineering and Computing* **25**, 57–62.

MICHAEL J. CAMPBELL

Clinical Significance Versus Statistical Significance

Study design, interpretation, and reporting often ignore the distinction between clinical significance and statistical significance. Statistical significance refers to whether or not the value of a statistical test exceeds some prespecified level. Clinical significance refers to the medical importance of a finding. The two often agree, but not always. The clearest example involves a study (either an **observational study** or an interventional study (*see* **Clinical Trials, Overview**)) that has a large number of participants. Statistical significance may be observed for small associations between exposures and disease or condition (in the case of observational studies) or differences between interventions (in the case of intervention studies). Whether or not these associations or differences are clinically significant or meaningful depends on the seriousness of the condition being studied, the prevalence of the condition, and the other benefits and risks of the intervention. An intervention effect shown in a clinical trial to be statistically significant and in the beneficial direction, but of small magnitude, may be clinically significant if the intervention is relatively nontoxic, easily administered, and useful for a condition that is of public health importance. Conversely, an intervention effect shown to be statistically significantly superior to control may not be clinically significant if the intervention has unacceptable adverse effects.

Sometimes, a combination of clinical events is used as the primary outcome of a study (*see* **Multiplicity in Clinical Trials**). The result may be statistically significant. Unless the combined outcome makes scientific or clinical sense, however, it will not be clinically relevant or significant. Clinical relevance may be achieved either because the outcome combines events that presumably reflect a common mechanism of action of an intervention or a risk factor or enables a clinician to summarize readily the effect of the intervention or risk factor.

In addition to the selection of the outcome (*see* **Outcome Measures in Clinical Trials**), clinical significance enters into the design of a trial when the expected size of the effect of the intervention and the choices of type I and type II error rates (*see* **Hypothesis Testing**) are determined. A study should be large enough to detect a clinically important difference (*see* **Sample Size Determination for Clinical Trials**). If the difference turns out not to be statistically significant, **confidence intervals** should be calculated to determine whether a clinically important difference has been excluded.

It should also be noted that the criterion for statistical significance is commonly, by convention, $P < 0.05$, though it need not be (*see* ***P*** **Value**). Clinical significance has no such convention; after the study is completed the interpretation is often individual, and will be viewed differently by different investigators, physicians, or other practitioners, and patients.

The concepts of **relative risk**, **absolute risk**, and population **attributable risk** are relevant to clinical significance. A clinical trial that yields a relative risk reduction of 30% may mean a change from a 60% event rate to 40%, or from 6% to 4%. Depending upon the size of the sample, either may be statistically significant. But also depending upon the severity of the disease and the nature and cost of the intervention, including frequency of adverse effects, the absolute reduction of 2%, from 6% to 4%, may not be clinically significant. If the condition is common and the intervention is simple, however, this 2% may be an important reduction in population attributable risk, resulting in the prevention of many clinically serious events.

The example of blood pressure reduction by nonpharmacologic means is instructive. Relatively small (3–4 mm Hg for systolic; less for diastolic), yet statistically significant, reductions in blood pressure have been obtained by means of weight or dietary sodium reductions. Although some patients will achieve greater reductions, this average reduction observed in clinical trials is frequently not seen as clinically meaningful for an individual hypertensive patient or the treating physician. However, from a public health standpoint, it is exceedingly important, as it is brought about without the need for drug treatment and translates into many thousands fewer strokes and myocardial infarctions per year in a population as large as that of the US.

Studies that are designed with appropriate **power** to detect clinically important differences in **binary** outcomes commonly also evaluate other outcomes that may be continuous variables. Examples are laboratory or physiologic measures, or quality of life

assessments. If these continuous variables are measured in all subjects, small, perhaps clinically meaningless, differences may be statistically significant. An example is a trial of a few thousand participants assessing occurrence of a clinical event as the primary outcome. Instruments measuring quality of life may be incorporated into this trial. Because quality of life is assessed in all participants, a difference of a few points on a scale that has a 40 or more point range will be statistically significant but probably not clinically significant.

Can there be clinical significance in the absence of statistical significance? Clinicians are always making judgments in the absence of optimal information. Many treatments are used, although a clinical trial may not be definitive, or may not even have been conducted, because a physician believes that the intervention is likely to be more beneficial than harmful. In addition, no clinical trial stands by itself. The results are always interpreted in light of other research (both basic and applied) findings, including other trials and observational studies. Sometimes, the totality of the information may be sufficiently persuasive, even though an individual trial may not show statistical significance, and perhaps even a **meta-analysis** may fail to do so. The judgment is made that the evidence for use is adequate, given relatively low toxicity of the intervention or the existence of a serious condition or disease. For example, the association of elevated serum cholesterol and ischemic heart disease has been well known for many years. Despite the fact that until recently there had not been evidence from clinical trials that cholesterol lowering was unambiguously beneficial, many clinicians acted as if that were the case. They may have prescribed lipid-lowering drugs only for patients at the highest risk, but recommended dietary changes for many more. This example points out that, when such judgments are made, clinicians are more comfortable employing interventions perceived to have low risk, relative to the condition being treated. Thus, dietary changes could be safely recommended even though trials of dietary intervention have not been conclusive, whereas drug therapy would await stronger evidence. Similarly, if the condition is life-threatening and there is no good standard therapy, as with some cancers, physicians often determine that a treatment merits consideration, even in the absence of statistically significant evidence.

A related issue concerns decisions based on **treatment–covariate interactions**, or subgroup findings. Overall, in a clinical trial, there may be a statistically significant beneficial finding, but even without clear evidence of interactions, clinicians may decide that the intervention should be reserved for those patients at the greatest risk. Alternatively, there may be a statistically nonsignificant trend overall in favor of an intervention, but an apparently large benefit in one or more subgroups of patients. If these results are consistent with other studies, they may be seen as evidence to use the intervention clinically in those subgroups. The reasonableness of such a course depends on the persuasiveness of the ancillary research data.

For a somewhat different view of statistical and clinical significance, see Feinstein.

Further Reading

Armitage, P. & Berry, G. (1994). *Statistical Methods in Medical Research*, 3rd Ed. Blackwell Science, Oxford, pp. 96, 98–99, 506.

Feinstein, A.R. (1985). *Clinical Epidemiology: The Architecture of Clinical Research*. W.B. Saunders Company, Philadelphia, pp. 396–406.

Fletcher, R.H., Fletcher, S.W. & Wagner, E. (1996). *Clinical Epidemiology: The Essentials*, 3rd Ed. Williams & Wilkins, Baltimore, p. 190.

Lindgren, B.R., Wielinski, C.L. Finkelstein, S.M. & Warwick, W.J. (1993). Contrasting clinical and statistical significance within the research setting, *Pediatric Pulmonology* **16**, 336–340.

Roberts, C.J. (1977). *Epidemiology for Clinicians*. Pittman Medical Publishing Company, Tunbridge Wells, pp. 162–165.

LAWRENCE M. FRIEDMAN

Clinical Trials Audit and Quality Control

A **multicenter** randomized clinical trial is a complex undertaking, requiring cooperation among a diverse group of participants to achieve a successful result. Patients who agree to participate in a trial must be properly registered and randomized to one of the available treatments (*see* **Randomized Treatment Assignment**), and data on baseline patient characteristics, **eligibility and exclusion** requirements, adherence to treatment (*see* **Compliance Assessment in Clinical Trials**), adverse events, laboratory or other measurements, and clinical **outcome measures** must be collected and analyzed. Throughout this process, there is the possibility of errors arising in the data, ranging from honest mistakes to sloppiness or, rarely, from deliberate fraud (i.e. fabrication or falsification of data). Because of this complexity, one of the important hallmarks of a successful trial is a well-developed system for data quality control (QC) and auditing [3, 5, 6, 14].

The purpose of a data QC system is to provide reasonable assurance to the organizers of the trial as well as to the "consumers" of the results that the data on which the conclusions are based are reliable. It is unreasonable to attempt to detect and eliminate all errors in the data. A small percentage of errors will not materially affect the scientific conclusions of a well-designed and well-conducted clinical trial. Also, the data QC system itself can be a major cost component of the trial, and there is a law of diminishing returns in the attempt to lower the error rate toward zero. On the other hand, it is also unreasonable to be unconcerned with data QC. Even if the effect of a small amount of data errors on scientific conclusions is minimal, the effect of discovered errors in the data on public perception and external acceptance of the results can be profound.

Why do we do data QC at all? There are at least four reasons.

1. *Scientific validity*. It is theoretically possible that systematic or random data errors may be of a sufficient magnitude to threaten the primary scientific conclusions of the trial. Thus, a primary function of data QC procedures is to ensure that the nature and magnitude of data errors are within acceptable limits.
2. *Prevention of future errors*. Data QC procedures, applied early in a trial, can serve as an educational tool to prevent future errors in the trial. Indeed, the feedback from early detection of errors can be an extremely important aspect in ensuring high-quality data.
3. *Public confidence*. The ability of a clinical trial to affect clinical practice depends on many factors, a key one of which is public confidence in the integrity of the trial process. Unfortunately, well-publicized cases of data problems can have a major negative impact on public confidence, far beyond that justified from a purely scientific viewpoint.
4. *Product licensing*. There are expectations and requirements with respect to data quality from regulatory agencies (e.g. the **Food and Drug Administration** in the US) in those trials supporting product or device licensing applications. For such trials, the need for data QC procedures is obvious [15].

In the following sections, the clinical trials data flow process is described, the primary targets for a QC program are identified, general clinical trials QC procedures are described, useful statistical QC procedures are discussed, site monitoring and audits are described, and the costs of QC procedures are assessed.

The Clinical Trials Data Flow Process

A simplified conceptual framework for considering the data flow process in a clinical trial is given in Figure 1, where there are three steps leading from the patients to the analysis of the trial. In the first

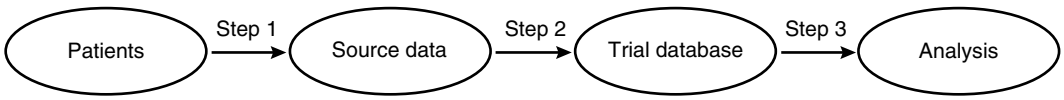

Figure 1 Data flow in a clinical trial

step, physical measurements on patients, laboratory tests, demographic data, and other patient data result in some "source data records" such as those in the official medical records or in some clinical or research laboratory database. In the second step, some subset of the source data is sent via a case report form (CRF) or in an electronic format to a centralized trial database. In some cases (e.g. questionnaire data), these two steps are combined and the data flow is directly from the patient to the trial database, and there are effectively no "source data" against which the trial database can be checked in an audit (*see* **Data Management and Coordination**). The third step in this process is the analysis step, in which various statistical procedures are applied to the trial database.

Data errors can arise at each step of the process. Thus, at step one, the source data may not reflect the "truth" of patient status or response (e.g. the birth date in the source data may not be correct if there was an error in recording it or, even, because of misrepresentation by the patient). At step two, the trial data may not be the same as the source data (in the birth date example this could be caused by an error in data entry from the CRF). At step three, there could be a programming error so that the analysis does not properly reflect the trial database. Each of these potential sources of error requires attention, but the types of QC procedures are quite different at each step. In the remainder of this article, attention will be focused primarily on step two, on data QC procedures aimed at ensuring that the trial database is an accurate reflection of the source data.

Primary Areas Targeted for QC

There are several areas that should receive special attention in any clinical trials QC plan. These are:

1. *Registration and randomization*. The process of registration and randomization of patients is of fundamental importance and must be tightly controlled.
2. **Eligibility and exclusion criteria**. The data required for assessing patient eligibility must be carefully collected and validated. For scientific reasons and for easing complexity, it is desirable to make the eligibility criteria as simple and broad as possible [7].
3. *Baseline patient characteristics*. Data available on patients at registration are often used as prognostic factors, as **stratification** factors, or for other reasons.
4. *Treatment delivery and* **compliance**. Despite the widely accepted principle of **intention-to-treat** in the statistical analysis of randomized trials, some idea of the extent of treatment delivery and compliance is desirable in interpreting the results.
5. *Response and toxicity*. Knowledge of the effect of treatments on clinical response and on the toxicity or other adverse events is essential. The clinical response is often a primary outcome measure in the trial.
6. *Laboratory values*. Laboratory data, from both clinical and research laboratories, are important in the analysis of many clinical trials [4, 18].
7. *Follow-up*. Long-term outcomes and survival are important in the analysis of most clinical trials. Indeed, in most serious chronic diseases (e.g. cancer), long-term follow-up data are the primary focus of the trial.

General QC Principles

One of the more important principles of any data QC plan is that the trial policies and procedures should be written down beforehand in a Standard Operating Procedures (SOP) document or equivalent, which should be sufficiently detailed to enable an external reader to be able to assess the data QC plan. The SOP document is also important in QC training and documentation of QC procedures at all levels. A well-trained staff following clearly articulated QC procedures provides strong protection against serious data problems. The prevention or early detection of data problems in a clinical trial is an effective way to avoid later, and very expensive, corrective measures. Indeed, some data errors cannot be corrected later.

The independence of the statistical center from the trial organizers and participants will help to ensure that QC procedures are above reproach. For example, the registration and randomization process is so important to the integrity of the trial that it must be carried out in an exemplary fashion, preferably by an independent statistical center or its equivalent.

Case report forms should be as simple as possible. There is often a temptation to include data that are not essential to the primary objectives of the trial. This temptation should be strongly resisted, since

it inevitably adds to the expenses of the trial and dilutes the QC procedures aimed at the key data. The data entry or data acquisition procedures need careful attention. For data entered into the trial database by data entry staff, one standard procedure to minimize data entry errors is independent double data entry [2]. This procedure reduces the number of data entry errors but only at increased cost. For automated data acquisition, procedures for validating the data prior to loading into the trial database are important.

Automated data editing procedures, including range and validity checks, as well as cross-field consistency checks, are an important part of any data QC program [8]. These procedures are usually deterministic checks on the data. Statistical procedures are discussed in the next Section. Range checks are of the form $a < x < b$, where a and b are prespecified constants, and x is the variable value. If x is outside the indicated interval, then a flag is raised indicating that the value x must be checked. A validity check is a test to ensure that a variable has a valid value. For example, a discrete variable must take one of a specified finite set of values. Cross-field consistency checks are checks of required relationships among two or more variables. These can be of the form $x < y$ for numerical or date variables or can simply reflect impossible configurations (e.g. females with prostate cancer).

A program of on-site monitoring and on-site auditing is important for ensuring that data in the trial database accurately reflect the source data [11]. Unfortunately, this aspect of the data QC plan is usually extremely costly, so careful thought is needed in designing this part of the plan. A separate Section below deals with monitoring and auditing in more detail.

Statistical QC Procedures

There are many traditional statistical procedures that are useful in a data QC plan [10, 12]. These are also relatively inexpensive to implement, especially compared to on-site visits to participating centers or other labor-intensive procedures, and thus should be employed extensively. Standard univariate and multivariate **outlier** detection techniques are useful (*see* **Multivariate Outliers**). The identified outliers can then be verified. Of course, unless some documented error is detected, no outlier should be deleted.

Another type of analysis that is important in multicenter studies is an analysis by center. This analysis can be as simple as an analysis of primary outcome to look for unusual discrepancies among centers. It is also important to look at bivariate plots of important data by center to spot unusual patterns. Such an analysis led to the detection of one case of scientific misconduct [1]. A more common result of this type of analysis would be to identify and correct some systematic, but unintentional, data problem at one or more centers.

Statistical random sampling schemes (*see* **Probability Sampling**) are an important part of industrial QC procedures, and these schemes can be applied profitably to data QC in clinical trials. For example, a periodic random resampling of a small percentage of records in the trial database and a check of selected fields in these records against the CRFs can ensure that the overall process is working well. Random sampling is also a key part of the auditing process described below.

On-Site Monitoring and On-Site Audits

In any multicenter clinical trial, the overall quality of the data is fundamentally dependent on the quality of data from the individual centers [9, 16]. Thus, there must be education and training of the key personnel at each center (physicians, nurses, data managers, clinical research associates, and others) and some procedure to compare the trial database with the source data from these centers. In principle, all of these activities could be conducted without visits to the individual centers, but on-site visits for these purposes are clearly preferable, although extremely expensive [17, 19]. In large simple trials, it may be impossible to visit the individual centers, whereas in some product licensing trials, extensive on-site monitoring and on-site auditing procedures at every center may be employed. Most trials fall between these extremes.

Cost–Benefit Analyses

It is a common complaint among clinical trialists that costs of data QC procedures are excessive relative to other costs of the trial and to the expected benefits. Surprisingly little has been written on this important topic [13]. As in other areas of economic analysis, there is clearly a law of diminishing returns in effect as additional QC procedures are added to reduce the

data error rate. The problem lies in deciding where to draw the line.

It seems clear from the earlier considerations that automated data editing procedures and statistical QC procedures are likely to have a large benefit for a relatively low cost. However, these procedures are largely ineffective in assuring that the trial database matches the source data. On-site monitoring and on-site auditing is almost certain to be the most expensive part of the QC procedures, but is a primary method for checking the source data.

Summary

Data quality control procedures are an important part of any well-designed and well-conducted clinical trial. Automated data editing and statistical QC procedures offer major benefits for relatively low costs. More expensive procedures such as on-site monitoring and on-site auditing may be essential to identify certain types of data errors, but their benefits relative to their costs have not been well studied.

References

[1] Bailey, K.R. (1991). Detecting fabrication of data in a multicenter collaborative animal study, *Controlled Clinical Trials* **12**, 741–752.

[2] Blumenstein, B.A. (1993). Verifying keyed medical research data, *Statistics in Medicine* **12**, 1535–1542.

[3] Buyse, M. (1984). Quality control in multi-centre cancer clinical trials, in *Cancer Clinical Trials Methods and Practice*, M.E. Buyse, M.J. Staquet & R.J. Sylvester, eds. Oxford University Press, Oxford, pp. 102–123.

[4] Dent, N.J. (1991). European good laboratory and clinical practices: their relevance to clinical pathology laboratories, *Quality Assurance* **1**, 82–88.

[5] Duchene, A.G., Hultgren, D.H., Neaton, J.D., Grambsch, P.V., Broste, S.K., Aus, B.M. & Rasmussen, W.L. (1986). Forms control and error detection procedures used at the Coordinating Center of the Multiple Risk Factor Intervention Trial (MRFIT), *Controlled Clinical Trials* **7**, 34S–45S.

[6] Gassman, J.J., Owen, W.W., Kuntz, T.E., Martin, J.P. & Amoroso, W.P. (1995). Data quality assurance, monitoring, and reporting, *Controlled Clinical Trials* **16**, 104S–136S.

[7] George, S.L. (1996). Reducing patient eligibility criteria in cancer clinical trials, *Journal of Clinical Oncology* **14**, 1364–1370.

[8] Karrison, T. (1981). Data editing in a clinical trial, *Controlled Clinical Trials* **2**, 15–29.

[9] Knatterud, G.L. (1981). Methods of quality control and of continuous audit procedures for controlled clinical trials, *Controlled Clinical Trials* **1**, 327–332.

[10] Liepins, G.E. & Uppuluri, V.R.R. (1990). *Data Quality Control*. Marcel Dekker, New York.

[11] Mowery, R.L. & Williams, O.D. (1979). Aspects of clinic monitoring in large-scale multiclinic trials, *Clinical Pharmacology and Therapeutics* **25**, 717–719.

[12] Naus, J.I. (1975). *Data Quality Control and Editing*. Marcel Dekker, New York.

[13] Neaton, J.D., Duchene, A.G., Svendsen, K.H. & Wentworth, D. (1990). An examination of the efficiency of some quality assurance methods commonly employed in clinical trials, *Statistics in Medicine* **9**, 115–124.

[14] Pollock, B.H. (1994). Quality assurance for interventions in clinical trials. Multicenter data monitoring, data management, and analysis, *Cancer* **74**, 2647–2652.

[15] Ransom, C., Zamora, G. & Jones, L. (1995). The development and implementation of a quality assurance Master Audit Plan, *Quality Assurance* **4**, 80–82.

[16] Severe, J.B., Schooler, N.R., Lee, J.H., Haas, G., Mueser, K.T., Rosen, P., Shortell, D. & Shumway, M. (1989). Ensuring data quality in a multicenter clinical trial: remote site data entry, central coordination and feedback, *Psychopharmacology Bulletin* **25**, 488–490.

[17] Shapiro, M.F. & Charrow, R.P. (1989). The role of data audits in detecting scientific misconduct: results of the FDA program, *Journal of the American Medical Association* **261**, 2505–2511.

[18] Vantongelen, K., Rotmensz, N. & van der Schueren, E. (1989). Quality control of validity of data collected in clinical trials. EORTC Study Group on Data Management (SGDM), *European Journal of Cancer and Clinical Oncology* **25**, 1241–1247.

[19] Weiss, R.B., Vogelzang, N.J., Peterson, B.A., Panasci, L.C., Carpenter, J.T., Gavigan, M., Sartell, K., Frei, E., III & McIntyre, O.R. (1993). A successful system of scientific data audits for clinical trials. A report from the Cancer and Leukemia Group B, *Journal of the American Medical Association* **270**, 459–464.

STEPHEN L. GEORGE

Clinical Trials of Antibacterial Agents

This article focuses on comparative **clinical trials** of antibacterial agents in the treatment of acute infections caused by bacteria. Many of the issues also relate to trials of antifungal and antiviral agents, and to prophylactic trials (*see* **Prevention Trials**; **Vaccine Studies**). Specific issues of antibacterial agents in neutropenic patients are not addressed.

Clinical trials of antibacterials differ from other clinical trials in several ways. There exists a three-way **interaction** between the antibacterial agent, the patient, and the organism causing infection, and these interactions affect the approach to the design, analysis, and interpretation of results. In addition, antibacterial studies often include patients with potentially life-threatening conditions, where treatment must commence immediately. This leads to practical problems such as entry of ineligible patients and inappropriate treatment (*see* **Eligibility and Exclusion Criteria**).

Terminology

An organism causing an infection is termed a *pathogen*. Infections may be caused by more than one pathogen (a *polymicrobial infection*). The pathogen causes signs and symptoms of infection in the patient, e.g. raised temperature, and the aim of treatment is to eliminate both these clinical signs and symptoms as well as to eradicate the pathogen(s). *Susceptibility* of pathogens to treatments is an important consideration in clinical trials and is an indication of whether the pathogen is likely to be eradicated by the antibacterial drug. Susceptibility is usually measured by either minimum inhibitory concentrations (MICs) or zone sizes, and pathogens are classified as either *sensitive* or *resistant* to the antibacterial agent. It may not be possible to identify a bacterial cause for the infection before treatment commences, and patients are sometimes found to have a viral or fungal infection after entry, or even no infection at all; such patients are referred to as *misdiagnoses*.

The *clinical response* is the investigator's assessment of the patient's clinical outcome (*see* **Outcome Measures in Clinical Trials**) for the infection for which the patient entered the trial. Categories of cure and failure are used; sometimes also others, such as improvement, but categories are usually combined at the analysis stage to give a binary outcome of satisfactory or unsatisfactory. A response of indeterminate is used when a patient cannot be assessed, e.g. they are lost to follow-up. Responses are described in detail in various Guidelines [1–3].

To identify the pathogen causing infection, an appropriate sample must be taken from the patient, e.g. sputum in the case of bronchitis, cerebrospinal fluid (CSF) in the case of meningitis. Other data which should be recorded are: date of collection, source of sample, quantitative evaluation of pathogens for certain infections, e.g. urinary tract infections, and susceptibility to trial treatments. A *microbiological response* should be given for each pathogen isolated pretreatment. This will be *eradicated* if the pathogen is no longer present at the time of assessment; *persisted* if it is still present. If it is not possible to take a sample, e.g. for ethical reasons as in meningitis trials where it may not be desirable to obtain a further CSF sample, the response may be *presumed eradicated* or *presumed persisted*, on the basis of whether the patient is clinically satisfactory or not. A new organism appearing during the trial which requires treatment is called a *superinfection*. These responses are detailed in the Guidelines.

Sometimes a *"by patient" microbiological response* is given, which gives a summary of the response for pathogens within a patient; this is particularly useful in polymicrobial infections. It is derived from the pathogen microbiological response, as shown in Table 1.

Assigning responses is not always straightforward, and clear guidance must be given in the protocol of how each of the categories for each of the responses should be used.

Assessments of all three responses are made at the end of treatment and also after a suitable follow-up period to detect relapses, which are important because they may indicate that the infection was only suppressed.

Historical Development

The first set of Guidelines for the conduct of clinical trials with antibacterial agents was published by the **Food and Drug Administration (FDA)** in 1977 [5]. The British Society of Antimicrobial Chemotherapy

Table 1 "By patient" microbiological response

Pathogen response	"By patient" response	Response in analysis
All eradicated	Success	Satisfactory
All presumed eradicated	Presumed success	Satisfactory
One or more persisted	Failure	Unsatisfactory
One or more presumed persisted	Presumed failure	Unsatisfactory
Superinfection and all pretreatment Pathogens eradicated	Superinfection	Unsatisfactory

(BSAC) published guidelines in 1989 [3]. The Infectious Diseases Society of America, under a contract with the FDA, and a European Working Party produced general and disease-specific guidelines in 1992 and 1993 respectively [1, 2].

Typical Study Design

Phase III clinical studies (*see* **Clinical Trials, Overview**) of antibacterials are typically parallel group, randomized, **multicenter**, active-controlled. Placebo-controlled trials are rarely ethical because effective treatment is already available. A double-blind (*see* **Blinding or Masking**) design can be difficult for studies of intravenous or intramuscular formulations, for example some drugs effervesce. At a minimum, studies should always be assessor-blinded. Studies are relatively short; treatment rarely lasts for more than 10 days, with a 2–4 week follow-up.

The majority of studies aim to demonstrate overall **equivalence** because for most infections effective treatment is already available and it would not be ethical to conduct a trial where the expectation of success was less for one drug than for another (*see* **Ethics of Randomized Trials**). Patients who are anticipated to be at different risks of failure with the study treatments should be excluded from these trials. Hence, generally the aim is to show equivalence by proving that the new treatment is at least as good as the control treatment. The principles of equivalence trials are addressed by Jones et al. [6].

Statistical Analysis of Endpoints

No new statistical techniques have been specifically developed to analyze data from antibacterial trials. For analysis purposes, assessments are combined to give a **binary** response of satisfactory or unsatisfactory. Standard techniques for analyzing binary data are then used e.g. **odds ratio** or difference in proportions. Strata (e.g. centers) (*see* **Randomized Treatment Assignment**) and risk factors defined at the design stage may need to be explored.

Analyses are generally performed on the clinical response and on the "by patient" microbiological response. Analysis of the microbiological response for pathogens must be undertaken with care. If patients have more than one pathogen, then an analysis of all pathogens violates the assumption of independence. An analysis of each individual pathogen may be uninformative owing to the small numbers of each pathogen which are likely to be obtained, and the difficulty in predicting which types of pathogens will occur in a given study. Usually these data are only summarized and are not subjected to statistical analysis. The "by patient" microbiological response is a more useful endpoint for analysis, and can be more easily interpreted.

The analyses should be performed at the end of therapy at a minimum. At follow-up, there are often a large number of patients without a response as they are lost to follow-up, and in such situations, a statistical analysis will not be informative. A "last-value carried forward" approach, i.e. carrying forward the response at end of treatment to follow-up, can be very misleading particularly if many patients are lost to follow-up.

Response rates of zero or 100% within centers may cause complications with some methods of analysis, although exact methods are sometimes used in such situations.

Equivalence is usually defined in terms of a lower limit for the **confidence interval** for the difference between response rates, typically -10%, i.e. equivalence is demonstrated if the lower limit of the confidence interval is above -10%. This infers that for equivalence trials, the analysis should be performed

in terms of differences in proportions. This can create a dilemma for the statistician since, if other methods of analysis are used, e.g. odds ratio, this definition of equivalence cannot be applied. Often the approach taken to this problem is to present the analysis by difference in proportions, but other methods are used to check the **robustness** of this analysis.

Patient Groups for Analysis

All patients exposed to either trial drug should be included in the assessment of safety (*see* **Data and Safety Monitoring**). In the assessment of efficacy, the aim should be to analyze the comparative efficacy of the treatments whilst minimizing the potential for bias. The conventional method is to consider **intention-to-treat** (ITT) and per protocol (PP) groups. These represent all patients entering the study and all patients who meet the study criteria, respectively, although the interpretation of these can differ, as discussed below.

Efficacy analyses are usually performed on the patient groups shown in Table 2. The patients included in the bacteriological PP group will be a subset of the bacteriological ITT group, which will be a subset of the clinical ITT group. The clinical PP group will be identical to the bacteriological PP group if the clinical PP group requires patients to have microbiologically documented evidence of an infection.

In studies of antibacterial agents, some problems arise more frequently than in other types of studies; patients can be inadvertently misrandomized, they can also be randomized but not treated, and there can be misdiagnoses. There is no standard method for handling these patients in either the ITT or PP analyses, and the proposed methods should be defined in the protocol for each study. However, such patients are unlikely to affect the outcome or interpretation of an analysis unless there are significant numbers of them or the number is unbalanced between the treatment groups. In such cases the reasons should be fully explored and the interpretation of the analysis carefully considered.

Patients with an indeterminate response will be excluded from all analysis populations except ITT, where they should be classified as unsatisfactory. If there are a significant number of patients added to the ITT analysis as failures, or the number is unbalanced between the treatment groups, then the interpretation of the analysis again needs to be carefully considered.

The purest view for the PP group is to exclude all patients who do not meet the entry criteria of the trial. A more pragmatic approach is to agree which criteria are unlikely to affect the outcome of treatment and to retain such patients in the PP analysis. For patients who deviate during the trial, rules should be agreed on what constitutes a deviation serious enough to warrant exclusion. Two common problems that arise in these studies are patients given concomitant therapy to treat an infection and patients who are found to have resistant pathogens. In either case it is not always ethical to withdraw the patient from the trial if he/she appears to be responding to the trial treatment.

Table 2 Efficacy analysis groups

Name of analysis group	Response	Patient group
Clinical ITT	Clinical	All patients
Clinical PP	Clinical	All patients who meet the trial criteria. This may or may not require patients to have documented evidence of a bacterial infection
Modified ITT	"By patient" microbiological	All patients who have documented evidence of a bacterial infection
Bacteriological PP	"By patient" microbiological	All patients who meet the trial criteria and have documented evidence of a bacterial infection

ITT = intention to treat.
PP = per protocol.

Patients taking concomitant therapy are sometimes classed as clinical failures or as indeterminate; sometimes all such patients are excluded from the analyses and sometimes a more judgmental approach is taken on the basis of whether the therapy taken is expected to affect the pathogen causing the infection. Patients found to have resistant pathogens are often omitted from analyses and sometimes this is applied if the pathogen is resistant to either study treatment or only resistant to the study drug received.

There is still confusion over whether the ITT or the PP analysis should be primary. The Committee for Proprietary Medicinal Products guidelines [4] state that in studies designed to show superiority of one drug over another, the ITT analysis is usually more conservative than the PP, since the noncompliers (*see* **Compliance Assessment in Clinical Trials**) included in an ITT analysis dilute the overall treatment effect. Hence, since antibacterial studies are usually designed as equivalence studies the PP should be the primary analysis because the ITT is less conservative in showing equivalence. However, whenever possible equivalence should be demonstrated in both the ITT and PP analyses. If results differ, then this must be explained [6].

Choice of Sample Size

It is usual to base patient numbers (*see* **Sample Size Determination for Clinical Trials**) on the clinical response, although this depends on the aims of the trial and the particular infection under study. There are many methods available for calculating patient numbers using binary data, several of which are based on the normal approximation of **binomial** probability. Since the primary analysis is usually the PP analysis, patient numbers will need to be inflated based on the estimate of the proportion of patients who will contribute to the PP analysis. This will be based on previous experience in this area, or published data, and usually ranges between 65% and 95%. In studies for US registration, samples sizes may need to be sufficient to ensure sufficient numbers of particular pathogens.

Unresolved Problems

There are still a significant number of issues that remain unresolved in this area. There is no standard and accepted way to analyze the data. There is also no agreement on whether the ITT analysis or the PP analysis is primary, or whether 90% or 95% confidence limits should be used, although this partly stems from a historical lack of understanding of the principles of equivalence trials. There is no standard approach to handling patients in the PP and ITT analyses, particularly with respect to such deviations as patients taking concomitant antibacterial therapy.

Publications of antibacterial trials are generally good in their statistical content, although issues in equivalence trials are not sufficiently understood yet. Papers often lack sufficient detail in defining which patients have been included in which analyses. The approaches to handling deviations can differ widely between trials and also within trials.

References

[1] Beam, T.R., Gilbert, D.N. & Kunin, C.M. (1992). General guidelines for the clinical evaluation of anti-infective drug products, *Clinical Infectious Diseases* **15**, Supplement 1.

[2] Beam, T.R., Gilbert, D.N. & Kunin, C.M. (1993). *European Guidelines for the Clinical Evaluation of Anti-Infective Drug Products*. European Society of Clinical Microbiology and Infectious Diseases.

[3] British Society for Antimicrobial Chemotherapy (1989). The clinical evaluation of antibacterial drugs, *Journal of Antimicrobial Chemotherapy* **23**, Supplement B.

[4] Committee for Proprietary Medicinal Products (1994). *Efficacy Working Party, Note for Guidance*. European Commission.

[5] Food and Drug Administration (1977). *Guidelines for the Clinical Evaluation of Anti-Infective Drugs (Systemic) (Adults and Children)*. Publication FDA 77–3046. US Department of Health, Education and Welfare, Washington.

[6] Jones, B., Jarvis, P., Lewis, J.A. & Ebbutt, A.F. (1996). Trials to assess equivalence: the importance of rigorous methods, *British Medical Journal* **313**, 36–39.

C. SMITH

Clinical Trials Protocols

A **clinical trial** protocol serves several purposes. First and foremost the protocol serves as a guideline for the conduct of the trial. It describes in a clear and detailed manner how the trial is performed so that all investigators know the procedures. This is particularly important in **multicenter trials** where it can be difficult to ensure that all centers and investigators conduct the study properly.

A second purpose of the protocol is to procure funding for the trial. Any funding source, such as government, a pharmaceutical company, or a private foundation, will generally require a protocol on which it can judge the merit of the proposal. Since funding institutions have different protocol formats, it is essential that the investigator obtains and follows the required format.

In addition to review committees of funding institutions, committees at the local institution, such as research committees or institutional review boards, need to review the trial to ensure that the trial participants' rights and safety are adequately protected, that the trial is in compliance with all of the local institution's regulations, and that the trial is feasible at the institution. The protocol serves as the basis for the review by these committees.

A fourth purpose is to provide guidelines to the groups responsible for monitoring the trial. At the local institution this may be the research committee or an institutional review board. For large studies or multicenter studies, formal data monitoring and safety committees (*see* **Data Monitoring Committees**) are also usually established. For all of these committees, the protocol provides the background information against which they determine whether the trial is progressing satisfactorily and that the investigators are complying with the intended procedures.

Finally, the protocol serves as a historical document for the trial to which trial investigators or outside investigators can refer should questions arise after its completion.

The clinical trial protocol provides broad detail about the trial and does not include the fine details of the day-to-day operations. These fine details, such as the steps to be taken at each visit, how to complete study forms, and how data will actually be processed, are usually described in a separate document, a Manual of Operations. If the trial is testing a particular technique, such as a specific surgical or a behavioral modification technique, or is using a complicated rating scale as a study variable, the fine details of the technique or administering the scale are not usually included in the protocol, but are provided in a separate document called a Training Manual.

Although each funding institution may specify its own format for a protocol, there are a number of items nearly always included. These include an abstract, the clinical background, the purpose of the trial, the methods, ethics and safety issues, organization, the budget, and copies of the proposed study forms. Some general texts discussing protocol issues are [4, 5, 7] and [8].

Abstract

A protocol usually begins with a short abstract that states the purpose and significance of the study and provides the most pertinent details, including the patient population to be studied, the treatments to be tested, the study design, the primary outcome measure(s), the number of patients, and the duration of treatment and follow-up.

Background

An in-depth summary with relevant references to published work on the study topic is included to justify the need for the trial. Any unpublished work that the investigators have done on the subject is also described. If drugs are involved, then pertinent pharmacological and toxicity data are included. In addition, if any new or nonstandard methods, such as the use of specific **surrogate endpoints**, are used in the study, then information about the method is provided.

Purpose

The purpose of the trial and its current importance are described in clear and concise terms. For purposes of funding review, this section is used to sell the importance of the trial, but it should not promise more than can actually be obtained. Reasons for

the trial might include (i) to test a new treatment regimen, (ii) to test an established treatment for a new indication, (iii) to determine the best of a number of standard treatments, and (iv) to provide additional data on the safety or efficacy of a treatment regimen for approval by a regulatory agency.

Methods

The following items are usually addressed in the methods section as appropriate for the type of trial proposed.

Hypotheses

The hypotheses that the trial is designed to test are clearly specified. Although on occasion one needs to specify more than one primary hypothesis, the number should be few so that the main aims of the trial can be kept in focus. Any secondary hypotheses that the investigators want to test are also specified. Listing the secondary hypotheses in a protocol prior to the conduct of the trial lends more credence to the results of testing such hypotheses as it shows that the results are not the product of a "fishing expedition", in which large numbers of unlikely hypotheses are tested and only those that meet some usually inappropriate criteria for statistical significance are reported (*see* **Multiplicity in Clinical Trials**). Careful consideration of secondary hypotheses in the protocol also ensures the collection of the data necessary to answer them.

Patient Population

The population of patients entered into the trial is described in detail. This is usually done by specifying criteria for inclusion and/or exclusion. Inclusion criteria are those characteristics that the patient must have to be considered for inclusion in the trial. For example, if the trial compares drug treatments in elderly male, alcoholic patients, the inclusion criteria might specify that patients be (i) men, (ii) $\geq$65 years of age, with (iii) a diagnosis of alcoholism. Exclusion criteria are the characteristics of a patient that prevent entry into a trial even though all inclusion criteria are met. For example, if the primary endpoint for the study is measured two years after entry in the trial, then patients with a life expectancy of less than two years might be excluded from participation. When defining a patient population, it should be kept in mind that, although having a large number of inclusion and exclusion criteria may ensure a more homogeneous trial population, it may also lead to results that are less generalizable and can make patient recruitment more difficult (*see* **Eligibility and Exclusion Criteria**).

Treatment Regimens

The treatments under study are described in broad detail. For drug studies, the dose administered, the dosing regimen, and the duration of dosing are given. For surgical studies, broad details about the surgical procedures are usually given; the fine details are provided in a separate training manual if needed. In trials of medical devices, the use and maintenance of the device are described. Noninvasive interventions, such as psychotherapies, require broad information such as references to the technique, how it is administered, who administers it, how the interventionist is trained, how often and how long the intervention is given, and how the intervention is monitored. Specific details on the intervention should usually be described in an appendix or separate training manual.

Trial Design

Trial design items may include the following.

1. *Randomization*. Will the treatment assignment be randomized? If so, details of the randomization method are included. Any strata identified (e.g. gender), to ensure that treatment assignments are equally distributed over important prognostic factors, are described (*see* **Randomized Treatment Assignment**).
2. *Control groups*. If a control group is used, then the protocol should define the control and justify its use. If a placebo is used rather than a standard treatment, then safety considerations for participants assigned to the placebo must be addressed.
3. *Masking (blinding)*. Knowledge of the treatment assignment can lead to bias in reporting of some outcome measures. The protocol should describe who knows the patient's treatment assignment and who is masked. Measures to maintain the

masking are also specified (*see* **Blinding or Masking**).

4. *Experimental design*. The protocol indicates the experimental design used. While most studies are two-group parallel designs in which a patient is assigned to one of two treatment groups for the entire study, other designs such as **crossover designs** and factorial designs are sometimes used (*see* **Factorial Designs in Clinical Trials**).

Pre-Study Procedures

This portion of the protocol describes the process for the recruitment and selection of trial patients as well as the pre-study evaluations and procedures that patients have before entering the trial. The data collected pre-study are described.

Treatment Phase

The clinical management of the patient over the period of treatment is described in this section. It includes items such as how often and by whom the patient is seen for treatment or monitoring, what tests or procedures are performed at each visit, and what data are collected at each visit.

Follow-Up Phase

In trials where the patient continues to be observed after treatment is completed, the follow-up procedures are described including the frequency and duration of the follow-up visits and the data collected at each visit.

Termination

Procedures for ending patients' participation in the trial, whether because they completed the planned schedule or because they need or wish to leave early, are described. Anticipated reasons for early termination are specified and methods for minimizing early terminations are described.

Study Flow Diagram

It is often desirable to include a flowchart describing how patients progress through the trial. The chart starts with patients' initial screening for recruitment and ends with their completion of the planned schedule giving pertinent highlights such as randomization, treatment assignment, treatment and follow-up visits, and possible early termination. For complicated studies, such a flowchart is useful, both to reviewers and trial investigators, as a concise reference for implementing the protocol.

Outcome Measures

The primary outcome measure(s) are those required to answer the trial's primary hypotheses. They are described in detail, including how they are collected, their validity, accuracy and reliability, the methods used to ensure that they are measured in a uniform and unbiased manner, and the methods used to minimize loss of these data in the trial. The secondary outcome measures for answering any secondary hypotheses are also listed (*see* **Outcome Measures in Clinical Trials**).

Statistical Issues

Three types of statistical issues may need to be addressed in the protocol.

1. *Sample size*. The number of patients required for the trial is determined and justified. A trial that is too small will have little chance to answer clearly the study hypotheses, while too large a trial will waste money and may subject patients needlessly to an inferior treatment. The trial sample size should be large enough to answer all of the primary hypotheses. Methods to calculate sample sizes are available for most trial designs [1, 3, 6]. The crucial issues that must be resolved usually involve estimating the expected results in the control group and determining the treatment difference to be detected (*see* **Sample Size Determination for Clinical Trials**).
2. *Statistical analysis*. The planned statistical analyses to analyze the trial data are outlined. These planned analyses are specific for each of the study hypotheses, especially for the primary hypotheses.
3. *Interim monitoring*. For trials over extended periods, the data are usually analyzed at regular intervals to determine whether a conclusion can be reached early or if there are any safety issues

that need addressing (*see* **Data and Safety Monitoring**). For such interim monitoring there are a number of statistical techniques [2] that have been developed that allow analysis of the accumulating data multiple times without affecting statistical inference at the end of the study. This section describes such interim monitoring plans.

Laboratories

If special, nonroutine laboratory tests are used in the trial, then they are described in the protocol. For multicenter studies that include central laboratories, procedures for obtaining and shipping specimens to the central laboratory are described.

Compliance

For many studies, such as those that require the administration of medication over long periods of time or that deal with difficult populations such as drug abusers, compliance with treatment regimens and with protocol procedures is a major concern. For such trials, the protocol describes how compliance is monitored (e.g. pill counts, blood serum levels, and/or missed visits) and methods used to improve compliance in noncompliant patients (*see* **Compliance Assessment in Clinical Trials**).

Ethics and Safety

One of the most important concerns of any clinical trial is the protection of the trial patients' rights and safety. These concerns are addressed in the protocol. The protocol discusses how the patient is approached for entry into the trial, how informed consent is obtained, and what safeguards are in place to ensure that the patient's participation in the trial is voluntary and confidential. Copies of the informed consent form to be used in the trial are included in an appendix. Procedures to monitor patient safety are also described, including the adverse events to be monitored, how and to whom the adverse events are reported, and a plan of action should a serious adverse event be detected during the trial (*see* **Ethics of Randomized Trials**).

Organization

The conduct of a clinical trial can be complex, particularly in **multicenter trials**. To ensure that the trial is conducted correctly, the protocol describes its organizational structure. This includes naming each investigator and describing his/her role, including his/her supervisory responsibilities, describing the roles and responsibilities of all trial support staff, and providing the composition and rules of any special committees, such as steering, endpoint adjudication, and data and safety monitoring committees, including their relationship to other components of the trial. **Data management and coordination** plans may also be described here.

Budget

The budget section lists the projected costs for the study by year of trial as well as the total budget, and provides a breakdown of costs including personnel, equipment, supplies, laboratory costs, and travel. Justification for most items in the budget is provided.

Study Forms

Many funding institutions require that the study forms be included in an appendix. This helps reviewers to determine whether the investigators are collecting the appropriate data.

Summary

Since a clinical trial protocol has many purposes and is used by so many people, it is important that it be written clearly, concisely, unambiguously, and in sufficient detail that it meets the requirements of all of its users. The protocol should provide enough detail that readers will know how the study is being conducted, but not so much detail as to overwhelm the reader. Although it is important that investigators adhere to the protocol, mechanisms should be in place for making changes if the need arises. If changes are made, then they must be well documented. In this manner, the different versions of the protocol provide a description of the evolution of the trial. In summary, the protocol acts as the main document describing the

design and conduct of a clinical trial and, as such, plays a central role in the trial.

References

[1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. Lawrence Erlbaum, Hillsdale.

[2] DeMets, D.L. (1987). Practical aspects in data monitoring: a brief review, *Statistics in Medicine* **6**, 753–760.

[3] Lachin, J.M. (1981). Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials* **2**, 93–113.

[4] Meinert, C.L. (1985). *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

[5] Schwartz, D., Flamant, R. & Lellouch, J. (1980). *Clinical Trials*. Academic Press, London.

[6] Shuster, J.J. (1990). *Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, Boca Raton.

[7] Spilker, B. (1984). *Guide to Clinical Studies and Developing Protocols*. Raven Press, New York.

[8] Weiner, J.M. (1979). *Issues in the Design and Evaluation of Medical Trials*. G.K. Hall Medical Publishers, Boston.

(*See also* **Clinical Trials, Overview**)

JOSEPH F. COLLINS

Clinical Trials, Early Cancer and Heart Disease

Early developments in controlled **clinical trials** at the **National Institutes of Health (NIH)** took place mainly at the National Cancer Institute (NCI) and what was then the National Heart Institute (NHI) [subsequently the National Heart, Lung, and Blood Institute (NHLBI)] beginning in the 1950s. This article reviews the developments from the early 1950s to the late 1960s at both institutes, summarizing the early efforts in clinical trials, the organizations set up to conduct and monitor the clinical trials, and the developments in statistical methodology that have formed the basis for conducting many of the present day randomized controlled trials. The early history of clinical trials at these institutes has been reviewed in more detail at NCI by Gehan & Schneiderman and at NHLBI by Halperin et al. [28, 32].

Developments in Clinical Trials at the National Cancer Institute (NCI)

A major advance in the development of chemical agents for the treatment of cancer came from observations of the treatment of children with acute lymphocytic leukemia, which was a rapidly fatal disease until 1948 when Sidney Farber, in a nonrandomized study of methotrexate, observed complete remissions and longer survival among some pediatric patients [21]. However, results did not meet with uniform acceptance and questions were raised about diagnosis, selection of patients, and reporting. There was a need for a more organized approach to treatment experimentation that would lead to **unbiased** evaluations of treatments. At about the same time, animal models of the major forms of cancer – sarcomas, carcinomas, and leukemias – were developed that could be used to screen candidate materials and, if the materials were effective and not overly toxic, ultimately lead to clinical trials in humans. By 1960, there was an annual screening of approximately 25 000–30 000 materials sponsored by NCI with only about 10 – 20 new agents having sufficient effectiveness in animal systems to merit consideration for testing in humans. Peter Armitage, of the London School of Hygiene and Tropical Medicine, was a visiting scientist at NCI in the late 1950s. His background in sequential statistical procedures (*see* **Sequential Analysis**) quickly found direct application in the development of two- and three-stage screening procedures for animal tumor systems that permitted rejection of an agent at any stage but acceptance only at the final stage [3, 43]. The object was to determine quickly which new compounds should be considered for further study in man. In the late 1950s, optimism was high that this screening program would lead to a new chemotherapeutic treatment that would make large clinical trials unnecessary. Also, there was a belief that different forms of cancer were sufficiently similar so that an agent active in one form of the disease would also be active in another.

Early Efforts in Clinical Trials

Dr C. Gordon Zubrod came to NCI in 1954 at about the time that Dr James Holland departed for Roswell Park Memorial Institute in Buffalo, NY. Drs Emil Frei and E.J. Freireich arrived at NCI in 1955. Under the leadership of Zubrod, this formed the key group of clinicians who initiated the clinical trials program at NCI. When Zubrod was at Johns Hopkins University in the early 1950s, he indicated that there "were two streams of influence (relating to developments in clinical trials) – infectious disease chemotherapy and comparative studies of analgesics and hypnotic drugs" [52]. Among those playing an important role in the conduct of clinical trials at Johns Hopkins were Dr James Shannon (later Director of the National Institutes of Health), the pharmacologist E.K. Marshall, Jr and **W.G. Cochran**. About this time, the studies of streptomycin in pulmonary tuberculosis by the Medical Research Council were published and had a profound influence on the Johns Hopkins group [41] (*see* **Medical Research Council Streptomycin Trial**). The first effort at a randomized trial was a comparison of the efficacy of tetracycline and penicillin in the treatment of lobar pneumonia [5]. At the same time, the Veterans Administration began its first randomized controlled trials in tuberculosis [50].

The Organization of Trials

In 1954, the US Congress created the Cancer Chemotherapy National Service Center (CCNSC) to stimulate research in the chemotherapy of cancer. A clinical panel was formed, headed by Dr I. Ravdin, and

included among others Drs Zubrod and Holland. At an early meeting, the clinical panel reviewed a paper by Louis Lasagna, which enunciated five principles of the controlled clinical trial, including **randomization** and the statistical treatment of data [38]. Over the next several years, the clinical panel of the CCNSC oversaw the organization of cooperative clinical trials groups for the conduct of clinical trials in cancer (*see* **Cooperative Cancer Trials**). By 1960, there were 11 cooperative clinical study groups (Table 1), each comprised of a number of universities and/or V.A. Hospitals and Medical Centers and a Statistical Coordinating Center [48]. The cooperative groups were funded by the NCI through the Chairman and a Statistical Center. Zubrod recruited the chairman of each group and **Marvin Schneiderman** recruited the biostatisticians and statistical centers. One of the statisticians, W.J. Dixon, had two graduate students who were writing general statistical programs for the analysis of biomedical data. NCI awarded a contract to carry out this work that subsequently became the Biomedical Data Processing Program (BMDP) package of statistical programs (*see* **Software, Biostatistical**).

In the establishment of a clinical cooperative group, CCNSC agreed that there should be adherence to the following principles: combination of data from all institutions to accumulate rapidly the necessary number of patients; standard criteria of diagnosis, treatment, and measurement of effect; statistical design of the study, with a randomized assignment of patients to the groups to be compared; and statistical analysis and collaborative reporting of the results.

The clinical trials effort involved more types of clinical studies than randomized trials. There was a sequence of trials with differing objectives: **Phase I** – to determine the maximum tolerated dose of a regimen that can be used in looking for therapeutic effect; **Phase II** – to determine whether a particular dosage schedule of an agent is active enough to warrant further study; and Phase III – a comparative trial, usually randomized, to decide whether a new therapy is superior to a standard therapy. The primary objective of the clinical trials program was to provide a means of testing in humans new agents that had previously demonstrated effectiveness in animal tumor systems.

Some Early Trials

Following some preliminary discussions between Dr Zubrod and **Jerome Cornfield**, a leading statistician at NIH, there was agreement that childhood leukemia was an ideal disease for testing some of the new agents objectively. The first randomized cooperative clinical trial in acute leukemia was planned in 1954, begun in 1955, and reported by Frei et al. in 1958 [23]. The trial involved two regimens of combination chemotherapy – 6-mercaptopurine and either intermittent or continuous methotrexate in 65 patients. The study had the following features: a uniform protocol at the four participating institutions (*see* **Clinical Trials Protocols**); uniform criteria of response (*see* **Outcome Measures in Clinical Trials**); adherence to the principles of the controlled clinical trial, especially the randomization of patients to therapies (*see* **Randomized Treatment Assignment**); and stratification of patients by age, type of leukemia, and history of prior therapy. Statistical methods used were a comparison of

Table 1 Cooperative clinical study groups in 1960

Group	Chairman	Statistician
Acute leukemia, Group A	M. Lois Murphy	I. Bross
Acute leukemia, Group B	E. Frei	M. Schneiderman
Eastern Solid Tumor Group	C.G. Zubrod	M. Schneiderman
Southeastern Group	R.W. Rundles	B.G. Greenberg
Western Group	F. Willett	
Southwestern Group	H.G. Taylor	E. MacDonald
Prostate Group	H. Brendler	D. Mainland
Breast Group A	A. Segaloff	M. Schneiderman
Breast Group B	G. Gordon	M. Schneiderman
V.A. Groups – various malignancies	J. Wolf et al.	M. Patno
University Groups – lung, breast, stomach, ovary, colon	A. Curreri et al.	R. Stiver, G. Beebe, W. Dixon

median survival times and median duration of remissions between therapies, **confidence intervals**, and **Fisher's exact test**.

The first randomized clinical trial in solid tumors was conducted by members of the Eastern Solid Tumor Group and reported by Zubrod et al. in 1960 [53]. The trial involved a randomized comparison of two alkylating agents (thiotepa vs. nitrogen mustard) in patients with solid tumors. One objective was to "study the feasibility and usefulness of collaborative clinical research in cancer chemotherapy". The trial involved 258 randomized patients, and notable features were: blind evaluation of response by vote of clinical investigators (*see* **Blinding or Masking**); objective procedures for measurement of tumors and determination of when a response began and ended; the importance of accounting for type I and type II statistical errors (*see* **Hypothesis Testing**); appropriate sample size for detection of differences between treatments (*see* **Sample Size Determination for Clinical Trials**); and statistical analysis in the reporting of results.

A subsequent trial demonstrated the value of combination chemotherapy in acute leukemia and the independent action of drugs to increase the probability that a patient achieves complete remission [24]. Freireich et al. [25] reported a prospective, randomized, double-blind, placebo-controlled, sequential study of 6-mp vs. placebo in the maintenance of remissions in pediatric acute leukemia. This study established that 6-mp maintenance treatment leads to substantially longer remissions than placebo and was a forerunner to many adjuvant studies in other forms of cancer, such as breast cancer, in which treatments are administered when the patients are in a disease-free state [25]. This study also was a motivation for the development of an extension of the Wilcoxon test for comparing survival distributions subject to censoring [27] (*see* **Wilcoxon–Mann–Whitney Test**) and was used as an example by Cox in his, now classic, paper on regression models (*see* **Cox Regression Model**) and **life tables** [16].

Developments in Methodology

In the clinical trials program at NCI prior to 1970, there were several developments in methodology that have influenced the conduct of subsequent clinical trials. Before 1960, the clinical testing of new agents often involved as few as five patients, with the agent discarded if no positive response was obtained in at least one patient. In 1961, Gehan proposed a plan for Phase II trials that determined the minimum number of consecutive patients to study when all patients are nonresponders, before one could reject a new agent for further study, at given levels of rejection error [26]. This plan, or now more commonly Simon's modification, continues in use today in Phase II studies [46].

Several philosophical issues arose from the drug development program. The practice of human experimentation could be questioned by "Doesn't a physician have an implied duty to give his patient the best treatment? If that is the case, how can one justify having the toss of coin (i.e. randomization) decide which treatment a patient should receive?" The reply was (and is), "If the physician really knows what is the best treatment for the patient, the patient must receive that treatment and not be randomized into a trial." The question then becomes, "How and when does a physician know what is the best treatment for a specific patient?" The major ethical issue then becomes one of learning quickly (i.e. with a minimum number of patients) what is the best treatment (*see* **Ethics of Randomized Trials**). There have been several proposals for establishing what one "knows" while minimizing the number of patients who will receive the less effective treatment. Armitage proposed closed sequential procedures with paired patients on each treatment, and with the trial terminated as soon as one could establish the superiority of one of the treatments over the other [2] (*see* **Data and Safety Monitoring**). A feature of the plans was an upper limit on the number of patients one could enter. Schneiderman & Armitage later described a family of sequential procedures, called wedge plans because of the shape of the acceptance boundary, which provided a bridge between the open plans derived from **Wald's** theory and the restricted procedures of Armitage [44, 45].

In the 6-mp vs. placebo study for maintaining remissions in pediatric leukemia, patients were paired according to remission status (complete or partial), one patient receiving 6-mp and the other placebo by a random allocation, and a preference was recorded for 6-mp or placebo depending upon the therapy which resulted in the longer remission. The trial was conducted sequentially according to one of the Armitage plans [2] and a sequential boundary favoring 6-mp

was reached after 18 preferences had occurred – 15 for 6-mp and 3 for placebo. There were 12 patients still in remission at the time the study was terminated, although one could record a preference for one or the other treatment because the pair-mate had relapsed at an earlier time. It was clear that a more efficient analysis could be obtained by using the actual lengths of remission. Gehan, while working on an NCI fellowship with D.R. Cox at Birkbeck College in London, developed a generalization of the Wilcoxon test for the fixed sample size problem with each sample subject to arbitrary right **censoring** [27]. **Halperin** had previously developed a generalization of the Wilcoxon test, when all times to censoring were equal to the longest observation time [30]. Mantel noticed that one could utilize the **chi-square test** for comparison of survival data between two or more groups, assuming that one constructs a **contingency table** of deaths and survivors at each distinct failure time in the groups of patients under study. This chi-square test was appropriate when the risk of failure in one group was a constant multiple of that in the other; this test was an extension of the earlier test developed by Mantel and Haenszel which measured the statistical significance of an observed association between a disease and a factor under study in terms of an increased **relative risk** of disease [39, 40]. This test subsequently became known variously as the **Mantel–Haenszel test**, the **logrank test** or the Cox–Mantel test, and has been studied by Cox and Peto, among others [16, 42].

Another development in the 1960s was the exponential **regression** model proposed by Feigl & Zelen [22]. Dr Robert Levin of NCI was interested in studying the relationship of the survival time of leukemia patients to the concomitant variate of white blood count, separately according to the presence or absence of auer rods and/or significant granulature of leukemia cells in the bone marrow at diagnosis. Feigl & Zelen proposed a model in which an exponential **survival distribution** is postulated for each patient and the expected value of the survival time is linearly related to the patient's white blood count. A more general **loglinear model** was subsequently given by Glasser [29], and there have been numerous subsequent developments in parametric regression models with censored survival data (*see* **Parametric Models in Survival Analysis**) [17, Chapters 5 and 6, pp. 62–90].

Developments in Clinical Trials at the National Heart, Lung, and Blood Institute (NHLBI)

Prior to 1960, the National Heart Institute (NHI), subsequently to become NHLBI, had little involvement in **multicenter clinical trials**. In a trial designed in 1951, there was a comparison of ACTH, cortisone, and aspirin in the treatment of rheumatic fever and the prevention of rheumatic heart disease. A total of 497 children were enrolled in 12 centers in the UK, the US, and Canada. Felix Moore, then Chief of the Biometrics Section of NHI, was a statistical consultant. There were no differences in treatment effectiveness in the study, and no statistical or methodologic problems mentioned in the final report [8].

Subsequently, there was a multicenter observational study of lipoproteins in atherosclerosis that had substantial impact on the methodology for coordinating studies performed at several sites [47]. The Statistical Center was led by Felix Moore and Tavia Gordon at NHI. Careful quality control procedures and standardization of methods across centers were emphasized.

Early Efforts in Clinical Trials

Jerome Cornfield joined the NHI in 1960 and strongly influenced the conduct of clinical trials at NHI and statistical research on methodologic issues arising in clinical trials. In the early 1960s, intensive planning for two clinical trials was begun at NHI to reduce risk factors for coronary heart disease – The Diet Heart Feasibility Study (DHFS) and the Coronary Drug Project (CDP) [14, 20]. These studies reflected the strong interest in both dietary and drug approaches to the prevention of coronary heart disease and the recurrence of myocardial infarction. For the DHFS, the NHI Biometrics Branch served as the statistical coordinating center, first under the supervision of Joseph Schachter and later of Fred Ederer. Max Halperin rejoined the NHI in 1966 and, upon Cornfield's retirement in 1968, became Chief of the Biometrics Research Branch until his retirement in 1977. Four areas of clinical trials and methodology can be traced to these early studies and the individuals responsible for them. These areas are: organizational structure for clinical trials at NIH; methodology for the interim analysis of accumulating data (*see* **Data and Safety Monitoring**), including the **Bayesian**

approach, group sequential and stochastic curtailment methods; design and analysis of clinical trials, including the effects of patient **noncompliance** on **power** and the **intention to treat** principle; and methods for analysis of data from longitudinal clinical trials (*see* **Longitudinal Data Analysis, Overview**).

The Organization of NHLBI Trials

The "NHLBI Model" for cooperative clinical trials evolved from discussion during the planning stage of the CDP among outside medical experts and NHI medical and statistical staff. In 1967, a report by a committee appointed by the National Advisory Heart Council and chaired by **Bernard Greenberg** described this structure [35]. The report, subsequently known as the "Greenberg Report", became the basis for a structure of nearly all subsequent NHLBI trials as well as for many other trials sponsored at NIH.

The major components of the organizational structure include a Steering Committee, a Policy Advisory Board, a Data Monitoring Committee (*see* **Data Monitoring Committees**), and a Statistical or Data Coordinating Center, as well as individual clinics, central laboratories, and various other committees which served the needs of the trial. These might include committees to develop **eligibility criteria**, to assign cause of death, to define methodology and standards, or to oversee the preparation of manuscripts (for more details of organizational structure).

From the biostatistical viewpoint, the Data Monitoring Committee has the responsibility of monitoring accumulating data on a periodic basis and analyzing results for evidence of early benefit or harm. Primary and secondary outcomes measures are reviewed, along with safety data, compliance to the protocol, and subgroup analyses which may identify particular risk groups (*see* **Treatment-covariate Interaction**). The Statistical Coordinating Center and the Data Monitoring Committee work closely together in performing the appropriate data analyses needed for fulfilling the Committee's responsibilities.

The Statistical and Data Coordinating Centers for early trials at the NHLBI are given in Table 2. Personnel at these coordinating centers have played an important role in the development of clinical trials and made numerous contributions to statistical methodology.

Developments in Methodology

These are considered under three headings: data monitoring, design and analysis, and longitudinal studies.

Data Monitoring. Jerome Cornfield was involved in the planning and conduct of two clinical trials – the DHFS and the CDP. Both Cornfield and Halperin served on the Data and Safety Monitoring Committee of the CDP. At least partly motivated by his involvement in these trials, Cornfield published papers in 1966 on sequential trials, sequential analysis, and the **likelihood** principle, from a Bayesian perspective [9, 10].

In 1966, Max Halperin worked jointly with Cornfield and Samuel Greenhouse (then at the National

Table 2 Early NHLBI coordinating centers

University of Maryland/Maryland Research Institute
Coronary Drug Project
University of Texas School of Public Health
Hypertension Detection and Follow-up Program
University of North Carolina – Chapel Hill, School of Public Health
Lipid Research Clinical Program
University of Minnesota School of Public Health, Biometry Division
Multiple Risk Factor Intervention Trial
University of Washington School of Public Health, Biostatistics Department
Coronary Artery Surgery Study
George Washington University Biostatistics Center
Intermittent Positive Pressure Breathing Trial
NHLBI Biometrics Research Branch
National Diet Heart Feasibility Study
Urokinase Pulmonary Embolism Trial
Urokinase Streptokinase Pulmonary Embolism Trial

Institute of Mental Health) to develop an adaptive allocation procedure that would assign an increasing proportion of patients to the better of two treatments as evidence accumulated [13] (*see* **Adaptive and Dynamic Methods of Treatment Assignment**). Their approach to the problem was Bayesian and generalized the earlier work of Anscombe and Colton [1, 7]. At around the same time, Cornfield published a general paper on the Bayesian approach that involved the use of a **prior probability distribution** with a mass of probability P at the **null hypothesis**, with a continuous density of total mass $1 - P$ over a set of **alternative hypotheses** [11]. A key feature of Cornfield's proposal was the rejection of the null hypothesis when the posterior odds (the relative betting odds or RBO) became small for H_0. The RBO was used in the CDP in the monitoring of mortality differences between the control and each of the drug treatment groups. Subsequently, Canner, of the CDP Coordinating Center, considered the determination of critical values for decision making at multiple time points during the conduct of the clinical trial from the **Neyman–Pearson** perspective [6]. Later, curtailment and stochastic curtailment methods were developed and applied to trials of the NHLBI in the 1970s and early 1980s [19, 31, 34, 37].

Statisticians working with the CDP were aware that, as the data accumulated, repeated testing for treatment differences using conventional statistical significance levels would increase the type I error (*see* **Level of a Test**) over the nominal alpha level associated with that critical value. Armitage et al. evaluated the impact of repeated testing on the type I error and demonstrated that multiple tests could increase the type I error substantially [4]. Interim analyses of clinical data are necessary for scientific and ethical reasons, but large type I errors are not acceptable. Canner developed a method for the CDP for determining the critical value at each interim analysis so that the overall type I error is close to the desired level [6]. Statisticians involved with NHLBI trials developed group sequential methods and applied them to trials starting with the CDP.

Design and Analysis. In the DHFS, it was projected that a reduction in cardiovascular risk would result from a reduction in cholesterol level. The original sample size projection was for the entry of 8000 patients into several treatment arms. Although a special review committee suggested that this sample size might be too large, Cornfield argued that there were too many inconclusive small studies already in the literature. Several aspects of the trial required consideration, including noncompliance with the treatment regimen. It was presumed that the maximum effect on risk would occur only after some period of time on treatment and that failure to adhere to the treatment regimen could mean a return to higher risk levels. Halperin et al. [33] incorporated these considerations into the design of clinical trials by proposing methods for adjusting sample size for noncompliance in the treatment group. Studies were considered with a fixed period of observation and a comparison of proportions as the main analysis. Implicit in this paper is the "**intention to treat**" principle, i.e. analysis of all randomized patients in their assigned treatment group regardless of compliance. Ultimately, the report of the CDP recognized this point [15]. Most primary and secondary **prevention trials** conducted by the NHLBI since 1970 have made use of sample size adjustments for noncompliance.

The **Framingham Heart Study** was begun in 1948 and has had an important influence on methodologic research at the NHLBI and the design of prevention trials. Over 5000 adult residents of Framingham, Massachusetts, were entered into a longitudinal study with the objective of evaluating the effects of various risk factors on the development of subsequent cardiovascular disease. The study has clarified the roles of high blood pressure, elevated total serum cholesterol, and cigarette smoking on the risk of cardiovascular disease [18, 36]. Occurrence or not of a cardiovascular event in a 2-year follow-up period is a **binary** outcome. Cornfield considered a regression approach to deal with the binary outcome variables. The problem was closely related to the discrimination problem between two samples from **multivariate normal distributions**. For a specific prior probability of belonging or not to a disease group, the posterior probability could be represented as a **logistic regression** function that was closely related to what could be obtained from a conventional **discriminant function analysis** [49]. Cornfield & Mitchell argued that one could use the logistic model to predict the impact on risk of specified changes in risk factors [12]. Subsequently, this logistic model approach was used in the design of several NHLBI prevention trials.

Longitudinal Studies. A methodology for analysis of longitudinal data was needed for the Framingham

Study which could be considered both a **cohort** and a longitudinal study. Cohorts of individuals were followed to observe patterns of morbidity and mortality, and biennial measurements of cardiovascular risk factors provided an opportunity to study patterns relating to aging. Early reports of the Framingham study used simple graphical and descriptive methods to describe patterns of aging. During the 1980s, there was much work on methodology for longitudinal studies (*see* **Longitudinal Data Analysis, Overview**) that ultimately led to NHLBI sponsorship of a workshop on methods for analysis of longitudinal and follow-up studies, whose proceedings have appeared as a special issue in ***Statistics in Medicine*** [51].

References

[1] Anscombe, F.J. (1963). Sequential medical trials, *Journal of the American Statistical Association* **58**, 365–383.

[2] Armitage, P. (1957). Restricted sequential procedures, *Biometrika* **44**, 9–26.

[3] Armitage, P. & Schneiderman, M. (1958). Statistical problems in a mass screening program, *Annals of the New York Academy of Science* **76**, 896–908.

[4] Armitage, P., McPherson, C.K. & Rowe, B.C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

[5] Austrian, R., Mirick, G., Rogers, D., Sessoms, S.M., Tumulty, P.A., Vickers, W.H., Jr. & Zubrod, C.G. (1951). The efficacy of modified oral penicillin therapy of pneumococcal lobar pneumonia, *Bulletin of Johns Hopkins Hospital* **88**, 264–269.

[6] Canner, P.L. (1977). Monitoring treatment differences in long-term clinical trials, *Biometrics* **33**, 603–615.

[7] Colton, T. (1963). A model for selecting one of two medical treatments, *Journal of the American Statistical Association* **58**, 388–400.

[8] Cooperative Clinical Trial of ACTH, Cortisone and Aspirin in the Treatment of Rheumatic Fever and the Prevention of Rheumatic Heart Disease (October 1960). *Circulation* **22**.

[9] Cornfield, J. (1966). Bayesian test of some classical hypotheses – with applications to sequential clinical trials, *Journal of the American Statistical Association* **61**, 577–594.

[10] Cornfield, J. (1966). Sequential trials, sequential analysis, and the likelihood principle, *American Statistician* **20**, 18–23.

[11] Cornfield, J. (1969). The Bayesian outlook and its application, *Biometrics* **25**, 617–657.

[12] Cornfield, J. & Mitchell, S. (1969). Selected risk factors in coronary disease. Possible intervention effects, *Archives of Environmental Health* **19**, 382–394.

[13] Cornfield, J., Halperin, M. & Greenhouse, S. (1969). An adaptive procedure for sequential clinical trials, *Journal of the American Statistical Association* **64**, 759–770.

[14] Coronary Drug Project Research Group (1973). The Coronary Drug Project. Design, methods, and baseline results, *Circulation* **47**, Supplement 1, 179.

[15] Coronary Drug Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project, *New England Journal of Medicine* **303**, 1038–1041.

[16] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[17] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[18] Dawber, T.R., Meadors, G.F. & Moor, F.E. (1951). Epidemiological approaches to heart disease: the Framingham Study, *American Journal of Public Health* **41**, 279–286.

[19] DeMets, D.L. & Halperin, M. (1981). Early stopping in the two-sample problem for bounded variables, *Controlled Clinical Trials* **3**, 1–11.

[20] Diet–Heart Feasibility Study Research Group (1968). The National Diet–Heart Study Final Report, *Circulation* **37**, Supplement 1, 428.

[21] Farber, S., Diamond, L.K., Mercer, R., Sylvester, R.F. Jr. & Wolff, J.A. (1948). Temporary remissions in children produced by folic acid antagonist aminopterin, *New England Journal of Medicine* **238**, 787–793.

[22] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[23] Frei, E., III, Holland, J.F., Schneiderman, M.A., Pinkel, D., Selkirk, G., Freireich, E.J., Silver, R.T., Gold, G.L. & Regelson, W. (1958). A comparative study of two regimens of combination chemotherapy in acute leukemia, *Blood* **13**, 1126–1148.

[24] Frei, E., III, Freireich, E.J., Gehan, E.A., Pinkel, D., Holland, J.F., Selawry, O., Haurani, F., Spurr, C.L., Hayes, D.M., James, W., Rothberg, H., Sodee, D.B., Rundles, W., Schroeder, L.R., Hoogstraten, B., Wolman, I.J., Tragis, D.G., Cooper, T., Gendel, B.R., Ebaugh, F. & Taylor, R. (1961). Studies of sequential and combination antimetabolite therapy in acute leukemia: 6-mercaptopurine and methotrexate, *Blood* **18**, 431–454.

[25] Freireich, E.J., Gehan, E.A., Frei, E., III, Schroeder, L.R., Wolman, I.J., Anbari, R., Bergert, O., Mills, S.D., Pinkel, D., Selawry, O.S., Moon, J.H., Gendel, B.R., Spurr, C.L., Storrs, R., Haurani, F., Hoogstraten, B. & Lee, S. (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy, *Blood* **21**, 699–716.

[26] Gehan, E.A. (1961). The determination of the number of patients required in a preliminary and follow-up trial of a new chemotherapeutic agent, *Journal of Chronic Diseases* **13**, 346.

[27] Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples, *Biometrika* **52**, 203–223.

[28] Gehan, E.A. & Schneiderman, M.A. (1990). Historical and methodological developments in clinical trials at the National Cancer Institute, *Statistics in Medicine* **9**, 871–880.

[29] Glasser, M. (1967). Exponential survival with covariance, *Journal of the American Statistical Association* **62**, 561–568.

[30] Halperin, M. (1960). Extension of the Wilcoxon-Mann-Whitney test to samples censored at the same fixed point, *Journal of the American Statistical Association* **55**, 125–138.

[31] Halperin, M. & Ware, J. (1974). Early decision in a censored Wilcoxon two-sample test for accumulating survival data, *Journal of the American Statistical Association* **69**, 414–422.

[32] Halperin, M., DeMets, D.L. & Ware, J.H. (1990). Early methodological developments for clinical trials at the National Heart Lung and Blood Institute, *Statistics in Medicine* **9**, 881–882.

[33] Halperin, M., Rogot, E., Gurian, J. & Ederer, F. (1968). Sample sizes for medical trials with special reference to long term therapy, *Journal of Chronic Diseases* **21**, 13–24.

[34] Halperin, M., Ware, J., Johnson, N.J., Lan, K.K. & Demets, D. (1982). An aid to data monitoring in long-term clinical trials, *Controlled Clinical Trials* **3**, 311–323.

[35] Heart Special Project Committee (1988). Organization, review, and administration of cooperative studies (Greenberg Report): A report from the Heart Special Project Committee to the National Advisory Heart Council, May 1967, *Controlled Clinical Trials* **9**, 137–148.

[36] Kannel, W.B., Dawber, T.R., Kagan, A., Nevotskie, N. & Stokes, J. (1961). Factors of risk in the development of coronary heart disease – six year followup experience: the Framingham Study, *Annals of Internal Medicine* **55**, 33–50.

[37] Lan, K.K.G., Simon, R. & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials, *Communications in Statistics – Stochastic Models* **1**, 207–219.

[38] Lasagna, L. (1955). The controlled clinical trial: theory and practice, *Journal of Chronic Diseases* **1**, 353–358.

[39] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports* **50**, 163–170.

[40] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[41] Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis, *British Medical Journal* **2**, 769–783.

[42] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series A* **135**, 185–206.

[43] Schneiderman, M.A. (1961). Statistical problems in the screening search for anti-cancer drugs by the National Cancer Institute of the United States, in *Quantitative Methods in Pharmacology*. North-Holland, Amsterdam.

[44] Schneiderman, M.A. & Armitage, P. (1962). A family of closed sequential procedures, *Biometrika* **49**, 41–56.

[45] Schneiderman, M.A. & Armitage, P. (1962). Closed sequential t-tests, *Biometrika* **49**, 359–366.

[46] Simon, R. (1989). Optimal two stage designs for Phase II trials, *Controlled Clinical Trials* **10**, 1–10.

[47] Technical Group and Committee on Lipoproteins and Atherosclerosis (1956). Evaluation of serum lipoproteins and cholesterol measurements as predictors of clinical complications of atherosclerosis, *Circulation* **14**, 691–742.

[48] The National Program of Cancer Chemotherapy Research (1960). *Cancer Chemotherapy Reports* **1**, 5–34.

[49] Truett, J., Cornfield, J. & Kannel, W.B. (1967). A multivariate analysis of the risk factors of coronary heart disease in Framingham, *Journal of Chronic Diseases* **20**, 511–524.

[50] Tucker, W.B. (1954). Experiences with controls in the study of the chemotherapy of tuberculosis, *Transactions of the 13th Veterans Administration Conference on the Chemotherapy of Tuberculosis*, Vol. 15.

[51] Wu, M., Wittes, J.T., Zucker, D. & Kusek, J. eds (1988). Proceedings of the Workshop on Methods for Longitudinal Data Analysis in Epidemiological and Clinical Studies, *Statistics in Medicine* **7**, 1–361.

[52] Zubrod, C.G. (1982). Clinical trials in cancer patients: an introduction, *Controlled Clinical Trials* **3**, 185–187.

[53] Zubrod, C.G., Schneiderman, M., Frei, E., III, Brindley, C., Gold, G.L., Shnider, B., Oviedo, R., Gorman, J., Jones, R., Jr, Jonsson, U., Colsky, J., Chalmers, T., Ferguson, B., Dederick, M., Holland, J., Selawry, O., Regelson, W., Lasagna, L. & Owens, A.H., Jr (1960). Appraisal of methods for the study of chemotherapy of cancer in man: Comparative therapeutic trial of nitrogen mustard and thiophosphoramide, *Journal of Chronic Diseases* **11**, 7–33.

MARVIN A. SCHNEIDERMAN* &
EDMUND A. GEHAN

* Deceased, April 1997

Clinical Trials, History of

Aspects of the history of **clinical trials** have been reviewed by, among others, Bull [10], Lilienfeld [49], Armitage [4], Meinert [59], and Gail [31]. In this article we survey the historical progression toward the modern clinical trial, as this method of research is practiced at the end of the twentieth century, by tracing the development of five of its requisite elements: **controls** (a comparison group), **randomization**, **blinding or masking, ethics**, and interim statistical analysis (*see* **Data and Safety Monitoring**).

Controls

The essence of the clinical trial is the control group, which provides the basis for comparing the outcomes of two or more treatments. The comparative concept of assessing therapeutic efficacy has been known from ancient times. Lilienfeld [49] cites a description of a nutritional experiment involving a control group in the Book of Daniel from the Old Testament:

> 1. In the third year of the reign of Jehoiakim king of Judah came Nebuchadnezzar king of Babylon unto Jerusalem, and besieged it. . .
> 2. And the king spoke unto Ashpenaz his chief officer, that he should bring in certain of the children of Israel, and of the seed royal, and of the nobles. . .
> 5. And the king appointed for them a daily portion of the king's food and of the wine which he drank that they should be nourished for three years. . .
> 8. But Daniel purposed in his heart that he would not defile himself with the king's food, nor with the wine which he drank; therefore he requested of the chief of the officers that he might not defile himself. . .
> 10. And, the chief of officers said unto Daniel: "I fear my lord the king who hath appointed your food and your drink; for why should he see your faces sad in comparison with the youths of your own age?". . .
> 11. Then said Daniel to the steward. . .
> 12. Try thy servants, I beseech thee, ten days; and let them give us pulse (leguminous plants) to eat and water to drink. . .
> 13. Then let our countenances be looked upon before thee, and the countenances of the youths that eat of the king's food. . .
> 14. So, he hearkened unto them and tried them in this matter, and tried them ten days. . .
> 15. And at the end of ten days their countenances appeared fairer, and they were fatter in the flesh, than all the youths that did eat of the king's food [72].

In this early example of a clinical trial, we note the presence not merely of a control group, but of a concurrent control group. These fundamental elements of clinical research did not begin to be widely practiced until the latter half of the twentieth century.

There appear to be no other recorded examples of thinking in comparative terms about the outcome of medical treatment in ancient or medieval times. Lilienfeld [49] provides an example from the fourteenth century, a letter from Petrarch to Boccaccio:

> I solemnly affirm and believe, if a hundred or a thousand of men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one followed the prescriptions of the doctors of the variety of those practicing at the present day, and that the other half took no medicine but relied on Nature's instincts, I have no doubt as to which half would escape [78].

The Renaissance provides an example of an unplanned experiment in the treatment of battlefield wounds. The surgeon Ambroise Paré was using the standard treatment of pouring boiled oil over the wound during the battle to capture the castle of Villaine in 1537. When he ran out of oil, he resorted to the alternative of a digestive made of egg yolks, oil of roses, and turpentine. The superiority of the new treatment became evident the next day.

> I raised myself very early to visit them, when beyond my hope I found those to whom I applied the digestive medicament feeling but little pain, their wounds neither swollen nor inflamed, and having slept through the night. The others to whom I had applied the boiling oil were feverish with much pain and swelling about their wounds. Then I determined never again to burn thus so cruelly by arquebusses [63] (as cited in [10]).

An oft-cited eighteenth-century example of a planned controlled clinical trial is the ship-board experiment in which Lind found oranges and lemons to be the most effective of six dietary treatments for scurvy.

> On the 20th of *May*, 1747, I took twelve patients in the scurvy, on board the *Salisbury* at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz. water-gruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times puddings, boiled biscuit with sugar etc. And for supper, barley and raisins, rice and currants, sago and wine, or the like. Two of these were ordered each a quart of cyder a day. Two others took twenty-five gutts of *elixir vitriol* three times a day, upon an empty stomach; using a gargle strongly acidulated with it for their mouths. Two others took two spoonfuls of vinegar three times a day, upon an empty stomach; having their gruels and their other food well acidulated with it, as also the gargle for their mouths. Two of the worst patients, with the tendons in the ham rigid (a symptom none of the rest had) were put under a course of sea-water. Of this they drank half a pint every day, and sometimes more or less as it operated, by way of gentle physic. Two others had each two oranges and one lemon given them every day. These they eat with greediness, at different times, upon an empty stomach. They continued but six days under this course, having consumed the quantity that could be spared. The two remaining patients, took the bigness of a nutmeg three times a day of an electuary recommended by a hospital-surgeon, made of garlic, mustard-feed, *rad. raphan*, balsam of *Peru*, and gum myrrh; using for common drink barley-water well acidulated with tamarinds; by a decoction of which, with the addition of *cremor tartar*, they were greatly purged three or four times during the course.
>
> The consequence was, that the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them, being at the end of six days fit for duty. The spots were not indeed at that time quite off his body, nor his gums sound; but without any other medicine, than a gargle of *elixir vitriol*, he became quite healthy before we came into Plymouth, which was on the 16th June. The other was the best recovered of any in his condition; and being now deemed pretty well, was appointed nurse, to the rest of the sick [50] (as cited in [10, 37, 49], and [59]).

Pierre-Charles-Alexandre Louis, a nineteenth-century clinician and pathologist, introduced the "numerical method" for comparing treatments. His idea was to compare the results of treatments on groups of patients with similar degrees of disease, i.e. to compare "like with like":

> I come now to therapeutics, and suppose that you have some doubt as to the efficacy of a particular remedy: How are you to proceed?. . . You would take as many cases as possible, of as similar a description as you could find, and would count how many recovered under one mode of treatment, and how many under another; in how short a time they did so; and if the cases were in all respects alike, except in the treatment, you would have some confidence in your conclusions; and if you were fortunate enough to have a sufficient number of facts from which to deduce any general law, it would lead to your employment in practice of the method which you had seen oftenest successful [51] (as cited in [10, 49, 3], and [59]).

It remained for **Bradford Hill** more than a century later to use a formal method for creating groups of cases that were "in all respects alike, except in the treatment".

Randomization

The use of randomization as a scientific tool was a brilliant contribution by the famous statistician **Ronald A. Fisher** [22, 23]. Fisher's early applications of randomization were in agriculture. To determine which fertilizers effected the greatest crop yields, Fisher divided agricultural areas into plots, and randomly assigned the plots to experimental fertilizers. A goal of Fisher's was to obtain, through independent replications, a valid test of statistical significance (*see* **Randomization Tests**). In previous systematic designs, the fertilities of adjoining plots, to which different treatments had been applied, were not independent.

In clinical trials there were early schemes to use "**group randomization**": after dividing the patients into two groups, the treatment for each group was randomly selected. This method does not involve replication, and therefore precludes estimation of error. Armitage [3] cites a challenge based on the notion of random assignment, though not of individuals, issued as early as 1662 by the Belgian medicinal chemist van Helmont:

> Let us take out of the hospitals, out of the Camps, or from elsewhere, 200, or 500 poor People that have Fevers, Pleurisies, &c, Let us divide them into halfes, let us cast lots, that one half of them may fall to my share, and the others to yours,. . . we shall see how many funerals both of us shall have: *But* let the

> reward of the contention or wager, be 300 florens, deposited on both sides [75].

Group randomization was used by Amberson et al. in a trial of sanocrysin in the treatment of pulmonary tuberculosis published in 1931 [1].

A great step forward was the use of systematic assignment by Fibiger [21], who alternately assigned diphtheria patients to serum treatment or an untreated control group. As noted by Armitage [3], alternate assignment "would be deprecated today on the grounds that foreknowledge of the future treatment allocations may selectively bias the admission of patients into the treatment groups". Diehl et al. [18] reported in 1938 a common cold vaccine study with University of Minnesota students as subjects:

> At the beginning of each year... students were assigned at random... to a control group or an experimental group. The students in the control groups... received placebos... All students thought they were receiving vaccines... Even the physicians who saw the students... had no information as to which group they represented.

Gail [31] points out that, although on its face this appears to be the first published report of a modern randomized clinical trial, a typewritten manuscript by Diehl clarifies that this is another instance of systematic assignment:

> At the beginning of the study, students who volunteered to take these treatments were assigned alternately and without selection to control groups and experimental groups.

Hill, in the study of streptomycin in pulmonary tuberculosis [53], used random sampling numbers in assigning treatments to subjects in clinical trials, so that the subject was the unit of randomization (*see* **Medical Research Council Streptomycin Trial**). This study is now generally acknowledged to be the "first properly randomized clinical trial" [3]. It is of interest to note, as did Meier [58], that what Fisher saw important in randomization was that it made possible a valid test of significance (*see* **Hypothesis Testing**), whereas what Hill found important was the creation of comparable groups.

After the streptomycin-pulmonary tuberculosis trial, Bradford Hill and the British **Medical Research Council** continued with further randomized trials: chemotherapy of pulmonary tuberculosis in young adults [55], antihistaminic drugs in the prevention and treatment of the common cold [54], cortisone and aspirin in the treatment of early cases of rheumatoid arthritis [56, 57], and long-term anticoagulant therapy in cerebrovascular disease [39].

In the US, the **National Institutes of Health** followed the lead of the British Medical Research Council, starting in 1951 its first randomized trial [34], a National Heart Institute study of ACTH, cortisone, and aspirin in the treatment of rheumatic heart disease [68] (*see* **Clinical Trials, Early Cancer and Heart Disease**). This was followed in 1954 by a randomized trial of retrolental fibroplasia (now known as retinopathy of prematurity), sponsored by the National Institute of Neurological Diseases and Blindness [44]. In that same year, members of the US Congress asked officials of the National Cancer Institute to organize a comprehensive program for research in cancer chemotherapy, which led the next year to the development of a rapidly growing program of clinical trials under the Cancer Chemotherapy National Service Center [32]. By fiscal year 1986, the annual cost of randomized clinical trials sponsored by 10 categorical institutes of the National Institutes of Health amounted to 300 million dollars; the National Cancer Institute bore 58% of that cost [35]. During the four decades following the pioneering trials of the 1940s and 1950s, there was a large growth in the number of randomized trials not only in Britain and the US, but also in Canada and on the European continent. This growth gave impetus to the formation in the 1970s of two societies, the **International Society of Clinical Biostatistics** and the Society for Clinical Trials, and the publication of two new journals, **Controlled Clinical Trials** and **Statistics in Medicine**.

Masking

The purpose of masking, or blinding, in experiments is to prevent personal **bias** from influencing study observations. An awareness that personal bias can affect observation and judgment has existed for at least 400 years. Francis Bacon (1561–1626) noted "for what a man would like to be true, he more readily believes" [5]. Investigator bias caused a remarkable scientific delusion in the early years of the twentieth century: n-rays [76]. N-rays were "discovered" in 1902 by the eminent French physicist Blondlot, who in *Comptes rendus*, the leading French scientific journal, reported properties of

these rays that far transcended those of X-rays. According to Blondlot, n-rays were given off spontaneously by many metals, such as copper, zinc, lead, and aluminum, and when the rays fell upon the eye, they increased the eye's ability to see objects in a nearly dark room. The existence of n-rays was soon confirmed in laboratories in various parts of France, and a number of noted French scientists soon applied n-rays to research in chemistry, botany, physiology, and neurology. In 1904, *Science Abstracts* listed 77 n-ray papers. The French Academy awarded Blondlot the Lalande prize of 20 000 francs and its gold medal "for the discovery of n-rays". That same year, however, the American Physicist R.W. Wood visited Blondlot in his laboratory to test the experiments:

> He [Blondlot] first showed me a card on which some circles had been painted in luminous paint. He turned down the gas light and called my attention to their increased luminosity when the n-ray was turned on. I said that I saw no change. He said that was because my eyes were not sensitive enough, so that proved nothing. I asked him if I could move an opaque lead screen in and out of the path of the rays while he called out the fluctuations of the screen. He was almost 100 percent wrong and called out fluctuations when I made no movement at all, and that proved a lot, but I held my tongue. He then showed me the dimly lighted clock, and tried to convince me that I could see the hands when he held a large flat file just above his eyes. I asked if I could hold the file, for I had noticed a flat wooden ruler on his desk, and remembered that wood was one of the substances that *never* emitted n-rays. He agreed to this, and I felt around for the ruler and held it in front of his face. Oh, yes, he could see the hands perfectly. This also proved something [70, 76].

After Wood published his account, n-ray publications diminished in number. *Science Abstracts* listed only eight n-ray papers in 1905, and none in 1909. The French Academy changed its announced reason for the award to Blondlot "for his life work taken as a whole". According to Seabrook [70], the exposure of the blunder led to Blondlot's madness and death.

A masked experiment by the Austrian physicist Pozdena contributed to the disproof of the existence of n-rays. At haphazard intervals Pozdena's assistant soundlessly operated a shutter which in its closed position blocked the transmission of the hypothetical n-rays. During a pretest, the assistant wrote "o" for *offen* (open) and "g" for *geschlossen* (closed) while Pozdena indicated when he could detect increased luminosity. Whereas the shutter's movements were silent, the assistant's pencil scratches were not. Pozdena was able to hear the difference between an "o" and a "g". In the definitive experiment the assistant switched to a coded notation, and in 150 trials Pozdena reported increased luminosity about as often when the shutter was open as when it was closed [67].

The common cold vaccine study published by Diehl et al. [18] in 1938 cited earlier, in which University of Minnesota students were alternately assigned to vaccine or placebo, was a masked clinical trial.

> The students in the control groups . . . received placebos . . . All students thought they were receiving vaccines . . . Even the physicians who saw the students . . . had no information as to which group they represented.

Masking was used in the early Medical Research Council trials in which Bradford Hill was involved. Thus, in the first of those trials, the study of streptomycin in tuberculosis, the X-ray films were

> viewed by two radiologists and a clinician, each reading the films independently and not knowing if the films were of C [bed-rest alone] or S [streptomycin and bed-rest] cases [53].

Hill's lesson from the experience:

> If it [the clinical assessment of the patient's progress and of the severity of the illness] is to be used effectively, without fear and without reproach, the judgments must be made without any possibility of bias, without any overcompensation for any possible bias, and without any possible accusation of bias [37].

In the second trial, the antihistamine–common cold study [54], placebos ("dummies indistinguishable" from the drug under test) were used. Hill's lesson:

> . . .in [this] trial. . ., feelings may well run high in the bosom (or should I say the mucosa?) either of the recipient of the drug or the clinical observer, or indeed of both. If either were allowed to know the treatment that had been given, I believe that few of us would without qualms accept that the drug was of value – if such a result came out of the trial [37].

The terms "blind" and "double-blind" have been used commonly in clinical trials, the latter indicating

that neither the doctor nor the patient knows what treatment the patient is getting. When these terms were recognized as being awkward in trials of eye disease, the terms "masking" and "double-masking" were introduced [19].

Ethics

Medical Research Abuses

Experimentation in medicine is as old as medicine itself, and since antiquity some experiments on humans have been conducted without concern for the welfare of the subjects, who were often prisoners or disadvantaged people [52]. Thus, in the flourishing days of intellectual and scientific achievement in ancient Alexandria, anatomists used criminals for dissection alive [6]. Katz [42] provides examples of 19th century studies in Russia and Ireland of the consequences of infecting persons with syphilis and gonorrhea. During the same time, in the US,

> physicians put slaves into pit ovens to study heat stroke, and poured scalding water over them as an experimental cure for typhoid fever. One slave had two fingers amputated in a "controlled trial", one finger with anesthesia and one finger without, to test the effectiveness of anesthesia [52].

Unethical experiments on human beings have continued into the twentieth century [7, 24, 52]. In 1932 the US Public Health Service began a study in Tuskegee, Alabama, of the natural progression of untreated syphilis in 400 black men. The study continued until 1972, when a newspaper reported that the subjects were uninformed or misinformed about the purpose of the study. Participants were told that painful lumbar punctures were given as treatment, when in fact treatment for syphilis was withheld even after penicillin became available [24].

During the Nazi regime, 1933–1945, German doctors conducted sadistic medical experiments, mainly on Jews, but also on Gypsies, mentally disabled persons, Russian prisoners of war, and Polish concentration camp inmates:

> The "experiments" were quite varied. Prisoners were placed in pressure chambers and subjected to high-altitude tests until they stopped breathing. They were injected with lethal doses of typhus and jaundice. They were subjected to "freezing" experiments in icy water or exposed naked in the snow outdoors until they froze to death. Poison bullets were tried out on them as was mustard gas... [71].

Codes of Ethics, Informed Consent

The fact that in 1931, two years before the Nazis acceded to power, Germany had enacted "Richtlinien" (regulations) to control human experiments adds irony to the German doctors' cruel abuse and exploitation of human subjects.

> Issued by the Reich's Health Department, these regulations remained binding law throughout the period of the Third Reich. Consent requirements formed two of fourteen major provisions in the guidelines, one dealing with "New Therapy" and the other with "Human Experimentation". It was demanded that in both cases consent (first party or proxy consent, as appropriate) must always be given "in a clear and undebatable manner" [20].

The Nazi doctors were tried for their atrocities by the Allied Forces in 1946–1947 at Nuremberg. Three US judges at the trial promulgated the Nuremberg Code [47], the first international effort to codify ethical principles of clinical research. Principle 1 of the Nuremberg Code states:

> The voluntary consent of the human subject is absolutely essential. This means that the person involved should have legal capacity to give consent; should be so situated as to be able to exercise free power of choice, without the intervention of any element of force, fraud, deceit, duress, over-reaching, or other ulterior form of constraint or coercion; and should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision [74] (cited in [47, Appendix 3]).

Other principles of the Code are that the experiment should yield results for the good of society, that unnecessary suffering and injury should be avoided, and that the subject should be free to end the experiment.

Informed consent (*see* **Ethics of Randomized Trials**) was used by Walter Reed in his studies of yellow fever at the turn of the twentieth century [6, 69]. Mosquitoes were known to be involved in the transmission of the disease, but their precise role was not clear. To clarify the mode of transmission, members of Reed's research team had themselves been bitten by mosquitoes. After a fellow worker died of yellow fever from a purposeful bite, Reed

recruited American servicemen and Spanish workers for the experiments, and drew up a contract with the Spanish workers:

> The undersigned understands perfectly well that in the case of the development of yellow fever in him, that he endangers his life to a certain extent but it being entirely impossible to avoid the infection during his stay on this island he prefers to take the chance of contracting it intentionally in the belief that he will receive... the greatest care and most skillful medical service [6, 69].

The contract specified that volunteers would each receive $100 in gold, and a $100 bonus if they contracted yellow fever. In the event 25 volunteers became ill, but none died.

In addition to Reed, other early advocates of informed consent were Charles Francis Withington and William Osler. Withington, noting in 1886 the "possible conflict between the interests of medical science and those of the individual patient", sided with "the latter's indefensible rights" [77]. Osler in 1907 insisted on informed consent in medical experiments: "For man absolute safety and full consent are the conditions which make such tests allowable" [62]. Despite this early advocacy, and despite the promulgation of the 1931 German doctors' code and the 1946–1947 Nuremberg Code, the application of informed consent to medical experiments did not take foothold during the first six decades of the twentieth century. Bradford Hill [38], based on his experience in a number of early randomized clinical trials sponsored by the Medical Research Council, believed that it was not feasible to draw up a detailed code of ethics for clinical trials that would cover the variety of ethical issues that came up in these studies, and that the patient's consent was not warranted in all clinical trials. Although the judges at Nuremberg evidently intended the Code to apply not only to the case before them, but "for the practice of human experimentation wherever it is conducted" [43], European and American clinical investigators were slow to adopt it [7]. Gradually the medical community came to recognize the need to protect the reputation and integrity of medical research. In 1955 a human experimentation code was adopted by the Public Health Council in the Netherlands [60], cited in [15], and in 1964 the World Medical Association issued the Declaration of Helsinki [47], essentially adopting the ethical principles of the Nuremberg Code, with consent "a central requirement of ethical research" [20].

Justification to Begin

The view of Bradford Hill [38] was that in starting a randomized clinical trial the doctor accepts that "he really has no knowledge that one treatment [in the trial] will be better or worse [than the other treatments]". This state of uncertainty, which has come to be known as "equipoise" [28], has remained the ethical standard for starting a randomized clinical trial. For completeness, Levine [47] has added the proviso that there must not be a treatment, other than those to be studied in the trial, that is known to be superior to the study treatments.

Peer Review

One can trace back to 1803 the notion that therapeutic innovation must be preceded by peer review:

> And no such trials [of new remedies and new methods of chirurgical treatment] should be instituted, without a previous consultation of the physicians or surgeons ... [64] (cited in [47]).

According to Levine [47], "not much more was said about peer review for about 150 years".

Research ethics committees, the US history of which is traced by McNeill [52] and Levine [48], came into being in the US in the 1950s. The 1946–1947 Nuremberg Code and the 1964 Declaration of Helsinki do not mention committee review. A requirement for such review was imposed in 1953 at the newly established Clinical Center at the National Institutes of Health, and peer review of clinical research was also practiced at some US medical schools in the 1950s. By the early 1960s, one-third of US university medical schools responding to a survey had established research ethics committees. Public outrage at highly publicized research abuses, such as those published by Beecher [7], or those committed in the Tuskegee syphilis study, gave impetus to the adoption of requirements for informed consent and peer review in research on human beings in the US [52]. In 1966 the US Public Health Service issued a policy requiring recipients of Public Health Service research grants to provide for prior committee review of studies involving human subjects to ensure that the study plans conform to ethical standards. Because the Public Health Service was then (and still is) sponsoring a large majority of medical research in the US, research ethics committees were established at

medical schools throughout the US soon after 1966. National recommendations, guidelines, or regulations for the establishment of research ethics committees in other countries soon followed: Canada in 1966, the UK in 1967, Australia in 1973, New Zealand in 1975, and Ireland in 1987 [52].

Data Monitoring by Peers

In the modern randomized clinical trial, the accumulating data are usually monitored for safety and efficacy by an independent *data monitoring committee* – also called *data and safety monitoring committee*, or **data and safety monitoring board**. In 1968 the first such committee was established, serving the Coronary Drug Project, a large multicenter trial sponsored in the United States by the National Heart Institute of the National Institutes of Health [11, 29]. The organization of the Coronary Drug Project included a policy board – a senior advisory group made up of five scientists who were not otherwise participating in the study. In 1967, after a presentation of interim outcome data by the study leadership to all participating investigators of the Coronary Drug Project, **Thomas Chalmers** addressed a letter to the policy board chairman expressing concern:

> that knowledge by the investigators of early nonstatistically significant trends in mortality, morbidity, or incidence of side effects might result in some investigators – desirous of treating their patients in the best possible manner, that is, with the drug that is ahead – pulling out of the study or unblinding the treatment groups prematurely [11].

In 1968 a data and safety monitoring committee was established for the Coronary Drug Project (apparently by the policy board) consisting of scientists who were not contributing data to the study, and thereafter the practice of sharing accumulating outcome data with the study's investigators was discontinued. The data safety and monitoring committee assumed responsibility for deciding when the accumulating data warranted changing the study treatment protocol or terminating the study (*see* **Data and Safety Monitoring**).

In 1971, for the first randomized clinical trial it sponsored, the recently established National Eye Institute of the US National Institutes of Health adopted the model of the Coronary Drug Project by including in its organization a policy board and data monitoring committee; the trial was the multicenter Diabetic Retinopathy Study [17]. In this study, as in the Coronary Drug Project, the accumulating outcome data were not shared with data-contributing investigators.

In subsequent trials sponsored by the National Heart Institute (later named National Heart, Lung, and Blood Institute) and the National Eye Institute the functions of the policy board and data and safety monitoring committee were combined in a single data monitoring committee. The example set by the Coronary Drug Project and the Diabetic Retinopathy Study established a pattern for monitoring interim clinical trials data by an independent committee that was gradually adopted by many trials in North America and Europe.

Interim Analysis

In the conduct of the modern randomized clinical trial, the ethical requirement for interim analysis of study outcomes is widely recognized, and the responsibility for such analysis is commonly delegated to an independent data monitoring committee. Bradford Hill does not mention interim analysis in his extensive writings about the clinical trials he worked on during the late 1940s and 1950s [37].

The first formal recognition of the need for interim analyses, and that such analyses affect the probability of the type I error (*see* **Hypothesis Testing**), came with the publication in the 1950s of papers on sequential clinical trials by Bross [9] and Armitage [2] (*see* **Sequential Analysis**). In sequential trials of two treatments, patients are enrolled in pairs, with members of each pair randomly assigned to one or the other treatment. The data are analyzed each time that both members of a pair reach an endpoint (e.g. treatment failure). The overall probability of the type I error is controlled at a predetermined level. The principal advantage of a sequential trial over a fixed-sample-size trial, apart from that of correcting the significance level for repeated data analyses, is that when the length of time needed to reach an endpoint is short, e.g. weeks or months, the sample size required to detect a substantial benefit from one of the treatments is less. Applications of sequential trials have been limited because when follow-up is long-term, as is required by most trials, the sequential design is less effective.

In the 1960s **Cornfield** argued that data analysis in clinical trials, because it is often marked by unforeseen developments, does not lend itself well to predetermined stopping rules [13, 16]. For the dilemma of repeated interim analyses of the accumulating data, and to address the issue of **multiplicity in clinical trials** in general, he proposed use of the **likelihood ratio**. In particular, he proposed a **Bayesian** solution in the form of "relative betting odds" [12] – a method that was applied alongside conventional frequentist methods in two trials [14, 73].

In the 1970s and 1980s frequentist solutions to interim analysis came about in the form of "group sequential trials" and "stochastic curtailment" [4, 30, 66]. In the group sequential trial, an analogue of the classical sequential trial [2, 9], the frequency of interim analysis is usually limited to a small number, say, between 3 and 6, while the overall type I error probability is controlled at a predetermined level. Pocock's boundaries use constant nominal significance levels for the individual tests; the Haybittle–Peto boundary [36, 65] uses stringent significance levels, except for the final test; in the O'Brien–Fleming boundary, stringency gradually decreases [61]; in the model by Lan & DeMets [45], the total type I error probability is gradually spent in a manner that does not require the timing of analyses be prespecified; there have also been proposals for methods of repeated **confidence intervals** [40]. Whereas group sequential designs are used to determine whether a trial should be stopped early because a treatment is efficacious, stochastic curtailment, which involves prediction of future events, is invoked when it appears that a treatment is unlikely to be shown to be efficacious even if the trial is continued to its planned conclusion [46]. Both group sequential methods and stochastic curtailment have been frequently applied to trials in the 1980s and 1990s.

Despite a renewed interest in Bayesian clinical trials since the 1980s [8, 25, 26, 33, 41], there have been few applications. Freedman et al. [27] provide an overview of Bayesian approaches to interim analysis, including a Bayesian analogue to stochastic curtailment.

References

[1] Amberson, B. & McMahon, P.M. (1931). A clinical trial of sanocrysin in pulmonary tuberculosis, *American Review of Tuberculosis* **24**, 401.

[2] Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials, *Quarterly Journal of Medicine* **23**, 255–274.

[3] Armitage, P. (1983). Trials and errors: the emergence of clinical statistics, *Journal of the Royal Statistical Society, Series A* **146**, 321–334.

[4] Armitage, P. (1991). Interim analysis in clinical trials, *Statistics in Medicine* **10**, 925–937.

[5] Bacon, F.V.S. (1961–1963). Novum organum, in *The Works of Francis Bacon* (transl. James Spedding, orig. publ. Longman, London, 1858–1874). Facsimile reprint, Frommann, Stuttgart, 1961–1963. Also cited in *The Oxford Dictionary of Quotations*, 3rd Ed. Oxford University Press, 1979.

[6] Bean, W.B. (1995). Walter Reed and the ordeal of human experiments, in *Ethics in Epidemiology and Clinical Research*, S.S. Coughlin, ed. Epidemiology Resources Inc., Newton, pp. 3–22.

[7] Beecher, H.K. (1966). Ethics and clinical research, *New England Journal of Medicine* **274**, 1354–1360.

[8] Berry, D.A. (1987). Interim analyses in clinical research, *Cancer Investigation* **5**, 469–477.

[9] Bross, I. (1952). Sequential medical plans, *Biometrics* **8**, 188–295.

[10] Bull, J.P. (1959). The historical development of clinical therapeutic trials, *Journal of Chronic Disease* **10**, 218–248.

[11] Canner, P. (1983). Monitoring of the data for adverse or beneficial treatment effects, *Controlled Clinical Trials* **4**, 467–483.

[12] Cornfield, J. (1969). The Bayesian outlook and its applications, *Biometrics* **24**, 617–657.

[13] Cornfield, J. (1976). Recent methodological contributions to clinical trials, *American Journal of Epidemiology* **104**, 408–421.

[14] Coronary Drug Project Research Group (1970). The Coronary Drug Project. Initial findings leading to a modification of its research protocol, *Journal of the American Medical Association* **214**, 1303–1313.

[15] Curran, W.J. & Shapiro, E.D. (1970). *Law, Medicine, and Forensic Science*, 2nd Ed. Little, Brown & Company, Boston.

[16] Cutler, S.J., Greenhouse, S.W., Cornfield, J. & Schneiderman, M.A. (1966). The role of hypothesis testing in clinical trials. Biometrics seminar, *Journal of Chronic Diseases* **19**, 857–882.

[17] The Diabetic Retinopathy Study Research Group (1981). Photocoagulation treatment of diabetic retinopathy. Design, methods, and baseline results. Report 6, *Investigative Ophthalmology and Visual Science* **21**, Part 2, 149–209.

[18] Diehl, H.S., Baker, A.B. & Cowan, D.W. (1938). Cold vaccines: an evaluation based on a controlled study, *Journal of the American Medical Association* **111**, 1168–1173.

[19] Ederer, F. (1975). Patient bias, investigator bias and the double-masked procedure, *American Journal of Medicine* **58**, 295–299.

[20] Faden, R.R., Beauchamp, T. & King, N.M.P. (1986). *A History of Informed Consent*. Oxford University Press, New York.
[21] Fibiger, J. (1898). Om Serum Behandlung of Difteri, *Hospitalstidende* **6**, 309–325, 337–350.
[22] Fisher, R.A. (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture* **33**, 503–513.
[23] Fisher, R.A., & McKenzie, W.A. (1923). Studies in crop variation: II. The manurial response of different potato varieties, *Journal of Agricultural Science* **13**, 315.
[24] Freedman, B. (1995). Research, unethical, in *Encyclopedia of Bioethics*, W.T. Reich, ed. Free Press, New York, pp. 2258–2261.
[25] Freedman, L.S. & Spiegelhalter, D.J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials, *Statistician* **32**, 153–160.
[26] Freedman, L.S., Lowe, D. & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion, *Biometrics* **40**, 575–586.
[27] Freedman, L.S., Spiegelhalter, D.J. & Parmar, M.K.B. (1994). The what, why, and how of Bayesian clinical trials monitoring, *Statistics in Medicine* **13**, 1371–1383.
[28] Fried, C. (1974). *Medical Experimentation: Personal Integrity and Social Policy*. North-Holland, Amsterdam.
[29] Friedman, L. (1993). The NHLBI model: a 25-year history, *Statistics in Medicine* **12**, 425–431.
[30] Friedman, L.M., Furberg, C.D., & DeMets, D.L. (1985). *Fundamentals of Clinical Trials*, 2nd Ed. Wright, Boston.
[31] Gail, M.H. (1996). Statistics in action, *Journal of the American Statistical Association* **91**, 1–13.
[32] Gehan, E.A. & Lemak, N.A. (1994). *Statistics in Medical Research: Developments in Clinical Trials*. Plenum Medical Book Company, New York.
[33] George, S.L., Chengchang, L., Berry, D.A. & Green, M.R. (1994). Stopping a trial early: frequentist and Bayesian approaches applied to a CALGB trial of non-small-cell lung cancer, *Statistics in Medicine* **13**, 1313–1328.
[34] Greenhouse, S.W. (1990). Some historical and methodological developments in early clinical trials at the National Institutes of Health, *Statistics in Medicine* **9**, 893–901.
[35] Hawkins, B.S. (1988). The National Institutes of Health and their sponsorship of clinical trials, *Controlled Clinical Trials* **9**, 103–106.
[36] Haybittle, J.L. (1971). Repeated assessment of results of cancer treatment, *British Journal of Radiology* **44**, 793–797.
[37] Hill, A.B. (1962). *Statistical Methods in Clinical and Preventive Medicine*. E & S Livingstone, Edinburgh.
[38] Hill, A.B. (1963). Medical ethics and controlled trials, *British Medical Journal* **1**, 1043–1049.
[39] Hill, A.B., Marshall, J. & Shaw, D.A. (1960). A controlled clinical trial of long-term anticoagulant therapy in cerebrovascular disease, *Quarterly Journal of Medicine* **29**, (NS), 597–608.
[40] Jennison, C. & Turnbull, B.W. (1989). Interim analysis: the repeated confidence interval approach, *Journal of the Royal Statistical Society, Series B* **51**, 305–361.
[41] Kadane, J.B. (1995). Prime time for Bayes, *Controlled Clinical Trials* **16**, 313–318.
[42] Katz, J. (1972). *Experimentation with Human Beings: The Authority of the Investigator, Subject, Professions, and State in the Human Experimentation Process*. Russell Sage Foundation, New York.
[43] Katz, J. (1996). The Nuremberg Code and the Nuremberg Trial. A reappraisal, *Journal of the American Medical Association* **276**, 1662–1666.
[44] Kinsey, V.E. (1956). Retrolental fibroplasia, *AMA Archives of Ophthalmology* **56**, 481–543.
[45] Lan, K.K.G., & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.
[46] Lan, K.K.G., Simon, R. & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials, *Communications in Statistics-Sequential Analysis* **1**, 207–219.
[47] Levine, R.J. (1986). *Ethics and Regulation of Clinical Research*, 2nd Ed. Urban & Schwarzenberg, Baltimore, pp. 187–190.
[48] Levine, R. (1995). Research ethics committees, in *Encyclopedia of Bioethics*, W.T. Reich, ed. Free Press, New York, pp. 2258–2261.
[49] Lilienfeld, A.M. (1982). Ceteris paribus: The evolution of the clinical trial, *Bulletin of the History of Medicine* **56**, 1–18.
[50] Lind, J. (1753). *A Treatise of the Scurvy*. Sands Murray Cochran, Edinburgh, pp. 191–193.
[51] Louis, P.C.A. (1837). The applicability of statistics to the practice of medicine, *London Medical Gazette* **20**, 488–491.
[52] McNeill, P.M. (1993). *The Ethics and Politics of Human Experimentation*. Press Syndicate of the University of Cambridge, Cambridge.
[53] Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis, *British Medical Journal* **2**, 769–782.
[54] Medical Research Council (1950). Clinical trials of antihistaminic drugs in the prevention and treatment of the common cold, *British Medical Journal* **ii**, 425–431.
[55] Medical Research Council (1952). Chemotherapy of pulmonary tuberculosis in young adults, *British Medical Journal* **i**, 1162–1168.
[56] Medical Research Council (1954). A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis-I, *British Medical Journal* **i**, 1223–1227.
[57] Medical Research Council (1955). A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis-II, *British Medical Journal* **ii**, 695–700.
[58] Meier, P. (1975). Statistics and medical experimentation, *Biometrics* **31**, 511–529.

[59] Meinert, C.L. (1986). *Clinical Trials. Design, Conduct, and Analysis*. Oxford University Press, New York.

[60] Netherlands Minister of Social Affairs and Health (1957). *4 World Medical Journal*, 299–300.

[61] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[62] Osler, W. (1907). The evolution of the idea of experiment, *Transactions of the Congress of American Physicians and Surgeons* **7**, 1–8.

[63] Packard, F.R. (1921). *The Life and Times of Ambroïse Paré*, 2nd Ed. Paul B. Hoeber, New York, pp. 27, 163.

[64] Percival, T. (1803). *Medical Ethics*. Russell, London.

[65] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. 1. Introduction and design, *British Journal of Cancer* **34**, 585–612.

[66] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.

[67] Pozdena, R.F. (1905). Versuche über Blondlot's "Emission Pesante", *Annalen der Physik* **17**, 104.

[68] Rheumatic Fever Working Party (1960). The evolution of rheumatic heart disease in children: five-year report of a cooperative clinical trial of ACTH, cortisone, and aspirin, *Circulation* **22**, 505–515.

[69] Rothman, D.J. (1995). Research, human: historical aspects, in *Encyclopedia of Bioethics*, W.T. Reich, ed. Free Press, New York, pp. 2258–2261.

[70] Seabrook, W. (1941). *Doctor Wood*. Harcourt, Brace & Company, New York, p. 234.

[71] Shirer, W.L. (1960). *The Rise and Fall of the Third Reich*. Simon & Schuster, New York.

[72] Slotki, J.J. (1951). *Daniel, Ezra, Nehemia, Hebrew Text and English Translation with Introductions and Commentary*. Soncino Press, London.

[73] Urokinase-Pulmonary Embolism Trial (1973). A National Cooperative Study, A.A. Sasahara, T.M. Cole, F. Ederer, J.A. Murray, N.K. Wenger, S. Sherry & J.M. Stengle, eds. *Circulation* **47**, Supplement 2, pp. 1–108.

[74] US Government Printing Office (1949). *Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10*, Vol. 2. US Government Printing Office, Washington, pp. 181–182.

[75] Van Helmont, J.B. (1662). *Oriatrike or Physik Refined* (translated by J. Chandler). Lodowick Loyd, London.

[76] Vogt, E.Z. & Hyman, R. (1959). *Water Witching, USA*. University of Chicago Press, Chicago, p. 50.

[77] Withington, C.F. (1886). *The Relation of Hospitals to Medical Education*. Cupples Uphman, Boston.

[78] Witkosky, S.J. (1889). *The Evil That Has Been Said of Doctors: Extracts From Early Writers*, Trans. with annotations by T.C. Minor. The Cincinnati Lancet-Clinic, Vol. 41, New Series Vol. 22, pp. 447–448.

FRED EDERER

Clinical Trials, Overview

Trial is from the Anglo–French *trier*, meaning *to try*. Broadly, it refers to the action or process of putting something to a test or proof. *Clinical* is from *clinic*, from the French *cliniqué* and from the Greek *klinike*, and refers to the practice of caring for the sick at the bedside. Hence, narrowly, a *clinical trial* is the action or process of putting something to a test or proof at the bedside of the sick. However, broadly it refers to any testing done on human beings for the sake of determining the value of a treatment for the sick or for preventing disease or sickness.

The broad definition of clinical trial includes definitions allowing for use of the term in references to studies involving a single treatment (e.g. as in most **phase I trials** and some **phase II drug trials**) and for studies involving use of an external **control** (e.g. studies involving historical controls) [66]. However, use herein will be in the stricter sense of usage; that is, to refer to trials involving two or more treatment groups comprised of persons enrolled, treated, and followed over the exact same time frame.

The *treatment* can be anything considered to hold promise in caring for the sick, in the prevention of disease, or in the maintenance of health. The term, in the context of a trial, refers to the experimental variable – the variable manipulated by the trialist. The variable may have just two states (e.g. as in a trial involving a single test treatment and single control treatment) or three or more states (e.g. as in a trial involving several different test treatments and one or more control treatments). The variable, in the case of drug trials, may serve to designate different drugs, different doses of the same drug, or different forms or routes of administration of the same drug. In other contexts, it may variously refer to different kinds or forms of surgery, different kinds or forms of care or management regimens, different kinds or forms of diagnostic tests, different kinds or forms of medical devices, different kinds or forms of counseling regimens to achieve some desired end, or combinations of the above.

The clinical trial, in its simplest form, involves the application of the experimental variable – treatment to a person or group of persons – and observation during or following application of the treatment to measure its effect. That measure **(outcome measure)** may be death, occurrence or recurrence of some morbid condition, or a difference indicative of change (e.g. difference in blood pressure measured for each person just prior to the start of treatment and again at some point during or after treatment).

There is no way to "test" a treatment or to "prove" its effectiveness in the absence of some absolute or relative measure of success. Trials are said to be *controlled* if the effect of a treatment is measured against a comparison treatment administered over the same time period and under similar conditions. That comparison treatment may be another test treatment or, depending on circumstances, a *control treatment* consisting of an accepted standard form of therapy, a placebo (*see* **Blinding or Masking**) or sham treatment, or observation only (no treatment).

A trial is said to be *uncontrolled* if it does not have a comparison treatment or if the enrollment to and administration of the test and comparison treatments is not concurrent (e.g. as with use of *historical controls* for evaluation of a treatment). The Book of Daniel (Chapter 1, verses 12–15) provides an account of what amounts to an uncontrolled trial involving a diet of pulse – edible seeds of certain pod-bearing plants, such as peas and beans (*see* **Clinical Trials, History of**).

> Prove thy servants, I beseech thee, ten days; and let them give us pulse to eat, and water to drink. Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the King's meat: and as thou seest, deal with thy servants. So he consented to them in this matter, and proved them ten days. And at the end of ten days their countenances appeared fairer and fatter in flesh than all the children which did eat the portion of the King's meat [1].

Fortuitous events can produce conditions reminiscent of the features of a trial. One such account is that given by Ambroise Paré (surgeon, 1510–1590) during the battle in 1537 for the castle of Villaine. The treatment for gunshot wounds in Paré's time was boiling oil poured over the wound. Because of the intensity of the battle, Paré ran out of oil and resorted to using an ointment made of egg yolks, oil of roses, and turpentine. The result of his "trial" is summarized by his observation the morning after the battle:

> I raised myself very early to visit them, when beyond my hope I found those to whom I had applied the digestive medicament, feeling but little pain, their wounds neither swollen nor inflamed, and having slept through the night. The others to whom I had

> applied the boiling oil were feverish with much pain and swelling about their wounds. Then I determined never again to burn thus so cruelly the poor wounded by arquebuses [72].

Many of the essential elements of the modern day *controlled trial* are contained in Lind's account of a trial performed aboard the *Salisbury* at sea in 1747:

> On the 20th of May 1747, I took twelve patients in the scurvy, on board the *Salisbury* at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz., watergruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times puddings, boiled biscuit with sugar, etc; and for supper, barley and raisins, rice and current, sago and wine, or the like. Two of these were ordered each a quart of cyder a day. Two others took twenty-five gutts of elixir vitriol three times a day, upon an empty stomach; using a gargle strongly acidulated with it for their mouths. Two others took two spoonfuls of vinegar three times a day, upon an empty stomach; having their gruels and their other food well acidulated with it, as also the gargle for their mouth. Two of the worst patients, with the tendons in the ham rigid, (a symptom none of the rest had), were put under a course of seawater. Of this they drank half a pint every day, and sometimes more or less as it operated, by way of gentle physic. Two others had each two oranges and one lemon given them every day. These they eat with greediness, at different times, upon an empty stomach. They continued but six days under this course, having consumed the quantity that could be spared. The two remaining patients, took the bigness of a nutmeg three times a-day, of an electuary recommended by an hospital surgeon, made of garlic, mustard-seed, rad raphan, balsam of Peru, and gum myrrh; using for common drink, barley-water well acidulated with tamarinds; by a decoction of which, with the addition of cremor tartar, they were gently purged three or four times during the course. ... the most sudden and visible good effects were perceived from the use of oranges and lemons, one of those who had taken them being at the end of six days fit for duty [62].

The Treatment Protocol

The treatment protocol (the general term, *study protocol* or *trial protocol* (*see* **Clinical Trials Protocols**) has broader meaning and refers to the constellation of activities involved in conducting a trial) of the trial specifies the treatments being studied, the manner and method of usage and administration, and conditions under which other treatments are called for when needed for the well-being of those enrolled. The treatment may be administrated in one application or multiple applications. The period of treatment may be short (e.g. as in trials involving a single application of treatment such as surgery) or extended (e.g. as in trials involving the treatment of a chronic condition with drugs) over a period of weeks, months, or years. The treatment, in the case of drug trials, may involve a fixed dose administered according to some schedule or dose titration in which each person ultimately receives the amount needed to achieve a desired effect (e.g. the amount of a hypoglycemic agent needed to bring blood glucose levels to within the normal range).

Protocols for all research involving human beings are subject to review and approval by institutional review boards (IRBs) or ethics review boards (ERBs) before implementation and at periodic intervals thereafter until the research is finished (*see* **Ethics of Randomized Trials**). Therefore, investigators undertaking trials have the obligation and responsibility to obtain IRB or ERB review and approval prior to initiation of a trial, and to seek its review and approval prior to implementing amendments to the protocol of the trial. They also have a responsibility to inform IRBs and ERBs of record of any untoward events in the conduct of the trial and to report to such boards any conditions or events believed to change the risk–benefit ratio for persons enrolled into the trial or still to be enrolled.

Only patients judged eligible (as determined by specified **eligibility criteria**) may be enrolled, and among those, only those who consent to participate in the trial. Persons are under no obligation to enroll or to continue once enrolled, and must be so informed prior to being enrolled. A person must be informed, as well, of what is entailed by enrollment, of the risks and benefits that may accrue by enrollment, and of such matters and details that might cause a reasonable person to decline enrollment when so informed (e.g. that treatments are randomly assigned and that they will be administered in masked fashion).

All trials involve data collection at various time points over the course of enrollment and follow-up of persons. The amount collected per person depends on the nature of the disease or condition being treated

and on the nature of the treatment process implied by the study treatments being used. The requirement for repeated observation of a person, as a rule (except for trials done in hospital or other settings involving resident populations or in which enrollment, follow-up, and treatment is directed or managed by telephone or mail), obligates a person to a series of visits to the study site. Usually, the purpose of the first visit or series of visits will be to determine eligibility, collect necessary baseline data, obtain consent, and initiate treatment. Visits thereafter will be to fine-tune or continue treatment and to collect necessary follow-up data. The schedule of follow-up visits will be timed from the point of **randomization** or initiation of treatment and, as a rule, will be on a defined time schedule (e.g. once every week) with provisions for interim (unscheduled) visits when necessary for the care of those enrolled.

Comparison of the different treatments tested for effect is done in different ways depending on the outcome measures used to assess effect. The comparison, in the case of an event, such as death or occurrence of a morbid event, will be based on the event rate (or the raw percentage of persons experiencing the event) as seen for the different treatment groups. In the case of a continuous variable, such as weight or blood pressure, the change from entry to some defined point after enrollment will be determined for each person studied and then summarized in some fashion (e.g. by calculating a **mean** or **median**). The treatment effect will be estimated by the difference obtained by subtracting the summary measure for the comparison treatment from the indicated test treatment.

Judging the safety or efficacy of a treatment is problematic in trials not involving a designed comparison group – often the case in Phase I, II, and I/II trials (see below for definitions). The problem is compounded by the typically short duration and small size of these trials. The problem is most acute in the testing of drugs in people having a life-threatening disease when the drugs themselves carry their own morbidity and increased risk of death. Are the morbid events observed the result of the disease or the drug? Even deaths become difficult to interpret in the presence of a high background death rate from the disease. Was a death the natural outcome of the disease, or was it induced by the treatment? The issue is rarely clear until sufficient information has accumulated to cause one to discount natural causes as the likely explanation, or to allow one to recognize an unusual **clustering** of deaths and morbid events, as with the case of a trial of fialuridine (FIAU) [63].

Classes of Trials

Most clinical trials involve *parallel treatment designs*, i.e. designs where an assignment unit (usually a person) is assigned to receive only one of the treatments under study. The word *parallel* indicates that two or more groups of assignment units are proceeding through the trial side by side, with the only ostensible difference (other than baseline differences in the composition of the groups) being the treatment administered. The goal in trials with parallel treatment designs is for each person enrolled to receive the assigned treatment and to have no exposure to any of the other treatments under study in the trial (except where the requirements for proper care are overriding and make such exposure necessary).

The assignment unit (randomization unit in randomized trials), in the case of parallel treatment designs, is usually a person but can be an aggregate of persons (e.g. members of the same household) (*see* **Group-randomization Designs**) or a subpart of a person (e.g. an eye, as in the Glaucoma Laser Trial [34] (*see* **Unit of Analysis**).

The treatment design in **crossover trials** is different. In this class of designs a person or treatment unit receives two or more study treatments in a specified order. Crossover trials are classified by the number of treatments to be administered to a person or treatment unit and by whether a given person or treatment unit receives all (complete or full crossover) or just some (partial or incomplete crossover) of the study treatments. For example, a two-period crossover design is one in which each person or treatment unit receives two study treatments in some order, usually random. An n-way crossover design is one in which a person or treatment unit receives n of the treatments represented in the design.

The utility of crossover designs is limited to settings in which it is feasible to administer different treatments to the same person or treatment unit, each for a short period of time, and in which it is possible to measure the effect at the end of each treatment period. They are not useful in settings in which the outcome of interest is a clinical event that can occur at any time after enrollment.

In a trial with a parallel treatment design, assignment determines the treatment to be administered

(except to the extent that other treatments are needed for proper care) whereas, in a crossover trial, assignment determines the order of treatments to be used. Typically, each treatment is administered for a designated period of time (e.g. 4 weeks). Often the last administration of one treatment and the first administration of the next treatment are separated in time (e.g. 1 week) to allow the effect of the preceding treatment to "wear off" ("washout period") before administering the next treatment.

Imagine a trial involving three study treatments (A, B, and C) with the same (uniform) assignment probabilities and 54 people. In a trial with a parallel treatment design, 18 people would be assigned to receive treatment A, 18 would be assigned to receive treatment B, and 18 would be assigned to receive treatment C. In a trial involving a complete (full) crossover of treatments, each of the 54 people would receive treatments A,B, and C. Assuming treatments are arranged in all possible orderings, there would be six different orderings of the treatments (ABC, ACB, BAC, BCA, CAB, and CBA), and nine patients would be randomly assigned to receive a given ordering.

While the goal of the two designs is the same, to find the most effective treatment, the methodology differs. With the parallel treatment design, the treatment is evaluated in comparable groups of treatment units (usually persons), and with the crossover treatment design, the treatment effect is evaluated within the same treatment unit (usually a person).

Trials involving parallel treatment designs are of two general types with regard to sample size design – fixed or sequential. The majority are of the fixed type. That is, the sample size is specified at the outset, as determined by pragmatic considerations (e.g. by the amount of money available for the trial) or by a formal **sample size calculation**. Trials are considered to have a fixed sample size even if they do not proceed to the desired sample size, e.g. are stopped early because of a treatment difference. The sample size is fixed in the sense that the intent is to enroll and follow the specified number of assignment units unless indicated otherwise by events transpiring during the course of the trial.

In sequential trials (also of two types – open and closed), enrollment and observation continue until a stopping boundary, constructed for the outcome of primary interest (usually a **binary** "success" or "failure" type event), is crossed. Open sequential designs involve two boundaries, one indicative of superiority and the other indicative of inferiority of a test treatment relative to a comparison treatment. Enrollment continues until the observation function for the outcome measure of interest crosses one of the two boundaries. The design has the advantage of providing a test of the **null treatment hypothesis** for given type I and II error levels (*see* **Hypothesis Testing**) that, on average, requires a smaller sample size than that for a fixed sample size design.

However, the actual sample size required for a boundary crossing can be larger (in theory, sometimes much larger) than that for a fixed sample size design. The possibility of the final sample size being much larger is ruled out with the closed sequential design. That design, in addition to the two boundaries mentioned above, involves a third boundary serving to place an upper bound on enrollment. If that boundary is crossed, because neither of the other two boundaries is crossed (signifying a difference in favor of one of the treatments), then the treatments being compared are considered to be of equivalent value as measured by the outcome observation function (*see* **Sequential Analysis**).

Sequential designs have limited utility in the context of clinical trials, partly because they require rigid adherence to a stopping rule. Use is limited to instances where the "success" or "failure" of a treatment can be determined shortly after administration. They are not useful in settings involving long-term treatment and with outcome measures requiring weeks, months, or years of observation. In general, more flexible methods of monitoring trials are more appropriate (*see* **Data and Safety Monitoring**).

Drug Trials

Compounds, no matter how promising or impassioned the pleas for use, have to go through a series of tests in animals before they can be tested in humans. Those considered to lack promise after animal testing do not come to testing in humans.

Typically, the testing in humans is done in a time-ordered sequence, as suggested by the phase label affixed to trials as defined below. However, in truth, adjoining phases overlap in purpose. Hence, the label, at best, serves only as a rough indicator of the stage of testing, especially when, as is often the case, drug sponsors, at any given point in time, may have several

trials under way carrying different phase labels. The definitions of the different phase labels follow:

Phase I: Usually the first stage of testing performed in anticipation of an Investigational New Drug Application (INDA or NDA); done to generate preliminary information on the chemical action and safety of the indicated drug and to find a safe dose; usually not randomized.

Phase II: Usually the second stage of testing; generally carried out on persons having the disease or condition of interest; done to provide preliminary information on efficacy of the drug and additional information on safety; may be designed to include a control treatment and random assignment of patients to treatment.

Phase I/II: A trial having some of the features of Phase I and II trials; designed to provide preliminary information on safety and efficacy.

Phase III: Usually the third and final stage in testing, prior to submission of an NDA; concerned with assessment of dosage effects, efficacy, and safety; usually designed to include a control treatment and random assignment to treatment. When the test is completed (or nearly completed), the drug manufacturer or sponsor may request permission to market the drug for the indication covered in the testing by submission of an NDA.

Phase II/III: A trial having some of the features of phase II and III trials; designed to provide information on safety and efficacy.

Phase IV: A fourth stage of testing, sometimes carried out. Usually controlled and performed after approval of the NDA. Typically done under circumstances approximating real-world conditions; usually has a clinical event as a basis for sample-size calculation and provides for extended treatment (where appropriate) and long-term follow-up, with efficacy and safety of the drug being measured against a control treatment.

Drugs, after marketing approval, remain under surveillance for serious adverse effects. The surveillance – broadly referred to as **postmarketing surveillance** – involves the collection of reports of adverse events via systematic reporting schemes and via sample surveys and **observational studies**.

Sample size tends to increase with the phase of the trial. Phase I and II trials are likely to have sample sizes in the 10s or low 100s compared to 100s or 1000s for Phase III and IV trials.

The focus shifts with phase. The aim in the early phases of testing is to determine whether the drug is safe enough to justify further testing in human beings. The emphasis is on determining the toxicity profile of the drug and on finding a proper, therapeutically effective dose for use in subsequent testing. The first trials, as a rule, are uncontrolled (i.e. do not involve a concurrently observed, randomized, control-treated group), of short duration (i.e. the period of treatment and follow-up is short), and conducted to find a suitable dose (usually via some traditional or **Bayesian** dose escalation design) for use in subsequent phases of testing. Trials in the later phases of testing, for the most part, involve traditional parallel treatment designs, randomization of patients to study treatments, a period of treatment typical for the condition being treated, and a period of follow-up extending over the period of treatment and beyond.

Most drug trials are done under an investigational new drug application (INDA or IND) held by the sponsor of the drug. The "sponsor" in the vernacular of the **Food and Drug Administration (FDA)** is typically a drug company, but can be a person or agency without "sponsorship" interests in the drug. Regulations require investigators to report adverse events to the FDA. The general guidelines regarding consent are similar, but not identical, to those promulgated by the Office for the Protection from Research Risks (OPRR) for IRBs.

The Randomized Trial

A *randomized trial* is a trial having a parallel treatment design in which treatment assignment for persons (treatment units) enrolled is determined by a

randomization process similar to coin flips or tossings of a die (*see* **Randomized Treatment Assignment**). The trialist's purpose in randomization is to avoid **selection bias** in the formation of the treatment groups. The bias is avoided because the treatment to which a person is assigned is determined by a process not subject to control or influence of the person being enrolled or those responsible for recruiting and enrolling the person. The comparison of one group to another for treatment effect will be biased if, for whatever the reason, one group is "healthier" or "sicker" on entry than the other. Schemes in which one knows or can predict treatment assignments in advance of issue are open to such bias. Clearly, that is the case with assignment schemes posted in a clinic and open for all to see prior to issue. The bias is likely as well with systematic schemes, such as those in which every other person is assigned to the test treatment or in which persons seen on odd-numbered days receive the test treatment and those seen on even-numbered days receive the control treatment.

The goal is to create groups that provide a valid basis for comparison. To achieve that end one has to ensure that the groups are similar (within the range of chance) and to avoid bias in the assignment process. The usual method for achieving both ends is randomization.

Randomization does not guarantee comparability of the treatment groups with regard to the various entry characteristics of interest. Indeed, one can, by chance, have differences among the treatment groups. A large difference (one yielding a small ***P* value**) can arise by chance and, hence, cannot be taken as prima facie evidence of a "breakdown" (e.g. "peeking" or other purposeful acts aimed at determining assignment before issue) of the randomization process, unless supported by other evidence of a "breakdown".

The hallmarks of a sound system of randomization are: reproducible order of assignment; documentation of methods for generation and administration of assignments; release of assignments only after essential conditions satisfied (e.g. only after a person has been judged eligible and has consented to enrollment); masking of assignments to all concerned until needed; inability to predict future assignments from past assignments; clear audit trail for assignments; and the ability to detect departures from established procedures [67] (*see* **Clinical Trials Audit and Quality Control**).

The randomization may be simple (complete) or restricted. The purpose of restriction is to force the assignments to satisfy the specified assignment ratio at intervals during enrollment. Those restrictions are typically referred to as **blocking**. For example, suppose a trial involves two treatments, A and B, and the desired assignment ratio is one-to-one. A simple (unrestricted) randomization scheme would involve the equivalent of repeated flips of an unbiased coin with a head leading to assignment to treatment A and a tail to treatment B. The design would, on average, yield the desired assignment ratio, but allows for wide departures from the desired mix, depending on the "luck" of the flips.

If such departures are of concern, then the randomization scheme can be restricted by blocking so as to ensure the desired mix after a specified number of assignments. For example, imposition of a blocking requirement after every eighth assignment would have the effect of "forcing" the randomization to yield the desired mix of one-to-one after every eighth assignment. The purpose of the blocking is to ensure a near desired assignment ratio so as to protect treatment comparisons against secular trends in the mix of patients as the trial proceeds.

The randomization also may be stratified. The purpose of **stratification** is to provide treatment groups comprised of persons or treatment units having identical (within the limits of the stratification) distributions of the stratification variable. It is useful only in so far as the variable used for stratification serves to influence or moderate the outcome of interest. The stratification has the effect of "controlling" the influence of the stratification variable on outcome by ensuring the same distribution of the variable across the different treatment groups. For example, suppose one wishes to stratify on gender in the trial described above (because, perhaps, of a belief that the treatment effect will be different in women than in men). The stratification would be achieved by creating two randomizations schedules, each with a one-to-one assignment ratio and with blocking to satisfy the assignment ratio after enrollment of the 8th, 16th, 24th, etc. person in each stratum. The effect of the stratification would be to ensure the same gender mix (within the limits of the blocking) for the two treatment groups, regardless of the underlying gender mix of the population to be studied. For example, suppose 96 patients are to be enrolled from a population with a 1:2 mix of males to females. In that case, one would

expect to enroll 32 males and 64 females, and to have 16 males and 32 females in each treatment group. If the underlying mix is one-to-one, then there would be 24 males and 24 females in each of the two treatment groups.

Clearly, the number of variables that can be controlled by stratification is limited. The more variables, the more subgroups for randomization and the less useful the process is as a reliable means of **variance** control [35]. In addition, there are logistic difficulties associated with use of variables whose values have to be determined by performing laboratory tests or other diagnostic procedures during the enrollment process. Even if one stratifies on a few selected variables, other variables may well be considered to be important determinants of outcome. Hence, the experienced trialist strives to "remove" the effect of such differences via analysis procedures, e.g. by assessing the treatment effect within defined subgroups (subgroup analysis, *see* **Treatment-covariate Interaction**); or by providing estimates of treatment effect that are adjusted for differences in the distribution of important demographic or baseline variables via **regression** procedures [90].

Masking

Masking is the purposeful concealment of some fact or condition and is done to keep knowledge of that fact or condition from influencing the behavior, observation, or reporting of persons so masked. Masking, in the context of trials, is imposed to reduce the likelihood of a treatment-related bias due to knowledge of treatment assignment (*see* **Blinding or Masking**).

That bias, after a person is enrolled, occurs whenever knowledge of that person's treatment assignment serves to color the way he or she is treated, followed, or observed. One way of reducing it is by masked treatment administration. In one form of such administration (single-masked), only one member of the subject–treater pair is masked to treatment, usually the subject. Another form of masking is one in which both members of the pair are masked – double-masked treatment administration. As a rule, double-masked treatment administration means that all persons in a clinic are masked and, therefore, that those responsible for data collection and generation are masked to treatment as well.

Generally, it is not possible or prudent to mask treatment administration in trials involving treatments requiring different routes or modes of administration (e.g. as in a trial involving a medical vs. a surgical form of treatment), where knowledge of treatment assignment is part of the effect being tested (e.g. as in trials aimed at modification of one's eating habits via different modes of dietary consulting), or where the masking carries risks for those enrolled. Therefore, the opportunities for double-masked treatment administration are limited largely to trials of drugs considered safe and that are reasonably free of side-effects and that can be administered at fixed dose levels. It is usually not wise or practical to administer treatments in a double-masked fashion when treatment doses are to be titrated to achieve desired effects.

Masked treatment administration has been used as a mark of "quality" for trials. There is, therefore, a tendency to view results from masked trials as more reliable than those from unmasked trials. In truth, however, masked treatment administration is rarely 100% effective. All forms of treatment, and especially those involving drugs, produce side-effects and telltale signs that may serve to unmask treatment. Hence, the protection provided by masking can be illusory. As a result, it is better to make assessments of "quality" in terms of the risk of treatment-related bias and the likely effect of such bias, if present, on the results reported. The risk of treatment-related bias is low for "hard" outcome measures and with explicitly defined treatment protocols, even in the absence of masked treatment administration.

The second line of defense, in the absence of double-masked treatment administration, is to mask as many groups of persons involved in the trial as is possible within the limits of practicality and safety. Hence, even if it is not possible to mask patients or those who treat them, it may be possible to mask those responsible for data collection or data generation (e.g. as with an arrangement as in the Glaucoma Laser Trial [34], where intraocular pressure was measured by masked readers, or as with laboratory personnel or readers of X-rays, ECGs, or fundus photographs masked to treatment assignment).

With or without treatment masking, trialists strive for objectively defined treatment and data collection procedures and for outcome measures as free from observer or respondent bias as is humanly possible. In addition, they are inclined toward continuing effort

over the course of a trial aimed at maintaining the training and certification of study personnel in regard to required study procedures, and toward establishing and maintaining standards of performance via ongoing monitoring and quality control surveillance (*see* **Clinical Trials Audit and Quality Control**).

Analysis

The protection provided against treatment-related bias by the assignment process is futile if the analysis is **biased**. Treatment comparisons, to be valid, must be based on analyses that are consistent with the design used to generate them. In the case of the randomized trial, this means that the primary analyses of the outcomes of interest must be by assigned treatment (also known as *analysis by* **intention-to-treat**). It means, for example, that observations relating to a morbid event are counted to a patient's assigned treatment regardless of whether or not the patient was still on the assigned treatment when the event occurred.

Analyses involving arrangements of data related to treatment administered may be performed, but only as supplements to the primary analyses. They should not and cannot serve as replacements for those analyses.

Analyses by treatment assignment, as a rule, serve to underestimate the treatment effect. Usually, analyses in which the requirement is relaxed will yield a larger estimate of the treatment difference than seen when evaluated under the intention-to-treat mode of analysis (e.g. as in the case of the **University Group Diabetes Program (UGDP)** trial) [90].

Designs allowing for termination of data collection when a person can no longer receive or be maintained on the assigned treatment are open to treatment-related bias. The goals of the primary analyses cannot be met when data collection for a person ceases when that person experiences a nonfatal "endpoint" or when the person's treatment is stopped or changed. The analysis requirement implies continued follow-up of all persons enrolled into a trial to the scheduled close of follow-up regardless of their treatment or outcome status.

Monitoring Treatment Effects

The randomized trial depends on a state of equipoise – a state of legitimate doubt regarding the test treatment relative to the control treatment(s) [4, 30, 61]. It cannot be undertaken without a proper ethical climate characterized by such a state of doubt (*see* **Ethics of Randomized Trials**). It does not matter whether that state has been dispelled by observation and data, by declaration, or in other ways. For example, it would not be possible to assess the value of coronary care units for persons appearing to be having a myocardial infarction (MI), even if their value has not been demonstrated by controlled trials. They are considered to be required for good care and, hence, the window of opportunity for testing via designed randomized trials has closed. Once closed, it may remain closed, or may open again years later if people start questioning the merits of the treatment. When the oral hypoglycemic agents appeared on the scene in the early 1950s, they were widely regarded as safe and effective and, hence, became a part of the armamentarium for care of the adult-onset diabetic. However, doubts raised in the late 1950s as to their value led to a climate of doubt suitable for initiation of the UGDP trial [89].

Trials are done because of the prospect of benefit associated with a new treatment, or to test the efficacy of an existing treatment. They are not undertaken to prove a treatment to be useless or harmful. Indeed, a trialist is obligated to stop a trial prior to its scheduled completion if the accumulated data indicate that the treatment of interest is inferior to the control or comparison treatment. In fact, some argue that there is an obligation to stop if it becomes clear that the test treatment is no better than the comparison treatment, even if one is uncertain whether it is harmful. Hence, for example, investigators in the UGDP opted to stop use of tolbutamide in that trial once they were certain it was no better than the control treatment – the usual antidiabetic dietary recommendations and placebo medication.

The need for ongoing monitoring exists for any trial in which the treatments carry risk of harm, and in which it is possible to reduce that risk by timely monitoring (*see* **Data and Safety Monitoring**). That need makes it necessary for the trialist to aim for an orderly and timely flow of data from the site of generation or collection to the processing and analysis site. Clearly, the best systems in this regard are those having real-time or near-real-time flows (e.g. as with systems requiring transmission of data related to a patient visit on completion of the visit or on occurrence of an outcome of interest).

Typically, treatment effects monitoring is entrusted to a group of people that together have the necessary skills and expertise to monitor effectively (*see* **Data Monitoring Committees**) [38, 67, 70, 94]. The group is usually comprised of 5–12 people with expertise in the disease under treatment, in the design, conduct, and analysis of clinical trials, or in other specialty areas. When the group comprises a mix of people from within the trial (e.g. the officers of the trial, such as the chair and vice chair, the director of the coordinating center, etc.) and outside the trial, the votes concerning recommendations for change, generally, are vested in those outside the study. The restriction is imposed, typically, because of concerns that persons associated with the trial may have conflicts of interest that could serve to influence their votes [17].

Monitoring proceeds under different constructs, depending on the philosophy of those doing the monitoring. Some constructs require stopping rules and restrictions on the type of data that may be monitored and the number of interim "looks" that can be made in relation to the monitoring. Other groups consider such restrictions unnecessary and rely instead on the collective judgment of the monitoring group [28].

The Office for Protection from Research Risks (OPRR) (an office within the **National Institutes of Health** responsible for the promulgation and administration of regulations regarding institutional review boards) and the set of rules relating to research on human beings obligates IRBs to be satisfied that risk to subjects is minimized. As part of this assurance in regard to clinical trials, investigators must have "adequate provision for monitoring the data collected to ensure the safety of subjects" (Section 46.111) [71] and must provide participants with information on ... "significant new findings developed during the course of the research which may relate to the subject's willingness to continue participation will be provided to the subject" (Section 46.116) [71]. This requirement makes it necessary to inform patients of results during the trial that bear on their willingness to continue. This requirement pertains to information from inside or outside the study, if the information is likely to cause patients to reconsider their decision to be enrolled in the trial. Formal reconsent procedures may be required if the treatment effects monitoring committee recommends changes to the treatment protocol (e.g. as discussed in [67]).

Representativeness, Validity, and Generalizability

Representativeness, in the context of a trial, refers to the degree or extent to which those enrolled can be considered representative of the general population of persons to whom the treatment may be applied, if shown to be useful. *Validity*, in the context of a treatment difference, refers to the extent to which that difference can be reasonably attributed to treatment assignment. *Generalizability* refers to the degree to which the findings of the trial can be extended to the general population of eligible persons.

The concepts of validity and generalizability are different. Validity derives from the design of the trial and from the way it is carried out, whereas generalizability is largely a matter of judgment. A treatment comparison is valid if it is based on comparable groups of persons treated and observed in such a way so as to make treatment assignment the most likely explanation of the result observed. "Representativeness" is deduced by comparison of the demographic and other host characteristics of the study population to that of the general population of eligible persons (or by comparison with all persons screened for enrollment).

The desire for representativeness arises from the belief that conclusions from a trial will be strengthened by having a broadly "representative" study population. The drive for demographic representativeness has been propelled in recent years by the belief that women and persons of ethnic minorities have been "underrepresented" or "understudied" relative to men and the prevailing ethnic majority in trials and other areas of clinical research. Those concerns have been sufficient to cause the US Congress, in the NIH Revitalization Act of 1993, to impose requirements on trials aimed at ensuring adequate numbers of women and ethnic minorities to determine whether the treatments being studied in a trial work differently in men than in women or in an ethnic minority than in the ethnic majority [88].

There is no way to ensure "representativeness" in the absence of a **sampling frame** for the eligible **study population** and a related sampling scheme aimed at providing a representative sample of that population. However, even if one were able to develop a sampling frame (usually impossible because to do so one would have to screen the general population to identify persons eligible for study), the

population ultimately enrolled, even if selected by sampling, would, at best, be representative only of those able and willing to be enrolled, because of the requirements of consent.

Hence, trials, by nature of their design, involve select, nonrepresentative populations. Even if a treatment is found effective in a trial, one has no direct way of knowing if it would be effective for those patients not agreeing to be studied. If the issues of consent and lack of a sampling frame were overcome, then one would still be left with the fact that most clinics, for practical and ethical reasons, have to rely on those who come to them. They do not have the ability or moral authority to go and seek out suitable patients for study, especially if doing so means that those who routinely come to them would be turned away. Such a "selective" approach would be viewed as violating the principle of justice as set forth in the Belmont Report [69].

That one needs to generalize is obvious. The need arises in regard to the route of treatment, amount of treatment, type of treatment, and type of patients. For example, if a trial involved a single fixed dose of a drug and failed to find a difference (e.g. as in the UGDP trial, regarding tolbutamide) [90] does one conclude that use of the same drug, under a different, more flexible dosing scheme would produce a more favorable result? Similarly, if one compound produces a benefit, does one conclude that other sister compounds will show the same effect? Or conversely, if one member of a drug family has a bad effect (e.g. fialuridine) or fails to show a benefit (e.g. tolbutamide), does one shy away from other related compounds? Also, if a trial involves mildly diseased people and shows a beneficial effect for the test treatment, does one conclude that the test treatment will have a similar effect in sicker people?

Last, if the drug tested works for the disease or condition being treated, is it not likely that it would be useful as well for a related condition or disease? So-called "off label" use (from the fact that drugs are approved for designated indications) accounts for a large number of treatment prescriptions [12, 36, 91, 93].

Whenever one generalizes, whatever the nature or direction, one is in effect answering one or more of the above questions. If as a treater, one chooses to use a sister compound of a drug shown to be ineffective in a trial, then one is in effect saying that the result from the trial, for whatever reason, is not generalizable. Generalizations depend on judgments regarding the trials and on prior beliefs regarding the treatment in question.

A trial can provide a valid basis for comparing one treatment to another if the differences in outcomes for the treatment groups being compared can be attributed reasonably to treatment. The general "laws of science" and "principles of **parsimony**" require that one defaults to the simplest explanation – usually the one requiring the fewest assumptions. Hence, in the case of the trial in which treatments are selected by the patient or physician, one is as a rule more inclined to attribute the difference to selection factors than to the test treatment. By the same principle, one should be more inclined to attribute a treatment difference to bias on the part of the observer rather than to the treatment when the opportunity for such bias exists. The degree of "reasonableness" of such an explanation will depend on the nature of the outcomes and whether one can reasonably ascribe it to biased observation. It becomes progressively more difficult to do so, even if the observer is not masked, the "harder" the outcome measure. For example, it is not reasonable to expect that one's opinion regarding the merits of a treatment will influence one's ability to report reliably whether a person is alive or dead, but such opinion may influence how one sees or reports on a person's **quality of life**. There is a responsibility on the part of trialists to rule out other lesser explanations of results before ascribing them to treatment.

Contrary to lay perceptions, trials and comparisons of treatments within the trial are made **robust** to selection bias and the consequences of "nonrepresentative" study populations by randomization. The assessment of treatment effect is achieved by having comparable groups of patients in the different treatment groups and by having procedures for observing and following patients that are independent of treatment assignment. The comparison is valid regardless of the study population and provides information on the relative value of one treatment to another. Hence, from the perspective of the trialist, it is far more important to have comparable treatment groups than to have "representative" treatment groups.

The drive for "representativeness", while perhaps of some social value, does little to make generalizations less risky or to increase the validity of trials. There are sound practical reasons to design trials with as few exclusions to enrollment as possible.

Table 1 References on methods and procedures of clinical trials

Topic	References
Specialty journals	
Applied Clinical Trials	[2]
Controlled Clinical Trials	[37]
Statistical Methods in Medical Research	[85]
Statistics in Medicine	[19, 20]
Textbooks	
Clinical trials	[15, 33, 44, 47, 67, 75, 82]
Data analysis	[40, 41]
Ethics	[51, 60]
History	[87
Dictionaries/Encyclopedias	
Clinical trials	[66]
Epidemiology	[58]
Statistics	[53]
Journal articles	
Analysis	[18, 24, 25, 50, 74, 83]
Bayesian methods	[6, 21, 31, 84]
Cost and efficiency	[95]
Design	[18]
Equipoise	[4], 30, 61, 76]
Ethics	[3, 4, 69]
Forms design and data management	[39, 81, 96, 97, 98]
History	[13, 62]
Meta-analysis and overviews	[5, 9, 14, 40, 48, 59, 86, 99]
Philosophy	[78, 79]
Randomization and stratification	[11, 35, 49, 56, 64, 73, 80, 92]
Sample size	[7, 8, 10, 27, 54, 55, 57, 77]
Subgroup analyses	[6, 23, 100]
Treatment effects monitoring	[3, 16, 26, 28, 29, 38, 43, 52, 59, 76]

The above list is due the efforts of Susan Tonascia, ScM, Johns Hopkins School of Public Health, Department of Epidemiology.

The fewer the restrictions the easier and faster it is to recruit. Any effort to make them more "representative" by selective recruitment and enrollment will make them more costly and will increase the time required to enroll them. The imposition of recruitment quotas to achieve a desired sample size for gender, age, and ethnic origin groups poses a far more complicated and costly recruitment effort than one involving the enrollment of all comers regardless of gender, age, or ethnic origin.

The goal of the trialist should be to strive for demographic neutrality in enrollment. That is to say, the trialist should not exclude potential participants on the basis of gender, ethnic origin, or age unless justified on scientific grounds. Scientific grounds include the knowledge or expectation of a qualitative treatment by demographic interaction (i.e. where treatment is believed to be beneficial for one demographic group and harmful for another) (*see* **Treatment-covariate Interaction**).

Another reason for exclusion is contraindication of treatment in a particular demographic group. If any one treatment is contraindicated in a trial involving multiple treatments, then the restriction has to apply to all treatments. For example, this requirement was one of the reasons why the Coronary Drug Project (CDP) involved only men. Two of the five test treatments in the trial could not have been administered to premenopausal women; thus this demographic group could not be included without making the trial much more complex [22].

As a rule, an anticipated low number in a specified demographic group is not a reason to exclude. Disease and extent of disease are much more likely to affect the response to treatment than are "demographic" characteristics. Analyses of treatment effects across the various demographic subgroups represented in a trial can help determine whether there are treatment by demographic interactions. In general, **interactions**, when noted in the context of treatment trials, are more likely to relate to disease characteristics than to demographics [32, 42, 45, 46, 65].

The mind-set regarding selection is different in **prevention trials**, where the goal is to determine whether a proposed prevention strategy works. One has to find a population suitable for testing the proposed strategy. Hence, unlike the treatment trial, risk factors predisposing to a disease and risk of an event are important. In this setting, one has to pay attention to both factors in trying to design a cost-effective trial (*see* **Health Economics**). Considerations of this sort led, for example, the designers of MRFIT [68] to exclude females from enrollment. The risk factors targeted (high blood pressure, high cholesterol, and smoking) occur less frequently in women than in men. Consequently, the effort required to find women for study would have been much greater than that required to find men. Further, for the age range studied, women have a markedly lower myocardial infarction rate (the outcome of primary interest) than do men. This lower event rate would have meant that the planned sample size with women included would have to have been considerably larger to detect the same relative difference at the power level specified for the trial. As it was, the trial required a sample size of 12 866 men (*see* **Validity and Generalizability in Epidemiologic Studies**).

Readings

The literature on the design, conduct, and analysis of clinical trials is ever-expanding. Students of trials need to monitor the literature of reported trials as they appear in medical journals and to read specialty journals, such as ***Biometrics***, ***Statistics in Medicine***, ***Controlled Clinical Trials***, *Applied Clinical Trials*, and ***Statistical Methods in Medical Research***. The list of citations given in Table 1 is but a snapshot of selected references dealing with the methods and procedures of trials.

References

[1] American Bible Society (1816). *The Holy Bible: Old and New Testaments, King James Version (1611)*. American Bible Society, New York.

[2] *Applied Clinical Trials* (1992–97),. Astor, Eugene.

[3] Ashby, D., ed. (1993). Conference on methodological and ethical issues in clinical trials, 27–28 June, 1991, *Statistics in Medicine* **12**, 1373–1534.

[4] Beecher, H.K. (1966). Ethics and clinical research, *New England Journal of Medicine* **274**, 1354–1360.

[5] Berlin, J.A., Laird, N.M., Sacks, H.S. & Chalmers, T.T. (1989). A comparison of statistical methods for combining event rates from clinical trials, *Statistics in Medicine* **8**, 141–151.

[6] Berry, D.A. (1985). Interim analysis in clinical trials: classical vs Bayesian approaches, *Statistics in Medicine* **4**, 521–526.

[7] Blackwelder, W.C. (1982). "Proving the null hypothesis" in clinical trials, *Controlled Clinical Trials* **3**, 345–353.

[8] Blackwelder, W.C. & Chang, M.A. (1984). Sample size graphs for "proving the null hypothesis", *Controlled Clinical Trials* **5**, 97–105.

[9] Boissel, J.P., Blanchard, J., Panak, E., Peyrieux, J.C. & Sacks, H. (1989). Considerations for the meta-analysis of randomized clinical trials: summary of a panel discussion, *Controlled Clinical Trials* **10**, 254–281.

[10] Bristol, D. (1989). Sample sizes for constructing CI's and testing hypotheses, *Statistics in Medicine* **8**, 803–821.

[11] Brown, B.W., Jr (1980). Designing for cancer clinical trials: Selection of prognostic factors, *Cancer Treatment Reports*, **64**, 499–502.

[12] Brosgart, C.L., Mitchell, T., Charlebois, E., Coleman, R., Mehalko, S., Young, J. & Abrams, D.I. (1996). Off-label drug use in human immunodeficiency virus disease, *Journal of Acquired Immune Deficiency Syndrome and Human Retrovirology* **12**, 56–62.

[13] Bull, J.P. (1959). The historical development of clinical therapeutic trials, *Journal of Chronic Diseases* **10**, 218–248.

[14] Buyse, M. & Ryan, L.M. (1987). Issues of efficiency in combining proportions of deaths from several clinical trials, *Statistics in Medicine* **6**, 565–576.

[15] Buyse, M.E., Staquet, M.J. & Sylvester, R.J. (1984). *Cancer Clinical Trials: Methods and Practice*. Oxford University Press, New York.

[16] Canner, P.L. (1977). Monitoring clinical trial data for evidence of adverse or beneficial treatment effects, *INSERM – Essais Controles Multicentres: Principes et Problemes* **76**, 131–149.

[17] Chalmers, T.C. & Amacher, P. (1982). Conference on recent history of randomized clinical trials (with discussion), *Controlled Clinical Trials* **3**, 299–309.

[18] Colton, T, Freedman, L.S., Johnson, A.L. & Machin, D. (1989). Recent issues in clinical trials, *Statistics in Medicine* **8**, 401–516.

[19] Colton, T., Freedman, L.S., Johnson, A.L. & Machin, D. (1991). Cumulative decade index, *Statistics in Medicine* **1–10**, 1–179.

[20] Colton, T., Freedman, L.S., Johnson, A.L. & Machin, D. (1991). Tenth anniversary issue, *Statistics in Medicine* **10**, 1789–2027.

[21] Cornfield, J. (1969). The Bayesian outlook and its application (including discussion by S. Geisser, H.O. Hartley, O. Kempthorne & H. Rubin), *Biometrics* **25**, 617–657.

[22] Coronary Drug Project Research Group (1973). The Coronary Drug Project: Design, methods, and baseline results, *Circulation* **47**(Supplement 1), I-1–I-50.

[23] Coronary Drug Project Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project, *New England Journal of Medicine* **303**, 1038–1041.

[24] Cox, D.R. (1972). Regression models and life tables (including discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[25] D'Agostino, R.B., Lange, N., Ryan, L.M. & Selwyn, M.R., eds (1992). Symposium on longitudinal data analysis, Fort Lauderdale, FL, June 19–21, 1991, *Statistics in Medicine* **11**, 1797–2040.

[26] DeMets, D.L. (1987). Practical aspects in data monitoring: a brief review, *Statistics in Medicine* **6**, 753–760.

[27] Donner, A. (1984). Approaches to sample size estimation in the designing of clinical trials – a review, *Statistics in Medicine* **3**, 199–214.

[28] Ellenberg, S., Geller, N., Simon, R. & Yusuf, S. (1993). Practical issues in data monitoring of clinical trials (Proceedings), Bethesda, MD, January 27–28, 1992, *Statistics in Medicine* **12**, 415–616.

[29] Fleming, T.R. & DeMets, D.L. (1993). Monitoring of clinical trials: issues and recommendations, *Statistics in Medicine* **14**, 183–197.

[30] Freedman, B. (1987). Equipoise and the ethics of clinical research, *New England Journal of Medicine* **317**, 141–145.

[31] Freedman, L.S. & Spiegelhalter, D.S. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials, *Controlled Clinical Trials* **10**, 357–367.

[32] Freedman, L.S., Simon, R., Foulkes, M.A., Friedman, L., Geller, N.L., Gordon, D.J. & Mowery, R. (1995). Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993-The perspective of NIH clinical trialists (with discussion and response), *Controlled Clinical Trials* **16**, 277–312.

[33] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1985). *Fundamentals of Clinical Trials*, 2nd Ed. PSG Publishing, Boston.

[34] Glaucoma Laser Trial Research Group (1991). The Glaucoma Laser Trial (GLT): 3. Design and methods, *Controlled Clinical Trials* **12**, 504–524.

[35] Grizzle, J.E. (1982). A note on stratifying versus complete random assignment in clinical trials, *Controlled Clinical Trials* **3**, 365–368.

[36] Grossman, E., Messerli, F.H., Grodzicki, T. & Kowey, P. (1996). Should a moratorium be placed on sublingual nifedipine capsules given for hypertensive emergencies and pseudoemergencies? *Journal of the American Medical Association* **276**, 1328–1331.

[37] Hawkins, B.S. (1991). Controlled clinical trials in the 1980s: a bibliography, *Controlled Clinical Trials* **12**, 1–272.

[38] Hawkins, B.S. (1991). Data monitoring committees for multicenter clinical trials sponsored by the National Institutes of Health: I. Roles and memberships of data monitoring committees for trials sponsored by the National Eye Institute, *Controlled Clinical Trials* **12**, 424–437.

[39] Hawkins, B.S. (1995). Data Management for multicenter studies: methods and guidelines, *Controlled Clinical Trials* **16**(Supplement 2), 1S–179S.

[40] Hedges, L. & Olken, I. (1987). *Statistical Methods for Meta-Analyses*. Academic Press, Orlando.

[41] Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

[42] Holbrook, J.T., Meinert, C.L. & Gilpin, A.K. (1994). On the likelihood of finding subgroup differences in clinical trials, *Controlled Clinical Trials* **15**, 129S–130S.

[43] Hughes, M.D. & Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trials, *Statistics in Medicine* **7**, 1231–1242.

[44] Hulley, S.B., Cummings, S.R., Browner, W.S., Newman, T.B. & Hearst, N. (1988). *Designing Clinical Research: An Epidemiologic Approach*. Williams & Wilkins, Baltimore.

[45] Institute of Medicine, Committee on the Ethical and Legal Issues Relating to the Inclusion of Women in Clinical Studies (1994). *Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies*, Vol. 1. National Academy Press, Washington.

[46] Institute of Medicine, Committee on the Ethical and Legal Issues Relating to the Inclusion of Women in Clinical studies (1994). *Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies, Workshop and Commissioned Papers*, Vol. 2. National Academy Press, Washington.

[47] Johnson, F.N. & Johnson, S., eds (1977). *Clinical Trials*. Blackwell Scientific, Oxford.

[48] Jones, D.R. (1995). Meta-analysis: weighing the evidence, *Statistics in Medicine* **14**, 137–149.

[49] Kalish, L.A. & Begg, C.B. (1985). Treatment allocation methods: a review, *Statistics in Medicine* **4**, 129–144.

[50] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[51] Katz, J. (1972). *Experimentation with Human Beings*. Russell Sage Foundation, New York.

[52] Kiri, A. & Meinert, C.L. (1995). Treatment effects monitoring practices as viewed through the published literature (abstract), *Controlled Clinical Trials* **16**(Supplement 3), 58S.

[53] Kotz, S., Johnson, N.L. & Read, B.C., eds (1982–1988). *Encyclopedia of Statistical Sciences*, Vols 1–9. Wiley, New York.

[54] Lachin, J.M. (1981). Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials* **2**, 93–114.

[55] Lachin, J.M. & Foulkes, M.A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance and stratification, *Biometrics* **42**, 507–519.

[56] Lachin, J.M., Matts, J.P. & Wei, L.J. (1988). Randomization in clinical trials: conclusions and recommendations, *Controlled Clinical Trials* **9**, 365–374.

[57] Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials, *Biometrics* **44**, 229–242.

[58] Last, J.M., ed. (1995). *A Dictionary of Epidemiology*, 3rd Ed. Oxford University Press, New York.

[59] Lee, Y.J., ed. (1996). Conference on meta-analysis in the design and monitoring of clinical trials, *Statistics in Medicine* **15**, 1233–1323.

[60] Levine, R.J. (1986). *Ethics and Regulation of Clinical Research*, 2nd Ed. Yale University Press, New Haven.

[61] Lilford, R.J. & Jackson, J. (1995). Equipoise and the ethics of randomization, *Journal of the Royal Society of Medicine* **88**, 552–559.

[62] Lind, J. (1753). *A Treatise of the Scurvy*. Sands, Murray, Cochran, Edinburgh. Reprinted in *Lind's Treatise on Scurvy*, C.P. Stewart & D. Guthrie, eds. Edinburgh University Press, Edinburgh, 1953.

[63] McKenzie, R., Fried, M.W., Sallie, R., Conjeevarm, H., Di Bisceglie, A.M., Park, Y., Savarese, B., Kleiner, D., Tsokos, M., Luciano, C., Pruett, T., Stotka, J.L., Straus, S.E. & Hoofnagle, J.H. (1995). Hepatic failure and lactic acidosis due to fialuridine (FIAU), an investigational nucleoside analogue for chronic hepatitis B, *New England Journal of Medicine* **333**, 1099–1105.

[64] Meier, P. (1981). Stratification in the design of a clinical trial, *Controlled Clinical Trials* **1**, 355–361.

[65] Meinert, C.L. (1995). The inclusion of women in clinical trials, *Science* **269**, 795–796.

[66] Meinert, C.L. (1996). *Clinical Trials Dictionary: Terminology and Usage Recommendations*. The Johns Hopkins Center for Clinical Trials, Baltimore.

[67] Meinert, C.L. & Tonascia, S. (1986). *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

[68] Multiple Risk Factor Intervention Trial Research Group (1977). Statistical design considerations in the NHLBI Multiple Risk Factor Intervention Trial (MRFIT), *Journal of Chronic Diseases* **30**, 261–275.

[69] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. US Government Printing Office, Washington.

[70] NIH Clinical Trials Committee (1979). Clinical activity, *NIH Guide for Grants and Contracts* **8**, 29 (R.S. Gorden, chair, June 5).

[71] Office for Protection from Research Risks (1991). *Code of Federal Regulations, Title 45: Public Welfare, Part 46: Protection of Human Subjects*, rev. June 18. Department of Health and Human Services, National Institutes of Health, Bethesda.

[72] Packard, F.R. (1921). *Life and Times of Ambroise Paré, 1510–1590*. Paul B. Hoeber, New York.

[73] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design, *British Journal of Cancer* **34**, 585–612.

[74] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K.,

Peto, J. & Smith, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. Analysis and examples, *British Journal of Cancer* **35**, 1–39.

[75] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.

[76] Pocock, S.J. (1993). Statistical and ethical issues in monitoring clinical Trials, *Statistics in Medicine* **12**, 1459–1475.

[77] Rubinstein, L.V., Gail, M.H. & Santner, T.J. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation, *Journal of Chronic Diseases* **34**, 469–479.

[78] Sackett, D.L. & Gent, M. (1979). Controversy in counting and attributing events in clinical trials, *New England Journal of Medicine* **301**, 1410–1412.

[79] Schwartz, D. & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutic trials, *Journal of Chronic Diseases*, **20**, 637–648.

[80] Senn, S.J. (1989). Covariate imbalance and random allocation in clinical trials, *Statistics in Medicine* **8**, 467–476.

[81] Singer, S.W. & Meinert, C.L. (1995). Format-independent data collection forms, *Controlled Clinical Trials* **16**, 363–376.

[82] Smith, P.G. & Morrow, R.H., eds (1991). *Methods for Field Trials of Interventions against Tropical Diseases: A "Toolbox"*. Oxford University Press, Oxford.

[83] Souhami, R.L. & Whitehead, J., ed, (1994). Workshop on early stopping rules in cancer clinical trials (Robinson College, Cambridge, UK, April 13–15, 1993), *Statistics in Medicine* **13**, 1293–1499.

[84] Spiegelhalter, D.S. & Freedman, L.S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion, *Statistics in Medicine* **5**, 1–13.

[85] *Statistical Methods in Medical Research* (1992–1996). Edward Arnold. Hodder & Stoughton Ltd, Oxford, Vols 1–5.

[86] Stewart, L.A. & Clarke, M.J. (on behalf of the Cochrane Working Group in Meta-Analysis) (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data, *Statistics in Medicine* **14**, 2057–2079.

[87] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press, Harvard University Press, Cambridge, Mass.

[88] United States Congress (1993). *National Institutes of Health Revitalization Act of 1993*, §131, Pub. L No. 103–43, 107 Stat. 133 (codified at 42 USC §289a-2).

[89] University Group Diabetes Program Research Group (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: I. Design, methods and baseline characteristics, *Diabetes* **19**(Supplement 2), 747–783.

[90] University Group Diabetes Program Research Group (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: II. Mortality results, *Diabetes* **19**(Supplement 2), 785–830.

[91] US General Accounting Office (1996). *Prescription Drugs: Implications of Drug Labeling and Off-Label Use* (testimony before US House of Representatives, 12 September, by Sarah F. Jagger). General Accounting Office (GAO/T-HEHS-96-212), Washington.

[92] Wei, L.J., Smythe, R.T. & Smith, R.L. (1986). K treatment comparisons with restricted randomization rules in clinical trials, *Annals of Statistics* **14**, 265–274.

[93] Winker, M.A. (1996). The FDA's decision regarding new indications for approved drugs: where's the evidence?, *Journal of the American Medical Association* **276**, 1342–1343.

[94] Wittes, J. (1993). Behind closed doors: The data monitoring board in randomized clinical trials (with discussion), *Statistics in Medicine* **12**, 419–434.

[95] Wittes, J., Duggan, J., Held, P. & Yusuf, S., eds (1990). Cost and efficiency in clinical trials (workshop proceedings; sponsored by the National Heart, Lung, and Blood Institute, Bethesda, MD, January 18–19, 1989), *Statistics in Medicine* **9**, 1–199.

[96] Wright, P. & Haybittle, J. (1979). Design of forms for clinical trials (1), *British Medical Journal* **2**, 529–530.

[97] Wright, P. & Haybittle, J. (1979). Design of forms for clinical trials (2), *British Medical Journal* **2**, 590–592.

[98] Wright, P. & Haybittle, J. (1979). Design of forms for clinical trials (3), *British Medical Journal* **2**, 650–651.

[99] Yusuf, S., Simon, R. & Ellenburg, S.S., eds (1987). Proceedings of the workshop on methodologic issues in overviews of randomized clinical trials, May 1986. Sponsored by the National Heart, Lung, and Blood Institute and the National Cancer Institute, Bethesda, MD, *Statistics in Medicine* **6**, 217–409.

[100] Yusuf, S., Wittes, J., Probstfield, J. & Tyroler, H.A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials, *Journal of the American Medical Association* **266**, 93–98.

CURTIS L. MEINERT

Clinical Trials

Clinical Trials: Journal of the Society for Clinical Trials, initiated publication in 2004, succeeding ***Controlled Clinical Trials*** (CCT) as the official journal of the **Society for Clinical Trials** (SCT, www.sctweb.org). The SCT provided editors for CCT since its initiation by Elsevier in 1980, the first editor being Curt Meinert (1980–1993), followed by Janet Wittes (1994–1998) and Jim Neaton (1999–2003). In 2004, the SCT founded a new journal, *Clinical Trials*, under the editorship of Steven Goodman. The SCT owns the copyright of its new journal, which is published by Hodder–Arnold of the United Kingdom.

The aims and scope of *Clinical Trials* mirror the makeup and interests of the sponsoring society, and are listed on the journal website, www.sctjournal.com. It is an interdisciplinary journal with coverage of virtually any method or issue related to or impacted by **clinical trials**. These include statistical methods, methods for design, conduct, monitoring (*see* **Data and Safety Monitoring**), or synthesis (*see* **Meta-analysis of Clinical Trials**) of clinical trials; research ethics (*see* **Ethics of Randomized Trials**); law, policy, and regulation (*see* **Drug Approval and Regulation**); **history of clinical trials**; impact of trials; and education and training. The journal also seeks to feature discussions of timely and important issues related to clinical trials. This will be facilitated by the SCTs initiation in 2004 of a process for producing society position papers on important methodologic or regulatory issues. In addition to articles that cover technical methods, the journal also publishes perspectives, profiles of prominent trialists, historical pieces, book reviews, and editorials. Full-length articles can be up to 7000 words, brief reports are limited to 1500 words, and research letters are 500 words or less. *Clinical Trials* publishes design articles, which describe the design and organization of particular clinical trials, focusing on how particularly interesting or challenging design issues were addressed. It will also be adding a feature popular in the first decade of CCT, a column that provides an overview of clinical trial–related articles, methodologies, and developments appearing in the wider scholarly and policy arena.

In addition to the Editor-in-Chief, the journal has a Deputy Editor, a 10-person Advisory Board, and 40 Associate Editors. The Associate Editors come from a broad array of disciplines and venues and have wide-ranging expertise: statistics, ethics, history, informatics, medicine, industry, and clinical trials. The editorial model is partially centralized, with Associate Editors making preliminary judgments about suitability of submitted manuscripts for peer review, identifying reviewers, and often handling revisions, but with their opinion being advisory to the editor. The journal is published bimonthly, although there will be supplements devoted to meeting proceedings or special topics. The journal accepts only electronic submissions (to clinicaltrials@jhmi.edu), and will be implementing the Manuscript Central online manuscript management and peer-review system in 2004. It aims to make all editorial decisions on articles not sent out for review within four weeks, and for those sent out for review within 10 weeks. Articles of particular timeliness can be fast-tracked for rapid review and publication. Articles can be made available in electronic form before the print version appears. The journal receives about 200 to 300 original manuscripts annually, of which it accepts roughly 20 to 25%.

The journal is published in both print and electronic form, with the electronic version on Ingenta (www.ingenta.com). Society members receive the journal as part of their membership fee; others can receive it either by joining the society, obtaining a nonmember individual subscription, or accessing it through an institutional subscription. Individual articles can also be downloaded by nonsubscribers for a fee.

The field of clinical trials is extremely challenging and exciting, as a field in which the unit of information is human lives or suffering, and how that information is both produced and analyzed has profound importance to medicine, patients, and society at large. *Clinical Trials* aims to be a forum where the science of statistics is advanced, as applied to this form of human research, and where the full implications and applications of these methodologies can be explored and understood by the broader scientific community with interests in this vital area.

STEVEN N. GOODMAN

Cluster Analysis of Subjects, Hierarchical Methods

Cluster analysis is concerned with investigating a set of data to discover whether or not it consists of relatively distinct groups of observations. The groups are unknown a priori so that cluster analysis is distinguished from the activity of allocating individuals to one of a set of existing groups, an activity usually referred to as *assignment* or *discrimination* (*see* **Discriminant Analysis, Linear**).

Uncovering the group structure (if any) of a set of data is clearly of considerable importance in understanding the data and using them to answer substantive subject-matter questions. In medicine, for example, separating diseases that require different treatments will often be the primary goal of any classification exercise.

A very broad division of cluster analysis techniques is into those that produce *hierarchical classifications* and those that produce *nonhierarchical classification* (*see* **Cluster Analysis of Subjects, Nonhierarchical Methods**). This article is concerned with the former and owes much to the two excellent review papers by Gordon [17, 18].

Data

The raw data to be analyzed by cluster analysis methods usually consist of a set of p variables for each of the n individuals to be clustered. Such data are commonly represented by an $n \times p$ matrix, $\mathbf{X}$, given by

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}. \qquad (1)$$

Many clustering methods first require the raw data matrix to be transformed into an $n \times n$ matrix of pairwise similarities, dissimilarities, or distances, $\mathbf{D}$. An example of a measure frequently used in practice is the Euclidean distance, where the elements of $\mathbf{D}$, d_{ij}, are defined as follows:

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2}. \qquad (2)$$

Many other dissimilarity and distance measures can be used, however, and details are given in another entry in this Encyclopedia (*see* **Similarity, Dissimilarity, and Distance Measure**). An important problem, not considered here, is if and how the raw data should be standardized before they are converted to dissimilarity or distance measures. Details of the issues involved are discussed in Everitt [13].

(Occasionally, particularly in psychology, the matrix $\mathbf{D}$ may arise directly from, for example, asking subjects to judge the pairwise dissimilarities of stimuli of interest.)

Hierarchical Classifications and Dendrograms

In a *hierarchical* classification the data are not partitioned into a particular number of groups or clusters at a single step. Instead, the result consists of a series of partitions of the data, P_n, $P_{n-1}, \dots, P_1$. The first, P_n, consists of n single-member "clusters", the last, P_1, consists of a single group containing all n individuals. Often an investigator will not be interested in the complete hierarchy but only in a single partition of the data into a particular number of groups, say g. Deciding on an appropriate value of g from a hierarchy is a problem which will be discussed briefly later in the article.

Hierarchic classifications may be represented by a diagram known as a **dendrogram**, which can, somewhat unhelpfully, formally be defined as a *rooted terminally-labeled weighted tree* in which all terminal nodes are equally distant from the root [29]. For the purpose of the article, however, dendrograms can be characterized less formally in terms of their topology or shape, a set of labels identifying the n individuals being classified and a set of "weights" or "heights" associated with the $n - 1$ internal nodes (i.e. clusters) of the dendrogram. An example of such a diagram is given in Figure 1; others will be given later in the article. The structure of Figure 1 resembles an *evolutionary tree* (see Figure 2) and it is in biological applications that hierarchical classifications are

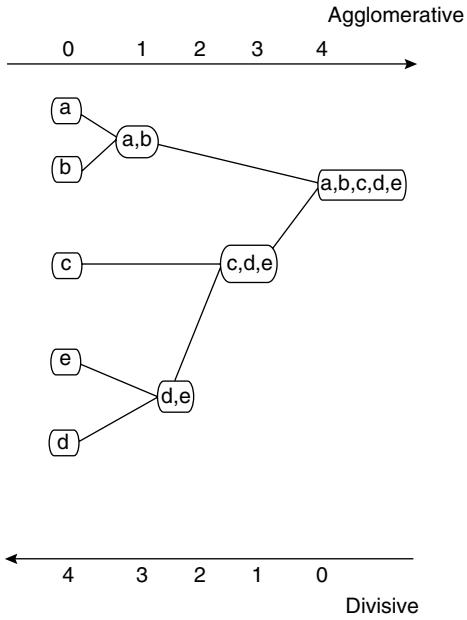

Figure 1 Example of a dendrogram

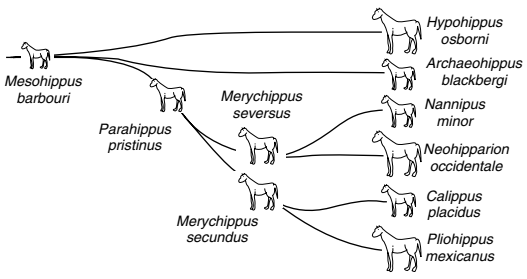

Figure 2 An evolutionary tree

perhaps most relevant. Rohlf [39], for example, suggests that a biologist, "all things being equal", aims for a system of nested clusters.

Hierarchical Clustering Methods

Many different ways have been proposed of transforming a dissimilarity or distance matrix into a dendrogram. Several of the methods, for example **single linkage** (see below), may be implemented by a variety of **algorithms**, and it is important to distinguish between algorithms and method (see [24]). Because different clustering strategies can produce different classifications of the same data set, careful appraisal of results is generally necessary, a point to which we shall return later in the article.

The two major types of algorithms that have been used to produce hierarchical classifications are *agglomerative* and *divisive*, with the former being far more commonly used in practice.

Agglomerative Algorithms

The basic operation of all such methods is similar, and is outlined in Figure 3. At each particular stage the methods fuse individuals or groups of individuals which are closest (or most similar). Differences between methods arise because of the variety of ways in which distance (or similarity) can be defined between two groups. In *single linkage clustering*, for example, intergroup distance is defined as the minimum of the interindividual values amongst pairs consisting of one individual from one group and one from the other. In *complete linkage*, the maximum value of the appropriate interindividual values is used. A method that uses more than a single interindividual distance value to define distance between groups is *group average clustering*, which uses the average of the appropriate distance values as illustrated in Figure 4. The dendrograms resulting from applying each of single linkage, complete linkage, and group average to the following small distance matrix are shown in Figure 5:

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}. \quad (3)$$

START: Clusters $C_1, C_2, \ldots, C_n$ each containing a single individual.

1. Find nearest pair of distinct clusters, say C_i and C_j, merge C_i and C_j,

delete C_j and decrement number of clusters by one.

If number of clusters equal one then stop, else return to 1.

Figure 3 Basic operation of hierarchical clustering procedures

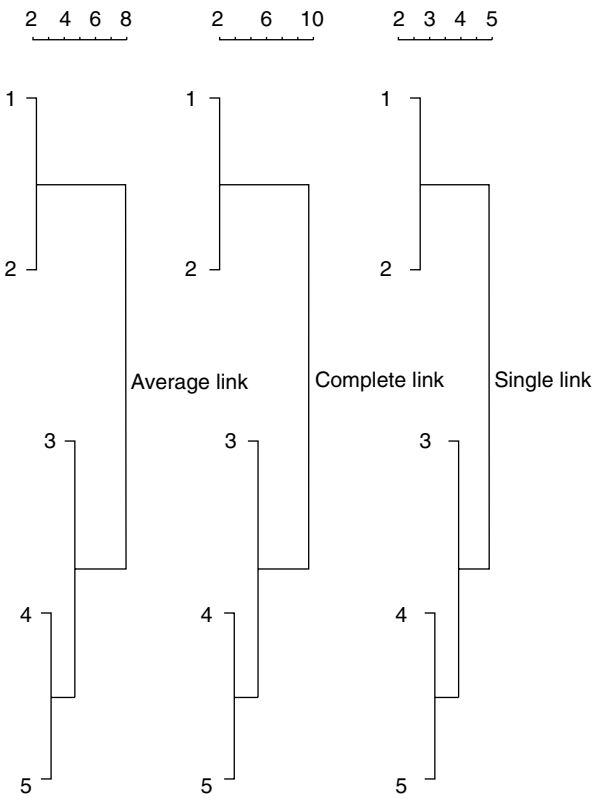

Figure 5 Single linkage, complete linkage, and group average dendrograms

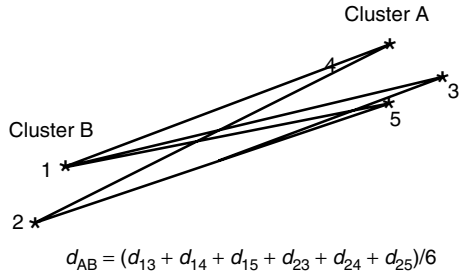

Figure 4 Group average distance

(Details of the steps in the progression from **D** to the appropriate dendrogram are given in [13].) The level at which two individuals join the same group, h_{ij}, depends on the cluster method used, but in general is different from their dissimilarity, d_{ij}. In fact, for single linkage clustering, $h_{ij} \leq d_{ij}$. The heights, h_{ij}, are always symmetric and satisfy the following ultrametric inequality:

$$h_{ij} \leq \max[h_{ik}, h_{jk}]. \tag{4}$$

The dendrograms in Figure 5 are constructed from the bottom up, but the natural way to use them is from the top down so that, for example, in the single linkage dendrogram the data are divided into two groups by "cutting" above the height value 4.

Many commonly used agglomerative algorithms are described by a general formula given originally by Lance & Williams [27, 28] and later extended by Jambu [23], in which the dissimilarity between a group (ij) formed by joining groups i and j, and some other group k is found from

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}| + \delta_i h_i + \delta_j h_j + \varepsilon h_k. \tag{5}$$

In this equation, h_i is the height in the dendrogram of group i and $\alpha_i, \alpha_j, \beta, \gamma, \delta_i, \delta_j$, and ε are parameters with particular values for particular clustering methods, as detailed in Table 1.

Methods C5 and C6 implicitly assume that the individuals are represented by points in some Euclidean space, and that the measure of pairwise dissimilarity is proportional to the squared Euclidean distance. The quantity d_{ij} has the following definitions: in the sum of squares method (C5) it is the within-group sum of squared distances of the group ij; in Ward's method (C6) it is the *increase* in the sum of squares that would be brought about by the amalgamation of groups i and j; in the centroid method (C7) it is the squared distance between the centroids of i and j; in the median method (C8) it is the squared distance between weighted centroids obtained by assigning each class the same number of individuals in evaluating the "centroid" of their union.

Direct use of the Lance–Williams–Jambu general agglomerative algorithm has $O(n^3)$ time complexity and $O(n^2)$ space complexity [47], and several authors have considered procedures for improving the efficiency of the approach. Day & Edelsbrunner [9] for example, show that by associating with each group a priority queue that orders the other groups by their distance to it, the time complexity can be reduced to $O(n^2 \log n)$. Further suggestions for improving efficiency are described in Bruynooghe [2] and Murtagh [38].

Divisive Algorithms

Divisive algorithms begin with a "group" containing all n individuals and at each succeeding stage divide an existing group into two. Algorithms that find the globally optimal divisions [12, 41] are computationally very demanding and practicable alternatives have been suggested by Macnaughton-Smith et al. [31], Vichi [44], and Hubert [22]. The only divisive algorithm that has been routinely used, however, is one applicable when the variables describing each individual are binary, and division of the data is now on the basis of the possession or otherwise of a single specified attribute, this being chosen to maximize the difference between the resulting two groups according to some particular criterion. Such procedures are known as *monothetic divisive algorithms* [7, 48]. The division criteria used are generally based on some **chi-square** type statistic – see Everitt [13] for details.

Table 1 Clustering strategies obtainable from the general agglomerative algorithm

Name	References	α_i	β	γ	δ_i	ε
C1	Single link (Florek et al. [14]; Sneath [42])	0.5	0	−0.5	0	0
C2	Complete link (McQuitty [32])	0.5	0	0.5	0	0
C3	Group average link (Sokal & Michener [43]; McQuitty [34])	$\frac{n_i}{n_i + n_j}$	0	0	0	0
C4	Weighted average link (McQuitty [33, 34])	0.5	0	0	0	0
C5	Sum of squares (Jambu [23])	$\frac{n_i + n_k}{n_+}$	$\frac{n_i + n_j}{n_+}$	0	$\frac{-n_i}{n_+}$	$\frac{-n_k}{n_+}$
C6	Incremental sum of squares (Ward [45]; Wishart [50])	$\frac{n_i + n_k}{n_+}$	$\frac{-n_k}{n_+}$	0	0	0
C7	Centroid (Sokal & Michener [43]; Gower [19])	$\frac{n_i}{n_i + n_j}$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0	0	0
C8	Median (Lance & Williams [27]; Gower [19])	0.5	−0.25	0	0	0
C9	Flexible (Lance & Williams [27])	$0.5(1 - \beta)$	$\beta(< 1)$	0	0	0

Note: n_i is the number of individuals in class C_i; $n_+ = n_i + n_j + n_k$

Properties and Problems of Hierarchical Clustering Techniques

Several hierarchical clustering techniques, for example single linkage and the median method, have a tendency to cluster together, at a relatively low level, individuals linked by a series of isolated intermediates. This property, known generally as *chaining*, may cause the method to fail to resolve relatively distinct clusters when there are a small number of such individuals between them. Methods based on single linkage which attempt to avoid the chaining problem are described in Wong [51] and Wong & Lane [52].

A notable advantage of both single and complete linkage clustering emphasized by Johnson [25] is their invariance under monotonic transformations of the original dissimilarity matrix. Consequently, only the ordinal properties of the dissimilarities are of concern, and the difficulties generally involved in scaling and combining different variables into a measure of dissimilarity are lessened. (Complete linkage is not satisfactory when the dissimilarity matrix contains ties.)

Jardine & Sibson [24] object to many hierarchical clustering methods on mathematical grounds. Briefly, what these authors show is that a cluster method that transforms a dissimilarity matrix into a hierarchic dendrogram may be regarded as a procedure that imposes the ultrametric inequality which is satisfied by the heights of the dendrogram, i.e. $h_{ij} \leq \max[h_{ik}, h_{kj}]$, on a dissimilarity coefficient which originally may have satisfied only the weaker metric inequality. Jardine & Sibson then specify certain simple conditions, for example continuity, minimum distortion, etc. that any such transformation should satisfy and demonstrate that only single linkage satisfies the specified requirements. Consequently, Jardine & Sibson recommend single linkage as the method with greatest mathematical appeal. Such a recommendation has been criticized by many authors, for example Williams et al. [49] and Gower [20], and certainly in practice, single linkage has often been found to be the *least* successful method.

The centroid and median methods have the unsatisfactory property that they can produce *reversals* in the dendrogram, in the sense that a group i may be contained in a group j, i.e. $i \in j$, but $h_i > h_j$. Reversals can be prevented by adding the requirement that d_{ij} be no less than $\max[h_i, h_j]$. Necessary and sufficient conditions for an absence of reversals in the dendrogram produced by the algorithm defined in (4) are

$$\begin{aligned} \gamma &\geq -\min(\alpha_i, \alpha_j), \\ \alpha_i + \alpha_j &\geq 0, \\ \alpha_i + \alpha_j + \beta &\geq 1. \end{aligned} \tag{6}$$

A number of empirical investigations of hierarchical clustering techniques have been performed to investigate the extent to which standard clustering algorithms recover known types of structure in data. Cunningham & Ogilvie [8], for example, compare seven hierarchical techniques and find that group average clustering performs most satisfactorily overall for the data sets considered; in addition, however, they find a large interaction in the results between types of input data and the particular clustering method used. Kuiper & Fisher [26] investigate six hierarchical techniques and find that for equal-sized groups, Ward's method classifies almost as well as Fisher's linear discriminant function (*see* **Discriminant Analysis, Linear**) when the groups are specified a priori; with unequal sized groups, however, centroid, group average, and complete linkage clustering are more successful. A study by Hands & Everitt [21] using **binary data** found similar results; Ward's method performed very well when the data contained approximately equally sized clusters, but poorly when the clusters were of different sizes. A review of **Monte Carlo** studies in this area is given in Milligan [35]. The only overall conclusion that can be reached is that no particular method can be claimed to be superior for all types of data.

Cheng & Milligan [4] investigate the influence of individual coordinate locations in a multivariate space and demonstrate that differences between clustering methods help to explain some of the results in previous validation research of these methods.

Validation

The clustering algorithms described in previous sections always produce a hierarchical classification and it is important to validate the results, in particular to investigate whether a hierarchical structure is appropriate and whether the derived

solution misrepresents the pattern in the data in any way. (Here we are referring to the entire dendrogram rather than to a solution corresponding to a particular number of groups.) Some authors have argued that the only relevant criteria for assessing a classification are its utility and interpretability. But as pointed out by Gordon [17], there are dangers in this approach; "human ingenuity is quite capable of providing a *post hoc* justification of dubious classifications".

According to Gordon [17], "there is no uniquely obvious way of specifying the absence (or presence) of class structure in data", which probably accounts for the lack of suitable, practical tests for the absence of structure hypothesis. Those tests that have been suggested by Fillenbaum & Rapoport [15], Ling [30], and Baker & Hubert [1], among others, have low power against some alternative hypotheses, and can be markedly influenced by a small number of outliers [16].

The suitability of a hierarchical classification for a data set can be assessed by comparing the original dissimilarities with the heights in the derived dendrogram. A data set can only be accurately represented by a hierarchical classification if little distortion is imposed in transforming from the dissimilarity matrix, d_{ij}, to the ultrametric matrix, h_{ij}. One common procedure for assessing the match between the dendrogram and the dissimilarity matrix is the *cophenetic correlation coefficient*. This is simply the product–moment **correlation** between the $n(n-1)/2$ entries in the lower half of the observed dissimilarity matrix and the corresponding terms in the so-called *cophenetic matrix*, **C**, containing the heights in the derived dendrogram at which individuals i and j first occur in the same cluster. Since the latter satisfy the ultrametric inequality, the match between dendrogram and data cannot be perfect unless the entries in the dissimilarity matrix are also ultrametric, a situation which seldom occurs in practice. For the example involving the application of three agglomerative algorithms given previously, the corresponding values of the cophenetic correlation coefficient are as follows:

$$\text{single linkage} = 0.82,$$

$$\text{complete linkage} = 0.85,$$

$$\text{average linkage} = 0.85.$$

Rohlf & Fisher [40] studied the distribution of the cophenetic correlation coefficient under the hypothesis that the individuals are randomly chosen from a *single* **multivariate normal distribution** (i.e. have no cluster structure). They found that the average value of the coefficient tends to decrease with n and to be almost independent of the number of variables. They also suggest that values of the cophenetic correlation above 0.8 indicate a nonartifactual hierarchical structure, although in a later paper, Rohlf [39] indicates that even values close to 0.9 are not necessarily a guarantee that the dendrogram serves as a sufficiently accurate summary of the relationships between the individuals.

Several other measure of the distortion produced by transforming a dissimilarity matrix into a dendrogram are described in Gordon [17].

The amount of support for hierarchical structure in a data set has also been assessed by *stability studies*, in which the original classification is compared with a classification of a modified version of the data obtained, for example by adding an error term to the dissimilarity matrix, or by deleting a small number of individuals. Large differences between the resulting two classifications would shed some doubt on the suitability of clustering for the data. A related approach is to divide the individuals randomly into two and compare the results of separate cluster analyses applied to each half.

Partitions from a Hierarchy – the Number of Groups Problem

As remarked earlier in the article, when the hierarchical clustering technique are used in practice the investigator is not usually interested in the complete hierarchy but only in one or two partitions found by "cutting" the dendrogram at a hopefully appropriate point. Deciding on the "best" fitting partition for a particular data set, i.e. choosing the correct number of clusters, is not straightforward, although a considerable number of methods *have* been suggested. One very informal procedure which is often used is simply to examine the dendrogram looking for "large" changes in level. The dendrogram shown in Figure 6, for example, suggests a two-group solution.

More formal approaches to the number of groups problem have been suggested by Duda & Hart [10], Calinski & Harabasz [3], and Mojena [37]. All such methods are considered in a detailed investigation reported by Milligan & Copper [36]; none appears completely satisfactory.

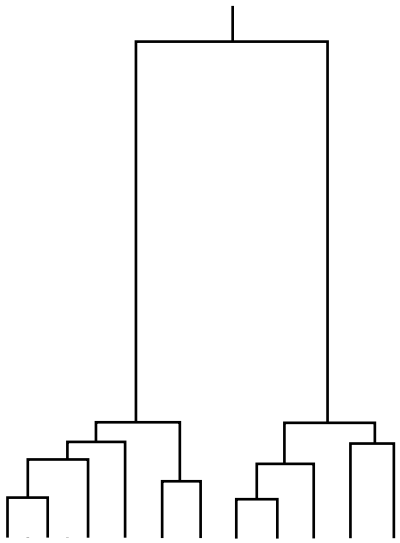

Figure 6 Dendrogram indicating two groups

Some Examples

Corbet et al. [5] describe a study to compare British populations of water voles with others from Europe. The original data consisted of recordings of the presence or absence of 13 characteristics on about 300 water vole skulls divided into samples from 14 populations; these data were converted into the matrix of *population* dissimilarities shown in Table 2 by a procedure involving the percentage incidence of each characteristic in each population (see the original paper for full details). The British populations are those numbered 1–6, and arise from the following areas:

1. Surrey
2. Shropshire
3. Yorkshire
4. Perthshire
5. Aberdeen
6. Eileen Gambia

The non-British populations are from two species, *Arvicola terrestris* (7–11) and *Arvicola sapidus* (12–14). The corresponding areas are:

7. Alps
8. Yugoslavia
9. Germany
10. Norway
11. Pyrenees I
12. Pyrenees II
13. North Spain
14. South Spain.

The dendrograms from applying single linkage, complete linkage, and average linkage to the dissimilarities are shown in Figures 7, 8, and 9. The two-group solutions given by single linkage and average linkage are identical and correspond to the following division of the 14 populations:
Group 1: 12, 13, 14
Group 2: 1–11

The two-group solution for complete linkage, however, is
Group 1: 1, 2, 3, 4, 5, 6, 9, 10
Group 2: 7, 8, 11, 12, 13, 14

Even on a small data set such as this, different hierarchical clustering methods may give different solutions.

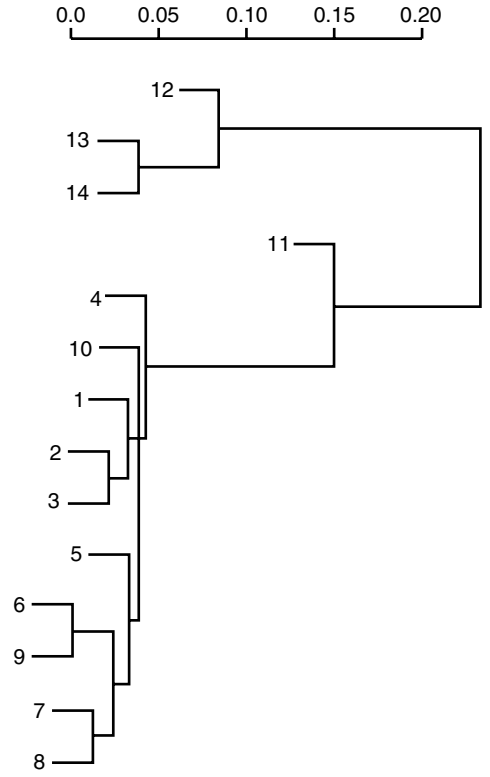

Figure 7 Single linkage dendrogram for water vole data

Table 2 Dissimilarities for water vole populations

	1	2	3	4	5	6	7	8	9	10	11	12	13
2	0.099												
3	0.033	0.022											
4	0.183	0.114	0.042										
5	0.148	0.224	0.059	0.068									
6	0.198	0.039	0.053	0.085	0.051								
7	0.462	0.266	0.322	0.435	0.268	0.025							
8	0.628	0.442	0.444	0.406	0.240	0.129	0.014						
9	0.113	0.070	0.046	0.047	0.634	0.002	0.106	0.129					
10	0.173	0.119	0.162	0.331	0.177	0.039	0.089	0.237	0.071				
11	0.434	0.419	0.339	0.505	0.469	0.390	0.315	0.349	0.151	0.430			
12	0.762	0.633	0.781	0.700	0.758	0.625	0.469	0.618	0.440	0.538	0.607		
13	0.530	0.389	0.482	0.570	0.597	0.498	0.374	0.562	0.247	0.383	0.387	0.084	
14	0.586	0.435	0.550	0.530	0.552	0.509	9.369	0.471	0.234	0.346	0.456	0.090	0.038

Wastell & Gray [46] describe the use of hierarchical clustering for the development of a classification of **pain** distribution in patients with temporomandibular joint pain dysfunction syndrome (TMJPDS). This refers to a complex symptom group involving facial pain, limitation and deviation of mandibular movements, joint noises, and muscle tenderness. Symptoms vary with the stage of the disease; etiology is equally complex and both physical and psychogenic factors have been implicated. Pain is the most commonly recorded symptom, but its facial distribution does not conform to a single pattern.

Wastell & Gray's main aim was to use clustering techniques to develop an objective typology for classifying facial pain in terms of its spatial distribution. Clinically the hope was that the derived classification would be useful in identifying different stages of the disease, which would be of help in defining more directed treatment plans.

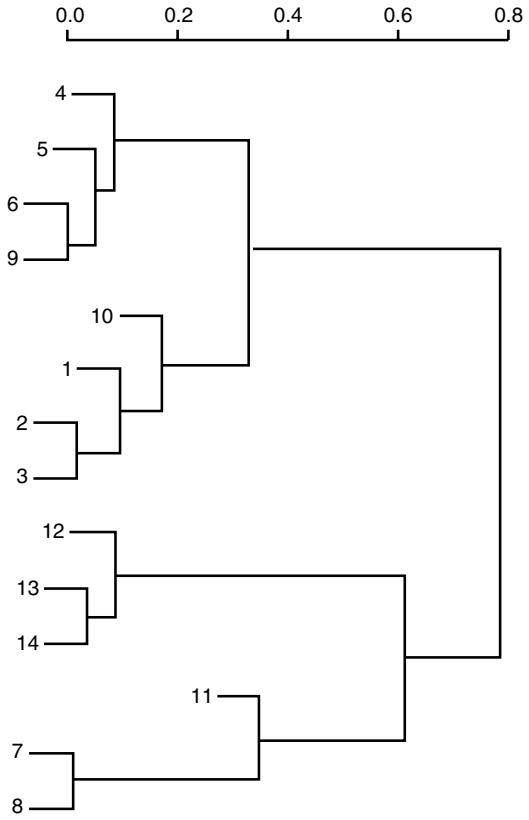

Figure 8 Complete linkage dendrogram for water vole data

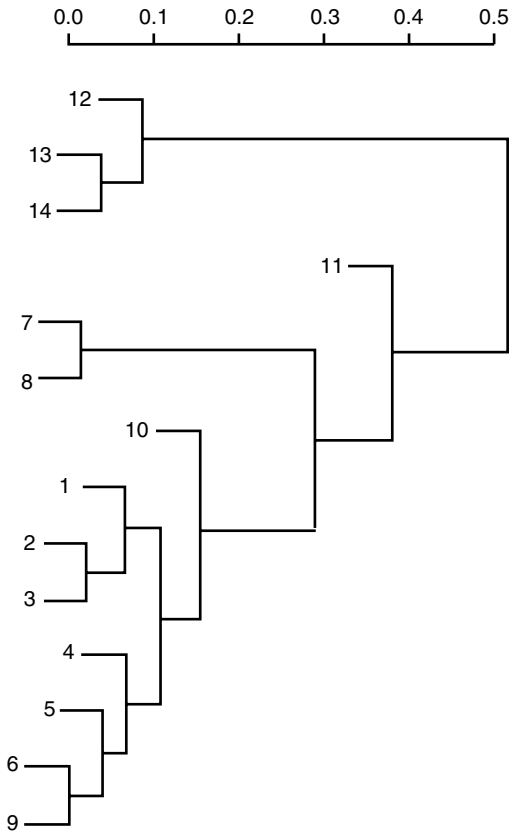

Figure 9 Average linkage dendrogram for water vole data

Data were collected from 127 patients attending the temporomandibular joint clinic of a university hospital complaining of classic TMJPDS. Patients were asked to trace the boundary of their pain-affected area with the tip of their index finger. The examiner recorded this outline on a diagram of the lateral profile of the face (see Figure 10) and this was adjusted until patients were satisfied that the outline matched their own pain area.

The squares of the grid shown in Figure 10 falling within the perimeter of a pain area were scored 1; those without, 0. In this way any patient's pain distribution may be described by a string of binary variables. In practice all distributions lay within a central rectangle with horizontal extension, J–T, and vertical extension, 11–28, giving $11 \times 18 = 198$ binary variables for analysis. The similarity between each pair of patients was calculated using Jaccard's coefficient, which is simply the proportion

of one-to-one matches for the two patients after ignoring the zero-to-zero matches. (See Everitt [13] for an explicit definition of this coefficient.)

Ward's method of clustering was used in this application and the resulting dendrogram is shown in Figure 11. The structure of the dendrogram appears to indicate three major classes, with a further possible subdivision of each of these into two. A composite pain distribution matrix for each class was constructed by simple matrix addition of the pain matrices of its constituent members. The authors' description of the pain classes was as follows:

Pain Class A:. The pain distribution of Class A was concentrated over the temporomandibular joint

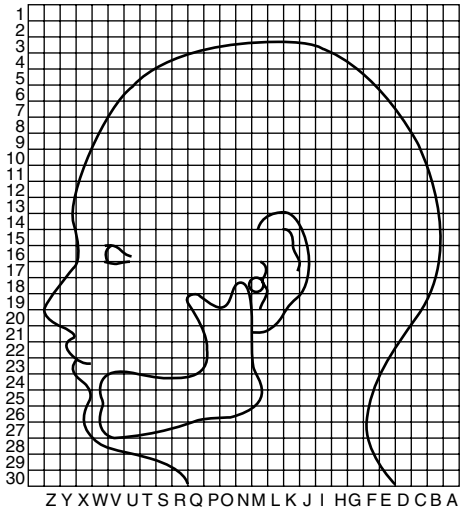

Figure 10 TMJPDS data collection method

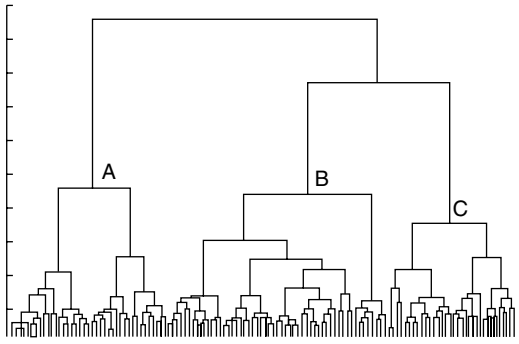

Figure 11 Ward's method dendrogram for TMJPDS data

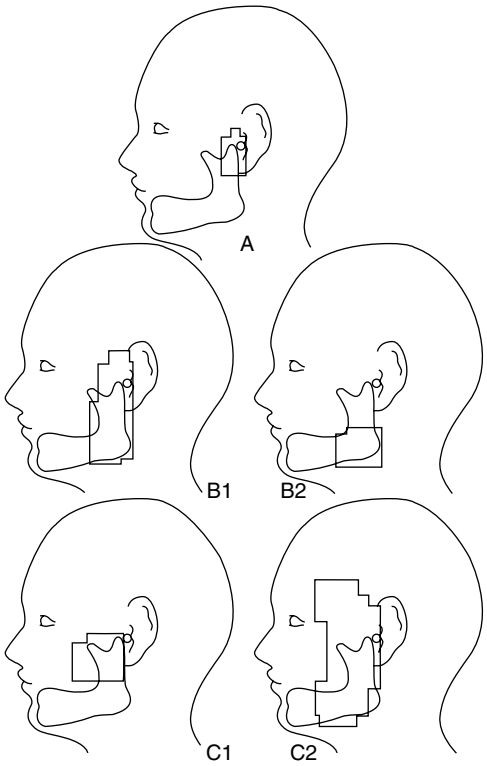

Figure 12 Pain distribution of cluster solution on TMJPDS data

(see Figure 12). The two subclasses were much alike, apart for a small vertical difference in their centroids.

Pain Class B:. The pain distribution of Class A differed from that of Class B in involving the vertical portion of the mandible (the ramus). The two subclasses of B were quite different: Class B2 showed a distribution to the lower point of the ramus and Class B1 showed a much wider distribution, covering all the ramus and the interior part of the temple (see Figure 12).

Pain Class C:. The pain distribution of Class C differed from the other two classes in involving an anterior projection over the zygomatic arch. The two subclasses were again distinct: Class C1 showed a distribution confined to the temporomandibular joint and the zygomatic arch; Class C2 showed a much wider distribution spreading over the temple as well as covering the ramus (see Figure 12).

The final conclusion, after careful validation, was that the groups could be interpreted in terms of a chronological model of the development of TMJPDS. Other interesting applications of hierarchical classification methods are described in Murtagh [38], Duflou et al. [11], and Coste et al. [6].

Summary

The concept of the hierarchical representation of a data set applies most satisfactorily in biology and related disciplines. Although any cluster analysis exercise with biological data need not necessarily replicate the structure implicit in the traditional structure of Linnaean Taxonomy, i.e. species, genera, etc., there nevertheless remains a strong inclination amongst biologists for hierarchical classifications. When used in other areas, however, the justification for a hierarchical rather than a nonhierarchical structure may be less clear, and there is then the danger of imposing rather than discovering structure. Careful validation of solutions is a clear requirement in any clustering exercise. No particular hierarchical clustering technique can be recommended as likely to be "best" in most situations, although Everitt [13] gives some suggestions as to which methods might be more generally useful, on the basis of the results of a number of Monte Carlo studies.

References

[1] Baker, F.B. & Hubert, L.J. (1976). A graph theoretic approach to goodness-of-fit in complete link hierarchical clustering, *Journal of the American Statistical Association* **71**, 870–878.

[2] Bruynooghe, M. (1978). Classification ascendante hiérarchique des grand ensembles des données: un algorithm rapide fondé sur la construction des voisinages réductibles, *Les Cahiers de L'Analyse des Données* **3**, 7–33.

[3] Calinski, T. & Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics* **3**, 1–27.

[4] Cheng, R. & Milligan, G.W. (1995). Mapping influence regions in hierarchical clustering, *Multivariate Behavioral Research* **30**, 547–576.

[5] Corbet, G.B., Cummins, J., Hedges, S.R. & Krzanowski, W.J. (1970). The taxonomic status of British water voles, genus *Arvicola*, *Journal of Zoology* **161**, 301–316.

[6] Coste, J., Spira, A., Ducimetiere, P. & Paolaggi, B. (1991). Clinical and psychological diversity of nonspecific low back pain. A new approach towards the classification of clinical subgroups, *Journal of Clinical Epidemiology* **44**, 1233–1245.

[7] Crawford, R.M.M. & Wishart, D. (1967). A rapid multivariate method for the detection and classification of groups of ecologically related species, *Journal of Ecology* **55**, 505–524.

[8] Cunningham, K.M. & Ogilvie, J.C. (1972). Evaluation of hierarchical grouping techniques: a preliminary study, *Computer Journal* **15**, 209–213.

[9] Day, W.H.E. & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods, *Journal of Classification* **1**, 7–24.

[10] Duda, R.O. & Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

[11] Duflou, H., Maenhaut, W. & De Reuck, J. (1990). Application of principal component and cluster analysis to the study of the distribution of minor and trace elements in normal human brain, *Chemometrics and Intelligent Laboratory Systems* **9**, 273–286.

[12] Edwards, A.W.F. & Cavalli-Sforza, L.L. (1965). A method for cluster analysis, *Biometrics* **21**, 363–375.

[13] Everitt, B.S. (1993). *Cluster Analysis*. Arnold, London.

[14] Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H. & Zubrzycki, S. (1951). Sur la liason et la division des points d'un ensemble fini, *Colloquium Mathematicum* **2**, 282–285.

[15] Fillenbaum, S. & Rapoport, A. (1971). *Structures in the Subjective Lexicon*. Academic Press, New York.

[16] Gordon, A.D. (1980). *Classification*. Chapman & Hall, London.

[17] Gordon, A.D. (1987). A review of hierarchical classification, *Journal of the Royal Statistical Society, Series A* **150**, 119–137.

[18] Gordon, A.D. (1996). Hierarchical classification, in *Clustering and Classification*, P. Arabie, L.J. Hubert & G. De Soete, Eds. World Scientific Publications, River Edge.

[19] Gower, J.C. (1967). A comparison of some methods of cluster analysis, *Biometrics* **23**, 623–628.

[20] Gower, J.C. (1975). Goodness-of-fit criteria for classification and other patterned structures, in *Proceedings of the Eighth International Conference on Numerical Taxonomy*, pp. 38–62.

[21] Hands, S. & Everitt, B.S. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques, *Multivariate Behavioural Research* **22**, 235–243.

[22] Hubert, L. (1973). Monotone invariant clustering procedures, *Psychometrika* **38**, 47–62.

[23] Jambu, M. (1978). *Classification Automatique pour L'Analyse des Données*, Tome 1. Dunod, Paris.

[24] Jardine, N. & Sibson, R. (1971),. *Mathematical Taxonomy*. Wiley, London.

[25] Johnson, S.C. (1967). Hierarchical clustering schemes, *Psychometrika* **32**, 241–254.

[26] Kuiper, F.K. & Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures, *Biometrics* **31**, 777–783.

[27] Lance, G.N. & Williams, W.T. (1966). A generalized sorting strategy for computer classifications, *Nature* **212**, 218.

[28] Lance, G.N. & Williams, W.T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems, *Computer Journal* **9**, 373–380.

[29] Lapointe, F.J. & Legendre, P. (1991). The generation of random ultrametric matrices representing dendrograms, *Journal of Classification* **8**, 177–200.

[30] Ling, R.F. (1973). Probability theory of cluster analysis, *Journal of the American Statistical Association* **68**, 159–164.

[31] Macnaughton-Smith, P., Williams, W.T., Dale, M.B. & Mockett, L.G. (1964). Dissimilarity analysis: a new technique of hierarchical sub-division, *Nature* **202**, 1034–1035.

[32] McQuitty, L.L. (1960). Hierarchical linkage analysis for the isolation of types, *Educational and Psychological Measurement* **20**, 55–67.

[33] McQuitty, L.L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data, *Educational and Psychological Measurement* **25**, 825–831.

[34] McQuitty, L.L. (1967). Expansion of similarity analysis by reciprocal pairs for discrete and continuous data, *Educational and Psychological Measurement* **27**, 253–255.

[35] Milligan, G.W. (1981). A review of Monte Carlo tests of cluster analysis, *Multivariate Behavioural Research* **16**, 379–407.

[36] Milligan, G.W. & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika* **50**, 159–179.

[37] Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation, *Computer Journal* **20**, 359–363.

[38] Murtagh, F. (1985). *Multidimensional Clustering Algorithms*. COMPSTAT Lectures 4. Physica-Verlag, Vienna.

[39] Rohlf, F.J. (1970). Adaptive hierarchical clustering schemes, *Systematic Zoology* **19**, 58–82.

[40] Rohlf, F.J. & Fisher, D.R. (1968). Test for hierarchical structure in random data sets, *Systematic Zoology* **17**, 407–412.

[41] Scott, A.J. & Symon, M.J. (1971). On the Edwards-Cavalli-Sforza method of cluster analysis, *Biometrics* **27**, 217–219.

[42] Sneath, P.H.A. (1957). The application of computers to taxonomy, *Journal of General Microbiology* **17**, 201–226.

[43] Sokal, R.R. & Michener, C.D. (1958). A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin* **38**, 1409–1438.

[44] Vichi, M. (1985). On a flexible and computationally feasible divisive clustering technique, *Rivista di Statistica Applicata* **18**, 199–208.

[45] Ward, J.H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* **58**, 236–244.

[46] Wastell, D.G. & Gray, R. (1987). The numerical approach to classification: a medical application to develop a typology of facial pain, *Statistics in Medicine* **6**, 137–164.

[47] Weide, B. (1977). A survey of analysis techniques for discrete algorithms, *ACM Computer Survey* **9**, 291–313.

[48] Williams, W.T. & Lambert, J.M. (1959). Multivariate methods in plant ecology I. Association analysis in plant communities, *Journal of Ecology* **47**, 83–101.

[49] Williams, W.T., Lance, G.N., Dale, M.B. & Clifford, H.T. (1971). Controversy concerning the criteria for taxonometric strategies, *Computer Journal* **14**, 162–165.

[50] Wishart, D. (1969). An algorithm for hierarchical classifications, *Biometrics* **25**, 165–170.

[51] Wong, M.A. (1982). A hybrid clustering method for identifying high-density clusters, *Journal of the American Statistical Association* **77**, 841–847.

[52] Wong, M.A. & Lane, T. (1983). A kth nearest neighbour clustering procedure, *Journal of the Royal Statistical Society, Series B* **45**, 362–368.

(*See also* **Classification, Overview**; **Cluster Analysis, Variables**; **Multidimensional Scaling**; **Pattern Recognition**; **Projection Pursuit**; **R- and Q-analysis**)

BRIAN S. EVERITT

Cluster Analysis of Subjects, Nonhierarchical Methods

Cluster analysis is concerned with investigating a set of data to discover whether or not it consists of relatively distinct groups of observations. The groups are unknown a priori so that cluster analysis is distinguished from the activity of allocating individuals to one of a set of existing groups, an activity usually referred to as *assignment* or *discrimination* (*see* **Discriminant Analysis, Linear**).

Uncovering the group structure (if any) of a set of data is clearly of considerable importance in understanding the data and in using them to answer substantive subject matter questions. In medicine, for example, separating diseases that require different treatments will often be the primary goal of any classification exercise.

A very broad division of cluster analysis techniques is into those that produce *hierarchical classifications* (*see* **Cluster Analysis of Subjects, Hierarchical Methods**) and those where the groupings are not necessarily constrained in this way, the so-called *nonhierarchical* methods. This article is concerned with the latter.

Data

The raw data to be analyzed by cluster analysis methods usually consists of a set of p variable values for each of the n individuals to be clustered. Such data are commonly represented by an $n \times p$ matrix, $\mathbf{X}$, given by

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}. \qquad (1)$$

When $p = 2$ simply plotting the data may provide, via the excellence of the eye–brain pattern recognition system, a simple method of "clustering". The three "clusters" in Figure 1, for example, are immediately apparent without the application of any formal method, or indeed, without making the meaning of the term "cluster" explicit.

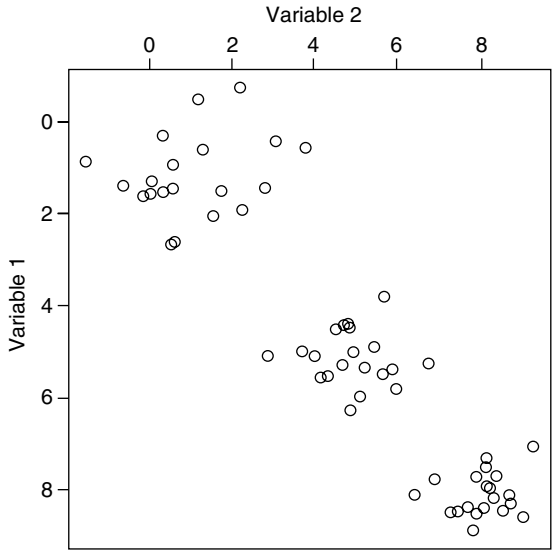

Figure 1 Data containing three distinct clusters

When $p > 3$ such a direct approach is not possible, but plots of the data in the space of the first two or three principal components may then be helpful (*see* **Principal Components Analysis**; **Multivariate Graphics**).

Optimization Methods

The cluster analysis methods described in this entry produce a partition of the individuals into a *particular* number of groups, by optimizing some numerical criterion. The basic idea behind these techniques is that associated with each partition of the n individuals into the required number of groups, g, is an index, $f(n, g)$, the value of which reflects how successful the partition is in describing the data. For some indices, partitions corresponding to high values are good and so the "best" partition is found by maximizing the index. For other indices, low values are sought and the search is for the partition with minimum value. Associating a numerical value with each partition allows competing partitions to be compared. Differences between the methods in this class arise both because of the different clustering indices that can be used, and the alternate **algorithms** which can be employed to optimize the selected index. In the next section a number of possible clustering indices

are described and in the following section the optimization problem is discussed. It will be assumed that the variables describing each individual are continuous, and that the spatial distribution of the individuals, represented as points in a p-dimensional space, can be meaningfully summarized by the location of the center of gravity of each cluster and by the sample scatter matrix of each cluster. (Clustering criteria suitable for binary and ordinal variables (*see* **Ordered Categorical Data**) are described in [23].)

A problem common to all the procedures to be described is that of selecting the most appropriate number of groups for a data set; some discussion and suggestions will be given later.

Clustering Indices

Many numerical indices for clustering have been suggested, but those most commonly used arise from consideration of three matrices, **T**, **B**, and **W**, which can be calculated from a partition of the data into g groups as follows:

$$\mathbf{T} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})',$$

$$\mathbf{W} = \frac{1}{n_g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{ij})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{j})', \qquad (1)$$

$$\mathbf{B} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})',$$

where $\mathbf{x}_{ij}$ represents the jth, p-dimensional observation in group i, $\bar{\mathbf{x}}_i$ is the mean vector of group i, $\bar{\mathbf{x}}$ is the mean vector of all the observations, and n_i is the number of observations in group i.

The $p \times p$ matrices defined above represent, respectively, total dispersion **T**, within-group dispersion **W**, and between-group dispersion **B**; they are related as follows:

$$\mathbf{T} = \mathbf{W} + \mathbf{B}. \qquad (2)$$

For $p = 1$ this equation represents a relationship between scalars, namely the division of the total sum of squares for a variable into the within- and between-group sum of squares familiar from one-way **analysis of variance**. Here a natural candidate for a clustering index is the within-group sum of squares, with a partition having minimum value being sought. Fisher [7] describes a procedure to find a partition into g groups for which the within-group sum of squares is minimized. For $p > 1$ the derivation of clustering criteria from (2) is not quite so clear-cut, and several possibilities have been suggested.

1. trace(**W**).

An obvious extension of the within-group sum of squares criterion when $p > 1$ is the sum of the separate within-group sum of squares for each variable, i.e. a partition minimizing trace(**W**) is looked for. [Minimizing trace(**W**) is equivalent to maximizing trace(**B**).] This criterion was suggested explicitly by Singleton & Kautz [22] and is also implicit in the clustering procedures described by Forgey [9], Jancey [13], MacQueen [16], and Ball & Hall [1].

Despite its popularity, the minimization of trace(**W**) approach to clustering suffers from a number of serious problems. One is that it is scale dependent so that different solutions may be obtained from the raw data than from the data standardized in some particular way. Since the question of the appropriate standardization in cluster analysis is a difficult one (see, for example, [8]), this lack of invariance can cause severe problems in practice.

A further problem with the use of this criterion is that it has a strong tendency to produce spherically shaped clusters, even when the natural clusters in the data are of other shapes – see [5] for an example. Consequently, its use may *impose* an artifactual structure on the data rather than uncover their true structure.

2. det(**W**).

In **multivariate analysis of variance** one of the tests for differences in population mean vectors is based on the ratio of the determinants of the within and total dispersion matrices (see [14]). Large values of det(**T**)/det(**W**) indicate a difference between the mean vectors. In a clustering context this implies that seeking a partition that maximizes the ratio of the two determinants would lead to clusters with widely separated mean vectors relative to their within-cluster dispersion. The maximization of det(**T**)/det(**W**) as a method of cluster analysis was first suggested by Friedman & Rubin [10].

Since, for all partitions of the n individuals into g groups, **T** remains the same, maximization of det(**T**)/det(**W**) is equivalent to minimizing det(**W**), the generalized variance of a set of multivariate data.

This particular criterion has been studied in some detail by Marriott [17, 18].

A major advantage of the det(**W**) criterion over the trace(**W**) criterion discussed earlier is that it is invariant under nonsingular linear transformations of the original data matrix. Consequently, problems with standardization do not arise. A further advantage of the det(**W**) criterion is that it does not restrict clusters to being "hyperfootballs", although it does assume that all the clusters have the same shape and orientation; Everitt [5] gives an example where the criterion can lead to the "wrong" clusters when this assumption does not hold.

3. trace($\mathbf{BW}^{-1}$).

A further criterion suggested by Friedman & Rubin [10] is the maximization of the matrix obtained from the product of the between-groups dispersion matrix and the inverse of the within-groups matrix. This function also appears in the context of multivariate analysis of variance (see [14]), and is equivalent to what Rao [20] calls the generalization of the **Mahalanobis distance** to more than two groups.

Both trace($\mathbf{BW}^{-1}$) and det(**T**)/det(**W**) may be expressed in terms of the **eigenvalues**, λ_i, of $\mathbf{BW}^{-1}$ as follows:

$$\text{trace}(\mathbf{BW}^{-1}) = \sum_{i=1}^{p} \lambda_i, \tag{3}$$

$$\frac{\det(\mathbf{T})}{\det(\mathbf{W})} = \prod_{i=1}^{p} (1 + \lambda_i). \tag{4}$$

For $g = 2$ partitions given by maximizing trace($\mathbf{BW}^{-1}$) and minimizing det(**W**) will be the same. In other cases the former procedure has a tendency to produce long thin clusters strung out along a single direction (see [10]).

Scott & Symons [21] and Banfield & Raftery [2] demonstrate how the clustering criteria described above arise from considering a likelihood approach to the clustering problem, an approach which also leads to a number of other possible indices for clustering. The probability model assumed is that the population of interest consists of g different subpopulations, and that the density of a p-dimensional observation $\mathbf{x}$ from the kth subpopulation is $f_k(\mathbf{x}; \boldsymbol{\theta})$ for some unknown vector of parameters $\boldsymbol{\theta}$. Given observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and defining identifying labels $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$, where $\gamma_i = k$ if $\mathbf{x}_i$ comes from the kth subpopulation, the required likelihood function can be written as

$$L(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i; \boldsymbol{\theta}). \tag{5}$$

In the so-called *classification maximum likelihood procedure*, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are chosen so as to maximize this likelihood.

When $f_k(\mathbf{x}; \boldsymbol{\theta})$ is a **multivariate normal** density with mean vector, $\boldsymbol{\mu}_k$, and covariance matrix, $\boldsymbol{\Sigma}_k$, then the likelihood in (5) becomes

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \text{const} \prod_{k=1}^{g} \prod_{i \in E_k} |\boldsymbol{\Sigma}_k|^{-1/2} \times \exp\left[-\tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right], \tag{6}$$

where $E_k = \{i; \gamma_i = k\}$ defines a particular partition of the data. Scott & Symons [21] and Banfield & Raftery [2] demonstrate the following:

1. If $\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}, k = 1, \ldots, g$ then the likelihood is maximized by choosing $\boldsymbol{\gamma}$ to minimize trace(**W**). Consequently, this particular clustering criterion is essentially only suitable when the clusters in the data are spherical in shape and of approximately equal sizes (equal number of observations).
2. If $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}, k = 1, \ldots, g$, then the likelihood is maximized by choosing $\boldsymbol{\gamma}$ to minimize det(**W**). Consequently, this clustering criterion is suitable only when the clusters have the same orientation and shape, although this is not constrained to being spherical.
3. When the $\boldsymbol{\Sigma}_k$ are not constrained in any way, the likelihood is maximized by choosing $\boldsymbol{\gamma}$ to minimize $\sum_{k=1}^{g} n_k \log[\det(\mathbf{W}_k / n_k)]$, where $\mathbf{W}_k$ is the sample cross-product matrix for the kth cluster. This particular criterion does not appear to have been used as a basis for cluster analysis, which Banfield & Raftery [2] suggest is due to its very generality and lack of parsimony.

Banfield & Raftery [2] use the classification likelihood formulation of the clustering problem as the basis for suggesting a number of other clustering criteria that allow some features of cluster distributions (orientation, size, and shape) to vary between clusters, while constraining others to be the same. The

key to this is a reparameterization of the covariance matrix $\mathbf{\Sigma}_k$ in terms of its eigenvalue decomposition

$$\mathbf{\Sigma}_k = \mathbf{D}_k \mathbf{\Lambda}_k \mathbf{D}_k', \tag{7}$$

where $\mathbf{D}_k$ is the matrix of **eigenvectors** and $\mathbf{\Lambda}_k$ is a diagonal matrix with the eigenvalues of $\mathbf{\Sigma}_k$ on the diagonal. The orientation of the principal components of $\mathbf{\Sigma}_k$ is determined by $\mathbf{D}_k$, while $\mathbf{\Lambda}_k$ specifies the size and shape of the density contours.

One example of Banfield & Raftery's approach leads to a generalization of the sum of squares criterion described previously, in which clusters, though still assumed spherical, are allowed to be of different sizes. The resulting criterion to be minimized is

$$\sum_{k=1}^{g} \log \left[\text{trace} \left(\frac{\mathbf{W}_k}{n_k} \right) \right]. \tag{8}$$

Other criteria allow cluster orientations to vary while keeping size and shape constant. All the suggested criteria can be implemented in the **S-PLUS** software available for applying this form of cluster analysis.

Banfield & Raftery [2] also consider the situation when the f_k are *not* multivariate normal and when there are "noise" points in the data that do not follow the general mixed distribution pattern.

Optimizing a Clustering Criterion

Once a suitable numerical clustering criterion has been selected, consideration needs to be given to how to choose the g group partition of the data which leads to its optimization. In theory, of course, the problem is simple, at least to "Dr. Idnozo Hcahscror-Tenib", that "super galactician hypermetrician" who featured in Thorndike's 1953 presidential address to the Psychometrika Society [24]: "Is easy. Finite number of combinations. Only 563 billion billion billion. Try all keep best."

Unfortunately, in practice the problem is not so straightforward since the number of partitions is enormously large and even with the fastest computers available, complete enumeration of *every* possible partition of n individuals into g groups is simply not possible. A general expression for the number of partitions, N, is given by Liu [15]:

$$N = \frac{1}{g!} \sum_{i=0}^{g} (-1)^{g-i} \binom{g}{i} i^n. \tag{9}$$

Table 1

n	g	N
15	3	2375101
20	4	45232115901
25	8	690223721118368580
100	5	10^{68}

Some specific examples are given in Table 1.

The impracticability of examining every possible partition has led to the development of algorithms designed to search for the optimum value of a clustering criterion by a nonexhaustive search procedure usually involving the rearrangement of an existing partition and keeping the new one *only* if it provides an improvement in the criterion value. Many such algorithms have been proposed, all differing in some more or less subtle ways; the *common* steps in the majority of these algorithms, however, are as follows:

1. Find some initial partition of the individuals into the required number of groups. Possible sources of such initial partitions are the solutions from a hierarchical clustering.
2. Calculate the change in the clustering criterion produced by moving each individual from its own to another cluster.
3. Make the change which leads to the greatest improvement in the value of the clustering criterion.
4. Repeat steps 2 and 3 until no move of a single individual causes the clustering criterion to improve.

Different initial partitions can lead to different *local* optima of the clustering criterion, although with well-structured data it is probably reasonable to expect convergence to the same, hopefully global, optimum from most starting configurations. Marriott [18] suggests that slow convergence and widely different partitions arising from different starting points may indicate that g has been wrongly chosen, in particular that there is no evidence of clustering.

Deciding on the "Best" Value for g

A variety of methods have been suggested for selecting the most appropriate number of clusters for a data set. The most commonly used procedure is simply to plot the value of the clustering criterion against the

number of groups and look for "large" changes of level, such changes perhaps being indicative of a particular number of groups. Clearly, such an approach may be very subjective.

More formal methods have been proposed by Beale [3], Calinski & Harabasz [4], Marriott [17], and Banfield & Raftery [2]. Marriott, for example, suggests taking the value of g for which $g^2 \det(\mathbf{W})$ is a minimum. For unimodel distributions, Marriott shows that this is likely to lead to accepting that $g = 1$, and for strongly grouped data it will indicate the appropriate value of g.

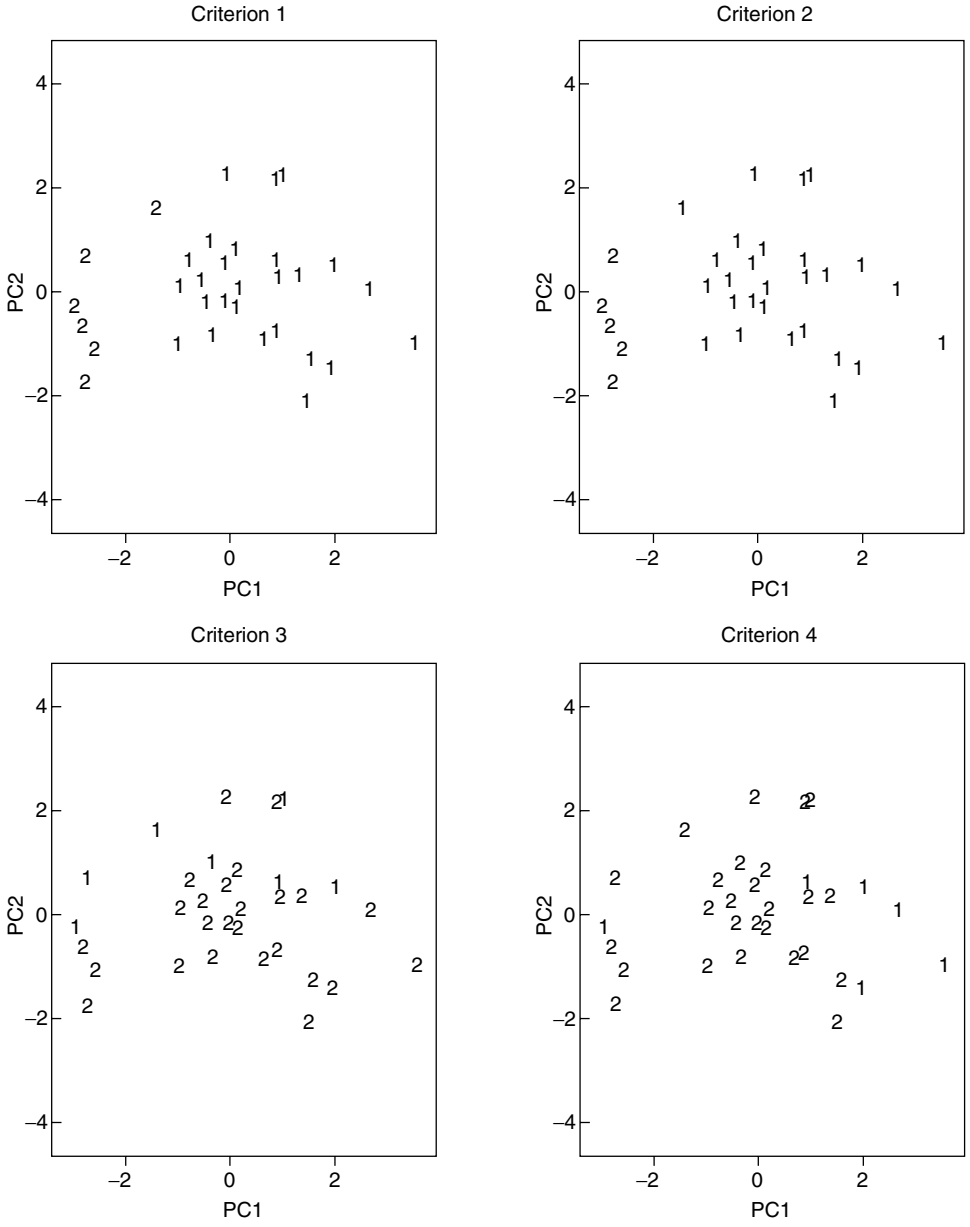

Figure 2 Two group cluster solutions of Tibetan skull data from four nonhierarchical clustering methods, displayed in the space of the first two principal components of the data

Banfield & Raftery [2] suggest an approximate **Bayesian** procedure for choosing the number of clusters and give a relatively crude estimate of $P(g|\mathbf{X})$. The maximum value of this estimate for different values of g could be used to estimate the number of clusters, although Banfield & Raftery recommend considering several values for g guided by the estimated posterior probabilities.

Some Applications of Nonhierarchical Cluster Analysis

Tibetan Skulls

Morant [19] describes data collected by Colonel L.A. Waddell on 32 skulls found in the south-western and eastern districts of Tibet (*see* **Anthropometry**). According to Morant the data can be divided into two groups. The first (type 1) comprises 17 skulls found in graves in Sikkim and neighboring areas of Tibet. The remaining 15 skulls (type 2) were picked up on a battlefield in the Lhasa district and were believed to be those of native soldiers from the eastern province of Khams. These skulls were of particular interest because it was thought at the time that Tibetans from Khams might be survivors of a particular fundamental human type, unrelated to the Mongolian and Indian types which surrounded them. On each skull the following five measurements (all in millimetres) were obtained:

x_1 : greatest length of skull
x_2 : greatest horizontal breadth of skull
x_3 : height of skull
x_4 : upper face height
x_5 : face breadth, between outermost points of cheek bones.

The data are reproduced in Hand et al. [11]. Here the a priori group structure will be ignored apart from comparing it with the two cluster solutions found from a number of optimization techniques described previously.

The clustering criteria used in this application were as follows:

1. minimization of trace($\mathbf{W}$)
2. minimization of $\sum_{k=1}^{g}, n_k \log[\text{trace}(\mathbf{W}_k/n_k)]$
3. minimization of det($\mathbf{W}$)
4. minimization of $\sum_{k=1}^{g} n_k \log[\det(\mathbf{W}_k/n_k)]$.

The resulting two group solutions found by each method are displayed graphically in Figure 2 by plotting them in the space of the first two principal components of the correlation matrix of the data. (These two components are the only ones with eigenvalues greater than one, and together they account for 60% of the variation in the data.) For comparison, the two groups as defined by Morant are shown, also in the space of the first two principal components, in Figure 3. Clearly, the four clustering solutions are considerably different from one another and each differs from the a priori grouping.

MRI Brain Scan

Banfield & Raftery [2] describe a fascinating application of optimization clustering methods to data collected from MRI scans of human brains. A randomly chosen set of 522 voxels (three-dimensional volume elements) were subjected to optimization clustering using a particular criterion (see original paper for details) and a seven-cluster solution

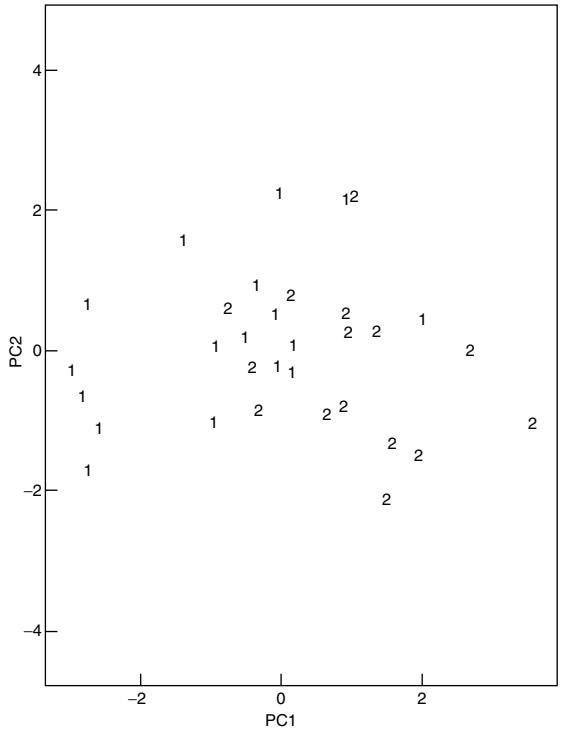

Figure 3 Original two groups in Tibetan skull data displayed in the space of the first two principal components of the data

was suggested by the Bayesian criterion mentioned earlier. The clusters were found to correspond to distinct anatomical structures. Simpler clustering techniques such as complete linkage and Ward's method (see [5]) tended to group together dissimilar anatomical voxels.

Other biostatistical applications of optimization clustering can be found in Everitt et al. [6] and Heinrich et al. [12].

Conclusions

Optimization clustering techniques produce a partition of a set of multivariate data into a particular number of groups by maximizing or minimizing some numerical index of clustering. Several such indices have been suggested, these differing in the implicit assumptions made about the shape of any clusters present. Many of the suggested criteria arise from considering a likelihood approach to the clustering problem.

References

[1] Ball, G.H. & Hall, D.J. (1967). A clustering technique for summarizing multivariate data, *Behavioural Science* **12**, 153–155.

[2] Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**, 803–822.

[3] Beale, E.M.L. (1969). Euclidean cluster analysis, *Bulletin of the International Statistical Institute* **43**, Book 2, 92–94.

[4] Calinski, T. & Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics* **3**, 1–24.

[5] Everitt, B.S. (1993). *Cluster Analysis*, 3rd Ed. Arnold, London.

[6] Everitt, B.S., Gourlay, A.J. & Kendell, R.E. (1971). An attempt at validation of traditional psychiatric syndromes by cluster analysis, *British Journal of Psychiatry* **119**, 399–412.

[7] Fisher, W.D. (1958). On grouping for maximum homogeneity, *Journal of the American Statistical Association* **53**, 789–798.

[8] Fleiss, J.L. & Zubin, J. (1969). On the methods and theory of clustering, *Multivariate Behavioural Research* **4**, 235–250.

[9] Forgey, E.W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometrics* **21**, 768–769.

[10] Friedman, H.P. & Rubin, J. (1967). On some invariant criteria for grouping data, *Journal of the American Statistical Association* **62**, 1159–1178.

[11] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994). *A Handbook of Small Data Sets*. Chapman & Hall, London.

[12] Heinrich, I., O'Hara, H., Sweetman, B. & Anderson, J.A.D. (1985). Validation aspects of an empirically derived classification for non-specific low back pain, *Statistician* **34**, 215–230.

[13] Jancey, R.C. (1966). Multidimensional group analysis, *Australian Journal of Botany* **14**, 127–130.

[14] Krzanowski, W.J. (1988). *Principles of Multivariate Analysis*. Oxford Science Publications, Oxford.

[15] Liu, G.L. (1968):. *Introduction to Combinational Mathematics*. McGraw-Hill, New York.

[16] MacQueen, J. (1967). Some method for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium*. University of California Press, Berkeley, Vol. 1, pp. 281–297.

[17] Marriott, F.H.C. (1971). Practical problems in a method of cluster analysis, *Biometrics* **27**, 501–514.

[18] Marriott, F.H.C. (1982). Optimization methods of cluster analysis, *Biometrika* **69**, 417–421.

[19] Morant, G.M. (1923). A first study of the Tibetan skull, *Biometrika* **14**, 193–260.

[20] Rao, C.R. (1952). *Advanced Statistical Methods in Biometrics Research*. Wiley, New York.

[21] Scott, A.J. & Symons, M.J. (1981). Clustering methods based on likelihood ratio criteria, *Biometrics* **27**, 387–398.

[22] Singleton, R.C. & Kautz, W. (1965). *Minimum Squared Error Clustering Algorithm*. Stanford Research Institute, Stanford.

[23] Spath, H. (1985). *Cluster Dissection and Analysis*. Ellis Horwood, Chichester.

[24] Thorndike, R.L. (1953). Who belongs in a family?, *Psychometrika* **18**, 267–276.

(*See also* **Classification, Overview**; **Cluster Analysis, Variables**; **Multidimensional Scaling**; **Pattern Recognition**; **Projection Pursuit**; **R- and Q-analysis**)

BRIAN S. EVERITT

Cluster Analysis, Variables

Cluster analysis is a process of clustering objects into groups where the groups (or clusters) are unknown a priori. Clusters are formed in such a way that objects in the same cluster are similar to each other, while members of different clusters are considerably different from each other. Similarity or dissimilarity of objects are often measured by some indices of association. These indices can be correlations, cosines, coefficients of agreement, covariances, Euclidean distances between standardized measures, **Mahalanobis generalized distances**, and other problem-oriented indices (*see* **Similarity, Dissimilarity, and Distance Measure**). The objects of clustering can be *subjects* or *variables*, but clustering subjects is probably more common than is clustering variables. For discussions on the cluster analysis of subjects, *see* **Cluster Analysis of Subjects, Hierarchical Methods** and **Cluster Analysis of Subjects, Nonhierarchical Methods**. In this article we consider techniques of performing *cluster analysis of variables*. The variables can be measurements, **ranks**, or dichotomies (*see* **Binary Data**).

Cluster analysis of variables can be applied to a wide range of problems in different fields. Popular applications relate to the construction of scorable subsets (*see* **Principal Components Analysis**; **Cluster Score**) and **battery reduction**. Most of the procedures and **algorithms** discussed in the cluster analysis of subjects can be applied to the cluster analysis of variables with the appropriate index of association, where the objects for clustering are now variables rather than subjects. Here we discuss only the type of methods that cluster variables on the basis of the **correlation** structure of the variables or on the factorial structures of the variables. Factorial structure means a structure obtained from a **factor analysis** or **principal component analysis**. Examples will be provided to demonstrate how to perform some of the techniques. The reader should note that cluster analysis of variables can differ from factor analysis even though they both examine the common factor structure of the variables. In factor analysis we are concerned with the underlying dimension of the data and in identifying latent factors that measure the dimension. In cluster analysis we desire to group variables together, often to produce sets from which cluster scores can be produced. These subsets may contain only part of the set of variables associated with a factor.

An Intuitive Approach

A simple intuitive approach to cluster analysis uses the results of a factor analysis or a principal component analysis by forming clusters of variables that have high loadings on the same factor. **Factor loadings** are the elements of the final factor matrix and they represent the correlations between the variables and the factors. We can use these loadings as indices of association for measuring the similarity or dissimilarity of the variables in question (*see* **Principal Components Analysis**; **Factor Analysis, Overview**, for procedures to obtain the final factor matrix).

Example

To illustrate this intuitive approach, we selected seven variables from the **Framingham** offspring data. These variables are the systolic blood pressure (SPF), diastolic blood pressure (DPF), height (HGT), total volume capacity (TVC), triglycerides (TG), total cholesterol (SCL), and high density lipoprotein cholesterol (HDL). These variables are selected solely for the purpose of illustration and not for substantive interpretation. There is a total of 2370 nondiabetic males included in the analysis. A factor analysis with **varimax rotation** was performed on this data. Three factors were retained. The initial factor matrix as well as the final rotated factor matrix are presented in Table 1. In this example we use the factor matrix from a factor analysis. In practice, we could also perform a principal components analysis and also we could use other types of rotation. Comparisons between a factor analysis and a principal components analysis can be found in the articles on **Principal Components Analysis** and **Factor Analysis, Overview**.

Using the rule of thumb of Cureton & D'Agostino [1] we use 0.3 as the minimum threshold for considering a loading "significant" and so also for putting a variable into a cluster. With a 0.3 threshold, three disjoint clusters are formed: (i) SPF and DPF (with their first factor loadings equal to 0.867 and 0.866); (ii) HGT and TVC (with their second factor

Table 1 Initial and rotated factor matrices

	Initial factor matrix		
Variable	*a*	*b*	*c*
SPF	0.815	0.236	−0.194
DPF	0.833	0.242	−0.125
HGT	−0.126	0.626	0.194
TVC	−0.309	0.599	0.123
TG	0.396	−0.146	0.497
SCL	0.391	−0.187	0.251
HDL	0.069	0.002	0.421

	Rotated factor matrix[a]		
Variable	*a*	*b*	*c*
SPF	0.867	−0.047	0.067
DPF	0.866	−0.028	0.136
HGT	0.043	0.665	0.030
TVC	−0.111	0.669	−0.095
TG	0.168	−0.107	0.621
SCL	0.218	−0.209	0.399
HDL	−0.055	0.096	0.412

[a]Varimax rotation.

loadings equal to 0.665 and 0.669); and (iii) TG, SCL, and HDL (with their third factor loadings equal to 0.621, 0.399, and 0.412). We do not always obtain disjoint clusters as presented in this example although this is usually considered ideal. It is possible to obtain clusters with overlapping variables.

One available **software** package in SAS, PROC VARCLUS, always produces clusters with nonoverlapping elements [8]. This is an oblique component analysis related to multiple group factor analysis. Standardized scoring coefficients are provided for the computations of cluster scores.

Graphical Cluster Analysis of Variables

Tryon [9] proposed finding clusters of variables on the basis of the profiles of the correlations of the variables. He suggested plotting each row of the correlation matrix, having the variable number on the abscissa and the numerical value of the correlation coefficient on the ordinate. We denote the correlation coefficient by r_{ij}, where $i, j = 1, \ldots, p$ and p is the number of the original variables. Each row of points (i.e. $r_{i1}, r_{i2}, \ldots, r_{ip}$) is connected by a jagged line, leaving a gap at r_{ii}. A cluster is then identified by the subset of *approximately parallel proportional* profiles. An example of a correlation profile plot is given in Figure 1. These are the correlations for the seven variables of the Framingham offspring data. The correlation matrix is displayed in Table 2.

From Figure 1, SPF and DPF can be seen to have high correlations with each other and their profiles are very close throughout. Thus, they form a cluster. HGT and TVC are moderately correlated with each other and they exhibit approximately parallel proportional profiles across the plot. It seems reasonable to group them in one cluster. TG and SCL have some correlations with each other and they have very similar correlation patterns with other variables except with HDL. Based on the general profiles, we can put them in a cluster. The profile of HDL does not show any resemblance to the profiles of any other variables, therefore it forms a cluster by itself. We sometimes call a variable in such a cluster an *outlier*. On the basis of this plot, we may conclude that there are four clusters: (i) SPF-DPF, (ii) HGT-TVC, (iii) TG-SCL, and (iv) HDL. In practice, it is not easy to identify the subsets of profiles because these profiles often exhibit similar shape but different height, and the gaps at the r_{ii}s make the profiles quite difficult to follow.

Table 2 Correlation matrix

Variable	SPF	DPF	HGT	TVC	TG	SCL	HDL
SPF	1.000	0.829	0.011	−0.146	0.197	0.205	−0.026
DPF	0.829	1.000	0.036	−0.141	0.228	0.250	0.018
HGT	0.011	0.036	1.000	0.558	−0.042	−0.097	0.084
TVC	−0.146	−0.141	0.558	1.000	−0.154	−0.196	0.006
TG	0.197	0.228	−0.042	−0.154	1.000	0.404	0.335
SCL	0.205	0.250	−0.097	−0.196	0.404	1.000	0.013
HDL	−0.026	0.018	0.084	0.006	0.335	0.013	1.000

Note: The matrix is computed from Framingham offspring data.

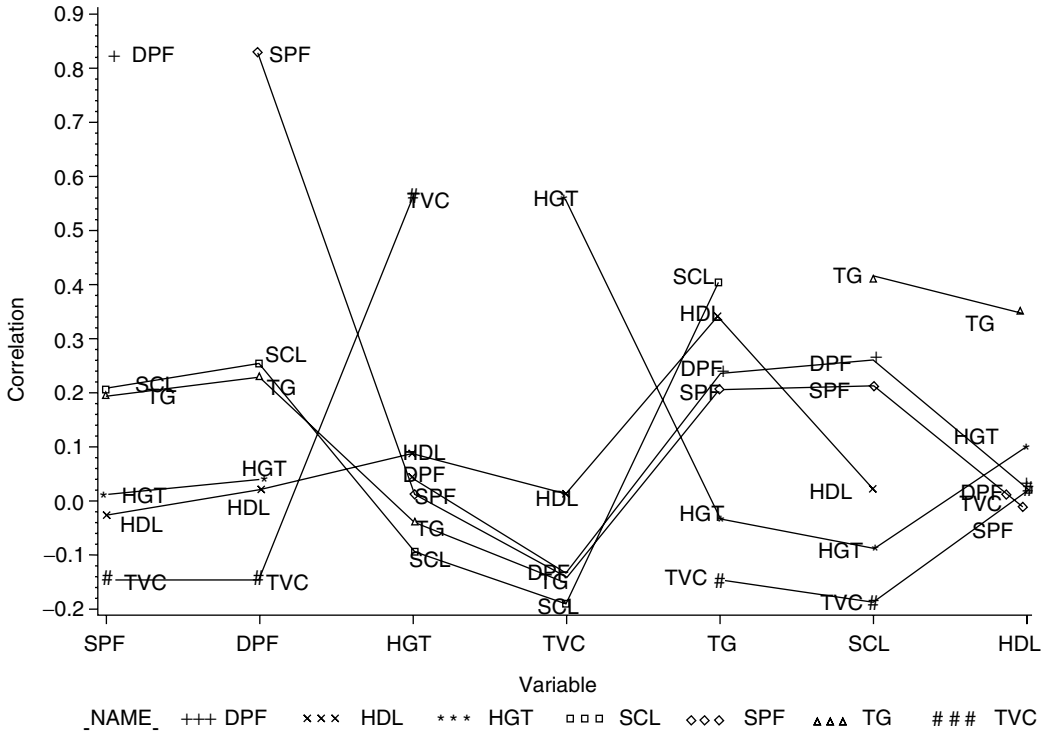

Figure 1 Graphical cluster analysis (correlation profiles), Framingham offspring data (nondiabetic men)

Cureton & D'Agostino [1] present another graphical method based on the profiles of the factorial structures of the variables which usually avoids some of the difficulties in Tryon's procedure. Using their approach, we first compute a row-normalized varimax factor matrix (see [5] and [1]) or a weighted-varimax factor matrix (see [2]). If these varimax factor matrices are not available, then a Landahl transformation to the row-normalized initial factor matrix can be used instead (see [6]). Then we plot each row of this transformed matrix, having the factor number (i.e. 1 to m) on the abscissa and the numerical value of the factor loadings on the ordinate. An example is given in Table 3 to illustrate the step-by-step computations of this Landahl transformed initial factor matrix. The advantages of using the profiles from a transformed row-normalized initial factor matrix over the profiles from a correlation matrix are twofold: (i) this transformed initial factor matrix provides the same information contained in the correlation matrix but there are only m points to plot for each variable instead of $p - 1$ and no gaps are left in the profiles, and (ii) in this factor matrix plot, the profiles belonging to the same cluster exhibit about the same average shape and height, so it is easier to visualize subsets of variables that belong to the same cluster. Sometimes, we may add the profiles of the primary axes (*see* **Primary Factors**). We should try to avoid breaking up the variables that have profiles close to those of the primary axes. If we are not sure about some of the subsets, then we can always quantify the association of the profiles by computing the cosines of angle between each pair of the row-normalized initial factor matrix vectors (*see* **Cosine of Angle Between Two Vectors**; **Matrix Algebra**). The complete cosine matrix, denoted by **K**, is of dimension $p \times p$. It is obtained by postmultiplying the row-normalized initial factor matrix by its transpose. Two variables presumed to lie in the same cluster should have a high cosine value. These cosines are derived as an index of association for the profiles of variables. They can be used to determine numerically the compactness of a cluster. We can also set an acceptance level to control the level of compactness we wish for a cluster. A

Table 3 Transformation of the row-normalized initial factor matrix

	Principal axes[a] (**F**)			Normalization factors (**f**)
Variable	*a*	*b*	*c*	$(a^2 + b^2 + c^2)^{-1/2}$
SPF	0.815	0.236	−0.194	1.14896
DPF	0.833	0.242	−0.125	1.14050
HGT	−0.126	0.626	0.194	1.49861
TVC	−0.309	0.599	0.123	1.46058
TG	0.396	−0.146	0.497	1.53342
SCL	0.391	−0.187	0.251	1.99721
HDL	0.069	0.002	0.421	2.34542

Normalized initial factor matrix with primary axes ($\mathbf{F}_n$)				
Variable	*a*	*b*	*c*	
SPF	0.936	0.271	−0.223	diag (**f**) ∗ **F**
DPF	0.951	0.276	−0.142	
HGT	−0.188	0.938	0.290	
TVC	−0.451	0.874	0.179	
TG	0.608	−0.224	0.762	
SCL	0.781	−0.374	0.501	
HDL	0.163	0.005	0.987	
A	0.899	0.336	−0.280	transpose of orthogonal
B	−0.263	0.927	0.268	transformation matrix
C	0.350	−0.167	0.922	(obtained from SAS output)

Landahl transformation matrix[b] ($\mathbf{L}_t$)			
	A	*B*	*C*
a	0.5774	0.5774	0.5774
b	0.8165	−0.4083	−0.4083
c	0.0000	0.7071	−0.7071

Normalized Landahl factor matrix ($\mathbf{L}_n = \mathbf{F}_n * \mathbf{L}_t$)			
Variable	*a*	*b*	*c*
SPF	0.762	0.272	0.587
DPF	0.774	0.335	0.537
HGT	0.657	−0.287	−0.697
TVC	0.454	−0.491	−0.744
TG	0.168	0.981	−0.096
SCL	0.146	0.957	0.249
HDL	0.098	0.790	−0.606
A	0.793	0.184	0.580
B	0.605	−0.341	−0.719
C	0.065	0.922	−0.382

Note: The matrices are based on the Framingham offspring data.
[a]The principal axes matrix is obtained from the SAS output.
[b]The Landahl transformation matrix is obtained from [6].

variable is accepted as a member of a cluster only if its cosine with the existing members of the cluster is higher than the acceptance level. When an acceptance level is used in the analysis, one should be aware that the choice for this threshold is often arbitrary, but the level we choose may affect the composition of the clusters. An acceptance level that is too high may create many outliers and also result in too many small clusters. But when the acceptance level is too low, the analysis tends to form large clusters with heterogeneous elements in them. Therefore, in practice we should try different acceptance levels to see which result gives the most sensible interpretation.

For illustration we use again the seven variables from the Framingham offspring data. We first perform a factor analysis, and decide to retain three

factors. Table 3 presents the computations of the transformation. A SAS macro [4] is available to perform all these calculations and plot the profile, as shown in Figure 2. In this example a varimax rotation is applied to the primary axes. In practice, we can also employ other types of **orthogonal rotation** or **oblique rotation** (*see* **Rotation of Axes**). A different rotation may produce a different set of primary axes.

In Figure 2, SPF and DPF are close together on each of the three factors so they form a cluster of two variables. HGT and TVC are close together on the scale for factor 3 but moderately apart on the scales for factors 1 and 2. We tentatively put them into the same cluster. TG, SCL, and HDL are close together for factor 1. HDL begins to drift slightly downwards from TG and SCL for factor 2 and then all three variables drift farther apart from each other for factor 3. So, we tentatively call them a cluster. The next step is to check the primary axes. The profile of the primary axis, A, is fairly close to the SPF–DPF cluster and the profile of the primary axis, B, is wholly contained within the HGT–TVC cluster. The profile of the primary axis, C, is close to the TG–SCL–HDL cluster for factor 1 but it moves away from HDL for factor 2. For factor 3, all these variables and the primary axis, C, are lying apart from each other. To verify the tentative clusters, we compute the cosines of angles between each pair of the normalized initial factor matrix vectors. An example of the cosine matrix is provided in Table 4, which is computed from the row-normalized initial factor matrix given in Table 3.

SPF and DPF are a compact cluster with a cosine of 0.997. HGT and TVC are reasonably compact with a cosine of 0.957. As reflected in Figure 2, TG, SCL, and HDL are not quite compact. TG and SCL have a cosine of 0.940, TG and HDL have a cosine of 0.849, while SCL and HDL have a cosine of 0.620. From this graphic cluster analysis we may conclude that SPF–DPF, HGT–TVC, and TG–SCL–HDL are three distinct clusters if no acceptance level is used. But if we use, for example, 0.9 as the acceptance level, then we need to break the last cluster into two smaller clusters and HDL will become an outlier and form a cluster by itself. We obtain the same clustering results for using acceptance levels 0.7 and 0.8. We can see that this clustering is consistent with the result given in the correlation profile.

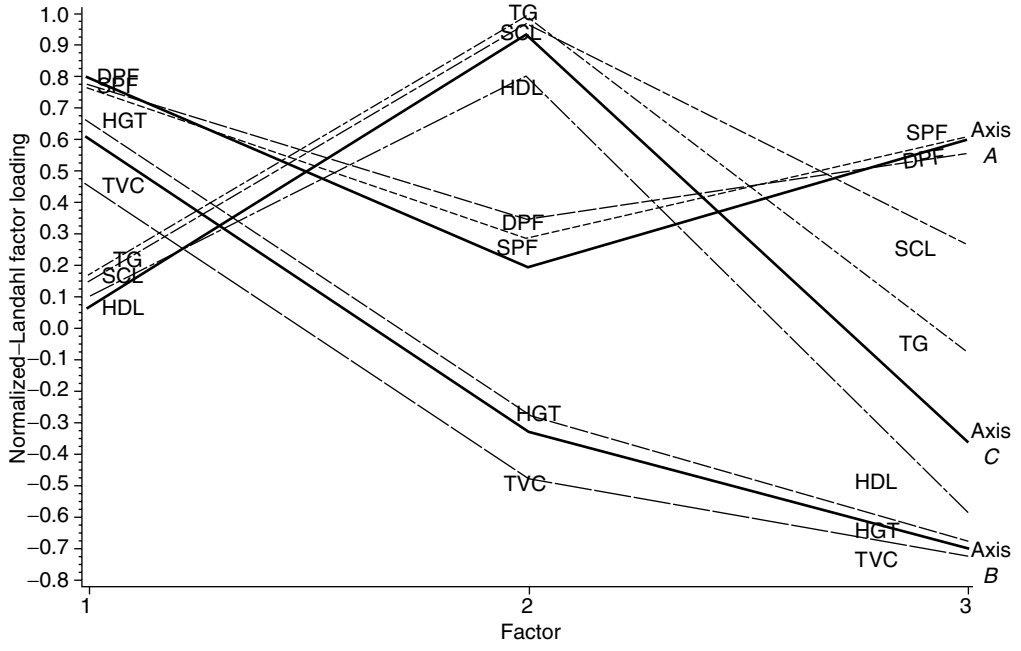

Figure 2 Graphical cluster analysis (normalized–Landahl factor profiles). Framingham offspring data (nondiabetic men)

Table 4 Cosine matrix ($\mathbf{K} = \mathbf{F}_n * \mathbf{F}_n{}'$)

Variable	SPF	DPF	HGT	TVC	TG	SCL	HDL
SPF	1.000	0.997	0.014	−0.225	0.339	0.518	−0.066
DPF	0.997	1.000	0.039	−0.213	0.408	0.568	0.015
HGT	0.014	0.039	1.000	0.957	−0.104	−0.352	0.260
TVC	−0.225	−0.213	0.957	1.000	−0.334	−0.589	0.108
TG	0.339	0.408	−0.104	−0.334	1.000	0.940	0.849
SCL	0.518	0.568	−0.352	−0.589	0.940	1.000	0.620
HDL	−0.066	0.015	0.260	0.108	0.849	0.620	1.000

Note: The matrix is computed from Framingham offspring data.

Elementary Linkage Analysis

The elementary linkage analysis is a crude but simple numerical method that can be applied in the cluster analysis of variables. This method was first described in McQuitty [7] and later modified by Cureton & D'Agostino [1]. Elementary linkage analysis is equivalent to performing a **single linkage** analysis (*see* **Cluster Analysis of Subjects, Hierarchical Methods**) using the cosine of angle between two vectors as the index of association. We start by underlining the highest cosine in each column of **K**, the complete cosine matrix. The general rule is that a variable, V_1, belongs in the same cluster with another variable, V_2, if V_1 has the highest cosine with V_2. If V_1 and V_2 have the same highest cosine with each other, then they are called a *reciprocal pair*. The cluster that is formed by the reciprocal pair is called a *nuclear cluster*, which we should avoid breaking. After finding the nuclear clusters, we look along the row of each member of the nuclear cluster to see if there is other underlined cosine on the row. If there is, then the member of the pair is said to bring in a new member. We continue to check the other members of the cluster and locate any "bring-in" members. If no further underlined cosines are found in all the current members of the cluster, then the cluster is complete. When all the clusters are complete, the standard elementary linkage analysis is done. Again, we can form compact clusters in this analysis by specifying an acceptance level. Using an acceptance level in this analysis, a variable can become an *intrinsic outlier*. It is a variable which has the highest cosine in a column of the cosine matrix but the value is below the acceptance level. A variable can also become a forced outlier if its cosine with one of the existing members of the cluster is below the acceptance level. There is no specified rule on how to handle these outliers in the analysis. It relies on the judgment of the investigator.

Recall the cosine matrix given in Table 4 and underline the highest cosine in each column. According to the definition, SPF–DPF, HGT–TVC, and TG–SCL are the reciprocal pairs and they form three distinct nuclear clusters. These clusters can be represented as follows:

$$\text{SPF} \overset{0.997}{\longleftrightarrow} \text{DPF HGT} \overset{0.957}{\longleftrightarrow} \text{TVC TG} \overset{0.997}{\longleftrightarrow} \text{SCL}$$

The next step is to locate other potential members for each nuclear cluster by checking other underlined cosines along the rows. We can see that TG brings in HDL to the TG–SCL cluster and the relationship is graphically presented as follows:

$$\begin{array}{c} \text{TG} \overset{0.941}{\longleftrightarrow} \text{SCL} \\ \downarrow 0.867 \\ \text{HDL} \end{array}$$

There are no other underlined cosines, so all the clusters are complete and so is the analysis. Of course, if we want to form compact clusters with an acceptance level, say 0.9, then HDL will become a forced outlier and form a cluster by itself. These results are consistent with the results found in the graphical cluster analyses.

Cycle Hunt Analysis

Cureton et al. [3] proposed a more complex system of clustering called a cycle hunt analysis. This analysis consists of two stages: cycle hunt and cycle change. Before we start the cycle hunt stage, we need to set an acceptance level. If the objective of the study is to form clusters from subsets of variables, then we

will also need an exclusion level. An exclusion level is the lowest acceptable **communality** for a variable to be included in a cluster. The communalities are computed from the initial factor matrix before its rows are normalized. We then compute the complete cosine matrix, **K**, as previously defined and delete from **K** the row and column corresponding to each variable whose cosine is below the acceptance level, and also whose communality is below the exclusion level. Then this reduced **K** matrix is used in the rest of the analysis.

The cycle hunt stage forms clusters by performing either one of the three operations: (i) combining two variables, neither of which belongs to any existing cluster; (ii) adding to an existing cluster a variable not previously in any cluster; and (iii) combining two clusters to form a larger cluster. At each step we perform the operation that will leave the lowest within-cluster cosine highest and make sure the cosines between pairs of variables in the clusters are above the acceptance level. At this stage, once a variable is assigned to a cluster, it stays with that cluster for the whole cycle. The cycle hunt terminates when no further operations of type (i) or (ii) can be done without the lowest cosine going beyond the acceptance level.

The cycle change stage starts after the cycle hunt is completed. This second stage is divided into two subcycles. In subcycle 1, we start arbitrarily with a variable, say V. For each cluster, including the one V is in, we find the variable with which V has the lowest cosine. Among all these lowest cosines we find the particular variable, say W, whose lowest cosine with V is the highest. Then, V will join the cluster of W. If V is a forced outlier itself, then we need to make sure that the lowest cosine is above the acceptance level. After the reassignment of V, we proceed to the next variable and repeat the same procedure. When all the variables are considered, we return to the first variable, V, and repeat the entire procedure. Subcycle 1 terminates when there are no changes made throughout the entire procedure. In subcycle 2 we compute the lowest within-cluster cosine that results from combining each pair of clusters. We make sure that the cosine is above the acceptance level, and combine the one pair whose lowest cosine is the highest. We then repeat subcycles 1 and 2 in sequence until, at subcycle 2, no two clusters can be combined without producing a within-cluster cosine below the acceptance level. At this point the cycle change terminates and the cycle hunt analysis is complete. Cureton & D'Agostino [1] suggested that, in practice, we should repeat the cycle hunt analysis using several different acceptance levels but the same exclusion level, if any. We then choose as the final result the one with the highest acceptance level that gives a "sensible" result.

A SAS macro [4] is available to perform this complex algorithm of clustering. Despite its complexity, the cycle hunt system does offer some advantages over the graphical cluster analysis and the elementary linkage analysis. It can handle large numbers of variables. Unlike the elementary linkage analysis the cycle hunt analysis considers not only the indices of association for variables that belong to the same cluster, but also the indices of association for variables that are in different clusters. The associations of the outliers with other variables also get examined.

We now revisit the example based on the Framingham offspring data. We apply the cycle hunt macro in [4] to perform a cluster analysis on the seven variables. Taking advice from Cureton & D'Agostino [1], we repeat the analysis using 0.9, 0.8, 0.7, 0.6, and 0.5 as the acceptance levels. No exclusion level is used in this example. All the analyses except the one using 0.5 as the acceptance level lead to the same clustering: SPF and DPF belong to one cluster; HGT and TVC form another cluster; TG and SCL form the third; and HDL is not included in any cluster. This result is consistent with the results obtained from both the graphical cluster analysis and the elementary linkage analysis using an acceptance level.

Conclusions

The factor matrix, correlation matrix, and cosine matrix are the major indices of association used in the above clustering methods. When the method is based on a factor matrix obtained from a factor analysis, the contributions of the common factors are emphasized in clustering the variables. But there are other alternatives on which we can base the cluster analysis. If we are interested in including the contributions of the unique factors, then we can form the cosine matrix from a component matrix. If we wish to include the specific factors as well as the common factors, but not the error factors, then we can start a cluster analysis from a correlation matrix with reliability coefficients on the diagonal (see the discussions on reliability coefficients in [1, p. 365]).

To summarize, it is clear that each method described above has its own advantages and disadvantages. For the intuitive approach, all we need is a final rotated factor matrix, which is easily obtained from most of the statistical packages. No extra computations are required. However, in practice it is not easy to interpret the final matrix and form clusters that make sense. The VARCLUS procedure produces logical groupings on the basis of various maximization criteria. It is a numerical procedure that does not need subjective input to obtain the clusters. However, to achieve its goal of producing disjoint clusters from a set of variables, the procedure sometimes forces variables to join exactly one cluster, even though the variables may not belong to one cluster exclusively. The graphical method provides an effective means of presenting the somewhat complicated concept of the factorial structure of the variables. Many find it easier to understand a pictorial representation than a display of numbers. However, when there are a large number of unreliable variables involved in the plot, the display of the structure can become confusing. The elementary linkage method is a simple numerical procedure. Reciprocal pairs are used as its basic building block for clusters. The cluster-forming process is rather crude. Some important associations among variables other than reciprocal pairs are not vigorously explored. The cycle hunt analysis remedies some of the weaknesses of the simpler methods. It is able to handle large numbers of variables and it is a more formal technique for finding the subsets of variables. But this system involves a complex algorithm and it is very computer-intensive. Also, the analysis requires the use of an acceptable level while there is no formal rule to decide what level should be used. In a nutshell, there is obviously no one best clustering method. In practice, we recommend applying different methods to the same data and interpreting the clusters carefully to obtain a "sensible" solution. We also recommend repeating the analyses with different acceptance levels, especially when there are large numbers of unreliable variables.

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[2] Cureton, E.E. & Mulaik, S.A. (1975). The weighted varimax rotation and the promax rotation, *Psychometrika* **40**, 183–195.

[3] Cureton, E.E., Cureton, L.W. & Durfee, R.C. (1970). A method of cluster analysis, *Multivariate Behavioral Research* **5**, 101–116.

[4] D'Agostino, R.B., Dukes, K.A., Massaro, J. & Zhang, Z. (1992). Data/variable reduction by principal components, battery reduction and variable clustering, in *NESUG '92 Proceedings*. Connecticut, pp. 464–474.

[5] Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

[6] Landahl, H.D. (1938). Centroid orthogonal transformation, *Psychometrika* **3**, 219–223.

[7] McQuitty, L.L. (1957). Elementary linkage analysis isolating orthogonal and oblique types and typal relevancies, *Educational and Psychological Measurement* **17**, 207–229.

[8] SAS Institute Inc. (1989). *SAS/STAT® User's Guide, Version 6*, 4th Ed., Vol. 2. SAS Institute Inc., Cary, pp. 1641–1659.

[9] Tryon, R.C. (1939). *Cluster Analysis*. Edwards, Ann Arbor.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Cluster Randomization

A cluster randomization trial is one in which intact social units, or clusters of individuals, rather than individuals themselves, are randomized to different intervention groups. Trials randomizing clusters, sometimes called **group randomization** trials, have become particularly widespread in the evaluation of nontherapeutic interventions, including lifestyle modification, educational programs and innovations in the provision of health care. The units of randomization in such studies are diverse, ranging from relatively small clusters, such as households or families, to entire neighborhoods or communities, but also including worksites, hospital wards, classrooms and medical practices. There are also reports of trials that have randomized more unusual units, including athletic teams [76], tribes [34], religious institutions [48] and sex establishments [28].

Proper accounting for the clustering of subjects' responses within the units of randomization is a key challenge faced by investigators adopting this design. We begin by describing how such clustering arises and how it reduces the efficiency of cluster randomization relative to individually randomized trials. A brief historical discussion of cluster randomization is then provided, followed by a discussion of ethical issues. More technical material that deals with the impact of cluster randomization on trial methodology, including the selection of a study design, sample size estimation and data analysis, is presented in the next few sections. We conclude by providing some guidelines for trial reporting. The reader interested in a more detailed discussion might wish to consult [20] from which this article was abstracted.

The Impact of Clustering

The degree of similarity among responses within a cluster is typically measured by a parameter known as the intracluster (intraclass) correlation coefficient. Denoted by the Greek letter ρ, this parameter may be interpreted as the standard Pearson correlation coefficient between any two responses in the same cluster. Stating that ρ is positive is equivalent to assuming that the variation between observations in different clusters exceeds the variation within clusters. Under these conditions, we may say that the design is characterized by "between-cluster variation".

The underlying reasons for variation between clusters will differ from trial to trial, but in practice may include the following.

1. Subject selection, where individuals are in a position to choose the cluster to which they belong. For example, in a trial randomizing medical practices, the characteristics of patients belonging to a practice could be related to age or sex differences among physicians. To the extent that these characteristics are also related to patient response, a clustering effect will be induced within practices.
2. The influence of covariates at the cluster level, where all individuals in a cluster are affected in a similar manner as a result of sharing exposure to a common environment. For example, infection rates in nurseries may vary owing to differences in temperature or other environmental conditions, while differences in bylaws between communities could influence the success of smoking cessation programs. Furthermore, when intact families or households are randomized, the combined effect of both environmental and genetic factors will contribute to the observed between-cluster variation.
3. The tendency of infectious diseases to spread more rapidly within than among families or communities. The possibility of outbreaks or epidemics in some clusters, the method by which the infectious agent is spread and its virulence will also affect the degree of between-cluster variation in rates of disease.
4. The effect of personal interactions among cluster members who receive the same intervention. For example, educational strategies provided in a group setting could lead to a sharing of information that create a clustering effect. More generally, as noted by Koepsell [45], just as infectious agents can be spread from person to person, the transmission of attitudes, norms and behaviors among people who are in regular contact can result in similar responses.

Without extensive empirical data, it is usually impossible to distinguish among the potential reasons for between-cluster variation. Regardless of the specific cause, however, such variation invariably leads to a reduction in precision in estimating the effect of intervention, where the size of the reduction increases

with both the magnitude of ρ and the average cluster size.

These effects of clustering may be expressed quantitatively in a fairly simple fashion. Consider an experimental trial in which k clusters of m individuals are randomly assigned to each of an experimental group and a control group. We suppose that the primary aim of the trial is to compare the groups with respect to their mean values on a normally distributed response variable Y having a common but unknown variance σ^2. Estimates of the population means μ_1 and μ_2 are given by the usual sample means $\overline{Y}_1$ and $\overline{Y}_2$ for the experimental and control groups, respectively. From a well-known result in cluster sampling (e.g. Kish [43, Chapter 5]), the variance of each of these means is given by

$$\text{var}(\overline{Y}_i) = \frac{\sigma^2}{km}[1 + (m-1)\rho], \quad i = 1, 2, \qquad (1)$$

where ρ is the intracluster correlation coefficient. If σ^2 is replaced by $P(1-P)$, where P denotes the probability of a success, then (1) also provides an expression for the variance of a sample proportion under clustering.

The intracluster correlation coefficient ρ may be interpreted as the proportion of overall variation in response that can be accounted for by the between-cluster variation. With this interpretation, we may write

$$\rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_W^2} \qquad (2)$$

where σ_A^2 is the between-cluster component of variance, σ_W^2 the within-cluster component, and $\sigma^2 = \sigma_A^2 + \sigma_W^2$. A sample estimate of ρ may be obtained using a standard one-way analysis of variance among and within clusters (e.g. Armitage & Berry [2, Section 8.7]).

For **sample size determination**, (1) implies that the usual estimate of the required number of individuals in each group should be multiplied by the variance inflation factor $IF = 1 + (m-1)\rho$ to provide the same statistical power as would be obtained by randomizing km individuals to each group when there is no clustering effect. This expression is also well known in the sample survey literature, in which it is referred to as a "design effect" [43, p. 162]. The special case $\rho = 0$ corresponds to that of statistical independence among members of a cluster. The case $\rho = 1$, on the other hand, corresponds to total dependence. In this case all responses in a cluster are identical, so that the total information supplied by the cluster is no more than that supplied by a single member, i.e. the "effective cluster size" is one. In general, the effective cluster size is given by the simple formula $m/[1 + (m-1)\rho]$.

Given this loss of efficiency relative to individual randomization, the reasons for adopting cluster randomization must clearly rest on other considerations. Basic feasibility considerations were cited by Bass et al. [4] in their choice of a cluster randomized design for evaluating a program to enhance the effectiveness of hypertension screening and management in general practice. It was recognized that such a program would not function effectively if some patients in a practice but not others were entered into it. Randomization at the practice level also enhanced physician **compliance** by avoiding the potential ethical challenges that could arise when not all patients in a practice are offered a new intervention. Finally, randomization at this level also avoids the contamination that could occur when knowledge of the experimental intervention influences the responses of subjects in the control group.

Investigator concerns regarding the possibility of contamination may be particularly acute in trials of infectious diseases where the aim of the study is to reduce the transmission of infection. Individual randomization might prove impractical in such trials because the dynamics of transmission might lead subjects who do not receive the intervention to nevertheless receive protection as a consequence of herd immunity [14, 38].

In some studies randomization by cluster is the only natural choice or a clear necessity, with no special justification required. This was arguably the case in the HIV prevention trial described by Grosskurth et al. [36]. As the authors note, randomization was necessary at the community level since the intervention involved the provision of improved services at designated health facilities, with these services available to the entire population served by each facility.

Historical Development of Cluster Randomization

The British **Medical Research Council's** 1946 study of streptomycin for the treatment of tuberculosis is generally considered to be the first publication

of a clinical trial with a properly randomized control group [54, 60, pp. 17–18]. The success of the streptomycin trial in instilling the virtues of random assignment among clinical researchers was at first quite modest. For example, none of the 29 trials reported in the *New England Journal of Medicine* in 1953 used randomized controls [11]. In spite of a fairly steady and dramatic increase, only 50% of clinical trials published in the late 1970s could claim to have employed **randomization**. There are also very few examples of properly designed and analyzed cluster randomization trials conducted by health care researchers prior to 1978.

However, there were several notable cluster randomization trials conducted by epidemiologists interested in evaluating methods for preventing tuberculosis (e.g. [26]). These investigators randomized 433 groups of hospital wards in double-blind fashion (*see* **Blinding or Masking**) to either an experimental or placebo control group. The test of intervention effect was adjusted for clustering by adapting a method described by Cochran [12] for the analysis of sample survey data.

The statistical features of cluster randomization were first brought to wide attention in the health research community by Cornfield [15]. His paper made it clear that such allocation schemes are less efficient, in a statistical sense, than designs that randomize individuals to intervention groups. This general result, however, was recognized much earlier in the statistical literature (see [75]).

The 1980s saw a dramatic increase in the development of methods for analyzing correlated outcome data in general (e.g. [3] and [77]) and methods for the design and analysis of cluster randomized trials in particular (e.g. [21] and [32]). As might be expected, publication of this work did not immediately translate into any marked improvement in the methodological quality of cluster randomized trials. The difficulties investigators continued to experience with the design and analysis of cluster randomization trials were demonstrated in several methodological reviews (e.g. [22], [70] and [71]). Similar results were reported in each of these reviews, with less than 25% of the studies considered accounting for between-cluster variation when determining trial power. The situation was somewhat improved with respect to data analysis, where the effects of clustering were seen to be accounted for by at least 50% of the trials considered in each review. It is reasonable to expect further improvements in the methodological quality of cluster randomization trials as appropriate methods of analysis and reporting are gradually being incorporated into standard checklists for reporting randomized controlled trials (e.g. [5]).

The Role of Informed Consent

Every randomized trial requires assurance that the proposed study meets commonly accepted ethical standards. This task is particularly complex for cluster randomization trials since almost all ethical guidelines have been written from the sole perspective of trials that randomize individuals. Only recently has attention been given to the unique ethical challenges posed by cluster randomization (e.g. [24] and [33]) (*see* **Medical Ethics and Statistics**).

For instance, according to the World Medical Association Declaration of Helsinki (World Medical Association [79]) consent must be obtained from each patient prior to random assignment. The situation is more complicated for cluster randomization trials particularly when randomizing larger units (e.g. schools, communities, worksites). Then school principals, community leaders or other key decision-makers will usually provide permission for both random assignment and implementation of the intervention. Individual study subjects must still be free to withhold their participation, although they may not even then be able to avoid completely the inherent risks of an intervention that is applied on a cluster-wide level.

The identification of individuals mandated to provide agreement for random assignment may not be a simple task. Typically it is elected or appointed officials who make such decisions. However, as Strasser et al. [73] point out, it is by no means certain when or even if securing the agreement of these officials is sufficient.

The practical difficulties of securing informed consent prior to random assignment do not necessarily arise when smaller clusters such as households or families are the unit of randomization. For instance, Payment et al. [59] evaluated the risk of gastrointestinal disease in households randomly assigned to receive domestic water filters as compared with households using tap water. One of the eligibility criteria was “willingness to participate in a longitudinal trial in which a random half of the households would have a filter installed”.

The relative absence of ethical guidelines for cluster randomized trials appears to have created a research environment in which the choice of randomization unit may determine whether or not informed consent is deemed necessary prior to random assignment. This phenomenon can be seen, for example, in the several published trials of vitamin A supplementation on childhood mortality. Informed consent was obtained from mothers prior to assigning children to either vitamin A or placebo in the household randomization trial reported by Herrera et al. [39]. This was not the case in the community intervention trial of vitamin A reported by the Ghana VAST Study Team [31], where consent to participate was obtained only after random assignment. It seems questionable, on both an ethical and methodological level, whether the unit of randomization should play such a critical role in deciding whether or not informed consent is required.

Selecting an Experimental Design

There are three designs that are most frequently adopted in cluster randomization trials:

1. completely randomized, involving no pre-stratification or matching of clusters according to baseline characteristics;
2. matched-pair, in which one of two clusters in a stratum are randomly assigned to each intervention;
3. stratified, involving the assignment of two or more clusters to at least some combinations of stratum and intervention.

An interesting example of the completely randomized design is given by the ACEH study, as reported by Abdeljaber et al. [1]. This trial was designed to evaluate the effectiveness of vitamin A supplementation on the one-year prevalence of cough, fever and diarrhea among Indonesian children. The completely randomized design was very appropriate for the ACEH trial since there were over 200 villages assigned to each of the experimental and control groups.

Matching or stratification is often considered for community intervention trials in which the numbers of clusters that can be enrolled may be limited by economic or practical considerations. The main advantage of this design is its potential to provide very tight and explicit balancing of important prognostic factors, thereby improving the power for detecting the effect of intervention. An illustrative example is the COMMIT trial which was designed to promote smoking cessation using a variety of community resources [13]. This trial involved 11 pairs of communities matched on the basis of community size, population density, demographic profile, community structure and geographic proximity.

In spite of its obvious potential for creating comparable groups of subjects, there are some important analytic limitations associated with the matched-pair design. These limitations arise because of the inherent feature of this design that there is exactly one cluster assigned to each combination of intervention and stratum. As a result, the natural variation in response between clusters in a matched pair is totally confounded with the effect of intervention. Thus it is impossible to obtain a valid estimate of ρ without making special assumptions (e.g. the intervention has no effect). This difficulty, which complicates both sample size determination and data analysis, is explored further by Klar & Donner [44].

The stratified design is an extension of the matched-pair design in which several clusters, rather than just one, are randomly assigned within strata to each of the intervention and control groups. An example of this design is provided by CATCH [52]. In this study the unit of randomization was the elementary school, while the strata consisted of four cities in the US with 24 schools randomly assigned to the experimental or control group within each city.

The stratified design has been used much less frequently than either the matched-pair or completely randomized design. However, for many studies it would seem to represent a sensible compromise between these two designs in that it provides at least some baseline control on factors thought to be related to outcome, while easing the practical difficulties of finding appropriate pair-matches and avoiding the special analytic challenges raised by the matched-pair design.

Sample Size Estimation

A quantitatively justified sample size calculation is almost universally regarded as a fundamental design feature of a properly controlled clinical trial. Yet, as noted above, methodologic reviews of cluster randomization trials have consistently shown that only

a small proportion of these studies have adopted a predetermined sample size based on formal considerations of statistical power. Moreover, some investigators have designed community intervention trials in which exactly one cluster has been assigned to the experimental group and one to the control group (e.g. [6] and [56]). Such trials invariably result in interpretational difficulties caused by the total confounding of two sources of variation: (i) the variation in response due to the effect of intervention and (ii) the natural variation that exists between communities.

There are several possible explanations for the difficulties investigators have faced in designing adequately powered studies. One obvious reason is that the required sample size formulas still tend to be relatively inaccessible, not being available, for example, in most standard texts or software packages. A second reason is that the proper use of these formulas requires some prior assessment of the intracluster correlation coefficient ρ, either directly or through comparable information on the value of σ_A^2, the between-cluster component of variation. Such information is not always easily available.

Difficulties in obtaining accurate estimates of intracluster correlation are slowly being addressed as more investigators begin publishing these values in the reporting of trial results (e.g. [68]). Summary tables listing intracluster correlation coefficients and variance inflation factors from a wide range of cluster randomization trials and complex surveys are also starting to appear (e.g. [20, Chapter 5]). In practice, estimates of ρ are almost always positive and tend to be larger in smaller clusters [37].

The calculation of sample size also requires the specification of the cluster size m. However, the actual size of the clusters randomized is frequently determined by the selected interventions. For example, households were the natural unit of randomization in the study reported by Payment et al. [59], which, as noted above, considered the effect of domestic water filters on subjects' risk of gastrointestinal disease. Consequently the average cluster size at entry in this trial was approximately four. When, on the other hand, relatively large clusters are randomized, subsamples of individual cluster members may be selected to reduce costs. For example, the end point for the community intervention trial COMMIT was the quit rate of approximately 550 heavy smokers selected from each cluster [13].

Equation (1) may be applied directly to determine the required sample size for a completely randomized design. Let $Z_{\alpha/2}$ denote the two-sided critical value of the standard normal distribution corresponding to the error rate α, and Z_β denote the critical value corresponding to β. Then, if $\overline{Y}_1 - \overline{Y}_2$ can be regarded as approximately normally distributed, then the number of subjects required per intervention group is given [21] by

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2(2\sigma^2)[1 + (m-1)\rho]}{(\mu_1 - \mu_2)^2}, \qquad (3)$$

and $\mu_1 - \mu_2$ denotes the magnitude of the difference to be detected. Equivalently the number of clusters required per group is given by $k = n/m$. With this allocation, the "effective sample size" for each group would be given by $n/[1 + (m-1)\rho]$. Thus at $\rho = 0$, (3) reduces to the usual sample size specification (e.g. [2, Section 6.6]).

The degree of variance inflation due to clustering can be profound even for very small values of ρ. For example, as part of a pilot study for a planned worksite intervention trial, the estimated intracluster correlation coefficient is given by $\hat{\rho} = 0.04$ [41]. Since approximately 70 subjects were eligible per worksite the required sample size is about four times that required for a comparable individually randomized trial.

In the case of unequal cluster sizes, we may replace m in (3) with the average cluster size $\overline{m}$. This approximation will tend to underestimate slightly the actual required sample size, but the underestimation will be negligible provided the variation in cluster size is not substantial. Further inaccuracy may result due to the inherent imprecision in the estimate of ρ. For example, the pilot study reported by Hsieh [41] included only four clusters, so that the estimated value of ρ computed from this study would be quite imprecise. It is therefore usually advisable to perform sensitivity analyses that explore the effect of different values of ρ on the estimated sample size.

Sensitivity analyses can also be useful for community intervention trials when assessing the effect of the number of clusters per intervention group and the subsample size on statistical efficiency. Such analyses will tend to reveal that greater gains in power will be obtained by increasing the number of clusters rather than the subsample size (see, for example, [41]), an easily demonstrated consequence of (1).

A similar approach may be used to determine sample size for completely randomized or stratified designs when the primary study endpoint (*see* **Outcome Measures in Clinical Trials**) is binary (see, for example, [18]). That is, the number of subjects required per intervention group may be calculated using standard sample size formulas applicable to individually randomized trials after multiplication by the inflation factor $IF = 1 + (m - 1)\rho$.

Unfortunately, the absence of appropriate measures of intracluster correlation for time to event and incidence rate data (see [66]) complicates sample size determination even for completely randomized designs. Alternative procedures for these outcomes have been described by Hayes & Bennett [38].

Difficulties in obtaining measures of intracluster correlation also complicate sample size determination for matched-pair designs. One simple alternative is to determine the sample size by assuming that the trial is completely randomized. Ignoring differences in degrees of freedom between the two designs, this approach will tend to be conservative [53]. Greater precision may be obtained when the effect of matching can be accurately estimated in advance, using, for example, the within-stratum component of variation $\sigma_{AM}^2 = \sigma_A^2(1 - \rho_M)$, where ρ_M measures the effectiveness of the matching. Unfortunately, except for a few notable exceptions (e.g. [29]) it can be quite difficult to match successfully. Donner & Klar [20, Chapter 3] cite examples of trials in which matching may even have resulted in a loss of efficiency.

Unit of Inference and Unit of Analysis

Many of the challenges of cluster randomization arise because inferences are frequently intended to apply at the individual level while randomization is at the cluster level. If inferences were intended to apply at the cluster level, implying that an analysis at the cluster level would be most appropriate, then the study could be regarded, at least with respect to sample size estimation and data analysis, as a standard clinical trial. For example, one of the secondary aims of the Child and Adolescent Trial for Cardiovascular Health (CATCH) was to assess the effect of training food service personnel on how to improve the dietary content of lunch menus. The resulting analyses of dietary content were then naturally conducted at the cluster (school) level.

Analyses are inevitably more complicated when data are available from individual study subjects where the investigator must account for the lack of statistical independence among observations within a cluster. An obvious method of simplifying the problem is to collapse the data in each cluster, followed by the construction of a meaningful summary measure, such as an average, which then serves as the unit of analysis. Standard statistical methods can then be directly applied to the collapsed measures. This removes the problem of nonindependence since the subsequent significance tests and confidence intervals would be based on the variation among cluster summary values rather than on variation among individuals.

An important special case arises in trials having a quantitative outcome variable when each cluster has a fixed number of subjects. In this case the test statistic obtained using the analysis of variance is algebraically identical to the test statistic obtained using a cluster-level analysis [40, 46]. Thus, the suggestion that is sometimes made that a cluster-level analysis intrinsically assumes $\rho = 1$ is misleading, since such an analysis can be efficiently conducted regardless of the value of ρ. It is important to note, however, that this equivalence between cluster-level and individual-level analyses, which holds exactly for quantitative outcome variables under balance, holds only approximately for other outcome variables (e.g. binary, time to event, count). A second implication of this algebraic identity is that concerns for the ecologic fallacy (e.g. [47]) cannot arise in the case of cluster-level **intention to treat analyses** since the assigned intervention is shared by all cluster members.

In practice, the number of subjects per cluster will tend to exhibit considerable variability, either by design or by subject attrition. Cluster-level analyses, which give equal weight to all clusters, may prove to be imprecise. However, the precision of appropriately weighted cluster-level analyses is asymptotically equivalent to individual-level analyses. If there is only a small number of clusters per intervention group, then the resulting imprecision in the estimation of these weights might even result in a loss of power (e.g. [67]). Moreover, the validity of approximate individual-level analyses may then become questionable. In this case it might be preferable to consider exact statistical inferences constructed at the cluster level based on the randomization distribution for

the selected experimental design (e.g. completely randomized, matched-pair, stratified).

We now present methods for the analysis of common study outcomes (e.g. binary, quantitative, count, time to event) typical of cluster randomization trials. The order and detail provided for each study outcome reflect the relative frequency with which these different types of outcome data are encountered in cluster randomization trials.

Analysis of Binary Outcomes

Methods for analyzing binary (dichotomous) outcome data in cluster randomization trials are not as well established as methods for analyzing quantitative outcomes. The analytic issues involved are complicated by the absence of a unique multivariate extension of the binomial distribution analogous to the multivariate normal distribution, and by the fact that there is no single approach that has uniformly superior properties.

It follows that there are several procedures that may be used to test $H_0 : P_1 = P_2$, where P_i is the true event rate for the ith intervention group, $i = 1, 2$. To make ideas concrete, the discussion will take place in the context of a completely randomized, school-based smoking prevention trial [55]. In this study, 12 schools were randomly assigned to each of four conditions including three experimental conditions and a control condition (existing curriculum). For purposes of illustration we are interested in comparing the effect of the Smoke Free Generation intervention with the existing curriculum on the proportion of children who report using smokeless tobacco after two years of follow-up.

The overall observed rates of tobacco use in the experimental and control groups are 0.043 (58/1341) and 0.062 (91/1479), respectively. Investigators frequently use inappropriate methods, such as the Pearson chi-square statistic, to test the effect of intervention in cluster randomization trials [70]. Application of this statistic to comparing the overall event rates in the two groups yields $\chi_P^2 = 4.69$ ($P = 0.03$). This result might be taken to imply that the use of smokeless tobacco among experimental group individuals is significantly reduced as compared with control group individuals. However, a fundamental assumption of this procedure is that the sample observations are statistically independent. This assumption is almost certainly violated here, since it is more reasonable to assume that responses taken on subjects within a school are more similar than responses taken on subjects in different schools, i.e. to assume that the intracluster correlation coefficient ρ is positive. This would imply that the computed P value of 0.03 is likely to be biased downward.

Donner & Donald [19] proposed an adjustment for Pearson's chi-square statistic that depends on computing correction factors for the effect of clustering. For clusters of fixed size m, the adjusted Pearson chi-square test statistic reduces to $\chi_P^2/[1 + (m - 1)\hat{\rho}]$, where $\hat{\rho}$ is the sample estimate of ρ.

For the data from the smoking prevention trial we have $\hat{\rho} = 0.01$, with group-specific correction factors $C_1 = 2.54$ and $C_2 = 2.60$. The adjusted chi-square statistic $\chi_A^2 = 1.83$ ($P = 0.18$), which is not statistically significant at any conventional level.

Similar conclusions can be reached using other simple modifications of the Pearson chi-square statistic, as in, for example, the ratio estimator approach described by Rao & Scott [61]. Cluster-level analyses based on the two-sample t-test comparing the mean event rates, or nonparametric alternatives (e.g. Fisher's two-sample permutation test, the Wilcoxon rank sum test) also provide essentially the same results for this example.

It is possible that the absence of a statistically significant effect of intervention in this example is due, at least in part, to chance imbalance on prognostically important baseline covariates (*see* **Covariate Imbalance, Adjustment for**). Such imbalance, of course, may arise in any randomized trial. However, for a given total number of individuals, the probability of a substantive imbalance will be higher in a trial randomizing clusters, owing to the smaller effective sample size. Provided the number of clusters is sufficiently large, multiple regression models may be used to account for such an imbalance, as well as to help increase the precision with which the intervention effect is estimated.

The effect of covariate adjustment on the estimated effect of intervention for a binary outcome variable may be explored using extensions of logistic regression adjusted for clustering. Two possible choices are the logistic–normal model and the generalized estimating equations (GEE) extension of logistic regression [50, 57]. An advantage of these extensions is that they may be used to account for an imbalance on individual-level (e.g. sex) as

well as cluster-level (percent male, cluster size) characteristics.

The logistic–normal model assumes that the logit transform of the probability of using smokeless tobacco follows a normal distribution across clusters. Likelihood ratio tests will have maximum power to detect the effects of intervention as statistically significant when such parametric model assumptions are satisfied.

In practice it may be difficult to ensure that the assumptions underlying the use of parametric models hold. We therefore limit attention here to the GEE approach, which has the advantage of not requiring specification of a fully parametric distribution. Two distinct strategies are available to adjust for the effect of clustering using the GEE approach. The first can be said to be model-based, since it requires the specification of a working correlation matrix that describes the pattern of correlation between responses of cluster members. For cluster randomization trials the simplest assumption to make is that the responses of cluster members are equally correlated, i.e. exchangeable. The second strategy to adjust for the effect of clustering employs "robust variance estimators" that are constructed using between-cluster information. These estimators consistently estimate the true variance even if the working correlation matrix is misspecified. Moreover, provided there are a large number of clusters, inferences obtained using robust variance estimators will become equivalent to those obtained using the model-based strategy when the working correlation matrix is correctly specified.

The sample odds ratio comparing the use of smokeless tobacco in the experimental vs. the control group may be obtained by fitting a GEE extension of logistic regression with a working exchangeable correlation matrix. The resulting odds ratio is given by 0.67 with a 95% robust confidence interval given by (0.41, 1.09). The corresponding one-degree-of-freedom chi-square test statistic is given by $\chi^2_{LZR} = 2.62$ ($P = 0.11$). Adjustment for subject age and sex results in a stronger observed effect of intervention, giving an adjusted odds ratio estimate of 0.65. As in earlier analyses, however, the association is not statistically significant.

The data from a matched-pair cluster randomization trial with a binary outcome variable and k strata may be summarized in a series of k 2×2 contingency tables. Methods used to evaluate the effect of intervention in matched-pair designs must rely on estimates of variance obtained using between-stratum information. Methods of analysis which may be used to test for an effect of intervention include a version of the one-sample permutation test as proposed by Liang [49] and a paired t-test as applied to the stratum-specific difference in event rates. Note that a naive application of the Mantel–Haenszel test statistic, χ^2_{MH}, is invalid since dependencies among responses of cluster members induce extra variability not accounted for in the test statistic. Thus, the true level of significance associated with χ^2_{MH} may be substantially greater than the nominal significance level if the clustering effect is ignored. These methods have been compared in the context of a hypertension screening and management trial by Donner [17].

Naive application of the Mantel–Haenszel test statistic is similarly inappropriate for analyses of data from stratified cluster randomization trials. The required generalization of the Mantel–Haenszel procedure is given by Donner [18].

Statistical analyses for the stratified design may also be conducted using GEE or other extensions of logistic regression. Unlike the adjusted Mantel–Haenszel procedure, the use of GEE in this context allows adjustment for individual-level as well as cluster-level covariates. However, these extensions of logistic regression are not directly applicable to the analyses of data from matched-pair designs unless the variance term is computed from between-stratum information and under the added assumption that there is no intervention by stratum interaction. A further limitation is that paired designs inevitably have fewer degrees of freedom for the estimation of error than either the completely randomized or stratified designs.

It is important to note, notwithstanding these difficulties, that if the aim of the analysis is limited to the adjustment or control of individual-level covariates in assessing the effect of intervention, then appropriate methods are readily available. These methods take the form of two-stage procedures based on standardizing the data with respect to individual-level covariates in advance of the primary analysis (e.g. [30]).

Analysis of Quantitative Outcomes

Analysis of quantitative outcome data from cluster randomization trials can often be accomplished using mixed-effects linear models [51, 74]. These models

can be most directly used with completely randomized or stratified designs to estimate the effect of intervention, to test if the observed effect is due to chance, and to adjust for imbalance on baseline prognostic factors.

For balanced designs having a fixed number of subjects per cluster, standard analysis of variance (ANOVA) may be used to test for the effect of intervention. As noted above, statistical inferences are then exact and identical to those obtained from cluster-level analyses (e.g. [46]). Of course, in practice the number of subjects per cluster may be highly variable. In this case iterative approaches, such as generalized least squares, are usually the method of choice for fitting mixed effects linear regression models, since the associated procedures provide maximum likelihood estimates of the effect of intervention [64, Section 6.8]. However, the resulting inferences will now be approximate rather than exact [23, Chapter 6], with no unique method for calculating the degrees of freedom [51, Chapter 2].

An interesting application of robust variance estimation is provided by Brook et al. [8] in their analyses of the Rand Health Insurance experiment [58]. In this study approximately 2000 families from the US were randomly assigned to one of 14 health insurance plans to evaluate the extent to which the provision of free care improves health. The analytic approach was based on a strategy suggested by Huber [42], a special case of the more general GEE procedure described by Liang & Zeger [50]. Further discussion is provided by Diggle et al. [16, p. 69].

An alternative to classical ANOVA or mixed effects linear regression is to apply a methodology often referred to in the behavioral and educational literature as multilevel modeling [35] or as hierarchical linear modeling [10]. Results obtained using any of these approaches are likely to be similar owing to their very close algebraic relationship (e.g. see [27]). Note that, in general, the relationship among the various extensions of logistic regression is more complicated, consistent with the greater challenge posed by analyses of binary outcome variables (e.g. [57]).

As noted above, the presence of only two clusters per stratum in the matched-pair design implies that estimates of between-cluster variability are totally confounded with the effect of intervention. Tests of the effect of intervention must therefore be calculated using between-stratum information on variability. One simple option is to apply the standard paired t-statistic to the stratum-specific mean differences. In the presence of obvious non-normality, reasonable alternatives are the Wilcoxon signed rank test or Fisher's one-sample permutation test.

These methods of analysis are illustrated by Donner & Klar [20, Chapter 7] using data from the British Family Heart Study [25]. The purpose of this trial was to examine the effect of a one-year, nurse-led lifestyle intervention on cardiovascular disease risk factors, with participating patients recruited from general practices.

Analysis of Count and Time to Event Outcomes

The primary outcome variable in most cluster randomization trials is either binary or continuous. However, this is not always the case. For example, Payment et al. [59], in a study involving count data, examined the effect of a domestic water filter on reducing the annual number of gastrointestinal episodes for each subject. Standard methods for the analysis of count data (e.g. [7]) were correctly recognized as being inappropriate, since each episode would then be counted as an independent event. Methods that allowed for correlation in the number of episodes per household were adopted instead.

Individual-level analyses for completely randomized designs for count data may be conducted using relatively simple procedures based on the ratio estimator approach [62]. More computer-intensive procedures based on extensions of Poisson regression models which adjust for clustering have also been described (e.g. see [69]).

Subjects in randomized trials are not always followed for the same length of time. Thus, variable follow-up times may occur because some subjects have the event of interest early in the trial while others survive to the end of the trial (i.e. are censored). This was the case in the community intervention trial reported by Sommer et al. [72], which examined the effect of vitamin A supplementation on childhood mortality. In such trials count data are often reported as the number of events that occur relative to the total follow-up time, yielding an incidence rate that accounts for the variable follow-up. Extensions of Poisson regression, which account for clustering, may again be used to assess the effect of intervention.

An alternative analytic approach is to view the primary outcome as a failure time random variable.

The effects of clustering may then be accounted for using, for example, extensions of exponential or Weibull failure time models [65]. A semi-parametric approach based on GEE extensions of the Cox model might also be considered [20, Chapter 8]. These regression models may be applied to the analysis of data from either completely randomized or stratified designs, provided a reasonably large number of clusters has been assigned to each intervention group.

Comparison of mortality experience across intervention groups can be graphically displayed using standard Kaplan–Meier (product–limit) survival curves. However, standard methods for obtaining corresponding variance estimators need to be adjusted for the impact of clustering, since otherwise the resulting statistical inferences will be biased (e.g. [78]).

As an example of a matched-pair design with time-to-event outcomes Ray et al. [63] reported on a randomized trial of a consultation service developed to help prevent falls and associated injuries in high-risk nursing home residents. Seven pairs of nursing homes were enrolled, matched by geographic proximity and the number of available beds. Time on study began when the program assessments were initiated in the nursing homes assigned to the intervention group, thus further matching the homes for calendar time. Study subjects were followed for up to one year after the index date, with observations censored for various reasons including death or discharge. One of the primary endpoints was the 12 month incidence rate of injurious falls (i.e. falls requiring medical attention) calculated separately for each of the 14 nursing homes. While the mean rate of injurious falls was lower for intervention facilities (13.7 falls per 100 person-years) than for control facilities (19.9 falls per 100 person-years) the difference was not statistically significant ($P = 0.22$) using a paired t-test. Similar conclusions were reached by Donner & Klar [20, Chapter 8] using exact permutation tests, as suggested by the results of simulation studies reported by Brookmeyer & Chen [9]. These authors also proposed a two-stage regression approach which may be used to adjust for imbalance on baseline covariates not accounted for by random assignment.

Reporting of Cluster Randomization Trials

Reporting standards for randomized clinical trials have now been widely disseminated (e.g. [5]). Many of the principles that apply to trials randomizing individuals also apply to trials randomizing intact clusters. These include a carefully posed justification for the trial, a clear statement of the study objectives, a detailed description of the planned intervention and the method of randomization and an accurate accounting of all subjects randomized to the trial. Unambiguous inclusion–exclusion criteria (*see* **Eligibility and Exclusion Criteria**) must also be formulated, although perhaps separately for cluster-level and individual-level characteristics. There are, however, some unique aspects of cluster randomization trials that require special attention at the reporting stage. We focus here on some of the most important of these. A more complete account is provided by Donner & Klar [20, Chapter 9].

The decreased statistical efficiency of cluster randomization relative to individual randomization can be substantial, depending on the sizes of the clusters randomized and the degree of intracluster correlation. Thus, unless there is obviously no alternative, the reasons for randomizing clusters rather than individuals should be clearly stated. This information, accompanied by a clear description of the units randomized, can help a reader decide if the loss of precision due to cluster randomization is in fact justified.

Having decided to randomize clusters, investigators may still have considerable latitude in their choice of unit. Since different levels of statistical efficiency are associated with different cluster sizes, it would seem important to select the unit of randomization on a carefully considered basis. However, it is apparent from published reports that this has not always been the case. For example, the review of cluster randomization studies reported by Donner et al. [22] found that only one-quarter of the trials considered provided reasons for their choice of randomization unit. An unambiguous definition of the unit of randomization is also required. For example, a statement that "neighborhoods" were randomized is clearly incomplete without a detailed description of this term in the context of the planned trial.

As noted previously, the consensus that exists in most clinical trial settings regarding the role of informed consent has not tended to apply to cluster randomization trials. By reporting the methods used (if any) to obtain informed consent in their own trials, it may gradually become possible for the research

community to develop reasonably uniform standards regarding this important issue.

The clusters that participate in a trial, simply owing to their consent to be randomized, may not be representative of the target population of clusters. Some indication of this lack of representativeness may be obtained by listing the number of clusters that met the eligibility criteria for the trial, but which declined to participate, along with a description of their characteristics.

A continuing difficulty with reports of cluster randomization trials is that justification for the sample size is all too often omitted. Investigators should clearly describe how the sample size for their trial was determined, with particular attention given to how clustering effects were adjusted for. This description should be in the context of the experimental design selected (e.g. completely randomized, matched-pair, stratified).

It would also be beneficial to the research community if empirical estimates of ρ were routinely published (with an indication of whether or not the reported values have been adjusted for the effect of baseline covariates).

It should be further specified what provisions, if any, were made in the sample size calculations to account for potential loss to follow-up. Since the factors leading to the loss to follow-up of individual members of a cluster may be very different from those leading to the loss of an entire cluster, both sets of factors must be considered here.

A large variety of methods, based on very different sets of assumptions, have been used to analyze data arising from cluster randomization trials. For example, possible choices for the analysis of binary outcomes include adjusted chi-square statistics, the GEE method and logistic–normal regression models. These methods are not as familiar as the standard procedures commonly used to analyze clinical trial data. This is partly because the methodology for analyzing cluster randomization trials is in a state of rapid development, with virtually no standardization and a proliferation of associated **software**. Therefore it is incumbent upon authors to provide a clear statement of the statistical methods used, and accompanied, where it is not obvious, by an explanation of how the analysis adjusts for the effect of clustering. The software used to implement these analyses should also be reported.

References

[1] Abdeljaber, M.H., Monto, A.S., Tilden, R.L., Schork, M.A. & Tarwotjo, I. (1991). The impact of vitamin A supplementation on morbidity: a randomized community intervention trial, *American Journal of Public Health* **81**, 1654–1656.

[2] Armitage, P. & Berry, G. (1994). *Statistical Methods in Medical Research*, 3rd Ed. Blackwell Scientific Publications, Oxford.

[3] Ashby, M., Neuhaus, J.M., Hauck, W.W., Bacchetti, P., Heilbron, D.C., Jewell, N.P., Segal, M.R. & Fusaro, R.E. (1992). An annotated bibliography of methods for analyzing correlated categorical data, *Statistics in Medicine* **11**, 67–99.

[4] Bass, M.J., McWhinney, I.R. & Donner, A. (1986). Do family physicians need medical assistants to detect and manage hypertension?, *Canadian Medical Association Journal* **134**, 1247–1255.

[5] Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. & Stroup, D.F. (1996). Improving the quality of reporting of randomized controlled trials, The CONSORT statement, *Journal of the American Medical Association* **276**, 637–639.

[6] Blum, D. & Feachem, R.G. (1983). Measuring the impact of water supply and sanitation investments on diarrhoeal diseases: problems of methodology, *International Journal of Epidemiology* **12**, 357–365.

[7] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Volume II: *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[8] Brook, R.H., Ware, J.E., Jr, Rogers, W.H., Keeler, E.B., Davies, A.R., Donald, C.A., Goldberg, G.A., Lohr, K.N., Masthay, P.C. & Newhouse, J.P. (1983). Does free care improve adults' health? Results from a randomized controlled trial, *New England Journal of Medicine* **309**, 1426–1434.

[9] Brookmeyer, R. & Chen, Y.-Q. (1998). Person-time analysis of paired community intervention trials when the number of communities is small, *Statistics in Medicine* **17**, 2121–2132.

[10] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Application and Data Analysis Methods*. Sage Publications, Newbury Park.

[11] Chalmers, T.C. & Schroeder, B. (1979). Controls in journal articles, *New England Journal of Medicine* **301**, 1293.

[12] Cochran, W.G. (1953). *Sampling Techniques*, 2nd Ed. Wiley, New York.

[13] COMMIT Research Group. (1995). Community Intervention Trial for Smoking Cessation (COMMIT): I.

Cohort results from a four-year community intervention, *American Journal of Public Health* **85**, 183–192.

[14] Comstock, G.W. (1978). Uncontrolled ruminations on modern controlled trials, *American Journal of Epidemiology* **108**, 81–84.

[15] Cornfield, J. (1978). Randomization by group: a formal analysis, *American Journal of Epidemiology* **108**, 100–102.

[16] Diggle, P.J., Liang, K.Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[17] Donner, A. (1987). Statistical methodology for paired cluster designs, *American Journal of Epidemiology* **126**, 972–979.

[18] Donner, A. (1998). Some aspects of the design and analysis of cluster randomization trials, *Applied Statistics* **47**, 95–114.

[19] Donner, A. & Donald, A. (1988). The statistical analysis of multiple binary measurements, *Journal of Chronic Diseases* **41**, 899–905.

[20] Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London.

[21] Donner, A., Birkett, N. & Buck, C. (1981). Randomization by cluster: sample size requirements and analysis, *American Journal of Epidemiology* **114**, 906–914.

[22] Donner, A., Brown, K.S. & Brasher, P. (1990). A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989, *International Journal of Epidemiology* **19**, 795–800.

[23] Dunn, O.J. & Clark, V.A. (1987). *Applied Statistics: Analysis of Variance and Regression*, 2nd Ed. Wiley, New York.

[24] Edwards, S.J.L., Braunholtz, D.A., Lilford, R.J. & Stevens, A.J. (1999). Ethical issues in the design and conduct of cluster randomized controlled trials, *British Medical Journal* **318**, 1407–1409.

[25] Family Heart Study Group (1994). The British Family Heart Study: its design and methods, and prevalence of cardiovascular risk factors, *British Journal of General Practice* **44**, 62–67.

[26] Ferebee, S.H., Mount, F.W., Murray, F.J. & Livesay, V.T. (1963). A controlled trial of isoniazid prophylaxis in mental institutions, *American Review of Respiratory Disease* **88**, 161–175.

[27] Ferron, J. (1997). Moving between hierarchical modeling notations, *Journal of Education and Behavioral Statistics* **22**, 119–123.

[28] Fontanet, A.L., Saba, J., Chandelying, V., Sakondhavat, C., Bhiraleus, P., Rugpao, S., Chongsomchai, C., Kiriwat, O., Tovanabutra, S., Dally, L., Lange, J.M. & Rojanapithayakorn, W. (1998). Protection against sexually transmitted diseases by granting sex workers in Thialand the choice of using the male or female condom: results from a randomized controlled trial, *AIDS* **12**, 1851–1859.

[29] Freedman, L.S., Gail, M.H., Green, S.B. & Corle, M.S. for the COMMIT Study Group (1997). The efficiency of the matched pairs design of the Community Intervention Trial for Smoking Cessation (COMMIT), *Controlled Clinical Trials* **18**, 131–139.

[30] Gail, M.H., Mark, S.D., Carroll, R.J., Green, S.B. & Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials, *Statistics in Medicine* **15**, 1069–1092.

[31] Ghana VAST Study Team (1993). Vitamin A supplementation in northern Ghana: effects on clinic attendances, hospital admissions, and child mortality, *Lancet* **342**, 7–12.

[32] Gillum, R.F., Williams, P.T. & Sondik, E. (1980). Some consideration for the planning of total-community prevention trials. When is sample size adequate?, *Journal of Community Health* **5**, 270–278.

[33] Glanz, K., Rimer, B.K. & Lerman, C. (1996). Ethical issues in the design and conduct of community-based intervention studies, in *Ethics and Epidemiology*, S.S. Coughlin & T.L. Beauchamp, eds. Oxford University Press, Oxford, Chapter 8.

[34] Glasgow, R.E., Lichtenstein, E., Wilder, D., Hall, R., McRae, S.G. & Liberty, B. (1995). The tribal tobacco policy project: working with Northwest Indian tribes on smoking policies, *Preventive Medicine* **24**, 434–440.

[35] Goldstein, H. (1995). *Multi-level Statistical Models*, 2nd Ed. Arnold, London.

[36] Grosskurth, H., Mosha, F., Todd, J., Mwijarubi, E., Klokke, A., Senkoro, K., Mayaud, P., Changalucha, J., Nicoll, A., ka-Gina, G., Newell, J., Mugeke, K., Mobey, D. & Hayes, R. (1995). Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomized controlled trial, *Lancet* **346**, 530–536.

[37] Hansen, M.H. & Hurwitz, W.N. (1942). Relative efficiencies of various sampling units in population inquiries, *Journal of the American Statistical Association* **37**, 89–94.

[38] Hayes, R.J. & Bennett, S. (1999). Simple sample size calculation for cluster-randomization trials, *International Journal of Epidemiology* **28**, 319–326.

[39] Herrera, M.G., Nestel, P., El Amin, A., Fawzi, W.W., Muhammad, K.A. & Weld, L. (1992). Vitamin A supplementation and child survival, *Lancet* **340**, 267–271.

[40] Hopkins, K.D. (1982). The unit of analysis: group means versus individual observations, *American Educational Research Journal* **19**, 5–18.

[41] Hsieh, F.Y. (1988). Sample size formulae for intervention studies with the cluster as unit of randomization, *Statistics in Medicine* **8**, 1195–1201.

[42] Huber, P.J. (1965). The behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, L.M. Lecam & J. Neyman, eds. University of California Press, Berkeley, pp. 221–233.

[43] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[44] Klar, N. & Donner, A. (1997). The merits of matching in community intervention trials, *Statistics in Medicine* **16**, 1753–1764.

[45] Koepsell, T.D. (1998). Epidemiologic issues in the design of community intervention trials, in *Applied Epidemiology, Theory to Practice*, R.C. Brownson & D.B. Petitti, eds. Oxford University Press, New York, Chapter 6, pp. 177–211.

[46] Koepsell, T.D., Martin, D.C., Diehr, P.H., Psaty, B.M., Wagner, E.H., Perrin, E.B. & Cheadle, A. (1991). Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach, *Journal of Clinical Epidemiology* **44**, 701–713.

[47] Kreft, I.G.G. (1998). An illustration of item homogeneity scaling and multilevel analysis techniques in the evaluation of drug prevention programs, *Evaluation Review* **22**, 46–77.

[48] Lasater, T.M., Becker, D.M., Hill, M.N. & Gans, K.M. (1997). Synthesis of findings and issues from religious-based cardiovascular disease prevention trials, *Annals of Epidemiology* **S7**, S46–S53.

[49] Liang, K.-Y. (1985). Odds ratio inference with dependent data, *Biometrika* **72**, 678–682.

[50] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[51] Littell, R.C., Milliken, R.C., Stroup, G.A., Walter, W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute Inc., Cary.

[52] Luepker, R.V., Perry, C.L., McKinlay, S.M., Nader, P.R., Parcel, G.S., Stone, E.J., Feldman, H.A., Johnson, C.C., Kelder, S.H. & Wu, M. for the CATCH Collaborative Group (1996). Outcomes of a field trial to improve children's dietary patterns and physical activity, *Journal of the American Medical Association* **275**, 768–776.

[53] Martin, D.C., Diehr, P., Perrin, E.B. & Koepsell, T.D. (1993). The effect of matching on the power of randomized community intervention studies, *Statistics in Medicine* **12**, 329–338.

[54] Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis, *British Medical Journal* **ii**, 769–782.

[55] Murray, D.M., Perry, C.L., Griffin, G., Harty, K.C., Jacobs, D.R. Jr, Schmid, L., Daly, K. & Pallonen, U. (1992). Results from a statewide approach to adolescent tobacco use prevention, *Preventive Medicine* **21**, 449–472.

[56] Murray, J.P., Stam, A. & Lastovicka, J.L. (1993). Evaluating an anti-drinking and driving advertising campaign with a sample-survey and time series intervention analysis, *Journal of the American Statistical Association* **88**, 50–56.

[57] Neuhaus, J.M. (1992). Statistical methods for longitudinal and clustered designs with binary responses, *Statistical Methods in Medical Research* **1**, 249–273.

[58] Newhouse, J.P. & the Insurance Experiment Group (1993). *Free for All? Lessons from the Rand Health Insurance Experiment: A Rand Study*. Harvard University Press, Cambridge.

[59] Payment, P., Richardson, L., Siemiatycki, J., Dewar, R., Edwardes, M. & Franco, E. (1991). A randomized trial to evaluate the risk of gastrointestinal disease due to consumption of drinking water meeting microbiological standards, *American Journal of Public Health* **81**, 703–708.

[60] Pocock, S.J. (1983). *Clinical Trials. A Practical Approach*. Wiley, New York, pp. 17–18.

[61] Rao, J.N.K. & Scott, A.J. (1992). A simple method for the analysis of clustered binary data, *Biometrics* **48**, 577–585.

[62] Rao, J.N.K. & Scott, A.J. (1999). A simple method for analysing overdispersion in clustered Poisson data, *Statistics in Medicine* **18**, 1373–1385.

[63] Ray, W.A., Taylor, J.A., Meador, K.G., Thapa, P.B., Brown, A.K., Kajihara, H.K., Davis, C., Gideon, P. & Griffin, M.R. (1997). A randomized trial of a consultation service to reduce falls in nursing homes, *Journal of the American Medical Association* **278**, 557–562.

[64] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

[65] Segal, M.R. & Neuhaus, J.M. (1993). Robust inference for multivariate survival data, *Statistics in Medicine* **12**, 1019–1031.

[66] Segal, M.R., Neuhaus, J.M. & James, I.R. (1997). Dependence estimation for marginal models of multivariate survival data, *Lifetime Data Analysis* **3**, 251–268.

[67] Shao, J. (1990). Ordinary and weighted least-squares estimators, *Canadian Journal of Statistics* **18**, 327–336.

[68] Siddiqui, O., Hedeker, D., Flay, B. & Hu, F.B. (1996). Intraclass correlation in a school-based smoking prevention study. Outcome and mediating variables, by sex and ethnicity, *American Journal of Epidemiology* **144**, 425–433.

[69] Siddiqui, O., Mott, J., Anderson, T. & Flay, B. (1999). The application of Poisson random-effects regression models to the analyses of adolescents' current level of smoking, *Preventive Medicine* **29**, 91–101.

[70] Simpson, J.M., Klar, N. & Donner, A. (1995). Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993, *American Journal of Public Health* **85**, 1378–1382.

[71] Smith, P.J., Moffatt, M.E.K., Gelskey, S.C., Hudson, S. & Kaita, K. (1997). Are community health interventions evaluated appropriately? A review of six journals, *Journal of Clinical Epidemiology* **50**, 137–146.

[72] Sommer, A., Tarwotjo, I., Djunaedi, E., West, K.P. Jr, Loeden, A.A., Tilden, M.L. & the ACEH Study Group. (1986). Impact of vitamin A supplementation on childhood mortality, *Lancet*, 24 May, 1169–1173.

[73] Strasser, T., Jeanneret, O. & Raymond, L. (1987). Ethical aspects of prevention trials, in *Ethical Dilemmas in Health Promotion*, S. Doxiadis, ed. Wiley, New York, Chapter 15.

[74] Verbeke, G. (1997). Linear mixed models for longitudinal data, in *Linear Mixed Models in Practice, A SAS-Oriented Approach*, G. Verbeke & G. Molenberghs, eds. Springer-Verlag, New York, Chapter 3.

[75] Walsh, J.E. (1947). Concerning the effect of intraclass correlation on certain significance tests, *Annals of Mathematical Statistics* **18**, 88–96.

[76] Walsh, M.W., Hilton, J.F., Masouredis, C.M., Gee, L., Chesney, M.A. & Ernster, V.L. (1999). Spit tobacco cessation intervention for college athletes: results after one year, *American Journal of Public Health* **89**, 228–234.

[77] Ware, J.H. & Liang, K.Y. (1996). The design and analysis of longitudinal studies: a historical perspective, in *Advances in Biometry*, P. Armitage & H.A. David, eds. Wiley, New York, Chapter 17, pp. 339–362.

[78] Williams, R.L. (1995). Product-limit survival functions with correlated survival times, *Data Analysis* **1**, 171–186.

[79] World Medical Association (1997). World Medical Association Declaration of Helsinki, 1996. Republished in the *Journal of the American Medical Association* **277**, 925–926.

NEIL KLAR & ALLAN DONNER

Cluster Sampling, Optimal Allocation

Optimal allocation refers to the use of a survey sampling project's resources in a manner that either minimizes sampling **variance** for a fixed cost or minimizes cost to obtain a fixed variance. When a survey has several objectives or population parameters to be estimated, the sampling statistician needs to consider the designs that are optimal for each objective. He or she should then present the alternatives to the client and provide statistical guidance in making an informed choice.

Sampling units may be allocated optimally either across strata (*see* **Stratified Sampling, Allocation in**) or by varying the cluster structure of the population. This article discusses the optimal number of sample clusters and sample units (or *elements*) within clusters [2, 3].

The simplest case is for a one-stage cluster sample (*see* **Cluster Sampling**), where each cluster contains the same number of elements, where we select a **simple random sample** of the clusters; and where we select and measure each unit within each of the sample clusters. An example of this scenario is in area sampling, where we are trying to determine the optimum number of units to group together to form a *chunk* or *segment*. The chunks or segments are the clusters and the number of clusters in a given arrangement of the population is denoted M_a. The number of units in the ith cluster, $i = 1, \ldots, M_a$, is N_{ai} ($= N_a$ for all i since the number of units or elements within clusters is assumed to be the same for all clusters). The total number of units in the population is $N = \sum_{i=1}^{M_a} N_{ai} = M_a N_a$. The sample consists of m_a clusters selected at random, which yields a total of $n_a = \sum_{i=1}^{m_a} N_a = m_a N_a$ sample elements. Suppose we wish to estimate (*see* **Estimation**) the **mean** over all elements of a characteristic, X, which takes on the values x_{ij} for the jth unit in the ith cluster, $j = 1, \ldots N_a$ and $i = 1, \ldots, M_a$. An **unbiased** estimate is $\overline{\overline{x}}_a = \sum_{i=1}^{m_a} \sum_{j=1}^{N_a} x_{ij} \big/ m_a N_a$, which has **variance**

$$\mathrm{var}(\overline{\overline{x}}_a) = \frac{S_{a1}^2}{m_a} \left(\frac{M_a - m_a}{M_a} \right)^2,$$

where S_{a1} is the **standard deviation** of the distribution of cluster means. That is,

$$S_{a1} = \left[\frac{N_a \sum_{i=1}^{M_a} (\overline{X}_{ai} - \overline{\overline{X}})^2}{M_a - 1} \right]^{1/2},$$

$$\overline{X}_{ai} = \frac{\sum_{j=1}^{N_a} X_{ij}}{N_a},$$

and

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{M_a} \sum_{j=1}^{N_a} X_{ij}}{N}.$$

The precision of the estimate depends on the relationship between S_{a1} and m_a. For estimating the mean per unit of a characteristic, the optimum arrangement of the population into clusters can be determined [1, p. 234] from the fact that both the relative cost for a fixed variance, $\mathrm{var}(\overline{\overline{x}})$, and the relative variance for a fixed cost, C, are proportional to $S_{a1}^2 C_a$, where C_a is the relative cost of measuring a cluster in the ath allocation scheme.

The usual procedure for determining N_a is to examine several allocation schemes and select the one that minimizes variance for a fixed cost. For example, suppose the cost to drive to a cluster of any size averages about \$10. Once at the cluster, it may take 5 minutes to measure a unit. If the measurement costs \$15 per hour, then we can measure 12 units in an hour at a cost of \$1.25 per unit. We may want to decide between a plan with 60 units per cluster and one with 30 units per cluster. Suppose that previous surveys have shown the large clusters to have a standard deviation of 10 across clusters and the smaller clusters have a standard deviation of 15. This is the standard deviation of the mean unit characteristic across clusters $S_{60\mathrm{B}}$ and $S_{30\mathrm{B}}$, respectively. The two costs are:

$$C_{30} = \$10 + 30(\$1.25) = \$47.5 \text{ per cluster},$$

and

$$C_{60} = \$10 + 60(\$1.25) = \$85 \text{ per cluster}.$$

The relative variances for a fixed cost are:

$$\text{var}_{30} \propto 47.5 \times 15^2 = 10\,687.5$$

and

$$\text{var}_{60} \propto 85 \times 10^2 = 8500.$$

For this characteristic, the larger cluster would be preferred. The process is then repeated for each survey objective and the client then needs to exercise judgment as to which allocation best fits the project's overall goals.

The next level of complication is to consider clusters of unequal size, N_{ai}. Let $\overline{N}_a = \sum_{i=1}^{M_a} N_{ai} \big/ M_a$ be the average cluster size. An estimator of the mean per unit is $\overline{\overline{x}}_a = \sum_{i=1}^{m_a} \sum_{j=1}^{N_{ai}} x_{ij} \big/ m_a \overline{N}_a$. One approach is to treat the units as if they have approximately equal size by replacing N_{ai} by $\overline{N}_a$ and proceeding as if the cluster sizes were equal. In some circumstances the clusters may be stratified to increase the homogeneity of size.

Another complication occurs when we observe that units within clusters have similar values for the characteristic of interest. This homogeneity of units within clusters implies that measuring all the units within a cluster is very unlikely to provide as much information as obtaining measurements from a few units selected from several clusters. This leads to a two-step or two-stage sampling design (*see* **Multistage Sampling**).

For a two-stage design, m_a clusters are selected at random. Within each selected cluster a sample of n_{ai} units, $i = 1, \ldots, m_a$, are measured. As with simple one-stage cluster sampling there are M_a clusters each containing N_a units. The subscript a indexes the particular allocation scheme. A complete sample contains $n_a = \sum_{i=1}^{m_a} n_{ai}$ units. It is reasonable to fix $n_{ai} = \overline{n}_a$ since the clusters are of equal size. The estimator of the population mean per unit is

$$\overline{\overline{x}}_a = \frac{\sum_{i=1}^{m_a} \sum_{j=1}^{\overline{n}_a} x_{ij}}{m_a \overline{n}_a}.$$

This estimator is unbiased [1, Theorem 10.1, p. 277] and has **standard error**:

$$\text{se}(\overline{\overline{x}}_a) = \left[\left(\frac{M_a - m_a}{M_a} \right) \frac{S_{a1}^2}{m_a} + \left(\frac{N_a - \overline{n}_a}{N_a} \right) \frac{S_{a2}^2}{m_a \overline{n}_a} \right]^{1/2},$$

where

$$S_{a1}^2 = \frac{\sum_{i=1}^{M_a} \left(\overline{X}_{ai} - \overline{\overline{X}} \right)^2}{M_a - 1}$$

and

$$S_{a2}^2 = \frac{\sum_{i=1}^{M_a} \sum_{j=1}^{\overline{N}_a} (X_{ij} - \overline{X}_{ai})^2}{\overline{N}_a - 1}.$$

These two variances have the following practical interpretation:

1. S_{a1} is the between-cluster standard deviation of the cluster means; thus, as clusters become larger this should become smaller.
2. S_{a2} is the within-cluster standard deviation; thus, as the clusters become larger this should become larger.

Optimal allocation means altering the cluster size so as to balance these two sources of variation. The balance is affected by the cost of preparing a cluster for subsampling and the cost of measuring units within the cluster.

Many different cost functions have been examined, but that proposed earlier provides an appropriate model for many situations. Let the total cost depend only on the cost of preparing a cluster, c_1, and the cost of measuring a unit, c_2. The relative survey cost for a particular allocation scheme is then $c_a = c_1 m_a + c_2 m_a \overline{n}_a$. Following **Cochran** [1, p. 280], to find the choice of sample cluster size, $\overline{n}_a$, that optimizes the survey, we write the variance as

$$\text{var}(\overline{\overline{x}}_a) = \frac{1}{m_a} \left(S_{a1}^2 - \frac{S_{a2}^2}{N_a} \right) + \frac{S_{a2}^2}{m_a n_a} - \frac{S_{a1}^2}{M_a}.$$

It follows that the problem reduces to minimizing the following product:

$$C \left(\text{var}(\overline{\overline{x}}_a) + \frac{S_{a1}^2}{M_a} \right) = \left(S_{a1}^2 - \frac{S_{a2}^2}{N_a} + \frac{S_{a2}^2}{n_a} \right) (c_1 + c_2 \overline{n}_a).$$

Using Lagrange multipliers and differentiating we obtain:

$$\overline{n}_{a,\text{opt}} = \frac{S_{a2}}{S_{au}} \left(\frac{c_1}{c_2} \right)^{1/2},$$

where $S_{au}^2 = S_{a1}^2 - (S_{a2}^2/N_a)$ and we assume $S_{au}^2 > 0$. Thus the within-cluster sample should be increased

as the within-cluster standard deviation increases and as the cost of preparing a cluster increases. S_{au} is the between-cluster standard deviation reduced by the relative within-cluster standard deviation. The optimum varies inversely with this term.

Suppose, in our earlier example, we learn that the size 30 segments yield $S_{30,2} = 1$ and $S_{60,2} = 4$. Then

$$S_{30,u} = \left(15^2 - \frac{1^2}{30}\right)^{1/2} \approx 15$$

and

$$S_{60,u} = \left(10^2 - \frac{4^2}{60}\right)^{1/2} \approx 10.$$

The optimum within cluster sizes are

$$n_{30,\text{opt}} = \frac{1}{15}\left(\frac{10}{1.25}\right)^{1/2} = 0.189$$

(or one unit per cluster) and

$$n_{60,\text{opt}} = \frac{4}{10}\left(\frac{15}{1.25}\right)^{1/2} = 1.39$$

or two units per cluster. Solving the cost model for m_a allows us to evaluate the alternative sample sizes:

$$m_{30} = \frac{C}{(10 + 1.25 \times 1)} = C \times 0.0889$$

and

$$m_{60} = \frac{C}{(10 + 1.25 \times 2)} = C \times 0.08.$$

The greatest reduction in variance for fixed cost occurs by maximizing m_a, so we would select a design with 30 units per cluster and select one unit per cluster at random.

The most complex situation for two-stage sampling is when the first-stage units (clusters) are of widely varying size. As a first step, the population of clusters can be stratified by size so as to create more homogeneous clusters (*see* **Stratified Sampling**). These are then sampled as above. Beyond this the allocation problem becomes quite complex. For additional discussion, the relevant material in Cochran [1] is quite complete.

References

[1] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[2] Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample Survey Methods and Theory*. Wiley, New York.

[3] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

DANIEL H. FREEMAN, JR

Cluster Sampling

The term *cluster sampling* has been used to categorize sampling designs in which the sampling units are groups (or *clusters*) of enumeration units and there is only one stage of sampling [2, 3]. In other words, cluster sampling involves taking a sample of clusters according to some sampling plan and then selecting every enumeration unit within each cluster that was sampled. In contrast, designs that sample clusters of enumeration units but involve more than one stage of sampling are generally referred to as **multistage sampling**. The term cluster sampling, as used here, is synonymous with terms such as *single-stage cluster sampling* or *one-stage cluster sampling* that have been used widely in the past and may still be seen occasionally in the literature.

The use of cluster sampling is motivated primarily by reasons that involve feasibility and economy. In sampling of human populations, for example, where estimates are desired on a per-person basis, it is not usually possible to construct a **sampling frame** that lists all individuals or even all households in the population. There may, however, be a list of city blocks or other geographical entities that could serve for sampling purposes as the cluster, and individual households (and ultimately individual persons) within each sample cluster can then be enumerated or "listed" and information obtained from them. Even when a frame can be constructed that consists of lists of individual households, it is often more economical to use these larger geographical entities as the sampling unit, especially when the **target population** is dispersed over a wide geographical area.

Our major objective here is to acquaint the reader with the basic concepts and formulation of cluster sampling. Our emphasis will be on the various types of **estimation** procedures and on issues related to the cost–effectiveness of cluster sampling. In the ensuing discussion, we will assume for the sake of simplicity that the enumeration units are also the elementary units.

Terminology

Let us suppose that a population contains N elementary units grouped into M clusters, and that each cluster contains N_i enumeration units ($\sum_{i=1}^{M} N_i = N$); that we are interested in estimating the level of some characteristic χ in this population; and that X_{ij} is the value of χ for enumeration unit j in the ith cluster. Some population parameters are given below for characteristic χ.

Population total for cluster i:

$$X_i = \sum_{j=1}^{N_i} X_{ij}.$$

Overall population total:

$$X = \sum_{i=1}^{M} X_i.$$

Population **mean** for cluster i:

$$\overline{X}_i = \frac{X_i}{N_i}.$$

Overall population mean per cluster:

$$\overline{X} = \frac{X}{M}.$$

Overall population mean per enumeration unit:

$$\overline{\overline{X}} = \frac{X}{N}.$$

Average number of elements per cluster:

$$\overline{N} = \frac{N}{M}.$$

Variance of distribution of cluster totals, X_i over all clusters:

$$\sigma_{1x}^2 = \frac{\sum_{i=1}^{M} (X_i - \overline{X})^2}{M}.$$

Let us also suppose that we are taking a sample of m clusters according to some sample design and, since this is cluster sampling, selecting all enumeration units within each sample cluster. Also, to simplify notation, let us assume that the clusters labeled $1, 2, \ldots, m$ are the sample clusters. The total number, n, of sample enumeration units is equal to $n = \sum_{i=1}^{m} N_i$, and the analogous sample statistics for characteristic χ are shown below.

Sample total for cluster i:

$$x_i = \sum_{j=1}^{N_i} x_{ij}.$$

Overall sample total:

$$x = \sum_{i=1}^{m} x_i.$$

Sample mean per cluster:

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i}{m}.$$

Sample variance of distribution of x_i over all sample clusters:

$$s_{1x}^2 = \frac{\sum_{i=1}^{m} (x_i - \overline{x})^2}{m-1}.$$

One of the characteristics of cluster sampling is that the formulas for estimates and their **standard errors** have a very similar appearance as the estimates for sample designs in which the enumeration units are sampled directly. The major difference is that the cluster totals, x_i, replace the individual values, x_{ij}. We illustrate this below for the scenario in which the clusters are sampled by **simple random sampling**.

Estimation Under Simple Random Sampling of Clusters

In cluster sampling, if simple random sampling is used to select clusters, the design is called *simple cluster sampling* or *simple one-stage cluster sampling*. Under simple cluster sampling, linear estimators such as means and totals are **unbiased** estimators of the corresponding population parameters as shown below. These estimators are shown below.

Estimated population mean per cluster:

$$\overline{x}_{\text{clu}} = \frac{\sum_{i=1}^{m} x_i}{m}. \tag{1}$$

Estimated standard error of $\overline{x}_{\text{clu}}$:

$$\widehat{SE}(\overline{x}_{\text{clu}}) = \left(\frac{1}{\sqrt{m}}\right) s_{1x} \left(\frac{M-m}{M}\right)^{1/2}. \tag{2}$$

Estimated population mean per enumeration unit:

$$\overline{\overline{x}}_{\text{clu}} = \frac{\sum_{i=1}^{m} x_i}{m\overline{N}}. \tag{3}$$

Estimated standard error of $\overline{\overline{x}}_{\text{clu}}$:

$$\widehat{SE}(\overline{\overline{x}}_{\text{clu}}) = \left(\frac{1}{\sqrt{m\overline{N}}}\right) s_{1x} \left(\frac{M-m}{M}\right)^{1/2}. \tag{4}$$

Estimated total:

$$x'_{\text{clu}} = \frac{M}{m} x. \tag{5}$$

Estimated standard error of estimated total:

$$\widehat{SE}(x'_{\text{clu}}) = \frac{M}{\sqrt{m}} s_{1x} \left(\frac{M-m}{M}\right)^{1/2}. \tag{6}$$

One can see that the above equations have the same form as those appropriate for simple random sampling of enumeration units with the cluster totals, x_i replacing the enumeration unit totals, x_{ij}, and the number, M, of clusters replacing the number, N, of enumeration units.

Illustrative Example

Suppose that a health insurance company conducts an audit of the claims reimbursed to a provider of medical care services over a calendar year. The provider was reimbursed for 14 claims submitted on behalf of 49 patients. Since the audit involves detailed examination of patient medical files, economy dictates that a simple cluster sample of patients be taken (implying that all claims reimbursed on behalf of each sampled patient be audited). The audit involved taking a cluster sample of $m = 7$ patients from the population of $M = 49$ patients on behalf of whom reimbursements were made. Table 1 shows the data obtained from this audit.

From the data on overpayment shown in the last column of Table 1, we obtain the following summary statistics:

$$x = \sum_{i=1}^{7} x_i = \$321.50,$$

$$s_{1x} = (\$3739.04)^{1/2} = \$61.15,$$

$$x'_{\text{clu}} = \frac{49}{7}(\$321.50) = \$2250.50,$$

Table 1 Findings of audit on claims made on behalf of seven patients

Patient	Claim ($)	Total reimbursed ($)	Overpayment ($)	Total overpayment per patient (x_i) ($)
1	1	25.00	15.00	15.00
2	1	37.50	18.00	
	2	75.00	26.00	
	3	235.00	95.00	139.00
3	1	87.00	28.00	28.00
4	1	24.00	0.00	
	2	87.00	10.00	10.00
5	1	145.00	49.00	
	2	123.00	40.50	
	3	89.00	40.00	129.50
6	1	37.00	0.00	0.00
7	1	167.00	0.00	
	2	193.00	0.00	
	3	12.00	0.00	0.00

$$\widehat{SE}(x'_{\text{clu}}) = \frac{49}{\sqrt{7}}(\$61.15)\left(\frac{49-7}{49}\right)^{1/2} = \$1048.46.$$

It should be noted that it is not necessary to know the total number, N, of enumeration units in order to perform a simple cluster sample. All that is required is identification of the clusters. Once the sample clusters are chosen, the enumeration units, if not already identified, can be identified as part of the field work, but this need only be done for the clusters actually sampled. Although the formulas (3) and (4), which estimate the mean per enumeration unit and its standard error, require knowledge of the total, N, of enumeration units in the population, an alternative estimator is available that does not require knowledge of N [3]. This alternative estimator is a **ratio estimator** in which the denominator is an estimate obtained from the sample of the total number of enumeration units in the population.

Precision of Estimates from Cluster Sampling

In practice, estimates obtained from cluster sampling often have higher standard errors than those that would have been obtained from a simple random sample of the same number, n, of enumeration units. This is because the variability among units *within* clusters with respect to a characteristic of interest in the survey may be considerably less than that among units in the population as a whole. For example, in a sample survey of a large city with the city blocks serving as clusters, there might be considerable homogeneity among residents of the same block with respect to characteristics such as income level, education level, ethnicity, etc., even though the city, when considered as a whole, has considerable diversity with respect to such characteristics. Since the process of cluster sampling involves sampling every enumeration unit within each sample cluster, there can be considerable "redundancy" of information. Another scenario might arrive in a manufacturing environment in which products are manufactured in "batches" with each batch containing $\overline{N}$ individual products. If the process is such that either all individuals in a particular batch are defective or none is defective, then a cluster sample of batches would be inefficient because it would entail sampling (unnecessarily) every product within each sample batch.

The ratio of the variance of an estimator under any sample design to that at comparable n under simple random sampling is known as its **design effect** and this index is generally used as an indicator of its precision (or lack thereof).

For cluster sampling, the design effect, $DEFF_{\text{clu}}(x'_{\text{clu}})$, of an estimated total is given by

$$DEFF_{\text{clu}}(x'_{\text{clu}}) = \left(\frac{1}{\overline{N}}\right)\left(\frac{\sigma^2_{1x}}{\sigma^2_x}\right)\left(\frac{N-1}{N-\overline{N}}\right). \quad (7)$$

This implies that if the variance, σ^2_{1x}, among cluster totals is large compared with the overall variability, σ^2_x, among enumeration units in the population, then estimated totals from cluster sampling will have high variances relative to those at the same n from simple

random sampling. For the example shown above, we have $\hat{\sigma}_{1x}^2 = 3662.73$, $\hat{\sigma}_x^2 = 715.11$, $\hat{N} = 98$, and $M = 49$, which from relation (7) gives a design effect equal to 2.587. This implies that an estimated total from a simple cluster sample will have a variance that is approximately 2.6 times as high as one that would be obtained from a simple random sample of the same number n of enumeration units.

Survey Costs

Although the variance of an estimate from cluster sampling will invariably be higher at the same sample size, n, than one obtained from simple random sampling, the cost of taking a cluster sample that would yield $n(= m\overline{n})$ enumeration units might be considerably lower than that of taking a simple random sample of n enumeration units. Thus, it is not particularly useful to compare cluster sampling with simple random sampling at the same sample size, n. A more appropriate basis for comparing cluster sampling with simple random sampling (or for that matter with any alternative design) would be at equivalent cost. One can do this by determining the number, m, of clusters that can be sampled at some fixed cost, C, and then determining the number, n, of enumeration units that can be sampled at this same cost, C. The design effect (at fixed cost rather than fixed sample size) would be the ratio of the variance of the estimate from a cluster sample of these m clusters to that of the estimate obtained from a simple random sample of the n enumeration units. Note that in this instance the n described above for the simple random sample is *not equal* to the entity $m\overline{N}$ that would be the total number of enumeration units obtained in the cluster sample.

Illustrative Example

Let us consider the example discussed earlier involving an audit of claims reimbursed to a provider of medical care services over a calendar year. The field costs involved in the survey involve: (i) construction of the sampling frame; (ii) abstraction of the required data; and (iii) processing of the data.

Let us suppose that the insurance company has made payments for $N = 98$ services that the provider made to $M = 49$ patients that calendar year, and that the names of these patients are in the insurance company's database. Let us suppose further that the list of individual claims is not computerized and that some clerical work is need to identify the individual claims, review the medical record, and abstract the data. We list the following cost components associated with these operations:

1. *Identify and list the individual claim*. Let us suppose that it takes 0.50 person hours to identify and list each individual claim.
2. *Abstract and process data from the patient medical record pertaining to each claim*. Let us suppose that it takes on the average 1.25 person hours to review, abstract, and process the relevant data.

From the above assumptions, we can now list the field costs associated with simple random sampling and with cluster sampling.

For simple random sampling, all 98 claims must be listed no matter what the size of the sample will be, and this will take 49.0 person hours. The abstraction and processing costs will be $1.25 \times n$, where n is the size of the sample to be taken. Thus, the total field costs, C, for a simple random sample of n enumeration units is given by the following:

$$C = 0.50 \times N + 1.25 \times n = 49 + 1.25 \times n. \quad (8)$$

For cluster sampling, all claims need to be listed for the m patients selected in the sample, but no listing has to be done for the $M - m$ patients not selected in the sampling. If m clusters are sampled, then the total field costs would be equal to

$$C = 0.50 \times m\overline{N} + 1.25 \times m\overline{N} = 3.50m \quad (9)$$

(since $\overline{N} = 2$). At a cost of 64 person-hours, a simple random sample of 12 claims can be taken based on the cost function for simple random sampling shown in relation (8). At the same cost, we can take a cluster sample of 18 patients (relation 9) which would give us an expected sample of 36 claims. From (6), using the estimated value of σ_{1x} computed earlier, we see that the estimated total overpaid based on a simple cluster sample of $m = 18$ patients would have an estimated standard error equal to \$555.96, whereas that based on a simple random sample of 12 claims would have an estimated standard error of \$708.67. Thus, in this example, at equivalent cost of 64 person-hours, cluster sampling would yield a more reliable estimate than simple random sampling.

The major reason that cluster sampling often produces more reliable estimates at equivalent cost than simple random sampling is that the enumeration units need to be listed only for the clusters that were selected in the sample, whereas in simple random sampling, all enumeration units in the population need to be listed before sampling can take place. If listing costs are high, this can put simple random sampling at a great disadvantage over cluster sampling.

For the purposes of illustration, the above discussion oversimplifies the costs associated with simple cluster sampling and simple random sampling. The specification of relevant cost components is a complex undertaking and requires considerable experience with the operations involved in the conduct of surveys. Groves [1] treats this topic at great length and identifies a number of important references.

Usefulness of Cluster Sampling

Cluster sampling is most useful when the homogeneity among enumeration units within clusters with respect to the levels of the variables being measured in the sample survey is no greater than the homogeneity among listing units in the population as a whole. In such instances, cluster sampling can result in considerable reductions in frame construction and travel costs without resulting in increased sampling errors. This, however, rarely occurs in practice since enumeration units within clusters generally show considerably more homogeneity than is reflected in the population as a whole. Thus, sampling every enumeration unit within a sample cluster (as is prescribed in cluster sampling) results in a considerable amount of redundancy. In practice, this is avoided by using multistage sampling instead of cluster sampling. In multistage designs, within each sample cluster, a subsample of enumeration units is taken rather than every enumeration, and this reduces the redundancy of the sampling. Thus, multistage sampling is a much more widely used procedure than cluster sampling.

Sometimes, however, it is not feasible to subsample within clusters. Levy et al. [4] describes such a situation that occurred in the planning of a sample survey in Shanghai, China, in 1986. The target population was persons 55 years of age and older residing in a District within Shanghai, and the clusters were administrative entities called *neighborhood groups* consisting of contiguous households that are organized for political and social purposes. For each neighborhood group, a listing of households with the names and ages of the members is available from the government.

Because the neighborhood groups consist of contiguous households, it was felt that there would be considerable homogeneity within neighborhood groups with respect to information gathered from each subject of the sample survey. Thus, in the initial planning, a multistage design was proposed with a sample taken of individuals within each sample neighborhood group. The Chinese collaborators, however, felt that within each neighborhood group sampled, any attempt to subsample individuals would seriously compromise the response rate and the overall cooperation of the residents. The major reason that this should be so is that Shanghai residents within the target age group would have difficulty understanding why the study would single out one person for interview but not a neighbor in the same age group. Also, the experience of this population, based on over 40 years of political campaigns under the socialist and earlier regimes, has led to considerable apprehensions among the elderly of their being "singled out" for any kind of interview. Thus, it was felt that a cluster sampling design that specified interviewing all individuals 55 years of age and older within each sample cluster would be the most feasible design.

References

[1] Groves, R. (1989). *Survey Methods and Survey Costs*. Wiley, New York.
[2] Kish, L. (1965). *Survey Sampling*. Wiley, New York.
[3] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.
[4] Levy, P.S., Yu, E., Liu, W.T., Wong, S., Zhang, M., Wang, Z. & Katzman, R. (1989). Single-stage cluster sampling with a telescopic respondent rule: a variation motivated by a survey of dementia in elderly residents of Shanghai, *Statistics in Medicine* **8**, 1537–1544.

PAUL S. LEVY

Cluster Score

In **cluster analysis of variables** we often obtain subsets or clusters of variables for the purpose of producing scores or cluster scores quantifying the subsets. A cluster score may be simply the sum of the original measurements on the variables in the cluster. Often, however, the variables have unequal standard deviations. In this case the variables are usually first standardized. Cluster scores can then be the sum of these standardized measurements. If the variables of the cluster are equally reliable, then this is a good strategy. For the case where the variables of the cluster are not equally reliable and are not too numerous, we can resort to two other methods of forming cluster scores. These involve forming a weighted sum of the original measurements or the standardized measurements on the variables in each cluster.

The first method starts by computing the loadings of the first **principal component** or the first **centroid** component separately for each cluster. A cluster score is then formed as a weighted sum of the standardized measurements of the variables in the cluster. The loadings of the first principal component or the first centroid component are used as the weights. In symbols, the cluster score for a cluster consisting of k variables is given by

$$C = a_1 X_1^* + a_2 X_2^* + \cdots + a_k X_k^*,$$

where $a_1, \ldots, a_k$ are the first component loadings of the variables in the cluster, and $X_1^*, \ldots, X_k^*$ are the standardized measurements of the k variables.

Kelley [3] proposed another cluster score, which is a weighted sum of the original variables in the cluster. The cluster score for a cluster with k variables is given as follows:

$$X = w_1 X_1 + w_2 X_2 + \cdots + w_k X_k,$$

where X_i is the original measurement of the ith variable and $w_i = \sqrt{r_{ii}}/[\sigma_i(1 - r_{ii})]$. r_{ii} is the reliability of the ith variable and σ_i is its standard deviation. This score is useful when the variables are educational and psychological tests, and reliability coefficients consistent with the intercorrelations are available. The reliability coefficients may be obtained by Kuder–Richardson's KR-20 [4], the split-half method, the Spearman–Brown formula, and a reliability approximation given by Cureton et al. [2]. Kelley's method maximizes the reliability of the cluster scores, so it is the preferred method whenever reliability coefficients are available.

Cureton & D'Agostino [1] suggest that cluster scores are more useful than component scores (*see* **Principal Components Analysis**) and the **factor scores**. This is because the cluster scores are measured with little error and the interpretation of a cluster is generally clearer than a factor or component (i.e. the interpretation of a cluster is simply the feature that is common to all the variables in the cluster). They are also more reproducible in other data sets from similar populations than are component or factor scores.

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[2] Cureton, E.E., Cook, R.T., Fischer, R.T., Laser, S.A., Rockwell, N.J. & Simmons, J.W., Jr (1973). Length of test and standard error of measurement, *Educational and Psychological Measurement* **33**, 63–68.

[3] Kelley, T.L. (1927). *Interpretation of Educational Measurements*. World Books, Yonkers-on-Hudson.

[4] Kuder, G.F. & Richardson, M.W. (1937). The theory of the estimation of test reliability, *Psychometrika* **2**, 151–160.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Clustering, Complete Linkage

Complete linkage clustering is one of several agglomerative algorithms used in hierarchical cluster analysis to partition a set of n observations into g groups or clusters based on the data collected on the n observations. (*See* **Cluster Analysis of Subjects, Hierarchical Methods**, in particular for determining an appropriate value of g.)

Specifically, suppose p variables are collected on the n observations. Let the $n \times p$ matrix $\mathbf{X}$ contain these data as follows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \ldots x_{1p} \\ x_{21} & x_{22} \ldots x_{2p} \\ \vdots & \vdots \quad \vdots \\ x_{n1} & x_{n2} \ldots x_{np} \end{bmatrix}.$$

Prior to performing complete linkage clustering, the matrix $\mathbf{X}$ must first be transformed into an $n \times n$ matrix $\mathbf{D}$ containing pairwise distances between observations. For example, one common choice for d_{ij}, the elements of $\mathbf{D}$, are the Euclidean distances

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2}$$

(*see* **Similarity, Dissimilarity, and Distance Measure**). The following steps are taken when performing hierarchical cluster analysis using complete linkage:

1. Initially, consider each of the n variables as n clusters. In other words, at the onset we have n clusters,

   $$C_1^{(1)} = S_1, \qquad C_2^{(1)} = S_2, \ldots, C_n^{(1)} = S_n,$$

   where $C_i^{(1)}$ is the ith cluster, S_i is the ith observation, $i = 1, 2, \ldots, n$, and where the superscript indicates that this is the first level of clustering.
2. Reduce the n observations (clusters) to $n - 1$ clusters by combining the two observations which have the smallest "distance" (d_{ij}) between them into one cluster. Assume, without loss of generality, that observations S_1 and S_2 are the two observations combined, and call the cluster into which they are combined $C_1^{(2)}$. We then have the $n - 1$ clusters

   $$C_1^{(2)} = (S_1, S_2), \qquad C_2^{(2)} = S_3,$$
   $$C_3^{(2)} = S_4, \ldots, C_{n-1}^{(2)} = S_n \tag{1}$$

   or

   $$C_1^{(2)} = (C_1^{(1)}, C_2^{(1)}), \qquad C_2^{(2)} = C_3^{(1)},$$
   $$C_3^{(2)} = C_4^{(1)}, \ldots, C_{n-1}^{(2)} = C_n^{(1)}. \tag{2}$$

3. Determine the distance between each of the clusters in (2) above. In complete linkage clustering, the "distance" between cluster $C_k^{(2)}$ and cluster $C_l^{(2)}$, $k, l = 1, 2, \ldots, n - 1$, is defined as the maximum of all pairwise distances between observations in cluster $C_k^{(2)}$ and observations in cluster $C_l^{(2)}$. This distance is denoted as $D_{kl}^{(2)}$. For example, $D_{13}^{(2)}$ for the clusters from (1) above is the maximum of (d_{13}, d_{23}), where d_{ij} is the distance between observations i and j. Note, at this step, $D_{kl}^{(2)} = d_{kl}$ for k not equal to 1.
4. Reduce the $n - 1$ clusters to $n - 2$ clusters by combining the two clusters in (2) with the smallest value of $D_{kl}^{(2)}$.
5. At subsequent steps, reduce the $n - j$ clusters to $n - j - 1$ clusters, $j = 2, 3, \ldots, n - 2$, until only one cluster (consisting of the entire sample) is obtained. Then choose an appropriate value of g, the final number of clusters into which the data are partitioned.

Note that a general complete linkage formula for determining the distance between a cluster $C_k^{(h)}$ and a cluster $C_l^{(h)}$, where $C_k^{(h)}$ was created by combining clusters $C_i^{(h-1)}$ and $C_j^{(h-1)}$ is

$$D_{kl}^{(h)} = 0.5 \times (D_{il}^{(h-1)} + D_{jl}^{(h-1)}) + 0.5 \times \left| D_{il}^{(h-1)} + D_{jl}^{(h-1)} \right|.$$

Further details can be found in Everitt [1] and McQuitty [2].

References

[1] Everitt, B.S. (1993). *Cluster Analysis*. Edward Arnold, London.
[2] McQuitty, L.L. (1996). Similarity analysis by reciprocal pairs for discrete and continuous data, *Educational and Psychological Measurement* **22**, 253–255.

(*See also* **Classification, Overview**; **Cluster Analysis, Variables**)

JOSEPH M. MASSARO

Clustering, Single Linkage

Single linkage clustering is one of several agglomerative algorithms used in hierarchical cluster analysis (*see* **Cluster Analysis of Subjects, Hierarchical Methods**) to partition a set n observations into g groups or clusters based on the data collected on the n observations (determining an appropriate value of g is covered elsewhere in this Encyclopedia).

Specifically, suppose that p variables are collected on the n observations. Let the $n \times p$ matrix $\mathbf{X}$ contain this data as follows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

Prior to performing single linkage clustering, the matrix $\mathbf{X}$ must first be transformed into an $n \times n$ matrix $\mathbf{D}$ containing pairwise distances between observations. For example, one common choice for d_{ij}, the elements of $\mathbf{D}$, are the Euclidean distances

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2}.$$

The following steps are taken when performing hierarchical cluster analysis using single linkage:

1. Initially, consider each of the n variables as n clusters. In other words, at the onset we have n clusters
$$C_1^{(1)} = S_1, \quad C_2^{(1)} = S_2, \dots, C_n^{(1)} = S_n,$$
where $C_i^{(1)}$ is the ith cluster, S_i is the ith observation, $i = 1, 2, \dots, n$, and where the superscript indicates that this is the first level of clustering.
2. Reduce the n observations (clusters) to $n-1$ clusters by combining the two observations which have the smallest "distance" (d_{ij}) between them into one cluster. Assume, without loss of generality, that observations S_1 and S_2 are the two observations combined, and call the cluster into which they are combined $C_1^{(2)}$. We then have the $n-1$ clusters
$$C_1^{(2)} = (S_1, S_2), \quad C_2^{(2)} = S_3,$$
$$C_3^{(2)} = S_4, \dots, C_{n-1}^{(2)} = S_n \qquad (1)$$
or
$$C_1^{(2)} = [C_1^{(1)}, C_2^{(1)}], \qquad C_2^{(2)} = C_3^{(1)},$$
$$C_3^{(2)} = C_4^{(1)}, \dots, C_{n-1}^{(2)} = C_n^{(1)}. \qquad (2)$$
3. Determine the distance between each of the clusters in (2) above. In single linkage clustering, the "distance" between cluster $C_k^{(2)}$ and cluster $C_l^{(2)}$, $k, l = 1, 2, \dots, n-1$, is defined as the minimum of all pairwise distances between observations in cluster $C_k^{(2)}$ and observations in cluster $C_l^{(2)}$. This distance is denoted as $D_{kl}^{(2)}$. For example, $D_{13}^{(2)}$ for the clusters from (1) above is the minimum of (d_{13}, d_{23}), where d_{ij} is the distance between observations i and j. Note, at this step, $D_{kl}^{(2)} = d_{kl}$ for k not equal to 1.
4. Reduce the $n-1$ clusters to $n-2$ clusters by combining the two clusters in (2) with the smallest value of $D_{kl}^{(2)}$.
5. At subsequent steps, reduce the $n-j$ clusters to $n-j-1$ clusters $(j = 2, 3, \dots, n-2)$ until only one cluster (consisting of the entire sample) is obtained. Then choose an appropriate value of g, the final number of clusters into which the data is partitioned.

Note that a general single linkage formula for determining the distance between a cluster $C_k^{(h)}$ and a cluster $C_l^{(h)}$, where $C_k^{(h)}$ was created by combining clusters $C_i^{(h-1)}$ and $C_j^{(h-1)}$, is

$$D_{kl}^{(h)} = 0.5 * (D_{ik}^{(h-1)} + D_{jk}^{(h-1)}) - 0.5 * |D_{il}^{(h-1)} + D_{jl}^{(h-1)}|.$$

Further details can be found in Everitt [1], Florek et al. [2], and Sneath [3], and in the articles **Classification, Overview**, and **Similarity, Dissimilarity, and Distance Measure**.

References

[1] Everitt, B.S. (1993). *Cluster Analysis*. Edward Arnold, London.

[2] Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H. & Zubrzycki, S. (1951). Sur la liason et la division des points d'un ensemple fini, *Colloquium Mathematicum* **2**, 282–285.

[3] Sneath, P.H.A. (1957). The application of computers to taxonomy, *Journal of General Microbiology* **17**, 201–226.

(*See also* **Cluster Analysis, Variables**)

JOSEPH M. MASSARO

Clustering

Clustering tests are of two basic types – cell count tests where the number of events in predefined regions are examined, and distance tests where the distance between nearest neighbors is examined. Use of matched control sets is a newer development. More work is needed on the definition and estimation of parameters for clustering.

Clustering can be defined as the "irregular" grouping of events from a **stochastic process** in either space or time or simultaneously in both space and time. More generally, it can be studied in any metric space that possesses a uniform or baseline "nonclustered" probability measure for the location of events (point process) under the **null hypothesis** of no clustering.

A major issue in this area, which is particularly acute in the medical context, is the differentiation between clustering as a general phenomenon, and the usually *post hoc* attempt to establish "significance" or "likelihood" of a real causal agent for a previously observed "cluster". The former question is amenable to reasonably standard statistical analysis. Usually, data are available from a large area or time period and a null hypothesis of "no clustering" can be formed. The description of a realistic **alternative "clustered" hypothesis** or family of hypotheses is usually more difficult, but a number of model cluster processes exist, and various *ad hoc* tests have been developed that will consistently detect a wide range of departures from the nonclustered null hypothesis. In many cases, these *ad hoc* tests have much to offer, since most clustering models are too inflexible to encompass the full range of possibilities seen in practice, and **likelihood ratio tests** based on them are usually complicated and may not have good **power** against other alternatives.

Unfortunately, it is far more common to encounter the latter situation, where some measure of "reality" is demanded for a "cluster" that has been identified before any statistical analysis has taken place. In this case, the temporal and/or spatial aspects of the cluster have been well circumscribed in advance and formal **hypothesis testing** is not valid. The problem is aptly illustrated by the Texas sharpshooter who fires his shots first and then positions the target to best advantage *afterwards*. Attempts to allow for this by adjusting for **multiple potential comparisons** are not very helpful, because the adjustment factor depends on the domain used for this, and by taking a large enough spatial or temporal domain, any cluster can be reduced to nonsignificance. In the end, the "reality" of any cluster ultimately depends on finding a cause and establishing that it also causes the disease in question in other circumstances. However, this is a costly and time-consuming process, and the "art of statistics" has an important role to play in helping to focus attention on "clusters" that still appear "highly unusual" after a range of statistical tests have been employed.

Such investigations usually start with readily available information, and clusters that are still interesting are then subjected to more stringent tests that may require the gathering of more data. A guideline for approaching this has been developed by the **Centers for Disease Control**, Atlanta, Georgia, USA [7, 49]. They propose a four-stage process of increasing complexity and cost, in which the investigation can be terminated after any stage if alternative explanations can be found or the evidence for clustering is no longer very compelling. In outline form, the stages are as follows:

Stage 1: Initial contact. The purpose at this stage is to collect information from the person(s) or group(s) first reporting a perceived cluster (hereafter referred to as the caller). The caller must be referred quickly to the responsible health agency unit, and the problem should never be dismissed summarily. The majority of potential cluster reports can be brought to successful closure at the time of initial contact, and the first encounter is often one of the best opportunities for communication with the caller about the nature of clusters.

Stage 2: Assessment. Once the decision has been made at Stage 1 to proceed further with an assessment, it is important to separate two concurrent issues: whether an excess number of cases has actually occurred, and whether the excess can be tied etiologically to some exposure. The first usually has precedence, and may or may not lead to the second. This stage initiates a mechanism for evaluating whether an excess has occurred. Three separate elements are identified:

1. a preliminary evaluation (Stage 2(a)) to assess quickly from the available data whether an excess may have occurred;
2. confirmation of cases (Stage 2(b)) to assure that there is a biological basis for further work; and

3. an occurrence investigation (Stage 2(c)) whose purpose is a more detailed description of the cluster through case-finding, interaction with the community, and **descriptive epidemiology**.

If an excess is confirmed, and the epidemiologic and biologic plausibility is compelling, proceed to Stage 3.

Stage 3: Major feasibility study. The major feasibility study examines the potential for relating the cluster to some exposure. It should consider all the options for geographic and temporal analysis, including the use of cases that are from a different geographic locale or time period and not part of the original cluster. In some instances the feasibility study itself may provide answers to the question under study.

If the feasibility study suggests that there is merit in pursuing an etiologic investigation, then the health agency should proceed to Stage 4. This may entail extensive resource commitment, however, and the decision to conduct a study will be tied to the process of resource allocation.

State 4: Etiologic investigation. Perform an etiologic investigation of a potential disease–exposure relationship. Using the major feasibility study as a guide, a specific protocol for the study should be developed and the study implemented.

Reporting of results. At whatever stage an investigation terminates, administrative closure is critical. Health authorities must remain aware that even internal reports are, in many circumstances, public documents, and can become part of legal proceedings. Even a brief memorandum to the record or a handwritten note summarizing a telephone call are subject to use in court and should be handled accordingly.

Issues of ***Statistics in Medicine*** (April–May 1996, Vol. 15) and the *American Journal of Epidemiology* (July 1990, Vol. 132) have been devoted exclusively to theoretical and practical issues related to clustering. These publications and the book edited by Elliot et al. [16] provide fuller details of points raised below.

Statistical Tests for Clustering

Tests for clustering, either spatially or temporally, fall into four main classes:

1. Methods based on cell occupancy counts for a partition of the region of interest.
2. Methods based on overlapping cells or adjacencies of cells with "high counts".
3. Distance methods.
4. Space–time clustering methods.

Methods can also be distinguished by the form of the null hypothesis. In nonmedical settings, often a uniform process is appropriate in which events occur according to a homogeneous **Poisson process** on the time axis (for temporal clustering) or the plane (for spatial clustering). For medical applications often some degree of inhomogeneity is appropriate for the nonclustered process. For **time series, seasonal** effects may need to be accounted for, and in studying spatial clustering, the nonuniform nature of the population density needs to be considered. In these circumstances, some form of **control** series is needed to allow for variations not related to short-range clustering.

Cell Occupancy Methods

These methods are equally appropriate to both temporal and spatial clustering as they take no account of the geometric structure of the region once it has been partitioned into nonoverlapping cells.

Dispersion Test. The simplest test for clustering is a heterogeneity or dispersion test [10]. When there are n cells, this is a simple **chi-square test** for an $n \times 1$ table. If E_i denotes the expected number of events in cell i, and N_i is the observed number, then the test is given by

$$T_{\mathrm{D}} = \sum_{i=1}^{n} \frac{(N_i - \mathrm{E}_i)^2}{\mathrm{E}_i}.$$

When all the E_i are large (≥ 5), the test has an approximately χ^2 distribution on n degrees of freedom. However, this is usually not the case. Nevertheless, T tends to a **normal distribution** as $n \to \infty$, in general. The **mean** and **variance** depend on whether one conditions on the total number of events $N = \sum N_i$. Usually, this is the case so that the N_i have a **multinomial distribution** with cell probabilities $p_i = \mathrm{E}_i/\mathrm{E}$, where $\mathrm{E} = \sum \mathrm{E}_i$, and total number of observations N. In that case [24, 38],

$$\mathrm{E}(T_{\mathrm{D}}) = n - 1,$$

$$\mathrm{var}(T_{\mathrm{D}}) = 2(n-1) + \sum \mathrm{E}_i^{-1} - \frac{1}{N}(n^2 - 2n + 2),$$

and $[T_D - E(T_D)]/\text{var}^{1/2}(T_D)$ is approximately a **standard normal deviate**.

Potthoff & Whittinghill's Test. Pothoff & Whittinghill [46, 47] considered a cell occupancy model in which the cell counts were Poisson with mean E_i under the null hypothesis. Under the alternative, the E_i were assumed to be multiplied by independent **gamma-distributed** variables with mean one, so that the observed counts had a **negative binomial distribution** (i.e. a compound Poisson-gamma distribution). They computed the **score** test as the variance of the gamma distribution tended to zero while keeping the means fixed, and arrived at a test of the form:

$$T_{PW} = \sum_{i=1}^{n} \frac{N_i(N_i - 1)}{E_i}.$$

Again, the N_i are multinomial and if $\max_i p_i \to 0$ as $n \to \infty$,

$$E(T_{PW}) = N - 1,$$
$$\text{var}(T_{PW}) \cong \frac{2(N-1)^2}{N},$$

and

$$\frac{[T_{PW} - E(T_{PW})]}{\text{var}^{1/2}(T_{PW})} \longrightarrow \mathcal{N}(0, 1).$$

Note that cells contribute nothing to T_{PW} unless there are at least two events in the cell. Also, when all the E_i are the same, this test is equivalent to the dispersion test given above. In general, T_{PW} gives less weight to cells with larger expected values than does T_D.

Dispersion Tests with Controls. In some circumstances, it is appropriate to consider whether the variability in a time series of events is greater than that in a control series. The control series may account for seasonal disease patterns or changing referral patterns and is used as a surrogate for the cell probabilities p_i defined above. An approach based on a chi-square test for independence in a $2 \times n$ table has been proposed by Fleiss & Cuzick [18] in which the case and control series comprise the two rows. The asymptotic distribution is *not* χ^2 on (n − 1) degrees of freedom (df) unless all cell counts are large, but as $n \to \infty$ the test does tend to be a normal with mean $(n - 1)$ and variance [25, 38]

$$\text{var}(T_D) \cong 2(n - 1 - N^*) + (p^* - 2)\left(N^* - \frac{n^2}{N}\right) + 2\left(\frac{n}{N}\right)(1 - p^*) - \left(\frac{n}{N}\right)^2 (5 - p^*) + O\left(\frac{1}{n}\right),$$

where

$$p^* = \overline{p}^{-1} + (1 - \overline{p})^{-1},$$
$$N^* = \sum_{i=1}^{n} N_i^{-1},$$

$\overline{p}$ is the proportion of all events that are in the case series, and N_i is the total number of events in period i for cases and controls combined.

Tests Based on the Maximum Cell Count. When looking for a single cluster, tests based on the maximum cell count are an obvious choice. When done in a *post hoc* fashion, that is, when the cluster has already been identified, the significance is completely dependent on the number of cells included in the sample. A more attractive option is to examine the maximum cell count in a number of short subsets of the data. This approach has an intuitive appeal, and will detect multiple clusters. A weakness is that the cells are chosen in advance and clusters that are split over two cells will not be fully scored. Also, the method assumes equal expected numbers in the different cells. The method was first proposed by Ederer et al. [15]. For their test, the cells are first partitioned into a number of nonoverlapping time series. For example, if the data consisted of quarterly event rates over a 20-year period in a number of localities, the individual time series might be chosen to be the event rates over a five-year period for each locality. These individual series are then self-normalized by conditioning on the total number of events in that subseries. Thus, if N_{ij} denotes the number of events in cell i of series $j, i = 1, \ldots, I, j = 1, \ldots, J$, one computes $M_j = \max_{1 \le i \le I} N_{ij}$ conditional on $N_j = \sum_{i=1}^{I} N_{ij}$.

The overall test statistic is then

$$T_{EMM} = \sum_{j=1}^{J} M_j,$$

and is normalized by the conditional means:

$$\mathrm{E}(T_{\mathrm{EMM}}) = \sum \mathrm{E}(M_j|N_j),$$

and conditional variances

$$\mathrm{var}(T_{\mathrm{EMM}}) = \sum \mathrm{var}(M_j|N_j)$$

to form an approximately standard normal deviate in the usual way.

Thus, this test only looks for evidence of clustering within the subseries and by appropriate choice of the space and time groups can be viewed as a space–time clustering method as well. However, by including only one geographic entity and allowing the subseries to be rather long, it takes on more the character of a purely temporal statistic. An underlying assumption is that all cells in the same subseries have the same expected event rate under the null hypothesis. Thus, some adjustment for seasonal variation can be made by creating subseries based on the same months in successive years, and temporal trends can be accounted for by using short time periods, but it is not possible to adjust for both simultaneously.

Ederer et al. rely on the asymptotic normality of their test and thus it is only necessary to compute $\mathrm{E}(M_i|N_i)$ and $\mathrm{var}(M_i|N_i)$, for which Mantel et al. [35] have given tables for small values of N_i and I. Grimson [22] has obtained the exact distribution of the maximum, based on factorial moments and the inclusion–exclusion formula. Specifically, he shows that when $N_i = r$ and $I = c$:

$$\Pr(M_i \geq m) = i^r \sum_{k=1} (-1)^{k+1} \binom{c}{k} \times \sum_{j_1,\ldots,j_k=m}^{r} (c-k)^{r-j_1-\cdots-j_k} \times \binom{r}{j_1,\ldots,j_k}.$$

Levin [31] has given a large sample approximation based on Edgeworth expansions.

Overlapping or Adjacent Cells Methods

One of the problems with cell occupancy methods is that clusters may span more than one cell. To be valid, the partitioning of space for these methods must be independent of the location of events so that it is likely that any clusters will not match up with the partitioning. Attempts to address this problem have led to tests based on overlapping or neighboring cells.

Scan Tests. For temporal clustering, where events are distributed on a line and cells are often based on equal length intervals (e.g. months), an obvious solution exists. Instead of looking only at the number of events in prespecified intervals (e.g. six months), it is also possible to look at overlapping intervals (e.g. six-monthly intervals beginning every quarter). The most thorough method is to look at all intervals of a fixed size (e.g. six-monthly intervals beginning every day), and consider the maximum number of events in any such interval. However, the fact that overlapping intervals are now being considered greatly complicates the problem of determining the distribution of the scan test. To simplify matters, one usually assumes that the time axis is broken into sufficiently small intervals (e.g. one day) so that one can safely assume that the scan is continuous. Under the null hypothesis, the problem is then transformed into determining the distribution of $n(t, N)$, the maximum number of events in an interval of length t when N events are uniformly distributed on the unit interval. Computations for this probability go back at least as far as the 1940s [2, 33], but Naus [39, 41, 42] was the first to develop this as a test for temporal clustering. Exact calculations require detailed combinational expressions and are very computationally time-consuming in the important case when t is small. Various approximations have been proposed [3, 13, 19, 21, 30, 42, 51] (*see* **Scan Statistics for Disease Surveillance**).

The primary quantity of interest is $p(n, t, N) = \Pr(n(t, N) \geq n)$, and approximations are given in terms of the **binomial** probabilities $b(j, n, p) = \binom{n}{j} p^j (1-p)^{n-j}$. Wallenstein & Neff [51] give the simple approximation:

$$p(n, t, N) \cong \left(\frac{n}{t} - N + 1\right) b(n, N, t) + 2 \sum_{j=n+1}^{N} b(j, N, t),$$

which is accurate when $p(n, t, N)$ is small and, in fact, exact when $n > N/2$ and $t < \frac{1}{2}$. Berman & Eagleson [3] give an upper bound of similar complexity based on a second-order inclusion–exclusion

(Bonferroni) approximation:

$$P(n,t,N) \le (N-n+1)\sum_{j=n-1}^{N} b(j,N,t) - \sum_{j=n-1}^{N}(-1)^{j+n-1}b(j,N,t),$$

which is also a good approximation in some cases. Glaz [20, 21] has developed more complicated expressions on the basis of higher-order **Bonferroni inequalities** and a product-type inequality, all of which have greater accuracy. Two series of expressions are given that increase in complexity (and hopefully accuracy) as L increases.

The first is an upper bound: for $1 \le L \le n \le N/2$:

$$P(n,t,N) \le \sum_{j=1}^{L-1} Q_j^* + (N-n+2-L)Q_L^*,$$

where

$$Q_1^* = \sum_{J=n-1}^{N} b(j,N,t),$$

$$Q_2^* = Q_1^* - \sum_{j=n-1}^{N}(-1)^{j+n-1}b(j,N,t),$$

and for $j \ge 3$

$$Q_j^* = b(n-1,N,t) - b(n,N,t) + \sum_{k=j}^{N-n+1}(-1)^k \prod_{i=1}^{j-2}\left[\frac{1-k(k-1)}{i(i+1)}\right] \times b(n+k-1,N,t).$$

A more accurate approximation is given by

$$P(n,t,N) \cong 1 - \left(1 - \sum_{j=1}^{L} Q_j^*\right) \times \left[\left(1-\sum_{j=1}^{L} Q_j^*\right) \Big/ \left(1-\sum_{j=1}^{L-1} Q_j^*\right)\right]^{N-n+1-L},$$

which is said to be most accurate when $L = n$. Neff & Naus [43] have tabulated $P(n,t,N)$ for $t < \frac{1}{2}$, $3 \le n < N \le 25$.

Approximations for $P(n,t,N)$ under various clustering alternatives are described in Wallenstein et al. [52], which are important for power calculations.

Glaz [21] gives references for approximations of the moments of $n(t,N)$ and various related quantities. He also discusses scan statistics based on the use of a range of different window widths, t_i. Nagarwalla [37] also discusses a scan test with variable window width. Simulation is required to approximate its null distribution.

Extensions of the scan statistic to the plane or higher dimensions have not been very fully developed. There are problems in the choice of metric (e.g. circles or squares) and the structure of overlapping cells is much more complicated in the plane. Naus [40] has some theoretical results for the plane, and Openshaw et al. [44] give an example of an application that uses circles of different radii. This approach is very descriptive, however.

Runs Test. The runs test is a well-established method for evaluating clustering in sequences of **binary** data. For temporal data, such sequences can be created by looking at intervals of fixed length and recording whether one or more events occur in each interval, or, when events are more common, intervals in which an excessive number of events have occurred. The elements of the sequence are then treated as independent Bernoulli trials with fixed success probabilities. As the success probability is usually unknown, one usually is interested in the probability of a run of length at least m in n trials in which there are known to be s successes overall. Denoting this as $P_R(m,s,n)$, the exact distribution is (cf. [5, p. 257])

$$P_R(m,s,n) = \sum_{i=1}^{[s/m]}(-1)^{i+1} \times \binom{n-s+1}{i}\binom{n-im}{n-s} \Big/ \binom{n}{s}.$$

Burr & Cane [6] have developed and surveyed various approximations and Naus [42], on the basis of ideas used for the scan test, suggested

$$P_R(m,s,n) \cong 1 - Q_2^{**}\left(\frac{Q_3^{**}}{Q_2^{**}}\right)^{(n/m)-2},$$

where

$$Q_2^{**} = \left[m \binom{n-m-1}{s-m} + \binom{n-m}{s-m} \right] \Big/ \binom{n}{s},$$

$$Q_3^{**} = \left[2m \binom{n-m-1}{s-m} + \binom{n-m}{s-m} - \binom{m}{2} \binom{n-2m-2}{s-2m} - m \binom{n-2m-1}{s-2m} \right] \Big/ \binom{n}{s}.$$

A simple, but less accurate, approximation is given by Feller [17]:

$$P_R(m, s, n) \cong 1 - \exp(-nqp^m),$$

where $p = 1 - q = s/n$. Tests have also been based on the number of runs greater than a certain size or the total number of runs (change from zero to one or vice versa) in a series [27].

Join–Count Statistics. This approach was first developed by Moran [36], and the test is sometimes called the Geary–Moran statistic. It has been developed for two-dimensional maps, but the ideas are quite general and suitable for other dimensions as well. It can be seen as a hybrid between a cell occupancy method and a distance method. The basic approach is as follows: one starts with a map in which a partition into cells is given. Cells with "large" numbers of events are determined by some scheme chosen separately by the investigator. Common approaches are to choose cells in which the observed number of events exceeds the expected number by a given percentage (standardized incidence ratio) (*see* **Standardization Methods**) or is significantly different at a predetermined level (say 5%). The number of pairs of "such" extreme cells that are adjacent, that is, that share a common boundary (denoted T_{JC}), is then determined and compared against expected numbers. The problem can be reduced to the analysis of a graph in which the cells are vertices and adjacent cells are connected by edges. The null hypothesis is that the "extreme cells" are chosen at random (permutational distribution). Under this hypothesis, the expected number of adjacent cells is

$$E(T_{JC}) = Np_1,$$

where N is the total number of joins (adjacent cells) and for $j = 0, 1, 2, \ldots$,

$$p_j = \prod_{i=0}^{j} \frac{n_0 - l}{n - l},$$

n = number of points, and n_0 = number of "extreme" points, and the variance is

$$\text{var}(T_{JC}) = N(p_1^2 - p_3) - N^2(p_1^2 - p_3) + p_2 \sum_{i=1}^{n} M_i(M_i - 1) - p_3 \sum_{j=1}^{n} K_j,$$

where M_i is the number of joins emanating from point i, and K_j is the number of points to which the points at the ends of join j are both joined.

Asymptotic normality of $[T_{JD} - E(T_{JC})]/\text{var}^{1/2}(T_{JC})$ can be established. Besag [4] has noted that even in the absence of clustering, the null hypothesis may not hold when extreme cells are determined by observed to expected ratios. When the populations in different cells are not the same, and low population cells are next to each other (e.g. rural areas), those cells are more likely to be extreme because of greater random fluctuations, and artefactual aggregation could arise. Contrariwise, if there is general, but unaccounted for, extra Poisson variation, and significance levels based on the Poisson distribution are used to determine extreme cells, then cells with large underlying populations could be over-represented. In these circumstances, more complicated expressions for the mean and variance of T_{JC} are needed, or **simulation** must be used to assess the significance level.

Distance Methods

Nearest Neighbor Tests for Uniform Populations. Under the assumption of a uniform (or homogeneous Poisson) distribution, tests for clustering have been based on the distance to the kth nearest neighbor of any particular case [9, 32]. If d_k denotes the distance to the kth nearest point from any arbitrary point in the plane, then when the number of points is large (so edge effects can be discounted), d_k has a gamma (μ, k) distribution:

$$P(d_k > t) = \frac{\sum_{j=0}^{k-1} \mu^j e^{-\mu}}{j!},$$

where $\mu = 2\pi\lambda d_k^2$ and λ is the event rate of the underlying Poisson process. Tests have been based on the mean of d_k, often with $k = 1$, that is, the average distance from an arbitrary point to its nearest neighbor. Other approaches based on more complicated sampling plans are surveyed in [48, Chapter 7].

Covariance Function. An alternative approach is to use the k function associated with homogeneous **point processes** [13]. Specifically, for a homogeneous point process with event rate λ per unit area, define $k(t) = \lambda^{-1}\mathrm{E}$ (number of events within distance t of an arbitrary event). For a Poisson process with unit intensity, $k(t) = \frac{1}{2}\pi t^2$. Empirical estimates $\hat{k}$ can be formed in an obvious way and compared with $k(t)$ to see if more or fewer events are occurring near to an arbitrary event.

Nearest Neighbor Tests for Nonuniform Populations. Work with homogeneous populations arose from questions in ecology and geography and is generally not applicable to questions of disease association in populations because of the nonhomogeneous distribution of the population. In this case, some estimate of the local population density is also required. Cuzick & Edwards [11, 12] have developed a variety of tests in this setting. When the population density is known precisely, this test consists of creating circles of different radius but constant expected number of events around each case and counting the number of observed events in all such circles. If the expected number is taken to be λ and O_i is the observed number of events in the circle around event i, then this takes the form

$$T_{1s} = \sum_i (O_i - \lambda),$$

with

$$\mathrm{E}(T_{1S}) = 0,$$
$$\mathrm{var}(T_{1S}) = n\lambda + 2n^{-1}N_S + n^{-1}N_t - n\lambda^2,$$

where n is the number of points, N_S is the number of pairs of points in each other's neighborhood, and $N_t = \sum M_i(M_i - 1)$, where M_i is the number of points for which point i falls in its neighborhood. In many circumstances, precise information about the geographic location of the population is unavailable and information about this must be replaced by a control series selected in an appropriate way. Cuzick & Edwards [11] developed a test on the basis of the number of cases in the k nearest neighbors ($k - NN$) to each case; see also Schilling [50] and Henze [26]. Specifically, cases and controls are labeled from $1, \ldots, N$ and

$$T_k = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\delta_i\delta_j,$$

where δ_i is the indicator for the ith event to be a case and a_{ij} is the indicator for event j to be among the $k - NN$s of event i. This test can also be viewed as formally similar to a join–count test, except that the a_{ij} is not in general symmetric (although it is easily converted by replacing a_{ij} with $\frac{1}{2}(a_{ij} + a_{ji})$). The permutational distribution of T_k can be computed as before (assuming cases are selected at random from the set of cases and controls) and yields

$$\mathrm{E}(T_k) = p_1 kn,$$

where p_1 is as in the section on join–count statistics, where n_o denotes the number of cases and n the total number of points, and

$$\begin{aligned}\mathrm{var}(T_k) = {} & (kn + N_s)p_1(1 - p_1) \\ & + [(3k^2 - k)n + N_t - 2N_s](p_2 - p_1)^2 \\ & - [k^2(n^2 - 3n) + N_s - N_t](p_1^2 - p_3),\end{aligned}$$

where N_s is the number of pairs of points for which the $k - NN$ relation is symmetric, that is, pairs that are $k - NN$s of each other and $N_t = \sum_{i=1}^{n} M_i(M_i - 1)$, and where M_i is the number of points for which point i is a $k - NN$.

Diggle [14] has suggested an alternative based on comparing the k function of the cases with that of the controls. His approach essentially uses circles of the same geographic size, whereas the Cuzick–Edwards approach uses circles of approximately equal population. The choice of approach depends on the type of alternative envisaged. Cuzick & Edwards [11] have also considered tests based on the number of cases encountered before k controls. When $k = 1$, this is similar to the runs test sometimes used to look for clustering in one dimension.

Space–Time Clustering Methods

When events are thought to be closely related to exposures both in time and distance, such as in

epidemics of disease associated with contagious infectious agents, then tests based on space–time clustering are most appropriate. Such tests are self-normalizing in the sense that any overall nonhomogeneity in the time course of the events or their spatial distribution is automatically accounted for. This has great advantages in terms of automatically cancelling out seasonal effects of nonhomogeneous population distributions, but, of course, limits the ability of the test to detect alternatives that vary both in time and space. Thus, a point source of events or a change that affected the entire geographic region under consideration would not be detected. Thus, these tests are appropriate for infectious diseases with short incubation times, but Chen et al. [8] have shown they have low power for the type of clustering expected in adult cancer or other chronic diseases where the **latent period** between exposure and disease is long.

The simplest space–time tests partition space and time separately and then perform a chi-square test for independence on the associated two-way **contingency table** in which spatial cells are rows and temporal cells are columns. However, more interest has been generated by the work of Knox [28, 29], in which distances and times are computed between all pairs of points and tests are developed to determine whether events that are close spatially are also close in time. By partitioning distances and time intervals, these data can be summarized in a two-way table, but the induced **correlations** arising from considering pairs of points means that the **chi-square distribution** is invalid for assessing independence. Knox only considered a **two by two table** formally by dichotomizing time and spatial pairs as close or distant in each variable. One problem with Knox's approach is the arbitrary dichotomy on the close and distant pairs. Related tests have been proposed by Pinkel & Nefzger [45], Barton & David [1] and Mantel [34]. Mantel [34] considered a more general approach giving weights to distances that decrease as the distance increases. Knox's tests is then a special case in which the weight changes from one to zero, once the cut-off is exceeded. Mantel's test can be written in the very general form:

$$T_{\mathrm{M}} = \sum \sum_{i \neq j} X_{ij} Y_{ij},$$

where $X_{ij} = a(i, j)$ is a score for the pair of spatial variables and $Y_{ij} = b(i, j)$ is a score for the pair of temporal variables. Under the assumption that the spatial variables are unrelated to the temporal variables, the permutational distribution of T_{M} can be simulated or approximated. For large samples, the permutational mean and variance can be computed and asymptotic normality assumed, although conditions for this are not clearly known [23]. Even the computation of the permutational variance can be quite involved [34].

References

[1] Barton, D.E. & David, F.N. (1966). The random intersection of two graphs, in *Research Papers in Statistics*, F.N. David, ed. Wiley, New York, pp. 455–459.

[2] Berg, W. (1945). Aggregates in one- and two-dimensional random distributions, *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **36**, 319–336.

[3] Berman, M. & Eagleson, G.K. (1985). A useful upper bound for the tail probabilities of the scan statistic when the sample size is large, *Journal of the American Statistical Association* **80**, 886–889.

[4] Besag, J. (1990). Discussion of the paper by Cuzick and Edwards, *Journal of the Royal Statistical Society, Series B* **52**, 96–104.

[5] Bradley, J.V. (1968). *Distribution-Free Statistical Tests*. Prentice-Hall, Englewood Cliffs.

[6] Burr, E.J. & Cane, G. (1961). Longest run of consecutive observations having a specified attribute, *Biometrika*, **48**, 461–465.

[7] CDC (1990). Guidelines for investigating clusters of health events, *Morbidity and Mortality Weekly Report* **39**, (RR-11), 1–23.

[8] Chen, R., Mantel, N. & Klingberg, M.A. (1984). A study of three techniques for time-space clustering in Hodgkin's disease, *Statistics in Medicine* **3**, 173–184.

[9] Clark, P.J. & Evans, F.C. (1955). On some aspects of spatial pattern in biological populations, *Science* **121**, 397–398.

[10] Cox, D.R. & Lewis, P.A.W. (1966). *The Statistical Analysis of Series of Events*. Methuen, London.

[11] Cuzick, J. & Edwards, R. (1990). Spatial clustering for inhomogeneous populations, *Journal of the Royal Statistical Society, Series B* **52**, 73–104.

[12] Cuzick, J. & Edwards, R. (1996). Clustering methods based on k nearest neighbour distributions, in *Methods for Investigating Localized Clustering of Disease*, F.E. Alexander & P. Boyle, eds. IARC Scientific Publication No. 135, Lyon.

[13] Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

[14] Diggle, (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity

of a pre-specified point, *Journal of the Royal Statistical Society, Series A* **153**, 349–362.

[15] Ederer, F., Myers, M.H. & Mantel, N. (1964). A statistical problem in space and time: do leukemia cases come in clusters?, *Biometrics* **20**, 626–638.

[16] Elliott, P., Cuzick, J., English, D. & Stern, R. (1992). *Geographical and Environmental Epidemiology Methods for Small-Area Studies*. Oxford University Press, Oxford.

[17] Feller, W. (1957). *An Introduction to Probability Theory and its Application*, Vol. 1, 2nd Ed. Wiley, New York.

[18] Fleiss, J.L. & Cuzick, J. (1979). The reliability of dichotomous judgments: unequal numbers of judges per subject, *Applied Psychological Measurement* **3**, 537–542.

[19] Gates, D.J. & Westcott, M. (1984). On the distributions of scan statistics, *Journal of the American Statistical Association* **79**, 423–429.

[20] Glaz, J. (1992). Approximations for tail probabilities and moments of the scan statistic, *Computational Statistics and Data Analysis* **14**, 213–227.

[21] Glaz, J. (1993). Approximations for the tail probabilities and moments of the scan statistic, *Statistics in Medicine* **12**, 1845–1852.

[22] Grimson, R. (1993). Disease cluster, exact distributions of maxima, and P-values, *Statistics in Medicine* **12**, 1773–1794.

[23] Guttorp, P. & Lockhart, R.A. (1988). On the asymptotic distribution of quadratic forms in uniform order statistics, *Annals of Statistics* **16**, 433–449.

[24] Haldane, J.B.S. (1937). The exact value of the moments of the distribution of χ^2, used as a test of goodness of fit, when expectations are small, *Biometrika* **29**, 133–143.

[25] Haldane, J.B.S. (1940). The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small, *Biometrika* **31**, 346–355.

[26] Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbour type coincidence, *Annals of Statistics* **16**, 772–783.

[27] Johnson, N.L. & Kotz, S. (1969). *Discrete Distributions*. Wiley, New York.

[28] Knox, E.G. (1964). The detection of space-time interactions, *Applied Statistics* **13**, 25–29.

[29] Knox, E.G. (1964). Epidemiology of childhood leukaemia in Northumberland and Durham, *British Journal of Preventive and Social Medicine* **18**, 17–24.

[30] Krauth, J. (1988). An improved upper bound for the tail probability of the scan statistic for testing nonrandom clustering, in *Classification and Related Methods of Data Analysis, Proceeding of the First Conference of the International Federation of Classification Society*. Technical University of Aachen, Germany, pp. 237–244.

[31] Levin, B. (1981). A representation for multinomial cumulative distribution functions, *Annals of Statistics* **9**, 1123–1126.

[32] Lewis, M.S. (1980). Spatial clustering in childhood leukaemia, *Journal of Chronic Diseases* **33**, 703–712.

[33] Mack, C. (1948). An exact formula for $Q_k(n)$, the probable number of k-aggregates in a random distribution of n points, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **39**, 778–790.

[34] Mantel, N. (1967). The detection of disease clustering and a generalized regression approach, *Cancer Research* **27**, 209–220.

[35] Mantel, M., Kryscio, R.J. & Myers, M.H., (1976). Tables and formulas for extended use of the Ederer-Myers-Mantel disease clustering procedure, *American Journal of Epidemiology* **104**, 576–584.

[36] Moran, P.A.P. (1948). The interpretation of statistical maps, *Journal of the Royal Statistical Society, Series B* **10**, 243–251.

[37] Nagarwalla, N. (1996). A scan statistic with a variable window, *Statistics in Medicine* **15**, 845–850.

[38] Nass, C.A.G. (1959). The χ^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism, *Biometrika* **46**, 365–385.

[39] Naus, J.I. (1965). The distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association* **60**, 532–538.

[40] Naus, J.I. (1965). Clustering of random points in two dimensions, *Biometrika* **52**, 263–267.

[41] Naus, J.I. (1966). Some probabilities, expectations and variances for the size of the largest clusters and smallest intervals, *Journal of the American Statistical Association* **61**, 1191–1199.

[42] Naus, J.I. (1982). Approximations for distributions of scan statistics, *Journal of the American Statistical Association* **77**, 177–183.

[43] Neff, N.D. & Naus, J.I. (1980). The distribution of the size of the maximum cluster of points on a line, in *IMS Series of Selected Tables in Mathematical Statistics*, Vol. 6, American Mathematical Society, Providence.

[44] Openshaw, S., Craft, A.W., Charlton, M. & Birch, J.M. (1988). Investigation of leukaemia clusters by use of a geographical analysis machine, *Lancet* **i**, 272–273.

[45] Pinkel, D. & Nefzger, D. (1959). Some epidemiological features of childhood leukemia in the Buffalo, N.Y. area, *Cancer* **12**, 351–358.

[46] Potthoff, R.F. & Whittinghill, M. (1966). Testing for homogeneity. I. The binomial and multinomial distributions, *Biometrika* **53**, 167–182.

[47] Potthoff, R.F. & Whittinghill, M. (1966). Testing for homogeneity. II. The Poisson distribution, *Biometrika* **53**, 183–190.

[48] Ripley, B.D. (1981). *Spatial Statistics*. Wiley, New York.

[49] Rothenberg, R.B. & Thacker, S.B. (1992). Guidelines for the investigation of clusters of adverse health events, in *Geographical and Environmental Epidemiology. Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. Oxford University Press, Oxford, 264–277.

[50] Schilling, M.F. (1986). Multivariate two-sample tests based on nearest neighbors, *Journal of the American Statistical Association* **81**, 799–806.

[51] Wallenstein, S. & Neff, N. (1987). An approximation for the distribution of the scan statistic, *Statistics in Medicine* **6**, 197–207.

[52] Wallenstein, S., Naus, J. & Glaz, J. (1993). Power of the scan statistic for detection of clustering, *Statistics in Medicine* **12**, 1829–1843.

JACK CUZICK

Coarsening at Random

Incomplete Data

Incomplete data often occur in biostatistical contexts. Typical examples include **censoring** in **survival analysis**, dropout in longitudinal studies (*see* **Nonignorable Dropout in Longitudinal Studies**), and missing values in multivariate data sets (*see* **Missing Data**). Generally, this can be modeled by replacing the incomplete observation by the subset of the sample space that is known to contain the observation one had wished to make. For instance, right censoring gives rise to half-lines from the censoring time to infinity. Observations where incompleteness is represented by subsets of the sample space in this way are called *coarsened*. Typically, the subsets are stochastic – for instance, determined by a stochastic censoring time – and the criteria for ignoring this randomness have been useful, particularly in the context of missing data and right censoring. In the general model, these criteria go under the name of *coarsening at random*.

Coarsening at random was introduced as a generalization of missing at random by Heitjan and Rubin [3]. Heitjan [2] gives several biostatistical examples and an application to data from the Stanford Heart **Transplantation** Program. Nielsen [6] discusses how to analyze survival data with coarsely observed covariates.

The Coarsening Model

Let X be a random variable we intend to observe but only observe incompletely. Instead we observe a subset, Y, of the sample space, $\mathcal{X}$; Y represents what is observed about X – in particular, $X \in Y$ – but may depend on further randomness. This is modeled by an auxiliary random variable, G, such that Y is a (nonrandom) function of (X, G), that is, $Y = Y(X, G)$, say. Often, G is incompletely observed as well – for instance, a censoring time is not observed unless the survival time is censored – and this is modeled by a subset, H, of the sample space, $\mathcal{Z}$, of G. Thus, the incomplete observation of X is represented by a subset $T = Y \times H$ of the product, $\mathcal{X} \times \mathcal{Z}$, of the sample spaces. This subset should represent all that is known about the intended observation, X, and the coarsening variable, G. Therefore, T contains exactly those elements, (x, g) of $\mathcal{X} \times \mathcal{Z}$ that are mapped to T; that is, $\{(x, g) \in \mathcal{X} \times \mathcal{Z} : T(x, g) = T\}$. Typically, there is a natural choice of the *coarsening variable*, G; otherwise G is just the observed subset, Y.

A slightly more general model is discussed by Nielsen [5] and Gill et al. [1].

Example

(1) Let X be a survival time and G, a censoring time. We observe the smaller of these two, $T^* = X \wedge G$ and which one is the smallest. Such data are called right censored. Here,

$$T(X, G) = \begin{cases} \{X\} \times [X; \infty] & \text{if } T^* = X \\ ]G; \infty[\times \{G\} & \text{if } T^* = G \end{cases}. \tag{1}$$

(2) Let X denote age, which is reported rounded either to the nearest month ($G = 0$) or the nearest year ($G = 1$). Such data are called heaped. Let T^* be the age reported, that is,

$$T^*(X, G) = \begin{cases} \dfrac{\lfloor 12X + 6 \rfloor}{12} & \text{if } G = 0 \\ \left\lfloor X + \frac{1}{2} \right\rfloor & \text{if } G = 1 \end{cases}, \tag{2}$$

where $\lfloor x \rfloor$ is the largest integer not greater than x. Then

$$T(X, G) = \begin{cases} [T^* - \frac{1}{24}; T^* + \frac{1}{24}[\times \{0\} & \text{if } G = 0 \\ [T^* - \frac{1}{2}; T^* + \frac{1}{2}[\times \{1\} & \text{if } G = 1 \end{cases}. \tag{3}$$

Let f_θ be the density of the distribution of X and let the conditional distribution of G given $X = x$ have density $h_\gamma(g|x)$. All densities – probability density functions or probability mass functions – here and in what follows are densities with respect to probability measures. A density with respect to a probability measure is obtained from an ordinary density by dividing it with a fixed density. The reference measure is then the probability measure given by the fixed density; see [4] for a discussion.

The conditional density of T given $X = x$ is then $k_\gamma(t|x) = E_x[h_\gamma(G|x)|T(x, \cdot) = t]$; here, the expectation is with respect to the reference measure of the

distribution of G given $X = x$; this may depend on x. The unconditional density of T is given by

$$\begin{aligned}\varphi_{\theta,\gamma}(t) &= E[f_\theta(X)h_\gamma(G|X)|T = t] \\ &= E[f_\theta(X)k_\gamma(T|X)|T = t], \end{aligned} \quad (4)$$

where the expectation is with respect to the reference measure of the joint distribution of (X, G).

Let $P_{\theta,\gamma}$ be the distribution of (X, G) defined above. A statistical model is obtained when (θ, γ) varies in a space, $\Theta \times \Upsilon$.

Coarsening at Random

Three different concepts of coarsening at random have been developed.

- A *probability*, $P_{\theta,\gamma}$, is *absolutely coarsened at random* or CAR(ABS) if for each pair (x, x') $P_{\theta,\gamma}\{T \in D|X = x\} = P_{\theta,\gamma}\{T \in D|X = x'\}$ for all sets $D \subseteq \{t = y \times h : x, x' \in y\}$.
- A *statistical model* $(P_{\theta,\gamma})_{(\theta,\gamma)\in\Theta\times\Upsilon}$ is *relatively coarsened at random* or CAR(REL) if for all $\gamma \in \Upsilon$ and all $t = y \times h$, the conditional density $k_\gamma(t|x)$ is constant for $x \in y$.
- An *observation* t is a *random coarsening* of x if for all $\gamma \in \Upsilon$ $k_\gamma(t|x)$ only depends on x through y, that is, is constant for $x \in y$.

Example

(1) Suppose the reference measure, ν, of the distribution of G given $X = x$ in the right-censoring example does not depend on x. Then

$$k_\gamma(t|x) = \begin{cases} \dfrac{P_\gamma([x;\infty])}{\nu([x;\infty])} & \text{if } t = \{x\} \times [x;\infty[\\ h_\gamma(g|x) & \text{if } t =]g;\infty[\times\{g\} \end{cases}. \quad (5)$$

The model is CAR(REL) if this expression is the same for all $x \in y$, that is, if $h_\gamma(g|x)$ does not depend on $x \in]g;\infty[$.

(2) In the heaping example, $k_\gamma(t|x) = h_\gamma(g|x)$, and the model is CAR(REL) if $h_\gamma(g|x)$ only depends on x through what is observed about x. Thus, $h_\gamma(1|x)$ may only depend on the age rounded to the closest year, $\lfloor x + 1/2 \rfloor$, whereas $h_\gamma(0|x)$ may depend on the month as well (subject to the total mass restriction, $h_\gamma(1|x)q_x + h_\gamma(0|x)(1 - q_x) = 1$, where q_x is the mass assigned by the reference measure to the set $\{G = 1\}$ given $X = x$).
The probabilities are CAR(ABS) if $P_\gamma\{G = 1|X = x\} = h_\gamma(1|x)q_x$ only depends on the age rounded to the nearest year. Owing to the total mass restriction also, the probability $P_\gamma\{G = 0|X = x\} = h_\gamma(0|x)(1 - q_x)$ can only depend on age rounded to the nearest year.

If all the probabilities $P_{\theta,\gamma}$ in a model are CAR(ABS), then the statistical model is CAR(REL); the converse is not generally true. However, if the reference measure is a product measure, then the probability $P_{\theta,\gamma}$ is CAR(ABS) if the statistical model is CAR(REL).

If CAR(REL) holds, then the **likelihood** factors

$$L_t(\theta, \gamma) = \varphi_{\theta,\gamma}(t) = E[f_\theta(X)|T = t]k_\gamma(t|x) \quad (6)$$

and the **profile likelihood** of θ is just

$$L_t(\theta) = E[f_\theta(X)|T = t]. \quad (7)$$

If furthermore the probabilities are CAR(ABS), then the profile likelihood simplifies to

$$L_t(\theta) = \begin{cases} f_\theta(x) & \text{if } t = \{x\} \times h \\ \displaystyle\int_y f_\theta(x)d\mu(x) & \text{if } \mu(y) > 0 \end{cases}, \quad (8)$$

where μ is the reference measure of the distribution of X. Notice that (8) does not cover all possible observations t; generally, we have to rely on (7).

Example

(1) In the right-censoring example, the profile likelihood becomes

$$L_t(\theta) = \begin{cases} f_\theta(x) & \text{if } t = \{x\} \times [x;\infty[\\ P_\theta\{X > g\} & \text{if } t =]g;\infty[\times\{g\} \end{cases} \quad (9)$$

if the probabilities are CAR(ABS).

(2) In the heaping example, the likelihood becomes

$$L_t(\theta) = \begin{cases} \displaystyle\int_{[t^*-\frac{1}{24};t^*+\frac{1}{24}[} f_\theta(x)(1 - q_x)\mathrm{d}\mu(x) & \text{if } g = 0 \\ \displaystyle\int_{[t^*-\frac{1}{2};t^*+\frac{1}{2}[} f_\theta(x)q_x\,\mathrm{d}\mu(x) & \text{if } g = 1 \end{cases} \quad (10)$$

where t^* is the observed value of T^*, that is, the observed rounded age, if CAR(REL) holds. If also CAR(ABS) holds, then the q_x and $1 - q_x$ terms disappear.

Ignorability

The concept of coarsening at random is useful because it allows one to discuss when and to what extent incompleteness of observations may be ignored when doing likelihood-based inference. In this section, we will discuss three different kinds of ignorability.

If a statistical model is CAR(REL), the **nuisance parameter**, γ, may be ignored for likelihood-based inference, because of the factorization of the likelihood. **Maximum likelihood** estimation of θ is performed as if the coarsening mechanism, that is, γ, is known.

Example In the heaping example, we see that the profile likelihood of θ is a weighted integral of the density over the observed set y if the model is CAR(REL). Thus, apart from the known weights (q_x and $1 - q_x$), we may ignore the coarsening mechanism when calculating the likelihood.

If the probabilities are CAR(ABS), then the coarsening mechanism can be ignored. The θ-part of the likelihood can be calculated from the marginal model of X treating the mapping $Y(X, G)$ as a function of X alone.

Example In the right-censoring example, if CAR(ABS) holds, the profile likelihood is just the same as if the censoring had been fixed.

It is clear that the second kind of ignorability is stronger than the first. CAR(ABS) implies that we may treat the data as if the coarsening is fixed rather than stochastic. When the model is CAR(REL), we can treat the coarsening as known (but stochastic).

CAR(ABS) will in many applications be a questionable assumption as discussed by Heitjan [2] (for instance, in the case of age heaping). Here the weaker assumption, CAR(REL), may be more reasonable.

For ignorability in a **Bayesian** sense, it suffices that the given observation is coarsened at random and that the parameters θ and γ are a priori independent; this implies that the parameters are a posteriori independent. In particular, the posterior distribution of θ can be calculated without having to calculate the posterior distribution of γ or specifying a **prior distribution** of γ.

Example In the right-censoring example, suppose X is **exponentially distributed** with intensity θ. If the prior distribution of θ is exponential with intensity a, then the posterior distribution given y is

$$\begin{aligned}\pi(\theta, \gamma|y) &\propto L_t(\theta, \gamma) \cdot \pi(\theta, \gamma) \\ &\propto (t^* + a)\exp(-\theta(t^* + a)) \\ &\quad \cdot k_\gamma(t|x)\pi(\gamma) \qquad (11)\end{aligned}$$

where t^* is the observed value of $X \wedge G$, if θ and γ are a priori independent. Thus, the posterior distribution of θ is exponential with intensity $t^* + a$. It should be noted that this result uses that the reference measure is a product measure; without this structure the posterior distribution would involve $E[f_\theta(X)|T = t]$ instead of $\exp(-\theta t^*)$.

References

[1] Gill, R.D., van der Laan, M.L. & Robins, J.M. (1997). Coarsening at random: characterisations, conjectures and counter-examples, in *Proceedings First Seattle Conference on Biostatistics*, D.-Y. Lin, ed. Springer, New York, pp. 255–294.

[2] Heitjan, D.F. (1993). Ignorability and coarse data: some biomedical examples, *Biometrics* **49**, 1099–1109.

[3] Heitjan, D.F. & Rubin, D.B. (1991). Ignorability and coarse data, *The Annals of Statistics* **19**, 2244–2253.

[4] Jacobsen, M. & Keiding, N. (1995). Coarsening at random in general sample space and random censoring in continuous time, *The Annals of Statistics* **23**, 774–786.

[5] Nielsen, S.F. (2000). Relative coarsening at random, *Statistica Neerlandica* **54**, 79–99.

[6] Nielsen, S.F. (2003). Survival analysis with coarsely observed covariates, *Statistics and Operations Research Transactions* **27**, 41–64.

(*See also* **Diggle–Kenward Model for Dropouts**)

SØREN FEODOR NIELSEN

Cochran, William Gemmell

Born: July 15, 1909, in Rutherglen, Scotland.
Died: March 29, 1980, in Orleans, Massachusetts.

William G. Cochran. Reproduced by permission of the Royal Statistical Society

William Cochran was a leading contributor to the British–American school of applied statistics during a period of rapid development of the field across the middle decades of the twentieth century. Coming from a modest family background, he early in life showed academic talents that carried him first to Glasgow University and then to Cambridge University, where he studied mathematics, applied mathematics and statistics. His formal education ended with an M.A. degree from Cambridge because **Frank Yates** made an offer, unusual in the depression year of 1934, for him to join the staff at the Rothamsted Experimental Station, where he carried out major analyses of long-term agricultural experiments, gained much practical experience, and became well known in the field. Responding to an invitation from **George Snedecor** of Iowa State College (now University) at Ames, Cochran emigrated to the US in 1939, carrying with him deep involvement with the extensive improvements in applied statistics then taking place in Britain. The US at the time had relatively little exposure to newer methods and theories, especially those deriving from **R.A. Fisher**. In 1943 and 1944, he worked with the Statistical Research Group at Princeton University, specifically on military problems of naval warfare and bomb efficiency. After the war, he was recruited by **Gertrude Cox** to the newly formed Institute of Statistics in North Carolina, where he organized and headed the graduate program in experimental statistics at North Carolina State College in Raleigh. In 1949, he became chairman of the Department of Biostatistics in the School of Hygiene and Public Health at the Johns Hopkins University, where a shift of his focus from agricultural to medical and biological applications took place. In 1957, Cochran moved to Massachusetts to join the Department of Statistics that Frederick Mosteller had just started at Harvard University, where he remained until his retirement in 1976.

Alongside the teaching and research that were the formal responsibilities of his academic positions, Cochran contributed to many panels, committees, and seminars. He was a leading statistician for nationally and internationally prominent reports on the effects of **radiation** in Hiroshima, the Kinsey Report of human sexual behavior, the **Salk polio vaccine** trials, the pivotal Surgeon General's 1964 Report on **Smoking and Health**, and equality of educational opportunity. He published more than 100 research papers, and important books that became widely used text and reference books through numerous editions, most notably: *Experimental Designs* (with Gertrude Cox) in 1950 (1957, 1992); *Sampling Techniques* in 1953 (1963, 1977), and a substantially revised version of George Snedecor's influential *Statistical Methods* in 1967 (1980, 1989). A posthumous text *Planning and Analysis of Observational Studies* was edited in 1983 by Lincoln Moses and Frederick Mosteller. Cochran's publisher, John Wiley & Sons, put out a thick volume of collected papers in 1982.

Bill Cochran was modest in demeanor, but penetrating and quick in his perceptions and analyses of statistical problems. He was a gifted and much respected teacher and speaker at scientific meetings who could always illustrate theory from a wealth of

applications often drawn from personal experience. Well-known leaders in many areas of statistics are among his 40 or so doctoral students. Above all, he was a distinguished scientist, and was fittingly recognized as such by election to the US National Academy of Sciences in 1974, an honor all too rare among biometricians.

A.P. DEMPSTER

Cochrane Collaboration

For all but the last 100 years, decisions on how to treat patients were almost always based on personal experience, anecdotal case histories, and comparisons between a group of patients who received one treatment with an entirely separate group of patients who did not receive that treatment. These processes, although subject to many biases, are still in use today but ways to minimize these biases are now available, accepted, and more easily adopted. Among these is the use of the randomized trial (*see* **Clinical Trials, Overview**) as a means of providing more reliable estimates of the relative effects of interventions, since the only difference between the patients in the groups being compared in a randomized trial will be that of most interest: namely, the interventions under investigation.

However, in part because of chance variations in the types of patients allocated to the different interventions in the randomized trial, the results of a single trial will rarely be sufficient. Most trials are too small and their results are not sufficiently robust against the effects of chance. In addition, small trials might be too focused on a particular type of patient to provide a result that can be either easily or reliably generalized to future patients. Added to this, the amount of information about health care, including that coming from individual randomized trials, is now overwhelming. Vast amounts of information are now readily available in journals, books, magazines, the media and, especially in recent years, on the Internet. However, people making decisions about health care – including patients, their carers, health care professionals, policy makers, and managers – need high quality information and, unfortunately, much of what is available is of poor quality.

To help identify which forms of health care work, which do not, and which are even harmful, results from similar randomized trials need to be brought together. Trials need to be assessed and those that are good enough can be combined to produce both a more statistically reliable result and one that can be more easily applied in other settings. This combination of trials needs to be done in as reliable a way as possible. It needs to be systematic. A systematic review uses a predefined, explicit methodology. The methods used include steps to minimize bias in all parts of the process: identifying relevant studies, selecting them for inclusion, and collecting and combining their data. Studies should be sought regardless of their results.

A systematic review does not need to contain a statistical synthesis of the results from the included studies. This might be impossible if the designs of the studies are too different for an averaging of their results to be meaningful or if the outcomes measured are not sufficiently similar. If the results of the individual studies are combined to produce an overall statistic, this is usually called a **meta-analysis** (*see* **Meta-analysis of Clinical Trials**). A meta-analysis can also be done without a systematic review, simply by combining the results from more than one trial. However, although such a meta-analysis will have greater mathematical precision than an analysis of any one of the component trials, it will be subject to any biases that arise from the study selection process, and may produce a mathematically precise, but clinically misleading, result.

The Cochrane Collaboration

The Cochrane Collaboration is the largest organization in the world engaged in the production and maintenance of systematic reviews. It has received worldwide support in its efforts to do something about the problems outlined above, by making systematic reviews accessible to people making decisions about health care (`www.cochrane.org`). The Collaboration aims to help people make well-informed decisions by preparing, maintaining, and promoting the accessibility of systematic reviews of the effects of interventions in all areas of health care. These reviews bring together the relevant research findings on a particular topic, synthesize this evidence, and then present them in a standard, structured way. One of their most important attributes is that they are periodically updated to take account of new studies and other new information, to help people be confident that the systematic reviews are sufficiently current to be useful in making decisions about health care.

The Cochrane Collaboration was established in 1993, founded on ideas and ideals that stem from earlier times. In October 1992, Iain Chalmers, Kay Dickersin, and Thomas Chalmers wrote an editorial in the *British Medical Journal* [1] that began with the following quote from the British epidemiologist, Archie **Cochrane**, published in 1972:

> It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, updated periodically, of all relevant randomised controlled trials. [3]

This editorial was published at the time of the opening of the first Cochrane Centre in Oxford, United Kingdom. This Centre, was funded by the National Health Service Research and Development Programme in the United Kingdom "to facilitate and co-ordinate the preparation and maintenance of systematic reviews of randomized controlled trials of healthcare". However, there was a clear need for the work to extend beyond this Centre, the United Kingdom and in some circumstances randomized trials.

A year after the UK Cochrane Centre opened, 77 people from 19 countries gathered at what was to become the first Cochrane Collaboration and established *The Cochrane Collaboration* as an international organization. There have been annual Cochrane Colloquia since then, with the most recent being in Barcelona, Spain, in October 2003, attended by more than 1000 people from 45 countries.

The Cochrane Collaboration is supported by hundreds of organizations from around the world, including health service providers, research funding agencies, departments of health, international organizations, industries, and universities. There are currently more than 10 000 people contributing to the work of The Cochrane Collaboration from over 80 countries, and this involvement continues to grow. The number of people involved has increased by about 20% year on year for each of the five years to 2004. The importance of involving people from low and middle income countries in the work of The Cochrane Collaboration is well recognized. This is reflected by the efforts of the centers based in these countries and the steady increase in the number of people actively involved in the preparation and maintenance of Cochrane reviews, from about 300 in the year 2000 to more than 700 in 2003.

The Cochrane Collaboration has 10 guiding principles:

- Collaboration, by internally and externally fostering good communications, open decision-making and teamwork.
- Building on the enthusiasm of individuals, by involving and supporting people of different skills and backgrounds.
- Avoiding duplication, by good management and coordination to maximize economy of effort.
- Minimizing **bias**, through a variety of approaches such as scientific rigor, ensuring broad participation, and avoiding conflicts of interest.
- Keeping up-to-date, by a commitment to ensure that Cochrane Reviews are maintained through identification and incorporation of new evidence.
- Striving for relevance, by promoting the assessment of health care interventions using outcomes that matter to people making choices in health care.
- Promoting access, by wide dissemination of the outputs of The Cochrane Collaboration, taking advantage of strategic alliances, and by promoting appropriate prices, content, and media to meet the needs of users worldwide.
- Ensuring quality, by being open and responsive to criticism, applying advances in methodology, and developing systems for quality improvement.
- Continuity, by ensuring that responsibility for reviews, editorial processes, and key functions is maintained and renewed.
- Enabling wide participation in the work of The Cochrane Collaboration by reducing barriers to contributing and by encouraging diversity.

The work of preparing and maintaining Cochrane reviews is done by the reviewers, of whom there are more than 4000 in 2004. Very few of these are paid to work on their reviews and the main motivation is a desire to answer reliably a question about the relative effects of interventions for people with particular conditions. The reviewers are supported by 50 Cochrane Collaborative Review Groups, who are responsible for reviews within particular areas of health and collectively providing a home for reviews in all aspects of health care. These Groups organize the refereeing of the drafts for Cochrane reviews, and the protocols that precede them, and the editorial teams in these Groups decide whether or not a Cochrane review should be published. As far as possible, they work with the reviewers to ensure that this happens, and the decision that a Cochrane review will be published depends on its quality not its findings. This is unlike the publication process elsewhere in the health care literature where journals rarely help authors with their reports and where the decision about whether or not a paper

will be published will often be dependent on the importance given to the paper, in the light of its findings.

The Collaborative Reviews Groups are based around the world and some have editorial bases in more than one country. There are also Cochrane Methods Groups, with expertise in relevant areas of methodology; Fields or Networks, with broad areas of interest and expertise spanning the scope of many Review Groups; and a Consumer Network helping to promote the interests of users of health care. The work of these Cochrane entities, and their members, is supported by 12 regional Cochrane Centres: Australasian, Brazilian, Canadian, Chinese, Dutch, German, IberoAmerican, Italian, Nordic, South African, UK and USA. The Cochrane Collaboration Steering Group, containing elected members from the different types of entity, is responsible for setting Collaboration-wide policy and, by working with the entities, the implementation of the Collaboration's strategic plan.

One of the important ways in which activity within The Cochrane Collaboration is supported is the Collaboration's Information Management System (IMS). This was developed initially by Update Software, the original publishing partner of The Cochrane Collaboration, before responsibility was transferred to the Nordic Cochrane Centre in Copenhagen, Denmark where much further development has taken place over the last few years. The IMS comprises the set of software tools used to prepare and maintain Cochrane reviews and to submit these for publication, and also to describe the work of each entity and to manage contact details of their members. For the Collaboration's first decade, the IMS worked mainly as standard software running on local computers, with reviewers sharing their files by disk or email attachment. As the Collaboration grew and the number of reviews and the vital task of keeping these up-to-date got bigger, a better way to share these documents and information was needed. In 2001, a software needs assessment survey was conducted. Nearly, all Cochrane entities and almost 500 individuals responded. The results were influential in planning the new IMS, which is being introduced from 2004 to 2006 and which increases the ability of people in The Cochrane Collaboration to work together by providing a central computer approach to the storage of documents such as the draft versions of Cochrane reviews.

The Collaboration has grown quickly through its first decade. Although there is a great deal of work that remains to be done, much has been accomplished already. Cochrane reviews are published in *The Cochrane Database of Systematic Reviews* (*CDSR*). As of mid-2004, this contains the full text of more than 2000 complete Cochrane reviews, each of which will be kept up-to-date as new evidence and information accumulates. There are a further 1500 published protocols for reviews in progress. These set out how the reviews will be done and provide an explicit description of the methods to be followed. The growth in Cochrane reviews is well illustrated by the following milestones. The first issue of CDSR, at the beginning of 1995, included 36 Cochrane reviews; there were 500 in 1999, 1000 in 2001, and 2000 in April 2004. Hundreds of newly completed reviews and protocols are added each year and a few hundred existing reviews are updated so substantively that they can be considered to be the equivalent of new reviews, and there are currently several hundred Cochrane reviews at earlier stages than the published protocol.

The Cochrane Database of Systematic Reviews is available on the Internet and on CD-ROM as part of *The Cochrane Library*. This is published by John Wiley and Sons Ltd and is available on a subscription basis. The establishment of national contracts means that *The Cochrane Library* is currently free at the point of use to everyone in Australia, Denmark, England, Finland, Ireland, Northern Ireland, Norway, and Wales. More countries are being added to this list each year.

The output of The Cochrane Collaboration also includes the *Cochrane Central Register of Controlled Trials (CENTRAL)*, the *Cochrane Database of Methodology Reviews* and the *Cochrane Methodology Register*. All of these are unique resources. In 1993, when the Collaboration was established, less than 20 000 reports of randomized trials could be found easily in MEDLINE, and one of the main tasks facing the Collaboration was the need to identify and make accessible information on reports of trials that might be suitable for inclusion in Cochrane reviews. It has done this through extensive programs of the hand searching of journals (in which a journal is checked from cover to cover to look for relevant reports) and of electronic searching of bibliographic databases such as MEDLINE and EMBASE. Suitable records are

then added to CENTRAL, with coordination by the US Cochrane Centre in Rhode Island, USA [4, 5]. By 2004, CENTRAL contained records for more than 400 000 reports of randomized (or possibly randomized) trials, many of which are not included in any other electronic database. The Cochrane Database of Methodology Reviews contains the full text for Cochrane methodology reviews, which are systematic reviews of issues relevant to the conduct of reviews of health care interventions or evaluations of health care more generally. Currently (mid-2004), there are 10 full Cochrane methodology reviews and published protocols for 8 more. The Cochrane Methodology Register, to a large extent, provides the raw material for the Cochrane methodology reviews, containing more than 5000 records, for example, records for reports of research, and also for ongoing, unpublished research, into the control of bias in health care evaluation.

Over the next few years, The Cochrane Collaboration will strive to ensure that its work is sustainable. Even with more than 4000 Cochrane reviews already underway, and results available from 2000 of these, there is still a large amount of work to be done. A recent estimate is that approximately 10 000 systematic reviews are needed to cover all health care interventions that have already been investigated in controlled trials, and such reviews would need to be assessed and, if necessary, updated at the rate of 5000 per year. If the growth in The Cochrane Collaboration continues at the pace of the last few years, this target will be reached within the coming 10 years. However, this will require continuing and evolving partnership and collaboration. The Cochrane Collaboration will need to continue to attract and support the wide variety of people who contribute to its work. It will also need to work together with funders and with providers of health care to ensure that the resources needed for the work grow and the output of the work is accessible to people making decisions about health care [2].

References

[1] Chalmers, I., Dickersin, K. & Chalmers, T.C. (1992). Getting to grips with Archie Cochrane's agenda, *British Medical Journal* **305**, 786–788.

[2] Clarke, M. & Langhorne, P. (2001). Revisiting the Cochrane Collaboration, *British Medical Journal* **323**, 821.

[3] Cochrane, A.L. (1979). 1931–1971: a critical review, with particular reference to the medical profession, in *Medicines for the Year 2000*. Office of Health Economics, London, (pp. 1–11).

[4] Dickersin, K., Manheimer, E., Wieland, S., Robinson, K.A., Lefebvre, C. & McDonald, S. (2002). Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials, *Evaluation and the Health Professions* **25**, 38–64.

[5] Lefebvre, C. & Clarke, M.J. (2001). Identifying randomised trials, in *Systematic Reviews in Health Care: Meta-analysis in Context*, 2nd Ed., M. Egger, G. Davey Smith & D. Altman eds. BMJ Books, London, (pp. 69–86).

(*See also* **Evidence-based Medicine**)

MIKE CLARKE

Cochrane Lectures

The legacy of Archie **Cochrane** in **epidemiology**, **clinical trials**, social medicine, and **health services research** is marked by at least two series of annual lectures in the United Kingdom.

The first of these – "The Cochrane Lecture" – was initiated two years after Cochrane's death, by the Society for Social Medicine at its annual meeting in September, a gathering that Cochrane attended regularly. The Cochrane lecturers between 1990 and 2004 have been as follows:

1990	Peter Elwood	"Archie Cochrane"
1991	Donald Acheson	"Health, cities and the future"
1992	Iain Chalmers	"Getting to grips with Archie Cochrane's agenda"
1993	Klim McPherson	"The best and the enemy of the good: assessing the role of patient choice in medical decision making"
1994	Stuart Kilpatrick	"Tuberculosis – yesterday and today"
1995	Kay Dickersin	"Consumer involvement in research"
1996	Alan Williams	"All cost-effective treatments should be free!"
1997	Julian Tudor Hart	"What sorts of evidence do we need for evidence-based medicine"
1998	Ann Oakley	"Social science and the experimenting society"
1999	Richard Lilford	"What use are qualitative data when decisions have to be made?"
2000	Nick Black	"Evidence, policy, and evidence-based policy"
2001	Richard Peto	"Halving premature death"
2002	Catherine Peckham	"Science to policy: HIV and other fetal and childhood infections"
2003	Mildred Blaxter	"Fish in water: social capital and the qualitative researcher"
2004	George Davey Smith	"Randomised by (your) god: robust evidence from an observational study design"

A second annual lecture series – "The Cochrane 'Effectiveness and Efficiency' Anniversary Lecture" – was established to mark the anniversary of Cochrane's seminal Rock Carling lecture, "Effectiveness and Efficiency: random reflections on health services", which he delivered in Edinburgh on March 20, 1972. The Nuffield Provincial Hospitals Trust (which published Cochrane's lecture) provided initial funding support for the lecture series, which was established by Iain Chalmers, under the official aegis of Green College, Oxford, a beneficiary of Cochrane's estate. Cochrane lecturers between 1993 and 2004 have been:

1993	Walter Holland	"Epidemiology, research, and how it can contribute to the development of health policy"
1994	William Silverman	"Effectiveness and efficiency... and subjective choice"
1995	David Sackett	"On the need for evidence-based health care"
1996	Richard Doll	"Cochrane and the benefits of wine"
1997	Peter Elwood	"Cochrane and the benefits of aspirin"
1998	Iain Chalmers	"Lord Rayleigh's injunction"
1999	Chris Silagy	"The challenge of the post-Cochrane agenda: consumers and evidence"
2000	Richard Peto	"Getting large-scale randomized evidence"
2001	Valerie Beral	"The causes of breast cancer"
2002	Rory Collins	"LDL cholesterol: from observational to randomized evidence"
2003	Sarah Lewington	"Doubling the importance of blood pressure: the Prospective Studies Collaboration"

2004	Peter Rothwell	"Effectiveness and efficiency in the prevention of the stroke"

(*See also* **Cochrane Collaboration**)

IAIN CHALMERS

Cochrane, Archibald ('Archie') Leman

Born: January 12, 1909, in Galashiels, UK.
Died: June 18, 1988, in Dorset, UK.

Archibald Leman Cochrane (1909–1988) was Director of the **Medical Research Council** (MRC) Epidemiology Unit in Cardiff from 1969 to 1974. Prior to this, from 1960 he had been honorary director of the unit and had also held the David Davies chair of tuberculosis and chest diseases in the Welsh National School of Medicine, later the University of Wales College of Medicine. Amongst the many honors he received and significant posts he occupied, the founding Presidency of the Faculty of Community Medicine from 1971 to 1973 ranks high.

Following the study of natural sciences in Cambridge and psychoanalysis in Vienna, he completed his medical studies in University College London and qualified M.B. in 1938. After service in the Royal Army Medical Corps he studied Public Health in the London School of Hygiene and Tropical Medicine.

Archie's career really began to develop when he spent eighteen months studying the epidemiology of tuberculosis at the Henry Phipps Institute in Philadelphia on a Rockefeller Fellowship. Following this, he worked on coal workers' pneumoconiosis and tuberculosis in the MRC Pneumoconiosis Research Unit in South Wales from 1947 to 1960. There he developed field epidemiology methods and the study of total defined communities (*see* **Population-based Study**), rather than just samples of working men. He became almost obsessional about practical aspects of field epidemiology: the representativeness of population samples; the response rate of subjects selected for study; the completeness of the follow-up in a **cohort study**; and the reproducibility of the measurements made. He became especially concerned about reproducibility in the reading of chest X-rays and he introduced the use of "standard films" to reduce observer error in the grading of pneumoconiosis and other chest lesions (*see* **Observer Reliability and Agreement**). In short, he was acutely concerned with every aspect of research methodology and he himself repeatedly demonstrated the potential of epidemiology as a highly accurate quantitative science. He argued that in its relevance to medical and social problems, epidemiology is second to no other research strategy, and he extended his own work to the study of a wide range of conditions, in addition to pneumoconiosis and tuberculosis, including iron deficiency anemia, arthritis, glaucoma, hypertension, bronchitis, and cardiovascular disease.

Archie's greatest love, however, was for the randomized controlled trial (*see* **Clinical Trials, Overview**). Teaching by **Sir Austin Bradford Hill** led him to comment later that "this innovation ... offered clinical medicine an experimental approach to the validation of its practices and treatments" [4]. Although Archie himself conducted very few trials, his encouragement was seminal in many randomized controlled trials in a wide range of medical situations. One trial of his, however, on the treatment of famine edema in men in a prisoner of war camp in Salonica, will certainly go down in history. This was eventually published under the title "Sickness in Salonica: my first, worst and most successful clinical trial" [3].

Archie saw an especially valuable role for the randomized controlled trial in the evaluation of clinical procedures, and the trials he stimulated on the best place of treatment and the optimum length of stay in hospital led to him being invited to give the Rock Carling lecture in 1970. This was later published as a monograph under the title: *Effectiveness and Efficiency – Random Reflections on Health Services* [1]. The provocative and challenging ideas he developed in this had a widespread international effect in stimulating critical evaluation of all aspects of the National Health Service in the UK, and health services in many other countries (*see* **Health Services Research, Overview**).

Perhaps the contribution of Archie to medical research that has had the widest effect, and is likely to have the most far, and long reaching effects, originated from a challenge he made in 1979:

> It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials [2].

Iain Chalmers, with others, took this challenge by Cochrane and turned it into the world-wide ongoing **Cochrane Collaboration**. By early 1997 there were 13 Cochrane Centers across the world which organize the searching of all the medical literature, the identification of randomized controlled trials, the preparation of databases (*see* **Database Systems**), and the preparation, publication and maintenance

of systematic reviews of the effects of health care interventions (*see* **Meta-analysis of Clinical Trials**).

These centers are a most fitting memorial to Cochrane. His inspired vision, together with the enthusiasm and sweat of Chalmers, has almost limitless potential benefit. A further development that has arisen from this initiative – the definition and promotion of **evidence based medicine** – is likely to have a profound and widespread effect on all clinical practice.

References

[1] Cochrane, A.L. (1972). *Effectiveness and Efficiency-Random Reflections on Health Services*. Nuffield Provincial Hospitals Trust, London. (Reprinted in 1989 in association with the *British Medical Journal*.).

[2] Cochrane, A.L. (1979). 1931–1971: a critical review, with particular reference to the medical profession, in *Medicines for the Year 2000*. Office of Health Economics, London, pp. 1–11.

[3] Cochrane, A.L. (1984). Sickness in Salonica: my first, worst and most successful clinical trial, *British Medical Journal* **289**, 1726–1727.

[4] Cochrane, A.L. & Blythe, M. (1989). *One Man's Medicine*. The Memoir Club, British Medical Association Press, London, p. 157.

P. ELWOOD & C. HUGHES

Cohabitation

Cohabitation has increasingly come to be the term used to describe the marital status of couples who are unmarried sexual partners and share the same household. With the rise in cohabitation, shorthand for "unmarried cohabitation", that has occurred in developed countries, the full extent of coresidential heterosexual partnerships are no longer captured by marriage data. Moreover, the rise in extramarital fertility that has also occurred across developed societies in recent decades is related to developments in cohabitation.

Men and women living together outside marriage is not a new phenomenon. Prior to the 1970s it was largely statistically invisible and probably socially invisible outside the local community or milieu. In some countries there were subgroups that were probably more prone to cohabitation than others: the very poor; those whose marriages had broken up but were unable to obtain a divorce, as there was no such legislation, or it was more stringent than nowadays or it was very expensive to obtain a divorce; certain groups of rural dwellers; and groups ideologically opposed to marriage. The form of cohabitation that came to the fore during the 1960s in Sweden and Denmark, and the 1970s in other Northern and Western European countries, North America, and Australia is new, and could be aptly termed "nubile cohabitation", whereby young people predominantly in their 20s and early 30s live together either as a prelude to, or as an alternative to, marriage. Additionally, with the growth in divorce, "postmarital cohabitation" is also likely to have become more prevalent, with the divorced cohabiting either in preference to, or as a prelude to, remarriage. In many data sources it is difficult to distinguish between "nubile" and "postmarital" cohabitation. The increased prevalence of cohabiting unions lies behind much of the decline in first marriage and remarriage rates that have occurred in recent decades.

To date, data on cohabitation tend to be scarce and generally emanate from surveys which can make any comparative analyses problematic, as sample sizes, coverage, and definitions may vary. Notwithstanding, in developed countries with a **time series** of data it is clear that there have been increases in the proportions of women cohabiting, particularly in their 20s. The peak ages for cohabitation tend to be the early 20s, and cohabitation at older ages, particularly in the 30s, is less common. There is some evidence that cohabiting unions tend to be more fragile and less fertile than marriages. What emerges from existing surveys is that cohabitation is a relatively youthful practice and marriage has not been rejected permanently on a wide scale, as even in Sweden and Denmark, where the practice is more long-standing, the majority of unions amongst women in their 30s are legal marital unions.

KATHLEEN KIERNAN

Coherence Between Time Series

Coherence is a measure of the strength of association between **time series**; it is a time series analog of the standard correlation coefficient. Association between time series is a more complex concept than that between scalar characteristics, since the time series data structure is much richer; for example, the association may include a leading or lagging relationship. Two examples are shown in Figure 1. Figure 1(a) shows daily mortality and SO_2 time series in London during the winter months of 1958, with an obvious question whether the pollution was in any way affecting mortality. The strong peaks toward the end of both series are suggestive of a strong association; see [8] for the data set and a detailed analysis, and [9] for a related work. Figure 1(b) shows traces of a person's respiration and heart-rate variability. Since a normal heart responds to the respiration cycles, while an abnormal heart does not, the correlation analysis of the two time series carries a high diagnostic value as to the state of health of the person's heart; see, for example, [1].

We first define some notations and terminology theoretically, then comment on how the quantities are estimated given some finite-length time series data. Given two stationary time series X_t and Y_t, for $t = 0, \pm 1, \ldots$, the autocovariance functions are given by

$$C_x(m) = \mathrm{cov}(X_t, X_{t+m}), \tag{1}$$

$$C_y(m) = \mathrm{cov}(Y_t, Y_{t+m}), \quad m = 0, \pm 1, \ldots. \tag{2}$$

The association between the two time series may be expressed by the cross-covariance function

$$C_{xy}(m) = \mathrm{cov}(X_t, Y_{t+m}), \quad m = 0, \pm 1, \ldots. \tag{3}$$

The cross-correlation function is defined similarly, so the measure of dependencies between X_t and Y_t is no longer a single number, but a function of lag m, which means that potentially it can carry a lot of information about the dependency between X_t and Y_t.

The spectra and cross-spectra of X_t and Y_t (*see* **Spectral Analysis**) are the Fourier transforms of the above quantities, i.e.

$$f_x(\omega) = \sum_{m=-\infty}^{\infty} C_x(m) \exp(-2\pi i \omega m), \tag{4}$$

$$f_y(\omega) = \sum_{m=-\infty}^{\infty} C_y(m) \exp(-2\pi i \omega m), \tag{5}$$

$$f_{xy}(\omega) = \sum_{m=-\infty}^{\infty} C_{xy}(m) \exp(-2\pi i \omega m). \tag{6}$$

The functions $f_x(\omega)$ and $f_y(\omega)$ are called the spectral density functions of X_t and Y_t, and $f_{xy}(\omega)$ the cross-spectrum between X_t and Y_t (*see* **Multiple Time Series**). Since C_x and C_y are symmetric, $f_x(\omega)$ and $f_y(\omega)$ are real functions, but $f_{xy}(\omega)$ is, in general, a complex function. These frequency-domain objects are, in a mathematical sense, equivalent to their time-domain analogs; however, they carry different

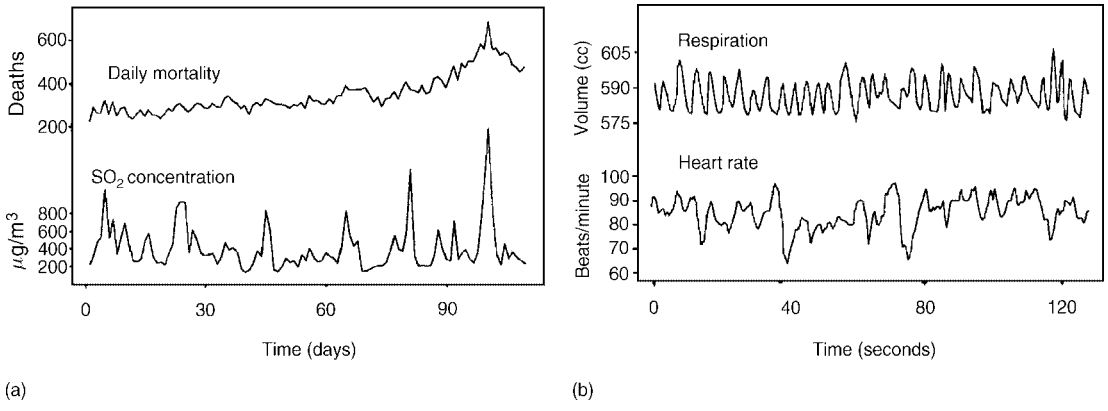

Figure 1 Examples of bivariate time series. (a) Pollution and respiratory mortality in London during winter 1958. (b) Respiration time series in terms of thoracic volume and the heart-rate time series

physical interpretations. To aid this interpretation, define the (squared) coherence as

$$\rho^2(\omega) = \frac{|f_{xy}(\omega)|^2}{f_x(\omega) f_y(\omega)}, \tag{7}$$

and the phase spectrum

$$\phi(\omega) = \tan^{-1}\left\{\frac{\mathrm{Im}[f_{xy}(\omega)]}{\mathrm{Re}[f_{xy}(\omega)]}\right\},$$

where Re $[\cdot]$ and Im $[\cdot]$ are the real and imaginary parts of a complex quantity.

Now the interpretations. It is sometimes natural to think of a time series X_t as a waveform, composed of different frequencies. The spectrum $f_x(\omega)$ is the variance or the power of the component of time series X_t at frequency ω. The coherence $\rho^2(\omega)$ is the proportion of variability of the Y_t component at frequency ω explained by the corresponding X_t component and the $\phi(\omega)$ is the phase shift in the X_t component. The interpretation of $\rho^2(\omega)$ is especially appealing as it corresponds to the usual R^2 interpretation in regression analysis. Let us consider some simple theoretical examples.

Example 1

Let $Y_t = -X_t$; then intuitively $\rho^2(\omega) = 1$, which means that every frequency component of Y_t is totally determined by X_t or there is no noise. (Note the analogy for scalar random variables: if $Y = -X$, then the squared correlation is $\rho^2 = 1$.) It may be shown that $\phi(\omega) = 1/2$, which means that the Y_t components are out of phase by half a cycle from the corresponding X_t components. For a theoretical treatment of this example, see [3, p. 213].

Example 2

Let $Y_t = X_{t-1} + Z_t$, where X_t and Z_t are independent, i.e. Y_t is a delayed version of X_t plus some noise. Let $f_z(\omega)$ be the spectrum of Z_t. Then the coherence between X_t and Y_t is

$$\rho^2(\omega) = \frac{f_x(\omega)}{f_x(\omega) + f_z(\omega)}. \tag{8}$$

(Note the analogy for scalar random variables: if $Y = X + Z$, then the squared correlation is $\rho^2 = \sigma_x^2/(\sigma_x^2 + \sigma_z^2)$.) The interpretation is immediate and useful in general: in frequencies where the noise level is low, i.e. $f_z(\omega)$ is small, the correlation between X_t and Y_t is high, and vice versa. It may be shown that the phase shift is $\phi(\omega) = \omega$. This means that the low-frequency components of X_t and Y_t (for example, the trend or large swings in the series) move in phase, but the faster components are out of phase.

Before we discuss a real data example, we remark that there is a large literature on the spectral estimation based on finite-length time series. Priestley [7] and Brillinger [2] are two main references in the area. An older reference, Jenkins & Watts [4, Chapter 9], gives the techniques and some detailed examples of coherence analysis. The concept of smoothing is central in the nonparametric estimation of the spectra and cross-spectra, where the amount of smoothing is traditionally left subjectively for the user. An objective and automatic smoothing of the spectrum was proposed, for example, in [11] and [6]. Pawitan [5] describes a fully automatic estimation of the cross-spectrum.

Standard statistical packages such as SAS and BMDP have procedures (`PROC SPECTRA` and `BMDP 1T`, respectively) for the traditional estimation of the spectra and cross-spectra, as well as the coherence and phase spectrum (*see* **Software, Biostatistical**). Venables & Ripley [10] show in Chapter 14 some examples of coherence analysis using the S-PLUS statistical language.

Example 3

This example shows the power of coherence analysis using real data on heart-rate variability. The estimates shown in Figure 2 are fully automatic estimates based on the time series in Figure 1(b); see [5] for a detailed description of the estimation technique. The coherence between respiration and heart rate shows that the heart responds to the natural respiration cycle between 0.25–0.5 cycles per second or between 2–4 seconds per cycle. This is expected from a healthy heart and has been shown to be mediated by the parasympathetic nervous system [1]. The phase spectrum indicates that the cardiac response is in phase with respiration if the respiration is slow enough at around 0.25 cycles per second, but is increasingly out of phase with faster respiration due to delay in the cardiac response. We note that none of these phenomena is apparent from the spectral analysis of the individual time series.

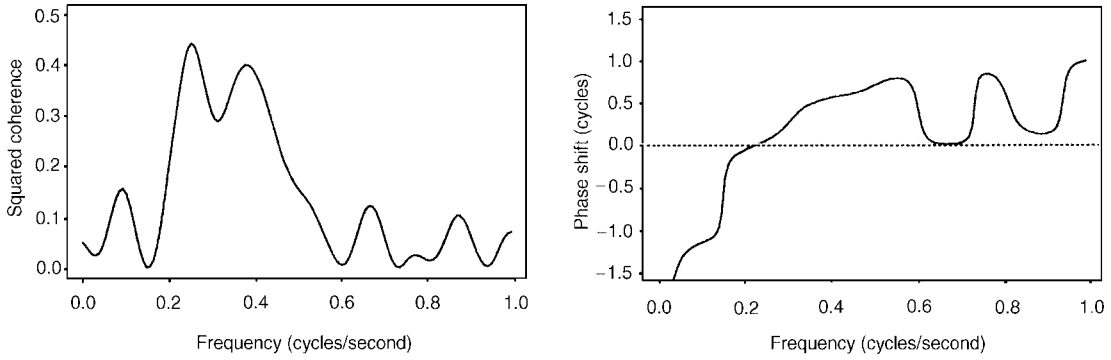

Figure 2 Coherence analysis of heart-rate variability. (a) Heart responds to the respiration cycle at frequencies 0.25–0.5 cycles per second. (b) Cardiac response to slow respiration at around 0.25 cycles per second is almost immediate

References

[1] Appel, M.L., Berger, R.D., Saul, P.S., Smith, J.M. & Cohen, R.J. (1989). Beat to beat variability in cardiovascular variables: noise or music?, *Journal of the American College of Cardiology* **14**, 1139–1148.

[2] Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. Holden Day, San Francisco.

[3] Diggle, P. (1990). *Time Series: A Biostatistical Introduction*. Oxford Science Publications, Oxford.

[4] Jenkins, G.M. & Watts, D. (1968). *Spectral Analysis and Its Applications*. Holden Day, San Francisco.

[5] Pawitan, Y. (1996). Automatic estimation of the cross-spectrum of a bivariate time series, *Biometrika* **83**, 419–432.

[6] Pawitan, Y. & O'Sullivan, F. (1994). Nonparametric spectral density estimation using penalized Whittle likelihood, *Journal of the American Statistical Association* **89**, 600–610.

[7] Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Academic Press, New York.

[8] Shumway, R.H. (1988). *Applied Statistical Time Series Analysis*. Prentice-Hall, Englewood Cliffs.

[9] Shumway, R.H., Azari, R. & Pawitan, Y. (1988). Modelling mortality fluctuations in Los Angeles as functions of pollution and weather effects, *Environmental Research* **45**, 224–241.

[10] Venables, W.N. & Ripley, B.D. (1994). *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York.

[11] Wahba, G. (1980). Automatic smoothing of the log-periodogram, *Journal of the American Statistical Association* **75**, 122–132.

YUDI PAWITAN

Cohort Study, Historical

In **cohort study** design, participants are enrolled, often with selection based on one or more exposures of interest, and observed over time for disease incidence or mortality. Cohort studies are further classified by the timing of the enrollment and follow-up in relation to actual calendar time. Cohort studies involving identification of participants and follow-up into the future are termed "prospective cohort studies", while those involving follow-up and events in the past are referred to as "historical cohort studies". Other designations used for historical cohort design include "retrospective cohort study" and "nonconcurrent cohort study". A study may be initiated as an historical cohort study, but subsequently follow-up could be maintained into the future. The study of bladder cancer in British chemical industry workers, conducted by Case et al. [1], represents one of the first comprehensive applications of historical cohort design. Beginning the study in the 1950s, Case et al. traced a group of chemical industry workers employed subsequent to 1920 and showed a clear excess of deaths from bladder cancer, which was attributed to exposures to anilines. The researchers compared the data with expected mortality, on the basis of the experience of males in the general population, over the same time interval. Frost [3] and others had previously applied a similar method in studying infectious diseases.

Historical cohort design can be applied if records are available for the retrospective identification of study participants, the classification of the exposure(s) of interest, and the follow-up of the participants for the relevant outcomes. For example, historical cohort design has been widely applied in investigating the effects of specific occupations and industries (*see* **Occupational Epidemiology**), because of the availability of records appropriate to these purposes. Employment records can be used to identify cohort members and, in some instances, to estimate exposures; follow-up for mortality can be accomplished using pension records and national death registries. In the absence of an internal reference population, comparison has been made in many studies to mortality in the general population. For example, Samet et al. [6] conducted an historical cohort study of lung cancer mortality in underground uranium miners who had worked in the state of New Mexico, USA, using industry and health clinic records to define the cohort. The investigation began in 1978; the records were used to identify men who had worked for at least 12 months in an underground uranium mine in New Mexico by December 31, 1976. Follow-up for mortality was accomplished by using listings of deaths in the state and by matching the study roster against two national databases, the files of the Social Security Administration and the National Death Index. The initial report of study findings involved follow-up through 1985; follow-up has continued.

A variety of approaches may be used to estimate exposures in historical cohort studies of occupational groups, depending on the nature of the exposure(s) and the extent and quality of data available on exposures [2]. The occupation or industry may serve as a surrogate for associated exposures, and job information may be used in a job–exposure matrix to link occupation and industry pairs to specific exposures. General systems have been created for this purpose and the same approach has been tailored to specific occupational groups. If data are available on concentrations of workplace contaminants, it may be possible to calculate estimates of exposure for specific study participants by combining the concentration data with the time spent in jobs involving the exposure. For example, in the study of New Mexico uranium miners, information on concentrations of radon progeny in specific mines was used in combination with data on time spent in the mines to estimate exposures to radon progeny [6].

Historical cohort design has the advantage of rapidity of execution. While the task of conducting an historical cohort study may be formidable, the investigator does not need to wait for follow-up time to accumulate, as in a prospective cohort study. Costs tend to be modest as a result, and many historical cohort studies can be completed in only a few years, depending on the status and complexity of the involved databases.

Historical cohort studies, similar to prospective cohort studies, are subject to limitation by information bias, selection bias (*see* **Selection Bias**; **Bias in Cohort Studies**) and confounding (*see* **Bias in Observational Studies**). The limitations of the design primarily reflect the availability of the relevant historical data to ascertain the cohort participants and to estimate exposures of interest and potential

confounding and modifying factors. Information bias is a particular concern. There is a strong potential for exposure misclassification (*see* **Misclassification Error**) and for bias from uncontrolled confounding. Because databases used to estimate exposures and covariates may be most complete in the more recent years, there is a potential for complex time-dependent exposure misclassification. The design may be further compromised by losses to follow-up and misclassification of the health outcome(s), because of reliance on historical records and death certificate assignment of cause of death. Historical cohort studies of workers involving mortality as the outcome measure are subject to a bias that has been widely termed "the healthy worker effect" [4]. Employed persons tend to be healthier than unemployed persons and consequently fewer deaths than expected are typically observed. The healthy worker effect has been characterized as a form of selection bias [2], although it can also be viewed as a reflection of confounding from uncontrolled differences between employed and unemployed persons. Selection bias could also be introduced in defining a cohort on the basis of incomplete records; for example, the findings of an historical cohort study could be affected by selection bias if records used to define the cohort were more complete in the most recent years of exposure and exposures had declined over the period of eligibility. Furthermore, investigating disease incidence may not be possible using historical cohort design, unless special mechanisms have been put in place to track outcomes of interest, or unless it is possible to match records against an incidence registry for the disease(s) of interest, e.g. a cancer registry.

Historical cohort design may be strengthened by the addition of complementary, nested studies that involve additional collection of data on exposures, confounders, or modifiers from samples of the cohort members. Using case-based sampling methods (*see* **Case–Cohort Study**), more detailed data may be obtained for participants who have developed the outcome of interest and for an appropriate sample of controls. For example, in the study of New Mexico uranium miners, the effect of silicosis (a chronic respiratory disease arising from silica dust exposure) on lung cancer risk was assessed using a nested case-control design [5]. Chest radiographs were interpreted for lung cancer cases ($N = 65$) and for controls ($N = 216$) sampled from a total cohort of 3400 miners.

Historical cohort design has proven to be useful for studying the effects of occupational and other exposures. We may see increasing application of this design as implementation of disease registries expands and large administrative databases developed by health care organizations are used for research on health care outcomes and effectiveness (*see* **Administrative Databases**).

References

[1] Case, R.A.M., Hosker, M.E., McDonald, D.B. & Pearson, J.T. (1954). Tumors of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry, *British Journal of Industrial Medicine* **11**, 75–104.

[2] Checkoway, H., Pearce, N.E. & Crawford, D.J. (1989). *Research Methods in Occupational Epidemiology*. Oxford University Press, New York.

[3] Frost, W.H. (1933). Risk of persons in familial contact with pulmonary tuberculosis, *American Journal of Public Health* **23**, 426–432.

[4] McMichael, A.J. (1976). Standardized mortality ratios and the "healthy worker effect": scratching below the surface, *Journal of Occupational Medicine* **18**, 165–168.

[5] Samet, J.M., Pathak, D.R., Morgan, M.V., Coultas, D.B. & Hunt, W.C. (1994). Silicosis and lung cancer risk in underground uranium miners, *Health Physics* **66**, 450–453.

[6] Samet, J.M., Pathak, D.R., Morgan, M.V., Key, C.R. & Valdivia, A.A. (1991). Lung cancer mortality and exposure to radon decay products in a cohort of New Mexico underground uranium miners, *Health Physics* **61**, 745–752.

JONATHAN M. SAMET

Cohort Study

Cohort studies constitute a central epidemiologic approach to the study of relationships between personal characteristics or exposures and the occurrence of health-related events, and hence to the identification of disease prevention hypotheses and strategies.

Consider a conceptually infinite population of individuals moving forward in time. A cohort study involves sampling a subset of such individuals, and observing the occurrence of events of interest, generically referred to as disease events, over some follow-up period. Such a study may be conducted to estimate the rates of occurrence of the diseases to be ascertained, but most often estimation of relationships between such rates and individual characteristics or exposures is the more fundamental study goal. If cohort study identification precedes the follow-up period, then the study is termed prospective, while a retrospective or historical cohort study involves cohort identification after a conceptual follow-up period (*see* **Cohort Study, Historical**). The subsequent presentation assumes a prospective design.

Other research strategies for studying exposure–disease associations, and for identifying disease prevention strategies, include **case–control studies** and randomized controlled disease **prevention trials**. Compared with case–control studies, cohort studies have the advantages that a wide range of health events can be studied in relation to exposures or characteristics of interest, and that prospectively ascertained exposure data are often of better quality than the retrospectively obtained data that characterize case–control studies. However, a cohort study of a particular association would typically require much greater cost and longer duration than would a corresponding case–control study, particularly if the study disease is rare. Compared with randomized controlled trials, cohort studies have the advantage of allowing the study of a broad range of exposures or characteristics in relation to health outcomes of interest, and typically of much simplified study logistics and reduced cost. Randomized intervention trials can also examine a broad range of exposures and disease associations in an observational manner, but the randomized assessments are necessarily restricted to examining the health consequences of a small number of treatments or interventions. However, disease prevention trials have the major advantage that these comparisons are not confounded (*see* **Confounding**) by pre-randomization disease risk factors, whether or not these are even recognized. The choice among these and other research strategies may depend on the distribution of the exposures in the study population and especially on the ability to measure such exposures reliably, on the knowledge and measurement of confounding factors, on the reliability of outcome ascertainment, and on study costs in relation to the public health potential of study results. These issues will be returned to in the final section of this article.

There are many examples of associations that have been identified or confirmed using cohort study techniques, including that between cigarette smoking and lung cancer (*see* **Smoking and Health**); between blood pressure, blood cholesterol, cigarette smoking, and coronary heart disease; between current use of the original combined oral contraceptives and the risk of various vascular diseases; and between atomic bomb radiation exposure and the risk of leukemia or of various solid tumors, to name a few. In recent years there have also been many examples of the use of cohort study designs to examine the association between exposures that are difficult to measure, or that may have limited within-cohort exposure variability, and the occurrence of disease. Such examples may involve, for example, physical activity, dietary, environmental, or occupational exposures. In these settings cohort studies seem often to yield weak or equivocal results, and multiple cohort studies of the same general association may yield contradictory results. It is important to be able to anticipate the reliability and power of cohort studies, to be aware of strategies for enhancing study power and reliability, and to consider carefully optimal research strategies for assessing specific exposure–disease hypotheses.

This article relies substantially on a recent review of cohort study design issues by the author [66]. The reader is also referred to a number of books and review articles focusing on cohort study methodology, including Kleinbaum et al. [41], Miettinen [52], Kelsey et al. [40], Rothman [80], Breslow & Day [7], Kahn & Sempos [38], Checkoway et al. [16], Willett [101], and Morganstern & Thomas [57].

Basic Cohort Study Elements

Exposure Histories and Disease Rates

A general regression notation can be used to represent the exposures (and characteristics) to be ascertained in a cohort study. Let $z_1(u)^{\mathrm{T}} = [z_{11}(u), z_{12}(u), \ldots]$ denote a set of numerically coded variables that describe an individual's characteristics at "time" u, where, to be specific, u can be defined as time from selection into the cohort, and "T" denotes vector transpose. Let $Z_1(t) = [z_1(u), u < t]$ denote the history of such characteristics at times less than t. Note that the "baseline" exposure data, $Z_1(0)$, may include information that pertains to time periods prior to selection into the cohort. Denote by $\lambda[t; Z_1(t)]$ the population **incidence rate** at time t for a disease of interest, as a function of an individual's preceding **"covariate"** history. A typical cohort study goal is the elucidation of the relationship between aspects of $Z_1(t)$ and the corresponding disease rate $\lambda[t; Z_1(t)]$. As mentioned above, a single cohort study may be used to examine many such covariate–disease associations.

The interpretation of the relationship between $\lambda[t; Z_1(t)]$ and $Z_1(t)$ may well depend on other factors. Let $Z_2(t)$ denote the history up to time t of a set of additional characteristics. If the variates $Z_1(t)$ and $Z_2(t)$ are related among population members at risk for disease at time t, and if the disease rate $\lambda[t; Z_1(t), Z_2(t)]$ depends on $Z_2(t)$, then an observed relationship between $\lambda[t; Z_1(t)]$ and $Z_1(t)$ may be attributable, in whole or in part, to $Z_2(t)$. Hence, toward an interpretation of causality (*see* **Causation**) one can focus instead on the relationship between $Z_1(t)$ and the disease rate function $\lambda[t; Z_1(t), Z_2(t)]$, thereby controlling for the "confounding" influences of Z_2. In principle, a cohort study needs to control for all pertinent confounding factors in order to interpret a relationship between Z_1 and disease risk as causal. It follows that a good deal must be known about the disease process and disease risk factors before an argument of causality can be made reliably. This feature places a special emphasis on the replication of results in various populations, with the idea that unrecognized or unmeasured confounding factors may differ among populations. As noted above, the principal advantage of a randomized disease prevention trial, as compared with a purely observational study, is that the randomization indicator variable $Z_1 = Z_1(0)$, where here $t = 0$ denotes the time of randomization, is unrelated to the histories $Z_2(0)$ of all confounding factors, whether or not such are recognized or measured. See, for example, Rubin [82], Robins [74, 76], and Greenland [27], for a fuller discussion of causal inference criteria and strategies.

The choice as to which factors to include in $Z_2(t)$, for values of t in the cohort follow-up period, can be far from straightforward. For example, factors on a causal pathway between $Z_1(t)$ and disease risk may give rise to "overadjustment" if included in $Z_2(t)$, since one of the mechanisms whereby the history $Z_1(t)$ alters disease risk has been conditioned upon. However, omission of such factors may leave a confounded association, since the relationship between Z_2 and disease risk may not be wholly attributable to the effects of Z_1 on Z_2. See Robins [76] for a detailed discussion of the assumptions and procedures needed to argue causality in such circumstances.

Cohort Selection and Follow-Up

Upon identifying the study diseases of interest and the "covariate" histories $Z(t) = [Z_1(t), Z_2(t)]$ to be ascertained and studied in relation to disease risk, one can turn to the estimation of $\lambda[t; Z(t)]$ based on a cohort of individuals selected from the study population. The basic cohort selection and follow-up requirement for valid estimation of $\lambda[t; Z(t)]$ is that at any $[t, Z(t)]$ a sample that is representative of the population in terms of disease rate be available and under active follow-up for disease occurrence. Hence, conceptually, cohort selection and censoring rates (*see* **Censored Data**) (e.g. loss to follow-up rates) could depend arbitrarily on $[t, Z(t)]$, but selection and follow-up procedures cannot be affected in any manner by knowledge about, or perception of, disease risk at specified $[t, Z(t)]$.

Cohort selection rates typically depend on a variety of pre-enrollment characteristics. Potential study subjects may be excluded if they fail to meet certain conditions, perhaps related to prior health events or to their likelihood of completing all study requirements. Similarly, potential study subjects may choose not to participate in a cohort study for a myriad reasons that may be impossible to quantify. How do such selection factors affect the validity or interpretation of estimates of $\lambda[t; Z(t)]$?

In the presence of selection factors the estimation of $\lambda[t; Z(t)]$ applies not to the original conceptual

population, but to a reduced population satisfying cohort exclusionary and "willingness" criteria. The magnitude of disease rates may well differ between these two populations, particularly if certain health criteria must be met for inclusion, or if study subjects tend to be more or less healthy than the broader population from which they arise. The magnitude of associations between $\lambda[t; Z(t)]$ and elements of $Z(t)$ may be affected by cohort selection, thereby limiting the ability to "generalize" the estimated association to the larger population. In general, issues of bias and **effect modification** can be addressed satisfactorily only if the selection factors are accurately measured and properly incorporated into the disease rate model. The ability to generalize to the larger population additionally requires knowledge about, or estimates of, cohort selection rates as a function of selection factor values (*see* **Validity and Generalizability in Epidemiologic Studies**).

As noted previously, censoring rates at a typical follow-up time t may depend on aspects of $Z(t)$ without biasing the estimation of $\lambda[t; Z(t)]$. Note, however, that elements of $Z(t)$ that relate to censoring then typically need to be included in the disease rate model and analysis in order to avoid bias. For this reason, as well as reasons of overall study **power**, it is important to strive to minimize losses to follow-up in cohort study conduct. Certainly, a dependence of selection or censoring rates on characteristics that may be affected by the exposures of interest can much complicate the interpretation of corresponding estimated relationships with disease risk.

The reader is referred to Miettinen [52] and the texts previously listed, as well as to Greenland [24], Miettinen [53, 54], and Poole [63], for discussion of the definition of the study population, and of the "study base" subset thereof from which a cohort is selected, and to these same sources and Greenland [25], Kalbfleisch & Prentice [39, Chapter 5], and Robins [75] for further discussion of cohort study follow-up bias (*see* **Bias in Cohort Studies**).

Covariate History Ascertainment

In general, valid estimation of $\lambda[t; Z(t)]$ within the subpopulation defined by the selection and follow-up procedures requires the accurate and timely ascertainment of the histories $Z(t)$ during the cohort study follow-up period. As before, let $t = 0$ denote the time of enrollment into the cohort. Then one seeks accurate ascertainment of $Z(t)$ for values of $t \geq 0$ in the follow-up period for each cohort member. Characteristics or exposures prior to cohort enrollment ($t < 0$) may be of considerable interest, but there may be a limited ability to obtain such information retrospectively. Hence, it may sometimes be necessary to restrict the covariate history $Z(0)$ to time-independent factors, or to the current or recent values of time-varying factors (*see* **Time-dependent Covariate**). Reliable measurement tools may not be available, even for current values of exposures of interest, or for corresponding confounding factor histories. Similarly, during cohort follow-up ($t > 0$) reliable means of updating covariate histories may or may not be available, and such updates would typically be practicable only at a few selected time points. Also, some of the measurements of interest to be included in $Z(t)$ may be too expensive to obtain on all cohort members. For example, such measurements may involve biochemical or molecular analysis of blood components, or hand extraction of occupational exposure histories from employer records. Hence, a covariate subsampling plan may be an important element of a cohort study concept and design.

Disease Event Ascertainment

A cohort study will often involve a system for regularly updating disease event information. This may involve asking study subjects to self-report a given set of diagnoses, or to self-report all hospitalizations. Hospital discharge summaries may then be examined for diagnoses of interest with confirmation by other medical and laboratory records. Sometimes disease events of interest will be actively ascertained by taking periodic measurements on all cohort members. For example, electrocardiographic tracings toward coronary heart disease diagnosis or screening breast mammograms toward breast cancer diagnosis may be a part of a basic study protocol. Diagnoses that require considerable judgment may be examined by a committee of experts toward enhancing the standardization and accuracy of disease event diagnoses. In spite of the application of the best practical outcome ascertainment procedures, there will usually be some misclassification (*see* **Misclassification Error**) of whether or not certain disease events have occurred with resulting bias in the estimation of $\lambda[t; Z(t)]$. A dependence of ascertainment rates on factors other

than those included in $Z(t)$ may be able to be accommodated by including such factors as control variables in $Z_2(t)$.

Unbiased ascertainment of the timing of disease events relative to $Z(t)$ is also important for valid inference. For example, if disease screening activities vary with aspects of $Z(t)$, leading to earlier reporting at some covariate values than at others, then biased associations will typically arise. Similarly, differential lags in the reporting of disease events may cause bias unless specifically accommodated in data analysis.

Data Analysis

Suppose now that covariate disease associations of interest have been identified, and that procedures for selecting a cohort and for accurately ascertaining pertinent covariate histories and disease event times have been established. What then can be said about the ability to detect an association between a particular characteristic or exposure and a corresponding disease risk? Typically, a test of association would be formulated in the context of a descriptive statistical model, though in some settings a mechanistic or biologically based model may be available (e.g. Armitage & Doll [2], Whittemore & Keller [99], and Moolgavkar & Knudson [56] (*see* **Multistage Carcinogenesis Models**).

A very useful and flexible descriptive modeling approach formulates the association in terms of relative risk. Specifically, one supposes [18] that

$$\lambda\{t; Z(t)\} = \lambda_0(t) \exp\{z(t)^{\mathrm{T}}\beta\}, \qquad (1)$$

where $z(t)^{\mathrm{T}} = \{z_1(t), \ldots, z_p(t)\}$ is a modeled regression vector formed from $Z(t)$, $\beta^{\mathrm{T}} = (\beta_1, \ldots, \beta_p)$ is a corresponding relative risk parameter to be estimated, and $\lambda_0(\cdot)$ is an unrestricted baseline disease rate (**hazard**) function corresponding to a modeled regression vector $z(t) \equiv 0$. A test of the null hypothesis of no association between, say, $z_1(t)$ and disease risk then corresponds to $\beta_1 = 0$. Estimation and testing can be conducted by applying standard likelihood procedures to the **partial likelihood** function

$$L(\beta) = \prod_{i=1}^{k} \left\{ \exp[z_i(t_i)^{\mathrm{T}}\beta] \Big/ \sum_{l \in R(t_i)} \exp[z_l(t_i)^{\mathrm{T}}\beta] \right\}, \qquad (2)$$

where $t_1, \ldots, t_k$ denotes the disease incidence times in the cohort, $z_i(t_i)$ is the modeled covariate at time t_i for the cohort member diagnosed at t_i, and $R(t_i)$ denotes the set of cohort members being followed for disease occurrence at time t_i. Consideration of the distribution of the score statistics $U_1(0) = \partial \log L(\hat{\beta}_0)/\partial \hat{\beta}_0^{\mathrm{T}}$, where $\hat{\beta}_0$ maximizes $L(\beta)$ subject to $\beta_1 = 0$, makes it clear that the power of the test for $\beta_1 = 0$ depends primarily on the magnitude, β_1, of the regression coefficient, the expected number of disease events, k, during cohort follow-up, and the "spread" of the primary regression variable distribution $[z_{1l}(t); l \in R(t)]$ across the cohort follow-up times. The power will also depend somewhat on the distributions of the other (control) variables included in $z(t)$ and on the sampling variation in $\hat{\beta}_0$. A useful generalization of (1) allows the baseline disease rate function $\lambda_0(\cdot)$ to differ arbitrarily among strata that may be time dependent, typically defined by categorizing the histories $Z_2(t)$. Estimation and testing can then be based on a likelihood function that is simply the product of terms (2) over strata.

Conditions for the avoidance of confounding and other biases in the estimation, of a **relative risk** parameter β_1 are naturally less restrictive than are those for accurate estimation of the entire disease rate process $\lambda[t; Z(t)]$. For example, selection, follow-up, and disease ascertainment rates can depend on factors not included in $Z(t)$ provided such factors are unrelated to $Z_1(t)$, conditional on $[t, Z_2(t)]$, without implying bias in the estimation of β_1, assuming that a relative risk model of the form (1) holds conditional on such factors. There is a considerable epidemiologic literature exploring such issues. In addition to the texts previously cited see, for example, Miettinen & Cook [55], Boivin & Wacholder [5], and Greenland & Robins [29]. Though the use of relative risk, and of closely associated **odds ratios** is ubiquitous in epidemiology, other measures of association, including disease rate difference measures, also have utility and their own criteria for valid estimation. See, for example, the texts previously cited and Greenland [25]. The evolution of the covariate histories $Z(t)$, over time, may also be of substantive interest. For example, the extent to which disease risk factors track over time may have clinical implications, or the extent to which an intermediate outcome in $Z_2(t)$ can explain an exposure disease association may provide insights into disease mechanisms. Joint

analyses of an exposure in relation to two or more disease processes may also be of considerable practical interest.

Subsequent sections expand upon these basic cohort study features.

Study Design

Cohort Study Power

A number of authors have provided a methodology for cohort study sample size and power determination (e.g. Gail [22], Casagrande et al. [15], Fleiss et al. [20], Whittemore [97], Brown & Green [11], Greenland [23], and Self et al. [85]. Breslow & Day [7, Chapter 7] provide a detailed account of this topic, including consideration of the impact of varying the exposure distribution, of confounding factor control, of **matching**, and of nested case–control sampling (*see* **Case–Control Study, Nested**), on study power. See also Whittemore & McMillan [100].

As suggested above, a comprehensive approach to the issue of study power for a particular association would require a range of design assumptions, including assumptions about the exposure distribution and its variation across time, about the magnitude of the regression parameter β_1, and concerning the baseline disease incidence rates. Though such a comprehensive approach may be useful, and flexible power calculation procedures permitting the use of complex assumptions are available (e.g. Self et al. [85]), power calculations for the simple odds ratio special case can provide valuable guidance concerning cohort study power and related design choices. Suppose that a baseline characteristic or exposure of interest is dichotomized into an "exposed" group ($Z_1 = 1$) comprised of the fraction, γ, of the population having a high value of an exposure, and an "unexposed" group ($Z_1 = 0$) comprised of the fraction, $1 - \gamma$, of the population having a comparatively low value. Let p_1 denote the probability that an exposed subject experiences a study disease of interest during a prescribed cohort follow-up period and let p_2 be the corresponding probability for an unexposed subject. Note that p_1 and p_2 can be thought of as average probabilities over the respective distributions of cohort follow-up times and over the exposure distributions within the exposed and unexposed categories. A simple sample size formula, based on the well-known approximate normality of logarithm of the simple odds ratio estimator, indicates that the cohort sample size must be at least

$$n = [p_2(1 - p_2)]^{-1}(\log \lambda)^{-2} Q, \qquad (3)$$

where $\lambda = p_1(1 - p_2)[p_2(1 - p_1)]^{-1}$ is the exposed vs. unexposed odds ratio, and $Q = [\gamma(1 - \gamma)]^{-1}\{W_{\alpha/2} - W_{1-\eta}[\gamma + \lambda^{-1}(1 - p_2 + \lambda p_2)^2(1 - \gamma)]^{1/2}\}^2$, to ensure that a two-sided α-level test (e.g. $\alpha = 0.05$) of the null hypothesis of no exposure effect ($\lambda = 1$) will be rejected with probability (power) η, where $W_{\alpha/2}$ and $W_{1-\eta}$ denote the upper $\alpha/2$ and $1 - \eta$ percentiles of the standard normal distribution, respectively.

Note that Q in Eq. (3) is a rather slowly varying function of λ and p_2 at specified α and η, so that the sample size necessary to achieve a specified power is approximately inversely proportional to $p_2(1 - p_2)$, where p_2 is again the unexposed disease probability, and inversely proportional to $(\log \lambda)^2$, the square of the exposed vs. unexposed log-odds ratio. Hence there is considerable sample size sensitivity to the magnitude of the odds ratio, with an odds ratio of 1.5 requiring about three times the cohort size of an odds ratio of 2.0, and an odds ratio of 1.25 requiring about ten times the sample size of that for an odds ratio of 2.0. The magnitude of the basic disease incidence rates (i.e. p_2) are also of considerable importance in the choice of a cohort size and average follow-up duration. Prentice [66] displayed selected power, η, calculations, developed in planning the cohort study component of the Women's Health Initiative [79, 103] which is currently enrolling 100 000 post-menopausal American women in the age range 50–79, based on (3) for selected cohort sizes, n, odds ratios, λ, and exposure fractions, γ. The power calculations are shown as a function of unexposed average incidence rates and average cohort follow-up duration, the product of which is the unexposed incidence rate p_2.

Measurement error in the modeled regression variable in (1) can involve a substantial loss of power and, except in idealized situations, can invalidate the null hypothesis test. The impact of covariate measurement error on the study design, conduct, and interpretation is one of the least developed, and potentially most important, aspects of cohort study methodology. For example, consider a binary exposure variable subject to misclassification. Suppose also that the misclassification rates do not vary within the exposed and unexposed

groups, according to quantitative exposure levels or other study subject characteristics, and that all necessary confounding variables are included in the analysis and are measured without error. Under this circumstance, misclassification in the binary exposure variable effectively reduces the odds ratio λ in the sample size–power relationship (3). To cite a specific example, suppose that $Z_1 = 1$ denotes values above the median for a specific exposure while $Z_1 = 0$ denotes values below the median, so that $\gamma = 0.5, \lambda = 2.0$, and $p_2 = 0.02$. Suppose that rather than Z_1 one can only measure a variable X_1 which, when dichotomized at its median, gives $p(X_1 = 1|Z_1 = 1) = p(X_1 = 0|Z_1 = 0) = 1 - \Delta$ and $p(X_1 = 1|Z_1 = 0) = p(X_1 = 0|Z_1 = 0) = \Delta$. Suppose that this common misclassification probability takes the value $\Delta = 0.2$. The odds ratio based on the measured dichotomous variate X can then be calculated to be 1.50, so that a 20% exposure misclassification leads in these circumstances to a substantially attenuated odds ratio and requires an increase in cohort sample size by a factor of about 3 to preserve power for the null hypothesis test. See, for example, Walter & Irwig [93] and Holford & Stack [33] for additional discussion of exposure measurement error effects on study design and power. In general, the effects of covariate measurement error may be much more profound than simply relative risk attenuation and loss of power, as will be discussed further below (*see* **Misclassification Error**; **Measurement Error in Survival Analysis**; **Measurement Error in Epidemiologic Studies**).

Study Population

Typically a cohort study will be conceived with a set of motivating hypotheses in mind. The study population may then be selected as one in which such hypotheses and related associations may be able to be efficiently and reliably tested. For example, a range of studies of the health risks following from human exposure to ionizing radiation have been carried out in cohorts of atomic bomb exposed populations in Hiroshima and Nagasaki. A principal "Life Span Study" cohort consists of over 100 000 persons with residence in either city as of 1 October 1950 [4]. Decisions needed to be made concerning the inclusion of such residents who were some distance from the epicenter at the time of bombing or who were "not-in-city" at the time of bombing. The latter group has been variably included in reports from this study. The generalizability of the Life Span Study results is somewhat impacted by selection factors related to survival of the acute exposure and factors related to continuing residence in either of the two cities during the time period 1945–1950, but otherwise appears to be representative of a hypothetical population "like that of Hiroshima and Nagasaki residents" at the time of radiation exposure.

There are several ongoing cohort studies that are motivated in part by hypotheses related to diet and cancer. Recent examples include studies of US nurses (e.g. Willett et al. [102]), studies of women participating in a randomized trial of breast screening to prevent breast cancer mortality (e.g. Howe et al. [34]), studies of Iowa women (e.g. Kushi et al. [43]), and studies of men of Japanese heritage living in Hawaii (e.g. Stemmerman et al. [89]). Such studies appear to be limited (e.g. Prentice et al. [69]) by modest within-population variability in nutrient exposures of interest, substantial exposure measurement error in dietary assessment, and many highly correlated dietary exposure variables. These factors may combine to cast doubt on study reliability. Recent diet and cancer cohort study developments attempt to address such concerns by studying populations having a broader than usual range in dietary habits, by enhancing dietary assessment methodology, and by employing multiple dietary assessment tools in subsets of the cohort. Specifically, recently initiated diet and cancer cohort studies include a study among members of the American Association of Retired Persons with an overrepresentation of persons having estimated fat intake within certain extreme percentiles of the overall fat intake distribution; a multiethnic cohort study taking place in Los Angeles and Hawaii; and a multipopulation cohort study in Europe entitled the European Prospective Investigation into Cancer and Nutrition (e.g. Riboli [73]).

Much valuable information on cardiovascular disease risk factors has arisen in the context of randomized prevention trials, including, for example, cohort studies of persons enrolled in the Multiple Risk Factor Intervention Trial (MRFIT) [58], the Lipid Research Clinic Primary Prevention Trial [49], and the Hypertension Detection and Follow-up Program Trial [35]. Persons screened for possible enrollment in such trials include another potential source

of cohort study enrollees. For example, the approximately 300 000 men screened for possible enrollment in MRFIT, a trial involving about 12 000 randomized men, yielded precise information on the relationship between blood cholesterol and mortality from various diseases (e.g. Jacobs et al. [37]).

In each of the cohort studies mentioned above consideration of inclusion and exclusion criteria is required in interpreting study results, particularly concerning the degree of generalizability of results to a broader source population.

Sample Size and Study Duration

One approach to establishing the size of the cohort is to list motivating hypotheses along with corresponding design assumptions, and to select a cohort size that will yield acceptable power (e.g. >80%) for all, or most, key hypotheses within a practical follow-up period. In fact, cohort studies are often initiated with the hope that active follow-up will continue for some decades, but for scientific and logistic reasons study planning exercises may need to be based on an average follow-up period of, say, 5–10 years. These reasons include funding cycle logistics, the desire of investigators to produce new information in a practicable time period, and a possible reduced relevance of baseline covariate data to disease risk determination beyond a few years of follow-up. Exercises to determine a cohort size should make provisions for the power-influencing factors mentioned previously, particularly exposure measurement error influences.

A more empirical approach to cohort size determination can be based on the consideration of previous cohort study sizes and of the corresponding range of associations tested. Specifically, cohort studies that have yielded much useful information on cardiovascular disease risk factors have often been in the range of 5000 to 20 000 persons, including, for example, the **Framingham Study**, observational studies within the MRFIT study and other coronary heart disease prevention trials, the Adult Health Study of atomic bomb survivors, and the Cardiovascular Health Study. Cohort sizes in the vicinity of 5000 may be adequate for cardiovascular disease studies among older persons, as in the Cardiovascular Health Study [21] that is restricted to persons of age 65 or older, whereas considerably larger cohort sizes may be indicated for studies among younger persons, as in the Royal College of General Practitioners' study of the health effects of oral contraceptive use [81] that enrolled 46 000 younger women.

Most cohort studies of diet and cancer to date have involved sample sizes in the vicinity of 50 000–100 000, including, for example, the Nurses Health Study, the Canadian National Breast Screening study, and the Iowa Women's study. Consideration of range of nutrient intake and likely magnitude of random measurement error in dietary assessment, however, suggests that some pertinent odds ratios following measurement error influences can be hypothesized to be in the range 1.1–1.2 (e.g. Prentice et al. [69], Prentice & Sheppard [70], Prentice [67]). These types of considerations support the recently initiated cohort studies involving larger sample sizes (e.g. 100 000–400 000) in populations having an unusual degree of diversity of dietary habits.

Study Conduct and Analysis

Protocol and Procedures

A cohort study requires a clear, concise protocol that describes study objectives, design choices, performance goals and monitoring and analysis procedures. A detailed manual of procedures, which describes how the goals will be achieved, is necessary to ensure that the protocol is applied in a standardized fashion. Carefully developed data collection and management tools and procedures, with as much automation as practicable, can also enhance study quality. Centralized training of key personnel may be required to ensure that the protocol is understood, and to enhance study subject recruitment and comparability of outcome ascertainment as a function of exposures and confounding factors.

Analysis and reporting procedures should acknowledge the large number of exposure–disease associations that may be examined in a given cohort study, as well as the multiple time points at which hypotheses concerning each such exposure may be tested. In fact, even though such multiple testing considerations are routinely acknowledged in randomized clinical trials, their inclusion in cohort study reporting seems to be uncommon.

Covariate Data Ascertainment and Reliability

The above framework assumes that pertinent covariate histories $Z(t)$ are available for all cohort members

at all times t during the cohort follow-up period. Conceptually, the availability of such data would require baseline covariate data collection to ascertain all pertinent exposures and characteristics prior to enrollment in the cohort study followed by the continuous updating of evolving covariates of interest during cohort follow-up. In practice, however, there may be a limited ability to ascertain retrospectively such covariate information at baseline, unless relevant specimens and materials had fortuitously been collected and stored for other reasons. Furthermore, there will be a limit to the frequency and extent of covariate data updating that can be carried out during covariate follow-up. Such evolving covariate data may be a key to adequate confounding control [76], and may be fundamental to such issues as the estimation of time lags between exposure and disease risk, and estimation of the relationship between covariate change and disease risk more generally.

The fact that the desired covariates may be poorly measured, or completely missing, can be a substantial impediment to the analysis and interpretation of study results. The effects of mismeasured or missing exposure or confounding factor data on relative risk parameter estimation may be much more profound than simple attenuation. In fact, if the measurement error variances are at all large (e.g. more than 10%–20%) relative to the variance of the true regression variables, then it will often be important to undertake additional data collection in the form of validation or calibration substudies, toward accommodating such measurement errors in data analysis. These substudies can also aid in the accommodation of missing covariate data, as it is otherwise necessary to make a missing at random assumption; that is, to assume that missingness rates are independent of the true covariate value, given the accumulated data at earlier time points (*see* **Missing Data in Epidemiologic Studies**).

Validation and Calibration Substudies

In some circumstances it may be possible to design a cohort study to include a validation subsample in which covariate measurements that are essentially without measurement error are taken in a random subset of the cohort. Such measurements may be too expensive or too demanding on study subjects to be practicable for more than a small subset of the cohort. See Greenland [26], Marshall [50], and Spiegelman & Gray [88] for discussion of the role and design of validation substudies (*see* **Validation Study**).

Consistent estimation of the relative risk parameter β is possible by making use of a validation substudy (e.g. Pepe & Fleming [59], Carroll & Wand [14], Lee & Sepanski [46]), though the loss of efficiency arising from substantial measurement errors presumably may be large. The validation sample permits nonparametric estimation of the expectation of $\exp[z(t)^{\mathrm{T}}\beta]$ given the measured covariate and study subject "at risk" status at t, obviating the need for specific measurement error assumptions, and giving rise to estimated likelihood or estimated score procedures for the estimation of β. An alternate data analysis strategy, in the presence of a validation subsample, simply replaces the mismeasured covariate by its estimated conditional expectation, given the accumulated data on the study subject at preceding times. This so-called regression calibration approach is quite convenient, and it performs well in a variety of circumstances even though typically technically inconsistent under (1) and other nonlinear models (e.g. Prentice [64], Rosner et al. [77, 78], Carroll et al. [12], and Wang et al. [94]). If a validation study is possible for some or all important covariates, then the inclusion of a validation sample of appropriate size may be a critically important aspect of cohort study design and conduct.

The regression calibration procedure just mentioned extends fairly readily even if the subsample measure is not the true covariate value but is contaminated by measurement error that is independent of both the routinely available measurement and the true covariate values. See, for example, Greenland & Kleinbaum [28], Kupper [42], Whittemore [98], Rosner et al. [78, 79], Pierce et al. [61], Carroll et al. [13], Armstrong [3], Clayton [17], Thomas et al. [90], and Sepanski et al. [86] for various approaches to cohort data analyses in the presence of calibration subsamples.

Unfortunately, the situation changes dramatically if the subsample measurement error does not satisfy such independence properties. For example, Prentice [67] considers a measurement model for two self-report measures of dietary fat in an attempt to interpret a combined cohort study analysis of dietary fat in relation to breast cancer. One self-report measure, based on food frequency assessment, was available in all cohort study members, while a more detailed measure, based on multiple days of food recording, was available on a small subset of the

cohorts in question. By using food frequency and food record data from the Women's Health Trial feasibility study [36] and allowing the dietary fat measurement errors for the two instruments to be correlated and to depend on an individual's body mass index, Prentice [67] showed that even the strong relative risk relationship between fat and postmenopausal breast cancer suggested by international correlation studies (e.g. Prentice & Sheppard [70] could be projected to be essentially undetectable in a cohort study using a food frequency instrument, regardless of its size. Specifically, relative risks of 1, 1.5, 2.0, 2.7, and 4.0 across fat intake quintiles if measurement errors were absent are reduced to projected values of 1, 1.0, 1.0, 1.1, and 1.1 upon allowing for both random and systematic aspects of dietary fat measurement error. This illustration suggests that covariate measurement errors may be at the root of many controversial associations in epidemiology, and motivates the importance of objective measures of exposure. Biomarker exposure measures may be quite valuable in such contexts even if such measures include considerable noise. For example, in the dietary area, total energy expenditure can be objectively measured over short periods of time using doubly labelled water techniques while protein expenditure can be measured by urinary nitrogen (e.g. Lichtman et al. [47], Heitmann & Lessner [32], Martin et al. [51], Sawaya et al. [84]). Plummer & Clayton [62] use urinary nitrogen data to demonstrate correlated measurement errors among dietary protein self-report measures.

Additional Sampling Strategies

It will often be efficient when conducting a cohort study to assemble the raw materials for covariate history assembly on the entire cohort, but to restrict the processing and analysis of such materials to appropriate subsamples. For example, a random sample, or stratified random sample, of the cohort may be selected, along with all persons experiencing disease events of interest, for covariate data processing. Such a case–cohort approach allows relative risk estimation (e.g. Prentice [65]) based on (2) with $R(t_i)$ consisting of the person developing a disease at t_i (the case) along with all "at risk" members of the selected sample (the subcohort), though (2) no longer has a likelihood function interpretation and specialized variance estimators are required (*see* **Case–Cohort Study**). Alternatively, a case of a specific disease at time t_i may be matched to one or more controls randomly selected from the risk set at t_i, with ordinary likelihood methods applied to (2) except that $R(t_i)$ is replaced by the case and time-matched controls (e.g. Liddell et al. [48], Prentice & Breslow [68]) (*see* **Case–Control Study, Nested**). Recently Samuelson [83] has proposed an alternate analysis of such nested case–control samples that appears to yield meaningful efficiency improvements relative to the standard procedure just described.

The nested case–control sampling approach allows cases and controls to be matched on various study subject characteristics, including time from enrollment into the cohort, as may be important if some covariate measurements (e.g. blood concentrations of selected nutrients) degrade with storage time. Such issues may be accommodated under case–cohort sampling by stratifying the subcohort selection on cohort enrollment data and by analyzing all case and subcohort specimens and materials at a common point in time. In fact, it may be useful to delay subcohort selection to the time of data analysis in order to match subcohort sizes in each stratum to the corresponding numbers of disease events. See Langholz & Thomas [44, 45] and Wacholder [91] for further discussion and comparison of nested case–-control and case–cohort sampling.

A related topic includes a two-stage process in establishing a cohort. A first stage would involve collecting information on the exposures of primary interest, or perhaps collecting fairly crude estimates of such exposures. The second stage would then involve selecting a subset of the stage 1 study subjects that give a desirable exposure distribution for more detailed data collection. The diet and cancer cohort study among members of the American Association of Retired Persons provides an example in which subjects are oversampled for the second stage if their food frequency estimated fat intakes are in certain extreme percentiles. A two-stage sampling approach is also natural for the study of rare exposures. A considerable literature exists on the design and analysis of two-stage sampling schemes, mostly in the context of case–control studies (e.g. White [96], Walker [92], Breslow & Cain [8], Breslow & Zhao [9], Flanders & Greenland [19] (*see* **Case–Control Study, Two-phase**).

Additional Data Analysis Topics

The above discussion assumes that the basic time variable is time from selection into the cohort. This choice is attractive in that relative risk estimation [i.e. each factor in (2)] is then based on comparisons among individuals at the same length of time since cohort entry. Other important time variables, such as study subject age and chronologic time, can be accommodated by regression modeling, or stratification with (2) replaced by a product of like terms over strata. Alternatively, if cohort eligibility criteria or recruitment strategies changed markedly over time, or if covariate data or outcome data ascertainment procedures changed markedly over time, one may prefer to define t to be chronologic time so that relative risk estimates are based on comparisons of measurement taken under common procedures, with age and time since study enrollment controlled by stratification or modeling. In this case a study subject begins contributing to (2) at the time (date) of study enrollment.

If the study is conducted to estimate disease rates, or cumulative disease rates and **absolute risks**, then it may be natural for interpretation to define t to be age, or time from some significant event (e.g. infection with human immunodeficiency virus). Depending on the means of study subject identification such significant event data may only be known to be earlier than a specified time, giving rise to interval censored event time data and a range of interesting statistical estimation issues (e.g. Brookmeyer & Gail [10, Chapter 5]) (*see* **Biased Sampling of Cohorts**).

The fact that multiple outcomes are typically ascertained in a cohort study allows for the possibility of relating an exposure jointly to two or more event rates and, of course, there may be multiple events of a given type during the cohort study follow-up of an individual. The literature includes generalization of (1) to repeat failure times on an individual study subject (e.g. Prentice et al. [72], Andersen et al. [1]), while for single occurrences of multiple types of events one can consider the use of (2) for each failure type in conjunction with a modified variance estimation procedure for the joint estimation of relative risks for several diseases (e.g. Wei et al. [95]). This latter procedure will have acceptable efficiency under most situations of practical interest.

The above presentation assumes that failure times among distinct cohort study members are independent. This assumption may be violated if study subjects share environmental or genetic factors. In fact, in **genetic epidemiology** one may use the relationships among the failure time data of family members in a pedigree cohort study in an attempt to identify the existence (aggregation analysis), inheritance pattern (segregation analysis) and physical location (linkage analysis) of genes that play a role in determining disease risk. Such studies or, more practically, case–control subsamples of population-based family studies, may also be used to study **gene–environment interaction**.

Cohort Study Role

Cohort studies properly play a central role in epidemiologic research. Disease associations that are relatively strong can often be reliably studied using cohort study techniques, especially if there is sufficient knowledge of disease risk factors and exposure correlates to permit comprehensive efforts to control confounding. Relative risk analyses are unlikely to be misleading under these circumstances since residual **confounding** will tend to be small compared with the relative risk trend under study. If the exposures and other covariates of interest can be reliably ascertained retrospectively, then a case–control design may introduce considerable economies relative to a cohort study; in fact, the case–control design is very commonly employed, particularly if the study diseases are rare and good disease registries are available for case ascertainment.

If the relative risk trends to be studied are more modest, then the reliability of the cohort study may be less clear, as uncontrolled confounding, or other biases, may have a salient impact on the estimated associations. If, in addition, the exposures of interest or strong confounding factors or disease outcomes involve measurement errors that are substantial, but of unknown properties, then the cohort study reliability may be poor.

In these circumstances one can turn to an experimental approach, as has been done to study the role of diet in disease in various studies of micronutrient supplementation, and of a low fat eating pattern. Such clinical trials permit a valid test of the intervention applied in relation to a range of diseases without concern about "baseline" confounding,

and without the need for precise dietary assessment. Dietary assessment, in such contexts, enters in a secondary fashion to document that a sufficiently powerful hypothesis test has been conducted, and in attempts to isolate intervention activities responsible for any observed disease risk difference. However, primary disease prevention clinical trials are logistically difficult and expensive, so that only a few can be conducted at any time point, preferably those that are motivated by hypotheses having great public health potential.

The recent movement mentioned above, toward multipopulation cohort studies (e.g. the EPIC study) to enhance exposure heterogeneity seems attractive in the type of circumstances alluded to above. Such a multipopulation cohort can be expected to involve a broadened exposure range, perhaps in a manner that does not increase covariate measurement errors, thereby enhancing both reliability and power. The relative risk information in such a multipopulation study can be partitioned into between-population and within-population components under standard **random effects** modeling assumptions (e.g. Sheppard & Prentice [87]). In fact, it may often happen that much of the retrievable information arises from between-population sources. Also, between-population, but not within-population, relative risk estimates tend to be highly robust to independent mean zero measurement errors in covariates, essentially because such estimates are based on covariate function averages over large numbers of study subjects that are little affected by such errors. Furthermore, the between-population relative risk information may be able to be extracted efficiently by relating covariate history data on modest numbers of persons in each population (e.g. 100–200) to corresponding population disease rates, as may be available from disease registers in each population [71]. These points suggest that studies of relatively modest relative risk trends associated with exposures that are measured with considerable noise may sometimes be efficiently studied using an aggregate data (ecologic) approach that involves covariate surveys in disease populations covered by good quality disease registers. Note, however, that confounding control across heterogeneous populations may pose particular challenges, and that careful data analysis will be required to avoid aggregation and other biases in such a study. See Piantadosi et al. [60], Greenland & Morganstern [30], Brenner et al. [6], and Greenland & Robins [31] for further discussion of these bias issues (*see* **Ecologic Study**; **Ecologic Fallacy**).

Statistical thinking and methodology have come to play an important role in the design, conduct, and analysis of cohort studies. Cox regression and closely related logistic regression methods play a central role in data analysis and reporting, and in related study planning efforts. Cohort sampling techniques are widely used and have enhanced cohort study efficiency. A topic of continuing importance relates to the methodology for measurement error assessment, and to analytic methods to reduce the sensitivity of results to measurement error influences.

Acknowledgment

This work was supported by grant CA-53996 from the National Institutes of Health.

References

[1] Andersen, P.K., Borgan, D., Gill, R.D. & Keiding, N. (1961). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Armitage, P. & Doll, R. (1961). Stochastic models for carcinogenesis, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. IV, University of California Press, Berkeley, pp. 19–38.

[3] Armstrong, B.G. (1991). The effect of measurement error on relative risk regressions, *American Journal of Epidemiology* **132**, 1176–1184.

[4] Beebe, G.W. & Usagawa, M. (1968). The major ABCC samples, *Atomic Bomb Casualty Commission Technical Report* 12–68.

[5] Boivin, J.F. & Wacholder, S. (1985). Conditions for confounding of the risk ratio and the odds ratio, *American Journal of Epidemiology* **121**, 152–158.

[6] Brenner, H., Savitz, D.A., Jöckel, K.H. & Greenland, S. (1992). The effects of non-differential exposure misclassification in ecological studies, *American Journal of Epidemiology* **135**, 85–95.

[7] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2: *The Design and Analysis of Cohort Studies*. IARC Scientific Publications No. 82, International Agency for Research on Cancer, Lyon, France.

[8] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case-control data, *Biometrika* **74**, 11–20.

[9] Breslow, N.E. & Zhao, L.P. (1988). Logistic regression for stratified case-control studies, *Biometrics* **44**, 891–899.

[10] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, New York.

[11] Brown, C.C. & Green, S.B. (1982). Additional power computations for designing comparative Poisson trials, *American Journal of Epidemiology* **115**, 752–758.

[12] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, New York.

[13] Carroll, R.J. & Stefanski, L.A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors, *Journal of the American Statistical Association* **85**, 652–663.

[14] Carroll, R.J. & Wand, M.P. (1991). Semiparametric estimation in logistic measurement error models, *Journal of the Royal Statistical Society, Series B* **53**, 573–585.

[15] Casagrande, J.T., Pike, M.C. & Smith, P.G. (1978). An improved approximate formula for calculating sample sizes for comparing two binomial distributions, *Biometrics* **34**, 483–486.

[16] Checkoway, H., Pearce, N. & Crawford-Brown, D.J. (1989). *Research Methods in Occupational Epidemiology*. Oxford University Press, New York.

[17] Clayton, D.G. (1992). Models for the analysis of cohort and case-control studies with inaccurately measured exposures, in *Statistical Models for Longitudinal Studies of Health*, J.H. Dwyer, P. Lippert, M. Feinleib & H. Hoffmeister, eds. Oxford University Press, Oxford, pp. 301–331.

[18] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[19] Flanders, W.D. & Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs, *Statistics in Medicine* **10**, 739–747.

[20] Fleiss, J.L., Tytun, A. & Ury, U.K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions, *Biometrics* **36**, 343–346.

[21] Fried, L.P., Borhani, N., Enright, P., Furberg, C., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T., Mittlemark, M.B., Newman, A., O'Leary, D.H., Psaty, B., Rautaharju, P., Tracy, R.P., Weiler, P.G. for the Cardiovascular Health Study Research Group (CHS) (1991). The cardiovascular health study: design and rationale, *Annals of Epidemiology* **1**, 263–276.

[22] Gail, M. (1974). Power computations for designing comparative Poisson trials, *Biometrics* **30**, 231–237.

[23] Greenland, S. (1985). Power, sample size and smallest detectable effect determination for multivariate studies, *Statistics in Medicine* **4**, 117–127.

[24] Greenland, S. (1986). Cohorts versus dynamic populations: A dissenting view, *Journal of Chronic Diseases* **39**, 565–566.

[25] Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses, *American Journal of Epidemiology* **125**, 761–768.

[26] Greenland, S. (1988). Statistical uncertainty due to misclassification: implications for validation substudies, *Journal of Clinical Epidemiology* **41**, 1167–1174.

[27] Greenland, S. (1990). Randomization, statistics and causal inference, *Epidemiology* **1**, 421–429.

[28] Greenland, S. & Kleinbaum, D.G. (1983). Correcting for misclassification in two-way tables and matched-pair studies, *International Journal of Epidemiology* **12**, 93–97.

[29] Greenland, S. & Robins, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding, *International Journal of Epidemiology* **15**, 412–418.

[30] Greenland, S. & Morganstern, H. (1989). Ecological bias, confounding and effect modification, *International Journal of Epidemiology* **18**, 269–274.

[31] Greenland, S. & Robins, J. (1994). Invited commentary: Ecologic studies. Biases, misconceptions and counterexamples, *American Journal of Epidemiology* **139**, 747–760.

[32] Heitman, B.L. & Lessner, L. (1995). Dietary underreporting by obese individuals – is it specific or non-specific? *British Medical Journal* **311**, 986–989.

[33] Holford, T.R. & Stack, C. (1995). Study design for epidemiologic studies with measurement error, *Statistical Methods in Medical Research* **4**, 339–358.

[34] Howe, G.R., Friedenreich, C.M., Jain, M. & Miller, A.B. (1991). A cohort study of fat intake and risk of breast cancer, *Journal of the National Cancer Institute* **83**, 336–340.

[35] Hypertension Detection and Follow-up Program Cooperative Group (1979). Five year findings of the Hypertension Detection and Follow-up Program I. Reductions in mortality of persons with high blood pressure, including mild hypertension, *Journal of the American Medical Association* **242**, 2562–2571.

[36] Insull, W., Henderson, M.M., Prentice, R.L., Thompson, D.J., Clifford, C., Goldman, S., Gorbach, S., Moskowitz, M., Thompson, R. & Woods, M. (1990). Results of a feasibility study of a low fat diet, *Archives of Internal Medicine* **150**, 421–427.

[37] Jacobs, D., Blackburn, H., Higgins, M., Reed, D., Iso, H., McMillan, G., Neaton, J., Nelson, J., Potter, J., Rifkind, B., Rossouw, J., Shekelle, R., Yusuf, S., for Participants in the Conference on Low Cholesterol: Mortality Associations (1992). Report of the conference on low blood cholesterol: Mortality associations, *Circulation* **86**, 1046–1060.

[38] Kahn, H.A. & Sempos C.T. (1989). *Statistical Methods in Epidemiology*. Oxford University Press, New York.

[39] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[40] Kelsey, J.L., Thompson, W.D. & Evans, A.S. (1986). *Methods in Observational Epidemiology*. Oxford University Press, New York.

[41] Kleinbaum, D.G., Kupper, L.L. & Morganstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont.

[42] Kupper, L.L. (1984). Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies, *American Journal of Epidemiology* **120**, 643–648.

[43] Kushi, L.H., Sellers, T.A., Potter, J.D., Nelson, C.L., Munger, R.G., Kaye, S.A. and Folson, A.R. (1992). Dietary fat and postmenopausal breast cancer, *Journal of the National Cancer Institute* **84**, 1092–1099.

[44] Langholz, B. & Thomas, D.C. (1990). Nested case-control and case–cohort methods for sampling from a cohort: a critical comparison, *American Journal of Epidemiology* **31**, 169–176.

[45] Langholz, B. & Thomas, D.C. (1991). Efficiency of cohort sampling designs: some surprising results, *Biometrics* **47**, 1553–1571.

[46] Lee, L.F. & Sepanski, J.H. (1995). Estimation in linear and nonlinear errors in variables models using validation data, *Journal of American Statistical Association* **90**, 130–140.

[47] Lichtman, S.W., Pisarka, K., Berman, E.R., Pestone, M., Dowling, H., Offenbacker, E., Weisel, H., Heshka, S., Matthews, D.E. & Heymsfield, S.B. (1992). Discrepancy between self-reported and actual calorie intake and exercise in obese subjects, *New England Journal of Medicine* **327**, 1893–1898.

[48] Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977). Methods for cohort analysis: appraisal by application to asbestos mining, *Journal of the Royal Statistical Society, Series A* **140**, 469–490.

[49] Lipid Research Clinic Program: The Lipid Research Clinic Coronary Primary Prevention Trial Results I (1984). Reduction in incidence of coronary heart disease, *Journal of the American Medical Association* **251**, 351–364.

[50] Marshall, R.J. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data, *Journal of Clinical Epidemiology* **43**, 941–947.

[51] Martin, L.J., Su, W., Jones, P.J., Lockwood, G.A., Tritchler, D.L. & Boyd, N.F. (1996). Comparison of energy intakes determined by food records and doubly labeled water in women participating in a dietary intervention trial, *American Journal of Clinical Nutrition* **63**, 483–490.

[52] Miettinen, O.S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. Wiley, New York.

[53] Miettinen, O.S. (1986). Response, *Journal of Chronic Diseases* **39**, 567.

[54] Miettinen, O.S. (1990). The concept of secondary base, *Journal of Clinical Epidemiology* **43**, 1017–1020.

[55] Miettinen, O.S. & Cook, E.F. (1981). Confounding: essence and detection, *American Journal of Epidemiology* **114**, 593–603.

[56] Moolgavkar, S.H. & Knudson, A. (1981). Mutation and cancer: a model for human carcinogenesis, *Journal of the National Cancer Institute* **66**, 1037–1052.

[57] Morganstern, H. & Thomas, D. (1993). Principles of study design in environmental epidemiology, *Environmental Health Perspectives* **101**, 23–38.

[58] Multiple Risk Factor Intervention Trial (MRFIT) Research Group (1982). Multiple risk factor intervention trial: risk factor changes and mortality results, *Journal of the American Medical Association* **248**, 1465–1477.

[59] Pepe, M. & Fleming, T.R. (1991). A nonparametric method for dealing with mis-measured covariate data, *Journal of the American Statistical Association* **86**, 108–113.

[60] Piantadosi, S., Byar, D.P. & Green, S.B. (1988). The ecological fallacy, *American Journal of Epidemiology* **127**, 893–904.

[61] Pierce, D.A., Stram, D. & Vaeth, M. (1990). Allowing for random errors in radiation dose estimates for the atomic bomb survivors, *Radiation Research* **126**, 36–42.

[62] Plummer, M. & Clayton, D. (1993). Measurement error in dietary assessment: an assessment using covariance structured models. Part II, *Statistics in Medicine* **12**, 937–948.

[63] Poole, C. (1990). Would vs. should in the definition of secondary study base, *Journal of Clinical Epidemiology* **43**, 1016–1017.

[64] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in Cox's failure time regression model, *Biometrika* **69**, 331–342.

[65] Prentice, R.L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[66] Prentice, R.L. (1995). Design issues in cohort studies, *Statistical Methods in Medical Research* **4**, 273–292.

[67] Prentice, R.L. (1996). Measurement error and results from analytic epidemiology: Dietary fat and breast cancer, *Journal of the National Cancer Institute* **88**, 1738–1747.

[68] Prentice, R.L. & Breslow, N.E. (1978). Retrospective studies and failure time models, *Biometrika* **65**, 153–158.

[69] Prentice, R.L., Pepe, M. & Self, S.G. (1989). Dietary fat and breast cancer: a quantitative assessment of the epidemiologic literature and a discussion of methodologic issues, *Cancer Research* **49**, 3147–3156.

[70] Prentice, R.L. & Sheppard, L. (1990). Dietary fat and cancer: consistency of the epidemiologic data and disease prevention that may follow from a practical reduction in fat consumption, *Cancer Causes and Control* **1**, 87–97.

[71] Prentice, R.L. & Sheppard, L. (1995). Aggregate data studies of disease risk factors, *Biometrika* **82**, 113–125.

[72] Prentice, R.L., Williams, B.J. & Peterson, A.V. (1981). On the regression analysis of multivariate failure time data, *Biometrika* **68**, 373–379.

[73] Riboli, E. (1992). Nutrition and cancer: background and rationale of the European Prospective Investigation

into cancer and nutrition (EPIC), *Annals of Oncology* **3**, 783–791.

[74] Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods, *Journal of Chronic Diseases, Suppl.* **2**, 139–161.

[75] Robins, J. (1989). The control of confounding by intermediate variables, *Statistics in Medicine* **8**, 679–701.

[76] Robins, J. (1997). Causal inference from complex longitudinal data, in *Latent Variable Modeling and Application to Causality, Springer Lecture Notes in Statistics*, Vol. 120, M. Berkane, ed. Springer-Verlag, New York, pp. 69–117.

[77] Rosner, B., Spiegelman, D. & Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error, *American Journal of Epidemiology* **132**, 734–745.

[78] Rosner, B., Willett, W.C. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error, *Statistics in Medicine* **8**, 1051–1070.

[79] Rossouw, J., Finnegan, L.P., Harlan, W.R., Pinn, V.W., Clifford, C. & McGowan, J.A. (1995). The evolution of the Women's Health Initiative: perspectives from the NIH, *Journal of the American Medical Association* **50**, 50–55.

[80] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown & Co., Boston, MA.

[81] Royal College of General Practitioners (RCGP) Oral Contraception Study (1974). *Oral Contraceptives and Health: An Interim Report*. Pitman, London.

[82] Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization, *Annals of Statistics* **6**, 34–58.

[83] Samuelson, S.D. (1996). A Pseudo-Likelihood Approach to Analysis of Nested Case-Control Studies. *Technical Report No. 2*. Institute of Mathematics, University of Oslo.

[84] Sawaya, A.L., Tucker, K., Tsay, R., Willett, W., Saltzman, E., Dallal, G.E. & Roberts, S.B. (1996). Evaluation of four methods for determining energy intake in young and older women: comparison with doubly labeled water measurements of total energy expenditure, *American Journal of Clinical Nutrition* **63**, 491–499.

[85] Self, S.G., Mauritsen, R.H. & Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models, *Biometrics* **48**, 31–39.

[86] Sepanski, J.H., Knickerbocker, R. & Carrol, R.J. (1994). A semiparametric correction for attenuation, *Journal of the American Statistical Association* **89**, 1366–1373.

[87] Sheppard, L. & Prentice, R.L. (1995). On the reliability and precision of within and between population estimates of relative risk parameters, *Biometrics* **51**, 853–863.

[88] Spiegelman, D. & Gray, R. (1991). Cost efficient study designs for binary response data with Gaussian covariate measurement error, *Biometrics* **47**, 851–870.

[89] Stemmerman, G.N., Nomura, A.M.Y. & Heilbrun, L.K. (1984). Dietary fat and risk of colorectal cancer, *Cancer Research* **44**, 4633–4637.

[90] Thomas, D., Stram, D. & Dwyer, J. (1993). Exposure measurement error: influence on exposure–disease relationships and methods of correction, *Annual Review of Public Health* **14**, 69–93.

[91] Wacholder, S. (1991). Practical considerations in choosing between the case–cohort and nested case-control designs, *Epidemiology* **2**, 155–158.

[92] Walker, A.M. (1982). Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known, *Biometrics* **38**, 1025–1032.

[93] Walter, S.D. & Irwig, L.M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, *Journal of Clinical Epidemiology* **41**, 923–927.

[94] Wang, C.Y., Hsu, L., Feng, Z.D. & Prentice, R.L. (1997). Regression calibration in failure time regression with surrogate variables, *Biometrics*, 131.

[95] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

[96] White, J.E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

[97] Whittemore, A.S. (1981). Sample size for logistic regression with small response probabilities, *Journal of the American Statistical Association* **76**, 27–32.

[98] Whittemore, A.S. (1989). Errors-in-variables regression using Stein estimates, *American Statistician* **43**, 226–228.

[99] Whittemore, A.S. & Keller, J.B. (1978). Quantitative theories of carcinogenesis, *SIAM Review* **20**, 1–30.

[100] Whittemore, A.S. & McMillan, A. (1982). Analyzing occupational cohort data: application to US uranium miners, in *Environmental Epidemiology: Risk Assessment*, R.L. Prentice & A.S. Whittemore, eds. SIAM, Philadelphia, pp. 65–81.

[101] Willett, W.C. (1989). *Nutritional Epidemiology*. Oxford University Press, New York.

[102] Willett, W.C., Hunter, D.J., Stampfer, M.J., Colditz, G., Manson, J.E., Spiegelman, D., Rosner, B., Hennekens, C.H. & Speizer, F.E. (1992). Dietary fat and fiber in relation to risk of breast cancer, *Journal of the American Medical Association* **268**, 2037–2044.

[103] Women's Health Initiative Study Group (1997). Design of the Women's Health Initiative Clinical Trial and Observational Study, *Controlled Clinical Trials*, in press.

ROSS L. PRENTICE

Collapsibility

Consider the $I \times J \times K$ **contingency table** representing the joint distribution of three discrete variables, X, Y, and Z, the $I \times J$ marginal table representing the joint distribution of X and Y, and the set of conditional $I \times J$ subtables (strata) representing the joint distributions of X and Y within levels of Z. A measure of association of X and Y is said to be *strictly collapsible* across Z if it is constant across the conditional subtables *and* this constant value equals the value obtained from the marginal table. Noncollapsibility (violation of collapsibility) is sometimes referred to as **Simpson's paradox** after a celebrated article by Simpson [12], but the same phenomenon had been discussed by earlier authors, including Yule [15]; see also Cohen & Nagel [3]. The term *collapsibility*, however, seems to have arisen later in the work of Bishop and colleagues; see Bishop et al. [1].

Table 1 provides some simple examples. The difference of probabilities that $Y = 1$ (the risk difference) is strictly collapsible. Nonetheless, the ratio of probabilities that $Y = 1$ (the risk ratio) is not strictly collapsible because the risk ratio (*see* **Relative Risk**) varies across the Z strata, and the **odds ratio** is not collapsible because its marginal value does not equal the constant conditional (stratum-specific) value. Thus, collapsibility depends on the chosen measure of **association**.

Now suppose that a measure is not constant across the strata, but that a particular summary of the conditional measures does equal the marginal measure. This summary is then said to be *collapsible* across Z. As an example, in Table 1 the risk ratio standardized to the marginal distribution of Z is

$$\frac{\begin{array}{c}\Pr(Z=1)\Pr(Y=1|X=1,Z=1)\\+\Pr(Z=0)\Pr(Y=1|X=1,Z=0)\end{array}}{\begin{array}{c}\Pr(Z=1)\Pr(Y=1|X=0,Z=1)\\+\Pr(Z=0)\Pr(Y=1|X=0,Z=0)\end{array}}$$
$$= \frac{0.50(0.80)+0.50(0.40)}{0.50(0.60)+0.50(0.20)} = 1.50,$$

equal to marginal (crude) risk ratio. Thus, this measure is collapsible in Table 1. Various tests of collapsibility and strict collapsibility have been developed [4, 7, 14].

The definition of collapsibility also extends to **regression** contexts. Consider a **generalized linear model** for the regression of Y on three regressor vectors $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Z}$:

$$g[\mathrm{E}(Y|\mathbf{W}=\mathbf{w},\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})] = \alpha + \mathbf{w}\boldsymbol{\beta} + \mathbf{x}\boldsymbol{\gamma} + \mathbf{z}\boldsymbol{\delta}.$$

The regression is said to be *collapsible* for $\boldsymbol{\beta}$ over $\mathbf{Z}$ if $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ in the regression omitting $\mathbf{Z}$ [2],

$$g[\mathrm{E}(Y|\mathbf{W}=\mathbf{w},\mathbf{X}=\mathbf{x})] = \alpha^* + \mathbf{w}\boldsymbol{\beta}^* + \mathbf{x}\boldsymbol{\gamma}^*.$$

Thus, if the regression is collapsible for $\boldsymbol{\beta}$ over $\mathbf{Z}$ and $\boldsymbol{\beta}$ is the parameter of interest, then $\mathbf{Z}$ need not be measured to estimate $\boldsymbol{\beta}$. If $\mathbf{Z}$ is measured, however, tests of $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ can be constructed [2, 9].

The preceding definition generalizes the original contingency-table definition to arbitrary variables. However, there is a technical problem with the regression definition: if the first (full) model is correct, then it is unlikely that the second (reduced) regression will follow the given form. If, for example, Y is Bernoulli and g is the logit link function, so that the full regression is first-order **logistic**, the reduced regression will not follow a first-order logistic model except in special cases. One way around this dilemma

Table 1 Examples of collapsibility and noncollapsibility in a three-way distribution

	$Z=1$		$Z=0$		Marginal	
	$X=1$	$X=0$	$X=1$	$X=0$	$X=1$	$X=0$
$Y=1$	0.20	0.15	0.10	0.05	0.30	0.20
$Y=0$	0.05	0.10	0.15	0.20	0.20	0.30
Risks[a]	0.80	0.60	0.40	0.20	0.60	0.40
Risk differences	0.20		0.20		0.20	
Risk ratios	1.33		2.00		1.50	
Odds ratios	2.67		2.67		2.25	

[a]Probabilities of $Y = 1$.

(and the fact that neither of the models is likely to be exactly correct) is to define the model parameters as the asymptotic **means** of the **maximum likelihood** estimators. These means are well defined and interpretable even if the models are not correct [13].

It may be obvious that, if the full model is correct, $\boldsymbol{\delta} = 0$ implies collapsibility for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ over $\mathbf{Z}$. Suppose, however, that neither $\boldsymbol{\beta}$ nor $\boldsymbol{\delta}$ is zero. In that case, independence of the regressors does not ensure collapsibility for $\boldsymbol{\beta}$ over $\mathbf{Z}$ except when g is the identity or log link [5, 6]; conversely, collapsibility can occur even if the regressors are dependent [14]. Thus, it is not correct to equate collapsibility over $\mathbf{Z}$ with simple independence conditions.

Consider a situation in which the full regression is intended to represent the causal effects of the regressors on Y. One point, overlooked in much of the literature, is that noncollapsibility over $\mathbf{Z}$ (that is, $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$) does not correspond to **confounding** of effects by $\mathbf{Z}$ unless g is the identity or log link. That is, it is possible for $\boldsymbol{\beta}$ to represent unbiasedly the effect of manipulating $\mathbf{W}$ within levels of $\mathbf{X}$ and $\mathbf{Z}$, and, at the same time, for $\boldsymbol{\beta}^*$ to represent unbiasedly the effect of manipulating $\mathbf{W}$ within levels of $\mathbf{X}$, even though $\boldsymbol{\beta}^* \neq \boldsymbol{\beta}$. Such a divergence is easily shown for logistic models, and points out that noncollapsibility does not always signal a **bias**. In the literature on **random effects** logistic models, the divergence corresponds to the distinction between cluster-specific and population-averaged effects [10]. The cluster-specific model corresponds to the full model in which $\mathbf{Z}$ is an unobserved cluster-specific random variable independent of $\mathbf{W}$ and $\mathbf{X}$, with mean zero and unit variance; $\boldsymbol{\delta}^2$ is then the vector of random-effects variances. For further discussion and an example in which confounding and noncollapsibility diverge, (*see* **Confounding**).

References

[1] Bishop, Y.M.M., Feinberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[2] Clogg, C.C., Petkova, E. & Shihadeh, E.S. (1992). Statistical methods for analyzing collapsibility in regression models, *Journal of Educational Statistics* **17**, 51–74.

[3] Cohen, M.R. & Nagel, E. (1934). *An Introduction to Logic and the Scientific Method*. Harcourt, Brace & Company, New York.

[4] Ducharme, G.R. & LePage, Y. (1986). Testing collapsibility in contingency tables, *Journal of the Royal Statistical Society, Series B* **48**, 197–205.

[5] Gail, M.H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds. Wiley, New York.

[6] Gail, M.H., Wieand, S. & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates, *Biometrika* **71**, 431–444.

[7] Greenland, S. & Mickey, R.M. (1988). Closed-form and dually consistent methods for inference on collapsibility in $2 \times 2 \times K$ and $2 \times J \times K$ tables, *Applied Statistics* **37**, 335–343.

[8] Greenland, S., Robins, J.M. & Pearl, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science* **14**, 29–46.

[9] Hausman, J. (1978). Specification tests in econometrics, *Econometrica* **46**, 1251–1271.

[10] Neuhas, J.M., Kalbfleisch, J.D. & Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data, *International Statistical Review* **59**, 25–35.

[11] Pearl, J. (2000). *Causality*, Ch. 6, Cambridge University Press, New York.

[12] Simpson, F.H. (1951). The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society, Series B* **13**, 238–241.

[13] White, H.A. (1994). *Estimation, Inference, and Specification Analysis*. Cambridge University Press, New York.

[14] Whittemore, A.S. (1978). Collapsing of multidimensional contingency tables, *Journal of the Royal Statistical Society, Series B* **40**, 328–340.

[15] Yule, G.U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* **2**, 121–134.

SANDER GREENLAND

Collinearity

Collinearity (or "multicollinearity") refers to a high level of **correlation** within a set of **explanatory variables**. In a **regression** modeling situation, if explanatory variables are highly correlated, then regression coefficient estimates may become unstable and not provide accurate measures of the individual effects of the variables. The estimate of the precision of these coefficient estimates is also affected and therefore confidence intervals and **hypothesis tests** are, likewise, affected.

For the estimation of regression coefficients, the columns of the design matrix (*see* **General Linear Model**) must be linearly independent. At an extreme, if two explanatory variables are perfectly linearly associated (i.e. their correlation is equal to 1), then such collinearity is an example of linearly dependent columns in the design matrix, $\mathbf{X}$. While two parameters require estimation (i.e. the regression coefficients for the two explanatory variables), information is not available in the design matrix to estimate both coefficients uniquely. The two individual effects cannot be distinguished as a result of this collinearity. While collinearity typically does not involve completely linearly related explanatory variables, high levels of correlation can still lead to difficulties in coefficient estimation.

It should be noted that this issue pertains to the relationship among explanatory variables which, ultimately, affects the ability to investigate simultaneously the relationship between the response variable and the explanatory variables. Therefore, the identification of potential collinearity problems is usually addressed by examination of the relationships among explanatory variables.

One simple technique for the identification of collinearity is presented in Kleinbaum et al. [1]. The computation of the variance inflation factor (VIF) is suggested. If there are p explanatory variables, each explanatory variable is, in turn, regarded as an outcome variable in a regression equation that includes the remaining $p-1$ explanatory variables. Then, R_j^2 represents the squared residual correlation obtained using explanatory variable j, $j = 1, \ldots, p$, as the response. The VIF is then defined for each such regression as:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}.$$

If there is a strong relationship between the explanatory variable j and the remaining $p-1$ explanatory variables, then R_j^2 is close to 1 and VIF_j is large. It is suggested, in [1], that values of VIF greater than 10 indicate serious collinearity that will affect coefficient and precision estimation.

Collinearity may also be indicated if coefficient estimates from fitting simple regression models of the response with each explanatory variable are substantially different from coefficient estimates from fitting a **multiple regression** model including all explanatory variables. Similarly, if the order in which certain terms are included in the model seriously affects the coefficient estimates for these terms, then collinearity is indicated. Of course, one of the primary purposes of multivariate regression models is to examine the role of explanatory variables having "adjusted" for other variables in the model so that such behavior is not necessarily a problem. However, serious collinearity problems may prohibit a multivariate model from being fitted at all.

If two or more explanatory variables are highly correlated because they represent measurements of the same general phenomenon (e.g. highest attained level of education and current salary are both aspects of socioeconomic status), then collinearity can be addressed by choosing one variable thought to be the most relevant. This variable would then be included in any models and the remaining, so-called redundant, variables would be excluded. The identification of such redundant variables may be difficult, so, alternately, a new variable that combines information on the correlated variables can be derived. This aggregate variable would be included in models instead of all of the component variables.

It is sometimes helpful, particularly when collinearity is created as a result of including polynomial terms (e.g. $\mathbf{X}$ and $\mathbf{X}^2$ are included in a model together) but also in general, to center the original explanatory variables. This is accomplished by computing new explanatory variables that are the original measurements with the means subtracted. Suppose there are n individuals and p explanatory variables measured on each individual, $X_{ji}, i = 1, \ldots, n$, $j = 1, \ldots, p$. Then the new explanatory variables are $Z_{ji} = X_{ji} - \overline{X}_j$. If a quadratic model is of interest, then one would include the terms Z_{ji} and Z_{ji}^2 in the model. In [1], an example of the effectiveness

of such an approach for correcting collinearity is presented.

When **polynomial regression** is being undertaken, then the further step of orthogonalization (*see* **Orthogonality**) of the explanatory variables is also possible and frequently used in some settings. Orthogonalization of more general sets of explanatory variables is possible but not as widely used.

Reference

[1] Kleinbaum, D.G., Kupper, L.L. & Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*, 2nd Ed. PWS-Kent, Boston.

(*See also* **Identifiability**)

G.A. DARLINGTON

Combining *P* Values

Empirical investigations typically involve many **hypotheses**. In **clinical trials**, one may test that the medical treatment is efficacious for each of a variety of medical endpoints such as survival, blood pressure, progression of disease, and so on; in **epidemiological** studies, one may test that an environmental toxin is associated with cancer at each of a number of different geographical locations and/or demographic groups.

Combinations of ***P* values** may be considered when the hypotheses are all related to a common question, such as "Does the drug work?" or "Does the environmental toxin cause cancer?"; such methods are called "**meta-analysis**" [4].

Suppose the **null hypotheses** are stated as H_{0i}: {no difference for test i}; and suppose the test statistic is Z_i (*see* **Hypothesis Testing**). The P value is obtained using the distribution of the test statistic Z_i under H_{0i}; assuming a **standard normal** distribution with cumulative probability distribution function $\Phi(.)$, the lower-tailed, upper-tailed, and two-tailed P values are $p_i = \Phi(z_i)$, $p_i = 1 - \Phi(z_i)$, and $p_i = 2(1 - \Phi(|z_i|))$, respectively.

Example: A Clinical Trial

A clinical study of a drug might result in five distinct measurements of **pain**. Upper-tailed (favoring efficacy of drug) P values are $p_1 = 0.078$, $p_2 = 0.091$, $p_3 = 0.213$, $p_4 = 0.121$, and $p_5 = 0.061$, suggesting a clear pattern of beneficial effects, as a P value of 0.50 corresponds to no difference between drug and placebo. However, none meet the standard 0.05 threshold (*see* **Level of a Test**), and the study might therefore be considered nonefficacious. The conclusion that there is no significant efficacy is troublesome from the "common-sense" standpoint, as it seems extremely unlikely that such a consistent pattern of P values could be observed if the drug truly had no effect.

This article addresses the following questions:

1. Can one combine the P values from the various sources to arrive at stronger evidence than is obtained when considering each P value individually?
2. What are the various methods of combining P values?
3. How do the **power** functions of the different methods compare, that is, which combination function should one use?
4. How does one accommodate **correlations** between test statistics and failed distributional assumptions?

Combining *P* Values to Get More Power

Combination of tests is most relevant when the hypotheses are related. For example, if a drug beneficially affects the progression of disease, then one might argue that it should also beneficially affect survival; thus, the hypotheses are related. Similarly, in an epidemiological study, if there were a carcinogenic effect of an environmental toxin in one geographic site, then one would expect that there would be an effect in other sites as well.

Continuing with the epidemiology example, suppose there are two Z-values, each independent, distributed as $N(0,1)$ when there is no effect of environmental toxin on cancer at site i, where $i = 1, 2$. Under the alternatives, the distributions are $N(\delta_i, 1)$, where δ_i is the noncentrality parameter. The 0.05-level test of H_{0i}: $\delta_i = 0$ in favor of H_{0i}: $\delta_i > 0$ rejects when $z_i > 1.645$; the power of the 0.05-level test of H_{0i}: $\delta_i = 0$, for a fixed alternative $\delta_i > 0$, is given by $1 - \Phi(1.645 - \delta_i)$.

One might rather combine the evidence from both regions into a single test using the sum of the Z's; in this case, the hypothesis is H_0: {no effect of environmental toxin at *either* site}. The test rejects H_0 when $(z_1 + z_2) > 1.645 \times 2^{1/2}$, and the power function is $1 - \Phi\{1.645 - (\delta_1 + \delta_2)/2^{1/2}\}$. The combined test has more power than either component test when $0.41\delta_1 < \delta_2 < 2.41\delta_1$. In the case where the hypotheses are related and the study designs comparable, the noncentrality parameters should also be comparable ($\delta_1 \approx \delta_2$), and thus the combined test should be more powerful.

Z-values are simple transformations of the P values, so the test given above is an example of a P value combination statistic $C = \Sigma\Phi^{-1}(1\text{-}p_i)$. This is called the "Liptak" P value combination test; the Liptak test and others are given in the following section.

P Value Combination Tests

One typically assumes that the P values p_i are independent and **uniformly distributed** under their

respective null hypotheses H_{0i}; these assumptions are made throughout this section. P values often are uniformly distributed when the correct distributional forms are used to model the data; for example, when normality, independence, and homoscedasticity (*see* **Scedasticity**) hold. Independence follows when the P values are computed from data values that can be assumed independent.

Under these assumptions, one can provide critical values for the various tests under the combined hypothesis H_0: {H_{01} true *and* H_{02} true *and* ... *and* H_{0k} true}, that is, under $H_0 = \cap H_{0i}$. Combination tests include:

Fisher Combination Test

This test is due to Fisher [3]. The combined test statistic is

$$C_F = -2\Sigma \ln(p_i). \tag{1}$$

C_F is distributed as **Chi-Square** with 2k **degrees of freedom** under H_0; therefore, the α-level test of H_0 rejects when $C_F \geq {\chi^2}_{1-\alpha,2k}$, and the P value of the combined test is $P({\chi^2}_{2k} \geq C_F)$.

In the example with P values 0.078, 0.091, 0.213, 0.121, and 0.061, $C_F = 22.8065$ and ${\chi^2}_{.95,10} = 18.307$; thus, we can reject the combined null at the $\alpha = 0.05$ level, despite the insignificance of every component P value at the 0.05 level. The P value for the combined test is 0.0115.

However, we should note that the assumption of independence is clearly violated in this example, as the multiple pain measurements are correlated, implying that the P values are also correlated. This Fisher combination procedure is therefore not valid; this example is used for illustrative purposes only. The section "Accommodating Nonuniform and/or Dependent P Values" shows how to accommodate **correlation**.

Liptak Combination Test

This test was originated by Liptak [6].

For this procedure, the P values should be one-sided. As shown in the previous section, one may define

$C_L = \Sigma\Phi^{-1}(1 - p_i)$, for upper-tailed P values, or

$C_L = \Sigma\Phi^{-1}(p_i)$, for lower-tailed P values.

C_L is distributed as N(0, k) under H_0; therefore, the α-level test of H_0 rejects when $C_L \geq k^{1/2}\Phi^{-1}(1 - \alpha)$ for upper-tail tests, and when $C_L \leq k^{1/2}\Phi^{-1}(\alpha)$ for lower-tail tests. The P value is either $1 - \Phi(C_L/k^{1/2})$ or $\Phi(C_L/k^{1/2})$ for upper- and lower-tail tests, respectively.

In the example with P values 0.078, 0.091, 0.213, 0.121, and 0.061, $C_L = 6.266$ and $5^{1/2}\Phi^{-1}(1 - .05) = 3.678$; thus, we reject the combined null at the $\alpha = 0.05$ level. The P value for the combined test is 0.0025.

Again, as in the case of the Fisher combination test, this procedure is not valid because these P values are realizations of correlated quantities. Thus, the example given here is for illustrative purposes only. The section "Accommodating Nonuniform and/or Dependent P Values" gives remedies for correlated P values.

Tippett Combination Test

Named after Tippett [13], this test is also known as the "MinP" test, and is the basis for the **Bonferroni** procedure for **multiple comparisons**.

The test statistic is

$$C_T = k \min(p_i). \tag{2}$$

While the distribution of C_T can be obtained easily under the assumptions of uniformity and independence, this test is most commonly used with the Bonferroni inequality; thus, the α-level test of H_0 rejects when $C_T \leq \alpha$. The test is slightly conservative under independence. The P value for the composite test is exactly C_T.

In the example, we have $C_T = 5(0.061) = 0.305$. This is also the P value for the combined test, and it is not statistically significant.

Despite insignificance, this test is valid in the presence of correlation, because of the Bonferroni inequality.

Sidak Combination Test

This test may be attributed to Sidak [11]. The test is very similar to the Tippett test, but is exact, and not conservative, under the assumptions of uniformity and independence.

The test statistic is

$$C_I = 1 - \{1 - \min(p_i)\}^k. \tag{3}$$

The null distribution of C_I is the uniform (0,1) distribution; thus, as with C_T, the α-level test of H_0 rejects when $C_I \leq \alpha$. The P value for the composite test is exactly C_I.

In our example, we have $C_I = 1 - (1 - 0.061)^5 = 0.2700$, which is also the P value of the combined test. This P value is smaller than the Tippett P value, but still insignificant.

This test is also valid in the presence of positive correlation by the Sidak [11] inequality.

Simes' Combination Test

The Sidak and Liptak tests are clearly inferior in that they exclude information in all but one of the P values. The Simes combination test is similar to the Tippett test, but uses information in all of the P values.

Let the ordered P values be $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$. The test statistic is

$$C_S = \min_i \left\{ \frac{kp_{(i)}}{i} \right\}. \tag{4}$$

Since $C_T = kp_{(1)}$, we have $C_S \leq C_T$, and the Simes test is thus uniformly more powerful than the Tippett test. Simes [12] showed that, under independence and uniformity, the distribution of C_S is uniform on (0,1); thus, the α-level test of H_0 rejects when $C_S \leq \alpha$ and the P value for the composite test is exactly C_S.

In the example, $C_S = \min\{5(0.061)/1,\ 5(0.078)/2,\ 5(0.091)/3,\ 5(0.121)/4,\ 5(0.213)/5\} = 0.1513$. This is also the P value for the test, and, while smaller than C_T and C_I, remains insignificant.

This test is valid in the presence of positive correlation by Sarkar's [10] inequality.

Weighted Combinations and other Generalizations

Some hypotheses might be more important than others, in which case weights $w_1, \ldots, w_k$ can be preassigned to the P values. A weighted Fisher combination test statistic can be given by $C_{wF} = -2\ \Sigma w_i \ln(p_i)$, a weighted Liptak test statistic by $C_{wL} = \Sigma w_i \Phi^{-1}(1 - p_i)$, and a weighted Tippett test statistic by $C_{wT} = \min(p_i / w_i)$. Weighted versions of the Tippett test are discussed by Westfall and Krishen [14], and weighted versions of Simes' test are discussed by Benjamini and Hochberg [1]. Another type of combined P value test is the "cutoff" test, which utilizes the ordered P values and may be considered a generalized Simes test [5]. The "truncated product" method has also been proposed recently as a modification of Fisher's combination test, which cures some of its deficiencies [17].

Critical values for generalized combination tests are obtained simply under the assumptions of independence and uniformity; in more complex cases, refer to the section "Accommodating Nonuniform and/or Dependent P Values" below.

Power Comparisons

Assume that the data are, for $i = 1, 2$, $Z_i \sim \mathrm{N}(\delta_i, 1)$, independent, with upper-tailed $\alpha = 0.05$ tests. Power for all procedures is then a function of (δ_1, δ_2).

We can dismiss the Tippett test since it is always less powerful than the Sidak and Simes tests. Further, the Simes test has a very similar power function as the Sidak test; but the Simes test is slightly more powerful than the Sidak test for all cases except $(\delta_1, \delta_2) \in \{\text{very large}, \sim 0+\}$ and $(\delta_1, \delta_2) \in \{\sim 0+, \text{very large}\}$; we thus exclude the Sidak test as well.

Figure 1 shows regions in the positive orthant where each of the methods Fisher, Liptak, and Simes dominate.

Larger Numbers of Tests and Many null effects

Figure 1 shows that the Liptak and Fisher tests are better when the evidence from the two tests is "reinforcing", while the Simes test is better when one effect is large and the other small. While this type of power behavior might suggest that Fisher and Liptak are generally superior in situations where combinations are desired, caution is urged. When the number of tests is larger, with many hypotheses nearly null, or even worse, with one-sided P values in the unanticipated direction, the power of the Liptak and Fisher tests can be much less than that of the Simes, Tippett, and Sidak tests.

Accommodating Nonuniform and/or Dependent *P* Values

In practice, the assumptions of uniformity and independence are not likely to hold. If the P values

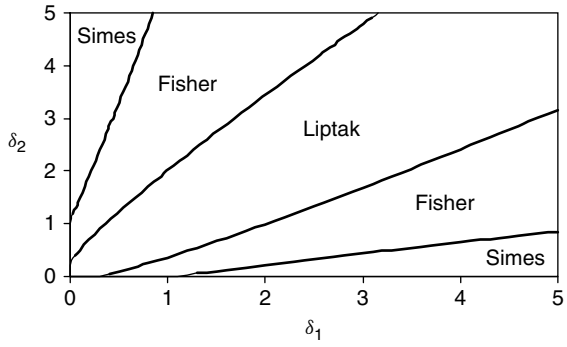

Figure 1 Regions of highest power

are uniformly distributed but dependent, some of the combination tests remain valid. The Tippett combination test is valid, in the sense that the type I error is less than the nominal α in these cases; this follows from the Bonferroni inequality. The Simes combination test has been shown to be conservative as well, under conditions of positive correlation of the P values [10]. However, whereas the Tippett test becomes extremely conservative in cases of highly correlated P values, the Simes test becomes less conservative, and is in fact exact in the case of perfect dependence. The Liptak test can be modified to account for dependence structure; in this case, it is often easier to work directly with the Z statistics rather than the P values; the resulting tests are often called "O'Brien tests" [7].

If the P values are nonuniformly distributed, as typically happens when the distributional assumptions about the data are wrong, it is harder to specify the effects on the combined tests. An exception is the case of discrete (hence nonuniform) **exact** tests, in which case combination tests are often conservative [9, 15].

One can use resampling techniques to obtain more robust, accurate, and in some cases, exact P values for combination tests that fail to satisfy uniformity, dependence, or both. In particular, if the tests come from multivariate data from two groups, and if the global hypothesis refers to equality of the two **multivariate distributions**, one may test the global hypothesis by permutation sampling (*see* **Randomization Tests**) as follows:

(a) Denote the value of the original combination test statistic by C (assume that larger C favor the alternative).

(b) Permute the group labels of the observation vectors, and recompute the combination test statistic from the resulting multivariate data set with permuted group labels. Call the recomputed test statistic C^*.

(c) Repeat (a) to (b) B times, noting whether $C^* \geq C$ for each permuted sample.

The permutation-based P value for the combination tests is then {number of samples for which $C^* \geq C\}/B$. This P value accommodates nonnormality via resampling (nonnormal characteristics of the data are reflected in the sample) and it also accommodates correlations, since the observation vectors remain intact in all permutation samples.

When $B = \infty$, or when the permutations can be enumerated completely, the resulting P value is the exact permutation P value, and the test is called an "exact" test, in the sense of the permutation methods supplied by the software "**StatXact**" [2]. Of course, we cannot take $B = \infty$ in practice, and often the permutations are too numerous to enumerate completely. In these cases, the use of a large B provides good accuracy: If the true P value (as obtained through complete enumeration) is denoted p_C, then the **Monte Carlo** standard error associated with B replications is simply binomial, $\{p_C(1 - p_C)/B\}^{1/2}$. When reporting this standard error, one can substitute the sample-based estimate for p_C.

An entire book is devoted to this subject; see [8] for further discussion and for **algorithms**.

In some cases the permutation approach is not simple, as there may be **covariates**, **survival** functions, and the like, and it may not be clear what should be permuted to represent the combined null hypothesis

H_0. In such cases, **bootstrapping**, either parametric or semiparametric, can help. One needs to model the (usually multivariate) distribution of the data that produced the P values under H_0, and then estimate that distribution under either parametric or semiparametric assumptions. Then one simulates data from the estimated distribution, and computes the resampling-based P value as shown above. Westfall and Young [16, p. 122–3; 214–6] provide examples and further discussion.

References

[1] Benjamini, Y. & Hochberg, Y. (1997). Multiple hypothesis testing and weights, *Scandinavian Journal of Statistics* **24**, 407–418.

[2] Cytel Software Corporation. (2004). `http://www.cytel.com/`.

[3] Fisher, R.A. (1932). *Statistical Methods for Research Workers*, 4th Ed. Oliver & Boyd, Edinburgh.

[4] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York.

[5] Kieser, M., Reitmeir, P. & Wassmer, G. (1995). Test procedures for clinical trials with multiple endpoints, in *Biometrie in der Chemisch-pharmazeutischen Industrie*, J. Vollmar, ed. Gustav Fischer Verlag, Stuttgart, pp. 41–60.

[6] Liptak, T. (1958). On the combination of independent tests, *Magyar Tudomanyos Akademia Matematikai Kutato Intezetenek Közlemenyei* **3**, 171–179.

[7] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1087.

[8] Pesarin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics*. John Wiley & Sons, Chichester.

[9] Rom, D.M. (1992). Strengthening some common multiple test procedures for discrete data, *Statistics in Medicine* **11**, 511–514.

[10] Sarkar, S. (1998). Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture, *Annals of Statistics* **26**, 494–504.

[11] Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions, *Journal of the American Statistical Association* **62**, 626–633.

[12] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **73**, 751–754.

[13] Tippett, L.H.C. (1931). *The Method of Statistics*. Williams and Northgate, London.

[14] Westfall, P.H. & Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures, *Journal of Statistical Planning and Inference* **99**, 25–40.

[15] Westfall, P.H. & Wolfinger, R.D. (1997). Multiple tests with discrete distributions, *The American Statistician* **51**, 3–8.

[16] Westfall, P.H. & Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P value Adjustment*. Wiley, New York.

[17] Zaykin, D.V., Zhivotsky, L.A., Westfall, P.H. & Weir, B.S. (2002). Truncated product method for combining P values, *Genetic Epidemiology* **22**, 170–185.

PETER H. WESTFALL

Commingling Analysis

A mixture distribution with c components ($c \geq 2$), each with probability density function (pdf) $f_i(x)$, has the pdf f_M given by

$$f_M(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \cdots + \pi_c f_c(x).$$

The parameters $\pi_i, i = 1, \ldots, c$, are called the mixing proportions, and they satisfy the constraints $\pi_i \geq 0$ and $\pi_1 + \pi_2 + \cdots + \pi_c = 1$. In this article we primarily discuss normal mixtures with two components. That is, $f_M(x)$ is given by

$$f_M(x) = \pi_1 \varphi(x; \mu_1, \sigma^2) + \pi_2 \varphi(x; \mu_2, \sigma^2),$$

where

$$\varphi(x; \mu, \sigma) = \frac{1}{(2\pi)^{1/2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right].$$

Everitt & Hand [2] and Titterington et al. [11] have published books on the mixture distribution with extensive examples.

Although the distribution is easy to describe and arises from a simple model, it presents formidable mathematical, computational, and statistical difficulties. **Karl Pearson** [9] was the first to study normal mixtures. He used the **method of moments** to estimate the parameters of each of the components. He pointed out that mixtures of **normal distributions** are hard to distinguish from skewed distributions. For normal components with equal **variances**, the **EM algorithm** [1] is effective for finding the **maximum likelihood** estimates of the parameters. Often, a specific sample has a number of solutions to its maximum likelihood equations, and researchers must take care that they have found the global maximum rather than a local maximum. Finch et al. [4] proposed a random search strategy that has an associated measure of the probability of finding an additional solution.

The most basic problem is to distinguish between a sample from a mixture of two normals and a sample from a single normal. That is, the **null hypothesis** is that a **random sample** was drawn from a normal distribution. The **alternative hypothesis** is that a random sample was drawn from a mixture of two normal distribution with equal variances. The asymptotic distribution of the **likelihood ratio test** is not known because the regularity condition that allows the distribution of the likelihood ratio statistic to be expanded in a series about its parameters does not hold [5]. Otherwise, the asymptotic distribution of the likelihood ratio test would be **chi-square** with two **degrees of freedom**. Thode et al. [10] used **simulation** techniques and found that the convergence rate of the distribution of the likelihood ratio statistic was very slow, if it converged at all. They could not exclude the chi-square distribution with two degrees of freedom as the asymptotic distribution. They reported an approximation to the percentiles of the distribution of the likelihood ratio statistic for finite samples. Very large sample sizes are required to detect a normal mixture with high probability [8]. In a study comparing the likelihood ratio test to other tests of normality, Mendell et al. [7] reported that Fisher's **skewness** test and other tests of symmetry are also powerful tests for detecting mixtures with proportions that are not 50–50 splits.

Maclean et al. [6] suggested the use of a (modified) Box–Cox transformation for this problem to reduce the sensitivity of the test to data from skewed distributions (*see* **Power Transformations**). Their alternative hypothesis is that the random sample was drawn from a mixture of two normals after applying a Box–Cox transformation, and their null hypothesis is that the random sample was drawn from a single normal after applying a possibly different Box–Cox transformation. This procedure reduces, but does not eliminate, the sensitivity of tests for mixtures to data from single skewed distributions.

The normal mixture distribution has great interest in exploratory genetic studies. Researchers interpret evidence of a mixture distribution in the measurement of a biological variable as supportive of the hypothesis that the variable is determined genetically. Commingled distributions have been proposed as a logical extension to the genetic model for metric traits that are either **polygenic** (determined by the additive effect of several genes) or multifactorial (polygenic traits determined by many environmental factors as well). For a polygenic trait (i.e. one determined by a large number of loci, each with small, equal, and additive effects), the distribution of the trait is approximately normal by the **central limit theorem**. Falconer [3] illustrates this in his chapter on continuous variation by deriving the distribution of the **genotype** values for a trait determined by dominant alleles (with gene frequency of

0.5) at each of 24 loci. If a trait is determined by codominant alleles, then the distribution would approximate a normal distribution with even fewer loci involved.

The model for continuous variation resulting from a major gene is a mixture of two or three normal distributions. The expected value for each component is the genotype value. That is, if the trait is determined by a major gene, then there are three major genotypes, A_1A_1, A_1A_2, and A_2A_2. Correspondingly, each genotype has, respectively, the expected values $E(A_1A_1)$, $E(A_1A_2)$, and $E(A_2A_2)$, where $E(A_1A_1) < E(A_2A_2)$. Then, if the allele for high values of the trait is dominant to the allele for low values of the trait, $E(A_1A_2) = E(A_2A_2)$. Similarly, if the allele for high values for the trait is recessive to the allele for a low value for the trait, then $E(A_1A_1) = E(A_1A_2)$. The variation of trait values of individuals who are identical in genotype is due to either minor genes or environmental factors. Thus, if a major gene is involved in the determination of a continuous trait, then there are necessarily two or three components. If there is complete dominance (of either A_1 to A_2 or A_2 to A_1), then there are two components in the distribution. If there is additivity (i.e. $E(A_1A_2) = [E(A_1A_1) + E(A_2A_2)]/2$), incomplete dominance (i.e. $E(A_1A_1) < E(A_1A_2) < E(A_2A_2)$ with $E(A_1A_2) \neq [E(A_1A_1) + E(A_2A_2)]/2$), or overdominance (i.e. $E(A_1A_2) < E(A_1A_1)$ or $E(A_1A_2) > E(A_2A_2)$ at this major locus), then there are necessarily three components. The mixing proportions will correspond to the relative frequencies of the genotypes. Thus, the **power** to detect a mixture distribution will depend on the relative frequencies of the A_1 and A_2 alleles and the differences between the genotype means relative to the variability within groups sharing the same genotype (standardized difference between the genotype means).

References

[1] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[2] Everitt, B.S. & Hand, D.J. (1981). *Finite Mixture Distributions*. Chapman & Hall, New York.

[3] Falconer, D.S. (1985). *Introduction to Quantitative Genetics*. Longman, London.

[4] Finch, S.J., Mendell, N.R. & Thode, H.C. (1989). Probabilistic measures of a numerical search for a global maximum, *Journal of the American Statistical Association* **84**, 1020–1023.

[5] Hartigan, J.A. (1985). A failure of likelihood asymptotics for normal mixtures, in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. II, L.M. LeCam & R.M. Olshen, eds. Wadsworth, Belmont, pp. 807–810.

[6] MacLean, C.J., Morton, N.E., Elston R.C. & Yee, S. (1976). Skewness in commingled distributions, *Biometrics* **32**, 695–699.

[7] Mendell, N.R., Finch, S.J. & Thode, H.C. (1993). Where is the likelihood ratio test powerful for detecting two component normal mixtures?, *Biometrics* **49**, 907–915.

[8] Mendell, N.R., Thode, H.C. & Finch, S.J. (1991). The likelihood ratio test for the two component mixture problem: Power and sample size analysis, *Biometrics* **47**, 1143–1148.

[9] Pearson, K. (1894). Contribution to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society, Series A* **185**, 71–110.

[10] Thode, H.C., Finch, S.J. & Mendell, N.R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals, *Biometrics* **44**, 1195–1201.

[11] Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1992). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

(*See also* **Segregation Analysis, Mixed Models**; **Segregation Analysis, Complex**)

NANCY ROLE MENDELL & STEPHEN J. FINCH

Committee of Presidents of Statistical Societies (COPSS)

The Committee of Presidents of Statistical Societies (COPSS) was established in 1963 to address common interests and concerns of North American Statistical Societies. The member societies are the **American Statistical Association** (ASA), the Eastern North American Region (ENAR) of the **International Biometric Society** (IBS), and the Institute of Mathematical Statistics (IMS), the Statistical Society of Canada (SSC), and the Western North American Region (WNAR) of the IBS. COPSS consists of the presidents, past presidents, and presidents-elect of each member society plus a chair and a secretary/treasurer appointed by the committee. The COPSS committee provides a forum for member societies to discuss issues important to statistics.

Historically, COPSS worked on shared problems of the member societies and improved intersociety communication. Many of the initial activities of COPSS were seed activities that later blossomed into programs at member societies. For example, COPSS initiated production of statistical directories. COPSS prepared information for students about statistics and profiles of career statisticians, material that later became the "ASA Careers in Statistics" brochure. COPSS sponsored a lecture series and coordinated the calendar of statistical meetings. COPSS helped to found institutions such as the National Institute of Statistical Sciences (NISS). However, the activity that COPSS is most famous for is that of the COPSS awards.

COPSS began by sponsoring one lectureship and now presents four additional awards. COPSS established the R. A. **Fisher Lectureship** in 1963, to honor the contributions of Sir Ronald Aylmer **Fisher** and the work of a present-day statistician. The Fisher Lectureship recognizes the importance of statistical methods for scientific investigations, and the list of past Fisher lecturers well reflects the prestige that COPSS and its member societies place on this award. The lecture is to be broadly based and is to emphasize aspects of statistics and probability that are closely related to scientific collection and interpretation of data, which are areas in which Fisher made outstanding contributions. The lecture is generally published in one of the COPSS society journals. In 1982, COPSS has added a cash prize and a certificate to the lectureship.

COPSS established the George W. Snedecor Award in 1976. This award honors an individual who is instrumental in the development of statistical theory in biometry. Dr. **Snedecor** was a pioneer in improving the quality of scientific methods concerning the use of statistical methodology. The Snedecor award is for a noteworthy publication in biometry within three years of the date of the award. It is awarded biannually. The presidents' Award, also established in 1976, is presented annually to a young member of the statistical community in recognition of outstanding contributions to the profession of statistics. COPSS defined "young" to mean as not yet having reached his or her 41st birthday during the calendar year of the award. The Elizabeth L. Scott Award, established in 1992, recognizes an individual who exemplifies the contributions of Elizabeth L. Scott's lifelong efforts to further the careers of women in academia. Most recently, COPSS, jointly with the Caucus for Women in Statistics, established the Florence Nightingale David Award in 2001. This award recognizes an individual who exemplifies the contributions of F. N. David. Both the Scott and David awards are presented biannually. All awards have a separate COPSS selection committee. All of the awards have a plaque, certificate, and cash prize. Further information on COPSS can be found at the COPSS website at `www.niss.org/copss/`

SALLIE KELLER-MCNULTY & APARNA V. HUZURBAZAR

Communality

Communality is the proportion of a variable's total variance that is accounted for by the common **factors**. In **factor analysis**, a mathematical model is used to explain the interrelationships of a set of p manifest variables by a smaller number of m underlying latent factors that cannot be observed or measured directly. In common factor analysis, the model relating the measurements on p correlated variables to the m postulated uncorrelated common factors ($m < p$) and p unique factors is

$$\begin{aligned} X_1 &= a_1 A + b_1 B + \cdots + m_1 M + u_1 U_1, \\ X_2 &= a_2 A + b_2 B + \cdots + m_2 M + u_2 U_2, \\ &\vdots \qquad\qquad \vdots \\ X_p &= a_p A + b_p B + \cdots + m_p M + u_p U_p, \end{aligned}$$

where $X_1, X_2, \ldots, X_p$ represent the standardized measurement of the p manifest variables, $A, B, \ldots, M$ represent the standardized scores in the uncorrelated common factors, $a_i, b_i, \ldots, m_i, i = 1, \ldots, p$, are the *common* **factor loadings**, $U_1, U_2, \ldots, U_p$ represent the standardized scores on the p unique factors, and $u_1, u_2, \ldots, u_p$ are the *unique factor loadings*. In the above model the m latent factors and the p unique factors are uncorrelated.

Given that the X_i are standardized and that the latent variables are uncorrelated the variance of X_i is

$$\operatorname{var}(X_i) = 1 = a_i^2 + b_i^2 + \cdots + m_i^2 + u_i^2.$$

The communality h_i^2 is defined as the variance "shared" by the common factors. That is,

$$h_i^2 = a_i^2 + b_i^2 + \cdots + m_i^2,$$

where $i = 1, 2, \ldots, p$. The uniqueness of a variable is u_i^2. It represents the proportion of a variable's variance that does not relate to the common factors.

Communalities can be used in a factor analysis to delineate the dimensions that account for the common variance space. In common factor analysis [1] this is achieved by replacing the unities on the diagonal of the correlation matrix with the communalities. There are various quantities that can be used to estimate the communalities. They include the highest correlation of a variable with the rest of the variables in the set, the average correlation of a variable, the squared multiple correlation of a variable with the others, or the communality of a variable obtained from an initial factor analysis with unities in the diagonal. Among these estimates, the squared multiple **correlation** is the most commonly used estimate for the communality. This squared multiple correlation is the square of the multiple correlation of the regression of X_i on all the other variables. It is the square of the Pearson product moment correlation of a variable, X_i, with $\hat{X}_i$, where $\hat{X}_i$ is the linear multiple regression estimate of X_i on the other $p - 1$ variables. Guttman [2] showed that the squared multiple correlation is a lower bound to its communality and it approaches the true communality as the number of variables increases. However, in the situation where there are a number of highly unreliable variables in the analysis, there may be weak relationships among two or three variables, generating small common factors with no considerable significant loadings. These small "real" common factors and the error factors may jointly account for a substantial proportion of the total variance. Since the squared multiple correlation is based on all the variance which a variable has in common with the others, the *effective* communalities, which are based only on the *salient* factors, may be less than the squared multiple correlations. In this case the effective communalities will be the communalities of interest. A detailed discussion on how to estimate and use these effective communalities in factor analysis can be found in Cureton & D'Agostino [1, Chapter 5].

Because the estimation of communalities introduces a source of error, some factor analyses procedures avoid estimating them and deal solely with the off-diagonal estimates of a correlation to estimate the number of factors (m) and the loadings [3].

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[2] Guttman, L. (1940). Multiple rectilinear prediction and the resolution into components, *Psychometrika* **5**, 75–99.

[3] Reyment, R. & Jöreskog, K.G. (1993). *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, Cambridge.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Communicable Diseases

The epidemiology of communicable diseases typically involves an interplay between the natural history of infective organisms, evolving largely within infected individuals, and their transmission dynamics, governed by direct or indirect contacts between individuals. While most fields of epidemiology comprise a social dimension, for infectious diseases this element enters at the level of mechanism, and is therefore central to our understanding of these diseases. The application of statistics to infectious diseases thus requires both a biological and a sociological (*see* **Social Sciences**) perspective. This dialectic is strikingly illustrated by the epidemiology of the human immunodeficiency virus (HIV) (*see* **AIDS and HIV**), the study of which motivated a large-scale investigation of sexual attitudes and lifestyles [70].

The Scope of Statistics in Infectious Disease Epidemiology

In spite of the distinct characteristics of infectious disease epidemiology, many of the statistical methods most commonly used are entirely standard. Thus, for instance, **observational studies** based on **surveillance** data, and epidemiological investigations using **case–control** or **cohort** designs, employ broadly the same statistical methodology whether the disease involved is measles or cancer. Nevertheless, statistical science has made distinct contributions to many areas of infectious disease epidemiology. The most prominent include catalytic and transmission models (*see* **Infectious Disease Models**), the study of the natural history of infectious disease, the detection of infectiousness, and the statistics of vaccination (*see* **Vaccine Studies**). These are the major topics around which this article is organized. In addition, statistics has played a key role in promoting our understanding of specific infectious diseases, as witnessed by the large statistical literature on the acquired immunodeficiency syndrome (AIDS) [21]. Historically, many statistical techniques which are today widely employed were originally developed in the context of infectious disease epidemiology. Thus John Snow's famous study [110] of cholera and his investigation of an outbreak of cases around the Broad Street pump is an early demonstration of space–time clustering. A more recent example is provided by **Bradford Hill's** pioneering influence on field trials of pertussis vaccines [88], of streptomycin against pulmonary tuberculosis [86], and antihistamines against the common cold [87], which did much to establish the randomized controlled trial (*see* **Clinical Trials, Overview**) as a basic tool in medical research. More broadly, many of the principles governing the transmission of infectious diseases are of a more general nature, as originally alluded to by Ross in his "theory of happenings" [103], and similar models have been applied to such diverse topics as the spread of drug addiction and of scientific ideas [32].

Historical Background

The application of formal mathematical methods to infectious diseases dates back to Daniel Bernoulli's 1760 publication [18] (*see* **Bernoulli Family**) of a mathematical model to evaluate the impact of smallpox on **life expectancy** [31]. Later, **William Farr** [36] sought to describe the course of epidemics by fitting curves to smallpox data, and used this technique to predict the course of an epidemic of rinderpest among cattle [37]. **Brownlee** [22] also pursued this approach, fitting **Pearson distribution** curves to outbreak data on numerous diseases in support of his theory, later disproved, that pathogens decline in infectiousness during the course of an epidemic [50] (*see* **Epidemic Models, Deterministic**).

This early phase of largely empirical investigations was followed by the more analytical work of Hamer [66] and Ross [103], who sought to understand the mechanisms governing disease transmission. Hamer first formulated what was later to be known as the mass action principle, according to which the number of cases in generation $t+1$ is proportional to the numbers of susceptibles and infectives in generation t. Ross developed transmission models for malaria, and formulated the first threshold theorem, on the critical density of mosquitoes required for malaria to remain endemic [49]. The ideas of Hamer and Ross were later elaborated by Kermack & McKendrick [74]. **Chain binomial models** of disease spread may be traced back to En'ko [35], who, as argued by Dietz [29], anticipated the Reed–Frost model (1928) (published as Frost [57]).

The fully stochastic treatment (*see* **Stochastic Processes**) of the chain binomial model is primarily due to **Greenwood** [59].

The pioneering approach of Bernoulli was extended by Muench [91, 92], who investigated the age distribution of susceptibles in a population using survival analytic methods (*see* **Survival Analysis, Overview**). Drawing upon an analogy from chemistry, Muench described his models as catalytic. Dietz [28] clarified the connections between transmission models and catalytic models, and demonstrated how key transmission parameters can be estimated from epidemiological data.

Catalytic Models

The great variety in the ecology and natural history of different infections generally requires different models for different diseases. However, some of the key features of all infectious disease models are captured by the simple case of person-to-person transmission of an infection conferring permanent immunity to those infected. The emphasis here is on endemic diseases that have reached a state of dynamic equilibrium. In this state, the long-term average incidence of infection is broadly constant.

The Basic Relationships

Let the random variable X denote the age at which individuals acquire infection. Let $S(x)$ denote the survivor function (*see* **Survival Analysis, Overview**):

$$S(x) = \Pr\{X > x\}$$

and $\lambda(x)$ the age-specific **hazard rate** of infection. In the context of infectious disease epidemiology, this is often called the force of infection, and depends on a variety of factors, including the rate at which individuals come into contact with one another, and the ease with which the organism is transmitted, given a suitable contact.

Assume, for simplicity, that all individuals in the population die at a fixed age, L, the life expectancy, and that the disease concerned is not fatal. In addition, we allow for the possibility that individuals are protected by maternally acquired immunity from birth to some age, M. These assumptions are broadly applicable to common childhood diseases such as measles, mumps, rubella, and whooping cough in developed countries. For these diseases, $\lambda(x)$ is typically a non-negative unimodal function, formally set to zero for $x \leq M$, and

$$S(x) = \exp\left(-\int_0^x \lambda(u)\,\mathrm{d}u\right).$$

The survivor function equals 1 for $x \leq M$, and hence differs from the probability that an individual of age x is susceptible, which is 0 for $x \leq M$ and $S(x)$ for $x > M$. Let A denote the expectation of X, or average age at infection. The average force of infection acting on susceptibles, $\overline{\lambda}$, is related to A by

$$\overline{\lambda} = \frac{\int_M^L \lambda(x)S(x)\,\mathrm{d}x}{\int_M^L S(x)\,\mathrm{d}x} = \frac{1 - S(L)}{A - M}.$$

For endemic infections, the proportion of the population escaping infection is negligible, and hence $S(L)$ is effectively zero. Thus,

$$\overline{\lambda} \doteq \frac{1}{(A - M)}.$$

It follows that if the average force of infection is reduced, for instance by vaccination, then the average age at infection will rise. Assuming a uniform age distribution, the proportion of the population remaining susceptible is

$$\pi = \int_M^L \frac{S(x)}{L - M}\,\mathrm{d}x = \frac{A - M}{L - M}. \qquad (1)$$

Since the duration of protection by maternal antibodies is usually much less than A, the following approximations hold:

$$\pi \doteq \frac{A}{L} \doteq \frac{1}{\overline{\lambda}L}.$$

So far, the methods described are essentially those of survival analysis, and apply to any event occurring with hazard $\lambda(x)$. We now make the connection with infectious diseases.

An important summary measure of the infectiousness of a disease in a given population is the basic **reproduction number**, R_0. This is the mean number of secondary cases generated by a single infective in a totally susceptible population. It is a parameter of fundamental importance in infectious disease

epidemiology. The higher the value of R_0, the more infectious the disease. However, if the basic reproduction number is equal to or below 1, transmission of the infection cannot be sustained and will eventually die out with probability 1.

R_0 depends on the effective contact rate, β. The term "contact" is taken to mean one of such a nature as to enable transmission of infection to occur. This of course depends on the mode of transmission of each organism. The effective contact rate, β, is defined as follows. Let η denote the rate at which contacts occur in the population, that is, the average number of contact that an individual makes per unit time, and let θ denote the conditional probability of infection, given a contact between an infective and a susceptible. Then the effective contact rate is

$$\beta = \eta\theta.$$

It follows from this definition that

$$R_0 = \beta D,$$

where D is the **mean** duration of the infectious period.

In general, the contact rate, η, and hence β and R_0, vary with age. For simplicity, we assume that the population mixes in a homogeneous fashion, so that η is independent of age. Then β, R_0, and the force of infection are also independent of age. For long-established endemic diseases in dynamic equilibrium, the proportion susceptible fluctuates around the constant value π given by (1). Thus the average number of secondary cases produced by one infective is πR_0. But since the disease is in equilibrium, this must on average equal 1, since otherwise the average proportion susceptible will not remain constant. It follows that

$$\pi R_0 = 1,$$

and hence

$$R_0 = \frac{1}{\pi} \doteq \frac{L}{A} \doteq \lambda L,$$

$$\beta = \frac{1}{\pi D} \doteq \frac{L}{AD} \doteq \frac{\lambda L}{D}. \qquad (2)$$

Thus, in a homogeneously mixing population, the basic reproduction number and the effective contact rate may be estimated from the average age at infection, the life expectancy, and the duration of the infectious period.

These fundamental relationships also have important consequences for the control of infectious diseases. If the proportion of susceptible individuals in the population is reduced by vaccination below the equilibrium level π, then the number of secondary cases produced by each infective will be reduced below 1. Thus the infection will no longer be self-sustaining, and will eventually die out. Assuming that vaccination confers complete protection, and letting V denote the minimum proportion that must be vaccinated for eradication of the infection, we therefore have

$$V = 1 - \pi \doteq 1 - \frac{A}{L} \doteq 1 - (\lambda L)^{-1}.$$

In particular, note that it is not necessary to vaccinate the entire population to eradicate infection. This is the phenomenon of herd immunity: the effect of vaccination is not simply to protect vaccinated individuals, but also to impart indirect protection to unvaccinated susceptibles by impeding the circulation of the infection in the population.

This discussion shows how key epidemiologic parameters, such as the basic reproduction number and the proportion to vaccinate for eradication of infection, may be estimated from observable quantities such as the hazard of infection. Clearly, the assumption of homogeneous mixing is untenable for many populations. More complex versions of the results described above may be derived for age-dependent mixing patterns [3] and [44]. Similarly, the concepts introduced above in the case of person-to-person transmission have direct counterparts for other types of infection. For instance, in the case of helminth infections, the force of infection is the rate at which uninfected individuals acquire parasites, and R_0 is the average number of offspring produced throughout the reproductive life span of a mature parasite that themselves survive to maturity, in the absence of density-dependent constraints on population growth [3].

Estimation of the Force of Infection

As shown above, knowledge of the age-specific force of infection, or infection hazard $\lambda(x)$, is central to control strategies for infectious diseases. In the case of endemic infections conferring long-lasting immunity, with no differential mortality, in unvaccinated

populations, it is most readily estimated from serological surveys, in which a **cross-sectional** sample of the population is tested for the presence or absence of relevant antibodies. Suppose that n_x individuals of age x are tested, with $x > M$, the duration of protection from maternal antibodies. The number r_x who display no evidence of past infection may be regarded as **binomial** $[n_x, S(x)]$, where $S(x)$ is the survivor function. Given a parametric form for the hazard function, the **likelihood** is proportional to

$$\prod_{x=1}^{L} S(x;\beta)^{r_x}[1 - S(x;\beta)]^{n_x - r_x}, \qquad (3)$$

where

$$S(x;\beta) = \exp\left(-\int_0^x \lambda(u;\beta)\,\mathrm{d}u\right).$$

Individuals below age M are excluded from this analysis because they may be protected by maternally acquired antibodies. The decline of maternal protection after birth is of course of interest in its own right, and may be modeled provided sufficient data are available on infants. Note also that in the presence of vaccination, the hazard function can no longer be interpreted as the hazard of infection, but is the combined effect of natural infection and immunization.

For diseases with short **incubation periods**, the force of infection may also be estimated from case reports or routine notification data, provided it is reasonable to assume that these are not subject to age-specific bias in diagnosis or reporting. It is also commonly assumed that the age distribution of the population from which the cases are drawn is **uniform**, at least up to some age x_k above which few infections occur. Let $\mathbf{n} = (n_1, \ldots, n_k)$ denote the numbers of cases in the different age groups, where the subscript i indexes the age range $[x_{i-1}, x_i), i = 1, \ldots, k$. Thus $\mathbf{n}$ is **multinomial** and the likelihood is proportional to

$$\prod_{i=1}^{k}\left[\frac{S(x_{i-1}|\beta) - S(x_i|\beta)}{1 - S(x_k|\beta)}\right]^{n_i}.$$

Parametric models for the force of infection are discussed by Griffiths [62], Grenfell & Anderson [61], and Farrington [38]. Keiding [73] discusses **nonparametric** estimation, with an application to hepatitis A data (*see* **Hepatology**). Alternately, the hazard may be assumed piecewise constant, taking the value λ_j in age group $[x_{j-1}, x_j)$, $j = 1 \ldots k$. This enables **regression** models to be fitted using standard modeling software. Given fixed **covariates z** and letting $t_j(x)$ denote the time spent by an individual of age x in the jth interval, we have

$$\ln[S(x|\beta, \lambda)] = \beta^{\mathrm{T}} z + \sum_{j=1}^{k} \lambda_j t_j(x),$$

and hence the likelihood (3) may be maximized by **generalized linear modeling** with logarithmic link function.

The force of infection is important in its own right, but also in estimating age-dependent contact rates and the basic reproduction number R_0. Statistical methods for estimating the R_0 from serological and other data are described in [19, 30, 44].

The methods described above apply to infections having reached a dynamic equilibrium, in which the age-specific incidence fluctuates around a constant value according to stable epidemic cycles. For emerging infections, such as HIV, the age-specific incidence will initially increase over time. For others, such as hepatitis A, it may decline as contact rates change. The methods described above may be extended to model secular changes in the hazard of infection, using data from sequential seroprevalence surveys in the same population. In this context the hazard $\lambda(x, t)$ depends on both age and time, and the survivor function is

$$S(x, t) = \exp\left(-\int_0^x \lambda(u, t - x + u)\,\mathrm{d}u\right).$$

Ades & Nokes [2] discuss an application to the incidence of toxoplasma infection. In some circumstances data may be available on repeat tests for the same individuals. One example is repeat tests for HIV on attenders at genito-urinary medicine clinics. This sampling scheme gives rise to **interval-censored** data, in which the time of infection is bracketed between the times of the last negative and the first positive tests. Denoting these, respectively, by U and V, where U may be zero (denoting a left-censored observation) and V may be infinite (denoting a right-censored observation), the likelihood is

$$\prod_{i=1}^{n}[S(U_i) - S(V_i)].$$

The estimation method based on case reports can similarly be extended to incorporate secular changes in incidence, provided the incubation period of the disease is short. For diseases with long incubation periods, for example AIDS and the chronic sequelae of hepatitis C infection, symptoms may appear many years after infection, and hence case reports alone provide little information on the date or age of infection. Data on case reports must be combined with information about the incubation period to estimate the incidence function. This is the **back calculation** approach, developed by Brookmeyer & Gail [20] to model the incidence of HIV infection.

For simplicity we ignore age effects. Let $\mu(t)$ denote the rate at which case reports arise, and $F(t)$ denote the distribution function for the incubation period. These are related to the force of infection by the convolution equation

$$\mu(t) = \int_{-\infty}^{t} \lambda(s)S(s)F(t-s)\,\mathrm{d}s.$$

The product $\lambda(s)S(s)$ is sometimes combined into a single term, $\alpha(s)$, denoting the incidence of infection in the population. For uncommon or emerging infections, virtually the entire population is susceptible, hence $\alpha(s)$ and $\lambda(s)$ are practically identical.

Many methods have been proposed to estimate $\lambda(t)$ from this basic equation, given observed case reports and knowledge of the incubation period distribution. A comprehensive account is given in Brookmeyer & Gail [21]. For instance, given a parametric form $\alpha(s;\beta)$ for the incidence curve and assuming that infections arise in a **Poisson process**, then the number of cases, Y_i, arising in time period $[t_{i-1}, t_i)$ is also Poisson with mean

$$\mu_i = \int_{t_{i-1}}^{t_i} \alpha(s;\beta)[F(t_i - s) - F(t_{i-1} - s)]\,\mathrm{d}s,$$

and hence the parameters β may be estimated by maximizing the Poisson log likelihood:

$$\sum_{i=1}^{n} [n_i \ln(\mu_i) - \mu_i - \ln(n_i!)].$$

Transmission Models

Catalytic models describe the age distribution of susceptibles for an endemic infection in dynamic equilibrium. However, they only capture the steady-state characteristics of the infection process, averaged over a long period of time. In particular, they fail to account for epidemic cycles. These constitute one of the most striking features of endemic diseases with short incubation periods, at least those conferring lifelong immunity and not involving a carrier state. In addition, catalytic models cannot be used for infections that have not reached a dynamic equilibrium, such as emerging diseases. By contrast, transmission models, whether deterministic or stochastic, attempt to incorporate the mechanism by which infection is spread in a population.

Dynamic Models for Large Populations

As seen above, the mechanism of disease spread is governed by the effective contact rate, β. This may be expressed more generally as a function $\beta(u, v)$ representing the number of effective contacts per unit time between an individual of age v and individuals of age u. The relationship between the contact rate and the force of infection at age x and at time t depends on the number of infectives in the population at time t. An intuitively appealing, though by no means unique, functional relationship is

$$\lambda(x,t) = \int_0^{\tau} \beta(u,x)Y(u,t)\,\mathrm{d}u,$$

where $Y(x,t)$ is the proportion of infectives of age x in the population at time t.

The basic ideas behind dynamic models may be illustrated using the simple example of a homogeneously mixing population, that is, one in which the contact rate is a constant β. It follows that the force of infection is also independent of age.

The population, of constant size, is divided into proportions $X(t)$ susceptible, $Y(t)$ infective, and $Z(t)$ recovered, who are immune from further infection. Individuals are born into the susceptible class with constant rate μ, and die at the same rate. There is no disease-associated mortality. The model is driven by the mass action principle, according to which the number of new infectives in a small time interval $[t, t+\delta t)$ is proportional to the number of infectives and to the number of susceptibles at time t. The dynamics of this three-stage model, often called a SIR (for susceptible–infectious–recovered) model, may be represented by the following system of differential

equations:

$$\begin{aligned}\dot{X}(t) &= -\beta X(t)Y(t) + \mu - \mu X(t),\\ \dot{Y}(t) &= \beta X(t)Y(t) - \gamma Y(t) - \mu Y(t),\end{aligned} \tag{4}$$

in which the dots represent derivatives with respect to t and γ is the reciprocal of the duration of the infectious period, D. Setting the derivatives to zero and solving for $X(t)$ gives the nontrivial equilibrium proportion susceptible:

$$\pi = \frac{\gamma + \mu}{\beta} \doteq \frac{1}{\beta D},$$

where the approximation is valid provided the duration of infectiousness is much smaller than life expectancy. Note that expression (2) is retrieved only approximately, owing to the different assumptions about the death rate.

Using standard methods for the analysis of small departures from equilibrium, it can be shown that these result in oscillations in the numbers of infectives. When the duration of the infectious period is small compared with the average age at infection, these have period

$$T = 2\pi(AD)^{1/2}. \tag{5}$$

This simple model thus exhibits the well-known phenomenon of epidemic cycles, typical of endemic immunizing infections with short infectious periods and no carrier state.

The second equation in (4) also illustrates the fact that the number of cases only grows if $\beta X(t) - \gamma - \mu > 0$, and hence in an initially susceptible population an epidemic can only occur if

$$R_0 = \frac{\beta}{\gamma + \mu} > 1.$$

Clearly, more sophisticated models can be developed to incorporate a latent period, protection by maternal antibodies, age-dependence, etc. In particular, if an infection has latent period E, D in expression (5) should be replaced by $D + E$. Anderson et al. [4] apply **spectral analysis** to the **time series** of some common childhood infections and show that the epidemic periods broadly correspond to those predicted by the model. More complex modeling techniques, combining time series methods with epidemic modeling, have been described by Finkenstädt & Grenfell [54].

The transmission of infection in a population clearly involves a stochastic component, which is not captured by deterministic models. In small populations, stochastic effects become dominant, and deterministic models are of little use. In particular, they cannot readily account for the phenomenon of extinction, in which the transmission of infection is interrupted.

The stochastic version of (4) may be developed in terms of transition probabilities. Thus, letting $X'(t)$ and $Y'(t)$ denote, respectively, the number (rather than the proportions) of susceptibles and infectives at time t in a population of fixed size N, the corresponding transition probabilities in a short interval $(t, t + \delta t)$ are

$$\begin{aligned}\Pr[X'(t+\delta t) &= X'(t) - 1; Y'(t+\delta t) = Y'(t) + 1]\\ &= N^{-1}\beta X'(t)Y'(t)\delta t,\\ \Pr[X'(t+\delta t) &= X'(t); Y'(t+\delta t) = Y'(t) - 1]\\ &= (\gamma + \mu)Y'(t)\delta t,\\ \Pr[X'(t+\delta t) &= X'(t) + 1; Y'(t+\delta t) = Y'(t)]\\ &= N\mu\delta t,\\ \Pr[X'(t+\delta t) &= X'(t) - 1; Y'(t+\delta t) = Y'(t)]\\ &= \mu X'(t)\delta t.\end{aligned}$$

Unfortunately the solution of such systems is far from straightforward. There is a substantial literature on the properties of general stochastic models, much of it of a highly mathematical nature [5–10, 26] and [112]. In practice, in large populations analytic solutions for the stochastic model become unmanageable. Using **Monte Carlo methods**, Bartlett [9, 10] showed that a minimum population size is required for an infection to remain endemic.

The main purpose of the deterministic and stochastic dynamic models discussed above is to exhibit the qualitative aspects of the spread of infectious disease and to explore threshold phenomena. One important application is in predicting the impact of vaccination strategies. However, for the statistical purposes of hypothesis testing and parameter estimation, different types of stochastic models are used.

Branching Processes

In some cases it is possible to ignore the depletion of susceptibles, for instance during the initial stages

of an epidemic. In these circumstances the spread of infection may be modeled by means of a **branching process** in discrete time [14, 15]. This approach is useful when the emphasis is on parameter **estimation**, rather than **prediction**. In such a model the epidemic begins with the introduction of an initial number of infectives, Y_0, at generation 0. These infect Y_1 individuals, who comprise the next generation of cases. In turn, these infect Y_2 individuals in the next generation of cases, and so on. Suppose that the number of infections directly caused by one individual is a random variable Z with probability distribution $g(z)$, the offspring distribution, and that the numbers of infections caused by two cases from the same generation are independent. Thus, for each i, $Y_i = Z_1 + \cdots + Z_{Y_{i-1}}$, where the Z_j are independent variables with density $g(z)$.

Let μ and σ^2 denote the mean and **variance**, respectively, of the offspring distribution. Thus μ is the expected number of infections caused by one case, and plays the same role as R_0 above. General results on branching processes show that if $\mu \leq 1$, then the process will become extinct with probability one [67]. Inference for branching processes is usually conditional on extinction or nonextinction.

For endemic diseases we proceed conditionally on nonextinction. Generation sizes $Y_0, \ldots, Y_k$ are observed for some value of k. In practice k is usually small, because generations soon become indistinguishable. Also the stock of susceptibles may be depleted, thus rendering the branching process model invalid. Harris [67] proposed the following **nonparametric maximum likelihood** estimator for μ:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} Y_i}{\sum_{i=1}^{k} Y_{i-1}}.$$

The properties of this estimator are discussed by Keiding [72]. Heyde [68] and Dion [33] discuss nonparametric interval estimation (*see* **Estimation, Interval**). Alternatively, a parametric assumption may be made about the offspring distribution $g(z)$. Becker [14] suggests the alternative estimator:

$$\hat{\mu} = \begin{cases} (y_k/y_0)^{1/k}, & \text{if } I_k > 0, \\ 1, & \text{if } y_k = 0. \end{cases}$$

For outbreaks of diseases for which $\mu \leq 1$, it can be shown that if the offspring distribution has a power series distribution, then the total size of the outbreak also has a power series distribution [12], and the total outbreak size is a **sufficient statistic** for μ. The **maximum likelihood** estimate of μ is

$$\hat{\mu} = 1 - \frac{Y_0}{Y_+},$$

where Y_+ is the outbreak size. In particular for **Poisson offspring distributions**, the total number of cases follows the Borel–Tanner distribution and the asymptotic variance of the maximum likelihood estimator of μ is $\mu(1-\mu)/Y_+$.

Heyde [69] discusses a **Bayesian** approach which allows the cases $\mu > 1$ and $\mu \leq 1$ to be treated without distinction. Becker [12, 13, 15] discusses several applications of branching processes to smallpox epidemics. Branching process models have also been proposed for the surveillance of vaccination programs [45].

Chain Binomial Models and Extensions

The branching process offers a simple framework in which to investigate the early stages of epidemics, and outbreaks of nonendemic infections. However, its applicability is limited to situations in which it is reasonable to assume an unlimited pool of susceptibles. In particular, branching processes are unsuitable for the analysis of disease spread within households and small communities. In this context, a more appropriate framework is provided by **chain binomial models**.

Consider a household with n individuals. At generation k there are X_k susceptibles exposed to Y_k infectives. The distribution of the number of cases in the next generation, Y_{k+1}, conditional on X_k and Y_k, is **binomial**:

$$\Pr(Y_{k+1} = z | X_k = x, Y_k = y) = \frac{x!}{z!(x-z)!} p_k^z (1-p_k)^{x-z},$$

where p_k is the probability that a susceptible of generation k will acquire infection from one of the y_k infectives. To parameterize p_k, some assumptions are required. A common assumption due to Reed &

Frost (1928), published as [57], is that contacts with infectives occur independently, so that

$$p_k = 1 - (1 - \pi)^{y_k},$$

where π is the probability of an effective contact between two individuals. In continuous time, the mass action principle as incorporated in (4) coincides with the Reed–Frost model, in which π is replaced by $\beta \cdot \delta t$, where β is the contact rate. An alternate assumption, due to Greenwood [59], is

$$p_k = \begin{cases} \pi, & \text{if } y_k \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

The Greenwood model may be valid for diseases such as measles, in which infectious material is spread by aerosol. In these circumstances the number of infectives present will have little bearing on the number of susceptibles infected.

In some cases, data may be available on the actual chains of infection within the household, thus enabling the full **likelihood** to be written down. Thus, for one household with x_0 initial susceptibles the likelihood is

$$L = \prod_{i=1}^{m} \Pr(y_i | x_0, y_0, \ldots, y_{i-1}),$$

where m is the total number of generations of cases in the household. For instance, in a household of size three, with one initial infective and two initial susceptibles, the possible chains of infection are {1}, {1, 1}, {1, 1, 1}, {1, 2}. The corresponding probabilities are, respectively, $(1 - \pi)^2$, $2\pi(1 - \pi)^2$, $2\pi^2(1 - \pi)$, π^2. In this case the probabilities are the same under the Reed–Frost and Greenwood assumptions, although this is not generally the case. Bailey [5] contains tables of probabilities for households of up to five.

The parameter π may be estimated by directly maximizing the likelihood L. Alternately it may be maximized using **generalized linear modeling** techniques, since y_{k+1} is conditionally binomial(x_k, p_k). This enables household characteristics to be modeled in a straightforward manner. In the case of the Reed–Frost model, the appropriate link function is the complementary log–log, used with the offset $\ln(y_k)$. Thus, if

$$\ln[-\ln(1 - \pi)] = \alpha + \beta^{\mathrm{T}} x$$

for covariates x, then

$$\ln[-\ln(1 - p_k)] = \ln(y_k) + \alpha + \beta^{\mathrm{T}} x. \qquad (6)$$

The Reed–Frost assumption may also be tested formally, since omission of the offset term $\ln(y_k)$ in (6) corresponds to the Greenwood assumption.

In practice it is exceedingly rare to have such detailed data. In some situations, however, data may be available on the size of outbreaks in households. Assuming that information is also available on the numbers of introductory cases and initial susceptibles, the likelihood of an outbreak of any given size may be obtained by summing the probabilities of all chains with that number of cases. Thus, for instance, in a household of size three with one initial infective and two initial susceptibles, the probability of an outbreak of total size three is $2\pi^2(1 - \pi) + \pi^2 = \pi^2(3 - 2\pi)$.

Bailey [5] discusses some of the problems involved in collecting data on household outbreaks. Bailey [5] and Becker [15] describe **model checking** for the chain binomial model, and extensions to random infectiousness, in which the parameter π may vary between households, for instance according to a **beta distribution**. Becker [15] gives a detailed application of chain binomial methods to the common cold.

The chain binomial models so far considered only seek to model the course of disease within households, and ignore the transmission of infection between households. This latter problem has been discussed by Longini et al. [81] and Longini & Koopman [80]. Suppose that an outbreak in a defined community of households is observed. Infections may be acquired within the household, or from the community, that is, from an individual from a different household. Let π_{c} and π_{h} denote, respectively, the probabilities that a susceptible is infected from the community and from the household during the outbreak. Let p_{is} denote the probability that i of the s initial susceptibles within a given household are infected during the outbreak. Expressions for the p_{is} in terms of π_{c} and π_{h} are obtained recursively using the formula

$$p_{is} = \frac{s!}{i!(s - i)!} p_{ii} (1 - \pi_{\mathrm{c}})^{s-i} (1 - \pi_{\mathrm{h}})^{i(s-i)}.$$

Given a **random sample** of households from the community, the likelihood is the product of the p_{is} over the households in the sample and may be

maximized to estimate the probabilities π_c and π_h. Becker [15], Longini & Koopman [80], and Longini et al. [81] give several applications of this method to data on various respiratory diseases. More complex stochastic models have also been proposed to take account of several levels of mixing, for example, within and between households [7].

Data Augmentation, Martingale and Markov Chain Monte Carlo Methods

Likelihood-based methods for data on infectious disease data can become notoriously cumbersome for even medium-size problems. This is because the infection and disease process is generally only partly observed. For example, in outbreaks of infectious diseases, infection times are rarely observed: only disease onsets are reported.

To remedy this problem, Martingale [15] and data augmentation techniques have been proposed [15–17]. Bayesian estimation methods implemented by Markov chain Monte Carlo are particularly suitable for use with missing data. The methods have been applied to infectious disease data, both to simplify existing likelihood techniques, and to extend them by weakening the assumptions [94–96].

The Natural History of Infectious Diseases

An important area of application of infectious disease statistics is in estimating key parameters describing the natural history of infection. This section describes methods for estimating the incubation, latent, and infectious periods, and the risk and severity of clinical disease following infection.

Estimation of the Incubation Period

The **incubation period** is defined as the interval between acquisition of infection and the appearance of symptoms. A related concept is that of generation time, also called the *serial interval*, which is the time between acquisition and transmission of infection. For many infections, such as measles, mumps, or chickenpox, the two are almost identical. Knowledge of the incubation period and generation time are important for several reasons. First, it enables cases in an epidemic to be classified into generations, thus allowing more detailed investigation of the spread of disease. Secondly, in the investigation of point source outbreaks, for instance involving food contaminated with Salmonella, knowledge of the incubation period is essential to define the time period over which food histories and other risk factor information should be collected. Thirdly, for evolving diseases with long incubation periods, the true incidence of infection can only be determined if the incubation period is known.

For diseases with short incubation periods, the incubation period distribution may be estimated directly using information on the time of exposure. In this way Sartwell [104] found that the incubation periods of many common diseases follow **lognormal distributions**. However, for diseases with long incubation periods, account must be taken of right-truncation: infected individuals who have not yet developed symptoms are not observed (*see* **Truncated Survival Times**). If the incidence of infection varies over time, then ignoring the truncation of the data would produce biased estimates.

Several methods have been proposed for estimating the incubation period in this context. One simple method applicable to grouped data is by linear modeling. Let r_{ij} denote the number of observed cases infected in time period i with incubation period j, where $i = 1, \ldots, k$ and $j = 0, \ldots, k-1$ range over discrete time intervals with $i + j \leq k$, where k denotes the most recent time interval on which data are available. The incubation period distribution is estimated conditionally on the maximum observed interval, $k - 1$. Taking r_{ij} as Poisson with mean μ_{ij}, define the **loglinear model**

$$\log(\mu_{ij}) = h(i) + g(j)$$

for some parametric or nonparametric incidence function $\exp[h(i)]$ and incubation period distribution $\exp[g(j)]$ satisfying $\sum \exp[g(j)] = 1$. This method is described in Zeger et al. [120] and provides a simple way of jointly estimating the incidence function and the (conditional) incubation period distribution.

When data on individuals are available, an alternate method is as follows. Suppose that onsets are observed up to time t_0, resulting in n observations (t_i, s_i), where t_i is the date of onset of symptoms and s_i is the interval between infection and onset of symptoms, that is, the observed incubation period. Letting $\alpha(t)$ denote the incidence of infection at time t and $f(s)$ the density function of the incubation period,

the log likelihood may be written

$$\sum_{i=1}^{n} \ln[\alpha(t_i)] + \sum_{i=1}^{n} \ln[f(s_i)] - \int_{-\infty}^{t_0} \alpha(u) F(t_0 - u)\, \mathrm{d}u, \qquad (7)$$

where $F(s)$ is the distribution function of the incubation period. Suitable parameterizations of $\alpha(t)$ and $f(s)$ may be inserted into (7) and estimates obtained by maximum likelihood. A slightly more general version of this approach was used by Medley et al. [89] to estimate the incubation period of AIDS.

A nonparametric approach has been suggested by Lagakos et al. [78]. In this approach the analysis is undertaken in "reverse time". Right-truncated observations in forward time become left-truncated in reverse time, and can be handled using **nonparametric methods** developed for left-truncated data. Suppose that n observations (t_i, s_i) are made over the time interval $[0, t_0]$, where t_i and s_i denote the same quantities as above. The distribution of the incubation period is estimated conditionally on the maximum observable interval, t_0. Thus, the distribution estimated is

$$F^*(s|t_0) = \frac{F(s)}{F(t_0)}, \quad s \le t_0.$$

Let $v_1, \ldots, v_k$ denote the distinct values of the s_i and define

$$n_j = \sum_{i=1}^{n} 1(s_i = v_j),$$

$$N_j = \sum_{i=1}^{n} 1(s_i \le v_j \le t_0 - t_i),$$

where $1(\cdot)$ is the indicator function. Then in reverse time measured from t_0 to $0, n_j$ is the number of events occurring at reverse time $t_0 - s_j$, and N_j is the number at risk. The nonparametric maximum likelihood estimate of $F^*(s|t_0)$ is then

$$\hat{F}^*(s) = \prod_{v_j \ge s} \left(1 - \frac{n_j}{N_j}\right)$$

for $0 \le s \le v_k$, and 1 for $v_k < s \le t_0$.

All three methods may be used in other contexts, such as estimating the distribution of delays between the onset of disease and reporting or ascertainment.

Estimation of the Latent and Infectious Periods

For some diseases it is possible to identify the end of the infectious period. Thus, for instance, for measles infectiousness is minimal 2 days after the appearance of the rash. For such diseases it is possible to estimate the latency and incubation period from home contact studies, provided the chains of infection within the household are known. The method is described for households of two susceptibles.

For simplicity, assume that the infectious period is of fixed length, μ, and let Z denote the latency period, with probability density function (pdf) $g(z)$. For each case the end of the infectious period is observed, and occurs at time Y. The beginning of the infectious period for each individual is thus $Y - \mu$. Assume also that infectiousness is constant over the infectious period, so that infectious contacts occur in a Poisson process with constant rate β.

The contribution to the likelihood from a household with two co-primary cases infected at the same time with $Y_1 = y_1$, $Y_2 = y_2$, and $y_1 < y_2$, is

$$\int_0^{\infty} g(z) g(z + y_2 - y_1)\, \mathrm{d}z.$$

In the same notation, the contribution to the likelihood of a household in which the primary case infects the remaining susceptible is

$$\beta \int_0^{\mu} g(y_2 - y_1 - z) \exp(-\beta z)\, \mathrm{d}z,$$

while the contribution of a household with a single case with $Y_1 = y_1$ is

$$\exp(-\beta\mu).$$

Letting n_1, n_{11}, and n_2 denote the numbers of households of size two with chains $\{1\}$, $\{1, 1\}$, and $\{2\}$, respectively, the log likelihood is then

$$-n_1\beta\mu + n_{11}\log(\beta) + n_{11}\log\left[\int_0^{\mu} g(y_2 - y_1 - z)\exp(-\beta z)\,\mathrm{d}z\right] + n_2 \log\left[\int_0^{\infty} g(z) g(z + y_2 - y_1)\,\mathrm{d}z\right].$$

Specification of a suitable parametric form $g(\cdot|\gamma)$ then enables the log likelihood to be maximized with respect to β, μ, and γ. This approach can be extended

to households with more than two susceptibles. The assumption of a constant infective period can also be relaxed. These and other extensions are discussed in Bailey [5] and Becker [15]. Becker [15] gives a detailed application of this method to measles data. Likelihood-based methods often rely on strong assumptions. Such assumptions can sometimes be weakened by making use of MCMC methods. O'Neill et al. [94] estimate the latent and infectious periods of measles using such methods.

Severity and Complications of Infectious Diseases

Infection and disease are not synonymous. Some infections are asymptomatic, and most produce a range of symptoms which can vary in severity. The clinical severity of disease may be directly quantified using clinical information. One important such measure is the **case fatality** rate, which is the proportion dying as a result of infection. Alternately, severity may be measured by proxy variables, such as the proportion of cases admitted to hospital, or socioeconomic variables, such as days off work. In most cases disease severity is age-dependent: for instance, for mumps and hepatitis A, severity of disease increases with age. Since the introduction of vaccination increases the average age at infection, vaccination programs can perversely produce an increase in the morbidity attributable to infection [3, 76].

The statistical methods commonly used to investigate risk factors for clinical disease typically involve **survival analysis** and **regression**. One perhaps distinctive application is to the estimation of the mortality attributable to influenza. For example, Serfling [105] applied regression techniques to model the seasonal (*see* **Seasonal Time Series**) and secular trends (*see* **Time Series**) in mortality in selected cities in the US. In influenza epidemic years, large positive residuals are observed, from which an estimate of the **excess mortality** attributable to influenza may be derived.

Some infections, which are otherwise relatively benign, can have devastating consequences on the unborn child. For instance, rubella, toxoplasma, and cytomegalovirus infection in pregnancy can result in congenital abnormalities. The estimation of the number of infections in pregnancy is therefore of critical importance in assessing the value of **screening** and other prevention policies (*see* **Preventive Medicine**). Letting $\lambda(x)$ denote the force of infection at age x, $\eta(x)$ the number of births to women of age x, and τ the duration of pregnancy, the expected number of women infected in pregnancy is

$$\int_0^\infty \eta(z) \int_{z-\tau}^{z} \left[\lambda(x) \exp\left(-\int_0^x \lambda(u)\, du \right) dx \right] dz.$$

The force of infection may be estimated using the methods described above, while information on $\eta(z)$ may be derived from **vital statistics**. This and other methods are discussed by Ades [1].

Some complications have long induction periods. When the incidence of the originating infections varies over time, the estimation of the risk of complications following infection must take account of truncation effects. Similarly, transient effects due to changing incidence of infection may distort the induction period distribution. To see this let $\alpha(t)$ denote the incidence of originating infections, and $f(s)$ denote the pdf of the induction period distribution. Then the observed induction period distribution at time t is

$$f_t^*(s) = \frac{\alpha(t-s) f(s)}{\int_0^\infty \alpha(t-u) f(u)\, du}.$$

Thus, if the incidence is declining over time, the observed distribution of induction times is biased towards longer intervals. These effects are analyzed by Cox & Medley [25] in the context of AIDS. Farrington [39] discusses an application to subacute sclerosing panencephalitis (SSPE) after infection by wild measles virus.

Clustering of Infectious Diseases

One of the distinguishing features of many infectious diseases is their tendency to arise in clusters in space and time (*see* **Clustering**). While *cluster analysis* is a preoccupation common to most areas of epidemiology, in the case of infectious diseases it stems directly from the transmissibility of disease, rather than the influence of shared risk factors. The clearest example of this is the epidemicity of endemic infections, discussed above. Spatial effects can also be important in disease transmission, and may be estimated using suitable statistical methods [79]. This section describes two further areas in which statistical methods have been developed specifically for detecting clusters of infectious diseases.

Detection of an Infectious Etiology

For some diseases, such as leukemia (*see* **Leukemia Clusters**), Hodgkin's disease, and multiple sclerosis, an infectious etiology has been suggested, but still remains unproven. Much effort has been devoted to detecting clustering of cases which might support such a hypothesis. From a statistical point of view, this requires the formulation of a suitable **null hypothesis** reflecting the distribution of cases expected if there was no infectious agent involved, and testing for departures from this hypothesis in a direction suggestive of infection.

There is a large literature on clustering of health events, much of which can be applied to the detection of infectiousness; see, for instance, Mantel [83]. This discussion is limited to some of the methods developed specifically for this purpose.

An early procedure was described by Mathen & Chakraborty [84], who considered the distribution of disease within a community of households. Regarding the total number of cases as fixed, they used as test statistic the total number of households containing at least one case. This idea was developed further by Walter [117] to take account of the actual numbers of cases within each household. Consider a population of n individuals in s households. The ith household comprises n_i individuals, of whom r_i become infected. Conditioning on the total $r = \sum r_i$ cases, the null distribution of r_i is **hypergeometric**. This corresponds to a null hypothesis of no clustering within households. The test statistic, T, is the number of distinct pairs of individuals, both of whom are infected, and both of whom are from the same household:

$$T = \frac{1}{2}\sum_{i=1}^{s} r_i(r_i - 1).$$

Walter [117] gives exact expressions for the null expectation and variance of T. Asymptotically, as $n \to \infty$ and $r/n \to p$, the proportion of the population infected, the null mean and variance of T tend to

$$\mathrm{E}(T) \to np^2(\nu_2 - 1)/2,$$

$$\mathrm{var}(T) \to np^2(1-p)[2p\nu_3 + (\nu_2 - 1)(1-p) - 2p\nu_2^2]/2.$$

where $\nu_2 = s\mathrm{E}(m_i^2)/n$ and $\nu_3 = s\mathrm{E}(m_i^3)/n$. The method may also be extended to the situation where only data on households with one or more infected individuals are collected. Various modifications have been proposed to this test statistic, by Smith & Pike [108] and Fraser [56].

Methods based on the distribution of cases within households are likely to lack **power** when applied to rare diseases. In addition, they do not use information on the times at which cases arise. These shortcomings are addressed by methods to detect space–time interactions. Many methods for detecting space–time clustering have been developed, the first being that of Knox [75]. Suppose that n cases of disease are identified, together with their locations and times of onset. Knox's test statistic is the number of distinct pairs of cases which lived within a distance d and had onsets within a time period t of each other, for fixed values of d and t. Barton & David [11] expressed Knox's statistic in graph-theoretic terms, proximity in space and time being represented by adjacency matrices, and derived the null expectation and variance of Knox's statistic.

Several enhancements of Knox's statistic have been proposed. In its original form it is applicable only to infections with short latent periods. To extend it to diseases with long latent periods, Pike & Smith [99] included information on the infectious and susceptible periods of each case. Evidence of contagion is given by any case being in the "right" place at the "right" time to have caught the disease from some other case. The test statistic measures the total effective contact between distinct pairs of cases, larger values providing evidence of infectiousness.

Pike & Smith [100] extend this approach further by including a control group, and apply the method to Hodgkin's disease. Given n cases of disease, let y_{ij} denote the presence ($y_{ij} = 1$) or absence ($y_{ij} = 0$) of an effective contact from case i to case j, that is, which may have resulted in case i transmitting the disease to case j, for $i, j = 1, \ldots, n$ and $i \neq j$. The test statistic is defined as the total contact between the n cases:

$$T = \sum_i \sum_{j \neq i} y_{ij}.$$

For each case a matched control is selected. The null distribution of T is obtained by calculating the total contact for each of the 2^n random selections of one individual from each of the n case–control pairs. This is most readily derived by **Monte Carlo** simulation. Pike & Smith [100] derive the exact null values of

$E(T)$ and $\text{var}(T)$, and propose a variety of related test statistics for the total numbers of linked patients.

The Detection of Outbreaks

A characteristic shared by many infectious diseases, at least those with short incubation periods, is the rapidity with which they evolve. It follows that if effective control measures are to be introduced, then outbreaks of infectious diseases must be detected in a timely fashion. In this context the emphasis is on the prospective detection of temporal clustering of disease, that is, as data accumulates, rather than the more usual retrospective identification of temporal clusters for epidemiologic analysis. Prospective outbreak detection is therefore necessarily based on incomplete data which are usually subject to delays in reporting, and is further complicated by fluctuations in the historical data series due to seasonal cycles, secular trends, and past outbreaks (*see* **Surveillance of Diseases**).

Several statistical approaches to prospective outbreak detection have been suggested. Cumulative sum statistics have been used to detect the onset of influenza epidemics [115]. **Time series** methods have been suggested for the detection of outbreaks of Salmonella [119]. Nobre & Stroup [93] describe a different approach using exponential smoothing of the time series (*see* **Nonparametric Regression**) to identify the points at which the first derivative of the series departs significantly from zero.

The methods described above involve organism-specific modeling and hence are best suited to monitoring small numbers of data series. For the purposes of routine monitoring of large databases of infectious disease reports, however, **robust** methods are required applicable to a wide variety of organisms. Stroup et al. [114] describe a simple method for routinely detecting aberrations in reports of notifiable diseases in the US. Their method is to compare the current month's reports with the average of those received in comparable baseline periods over previous years. The current month's report is declared aberrant if it lies outside the limits $\hat{\mu} \pm 2\hat{\sigma}$, where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation, respectively, of the baseline values.

This approach corrects for seasonal variation, though not for past outbreaks. Also, the method is applicable only for organisms with substantial monthly counts. An algorithm based on **Poisson regression** modeling, applicable to rare as well as frequent organisms, and incorporating an adjustment for past outbreaks, has been described by Farrington et al. [46]. A two-thirds **power transformation** is applied to preserve a roughly constant **false positive** probability for different organism frequencies.

Increasing the availability of good data on geographical locations of the incident cases also make it possible to envisage incorporating a spatial element to the detection process [77, 102].

Vaccination

One of the distinguishing features of infectious disease epidemiology is the ability to prevent disease by vaccination. Vaccination programs are generally acknowledged as among the most effective and cheapest public health measures available. It follows that the statistical issues raised by vaccination are central to the statistics of communicable diseases.

Vaccine Trials

The **clinical trial** is the method of choice for the evaluation of vaccines. (*see* **Vaccine Studies**) Since the 1940s, when the method was first applied systematically to the evaluation of vaccines, a vast body of experience and methodology has developed. This section briefly reviews some of those aspects specific to vaccine trials. More detailed discussions may be found in Smith & Morrow [107] and Farrington & Miller [47].

Vaccine trials broadly fit within the **Phase I**, **Phase II**, and Phase III sequence of trial methodology. Most Phase I and II trials may be regarded as preliminary investigations, laying the groundwork for Phase III protective efficacy trials. However, in some cases Phase II trials take on a different purpose in that they are used to underpin major decisions about vaccination policy. This is the case, for instance, when a vaccine has already undergone a successful evaluation in a Phase III trial, possibly in another country. Additional data are required to support the introduction of the vaccine in a different population, possibly under a different immunization schedule from that used in the Phase III trial. Such Phase II trials may thus be described as confirmatory rather than exploratory.

The primary purpose of a Phase III trial is to assess the protective efficacy of the vaccine in the **target**

population. Phase III trials may be large, especially when the disease concerned is rare. The 1954 field trial of **Salk vaccine** for polio involved 1.8 million children in the US, over 400 000 of whom were randomly assigned to vaccine or placebo and a much larger number enrolled in an open study [55]. Trials on such a gigantic scale are rare, but many nevertheless require sample sizes running into thousands. Vaccine efficacy trials are thus considerable logistic undertakings, and require clear procedures for handling vaccines, including their labeling, storage and transport, and monitoring their condition, for instance using temperature-sensitive devices.

The first stage in designing an efficacy trial is to define the outcome of primary interest (*see* **Outcome Measures in Clinical Trials**). Many vaccines, such as those against pertussis [52] or rotavirus [116], alter the clinical course of the disease. Thus, a vaccine that is effective in preventing clinical disease may have a considerably lower efficacy against milder or asymptomatic infection. Clarity about the purpose of the trial and the intended use of the vaccine is, therefore, essential from the start if confusion resulting from contradictory interpretations of the trial results is to be avoided. In the special circumstances of measles vaccines in developing countries, it has been argued that it is more appropriate to use total mortality from any cause as the primary outcome, rather than measles morbidity, since the effect of vaccination may have nonspecific immunological consequences [64].

These considerations will guide the choice of primary case definitions, which should be chosen with due regard to potential **biases**. Clinical case definitions may lack **specificity**, unless corroborated by laboratory evidence, and hence will bias efficacy towards zero (*see* **Bias Toward the Null**), since the vaccine cannot be expected to protect against infections other than that for which it was developed. However, the use of laboratory methods for confirmation should be validated, since the **sensitivity** of the method may vary between vaccinated and unvaccinated cases. For example, there is some evidence that bacterial isolation rates of *B. pertussis* are lower in vaccinated than unvaccinated cases [113] which would result in an artificially high estimate of vaccine efficacy. As in other studies of vaccine efficacy, special care must be taken to ensure that individuals allocated to the different vaccine groups have the same probability of exposure to infection, for instance by using block randomization within suitably defined units of space and time (*see* **Randomized Treatment Assignment**).

Estimation of Vaccine Efficacy

Vaccine efficacy is defined as the percentage reduction in the attack rate attributable to the vaccine. This may be written:

$$VE = \left(\frac{p_u - p_v}{p_u}\right) \times 100, \tag{8}$$

where p_u and p_v denote, respectively, the risk of infection in unvaccinated and vaccinated individuals over a specified observation period $[0, t]$ (*see* **Attributable Risk**). For notational simplicity, we omit the percentage multiplier and write

$$VE = 1 - \rho,$$

where ρ is the **relative risk** of infection, p_v/p_u. Alternately, one can define vaccine efficacy in terms of the **relative hazard** of infection:

$$VE = 1 - \frac{\lambda_v}{\lambda_u}, \tag{9}$$

where

$$p_u = 1 - \exp(-\lambda_u t), \qquad p_v = 1 - \exp(-\lambda_v t).$$

The two measures do not differ appreciably when $\lambda_u t$ is small.

These measures are used both to quantify vaccine efficacy in clinical trials and to evaluate vaccine effectiveness in the field. The term "effectiveness", rather than "efficacy", is often used to underline the distinction between estimates obtained in controlled experiments and those achieved under field conditions. The latter may be influenced by vaccine storage, variability of vaccination schedules, herd immunity, and other factors not directly attributable to the vaccine's direct biological effect. Henceforth, for reasons of economy, we use the term "efficacy" to cover both biological efficacy and field effectiveness (*see* **Pharmacoepidemiology, Adverse and Beneficial Effects**).

Vaccine efficacy may be estimated directly in a **cohort study** involving vaccinated and unvaccinated individuals. This was the original approach of Greenwood & Yule [60]. Let n_v and n_u denote, respectively,

the numbers of vaccinated and unvaccinated individuals. Suppose that r_v vaccinated and r_u unvaccinated cases arise during a specified observation period. The vaccine efficacy may then be estimated as

$$\widehat{VE} = 1 - \frac{r_v/n_v}{r_u/n_u}.$$

Confidence limits may be derived from those for the estimated relative risk $\hat{\rho} = (r_u/n_u)/(r_v/n_v)$.

Covariate effects may be estimated by modeling the number of cases as **binomial**, using a **generalized linear model** with logarithmic link. Thus, for instance, if each group is stratified according to covariates $\mathbf{x}_i, i = 1, \ldots, k$, then the main effects model for vaccine and covariate effects has the structure:

$$\ln(p_{ui}) = \alpha + \beta^T X_i,$$
$$\ln(p_{vi}) = \alpha - \gamma + \beta^T X_i,$$

and the corrected estimate of vaccine efficacy, allowing for covariate effects, is $\widehat{VE} = 1 - e^{-\hat{\gamma}}$. The effect of the covariates on vaccine efficacy may also be investigated by fitting the relevant **interaction** terms.

A second method for estimating vaccine efficacy is by means of a **case–control study** [106]. A sample of cases is identified along with suitable controls, typically from the same age group and locality. Vaccination histories are obtained for both cases and controls. The **odds ratio** of vaccination in cases and controls is equal to the odds ratio of disease in vaccinated and unvaccinated children. Provided attack rates are low, this approximates the relative risk, so that

$$VE \doteq 1 - \frac{p_v/(1-p_v)}{p_u/(1-p_u)}.$$

The analysis uses standard case control methodology, regression models being fitted with conditional or unconditional **logistic regression** techniques.

A third method for estimating vaccine efficacy, called the *screening method*, is commonly used for routine monitoring purposes, or in circumstances in which only data on cases are available. Suppose that all or a random sample of cases of disease arising over a given period in a defined population are available. Let θ denote the proportion of cases vaccinated, and suppose that the proportion of the population vaccinated, π, is known. The vaccine efficacy is then

$$VE = 1 - \left(\frac{\theta}{1-\theta}\right)\left(\frac{1-\pi}{\pi}\right). \qquad (10)$$

In the screening method, the vaccination coverage, π, is fixed, while θ is estimated.

Stratified analyses (*see* **Stratification**) using the screening method are possible provided suitably stratified vaccine coverage statistics are available. Suppose that cases are observed and classified into m strata, with n_i cases in stratum $i, i = 1, \ldots, m$. Suppose that of the n_i cases in stratum i, r_i, are vaccinated. For each stratum $i = 1, \ldots, m$, let θ_i denote the probability that a case is vaccinated, π_i the population vaccine coverage, and ρ_i the relative risk of disease. Suppose also that k covariates on each stratum are also available, the value of the jth covariate in the ith stratum being denoted by $x_{ij}, j = 1, \ldots, k$. Then, given a linear model for the relative risk,

$$\ln(\rho_i) = \alpha + \sum_{j=1}^{k} \beta_j X_{ij},$$

(10) may be re-expressed for each stratum as

$$\text{logit}(\theta_i) = \text{logit}(\pi_i) + \alpha + \sum_{j=1}^{k} \beta_j X_{ij}. \qquad (11)$$

Assuming that cases of disease arise in a **Poisson process**, r_i is **binomial** (n_i, θ_i). Thus, the model specified by (11) may be fitted as a **generalized linear model** with binomial error and logistic link, with offsets $\text{logit}(\pi_i)$. Clearly, this method depends for its validity on the availability of accurate data on vaccine coverage. Given such data, the method allows vaccine efficacies to be calculated very simply from case reports or notifications. Applications of the method to measles and pertussis vaccine efficacy are given in [42].

Alternatively, if the population vaccine coverage is not known, then it may be estimated from a sample. This approach leads to **case–cohort** designs [90].

In all the methods described above, care must be taken to adjust for potential **confounders**. For instance, both vaccine coverage and incidence of infection may vary with age and location, which may therefore confound vaccine efficacy. In field investigations it is essential to document vaccination

histories, and to ascertain cases independently of vaccination status. These and other methodological issues are discussed in detail in Clarkson & Fine [23], Fine & Clarkson [52], and Orenstein et al. [97].

In estimating vaccine efficacy it is also important to ensure that vaccinated and unvaccinated individuals have the same probability of exposure. In an attempt to control for exposure, vaccine efficacy is sometimes estimated from attack rates in household contacts of infected cases. Some of the methodological issues surrounding such studies are discussed in Fine et al. [53]. Concern about this issue has led to a further measure of vaccine efficacy being proposed, which controls for exposure:

$$VE = 1 - \frac{\theta_v}{\theta_u}, \tag{12}$$

where θ_u and θ_v are, respectively, the transmission probabilities of infection to unvaccinated and vaccinated individuals given contact with a single infective [63].

The various definitions of efficacy given above all represent direct efficacy, that is, they measure the individual benefit gained from vaccination, in a vaccinated population. In addition, the vaccine may confer an indirect benefit through herd immunity, by reducing the circulation of the infection, and hence indirectly protecting unvaccinated individuals. Direct and indirect measures of vaccine efficacy are discussed in [65].

In addition to reducing the susceptibility of those vaccinated, vaccines may also reduce the infectiousness of individuals infected. The estimation of the effect of vaccination on infectiousness is not straightforward, but can be undertaken in the context of household situations [82], or in several populations with different levels of vaccine coverage [27].

Vaccine Models

Two contrasting models of vaccine mechanisms have been suggested [109]. In the first, the so-called "all-or-nothing" model, the vaccine imparts total, long-lasting protection to a proportion of vaccinees, and gives no protection to the remainder. The proportion protected may be estimated unbiasedly as one minus the relative risk of disease, that is, using (8) above. It has been suggested that this model of vaccine protection may apply to live viral vaccines such as measles vaccine. In the second model, sometimes called the "leaky" vaccine model, vaccination does not impart complete protection on any individual, but reduces the hazard of infection by a constant factor. This relative hazard corresponds to $1 - VE$, where the vaccine efficacy is estimated using (9) above. This mechanism might be appropriate for bacterial vaccines, such as whole cell pertussis vaccine. In a randomly mixing population, the vaccine efficacy measure, (12), based on transmission probabilities, has been shown to equal that based on attack rates, (8), for all-or-nothing vaccines, and (9) for leaky vaccines [63].

In practice it is extremely difficult to distinguish between the two vaccine mechanisms, since changes in one or other measure of vaccine efficacy over time can be attributed either to the vaccine mechanism or to waning vaccine efficacy. This confounding of vaccine mechanism and changes in efficacy over time poses particular problems for the evaluation of age-specific efficacy of the vaccine, since a decline in age-specific efficacy may be attributed to the vaccine mechanism, to bias in identifying susceptible individuals, or to waning efficacy of the vaccine. The problem of estimating and interpreting age-specific vaccine efficacy measures is discussed in [41] and [71].

Vaccine Safety Evaluation

Vaccine safety is clearly a critical issue, often generating considerable public interest, particularly in the case of vaccines administered to children on a large scale for preventive purposes. Clinical trials are usually too small to demonstrate the safety of vaccines with respect to rare, but potentially serious, adverse events. Instead, vaccine safety is monitored by surveillance and epidemiologic investigations after the vaccine is in widespread use. Such studies present considerable statistical challenges. For instance, for vaccines administered as part of a routine immunization program, unvaccinated individuals are a selected group which cannot be assumed representative of the population as a whole, and hence should not be included in the control group once high vaccine coverage is achieved [27]. Instead, for acute adverse reactions, such as febrile convulsions following measles or pertussis vaccine, or aseptic meningitis following mumps vaccine, the focus is to detect a clustering of events in the period following vaccination. This is achieved by estimating the relative

rate at which events occur in a specified period following vaccination, compared with the background rate in control time periods, correcting for age effects which may confound the relationship between vaccination and adverse events. Commonly used methods include cohort and case–control studies. In the cohort approach, the analysis is conditional on vaccination times, the time of observation for each individual being divided into "at risk" and "control" periods. In the case–control approach, cases are matched to controls by date of birth and other relevant variables. The date of the reaction in the case is taken as the index date, and exposure in both case and control is defined as vaccination in a specified period prior to the index date. The methodologic issues associated with such studies are discussed in Ray et al. [101] and Fine & Chen [51].

A third method specifically designed for the analysis of acute, transient vaccine reactions combines the economy of the case–control method and the power of the cohort method. This is the case–series method [43]. The method is derived from a cohort model, by conditioning on the total number of events experienced by each individual. Individuals who do not experience any reactions thus make no contribution to the likelihood, and hence only data on cases are required. Thus, suppose that n cases are observed over a defined observation period. For the ith individual, let e_{ijk} denote the time spent in age group j and risk group k. The risk groups are defined in relation to vaccination: for instance, for febrile convulsions after pertussis vaccine, one might use the intervals 0–3 days, 4–7 days, 8–14 days after vaccination as distinct risk groups, all other times being included in the reference control group. The number of adverse events experienced by individual i in age group j and risk group k, r_{ijk}, is assumed Poisson with mean μ_{ijk}. A simple cohort model may be written

$$\ln(\mu_{ijk}) = \ln(e_{ijk}) + \alpha^{\mathrm{T}} x_i + \beta_j + \gamma_k,$$

where x_i are fixed covariates. Conditioning on the total number of events observed for individual i results in the product multinomial likelihood:

$$L = \prod_i \prod_{j,k} \left(\frac{e_{ijk} \exp(\beta_j + \gamma_k)}{\sum_{r,s} e_{irs} \exp(\beta_r + \gamma_s)} \right)^{r_{ijk}},$$

to which individuals with $r_{i..} = 0$ contribute 1, and hence may be ignored. This greatly simplifies the study of adverse events, since only a sample of individuals experiencing an event over a given period is required. This model assumes that events are potentially recurrent. However, for rare nonrecurrent events, such as sudden infant death syndrome (SIDS) or encephalopathy, the method can be applied with little bias. Note that with this approach the effect of fixed covariates on the incidence of adverse events cannot be estimated. However, their effect on the relative incidence of adverse events after vaccine can be estimated by including suitable interaction terms in the model. This method has been shown to be as powerful as the cohort method when vaccine coverage is high [48].

Other Methods

The collection of infectious disease data frequently relies on complex laboratory techniques, such as serological and other assays, electron microscopy, electrophoresis, typing, and other identification methods. Many of these involve statistical techniques, such as statistical taxonomy. A thorough discussion of laboratory methods is beyond the scope of this article. However, the laboratory methods used often also have a direct bearing on the analysis and interpretation of epidemiologic data. This section aims to illustrate this point using two examples.

Prevalence Estimation by Group Testing

In some situations it is necessary, for instance for reasons of economy, to pool samples of material for analysis, and test the combined pool rather than the individual samples. The statistical properties of this approach were originally investigated by Dorfman [34] to reduce the number of tests required to identify cases of syphilis. Today, group testing is often used to test for HIV in large population surveys [21]. In some cases it is possible, and indeed, for diagnostic tests, necessary, to retest the individual components of a positive pool. However, the approach may also be used directly to estimate population prevalence [111]. In this setting the retesting of individual samples is not necessary, and in some cases not possible, as for instance in the study of vertical transmission of yellow fever virus by mosquitoes discussed by Walter et al. [118].

Suppose that n pools have been formed, the ith pool including m_i samples from individuals with a common covariate vector $\mathbf{x}_i$, $i = 1, \ldots, n$. Let $r_i = 1$ if the ith pool is positive, 0 if it is negative. Let π_i be the probability that the ith pool tests positive, and θ_i be the probability that an individual in the ith pool is positive. Thus θ_i is the prevalence in the subpopulation with characteristics $\mathbf{x}_i$. Provided that the individuals within each pool are independent, the pool and population prevalences are related by

$$\pi_i = 1 - (1 - \theta_i)^{m_i}.$$

Thus, if the population prevalences satisfy the linear model

$$\log(-\log(1 - \theta_i)) = \beta^{\mathrm{T}} x_i,$$

with regression parameters β, then the pool prevalences satisfy

$$\log(\pi_i) = \log(m_i) + \beta^{\mathrm{T}} x_i.$$

Thus, the population prevalences and the regression parameters may be estimated by fitting a generalized model with dependent variable r_i, binomial error, complementary log–log link function, and fixed offset $\log(m_i)$. This method is further discussed in [40], with an application to estimating the prevalence of Salmonella contamination in eggs.

The concept of group testing may also be applied to estimating the most probable number (MPN) (*see* **Serial Dilution Assay**) of coliform organisms in water samples, using the multiple fermentation tube method of McCrady [85]. A sample of water from a given source is taken and subdivided, with or without dilution, into subvolumes. These are then incubated in separate tubes, which are examined for evidence of growth, indicating that at least one organism was present in the subvolume. Let n_i denote the number of tubes containing a volume v_i of the original water sample, and r_i the number of positive tubes among these n_i. If coliforms are homogeneously distributed with density λ per unit volume, then r_i is binomial (n_i, π_i), where

$$\log(-\log(1 - \pi_i)) = \log(v_i) + \log(\lambda).$$

Hence the density, λ, and the MPN, λV, where V denotes the original volume of water, may be estimated using a binomial model with a complementary log–log link function and offset $\log(v_i)$.

Mixture Models for Quantitative Assay Data

Laboratory assays are widely used for diagnostic purposes on samples from individual patients. They may also be used in serological surveys to determine immunity levels in a population, and hence to monitor or design vaccination programs. Many commonly used assays, such as enzyme-linked immunosorbent assays (ELISA), give quantitative results. When used for diagnostic purposes, it is necessary to define one or more cutoff values to classify the test results as negative, positive, or equivocal. These cutoff values also determine the diagnostic sensitivity and specificity of the assay. In serological surveys, on the other hand, the aim is not to classify individual test results, but to estimate the age-specific prevalence in the population.

The determination of cutoff values is problematic when there is no objective criterion by which to classify samples as "true" positives or negatives. For common infections, however, this may be achieved by fitting mixture models to data obtained from population-based serological surveys. Furthermore, when used for determining population prevalence, cutoff values are not required using these methods.

Each age group i is assumed to be a mixture of positives (immunes) and negatives (nonimmunes) in the proportions π_i and $1 - \pi_i$, respectively, where π_i is the proportion positive, which may be constrained to increase with age. Assay results in age group i are distributed with density:

$$f(x|\pi_i, \alpha_i, \beta_i) = (1 - \pi_i) g(x|\alpha_i) + \pi_i h(x|\beta_i),$$

where $g(x|\alpha_i)$ and $h(x|\beta_i)$ are the densities of assay results from negative and positive individuals, respectively. For instance, for an ELISA assay, the random variable X may be taken to be the logarithm of the optical density reading, and $g(\cdot)$ and $h(\cdot)$ may be normal with age-dependent means and constant variance. Let $x_0 = -\infty < x_1 < \cdots < x_k = \infty$ subdivide the range of X and suppose that there are n_{ij} values from age group i in interval $[x_{j-1}, x_j)$. The parameters π_i, α_i, and β_i may be estimated by maximizing the product multinomial log likelihood:

$$\sum_i \sum_j n_{ij} \log \left(\int_{x_{j-1}}^{x_j} f(x|\pi_i, \alpha_i, \beta_i)\, \mathrm{d}x \right).$$

Note that the parameters π_i are estimated without the need to specify cutoff values. Appropriate cutoff

values may be derived by examining the **receiver operating characteristic (ROC) curve** using the estimated specificities and sensitivities corresponding to different cutoff values c:

$$\text{spec}_i(c) = \int_{-\infty}^{c} g(x|\hat{\alpha}_i)\,\mathrm{d}x,$$

$$\text{sens}_i(c) = \int_{c}^{\infty} h(x|\hat{\beta}_i)\,\mathrm{d}x.$$

The mixture modeling approach to the determination of cutoff values is discussed in [98] and applied to estimating prevalences in serological surveys by Gay [58], both in relation to parvovirus B19 infection.

Future Challenges

The substantial statistical literature on applications to **AIDS and HIV** demonstrates the vitality of infectious disease statistics. By the beginning of the twenty-first century, smallpox remained the only infectious disease to have been eradicated. The continued toll of infectious disease throughout the world–particularly the increasing impact of **AIDS** in many countries–along with the emergence of new communicable diseases such as **vCJD**, and of resistant forms of diseases like tuberculosis, combine with new concerns over bioterrorism to unfortunately guarantee the continued relevance of infectious disease statistics.

It is perhaps surprising that our understanding of the mechanism of transmission of infectious diseases has not changed fundamentally since the work of the pioneers at the start of the twentieth century. Further work is required on understanding the epidemicity and seasonality of infectious diseases. Work is also needed on the geographic spread of infectious diseases. Though much mathematical modeling and some statistical work has been done in this area (see, for instance, Cliff & Haggett [24]), little data have so far been available. The development of combination vaccines will also bring new challenges, requiring the assessment of new vaccines against old in **equivalence trials** with **multiple comparisons** and **surrogate endpoints**. The potential risks of vaccines are likely to come under ever closer scrutiny, raising the difficult statistical issue of evaluating vaccine safety with respect to adverse events with long induction periods, in highly vaccinated populations.

References

[1] Ades, A.E. (1992). Methods for estimating the incidence of primary infection in pregnancy: a reappraisal of toxoplasmosis and cytomegalovirus data, *Epidemiology and Infection* **108**, 367–375.

[2] Ades, A.E. & Nokes, D.J. (1993). Modelling age- and time-specific incidence from seroprevalence: toxoplasmosis, *American Journal of Epidemiology* **137**, 1022–1034.

[3] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[4] Anderson, R.M., Grenfell, B.T. & May, R.M. (1984). Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis, *Journal of Hygiene, Cambridge* **93**, 587–608.

[5] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd Ed. Griffin, London.

[6] Ball, F. & Lyne, O. (2001). Stochastic multi-type SIR epidemics among a population partitioned into households, *Advances in Applied Probability* **33**, 99–123.

[7] Ball, F., Mollison, D. & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing, *The Annals of Applied Probability* **7**, 46–89.

[8] Bartlett, M.S. (1956). Deterministic and stochastic models for recurrent epidemics, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. University of California Press, Berkeley, pp. 81–109.

[9] Bartlett, M.S. (1957). Measles periodicity and community size, *Journal of the Royal Statistical Society, Series A* **120**, 48–70.

[10] Bartlett, M.S. (1960). The critical community size for measles in the United States, *Journal of the Royal Statistical Society, Series A* **123**, 37–44.

[11] Barton, D.E. & David, F.N. (1966). The random intersection of two graphs, in *Research Papers in Statistics: Festschrift for J. Neyman*, F.N. David, ed. Wiley, New York, pp. 445–459.

[12] Becker, N. (1974). On parametric estimation for mortal branching processes, *Biometrika* **61**, 393–399.

[13] Becker, N. (1976). Estimation for an epidemic model, *Biometrics* **32**, 769–777.

[14] Becker, N. (1977). Estimation for discrete time branching processes with applications to epidemics, *Biometrics* **33**, 515–522.

[15] Becker, N.J. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

[16] Becker N.G. & Britton T. (1999). Statistical studies of infectious disease incidence, *Journal of the Royal Statistical Society Series B* **61**, 287–307.

[17] Becker N.G. & Hasofer A.M. (1998). Estimating the transmission rate for a highly infectious disease, *Biometrics* **54**, 730–738.

[18] Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir, *Mémoires de Mathematiques et de Physique*, pp. 1–45. In *Histoire de l'Académie Royale des Sciences, Paris* (1760).

[19] Britton T. (2001). Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage, *Journal of the Royal statistical Society Series B* **63**, 705–715.

[20] Brookmeyer, R. & Gail, M.H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic, *Journal of the American Statistical Association* **83**, 301–308.

[21] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, Oxford.

[22] Brownlee, J. (1907). Statistical studies in immunity: the theory of an epidemic, *Proceedings of the Royal Society of Edinburgh* **26**, 484–521.

[23] Clarkson, J.A. & Fine, P.E.M. (1987). An assessment of methods for routine local monitoring of vaccine efficacy, with particular reference to measles and pertussis, *Epidemiology and Infection* **99**, 485–499.

[24] Cliff, A.D. & Haggett, P. (1982). Methods for the measurement of epidemic velocity from time series data, *International Journal of Epidemiology* **11**, 82–89.

[25] Cox, D.R. & Medley, G.F. (1989). A process of events with notification delay and the forecasting of AIDS, *Philosophical Transaction of the Royal Society of London, Series B* **925**, 135–145.

[26] Daley D.J. & Gani, J. *Epidemic Modelling: an Introduction*. Cambridge University Press, Cambridge 1999.

[27] Datta S., Halloran M.E. & Longini I.M. (1999). Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: Randomization by individual versus household, *Biometrics* **55**, 792–798.

[28] Dietz, K. (1975). Transmission and control of arbovirus diseases, in *Epidemiology*, D. Ludwig & K.L. Cooke, eds. SIAM, Philadelphia.

[29] Dietz, K. (1988). The first epidemic model: a historical note on P.D. En'ko, *Australian Journal of Statistics* **30A**, 56–65.

[30] Dietz K. (1993). The estimation of the basic reproduction number for infectious diseases, *Statistical Methods in Medical Research* **2**, 23–41.

[31] Dietz K. & Hesterbeek J.A.P. (2000). Bernoulli was ahead of modern epidemiology, *Nature* **408**, 513–514.

[32] Dietz, K. & Schenzle, D. (1985). Mathematical models for infectious disease statistics, in *A Celebration of Statistics*, A.C. Atkinson & S.E. Fienberg, eds. Springer-Verlag, New York.

[33] Dion, J.P. (1975). Estimation of the variance of a branching process, *Annals of Statistics* **3**, 1184–1187.

[34] Dorfman, R. (1943). The detection of defective members of large populations, *Annals of Mathematical Statistics* **14**, 436–440.

[35] En'ko, P.D. (1889). The epidemic course of some infectious diseases, *Vrach, St Petersburg* **10**, 1008–1010; 1039–1042; 1061–1063 (in Russian).

[36] Farr, W. (1840). *Progress of Epidemics. Second Report of the Registrar General of England and Wales.* HMSO, London, pp. 91–98.

[37] Farr, W. (1866). On the cattle plague, *Letter to the Editor of the Daily News*, February 19, London.

[38] Farrington, C.P. (1990). Modelling forces of infection for measles, mumps and rubella, *Statistics in Medicine* **9**, 953–967.

[39] Farrington, C.P. (1991). Subacute sclerosing panencephalitis in England and Wales: transient effects and risk estimates, *Statistics in Medicine* **10**, 1733–1744.

[40] Farrington, C.P. (1992). Estimating prevalence by group testing using generalized linear models, *Statistics in Medicine* **11**, 1591–1597.

[41] Farrington, C.P. (1992). The measurement and interpretation of age-specific vaccine efficacy, *International Journal of Epidemiology* **21**, 1014–1020.

[42] Farrington, C.P. (1993). Estimation of vaccine efficacy using the screening method, *International Journal of Epidemiology* **22**, 742–746.

[43] Farrington, C.P. (1995). Relative incidence estimation from case series for vaccine safety evaluation, *Biometrics* **51**, 228–235.

[44] Farrington C.P., Kanaan M.N. & Gay N.J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data, *Applied Statistics* (with Discussion), **50**, 251–283.

[45] Farrington C.P., Kanaan M.N. & Gay N.J. (2002). Branching process models for surveillance of infectious diseases controlled by mass vaccination, *Biostatistics* (in press).

[46] Farrington, C.P., Andrews, N.J., Beale, A.D. & Catchpole, M.A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease, *Journal of the Royal Statistical Society, Series A* **159**, 547–563.

[47] Farrington, P. & Miller, E. (1996). Clinical trials, in *Methods in Molecular Medicine: Vaccine Protocols*, A. Robinson, G. Farrar & C. Wiblin, eds. Humana Press, Totowa.

[48] Farrington, C.P., Nash, J. & Miller, E. (1996). Case series analysis of adverse reactions to vaccine: a comparative evaluation, *American Journal of Epidemiology* **143**, 1165–1173.

[49] Fine, P.E.M. (1975). Ross's *a priori* pathometry: a perspective, *Proceedings of the Royal Society of Medicine* **68**, 547–551.

[50] Fine, P.E.M. (1979). John Brownlee and the measurement of infectiousness: an historical study in epidemic theory, *Journal of the Royal Statistical Society, Series A* **142**, 347–362.

[51] Fine, P.E. & Chen, R.T. (1992). Confounding in studies of adverse reactions to vaccines, *American Journal of Epidemiology* **136**, 121–135.

[52] Fine, P.E.M. & Clarkson, J.A. (1987). Reflections of the efficacy of pertussis vaccines, *Reviews of Infectious Diseases* **9**, 866–883.

[53] Fine, P.E.M., Clarkson, J.A. & Miller, E. (1988). The efficacy of pertussis vaccines under conditions of household exposure, *International Journal of Epidemiology* **17**, 635–642.

[54] Finkenstädt B.F. & Grenfell B.T. (2000). Time series modelling of childhood diseases: a dynamical systems approach, *Applied Statistics* **49**, 187–205.

[55] Francis, T., Korns, R.F., Voight, T., Boisen, M., Hemphill, F.M., Napier, J.A. & Tochinsky, E. (1955). An evaluation of the 1954 poliomyelitis vaccine trials, *American Journal of Public Health, Part 2* **45**, 1–63.

[56] Fraser, D.W. (1983). Clustering of disease in population units: an exact test and its asymptotic version, *American Journal of Epidemiology* **118**, 732–739.

[57] Frost, W.H. (1976). Some conceptions of epidemics in general, *American Journal of Epidemiology* **103**, 141–151.

[58] Gay, N.J. (1996). Analysis of serological surveys using mixture models: application to a survey of parvovirus B19, *Statistics in Medicine* **15**, 1567–1573.

[59] Greenwood, M. (1931). On the statistical measure of infectiousness, *Journal of Hygiene, Cambridge* **31**, 336–351.

[60] Greenwood, M. & Yule, U.G. (1915). The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general, *Proceedings of the Royal Society of Medicine* **8**(2), 113–194.

[61] Grenfell, B.T. & Anderson, R.M. (1985). The estimation of age-related rates of infection from case notifications and serological data, *Journal of Hygiene, Cambridge* **95**, 419–436.

[62] Griffiths, D.A. (1974). A catalytic model of infection for measles, *Applied Statistics* **23**, 330–339.

[63] Haber, M., Longini, I.M. & Halloran, M.E. (1991). Measures of the effects of vaccination in a randomly mixing population, *International Journal of Epidemiology* **20**, 300–310.

[64] Hall, A.J. & Aaby, P. (1990). Tropical trials and tribulations, *International Journal of Epidemiology* **19**, 777–781.

[65] Halloran, M.E., Haber, M., Longini, I.M. & Struchiner, C.J. (1991). Direct and indirect effects in vaccine efficacy and effectiveness, *American Journal of Epidemiology* **133**, 323–331.

[66] Hamer, W.H. (1906). Epidemic disease in England: the evidence of variability and persistency of type, *Lancet* **ii**, 733–739.

[67] Harris, T.E. (1948). Branching processes, *Annals of Mathematical Statistics* **19**, 474–494.

[68] Heyde, C.C. (1974). On estimating the variance of the offspring distribution in a simple branching process, *Advances in Applied Probability* **6**, 421–433.

[69] Heyde, C.C. (1979). On assessing the potential severity of an outbreak of a rare infectious disease: a Bayesian approach, *Australian Journal of Statistics* **21**, 282–292.

[70] Johnson, A.M., Wadsworth, J., Wellings, K. & Field J. (1994). *Sexual Attitudes and Lifestyles*. Blackwell, Oxford.

[71] Kanaan M.N. & Farrington C.P. (2002). Estimation of waning vaccine efficacy, *Journal of the American Statistical Association* **97**, 389–397.

[72] Keiding, N. (1975). Estimation theory for branching processes, *Bulletin of the International Statistical Institute* **46**(4), 12–19.

[73] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective, *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[74] Kermack, W.O. & McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society, Series A* **115**, 700–721.

[75] Knox, E.G. (1964). The detection of space–time interactions, *Applied Statistics* **13**, 25–29.

[76] Knox, E.G. (1980). Strategy for rubella vaccination, *International Journal of Epidemiology* **9**, 13–23.

[77] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society, Series A* **164**, 61–72.

[78] Lagakos, S.W., Barraj, L.M. & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS, *Biometrika* **75**, 515–523.

[79] Lawson A.B. & Leimich P. (2000). Approaches to the space-time modelling of infectious disease behavior, *IMA J. Math. Appl. Med.* **17**, 1–13.

[80] Longini, I.M. & Koopman, J.S. (1982). Household and community transmission parameters from final distributions of infections in households, *Biometrics* **38**, 115–126.

[81] Longini, I.M., Koopman, J.S., Monto, A.S. & Fox, J.P. (1982). Estimating household and community transmission parameters for influenza, *American Journal of Epidemiology* **115**, 736–751.

[82] Longini I.M., Sagatelian, K., Rida W.N. & Halloran, M.E. (1998). Optimal vaccine trial design when estimating vaccine efficacy for susceptibility and infectiousness from multiple populations, *Statistics in Medicine* **17**, 1121–1136.

[83] Mantel, N. (1967). The detection of disease clustering and a generalized regression approach, *Cancer Research* **27**, 209–220.

[84] Mathen, K.K. & Chakraborty, P.N. (1950). A statistical study on multiple cases of disease in households, *Sankhyā* **10**, 387–392.

[85] McCrady, M.H. (1918). Tables for rapid interpretation of fermentation-tube results, *Public Health Journal, Toronto* **9**, 201–220.

[86] Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis, *British Medical Journal* **ii**, 769–782.

[87] Medical Research Council (1950). Clinical trials of antihistaminic drugs in the prevention and treatment of the common cold, *British Medical Journal* **ii**, 425–429.

[88] Medical Research Council (1951). The prevention of whooping cough by vaccination, *British Medical Journal* **i**, 1463–1471.

[89] Medley, G.F., Billard, L., Cox, D.R. & Anderson, R.M. (1988). The distribution of the incubation period for the acquired immunodeficiency syndrome (AIDS), *Proceedings of the Royal Society, Series B* **233**, 367–377.

[90] Moulton, L.H., Wolff, M.C., Brenneman, G. & Santosham, M. (1994). Case–cohort analysis of case-coverage studies of vaccine effectiveness, *American Journal of Epidemiology* **142**, 1000–1006.

[91] Muench, H. (1934). Derivation of rates from summation data by the catalytic curve, *Journal of the American Statistical Association* **29**, 25–38.

[92] Muench, H. (1959). *Catalytic Models in Epidemiology*. Harvard University Press, Cambridge, Mass.

[93] Nobre, F.F. & Stroup, D.F. (1994). A monitoring system to detect changes in public health surveillance data, *International Journal of Epidemiology* **23**, 408–418.

[94] O'Neill, P.D., Balding, D.J., Becker, N.G., Eerola, M. & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods, *Applied Statistics* **49**, 517–542.

[95] O'Neill, P.D. & Becker, N.G. (2001). Inference for an epidemic when susceptibility varies, *Biostatistics* **2**, 99–108.

[96] O'Neill, P.D. & Roberts, G.O. (1999). Bayesian inference for partially observed stochastic epidemics, *Journal of the Royal Statistical Society Series A* **162**, 121–129.

[97] Orenstein, W.A., Bernier, R.H. & Hinman, A.R. (1988). Assessing vaccine efficacy in the field, *Epidemiologic Reviews* **10**, 212–241.

[98] Parker, R.A., Erdman, D.D. & Anderson, L.J. (1990). Use of mixture models in determining laboratory criterion for identification of seropositive individuals: application to parvovirus B19 serology, *Journal of Virological Methods* **27**, 135–144.

[99] Pike, M.C. & Smith, P.G. (1968). Disease clustering: a generalization of Knox's approach to the detection of space–time interactions, *Biometrics* **24**, 541–556.

[100] Pike, M.C. & Smith, P.G. (1974). A case–control approach to examine diseases for evidence of contagion, including diseases with long latent periods, *Biometrics* **30**, 263–279.

[101] Ray, W.A. & Griffin, M.R. (1989). Use of Medicaid data for pharmacoepidemiology, *American Journal of Epidemiology* **129**, 837–849.

[102] Rogerson, P.A. (2001). Monitoring point patterns for the development of space-time clusters, *Journal of the Royal Statistical Society, Series A* **164**, 87–96.

[103] Ross, R. (1911). *The Prevention of Malaria*, 2nd Ed. John Murray, London.

[104] Sartwell, P.E. (1950). The distribution of incubation periods of infectious disease, *American Journal of Hygiene* **51**, 310–318.

[105] Serfling, R.E. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths, *Public Health Reports* **78**, 494–506.

[106] Smith, P.G. (1982). Retrospective assessment of the effectiveness of BCG vaccination against tuberculosis using the case–control method, *Tubercle* **63**, 23–35.

[107] Smith, P.G. & Morrow, R.H. eds. (1991). *Methods for Field Trials of Interventions Against Tropical Diseases: A Toolbox*. Oxford University Press, Oxford.

[108] Smith, P.G. & Pike, M.C. (1976). Generalization of two tests for the detection of household aggregation of disease, *Biometrics* **32**, 817–828.

[109] Smith, P.G., Rodrigues, L.C. & Fine, P.E.M. (1984). Assessment of the protective efficacy of vaccines against common diseases using case–control and cohort studies, *International Journal of Epidemiology* **13**, 87–93.

[110] Snow, J. (1855). *The Mode of Communication of Cholera*, 2nd Ed. Churchill, London.

[111] Sobel, M. & Elashoff, R.M. (1975). Group testing with a new goal, estimation, *Biometrika* **62**, 181–193.

[112] Stirzaker, D.R. (1975). A perturbation method for the stochastic recurrent epidemic, *Journal of the Institute for Mathematics and its Applications* **15**, 135–160.

[113] Storsaeter, J., Hallander, H., Farrington, C.P., Olin, P., Mollby, R. & Miller E. (1990). Secondary analyses of the efficacy of two acellular pertussis vaccines evaluated in a Swedish phase III trial, *Vaccine* **8**, 457–461.

[114] Stroup, D.F., Williamson, G.D. & Herndon, J.L. (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data, *Statistics in Medicine* **8**, 323–329.

[115] Tillett, H.E. & Spencer, I.-L. (1982). Influenza surveillance in England and Wales using routine statistics, *Journal of Hygiene, Cambridge* **88**, 83–94.

[116] Vesikari, T. (1993). Clinical trials of live oral rotavirus vaccines: the Finnish experience, *Vaccine* **11**, 255–261.

[117] Walter, S.D. (1974). On the detection of household aggregation of disease, *Biometrics* **30**, 525–538.

[118] Walter, S.D., Hildreth, S.W. & Beaty, B.J. (1980). Estimation of infection rates in populations of organisms using pools of variable size, *American Journal of Epidemiology* **112**, 124–128.

[119] Watier, L., Richardson, S. & Hubert, B. (1991). A time series construction of an alert threshold with application to *S. bovismorbificans* in France, *Statistics in Medicine* **10**, 1493–1509.

[120] Zeger, S.L., See, L.-C. & Diggle, P.J. (1989). Statistical methods for monitoring the AIDS epidemic, *Statistics in Medicine* **8**, 3–21.

C.P. FARRINGTON

Community Medicine

Community medicine is a broad medical specialty developed during the twentieth century to cover various aspects of medicine and care in relation to populations (communities) rather than individuals. It embraced the organization and provision of health care throughout a community (or region), including the identification of health problems and needs, and how they were dealt with and met; it involved extensive use of epidemiology, **preventive medicine**, public health, and **health services research**, and more recently, audit and **health economics**. Although still used as a title for some Clinical or Medical School Departments, frequently in conjunction with Family Medicine, Social Medicine, Public Health, or Health Case, sole usage declined towards the end of the twentieth century; it has been replaced by other terms including those noted above.

ANTHONY L. JOHNSON

Co-morbidity

The term "co-morbidity" of two disorders, R and S [1, 3], in its most general sense, indicates only the potential co-occurrence of those disorders in the same unit (e.g. patient, family). Except in those rare cases where the diagnosis of one disorder, say R, explicitly rules out the possibility of the other, S, most disorders are in this sense potentially "co-morbid". There is, however, growing interest in co-morbidity, defined in a more technical sense. In this sense, the co-morbidity of R and S has come to mean some type of nonrandom association between R and S in a population of subjects. There are at least three distinct types of nonrandom such co-morbidity [2], of wide interest in medical research and important to biostatistical issues in research design and analysis.

Clinical Co-morbidity

R and S are said to have "clinical co-morbidity" if the etiology, time course, prognosis, or response to treatment of R is different depending on whether S is or is not present. For example, Feinstein [1, p. 154] points out that a patient's response to treatment of acute pneumococcal pneumonia may be influenced by whether or not the patient has poorly controlled diabetes mellitus or underlying chronic bronchitis at the same time.

This type of co-morbidity is of biostatistical concern in studies of the epidemiology of R, in **clinical trials** assessing the efficacy or effectiveness of treatment of R, and in **health services research** regarding costs related to course, treatment, or prognosis of R. Whether the population of concern in a study in any of these areas should include or exclude subjects with co-morbid S, and if they are included, whether the sample should or should not be stratified on presence or absence of co-morbid S (*see* **Stratification**), makes a major difference to the **power** of statistical **hypothesis testing**, precision of estimating parameters (*see* **Estimation**), and, above all, the clarity of the conclusions.

Epidemiologic Co-morbidity

R and S are said to have "epidemiologic co-morbidity" in a population if the probability of S in a unit with R from that population is different from the probability of S in a unit without R, i.e. the occurrence of R and S in a unit are not independent events. For example, those in a population with depression may be more likely to be alcoholics than those without depression, i.e. there is some correlation between depression and alcoholism.

The most common sources of epidemiologic co-morbidity are shared risk factors. For example, many disorders are more (or less) likely to affect men than women, to affect older subjects than younger, or to affect socially disadvantaged subjects than advantaged ones. Any pair of such disorders is likely to have epidemiologic co-morbidity in a population heterogeneous on those factors, even if, in a subpopulation matched on age, gender, and social class, the two disorders are completely independent. In such a case, the **correlation** between R and S in the heterogeneous population has often been called "pseudocorrelation", since it is a correlation completely explained by a third factor (*see* **Confounding**).

Thus one is likely to find epidemiologic co-morbidity between tobacco and alcohol dependency in the general adult population, a portion of which may arise simply because both tobacco and alcohol abuse is more common among men than among women, and more common in lower socioeconomic status groups than in higher. Within an age–gender matched subpopulation, it may be that tobacco and alcohol use may be independent, that is, not epidemiologically co-morbid. Consequently, the study of epidemiologic co-morbidities can help distinguish those risk factors that might be causal for a disorder from those that are merely markers.

However, epidemiologic co-morbidities may also arise because of "fuzzy" diagnostic boundaries. For example, it is difficult to identify measures of depression and anxiety that are not very highly correlated. When such measures play some role in diagnosing depression and anxiety disorders, one may find epidemiologic co-morbidity between these disorders because the boundaries are so indistinct. What one diagnostician might see as depression, another might classify as anxiety disorder, and vice versa. Consequently, study of epidemiologic co-morbidity also plays a role in the study of diagnostic reliability and validity.

Finally, epidemiologic co-morbidity may arise simply because S is a risk factor for R or vice versa. It may be, for example, that the well-recognized

epidemiologic co-morbidity between depression and alcoholism arises because those suffering depression attempt to self-medicate with alcohol, or because alcoholism induces depression. This is yet another reason why epidemiologic co-morbidity is an important issue in assessing risk factors and identifying possible causal factors in epidemiologic studies.

Familial Co-morbidity

Both clinical and epidemiologic co-morbidity typically tend to refer to co-occurrences of the two disorders in the same person. In contrast, R and S are said to have "familial co-morbidity" if the prevalence of R in the relatives of probands matched on R (either all have R or all do not have R) differs depending on whether the proband does or does not have S. For example, in families of probands, none of whom is an alcoholic, the prevalence of alcoholism among relatives of those probands with major depression may be higher that among relatives of those probands without major depression. In this case, the occurrences of R in the relatives of the probands, not in the probands themselves, is at issue. Indeed, it is possible that R and S cannot, by definition, occur in the same person at the same time, but R and S may still have familial co-morbidity. Such familial co-morbidity has become of interest with the growing interest in the identification of genetic linkages between disorders (*see* **Familial Correlations**).

Two disorders, R and S, may have clinical co-morbidity or epidemiologic co-morbidity, or familial co-morbidity, any two or these, or all three. For example, not only is there epidemiologic co-morbidity among tobacco and alcohol dependency, but evidence is growing that it is more difficult to induce smoking cessation among those who are alcohol-dependent than among those not, which would be clinical co-morbidity. There is no reason why two disorders R and S that have epidemiologic co-morbidity must necessarily have clinical co-morbidity, and presence of either of these types of co-morbidity gives no indication of whether familial co-morbidity pertains as well.

Methods of assessing these types of co-morbidity differ completely. To establish clinical co-morbidity, one samples the population having one or both disorders and assesses some parameter related to etiology, course, prognosis, or response to treatment. Those with neither R nor S are irrelevant to the issue. To establish epidemiologic co-morbidity, one samples the general population, including those with neither R nor S, and assesses incidence or prevalence of both. To establish familial co-morbidity, one identifies probands matched on R, some of whom have and others of whom do not have S, and assesses incidence or prevalence of R among their relatives. One either excludes probands with R (if it is decided to match by sampling probands without R) or excludes probands without R (if it is decided to match by sampling probands with R).

The statistical testing and estimation methods to be used in such assessments comprise fundamentally two-group comparisons of all types. For example, to establish clinical co-morbidity, one might compare survival curves to age of onset of S between those with and without R. To establish epidemiologic co-morbidity, one might estimate or test the **odds ratio** between the occurrences of R and S in the population. To establish familial co-morbidity, one might sample m first-degree relatives for each proband without R and count the number of relatives with R, then compare the distribution of these counts between those probands with or without S.

Co-morbidity has been characterized as "the single most important concept for psychiatric research and practice" [3], and its importance has been stressed in other fields of medical research as well. However, it is a relatively new construct. The majority of references to "co-morbidity" in the medical research literature continue to mean co-occurrence of disorders, whether random or not. Studies specifically related to these types of nonrandom co-morbidity remain relatively rare [4].

References

[1] Feinstein, A.R. (1967). *Clinical Judgment*. Krieger, Huntington.

[2] Kraemer, H.C. (1995). Statistical issues in assessing comorbidity, *Statistics in Medicine* **14**, 721–733.

[3] Maser, J.D. & Cloninger, C.R., eds (1990). *Comorbidity of Mood and Anxiety Disorders*. American Psychiatric Press, Washington.

[4] Merikangas, K., Angst, J., Eaton, W., Canino, G.R.-S.M., Wacker, H., Wittchen, H.U., Andrade, L., Essau, C., Kraemer, H., Robins, L. & Kupfer, D. (1996). Comorbidity and boundaries of affective disorders with anxiety disorders and substance abuse: results of an international task force, *British Journal of Psychiatry* **168**, 58–67.

HELENA CHMURA KRAEMER

Comparative Genomic Hybridization

Comparative genomic hybridization (CGH) is a method for directly identifying regions of gains and losses of genomic material in chromosomes. It is accomplished by extracting deoxyribonucleic acid (**DNA**) from both test and reference cells, each labeled with a different colored fluorescent dye (e.g. red and green). A pool of test and reference DNA is hybridized to a set of normal chromosomes. The result is measured as a series of test-to-reference signals from the fluorescent dyes along each chromosome. An excess of test signal in a chromosomal region indicates gain of genomic material in test relative to reference in that region, while an excess of reference signal indicates loss of genomic material in test relative to reference [3, 4].

Since CGH is applied to the entire genome (i.e. all the chromosomes), the data consist of test-to-reference ratios for a large number of distinct chromosomal regions, called loci (*see* **Gene**). These loci can be ordered from the tip of the short arm of chromosome 1 (called 1pter) to the tip of the long arm of chromosome 22 (called 22qter) in humans. Thus, the human genome is characterized by CGH as a series of such ratios, called a profile. These ratios are first standardized by dividing each ratio by the mean (or median) of all the ratios so that the standardized ratios will have mean (or median) equal to 1.0.

Statistical methods are used to identify chromosomal regions where CGH ratios differ significantly from 1.0. Ideally, there will be replicate samples of the same test material that can be used to estimate a standard deviation for each locus along the CGH profile. Naive methods define CGH loss at a locus if the (standardized) mean of the replicate ratios is more than 2 standard deviations (sd) below 1.0 and a gain if the mean is more than 2 sd above 1.0 [2]. Unfortunately, observations from several data sets suggest that ratios vary systematically from 1.0 along the CGH profile so that the naive method is prone to error. When CGH is applied to whole (intact) chromosomes, the recommended procedure is to obtain replicate samples of reference vs. reference as well as replicate samples of test vs. reference so that a *t*-test can be used to determine at which loci DNA is gained or lost [5, 9].

CGH is now being applied to microarrays, where genomic DNA is represented by thousands of small segments [1, 6, 7]. Experiments have shown that when used to define abnormality (i.e. gain or loss) in tumors based on replicate sets of reference vs. reference CGH means and sd, this method does not work because of hybridization-to-hybridization variability, especially variability between tumor and normal samples. Statistical methods for dealing with these large arrays of CGH ratios are currently under development. One method circumvents the problem of defining gains or losses by using two-sample *t*-tests applied to the (standardized) CGH ratios themselves to find loci where CGH ratios differ between chromosomes in two groups of tumors (for example, invasive vs. *in situ*) or between tumor and normal chromosomes. Because of the large numbers of *t*-tests applied (one for each of the thousands of loci) and lack of independence among the different loci, permutation methods are used to define the expected distribution of the *t*-statistics [8].

References

[1] Albertson, D.G., Ulstra, B., Segraves, R., Collins, C., Dairkee, S.H., Kowbel, D., Kuo, W.-L., Gray, J.W. & Pinkel, D. (2000). Quantitative mapping of amplicon structure by array CGH identifies vitamin D-24 hydroxylase (CYP24) as a candidate oncogene, *Nature Genetics* **25**, 144–146.

[2] du Manoir, S., Speicher, M.R., Joos, S., Schrock, E., Popp, S., Dohner, H., Kovacs, G., Robert-Nicoud, M., Lichter, P. & Cremer, T. (1993). Detection of complete and partial chromosome gains and losses by comparative genomic *in situ* hybridization, *Human Genetics* **90**, 590–610.

[3] Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F.M. & Pinkel, D. (1992). Comparative genomic hybridization: a powerful new method for cytogenetic analysis of solid tumors, *Science* **258**, 818–821.

[4] Kallioniemi, A., Kallioniemi, O.-P., Piper, J., Chen, L., Smith, H.S., Gray, J.W., Pinkel, D. & Waldman, F.M. (1994). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization, *Proceedings of the National Academy of Sciences* **91**, 2156–2160.

[5] Moore, D.H. II, Pallavicini, M., Cher, M.L. & Gray, J.W. (1997). A *t*-statistic for objective interpretation of comparative genomic hybridization (CGH) profiles, *Cytometry* **28**, 183–190.

[6] Pinkel, D., Segraves, R., Sudar, S., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Z.,

Dairkee, S., Ljung, B.-M., Gray, J.W. & Albertson, D.G. (1998). High resolution analysis of DNA copy number variation in breast cancer using comparative genomic hybridization to DNA microarrays, *Nature Genetics* **20**, 207–211.

[7] Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. & Lichter, P. (1997). Matrix-based comparative genomic hybridization:biochips to screen for genomic imbalances, *Genes, Chromosomes and Cancer* **20**, 399–407.

[8] Tusher, V.G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences* **98**, 5116–5121.

[9] Yu, L.-C., Moore, D.H., Magrane, G., Cronin, C., Pinkel, D., Lebo, R.V. & Gray, J.W. (1997). Objective aneuploidy detection for fetal and neonatal screening using comparative genomic hybridization (CGH), *Cytometry* **28**, 191–197.

DAN H. MOORE II

Compartment Models

Compartmental modeling is a well-established paradigm for describing system kinetics in the natural and biomedical sciences. One early use of such modeling began with attempts to analyze data on the distribution and metabolism of tracer-labeled compounds in the 1920s [5]. The field gained great impetus with the widespread availability of radioactive tracers in the late 1940s and early 1950s. Sheppard [13] is credited with the first use of the term "compartment" in 1948 to describe a kinetic entity. He conceptualized physiological systems as "well-stirred" hydrology models, noting that "real compartments may exist whose contents are homogeneous and are separated from one another by real boundaries". Berman and Schoenfeld [2] and others in the 1960s established the basic mathematical properties of the underlying linear compartmental model with constant coefficients. This early and the subsequent development of compartmental modeling is contained in the classic books by Jacquez [6], which also describe many diverse applications of the methodology (see also [4]). A parallel but virtually independent development of compartmental modeling occurred in pharmacokinetics, as outlined in [3], (*see* **Pharmacokinetics and Pharmacodynamics**).

Standard Deterministic Model

The basic structure of a compartmental model in a physiological context is illustrated in the schematic in Figure 1. The following notation is standard for an n-compartment model. Let

1. $X_i(t)$ be the amount of substance in compartment i at time t,
2. $\mathbf{X}(t) = [X_1(t), \ldots, X_n(t)]'$ be the column-vector of amounts at time t,
3. k_{ij}, for $i = 0, 1, \ldots, n; j = 1, \ldots, n; i \neq j$; denote the fractional flow rate to i from j, where 0 represents the system exterior,
4. $k_{jj} = -\sum_{i \neq j} k_{ij}$, denote the total outflow rate from j,
5. $\mathbf{K} = (k_{ij})$, for $i, j > 0$; be the $n \times n$ **matrix** of k_{ij} coefficients,
6. $\lambda_1, \ldots, \lambda_n$ be the **eigenvalues** of $\mathbf{K}$, with $\mathbf{\Lambda}$ as the diagonal matrix of λ_i, and
7. $T_1, \ldots, T_n$ be the corresponding **eigenvectors** of $\mathbf{K}$ with $n \times n$ eigenvector matrix $\mathbf{T} = (T_1, \ldots, T_n)$.

A (linear) compartment model assumes that each derivative, $\dot{X}_i(t)$, is a linear function of the $X_i(t)$. For example, the most general two-compartment model assumes the following:

$$\begin{aligned} \dot{X}_1(t) &= -(k_{01} + k_{21})X_1(t) + k_{12}X_2(t), \\ \text{and} \quad \dot{X}_2(t) &= k_{21}X_1(t) - (k_{02} + k_{12})X_2(t). \end{aligned} \tag{1}$$

Compartment models may be expressed in matrix form as follows:

$$\dot{\mathbf{X}}(t) = \mathbf{K}\mathbf{X}(t). \tag{2}$$

The above vector of differential equations has the following solution:

$$\mathbf{X}(t) = \exp(\mathbf{K}t)\mathbf{X}(0) \tag{3}$$

with the matrix exponential defined as $\exp(\mathbf{K}t) = \mathbf{I} + \sum_{i=1}^{\infty} \mathbf{K}^i t^i / i!$.

Two corollaries are helpful in practical applications. In most natural applications, the system is "open" and at least "weakly connected", so that no compartment is a "sink". Assuming these conditions, and that $\mathbf{K}$ admits a spectral decomposition (*see* **Matrix Algebra**), a sufficient condition for which is distinct λ_i, one can show

(a) the λ_i have negative real parts, and
(b) $\mathbf{X}(t) = \mathbf{T}\exp(\mathbf{\Lambda}t)\mathbf{T}^{-1}\mathbf{X}(0)$,

where $\exp(\mathbf{\Lambda}t)$ is a diagonal matrix with elements $\exp(\lambda_i t)$.

Assuming also that the λ_i are real, as they are for all two-compartment and most three-compartment

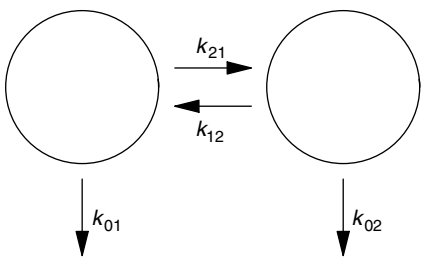

Figure 1 Schematic of general two-compartment model

models, it follows that

$$X_i(t) = \sum_{j=1}^{n} A_{ij} \exp(\lambda_j t), \quad \text{for } i = 1, \ldots, n. \quad (4)$$

In such "sum of exponentials" models, the A_{ij} and λ_j are often called the "macroparameters", and they are usually involved functions of the k_{ij} microparameters. The equivalent solutions based on the k_{ij} parameters are given explicitly for the common two- and three-compartment models in many books, including [4] and [6]. The pharmacokinetic applications share the sum of exponential formulation; however, one notable difference is its reversed order of subscripts in the k_{ij} flow rates.

In practice, often data are concentrations, $C_i(t) = X_i(t)/V_i$, where the V_i are compartment "volumes". In many applications, there are "inputs", leading to a more general matrix model

$$\dot{\mathbf{X}}(t) = \mathbf{K}\mathbf{X}(t) + \mathbf{U},$$
$$\mathbf{Y}(t) = \mathbf{D}\mathbf{X}(t), \quad (5)$$

where $\mathbf{U}$, $\mathbf{Y}(t)$ and $\mathbf{X}(t)$ are vectors of inputs to, outputs from, and amounts in the compartments, respectively, and $\mathbf{K}$ and $\mathbf{D}$ are conformable matrices. These extensions preserve the "sum of exponentials" model. More general formulations, including time-varying $\mathbf{K}(t)$, $\mathbf{U}(t)$, and $\mathbf{D}(t)$ matrices, are considered in general linear systems theory [8].

These models may be fitted to data, using either weighted or unweighted nonlinear **least squares**, to estimate the macro- or microparameters. The underlying theory is described in [1]. Many computer packages are available, including WinSAAM [14] (*see* **Software for Clinical Trials**).

As an illustration, a standard model for calcium kinetics in humans illustrated in Figure 2 has three compartments, namely, plasma (1), soft tissue (2), and bone (3). Weiss et al. [15] describe an experiment in which labeled calcium was introduced as a bolus injection into the plasma compartment of an adult woman, after which the concentration of labeled calcium was observed usually every 8 hours for almost 24 days, as illustrated in Figure 3(a). The fitted concentration-time curve,

$$\mathbf{C}_1(t) = 351.39e^{-5.835t} + 197.20e^{-0.3754t} + 145.31e^{-0.0144t}, \quad (6)$$

fits the data well, and gives estimated rates for $k_{01}, k_{21}, k_{12}, k_{31}$, and k_{13}, respectively, of 0.0652, 2.611, 2.998, 0.389, and 0.162/h, with estimated approximate standard errors of less than 10% for each rate coefficient.

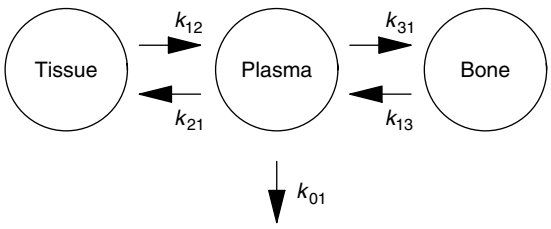

Figure 2 Schematic of standard three-compartment model of calcium clearance

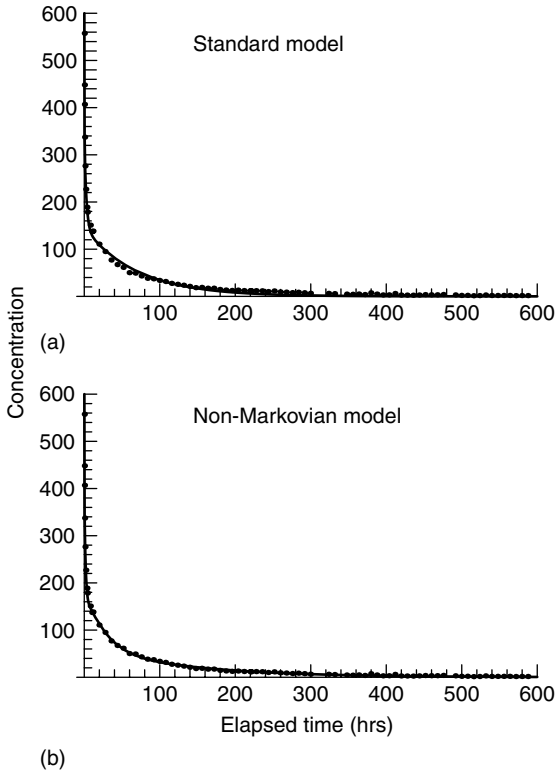

Figure 3 Observed calcium clearance data with fitted curve

An Analogous Stochastic Model

Jacquez [6, p. 235] argues that "In deterministic theory, ... the material in a compartment is treated as a continuum. But matter ... comes in discrete units ... Consequently it is important to develop the theory ... in terms of the probabilities of transfers of unit(s)." A compelling example is modeling the passage of hay particles in ruminants [11]. An analogous stochastic development (*see* **Migration Processes**) requires the following additional notation. Let

8. $P_{ij}(t)$ be the probability that a particle starting in j at time 0, will be in i at time t,
9. $\mathbf{P}(t) = [P_{ij}(t)]$ be an $n \times n$ matrix of probabilities.

Also, let

10. R_{ij} be the retention time during a single visit in j of a particle whose next transfer will be to i,
11. S_{ij} be the total residence time that a particle originating in j will accumulate in i during all of its visits,
12. $\mathbf{E}[S]$ denote the $n \times n$ matrix of mean residence times, with the elements $E[S_{ij}]$.

The key assumptions in the stochastic model are

(a) the conditional probability that, for small Δt, a random particle in j at time t will transfer to i by time $t + \Delta t$ is

$$k_{ij}\,\Delta t,$$

(b) all particles are independent.

The former assumption with the constant conditional flow probability (i.e. **hazard rate**) is equivalent to assuming **exponentially distributed** R_{ij} retention times for all i and j. Under these assumptions, one can show that

$$\mathbf{P}(t) = \exp(\mathbf{K}t), \tag{7}$$

which, under the same regularity conditions as in the deterministic model, implies that each "occupancy" probability function is a sum of exponentials model, that is,

$$P_{ij}(t) = \sum_{\ell} A_{ijl}\,\exp(\lambda_l t) \quad \text{for } i, j = 1, \ldots, n. \tag{8}$$

Thus, this stochastic formulation gives the same sum of exponentials regression model for data.

However, the stochastic model gives additional insight into particle kinetics. The underlying process is **Markovian**, whereupon the matrix of mean residence times (MRT) is

$$\mathbf{E}(S) = -\mathbf{K}^{-1}, \tag{9}$$

with corresponding results available for higher-order moments of the S_{ij} variables, as well as for the number of particle cyclings. As an illustration, the estimated $\mathbf{K}$ matrix for the previous calcium data is

$$\mathbf{K} = \begin{bmatrix} -3.0652 & 2.999 & 0.162 \\ 2.611 & -2.999 & 0 \\ 0.389 & 0 & -0.162 \end{bmatrix}. \tag{10}$$

The first column of $-\mathbf{K}^{-1}$ gives the MRT in the plasma, soft tissue, and bone compartments of a calcium particle introduced into the plasma. The results are 15.34, 13.36, and 36.86 h, respectively, for a total expected residence time in the body of 50.22 h.

Deterministic Model with Time Lags

A generalization of the deterministic model recognizes that transfer of material between compartments may not be instantaneous. Time lags could be introduced, either of fixed size, say τ_{ij}, or more commonly in modeling, of random size with a density function, $h_{ij}(\tau)$. In the latter case, the two-component deterministic model would be

$$\begin{aligned} \dot{X}_1(t) &= k_{11}X_1(t) + k_{12}\int_{-\infty}^{t} X_2(\tau)h_{12}(t-\tau)\mathrm{d}\tau \\ \dot{X}_2(t) &= k_{21}\int_{-\infty}^{t} X_1(\tau)h_{21}(t-\tau)\mathrm{d}\tau + k_{22}X_2(t). \end{aligned} \tag{11}$$

In practice, tractable $h_{ij}(\tau)$ functions are obtained using the "hidden variables" approach, which introduces subsystems of compartments to generate the desired lag distributions. In so doing, "compartmental systems with lags are equivalent to larger compartmental systems without lags" [7], and all of the basic mathematical properties of linear systems are preserved. The procedure is also illustrated in [7].

Analogous Non-Markovian Stochastic Model

The compartment paradigm provides an immediate and tractable stochastic analog of time lags. Neuts [12] shows that any nondegenerate distribution of a positive variable may be represented as a phase-type (PH) distribution. A PH distribution is generated by definition as the time to absorption in a continuous time **Markov chain**, which in turn implies that PH distributions may be generated by stochastic compartmental submodels. Hence, the assumption of an exponentially distributed retention time in a given compartment may be generalized by utilizing an (expanded) compartmental subsystem to generate the desired nonexponential variable. As a simple illustration, setting $k_{01} = k_{12} = 0$ and $k_{21} = k_{02}$ in Figure 1 yields the special case of an Erlang (2) PH distribution of particle retention times.

In practice, any assumed nonexponential variable in the model is approximated to the desired accuracy by utilizing appropriate compartmental submodels, and the resulting larger compartmental systems are fitted to data using standard software. As an illustration, suppose that the previous exponential retention time in the bone compartment, comp 3, is replaced with a PH distribution with four (pseudo)compartments, namely, 3*, 4*, 5*, and 6*, as illustrated in Figure 4. The new flow rates in the submodel are created from the previous two rates, k_{31} and k_{13}, as follows: $k_{3^*1} = k_{4^*3^*} = k_{31} + k_{13}$, $k_{5^*4^*} = k_{6^*5^*} = k_{4^*6^*} = k_{31}$ and $k_{14^*} = k_{13}$. Instead of the previously assumed homogeneous (well-stirred) bone compartment, this expanded model creates a physiologically more realistic short cycle of particles in "soft bone" with possible additional cycles for "hard bone". This yields a natural long-tailed PH distribution without increasing the number of parameters in the model. The estimated parameters k_{01}, k_{21}, k_{12}, k_{31}, and k_{13}, for the expanded six-compartment model are 0.0607, 3.131, 3.735, 0.030, and 0.414 [10]. Some of the eigenvalues of this expanded system with cycling are complex, as expected. The fitted curve,

$$\begin{aligned} C_1(t) = {} & 356.16e^{-7.105t} + 68.50e^{-0.0075t} \\ & + e^{-0.566t}[44.98 \sin b_1 t + 187.85 \cos b_1 t] \\ & + e^{-0.038t}[32.85 \sin b_2 t + 110.26 \cos b_2 t] \end{aligned} \tag{12}$$

with $b_1 = 0.2852$ and $b_2 = 0.0162$, fits the data better as illustrated in Figure 3(b), with an 80% reduction in **mean squared error** (MSE), due to its longer tail. The MRT for the expanded model may be obtained from the negative inverse of the corresponding 6×6 **K** matrix. After summing MRT for the bone submodel, the estimated MRT in the plasma, soft tissue, and bone compartments of a calcium particle introduced into the plasma are 16.47, 13.81, and 69.47 hours, respectively. The latter represents a substantial increase in the MRT from the simple linear model without lags, or correspondingly from the simple Markovian model.

The compartment model construct has also been generalized to include birth and nonlinear rate features, thus broadening its application to broad areas of population biology and other biomedical problems [9]. It is expected that its clear linkage between the mathematical formalism and the underlying biological system, and its ease of application will sustain the widespread use of compartment modeling in practice.

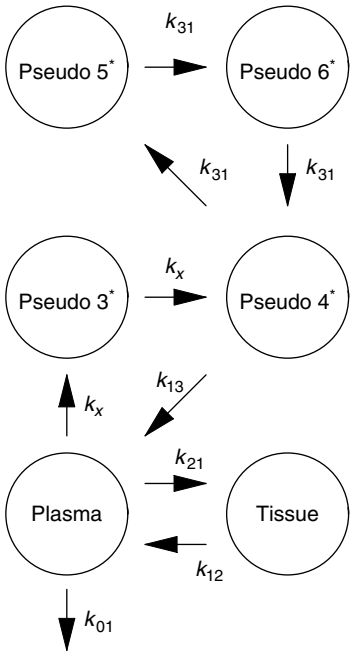

Figure 4 Schematic of non-Markovian model, with four pseudo compartments generating a PH retention time distribution. (Note $k_x = k_{13} + k_{31}$)

References

[1] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

[2] Berman, M. & Schoenfeld, R. (1956). Invariants in experimental data on linear kinetics and the formulation of models, *Journal of Applied Physics* **27**, 1361–1370.

[3] Gibaldi, M. & Perrier, D. (1982). *Pharmacokinetics*, 2nd Ed., M. Dekker, New York.

[4] Godfrey, K.R. (1983). *Compartmental Models and Their Applications*. Academic, London.

[5] Hevesy, G. (1962). *Adventures in Radioisotope Research. The Collected Papers of G. Hevesy*, Vols. 1 and 2, Permagon, New York.

[6] Jacquez, J.A. (1972, 1985, 1996). *Compartmental Analysis in Biology and Medicine*, 1, 2, and 3 Ed., Elsevier, University of Michigan Press, and Biomedware, Ann Arbor, MI, respectively.

[7] Jacquez, J.A. & Simon, C.P. (2002). Qualitative theory of compartmental systems with lags, *Mathematical Biosciences* **180**, 329–362.

[8] Kalman, R.E., Falb, P.L. & Arbib, M.A. (1969). *Topics in Mathematical Systems Theory*. McGraw-Hill, New York.

[9] Matis, J.H. & Kiffe, T.R. (2000). *Stochastic Population Models: A Compartmental Perspective*, Lecture Notes in Statistics 145, Springer, New York.

[10] Matis, J.H. & Wehrly, T.E. (1998). A general approach to non-Markovian compartmental models, *Journal of Pharmacokinetics and Biopharmaceutics* **11**, 77–92.

[11] Matis, J.H., Wehrly, T.E. & Ellis, W.C. (1989). Some generalized stochastic compartment models for digesta flow, *Biometrics* **45**, 703–720.

[12] Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press, Baltimore.

[13] Sheppard, C.W. (1962). *Basic Principles of the Tracer Method*. Wiley, New York.

[14] Wastney, M.W., Patterson, B.H., Linares, O.A., Greif, P.C. & Boston, R.C. (1998). *Investigating Biological Systems Using Modeling*. Academic, New York.

[15] Weiss, G.H., Goans, R.E., Gitterman, P.D., Abrams, S.A., Vieira, N.E. & Yergey, A.L. (1994). A non-Markovian model for calcium kinetics, *Journal of Pharmacokinetics and Biopharmaceutics* **22**, 367–379.

(*See also* **Dose-response in Pharmacoepidemiology**; **Model, Choice of**)

JAMES H. MATIS, THOMAS R. KIFFE & THOMAS E. WEHRLY

Competing Risks

"Competing risks" refers to the study of mortality patterns in a population of individuals, all subject to the same $k \geq 2$ competing risks or **causes of death**. Specifically, the objective is to isolate the effect of a given risk, or a subset of risks, acting on a population. The use of competing risks dates back to 1760 and evolved out of a controversy over smallpox inoculation.

According to Karn [22] and Todhunter [30], smallpox inoculation in the 1700s was administered by applying leeches to the body, a practice that could lead to acute illness and death. Physicians argued whether the benefits of inoculation outweighed the initial risk of death. Daniel **Bernoulli** [9], in a 1760 memoir entitled "Essai d'une nouvelle analyse de la mortalité causée par le petite vérole; et des advantages de l'inoculation pour le prévenir", tried to estimate the expected increase in lifespan (*see* **Life Expectancy**) if smallpox were eliminated. This calculation could then be used to weigh the pros and cons of smallpox inoculation.

Similarly, in the modern treatment of competing risks we are interested in isolating the effect of individual risks. For example, suppose we wish to assess a new treatment for heart disease. In a long-term study of this treatment on a sample of individuals, some will die of causes other than heart disease. The appropriate analysis of this problem must account for the competing effects of death from other causes.

Various methods have been proposed to study the problem of competing risks. For example, Makeham [24] formulated the law of composition of decremental forces and applied it to competing risks theory. A multiple decrement model is a time-continuous **Markov model** with one transient state and k absorbing states. An excellent account of the use of multiple decrement theory to explain competing risks may be found in Chiang [12].

Another approach to modeling competing risks is through the use of latent failure times. This method was first advocated by Sampford [28] who proposed an "accidental death model". In this approach each individual has latent failure times T_1 and T_2, where T_1 corresponds to time of natural death and T_2 to time to accidental death. Sampford assumed that T_1 and T_2 are independent and **normally distributed** and death occurred at time X equal to the minimum of T_1 and T_2. Berkson & Elveback [8] considered a similar model to study the effect of smoking on lung cancer assuming that the latent failure times were independent **exponentially** distributed **random variables**. Moeschberger & David [25] generalized these ideas to k causes of death with general **survival distributions**. Excellent reviews of the theory of competing risks are given by Gail [18, 19], David & Moeschberger [15], and Birnbaum [10].

In this article, latent failure times are used to describe competing risks models. We assume that all individuals in a population are subject to k competing causes of death, $D_1, \ldots, D_k$. For each possible cause of death, D_i, there corresponds a latent failure time, T_i, a positive random variable representing the age at death in the hypothetical situation in which D_i is the only possible cause of death. The joint distribution of the latent failure times is given by the **multivariate survival distribution**

$$H^C(t_1, \ldots, t_k) = P(T_1 > t_1, \ldots, T_k > t_k), \qquad (1)$$

defined for all nonnegative values $t_1, \ldots, t_k$. We use a superscript C to highlight that this is the joint distribution of the complete set of risks acting on the population. The latent failure times are mostly unobservable and serve only as a theoretical construct. In contrast, the observable random variables for each member of a population of individuals are the actual times to death, denoted by the positive random variable X, and the cause of death, Δ, which may take one of the integer values $1, \ldots, k$. The observed time of death, X, is taken to be the minimum of $T_1, \ldots, T_k$, and Δ indexes this cause of death, i.e. $\Delta = i$ if $X = T_i$. For simplicity we assume the joint distribution is absolutely continuous so that Δ is uniquely defined.

The study of competing risks considers the interrelationship of three types of probabilities of death from specific causes. These are:

1. *The crude probability*: the probability of death from a specific cause in the presence of all other risks acting on the population. This is also referred to as **absolute risk**. An example of a crude probability is the answer to the question: What is the chance that a woman will die of breast cancer between ages 40 and 60?
2. *The net probability*: the probability of death if a specific risk is the only risk acting on a population, or conversely, the probability of death if a

specific cause is eliminated from the population. For example, what is the chance of surviving to age 60 if cancer were the only cause of death?

3. *The partial crude probability*: the probability of death from a specific cause when some risks are eliminated from the population. For example, what is the chance that a woman would die from breast cancer between ages 40 and 60 if smallpox were eliminated?

In the next section we define notation and give some fundamental relationships between the three different types of probabilities. Then we consider the issue of **identifiability** of these probabilities and discuss some philosophical issues regarding the study of competing risks in light of nonidentifiability. Finally, we address statistical issues of **estimation** and **hypotheses testing** based on a sample of observable data.

Notation and Relationships

Crude Probability

Crude probability is a way of describing the probability distribution for a specific cause of death in the presence of all causes. Crude probability refers to quantities derived from the probability distribution of the observable random variables, X and Δ, where X is time to death, and $\Delta = 1, \ldots, k$ is cause of death. Two approaches have been used to describe the distribution of X and Δ. The first is through subdistribution functions:

$$F_i^C(x) = \Pr(X \le x, \Delta = i), \quad i = 1, \ldots, k.$$

The function $F_i^C(x)$ denotes the proportion of all individuals who are observed to die from cause D_i at or before time x in the presence of all causes of death. We use the superscript C to denote all causes of death, i.e. $C = \{1, \ldots, k\}$. For example, if D_1 represents death from breast cancer, then the chance that a woman dies from breast cancer between ages 40 and 60 would be equal to $[F_1^C(60) - F_1^C(40)]$. Note that $F_i^C(\infty)$ is the proportion of individuals who will be observed to die from cause D_i, and $\sum_{i=1}^k F_i^C(x) = F^C(x)$ defines the distribution function for death from any cause, i.e. $F^C(x) = \Pr(X \le x)$. We denote the overall survival distribution as $S^C(x) = 1 - F^C(x)$.

Another way to define the distribution of X and Δ is through the use of k cause specific **hazard rate** functions given by

$$\lambda_i^C(x) = \lim_{h \to 0} \left[\frac{\Pr(x \le X < x + h, \Delta = i | X \ge x)}{h} \right], \quad i = 1 \ldots, k.$$

The ith cause-specific hazard is the rate of death at time x from cause i among individuals who are still alive at time x. Calculus yields the following relationships:

$$\lambda_i^C(x) = \frac{\mathrm{d}F_i^C(x)}{\mathrm{d}x} \Big/ S^C(x),$$

$$\lambda^C(x) = \sum_{i=1}^k \lambda_i^C(x) = \frac{\mathrm{d}F^C(x)}{\mathrm{d}x} \Big/ S^C(x), \tag{2}$$

$$S^C(x) = \exp[-\Lambda^C(x)]; \qquad \Lambda^C(x) = \int_0^x \lambda^C(u)\,\mathrm{d}u,$$

$$F_i^C(x) = \int_0^x \exp[-\Lambda^C(u)]\lambda_i^C(u)\,\mathrm{d}u.$$

Note that $\Lambda^C(x)$ is defined as the **cumulative hazard** function of death from any cause and is the sum of the individual cause-specific integrated hazards. The relationship given in (2) illustrates that there is a one-to-one relationship between subdistribution functions and cause-specific hazard functions.

The crude probability distributions may be derived from the joint distribution of the latent failure times as follows. Because $X = \min(T_1, \ldots, T_k)$, it follows that $S^C(x) = H^C(x, \ldots, x)$; hence, it is straightforward to show that

$$\frac{\mathrm{d}F_i^C(x)}{\mathrm{d}x} = -\frac{\partial H^C(t_1, \ldots, t_k)}{\partial t_i}\bigg|_{t_1 = \cdots = t_k = x}.$$

Using (2), the cause-specific hazard function is given by

$$\lambda_i^C(x) = \frac{-\dfrac{\partial H^C(t_1, \ldots, t_k)}{\partial t_i}\Big|_{t_1 = \cdots = t_k = x}}{H^C(x, \ldots, x)}. \tag{3}$$

This relationship was derived by Gail [18] and Tsiatis [31].

Cause-specific hazard functions and cause-specific subdistribution functions may also be defined for a subset of risks. We use italicized capital letters to

index a subset of the risks $1, \ldots, k$; for example, J may be used to denote such a subset of risks. The complement of J is equal to $C - J$ and is denoted by $\overline{J}$. The subdistribution function for failing from any of the causes in J is given by

$$F_J^C(x) = \Pr(X \leq x, \Delta \in J) = \sum_{i \in J} F_i^C(x),$$

and the cause-specific hazard of failing from any of the causes in J is

$$\lambda_J^C(x) = \lim_{h \to 0} \left[\frac{\Pr(x \leq X < x + h, \Delta \in J | X \geq x)}{h} \right] = \sum_{i \in J} \lambda_i^C(x).$$

The Net Probability

The net probability is the probability distribution of time to death if only one cause of death acted on a population. If we are interested in the net probability distribution from cause D_i, then this would be the **marginal probability** distribution of the latent failure time, T_i, given by

$$S_i^i(x) = \Pr(T_i > x) = H^C(t_1, \ldots, t_k) | t_i = x, \quad t_j = 0, j \neq i.$$

We use superscript i to highlight the fact that we consider only the case where D_i is acting on a population. For example, if D_1 denotes death from cancer, then the chance of surviving to age 60 if cancer were the only cause of death would be given by S_1^1 (60).

The net distribution may be defined through the net or marginal hazard function for T_i, that is,

$$\lambda_i^i(x) = \lim_{h \to 0} \left[\frac{\Pr(x \leq T_i < x + h | T_i \geq x)}{h} \right].$$

The net hazard function and net survival distribution are related to each other as follows:

$$\lambda_i^i(x) = -\frac{\mathrm{d}S_i^i(x)}{\mathrm{d}x} \Big/ S_i^i(x), \qquad S_i^i(x) = \exp[-\Lambda_i^i(x)], \tag{4}$$

where $\Lambda_i^i(x) = \int_0^x \lambda_i^i(u)\, \mathrm{d}u$.

One of the key results in competing risks theory is for the case where the latent failure times are assumed to be statistically independent, i.e.

$$H^C(t_1, \ldots, t_k) = \prod_{i=1}^{k} S_i^i(t_i).$$

From (1) it is a simple exercise to show that the ith cause-specific hazard function, $\lambda_i^C(x)$, is equal to the ith net-specific hazard function, $\lambda_i^i(x)$. This important fact allows one to use the crude probability distribution of the observables to obtain net probabilities. Specifically, formulas (1) and (2) may be used to show that the net survival distribution is related to the crude subdistribution functions by

$$H_i^i(x) = \exp\left[-\int_0^x \frac{\mathrm{d}F_i^C(u)}{S^C(u)} \right]. \tag{5}$$

Because $F_i^C(u)$ and $S^C(u)$ may be estimated from a sample of observable data, (5) suggests obvious methods for estimating net survival probabilities when the latent failure times are assumed independent, which are described in detail later. Although the crude cause-specific hazard is equal to the net-specific hazard when the latent failure times are independent, the converse is not true. Examples where non-independent latent failure times have cause-specific hazards equal to the net-specific hazards, although mathematically possible, are generally artificial constructs and not important from an applied perspective.

For many applications it may not be reasonable to assume that the latent failure times are independent. In such cases the relationship between net and crude probabilities becomes more complicated. Without additional assumptions, there is a problem of nonidentifiability discussed in greater detail later.

Partial Crude Probability

We now show how to characterize the distribution of probability of death from a subset of causes acting on a population in the hypothetical situation where all other causes of death are eliminated. Similar to crude probabilities, partial crude probabilities may be expressed through partial crude subdistribution functions or partial crude cause-specific hazard functions. Define X^J and Δ^J respectively as the time of death and cause of death in the hypothetical case where individuals are only subject to the causes of death in J, i.e. the causes $\overline{J}$ are eliminated. In terms of latent failure times, $X^J = \min(T_i, i \in J)$ and $\Delta^J = i$, if

$X^J = T_i,\ i \in J$. The partial crude subdistribution function is given by

$$F_i^J(x) = \Pr(X^J \leq x, \Delta^J = i), \quad i \in J,$$

and the partial crude cause-specific hazard is given by

$$\lambda_i^J(x) = \lim_{h \to 0} \left(\frac{\Pr(x \leq X^J < x + h, \Delta^J = i | X^J \geq x)}{h} \right), \quad i \in J.$$

These definitions may be extended in a natural way to subsets K of J, i.e.

$$F_K^J(x) = \sum_{i \in K} F_i^J(x)$$

and

$$\lambda_K^J(x) = \sum_{i \in K} \lambda_i^J(x).$$

If $J = C$, then partial crude probabilities are the same as crude probabilities, and if $J = i$, so that there is only one cause of death, then partial crude probability is the same as net probability.

Using the same logic as for crude probabilities, we can derive the partial crude cause-specific hazard function from the joint distribution of the latent failure times in a manner similar to that for (3). The partial crude cause-specific hazard is given by

$$\lambda_i^J(x) = \frac{-\left. \dfrac{\partial H^C(t_1, \ldots, t_k)}{\partial t_i} \right|_{t_j = x, j \in J; t_j = 0, j \in \overline{J}}}{H^C(t_1, \ldots, t_k)|_{t_j = x, j \in J; t_j = 0, j \in \overline{J}}}, \quad (6)$$

and the partial crude subdistribution function may be expressed as

$$F_i^J(x) = \int_0^x \exp[-\Lambda_J^J(u)] \lambda_i^J(u)\, \mathrm{d}u, \quad i \in J, \quad (7)$$

where $\Lambda_J^J(u) = \int_0^u \lambda_J^J(v)\, \mathrm{d}v$.

Of particular interest is the case when the latent failure times in the set J are independent of the latent failure times in $\overline{J}$. Comparing (6) with (3) we see that the ith partial crude cause-specific hazard function, $\lambda_i^J(x)$, is equal to the overall crude cause-specific hazard function, $\lambda_i^C(x)$, $i \in J$. This allows us to express the unobservable partial crude probabilities in terms of the observable crude probabilities. So, for example, the partial crude subdistribution function may be expressed in terms of the observable crude subdistribution functions as follows:

$$F_i^J(x) = \int_0^x \exp[-\Lambda_J^C(u)] \lambda_i^C(u)\, \mathrm{d}u, \quad i \in J, \quad (8)$$

where

$$\lambda_i^C(u) = \frac{\mathrm{d}F_i^C(u)}{\mathrm{d}u} \Big/ S^C(u).$$

The above relationships hold whenever the latent failure times in J and $\overline{J}$ are independent. It is not necessary that the failure times within J or $\overline{J}$ be independent.

Issues Regarding the Use and Interpretation of Competing Risks

A major aim in many competing risks studies is the estimation of net survival probabilities. The ability to isolate the effect of one risk acting on a population is intuitively attractive, especially if the focus of a study is to evaluate the effect of an intervention that is targeted at reducing mortality from that specific cause. Of course, net survival probabilities are hypothetical quantities and not directly observable in a population; therefore they must be computed from the available information on the distribution of observables, or what we refer to as *crude probabilities*. Previously, we derived the net survival distribution for a specific risk D_i as a function of the observable crude probabilities under the assumption that the different latent failure times were independent of each other. The independence assumption is critical, because in this case the crude cause-specific hazard function is equal to the net hazard function, which leads to the important relationship given by (5).

In some situations such an assumption of independence may be reasonable. For example, when studying cause of death from a specific disease, it may be reasonable to assume that death from accidental causes is independent of those causes associated with the disease. Of course, there are other scenarios for which the independence assumption is not plausible. It is therefore important to consider the relationship of net probabilities to crude probabilities in the case where the latent failure times are not independent.

As we showed in (3), given any joint distribution of latent failure times, there exists a corresponding set of crude cause-specific hazard functions, or equivalently a set of crude cause-specific subdistribution

functions. Unfortunately, the converse is not true, as there exist many joint distributions, $H^C(t_1, \ldots, t_k)$, that would result in the same set of crude subdistribution functions, $F_i^C(x)$, $i = 1, \ldots k$. These different joint distributions of latent failure times, each resulting in the same set of subdistribution functions, would lead to different net survival probabilities. Consequently we cannot identify net survival probabilities from corresponding crude probabilities. Because crude survival distributions define the observable random variables, we cannot estimate the net survival probabilities from observable data without making additional assumptions that cannot be verified from the observable data. Independence of the latent failure times is one assumption that would resolve the problem of identifiability and permit estimation of net probabilities; however, this assumption can never be verified. This problem of nonidentifiability was pointed out by Cox [13] and Tsiatis [31].

To get a sense of the extent of the nonidentifiability problem, Peterson [26] computed sharp bounds for net survival probabilities as a function of crude subdistribution functions. Specifically, he showed that

$$S^C(x) \leq S_i^i(x) \leq 1 - F_i^C(x).$$

Heuristically, these inequalities may be explained as follows. First, consider the hypothetical case that the causes of death are so highly correlated that an individual dying at time x from any cause other than D_i would have died from cause D_i immediately thereafter. For such a scenario the net survival probability at time x, $S_i^i(x)$, would be equal to the probability of surviving until time x from any cause, $S^C(x) = \Pr(X > x)$. At the other extreme, consider the hypothetical case where an individual who would die from any cause other than D_i would never die from cause D_i. Here, $\Pr(T_i \leq x) = 1 - S_i^i(x)$ would be equal to $F_i^C(x) = \Pr(X \leq x, \Delta = i)$. The upper and lower bounds for net survival probabilities may be quite substantial, as shown by Tsiatis [32].

This creates a philosophical dilemma in competing risks theory. Knowledge of the distribution of observable causes of death does not suffice to determine net survival probabilities. Only if additional assumptions are made on the joint distribution of the latent failure times are we able to identify uniquely the net survival probabilities. Two points of view have been taken in the literature. One is to restrict attention to certain dependency structures on the latent failure times that allow for identification or, at least, restrict to a class of joint distributions where the bounds for the net survival probability are much tighter than the Peterson bounds. This has been the focus of research by Slud & Rubinstein [29], Klein & Moeschberger [23], and Zheng & Klein [33].

Another perspective is as follows. Because nonidentifiability problems can only be handled by making additional assumptions that cannot be verified from the data, perhaps we should only consider making **inference** on the distribution of the observable random variables. That is, the focus should be on estimating cause-specific hazard and subdistribution functions and the comparison of such quantities under a variety of conditions that have practical importance. For example, comparisons may be made among different treatments or varying environmental conditions. This pragmatic point of view suggests that there is no reason to consider hypothetical quantities, such as net survival probabilities, because in fact we will never be in a position to evaluate one cause of death acting in isolation on a population. This point of view was eloquently presented by Prentice et al. [27].

This idea may be modified slightly in the case where a subset of the causes of death that are not of primary interest, denoted by $\overline{J}$, are thought a priori to be independent of the other causes of death, J, that are of interest. For example, certain accidental causes of death may fall into this category when studying treatment of disease. For these problems, inference using partial crude probabilities may be appropriate. We showed before how partial crude probabilities can be defined in terms of the distribution of the observable crude probabilities when causes J are independent of $\overline{J}$.

The Statistical Analysis of Competing Risk Data

Often, the data available for the analysis of competing risks are incomplete or right **censored**. This may be due to the termination of the study before all individuals fail, or to individuals who drop out of the study and subsequently are lost to follow-up. To accommodate this situation we extend the definition of competing risks to include censoring, i.e. we include an additional random variable, T_0, that denotes the latent time to censoring. With this extended definition of competing risks, the observable data are defined by

X^* and Δ^*, where $X^* = \min(T_0, \ldots, T_k)$ and $\Delta^* = i$ if $X^* = T_i$, $i = 0, \ldots, k$. We note that $\Delta^* = 0$ means that the failure time was censored at time X^*.

In a typical competing risks study we observe a sample of data (X_j^*, Δ_j^*, Z_j), $j = 1, \ldots, n$, where for the jth individual, X_j^* denotes the time to failure or censoring, Δ_j^* corresponds to cause of death or censoring, and Z_j corresponds to **covariate(s)** which we may use for modeling the distribution of competing risks. Using this extended notation, the observable data include censoring as a competing risk. We use an asterisk to denote the competing risks model that includes censoring. Therefore, the complete set of observable risks will be denoted by $C^* = 0, \ldots, k$, in contrast to the risks of interest, $C = 1, \ldots, k$, or perhaps some subset, J. In the previous section we denoted the complement of the subset J by $\overline{J} = C - J$; in the extended definition of competing risks we denote the complement of J by $\overline{J}^* = C^* - J$. In what follows it will be assumed that censoring, or risk 0, is independent of the other risks C. Without this assumption, nonidentifiability problems would not allow for **estimation** of the competing risk probabilities of interest regarding causes C.

One Sample Problems

Here we consider the problem of estimating relevant competing risk probabilities from a single sample of data (X_j^*, Δ_j^*), $j = 1, \ldots, n$.

Estimating Cause-Specific Hazard Functions

We showed before that the partial crude cause-specific hazard function is equal to the observable crude cause-specific hazard function whenever the risks in J are independent of the risks in $\overline{J}^*$, i.e.

$$\lambda_i^J(x) = \lambda_i^{C^*}(x). \tag{9}$$

Because censoring, or risk 0, is always assumed independent of the other risks, (9) will follow as long as the risks in J are independent of $\overline{J}$. It is important to note that the crude cause-specific hazard functions discussed in the previous section, $\lambda_i^C(x)$, are actually partial crude cause-specific hazard functions when we include censoring as a competing risk. However, because of (9) applied to $J = C$, $\lambda_i^C(x)$ is equal to the observable $\lambda_i^{C^*}(x)$. In the case when cause of death D_i is independent of the other risks, the net-specific hazard function, $\lambda_i^i(x)$, is equal to $\lambda_i^{C^*}(x)$.

For certain independence assumptions, the cause-specific hazard functions are related to the observable crude cause-specific hazard functions, which by (2) is equal to

$$\lambda_i^{C^*}(x) = \frac{\mathrm{d}F_i^{C^*}(x)}{\mathrm{d}x} \Big/ S^{C^*}(x),$$

where

$$F_i^{C^*}(x) = \Pr(X^* \le x, \Delta^* = i)$$

and

$$S^{C^*}(x) = \Pr(X^* > x).$$

The natural estimate for the crude subdistribution function is the empirical subdistribution function, i.e.

$$\hat{F}_i^{C^*}(x) = n^{-1} \sum_{j=1}^{n} I(X_j^* \le x, \Delta_j^* = i),$$

where $I(\cdot)$ denotes the indicator function. This estimate puts mass $1/n$ at each observed event time from cause i. Similarly,

$$\hat{S}^{C^*}(x) = n^{-1} \sum_{j=1}^{n} I(X_j^* > x)$$

puts mass $1/n$ at each event time.

Because crude cause-specific hazards are functions of the crude subdistribution probabilities, the obvious estimates are obtained by substituting the corresponding functions of the empirical subdistribution probabilities. For example, the estimate of the cumulative cause-specific hazard function is

$$\hat{\Lambda}_i^{C^*}(x) = \int_0^x \frac{\mathrm{d}\hat{F}_i^{C^*}(u)}{\hat{S}^{C^*}(u)} = \sum_{j=1}^{n} \frac{I(X_j^* \le x, \Delta_j^* = i)}{Y(X_j^*)},$$

where $Y(u) = \sum_{j=1}^{n} I(X_j^* > u)$ denotes the number of individuals in the sample who are at risk at time u, i.e. neither died nor were censored. This estimator is the so-called **Nelson–Aalen estimator**; see Aalen [1]. Aalen [2, 3] derived the theoretical **large-sample** properties, including consistency and asymptotic normality, using the theory of **counting processes**.

This estimator of the ith crude cause-specific cumulative hazard is the appropriate estimator for the

partial crude cause-specific cumulative hazard whenever the causes in J are independent of the causes in $\overline{J}^*$, i.e.

$$\hat{\Lambda}_i^{J^*}(x) = \hat{\Lambda}_i^{C^*}(x), \quad i \in J.$$

In the special case where cause i is assumed independent of all other causes, the ith net-specific cumulative hazard function, $\Lambda_i^i(x)$, is estimated by $\hat{\Lambda}_i^{C^*}(x)$. The ith net survival distribution, $S_i^i(x)$, is equal to $\exp[-\Lambda_i^i(x)]$. Therefore, a natural estimator is the exponentiated negative of the Nelson–Aalen estimator. This estimator is

$$\hat{S}_i^i(x) = \exp - \left[\sum_{j=1}^{n} \frac{I(X_j^* \leq x, \Delta_j^* = i)}{Y(X_j^*)} \right].$$

Noting that this is equal to

$$\prod_{j=1}^{n} \exp \left[\frac{-I(X_j^* \leq x, \Delta_j^* = i)}{Y(X_j^*)} \right]$$

and that

$$\exp\left[\frac{-1}{Y(u)}\right] \approx \left[\frac{1-1}{Y(u)}\right],$$

yields the approximation

$$\hat{S}_i^i(x) \approx \prod_{j=1}^{n} \left[\frac{1-1}{Y(X_j^*)} \right]^{I(X_j^* \leq x, \Delta_j^* = i)}.$$

This is the well known **Kaplan–Meier** [21], or product-limit, estimator. The asymptotic equivalence of the exponentiated Nelson–Aalen estimator and the Kaplan–Meier estimator, and the large-sample properties of these estimators, are given by Breslow & Crowley [11].

It is important to note that the Kaplan–Meier estimator, by construction, is a **consistent estimator** of the exponentiated cumulative crude cause-specific hazard function. That this corresponds to an estimator of the net survival distribution follows only when the net hazard function is equal to the crude cause-specific hazard, i.e. when cause i is independent of all the other causes, including censoring. Without this assumption, the Kaplan–Meier estimator of the ith net-specific survival distribution does not estimate any interesting or relevant probability.

If we consider death from any cause, i.e. $\Delta \in C$, then the estimate of the corresponding survival distribution, $S^C(x)$, from a sample of potentially censored data (X_j^*, Δ_j^*), $j = 1, \ldots, n$, follows from applying the same logic:

$$\hat{S}^C(x) = \prod_{j=1}^{n} \left[\frac{1-1}{Y(X_j^*)} \right]^{I(X_j^* \leq x, \Delta_j^* \in C)}.$$

This estimator for the survival distribution from any cause of death in the presence of censoring is the Kaplan–Meier estimator as originally presented in the seminal paper [21] in 1958. Failure is considered a death from any cause, and an incomplete observation is a censored observation. The estimator of the ith net survival function given above is also referred to as a Kaplan–Meier estimator, since it may be derived via the same formula, letting failure be death from cause i and an incomplete observation be death from any cause other than i or censoring.

Estimating Subdistribution Functions

We may use the above results to derive **nonparametric** estimators for crude and partial crude subdistribution functions. Using (2), the ith crude subdistribution function may be expressed as

$$F_i^C(x) = \int_0^x S^C(u)\lambda_i^C(u)\,\mathrm{d}u.$$

Because censoring is independent of the other causes of death, $\lambda_i^C(u) = \lambda_i^{C^*}(u)$. Therefore a natural estimator for the ith subdistribution function is given by

$$\hat{F}_i^C(x) = \int_0^x \hat{S}^C(u) \frac{\mathrm{d}\hat{F}_i^{C^*}(u)}{\hat{S}^{C^*}(u)},$$

where $\hat{S}^C(u)$ is the Kaplan–Meier estimator for the survival distribution of time to death from any cause.

The large-sample statistical properties of this estimator may be derived using the theory of counting processes. Details may be found in Aalen [2, 3], Fleming [16, 17], Benichou & Gail [6], and Andersen et al. [5] when using **cohort** data, and in Benichou & Gail [7] when using **population-based case–control data**.

The Relationship of Competing Risks to Covariates

Often, we are interested in studying the relationship of time to death from one or many causes to other

covariates. For example, we may be interested in the effect of different treatments on reducing the risk of death from specific causes, or we may wish to model the relationship of competing risk probabilities to other **prognostic factors**. These problems are generally posed in terms of hypothesis testing or estimation of **regression** parameters. There is a wide literature on inferential techniques for hypothesis testing and regression modeling for survival problems with censored data. Because of the close relationship between censoring and competing risks, many of the methods developed for analyzing censored survival data may also be applied to competing risks data (*see* **Survival Analysis, Overview**).

Hypothesis Testing

The most widely used methods for testing the **null hypothesis** of no treatment effect among K treatments with censored survival data are the **logrank** or weighted logrank tests. These tests were designed to test the equality of the hazard functions for death among K treatments when the censoring time is independent of time to death within each treatment group. If we study these tests carefully, then we realize that they actually compare the observable cause-specific hazard functions among the different treatment groups. Therefore, we can immediately apply these methods for testing equality of cause-specific hazard functions among different treatments. To be more precise, we denote by $\lambda_{il}^{C^*}(x)$, $l = 1, \ldots, K$, the ith cause-specific hazard function within treatment group l. The weighted logrank tests may then be used to test the null hypothesis that

$$\lambda_{i1}^{C^*}(x) = \cdots = \lambda_{iK}^{C^*}(x), \quad x > 0.$$

The theoretical development for these tests is given by Andersen et al. [4]. This is carried out by letting failure correspond to death from cause i and an incomplete observation to correspond to death from any cause other than i or censoring ($\Delta^* = 0$).

We reiterate the interpretation of this null hypothesis and the results of the logrank test. If we are willing to assume that time to death from cause i is independent of the times to death from other causes as well as time to censoring, within each treatment group $l = 1, \ldots, K$, then the cause-specific hazard function, $\lambda_{il}^{C^*}(x)$, is equal to the net-specific, or marginal, hazard function, $\lambda_{il}^{i}(x)$. Equality of net-specific hazard functions implies equality of net-specific survival probabilities. Therefore, with the assumption of independence, the logrank test is a test of the null hypothesis that the K net-specific survival distributions are equal. This is often the hypothesis of interest.

To illustrate, consider a **clinical trial** of several treatments to reduce breast cancer mortality. Because breast cancer clinical trials generally occur over many years, some patients may die from causes other than breast cancer. Because the treatments were targeted to reduce breast cancer mortality, the investigators are not interested in the effect that treatment may have on other causes of death; rather, they are mainly interested in the effect of the treatments on breast cancer mortality in the absence of causes of death other than breast cancer. This is the classic competing risks problem of comparing net survival distributions. When the logrank test is used, patients not dying from breast cancer are treated as censored observations. As previously discussed, this is an appropriate test for the equality of net survival probabilities when the time to death from other causes is independent of time to death from breast cancer within each treatment group. This assumption may not be true, and in fact cannot be verified with the data because of nonidentifiability problems alluded to above. If this independence assumption is not true, then it is not clear what we are testing when we use the logrank test.

One way around this philosophical dilemma is to consider only tests of observable population parameters. An important observable population parameter is the crude cause-specific hazard function, $\lambda_i^C(x)$. We again emphasize that the population cause-specific hazard function, $\lambda_i^C(x)$, is observable only if there is no additional censoring introduced. With the introduction of censoring, the observable parameter is $\lambda_i^{C^*}(x)$. However, by assumption, the censoring ($\Delta^* = 0$) is independent of the other causes of death, in which case $\lambda_i^C(x) = \lambda_i^{C^*}(x)$.

As we pointed out, the logrank test tests the equality of the cause-specific hazard functions, $\lambda_{il}^{C^*}(x)$, and, with independent censoring, the equality of $\lambda_{il}^{C}(x)$. Therefore, the logrank test would be a valid test of the equality of the breast cancer specific hazard functions among the K treatments. Although this cause-specific hazard function may not be directly related to net-specific breast cancer mortality, if independence does not hold, then it still may be an important comparison. This point of view is given by Prentice et al. [27].

Another observable quantity is the subdistribution function $F_{il}^C(x)$ for cause i within treatment group l. Very little work has been done on deriving tests for the equality of these K-sample subdistribution functions. One exception is a class of tests derived by Gray [20] to test the null hypothesis that

$$F_{il}^C(x) = \cdots = F_{iK}^C(x).$$

Regression Modeling

The most popular framework for modeling the association of censored survival data to prognostic variables is with the **proportional hazards** model of Cox [14] (*see* **Cox Regression Model**). In this model the hazard for death is related to a vector of covariates by

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp(\beta^{\mathrm{T}}\mathbf{z}),$$

where $\mathbf{z}$ represents a vector of covariates, and $\lambda(t|\mathbf{z})$ is the hazard rate of death at time t given covariates $\mathbf{z}$. In this model, censoring is assumed to be independent of the failure time, conditional on the covariates. A careful study of the inferential procedure for estimating parameters in the Cox model reveals that this is actually the observable cause-specific hazard of death in the presence of censoring. That this corresponds to the actual net hazard function of death holds only when we add the assumption of independence of censoring time and failure time.

Consequently, this model may also be applied to competing risks data; that is, we may use the same inferential procedures to estimate the parameter β when considering the model

$$\lambda_i^{C^*}(t|\mathbf{z}) = \lambda_{i0}^{C^*}(t)\exp(\beta^{\mathrm{T}}\mathbf{z}).$$

To apply software for the Cox model (*see* **Survival Analysis, Software**), we must define a failure as death from cause i, and an incomplete observation as either death from a cause other than i or censoring. The interpretation of this model and the parameters is the same as discussed above. That is, if we are willing to assume that time to death from cause i is independent of the times to death from other causes and time to censoring, then the observable cause-specific hazard, $\lambda_i^{C^*}(t|\mathbf{z})$, is equal to the net-specific hazard, $\lambda_i^i(t|\mathbf{z})$.

Even if we are unwilling to make this nonidentifiable assumption, the relationship of the observable cause-specific hazard to covariates may be of interest. By assumption, censoring is independent of all other causes of death. This implies that $\lambda_i^C(t|\mathbf{z}) = \lambda_i^{C^*}(t|\mathbf{z})$. Therefore, the results of the Cox regression analysis may be used to estimate the parameters in the model of the cause-specific hazard function, given by

$$\lambda_i^C(t|\mathbf{z}) = \lambda_{i0}^C(t)\exp(\beta^{\mathrm{T}}\mathbf{z}).$$

Using cause-specific hazards thus allows useful interpretation of relevant observable quantities without an additional assumption of independence of the different causes of death.

References

[1] Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models, *Scandinavian Journal of Statistics* **3**, 15–27.

[2] Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models, *Annals of Statistics* **6**, 534–545.

[3] Aalen, O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[4] Andersen, P.K., Borgan, O., Gill, R.D. & Kieding, N. (1982). Linear nonparametric tests for comparison of counting processes, with applications to censored survival data, *International Statistical Review* **50**, 219–258.

[5] Andersen, P.K., Borgan, O., Gill, R.D. & Kieding, N. (1992). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, pp. 299–301.

[6] Benichou, J. & Gail, M.H. (1990). Estimates of absolute risk in cohort studies, *Biometrics* **46**, 813–826.

[7] Benichou, J. & Gail, M.H. (1995). Methods of interferences for estimates of absolute risk derived from population-based case–control studies, *Biometrics* **51**, 182–194.

[8] Berkson, J. & Elveback, L. (1960). Competing exponential risks with particular reference to the study of smoking and lung cancer, *Journal of the American Statistical Association* **55**, 415–428.

[9] Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour le prévenir. *Historie avec les Mémoires, Académie Royal des Sciences*. Paris, pp. 1–45.

[10] Birnbaum, Z.W. (1979). *On the Mathematics of Competing Risks*. US Department of Health, Education and Welfare, Washington.

[11] Breslow, N. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics* **2**, 437–453.

[12] Chiang, C.L. (1968). *Introduction to Stochastic Processes in Biostatistics*. Wiley, New York, Chapter 11.

[13] Cox, D.R. (1959). The analysis of exponentially distributed life-times with two types of failure, *Journal of the Royal Statistical Society, Series B* **21**, 411–421.

[14] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[15] David, H.A. & Moeschberger, M.L. (1978). The Theory of Competing Risks, Griffin's Statistical Monograph No. 39. Macmillan, New York.

[16] Fleming, T.R. (1978). Nonparametric estimation for nonhomogeneous Markov processes in the problem of competing risks, *Annals of Statistics* **6**, 1057–1070.

[17] Fleming, T.R. (1978). Asymptotic distribution results in competing risks estimation, *Annals of Statistics* **6**, 1071–1079.

[18] Gail, M. (1975). A review and critique of some models used in competing risk analysis, *Biometrics* **31**, 209–222.

[19] Gail, M. (1982). Competing risks, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 75–81.

[20] Gray, R.J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk, *Annals of Statistics* **16**, 1141–1154.

[21] Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[22] Karn, M.N. (1931). An inquiry into various death rates and the comparative influence of certain diseases on the duration of life, *Annals of Eugenics* **4**, 279–326.

[23] Klein, J.P. & Moeschberger, M.L. (1988). Bounds on net survival probabilities for dependent competing risks, *Biometrics* **44**, 529–538.

[24] Makeham, W.M. (1874). On the law of mortality, *Journal of the Institute of Actuaries* **18**, 317–322.

[25] Moeschberger, M.L. & David, H.A. (1971). Life tests under competing causes of failure and the theory of competing risks, *Biometrics* **27**, 909–933.

[26] Peterson, A.V. (1976). Bounds for a joint distribution function with fixed subdistribution functions: applications to competing risks, *Proceedings of the National Academy of Sciences* **73**, 11–13.

[27] Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Flournoy, N., Farewell, V.T. & Breslow, N.E. (1978). The analysis of failure time data in the presence of competing risks, *Biometrics* **34**, 541–554.

[28] Sampford, M.R. (1952). The estimation of response time distributions. II. Multistimulus distributions, *Biometrics* **8**, 307–353.

[29] Slud, E.V. & Rubinstein, L.V. (1983). Dependent competing risks and summary survival curves, *Biometrika* **70**, 643–650.

[30] Todhunter, J. (1949). *A History of the Mathematical Theory of Probability*. Chelsea, New York.

[31] Tsiatis, A.A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences* **72**, 20–22.

[32] Tsiatis, A.A. (1978). An example of nonidentifiability in competing risks, *Scandinavian Actuarial Journal* **1978**, 235–239.

[33] Zheng, M. & Klein, J.P. (1994). A self-consistent estimator of marginal survival functions based on dependent competing risk data and the assumed copula, *Communications in Statistics – Theory and Methods* **23**, 2299–2311.

A.A. TSIATIS

Complex Diseases

The principal distinction between "simple" monogenic diseases and "complex" genetic diseases is that the latter do not exhibit classic Mendelian patterns of inheritance (*see* **Mendel's Laws**) and characteristically involve multiple **genes** that interact in complex ways with multiple environmental factors [5] (*see* **Gene-environment Interaction**). For instance, asthma and many of the traits associated with asthma exhibit non-Mendelian patterns of inheritance and substantial heterogeneity [6]. The intricacy of the disparate pathogenic mechanisms associated with asthma suggests multiple environmental and genetic determinants, and has led to the definition of asthma as a "complex" phenotype [4].

Other examples of complex human diseases include type 1 and type 2 diabetes, Alzheimer's disease, rheumatoid arthritis, and many cancers and psychiatric disorders. Such diseases tend to be common relative to single-gene disorders, to be chronic conditions that are responsible for significant morbidity and mortality, and to be associated with very substantial economic costs. For example, asthma is the most common chronic childhood disease in developed nations [1], and carries a very substantial direct and indirect economic cost worldwide [8]. As a result of these factors, such diseases have become a major focus of bioscience research in both industry and academia.

Complex diseases tend to involve many difficulties in phenotypic definition and are not generally amenable to investigation using techniques that assume monogenic inheritance. Relative to monogenic disorders, the study of genetically complex diseases presents many additional challenges for **genetic epidemiology**. For instance, classic **segregation analysis** was designed for studies of monogenic diseases and assumes etiologic homogeneity – an assumption unlikely to be met in analyses of complex phenotypes, which are likely to be under the control of multiple environmental and genetic factors. Different genes may segregate in different families, which may in turn be exposed to different environmental factors. The mapping of human susceptibility loci for complex disease is further made difficult by a high population frequency, incomplete **penetrance**, phenocopies (environmentally determined, nongenetic variation in a phenotype that resembles genetically determined variation), **genetic heterogeneity**, and pleiotropy. The substantial genetic heterogeneity that characterizes such diseases is likely to involve both different susceptibility loci and different susceptibility alleles within a locus. Strategies to minimize the effects of genetic heterogeneity in studies of complex diseases have included the use of large pedigrees, genetically isolated populations likely to exhibit **founder effects** (*see* **Inbreeding**), and phenotypically homogeneous subgroups such as early-onset forms of disease [7].

A further characteristic of many complex human diseases is that they are closely associated with one or more "intermediate" phenotypes, e.g. hyperlipidemia in type 2 diabetes [13] and measures of lung function in asthma [16]. Intermediate phenotypes are often quantitative, pathophysiological traits assumed – on the basis of prior clinical, epidemiologic, or laboratory evidence – to be on an etiologic pathway leading to disease. Due to difficulties in phenotype definition and related low statistical power, and the complexity of the interrelationships between clinical disease, intermediate phenotypes and epidemiologic covariates, dichotomous disease affection as an outcome has proven difficult to dissect genetically in many complex diseases. Disease-associated intermediate phenotypes are often themselves highly **heritable**, and power calculations have suggested that testing for **linkage** is significantly enhanced by the use of quantitative traits in preference to a categoric affection phenotype [2, 15]. The use of quantitative intermediate phenotypes also permits selection of subjects in the extremes of the distribution, with increased power to detect linkage [17]. For these reasons, objectively-measured intermediate phenotypes have increasingly become the focus of genetic studies of complex human diseases.

In an attempt to simplify etiologic heterogeneity, many researchers investigating the genetics of complex human disease have turned to inbred animal models [10, 18]. Animals have the advantages of permitting control of both breeding and environmental conditions. Assuming a valid model, these advantages potentially allow complex genetic traits to be more easily dissected using animal models than using human study populations. The extent to which results obtained from experimental animals can be generalized to humans may be problematic, however.

As part of the intense research effort to improve our ability to discover the genetic determinants of complex human disease over the last decade, technologic advances in the laboratory related to sequencing and single nucleotide **polymorphism** (SNP) genotyping have proceeded at a very rapid rate (*see* **Bioinformatics**; **Genetic Markers**). Catalyzed in part by the vast amounts of data generated by the **Human Genome Project** and the SNP genotyping efforts in complex human disease, it has become clear that concomitant statistical advances in the mapping of complex traits will also be required [11, 22, 25] (*see* **Disease-marker Association**; **Linkage Analysis, Model-based**; **Linkage Disequilibrium**). The Human Genome Project and SNP genotyping efforts have caused a broad reexamination of mapping methodologies and study designs in complex human disease [7, 17, 20, 23]. The testing of large numbers of genotypes for association with one or more traits raises important statistical issues regarding the appropriate false positive rate of the tests and the level of statistical significance to be adopted given the multiple testing involved [17, 20]. The required methodologic development in genetic statistics is nontrivial given the complexity of most common human diseases. Some current areas of methodologic development include **haplotype analysis** [9, 22, 25], distance-based mapping measures [3, 19], combined linkage and association analyses [12], techniques for modeling linkage disequilibrium and population history [26] (*see* **Population Genetics**) and **Markov Chain Monte Carlo** based approaches [14].

Genetic approaches to complex diseases offer great potential to improve our understanding of their etiology, but they also offer significant challenges. Despite much progress in defining the genetic basis of diseases such as asthma in the last decade, accompanied by rapid technical progress in sequencing and SNP genotyping technologies (*see* **Sequence Analysis**), further research is required. In particular, genetic localization of most susceptibility loci is still insufficiently precise for the positional cloning of new genes influencing such diseases. Furthermore, there are technical, statistical, ethical, and psychosocial issues that remain unresolved in the investigation of the genetics of complex human diseases. However, a large number of groups are currently active in addressing methodologic problems in genetic statistics, and methodologic progress, together with technologic advances in positional cloning and candidate loci linkage-disequilibrium mapping techniques will likely accelerate our ability to investigate the genetic basis of complex diseases.

References

[1] Asher, M.I., Keil, U., Anderson, H.R., Beasley, R., Crane, J., Martinez, F., Mitchell, E.A., Pearce, N., Sibbald, B., Stewart, A.W., Strachan, D., Weiland, S.K., & Williams, H.C. (1995). International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods, *European Respiratory Journal* **8**, 483–491.

[2] Allison, D.B. & Schork, N.J. (1997). Selected methodological issues in meiotic mapping of obesity genes in humans: issues of power and efficiency, *Behavioral Genetics* **27**, 401–421.

[3] Collins, A. & Morton, N.E. (1998). Mapping a disease locus by allelic association, *Proceedings of the National Academy of Sciences* **95**, 1741–1745.

[4] Elston, R. (1993). Genetic analysis of complex phenotypes: family studies of total IgE and bronchial hyperreactivity, in *The Genetics of Asthma*, D. Marsh, A. Lockhart & S. Holgate, eds. Blackwell Scientific Publications, Oxford, Chapter 12.

[5] Elston, R. (1995). The genetic dissection of multifactorial traits, *Clinical and Experimental Allergy* **2**, 103–106.

[6] Hopkins, J. (1995). Genetics of atopy, *Pediatric Allergy and Immunology* **6**, 139–144.

[7] Lander, E. & Schork, N. (1994). Genetic dissection of complex traits, *Science* **265**, 2037–2048.

[8] Lenney, W. (1997). The burden of pediatric asthma, *Pediatric Pulmonology Supplement* **15**, 13–16.

[9] Li, T., Ball, D., Zhao, J., Murray, R.M., Liu, X., Sham, P.C. & Collier, D.A. (2000). Family-based linkage disequilibrium mapping using SNP marker haplotypes: application to a potential locus for schizophrenia at chromosome 22q11, *Molecular Psychiatry* **5**, 452.

[10] Linder, C.C. (2001). The influence of genetic background on spontaneous and genetically engineered mouse models of complex diseases, *Laboratory Animals (New York)* **30**, 34–39.

[11] Long, A.D. & Langley, C.H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits, *Genome Research* **9**, 720–731.

[12] MacLean, C.J., Morton, N.E. & Yee, S. (1984). Combined analysis of genetic segregation and linkage under an oligogenic model, *Computers and Biomedical Research* **17**, 471–480.

[13] Marklova, E. (2001). Genetic aspects of diabetes mellitus, *Acta Medica* **44**, 3–6.

[14] Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms, *Genetics* **154**, 931–942.

[15] Olson, J.M., Witte, J.S. & Elston, R.C. (1999). Genetic mapping of complex traits, *Statistics in Medicine* **18**, 2961–2981.

[16] Palmer, L.J., Burton, P.R., James, A.L., Musk, A.W. & Cookson, W.O. (2000). Familial aggregation and heritability of asthma-associated quantitative traits in a population-based sample of nuclear families, *European Journal of Human Genetics* **8**, 853–860.

[17] Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science* **273**, 1516–1517.

[18] Sibal, L.R. & Samson, K.J. (2001). Nonhuman primates: a critical role in current disease research, *ILAR Journal* **42**, 74–84.

[19] Terwilliger, J.D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci, *American Journal of Human Genetics* **56**, 777–787.

[20] Terwilliger, J.D. & Goring, H.H. (2000). Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design, *Human Biology* **72**, 63–132.

[21] Terwilliger, J.D. & Weiss, K.M. (1998). Linkage disequilibrium mapping of complex disease: fantasy or reality?, *Current Opinion in Biotechnology* **9**, 578–594.

[22] Toivonen, H.T., Onkamo, P., Vasko, K., Ollikainen, V., Sevon, P., Mannila, H., Herr, M. & Kere, J. (2000). Data mining applied to linkage disequilibrium mapping, *American Journal of Human Genetics* **67**, 133–145.

[23] Weeks, D. & Lathrop, G. (1995). Polygenic disease: methods for mapping complex disease traits, *Trends in Genetics* **11**, 513–519.

[24] Zhao, L.P., Aragaki, C., Hsu, L. & Quiaoit, F. (1998). Mapping of complex traits by single-nucleotide polymorphisms, *American Journal of Human Genetics* **63**, 225–240.

[25] Zollner, S. & von Haeseler, A. (2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms, *American Journal of Human Genetics* **66**, 615–628.

LYLE J. PALMER

Compliance and Survival Analysis

Compliance: Cause and Effect

Today, new treatments must prove their worth in comparative (double blind) randomized clinical trials, the gold standard design for causal inference (*see* **Clinical Trials, Overview**). With noninformatively right-**censored** survival outcomes, a typical robust **intention-to-treat analysis** compares groups as randomized using the popular (weighted) **logrank test**. Accompanying **Kaplan–Meier**, curves describe nonparametrically how survival chances differ between arms. A one-parameter summary of the contrast follows from a **semiparametric** Cox **proportional hazards** (PH) model (*see* **Cox Regression Model**) or **Accelerated Failure-Time Model** [6].

In general, and especially with long-term treatments, patients tend to deviate from their prescribed treatment regime. Varying patterns of observed exposure relative to the assigned are called "compliance (levels)" and recognized as a likely source of variation in treatment effect (*see* **Compliance Assessment in Clinical Trials**). Because deviations from prescribed regimes occur naturally in clinical practice, it is wise to learn about them within the trial context rather than restrict the study population to perfect compliers, an atypical and sometimes small and unobtainable subset of the future patient horizon [12].

Treatments that are stopped or switched or are less dramatic lapses in dosing happen in response to a given assignment. Different exposure patterns between treatment arms therefore point to (perceived) differences following alternative assignments. Studying compliance patterns as an outcome can yield valuable insights [15].

Of course, actual treatment regimes may also influence primary outcomes. From the intent-to-treat perspective, underdosing causes reduced **power** and a requirement for larger samples. Fortunately, the strong **null hypothesis**, where treatment and its assignment have no impact on outcome, is consistently tested irrespective of compliance levels. Under the alternative, we expect, however, different (smaller) intent-to-treat effects than the prescribed regime would create when it materializes. This happens as the treatment group becomes a mix of varying (lower) degrees of exposure [2, 8].

Estimation of the causal effect of actual dose timing becomes challenging, when observed exposure patterns are no longer randomized. (Un)measured patient characteristics and earlier experience may determine exposure levels that become confounded with the natural treatment-free hazard of the patient. The association between compliance which induces treatment exposure levels, and treatment-free hazards is often called a *selection effect* in line with **missing data** terminology. An "as-treated" analysis, such as a PH analysis, with the currently received treatment as a **time-dependent covariate**, compares hazards between differently treated groups at a given time and thus estimates a mix of selection and causal effects [11]. An "on-treatment" analysis censors patients as soon as they go off the assigned treatment and thus generates informative censoring for the same reason. Structural accelerated failure time (SAFT) models and structural PH models have been designed to avoid these biases. We explain their key features and potential through a simple example first.

All-or-nothing Compliance

In randomized studies that evaluate a one-shot treatment, such as surgery [7], vaccination [4], or an invitation to undergo screening [1], all-or-nothing compliance with experimental assignment arises naturally. Let $R_i = 1(0)$ indicate whether individual i, with baseline covariates $\mathbf{X}_i$, gets randomized to the experimental (control) arm. When experimental treatment is not available outside the treatment arm the control arm remains uncontaminated by experimental exposure and its outcomes can serve as reference outcomes for causal inference. Let T_{0i} denote such a potential survival time for individual i following a control trajectory free of experimental exposure. Let T_{1i} and C_i respectively be survival time and compliance following a possible treatment assignment.

R_i operates independently of the vector (T_{0i}, T_{1i}, C_i, $\mathbf{X}_i$) but determines, which components are observed. Observed survival time and exposure are simply denoted T_i, E_i for all. With an uncontaminated control arm, $E_i = C_i R_i$. One goal of causal inference is to estimate how the contrast between potential survival times T_{1i} and T_{0i} varies over the subpopulations determined by different C_i-levels and their induced level of experimental exposure on the treatment arm, E_i.

The sharp null hypothesis assumes it makes no difference what arm one is assigned to and hence:

$$T_{0i} \stackrel{d|\mathbf{X}_i}{=} T_{1i}, \quad (1)$$

where $\stackrel{d|\mathbf{X}_i}{=}$ indicates equality in distribution conditional on $\mathbf{X}_i$.

The most obvious violation of (1) occurs when (some) patients on the experimental arm become exposed and exposure alters survival chances. This is called a direct causal effect of exposure [10]. When an assignment influences survival through mechanisms of action operating independently from exposure levels, we have an indirect effect. Below, we consider a cancer clinical trial [7], where the experimental intervention consists of implanting an arterial device during surgical resection of metastases. A planned implant could lead to an operation scheduled earlier in the day and timing may create its own set of prognostic circumstances. In addition, the news that a planned implant did not happen could be depressing to the patient and diminish survival chances beyond what would have happened on the control arm. Both mechanisms can lead to an indirect (clinical) effect of exposure assignment. Double blind studies are carefully designed to avoid indirect effects, so they satisfy: $T_{0i} \stackrel{\{d|C_i=0,\,\mathbf{X}_i\}}{=} T_{1i}$ and hence, $P(T_{1i} > t|C_i = 0, R_i = 1, \mathbf{X}_i) = P(T_{0i} > t|C_i = 0, R_i = 0, \mathbf{X}_i)$, for all t. The contrast between $P(T_{1i} > t|C_i = e, R_i = 1, \mathbf{X}_i)$ and $P(T_{0i} > t|C_i = e, R_i = 0, \mathbf{X}_i)$ then represents the causal effect of exposure level e. In general, however, this combines direct and indirect effects of assignment in the population with compliance level e.

In what follows, we ignore $\mathbf{X}_i$ for simplicity, but stronger inference can be drawn when assumptions condition on $\mathbf{X}_i$. To estimate $P(T_{0i} > t|C_i = 1, R_i = 1)$, one can solve $P(T_{0i} > t|C_i = 1|R_i = 1)P(C_i = 1, \ R_i = 1) + P(T_{0i} > t|C_i = 0, \ R_i = 1)P(C_i = 0|R_i = 1) = (P(T_{0i} > t|R_i = 1) =)P(T_{0i} > t|R_i = 0)$ after substituting empirical means or (Kaplan–Meier) estimates for the other unknown terms. **Isotonic regression** can turn the pointwise estimates in a monotone survival curve. To evaluate the treatment effect among the exposed, one compares $\hat{P}(T_{1i} > t|C_i = 1, R_i = 1)$ with $\hat{P}(T_{0i} > t|C_i = 1, R_i = 1)$. The selective nature of exposure is seen by contrasting treatment-free survival probabilities for the exposed and nonexposed subpopulations: $\hat{P}(T_{0i} > t|C_i = 1, R_i = 1)$ and $\hat{P}(T_{0i} > t|C_i = 0, R_i = 1)$.

More General Exposure Patterns

A structural model parameterizes the shift in distribution from observed survival time T_i to a reference time, T_{ei} following a specific exposure regime e, in terms of observed (possibly time-dependent) exposures $\mathbf{E}_i$ and covariates $\mathbf{X}_i$. One can thus perform parameter-specific $(\mathbf{E}_i, \mathbf{X}_i)$-dependent back transformations of observed survival times (or distributions). The parameter value that solves estimating equations demanding equality of estimated T_{ei} distributions (conditional on baseline **covariates**) between arms is our point estimate.

The procedure is illustrated in Figure 1 for the SAFT model $T_i \exp\{-\beta_0 E_i\} \stackrel{d|R_i}{=} T_{0i}$ in our trial, where E_i indicates an actual implant of the arterial device. For time-dependent implants $E_i(t)$, we could have used the SAFT model $\int_0^{T_i} \exp(-\beta_0 E_i(u))\,\mathrm{d}u \stackrel{d|R_i}{=} T_{0i}$. For technical points concerning the specific treatment of censored data, we refer the reader to [5, 12]. The left-hand panel shows ITT Kaplan–Meier curves in the standard and intervention arm. In the right-hand panel, the survival curve for the standard arm is compared with KM-curves following the transformations $T_i \exp\{-\beta E_i\}$ with $\beta = -1.5$ and $\beta = -0.36$ on the intervention arm. Reducing treated failure times by the factor $\exp(-1.5)$ overcompensates for the observed harmful treatment effect as survival chances on the intervention arm are now higher than on the standard arm. This is confirmed by the logrank chi-squared value of 9.326, plotted in the middle panel. The survival curve corresponding to the point estimate $\hat{\beta} = -0.36$ ($\exp(\hat{\beta}) = 70\%$) is convincingly close to the observed survival in the standard arm. The middle panel reveals chi-squared values for a range of hypothesized structural parameter values. Those that do not lead to significantly different curves at the 5% level form the 95% **confidence interval** $[-1.07, 0.39]$ for β_0.

Other Structural Modeling Options

One can propose many different maps of the observed into the treatment-specific survival distributions. This may happen on a PH scale [7] or involve time-dependent measures of effect in the SAFT setting [12]. Estimation methods, which rely on the instrument of randomization, protect the α–level (*see* **Hypothesis Testing**) just like the intent-to-treat test,

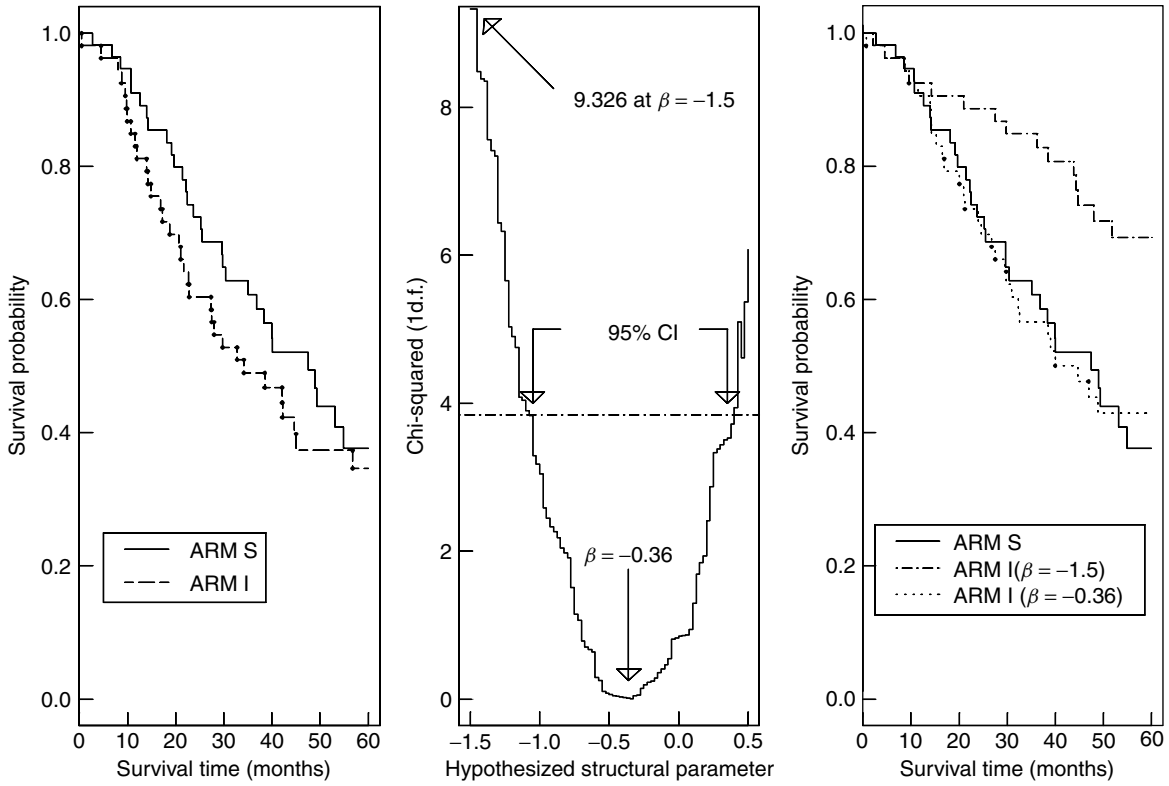

Figure 1 Estimation of structural parameters

but shift the point estimate away from a diluted average. To achieve this, they rely on the postulated structural model, which can sometimes be rejected by the data, but generally not confirmed owing to a lack of observed degrees of freedom. Special care is thus required when interpreting these models and their results. Some diagnostic procedures have been proposed (*see* **Diagnostics**) and forms of **sensitivity analyses** [13, 14].

To explicitly account for measured time-dependent confounders, **structural nested failure-time models** can be used as an alternative, or marginal structural models for T_{ei} as in [3]. The estimation process then relies on the assumption of "no residual confounding", ignores the instrument R_i, and loses its robust protection of the $\alpha-$level.

Structural modeling of failure time distributions has opened a world of practical and theoretical developments for the analysis of compliance and survival time. The field of research is very much alive today. Recent work [9], for instance, proposes to estimate optimal treatment regimes from compliance data. Our brief account can give but a flavor of this wealth.

Acknowledgment

We thank Tom Loeys for drawing the figure.

References

[1] Baker, S.G. (1999). Analysis of Survival data from a randomized trial with all-or-nothing compliance: estimating the cost-effectiveness of a cancer screening program, *Journal of the American Statistical Association* **94**, 929–934.

[2] Frangakis, C.E. & Rubin, D.B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes, *Biometrika* **80**, 365–379.

[3] Hernan, M.A., Brumback, B. & Robins, J.M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments, *Journal of the American Statistical Association* **96**, 440–448.

[4] Hirano, K., Imbens, G., Rubin, D. & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design, *Biostatistics* **1**, 69–88.

[5] Joffe, M.M. (2001). Administrative and artificial censoring in censored regression models, *Statistics in Medicine* **20**, 2287–2304.

[6] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. Wiley, New Jersey, Hoboken.

[7] Loeys, T. & Goetghebeur, E. (2003). A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance, *Biometrics* **59**, 100–105.

[8] Mark, S.D. & Robins, J.M. (1993). A method for the analysis of randomized trials with compliance information: an application to the multiple risk factor intervention trial, *Controlled Clinical Trials* **14**, 79–97.

[9] Murphy, S. (2003). Optimal dynamic treatment regimes, *Journal of the Royal Statistical Society, Series B* **65**, 331–355.

[10] Pearl, J. (2001). Causal inference in the health sciences: a conceptual introduction, *Health Services and Outcomes Research Methodology* **2**, 189–220.

[11] Robins, J.M. & Greenland, S. (1994). Adjusting for differential rates of PCP prophylaxis in high-versus low-dose azt treatment arms in an aids randomized trial, *Journal of the American Statistical Association* **89**, 737–749.

[12] Robins, J.M. & Tsiatis, A.A. (1991). Correcting for noncompliance in randomized trials using rank preserving structural failure time models, *Communications in Statistics, A* **20**, 2609–2631.

[13] Scharfstein, D., Robins, J.M., Eddings, W. & Rotnitzky, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints, *Biometrics* **57**, 404–413.

[14] White, I. & Goetghebeur, E. (1998). Clinical trials comparing two treatment arm policies: which aspects of the treatment policies make a difference? *Statistics in Medicine* **17**, 319–340.

[15] White, I. & Pocock, S. (1996). Statistical reporting of clinical trials with individual changes from allocated treatment, *Statistics in Medicine* **15**, 249–262.

(*See also* **Noncompliance, Adjustment for**; **Survival Analysis, Overview**)

ELS GOETGHEBEUR

Compliance Assessment in Clinical Trials

If the results of an **intention-to-treat analysis** in a randomized clinical trial are not statistically significant, then there are several possible explanations. In addition to the obvious explanations of lack of treatment effect and low statistical **power** due to inadequate sample size, treatment differences may be underestimated, and thus statistical power reduced, by poor compliance (sometimes called adherence) to the intervention on the part of the trial participants.

As a simple example, consider a randomized clinical trial in which participants are assigned by randomization to receive active treatment (A) or control (C). We assume that the response to treatment is a binary variable with expected value p_C in the control group and p_A in the active group. The treatment difference is $\delta = p_C - p_A$. If a proportion, π, of the participants assigned to active treatment do not comply with the therapy and we assume that noncompliers respond as control participants, then the observed treatment difference has expected value $\delta^* = p_C - (1-\pi)p_A - \pi p_C = (1-\pi)\delta$. Thus to maintain power, the sample size should be increased by a factor of $1/(1-\pi)^2$ (*see* **Sample Size Determination for Clinical Trials**).

Because poor participant compliance can adversely affect the outcome of a trial, it is important to use methods both to improve and monitor the level of compliance. The methods for improving compliance are largely applications of behavioral methods [4]. However, one widely used method for improving compliance is the use of a placebo run-in (see, for example, [12]). In a placebo run-in, potential participants in a trial are asked to take placebo pills (single masked) (*see* **Blinding or Masking**) for a short period of time. If the participant takes the pills as instructed, then he or she is entered into the trial. However, if the participant does not comply with the instructions on pill taking, then he or she is excluded from the trial. The assumption underlying the placebo run-in is that participants who do not comply in short term run-ins are more likely not to comply to long-term therapy. Two reports [2, 13] have challenged this assumption using either an empirical test or meta-analytic methods.

The most frequently used measure of compliance is the pill count. Participants in a clinical trial are given a specific number of doses of treatment and asked to return unused pills at the time of their next visit to the clinic. A simple calculation determines the proportion of medication unused and, by inference, the amount of medication taken. This measure has obvious problems related to the assumption that the pills not returned have been consumed. For example, a participant may have discarded some pills in anticipation of the clinic visit or pills may have been lost. Thus it is likely that pill counts overestimate compliance [8].

An alternative measure of compliance can be obtained by measuring the level of the agent in body fluids, such as blood or urine. For example, serum and urine measures of zidovudine levels are used in studies of zidovudine in HIV infected patients [10]. In this study, the association between pill count and zidovudine levels was good. This direct measure is superior to pill counts, but the cost of assays may be prohibitive in many settings.

In some lifestyle interventions, such as dietary change, such measures may be the most objective measure available for measuring compliance. Nevertheless, the correlation between the body fluid measure and compliance to the intervention may be low at the individual level, and the measure may be useful only at the group level. An example is the use of serum cholesterol levels to measure compliance with a low-fat dietary intervention.

An alternative to measuring actual blood levels is to add riboflavin to the compound. Urine samples are then collected from participants and the presence of riboflavin in the urine is measured by fluorescence. A major drawback to this method is the possible interaction of the riboflavin with the pharmacological action of the medication. Tests for the possibility of this interaction are too complex and expensive in most cases, rendering the method of little practical utility.

In recent years microprocessors have been used to monitor compliance. The microprocessor is installed in the bottle cap and programmed to record the dates and times of opening of the bottle [3, 5, 7–9]. A similar timing device has been applied in a study using an inhaler for delivery of the drug [1]. The basic assumption for the use of these devices is that the medication is taken at the time the bottle is opened. It is difficult to verify this assumption

in practice, but the method does appear to be more valid than the pill count. These devices have an added advantage over pill counts in that they provide data on compliance with the timing of the medication. Comparison of the electronic measure with pill counts indicates that compliance is estimated to be greater with the pill count than with the electronic devices [3, 11]. The electronic devices have also shown that noncompliance with the timing of doses is a sizable problem [1, 5, 6]. The major drawback to the use of the electronic compliance measures in a large clinical trial is cost.

References

[1] Coutts, J.A.P., Gibson, N.A. & Paton, J.Y. (1992). Measuring compliance with inhaled medication in asthma, *Archives of Disease in Childhood* **67**, 332–333.

[2] Davis, C.E., Applegate, W.B., Gordon, D.J., Curtis, C. & McCormick, M. (1995). An empirical evaluation of the placebo run-in, *Controlled Clinical Trials* **16**, 41–50.

[3] Dunbar-Jackob, J., Burke, L.E., Rohay, J.M., Sereika, S., Schlenk, E.A., Lipello, A. & Muldoon, M.F. (1996). Comparability of self-report, pill count, and electronically monitored adherence data, *Controlled Clinical Trials* **17**(Supplement 2), 80S–81S.

[4] Gorkin, L., Goldsten, M.G., Follick, M.J. & Lefevre, R.C. (1990). Strategies for enhancing adherence in clinical trials, in *The Handbook of Health Behavior Change*, S.A. Shumaker, E.B. Schron & J.K. Ockene, eds. Springer-Verlag, New York, pp. 361–375.

[5] Kruse, W., Eggert-Kruse, W., Rampmaier, J., Runnebaum, B. & Weber, E. (1993). Compliance and adverse drug reaction: a prospective study with ethinylestradiol using continuous compliance monitoring, *Clinical Pharmacology* **71**, 483–487.

[6] Mengden, T., Binswanger, B., Spuhler, T., Weisser, B. & Vetter, W. (1993). The use of self-measured blood pressure determinations in assessing dynamics of drug compliance in a study with amlodipine once a day, morning versus evening, *Journal of Hypertension* **11**, 1403–1411.

[7] Rohay, J.M., Dunbar-Jacob, J., Sereika, S., Kwoh, K. & Burke, L.E. (1996). The impact of method of calculation of electronically monitored adherence data, *Controlled Clinical Trials* **17**, 82S–83S.

[8] Rudd, P., Byyny, R.L., Zachary, V., Lo Verde, M.E., Mitchell, W.D., Titus, C. & Marshall, G. (1988). Pill count measures of compliance in a drug trial: variability and suitability, *American Journal of Hypertension* **1**, 309–312.

[9] Rudd, P., Ahmed, S., Zachary, V., Barton, C. & Bonduelle, D. (1990). Improved compliance measures: applications in an ambulatory hypertensive drug trial, *Clinical Pharmacology and Therapeutics* **48**, 676–685.

[10] Samet, J.H., Libman, H., Steger, K., Dhawan, R.K., Chen, J., Shevitz, A.H., Dewees-Dunk, R., Levenson, S., Kufe, D. & Craven, D.E. (1992). Compliance with Zidovudine therapy in patients infected with human immunodeficiency virus, type I: a cross-sectional study in a municipal hospital clinic, *American Journal of Medicine* **92**, 495–502.

[11] Schlenk, E.A., Dunbar-Jacob, J.M. & Rohay, J.M. (1996). Concordance of medication adherence measures in primary Raynaud's disease, *Controlled Clinical Trials* **17**, 123S.

[12] The Steering Committee of the Physicians' Health Study Research Group (1988). Findings from the aspirin component of the ongoing Physicians' Health Study, *New England Journal of Medicine* **318**, 262–264.

[13] Trivedi, M.H. & Rush, H. (1994). Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications?, *Neuropsychopharmacology* **11**, 33–43.

(*See also* **Noncompliance, Adjustment for**)

C.E. DAVIS

Composite Estimators

A composite estimator is a weighted combination of two (or more) component estimators. It is important in **sample surveys** because, when appropriate weights are used, its **mean square error** is smaller than that of either component estimator. This decrease in mean square error can be considerable when the two component estimators are independent and their mean square errors are of similar size.

Studies on the use of composite estimators in specific applications often explore different component estimators and different approaches to weighting them in constructing the composite estimator. In a given application, the sample design and available auxiliary data restrict the choice of component estimators and, although, there are a number of approaches to the selection of the composite estimator weight, the mean square error of the resulting composite estimator is often **robust** to modest deviations from the optimum weight.

Early applications of composite estimators appear in the survey sampling literature in conjunction with rotational designs in which some units provide data for more than one time period (see, for example, [18] and [19]). In such situations, two **unbiased** component estimators are available: a standard estimator using data only from the time period of interest, and an estimator that adjusts the estimate for the previous time period forward to the current time using data from units surveyed in both time periods. In these applications, the sample is often designed with a composite estimator in mind.

The majority of the literature on composite estimators addresses a second type of application. When data from a sample survey become available, invariably there is demand for estimates from domains whose sample sizes were not controlled in the design. Generally, these sample sizes are too small to make reliable estimates using standard direct estimators, and indirect estimators are considered. Indirect estimators "borrow strength" through models that link the domain and/or time period of interest to others so that the small sample size in the domain of interest is supplemented by sample observations from other domains and/or times. Although domains defined by geographic boundaries have received, by far, the most attention, these methods are applicable to any arbitrary domain. For example, a synthetic estimator was used to produce selected health statistics for states in an early application of indirect estimators at the US **National Center for Health Statistics** [9, 11]. Its use is justified under a model relating units across domains within poststrata (*see* **Poststratification in Survey Sampling**) that are defined on variables correlated with the one of interest.

A composite estimator was considered in this early application but not implemented because of difficulties in defining a weighting scheme. Further research helped specify weights [15], after which a composite estimator combining the unbiased, high-variance direct estimator with the biased, low-variance indirect estimator was implemented [12]. In this composite estimator, the first component estimator was a standard direct estimator that incorporated observations on the variable of interest only from the state for which the estimate was being made. The second was a synthetic estimator that used observations on the variable of interest from the region comprising the state within poststrata defined by age, race, sex, and other related variables. The weights for the two-component estimators were chosen to minimize the mean square error of the composite estimator and were approximated empirically.

In another early application, the US Bureau of the Census used a composite estimator, composed of a direct sample estimator and an indirect regression estimator (*see* **Ratio and Regression Estimates**), to produce state estimates of median income for four-person families [4]. In yet another application, the US Department of Agriculture used a composite estimator to produce county estimates of livestock inventories, crop production, and acreage planted in selected crops [8]. In addition, a number of other government statistical agencies have considered composite estimators for the production of subnational estimates (see, for example, [1], [3], [10], and [16]).

Theoretical considerations have guided the applications of composite estimators; but, not surprisingly, different theoretical approaches lead to different component estimators and weighting schemes. In the health statistics application described above, the component estimators were treated as given and the problem was approached as a simple one of determining weights to minimize the mean square error of the composite estimator [14].

There are a number of model-based theoretical approaches to the domain estimation problem, in general, and to the derivation of composite estimators,

specifically. One such approach is provided by Royall [13] within the framework of prediction theory, a model-based approach to finite population sampling. **Regression** models relating the variable of interest to auxiliary variables lead to best linear unbiased predictors (*see* **Minimum Variance Unbiased (MVU) Estimator**) of the specified finite population quantity. Models with domain-specific parameters produce direct estimators, whereas models with parameters common across domains produce indirect estimators. In certain models, composite estimators result when a **correlation** among units within poststrata is introduced, with both component estimators and the associated weights specified by the theoretical development.

Whereas the prediction approach uses models at the unit level within domains, nested error linear models (*see* **Multilevel Models**) [5], **Bayes** [17], **empirical Bayes** [2], and hierarchical Bayes [6] methods all incorporate models that are specified at the domain level. These approaches address the problem of estimating, for example, a population mean for each of a number of domains and provide estimators that, under the model, minimize the average squared error over all domains. In fact, such approaches predominate in theoretical investigations of small domain estimation problems (see [7] for a review with discussion).

References

[1] Drew, J.D., Singh, M.P. & Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian labour force survey, *Survey Methodology* **8**, 17–47.

[2] Effron, B. & Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach, *Journal of the American Statistical Association* **68**, 117–130.

[3] Falorsi, P.D., Falorsi, S. & Russo, A. (1994). Empirical comparison of small area estimation methods for the Italian labour force survey, *Survey Methodology* **20**, 171–176.

[4] Fay, R.E. & Herriot, R.A. (1979). Estimates of income for small places: an application of James–Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269–277.

[5] Fuller, W.A. & Harter, R.M. (1987). The multivariate components of variance model for small area estimation, in *Small Area Statistics*, R. Platek, J.N.K. Rao, C.E. Sarndal, & M.P. Singh, eds. Wiley, New York, pp. 103–123.

[6] Ghosh, M. & Lahiri, P. (1992). A hierarchical Bayes approach to small area estimation with auxiliary information, in *Bayesian Analysis in Statistics and Econometrics*, P.K. Goel & N.S. Iyengar, eds. Springer-Verlag, Berlin, pp. 107–125.

[7] Ghosh, M. & Rao, J.N.K. (1994). Small area estimation: an appraisal, *Statistical Science* **9**, 55–93.

[8] Iwig, W.C. (1996). The National Agricultural Statistics Service County Estimates Program, in *Indirect Estimators in U.S. Federal Programs*, W.L. Schaible, ed. Springer-Verlag, New York, pp. 129–144.

[9] Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates, in *American Statistical Association 1971 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 328–331.

[10] Lundström, S. (1987). An evaluation of small area estimation methods: the case of estimating the number of nonmarried cohabiting persons in Swedish municipalities, in *Small Area Statistics*, R. Platek, J.N.K. Rao, C.E. Sarndal & M.P. Singh, eds. Wiley, New York, pp. 239–253.

[11] National Center for Health Statistics (1969). *Synthetic State Estimates of Disability*, PHS Publication No. 1759. US Government Printing Office, Washington.

[12] National Center for Health Statistics (1978). *State Estimates of Disability and Utilization of Medical Services, United States, 1974–76*, *DHEW Publication* No. (PHS) 78–1241. US Government Printing Office, Washington.

[13] Royal, R.A. (1979). Prediction models in small area estimation, in *Synthetic Estimates for Small Areas*, *National Institute on Drug Abuse, Research Monograph* 24. US Government Printing Office, Washington, DC, pp. 63–83.

[14] Schaible, W.L. (1978). Choosing weights for composite estimators for small area statistics, in *American Statistical Association 1978 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 741–746.

[15] Schaible, W.L., Brock, D.B. & Schnack, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics, in *American Statistical Association 1977 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 1017–1021.

[16] Spjøtvoll, E. & Thomsen, I. (1987). Application of some empirical Bayes methods to small area statistics, in *Proceedings of the 46th Session of the International Statistical Institute*, Vol. 52, Book 4. Voorburg, Netherlands, pp. 435–450.

[17] Stroud, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation, in *Small Area Statistics*, R. Platek, J.N.K. Rao, C.E. Särndal & M.P. Singh, eds. Wiley, New York, pp. 124–137.

[18] Woodruff, R.S. (1963). A class of ratio composite estimators, *Journal of the American Statistical Association* **58**, 454–467.

[19] Woodruff, R.S. (1966). Use of a regression technique to produce area breakdowns of the monthly national estimates of retail trade, *Journal of the American Statistical Association* **79**, 496–504.

WES SCHAIBLE

Computer Algebra

Conventional scientific **computer languages** deal primarily with the manipulation of fixed-length integers and fixed-precision **floating-point** numbers. *Computer Algebra* packages permit one to program using mathematical expressions as well, so that it is straightforward to perform such tasks (*see* **Numerical Analysis**) as:

- arbitrary-precision arithmetic,
- polynomial factorization,
- differentiation of complicated functions,
- integration of wide classes of functions.

This article begins by presenting basic common features, and by listing and giving references for current commercial computer algebra packages. There follow two simple introductory examples of the use of computer algebra in statistical contexts, and then a discussion of some typical constraints and features of computer algebra systems: it is important to be aware of these, since in this area, unrealistic expectations lead all too rapidly to frustration and disappointment. Finally, we list a few representative examples of research uses of computer algebra in statistical science, and conclude with suggestions for further reading.

Basic Common Features of Computer Algebra Systems

What might a computer algebra system have to offer to statistical users? A typical feature list includes:

- a user-interface allowing input and manipulation of mathematical formula, so that one's activity on the computer connects very directly to the underlying mathematics;
- multiple-precision arithmetic (invaluable when one has to check whether a very small value is actually positive or negative!);
- the ability to program the computer algebra system to perform routine tedious formula-manipulation tasks (e.g. computation of the **matrix** of second partial derivatives of a log-**likelihood** function);
- (for most computer algebra packages) integrated graphical and numerical facilities, thus mixing the benefits of computer algebra and more conventional computing (that is to say, based on floating point, rather than symbolic, calculations);
- almost all computer algebra systems are interactive, allowing one to experiment and to explore. This is particularly powerful when combined with numerical and graphical facilities.

Thus, a typical computer algebra system can support all kinds of mathematical calculation, particularly including exact formula-based calculations. This versatility makes computer algebra a most useful resource.

Currently Available Computer Algebra Systems

At the time of writing, commonly used commercial computer algebra systems included *Maple*, *Mathematica*and *REDUCE*. There is little to choose between them in terms of their *basic* computer algebra capabilities, though more advanced users will discover pronounced differences in underlying design philosophies. *Mathematica* and *Maple* both possess hypertext help facilities and have sophisticated "notebook" or "worksheet" front ends, which allow presentation of symbolic, graphical, and numerical results mixed with formatted text in cells in a scrollable display and permit online recalculation and modification of the results. In this as in other features, *REDUCE* takes a minimalist approach (reflecting a design philosophy which emphasizes ease of portability to many platforms).

The first edition of this article also mentioned *AXIOM* and *MACSYMA*. Sadly the innovative *AXIOM* [11] is no longer supported commercially (though readers may care to visit the website

`http://www.aldor.org/`

for the associated compiler `Aldor`). *MACSYMA* [10] is one of the very earliest computer algebra systems, and is still available in a variety of implementations both commercial and noncommercial.

There also exist largely nonprogrammable computer algebra systems such as `Derive` (though this particular system is being continuously enhanced and now offers programming capability), and free computer algebra software such as `CoCoA` (a system for calculations in algebra). However, we limit ourselves here mainly to discussion relating to the commercial systems mentioned above. Table 1 gives references

Table 1 Computer algebra systems: some World Wide Web and literature references
The World Wide Web is a volatile entity: links may change from those given here!

SYSTEM	WWW reference	References
Maple	`http://www.maplesoft.com`	[9]
See also `http://www.mapleapps.com/List.asp?CategoryID=36&Category=Statistics`		
Mathematica	`http://www.wolfram.com`	[14], [21]
See also `http://www.wolfram.com/solutions/statistics/books.html`		
REDUCE	`http://www.uni-koeln.de/REDUCE/`	[15], [18]
`Derive`	`http://www.derive.com`	
`CoCoA`	`http://cocoa.dima.unige.it`	

(World Wide Web and literature) for some of these packages (*see* **Internet**).

Introductory Examples

Here are two simple introductory examples, illustrating the use of computer algebra in a broadly statistical context. A different computer algebra package is used for each example, and the meaning of each of the various code fragments is briefly discussed. It should be emphasized that at this level computer algebra packages are largely interchangeable: one could as well use one as the other on such basic problems as these.

The *first example* is an application to probability **generating functions** (pgf). Since pgf's are used to translate basic operations on **random variables** into algebraic operations, they are natural candidates for computer algebra. We use *Mathematica* for this example.

Consider a (hypothetical) animal that hosts N parasites, where $\Pr[N = n] = \lambda^n(1-\lambda)$ for $n = 0, 1, \ldots,$ (and $0 < \lambda < 1$) so N is geometric. Independently each parasite gives birth to M daughter parasites, where $\mathbf{P}[M = m] = \mu^m e^{-\mu}/m!$ for $m = 0, 1, \ldots,$ (and $\mu > 0$) so M is **Poisson**. We know N has pgf $f(s) = (1-\lambda)/(1-\lambda s)$, while M has pgf $g(s) = \exp(-\mu(1-s))$. Then pgf theory tells us the total number T of descendants has pgf $h(s) = f(g(s))$ (*see* **Contagious Distributions**).

We may compute the mean by $\mathbf{E}[T] = h'(1)$, and the probabilities $\mathbf{P}[T = t] = (1/t!)[d^t h(s)/ds^t]_{s=0}$. These operations can be algebraically tedious, but are carried out without much labor if computer algebra is used. Define the functions f, g, h by

```
f[s_] := (1-lambda)/(1-lambda s)
g[s_] := Exp[-mu (1-s)]
h[s_] := f[g[s]]
```

(employing the usefully succinct *Mathematica* notation `f[s_] := ...` for defining a function of `s`), and can then compute the mean by

```
D[h[s],s] /. s->1
```

```
lambda mu
----------
1 - lambda
```

Here `D` carries out the differentiation, and `/.s->1` performs the substitution $s = 1$.

The probabilities can be computed similarly. First define the probability of there being a total of `i` parasites:

```
prob[i_] := Module[{s}, Simplify
       [D[h[s],{s,i}]/(i!) /. s->0] ]
```

This uses *Mathematica*'s `Module` command (because, purely for reasons of programming style, we want `s` to be a local variable here) and `Simplify`, because the output is lengthy enough even when simplified! Note also the iterated derivative form of `D` used here.

For output, we use an iterative `Do` loop and also *Mathematica*'s `TeXForm` facility for producing TEX output:

```
Do[ Print["p_",i," = ",TeXForm
          [prob[i]]], {i, 0, 4}]
```

We present only the last line of output:

$$p_4 = \frac{e^{\mu}\ (-1+\lambda)\ \lambda}{\times \left(e^{3\,\mu} + 11\, e^{2\,\mu}\,\lambda + 11\, e^{\mu}\,\lambda^2 + \lambda^3\right)\ \mu^4} \over 24\,(-e^{\mu}+\lambda)^5}. \quad (1)$$

(Notice that there is still cosmetic work to be done in beautifying this expression; for example, $(-1+\lambda) \to (\lambda - 1)$.)

This last command is easily altered to produce as many probabilities as might be required, and indeed

one can also alter the definitions of $f(s)$, $g(s)$ without difficulty. This demonstrates the flexibility of computer algebra for simple problems involving large quantities of calculation.

The *second example* is an application of computer algebra to the differential equations underlying deterministic **SIR epidemics** (*see* **Epidemic Models, Deterministic**). We use *Maple* for this example.

The underlying differential equations are given by

$$\begin{aligned} \frac{dx}{dt} &= -\alpha x y \\ \frac{dy}{dt} &= \alpha x y - y \\ \frac{dz}{dt} &= y \end{aligned} \tag{2}$$

This is a nonlinear system, so we do not expect to be able to solve it explicitly. However notice that the differential equations for dx/dz and dy/dz are linear, so we may hope to obtain a partial solution using the linear ODE solver `dsolve` of *Maple*. First define the differential equations:

```
SIR :=  diff( x(z), z ) + alpha *
                x(z),
        diff( y(z), z ) - alpha *
                x(z) + 1;
```

We then employ the solver, adding in initial-value conditions and the variables for the differential equation:

```
XYsoln := dsolve( {SIR, x(0) = 1,
        y(0) = n-1}, { x(z),
        y(z) }, 'laplace');
```

where ``laplace'' signifies that the solver `dsolve` is to use the method of Laplace transforms. The result is

```
XYsoln := {y(z) = n - z - exp(-
      alpha z), x(z) = exp(- alpha z)}
```

We can now substitute this back into the equation for dz/dt:

```
Zequation := diff(z(t),t) =
    subs(XYsoln, z = z(t), y(z));
                d
Zequation := ---- z(t) = n - z(t)
                dt
              - exp(-alpha z(t))
```

The resulting ODE cannot be solved for z, but we can obtain a series solution by altering the ``laplace'' directive to ``series'':

```
  Zseries := dsolve( {z(0)=0,
        Zequation }, z(t), 'series');
z(t) = (- 1 + n) t + (- 1/2 alpha +
  1/2 alpha n + 1/2 - 1/2 n) t^2 + O(t^3)
```

Here we have truncated the series to $O(t^3)$ simply to fit the output on the page. Without much further work, one can produce a higher-order series expansion, and use it to further expand the previous solution to x and y. (It is also possible to solve the differential equations numerically within *Maple*, but this leads us too far from our remit.)

Basic Considerations for the Use of Computer Algebra

When choosing which computer algebra system to learn and to use, the dominant consideration should be to find out what system is used by the nearest friendly expert. Readily available expert help is enormously important in getting the most from sophisticated mathematical software. There are differences between systems (further insight on this can be gained by reading some of the critiques and comparisons referenced below), but these are usually of less practical significance than the immediate availability of helpful advice. Note that these systems can all be augmented by writing programs in computer languages specific to the respective systems, and are being actively and vigorously extended by enthusiastic user communities, so that comparison of feature lists is not as relevant as might at first appear.

However, even the casual user, of whatever system, needs to be aware of typical features and constraints of computer algebra computation. Some of these arise out of the very nature of the tasks being performed. For example, the basic unit of computation in a computer algebra system is the representation of a mathematical expression as a list of symbols (for example, $1 + 2x + x^2$ as {+,1,{*,2,x},{^,x,2}}) (for this reason, a common synonym for the term "computer algebra" is *symbolic computation*). These lists can grow to extraordinary lengths in the course of a calculation! (Consider the result of dividing $1 - x^n$ by $1 - x$ when n is some large integer.) This contrasts with conventional scientific

computation, in which the basic unit of computation is a fixed-length bit-sequence representing an integer or floating-point number represented up to machine accuracy. In general computer algebra can be very memory-hungry, and it is strikingly easy to attempt manipulations which in their intermediate stages require huge amounts of memory even if the final result is concise ("intermediate expression swell"). In practice such problems can often be evaded by breaking the task down into more manageable subtasks: success in computer algebra therefore requires persistence and flexibility.

A related problem is that ill-considered computer algebra computation can produce a vast mass of uninformative output, for example, in statistical asymptotics. Of course this is a problem in more conventional computing too, but Hamming's remark "The purpose of computing is insight not numbers" applies with particular force to the practice of computer algebra.

There are other issues: square roots have to be handled with care (since the sign of the root is ambiguous); numerical stability when using numerical interfaces; possible software bugs (computer algebra systems are as prone to bugs as any other large computer program!); some tasks (such as definite integration) do not admit algorithmic solutions and so cannot be implemented in a completely satisfactory way.

Most of the above issues are magnified versions of problems that have to be confronted by anyone undertaking a large calculation: and if they cause more trouble for computer algebra then it is largely because the calculations tend to be larger in scale. In particular, the question of error-free computation is just as pressing for humans as for computer programs! Neither computer algebra nor any other technique can be a replacement for careful thought about a problem.

Examples of Computer Algebra in Statistical Science

To make the case for computer algebra as a useful tool, here is a list of some recent applications drawn from the statistical research literature.

Currie [4] describes the solution of several statistical **maximum likelihood** estimation problems using *Mathematica* in a conceptually simple way (compute the likelihood, then maximize it etc.). As the author points out, this is limited in scope because it makes no use of any special structure which might be present. Nevertheless, two of the statistical problems presented are themselves taken from the recent statistical literature.

This sort of application might equally be carried out using numerical software such as *MATLAB*, except that computer algebra allows one to use the computer to manipulate expressions algebraically (and thus exactly) before committing to floating-point computation. In fact, there are now several examples of symbiotic relationships between computer algebra packages and general purpose numerical analysis software: for example, *MATLAB* can be interfaced to *Maple*.

A more extensive example of *conceptually* simple computer algebra (although possessing formidable technical content) is described by Mannion [16], who uses *REDUCE* to solve a problem in geometric probability posed by Klee, namely, the expected value of the volume of a tetrahedron whose vertices are chosen independently and uniformly at random from within a parent tetrahedron of unit volume.

Use of more sophisticated computer algebra algorithms is discussed in the monograph of Pistone, Riccomagno, and Wynn [17], who consider the use of computational algebra in, for example, issues of **confounding** for statistical models arising in **experimental design**. Let M be a family of **polynomial regression** models $y = p(x) + \varepsilon$ parameterized by $p(x)$, which is the corresponding (multivariate) polynomial in $x = (x_1, \ldots, x_n)$. Statistical identifiability of M corresponds to the question of whether $p(x)$ and $q(x)$ from M agree on all design points x, and this is related in turn to questions of computational algebraic geometry, namely, the theory of Gröbner basis algorithms. Pistone and Wynn discuss the use of the Gröbner basis algorithms in *Maple* and CoCoA in this and other statistical contexts.

Expansions in statistical asymptotics are notorious for leading to lengthy and laborious calculations. Andrews and Stafford [2, 3] describe a cooperating family of *Mathematica* procedures serving as tools to perform such expansions. Collectively the tools provided by Andrews and Stafford provide an environment which the researcher can use to speak statistical asymptotics to the computer.

An example of computer algebra in applied probability is its application to Itô calculus. Recall that Itô calculus allows one to do calculus with **Brownian**

motion instead of smooth functions, thus providing a flexible means of representing diffusions and other continuous random processes. The price to be paid for this flexibility is that the usual fundamental theorem of calculus, namely,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t) = f'(t) \tag{3}$$

has to be replaced by the celebrated Itô formula, which takes the following form for Brownian motion B:

$$\mathrm{d}\, f(B) = f'(B)\, \mathrm{d}B + \frac{1}{2}\, f''(B)\, \mathrm{d}t \tag{4}$$

This leads to complicated formulae, which has motivated a number of workers to program the resulting structure into a variety of computer algebra packages. The earliest example is to be found in [12], which used a *REDUCE* implementation to solve problems in the statistical theory of **shape**. This implementation can be understood as a formal translation of the underlying algebra of stochastic differentials into the computer algebra package. It has been translated into *Mathematica* [13] and applied to mathematical finance problems, as well as to a number of other research problems.

Further Reading

We have already referred in Table 1 to some books introducing various computer algebra packages. Other valuable literature resources for computer algebra users include a number of critiques and comparisons of computer algebra systems. These are useful not so much at the stage of choosing which computer algebra system to use (as indicated above, this is more likely to be controlled by local availability) but rather after a certain amount of computer algebra experience has been gained. It can then be very helpful to get some insight into the effects of varying design philosophies, which helps one better understand the features and strengths of one's chosen computer algebra system.

Good general introductions to computer algebra are (naturally enough) specific to particular computer algebra systems: [9, 10, 18] and the first parts of [11, 21] all provide very good guidance for beginners using the respective systems.

Fateman has written an excellent critique of *MACSYMA* [6], and also of *Mathematica* [7], which should be required reading for all serious users of computer algebra. Comparisons of various systems are to be found in [8 , 19, 20, 22]. In all these cases, readers should be aware that, for obvious reasons, the computer algebra systems reviewed are usually not the versions currently on market.

Finally, discussion of the basic algorithms employed by computer algebra systems can be found in [1, 5]. This will be useful particularly for workers needing to use advanced features such as Gröbner basis algorithms.

References

[1] Akritas, A.G. (1989). *Elements of Computer Algebra with Applications*. John Wiley & Sons, Chichester and New York.

[2] Andrews, D.F. & Stafford, J.E. (1993). Tools for the symbolic computation of asymptotic expressions, *Journal of the Royal Statistical Society* **B55**, 613–627.

[3] Andrews, D.F. & Stafford, J.E. (2000). *Symbolic Algorithms for Statistical Inference*. Oxford University Press, Oxford.

[4] Currie, I.D. (1995). Maximum likelihood estimation and *Mathematica*, *Applied Statistics* **44**, 379–394.

[5] Davenport, J.H., Siret, Y. & Tournier, E. (1988). *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York.

[6] Fateman, R. (1989). A review of *MACSYMA*, *Institute of Electrical and Electronics Engineers. Transactions on Knowledge and Data Engineering* **1**, 133–145.

[7] Fateman, R. (1992). A review of *Mathematica*, *Journal of Symbolic Computation* **13**, 545–579.

[8] Harper, D., Wooff, C. & Hodgkinson, D. (1991). *A Guide to Computer Algebra*. John Wiley & Sons, Chichester and New York.

[9] Heck, A. (1996). *Introduction to Maple*, 2nd Ed. Springer-Verlag, New York.

[10] Heller, B. (1991). *MACSYMA for the Statistician*. John Wiley & Sons, Chichester and New York.

[11] Jenks, R.D. & Sutor, R.S. (1992). *AXIOM: The Scientific Computation System*. Springer-Verlag, New York.

[12] Kendall, W.S. (1988). Symbolic computation and the diffusion of shapes of triads, *Advances in Applied Probability* **20**, 775–797.

[13] Kendall, W.S. (1993). *Itovsn3*: Doing stochastic calculus with *Mathematica*, in *Economic and Financial Modeling with Mathematica*, H. Varian, ed. Springer-Verlag, New York, pp. 214–238.

[14] Maeda, R.E. (1991). *Programming in Mathematica*, 2nd Ed. Addison-Wesley, Reading, MA.

[15] MacCallum, M.A.H. & Wright, F.J. (1991). *Algebraic Computing with REDUCE*. Clarendon Press, Oxford.

[16] Mannion, D. (1994). The volume of a tetrahedron whose vertices are chosen at random in the interior of a

parent tetrahedron, *Advances in Applied Probability* **26**, 577–596.

[17] Pistone, G., Riccomagno, E. & Wynn, H.P. (2000). *Algebraic Statistics*. Chapman & Hall, CRC press, New York.

[18] Rayna, G. (1987). *REDUCE: Software for Algebraic Computation*. Springer-Verlag, New York.

[19] Stoutemeyer, D. (1991). Crimes and misdemeanors in the computer algebra trade, *Notices of the American Mathematical Society* **33**, 40–43.

[20] Wester, M. (1994). A review of CAS mathematical capabilities, *Computer Algebra Nederland Nieuwsbrief* **13**, 41–48.

[21] Wolfram, S. (1999). *The Mathematica™ Book, Version 4*. CUP, Cambridge.

[22] Zimmerman, P. (1996). A Comparison of *Maple* V.3 and *Mathematica* 2.2.2, Preprint of École Polytechnique Palaiseau France.

(*See also* **Software, Biostatistical**; **Matrix Computations**; **Numerical Integration**; **Optimization and Nonlinear Equations**; **Polynomial Approximation**)

WILFRID S. KENDALL

Computer Architecture and Organization

Introduction

Computer technology has made amazing progress since the first general-purpose electronic computer was created half way through the last century. Today less than a thousand US dollars will purchase a computer that has more performance, more memory, and more communication capabilities than a computer bought in 1980 at a cost of one million US dollars. This improvement is driven both by advances in the technology used to build computers and increasingly by innovations in computer designs harnessing the tremendous capabilities of integrated circuits.

A *computer architecture* refers to those parts of a computer system that are visible to a programmer, whereas a *computer organization* refers to those operational units and the interconnects, not visible to the programmer, that realize the architectural specifications. As a simple example, it is an architectural design issue as to whether or not a computer will have a multiply instruction. It is an organizational issue whether that instruction will be implemented by a special multiply unit or by a mechanism that makes repeated use of the add unit of the system. The distinction is important as one can have machines with the same architecture but different organization. Such machines will typically have different price and performance characteristics but will be able to run the same programs – in which case they are said to be compatible.

The aim here is two-fold: One is to give an overview of modern computer architectures and their organization. The second is to outline programming practices that will lead to efficient use of commonly available computer systems.

Building Blocks of a Typical Computer

A computer is a complex system containing millions of electronic components that can interact in billions of configurations. The key to understanding and describing such a system is to recognize that most complex systems, including the computer, possess a hierarchical structure [11]. With a hierarchical system, we need only deal with one particular level at a time. The behavior at each level depends only on a simplified, abstracted characterization of the system at the next lower level. This model is applicable to both hardware and software design.

At the highest level, we can view the computer as a system consisting of four components: The central processing unit (CPU), main memory, input/output (I/O), and control.

- **Central processing unit (CPU):** This is the "brain" of the computer. Its function is to execute programs by fetching their instructions, examining them, and then executing them. Execution of one instruction typically involves loading of data into main memory or adding two numbers. Distinct components of a CPU are the arithmetic and logical unit (ALU), which performs the computer's data processing functions on integers, floating point unit (FPU), which takes care of computations on real numbers, registers, which provide very fast storage within the CPU, control unit, and CPU internal communication.
- **Memory:** The function of memory units is to store programs and data for retrieval by the CPU. Memory can be classified as primary storage and secondary storage. Primary storage, also known as *Random Access Memory* (RAM), is electronic memory that operates at a relatively high speed but requires a power supply to retain data. Secondary storage, typically magnetic disks, tapes, and optical disks (CDs and DVDs), is mechanical and therefore generally slower. However, it is significantly cheaper and also persistent in the sense that data is retained more or less indefinitely.
- **Input/Output:** Moves data between the computer and its external environment, for example, keyboard, mouse, monitor, printer, networks, or other computers.
- **Control:** Provides for control and communication among CPU, memory, and I/O.

Each of these components constitute complex hierarchies of their own and there may be one or more of each of them in one computer.

Performance

The components described above each have their own set of performance characteristics and achieving

a balanced system in which all components make comparable contributions to the overall performance is a complex and delicate task. An overview of the most important performance issues is given in the following.

CPU Speed and Moore's Law

The speed with which a CPU performs its tasks is normally given as *clock frequency*, these days measured in GHz (gigahertz). The clock defines regular time intervals, *clock cycles*, used to synchronize the various subsystems of the CPU. To execute an instruction, the CPU divides the action to be performed into a series of basic steps such that each step can be completed in one clock cycle. A clock frequency of 1 GHz, therefore, means that the processor can complete one step every nanosecond (10^{-9} s). Be aware, though, that the clock frequency is a measure of the speed of which the processor is capable; the maximal performance will only be realized if the data to be processed are available when needed. For more information on this topic, see below and, for example, [5, p. 16, 329ff] or [13, pp. 39–56].

Gordon Moore [8] observed as early as 1965 that the number of transistors that could be placed onto a single chip was doubling every year – a phenomenon now known as *Moore's law*. To the surprise of many, including Moore, this exponential pace continued year after year into the subsequent decades (after 1970, the number of transistors roughly doubled every 18 months). The consequences of Moore's law are that CPU speeds increase due to the shorter distances between transistors, devices shrink in size, power and cooling requirements are reduced, reliability increases with fewer interchip connections, but the cost of making a single chip remains nearly constant, which, overall, leads to dramatic reductions in the costs of computer circuitry. [12, pp. 31–32].

Caching

Clock rates have roughly doubled every 18 months over the past three decades according to Moore's law. However, while processor power has raced ahead at breakneck speed, other critical components (e.g. memory, disks, communication speeds) of the computer have not kept up. Consequently, there is a constant need for *performance balance*, an adjustment of the computer organization to compensate for the mismatch among the capabilities of the various components. Nowhere is the problem created by such mismatches greater than in the interface between the processor and main memory (RAM).

While the *size* of computer memory has increased according to Moore's law, the typical storage and retrieval *speed* of memory has only increased by about 10% per year [9, p. 243], leading to an increasing gap between the speed with which the CPU can process data, and the speed with which the memory can feed the "monster". Figure 1 illustrates that while clock frequencies have increased from kHz to GHz ranges over the last 30 years, memory access times have only increased from their speeds in 1970 by a factor of approximately 7% per year to a couple of hundred MHz. If memory access fails to keep pace with the processor's demands, the processor will stall in a *wait state* and valuable computational time is lost. For more details, please consult [5, p. 304] or [6, pp. 391, 454–460, 501].

It is technologically possible to make memory that is sufficiently fast to match current clock rates but such memory is very expensive and must be located inside (or very close) to the CPU. This puts a limit on the size of such fast memory. The primary strategy for dealing with this dilemma is to combine a small amount of fast memory (called a *cache* and pronounced "cash") with the relatively slow but large RAM memory.

The basic observation is that memory references made in any short time interval typically use only a

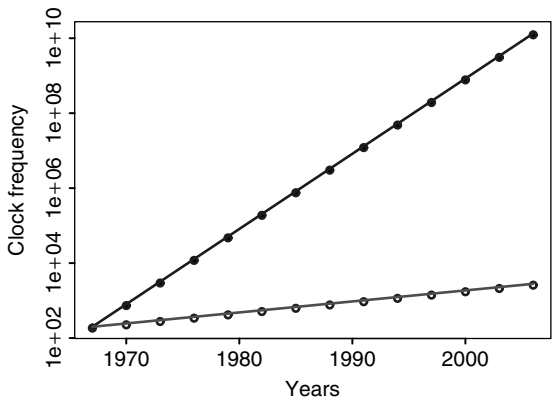

Figure 1 Gap in performance between memory (circles) and CPUs (dots) plotted over time

small fraction of the total memory. This leads to the *locality principle* that is fundamental to all caching systems. When a piece of data is referenced, the data and some of its neighbors are brought from slow memory into the quicker cache so that next time some, or (even better) all, of them are used, they can be accessed quickly.

Design issues such as ensuring that cached data agrees with data in memory, determining optimal cache size, associativity (where in the cache memory blocks can go), whether or not to cache instructions and data separately, are beyond the scope of this treatise. The message for the programmer is that good performance requires respect for the locality principle. It is better to reference elements in an array consecutively rather than in strides greater than one or, even worse, randomly. In addition, one should process a large amount of data block-wise rather than trying to load all of it and then, naively, work from one end to the other. The goal of both strategies is to maximize access to data loaded into the cache before that data gets replaced.

Most operating systems are designed so that data in memory, which is not currently in use, may be temporarily moved to disk in order to free up space. This process, called *swapping*, is necessary for the normal operation of a computer. However, excessive swapping will eventually lead to a phenomenon called *thrashing*, which is an undesirable situation in which the computer spends a large fraction of its time needlessly moving data around. Although thrashing rhymes with "caching", it is extremely detrimental to performance. A good blocking strategy will not only make good use of the memory hierarchy but also help avoid thrashing.

Modern compilers generally scrutinize source code at compile time to ensure that the executable code makes good use of the memory hierarchy. However, it is often necessary to help the compiler in this aspect by observing the locality principle; see for example [6, Chapter 5], [12, p. 41], or [13, p. 65] for thorough discussions of caching.

Memory Hierarchies

> No matter how big the main memory is, it is always too small.
>
> Andrew S Tanenbaum

The caching idea is usually applied at many levels: External cache memory itself is cached in even faster memory inside the CPU, which in turn is cached in registers. Parts of the main memory often serve as disk cache and the disk, in turn, may cache data from a web site or external media such as magnetic tapes. This entire set is known as the memory hierarchy. At the top, we have the CPU registers, which can be accessed at full speed. Next comes the cache memories, internal as well as external to the CPU, which are currently of the order 32 kB to a few megabytes. These cache memories are often referred to as level 1 and level 2 cache, respectively. Main memory is next with sizes ranging from 64 MB to tens of gigabytes. This is followed by magnetic disks and then slower devices such as tapes, CDs, DVDs, floppy disks, and even the whole **internet**. Figure 2 illustrates a conceptualized memory hierarchy.

Achieving Performance

Even though today's computers are capable of performing tasks at unprecedented speeds, there is no guarantee that this will hold true for any individual program. In fact, very few programs utilize the CPU fully and most computers spend their life moving data around rather than processing it. There are a few basic rules and pitfalls that, if observed properly, can increase the performance of a given program dramatically.

- **Make the common case fast:** This is perhaps the most important principle in both hardware and software design. By favoring the frequent event over the infrequent event, overall performance can be greatly improved. An example would be to replace a general solver of linear equations by a faster banded solver in (frequent) cases where equations have a banded structure; see, for example, [6, p. 39] for details.
- **Observe the locality principle:** Work with the cache memory by accessing data consecutively and process large amounts of data block-wise rather than from one end to the other; see for example [6, pp. 432–433] for examples of both strategies.

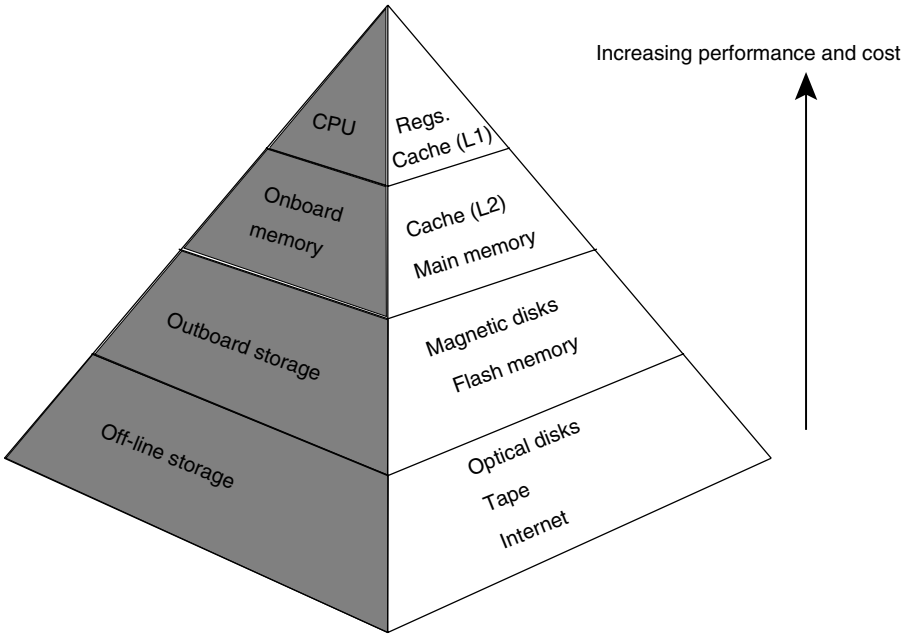

Figure 2 A typical memory hierarchy. Memory media at the top are fast but expensive. Media at the bottom are slow and cheap

Parallel Architectures and Distributed Computing

Increasing clock speeds and making CPUs smarter is one way of achieving greater performance. However, at any given point in time, there is a limit to the performance current technology can achieve for one processor.

An important improvement is to look for concurrency or parallelism where more than one functional unit is employed in order to exceed the performance of one unit. Parallelism comes in two general forms: *Instruction-level parallelism* and *Processor-level parallelism*. In the former, parallelism is exploited within individual instructions to yield more instructions per second out of the machine. In the latter, multiple CPUs work together on the same problem.

Instruction-level Parallelism

To execute a typical instruction, the CPU goes through a number of stages:

1. Fetch the instruction
2. Decode the instruction
3. Fetch the operand
4. Execute the instruction
5. Store the result.

Assuming that each stage can be completed in one clock cycle, the total execution time will take five clock cycles to complete.

Pipelining is a technique whereby multiple instructions are overlapped as in an assembly line. Each stage is carried out by a separate specialized unit operating in *parallel* with other units, although on a different instruction. As an example consider five instructions: During clock cycle 1, Unit A is fetching instruction 1, during cycle 2 Unit A is fetching instruction 2 while Unit B is decoding instruction 1 and so forth. Once the pipeline is full, that is, when the first instruction has been completed, this scheme will complete one instruction for every clock cycle instead of one for every five cycles. Pipelining is a key method for increasing CPU performance [6, p. A-2].

A higher degree of concurrency can be achieved by using multiple pipelines (possibly designed for specific operations such as floating point instructions) to fetch and execute several instructions in

parallel. This is known as *superscalar execution* and can provide execution of more than one instruction per clock cycle. The Pentium family, for example, employs multiple pipelines [13, p. 52]. Of course, the execution must preserve the logical correctness of programs, so extra control hardware and sophisticated compilers are needed.

Vector processors are specialized to operate on long arrays of numbers using pipelining [12, pp. 655–656], [13, pp. 54, 556]. Prominent examples of vector computers are members of the Cray family and the Fujitsu VPP family. Here, the programmer needs to pay special attention to the time taken to "fill up" the pipeline (latency) because it can be significant in the scheme of things. Fortunately, practices similar to those for utilizing caching apply: Operate on long consecutive vectors whenever possible.

Processor-level Parallelism

When multiple CPUs or even multiple computers are combined to solve a task, we talk about processor-level parallelism. Each individual processor may (and typically will) still utilize elements of instruction-level parallelism as described above.

One way of classifying these architectures is according to whether the processors have access to the same memory address space or, alternatively, each CPU has its own memory but communicates with others through a (typically high speed) network. In the former case, we talk about *shared memory machines* or *multiprocessors* while in the latter we talk about *distributed memory machines* or *multicomputers*. Examples of shared memory multiprocessors are the SUN Enterprise, which typically has up to 12 processors sharing a common memory or Pentium PCs with more than one processor. Examples of distributed memory multicomputers are IBM SP2, Compaq Alpha, and all configurations of the *Beowulf* type. The latter comprises a number of off-the-shelf PCs connected with a fast network to form a relatively inexpensive parallel computer; see for example `http://www.beowulf.org` for further details. With shared memory architectures, processors communicate by reading and writing to the same memory, whereas distributed memory architectures rely on explicit communication (often implemented as *message passing*) for interprocessor communication.

Programming shared memory machines is often considered easier than programming distributed memory machines, the reason being that one does not have to worry about where data is located. However, this is somewhat deceptive as *parallel performance* depends strongly on the spatial location of data. The locality principle applies here as well – and in the case of parallel computers, it has to be observed for every processor. Hamacher et al. [5, pp. 648–653] has a worked example of a parallel program written for both paradigms.

Fortunately, the same general guidelines apply to both parallel architectures and the goal is the same, namely, to be able to complete a certain task faster and/or be able to deal with larger problems.

Achieving Parallel Performance

One measure of parallel performance is the *speedup* defined as $S_P = T_1/T_P$, where T_P is the time required to execute a specific task on P processors. The ultimate aim is to be P times faster with P processors, that is $S_P = P$. However, observed speedup is usually less than this ideal and any speedup above $0.75P$ is considered to be good. One must address three critical issues to achieve good speedup:

- **Interprocessor communication:** The amount and the frequency of communications should be kept as low as possible in order to minimize the time processors spend waiting for data. This is the same theme as that of slow memory access preventing full utilization of processors, only much worse in this case due to network speeds being slower than memory speeds. Each communication has a fixed startup time (*latency*), which is independent of the amount of data to be transferred. Frequent transfers of a number of small blocks will, therefore, take longer than transferring all blocks in one bundle. Ensure that interprocessor communications are as infrequent as possible and that processors are busy between transfers; see, for example, [3, p. 23] or [6, p. 546] for details.
- **Data distribution and load-balancing:** The total execution time of a parallel program is determined by that of the slowest processor. If some processors finish much sooner than others, we say that the program is poorly load-balanced. Each processor should get its fair

share of the work load; see, for example, [3, p. 26] for details.

- **Sequential parts of a program:** If half of a program, say, is inherently sequential, the speedup can never exceed 2 no matter how well the remaining half is parallelized. This is a special case of what is known as Amdahl's law. Ensure that all the cost intensive parts are parallelized; see, for example, [3, p. 24], [5, p. 654] or [6, pp. 40, 537] for details.

Consult [10] for a representative example of how to analyze parallel performance.

Examples of Parallel Applications

Of particular interest to the readers are the following examples of the use of parallelism in biostatistics.

- *Parallel Inference Machine* (PIM) is a massively parallel controller designed to execute thousands of control programs concurrently and in real time. Several **bioinformatics** algorithms including *Basic Local Alignment Search Tool* (BLAST) have been implemented on this specialized architecture; see `http://www.paracel.com/faq/faq_algorithm_primer.html`.
- Apple's version of BLAST is developed in collaboration with Genentech and the Stanford University Genetics Department. It takes advantage of algorithmic improvements, advanced memory management and the ability of Apple's Power PC G4 processor to perform multiple operations per clock cycle. This is an example of how performance can be increased by writing and tuning software specifically for a particular computer architecture. However, this often comes at the expense of portability since such software rarely performs well (if at all) on other architectures; see `http://developer.apple.com/hardware/ve/acgresearch.html`.
- Turbogenomic's *Turboblast* delivers BLAST jobs across Beowulf Clusters and other distributed architectures in parallel; see `http://www.turbogenomics.com/products/turboblast_index.html`.

Future Directions

> It is difficult to make predictions, especially about the future.
>
> Robert Storm Petersen

For as long as people have attempted to predict the future of computer architectures, the view has been common, that growth of uniprocessor performance must soon end. Yet as various physical limits were reached, a new technology or approach would emerge yielding continued performance growth according to Moore's law. However, we may at last be approaching a fundamental limitation of silicon and many believe that improvements will slow down over the next 10 to 15 years [6, pp. 528, 648]. In any case, it is almost certain that parallel architectures will play an increasing role in the future, both in the form of tightly knit clusters (such as traditional parallel computers) and in the form of widely distributed networks. One important factor is that the available bandwidth, due to advances of fiber optics, is now growing even faster than Moore's law. This suggests a paradigm shift from a "CPU-centric" view to one where the network itself is central and processors are regarded as peripherals; see [4] or `http://www.gildertech.com`.

A new initiative dubbed *The Grid* promises to fundamentally change the way we think about and use computing. This infrastructure would connect multiple regional and national computing grids creating a universal source of ubiquitous and dependable computing power similar to the electricity and telephone utilities that have existed for almost a century; see [2], `http://www.gridcomputing.com`, or (for a bioinformatics example) `http://www.ncbiogrid.org`.

Other related trends are the emergence of *wireless interconnects* such as Bluetooth (`http://www.bluetooth.com`) and the use of small dedicated (as opposed to general purpose) computers, such as personal organizers, palmtops, fitness/diving/cycling computers, global positioning systems, and so on. Together with grid projects such as Globus, NetSolve, Condor, CUMULVS, WebFlow, they promise to make computing in the future less dependent on local resources and more universally available. For more recent thoughts about the future of computer architectures, see [1], [6, pp. 528, 644–645], or [7].

References

[1] Dongarra, J. et al. (2002). *Sourcebook of Parallel Computing*. Morgan Kaufmann Publishers.

[2] Foster, I. & Kesselman, C. (2002). *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers.

[3] Freeman, T.L. & Philips, C. (1992). *Parallel Numerical Algorithms*, Series in Computer Science. Prentice Hall International.

[4] Gilder, G. (2002). *Telecosm: The World After Bandwidth Abundance*. Touchstone Books.

[5] Hamacher, C., Vranesic, Z. & Zaky, S. (2002). *Computer Organization*, 5th Ed. McGraw Hill.

[6] Hennessy, J.L. & Patterson, D.A. (2003). *Computer Architecture – A Quantitative Approach*, 3rd Ed. Morgan Kaufmann.

[7] Khosrowpour, M. (2003). *Information Technology and Organizations: Trends, Issues, Challenges and Solutions*. Idea Group Publishing.

[8] Moore, G. (1965). Cramming more components onto integrated circuits, *Electronics Magazine* **38**(8) April.

[9] Murdocca, M.J. & Heuring, V.P. (2000). *Principles of Computer Architecture*. Prentice Hall.

[10] Nielsen, O.M. & Hegland, M. (2000). Parallel performance of fast wavelet transforms, *International Journal of High Speed Computing* **11**(1), 55–74.

[11] Simon, H. (1969). *The Sciences of the Artificial*. MIT Press, Cambridge.

[12] Stallings, W. (2000). *Computer Organisation and Architecture*, 5th Ed. Prentice Hall.

[13] Tanenbaum, A.S. (1999). *Structured Computer Organisation*, 4th Ed. Prentice Hall.

OLE M. NIELSEN

Computer Languages and Programs

Introduction

Computer languages are artificial languages that enable humans to give instructions to computer systems. Texts written in these languages are called *computer programs* or *code*. To run a program, a *compiler* or *interpreter* (themselves computer programs) must translate the program into *machine code*, which the hardware is able to execute directly. It is possible, but nowadays uncommon, to write programs directly in machine code, or rather in its mnemonic equivalent, *assembly language*.

The history of programming languages has seen the development of languages ever further removed from the machine code level. Not only is assembly language difficult to program but also every variety of computer central processing unit (CPU) has a different code. By contrast, modern high-level languages such as Java [2], Python [16], Matlab [10], and S (*see* **S-PLUS and S**; **R**) [21], are easier to understand; each command represents many lines of code in a lower level language and the code is not dependent on the CPU.

Classifications of Languages

Languages are classified in several different ways, such as by their historical development or computational level, by the sorts of applications for which they are commonly used, by the style of programming that they support, and by how they are translated to the computer's machine code.

Historical and Computational Level Classification

At one time, it was common to speak of "generations" of language development, with machine and assembly language being the first two. At the next level up, the older style of "high-level" languages, such as Fortran and C, were third generation or 3GL languages. In subsequent development beyond such "3GL" languages, any attempt at classification according to language generations breaks down; further development has gone in several different directions. The advances brought by new language developments add to human usability, both by compacting major functions from an earlier generation to a single command in the newer generation and by the use of more powerful conceptualizations. For complex tasks, the reduction in programming time and effort can be spectacular, leading at the same time to programs that are simpler and easier to maintain. Advances may, additionally, be in complexity of data structure, in orientation towards specific applications, in the generation of code from graphical interfaces, in scripting languages, in systems that allow the integration of components that may be written in different languages, in natural language recognition, and in so-called expert systems.

Classification According to Type of Application

C was attractive to programmers because it allowed much of the fine control that is available from assembly language. C^{++} is an extension of C that incorporates more modern language design features (see [2, 8]). C/C^{++} have found extensive use in writing compilers for other languages and in writing operating systems. Java, developed from C^{++}, has quickly gained a key role in **internet** applications. It has well-developed mechanisms for handling computing tasks, including graphics and animation, which involve multiple networked computers (*see* **Computer Architecture and Organization**). For example, the S language [3, 21] that is implemented in S-PLUS and R had in mind, data analysis, graphics (*see* **Graphical Displays**), and related scientific applications (*see* **Software for Clinical Trials**; **Software, Epidemiological**). Mathematica [24] was designed as a language and an environment for handling mathematical tasks, including symbol manipulation. Perl [15] was designed, initially, for use in system administration. It has found wide use in text searching and manipulation.

Programming Style Classification

Perhaps the most insightful classification of languages is by the style of programming that they support. A common distinction is between structured, logical, functional, and object-oriented programming, with modern languages often incorporating elements of all these approaches.

In structured programming, a problem is broken down into a sequence of steps that manipulate the

input data to produce the desired output. The program expresses these as commands to the computer, so that this is an *imperative* approach. The computer then runs this set of commands sequentially from beginning to end. Within this sequence of steps, discrete subtasks are usually identified and written into *procedures*. Hence, this approach is sometimes called *procedural*. Fortran and Algol pioneered this approach in the mid-1950s, and it continues to be in major use today. Languages that are in this tradition include Java, Delphi, Python, C, C^{++}, and Fortran; see [2, 8]. Usually, programs can share procedures that are located separately from the main program in a *library*.

Functional programming takes a different conceptual approach, where instead of executing commands, the program evaluates expressions or functions. Such languages have found wide use in the writing of systems for the text processing that is required for compilers and for the processing of natural languages. Languages that support this style of programming include Lisp and Scheme, which is a dialect of Lisp; see [8] for further details and background. The Emacs editor, which is popular as an interface to statistical systems such as S-PLUS and R, is embedded in a Lisp implementation. Lisp is a mnemonic for "list processing"; all operations are performed by modifying lists.

Prolog implements logical programming. Like functional programming, this is *declarative* rather than imperative, with a program closely resembling a logical proof. Both predicate and propositional logic are included in the language. While mainly used in **artificial intelligence** research, applications have included chemical structure databases (*see* **Chemometrics**) [11].

Object-oriented programming has recently come into prominence. In this approach, models for data are central. A data *object* encapsulates knowledge, both of data *attributes* and of procedures (*methods*) that may be applied to the object. For example, a "distribution" object may have attributes such as the number of data points, and the value of each, and methods to print histograms, estimate density curves, calculate standard deviations, and so on. The Java and C^{++} languages are the best known implementations, while many others, including S, have incorporated object-oriented features.

Excel has its own built-in language that allows its use for quite complex programming tasks. The key idea here is that of a spreadsheet whose entries are progressively modified as calculations proceed. The spreadsheet model is suitable for many simple accounting and data manipulation tasks, but is easily pressed beyond its proper limits for use in tasks for which Python or R or Mathematica would be more appropriate.

Quite different from any of the languages just noted are *markup languages*, of which HTML, XML, and TeX/LaTeX are well-known examples. These are texts that are designed to be "read" by a computer program, which then transforms them in some useful manner. For example, a web browser is a program that, among other functions, translates a file of HTML text into a display showing some portions of the text, while using other portions as instructions for layout, fonts, colors, and so on. The interchange of texts between these different systems can be a challenge [4].

Computer Translation Classification

The method of translating the language from its human readable form to a machine-usable form also varies between languages. There was once a simple choice between *interpreted* and *compiled* languages, with programs in interpreted languages referred to as *scripts*. The general wisdom was that interpreted languages were useful for small quick jobs, and could be quick to develop, but that a compiled language would be best for speed and stability.

When a language is compiled, a program called a compiler scans the human readable *source code*, and translates it into an *object file*, which the machine can execute. A pure interpreter, by contrast, does not preprocess, but scans the script line-by-line, translating into machine code on the fly. This removes many of the possibilities of optimization that a compiler can use, as well as many opportunities for error checking. Pure interpreters still exist in UNIX shell scripting. Compilers, and to some extent, interpreters, are specific to the particular type of machine or CPU. Language features may also be specific to particular types of machines, or to the operating system.

While compiled languages remain more efficient, the increased speed of computers, and access to segments of compiled code for computationally intensive tasks, has made this efficiency less important than earlier. Scripting languages such as Perl and Python [16] are now used for major applications, and

especially for web-based applications. Such scripts can, if necessary, be compiled.

More recently, languages have become hybrids, with both compiled and interpreted implementations or with multipass interpretation where code is translated into an intermediate state. The intermediate state may be virtual machine code – an approach that was pioneered in Smalltalk [8] and best known in Java. In this case, the same code can be used on any system that has its own virtual machine. A closely related approach is used in other languages including Perl and Python, where the intermediate state is a *parse tree* – a representation of the structure of the whole program in a form that is then executed by an interpreter. This is sometimes called "compiling", as parsing is usually a function of compilers. However, because no machine code is generated, this usage is disputed.

Abstract Representations of Problems

A modern view of computer languages is that they support abstract representations of problems in a form adapted to computer implementation. Thus, for the production of a simple form of rhyming dictionary, we may replace "riming" by "gnimir", "timing" by "gnimit", "liver" by "revil", and so on. After sorting, we restore the order in which the letters initially appeared within strings. The use of an abstract representation of an array of words is just as crucial as the procedural abstractions that reverse the order within words and sort words in a lexicographic order. All languages support some level of data abstraction, control abstraction, and procedural abstraction. They may allow entities that combine two or more of these basic forms of abstraction. Further detailed discussion of computer languages from this perspective is beyond the scope of this article.

Nonverbal Communication

Human communication makes extensive use of visual and tactile signals that supplement spoken language. Similarly, the point and click mechanisms of windowing systems supplement rather than replace keyboard or spoken language, and will develop into forms of communication that are richer and more versatile as yet. Visual programming approaches, where a symbolic visual representation describes the computation, can reduce the risk of logical errors in the written code.

Program Design and Implementation

Computer languages are used to construct computer programs. The range of applications is wide – examples are accounting, inventory management, maintenance of patient records, medical and other instrumentation, electronic mailing systems, statistical analysis, and so on. In these and other applications, computer programs may be expected to respond correctly to a huge range of possible inputs.

For simple problems, the writing of a program may be a straightforward use of computer language skills. For large and complex computer programs, program development must be carefully designed and managed. The article on **algorithms** discusses issues that are important for the design of individual components – functions or subroutines – of a computer program. The remainder of this section will comment briefly on software engineering issues that are important for the design and execution of large computer programs.

Steps in a large software engineering project will include: determination of requirements, construction of a specification, design of a computer program, and the writing of code that will implement the design. There should be careful testing at each step. When programs demand the combined efforts of a number of programmers, there must be an effective overall human management.

There have been spectacular failures that emphasize the considerable challenge to software designers and programmers (*see* **Software Reliability**, and [9, 13]). Checks in both hardware and software must ensure that inevitable occasional failure, whether from human error, from an operator error, or from an unanticipated interaction between operator and software, will not have catastrophic consequences.

Program Design Concepts

Key ways to manage complexity, additional to those that we have already described, include the use of encapsulation to hide the information not needed outside of the program module or object, the use of modularity to break programs into separate identifiable components, and the use of hierarchy to impose

a readily intelligible structure on the abstractions used for describing and implementing the program. Additionally, reuse of program modules assists documentation and maintenance, and reduces the burden of testing. Reuse strategies seem to be a particular strength of object-oriented programming.

Until the program is complete, no final check is possible. A prototype, typically an executable program shell from which many individual modules are incomplete or missing, may allow limited early testing. A prototype may be invaluable for helping clarify user requirements and design issues, in checking major aspects of program structure, and as an evolutionary step towards the development of the final product. Efficiency may not be an important consideration at the prototype stage.

Computer Language Use in Statistics

Statisticians have taken an interest in computer language development since its beginnings in the 1950s [5]. Efforts to bring together a code for frequently repeated tasks led to the development of libraries of subroutines or programs. The demand for coherence and uniformity led to the full-fledged package in which a master program handled major aspects of input and output and gave a common interface to the separate routines. Some packages quickly developed abilities for looping, branching, and conditional execution, which match those in, for example, Fortran. They may have added new features as occasion demanded, often without adequate regard to good overall design.

An alternative approach is to begin by developing a statistical language, then using the language to implement statistical analysis abilities. Lisp-Stat [20], which embedded statistical analysis abilities within the Xlisp dialect of Lisp, was an early example of this approach. The language S [3, 21], available commercially as S-PLUS (*see* **S-PLUS and S**) was designed and developed as a language for interactive data analysis and graphics, into which statistical abilities were then embedded. The **R** system [6, 17], which is a free (General Public Licence) implementation of a dialect of the S language, has become a popular environment for developing and testing new statistical methods.

Other languages that have been used by statisticians, and that are now mainly of historical importance, at least for statistical computing, include BASIC and APL [1]. BASIC was adapted to provide an interactive operating, editing, and programming environment on the first generation of microcomputers, beginning in 1975. APL (1962) was characterized by a heavy use of vector and matrix operators, and severe notational complexity. APL and BASIC went some way to providing, in their different ways, demands for an interactive programming environment in which it was easy to move between code development and execution. Much improved responses to these demands are now available.

Data Analysis as Experimental Programming

Statistical analysis problems frequently demand substantial adaptation of the analytical abilities that are immediately available in statistical software systems. Additionally, what emerges from earlier stages of an analysis will typically affect what is done at later stages. Oldford and Peters [14] describe the interactive programming needed for such tasks as experimental programming, a style of programming that S-PLUS and R are specifically designed to support. Depending on what each new computational step reveals and on what follows on from it, it may or may not become part of the final analysis.

The best environments for experimental programming link analysis closely with graphical presentation. Additionally, they automate much of the data and diagnostic testing that humans find tedious. They offer powerful and unified ways of conceptualizing and describing calculations that free the user from the need to worry about the details of implementation. Thus, Nelder and Wedderburn's **generalized linear models** [12] brought a large variety of models together into a common conceptual framework. The GLIM statistical package combined this conceptual framework with the Wilkinson and Rogers syntax [23] for **regression** and **analysis of variance**. The S-PLUS and R systems provide effective and natural object-oriented implementations.

Concepts from object-oriented programming can be important, as in the S language, in simplifying and unifying the description of the computing tasks. Thus, the same print and plot and other commands may be used, with widely differing effects, for a wide variety of objects – analysis of variance objects, linear model objects, and so on.

Experimental programming requires good strategies that will minimize the risk of undetected programming errors, allow retracing of steps, and facilitate documentation of what has been achieved.

Further Reading

Appleby and Vandekopple [2] give an overview of computer languages and computer language concepts. Levenez [8] has links to pages that give a large amount of information on computer languages. For program design and software engineering concepts, see [18]. Wexelblatt [22] is a useful reference for the history of computer languages; see also [8]. The article by Lang [7] (see also [19]) is both a critique of the existing environments for statistical computing and a proposal for the creation of a new generation of tools.

References

[1] Anscombe, F.J. (1981). *Computing in Statistical Science through APL*. Springer-Verlag, New York.

[2] Appleby, D. & Vandekopple, J. (1997). *Programming Languages: Paradigm and Practice*. McGraw-Hill, New York.

[3] Chambers, J.M. (1998). *Programming with Data*. Springer-Verlag.

[4] Goosens, M., Rahtz, S., Gurari, E.M., Moore, R. & Sutor, R.S. (1999). *The LaTeX Web Companion: Integrating TeX, HTML and XML*. Addison-Wesley, Reading, MA.

[5] Gower, J.C., Simpson, H.R. & Martin, A.H. (1967). A statistical programming language, *Applied Statistics* **16**, 87–89.

[6] Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**, 299–314.

[7] Lang, D.T. (2000). The Omegahat environment: new possibilities for statistical computing, *Journal of Computational and Graphical Statistics* **9**, 423–451.

[8] Levenez, E. (2003). Computer Languages Web Page. `http://www.levenez.com/lang`.

[9] Leveson, D.G. & Turner, C.S. (1995). CASE: an investigation of the Therac-25 accidents, in *Computers, Ethics and Social Values*, D.G. Johnson & H. Nissenbaum, eds. Prentice-Hall, Upper Saddle River, NJ, pp. 474–514.

[10] Matlab home page (2003). `http://www.mathworks.com`.

[11] McIlwain, S., Page, D., Spatola, A., Vogel, D. & Blondelle, S. (2002). Mining three-dimensional chemical structure data, in The 10th International Conference on Intelligent Systems for Molecular Biology (ISMB 2002). Edmonton.

[12] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

[13] Neumann, P. (1995). *Computer-Related Risks*. Addison-Wesley, New York.

[14] Oldford, R.W. & Peters, S.C. (1988). DINDE: towards more sophisticated environments for statistics, *SIAM Journal for Scientific and Statistical Computation* **9**, 191–211.

[15] Perl home page (2003). `http://www.perl.com/perl`.

[16] Python home page (2003). `http://www.python.org`.

[17] R home page (2003). `http://www.r-project.org`.

[18] Sommerville, I. (2000). *Software Engineering*, 6th Ed. Addison_Wesley, Harlow, UK.

[19] The Omegahat home page. (2003). `http://www.omegahat.org`.

[20] Tierney, L.(1990). *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York.

[21] Venables, W.N. & Ripley, B.D. (2000). *S Programming*. Springer-Verlag, New York.

[22] Wexelblatt, R.L., ed. (1981). *History of Programming Languages*. Addison-Wesley, New York.

[23] Wilkinson, G.N. & Rogers, C.E. (1973). Symbolic description of models for analysis of variance, *Applied Statistics* **22**, 392–399.

[24] Wolfram, S. (1999). *The Mathematica Book*, 4th Ed. Cambridge University Press, New York `http://documents.wolfram.com/v4/`.

CATHERINE LAWRENCE & JOHN H. MAINDONALD

Computer-aided Diagnosis

In the past, a clinician made a diagnosis by supplementing accumulated knowledge with information from written notes, printouts of diagnostic test results, and information from the literature, often obtained by reading medical textbooks and scanning printed versions of *Index Medicus*. Though all the necessary information was contained in this paper trail, a key piece of data could be easily missed. Given the growth in medical knowledge and the large amount of clinical data available, clinicians using paper records find it even more difficult to consider all diagnostic possibilities, particularly with unusual diseases and patient presentations.

Changing technology has created new opportunities for helping the clinician make an accurate diagnosis. Most hospitals and many outpatient practices now use computers for the systematic processing and storage of clinical data, making information retrieval more efficient and accurate. Many medical texts are now available in electronic form, allowing expedient and robust searches of large quantities of material. Also, the availability of MEDLINE, the computerized database from which *Index Medicus* is derived, allows clinicians to cross-reference medical topics from 1966 to the present.

Although these electronic systems allow more efficient retrieval of data, processing the information once it is retrieved usually is still left to the clinician. A clinician traditionally approaches the diagnostic task by compiling available data and developing a list of one or more diagnostic possibilities in a list called a differential diagnosis. Elements of this list are sequentially ruled out on the basis of their appropriateness to the overall clinical scenario. Diagnostic possibilities on this revised list are then tested and ranked to determine appropriate management for a patient.

Over the years, computer aids have evolved to assist with information processing and thus improve the diagnostic process further. The origin of computer-aided diagnostic systems often is credited to Ledley and Lusted [13]. Their 1959 paper described symbolic logic and **probability theory** that led to diagnoses similar to those produced by clinicians' complex reasoning, though probably without reproducing that reasoning exactly. Since then, there have been many refinements of the scope, methods, and capabilities of computer-aided diagnostic systems. In this chapter, we describe these systems and focus on their advantages and limitations and present an overall evaluation of the role of computer-aided diagnosis in clinical practice. The systems are

1. **Algorithms**,
2. **Bayesian** analysis,
3. Belief networks,
4. **Prediction** rules,
5. Rule-based systems,
6. Decision trees,
7. **Artificial intelligence/**Causal reasoning,
8. **Neural networks**.

In addition to these established methods of computer-aided diagnosis, two emerging technologies are being developed and analyzed for their role in computer-aided diagnosis:

9. Microarray technology,
10. Syndromic surveillance.

Algorithms

Algorithmic methods [18] are suitable when a flow chart can be constructed that represents the logic used by a clinician to make a diagnosis. To implement the algorithm, the computer asks the user for information, processes the information, compares the result to the criteria at a branch point in the flow chart, selects a branch, and then moves to the next branch point. This process continues until a terminal branch is reached and a decision can be made. For example, an algorithmic method has been useful for the diagnosis of acid-base disorders [3]. In this situation, the user enters laboratory values, and then answers a series of questions generated by the computer about the patient's clinical condition. When enough information is available, the computer responds with a diagnosis.

Figure 1 shows a simple algorithm to determine the nature of a primary acidosis based solely on laboratory parameters. In this example, the computer would ask the user for the value of the blood pH and the pCO2. The computer then could make diagnostic suggestions, based on the numerical relationship between the pCO2 and the HCO3.

In comparison with other decision aids, algorithmic methods have the advantage of being understood easily by the user. Because the stepwise design

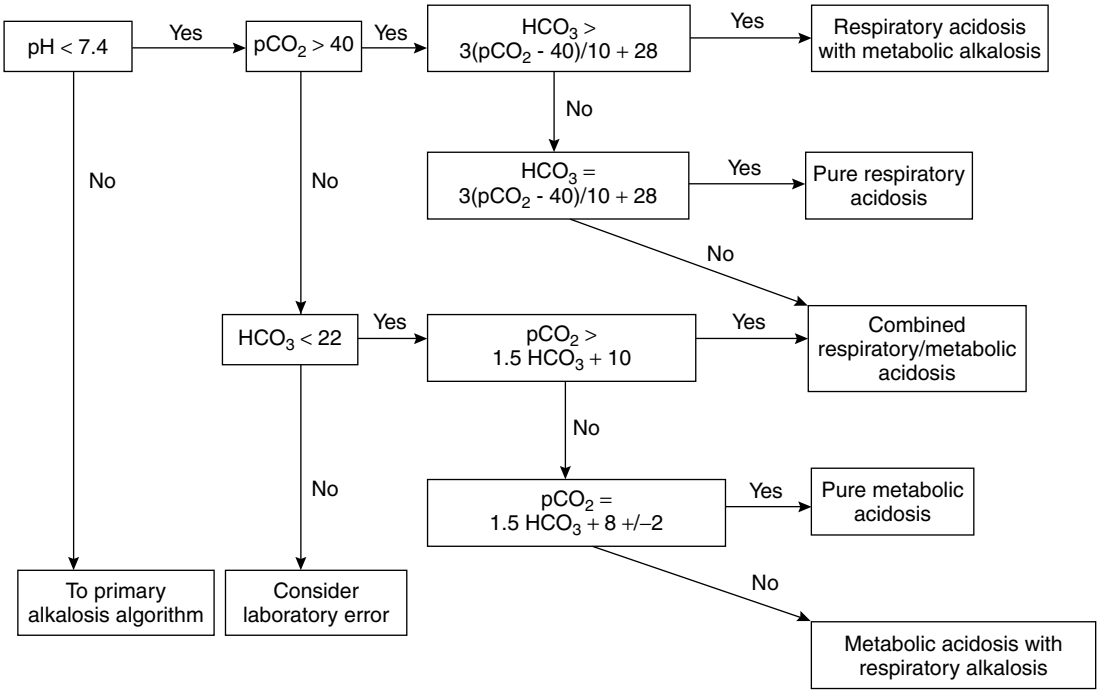

Figure 1 An algorithm for diagnosing classifications of acid disorders. On the basis of the results of laboratory parameters, the algorithm is traversed until an end point is reached. This end point may be a diagnosis, a suggestion to recheck results, or a recommendation to continue traversing a different algorithm

directly follows the clinician's logic, the program's conclusions can be explained so that the clinician can understand easily. Algorithmic methods, however, are not suitable for complex problems. Additionally, clinical diagnoses often involve ambiguous and **missing data**, which cannot be processed by the algorithmic method.

Bayesian Analysis

Bayesian analysis can partially overcome the problems created by missing data and uncertain associations. Instead of requiring a complete data set to provide a definitive conclusion, the Bayesian approach starts with the probability of disease in similar patients, and then uses the conditional probability of a symptom or a test result given the presence of disease to calculate the probability of disease with a symptom or a test result. In this approach, the presence or absence of a clinical finding will not necessarily eliminate a diagnostic possibility from contention, but may change its probability. The mathematical relationship can be expressed as follows:

$$\Pr(D|F) = \frac{\Pr(F|D) \times \Pr(D)}{\Pr(F|D) \times \Pr(D) + \Pr(F|\text{not } D) \times \Pr(\text{not } D)} \quad (1)$$

where $\Pr(D|F)$ is the probability of having disease D given the presence of a set of findings F; $\Pr(D)$ is the prior probability of disease D in the population; $\Pr(F|D)$ is the probability of observing a set of findings F, given the presence of disease D; and $\Pr(F|\text{not } D)$ is the probability of observing a set of findings, F, in a population without disease D (*see* **Bayes' Theorem**).

The calculation is trivial when dealing with a disease having only one finding. If a disease D is associated with many clinically independent findings,

$$P(F|D) = P(f1|D) \times P(f2|D) \times \cdots \times P(Fn|D) \quad (2)$$

where $f1, f2, \ldots, fn$ are individual findings of set F, $P(fi|D)$ is the probability of having finding fi given that a patient is known to have disease D.

There are many successful examples of the Bayesian approach. One involved the diagnosis of patients presenting with acute abdominal pain in which the computer accurately diagnosed 279 out of 304 patients, including 84 out of 85 with appendicitis [9]. The surgeons involved in the study made only 242 correct diagnoses.

There are many limits to the Bayesian approach, however. Most important is the assumption that all signs and symptoms are independent – a feature not typically seen in patient presentations. For instance, both the presence of fever and elevated white blood cell count may be predictors for a bacterial infection. However, the findings of fever and elevated white blood cell count are often associated despite the presence of bacterial infection. These two findings, therefore, are not clinically independent and the Bayesian calculation will be inaccurate. Mathematically, given the assumption of nonindependence of findings $f1$ and $f2$,

$$P(D|F) \neq P(D|f1) \times P(D|f2) \qquad (3)$$

Secondly, the approach assumes that outcomes are mutually exclusive, when in fact, patients can present with multiple disorders. For instance, if the probability of bacterial infection is increased by the presence of an increased white blood cell count, then this probability would be falsely increased by the presence of noninfectious causes of an increased white blood cell count such as leukemia. Thirdly, often the exact **conditional probabilities** are unknown and subjective estimations may decrease accuracy.

Belief Networks

Belief networks [6] were devised to address the Bayesian limitation of conditional nonindependence. The belief network is a lattice in which nodes are used to represent symptoms or diseases, and links among the nodes are structured as a directed graph in which there is a defined association or causal relationship between connected nodes. In this lattice structure, a node representing a disease can "point" to two independent clinical findings caused directly by the disease, which in turn can point to one or more other findings related directly to the first finding and indirectly related to the original disease entity. Conditional probabilities in the direction of the graph, such as the probability of observing a clinical finding given the presence of disease, are presumed known. Bayesian-like logic is then used to create a differential diagnosis by calculating the conditional probabilities of a set of diseases given a set of symptoms or clinical findings. The limitation of belief networks is that they are computationally complex, requiring exponential growth in the number of calculations for each additional node. This limitation, however, is partially mitigated by the use of heuristics and algorithms that can simplify the calculations of some structured networks.

To highlight the difference between a Bayesian approach and a belief network, consider the mathematical relationship between a diagnosis of metastatic cancer and several known findings such as increased total serum calcium, brain tumor, coma, and papilledema. If each of these findings was considered independent of any other finding, the relationship could be diagrammed as in Figure 2.

The probability of having Disease A in the setting of positive findings B through E would be based upon the Bayesian formula with $\Pr(F|A) = \Pr(B|A) \times \Pr(C|A) \times \Pr(D|A) \times \Pr(E|A)$, where F is the set of findings such that $B = T$, $C = T$, $D = T$, $E = T$.

However, medical experience has shown that the presence of coma and papilledema may be related to the presence of a brain tumor, which in turn is related directly to the presence of metastatic cancer; so these findings are not conditionally independent. As a result, coma and papilledema must fall to secondary positions, related directly to brain tumors and indirectly to metastatic cancer. Medical experience also shows that the presence of coma is often related to increased total serum calcium. The belief network structure illustrated in Figure 3 reflects these hierarchical relationships. With this structure, it is possible to calculate the probability of having Disease A with

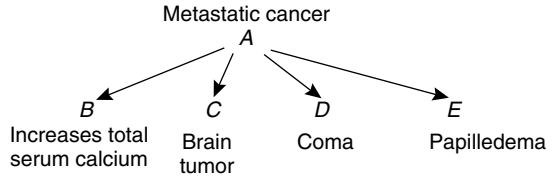

Figure 2 A Bayesian approach to determining the likelihood of metastatic cancer given the presence of associated findings. The structure presumes independence of findings B through E

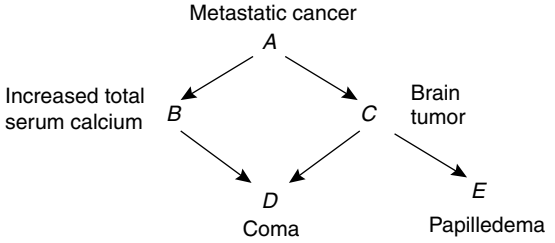

$\Pr(A) = 0.001$

$\Pr(B = T \mid A = T) = 0.3$
$\Pr(B = T \mid A = F) = 0.001$

$\Pr(C = T \mid A = T) = 0.2$
$\Pr(C = T \mid A = F) = 0.00005$

$\Pr(E = T \mid C = T) = 0.4$
$\Pr(E = T \mid C = F) = 0.002$

$\Pr(D = T \mid B = T, C = T) = 0.1$
$\Pr(D = T \mid B = T, C = F) = 0.01$
$\Pr(D = T \mid B = F, C = T) = 0.05$
$\Pr(D = T \mid B = F, C = F) = 0.00001$

Figure 3 A belief network of metastatic cancer with associated clinical findings. In this structure, the disease, metastatic cancer, is related directly to the finding of brain tumor, which, in turn, influences the presence of coma and papilledema. Increased total serum calcium is also related directly to metastatic cancer, and, along with brain tumor, may influence the finding of coma

any combination of associated findings. The calculation is too complex to be illustrated here [6].

Prediction Rules

Despite the mathematical rigor of Bayesian analysis and belief networks, studies have shown that medical experts do not apply this logic. Instead, they often apply prediction rules, which are simple declarative statements of the form "if antecedent then consequent", where the antecedents are one or more patient characteristics and the consequent is the potential diagnosis. Patient characteristics can include demographic characteristics, symptoms, physical examination findings, or laboratory findings. For example, the following prediction rule derives from experience and often is invoked in medicine clinics: "If the patient has a sore throat associated with fever; painful, swollen lymph nodes; tonsillar exudates; and a lack of cough, the diagnosis is more likely streptococcal than viral." Clinicians do not necessarily know how predictive this rule may be, but it is often applied on the basis of the empirical observation that the rule is frequently true.

With the availability of large clinical databases and the diffusion of advanced statistical methods into clinical research, prediction rules are increasingly being created using more sophisticated, mathematically based methods. In these approaches, the statistical analysis identifies which clinical variables are related to the outcome, and then calculates the strength of that relationship. Once this information is known, a rule can be created that contains an added degree of certainty. For example, several analyses of patients with sore throats have been done. These analyses attempted to find a relationship between the empirically observed findings involving patients with sore throats. In one such analysis, **logistic regression** demonstrated that the strength of the relationship between each finding and streptococcal sore throat was about the same [5]. Therefore, the prediction rule specifies counting how many findings are present and then comparing the result with the measured probability of a streptococcal sore throat. If none of the findings is present, the probability of a streptococcal sore throat is 2.5%. The probability of streptococcal sore throat increases when more findings are present: 6.5% with one, 14.8 % with two, 32.0% with three, and 55.7% with all four.

Although this prediction rule was developed with logistic regression, a similar rule for determining the likelihood for a strep throat was developed with **discriminant analysis** [25]. Using this rule, varying point scores are given for the presence and degree of a finding and probabilities are assigned on the basis of the total score. In fact, many other statistical methods that measure the association between variables can be used to develop prediction rules, including such simple methods as **contingency tables** and branching algorithms. Because the relationships among variables are often complex, however, more

complex methods often are appropriate. The more common methods used in clinical medicine are ordinary **least-squares** regression, logistic regression, and discriminant analysis, although recursive partitioning methods (*see* **Tree-structured Statistical Methods**) are becoming increasingly popular, and event-history methods like **Cox regression** are being applied to problems that involve predicting the time to an event, such as death.

Rule-based Systems

Individual prediction rules, both empirically and mathematically derived, have been grouped together in rule-based systems to solve diagnostic dilemmas in a variety of clinical settings [8, 26]. In this approach, the system stores many rules, sometimes several hundred or more, which are processed by a rule interpreter. This interpreter can be designed to work in a top-down fashion in which a disease of interest is tested, by attempting to establish the presence of clinical characteristics known to be linked to the disease by rules in the knowledge base. The top-down approach can be useful if a diagnosis needs to be "ruled-out". For instance, if the computer were considering the diagnosis of streptococcal pharyngitis in a patient with a sore throat, the computer would ask the user if the clinical characteristics associated with streptococcal disease were present. If not, the computer would move on to ask other questions looking for different causes for a sore throat. The advantage to this method is that it may prompt the clinician to find additional information necessary to make a diagnosis. The disadvantage is that it only considers one diagnosis at a time, and a possibly correct diagnosis may be overlooked if the heuristics for identifying that diagnosis are incomplete.

Alternatively, the rule interpreter can work in a bottom-up fashion by accumulating all known patient characteristics and then applying the rules to determine which diseases can be inferred. The bottom-up method is useful because it provides a breadth of diagnostic possibilities, though its main disadvantage is that an incomplete or inaccurate input list of symptoms or patient characteristics may lead to an inaccurate differential diagnosis.

Rule-based systems have several advantages. The rules are modular, declarative statements of medical knowledge that can be used to explain to a user the logic involved in reaching a conclusion. This capability helps the clinician to trust the answer more. However, medical knowledge is not organized into discrete rules, and rule-based systems do not necessarily organize medical knowledge into formats that are intuitive to clinicians. Additionally, the correct application of a rule depends on the clinical context, and specifying all the possible contexts requires an exponential growth in the number of rules. Successful rule-based systems, therefore, encompass limited clinical domains. One such example is MYCIN, which was designed to provide consultative advice on the diagnosis and therapy of infectious disease [8].

Decision Trees

Decision trees constitute still another approach to computer-aided diagnosis. Though decision trees do not provide lists of differential diagnoses, they are used to choose among diagnostic strategies logically and consistently.

Decision trees are structures that rigorously define and link choices and possible outcomes [18]. The tree consists of a linked series of nodes. At the origin is a decision node that reflects the diagnostic choices. Each choice may have one or more intermediate outcomes, which are represented as chance nodes. These nodes may connect to deeper elements of the tree that represent secondary events or further decisions. The branches of the tree end in terminal nodes that represent terminal outcomes and are assigned values, usually based on the clinical importance of the outcome, the cost of the outcome, or the patient's preference for the outcome. Simple calculations based on the probability of each outcome combined with its value determine the expected value of making any given choice. The choice with the highest expected value is preferred over other choices. **Sensitivity analyses** can be performed to determine whether the relationship among expected values changes when inputs are varied over reasonable ranges. In the decision tree shown in Figure 4, the label above each branch describes the action represented by the branch and the label below the branch describes the probability that the action will occur.

A common criticism of decision trees is that many probabilities and values in the trees are not known with certainty. Often, however, the relative ranking of expected values for alternative choices can be shown

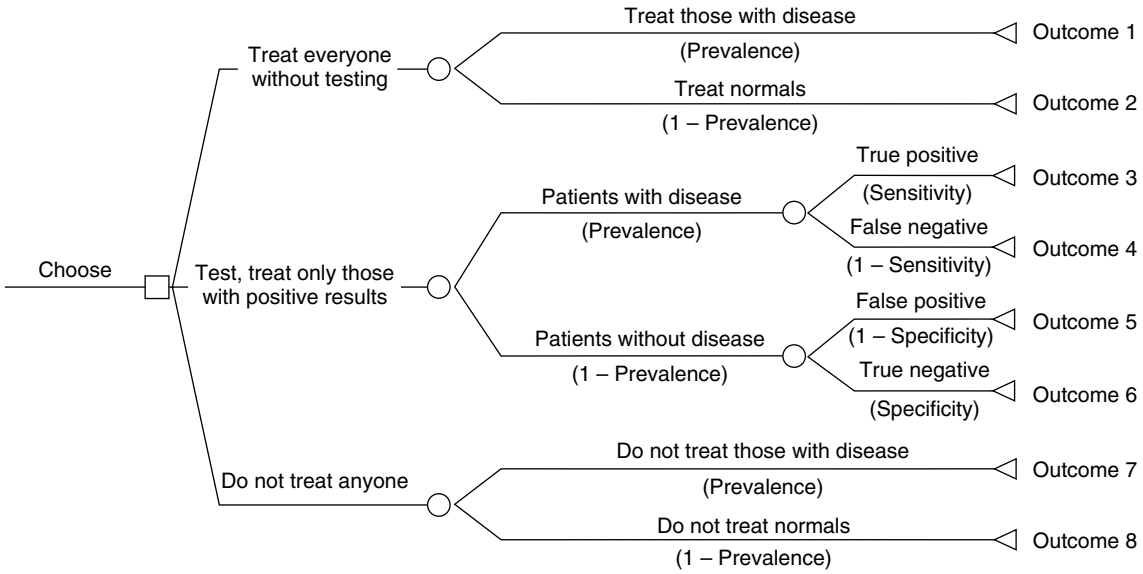

Figure 4 A generalized decision tree for deciding whether to perform a diagnostic test or to give or withhold treatment without testing. The label above each branch describes the action represented by the branch and the label below the branch describes the probability that the action will occur

not to change over a wide range of probabilities and outcome values. Additionally, the construction of a decision tree and its calculation can be time consuming. However, the design process may raise possibilities not considered in a more superficial approach to the diagnostic process and as a result, change a diagnostic approach even without a formal calculation of the decision tree.

Artificial Intelligence/Causal Reasoning

Artificial intelligence (AI) adds a layer of sophistication to computer-assisted diagnosis. These systems not only contain structured sets of rules and mathematical relationships, but they are also programmed with the causal reasoning underlying these associations and many contain mechanisms for self-expansion of the system's knowledge base.

Causal reasoning mimics the thought process of a clinician and is central to the concept of artificial intelligence. Although the algorithmic and prediction-rule approaches depend on quantitative measures, causal reasoning in medicine is mainly qualitative [15, 17]. When deciding blood pressure management, a clinician knows, based on physiological principles, that increases in cardiac contractility cause the blood pressure to increase and that relaxation of blood vessels causes the blood pressure to decrease. The clinician cannot necessarily predict the precise quantities of these values, but can still manage the patient appropriately. However, qualitative reasoning makes it difficult to determine the overall effect of conflicting forces. For example, if cardiac contractility increases and blood vessels relax, the effect on blood pressure will be uncertain.

Causal reasoning can approach a problem from different levels of detail [15, 21]. On a superficial level, interactions can be viewed solely as clinical observations: "The patient has a fever because he has an infection." This explanation may be satisfactory, though other levels of detail are possible if the problem is viewed as a pathophysiological process. For example, "The presence of infection causes white blood cells to release interleukins that stimulate the hypothalamus to raise body temperature." The value of this deeper level of understanding is that it affords a more refined explanation when requested, and it allows use of the rule for other relationships that have a common pathway, for example, other processes that stimulate release of interleukins and result in fever. Though the deep level of understanding allows a more rigorous analysis, separation of the levels

enables the computer to deal with a more manageable set of facts and more closely mimic the manner in which a clinician approaches a problem. For example, though the relationship between insulin and glucose balance can be explained through pathophysiological mechanisms, a clinician can manage diabetes knowing only that administering insulin will lower blood sugar and prevent undesirable complications.

Though artificial intelligence may incorporate recognized statistical methods, the methods in which the formulas are applied require modification to more closely mimic human reasoning. For instance, a strictly Bayesian approach to causal modeling for computer-aided diagnoses can be criticized for its equal weighting of all presenting signs and symptoms. In biologic systems, clinical phenomena such as symptoms and laboratory test results are variably associated with disease. Additionally, our ability to measure these phenomena are inexact. A clinician intuitively considers this variability and uncertainty when making medical decisions. To more closely mimic human reasoning, Bayesian models need to be combined with **multivariate analysis** to take into account the intensity, distribution, and validity of a relationship [4]. Intensity is defined as the expected change in the effect given the cause. For instance, how does the likelihood of a diagnosis change as a laboratory value varies out of the normal range? Distribution refers to the variability of the intensity across patients. Validity is an assigned value reflecting the certainty of relationship. Defining these variables and incorporating them into statistical models is the challenge of artificial intelligence.

The statistical association between a clinical finding and disease entity is only one part of the hierarchy within a causal relationship [16]. The relationship must also have a temporal association. The presence of a historical finding may be significant only if it occurs within the correct interval in relation to the disease. For instance, streptococcal pharyngitis can be a cause of rheumatic fever, but only if the infection occurred a few weeks before the onset of rheumatic fever. Similarly, there must be a functional association. For example, diarrhea may be the functional cause of hypokalemia, but the degree of diarrhea should match the degree of hypokalemia, or else other causes need to be considered.

Incorporation of these mathematical methods into artificial intelligence requires collaboration among diverse disciplines such as psychology, linguistics, and the computer and decision sciences. A true artificial intelligence system would incorporate elements of human reasoning with an automated ability to acquire knowledge and exchange information with its human counterparts naturally. Working examples of artificial intelligence in the medical field have tested each of these features individually, with varying success, but a complete artificial intelligence system has yet to be developed.

Neural Networks

Neural networks are computer programs with features similar to biologic nervous systems. Proponents of neural networks [1, 7, 14] believe other approaches to medical diagnosis can never reflect the complexity of the relationships among symptoms and diseases. Neural networks utilize complex, nonlinear statistical methods to form mathematical relationships between the presence or the absence of clinical findings and the presence or the absence of disease. Paths in the computer representation of a neural network are activated when multiple inputs to nodes reach a certain threshold, causing the node to "fire", analogous to a neuron generating an action potential when a sufficient stimulus is provided. The firing node then stimulates other nodes in the network with a "weight" that is set through a learning process. The neural network "learns" by examining a set of patients with known symptoms and diagnoses. The generated "weights" are not constant and do not apply to a single clinical finding. The mathematical structure formed from this learning set can be applied to a test set of patients whose diagnoses are unknown.

Technically, neural networks are not true artificial intelligence because they lack an underlying causal structure with a deep layer of pathophysiological detail. As a result, a neural network typically cannot explain to clinicians the reasoning that underlies the conclusions the system generates. However, recent developments have enabled extraction of rules from neural networks that can address this limitation within a medical domain [11]. Statisticians have expressed criticism of the mathematical methods that underlie neural networks [19]. For example, there is no equivalent of a **power** calculation to determine how many patients should be included in the learning set. Though intuition would suggest that a larger sample is better, sometimes too large a learning set, with too little variety, imparts an inability to

generalize. However, despite these limitations, neural networks are more accurate than clinicians and other computer diagnostic aids in their prediction of some disease states including myocardial infarction, pulmonary embolism, appendicitis, radiographic analysis, and the analysis of electrocardiograms.

Microarray Technology

While a significant degree of human disease likely has a genetic origin, until recently, techniques to find genetic markers of disease generally have been limited to diseases caused by point mutations of single genes. However, the genetic causes of common diseases such as hypertension, or cancer with its varied clinical presentation and aggressiveness is likely related to the interplay of hundreds or even thousands of genes. Microarray technology is a recent development within molecular biology that has enabled simultaneous analysis of thousands of genes. A microarray is a rectangular grid of small drops (nanoliters) of genetic material that are bound to a solid surface, typically glass. Each spot of genetic material represents a different gene fragment. Since these drops are less than 250 μm in diameter, literally thousands of drops can be present on a glass slide only a few centimeters square. The small scale of these drops, and the mechanism by which the genetic material is "printed" on the slide is analogous to the development of computer microchips in which digital electronic components have been made progressively smaller and now etched into a silicon wafer as a microchip. As a result, these microarray slides are also known as gene chips [22].

In a microarray experiment, genetic material from patients (either DNA, mRNA, or cDNA) is labeled with a fluorescent dye and exposed to the genetic material on the microarray. On spots where the patient's genetic material is complementary to the genetic material on the slide, the patient's genetic material binds to the spot through a process called hybridization. The fluorescent label of the attached genetic material imparts a color change to the spot. Since each spot is so small, and there are thousands of spots to review, computer controlled lasers are used to determine the color change of each spot of the microarray. Genetic material from a patient with a disease and without a disease can be labeled with different colored dyes and exposed to a microarray. Given the genetic similarities of all people, genetic material from both patients will hybridize to many of the same spots on the microarray. However, some spots will only bind the genetic material from the diseased patient, while other spots will bind the material from the unaffected patient. As a result, the presence of a genetic disease can be assessed by observing patterns of colored spots on a microarray slide when the slide is exposed to a patient's genetic material. It is not necessary *a priori* to fully understand the function of the genetic material in each spot of the microarray. Rather, a determination that the pattern of colored spots is similar to that of a diseased individual suggests that a genetic disease is present. When areas of differential coloring are discovered between affected and unaffected individuals, the genetic material associated with those spots can be examined in more detail, thereby helping to focus attention on an area of the genome likely to be associated with disease.

Each microarray experiment generates a large data set with thousands of numbers that represent the degree to which each spot of genetic fragments is complementary to the patient's genetic material. **Cluster analysis** is a common technique used to discover the pattern of coloring that is associated with disease [22]. The computer is provided a training set of data, and it attempts to find patterns of numbers that are differentially associated with disease and nondisease states.

When samples are compared, the distinguishing characteristic does not necessarily have to be the presence or the absence of disease. Microarray analysis can compare genetic samples from patients with the same disease but different levels of aggressiveness or response to treatment. These comparisons might provide a more accurate prognosis or suggest an individualized treatment that is based on the individual's genetic characteristics.

Syndromic Surveillance

For years, technical and administrative systems have been in place, which detect outbreaks of common diseases such as the flu or chicken pox (*see* **Surveillance of Diseases**). In general, these systems have been used to direct resources to areas of need as soon as possible after an outbreak occurs. For example, a growing flu epidemic in a region may help to focus resources on ensuring immunological protection of as-yet-unaffected individuals. These systems

depend on the voluntary reporting of occurrences of symptoms that are consistent with a disease and the submission of patient specimens for more formal analysis. With the acts of bioterrorism of late 2001, it has become increasingly apparent that similar systems need to be implemented that can detect smaller outbreaks of more lethal diseases at an earlier stage when the condition may be more treatable and fewer people have been affected [23]. The difficulty with a practical implementation of such a system is that the early stages of a bioterrorism agent exposure may masquerade as a serious form of a more benign condition such as a "bad" flu or rash. On any given day, a large number of patients present to primary care offices and emergency departments with these symptoms, so it can be difficult to detect the signal of a few patients presenting with symptoms caused by a bioterrorism agent. Syndromic surveillance is the term used to describe the automated data collection and analysis necessary to detect disease caused by bioterrorism agents and changes in the rates of other diseases.

This automated process is currently infeasible in most parts of the country where clinical findings are recorded on paper. However, where electronic medical records exist in ambulatory medical clinics and emergency departments, clinical information systems can transmit instantaneously and automatically a presenting patient's chief complaint, assigned diagnosis, and laboratory information to central processing computers. These central computers integrate data both regionally and nationally. Analysis of this data using spatial cluster detection methods, recursive least squares(RLS) and probabilistic inference techniques may reveal otherwise unrecognized patterns of disease occurrence that could alert public health officials about exposure to a bioterrorism agent [12]. One example of a syndromic surveillance system is the real-time outbreak and disease surveillance (RODS) system that was developed at the University of Pittsburgh and implemented at the 10 emergency departments and 20 acute care facilities comprising the University of Utah Health Sciences Center and Intermountain Health Care during the Winter Olympics in 2002 [10, 24].

Evaluation of Computer Diagnostic Aids

Evaluation of computer-assisted diagnostic systems has been inconsistent. While current studies of neural networks show levels of accuracy in the 90% range, many older studies showed bias in case selection and counted the computer's differential diagnosis as correct even if the true diagnosis was assigned a low probability [20]. A more recent study [2] compared the performance of four differential diagnosis programs chosen for their relatively common use and the breadth of information they covered. In this study, 105 cases were chosen from actual clinic experiences because they were diagnostically challenging and the diagnoses were known with certainty. The clinical findings were entered into the diagnostic programs and the resulting differential diagnoses were analyzed on several objective and subjective scales. Objective measures of success included the presence of the actual diagnosis anywhere on the list of differential diagnoses, the presence of the diagnosis within the top 10 diagnoses, and the mere presence of the diagnosis within the program's knowledge base. Subjective measures of success included the relevance of the top 20 diagnoses, which were defined as diagnoses that were on a clinician's differential diagnosis, but were nonetheless incorrect, and additional diagnoses that were considered relevant, but not included on the clinicians' original differential. The correct diagnosis score, representing the presence of the correct diagnosis anywhere on the list, ranged from 52 to 71%. Only 37 to 44% of correct diagnosis appeared within the top 10 most likely diagnoses. The number of additional diagnoses found by the programs that seemed relevant, but were not, on the clinician's original differential ranged from 1.8 to 2.3.

To make correct and relevant diagnoses, the computer diagnostic aid requires a clinician to be certain of the clinical findings that are entered. These findings may range from answers to simple historical questions to results of invasive and expensive tests. Ideally, all results can be known and the values are accurate. In reality, this is not true. For instance, the differential diagnosis of abdominal pain is quite broad, and a specific diagnosis, such as appendicitis, depends on correctly identifying the presence or absence of presenting signs and symptoms. Signs and symptoms are inherently subjective. Is the patient's pain severe, or does he have a low pain tolerance? Does the patient have a firm abdominal wall, or is there localized tenderness? Are bowel sounds hypoactive, or is this a normal finding for this patient? Even laboratory studies are not definitive. Is the patient's white blood cell count "normal" for the patient despite it being outside

the usual range? The computer diagnostic aid cannot make these distinctions.

Whatever the analytical technique used by a computer diagnostic aid, if a clinician enters the presence of irrelevant signs or symptoms into a diagnostic program, the program cannot rank elements of the extensive differential diagnosis effectively. If important findings are inappropriately omitted, then important diagnoses will be missed.

Lastly, the knowledge base of the system must be considered. Most successful systems encompass a narrow domain of medicine. A complete medical knowledge base must have not only the presenting signs, symptoms, and special study results associated with disease, but also be aware of the complex interrelationships among diseases. The breadth and depth of medicine make this a difficult task even for a seasoned clinician. As with clinicians, the computer-aided diagnostic program can only be as knowledgeable as its knowledge base allows.

These limitations of computer diagnostic decision aids suggest that current technology does not allow them to substitute for expert clinical judgment. Therefore, there is ongoing work to establish appropriate contexts in which these tools can be used. In a general medicine setting, the true diagnosis was not always at the top of the differential diagnosis list, but computer aids identified an average of two relevant diagnoses missed by an expert clinician [2]. This result suggests that sometimes the computer would have motivated a clinician to alter a diagnostic work up strategy, and possibly improve patient outcomes. The value of these diagnostic aids would be different in a subspecialty or general surgery clinic. There is the possibility that they might be most useful in situations where the clinician has less training or experience. For instance, a decision aid could be used by a school nurse when deciding whether a child has a benign condition and can be kept at school or has a more serious condition and should be sent to a doctor's office or an emergency room.

Though there has been much attention directed at comprehensive diagnostic aids, many successful diagnostic aids are used in an appropriately narrow domain. Computers have demonstrated accuracy in EKG interpretation, rheumatologic diagnosis, and the classification of patients with chest pain. Outcomes are presumed to be improved with the use of these tools, though this has not been demonstrated rigorously.

Because they formalize the logic used by clinicians, diagnostic aids can help standardize the practice of medicine. For example, computer-assisted diagnostic aids are being used to develop clinical guidelines and critical care pathways. Also, a clinician using the diagnostic aid may benefit because the aid may suggest an unfamiliar diagnosis and thus redirect the clinician's line of questioning or lead to a special study. The value of this potential benefit is difficult to quantify or evaluate.

The realization of these possibilities and limitations has changed the roles for computer-aided diagnostic systems. The fabled model based on the Greek Oracle in which a computer substituted for the clinical judgment of a clinician has been discarded. Many analysts now believe that a diagnostic system should be supportive, not authoritative. In the future, the appropriate clinical contexts for these systems will be identified. Improvements in technology should enable clinicians to engage these systems more naturally and enable knowledge bases to grow at the same rate as medical knowledge. In this manner, clinicians will retain their central role in the diagnosis and care of patients with the computer diagnostic aid as a resource for difficult diagnostic dilemmas.

References

[1] Baxt, W.G. (1995). Application of artificial neural networks to clinical medicine, *Lancet* **346**, 1135–1138.

[2] Berner, E.S., et al. (1994). Performance of four computer-based diagnostic systems, *The New England Journal of Medicine* **330**, 1792–1796.

[3] Bleich, H.L. (1972). Computer-based consultation: electrolyte and acid base disorders, *The American Journal of Medicine* **53**, 285–291.

[4] Blum, R.L. (1983). Modeling and encoding clinical causal relations in a medical knowledge base, *Proceedings of the 7th Annual Symposium on Computer Applications in Medical Care* **7**, 837–841.

[5] Centor, R.M., Witherspoon, J.M., Dalton, H.P., Brody, C.E. & Link, K. (1981). The diagnosis of strep throat in an emergency room, *Medical Decision Making* **1**, 239–246.

[6] Cooper, G.F. (1988). Computer-based medical diagnosis using belief networks and bounded probabilities, in *Selected Topics in Medical Artificial Intelligence*, P.L. Miller, ed. Springer-Verlag, New York.

[7] Cross, S.S., Harrison, R.F. & Kennedy, R.L. (1995). Introduction to neural networks, *Lancet* **346**, 1075–1079.

[8] Davis, R., Buchanan, B. & Shortliffe, E. (1977). Production rules as a representation for a knowledge-based consultation program, *Artificial Intelligence* **8**, 15–45.

[9] de Dombal, F.T., Leaper, D.J., Staniland J.R., McCann A.P. & Horrocks, J.C. (1972). Computer-aided diagnosis of Abdominal pain, *British Medical Journal* **2**, 9–13.

[10] Gesteland, P.H., Wagner, M.M., Chapman, W.W., Espino, J.U., Tsui, F.C., Gardner, R.M., et al. (2002). Rapid deployment of an electronic disease surveillance system in the state of Utah for the 2002 Olympic winter games, in *Proceedings of AMIA 2002 Annual Symposium*, pp. 285–289.

[11] Hayashi, Y., Setiono, R. & Yoshida, K. (2000). A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders, *Artificial Intelligence in Medicine* **20**, 205–216.

[12] Kohane, I.S. (2002). The contributions of biomedical informatics to the fight against bioterrorism. (2002), *Journal of the American Medical Informatics Association* **9**, 116–119.

[13] Ledley, R.S. & Lusted, L.B. (1959). Reasoning foundations of medical diagnosis, *Science* **130**, 9–21.

[14] Lisboa PJG (2002). A review of evidence of health benefit from artificial neural networks in medical intervention, *Neural Networks* **15**, 11–39.

[15] Miller, P.L. & Fisher, P.R. (1987). Causal models for medical artificial intelligence, *Proceedings, 11th Symposium on Computer Applications in Medical Care* **11**, 17–22.

[16] Patel, R.S. & Semyk, O. (1987). Compiling causal knowledge for diagnostic reasoning, *Proceedings of the 11th Symposium on Computer Applications in Medical Care* **11**, 23–29.

[17] Patel, R.S., Szolovits, P. & Schwartz, W.B. (1981). Causal understanding of patient illness in medical diagnosis, *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* **7**, 893–899.

[18] Reggia, J.A. & Tuhrim, S. (1985). An overview of methods for computer assisted medical decision making, in *Computer-Assisted Medical Decision Making*, Vol. 1, J.A. Reggia & S. Tuhrim, eds. Springer Verlag, New York.

[19] Schwartzer, G., Vach, W. & Schumacher, M. (2000). On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology, *Statistics in Medicine* **19**, 541–551.

[20] Sutton, G.C. (1989). Computer-aided diagnosis: a review, *British Journal of Surgery* **76**, 82–85.

[21] Szolovits, P., Patil, R.S. & Schwartz, W.B. (1988). Artificial intelligence in medical diagnosis, *Annals of Internal Medicine* **108**, 80–87.

[22] Tefferi, A., Bolander, M.E., Ansell, S.M., Wieben, E.D. & Spelsberg, T.C. (2002). Primer on medical genomics, part III: Microarray experiments and data analysis, *Mayo Clinic Proceedings* **77**, 927–940.

[23] Teich, J.M., Wagner, M.M., Mackenzie, C.F. & Schafer, K.O. (2002). The informatics response in Disaster, Terrorism and War, *Journal of the American Medical Informatics Association* **9**, 97–104.

[24] Tsui, F.C., Espino, J.U., Wagner, M.M., Gesteland, P., Ivanov, O., Olszewski, R.T., et al. (2002). Data, network and application: technical description of the Utah RODS Winter Olympic Biosurveillance System, in *Proceedings of AMIA 2002 Annual Symposium*, pp. 815–819.

[25] Walsh, B.T., Bookheim, W.W., Johnson, R.C. & Tompkins, R.K. (1975). Recognition of streptococcal pharyngitis in adults, *Archives of Internal Medicine* **135**, 1493–1497.

[26] Wasson, J.H., Sox, H.C., Neff, R.K. & Goldman, L. (1985). Clinical prediction rules: applications and methodological standards, *The New England Journal of Medicine* **313**, 793–799.

Further Reading

Schwartz, W.B., Gorry, G.A., Kassirer, J.P. & Essig, A. (1973). Decision analysis and clinical judgment, *The American Journal of Medicine* **55**, 459–472.

Waxman, H.S. & Worley, W.E. (1990). Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer based system, *Medicine* **69**, 125–136.

M.G. WEINER, ERIC PIFER &
SANKEY V. WILLIAMS

Computer-assisted Interviewing

Computer-assisted Interviewing (CAI) is a term that describes the use of a computer to aid the interview process in a survey data collection. The computer generally presents the question text on the screen, along with the allowable response categories, and the interviewer or respondent records the answers directly into the computer. The computer can also be used to undertake various forms of processing at the time of the interview, and it can facilitate electronic data transfer.

CAI is an alternative methodology to traditional paper and pencil interviewing (PAPI), and it has many advantages to offer the researcher over the paper-based approach. The advantages and disadvantages of CAI relative to PAPI are discussed briefly below, with the emphasis on the most commonly occurring situation where interviews are conducted between an interviewer and respondents.

Variations

There are several variations in the way CAI can be administered. The most commonly known form is computer-assisted telephone interviewing (CATI) where the interview is conducted over the **telephone**. CATI is good for shorter, simpler surveys, as it can be more difficult to maintain rapport with the respondent over the telephone and respondent fatigue can occur more quickly [6]. CATI is most commonly conducted in a centralized environment, but there are also significant decentralized applications [2]. Centralized CATI offers certain benefits not available in a decentralized environment, such as automated call scheduling and more timely monitoring capabilities.

Computer-assisted personal interviewing (CAPI) is another variation of CAI. This involves the conduct of face-to-face interviews between the interviewer and respondents. CAPI is good for longer or more complex topics where it is more difficult to maintain respondent cooperation. A distinct advantage of CAPI is that visual aids can be used (e.g. prompt cards), and respondents can complete any associated paperwork (e.g. consent forms). CAPI, by its nature, is usually conducted as a decentralized operation.

A further variation is computer-assisted self-interviewing (CASI) where the respondent completes the survey instrument for him/herself. CASI has numerous applications. One example is in health research where a respondent can sit down at a computer and complete a simple diagnostic questionnaire. Another is in surveys of businesses where the instrument can be emailed or sent on floppy disk for completion. Other less common types of CASI include the use of telephone touchtone or voice recognition software [10]. If CASI is used, it is essential that the instrument provide adequate guides to respondents and that respondents have some basic computer skills.

Advantages

CAI provides greater scope for surveys to deal with issues in depth and to target specific subgroups in the population. This is because the complexity of determining the next relevant question sequence or identifying subgroups can be managed quickly and reliably by the computer program. This is not only an advantage for data quality through such things as lower item **nonresponse**, more consistent response, and so on, but it is also a cost-saving measure, as fewer interviewer errors will need to be corrected in the office. In a large-scale survey operations environment, several CAI surveys can be loaded onto the computer at one time for accurate and efficient enumeration in the field.

CAI enables the researcher to tailor the wording of questions based on information collected earlier in the instrument (e.g. inserting names into relevant questions about family members, using the name of a medication in questions about treatment for a health condition, etc.). This helps make the survey more specific and personalized, and this in turn aids respondent comprehension and rapport [10]. In addition, other information already known about the respondent can be used to assist with the interview (this practice is known as dependent interviewing). This is particularly useful in longitudinal surveys where information gathered in earlier interviews or from other sources (e.g. long-term health conditions, episodes in hospital, etc.) can be used in subsequent interviews. Another example is for monitoring change over time and responding appropriately when change occurs.

Processing time and costs can be reduced significantly with CAI because the data is in an electronic

format as soon as the survey is completed. Data in that format can be transmitted and loaded into the processing system within hours, compared to days or weeks for paper systems, where the forms have to be mailed and the data entered using data entry facilities. CAI also offers the facility of programming simple edits and consistency checks into the instrument, yielding further savings in processing time because some of the checking work is done while the interview is being carried out in the field. Field edits also provide for increased **data quality**, as any data conflicts can be reconciled immediately with the respondent.

Another distinct advantage offered by CAI is the ability to assist the interviewer to code complex responses directly into the computer, rather than recording the answers verbatim and coding them later. A number of alternative coding methods are available, ranging from a simple pick list (shown on the screen) to more sophisticated coding tools that guide the selection of an appropriate code according to prespecified rules and which require only a few letters to be typed. This not only reduces the costs of office coding, it also allows for greater data quality [4, 6]. A good example of this is an item like country of birth, where numerous responses are possible. With a coding tool, the interviewer can code the item in the field while still having access to the respondent, if there is a problem coding a particular response.

CAI also provides some secondary benefits, as the computer can be used for purposes other than just interviewing. For example, the interviewers can be given access to email or news from the office, training, and documentation can be provided electronically, and the computer can be used for other survey tasks such as coding of previously entered data. Sample management and cost monitoring tasks can also be timelier and less onerous in a CAI environment.

Disadvantages

The most obvious disadvantage of CAI relative to PAPI is the initial cost of setting up. There is a large capital investment involved whether you invest in personal computers linked on a network, or individual laptop, or handheld computers (for decentralized interviewing). Indeed, the equipment costs can be so large that despite the many efficiencies of CAI outlined above, CAI can be more expensive than PAPI overall. CAI is most suitable for an environment where the cost of the equipment can be amortized over a number of surveys.

CAI generally requires more time and effort to develop and test the survey instrument and associated field systems prior to the start of fieldwork. Indeed, the total time taken from the start of the survey to the analysis stage is not necessarily any shorter for CAI, especially for one-off surveys. For instance, CAI requires more rigorous **validation** processes to check the instrument [1, 8], as it is most important for the CAI instrument to be error-free (as far as possible) in the field because of the relative inflexibility of CAI for dealing with problems or unusual situations. A good example to illustrate the validation issue is questions with multiple response categories. In a PAPI questionnaire, it is easy to see at a glance that all the response categories are sequenced appropriately. However, in a CAI instrument, it is necessary to check the detailed computer code (a method that can be error prone), or to physically select each possible response category when testing the instrument to ensure that all sequencing is correct. Training for CAI is more complex than for PAPI, as interviewers require basic computer skills in addition to an understanding of the survey instrument. Training may consequently take longer and be more expensive [7]. As an aid to testing interviewer instructions, as well as the effectiveness of CAI instrument layout and design, and the associated field systems, usability testing has gained prominence in recent years. Usability testing focuses on the interviewer's interaction with the survey instrument and CAI systems, and generally involves people being observed in a controlled "laboratory" setting as they use the computer systems. Usability testing is effective in identifying serious and recurring usability problems [5]. As a related training issue, there is also a need to support the CAI technology, both hardware and software, in the field.

Decentralized CAI has the added difficulty of transferring survey instruments to interviewers in the field and the collected responses back to the office. There are several methods available to transfer survey instruments and collected records. One is to make the transfer completely electronic using a fax/modem. With this approach, it is wise to set up an electronic despatch and receipt system to ensure that the collection operations can be suitably managed. Another method is to send floppy disks through the

post. Both these methods (electronic transmission and mailing floppy disks) have security implications for the confidentiality of the survey results. There are good software solutions that make use of passwords and encryption to deal with these security problems, both on the computers and during transmission, but they can add further complexity and costs to a survey. An alternative is to have the interviewers bring the computers to an office to have the information loaded and downloaded, but this method is only really feasible if the interviewers are relatively close to the central location.

The use of computers also introduces occupational health and safety issues for the management of an interviewer work force. With centralized CAI, there are issues of time spent keying, and for field interviewers, there are issues of the size, weight, visibility, and positioning of equipment, which need to be dealt with [10]. Such issues may increase costs and/or restrict the kind of equipment that can be used. These issues may even have an impact on nonresponse (e.g. interviews standing at the doorstep are considerably more difficult to conduct with CAI).

CAI is generally best conducted with software specifically developed for that purpose. There are clearly costs associated with the purchase of the software that must be considered, but software can also have limitations that may restrict the flexibility of the **questionnaire design**. When interviewers first see the computer screen, they should be drawn immediately to the key features needed for successful delivery of questions and accurate recording of responses [3]. To do so requires consistent design, visual discrimination among the different elements (e.g. questions, interviewer's instructions, response categories, navigation tools, etc.), adherence to normal reading behavior (i.e. starting in the upper left-hand corner) and removal of unnecessary information or display features that distract from the task (i.e. clean screen designs with more "white" space). Software should also cater for emerging CAI "best practice" relating to font size and color, line length and spacing, word emphasis, and so on. [3, 9]. The software chosen to write the survey instrument should also ideally be compatible with the survey processing system in order to realize the full benefits of reduced processing costs [6].

An additional disadvantage of CAI software is that it is often unable to produce a suitable paper version of the instrument that includes all the functionality that can be programmed in the software. This can be a particular problem when dealing with people who want to understand the question sequencing and data item derivations.

Related Topics

There are other computer-assisted methodologies used in processing surveys. One is computer-assisted coding (CAC). CAC is similar to an autocoder in a CAI instrument, but it is generally used for more complex coding operations by specialist coders after the main enumeration. Applications for this could include such things as coding of a respondent's medical conditions, medications taken, or possibly their dietary intake.

Another computer-assisted methodology for processing surveys is computer-assisted data input (CADI). CADI is a sophisticated form of data entry whereby on-line edits (e.g. logical and range edits) are applied to the instrument as the data is keyed in. Once the instrument is keyed, the data is internally consistent, thus obviating the need for batch edits checking the internal consistency of forms. Batch edits are only required for checking across forms. Another advantage to CADI is that a CAC can be embedded, thus reducing the costs involved in office coding.

References

[1] Baker, R.P., Bradburn, N.M. & Johnson, A. (1995). Computer-assisted personal interviewing: an experimental evaluation of data quality on costs, *Journal of Official Statistics* **11**(4), 415–431.

[2] Bergman, L.R., Kristansson, K.E., Olofsson, A. & Safstrom, M. (1994). Decentralised CATI versus Paper and Pencil Interviewing: effects on the results of the Swedish labour force survey, *Journal of Official Statistics* **10**(2), 181–195.

[3] Couper, M.P., Beatty, P., Hansen, S.E., Lamias, M. & Marvin, T. (2000). *CAPI Design Recommendations*. Report submitted to the Bureau of Labour Statistics, US. Interface Design Group, Survey Methodology Program, Survey Research Center, University of Michigan.

[4] Groves, R.M. & Nicholls, W.L. (1986). The status of computer-assisted telephone interviewing: part II – data quality issues, *Journal of Official Statistics* **2**(2), 117–134.

[5] Hansen, S.E., Couper, M.P. & Fuchs, M. (1998). Usability evaluation of the NIHS instrument, presented at the *Annual Meeting of the American Association of Public Opinion Research*, May, St. Louis.

[6] Martin, J. and Manners, T. (1995). Computer assisted personal interviewing in survey research, in *Information Technology for the Social Scientist*, R.M. Lee, ed. UCL Press, London, pp. 52–71.

[7] Nicholls, W.L. & Groves, R.M. (1986). The status of computer-assisted telephone interviewing: part I – introduction and impact on cost and timeliness of survey data, *Journal of Official Statistics* **2**(2), 93–115.

[8] Peters, L., Morris-Yates, A. & Andrews, G. (1994). Computerised diagnosis: the CIDI-auto, *Computers in Medical Health* **1**, 103–107.

[9] Pierzchala, M. & Manners, T. (1998). Producing CAI instruments for a program of surveys, in *Computer Assisted Information Collection*, M.P. Couper, R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls & J.M. O'Reilly, eds. Wiley Series in Probability and Statistics, New York, pp. 125–145.

[10] Weeks, M.F. (1992). Computer-assisted survey information collection: a review of CASIC methods and their implications for survey operations, *Journal of Official Statistics* **8**(4), 445–465.

(*See also* **Data Management and Coordination**; **Interviewer Bias**; **Sample Surveys in the Health Sciences**; **Surveys, Health and Morbidity**)

EDEN BRINKLEY, ELIZABETH FINLAY & FRED WENSING

Computer-intensive Methods

One sense of "computer-intensive" statistics is just statistical methodology that makes use of a large amount of computer time. (Examples include the **bootstrap, jackknife,** smoothing, **image analysis**, and many uses of the **EM algorithm**.) However, the term is usually used for methods which go beyond the minimum of calculations needed for an illuminating analysis, for example by replacing analytic approximations by computational ones, or requiring numeric **optimization** or integration over high-dimensional spaces. We introduce the subject by a very simple yet useful example, and then consider some of the areas in which computer-intensive methods are used, to give a flavor of current research.

A Simple Example

Let us examine the idea of a **Monte Carlo** test (see, for example, [60]). Suppose that we have a test statistic T, large values of which indicate a departure from a simple null hypothesis. The traditional analysis is to report *either* the ***P* value** $p = \Pr(T > T_{\text{obs}})$ *or* whether p falls into one of the conventional ranges. However, to compute p we do need to know the distribution of T under the null hypothesis. Traditionally, either T was chosen because its exact distribution was tractable, or some large-sample approximation was used, perhaps a normal distribution.

The Monte Carlo test depends on our being able to **simulate** under the null hypothesis; for each of m replications we generate some artificial data from the null hypothesis and compute the value T_i of the test statistic. One idea is to use the empirical distribution of the T_i as an approximation to the null-hypothesis distribution of T; that is, to compute $\hat{p} = \#\{i : T_i > T_{\text{obs}}\}/m$. However, Monte Carlo tests use a clever variation. If the null hypothesis is true, we have not m but $m + 1$ samples from the null hypothesis, the $T_1, \ldots, T_m$ that we generated plus T itself. A simple counting argument shows that

$$\Pr(T \text{ is amongst the } r \text{ largest}) = \frac{r}{(m+1)}.$$

Thus, we can obtain an *exact* 5% test by taking $r = 1$, $m = 19$ or $r = 5$, $m = 99$ or $r = 25$, $m = 499, \ldots$.

When Monte Carlo tests were first proposed by Barnard [4] in 1963, they would have been rather slow to compute, and needed extensive programming. When the idea was rediscovered in 1975 for some problems in spatial statistics (see [59]), increased computing power had made Monte Carlo tests much more attractive, although the choice of m was still severely limited by computing resources. Eventually good analytic approximations were found for those test statistics under the simpler null hypotheses, but they are hardly ever used, as it has become easy to use the exact Monte Carlo test.

This example has many of the key features of computer-intensive methods: it makes use of a simple calculation repeated many times, it relaxes the distributional assumptions needed for analytical results, and it is in principle exact given an infinite amount of computation. Rather than considering large amounts of data, we consider large amounts of computation, as the ratio of cost of computation to data collection is continually falling.

It is also important that it allows us to avoid asymptotics. Large-sample results are only useful in practice by providing an approximate distribution theory (*see* **Large-sample Theory**). In the spatial-statistics context there are "several ways to infinity" [61]–that is, several possible asymptotic regimes–and it is only possible to find good enough approximations by combining leading terms from all of them. When we consider **neural networks** below, we will see that it is common to increase the complexity of the model with the amount of data to hand, so large-sample results are never appropriate (and in the real world models are always false).

Graphics

A considerable amount of computer power is used to plot or print graphs, and even more is needed for dynamic graphics; for example, to rotate views of data, interactively change the bin size of a histogram, or highlight or identify points (*see* **Graphical Displays**). We now take this for granted (although systems to do this are described later, in the section on Programming Environments). As computer power grows, researchers are exploring ways in which to

compute large numbers of views and let the software arrange the "interesting" ones to show to the user.

The Trellis system (see Becker et al. [7]; this is now part of **S-PLUS** [45]) is based on presenting graphical summaries of many "slices" of the data in a systematic way. Asimov's *grand tour* [3, 18] shows the user a continually rotating series of views (projections) of multidimensional data of real-valued variates, but with four or more variables it is almost impossible for the user to screen the series for interesting views. (See the discussion in [38].) **Projection pursuit** [17, 18, 25, 26, 38, 39, 42, 43] replaces human inspection by optimizing (locally) the "interestingness" of a view. Sometimes the results can be frustrating, but at other times very interesting structures are revealed (see, for example, [64]).

Simulation-Based Approaches

A *reductio ad absurdum* view of statistical methods is that they reduce to either the optimization or the integration of some function, with **Bayesian methods** majoring on integration. For reasonably realistic models the integration is often (extremely) computer-intensive. (Evans & Swartz [23] review analytical as well as Monte Carlo methods to approximate integrals in statistics.)

Simulation provides a very simple way to perform an integration such as $\phi = \mathrm{E}f(X)$. Just generate m examples, $X_1, \ldots, X_m$, from the distribution of X and report the average of $f(X_i)$. It is not usually a good way to find an accurate estimate of ϕ, for the **central limit theorem** (if applicable) suggests that the average error decreases at a rate $1/\sqrt{m}$. For smooth functions in a small number of dimensions, numerical quadrature can do better, and in a moderate number of dimensions it may be better to use nonindependent samples X_i, the so-called quasi-Monte Carlo method [24, 50, 60, 66, 67].

In many statistical applications we do not need to know the integral ϕ at all accurately; in the Monte Carlo test a significance level of 5% is somewhere between, perhaps, 1% and 10%. Thus, in many cases, simulation is a good choice for approximate integration. Remember that a simulation experiment *is* an experiment, and there are a number of techniques [60, 23] to design and analyze it to obtain maximum precision for the computer time spent.

Markov Chain Monte Carlo

The simulation-based methods are of course only easy if we have a simple way to simulate from the assumed model. In highly structured situations we can find that everything depends on everything else. This was first encountered in statistical physics [46] and spatial statistics [30, 59]. Those authors devised iterative algorithms that only used the conditional distributions of small groups of random variables given the rest. As successive samples are not independent but form a **Markov chain** (on a very large state space), these methods are known as MCMC, short for **Markov Chain Monte Carlo**.

The same ideas were taken up for hierarchical Bayesian models by Gelfand & Smith [28, 29] some years later, under the name of Gibbs sampling (precisely the algorithm published in 1977 [59] and 1984 [30], and in the Bayesian context by Pearl [51] in 1987), and the earlier work and many of its lessons have been overlooked. In particular, although the algorithms are guaranteed to converge under mild conditions, that convergence can be far too slow to be useful without clever design, and possibly even then; see [65].

Recent reviews of MCMC from several viewpoints are given in [9], [33], [35], and [75].

Simulated Annealing

Simulation can also be used to do optimization! Most optimization methods can only find a local optimum of a function f on a domain $\mathcal{X}$, and can have great difficulty in combinatorial optimization problems (in which some of the components of $\mathcal{X}$ are categories rather than numbers).

Suppose that we want to minimize f, and we know $f \geq 0$. Consider drawing samples X_i from a distribution over $\mathcal{X}$ with density proportional to $\exp[-f(x)/T]$ for $T > 0$. As we increase T, the distribution becomes more and more concentrated on the region in which $f(x)$ is near its global minimum. Suppose that we run an MCMC simulation and decrease T; we might hope that the sample at time N converges to a global minimum. This idea was suggested by Pincus [55, 56], but was rediscovered and used by Kirkpatrick et al. [41] under the name of

simulated annealing. The idea that the sample might converge proved to be too optimistic in practice (convergence occurs at rate $1/\log N$). Aarts & Korst [1] discuss the theory, and Cantoni [15] provides some useful results on how to vary T if a bounded amount of computing is to be used.

Despite the lack of a convincing theory, simulated annealing has been widely used in large-scale optimization problems. Perhaps the most familiar application in statistics is to the segmentation of noisy images [8, 30, 61].

There are a number of other ways in which to use simulation to approximate maximum likelihood estimates using MCMC; for example, in spatial statistics by Pentinnen [53] and Ripley [61] and in genetics by Geyer [33, 34]).

Relaxing Linearity

One major use of computer-intensive methods has taken place largely outside statistics, in a drive to relax linearity and use much more complex models. In 1993, Ripley [62] wrote about neural networks that

> Their pervasiveness means that they can not be ignored. In one way their success is a warning to statisticians who have worked in a simply-structured linear world for too long.

The extent of nonlinearity which is not just being contemplated, but is being built into consumer hardware, goes far beyond **nonlinear regression** in the sense of, for example, [5].

The most influential developments have been in decision trees (*see* **Computer-aided Diagnosis**) [13, 57, 64], Bayesian networks [2, 27, 36, 37, 49, 52, 64, 68], and neural networks [10, 62–64]. The emphasis in the first and last is on good prediction, and the complexity of the models which are built is normally limited only by the amount of data to hand, even when this is 10^4 or more samples.

Selecting and Combining Models

Another way in which a large amount of computer power has been harnessed is to consider a very large number of models. We should distinguish two uses of statistical models, for *explanation* and for **prediction**. Although statistical training has usually emphasized explanation, that the purpose of the analysis is to discover structure in the data set to hand, in many other fields the emphasis is entirely on prediction. Even in a field such as medical diagnosis (*see* **Decision Analysis in Diagnosis and Treatment Choice**), in which it is helpful to be able to explain the basis of the diagnosis (and this is the aim of work in Bayesian networks), it would be useful to have an accurate diagnosis system without explanation. And for a voice recognition system, to read vehicle license plates, for biometric security checking, and so on, all that is required is accuracy and perhaps speed.

This suggests that for prediction we should either choose a model on the basis of its predictions, or even choose to combine the predictions from several models. The traditional way to proceed was to divide the data into training, validation, and test sets. The training set is used to optimize the parameters in each of the candidate models, the validation set to choose the best one (or combination), and the test set to estimate the performance of the selected prediction method (necessary as the performance on the validation set will be an optimistic measure of future performance)

Rarely will we have enough data to afford separate training, validation, and test sets, and in many problems increasing the size of the training set will result in a significant gain in performance. Enter K-fold **cross-validation**, in which the data set is divided into K sections of roughly equal size, and each section in turn is used as the validation set for the candidate models trained on the rest of the data. When all K fits are done, we have used the whole data set as a validation set, and can choose the best model (or combination). To assess the future performance we also have to take into account the variability of the selection procedure, so we nest the K-fold cross-validation inside a V-fold split into training set and training/validation set. Typically, K and V will be chosen of the order of ten, so each model is fitted of the order of 100 times, and it is not uncommon to entertain 25 models. The procedure rapidly becomes computer-intensive, but only by repeating a simple building block (fit a model on one data set, predict on another) many times. Such procedures can easily be done in parallel, or on separate machines.

The idea of using a weighted average $\Sigma_i w_i f_i(x)$ of the predictions for all the candidate models goes back at least to Stone [71], and arises from Bayesian theory if we believe that one of the models is true,

but do not know which one. In the Bayesian context the weights are the posterior probabilities of the models, the computation of which usually involves a very high-dimensional integration over the parameters (see, for example [54], and [58]). Alternatively, we can use K-fold cross-validation to choose the weights w_i from the performance on the validation sets. Some examples of the use of model averaging in the statistical literature are [21, 31, 32, 44, 48], and [70], but there are many in other fields.

There is no reason why the weights we give to the various models should not vary (slowly) with x, which leads to the idea of regarding the predictions of the candidate models together with x as inputs to some nonlinear function g, so the prediction system becomes

$$g(x, f_i(x), \ldots, f_M(x); \theta)$$

and choosing θ by simultaneous fitting or cross-validation. (In the neural networks literature, variants are known as *stacked generalization* [77] and *hierarchical mixtures of experts* [40].)

Many of the nonlinear procedures such as decision tree induction are rather sensitive to the training set used, and many methods have multiple optima and so are sensitive to the starting values. We can apply the ideas of combining *models* to combining the predictions of the same model fitted on different training sets. Breiman [11, 12] called one procedure *bagging*, which averages the predictions of a model fitted to bootstrapped training sets. We can go further and design subsets of the training data to produce rather different fits to average, a process known as *boosting* [22].

This is a very active area of research, and some of the results of combining simple models have shown very appreciable gains in performance.

Programming Environments

A major attraction of computer-intensive methods is their conceptual simplicity, but most biostatisticians are no keener on computer programming than on deriving asymptotic approximations. The S language (see Becker et al. [6] and Chambers & Hastie [16]) is the preferred working environment of many of the researchers in the field. S is currently marketed exclusively as part of the commercial S-PLUS system [45]. Despite a steep initial learning curve, sophisticated analyses can be performed in S-PLUS very easily [76], and many researchers donate their S software to public archives, notably statlib [47]. The free XLispStat system (see Tierney [74] and also Cook & Weisberg [19], available from statlib [47]) is also popular and has a substantial archive of user-contributed software at UCLA [20]. The package XGobi [14, 17, 18, 72, 73] implements dynamic graphics and projection pursuit, and BUGS [69] is a (currently free) system for using Gibbs sampling for a class of Bayesian methods (*see* **Software, Biostatistical**).

References

[1] Aarts, E. & Korst, J. (1989). *Simulated Annealing and Boltzmann Machines*. Wiley, New York.

[2] Almond, R.G. (1995). *Graphical Belief Modeling*. Chapman & Hall, London.

[3] Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data, *SIAM Journal on Scientific and Statistical Computing* **6**, 128–143.

[4] Barnard, G. (1963). Contribution to the discussion of Bartlett's paper, *Journal of the Royal Statistical Society, Series B* **25**, 294.

[5] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*. Wiley, New York.

[6] Becker, R.A., Chambers, J.M. & Wilks, A.R. (1988). *The NEW S Language*. Chapman & Hall, New York.

[7] Becker, R.A., Cleveland, W.S. & Shyu, M.-J. (1996). The visual design and control of Trellis display, *Journal of Computational and Graphical Statistics* **5**, 123–155.

[8] Besag, J. (1986). The statistical analysis of dirty pictures (with discussion), *Journal of the Royal Statistical Society, Series B* **48**, 259–302.

[9] Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science* **10**, 3–66.

[10] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

[11] Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**, 123–140.

[12] Breiman, L. (1996). The heuristics of instability in model selection, *Annals of Statistics* **24**, 2350–2383.

[13] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey.

[14] Buja, A., Cook, D. & Swayne, D.F. (1996). Interactive high-dimensional data visualization, *Journal of Computational and Graphical Statistics* **5**, 78–99.

[15] Cantoni, O. (1992). Rough large deviation estimates for simulated annealing: application to exponential schedules, *Annals of Probability* **20**, 1109–1146.

[16] Chambers, J.M. & Hastie, T.J., eds. (1992). *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove.

[17] Cook, D., Buja, A. & Cabrera, J. (1993). Projection pursuit indices based on orthonormal function expansions, *Journal of Computational and Graphical Statistics* **2**, 225–250.

[18] Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995). Grand tour and projection pursuit, *Journal of Computational and Graphical Statistics* **4**, 155–172.

[19] Cook, R.D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

[20] de Leeuw, J. Archive of XLispStat software. Use WWW or anonymous ftp to `www.stat.ucla.edu`.

[21] Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion), *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

[22] Drucker, H., Cortes, C., Jaeckel, L.D., LeCun, Y. & Vapnik, V. (1994). Boosting and other ensemble methods, *Neural Computation* **6**, 1289–1301.

[23] Evans, M. & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems, *Statistical Science* **10**, 254–272.

[24] Fang, K.-T. & Wang, Y. (1994). *Number-theoretic Methods in Statistics*. Chapman & Hall, London.

[25] Friedman, J.H. (1987). Exploratory projection pursuit, *Journal of the American Statistical Association* **82**, 249–266.

[26] Friedman, J.H. & Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers* **23**, 881–890.

[27] Fung, R. & Del Favarro, B. (1995). Applying Bayesian networks to information retrival, *Communications of the ACM* **38**, 42–48, 57.

[28] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[29] Gelfand, A.E., Hills, S.E., Racine-Poon, A. & Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *Journal of the American Statistical Association* **85**, 972–985.

[30] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[31] George, E.I. & McCulloch, R.E. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association* **88**, 881–889.

[32] George, E.I. & McCulloch, R.E. (1996). Stochastic search variable selection, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 203–214.

[33] Geyer, C. (1992). Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **7**, 473–511.

[34] Geyer, C.J. (1996). Estimation and optimization of functions, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 241–258.

[35] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[36] Heckerman, D. & Wellman, M.P. (1995). Bayesian networks, *Communications of the ACM* **38**, 26–30.

[37] Heckerman, D., Breese, J.S. & Rommelse, K. (1995). Decision-theoretic troubleshooting, *Communications of the ACM* **38**, 49–57.

[38] Huber, P.J. (1985). Projection pursuit (with discussion), *Annals of Statistics* **13**, 435–525.

[39] Jones, M.C. & Sibson, R. (1987). What is projection pursuit? (with discussion), *Journal of the Royal Statistical Society, Series A* **150**, 1–36.

[40] Jordan, M.I. & Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm, *Neural Computation* **6**, 181–214.

[41] Kirkpatrick, S., Gellat, C.D., Jr & Vecchi, M.P. (1983). Optimization by simulated annealing, *Science* **220**, 671–680.

[42] Kruskal, J.B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation", in *Statistical Computation*, R.C. Milton & J.A. Nelder, eds. Academic Press, New York, pp. 427–440.

[43] Kruskal, J.B. (1972). Linear transformation of multivariate data to reveal clustering, in *Multidimensional Scaling: Theory and Application in the Behavioural Sciences*, R.N. Shephard, A.K. Romney & S.K. Nerlove, eds. Seminar Press, New York, pp. 179–191.

[44] Madigan, D. & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of the American Statistical Association* **89**, 1535–1546.

[45] MathSoft Inc. S-PLUS. Data Analysis Products Division, MathSoft Inc., Seattle.

[46] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1091.

[47] Meyer, M.J. statlib. On-line archive of data and computer software. Use WWW, anonymous ftp, or Gopher to `lib.stat.cmu.edu`.

[48] Moulton, B.R. (1991). A Bayesian-approach to regression selection and estimation with application to a price-index for radio services, *Journal of Econometrics* **49**, 169–193.

[49] Neapolitan, E. (1990). *Probabilistic Reasoning in Expert Systems. Theory and Algorithms*. Wiley, New York.

[50] Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia.

[51] Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models, *Artificial Intelligence* **32**, 245–257.

[52] Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems. Networks of Plausible Inference*. Morgan Kaufmann, San Mateo.

[53] Pentinnen, A. (1984). *Modelling Interaction in Spatial Point Patterns: Parameter Estimation in the Maximum Likelihood Method*. Jyväskylä Studies in Computer Science, Economics and Statistics, Vol. 7. Jyväskylän Yliopisto, Jyväskyä.

[54] Philips, D.B. & Smith, A.F.M. (1996). Bayesian model comparison via jump diffusions, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 215–239.

[55] Pincus, M. (1968). A closed form solution of certain programming problems, *Operations Research* **16**, 690–694.

[56] Pincus, M. (1970). A Monte Carlo method for the approximate solution of certain types of constrained optimization problems, *Operations Research* **18**, 1225–1228.

[57] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.

[58] Raftery, A.E. (1996). Hypothesis testing and model selection, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 163–187.

[59] Ripley, B.D. (1977). Modelling spatial patterns (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 172–212.

[60] Ripley, B.D. (1987). *Stochastic Simulation*. Wiley, New York.

[61] Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

[62] Ripley, B.D. (1993). Statistical aspects of neural networks, in *Networks and Chaos–Statistical and Probabilistic Aspects*, O.E. Barndorff-Nielsen, J.L. Jensen, & W.S. Kendall, eds. Chapman & Hall, London, pp. 40–123.

[63] Ripley, B.D. (1994). Neural networks and related methods for classification (with discussion), *Journal of the Royal Statistical Society, Series B* **56**, 409–456.

[64] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

[65] Ripley, B.D. & Kirkland, M.D. (1990). Iterative simulation methods, *Journal of Computational and Applied Mathematics* **31**, 165–172.

[66] Shaw, J.E.H. (1988). A quasirandom approach to integration in Bayesian statistics, *Annals of Statistics* **16**, 895–914.

[67] Spanier, J. & Maize, E.H. (1994). Quasi-random methods for estimating integrals using relatively small samples, *SIAM Review* **36**, 18–44.

[68] Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L. & Cowell, R.G. (1993). Bayesian analysis in expert systems (with discussion), *Statistical Science* **8**, 219–283.

[69] Spiegelhalter, D.J., Thomas, A., Best, N.G. & Gilks, W.R. BUGS. Bayesian inference Using Gibbs Sampling. Version 0.5. MRC Biostatistics Unit, Cambridge, UK. Available from URL `http://www.mrc-bsu.cam.ac.uk` or by anonymous ftp from `ftp.mrc-bsu.cam.ac.uk`.

[70] Stewart, L. (1987). Hierarchical Bayesian analysis using Monte Carlo integration: computing posterior distributions when there are many possible models, *Statistician* **36**, 211–219.

[71] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *Journal of the Royal Statistical Society, Series B* **36**, 111–147.

[72] Swayne, D.F., Cook, D. & Buja, A. XGobi. Use WWW, anonymous ftp, or Gopher to `lib.stat.cmu.edu`.

[73] Swayne, D.F., Cook, D. & Buja, A. (1991). XGobi: interactive dynamic graphics in the X window system with a link to S, in *Proceedings of the ASA Section on Statistical Graphics*. American Statistical Association, Alexandria, pp. 1–8.

[74] Tierney, L. (1990). *LISP-STAT*. Wiley, New York.

[75] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics* **22**, 1701–1762.

[76] Venables, W.N. & Ripley, B.D. (1997). *Modern Applied Statistics with S-Plus*, 2nd Ed. Springer-Verlag, New York.

[77] Wolpert, D.H. (1992). Stacked generalization, *Neural Networks* **5**, 241–259.

B.D. RIPLEY

Computerized Therapeutic Decision Support

Introduction

A recent report by the Institute of Medicine examined some of the problems in the United States' health care system and concluded that 44 000 to 98 000 Americans die each year as a result of medical errors [5]. Although some analysts have questioned the accuracy of these numbers, most would agree that more needs to be done to reduce the number of errors. The report's recommendations specifically recognized the potential for error reduction when decision support systems are incorporated into medication-prescribing software and provider order-entry systems. Although this report brought greater attention to the importance of medical errors, we have known for some time that medical errors are a serious problem, and the notion that computer systems can improve the delivery of care and thus reduce errors is not new [9]. Because errors are such a serious problem and decision support systems are so important to their prevention, this chapter focuses on the nature of medical errors and the role of computerized therapeutic decision support in promoting patient safety.

Errors can be defined in two broad categories: errors of execution and errors of planning. Errors of planning occur when an inappropriate plan is selected for a patient. Errors of execution occur when an appropriate plan is carried out incorrectly. Errors of execution can be divided into errors of omission, in which a needed action is not performed, and errors of commission, in which the action performed is the error, either because the wrong action is performed or the correct action is performed incorrectly.

Decision Support

In the broadest sense, "computerized decision support" refers to any assistance that a computer provides when clinicians make decisions about patient care. Thus, any system that presents data in an advantageous way is a form of computerized decision support. For example, a hospital information system can be set up to display a patient's laboratory values at the same time that physicians use the computer to order the patient's medications. This setup might decrease the likelihood that physicians would prescribe a drug that could become toxic in the setting of renal insufficiency because of the prominent display of lab indicators of renal function. Designs that provide this type of assistance to clinicians are valuable, and should be an important part of any clinical system. This chapter, however, focuses on more sophisticated designs. In these designs, the computer system uses more patient information, analyzes the information in complex and sometimes novel ways, and often renders advice specific to the patient and the situation.

These sophisticated decision support systems can be understood best by examining four of their characteristics [2]:

1. Support goal refers to the overall goal of the advice being given.
2. Type of intervention describes how the clinician accesses the advice.
3. Type of knowledge refers to the information being evaluated when the advice is rendered.
4. Method of reasoning refers to the logic used by the system to interpret that knowledge and render advice.

Support Goal

One way to think of the goals of a therapeutic decision support system is to think of the system as either a state-based system or an action-oriented system.

State-based decision support systems attempt to answer the question, "What is true about this patient?" These systems synthesize multiple pieces of information in an attempt to determine diagnosis, prognosis, or specific condition along the continuum of care.

Action-oriented systems attempt to answer the question, "What should be done?" by suggesting a strategy or course of action. For example, a system may take two pieces of data, apply simple logic, and render advice, as follows:

```
IF an angiotensin-converting enzyme
(ACE) inhibitor is prescribed
      AND patient is pregnant
          THEN display
  ''ACE inhibitor is contraindicated in
pregnant patient.
   Consider discontinuing medication.''
```

Action-oriented systems can be extremely useful, especially when they have access to a wide variety of clinical information. They should be a primary focus for dealing with errors of execution when the decision support system operates in conjunction with an electronic medical record or another computer information system.

In practice, the best therapeutic decision support systems combine both of these strategies. These systems first synthesize data from multiple sources to determine the state of the patient and then use their knowledge bases to recommend an action.

Type of Intervention

The type of intervention refers to how much the user is required to do to get a recommendation. In passive systems, the user has to activate the system to get a recommendation. Passive systems often require the user to sit at a computer terminal and respond to questions so that the system can complete its knowledge base before providing a recommendation. These systems are more helpful when clinical management is complicated and requires a small number of difficult decisions, for example, for the selection of an antibiotic in patients with infection [11] and the selection of a chemotherapeutic agent in patients with cancer [4]. Passive systems are more difficult to use and require more clinician time than active forms of decision support.

Semi-active systems provide recommendations without requiring the user to activate the system, and some systems facilitate the action being recommended. For example, a semi-active system might display a reminder message advising against the renewal of an angiotensin converting enzyme ACE inhibitor for a pregnant patient and also provide a convenient way for the physician to discontinue the drug. A semi-active system, however, would never discontinue the medication without the consent of the physician.

Semi-active systems can be divided into reminder systems and alert systems. Alerts can be defined as messages that arise in response to an action. They can be extremely useful in altering errors of commission that might occur while a clinician is using a health care information system. Reminders can be defined as messages that arise spontaneously from an information system or other source. They can be useful for errors of omission where a clinician forgets to perform a needed action. Alerts and reminders can take many forms, in some cases, they would print from an information system at a predetermined time, and in other cases, they would appear on screen while the clinician was interacting with the information system.

Active systems perform actions without any need for direct participation by a clinician. For example, an active system might control a ventilator or a pacemaker. Also, active systems might operate when one action logically follows another. For example, an active system might order a serum drug level automatically when a clinician orders a drug that requires lab monitoring. Active systems promise to reduce errors for activities that require constant monitoring and adjustment, especially when the underlying mechanisms are understood and the need for adjustments are based on well-known parameters. Because medical decision-making (*see* **Decision Analysis in Diagnosis and Treatment Choice**) is so complex and poorly understood, however, active systems are little used.

Some decision support systems combine active, semi-active, and passive elements. Such systems might automatically ask for more information when a clinician orders a test or a patient's status changes and then provide a recommendation. The majority of therapeutic decision support systems, however, are semi-active systems that are used with an electronic medical record or some other type of information system.

Types of Knowledge

Decision support systems require several types of information, or "knowledge", to provide recommendations. One way of categorizing types of knowledge divides them into dynamic facts, static facts, and judgmental knowledge [14]. Dynamic facts change over time. For example, heart rate and blood pressure change rapidly in an intensive care unit and a patient's weight, drug doses, and medication list change more slowly in the outpatient setting. Other dynamic facts reflect changes in patient populations. For example, a decision support system could perform frequent statistical analyses of microbiology data to determine drug resistance patterns and then store the results as a dynamic fact.

Static facts include basic truths about medical care. "E. Coli is a gram-negative rod" is an example of a static fact.

Judgmental knowledge is a combination of formal knowledge about medicine, like the knowledge in textbooks, and experiential knowledge, such as knowing how to perform a procedure. Rules that would be contained in a decision support system typically comprise this type of knowledge. Figure 1 shows how the various types of knowledge might interact in a decision support system. In this case, the electronic medical record stores dynamic facts in the "factual database". The "knowledge base" contains static facts and judgmental knowledge in the form of rules. The "inference engine" determines how the knowledge base and the factual database combine to produce a recommendation.

For example, as the clinician orders new drugs for a patient, the decision support system stores these changes in the factual database as dynamic facts. The inference engine constantly checks dynamic facts to see whether there are related rules in the knowledge base. If the drug "coumadin" is entered into the factual database, the inference engine might identify three related rules.

1. Coumadin is warfarin (anticoagulant) (static fact)
2. IF warfarin AND amiodarone are present on the medication list, THEN display "Amiodarone may increase prothrombin time in patients also receiving warfarin. (priority 5)" (judgmental knowledge)
3. IF warfarin is on the medication list AND the international normalized ratio (INR) of the patient's prothrombin time is greater than 3.5, THEN display "INR is elevated, consider changing the dose of warfarin. (priority 7)" (judgmental knowledge).

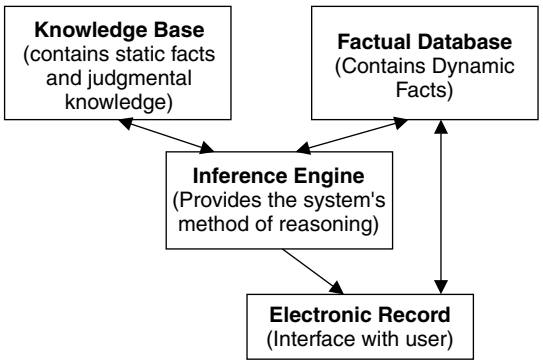

Figure 1 The relationships among components of a typical decision support system

The inference engine runs the rules in sequence and identifies that rule 1 is true but does not require action, rule 2 requires an action with a priority of 5 and rule 3 requires an action with a priority of 7. The inference engine sends a message to the electronic record to display an alert with the text for rule number 3 and a reminder with the text for rule number 2. Since rule 3 has higher priority, it is displayed first in a larger font and rule 2 is displayed second in a smaller font.

Methods of Reasoning

Therapeutic decision support systems can use several methods to process the information contained in their knowledge bases. Although the goal in a therapeutic system is always to recommend an action, the system may first determine the state of the patient in sophisticated ways prior to rendering that advice. For example, **Bayesian** analysis and belief networks can be used to determine a patient's state before the support system makes a recommendation. Most therapeutic decision support systems, however, use rule-based systems (*see* **Computer-aided Diagnosis**). The rules contain the judgmental knowledge that facilitates error reduction and most often take the following form: "IF antecedent THEN consequent." The rules in a therapeutic decision support system are typically run sequentially. The inference engine begins at the top of the list of rules and processes the list until it identifies a relevant rule. Then actions are taken as dictated by the rule. A rule might lead to the evaluation of a new set of rules, or it might initiate some action directed at the clinician, such as a reminder. This rule-based approach is especially useful for preventing errors of execution, because the rules require only a few of the thousands of loosely related bits of knowledge to give advice. For example, the presence of one drug that interacts with another drug is important by itself and does not depend on the other information in the factual database. Because rules work so well, better support systems are distinguished not so much by their rules but by how many rules activate alerts and reminders, which rules are prioritized over others, and whether the factual database is adequate for rule activation.

Evaluation of Therapeutic Decision Support Systems

Increasing use of more comprehensive health care information systems, such as an electronic medical record, has greatly facilitated the use of therapeutic decision support systems. Comprehensive health care information systems tightly integrate computerized advice with the clinician's workflow. Most analysts believe that this integration is critical to having the advice followed.

Drug Dosing and Therapy

Much of the focus for errors in medicine has been on errors in drug dosing and administration. The greatest promise of computer systems may be in decreasing and preventing these types of errors. Several studies have shown that drugs with severe side effects can be more accurately dosed using Bayesian estimates of the drug level following initiation or dose change [3, 13]. Early trials looking at the use of reminders with an electronic order-entry system showed a great deal of promise. In one study, the likelihood of changing the dose of potassium when indicated by abnormal renal function or an elevated potassium level went from 0 to 53%. The health care information systems used in these early studies, however, did not always provide advice at the ideal time and relied instead on paper printouts that were produced and distributed after the clinician had decided about drug dose. Several studies using more sophisticated healthcare information systems have shown that reminder and alert systems improve care by providing recommendations at the point of care. Dosage guidance, alternative medication selection, and prompts for order suggestion have all been shown to be facilitated by reminder systems used with an order-entry system [12]. One study, using a health care information system, compared therapeutic reminders to a "team" approach, which more heavily involved pharmacists and other providers in the drug dosing and administration process [1]. The computerized order-entry system checked for drug allergies, serious drug–drug interactions, and some drug-laboratory interactions. It prevented more than half of the serious medication errors that would otherwise have occurred if there were no intervention at all. Some of this effect was due to decision support and some due to improved legibility and process changes introduced by the order-entry system. No effect over the baseline error rate was seen for the team intervention in this study.

Preventive Care

When a reminder system that focused on preventive care (*see* **Preventive Medicine**) in the inpatient setting was coupled with an order-entry system, no significant differences were found between control physicians and test physicians [10]. When clinicians received reminders, for preventive care in outpatients, that were printed on patient encounter forms, some, but not all, activities increased in frequency [6]. For example, increases were found in fecal occult blood testing (49 to 61%, $p = 0.0007$) and mammography (47 to 54%, $p = 0.036$) but not in pap testing (18 to 21%, $p = 0.2$). The most common reason physicians gave for not performing the study was that it was "not applicable" (22.6%) to the patient because the test had been performed elsewhere (69%) [7], which emphasizes that when preventing errors of omission, the decision support system must have access to all information about a patient to be most effective.

Care Guidelines

Decision support systems for chronic diseases are similar to those for preventive care. One study that examined the use of reminders for performing recommended maintenance studies for patients with diabetes found improvements in foot exams (30 to 55%), urine protein determinations (3.9 to 73.3%), ophthalmologic exams (3 to 19%), and pneumococcal vaccinations (0 to 19.8%) [8].

Conclusion

Computerized therapeutic decision support systems offer tremendous promise in solving the problem of errors in medicine. They seem to function best when they are used in concert with a healthcare information system designed to store patient data in a machine-readable format. In these cases, relatively simple semi-active systems that use a rule-based approach to decrease the likelihood of errors in execution seem to be particularly effective. However, the culture of

medicine is slow to change and these information systems have not been adopted by the vast majority of clinicians and organizations. It will take alignment of financial incentives and a great deal more effort before patients can begin to reap the benefits of this new, decision-making process.

References

[1] Bates, D.W., Teich, J.M., Lee, J., Seger, D., Kuperman, G.J., Ma'Luf, N., Boyle, D. & Leape, L. (1999). The impact of computerized physician order entry on medication error prevention, *Journal of the American Medical Informatics Association* **6**(4), 313–321.

[2] Degoulet, P. & Fieschi, M. (1997). Medical decision support systems, *Introduction to Clinical Informatics*, Springer-verlag, New York.

[3] Gonzalez, E.R., Vanderheyden, B.A., Ornato, J.P. & Comstock, T.G. (1989). Computer-assisted optimization of aminophylline therapy in the emergency department, *American Journal of Emergency Medicine* **7**(4), 395–401.

[4] Hickam, D.H., Shortliffe, E.H., Bischoff, M.B., Scott, A.C. & Jacobs, C.D. (1985). The treatment advice of a computer-based cancer chemotherapy protocol advisor, *Annals of Internal Medicine* **103**(6), 928–936.

[5] Kohn, L.T., Corrigan, J.M. & Donaldson, M.S. (2000). Errors in healthcare, a leading cause of death and injury, *To Err is Human*. National Academy Press, Washington D.C.

[6] Litzelman, D.K., Dittus, R.S. Miller, M.E. & Tierney, W.M. (1993). Requiring physicians to respond to computerized reminders improves their compliance with preventive care protocols, *Journal of General Internal Medicine* **8**(6), 311–317.

[7] Litzelman, D.K. & Tierney, W.M. (1996). Physicians' reasons for failing to comply with computerized preventive care guidelines, *Journal of General Internal Medicine* **11**(8), 497–499.

[8] Lobach, D.F. & Hammond, W.E. (1997). Computerized decision support based on a clinical practice guideline improves compliance with care standards, *American Journal of Medicine* **102**(1), 89–98.

[9] McDonald, C.J. (1976). Protocol Based computer reminders, the quality of care and the non-perfectability of man, *New England Journal of Medicine* **295**, 1351–1355.

[10] Overhage, J.M., Tierney, W.M. & McDonald, C.J. (1996). Computer reminders to implement preventive care guidelines for hospitalized patients, *Archives of Internal Medicine* **156**(14), 1551–1556.

[11] Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C. & Cohen, S.N. (1973). An artificial intelligence program to advise physicians regarding antimicrobial therapy, *Computers and Biomedical Research* **6**(6), 544–560.

[12] Teich, J.M., Merchia, P.R., Schmiz, J.L., Kuperman, G.J., Spurr, C.D. & Bates, D.W. (2000). Effects of computerized physician order entry on prescribing practices, *Archives of Internal Medicine* **160**(18), 2741–2747.

[13] White, R.H. & Mungall, D. (1991). Outpatient management of warfarin therapy: comparison of computer-predicted dosage adjustment to skilled professional care, *Therapeutic Drug Monitoring* **13**(1), 46–50.

[14] Wraith, S.M., Aikins, J.S., Buchanan, B.G., Clancey, W.J., Davis, R., Fagan, L.M., Hannigan, J.F., Scott, A.C., Shortliffe, E.H., van Melle, W.J., Yu, V.L., Axline, S.G. & Cohen, S.N. (1976). Computerized consultation system for selection of antimicrobial therapy, *American Journal of Hospital Pharmacy* **33**(12), 1304.

Further Reading

Brennan, T.A., Leape, L.L., Laird, N.M., Hebert, L., Localio, A.R., Lawthers, A.G., Newhouse, J.P., Weiler, P.C. & Hiatt, H.H. (1991). Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard medical practice study I, *New England Journal of Medicine* **324**(6), 370–376.

ERIC PIFER, M.G. WEINER &
SANKEY V. WILLIAMS

Conception, Models for

The study of human fertility is of enduring interest to demographers who wish to forecast and characterize the dynamics of population change, to reproductive epidemiologists who wish to identify factors that adversely affect human reproduction, and to biologists who wish to improve our understanding of the fundamental processes that underlie human reproduction. The statistical models required necessarily depend on the level of detail available in the data and the scientific purpose of the modeling.

The crudest and most widely available data are on time from marriage to first birth, and time between successive births. Mathematical models for the distribution of such intervals in a population are plentiful in the **demography** literature [8, 19]. Such models account for the heterogeneity across couples in their fertility, sometimes by assuming that the monthly probability of conceiving a clinically recognized pregnancy is a property of the couple that varies across couples according to a beta distribution (*see* **Beta-binomial Distribution**).

In an industrialized society in which contraception and abortion are widely available, fertility is primarily under volitional control and, consequently, the study of such intervals carries little information about the biological capacity to reproduce. Trends, such as increases in the age at a woman's first birth and (consequent) increases in the use of infertility services, largely reflect social forces, and are of greater interest to the demographer than to the reproductive biologist.

A biologically more informative level of detail is provided by studies of the time required to achieve pregnancy in couples not using contraception. The earliest such studies recruited sexually active couples at the time at which they discontinued contraception in order to conceive [10] and followed them up to pregnancy or some maximum follow-up time.

The time-to-event data from such couples can be regarded as discrete failure-time data (*see* **Survival Analysis, Overview**), where "failure" is here a misnomer referring to the occurrence of a recognized pregnancy. While sometimes approximated as smooth for modeling [2, 3], time is discrete in this context, because each menstrual cycle provides a single opportunity to conceive. Consequently, time-to-pregnancy is best measured in integer units based on the number of menstrual cycles from discontinuation of contraception to the achievement of a recognized pregnancy. If reported in units of calendar time, these intervals can be converted to numbers of menstrual cycles, by dividing by the woman's usual cycle length. If all couples had the same constant probability of conceiving per menstrual cycle ("fecundability"), then the distribution of times to pregnancy would be **geometric**.

In practice, however, these intervals are **overdispersed** compared to the geometric, reflecting underlying heterogeneity in fecundability across couples in the population. In the first menstrual cycle at risk, about a third of couples conceive; in the second cycle, about a fourth, and so on. The cycle-specific conception rate continues to decline over time. This is not an effect of time *per se*, but reflects sorting within a heterogeneous population of couples, where the couples with the highest fecundability conceive early and are not present in subsequent **risk sets**. The pattern is not simply due to the presence of a sterile subpopulation, but is evident even among those couples who ultimately do conceive. Following the modeling traditions established in demography, Weinberg & Gladen [11] modeled time-to-pregnancy data according to a beta-geometric, by assuming that each couple has a characteristic fecundability parameter, drawn from a beta distribution.

Parameter estimation is straightforward, because the resulting beta-geometric distribution can be expressed as **generalized linear model**. If p_j denotes the conception rate at cycle j, among couples still at risk, then one can show that

$$\frac{1}{p_j} = c + d(j-1),$$

and hence the maximum likelihood estimates for the beta parameters can be developed using standard software. The parameter, c, is interpretable as the inverse of the mean fecundability across couples, while d depends on the beta variance. One can also allow for a subpopulation of completely sterile couples by means of the **EM algorithm** [4], where the underlying beta distribution for fecundability is now contaminated by a degenerate distribution with mass at fecundability 0 [11]. Thus, relative to this parametric model, one can estimate the prevalence of sterility in a population.

One could also model the underlying heterogeneity distribution nonparametrically, since in general

the cycle-specific conception rates depend in a simple way on the moments of the underlying distribution, as described by Weinberg & Gladen [11]. With a prospective study of cycles up to at most K, the cycle-specific conception rates can be shown to depend on the first K **moments** of the fecundability distribution. Because the number of parameters (moments) to be estimated can be quite large, and the alternative approach is based on a rich family of density functions, the beta-geometric will often be preferred to the nonparametric approach in practice. Also, the natural linear extension of the above generalized linear model allows effects of covariates to be incorporated in the beta-geometric approach.

Because the beta-geometric model does not yield any summary measure of the effect of an exposure that has much intuitive appeal, investigators have turned to other modeling approaches. One can apply the model that Cox suggested for **discrete survival time** data [5], where each cycle's binary outcome is modeled as logit-linear and a cycle-specific baseline serves as the discrete-time analog of the baseline hazard function for continuous failure-time data.

A model that is **loglinear** in covariates can also be fitted [14, 16]. With this model, the effect of a dichotomous exposure is assumed to be multiplicative on each cycle-specific conception probability. The exponentiated coefficient is interpretable as a "fecundability ratio," analogous to a **hazard ratio**, except that it is a ratio of the probability of conception in a cycle for those exposed divided by the probability for those unexposed. Adjustments for **confounding** factors are easily included. Because the outcome here is not rare, the fecundability **odds ratio** estimated in the discrete-time Cox model cannot be seen as an approximation to the fecundability ratio estimated in the loglinear model. The loglinear model can be fitted using standard generalized linear model software, provided that the model incorporates a baseline conception-rate parameter for each cycle number. One complication with using the loglinear model is that the predictive linear function can sometimes exceed zero, implying a probability above 1.0. The logistic formulation is less intuitive to the scientist but avoids this annoying pitfall.

Time to pregnancy can be ascertained either prospectively or retrospectively. In a prospective study, couples are entered either at the time at which they discontinue contraception, or at some point in the middle of their attempt. In the latter case, the left **truncation** must be taken into account by ascertaining the number of months at risk prior to study entry, and delay-entering the couple into the appropriate risk set (cycle number). Thus a couple who have been trying to conceive for six months prior to recruitment into the study would not be credited for their first six failures, but would be entered into their first risk set at cycle seven. Assuming that late entry is not related to fecundability, allowing such couples into the study should produce no bias.

In the retrospective approach, couples are asked about time to pregnancy, based on reconstructing their history of contraceptive use. If every attempt at pregnancy ended in conception, the retrospective design would yield data equivalent to that from a prospective design. In practice, populations include couples with very low or zero fecundability. Such couples contribute to prospective studies but are under-represented in retrospective studies based on achieved conceptions.

Retrospective studies based on a current pregnancy or the most recent pregnancy are also prone to bias in estimating effects of exposures, if the prevalence of those exposures has changed systematically over calendar time (*see* **Bias in Observational Studies**). This is because the couples who required a long time to conceive are reporting exposures that took place long ago, when the opportunity for exposure might have been very different from what it was more recently. Weinberg et al. [13] describe an instance of such bias, where the use of latex gloves by dental assistants was shown (spuriously, we presume) to enhance their fertility. This arose because dental workers wore gloves more often after the AIDS epidemic began, and women who conceived quickly were more likely to have begun their attempt recently, after glove use had become widespread.

A less bias-prone and more detailed approach to studying fertility is provided by a **cohort study** where not only menstrual cycles and pregnancies are recorded, but some marker for the day of ovulation is available for each cycle, together with daily data indicating whether there was unprotected intercourse. Markers for ovulation can be based on hormonal assays of excreted hormones, or on daily records of basal body temperatures. When both intercourse data and a benchmark for ovulation are available, one can ask some interesting questions about reproductive biology. What is the relationship between the timing of intercourse *vis-à-vis* ovulation and the probability

of conception? For how many days in a month is a woman fertile? Do the Y-bearing (boy-producing) sperm and the X-bearing (girl-producing) sperm have similar survival and potency, or is there a relationship between the timing of intercourse and the sex of the baby?

The first extensive data of this sort were described by Barrett & Marshall [1], who studied couples using natural family planning (the rhythm method). Basing the identification of day of ovulation on the rise in basal body temperature that accompanies ovulation, Barrett & Marshall applied a model asserting that the cohorts of sperm introduced to the woman's reproductive tract on different days pose independent **competing risks** (of fertilization) to the ovum. Under this simple independence model, the probability of conception can be written as

$$\Pr[\text{conception in cycle } j \,|\, \text{intercourse pattern}] = 1 - \prod_k (1 - p_k)^{X_{j,k}},$$

where $X_{j,k}$ is an indicator that has a unit value if there was intercourse on day k in cycle j. The indexing specified by k is relative to the day of ovulation, which is usually taken to be day "zero". The parameters p_k are interpretable as the probability that conception would result had there been intercourse only on day k.

Schwartz et al. [7] modified this model in a way that allows for the fact that timing is not everything: a broad constellation of factors must be favorable for conception even to be possible in a given cycle. The ovum must mature properly and be viable, and must be transported through the oviduct; the uterine endometrium must be adequately prepared under hormonal stimulation; the immune system must function properly and not reject the embryonic foreign tissue, and so on. The constellation of such factors is sometimes referred to [15] as "cycle viability," and under the Schwartz et al. model specification, each cycle is either viable or not. Without cycle viability, conception will not occur, regardless of the timing of intercourse. If we denote the probability that the cycle is viable by A, then the model becomes

$$\Pr[\text{conception in cycle } j \,|\, \text{intercourse pattern}] = A \left\{ 1 - \prod_k (1 - p_k)^{X_{j,k}} \right\}.$$

When this model was fitted to data from a cohort study carried out in North Carolina [18], the p_k parameters exceeded zero only over a six-day interval, ending on the estimated day of ovulation [17] and the A parameter was estimated to be 0.37, suggesting that more than half of apparently ovulatory menstrual cycles in healthy women of reproductive age are nonviable.

In fact, as was recognized by Schwartz et al., the A parameter requires a broader interpretation than suggested by "cycle viability". Fertility studies do not detect all conceptions, only those that survive to the point at which the methods applied are capable of recognizing that pregnancy has begun. Studies in which only clinically recognized pregnancies are found in effect include as part of A an additional factor corresponding to the probability that the pregnancy survives up to clinical detectability. Even studies that make use of a very sensitive assay for the pregnancy hormone, hCG, are only detecting conceptions that survived long enough to successfully implant and establish communication with the maternal circulation. While the survival probability is embedded within A and is not statistically identifiable under this model, it is important to recognize that A can incorporate effects of male factors. For example, if husbands are exposed to a mutagen and produce some sperm that are capable of fertilizing a viable ovum but do not produce a viable embryo, then the A parameter would be reduced accordingly.

This means that the models to be discussed that allow the parameters A and p_k to depend on covariates should allow for the possibility that male exposures can affect both A and the p_k. Similarly, female exposures could theoretically affect both A and the p_k. The woman provides the environment in which the sperm must temporarily live before encountering the ovum, so an exposure that renders the female reproductive tract hostile to sperm survival could reduce the p_k, especially for days prior to ovulation. It is also possible that the pattern of intercourse itself could affect A, if certain patterns are associated with aging of the gametes and consequent effects on viability of the conceptus.

Schwartz et al. [7], and later Royston [6], who reanalyzed the same data, allowed for the possibility that A could decline with advancing maternal age. Weinberg et al. [15] developed methods for fitting more complex models, by using the EM algorithm [4] to model cycle viability as an unobservable **Bernoulli**

outcome, whose probability can depend on **covariates** through a generalized linear model.

Parametric models that explicitly estimate parameters related to survival of the sperm and ova have also been proposed [6, 12]. The model developed by Weinberg & Wilcox specifies that the instantaneous probability (hazard) for fertilization at time t (where $t = 0$ at the moment of ovulation), given that the ovum is still viable at t, is proportional to the number of sperm that are still viable at time t. The competing risks for fertilization due to batches of sperm that were introduced on different days are still treated as independent, so that the surviving sperm from the various batches simply commingle. The viable lifetime of individual sperm is assumed to be **exponential**, while ovum survival is taken as fixed. Applying this model to data from the North Carolina study, the mean viable lifetime for sperm was estimated to be 1.4 days, while the ovum appears to survive for less than a day. While it fits the available data very well, this model oversimplifies the underlying biology, because it does not take into account female factors, most notably changes in the cervical mucus that can, depending on the day of the cycle, alternately impede or facilitate the entry of sperm into the upper female reproductive tract.

These variations on the model proposed by Schwartz et al. all have certain doubtful assumptions in common. They assume that sperm introduced on different days present independent risks of fertilization to the ovum. They assume that the time between successive acts of intercourse has no effect on the potency of the second batch of sperm. They assume that the pattern of intercourse has no bearing on the survival-to-detection probability for the embryo, so that embryos formed from relatively aged gametes are not of reduced inherent viability. Finally, they assume that the outcomes for successive menstrual cycles within a couple are independent. Of these, the latter assumption is the most easily demonstrated to be false [22].

Subsequent work has allowed the dependency among successive outcomes from a single couple to be handled by a **generalized estimating equation** (GEE) approach. The estimating equations are taken to be the likelihood equations, so the estimated effects are not modified, but the standard errors are adjusted, usually upward, to properly reflect the dependencies in the data. The extended model now also allows for exposures that may vary from day to day and directly influence the day-specific p_k [20].

The capacity to handle day-specific exposures means that the model can now be used to assess certain direct effects on fertility. One example is in assessment of contraceptive efficacy, where one now would not need to exclude menstrual cycles in which the method being assessed, for example the female condom, was not used for every act of intercourse. Another example is in trying to assess the effect of day-specific characteristics of the cervical mucus on the likelihood of conception.

The heterogeneity among couples is treated as a nuisance by the GEE approach, whereas this variability may be of interest in itself. Accordingly, a subject-specific approach was also developed, in which the cycle viability probability is now taken to be a couple-specific parameter and these are assumed to have been sampled from a beta distribution [22]. The existence of significant heterogeneity in cycle viability among couples demonstrates that the heterogeneity in fecundability long known to demographers is not simply secondary to variation in the frequency and patterns of intercourse, but reflects determinants that are more biologically innate.

The contexts discussed so far have included three levels of detail available in studies of human fertility: that provided by the intervals between marriage and first birth, or between successive births; that provided by studying the number of menstrual cycles from discontinuance of contraception to the onset of a detectable pregnancy; and that provided by daily hormonal assays and intercourse records, allowing the investigator to match up the pattern of intercourse to the time of ovulation itself. An even more refined level of detail is provided by certain clinical protocols used to treat infertile couples.

In *in vitro* fertilization (IVF), the woman's ovaries are hyperstimulated by exogenous hormones to produce many ova at once. When mature, these ova are withdrawn surgically from her ovaries and then fertilized *in vitro*, usually by her husband's sperm. A selection of the resulting embryos is then transferred to her uterus in hopes that pregnancy will ensue. A model closely related to that proposed by Schwartz et al. can be applied to assess the effect of exposures and clinical markers on the likelihood of conception in IVF. The occurrence of conception depends on a susceptibility factor, the receptivity of the uterus to implantation, together with

an aggregation of Bernoulli trials: at least one of the transferred embryos must be viable. The resulting model, an analog of the Schwartz et al. model described above, was proposed by Speirs et al. [9], and more recently reconsidered by Zhou & Weinberg [21].

References

[1] Barrett, J.C. & Marshall, J. (1969). The risk of conception on different days of the menstrual cycle, *Population Studies* **23**, 455–461.

[2] Boldsen, J.L. & Schaumburg, I. (1990). Time to pregnancy – a model and its application, *Journal of Biosocial Science* **22**, 255–262.

[3] Bolumar, F., Olsen, J., Boldsen, F. and the European Study Group on Infertility and Subfecundity (1996). Smoking reduces fecundity: a European multicenter study on infertility and subfecundity, *American Journal of Epidemiology* **143**, 578–587.

[4] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[5] Kalbfleisch, J.D. & Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model, *Biometrika* **60**, 267–278.

[6] Royston, J.P. (1982). Basal body temperature, ovulation and the risk of conception, with special reference to the lifetimes of sperm and egg, *Biometrics* **38**, 397–406.

[7] Schwartz, D., MacDonald, P.D.M. & Heuchel, V. (1980). Fecundability, coital frequency and the viability of ova, *Population Studies* **34**, 397–400.

[8] Sheps, M.C. & Mencken, J.A. (1973). *Mathematical Models of Conception and Birth*. University of Chicago Press, Chicago.

[9] Speirs, A.L., Lopata, A., Gronow, M.J., Kellow, G.N. & Johnston, W.I.H. (1983). Analysis of the benefits and risks of multiple embryo transfer, *Fertility and Sterility* **39**, 468–471.

[10] Tietze, C. (1968). Fertility after discontinuation of intrauterine and oral contraception, *International Journal of Fertility* **13**, 385–389.

[11] Weinberg, C.R. & Gladen, B.C. (1986). The beta-geometric distribution applied to comparative fecundability studies, *Biometrics* **42**, 547–560.

[12] Weinberg, C.R. & Wilcox, A.J. (1995). A model for estimating the potency and survival of human gametes *in vivo*, *Biometrics* **51**, 405–412.

[13] Weinberg, C.R., Baird, D.D. & Rowland, A. (1992). Pitfalls inherent in retrospective time-to-event data: the example of time to pregnancy, *Statistics in Medicine* **12**, 867–879.

[14] Weinberg, C.R., Baird, D.D. & Wilcox, A.J. (1994). Sources of bias in studies of time to pregnancy, *Statistics in Medicine* **13**, 671–681.

[15] Weinberg, C.R., Gladen, B.C. & Wilcox, A.J. (1994). Models relating the timing of intercourse to the probability of conception and the sex of the baby, *Biometrics* **50**, 358–367.

[16] Wilcox, A.J., Weinberg, C.R. & Baird, D.D. (1988). Caffeinated beverages and decreased fertility, *Lancet* **2**, 1453–1456.

[17] Wilcox, A.J., Weinberg, C.R. & Baird, D.D. (1996). Timing of sexual intercourse in relation to ovulation – effects on the probability of conception, survival of the pregnancy, and sex of the baby, *New England Journal of Medicine* **333**, 1517–1521.

[18] Wilcox, A.J., Weinberg, C.R., O'Connor, J.F., Baird, D.D., Schlatterer, J.P., Canfield, R.E., Armstrong, E.G. & Nisula, B.C. (1988). Incidence of early pregnancy loss, *New England Journal of Medicine* **319**, 189–194.

[19] Wood, J.W. (1994). *Dynamics of Human Reproduction: Biology, Biometry, Demography*. Aldine De Gruyter, New York.

[20] Zhou, H. & Weinberg, C.R. (1996). Modeling conception as an aggregated Bernoulli outcome with latent variables, via the EM algorithm, *Biometrics* **52**, 945–954.

[21] Zhou, H. & Weinberg, C.R. (1998). Statistical methods for evaluating effects of exposures and clinical factors on embryo viability and uterine receptivity in *in vitro* fertilization, *Biometrics*, in press.

[22] Zhou, H., Weinberg, C.R., Wilcox, A.J. & Baird, D.D. (1996). A random effects model for cycle viability in fertility studies, *Journal of the American Statistical Association* **91**, 1413–1422.

CLARICE R. WEINBERG

Conception

Any attempt to count the numbers of pregnancies conceived in a population and then use these figures to derive conception rates is bound to be incomplete. Clearly, the pregnancies that end very early before the woman realizes she is pregnant will be missing. Those that end before help is sought from a midwife or doctor are also likely to be left out. Once maternity care has started, documents confirming pregnancy are produced for various official purposes, such as to prove entitlement for maternity benefit payments, but these documents do not usually end up in **vital statistics** systems.

Nevertheless, the idea of trying to count pregnancies directly through a notification system is not a new one. For example in Huddersfield, England, in 1916 a system was introduced by which the doctor or midwife booked by the woman for delivery notified the pregnancy to the Public Health department. By 1934, the proportion of pregnancies notified had reached 77% [1].

The other approach is to estimate the numbers of conceptions indirectly through data collection systems designed to count the outcomes of pregnancy in terms of miscarriage, legal abortion, and registrable birth.

In most countries, the biggest gap is in statistics about miscarriage. Miscarriages once a pregnancy is well established are likely to lead to hospital admission and thus be counted in hospital inpatient statistics. Earlier miscarriages may lead to care outside hospital by general or family practitioners whose work is less likely to appear in national systems. If these miscarriages are reported, then there may be double counting in cases where the woman is subsequently referred to hospital. Most countries have routine data about birth and legal abortion, from which estimates of conceptions can be derived, however. This can be difficult if the records do not include the **gestational age** at which the events occurred.

In England and Wales, estimates are made of the numbers of conceptions leading to either a legal abortion under the 1967 Act or a maternity with one or more registrable live or still births [3]. These estimates are then used to derive age-specific rates per thousand women. Because of the lack of adequate data, pregnancies leading to miscarriages or other outcomes are not included. In Scotland, data about miscarriages in hospital are combined with those about birth and legal abortion to derive estimated teenage conception rates [1].

Gestational age is stated on notifications of abortion, and from 1974–80, the date of the last menstrual period was also included. In deriving the estimated date of conception, it is assumed that the gestational age is determined from the date of the last menstrual period and that, on average, conception takes place 14 days after this. Gestational age is stated on registrations of fetal deaths (*see* **Vital Statistics, Overview**) as stillbirths, and the stated gestational age is used to derive the estimated date of conception.

Gestational age is not recorded on registrations of live births in England and Wales. The date of conception is estimated to have been 38 weeks before the date of birth, on the assumption that the average time between the first day of the last menstrual period and the date of birth is 40 weeks. This can lead to **bias** when tabulating conceptions according to age and other characteristics of the mother, as it assumes that women having preterm births are broadly similar to childbearing women as a whole. This is not the case.

References

[1] ISD Online. Sexual and reproductive health. Teenage pregnancy. ISP, Edinburgh. `http://www.show.scot.nhs.uk/isd/sexual_health/Teenpregs/Teenpregs_homepage.htm`. Accessed 29/10/02.

[2] Ministry of Health (1937). *Report on an Investigation Into Maternal Mortality*, Cmd 5422. HMSO, London.

[3] Office for National Statistics (2001). *Birth Statistics*, Review of the Registrar General on births and patterns of family building in England and Wales, 2000, Series FM1, No. 29, office for National Statistics, London.

ALISON MACFARLANE

Conditional Probability

Conditional probability has been one of the least understood and most controversial concepts in the history of science. It was introduced by **Thomas Bayes** in 1764, who defined, motivated, and applied this concept much as is done today in post-calculus, but not measure-theoretic, courses in probability theory. For a finite space of possible outcomes, the conditional probability $\Pr(E|F)$ is defined as

$$\Pr(E|F) = \frac{\Pr(E \wedge F)}{\Pr(F)},$$

provided that $\Pr(F) > 0$. (Here "$\wedge$" means "and".) This definition is not arbitrary, but is requisite in both the **Bayesian** and classical frequentist theories of **probability**, in order that such theories do not lead to absurd results.

First consider the classical frequentist approach that stemmed from the work of Jakob Bernoulli (*see* **Bernoulli Family**). In repeated trials, such as arise in games of chance, by the **law of large numbers** it follows that amongst the cases in which the event F occurs, there will be a limiting proportion $\Pr(E|F)$ of cases in which also E occurs, provided that $\Pr(F) > 0$. For a specified event A and sequence of trials on each of which A may or may not occur, define $f_N(A)$ to be the number of times that A occurs in the first N trials. Now suppose that E and $E \wedge F$ are such events, and consider a sequence of independent trials concerning these events, always with the same fixed probability distribution. Then the proportion of times in which the event F occurs amongst the first N trials converges almost surely to its probability $\Pr(F)$, which we assume to be positive (*see* **Convergence in Distribution and in Probability**). The relative frequency of cases in which both E and F occur, amongst those cases in which F occurs, in the first N trials, is

$$\frac{f_N(E \wedge F)}{f_N(F)} = \frac{\left[\dfrac{f_N(E \wedge F)}{N}\right]}{\left[\dfrac{f_N(F)}{N}\right]},$$

with $f_N(F) > 0$ for sufficiently large N. Hence in the frequentist theory, as $N \to \infty$ this relative frequency converges almost certainly to $\Pr(E \wedge F)/\Pr(F)$.

Next, to justify his definition of conditional probability, Bayes had already presented a version of the fundamental coherency argument in terms of a called-off gamble on the event E given F, which was later to be developed in great detail by **de Finetti**. A coherency theorem of de Finetti (Theorem 1 below) proves that unless the value of a called-off gamble is assessed in accord with the conventional $\Pr(E|F)$, a person who bets on all such gambles according to the probability assessments he has asserted can be made a sure loser.

De Finetti's theory of coherence was developed as a consequence of the basic notion that probabilities, if acted upon, should not give rise to certain loss. Any contract concerning an unknown event or variable is called a gamble, so that the taking out of insurance, and indeed investments of any sort, will here be referred to as gambles, without the usual negative connotation for highly speculative activities. A gamble that gives rise to certain loss, no matter what actually occurs, is traditionally called a Dutch book. (More recently, such gambles are discussed in terms of "arbitrage".) To be precise, by a "Dutch book" is here meant a finite collection of gambles and conditional gambles such that no matter how the individual events turn out, whether true or false, one is logically certain to lose a positive amount greater than some $\varepsilon > 0$. The primary result in the theory of coherence is de Finetti's theorem giving necessary and sufficient conditions to avoid a Dutch book. The context for de Finetti's theorem concerns a set of simple gambles and simple conditional gambles. By a simple gamble **G** we mean the following. There exists an event E, which will be verified to be either true (1) or false (0), and the gamble consists in a contract under which if E occurs (or is true) one receives a specified monetary stake S, while if E does not occur (or is false) one receives nothing. We define $\Pr(E)$ to be the price at which a particular person evaluates the worth of this gamble, in the sense that the person would either buy or sell the gamble at the price $\Pr(E) \times S$. If S is positive, and if one is to avoid sure loss, then plainly the gamble has some nonnegative value between 0 and S. It is customary to choose small monetary values for S to aid in the evaluation of $\Pr(E)$, so that one is not overly influenced by considerations that arise when dealing with large sums of money, and which can be dealt with by the theory of **utility**, as for example by F.P. Ramsey [16] and by **L.J. Savage** [19].

Next, a simple conditional gamble concerning the event E given the event F, which is written as (**E**|**F**), is a gamble under which one receives the stake S if both E and F occur, one receives nothing if F but not E occurs, and the gamble is called off if F does not occur. If S is positive then such a gamble has again some nonnegative value, say $p \times S$, which is the price at which one evaluates the worth of the conditional gamble, again in the sense that one would either buy or sell the conditional gamble at this price. It is understood that if F does not occur, then this price is returned. The following theorem, due to de Finetti [4, p. 109], shows that the p obtained in this way for the conditional gamble must be precisely $p = \Pr(E \wedge F)/\Pr(F)$, if sure loss is to be avoided.

Theorem 1. For simple gambles on events $E \wedge F$ and on F, and a simple conditional gamble (**E**|**F**), to avoid a Dutch book it is necessary and sufficient that $\Pr(E \wedge F) = p \times \Pr(F)$, with $0 \leq \Pr(E \wedge F) \leq \Pr(F) \leq 1$. In this case $0 \leq p \leq 1$ whenever $\Pr(F) > 0$.

The proof of this theorem is obtained by considering the payoff for simultaneous bets on each of $E \wedge F$, F, and the conditional gamble (**E**|**F**). The avoidance of a Dutch book is equivalent to the singularity of the matrix that represents the payoff on these three bets as a function of the separate stakes on each bet. It is worth noting that when $\Pr(F) = 0$ the matrix is necessarily singular, and so a conditional probability given an event of probability 0 can be evaluated arbitrarily, without giving rise to sure loss. Thus, in both the de Finetti coherency theory and the classical frequentist theory, standard arguments for the conventional definition of a conditional probability do not apply when the event F has probability 0. When there are only a finite number of possible outcomes, of course, no one takes seriously the possibility that an event of probability 0 might occur. On the other hand, when there are a nonfinite number of possible outcomes, some delicate issues arise for all approaches, which will be discussed below.

Theorem 2 combines related results of de Finetti [4; 5, p. 111] into a single theorem.

Theorem 2. Let $\mathcal{W}$ be a finite space of points w_i, for $i = 1, \ldots, N$. Suppose that a nonnegative function $\Pr(E|F)$ is defined for some pairs of subsets of $\mathcal{W}$ with $F \neq \emptyset$. Suppose further that this function is used to determine prices for conditional gambles, with $\Pr(E|F)$ the price for the conditional gamble (**E**|**F**) with stake unity; and that when $F = \mathcal{W}$, we define $\Pr(E) \equiv \Pr(E|\mathcal{W})$ to be the price for the unconditional gamble on E. Then in order that there be no Dutch book possible on the collection of conditional and unconditional gambles for which prices have already been specified, it is necessary and sufficient that the already specified $\Pr(E|F)$ can be extended to a probability distribution π on $\mathcal{W}$, such that for all E and H with $\pi(H) \neq 0$ we have

$$\Pr(E|H) = \frac{\pi(E \wedge H)}{\pi(H)}.$$

Now let $\mathcal{W}$ be any finite space of outcomes or points. Suppose that one specifies prices for some simple gambles and conditional gambles involving the events of $\mathcal{W}$. Then it follows from the theorem that either one is already subject to a Dutch book based upon these gambles, or else one can extend the original specification to a probability distribution π on all of $\mathcal{W}$. In the latter case, if one uses **Bayes' theorem** to obtain posterior probabilities in any such extension, then the theorem guarantees coherency (the impossibility of a Dutch book) within this framework; conversely, if the original specifications violate Bayes's theorem, in the sense that they are not consistent with *any* Bayesian analysis, one can always be made subject to a Dutch book no matter what the actual outcomes.

In this century, a new understanding of conditional probability arose from the Borel – von Neumann – Wald theory of games and statistical decision functions (*see* **Decision Theory**), which we will refer to as BNW theory. In this context one considers mappings of the data into a space of terminal actions. As will be seen, both the coherency theory of Bayes and de Finetti, as well as the classical frequentist theory of conditional probability, can be expressed in decision-theoretic terms. Since even in logic the interpretation of a conditional statement (such as a counterfactual) is controversial, it is important to provide an operational meaning for conditional probability statements that makes it clear what is gained or lost by various methods for specifying such conditional probabilities. Otherwise, the theory would be largely arbitrary and all assessments would be subject to controversy and doubt. Fortunately, this can be done both simply and forcefully, within the BNW framework, both for conventional decision problems concerning an unknown

parameter, and for prediction problems. In this framework it becomes clear that only decision procedures that are based upon conditional probability cannot be ruled out as objectively defective.

Conditional Probability and Decision Theory

Beginning with the **Neyman–Pearson lemma**, it has been argued that a procedure that has a risk function that can be decreased in some or all components of risk, without increasing other components of risk, is undesirable. For example, in choosing to minimize the type two error probability β of a simple versus simple **hypothesis test**, for a specified type one error probability α, one is implicitly replacing a particular risk function by one that is generally regarded as better in an objective sense. Based upon the theory of games and statistical decision functions, it is shown below that any real-world decision procedure that is not based upon conditional probability is objectively defective in its performance in precisely the same sense as for such type one and two errors.

In decision theory, it is conventional to introduce a space of terminal actions $\mathcal{A}$, a parameter θ, and a loss function $L(\theta, a) \geq 0$, which specifies the loss when action a is taken and θ is the true value of the unknown. A probability model for the data X is specified, which depends only upon the value of θ. If $\mathcal{X}$ is the space of possible data observations, then a pure decision rule d is a mapping from $\mathcal{X}$ into $\mathcal{A}$. A randomized decision rule is a finite mixture of the pure decision rules; for example, $\delta(X) = \sum_{i=1}^{J} [\alpha_j] d_j(X)$ is the randomized decision rule that takes the decision specified by pure decision rule d_j with probability α_j. Pure decision rules are identified with such degenerate probability distributions. The space of randomized decision rules consists of all such finite mixtures of the pure decision rules. The performance of a particular randomized decision rule δ is measured by its risk function

$$R_\delta(\theta) = \mathcal{E}\, L[\theta, \delta(X)],$$

which is the expected loss for δ when θ is the value of the parameter or unknown. The value of $R_\delta(\theta)$ at a particular θ is known as a component of risk for δ.

A procedure is said to be admissible if there is no other available procedure with risk at least as small for all θ and strictly smaller somewhere. A procedure is said to be extended admissible if there is no other decision procedure available that has uniformly smaller risk by some $\varepsilon > 0$. The collection Γ of risk functions for all available randomized decision procedures is known as the risk set, and consists of all mixtures $\sum_{j=1}^{J} \alpha_j R_{d_j}(\theta)$ of the risk functions of the pure decision rules d_j. This set is the convex hull of the risk functions for pure decision rules.

A decision procedure δ is said to be Bayes with respect to a **prior distribution** π for θ if its risk function $R_\delta(\theta)$ is such that

$$\int R_\delta(\theta)\, \pi(\mathrm{d}\theta) \leq \int R_{\delta_1}(\theta)\, \pi(\mathrm{d}\theta)$$

for all other available decision rules δ_1. In other words, a procedure δ is Bayes if for some a priori probability distribution π for θ, no other procedure has smaller expected risk when θ has distribution π. The Bayes risk of a decision procedure δ when π is the a priori distribution is by definition

$$B(\pi, \delta) = \int R_\delta(\theta)\, \pi(\mathrm{d}\theta).$$

The Bayes boundary of the risk set consists of all those risk functions for which no uniform improvement is possible; that is, no improvement by some $\varepsilon > 0$ uniformly in the parameter space.

Next, if a procedure is to be appropriate for real-world applications, in the sense of being implementable on a computer with finite memory, it is necessary that both the data space $\mathcal{X}$ and the action space $\mathcal{A}$ be finite. Even if the original data space were of infinite cardinality, it would be necessary to finitize it in order to put all possible observations into such a computer. Such finitization procedures are of course quite customary, even with integer data, such as time to death as measured to the nearest year, where one puts in an upper bound, for example so that all deaths beyond 200 years are lumped into a single category. Similarly, if time of death is measured in terms of fractions of years, it is customary to round these also, both because no one is seriously interested in measuring such times to death too finely (such as 67.3487532137 years), and also because even if it were possible to do so in a meaningful sense, it would be impossible either to measure or record in a computer too many such decimal points.

In real-world decision problems, the action space must also be finite. No decision maker seriously

contemplates taking more than a finite number of actions. For example, if the problem is to forecast an interest rate a year from now, then real-world forecasts are not given to more than a few decimal points. With regard to the cardinality of the parameter space, in typical real-world problems the parameter too is rounded, and typically loses its meaning beyond a certain known finite number of decimal points. For example, the weight of a whale changes nontrivially whenever the whale spouts, and the height of a person changes during the course of a day. Hence, it is meaningless to define such parameters to too great precision. Only in certain (typically exotic) problems arising in the physical sciences can parameters be taken seriously to many decimal places, and even here the uncertainty principle of quantum mechanics suggests limitations on the ability to measure such parameters.

In the case of a finite data space and a finite parameter space, the conditional probabilities for θ given $X = x$ are well defined, and it is easy to show that the Bayes procedures are those that are equivalent to first updating the initial distribution π to a posterior distribution π^* by using Bayes's theorem, and then choosing a terminal action that minimizes the expected loss with respect to this posterior distribution: see Savage [19, Chapter 3] or DeGroot [7, p. 138]. In the finite case there always exists a pure decision rule δ_π that is a Bayes decision rule for π.

When the parameter space is not finite, there are some technical issues to mention. First, in the original **Kolmogorov** theory, it is conventional to allow only countably additive probability distributions for θ. Such countably additive distributions are necessarily finitely additive, in the sense that the probability of a finite union of disjoint events is the sum of probabilities. However, in recent years it has become understood that the collection of finitely additive distributions, which is a much larger collection, is also of importance. (Indeed, improper prior distributions, such as are widely used by Bayesian statisticians after the fashion of **H. Jeffreys** [12], and implicitly used by some non-Bayesians, can be rigorously interpreted as finitely additive distributions.) By a finitely additive distribution, we mean one in which additivity of probabilities is required to hold for all *finite* unions of disjoint events, but not necessarily for nonfinite unions. If the parameter has only a finite number of values, then finite additivity and countable additivity are equivalent. To say that a procedure is Bayes with respect to a finitely additive probability distribution π means the same as previously, in terms of minimization of expected risk with respect to that distribution, except that now the distributions may be only finitely additive. It is easy to show that the space of all possible finitely additive distributions is equivalent to the space of all nonnegative linear functionals $\pi(f)$ defined on the collection of bounded functions f of the parameter θ. The probability of a subset A of the parameter space is simply the value of the functional π at the indicator of the set A. See Kolmogorov & Fomin [15] for an elegant presentation of the theory of such linear functionals.

The next theorem proves that any procedure that is not based upon conditional probability is objectively defective. This is meant in precisely the same sense as in the Neyman–Pearson lemma, where it is foolish not to minimize the type two error probability for a specified type one error probability. This theorem is a slight strengthening of Theorem 1 of Hill [9] to allow for a possibly infinite parameter space, and the proof is essentially the same. When the parameter space is finite the theorem holds for any loss function whatsoever.

Theorem 3. Suppose that the space of terminal actions $\mathcal{A}$ and the data space $\mathcal{X}$ are finite, with the parameter space arbitrary, and the loss function nonnegative and bounded. Let $\mathcal{D}$ be the class of randomized decision rules of the form $\delta(X) = \sum_{i=1}^{J} [\alpha_j] d_j(X)$. If a decision rule $\delta_0 \in \mathcal{D}$ is not Bayes with respect to some finitely additive a priori distribution, then it can be improved upon, uniformly in the parameter, by an admissible and computable Bayes procedure in $\mathcal{D}$.

The proof of this theorem is obtained by using the fact that under our assumptions any non-Bayes procedure δ_0 can be uniformly improved upon by some other available procedure, say δ_1, not necessarily a Bayes procedure. Now consider the restricted decision problem, in which only decision rules at least as good in risk as δ_1 are considered, and find a Bayes procedure δ_π in this restricted problem for some prior distribution π. Any such Bayes procedure has a risk function less than or equal to that of δ_1 for all θ, and so is uniformly better than δ_0. It is straightforward to show that π can always be chosen so that δ_π is admissible both in the restricted and the original decision problem.

This theorem shows that any real-world non-Bayes procedure can always be improved upon uniformly in the parameter by some positive amount. This is not merely a theoretic possibility, but Theorem 3 suggests a concrete **algorithm** for obtaining such improvements. Indeed, they can be routinely provided by means of existing computational methods for solving **linear programming** problems. The equivalence of the above minimization problem with those arising in linear programming follows from the fact that both problems can be formulated mathematically in terms of the minimization of an inner product of a fixed vector π with a variable vector γ that lies in a known closed convex set. In our problem the vector π is the a priori distribution, while in linear programming problems it is known as the objective function. It should be noted that when the non-Bayes procedure δ_0 is close to the Bayes boundary, then the improvement although uniform is small. On the other hand, many non-Bayes procedures in common use are very remote from the Bayes boundary, and substantial improvement is possible.

The theorem also suggests a new way to resolve issues about subjectivity of Bayes procedures. One can simply take any standard non-Bayes procedure δ_0 and use linear programming to replace it by a uniformly better Bayes procedure. In this way, by restricting the choice to be amongst only those Bayes procedures that are uniformly better than the standard procedure, one can avoid the more subtle and controversial issues concerning comparisons within the full class of Bayes procedures. Of course, the fact that any non-Bayes procedure can be uniformly improved upon by a Bayes procedure makes it clear that one can restrict attention to the class of Bayes procedures, without any loss in so doing. When there is some compelling case for the standard procedure δ_0, then this provides a motivation for giving particular attention to those Bayes procedures that uniformly dominate it, and therefore greatly simplifies the decision problem. Typically there will be several pure Bayes procedures that are uniformly better than δ_0. To choose amongst them, one can either use subjective judgment to select an appropriate a priori distribution π; or alternately use some more objective method to choose amongst the Bayes procedures that are uniformly better than δ_0. For example, one could use a **minimax** procedure in the restricted problem.

It is important to note that many conventional statistical procedures map the data into estimates or tests or predictions, without interpreting such a mapping as being in any sense a conditional procedure, or even as being conditional upon the data. Examples include the product-limit estimator in **survival analysis** (*see* **Kaplan–Meier Estimator**), **proportional hazards** models, the various **bootstraps**, and many other such well-known statistical procedures that are routinely used in the analysis of data. Theorem 3 shows that all such mappings can be assessed with respect to their unconditional performance, and only those procedures which possess the internal consistency properties of procedures derivable from conditional probability distributions are not objectively defective. The theorem even applies to the so-called group decision problem, in which a group must arrive at some decision procedure. No matter how arrived at, if that procedure is not a Bayes procedure based upon conditional probability, then it can be uniformly improved upon by a Bayes decision rule.

Theorem 3 relies upon the existence of a probability model for the data, given the parameter, and also of a **loss function**. When there is no accepted such probability model, then of course everything becomes subjective, and there is essentially no serious role for theory at all. With regard to the loss function, there exist both statistical problems with a generally accepted loss function, and others in which the loss function does not exist or is unknown or is controversial. If there is no loss function at all, then there is really no problem, since anything can be done whatsoever without any punishment for even the most absurd procedures. When a loss function exists but is not entirely known, or alternately when losses or utilities are difficult to assess, it is possible to make use of **robustness** properties of decision procedures. For example, one can obtain an optimal procedure for several different loss functions under consideration, and if these are nearly in agreement then for practical purposes the decision problem is solved.

The Evaluation Game

While it is clear that a decision procedure that can always be uniformly improved upon in risk is not particularly desirable, it is important to point out precisely how the latter leads in practice to poor decisions.

Let θ be a conventional parameter that determines the distribution of a random variable, X, and

let $\delta_i(X)$, $i = 0, 1$, be two decision functions. Suppose that there is a referee who generates couples (θ_j, X_j), on a computer, for $j = 1, \ldots, M$, generating the θ_j in any way whatsoever (not necessarily probabilistically), and then using the specified conditional distribution to generate X_j, given θ_j, with the X_j conditionally independent. Let the referee generate M couples in this way. Assume that the conditional distribution for X_j, given θ_j, is known to all concerned. Consider a statistician or decision-maker who must choose between $\delta_1(X_j)$ and $\delta_0(X_j)$ to **estimate** θ_j using the same decision rule on each of the M occasions. Let $L(\theta_j, \delta_i(X_j))$ be the loss if θ_j is the true value of the parameter and $\delta_i(X_j)$ is used on the jth occasion, $1 \leq j \leq M$. We shall assume that all decision functions are to be mechanically implemented on a computer, without any data analysis or learning from one occasion to another.

Suppose, as in Theorem 3, that δ_1 is uniformly better than δ_0 with $R_{\delta_1}(\theta) \leq R_{\delta_0}(\theta) - \varepsilon$ for all θ with $\varepsilon > 0$. Summing over the M occasions, the actual increment in loss if δ_0 were used on each occasion instead of δ_1, would be $\sum_{j=1}^{M}[L(\theta_j, \delta_0(X_j)) - L(\theta_j, \delta_1(X_j))]$. The conditional expectation of this incremental loss due to use of δ_0, from the perspective of the referee who knows the θ_j, *is then, for any* θ_j *whatsoever*,

$$\begin{aligned} K(\theta_1, \ldots, \theta_M) &= \sum_{j=1}^{M} E_{X_j|\theta_j}\{L[\theta_j, \delta_0(X_j)] \\ &\quad - L[\theta_j, \delta_1(X_j)]\} \\ &= \sum_{j=1}^{M}[R_{\delta_0}(\theta_j) - R_{\delta_1}(\theta_j)] \geq M\varepsilon. \end{aligned}$$

This proves that in repeated usage of δ_0 instead of δ_1 with M large, one typically anticipates enormous extra loss. If, in addition, the referee uses some probability distribution π to generate the θ_j, then

$$\mathcal{E}\, K(\theta_1, \ldots, \theta_M) = M[B(\pi, \delta_0) - B(\pi, \delta_1)],$$

which is M times the difference in Bayes risks for the two procedures. Of course, the best possible decision procedure would be a procedure that is Bayes with respect to the π used by the referee, but even if this is unknown (or does not exist) it is still the case that one can enormously improve upon any specified non-Bayes procedure δ_0 by means of Theorem 3. This is the operational sense in which usage of non-Bayes procedures, that is, those not based upon conditional probability assessments, are objectively defective in performance. This argument also reveals the intimate connection between the fundamental frequentist argument that procedures are to be assessed in terms of long-run performance, and the Bayesian algorithm for obtaining optimal decisions by optimizing conditionally upon the data.

Next, suppose that instead of estimation of an unknown parameter θ, one is using the data X to **predict** another random variable Y. In other words, on the jth occasion one is given X_j and must now predict Y_j. Then the above argument goes through in exactly the same way, with now the usual estimative risk function replaced by the predictive risk function, which is a function of the unknown value of Y. The theory of predictive risk functions is presented in Hill [10]. If the decision function used to predict Y is $\delta(X)$ then the predictive risk function is

$$R_\delta(y) = \mathcal{E}\, L(y, \delta(X)).$$

In other words, the parameter θ is replaced by the true value of Y to be observed. It is assumed that a joint probability distribution has been specified for (X, Y), and $R_\delta(y)$ is the expectation of $L(Y, \delta(X))$, conditional upon $Y = y$.

A particularly important and challenging example, to which Theorem 3 applies, concerns data in a large sparse contingency table, such as is common in medical diagnosis. Suppose that one must put forth a conditional probability that a new patient with symptoms given by $X = x$ has disease y, based upon the data in the table. This is precisely the type of prediction problem just discussed. Many *ad hoc* methods exist for estimating such probabilities. Only those, however, for which there exists a joint distribution for (X, Y) such that the probabilities put forth are truly conditional probabilities based upon that joint distribution, cannot be improved upon uniformly by the algorithm of Theorem 3 (*see* **Decision Analysis in Diagnosis and Treatment Choice**).

The Infinite Case

Kolmogorov [13] put forth a theory of probability in which conditional probabilities were defined as Radon–Nikodym derivates of one bounded signed measure with respect to another. Briefly, let Y be a

random variable with respect to the probability space $(\Omega, \mathcal{A}, P)$ for which the expectation $\mathcal{E}(Y)$ exists, and let $\mathcal{B}$ be a σ-algebra of subsets of Ω such that $\mathcal{B} \subset \mathcal{A}$. Then the signed measure $\mu(B) = \int_B Y \, dP$ defined for $B \in \mathcal{B}$ is absolutely continuous with respect to P. According to the **Radon–Nikodým theorem**, there exists a $\mathcal{B}$-measurable function $f(\omega)$, often written as $f(\omega) = (d\mu/dP)(\omega)$, such that

$$\int_B Y(\omega) \, dP(\omega) = \int_B f(\omega) \, dP(\omega),$$

for all $B \in \mathcal{B}$. Any two such functions $f_i(\omega)$ can differ only on a set in $\mathcal{B}$ of P measure 0, and the conditional expectation of Y given $\mathcal{B}$ is defined to be $\mathcal{E}(Y|\mathcal{B})(\omega) = f(\omega)$ for any such function. Kolmogorov thus attempted to extend the classical concept of conditional probability to the nonfinite case, by requiring that the generalized law of total probability $\mathcal{E}Y = \mathcal{E}\mathcal{E}[Y|X]$ remain true. For Y the indicator of an event, this provides a definition of conditional probability given $\mathcal{B}$. An alternate method to obtain conditional probabilities in the sense of Kolmogorov, more directly related to standard mathematics and expectation, is to use the theory of projections in Hilbert space developed by von Neumann, as for example presented in Rényi [17, p. 262].

The theory of Kolmogorov has proved fruitful in allowing many elegant theorems concerning martingales (*see* **Counting Process Methods in Survival Analysis**) and **Markov processes** to be proved rigorously, in accord with the usual mathematical conventions. On the other hand, this theory is based on a number of idealizations, and in particular rests strongly upon the Axiom of Continuity (or Countable Additivity). Kolmogorov [13, p. 15] states:

> For infinite fields, on the other hand, the Axiom of Continuity, VI, proved to be independent of Axioms I–V. Since the new axiom is essential for infinite fields of probability only, it is almost impossible to elucidate its empirical meaning, as has been done, for example, in the case of Axioms I–V in 2 of the first chapter. For, in describing any observable random process, we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes. *We limit ourselves, arbitrarily, to only those models which satisfy Axiom VI.* [Author's italics.] This limitation has been found expedient in researches of the most diverse sort.

Conditional probability in the sense of Kolmogorov is an extension of the classical concept of Bayes in the sense that the two are in agreement whenever $\mathcal{B}$ is purely atomic, as for countable spaces of outcomes. For it is a standard result that whenever $\mathcal{B}$ is purely atomic with atoms B_i having positive probability, then $(d\mu/dP)(\omega) = \mathcal{E}(Y|B_i)$ for $\omega \in B_i$. See Rényi [17, p. 261]. However, the theory of Kolmogorov also applies to cases in which the underlying space Ω is a finite-dimensional Euclidean space, a Hilbert space, a pseudo-metric space, and even to appropriately defined Borel sets in an abstract space of points. Alternately, this theory can be based upon the Daniell integral, as in Riesz–Sz.-Nagy [18, p. 132], and is then closely related to the theory of nonnegative linear functionals. However, it is not necessarily harmless to generalize the concrete and clear concept of conditional probability in finite spaces to such idealized spaces. Kolmogorov [13, p. 17] puts it well in discussing the Borel field $\mathcal{BF}$:

> Even if the sets (events) A of $\mathcal{F}$ can be interpreted as actual and (perhaps only approximately) observable events, it does not, of course, follow from this that the sets of the extended field $\mathcal{BF}$ reasonably admit of such an interpretation.
>
> Thus there is the possibility that while a field of probability $(\mathcal{F}, P)$ may be regarded as the image (idealized, however) of actual random events, the extended field of probability $(\mathcal{B}, P)$ will still remain merely a mathematical structure.

If one allows the possibility of the realization of an irrational number as the actual outcome of an experiment, with this number obtained by direct measurement, then conditional probability in the sense of Kolmogorov can disagree with the classical concept of both Bayes and the frequentist theory. Of course, no such outcome has ever been observed, or could ever be observed, in finite time, and even such irrational numbers as π and e are at any given time known only up to a finite number of decimal points. Furthermore, the use of transformations such as $\sqrt{x}$ that can lead to irrational numbers does not alter things, since when used operationally by a computer these must be replaced by some finite approximation. If the data space $\mathcal{X}$ consisted of all points in even one of the most simple idealized spaces, the real line, then no real-world observation could ever consist of the exact value of the observation, since it would require infinite time to determine all the decimal points for even a single such measurement. Hence the actual observations upon which one conditions, as in Bayes' theorem, are necessarily very special subsets of $\mathcal{X}$,

such as for example that the observation lies between two rational numbers. Borel, who initiated modern measure theory, was particularly concerned about the misuse of mathematics in connection with real-world data, as for example in Borel [2, Chapters 5–8], and emphasized the importance of approximations in the *evaluation* of real-world probabilities. The theory of Kolmogorov, in the case of even such simple sample spaces as the real line, has no direct relevance for the question of updating of opinions, as in the Bayesian theory, or for decision theory, since the datum x upon which the decision is to be based will always be finitized. Rather, it includes an assumption (countable additivity) which, although useful in proving limit theorems, according to Kolmogorov cannot be justified other than by pragmatic reasons even for this purpose, much less for real-world decision problems.

De Finetti also attempted to extend the classical concept of conditional probability, and proposed a third axiom in de Finetti [6, p. 338] to allow for conditional probability given an event of probability 0. His third axiom states that probability evaluations are to be in accord with the axioms of finitely additive probability theory, even conditional upon an event of probability 0. To obtain conditional probabilities in the general finitely additive setting, Dubins & Savage [8] developed the concept of a finitely additive strategy, under which probability distributions are attached to each history of a process. These specify the probability for the future, given the past of the process, and allow arbitrary observational data, such as irrational numbers or a point in Hilbert space. At this level of generality, it is not necessarily possible to reverse the order of integration, so that a strategy may presume a definite ordering of the observations.

For infinite spaces, the finitely additive theory contains paradoxes of nonconglomerability, a concept due to de Finetti. Nonconglomerability means that for some event A and partition B_t it is the case that $P(A) > \sup_t P(A|B_t)$. In denumerable spaces it is known that countable additivity is equivalent to conglomerability. See Hill & Lane [11] for an elementary proof. Thus the countably additive theory builds in assumptions regarding conditional probability. Kolmogorov, in assuming countable additivity, was implicitly ruling out nonconglomerability, at least in the discrete case. At the present time there is no theory, free of paradoxes, that can seriously deal with the nondenumerable case, as when irrational numbers are taken literally. Borel [3, p. 60, p. 175] gives illuminating discussions of ways in which some mathematicians, unaware of the questions already raised by himself and Poincaré, and later by Kolmogorov; and with limited knowledge of science, have often confused the basic issues when dealing with the nonfinite case. Borel [7, Chapters 2, 3, & 5] discusses the subtle issues that arise in attempting to apply the theory of probability to real-world problems, such as arise in the analysis of mortality tables.

In serious mathematics, an irrational number is viewed as the idealized limit obtained by means of a certain procedure, such as for example the limit of a sequence of partial sums. This point of view can also be taken regarding procedures involving randomness, such as draws from an urn. Prior to the work of Cantor, the realized infinite was regarded as nonsense by most major mathematicians; for example, Gauss and Kronecker. Related viewpoints continued into this century, as represented by Borel. Poincaré, Brouwer, Weyl, and others. Kolmogorov, also a major mathematician who made serious contributions to logic as well as to probability, was concerned with such issues, and his opinions evolved over time. For example, in his book with Fomin, a measure is so defined as not necessarily to be countable additive, and some standard finitely additive measures are studied. The theory of Kolmogorov was elegantly extended by Rényi [17, p. 38] to conditional probability spaces. This extension allows one to deal rigorously with σ-finite measures such as counting measure on a denumerable space, and is a major step toward the finitely additive theory, although Rényi did not choose to make the final extension. There is, however, a clear recognition both by Kolmogorov & Fomin [15, p. 206] and by Rényi [17, p. 60] that generalized functions such as the Dirac delta function are important and legitimate objects for mathematics (and probability) to study. The finitely additive theory can be regarded as the extension of that of Kolmogorov to include such objects.

Kolmogorov [14, p. 1] asserted his continuing belief that

> The frequency concept based on the notion of *limiting frequency* [author's italics] as the number of trials increases to infinity, does not contribute anything to substantiate the applicability of the results of probability theory to real practical problems where we have always to deal with a finite number of trials.

He then proposed a theory of complexity and information based upon admissible algorithms for selecting a subset of a random table, as a possible justification. With a return to the finite case, however, as the critical case for real-world use of probability, the axioms of finitely additive probability (Kolmorogov's axioms I–V) can be strongly motivated, as by himself, or by the coherency theory of de Finetti, or by the BNW theory of statistical decision functions. Consequently, one is led back to the use of conditional probability and Bayes procedures for real-world decision-making. To the extent that nonfinite spaces arise at all in real-world problems, as suggested by Kolmogorov [13, p. 18; 14], it is in giving insight as to approximations that arise when the data space is large but finite, and in providing answers to finite problems by means of methods of analysis available in the infinite case.

References

[1] Borel, E. (1962). *Probabilities and Life*. Dover, New York.

[2] Borel, E. (1963). *Probability and Certainty*. Walker, New York.

[3] Borel, E. (1965). *Elements of the Theory of Probability*. Prentice-Hall, Englewood Cliffs, New Jersey.

[4] De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives, *Annales de l'Institut Henri Poincaré* **7**, 1–68.

[5] De Finetti, B. (1974). *Theory of Probability*, Vol. I. Wiley, London.

[6] De Finetti, B. (1975). *Theory of Probability*, Vol. II. Wiley, London.

[7] DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

[8] Dubins, L.E. & Savage, L.J. (1976). *Inequalities for Stochastic Processes*. Dover, New York.

[9] Hill, B.M. (1994). On Steinian shrinkage estimators: the finite/infinite problem and formalism in probability and statistics, in *Aspects of Uncertainty*, P. Freeman & A.F.M. Smith, eds. Wiley, Chichester, pp. 223–260.

[10] Hill, B.M. (1994). Bayesian forecasting of economic time series, *Econometric Theory* **10**, 483–513.

[11] Hill, B.M. & Lane, D. (1985). Conglomerability and countable additivity, *Sankhyā, Series A* **47**, 366–379.

[12] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Oxford University Press, London.

[13] Kolmogorov, A. (1950). *Foundations of the Theory of Probability*. Chelsea, New York.

[14] Kolmogorov, A.N. (1963). On tables of random numbers, *Sankhyā, Series A* **25**, 369–376.

[15] Kolmogorov, A.N. & Fomin, S.V. (1970). *Introductory Real Analysis*. Dover, New York.

[16] Ramsey, F.P. (1926). Truth and probability, reprinted in *The Foundations of Mathematics and Other Logical Essays*, R.B. Braithwaite, ed. Humanities Press, New York, 1950.

[17] Rényi, A. (1970). *Probability Theory*. American Elsevier, New York.

[18] Riesz, F. & Sz.-Nagy, B. (1955). *Functional Analysis*. Frederick Ungar, New York.

[19] Savage, L.J. (1972). *The Foundations of Statistics*, 2nd Rev. Ed. Dover, New York.

(*See also* **Axioms of Probability**; **Foundations of Probability**)

B. HILL

Conditionality Principle

The conditionality principle of statistical inference is usually interpreted to mean that inference about $\boldsymbol{\theta}$ in the model $f(\boldsymbol{y};\boldsymbol{\theta})$ should be conditional on any **ancillary statistic** for $\boldsymbol{\theta}$. This principle has caused a great deal of discussion in the literature on the foundations of statistics: for an introduction see [1]. As part of this discussion, examples have been constructed for which there are nonunique ancillary statistics, for which no ancillary statistics exist, and for which there exist ancillary statistics but no maximal ancillary statistic.

From a foundational point of view the conditionality principle entails quite a few difficulties. One of these is Birnbaum's theorem [3], which shows that **sufficiency** and the conditionality principle imply the so-called likelihood principle, which states that inference should be based only on the likelihood function, and not, for example, on the sampling properties of the likelihood function under the model, which would be the usual frequentist approach to likelihood-based inference (*see* **Foundations of Probability**). In fact, the conditionality principle alone entails the likelihood principle, as shown in [5].

Berger & Wolpert [2] is the most comprehensive book treatment of the conditionality principle. It is argued there that the most satisfactory implementation of the likelihood principle is a Bayesian approach to inference. Discussion of other foundational issues arising in conditioning appears in [4]. Recently, McCullagh [6] has considered in detail an unusual (although possibly artificial) class of examples which has more than one ancillary statistic, and in which the conditional inference is quite dependent on the choice of ancillary on which to condition.

References

[1] Berger, R.L. & Casella, G. (1990). *Statistical Inference*. Wadsworth & Brooks/Cole, Belmont.

[2] Berger, J.O. & Wolpert, R.L. (1984). *The Likelihood Principle*. IMS Lecture Notes - Monograph Series, Vol. 6 Institute of Mathematical Statistics, Hayward.

[3] Birnbaum, A. (1962). On the foundations of statistical inference (with discussion), *Journal of the American Statistical Association* **57**, 269–306.

[4] Brown, L.D. (1990). An ancillarity paradox in multiple regression (with discussion), *Annals of Statistics* **18**, 471–533.

[5] Evans, M.J., Fraser, D.A.S. & Monette, G. (1986). On principles and arguments to likelihood (with discussion), *Canadian Journal of Statistics* **14**, 181–200.

[6] McCullagh, P. (1992). Conditional inference and Cauchy models, *Biometrika* **79**, 247–259.

(*See also* **Inference**; **Likelihood**)

N. REID

Confidence Intervals, Binomial, when no events are observed

The appeal of the *rule of three* to clinicians is in its simplicity and usefulness in safety evaluation of adverse events. Specifically, consider a scenario in which the **probability**, p, of an event (generally an adverse reaction to a drug or clinical procedure) is known, a priori, to be small, and in a study conducted on n patients, no events have occurred. The problem is to find an upper bound for the unknown probability, p. The *rule of three* states that the 95% upper confidence bound for p is approximately $3/n$ (*see* **Confidence Intervals and Sets**).

In the context of safety evaluation in clinical research, in which a clinician has to demonstrate the safety of a new procedure (i.e. show that the probability of an adverse event is lower than some small acceptable probability), the *lower* bound for p is not of practical interest; one is mainly concerned with the *upper* bound on p, since it represents the "worst case scenario", or the largest probability for placing a patient at risk. Thus, the lower bound for p may a priori be taken to equal zero.

Derivation of the Rule of Three

Let the **random variable** X have a **binomial distribution** with parameters n and p. If $X = x$ is the observed number of events in n trials, then the Clopper–Pearson (max-P) upper $100(1-\alpha)\%$ bound for p may be obtained as a solution to

$$\sum_{t=0}^{x} \binom{n}{t} p^t (1-p)^{n-t} = \alpha.$$

When $x = 0$, the expression reduces to $(1-p)^n = \alpha$. Then $P(X = 0|n, p) = (1-p)^n$, and one can obtain the $100(1-\alpha)\%$ upper bound for p by solving $(1-p)^n \leq \alpha$ for p. This yields $p \geq 1 - \alpha^{1/n}$, and by taking $p_u = 1 - \alpha^{1/n}$ for the least upper bound for p, the interval $(0, p_u)$ provides $100(1-\alpha)\%$ coverage for p. Now, $3/n$ appears for the following reason. From the Taylor expansion $\alpha^{1/n} = 1 + \ln(\alpha)/n + [\ln(\alpha)]^2/2n^2 + \cdots$, one obtains, by retaining only the linear portion,

$$1 - \alpha^{1/n} = \frac{-\ln(\alpha)}{n}.$$

For $\alpha = 0.05$, $-\ln(\alpha) = 2.996$, and thus p_u is numerically close to $3/n$.

A similar argument using a **Poisson** random variable, with parameter $\lambda = np$, yields $P(X = 0|np) = \exp(-np) \leq \alpha$, which, after taking the natural log of both sides, produces the least upper bound $p_u = -\ln(\alpha)/n$ which yields $3/n$ as in the binomial case.

The *rule of three* may be derived using the **Bayesian** approach as well. Assume a **Beta**$(1, b)$ **prior** on p, i.e. $\pi(p) = (1-p)^{b-1}/B(1, b)$. Figure 1 presents a sequence of kernels (*see* **Density Estimation**) of Beta$(1, b)$ priors for various values of b. With a Beta$(1, b)$ prior, simple integration yields a posterior credibility interval for p:

$$P(0 < p < p_u | X = 0, b, n) = 1 - (1 - p_u)^{(n+b)} \geq (1-\alpha),$$

which simplifies to

$$p_u \geq 1 - \alpha^{1/(n+b)},$$

and the right-hand side can be approximated via Taylor expansion by

$$\frac{-\ln(\alpha)}{(n+b)}.$$

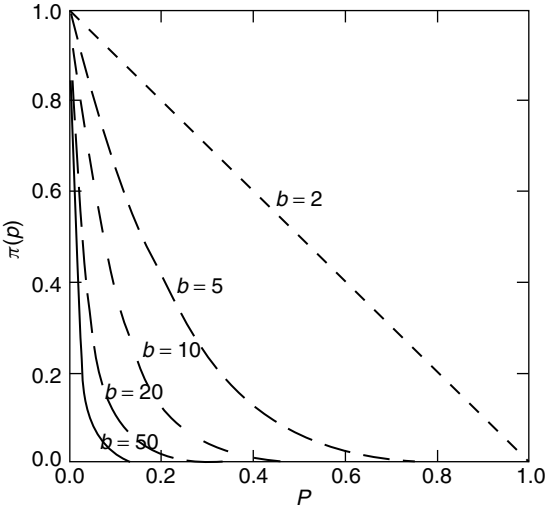

Figure 1 Beta$(1, b)$ prior kernels for various values of b

Table 1 Upper bounds on p when $x = 0$: Poisson (2); exact binomial (3); Rule of Three (4); Bayesian upper bound for uniform (Beta(1, b), $b = 1$); prior (5); improved Rule of Three (6); and Bayesian Rule of Three for $b = 20$ (7)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
n	$-\ln(\alpha)/n$	$1-\alpha^{1/n}$	$3/n$	$1-\alpha^{1/(n-1)}$	$3/(n+1)$	$3/(n+20)$
3	0.99858	0.63160	1.00000	0.52713	0.75000	0.13043
4	0.74893	0.52713	0.75000	0.45072	0.60000	0.12500
5	0.59915	0.45072	0.60000	0.39304	0.50000	0.12000
6	0.49929	0.39304	0.50000	0.34816	0.42857	0.11538
7	0.42796	0.34816	0.42857	0.31234	0.37500	0.11111
8	0.37477	0.31234	0.37500	0.28313	0.33333	0.10714
9	0.33286	0.28313	0.33333	0.25877	0.30000	0.10345
10	0.29957	0.25877	0.30000	0.23840	0.27273	0.10000
20	0.14979	0.13911	0.15000	0.13295	0.14286	0.07500
50	0.05991	0.05816	0.06000	0.05705	0.05882	0.04285
100	0.02996	0.02951	0.03000	0.02923	0.02970	0.02500

This gives us a Bayesian Rule of Three as $3/(n + b)$, with $b \geq 1$. Obviously, for $b = 1$, $3/(n + 1)$ is the largest such upper bound, corresponding to the uniform prior (see Table 1).

If no events occur in k studies of sizes n_1, $n_2, \ldots, n_k$, then one has a more general *rule of three* as $3/(n_1 + n_2 + \cdots + n_k + 1)$, which may be derived using a Bayesian approach.

Bibliography

Hanley, J.A. & Lippman-Hand, A. (1983). If nothing goes wrong, is everything allright? Interpreting zero numerators, *Journal of the American Medical Association* **249**, 1743–1745.

Jovanovic, B. & Levy, P.S. (1997). A look at the rule of three, *American Statistician* **51**, 137–139.

Jovanovic, B. & Viana, M.A.G. (1997). Upper confidence bounds for binomial probability in safety evaluation. *American Statistical Association 1996 Proceedings of the Section on Biopharmaceuticals*. American Statistical Association, Alexandria.

Jovanovic, B. & Zalensky, R. (1997). Upper bound on binomial probability when the number of observed events is small or zero, *Annals of Emergency Medicine* **30**, 301–306.

Kerns, J.R., Shaub, T.B. & Fontanarosa, P.B. (1993). Emergency cardiac testing in the evaluation of emergency department patients with atypical chest pain, *Annals of Emergency Medicine* **22**, 794–798.

Louis, T.A. (1981). Confidence intervals for a binomial parameter after observing no successes, *American Statistician* **35**, 154.

Press, S.J. (1989). *Bayesian Statistics: Principles, Models and Applications*. Wiley, New York.

Vollset, S.E. (1993). Confidence intervals for a binomial proportion, *Statistics in Medicine* **12**, 809–824.

BORKO D. JOVANOVIC

Confidence Intervals and Sets

A *confidence interval* for a fixed parameter θ, or a *confidence set* for a multidimensional parameter $\boldsymbol{\theta}$, represents a plausible range of values for the parameter(s) that is consistent with the observed data. Specifically, for a single parameter θ, the interval (L, U) is a $100(1-\alpha)\%$ confidence interval for θ if $\Pr(L \leq \theta \leq U) = 1-\alpha$. The quantity $1-\alpha$ is called the *confidence coefficient* or *confidence level*, and is equal to the probability that the random interval (L, U), contains the fixed parameter θ. The confidence limits L and U are constructed from the observed data in such a way that in infinite replications of the study, the proportion of such intervals that contain the parameter θ, or the *coverage probability*, is $1-\alpha$. For an r-dimensional parameter vector $\boldsymbol{\theta}$, the confidence set $\mathbf{I}$ is defined as the r-dimensional space $\mathbf{I} = [\boldsymbol{\theta} : L_j \leq \theta_j \leq U_j,\ j = 1, \ldots, r]$ such that $\Pr(\boldsymbol{\theta} \in \mathbf{I}) = 1-\alpha$.

Confidence intervals need not be symmetric, but could reflect upper or lower bounds for a parameter in a *one-sided confidence interval*, as opposed to the two-sided interval described above. A $100(1-\alpha)\%$ upper one-sided confidence limit or bound U' for θ has the property that $\Pr(U' \geq \theta) = 1-\alpha$. The interval $(-\infty, U')$ is sometimes called a lower one-sided confidence interval. The corresponding lower one-sided bound L' is such that $\Pr(L' \leq \theta) = 1-\alpha$. The definitions of coverage probabilities and confidence coefficients are equally applicable to one-sided limits.

Confidence intervals are often centered around an estimate of the parameter of interest θ, and give an indication of the precision of the estimate. They incorporate the random variation inherent in the data into the estimation procedure. Various methods of constructing confidence intervals are used, depending on the distribution of the data, and the particular parameter of interest (*see* **Estimation, Interval**). The length of the confidence interval, a reflection of the precision of the estimate, is influenced by the sample size, the variability in the data, and the confidence coefficient. The higher the level of confidence, the greater the variability in the data, or the smaller the sample size, the wider is the interval. The tighter the interval, the more certain is the parameter estimation.

The confidence interval is based on frequentist statistical theory in which the parameter θ is considered fixed but unknown, and was developed by Neyman [2, 3]. The alternative **Bayesian** theory considers the parameter θ to be a realization of a random variable Θ, and instead defines a probability distribution for Θ [1]. Probability statements about Θ, conditional on the observed data, are based on the posterior distribution. Upper and lower percentage points of the posterior distribution would correspond to the frequentist confidence interval. For multidimensional parameters, a region of high posterior density would correspond to the confidence set.

References

[1] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[2] Neyman, J. (1935). On the problem of confidence intervals, *Annals of Mathematical Statistics* **6**, 111–116.

[3] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transaction of Royal Society of London, Series A* **236**, 333–380.

(*See also* **Estimation**; **Inference**)

NANCY R. COOK

Confidentiality and Computers

Introduction

This updates an article with the same title that appeared in the first edition of EOB.

Confidentiality is defined as "the characteristic of data and information being disclosed only to authorized persons, entities and processes at authorized times and in the authorized manner" [24]. Essentially, confidentiality relates to control over information. Who should have access to information, and under what circumstances?

While confidentiality concerns apply also to paper-based or manual records, the widespread utilization of information technology, and especially the networking of computers (*see* **Computer Architecture and Organization**), has led to new and difficult challenges for maintaining confidentiality and privacy. Confidentiality issues arise both for personal information and for information that is important for the functioning of business and other organizations. Personal information may include facts and figures about ourselves, our lives, our personal work and financial situations, and our medical history. Most people do not however regard information on telephone number or street address as confidential, and are happy for it to be printed in a phone directory, that is, it is placed "in the public domain".

When information is provided to people in trusted positions – doctors, lawyers, researchers, and government officials – there is an expectation that it will be kept secure, used only for the purpose for which it has been collected. Any breach of this trust may compromise future data collection [4].

Social and cultural value systems strongly influence attitudes to privacy and confidentiality. Different societies have different views on where the balance should lie between individual rights and public good.

Privacy and Confidentiality

The confidentiality of personal information is closely aligned with the broader issue of privacy. Gostin et al. [12] suggest that privacy is "the right of an individual to limit access by others to some aspect of the person", while confidentiality is "a form of informational privacy characterized by a special relationship, such as the physician-patient relationship". Privacy protection is about individuals being informed why their information is being collected, having access to their information and having as much say as possible about how their information is used and to whom it may be disclosed [22].

Laws, Standards, Principles, and Practicalities

Rights to privacy and confidentiality, and principles that guide the handling of private information, may be enshrined in law. Public and other bodies who hold data may have their own standards, which interpret and/or supplement legal requirements. Such laws and standards provide a framework for deciding when and under what conditions data can be made available to third parties.

A major impetus to the passing of laws on the privacy of data, in a number of countries, was the *1995 Privacy Directive of the European Union* [11], which applies both to private and to public data. This obliges the EU's member states to ensure that their national legislation is in accordance with the directive, and prevents the exchange of data with nations that do not have "adequate" privacy protections.

A basic principle of most privacy legislation is the protection of the "need to know", or the limiting of access to the minimum required to perform a task. Implicit in this is the need to exclude access for those who have no genuine requirement to access the data. Thus for many (but not all) research uses of medical data, it will be enough to make data available without identifying information, usually with some form of coding or **record linkage** that is unique to the individual. This is not as straightforward as might appear at first glance; thus mechanisms will be required that allow a check on apparent anomalies or suspected errors.

Confidentiality and Computer Systems

Privacy and confidentially measures can work only when the data holder has the will, technical capacity, and moral or legal authority to keep such information secure [8]. To be effective, such measures require the informed cooperation of users. Education thus has an important role.

Stand-alone systems are in principle relatively easy to secure. Broadly, access can be limited to trusted individuals for authorized purposes, and all use can be monitored and logged. It will be possible to attribute any security lapse to one of a small number of known individuals. With smaller systems the same person may have more than one role, that is, general user and system manager. Such a user may be given separate accounts for the separate roles, and be required to access the system with the account that is relevant for the role or task that they are performing at the time.

Networked systems, systems that use wireless connections and (even more) systems that are connected to the **internet**, pose new and difficult security problems. Data collected by the CERTC Coordination Center of Internet security expertise, shows a sharp increase in intrusion incidents between 1999 and 2001 [15]. There may be denials of service, files may be destroyed, damaged, or altered, and the security of data may be compromised.

Current systems may be testing to breaking point current design approaches for networked systems, perhaps inevitable in the rush to create a networked world. The issues are so important that a recently established journal (January 2003), IEEE Security and Privacy, is devoted to them. System design is important, but gives only a first line of defence. System design, however careful and informed, is likely to lag behind technical innovation, for which the crucial test is day to day use.

Planning should therefore include measures that may detect unauthorized intrusion or use of data, should preventative measures fail, and have in place strategies that will respond rapidly and effectively [6, 16, 19, 25]. It should aim to mitigate the potential damage from unauthorized access. An "outsider" who gains access to data, whether by intrusion, from a mix-up, or from an incompletely erased hard drive that has been sent for disposal, will, in general, be unable to do much damage with data that lacks identification information. This emphasizes the importance of the trust that is placed in "insiders", who may be well placed to reconstruct the missing connections.

Security systems that place undue obstacles in the way of legitimate users can be self-defeating. They place obstacles in the way of legitimate use of the data. For medical data, they may compromise the obtaining of information that has strong public health implications. They may make it difficult or impossible to check on apparently anomalous data (*see* **Outliers**). At the same time, unduly onerous systems invite practices that render them partially ineffective. For example, access codes for a supposedly secure system may be made widely available, avoiding the time and complication associated with creating any new access codes.

Preventing and Responding to Incursion

Steps that may reduce the risk of unauthorized access, or to reduce its impact when it does occur, are:

- The use of "secure" software systems, that is, systems that are relatively invulnerable to unauthorized intrusion or to misuse.
- The use of "secure" forms of user identification, and of data transfer (*see* **Sample Surveys in the Health Sciences**).
- The use of a system of permissions that operates on a per file or per directory basis, with access restricted on a "need to access" basis.
- Avoiding or preventing storage of sensitive data on inherently insecure hardware, such as removable devices and laptops.
- Monitoring and logging of access to sensitive data.
- Education of managers and users.
- Regular auditing, checking that physical and system security measures are effective, that users have the correct level of access for their roles, that the rules defined by data providers are being met, and that managers and users understand their roles and responsibilities.
- The use, for sensitive data, of suitable forms of encryption, for data storage as well as for data transmission.
- Use of a system of record linkage that limits the need for access to identification information.
- Storage of identification information separately from the data to which it relates.

It is essential to get expert advice for the setting up of highly secure systems. A threat model, which balances security against the costs and user inconvenience of harsher security implementations, can be helpful [3, 19]. Security breaches such as are documented in Neumann [18, 20] provide data against which any such model can be validated.

Confidentiality, Computers, and Statistical Analysis

In most official statistical agencies, the protection from disclosure of individually identifiable records is guaranteed by legislation. The situation in Australia provides a good example of the issues that this raises. The Census and Statistics Act requires that the Australian Bureau of Statistics does not release statistics "in a manner that is likely to enable the identification of a particular person or organization". This has particular relevance to the release of computer files containing unidentifiable unit records, or microdata. The Australian Statistician [17] has discretionary powers to release unidentifiable individual statistical records, provided a recipient gives a legally binding undertaking that no attempt will be made to identify particular persons or organizations; that the information will be used for statistical purposes only; and that the information will not be disclosed to any other person or organizations. Other conditions may also be imposed.

The Australian Statistician is advised by a panel that assesses all proposals, to ensure that the data are unlikely to enable unit identification. The panel takes into account the level of detail in each record and the extent of disclosure avoidance techniques (e.g. releasing values not as collected but as classes, and randomly perturbing values by some small number). In this way, confidentiality provisions are satisfied while permitting legitimate demands for secondary data analysis, although users sometimes express concern that such techniques significantly reduce the value of the data that are released.

At times, statistical tables can also threaten confidentiality. For example, when a large enterprise dominates an industry, publishing information about that industry could be commercially sensitive. Similarly, detailed or multidimensional tables could contain information about individuals living in small communities. Possible ways to modify tables so that confidentiality is protected include:

a. Perturbing the table, to change the values of the data, for example, randomly rounding the values.
b. Grouping categories together (aggregating). Adding columns or rows together where there are small or confidential numbers, reduces the risk that it will be possible to infer confidential information.
c. Deleting values, or cell suppression. The value in a cell from the table may be suppressed. In order to avoid the value being calculated from subtotals/marginal totals, it is necessary to take other steps as well, usually deleting another set of cell values. Deleting these other cells is known as secondary or complementary cell suppression.

There is an extensive literature that discusses these issues [9, 26, 27, 28].

Confidentiality, Computers, and Epidemiological Research

The quantity and type of health information collected, transmitted, and stored electronically has increased dramatically in recent years. This reflects the widespread use of computers to store health related information systems (*see* **Health Care Utilization Data**), to satisfy accountability requirements and assist quality and continuity of care. Also, the number of procedures and treatments performed has increased and information on lifestyle, **risk factors**, family medical history (*see* **Family History Validation**), health and functional status (*see* **Health Status Instruments, Measurement Properties of**), and genetic data (*see* **Genetic Epidemiology**) are increasingly likely to be recorded. In addition, most countries now maintain registries of **vital statistics**, and registries for particular diseases (*see* **Disease Registers**), for example, **cancer registries**.

In some types of epidemiological research, identification of individuals to the researcher may be unavoidable (*see* **Confidentiality in Epidemiology**). In such cases, the benefits to society must be carefully weighed and justified against privacy principles and against guidelines on the use of identifying information for research purposes. Such guidelines may address the issues of informed consent of subjects (*see* **Medical Ethics and Statistics**), steps to preserve confidentiality, the use of the information obtained only for the purpose for which it was collected, and approval from an institutional ethics committee. Past Australian examples where the public good has outweighed personal privacy principles include epidemiological research into the effects of cigarette **smoking**, use of medical record linkage in research into the side-effects of the oral contraceptive pill, and research into the long-term consequences of chemical

exposures (*see* **Risk Assessment for Environmental Chemicals**) and of war service.

Medical research with human subjects relies heavily both on the trust of patients and on public support for research funding. That trust can be helped by making those whose data are collected aware of their rights, of the intended use of the data, and of potential benefits for medical research. Their attention may be drawn to web sites where they can follow the progress of research. Measures that keep subjects informed may be more important, for the maintenance of public trust, than unduly stringent and onerous privacy guarantees that may impede the use of data for research purposes.

Links on the web site [2] give access to a wide range of information on privacy, confidentiality, and data security, in the United States and internationally. For summary information on selected legal norms that relate to the protection of personal information in health research, with extensive references, see [5]. As examples of requirements, see [5, 7, 13, 14, 21, 22, 23]. See [1] for a set of standards that are intended, in the first place, for medical clinicians. The article [10] has practical advice that is relevant to anyone whose work involves the processing and analysis of data.

Conclusion

Confidentiality has social, cultural, ethical, and legal dimensions. While it is not specific to the use of information on computers, the ability of information systems to store, process, and transmit large amounts of data makes attention to confidentiality in computer systems an issue of major importance. The linking of computer systems via the Internet raises has raised new technical difficulties for the maintaining of security and confidentiality.

References

[1] American Medical Association. (2003). E-5.07 Confidentiality: Computers. `http://www.ama-assn.org/ama/pub/category/8360.html`,.

[2] American Statistical Association. (2003). Privacy, Confidentiality, and Data Security Website. `http://www.amstat.org/comm/cmtepc`.

[3] Barrows, R.C. & Clayton, P.D. (1996). Privacy, confidentiality, and electronic medical records, *Journal of the American Medical Informatics Association* **3**, 139–148.

[4] Butz, W.P. (1985). Data confidentiality and public perceptions: the case of European censuses, *Proceedings of the Section on Survey Research Methods*. American Statistical Association, Washington, DC, pp. 90–97.

[5] Canadian Institutes of Health Research. (2001). Selected International Legal Norms on the Protection of Personal Information in Health Research. `http://www.cihr.ca/`.

[6] Cybenko, G. (2002). Editor's message: the long march, *IEEE Computer* **35**(Suppl. 1), `http://computer.org/security/supp_toc.htm`.

[7] Department of Health (UK). (2003). Patient Confidentiality and Caldicott Guardians. `http://www.dog.gov.uk/ipu/confiden/`.

[8] Donaldson, M.S. & Lohr, K.N., eds. (1994). *Health Data in the Information Age. Use, Disclosure and Privacy*. National Academy Press, Washington, pp. 152–153.

[9] Doyle, P., Lane, J.I., Theeuwes, J.M. & Zayatz, L.V., eds. (2001). *Confidentiality, Disclosure and Data Access – Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam.

[10] Earnhart, B. (2003). Respect your data, *Amstat News* (309), 36–38. `http://www.uiowa.edu/~soc/datarespect/data_training_frm.html`.

[11] European Parliament and Council of Europe. (1995). Directive 95/46/EC of the European and of the Council 24 October 1995 on the protection of individuals with regard to the processing of personal data and the free movement of such data, *Official Journal L* **281**, 0031–0050. `http://europa.eu.int/eur-lex/en/lif/dat/1995/en_395L0045.html`.

[12] Gostin, L.O., Turek-Brezina, J., Powers, M., Kozloff, R., Faden, R. & Steinauer, D.D. (1993). Privacy and security of personal information in a new health care system, *Journal of the American Medical Association* **270**, 2487–2493.

[13] Guidelines Under Section 95 of the Privacy Act 1998. (2000). Canberra. `http://www.health.gov.au/nhmrc/issues/researchethics.htm`.

[14] Guidelines Under Section 95A of the Privacy Act 1998. (2001). Canberra. `http://www.health.gov.au/nhmrc/issues/researchethics.htm`.

[15] Householder, A., Houle, K. & Dougherty, C. (2002). Computer attack trends challenge internet security *IEEE Computer* **35**(Suppl. 1), `http://computer.org/security/supp_toc.htm`.

[16] McConnell, M. (2002). Information assurance in the twenty-first century, *IEEE Computer* **35**(Suppl. 1) `http://computer.org/security/supp_toc.htm`.

[17] McLennan, W. (1996). The product of the Australian Bureau of Statistics, *Australian Journal of Statistics* **38**, 1–14.

[18] Neumann, P. (1995). *Computer-Related Risks*. Addison-Wesley, Reading, MA.

[19] Neumann, P. (1999). The Challenges of Insider Misuse. `http://www.csl.sri.com/users/neumann/pgn-misuse.html`.

[20] Neumann, P. (2003). Illustrative Risks to the Public in the Use of Computer Systems and Related Technology.

http://www.csl.sri.com/users/neumann/illustrative.html.

[21] Thompson, C. (2001). *NHMRC Human Research Ethics Handbook*, Section 18, Privacy of Information. http://www.health.gov.au/nhmrc/hrecbook/01_commentary/18.htm.

[22] O'Connor, K. (1996). Privacy Issues Facing a Networked Health Environment, from the text of a speech to the Health Issues Centre Discussion Forum, Melbourne, March 18. Privacy Commissioner, Human Rights Australia.

[23] Office for Civil Rights – HIPAA. (2003). Medical Privacy–National Standards to Protect the Privacy of Personal Health Information. http://www.hhs.gov/ocr/hipaa/bkgrnd.html.

[24] Organisation for Economic Co-operation and Development. (1992). *Guidelines for Security of Information Systems*. OECD, Paris.

[25] Stajano, F. & Anderson, R. (2002). The resurrecting duckling: security issues for ubiquitous computing, *IEEE Computer* **35**(Suppl. 1), http://computer.org/security/supp_toc.htm.

[26] U.S. Office of Federal Statistical Policy and Standards. (1980). Report on Statistical Disclosure and Disclosure-Avoidance Techniques, Statistical Working Paper 2, U.S. Department of Commerce.

[27] U.S. Office of Federal Statistical Policy and Standards. (1994). Report on Statistical Disclosure and Limitation Methodology, Statistical Working Paper 22, U.S. Department of Commerce.

[28] Willenborg, L.C.R.J. & de Waal, A.G. (1996). *Statistical Disclosure in Practice*, Springer Lecture Notes in Statistics 111, Springer-Verlag, New York.

JOHN H. MAINDONALD & HELEN P. STOTT

Confidentiality in Epidemiology

Respecting **confidentiality** is an important prerequisite in clinical and epidemiologic research. The term *confidentiality* is closely related to informational "privacy" and states the principle of the individual wish and right to decide about the disclosure of personal health data [1, 8]. Based on the Declarations of Helsinki (1964 and 1975) from the World Medical Association, it is a basic right of the patient to be assured that all his medical and personal data are confidential. The health professional who has obtained such data has a primary obligation to respect the confidentiality of the data and to safeguard them against any disclosure. Only in the case of a few well-defined exceptions is disclosure allowed, e.g. prevention of serious risk to public health, order by a court of law in a crime case, and, under certain safeguards, health research (including epidemiologic inquiry) [1, 10] (*see* **Epidemiology as Legal Evidence**).

According to Thompson, the concept of confidentiality refers to three principal values, namely "privacy", "confidence", and "secrecy" [8]. Privacy – and in epidemiology we mean in most cases "informational" privacy – deals with the right of individuals to control their own lives, while confidence is an essential requirement of the doctor–patient relationship. Abuse of the trust of confidence the patient places in his doctor would make the practice of medicine impossible. "Secrecy" can be seen as a complementary factor to individual privacy, but from the perspective of the professional dealing with the question of what the patient is allowed to know about his own records.

Ethical principles may affect research and the way we deal with confidentiality in many ways [2, 10, 11]. First, there is the question of the decision to do or not to do a study. Then the legal framework, including guidelines on how to conduct a study, is important. There is a considerable range of types of legislation to protect individual confidentiality in time and among countries. This perspective deals also with the **data management** and disclosure approaches in balancing the interests of science while respecting confidentiality rules and guidelines [6] (*see* **Ethics of Randomized Trials**).

Ethical Principles

Respect for confidentiality of persons involved in clinical and epidemiologic research has its origins in the fulfillment of relevant ethical principles. In general, four ethical principles can be distinguished [2, 6].

1. *Beneficence*. People in general, and the same is true for epidemiologists, have a moral obligation to do "right", i.e. to benefit individuals and society. The results of a research project should add to the existing knowledge base of medicine in order to make patients better, to prevent health hazards, or to decrease mortality.
2. *Nonmaleficence*. This principle reflects a moral obligation not to do harm to the persons involved in a scientific study. Harm can, under certain circumstances, be justified when the population benefits outweigh the individual harm: e.g. a **Phase I trial** in oncology almost never benefits the patients in the study, but may benefit other patients in the future.
3. *Autonomy*. The principle of autonomy states the moral obligation to respect the right to self-determination. Autonomy is the key principle for respecting confidentiality [1]. The demand for informed consent given by the persons involved in a study reflects the fulfillment of the principle of autonomy.
4. *Justice*. Justice can be considered as the principle of a fair distribution of burdens and benefits between individuals, and between groups in society. This principle may mean equal access to study participation and subsequent benefit (e.g. in **AIDS** trials), as well as equal exposure when certain outcomes are still uncertain (e.g. **post-marketing** studies with new drugs).

A useful approach in applying such principles is the assessment of each ethical principle in the context (or scope) of a specific study. Nilstun & Westrin have proposed to apply these principles from the perspective of each of the parties involved and then to assess the ethical "benefits" and "costs" in the event the study is or is not conducted [6]. This process can be illustrated by an example, in deciding whether to do or not a **pharmacoepidemiologic** study on the risk of hip fracture in patients using benzodiazepines [3]. In this example we may identify two relevant parties, i.e. the persons included in the study and society at

Table 1 The most important possible "benefits" and "costs" when the study is done [3]

	Bene-ficence	Nonmale-ficence	Auto-nomy	Justice
Persons in the study			Costs	
Society at large	Benefit			Benefit

large. In Table 1 a possible outcome of an analysis of the most relevant "benefits" and "costs" is listed concerning the two dimensions of ethical principles and parties involved in the event that the study is conducted. "Benefits" and "costs" are essentially exchanged if the study is not conducted (*see* **Health Economics**).

If the study is done, there are possible "benefits" for society at large because the study strengthened a hypothesis based on earlier findings and its results can guide prescribers in rationing the use of these drugs. There could be potential costs with respect for autonomy by violating the privacy of the patients in the study. Data on prescription drug use had to be linked to cases of hospitalizations for hip fracture without knowing the patient's identity. All this was done with existing data and applying a probabilistic approach in relating different datafiles to the same individual using dates of birth, gender, and physician practice [3]. For society at large, potential "benefits" to the principle of justice can be stated. Justice means a fair contribution to the gain of relevant medical knowledge, obviously within the boundaries of economics and other structural conditions. By "participating" in such a study the population involved took its share in the solidarity of bringing together relevant pieces of clinical and epidemiologic insight.

Legal Framework and Guidelines

There is great international variety in legislation and practice guidelines on protecting violation of confidentiality. Several international professional organizations have developed ethical guidelines and recommendations for epidemiologic studies, including those produced by the **World Health Organization (WHO)**, the Industrial Epidemiology Forum (IEF), the International Epidemiological Association (IEA), and the Council of International Organizations of Medical Sciences (CIOMS), and the guidelines of the International Society for Pharmacoepidemiology (ISPE).

Recently, the greatest attention has been paid to the directives of the European Union [4]. There has been ample expression of public and professional concern against these directives, principally because "privacy" is deemed to be violated even in epidemiologic studies where confidentiality is assured, unless the particular purpose is approved by all individuals [5, 7, 9]. The intent of the European directives is to protect individuals from improper administrative use of personal data, including medical data, although no specific details in this direction were given. The definition of "personal data" is crucial here, because it relates to the question of how much effort is needed to identify an individual person within the format of the data. Apart from many other criteria for and conditions about data confidentiality, "express and written informed consent" and "personal data" are the two basic features of European directives that are most critical.

Record linkage is a key methodology in epidemiology and various techniques (e.g. **Probabilistic matching** and encryption techniques) have been developed to cope with the confidentiality issue [3, 5]. Some privacy advocates, however, continue to argue that written informed consent from all patients would be required to do such linkage studies, even if all the data are processed in a fully anonymous fashion. The creation of unbiased personal histories (including both data on various exposure and outcomes) is a crucial requirement in epidemiology. Record linkage, as was done in the case study on hip fractures, would be virtually impossible if a requirement of full written informed consent were imposed.

Discussion

Protecting confidentiality in the context of scientific inquiry is one of today's paradoxes. The paradox confuses us because it requires us to live simultaneously with opposites. The paradox here is the growing ability and need to apply advanced data systems to investigate health hazards related to various exposures (*see* **Administrative Databases**), and on the other hand the increase of legal control over data collection and procedures to use these data for epidemiologic research. Major progress is being made in automated databases and information technology

to establish effective strategies in the use of data for epidemiologic research. Linking existing data can be an effective and efficient way to study various exposures and population outcomes. However, society is increasingly concerned about violating the privacy of individuals. Ethical controversies on confidentiality may affect and obstruct epidemiologic research in a significant way [4, 9]. On the other hand, ethical principles may be important to support the conduct of research and to guide decision making in this respect.

Westrin & Nilstun have compared the aims and tasks of both epidemiologists and journalists in terms of their responsibility towards the protection of confidentiality [11]. Society seems willing to accept that, in the interests of wider public good, journalism may sometimes invade individuals' privacy and do them harm, but it is not prepared to offer epidemiology an equal measure of tolerance. Confusion still surrounds the question of whether confidentiality can be fundamentally violated by data drawn from personal records. Ethical conflicts between moral principles and methodologic standards affecting epidemiologic research will remain. The future lies in thoughtful weighing of the various "costs" and "benefits" in such conflicts.

References

[1] Beauchamp, T.L. & Childres, J.F. (1989). *Principles of Biomedical Ethics*, 3rd Ed. Oxford University Press, New York.

[2] Gillon, R. (1994). Medical ethics: four principles plus attention to scope, *British Medical Journal* **309**, 184–188.

[3] Herings, R.M.C., Stricker, B.H.Ch., Boer, de A., Bakker, A. & Sturmans, F. (1995). Benzodiazepines and the risk of falling leading to femur fractures: dosage more important than elimination half-life, *Archives of Internal Medicine* **155**, 1801–1807.

[4] Knox, E.G. (1992). Confidential medical records and epidemiological research, *British Medical Journal* **304**, 727–728.

[5] Newcombe, H.B. (1995). When "privacy" threatens public health, *Canadian Journal of Public Health*, **86**, 188–192.

[6] Nilstun, T. & Westrin, C.G. (1994). Analyzing ethics, *Health Care Analysis*, **2**, 43–46.

[7] Olsen, J., Breart, G., Feskens, E., Grabauskas, V., Noah, N., Olsen, J., Porta, M. & Saracci, R. The International Epidemiological Association-IEA European Epidemiological Group (1995). Directive of the European Parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data, *International Journal of Epidemiology* **24**, 462–463.

[8] Thompson, I.E. (1979). The nature of confidentiality, *Journal of Medical Ethics*, **5**, 5.

[9] Vandenbroucke, J.P. (1992). Privacy, confidentiality and epidemiology: the Dutch ordeal, *International Journal of Epidemiology* **21**, 825–826.

[10] Weed, D.L. (1994). Science, ethics guidelines, and advocacy in epidemiology, *Annals of Epidemiology* **4**, 166–171.

[11] Westrin, C.G. & Nilstun, T. (1994). The ethics of data utilisation: a comparison between epidemiology and journalism, *British Medical Journal* **308**, 522–523.

HUBERT G. LEUFKENS

Confidentiality

Confidentiality in the context of health and biostatistical research concerns the avoidance of disclosure of sensitive and identifiable information about individual patients to a third party. Confidentiality procedures have been subject to change during recent decades. Patient treatment has increased in complexity, frequently involving both primary health care workers and specialists, often in several hospitals, and more persons thus have a need for access to data on an individual patient, but they may also obtain access to data which are irrelevant for the medical service to be provided. Patient data that were once used almost exclusively by the treating physician are thus often shared with others, including nonmedical persons, to a far greater extent than in the past, particularly for research purposes. This research is generally intended for the benefit of the whole population in identifying causes of disease, evaluating the outcome of treatment, assessing equity in health and in access to treatment services (*see* **Health Services Research, Overview**), etc. The availability and widespread use of computer systems to store, analyze, and transmit large volumes of data, sometimes over public data networks, have radically altered the climate in which patient confidentiality must be maintained (*see* **Administrative Databases**). These changes in the use of confidential data have coincided with heated debate on the **ethics of randomized trials**. Randomized **clinical trials** are essential research tools for identifying optimal future treatments, but they are not always readily understood by the general public. Prospective research such as randomized trials, involving direct recruitment of living patients, requires the informed consent of the patient in many countries today. This consent will usually also set the terms for use of the data.

The ethical issue of confidentiality is more complex where the data subject is not contacted, and may no longer be alive, even if the results of the research do not enable the individual to be identified. This situation arises when data that are collected, for purposes such as routine **surveillance of disease** or **vital statistics**, often under the aegis of government, or for hospital administration or **occupational health**, are collated from available sources and the records for a given individual linked for analysis (*see* **Record Linkage**). The results of such research may provide powerful new insights into trends in the health of the population without any need for individuals to be identified. The possibilities for computerized linkage of data for individuals, even for very large volumes of data, have increased public fears of misuse and of error, and they have stimulated a continuing public debate on ethics and confidentiality in health research (*see* **Confidentiality in Epidemiology**).

What Should be Kept Confidential?

Data given in confidence must clearly be treated as confidential. Health data obtained during the doctor's management of a patient should be regarded as confidential, but may, in the interests of the patient, be transmitted to other physicians involved in the treatment of that patient, and physicians will be expected to observe professional confidentiality. Responses given to an interviewer or written on a form as part of a health survey (*see* **Surveys, Health and Morbidity**) should also be treated as confidential information. Often, a promise of confidentiality is explicitly given to survey respondents to improve the quality and completeness of the information being collected.

The interests of society may override the individual's perceived right to absolute confidentiality when, for instance, the health effects of environmental pollution (*see* **Environmental Epidemiology**), or occupational exposure (*see* **Occupational Epidemiology**) need to be assessed and controlled. Publication and interpretation of the results of such research do not require identification of individuals, whereas it is crucial to the underlying analysis. In such cases, the confidentiality of data about individuals must be observed to the fullest extent possible, and preferably regulated by an independent body. Almost any information may be considered confidential by an individual – such as memberships of unions, income, tax, childbirth, etc. Some obviously identifiable data in the public domain are not considered confidential (e.g. names, addresses, and telephone numbers). Often, however, when joining a union or becoming a member of an organization, the individual may allow use of the data for purposes relevant to their membership.

Who is Entitled to Confidentiality and How is it Preserved?

We are all entitled to confidentiality, to have a space of our own. The problem arises when this space is

of interest to others, in particular to the society in which we have chosen to be members, for example when dealing with public health issues. Confidentiality not only concerns relations with the patient – the data subject – but also the data providers, and it must be preserved in all aspects of data collection, storage, research use, and transmission. Consequently, all persons who have access to personal data must be expected to obey the same constraints with regard to confidentiality (*see* **Data Management and Coordination**). In health, medical confidentiality is part of the professional ethic embodied in the Hippocratic oath, and is set out in most publications on good medical practice. Physicians may lose their license to practice, as may lawyers, if they breach patient–client confidentiality. Other persons such as epidemiologists with access to confidential data should also be subject to such rules. This is not always the case at present [3].

Existing Guidelines on Confidentiality and Security in Manual and Automated Systems

Preservation of confidentiality has been a long-standing tradition in epidemiologic research, in particular in cancer registration (*see* **Disease Registers**), dating back to the 1940s. Automated cancer registration based on electronically available data presents a new challenge, and this also needs to be addressed from the point of view of confidentiality and data security. In principle, security and confidentiality for electronic data collections should follow exactly the same rules and standards as for traditional (manual) registry systems.

Guidelines exist from the International Association of Cancer Registries [2]. These complement international legislation such as the European Union directive "On the Protection of Individuals with Regard to the Processing of Personal Data and the Free Movement of Such Data" [5], recommendations from the Council of Europe, national legislation in the form of data protection acts, and international recommendations on ethics [6, 7].

In manual systems, security depends on dispersion of the data, and the relative difficulty of getting access to data and in linking these with other data. This does not safeguard individual data, which may be disclosed to third parties if the physical security of premises and files is not carefully observed. Coleman et al. [2] list a number of situations where precautions must be taken to avoid accidental breaches of confidentiality by staff, such as use of the telephone and telefax for communicating confidential information, improper disposal of paper records, and insecure transport of data by mail in addition to access control, and so on.

In electronic systems, large volumes of data are gathered in a structured manner, and if confidentiality is not observed, disclosure may concern a much larger number of individuals. It is possible to safeguard electronic systems much better than manual systems, however, not only against unauthorized access, but also by monitoring both authorized access and data in transit between provider and registry (*see* **Confidentiality and Computers**).

It is obvious that if researchers lose the ability to link data on individuals unequivocally, both maintenance of registries such as those on cancer and a great deal of epidemiologic and public health research will be impossible to perform. It is thus incumbent on registries and researchers to preserve the confidentiality of identifiable data in the interest of their own professional activities, irrespective of any official requirements. If confidentiality is breached, years of planning and work on specific research projects may be lost.

The Threat of Breach of Confidentiality

It should also be clear that improper disclosure of data depends on two things – the value of the data, and the number of people who have access to them. Those who see a value in identifiable data are also likely to be those through whom a breach in confidentiality may occur. Registries and researchers may be pressurized to disclose information to parties who believe they have a legal interest in knowing the details – one example being to check if persons with a disease possibly related to occupational asbestos exposure have been reported by the physicians to all relevant bodies [4, 9]. Here, the confidentiality of the data provider and the patient is at stake, competing with the demands of society to ensure that rules and regulations are followed. Another situation is a wish for access from insurance companies, which may provide no benefit to either the individual or society, but rather to the company.

If trust in registries or researchers is lost, there will be an erosion of data quality, inasmuch as sensitive

facts may be suppressed. The value of the registry and the data will drop, and the use for which it was intended may become impossible. It is thus important to have written rules that apply to health researchers using identifiable data, and these must regulate access, specify duties, and impose penalties for any breach, unless this is covered by national legislation.

Practical Means of Preserving Confidentiality

In systems with automated (computerized) data collection from a variety of sources, the number of persons involved with some access to the electronic data increases significantly. It is therefore necessary to control access; to impose passwords and restrictions for various user types; to take further precautions when transmitting, collecting, and analyzing data; and to consider failures in both software and hardware that might corrupt data. Table 1 summarizes the actions that can be taken to preserve confidentiality and data security.

Monitoring Use and Access

Control and monitoring (logging) of access and notification of data subjects and regulatory bodies on access is one basic measure in preserving confidentiality. Even if this does not prevent criminal actions, the fact that access is monitored and that relevant authorities will be notified of any breach of rules should decrease the likelihood of misuse due to carelessness. If misuse occurs, adequate action must be taken. Access to data can also be made user-specific, a mechanism by which only persons with a specific need for knowing the identity of individuals can obtain access to such information – and only for the time period for which this access is needed – whereas access to statistical, tabular, or anonymous person records can be more liberal. In other words, there can be different levels of access and of access control. Users should learn that all access will be monitored and logged.

Anonymous Data File Separation

Various methods have been proposed to safeguard confidentiality. It is relatively easy to make data

Table 1 Actions for preserving confidentiality and data security

Area of concern	Actions
Access control	Passwords Limited access (certain data or certain periods) Access logs – notification (regulatory bodies/data subject)
ID safeguards (data subject and data provider)	Clear rules for authorizing access to identifiable information Making data anonymous – use of keys and file separation Encryption of identifiers Statistical data base (no identifiers) for "public" use
Data use	Monitoring of users and usage
Data security	Transmission – encryption/decryption Stand-alone registry computer system Restricted access from outside – "fire-wall" Dialback methods (combined with encryption) Confidential destruction of used equipment – tapes, hard disks, etc. Test system without real data for development, etc.
Consent	From the data subjects From data subject group representatives (unions, etc.) From institutions (employer, etc.) From ethical committees From data inspection agencies
Legal actions	"Hippocratic oath" or equivalent for all professions Loss of license to practice if violation of confidentiality rules is proven Statutory penalties

anonymous by separating identity information from the data file, and keeping a code detached from the computer system by which the identity of an individual can be linked back to the medical information. This is advisable for research data sets and for PCs in hospitals and clinics where theft of equipment (data on hard disks) poses a major risk for unwarranted breach in confidentiality. Another way of doing this is by means of encryption of the identifying information, with the possibility of subsequent decryption by the researcher holding a secret key or algorithm.

Encryption

Encryption of data is a valuable tool for preserving confidentiality in the communication of data from one place to another. Today, encryption systems exist that allow users to have separate keys for encryption and decryption [1]; the first can be published while the second can be kept secret. Decryption can thus be performed where the necessary precautions for data security are in place, and linkage procedures and quality control can be performed on data with no question about the identity and correctness of linkages.

A cancer registry system based on a complicated system of data encryption has been proposed in Germany [8]. All linkages would be done on encrypted data on the basis of identifying information associated with name and date of birth. Although **false positive** matches were uncommon (<1%), **false negative** matches (duplicates) were higher. Whereas these errors may have little impact on rates and **descriptive epidemiology**, the inability to link cases with absolute reliability for follow-up studies (where more linkages are often involved) is much more serious, since the observed numbers of events may be wrong, and the expected number may also be biased by failure to censor individuals at death, since no link is obtained between data stored at entry and the date of exit (death). It may not be meaningful to conduct registry-based studies in such a setting. It will be of both practical and ethical concern that the researcher cannot control the quality of data, and confidence in the results will be low.

Isolation

Another approach to security is to use technical solutions that make unauthorized access impossible or very difficult. The ultimate solution is the "stand-alone" registry computer system. This may be designed in such a way that the registry still has access to the outside world, whereas external contact with the registry is prohibited by a so-called "firewall" that only allows one-way traffic. Another measure is dialback, where only certain telephone numbers and users (access codes) are allowed a line to the registry, and the request is processed by the registry computer, which dials back the authorized person who requested information or access. Data transmission in such cases can be safeguarded by encryption and decryption, using a common key at both ends.

Confidentiality and data security go hand in hand. Measures need to be taken when implementing new **software** and hardware, where all testing must be performed on test data, not real data. Any electronic media – hard disks, floppy disks, and tapes – must be suitably erased or destroyed if taken out of use. Data discipline is also needed, in the sense that old data are not deleted but changes are appended with dates and a record of who carried out the addition. Precautions must be taken to avoid corruption of data.

Conclusion

Confidentiality must continue to be taken seriously. It should not become prohibitive for research, however, since the results will benefit future generations. We must answer the question "Who are we protecting, and why?" If, in our desire for absolute privacy, we allow unjustifiable concerns for confidentiality to prevent ethical research, we risk protecting the "criminal" rather than our society. Thus we may protect the employer who is exposing employees to carcinogenic substances, or health services that perform poorly, or industries that pollute our environment. We may thus be unable to demonstrate the effect of powerful carcinogens such as tobacco smoking in the future (*see* **Smoking and Health**). We must also balance individual rights with the rights of the society in which we have chosen to live. This balance between the individual's right to privacy and rights of society to uncover hazards is not easy to strike, but if we accept regulation and surveillance by independent bodies, such as democratically elected ethical committees and data protection agencies, we may avoid reintroduction of the Middle Ages in health research. It is often said that those who demand complete confidentiality and privacy are also those who demand control of every

conceivable environmental or occupational hazard to health. Adequate control without prior study is impossible, and such contradictory demands are thus equally impossible to meet. So far, we have not experienced any major breach in confidentiality of identifiable data entrusted to the health research community. Ironically, breaches of confidentiality for criminal records and credit ratings have been widely reported, as have breaches of highly sensitive defense computers. The good record on confidentiality in the domain of health research is due in part to the wide respect for professional ethics, but also to the fact that health researchers are well aware that such a breach could spell the end of much **population-based** research.

References

[1] Anderson, R.J. (1996). *Security in Clinical Information Systems*. University of Cambridge (Internet publication `http://hypatia.dcs.qmw.ac.uk/authors/A/AndersonRJ/papers/policy.txt`).

[2] Coleman, M.P., Ménégoz, F. & Muir, C.S. (1992). Guidelines on confidentiality in the cancer registry, *British Journal of Cancer* **66**, 1138–1149.

[3] Coughlin, S.S. (1996). Advancing professional ethics in epidemiology, *Journal of Epidemiology and Biostatistics* **1**, 71–77.

[4] Danø, H., Skov, T. & Lynge, E. (1996). Underreporting of occupational cancers in Denmark, *Scandinavian Journal of Work Environment and Health* **22**, 55–57.

[5] European Union (1995). Directive 9/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data. European Union, Brussels.

[6] Gordis, L., Gold, E. & Seltser, R. (1977). Privacy protection in epidemiologic and medical research, *American Journal of Epidemiology* **105**, 163–168.

[7] Last, J. (1996). Professional standards of conduct for epidemiologists, in *Ethics and Epidemiology*, S.S. Coughlin & T.L. Beauchamp, eds. Oxford University Press, New York.

[8] Michaelis, J., Miller, M., Pommering, K. & Schmidtmann, I. (1995). A new concept to ensure data privacy and data security in cancer registries, *Medinfo* **8**, 661–665.

[9] Skov, T., Mikkelsen, S., Svane, O. & Lynge, E. (1990). Reporting of occupational cancer in Denmark, *Scandinavian Journal of Work Environment and Health* **16**, 401–405.

HANS H. STORM

Confounder Summary Score

Consider an observational study of the effect of an "exposure" variable X on an outcome variable Y in which multiple **confounders** must be controlled (*see* **Confounding**). Simultaneous **stratification** on all observed confounder combinations may lead to many uninformative strata (i.e. strata in which there is either no variation in the exposure or no variation in the outcome). The usual method of coping with this problem is to estimate exposure effects from coefficients in a **regression** model for the dependence of the outcome on the exposure and confounders. Confounder summary scores are alternatives that use fitted models to define strata.

Parametric modeling raises concerns about the dependence of the resulting effect estimates on the model specification (*see* **Model, Choice of**). In one analysis of the National Halothane Study (NHS), the outcome was regressed on confounders, and the data were then stratified on the fitted values from the regression model [2]. In this manner, the problem of multiple confounders was reduced to stratification on just one variable, the fitted outcome $\hat{Y}$. It was later noted that this scoring procedure produces **biased** effect estimates unless the resulting fitted values are modified by removing the estimated exposure effect [3, 4]. To illustrate these ideas, let X denote the treatment or exposure of interest and let $\mathbf{Z}$ denote the row vector of confounders. The National Halothane procedure involved fitting a model such as

$$g[\mathrm{E}(Y|\mathbf{Z} = \mathbf{z})] = \alpha^* + \mathbf{z}\boldsymbol{\gamma}^*,$$

and then stratifying subjects on their fitted values:

$$\hat{Y} = g^{-1}(\hat{\alpha}^* + \mathbf{z}\hat{\boldsymbol{\gamma}}^*).$$

Miettinen instead fit

$$g[\mathrm{E}(Y|X = x, \mathbf{Z} = \mathbf{z},)] = \alpha + x\beta + \mathbf{z}\gamma.$$

He then stratified subjects on the modified score $g^{-1}(\hat{\alpha} + \mathbf{z}\hat{\boldsymbol{\gamma}})$ (or, equivalently, on $\hat{\alpha} + \mathbf{z}\hat{\boldsymbol{\gamma}}$), the fitted value obtained by deleting the estimated exposure effect $x\hat{\beta}$. The exposure effect $x\hat{\beta}$ is left out of the scoring to ensure that the strata are not defined in part by exposure. The inclusion of exposure when fitting the model also serves this purpose: note that the confounder coefficient, $\boldsymbol{\gamma}^*$, in the model without exposure may carry some of the exposure effects unless X and $\mathbf{Z}$ are independent.

These modified fitted values or linear predictors are an example of *confounder scores*, and the scoring process is called *confounder summarization*. Assuming the fitted model is correct, it has been shown that adjustments using these modified scores could lead to effect estimates unconfounded by $\mathbf{Z}$, but could also overstate significance (i.e. yield downwardly biased ***P* values**) for testing the **null hypothesis** of no exposure effect ($\beta = 0$), whereas the unmodified score used in the original National Halothane approach would generally yield biased estimates of β, but could yield valid significance levels for testing the null [4].

An alternative approach is to control confounding by stratifying on the fitted exposure values obtained by regressing exposure X on the confounders $\mathbf{Z}$. For binary X, Rosenbaum & Rubin [6] termed the resulting fitted values ($\hat{X}$) **propensity scores** and showed that these scores have a number of desirable properties. In particular, Rosenbaum & Rubin showed that, by stratifying on propensity scores, one could obtain valid effect estimates and significance levels from the same model. Scores based on outcome regression are sometimes referred to as *risk scores* or *prognostic scores*, while scores based on exposure regression are sometimes called *exposure scores*. Propensity scores are sometimes referred to as *balancing scores*, reflecting their use in creating strata such that **covariate** distributions are "balanced" across exposure groups.

Apart from propensity scores, confounder summarization methods have seen relatively little use since their initial development. This disuse may be attributable to a number of factors. One problem is that confounder summarization methods are not as insensitive to model **misspecification** as was hoped [1]. As an extreme but transparent example, suppose there is but one confounder, Z, that X given Z has a standard **normal distribution**, that Z has a standard normal marginal distribution, and that the regression of X on Z is

$$\mathrm{E}(X|Z = z) = z^2.$$

If the model fitted for exposure scoring is

$$\mathrm{E}(X|Z = z) = \alpha + \beta z,$$

then the ordinary **least squares** estimate β will have zero expectation and the resulting exposure

scores will stratify subjects randomly, with little confounder control achieved by the stratification. Similarly discouraging examples can be constructed for risk scores.

Of course, careful modeling should detect misspecification as gross as just illustrated. Nonetheless, the example points out that confounder summarization may require as much modeling effort as ordinary analysis. Even more effort may be needed if multiple exposures are studied, for then separate exposure scores must be constructed for each exposure. Thus, in observational studies, there may be no convenience and little **robustness** advantage of confounder summarization over direct model-based estimation. Any robustness advantage may be further diminished when **nonparametric regression** methods can be used to estimate exposure effects.

The interpretation of strata and estimates constructed from confounder scores can also be difficult. While the confounder distributions may be balanced within strata, the strata will usually contain subjects with a heterogeneous mix of confounder profiles, the comparability of which may not be immediately obvious to a clinical reader. When the exposure effect varies across strata, it may be necessary to return to standard methods to identify the source of the variation.

Exposure regression can be used directly as part of a system of models for control of confounding in effect **estimation** [5]. This use, however, is distinct from its use in confounder summarization.

References

[1] Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect, *Biometrics* **49**, 1231–1236.

[2] Halpern, J., Moses, L.E. & Bishop, Y.M.M. (1969). Analysis by regression methods, in *The National Halothane Study*, J.P. Bunker, W.H. Forrest & F. Mosteller, eds. National Institute of General Medical Sciences, Bethesda, Chapter IV-5.

[3] Miettinen, O.S. (1976). Stratification by a multivariate confounder score, *American Journal of Epidemiology* **104**, 609–620.

[4] Pike, M.C., Anderson, J. & Day, N.E. (1979). Some insights into Miettinen's multivariate confounder score approach to case–control study analysis, *Journal of Epidemiology and Community Health* **33**, 104–106.

[5] Robins, J.M. & Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial, *Journal of the American Statistical Association* **89**, 737–749.

[6] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

SANDER GREENLAND

Confounder

As used in epidemiology, a confounder is a factor that is associated with the risk of disease in subjects unexposed to the exposure of interest, that is not affected by exposure or disease, and that is associated with exposure in the source population from which cases arise. For example, the risk of cancer increases with age. To study an exposure that is associated with age, such as cumulative coffee consumption, as a risk factor for cancer, one needs to control for the **confounding** effects of age. Because the confounder is associated both with disease risk and with exposure status, failure to account for the confounder either by appropriate choice of study design, such as a restricted design or a stratified design (*see* **Stratification**), or by analytical adjustments (*see* **Standardization Methods**) can lead to misleading estimates of the relationship between the exposure of interest and the risk of disease (*see* **Confounding**; **Matched Analysis**; **Matching**). However, applying such adjustment methods to a factor that is affected by exposure or disease, such as an intermediate effect of exposure on the pathway leading from exposure to disease (*see* **Causation**) can misleadingly reduce estimates of the strength of **association** between exposure and disease (*see* **Odds Ratio**; **Relative Risk**).

Confounding can be described more fundamentally as a distortion in estimates of **exposure effect** that results when responses from an unexposed **control** population are used to estimate the hypothetical responses that would have been observed in the exposed population had that population been unexposed. In this context, confounders are factors that account for differences between observed control responses and the hypothetical responses in the exposed group that would have been observed had that group been unexposed. A disadvantage of this formulation is that one cannot verify directly that confounding is present, because one does not observe the hypothetical responses of the exposed population had it been unexposed. The criteria for confounding in the previous paragraph, although less fundamental than the definition just given, are at least useful indicators of confounding, although they can be misleading in some situations [1, 2].

References

[1] Greenland, S. & Robin, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding, *International Journal of Epidemiology* **15**, 413–419.

[2] Pearl, J. (1995). Causal diagrams for empirical research (with discussion), *Biometrika* **82**, 669–710.

MITCHELL H. GAIL

Confounding

The word *confounding* has been used to refer to at least three distinct concepts. In the oldest usage, confounding is a **bias** in estimating causal effects (*see* **Causation**). This bias is sometimes informally described as a mixing of effects of extraneous factors (called **confounders**) with the effect of interest. This usage predominates in nonexperimental research, especially in epidemiology and sociology. In a second and more recent usage, confounding is a synonym for noncollapsibility (*see* **Collapsibility**), although this usage is often limited to situations in which the parameter of interest is a causal effect. In a third usage, originating in the **experimental-design** literature, confounding refers to inseparability of main effects and **interactions** under a particular design. The term *aliasing* is also sometimes used to refer to the latter concept; this usage is common in the **analysis of variance** literature.

The three concepts are closely related and are not always distinguished from one another. In particular, the concepts of confounding as a bias in effect estimation and as noncollapsibility are often treated as identical, although there are many examples in which the two concepts diverge [8, 9, 14]; one is given below.

Confounding as a Bias in Effect Estimation

Confounding

A classic discussion of confounding in which explicit reference is made to "confounded effects" is Mill [15, Chapter X] (although in Chapter III Mill lays out the primary issues and acknowledges Francis Bacon as a forerunner in dealing with them). There, he lists a requirement for an experiment intended to determine causal relations:

> ... none of the circumstances [of the experiment] that we do know shall have effects susceptible of being *confounded with* those of the agents whose properties we wish to study (emphasis added).

It should be noted that, in Mill's time, the word "experiment" referred to an observation in which some circumstances were under the control of the observer, as it still is used in ordinary English, rather than to the notion of a comparative trial. Nonetheless, Mill's requirement suggests that a comparison is to be made between the outcome of his experiment (which is, essentially, an uncontrolled trial) and what we would expect the outcome to be if the agents we wish to study had been absent. If the outcomes is not as one would expect in the absence of the study agents, then his requirement ensures that the unexpected outcome was not brought about by extraneous circumstances. If, however, those circumstances do bring about the unexpected outcome, and that outcome is mistakenly attributed to effects of the study agents, then the mistake is one of confounding (or confusion) of the extraneous effects with the agent effects.

Much of the modern literature follows the same informal conceptualization given by Mill. Terminology is now more specific, with "treatment" used to refer to an agent administered by the investigator and "exposure" often used to denote an unmanipulated agent. The chief development beyond Mill is that the **expectation** for the outcome in the absence of the study exposure is now almost always explicitly derived from observation of a **control** group that is untreated or unexposed. For example, Clayton & Hills [2] state of **observational studies**,

> ... there is always the possibility that an important influence on the outcome ... differs systematically between the comparison [exposed and unexposed] groups. It is then possible [that] part of the apparent effect of exposure is due to these differences, [in which case] the comparison of the exposure groups is said to be *confounded* (emphasis in the original).

In fact, confounding is also possible in randomized experiments (*see* **Clinical Trials, Overview**), owing to systematic improprieties in treatment allocation, administration, and compliance. A further and somewhat controversial point is that confounding (as per Mill's original definition) can also occur in perfect randomized trials due to *random* differences between comparison groups [6, 8].

Various mathematical formalizations of confounding have been proposed. Perhaps the one closest to Mill's concept is based on a formal counterfactual model for causal effects. Suppose our objective is to determine the effect of applying a treatment or exposure x_1 on a parameter μ of population A, relative to applying treatment or exposure x_0. For example, A could be a cohort of breast-cancer patients, treatment x_1 could be a new hormone therapy, x_0 could be a placebo therapy, and the parameter μ could be the 5-year survival probability. The population A

is sometimes called the **target population** or *index population*; the treatment x_1 is sometimes called the *index* treatment; and the treatment x_0 is sometimes called the *control* or *reference* treatment (which is often a standard or placebo treatment).

The counterfactual model assumes that μ will equal μ_{A1} if x_1 is applied, μ_{A0} if x_0 is applied; the causal effect of x_1 relative to x_0 is defined as the change from μ_{A0} to μ_{A1}, which might be measured as $\mu_{A1} - \mu_{A0}$ or μ_{A1}/μ_{A0}. If A is observed under treatment x_1, then μ will equal μ_{A1}, which is observable or estimable, but μ_{A0} will be unobservable. Suppose, however, we expect μ_{A0} to equal μ_{B0}, where μ_{B0} is the value of the outcome μ observed or estimated for a population B that was administered treatment x_0. The latter population is sometimes called the *control* or *reference* population. *Confounding* is said to be present if in fact $\mu_{A0} \neq \mu_{B0}$, for then there must be some difference between populations A and B (other than treatment) that is affecting μ.

If confounding is present, a naive (crude) **association** measure obtained by substituting μ_{B0} for μ_{A0} in an effect measure will not equal the effect measure, and the association measure is said to be *confounded*. For example, if $\mu_{B0} \neq \mu_{A0}$, then $\mu_{A1} - \mu_{B0}$, which measures the *association* of treatments with outcomes *across* the populations, is confounded for $\mu_{A1} - \mu_{A0}$, which measures the *effect* of treatment x_1 on population A. Thus, saying a measure of association such as $\mu_{A1} - \mu_{B0}$ is confounded for a measure of effect such as $\mu_{A1} - \mu_{A0}$ is synonymous with saying the two measures are not equal.

The preceding formalization of confounding gradually emerged through attempts to separate effect measures into a component due to the effect of interest and a component due to extraneous effects [1, 4, 10, 12, 13]. These decompositions will be discussed below.

One noteworthy aspect of the above formalization is that confounding depends on the outcome parameter. For example, suppose populations A and B have a different 5-year survival probability μ under placebo treatment x_0; that is, suppose $\mu_{B0} \neq \mu_{A0}$, so that $\mu_{A1} - \mu_{B0}$ is confounded for the actual effect $\mu_{A1} - \mu_{A0}$ of treatment on 5-year survival. It is then still possible that 10-year survival, ν, under the placebo would be identical in both populations; that is, ν_{A0} could still equal ν_{B0}, so that $\nu_{A1} - \nu_{B0}$ is not confounded for the actual effect of treatment on 10-year survival. (We should generally expect no confounding for 200-year survival, since no treatment is likely to raise the 200-year survival probability of human patients above zero.)

A second noteworthy point is that confounding depends on the target population of **inference**. The preceding example, with A as the target, had different 5-year survivals μ_{A0} and μ_{B0} for A and B under placebo therapy, and hence $\mu_{A1} - \mu_{B0}$ was confounded for the effect $\mu_{A1} - \mu_{A0}$ of treatment on population A. A lawyer or ethicist may also be interested in what effect the treatment x_1 would have had on population B. Writing μ_{B1} for the (unobserved) outcome of B under treatment x_1, this effect on B may be measured by $\mu_{B1} - \mu_{B0}$. Substituting μ_{A1} for the unobserved μ_{B1} yields $\mu_{A1} - \mu_{B0}$. This measure of association is confounded for $\mu_{B1} - \mu_{B0}$ (the effect of treatment x_1 on 5-year survival in population B) if and only if $\mu_{A1} \neq \mu_{B1}$. Thus, the same measure of association, $\mu_{A1} - \mu_{B0}$, may be confounded for the effect of treatment on neither, one, or both of populations A and B.

Confounders

A third noteworthy aspect of the counterfactual formalization of confounding is that it invokes no explicit differences (imbalances) between populations A and B with respect to circumstances or **covariates** that might influence μ [8]. Clearly, if μ_{A0} and μ_{B0} differ, then A and B must differ with respect to factors that influence μ. This observation has led some authors to define confounding as the presence of such covariate differences between the compared populations. Nonetheless, confounding is only a consequence of these covariate differences. In fact, A and B may differ profoundly with respect to covariates that influence μ, and yet confounding may be absent. In other words, a covariate difference between A and B is a necessary but not sufficient condition for confounding. This point will be illustrated below.

Suppose now that populations A and B differ with respect to certain covariates, and that these differences have led to confounding of an association measure for the effect measure of interest. The responsible covariates are then termed *confounders* of the association measure. In the above example, with $\mu_{A1} - \mu_{B0}$ confounded for the effect $\mu_{A1} - \mu_{A0}$, the factors responsible for the confounding (i.e. the factors that led to $\mu_{A0} \neq \mu_{B0}$) are the confounders.

It can be deduced that a variable cannot be a confounder unless it can affect the outcome parameter μ within treatment groups and it is distributed differently among the compared populations (e.g. see Yule [23], who however uses terms such as "fictitious association" rather than confounding). These two necessary conditions are sometimes offered together as a definition of a confounder. Nonetheless, counterexamples show that the two conditions are not sufficient for a variable with more than two levels to be a confounder as defined above; one such counterexample is given in the next section.

Prevention of Confounding

Perhaps the most obvious way to avoid confounding in estimating $\mu_{A1} - \mu_{A0}$ is to obtain a reference population B for which μ_{B0} is known to equal μ_{A0}. Among epidemiologists, such a population is sometimes said to be *comparable to* or *exchangeable with* A with respect to the outcome under the reference treatment. In practice, such a population may be difficult or impossible to find. Thus, an investigator may attempt to construct such a population, or to construct exchangeable index and reference populations. These constructions may be viewed as *design-based* methods for the control of confounding.

Perhaps no approach is more effective for preventing confounding by a known factor than *restriction*. For example, gender imbalances cannot confound a study restricted to women. However, there are several drawbacks: restriction on enough factors can reduce the number of available subjects to unacceptably low levels, and may greatly reduce the generalizability of results as well. **Matching** the treatment populations on confounders overcomes these drawbacks and, if successful, can be as effective as restriction. For example, gender imbalances cannot confound a study in which the compared groups have identical proportions of women. Unfortunately, differential losses to observation may undo the initial covariate balances produced by matching.

Neither restriction nor matching prevents (although it may diminish) imbalances on unrestricted, unmatched, or unmeasured covariates. In contrast, **randomization** offers a means of dealing with confounding by covariates not accounted for by the design. It must be emphasized, however, that this solution is only probabilistic and subject to severe constraints in practice. Randomization is not always feasible, and (as mentioned earlier) many practical problems, such as differential loss and noncompliance, can lead to confounding in comparisons of the groups actually receiving treatments x_1 and x_0. One somewhat controversial solution to noncompliance problems is **intention-to-treat analysis**, which *defines* the comparison groups A and B by treatment assigned rather than treatment received. Confounding may, however, affect even intention-to-treat analyses. For example, the assignments may not always be random, as when **blinding** is insufficient to prevent the treatment providers from protocol violations. And, purely by bad luck, randomization may itself produce allocations with severe covariate imbalances between the groups (and consequent confounding), especially if the study size is small [6, 8, 19]. Block randomization (*see* **Randomized Treatment Assignment**) can help ensure that random imbalances on the **blocking** factors will not occur, but it does not guarantee balance of unblocked factors.

Adjustment for Confounding

Design-based methods are often infeasible or insufficient to prevent confounding. Thus there has been an enormous amount of work devoted to analytic adjustments for confounding. With a few exceptions, these methods are based on observed covariate distributions in the compared populations. Such methods can successfully control confounding only to the extent that enough confounders are adequately measured. Then, too, many methods employ parametric models at some stage, and their success may thus depend on the faithfulness of the model to reality. These issues cannot be covered in depth here, but a few basic points are worth noting.

The simplest and most widely trusted methods of adjustment begin with **stratification** on confounders. A covariate cannot be responsible for confounding within internally homogeneous strata of the covariate. For example, gender imbalances cannot confound observations within a stratum composed solely of women. More generally, comparisons within strata cannot be confounded by a covariate that is constant (homogeneous) within strata. This is so regardless of whether the covariate was used to define the strata. Generalizing this observation to a **regression** context, we find that any covariate with a residual variance of zero conditional on the regressors cannot confound regression estimates of effect (assuming

that the regression model is correct). A broader and more useful observation is that any covariate that is unassociated with treatment conditional on the regressors cannot confound the effect estimates; this insight leads directly to adjustments using a **propensity score**.

Some controversy has existed about adjustment for covariates in randomized trials. Although **Fisher** asserted that randomized comparisons were **unbiased**, he also pointed out that they could be confounded in the sense used here (e.g. see Fisher [6, p. 49]). Fisher's use of the word "unbiased" was unconditional on allocation, and therefore of little guidance for analysis of a given trial. The ancillarity of the allocation naturally leads to conditioning on the observed distribution of any pretreatment covariate that can influence the outcome parameter. Conditional on this distribution, the unadjusted treatment–effect estimate will be biased if the covariate is associated with treatment; this conditional bias can be removed by adjustment for the confounders [8, 18]. Note that the adjusted estimate is also unconditionally unbiased, and thus is a reasonable alternative to the unadjusted estimate even without conditioning.

Measures of Confounding

The parameter estimated by a direct unadjusted comparison of cohorts A and B is $\mu_{A1} - \mu_{A0}$. A number of authors have measured the bias (confounding) of the unadjusted comparison by [10, 12]

$$(\mu_{A1} - \mu_{B0}) - (\mu_{A1} - \mu_{A0}) = \mu_{A0} - \mu_{B0}.$$

When the outcome parameters, μ, are **risks** (probabilities), epidemiologists use instead the analogous ratio

$$\frac{\mu_{A1}/\mu_{B0}}{\mu_{A1}/\mu_{A0}} = \frac{\mu_{A0}}{\mu_{B0}}$$

as a measure of bias [1, 4, 14]; μ_{A0}/μ_{B0} is sometimes called the *confounding risk ratio*. The latter term is somewhat confusing because it is sometimes misunderstood to refer to the effect of a particular confounder on risk. This is not so, although the ratio does reflect the net effect of the differences in the confounder distributions of A and B.

Residual Confounding

Suppose now that adjustment for confounding is done by subdividing the total study population (A + B) into K strata indexed by k. Let μ_{A1k} be the parameter of interest in stratum k of populations A and B under treatment x_0. The effect of treatment x_1 relative to x_0 in stratum k may be defined as $\mu_{A1k} - \mu_{A0k}$ or μ_{A1k}/μ_{A0k}. The confounding that remains in stratum K is called the *residual confounding* in the stratum, and is measured by $\mu_{A0k} - \mu_{B0k}$ or μ_{A1k}/μ_{B0k}.

Like effects, stratum-specific residual confounding may be summarized across the strata in a number of ways, for example by **standardization methods** or by other weighted-averaging methods. As an illustration, suppose we are given a standard distribution $p_1, \ldots, p_K$ for the stratum index k. In ratio terms, the standardized effect of x_1 vs. x_0 on A under this distribution is

$$R_{AA} = \frac{\sum_k p_k \mu_{A1k}}{\sum_k p_k \mu_{A0k}},$$

whereas the standardized ratio comparing A with B is

$$R_{AB} = \frac{\sum_k p_k \mu_{A1k}}{\sum_k p_k \mu_{B0k}}.$$

The overall residual confounding in R_{AB} is thus

$$\frac{R_{AB}}{R_{AA}} = \frac{\sum_k p_k \mu_{A0k}}{\sum_k p_k \mu_{B0k}},$$

which may be recognized as the standardized ratio comparing A and B when both are given treatment x_0, using $p_1, \ldots, p_K$ as the standard distribution.

Regression Formulations

For simplicity, the above presentation has focused on comparing two populations and two treatments. The basic concepts extend immediately to the consideration of multiple populations and treatments. Paired comparisons may be represented using the above formalization without modification. Parametric models for these comparisons then provide a connection to more familiar regression models.

As an illustration, suppose population differences and treatment effects follow the model

$$\mu_k(x) = \alpha_k + x\beta,$$

where the treatment level x may range over a continuum, and k indexes populations. Suppose population k is given treatment x_k, even though it could have been given some other treatment. The absolute effect of x_1 vs. x_2 on μ in population 1 is

$$\mu_1(x_1) - \mu_1(x_2) = (x_1 - x_2)\beta.$$

Substitution of $\mu_2(x_2)$, the value of μ in population 2 under treatment x_2, for $\mu_1(x_2)$ yields

$$\mu_1(x_1) - \mu_2(x_2) = \alpha_1 - \alpha_2 + (x_1 - x_2)\beta,$$

which is biased by the amount

$$\mu_1(x_2) - \mu_2(x_2) = \alpha_1 - \alpha_2.$$

Thus, under this model no confounding will occur if the intercepts α_k equal a constant α across populations, so that $\mu_k(x) = \alpha + \beta x$.

When constant intercepts cannot be assumed and nothing else is known about the intercept magnitudes, it may be possible to represent our uncertainty about α_k via the following mixed-effects model:

$$\mu_k(x) = \alpha + x\beta + \varepsilon_k.$$

Here, α_k has been decomposed into $\alpha + \varepsilon_k$, where ε_k has mean zero, and the confounding in $\mu_1(x_1) - \mu_2(x_2)$ has become an unobserved random variable, $\varepsilon_1 - \varepsilon_2$. **Correlation** of population membership k with x_k leads to a correlation of ε_k with x_k, which in turn leads to bias in estimating β. This bias may be attributed to or interpreted as confounding for β in the regression analysis. Confounders are now covariates that causally "explain" the correlation between ε_k and x_k. In particular, confounders normally reduce the correlation of x_k and ε_k when entered in the model. The converse is false, however: a variable that reduces the correlation of x_k and ε_k when entered need not be a confounder; it may, for example, be a variable affected by both the treatment and the exposure.

Confounding and Noncollapsibility

Much of the statistics literature does not distinguish between the concept of confounding as described above and the concept of noncollapsibility. Nonetheless, the two concepts are distinct: for certain outcome parameters, confounding may occur with or without noncollapsibility and noncollapsibility may occur with or without confounding [8, 9, 14, 17, 20, 22]. Mathematically identical conclusions have been reached by other authors, albeit with different terminology in which noncollapsibility corresponds to "bias" and confounding corresponds to **covariate imbalance** [7, 11].

As an example of no **collapsibility** with no confounding, consider the response distributions under treatments x_1 and x_0 given in Table 1 for a hypothetical index population A, and the response distribution under treatment x_0 given in Table 2 for a hypothetical reference population B. If we take the odds of response as the outcome parameter μ, we get

$$\mu_{A1} = \frac{1460}{540} = 2.70$$

and

$$\mu_{A0} = \mu_{B0} = \frac{1000}{1000} = 1.00.$$

There is thus no confounding of the **odds ratio**: $\mu_{A1}/\mu_{A0} = \mu_{A1}/\mu_{B0} = 2.70/1.00 = 2.70$. Nonetheless, the covariate Z is associated with response and is distributed differently in A and B. Furthermore,

Table 1 Distribution of responses for population A, within strata of Z and ignoring Z, under treatments x_1 and x_0

Subpopulation	Number of responses under x_1	Number of responses under x_0	Subpopulation size
$Z = 1$	200	100	400
$Z = 2$	900	600	1200
$Z = 3$	360	300	400
Totals	1460	1000	2000

Table 2 Distribution of responses for population B, within strata of Z and ignoring Z, under treatment x_0

Subpopulation	Number of responses under x_0	Subpopulation size
$Z = 1$	200	800
$Z = 2$	200	400
$Z = 3$	600	800
Totals	1000	2000

the odds ratio is not collapsible: within levels of Z, the odds ratios comparing A under treatment x_1 with either A or B under x_0 are $(200/200)/(200/600) = (900/300)/(200/200) = (360/40)/(600/200) = 3.00$, a bit higher than the odds ratio of 2.70 obtained when Z is ignored.

The preceding example illustrates a peculiar property of the odds ratio as an effect measure: treatment x_1 (relative to x_0) elevates the odds of response by 170% in population A, yet within each stratum of Z it raises the odds by 200%. When Z is associated with response conditional on treatment but unconditionally unassociated with treatment, the stratum-specific effects on odds ratios will be further from the null than the overall effect if the latter is not null [7]. This phenomenon is often interpreted as a "bias" in the overall odds ratio, but in fact there is no bias if one does not interpret the overall effect as an estimate of the stratum-specific effects.

The example also shows that, when μ is the odds, the "confounding odds ratio" $(\mu_{A1}/\mu_{B0})/(\mu_{A1}/\mu_{A0}) = \mu_{A0}/\mu_{B0}$ may be 1 even when the odds ratio is not collapsible over the confounders. Conversely, we may have $\mu_{A0}/\mu_{B0} \neq 1$ even when the odds ratio is collapsible. More generally, the ratio of crude and stratum-specific odds ratios does not equal μ_{A0}/μ_{B0} except in some special cases. When the odds are low, however, the odds will be close to the corresponding risks, and so the two ratios will approximate one another.

The phenomenon illustrated in the example corresponds to the differences between cluster-specific and population-averaged (marginal) effects in **nonlinear mixed-effects regression** [16]. Specifically, the clusters of correlated outcomes correspond to the strata, the cluster effects correspond to covariate effects, the cluster-specific treatment effects correspond to the stratum-specific log odds ratios, and the population-averaged treatment effect corresponds to the crude log odds ratio.

Results of Gail [7] imply that if the effect measure is the difference or ratio of response proportions, then the above phenomenon – noncollapsibility over Z without confounding by Z – cannot occur, nor can confounding by Z occur without noncollapsibility over Z. More generally, when the effect measure is an expectation over population units, confounding by Z and noncollapsibility over Z are algebraically equivalent. This equivalence may explain why the two concepts are often not distinguished.

Confounding in Experimental Design

Like the bias definition, the third usage of confounding stems from the notion of mixing of effects. However, the effects that are mixed are main (block) effects and interactions (or different interactions) in a linear model, rather than effects in the **nonparametric** sense of a counterfactual model. This definition of confounding differs even more markedly from other definitions in that it refers to an intentional design feature of certain experimental studies, rather than a bias.

The topic of confounded designs is extensive; some classic references are Fisher [6], Cochran & Cox [3], Cox [5], and Scheffé [21]. Confounding can serve to improve efficiency in estimation of certain contrasts and can reduce the number of treatment groups that must be considered. The price paid for these benefits is a loss of **identifiability** of certain parameters, as reflected by aliasing of those parameters.

As a simple example, consider a situation in which we wish to estimate three effects in a single experiment: that of treatments x_1 vs. x_0, y_1 vs. y_0, and z_1 vs. z_0. For example, in a smoking cessation trial these treatments may represent active and placebo versions of the nicotine patch, nicotine gum, and buspirone. With no restrictions on number or size of groups, a fully crossed design would be reasonable. By allocating subjects to each of the $2^3 = 8$ possible treatment combinations, one could estimate all three main effects, all three two-way interactions, and the three-way interaction of the treatments.

Suppose, however, that we were restricted to use of only four treatment groups (e.g. because of cost or complexity considerations). A naive approach would be to use groups of equal size, assigning one group to placebos only (x_0, y_0, z_0) and the remaining three groups to one active treatment each: (x_1, y_0, z_0), (x_0, y_1, z_0), and (x_0, y_0, z_1). Unfortunately, with a fixed number N of subjects available, this design would provide only $N/4$ subjects under each active treatment.

As an alternative, consider the design with four groups of equal size with treatments (x_0, y_0, z_0), (x_1, y_1, z_0), (x_1, y_0, z_1), and (x_0, y_1, z_1). This **fractional factorial design** would provide $N/2$ subjects

under each active treatment, at the cost of confounding main effects and interactions. For example, no linear combination of group means containing the main effect of x_1 vs. x_0 would be free of interactions. If one could assume that all interactions were negligible, however, this design could provide considerably more precise estimates of the main effects than the naive four-group design.

To see these points, consider the following linear model:

$$\mu_{XYZ} = \alpha + \beta_1 X + \beta_2 Y + \beta_3 Z + \gamma_1 XY + \gamma_2 XZ + \gamma_3 YZ + \delta XYZ,$$

where X, Y, and Z equal 1 for x_1, y_1, and z_1, and 0 for x_0, y_0, and z_0, respectively. The group means, in the fractional factorial design are then

$$\mu_{000} = \alpha,$$
$$\mu_{110} = \alpha + \beta_1 + \beta_2 + \gamma_1,$$
$$\mu_{101} = \alpha + \beta_1 + \beta_3 + \gamma_2,$$
$$\mu_{011} = \alpha + \beta_2 + \beta_3 + \gamma_3.$$

Treating the means as observed and the coefficients as unknown, the above system is underidentified. In particular, there is no solution for any main effect β_j in terms of the means μ_{ijk}. Nonetheless, assuming all $\gamma_j = 0$ yields immediate solutions for all the β_j. Additionally assuming a **variance** of σ^2 for each estimated group mean yields that the main-effect estimates under this design would have variances of σ^2, as opposed to $2\sigma^2$ for the main-effect estimates from the naive four-group design of the same size. For example, under the confounded fractional factorial design (assuming no interactions)

$$\hat{\beta}_1 = \frac{\hat{\mu}_{110} + \hat{\mu}_{101} - \hat{\mu}_{000} - \hat{\mu}_{011}}{2},$$

so $\text{var}(\hat{\beta}_1) = 4\sigma^2/4 = \sigma^2$, whereas under the naive design, $\hat{\beta} = \hat{\mu}_{100} - \hat{\mu}_{000}$ so $\text{var}(\hat{\beta}_1) = 2\sigma^2$. Of course, the precision advantage of the confounded design is purchased by the assumption of no interaction, which is not needed by the naive design.

References

[1] Bross, I.D.J. (1967). Pertinency of an extraneous variable, *Journal of Chronic Diseases* **20**, 487–495.

[2] Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, New York.

[3] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York.

[4] Cornfield, J., Haenszel, W., Hammond, W.C., Lilienfeld, A.M., Shimkin, M.B. & Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.

[5] Cox, D.R. (1958). *The Planning of Experiments*. Wiley, New York.

[6] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[7] Gail, M.H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds. Wiley, New York.

[8] Greenland, S. & Robins, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding, *International Journal of Epidemiology* **15**, 413–419.

[9] Greenland, S., Robins, J.M. & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29–46.

[10] Groves, E.R. & Ogburn, W.F. (1928). *American Marriage and Family Relationships*. Henry Holt & Company, New York, pp. 160–164.

[11] Hauck, W.W., Neuhaus, J.M., Kalbfleisch, J.D. & Anderson, S. (1991). A consequence of omitted covariates when estimating odds ratios, *Journal of Clinical Epidemiology* **44**, 77–81.

[12] Kitagawa, E.M. (1955),. Components of a difference between two rates, *Journal of the American Statistical Association* **50**, 1168–1194.

[13] Miettinen, O.S. (1972). Components of the crude risk ratio, *American Journal of Epidemiology* **96**, 168–172.

[14] Miettinen, O.S. & Cook, E.F. (1981). Confounding: essence and detection, *American Journal of Epidemiology* **114**, 593–603.

[15] Mill, J.S. (1843). *A System of Logic, Ratiocinative and Inductive*. Reprinted by Longmans, Green & Company, London, 1956.

[16] Neuhaus, J.M., Kalbfleisch, J.D. & Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data, *International Statistical Review* **59**, 25–35.

[17] Pearl, J. (2000). *Causality*. Cambridge University Press, New York, Ch. 6.

[18] Robins, J.M. & Morgenstern, H. (1987). The mathematical foundations of confounding in epidemiology, *Computers and Mathematics with Applications* **14**, 869–916.

[19] Rothman, K.J. (1977). Epidemiologic methods in clinical trials, *Cancer* **39**, 1771–1775.

[20] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd ed. Lippincott, Philadelphia, Ch. 4.

[21] Scheffé, H.A. (1959). *The Analysis of Variance*. Wiley, New York.

[22] Wickramaratne, P. & Holford, T. (1987). Confounding in epidemiologic studies: the adequacy of the control groups as a measure of confounding, *Biometrics* **43**, 751–765.

[23] Yule, G.U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* **2**, 121–134.

SANDER GREENLAND

Consistent Estimator

Scientists in the fields of medicine and the health sciences frequently collect experimental or observational data to address questions considered to be relevant to the advancement of human health. Examples include **clinical trials** to examine the efficacy of new treatments and epidemiologic studies to better understand the etiology of diseases (*see* **Analytic Epidemiology**). Given that the instruments used for data collection – or the data themselves – are less than perfect, uncertainty is acknowledged in the data analysis stage of scientific inquiry by assuming that the observed data of size n, $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, are generated from a probability (density) function of the form $f(y; \boldsymbol{\theta})$. Here, $\boldsymbol{\theta}$ is a vector of p-dimension, representing parameters characterizing the random mechanism which generated the data. Presumably the parameters $\boldsymbol{\theta}$, or some functions thereof, describe the scientific objectives in a meaningful way. For example, in a clinical trial $\boldsymbol{\theta}$ might characterize the treatment effect for the targeted population (i.e. patients diagnosed with a specific disease). Meanwhile, in an epidemiologic study of the etiology of a disease, $\boldsymbol{\theta}$ might represent the strength of association between a potential risk factor and risk for the disease, often quantified in terms of the **odds ratio**. One goal of data analysis is to learn about the magnitude of $\boldsymbol{\theta}$ from the observed data $\mathbf{Y}$; that is, to *estimate* $\boldsymbol{\theta}$ using $\mathbf{Y}$. **Estimation** is one of several important aspects of statistical **inference**.

An immediate question is how to select a $(p \times 1)$ statistic $\boldsymbol{\delta}(\mathbf{Y})$ to use as an estimator of the unknown value $\boldsymbol{\theta}$. While many criteria for selecting an estimator have been proposed and investigated, one of the most popular is that of **unbiasedness**. Specifically, $\boldsymbol{\delta}(\mathbf{Y})$ is an *unbiased estimator* of $\boldsymbol{\theta}$ if

$$\mathrm{E}(\boldsymbol{\delta}(\mathbf{Y})) = \boldsymbol{\theta}, \tag{1}$$

where the expectation is taken with respect to the true mechanism $f(\cdot; \boldsymbol{\theta})$. A popular interpretation of (1) in nontechnical terms is that if one could repeat the study under the same conditions many times, then $\boldsymbol{\delta}(\mathbf{Y})$, on average, would be equal to the true but unknown $\boldsymbol{\theta}$ value. A further consequence of (1) is that in many cases as the sample size grows, the value of $\boldsymbol{\delta}(\mathbf{Y})$ will grow closer and closer to the true parameter value, $\boldsymbol{\theta}$. This property is due to the **law of large numbers**.

The following two problems, however, hamper the wide utility of unbiasedness as a criterion for selecting estimators of $\boldsymbol{\theta}$. First, unbiasedness is not an invariant property in that, while $\boldsymbol{\delta}(\mathbf{Y})$ is unbiased for $\boldsymbol{\theta}$, it is not true in general that a function $\mathbf{g}(\boldsymbol{\delta}(\mathbf{Y}))$ be unbiased for $\mathbf{g}(\boldsymbol{\theta})$. Secondly, except in the special case of probability models from the **exponential family**, there is very little guidance as to how to obtain unbiased estimators. Indeed, unbiased estimators for $\boldsymbol{\theta}$ may not exist for general probability models.

The **binomial distribution** for independent **binary** observations, which are commonly seen in biomedical research, serves as an excellent example. Suppose for $i = 1, \ldots, n$, that the Y_is are independent and binomially distributed with size one and probability θ. While $\overline{Y} = (Y_1 + \cdots + Y_n)/n$ is known to be unbiased for θ, $\log[\overline{Y}/(1 - \overline{Y})]$ is not unbiased for $\log[\theta/(1 - \theta)]$, the log **odds** of Y_i. Indeed, the expectation of $\log(\overline{Y}/(1 - \overline{Y}))$ does not exist, and no unbiased estimator of $\log[\theta/(1 - \theta)]$, which forms the basis for **logistic regression** models, is known to exist.

Thus, the utility of unbiasedness as a criterion for choosing estimators is limited by the lack of available unbiased estimators. Faced with this situation, statisticians appeal to **large-sample theory** and relax the unbiasedness criterion, while maintaining the requirement that, as the sample size becomes large, the estimator will grow closer and closer to the true $\boldsymbol{\theta}$. Formally, we say that $\boldsymbol{\delta}(\mathbf{Y})$ *converges in probability* to $\boldsymbol{\theta}$ if for any $\varepsilon > 0$,

$$\Pr\{[\boldsymbol{\delta}(\mathbf{Y}) - \boldsymbol{\theta}]'[\boldsymbol{\delta}(\mathbf{Y}) - \boldsymbol{\theta}] > \varepsilon\} \to 0 \quad \text{as } n \to \infty,$$

(*see* **Convergence in Distribution and in Probability**). We define $\boldsymbol{\delta}(\mathbf{Y})$ to be a *consistent estimator* if $\boldsymbol{\delta}(\mathbf{Y}) \to \boldsymbol{\theta}$ in probability as $n \to \infty$. Consistent estimators are desirable for several reasons. First, they are "approximately unbiased" in that, when the sample size is sufficiently large, $\boldsymbol{\delta}(\mathbf{Y})$ will, on average, be close to $\boldsymbol{\theta}$ from the viewpoint of repeated sampling. More importantly, the variability of $\boldsymbol{\delta}(\mathbf{Y})$ *around the true* $\boldsymbol{\theta}$ will vanish as $n \to \infty$. Thirdly, as a result of Slutsky's theorem, consistent estimators are invariant; that is, if $\boldsymbol{\delta}(\mathbf{Y})$ is a consistent estimator of $\boldsymbol{\theta}$, then $\mathbf{g}[\boldsymbol{\delta}(\mathbf{Y})]$ is consistent for $\mathbf{g}(\boldsymbol{\theta})$ as well. Finally, consistency is usually the first step in establishing the asymptotic distribution of an estimator (*see* **Large-sample Theory**). Consistency is further attractive as a criterion because consistent estimators are available in a wide variety of problems.

Returning to the binomial example, the law of large numbers and Slutsky's theorem immediately imply that both $\overline{Y}/n$ and $\log[\overline{Y}/(n-\overline{Y})]$ are consistent estimators of $\boldsymbol{\theta}$ and $\log[\theta/(1-\theta)]$, respectively. Continuing with the binomial example, approximate variances of $\overline{Y}$ and $\log[\overline{Y}/(n-\overline{Y})]$ are, respectively, $\theta(1-\theta)/n$ and $1/(n\theta)+1/[n(1-\theta)]$, decreasing at the rate of n^{-1}. As this example demonstrates, if we accept consistency as a criterion, a variety of estimators are available to the user of statistical methods.

The remaining question is, then: How does one derive consistent estimators for $\boldsymbol{\theta}$? While many estimation methods have been developed in the past century, most of them can be classified into one of the following two types. In the first type, estimators are derived by minimizing with respect to $\boldsymbol{\theta}$ a prespecified objective function of the data $\mathbf{y}$ and parameters $\boldsymbol{\theta}$. Type I methods include as special cases the **maximum likelihood** method, the M-estimation method pioneered by Huber [12] (*see* **Robustness**), and the weighted **least squares** method. The second type of estimators are obtained by simultaneously solving (for $\boldsymbol{\theta}$) p equations containing $\mathbf{y}$ and $\boldsymbol{\theta}$. Examples include the **method of moments**, the **quasi-likelihood** method, and the **estimating functions** method advocated by Godambe [9] and Durbin [6]. One advantage of the first approach is that no differentiability assumption on $\boldsymbol{\theta}$ is required. This is particularly useful when the sampling space depends on $\boldsymbol{\theta}$ [e.g. the **uniform distribution** on $(0,\theta)$] or when the parameter space is discrete. On the other hand, the second approach offers a useful alternative when the probability mechanism generating the data is unclear, and yet the scientific focus is well described by $\boldsymbol{\theta}$. Furthermore, this approach naturally leads to the derivation of the asymptotic distribution of the estimator, an issue addressed elsewhere.

We now discuss briefly how to establish consistency for either type of estimator. The focus is on the basic ideas that are crucial to establishing consistency, and we provide key references for the detailed technical derivations.

Consistency: Type I

This approach aims to find an estimator of $\boldsymbol{\theta}$ by minimizing, with respect to $\boldsymbol{\theta}$, a prespecified objective function $Q_n(\mathbf{y},\boldsymbol{\theta})$ of data $\mathbf{y}=(y_1,\ldots,y_n)'$ and parameter $\boldsymbol{\theta}$. Some familiar examples are given as follows.

Example 1 (Maximum Likelihood Method)

By choosing

$$Q_n(\mathbf{y},\boldsymbol{\theta}) = -\ln f(\mathbf{y};\boldsymbol{\theta}), \tag{2}$$

one has the conventional maximum likelihood estimator (MLE) originated by Fisher [8].

Example 2 (Pseudo Maximum Likelihood Method)

Let $h(\cdot;\theta)$ be a probability (density) function that is specified by the investigator. The idea is to minimize

$$Q_n(\mathbf{y},\boldsymbol{\theta}) = -\ln h(\mathbf{y};\boldsymbol{\theta}), \tag{3}$$

while recognizing that h may not correctly specify the probability mechanism for Y; that is, $h(\mathbf{y};\boldsymbol{\theta}) \neq f(\mathbf{y};\boldsymbol{\theta})$ for some $\mathbf{y}$. This approach was first considered by Huber [13]. The term "pseudo maximum likelihood" was coined in the econometric literature by Gourieroux et al. [10] to emphasize that, due to the complexity of economic phenomena, there is little guarantee that probability models obtained from economic theory are necessarily correct [3]. This concern is a legitimate one in the health sciences as well due to the complexity of bio-psychosocial factors in the human disease process.

Example 3 (M-Estimation)

The MLE for the location parameter under the normality assumption is known to be sensitive to the presence of a small number of **outliers** (**extreme values**) [24]. For example, suppose that interest were on θ, the mean of y. Instead of minimizing $\sum_i (y_i-\theta)^2/2$, which gives rise to $\overline{y}$ as the MLE of θ, Huber [12] proposes to address the problem of such outliers by minimizing

$$Q_n(\mathbf{y},\theta) = \sum_{i=1}^{n} \rho(y_i;\theta),$$

where

$$\rho(y_i,\theta) = \begin{cases} (y_i-\theta)^2/2, & \text{if } |y_i-\theta| \le k, \\ k|y_i-\theta| - k^2/2, & \text{if } |y_i-\theta| > k, \end{cases}$$

and k is a constant pre-specified by the investigator. Many other choices of the function ρ are available and discussed in detail in, for example, Huber [14] and Serfling [23]. Approaches of this kind have

been termed M-estimation and studied extensively, especially in the context of **robust regression** techniques [14].

Example 4 (Weighted Least Squares Method)

The least squares method was invented almost two centuries ago by Legendre [16]. Its purpose is to estimate the regression coefficients θ for the mean of the Y_is conditional on the x_is by minimizing

$$Q_n(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\theta})^2, \tag{4}$$

where $\mathbf{x}_i$ is a $p \times 1$ vector of covariates thought to be related to the response variable, $Y_i, i = 1, \ldots, n$. Implicit behind this procedure is the assumption that the x_is and the "error term", $e_i = Y_i - x_i'\theta$, are uncorrelated with each other for the true θ value. In the situation in which the variance varies with $\mathbf{x}_i'\boldsymbol{\theta}$, the mean value of Y_i given x_i, one might modify (4) by minimizing

$$Q_n(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i' \boldsymbol{\theta})^2}{V_i(\boldsymbol{\theta})}, \tag{5}$$

where $V_i(\boldsymbol{\theta}) = V(\mathbf{x}_i'\boldsymbol{\theta}) = \mathrm{var}(e_i)$. Consistency of the resulting estimator is very sensitive to this last assumption, since the $\boldsymbol{\theta}$ in $V_i(\boldsymbol{\theta})$ now figures in the objective function.

To demonstrate consistency, let $\hat{\boldsymbol{\theta}}_n \equiv \hat{\boldsymbol{\theta}}_n(\mathbf{y})$ be a statistic which minimizes $Q_n(\mathbf{y}, \boldsymbol{\theta})$ over Θ, the parameter space for $\boldsymbol{\theta}$, and let $\boldsymbol{\theta}_0$ be the true but unknown value of $\boldsymbol{\theta}$. The question is: Can we claim that $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$; that is, that $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability? Suppose that Q_n/n converges in probability to $Q_0(\boldsymbol{\theta})$ as $n \to \infty$, where $Q_0(\boldsymbol{\theta})$ is a function determined by the chosen objective function, Q_n, and the true probability model, $f(\cdot; \boldsymbol{\theta}_0)$. Assuming that $Q_0(\boldsymbol{\theta})$ is uniquely minimized by $\boldsymbol{\theta}^*$, intuition (which needs to be made rigorous by additional assumptions and mathematical proofs, of course) suggests that $\hat{\boldsymbol{\theta}}_n$ ought to converge to $\boldsymbol{\theta}^*$ in probability. The claim of consistency is then warranted if $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. Can one be sure that $\boldsymbol{\theta}^*$ is indeed equal to $\boldsymbol{\theta}_0$? While there is no simple answer to this question, many sufficient conditions exist to ensure the consistency of $\hat{\boldsymbol{\theta}}_n$ as we will briefly discuss in a few of our example cases.

Example 1 (continued)

In the special case that the Y_is are independent and identically distributed (iid),

$$Q_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \ln f_1(y_i; \boldsymbol{\theta}),$$

where $f_1(\cdot; \boldsymbol{\theta})$ is the probability (density) function for a single observation. Then the law of large numbers implies that Q_n/n converges in probability to $Q_0(\boldsymbol{\theta}) = -\mathrm{E}_0[\ln f_1(Y_1; \boldsymbol{\theta})]$ at each θ, where E_0 denotes expectation taken with respect to $f_1(\cdot; \boldsymbol{\theta}_0)$. However, for each $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned} Q_0(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta}_0) &= -\mathrm{E}\left[\ln\left(\frac{f_1(Y_1; \boldsymbol{\theta})}{f_1(Y_1; \boldsymbol{\theta}_0)}\right); \boldsymbol{\theta}_0\right] \\ &\geq -\ln \mathrm{E}\left(\frac{f_1(Y_1; \boldsymbol{\theta})}{f_1(Y_1; \boldsymbol{\theta}_0)}; \boldsymbol{\theta}_0\right) = 0, \end{aligned}$$

known as Kullback's inequality. This inequality is strict if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and if $f(\cdot; \boldsymbol{\theta})$ is identifiable. Therefore, if $f(\cdot; \boldsymbol{\theta})$ is identifiable, then $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and hence $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$. For detailed proof of the consistency of MLE, see, for example, [25, 17], and [5, pp. 256–257].

Example 2 (continued)

In the iid. case, Q_n/n converges to $Q_0(\boldsymbol{\theta}) = -\mathrm{E}_0(\ln h_1(Y_1; \boldsymbol{\theta}))$. Whether $\boldsymbol{\theta}_0$ minimizes $Q_0(\boldsymbol{\theta})$ depends on the probability model specified and its interaction with the true probability model. In the situation in which $\boldsymbol{\theta} = \mathrm{E}(Y_1)$, Gourieroux et al. [10] provide a necessary and sufficient condition for $\boldsymbol{\theta}_0$ to minimize $Q_0(\boldsymbol{\theta})$, namely that $h_1(\cdot; \boldsymbol{\theta})$ take the form of

$$h_1(y_1; \boldsymbol{\theta}) = \exp[a(\boldsymbol{\theta}) y_1 + b(\boldsymbol{\theta}) + c(y_1)],$$

known as a linear exponential family. In this case

$$Q_0(\boldsymbol{\theta}) = -a(\boldsymbol{\theta})\boldsymbol{\theta}_0 - b(\boldsymbol{\theta}) - \mathrm{E}_0[c(Y_1)]$$

is minimized at $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_0$, a simple consequence of Kullback's inequality and the fact that $h_1(\cdot; \boldsymbol{\theta})$ is a probability (density) function.

Example 4 (continued)

Assuming that the x_is are iid, Q_n/n in (5) converges to

$$Q_0(\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{X}_1}\left\{\frac{[\mathbf{X}_1'(\boldsymbol{\theta}_0 - \boldsymbol{\theta})]^2}{V(\mathbf{X}_1'\boldsymbol{\theta})}\right\}$$

$$+ \mathrm{E}_{\mathbf{X}_1}\left[\frac{V(\mathbf{X}_1'\boldsymbol{\theta}_0)}{V(\mathbf{X}_1'\boldsymbol{\theta})}\right].$$

If $\mathrm{var}(Y_i|x_i) = V(\mathbf{x}_i'\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$ as in (4), then it is obvious that $\boldsymbol{\theta}_0$ minimizes $Q_0(\boldsymbol{\theta})$ and hence the consistency of $\hat{\boldsymbol{\theta}}_n$ which minimizes the weighted least squares in (4) is established. This is not true, in general, if $V(\cdot)$ does depend on $\boldsymbol{\theta}$. Consider a special case that x_i is dichotomous (1 or 0) with $\lambda = \Pr(X_i = 1)$, so that

$$\mathrm{E}(Y_i|x_i) = \theta_1 + \theta_2 x_i.$$

Furthermore, let $\mathrm{var}(Y_i|x_i) = \mathrm{E}^2(Y_i|x_i)$. It is easy to see that

$$Q_0(\theta) = \lambda\frac{(\theta_{01}+\theta_{02})^2 + (\theta_{01}-\theta_1+\theta_{02}-\theta_2)^2}{(\theta_1+\theta_2)^2} + (1-\lambda)\frac{\theta_{01}^2 + (\theta_{01}-\theta_1)^2}{\theta_1^2},$$

and that $Q_0(\theta)$ is minimized at

$$(\theta_1^*, \theta_2^*) = (2\theta_{01}, 2\theta_{02}).$$

Consistency: Type II

An alternative way of obtaining estimators is by solving an equation

$$\mathbf{g}_n(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{n}\mathbf{g}(y_i;\boldsymbol{\theta}) = 0, \qquad (6)$$

where $\mathbf{g}_n$, a $(p \times 1)$ vector-valued function, is called an *estimating function* for $\boldsymbol{\theta}$. The MLE may be viewed as a special case if $f(\mathbf{y};\boldsymbol{\theta})$ is first-order differentiable, in which case $\mathbf{g}_n(\mathbf{y};\boldsymbol{\theta}) = \partial \ln f(\mathbf{y};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, known as the score function for $\boldsymbol{\theta}$ (*see* **Likelihood**). The utility of the estimating function approach, however, is that it provides a sensible alternative to the MLE when the available substantive knowledge is insufficient to formulate a model $f(\mathbf{y};\boldsymbol{\theta})$ for the probability mechanism or, to a lesser extent, when the model $f(\mathbf{y};\boldsymbol{\theta})$ is too complicated to compute.

Example 5

The **quasi-likelihood** method proposed by Wedderburn [26] offers an excellent example. Here, $\boldsymbol{\theta}$ characterizes the relationship between y_i and covariates $\mathbf{x}_i$ through

$$\mu_i(\boldsymbol{\theta}) = \mathrm{E}(Y_i|\mathbf{x}_i) = h^{-1}(\mathbf{x}_i'\boldsymbol{\theta}),$$

where the one-to-one continuous function $h(\cdot)$ is known as the "link" function (*see* **Generalized Linear Model**). The contribution to g_n in (6) from the ith observation is

$$\mathbf{g}(y_i, \mathbf{x}_i;\boldsymbol{\theta}) = \left(\frac{\partial\mu_i}{\partial\theta}\right)' V_i^{-1}(\boldsymbol{\theta})(y_i - \mu_i(\boldsymbol{\theta})), \qquad (7)$$

which depends only on the first two conditional moments of y_i, namely $E(Y_i|\mathbf{x}_i) = \mu_i(\boldsymbol{\theta})$ and $\mathrm{var}(Y_i|\mathbf{x}_i) = V_i(\boldsymbol{\theta})$, where $V_i(\boldsymbol{\theta}) = V(\mu_i(\boldsymbol{\theta}))$.

Note that the expectation of $\mathbf{g}$ in (7) is zero so long as the true probability model $f(\cdot;\boldsymbol{\theta})$ has $\mu_i(\boldsymbol{\theta})$ as the mean for Y_i, and this is true even if $V_i(\boldsymbol{\theta})$ is misspecified. Note also that the score function, $\partial \ln f(\mathbf{y};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, has zero expectation. This zero expectation property for $\mathbf{g}_n$ is a natural one for the following reason: if the purpose is to find a solution $\hat{\boldsymbol{\theta}}_n$ such that $\mathbf{g}_n(\mathbf{y}, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}$, then it is intuitively desirable that the $\mathbf{g}$s be close to zero, at least in average, at the true $\boldsymbol{\theta}$ value. Similarly to the type I consistent estimators, in the simple case in which the Y_is are iid, the law of large numbers shows that $\overline{\mathbf{g}}_n = \mathbf{g}_n/n$ converges in probability to $\mathbf{g}_0(\boldsymbol{\theta}) = \mathrm{E}_0[\mathbf{g}_1(Y_1;\boldsymbol{\theta})]$ as $n \to \infty$. Following this argument, intuition suggests that $\hat{\boldsymbol{\theta}}_n$ may converge to some $\boldsymbol{\theta}^*$, a solution of $\mathbf{g}_0(\boldsymbol{\theta}) = \mathbf{0}$ and, consequently, $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$ if $\boldsymbol{\theta}^* \equiv \boldsymbol{\theta}_0$. There will generally exist a consistent estimator $\hat{\boldsymbol{\theta}}_n$ if the $\mathbf{g}$s are unbiased (i.e. if $g_0(\theta_0) = 0$) and the following conditions hold:

1. The true parameter $\boldsymbol{\theta}_0$ is an isolated root of $g_0(\boldsymbol{\theta})$ in the sense that there exists an open neighborhood $N \subset \Theta$ containing $\boldsymbol{\theta}_0$ such that $g_0(\boldsymbol{\theta}) \neq 0$ for all $\boldsymbol{\theta} \in N$ except $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.
2. $\mathbf{g}_n(\mathbf{y}, \boldsymbol{\theta})$ is continuous on Θ for all n.
3. $g_0(\boldsymbol{\theta})$ is either continuous or monotone (nonincreasing) on Θ.

In the case where θ is one-dimensional, condition 1 is generally satisfied if $n^{-1}\mathrm{E}_0(-\partial g_n/\partial\theta)$ and $n^{-1}\mathrm{var}_0(g_n)$ converge to positive constants at $\theta = \theta_0$. Condition 3 holds if g_n is continuous in θ and bounded or if g_n is a nonincreasing function of θ, but other sufficient conditions may be available in specific problems or cases in which $g_0(\theta)$ can be computed directly. Under the further condition that g_n is monotone in θ, all solution sequences $\hat{\theta}_n$ will converge to θ_0. Rigorous proofs of the consistency of $\hat{\theta}_n$ based on condition 1 are given in, for example,

[23] and [13]. An alternative approach when g_n is the score function is given by Cramer [4, pp. 501–503].

Example 5 (continued)

The "quasi-score function" defined in (7) generally satisfies the above three conditions if the unconditional (with respect to $\mathbf{X}_1$) **information matrix**

$$\mathrm{E}_{\mathbf{X}_1}\left[\frac{\partial \mu_1}{\partial \boldsymbol{\theta}} V_1^{-1}(\boldsymbol{\theta}) \frac{\partial \mu_1}{\partial \boldsymbol{\theta}}\right]$$

is positive definite for all $\boldsymbol{\theta}$. See detailed derivations for this important special case in [20] and [7].

Example 6 (Mantel–Haenszel Estimator)

Let $y_i = (a_i, b_i, c_i, d_i)$ be the entries from the ith **two-by-two table** stratified according to some arbitrary **confounding** variable, x_i. For example, this sort of data structure could arise in a retrospective **case–control study**, where i indexes a **stratifying** variable for which we want to control. Interest is on the odds ratio describing the association between some exposure of interest and risk for disease, adjusting for varying risk by stratum. In particular, we want to estimate the conditional (on x_i) odds ratio

$$\theta = \frac{\Pr(A_i = a_i|x_i)\Pr(D_i = d_i|x_i)}{\Pr(B_i = b_i|x_i)\Pr(C_i = c_i|x_i)},$$

which measures the strength of association between two dichotomous variables. The **Mantel–Haenszel** estimator [19] is seen as the solution of

$$g_n(\mathbf{y};\theta) = \sum_{i=1}^{n} \frac{1}{N_i}(a_i d_i - \theta b_i c_i) = 0,$$

where $N_i = a_i + b_i + c_i + d_i$ is the number of patients in the ith clinic. Assuming that the X_is are iid and the N_is are constant, one has immediately

$$\mathrm{E}_0[g_1(Y_1, X_1;\theta)] = \frac{1}{N_1}\mathrm{E}_{X_1}\{[\mathrm{E}_0(A_1 D_1|X_1) - \theta \mathrm{E}_0(B_1 C_1|X_1)]\},$$

which is linear in θ with negative slope $\mathrm{E}_{x_1,0}(B_1 C_1)/N_1$.

Some Final Remarks

First, in discussing the consistency of type II estimators, we dealt principally with one-dimensional θ. In models with multidimensional parameter, satisfying condition 1 requires more care. In particular, this condition is generally satisfied if $n^{-1}\mathbf{i}_n(\boldsymbol{\theta}) \to \mathbf{i}(\boldsymbol{\theta})$ as $n \to \infty$, where $\mathbf{i}(\boldsymbol{\theta})$ is a positive definite matrix and continuous in $\boldsymbol{\theta}$, and $\mathbf{i}_n(\boldsymbol{\theta})$ is the information in $\mathbf{g}_n$ given by

$$\mathbf{i}_n(\boldsymbol{\theta}_0) = \left[\mathrm{E}\left(-\frac{\partial \mathbf{g}_n}{\partial \boldsymbol{\theta}}\right)\mathrm{var}^{-1}(\mathbf{g}_n)\mathrm{E}\left(-\frac{\partial \mathbf{g}_n}{\partial \boldsymbol{\theta}}\right)'\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Secondly, we have dealt principally with the case in which the Y_is are iid and incorporated regression into our framework by considering the data (Y_i, X_i) to be iid random vectors, implicitly assuming that the distribution of the X_is exists, but is left unspecified. Extension to the independent but not identically distributed case, which may be a more natural approach to regression, is established by setting $\mathbf{x} = (x_1, \ldots, x_n)'$ and assuming that the normed conditional information

$$n^{-1}\mathbf{i}_n(\boldsymbol{\theta}_0) = n^{-1}\left[\mathrm{E}\left(-\frac{\partial \mathbf{g}_n}{\partial \boldsymbol{\theta}}|\mathbf{x}\right)\mathrm{var}^{-1}(\mathbf{g}_n|\mathbf{x}) \times \mathrm{E}\left(-\frac{\partial \mathbf{g}_n}{\partial \boldsymbol{\theta}}|\mathbf{x}\right)'\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

converges to a positive definite matrix.

Thirdly, it is not required that the Y_is be independent, and models with dependent Y_is are common in biomedical research. In many cases, we still have consistency of $\hat{\boldsymbol{\theta}}_n$. Again, the main requirement is that $n^{-1}\mathbf{i}_n(\boldsymbol{\theta}) \to \mathbf{i}(\boldsymbol{\theta})$ as $n \to \infty$, where $\mathbf{i}(\boldsymbol{\theta})$ is a positive definite matrix.

Fourthly, we have assumed throughout that the elements of $\boldsymbol{\theta}$ are the sole parameters necessary to specify the probability model for the data. Very often in practice, one needs additional parameters, $\boldsymbol{\phi}$ say, to completely specify the model, that is $f(\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta},\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a $q \times 1$ vector of additional parameters. For example, the variance of Y_i in (7) is likely to depend on parameters other than $\boldsymbol{\theta}$ which characterizes the first moment of Y_i. Or, in the retrospective study example (Example 6), the parameter of interest characterizes the within-stratum ratio of disease risk between the exposed and unexposed persons, so that we must control for confounding factors

relating to the stratum. The "stratum effect" is represented by the parameter $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$. Parameters of this kind are called **nuisance parameters**. Although in the particular case of Example 6, due to the elegance of the Mantel–Haenszel estimator, it is possible to formulate the problem independently of $\boldsymbol{\phi}$, in more general problems, the objective function Q_n to be minimized or the estimating function $\mathbf{g}_n$ to be solved, will depend on $\boldsymbol{\phi}$. We distinguish two cases.

In the first, assuming that one can find a well-behaved estimator of $\boldsymbol{\phi}$, $\hat{\boldsymbol{\phi}}$ say, in that

$$\sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) = O_p(1), \tag{8}$$

no additional complication is introduced, so far as the consistency of $\hat{\boldsymbol{\theta}}_n$ is concerned, by either minimizing $Q_n(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_n)$ or solving $\mathbf{g}_n(\mathbf{y}; \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_n) = 0$. In the second case, however, $\hat{\boldsymbol{\phi}}_n$ exists but does not meet the requirement given in (8). This often occurs when the number of nuisance parameters increases with n and the number of observations that is informative for each nuisance parameter is small even as n increases. In the retrospective study example, each y_i is a two-by-two table, and $q = n$. This is known as the Neyman & Scott [21] problem, and in such cases the maximum likelihood estimator might not be consistent (*see* **Estimating Functions**). Kalbfleisch & Sprott [15] and Basu [2] discuss a variety of likelihood methods to construct consistent estimators in the presence of many nuisance parameters. Andersen [1] provides proof of the consistency of conditional maximum likelihood estimators under the exponential family. Lindsay [18] extends Andersen's work to a broader class of probability models by further developing the concept of a conditional score function, which provides a zero-unbiased estimating function for θ even if ϕ is poorly estimated.

Fifthly, another assumption often made to show consistency is that the true parameter value θ_0 is an interior point of Θ. There are situations in which the parameter values of interest, θ_0, are on the boundary of the parameter space. A typical example is the application of **variance component** models to **pedigree** data in which the contribution from a particular genetic component is expressed in terms of its variance, which is necessarily nonnegative. Here a value of zero variance corresponds to the lack of contribution from this hypothesized component. In many cases, consistency turns out to hold even for θ_0 on the boundary [22].

Sixthly, one concern for the second (estimating function) approach is the issue of multiple roots when solving $\mathbf{g}_n = \mathbf{0}$. This is a potential problem in all but the simplest probability models (e.g. an exponential family with canonical link function). The question as to which root corresponds to a consistent estimator can be addressed in a variety of ways. For example, if a known consistent estimator is available, the root closest to that estimator will also be consistent [17, p. 421]. The flexibility of the estimating function approach often permits construction of a function with a unique root, which would provide such an estimator. This problem is further addressed by Hanfelt & Liang [11], who integrate estimating functions with respect to $\boldsymbol{\theta}$ to obtain an objective function, which is then minimized to select one $\hat{\boldsymbol{\theta}}_n$ among several solutions to $\mathbf{g}_n = \mathbf{0}$.

Finally, the task of parameter estimation allows scientists to learn about the magnitude of the quantities that reflect scientific objectives. The notion of consistency provides a means to evaluate the quality of proposed estimators. However, one should not lose sight of the fact that point estimation is only an intermediate part of statistical analysis. Equally important is to assess the precision of the estimator $\hat{\boldsymbol{\theta}}_n$, an issue not addressed here. The precision, along with the magnitude of $\hat{\boldsymbol{\theta}}_n$, allows one to construct **hypothesis tests** on $\boldsymbol{\theta}$ or to provide a range of plausible values for $\boldsymbol{\theta}$, given the data $\mathbf{Y} = \mathbf{y}$.

References

[1] Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators, *Journal of the Royal Statistical Society, Series B* **32**, 283–301.

[2] Basu, D. (1977). On the elimination of nuisance parameters, *Journal of the American Statistical Association* **72**, 355–366.

[3] Bates, C. & White, H. (1985). A unified theory of consistent estimation for parametric models, *Econometric Theory* **1**, 151–178.

[4] Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[5] Davidson, R. & MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, Oxford.

[6] Durbin, J. (1960). Estimation of parameters in time-series regression models, *Biometrika* **47**, 139–153.

[7] Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *Annals of Statistics* **13**, 342–368.

[8] Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves, *Messenger of Mathematics* **41**, 155–160.

[9] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation, *Annals of Mathematical Statistics* **31**, 1208–1212.

[10] Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudo maximum likelihood methods: theory, *Econometrica* **52**, 681–700.

[11] Hanfelt, J.J. & Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions, *Biometrika* **82**, 461–477.

[12] Huber, P.J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**, 73–101.

[13] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* Vol. 1. University of California Press, Berkeley, pp. 221–233.

[14] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

[15] Kalbfleisch, J.D. & Sprott, D.A. (1970). Application of likelihood methods to models involving a large number of nuisance parameters (with discussion), *Journal of the Royal Statistical Society, Series B* **32**, 175–208.

[16] Legendre, A.M. (1805). *Nouvelles Méthodes pour la Detérmination des Orbites des Cometès*. Courcier, Paris.

[17] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[18] Lindsay, B. (1982). Conditional score functions: some optimality results, *Biometrika* **69**, 503–512.

[19] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[20] McCullagh, P. (1983). Quasi-likelihood functions, *Annals of Statistics* **11**, 59–67.

[21] Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations, *Econometrica* **16**, 1–32.

[22] Self, S.G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association* **82**, 605–610.

[23] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[24] Tukey, J.W. (1960). A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics*, I. Olkin, ed. Stanford University Press, Stanford.

[25] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *Annals of Mathematical Statistics* **20**, 595–601.

[26] Wedderburn, R.W.M. (1974). Quasi likelihood functions, generalized linear models and the Gauss–Newton method, *Biometrika* **61**, 439–447.

(*See also* **Inference, Foundations of**)

KUNG-YEE LIANG & P.J. RATHOUZ

CONSORT

A report of a randomized controlled trial (RCT) should convey to the reader, in a transparent manner, why the study was undertaken, and how it was conducted and analyzed. To assess the strengths and limitations of an RCT, the reader needs and deserves to know the quality of its methodology. Despite several decades of educational efforts, RCTs still are not being reported adequately [2, 5, 10].

The Consolidated Standards of Reporting Trials (CONSORT) statement, published in the *Journal of the American Medical Association* in 1996 [1], was developed to try to help rectify this problem. The CONSORT statement was developed by an international group of clinical trialists, statisticians, epidemiologists and biomedical editors. The CONSORT statement is one result of previous efforts made by two independent groups, the Standards of Reporting Trials (SORT) group [9] and the Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature [11]. The CONSORT statement consists of two components, a 21-item checklist (Table 1) and a flow diagram (Figure 1). The checklist has six major headings that pertain to the contents of the report of a trial, namely Title, Abstract, Introduction, Methods, Results and Discussion. Within these major headings there are subheadings that pertain to specific items that should be included in any clinical trial manuscript.

These items constitute the key pieces of information necessary for authors to address when reporting

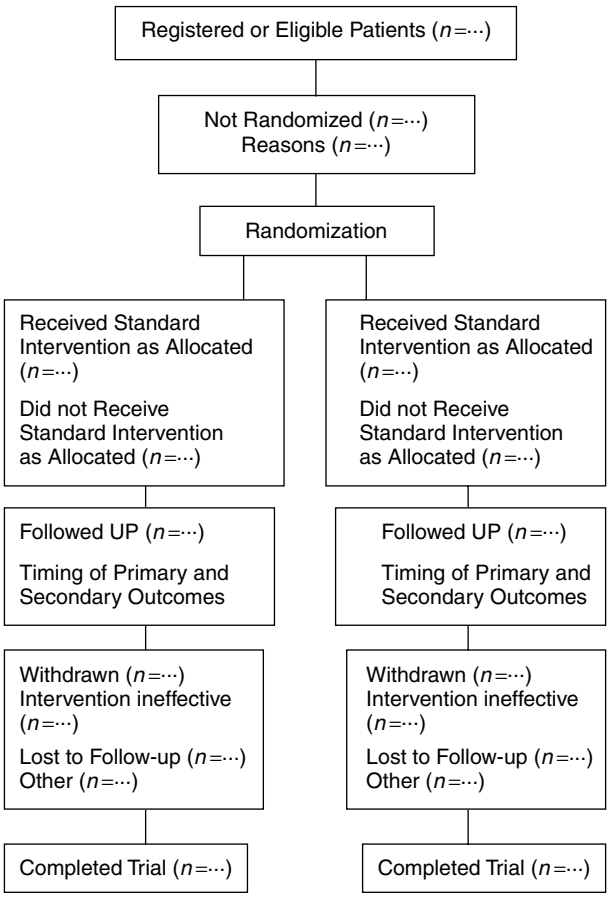

Figure 1 CONSORT flowchart. Reproduced with permission from the *Journal of the American Medical Association*, 1996, Volume 276, 637–665. Copyrighted (1996), American Medical Association

Table 1 CONSORT checklist

Heading	Subheading	Descriptor	Was it reported?	Page no.?
Title		Identify the study as a randomized trial.		
Abstract		Use a structured format.		
Introduction		Identify the study as a randomized trial. Use a structured format. State prospectively defined hypothesis, clinical objectives, and planned subgroup or covariate analyses		
Methods	Protocol	Describe 1. Planned study population, together with inclusion/exclusion criteria. 2. Planned interventions and their timing. 3. Primary and secondary outcome measure(s) and the minimum important difference(s), and indicate how the target sample size was projected. 4. Rationale and methods for statistical analyses, detailing main comparative analyses and whether they were completed on an intention-to-treat basis. 5. Prospectively defined stopping rules (if warranted)		
	Assignment	Describe 1. Unit of randomization (e.g. individual, cluster, geographic). 2. Method used to generate the allocation schedule. 3. Method of allocation concealment and timing of assignment. 4. Method to separate the generator from the executor of assignment.		
	Masking (Blinding)	Describe mechanism (e.g. capsules, tablets); similarity of treatment characteristics (e.g. appearance, taste); allocation schedule control (location of code during trial and when broken); and evidence for successful blinding among participants, person doing intervention, outcome assessors, and data analysts.		
Results	Participant Flow and Follow-up	Provide a trial profile (Figure 1) summarizing participant flow, numbers and timing of randomization assignment, interventions, and measurements for each randomized group.		
	Analysis	State estimated effect of intervention on primary and secondary outcome measures, including a point estimate and measure of precision (confidence interval). State results in absolute numbers when feasible (e.g. 10/20, not 50%). Present summary data and appropriate descriptive and inferential statistics in sufficient detail to permit alternative analyses and replication. Describe prognostic variables by treatment group and any attempt to adjust for them. Describe protocol deviations from the study as planned, together with the reasons.		
Comment		State specific interpretation of study findings, including sources of bias and imprecision (internal validity) and discussion of external validity, including appropriate quantitative measures when possible. State general interpretation of the data in light of the totality of the available evidence.		

the results of an RCT. Their inclusion is based on evidence, whenever possible. For example, authors are asked to report on the methods they used to achieve allocation concealment, possible in every randomized trial. There is growing evidence that inadequately concealed trials, compared with adequately concealed ones, exaggerate the estimates of intervention benefit by 30%–40%, on average [7, 8]. Additional benefits of the checklist (and flow diagram) include facilitating editors, peer reviewers and journal readers to evaluate the internal and external validity of a clinical trial report.

The flow diagram pertains to the process of winnowing down the number of participants from those eligible or screened for a trial to those who ultimately completed the trial and were included in the analysis. The flow diagram pertains particularly to a two-group, parallel design, as stated in the CONSORT statement. Other checklists and flow diagrams have been developed for reporting cluster randomized trials [4] and other designs (see `http://www.consort-statement.org`). The flow diagram, in particular, requests relevant information regarding participants in each of the intervention and control groups who did not receive the regimen for the group to which they were randomized, those who during the course of the trial were discontinued, withdrew, became lost to follow-up, and those who have incomplete information for some other reason.

There is emerging evidence to suggest that the quality of reporting of RCTs, based on the use of the CONSORT statement, compared with not using it, is higher on several dimensions, such as reduced reporting of unclear allocation concealment [6]. Similarly, use of the flow diagram was associated with better overall reporting of RCTs [3].

The CONSORT statement (checklist and flow diagram) is available on the CONSORT website (`www.consort-statement.org`). This site includes information on the growing number of health care journals and biomedical editorial groups, such as the International Council of Medical Journal Editors (ICMJE), who support the use of the CONSORT statement for reporting RCTs.

At this writing the CONSORT statement is undergoing revision. Present plans call for the revised Statement to appear in Spring 2001 along with an extensive explanatory and elaboration document to overcome some of the shortcomings of the original statement, both of which will be available on the above website.

References

[1] Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. & Stroup, D.F. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT Statement, *Journal of the American Medical Association* **276**, 637–639.

[2] Dickinson, K., Bunn, F., Wentz, R., Edwards, P. & Roberts, I. (2000). Size and quality of randomized controlled trials in head injury: review of published studies, *British Medical Journal* **320**, 1308–1311.

[3] Egger, M., Jüni, P., Bartlett, C. for the CONSORT Group (2001). The value of CONSORT flow charts in reports of randomized controlled trials: bibliographic study, *Journal of the American Medical Association*, in press.

[4] Elbourne, D.R. & Campbell, M.K. (2001). Extending the CONSORT statement to cluster randomized trials: for discussion, *Statistics in Medicine* **20**, 489–496.

[5] Hotopf, M., Lewis, G. & Normand, C. (1997). Putting trials on trial – the costs and consequences of small trials in depression: a systematic review of methodology, *Journal of Epidemiology and Community Health* **51**, 354–358.

[6] Moher, D., Jones, A., Lepage, L. for the CONSORT Group (2001). Does the CONSORT statement improve the quality of reports of randomized trials: a comparative before and after evaluation?, *Journal of the American Medical Association*, in press.

[7] Moher, D., Pham, B., Jones, A., Cook, D.J., Jadad, A.R., Moher, M. & Tugwell, P. (1998). Does the quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analyses?, *Lancet* **352**, 609–613.

[8] Schulz, K.F., Chalmers, I., Hayes, R.J. & Altman, D.G. (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association* **273**, 408–412.

[9] The Standards of Reporting Trials Group (1994). A proposal for structured reporting of randomized controlled trials, *Journal of the American Medical Association* **272**, 1926–1931. Correction: *Journal of the American Medical Association* **273**, 776.

[10] Thornley, B. & Adams, C.E. (1998). Content and quality of 2000 controlled trials in schizophrenia over 50 years, *British Medical Journal* **317**, 1181–1184.

[11] Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature (1994). Call for comments on a proposal to improve reporting

of clinical trials in the biomedical literature: a position paper, *Annals of Internal Medicine* **121**, 894–895.

(*See also* **QUORUM**)

DAVID MOHER

Contagious Distributions

The term "contagion" entered the statistical literature in a paper by Pólya [14] that reexpounded certain ideas previously put forward by Eggenberger & Pólya [3] (*see* **Polya's Urn Model**). During the spread of an infection each new case was considered to release more germs into the atmosphere; the probability of catching the disease was therefore thought to depend on the existing number of cases. Eggenberger & Pólya modeled this via an urn initially containing R red balls (cases) and S black balls (others), where $R + S = N$; balls are drawn one at a time at random from the urn and each drawn ball is immediately replaced together with Δ new balls of the same color. They showed that the probability that, after n draws, r red balls (representing cases of infection) and $n - r = s$ black balls (individuals without the infection) have been drawn is

$$\mathrm{P}_r = \frac{n!}{r!s!}\frac{\rho(\rho+\delta)\cdots[\rho+(r-1)\delta]}{1(1+\delta)\cdots[1+(r-1)\delta]} \times \frac{\sigma(\sigma+\delta)\cdots[\sigma+(s-1)\delta]}{(1+r\delta)[1+(r+1)\delta]\cdots[1+(n-1)\delta}, \quad (1)$$

where $\rho = R/N, \sigma = S/N$, and $\delta = \Delta/N$. In [3] they set $n\rho = h, n\delta = d$, and let $n \to \infty$, with h and d remaining finite, giving the limiting form

$$\mathrm{P}_r = \frac{h(h+d)(h+2d)\cdots[h+(r-1)d]}{r!(1+d)^{(h/d)+r}}, \quad (2)$$

i.e. a **negative binomial distribution**. They fitted this and a **Poisson distribution** to Swiss monthly mortality figures for smallpox over the period 1877 to 1900, as shown in Table 1. The negative binomial distribution fits the data very much better than a Poisson distribution (which is the model for deaths occurring completely at random), supporting their hypothesis of contagion.

The same type of contagion concept, concerning the changing probability of an event during a sequence of events, had arisen previously in a paper by Greenwood & Yule [6]. These authors put forward the "burnt-fingers" hypothesis that after an individual has suffered one accident he/she will have a lower probability of suffering future accidents of the same type (*see* **Accident Proneness**). They obtained a good fit when they applied their model to data on accidents in the manufacture of high-explosive shells.

Neyman [12] had a different concept of contagion. He sought an explanation for the failure of the Poisson distribution to model the distribution of insect larvae in an agricultural field or the distribution of cells in a haemocytometer and observed that such data generally have a long tail and an unexpectedly high variance. He said that such data have a "contagious distribution". The NTA (Neyman's type A) distribution that he fitted can be regarded as arising from a Poisson distribution of clumps (with mean λ), with the number of cells for each clump having independent and identical Poisson distributions (with mean ϕ). It gives a better fit than the Poisson distribution, as shown in Table 2.

Table 1

Number of deaths	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	≥15	Total
Observed number of months	100	39	28	26	13	6	11	5	5	6	1	6	2	2	3	35	288
Negative binomial fit	100.4	36.3	23.5	17.5	13.8	11.3	9.5	8.1	7.0	6.1	5.3	4.7	4.2	3.7	3.3	33.3	288.0
Poisson fit	1.2	6.5	17.8	32.6	44.9	49.4	45.2	35.5	24.5	15.0	8.2	4.1	1.9	0.8	0.3	0.1	288.0

Table 2

Number of cells/square	0	1	2	3	4	≥5	Total
Observed number of squares	213	128	37	18	3	1	400
NTA fit	214.8	121.3	45.7	13.7	3.6	0.9	400.0
Poisson fit	202.2	138.0	47.1	10.7	1.8	0.2	400.0

The expressions for the NTA probabilities are very complicated. The distribution is easier to understand via its probability generating function (pgf) (*see* **Generating Functions**)

$$G_{\mathrm{NTA}}(z) = G[g(z)], \tag{3}$$

where $G(z) = \exp[\lambda(z-1)]$, which is the Poisson pgf for the number of clumps per sample, and $g(z) = \exp[\phi(z-1)]$, the Poisson pgf for the number of larvae per clump.

Many other contagious distributions have been constructed in a similar manner. For example, the Lagrangian Poisson distribution [1] has the pgf

$$G_{\mathrm{LP}}(z) = \exp\{\lambda[t(z)-1]\}, \tag{4}$$

where $t(z)$ is the pgf that is the solution of the equation $t(z) = z\exp\{\psi[t(z)-1]\}$; Consul and his coworkers have fitted this successfully to many types of data including lesions in rabbit lymphoblasts and chromosome aberrations in cultures of human leucocytes. For further examples of such distributions, see Douglas [2] and Johnson et al. [7].

The heterogeneity model considered by Greenwood & Yule [6] in the **accident proneness** context has also been called "contagious". Instead of all individuals in a population having identical Poisson distributions of accidents, the Poisson parameter θ is assumed to vary among individuals. This is a mixture model and θ is the mixing variable. In terms of pgfs we have

$$G_{\mathrm{mixed}}(z) = \int g(z|\theta) f(\theta)\,\mathrm{d}\theta, \tag{5}$$

where $g(z|\theta)$ is a Poisson distribution with parameter θ, $f(\theta)$ is the pdf of the distribution of θ, and integration is over all values of θ. Greenwood & Yule showed that if θ has a **gamma distribution**, then the resultant mixed Poisson distribution is a negative binomial distribution.

The discrete analog of (5) is

$$G_{\mathrm{dmixed}}(z) = \sum_{x=0}^{\infty} g_x(z)\omega_x, \tag{6}$$

were x takes integer values, $g_x(z)$ is a pgf, $\omega_x \geq 0$, and $\sum_{x=0}^{\infty}\omega_x = 1$.

The pgf (3) has the form

$$\begin{aligned} G_{\mathrm{dmixed}}(z) &= \omega_0 + \omega_1 g(z) + \omega_2[g(z)]^2 + \cdots \\ &= G[g(z)], \end{aligned} \tag{7}$$

where $G(z) = \omega_0 + \omega_1 z + \omega_2 z^2 + \cdots$ and $g_x(z) = [g(z)]^x$. It can therefore be regarded as arising from a special kind of mixture of distributions, since $[g(z)]^x$ is itself a pgf when x is an integer (Neyman used this method of expansion when obtaining the NTA probabilities).

Feller [4] referred to the Pólya [14] model as "true contagion" and to the mixture models as "apparent contagion" (in his book, [5, pp. 121–123], he also used the term "spurious contagion"). The "true" and "apparent" terminology is still used. His opinion was that statisticians "speak of contagion (or contagious probability distributions) in a vague and misleading manner".

A major reason for this vagueness is the variety of models that may give rise to a single distribution. Lüders [10] showed that a negative binomial distribution results if $G(z)$ is the pgf of a Poisson distribution and $g(z)$ is the pgf of a logarithmic distribution in (7). The negative binomial distribution therefore arises from all three models discussed above (it is also the outcome of several models unrelated to contagion).

A second instance of a distribution arising from several models is the Hermite distribution. This was first derived by McKendrick [11] via the sum of two correlated Poisson random variables. He fitted it to data on the number of bacteria ingested by phagocytes. Table 3 shows the fit. Kemp & Kemp [9] stated the pgf in the form

$$G_{\mathrm{H}}(z) = \exp\{\lambda[2pq(z-1) + p^2(z^2-1)]\}, \tag{8}$$

and hence showed that the data could also have arisen from a Poisson distribution of phagocytes with each phagocyte containing 0, 1, or 2 bacteria according to a binomial distribution with $n = 2$ (a clustering

Table 3

Number of bacteria/phagocyte	0	1	2	3	≥4	Total
Observed number of phagocytes	269	4	26	0	1	300
Hermite fit (by mle)	269.6	3.7	25.2	0.3	1.2	300.0

model). Alternatively, (8) can be restated as

$$G_{\mathrm{H}}(z) = \mathrm{e}^{-\lambda} + \mathrm{e}^{-\lambda}\lambda(q + pz)^2 + \mathrm{e}^{-\lambda}\lambda^2(q + pz)^4/2! + \mathrm{e}^{-\lambda}\lambda^3(q + pz)^6/3! + \cdots, \quad (9)$$

i.e. as a *mixture* of binomial distributions with exponent parameters taking the values $0, 2, 4, 6, \ldots$.

Another example is the Gegenbauer distribution. Plunkett & Jain [13] obtained this as a gamma mixture of Hermite distributions, giving an "apparent contagion" model. They fitted it to the haemocytometer data quoted above (Table 2) and obtained expected frequencies very similar to those obtained using the NTA distribution. Johnson et al. [7] give a number of other models for the Gegenbauer distribution, including McKendrick's [11] nonhomogeneous birth and death process (*see* **Stochastic Processes**) and Kemp's [8] field observation model involving a Rao damage process.

The multitude of models that may exist for a particular distribution means that it is very much more informative to use the term "contagion model", specifying the model and its form of contagion precisely, than it is to call a particular distribution a "contagious distribution".

References

[1] Consul, P.C. (1989). *Generalized Poisson Distributions*. Marcel Dekker, New York.

[2] Douglas, J.B. (1980). *Analysis with Standard Contagious Distributions*. International Co-operative Publishing House, Burtonsville.

[3] Eggenberger, F. & Pólya, G. (1923). Über die Statistik verketteter Vorgänge, *Zeitschrift für angewandte Mathematik und Mechanik* **3**, 279–289.

[4] Feller, W. (1943). On a general class of "contagious" distributions, *Annals of Mathematical Statistics* **14**, 389–400.

[5] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd Ed. Wiley, New York.

[6] Greenwood, M. & Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society, Series A* **83**, 255–279.

[7] Johnson, N.L., Kotz, S. & Kemp, A.W. (1992). *Univariate Discrete Distributions*, 2nd Ed. Wiley, New York.

[8] Kemp, A.W. (1992). On counts of individuals able to signal the presence of an observer, *Biometrical Journal* **34**, 595–604.

[9] Kemp, C.D. & Kemp, A.W. (1965). Some properties of the "Hermite" distribution, *Biometrika* **52**, 381–394.

[10] Lüders, R. (1934). Die Statistik der seltenen Ereignisse, *Biometrika* **26**, 108–128.

[11] McKendrick, A.G. (1926). Applications of mathematics to medical problems, *Proceedings of the Edinburgh Mathematical Society* **44**, 98–130.

[12] Neyman, J. (1939). On a new class of "contagious" distributions applicable in entomology and bacteriology, *Annals of Mathematical Statistics* **10**, 35–57.

[13] Plunkett, I.G. & Jain, G.C. (1975). Three generalized negative binomial distributions, *Biometrische Zeitschrift* **17**, 276–302.

[14] Pólya, G. (1930). Sur quelques points de la théorie des probabilités, *Annales de l'Institut H. Poincaré* **1**, 117–161.

ADRIENNE W. KEMP & C.D. KEMP

Contingency Table

We start by defining a two-way contingency table. Suppose that each of a sample of n individuals is classified according to each of two separate criteria, and each of these criteria can take only a finite number of distinct values. Let us take as an example a study to assess factors associated with women's attitudes toward mammography [6, p. 220]. Each of 309 women has been classified according to:

1. Her "Mammography Experience", which has as its possible values "Never", "Over one year ago", and "Within the past year".
2. Her response to the question "How likely is it that a mammogram could find a new case of breast cancer?", which has as its possible values "Not likely", "Somewhat likely", and "Very likely".

Let the Mammography Experience correspond to the *Rows* of the table and the response to the question about the detection of breast cancer to the *Columns*. Although it happens in this particular example that each of the Rows and Columns has a natural ordering, this is not necessary to the definition of a contingency table. Then the resulting 3×3 *table* of frequencies or counts is, say, (n_{ij}), where:

1. n_{ij} is the number of individuals in Row i and Column j for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.
2. I and J are, respectively, the total numbers of Rows and Columns of the table, here 3, 3. Let n be the total sample size, so that n is the sum of entries of the table; we can then write $n = \sum\sum n_{ij}$.

Note that in our construction of this *cross tabulation* or contingency table, we assume that each individual of the sample of size n is classified into exactly one of the IJ categories of the table, these categories being mutually exclusive and exhaustive. Written less formally, this means that the IJ categories do not overlap in any way, and together they cover all possibilities.

Of course, exactly how the sample is collected for a particular study is of great importance in the subsequent statistical analysis. In classical contingency table analysis, it is assumed that the n individuals form a *random sample* from a population: this means that we assume that each individual is classified independently of the others, and each has the same probability, say p_{ij}, of being classified into row i and column j. Thus $\sum\sum p_{ij} = 1$. Our assumption that we have a random sample has the consequence that the probability distribution of the variable (n_{ij}) is **multinomial**; that is, the frequency function $p[(n_{ij})|p]$ is

$$n! \prod_i \prod_j \left(\frac{p_{ij}^{n_{ij}}}{n_{ij}!} \right),$$

provided that $n_{ij} \geq 0$ and $\sum\sum n_{ij} = n$.

In this case we say that (n_{ij}) is multinomial, with parameters n and (p_{ij}).

To continue with the above example of the attitudes toward mammography data, the results of our survey are shown in Table 1.

These data show a dramatic and self-evident association between rows and columns: in other words, the women's responses to the question about the detection of breast cancer depend strongly on their mammography experience. This can be seen even more clearly by presenting the data from this sample in terms of the *row percentages*. Rounded to the nearest whole number for clarity, these are shown in Table 2.

Presenting the data in this way shows more clearly how the response to the question depends on the mammography experience. Another illuminating way to show this effect would be by using three *bar charts*,

Table 1 Mammography: the counts

	Detection of breast cancer		
Mammography experience	Not likely	Somewhat likely	Very likely
Never	13	77	144
Over one year ago	4	16	54
Within the past year	1	12	91

Table 2 Mammography: the row percentages

	Detection of breast cancer			
Mammography experience	Not likely	Somewhat likely	Very likely	Total
Never	6	33	61	100
Over one year ago	5	22	73	100
Within the past year	1	11	87	100

overlaid on the same graph, with each corresponding to a different row of the table (*see* **Graphical Displays**).

There are several possible analyses we can carry out on this particular table, such as a simple summary in terms of percentages, a graphical summary, or a statistical test of independence which we describe below. None of these analyses can establish whether the association between rows and columns is a *causal* one. This is an inherent limitation of the purely statistical approach. In the current example, the Column variable might be thought of as a **response** to the Row variable, which is an **explanatory** variable or **covariate**. But the simple two-way contingency table cannot directly tell us the mechanism by which the explanatory variable affects the response variable. In any case, in the current example, there could be a mixture of several complex mechanisms which act simultaneously.

The Hypothesis of Independence, H_0, and the Chi-square Test

With the notation introduced above, that

$$p_{ij} = \Pr(\text{Row} = i, \text{Column} = j) \tag{1}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$, we see that the hypothesis of independence of rows and columns may be written as

$$H_0 : p_{ij} = p_{i+}p_{+j} \tag{2}$$

for all i, j, where $p_{i+} = \Sigma_j p_{ij}$ is the probability that Row $= i$, and $p_{+j} = \Sigma_i p_{ij}$ is the probability that Column $= j$.

Observe that an equivalent way of writing H_0 is

$$H_0 : \frac{p_{ij}}{p_{i+}} = p_{+j} \tag{3}$$

for all i, j. Now p_{ij}/p_{i+} is the probability that the Column is j, conditional on the Row being i: we write this as $P(\text{Column} = j|\text{Row} = i)$. Thus the hypothesis of independence of rows and columns is equivalent to the statement that the distribution of the variable Column, conditional on the value of the variable Row being i, is the same for all values of i: in this case we say that these conditional distributions are *homogeneous*.

To test the **null hypothesis** of independence H_0, we compute Pearson's chi-square statistic X^2, which is defined as

$$X^2 = \sum\sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{4}$$

where e_{ij} are conventionally called the "expected values", and are defined by

$$e_{ij} = \frac{(n_{i+}n_{+j})}{n} \quad \text{for all } i, j, \tag{5}$$

and n_{i+} and n_{+j} are the row and column totals, respectively, defined by $n_{i+} = \Sigma_j n_{ij}$ for all i and $n_{+j} = \Sigma_i n_{ij}$ for all j. (A pedantic but more accurate description of (e_{ij}) is "the **maximum likelihood** estimates of the expected values of (n_{ij}) under H_0".) We know that, under the null hypothesis H_0, the distribution of X^2 is, for large n, approximately χ^2_f, where f, the *degrees of freedom*, are $(I - 1)(J - 1)$. Hence for a significance test of H_0 with approximate significance level α, we reject H_0 if

$$X^2 \geq \chi^2_f(\alpha), \tag{6}$$

the right-hand side being the upper α point of the χ^2_f distribution, which we can find from statistical tables: typically we would take α as 0.01 or 0.05. Thus we have approximately arranged that, for large n, our test gives

$$\Pr(\text{reject } H_0|H_0 \text{ true}) \leq \alpha. \tag{7}$$

This is what we mean by saying that the test is of approximate *size* α.

Chi-square tests are described in depth in another article. The value of the X^2 statistic for the numerical example given above is 24.15, which lies well above 18.47, the 0.001 point of the χ^2_4 distribution. Thus for these data, the null hypothesis of independence of the women's response to the question about the detection of breast cancer and their "Mammography Experiences" is rejected, as our preliminary look at the table suggested would be the case.

From the point of view of scientific enquiry, a statistical significance test is a crude and blunt instrument. If we found that the value of X^2 was significant, we would usually want to investigate in some detail *why* the null hypothesis of independence fails to fit. It is good statistical practice to compare the *adjusted* **residuals**

$$\frac{(n_{ij} - e_{ij})}{[e_{ij}(1 - \hat{p}_{i+})(1 - \hat{p}_{+j})]^{1/2}}, \tag{8}$$

where $\hat{p}_{i+} = (n_{i+}/n)$, $\hat{p}_{+j} = (n_{+j}/n)$, with the standard normal distribution. Thus any residual greater than about 1.5 in absolute value is "interesting", and the corresponding cell (i, j) of the table is special in some way and deserves investigation.

Under the null hypothesis H_0, the **sufficient statistics** for the unknown parameters $((p_{i+}), (p_{+j}))$ are $((n_{i+}), (n_{+j}))$. One consequence of this is that the observed values of these statistics must agree exactly with the corresponding expected values: thus

$$n_{i+} = e_{i+} \text{ for each } i, \text{ and } n_{+j} = e_{+j} \text{ for each } j. \tag{9}$$

This is a special case of a general result for **exponential families**, and we shall see examples of the same result when we consider multiway tables in the final section.

A test statistic which is equivalent to X^2 for large sample size is the *deviance*, conventionally denoted by G^2. It is derived from the ratio of maximized likelihoods (*see* **Likelihood Ratio Tests**). For the example of testing H_0, G^2 is defined by

$$G^2 = 2 \sum \sum n_{ij} \log \left(\frac{n_{ij}}{e_{ij}} \right) \tag{10}$$

and for the numerical example given above G^2 has the value 26.80; recall that the value of X^2 was 24.15.

Both X^2 and G^2 are appropriate test statistics for sampling setups other than the straightforward single multinomial one introduced above. For example, suppose the contingency table (n_{ij}) were collected with the row totals

(n_{i+}) fixed, and

$(n_{ij}|n_{i+})$ independently and multinomially distributed, with

(n_{ij}) multinomial parameters n_{i+}, (θ_{ij}),

where $\sum_j \theta_{ij} = 1$ for each i.

Then the statistics X^2 and G^2 are the appropriate test statistics for testing H: $\theta_{ij} = \phi_j$ for all i, j, for some (unknown) ϕ_j such that $\sum \phi_j = 1$; in other words, the null hypothesis of homogeneity of row distributions.

Yet another way of writing H_0 is

$$H_0 : \log(p_{ij}) = \alpha_i + \beta_j \quad \text{for all } i, j. \tag{11}$$

for some α, β such that $\Sigma p_{ij} = 1$. One reason for writing H_0 in this form is that it enables us to see that it is a *loglinear* hypothesis; that is, the log of the probabilities can be written as a function which is linear in the unknown parameters, which in this case are (α_i), (β_j) (*see* **Loglinear Model**).

Computational Aspects

Contingency table analysis, both by chi-square tests and by small-sample "exact" methods, is easily achieved in many statistical software packages. Usually we can ask for direct computation of the appropriate chi-square statistic, with associated residuals and significance tests. More sophisticated software will allow us to test complex hypotheses of independence by fitting models using one or both of **iterative proportional fitting** (IPF) or **generalized linear modeling** (GLM). IPF may be computationally more efficient than the GLM approach, since in using GLM for contingency table analysis we are ignoring the particular structure of the parameters. For example, in using GLM to test for independence in a two-way contingency table, we use an iterative technique (effectively the *Newton–Raphson*) to solve the maximum likelihood equations when in fact five minutes with a pencil and paper would show us that iteration is unnecessary: there is a closed form solution. But an advantage of the GLM approach is that it is more suited to the problem of choosing the simplest possible model consistent with a given data set. Most statisticians are familiar with the use of GLM in any case, and so this advantage outweighs the possible gain in computational efficiency of IPF. The GLM approach may be rather more forbidding for the less mathematical user, but has definite advantages when we come to consider cross tabulations which are more complex than just the two-way ones. We discuss *multi-way* contingency tables in the last section of this article. Although there are several possible probability distributions available for GLMs, there is no multinomial distribution. This is not a problem, provided that we are fitting a *loglinear* model, such as H_0 as written in its final form above. Thus for testing independence in a two-way contingency table we can use the **Poisson distribution** as a "surrogate" for the multinomial, so that we can compute the appropriate deviance for testing independence for the multinomial model by pretending that (n_{ij}) are

observations on independent Poisson variables. This is a special case of a general result, relating Poisson and multinomial loglinear models.

Quasi-Independence

For some two-way contingency tables with a special structure, it may be obvious that the null hypothesis of independence H_0 cannot possibly be expected to fit, because it is far too "severe". However, a weaker hypothesis which is chosen to represent a version of independence, or **quasi-independence**, may be helpful in the interpretation of the data. For example, Altham [3] discusses the data on initial and final ratings of 121 stroke patients in Table 3, taken from [5].

The rows of the 5×5 table correspond to the patient's state on admission to hospital, and the columns to his state when discharged. These states have possible values A, B, C, D, and E, ranging from A as the least severe to E as the most severe. The resulting contingency table is *triangular*, because the patient's state on being discharged is never worse than his initial state. Because of this constraint, the hypothesis of independence of rows and columns cannot possibly hold. But if we assume that the frequencies (n_{ij}) for $1 \leq j \leq i \leq 5$ come from a multinomial distribution with corresponding parameters (p_{ij}), then it may still be helpful to fit a hypothesis of *quasi-independence* H_{qi}, say:

$$H_{qi} : p_{ij} = \alpha_i \beta_j, \quad \text{for all } j \leq i, \tag{12}$$

for some (α_i), (β_j). This is another loglinear hypothesis, and so may be tested by treating the (n_{ij}) as independent Poisson variables. For the data given here, the deviance for testing H_{qi} has value 9.60, $df = 6$, indicating a reasonable fit. Similar models may be appropriate for the type of "triangular" data arising from **capture–recapture** studies in ecology, where, for example, a bird first ringed in 1975 may be observed in any one of the 10 successive years, but of course a bird first ringed in 1977 could not have appeared in the 1975 count.

Quasi-independence models for square contingency tables (omitting the diagonal entries) are known as "mover–stayer" models in the sociological context. This hypothesis of quasi-independence is one of several models especially suitable for a **square contingency table**; others in this class are models of symmetry, marginal homogeneity, and **quasi-symmetry**.

Table 3 Stroke patients. Reproduced from [5] by permission of the Biometrics Society

	A	*B*	*C*	*D*	*E*	
A	5	–	–	–	–	5
B	4	5	–	–	–	9
C	6	4	4	–	–	14
D	9	10	4	1	–	24
E	11	23	12	15	8	69
	35	42	20	16	8	121

Exact Tests of Independence

For tables with small frequencies, the large-sample approximation of the distribution of X^2, or equivalently of the distribution of G^2, to the χ^2 distribution may be doubtful. Most software will give appropriate "warning messages" if the "expected counts" e_{ij} are too small, less than five being a conservative interpretation of "too small" here. In this case the statistician will need to use an exact method (*see* **Exact Inference for Categorical Data**). For a **2 × 2 table**, **Fisher's exact test**, which is based on the **hypergeometric distribution**, is appropriate, and it has been generalized, say, to $I \times J$ tables by using the multivariate hypergeometric distribution: this is the result of conditioning the multinomial distribution on both the row and column frequency totals, under the null hypothesis of independence of rows and columns.

For example, consider the 2×2 contingency table in Table 4.

In carrying out a χ^2 test on this table, we are essentially asking whether the proportion 8/16 is different from the proportion 20/23. The X^2 statistic is 4.67, with corresponding P value $= 0.0307$. But the software warns that the expected counts may be too low for the large-sample approximation to be valid, and it is easy to see that e_{11}, which is $16 \times 11/39$, is 4.5. For the sake of comparison, we do a two-sided Fisher exact test, and find that the corresponding P value is 0.0272. So either

Table 4 Example for Fisher's exact test

8	8
3	20

method would lead us to reject the null hypothesis of independence of rows and columns.

Multiway Contingency Tables

An example of a four-way contingency table is given in Table 5. These data have been slightly adapted for pedagogic purposes from a data set on adolescents supplied by Professor I.J. Goodyer of Cambridge University Developmental Psychiatry. For further discussion of the original data, see [4].

One purpose in collecting such a data set is to find a model which explains how the four variables "Gender", "Depression status", "Behavior", and "Anxiety" are interrelated.

For example, we can see from the two-way marginal Table 6 that the prevalence of depression among girls is 17%, compared with only 8% as the corresponding figure for boys. (The X^2 statistic for this 2×2 table is 4.82, which is significant when referred to χ^2_1.) Similarly, the two-way marginal Table 7, for which the corresponding X^2 statistic is 7.42, suggests that there is a strong positive association between anxiety and behavioral symptoms.

Table 5 2^4 table from psychiatry

	Behavioral symptoms:	No		Yes	
	Anxiety symptoms:	No	Yes	No	Yes
Girls	Depression = no	85	14	10	2
Girls	Depression = yes	8	4	7	4
Boys	Depression = no	107	13	8	3
Boys	Depression = yes	2	3	3	4

Table 6 A pairwise summary from Table 5

	Depression	
	No	Yes
Girls	111	23
Boys	131	12

Table 7 Another pairwise summary from Table 5

		Anxiety symptoms	
		No	Yes
Behavioral symptoms	No	202	34
Behavioral symptoms	Yes	28	13

However, we can see that this piecemeal and *ad hoc* approach to the data analysis is rather unsatisfactory. Looking at the pairwise marginal tables can raise some interesting suggestions, but it may also be misleading because it possibly conceals important features of the data, as we shall see later. Furthermore, even for a four-way contingency table there are six possible pairwise marginal tables to be examined, so that it is clear that for large tables, for example seven-dimensional, we need a more focused modeling strategy.

We return to the above practical example when we have discussed types of independence for multiway contingency in a formal way.

Suppose that the rows, columns, layers, etc. of the multiway table are labeled A, B, C, etc. There are many different sorts of independence between these variables that are possible. This makes analysis of multiway contingency tables interesting and complex. Fortunately, the relationship between the variety of types of independence and loglinear models fits naturally within the GLM framework. We will once again make use of the relationship between the Poisson and the multinomial in the context of loglinear models.

An example with only three variables, say A, B, and C, serves to illustrate the methods used in tables of dimension higher than two. Suppose that A, B, and C correspond, respectively, to the rows, columns, and layers of the three-way table. Let

$$p_{ijk} = \Pr(A = i, B = j, C = k) \quad \text{for } i = 1, \ldots, I, \; j = 1, \ldots, J, k = 1, \ldots, K, \tag{13}$$

so that $\sum p_{ijk} = 1$, and let (n_{ijk}) be the corresponding observed frequencies, assumed to be observations from a multinomial distribution, parameters n, (p_{ijk}).

There are eight different hypotheses corresponding to types of independence between A, B, and C that we now consider. Assume in all of these that the parameters given are such that $\sum p_{ijk} = 1$.

$$\mathrm{H}_0 : p_{ijk} = p_{i++}p_{+j+}p_{++k} \quad \text{for all } i, j, k, \tag{14}$$

thus H_0 corresponds to A, B, and C independent.

$$\mathrm{H}_1 : p_{ijk} = p_{i++}p_{+jk} \quad \text{for all } i, j, k, \tag{15}$$

thus H_1 corresponds to A independent of (B, C).

(Likewise, we could consider the hypothesis B independent of (A, C), and the hypothesis C independent of (A, B).)

$$\mathrm{H}_2 : \frac{p_{ijk}}{p_{i++}} = \left(\frac{p_{ij+}}{p_{i++}}\right)\left(\frac{p_{i+k}}{p_{i++}}\right) \quad \text{for all } i, j, k. \tag{16}$$

Thus H_2 is equivalent to

$$\Pr(B = j, C = k|A = i) = \Pr(B = j|A = i) \times \Pr(C = k|A = i) \quad \text{for all } i, j, k, \tag{17}$$

and so H_2 corresponds to the hypothesis that, for each i, conditional on $A = i$, the variables B and C are independent. In this case we say that B and C are independent, conditional on A. (Likewise, we can define two similar hypotheses by interchanging A, B, and C.)

Independence and conditional independence between variables can be more clearly understood by showing suitable *association graphs* (*see* **Interaction Model**). In this graphical representation of the hypotheses, we represent the variables A, B, and C by the *vertices* of the graph, with links between these vertices representing dependence, and absence of links representing independence. Thus

H_0 corresponds to no links between any of A, B, and C.

H_1 corresponds to a link between B and C only, and A as an isolated point.

H_2 corresponds to a link between A and B, and a link between A and C: thus B and C are linked only through A; see the diagram below.

The theory of association graphs is a useful and elegant one, particularly suitable for displaying and understanding the structure of high-dimensional tables.

Finally, we consider

$$\mathrm{H}_3 : p_{ijk} = \alpha_{jk}\beta_{ik}\gamma_{jk} \quad \text{for all } i, j, k, \text{ for some } \alpha, \beta, \gamma. \tag{18}$$

This hypothesis, which is symmetric in A, B, and C, cannot be given an interpretation in terms of conditional probability (see Figure 1). It corresponds to saying that the *interaction* between any two of the three factors, say A and B, given the level of the third factor, say C, is independent of the level of C. We say that H_3 corresponds to "no three-way interaction" between A, B, and C.

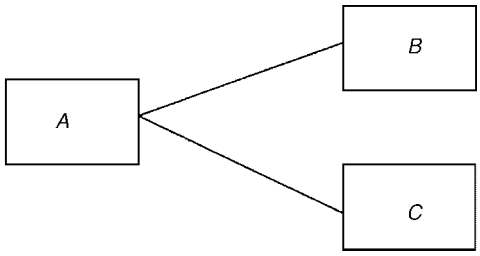

Figure 1 B and C conditionally independent, given A

One way of writing this formally is to say that for each i, j the **odds ratio** describing the dependence between A and B

$$\frac{(p_{ijk}p_{IJk})}{(p_{iJk}p_{Ijk})} \tag{19}$$

is the same for all k.

The eight hypotheses are related to one another as follows: for any given probabilities (p_{ijk}),

H_0 implies H_1, which in turn implies H_3, and

H_0 implies H_2, which in turn implies H_3, and

$\mathrm{H}_1$and H_2 are together equivalent to H_0.

Thus H_0 is the strongest hypothesis and H_3 is the weakest.

All of the eight hypotheses above may be written as loglinear hypotheses and hence tested within the GLM framework with the Poisson distribution and log link function. Since the log is the *canonical link function* for the Poisson distribution, it is the *default* link function for the Poisson in GLM terms.

For example, we may rewrite H_2 as

$$\mathrm{H}_2 : \log(p_{ijk}) = \phi_{ij} + \psi_{ik} \quad \text{for all } i, j, k, \tag{20}$$

for some ϕ, ψ. In the GLM notation for interactions between factors, this corresponds to the model

$$A * B + A * C \text{ or, equivalently, } A * (B + C).$$

We pause to explain briefly the GLM notation for interactions between factors. For example, consider the model

$$A * B + A * C.$$

In the context of loglinear models, this means that we can write $\log(p_{ijk})$ as

$$\log(p_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} \tag{21}$$

for all i, j, k. Here the set of parameters $[(\alpha\beta)_{ij}]$ is termed $A.B$, the AB interaction, the set $[(\alpha\gamma)_{ik}]$ the AC interaction, and the sets (α_i), (β_j), (γ_k) are, respectively, the main effects of A, B, and C denoted by A, B, C.

For parameter identifiability in our computations we will need to impose constraints on (α_i), etc. We assume in the numerical examples that follow that the *corner-point* constraints are imposed; thus

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad \gamma_1 = 0, \quad (\alpha\beta)_{1j} = 0, \tag{22}$$

and so on, so that any parameter with a subscript 1 anywhere is set to zero. Different software may impose a different set of constraints: this is always a confusing point for a beginner.

In the same GLM notation, H_0, H_1, and H_3 correspond, respectively, to $A + B + C$, $A + B * C$, and $(B * C + A * B + A * C)$.

For completeness, we will also define the *saturated* model H_s say, which makes no independence statement about A, B, and C:

$$H_s: \log(p_{ijk}) = \rho_{ijk} \text{ for all } i, j, k, \tag{23}$$

for some ρ: in GLM notation this is simply

$$A * B * C.$$

In terms of our association graph, this corresponds to all three links being present between the vertices A, B, and C.

Example

Consider the $2 \times 2 \times 2$ table shown in Table 8.

For example, A might correspond to "agree with a controversial statement" (such as "Women are not inherently better than men at looking after children", B might be the educational level (below or above average), and C might be the gender (men/women).

Table 8 A 2^3 table

	$C = 1$		$C = 2$	
	$A = 1$	$A = 2$	$A = 1$	$A = 2$
$B = 1$	17	23	36	50
$B = 2$	29	14	59	24

As our baseline model, we first fit the saturated model

$$A * B * C.$$

This is bound to give a perfect fit, with deviance and df both zero. So this first step in the model fitting might not appear to be a useful exercise. However, inspection of the resulting *parameter estimates* together with their *standard errors* suggests that the three-factor interaction term, $A.B.C$ can be dropped from the model, since the estimate of $A.B.C$ is -0.197 with se $= 0.561$. So our next step is to fit $(A + B + C) * (A + B + C)$ which gives a deviance of 0.12, with 1 df. Comparison with χ^2_1 shows that this model, which is H_3, is a good fit. Thus there is no three-way interaction between A, B, and C, and so, for example, we can say that the way in which the response to the question A depends on educational level is the same for both men and women.

Again, we compare the parameter estimates with their standard errors, and find that the pairwise interaction $A.C$ is -0.068, with se $= 0.28$. This suggests that $A.C$ should now be dropped from the model, and indeed the resulting model $(A + C) * B$ has deviance 0.18 (2 df). So this model is a good fit. It states that conditional on the educational level, the response to A is independent of the gender.

This model can be rewritten as $A * B + B * C$, and this has the consequence that the *observed and fitted frequencies* for the two-way marginal table $A * B$ must agree exactly, and similarly for the $B * C$ marginal table.

Inspection of the parameter estimates and their standard errors for this model shows that $B.C$ can also be dropped from the model, leading to the model $A * B + C$, which has deviance 0.34 (3 df). But this final step would not make sense if the data were collected with *fixed* totals in the $B.C$ marginal table, since under the model $A * B + C$ we would not necessarily have exact agreement between the observed and fitted frequencies of the $B.C$ table.

The Relationship between Binomial Logistic Regression and Loglinear Models in a Multiway Contingency Table

In a multiway contingency table, it may not be appropriate to treat the variables, say A, B, C, etc. symmetrically. For example, it may be more natural to treat A as a *response* variable, and B, C, etc. as *explanatory* variables.

In particular, if the number of levels of A is two, for example, corresponding to "yes, no", then it may make the analysis easier to interpret if we carry out a binomial logistic regression of A on the factors B, C, etc.

Such an analysis is not essentially different from a loglinear analysis. We can see from the following considerations that there must be certain exact correspondences between the two approaches. To be specific, take the case in which (Y_{ijk}) is multinomial, parameters n, (p_{ijk}) and suppose $i = 1, 2$. Write y_{+jk} as $y_{1jk} + y_{2jk}$. Then $Y_{1jk}|y_{+jk}$ are independent binomial variables, parameters y_{+jk}, θ_{jk}, where

$$\theta_{jk} = \frac{p_{1jk}}{p_{+jk}}. \tag{24}$$

So, for example, the model $A * B + B * C + C * A$ for (p_{ijk}) can be shown to be equivalent to the model for $\text{logit}\,\theta_{jk} = \log[\theta_{jk}/(1 - \theta_{jk})]$,

$$\text{logit}\,\theta_{jk} = \beta_j + \gamma_k. \tag{25}$$

For example, we could use the data from Table 8 above, with A as the response variable, so that we use the binomial proportions 17/40, 29/43, 36/86, and 59/83 as the responses corresponding to the factors (B,C) as (1,1), (2,1), (1,2), and (2,2). In this case it may be seen that the deviance and the fitted frequencies for the logit model $B + C$ are *exactly* the same as those for loglinear model $A * B + B * C + C * A$ with data for the $2 \times 2 \times 2$ and the multinomial model, as above.

We now return to the 2^4 contingency table, Table 5. The simplest model consistent with these data is (gender + anxiety symptoms + behavioral symptoms) * depression. This model has deviance 4.91, with 8 df. For a relatively complex table, a good way to find the simplest reasonable model is to start by fitting the saturated model, and then "step down", discarding the unnecessary parameters, using the **Akaike information criterion** (AIC) as the guide to when to stop the stepping down process.

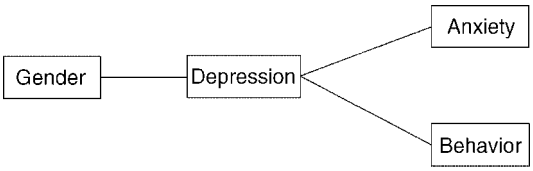

Figure 2 A graphic model for psychiatry data

The final model here shows the key role of depression in explaining the dependence in the four-way table: conditional on the depression status, the two variables anxiety and behavior are independent of each other and are also independent of gender. Its graphical representation is a graph in which the three vertices "gender", "anxiety symptoms", and "behavioral symptoms" are linked to depression, and no other links are present (see Figure 2).

This final model also shows that summarizing the data by certain 2×2 tables, for example by *collapsing* over the variable gender and depression, may be a misleading way to investigate the association between variables. To put this more technically, for the current example with the final model, the two-way marginal table anxiety symptoms × behavioral symptoms is not among the sufficient statistics for the unknown parameters.

The danger in obtaining misleading results by collapsing a contingency table is known as **Simpson's paradox** (also known as *Yule's paradox*).

This brief article has introduced the topic of contingency tables, with particular discussion of types of independence for multiway tables. The reader will have realized that in none of the methods described above is any attention paid to the *order* of the categories of a variable: for example, if A takes as possible values 1, 2, 3, and 4, the above analyses would all remain the same if these levels were renamed as *apples*, *pears*, *oranges*, *bananas*. In this sense our treatment of the variables so far has been entirely *nominal* (*see* **Nominal Data**). This approach may not always be the most sensible one to follow. For example, both of the rows and columns of Table 1 are ordered. We might prefer our analysis to recognize this fact (*see* **Ordered Categorical Data**).

This article represents a brief introduction to the topic of contingency tables, and is not an exhaustive

coverage of the available methods. For more comprehensive coverage, the reader is referred to the two textbooks by Agresti [1, 2].

References

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. Wiley, New York.

[2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

[3] Altham, P.M.E. (1975). Quasi-independent triangular contingency tables, *Biometrics* **31**, 233–238.

[4] Altham, P.M.E. (1992). Fitting graphical models to multiway contingency tables in GLIM, *GLIM Newsletter* **21**, 4–8.

[5] Bishop, Y.M.M. & Fienberg, S.E. (1969). Incomplete two-dimensional contingency tables, *Biometrics* **25**, 119–128.

[6] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

(*See also* **Categorical Data Analysis**)

P.M.E. ALTHAM

Continuity Correction

The term "continuity correction" has traditionally referred to an adjustment made when using a continuous probability distribution to approximate a discrete distribution [34]. Statisticians realized early that large-sample approximations to the exact distributions often arising in discrete data settings are poor in small samples and proposed simple adjustments to existing summary statistics in order to provide more reliable inferences. Although the earliest proposals involved corrections of **normal** approximations for making probability statements about **binomial**, **Poisson**, or **negative binomial** random variables [36], continuity corrections have been most prominent in the analysis of **contingency tables**. Recently, modern computing and the availability of exact methods in several software packages [42] have lessened the need for these specific procedures [3] (*see* **Exact Inference for Categorical Data**). However, the term "continuity correction" is also used more generally to refer to any method designed to address the effects on inferential procedures of the discreteness of a small-sample exact distribution (e.g. [21], with reference to the mid-***P* value**). Development of continuity corrections in this broader sense is an active area of statistical research. In this entry, we briefly outline the historical development of the traditional continuity correction in contingency tables and review the performance of continuity corrections in the broader context of small-sample strategies for categorical data.

Hypothesis Testing

The simplest continuity correction involves the normal approximation to a discrete distribution, such as the binomial probability mass function. Consider a binomial random variate X for sample size n and success probability π. The normal approximation to $P(X = x)$ is $P(x - 0.5 \le Z \le x + 0.5)$, where Z is normally distributed with mean and variance matching those from the true $\text{Bin}(n, \pi)$ distribution. The presence of 0.5 in the probability statement constitutes a *continuity correction* in that it accounts for the fact that we use the continuous normal approximation, and not the discrete binomial distribution, to make probability statements about X.

Now consider a **2 × 2 table** with cell frequencies $\{n_{ij}\}$ formed by cross-classifying binary variables X and Y. Denote the marginal row and column totals as $\{n_{i+}\}$ and $\{n_{j+}\}$, respectively, and let n be the total sample size. The well-known chi-squared test of independence uses the Pearson statistic

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - m_{ij}\right)^2}{m_{ij}}, \tag{1}$$

where $m_{ij} = n_{i+}n_{+j}/n$ are the fitted values under independence (*see* **Chi-square Tests**). Under the **null hypothesis**, X^2 has a large-sample chi-squared distribution with 1 df. For small samples, Yates [48] proposed the correction

$$X_c^2 = \sum_i \sum_j \frac{\left(|n_{ij} - m_{ij}| - 0.5\right)^2}{m_{ij}} \tag{2}$$

and used the statistic X_c^2 to test the independence of X and Y by comparing it to the same reference chi-squared distribution (*see* **Yates's Continuity Correction**). The continuity correction 0.5 adjusts for using the continuous χ^2 distribution to approximate the exact discrete distribution of X^2 and produces P values that approximate those obtained from **Fisher's exact test**.

Mantel and Haenszel [35] proposed the same correction for testing conditional independence in stratified 2×2 tables (*see* **Stratification**). Let n_{ijk} be the cell counts in stratum k, $k = 1, \ldots, K$, formed by stratifying the relationship of X and Y by a control variable Z, and as above, let subscripted "+" denote marginal summation. These authors considered the exact null product **hypergeometric distribution** of $\{n_{11k}\}$ obtained by treating the observations in different strata as independent and the row and column totals in each stratum as fixed. Specifically, the Mantel–Haenszel statistic is

$$\begin{aligned} M^2 &= \frac{\left(\left|\sum_{k=1}^K n_{11k} - \sum_{k=1}^K m_{11k}\right| - 0.5\right)^2}{\sum_{k=1}^K V(n_{11k})} \\ &= \frac{(|n_{11+} - m_{11+}| - 0.5)^2}{\sum_{k=1}^K V(n_{11k})}, \end{aligned} \tag{3}$$

where

$$m_{11k} = \mathrm{E}(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}}$$

and $$V(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}$$

are the mean and variance of $\{n_{11k}\}$ based on this exact product hypergeometric distribution under the null hypothesis of conditional independence of X and Y given Z (*see* **Mantel–Haenszel Methods**). Under this null hypothesis, M^2 has approximately a chi-squared distribution with 1 df. Cochran [15] had earlier proposed a closely related statistic but conditioning only on the row totals and without continuity correction. As a result, others (e.g. Landis et al. [32] and Agresti [3]) have labeled (3) the Cochran–Mantel–Haenszel (C-M-H) statistic.

Like the Yates continuity-corrected statistic (2), the C-M-H statistic yields inferences that closely approximate those obtained from the corresponding exact conditional test, which also bases inference on the statistic n_{11+}. As above, assume fixed n_{i+k}. Let π_{ik}, $i = 1, 2$, $k = 1, \ldots, K$, be the probability of a success at level i of X and stratum k of Z. A **logistic regression** model for π_{ik} that specifies a homogeneous log **odds ratio** β across strata is

$$\mathrm{logit}(\pi_{ik}) = \alpha_k + \beta(\mathrm{I}[i = 2]), \tag{4}$$

where $I[\ ]$ is the indicator function. Under this model, the null hypothesis of conditional independence corresponds to $H_0: \beta = 0$, and the **sufficient statistic** for β is n_{11+}. The exact approach bases inference on the distribution of this statistic that is free from the **nuisance parameters** $\{\alpha_k\}$ by conditioning on their sufficient statistics $\{n_{+1k}\}$. The resulting exact joint distribution for the $\{n_{11k}\}$ is the product of the K hypergeometric distributions considered by Mantel and Haenszel, from which one can compute the exact distribution of n_{11+}.

One can now conduct both the C-M-H test and the exact test, as well as compute the corresponding **confidence intervals** for a common odds ratio, using the statistical package **StatXact** [21]. Because exact results are now routinely available in StatXact, there is less of a need for the approximation (3) to the exact result. As a result, StatXact does not report results based on this continuity-corrected version but instead makes a clear distinction between asymptotic and exact tests in this setting by reporting the uncorrected version of (3).

There has been a long running controversy [19, 25, 26] as to the appropriateness of the continuity corrections in (2) and (3). Much of the criticism of these corrections arises from the fact that they lead to tests that approximate exact results, which for small or highly unbalanced samples can yield conservative inferences due to the discreteness of the exact distribution [3, 4]. As a result, the continuity-corrected approximations of the exact tests are also typically conservative. For instance, D'Agostino [22] noted that it is not uncommon for the actual level of a **hypothesis test** with nominal 5% level based on the continuity-corrected Pearson statistic to actually be 1%. Thus, less conservative alternative strategies may be preferable when the exact distribution is highly discrete. For instance, rather than rejecting a null hypothesis based on a preset significance level, Yates [49] and the discussants of his article recommended either (a) simply reporting the P value associated with an exact test or (b) selecting as the type I error rate one of the values having positive probability mass in the discrete P value distribution. For a single 2×2 table, D'Agostino, Chase, and Belanger [23] recommended using either the two-sample t-test or the uncorrected X^2 (*see* **Student's *t* Statistics**). In principle, one can avoid this conservatism by using supplementary **randomization** on the boundary of the **critical region** to construct a uniformly most powerful unbiased (UMPU) α-level test [13, Chapter 8] (*see* **Power**). However, this strategy is unattractive in practice, as two investigators observing the same results can arrive at different conclusions.

Another approach uses the mid-P value [31] to adjust for discreteness. Consider a test statistic T with observed value t_o and an alternative hypothesis such that large values of T reject H_0. The mid-P value is

$$\begin{aligned}\text{mid-}P &= P_0(T > t_o) + \tfrac{1}{2}P_0(T = t_o) \\ &= \text{exact P} - \tfrac{1}{2}P_0(T = t_o),\end{aligned} \tag{5}$$

where P_0 denotes probabilities computed under the null hypothesis. Because this adjustment subtracts one-half of the null probability of an observed response from the ordinary exact P value, mid-P is always less than the ordinary exact P value, leading to a less conservative test. Agresti [3] noted that mid-P behaves more like the P value for a test statistic having a continuous distribution, with the null

distribution of mid-P being more like a **uniform distribution**. For instance, E(mid-P) = 0.5, which is not the case for the ordinary P value based on a discrete distribution. It is in this sense that some consider mid-P a type of continuity correction [21]. Although tests based on mid-P sacrifice precise error control in the sense that the type I error rate is not theoretically guaranteed to be below the nominal level, empirical investigations have shown that in some situations mid-P actually does preserve this nominal level. For instance, Mehta and Walsh [38] reported that mid-P often preserves nominal levels when testing a common odds ratio in stratified 2×2 tables. Overall, for small samples, tests based on mid-P result in actual levels that are on average closer to the nominal levels than those of the corresponding fully exact tests. In addition, this strategy may be applied to any discrete problem for which the exact distribution is obtainable – single binomial proportion, two-sample binomials, paired binomials [27], stratified tables, or more generally small-sample logistic regression [28, 37]. Hwang and Yang [29] have presented additional theoretical arguments for the use of mid-P, and Agresti [2, 3] has recommended the use of the mid-P value as a useful, general strategy for inference in discrete problems.

Several authors have proposed other modifications of the ordinary exact P value that result in less conservative inference. Cohen and Sackrowitz [18] proposed a P value that, like mid-P, adds only a portion of the probability mass $P_0(T = t_o)$ to $P_0(T > t_o)$. In particular, these authors proposed refining the sample space satisfying $T = t_0$ by calculating the null probability of each sample in this set and adding to $P_0(T > t_o)$ only those sample probabilities that are less than or equal to the probability of the observed sample. Similar to the mid-P approach, this alternative strategy results in a test with a smaller P value, and hence greater power, than its exact counterpart. This advantage increases when the distribution of T is highly discrete. Also like mid-P, in theory, this strategy can be applied in a large number of discrete settings. For example, Corcoran, Mehta, and Senchaudhuri [16] used this modified P to test for trend in a $2 \times c$ table with ordered columns (*see* **Trend Test for Counts and Proportions**). The actual level of the test based on this modified P cannot exceed the nominal level for any value of the unknown parameter, by the same reasoning that proves this property of the test based on the ordinary exact P value. Thus, unlike a test based on mid-P, the test based on this modified P retains the strong error control of the exact test.

Interval Estimation

Recently, a large research effort has focused on estimation strategies when the exact distribution of interest is highly discrete. Just as an exact test (without supplementary randomization on the boundary of the critical region) is conservative, confidence intervals based on exact probabilities are conservative in that, for any fixed parameter value, the actual coverage probability can be much larger than the nominal confidence level. This high coverage comes at the price of a loss of precision, in the form of wider than necessary confidence intervals. Several authors have shown that, if one is willing to relax the requirement that the actual confidence level of the interval always be no less than the nominal level, one may construct intervals that, when compared to the exact interval, have coverage probabilities much closer to the nominal confidence level for most parameter values.

For instance, for a single binomial proportion, Agresti and Coull [6] argued that the Clopper–Pearson (1934) "exact" confidence interval for the success probability, based on inverting equal-tailed binomial tests of $H_0: \pi = \pi_0$, is not necessarily optimal in small samples because of this conservatism. Several authors (e.g. [6, 10, 39, 46]) have recommended the score interval as an alternative (*see* **Likelihood**). Let X denote a binomial variate for sample size n, and let $\hat{\pi}$ denote the sample proportion. This score confidence interval, apparently first discussed by Edwin B. Wilson [47], takes the form

$$\frac{\left(\hat{\pi} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{[\hat{\pi}(1-\hat{\pi}) + z^2_{\alpha/2}/4n]/n}\right)}{(1 + z^2_{\alpha/2}/n)}, \tag{6}$$

where z_c denotes the 1-c quantile of the standard normal distribution. Expression (3) results from inverting the approximately normal test that uses the null **standard error**, that is, its endpoints are the π_0 solutions to the equations $(\hat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n} = \pm z_{\alpha/2}$. For a 95% confidence interval, Agresti and Coull [6] also proposed a simple approximation to the score interval that is simple to compute. This approximation

constructs the Wald interval (*see* **Likelihood**)

$$\tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}} \tag{7}$$

using a new sample proportion $\tilde{\pi}$ and sample size $\tilde{n}$ formed after adding two successes and two failures to the data. Both of these alternative intervals yield actual confidence levels that are typically much closer to nominal levels than those of the exact Clopper–Pearson interval.

Others have proposed additional strategies for estimation of a binomial proportion. Vollset [46] showed that the interval obtained by inverting the test based on mid-P is less conservative than the exact interval but slightly more conservative than the score interval. This interval may be preferable in situations where the score interval is slightly liberal, namely, when π is near 0 or 1. Brown, Cai, and Das Gupta [10] recommended an interval based on Jeffrey's **prior** in a **Bayesian** setting. Interestingly, this interval approximates the mid-P–corrected Clopper–Pearson interval. Still another effective strategy for reducing conservatism in this setting is inversion of a single two-sided rather than two one-sided tests [4, 7]. That is, let $f(X;\pi)$ be the probability mass function of X given π, and let x be the observed value of X. Sterne [45] showed that the $100 \times (1-\alpha)$ confidence interval for π obtained by inverting a single two-sided test consists of those values of π that satisfy

$$P_\pi\left[f(X;\pi) \le f(x;\pi)\right] > \alpha. \tag{8}$$

That is, the confidence interval consists of inverting a test that uses as a P value the sum of null probabilities less than or equal to that for the observed response. The resulting interval is exact, yet uniformly shorter than the Clopper–Pearson interval. Blyth and Still [9] and Casella [11] refined this approach to satisfy several optimality properties, and Blaker [8] proposed a related exact interval. The Blyth–Still–Casella intervals are available in the latest version of the StatXact software [21].

It is worth noting that Vollset considered continuity-corrected versions of both the score interval (6) and the normal-theory Wald interval [i.e. equation (7) using the original sample proportion $\hat{\pi}$ and sample size n]. Both of these corrected intervals use the normal continuity correction to the binomial observation x mentioned earlier. Vollset noted that the correction does not significantly improve the well-known horrible performance [10] of the Wald interval. However, the continuity-corrected score interval approximates the exact interval, leading to conservative inference. Casella [12] expressed a preference for this corrected score interval in the classroom. Brown, Cai, and Das Gupta [10] explicitly showed the effect of this correction on the score interval by plotting the coverage probabilities and expected lengths of the score intervals with and without the correction. For other general reviews of inference for a single binomial proportion, see Agresti [2, 4], Agresti and Min [7], and Newcombe [39].

Similar patterns hold in other discrete settings. For a review, see Agresti [2, 4, 7]. For the difference between two independent binomial proportions, Nurminen [41] proposed inverting the large-sample score test. Let $\delta = \pi_1 - \pi_2$ denote the difference between two independent proportions. This test for $H_0\colon \delta = \delta_0$ treats the test statistic

$$T = \frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta_0}{\sqrt{\tilde{\pi}_1(1-\tilde{\pi}_1)/n_1 + \tilde{\pi}_2(1-\tilde{\pi}_2)/n_2}}, \tag{9}$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ denote the ML estimates of π_1 and π_2 subject to $\pi_1 - \pi_2 = \delta_0$, as a standard normal random variable. Agresti [4] noted that the score approach yields reasonable coverage probabilities in this context as well. Chan and Zhang [14] proposed an exact test formed by inverting two one-sided tests based on the null probability ordering of (9). Because either π_1 or π_2 is a nuisance parameter, the exact intervals in this setting are unconditional. That is, because the difference $\delta = \pi_1 - \pi_2$ does not correspond to a canonical parameter in an **exponential family** model, it is not possible to obtain a likelihood free of π_1 by conditioning on its sufficient statistic. Instead, the exact approach inverts the test for $H_0\colon \delta = \delta_0$ based on the unconditional P value, $\sup_{\pi_1}\left(P_{\delta=\delta_0,\pi_1}(T \ge t_o)\right)$. Agresti and Min [7] showed that inverting the analogous two-sided score test can lead to shorter intervals, and Agresti and Caffo [5] showed that a simple adjustment that constructs the normal-theory Wald interval after adding a success and a failure to each sample is also effective. Newcombe [40] considered approaches that form intervals by inverting two test statistics, one for each proportion, both with and without continuity corrections. As is the case for a single proportion, these investigations showed that continuity-corrected Wald-type

intervals are too liberal, whereas the continuity-corrected score intervals are quite conservative. Santner and Snell [43] proposed an unconditional exact interval for δ based on the unstandardized statistic $\widehat{\pi}_1 - \widehat{\pi}_2$, although recent results have shown this statistic to be extremely conservative [14, 21]. For related approaches, see [17, 44]. Recent software now reflects the increased number of choices for inference in this setting. For example, StatXact-5 gives the user a choice of constructing the Agresti-Min, the Chan–Zhang, or the Santner–Snell interval for δ, although Cytel notes [21] that the latter conservative interval is included largely for historical reasons.

These general trends also hold when interest focuses on the odds ratio in 2×2 tables. For a single table, Cornfield [20] proposed the exact conditional interval for the odds ratio θ in a single 2×2 table. One obtains this interval by conditioning on the row and column totals of the table and by inverting two one-sided $\alpha/2$ tests for θ based on the resulting exact hypergeometric distribution. Several authors (e.g. [1, 33]) have shown this interval to be highly conservative for small or highly unbalanced samples. Cornfield [20] and Fisher [24] proposed a continuity-corrected approximation to Cornfield's exact interval. This interval, formed by inverting the distribution of the continuity-corrected Pearson chi-squared statistic (2) after conditioning on the observed row and column totals, is the continuity-corrected score interval in this case. As expected, this strategy is also conservative, as it once again approximates the exact interval [1, 33]. Agresti [1] noted that the interval obtained by inverting the uncorrected score interval yields much shorter intervals, while producing coverage probabilities at or near nominal levels. Less work exists for the stratified 2×2 setting, although Kim and Agresti [30] proposed an exact confidence interval for a common odds ratio formed by inverting a test based on the modified P value discussed above.

In summary, the traditional continuity corrections of score statistics typically yield results that closely approximate those from the corresponding exact method. As a result, tests and intervals based on these continuity-corrected statistics yield reliably conservative inferences but with actual confidence levels often far above nominal levels. In applications in which maintaining either the type I error rate or nominal confidence level is absolutely necessary, exact methods are preferred to other approaches, and recent advances in computational algorithms make such methods accessible in commercial **software**. Thus, the continuity-corrected score methods are perhaps useful when one is interested in exact inference but does not have access to software with exact capabilities. In this sense, one can obtain a good approximation to an exact result using simple formulas. In situations in which maintaining nominal levels is not as crucial, one can use an alternative method, such as an approach based on the uncorrected score statistic or the mid-P value, that will yield an actual level close to, but not always bounded by, the nominal level.

References

[1] Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples, *Biometrics* **55**, 597–602.

[2] Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies, *Statistics in Medicine* **20**, 2709–2722.

[3] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. Wiley, New York.

[4] Agresti, A. (2003). Dealing with discreteness: making 'exact' confidence intervals for proportions, difference of proportions, and odds ratios more exact, *Statistical Methods in Medical Research* **12**, 3–21.

[5] Agresti, A. & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, *The American Statistician* **54**, 280–288.

[6] Agresti, A. & Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions, *The American Statistician* **52**, 119–126.

[7] Agresti, A. & Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions, *Biometrics* **57**, 963–971.

[8] Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions, *Canadian Journal of Statistics* **28**, 783–798.

[9] Blyth, C.R. & Still, H.A. (1983). Binomial confidence intervals, *Journal of the American Statistical Association* **78**, 108–116.

[10] Brown, L.D., Cai, T.T. & Das Gupta, A. (2001). Interval estimation for a binomial proportion (with discussion), *Statistical Science* **16**, 101–133.

[11] Casella, G. (1986). Refining binomial confidence intervals, *The Canadian Journal of Statistics* **14**, 113–129.

[12] Casella, G. (2001). Comment on "Interval estimation for a binomial proportion" by Brown, Cai, and Das Gupta, *Statistical Science* **16**, 120–122.

[13] Casella, G. & Berger, R.L. (1990). *Statistical Inference*. Wadsworth & Brooks Cole, Pacific Grove.

[14] Chan, I.S.F. & Zhang, Z.X. (1999). Test-based exact confidence intervals for the difference of two binomial proportions, *Biometrics* **55**, 1202–1209.

[15] Cochran, W.G. (1954). Some methods of strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.

[16] Cochran, C., Mehta, C.R. & Senchaudhuri, P. (2000). Power comparisons for tests of trend in dose-response studies, *Statistics in Medicine* **19**, 3037–3050.

[17] Coe, P.R. & Tamhane, A.C. (1993). Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities, *Communications in Statistics, Part B – Simulation and Computation* **22**, 925–938.

[18] Cohen, A. & Sackrowitz, H.B. (1992). An evaluation of some tests of trend in contingency tables, *Journal of the American Statistical Association* **87**, 470–475.

[19] Conover, W.J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables (with discussion), *Journal of the American Statistical Association* **69**, 374–382.

[20] Cornfield, J. (1956). A statistical problem arising from retrospective studies, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **4**, 135–148.

[21] Cytel Software Corporation. (2001). *StatXact 5: Statistical Software for Exact Nonparametric Inference*. Cytel, Cambridge.

[22] D'Agostino, R.B. (1990). Comment on "Yates's correction for continuity and the analysis of 2×2 contingency tables" by M. G. Haviland, *Statistics in Medicine* **9**, 363–367.

[23] D'Agostino, R.B., Chase, W. & Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations, *The American Statistician* **42**, 198–202.

[24] Fisher, R.A. (1962). Confidence limits for a cross-product ratio, *Australian Journal of Statistics* **4**, 41.

[25] Grizzle, J.E. (1967). Continuity correction in the χ^2-test for 2×2 tables (with discussion), *The American Statistician* **4**, 28–32.

[26] Haviland, M.G. (1990). Yates's correction for continuity and the analysis of 2×2 contingency tables, *Statistics in Medicine* **9**, 363–367.

[27] Hirji, H.F. (1991). A comparison of exact, mid-P, and score tests for matched case-control studies, *Biometrics* **47**, 487–496.

[28] Hirji, K.F., Mehta, C.R. & Patel, N.R. (1987). Computing distributions for exact logistic regression, *Journal of the American Statistical Association* **82**, 1110–1117.

[29] Hwang, J.T.G. & Yang, M.-C. (2001). An optimality theory for mid-P values in 2×2 contingency tables, *Statistica Sinca* **11**, 807–826.

[30] Kim, D. & Agresti, A. (1995). Improved exact inference about conditional association in three-way contingency tables, *Journal of the American Statistical Association* **90**, 632–639.

[31] Lancaster, H.O. (1961). Significance tests in discrete distributions, *Journal of the American Statistical Association* **56**, 223–234.

[32] Landis, J.Richard, Cooper, Murray.M., Kennedy, Thomas. & Koch, Gary.G. (1979). A computer program for testing average partial association in three-way contingency tables (PARCAT), *Computer Methods and Programs in Biomedicine* **9**, 223–246.

[33] Lui, K.-J. & Lin, C.-D. (2003). A Revisit on comparing the asymptotic interval estimators of odds ratio in a single 2×2 table, *Biometrical Journal* **45**, 226–237.

[34] Mantel, N. & Greenhouse, S.W. (1967). What is the continuity correction? *The American Statistician* **22**, 27–30.

[35] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[36] Maxwell, E.A. (1988). Continuity corrections, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 172–174.

[37] Mehta, C.R., Patel, N.R. & Senchaudhuri, P. (2000). Efficient Monte Carlo methods for conditional logistic regression, *Journal of the American Statistical Association* **95**, 99–108.

[38] Mehta, C.R. & Walsh, S.J. (1992). Comparison of exact, Mid-p, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables, *The American Statistician* **46**, 146–150.

[39] Newcombe, R.G. (1998a). Two-sided confidence intervals for the single proportion: comparison of seven methods, *Statistics in Medicine* **17**, 857–872.

[40] Newcombe, R.G. (1998b). Interval estimation for the difference between independent proportions: comparison of eleven methods, *Statistics in Medicine* **17**, 873–890.

[41] Nurminen, M. (1986). Confidence intervals for the ratio and difference of two binomial proportions, *Biometrics* **42**, 675–676.

[42] Oster, R.A. (2002). An examination of statistical software packages for categorical data analysis using exact methods, *The American Statistician* **56**, 235–246.

[43] Santner, T.J. & Snell, M.K. (1980). Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables, *Journal of the American Statistical Association* **75**, 386–394.

[44] Santner, T.J. & Yamaguchi, S. (1993). Invariant small sample confidence-intervals for the difference of 2 success probabilities, *Communications in Statistics, Part B – Simulation and Computation* **22**, 33–59.

[45] Sterne, T.E. (1954). Some remarks on confidence or fiducial limits, *Biometrika* **41**, 275–278.

[46] Vollset, S.E. (1993). Confidence intervals for a binomial proportion, *Statistics in Medicine* **12**, 809–824.

[47] Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association* **22**, 209–212.

[48] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society* Supplement **1**, 217–235.

[49] Yates, F. (1984). Tests of significance for 2×2 contingency tables (with discussion), *Journal of the Royal Statistical Society, Series A* **147**, 426–463.

Further Reading

Cohen, A. & Sackrowitz, H.B. (2003). Methods of reducing loss of efficiency due to discreteness of distributions, *Statistical Methods in Medical Research* **12**, 3–21.

BRENT A. COULL

Contrasts

The objectives of most studies include comparisons among population characteristics (e.g. **means**). Most such comparisons are *contrasts*. Consider a **clinical trial** involving a control and two test treatments, with parameters μ_1, μ_2, and μ_3 the control and test treatment (population) means, respectively. The *pairwise* contrast (*see* **Paired Comparisons**) $\Psi_1 = \mu_2 - \mu_3$ compares the means of the two test treatments, while the contrast $\Psi_2 = (\mu_2 + \mu_3)/2 - \mu_1$ compares the average of the means of the two test treatments with the mean of the control. More generally, collections of contrasts may be formulated to represent research comparisons of interest. For example, certain specialized classes of contrasts determine the main effect and **interaction** sums of squares reported in **analysis of variance** (ANOVA) tables (see below).

Contrasts are formally defined and discussed here using the matrix representation of a linear model (*see* **General Linear Model**), although the concept of a contrast is not limited to experiments analyzed with such models. Let $\mathbf{Y}^{(n\times 1)}$ denote the vector of n observed responses, $\mathbf{X}^{(n\times p)}$ the $(n \times p)$ regression or design matrix of rank $r \le p$, where x_{ij} is the value of the jth independent variable for the ith observation, $i = 1, \ldots, n$, $j = 1, \ldots, p$, and $\beta^{(p\times 1)}$ the parameter vector. The linear model representation is: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon^{(n\times 1)}$ is the vector of **unbiased** (mean zero) **random errors**. A *contrast* in the parameters $\beta_1, \ldots, \beta_p$ is a linear combination

$$\Psi = \mathbf{c}'\beta = c_1\beta_1 + \cdots + c_p\beta_p \quad \text{subject to}$$
$$c_1 + \cdots + c_p = 0. \tag{1}$$

For the clinical study example above, with $\beta = (\mu_1, \mu_2, \mu_3)'$, $\Psi_1 = \mathbf{c}_1'\beta$ where $\mathbf{c}_1 = (0, 1, -1)'$ and $\Psi_2 = \mathbf{c}_2'\beta$ where $\mathbf{c}_2 = (-1, 0.5, 0.5)'$.

Estimability of Contrasts

A contrast $\Psi = \mathbf{c}'\beta$ is *estimable* if there exists an $(n \times 1)$ vector of constants, $\mathbf{b}$, such that $\mathbf{b}'\mathbf{Y}$ is an *unbiased* estimator of Ψ. Equivalently, Ψ is estimable if there exists a constant vector $\mathbf{a}$ such that $\mathbf{c}' = \mathbf{a}'\mathbf{X}$. An estimable Ψ has a *unique* **least squares** *estimate* $\hat{\Psi} = \mathbf{c}'\hat{\beta}$, where $\hat{\beta}$ is *any* solution to the normal equations $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$ [4]. Furthermore, $\hat{\Psi}$ is the Best Linear Unbiased Estimate (BLUE) of Ψ [4] (*see* **Minimum Variance Unbiased (MVU) Estimator**). Note that the preceding estimability results hold whether the variance–**covariance matrix** of ε has the standard "uncorrelated, constant variance" form, $\sigma^2\mathbf{I}_n$, or the more general form $\sigma^2\mathbf{V}$, where $\mathbf{V}$ is a known $(n \times n)$ positive definite matrix, except that in the latter case $\hat{\beta}$ is any solution of $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$. For the remainder of this discussion, the standard case will be assumed.

Variance of Contrasts

If $\Psi_1 = \mathbf{c}_1'\beta, \ldots, \Psi_q = \mathbf{c}_q'\beta$ are q estimable contrasts, then the $(q \times q)$ variance–covariance matrix of $(\Psi_1, \ldots, \Psi_q)$ is given by

$$\sigma^2\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}, \tag{2}$$

where the ith column of $\mathbf{C}$ is $\mathbf{c}_i$, $i = 1, \ldots, q$, and $(\mathbf{X}'\mathbf{X})^-$ is the same generalized inverse of the information matrix $\mathbf{X}'\mathbf{X}$ used to produce an estimated parameter vector, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$, that solves the normal equations. An estimate of the error variance, σ^2, is usually provided by the **mean square for error**.

Example: Clinical Study

For the clinical study example introduced above, the *means* model is: $Y_{ij} = \mu_i + \varepsilon_{ij}$, $j = 1, \ldots, n_i, i = 1, 2, 3$. With this parameterization, *every* linear combination of the three means is estimable, including all contrasts. With $\hat{\beta} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3)'$, where $\hat{\mu}_i = \sum_j Y_{ij}/n_i$, the unique least squares estimator of a contrast $\Psi = \mathbf{c}'\beta$ defined by (1) is $\hat{\Psi} = \sum c_i\hat{\mu}_i$. Since $\mathbf{X}'\mathbf{X}$ has inverse $(\mathbf{X}'\mathbf{X})^{-1} = \text{diag}(n_1^{-1}, n_2^{-1}, n_3^{-1})$, using (2), the variance of $\hat{\Psi}$ is $\text{var}(\hat{\Psi}) = \sigma^2 \sum c_i^2/n_i$.

In the equivalent *effects* model parameterization, $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, the individual components of $\beta = (\mu, \tau_1, \tau_2, \tau_3)'$ are not estimable and $\mathbf{X}'\mathbf{X}$ does not have an inverse. However, a contrast in the means, $\Psi = \sum c_i\mu_i$, is (i) estimable and (ii) equal to the *same* contrast in the treatment effects, $\Psi = \sum c_i\tau_i$. The unique least squares estimator of Ψ is $\hat{\Psi} = \sum c_i\hat{\tau}_i$, irrespective of the choice of generalized inverse used to find a solution $\hat{\beta} = (\hat{\mu}, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3)'$ of the normal equations [4]. The variance of $\hat{\Psi}$ is given by (2) and is the same in both the means and effect model parameterizations. See the *SAS/STAT User's*

Guide [2] entry for procedure GLM for solutions corresponding to the generalized inverse of $\mathbf{X}'\mathbf{X}$ obtained with the *set-to-zero* side conditions (*see* **Software, Biostatistical**). In particular, the *only* estimable linear combinations of the effects parameters [i.e. the coefficient of μ in (1) is zero] are contrasts [3]. Furthermore, if Ψ is a contrast in the effects parameters, then the coefficient of μ must be zero.

Main Effect and Interaction Contrasts

The results in the preceding example hold for any one-way treatment structure, $\beta = (\mu_1, \ldots, \mu_p)'$. Contrasts that represent main effect and interaction comparisons are discussed here in the context of the two-way **factorial** treatment structure, where factor A has levels $i = 1, \ldots, I$ and factor B has levels $j = 1, \ldots, J$. There are $n_{ij} > 0$ observations on each *treatment combination* (i.e. no missing cells). Extensions to three or more factors follow accordingly but with increased technical and notational complexity.

The population mean for the (i, j)th treatment combination is μ_{ij}, and is modeled by $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ in the corresponding *full* effects model. The linear model is $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, $k = 1, \ldots, n_{ij}$. Any contrast $\Psi = \sum_i \sum_j c_{ij}\mu_{ij}$, $\sum_i \sum_j c_{ij} = 0$, in the means

$$(\mu_{11}, \ldots, \mu_{1J}, \mu_{21}, \ldots, \mu_{2J}, \ldots, \mu_{I1}, \ldots, \mu_{IJ})', \tag{3}$$

is estimable and has the same estimate (and variance) whether the means or full effects model parameterization is used.

Following Milliken et al. [1], a contrast $\Psi = \sum_i \sum_j c_{ij}\mu_{ij}$ is an *A main effects* contrast if $c_{ij} = c_i$, for each $j = 1, \ldots, J$, in which case $\sum c_i = 0$, as usual. In this context, Ψ may be expressed as $\Psi = \sum c_i \overline{\mu}_i.$, where $\overline{\mu}_i. = \sum_j \mu_{ij}/J$ is the *population marginal mean* (also referred to as the "least squares mean", LSMEAN, in the SAS [2] procedure GLM) for level i of factor A. Similarly, a *B main effects* contrast requires that $c_{ij} = c_j$, for each $i = 1, \ldots, I$. Finally, a contrast Ψ is an *interaction contrast* if (i) $\sum_i c_{ij} = 0$ for each $j = 1, \ldots, J$ and (ii) $\sum_j c_{ij} = 0$ for each $i = 1, \ldots, I$.

A contrast Ψ in the IJ means (3) is equivalently written as a contrast in the full effects parameters

$$\Psi = \sum_i c_{i\cdot}\alpha_i + \sum_j c_{\cdot j}\beta_j + \sum_i \sum_j c_{ij}\gamma_{ij}, \tag{4}$$

where $c_{i\cdot} = \sum_j c_{ij}$ and $c_{\cdot j} = \sum_i c_{ij}$. Consequently, if Ψ is an A main effects contrast, $c_{\cdot j} = 0$ and $c_{i\cdot} = Jc_i$, in which case Ψ in (4) may be written $\Psi = J\sum_i c_i\alpha_i + \sum_i c_i\gamma_{i\cdot}$, where $\gamma_{i\cdot} = \sum_j \gamma_{ij}$. Similarly, if Ψ is a B main effects contrast, $c_{i\cdot} = 0$ and $c_{\cdot j} = Ic_j$, whence $\Psi = I\sum_j c_j\beta_j + \sum_j c_j\gamma_{\cdot j}$. Finally, an interaction contrast in the means (3) has exactly the same form in the interaction parameters $\{\gamma_{ij}\}$, that is, $\Psi = \sum_i \sum_j c_{ij}\mu_{ij} = \sum_i \sum_j c_{ij}\gamma_{ij}$ provided Ψ is an interaction contrast.

Orthogonal Contrasts

Suppose that $\Psi_k = \mathbf{c}_k'\beta$, $k = 1, \ldots, q (2 \le q \le p - 1)$ are q *linearly independent* contrasts in the parameters, β. Then these q contrasts are *mutually orthogonal* if they are uncorrelated with each other (*see* **Orthogonality**). Thus, if $\mathbf{C}$ is the $(p \times q)$ matrix whose kth column is $\mathbf{c}_k$, then these q contrasts are mutually orthogonal if, by (2), $\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}$ is a diagonal matrix. Orthogonality of contrasts is useful when the errors, ε_i, are **normally distributed**, in which case the contrast estimators are statistically independent, as are the sums of squares for each contrast (see below), which may simplify practical interpretation of results. For example, in an experiment with a factor at four quantitative and equally spaced levels, and equal replication at each factor level, linear, quadratic and cubic orthogonal polynomial contrasts defined by $\mathbf{c}_1 = (-3, -1, 1, 3)'$, $\mathbf{c}_2 = (1, -1, -1, 1)'$, and $\mathbf{c}_3 = (-1, 3, -3, 1)'$ may be used to test for the degree of a polynomial describing the mean response. See Milliken et al. [1] for a data analysis using orthogonal polynomial contrasts.

For the clinical study example above, $q = 2$ contrasts are orthogonal if $\sum_i c_{i1}c_{i2}/n_i = 0$. If the experiment is *balanced*, i.e. if $n_i = n$, $i = 1, 2, 3$, then orthogonality is equivalent to $\mathbf{c}_1'\mathbf{c}_2 = \sum_i c_{i1}c_{i2} = 0$, which states that the contrast coefficient vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ are at right angles. This is true for the contrasts Ψ_1 and Ψ_2 in this example. However, if $n_2 \neq n_3$, then Ψ_1 and Ψ_2 are not orthogonal in this example. If Ψ_2 is defined by $\mathbf{c}_2 = (n_2 + n_3, -n_2, -n_3)'$ instead, then Ψ_1 and Ψ_2 are orthogonal, although it is unlikely that

Ψ_2 is a meaningful comparison in this experiment. Typically, sets of orthogonal contrasts are of more theoretical than practical convenience.

For the two-way treatment structure (3), two contrasts Ψ_1 and Ψ_2 are orthogonal if

$$\sum_i \sum_j \frac{c_{ij1}c_{ij2}}{n_{ij}} = 0. \quad (5)$$

This simplifies to $\sum_i \sum_j c_{ij1}c_{ij2} = 0$ in the equally replicated case, that is, $n_{ij} = n$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. For example, in this latter case of equal replication, two A main effect contrasts are orthogonal if $\sum_i c_{i1}c_{i2} = 0$, where $c_{i1} = c_{ij1}$ and $c_{i2} = c_{ij2}$. A similar statement holds for orthogonality of two B main effects contrasts. Two interaction contrasts are orthogonal if their coefficients satisfy (5).

Sums of Squares from Contrasts

If $\Psi_k = \mathbf{c}_k'\beta$, $k = 1, \ldots, q (1 \le q \le p-1)$ are q linearly independent estimable contrasts, then the sum of squares for testing the **null hypothesis** $H_0 : \mathbf{C}'\beta = \mathbf{d}$, where the kth column of the $(p \times q)$ matrix $\mathbf{C}$ is $\mathbf{c}_k$ and $\mathbf{d}$ is a $(q \times 1)$ vector of constants (usually $\mathbf{d} = \mathbf{0}$) is [1]

$$\mathrm{SSH}_0 = (\mathbf{C}'\hat{\beta} - \mathbf{d})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}]^{-1}(\mathbf{C}'\hat{\beta} - \mathbf{d}). \quad (6)$$

This sum of squares has q df. If the errors ε_i are iid $N(0, \sigma^2)$, then SSH_0 has a $\chi^2(q)$ distribution (central under H_0) (*see* **Chi-square Distribution**) and is independent of the error mean square [4].

For example, in the one-way ANOVA model for a single treatment factor with p levels, the contrasts $\Psi_i = \mu_i - \mu_p = \tau_i - \tau_p$ $(i = 1, \ldots, p-1)$ are linearly independent and estimable. Their sum of squares given by (6), with $\mathbf{d} = \mathbf{0}$, reproduces the treatment/model sum of squares reported in an ANOVA table. See the *contrast* option to the *repeated* statement in the SAS [2] procedure GLM for the polynomial, Helmert, mean, and profile sets of linearly independent, estimable contrasts that achieve the same result. Alternatively, a set of $p-1$ orthogonal contrasts may be chosen instead, with the added benefit that the sum of squares for each individual contrast will be uncorrelated with the others and the $p-1$ individual sums of squares will add to the treatment/model sum of squares.

Generating ANOVA table sums of squares for main effects and interactions in multiway treatment structures is more complicated and depends on the replication of each treatment combination. SAS [2] procedure GLM defines and computes the sum of squares for four *types* of contrasts, types I–IV, for each class of effects (main effects, two-factor interactions, etc.). Type IV contrasts/sums of squares are relevant only when one or more treatment combinations is not observed [1] and will not be discussed further here. Types I–II contrasts have coefficients that depend on the replication numbers n_{ij} and are, in general, not meaningful in unbalanced experiments. For example, the A main effect type I contrasts (for the usual order of A, B, and AB effects classes) are

$$\frac{1}{n_{i\cdot}}\sum_j n_{ij}\mu_{ij} - \frac{1}{n_{I\cdot}}\sum_j n_{Ij}\mu_{Ij} = 0,$$
$$i = 1, \ldots, I-1.$$

Note that these contrasts do not satisfy the more restrictive definition for main effects contrasts given above; also note that they have coefficients that are functions of the amount of data collected (for details, see Milliken et al. [1]).

Type III contrasts and their sums of squares are appropriate for most experiments, whether the experiment is balanced or not. The type III contrasts for each class of effects result from equal averaging over the levels of all factors that are not part of the class of effects of interest. For treatment structure (3), type III A main effects contrasts are $\overline{\mu}_{i\cdot} - \overline{\mu}_{I\cdot} = 0$, $(i = 1, \ldots, I-1)$, or *any* contrast formed from a linear combination of these $I-1$ contrasts. Similarly, type III B main effect contrasts are any linear combinations of $\overline{\mu}_{\cdot j} - \overline{\mu}_{\cdot J} = 0$ $(j = 1, \ldots, J-1)$. The AB interaction contrasts are $\mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'}$, $(i \ne i', j \ne j')$. These type III contrasts in the population marginal means extend to higher-factor interactions in an analogous manner.

If each treatment combination is equally replicated, so $n_{ij} = n$, then type I–IV contrasts and sums of squares are the same for each class of effects. Under normality, the sums of squares for each class of effects have a χ^2 distribution and they are mutually independent [4].

Extensions

Contrasts may represent meaningful comparisons of interest in experiments analyzed by models other than the linear model discussed here. For example, **generalized linear models** (GLMs), which include the normal-theory linear models used above with the identity link between mean response and the linear predictors, are appropriate for a wide variety of non-Gaussian error distributions. If a GLM model contains treatment factors, then contrasts among the levels of these factors may be estimated and tested. For example, with a dichotomous response, a **logistic regression** (logit analysis) may be used. If a factor has two levels (e.g. gender), then a contrast in the parameters between the two levels of this factor would represent the log-**odds ratio** between the two genders.

Contrasts among fixed factor parameters may be estimated in *mixed models* that have two or more **random effects** (e.g. **split plot designs**, repeated measures). In mixed models, unbiased estimators of contrasts are typically available. However, estimates of the variances of such contrasts usually require iterative computation, may be at best approximate (and biased, especially in unbalanced experiments), and may have nonstandard distributions even when all errors are Gaussian. Milliken et al. [1] provide examples for mixed ANOVA models.

References

[1] Milliken, G.A. & Johnson, D.E. (1992). *Analysis of Messy Data*, Vol. 1. Chapman & Hall, New York.
[2] SAS Institute Inc. (1990). *SAS/STAT User's Guide, Version 6*, 4th Ed. Vol. 2. SAS Institute Inc., Cary.
[3] Scheffe, H. (1959). *The Analysis of Variance*. Wiley, London.
[4] Seber, G.A.F. (1977). *Linear Regression Analysis*. Wiley, New York.

D. COSTER

Controlled Clinical Trials

The need for the Journal developed from the emergence of the controlled trial as the ultimate test of a treatment touted for use in human beings (*see* **Clinical Trials, Overview**). The work of **Sir Ronald A. Fisher** [9], in the 1930s and 1940s, and the teachings and writings of **Bradford Hill** [11] in the 1940s and 1950s, along with the **Medical Research Council (MRC)** of the UK, calling in 1931 for trials of new remedies [17], created the basis for modern-day trials. That basis was strengthened, following World War II, by unprecedented expansion of Federal funding for medical research in the US, starting in the 1950s and continuing into the 1980s. The US Congress, concerned with the safety and efficacy of drugs approved for use on human beings, acted to strengthen the **Food and Drug Administration** and to set approval standards based on evidence of efficacy as provided by "adequate and well-controlled trials" [36]. As a result, the clinical trial came to be seen as the "indispensable ordeal" for evaluating new and old treatments, even if protracting the "moment of truth to excruciating limits" [10].

The resources and resolve of the 1960s led to the emergence of **National Institutes of Health** sponsored long-term, **multicenter trials** designed to investigate treatments for chronic diseases (*see* **Clinical Trials, Early Cancer and Heart Disease**; **Cooperative Cancer Trials**; **Cooperative Heart Disease Trials**). In order to enroll and treat the numbers of patients necessary to evaluate therapies for treatment or prevention of these diseases, the trials sometimes involved thousands of patients and a decade or more to complete. One of the first trials in this class was the **University Group Diabetes Program (UGDP)**, started in 1960 and aimed at answering the question of whether drug control of blood glucose levels in persons with adult-onset diabetes is useful [34, 35]. It was followed in the mid-1960s by the Coronary Drug Project (CDP) [8]; and thereafter by a series of primary and secondary prevention heart disease trials [2, 3, 6, 7, 12, 13, 15, 16, 24, 25]. Although there was record growth in the numbers of these multicenter trials, there was little or no interchange among the different centers funded to coordinate the trials and to receive, process, and analyze the data generated by them.

Efforts to improve communication and to encourage the exchange of information on methods and procedures for designing and conducting large-scale multicenter trials commenced with a meeting of representatives from several coordinating centers held in Columbia, Maryland, in 1973. That meeting was followed by a second one in 1975 and thereafter by annual meetings through 1981 [4, 5]. The format of the meetings changed from "show and tell" descriptions of processes or procedures, to one reminiscent of meetings of professional societies.

The interchanges helped to underscore the fact that there is both an art and a science to coordination and data processing and to the realization that there was no obvious "home" for papers having to do with the methods and science of the design and conduct of trials. That realization gave rise to the urge to create the means for receiving and publishing such papers. The call for a journal issued from a workshop entitled "National Conference on Clinical Trials Methodology", held at the NIH on October 3–4, 1977 [14, 23].

Following the workshop, representatives from Elsevier [1, 18] approached Robert Gordon, then Director of the NIH Clinical Trials Committee, for advice as to persons they might approach to head such a journal. They were directed to one of us (C.L.M.). The conversations regarding assumption of the editorship took place in the summer of 1979. The first issue was published in May of 1980, followed by three other issues in that year and by four issues thereafter through 1989. The Journal has been published six times yearly since then [19–22, 26].

The meetings of coordinating centers that started in 1973 and the NIH workshop in 1977 gave rise, as well, to the Society for Clinical Trials. It emerged from a working group created following the 1977 NIH workshop [14, 23, 29] and was chartered on October 5, 1978, in the State of Maryland. The last two meetings of coordinating centers (Annual Symposium on Coordinating Clinical Trials) were, in fact, held in conjunction with the first two meetings of the Society for Clinical Trials. The first such joint meeting coincided with the Seventh Annual Symposium on Coordinating Clinical Trials and the first meeting of the Society for Clinical Trials in Philadelphia in 1980 [27, 28, 30–33]. The Society adopted the Journal as its official organ by virtue of a contract, dated July 19, 1979, between the Publisher, the Society, and, the Editor-in-Chief of the Journal.

The Journal, as reflected in its aims and scope, is intended to attract and publish papers having to do with issues related to the design, conduct, organization or analysis of trials. It is not a traditional results journal, as such. It has, over the years, published papers dealing with the basic design and operating features of trials and in some cases has provided detailed descriptive information as to the baseline characteristics of study population of trials. It has published several monographs with papers related to a particular trial and on topics such as data processing or recruitment. Some of the pages of the Journal are devoted to commentary, letters to the editor, book reviews, and software reviews. The **Society for Clinical Trials** provided editors for *CCT* from its inception in 1980: Curtis Meinert (1980–1993), Janet Wittes (1994–1998), and Jim Neaton (1999–2003). The SCT terminated its relationship with the journal in 2004, founding a new journal, ***Clinical Trials***. The current editor of *CCT* is Kathleen B. Drennan.

References

[1] Altman, Y. (1977). *Elsevier Letter to Various People Regarding Possible Clinical Trials Journal*, Personal Communication, October 28.

[2] Aspirin Myocardial Infarction Study Research Group (1980). A randomized, controlled trial of aspirin in persons recovered from myocardial infarction, *Journal of the American Medical Association* **243**, 661–669.

[3] Aspirin Myocardial Infarction Study Research Group (1980). *Aspirin Myocardial Infarction Study: Design, Methods, and Baseline Results*, Publ. No. 80–2106, National Heart, Lung, and Blood Institute, Bethesda.

[4] Coordinating Center Models Project Research Group (1979). *Coordinating Center Models Project: a Study of Coordinating Centers in Multicenter Clinical Trials: I. Design and Methods* (in two parts), Division of Heart and Vascular Diseases, National Heart, Lung, and Blood Institute, Bethesda, Maryland, March.

[5] Coordinating Center Models Project Research Group (1979). *Coordinating Center Models Project: a Study of Coordinating Centers in Multicenter Clinical Trials: XVI: CCMP Manuscripts Presented at the Annual Symposium on Coordinating Clinical Trials*, Division of Heart and Vascular Diseases, National Heart, Lung, and Blood Institute, Bethesda, Maryland.

[6] Coronary Artery Surgery Study Research Group (1983). Coronary Artery Surgery Study (CASS): a randomized trial of coronary artery bypass surgery: survival data, *Circulation* **68**, 939–950.

[7] Coronary Artery Surgery Study Research Group (1981). National Heart, Lung, and Blood Institute Coronary Artery Surgery Study: a multicenter comparison of the effects of randomized medical and surgical treatment of mildly symptomatic patients with coronary artery disease, and a registry of consecutive patients undergoing coronary angiography (edited by T. Killip, L.D. Fisher & M.B. Mock), *Circulation* **63**, Supplement, I-1–I-81.

[8] Coronary Drug Project Research Group (1973). The Coronary Drug Project: design, methods, and baseline results, *Circulation* **47**, I-1–I-50.

[9] Fisher, R.A. (1946). *Statistical Methods for Research Workers*, 10th Ed. Oliver & Boyd, Edinburgh.

[10] Fredrickson, D.S. (1968). The field trial: some thoughts on the indispensable ordeal, *Bulletin of the New York Academy of Medicine* **44**, 985–993.

[11] Hill, A.B. (1962). *Statistical Methods in Clinical and Preventive Medicine*. Oxford University Press, New York.

[12] Hypertension Detection and Follow-up Program Cooperative Group (1976). The Hypertension Detection and Follow-up Program, *Preventive Medicine* **5**, 207–215.

[13] Hypertension Detection and Follow-up Program Cooperative Group (1977). Blood pressure studies in 14 communities: a two-stage screen for hypertensives, *Journal of the American Medical Association* **237**, 2385–2391.

[14] Kolata, G.B. (1977). Clinical trials: methods and ethics are debated, *Science* **198**, 1127–1131.

[15] Lipid Research Clinics Program (1979). The Coronary Primary Prevention Trial: design and implementation, *Journal of Chronic Diseases* **32**, 609–631.

[16] Lipid Research Clinics Program (1984). The Lipid Research Clinics Coronary Primary Prevention Trial Results, I. Reduction in incidence of coronary heart disease, *Journal of the American Medical Association* **251**, 351–364.

[17] Medical Research Council (1931). Clinical trials of new remedies (annotations), *Lancet* **2**, 304.

[18] Meinert, C.L. (1977). *Prospectus for a Journal on the Methodology of Controlled Clinical Trials*, Personal Communication, October 26.

[19] Meinert, C.L. (1980). Why another journal?, *Controlled Clinical Trials* **1**, 1–2.

[20] Meinert, C.L. (1981). The first year, *Controlled Clinical Trials* **2**, 1.

[21] Meinert, C.L. (1990). The Journal after 10 years, *Controlled Clinical Trials* **11**, 1–3.

[22] Meinert, C.L. (1994). Farewell swan song, *Controlled Clinical Trials* **15**, 235–237.

[23] Meinert, C.L. & Hawkins, B.S. (1979). Methodology: the case for improved communications, *Clinical Pharmacology and Therapeutics* **25**, 754–757.

[24] Multiple Risk Factor Intervention Trial Research Group (1977). Statistical design considerations in the NHLBI Multiple Risk Factor Intervention Trial (MRFIT), *Journal of Chronic Diseases* **30**, 261–275.

[25] Multiple Risk Factor Intervention Trial Research Group (1982). Multiple Risk Factor Intervention trial: risk factor changes and mortality results, *Journal of the American Medical Association* **248**, 1465–1477.

[26] Resource Committee of the Coordinating Center Models Project (1979). *Minutes of the New Orleans Meeting*, New Orleans, March.

[27] Roth, H.P. (1979). *Memo to Prospective Participants in the Scientific Sessions of the Society for Clinical Trials and Seventh Annual Symposium on Coordinating Clinical Trials*, Personal Communication, October 26.

[28] Roth, H.P. (1980). On the Society for Clinical Trials, *Controlled Clinical Trials* **1**, 81–82.

[29] Schimmel, J. (1978). *Memo to Registrants – National Conference on Clinical Trials Methodology*, regarding Need for Society, Journal, Annual Meeting.

[30] Society for Clinical Trials (1980). Abstracts of the Combined Annual Scientific Sessions of the Society for Clinical Trials and the Seventh Annual Symposium for Coordinating Clinical Trials, *Controlled Clinical Trials* **1**, 167–180.

[31] Society for Clinical Trials (1980). Society for Clinical Trials, Inc.: bylaws, *Controlled Clinical Trials* **1**, 83–89.

[32] Society for Clinical Trials (1981). Abstracts of the Combined Annual Scientific Sessions of the Society for Clinical Trials and the Eighth Annual Symposium for Coordinating Clinical Trials, *Controlled Clinical Trials* **2**, 67–89.

[33] Society for Clinical Trials (1981). Preliminary Program of the Combined Annual Scientific Sessions of the Society for Clinical Trials and the Eighth Annual Symposium for Coordinating Clinical Trials, *Controlled Clinical Trials* **2**, 59–65.

[34] University Group Diabetes Program Research Group (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: I. Design, methods, and baseline characteristics, *Diabetes* **19**, 747–783.

[35] University Group Diabetes Program Research Group (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: II. Mortality results, *Diabetes* **19**, 785–830.

[36] US Congress (1962). *Public Law No. 87–781: Drug Amendments of 1962*, 76 Stat. 780.

CURTIS L. MEINERT & SUSAN TONASCIA

Controls

Controls are subjects against whom a comparison is made in experimental or **observational studies**. In randomized **clinical trials**, the control group may receive no treatment, a placebo treatment, or the best currently accepted active treatment. This group is used as a basis of comparison against the group that receives the new experimental treatment. Sometimes the study subjects are stratified into risk groups and are allocated into experimental or control groups in such as way as to assure roughly equal numbers of subjects in the experimental and control groups within each stratum (*see* **Randomized Treatment Assignment**). In observational **cohort studies**, a comparison may be drawn with unexposed or less exposed members of the cohort ("internal" controls). If all members of a cohort are exposed, however, as in some studies of occupational risk (*see* **Occupational Epidemiology**), then an external unexposed or only slightly exposed control population, such as the general population of the US, may be taken as a basis of comparison (*see* **Standardization Methods**).

Controls chosen for comparison with cases in **case–control studies** may be selected from the general population (*see* **Case–Control Study, Population-based**; **Case–Control Study, Prevalent**) or from a selected source, as in **hospital-based case–control studies**. Controls may also be matched to cases on characteristics that may **confound** the **association** between exposure and disease status (*see* **Matching**).

MITCHELL H. GAIL

Convergence in Distribution and in Probability

To motivate convergence in distribution and in probability, suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample of size n taken from a population with unknown mean μ and variance σ^2. The sample mean $\overline{Y}_n = (Y_1 + \cdots + Y_n)/n$ is typically used to estimate μ. As the sample size n grows to infinity, does $\overline{Y}_n$ converge to μ (if so, in what sense) and what can one say about the deviations of $\overline{Y}_n$ from μ? It turns out that $\overline{Y}_n$ converges to μ in probability (in fact, even "almost surely") and $\sqrt{n}(\overline{Y}_n - \mu)/\sigma$ converges in distribution to a Gaussian random variable with mean 0 and variance 1. We shall now define these two modes of convergence.

Let $X_n, n = 1, 2, \ldots$ be a sequence of **random variables** (for example, the sample means $\overline{Y}_n$ considered earlier). The cumulative distribution function $F_{X_n}(x) = \Pr[X_n \le x], -\infty < x < \infty, n = 1, 2, \ldots$ describes the probability law of X_n. Let Z be another random variable with cumulative distribution function $F_Z(x) = \Pr[Z \le x], -\infty < x < \infty$.

Definition. The random variables X_n converge to Z in distribution as n tends to infinity, if the functions $F_{X_n}(x)$ converge to $F_Z(x)$ for all $-\infty < x < \infty$ that are points of continuity of F_Z.

To understand why we do not require that $F_{X_n}(x)$ converge to $F_Z(x)$ for all x, suppose that X_n equals $1/n$ and hence is not random. As n tends to infinity, $1/n$ tends to 0 and hence we expect the distribution of X_n to converge to that of Z, where Z is also not random and equals 0. Note that $F_{X_n}(x) = 0$ if $x < 1/n$ and $F_{X_n}(x) = 1$ if $x \ge 1/n$. Similarly, $F_Z(x) = 0$ if $x < 0$ and $F_Z(x) = 1$ if $x \ge 0$. It is easy to see that $F_{X_n}(x) \to F_Z(x)$ for all $x \ne 0$ but not at $x = 0$, because $F_{X_n}(0) = 0$, whereas $F_Z(0) = 1$. It is therefore not natural to require that F_{X_n} converge to F_Z at points of discontinuity of F_Z.

Let us now turn to convergence in probability.

Definition. The random variables X_n converge to Z in probability as n tends to infinity if, for all $\varepsilon > 0$, $\lim_{n\to\infty} \Pr[|X_n - Z| > \varepsilon] = 0$.

In other words, X_n converges to Z in probability if the probability that X_n differs from Z by any given small quantity tends to zero as n tends to infinity.

Convergence in probability always implies convergence in distribution, and if Z is nonrandom, then the two modes of convergence are equivalent.

Technically, since convergence in probability involves a joint probability statement about X_n and Z, the random variables X_n and Z have to be defined in the same probability space. This is not necessary if X_n merely converges to Z in distribution.

In practice, to verify convergence in probability, one uses Chebychev's inequality:

$$\Pr[|X_n - Z| > \varepsilon] \le \frac{\operatorname{var}(X_n - Z)}{\varepsilon^2}.$$

Thus X_n converges to Z in probability if X_n and Z have the same mean and $\lim_{n\to\infty} \operatorname{var}(X_n - Z) = 0$.

Returning to the example given at the beginning of the article, to verify that the sample mean $\overline{Y}_n$ converges to the population mean μ in probability, it is sufficient to check that $\lim_{n\to\infty} \operatorname{var}(\overline{Y}_n - \mu) = 0$. This relation holds because

$$\operatorname{var}(\overline{Y}_n - \mu) = \operatorname{var}\overline{Y}_n = \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n} Y_i\right)$$
$$= \frac{1}{n^2} n\sigma^2 \to 0 \quad \text{as } n \to \infty.$$

The convergence in distribution of $\sqrt{n}(\overline{Y}_n - \mu)/\sigma$ to a Gaussian random variable is a consequence of the **central limit theory**. While $\overline{Y}_n - \mu \to 0$ (in probability), multiplying $\sqrt{n}/\sigma$ by $(\overline{Y}_n - \mu)$ yields an $\infty \times 0$ situation. The result (not a priori obvious) is that the limit is a well-defined random variable the range of which is $(-\infty, \infty)$ and the density function has the Gaussian bell-shaped curve centered at zero; that is, has the normal distribution N(0, 1).

Large-sample properties of estimators typically involve convergence in distribution (*see* **Large-sample Theory**). Hence, referring again to the example given at the beginning of this article, for large n, the distribution of $\sqrt{n}(\overline{Y}_n - \mu)/\sigma$ is approximately N(0,1) and that of $\sqrt{n}(\overline{Y}_n - \mu)/s$, where $s^2 = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2/(n-1)$ is the sample variance, is approximately a **Student's *t* distribution** with $n - 1$ degrees of freedom. Similarly, the distribution of $(n-1)s^2/\sigma^2$ is approximately χ^2 (**chi-square distribution**) with $n - 1$ degrees of freedom. These limiting distributions are used to compute approximate confidence intervals for μ and σ^2, respectively.

Convergence in probability, on the other hand, is often used to show that the sample **moment** converges to the population moment. Thus, as n tends to infinity, $\overline{Y}_n \to \mu$ and $s^2 \to \sigma^2$, both in probability; that is, the sample mean converges to the population mean and the sample variance converges to the population variance.

Here are some additional facts that are useful to establish convergence results:

1. If $X_n \to Z$ in distribution and $X_n - W_n \to 0$ in probability, then $W_n \to Z$ in distribution.
2. $X_n \to Z$ in distribution if its **characteristic function** $\phi_{X_n}(t) = \mathrm{E}[\exp(\mathrm{i}tX_n)]$, where $\mathrm{i} = \sqrt{-1}$, tends, for all $-\infty < t < \infty$, to the characteristic function $\phi_Z(t) = \mathrm{E}[\exp(\mathrm{i}tZ)]$ of Z.
3. If X_n and Z are random vectors, then $X_n \to Z$ in distribution if all linear combinations of the components of X_n tend in distribution to the corresponding linear combination of the components of Z.

(*See also* **Limit Theorems**)

M.S. TAQQU

Cooperative Cancer Trials

The last four decades have witnessed a remarkable symbiotic relationship between clinical **oncology** and biostatistics. Research oncologists have become increasingly sophisticated with regard to statistical concepts and tools, and many are now quite familiar with terms such as **Mantel–Haenszel logrank test**, Gehan–Wilcoxon statistic, O'Brien and Fleming boundaries, and Simon optimal Phase II design, all of which relate to statistical techniques developed by biostatisticians working in cancer research. The magnitude of the cancer challenge is enormous. One-third of Americans will develop cancer of at least one of the 65 specific types; 1.4 million new cases and 550 000 deaths occurred in 1996. However, after observing persistent rises in age-adjusted mortality ever since statistics began to be maintained in the 1930s, there has been a drop of 2%–4% over the last 5 years [20]. Biostatisticians have played a crucial role in defining, refining, and applying the methodology of cancer **clinical trials**. In this article we emphasize the development, status, and achievements of the US clinical trials cooperative group program over the last decade (approximately 1986–1996), and we will briefly review some of the important methodological advances that have been made in biostatistics, in the same time period, in response to the challenges posed by these trials [42]. (*See* **Clinical Trials, Early Cancer and Heart Disease** for developments from approximately 1955 to 1965.)

The clinical trials cooperative group program grew out of the Cancer Chemotherapy National Service Center, established in 1955 [18] (*see* **Multicenter Trials**). By 1960, there were 11 cooperative oncology groups, with 10–20 new agents taken to clinical trial per year, of the 25 000–30 000 agents screened per year in the laboratory, in tissue culture and in animals. This pattern has persisted, with approximately 20 000 agents per year screened currently in the National Cancer Institute (NCI) *in vitro* human tumor cell line assay, and with 20–25 per year taken to clinical trial. Also by 1960, the current clinical developmental pathway had been defined: **Phase I trials** (of approximately 20 patients) to determine the appropriate dose of an agent, followed by **Phase II trials** (of 30–50 patients) to provide an initial test of its ability to shrink tumors, followed by randomized Phase III trials to test its efficacy, usually defined in terms of ability, alone or in combination, to prolong survival, compared with a control therapy. As they do now, the cooperative groups in 1960 collaboratively accrued to standardized **clinical trials protocols**, with standardized criteria of diagnosis, treatment, and measurement of effect, with **randomization** (that eventually became centralized), and with a prospective statistical design and collaborative analysis and reporting. The early development of the statistical centers was led by NCI statistician **Marvin Schneiderman** [18].

In the mid 1960s the focus of cooperative group cancer trials was expanded from new agent testing to encompass the development of new disease-oriented combination chemotherapy regimens. In 1972, the trials program formally entered a period of further expansion with the establishment by Congress of the National Cancer Program. In particular, patients with early stage disease were included in larger numbers (up to 1974, at least 85% of trial patients had advanced disease). By 1979, the configuration of the cooperative groups approximated the current one. Of the 31 groups that had been established since 1955, 14 remained under NCI sponsorship. A proliferation of small regional and single-disease groups had given way to primarily national multidisease groups [6], with correspondingly larger statistical centers, made up of statisticians and data managers under the supervision of a "group statistician" (Table 1 lists the extant cooperative groups). Statisticians played an increasingly prominent part in the groups. In addition to their lead role in the monitoring, analysis and reporting of group studies, the statisticians played a crucial role in their design.

Table 1 US cooperative cancer groups

Group	Acronym
Cancer and Leukemia Group B	CALGB
Children's Cancer Group	CCG
Eastern Cooperative Oncology Group	ECOG
Gynecologic Oncology Group	GOG
Intergroup Rhabdomyosarcoma Study Group	IRSG
National Surgical Adjuvant Breast and Bowel Project	NSABP
National Wilms' Tumor Study Group	NWTSG
North Central Cancer Treatment Group	NCCTG
Pediatric Oncology Group	POG
Radiation Therapy Oncology Group	RTOG
Southwest Oncology Group	SWOG

By the mid 1970s, NCI Biometric Research Branch (BRB) statisticians, then and now under the leadership of Richard Simon, had taken over statistical oversight of NCI-sponsored cancer treatment trials. The design, monitoring, and analysis of the trials was conducted at the statistical centers, with the exception of some of the prostate, central nervous system, and lung cancer trials, with statisticians in the NCI Clinical Trials Section under the leadership of **David Byar**. Between 1979 and 1982 the cooperative groups' grants and contracts were replaced by "cooperative agreements", which involved an expanded role for the NCI staff, in particular, for the BRB. By the early 1980s, as now, BRB statisticians were prospectively reviewing the design and conduct of all treatment studies and were acting as liaisons to the group statistical centers. In addition, they were deeply involved in the conception of drug development and disease-oriented strategies and in the design of many particular phase III studies. NCI and cooperative group statistical center statisticians have been extensively involved in the development of new statistical methodology concerning the design, conduct, and analysis of clinical studies; these contributions are discussed below.

The number of cooperative groups increased to 18 by 1985 and then decreased, from 1985 to 1988, to the 11 groups currently in existence. In the same period, total accrual and total number of trials for the cooperative group program increased by 15%, as emphasis was placed on increasing accrual through enhanced involvement of community hospitals and increased enrollment of minority patients [16, 17].

At this time, there was also substantial emphasis on studies to correlate potentially prognostic laboratory measures with clinical results and on translation of the rapidly emerging basic science findings to clinical advances (at present approximately 50% of treatment trials have explicit correlative science objectives). At the heart of this were the cancer centers, primarily large academic medical centers where advanced multidisciplinary basic, clinical, and translational research could be conducted. The cancer centers program developed in the early 1960s, with 12 institutions and a combined budget of 6 million dollars. The program was formally established through the National Cancer Act of 1971, and in 1996 there were 55 centers with a combined budget of 150 million dollars. The cancer centers participate in cooperative clinical trials through membership in the cooperative groups (and also conduct studies independently).

As a result of the community outreach efforts of the mid 1980s and later, 80% of cooperative group patients are currently treated outside of university and cancer center hospitals, through the Cooperative Group Outreach Program (CGOP) and the Community Clinical Oncology Program (CCOP). Approximately 50% of patients are accrued through CGOP, which was started in 1976 and involves small hospitals and practices that enter patients through cooperative group member institutions [7]. Approximately 30% of patients are accrued through CCOP, which was started in 1983 and involves larger hospitals that may independently affiliate with more than one cooperative group [7]. Studies have shown little or no differences in numbers of ineligible patients or protocol violations between academic and community hospitals [16].

The magnitude, structure, and cost of the US clinical trials cooperative group program have been relatively stable over the last decade. Of the 38 groups that have been established since 1955, the present 11 groups (Table 1) are those that constituted the program in 1988. These include seven adult patient groups – three multidisease national groups (CALGB, ECOG, SWOG), one regional group (NCCTG), one radiotherapy group (RTOG), one surgical adjuvant group (NSABP), and one gynecological cancer group (GOG) – and four pediatric patient groups – two multidisease national groups (CCG, POG) and two single-disease intergroup mechanisms (IRSG, NWTSG). In 1995, they comprised 6600 investigators at 1500 institutions, entering 16 000 patients on 340 open treatment studies, with a collective budget of 100 million dollars. (Over the past decade, the cooperative groups have accrued an average of 20 000 patients per year.) Also in 1995, they placed approximately 10 000 of these patients on 90 open nontherapeutic studies (ancillary **quality of life** and laboratory studies, associated with the therapeutic studies). To handle this work load, the statistical center of a typical large multidisease group employs approximately 12 statisticians and 20 data managers [6]. Typically, group statistical centers are associated with major university biostatistics departments, such as at Duke (CALGB), Harvard (ECOG), and the University of Washington (SWOG). In addition, the

NCI provides partial support to the data center of the **European Organization for Research and Treatment of Cancer (EORTC)**, under the statistical leadership of Richard Sylvester since 1975, which represents 350 institutions in 31 countries and places approximately 6000 patients per year on studies [46]. Finally, the National Cancer Institute of Canada, under the statistical leadership of Benny Zee since 1987, representing 50 institutions and placing 1500 patients per year on studies, has been very active in collaborating with the NCI-sponsored cooperative groups, as well as in conducting studies independently.

Over the past decade, the efforts of the cooperative groups have been divided roughly as follows with respect to treatment studies [17]. Phase I efforts have varied among the groups, but overall Phase I trials (dose–finding studies) made up only 10% of the total and accounted for only 2%–3% of the patients. Phase I trials have been primarily the domain of the cancer centers, working independently. Phase II trials (initial nonrandomized assessments of clinical benefit) made up 50%–60% of the total, and they accounted for 20%–25% of the patients. Phase III trials (randomized comparisons, usually against a control treatment) made up 35%–40% of the total, but because of their much larger size accounted for 70%–80% of the patients. Definitive Phase III trials are the "jewel in the crown" of the US cooperative group program; in Table 2 we give 15 of the most significant completed in the past decade. Accrual of the major histologic subgroups was approximately as follows [17]: breast cancer patients accounted for 4000 of the total annual accrual, colorectal cancer patients accounted for 2000, and lung cancer, ovarian cancer, central nervous system cancer, and prostate cancer patients accounted for 1000 each.

Pediatric cancer patients have accounted for approximately 5000 of the total annual accrual to treatment studies over the past decade, and this subgroup is remarkable in two respects. First, 95% of pediatric cancer patients in the US are seen at institutions belonging either to CCG or POG [36], with 80% of potentially eligible patients actually placed on a cooperative group protocol [7]. Secondly, there have been striking clinical advances in pediatric cancer, in particular since the advent of the National Cancer Program; for example, 4-year survival in pediatric leukemia on CCG protocols rose from 20% in the late 1960s to 80% in the mid 1980s [2], and dramatic survival increases in this disease and in other pediatric cancers have been seen on POG protocols [38].

Although the magnitude and structure of the clinical trials cooperative group program has been stable over the last decade, there have been two dramatic developments in the way in which Phase III trials are conducted. The first development involves the remarkable increase in the number of "intergroup" (involving more than one cooperative group) Phase III studies, particularly among the adult disease groups, where currently 80%–90% of phase III studies are intergroup efforts (accounting for 80%–90% of Phase III patients). Precise guidelines for the management of these trials were formulated in the early 1990s under the leadership of Eleanor McFadden, at Harvard. The second development involves the formalization of interim monitoring schemes and the establishment of **data and safety monitoring boards**. Up to the late 1970s it was common practice to report annual or semiannual interim outcome analyses of randomized studies to the entire cooperative group. By the mid 1980s the statistical dangers in such a practice were widely recognized, and interim outcome reporting was limited to a trial steering committee. Precise predefined interim monitoring schemes were used, which were developed by O'Brien et al. [15] to allow early termination in the case of dramatic treatment differences (*see* **Data and Safety Monitoring**) and then extended by Wieand & Therneau [49] to also allow early termination in the case of convincing evidence of lack of treatment differences. In the 1990s, under the leadership of NCI staff, in particular BRB statisticians, data safety monitoring committees were established. From the beginning these provided interim monitoring independent of individual trial leadership, and they now include a majority of members from outside the cooperative group itself.

Another dramatic recent development has been the increase in cancer prevention and control trials (*see* **Prevention Trials**), which have been conducted through the clinical trials cooperative groups and the CCOP. Over the last decade 30–40 such NCI-sponsored trials have been active at a given time. These trials tend to be much larger than the treatment trials, and they have their own particular statistical problems arising from low event rates and potentially high noncompliance rates [4] (*see* **Noncompliance, Adjustment for**). The two most

Table 2 Noteworthy US randomized trials of the last decade

Protocol no. Accretion dates Participants	Protocol name	Protocol significance
NSABP-B06 1976–84	A randomized clinical trial comparing total mastectomy vs. lumpectomy with or without irradiation in the treatment of breast cancer [12]	This study provided important data supporting the use of lumpectomy in patients with stage I or II breast cancer, and demonstrated that irradiation reduces the probability of local recurrence of tumor in patients treated with lumpectomy.
CALGB-8251 1982–87	Treatment of advanced Hodgkin's disease: randomized Phase III trial comparing MOPP vs. ABVD vs. MOPP alternating with ABVD [5]	While MOPP had been the standard chemotherapy regimen for advanced stage Hodgkin's disease, this study demonstrated that both ABVD and MOPP/ABVD were superior to MOPP. Additional advantages of ABVD compared to MOPP include lower risk of secondary leukemia and decreased incidence of sterility.
INT-0035 1985–87 ECOG NCCTG SWOG	Intergroup Phase III surgical adjuvant trial for stage B2 and C colon cancer [35]	The early results (the 1990 report) served as the basis for the National Institutes of Health Consensus Panel recommendation of 5-FU/levamisole as standard treatment in the US.
NWTS-3 1979–87 CCG POG	National Wilms' tumor study 3 [9] and subsequent National Wilms' Tumor Study Group trials have utilized the less intensive (yet efficacious) therapies identified in this study.	This study demonstrated that less intensive therapy does not appear to worsen results for low-risk patients,
POG-8314 1983–87	Localized non-Hodgkin's lymphoma: chemotherapy +/− radiotherapy [31]	This study demonstrated that radiotherapy can be safely omitted from the therapy of most children with localized non-Hodgkin's lymphoma without substantially jeopardizing their excellent chance of cure.
NSABP-B13 1981–88	A clinical trial to assess sequential MTX, 5-FU in patients with axillary node- and estrogen receptor + primary breast cancer [14]	This was one of the trials that resulted in the National Cancer Institute issuing a clinical alert in 1989 about the positive role of chemotherapy for selected subsets of women with node-negative breast cancer.
CALGB-8525 1985–90	Phase III comparison of post-remission intensive ara-C in patients with acute myelogenous leukemia in first remission [32]	This study demonstrated the importance of post-remission dose intensity of cytarabine for patients younger than 60 years of age, establishing high-dose cytarabine as standard consolidation therapy for this patient population.

Table 2 (*continued*)

Protocol no. Accretion dates Participants	Protocol name	Protocol significance
GOG-97 1986–90	Phase III randomized study of cyclophosphamide and cisplatin in patients with suboptimal stage III, IV ovarian carcinoma comparing intensive and nonintensive schedules [33]	Although preclinical and clinical data based on historical comparisons suggested that increased dose-intensity might improve outcome for women with ovarian cancer, this study demonstrated that a doubling of the dose-intensity in the treatment of bulky ovarian epithelial cancers led to no discernible improvement in patient outcome and was associated with more severe toxicity.
CALGB-8541 1985–91	Adjuvant CAF for pathologic stage II, node+ breast cancer: randomization among intensive CAF for 4 mo. vs. low-dose CAF for 4 mo. vs. standard-dose CAF for 6 mo. [50]	This study demonstrated a dose–response effect for adjuvant chemotherapy within the standard dose range. An important translational science finding that resulted from the clinical trial was the demonstration that the dose–response effect was limited to patients whose tumors overexpressed c-erbB-2.
INT-0032 1984–91 CCG IRSG POG	Intergroup rhabdomyosarcoma study III [8] Additionally, a subset of patients with favorable prognosis was identified for whom relatively non-toxic two-drug therapy resulted in very good outcome.	Intensification of therapy for most patients in this study, using a risk-based study design, significantly improved treatment outcome overall.
INT-0067 1986–91 ECOG SWOG	Phase III comparison of CHOP vs. M-BACOD vs. ProMACE-CytaBOM vs. MACOP-B in patients with intermediate or high grade non-Hodgkin's lymphoma [13]	Although single-institution, uncontrolled studies had suggested an advantage for third-generation regimens, this study demonstrated that CHOP remains the best available treatment for patients with advanced-stage intermediate-grade or high-grade non-Hodgkin's lymphoma, with decreased toxicity and expense compared to the third-generation regimens.
INT-R8501 1986–91 NCCTG RTOG SWOG	Phase III prospective trial for localized cancer of the esophagus comparing radiation as a single modality with radiation therapy plus chemotherapy [26]	This trial demonstrated an unequivocal survival advantage for combined modality therapy, specifically the addition of chemotherapy to radiation-based (nonsurgical) approaches in this rarely curable cancer. Studies to optimize combined modality therapy have followed this lead.

(*continued overleaf*)

Table 2 (*continued*)

Protocol no. Accretion dates Participants	Protocol name	Protocol significance
GOG-111 1990–92	A Phase III randomized study of cyclophosphamide and cisplatin vs. paclitaxel and cisplatin in patients with suboptimal stage III and IV epithelial ovarian cancer [34]	This study was the first phase III trial of paclitaxel in the primary treatment of patients with cancer. It demonstrated that incorporating paclitaxel into first-line therapy improves both the duration of progression-free survival and overall survival in women with incompletely resected stage III and IV ovarian cancer.
RTOG-8808 1989–92 ECOG SWOG	Phase III study of radiation therapy alone or in combination with chemotherapy for patients with non-small cell lung cancer [40]	Patients with regionally advanced, surgically unresectable non-small cell lung cancer have a poor prognosis, and the standard therapy had been external beam irradiation to the primary tumor and regional lymphatics. The results from this trial confirmed the survival benefits and acceptable toxicity seen in a smaller randomized trial, and the standard of care for this patient population now includes a combination of radiotherapy and chemotherapy.
CCG-2891 1989–95	Treatment of children <21 with newly diagnosed acute myeloid leukemia and myelodysplastic syndrome [51]	This study is important for demonstrating that although early intensive therapy is more toxic than standard timing therapy, eventual outcome is improved by use of the timing intensive strategy.

noteworthy studies are the Breast Cancer Prevention Trial, conducted by NSABP and accruing 13 000 women from 1992 to 1997, which tests the ability of tamoxifen to prevent breast cancer in a high-risk population, and the Prostate Cancer Prevention Trial, conducted by SWOG and accruing 18 000 elderly men from 1993 to 1996, which tests the ability of dihydrotestosterone to prevent or delay prostate cancer [21].

Finally, we summarize some of the important developments in clinical trials methodology supported by the NCI in the last decade. Simon [42] has provided an excellent review of advances in clinical trials methodology in the 1980s, in which many of the NCI and cooperative group statisticians appear prominently, and we try to minimize overlap with this review. We organize the contributions into nine areas:

1. New Phase I trial designs: traditionally, phase I trial designs have used cohorts of three to six patients, treated at escalating doses of a new chemotherapeutic agent, to identify a maximum tolerated dose (MTD). New designs, with their associated methods for estimating the MTD, have been proposed by Storer [45], O'Quigley et al. [37], and Goodman et al. [22], and have had their statistical properties reviewed by Korn et al. [28].
2. New Phase II trial designs: Simon [41] proposed and studied two-stage Phase II designs that are optimal in the sense of minimizing the expected number of patients exposed to ineffective therapy.
3. Flexible interim monitoring rules: Lan & DeMets [29] further developed the "spending function" approach, which allows flexibility in the timing of formal interim analyses.

4. Stopping early for "negative" results: Thall et al. [47] and Wieand et al. [48] developed simple rules to allow early stopping if, with high probability, the treatment difference would not be statistically significant were the trial to continue to its originally planned end.
5. **Bayesian methods** in randomized trials: Breslow [3], Dixon & Simon [10], Spiegelhalter et al. [43], and Gray [24] used Bayesian methods to address some of the difficulties standard frequentist methods have with combining information from diverse sources for use in **sample size** calculations, inference after early stopping of trials, and **multiple comparison** problems.
6. **Surrogate endpoints**: Prentice [39] elucidated the statistical conditions necessary for an intermediate outcome to be a valid surrogate, thereby saving trial time or avoiding **confounding** because of treatment cross-over prior to the final outcome.
7. Innovative methods for identifying **prognostic factors**: Durrleman & Simon [11], LeBlanc & Crowley [30], and Gray [25] developed new methods to identify variables that may predict survival, including **spline** smoothing of variables, **quantile regression** and **tree-structured statistical methods**.
8. **Quality-of-life** endpoints: Gelber et al. [19] and Korn [27] proposed and studied new methods for analyzing quality-of-life endpoints, which have assumed increasing importance in clinical trials.
9. **Cure models**: Gray & Tsiatis [23] and Sposto et al. [44] developed and applied models for the case where treatment is anticipated to cure disease in addition to increasing survival for those not cured.

References

[1] Alberts, D.S. & Garcia, D.J. (1995). An overview of clinical cancer chemoprevention studies with emphasis on positive Phase III studies, *Journal of Nutrition* **125**, 692S–697S.

[2] Bleyer, W.A. (1990). Acute lymphoblastic leukemia in children: Advances and prospectus, *Cancer* **65**, 689–695.

[3] Breslow, N. (1990). Biostatistics and Bayes, *Statistical Science* **5**, 269–298.

[4] Byar, D.P. & Freedman, L.S. (1990). The importance and nature of cancer prevention trials, *Seminars in Oncology* **17**, 413–424.

[5] Canellos, G.P., Anderson, J.R., Propert, K.J., Nissen, N., Cooper, M.R., Henderson, E.S., Green, M.R., Gottlieb, A. & Peterson, B.A. (1992). Chemotherapy of advanced Hodgkin's disease with MOPP, ABVD, or MOPP alternating with ABVD, *New England Journal of Medicine* **327**, 1478–1484.

[6] Carbone, P.P. & Tormey, D.C. (1991). Organizing multicenter trials: lessons from the cooperative oncology groups, *Preventive Medicine* **20**, 162–169.

[7] Cheson, B.D. (1991). Clinical trials programs, *Seminars in Oncology Nursing* **7**, 235–242.

[8] Crist, W., Gehan, E.A., Ragab, A.H., Dickman, P.S., Donaldson, S.S., Fryer, C., Hammond, D., Hays, D.M., Herrmann, J., Heyn, R., Jones, P.M., Lawrence, W., Newton, W., Ortega, J., Raney, B., Ruymann, F.B., Tefft, M., Webber, B., Wiener, E., Wharam, M., Vietti, T. & Maurer, H.M. (1995). The Third Intergroup Rhabdomyosarcoma Study, *Journal of Clinical Oncology* **13**, 610–630.

[9] D'Angio, G., Breslow, N., Beckwith, J.B., Evans, A., Baum, E., de Lorimier, A., Fernbach, D., Hrabovsky, E., Jones, B., Kelalis, P., Othersen, B., Tefft, M. & Thomas, P.R.M. (1989). Treatment of Wilms' Tumor. Results of the Third National Wilms' Tumor Study, *Cancer* **64**, 349–360.

[10] Dixon, D.O. & Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial, *Statistics in Medicine* **11**, 13–22.

[11] Durrleman, S. & Simon, R. (1989). Flexible regression models with cubic splines, *Statistics in Medicine* **8**, 551–561.

[12] Fisher, B., Anderson, S., Redmond, C.K., Wolmark, N., Wickerham, D.L. & Cronin, W.M. (1995). Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer, *New England Journal of Medicine* **333**, 1456–1461.

[13] Fisher, R.I., Gaynor, E.R., Dahlberg, S., Oken, M.M., Grogan, T.M., Mize, E.M., Glick, J.H., Coltman, C.A., Jr & Miller, T.P. (1993). Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma, *New England Journal of Medicine* **328**, 1002–1006.

[14] Fisher, B., Redmond, C., Dimitrov, N.V., Bowman, D., Legault-Poisson, S., Wickerham, L., Wolmark, N., Fisher, E.R., Margolese, R., Sutherland, C., Glass, A., Foster, R. & Caplan, R. (1989). A randomized clinical trial evaluating sequential methotrexate and fluorouracil in the treatment of patients with node-negative breast cancer who have estrogen-receptor-negative tumors, *New England Journal of Medicine* **320**, 473–478.

[15] Fleming, T., Harrington, D. & O'Brien, P. (1984). Designs for group sequential tests, *Controlled Clinical Trials* **5**, 348–361.

[16] Frelick, R.W. (1994). The Community Clinical Oncology Program (CCOP) story: review of community oncologists' experiences with clinical research trials in cancer

with an emphasis on the CCOP of the National Cancer Institute between 1982 and 1987, *Journal of Clinical Oncology* **12**, 1718–1723.

[17] Friedman, M.A. & Cain, D.F. (1990). National Cancer Institute sponsored cooperative clinical trials, *Cancer* **65**, 2376S–2382S.

[18] Gehan, E.A. & Schneiderman, M.A. (1990). Historical and methodological developments in clinical trials at the National Cancer Institute, *Statistics in Medicine* **9**, 871–880.

[19] Gelber, R.D., Gelman, R.S. & Goldhirsch, A. (1989). A quality-of-life oriented endpoint for comparing therapies, *Biometrics* **45**, 781–796.

[20] Goldberg, B.K. & Goldberg, P. (1996). Cancer mortality rate declining in 1990's, NCI, ACS say, *Cancer Letter* **22**, 5.

[21] Goldberg, B.K. & Goldberg, P. (1996). Enrollment complete in PCPT, three years after beginning, *Cancer Letter* **22**, 6–7.

[22] Goodman, S.N., Zahurak, M.L. & Piantadosi, S. (1995). Some practical improvements in the continual reassessment method for phase I studies, *Statistics in Medicine* **14**, 1149–1161.

[23] Gray, R.J. & Tsiatis, A.A. (1989). A linear rank test for use when the main interest is in differences in cure rate, *Biometrics* **45**, 899–904.

[24] Gray, R.J. (1994). A Bayesian analysis of institutional effects in a multicenter cancer clinical trial, *Biometrics* **50**, 244–253.

[25] Gray, R.J. (1994). Spline-based tests in survival analysis, *Biometrics* **50**, 640–652.

[26] Herskovic, A., Martz, K., al-Sarraf, M., Leichman, L., Brindle, J., Vaitkevicius, V., Cooper, J., Byhardt, R., Davis, L. & Emami, B. (1992). Combined chemotherapy and radiotherapy compared with radiotherapy alone in patients with cancer of the esophagus, *New England Journal of Medicine* **326**, 1593–1598.

[27] Korn, E.L. (1993). On estimating the distribution function for quality of life in cancer clinical trials, *Biometrika* **80**, 535–542.

[28] Korn, E.L., Midthune, D., Chen, T.T., Rubinstein, L.V., Christian, M.C. & Simon, R.M. (1994). A comparison of two phase I trial designs, *Statistics in Medicine* **13**, 1799–1806.

[29] Lan, K.K.G. & DeMets, D.L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics* **45**, 1017–20.

[30] LeBlanc, M. & Crowley, J. (1995). Semiparametric regression functionals, *Journal of the American Statistical Association* **90**, 95–105.

[31] Link, M.P., Donaldson, S.S., Berard, C.W., Shuster, J.J. & Murphy, S.B. (1990). Results of treatment of childhood localized Non-Hodgkin's lymphoma with combination chemotherapy with of without radiotherapy, *New England Journal of Medicine* **322**, 1169–1174.

[32] Mayer, R.J., Davis, R.B., Schiffer, C.A., Berg, D.T., Powell, B.L., Schulman, P., Omura, G.A., Moore, J.O., McIntyre, O.R. & Frei, E. III (1994). Intensive post-remission chemotherapy in adults with acute myeloid leukemia, *New England Journal of Medicine* **331**, 896–903.

[33] McGuire, W.P., Hoskins, W.J., Brady, M.F., Homesley, H.D., Creasman, W.T., Berman, M.L., Ball, H., Berek, J.S. & Woodward, J. (1995). Assessment of dose-intensive therapy in suboptimally debulked ovarian cancer: a Gynecologic Oncology Group study, *Journal of Clinical Oncology* **13**, 1589–1599.

[34] McGuire, W.P., Hoskins, W.J., Brady, M.F., Kucera, P.R., Partridge, E.E., Look, K.Y., Clarke-Pearson, D.L. & Davidson, M. (1996). Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer, *New England Journal of Medicine* **334**, 1–6.

[35] Moertel, C.G., Fleming, T.R., McDonald, J.S., Haller, D.G., Laurie, J.A., Tangen, C.M., Ungerleider, J.S., Emerson, W.A., Tormey, D.C., Glick, J.H., Veeder, M.H. & Mailliard, J.A. (1995). Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report, *Annals of Internal Medicine* **122**, 321–326.

[36] Murphy, S.B. (1995). The national impact of clinical cooperative group trials for pediatric cancer, *Medical and Pediatric Oncology* **24**, 279–280.

[37] O'Quigley, J., Pepe, M. & Fisher, L. (1990). Continual reassessment method: a practical design for Phase I clinical trials in cancer, *Biometrics* **46**, 33–48.

[38] Pediatric Oncology Group (1992). Progress against childhood cancer: the Pediatric Oncology Group experience, *Pediatrics* **89**, 597–600.

[39] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.

[40] Sause, W.T., Scott, C., Taylor, S., Johnson, D., Livingston, R., Komaki, R., Emami, B., Curran, W.J., Byhardt, R.W., Turrisi, A.T., Dar, A.R. & Cox, J.D. (1995). Radiation Therapy Oncology Group (RTOG) 88-08 and Eastern Cooperative Oncology Group (ECOG) 4588: preliminary results of a phase III trial in regionally advanced, unresectable non-small-cell lung cancer, *Journal of the National Cancer Institute* **87**, 198–205.

[41] Simon, R. (1989). Optimal two-stage designs for phase II clinical trials, *Controlled Clinical Trials* **10**, 1–10.

[42] Simon, R. (1991). A decade of progress in statistical methodology for clinical trials, *Statistics in Medicine* **10**, 1789–1817.

[43] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society, Series A* **157**, 357–416.

[44] Sposto, R., Sather, H.N. & Baker, S.A. (1992). A comparison of tests of the difference in the proportion of patients who are cured, *Biometrics* **48**, 87–100.

[45] Storer, B.E. (1989). Design and analysis of Phase I clinical trials, *Biometrics* **45**, 925–937.

[46] Sylvester, R. & Meunier, F. (1994). The EORTC central office/data center, *European Journal of Cancer* **30A**, 229–232.

[47] Thall, P.F., Simon, R. & Ellenberg, S.S. (1989). A two-stage design for choosing among several experimental treatments and a control in clinical trials, *Biometrics* **45**, 537–547.

[48] Wieand, S., Schroeder, G. & O'Fallon, J.R. (1994). Stopping when the experimental regimen does not appear to help, *Statistics in Medicine* **13**, 1453–1458.

[49] Wieand, S. & Therneau, T. (1987). A two-stage design for randomized trials with binary outcomes, *Controlled Clinical Trials* **8**, 20–28.

[50] Wood, W.C., Budman, D.R., Korzun, A.H., Cooper, M.R., Younger, J., Hart, R.D., Moore, A., Ellerton, J.A., Norton, L., Ferree, C.R., Ballow, A.C., Frei, E. & Henderson, I.C. (1994). Dose and dose intensity of adjuvant chemotherapy for stage II, node-positive breast carcinoma, *New England Journal of Medicine* **330**, 1253–1259.

[51] Woods, W.G., Kobrinsky, N., Buckley, J.D., Lee, J.W., Sanders, J., Neudorf, S., Gold, S., Barnard, D.R., DeSwarte, J., Dusenbery, K., Kalousek, D., Arthur, D.C. & Lange, B.J. (1996). Timed-sequential induction therapy improves postremission outcome in acute myeloid leukemia: a report from the Children's Cancer Group, *Blood* **87**, 4979–4989.

LAWRENCE V. RUBINSTEIN &
RICHARD S. UNGERLEIDER

Cooperative Heart Disease Trials

The National Heart, Lung, and Blood Institute (NHLBI) of the **National Institutes of Health**, initially the National Heart Institute (NHI) and then the National Heart and Lung Institute (NHLI), has undertaken cooperative (**multicenter) clinical trials** in heart disease since the early 1960s (*see* **Cardiology and Cardiovascular Disease**). This article describes the evolution of trial structure, and draws on several of the best known trials as examples of how these large, cooperative trials were organized and carried out. A summary of early methodological developments for clinical trials at the National Heart, Lung, and Blood Institute, with greater emphasis on methodology, was given by Halperin et al. [43] (*see* **Clinical Trials, Early Cancer and Heart Disease**).

Prior to 1960, the National Heart Institute supported one multicenter clinical trial, a cooperative study of the relative merits of adrenocorticotropic hormone (ACTH), cortisone, and aspirin in the treatment of rheumatic fever and the prevention of rheumatic heart disease. This trial enrolled 497 children, a small number by present standards, at centers in Great Britain, Canada, and the United States. After six weeks of treatment and another three weeks of observation, it was concluded there was no evidence that any of the three agents cured the disease. At five years, there was no difference in the amount and severity of rheumatic heart disease, but it was observed that the status of the heart at the time treatment was begun was the major factor in determining prognosis. The final report (Rheumatic Fever Working Party of the Medical Research Council of Great Britain and the Subcommittee of Principal Investigators of the American Council on Rheumatic Fever and Congenital Heart Disease, American Heart Association [67]) contained extensive data tabulations by various subgroups, but virtually no statistical analysis. This trial gave little indication of the number of methodological developments that were to come out of trials sponsored by the NHI and its successors.

The Greenberg Report

In the mid-1960s, the National Advisory Heart Council appointed a Heart Special Project Committee chaired by Dr. Bernard **Greenberg** of the University of North Carolina School of Public Health, to set down guidelines for the organization, review, and administration of cooperative studies. The resulting "Greenberg Report", completed in 1967, was published in 1988 [64] and laid down a structure for NHLBI-sponsored clinical studies. The foresight of this committee is remarkable, as the procedures they laid down have been followed ever since, with only minor modifications.

The report defined a cooperative study as "an identified activity in which two or more investigators in separate institutions contribute toward a common research goal. . ., follow a common protocol and work within a clearly defined structure for the project as a whole." The pooled resources would minimize the length of time it would take to accrue subjects in order to obtain a significant answer to a clinical or epidemiologic problem.

There were four criteria for good studies:

"1. The problem to be studied is an important one that must be resolved (a) from a purely scientific point of view, or (b) for the benefit of mankind through improved methods of prevention, diagnosis, and/or therapy;
2. An answer must be obtained in a relatively short time, and a multiinstitutional collaborative effort is the best way to reach a solution in the briefest period;
3. The study is feasible within the potential cooperating institutions, and likely to lead to an answer; and
4. There is assurance of adequate leadership and control of performance for the duration of the study."

Ideally, the question to be answered would be clearly defined and simple. Competent biometric advice would be sought early for assistance in protocol design (*see* **Clinical Trials Protocols**). This set the stage for the active participation of the Biometrics Branch (later the Biostatistics Research Branch and now the Office of Biostatistics Research of NHLBI) as collaborators involved in the planning, design, implementation, and analysis of cooperative studies. The protocol itself should be clear in order to maintain a high scientific level in the project. Continuing strong leadership with a well-defined administrative structure and control of performance at all levels

would assure that the project was carried out to completion (*see* **Clinical Trials Audit and Quality Control**).

The basic units that were essential to achieve the aims of a cooperative study are the local units, which see the participants and collect the data. The Coordinating Center would receive the data, assure their quality, and undertake appropriate, and timely analyses. The Director of the Coordinating Center would monitor performance of the local units and have the authority to carry out policing activities (*see* **Data Management and Coordination**). An Executive or Steering Committee, comprised of a limited number of strong investigators, should supervise the Coordinating Center. It should be led by a study chair, the most important position in a cooperative project. An independent Policy Board or Advisory Committee of experts in the field of the study, but not contributing to the study should review study plans and offer substantial advice. Last, the Institute staff would interact with both the Steering Committee and the Policy Board, as well as with other advisors to the Institute (such as the National Advisory Heart Council as it was then called).

The role of the Institute in cooperative studies was also described. The staff should play an active role during the early planning phases, drawing on past experience, and acting as a liaison between review committees, the National Advisory Heart Council, and investigators. Continuing communication with investigators was also advocated, as was exertion of a considerable degree of control over the study itself. The suggestion was made that one staff member serve as the Project Officer for each study and that an appropriate means of funding such studies would be a phased contract mechanism.

Major Clinical Trials Sponsored by NHLBI

The major multicenter heart disease clinical trials sponsored by the NHLBI and initiated before 1980 are listed in Table 1. Several of the trials are discussed in detail below and references to the others are given for the interested reader [4–6, 12–14].

The Coronary Drug Project (CDP)

Preceding the Greenberg Report in design, but not completion, the Coronary Drug Project (CDP) was the first large, multicenter clinical trial sponsored by

Table 1 Multicenter clinical trials in heart disease sponsored by NHLBI with at least 500 patients initiated prior to 1980

Trial	No. centers	Treatments (sample size)	Dates
Coronary Drug Project (CDP)	53	Clofibrate (1103) Niacin (1119) Dextrothyroxine (1110) Estrogen - low dose (1101) Estrogen - high dose (1119) Placebo (2789)	1966–1969
Hypertension Detection and Follow-up Program (HDFP)	14	Stepped care (5485) Referred care (5455)	1971–1975
Multiple Risk Factor Intervention Trial for the Prevention of Coronary Heart Disease (MRFIT)	22	Special intervention (6428) Usual care (6438)	1972–1983
Lipid Research Clinics Coronary Primary Prevention Trial (LRC)	12	Cholestyramine (1906) Placebo (1900)	1973–1983
Coronary Artery Surgery Study (CASS)	15	Surgery (390) Medical (390)	1973–1983
Program on Surgical Control of Hyperlipidemias (POSCH)	4	Surgery (421) Control (417)	1973–1990
Aspirin Myocardial Infarction Study (AMIS)	30	Aspirin (2267) Placebo (2257)	1974–1979
β-Blocker Heart Attack Trial (BHAT)	31	Propranolol (1916) Placebo (1921)	1977–1981
Multicenter Investigation of Limitation of Infarct Size (MILIS)	5	Propranolol (134) Hyaluronidase (420) Placebo (431)	1978–1988

NHI and was designed to assess the efficacy of several lipid modifying drugs in men between 30 and 64 years of age who had previously had a myocardial infarction (MI). The initiation of such a trial was recommended by the National Heart Advisory Council in 1960 and planning began in the Institute in early 1961 [78].

Several years of budgetary and political negotiations followed. Finally, with a Policy Board, a Coordinating Center, and five clinics in place, the first participant was enrolled in 1966, and with 48 additional clinics, enrollment of 8341 participants was completed in 2.5 years. All participants were followed for a minimum of 54 months, with 96% followed up for at least five years [17].

Randomization was performed by the Coordinating Center, stratified on clinic and risk (whether the participant had one prior MI with no complications or was of higher risk) (*see* **Randomized Treatment Assignment**). Double-blinded (neither participant nor treating physician knew which treatment a participant received) and placebo controlled (*see* **Blinding or Masking**), the study treatments included two doses of equine estrogens, clofibrate, dextrothyroxine, nicotinic acid, and a lactose placebo. Two and a half times as many subjects were assigned to placebo as to any other treatment, which is optimal allocation when five treatments are being compared with a single **control**. The sample size was set to detect a 30% mortality in the control group and 22.5% (a 25% reduction in mortality) in any treatment group when the mortality percentage in any treatment group was compared to that of the control group at the 1% significance level (one-sided test using an arc sine **transformation** with 95% **power** (*see* **Sample Size Determination for Clinical Trials**). Further sample size adjustments were made for the possibility of dropouts and treatment withdrawals as well as the time necessary to achieve maximum benefit of treatment [31, 45].

The CDP investigators had the foresight to consider many important aspects of trial design: optimal allocation when multiple treatments are compared to a single control; that there would be dropouts, which would cause loss of power and so the sample size should be increased to compensate; that treatment benefit was not likely to be attained instantaneously; and that there should be an adjustment of the significance level at which tests were conducted because there would be **multiple comparisons** [31]. Although methodology for sample size calculations has greatly advanced, new methods consider these same principles. The final results paper of the CDP, reported results based on an **intention to treat analysis**: all patients randomized were included, no matter how well or poorly they adhered to the treatment plan [35]. The final analysis reported on both the planned analysis (the proportion of events at the end of the trial, which was used for trial design), and the **Cox proportional hazards** model [33]. At the time of the design of the CDP, no methods for calculating sample size for time-to-event data were known (*see* **Sample Size Determination in Survival Analysis**).

The Steering Committee of the CDP consisted of the study chair, the director of the Coordinating Center, several study principal investigators (some permanent and others rotating), the NHI Medical Liaison Officer, directors of the ECG reading center and the central laboratory, and biostatistical representation from the Biometrics Research Branch. It met twice a year throughout the study. The CDP technical group represented all operational units participating in the CDP. Initially, unblinded data reports by study treatment were presented to the CDP technical group but, recognizing the potential for compromising the trial, in 1968, a data and safety monitoring committee (DSMC) (*see* **Data Monitoring Committees**) was appointed to review the accumulating data by treatment group. The Policy Board advised on policy matters and consisted of five voting members, all of whom were independent of the study investigators. In addition, nonvoting members from the Coordinating Center, Steering Committee, and Institute attended meetings. The Policy Board acted on the recommendation of the DSMC regarding continuation of the study or discontinuation of one or more treatment groups. Aside from these structural committees, there were trial committees, such as the Editorial Review Committee that reviewed and approved all plans for papers and composition of writing committees and all oral presentations [17].

As for the results of the CDP, adverse effects of the two estrogen treatments and of dextrothyroxine led to successive halting of these treatments [29, 30, 32] and continuation of the others. Several monitoring procedures were used as the data accumulated [16, 28, 46] (*see* **Data and Safety Monitoring**). A good summary is given by Canner [18]. Of participants who had received the treatments that were stopped

early, 1529 were randomized into a double-blind trial of aspirin versus placebo [34].

At the end of the CDP, no evidence of efficacy was found for clofibrate, and some was found for nicotinic acid with respect to definite, nonfatal MI, but not for total mortality [33]. Subsequent 15-year follow-up of the study participants showed a statistically significant reduction in total mortality from nicotinic acid [19].

Hypertension Detection and Follow-up Program (HDFP)

While the CDP developed or implemented many of the designs and biostatistical features that are currently used in multicenter clinical trials, modifications and new concepts were incorporated into subsequent trials. The Hypertension Detection and Follow-up Program (HDFP) began in the early 1970s [47–50]. This trial of the effects of antihypertensive treatment on mortality in 10 940 high blood pressure participants was novel in several respects. First, it had broad **eligibility criteria**: anyone aged between 30 and 69, with diastolic blood pressure 90 mm Hg or over was eligible. As a result, some of the participants had preexisting diseases (stroke, myocardial infarction, and renal disease), although most did not. Secondly, all but one of the collaborating centers enrolled participants in residential areas, thus making it a community-based study. Thirdly, almost half (about 45%) of the participants were women and about the same number were black. Fourthly, the HDFP used a stepped care approach to intervention, with medication being prescribed in a standard sequence in order to achieve the blood pressure goal. The control group received "referred care", that is, usual medical care, and the study was not blinded; in other words both participants and investigators were aware of the treatment arm to which each participant was assigned. To assure no bias was introduced as a result of treatment being known to participants (*see* **Bias, Overview**), the primary **outcome** was five-year all-cause mortality, and secondary outcomes were objective measures of heart disease and stroke, which were assessed at regular intervals. It might be noted that, like the CDP, the HDFP had two monitoring groups. One, called the *Toxicity and Endpoints Evaluation Committee*, reviewed data, endpoints, and toxicity reports at specified intervals. Membership on this committee included the chair of the Policy Advisory Board, the head of the Coordinating Center, several clinical investigators (some of whom were involved in the study, but did not see study patients), and Institute staff. The other monitoring group, the Policy Advisory Board, was entirely external to the study, and monitored the trial and advised the Institute on its progress.

The conclusion of HDFP was that stepped care was superior to referred care both with respect to control of diastolic blood pressure and with respect to overall mortality at five years. Five-year mortality from all causes was 17% lower in the stepped care group compared to the referred care group. The trial led to a recommendation for systematic, effective management of hypertension for its great potential to reduce mortality from hypertension, including those with "mild" hypertension (diastolic blood pressure 90–104 mm Hg) [48]. Because it was recognized that power was limited to detect differences in various subgroups of interest, substantial care was taken in reporting results in race, sex, and age subgroups (*see* **Treatment-covariate Interaction**); no ***P*-values** were reported [49, 50].

Multiple Risk Factor Intervention Trial (MRFIT)

A third major multicenter clinical trial was the Multiple Risk Factor Intervention Trial (MRFIT) [61], which was started in 1972. This primary **prevention trial** selected participants at high risk of developing heart disease, based on a **logistic regression** equation developed by **J. Cornfield**, and others of **prognostic factors** based on the **Framingham Heart Study** [71]. A combination of serum cholesterol and blood pressure levels as well as number of cigarettes smoked per day determined eligibility. The intervention consisted of efforts to reduce those risk factors (diet for cholesterol, drug for blood pressure, and behavioral intervention for smoking). The power to detect benefit from any one of the interventions was quite low, and therefore, the study was designed to compare the intervention to the control group. As with the HDFP, the control group received only usual medical care, the study was not blinded, and the outcomes were measured only annually, so the more frequent evaluations of the intervention group would not bias the endpoint assessment. This trial began with a dual external data monitoring/oversight organization, but evolved to a single oversight committee because of greatly overlapping responsibilities [41]. Despite the

use of the best available data for outcome incidence during the design phase, the event rates at the end of the trial were lower than expected and no difference was found in the primary endpoint, coronary heart disease deaths, at six years [62]. It is likely that a combination of self-selection factors for volunteer studies (*see* **Selection Bias**) and secular trends, due to enhanced appreciation in the community of the need to treat risk factors, were largely responsible. Efforts have been made in subsequent studies to account for such possibilities. Long-term assessment of mortality in MRFIT was continued, and a trend in favor of the special intervention group was seen after 10.5 years [63].

Lipid Research Clinics Cholesterol Primary Prevention Trial (LRC-CPPT)

While HDFP had broad eligibility criteria, the Lipid Research Clinics Cholesterol Primary Prevention Trial (LRC-CPPT) [54], started two years later, had much narrower eligibility criteria. It was entirely a primary prevention study in the sense that all participants were to be free of existing heart disease. Also, it used a combination primary outcome, definite coronary heart disease death and/or a definite nonfatal myocardial infarction. The bile acid sequestrant, cholestyramine, was compared with placebo after 7.4 years of double-blinded treatment of 3806 asymptomatic 35 to 59 year old men with hypercholesterolemia.

Of note, according to the original analysis plan, a one-sided significance level of 5% was used to declare statistical significance (*see* **Hypothesis Testing**). The one-sided **logrank** test (stratified on eight baseline risk factors and adjusted for multiple looks at the data: *see* **Multiplicity in Clinical Trials**) yielded a P-value just under 5% in favor of the active treatment [55, 56]. Because the results would not have been statistically significant, had a two-sided significance level of 5% been used, they generated some controversy [51]. Many in the cardiovascular clinical trials community prefer two-sided hypothesis testing to one-sided testing and believe that if a one-sided test is done, the significance level should be 2.5% rather than 5%. Had the LRC-CPPT been designed with a two-sided significance level of 5% (or a one-sided significance level of 2.5%), the sample size would have had to be larger. Despite the controversy, it was clear that cholestyramine decreased primary outcome events compared with placebo, and the results of this trial were instrumental in public campaigns to reduce serum cholesterol levels.

Other Secondary Prevention Trials of the 1970s and 1980s

Several secondary prevention trials that began in the mid- to late 1970s, and the 1980s made further design and biostatistical advances. New stopping guidelines for data monitoring were incorporated. In the Multicenter Investigation of Limitation of Infarct Size (MILIS), patients who had had a myocardial infarction in the previous 18 hours were randomized to propranolol or placebo, or in patients in whom propranolol was contraindicated, to hyaluronidase or placebo [60]. Conditional power was used to declare propranolol ineffective before the scheduled end of the trial [68]. Conditional power assesses the probability that a difference will be found at the end of a trial considering the results up to the time of the calculation, and making a variety of assumptions about the future, including continuation of the current trend (leading to "stochastic curtailment"; *see* **Data and Safety Monitoring**). The final results, that hyaluronidase was no better than placebo, were later reported [59].

The β-Blocker Heart Attack Trial (BHAT) randomized 3837 patients recovering from myocardial infarction to propranolol or placebo administered in a double-blind manner to assess overall mortality during a two- to four-year period [8–10]. The sample size was set by a new method to allow for **noncompliance** [8, 15, 37, 76]. Both conditional power and group sequential boundaries (*see* **Sequential Methods for Clinical Trials**) were used to stop the study nine months early because of significant benefit of propranolol [9, 10, 37, 44, 52, 53]. Also, assuming the current trends, the gain in precision in estimating survival distributions had BHAT continued to its planned termination was examined [37].

The Coronary Artery Surgery Study (CASS) compared coronary artery bypass surgery to medical therapy in participants with either angina or previous myocardial infarction [24]. No significant difference was found between the two treatments with respect to survival or to nonfatal myocardial infarction, although the bypass group had a lower event rate than did the medical group. The results were explained by the event rate in the medical group being lower than

others had reported. This was perhaps due to entry of participants with better prognostic factors or due to the method used to confirm nonfatal myocardial infarction [25, 26].

One of the concerns about clinical trials is that the **study population** is not representative of the general population of those with the condition (*see* **Clinical Trials, Overview**). To address this criticism, CASS used a registry of participants not enrolled in the randomized trial to assess generalizability of the trial results. It found that results from the trial were comparable with findings from the registry [27], and were certainly reassuring.

The Cardiac Arrhythmia Suppression Trial (CAST) was designed in 1986 to assess the effectiveness of antiarrhythmic therapy in postmyocardial infarction patients with ventricular arrhythmias. In preparation, a pilot study, the Cardiac Arrhythmia Pilot Study (CAPS) was conducted to assess feasibility, and answer certain other design questions essential to the optimal design of the full-scale trial [20, 21]. This pilot study was separate from the full-scale trial. Other studies have since employed an internal pilot [7]; that is, the study is designed so that the pilot data can be included in the final trial's analysis. The sample size for the whole study can be based on certain pilot information (such as variability and estimation of control group event rates) and if there are no changes in the eligibility or interventions, there is little impact on the trial's significance level. Some have argued that an internal pilot approach increases study efficiency [72].

One of the results of CAPS was that it was recognized that some patients with arrhythmias could have them suppressed by at least 80% by some, but not all, of the drugs under study. In CAST, during a run-in period, patients received one of the three active treatments. If that first drug suppressed their arrhythmia, they were randomized between it and placebo. If the first drug did not suppress their arrhythmia (either at the initial dose or at a higher dose), they received another active drug. After this run-in, if their arrhythmias were not suppressed, they were not randomized. Thus, CAST did not compare three active treatments and placebo, but tested the hypothesis that suppressing arrhythmias with *any* of the three drugs would reduce the risk of arrhythmic death or cardiac arrest when compared to placebo. This design was an attempt to mimic treatment in clinical practice.

In CAST, the formal test of significance was one-sided (for benefit) with a 2.5% significance level, but there were advisory stopping guidelines for harm. The CAST Data and Safety Monitoring Board decided initially to remain blinded to which group was intervention and which was control. In the first part of CAST, the advisory stopping boundary for harm was symmetrical to the guideline for benefit [65]. A strong trend for difference in mortality was noted early, but the Data and Safety Monitoring Board decided that regardless of the direction, the study should continue. Because of an increasing difference, the Board was unblinded, and learned that it was in a direction harmful for the treatment group, with the advisory boundary for harm being crossed [11, 42]. As a result, encainide and flecainide, two of the three drugs being used in the treatment arm, were discontinued [22].

In CAST II, randomization between the third antiarrhythmic drug, moricizine, and placebo continued. The advisory stopping boundary for harm was less extreme than the boundary for benefit, reflecting the view that less evidence would then be required to stop if the trend were going in an unfavorable direction. Another difference in the second part of CAST was that assessment of the proper dose of antiarrhythmic drug, which had been done in an open fashion in the first part of CAST, was done blinded, with a placebo control. CAST II was also stopped ahead of schedule. Conditional power calculations showed that the remaining antiarrhythmic drug was extremely unlikely to be proven beneficial. In addition, during the dose-ranging incorporated in the study, there was strong evidence of harm from the active agent [23].

As has been noted [42], CAST provided several data monitoring lessons. First, if a one-sided test of hypothesis is used, a plan to monitor the data should be in place, even for trends in the unanticipated direction. Second, adverse events can accumulate quickly, and the monitoring guidelines need to be in place from the beginning of the trial. Third, in placebo-controlled trials, monitoring adverse events is not a symmetric process, suggesting that it is not wise for a data monitoring committee to be blinded to treatment assignment. Fourth, a variety of factors, both statistical and nonstatistical, are considered in monitoring. Statistical monitoring guidelines can assist, but cannot replace, judgment about when to stop, continue, or modify a study protocol.

Designs of Other Multicenter Trials

Multicenter clinical trials since the late 1980s evolved in other directions, with more studies including **cluster randomization** designs, large simple designs, and **factorial designs**.

Cluster (group) randomization has been undertaken in several trials in which special health-promoting interventions in schools or communities are compared with usual procedures. The Child and Adolescent Trial of Cardiovascular Health (CATCH) randomized schools and intervened in the third through fifth grades with respect to lower fat content school lunches, enhanced physical education, and enhanced classroom health curricula [57, 77]. The Rapid Early Action for Coronary Treatment (REACT) trial randomized 10 matched pairs of cities to evaluate a community-based intervention aimed at reducing delay time from onset of heart attack symptoms to hospital arrival [58]. In such trials, the cluster is the unit of randomization and it is imperative to consider that in both the design and analysis (*see* **Unit of Analysis**).

The large, simple trial design has been adopted, when appropriate, as in the Digitalis Investigation Group (DIG) trial, a trial of digitalis in patients with heart failure [38, 39]. Although earlier studies had collected limited data and had many clinics, they had short-term interventions and follow-up. DIG was one of the first to use that approach in a trial requiring long-term drug administration and follow-up. Other trials, such as the Studies of Left Ventricular Dysfunction (SOLVD), a trial of an angiotensin-converting enzyme inhibitor in patients with poor left ventricular ejection fraction, combined a relatively simple protocol with numerous substudies aimed at assessing detailed physiologic, biochemical, or behavioral mechanisms [69, 70].

Factorial designs have become more commonly used. Sometimes this design has been used in the traditional manner, as in the Post-Coronary Artery Bypass Graft Trial (Post-CABG), a two-by-two factorial trial of intensive versus moderate lipid-lowering and low-dose anticoagulation versus placebo in patients who had had coronary bypass surgery [66]. Other trials have had a "partial factorial" design. These have consisted of separate trials of interventions believed to be independent (as in a factorial design), with some, but not all of the subjects enrolled in more than one of the trials. These partial factorial designs are often analyzed as independent trials as they would be in a factorial design.

The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) consisted of two large, simple trials with a partial factorial design. One of the trials compared a diuretic to three newer antihypertensive agents (amlodipine, a calcium agonist, lisinopril, an ACE inhibitor and doxazosin, an α-adrenergic blocker) in high-risk hypertensive people to see if there were differences in the occurrence of fatal coronary heart disease or nonfatal myocardial infarction. This trial was designed to enroll 40 000 eligible participants with hypertension, and intervene and follow them for a mean duration of six years. The second ALLHAT trial offered those who were moderately hypercholesterolemic the opportunity to be randomly assigned to open-label pravastatin, a lipid-lowering drug, or usual care, with total mortality as the primary outcome. Not only was this trial a partial factorial design, intended to be conducted with limited data collection, it was designed to be carried out in clinical practice settings. Over 600 practices took part in the antihypertensive component and over 500 in the lipid-lowering component [36].

The α-adrenergic blocker arm of ALLHAT was stopped early because of an excess of combined cardiovascular events, in particular, heart failure. The other three hypertensive treatment arms continued, with the result that each of the newer agents was no better than the diuretic [1, 2]. The results of the lipid-lowering component were that pravastatin did not reduce either all-cause mortality or CHD significantly when compared with usual care, perhaps due to the modest differential in total cholesterol between pravastatin and usual care groups [3]. This study demonstrated that given important clinical questions and a sufficiently simple protocol, many physicians and nurses not traditionally associated with clinical research could effectively participate.

A large study that is far from simple is the Women's Health Initiative (WHI) (*see* **Women's Health Initiative: Statistical Aspects and Selected Early Results**). This study in postmenopausal women was started in 1992, and includes four clinical trials with a partial factorial design. Two trials, in women with and without a uterus, randomized participants to hormone replacement therapy (combined estrogen and progesterone in women with a uterus and estrogen alone in women without a uterus) or placebo

to test whether coronary heart disease and other cardiovascular diseases and hip and other fractures could be reduced. A third trial of a low-fat eating pattern versus usual dietary advice was designed to see if breast cancer, colorectal cancer, and coronary heart disease could be reduced. Subjects could enter both a hormone replacement trial and the diet trial. The fourth trial of the WHI randomized women to calcium and vitamin D supplementation or placebo to see if hip and other fractures and colorectal cancer could be reduced. To enter the calcium and vitamin D supplementation trial, a subject had to be enrolled in either one of the hormone replacement therapy trials or the diet trial. In addition, a large observational study component was conducted along with the clinical trial [74].

The WHI is notable for several reasons. First, it is a very large undertaking, with approximately 68 000 women in one or another of the trials. Second, a partial factorial design was employed and the trials are treated as independent trials. Third, the observational component has over 93 000 women, providing considerable information along with the trial. Fourth, a trial of such complexity and societal import requires careful monitoring of the many endpoints of interest. The statistical stopping rules for the hormone replacement trial not only monitored the primary and secondary endpoints, but also considered adverse events. An overall measure of risk and benefit was also assessed [40].

The first results from the WHI were published in 2002. The hormone replacement trial in women with an intact uterus was stopped early due to an early increase in breast cancer, the primary adverse event. Somewhat surprisingly, the combination of estrogen and progestin not only led to an expected increase in breast cancer and reduction in bone fractures, it caused an increase in cardiovascular disease [75]. This last finding was contrary to expectation based on many epidemiology studies, showing the hazards of relying only on **observational studies**. In 2004, the WHI trial of estrogen-alone, in women who had had a hysterectomy, was also stopped ahead of schedule. This was a more difficult decision. The study showed the expected reduction in bone fractures, and there was an unexpected non-significant trend in the direction of fewer cases of breast cancer. Importantly, however, estrogen increased the risk of stroke and did not reduce the risk of coronary heart disease [73].

Summary

This review of NHLBI-sponsored multicenter trials in heart disease illustrates some design and biostatistical advances that selected trials have developed or used effectively to answer important public health questions. Many of the design, organizational characteristics, and methods for monitoring modern clinical trials were implemented as a result of the Greenberg Report and were first used in the CDP. While more recent studies employed more efficient organizations, newer designs, and newer monitoring guidelines, the genesis of multicenter clinical trial practice in heart disease was outlined in the 1960s and that early structure has endured the test of time.

References

[1] ALLHAT Collaborative Group (2000). Major cardiovascular events in hypertensive patients randomized to doxazosin vs chlorthalidone: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), *Journal of the American Medical Association* **283**, 1967–1975.

[2] ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group (2002). Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), *Journal of the American Medical Association* **288**, 2981–2997.

[3] ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group (2002). Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT-LLT), *Journal of the American Medical Association* **288**, 2998–3007.

[4] Aspirin Myocardial Infarction Research Group (1977). An intervention study - the aspirin myocardial infarction study, *Lipids* **12**, 59–63.

[5] Aspirin Myocardial Infarction Research Group (1980a). A randomized controlled trial of aspirin in persons recovered from myocardial infarction, *Journal of the American Medical Association* **243**, 661–669.

[6] Aspirin Myocardial Infarction Study Research Group (1980b). The aspirin myocardial infarction study: final results, *Circulation* **62**(Suppl. V), V79–V84.

[7] AVID Investigators (1995). Antiarrhythmics Versus Implantable Defibrillators (AVID) - Rationale, design, and methods, *The American Journal of Cardiology* **75**, 470–475.

[8] β-Blocker Heart Attack Research Group (1981a). β-Blocker heart attack trial: design features, *Controlled Clinical Trials* **2**, 275–285.

[9] β-Blocker Heart Attack Study Group (1981b). A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *Journal of the American Medical Association* **247**, 1707–1714.

[10] β-Blocker Heart Attack Study Group (1983). A randomized trial of propranolol in patients with acute myocardial infarction. II. Morbidity results. *Journal of the American Medical Association* **250**, 2814–2819.

[11] Bigger, J.T. Jr. (1990). The events surrounding the removal of encainide and flecainide from the Cardiac Arrhythmia Suppression Trial (CAST) and Why CAST is continuing with moricizine, *Journal of the American College of Cardiology* **15**, 243–245.

[12] Buchwald, H., Matts, J.P., Fitch, L.L., Varco, R.L., Campbell, G.S., Pearce, M., Yellin, A., Smink, Jr., R.D., Sawin, Jr., H.S., Campos, C.T., Hansen, B.J., Long, J.M., the POSCH Group. (1989). Program on the Surgical Control of the Hyperlipidemias (POSCH): design and methodology, *Journal of Clinical Epidemiology* **42**, 1111–1127.

[13] Buchwald, H., Moore, R.B. & Varco, R.L. (1982). Surgery in the therapy of atherosclerosis: partial ileal bypass, in *Atherosclerosis: Clinical Evaluation and Therapy*. MTP Press, International Medical Publication, Lancaster.

[14] Buchwald, H., Varco, R.L., Matts, J.P., Long, J.M., Fitch, L.L., Campbell, G.S., Pearce, M.B., Yellin, A.E., Edmiston, W.A., Smink, Jr., R.D., Sawin, Jr., H.S., Campos, C.T., Hansen, B.J., Tuna, N., Karnegis, J.N., Sanmarco, M.E., Amplatz, K., Castaneda-Zuniga, W.R., Hunter, D.W., Bissett, J.K., Weber, F.J., Stevenson, J.W. Leon, A.S., Chalmers, T.C., the POSCH Group (1990). Effect of partial ileal bypass surgery on mortality and morbidity from coronary heart disease in patients with hypercholesterolemia. Report of the Program on the Surgical Control of the Hyperlipidemias (POSCH), *New England Journal of Medicine* **323**, 946–955.

[15] Byington, R.P. (1984). Beta-Blocker heart attack trial: design, methods, and baseline results, *Controlled Clinical Trials* **5**, 382–437.

[16] Canner, P.C. (1977). Monitoring differences in long term clinical trials, *Biometrics* **33**, 603–613.

[17] Canner, P., ed. (1983a). Coronary drug project: methods and lessons of a multicenter clinical trial, *Controlled Clinical Trials* **4**, 273–536.

[18] Canner, P. (1983b). Further aspects of data analysis in the coronary drug project, *Controlled Clinical Trials* **4**, 485–503.

[19] Canner, P., Berge, K.G., Wenger, N.K., Stamler, J., Friedman, L., Prineas, R.J., Friedewald, W. for the Coronary Drug Project Research Group (1986). Fifteen year mortality in coronary drug project patients: long-term benefit with niacin, *Journal of the American College of Cardiology* **8**, 1245–55.

[20] Cardiac Arrhythmia Pilot Study Investigators (1986). The cardiac arrhythmia pilot study, *American Journal of Cardiology* **57**, 91–95.

[21] Cardiac Arrhythmia Pilot Study Investigators (1988). Effects of encainide, flecainide, imipramine, and moricizine on ventricular arrhythmias during the year after acute myocardial infarction, *American Journal of Cardiology* **61**, 501–509.

[22] Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989). Special Report. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction, *New England Journal of Medicine* **321**, 406–412.

[23] Cardiac Arrhythmia Suppression Trial II Investigators (1992). Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction, *New England Journal of Medicine* **327**, 227–233.

[24] CASS Principal Investigators and their Associates (1981). National heart, lung, and blood institute coronary artery surgery study, *Circulation* **63**, I1–I81.

[25] CASS Principal Investigators and their Associates (1983). Coronary Artery Surgery Study (CASS): a randomized trial of coronary artery bypass surgery. Survival data, *Circulation* **68**, 939–950.

[26] CASS Principal Investigators and their Associates (1984a). Coronary Artery Surgery Study (CASS): a randomized trial of coronary artery bypass surgery: comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomization criteria, *Journal of the American College of Cardiology* **3**, 114–128.

[27] CASS Principal Investigators and their Associates (1984b). Myocardial infarction and mortality in the Coronary Artery Surgery Study (CASS) randomized trial, *New England Journal of Medicine* **310**, 750–758.

[28] Cornfield, J. (1969). The Bayesian outlook and its application, *Biometrics* **25**, 617–657.

[29] Coronary Drug Project Research Group (1970). The coronary drug project: initial findings leading to modifications of its research protocol, *Journal of the American Medical Association* **214**, 1303–1313.

[30] Coronary Drug Project Research Group (1972). The coronary drug project: findings leading to further modification of its protocol with respect to dextrothyroxine, *Journal of the American Medical Association* **220**, 996–1008.

[31] Coronary Drug Project Research Group (1973a). The coronary drug project: design, methods, and baseline results, *Circulation* **47**(Suppl. 1), I1–I79.

[32] Coronary Drug Project Research Group (1973b). The coronary drug project: findings leading to discontinuation of the 2.5 mg/day estrogen group, *Journal of the American Medical Association* **226**, 652–657.

[33] Coronary Drug Project Research Group (1975). Clofibrate and niacin in coronary heart disease, *Journal of the American Medical Association* **231**, 360–381.

[34] Coronary Drug Project Research Group (1976). Aspirin in coronary heart disease, *Journal of Chronic Diseases* **29**, 625–642.

[35] Coronary Drug Project Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project, *New England Journal of Medicine* **303**, 1038–1041.

[36] Davis, B.R., Cutler, J.A., Gordon, D.J., Furberg, C.D., Wright, Jr., J.T., Cushman, W.C., Grimm, R.H., LaRosa, J., Whelton, P.K., Perry, H.M., Alderman, M.H., Ford, C.E., Oparil, S., Francis, C., Proschan, M., Pressel, S., Black, H.R., Hawkins, C.M., for the ALLHAT Research Group (1996). Rationale and design for the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), *American Journal of Hypertension* **9**, 342–360.

[37] DeMets, D.L., Hardy, R., Friedman, L.M. & Lan, K.K.G. (1984). Statistical aspects of early termination in the beta-blocker heart trial, *Controlled Clinical Trials* **5**, 362–372.

[38] Digitalis Investigation Group (1996). Rationale, design, implementation and baseline characteristics of patients in the DIG trial: a large, simple trial to evaluate the effect of digitalis on mortality in heart failure, *Controlled Clinical Trials* **17**, 77–97.

[39] Digitalis Investigation Group (1997). The effect of digoxin on mortality and morbidity in patients with heart failure, *New England Journal of Medicine* **336**, 525–533.

[40] Freedman, L., Anderson, G., Kipnis, V., Prentice, R., Wang, C.Y., Rossouw, J., Wittes, J. & DeMets, D. (1996). Approaches to monitoring the results of long-term disease prevention trials: examples from the Women's Health Initiative, *Controlled Clinical Trials* **17**, 509–525.

[41] Friedman, L. (1993). The NHLBI model: a 25 year history, *Statistics in Medicine* **12**, 425–431.

[42] Friedman, L.M., Bristow, J.D., Hallstrom, A., Schron, E., Proschan, M., Verter, J., DeMets, D., Fisch, C., Nies, A.S., Ruskin, J., Strauss, H. & Walters, L. (1993). Data monitoring in the cardiac arrhythmia suppression trial, *Online Journal of Current Clinical Trials* Doc. No 79.

[43] Halperin, M., DeMets, D.L. & Ware, J.H. (1990). Early methodological developments for clinical trials at the National Heart, Lung and Blood Institute, *Statistics in Medicine* **9**, 881–891.

[44] Halperin, M., Lan, K.K.G., Ware, J.H., Johnson, N.J. & DeMets, D.L. (1982). An aid to data monitoring in long-term clinical trials, *Controlled Clinical Trials* **3**, 311–323.

[45] Halperin, M., Rogot, E., Gurian, J. & Ederer, F. (1968). Sample sizes for medical trials with special reference to long-term therapy, *Journal of Chronic Diseases* **21**, 13–24.

[46] Halperin, M. & Ware, J. (1974). Early decision in a censored Wilcoxon two-sample test for accumulating survival data, *Journal of the American Statistical Association* **69**, 414–422.

[47] Hypertension Detection and Follow-up Program Cooperative Group (1976). Hypertension detection and follow-up program, *Preventive Medicine* **5**, 207–215.

[48] Hypertension Detection and Follow-up Program Cooperative Group (1979a). Five-Year findings of the hypertension and follow-up program: I. Reduction in mortality of persons with high blood pressure, including mild hypertension, *Journal of the American Medical Association* **242**, 2562–2571.

[49] Hypertension Detection and Follow-up Program Cooperative Group (1979b). Five-year findings of the hypertension and follow-up program: II. Mortality by race-sex and age, *Journal of the American Medical Association* **242**, 2572–2577.

[50] Hypertension Detection and Follow-up Program Cooperative Group (1982). The effect of treatment on mortality in "mild" hypertension. Results of the hypertension detection and follow-up program. *New England Journal of Medicine* **307**, 976–980.

[51] Kronmal, RA (1985). Commentary on the published results of the lipid research clinics coronary primary prevention trial, *Journal of the American Medical Association* **253**, 2091–2093.

[52] Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.

[53] Lan, K.K.G., Simon, R. & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials, *Communications in Statistics C* **1**, 207–219.

[54] Lipid Research Clinics Program (1979). The coronary primary prevention trial: design and implementation, *Journal of Chronic Diseases* **32**, 609–631.

[55] Lipid Research Clinics Program (1984a). The lipid research clinics coronary primary prevention trial results. I. Reduction in the incidence of coronary heart disease, *Journal of the American Medical Association* **251**, 351–364.

[56] Lipid Research Clinics Program (1984b). The lipid research clinics coronary primary prevention trial results. II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering, *Journal of the American Medical Association* **251**, 365–374.

[57] Luepker, R.V., Perry, C.L., McKinlay, S.M., Nader, P.R., Parcel, G.S., Stone, E.J., Webber, L.S., Elder, J.P., Feldman, H.A., Johnson, C.C., Kelder, S.H., Wu, M., for the CATCH Collaborative Group (1996). Outcomes of a field trial to improve children's dietary patterns and physical activity. The child and adolescent trial for cardiovascular health, *Journal of the American Medical Association* **275**, 768–776.

[58] Luepker, R.V., Raczynski, J.M., Osganian, S., Goldberg, R.J., Finnegan, J.R. Jr., Hedges, J.R., Goff, D.C. Jr., Eisenberg, M.S., Zapka, J.G., Feldman, H.A., Labarthe, D.R., McGovern, P.G., Cornell, C.E., Proschan, M.A. & Simons-Morton, D.G. (2000). *Journal of the American Medical Association* **284**, 60–67.

[59] MILIS Study Group (1986). MILIS: Hyaluronidase therapy for acute myocardial infarction: results of a

randomized, blinded multicenter trial, *American Journal of Cardiology* **57**, 1236–1243.

[60] Muller, J.E. (Ed.), Braunwald, E., Mock, M.B., Mullin, S.M., Passamani, E.R., Poole, W.K. & Scheiner, E. (Assoc. Eds.) (1984). National Heart, Lung, and Blood Institute Multicenter Investigation of the Limitation of Infarct Size (MILIS). Design and methods of the clinical trial. An investigation of beta-blockade and hyaluronidase for treatment of acute myocardial infarction, *American Heart Association Monograph* **100**, 1–134.

[61] Multiple Risk Factor Intervention Trial Group (1977). Statistical design considerations in the NHLI Multiple Risk Factor Intervention Trial (MRFIT), *Journal of Chronic Diseases* **30**, 261–275.

[62] Multiple Risk Factor Intervention Trial Group (1982). Multiple risk factor intervention trial: risk factor changes and mortality results, *Journal of the American Medical Association* **248**, 1465–1477.

[63] Multiple Risk Factor Intervention Trial Research Group (1990). Mortality rates after 10.5 years for participants in the multiple risk factor intervention trial. Findings related to a priori hypotheses of the trial. *Journal of the American Medical Association* **263**, 1795–1801.

[64] Organization, Review and Administration of Cooperative Studies (Greenberg Report): A Report from the Heart Special Project Committee to the National Advisory Heart Council, May, 1967 (1988); *Controlled Clinical Trials* **9**, 137–148.

[65] Pawitan, Y. & Hallstrom, A. (1990). Statistical interim monitoring of the cardiac arrhythmia suppression trial, *Statistics in Medicine* **9**, 1081–1090.

[66] Post-CABG Investigators (1997). The effect of aggressive LDL cholesterol lowering and low dose anticoagulation on obstructive changes in saphenous vein coronary bypass grafts, *New England Journal of Medicine* **336**, 153–162.

[67] Rheumatic Fever Working Party of the Medical Research Council of Great Britain and the Subcommittee of Principal Investigators of the American Council on Rheumatic Fever and Congenital Heart Disease, American Heart Association (1960). The evolution of rheumatic heart disease in children five-year report of a cooperative clinical trial of ACTH, cortisone, and aspirin, *Circulation* **22**, 503–515.

[68] Roberts, R., Croft, C.H., Gold, H.K., Hartwell, T.D., Jaffe, A.S., Muller, J.E., Mullin, S.M., Parker, C., Passamani, E.R., Poole, W.K., Raabe, D.S., Rude, R.E., Stone, P.H., Turi, G., Sobel, B.E., Willerson, J.T., Braunwald, E., the MILIS Study Group (1984). Effect of propranolol on myocardial-infarct size in a randomized blinded multicenter trial, *New England Journal of Medicine* **311**, 218–225.

[69] SOLVD Investigators (1991). Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure, *New England Journal of Medicine* **325**, 293–302.

[70] SOLVD Investigators (1992). Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions, *New England Journal of Medicine* **327**, 685–691.

[71] Truett, J. Cornfield, J. & Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham, *Journal of Chronic Diseases* **20**, 511–524.

[72] Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials, *Statistics in Medicine* **9**, 65–72.

[73] The Women's Health Initiative Steering Committee (2004). Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial, *Journal of the American Medical Association* **291**, 1701–1712.

[74] Women's Health Initiative Study Group (1998). Design of the Women's Health Initiative, *Controlled Clinical Trials* **19**, 61–109.

[75] Writing Group for the Women's Health Initiative Investigators (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial, *Journal of the American Medical Association* **288**, 321–333.

[76] Wu, M., Fisher, M. & DeMets, D. (1980). Sample sizes for long-term medical trials with time-dependent dropout and event rates, *Controlled Clinical Trials* **1**, 109–120.

[77] Zucker, D.M., Lakatos, E., Webber, L.S., Murray, D.M., McKinlay, S.M., Feldman, H.A., Kelder, S.H., Nader and Nader, P.R. (1995). Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH): implications of cluster randomization, *Controlled Clinical Trials* **16**, 96–118.

[78] Zukel, W.J. (1983). Evolution and funding of the coronary drug project, *Controlled Clinical Trials* **4**, 281–298.

NANCY L. GELLER &
LAWRENCE M. FRIEDMAN

Cooperative Studies Program, US Department of Veterans Affairs

The Department of Veterans Affairs (VA) is in a unique position in the US, and perhaps the world, in conducting **multicenter clinical trials**. This is due to several factors: (1) its network of 172 medical centers geographically dispersed throughout the country, under one administrative system; (2) a dedicated group of talented physicians and other health professionals serving at these medical centers; (3) a loyal and compliant patient population of nearly four million veterans; (4) a system of experienced coordinating centers that provide biostatistical, data processing, pharmacy and administrative support; and (5) a research service that recognizes the uniqueness and importance of the program and strongly supports its mission. The VA has conducted multicenter clinical trials for more than half a century, beginning with its first trial, which was organized in 1945 to evaluate the safety and efficacy of chemotherapeutic agents for tuberculosis. This article describes the history of the program, its organization and operating procedures, some of its noteworthy trials, and current challenges and opportunities.

History of the Cooperative Studies Program (CSP)

The first cooperative clinical trial conducted by the VA was a joint study with the US Armed Forces to evaluate the safety and efficacy of chemotherapeutic agents for tuberculosis. Drs John B. Barnwell and Arthur M. Walker initiated a clinical trial to evaluate various drugs in the treatment of tuberculosis, including the antibiotic streptomycin [3, 48]. The challenge of caring for 10 000 veterans suffering from the disease following World War II was the impetus for the study. Not only did the results revolutionize the treatment of tuberculosis, they also led to the development of an innovative method for testing the effectiveness of new therapies – the cooperative clinical trial.

A VA Program for conducting cooperative studies in psychiatry was started in 1955 and supported by a newly developed Central Neuropsychiatric Research Laboratory at the Perry Point, Maryland VA Medical Center (VAMC). This Program emphasized the design and conduct of randomized trials for the treatment of chronic schizophrenia. Trials were completed evaluating the efficacy of prefrontal lobotomy [2], chlorpromazine and promazine [8], phenothiazine derivatives [10], other psychotropic drugs [9, 31], the reduction or discontinuation of medication [6], the combination of medication and group psychotherapy [20], brief hospitalization and aftercare [7], the need for long-term use of antiparkinsonian drugs [30], and intermittent pharmacotherapy [43].

Noteworthy VA cooperative clinical trials in other disease areas were started in the late 1950s and 1960s. A VA cooperative study group on hypertension was started in the 1950s (and still exists today). This group was the first to show that antihypertensive drug therapy reduces the long-term morbidity and mortality in patients with severe [54] and moderate [55] elevations of blood pressure. Other areas researched by the early VA cooperative studies included: use of long-term anticoagulants after myocardial infarction; lipid lowering drugs to prevent myocardial and cerebral infarction; treatment of gastric ulcer disease; efficacy of gamma globulin in posttransfusion hepatitis; analgesics to reduce postoperative pain; surgical treatment of coronary artery disease; the effect of portal caval shunt in esophageal varices; and the effects of radical prostatectomy, estrogens, and orchiectomy in the treatment of prostate cancer.

In 1962, the VA developed a concept, novel in Federal Government medical research programs at that time, of providing its investigators access to techniques and specialized help and information essential to their research. Four regional research support centers were established: the Eastern Research Support Center at the West Haven, CT VAMC; the Midwest Research Support Center at the Hines, IL VAMC; the Southern Research Support Center at the Little Rock, AR VAMC; and the Western Research Support Center at the Sepulveda, CA VAMC (*see* **Data Management and Coordination**). Individual investigators were assisted in such areas as research design, statistical methods, data management, computer programming, and biomedical engineering. The early VA cooperative studies were coordinated by VA Central Office staff in Washington, DC, by these regional research support

centers, and by contracts with university coordinating centers. The program was led by Mr Lawrence Shaw.

Beginning in 1972, a special emphasis was placed on the CSP in the VA's Medical Research Service and its budget was quadrupled over the next decade. Under the leadership of James Hagans, MD, PhD, the program's current organization and structure were developed and codified in the *Cooperative Studies Program Guidelines* [1]. This included the establishment of four statistical/data processing/management coordinating centers and a research pharmacy coordinating center solely dedicated to conducting cooperative studies; central human rights committees attached to each of the statistical coordinating centers; a standing central evaluation committee for the review of all new proposals for VA cooperative studies and all ongoing studies every three years; and clearly defined procedures for the planning, implementation, conduct, and closeout of all VA cooperative studies. The Central Neuropsychiatric Research Laboratory at the Perry Point, MD VAMC; the Eastern Research Support Center at the West Haven, CT VAMC; the Midwest Research Support Center at the Hines, IL VAMC; and a new center at the Palo Alto, CA VAMC were established as the four new VA Cooperative Studies Program Coordinating Center (CSPCCs). The Cooperative Studies Program Clinical Research Pharmacy Coordinating Center (CSPCRPCC) was established at the Washington, DC VAMC, but later relocated to the Albuquerque, NM VAMC in 1977.

Daniel Deykin, MD was the first person to head simultaneously the VA research programs both in Health Services Research and Development and the CSP, from 1985 to 1996. He took advantage of this opportunity to promote the development of a series of multicenter clinical trials in the organization and delivery of health services. These trials represented unique challenges in design and conduct. Some of these trials have recently been completed and are in the process of being published [25, 44, 57].

In 1996, John Feussner, MD, MPH was appointed as the VA's Chief Research & Development Officer, and simultaneously assumed leadership of the CSP. Up until the time Dr Feussner was appointed, the VA Research Service was composed of three major research programs – Medical Research (of which the CSP was a part), Rehabilitation Research & Development, and Health Services Research & Development. Dr Feussner moved the CSP out of the VA Medical Research Service and elevated the Program to an equal level with the three other major VA research programs. New emphases brought to the Program by Dr Feussner include: initiation of a strategic planning process; more integration and interdependence of the coordinating centers; institution of good clinical practices and standard operating procedures at the coordinating centers; pharmaceutical manufacturing; experimentation to improve the process of informed consent [32]; educational programs in clinical research for VA investigators; partnering with industry, National Institutes of Health (NIH), and international clinical trials groups; the development of three new Epidemiology Research and Information Centers at the VAMCs in Seattle, WA, Boston, MA, and Durham, NC [5]; and Intranet and Internet communications. The strategic planning process initiated in 1997 defined the vision, mission, and specific goals for the Program (Table 1).

Organization and Functioning of the CSP

This section describes how a VA cooperative study evolves and the support provided by the VACSP.

Table 1 Vision/mission/goals of the VACSP

Vision
- The CSP is a premier research program conducting multicenter studies with world-wide impact on health care

Mission
- To advance the health and care of veterans through education, training, and collaborative research studies that produce innovative and effective solutions to national healthcare problems

Goals
- To enhance the proficiency of CSP staff and CSP partners (chairpersons, participating investigators) in the conduct of multicenter trials
- To enhance the consistency of management support for the CSP
- To increase the flow of new research ideas for cooperative studies
- To increase the application of research products into clinical practice
- To enhance the interdependence of the CSP coordinating centers
- To improve the capabilities of dissemination of research findings

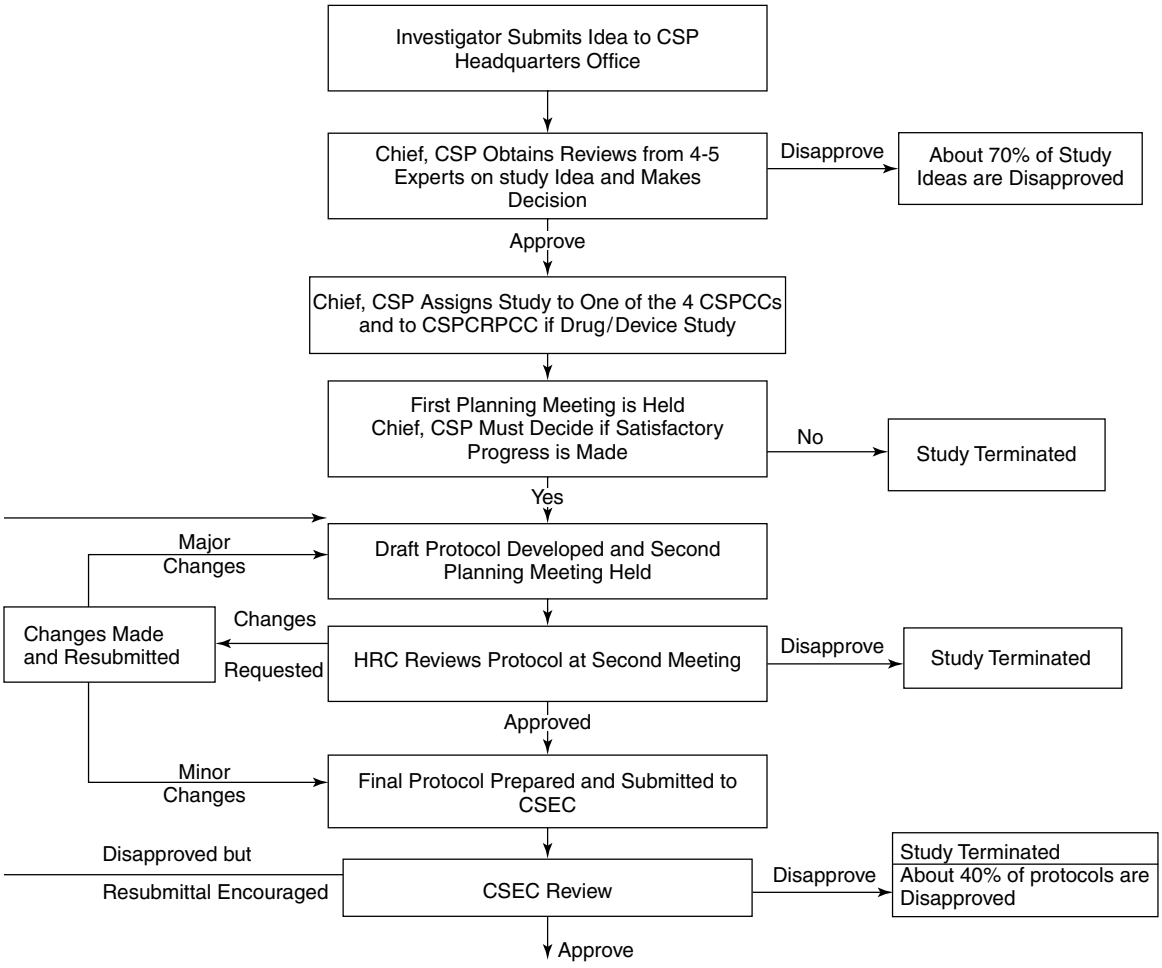

Figure 1 Development of a VA cooperative study

These aspects of the Program have been reported previously [24, 27, 28].

Planning Request

Initiation of a planning request through the evaluation phase is outlined in Figure 1. The VA Research Program, including the CSP, involves strictly intramural research. To receive VA research funding, the investigator must be at least five-eighths time VA. One of the strengths of the CSP is that most of its studies are investigator-initiated. The research questions come from investigators throughout the VA health care system who are on the front lines in providing health care for veterans.

To start the process, the investigator submits to VA Headquarters a 5–10 page planning request outlining the background of the problem, the hypothesis, a brief overview of the design, and anticipated size of the study. The planning request is given a CSP number to aid in tracking the study through its evolutionary phases. The planning request is sent to four or five independent experts in the field who initially judge the importance and feasibility of the study. If this review is sufficiently positive, the study is put into planning and assigned to one of the CSPCCs (and the CSPCRPCC if it involves drugs or devices) for development of the full proposal.

This process has evolved to satisfy two important needs. First, the CSP recognizes that the ability to

come up with a good idea needing rigorous test does not necessarily carry with it the ability to pull together all the expertise necessary to plan a clinical trial. This was especially true in the early days, when "trialists" were few and far between, and clinical researchers seldom had training in modern statistical trials design. So it is important to provide access to this expertise early in the planning process. However, such aid is expensive and scarce, so it is important not to waste it on ideas that do not show promise. Thus, the second need is for an initial concept review. This has proved to be a very efficient allocation method; about 70% of all initial proposals are not approved to go on to planning, and of the surviving 30%, about two-thirds complete the planning process. Of those that are successfully planned, about three-quarters are approved and funded. Thus, the method helps to avoid the problem of insufficiently developed **protocols**, while conserving the scarce resources of planning.

Planning Phase

Once the study is approved for planning, the resources of the CSPCCs are applied to the development of the full proposal. Within the coordinating centers, the study is assigned to a specific biostatistician and clinical research pharmacist. These individuals work with the principal proponent in nominating a planning committee which is reviewed and approved by the CSPCC and CSP Directors. The planning committee generally consists of the principal proponent, study biostatistician, study clinical research pharmacist, CSPCC Director, two or three potential participating site investigators, and outside consultants, as needed. The planning committee is funded for two planning meetings. At the first meeting, the basic design features of the study are agreed upon (hypothesis, patient population, treatment groups, primary and secondary endpoints, pharmacologic and drug handling issues, baseline and follow-up data collection and frequency, treatment effect size, sample size, number of sites, duration of study, publication plan, and budget). The full proposal is then written, and at the second planning meeting a draft of the protocol is fine-tuned. Development of a full proposal generally requires six to nine months.

Evaluation Phase

The completed proposal is first reviewed by the Human Rights Committee (HRC) attached to the CSPCC. This committee is comprised of scientists and laypeople from the community who review proposals for all new VA cooperative studies and all ongoing studies annually. The committee serves as a central Institutional Review Board (IRB) for studies assigned to the CSPCC and considers such aspects of the proposal as risks versus benefits to the patients, patient management, burdens placed on the patients from participation in the study, community equipoise with regard to the treatments being compared, and the informed consent procedures. This committee has absolute authority over approval or disapproval of the study. Only the HRC has the authority to change its own decisions. The composition of the committee follows VA regulations and is consistent with Food and Drug Administration (FDA) guidelines. Minimum membership of the HRC includes a VA chairperson, a practicing physician from the community, a nonphysician scientist, a veteran representative, a member of a recognized minority group, a clergyman or ethicist, and an attorney.

If the HRC approves the study, then the proposal is submitted to VA Headquarters for scientific review and a funding decision. The proposal is sent to four or five experts in the field and a biostatistician for written reviews. All cooperative study proposals are reviewed by a standing scientific review committee, called the Cooperative Studies Evaluation Committee (CSEC). This committee is composed of senior physician scientists and biostatisticians who have had extensive experience in cooperative studies and clinical trials. The CSEC meets in the spring and fall of each year. The principal proponent and study biostatistician present the study to the CSEC in person (reverse site visit), defend their proposal and answer questions from the CSEC members. The CSEC then goes into executive session, and decides to recommend approval or disapproval of the study and, for approvals, gives a scientific priority score, ranging from 10 to 50 with 10 being the best score. The final funding decision is made by the CSP Director. The advantages of this review process are that the study investigators have the opportunity to interact personally with the review body to answer their criticisms and concerns, and the final decision is known immediately following the review and executive session.

Implementation of the Trial

Implementation and conduct of a VA cooperative trial are outlined in Figure 2. Once the CSP Director

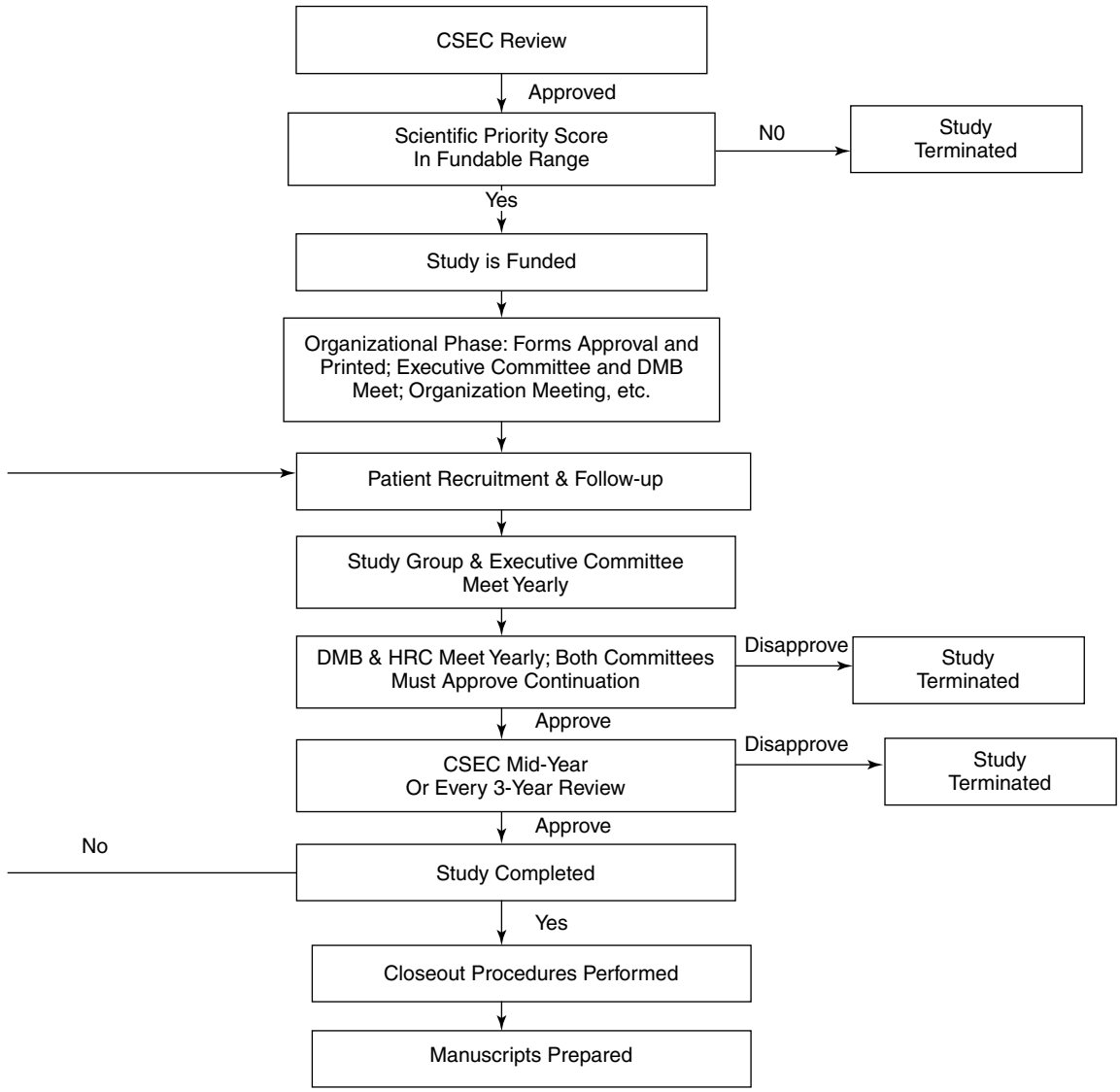

Figure 2 Conduct of a VA cooperative study

approves funding, the implementation phase of the cooperative study begins. All activities in this process are closely coordinated by the CSPCC, CSPCRPCC and the Study Chairperson's Office.

The necessity for carefully controlled medical treatment and data collection procedures for the successful conduct of multicenter clinical trials is well recognized. Because of its administrative structure, the VA provides an environment that is uniquely suited to this type of research. Each participating facility is funded by one control point for the entire period of the study, and the VAMC system provides a structure in which a relatively high degree of medical, scientific, and administrative control can be exercised. This same degree of control is often more difficult to obtain in studies that involve participating sites from different administrative systems [27].

The CSPCC recommends funding levels and monitors the performance of the individual medical centers. This information is reviewed regularly by the

study biostatistician, center director, the Executive Committee, and at least annually by the **Data and Safety Monitoring Board**. This integrated monitoring of scientific, biostatistical, and administrative aspects by the CSPCC provides a comprehensive approach to the management of multicenter clinical trials, in contrast to other clinical trials biostatistics groups that are responsible only for the analytical and data processing aspects and exercise no administrative control [27].

The Research and Development Committee and the IRB of each participating medical center must review and approve a cooperative study before it can be implemented at that facility. They are able to make modifications to the prototype consent form approved by the CSPCC HRC, but all local modifications must be reviewed and approved by the CSPCC.

Included in the implementation component of a cooperative study is the establishment of the Executive Committee and Data and Safety Monitoring Board (DSMB) who share responsibility for the conduct and monitoring of the cooperative study in the ongoing phase. The Executive Committee, which often includes several members of the original Planning Committee, consists of the study chairperson who heads the committee, the study biostatistician, the clinical research pharmacist, two or three participating investigators, and one or two consultants. This committee is responsible for the detailed operational aspects of the study during its ongoing phase and ensures adherence to the study protocol, including aspects relating to patient recruitment, treatment, laboratories, data collection and submission, biostatistical analysis, data processing, subprotocols and reporting. The Executive Committee sometimes recommends probation or termination of sites whose performance is poor.

The DSMB consists of five to eight individuals who have not been involved in the planning and development of the proposal, and includes one or two biostatisticians and two or more subject-matter experts in the field(s) of the cooperative study. This committee is charged with the responsibility of monitoring and determining the course of the ongoing study and considers such aspects as patient accrual; performance of the participating sites, CSPCC, and chairperson's office; and safety and efficacy data. Perhaps a unique feature of the CSP is that the CSPCC HRC also reviews each ongoing study annually and receives the same data reports as presented to the DSMB. The study chairperson, participating site investigators, and other members of the Executive Committee are masked to the outcome data during the course of the study. Only the DSMB, HRC, CSPCC and CSPCRPCC see the outcome data during the conduct of the study.

The fourth body involved in the conduct of the cooperative study is the Study Group, which consists of all participating investigators, the study chairperson (co-chairpersons), biostatistician, clinical research pharmacist, and consultants. This body meets once annually to consider the progress of the study and to resolve problems that may arise at the participating centers.

Within the CSPCC, the biostatistician heads a team of administrative, programming and data management personnel that provides regular monitoring of the study. This team develops an operations manual (*see* **Clinical Trials Protocols**), in conjunction with the chairperson's office, to train study personnel in the day-to-day conduct of the trial. They also develop a computer data management system to edit, clean, and manage the study data. Automated query reports are generated by the computer system and sent to the participating sites for data checking and cleaning. Statistical progress reports are published by the CSPCC and distributed to the Study Group, Executive Committee, and DSMB at scheduled meetings.

An initial kickoff meeting is held before the study starts to train site personnel in the conduct of the study. Annual meetings are held thereafter to refresh training and discuss issues in the conduct of the study. Frequent conference calls of the study committees are also used to facilitate communication and training.

Another unique aspect of the CSP is the CRPCC, which operationalizes the pharmaceutical aspects of the clinical trials (Table 2). In the planning stages of the study the clinical research pharmacist designs the drug or device treatment and handling protocol and works with the study chairperson and pharmaceutical and device industries to purchase or obtain donations of clinical supplies for the study. The CRPCC coordinates the development of appropriate drug dosage formulations and the manufacture of study drugs or devices. In the event that drug donations are not possible, the CRPCC has the expertise and capability to provide the in-house manufacture of active drugs and matching placebo. Drugs for all cooperative studies

Table 2 Unique functions and roles of the CSP pharmacy coordinating center

- Design of a drug or device handling protocol for each study involving drugs or devices
- Preparation and submission of INDAs or IDEs
- Obtaining donations or purchase of clinical supplies for study
- Coordination of appropriate drug dosage formulations and manufacture of study drugs or devices
- Quality control testing of drugs
- Development of blinding methods
- Storage, packaging and shipment of clinical supplies to pharmacies at the participating sites
- Computerized drug inventory and information system to track and replenish supplies at site pharmacies
- Monitoring adverse medical events and reporting to appropriate authorities
- Monitoring, auditing and education services to ensure sites are in compliance with GCP
- Preparation of final drug/device accountability reports

must pass the testing of the CRPCC's quality control testing laboratory. The CRPCC also assesses study product **blinding** methods.

At the CRPCC, study medications are stored in an electronically controlled and secured environment. The CRPCC customizes labels and packages all study medications, which are centrally distributed to the pharmacies at the participating sites. In doing so, the CRPCC provides a computerized drug inventory and information system for complete accountability of clinical supplies. This includes automated study supply tracking and replenishment systems for maintaining adequate study supplies at participating sites as well as automated telephone randomization and drug assignment systems. The clinical research pharmacist is then able to direct and monitor the study prescribing and dispensing activities as well as to monitor the **compliance** with the study protocol treatments at the participating sites. At the end of the study the CRPCC directs the retrieval and disposition of unused clinical supplies and prepares a final drug/device accountability report.

The clinical research pharmacist also works closely with the study chairperson and the manufacturers to prepare, submit, and maintain Investigational New Drug Application (INDAs) or Investigational Device Exemption (IDEs), which includes preparing and submitting annual and special reports to the FDA. Along with this responsibility, the clinical research pharmacist coordinates the monitoring and reporting of all adverse medical events to study management, FDA and associated manufacturers. Recently the CRPCC established a central Good Clinical Practices (GCP) Assurance Program. The Program provides monitoring, auditing, and educational services for all VA cooperative studies to ensure that the participating sites are in GCP compliance. If needed, the Program is capable of providing full GCP monitoring for studies under regulatory (FDA) scrutiny.

Final Analysis and Publication Phase

Upon completion of patient intake and follow-up, the study enters the final analysis and publication phase. If the Executive Committee, the CSPCC, and the study biostatistician have performed their tasks well, this phase should be quite straightforward. It requires an updating of all study files and the processing and analysis of the complete data set. The interim statistical analyses that were run during the ongoing phase of the study are now executed on the complete data. In addition, some analyses may point to additional questions that would be of interest; however, it is anticipated that the majority of final analyses and interpretation of results will occur within 6 to 12 months after study termination. All publications emanating from the cooperative study must be approved by the Executive Committee. Although the responsibility of the DSMB terminates at the end of data collection, the Board is at times requested to review manuscripts and give advice prior to submission for publication [27].

Usually each trial generates a number of manuscripts. The Executive Committee establishes priorities for statistical analyses and manuscript development and appoints writing committees composed of members of the Executive Committee and Study Group for each paper. Authorship of the main paper(s) usually consists of the chairperson, study biostatistician, study clinical research pharmacist, members of the Executive Committee, and, in some cases, the participating site investigators. Secondary papers are often written by other members of the Executive Committee and site investigators.

The CSPCC serves as the final data repository for the study. The study database, protocol, operations manuals, forms, study correspondence and interim and final statistical progress reports are archived at the CSPCC.

Role of the Biostatistician and Pharmacist in the CSP

One of the unique features of the CSP is that the biostatistician at the CSPCC plays a major organizational, fiscal, and administrative role, in addition to the usual technical role. In recent times, as the administration of studies has become more complex, the biostatistician may be assisted by a study coordinator but, as in the past, the greater part of the burden of management falls on the biostatistician. In contrast to the pharmaceutical industry and to many Contract Research Organization (CROs), the biostatistician is responsible for monitoring site adherence to protocol, recruitment, and many other aspects of the study conduct. In addition, the study pharmacist plays a key role in monitoring adverse effects, maintaining supplies of the study drug, regulatory reporting, and the like. In a sense, the study team is deployed to support the investigators, but has independent authority and responsibility as well.

One of the strengths of this approach to study management is that it is possible to guarantee some degree of uniformity in the conduct of the studies, independent of the varying managerial skills and style of the study chairs. The biostatistician and pharmacist, together with the coordinating centers of which they are a part, provide institutional memory and continuity. Their central position on the study teams reinforces the key idea that the studies mounted by the VACSP are the joint responsibility of the program and the investigators. Such an intramural program can only succeed on a limited budget if issues of cost and complexity are kept to the forefront during the planning process. A consequence that is easily observed is that the typical CSP trial is a lean, focused attack on a single important clinical question, rather than a broad-based research project with many interwoven strands of investigation.

In contrast to the much larger NIH clinical trials efforts, which are organized along disease lines, the CSP biostatisticians and CSPCCs are generalists, doing studies in all areas of medicine with relevance to the VA. Along the way, some centers have developed some special experience in certain areas, but there has never been a "heart" center or a "cancer" center. Because the CSP has such a broad medical purview, but a relatively low volume of studies, it has not made economic sense to specialize. The scarce resource of statistical and data management expertise has needed to be allocated efficiently to support the proposals that were emerging from the field. Since VA resources have followed the strength of the proposals rather than disease areas, the CSPCCs have not specialized to any large degree.

While there are undoubted advantages to specialization, as shown by the contributions made by the National Cancer Institute (NCI) (*see* **Cooperative Cancer Trials**) and the National Heart, Lung, and Blood Institute (NHLBI) (*see* **Cooperative Heart Disease Trials**) statisticians to the statistical science of their disease areas, there are some advantages to generalizing. In particular, it has been possible to transplant methods and lessons learned from well-studied areas such as cancer and heart disease, to other areas such as psychopharmacology, device research, health services research, and trials of surgical procedures. The absence of disease-area "stovepiping" has facilitated a high general level of sophistication in the conduct of trials, with techniques travelling readily across borders.

This cross-pollination has also been facilitated by the structure of the VACSP scientific peer review. The standing committee that reviews and recommends studies for funding mixes disciplines with common expertise in multisite studies. *Ad hoc* reviewers provide the crucial discipline-specific input to the committee, but the same committee may review a heart failure trial in the morning and a psychopharmacology trial in the afternoon. The result is a high degree of uniformity in the standards for the research across disease areas, and this has been an enduring strength of the program.

Ongoing and Completed Cooperative Studies (1972–2000)

One hundred and fifty-one VA cooperative studies were completed or are currently ongoing in the period 1972–2000. Table 3 presents the health care areas

Table 3 Health care areas of ongoing and completed VA cooperative studies (1972–2000)

Health care area	Number of studies	Percent of studies
Cardiology/cardiac surgery	24	15.9
Hypertension	15	10.0
Gastrointestinal	14	9.3
Substance abuse	11	7.3
Mental health	10	6.6
Infectious diseases	9	6.0
Cancer	8	5.3
Dental	6	4.0
General surgery/anesthesia	6	4.0
Cerebrovascular	5	3.3
Peripheral vascular	5	3.3
Military service effects	4	2.6
Ambulatory care	4	2.6
Epilepsy	4	2.6
Genitourinary	4	2.6
Diabetes	3	2.0
Renal	3	2.0
Sleep	3	2.0
Pulmonary	2	1.3
Hematology	2	1.3
Hearing	2	1.3
One each in seven areas[a]	7	4.7
Total	151	100.0

[a]Analgesics, arthritis, geriatrics, hospital-based home care, laboratory quality control, computerized neuropsychological testing, ophthalmology

of these studies. These areas are generally reflective of the major health problems of the US veteran population, consisting mainly of middle-aged and senior adult males. Studies in cardiology and cardiac surgery represent 15.9% of the 151 studies, followed by hypertension (10.0%), gastrointestinal diseases (9.3%), substance abuse (7.3%), mental health (6.6%), infectious diseases (6.0%), and cancer (5.3%).

There are a few notable disease areas that are prevalent in the VA population and yet might be considered underrepresented in the CSP. These include diabetes (2.0%), renal diseases (2.0%), pulmonary diseases (1.3%), hearing diseases (1.3%), arthritis (0.7%), and ophthalmologic diseases (0.7%). Because the CSP mainly relies on investigator-initiated studies, the conclusion might be drawn that these subspecialties have underutilized the Program. Although studies on effects of military service represent only 2.6% of the 151 studies, studies listed in other categories have investigated treatments for service-connected illnesses (e.g. posttraumatic stress disorder studies are categorized under mental health, and the substance abuse studies could be considered consequences of military service).

Table 4 briefly summarizes some of the noteworthy VA cooperative clinical trials that were completed in the 1980s and 1990s. Many of these trials resulted in advances in clinical medicine that could immediately be applied to improve the health care of US veterans and the US population in general.

Current Challenges and Opportunities

Although the VACSP has had numerous past successes, it faces many challenges and opportunities in the future. These include: (1) changes in the VA health care system and their effects on research; (2) nationwide concerns about violations of patients' rights in research; (3) increasing the efficiency and interdependence among the coordinating centers and standardizing procedures; (4) ensuring the adequacy of flow of research ideas and training of investigators; and (5) partnering with industry, other federal agencies, nonprofit organizations, and international clinical trial groups to enhance the capacity of the Program.

Changes in the VA Health Care System

The VA health care system has been undergoing substantial changes that could adversely affect research. In 1996, the VA reorganized into 22 geographically defined Veterans Integrated Service Networks (VISNs). Much of the central authority, decision-making, and budgeting once performed in VA Headquarters in Washington, DC, has been delegated to the 22 VISN offices. Within the VISNs, administrative and health care services and in some cases entire VAMCs are being consolidated. The largest component of the VA patient population, the World War II veterans, is rapidly declining. Health care personnel in some VISNs are experiencing reductions in force, with the result that the remaining personnel have limited time to devote to research. These factors may already be adversely affecting the Program's ability to meet recruitment goals in ongoing trials [23].

Concerns About Patients' Rights in Research

The nature of the veteran population treated at VA hospitals raises some special issues in human rights

Table 4 Noteworthy VA cooperative studies

- 80% of strokes in patients with atrial fibrillation can be prevented with low-dose warfarin [15]
- Carotid endarterectomy is effective in preventing strokes in symptomatic and transient ischemic attacks in asymptomatic patients [26, 39]
- Aggressive treatment of moderate hypertension works well in elderly patients [19, 37]
- Age and racial groupings can be used to optimize selection of first line drugs in hypertension [38]
- Coronary artery bypass surgery prevents mortality in patients with left main disease and in high-risk patients without left main disease [42, 49]
- Low dose aspirin reduces heart attacks and death in 50% of patients with unstable angina [34]
- Vasodilators and angiotensin converting enzyme inhibitors prevent deaths in patients with congestive heart failure [12, 13]
- Low dose aspirin started 6 hours after coronary artery bypass surgery and continued for one year prevents the occlusion of the bypass grafts [17, 18]
- Mechanical artificial aortic heart valves prolong survival more than bioprosthetic aortic heart valves [22]
- A conservative, ischemia-guided strategy is safe and effective for management of patients with non-Q-wave myocardial infarction [4]
- Digoxin does not reduce mortality but does reduce hospitalizations in patients with congestive heart failure [50]
- The rate of coronary events (myocardial infarction or death) in men with coronary heart disease can be reduced by 22% with Gemfibrozil therapy, which increases high density lipoprotein cholesterol and lowers triglyceride levels [45]
- Progression of human immunodeficiency virus (HIV) infection to full blown acquired immune deficiency syndrome (AIDS) can be delayed with the drug zidovudine [21]
- Steroid therapy does not improve survival of patients with septic shock [52]
- Patients with advanced laryngeal cancer can be treated with larynx-sparing chemotherapy and radiation compared with standard surgical removal of the larynx and have equivalent long-term survival [14]
- The drug Terazosin is more effective than Finasteride in relieving the symptoms of benign prostatic hyperplasia [33]. Transurethral resection of the prostate is an effective operation, but Watchful Waiting can be effective in many patients [56]
- An implantable insulin pump is more effective than multiple daily insulin injections in reducing hypoglycemic side-effects, and enhancing quality of life in adult-onset Type II diabetes mellitus [46]
- Multi-channel are superior to single-channel cochlear implants in restoring hearing to patients with profound hearing loss [11]
- Sclerotherapy is an effective treatment for esophageal varices in patients who have had prior bleeds but not in patients without prior bleeds [51]
- Antireflux surgery is more effective than medical therapy in patients with complicated gastroesophageal reflux disease [47]
- Severely malnourished patients benefit from pre-operative total parenteral nutrition but mildly malnourished patients do not [53]
- Clozapine is a cost-effective treatment for patients with refractory schizophrenia who have high hospital use [44]
- Erythropoietin administered subcutaneously compared with intravenously can significantly reduce the costs of hemodialysis [29]
- Use of intrapleural tetracycline reduces recurrence rate by 39% in patients with spontaneous pneumothorax [35]
- Rapid access to high quality primary care for patients with severe chronic illnesses greatly improves patient satisfaction with care but may lead to an increase in hospital readmissions [57]
- Levomethadyl acetate (LAAM) is a safe and efficacious drug to use for heroin addiction. Studies were used to gain FDA approval for LAAM as treatment for heroin addiction [16, 36]
- Systemic corticosteroids improve clinical outcomes up to three months in patients with chronic obstructive pulmonary disease [40]

protections (*see* **Medical Ethics and Statistics**). The VA treats about four million veterans, who tend to be less well off than the average veteran (or the average citizen). They are often more severely ill than non-VA patients with the same age and diagnosis, and often have multiple co-morbidities. They are on average more dependent on the VA for their health care than the typical non-VA patient is on his or her usual health care provider. Against this background we note the extraordinary willingness of the veteran patient to engage in research, trusting the clinical researcher to an astonishing degree. Such trust demands an extraordinary level of protection in response.

The CSP has instituted a unique framework of human subjects' protections, going beyond the usual procedures that other federal sponsors and drug companies require. This begins in the planning stage, when each proposal must undergo a rigorous review by the HRC attached to the coordinating center. It typically meets for several hours over a single protocol, reviewing it in fine detail. The protocol cannot go forward without their independent approval.

The CSP also requires the usual individual site IRB approval, and other reviews that are mandated at the local site, before a study can start at a site. The ongoing IRB reviews at the local sites (annually, or more often, as stipulated in the initial review) are monitored by the CSP staff. As has become standard in multisite trials, each CSP study has its own independent DSMB that meets at least annually to review the progress of the study.

The unique CSP innovation to this process is the joint review by the HRC and DSMB. Thus, after every DSMB meeting, the two groups meet to review and recommend, with the same basis of information on study progress. The CSP has found that the HRC is able to hear the recommendation of the DSMB, which is typically heavily weighted with subject-matter expertise, and interpret it in the light of the other perspectives they bring. The CSP believes that this has been a successful experiment in resolving the knotty issue of how to obtain full and informed ongoing review of studies where investigators are kept blind, and site-level information must be far less informative than the big picture presented to the DSMB. We believe that such joint reviews add considerably to the level of protection of human subjects.

In addition, members of the central HRCs conduct three site visits per year during which patients are interviewed about their participation in the trials. Thus, the Program as a whole conducts 12 such visits per year. The Albuquerque auditing group periodically site visits VAMCs participating in cooperative studies and performs audits to ensure that the sites are complying with GCP guidelines. The CSPCCs also receive copies of consent forms from all patients in all of the trials as further evidence of proper consent procedures.

The VA recently established its own office to oversee the protection of patients' rights in VA research, performing functions similar to those of the Office of Protection from Research Risks (OPRR) of the Department of Health and Human Services. IRBs at VAMCs currently are required to be accredited by an external, non-VA entity.

In addition to these standard procedures, followed in all studies, the CSP has recognized two other areas of human subjects' protection in which it can make a contribution. The Enhancing the Quality of Informed Consent (EQUIC) program [32] is designed to institutionalize the process of testing innovations in methods for obtaining informed consent. It piggybacks tests of new methods on ongoing CSP studies, and provides a centralized assessment of the quality of informed consent encounters (by remote telephone interview of patients). In the spirit of EQUIC, a substudy is being conducted in one ongoing VA cooperative study to evaluate the utility of an informed consent document developed by a focus group of subjects eligible for the trial [41].

The second topic that the CSP has engaged are the ethical, legal, and social implications of genetics research, specifically of deoxyribonucleic acid (DNA) banking with linked clinical (phenotype) data. The CSP has begun a project to provide uniform methods for obtaining and banking such samples.

Steps to ensure human subjects' protection in VA cooperative studies are listed in Table 5.

Efficiency and Interdependence of the CSPCCs

The VACSP recently contracted with an outside vendor to help develop standard operating procedure (SOPs) for the CSPCCs. Twenty-two SOPs were developed in the areas of administration, planning and implementing clinical trials, data management,

Table 5 Steps to ensure human subjects' protection in the VACSP

- Investigator's integrity
- Development of proposal through collaboration between investigators and CSPCCs
- HRC review of proposal initially
- Site Monitoring and Review Team (SMART) audit of consent form contents
- CSEC review of proposal
- Initial review of proposal by participating site R&D and IRB
- Annual central reviews of trial by DSMB and HRC
- Annual reviews of study by local R&D committee and IRB
- SMART audit of participating sites
- HRC site visits and interviews of study patients
- Receipt of copies of patient consent forms by CSPCC, local research offices, and local pharmacies
- Implementation of SOPs and good clinical practices
- Compliance with all FDA and VA regulations
- Innovative studies on improving informed consent

Table 6 Recently adopted SOPs for the VACSP

Administration

- Preparing, issuing and updating SOPs
- Training and training documentation

Planning and implementation of clinical trials

- Developing, approving and amending protocols
- Study/training meetings
- Preparing and approving operations manuals
- Study initiation
- Developing and approving case report forms
- Creating and validating data entry screens and programs
- Preparing, documenting and validating data checking programs
- Preparing, documenting and validating statistical programs
- Developing and conducting statistical analyses

Handling data from clinical trials

- Randomization, blinding and unblinding
- Central monitoring
- Case report form flow and tracking
- Data entry and verification
- Data cleaning
- Reporting adverse events

Study closeout

- Archiving study documentation
- Study closeout

Study oversight

- Assuring site R&D and IRB approvals
- DSMB
- HRC

study closeout, and study oversight (Table 6). By standardizing among and within the coordinating centers certain procedures that are performed in every study, we will achieve an even higher level of support to all studies more efficiently than previously done. The SOPs will also enable the CSPCCs to be in better compliance with GCP principles and International Conference on Harmonization (ICH) **guidelines**.

Since 1996, the Directors of the Program and centers have been meeting as a group semiannually to identify current and future challenges and opportunities, and to develop annual strategic plans to respond to these challenges and opportunities. This has enhanced the development of mutual projects which the centers can work on together to further the goals of the organization as a whole, such as the development of a Clinical Research Methods Course, a one-year sabbatical program for clinical investigators to enhance their training and skills, and SOPs for the central HRCs.

Ensuring the Adequacy of Flow of Ideas and Training of Investigators

In recent years, the CSP has developed several educational opportunities to help train VA investigators in clinical research and to encourage utilization of the Program to answer important clinical questions. These include a five-day course in clinical trials and sabbatical and career development programs focused on clinical research methodology.

The five-day course is taught once each year and involves 10 faculty members (two from each of the five coordinating centers) and 60 VA investigators selected from applications from throughout the country. The course consists of 15 lecture/discussion sessions on various aspects of designing a clinical trial, interspersed with breakout sessions during which the students are divided into five planning committees to design a clinical trial. On the last day of the course, the student groups take turns in presenting their clinical trials and receiving critiques from the audience. The course has been taught twice and has received excellent feedback from the students.

The CSP Career Development Program provides protected time to clinician–investigators for a period of concentrated clinical research activity. The objective is to build capacity in a wide geographic distribution for the Department of Veterans Affairs to conduct clinical research in acute-care hospitals, long-term care facilities, or outpatient settings. The Program is designed to foster the research careers of clinician–scientists who are not yet fully independent but who seek to become independent clinical researchers. The award provides three years of salary and some supplemental research support, and the awardees are expected to work at least part of the time at one of the five CSPCCs or three Epidemiology Research and Information Center (ERICs).

In 1999 CSP announced a sabbatical program for established clinician–scientists to train at one of the CSPCCs or ERICs for up to one year. The purpose of the program is to support clinician–investigators who wish to secure training time to learn about the conduct of cooperative studies and epidemiologic research.

Partnering with Outside Organizations

The VACSP has partnered with NIH and industry for many years in conducting multicenter clinical trials. In recent years, a special emphasis has been placed on partnering with outside agencies to enhance the effect of the limited VA research funding, and these efforts have been fruitful.

Recent examples of this partnering include: the Digitalis in Heart Failure Trial, sponsored by the VA, NHLBI, and Burroughs–Wellcome Company and conducted in 302 VA and non-VA sites in the US and Canada; a series of trials sponsored by the VA and the National Institute of Drug Abuse (NIDA) to evaluate new treatments for drug abuse; the Prostate Cancer Intervention Versus Observation Trial (PIVOT), sponsored by the VA, Agency for Healthcare Quality and Research (AHQR) and NCI; the Beta-Blocker Evaluation of Survival Trial (BEST), funded by the VA, NHLBI, and industry; the Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation (COURAGE) trial, supported by the VA and 10 pharmaceutical companies; the VA/National Institute of Deafness and Other Communication Disorders (NIDCD) Hearing Aid Clinical Trial; and the Shingles Prevention Study sponsored by the VA, NIH, and a pharmaceutical company.

The VACSP has been working with the American College of Surgeons to promote clinical trials evaluating new surgical operations and technologies. This collaboration has resulted in a VA trial comparing the outcomes of laparoscopic vs. open tension-free inguinal hernia repair, a trial comparing open tension-free hernia repair vs. watchful waiting funded by AHQR, and a trial comparing pallidotomy vs. deep brain stimulation in Parkinson's Disease.

The VACSP has also issued a program announcement for the development of multinational clinical trials between the VA and the Medical Research Councils of Canada and the UK. As the field of clinical trials matures, it is likely that achievable treatment effect sizes will decrease, necessitating larger and larger trials, or "mega" trials. These types of collaborations will be important in the future, as the larger trials will exceed the capacity of any single clinical trials program.

Concluding Remarks

In summary, we believe that there are considerable strengths to conducting multicenter clinical trials in

Table 7 Strengths of the VACSP

Related to VA health care system

- Large veteran population willing to participate in research, well-represented by minority groups
- Largest integrated healthcare system in US with 172 medical centers under single administrative system
- High-quality physician–investigators
- National administrative databases that allow for tracking of patients
- Supportive management in VA Headquarters
- System of local research offices and IRBs at participating sites that facilitate multicenter research

Related to the CSP

- Quality and experience of the coordinating centers
- Well-established mechanisms for conducting multi-site trials
- Planning process usually produces tightly focused, cost-effective protocols
- Rigorous review process by HRC and CSEC
- Guidelines and SOPs for conducting trials
- Multiple levels of protection of research subjects
- Ability to conduct trials with high power and generalizability, so the impact on changing health care practices is maximized compared with other research programs

Table 8 Limitations of the VACSP

Related to VA health care system

- Primarily male population, limiting generalizability of results
- Large studies in female and childhood diseases are not possible
- Changes in the health care system, including aging and declining of veteran population, decentralization and consolidation of facilities
- Reduction in dedicated research time for physician–investigators

Related to the CSP

- Long duration from submission of planning request to publication of main results raises the risk of study becoming outdated
- Limitation of funding
- Limited capacity to conduct mega trials within VA system

the VA health care system, as enumerated in Table 7. There are also some acknowledged limitations of the Program, some of which can be addressed in the future (Table 8).

This article has described the history, organization and productivity of a clinical trials program designed as an integral part of a large health care system. The biostatistical and pharmacy positions in the Program are ideal from the standpoint that these people are integrally involved in the research from beginning to end and play a major role in the conduct of the trials. The Program is an example of how clinician–investigators and methodologists can work together successfully to design and conduct large-scale clinical research.

Acknowledgments

We are extremely indebted to the foresight and support of the US Congress, Executive Branch, and VA management, and to the dedication of the health care providers and veteran patients in the VA system to enable us to carry out this important research.

References

[1] Anonymous (1997). Cooperative Studies Program Guidelines for the Planning and Conduct of Cooperative Studies. Office of Research and Development, Department of Veterans Affairs, Washington, DC.

[2] Ball, J., Klett, C.J. & Gresock, C.J. (1959). The Veterans Administration study of prefrontal lobotomy, *Journal of Clinical and Experimental Psychopathology* **20**, 205–217.

[3] Barnwell, J.B., Bunn, P.A. & Walker, A.M. (1947). The effect of streptomycin upon pulmonary tuberculosis, *American Review of Tuberculosis* **56**, 485–507.

[4] Boden, W.E., O'Rourke, R.A., Crawford, M.H., et al. (1998). Outcomes in patients with acute non-Q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative management strategy, *New England Journal of Medicine* **338**, 1785–1792.

[5] Boyko, E.J., Koepsell, T.D., Gaziano, J.M., et al. (2000). U.S. Department of Veterans Affairs medical system as a resource to epidemiologists, *American Journal of Epidemiology* **151**, 307–314.

[6] Caffey, Jr, E.M., Diamond, L.S., Frank, T.V., et al. (1964). Discontinuation or reduction of chemotherapy in chronic schizophrenics, *Journal of Chronic Diseases* **17**, 347–358.

[7] Caffey, Jr, E.M., Galbrecht, C.R. & Klett, C.J. (1971). Brief hospitalization and aftercare in the treatment of schizophrenia, *Archives of General Psychiatry* **24**, 81–85.

[8] Casey, J.F., Bennett, I.F., Lindley, C.J., et al. (1961). Drug therapy in schizophrenia: a controlled study of the relative effectiveness of chlorpromazine, promazine, phenobarbital, and placebo, *Archives of General Psychiatry* **4**, 381–389.

[9] Casey, J.F., Hollister, L.E., Klett, C.J., et al. (1961). Combined drug therapy of chronic schizophrenics: a controlled evaluation of placebo, dextroamphetamine, imipramine, isocarboxazid, and trifluoperazine added to maintenance doses of chlorpromazine, *American Journal of Psychiatry* **117**, 997–1003.

[10] Casey, J.F., Lasky, J.J., Klett, C.J. & Hollister, L.E. (1960). Treatment of schizophrenic reactions with phenothiazine derivatives: A comparative study of chlorpromazine, triflupromazine, mepazine, prochlorperazine, perphenazine and phenobarbital, *American Journal of Psychiatry* **117**, 97–105.

[11] Cohen, N.L., Waltzman, S.B., Fisher, S.G., et al. (1993). A prospective, randomized study of cochlear implants, *New England Journal of Medicine* **328**, 233–237.

[12] Cohn, J.N., Archibald, D.G., et al. (1986). Effect of vasodilator therapy on mortality in chronic congestive heart failure. Results of a Veterans Administration Cooperative Study, *New England Journal of Medicine* **314**, 1547–1552.

[13] Cohn, J.N., Johnson, G., Zeische, S., et al. (1991). A comparison of enalapril with hydralazine isorbide dinitrate in the treatment of chronic congestive heart failure, *New England Journal of Medicine* **325**, 303–310.

[14] Department of Veterans Affairs Laryngeal Cancer Study Group (1991). Induction chemotherapy plus radiation compared with surgery plus radiation in patients with advanced laryngeal cancer. *New England Journal of Medicine* **324**, 1685–1690.

[15] Ezekowitz, M.D., Bridges, S.L., James, K.E., et al. (1992). Warfarin in the prevention of stroke associated

with nonrheumatic atrial fibrillation, *New England Journal of Medicine* **327**, 1406–1412.

[16] Fudala, P.J., Vocci, F., Montgomery, A. & Trachlenberg, A.I. (1997). Levomethyl acetate (LAAM) for the treatment of opioid dependence: a multisite, open-label study of LAAM safety and an evaluation of the product labeling and treatment regulations, *Journal of Maintenance in the Addictions* **1**, 9–39.

[17] Goldman, S., Copeland, J., Moritz, T., et al. (1989). Saphenous vein graft patency 1 year after coronary artery bypass surgery and effects of antiplatelet therapy. Results of a Veterans Administration Cooperative Study, *Circulation* **80**, 1190–1197.

[18] Goldman, S., Copeland, J., Moritz, T., et al. (1991). Starting aspirin therapy after operation. Effects on early graft patency. *Circulation* **84**, 520–526.

[19] Goldstein, G., Materson, B.J., Cushman, W.C., et al. (1990). Treatment of hypertension in the elderly: II. Cognitive and behavioral function. Results of a Department of Veterans Affairs Cooperative Study, *Hypertension* **15**, 361–369.

[20] Gorham, D.R., Pokorny, A.D. & Moseley, E.C. (1964). Effects of a phenothiazine and/or group psychotherapy with schizophrenics, *Diseases of the Nervous System* **25**, 77–86.

[21] Hamilton, J.D., Hartigan, P.M., Simberkoff, M.S., et al. (1992). Early versus later zidovudine therapy of patients with symptomatic human immunodeficiency virus infection: results of a randomized, double-blind VA Cooperative Study, *New England Journal of Medicine* **326**, 437–443.

[22] Hammermeister, K., Sethi, G.K., Henderson, W.G., et al. (1999). Outcomes 15 years after valve replacement with a mechanical versus a bioprosthetic valve: final report of the VA randomized trial, *Journal of the American College of Cardiology* **36**, 1152–1158.

[23] Henderson, W.G. (2000). Is it becoming more difficult to attain target sample sizes in clinical trials? Presentation at the 21st Annual Meeting of the Society for Clinical Trials. Toronto, Canada, April 16–19.

[24] Henderson, W.G. (1980). Some operational aspects of the Veterans Administration Cooperative Studies Program from 1972–1979, *Controlled Clinical Trials* **1**, 209–226.

[25] Henderson, W.G., Demakis, J., Fihn, S.D., et al. (1998). Cooperative studies in health services research in the Department of Veterans Affairs, *Controlled Clinical Trials* **19**, 134–148.

[26] Hobson, R.W., Weiss, D.G., Fields, W.S., et al. (1993). Efficacy of carotid endarterectomy for asymptomatic carotid stenosis, *New England Journal of Medicine* **328**, 221–227.

[27] James, K.E. (1980). A model for the development, conduct, and monitoring of multicenter clinical trials in the Veterans Administration, *Controlled Clinical Trials* **1**, 193–207.

[28] Kathe, B.A., Chan, Y.-K., Buehler, D.A., et al. (1981). Protection of patient rights and welfare in the VA Cooperative Studies Program, *Controlled Clinical Trials* **2**, 267–274.

[29] Kaufman, J.S., Reda, D.J., Fye, C.L., et al. (1998). Subcutaneous compared with intravenous epoetin in patients receiving hemodialysis, *New England Journal of Medicine* **339**, 578–583.

[30] Klett, C.J. & Caffey, Jr, E.M. (1972). Evaluating the long-term need for antiparkinson drugs by chronic schizophrenics, *Archives of General Psychiatry* **26**, 374–379.

[31] Lasky, J.J., Klett, C.J., Caffey, Jr, E.M., et al. (1962). Drug treatment of schizophrenic patients: a comparative evaluation of chlorpromazine, chlorprothixene, fluphenazine, reserpine, thioridazine and triflupromazine, *Diseases of the Nervous System* **23**, 698–706.

[32] Lavori, P.W., Sugarman, J., Hays, M.T. & Feussner, J.R. (1999). Improving informed consent in clinical trials: a duty to experiment, *Controlled Clinical Trials* **20**, 187–193.

[33] Lepor, H., Williford, W.O., Barry, M.J., et al. (1996). The efficacy of terazosin, finasteride, or both in benign prostatic hyperplasia, *New England Journal of Medicine* **335**, 533–539.

[34] Lewis, H.D., Davis, J.W., Archibald, D.G., et al. (1983). Protective effects of aspirin against acute myocardial infarction and death in men with unstable angina: results of a Veterans Administration Cooperative Study, *New England Journal of Medicine* **309**, 396–403.

[35] Light, R.W., O'Hara, V.S., Moritz, T.E., et al. (1990). Intrapleural tetracycline for the prevention of recurrent spontaneous pneumothorax. Results of a Department of Veterans Affairs Cooperative Study, *Journal of the American Medical Association* **264**, 2224–2230.

[36] Ling, W., Charuvastra, C.V., Kaim, S.C. & Klett, C.J. (1976). Methyl acetate and methadone as maintenance treatment for heroin addicts: a Veterans Administration cooperative study, *Archives of General Psychiatry* **33**, 709–720.

[37] Materson, B.J., Cushman, W.C., Goldstein, G., et al. (1990). Treatment of hypertension in the elderly: I. Blood pressure and clinical changes. Results of a Department of Veterans Affairs Cooperative Study, *Hypertension* **15**, 348–360.

[38] Materson, B.J., Reda, D.J., Cushman, W.C., et al. (1993). Single-drug therapy for hypertension in men. A comparison of six antihypertensive agents with placebo, *New England Journal of Medicine*, **328**, 914–921.

[39] Mayberg, M.R., Wilson, S.E., Yatsu, F., et al. (1991). Carotid endarterectomy and prevention of cerebral ischemia in symptomatic carotid stenosis, *Journal of the American Medical Association* **266**, 3289–3294.

[40] Niewoehner, D., Erbland, M.L., Deupree, R.H., et al. (1993). Effect of systemic glucocorticoids on exacerbations of chronic obstructive pulmonary disease, *New England Journal of Medicine* **340**, 1941–1947.

[41] Peduzzi, P., Guarino, P., Donta, S., et al. (2000). Design of an informed consent study to evaluate the utility of a

focus group consent document in the VA cooperative study. A randomized multicenter controlled trial of multi-modal therapy in veterans with Gulf War illness (CSP 470). Presentation at the 21st Annual Meeting of the Society for Clinical Trials. Toronto, Canada. April 16–19.

[42] Peduzzi, P., Kamina, A. & Detre, K. (1998). Twenty-two year follow-up in the VA cooperative study of coronary artery bypass surgery for stable angina, *American Journal of Cardiology* **81**, 1393–1399.

[43] Prien, R.F., Gillis, R.D. & Caffey, Jr, E.M. (1973). Intermittent pharmacotherapy in chronic schizophrenia, *Hospital and Community Psychiatry* **24**, 317–322.

[44] Rosenheck, R., Cramer, J., Xu, W., et al. (1997). A comparison of clozapine and haloperidol in hospitalized patients with refractory schizophrenia, *New England Journal of Medicine* **337**, 809–815.

[45] Rubins, H.B., Robins, S.J., Collins, D., et al. (1999). Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol, *New England Journal of Medicine* **341**, 410–418.

[46] Saudek, C.D., Duckworth, W.C., Giobbie-Hurder, A., et al. (1996). Implantable insulin pump vs. multiple-dose insulin for non-insulin-dependent diabetes mellitus. A randomized clinical trial, *Journal of the American Medical Association* **276**, 1322–1327.

[47] Spechler, S.J. & the Department of Veterans Affairs Gastroesophageal Reflux Disease Study Group (1992). Comparison of medical and surgical therapy for complicated gastroesophageal reflux disease in veterans, *New England Journal of Medicine* **326**, 786–792.

[48] Streptomycin Committee (1947). The effects of streptomycin on tuberculosis in man, *Journal of the American Medical Association* **135**, 634–641.

[49] Takaro, T., Hultgren, H.N., Lipton, M.J., et al. (1976). The VA cooperative randomized study of surgery for coronary arterial occlusive disease. II. Subgroup with significant left main lesions, *Circulation* **54**, (Suppl. III), 107–117.

[50] The Digitalis Investigation Group (1997). The effect of digoxin on mortality and morbidity in patients with heart failure, *New England Journal of Medicine* **336**, 525–533.

[51] The VA Cooperative Variceal Sclerotherapy Group (1991). Prophylactic sclerotherapy for esophageal varices in male alcoholics with cirrhosis: a randomized single blind multi-center clinical trial, *New England Journal of Medicine*, **324**, 1779–1784.

[52] The Veterans Administration Systemic Sepsis Cooperative Study Group (1987). Effect of high-dose glucocorticoid therapy on mortality in patients with clinical signs of systemic sepsis, *New England Journal of Medicine* **317**, 659–665.

[53] The Veterans Affairs Total Parenteral Nutrition Cooperative Study (1991). Perioperative total parenteral nutrition in surgical patients, *New England Journal of Medicine* **325**, 525–532.

[54] VA Cooperative Study Group on Antihypertensive Agents (1967). Effects of treatment on morbidity in hypertension. Results in patients with diastolic blood pressures averaging 115 through 129 mm Hg, *Journal of the American Medical Association* **202**, 1023–1034.

[55] VA Cooperative Study Group on Antihypertensive Agents (1970). Effects of treatment on morbidity in hypertension. II. Results in patients with diastolic blood pressure averaging 90 through 114 mm Hg, *Journal of the American Medical Association* **213**, 1143–1152.

[56] Wasson, J.H., Reda, D.J., Bruskewitz, R.C., et al. (1995). A comparison of transurethral surgery with watchful waiting for moderate symptoms of benign prostatic hyperplasia, *New England Journal of Medicine* **332**, 75–79.

[57] Weinberger, M., Oddone, E.Z., Henderson, W.G., et al. (1996). Does increased access to primary care reduce hospital readmissions?, *New England Journal of Medicine* **334**, 1441–1447.

WILLIAM G. HENDERSON, PHILIP W. LAVORI,
PETER PEDUZZI, JOSEPH F. COLLINS,
MIKE R. SATHER & JOHN R. FEUSSNER

Copula

The term *copula* was introduced by Sklar [18] to denote a bivariate distribution function with uniform marginals. If X and Y are random variables with joint distribution function $H(x, y)$ and marginals $F(x) = H(x, \infty)$, $G(y) = H(\infty, y)$, then there exists a copula $C(u, v)$, uniquely determined on (Range F) $\times$ (Range G), such that

$$H(x, y) = C[F(x), G(y)] : \tag{1}$$

see [18], [15], and [16]. If X and Y are independent, $C(u, v) = uv$. The copula of a continuous bivariate distribution is invariant under separate continuous monotone transformations of each marginal distribution, and is maximal in the sense that any **measure of association** that is invariant under all such **transformations**, such as Kendall's τ, or Spearman's ρ, depends on H only through C. If $C(u, v)$ is a copula, then so also is $\overline{C}(u, v) = C(1 - u, 1 - v) + u + v - 1$: it corresponds to the distribution of $(-X, -Y)$ and represents the joint survivor function $\overline{H}(x, y) = 1 - F(x) - G(y) + H(x, y)$ of (X, Y) in terms of the marginal survivor functions $\overline{F}(x) = 1 - F(x)$ of X and $\overline{G}(y) = 1 - G(y)$ of Y (*see* **Survival Distributions and Their Characteristics**).

In statistical work, parameterization of the joint distribution H via distinct parameters α, β, and γ for the marginals F and G and the copula C allows specification of the marginals of the distributions to be separated from specification of the dependence structure, which is often desirable for interpretation. However, inferences about the three parameters are not generally orthogonal unless X and Y are actually independent.

Frank [6] described an important subclass, called "archimedean copulas". These arise mathematically as the class of associative copulas; that is those that satisfy $C[u_1, C(u_2, u_3)] = C[C(u_1, u_2), u_3]$ and have the general form $C(u, v) = \phi^{-1}[\phi(u) + \phi(v)]$, where ϕ decreases monotonically with u, $\phi(0) = 1$, and has an increasing first derivative. Genest & MacKay [7] discussed applications of this class in statistics. See also [14] for applications in survival analysis, and [10] and [1] for applications in **extreme value** theory. Oakes [14] showed how many members of this class, or more precisely of the complementary class

$$\overline{C}(u, v) = \phi^{-1}[\phi(u) + \phi(v)], \tag{2}$$

can arise from frailty models. Specifically, suppose that X and Y are conditionally independent given the value of a third, unobserved, variable W, called a frailty, and that each follows a (continuous) **proportional hazards** model in W, so that $\Pr(X > x|W = w) = A(x)^w$, $\Pr(Y > y|W = w) = B(y)^w$, for some *baseline* survivor functions $A(x)$ and $B(y)$. Then the joint survivor function of (X, Y) is

$$\begin{aligned}\Pr(X > x, Y > y) &= \mathrm{E}\left[\Pr(X > x, Y > y|W)\right] \\ &= \mathrm{E}\left[A(x)^W B(y)^W\right] \\ &= p\{-\ln[A(x)] - \ln[B(y)]\},\end{aligned} \tag{3}$$

where $p(s) = \mathrm{E}[\exp(-sW)]$ denotes the Laplace transform of the distribution of W. This is an archimedean copula model for $\overline{C}$ with $\phi^{-1} = p$, $\overline{F}(x) = p\{-\ln[A(x)]\}$ and $\overline{G}(y) = p\{-\ln[B(y)]\}$. Important examples include the gamma frailty model of Clayton [4] and Oakes [13], with $p(u) = (1 + u)^{-\kappa}$, and the positive stable model of Hougaard [11], with $p(u) = \exp(-u^\alpha)$ $(0 < \alpha < 1)$.

Models of the form given in (3) are natural in survival analysis, because of the close analogy with Cox's [5] proportional hazards regression model. Clayton [4] fitted the gamma frailty model to ages of occurrence of heart attacks in fathers and sons. He pointed out an appealing interpretation in terms of bivariate hazard functions. For the gamma frailty model the ratio of the hazards at y of the conditional distribution of Y given $X = x$ and of Y given $X > x$ is the same for all points (x, y). Oakes [14] extended this result to the general archimedean copula model, showing that the ratio of hazards in (3) depends on (x, y) only through $\overline{H}(x, y)$, and that this ratio of hazards characterizes the joint distribution. Genest & Rivest [8] explored a different characterization of archimedean copula models using Kendall's τ.

Bickel et al. [3, Chapter 4] discussed the challenging problems of semiparametric inference in (2) or (3). Genest et al. [9] and Shih & Louis [17] discussed a **pseudo-likelihood** approach to estimation of association in special cases of the model given in (2) from **censored data**.

There has been little work to date on the inclusion of observed **explanatory variables** (**covariates**) in copula models. For frailty models such as that given

in (2), covariate effects can be modeled either conditionally on the unobserved frailty, or unconditionally; that is, on the marginal distributions.

Extensions to higher dimensions have been considered in [2]. For a book-length treatment of copulas, see [12].

References

[1] Ballerini, R. (1994). Archimedean copulas, exchangeability and max-stability, *Journal of Applied Probability* **31**, 383–390.

[2] Bandeen-Roche, K.J. & Liang, K.-Y. (1996). Modelling failure-time associations in data with multiple levels of clustering, *Biometrika* **83**, 29–39.

[3] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[4] Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiologic studies of familial tendency in chronic disease incidence, *Biometrika* **65**, 141–151.

[5] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[6] Frank, M.J. (1975). Associativity in a class of operations on spaces of distribution functions, *Aequationes Mathematicae* **12**, 121–144.

[7] Genest, C. & MacKay, R.J. (1986). Copules archimediennes et familles de lois bidimensionelles dont les marges sont donnees, *Canadian Journal of Statistics* **14**, 145–159.

[8] Genest, C. & Rivest, L.P. (1993). Semiparametric inference procedures for bivariate archimedean copulas, *Journal of the American Statistical Association* **88**, 1034–1043.

[9] Genest, C., Ghoudi, K. & Rivest, L.P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika* **82**, 543–552.

[10] Gumbel, E.J. (1960). Bivariate exponential distributions, *Journal of the American Statistical Association* **55**, 698–707.

[11] Hougaard, P. (1986). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678; Correction: *Biometrika* **75**, (1988). 395.

[12] Nelsen, R.B., (1999). *An Introduction to Copulas*. Springer-Verlag, New York.

[13] Oakes, D. (1982). A model for association in bivariate survival data, *Journal of the Royal Statistical Society, Series B* **44**, 414–422.

[14] Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**, 487–493.

[15] Schweizer, B. & Sklar, A. (1974). Operations on distribution functions not derivable by operations on random variables, *Studia Mathematica* **52**, 43–52.

[16] Schweizer, B. & Wolff, E.F. (1981). On nonparametric measures of dependence for random variables, *Annals of Statistics* **9**, 879–885.

[17] Shih, J.H. & Louis, T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data, *Biometrics* **51**, 1384–1399.

[18] Sklar, A. (1959). Fonctions de répartition à *n* dimensions et leurs marges, *Institute Statistique Université Paris Publications* **8**, 229–231.

DAVID OAKES

Cornfield, Jerome

Born: October 30, 1912, in New York City, New York.
Died: September 17, 1979, in Herndon, Virginia.

Reproduced by permission of the Royal Statistical Society

Jerome Cornfield was arguably the most influential statistician in the biomedical sciences in the US from the 1950s until his death. He was the consummate statistical scientist. His understanding of the nature of the subject-matter of statistics and of its essential role in the inductive process of integrating data into a body of empirical knowledge, particularly in the biomedical sciences, was outstanding. This thorough view of statistics and scientific research enabled him to identify essential statistical problems. He exercised considerable influence as an advisor and consultant, and for over two decades was a major advocate for statistical reasoning in clinical research.

After attending elementary and high schools in the Bronx, New York, he entered New York University, graduating in 1933 with a major in history. Cornfield did not receive any advanced degrees. He did, however, take some formal graduate courses in history at Columbia University. After moving to Washington, DC, in 1935, Cornfield took a number of courses in statistics at the US Department of Agriculture Graduate School during the period 1936–1938, including courses with M.A. Girshick in general statistics and **multivariate analysis**. He also had a course in sampling which, together with what he learned on the job from Duane Evans, enabled him to advance the cause of getting **probability sampling** accepted by several Federal Agencies. Although his formal training was minimal, most of what he had to learn about statistical theory, reasoning, and methodology was self-taught from a continually expanding literature. This enabled him to be discriminatingly selective both as to subject-matter and to the time at which he felt it necessary to learn about a subject. In later years, biomedical associates and statistical colleagues were surprised to discover that he had no doctorate.

A brief review of the major positions he held begins with the Bureau of Labor Statistics, where he was a statistician from 1935 to 1947. In 1947 he joined **Harold Dorn's** methods unit in the Public Health Service. This unit was shortly transferred to the National Cancer Institute on the campus of the **National Institutes of Health (NIH)**. Cornfield remained in the Cancer Institute until 1955 or 1956 when both he and Dorn moved over to a new Division of Research Services. Here, he consulted with investigators in various Institutes of the NIH. In 1958 he was invited to succeed **William Cochran** as Chairman of the Department of Biostatistics in the School of Hygiene and Public Health of the Johns Hopkins University. He was also appointed Professor of Biomathematics in the School of Medicine. He returned to the NIH in 1960 as Assistant Chief of the Biometrics Research Branch of the National Heart Institute, became Branch Chief in 1963, and served in that position until his retirement from the NIH in 1967. In 1968 he joined the Graduate School of Public Health of the University of Pittsburgh as a Research Professor of Biostatistics. At the same time he founded a biostatistics research group with offices in the Washington, DC, area. In 1972 he joined the Department of Statistics at the George Washington University as Professor of Statistics and brought his research group into the Department as the Biostatistics Center. He served as Chairman of the Department from 1973 to 1976 and continued as Professor of Statistics and Director of the Center until his terminal illness.

Over a span of three decades, from 1947 to 1979, Professor Cornfield was one of the leading statisticians working in the biomedical area. He made

many original contributions to biostatistics, epidemiology, **clinical trials**, and to quantitative methods in the design and analysis of experiments (*see* **Experimental Design**) conducted in clinical and laboratory research. In addition, he wrote a number of papers on Bayesian **inference** and on the application of **Bayesian methods** in the biomedical sciences. Before presenting the highlights of the work in this period, it is important to comment on his contributions to economic statistics and sampling while at the Bureau of Labor Statistics (BLS).

From the very beginning of his career, Cornfield was a creative and original thinker, motivated by important real-world problems. He made a number of important contributions to economics and economic statistics during his work at the BLS. He played a major role in the revision of the Consumer Price Index, 1938–1940, introducing several new procedures. He developed a keen interest in sampling, which led to the development of a survey using probability sampling for a study of Family Spending and Saving in wartime. This complex design, according to Duncan & Shelton [26, pp. 46–49] "represented a significant advance in a number of respects. Indeed, it was the precursor of several ideas which were worked out more fully and justified mathematically a year later by Hansen and Hurwitz". In 1941 Cornfield consulted with the Bureau of Home Economics on a nutrition-related problem which was known as the "diet" problem. The mathematical problem requires the minimization of linear functions subject to a set of given inequality constraints, the problem of **linear programming**. Zelen [31, p. 12] refers to a 1958 book on linear programming by Dorfman et al. as crediting Cornfield "as being the first person to formulate the linear programming problem and find an approximate solution". His work appeared in 1941 in an unpublished BLS memorandum. It was also at the BLS that Cornfield made his first contribution to statistical theory. He developed a method using indicator variables for easily obtaining the first few moments of the sample mean when sampling from finite populations. He thus obtained an unbiased estimate of the sample variance and of the variance of the sample mean [2].

From 1948 to his death 31 years later, Cornfield devoted the major portion of his career to the development and application of statistical theory and methods to the biomedical sciences. His contributions were diverse both in the nature of his statistical interests and in the areas of biostatistical applications. He was involved in and touched upon every major public health issue that arose in that period – the polio vaccines [23], smoking and lung cancer (*see* **Smoking and Health**) [22, 29], risk factors for cardiovascular disease [5, 30], and the difficult statistical issues of estimating the low-dose carcinogenic effects in humans (*see* **Extrapolation, Low Dose**) of a food additive that becomes suspect because it produces cancer in animals at much higher doses [14, 20].

In the broad area of biomedical research, Cornfield was involved in a wide variety of problems, in each of which he made significant and lasting contributions. These studies and problems include the following: an imaginative method for estimating the volume–surface ratio of individual cells as observed under the microscope [15], the statistics of bioassay (*see* **Biological Assay, Overview**) [3, 6, 19], photosynthesis [1], the analysis of the toxicity of mixtures of the essential amino acids [28], chemical kinetic experiments using radioactive compounds (*see* **Pharmacokinetics and Pharmacodynamics**) [25], the physiological and biological effects of irradiated animals (*see* **Radiation**) [17], and the computer diagnosis of electrocardiograms [12] (*see* **Clinical Signals**).

In the amino acid problem, the question was: Which mixtures of the 10 essential amino acids were toxic? The investigators called on Cornfield for help when they were confronted with the impractical task of conducting 1013 experiments with two or more mixtures. Cornfield considered the issue of measuring the joint effects of two or more drugs administered in combination. The method usually employed was to assume the joint effects were additive in their individual responses. Cornfield saw that this simple method could give strange results. Instead, he chose a measure of additivity introduced by Gaddum, namely additivity of doses conditioned on a given response, a concept which Cornfield called dose-wise additivity. After some persuasion, the biochemists proceeded to conduct experiments implied by dose-wise additivity. These turned out to be highly successful, leading to the previously unknown result that L-arginine was essential for the combination of the 10 amino acids to be nontoxic in the human [28].

The animal radiation study is noteworthy for the development of a methodology that would later become a fundamental tool in epidemiologic research,

i.e. multiple **logistic regression**. The issue in the animal data was the effect on survival of irradiated mice as a function of certain observed blood characteristics, such as lymphocytes and granulocytes. This is clearly a regression problem with a straightforward solution if the traits could be controlled at a set of fixed values. Since survival is a 0, 1 variable, the solution would be a multiple logistic function. (Of course, since that period much work has been done on regression with variables subject to error.) Since the observed blood properties were uncontrolled, Cornfield chose to adopt the method of analysis as that of discrimination between two multivariate populations for surviving and nonsurviving animals. With the additional assumption of **multivariate normality** and equal **covariance matrices**, he derived the multiple logistic risk function whose coefficients were the same as those found by **R.A. Fisher** in the linear discrimination problem (*see* **Discriminant Analysis, Linear**) [5, 21]. This solution is obtained directly, requiring only the inversion of a matrix but no iterations. Cornfield would later say that the simplicity of the solution appealed to him and he believed that if the assumptions were reasonable, then the solution would be close to that of the regression approach. Cornfield later applied the same reasoning to use the multiple risk function to identify cardiovascular risk factors on the basis of data obtained from the famous **Framingham Study** [5, 30].

Cornfield made another very important contribution to epidemiology. When epidemiologists began turning their attention to the study of chronic diseases, prospective **cohort** designs for finding causes of, or risk factors for, chronic diseases were in many instances impractical. They therefore turned to **case–control** or retrospective types of strategies. A problem with these designs, assuming they are well planned, is that they do not yield traditional estimates of **absolute risk** or **relative risk**. Cornfield, in 1955 at the Third Berkeley Symposium in Mathematical Statistics and Probability [4, 18], presented a derivation which demonstrated that under a rather strong assumption (but rather reasonable in the case of chronic diseases) the **odds ratio** or cross product ratio (in a 2×2 table) is a fairly good approximation of the relative risk. The assumption was that the incidence of the disease under study should be small. This result strengthened and increased the use of the case–control design, since it set this research strategy on a much more solid inferential foundation.

In an important paper [22], responding to critics of the purported causal relationship between smoking and lung cancer, Cornfield argued for the preference of measures of association based on relative risk as opposed to differences of absolute risk, at least for scientific purposes. However, the significant matter here is not the issue of risks but the example he used to justify his position. The illustration bears on the question of the effect of latent, unobservable variables. Sir Ronald Fisher, in arguing against the smoking–lung cancer relationship, had offered an hypothesis that postulated the existence of some constitutional factor (latent and unobservable), e.g. genetic, that caused cancer and that was also associated with the need to smoke. Without giving the details of his argument here, Cornfield demonstrated that if cigarette smokers are shown to have nine times the risk of nonsmokers of getting lung cancer, but that this elevated risk is due, not to cigarettes, but to some latent factor X, then the proportion of smokers having X must be larger than nine times the proportion of nonsmokers having X. Cornfield's conclusion was that if X was a causative agent of this magnitude, then the relationship between the latent factor X and the observed agent would probably have been detected much before that of the agent and the disease. No such factor has been found.

In addition to epidemiologic methods, Cornfield devoted a substantial portion of his career to the theory and practice of randomized, controlled clinical trials (RCTs). His influence was far-reaching. He wrote papers on aspects of design of RCTs both for therapeutic and **prevention trials** [10, 13, 17, 24], on statistical problems in the interpretation of results [13], and on a Bayesian test of hypotheses arising in RCTs [7]. But the totality of his publications constituted only a small part of his vast influence as an advisor and consultant. He was personally involved in the Coronary Drug Project, one of the earliest **multicenter trials** sponsored by the NHLBI. It was in this trial that Cornfield introduced the Bayesian concept of relative betting odds (a measure related to the Bayes Factor) as a measure to assess the efficacy of a therapy instead of the classical ***P* value**. He was personally involved in many major multicenter trials, serving in various capacities as a member of planning committees, steering

committees, policy advisory boards, and data monitoring and safety committees. These RCTs include the National-Diet Study, a trial of urokinase in the treatment of myocardial infarction, the Coronary Drug Project (CDP), the **University Group Diabetes Program (UGDP)**, the Urokinase Pulmonary Embolism Trial (UPET), the Diabetic Retinopathy Study (DRS), the Multiple Risk Factor Intervention Trial (MRFIT), the Program for the Surgical Control of the Hyperlipidemias (POSCH), and the Persantin Aspirin Reinfarction Study (PARIS).

Throughout his career in statistics Cornfield was interested in, and contributed to, the foundations of statistics, first as a frequentist and then as a Bayesian. The first manifestation of his interest in Bayesian inference was his joint work with Geisser on deriving the posterior distribution for the multivariate normal parameters [27]. He then followed with a number of papers on the theory of Bayesian inference and on its practice and application to clinical trials [8, 11], to estimation in higher order cross-classifications [9], and to the analysis of **life tables** [16].

Cornfield was also actively engaged as a consultant in areas other than clinical trials and epidemiology. He was a member of the Three Mile Island Advisory Committee, on the NHLBI Policy Advisory Board on Coronary Bypass Surgery, Chairman of the Committee on Biometry and Epidemiology for the Food and Drug Administration, on the Scientific Advisory Board for the Sloan–Kettering Institute for Cancer Research, etc. He also served in a number of editorial roles, the principal ones being Associate Editor for the ***Journal of the American Statistical Association*** and Consulting Editor for the *Journal of Chronic Diseases*. Cornfield was President of the **American Statistical Association**, the American Epidemiological Society, Vice-President of the American Heart Association, and President of the Eastern North American Region of the **International Biometric Society**.

For a more detailed review of Cornfield's contributions to the theory of statistics, laboratory research, clinical trials, and epidemiology, the reader is referred to the March 1982 supplement to ***Biometrics*** vol. 38. Furthermore, Cornfield's American Statistical Association Presidential Address is a wonderful account in his own words of his contributions to statistics and science, and his personal perspective on being a statistician.

Cornfield married Ruth Bittler and they have two daughters, Ann and Ellen.

References

[1] Burk, D., Cornfield, J. & Schwarz, M. (1951). The efficient transformation of light into chemical energy in photosynthesis, *Scientific Monthly* **73**, 213–233.

[2] Cornfield, J. (1944). On samples from finite populations, *Journal of the American Statistical Association* **39**, 236–239.

[3] Cornfield, J. (1955). Review: the statistics of bioassay, *Journal of the American Statistical Association* **50**, 1368–1371.

[4] Cornfield, J. (1956). A statistical problem arising from retrospective studies, in *Proceedings of the Third Berkeley Symposium*, Vol. 4, J. Neyman, ed. University of California Press, Berkeley, pp. 135–148.

[5] Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure, *Federation Proceedings* **21**, Supplement 11, Part 2, 58–61.

[6] Cornfield, J. (1964). Comparative bioassays and the role of parallelism, *Journal of Pharmacology and Experimental Therapeutics* **144**, 143–149.

[7] Cornfield, J. (1966). A Bayesian test of some classical hypotheses, *Journal of the American Statistical Association* **61**, 577–594.

[8] Cornfield, J. (1969). The Bayesian outlook and its applications, *Biometrics* **25**, 617–657.

[9] Cornfield, J. (1970). Bayesian estimation of higher order cross-classifications, *Milbank Memorial Fund Quarterly* **48**, 57–70.

[10] Cornfield, J. (1970). Design of primary and secondary prevention trials, in *Atherosclerosis: Proceedings of the Second International Symposium*, R.J. Jones, ed. Springer-Verlag, New York, pp. 566–571.

[11] Cornfield, J. (1970). The frequency theory of probability, Bayes' theorem, and sequential clinical trials, in *Bayesian Statistics*, D.L. Myers & R.O. Collier, Jr, eds. Peacock, Ithaca, pp. 1–28.

[12] Cornfield, J. (1972). Statistical classification methods, in *Computer Diagnosis and Diagnostic Methods*, J.A. Jacquez, ed. Charles C. Thomas, Springfield, pp. 108–130.

[13] Cornfield, J. (1976). Recent methodological contributions to clinical trials, *American Journal of Epidemiology* **104**, 408–421.

[14] Cornfield, J. (1977). Carcinogenic risk assessment, *Science* **198**, 693–699.

[15] Cornfield, J. & Chalkley, H.W. (1951). A problem in geometric probability, *Journal of the Washington Academy of Sciences* **41**, 226–229.

[16] Cornfield, J. & Detre, K. (1977). Bayesian life table analysis, *Journal of the Royal Statistical Society, Series B* **39**, 86–94.

[17] Cornfield, J. & Greenhouse, S.W. (1967). On certain aspects of sequential clinical trials, in *The Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, J. Neyman & L.M. Le Cam, eds. University of California Press, Berkeley, pp. 813–829.

[18] Cornfield, J. & Haenszel, W. (1960). Some aspects of retrospective studies, *Journal of Chronic Diseases* **11**, 523–534.

[19] Cornfield, J. & Mantel, N. (1950). Some new aspects of the application of maximum likelihood to the calculation of the dosage response curve, *Journal of the American Statistical Association* **45**, 181–210.

[20] Cornfield, J. & Mantel, N. (1977). "Safe doses" in carcinogenic experiments, *Biometrics* **33**, 21–30.

[21] Cornfield, J., Gordon, T. & Smith, W.W. (1961). Quantal response curves for experimentally uncontrolled variables, *Bulletin of the International Statistical Institute* **37**(3), 97–115.

[22] Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. & Wynder, E.L. (1959). Smoking and lung cancer, *Journal of the National Cancer Institute* **22**, 173–203.

[23] Cornfield, J., Halperin, M. & Moore, F. (1956). Some statistical aspects of safety: testing the Salk poliomyelitis vaccine, *Public Health Reports* **71**, 1045–1056.

[24] Cornfield, J., Halperin, M. & Greenhouse, S.W. (1969). An adaptive procedure for sequential clinical trials, *Journal of the American Statistical Association* **64**, 759–770.

[25] Cornfield, J., Steinfeld, J. & Greenhouse, S.W. (1960). Models for the interpretation of experiments using tracer compounds, *Biometrics* **16**, 212–234.

[26] Duncan, J.W. & Shelton, W.C. (1978). *Revolution in United States Government Statistics 1926–1976*. US Government Printing Office, Washington.

[27] Geisser, S. & Cornfield, J. (1963). Posterior distributions for multivariate parameters, *Journal of the Royal Statistical Society, Series B* **25**, 368–376.

[28] Gullino, P., Winitz, M., Birnbaum, S.M., Cornfield, J., Otey, M.C. & Greenstein, J.P. (1956). Studies on the metabolism of amino acids and related compounds *in vivo*, *Archives of Biochemistry and Biophysics* **64**, 319–332.

[29] Sadowsky, D.A., Gilliam, A.G. & Cornfield, J. (1953). Statistical association between smoking and carcinoma of the lung, *Journal of the National Cancer Institute* **13**, 1237–1258.

[30] Truett, J., Cornfield, J. & Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham, *Journal of Chronic Diseases* **20**, 511–524.

[31] Zelen, M. (1962). Contributions of Jerome Cornfield to the theory of statistics, *Biometrics Supplement* **38**, 11–15.

SAMUEL W. GREENHOUSE &
JOEL B. GREENHOUSE

Cornfield's Inequality

In response to claims that the relationship between smoking and lung cancer could be explained by a genetic or other omitted variable (OV), **Cornfield** et al. [6] developed an inequality linking the observed risk ratio (*see* **Relative Risk**) to the **prevalence** of the omitted variable in smoking and nonsmoking groups. They wrote:

> If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r.
>
> Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least 9 times greater than among nonsmokers. If the relative prevalence of hormone X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect.

See [13, p. 40] for a discussion of the origins of the inequality.

Formally, the analysis involves three binary variables: (i) $Z = 1$ for treatment (smoker) and $Z = 0$ for control (nonsmoker), (ii) $D = 1$ for positive response (lung cancer) and $D = 0$ for negative response, (iii) $U = 1$ for presence of the unobserved omitted variable and $U = 0$ for its absence. We observe the joint distribution of Z and D, specifically $\pi = \Pr(Z = 1)$, $p_1 = \Pr(D = 1|Z = 1)$ and $p_0 = \Pr(D = 1|Z = 0)$, from which we calculate the observed risk ratio $R_{\mathrm{O}} = p_1/p_0$. Could the observed risk ratio R_{O} deviate from one solely because of the unobserved variable, U? If this were the case, then D would be independent of Z given U. Hence,

$$\begin{aligned} p &\equiv \Pr(D = 1|Z = 1, U = 0) \\ &= \Pr(D = 1|Z = 0, U = 0) \\ &= \Pr(D = 1|U = 0) \end{aligned}$$

and

$$\begin{aligned} pR_{\mathrm{U}} &= \Pr(D = 1|Z = 1, U = 1) \\ &= \Pr(D = 1|Z = 0, U = 1) \\ &= \Pr(D = 1|U = 1), \end{aligned}$$

where

$$R_{\mathrm{U}} = \frac{\Pr(D = 1|U = 1)}{\Pr(D = 1|U = 0)}.$$

Here, R_{U} is the unobserved risk ratio linking the response, D, with the unobserved variable, U. We may assume that the two categories of the variable U, $U = 1$, and $U = 0$, have been labeled so that $R_{\mathrm{U}} \geq 1$. Writing $f_1 = \Pr(U = 1|Z = 1)$ and $f_0 = \Pr(U = 1|Z = 0)$ gives

$$\begin{aligned} R_{\mathrm{O}} = \frac{p_1}{p_0} &= \frac{p(1 - f_1) + pR_{\mathrm{U}}f_1}{p(1 - f_0) + pR_{\mathrm{U}}f_0} \\ &= \frac{f_1R_{\mathrm{U}} + 1 - f_1}{f_0R_{\mathrm{U}} + 1 - f_0}. \end{aligned} \tag{1}$$

For a fixed $R_{\mathrm{U}} \geq 1$, expression (1) is maximized when $f_1 = 1$ and $f_0 = 0$, leading to the inequality

$$R_{\mathrm{O}} \leq R_{\mathrm{U}}. \tag{2}$$

Similarly, for fixed values of f_0 and f_1, expression (1) is maximized by letting $R_{\mathrm{U}} \to \infty$, yielding the inequality

$$R_{\mathrm{O}} \leq \frac{f_1}{f_0} = \theta, \text{ say.} \tag{3}$$

Eqs. (2) and (3) say that the unobserved risk ratio, R_U, must exceed both the observed risk ratio, R_{O}, and the unobserved prevalence ratio, θ, if U is to explain away the association between treatment Z and response D. Expressions (2) and (3) are the inequalities of Cornfield et al. [6], and (3) is described in the quotation above.

Gastwirth [9] gave a sharper version of (3) by solving (1) for θ to obtain

$$\theta = R_{\mathrm{O}} + \frac{R_{\mathrm{O}} - 1}{R_{\mathrm{U}} - 1}\frac{1}{f_0}, \tag{4}$$

or, equivalently,

$$f_1 = R_{\mathrm{O}}f_0 + \frac{R_{\mathrm{O}} - 1}{R_{\mathrm{U}} - 1}. \tag{5}$$

We illustrate the result on data from the cohort study [29] of lung cancer in asbestos workers, as described in [9, p. 807]. Over the entire period of the study, the relative risk of exposed workers dying from lung cancer was 6.8 times their

expected number, assuming workers had the rate of lung cancer in the general male population. As smoking is another risk factor for lung cancer, we apply Cornfield's inequality to see whether smoking could explain the asbestos–lung cancer association. It is known that blue-collar workers have a greater prevalence of smoking than in the general male population. At the time of the study, 60% of all males smoked, in contrast to about 80% of males in asbestos-related occupations. The prevalence ratio, $\theta = 0.8/0.6 = 1.33$, is much less than $R_{\mathrm{O}} = 6.8$, so Cornfield's inequality implies that smoking cannot explain the entire association between asbestos and lung cancer.

When information about R_{U} is available, (4) can provide a substantially stronger statement. Suppose that a large study of workers exposed to chemical A found a relative risk of three for lung cancer. While smoking was controlled for in this imagined study, prior substantial exposure to asbestos, U, was not, although the literature indicates that R_{U} is, at most, 10. The original inequality (3) implies that the prevalence of asbestos exposure (U) in the exposed workers needs to be at least three times its prevalence among workers not exposed to chemical A. From the job histories of the workers one might estimate that the prevalence, f_0, of U among the unexposed group was 0.05, say. Then (5) implies that the prevalence of U among workers exposed to chemical A would need to reach 0.374 in order for the observed association between lung cancer and chemical A to be explained by prior substantial exposure to asbestos. This is much larger than $3 \times 0.05 = 0.15$ implied by the original inequality. Indeed, inequality (3) is obtained from (4) by letting R_{U} become arbitrarily large.

While Cornfield et al. [6] preferred the relative risk measure for assessing causality, the difference in proportions, or the absolute risk difference, is useful in public health. Write $\Delta_U = \Pr(D = 1|U = 1) - \Pr(D = 1|U = 0)$ for the difference in mortality rates associated with the unobserved variable, and write $\Delta_Z = \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)$ for the difference in mortality associated with the exposure. The corresponding inequality is given in the following lemma.

Lemma. If U is to explain entirely the observed difference Δ_Z, then one must have $(f_1 - f_0)\Delta_U \geq \Delta_Z$, and, in particular, one must have both $\Delta_U \geq \Delta_Z$ and $f_1 - f_0 \geq \Delta_Z$.

Inequalities closely related to Cornfield's inequality have been proposed by Bross [3, 4] and Schlesselmann [28]. Related equalities are discussed by Miettinen [19], Breslow & Day [2, p. 96], and Gail et al. [7], and the equalities might be used to calculate adjusted risk ratios. Gastwirth [10] suggests that the inequalities (2) and (3) be used in conjunction with Koopman's [15] one-sided **confidence interval** for the risk ratio $\Pr(D = 1|Z = 1)/\Pr(D = 1|Z = 0)$ to account for sampling error. Gastwirth [11] uses the reasoning underlying the inequality of Cornfield et al. [6] to examine the potential effect of **nonresponse** or **missing data**. Gail et al. [8] discuss the effect of failing to adjust for a covariate in a clinical trial.

Cornfield's inequality was the first formal method of **sensitivity analysis** in **observational studies** or **nonrandomized** experiments. The inequality may be viewed as asking how the conclusions of an observational study might be altered by departures of various magnitudes from the random assignment of treatments (*see* **Randomized Treatment Assignment**), where the departure is measured by the prevalence ratio, θ. Viewed in this way, sensitivity analysis based on Cornfield's inequality is identical in purpose, similar in spirit, though different in technical detail, to the method of permutational sensitivity analysis proposed later by Rosenbaum [21–25] and Rosenbaum & Krieger, [26]. The latter approach applies not only to **binary** responses, but also to continuous responses, discrete scores, censored survival times (*see* **Censored Data**) and multivariate outcomes. It permits sensitivity analysis for **quantiles**, Wilcoxon's [30] rank sum test (*see* **Wilcoxon–Mann–Whitney Test**) and **signed rank** test, the **logrank test** and Gehan test [12] for survival times, the Hodges–Lehmann [14] point estimates of an additive effect, **McNemar test** [18], the **Mantel–Haenszel method** [17], and Mantel's extension for discrete scores [16], among others. In one very special case, Cornfield's inequality and Rosenbaum's sensitivity analysis give identical results. Specifically, with a binary response in a **case–control study** that approximates the relative risk by the **odds ratio**, the lower endpoint of the $1 - \alpha$ confidence interval for the relative risk in Cornfield's inequality occurs at the value of the sensitivity parameter yielding an upper bound of α for the significance level for testing no treatment effect; see Rosenbaum [22] for specifics. Other methods

of sensitivity analysis are discussed in [1, 5, 20], and [27].

References

[1] Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with discussion), *Journal of the American Statistical Association* **91**, 444–472.

[2] Breslow, N. & Day, N. (1980). *The Analysis of Case–Control Studies*, Vol. 1: *Statistical Methods in Cancer Research*. International Agency for Research on Cancer of the World Health Organization, Lyon.

[3] Bross, I.D.J. (1966). Spurious effects from an extraneous variable, *Journal of Chronic Diseases* **19**, 637–647.

[4] Bross, I.D.J. (1967). Pertinency of an extraneous variable, *Journal of Chronic Diseases* **20**, 487–495.

[5] Copas, J.B. & Li, H.G. (1997). Inference in nonrandom samples (with discussion), *Journal of the Royal Statistical Society, Series B* **59**, 55–95.

[6] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.

[7] Gail, M., Wacholder, S. & Lubin, J. (1988). Indirect corrections for confounding under multiplicative and additive risk models, *American Journal of Industrial Medicine* **13**, 119–130.

[8] Gail, M., Wieand, S. & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and missing covariates, *Biometrika* **71**, 431–444.

[9] Gastwirth, J. (1988). *Statistical Reasoning in Law and Public Policy*. Academic Press, New York.

[10] Gastwirth, J. (1992). Method for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables, *Jurimetrics* **33**, 19–34.

[11] Gastwirth, J. (1992). Employment discrimination: a statistician's look at analysis of disparate impact claims, *Law and Inequality* **11**, 151–179.

[12] Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples, *Biometrika* **52**, 203–223.

[13] Greenhouse, S. (1982). Jerome Cornfield's contributions to epidemiology, *Biometrics* **28**, Supplement, 33–46.

[14] Hodges, J. & Lehmann, E. (1963). Estimates of location based on rank tests, *Annals of Mathematical Statistics* **34**, 598–611.

[15] Koopman, P.A.R. (1984). Confidence intervals for the ratio of two binomials, *Biometrics* **40**, 513–517.

[16] Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, *Journal of the American Statistical Association* **58**, 690–700.

[17] Mantel, N. & Haenszel, W. (1959). Statistical aspects of retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[18] McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages, *Psychometrika* **12**, 153–157.

[19] Miettinen, O. (1972). Components of the crude risk ratio, *American Journal of Epidemiology* **96**, 168–172.

[20] Rosenbaum, P. (1986). Dropping out of high school in the United States: an observational study, *Journal of Educational Statistics* **11**, 207–224.

[21] Rosenbaum, P.R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies, *Biometrika* **74**, 13–26.

[22] Rosenbaum, P.R. (1991). Sensitivity analysis for matched case–control studies, *Biometrics* **47**, 87–100.

[23] Rosenbaum, P.R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies, *Journal of the American Statistical Association* **88**, 1250–1253.

[24] Rosenbaum, P.R. (1995). Quantiles in nonrandom samples and observational studies, *Journal of the American Statistical Association* **90**, 1424–1431.

[25] Rosenbaum, P.R. (1995). *Observational Studies*. Springer-Verlag, New York.

[26] Rosenbaum, P. & Krieger, A. (1990). Sensitivity analysis for two-sample permutation inferences in observational studies, *Journal of the American Statistical Association* **85**, 493–498.

[27] Rosenbaum, P. & Rubin, D. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, *Journal of the Royal Statistical Society, Series B* **45**, 212–218.

[28] Schlesselmann, J.J. (1978). Assessing the effects of confounding variables, *American Journal of Epidemiology* **108**, 3–8.

[29] Selikoff, I., Hammond, E. & Churg, J. (1964). Asbestos exposure, smoking and neoplasia, *Journal of the American Statistical Association* **188**, 22–26.

[30] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

(*See also* **Confounding**)

JOSEPH L. GASTWIRTH, ABBA M. KRIEGER & PAUL R. ROSENBAUM

Correlated Binary Data

In many studies, the primary measure of interest can take on one of two possible values, and is known as a binary response measure (*see* **Binary Data**). An example might be disease status (i.e. diseased/not diseased), in which the two categories are the only ones possible. In other situations, we might form the two categories by dichotomizing a continuous response, for example, whether or not systolic blood pressure is greater than 140 (*see* **Categorizing Continuous Variables**). The term *correlated binary response data* refers to two or more such binary outcome measures, which we assume are correlated. This situation could arise in a **longitudinal** study, where disease state (diseased/not diseased) is measured over time on the same person. Spatial proximity can also induce **correlation**, with measurements taken closer together assumed more highly correlated than those further apart. The similarity of sampling units can induce the correlation, such as when litter-mates or siblings are sampled, or when a matching process is used to build sets of similar sampling units. A fourth setting of correlated responses arises when the outcome variable is measured on subunits of a single, primary sampling unit, such as measuring the disease status of a person's right and left eyes separately.

The term *cluster* is often used in the literature to describe a set of correlated measurements (*see* **Cluster Sampling**). A measurement on each of two eyes on the same person would constitute a cluster of size two. Measurements on the same person taken at 6, 12, 18, and 24 months would constitute a cluster of size four. In some settings, there are multiple levels of clustering, otherwise described as a hierarchy of clusters. For example, measurements over time (level 1) on each of the two eyes (level 2) of a person (level 3) would constitute a cluster of two eyes of one person and a subcluster of multiple measurements over time on each of those two eyes. The individual measurements within the cluster are called subunits (or sub-subunits) of the cluster. Positive correlation among subunits is manifested by more homogeneous subunits and more variable cluster totals than would be expected with no correlation (i.e. simple Bernoulli sampling or **binomial** data). This effect is called extrabinomial variation or **overdispersion**. Ignoring this positive correlation in the analysis will result in statistical tests that overstate the significance of the differences seen among subunit responses.

This discussion is restricted primarily to settings in which there is a single random sample of clusters, such as a sample of different people in the **ophthalmologic** and longitudinal data settings, with some discussion of hierarchical clustering as well.

In some settings, the investigator may have other **explanatory variables** (or **covariates**, which can be used to model the probability of a binary outcome. When the covariate value must stay the same for all subunits of a cluster, it is called a *between-cluster or subunit-independent covariate*. When it can change from one member of the cluster to another, it is called a *within-cluster or subunit-dependent covariate*. In the context of longitudinal studies, the terms *time-independent* (for between-cluster) and *time-dependent* (for within-cluster) covariates have often been used. In hierarchical clustering, covariates could change or be constant at any level of the hierarchy.

We start by laying the groundwork in the context of a very simple setting with no covariates. The rest of the discussion is focused on **regression**-type methods that can accommodate covariates, separated into the major categories of response feature models, conditionally specified models, **transitional models**, **marginal models**, and cluster-specific models, and how **hierarchical models** can be accommodated in the different categories. Emphasis is placed on similarities and differences among them in terms of the questions that can be addressed, methods for fitting the model, interpretations of parameters, making inferences, and, to a lessor degree, computational aspects. Examples of many of these methods can be found in books by Diggle et al. [15] Goldstein, [20], Verbeke, and Molenberghs [60, 61] and Davis [14]. Our intent is to give the reader an overview of the work in this area and to point to key references in which more detail can be found.

Simple Setting with No Covariates

In the simplest setting, the cluster is of size two and there are no explanatory variables (*see* **Matched Pairs With Categorical Data**). The data could be laid out in a **two-by-two contingency table**, such as that given in Figure 1 for binary measurements taken at two time points.

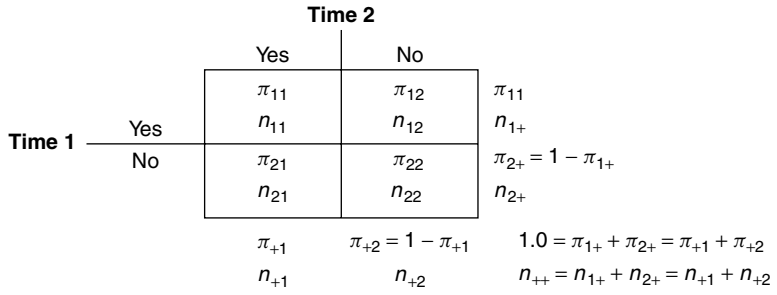

Figure 1 Notation for paired binary responses in contingency table format

Here, π_{11} represents the underlying joint probability of responding Yes at both time points, π_{12} and π_{21}, the probabilities of a Yes at one time point and a No at the other, and π_{22} the probability of responding No at both time points. The π_{1+} and π_{+1} are **marginal probabilities** – that is, π_{1+} is the probability of responding Yes at the first time point, regardless of whether a Yes or No was the response at the second time point.

The corresponding sample proportions, $p_{i,j} = n_{i,j}/n_{++}$ and $p_{1+} = (n_{11} + n_{12})/n_{++}$, can be used to address research questions. Different questions require different frameworks for modeling the data. If the goal is to assess the strength of the relationship between responses at the two time points, measures, such as an **odds ratio**, a **relative risk**, a tetrachoric correlation, or one of many other measures of **association** would be useful. If one were interested in determining if the proportion who answered Yes changed from time 1 to time 2, a **confidence interval** could be constructed on the difference in marginal proportions, $\pi_{1+} - \pi_{+1}$, or one could use the **McNemar test** of the **null hypothesis** $\pi_{1+} = \pi_{+1}$.

Cox [11] suggested a **logistic regression** approach to this problem of comparing marginal proportions, which only uses the clusters where a difference in responses was observed. Let $(y_{i,1}, y_{i,2})$ represent the pair of binary responses for the ith pair and let y take on the value 1 for Yes or success and 0 for No or failure. The model has the form

$$\log\left[\frac{P(y_{i,1}=1)}{P(y_{i,1}=0)}\right] = \alpha_i \qquad (1)$$

for the first member of the ith pair, and

$$\log\left[\frac{P(y_{i,2}=1)}{P(y_{i,2}=0)}\right] = \alpha_i + \beta \qquad (2)$$

for the second. The α_i parameters in both models simply measure variation from cluster to cluster. They are usually not of interest and are called **nuisance parameters**. The β parameter quantifies the difference between the two time points. The value exp (β) is called an odds ratio, meaning that the odds of a Yes (1) response (versus a No (0) response) are exp (β) times higher at time 2 than at time 1. The probability of a Yes response for the ith cluster at time 1 is $\exp(\alpha_i)/(1+\exp(\alpha_i))$ and $\exp(\alpha_i+\beta)/[1+\exp(\alpha_i+\beta)]$ at time 2. Because this model has as many α_i parameters as there are independent clusters of data, we could not estimate each of them. However, by conditioning on the number of Yes responses in the cluster, we find that the conditional joint distribution of responses only depends on those clusters with exactly one positive response and has the form of a **binomial distribution**, with mean parameter $\exp(\beta)/(1+\exp(\beta))$ and sample size parameter equaling the number of clusters, where $y_{i,1} + y_{i,2} = 1$ (*see* **Conditionality Principle**). Estimation and inference follow the usual theory for binomial data. We note that testing the mean parameter equal to 1/2 is equivalent to testing β equal to 0, a hypothesis of no difference between the responses at time 1 and time 2. This is an example of a **random-effects** model in which conditioning is used to enable **estimation**. Extensions of this type of model are further discussed in the section "Cluster-specific Models".

When $\pi_{1+} = \pi_{+1}$, it follows that $\pi_{2+} = \pi_{+2}$. This equality of all the marginal probabilities is known as marginal homogeneity (*see* **Marginal Models**). In a 2×2 table, marginal homogeneity also implies $\pi_{12} = \pi_{21}$. For the **square contingency table** in general, the condition $\pi_{ij} = \pi_{ji}$ for all $i \neq j$ is called symmetry of probabilities across the main

diagonal. The symmetry condition implies marginal homogeneity, but the reverse is not true for tables larger than 2×2.

In the case of binary responses at more than two time points, Cochran's Q test or, more generally, the Cochran–Mantel–Haenszel test, can be used to test for marginal homogeneity (*see* **Mantel–Haenszel Methods**).

Models with Covariates

In many settings, the researcher will be interested in learning not only if the response probabilities change over time (or among cluster subunits), but whether those changes are related to changes in covariates. For example, in **clinical trial** and **epidemiologic** settings, the goal may be to assess whether the probability of obtaining a disease differs across various randomized treatment regimens or is associated with observed **risk factors**. In other studies, however, the primary goal might be to investigate the structure of the within-cluster correlation or association. Models for these parameters could be set up to determine if the correlation/association within clusters is affected by between-cluster covariates. This type of analysis could, for example, have an application in determining the efficacy of a drug in preventing the spread of a contagious disease to other members of the cluster.

The types of models that can be fit in such situations depend upon the **measurement scale** of the covariates (nominal, ordinal, or continuous), the type and cluster level of covariates (between-cluster, within-cluster, or both), the types of assumptions the researcher is willing to make on the underlying joint distribution of the correlated responses, and the research questions to be answered. We focus primarily on regression-type methods that can accommodate both continuous and categorical covariates. Linear logistic regression models and probit models, which fall within a broad class of models known as **generalized linear models** (GLMs), play a central role in many analytic approaches to correlated binary data. Characterizing the underlying joint distribution of the responses within a cluster requires measures of the strength of the correlation or association within the cluster. Either first- and higher-order product-moment correlations or conditional odds ratios have typically been used to capture those intracluster relationships.

Modeling Strategies

Most modeling strategies for correlated binary data fall into at least one of five categories:

1. *Response feature models.* Collapse response information into one measure per cluster, and model using methods appropriate for independent univariate measures, such as logistic regression.

2. *Conditionally specified models.* The probability of a positive response for one member of the cluster is modeled conditionally on all other outcomes in the same cluster.

3. *Transitional models.* Model the probability of a positive response for one cluster member as a function of previous outcomes and covariates. This approach is applicable if the cluster members have a natural ordering, such as in longitudinal studies or studies concerning birth order.

4. *Marginal models.* Model the marginal probabilities in terms of covariates, often treating the correlation among cluster members as nuisance parameters, while focusing primarily on marginal mean parameters.

5. *Cluster-specific models.* Allow the model for each cluster to differ by including cluster-specific parameters, which can describe the correlation structure within the cluster. Since the number of such parameters grows along with the number of clusters, a popular approach is to consider these cluster-specific parameters as a **random sample** from some underlying distribution.

Once a modeling strategy has been chosen, there is also the issue of which method or methods can be used to fit the model. Because of the complexity of specifying a complete joint distribution for the set of correlated responses and the associated computational burdens, **maximum likelihood** estimation within a classical statistical approach is not always feasible. However, weighted **least squares**, conditional **likelihood**, **quasi-likelihood**, and different types of approximations to the desired likelihoods have been used. **Bayesian** approaches to specification and estimation in such models have gained a lot of attention recently, due to improved computational approaches to implementing the analyses and some attractive properties. However, they too provide

some challenges, such as the difficulty in specifying hyperparameters and the potential for badly behaved posterior distributions.

Naïve Approaches and Response Feature Models

Naïve approaches refer to the analysis of clustered binary data, which ignore the correlation between subunits. The advantage of such a method is that standard tools and familiar models can be used for the analysis. When all data are used but the association among subunits is ignored, the regression estimators will still be consistent for the parameters of a marginal model.

In some situations, the multivariate response from each cluster might be reduced to a univariate response without major loss, thus simplifying the situation considerably. Once the data have been reduced to a univariate response per cluster, the usual techniques applicable to independent univariate observations are appropriate. For example, an ophthalmologist may only be interested in analyzing binary outcomes from the "worst eye" or "best eye", which would allow the use of standard logistic regression techniques. In the case of longitudinal data, one might be interested in the peak response across a series of repeated measurements or the time until the first positive response (*see* **Summary Measures Analysis of Longitudinal Data**). It is important that this feature of the data be well-defined before the data were collected, rather than suggested by a particular sample.

There are a number of disadvantages of naïve approaches. Ignoring some of the data, such as choosing only one member of the cluster, is not fully **efficient** and **standard error** estimates will be incorrect. By summarizing the responses, we can lose valuable information and the chance to capitalize on the relationships among cluster members. Another difficulty is how to properly use within-cluster covariates. With response feature transformations that only take the data from one subunit, such as analyzing the "worst eye" of a patient with measurements on two eyes, using the covariate corresponding to the worst eye might have the undesirable effect of selecting the covariate on the basis of the response. However, with the idea of creating a "better" person-specific covariate value, sometimes functions of the subunit-dependent covariates are used, such as taking the sum or difference.

Conditional Specification of Models

Conditioning plays an important role in statistical methodology when independence assumptions are not applicable. We say that a model is *conditionally specified* if the joint distribution of the data is built up from a set of conditional distributions. In some settings, it is conceptually easier and more natural to specify such conditional distributions, viewing the joint distribution as merely a consequence of their synthesis. A key implication is that the model parameters are interpreted in terms of conditional probabilities, rather than in terms of joint or marginal probabilities.

Conditioning enters only on the cluster level, so it is reasonable to simplify the discussion and to consider models for one representative cluster of size n_i, that is, $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n_i})$, where $y_{i,j}$ is the binary response of the jth member of the ith cluster.

General Loglinear Model

The joint distribution of $\mathbf{y}_i$ in a general **loglinear model** [17, 66], is

$$\log(\Pr(\mathbf{y}_i)) = \gamma_i' \mathbf{y}_i + \delta_i' \mathbf{w}_i - A(\gamma_i, \delta_i). \qquad (3)$$

Here, $\mathbf{w}_i$ is a vector of length $(2^{n_i} - n_i - 1)$ that contains all cross-products of y_i, that is, all pairwise products $(y_{i,j} y_{i,k},\ j \neq k)$, three-way cross-products $(y_{i,j} y_{i,k} y_{i,l},\ j \neq k \neq l)$, up to the n_i-way cross-product $(y_{i,1} y_{i,2\ldots}, y_{i,n_i})$. $\boldsymbol{\gamma}_i$ and $\boldsymbol{\delta}_i$ are vectors of parameters, and $A(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$ is a normalizing constant.

The components of $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \ldots, \gamma_{i,n_i})$ have an interpretation in terms of the **conditional probability**:

$$\gamma_{i,j} = \text{logit}[\Pr(y_{i,j} = 1 | y_{i,k} = 0, k \neq j)]. \qquad (4)$$

Rather than using correlation parameters to characterize within-cluster relationships, this parameterization uses conditional odds ratios. The components of $\boldsymbol{\delta}_i = (\delta_{i,12}, \delta_{i,13}, \ldots, \delta_{i,12\ldots n_i})$ can be interpreted in terms of contrasts of log conditional odds ratios, which determine the second-order and higher-order associations between the subunits within a cluster.

The general regression strategy is to model $(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$ as a function of covariates $\mathbf{x}_{i,j}$ and unknown parameters $\boldsymbol{\theta}$. Depending on the setting and the questions to be addressed, modeling either the conditional probabilities or the marginal probabilities could be the more

natural approach. If the marginal model approach is preferred, the $\boldsymbol{\gamma}_i$ could be transformed to marginal probabilities, $\Pr(y_{i,j} = 1)$, and the log-conditional odds ratios to pairwise and higher-order correlations or other marginal measures of association. We note that such correlations will have a range of possible values constrained by the marginal probabilities and other correlation parameters. Because the number of association or correlation parameters in a fully specified joint distribution becomes large quickly as the cluster size increases, another issue in our modeling strategy is the question of which higher-order associations we can set to zero to simplify matters. For example, if we set all the associations to be equal to zero, that is $\boldsymbol{\delta}_i = 0$ for all i, then all the subunits within clusters are assumed independent. If we then model $\gamma_{i,m} = \mathbf{x}'_{i,m}\boldsymbol{\beta}$, we obtain the independent logistic regression model. For correlated data, a reasonable choice might be to select a set of higher-order association parameters to be zero. One computational problem that affects models based on (3) is that the normalizing constant $A(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$ often needs to be calculated explicitly.

These two modeling choices generate four different strategies for conditional modeling:

1. Model the conditional probabilities and not assume any of the associations are zero. Examples of this sort of model include saturated log-linear models.
2. Model the conditional probabilities directly, and fix the three- and higher-way moments to zero. Examples of these models are autologistic and response conditional models.
3. Model the marginal probabilities that are transformations of the conditional probabilities, and not assume that any of the associations are zero. Examples of this sort of model include mixed parameter models.
4. Model the marginal probabilities and assume the three- and higher-way moments are zero. Models that could accommodate this assumption include quadratic exponential models.

A particular research question may require specification of a model for the joint distribution, while a second research question might be better addressed by a model for the marginal distributions, or possibly marginal means and first-order associations. What method is used to fit a model depends not only on the assumptions the investigator is willing to make and the questions to be answered, but also (to some extent) on the availability of computational **algorithms**.

Marginal models for categorical data have been commonly fitted using weighted least squares methods, but this method has difficulty with sparse data. Traditionally, maximum likelihood methods for fitting categorical data marginal models have not been widely used, due to their perceived complexity. However, this is changing with more recent work. In the more general setting of modeling marginal means using any generalized linear model, the **generalized estimating equation** approach (discussed further in the Marginal Models section) has become common.

Autologistic Model

Consider a conditionally specified model where third- and higher-order associations are set to zero. In developing methods for **spatial** statistics, Besag [7] introduced the autologistic model having precisely this form:

$$\text{logit}[\Pr(y_{i,j} = 1 | y_{i,1}, y_{i,2}, \ldots, y_{i,j-1}, y_{i,j+1}, \ldots, y_{i,n_i})] = \alpha_j + \sum_{k \neq j} \beta_{j,k} y_{i,k}. \quad (5)$$

Although this proposal did not include covariates, it was perhaps the first extension of the linear logistic regression model to dependent binary data. Likelihood equations can be obtained from the joint distribution of the data and solved iteratively (*see* **Optimization and Nonlinear Equations**). However, as mentioned above, the requirement that the normalizing constant be evaluated during each iteration often causes difficulties. As an alternative estimation strategy, Besag [7] suggested the use of a **pseudo-likelihood** formed from the full conditional distributions. Logistic regression **software** can be used to compute the pseudo-likelihood estimates. Simply, the responses $y_{i,k}$, which are assumed to have a nonzero association with $y_{i,j}$, are used as covariate values for that case. Thus, while the model may properly account for dependence, the pseudo-likelihood estimation differs from likelihood estimation in that it ignores certain aspects of this dependence.

Given recent computational advances, exact likelihood evaluation is often possible, at least with relatively small clusters. In special cases, the normalizer

A collapses, or can be absorbed into a computable quantity [1]. Geyer & Thompson [19] present a **Markov chain Monte Carlo** method that enables likelihood inference.

Response Conditional Models

A flexible extension to the autologistic model in (5) allows covariates and a nonlinear contribution from associated responses.

Rosner [50] presents a model where dependence on the other responses in the ith cluster comes through their sum $s_{i,j} = \Sigma_{k \neq j} y_{i,k}$. The full conditional specification is

$$\begin{aligned} &\text{logit}[\Pr(y_{i,j} = 1 | y_{i,k}, k \neq j)] \\ &\quad = F(s_{i,j}, n_i, \theta_1, \theta_2) + \mathbf{x}'_{i,j}\boldsymbol{\beta}, \end{aligned} \tag{6}$$

where $\mathbf{x}_{i,j}$ is a covariate vector for the jth cluster member, $\boldsymbol{\beta}$ is a regression parameter, n_i is the size of the ith cluster, and

$$F(s, n, \theta_1, \theta_2) = \log \left\{ \frac{(\theta_1 + s\theta_2)}{[1 - \theta_1 + (n - 1 - s)\theta_2]} \right\}. \tag{7}$$

When the term $\mathbf{x}'_{i,j}\boldsymbol{\beta}$ equals 0 for all subunits of the cluster, Rosner's model reduces to the **beta-binomial distribution**, a model commonly used to account for overdispersion in binary data measured without covariates.

Several authors have proposed more general forms for F in this model, allowing negative intraclass correlation and extensions to multivariate **time series** [10, 37, 49], (*see* **Multiple Time Series**).

Transitional Models

The previous section details general strategies and models via the conditional distribution, where the underlying probabilities of interest are conditioned on all other responses in the cluster: $\Pr(y_{ij} | y_{ik}, k \neq j)$. An important special case occurs when the subunit responses within a cluster have an inherent ordering. The obvious classical example involves ordering over time, such as longitudinal data collected repeatedly and with familial data with an ordering over generations or birth order. In these cases, the analyst would not be interested in comparing or modeling the full conditional probabilities, since each would be conditional on events occurring both in the past and in the future. Instead, models that retain the inherent ordering and logical consistency of conditioning on only past responses are of interest. These models are generally termed *transitional* models.

The cornerstone of transitional models lies with **Markov chains** and **Markov processes**, which model the conditional probability given q prior outcomes. With binary responses, the collection of conditional probabilities can be combined into a transition matrix. To model the conditional probabilities as a function of covariates, regression models that explicitly use the past responses as additional covariates have been used. In particular, these build on the model by Cox [11], who is recognized as among the first to describe a transitional model for binary data. Specifically, given a first-order Markov structure, the log likelihood could be written as the sum of marginal and conditional likelihoods: $l(y_1) + l(y_2|y_1) + \cdots + l(y_n|y_{n-1})$. Ignoring the initial term, each log-likelihood component could be easily seen to be the same as sampling from a **multinomial distribution**, with the matrix of transition counts sufficient for the unknown transitional probabilities. Models for the probabilities could be formed by considering logistic equations of the form

$$\begin{aligned} &\text{logit}[\Pr(y_{i,t} = 1 | y_{i,(t-1)})] \\ &\qquad = \alpha + \beta_1 y_{i(t-1)}, + \text{ other terms.} \end{aligned} \tag{8}$$

Higher-order Markov models could be fitted, enhancing $\beta_1 y_{i,(t-1)}$ with other functions of $y_{i,(t-2)}$, $y_{i,(t-3)}, \ldots$. This model can be fitted using standard logistic regression software, and the interpretation of covariate regression coefficients is as a log odds ratio of outcome, given two people with identical past responses in addition to holding other covariates constant. **Interaction** terms of prior outcomes and covariates can be incorporated to see if the effect of a covariate is dependent on past responses.

Subsequent work by others has extended this idea in conjunction with random-effects models and generalized estimating equation approaches. When transitions between states do not occur at equally spaced (time) intervals, continuous time analogs of Markov chains can be used. Kalbfleisch & Lawless [33] provide a comprehensive discussion of extension of these models with covariates and efficient computation of parameter estimates. Heagerty [29]

takes a different approach by using a generalized linear model to relate the marginal means (i.e. $\Pr(y_{i,j} = 1|\mathbf{x}'_{i,j})$) to the covariates for that time point, assuming that the responses in the past do not influence covariate values of future observations. A separate pth order dependence model (e.g. Markov model) is used to capture the serial dependence, and together they fully specify the parametric model, which he describes as a *marginalized transition model of order p* (MTM(p)). Maximum likelihood estimation is possible, and the method allows for data both missing completely at random (MCAR) and missing at random (MAR) [39] (*see* **Missing Data**).

Marginal Models

In the previous section, the natural parameterization of the joint distribution was in terms of conditional probabilities. Often, the research question of interest is more appropriately described in terms of marginal parameters for modeling the relationship between the response and covariates. The parameters of such models have a "population-averaged" interpretation in the sense that the effect of the covariates is averaged across clusters or subsets with different values of cluster-level covariates which form the population, rather than conditional on parameter(s) identifying one particular cluster. Bahadur [6] specified the joint distribution of correlated binary data $(y_{i,1}, \ldots y_{i,n_i})$ in terms of marginal means, $\pi_{i,j} = \mathrm{E}(y_{i,j})$ standardized **residuals**, $e_{i,j} = (y_{i,j} - \pi_{i,j})/[\pi_{i,j}(1 - \pi_{i,j})]^{1/2}$, and correlation-type association parameters, $\rho_{i,12\ldots q} = \mathrm{E}(e_{i,1} \ldots e_{i,q})$, $q = 2, 3, \ldots, n_i$. A regression-type model could be used to model $\pi_{i,j}$ as a function of the covariates, and the usual maximum likelihood methods could be applied. Unfortunately, these association parameters are functions of the marginal probabilities $\pi_{i,j}$ and other association parameters, which constrains their values. The complexity of these interrelationships has deterred efforts to model all these parameters in terms of the covariates, particularly when only the marginal means, not the association parameters, are of scientific interest. To overcome this problem, models and methods have been formulated, which separate the marginal mean model from the specification of the dependency structure, often dealing specifically with only the (first-order) mean and second-order correlation components, operating under the assumption that the higher-order association parameters are not of critical importance.

Quadratic Exponential Models

The model introduced by Gourieroux et al. [22] is based on the loglinear model in (3), where the conditional probabilities $\boldsymbol{\gamma}_i$ are transformed to marginal probabilities, the second-order log conditional odds ratios in $\boldsymbol{\delta}_i$ are transformed to correlation parameters, and the three- and higher-way associations of $\boldsymbol{\delta}_i$ are set to zero. These authors show that maximum likelihood estimation yields consistent and asymptotically normal estimates.

Unfortunately, this estimation method requires calculation of third and fourth **moments**, which requires estimation of a normalizing constant. Zhao & Prentice [66] proposed solving a set of related estimating equations in lieu of the score equations. Further discussion of this and related models can be found in the section "Modifications and Extensions of the GEE method".

Generalized Estimating Equation (GEE) Approach

An estimation approach that places the emphasis on estimating marginal mean parameters, while treating the association parameters as nuisance parameters, is called the *generalized estimating equation (GEE) approach* [36], for which software is readily available [9, 51, 54, 58], The marginal means are modeled via any generalized linear model (GLM), which includes the familiar **linear regression** and logistic regression models. For binary data, the logit (i.e. logistic regression), probit, or complementary log–log links are commonly used to relate the marginal mean to the linear combination of the covariates (i.e. the linear predictor $\mathbf{x}'_{i,j}\ \boldsymbol{\beta}$).

If the analyst incorrectly assumed that all observations, both within and between clusters, were independent, maximum likelihood estimation of the β regression coefficients using standard software for generalized linear models would result in estimates that were **consistent**, but not efficient. To obtain better efficiency, the association within clusters must be built into the estimation method. The GEE method provides a way to do this.

The introduction of GLMs expanded the classical regression model by allowing the expected value of the response to be a nonlinear function of the

linear predictor and the variance of the responses to depend on the expected value. The relationship between the variance and the expected value, however, is restricted to those found in **exponential family** distributions. The data distribution in GLMs is completely specified, and thus, maximum likelihood estimation is possible. This estimating strategy is optimal in the sense that the solution to the score equations has minimum asymptotic variance among all estimates that are obtained from **unbiased** estimating equations.

If the constraint that the marginal distributions have exponential family form is relaxed so that the variance can be an arbitrary function of the mean, then we obtain quasi-likelihood models [62]. For these models, which still make between- and within-cluster independence assumptions, a quasi-score equation is derived via a quasi-likelihood function. The estimators obtained by solving the quasi-score function (estimating equation) are "optimal" in the sense that they have smallest variance among a class of linear unbiased estimators.

The GEE approach extends quasi-likelihood models by including a within-cluster "working" correlation matrix in the quasi-score (estimating) equations. The analyst can specify a form of this within-cluster correlation matrix or allow it to be completely unspecified. For example, one could specify a common correlation for every pair of cluster members (which would be assumed the same in every cluster), called an "exchangeable" correlation structure, or an autoregressive structure, where cluster members closer in time or space would be assumed to be more highly correlated than those further apart (*see* **ARMA and ARIMA Models**). This method allows unequal cluster sizes, but any missing data are assumed missing completely at random (MCAR) in the sense of Little & Rubin [39] (*see* **Nonignorable Dropout in Longitudinal Studies**).

The GEE method will produce consistent and asymptotically normal estimates of the $\boldsymbol{\beta}$ parameters (assuming some weak regularity conditions and the correct specification of the mean), even if the working correlation structure is specified incorrectly. The stronger the within-cluster correlation and the closer the working correlation is to the true underlying correlation, the higher the gain in efficiency. The resulting estimates of correlation parameters generally will not have good statistical properties. If they are considered nuisance parameters of no scientific interest, this lack of useful estimates is of little concern. Wald tests are used to assess the magnitude of the $\boldsymbol{\beta}$ parameters (*see* **Chi-square Tests**).

Modifications and Extensions of the GEE Method

The association parameters in the original GEE method were parameterized by pairwise moment correlations, which are constrained by the marginal probabilities, and other higher-order correlation parameters in the binary data setting. Because of this constraint and their treatment as nuisance parameters, other measures of association have been suggested and explored. Pairwise odds ratios, relative risks, and tetrachoric correlation have been considered. In practice, little difference has been found among the estimates of regression coefficients obtained using different measures of within-cluster dependencies.

As mentioned previously, an alternative approach to modeling marginal parameters of a multivariate binary distribution is to start with a characterization of the joint distribution in terms of conditional parameters and then transform to marginal parameters. An estimate of the normalizing constant is required to calculate third and fourth moments and the computation burden increases as the cluster size increases. As an alternative, an extension of the GEE method was proposed, which includes estimating equations for association parameters, treating them as scientific parameters of interest and allowing them to change as a function of the covariates. It uses a "working" structure that replaces the actual third and fourth moments with working estimates [48]. If the correlation (or other measure of association) between subunits changes for different covariate values, the parameters and their variance estimates should be closer to the truth using this approach and therefore more efficient. This extension of the GEE method, in which the estimation of the modeled mean and second-order association/covariance parameters are done simultaneously, is often referred to as the GEE2 method.

An advantage of the GEE2 over the GEE method is that the association parameters can now be treated as parameters of interest, with the same asymptotic normality and consistency properties of the GEE method. Like GEE, efficiency is improved if the working third- and fourth-order moments are approximately correct. Both GEE and GEE2 assume that the model for the mean is correctly specified. The GEE2

also assumes that the model for the dependency parameters is correct. If not, the consistency property of the mean regression parameters is lost – a trade off towards improved efficiency if both are correct.

Marginalized and Mixed Parameter Models

Several authors have taken a likelihood-based approach rather than a GEE estimation approach to the estimation of marginal means as a function of covariates (e.g. [17, 24, 26, 29, 43]). Likelihood-based methods have the advantages of a well-established framework for inference, **goodness-of-fit** measures, and the ability to accommodate data both MCAR and MAR [39]. This is in contrast with GEE estimation, which can only handle data MCAR. The basic approach is to characterize the relationship between the marginal mean and the covariates via a generalized linear regression model and complete the specification of the full multivariate distribution via canonical or marginal higher-order correlation parameters. In contrast to quasi-likelihood and GEE estimation approaches, the complete specification of the multivariate distribution allows the use of maximum likelihood estimation (when computationally feasible). For example, the mixed parameter model of Fitzmaurice & Laird [16] was built on the same family of distributions as described in (3), but uses a different parameterization. Instead of transforming from the canonical parameters ($\boldsymbol{\gamma}_i$, $\boldsymbol{\delta}_i$) to the marginal moments ($\boldsymbol{\mu}_i$, $\boldsymbol{\Delta}_i$), as was done by Zhao & Prentice [66], they use a one-to-one transformation to ($\boldsymbol{\mu}_i$, $\boldsymbol{\delta}_i$), forming a "mix" of marginal mean and canonical association parameters. The marginal mean is modeled as a function of the covariates via a link function. The canonical association parameters can also be modeled as a function of the linear predictor $\mathbf{z}'_i\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a parameter vector and $\mathbf{z}_i$ is a set of covariates, as in the GEE2 method. The elements of $\mathbf{z}_i$ could, for example, be indicator variables setting some higher-order associations to zero or a subset of the between-cluster covariates for the mean. Similarly, Azzalini [5] reparameterized Markov models to allow regression modeling of the induced marginal means. In this sense, both Fitzmaurice & Laird [16] and Azzalini [5] "marginalized" the response conditional model of (3).

Heagerty [26], generalized by Heagerty & Zeger [27], builds on this approach, "marginalizing" the latent variable model (**generalized linear mixed model**) described in the "Cluster-specific Models" section. These marginally specified mixed models use a general linear model to characterize the relationship between the marginal means and covariates, for example, $\text{logit}(\Pr(y_{i,j} = 1|\mathbf{x}'_{i,j})) = \mathbf{x}'_{i,j}\boldsymbol{\beta}$, and a general linear mixed model for the dependency parameters, captured by modeling the conditional mean $\mu^b_{i,j} = \Pr(y_{i,j} = 1|x'_{i,j}, b_{i,j})$, conditioned on unobserved subcluster specific latent variables $b_{i,j}$:

$$\text{logit}(\Pr(y_{i,j} = 1|\mathbf{x}'_{i,j}, b_{i,j})) = \Delta(\mathbf{x}'_{i,j}) + b_{i,j}. \quad (9)$$

Here, the $b_{i,j}$ represent cluster subunit-specific random effects with joint distribution $f_\alpha(\mathbf{b}_i|\mathbf{X}_i)$ characterized by the parameter $\boldsymbol{\alpha}$, and $\Delta(\mathbf{x}'_{i,j})$ is a parameter implicitly defined by both the marginal linear predictor $\mathbf{x}'_{i,j}\beta$ and the random-effects distribution $f_\alpha(b_{i,j})$. The $b_{i,j}$ can be modeled further, such as $b_{i,j} = b_{i,0}$ (a cluster-level random intercept) inducing variability from cluster to cluster or $b_{i,j} = b_{i,0} + b^*_{i,1}x^*_i$ (different variability depending on value of cluster-level covariate x^*_i – for example, $b_{i,j} = b_{i,0}$ for control clusters ($x^*_i = 0$) and $b_{i,j} = b_{i,0} + b^*_{i,1}$ for treatment clusters ($x^*_i = 1$)). In longitudinal data settings, one might assume that the $b_{i,j}$ have an autoregressive covariance structure, to model decreasing correlation for measurements further apart. Such approaches assume that the observations $(y_{i,1}, y_{i,2}, \ldots, y_{i,n})$ are conditionally independent, given the $\boldsymbol{b}_i$. Using the fact that the marginal mean can be expressed as the expected value of the cluster-specific conditional mean

$$\begin{aligned}\Pr(y_{i,j} = 1|x'_{i,j}) &= \mathrm{E}_{f_\alpha}(\mu^b_{i,j}) \\ &= \mathrm{E}_{f_\alpha}[\Pr(y_{i,j} = 1|x'_{i,j}, b_{i,j})], \end{aligned} \quad (10)$$

$\Delta(\mathbf{x}'_{i,j})$ can be estimated via the convolution equation

$$\Pr(y_{i,j} = 1|x'_{i,j}) = \int (\mu^b_{i,j}) f_\alpha(b_{i,j}|x'_{i,j})\, \mathrm{d}b_{i,j}. \quad (11)$$

Letting $h(x) = \mathrm{g}^{-1}(x)$ as the inverse link function (inverse logit function here), this convolution equation can be written

$$h(x'_{i,j}\beta) = \int h(\Delta(x'_{i,j}) + b_{i,j}) f_\alpha(b_{i,j}|x'_{i,j})\, \mathrm{d}b_{i,j}, \quad (12)$$

thus more explicitly showing the relationship between the parameter $\Delta(\mathbf{x}'_{i,j})$, the linear predictor $x'_{i,j}\beta$, and the distribution of the random effects $f_\alpha(b_{i,j}|x'_{i,j})$.

A key feature of the marginal model approach of Fitzmaurice & Laird [16] is that the joint likelihood for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ separates into distinct likelihoods for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, due to the **orthogonality** of the parameters' spaces. This implies that the information about the within-cluster covariance structure will not asymptotically help estimate $\boldsymbol{\beta}$ or be detrimental if misspecified, as would be the case with GEE2. The inverse of the Fisher **information matrix** can be used to approximate $\text{cov}(\boldsymbol{\beta}, \boldsymbol{\alpha})$. If the marginal model for the mean is specified correctly, but the model for the dependency structure is not, the inverse of the Fisher information matrix can give inconsistent estimates of $\text{cov}(\boldsymbol{\beta})$.

Because of the conditional interpretation of the canonical parameters in $\boldsymbol{\delta}_i$ in the mixed model of Fitzmaurice & Laird, they are most applicable when the cluster sizes are the same. Varying cluster sizes require the estimation of a set of canonical parameters for each cluster size, and the canonical parameterization is not reproducible – meaning that the distribution of a subset of $\boldsymbol{y}_i$ cannot be written in the same form as that of the complete $\boldsymbol{y}_i$ by using a subset of the canonical parameters. The parameterization of $\text{cov}(\boldsymbol{Y}_i)$ and higher-order moments in the marginalized mixed model of Heagerty and Zeger [27], however, does not depend on the dimension of $\boldsymbol{Y}_i$ and thus presents no interpretation difficulties when the number of responses varies across clusters. Since the focus of these models is often the relationship between the marginal means and covariates, one might argue that the association/dependency parameters could simply be viewed as nuisance parameters, whose interpretation is of little interest. What might be of more concern, however, is the appropriateness of setting some higher-order association terms equal to zero to simplify the analysis, because the effect of such assumptions on the marginal mean parameters has not been fully explored. Several authors have investigated the effect of **misspecifying** the distribution of the random effects $f_\alpha(b_{i,j}|x'_{i,j})$ or failing to recognize its dependence on covariates in the marginally specified mixed model approach, and found it can cause substantial bias in the $\boldsymbol{\beta}$ parameters. However, the impact is less in the marginally specified mixed model than that observed in the conditional mixed model, which is discussed in the next section ([28, 42]).

Cluster-specific Models

Cluster-specific models are differentiated from population average or marginal models by the inclusion of parameters that are specific to cluster. The Cox model for a 2×2 contingency table described at the beginning of this article is a simple example. A somewhat more complex cluster-specific model includes a covariate that is linearly related to the log odds of the marginal probability of a positive response. We might also expect the intercept and slope of the relationship to vary from cluster to cluster. A model for this situation would take the following form:

$$\text{logit}[\Pr(y_{i,j} = 1|x_{i,j})] = \beta_{i,1} + \beta_{i,2}x_{i,j}, \qquad (13)$$

where $\beta_{i,1}$ and $\beta_{i,2}$ are the intercept and slope parameters for cluster i. Inference under this model is complicated by the fact that the number of parameters grows with the number of clusters.

Random-effects Models

A popular approach to reducing the number of parameters in a cluster-specific model is to assume that the clusters are a random sample from some underlying population of clusters and that the parameter values for the clusters follow a distribution. A typical choice is the **multivariate normal** (Gaussian) distribution:

$$\begin{bmatrix} \beta_{i,1} \\ \beta_{i,2} \end{bmatrix} \sim \text{N}\left(\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \mathbf{D} \right), \qquad (14)$$

where α_1 and α_2 are the mean intercept and slope values for the population of clusters, and $\mathbf{D}$ is a 2×2 **covariance matrix**. This assumed distribution on the parameters makes this cluster-specific model a random-effects model. Models of this type are also commonly called **mixed-effects**, **hierarchical**, two-stage, and **empirical Bayes** models (*see* **Multilevel Models**).

If we define

$$b_{i,1} = \beta_{i,1} - \alpha_1 \text{ and } b_{i,2} = \beta_{i,2} - \alpha_2, \qquad (15)$$

then $b_{i,1}$ and $b_{i,2}$ are the deviations from the mean intercept and slope term for the ith cluster, sometimes described as *latent variables,* capturing an unmeasured, underlying variable. We can rewrite the model for the ith cluster's response in terms of the mean intercept and slope α_1 and α_2 (fixed effects) and the

(unobserved) individual deviations $b_{i,1}$ and $b_{i,2}$ (random effects) as:

$$\text{logit}[\Pr(y_{i,j} = 1|b_{i,1}, b_{i,2}, x_{i,j})] = (\alpha_1 + b_{i,1}) + (\alpha_2 + b_{i,2})x_{i,j}, \begin{bmatrix} b_{i,1} \\ b_{i,2} \end{bmatrix} \sim \text{N}(\mathbf{0}, \mathbf{D}). \quad (16)$$

In this formulation, the fixed effects are interpreted as the typical parameter values for the population, while the random effects modify the average parameters to be specific to that cluster. Unless the link is linear, predicted values from the fixed effects only will not produce a prediction of the mean response.

Conditional independence within the cluster is commonly assumed in random-effects models. That is, $y_{i,j}$ (given $b_{i,1}$, $b_{i,2}$, $x_{i,j}$) and $y_{i,j'}$ (given $b_{i,1}, b_{i,2}, x_{i,j'}$) are assumed independent for all $j \neq j'$, where j indexes the member within the cluster. Follmann & Wu [18] have proposed a random-effects model that accounts for informative missing data.

Inference in random-effects models is always based on a marginal or conditional distribution of the data, which does not include the random effects $b_{i,1}$ and $b_{i,2}$. This solves the problem of the number of parameters depending on the number of clusters.

Marginal Inference

If a parametric distribution G is assumed for the cluster-specific parameters, then the usual way to proceed is to obtain the marginal distribution of $\boldsymbol{y}_i$ by integrating out the random effects. This resulting distribution will depend on G but not on the random effects themselves. While G must be specified and estimated, inference about the regression coefficients tends not to be sensitive to misspecification of G if the marginal distribution of $\mathbf{y}_i$ is rich enough [44]. This should also hold for sensitivity to the assumption of conditional independence within the cluster. However, some recent work has shown greater sensitivity to unrecognized dependence of the distribution G on covariate values. [28]. In most cases, there is no closed-form expression for the marginal distribution, and direct maximum likelihood estimation requires either **numerical integration** or **Monte Carlo** integration. Alternatively, estimation can be accomplished using a maximum likelihood or **restricted maximum likelihood** approach based on closed-form approximations to the relevant distributions. These include implementations of the **EM algorithm** and generalized estimating equations [56, 65]. At this time, publicly available software for marginal estimation in cluster-specific models is becoming more readily available (MIXOR [30], PROC NLMIXED [52], GLMMIX Macro [53], EGRET [55].) When there are sufficient observations per cluster, a simple two-stage estimation procedure can be used as an approximation to maximum likelihood estimation [35]. First, individual logistic regressions are fit to each cluster. Secondly, weighted averages of these estimates are used to estimate the fixed effects.

Conditional Inference

Conditional estimation avoids the need to estimate the parameters in the random-effects distribution (*see* **Conditionality Principle**). This is accomplished by deriving the distribution of $\mathbf{y}_i$ conditional on **sufficient statistics** for the cluster-specific parameters. When the canonical link is used in a generalized linear model (such as the logit link with binomial errors), a sufficient statistic can be found. In general, however, such a sufficient statistic may not exist.

One simple model for which conditional estimation is tractable is a model with only one random effect corresponding to a cluster:

$$\text{logit}[\Pr(y_{i,j} = 1|\mathbf{x}_{i,j})] = \alpha_i + \beta_1 x_{i,j,1} + \beta_2 x_{i,j,2} + \cdots + \beta_q x_{i,j,1}. \quad (17)$$

Conditioning on the total number of responses in the ith cluster removes the α_i from the likelihood. The resulting conditional likelihood derives from a permutation argument [8]. Small-sample inferences for this model, which are similar to the well-known **Fisher's exact test** for 2×2 tables are also available and have been implemented in the LogXact and Proc-LogXact packages [12] (*see* **Exact Inference for Categorical Data**; **Logistic Regression, Conditional**; **StatXact**).

Interpretation of Regression Parameters

Since the models discussed above will often lead to different interpretations of the $\boldsymbol{\beta}$ parameters, an understanding of their differences is crucial. The choice of an appropriate model will be guided by how

well the interpretations address the research question of interest.

In the response conditional models, the interpretation of the $\boldsymbol{\beta}$ requires both the responses and the covariate values of the other cluster members to be held fixed. Also, the parameter interpretations change with cluster size, so it is more appropriate for data with common cluster sizes. For example, in a study of twins (cluster size of two) (*see* **Twin Analysis**), the regression coefficient for the kth covariate, β_k, represents the increase in log odds of the response related to a one-unit increase in the kth covariate for a particular twin, when holding all other covariates fixed *as well as* the response of the other twin. This dependence of the interpretation on the outcome of the sibling makes it different from the usual interpretation of a regression coefficient in logistic regression, underscoring the distinction between conditional and unconditional models. If the intent is to predict the outcome of one sibling (say, the one born second), conditioning on the information provided by the first-born may be highly desirable. However, if the intent is to study the association between outcome and a covariate, such conditioning would not be desirable.

The issue of how to interpret and compare the kth regression parameter, β_k, from a response conditional model relative to its counterpart in a marginal or cluster-specific model has been investigated by several authors [29, 45, 47]. Key factors are the magnitude and direction of the within-cluster dependencies both of the response measure and of the covariates.

Comparing the interpretations of β_k in the marginal model and as a fixed-effects parameter in a cluster-specific model has also received a great deal of attention. The difference between these two models may be difficult to internalize, partly because intuition carried over from classical linear regression models (correlated Gaussian data with the identity link) breaks down when considering non-Gaussian data (e.g. binary) and nonidentity links (e.g. logit link). We illustrate this with an example presented by Zeger et al. [65]. In a study of respiratory infections in children aged 7 to 11, the presence or absence of a respiratory infection in the previous year was recorded each year for 5 years. Thus, the child constitutes a cluster and the five observations over time per child are the subunits of the cluster. The mother's smoking status (Yes or No) was recorded at the beginning of the study, but not updated at follow-up visits. Thus the mother's smoking status would be a between-cluster covariate, assumed not to change over the five years.

Consider first a marginal model with an intercept, mother's smoking status, and age of the child at the time of measurement. If there were independence among responses across time, the interpretation of the regression coefficient would be precisely that of a simple logistic regression model. That is, every response would form its own cluster of size one. In the marginal model, the dependence is recognized in the estimation methodology, but not in the model for the marginal mean. The parameter β_k for smoking status represents the difference in the log odds ratio for respiratory infection between children with a smoking mother at baseline and those whose mother did not smoke at baseline. Mathematically, this is the difference in the log odds of the mean risk between these two groups, where the mean is taken over all children (clusters) and all observations (subunits), weighted by the working dependency structure used in the estimation method.

In a cluster-specific random-effects model with fixed effects for the same covariates as listed above, plus a random effect for child (i.e. a random cluster effect), the interpretation is conditional on that random effect for the child. Within this child, the coefficient represents the magnitude of change in the log odds one would expect with his mother smoking at baseline versus his mother not smoking at baseline. (Of course, this effect is unobservable because we cannot go back in time and change the mother's baseline smoking status.) Since the model specifies that this coefficient is the same for all children, it is estimated by combining information from different children, such as averaging over all children according to the distribution of that random effect for the child. Because the effect we are trying to measure is not observable, its estimation is heavily model-based.

If the mother's smoking status was recorded at each visit, thus considering smoking a within-cluster covariate, some mothers might change smoking status during the course of the study; hence, information on the effect of that change would be available. The cluster-specific model would capitalize on this observable change and allow estimation of the average effect (in terms of log odds ratios) of the change in smoking status on respiratory disease status, assuming there were clusters in which such changes took place. Mathematically, the collapse

of information across different children occurs after taking the difference in log odds at time points where the mother did and did not smoke, as opposed to first averaging across children and time points to obtain mean risks and then computing the log odds. The beta coefficient then represents the assumed common log odds ratio for respiratory disease of mother's smoking status across children, rather than the log odds ratio of the mean risk of respiratory disease.

The marginal model in this setting, on the other hand, would ignore the fact that the effect of change in smoking status was directly observable in some children, and persist in estimating only the odds ratio between smoking and nonsmoking mothers. Mothers who had changed smoking status would appear in both groups. Information obtained when individuals serve as their own controls when a change is observed is not used. In a marginally specified mixed model, the effect of a smoking mother versus a nonsmoking mother also involves averaging over the unobserved latent variables (random effects) in each group.

Extensions for Correlated Ordinal Responses

The focus of this article has been to review modeling strategies for correlated binary data by building upon the standard logistic regression model with a binomial sampling distribution. Models have also been introduced that extend regression models for ordinal responses with a multinomial sampling distribution. A brief review of two classes of extensions, marginal and random-effects, is discussed below. Agresti & Natarajan [3] provide a more complete review of these models.

The correlation structure of repeated ordinal data may be of interest but can be difficult to model. Dependence between ordinal responses has been described by Dale [13], and Heagerty & Zeger [25] discuss an approach extending the correlogram to categorical responses using log odds ratios.

Marginal Models

Stram et al. [57] introduced a marginal model for repeated ordinal measurements (*see* **Ordered Categorical Data**). Assuming that all patients are measured at common time points, they compute time-specific parameters for the **proportional odds model** proposed by McCullagh [40] and then combine these parameters using linear combinations weighted by the variance. However, as with the model of Wei & Stram [63], their model falls short in not allowing for **parsimonious** working correlation structures to be used. This has lead to the extensions of the generalized estimating equations methods for correlated ordinal measurements using cumulative logit models of McCullagh [40] ([24, 38, 41]). The current versions of SAS PROC GENMOD [51] and Stata [54] allow estimation of ordinal GEE models.

Because the GEE models are only valid with datasets that may have data missing completely at random (MCAR), there has been some concern about using these methods when data are missing at random (MAR). Kenward et al. [34] demonstrate potential problems as compared to likelihood-based approaches, such as those proposed by Molenberghs and Lesaffre [43].

Random-effects Models

Random-effects models for ordinal data were first proposed by Harville & Mee [23], whose purpose was best linear unbiased predictors (BLUP) of parameters of an underlying (latent) distribution. A more general approach was discussed by Hedeker & Gibbons [31], which uses a faster Fisher-scoring algorithm for estimating parameters of a model with multiple random effects. This model has been implemented in publicly available software [32]. An alternative method, which has not been specifically implemented, might be to use the Gibbs sampling methods of Zeger & Karim [64] to avoid direct numerical integration.

As mentioned above, another method for overcoming the increasing number of parameters in cluster-specific models is to treat the cluster-specific parameters as nuisance parameters, condition on their sufficient statistics, then maximize this conditional likelihood. Agresti & Lang [2] investigate this method for categorical, cluster-specific covariates.

Summary

Our focus has been on different estimation and modeling approaches for the general problem of clusters of correlated binary responses in the presence of both within- and between-cluster covariates. We have also generally restricted attention to methods

that can handle continuous covariates, rather than open the door to the large and dynamic literature on this topic presented within the general framework of categorical data models. We do not claim to have thoroughly covered or even mentioned all the relevant literature on this topic. For example, more could be said on goodness-of-fit testing and model **diagnostics**, Markov Chain models, estimation in the presence of missing data, design issues, Bayesian methods, small sample properties, adaptations to make these models more **robust**, and computational issues. Interested readers are referred to several books and survey articles that place emphasis on these and related models [4, 14, 15, 20, 21, 44, 46, 59].

References

[1] Agresti, A. (1993). Distribution-free fitting of logit models with random effects for repeated categorical responses, *Statistics in Medicine* **12**, 1969–1987.

[2] Agresti, A. & Lang, J.B. (1993). A proportional-odds model with subject-specific effects for repeated ordered categorical responses, *Biometrika* **80**, 527–534.

[3] Agresti, A. & Natarajan, R. (2001). Modeling clustered ordered categorical data, *International Statistical Review* **69**, 345–372.

[4] Ashby, M., Neuhaus, J.M., Hauck, W.W., Bacchetti, P., Heilbron, D.C., Jewell, N.P., Segal, M.R. & Fusaro, R.E. (1992). An annotated bibliography of methods for analysing correlated categorical data, *Statistics in Medicine* **11**, 67–99.

[5] Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures, *Biometrika* **81**, 767–775.

[6] Bahadur, R.R. (1961). A representation of the joint distribution of responses to *n* dichotomous items, in *Studies in Item Analysis and Prediction*, H. Solomon, ed. Stanford University Press, Stanford.

[7] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

[8] Breslow, N.E. & Day, N.E. (1980). *Statistical methods in Cancer Research*, Scientific Publications No. 32, Vol. 1: The Analysis of Case-Control Studies. International Agency for Research on Cancer, Lyon.

[9] Carey, V. & McDermott, A. (1996). gee() S-Plus function. *StatLib Archives*. Carnegie Mellon University, Pittsburgh.

[10] Connolly, M.A. & Liang, K.-Y. (1988). Conditional logistic regression models for correlated binary data, *Biometrika* **75**, 501–506.

[11] Cox, D.R. (1970). *The Analysis of Binary Data*. Methuen, London.

[12] Cytel Software Corporation (2000, 2001). *LogXact-4 and Proc-LogXact*. Cytel Software Corporation, Cambridge.

[13] Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses, *Biometrics* **42**, 909–917.

[14] Davis, C. (2002). *Statistical Methods of the Analysis of Repeated Measurements*. Springer-Verlag, New York.

[15] Diggle, P.J., Heagerty, P., Liang, K.-Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd Ed. Oxford University Press, New York.

[16] Fitzmaurice, G.M. & Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses, *Biometrika* **80**, 141–151.

[17] Fitzmaurice, G.M., Laird, N.M. & Rotnitzky, A.G. (1993). Regression models for discrete longitudinal responses (with discussion), *Statistical Science* **8**, 284–309.

[18] Follmann, D. & Wu, M. (1995). An Approximate generalized linear model with random effects for informative missing data, *Biometrics* **51**, 151–168.

[19] Geyer, C.J. & Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 657–699.

[20] Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd Ed., Kendall's Library of Statistics 3, Oxford University Press, New York, p. 172.

[21] Goldstein, H., Browne, W. & Rasbash, J. (2002). Tutorial in Biostatistics: mulitlevel modelling of medical data, *Statistics in Medicine* **21**, 3291–3315.

[22] Gourieroux, C., Montfort, A. & Trognon, A. (1984). Pseudomaximum likelihood methods: theory, *Econometrica: Journal of the Econometric Society* **52**, 681–700.

[23] Harville, D.A. & Mee, R.W. (1994). A mixed-model procedure for analyzing ordered categorical data, *Biometrics* **40**, 393–408.

[24] Heagerty, P. & Zeger, S.L. (1996). Marginal regression models for clustered ordinal measurements, *Journal of the American Statistical Association* **91**, 1024–1036.

[25] Heagerty, P. & Zeger, S.L. (1998). Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses, *Journal of the American Statistical Association* **93**, 150–162.

[26] Heagerty, P. (1999). Marginally specified logistic-normal models for longitudinal binary data, *Biometrics* **55**, 688–698.

[27] Heagerty, P. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data, *Biometrics* **58**, 342–351.

[28] Heagerty, P. & Kurland, B. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models, *Biometrika* **88**, 973–985.

[29] Heagerty, P. & Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference, *Statistical Science* **15**, 1–26.

[30] Hedeker, D.R. (1993). *MIXOR – A FORTRAN Program for Mixed-Effects Ordinal Probit and Logistic*

Regression. University of Illinois at Chicago Prevention Research Center, Chicago.

[31] Hedeker, D.R. & Gibbons, R.D. (1994). A random-effects ordinal regression model for multi-level analysis, *Biometrics* **50**, 933–944.

[32] Hedeker, D.R. & Gibbons, R.D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis, *Computer Methods and Programs in Biomedicine* **49**, 157–176.

[33] Kalbfleisch, J.D. & Lawless, J.F. (1985). The analysis of panel data under a Markov assumption, *Journal of the American Statistical Association* **80**, 863–871.

[34] Kenward, M.G., Lesaffre, E. & Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random, *Biometrics* **50**, 945–953.

[35] Korn, E.L. & Whittemore, A.S. (1976). Methods for analyzing panel studies of acute health effects of air pollution, *Biometrics* **35**, 795–802.

[36] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[37] Liang, K.-Y. & Zeger, S.L. (1989). A class of logistic regression models for multivariate binary time series, *Journal of the American Statistical Association* **84**, 447–451.

[38] Lipsitz, S.R., Kim, K. & Zhao, L.P. (1994). Analysis of repeated categorical-data using generalized estimating equations, *Statistics in Medicine* **13**, 1149–1163.

[39] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[40] McCullagh, P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[41] Miller, M.E., Davis, C.S. & Landis, J.R. (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares, *Biometrics* **49**, 1033–1044.

[42] Mills, J.E., Field, C.A. & Dupuis, D.J. (2002). Marginally specified generalized linear mixed models: a robust approach, *Biometrics* **58**, 727–734.

[43] Molenberghs, G. & Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution, *Journal of the American Statistical Association* **89**, 633–644.

[44] Neuhaus, J.M. (1992). Statistical methods for longitudinal and clustered designs with binary responses, *Statistical Methods in Medical Research* **1**, 249–273.

[45] Neuhaus, J.M. (1993). Estimation efficiency and tests of covariate effects with clustered binary data, *Biometrics* **49**, 989–996.

[46] Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M. & Fisher, M.R. (1996). A survey of methods for analyzing clustered binary response data, *International Statistical Review* **64**, 89–118.

[47] Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**, 1033–1048.

[48] Prentice, R.L. & Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics* **47**, 825–839.

[49] Qu, Y., Williams, G.W., Beck, G.J. & Goormastic, M. (1987). A generalized model of logistic regression for clustered data, *Communications in Statistics* **16**, 3447–3476.

[50] Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations, *Biometrics* **40**, 1025–1035.

[51] SAS Institute (1999a). PROC GENMOD. SAS Institute, Cary.

[52] SAS Institute (1999b). PROC NLMIXED. The SAS Institute, Cary.

[53] SAS Institute (1999c). SAS Macro GLMMIX. The SAS Institute, Cary NC.

[54] STATA (2003). *Stata Base Reference Manual*. Stata Press, College Station, TX.

[55] Statistics and Epidemiology Research Corporation (1993). *EGRET Reference Manual*. Statistics and Epidemiology Research Corporation and Cytel Software Corporation, Cambridge, MA.

[56] Stiratelli, R., Laird, N. & Ware, J.H. (1984). Random-effects models for serial observations with binary response, *Biometrics* **40**, 961–971.

[57] Stram, D.O., Wei, L.J. & Ware, J.H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates, *Journal of the American Statistical Association* **83**, 631–637.

[58] SUDAAN (2001). *SUDAAN Users Manual, Release 8.0*. Research Triangle Institute, Research Triangle Park.

[59] Sullivan, L.M., Dukes, K.A. & Losina, E. (1999). Tutorial in Biostatistics: an introduction to hierarchical modelling, *Statistics in Medicine* **18**, 855–888.

[60] Verbeke, G. & Molenberghs, G. (1997). *Linear Mixed Models in Practice*. Springer-Verlag, New York.

[61] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

[62] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**, 439–447.

[63] Wei, L.J. & Stram, D.O. (1988). Analysing repeated measurements with possibly missing observations by modelling marginal distributions, *Statistics in Medicine* **7**, 139–148.

[64] Zeger, S.L., Liang, K.-Y. & Albert, P. (1988). Models for longitudinal data, a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

[65] Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association* **86**, 79–86.

[66] Zhao, L.P. & Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika* **77**, 642–648.

JANE F. PENDERGAST, STEPHEN J. GANGE & MARY J. LINDSTROM

(*See also* **Generalized Linear Models for Longitudinal Data**)

Correlation

In a rather loose sense, two characteristics or variables are said to be correlated if changes in one variable tend to be accompanied by changes in the other, in either the same or the opposite direction. Thus, the incidence of ischemic heart disease is *positively* correlated with the softness of drinking water, since many epidemiologic studies have shown that the incidence tends to be higher in areas with softer water; and, conversely, the incidence is *negatively* correlated with water hardness.

In view of the more specific definitions to be discussed below, it would perhaps be preferable to use the term **association** for this informal usage. Correlation implies a *linear* relationship with superimposed random variation; association often loosely implies a monotone relationship, but the term may also be applied to nominal data where rank order is undefined.

For an account of the early history of the term *correlation*, see [10], especially pp. 297–299. The term was current during the middle of the nineteenth century, but its statistical usage is rightly attributed to Francis **Galton**, who initially used the spelling "co-relation". Galton was concerned with the correlation between characteristics of related individuals; for example, between an individual's height and the mean height of the two parents. The *correlation coefficient* emerged from Galton's work, after further elucidation by F.Y. **Edgeworth** and Karl **Pearson**, to become a central tool in the study of relationships between variables, especially (in Pearson's work) between physiological and behavioral measurements on human beings.

The Product–Moment Correlation Coefficient

The *product–moment correlation coefficient* (normally abbreviated to *correlation coefficient*) is a measure of the closeness of the association to a straight line. If the variables X and Y are **random variables** with a joint probability distribution, then the correlation coefficient is defined as

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{1}$$

where σ_{XY}, σ_X, and σ_Y are, respectively, the covariance of X and Y and the standard deviations of X and Y. The value of ρ is bounded between -1 and $+1$, taking these extreme values only when there is a linear functional relation between X and Y. Thus, if $Y = \alpha + \beta X$ exactly (as, for instance, with temperatures recorded in Fahrenheit for Y and centigrade for X), then $\rho = 1$ if $\beta > 0$ (as in this example), and $\rho = -1$ if $\beta < 0$.

Biologic variables are not normally connected by linear functional relations, and the correlation coefficient usually lies between the two extremes. There is a close connection between the concept of correlation and that of **linear regression**. Let $\beta_{Y.X}$ be the slope of the linear regression of Y on X, and $\beta_{X.Y}$ that of the regression of X on Y. Then

$$\beta_{Y.X} = \frac{\sigma_{XY}}{\sigma_X^2}, \qquad \beta_{X.Y} = \frac{\sigma_{XY}}{\sigma_Y^2}$$

and, from (1), $\beta_{Y.X}\beta_{X.Y} = \rho^2$. Since $\rho^2 \leq 1$, $|\beta_{Y.X}| \leq |1/\beta_{X.Y}|$ (the latter expression being the slope of the regression of X on Y in a diagram with Y as the ordinate), equality being achieved only for perfect correlation. Thus, the two regression lines are in general inclined at an angle. When the correlation coefficient $\rho = 0$, both $\beta_{Y.X}$ and $\beta_{X.Y}$ are zero, and the two regression lines are at right angles.

For a particular value of X, define Y_0 to be the value predicted by the linear regression of Y on X. Then the variance of the residuals about regression, $\mathrm{E}[(Y - Y_0)^2]$ is equal to $\sigma_Y^2(1 - \rho^2)$. One interpretation of the correlation coefficient is, therefore, the fact that its square is the proportion of the variance of one variable that is "explained" by linear regression on the other. The relationship is symmetric, being equally true for the other regression.

The Sample Correlation

The correlation coefficient may also be defined, in similar manner, for a finite set of n paired quantitative observations $(x_1, y_1), \ldots, (x_n, y_n)$. The correlation coefficient, denoted now by r, may be calculated as

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2\right]^{1/2}}. \tag{2}$$

Here, $\overline{x}$ and $\overline{y}$ are the mean values of the x_i and y_i, respectively, and the summations run from 1 to n.

The basic properties of the sample correlation coefficient are essentially those outlined above for random variables. The relationship with regression now applies to the slopes of the **least squares** regression lines, $b_{y.x}$ and $b_{x.y}$. The squared correlation coefficient, r^2, is the proportion of the total sum of squares $\Sigma(y_i - \overline{y})^2$ "explained" by the regression of y on x.

Some scatter diagrams representing simple situations are shown schematically in Figure 1. In Figures 1(a) and (e) the points lie on a straight line, and $r = +1$ and -1, respectively. In Figure 1(c) the variation in one variable is approximately independent of the value of the other variable, and $r = 0$. In Figures 1(b) and (d) there is an intermediate degree of correlation, the variance of one variable being reduced when the value of the other variable is fixed, so $0 < r < 1$ for (b) and $-1 < r < 0$ for (d).

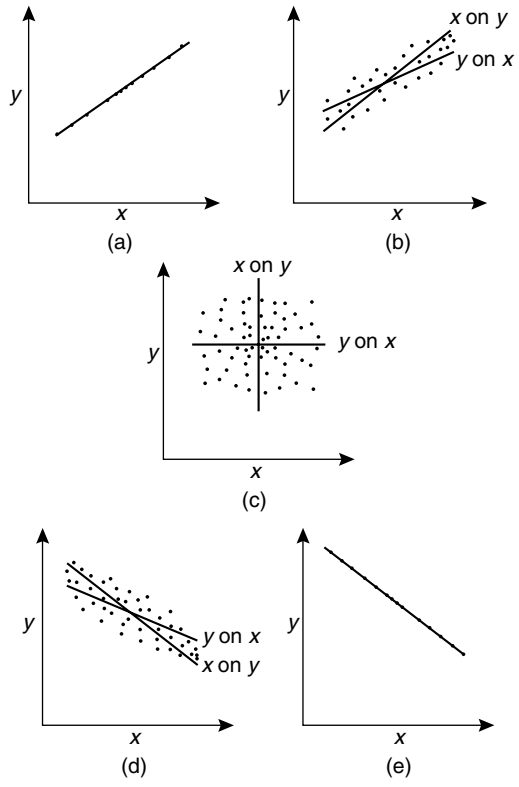

Figure 1 A schematic representation of scatter diagrams with regression lines, illustrating different values of the correlation coefficient. Reproduced from [1] by permission of Blackwell Science, Oxford

The sample correlation coefficient given in (2) is sometimes referred to as the *Pearson product–moment correlation coefficient*, the reference here being to Karl Pearson [7]. It provides the method of calculation for any finite set of paired values, and invites consideration of its sampling error.

Sampling Error

Suppose that the n pairs of observations are drawn at random from a **bivariate distribution** of random variables X and Y. The **sampling distribution** of r will depend on the characteristics of the parent distribution, in particular on the population correlation coefficient ρ. In general, r is a consistent but biased estimator of ρ (*see* **Unbiasedness**), but the bias is of order $1/n$ and likely to be small except for very small values of n (*see* **Estimation**). More specific results are available if stronger assumptions are made about the nature of the parent distribution, and the traditional assumption is that of a **bivariate normal distribution**. This model was widely used in the early work of Galton and Karl Pearson; appropriately enough, since it provides a reasonable description of many of the biometric variables studied by them.

Under the bivariate normal assumption, the distribution of r depends only on ρ and n. The density was first derived in 1915 by Fisher [4] and subsequently tabulated by David [3]. If $\rho = 0$, then the statistic

$$\frac{(n-2)^{1/2} r}{(1-r^2)^{1/2}} \tag{3}$$

follows a **Student's *t* distribution** with $n - 2$ degrees of freedom, and can be used to test the null hypothesis that $\rho = 0$ (*see* **Hypothesis Testing**). In fact, this test is valid more generally, for the standard model for linear regression (*see* **Linear Regression, Simple**), in which the values x of X are chosen arbitrarily but Y is distributed normally with constant variance around a linear function of x.

Returning to the bivariate normal model, Fisher derived a variance-stabilizing **transformation** of r,

$$z = \tanh^{-1} r = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right),$$

the distribution of which approaches normality more rapidly than that of r, as the sample size, n, increases. Asymptotically, $\mathrm{E}(z) = \tanh^{-1}\rho$, and, approximately, $\mathrm{var}(z) = 1/(n-3)$. An alternative transformation [9] provides a generalization of (3): for $\rho \neq 0$, the statistic

$$\frac{(n-2)^{1/2}(r-\rho)}{[(1-r^2)(1-\rho^2)]^{1/2}}$$

follows approximately a t distribution with $n-2$ degrees of freedom. For further details about the distribution of r, see [6, Chapter 10].

Intraclass Correlation

Suppose that observations on a single variable y are arranged in n groups, each containing m observations, and that there is reason to expect possible differences in the mean level of y between groups. If such differences exist, observations in the same group will tend to be positively correlated. This phenomenon is called *intraclass correlation*.

Fisher [5] illustrates this for the case $m = 2$ by referring to measurements on pairs of brothers. He distinguishes between situations in which the brothers fall into labeled categories such as "elder" and "younger", and those in which no such categorization is required. In the first case, there are two variables – measurements for elder and younger brothers – and the standard product–moment correlation coefficient may be calculated. This is called the *interclass correlation*. In the second case, each pair would enter into the calculation twice, since they are not naturally ordered, and the denominator of the correlation coefficient may be based on a single sum of squares about the mean for all the $2n$ observations. This is the *intraclass correlation*. It is interesting to note that Fisher used the term "class" to denote the possible labeling categories ("elder" and "younger" here), whereas many modern writers use it to denote the groups into which the observations are clustered ("sibships" here).

The intraclass correlation coefficient, r_I, may be calculated as a modified variant of (2) for all the $nm(m-1)$ pairs of observations in the same group, each pair being counted twice. In the cross product in the numerator in (2), all deviations are taken from the mean, $\overline{y}$, of all mn observations. Since each observation appears $m-1$ times in the cross product, the denominator of (2) is $m-1$ times the sum of squares of deviations of all mn observations about $\overline{y}$.

An equivalent formula for r_I clarifies the relation between intraclass correlation and variation between the group means. Denote by $\overline{y}_i$ the mean for the ith group, and by v the variance of the mn observations with divisor mn rather than $mn-1$. Then

$$r_I = \frac{m\sum(\overline{y}_i - \overline{y})^2 - nv}{(m-1)nv},$$

the summation running from 1 to n. It follows that

$$-\frac{1}{(m-1)} \leq r_I \leq 1,$$

the lower limit of $-1/(m-1)$ being achieved when all the y_i, are equal, and the upper limit of 1 when there is no variation within the groups so that $\Sigma(\overline{y}_i - \overline{y})^2 = nv$.

Data of the type considered here would normally be analyzed by a one-way **analysis of variance**, and, as Fisher [5] showed, there is a close connection between the two approaches. If, in the analysis of variance, the mean squares between and within groups are denoted by s_b^2 and s_w^2, respectively, then

$$r_I = \frac{s_\mathrm{b}^2 - \left[\frac{n}{(n-1)}\right]s_\mathrm{w}^2}{s_\mathrm{b}^2 + (m-1)\left[\frac{n}{(n-1)}\right]s_\mathrm{w}^2},$$

so, for large n,

$$r_I \sim \frac{s_\mathrm{b}^2 - s_\mathrm{w}^2}{s_\mathrm{b}^2 + (m-1)s_\mathrm{w}^2}.$$

Equivalently, $r_I \sim (F-1)/(F+m-1)$, where F is the usual variance ratio statistic, $s_\mathrm{b}^2/s_\mathrm{w}^2$.

Two approaches may be followed in discussion of the sampling error of the intraclass correlation coefficient, using either **finite population** theory as is usual for **multistage sampling**, or the **random effects** model more usual in biologic applications.

The first approach assumes that the n groups are randomly selected from a larger set of N, and that each set of m observations within a group is randomly selected from M. If the intraclass correlation coefficient for the finite population of MN observations is denoted by ρ_I, the sample value r_I may be regarded as an estimator of ρ_I.

The random effects model effectively assumes infinite values for M and N. The group means are assumed to be distributed with a **variance component** σ_b^2, and the within-group deviations with a component σ_w^2. Then

$$\rho_I = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}.$$

Inferences about ρ_I may be made using standard results for the ***F* distributions** in the one-way analysis of variance.

Some examples of the use of intraclass correlation as a descriptive tool are as follows:

1. In sample survey theory (*see* **Cluster Sampling**; **Multistage Sampling**; **Cluster Sampling, Optimal Allocation**), to indicate the correlation between observations in the same cluster due to systematic between-cluster variation.
2. In statistical genetics, to indicate the correlations in genetic traits between members of the same family: see the articles on **familial correlations** (which deals in detail with the situation in which the family groups vary in size) and **genetic correlations and covariances**. The numerical values of intraclass correlation coefficients are more meaningful in genetics than in most other applications, because of the predictions of Mendelian theory (*see* **Mendel's Laws**), although the predicted values for familial correlations, for example between siblings, may be distorted by the additional effects of environmental correlation.
3. In studies of the reliability of repeated measurements, or the agreement between observers in measuring characteristics of the same subject (*see* **Kappa**; **Observer Reliability and Agreement**). Note that when the same observers are used for each subject, there may be systematic differences in the level of recording for different observers, and the one-way analysis of variance analogue is no longer valid [2].

Some Generalizations

The concepts underlying the product–moment correlation coefficient may be generalized or modified in various ways. The (x, y) pairs may not be closely *linearly* related, but may nevertheless be perfectly associated through a nonlinear relation. If this relation is monotone, the observations will be ranked in the same order by both x and y. Two commonly used coefficients of **rank correlation** are **Spearman's** ρ (essentially the product–moment correlation of the ranks), and Kendall's τ (based on the number of discrepancies in the ranking of paired observations by the two variables). Both coefficients are bounded by the values -1 and $+1$ for perfect negative and positive agreement between the rankings.

When there are more than two quantitative variables, the correlations between pairs play an important part in various methods of multivariate analysis (*see* **Multivariate Analysis, Overview**). In **multiple linear regression**, two generalizations of r are commonly used. The *multiple correlation coefficient, R*, generalizes a property of r, in that the proportion of the variance of the dependent variable y that is "explained" by the multiple regression on $x_1, \ldots, x_k$ is R^2. (Since the squared form carries the essential information, R is never given a negative sign.) The *partial correlation coefficient* measures the product–moment correlation between y and one of the predictor variables, x_i say, when all the other predictors are kept constant. It is therefore useful in assessing the separate effects of different predictors, especially if they are themselves closely correlated.

In the early work of the Galton–Pearson school, much effort was put into the estimation of correlation coefficients when the observed values were nominal – perhaps even binary – but were supposed to represent divisions of some underlying, unobserved, continuous variables. The correlation between the presumed continuous variables had to be estimated from the discrete observations. It was usually assumed that the underlying distribution was bivariate normal. For two binary classifications, the measure is called *tetrachoric correlation*. When one variable is binary and the other is quantitative, the measure is called *biserial correlation*. These methods are less frequently used now, as alternate models for categorical data seem more appropriate (*see* **Association, Measures of**; **Categorical Data Analysis**).

Interpretation

The high profile assumed by the concept of correlation during the early part of the twentieth century has now largely vanished. This is partly due

to the emergence of more penetrating methods of statistical analysis. In particular, the emphasis has gradually moved away from an index measuring a degree of association, to an attempt to describe more explicitly the nature of that association. In a word, the emphasis has moved from correlation toward regression.

Apart from this general shift in viewpoint, there are some specific problems in the interpretation of correlation coefficients, which need to be taken into account in any data analysis in which they are used:

1. Correlation does not imply **causation**. Two variables may be highly correlated because they are both causally related to a third variable or a group of such variables, and yet have no causative relation to each other. Relations of this type are often called *nonsense* or *spurious correlations*. Often, the intervening variable is time. That is, two variables x and y may both be steadily increasing with time, over a certain time period, or one may be increasing while the other decreases. The two variables are then likely to be highly correlated. For instance, during the first two-thirds of the twentieth century, imports of tobacco into the UK increased steadily, as did the number of divorces granted. The two variables, measured in successive decades, are highly correlated. It would certainly not be correct to assume that either variable *caused* the other. Nonsense correlations are among the most prevalent causes of injudicious inferences from statistical data by the general public and the media.
2. Correlation measures closeness to a *linear* relationship. Two variables may be very closely associated by a nonlinear relation, and yet have a low correlation coefficient. As an extreme example, in Figure 2 (reproduced from [8]) is shown a scatter diagram between randomly drawn **standard normal deviates** and their squares. The population correlation coefficient is exactly zero, but it would have been entirely wrong to assume a lack of dependence. Independent random variables have zero correlation; but zero correlation does not imply independence.
3. The correlation between two biologic variables may be affected by selection of particular values of one variable. For instance, if X and Y have a distribution approximating to a bivariate normal distribution, with correlation coefficient ρ, restriction of the range of X by removal of extreme values in both directions will tend to decrease the correlation coefficient below the original value ρ. Thus, the correlation between height and age of children is higher for the age range 5–12 years than for the range 7–8 years. Conversely, omission of central values of X with retention of extreme values will tend to increase the correlation coefficient. This phenomenon may make it difficult to compare correlation coefficients in different populations differing in the degree and type of selection.
4. The effect of sampling variation is often underestimated, so that undue importance is given to moderately high correlations based on few observations. The upper 5 percentile of the distribution of $|r|$ when the population value $\rho = 0$, from (3), is 0.878 for $n = 5$, and 0.632 for $n = 10$. In this sense, moderately large correlations from small numbers of observations are inherently unreliable.

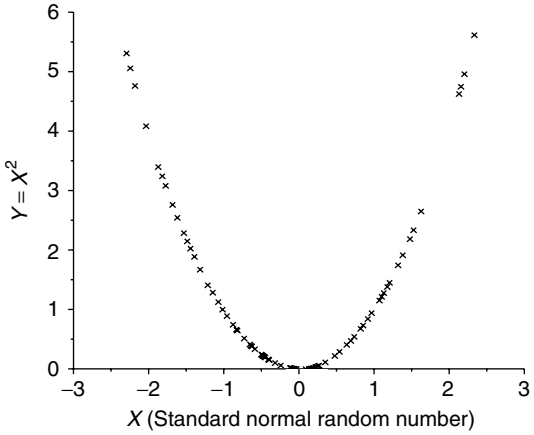

Figure 2 A scatter plot of $Y = X^2$ for 100 standard normally distributed random numbers and their squares. Reproduced from [8] by permission of Wiley, New York

References

[1] Armitage, P., Berry, G. & Matthews, J.N.S. (2002). *Statistical Methods in Medical Research*, 4th Ed. Blackwell Science, Oxford.

[2] Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability, *Psychological Reports* **19**, 3–11.

[3] David, F.N. (1938). *Tables of the Correlation Coefficient*. Cambridge University Press, Cambridge.

[4] Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika* **10**, 507–521.

[5] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[6] Patel, J.K. & Read, C.B. (1982). *Handbook of the Normal Distribution*. Marcel Dekker, New York.

[7] Pearson, K. (1896). Mathematical contributions to the theory of evolution, III: regression, heredity and panmixia, *Philosophical Transactions of the Royal Society of London, Series A* **187**, 253–318.

[8] Rodriguez, R.N. (1982). Correlation, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 193–204.

[9] Samiuddin, M. (1970). On a test for an assigned value of correlation in a bivariate normal distribution, *Biometrika* **57**, 461–464.

[10] Stigler, S.M. (1986). *The History of Statistics.* Belknap Press, Cambridge, Mass.

PETER ARMITAGE

Correlational Study

A correlational study is an **ecologic study** in which rates of disease in populations are correlated with average exposures or other features of such populations. The populations may be defined by geographic regions of residence, for example. Such correlations are useful for generating etiologic hypotheses, but because individual-level information is not available on exposure, disease outcome, and potential **confounders**, such **correlations** are subject to the **ecologic fallacy** and may be misleading.

MITCHELL H. GAIL

Correspondence Analysis of Longitudinal Data

Correspondence analysis is an exploratory tool for the analysis of **association(s)** between **categorical** variables. Usually, the results are displayed in a **graphical** way.

There are many interpretations of correspondence analysis. Here we make use of two of them. A first interpretation is that the observed categorical data are collected in a **matrix**, and correspondence analysis approximates this matrix by a matrix of lower rank [12]. This lower rank approximation of, say, rank $M+1$ is then displayed graphically in a M-dimensional representation in which each row and each column of the matrix is displayed as a point. The difference in rank between the rank $M+1$ matrix and the rank M representation is matrix of rank 1, and this matrix is the product of the marginal counts of the matrix, that is most often considered uninteresting. This brings us to the second interpretation, that is, that when the two-way matrix is a **contingency table**, correspondence analysis decomposes the departure from a matrix where the row and column variables are independent [8, 9]. Thus, correspondence analysis is a tool for **residual** analysis. This interpretation holds because for a contingency table estimates under the independence model are obtained from a product of the margins of the table (divided by the total sample size).

Longitudinal data are data where observations (e.g. individuals) are measured at least twice using the same variables. We consider here only categorical (i.e. nominal or ordinal) variables, as only this kind of variables is analyzed in standard applications of correspondence analysis [7].

Two Time Points

When there is one categorical variable measured at two time points, a so-called transition matrix can be constructed [1]. In this transition matrix, the row variable is the categorical variable measured at time 1, and the column variable is the categorical variable at time 2. The aim of a correspondence analysis of a transition matrix is to get an insight into the transitions from time 1 to time 2. Different questions about these transitions exist, and these lead to different form of correspondence analysis.

We index the levels of the row variable (time 1) with i, $(i = 1, \ldots, I)$ and the levels of the column variable (time 2) with j, $(j = 1, \ldots, J)$. We denote relative frequencies by p_{ij}, probabilities by π_{ij}, and estimates of probabilities by $\hat{\pi}_{ij}$. Marginal elements are found by replacing the index by "+", for example, row marginal elements of the matrix with relative frequencies are p_{i+} and column marginal elements are p_{+j}.

A first analysis would be a standard correspondence analysis of the contingency table with elements p_{ij}. The interpretation discussed above shows that the resulting graphic display can be interpreted as showing a decomposition of the residuals from the independence model, that is, $\hat{\pi}_{ij} = p_{i+}p_{+j}$ [7–9].

A problem with this standard analysis is that often interest goes out to the off-diagonal elements (i.e. the cells for which $i \neq j$) in the contingency table, as these represent the individuals that change. In a standard correspondence analysis, the view on these cells might be blurred by the diagonal cells, especially, when $p_{ij} \gg p_{i+}p_{+j}$ (which is the case when many individuals remain in the same level of the categorical variable from time point 1 to 2). A solution to this problem is not to study the residuals from the independence model, but from the so-called quasi-independence model, defined here as $\pi_{ii} = p_{ii}$ for $i = j$ and $\pi_{ii} = \alpha_i \beta_j$ for $i \neq j$ [1]. It is possible to adjust correspondence analysis so that residuals from quasi-independence are decomposed. This can be done in two ways: by adjusting the computer program or by changing the input data. The last option seems most simple, and the way to do it is as follows: the diagonal elements p_{ii} have to be replaced by elements for which independence holds. This can be accomplished by filling in elements $p_{i+}p_{+i}$ for the diagonal. By doing this, the margins of the new table have changed so that the elements on the diagonal are not independent, and therefore, using the new margins, again elements $p_{i+}p_{+i}$ have to be filled in. After a few iterations, these elements have stabilized, and a correspondence analysis of the resulting table can be interpreted as a decomposition of quasi-independence [7, 8, 11].

This approach can be extended further by adjusting correspondence analysis so that it can decompose residuals from the symmetry model or from the quasi-symmetry model [7, 8]. Another development

is to use statistical models instead of the exploratory approach described here. There are also close connections between correspondence analysis and **latent class analysis** [12].

We give a small example to illustrate an analysis of the departure from independence. Space limitations withhold us from a detailed interpretation, and for interpretation principles, we refer to **correspondence analysis**. The data are 5 import car types out of 16 car types published in [8]: subcompacts (subi), small specialties (smai), compacts (comi), midsize (midi), and luxury (luxi). In the rows of Table 1, we find the cars disposed of, and in the columns the new cars. Notice the dominant observed frequencies on the diagonal. These values dominate the first dimensions of a correspondence analysis (see Figure 1), especially, the diagonal luxi-cell compared with the rest. In a second analysis, we decompose the residuals from quasi-independence. Such an analysis can be accomplished by filling in "independent" values for the diagonal. These values are 12 790, 1381, 1033, 503 and 71. The interpretation of this correspondence analysis uses the same principles as for standard correspondence analysis of the table with the adjusted margins. For the margins, the residuals are zero, and therefore, the graph only shows car type changes. The car order for cars disposed off is luxi, midi, comi, subi, and smai, but for new cars it is luxi, midi, smai, comi, and subi (see Figure 2). Notice, for example, the different position of smai. It is due to asymmetries in the data that become visible now that the dominance of the diagonal elements has been suppressed. For example, when people dispose of a smai, they buy a luxi very often (relative to the margins of the adjusted table, i.e. observed 459 but predicted by margins 239) but the reverse does not hold (observed 341 but predicted by margins 413).

Table 1 1979 car changing data

	subi	smai	comi	midi	luxi	Total
subi	25 986	5400	2257	1307	288	35 238
smai	3622	5249	738	1070	459	11 138
comi	6981	1023	1536	1005	127	10 672
midi	2844	772	565	3059	595	7835
luxi	997	341	176	589	3124	5227
Total	40 430	12 785	5272	7030	4593	70 110

Rows denote cars disposed, columns denote new cars. Abbreviations are in the text.

More than Two Time Points

When there is one categorical variable measured at more than two time points, it is usual to code the response profiles into a so-called superindicator matrix (*see* **Correspondence Analysis**). Correspondence analysis of a superindicator matrix is also known as multiple correspondence analysis. A superindicator matrix has N individuals in the rows and the categories for each of the time points in the columns. This correspondence analysis has the aim to

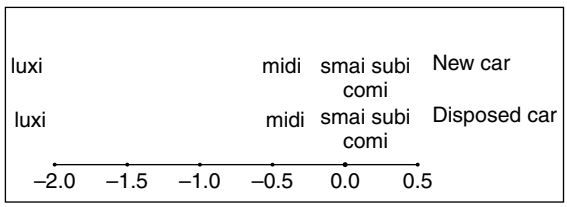

Figure 1 Ordinary CA of car changing data

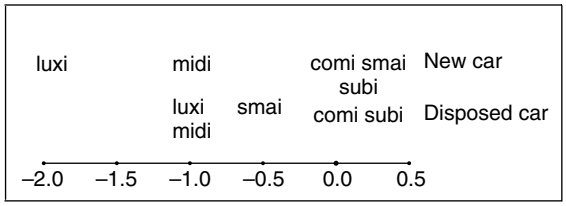

Figure 2 Generalized CA decomposing residuals from quasi-symmetry

get insight into the transitions between all time points simultaneously. The analysis also yields quantifications for the individuals, and the quantifications for an individual can be considered as summaries of the response profile of this individual that can be used, but it can also be used to obtain a classification of the response profiles of the individuals [2–7, 10].

As an example, we give a superindicator matrix of one dichotomous variable measured at three time points for $N = 101$ individuals (see Table 2). (In many computer programmes, the column vector with frequencies cannot be specified, but instead a matrix with 101 rows will serve as the data input file.) The matrix can be made larger in a straightforward way when the number of categories is larger than two, when there are more time points, or when there are more individuals. A correspondence analysis of this matrix will yield a three-dimensional display with 101 points, one for each individual, and a graphical display with 8 points, one for each category at each time point. Without going into technical details (see [5, 10]), individuals with similar profiles will be close together, categories that are often used by the same individuals will be close together, and, when we overlay the two graphs, individuals will be close to the categories that they use. It is also important to notice that, since correspondence analysis displays the departure from the row and from the column margin of a table, it follows that correspondence analysis will *not* show the trend in "a" and "b" over the three time points. This trend can be studied from the counts in the 3×2 table of time points by categories [7, 10].

Another way to interpret this analysis is when we realize that a correspondence analysis of the superindicator matrix, say $\boldsymbol{G}$, is mathematically related to a correspondence analysis of the so-called Burt matrix $\boldsymbol{G'G}$ (*see* **Correspondence Analysis**). The Burt matrix for this example is shown in Table 3. This matrix is a concatenation of a two-way contingency table for each pair of time points, and diagonal matrices with marginal frequencies. This shows that the solution of correspondence analysis only uses two-way **interactions**, and ignores higher-way interactions. Thus, a Burt matrix contains sufficient information for a nonstationary **Markov chain** (the table of time points 1 and 3 is the matrix product of the tables of time points 1 and 2, and 2 and 3) [7, 10].

Examples of such analyses can be found in [2–4, 7, 10]. If the number of individuals is not very large, the estimates for the category points will be unstable. More stability is obtained by constraining category points of adjacent time points to be the same. Such a solution can be obtained by adding up the indicator matrices of the adjacent time points [6]. This is also the way to go when the data to be analyzed are **event history** data, where the observations are in continuous time, or career data. Examples of unconstrained and constrained analyses are in [3–7, 10].

Table 2 A small example of a categorical data matrix (panel A) and its superindicator matrix (panel B)

Panel A		Panel B		
t		t_1	t_2	t_3
1 2 3	Freq	a b	a b	a b
a a a	40	1 0	1 0	1 0
a a b	16	1 0	1 0	0 1
a b a	4	1 0	0 1	1 0
a b b	12	1 0	0 1	0 1
b a a	8	0 1	1 0	1 0
b a b	3	0 1	1 0	0 1
b b a	6	0 1	0 1	1 0
b b b	12	0 1	0 1	0 1

Table 3 The Burt matrix for the example in Table 2

		t_1		t_2		t_3	
		a	b	a	b	a	b
t_1	a	72	0	56	16	44	28
	b	0	29	11	18	14	15
t_2	a	56	11	67	0	48	19
	b	16	18	0	34	10	24
t_3	a	44	14	48	10	58	0
	b	28	15	19	24	0	43

References

[1] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete multivariate analysis*. M.I.T.Press, Cambridge.

[2] de Leeuw, J., van der Heijden, P.G.M. & Kreft, I. (1985). Homogeneity analysis of event history data, *Methods of Operations Research* **50**, 299–316.

[3] Deville, J.-C. & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series, *Journal of econometrics* **22**, 169–189.

[4] Martens, B. (1994). Analyzing event history data by cluster analysis and multiple correspondence analysis: an example using data about work and occupations of scientists and engineers, in *Correspondence analysis in*

the social sciences, M. Greenacre & J. Blasius, eds. Academic Press, London, 233–251.

[5] Saporta, G. (1985). Data analysis for numerical and categorical individual time-series, *Applied stochastic models and data analysis* **1**, 109–119.

[6] van Buuren, S. & de Leeuw, J. (1992). Equality constraints in multiple correspondence analysis, *Multivariate behavioral research* **27**, 567–583.

[7] van der Heijden, P.G.M. (1987). *Correspondence analysis of longitudinal categorical data*. D.S.W.O.-Press, Leiden.

[8] van der Heijden, P.G.M., de Falguerolles, A. & de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and loglinear analysis, *Applied Statistics* **38**, 249–292.

[9] van der Heijden, P.G.M. & de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis, *Psychometrika* **50**, 429–447.

[10] van der Heijden, P.G.M. & de Leeuw, J. (1989). Correspondence analysis, with special attention to the analysis of panel data and event history data, in *Sociological Methodology 1989*, C.C. Clogg, ed. Basil Blackwell, Oxford, 43–87.

[11] van der Heijden, P.G.M., de Vries, H. & van Hooff, J.A.R.A.M. (1990). Correspondence analysis of transition matrices, with special attention to missing entries and asymmetry, *Animal Behaviour* **40**, 49–64.

[12] van der Heijden, P.G.M., Gilula, Z. & van der Ark, L.A. (1999). An extended study into the relationships between correspondence analysis and latent class analysis, in *Sociological Methodology 1999*, M. Sobel & M. Becker eds. Blackwell, Cambridge, pp 147–186.

PETER G.M. VAN DER HEIJDEN

Correspondence Analysis

Correspondence analysis facilitates the exploration and display of interrelations among two or more sets of variables. Historically, it has been identified by a variety of labels including canonical analysis and dual or optimal scaling [9]. A core operation in correspondence analysis and other metric scaling procedures is the decomposition of a matrix of data to find the underlying characteristic vectors and roots [3, 6, 8, 9, 14] (*see* **Eigenvector**; **Eigenvalue**). This mathematical procedure is equivalent to the method of reciprocal averages, **analysis of variance, principal components analysis**, and generalized canonical analysis [11]. **Transformations** of the data before and of the component vectors after decomposition, however, allows for the scaling of both row and column variables in the same spatial configuration. This latter feature, the joint scaling of both row and column variables in the same space, differentiates correspondence analysis from multivariate analyses which focus on either a row or a column variable representation. Thus, correspondence analysis provides information on the interrelationships among variables *within* a set (among the column or among the row variables, as does principal components analysis) and on the interrelationship *between* the row and column variables. Furthermore, correspondence analysis can be used on either qualitative (**categorical**) or quantitative (continuous) multivariate data. A data matrix, $\mathbf{X}$, may contain test scores for individuals (where the entry x_{ij} indicates subject i's score on test j, as is appropriate for a principal components analysis); it may be a cross-classification of cities by types of crimes (where entry x_{ij} indicates city i's frequency count for crime j, as is appropriate for a **contingency table** analysis with a **chi-square test**); or the matrix may contain proximity data (where the entry x_{ij} indicates the distance or similarity between items i and j, as is appropriate for **multidimensional scaling** (*see* **Similarity, Dissimilarity, and Distance Measure**)). Correspondence analysis can be used descriptively with a wide variety of data types, but the use of **inference** requires **random sampling** and frequency count data.

Because correspondence analysis is appropriate for contingency table data, it may be used to visualize and help interpret complex interactions as detected in a **loglinear** analysis [12]. By removing the effects of unequal marginal totals from the data, correspondence analysis is able to provide a detailed description of the **interaction** or **association** among variables or categories. Spatial plots of variables on the interactive factors aid in the interpretation of complex data. They can be used to answer questions about the nature of the association: whether the association is constant across categories and if categories are ordered and equally spaced [7] (*see* **Ordered Categorical Data**). Correspondence analysis also can be used to determine which categories should be combined [4, 5].

Predecomposition Transformation of Data

Correspondence analysis consists of three steps: an initial transformation of the raw data, a singular value decomposition of the transformed data, and a rescaling of the resultant eigenvectors. While metric scaling methods share a core decomposition procedure, they vary in terms of predecomposition transformations of the data and postdecomposition transformation of the eigenvectors. For example, a principal components analysis on a **correlation** matrix of column variables parallels a singular-value decomposition of a column-standardized matrix [14]. In correspondence analysis, the data are first transformed by dividing each x_{ij} entry by the square root of the product of the corresponding row and column marginal totals ($x_{i.}$ and $x_{.j}$, respectively):

$$h_{ij} = \frac{x_{ij}}{\left(\sum_j x_{ij} \sum_i x_{ij}\right)^{1/2}} = \frac{x_{ij}}{(x_{i.}x_{.j})^{1/2}}, \qquad (1)$$

where the matrix $\mathbf{H}$ contains the transformed data. In matrix notation, $\mathbf{H} = \mathbf{S}^{-1/2}\mathbf{X}\mathbf{C}^{-1/2}$, where $\mathbf{S}^{-1/2}$ and $\mathbf{C}^{-1/2}$ are diagonal matrices containing the reciprocal of the square root of the row and column marginal totals. This transformation removes the "magnitude" effects due to differences in marginal totals, so that the pattern of "interaction" may be examined in detail. In the analysis of cross-classified data this is equivalent to removing the Pearson chi-square expected values for the independence model. In fact, an alternative transformation parallels the calculation of each cell's contribution to the total chi-square $[(o_{ij} - e_{ij})^2/e_{ij}]$, and is each cell's contribution to

the overall association:

$$t_{ij} = \left[\frac{(o_{ij} - e_{ij})}{(e_{ij})^{1/2}}\right]\left[\frac{1}{n^{1/2}}\right], \qquad (2)$$

where $\mathbf{T}$ contains the transformed data, o_{ij} indicates the observed frequencies (x_{ij}), and e_{ij} indicates the chi-square expected values ($x_{i.}x_{.j}/x_{..}$). Either (1) or (2) may be used to transform the data without affecting the results, except as noted below.

Decomposition to Basic Structure

In the second step, a singular-value decomposition is used to find the underlying or characteristic vectors in the data. A matrix $\mathbf{X}$ with n rows and m columns, where $n \geq m$, may be represented as the product of three matrices:

$$\mathbf{X}_{(n\times m)} = \mathbf{U}_{(n\times m)}\mathbf{d}_{(m\times m)}\mathbf{V}^{\mathrm{T}}_{(m\times m)}. \qquad (3)$$

Each of the three new matrices contains information regarding the data in $\mathbf{X}$. The $\mathbf{U}$ matrix summarizes the rows, e.g. the row variables, of $\mathbf{X}$: the rows in $\mathbf{U}$ correspond to the rows in $\mathbf{X}$ and the columns in $\mathbf{U}$ are the underlying or characteristic vectors of the row variables. The $\mathbf{V}$ matrix similarly summarizes the information in the columns of $\mathbf{X}$: each row in $\mathbf{V}$ corresponds to a column in $\mathbf{X}$, and the columns of $\mathbf{V}$ are the characteristic vectors of those variables. The columns of $\mathbf{U}$ and $\mathbf{V}$ are all of unit length, so that $(\sum_i u_{ik}^2)^{1/2} = (\sum_j v_{jk}^2)^{1/2} = 1.0$. Also, when the $\mathbf{X}$ matrix is square and symmetric, the $\mathbf{U}$ and $\mathbf{V}$ matrices are equal. In this special case, $\mathbf{X} = \mathbf{U}\mathbf{d}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{d}\mathbf{U}^{\mathrm{T}} = \mathbf{V}\mathbf{d}\mathbf{V}^{\mathrm{T}}$. The $\mathbf{d}$ matrix is a square diagonal matrix with entries along the main diagonal and zeros elsewhere. The main diagonal cell entries are the singular values or roots corresponding to the columns of $\mathbf{U}$ and $\mathbf{V}$, so that the entry d_{jj} corresponds to the jth column of $\mathbf{U}$ and the jth column of $\mathbf{V}$. The singular values are ordered from largest to smallest, and so the dimensions in $\mathbf{U}$ and $\mathbf{V}$ are ordered in terms of their relative importance in accounting for the variance or shape of the configuration of data points.

Spatially, the initial configuration of data points is stretched and/or shrunk and rotated onto new, Euclidean, coordinate axes *without any loss of information*. The new axes (the columns of $\mathbf{U}$ and $\mathbf{V}$) are orthonormal (*see* **Orthogonality**); that is, they are orthogonal or perpendicular to one another and are normalized to unit length. The $\mathbf{U}$ and $\mathbf{V}$ matrices represent the data in normalized space, with each dimension weighted equally. The normalization changes a (Rugby) football-shaped cluster of points into a round spherical shape. The $\mathbf{d}$ matrix entries indicate the relative importance of each dimension and may be used to stretch or shrink dimensions of the configuration to return it to its original shape.

If the data do not contain redundant information, the number of columns in $\mathbf{U}$ and $\mathbf{V}$ will equal the minimum dimension m of $\mathbf{X}$. If the data contain mathematically redundant information, the dimensionality of the solution will be less than m. The dimensionality of the subspace spanned by the configuration of points is equal to the number of nonzero elements in $\mathbf{d}$ (the rank of the matrix). Models of reduced dimensionality may be used to represent the data by using fewer dimensions of the $\mathbf{U}$, $\mathbf{d}$, and $\mathbf{V}$ matrices. For example, a k-dimensional estimate of the data can be obtained by multiplying together the first k dimensions of the $\mathbf{U}$ and $\mathbf{V}$ matrices and the first k weights in $\mathbf{d}$. Note that in correspondence analysis the matrix $\mathbf{H}$ of (1) is factored and not the observed data, $\mathbf{X}$. Because of this, the maximum dimension of the solution is $m - 1$ and multiplication of the $\mathbf{U}$, $\mathbf{d}$, and $\mathbf{V}$ matrices results in the matrix $\mathbf{H}$. Correspondence analysis contains procedures for retrieving models of the original, observed data ($\mathbf{X}$).

One effect of the transformation in (1) is that the first column in the $\mathbf{U}$ and $\mathbf{V}$ matrices, corresponding to the independence model, contains a constant equal to 1.0. This factor is ignored and is sometimes referred to as the "trivial" factor. The first singular value, d_{11}, is also equal to 1.0 and is ignored for most purposes. Subsequent singular values in $\mathbf{d}$ are called **canonical correlations**. When (2) is used to transform the data, the trivial factor and the first singular value are not evident in the solution.

The $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{d}$ matrices may be found directly with a singular-value decomposition of $\mathbf{X}$ or indirectly by performing eigenanalyses on the cross-product matrices of $\mathbf{X}$. Pre- or postmultiplication of a matrix by its transpose results in a square, symmetric matrix, so that the characteristic vectors of a rectangular (nonsquare, nonsymmetric) matrix may be found by performing eigenanalyses on $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}} = \mathbf{U}\mathbf{d}^2\mathbf{U}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\mathbf{d}^2\mathbf{V}^{\mathrm{T}}$. The singular values are equal to the square root of the eigenvalues, so that

$\mathbf{X} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{d}\mathbf{V}^{\mathrm{T}}$. Decomposition of $\mathbf{X}, \mathbf{X}\mathbf{X}^{\mathrm{T}}$, or $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ results in solutions of the same rank. While the characteristic vectors and singular values can be derived from either an eigenanalysis or a singular-value decomposition, the singular-value decomposition algorithm offers greater numerical stability, especially with ill-conditioned matrices. The solution is a **least squares** solution with results typically providing a close approximation to the **maximum likelihood** solution [5, 13].

Transformation of Component Vectors to Obtain Optimal Scores

In the last step, the columns of the **U** and **V** matrices are rescaled. The rescaled characteristic vectors are referred to as optimal scores, canonical scores, or canonical variates. Rescaling is performed as follows:

$$u_{ik}{}^{*} = u_{ik}\left(\frac{x_{..}}{x_{i.}}\right)^{1/2} \tag{4}$$

and

$$v_{jk}{}^{*} = v_{jk}\left(\frac{x_{..}}{x_{.j}}\right)^{1/2}, \tag{5}$$

where k indexes the component vectors (ignoring the first trivial vector if (1) was used) and matrices $\mathbf{U}^{*}$ and $\mathbf{V}^{*}$ contain the optimal scores for **U** and **V**, respectively. The optimal scores are normalized, Euclidean spatial coordinates. Multiplication of the scores by the canonical correlations stretches or shrinks the configuration so that the relative importance of the factors is evident. For comparing scores between rows and columns and between dimensions, the scores must be weighted relative to the singular values. A common practice is to multiply the scores by the square root of the corresponding singular value, although other methods are available [2].

Modeling Observed Data

The number of interactive factors ($\leq m - 1$) needed to represent the data can be determined by sequentially creating models of increased dimensionality and testing the **goodness of fit** of each model to the observed data. With cross-classified data, this should be preceded by a chi-square test for independence. A significant chi-square indicates that the observed data are significantly different from the model of independence and that an analysis of the interaction may then be pursued. Models of successively higher dimensionality are constructed with the interactive factors until a model is obtained that is not significantly different from the observed data. Chi-square tests are used to test the goodness of fit between the models and observed data. The expected values (e_{ij}) for a k-dimensional model are calculated as

$$e_{ij} = \sum_{k}\left[\left(\frac{x_{i.}x_{.j}}{x_{..}}\right)d_{kk}u_{ik}{}^{*}v_{jk}{}^{*}\right], \tag{6}$$

where k is the number of dimensions (beyond the trivial factor) to be used in reconstructing the data, $\mathbf{U}^{*}$ and $\mathbf{V}^{*}$ are the optimal scores, and **d** contains the singular values or canonical correlations. Thus, data are reconstructed from the sum of successive models: the chi-square independence model (the trivial factor and the first singular value) and then the first nontrivial interactive factor and its canonical correlation, and so on. There may be as many factors as the minimum dimension of the matrix **X** minus one ($\leq m - 1$ pairs of factors).

With cross-classified data, goodness of fit is tested directly with the Pearson chi-square. With other types of data, goodness of fit can be expressed descriptively as a function of the singular values. Ratios of squared canonical correlations may be used in a correspondence analysis to describe the "proportion of explained variance" in the same way that eigenvalues are used in a principal components analysis. The ratio of the sum of the first k squared canonical correlations to the sum of all squared canonical correlations provides a descriptive index of explained variance. In correspondence analysis this is the proportion of the total association captured by a k-factor representation and in some applications is called "inertia" [1, 8]. In fact, the sum of the squared canonical correlations is equal to the total degree of association in the data (the Pearson chi-square score divided by the sample size). Since applications of correspondence analysis to nonfrequency data cannot use a Pearson chi-square to test how much data deviate from marginal totals, the ratio of the first (trivial) singular value to the sum of all squared singular values (including the trivial one) may be used to estimate the proportion of total variance in the original data that is explained by magnitude differences in marginal totals [14].

Alternative Data Formats: Indicator Variables

Qualitative data also may be analyzed as indicator variables (*see* **Dummy Variables**). If data are expressed as dichotomous variables (1 = characteristic present and 0 = absent), with a dichotomous indicator variable for *each* category of each variable, and if the cases with identical profiles are summed together, then results are parallel to those obtained from the cross-classification table of the same data [10, 14]. Analysis of indicator variables has the interesting advantage of scaling not only the categories of each variable but also the types of cases. The unique case profiles become an additional set of variables in the analysis, and thus the relationship between the type of case and the categories of each variable may be seen.

Ordination, Seriation, and Guttman Scaling

Correspondence analysis can also be used to find the optimal ordering of variables for a given set of characteristics. This problem has different names in different fields of study: ordination, seriation, and **Guttman scaling**. Because items that occur closer in time (or space) have more similar profiles than items further apart, rearrangement of the data by similarity in profiles helps to establish their optimal ordering. Guttman scaling is a specific type of ordering, wherein the order is cumulative and transitive: if someone has an object on the list, then they tend to have the objects that precede it; and if they lack an item on the list, then they tend not to have subsequent items. Correspondence analysis provides a weighted least squares solution to this problem. Analysis of a matrix of cross-classified data (location by type of archeological specimen) or of indicator variables (households by the presence or absence of specific consumer goods) provides optimal scores for row and column variables that establish the optimal order of variables. The scores may be used to reorder the original data matrix to see the pattern or they may be plotted spatially. The plot of the scores will often yield a curvilinear rather than a strictly linear result.

Multiple Correspondence Analysis

Multiple correspondence analysis involves three or more variables or sets of variables. The analysis is performed on indicator matrices or stacked contingency tables. Continuous variables can be recoded into categories for the analysis with a variable for each category. Indicator matrices, where the matrix rows represent cases (or unique case profiles) and the columns represent *all* categories of all variables, can be used. Guttman scaling is an example of a multiple correspondence analysis on indicator variables. Similarly, a cross-product or Burt [8] matrix ($\mathbf{X}^{\mathrm{T}}\mathbf{X}$) of the indicator variables can be analyzed to obtain the solution. In a stacked table analysis, a k-way contingency table is arrayed as a series of two-way contingency tables. The tables are analyzed as a single two-way table with the tables stacked side by side vertically or horizontally. In an analysis of age, gender, and methods of suicide the different ways of arraying the two-way tables emphasizes different aspects of the data: gender and age patterns for suicide methods; or gender differences in choice of a suicide method for different age groups [12].

References

[1] Benzecri, J.P. (1969). Statistical analysis as a tool to make patterns emerge from data, in *Methodologies of Pattern Recognition*, S. Watanabe, ed. Academic Press, New York, pp. 35–74.

[2] Carroll, J.D., Green, P.E. & Schaffer, C.M. (1986). Interpoint distance comparisons in correspondence analysis, *Journal of Marketing Research* **23**, 271–280.

[3] Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, New York.

[4] Gilula, Z. (1986). Grouping and association in contingency tables: an exploratory canonical correlation approach, *Journal of the American Statistical Association* **81**, 773–779.

[5] Gilula, Z. & Haberman, S.J. (1986). Canonical analysis of contingency tables by maximum likelihood, *Journal of the American Statistical Association* **81**, 780–788.

[6] Gittins, R. (1985). *Canonical Analysis: A Review with Applications in Ecology*. Springer-Verlag, Berlin.

[7] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association* **74**, 537–552.

[8] Greenacre, J.M. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, New York.

[9] Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*. University of Toronto Press, Toronto.

[10] Nishisato, S. & Sheu, W. (1980). Piecewise method of reciprocal averages for dual scaling of multiple-choice data, *Psychometrika* **45**, 467–478.
[11] Tenehaus, M. & Young, F.W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data, *Psychometrika* **50**, 91–119.
[12] Van der Heijden, P.G.M. & De Leeuw, J. (1985). Correspondence analysis used complimentary to loglinear analysis, *Psychometrika* **50**, 429–447.
[13] Wasserman, S. & Faust, K. (1989). Canonical analysis of the composition and structure of social networks, in *Sociological Methodology*, C.C. Clogg, ed. Basil Blackwell, Cambridge, Mass, pp. 1–42.
[14] Weller, S.C. & Romney, A.K. (1990). *Metric Scaling: Correspondence Analysis*. Sage, Newbury Park.

(*See also* **Inference, Foundations of**; **Matrix Algebra**; **Multivariate Analysis, Overview**)

SUSAN C. WELLER

Cosine of Angle Between Two Vectors

Two vectors in 2-space are shown in Figure 1, with (x_1, x_2) being a point on one vector and (y_1, y_2) on the other. Let A_x and A_y be the angles the vectors make with the horizontal axis, and define B as the angle between the vectors. Thus

$$\begin{aligned} B &= A_x - A_y, \\ \cos B &= \cos(A_x - A_y) \\ &= \cos A_x \cos A_y + \sin A_x \sin A_y. \end{aligned}$$

On dropping perpendiculars from the points to the horizontal axis, it is then easily seen from right-angle triangle geometry that

$$\begin{aligned} \cos B &= \frac{x_1}{(x_1^2 + x_2^2)^{1/2}} \frac{y_1}{(y_1^2 + y_2^2)^{1/2}} \\ &\quad + \frac{x_2}{(x_1^2 + x_2^2)^{1/2}} \frac{y_2}{(y_1^2 + y_2^2)^{1/2}} \qquad (1) \\ &= \frac{x_1 y_1 + x_2 y_2}{d_x d_y} \qquad (2) \end{aligned}$$

for

$$d_x^2 = x_1^2 + x_2^2 \quad \text{and} \quad d_y^2 = y_1^2 + y_2^2. \qquad (3)$$

Thus (2) is the formula for the cosine of the angle between two vectors in 2-space.

Another derivation of (2) with d_x and d_y being the distances from the origin to the points (x_1, x_2) and (y_1, y_2), respectively, is to apply the cosine rule to the angle B in the triangle formed by the origin and the points (x_1, x_2) and (y_1, y_2). This gives

$$\begin{aligned} &\cos B \\ &= \frac{d_x^2 + d_y^2 - [\text{the distance from } (x_1, x_2) \text{ to } (y_1, y_2)]^2}{2 d_x d_y} \\ &= \frac{d_x^2 + d_y^2 - [(x_1 - x_2)^2 + (y_1 - y_2)]^2}{2 d_x d_y}, \end{aligned}$$

which with (3) reduces to (2).

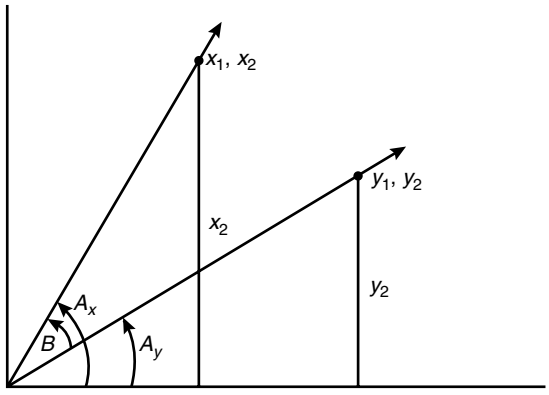

Figure 1 Two vectors in 2-space

Three-Space

For two vectors in 3-space a diagram analogous to Figure 1 can be drawn (see Searle [3]). Applying to that 3-space diagram some triangle geometry more complicated than that used for deriving (1) yields the result

$$\cos B = \frac{x_1 y_1 + x_2 y_2 + x_3 y_3}{d_x d_y}, \qquad (4)$$

with

$$d_x^2 = x_1^2 + x_2^2 + x_3^2 \quad \text{and} \quad d_y^2 = y_1^2 + y_2^2 + y_3^2, \quad (5)$$

where (x_1, x_2, x_3) is a point on one vector and (y_1, y_2, y_3) is on the other. Details are in Searle [3].

n-Space

Let $\mathbf{x}' = [x_1 \ x_2 \ldots x_i \ldots x_n]$ and $\mathbf{y}' = [y_1 \ y_2 \ldots y_i \ldots y_n]$ be two points in n-space, one on one vector and one on another. Then results (2), (3) and (4), extend very directly for n-space to

$$\cos B = \sum_{i=1}^{n} \frac{x_i y_i}{d_x d_y} \qquad (6)$$

for

$$d_x^2 = \sum_{i=1}^{n} x_i^2 \quad \text{and} \quad d_y^2 = \sum_{i=1}^{n} y_i^2 \qquad (7)$$

so that, in terms of the vectors $\mathbf{x}'$ and $\mathbf{y}'$,

$$\cos B = \frac{\mathbf{x}'\mathbf{y}}{(\mathbf{x}'\mathbf{x})^{1/2} (\mathbf{y}'\mathbf{y})^{1/2}}. \qquad (8)$$

Viewed from the geometry of two and three dimensions, (6) may not seem very satisfying. Moreover,

its derivation demands arguments in the geometry of n-space. These arguments are more theoretical than those for deriving (2) and (4) of 2-space and 3-space, respectively. Thus it is easier to simply take (6) as an algebraic definition of B as the angle between two vectors in n-space. Indeed, some books on **multivariate analysis** do just that, e.g. [2, p. 16] and [1, p. 99].

Invariance

The prime property of a vector is its direction, not its length. Yet each of (2), (4) and (6) seems to depend upon the actual values of the xs and the ys, i.e. their lengths. Fortunately this is not so. For example, with (2), if on the vector through (x_1, x_2) some other point (x_1^*, x_2^*) is taken, it will be found from the geometry of congruent triangles that if $x_1^* = \lambda_x x_1$, then $x_2^* = \lambda_x x_2$. Therefore $\cos B$ of (2), with the x^*s and y^*s replacing the xs and ys, becomes

$$\begin{aligned}\cos B &= \frac{x_1^* y_1^* + x_2^* y_2^*}{(x_1^{*2} + x_2^{*2})^{1/2}(y_1^{*2} + y_2^{*2})^{1/2}} \\ &= \frac{\lambda_x \lambda_y (x_1 y_1 + x_2 y_2)}{\lambda_x (x_1^2 + x_2^2)^{1/2} \lambda_y (y_1^2 + y_2^2)^{1/2}} \\ &= \frac{x_1 y_1 + x_2 y_2}{d_x d_y}\end{aligned}$$

as before; i.e. $\cos B$ of (2) is unchanged. Similar geometry also leaves (4) unchanged. And arguing in n-space that changing x_1 to $x_1^* = \lambda_x x_1$ leads to $x_i^* = \lambda_x x_i$ for all i, then (6) will be unchanged also.

A second form of invariance is when rotation of axes is considered. Although coordinates of a point will change under rotation, the angle between two vectors will not; and the cosine of that angle is unchanged and is the same function of the new coordinates as the old. Suppose for 2-space that the axes are rotated counterclockwise through an angle θ. It can then be shown (as in Searle [3]) that the coordinates x_1 and x_2 of Figure 1 become

$$\begin{aligned}x_1' &= x_1 \cos\theta + x_2 \sin\theta \quad \text{and} \\ x_2' &= x_2 \cos\theta - x_1 \sin\theta. \end{aligned} \tag{9}$$

From these it is easily seen that $d_x' = x_1'^2 + x_2'^2 = x_1^2 + x_2^2 = d_x$ (as one would expect). Furthermore, with y_1' and y_2' being the same function of y_1 and y_2 as the x's are of the xs in (9), it is straightforward to show that $x_1' y_1' + x_2' y_2' = x_1 y_1 + x_2 y_2$. Hence, comparable with (2),

$$\cos B = \frac{x_1 y_1 + x_2 y_2}{d_x d_y} = \frac{x_1' y_1' + x_2' y_2'}{d_{x'} d_{y'}}.$$

Thus, rotating the axes gives the cosine of the angle between two vectors being the same expression of the new coordinates as it does of the old ones.

Correlation

When the entries in $\mathbf{x}$ and $\mathbf{y}$ are data (e.g. height and weight of each member of a rowing club), define $\overline{x}$ and $\overline{y}$ as the observed averages:

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} \quad \text{and} \quad \overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}.$$

In $\mathbf{x}$ and $\mathbf{y}$ replace each element x_i by $x_{i0} = x_i - \overline{x}$ and y_i by $y_{i0} = y_i - \overline{y}$. Define $\mathbf{x}_0$ and $\mathbf{y}_0$ as the vectors of elements x_{i0} and y_{i0}. Then $\cos B$ for $\mathbf{x}_0$ and $\mathbf{y}_0$ is

$$\begin{aligned}\cos B &= \frac{\mathbf{x}_0' \mathbf{y}_0}{(\mathbf{x}_0' \mathbf{x}_0)^{1/2} (\mathbf{y}_0' \mathbf{y}_0)^{1/2}} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2\right]^{1/2}}\end{aligned}$$

is the product–moment **correlation** between the two variables.

References

[1] Johnson, R.A. & Wichern, D.W. (1988). *Applied Multivariate Statistical Analysis*, 3rd Ed. Prentice Hall, New York.

[2] Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, New York.

[3] Searle, S.R. (1996). 3-D geometry: a triangle-oriented proof of the cosine of the angle between two vectors, *Technical Report BU-1342-M.* Biometrics Unit, Cornell University.

(*See also* **Matrix Algebra**)

SHAYLE R. SEARLE

Cost–Benefit Analysis, Willingness to Pay

An important distinguishing feature between different methods of health care economic evaluation (*see* **Health Economics**) is the way in which health-related outcomes are defined, measured, and valued [10]. Cost–benefit analysis (CBA) is a technique where health benefits are valued in monetary units for comparison against program costs [16]. CBA holds appeal for at least three reasons: (i) it has a theoretical foundation in welfare economics [22]; (ii) by enabling direct comparison of program costs and benefits it permits the calculation of a program's net benefit and thereby avoids much of the ambiguity associated with cost-effectiveness "league tables" [4, 5, 21]; and (iii) the same principle of net benefit can be applied to other sectors such as transport and environment so that intersectoral comparisons of resource use can be considered.

CBA might become the method of choice for more researchers in health care if there was one unambiguous and generally agreed-upon method for estimating money values for health outcomes. Debate over CBA in health care and other areas of project appraisal requiring monetary valuation of health outcomes has a controversial history. In the 1960s and 1970s the prevalent notion of CBA was restricted simply to comparing health care costs with and without the program being evaluated. Studies such as those by Koplan [18] on pertussis vaccination, although labeled as CBA, are essentially cost comparisons because they make no attempt to place dollar values on health benefits. Following Becker [2], it became popular to quantify health benefits in monetary units, characterized as a return on human capital investment, using discounted future earnings streams (for an example in rubella vaccination, see Schoenbaum [28]).

In a seminal contribution, Mishan [23] argued that the human capital approach in CBA studies was flawed for two general reasons: (i) it builds upon the questionable assumption that society's main goal is the maximization of national income, and this has some worrisome implications for valuing programs that improve the health of persons with low or zero future earnings; and (ii) the method is inconsistent with the foundation of CBA from welfare economic theory. Mishan argued that the monetary valuation of programs that reduced **risks** to "life and limb", as with other goods, should be based upon the concept of consumer surplus; that is, the relevant economic notion of value for a program is the most that consumers are willing to give up to receive it – i.e. their maximum willingness to pay. Such measurement is based on the decision rule for CBA, where the goal is to determine whether the value of the program to those who gain is sufficiently large that they could, in principle, compensate in full all those who lose and still remain better off themselves (the so-called Potential Pareto improvement criterion).

Subsequent empirical work has taken two main directions. First, there are *revealed-preference studies* which document observed trade-offs between health risks and money. The principle here is that the analyst can only assess a person's preferences for trade-offs between money and health by actually observing market behavior. For example, this might be in the form of a person's willingness to pay for extra safety features on a new car. A major focus of revealed preference studies has been labor market studies, relating wage premiums to health risks for particular occupations [20, 31]. Secondly, there are *stated preference surveys* of hypothetical dollar – health risk choices such as road transport safety and consumer decisions [17]. This second approach is not based on observing actual market behavior and money – health trade-offs, but operates by offering hypothetical choices to respondents in a survey. This survey-based approach has been termed *contingent valuation*, because the respondent is being asked to consider the contingency of a market existing for the thing being valued, even though an actual market may not exist.

Much of the conceptual and empirical development of contingent valuation methods has been done in the areas of transport economics with a predominant focus on the value of life [17], and lately in environmental economics with application to the valuation of environmental and health goods such as improved air quality [24] (*see* **Environmental Epidemiology**). Contingent valuation studies are becoming more widespread in health care and have been undertaken in areas such as arthritis management [30], ultrasonography [3], care of the elderly [9], management of hypertension [15], blood transfusion [11], the use of ionic vs. nonionic contrast media

[1], and *in vitro* fertilization [25] (*see* **Clinical Epidemiology**; **Health Services Research, Overview**).

We review some of the statistical issues that have arisen for researchers who seek to measure willingness to pay (WTP). We will argue that choosing among the various measurement techniques presents the analyst with a variety of practical trade offs between **bias** and precision.

Measurement Objectives

There are numerous approaches to assessing WTP in health care, and in this limited discussion of selected statistical issues we cannot review all possible approaches. A conceptual framework and tutorial on WTP in health care has recently been published [26]; it gives a deeper understanding of linkages between the theory and practice of WTP.

To explore statistical issues we will focus on a hypothetical example of a **program evaluation** where WTP in the context of CBA is to be determined. An insurance-based WTP problem is described in Figure 1. Using this example, there are advantages and disadvantages to a number of measurement techniques that are available.

WTP Estimation Techniques

Open-Ended Questions

The measurement task is to find out the *maximum* that an individual (or group) would be willing to pay for the new program. The open-ended question format is the most direct format for determining this value from an individual. But simply asking the maximum a person would pay poses a difficult cognitive task for a respondent who may be unfamiliar with the program being valued and not accustomed to buying similar things in a market without price tags. Consequently, the open-ended format tends to produce large numbers of nonresponses or protest zero responses to WTP questions: this is true for both environmental valuation (e.g. [8]) and health (e.g. [15] and [30]). Such experience led researchers to try to simplify the choices presented to the respondents by making the market scenario more realistic.

Bidding Games

Mitchell & Carson [24] report that the oldest and most widely used elicitation method in contingent valuation surveys until recently has been the *bidding game*. Similar to an auction, an initial starting money value is bid up or down by the respondent. Some market realism is instilled because each bid level requires only a Yes/No response. It has been argued that the advantage of the bidding game is that the process of iteration and search enables the respondent to consider more fully the value of the program [14]. A major disadvantage is the potential for **bias** because the starting bid (chosen by the researcher) tends to imply a value for the good. In a health care contingent valuation study, O'Brien & Viramontes [27] tested the hypotheses of starting point bias using random assignment to different starting bias (*see* **Randomization**) in a bidding game. Although not statistically significant, these data suggest that higher starting bids were associated with higher final WTP. More recently, Stalhammer [29] has found evidence to support the hypothesis of starting point bias.

Suppose we wish to conduct a CBA of a new treatment program that is proposed to be added to the benefits covered in a Health Maintenance Organization (HMO) in the United States. Based on utilization projections and the costs of the new treatment we can estimate the total additional cost of the new program for the next year. To assess the dollar value of the new program, we want to estimate the total willingness to pay (WTP) among enrollees of the HMO. The proposal is to contact a random sample of enrollees, inform them of the benefits of covering the new program as an additional insured benefit, and then elicit the maximum they would be willing to pay – in additional monthly insurance premiums – to have the new treatment covered. These sample responses would then be projected to the total HMO population so that total benefits (i.e. WTP) could be compared with total cost.

Figure 1 Hypothetical willingness-to-pay problem

Payment Cards

The payment card method was developed by Mitchell & Carson [24] as an alternative to the bidding game. The payment card is a visual aid which contains a large array of potential WTP amounts ranging from \$0 to some large amount. According to Mitchell & Carson, this method "... circumvents the need to provide a single starting point, yet offers the respondent more of a context for her bid than the direct question method provides" [24]. A related technique is the checklist method, where respondents indicate which of a list of payment *ranges* includes their WTP amount. Neumann & Johannesson [25] used a form of payment card in their study of *in vitro* fertilization; they found evidence that the range of values listed influenced the final WTP.

Dichotomous Choice (Take it or Leave it)

Open-ended questions, bidding games, and payment cards are all approaches that have been used as methods for finding maximum WTP *for each sampled individual*. In contrast to this within-person search strategy, many researchers in environmental economics have moved to a strategy of a between-person search to find maximum WTP *for a sample of respondents*.

Using this approach, Bishop & Heberlein [6] developed another elicitation method known as the "take-it-or-leave-it" approach. This method uses a large number of predetermined prices and each respondent is asked if she is willing to pay a single one of these prices (Yes or No) for the program, with no further iteration. The prices are randomly assigned to respondents so that one can use statistical techniques such as probit analysis (*see* **Quantal Response Models**) to model bid acceptance as a function of respondent characteristics and determine **median** WTP [7]. The main problem is that, relative to other elicitation methods, the take-it-or-leave-it method is inefficient, requiring a much larger sample size for the same level of statistical precision as other methods. However, the method has been used with some success in health care [15].

To illustrate the basic approach of the dichotomous choice method, consider the bid curve in Figure 2. The goal of sampling is to be able to plot the bid curve which models the probability of accepting a bid (price) at different levels of bid. From the derived bid curve one can determine either the median WTP (i.e. bid at which 50% of the sample would pay) or the mean WTP, this being the area under the bid curve. Given the variation among respondents in demographic characteristics, the analyst would typically use **logistic (or probit) regression** to model the accept/reject probability as a function of bid level and demographics. Using this function it is possible to project the total WTP for a population if one knows its demographic characteristics.

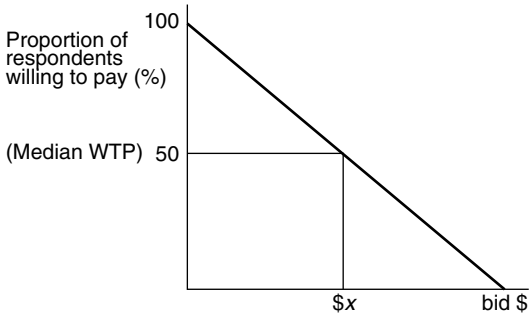

Figure 2 The population of respondents willing to pay as a function of the bid in a WTP study. Note that median WTP is \$$x$ and mean WTP is the area below the curve. The method is analogous to survival analysis and the calculation of life expectancy in biostatistics. For simplicity we have drawn the bid curve as a straight line, but in practice this observed "survival curve" will not be a linear, nor necessarily smooth, relationship

The main disadvantage of the simple dichotomous choice approach is that each individual is asked one question, i.e. "would you pay an additional insurance premium of \$2 to have this program covered?" Hence, to derive the bid curve one may need quite large sample sizes. Also note that if the upper end of the bid range is not sufficiently high, then the bid curve will not reach zero, making it problematic to estimate the area under the curve for the expected value.

Double-Bounded Dichotomous Choice

An extension to the previous method, to improve its statistical efficiency, is that of *take it or leave it with follow-up*. Here a follow-up bid question is asked of the respondent, higher or lower, conditional upon the response to the first bid; the higher or lower bid is randomly selected from a range. The increased statistical efficiency of this method using probit analysis

The probability of a positive response to the close-ended WTP question is the probability that the respondent's WTP exceeds the bid amount. The double-bounded, dichotomous choice model as derived by Hanemann et al.. [13] is described below. When confronted with a bid or dollar amount t, the respondent will accept the bid with a probability

$$\Pr(\text{Yes}|t; \mathbf{X}, \boldsymbol{\beta}) = 1 - F(t; \mathbf{X}, -\boldsymbol{\beta}),$$

where $F[\cdot]$ is the cumulative distribution function, t is the bid (or dollar) amount, $\mathbf{X}$ is a vector of socioeconomic characteristics or other demand shift variables, and $\boldsymbol{\beta}$ is a vector of parameters to be estimated.
Substituting the cumulative normal distribution function, $\Phi[\cdot]$, for $F[\cdot]$, we get

$$F(t; \mathbf{X}, -\boldsymbol{\beta}) = \begin{cases} \Phi(t_U; \mathbf{X}, -\boldsymbol{\beta}), & \text{if Yes–Yes response,} \\ \Phi(t_L; \mathbf{X}, -\boldsymbol{\beta}) - \Phi(t_U; \mathbf{X}, -\boldsymbol{\beta}), & \text{if Yes–No or No–Yes response,} \\ 1 - \Phi(t_L; \mathbf{X}, -\boldsymbol{\beta}), & \text{if No–No response,} \end{cases}$$

where t_U is the higher bid amount, t_L is the lower bid amount, and $\boldsymbol{\beta}$ is a vector of coefficients to be estimated.
The probability distribution of WTP is simply the derivative of Pr(Yes) or the change in Pr(Yes) at various bid levels. The mean WTP is the area under the Pr(Yes) function, whereas the median WTP is the bid value for which the estimated probability of answering "Yes", [Pr(Yes)] equals 0.5. In this article, we report median WTP values.

Figure 3 Double-bounded dichotomous choice WTP model

has been shown by Hanemann [13]. As indicated by the algebraic exposition in Figure 3, the goal of this method is to select ranges so as to "bound" as many respondents as possible between the extremes of their "Yes–Yes" and "No–No" responses. To minimize framing biases one can randomly select both the starting bid and the second conditional bid.

Future Trends: Conjoint Analysis

The most recent approach under evaluation for WTP in environment and health is conjoint analysis. Used widely in consumer economics to establish consumer preferences over attributes of commodities (e.g. a car's safety, fuel consumption, performance, and price), conjoint analysis is very similar in origin to multiattribute utility theory [12]. The method works by first defining a number of attributes of the product or program and then asking the respondent to choose between hypothetical pairs (e.g. program A vs. program B) that vary in their attribute composition and where one of the attributes is how much the individual would have to pay. As reviewed by Green & Krieger [12], there are various adaptive search algorithms that determine the modification of attributes conditional upon individuals' responses. Conjoint analysis has been used successfully in the evaluation of consumer products where health outcomes are attributes [19].

Concluding Remarks

In summary, there are a number of elicitation methods for WTP, and each approach has strengths and weaknesses. There are no obvious conceptual reasons, from economic theory, to predict that these measurement approaches will yield different valuations; each is an approach to revealing an underlying monetary valuation. In practice, for some of the reasons outlined above, estimates do vary when different methods are used in the same respondents (see [15]) on open-ended vs. take-it-or-leave-it methods in hypertension). Each estimation technique varies in terms of measurement properties of precision and bias. A further practical consideration is that formats such as bidding games necessitate in-person interviewing, perhaps even with computer-based bidding algorithms with random starting points. Other formats such as open-ended questions can be done by mail survey but suffer from low completion rates and greater **variance** due to the cognitive burden for respondents.

Currently there is considerable conceptual and statistical variation in the assessment of WTP, both in environmental health and health care program evaluations. Guidelines for WTP studies in environment were issued in 1993 in the US by the National Oceanic and Atmospheric Administration, who propose that WTP studies should use a dichotomous choice format and that values should be elicited by in-person interviews. Whether such guidance can be generalized to health care WTP studies is unclear. The nature of the health care commodity for WTP studies is very different from many environmental program benefits, and methods appropriate for the former may not be suited to the latter. Ongoing experiments using computer-based interviewing with the approach of conjoint analysis hold great promise for the future of WTP in health care.

Acknowledgments

Dr Bernie O'Brien is supported by a Senior Investigator award from the Canadian Institute for Health Research. He is grateful to Stephen Walter for helpful comments on an earlier draft.

References

[1] Appel, L.J., Steinberg, E.P., Powe, N.R., Anderson, G.F., Dwyer, S.A. & Faden, R.R. (1990). Risk reduction from low osmolality contrast media. What do patients think it is worth?, *Medical Care* **28**, 324–334.

[2] Becker, G.S. (1964). *Human Capital*. Columbia University, New York.

[3] Berwick, D.M. & Weinstein, M.C. (1985). What do patients value? Willingness-to-pay for ultrasound in normal pregnancy, *Medical Care* **23**, 881–893.

[4] Birch, S. & Gafni, A. (1992). Cost effectiveness/utility analysis: do current decision rules lead us to where we want to be?, *Journal of Health Economics* **11**, 279–286.

[5] Birch, S. & Gafni, A. (1994). Cost effectiveness ratios in a league of their own, *Health Policy* **28**, 133–141.

[6] Bishop, R.C. & Heberlein, T.A. (1979). Measuring values of extra-market goods: are indirect measures biased?, *American Journal of Agricultural Economics* **61**, 926–930.

[7] Cameron, T.A. & James, M.D. (1987). Efficient estimation methods for "closed-ended" contingent valuation surveys, *Review of Economics and Statistics* **69**, 269–276.

[8] Desvousges, W.H., Smith, V.K. & McGivney, M.P. (1983). *A Comparison of Alternative Approaches for Estimating Recreation and Related Benefits of Water Quality Improvements*. EPA-230-05-83-001. Office of Policy Analysis, US Environmental Protection Agency, Washington.

[9] Donaldson, C. (1990). Willingness to pay for publicly-provided goods: a possible measure of benefit, *Journal of Health Economics* **9**, 103–118.

[10] Drummond, M.F., Stoddard, G.L. & Torrance, G.W. (1987). *Methods for the Economic Evaluation of Health Care Programs*. Oxford University Press, Oxford.

[11] Estaugh, S.R. (1991). Valuation of the benefits of risk-free blood. Willingness to pay for hemoglobin solutions, *International Journal of Technology Assessment in Health Care* **7**, 51–57.

[12] Green, P.E. & Krieger, A.M. (1996). Individualized hybrid models for conjoint analysis, *Management Science* **42**, 850–867.

[13] Hanemann, M., Loomis, J. & Kanninen, B. (1991). Statistical efficiency of double-bounded dichotomous choice contingent valuation, *American Agricultural Economics Association* **73**, 1255–1263.

[14] Hoehn, J.P. & Randall, A. (1987). A satisfactory benefit cost indicator from contingent valuation, *Journal of Environmental Economics and Management* **14**, 226–247.

[15] Johannesson, M. & Jonsson, B. (1991). Willingness to pay for antihypertensive therapy: results of a Swedish pilot study, *Journal of Health Economics* **10**, 461–474.

[16] Johannesson, M. & Jonsson, B. (1991). Economic evaluation in health care: is there a role for cost–benefit analysis?, *Health Policy* **17**, 1–23.

[17] Jones-Lee, M.W. (1976). *The Value of a Life: An Economic Analysis*. University of Chicago Press, Chicago.

[18] Koplan, J.P., Schoenbaum, S.C., Weinstein, M.C. & Fraser, D.W. (1979). Pertussis vaccine – an analysis of benefits, risk and costs, *New England Journal of Medicine* **301**, 906–911.

[19] Magat, W.A., Viscusi, W.K. & Huber, J. (1996). Paired comparison and contingent valuation approaches to morbidity risk valuation, *Journal of Environmental Economics and Management* **15**, 395–411.

[20] Marin, A. & Psacharopoulos, G. (1982). The reward for risk in the labor market: evidence from the United Kingdom and a reconciliation with other studies, *Journal of Political Economy* **90**, 827–853.

[21] Mason, J., Drummond, M.F. & Torrance, G.W. (1993). Some guidelines on the use of cost–effectiveness league tables, *British Medical Journal* **306**, 570–572.

[22] Mishan, E.J. (1971). Evaluation of life and limb: a theoretical approach, *Journal of Political Economy* **79**, 687–705.

[23] Mishan, E.J. (1971). *Cost–Benefit Analysis*. Allen & Unwin, London.

[24] Mitchell, R.C. & Carson, R.T. (1989). In *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Resources for the Future, Washington.

[25] Neumann, P.J. & Johannesson, M. (1994). The willingness to pay for in vitro fertilization: a pilot study using contingent valuation, *Medical Care* **32**, 686–699.

[26] O'Brien, B.J. & Gafni, A. (1996). When do the "dollars" make sense? Toward a conceptual framework for contingent valuation studies in health care, *Medical Decision Making* **16**, 288–299.

[27] O'Brien, B. & Viramontes, J.L. (1994). Willingness to pay: a valid and reliable measure of health state preference?, *Medical Decision Making* **14**, 289–297.

[28] Schoenbaum, S.C., Hyde, J.N., Bartoshesky, L. & Crampton, K. (1967). Benefit–cost analysis of rubella vaccination policy, *New England Journal of Medicine* **294**, 306–310.

[29] Stalhammer, N.O. (1996). An empirical note on willingness to pay and starting-point bias, *Medical Decision Making* **16**, 242–247.

[30] Thompson, M.S. (1986). Willingness-to-pay and accepts risks to cure chronic disease, *American Journal of Public Health* **76**, 392–396.

[31] Viscusi, W.K. (1978). Labor market valuations of life and limb: empirical estimates and policy implications, *Public Policy* **26**, 359–389.

(*See also* **Standard Gamble Technique**; **Time Trade-off Technique**; **Utility in Health Studies**)

BERNIE O'BRIEN

Cost-effectiveness in Clinical Trials

Cost-effectiveness analysis (CEA) is a method used to evaluate the outcomes and costs of a project or intervention. It is a tool aimed at assisting decision-makers in judging the efficiency consequences of various courses of action. In the health care arena CEA is used to assess the costs of alternative approaches to achieving certain health outcomes [11]. Usually the results of a CEA are summarized in a series of cost-effectiveness ratios that show the difference in costs divided by the difference in outcomes of one treatment relative to another. This ratio of differences is called the incremental cost-effectiveness ratio (ICER). When cost-effectiveness studies use the same **outcome measures** (such as quality-adjusted life years), researchers can compare ICERs across different types of interventions and populations. In an idealized world, if a health plan wanted to maximize the health of its enrolled population generated by a given budget, then it would order interventions so that those with the lowest ICERs would be adopted first. It would distribute funds across diseases and service categories so that the ICERs in each category would be equal. This would ensure that each dollar of expenditures worked equally hard in producing health effects.

In the context of evidence-based medicine, CEA can be viewed as a tool for quantifying and interpreting data on costs and outcomes. The modern era of managed care in the US and constrained budgets in publicly funded systems has led to clinical decisions being more influenced by budgetary considerations than was commonly the case in the 1960s and 1970s. If the design of clinical programs, clinical pathways and other decision-making support systems are to be founded on evidence, then it is important to use all types of evidence in designing them. It is therefore logical to recommend including evidence on cost-effectiveness in the evidence base used in supporting therapeutic choices. Such information can be (and is) used by national health systems and health plans in deciding which interventions to introduce and pay for as well as how to structure the treatment process.

There are a number of different approaches to obtaining the necessary data for conducting CEAs [7, 11]. Since good outcome measures are a fundamental component of all CEAs, and because clinical trials are considered the gold standard for evaluating outcomes stemming from clinical interventions, there is interest in incorporating cost-effectiveness studies into clinical trials [7, 9, 19, 21] This interest includes both adding a cost-effectiveness component to a clinical trial that is being planned to test the efficacy of a specific intervention, as well as conducting clinical trials for the purpose of determining whether a given intervention is cost-effective. (The former is sometimes described as adding an economic component to a clinical trial while the latter are sometimes called pragmatic trials.)

In this article we focus on how the design and measurement strategies of clinical trials need to be adapted to successfully add a cost-effectiveness component. Key considerations include the appropriateness of adding a cost-effectiveness component to the type of clinical trial that is being planned, sample size considerations, and the nature of the additional data that would need to be collected [4, 6]. Although these are identified as separate issues, they are obviously interrelated. Detailed discussions of various approaches to conducting cost-effectiveness analyses *per se* can be found elsewhere [7, 11, 13].

Definition of Terms

Before examining these issues, it is useful to set out some definitions. The term *treatment* is construed to include a range of clinical interventions (such as drugs and surgical procedures), settings (inpatient, outpatient), providers (such as physicians and nurse practitioners) and management strategies. The term *efficacy* generally describes the benefits generated by a treatment administered under ideal conditions, such as those found in clinical trials where patients are carefully selected according to specified criteria and then randomized to treatment and control groups to eliminate the confounding impacts of patient self-selection into treatment and clinical outcomes. *Effectiveness* describes the benefits of efficacious treatments that are realized under clinical conditions that reflect the usual circumstances under which medical care is delivered. Using a different terminology, efficacy studies are generally designed to have a high degree of internal validity, while effectiveness studies must have a high degree of external validity. Although there is some overlap, the endpoints of clinical trials and effectiveness studies tend

to differ. Traditionally clinical trials have focused on end-points such as mortality and clinical symptoms (blood pressure and cholesterol levels). Effectiveness studies often include these end-points. However, in recent years there has been a movement towards including outcome measures that are more meaningful to patients, such as quality of life. Wells [37] provides a good overview of the major differences between efficacy and effectiveness research.

A *cost-effectiveness* study examines both the use of resources (the costs) and the resulting changes in patients' health status (the outcomes) associated with the two or more treatments that are being compared. It is worth pointing out that the terms cost-effectiveness and cost-saving are distinct. A new treatment that produces outcomes identical to usual care but at lower cost would be both cost-saving and cost-effective (more health gain per dollar spent relative to usual care). However, a new intervention that both increased costs and improved health outcomes would be cost-effective if the ICER is similar to that of treatments that are part of usual care. Although cost-effectiveness studies can be narrowly focused (e.g. the determination of the incremental cost of lowering systolic blood pressure by one unit), the greatest interest lies in estimating cost-effectiveness ratios in terms of the cost per quality-adjusted life year (QALY) gained from an intervention [10]. As a measure of health outcome, a QALY assigns a weight to each health state in each time period. These weights range from 0 to 1, where a weight of 1 corresponds to optimal health and a weight of 0 corresponds to death. The weights reflect patient preferences for a particular health state. Because the states of health are weighted by the preferences or utility associated with a health state, a QALY is also called a utility-based quality of life indicator.

Cost-effectiveness studies can be conducted from a number of perspectives. The perspective taken determines which outcomes and costs are taken into consideration and how they are measured. The three most common perspectives are those of the individual, the health plan or society at large. When a CEA is conducted from a social perspective, the analyst considers all the effects of the interventions, and counts all health outcomes and costs associated with the intervention regardless of who benefits and who bears the costs. In what follows we focus on CEAs from a social perspective.

Considerations for Adding a Cost-effectiveness Component to a Clinical Trial

General

Clinical trials are undertaken to study a wide range of issues from testing new drugs to evaluating new treatment methods for breast cancer. Some trials are very large and involve thousands of patients and multiple sites (*see* **Multicenter Trials**), while others are more limited in terms of their size and patient population. An investigator would add a cost-effectiveness component to a clinical trial in pursuit of answers to two questions:

1. What is the ICER of the intervention?
2. Does the estimated ICER indicate that the treatment is cost effective; that is, is the ICER equal to or less than some benchmark? (For example, a Canadian research team [18] assigned grades to different ICERs. It assigned a "B" to technologies that were more effective than existing ones and cost about $20 000 per QALY gained. It recommended that Grade B technologies be adopted.)

In deciding whether it makes sense to add a cost-effectiveness component to a clinical trial four questions should be posed [5, 11, 20]:

1. Does the trial address an issue that is likely to be of economic importance? Will the information generated by adding a cost-effectiveness component be of sufficient value to decision-makers to justify the cost of adding the component?
2. Is the trial testing something that is relevant to community practice? For example, a trial that compares an innovative treatment to usual care is more informative than one that compares a new drug to a placebo (unless of course there are no existing treatments for the condition under study).
3. Are the results likely to be of interest to decision-makers? Decision-makers include patients, physicians, large-scale provider organizations (hospitals, integrated delivery systems), health plans and government agencies.
4. Will the results of the trial have external validity? The more highly selected the patient population to be studied and the more closely monitored

the treatment protocol, the less likely it is that the outcomes would be representative of those observed in the community under conditions of community care.

Sample Size Considerations

It is likely that adding a cost-effectiveness component will raise issues with respect to sample size. There are two main problems that have to be addressed: (i) the fact that cost data typically display different properties than common clinical indices and (ii) the challenge of calculating confidence intervals for a ratio such as an ICER.

The ICER is a ratio that is composed of parameters that measure costs and outcomes. As such, the variance around an ICER encompasses the variance in costs and the variance in outcomes. Typically, a clinical trial is designed so that the sample size is based on the minimum number of subjects required to detect a given (predetermined) clinically important difference in outcomes with a given power (e.g. 0.90) at a conventional level of significance (e.g. $P < 0.05$) (*see* **Sample Size Determination**). The designers of clinical trials often impose strict exclusion and inclusion criteria (*see* **Eligibility and Exclusion Criteria**) to reduce the variation in outcomes in order to reduce the necessary sample size. In addition, many clinical measures have been designed to display particular statistical properties such as a Normal distribution.

Data on health care costs have different properties than do clinical data such as symptom counts and blood pressure measurements. Cost data typically do not follow a standard Normal distribution. They tend to exhibit density masses at minimum levels of expenditure and are highly skewed to the right. (For instance, hospitalizations related to side-effects associated with treatment may be very rare. However, the costs associated with the few hospitalized episodes are very high.) The implication of this is that the variance in costs will tend to be considerably greater than the variance in clinical outcome measures. Thus, it is likely that the sample size requirements for testing the significance of an estimated ICER will be larger than those for testing the significance of a clinical outcome.

There are other issues related to determining the variance of the ICER, and thus the confidence intervals. For instance, cost and outcome estimates in an ICER and their variances may not be independent. On the one hand, if the correlation between costs and outcomes is negative (higher costs associated with worse outcomes), then the variance of the ICER will increase. On the other hand, if the correlation between costs and outcomes is positive, then the variance will be reduced. Furthermore, the ICER is a ratio. Since the difference in outcome measure enters in the denominator of the ICER, small differences in outcomes between interventions can create very large differences in the ICER.

Several empirical strategies have been advanced for estimating confidence regions for ICERs [24, 34]. These methods involve different levels of computational burden and assumptions about the underlying relations between treatments and effects. Briggs & Gray [2], for example, use this literature to explore techniques for deriving a sample size formula for a CEA based on simple combinations of the confidence limits on costs and outcomes.

Effectiveness Data

In most cases the outcome measures collected under the trial will have to be expanded to include information on the quality of life experienced by subjects in all treatment arms. These measures should include generic measures of quality of life so that ICERs can be compared across disease types and interventions.

Ideally, utility-based quality of life information should be obtained [10]. There is a variety of methods for collecting this utility-based quality of life information. One method is to administer a general preference rated instrument for assessing quality of life. These instruments, which measure a number of concepts (health perceptions, social functioning, psychological functioning, physical functioning and impairment), are administered to the subjects periodically during the trial. Preference scores that have been obtained through earlier research projects are then used to weight the responses. Examples of such instruments are the EuroQol instrument [3, 8, 17], Quality of Well Being [15, 16] and the health utilities index [31, 33]. A second method is to map clinical indicators (such as the Hamilton rating scale, an instrument used to measure depressive symptoms [12]), into QALYs by using published information on the quality of life associated with that condition [20]. A third method is to obtain the utility estimates from the people who are actually enrolled in the trial. There are three standard ways of obtaining these. These are

asking subjects to: (i) respond to standard gamble questions (where subjects are asked to compare life in a given health state that is a sure thing to a gamble with a probability P that perfect health is the outcome and $1 - P$ that death is the outcome) [33]; (ii) respond to time–trade off questions (where subjects are asked to trade off life years in a state of less-than-perfect health for a shorter life span in a state of perfect health) [32]; or (iii) complete a visual analog (a direct rating scale where they are asked to place a mark at some point between two anchor points that indicates their rating of a give health state) [22].

The second approach to obtaining information on quality of life is to administer a nonpreference weighted quality of life instrument. There are several such instruments available. These instruments generally assess the same domains as the utility-based quality of life instruments, but the scores are not utility weighted. Examples of such instruments are the Medical Outcomes Scale (MOS) or SF-36 [29, 35], the Nottingham Health Profile [14] and the Sickness Impact Profile [1].

The third approach is to use a disease-specific instrument such as the Hamilton Rating Scale for Depression [12] and the McGill Pain Questionnaire [23]. As indicated above, it is not possible to compare the ICERs found in these studies with those of other studies that examine other disease conditions.

It should be noted that the field of health assessment is still developing and there is no general agreement on which is the best instrument. A general overview of rating scales and questionnaires can be found in Hunt et al. [14], McDowell & Newell [22] and Patrick & Erickson [26].

A quite different approach to measuring the "value" of outcomes on a common scale is to use what economists refer to as the "willingness-to-pay" approach. This involves estimating individuals' willingness-to-pay for new medical interventions that may improve health [28, 30]. Under this approach, the researcher still has to assess the outcomes of the trial; that is, the effect of the intervention relative to the status quo in terms of change in physical functioning, mental functioning and pain would still have to be assessed. However, the weights to be applied to these effects would come from studies that estimate the willingness-to-pay for these effects. The theoretical advantage of the willingness-to-pay measure is that it allows for a comprehensive consideration of what individuals value about medical treatments, including intangible benefits such as pain and discomfort which are not easily captured by most outcome measures. Some economists have suggested that the additional premiums that people are willing to pay in order to have a new intervention added to a health plan's set of reimbursable procedures summarizes the value of benefits from new interventions [27]. However, many are offended by valuing the gain in health in monetary terms. Furthermore, willingness-to-pay measures typically favor the health of the wealthy over the poor.

Adding additional outcome measures to the set of outcomes measures to be gathered during a clinical trial will both increase the cost of the trial and increase patient burden. Therefore, the designers of the cost-effective component will have to assess the impact of adding these outcome variables and to select the outcome variables carefully.

Cost Data

It is necessary to collect information on the costs incurred by subjects in all treatment arms. However, there is some discretion in determining the exact cost data to be collected. The range of the data to be collected will depend on the nature of the intervention and the health condition being treated. For instance, if the treatment is narrowly focused, and if it is unlikely that there is any relationship between the health condition that is being targeted and other health conditions, then it may be sufficient to focus only on the costs associated with the targeted condition. For example, in examining a trial that compared hospital with outpatient treatment for pelvic inflammatory disease, the researchers focused on health care costs associated with conditions related to fertility and gynecological problems [25]. However, treating some conditions may have broad health consequences. In this case it is necessary to measure the costs associated with treating not only the targeted condition but also other conditions that may be affected. For example, in looking at alternative treatments for depression, the researchers captured both mental and physical health costs [20] because there were strong arguments in the literature that the costs of good treatment for mental disorders would be offset by a decrease in the costs of treating physical health problems. Finally, it is possible that treating specific conditions may have effects on sectors of the economy other than the health care sector. For

example, in a study that compared intensive outpatient therapy with usual inpatient care for people with severe mental illness, Weisbrod [36] collected information on the costs associated with treating mental health conditions, the criminal justice system, family burden and care-taking costs.

In most cases costs are estimated by obtaining information on the quantity of services used and then by applying cost weights to those data. There are a number of ways to gather the utilization data. These include: obtaining information from the records maintained by the clinical trial, using administrative records maintained by the health care centers in which the subjects receive their medical care services, using the administrative records maintained by health plans or government agencies that pay for care, and from direct patient interview. The nature of the administrative records will vary by provider setting, agencies and across countries. Some records (such as medical records and hospital discharge records) will include only utilization information. Other administrative records systems (such as the hospital billing records in the US and the information maintained by insurance plans) will include information on the use of specific services, the charges (prices) associated with those services and the payments made.

The use of services must be assigned a monetary value. Ideally the researcher should estimate the social cost of resources used, or equivalently the marginal cost of those resources. Economic theory indicates that in efficient markets, prices are equal to marginal cost. However, health care markets are not competitive in the theoretical sense and thus the prices do not typically reflect costs. (This is true both in national health systems as well as in more market-based systems such as the US.) Thus, other methods must be used to assign cost weights to the service used and other resources. In some cases the analysts may conduct specific cost-finding studies. However, in most cases they use information from other sources – the hospital cost accounting systems and external fee schedules – to assign cost weights. For example, in the US researchers often use the fees from the Medicare fee schedule to assign cost weights to the different types of physician services. They use prices published in standard references such as the *Red Book* to assign prices to pharmaceuticals. They may use the information from the hospital bill to estimate the cost of a hospital stay; however, they will usually adjust the charges by the hospital cost to charge ratio in order to obtain a measure that is a continuing proxy for costs. In other countries costs are estimated by valuing the inputs into the treatment process by using salaries and input purchase prices (drugs).

In general, researchers need to collect information not only on the medical care costs but also on the time and transportation costs incurred in accessing and receiving care. If volunteers are used in any portion of their treatment interventions, then their time must be measured and costs impacted. The costs associated with actually conducting the trial itself (the design costs, data-collection costs, administrative/managerial costs and analyses costs) should not be included since they are not part of delivering care.

Gathering the cost data necessary to conduct a cost-effectiveness study will increase the cost of conducting the study and may increase patient burden. Therefore, it is necessary to select cost measures carefully.

Extending the Analysis to the Posttrial Period

If there are significant differences in the outcomes between the people who are enrolled in the clinical trial, it may be necessary to model the differences that would expected to be observed after the completion of the trial. The short timeframe of many clinical trials makes this an important consideration. For example, suppose that a trial is conducted to test the efficacy of a new antidepressant medication. Suppose that it is a six-month trial and that at the end of the sixth month a higher proportion of subjects who have been randomized to the experimental arm have recovered. The cost-effectiveness of that drug will differ depending on whether that different recovery rate can be expected to be maintained or whether a comparable number of subjects in each treatment arm will have recovered at the end of eight months or a year. Thus, an ICER ratio based only on the trial data will not be a good indicator of the true ICER of one treatment versus another. In this case, simulation and modeling methods must be used to extend the data beyond the trial period [7].

Sensitivity Analyses

It is clear that a number of judgments have to be made in conducting a CEA. Researchers must make

more judgments in estimating the ICERs than they do in analyzing the clinical findings of the trial. These judgments are related to the selection of the approach used to measure quality of life, the cost weights, the range of costs to include in the analyses and so forth. Judgments also have to be made about the reliability of the outcome treatment effect. Since the researchers often monitor the treatment being applied to the experiment group very closely, it is likely that the outcomes observed during the trial may be better than those that will be observed in practice. Thus, the researchers have to make some assumptions either about the costs that a practice would have to incur if care were to be monitored or about the decreases in effects that will be observed in practice without the monitoring that took place during the trial. As a result, researchers who conduct CEAs often estimate a number of ICERs in which they make different assumptions about several of the values. They then determine how sensitive their conclusions are to these assumptions.

Interpretation of CEAs

CEAs conducted in the context of clinical trials offer important information about the *potential* efficiency of a new clinical intervention or technology. CEAs in clinical trials are conducted under a particular set of allocation rules. That is, particular patient populations are selected to participate in the trials; and participants are randomly assigned to treatment independent of clinician or patient preferences. In the community setting, clinical technologies are put into practice under very different allocation rules and therefore realized effectiveness may differ markedly from those found in a clinical trial even if the technologies are appropriately delivered. An example may illustrate the point. Proton Pump Inhibitors have been shown to be cost-effective interventions for severe esophagitis relative to older H-2 Blockers. However, in practice a substantial portion of Proton Pump Inhibitors are used for the treatment of milder forms of esophagitis and in those cases offer few clinical benefits over the H-2 Blockers at a considerably higher cost. Thus, in practice the ICER of Proton Pump Inhibitors is likely to be considerably higher than that found in the clinical trials. For this reason, a CEA based on clinical trials offers important yet incomplete information on the efficiency of adopting a new clinical intervention into practice.

Conclusions

CEAs can offer important insights into the "value" of a new medical intervention or even into old interventions. Increasingly payers, health plans and regulators are interested in understanding the budgetary demands associated with new clinical technologies that promise enhanced health outcomes. CEAs linked to clinical trials can sometimes offer a rigorous method of informing such questions.

It is clear that adding cost-effectiveness components to clinical trials will increase the costs of conducting the trial. Not only is it likely that the sample size will have to be increased, but also additional data will have to be collected. There will be additional patient burden. Thus, adding CEAs to all clinical studies would not itself be a cost-effective use of society's resources. Clinical trials that are good candidates for CEA augmentation include those that are testing new treatments that may be significant from an economic standpoint in that they offer significant improvements in outcomes at a significant increase in cost or that they offer significant cost savings for similar levels of outcomes when compared with existing interventions.

Acknowledgments

Source support from NIMH Grants MH43703, NIMH (MH56925) and the John D. and Catherine T. MacArthur Foundation is gratefully acknowledged.

References

[1] Bergner, M., Bobbitt, R.A., Carter, W.B. & Gilson, B.S. (1981). The sickness impact profile: development and final revision of a health status measure, *Medical Care* **19**, 787–805.

[2] Briggs, A.H. & Gray, A.M. (1998). Power and sample size calculations for stochastic cost-effectiveness analysis, *Medical Decision Making* **18**, S81–S92.

[3] Brooks, R. with the EuroQol Group (1996). EuroQol: the current state of play, *Health Policy* **36**, 53–72.

[4] Coyle, D., Davies, L. & Drummond, M.F. (1998). Emerging issues in designing economic evaluations alongside clinical trials, *International Journal of Health Technology Assessment in Health Care* **14**, 135–144.

[5] Drummond, M. (1995). Economic analysis alongside clinical trials: problems and potential, *Journal of Rheumatology* **22**, 1403–1407.

[6] Drummond, M. & Davies, L. (1992). Economic analysis alongside clinical trials: revisiting and methodological issues, *International Journal of Health Technology Assessment in Health Care* **7**, 561–573.

[7] Drummond, M., O'Brien, B.J., Stoddart, G.L. & Torrance, G.W. (1997). *Methods for the Economic Evaluation of Health Care Programmes*, 2nd Ed. Oxford University Press, New York.

[8] EuroQol Group (1990). EuroQol: a new facility for the measurement of health-related quality of life, *Health Policy* **16**, 199.

[9] Fitzpatrick, R. & Davies, L. (1998). Health economics and quality of life in cancer trials: report based on a UKCCCR workshop, *British Journal of Cancer* **77**, 1543–1548.

[10] Gold, M.R., Patrick, D.L., Torrance, G.W., Fryback, D.G., Hadorn, D.C., Kamlet, M.S., Daniels, N. & Weinstein, M.C. (1996). Identifying and valuing outcomes, in *Cost-Effectiveness in Health and Medicine*, M.R. Gold, J.E. Siegel, L.B. Russel & M.C. Weinstein, eds. Oxford University Press, New York, pp. 82–134.

[11] Gold, M.R., Siegel, J.E., Russell, L.B. & Weinstein, M.C., eds. (1996). *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York.

[12] Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry* **23**, 56–62.

[13] Hargreaves, W.A., Shumway, M., Hu, T.W. & Currel, B. (1998). *Cost Outcomes Methods for Mental Health*. Academic Press, San Diego.

[14] Hunt, S., McEwen, J. & McKenna, S.P. (1986). *Measuring Health Status*. Croom Helm, London.

[15] Kaplan, R.M. (1999). Health-related quality of life in mental health services evaluation, in *Cost-Effectiveness of Psychotherapy*, N.E. Miller & K.M. Magruder, eds. Oxford University Press, New York, Chapter 16, pp. 160–173.

[16] Kaplan, R.M. & Anderson, J.P. (1988). A general health policy model: update and applications, *Health Services Research* **23**, 203–235.

[17] Kind, P. (1996). The EuroQol instrument: an index of health-related quality of life, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, B. Spilker, ed. Lippincott-Raven, Philadelphia.

[18] Laupacis, A., Fenny, D., Detsky, A.S. & Tugwell, P.X. (1992). How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluation, *Canadian Medical Association Journal* **146**, 473–481.

[19] Lave, J.R. & Schulberg, H.C. (1999). Integrating cost-effectiveness analyses within clinical trials of treatment for major depression in primary-care practice, in *Cost-Effectiveness of Psychotherapy*, N.E. Miller & K.M. Magruder, eds. Oxford University Press, New York, pp. 75–84.

[20] Lave, J.R., Frank, R.G., Schulberg, H.C. & Kamlet, M.S. (1998). Cost-effectiveness of treatments for major depression in primary care practice, *Archives of General Psychiatry* **55**, 645–651.

[21] Manheim, L.M. (1998). Health services research clinical trials: issues in the evaluation of economic costs and benefits, *Controlled Clinical Trials* **19**, 149–158.

[22] McDowell, I. & Newell, C. (1996). *Measuring Health: A Guide to Rating Scales and Questionnaires*, 2nd Ed. Oxford University Press, New York.

[23] Melzack, R. (1975). The McGill Pain Questionnaire: major properties and scoring methods. *Pain* **1**, 277–299.

[24] Mullahy, J. & Manning, W.G. (1994). Statistical issues in cost effectiveness analyses, in *Valuing Health Care: Costs, Benefits and Effectiveness of Pharmaceuticals and Other Medical Technologies*, F. Sloan, ed. Cambridge University Press, New York.

[25] Ness, R.B., Soper, D.R., Peipert, J., Sondheimer, S.J., Holley, R.L., Sweet, R.L., Hemsell, D.L., Randall, H., Hendrix, S.L., Bass, D.C., Kelsey, S.F., Songer, T.J. & Lave, J.R. (1998). Design of the PID Evaluation and Clinical Health (PEACH) Study, *Controlled Clinical Trials* **19**, 499–514.

[26] Patrick, D.L. & Erickson, P. (1993). *Health Status and Health Policy: Allocating Resources to Health Care*. Oxford University Press, New York.

[27] Pauly, M.V. (1994). Valuing health care benefits in monetary terms, in *Valuing Health Care: Cost, Benefits and Effectiveness of Pharmaceuticals and Other Medical Technologies*, F. Sloan, ed. Cambridge University Press, New York.

[28] Schelling, T.C. (1984). The life you save may be your own, in *Choice and Consequence*, T.C. Schelling, ed. Harvard University Press, Cambridge.

[29] Stewart, A.L. & Ware, J.E. (1992). *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*. Duke University Press, Durham.

[30] Tolley, G., Kenkel, D. & Fabian, R. (1994). *Valuing Health for Policy*. University of Chicago Press, Chicago.

[31] Torrance, G.W. (1986). Measurement of health utilities for economic appraisal, *Journal of Health Economics* **5**, 1–30.

[32] Torrance, G.W., Furlong, D., Fenny, D. & Boyle, M. (1995). Multi-attribute preference functions: health utilities index, *Pharmacoeconomics* **9**, 503–520.

[33] Torrance, G.W., Thomas, W.H. & Sackett, D.L. (1972). A utility maximizing model for evaluation of health care programs, *Health Services Research* **7**, 118–133.

[34] van Hout, B.A., Al, M.J., Gordon, G.S. & Ruten, F.F. (1994). Cost, effects and C.E-ratios alongside a clinical trial, *Health Economics*, **3**, 309–319.

[35] Ware, J. & Sherbourne, C. (1992). The MOS 36-item short form health survey I: conceptual framework and item selection, *Medical Care* **30**, 473–483.

[36] Weisbrod, B.A. (1981). Benefit-cost analysis of a controlled experiment: treating the mentally ill, *Journal of Human Resources* **16**, 523–548.

[37] Wells, K.B. (1999). Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research, *American Journal of Psychiatry* **156**, 5–10.

JUDITH R. LAVE & RICHARD G. FRANK

Counter-matching

Counter-matching is nested case–control study design (*see* **Case–Control Study, Nested**) in which a **covariate** is known on all cohort members, and controls are sampled to yield covariate-stratified case–control sets. The design is advantageous when a major analysis variable, or a correlate, is available on all cohort members and additional information is to be collected on a sample. Unbiased estimation requires the numbers of risk set members in each counter-matched sampling stratum.

The Design

Counter-matching was originally proposed as an exposure-stratified, individually matched nested case–control study method in the context of continuous failure-time (cohort) data [13, 15]. Counter-matched sets are characterized by the number of subjects m_l from each of the L *sampling strata* defined by the counter-matching variable. It is required that the counter-matching variable is known for all risk set members and, in addition to the case, controls are randomly sampled without replacement (*see* **Sampling With and Without Replacement**) from each of the sampling strata in the risk set to yield the required m_l subjects. As illustrated in Table 1, when the case is from sampling stratum 2, m_l controls are sampled from the n_l in risk set sampling stratum l except for stratum 2, from which $m_2 - 1$ controls are sampled. In the special case of two sampling strata, with one subject from each stratum (the 1 : 1 design), the control is sampled from the opposite sampling stratum of the case; the opposite of matching and thus motivating the name.

For grouped failure-time or simple **binary data** (multiple cases in the case–control set) counter-matching, the m_l would generally depend on total number of cases $|\mathbf{D}|$ in the study base [17]. (i.e. the design is characterized by $m_1(|\mathbf{D}|), \ldots, m_L(|\mathbf{D}|.)$ The actual number of cases that fall into counter-matched stratum l, $|\mathbf{D}_l|$, is random and determines the number of controls to be sampled from stratum l, $m_l - |\mathbf{D}_l|$. This is illustrated in Table 2.

Statistical Analysis

Estimation of Rate (Odds) Ratio Parameters. The analysis of the counter-matched data must take into account the **stratification** of sampled sets. For individual **matching** (continuous time risk sets), the **partial likelihood** is based on the probability that a subject is the case given the counter-matched set and requires the control sampling probabilities. In particular, with l_j indexing the sampling stratum for subject j, the probability of drawing the counter-matched sample if j were the case is given by $\pi_j = n_{l_j}/m_{l_j}\left[\prod_{l=1}^{L}\binom{n_l}{m_l}\right]^{-1}$. This leads to the **likelihood**

$$\prod_{\text{sets}} \frac{r_{case}(\beta)\dfrac{n_{l_{case}}}{m_{l_{case}}}}{\sum_{j\in\text{set}} r_j(\beta)\dfrac{n_{l_j}}{m_{l_j}}}, \tag{1}$$

where $r_j(\beta) = r(Z_j;\beta)$ is the rate ratio associated with Z_j and β is the rate ratio parameter from a **proportional hazards model**. This likelihood can be fitted using standard **conditional logistic regression** software that allows for fixing a regression parameter. For instance, for the standard **loglinear model** $(n_{l_j}/m_{l_j})r_j(\beta) = \exp(Z_j\beta + \log w_j)$ where $w_j = n_{l_j}/m_{l_j}$. So, the log weight can be included in the model with fixed parameter equal to one (an *offset* in the model). Aside from this offset, analysis proceeds as in any standard conditional logistic regression analysis for individually matched case–control studies. The likelihood (1) has the usual likelihood properties so that the standard likelihood inference techniques apply, with no additional modeling assumptions other than appropriate specification of the sampling weights [9, 15]. The full asymptotic theory has been derived [4, 13] and the performance,

Table 1 Individually matched counter-matched study (one case per counter-matched set). In this example, the case is in sampling stratum 2 so m_l controls are sampled from each stratum except stratum 2 for which $m_2 - 1$ controls are sampled.

	Sampling stratum				
	1	2	$\cdots$	L	Total
Cases	0	1	$\cdots$	0	1
Controls	m_1	$m_2 - 1$	$\cdots$	m_L	$\sum m_l - 1$
Total in sample	m_1	m_2	$\cdots$	m_L	$\sum m_l$
Total in risk set	n_1	n_2	$\cdots$	n_L	$\sum n_l$

Table 2 Unmatched counter-matched study (multiple cases per counter-matched set). The m_l are counter-matching design parameters representing the total number from stratum l. With $|\mathbf{D}_l|$ number of cases in stratum l, $m_l - |\mathbf{D}_l|$ controls are randomly sampled from stratum l to make a total of m_l subjects.

	Sampling stratum				
	1	2	$\cdots$	L	Total
Cases	$\lvert\mathbf{D}_1\rvert$	$\lvert\mathbf{D}_2\rvert$	$\cdots$	$\lvert\mathbf{D}_L\rvert$	$\lvert\mathbf{D}\rvert = \sum \lvert\mathbf{D}_l\rvert$
Controls	$m_1 - \lvert\mathbf{D}_1\rvert$	$m_2 - \lvert\mathbf{D}_2\rvert$	$\cdots$	$m_L - \lvert\mathbf{D}_L\rvert$	$\sum m_s - \lvert\mathbf{D}\rvert$
Total in sample	m_1	m_2	$\cdots$	m_L	$\sum m_l$
Total in risk set	n_1	n_2	$\cdots$	n_L	$\sum n_l$

compared with other designs, has been evaluated in a number of situations [1, 7, 10, 11, 16].

For counter-matching with multiple cases per set, the likelihood requires the (control selection) probability of picking a particular counter-matched set if a set of subjects $\mathbf{s}$ (of the same size as the actual set of cases $\mathbf{D}$) were the set of cases. With $\mathbf{s}_l$ the set of subjects from $\mathbf{s}$ in sampling stratum l, and $|\mathbf{s}_l|$ the number of subjects in $\mathbf{s}_l$, the counter-matching control selection probability is given by

$$\pi_{\mathbf{s}} = \left[\prod_{l=1}^{L} \frac{n_l(n_l-1)\cdots}{\frac{(n_l-|\mathbf{s}_l|+1)}{m_l(m_l-1)\cdots}\atop(m_l-|\mathbf{s}_l|+1)}\right]\left[\prod_{l=1}^{L}\binom{n_l}{m_l}\right]^{-1}. \quad (2)$$

This leads to the **likelihood** [17]:

$$\prod_{\text{sets}} \frac{r_{\mathbf{D}}(\beta)\left[\prod_{l=1}^{L} \frac{n_l(n_l-1)\cdots}{\frac{(n_l-|\mathbf{D}_l|+1)}{m_l(m_l-1)\cdots}\atop(m_l-|\mathbf{D}_l|+1)}\right]}{\sum_{\mathbf{s}\subset\tilde{\mathcal{R}}:|\mathbf{s}|=|\mathbf{D}|} r_{\mathbf{s}}(\beta)\left[\prod_{l=1}^{L} \frac{n_l(n_l-1)\cdots}{\frac{(n_l-|\mathbf{s}_l|+1)}{m_l(m_l-1)\cdots}\atop(m_l-|\mathbf{s}_l|+1)}\right]},$$

where $r_{\mathbf{s}}(\beta) = \prod_{j\in\mathbf{s}} r(Z_j;\beta)$ is the product of **odds ratios** associated with the Z_j in a **proportional odds (logistic) model** (*see* **Logistic Regression**). Although, in general, standard software does not accommodate this likelihood, conditional logistic software can be "tricked" to estimate the odds ratio parameters when the odds model is log-linear [17]. Because of the inherently correlated structure, derivation of the asymptotic properties of likelihood (2) poses some theoretical challenges that have not yet been addressed [2]. However, limited derivation and **simulation** studies indicate that the efficiency performance of (2) for the grouped data counter-matched is similar to that of (1) for individually matched data [17].

Estimation of Other Parameters. Methods for estimation of the cumulative baseline hazard and **absolute risk** from counter-matched data have been described [14] as well as methods for the estimation of regression parameters in the **Aalen linear model** [5]. A weighted unconditional logistic regression can be used to estimate baseline odds parameters from grouped data [17].

Examples

Crystalline Silica Exposure and Silicosis in Gold Miners. In a comparison of nested case–control study design options in an occupational cohort study of 3000 gold miners, counter-matching was compared with **random sampling** of controls [20]. A major cost component in this study was in obtaining silica exposure data from dust samples taken from the mines; a nested case–control study could have avoided much of this expense. Investigators compared random sampling and years-of-employment counter-matching of controls. The correlation between years of employment and cumulative silica dust exposure is about 0.7, and it was found that three counter-matched controls yielded the same statistical efficiency as 15 randomly sampled controls at the same cost. The situation considered in this example is typical of many **cohort studies** in which a "broad" measure (e.g. years of employment) is associated with disease and the nested case–control study is undertaken to identify better the possible causative agents (e.g. cumulative silica dust exposure). Counter-matching incorporates the cohort "broad measure" into the

sampling in order to obtain a sample that is more informative about the specific exposure, compared to random sampling.

Radiation, Hormones, and Breast Cancer in a Cohort of Japanese Atomic Bomb Survivors. A strong association of premenopausal breast cancer risk and radiation dose has been observed in the Radiation Effects Research Foundation's Life Span Study (LSS) of atomic bomb survivors [21] (*see* **Radiation Epidemiology**). For the Adult Health Study (AHS) cohort, a subgroup of LSS volunteers who participated in biennial clinical examinations, stored blood serum was available for 5724 women, from which estradiol levels could be measured, at some expense. The radiation-dose counter-matched study was undertaken to investigate associations with estradiol (and other hormonal and antioxidant factor) levels and radiation dose jointly on breast cancer risk. For each of the 80 premenopausal breast cancer cases, two controls were sampled with the counter-matching strata defined by radiation dose with a zero dose category and two exposure groups defined by the median of the distribution of the combined cases; that is, a control was randomly sampled from each of the (noncase) sampling strata [11, 19]. Given the actual radiation doses and a likely distribution of estradiol levels, the counter-matching design was compared with random sampling and radiation dose matching of controls. It was found that counter-matching was much more efficient than random sampling and of about equal efficiency to matching for a range of positive multiplicative radiation-estradiol **interactions**. But, unlike matching, counter-matching still allows for estimation of the radiation main effects so that a wider range of questions about the variation of breast cancer risk with radiation dose and estradiol levels can be addressed; in particular, about potential **confounding** [11]. In this study, the counter-matching variable was based on the actual exposure and the goal of the study is to investigate **effect modification** of the exposure-disease risk relationship.

Gene Susceptibility to Radiation Exposure for Second Breast Cancer Risk: The WECARE Study. The main goal of this study is to determine whether the risk of breast cancer after exposure to radiation is higher in women possessing **polymorphisms** of genes involved in double-strand break repair. The cohort consists of 31 243 women diagnosed with breast cancer identified by five **cancer registries**. There were 801 women with asynchronous bilateral breast cancer who were the cases in this study. A cohort of women with breast cancer is advantageous for addressing the study questions for two main reasons. First, women who have had a breast cancer are likely to have a higher **prevalence** of **genotypes** that cause the disease. Second, a large percentage of the women (about 40%) underwent radiation therapy for their first breast cancer. The "scatter" from the therapeutic radiation can result in significant exposure to the contralateral breast that is often well documented in treatment records. A nested case–control study with two controls per case was dictated by cost considerations. Now, although it may be imperfect, all the cancer registries record whether radiation therapy was part of the treatment regimen (RRT+) or not (RRT−). This was used in an RRT counter-matched design in which two controls were sampled so that the case–control set would possess two RRT+ subjects and one RRT− subject. From each enrolled subject, a blood sample was obtained for the genotyping; medical treatment records were obtained (for all participants) to determine if they had had radiation treatment and, if so, the dose to the contralateral breast was determined; and the women filled out a mailed questionnaire that asked about other treatments and breast cancer risk factors. In this study, the counter-matching variable is correlated to the exposure of interest. Intuitively, there is "more variability" in radiation dose among RRT+ than RRT− subjects suggesting that 2 RRT+, 1 RRT− allocation would be more efficient for assessing radiation dose response and radiation-gene interaction than random sampling two controls. This intuition was confirmed in a simulation study comparison [3]. In this study, the counter-matching variable was based on the a dichotomous correlate of exposure and the goals of the study include characterization of the dose response for exposure and to investigate effect modification of the exposure-disease risk relationship.

Early Asthma Risk Factors Study (EARS) of In Utero and Early Life Exposures and Asthma. In a cohort study of determinants of respiratory health, over 5000 children from 12 communities and three grade levels were surveyed for "baseline" data [18]. Information collected at enrollment to the study included whether the student had ever been diagnosed with asthma, exposed to tobacco smoke *in utero* and during

childhood, and other factors that are potentially related to respiratory health. Using these baseline data, it was found that an asthma diagnosis at age five or younger was associated with maternal smoking during pregnancy (*in utero* smoke exposure) but not with environmental tobacco smoke exposure in early childhood [12]. The EARS follows up on this finding, first, to augment the smoking during pregnancy information (this was just a yes/no question in the baseline questionnaire) to assess dose–response and within-pregnancy timing of exposure and, second, to ascertain the child's GST-T1 and GST-M1 genotypes and to assess gene susceptibility. For the purpose of this study, the cohort (or study base) consists of subjects enrolled into the longitudinal study, followed from birth to age five. Since *in utero* exposure (yes/no) information is available for the cohort members, children diagnosed with asthma at age less than five years, the cases, were counter-matched on *in utero* smoke exposure, with the number sampled from exposed and unexposed approximately equal to the number of cases within matching strata defined by community, grade, and gender. The additional maternal smoking exposure and other information was obtained in a short interview and genotype status was assessed using standard PCR methods from buccal cells collected from subjects by swabbing the inside of the mouth. In this study, "yes/no" *in utero* smoke exposure information was available on all cohort members, and it is of interest both to obtain more precise maternal smoking information to assess timing and dose-response, as well as joint effects with genetic factors. Because the counter-matching factor is fairly correlated with the number of packs smoked and other smoking information, the study has much more statistical information for inference about such factors than would a comparably sized study with randomly sampled controls [17]. In contrast with the studies described above that are individually matched, this study implements the grouped data version of counter-matching.

Design Considerations

General Considerations. Relative to random sampling, counter-matching enhances statistical efficiency for analyses involving the counter-matching variable or correlates. However, statistical efficiency is reduced for analyses of factors that are not correlated to the counter-matched variable. Thus, counter-matching is appropriate when the study is focused on questions related to the counter-matching (generally exposure-related) factor. Situations for which there is a large efficiency gain for the counter-matching variable appear to be the situations for which there is a large efficiency loss for factors uncorrelated to the counter-matching variable. In particular, the degree of this gain/loss depends on the rarity of exposure, so that counter-matching on a rare exposure can be very advantageous for exposure-related analyses, to the great detriment of analyses of (main effects) of other factors. Whether this trade off is worthwhile depends on the specific goals of the study.

Counter-matching on an Exposure Correlate. Increased variability of exposure from the exposure-correlate stratified sampling provides some intuition for why counter-matching on an exposure correlate can increase efficiency relative to random sampling, as well as suggest a favorable allocation of subjects across the sampling strata. However, this increase in variability is tempered by the need for a weighted analysis that, in the absence of adequate correlation, works against increased efficiency [9]. Some insight into the relative efficiency of counter-matching to simple random sampling is provided in the dichotomous exposure/correlate situation with $1:1$ counter-matching on the correlate. Under the "null" situation of no association between exposure and disease, and denoting the **sensitivity** and **specificity** of the correlate for exposure by η and γ, respectively, the asymptotic efficiency of $1:1$ counter-matching relative to $1:1$ random sampling is $2[\eta\gamma + (1-\eta)(1-\gamma)]$. Counter-matching on the correlate is more efficient when the correlate is both more (or both less) than 50% sensitive and specific. Further, if the "correlate" and exposure are independent, then the counter-matching efficiency is always less than or equal to 1; always worse than random sampling [15, 16]. This illustrates a general principle that the counter-matching factor must be "somewhat correlated" to the exposure in order to realize an efficiency gain.

Allocation of Subjects in Sampling Strata. Although analyses based on (1) or (2) with the appropriate weights are valid for any allocation of subjects in sampling strata, efficiency depends on how the sampling strata are formed and the m_l. As a general guideline, when there are more counter-matched

subjects than strata, an allocation that will yield the greatest exposure variability appears to be most desirable [3]. When the exposure or correlate has more categories than subjects to be sampled, then it is advantageous to create sampling strata that approximately results in equal numbers of cases in each stratum [11, 13, 16, 20]. Determination of the "best" counter-matched design for a given study can be addressed using asymptotic variance calculations and computer simulation.

Counter-matching and Studies of Effect Modification. Although the relative performance of case–control designs for assessing effect modification depends on the distributions of the factors involved and the relationships between these factors and disease risk in a complex way, the increased variability in one or both of the factors in a counter-matched design generally results in enhanced efficiency. An efficiency comparison of random sampling and matching or counter-matching on one of the exposure variables indicated that counter-matching was similar or superior to matched or random sampling over a wide range of situations [10]. A study of feasibility of nested case–control studies for investigation of gene-susceptibility studies compared designs using three controls per case including counter-matched designs with sampling strata defined by exposure only, family history only, and both exposure and family history, and found the latter to be the most efficient in a wide range of circumstances [1]. Other efficiency comparisons have been done in the context of the WECARE and the Radiation, Hormone, and Breast Cancer studies described in the section "Examples" [3, 19].

Other Issues

Marginal Information of the Counter-matching Variable. If the only analysis variable is a function of the counter-matching stratum variable, then counter-matching likelihood is proportional to that of the full cohort. To see this, let $Z(l)$ be a function of the counter-matching stratum l. Then, because there are m_l subjects from stratum l, contributions to (1) become

$$\frac{r_{\text{case}}(\beta)\dfrac{n_{l_{\text{case}}}}{m_{l_{\text{case}}}}}{\sum_{l=1}^{L} m_l r(Z(l);\beta)\dfrac{n_l}{m_l}} \propto \frac{r_{\text{case}}(\beta)}{\sum_{l=1}^{L} n_l r(Z(l);\beta)},$$

which is the full cohort contribution. This can be similarly shown for grouped time likelihood (2).

Counter-matching and Matching. Counter-matching is essentially the opposite of matching. Matching is a technique to create case–control sets that are *similar* in the matching factor. Counter-matching is a technique to create case–control sets that are *diverse* in the counter-matching factor. The analytic consequences of the two methods are also opposite. In particular, exact matching results in no statistical information for inference about the main effect of the matching factor, while counter-matching brings the full cohort "marginal" information for the counter-matching factor main effect into the sample. In the context of a nested case–control study, the application of the two techniques have a natural orthogonality. Matching is a natural method to incorporate information related to confounding, while counter-matching is a natural method to incorporate information related to exposure. Both methods can be used in a study by counter-matching within matching strata.

Related Designs. Two-phase exposure-stratified sampling (*see* **Case–Control Study, Two-phase**) differs from counter-matching in that case–control/exposure strata are sampled independently [8]. The design is appropriate for "large strata", that is, for grouped data with sufficient numbers of cases. A comparison of the two-phase approach and counter-matching is given in [17]. An exposure-stratified **case–cohort study** has been described [6].

References

[1] Andrieu, N., Goldstein, A.M., Thomas, D.C. & Langholz, B. (2000). Counter-matching in gene-environment interaction studies: efficiency and feasibility, *American Journal of Epidemiology* **153**, 265–274.

[2] Arratia, R., Goldstein, L. & Langholz, B. (2005). Local central limit theorems, the high order correlations of rejective sampling, and logistic likelihood asymptotics, *Annals of Statistics*; to appear.

[3] Bernstein, J.L., Langholz, B., Haile, R.W., Bernstein, L., Thomas, D.C., Stovall, M., Malone, K.E., Lynch, C.F., Olsen, J.H., Anton-Culver, H., Shore, R.E., Boice J.D., Jr., Berkowitz, G.S., Gatti, R.A., Teitelbaum, S.L., Smith, S.A., Rosenstein, B.S., Børresen-Dale, A.-L., Concannon, P. & Thompson, W.D. (2004). Study design: evaluating gene-environment interactions

in the etiology of breast cancer - the WECARE study, *Breast Cancer Research* **6**, R199–R214 .

[4] Borgan, Ø., Goldstein, L. & Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model, *Annals of Statistics* **23**, 1749–1778.

[5] Borgan, O. & Langholz, B. (1997). Estimation of excess risk from case-control data using Aalen's linear regression model, *Biometrics* **53**, 690–697.

[6] Borgan, O., Langholz, B., Samuelsen, S.O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs, *Lifetime Data Analysis* **6**, 39–58.

[7] Borgan, O. & Olsen, E.F. (1999). The efficiency of simple and counter-matched nested case-control sampling, *Scandinavian Journal of Statistics* **26**, 493–509.

[8] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two stage case-control data, *Biometrika* **75**, 11–20.

[9] Cologne, J. (1997). Counterintuitive matching, *Epidemiology* **8**, 227–229.

[10] Cologne, J. & Langholz, B. (2003). Selecting controls for assessing interaction in nested case-control studies, *Journal of Epidemiology* **13**, 193–202.

[11] Cologne, J.B., Sharp, G.B., Neriishi, K., Verkasalo, P.K., Land, C.E., & Nakachi, K. (2004). Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure, *International Journal of Epidemiology* **33**, 485–492.

[12] Gilliland, F.D., Li, Y.F. & Peters, J.M. (2001). Effects of maternal smoking during pregnancy and environmental tobacco smoke on asthma and wheezing in children, *American Journal of Respiratory and Critical Care Medicine* **163**, 429–936.

[13] Langholz, B. & Borgan, O. (1995). Counter-matching: a stratified nested case-control sampling method, *Biometrika* **82**, 69–79.

[14] Langholz, B. & Borgan, O. (1997). Estimation of absolute risk from nested case-control data, *Biometrics* **53**, 767–774.

[15] Langholz, B. & Clayton, D. (1994). Sampling strategies in nested case-control studies, *Environmental Health Perspectives* **102**(Suppl. 8), 47–51.

[16] Langholz, B. & Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies, *Statistical Science* **11**, 35–53.

[17] Langholz, B. & Goldstein, L. (2001). Conditional logistic analysis of case-control studies with complex sampling, *Biostatistics* **2**, 63–84.

[18] Peters, J.M., Avol, E., Navidi, W., London, S.J., Gauderman, W.J., Lurmann, F., Linn, W.S., Margolis, H., Rappaport, E., Gong, H. & Thomas, D.C. (1999). A study of twelve Southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity, *American Journal of Respiratory and Critical Care Medicine* **159**(3), 760–767.

[19] Sharp, G.B., Neriishi, K., Hakoda, M., Suzuki, G., Akahoshi, M., Cologne, J.B., Imai, K., Eguchi, H., Nakachi, K., Key, T.J., Stevens, R.G., Kabuto, M. & Land, C.E. A Nested Case-control Study of Breast and Endometrial Cancer in the Cohort of Japanese Atomic Bomb Survivors. Research Protocol RP-6-02, RERF, 2002.

[20] Steenland, K. & Deddens, J.A. (1997). Increased precision using counter-matching in nested case-control studies, *Epidemiology* **8**, 238–242.

[21] Tokunaga, M., Land, C.E., Tokuoka, S., Nishimori, I., Soda, M. & Akiba, S. (1994). Incidence of female breast cancer among atomic bomb survivors, 1950–1985, *Radiation Research* **138**, 209–223.

BRYAN LANGHOLZ

Counting Process Methods in Survival Analysis

Event history analysis or generalized survival analysis finds applications in **actuarial** science, **demography**, epidemiology, medical research, and many other fields. This theory studies a collection of individuals, each moving between a finite (usually small) number of states. The exact transition times in continuous time form the modeling basis of the phenomena, although often these times are only incompletely observed. This article describes how counting processes provide a useful mathematical framework when studying event history data.

In survival analysis, often a model is needed when studying the occurrence of a recurrent phenomenon or events of different types. Such models can be studied within the framework of counting processes. A counting process $N(t)$ can be thought of as counting observed events up to time t.

The simplest and most important model for event history data is the following model of survival data. More complicated event history data include, among others, data on **competing risks**, the so-called multistate survival data, and the data that may be modeled by illness–death process or disability–death processes (*see* **Stochastic Processes**).

Suppose that a group of n patients is followed at some hospital from the time of diagnosis of a certain disease to the time of death or to the date last known in follow-up. We note that these are often patients who are alive at the time of data analysis. For the ith patient, we observe a disease duration $\overline{T}_i$, which is either his true survival time T_i, that is, the length of time from diagnosis to death, or a censoring time, that is, the length of time from diagnosis to the date last known in follow-up (*see* **Censored Data**). Let $D_i = 1$ if $\overline{T}_i$ is a true survival time and $D_i = 0$ otherwise. We assume that the pairs (T_i, D_i) are independent for $i = 1, \ldots, n$.

Let

$$N_i(t) = I(\overline{T}_i \le t, D_i = 1), \tag{1}$$

where $I(\cdot)$ is the indicator function. Thus, N_i is 0 before $\overline{T}_i$ and jumps to 1 at $\overline{T}_i$ if and only if $\overline{T}_i$ is a true survival time. At any time t, we know that the ith patient either has been observed to die, or has been censored because of incomplete follow-up, or is still alive and at risk. For the first two cases, the conditional probability of observing N_i to jump in a small interval near t is 0. For the latter, this conditional probability is near $\alpha_i(t)\mathrm{d}t$, where $\alpha_i(\cdot)$ is the hazard function of T_i (*see* **Survival Distributions and Their Characteristics**). Let

$$Y_i(t) = I(\overline{T}_i \ge t), \tag{2}$$

which indicates whether the individual is still alive just before time t. The previous remarks indicate that

$$Pr[\mathrm{d}N_i(t) = 1 \mid \mathcal{F}_{t-}] = \alpha_i(t)Y_i(t)\,\mathrm{d}t. \tag{3}$$

Here $\mathrm{d}N_i(t)$ is the increment of N_i in a small interval near t, and $\mathcal{F}_{t-}(\mathcal{F}_t)$ represents all the information available on the course of the disease just before (up to) time t. $\mathcal{F}_t$ is called a *filtration*.

We note that both the deterministic function α_i and the random process Y_i are predictable processes in the sense that their values at any time t are known just before t. If we define **stochastic processes** M_i by having increments

$$\mathrm{d}M_i(t) = \mathrm{d}N_i(t) - \alpha_i(t)Y_i(t)\,\mathrm{d}t, \tag{4}$$

then (3) says

$$\mathrm{E}[\mathrm{d}M_i(t) \mid \mathcal{F}_{t-}] = 0, \tag{5}$$

which means

$$M_i(t) = N_i(t) - \int_0^t \alpha_i(s)Y_i(s)\,\mathrm{d}s \tag{6}$$

is a *martingale* (see, for example, [1, 4, 26]). In particular, $\mathrm{E}M_i(t) = 0$ for all t. Here $\lambda_i(t) \equiv \alpha_i(t)Y_i(t)$ is called the *intensity process* of N_i. More generally, under some regularity conditions, a counting process N has a predictable process λ such that

$$M(t) \equiv N(t) - \int_0^t \lambda(s)\,\mathrm{d}s \tag{7}$$

is a martingale. λ is called the *intensity process* for N, and

$$\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s \tag{8}$$

the *cumulative intensity process* because of $\lambda(t+) = \lim_{\triangle t \to 0}(1/\triangle t)P\{N(t + \triangle t) - N(t) = 1 \mid \mathcal{F}_t\}$.

An important example arises when the n subjects in (1) are independent and identically distributed

(i.i.d.). In this case, let $N_\cdot = \sum_{i=1}^{n} N_i$, which is a univariate counting process that counts the number of observed deaths. Its intensity process is $\lambda_\cdot(t) = \alpha(t)Y_\cdot(t)$, with $\alpha(t)$ being the hazard function of T_i, and $Y_\cdot(t) = \sum_{i=1}^{n} Y_i(t)$ being the number of individuals observed to be at risk just before time t. We note that $Y_\cdot(t)$ increases as the number of subjects increases.

The relation (7) is the key to the counting process approach to event history analysis. We will see below that many important estimators and test statistics in survival analysis can be expressed as, or approximated by, stochastic integrals with respect to the martingales (7). This, together with martingale theory, forms the basis to the analysis of these statistics. Properties of these statistics such as unbiasedness and estimators of variability are obtained by applying results on stochastic integration with respect to the basic martingales like (7). Asymptotic statistical theory follows from martingale central limit theorems. Another important ingredient of the counting process methods in survival analysis is the fact that likelihoods are product-integrals of conditional terms for infinitesimal time intervals (see, for example, [6]).

We now formalize the above discussion by presenting some relevant martingale theory. A *multivariate counting process* $N = \{[N_1(t), \ldots, N_k(t)], \ t \in [0, 1]\}$ is a k-dimensional stochastic process with components N_h whose sample functions are nondecreasing, right-continuous step functions, $N_h(0) = 0$, and with jumps of unit size. Moreover, it is assumed that, with probability 1, no two components jump simultaneously, and that each $N_h(1)$ is almost surely finite.

A stochastic process M adapted to a filtration $\mathcal{F}_t$, satisfying $M(0) = 0$, $\mathrm{E}|M|(t) < \infty$ for $t \in [0,1]$, and having sample functions, which are right continuous with left-hand limits, is called a *martingale* if $\mathrm{E}[M(t) \mid \mathcal{F}_s] = M(s)$ a.s. for $s \leq t$ [cf. (4)]. A martingale is square-integrable if $\sup_{t\in[0,1]} \mathrm{E}M^2(t) < \infty$. A stochastic process $M(t)$ is called a local martingale if $M(t \wedge T_n)$ is a martingale for some sequence of stopping times T_n tending to infinity.

Note that if a process is adapted and has left-continuous sample paths, then it is predictable and locally bounded. Moreover, any Borel measurable deterministic function is predictable.

A process X has a *compensator* Λ if $X - \Lambda$ is a local martingale, and Λ is predictable and has paths of locally bounded variation. According to the Doob–Meyer decomposition, each component N_h of a multivariate counting process has a unique compensator Λ_h. Hence

$$M_h(t) = N_h(t) - \Lambda_h(t) \tag{9}$$

is a local martingale.

Let H_h be a predictable and locally bounded process and M_h a local martingale. We define a new process $\widetilde{M}_h$ by the stochastic integral

$$\widetilde{M}_h(t) = \int_0^t H_h(s)\,\mathrm{d}M_h(s). \tag{10}$$

Then $\widetilde{M}_h$ is a local martingale itself, because the increment $\mathrm{d}\widetilde{M}_h(t) = H_h(t)\mathrm{d}M_h(t)$ has zero conditional expectation:

$$\mathrm{E}[H_h(t)\,\mathrm{d}M_h(t) \mid \mathcal{F}_{t-}] = H_h(t)\mathrm{E}[\mathrm{d}M_h(t) \mid \mathcal{F}_{t-}] = 0 \tag{11}$$

Here the first equality is due to the predictability of H_h. The variance of $\widetilde{M}_h(t)$ equals $\mathrm{E}\int_0^t H_h^2(s)\,\mathrm{d}\Lambda_h(s)$.

Counting Process Models

Statistical models based on counting processes are specified according to their intensity processes. Here are some important models for which the previous mathematical framework is the most useful. From the point of view of statistical modeling, there are in the realm of multistate models. Important recent reviews of multistate models include [8, 31] (*see* **Event History Analysis**).

Multiplicative Intensity Models

If the intensity $\lambda_h(t)$ in (8) can be written as

$$\lambda_h(t) = \alpha_h(t, \theta)Y_h(t) \tag{12}$$

with an unknown nonnegative deterministic function $\alpha_h(\cdot, \theta)$, where θ is a parameter, and a nonnegative observable predictable process Y_h, then we say the multivariate counting process N is a multiplicative intensity model [2]. We note that θ may be infinitely dimensional.

Equation (3) shows that censored survival data is a multiplicative intensity model.

Markov Process Models

Another important example of a multiplicative intensity model is provided by the **Markov processes** with finite state space. Let $\{X(t) \mid t \in (0, 1]\}$ be a Markov process with finite state space and right-continuous sample paths. Let $N_{hj}(t)$ be the number of direct transitions for X from state h to state j, $h \neq j$, in $[0, t]$. Then the counting process $N = [N_{hj}(\cdot) \mid h \neq j]$ and $X(0)$ are equivalent to X in the sense that observation of $X(u)$ for $0 \leq u \leq t$ gives the same data as observing $X(0)$ and N on $[0, t]$. Assuming that locally integrable transition intensities $\alpha_{hj}(t, \theta)$ from state h to state j, $h \neq j$, exist, then the intensity process for N with respect to the filtration $\mathcal{F}_t \equiv \sigma[X(0), N(s) \mid s \leq t]$ is

$$\alpha_{hj}(t, \theta) Y_h(t),$$

where $Y_h(t) = 1_{(X(t-)=h)}$ is the indicator for X being in the state h just before time t (cf. [6, pp. 92–94, 126]).

A special case of the Markov process example is the **competing risks** model, obtained by considering one transient state 0 (alive) and absorbing states $h = 1, \ldots, k$. State h corresponds to "dead by cause h". Equivalently, let $X_i = (X_{i1}, \ldots, X_{ik})$ consist of k independent random variables with respective hazard functions $\alpha_{01}(t, \theta), \ldots, \alpha_{0k}(t, \theta)$, the multivariate counting process $N(t) = (N_1(t), \ldots, N_k(t))$ with $N_h(t) = \sum_{i=1}^{n} I(\min_l X_{il} = X_{ih} \leq t)$ is of main interest in the competing risks model. Here $X_1, \ldots, X_n$ are assumed independent and thus, $N_h(t)$ has intensity process $\alpha_{0h}(t, \theta) \sum_{i=1}^{n} I_{[\min_l X_{il} \geq t]}$ (cf. [6, p. 127]).

Illness–death Model

A more detailed event history analysis may be performed when individuals switch between the states "healthy" and "diseased" before the absorbing state "death". The model is known as the illness–death model; the illness may be recurrent or not.

For the case where there is no recovery, let states 0, 1, and 2 denote healthy, diseased, and dead, respectively, and define the counting processes of transitions between these states by $N(t) = [N_{01}(t), N_{02}(t), N_{12}(t)]$. $N_{0h}(t)$, $h = 1, 2$, has intensity process $\alpha_{0h}(t) Y_0(t)$ with $Y_0(t) = 1 - N_{01}(t-) - N_{02}(t-)$, which indicates that the individual is in state 0 at time $t-$, whereas $N_{12}(t)$ has intensity process $\alpha_{12}(t, d) Y_1(t)$ with $Y_1(t) = N_{01}(t-) - N_{12}(t-)$, indicating that the individual is in state 1 at time $t-$.

If the intensity of dying while diseased, denoted by $\alpha_{12}(t, d)$, only depends on time t, the process corresponds to a Markov illness–death process; when $\alpha_{12}(t, d)$ only depends on d, the duration in the disease states, one has a special case of a **semi-Markov process** (*see* **Event History Analysis**).

Regression Models

Let (X_i, Z_i), $i = 1, \ldots, n$, be random variables with nonnegative X_i denoting the survival time of the ith subject and $Z_i \equiv (Z_{i1}, \ldots, Z_{ip})'$ be p-dimensional random vectors denoting the covariates. We assume that $X_1, \ldots, X_n$ are conditionally independent given $Z = (Z_1, \ldots, Z_n)$.

In the **Cox regression model** for survival data [22], the conditional hazard of X_i given the covariates $Z = z = (z_1, \ldots, z_n)$ has the form

$$\alpha_i(t, \theta) = \alpha_0(t, \gamma) \exp(\beta' z_i), \tag{13}$$

where $\theta = (\gamma, \beta)$, and $\alpha_0(\cdot, \gamma)$ is a nonnegative deterministic function depending on a parameter γ, which can be infinite dimensional or finite dimensional. $\beta \in \mathcal{R}^p$ is called the relative risk coefficient and β' is its transpose.

Let $\widetilde{X}_1, \ldots, \widetilde{X}_n$ be the observed right censored survival times corresponding to $X_1, \ldots, X_n$. Let $D_i = I(X_i = \widetilde{X}_i)$, $N_i(t) = I(\widetilde{X}_i \leq t, D_i = 1)$, $Y_i(t) = I(\widetilde{X}_i \geq t)$ be defined as in the introductory example of censored survival times. Let $\mathcal{F}_t$ be the filtration generated by Z and $[(N_1, \ldots, N_n)(s), s \leq t]$. Then the counting process $N = (N_1, \ldots, N_n)$ has the intensity $\lambda = (\lambda_1, \ldots, \lambda_n)$ with $\lambda_i(t) = \alpha_i(t, \theta) Y_i(t)$ relative to $\mathcal{F}_t$.

Extensions of the Cox regression model using counting process formulation, initially studied by Andersen and Gill [7], are discussed in (44).

In **Aalen's additive regression model** [3] the conditional hazard of X_i given $Z = z = (z_1, \ldots, z_n)$ has the form

$$\beta_0(t) + \beta(t)' z_i. \tag{14}$$

Here $\beta_0(\cdot)$ is an $\mathcal{R}$-valued function and $\beta(\cdot) = (\beta_1(\cdot), \ldots, \beta_p(\cdot))'$ is an $\mathcal{R}^p$-valued function and $\beta(t)'$ is its transpose. This model is a special case of the matrix version of the multiplicative intensity model studied in (56).

Right censoring, Left truncation, and Filtering

Right censoring, left **truncation**, and filtering are important patterns of incomplete observation that can be handled quite satisfactorily with the counting process methods. This discussion indicates situations in which the multiplicative intensity model is retained under these patterns of incomplete observation. A similar discussion can be made with the regression models mentioned above. A quite thorough discussion of these concepts in specifying statistical models can be found in [6, Chapter III]. Important works in this regard include [35–37].

Assume we have a multiplicative intensity model (12). Let $0 \leq V_h \leq U_h$ be random variables independent of N_h. Then

$$\int_0^t 1_{(V_h, U_h]}(s)\, \mathrm{d}N_h(s) \tag{15}$$

has the compensator

$$\int_0^t 1(V_h, U_h](s)\lambda_h(s)\, \mathrm{d}s = \int_0^t \alpha_h(s, \theta)Y_h(s)1_{(V_h, U_h]}(s)\, \mathrm{d}s. \tag{16}$$

Thus, if (12) holds and both (15) and $Y_h(\cdot)1_{(V_h, U_h]}(\cdot)$ are observable, then we again have a multiplicative intensity model. When $U_h = \infty$, we say (15) is the left-truncated process of N_h. When $V_h = 0$, we say (15) is the right-censored process of N_h. The independent filtering process $1_{(V_h, U_h]}(\cdot)$ is called an Aalen filter.

Nonparametric Estimation

Let $N = (N_1, \ldots, N_k)$ be a multivariate counting process with the intensity process $\lambda = (\lambda_1, \ldots, \lambda_k)$ satisfying the multiplicative intensity model $\lambda_h(t) = \alpha_h(t)Y_h(t)$, where $\alpha_h(\cdot)$ is a nonnegative deterministic function, and $Y_h(t)$ is a predictable and observable process. Often α_h is a force of transition, whereas Y_h counts the number at risk.

Nelson–Aalen Estimator

An important statistical problem is to estimate the cumulative intensity

$$A_h(t) = \int_0^t \alpha_h(s)\, \mathrm{d}s \tag{17}$$

based on the data $[N_h(t), Y_h(t) \mid 0 \leq t \leq 1, h = 1, \ldots, k]$. To derive estimators for (17), we use (6) to write symbolically "$\mathrm{d}N_h(t) = \alpha_h(t)Y_h(t) + \text{noise}$". With this in mind, we let $J_h(t) = I[Y_h(t) > 0]$, and define the estimator

$$\widehat{A}_h(t) = \int_0^t \left[\frac{J_h(s)}{Y_h(s)}\right] \mathrm{d}N_h(s), \tag{18}$$

where $J_h(t)/Y_h(t)$ is interpreted as 0 whenever $Y_h(t) = 0$. Equation (18) is called the **Nelson–Aalen estimator** [1, 2, 45]; (*see* **Event History Analysis**).

Let $T_{h1} < T_{h2} < \cdots$ be the successive jump times for N_h. Then

$$\widehat{A}_h(t) = \sum_{(j:\, T_{hj} \leq t)} [Y_h(T_{hj})]^{-1}. \tag{19}$$

Thus, $\widehat{A}_h$ is an increasing, right-continuous step function with increment $1/Y_h(T_{hj})$ at the observed jump time T_{hj} of N_h.

Suitably normalized, $\widehat{A}_h(t)$ is asymptotically normally distributed with mean $A_h(t)$ and a variance which may be estimated by $\int_0^t J_h(s)[Y_h(s)]^{-2}\, \mathrm{d}N_h(s)$. Properties of the **Kaplan–Meier estimator** $\widehat{S}_h(t)$ can be obtained from that of $\widehat{A}_h(t)$. Note that $\widehat{S}_h(t) = \prod_{0<s<t}[1 - \mathrm{d}\widehat{A}_h(s)]$, which is the **product integral** of the Nelson–Aalen estimator $\widehat{A}_h$.

Kernel Function Smoothing

Another important problem is to estimate the intensity $\alpha_h(t)$. Ramlau–Hansen [53] proposed the following kernel smoothing estimator:

$$\widehat{\alpha}_h(t) = \frac{1}{b}\int_0^1 K\left(\frac{t-s}{b}\right) \mathrm{d}\widehat{A}_h(s), \tag{20}$$

as estimators for $\alpha_h(t)$, for $h = 1, \ldots, k$. Here $\widehat{A}_h(\cdot)$ is defined in (18). The kernel function K is a bounded nonnegative function that is 0 outside $[-1, 1]$ and has integral 1. The window (bandwidth) b is a positive number. The kernel function and the window have to be chosen in concrete applications and may depend on h. We note that (20) is equal to

$$\widehat{\alpha}_h(t) = \frac{1}{b}\sum_{T_{hj}} K\left(\frac{t - T_{hj}}{b}\right)\frac{1}{Y_h(T_{hj})}. \tag{21}$$

Note that only values of T_{hj} satisfying $t - b \leq T_{hj} \leq t + b$ contribute to this sum.

If Y_h increase uniformly in a neighborhood of t, and at the same time the window b tends to 0, then, subject to some regularity conditions, $\widehat{\alpha}_h(t)$ is asymptotically normally distributed with mean $\alpha_h(t)$ and a variance that may be estimated by

$$\frac{1}{b^2}\int_0^1 K^2\left(\frac{t-s}{b}\right)\frac{J_h(s)}{Y_h^2(s)}\,\mathrm{d}N_h(s).$$

To apply the kernel smoothing estimator (20), one has to decide upon a choice of the kernel function K and the window b. Some guidelines to the choice of K are given in [54], and the choice of the window b was discussed in [6, Chapter IV 2.2] (*see* **Smoothing Hazard Rates**).

Nonparametric Hypothesis Testing

One-sample Tests

Consider a univariate counting process $N(t)$, with intensity process $\alpha(t)Y(t)$, where $\alpha(\cdot)$ is a nonnegative deterministic function and $Y(\cdot)$ a nonnegative predictable and observable process. The null hypothesis $\alpha = \alpha_0$ is to be tested, where α_0 is a known intensity function. For example, in a mortality study for a certain population, one may like to know if it equals the general population mortality.

The idea behind the test statistic to be proposed comes from the properties of the Nelson–Aalen estimator. The key is to compare the increments of the Nelson–Aalen estimator $\mathrm{d}\widehat{A}(t)$ with $\alpha_0(t)\mathrm{d}t$ using the weight process $K(t)$ (cf. [5]). Formally, let

$$\widehat{A}(t) = \int_0^t \left[\frac{J(s)}{Y(s)}\right]\mathrm{d}N(s), \tag{22}$$

$$A_0^*(t) = \int_0^t \alpha_0(s)J(s)\,\mathrm{d}s, \tag{23}$$

where $J(s) = I(Y(s) > 0)$.

Then, under the **null hypothesis**, $\widehat{A} - A_0^*$ is a local square-integrable martingale. Its expected variation is $\mathrm{E}(\widehat{A}(t) - A_0^*(t))^2 = \mathrm{E}\int_0^t J(s)\alpha_0(s)/Y(s)\,\mathrm{d}s$. Our general test statistic is based on the following stochastic process:

$$Z(t) = \int_0^t K(s)\,\mathrm{d}[\widehat{A}(s) - A_0^*(s)], \tag{24}$$

where K is a locally bounded predictable nonnegative stochastic process. It is assumed throughout that $K(s) = 0$ whenever $Y(s) = 0$. It follows immediately from the definition that, under the null hypothesis, Z is a local square-integrable martingale with $\mathrm{E}(Z(t))^2 = \mathrm{E}\int_0^t K^2(s)\alpha_0(s)/Y(s)\,\mathrm{d}s = \mathrm{E}\int_0^t K^2(s)Y^{-2}(s)\,\mathrm{d}N(s)$. The hypothesis $\alpha = \alpha_0$ may now be tested using the standardized test statistic based on $Z(t)$, which can be shown to have an approximate standard normal distribution. Usually, t is chosen as some "large" time. By choosing a different weight process K, one can obtain a number of test statistics of one-sample tests. For example, the choice $K = Y$ corresponds to the one-sample **logrank test** [16] (*see* **Linear Rank Tests in Survival Analysis**).

k-Sample Tests

One of the most commonly encountered problems in clinical trails is the comparison of treatments on survivals. This is a special case of the following k-sample problem.

Consider a k-variate counting process N satisfying the multiplicative intensity model (12). We want to derive a test for the hypothesis

$$H_0\colon \alpha_1 = \alpha_2 = \cdots = \alpha_k. \tag{25}$$

The common value of the α_hs will be denoted by α.

The idea is to construct a test statistic by comparing the Nelson–Aalen estimators $\widehat{A}_h(t)$ [cf. (18)] with an estimator of the hypothesized common value

$$A(t) = \int_0^t \alpha(s)\,\mathrm{d}s. \tag{26}$$

This latter quantity can be estimated by

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y.(s)}\,\mathrm{d}N.(s), \tag{27}$$

where

$$N. = \sum_{h=1}^k N_h, \quad Y. = \sum_{h=1}^k Y_h, \tag{28}$$

and

$$J(t) = I[Y.(t) > 0]. \tag{29}$$

Under the hypothesis, we know that $N.(t)$ is a (univariate) counting process with the intensity process $\alpha(t)Y.(t)$. Let

$$\overline{A}_h(t) = \int_0^t J_h(s)\,\mathrm{d}\widehat{A}(s) = \int_0^t \frac{J_h(s)}{Y.(s)}\,\mathrm{d}N.(s),$$

and note that, when (25) holds true, we have

$$\widehat{A}_h(t) - \overline{A}_h(t) = \int_0^t \frac{J_h(s)}{Y_h(s)}\,\mathrm{d}M_h(s) - \int_0^t \frac{J_h(s)}{Y_\cdot(s)}\,\mathrm{d}M_\cdot(s), \quad (30)$$

where

$$M_\cdot = \sum_{h=1}^{k} M_h.$$

Thus, except for random variations, $\widehat{A}_h$ and $\overline{A}_h$ are equal under the hypothesis. Let K_h be a nonnegative locally bounded predictable weight process, and

$$Z_h(t) = \int_0^t K_h(s)\,\mathrm{d}(\widehat{A}_h - \overline{A}_h)(s). \quad (31)$$

When (25) holds true, (30) shows that the Z_hs are linear combinations of stochastic integrals and hence $\mathrm{E}Z_h(t) = 0$ for all h and $t \in [0, 1]$.

It turns out that the special choice of weight processes,

$$K_h(t) = Y_h(t)L(t), \quad (32)$$

where L is a locally bounded predictable process that only depends on $(N_\cdot, Y_\cdot)$, covers most relevant examples. Under (32), we have

$$Z_h(t) = \int_0^t L(s)\,\mathrm{d}N_h(s) - \int_0^t L(s)\frac{Y_h(s)}{Y_\cdot(s)}\,\mathrm{d}N_\cdot(s). \quad (33)$$

We note that

$$\sum_{h=1}^{k} Z_h = 0. \quad (34)$$

It follows from the martingale central limit theorem that, under the hypothesis H_0, the Z_hs, properly normalized, converge weakly to a k-variate Gaussian martingale, as the Y_hs increase. In particular, $\mathbf{Z}(1) = \{Z_1(1), \ldots, Z_k(1)\}'$ is asymptotically multinormally distributed (*see* **Multivariate Normal Distribution**) with mean 0 and a (singular) **covariance matrix** that can be estimated by $\mathbf{V}(1) = \{V_{hj}(1)\}$, where

$$V_{hj}(1) = \int_0^1 L^2(s)\frac{Y_h(s)}{Y_\cdot(s)}\left(\delta_{hj} - \frac{Y_j(s)}{Y_\cdot(s)}\right)\mathrm{d}N_\cdot(s), \quad (35)$$

and δ_{hj} is the Kronecker delta [5, Theorem 3.1].

Thus, under the hypothesis (25), the statistic

$$\chi^2 = \mathbf{Z}(1)'\mathbf{V}(1)^-\mathbf{Z}(1), \quad (36)$$

is asymptotically **chi-square distributed** with $k - 1$ **degrees of freedom**, where $\mathbf{V}(1)^-$ is a generalized inverse [5, Section 9.1] (*see* **Matrix Algebra**). Note that we may denote by $\mathbf{Z}_0(1)$ and $\mathbf{V}_0(1)$, respectively, the vector and matrix obtained from $\mathbf{Z}(1)$ and $\mathbf{V}(1)$ by deleting the last component of $\mathbf{Z}(1)$ and the last row and column of $\mathbf{V}(1)$, and then using the relation $\mathbf{Z}(1)'\mathbf{V}(1)^-\mathbf{Z}(1) = \mathbf{Z}_0(1)'\mathbf{V}_0(1)^{-1}\mathbf{Z}_0(1)$ for (36).

Equation (36) covers not only many classical nonparametric tests but their generalizations to censored data as well. For example, the choice $L(t) = I[Y_\cdot(t) > 0]$ corresponds to the logrank (or Savage) test [48], while $L(t) = Y_\cdot(t)$ gives a generalization of the Kruskal–Wallis test [15]. Also the tests suggested by Tarone and Ware [61], Prentice [49], and Harrington and Fleming [30] are special cases of (36). More specifically, Harrington and Fleming [30] introduced a class of test statistics for censored survival data by letting $L(t) = [\widehat{S}(t-)]^\rho I[Y(t) > 0]$, where

$$\widehat{S}(t) = \prod_{s \le t}\left(1 - \frac{\Delta N_\cdot(s)}{Y_\cdot(s)}\right) \quad (37)$$

is the Kaplan–Meier estimator based on the combined sample and $0 \le \rho \le 1$. It is seen that $\rho = 0$ gives the logrank test, whereas $\rho = 1$ gives a test similar to Peto and Peto's and Prentice's generalization of the **Wilcoxon–Mann–Whitney** and Kruskal–Wallis tests.

Parametric Models

Let $N = (N_1, \ldots, N_k)$ be a multivariate counting process satisfying the multiplicative intensity model (12) with parametric α_hs, that is, the intensity process is given by

$$\lambda_h(t) = \alpha_h(t; \theta_0)Y_h(t), \quad (38)$$

where $\theta_0 = (\theta_{10}, \ldots, \theta_{q0})'$ is a q-dimensional parameter belonging to some open subset Θ of $\mathcal{R}^q$, and α_h are known functions. Under some regularity conditions, the log-**likelihood** function now takes

the form

$$l(\theta) = \sum_{h=1}^{k} \int_0^1 \log\{\alpha_h(s;\theta)\}\, \mathrm{d}N_h(s) - \sum_{h=1}^{k} \int_0^1 \alpha_h(s;\theta) Y_h(s)\, \mathrm{d}s, \qquad (39)$$

and the **maximum likelihood** estimator $\widehat{\theta}$ is defined as a solution to the set of equations

$$\sum_{h=1}^{k} \int_0^1 \frac{(\partial/\partial\theta_j)\alpha_h(s;\theta)}{\alpha_h(s;\theta)}\, \mathrm{d}N_h(s) - \sum_{h=1}^{k} \int_0^1 \frac{\partial}{\partial\theta_j}\alpha_h(s;\theta) Y_h(s)\, \mathrm{d}s = 0, \qquad (40)$$

where $j = 1, \ldots, q$. Under certain regularity conditions of the $\alpha_h s$, the likelihood equations (40) have, with probability tending to 1, exactly one consistent solution $\widehat{\theta}$ as the $Y_h s$ increase. Moreover, $\widehat{\theta}$ is asymptotically multinormally distributed with mean θ_0 and a covariance matrix that may be estimated by $-I(\widehat{\theta})^{-1}$, where $I(\theta) = \partial^2 l(\theta)/\partial\theta^2$ [14, Theorems 1 and 2]. Thus, the usual results for maximum likelihood estimation in the classical i.i.d. case continue to hold under the more general model (38).

With (6) and (38), we know that the left-hand side of (40), evaluated at the true parameter value θ_0, equals the stochastic integral:

$$\sum_{h=1}^{k} \int_0^1 \frac{(\partial/\partial\theta_j)\alpha_h(s;\theta_0)}{\alpha_h(s;\theta_0)}\, \mathrm{d}M_h(s). \qquad (41)$$

This makes it possible to use properties of martingales to derive the above-mentioned asymptotic results.

As in the i.i.d. case in which minus twice the logarithm of the **likelihood ratio test** statistic is asymptotically chi-square distributed, this is true also for the more general model (38). Other closely related test statistics, like Wald's test and the score test (*see* **Likelihood**), also have the desired properties as their counterparts in the i.i.d. case [14] (*see* **Parametric Models in Survival Analysis**).

Regression Models

Regression models are useful when it is desired to assess the effect of risk factors (**prognostic factors**) on survival. We discuss here some commonly used regression models in survival analysis where counting process and martingale techniques have played a central role. Readers are referred to [42] for general nonparametric models, to [67] for **marginal models**, and to [38, 55], and [63] for other **semiparametric regression** models.

The Cox Regression Model

Consider the multivariate counting process $[N_{hi}(t), h = 1, \ldots, k; i = 1, \ldots, n], t \in [0, 1]$, where $N_{hi}(t)$ counts the number of type h events in $[0, t]$ for individual i.

We assume that $N_{hi}(t)$ has an intensity process of the form

$$\lambda_{hi}(t) = \alpha_{0h}(t) \exp[\beta_0' Z_{hi}(t)] Y_{hi}(t). \qquad (42)$$

Here α_{oh} is an unspecified type-specific baseline whose integral

$$A_{0h}(t) = \int_0^t \alpha_{0h}(s)\, \mathrm{d}s \qquad (43)$$

satisfies $A_{0h}(1) < \infty$. Furthermore, $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})'$ is a vector of unknown regression coefficients, $Y_{hi}(t)$ is a predictable indicator process and $Z_{hi}(t) = [Z_{hi1}(t), \ldots, Z_{hip}(t)]'$ a vector of predictable and locally bounded (type-specific) observable time-dependent covariate processes.

The relative risk function $\exp[\beta_0' Z_{hi}(t)]$ can be replaced by a general form of function (cf. [51]). Also, the baseline function $\alpha_{0h}(t)$ can be replaced by various forms of random processes [19, 52].

The basic assumption in the extended Cox model (42) is that each covariate $Z_{hij}(t)$ has a multiplicative effect on the intensity [7, 23]; in particular, for time-independent covariates, we have a model with proportional intensities.

The estimator $\widehat{\beta}$ of β_0 is defined to be the solution to the equations $U(\beta, 1) = 0$, which is called the Cox's partial score and defined below (cf. [7]). The key to the derivation of the statistical properties of $\widehat{\beta}$ is to notice that the Cox's partial score $U(\beta_0, \cdot)$, evaluated at the true value β_0, is a local square integrable martingale.

Let

$$S_h^{(0)}(\beta, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{hi}(t) \exp[\beta' Z_{hi}(t)],$$

$$S_{hj}^{(1)}(\beta, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{hi}(t) Z_{hij}(t) \exp[\beta' Z_{hi}(t)], \qquad (44)$$

$$S^{(2)}_{hjl}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} Y_{hi}(t) Z_{hij}(t) Z_{hil}(t) \exp[\beta' Z_{hi}(t)],$$

and

$$\mathrm{E}_{hj}(\beta, t) = \frac{S^{(1)}_{hj}(\beta, t)}{S^{(0)}_{h}(\beta, t)},$$

where $h = 1, \ldots, k$ and $j, l = 1, \ldots, p$. Then the jth component of $U(\beta, t)$ is given as

$$U_j(\beta, t) = \sum_{h=1}^{k}\left[\int_0^t \sum_{i=1}^{n} Z_{hij}(s)\, \mathrm{d}N_{hi}(s) - \int_0^t \mathrm{E}_{hj}(\beta, s)\, \mathrm{d}N_h(s)\right], \qquad (45)$$

and using (6) and (42) we see that

$$U_j(\beta_0, t) = \sum_{h=1}^{k}\sum_{i=1}^{n}\int_0^t [Z_{hij}(s) - \mathrm{E}_{hj}(\beta_0, s)]\, \mathrm{d}M_{hi}(s) \qquad (46)$$

are linear combinations of stochastic integrals. Here $N_h = \sum_{i=1}^{n} N_{hi}$. Thus, with some regularity conditions, the martingale central limit theorem can be applied to prove that the process $n^{-l/2}U(\beta_0, \cdot)$, as n tends to ∞, is asymptotically distributed as a mean zero Gaussian martingale.

By a Taylor expansion technique, this result can be transformed into a theorem concerning the asymptotic distribution of $\widehat{\beta}$, much in the same way as for standard maximum likelihood estimation. Under certain regularity conditions, it can be shown that $n^{l/2}(\widehat{\beta} - \beta_0)$ is asymptotically multinormally distributed $N_p(0, \Sigma^{-1})$, where $\Sigma = \{\sigma_{jl}\}$ is positive definite, and σ_{jl} can be estimated consistently by $-I_{jl}(\widehat{\beta})/n$. Here $I_{jl}(\beta)$ is the partial derivative of $U(\beta, 1)$ with respect to β_j, that is,

$$I_{jl}(\beta) = -\sum_{h=1}^{k}\int_0^1 \left[\frac{S^{(2)}_{hjl}(\beta, s)}{S^{(0)}_{h}(\beta, s)} - \mathrm{E}_{hl}(\beta, s)\mathrm{E}_{hj}(\beta, s)\right] \times \mathrm{d}N_h(s). \qquad (47)$$

On the basis of the above results, one can draw inference on the regression parameter β even with the presence of the nuisance functions $\alpha_{0h}(t)$ in the semiparametric model (42).

In some cases, the underlying intensities are also of interest. Under the same set of regularity conditions, the estimates for the cumulative intensities $A_{0h}(t)$:

$$\widehat{A}_{0h}(t) = \int_0^t J_h(u)[nS^{(0)}_h(\widehat{\beta}, s)]^{-1}\, \mathrm{d}N_h(s) \qquad (48)$$

with $J_h(u) = I[Y_h(u) > 0]$, is distributed asymptotically as a Gaussian process.

Notice that for a homogeneous group of individuals, that is, when all $Z_{hi} \equiv 0$, the estimator $\widehat{A}_{0h}$ reduces to the Nelson–Aalen estimator.

Parametric Regression

As in (42), we now assume that N_{hi} has the intensity process of the form

$$\lambda_{hi}(t) = \alpha_{0h}(t; \theta_0) \exp[\beta_0' Z_{hi}(t)] Y_{hi}(t), \qquad (49)$$

with unknown parameter θ_0 belonging to an open subset of $\mathcal{R}^q$, and α_{0h} being some known functions [14].

In the case of survival data, (49) covers the exponential regression where $\alpha_{0h}(t, \theta) = \theta_h$, the Weibull regression where $\alpha_{0h}(t, \theta) = \theta_h t^{\rho h}$, and the piecewise exponential where $\alpha_{0h}(t, \theta)$ is piecewise constant. References in which parametric models of the form (49) have been studied for survival data can be found in Kalbfleisch and Prentice [33], for example, (*see* **Parametric Models in Survival Analysis**).

Inferences from parametric regression models will be based on the log-likelihood function

$$l(\theta, \beta) \equiv \log L(\theta, \beta) = \sum_{h=1}^{k}\sum_{i=1}^{n}\left\{\int_0^1 [\log\alpha_{0h}(s; \theta) + \beta' Z_{hi}(s)]\, \mathrm{d}N_{hi}(s) - \int_0^1 \alpha_{0h}(s, \theta)\exp[\beta' Z_{hi}(s)] Y_{hi}(s)\, \mathrm{d}s\right\}, \qquad (50)$$

and the maximum likelihood estimators $\widehat{\theta}$ and $\widehat{\beta}$ are defined as solutions to the set of equations

$$\frac{\partial}{\partial\theta_j} l(\theta, \beta) = \sum_{h=1}^{k}\sum_{i=1}^{n}\left\{\int_0^1 \frac{(\partial/\partial\theta_j)\alpha_{0h}(s, \theta)}{\alpha_{0h}(s, \theta)}\, \mathrm{d}N_{hi}(s) - \int_0^1 \frac{\partial}{\partial\theta_j}\alpha_{0h}(s, \theta)\exp[\beta' Z_{hi}(s)] Y_{hi}(s)\, \mathrm{d}s\right\} = 0,$$

and

$$\frac{\partial}{\partial \beta_l} l(\theta, \beta) = \sum_{h=1}^{k} \sum_{i=1}^{n} \left\{ \int_0^1 Z_{hil}(s)\, \mathrm{d}N_{hi}(s) - \int_0^1 \alpha_{0h}(s, \theta) Z_{hil}(s) \exp[\beta' Z_{hi}(s)] Y_{hi}(s)\, \mathrm{d}s \right\} = 0, \tag{51}$$

for $j = 1, \ldots, q$ and $l = 1, \ldots, p$. Since the left-hand sides of the likelihood equations (51), evaluated at the true parameter values θ_0 and β_0, are linear combinations of stochastic integrals, the asymptotic properties of $\widehat{\theta}$ and $\widehat{\beta}$ can be obtained by the martingale central limit theorem.

Nonparametric Additive Hazard Models

Let $\mathbf{N}(t) = [N_i(t); i = 1, \ldots, n]$ be a multivariate counting process. Assume that the individual process N_i has an $\mathcal{F}_t$-intensity process

$$\lambda_i(t) = \alpha_i[t; Z_i(t)] Y_i(t), \tag{52}$$

with

$$\alpha_i[t; Z_i(t)] = \beta_0(t) + \beta_1(t) Z_{i1}(t) + \cdots + \beta_p(t) Z_{ip}(t). \tag{53}$$

Here $\alpha_i(\cdot, \cdot)$ is nonnegative, whereas the regression functions $\beta_j(t)$ are completely unspecified. A major problem for the **additive hazard model** of Aalen [3] is to estimate the integrated regression functions

$$B_j(t) = \int_0^t \beta_j(u)\, \mathrm{d}u, \tag{54}$$

for $j = 0, 1, \ldots, p$. Let $\beta(t) = [\beta_0(t), \beta_1(t), \ldots, \beta_p(t)]'$, $\mathbf{Y}$(t) be the $n \times (p + 1)$ matrix with the ith row, $i = 1, \ldots, n$, given by $Y_i(t)[1, Z_{il}(t), \ldots, Z_{ip}(t)]$ and $\mathbf{B}(t) = [B_0(t), B_1(t), \ldots, B_p(t)]'$. The model given by (52) and (55) can be written in matrix form as

$$\mathbf{N}(t) = \int_0^t \mathbf{Y}(u) \beta(u)\, \mathrm{d}u + \mathbf{M}(t) \tag{55}$$

and is sometimes called the matrix multiplicative intensity model, where $\mathbf{M} = (M_1, \ldots, M_n)$ is an n-vector of local square-integrable martingales. Then

$$\widehat{\mathbf{B}}(t) = \int_0^t J(u) \mathbf{Y}^-(u)\, \mathrm{d}\mathbf{N}(u) \tag{56}$$

can be viewed as a generalized Nelson–Aalen estimator for $\mathbf{B}(t)$. Here $\mathbf{Y}^-(t)$ is the predictable generalized inverse of $\mathbf{Y}(t)$, that is, a $(p + 1) \times n$ matrix satisfying $\mathbf{Y}^-(t)\mathbf{Y}(t) = I$, the $(p + 1) \times (p + 1)$ identity matrix, and $J(t) = I[\mathrm{rank}\, \mathbf{Y}(t) = p + 1]$ is the predictable indicator of $\mathbf{Y}(t)$ having full rank (assuming $p + 1 \leq n$).

To achieve some kind of optimality, we need to choose the generalized inverse $\mathbf{Y}^-(t)$ properly [32, 41].

Frailty Models

Let $N_{ik}(t)$ denote the number of certain events experienced up to time t by the kth member of the ith family in an experiment. Suppose that the hazard rate $\lambda_{ik}(t)$ of $N_{ik}(t)$ has the proportional form

$$\lambda_{ik}(t) = \Lambda_i(t) Y_{ik}(t) \exp[\beta Z_{ik}(t)] \tag{57}$$

for $k = 1, 2, \ldots, K$. Here $Y_{ik}(\cdot)$ is an observable nonnegative predictable process, $Z_{ik}(\cdot)$ is a predictable process representing an observable covariate, β is the relative risk coefficient to be estimated, and $\Lambda_i(\cdot)$ is the unknown baseline hazard rate for the ith family, termed random **frailty** and shared by the K members in the ith family.

We note that, when $\Lambda_i(t)$ is a deterministic function not varying from family to family, (57) is a special case of (42).

Nielsen et al. [46] assumed that

$$\Lambda_i(t) = \alpha_i \lambda_0(t), \tag{58}$$

where $\lambda_0(\cdot)$ is an unknown nonnegative deterministic function (parametric or nonparametric) and the $\alpha_i s$ are independent unobservable nonnegative random variables with a gamma distribution. In certain medical examples, the frailty variable α_i is thought of as a way to describe susceptibility to accident of members of the ith family, who share a common genetic background.

Assume now that $K > 1$ and $\Lambda_1(\cdot), \Lambda_2(\cdot), \ldots$ is a sequence of i.i.d. random elements valued in the

space of nonnegative measurable functions on $[0, \infty)$, not necessarily of the form (58). In some medical contexts, this means, for example, that genetic background specifies completely the baseline hazard rate function of members of a family at any age point, not just a multiplicative factor of it.

The true parameter β_0 in (57) can be estimated by $\hat{\beta}_n$, which is the root of

$$G_n(\beta, t) = \sum_{i=1}^{n} \sum_{k=1}^{K} \times \int_0^t \left\{ Z_{ik}(s) - \frac{\sum_{l=1}^{K} Y_{il}(s) \exp[\beta Z_{il}(s)] Z_{il}(s)}{\sum_{l=1}^{K} Y_{il}(s) \exp[\beta Z_{il}(s)]} \right\} \times \mathrm{d}N_{ik}(s). \quad (59)$$

Following the approach for (42), asymptotic normality for G_n and $\hat{\beta}_n$ can be obtained [20].

S-I-R Epidemic Model

A major parameter in the study of an infectious disease is the so-called basic **reproduction number**. Various models have been proposed to calculate this number. Here we give a very brief account of the counting process approach (*see* **SIR Epidemic Models**).

An individual experiences a sequence of events in case of an infectious disease. First there is the infection time t_A that one gets infected, then the time t_B that one enters the infectious period, the time t_C that one shows symptoms, the time t_E that marks the end of the infectious period. It is known that $t_A \le t_B \le t_E$, and $t_A \le t_C \le t_E$. Usually, only t_C and t_E are observable.

At time t, an individual is called a susceptible if $t < t_A$, an infective if $t_B < t < t_E$, and removed if $t > t_A$. A susceptible becomes an infective through contact with an infective.

Suppose we have a closed community in the sense that there is no immigrant nor emigrant. Suppose at time 0, there are s susceptibles and i infectives in this community.

Let $S(t)$, $I(t)$, and $R(t)$ denote respectively the number of susceptibles, infectives, and removed ones at time t. Let $N(t)$ denote the number of individuals infected during $(0, t]$. It is clear that these are counting processes with different jump times. In view of the law of mass action, one assumes $N(t)$ has intensity

$$\beta \overline{S}(t-) I(t-),$$

where $\overline{S}(t) = S(t)/S(0)$. Another convenient assumption is that $R(t)$ has intensity

$$\gamma I(t-).$$

β is called the infection rate and γ the removal rate. The basic reproduction number $\theta = \beta/\gamma$ gives the expected number of susceptibles infected by one infective. Depending on the data available, several estimation procedures have been proposed. Here we present one of them (cf. [10], Chapter 7). Suppose $n = S(t) + I(t) + R(t)$. We consider inference based on the data $\{S(0), I(0), R(0), R(\tau)\}$, where τ is the time that the epidemic ends with $I(\tau) = 0$. This is the situation that only the final state of the epidemic is observed.

Then estimation can be based on the following zero-mean martingale

$$\begin{aligned} M(t) &= \int_0^t \frac{1}{\overline{S}(u-)} [\mathrm{d}N_1(u) - \beta \overline{S}(u-) I(u-) \, \mathrm{d}u] \\ &\quad - \frac{\beta}{\gamma} \left[R(t) - R(0) - \int_0^t \gamma I(u-) \, \mathrm{d}u \right] \\ &= \frac{n}{S(0)} + \frac{n}{S(0)-1} + \cdots + \frac{n}{S(t)+1} \\ &\quad - \frac{\beta}{\gamma} [R(t) - R(0)]. \end{aligned} \quad (60)$$

Thus

$$\hat{\theta} = \frac{\left\{ \frac{n}{S(0)} + \frac{n}{S(0)-1} + \cdots + \frac{n}{S(\tau)+1} \right\}}{\{R(\tau) - R(0)\}}, \quad (61)$$

and the standard derivation is estimated by

$$se(\hat{\theta}) = \left[\frac{1}{S(0)^2} + \frac{1}{(S(0)-1)^2} + \cdots + \frac{1}{(S(\tau)+1)^2} + \frac{\hat{\theta}^2}{n} \{R(\tau)/n - R(0)/n\} \right]^{\frac{1}{2}} \Big/ \{R(\tau)/n - R(0)/n\}. \quad (62)$$

Becker and Hasofer [12] showed that if the removal process $R(t)$ is observed continuously, then both β and γ can also be estimated.

The previous discussion assumes a population of homogeneous individuals who mix uniformly. Extensions to more general models with different observables, including populations with heterogeneity between individuals, are important in health-case studies, and are discussed in [11, 17], and so on.

Some other Counting Process Models

Some Alternative Approaches

The statistical models discussed so far are models for which counting process methods are the most successful in formulating the problem, proposing statistical procedures, analyzing their distributional properties, and studying their efficiency. We now mention some examples in which counting processes provide useful modeling tools but extra work or even entirely different method is needed for the analysis of the models. More examples can be found in Kalbfleisch and Prentice [33]. In fact, some **goodness-of-fit** tests for certain counting process models were analyzed with techniques other than martingale central limit theorems (see, for example, [42] and [40]). In general, asymptotic theories, including **efficiency**, are often established using empirical process theory. (cf. [64, 65])

Correlated Gamma-Frailty Models and Cox-gene Models

To solve the problem of estimating the relative risk coefficient β_0 when $K = 1$ in (57) and (58), Nielsen et al. [46] suggested an estimator using the **EM algorithm** based on parametric and nonparametric maximum likelihood in the case where the $\alpha_i s$ are **gamma** random variables. Murphy [43, 44] established the **consistency** and asymptotic normality of the estimator in a one-sample frailty model. Parner [47] extended theory to correlated gamma-frailty model with covariates. Compared with (57) and (58), this means

$$\lambda_{ik}(t) = (\alpha_{i0} + \alpha_{ik})\lambda_0(t)Y_{ik}(t)\exp[\beta Z_{ik}(t)], \quad (63)$$

with $\alpha_{i0}, \alpha_{i1}, \ldots,$ and α_{ik} being independent, unobservable, gamma-distributed random variables with parameters of the form (γ, η), $(\gamma^*, \eta), \ldots,$ and (γ^*, η). Both Murphy [43, 44] and Parner [47] employed the empirical process approach.

Following the notation in (63), the so-called Cox-gene model means

$$\lambda_{ik}(t) = \lambda_0(t)Y_{ik}(t)\exp[\beta Z_{ik}(t) + \mu S_{ik}]. \quad (64)$$

Here for kth member in the ith human family, $N_{ik}(t) = I(\overline{T}_{ik} \le t, D_{ik} = 1)$ with $\overline{T}_{ik}$ being the observed disease duration, and $D_{ik} = 1$ if $\overline{T}_{ik}$ is the true survival time and $D_{ik} = 0$ otherwise; S_{ik} denotes the unobservable genotype at certain locus. In a sense, (64) is a discrete version of (58). Monte Carlo methods were studied by Li, Thompson and Wijsman [39] and Siegmund and McKnight [58]; and asymptotic theory was established by Chang, Hsiung, Wang, and Wen [21], using empirical process theory.

Multivariate Survival Functions

Nonparametric estimation of a multivariate distribution function under censoring has important applications in many areas. The aim is to estimate the multivariate survival function on the basis of multivariate censored data without assuming any special structure among the different time coordinates. A major difference with the one-dimension case is that in higher dimensions there are actually many nonequivalent representations of the survival function in terms of hazard, leading to many different estimators (see, for example, [24, 25, 27, 28, 50, 62]); (*see* **Multivariate Survival Analysis**.)

Semi-Markov Models

For **Semi-Markov** models, or Markov **renewal processes**, the intensity for a transition between two states depends on the time elapsed since the entry into the current state (cf. [34]). Thus, "time" starts anew at zero after each transition into a new state. Voelkel and Crowley [66] showed how, via a random time change, one may apply the counting process methods for some semi-Markov processes.

We note that for semi-Markov models the "counting process approach" cannot be used directly to study the large-sample properties of nonparametric estimators of the transition intensities. The reason for this is that there exists no filtration relative to "duration time".

For parameter models, Arjas' **"real time" approach** [9] can work. For semiparametric generalized proportional hazards models for counting processes, Chang and Hsiung [19] identify a useful filtration so that martingale theory is still useful in the derivation of large-sample properties of the efficient estimators.

Sequential Analysis of Censored Survival Data with Staggered Entry

In most clinical trials, the patients enter trial sequentially in calendar time (*see* **Sequential Analysis**), whereas the most relevant time dimension is the duration time on trial.

In most of the above discussion, we have assumed that in the statistical analysis, these duration variables have been all realigned to start at time (duration) zero, and that the counting process martingales were then defined in the duration time scale after alignment. This device is not completely satisfactory when calendar time monitoring is desired.

For sequential analysis with **staggered entry**, one wants to stop the trial according to calendar time filtration, instead of stopping the trial according to duration time filtration. If the duration time is modeled parametrically, this problem can be studied by martingale method completely [18]. If it is not parametric, one has to consider two time scales, which makes the theoretical problems complicated. For Cox method with staggered entry, Bilias, Gu, and Ying [13] provided a general asymptotic theory for both test and estimation, generalizing the theory of Sellke and Siegmund [56], Slud [59], and Gu and Lai [29]. The approach of Bilias, Gu, and Ying [13] employs empirical process theory and considers the Cox score process in two time scales.

One may specify in advance a finite number of time points $t_1, \ldots, t_n$ at which tests are to be performed, so that periodic reviews of the trial may be performed while controlling the total significance level (*see* **Interim Analysis of Censored Data**). For specific procedures, see [59, 60] and [57, Section V.6].

Conclusion

Detailed life history data may be given a thorough analysis using the methods based on counting processes. However, there are situations in which martingale theory is not sufficient and alternative techniques are necessary. Some such techniques can also be based on counting process ideas, but not in the simple form.

References

[1] Aalen, O.O. (1975). *Statistical Inference for a Family of Counting Processes*. PhD thesis, University of California, Berkeley.

[2] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics* **6**, 701–726.

[3] Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes, in *Mathematical Statistics and Probability Theory*, Springer Lecture Notes on Statistics, Vol. 2, W. Klonecki, A. Kozek. & J. Rosiński, eds. Springer-Verlag, New York, pp. 1–25.

[4] Andersen, P.K. & Borgan, O. (1985). Counting process models for life history data: a review (with discussion), *Scandinavian Journal of Statistics* **12**, 97–158.

[5] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1982). Linear non-parametric tests for comparison of counting processes, with application to censored survival data (with discussion), *International Statistical Review* **50**, 219–258; Amendment **52**, (1984). 225.

[6] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[7] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *The Annals of Statistics* **10**, 1100–1120.

[8] Andersen, P.K. & Keiding, N. (2002). Multi-state models for event history analysis, *Statistical Method in Medical Research* **11**, 91–115.

[9] Arjas, E. (1986). Stanford heart transplantation data revisited: a real time approach, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds. Wiley, New York, pp. 65–81.

[10] Becker, N.G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

[11] Becker, N.G. & Britton, T. (1999). Statistical studies of infectious disease incidence, *Journal of the Royal Statistical Society, Series B* **61**, 287–307.

[12] Becker, N.G. & Hasofer, A.M. (1997). Estimation in epidemics with incomplete observations, *Journal of the Royal Statistical Society, Series B* **59**, 415–429.

[13] Bilias, Y., Gu, M. & Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry, *The Annals of Statistics* **25**, 662–682.

[14] Borgan, O. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data, *Scandinavian Journal of Statistics* **11**, 1–16; Correction **11**, 275.

[15] Breslow, N.E. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship, *Biometrika* **57**, 579–594.

[16] Breslow, N.E. (1975). Analysis of survival data under the proportional hazards model, *International Statistical Review* **43**, 45–58.

[17] Britton, T. (1998). Estimation in multitype epidemics, *Journal of the Royal Statistical Society, Series B* **60**, 663–679.

[18] Chang, I.S. & Hsiung, C.A. (1988). Likelihood process in parametric model of censored data with staggered entry– Asymptotic properties and applications, *Journal of Multivariate Analysis* **24**, 31–45.

[19] Chang, I.S. & Hsiung, C.A. (1994). Information and asymptotic efficiency in generalized proportional hazards models for counting processes, *The Annals of Statistics* **22**, 1275–1298.

[20] Chang, I.S. & Hsiung, C.A. (1996). An efficient estimator for proportional hazards models with frailties and applications, *Scandinavian Journal of Statistics* **23**, 13–26.

[21] Chang, I.S., Hsiung, C.A., Wang, M.C. & Wen, C.C. (2004). An asymptotic theory for the nonparametric maximum likelihood estimator in the Cox-gene model, (revised for *Bernoulli*).

[22] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[23] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[24] Dabrowska, D.M. (1988). Kaplan-Meier estimate on the plane, *The Annals of Statistics* **16**, 1475–1489.

[25] Dabrowska, D.M. (1989). Kaplan-Meier estimate on the plane: weak convergence, LIL, and the bootstrap, *Journal of Multivariate Analysis* **29**, 308–325.

[26] Gill, R.D. (1984). Understanding Cox's regression model: a martingale approach, *Journal of the American Statistical Association* **79**, 441–447.

[27] Gill, R.D. (1992). Multivariate survival analysis, *Theory of Probability and its Applications* **37**, 18–31, 284–301.

[28] Gill, R.D., van der Laan, M.J. & Wellner, J.A. (1995). Inefficient estimators of the bivariate survival function for three models. *Annales De L Institut Henri Poincare-Probabilites Et Statistiques* **31**, 547–597.

[29] Gu, M.G. & Lai, T.L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials, *The Annals of Statistics* **19**, 1403–1433.

[30] Harrington, D.P. & Fleming, T.R. (1982). A class of rank test procedures for censored survival data, *Biometrika* **69**, 133–143.

[31] Hougaard, P. (1999). Multi-state models: a review, *Lifetime Data Analysis* **5**, 239–264.

[32] Huffer, F.W. & McKeague, I.W. (1991). Weighted least squares estimation for Aalen' s additive risk model, *Journal of the American Statistical Association* **86**, 114–129.

[33] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition*. Wiley, New York.

[34] Keiding, N. (1986). Statistical analysis of semi-Markov models based on the theory of counting process, in *Semi-Markov Models, Theory and Applications*, J. Janssen, ed. Plenum, New York. pp. 301–315.

[35] Keiding, N. (1992). Independent delayed entry, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Dordrecht, pp. 309–326.

[36] Keiding, N. & Gill, R.D. (1990). Random truncation models and Markov processes, *The Annals of Statistics* **18**, 582–602.

[37] Lai, T.L. & Ying, Z. (1991a). Estimating a distribution function with truncated and censored data, *The Annals of Statistics* **19**, 417–442.

[38] Lai, T.L. & Ying, Z. (1991b). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, *The Annals of Statistics* **19**, 1370–1402.

[39] Li, H., Thompson, E.A. & Wijsman, E.M. (1998). Semiparametric estimation of major gene effects for age of onset, *Genetic Epidemiology* **15**, 279–298.

[40] Lin, D.Y., Wei, L.J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale based residuals, *Biometrika* **80**, 557–572.

[41] McKeague, I.W. (1988). Asymptotic theory for weighted least squares estimators in Aalen's additive risk model, *Contemporary Mathematics* **80**, 139–152.

[42] McKeague, I.W. & Utikal, K.J. (1990). Inference for a nonlinear counting process regression model, *The Annals of Statistics* **18**, 1172–1187.

[43] Murphy, S.A. (1994). Consistency in a proportional hazards model incorporating a random effect, *The Annals of Statistics* **22**, 712–731.

[44] Murphy, S.A. (1995). Asymptotic theory for the frailty model, *The Annnals of Statistics* **23**, 182–198.

[45] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**, 945–965.

[46] Nielsen, G.G., Gill, R.D., Andersen, P.K. & Sørensen, T.I.A. (1992). counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics* **19**, 25–43.

[47] Parner, E. (1998). Asymptotic Theory for the Correlated Gamma-Frailty Model, *The Annnals of Statistics* **26**, 183–214.

[48] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series A* **135**, 185–206.

[49] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika* **65**, 167–179; Correction **70**, (1983). 304.

[50] Prentice, R.L. & Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika* **79**, 495–512.

[51] Prentice, R.L. & Self, S.G. (1983). Asymptotic distribution theory for Cox-type regressions models with general relative risk form, *The Annals of Statistics* **11**, 804–813.

[52] Prentice, R.L., Williams, B.J. & Peterson, A.V. (1981). On the regression analysis of multivariate failure time data, *Biometrika* **68**, 373–379.

[53] Ramlau-Hansen, H. (1983a). Smoothing counting process intensities by means of kernel functions, *The Annals of Statistics* **11**, 453–466.

[54] Ramlau-Hansen, H. (1983b). The choice of a kernel function in the graduation of counting process intensities, *Scandinavian Actuarial Journal* **1983**, 165–182.

[55] Ritov, Y. (1990). Estimation in a linear regression model with censored data, *The Annals of Statistics* **18**, 303–328.

[56] Sellke, T. & Siegmund, D. (1983). Sequential analysis of the proportional hazards model, *Biometrika* **70**, 315–326.

[57] Siegmund, D. (1985). *Sequential Analysis. Tests and Confidence Intervals*. Springer-Verlag, New York.

[58] Siegmund, K. & McKnight, B. (1998). Modeling hazard functions in families, *Genetic Epidemiology* **15**, 147–171.

[59] Slud, E.V. (1984). Sequential linear rank tests for two-sample censored survival data, *The Annals of Statistics* **12**, 551–571.

[60] Slud, E.V. & Wei, L.J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic, *Journal of the American Statistical Association* **77**, 862–868.

[61] Tarone, R.E. & Ware, J.H. (1977). On distribution-free tests for equality for survival distributions, *Biometrika* **64**, 156–160.

[62] Tsai, W.Y., Leurgans., S. & Crowley, J.J. (1986). Nonparametric estimation of a bivariate survival function in presence of censoring, *The Annals of Statistics* **14**, 1351–1365.

[63] Tsiatis, A.A. (1990). Estimating regression parameters using linear rank tests for censored data, *The Annals of Statistics* **18**, 354–372.

[64] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[65] van der Vaart, A.W. & Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag, New York.

[66] Voelkel, J.G. & Crowley, J.J. (1984). Nonparametric inference for a class of semi-Markov processes with censored observations, *The Annals of Statistics* **12**, 142–160.

[67] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

Bibliography

Andersen, P.K. & Borgan, O. (1985). Counting process models for life history data: a review (with discussion), *Scandinavian Journal of Statistics* **12**, 97–158.

Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1982). Linear non-parametric tests for comparison of counting processes, with application to censored survival data (with discussion), *International Statistical Review* **50**, 219–258; Amendment **52**, (1984). 225. (Review paper on nonparametric one and k-sample tests for survival data.).

Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York. (A comprehensive book that may interest a wide range of readers, including biostatisticians, reliability engineers, mathematical statisticians, and probabilists.).

Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York. (An excellent graduate-level textbook.).

Gill, R.D. (1980). *Censoring and Stochastic Integrals*, Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam. (A rigorous treatment of the counting processes approach to survival data.).

Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure lime Data*. Wiley, New York. (An updated comprehensive account of survival data analysis.).

(*See also* **Survival Analysis, Overview**)

I-SHOU CHANG & CHAO AGNES HSIUNG

Covariance Matrix

Suppose that p variables have been measured on each of n sample individuals and the results have been displayed in an $n \times p$ data matrix. Write $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ for the vector of p values observed on the ith individual, so that $\mathbf{x}_i'$ constitutes the ith row of the data matrix ($i = 1, \ldots, n$). In order to provide a framework for parametric **inference**, the $\mathbf{x}_i$ are generally viewed as independent realizations of a random vector $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$, the distribution of which specifies the population from which the sample has been taken. Geometrically, the population can be represented as a swarm of points in p-dimensional space by associating each variable X_j with an orthogonal axis in this space and assigning the observed value $\mathbf{x}_i$ to the point with coordinates $(x_{i1}, x_{i2}, \ldots, x_{ip})$ on these axes (*see* **Axes in Multivariate Analysis**). The main characteristics of this swarm are its location in space and its dispersion. The former is specified by the mean vector

$$\boldsymbol{\mu} = \mathrm{E}(\mathbf{X}) = [\mathrm{E}(X_1), \mathrm{E}(X_2), \ldots, \mathrm{E}(X_p)]',$$

and the latter by the matrix

$$\boldsymbol{\Sigma} = \mathrm{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})',$$

which contains variances of the variables down its principal diagonal and covariances between every pair of variables in its off-diagonal positions.

Whether any more parameters are required fully to specify the population depends on the specific assumptions made about the distributional form of $\mathbf{X}$. In the vast majority of practical applications, however, multivariate **central limit** arguments suggest that **multivariate normality** is a suitable assumption. In this case, the probability density function *only* depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, so interest has focused very heavily on these two parameters within multivariate inference. In this article we concern ourselves with questions about $\boldsymbol{\Sigma}$.

First, we consider its estimation. **Maximum likelihood** is the most commonly adopted method of obtaining estimates of parameters in a frequentist approach to inference. Assuming normality, the **likelihood** of the sample is

$$L = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \times \exp\left[-\tfrac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]$$

and a little algebra (see, for example, [1, pp. 60–65]) establishes that the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)',$$

i.e. the sample mean vector (where $\bar{x}_j = 1/n \sum_{i=1}^{n} x_{ij}$), and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

If we write

$$\mathbf{A} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

then the diagonal elements of $\mathbf{A}$ are the corrected sums of squares $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ of each variable, the off-diagonal elements are the corrected sums of products $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ between every pair of variables, and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{A}.$$

The sampling distribution of $\mathbf{A}$ was derived first by Wishart [17] (*see* **Wishart Distribution**), whose results show that $\mathrm{E}(\mathbf{A}) = (n-1)\boldsymbol{\Sigma}$. Hence the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is biased. When corrected for bias we obtain the estimator

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A},$$

which has the usual sample variances of each variable down the main diagonal and sample covariances between every pair of variables in the off-diagonal positions. This is the *sample covariance matrix*; it is the estimator of $\boldsymbol{\Sigma}$ preferred by many practitioners.

Adopting approaches to inference other than the frequentist produces different estimators of $\boldsymbol{\Sigma}$. The two main approaches are **decision theoretic** and **Bayesian**, and we now briefly summarize the competitor estimates under each of these philosophies.

In decision theory we are required to supply a **loss function** $l(\boldsymbol{\Sigma}, \mathbf{T})$ that quantifies the "loss" incurred when $\boldsymbol{\Sigma}$ is estimated by $\mathbf{T}$. The expectation of this loss over the distribution of the data

defines the **risk** function $R(\boldsymbol{\Sigma}, \mathbf{T})$ associated with that loss, and this risk function is used to compare different estimators $\mathbf{T}_i$. An estimator $\mathbf{T}_1$ *beats* another estimator $\mathbf{T}_2$ if $R(\boldsymbol{\Sigma}, \mathbf{T}_1) \leq R(\boldsymbol{\Sigma}, \mathbf{T}_2)$ for all $\boldsymbol{\Sigma}$ and $R(\boldsymbol{\Sigma}, \mathbf{T}_1) < R(\boldsymbol{\Sigma}, \mathbf{T}_2)$ for at least one $\boldsymbol{\Sigma}$, and an estimator is *admissible*, i.e. "best", if no other estimator beats it. The **unbiased** estimator $\mathbf{S}$ turns out to be the best estimator of the form $\alpha\mathbf{A}$ under the loss function

$$l(\boldsymbol{\Sigma}, \mathbf{T}) = \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T}) - \log\ \det(\boldsymbol{\Sigma}^{-1}\mathbf{T}) - p,$$

but other (more complicated) estimators are best if we either look outside the class of estimators of the form $\alpha\mathbf{A}$ or consider other loss functions. Muirhead [11, pp. 128–136] summarizes the main results.

Turning to the Bayesian approach, it is first necessary to specify a joint **prior distribution** for all the unknown parameters. This is combined with the likelihood of the data to yield a joint posterior distribution of the parameters. Any parameters not of direct interest are then integrated out to give a marginal distribution of the parameters to be estimated, and a suitable summary measure of this marginal distribution (typically the mode) provides the estimator of the parameters. Assuming again that sampling is from a normal distribution, Press [14, p. 168] suggests using the "natural conjugate" prior

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(m+1)/2} \exp\{-\tfrac{1}{2}[\text{tr}\ \boldsymbol{\Sigma}^{-1}\mathbf{G} + (\boldsymbol{\mu} - \boldsymbol{\phi})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\phi})]\},$$

where $\boldsymbol{\phi}, \mathbf{G}$, and $m > 2p - 1$ are parameters the values of which quantify the prior knowledge about $\boldsymbol{\Sigma}$. Following through the above steps, Press then shows that the Bayes estimator of $\boldsymbol{\Sigma}$ is

$$\frac{\left[\dfrac{n\mathbf{A} + \mathbf{G} + n(\bar{\mathbf{x}} - \boldsymbol{\phi})(\bar{\mathbf{x}} - \boldsymbol{\phi})'}{1 + n}\right]}{(n + m - 2p - 2)}.$$

If there is no prior knowledge about $\boldsymbol{\Sigma}$, then a suitable choice of prior distribution is

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2},$$

which yields the Bayes estimator $[1/(n - p - 2)]\mathbf{A}$ on working through the same steps as before. Further results concerning Bayes estimation of $\boldsymbol{\Sigma}$ are given by Dickey et al. [3] and Leonard & Hsu [9].

All of the above results are appropriate when sampling from a multivariate normal distribution. In recent years there has been some interest in theory associated with *elliptic* distributions. These distributions share many of the features of the normal distribution (which is itself a member of this class of distributions), but they encompass distributions such as the **multivariate *t***, the multivariate Cauchy, and the multivariate **logistic**, all of which have heavier tails than the multivariate normal. Elliptic distributions thus provide good models for data involving either outliers or other contaminants. An elliptic distribution with mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$ has a density function of the form

$$f(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2}\psi[(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]$$

for some function $\psi(\cdot)$. Fang & Zhang [4] show that if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ form a random sample from this distribution, then the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\lambda_0\mathbf{A}$, where λ_0 is the maximum of the function $\phi(\lambda) = \lambda^{-np/2}\psi(p/\lambda)$. It is easy to check that in the case of a normal distribution, $\psi(z) = (2\pi)^{-p/2}\exp(-z/2)$, so that $\lambda_0 = 1/n$ and we recover the maximum likelihood estimator $\hat{\boldsymbol{\Sigma}}$.

Despite all of the above results, in the overwhelming number of practical applications $\boldsymbol{\Sigma}$ is routinely estimated in frequentist fashion either by the maximum likelihood estimator $\hat{\boldsymbol{\Sigma}} = (1/n)\mathbf{A}$ or by the unbiased matrix $\mathbf{S} = [1/(n - 1)]\mathbf{A}$, so we restrict our attention to these estimators for the rest of the present section.

The asymptotic distribution of $\mathbf{S}$ provides a mechanism for obtaining large-sample inferences about $\boldsymbol{\Sigma}$ without making any assumptions of normality for the data. We simply require $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to be independent realizations of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, the distribution of which has mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. There are $\frac{1}{2}p(p + 1)$ distinct elements of $\mathbf{S}$ and these elements can be written as a vector $\mathbf{s}$. [One common way of doing this is by stacking successive columns of the lower-triangular portion of $\mathbf{S}$ on top of each other in a column vector; such a vector is denoted vech($\mathbf{S}$)]. The corresponding vector representation of $\boldsymbol{\Sigma}$ can be denoted $\boldsymbol{\sigma} = \text{vech}(\boldsymbol{\Sigma})$. Layard [8] has studied the joint distribution of elements of $\mathbf{s}$. He uses the multivariate central limit theorem to show that, asymptotically for $n \to \infty$, $\mathbf{s}$ has a multivariate normal distribution in which the mean vector is $\boldsymbol{\sigma}$, the

variance of any element s_{jk} is

$$\frac{1}{n}[\mathrm{E}(Z_j^2 Z_k^2) - \mathrm{E}(Z_j^2)\mathrm{E}(Z_k^2)],$$

and the covariance between any two elements s_{jk} and s_{ms} is

$$\frac{1}{n}[\mathrm{E}(Z_j Z_k Z_m Z_s) - \mathrm{E}(Z_j Z_k)\mathrm{E}(Z_m Z_s)],$$

where $Z_i = X_i - \mu_i$ for $i = 1, \ldots, p$. **Convergence** to normality can be speeded up by transforming the elements of **s**, taking logarithms of the variances s_{jj} and using $\tanh^{-1}[s_{jk}/(s_{jj}s_{kk})^{1/2}]$ in place of the covariances s_{jk}. However, this improvement in speed of convergence comes at the expense of complicating the terms in the asymptotic dispersion matrix. Details are given by Seber [16, pp. 99–101]. Asymptotically, of course, $\hat{\boldsymbol{\Sigma}}$ has the same distribution as **S**.

This asymptotic distribution enables large-sample (approximate) confidence regions and hypothesis tests to be constructed for elements of $\boldsymbol{\Sigma}$, irrespective of the distribution from which the sample has been drawn (*see* **Large-sample Theory**). However, for small sample exact tests or for tests of specified structure of $\boldsymbol{\Sigma}$ we need to assume normality of data. Moreover, even then it is virtually impossible to employ optimal theory of **hypothesis testing** as uniformly **most powerful tests** are derivable only in rather artificial circumstances. In most practical circumstances, therefore, recourse must be made to some general principle that can be relied on to produce a "good" test. The principle of *invariance* will often focus attention on a particular class of test statistics within which to search, but may not necessarily pinpoint one specific test. To do this, the most common approaches are to use either the **likelihood ratio** or the **union–intersection principles** of test construction.

Suppose that the null hypothesis H_0 imposes a set of d constraints on the parameters, say $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}$ has d elements, and the alternative hypothesis H_a is the general "not H_0". Usually, also, there will be other (nuisance) parameters $\boldsymbol{\psi}$. Write $l(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}})$ for the log likelihood of the sample when $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\psi}}$ are unconstrained maximum likelihood estimates of all the parameters, and $l(\boldsymbol{\theta}_0, \hat{\boldsymbol{\psi}}_0)$ for the log likelihood when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\psi}}_0$ is the maximum likelihood estimate of $\boldsymbol{\psi}$ conditional on $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, under regularity conditions, the **likelihood ratio test** statistic is

$$\omega = 2l(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) - 2l(\boldsymbol{\theta}_0, \hat{\boldsymbol{\psi}}_0)$$

(or some monotonic function of ω).

On the other hand, any null hypothesis involving $d > 1$ constraints can be regarded as the union of an infinite set of simpler hypotheses. For example, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ implies $\mathbf{a}'\boldsymbol{\theta} = \mathbf{a}'\boldsymbol{\theta}_0$ for any vector $\mathbf{a}$. This is a univariate hypothesis, and univariate theory will generally supply some test statistic, V say, for this hypothesis. Finding the value of $\mathbf{a}$ (up to a multiplying factor) that maximizes V, choosing this hypothesis and then testing it (making due allowance for the maximization) is the basis of the union-intersection principle of test construction.

A full account of these principles of test construction can be found in most multivariate textbooks; see, for example, Mardia et al. [10]. Here we simply summarize the test statistics and their null distributions for the most common tests about $\boldsymbol{\Sigma}$. In all of these tests we assume normality of data, unknown population mean vector $\boldsymbol{\mu}$, and the general alternative H_a: not H_0. Thus, for deriving the likelihood ratio test statistic, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ above are $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ respectively, while unconstrained maximum likelihood estimators are given by $\hat{\boldsymbol{\Sigma}} = (1/n)\mathbf{A}$ and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ as above.

1. $\mathrm{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, a specified matrix. The likelihood ratio test statistic is $\omega = n\,\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}) - n\log|\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}| - np = np(a - \log g - 1)$, where a and g are the arithmetic and geometric means of the **eigenvalues** of $\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}$. For the exact null distribution of this statistic, see Anderson [1] and Korin [6]. However, this distribution is not easy to use, so recourse has to be made to the general result that asymptotically, ω has a $\chi^2_{(1/2)p(p+1)}$ distribution (**chi-square distribution** with $\frac{1}{2}p(p+1)$ **degrees of freedom**) under H_0. The union–intersection statistic, on the other hand, is a function of just the extreme eigenvalues of $\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}$. This test rejects H_0 if either $\lambda_p < c_1$ or $\lambda_1 > c_2$, where λ_i is the ith largest eigenvalue of $\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}$ and c_1 and c_2 are chosen to make the size of test α (*see* **Level of a Test**). Tables for carrying out this test are given in Pearson & Hartley [13].

2. $H_0 : \boldsymbol{\Sigma} = k\boldsymbol{\Sigma}_0$, for unknown k. The maximum likelihood estimate of k is given by $\hat{k} = \text{tr}(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}})/p$ and the likelihood ratio statistic is $\omega = np\log(a_0/g_0)$, where a_0 and g_0 are the arithmetic and geometric means of the eigenvalues of $\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}$. Asymptotically this statistic has a $\chi^2_{(1/2)(p-1)(p+2)}$ distribution under H_0. The special case of $\boldsymbol{\Sigma}_0 = \mathbf{I}$ leads to the **sphericity test**, for which we have $a_0 = (1/p)\,\text{tr}\hat{\boldsymbol{\Sigma}}$ and $g_0 = |\hat{\boldsymbol{\Sigma}}|^{1/p}$. No straightforward union–intersection tests exist in these situations, but Olkin & Tomsky [12] give some modified versions. The test of sphericity plays an important role in **analysis of variance**. In general, the data vector should have a covariance matrix consonant with the sphericity hypothesis for the F tests on means in this analysis to be valid. Also, more particular model structures can be reduced to this hypothesis and tested. The most important of these is the usual covariance structure assumed for repeated measures data (*see* **Longitudinal Data Analysis, Overview**), in which all variances (diagonal elements of $k\boldsymbol{\Sigma}_0$) are assumed to be equal to σ^2 and all covariances (off-diagonal elements of $k\boldsymbol{\Sigma}_0$) are assumed to be equal to $\rho\sigma^2$. It can be shown that any p-element random vector $\mathbf{X} = (X_1, \ldots, X_p)'$ satisfies this covariance structure if and only if the $(p-1)$-element vector $\mathbf{Y} = \mathbf{CX}$ satisfies the sphericity hypothesis, where $\mathbf{C}$ is any $(p-1) \times p$ matrix the rows of which are orthogonal to each other and to the vector $\mathbf{1} = (1, 1, \ldots, 1)'$.
3. $H_0 : \boldsymbol{\Sigma}$ is diagonal. This is the hypothesis that all the variables are uncorrelated with each other. Under H_0, the mean and variance of each variable are estimated separately, whence $\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}} = \mathbf{R}$, the sample **correlation** matrix. This has trace p, so $\omega = -n\log|\mathbf{R}|$. Under H_0, ω has an asymptotic $\chi^2_{(1/2)p(p-1)}$ distribution; Box [2] showed that the χ^2 approximation is improved if n is replaced by $n' = n - \frac{1}{2}(2p + 11)$. There is no straightforward union-intersection test in this case either.

Of course, the maximum likelihood estimate $\hat{\boldsymbol{\Sigma}}$ is equal to $[(n-1)/n]\mathbf{S}$, so each of the above test statistics can be expressed in terms of $\mathbf{S}$ if so desired.

Properties such as the unbiasedness and invariance of these statistics are discussed by Giri [5, Chapter 8] and Muirhead [11, Chapter 8 and 11]. Muirhead also details modifications to the statistics in order to insure unbiased tests, and establishes asymptotic null and nonnull distributional results for samples from elliptic as well as from normal distributions. A general review of all the tests, along with some significance levels, is provided by Krishnaiah & Lee [7].

One other problem of common interest is the testing of equality of dispersion matrices in several multivariate populations, since the assumption of equal dispersion matrices is made in multivariate techniques such as canonical variate analysis (*see* **Canonical Correlation**) and **multivariate analysis of variance**. The likelihood ratio test is a generalization of **Bartlett's test** of homogeneity of variance in univariate populations. We assume that random samples of sizes $n_1, n_2, \ldots, n_g$ are available from each of g populations, and we write $N = \sum_{i=1}^{g} n_i$. Suppose that $\mathbf{A}_i$ is the sums of squares and products matrix for the sample from population i, so that $\hat{\boldsymbol{\Sigma}}_i = (1/n_i)\mathbf{A}_i$ is the maximum likelihood estimator of the dispersion matrix for this population and $\mathbf{S}_i = [1/(n_i - 1)]\mathbf{A}_i$ is the unbiased version. Under the null hypothesis that all dispersion matrices are equal to $\boldsymbol{\Sigma}$, we have $\hat{\boldsymbol{\Sigma}} = (1/N)\sum_{i=1}^{g}\mathbf{A}_i$ and the corresponding unbiased version $\mathbf{S} = [1/(N-g)]\sum_{i=1}^{g}\mathbf{A}_i$. (The latter matrix is known as the pooled within-sample covariance matrix.) Then the likelihood ratio test statistic for testing the null hypothesis against the general alternative that at least one dispersion matrix differs from the rest is $N\log|\hat{\boldsymbol{\Sigma}}| - \sum_{i=1}^{g} n_i \log|\hat{\boldsymbol{\Sigma}}_i|$, and under the null hypothesis this statistic is asymptotically distributed as $\chi^2_{(1/2)p(p+1)(g-1)}$. Box [2] proposed the alternate statistic $(N-g)\log|\mathbf{S}| - \sum_{i=1}^{g}(n_i - 1)\log|\mathbf{S}_i|$, which has the same asymptotic chi-square distribution under the null hypothesis. He also gave an F approximation to the null distribution, and tables based on this latter approximation are given by Seber [16].

The union-intersection approach is viable in the special case $g = 2$, and produces a test based on the largest and smallest eigenvalues of $\mathbf{S}_1\mathbf{S}_2^{-1}$, with tables given by Schurrmann et al. [15]. However, this test does not generalize easily to the case $g > 2$.

This section has been concerned with inferential aspects of the sample covariance matrix. This matrix is at the heart of many multivariate techniques; see especially **principal components analysis** and **factor analysis**.

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika* **36**, 317–346.

[3] Dickey, J.M., Lindley, D.V. & Press, S.J. (1985). Bayesian estimation of the dispersion matrix of a multivariate normal distribution, *Communications in Statistics – Theory and Methods* **14**, 1019–1034.

[4] Fang, K.-T. & Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*. Science Press, Beijing/Springer-Verlag, Berlin.

[5] Giri, N.C. (1977). *Multivariate Statistical Inference*. Academic Press, New York.

[6] Korin, B.P. (1968). On the distribution of a statistic used for testing a covariance matrix, *Biometrika* **55**, 171–178.

[7] Krishnaiah, P.R. & Lee, J.C. (1980). Likelihood ratio tests for mean vectors and covariance matrices, in *Handbook of Statistics*, Vol. 1, P.R. Krishnaiah. ed. North-Holland, Amsterdam, pp. 513–570.

[8] Layard, M.W.J. (1972). Large sample tests for the equality of two covariance matrices, *Annals of Mathematical Statistics* **43**, 123–141.

[9] Leonard, T. & Hsu, J.S.J. (1992). Bayesian inference for a covariance matrix, *Annals of Statistics* **20**, 1669–1696.

[10] Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, London.

[11] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

[12] Olkin, I. & Tomsky, T.L. (1975). A new class of multivariate tests based on the union-intersection principle, *Bulletin of the International Statistical Institute* **46**, Part 4, 202–204.

[13] Pearson, E.S. & Hartley, H.O. (1972). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge University Press, Cambridge.

[14] Press, S.J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart & Winston, New York.

[15] Schurrmann, F.J., Waikar, V.B. & Krishnaiah, P.R. (1973). Percentage points of the joint distribution of the extreme roots of the random matrix $(\mathbf{S}_1 + \mathbf{S}_2)^{-1}$, *Journal of Statistical Computation and Simulation* **2**, 17–38.

[16] Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

[17] Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate distribution, *Biometrika* **20A**, 32–52 (correction: **20A**, 424).

(*See also* **Inference, Foundations of**; **Multivariate Analysis, Overview**; **Multivariate Bartlett Test**)

W.J. KRZANOWSKI

Covariate Imbalance, Adjustment for

Patients in a **clinical trial** tend to vary considerably with respect to clinical and demographic characteristics, some of which may affect their prognosis. It is clearly desirable that the characteristics of the patients in each group are as similar as possible, especially with respect to those characteristics which are prognostic. The enormous strength of randomized controlled trials for drawing **inferences** about treatments is very largely a direct consequence of the use of **randomization** to decide which patients receive each treatment (*see* **Randomized Treatment Assignment**). While randomization eliminates **bias**, it does not guarantee comparable baseline characteristics of the patients in the different treatment groups in a particular trial. Simple randomization is quite likely to yield some differences, especially in small trials. The use of stratified randomization or minimization (*see* **Adaptive and Dynamic Methods of Treatment Assignment**) will reduce such imbalances for selected variables.

Baseline balance is not a requirement. Because of the use of randomization, standard methods of analysis (**estimation** and **hypothesis testing**) will yield valid results regardless of the distribution of baseline variables. Nonetheless, it is often wise to try to avoid imbalance, using the simple design modifications indicated above, and to allow for imbalance if it arises.

Comparison of Baseline Characteristics

An important part of reporting the results of a clinical trial is to describe the patient characteristics for the different treatment groups. As well as characterizing the whole study sample, these data indicate how similar were the groups produced by randomization. In some sense the information is used to determine if the randomization has "worked". Here many investigators behave illogically, by using statistical tests to compare the groups. The aim is probably to attempt to establish that the groups really are comparable, thus strengthening the credibility of the trial. However, the **null hypothesis** for such tests is in essence that the data come from groups which are **random samples** from the same population. Because the treatment groups were indeed random samples, any differences observed between them are necessarily due to chance, and so the use of hypothesis tests is absurd [1]. Yet it is quite common to see groups described as having the "same" characteristics simply because no significant differences were observed, even when, say, there was a notable difference in mean age or the **prevalence** of smoking. A nonsignificant imbalance between groups can be quite important if that **covariate** is highly **prognostic**. Note that such imbalance can work in either direction, masking or overstating the true treatment difference. Significance testing for baseline differences does have one potential use, which is to see if the patients were indeed randomized. However, there is minimal **power** to test this hypothesis. As Senn [17] has noted, a significant imbalance ought really to lead to the conclusion that the trial was not properly randomized – not a conclusion that researchers are likely to draw about their own study. This test can be useful, however, in a **multicenter trial**. Trial coordinators may be able to detect a center which has not adhered to the protocol [8] (*see* **Clinical Trials Protocols**).

Over recent years many authors have examined the practice of testing baseline differences, with unanimous criticism of the practice [1, 3–5, 17]. These papers were mostly published in statistical journals, however, and hypothesis testing of baseline characteristics remains common in reports of trials in medical journals. Baseline testing was found in about 60% of trial reports in recent reviews of general and specialist journals [14]. Such tests are probably quite a recent development – testing was not mentioned in an early paper on baseline imbalance [12]. They are not a requirement of the regulatory bodies [17] (*see* **Drug Approval and Regulation**).

It might be thought that such testing is largely harmless, as it rarely has much impact on how the trial is analyzed and interpreted. However, carrying out one form of analysis conditional on the results of another can distort the results obtained, as described below. (A similar issue arises in other areas of statistics, such as in the analysis of **crossover designs**.) Baseline testing can have other adverse consequences. For example, it has been found that authors selectively report these tests: only 2% of about 1000 tests reported in 206 trial reports in obstetric journals gave results significant at the 5% level [14]. Some of this effect might be due to undisclosed stratification, but it appears that there is a tendency

to suppress the results of these tests if the imbalance is significant.

Effect of Imbalance

Imbalance in a patient characteristic will matter only if that characteristic is related to patient outcome, i.e. it is prognostic. In most situations gender will not be prognostic, but age often will be, especially in chronic diseases. Prognostic covariates are most often clinical or biochemical variables, some of which can have major importance. The effect of imbalance for a variable will depend both upon the size of the imbalance (e.g. difference in **means** or proportions) and the strength of the relation between that variable and the outcome.

When randomization leads to baseline imbalance in a prognostic variable, one group will have a poorer prognosis than the other before treatment starts. Thus chance imbalance will lead to a biased estimate of the treatment effect, in either direction according to the direction of the imbalance, when using a simple, unadjusted analysis. Also, a test of significance may yield a significant result when there is no true treatment difference or a nonsignificant result when there is. To take a specific example, Christensen et al. [7] carried out a randomized trial of azathioprine vs. placebo in patients with primary biliary cirrhosis. The unadjusted analysis gave $P = 0.2$ for the treatment comparison. There was some imbalance in serum bilirubin, which is a very strong prognostic variable in such patients. The azathioprine group had higher levels on average and hence a worse prognosis. An adjusted analysis gave $P = 0.02$ for the treatment effect. In practice some imbalance is likely in several prognostic variables, but the overall effect will be much the same as just outlined, especially when, as in this example, one variable is of primary prognostic importance.

Some authors [1, 10, 11] have suggested that imbalance is not so much of a problem for large trials, while others state the opposite [5, 15]. The apparent disagreement arises from the fact that there are several aspects that might be considered – the size of the test of treatment effect, the power of the test, the bias in estimating the treatment effect, and the precision of the estimated treatment effect. Some of these features diminish with increasing sample size, while others apply even to large trials. We certainly cannot rely on large sample size to overcome all of the problems associated with imbalance.

Rationale for Adjusting for Baseline Covariates

There are several reasons why investigators might wish to adjust for baseline characteristics when analyzing the data from a randomized trial, some of which have been mentioned already.

First, as already discussed, the aim of a randomized trial is to compare groups of patients who differ only in that they received different therapies. Imbalance in baseline variables may reduce the credibility of the results, both in the correctness of the randomization procedure and, more importantly, in the validity of the results. Even though such worries may not be well founded, this possibility should be regarded as reasonable grounds for concern. However, the necessary leap in complexity of the methodology (as described below) when adjusting for covariates may be disconcerting to medical readers [9] even though to statisticians it will not cause concern. To some extent transparency is replaced by opaqueness.

Given that chance imbalance in a prognostic variable will lead to some bias in the estimated treatment effect, one of the best reasons for adjusting is to remove this bias. We surely wish to obtain the most reliable estimate of treatment effect. Adjustment will also increase the power to detect a real treatment effect.

Another reason often given for adjusting is to increase the precision with which the treatment effect is estimated. However, while this is the case for normal **regression** models, it will not improve precision in **logistic** [13] or **Cox regression models** [6]. However, in these models, failure to adjust for prognostic variables will lead to underestimation of the treatment effect and hence a reduction in power [6, 13]. The bias associated with not adjusting in non-normal models applies even when there is perfect balance in a prognostic variable.

Methods of Adjusting for Baseline Covariates

The idea behind adjustment for baseline differences is to estimate what the treatment effect would have been if the groups had identical baseline variables, i.e. with

identical means for continuous variables and identical frequencies in each group for categorical variables. The generally recommended approach to adjustment is to use regression modeling, with treatment (as a **binary** variable) and prognostic variables included as the explanatory variables. I will follow convention in this context and refer to this approach as **analysis of covariance**, even though these analyses are not all encompassed within the usual idea of that analysis. Analysis of covariance can be used for all types of outcome measure – continuous, binary, and survival times. Its particular strength is that it gives a result that is **unbiased** regardless of the baseline distribution of prognostic variables – i.e. it is conditionally unbiased [16]. In addition, by comparison with an unadjusted analysis, analysis of covariance provides increased precision for the treatment effect (for normal models), an increase in the power of the trial, and a constant conditional size of the test comparing the treatment groups [15].

An alternative is to use a stratified analysis. This approach is more common in epidemiology, where outcomes are usually binary. Pocock [11] gives an example of the use of the **Mantel–Haenszel** test to perform a stratified analysis of a clinical trial. This method is appropriate for categorical covariates, but may not adjust fully for imbalance in continuous covariates [1]. This method may be seen as a special form of analysis of covariance.

There is rather greater difficulty associated with deciding for which covariates to adjust. I consider several possibilities.

Selection Based on Observed Imbalance

The first approach is to focus on the imbalance: two-sample tests, as discussed above, can be used in turn for each prognostic variable to compare the groups at baseline, with no regard to patient outcome. Those which are statistically significant can be used to adjust the treatment effect in a **multiple regression** analysis. While this strategy is very common, its use is unwise. Those variables with significant imbalance may or may not be prognostic, and by including variables conditionally on simple tests, the adjusted analysis is likely to lead to a biased estimate of the treatment effect. Also, as noted above, nonsignificant imbalance may be quite important, even in a normal model.

Selection Based on Relation to Patient Outcome

A second approach is to focus on patient outcome. A multiple regression model can be derived using stepwise selection to see which variables are significant predictors of the outcome, taking no account of baseline balance. While this analysis could be done ignoring treatment or separately within each treatment group, it is most sensible to include all patients and to include in the model an indicator for treatment. This analysis thus yields both the choice of important prognostic variables and the adjusted treatment effect. While far preferable to adjustment based on observed imbalance, this method is not fully satisfactory. Apart from the known overoptimism of regression models based on stepwise selection, adjustment is made for a data-dependent selection of prognostic variables using an arbitrary inclusion rule. We might instead choose those variables which have the largest effect on the estimated treatment effect, either as assessed by the change in the test statistic, as proposed by Canner [5], or the change in the magnitude of the estimated treatment effect (e.g. by 15%). While more reasonable than using ***P* values**, it is unclear here what the criterion should be for deciding which variables have a large enough effect to need adjustment.

An approach proposed by Tukey [19] can be outlined only briefly here. The idea is to minimize the number of regression coefficients without reducing the number of covariates. The outcome variable is regressed on each covariate in turn, with the patients in each group pooled. From each analysis a score is derived from the P value – he suggested scores of 1 to 4 corresponding to $P < 0.05$, $P < 0.01$, $P < 0.001$, and $P < 0.0002$, the scores being signed according to the direction of the effect. The method is easiest to explain with binary covariates each coded as "high" or "low". For a set of covariates, a "composite" is constructed for each patient as a weighted sum of the scores, where the weight is 0 if the variable is low and 1 if high. This composite is then treated as a single covariate to adjust the treatment effect. The weaknesses of this method include the use of the P value as a measure of the strength of the effect, the treatment of all covariates as providing independent information, and the lack of transparency.

Prespecified List of Variables

There are problems associated with all data-derived decisions about which variables to include in an

analysis. In particular, the use of significance tests to determine which variables to adjust for is not recommended. It seems far preferable to choose which variables to adjust for without regard to the actual data set to hand.

What criteria should be used to select such variables? Primarily one would wish to consider known important prognostic variables that have not been controlled by the design. It is advisable also to include any variables used for stratification. In addition, in multicenter trials it may be desirable to include centers. The prespecified strategy has the advantage of focusing attention on prognostic factors at the design stage, rather than leaving this issue to be dealt with in an ad hoc manner in the analysis. It means that for some trials the analysis will make adjustment for covariates which are in fact balanced. This will not matter greatly in the case of a **normally distributed** outcome, and is desirable, as noted above, for non-normal outcomes.

Baseline Measurements of the Outcome Variable

In many clinical trials where the object of treatment is to change the value of a continuous measurement (such as blood pressure), it is possible to measure the variable of interest at the start of the trial. The undesirability of baseline imbalance in the variable of primary interest is especially clear. One approach to the analysis of such trials is to analyze change from baseline. This would seem to solve the baseline imbalance problem, but it does not. The change from baseline within each group will usually be highly **correlated** with the baseline values (*see* **Regression to the Mean**), so the difference between the groups in change from baseline will be negatively correlated with the imbalance at the baseline [18]. In other words, while analyzing change from baseline seems to remove the problem associated with baseline imbalance, in fact the chance imbalance will still affect the difference in outcome, but in the opposite direction. In the case where we are seeking to increase lung function, say, and if by chance patients receiving treatment A have higher baseline values than those receiving treatment B, then the analysis of change from baseline will be biased in favor of group B, and vice versa if the imbalance goes the other way. Thus it can be seen that analysis of change from baseline does not deal adequately with baseline imbalance. It is often argued that in such trials change from the baseline is a clinically more relevant outcome measure. Senn [16] has argued strongly against this view. In any case, one can use analysis of covariance to adjust change from baseline for baseline values, with exactly equivalent answers, so the debate is irrelevant if analysis of covariance is used [16, 18] (*see* **Baseline Adjustment in Longitudinal Studies**).

Such trials may cause a further error. It is quite common to see authors report separate tests to assess whether each group has changed from the baseline. The resulting P values are compared and a difference claimed when one P value is significant and the other is not. This is not a valid form of statistical inference, and is likely to mislead [2, 15].

Comments

The main issues here are the proper analysis of randomized trials, and the distinction between substantive and **exploratory analyses**. A clear recommendation may be made for the analysis of trials. Ideally, a prespecified strategy should be developed as part of the protocol in which either no adjustment will be made for baseline variables or adjustment will be made for nominated variables using analysis of covariance.

Good statistical practice requires investigators to prespecify in the study protocol their intentions with regard to sample size (*see* **Sample Size Determination**), primary (and subsidiary) endpoints (*see* **Outcome Measures in Clinical Trials**), subgroup analyses (*see* **Treatment-covariate Interaction**), and so on. It is no different to suggest that the analysis strategy should also be prespecified, in particular intentions regarding adjusted analyses. It is usually known in advance which are the variables that are most prognostic of patient outcome. Whether or not these are used as stratifying variables, the trial protocol should specify which ones will be adjusted for in the analysis, and that this adjustment will not be conditional on the distribution of those variables across the treatment groups [4, 5, 15]. It would not be acceptable to specify in the protocol that adjustment would be made for any variables showing statistically significant imbalance; this would not circumvent the problems described above.

The protocol should also specify which will be the primary analysis. My view is that this should usually

be the adjusted analysis, otherwise there is little point in performing it. Sometimes, however, the adjusted analysis may be performed in order to strengthen belief in the results of the unadjusted analysis. Here it becomes unclear how similar the results need to be for the unadjusted analysis to be confirmed. It is not desirable for the choice of primary analysis to be conditional on the results.

It may not be as simple as just suggested to identify the "most prognostic variables". Clinicians may argue that all information being collected is potentially prognostic, which is why it is being collected. It may prove difficult to persuade them to identify in advance which variables will and which will not be adjusted for in the analysis, and harder still to get them to comply with this strategy when imbalance is seen within the latter group. Despite the wide recommendation of this general strategy, it is not common to see published studies reporting this as the basis for their chosen analysis. Partly, though, this may be because balance has been achieved through the design.

In practice, imbalance may arise when the possible need for adjustment has not been anticipated. What should the researchers do? They might choose to ignore the imbalance; as noted, this would be entirely proper. The difficulty then is one of credibility. Readers of their paper (including reviewers and editors) may question whether the observed finding has been influenced by the unequal distribution of one or more baseline covariates. It is still possible, and arguably advisable, to carry out an adjusted analysis, but now with the explicit acknowledgment that this is an exploratory rather than definitive analysis, and that the unadjusted analysis should be taken as the primary one. Obviously, if the simple and adjusted analyses yield substantially the same result, then there is no difficulty of interpretation. This will usually be the case. However, if the results of the two analyses differ, then there is a real problem. The existence of such a discrepancy must cast some doubt on the veracity of the overall (unadjusted) result. The situation is similar to the difficulties of interpretation that arise with unplanned subgroup comparisons. One suggestion in such circumstances is to try to mimic what would have been done if the problem *had* been anticipated, namely to adjust not for variables that are observed to be unbalanced, but for all variables that would have been identified in advance as prognostic. An independent source could be used to identify such variables. Alternatively, the trial data could be used to determine which variables are prognostic. This strategy too could be prespecified in the study protocol. Because this analysis would be performed conditionally on the observed imbalance, it does not remove bias and thus cannot be considered fully satisfactory.

Finally, I have assumed implicitly that the treatment effect is the same on average regardless of the values of the covariates. It may be desirable to examine whether there are any **treatment–covariate interactions**; here, too, prespecification of intentions is strongly advisable.

References

[1] Altman, D.G. (1985). Comparability of randomised groups, *Statistician* **34**, 125–136.

[2] Altman, D.G. & Doré, C.J. (1990). Randomisation and baseline comparisons in clinical trials, *Lancet* **335**, 149–153.

[3] Beach, M.L. & Meier, P. (1989). Choosing covariates in the analysis of clinical trials, *Controlled Clinical Trials* **10**, 161S–175S.

[4] Begg, C.B. (1990). Significance tests of covariate imbalance in clinical trials, *Controlled Clinical Trials* **11**, 223–225.

[5] Canner, P. (1991). Covariate adjustment of treatment effects in clinical trials, *Controlled Clinical Trials* **12**, 359–366.

[6] Chastang, C., Byar, D. & Piantadosi, S. (1988). A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models, *Statistics in Medicine* **7**, 1243–1255.

[7] Christensen, E., Neuberger, J., Crowe, J., Altman, D.G., Popper, H., Portmann, B., Doniach, D., Ranek, L., Tygstrup, N. & Williams R. (1985). Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial, *Gastroenterology* **89**, 1084–1091.

[8] Collins, R., Gray, R., Godwin, J. & Peto, R. (1987). Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews, *Statistics in Medicine* **6**, 245–250.

[9] Greenberg, E.R., Baron, J.A. & Colton, T. (1983). Reporting the results of a clinical trial, in *Clinical Trials: Issues and Approaches*, S.H. Shapiro & T.A. Louis, eds. Marcel Dekker, New York, pp. 191–204.

[10] Grizzle, J.E. (1982). A note on stratifying versus complete random assignment in clinical trials, *Controlled Clinical Trials* **3**, 365–368.

[11] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester, pp. 211–221.

[12] Radhakrishna, S. & Sutherland, I. (1962). The chance occurrence of substantial initial differences between

groups in studies based on random allocation, *Applied Statistics* **11**, 47–54.

[13] Robinson, L.D. & Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models, *International Statistical Review* **58**, 227–240.

[14] Schulz, K.F., Chalmers, I., Grimes, D.A., Altman, D.G. & Doré, C.J. (1995). The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals, *Online Journal of Current Clinical Trials* **4**, Doc. No. 197.

[15] Senn S. (1989). Covariate imbalance and random allocation in clinical trials, *Statistics in Medicine* **8**, 467–75.

[16] Senn, S. (1991). Baseline comparisons in randomized clinical trials, *Statistics in Medicine* **10**, 1157–1159.

[17] Senn, S. (1995). Base logic: tests of baseline balance in randomized clinical trials, *Clinical Research and Regulatory Affairs* **12**, 171–182.

[18] Senn, S. (1997). *Statistical Issues in Drug Development*. Wiley, Chichester, pp. 95–109.

[19] Tukey, J.W. (1991). Use of many covariates in clinical trials, *International Statistical Review* **59**, 123–137.

(*See also* **Variable Selection**)

DOUGLAS G. ALTMAN

Covariate

Quantification of the relationship between a response variable and a group of **explanatory variables** is the goal of fitting **regression** models. Some explanatory variables may be the main focus of a study, such as treatment variables in experimental studies or **risk** factors in epidemiologic studies. Other variables may be measurements that must be controlled for in the analysis but are not of specific interest, such as **confounders**. These latter variables are often termed covariates. Other terms for these variables are covariables or concomitant variables.

For illustration, consider that often in epidemiologic studies it is known that age is associated with the outcome of interest and that age is also associated with the exposure under investigation. The primary hypothesis focuses on the relationship between the outcome and the exposure with age included in the model because of the possible confounding effect of age. Thus, age is a covariate since the relationship between the outcome and age is not a focus of the study but, nonetheless, age is included in the statistical model.

In the context of experimental design, the **analysis of covariance** is used so that one or more covariates are taken into consideration along with the treatment variables of primary interest. The inclusion of the covariate information is viewed as necessary to assess appropriately the relationship between the outcome measure and the treatments. An example is an investigation of the effect of dietary components on weight in experimental animals. The initial weight of the animals is a covariate, or covariable, and the quantification of the effect of the dietary components is more accurately assessed after controlling for initial weight.

As with other explanatory variables, a covariate can be quantitative or qualitative. Likewise, in the case of **survival analysis**, covariates, and other explanatory variables, can be time-independent (e.g. sex) or **time-dependent** (e.g. marital status).

It should be noted that, while the term *covariate* does have a specific meaning, it is often used interchangeably with the terms *explanatory variable, predictor variable*, or *independent variable*.

G.A. DARLINGTON

Cox Regression Model

The Cox or **proportional hazards** regression model [21] is used to analyze **survival** or failure time data. It is now perhaps the most widely used statistical model in medical research. Whenever the outcome of a **clinical trial** is the time to an event, the Cox model is the first method considered by most researchers. The model has also inspired an enormous statistical literature, ranging from the mathematical study of estimating the model parameters, to applied techniques for validating the model assumptions.

This article is divided into sections touching on some of the vast literature that has developed around the model:

1. model definition
2. history
3. using the Cox model–the basics
4. estimators and **algorithms**
5. asymptotic properties (*see* **Large-sample Theory**)
6. **time-dependent** explanatory variables
7. **model checking**
8. alternatives and extensions.

Several books have now been published on survival analysis that devote major sections to the Cox model. The first of these appeared in the early 1980s [23, 50]. Of the more recent books some are mathematically rigorous [6, 29], while others are more applied [20, 53, 60]. The book by Andersen et al. [6] is the most comprehensive.

Model Definition

Cox's essential novelty was to model the hazard function (*see* **Hazard Rate**) rather than the mean or some other measure of location. Let X denote a random failure time and $\mathbf{Z}$ a vector of **explanatory variables**. The conditional hazard of X given $\mathbf{Z} = \mathbf{z}$ at time t is defined as

$$\lambda(t|\mathbf{z}) = \lim_{\Delta t \downarrow 0} \frac{\Pr(X \leq t + \Delta t | X > t, \mathbf{z})}{\Delta t}. \quad (1)$$

The hazard function is sometimes called the intensity function or the force of mortality. Roughly, the hazard function is the probability that someone who is alive now will die in the next small unit of time. Cox proposed that the conditional hazard be modeled as the product of an arbitrary baseline hazard $\lambda_0(t)$ and an exponential form that is linear in $\mathbf{z}$:

$$\lambda(t|\mathbf{z}) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{z}). \quad (2)$$

Here $\boldsymbol{\beta}$ is a vector of regression parameters and the infinite-dimensional parameter $\lambda_0(\cdot)$ is the hazard function for an individual with $\mathbf{Z} = \mathbf{0}$. The model in (2) forces the hazard ratio between two individuals to be constant over time:

$$\frac{\lambda(t|\mathbf{z}_2)}{\lambda(t|\mathbf{z}_1)} = \exp[\boldsymbol{\beta}'(\mathbf{z}_2 - \mathbf{z}_1)].$$

The exponential form of the relative risk function has become standard and is the most stable computationally, but it is not the only possibility. The more general model,

$$\lambda(t|\mathbf{z}) = \lambda_0(t) r(\boldsymbol{\beta}'\mathbf{z}),$$

for some known function r has also been considered [67, 82].

History

A distinguishing feature of survival data is that it is subject to **censoring**. Very often one does not observe the survival time for all individuals in a study. One may only know that a certain individual was still alive at some time T^*. If T_i^* is the last time at which individual i is known to be alive, it is called a censoring time – the individual's follow-up was censored at T_i^*. In 1958, Kaplan & Meier [51] studied the product-limit estimator of a survival function based on censored data (*see* **Kaplan–Meier Estimator**). The key concept of viewing the data as a process that reveals itself over time can be seen in their paper. Test statistics for censored data were considered a few years later [31, 59], and some may view the Cox model as the natural generalization to a regression setting of ideas present in Mantel's writing [59]. At about the same time, Feigl & Zelen [28] considered various **exponential** regression models. One of their models is equivalent to the Cox model with the baseline hazard constrained to be constant for all time, so that $\lambda(t|\mathbf{z})$ is a function of $\mathbf{z}$ but not t. However, unlike Cox [21], they formulate the model in terms of a parameterization of the mean

survival time, even though they use the exponential assumption to predict the entire survival distribution.

Cox's 1972 paper [21] was instantly acclaimed as a breakthrough in the analysis of right censored data, as can been seen from the enthusiastic discussion published together with the article. The model was rapidly adopted by applied statisticians, particularly in clinical trials. Its use became widespread once user-friendly software became readily available. Today, one can hardly open a leading medical or statistical journal without finding at least one reference to Cox (1972)! It is one of the most widely cited papers in scientific literature.

The original paper introduced a model that was to revolutionize the field, and provided the estimator that is today programmed into many statistical software packages. There were, however, several issues that were to challenge the statistical community. Some of these, such as how to deal with ties (two or more individuals with the same failure time) [63] (*see* **Tied Survival Times**), and the basis for the proposed estimator, were addressed at the Royal Statistical Society meeting. Cox provided justification for the estimator himself by introducing the concept of a **partial likelihood** [22]. But it was not until later that the estimators were shown to be **efficient** [11, 27]. Formal proofs of **consistency** and asymptotic normality took nearly a decade [5, 83]. Another topic of considerable interest to statisticians is the effect of **misspecification** on the estimates [80], and model interpretation. Various types of misspecification have been considered: explanatory variables measured with error [65] (*see* **Errors in Variables**); omission of important explanatory variables [16, 54, 78]; and rare but gross data contamination [9, 72] (*see* **Outliers**).

Parallel with the theoretical progress was work on model building and model checking. The results were less satisfactory than the elegant theory that developed around **counting processes** and martingales, but a variety of tools are now available. These included **goodness of fit** tests, as well as **residuals** and other **diagnostics**. Andersen [4] and others have discussed the quality of presentation of Cox regression analyses in the medical literature. Despite their constructive suggestions, the "Methods" sections of many papers are still no more informative than "we used the Cox model".

The basic model, (2), has been generalized in various directions. Even the original paper [21] considered time-dependent covariates, but these still cause a variety of difficulties [3]. A simple generalization is to permit different baseline hazard functions in each of a number of strata (*see* **Stratification**). The stratified Cox model assumes that, within each stratum, the proportional hazards assumption is justified and that the effect of the variable $\mathbf{Z}$ is the same in all strata:

$$\lambda_j(t|\mathbf{z}) := \lambda(t|\mathbf{z},\ \text{stratum}\ j) = \lambda_{0j}(t)\exp(\boldsymbol{\beta}'\mathbf{z}). \quad (3)$$

By incorporating constructed variables, that are constant in some strata, the stratified model, (3), can be used to model interactions between explanatory variables and strata. Suppose, for example, that one is stratifying by sex and including age as an explanatory variable. Let $z_1 = (\text{age} - 50)$ for men, $= 0$ for women; and let $z_2 = (\text{age} - 50)$ for women, $= 0$ for men. Then a model stratified on sex that includes z_1, z_2, and a treatment indicator z_3 permits interactions between age and sex, but assumes that the treatment acts proportionately on the hazards for any age–sex combination.

Many models used for analysis of **multivariate survival data** are generalizations of the Cox model, but they are not discussed here.

Using the Cox Model – the Basics

Before using the Cox model, or even attempting to interpret a published analysis, one must have some understanding of the assumptions that underlie the analysis. This section discusses those assumptions and explains a typical output from fitting the model in a statistical package.

There are three components to the data on each individual: the possibly censored failure time T; an indicator δ (*see* **Dummy Variables**) equal to 1 if T is a true failure time, 0 if it is censored; and $\mathbf{Z}$, the vector of explanatory variables. The model is flexible enough to incorporate explanatory variables that change value over the course of the study, but in this section we assume that $\mathbf{Z}$ is fixed and measured at time $t = 0$. The key censoring assumption is that the observation $(T = t, \delta = 0)$ tells us nothing more than that the true failure time X is greater than t.

In a clinical trial, the time origin for each individual will usually be his or her time of entry into the trial. If the trial ends at a particular calendar time, censoring all individuals who are not yet dead, then the censoring times are the times from entry until

the end of the trial and will vary from one individual to another. This is called administrative (or progressive type I) censoring. In such situations, it is necessary for survival to be independent of entry time for the above condition to be satisfied. To some extent this can be examined by including entry time as a covariate or by stratifying on the date of entry. Other forms of censoring are more problematic. If, for instance, a patient emigrates, one needs to consider whether this implies that the patient had in fact recovered. Conversely, a patient who fails to attend a follow-up clinic might be too sick to get out of bed. In such cases, the fact that the patient was censored at t tells us rather more than that she was alive at t.

The Cox model itself makes three assumptions: first, that the ratio of the hazards of two individuals is the same at all times; secondly, that the explanatory variables act multiplicatively on the hazard; and thirdly, that, conditionally on $\mathbf{Z}_i$ and $\mathbf{Z}_j$, the failure times of individuals i and j are independent. As with all regression models, one also assumes that the explanatory variables have been transformed so that they may be entered without further transformation and that all interactions have been included explicitly. We will see in the section on asymptotics that the independence assumption can be relaxed.

Table 1 presents the results of fitting a Cox model to data from 216 patients with primary biliary cirrhosis in a clinical trial of azathioprine vs. placebo [18]. The six variables were selected from an initial set of 25 partly using forward stepwise selection. An additional 32 patients were excluded because they had missing values of one or more of the six variables. Recruitment was over 6 years and follow-up a further 6 years. Of the 216 patients, 113 had censored survival times. The regression coefficients may be combined with their standard errors to obtain **confidence intervals** that rely on the asymptotic normality of the estimates.

The positive coefficient associated with treatment implies that patients on the placebo ($Z = 1$) had poorer prognosis than those on azathioprine ($Z = 0$): the hazard of those on placebo is about 1.7 times greater than that of those on active treatment. Similarly, older patients had poorer prognosis. The hazard ratio associated with two patients aged 50 and 30 is $\exp[0.0069(\exp 3 - \exp 1)] = 1.13$. Notice, however, that the effect on survival is not fully described by the information in Table 1 because, without estimating the baseline hazard, one cannot translate the regression coefficients into effects on 5-years survival nor on **median survival**.

Most statistical software for Cox regression will also estimate the cumulative baseline hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u \tag{4}$$

(*See* **Survival Distributions and Their Characteristics**), and from this one can calculate the estimated survival function for a given $\mathbf{z}$:

$$\Pr(X > t|\mathbf{z}) = \prod_{\{i:T_i \le t\}} [1 - \mathrm{d}\hat{\Lambda}_0(T_i)\exp(\boldsymbol{\beta}'\mathbf{z})].$$

Plots of the estimated survival function can be made for various $\mathbf{z}$s, and these can be viewed like **Kaplan–Meier** graphs. Alternatively, the estimated survival function can be used to estimate 5-year survival, say, as a function of the prognostic index $\boldsymbol{\beta}'\mathbf{z}$ (*see* **Prognosis**).

Estimators and Algorithms

The regression coefficients $\boldsymbol{\beta}$ are estimated by maximizing the so-called partial likelihood $L(\boldsymbol{\beta})$ [22]. An

Table 1 Cox model fitted to data from a clinical trial comparing the effects of azathioprine and placebo on the survival of 216 patients with primary biliary cirrhosis [18]. The six variables shown were selected, partly by a forward stepwise procedure, from 25 candidate variables

Variable	Coding	Coeff. $\hat{\beta}$	se($\hat{\beta}$)	exp($\hat{\beta}$)
Serum bilirubin	$\log_{10}$ (value in μmol/l)	2.51	0.316	12.3
Age	exp[(age in years − 20)/10]	0.0069	0.0016	1.0
Cirrhosis	0 = No; 1 = Yes	0.88	0.216	2.4
Serum albumin	value in g/l	−0.0504	0.018	0.95
Central cholestasis	0 = No; 1 = Yes	0.68	0.275	2.0
Therapy	0 = azathioprine; 1 = placebo	0.52	0.201	1.7

individual is said to be at risk at t if he has not yet failed nor been censored. This concept can be generalized to allow for individuals who do not enter the study at time 0. Such **delayed entry**, or left **truncation**, as it is called, often arises when t is the age of a patient or the time from infection, so that patients enter the study at some time $T_i^0 > 0$. Consider $L_i(\boldsymbol{\beta})$, the conditional probability that individual i fails at time T_i given that exactly one individual fails at T_i and knowing the values of $\mathbf{Z}$ for all individuals at risk at T_i:

$$L_i(\boldsymbol{\beta}) = \frac{\lambda(T_i|\mathbf{Z}_i)}{\sum_{j\in R_i}\lambda(T_i|\mathbf{Z}_j)} = \frac{\exp(\boldsymbol{\beta}' Z_i)}{\sum_{j\in R_i}\exp(\boldsymbol{\beta}'\mathbf{Z}_j)}, \qquad (5)$$

where $R_i = \{j : T_j^0 < T_i \le T_j\}$ is the **risk set** just prior to T_i. The partial likelihood is the product of these **conditional probabilities** over all failure times: $L(\boldsymbol{\beta}) = \prod_i L_i(\boldsymbol{\beta})$. Notice that the partial likelihood is a function of $\boldsymbol{\beta}$ only – it does not depend on the baseline hazard $\lambda_0(\cdot)$. With certain types of censoring (or no censoring) the partial likelihood is just the marginal **likelihood** of the ranks of the failure times. If there are ties in the data (two or more individuals failing at the same time), then both the partial likelihood and the marginal likelihood become difficult computationally [50, pp. 74–78]. Instead, most packages use an approximation [13, 63]:

$$L_i(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{S}_i)}{\left[\sum_{j\in R_i}\exp(\boldsymbol{\beta}'\mathbf{Z}_j)\right]^{d_i}}, \qquad (6)$$

where d_i is the number of individuals failing at T_i and $\mathbf{S}_i$ is the sum of the $\mathbf{Z}_j$ for these d_i individuals. The approximation is reasonable provided the number of ties at any failure time is small compared to the number in the risk set. Note that i indexes the N distinct failure times, whereas j indexes the n individuals ($n \ge N$).

It is standard practice to maximize the partial likelihood using Newton–Raphson to find a $\boldsymbol{\beta}$ at which the derivative of its logarithm is zero (*see* **Optimization and Nonlinear Equations**). Indeed, Jacobsen [47] has shown that, when the relative risk function $r(\boldsymbol{\beta}'\mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z})$, $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$ is concave. (It is strictly concave provided there is no exact **collinearity** among the explanatory variables and that no linear combination of the variables is a perfect predictor of failure. The latter would imply an infinite observed hazard ratio.)

We use the following notation: let

$$\mathbf{S}^{(k)}(\boldsymbol{\beta}, T_i) = \sum_{j\in R_i}\mathbf{Z}^{\otimes k}\exp(\boldsymbol{\beta}'\mathbf{Z}_j),$$

where $\mathbf{Z}^{\otimes 0} = 1$, $\mathbf{Z}^{\otimes 1} = \mathbf{Z}$, and $\mathbf{Z}^{\otimes 2} = \mathbf{Z}\mathbf{Z}'$. Let $\mathbf{U}(\boldsymbol{\beta})$ denote the score

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_i \frac{\mathrm{d}\log L_i(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}} = \sum_i \left[\mathbf{S}_i - d_i\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, T_i)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, T_i)}\right], \qquad (7)$$

and $\mathbf{I}(\boldsymbol{\beta})$ minus the Hessian:

$$\mathbf{I}(\boldsymbol{\beta}) = -\frac{\mathrm{d}\boldsymbol{U}(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}} = \sum_i d_i \times \left\{\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}, T_i)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, T_i)} - \left[\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, T_i)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, T_i)}\right]^{\otimes 2}\right\}. \qquad (8)$$

Given an estimate $\boldsymbol{\beta}^{(m)}$, one step of the algorithm gives

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \mathbf{I}(\boldsymbol{\beta}^{(m)})^{-1}\boldsymbol{U}(\boldsymbol{\beta}^{(m)}).$$

The algorithm is generally started from $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and convergence is determined by the magnitude of $|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}|$.

When there are S strata, one considers those at risk in each stratum separately. Let R_{si} denote the set of indices of individuals in stratum s at risk at time T_i, and let $L_{si}(\boldsymbol{\beta})$ be the partial likelihood contribution from stratum s and time T_i. Note that $\mathbf{S}_{si}$ is the sum of the $\mathbf{Z}_j$ of the d_{si} individuals in stratum s who fail at time T_i. The partial likelihood is then simply the product of the stratum specific partial likelihoods:

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{S}\prod_i L_{si}(\boldsymbol{\beta}).$$

Although the partial likelihood is not in general a likelihood, it is usually treated as such. It is standard practice to report the value of the logarithm of the partial likelihood and to compare the partial likelihood ratio statistic to a **chi-square distribution** for testing between nested regression models (*see* **Likelihood Ratio Tests**). Similarly, the covariance of $\hat{\boldsymbol{\beta}}$

is estimated by $\mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}$ and score tests (*see* **Likelihood**) are based on $\mathbf{U}(\mathbf{0})'\mathbf{I}(\mathbf{0})^{-1}\mathbf{U}(\mathbf{0})$. Indeed, in the absence of ties ($d_{si} = 1$ for all s and i), the score test from the Cox model with $K - 1$ dummy variables corresponding to a factor with K levels is identical to the K-sample log rank test. Further, the stratified log rank test is identical to the score test from the stratified Cox model.

Having computed $\hat{\boldsymbol{\beta}}$, the estimated regression coefficients, one can calculate the Breslow estimate of the cumulative baseline hazard [13] explicitly. The estimator for stratum s is

$$\hat{\Lambda}_{s0}(t) = \sum_{i:T_i \le t} \frac{d_{si}}{\sum_{j \in R_{si}} \exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_j)}. \tag{9}$$

Estimation of the hazard function itself can be done by taking a smooth derivative of the cumulative hazard. This is usually achieved by the kernel method [68] (*see* **Density Estimation**). The jumps in the Breslow estimate should not be used without **smoothing**. The jump at T_i crudely approximates $\lambda_0(T_i)(T_i - T_{i-1})$ not $\lambda_0(T_i)$. Breslow [13] also showed that the maximum partial likelihood estimate of $\boldsymbol{\beta}$ and the estimated cumulative baseline hazard, (5), can also be obtained by maximizing the full likelihood for $\boldsymbol{\beta}$ and Λ_0 simultaneously, assuming that Λ_0 is piecewise linear **spline**, i.e. the hazard $\lambda_0(t)$ is constant between each pair of ordered failure times. This heuristic argument was made precise by Johansen [48]. He showed that, in certain circumstances, the partial likelihood is formally the profile likelihood for $\boldsymbol{\beta}$. He permitted Λ_0 to be a step function and assumed that at the jumps $\mathrm{d}\Lambda(t|\mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z})\,\mathrm{d}\Lambda_0(t)$.

During the 1970s anyone wishing to fit a Cox model had to use a stand-alone computer program such as the FORTRAN code provided in the book by Kalbfleisch & Prentice [50]. Today, however, the situation is very different and there are many commercially available general statistical packages that will fit a Cox model to large data sets (*see* **Software, Biostatistical**).

Asymptotic Properties

The large sample properties of the maximum partial likelihood estimator of $\boldsymbol{\beta}$ and of the Breslow estimator of Λ_0 are unsurprising, but proofs of these results took some time. When the Cox model holds with parameters $\boldsymbol{\beta}_0$ (and Λ_0), the distribution of $\hat{\boldsymbol{\beta}}$ can be approximated by **multivariate normal** with mean $\boldsymbol{\beta}_0$ and a **covariance matrix** that can be estimated by $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$.

Two quite different approaches were successful. The first due to Tsiatis [83] was to consider independent and identically distributed triples $(X_i, \mathbf{Z}_i, C_i)$, where X_i is the failure time and C_i is the censoring time. It is assumed that the X_i are generated from a Cox model with covariates $\mathbf{Z}_i$ and that X_i are conditionally independent of C_i given $\mathbf{Z}_i$. The observed data are $(T_i, \mathbf{Z}_i, D_i)$, $i = 1, \ldots, n$, where $T_i = \min(X_i, C_i)$ and $D_i = 1$ if $T_i = X_i$ (the event is observed), and $D_i = 0$ otherwise (the event is censored). The estimators are functionals of the observed data, and classical large sample theory is applied. Under this model it can be shown that

$$\frac{\mathbf{S}^{(1)}(\hat{\boldsymbol{\beta}}, t)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, t)} \to \mathbf{E}(\mathbf{Z}|T = t, D = 1)$$

and that $\mathbf{I}(\hat{\boldsymbol{\beta}})/n \to \mathbf{E}[D\mathrm{var}(\mathbf{Z}|T, D)]$ [70]. By viewing the estimators as functionals of the empirical distribution of the unobserved triples and using results from the theory of empirical processes, it is possible to study the large sample properties of the Cox estimators even when the data come from some other model [72].

The other approach to large sample theory using a martingale **central limit theory** requires reformulating the model. This approach adds much insight to the model and will be outlined here. The counting process view of survival analysis is due to Aalen [1]. Andersen & Gill [5] redefined the Cox model and provided elegant proofs of its large-sample properties under mild regularity conditions.

Counting Process Formulation

A multivariate counting process

$$N = \{N_i(t) : 0 \le t < \infty; i = 1, \ldots, n\}$$

is a nondecreasing integer-valued stochastic process with n components. It is assumed that $N_i(0) = 0$ for all i and that the jumps are all of size +1. The process may count the number of events that have occurred in each of n individuals by time t. If the event is the death of a person, then $N_i(t) \in \{0, 1\}$ since people only die once! For technical reasons, N_i is taken to be

right continuous (so that $N_i(t)$ represents the number of events in $[0, t]$) and no two components of N jump at the same time.

Associated with such a counting process is a cumulative intensity process A with components defined by

$$A_i(t + \mathrm{d}t) - A_i(t) = \Pr\{N_i(t + \mathrm{d}t) - N_i(t) = 1 | \mathcal{F}_{t-}\},$$

where $\mathcal{F}_{t-}$ represents everything that has happened until just before t. The history $\mathcal{F}_{t-}$ will certainly include the paths of $N_j(\cdot)$ on $[0, t)$, $j = 1, \ldots, n$, and may include other information such as censoring or explanatory variables from $[0, t)$. $M = N - A$ is a multivariate martingale with respect to the history (filtration) $\{\mathcal{F}_t : t \geq 0\}$. The Andersen & Gill [7] generalization of the Cox model is that

$$A_i(t + \mathrm{d}t) - A_i(t) = \alpha_i(t)\,\mathrm{d}t = Y_i(t)\lambda_0(t) \times \exp[\boldsymbol{\beta}_0' \mathbf{Z}_i(t)]\,\mathrm{d}t,$$

where $Y_i(t)$ is equal to 1 if individual i is under observation just before time t, and is equal to 0 otherwise. $Y_i(\cdot)$ is called the ith "at-risk" indicator process. Here we are assuming that the process A is absolutely continuous with derivative α. Note that we have written the explanatory variables as processes depending on t, and that the definition of the intensity process requires $\{\mathbf{Z}_i(u) : 0 \leq u \leq t, i = 1, \ldots, n\}$ to be in the history $\mathcal{F}_{t-}$. This means that the value of $\mathbf{Z}(t)$ should be known just before t.

The classical Cox model corresponds to a very simple counting process, each component of which jumps at most once. We have

$$N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$$

and

$$Y_i(t) = I\{X_i \geq t, C_i \geq t\} = I\{T_i \geq t\}.$$

N_i starts at 0 and jumps to one when individual i is observed to die. If individual i is censored, N_i remains 0 for ever. Recall that $\alpha_i(t)\,\mathrm{d}t$ is the probability of N_i jumping in the interval $[t, t + \mathrm{d}t]$. If individual i has died or been censored before time t, then there is no chance of observing a death in the interval $[t, t + \mathrm{d}t]$, so $\alpha_i(t) = 0$. Otherwise $\alpha_i(t) = \lambda(t|\mathbf{Z}_i)$ by the definition of the hazard function. Hence in general $\alpha_i(t) = Y_i(t)\lambda(t|\mathbf{Z}_i)$.

Using the new notation, we define the log partial likelihood using information up to time u as

$$l(\boldsymbol{\beta}, u) = \int_0^u \sum_{i=1}^n \left(\boldsymbol{\beta}' \mathbf{Z}_i(t)\,\mathrm{d}N_i(t) - \log\left\{ \sum_{j=1}^n Y_j(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t)] \right\} \mathrm{d}N_i(t) \right).$$

Note that $\mathrm{d}N_i(t)$ is equal to either 0 or 1, because N_i is a counting process. Thus integration with respect to $\mathrm{d}N_i(t)$ is simple: in the classical Cox model $\int f(t)\,\mathrm{d}N_i(t) = D_i f(T_i)$. Differentiate l with respect to $\boldsymbol{\beta}$ to get the score process

$$\mathbf{U}(\boldsymbol{\beta}, u) = \int_0^u \sum_{i=1}^n [\mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}, t)]\,\mathrm{d}N_i(t),$$

where

$$\mathbf{E}(\boldsymbol{\beta}, t) = \frac{\sum_{j=1}^n Y_j(t) \mathbf{Z}_j(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t)]}{\sum_{j=1}^n Y_j(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t)]}. \quad (10)$$

It is easy to show that at the true $\boldsymbol{\beta}$, integration with respect to the intensity process is identically zero (for all u). Hence, at $\boldsymbol{\beta}_0$, one may replace $\mathrm{d}N_i(t)$ by $\mathrm{d}M_i(t)$:

$$\mathbf{U}(\boldsymbol{\beta}_0, t) = \int_0^u \sum_{i=1}^n [\mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}_0, t)]\,\mathrm{d}M_i(t).$$

It follows from the theory of martingale transforms that $\mathbf{U}(\boldsymbol{\beta}_0, \cdot)$ is a martingale since the integrand $[\mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}_0, t)]$ is predictable (i.e. its value is known just prior to t). Under mild regularity conditions [5] one can apply a martingale central limit theorem to show that $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0, \cdot)$ converges in distribution to a Gaussian process.

Extending the counting process notation in the obvious way to permit strata, so that, for instance, $Y_{si}(u)$ indicates whether individual i is at risk in

stratum s at time u, the Breslow estimator is

$$\hat{\Lambda}_{s0}(t) = \int_0^t \frac{\sum_{i=1}^{n} \mathrm{d}N_{si}(u)}{\sum_{i=1}^{n} Y_{si}(u) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(u)]}$$

$$= \int_0^t \sum_{i=1}^{n} \mathrm{d}N_{si}(u) / S_s^{(0)}(\hat{\boldsymbol{\beta}}, u)$$

Let $J_s(t) = I\{\sum_{i=1}^{n} Y_{si}(t)\} > 0$. Then

$$\int_0^t J_s(u)\, \mathrm{d}[\hat{\Lambda}_{s0}(u) - \Lambda_{s0}(u)] = \int_0^t \frac{J_s(u)}{S_s^{(0)}(\hat{\boldsymbol{\beta}}, u)}$$

$$\times \sum_{i=1}^{n} \{\mathrm{d}N_{si}(u) - Y_{si}(u) \exp[\boldsymbol{\beta}_0' \mathbf{Z}_i(u)]\, \mathrm{d}\Lambda_{s0}(u)\}$$

$$= \int_0^t \frac{J_s(u)}{S_s^{(0)}(\hat{\boldsymbol{\beta}}, u)} \sum_{i=1}^{n} \mathrm{d}M_{si}(u).$$

Thus, once again, the asymptotics can be proved using a martingale central limit theorem.

Time-Dependent Explanatory Variables

The possibility of including explanatory variables that change with time was realized by Cox in his original article [21]. There it is suggested that the inclusion of a user-defined variable $Z_2(t) = tZ_1$ might be used as a test of the proportional hazards assumption. Other authors have included explanatory variables that change value at possibly random times. The classical example of this sort of covariate is one that indicates whether a patient has received a heart transplant before time t [25]. The uses and interpretations of these two types of time-dependent variables are quite different. In this section they will be discussed relying heavily on the ideas presented by Kalbfleisch & Prentice [50].

External or Ancillary Variables

An external variable is one that is not affected by the failure process. The simplest sort of external variable is a fixed or time-independent one. A second type is a defined variable such as $Z_2(t) = tZ_1$. Although Z_2 is not fixed, its entire path is known from the outset. A more general example of an external variable is a measure of air pollution as a predictor of severe asthma attacks. Although the level of air pollution is not known in advance, it is "external" to the individuals in the study. Furthermore, the marginal distribution of the variable does not involve the parameters of the failure time model. The whole history of an external variable can be included in $\mathcal{F}_0$ and the hazard or intensity process can be related to the survival function $\Pr(T \geq t|\mathcal{F}_0)$ in the usual way.

Internal Variables

An internal explanatory variable is the output of a **stochastic process** that is generated by the individual under study and so is observed only so long as the individual survives and is uncensored [50]. An example might be the level of β-2 microglobulin in a patient's sera. In practice, the actual level at any given time will be unknown. Instead one uses the level as measured in the most recent blood sample. Typically blood will be taken at most a dozen times during a trial. In such circumstances, the term "updated" may be preferred to "time-dependent".

The key point is that although one may include the history of an internal process up to time t in the filtration $\mathcal{F}_t$ and so define the hazard or intensity function, the intensity function is itself a random process and is not simply a function of the survival function. In general survival from u to t depends on $\{Z(s) : u \leq s \leq t\}$ and this is unknown at u. Furthermore, if Z is only observed when an individual is alive, then $\Pr(T \geq t|Z(t)$ is not missing$) = 1$. Thus it is not possible to make predictions of survival from models that include internal explanatory variables. To do that one must jointly model the survival process and the explanatory variable trajectory.

In a clinical trial with primary focus on a treatment which is fixed by randomization at time 0, internal variables may change in response to treatment. If the effect of treatment is predominantly reflected in the changing value of the explanatory variable, a Cox model of survival that includes both treatment and the updated measurements of the explanatory variable will show little or no treatment differences. Clearly, then, one must be very careful when interpreting the output of a Cox model that includes an internal explanatory variable. Treatment differences in a

model that includes the values of explanatory variables only at time 0 may be inferred to be causative (because of **randomization**). When a large treatment difference is attenuated by inclusion of updated measurements of an internal variable, one may gain useful insights into the mechanism through which the treatment is effective. In such circumstances, it is sensible to also explore the effect of treatment on the internal variable directly.

As with censoring, the value of a variable may depend on the history of the trial so far, without depending on the history of a given individual. Thus, for instance, one might decide to change the environment of a controlled experiment after every 15 deaths. Such a variable is neither internal nor external, but for the purpose of making inference it is closer to an external process.

Computing with Time-Dependent Variables

There are many practical issues in fitting models using time-varying regressors, such as how to deal with missing values, that are not discussed here [3].

The Cox model does not distinguish between a single individual who enters a trial at time 0 and dies at time T_i with fixed regressors $\mathbf{Z}_i$, from two individuals both with regressors $\mathbf{Z}_i$ one of whom enters at time 0 and is censored at time u and one of whom enters at u and dies at T_i. This may sound surprising, but it is true; the likelihood contributions from $[0, u]$ and $(u, T_i]$ are $\Pr(X > u|\mathbf{Z}_i)$, and

$$\frac{\Pr(T_i \le X < T_i + \mathrm{d}t | X > u, \mathbf{Z}_i)}{\mathrm{d}t} = \frac{\Pr(T_i \le X < T_i + \mathrm{d}t|\mathbf{Z}_i)/\mathrm{d}t}{\Pr(X > u|\mathbf{Z}_i)},$$

respectively. Furthermore, in the partial likelihood, all that matters is the $\mathbf{Z}$ values of the members of the risk set at each failure time, not whether a given individual happens to appear in several different risk sets. Thus, if $\mathbf{Z}(t)$ is only updated at a few times per person, it is simplest to treat each person as several "individuals" each with a time fixed covariate. Let the vector $(T_0, T, D, \mathbf{Z})$ denote the entry and exit times, the censoring indicator, and the value of $\mathbf{Z}(t)$ for $t \in (T_0, T]$, respectively. Then an individual who enters a trial at time 0 with $Z(t) = -2$ for $0 \le t \le 1$, $Z(t) = -3$ for $1 < t \le 2$, $Z(t) = 2.5$ for $2 < t \le 3$, and $Z(t) = 2$ for $3 < t \le 3.6$ and dies at $T = 3.6$ is represented by the four data points (0, 1, 0, −2), (1, 2, 0, −3), (2, 3, 0, 2.5), and (3, 3.6, 1, 2).

When computing the likelihood with fixed regressors, it makes sense to use an updating formula. As one moves from one time point to the next, the risk set changes slightly due to the entry or the exit (due to death or censoring) of "individuals". The values for those "individuals" who remain in the risk set do not change and need not be recalculated. In this way the calculation is kept to order n (albeit $4n$ if each individual is treated as four because of changing covariate values).

By contrast, when using continuously varying regressors, one has no choice but to recalculate the partial likelihood contribution from each time point from scratch. This makes the calculation order n^2.

Many software packages that will handle updated regressors will not (easily) handle continuously varying regressors. It is difficult to fit models with user-defined variables such as $Z_2(t) = tZ_1$ using such packages. One might wish to compare the models with hazards $\lambda_0(t)\exp(\beta_1 Z_1)$ and $\lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 t Z_1)$. Of course, for the purpose of testing $\beta_2 = 0$, it is not necessary to fit the latter model. Instead, one may calculate the score statistic for $\beta_2 = 0$ evaluated at the maximum partial likelihood estimate of β_1 from the model with the single (fixed) regressor.

Model Checking

An important aspect of modeling any set of data is assessing the adequacy of the fit and checking to see that the resulting inference is not unduly influenced by a few observations. In general the iterative process of model building and checking may be considered an art rather than a science. Here we review some of the tools available to the statistical artisan analyzing survival data by means of a Cox model.

The simplest form of graphical check comes from dividing the data into groups based on some explanatory variable and fitting a stratified Cox model. If the explanatory variable "$Z = s$" is well modeled by the Cox model, one has $\Lambda_{s0} = \Lambda_0 \exp(\gamma s)$, say. Thus, plotting the logarithm of the cumulative hazard estimate from each strata should reveal parallel curves. That is, the vertical distance between the two curves $\log \Lambda_{r0}(t)$ and $\log \Lambda_{s0}(t)$ should be the same for all t. The common distance should be $\gamma(r - s)$. In practice, such graphics, while intuitively appealing, are not particularly useful.

A closely related, but rather more useful, graph for two strata is obtained by plotting one cumulative hazard $\Lambda_{r0}(t)$ against the other $\Lambda_{s0}(t)$ for all or some selected values of t. Under proportional hazards, such an H–H plot should approximate a straight line through the origin with slope $\exp(\gamma s)$ [6, Section VII.3.1]. The method is easily extended to multiple strata. The disadvantages of the H–H plot are that they do not record the actual time t, and that, if the proportional hazards assumption is seen to be violated, it is difficult to know how to modify the proportional hazards model other than by using a stratified model. Hess [42] reviews a number of variants on these two simple graphical checks of proportional hazards and compares eight graphical methods on each of three data sets. He recommends smoothed plots of scaled Schoenfeld residuals. These are described in the subsection on residuals.

Goodness-of-Fit Tests

Several authors have developed formal goodness-of-fit tests. These can be divided into those designed to be able to detect global alternatives and those with greater power at detecting some specified alternative. Virtually all the tests are asymptotically equivalent to tests based on a defined time-dependent explanatory variable. We saw earlier that the first such tests were proposed by Cox himself [21]. One may add an additional regressor $Z_*(t) = Zg(t)$ for some function $g(t)$. Common choices for g included the identity function $g(t) = t$ and its logarithmic transform $g(t) = \log t$. Other authors use step functions that may jump at either a fixed or a random (but predictable) time. If the partial likelihood is maximized with $Z(t)$, then the partial likelihood ratio test is the statistic of choice. But for testing the goodness of fit, the score test is simpler to compute because it does not require fitting a model with a time-dependent regressor.

Gill & Schumacher [34] proposed a family of tests of the proportional hazards assumption between two samples, A and B. Their tests are motivated by comparing two different estimates of the relative hazard between the two samples. Under proportional hazards the two estimates will be similar, but they need not be in general. The estimates of relative hazard used are derived from **linear rank tests**, which are themselves equivalent to score tests from the Cox partial likelihood with specially defined time-dependent regressors [33]. The family of tests proposed by Gill & Schumacher [34] are thus similar in spirit to those proposed by Breslow et al. [14]. The latter consider the score test for $\beta_2 = 0$ in the model

$$\lambda_B(t) = \lambda_A(t) \exp[\beta_1 + \beta_2 g(t)],$$

corresponding to covariates $Z_1 = I(B)$ and $Z_2(t) = I(B)g(t)$. A popular choice is $g(t) = \hat{S}(t)$, the Kaplan–Meier estimate of survival in the combined sample at t. O'Quigley & Pessione [62] suggest using a step function for $g(t)$. For a one degree of freedom test, one must choose both the cut points and the values of the step function. For a more general alternative hypothesis, one could partition the time axis into J intervals. The null hypothesis is that the relative hazards $\exp \beta_j$ in all $j = 1, \ldots, J$ intervals are the same, and this can be tested with $J - 1$ degrees of freedom. Wei [85] proposes an omnibus goodness-of-fit test for the two-sample problem based on the supremum of the score statistic $\sup_t |U(\hat{\beta}, t)|$.

Schoenfeld [75] was interested in a more general goodness-of-fit test for the Cox model. He suggested embedding a Cox model with regressor Z in a much larger model with regressors Z and $\mathbf{Z}_*(t)$, where the $\mathbf{Z}_*(t)$ are a set of indicator variables that partition the regressor–time space. Thus, for instance, one might divide the time axis into three parts and the covariate space into four, and form the Cartesian product with 12 cells. In addition to the score test for the coefficients of $\mathbf{Z}_*$ being all zero, one can examine the "residuals", i.e. the difference between the observed and expected (under the basic model with covariate Z) number of events in each of the 12 cells. Lin et al. [58] avoid the need for an arbitrary partition of the space by deriving a supremum test based on the cumulative sum

$$W(t, z) = \sum_{Z_i \le z} [O_i(t) - E_i(t)],$$

where $O_i(t) = N_i(t)$ and $E_i(t) = \int_0^t Y_i(u)\, \mathrm{d}\hat{\Lambda}_i(t)$ are, respectively, the observed and expected number of events in individual i, by time t.

Residuals

There have been numerous attempts to define residuals and to propose diagnostic plots for the Cox model

(*see* **Diagnostics**). The situation is complicated by both the **semiparametric model** and the presence of censoring. Some of the proposed techniques are decidedly less useful than one might have hoped. In particular, attempts to define residuals that (under the Cox model) look like a random sample from a specified distribution, so that $Q-Q$ plots can be drawn (*see* **Normal Scores**), have failed. Graphical assessment of the functional form of a covariate and of the constancy of the regression parameters over time have been more successful.

An early definition of residual for the Cox model was the estimated cumulative intensity for each individual:

$$\hat{A}_i(\infty) = \int_0^\infty Y_i(u)\,\mathrm{d}\hat{\Lambda}_i(u) = \int_0^\infty Y_i(u)\exp[\hat{\boldsymbol{\beta}}'\mathbf{z}_i(u)]\,\mathrm{d}\hat{\Lambda}_0(u) \quad (11)$$

[if $Y_i(u) = I(T_i \geq u)$ and $\mathbf{z}_i(u) = \mathbf{z}_i$, then $\hat{A}_i(\infty) = \hat{\Lambda}_i(T_i) = \exp(\hat{\boldsymbol{\beta}}'\mathbf{z}_i)\hat{\Lambda}_0(T_i)$] [24, 52]. Later authors made an adjustment to the residual depending on whether the individual was censored or not. The resulting residual $r_i = D_i - \hat{A}_i(\infty)$ is called the martingale residual and is a special case of the general family of residual processes defined by

$$\int_0^t H_i(u)\,\mathrm{d}\hat{M}_i(t), \quad (12)$$

where $\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\,\mathrm{d}\hat{\Lambda}_i(u)$ and H_i is a predictable process [8, 81]. Thus r_i is the estimated martingale transform, (12), with $H_i = 1$ and $t = \infty$. The martingale residual may be thought of as the difference between the observed and the expected number of events for the ith individual. The distribution of martingale residuals in a survival setting is very skewed since they have mean zero (under the true model) but range from 1 (for someone who fails at time 0) to minus a very large number (for someone who survives much longer than "expected"). Summing over individuals with similar covariate values $\{i : \mathbf{z}_i \in \mathcal{Z}\}$, say, one obtains the residual number of events for individuals with $\mathbf{z} \in \mathcal{Z}$. Thus, smoothing the martingale residuals against a regressor (or a potential regressor) gives an indication as to how well the model fits the data. Systematic departures from zero indicate that there is an excess (or deficit) in the modeled hazard for that group of individuals. Heuristically one has

$$\mathbf{E}\{N_i(\infty)|\mathbf{z},z^*\} = A(\infty|\mathbf{z}, z^*) \approx \hat{A}(\infty|\mathbf{z}) + \text{smooth}(r_i|z^*).$$

More recently, Grambsch et al. [36] have considered the model

$$\lambda(t|\mathbf{z},z^*) = \lambda_0(t)\exp[\boldsymbol{\beta}'\mathbf{z} + f(z^*)]. \quad (13)$$

They propose fitting the Cox model with prognostic index $\boldsymbol{\beta}'\mathbf{z} + \gamma z^*$ and plotting $\log\{\text{smooth}[N_i(\infty)]\} - \log\{\text{smooth}\,[\hat{A}_i(\infty)]\} + \hat{\gamma}z^*$ vs. z^*. The smooth curve will approximate $f(z^*)$ to first order. In practice, the approximation seems to work well even when Z^* is correlated with the other regressors $\mathbf{Z}$.

The martingale residuals were defined by integrating the martingale difference array $\mathrm{d}\hat{M}_i(t)$ over the time axis to give a single residual per individual. To examine the proportional hazards assumption, one is more interested in obtaining a separate residual for each failure time. This can be done by summing the martingale differences, at a given time, over all individuals. Now $\Sigma_i\,\mathrm{d}\hat{M}_i(t) = 0$ for all t by the definition of the Breslow estimator $\hat{\Lambda}_0$. Nevertheless, one can use the martingale transform, (12), with $\mathbf{H}_i = \mathbf{Z}_i$. Then at each failure time one is comparing the observed value of $\mathbf{Z}$ in the individual that fails with its expected value. Such a residual,

$$\mathbf{r}^*(T_j) = \sum_i \mathbf{Z}_i(T_j)[\,\mathrm{d}N_i(T_j) - Y_i(T_j)\,\mathrm{d}\hat{\Lambda}_i(T_j)] = \mathbf{S}_j - d_j\frac{\mathbf{S}^{(1)}(\hat{\boldsymbol{\beta}}, T_j)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, T_j)},$$

was first proposed by Schoenfeld [76]. It is seen that the sum of the Schoenfeld residuals evaluated at $\boldsymbol{\beta}$ is equal to the score $\mathbf{U}(\boldsymbol{\beta})$. It is not difficult to show that, even under the model

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[\boldsymbol{\beta}(t)'\mathbf{z}], \quad (14)$$

$\mathbf{S}^{(1)}\,[\boldsymbol{\beta}(t), t]/\boldsymbol{S}^{(0)}\,[\boldsymbol{\beta}(t), t] \rightarrow \mathbf{E}(\mathbf{Z}|T = t,\ D = 1)$. Thus, using a one-step Taylor series expansion about $\boldsymbol{\beta}(t) = \hat{\boldsymbol{\beta}}$, one has

$$\boldsymbol{\beta}(t) \approx \hat{\boldsymbol{\beta}} + \hat{\mathbf{V}}(t)^{-1}\mathbf{r}^*(t),$$

where

$$\hat{\mathbf{V}}(t) = \frac{\mathbf{S}^{(2)}(\hat{\boldsymbol{\beta}}, t)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, t)} - \left(\frac{\mathbf{S}^{(1)}(\hat{\boldsymbol{\beta}}, t)}{\mathbf{S}^{(0)}(\hat{\boldsymbol{\beta}}, t)}\right)^{\otimes 2}.$$

Hence, Grambsch & Therneau [35] have proposed plotting a smooth of $\hat{\boldsymbol{\beta}} + \hat{\mathbf{V}}(t)^{-1}\mathbf{r}^*(t)$ against t in order to get a feel of $\boldsymbol{\beta}(t)$. Often $\mathbf{V}(t) = \lim_{n\to\infty} \hat{\mathbf{V}}(t)$ does not vary much as a function of t, so for exploratory purposes it may be enough to use $\mathbf{I}(\hat{\beta})/\Sigma N_i(\infty)$ in place of $\hat{\mathbf{V}}(t)$. This has the advantage of not having to store and invert a different covariance matrix at each failure time. In practice, $\mathbf{V}(t)$ will vary most when a variable Z has a skewed distribution and those in the tail are at greatest risk. In all cases it will be difficult to estimate $\hat{\mathbf{V}}(t)$ if the risk set is small at time t, and it is also for large values of t that the $V(t)$ is most likely to be substantially different from its average value.

Influence Diagnostics

Various measures of influential observations have been suggested for the Cox model. The influence **diagnostic** is intended to approximate the amount by which the regression estimate $\hat{\boldsymbol{\beta}}$ would change if the ith individual were removed from the data set [69]. One such approximation is the infinitesimal **jackknife**, first proposed by Cain & Lange [17]. Their residuals are equal to the components of the scaled efficient score statistic. This can be written as a martingale transform residual with $\mathbf{H}_i(t) = \mathbf{Z}_i - \mathbf{E}(\hat{\boldsymbol{\beta}}, t)$. The scaling is done by $\mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}$. One has

$$\tilde{\mathbf{r}}_i = I(\hat{\boldsymbol{\beta}})^{-1} \int_0^\infty [\mathbf{z}_i - \mathbf{E}(\hat{\boldsymbol{\beta}}, t)]\, \mathrm{d}\hat{M}_i(t).$$

An alternative estimate of the influence of an individual is given by Storer & Crowley [79].

Alternatives and Extensions

We have already discussed many extensions of the basic Cox model. We have permitted nonproportional hazards through the stratified Cox model and through user-defined time-dependent variables. We have considered diagnostics to detect data that appear to come from more nonparametric models, such as the additive Cox model $\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[\sum_k f_k(z_k)]$, in which some of the functions f_k may be assumed to be linear while others are left unspecified [32, 37, 39, 70], and the multiplicative hazards model $\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[\boldsymbol{\beta}(t)'\mathbf{z}]$ [30, 40, 41, 86, 84].

We have also seen how the model that was originally perceived for survival data can be generalized quite naturally to event data in which a single individual may have multiple events. The events need not even all be of the same type. They may represent competing risks or more generally the various states in a multistate model. In the classic heart transplant situation, for example, one might use Cox regression to model the transition from identification as a potential recipient (state 0) to transplant (state 1); from state 0 to death (state 2); and from state 1 to death [26]. Three state models in which transitions from state 1 (diseased) back to state 0 (healthy) are possible are also common (see, for example, Andersen et al. [6, Example VII.2.10]). The study of (i) acute graft-vs.-host disease, (ii) chronic graft-vs.-host disease, (iii) leukemia relapse, and (iv) death following bone marrow transplantation (state 0) is also considered by Andersen et al. [6, Example VII.2.18].

We briefly mention a few alternatives to the semiparametric Cox model for regression analysis of censored survival data. Naturally one can try to adapt estimation in any parametric regression model to cope with right censored data. Loglinear models with **Weibull** or **Gamma** errors [50, Section 3.6] tend to be more popular in reliability (engineering) than in biostatistics. Particularly in epidemiology, one sometimes has a known population mortality rate that one wants to use in place of the baseline hazard function. The hazard for individual i is given by

$$\lambda_i(t) = \mu_i(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_i(t)],$$

where μ_i is the population mortality corresponding to individual i [7, 15]. Fully parametric models have been studied using counting process techniques by Borgan [12]. Another Cox-like model that uses a known rate is the proportional excess hazards model [73] (*see* **Excess Risk**),

$$\lambda_i(t) = \mu_i(t) + \lambda_0(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_i(t)],$$

in which the excess mortality is modeled by a Cox model.

A general family, known as the **accelerated failure-time model**, is a linear regression model for

the logarithm of the survival time,

$$\log X = \boldsymbol{\beta}'\mathbf{Z} + \varepsilon,$$

where the error ε may be either from a specified distribution or from an unknown distribution. Gaussian errors and no censoring simply correspond to linear regression of $\log X$. If the errors are Weibull, then the model is also a proportional hazards model. Theoretical attention has focused on the semiparametric model with unknown error distribution. Another family of models that include the Cox model as a special case are the transformation models in which an unknown monotone transformation of the survival time is assumed to have a linear regression:

$$\psi(X) = \boldsymbol{\beta}'\mathbf{Z} + \varepsilon.$$

If the error distribution is **extreme value** ($\exp\varepsilon$ distributed exponential with mean 1), then the transformation model is a Cox model with cumulative baseline hazard given by $\exp[\psi(t)]$ and regression parameter $s - \boldsymbol{\beta}$.

The Cox model is a multiplicative hazards model. Aalen [2] introduced an additive hazards model (*see* **Aalen's Additive Regression Model**). A semiparametric version of the model [61] is given by

$$\lambda(t|\mathbf{Z}_1, \mathbf{Z}_2) = \boldsymbol{\theta}_1(t)'\mathbf{Z}_1(t) + \boldsymbol{\theta}_2'\mathbf{Z}_2(t).$$

If the variables $\mathbf{Z}_1$ include a constant, then we may pull out a baseline hazard and write the first term on the right of the equation as $\lambda_0(t) + \alpha_{11}(t)Z_{11}(t) + \cdots + \alpha_{1p}(t)Z_{1p}(t)$. The cumulative components of hazard $A_{1j}(t) = \int_0^t \alpha_{1j}(u)\,\mathrm{d}u$ can be estimated at parametric rates, and these must be smoothed to estimate the $\alpha_{1j}(u)$. The model extends naturally to the more general counting process formulation.

Several authors have considered "special" Cox models. These include models for matched pairs [38, 44] (*see* **Matching**), and for **interval censored** survival data [46], a model for periodic data [64] (*see* **Seasonal Time Series**), and a model for **case–cohort** data [66, 77]. **Bayesian analysis** of the Cox model was first considered by Kalbfleisch [49; [50], Section 8.4] and later by Hjort [43].

There has been relatively little written about robust estimation in the Cox model (*see* **Robustness**). Estimators that maximize a weighted partial likelihood have been proposed independently at least three times [56, 71, 72, 74]. The weights may be random and may depend on the regressors $\mathbf{Z}$, but they should be predictable (or at least asymptotically equivalent to predictable weights). A slightly different estimator which essentially corresponds to the efficient score function from a weighted full likelihood has also been studied [10].

Consideration of the Cox estimator for $\hat{\boldsymbol{\beta}}$ when the data do not come from a Cox model leads naturally to adoption of the sandwich estimator of the variance of $\hat{\boldsymbol{\beta}}$ [57, 69]. This is the usual infinitesimal jackknife estimator that can be obtained from the influence residuals

$$\tilde{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \sum_i \tilde{r}_i \tilde{r}_i'.$$

The estimator is perhaps most useful when the data are clustered (*see* **Clustering**). Suppose that $\tilde{\mathbf{r}}_{ki}$ is the influence residual from individual i in cluster k. Then define $\tilde{\mathbf{r}}_k = \Sigma_i \tilde{\mathbf{r}}_{ki}$ and estimate the variance of $\hat{\boldsymbol{\beta}}$ by $\Sigma_k \tilde{\mathbf{r}}_k^{\otimes 2}$ [55]. This may be a simple technique for adjusting inference when using the Cox model with multivariate survival data. For instance, if each person could have several events, then one might wish to treat the person as a cluster. In another example, the clusters might be formed from survival data on individuals within families.

Another approach adapting the Cox model to multivariate data is through latent variables or **frailties**. The idea is that, conditionally on an unobserved variable or frailty, the survival times follow a Cox model. The value of the frailty W_i is assumed to be the same for all survival times within a cluster. Two frailty distributions have received the most attention: Clayton & Cuzick [19] considered the hazard model $\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z} + W)$ in which $\exp W$ has a gamma distribution; Hougaard [45] favors using the positive stable distribution, as this is the only choice that yields proportional hazards both marginally (integrating over the unobserved variable) and conditionally.

References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[2] Aalen, O.O. (1980). A Model for Nonparametric Regression Analysis of Counting Processes, *Springer Lecture Notes in Statistics*, Vol. 2. Springer-Verlag, New York, pp. 1–25.

[3] Altman, D.G. & De Stavola, B.L. (1994). Practical problems in fitting a proportional hazards model to data

with updated measurements of the covariates, *Statistics in Medicine* **13**, 301–341.

[4] Andersen, P.K. (1991). Survival analysis 1982–1991: the second decade of the proportional hazards regression model, *Statistics in Medicine* **10**, 1931–1941.

[5] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[6] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[7] Andersen, P.K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N. & Kreiner, S. (1985). A Cox regression model for the relative mortality and its application to diabetes mellitus survival data, *Biometrics* **41**, 921–932.

[8] Barlow, W.E. & Prentice, R. (1988). Residuals for relative risk regression, *Biometrika* **75**, 65–74.

[9] Bednarski, T. (1989). On sensitivity of Cox's estimator, *Statistics and Decisions* **7**, 215–228.

[10] Bednarski, T. (1993). Robust estimation in Cox's regression model, *Scandinavian Journal of Statistics* **20**, 213–225.

[11] Begun, J.M., Hall, W.J., Huang, W.M. & Wellner, J.A. (1983). Information and asymptotic efficiency in parametric - nonparametric models, *Annals of Statistics* **11**, 432–452.

[12] Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data, *Scandinavian Journal of Statistics* **11**, 1–16. Correction **11** (1984) 275.

[13] Breslow, N. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–99.

[14] Breslow, N.E., Edler, L. & Berger, J. (1984). A two-sample censored-data rank test for acceleration, *Biometrics* **40**, 1049–1062.

[15] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[16] Bretagnolle, J. & Huber-Carol, C. (1988). Effects of omitting covariates in Cox's regression model for survival data, *Scandinavian Journal of Statistics* **15**, 125–138.

[17] Cain, K.C. & Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data, *Biometrics* **40**, 493–499.

[18] Christensen, E., Neuberger, J., Crowe, J., Altman, D.G., Popper, H., Portmann, B., Doniach, D., Ranek, L., Tygstrup, N. & Williams, R. (1985). Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial, *Gastroenterology* **89**, 1084–1091.

[19] Clayton, D. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion), *Journal of the Royal Statistical Society, Series A* **148**, 82–117.

[20] Collett, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, London.

[21] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[22] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[23] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[24] Cox, D.R. & Snell, E.J. (1968). A general definition of residuals (with discussion), *Journal of the Royal Statistical Society, Series B* **30**, 248–275.

[25] Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data, *Journal of the American Statistical Association* **72**, 27–36.

[26] Crowley, J. & Storer, B.E. (1983). Comment on "A reanalysis of the Stanford heart transplant data", by Aitkin, Laird and Francis, *Journal of the American Statistical Association* **78**, 277–281.

[27] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association* **72**, 557–565.

[28] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[29] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[30] Gamerman, D. (1991). Dynamic Bayesian models for survival data, *Applied Statistics* **40**, 63–79.

[31] Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples, *Biometrika* **52**, 203–223.

[32] Gentleman, R. & Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model, *Biometrics* **47**, 1283–1296.

[33] Gill, R.D. (1984). Understanding Cox's Regression Model: a martingale approach, *Journal of the American Statistical Association* **79**, 441–447.

[34] Gill, R.D. & Schumacher, M. (1987). A simple test for the proportional hazards assumption, *Biometrika* **74**, 289–300.

[35] Grambsch, P.M. & Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika* **81**, 515–526.

[36] Grambsch, P.M., Therneau, T.M. & Fleming, T.R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models, *Biometrics* **51**, 1469–1482.

[37] Gray, R.J. (1992). Flexible methods for analysing survival data using splines, with applications to breast cancer prognosis, *Journal of the American Statistical Association* **87**, 942–951.

[38] Gross, S.T. & Huber, C. (1987). Matched pair experiments: Cox and maximum likelihood estimation, *Scandinavian Journal of Statistics* **14**, 27–41.

[39] Hastie, T. & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model, *Biometrics* **46**, 1005–1016.

[40] Hastie, T. & Tibshirani, R. (1993). Varying coefficient models (with discussion), *Journal of the Royal Statistical Society, Series B* **55**, 757–797.

[41] Hess, K.R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions, *Statistics in Medicine* **13**, 1045–1062.

[42] Hess, K.R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression, *Statistics in Medicine* **14**, 1707–1723.

[43] Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data, *Annals of Statistics* **18**, 1259–1294.

[44] Holt, J.D. & Prentice, R.L. (1974). Survival analysis in twin studies and matched pair experiments, *Biometrika* **65**, 159–166.

[45] Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity, *Biometrika* **71**, 75–83.

[46] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring, *Annals of Statistics* **24**, 540–568.

[47] Jacobsen, M. (1989). Existence and unicity of MLEs in discrete exponential family distributions, *Scandinavian Journal of Statistics* **16**, 335–349.

[48] Johansen, S. (1983). An extension of Cox's regression model, *International Statistical Review* **51**, 165–174.

[49] Kalbfleisch, J.D. (1978). Nonparametric Bayes analysis of survival data, *Journal of the Royal Statistical Society, Series B* **40**, 214–221.

[50] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[51] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[52] Kay, R. (1977). Proportional hazards regression models and the analysis of censored survival data, *Applied Statistics* **26**, 227–237.

[53] Kleinbaum, D.G. (1995). *Survival Analysis: A Self-Learning Text*. Springer-Verlag, New York.

[54] Lagakos, S. & Schoenfeld, D. (1984). Properties of proportional hazards score tests under misspecified regression models, *Biometrics* **40**, 1037–1048.

[55] Lee, E.W., Wei, L.J. & Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in *Survival Analysis: State of the Art*, J.P. Klein, & P.K. Goel, eds. Kluwer, Dordrecht, pp. 237–247.

[56] Lin, D.Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, *Journal of the American Statistical Association* **86**, 725–728.

[57] Lin, D.Y. & Wei, L.J. (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association* **84**, 1074–1078.

[58] Lin, D.Y., Wei, L.J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residual, *Biometrika* **80**, 557–572.

[59] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer and Chemotherapy Reports* **50**, 163–170.

[60] Marubini, E. & Valsecchi, M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley, Chichester.

[61] McKeague, I. & Sasieni, P. (1994). A partly parametric additive risk model, *Biometrika* **81**, 501–514.

[62] O'Quigley J. & Pessione F. (1989). Score tests for homogeneity of regression effects in the proportional hazards model, *Biometrics* **45**, 135–144.

[63] Peto, R. (1972). Contribution to the discussion of paper by D.R. Cox, *Journal of the Royal Statistical Society, Series B* **34**, 205–207.

[64] Pons, O. & de Turckheim, E. (1988). Cox's periodic regression model, *Annals of Statistics* **16**, 678–693.

[65] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[66] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[67] Prentice, R. & Self, S. (1983). Asymptotic distribution theory for Cox-type regression models with general risk form, *Annals of Statistics* **11**, 804–813.

[68] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions, *Annals of Statistics* **11**, 453–466.

[69] Reid, N. & Crépeau, H. (1985). Influence functions for proportional hazards regression, *Biometrika* **72**, 1–9.

[70] Sasieni, P. (1992). Information bounds for the conditional hazard ratio in a nested family of regression models, *Journal of the Royal Statistical Society, Series B* **54**, 617–635.

[71] Sasieni, P.D. (1993). Maximum weighted partial likelihood estimates for the Cox model, *Journal of the American Statistical Association* **88**, 144–152.

[72] Sasieni, P.D. (1993). Some new estimates for Cox regression, *Annals of Statistics* **21**, 1721–1759.

[73] Sasieni, P.D. (1996). Proportional excess hazards, *Biometrika* **83**, 127–141.

[74] Schemper, M. (1992). Cox analysis of survival data with non proportional hazards functions, *Statistician* **41**, 455–465.

[75] Schoenfeld, D. (1980). Chi-squared goodness of fit tests for the proportional hazards regression model, *Biometrika* **67**, 145–153.

[76] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**, 239–241.

[77] Self, S.G. & Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies, *Annals of Statistics* **16**, 64–81.

[78] Solomon, P.J. (1984). Effect of misspecification of regression models in the analysis of survival data, *Biometrika* **71**, 291–298. Amendment. **73** (1986) 245.

[79] Storer, B.E. & Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods, *Journal of the American Statistical Association* **80**, 139–147.

[80] Struthers, C.A. & Kalbfleisch, J.D. (1986). Misspecified proportional hazards models, *Biometrika* **73**, 363–369.

[81] Therneau, T.M., Grambsch, P.M. & Fleming, T.R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**, 147–160.

[82] Thomas, D.C. (1981). General relative risk models for survival time and matched case-control analysis, *Biometrics* **37**, 673–686.

[83] Tsiatis, A.A. (1981). A large sample study of Cox's regression model, *Annals of Statistics* **9**, 93–108.

[84] Verweij, P.J.M. & van Houwelingen, H.C. (1995). Time-dependent effects of fixed covariates in Cox regression, *Biometrics* **51**, 1550–1556.

[85] Wei, L.J. (1984). Testing goodness of fit for the proportional hazards model with censored observations, *Journal of the American Statistical Association* **80**, 139–147.

[86] Zucker, D.M. & Karr, A.F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach, *Annals of Statistics* **18**, 329–353.

PETER SASIENI

Cox, Gertrude Mary

Born: January 13, 1900, in Dayton, Iowa.
Died: October 17, 1978, in Durham, North Carolina.

Gertrude Mary Cox was one of the twentieth century's pioneers in statistics. She wrote the following notes on her early years:

> I was raised on a farm where I had several years for roaming in the woods by the river and over the hills. I learnt from my mother the value and joy of doing for other people. She nursed the sick and raised us to be active church workers. There were four of us, two boys and two girls. We had responsibilities at home. I liked best making the homemade bread for our family because I was allowed to sell one pan of biscuits. My main ambition was to help others so after high school, I took a two-year special social service course of study and worked two years as housemother for 16 little orphan boys in Montana.

The Cox family moved to Perry, Iowa, where Gertrude graduated from high school; her major interests were in arithmetic and mathematics. The social service and orphanage work were in preparation to become a deaconess in the Methodist Episcopal church; however, she finally decided that academic training was more to her liking. She enrolled at Iowa State College (ISC) where she majored in mathematics, but elected courses in psychology, sociology, and craft work and did computing to help pay school expenses, receiving a B.S. degree in 1929.

Gertrude secured a master's degree in statistics from ISC in 1931 (supervised by **George Snedecor**), which was the first in statistics at that institution. She then began work on a Ph.D. in psychological statistics at the University of California at Berkeley; she gave up the latter in 1933 to return to Iowa State to direct the Computing Laboratory of the newly created Statistical Laboratory under Professor Snedecor. She became interested in the design of experiments, in which she developed and taught graduate courses. Her courses were built around a collection of real-life examples in a variety of experimental areas. She taught from mimeographed materials, which formed part of the famous *Experimental Designs* by **W.G. Cochran** and her [4]. She had three major principles in setting up an experiment: (i) the experimenter should clearly set forth his or her objectives before proceeding; (ii) the experiment should be described in detail; and (iii) an outline of the analysis should be drawn up before the experiment is started. She emphasized the role of **randomization** and stressed the need to ascertain if the size of the experiment was sufficient to demonstrate treatment differences if they existed (*see* **Sample Size Determination**).

In 1940, Snedecor responded to a request for suggestions on possible candidates to head the new Department of Experimental Statistics in the School of Agriculture at North Carolina State College (in Raleigh); upon seeing his list of all males, Miss Cox asked why he had not included her name. He then inserted a footnote which stated that if a woman could be considered, he recommended her. This footnote has become a statistical landmark, because Miss Cox was selected. She started staffing her department with statisticians who had majors or strong minors in applied fields. In 1944 the President of the Consolidated University of North Carolina established an all-University Institute of Statistics with Gertrude Cox as head, and in 1945 she obtained funds from the General Education board to establish graduate programs at North Carolina State and in 1946 a new Mathematical Statistics Department at Chapel Hill.

In 1949, Gertrude Cox gave up the Headship at North Carolina State to devote full time to the Institute, including the development of strong statistics programs throughout the South. This latter development was augmented by an arrangement with the Southern Regional Education Board to establish a Committee on Statistics. From 1954 to 1973 the Committee sponsored a continuing series of six-week summer sessions and is now co-sponsoring (with the **American Statistical Association**) a Summer Research Conference.

Of special interest to biostatistics was the founding in 1949 of the Department of Biostatistics in the School of Public Health at UNC, Chapel Hill, chaired by **B.G. Greenberg**. In 1953, the Statistics Section of the **American Public Health Association** and the Biostatistics Department sponsored a Biostatistics Conference on procedures to provide field training for health statisticians.

One of Gertrude Cox's strongest points was her ability to obtain outside financial support. The Rockefeller Foundation supported a strong program in statistical genetics (*see* **Genetic Epidemiology**) at North Carolina State and the Ford Foundation supported one in dynamic economics. She was a strong advocate of the development of powerful computer

programs; North Carolina State was a leader in the use of high-speed computers, especially the IBM650 and the initial SAS programs (*see* **Software, Biostatistical**).

Iowa State bestowed on her an honorary Doctorate of Science in 1958 as a

> stimulating leader in experimental statistics... outstanding teacher, researcher, leader, and administrator... Her influence is worldwide, contributing to the development of national and international organizations, publications, and councils of her field.

Starting in 1958, Dr Cox and other members of the North Carolina State statistics faculty developed procedures to establish a Statistical Division in the not-for-profit Research Triangle Institute (RTI) in the Research Triangle Park (RTP) between Raleigh, Chapel Hill, and Durham; Gertrude Cox retired from the University in 1960 to direct this division. She retired from RTI in 1965, but continued to teach at North Carolina State and consult on research projects. RTP has developed into a world-recognized research park.

Dr Gertrude Cox was a consultant before and after retirement to many organizations, including the **World Health Organization**, the US. Public Health Service, and the government of Thailand, and on a number of US Government committees for the Bureau of the Budget, **National Institutes of Health**, National Science Foundation, Census Bureau, and Agricultural Department. She was a founding member of the **International Biometric Society** in 1947 for which she served as President in 1968–69, Council three times, and first editor of its journal, ***Biometrics***. She was an active member of the **International Statistical Institute** and was President of the American Statistical Association in 1956. Her Presidential address in 1957, "Statistical frontiers", was an affirmation of her ethical concepts of moral uprightness and hard work [5]. She emphasized that

> The fact that you, as an individual, are classified as a statistician does not free you from obligations and responsibilities toward other human beings. You have an obligation to clarify the foundations of your techniques and methods for your clients. I want you young statisticians not to become men of success but rather become men of value.

In 1970, North Carolina State University designated the building in which statistics was housed as Cox Hall, and in 1977 a Gertrude M. Cox Fellowship Fund was established for outstanding graduate students in statistics. Her election to the National Academy of Sciences in 1975 was a treasured recognition of her many contributions.

In 1976 Gertrude learned that she had leukemia but remained sure that she would conquer it up to the end. She even continued construction of a new house, unfortunately not completed until a week after her death. While under treatment at Duke University Hospital she kept detailed records of her progress, and her doctor often referred to them. With characteristic testy humor she called herself "the experimental unit", and died as she had lived, fighting to the end. To those who were fortunate to be with her through so many years, Raleigh will never be the same.

There are published obituaries and biographies of Gertrude Cox [1–3].

References

[1] Anderson, R.L. (1983). Biography of Gertrude Cox, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York.

[2] Anderson, R.L. (1990). Biography of Gertrude Cox, *Biographical Memoirs* **59**, 117–132.

[3] Anderson, R.L., Nelson, L. & Monroe, R. (1979). Obituary of Gertrude Cox, *Biometrics* **35**, 2–7.

[4] Cochran, W.G. & Cox, G. (1950). *Experimental Designs*. Wiley, New York.

[5] Cox, G. (1957). Statistical frontiers, *Journal of the American Statistical Association* **52**, 1–12.

R.L. ANDERSON

Cox's Test of Randomness

There are occasions when it is necessary to determine whether a series of events has occurred at random. Barnard [1] provides a simple test using the result that if the series is random, then the instants of occurrence of events in a finite interval constitute a random sample from a rectangular distribution. Cox [3] gives a test of randomness against the alternative that the series has some trend in the rate of occurrence of the events. Let n events occur at times $t_1, \ldots, t_n$ in the interval $(0, T)$ and assume that Pr[an event occurs in $(t, t + \delta t)] = \lambda(t) + o(\delta t)$, where δt is a small interval of time. Cox demonstrated that, if $\lambda(t) = \alpha \exp(\beta t)$, then an appropriate statistic to test $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ is $\hat{\beta} = \sum_{i=1}^{n} t_i / nT$. The probability distribution of this statistic under H_0 is the Irwin–Hall distribution with mean $\frac{1}{2}$ and variance $1/12n$. Bartholomew [2] shows that for $n \geq 20$ the distribution of $\hat{\beta}$ tends rapidly to normality, and calculates the power function of the test for $n = 5$. Mansfield [4] gives further power function values for $n = 20(10)80, 100, 200$, for both one- and two-tailed tests. Details of test statistics that can be used to detect other types of nonrandomness are given in [3]. For example, a statistic is given for the case where successive intervals between events are assumed to be correlated.

References

[1] Barnard, G.A. (1953). Time intervals between accidents, *Biometrika* **40**, 212–213.

[2] Bartholomew, D.J. (1956). Tests for randomness in a series of events when the alternative is a trend, *Journal of the Royal Statistical Society, Series B* **18**, 234–239.

[3] Cox, D.R. (1995). Some statistical methods connected with series of events (with discussion), *Journal of the Royal Statistical Society, Series B* **17**, 129–157.

[4] Mansfield, E. (1962). Power functions for Cox's test of randomness against trend, *Technometrics* **4**, 430–432.

(*See also* **Durbin–Watson Test**; **Randomness, Tests of**)

CLIVE J. LAWRENCE

Cramér–Rao Inequality

The Cramér–Rao inequality provides a lower bound for the variance of any **unbiased** estimator of a one-dimensional parameter θ in the probability density function (pdf) of the observed **random variable** or, more generally, of a given parametric function $g(\theta)$. The following regularity assumptions are made:

Assumption 1

The relation $\int p(x,\theta)\,\mathrm{d}x = 1$ for the pdf $p(\cdot,\theta)$ of random variable X can be differentiated twice with respect to θ under the integral sign.

Assumption 2

The relation $\int t(x)p(x,\theta)\,\mathrm{d}x = g(\theta)$ for any unbiased estimator $T = t(X)$ of $g(\theta)$ can be differentiated with respect to θ under the integral sign.

Assumption 1 requires that the support of the distribution (i.e. the range of the random variable X) does not depend on θ. If the random variable X above happens to be multi-dimensional, say $(X_1, X_2, \ldots, X_n)$, of continuous components, then the integral in Assumptions 1 and 2 has to be interpreted as an n-dimensional integral. However if X, or its components X_i, are discrete random variables, then the integral in Assumptions 1 and 2 is to be replaced by the summation symbol.

For the unbiased estimator $T = t(X)$ of $g(\theta)$ the basic form of the Cramér–Rao inequality is

$$\mathrm{var}_\theta(T) \geq \frac{[g'(\theta)]^2}{I(\theta)}, \tag{1}$$

due to Cramér [4] and Rao [7]; the lower bound on the right-hand side of the inequality is usually referred to as the *Cramér–Rao lower bound (CRLB)* for the variance of the unbiased estimator of $g(\theta)$. In (1), $g'(\theta)$ is the first derivative of g, and $I(\theta)$ is the Fisher **information** in (the pdf of) X, defined by

$$I(\theta) = \mathrm{E}_\theta\left[\frac{\partial \log p(X,\theta)}{\partial\theta}\right]^2. \tag{2}$$

In the special case $X = (X_1, \ldots, X_n)$, when we have independent and identically distributed components X_i in the *random sample* X, each with pdf $f(\cdot,\theta)$, $I(\theta) = ni(\theta)$, where

$$i(\theta) = \mathrm{E}_\theta\left[\frac{\partial \log f(X_1,\theta)}{\partial\theta}\right]^2, \tag{3}$$

the Cramér–Rao inequality (1) reduces to the form

$$\mathrm{var}_\theta(T) \geq \frac{[g'(\theta)]^2}{ni(\theta)}. \tag{4}$$

The Cramér–Rao inequality (1), and its special form (4), sometimes have been referred to as the information inequality, and the CRLB as the information bound. For the particular case $g(\theta) = \theta$, the numerator on the right-hand sides of both (1) and (4) reduces to 1.

The Cramér–Rao inequality has been generalized in a number of directions. Of special interest is the generalization for the case of d-dimensional parameter $\boldsymbol{\theta}$ with components θ_j, $j = 1, \ldots, d$ for $d \geq 1$. The regularity assumptions (Assumptions 1 and 2 above) are now needed with respect to elements θ_j of $\boldsymbol{\theta}$ for unbiased estimators $T_i = t_i(X)$ of the parametric function $g_i(\boldsymbol{\theta})$, $i = 1, 2, \ldots, r$.

Let $\mathbf{I}(\boldsymbol{\theta})$ be the $(d \times d)$ *Fisher information matrix* with elements

$$I_{jk}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{\partial \log p(X,\boldsymbol{\theta})}{\partial\theta_j}\,\frac{\partial \log p(X,\boldsymbol{\theta})}{\partial\theta_k}\right],$$

for $j, k = 1, \ldots, d$, and $\mathbf{G}(\boldsymbol{\theta})$ the $r \times d$ matrix of partial derivatives

$$G_{ij}(\boldsymbol{\theta}) = \frac{\partial g_i(\boldsymbol{\theta})}{\partial\theta_j}, \quad i = 1, \ldots, r; j = 1, \ldots, d.$$

In addition to generalized versions of the regularity assumptions (Assumptions 1 and 2) we now need also the following:

Assumption 3

The information matrix $\mathbf{I}(\boldsymbol{\theta})$ is positive-definite.

The generalized form of the Cramér–Rao inequality (1) is (see, for example, [8])

$$\mathrm{cov}_{\boldsymbol{\theta}}[\mathbf{T}] \geq \mathbf{G}(\boldsymbol{\theta})\mathbf{I}^{-1}(\boldsymbol{\theta})\mathbf{G}'(\boldsymbol{\theta}). \tag{5}$$

Here $\mathbf{T}$ is the vector of unbiased estimators T_i of $g_i(\boldsymbol{\theta})$, $i = 1, \ldots, r$, and the notation $\mathbf{A} \geq \mathbf{B}$ for any two nonnegative-definite matrices means that the matrix $\mathbf{A} - \mathbf{B}$ is nonnegative-definite.

For the special case $r = d$ with $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, the matrix inequality (5) reduces to the simpler form

$$\mathrm{cov}_\theta(\mathbf{T}) \geq \mathbf{I}^{-1}(\boldsymbol{\theta}), \tag{6}$$

for the covariance matrix $\mathbf{T}$ of unbiased estimators of $\boldsymbol{\theta}$.

When equality is attained for the two sides of inequality (1) for some unbiased estimator T of $g(\theta)$, it then follows that T is indeed the uniformly **minimum variance unbiased estimator** (UMVUE) of $g(\theta)$. Furthermore, under the regularity conditions (Assumptions 1 and 2), and some further technical conditions, it can be shown that the pdf of X has to belong to the one-parameter **exponential family** of the form

$$p(x, \theta) = \exp[a(\theta)T(x) + b(\theta) + c(x)], \tag{7}$$

for some real-valued functions $a(\cdot)$, $b(\cdot)$ of θ, and $c(\cdot)$ of x.

It should be noted that, when the form (7) holds for the pdf $p(\cdot, \theta)$, the equality is attained for the two sides in (1) only for the statistic T in (7) or its linear transforms $T^*(x) = c_1T(x) + c_2$, for some constants c_1 and c_2. Thus, the CRLB is attained only for unbiased estimation of parametric functions of the form $g(\theta) = c_1\mathrm{E}_\theta[T(X)] + c_2$, when X has pdf (7).

It has to be emphasized, however, that UMVUE estimates $T = T(X)$ can exist for unbiased estimation of $g(\theta)$ without attaining the CRLB; thus, the two sides of (1) might not be equal even when the pdf $p(\cdot, \theta)$ satisfies the form (7).

An analog of the Cramér–Rao inequality, when the regularity assumptions (Assumptions 1 and 2) are not necessarily met, is given by Chapman & Robbins [3], while Wolfowitz [9] has given the extension to sequential sampling situations. Bhattacharya [2] has given a more stringent inequality, using higher-order derivatives along with the first-order derivative of log $p(\cdot, \theta)$ used in the basic inequality (1). Similarly, for estimation of functions of parameters of interest $\boldsymbol{\theta}_1$, of d_1 dimensions, in the presence of unknown nuisance parameters $\boldsymbol{\theta}_2$, of d_2 dimensions, Bhapkar & Srinivasan [1] have given a more stringent inequality than (5) for the general case $d_1 \geq 1$, $r \leq d_1$ with $d_2 \geq 1$; the special case $r = d_1 = 1$ follows from the generalized information function developed by Godambe [6]. Gart [5] has given an extension of the inequality when the parameters $\boldsymbol{\theta}$ are themselves random variables.

References

[1] Bhapkar, V.P. & Srinivasan, C. (1994). On Fisher information inequalities in the presence of nuisance parameters, *Annals of the Institute of Statistical Mathematics* **46**, 593–604.

[2] Bhattacharya, A. (1946). On some analogues of the amount of information and their uses in statistical estimation, *Sankhyā* **8**, 1–14.

[3] Chapman, D.G. & Robbins, H. (1951). Minimum variance estimation without regularity assumptions, *Annals of Mathematical Statistics* **22**, 581–586.

[4] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[5] Gart, J.J. (1959). An extension of the Cramér–Rao inequality, *Annals of Mathematical Statistics* **30**, 367–380.

[6] Godambe, V.P. (1984). On ancillarity and Fisher information in the presence of nuisance parameter, *Biometrika* **71**, 626–629.

[7] Rao, C.R. (1945). Information and accuracy attainable in estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society* **37**, 81–91.

[8] Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd Ed. Wiley, New York.

[9] Wolfowitz, J. (1947). The efficiency of sequential estimates, and Wald's equation for sequential processes, *Annals of Mathematical Statistics* **18**, 215–230.

(*See also* **Efficiency and Efficient Estimators**; **Estimation**)

V.P. BHAPKAR

Critical Care

Over the last three decades, critical care has emerged as one of the most important and expensive aspects of medicine. There are over 6000 intensive care units (ICUs) in the United States today, caring for 55 000 patients per day [15]. The cost of this care is approximately 180 billion dollars, or almost 1% of the United States gross domestic product. The randomized controlled trial (RCT) has become the "gold" standard for clinical research (*see* **Clinical Trials, Overview**). This is also true for research in critical care; however, the complex nature of critical illness has made the design and conduct of RCTs difficult. Historically this has led to clinical decision based on **observational studies** or poorly controlled clinical trials. A recent review of RCTs published in a prominent critical care journal in the two decades up to 2000 found that only 25% of 173 trials could be considered adequate [9]. Over the last several years, several large well-designed RCTs have been published, which have resulted in major therapeutic advances in the treatment of the critically ill.

Research in critically ill patients presents unique challenges [7]. The disease processes are often not well defined but are described as a constellation of clinical findings, or syndrome. As such, definitions are often imprecise and variable. There are also multiple therapeutic interventions being administered and the disease processes themselves often have a variable clinical course. Finally, **outcome** selection is a challenge, mortality (short or long) versus nonmortality outcomes. These latter decisions have a major impact on ultimate determination of efficacy or effectiveness (*see* **Pharmacoepidemiology, Adverse and Beneficial Effects**).

Adding to the difficulty in conducting studies in the critically ill is the identification and enrollment of patients. To answer many of the clinical questions in critical care, large numbers of patients are required. While a particular clinical syndrome may be common in critically ill patients in general, often no single institution has enough patients to achieve a study of adequate size. Therefore, it is uncommon for any single center to be able to recruit patients in sufficient numbers within a reasonable time frame. This has led to a trend to large **multicenter studies**. In addition to improving the ability to recruit patients, multicenter trials potentially have greater generalizability as they account for practice difference across ICUs. A final issue is the difficulty in obtaining informed consent (*see* **Ethics of Randomized Trials**). This issue contributes to increasing difficulty in conducting research in a critically ill patient population [4, 11]. Regulations governing who can serve as a surrogate to provide consent for participation in a research study in those circumstances when patients are unable to give consent for themselves are becoming stricter. This is a particularly important issue for research in the ICU where patients often are not able to provide consent for themselves.

The last few years have seen the publication of the results of five RCTs that have had a major impact on the practice of critical care. These trials are notable in that they have all demonstrated mortality benefit for large populations of critically ill and in four of the five trials, the intervention involved simple changes in clinical practice rather than novel therapeutic agents.

Blood Transfusion

Critically ill patients are often anemic and as a result receive a large number of red blood cell (RBC) transfusions [5, 17]. However, over recent years, questions have arisen over both the safety and efficacy of RBC transfusion. Hebert and colleagues [8] compared two transfusion strategies: a restrictive strategy, maintaining a hemoglobin level between 7 and 9 $g\,dL^{-1}$ with a transfusion threshold of 7 $g\,dL^{-1}$; and a liberal strategy, maintaining a hemoglobin level between 10 and 12 $g\,dL^{-1}$ with a transfusion threshold of 10 $g\,dL^{-1}$. Because the entry criteria for the study specified a hemoglobin $\leq 9\,g\,dL^{-1}$, all patients in the latter group received an RBC transfusion. Patients in the restrictive group received 50% less RBC transfusions. All results favored the restrictive group, and in those patients who were younger ($\leq$55 years of age) and less sick (APACHE score $<$20), a liberal strategy resulted in a significant increase in mortality. The conclusion from this study was that a restrictive transfusion strategy was at least equivalent and in some patients superior to a more liberal transfusion strategy. This study has challenged long-standing beliefs regarding transfusion thresholds and transfusion practice.

Goal-directed Resuscitation

There is a long history of attempting to target resuscitation to physiologic endpoints in the critically ill patient with sepsis. The rationale for this is that there is an imbalance at the tissue level between oxygen delivery and demand and that by "optimizing" delivery, end organ damage could be prevented and survival thereby improved [2]. Unfortunately, these efforts have met with little success. Rivers and colleagues [14] in an RCT evaluated the efficacy of early goal-directed therapy initiated in the emergency room, prior to their ICU admission. As compared to "standard" therapy, early goal-directed therapy resulted in a significant reduction in mortality and organ failure (30.5% versus 46.5%, $p < 0.009$). The results from this study have major implications for the practice of critical care. The study highlights the importance of early identification of critically ill patients and early initiation of appropriate therapy, including initiation of therapy prior to arrival into the ICU.

Insulin Therapy

Hyperglycemia is common in critically ill patients, whether or not patients have a history of diabetes. Although it has been suggested that hyperglycemia may be associated with an increase in complications, there is little data on the impact of glucose control on clinical outcomes in the critically ill [12]. Van den Berghe et al. [16] in an RCT of patients receiving mechanical ventilation compared intensive insulin therapy (to maintain blood glucose between 80 and 110 $\mathrm{mg\,dL^{-1}}$) with standard therapy (insulin infusion for blood glucose $>215\,\mathrm{mg\,dL^{-1}}$ with maintenance between 180 and 200 $\mathrm{mg\,dL^{-1}}$). Intensive insulin therapy was associated with reduced mortality versus standard therapy (4.6% versus 8.0%, $p < 0.04$). This was a result of the mortality benefit in patients who remained in the ICU for more than five days (10.6% versus 20.2%, $p < 0.005$). This study demonstrates that normalization of blood glucose with insulin therapy resulted in a dramatic improvement in morbidity and mortality.

Acute Respiratory Distress Syndrome

Mortality for patients with acute respiratory distress syndrome (ARDS) is 40 to 50% [6]. The traditional approach to mechanical ventilation in these patients has involved the use of tidal volumes of 10 to 15 $\mathrm{ml\,kg^{-1}}$. However, it has been suggested that mechanical ventilation with tidal volumes in this range may in fact exacerbate lung injury. The ARDS Network conducted an RCT in patients with ARDS comparing a low tidal volume strategy with traditional mechanical ventilation [1]. This trial was terminated early because of mortality reduction in the lower tidal volume group (31% versus 39.8%, $p < 0.007$). In addition, patients in the low tidal volume group increased the number of days without ventilator use.

Sepsis

In the United States, 750 000 cases of sepsis occur each year of which 225 000 are fatal [10]. Over the last two decades, over 20 clinical trials have been conducted on therapeutic agents for sepsis [13]. These trials have failed to demonstrate any clinical benefit. Bernard et al. [3] performed an RCT evaluating the efficacy of activated protein C in severe sepsis. Activated protein C has antithrombotic, anti-inflammatory, and profibrinolytic properties. The mortality rate in the **placebo** group was 30.8% as compared to 24.7% with activated protein C. This was the first trial to demonstrate a benefit to therapeutic interventions directed towards interrupting the pathologic cascade initiated by sepsis.

References

[1] ARDS Network. (2000). Ventilation with low tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *The New England Journal of Medicine* **342**, 1301–1308.

[2] Beal, A.L. & Cerra, F.B. (1994). Multiple organ failure syndrome in the 1990s: systemic inflammatory response and organ dysfunction, *JAMA* **271**, 226–233.

[3] Bernard, G.R., Vincent, J.L., Laterrie, P.F., LaRosa, S.P., Dhainaut, J.F., Lopez-Rodriguez, A., Steingrub, J.S., Garber, G.E., Helterbrand, J.D., Ely, E.W. & Fisher, C.J. (2001). Efficacy and safety of recombinant human activated protein C for severe sepsis, *The New England Journal of Medicine* **344**, 699–709.

[4] Burck, R. (2002). Minimal risk: the debate goes on, *Critical Care Medicine* **30**, 1180–1181.

[5] Corwin, H.L., Abraham, E., Fink, M.P., Gettinger, A., MacIntyre, N., Pearl, R., Shabot, M. & Shapiro, M.J. (2004). Anemia and blood transfusion in the critically ill: current clinical practice in the US – The CRIT Study [abstract], *Critical Care Medicine* **32**, 39–52.

[6] Doyle, R.L., Szaflarski, N., Modin, G.W., Weiner-Kronish, J.P. & Matthay, M.A. (1995). Identification of patients with acute lung injury: predictors of mortality, *American Journal of Respiratory and Critical Care Medicine* **152**, 1818–1824.

[7] Hebert, P.C., Cook, D.J., Wells, G. & Marshall, J. (2002). The design of randomized clinical trials in critically ill patients, *Chest* **121**, 1290–1300.

[8] Hebert, P.C., Wells, G., Blajchman, M.A., Marshall, J., Martin, C., Pagliarello, G., Tweeddale, M., Schweitzer, I. & Yetisir, E. (1999). A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group, *The New England Journal of Medicine* **340**, 409–417.

[9] Latronico, N., Botteri, M., Cosetta, M., Zanotti, C., Bertoloni, G. & Candiani, A. (2002). Quality of reporting of randomized controlled trials in the intensive care literature, *Intensive Care Medicine* **28**, 1316–1323.

[10] Linde-Zwirble, W.T., Angus, D.C., Carcillo, J., Lidicker, J., Clermont, G. & Pinsky, M.R. (1999). Age-specific incidence and outcome in sepsis in the US, *Critical Care Medicine* **27**(Suppl. 1), A33.

[11] McRae, A.D. & Weijer, C. (2002). Lessons from everyday lives: a moral justification for acute care research, *Critical Care Medicine* **30**, 1146–1151.

[12] Mizock, B.A. (1995). Alterations in carbohydrate metabolism during stress: a review of the literature, *The American Journal of Medicine* **98**, 75–84.

[13] Nasraway, S.A. Jr. (1999). Sepsis research we must change course, *Critical Care Medicine* **27**, 427–430.

[14] Rivers, E., Nguyen, B., Havstad, S., Ressler, J., Muzzin, A., Knoblich, B., Peterson, E. & Tomlanovich, M. (2001). Early goal-directed therapy in the treatment of severe sepsis and septic shock, *The New England Journal of Medicine* **345**, 1368–1377.

[15] Schmitz, R., Lantin, M. & White, A. (1999). *Future Workforce Needs in Pulmonary and Critical Care Medicine.* Abt Associates, Cambridge.

[16] Van den Berghe, G., Wouters, P., Weekers, F., Verwaest, C., Bruyninckx, F., Schietz, M., Vlasselaers, D., Ferdinande, P., Lauwers, P. & Bouillon, R. (2001). Intensive insulin therapy in critically ill patients, *The New England Journal of Medicine* **345**, 1359–1367.

[17] Vincent, J.L., Baron, J.F., Gattinoni, L., Reinhart, K., Thijs, L., Webb, A., Meier-Hellmann, A., Nollet, G. & Peres-Bota, D. (2002). Anemia and blood transfusion in critically ill patients, *JAMA* **288**, 1499–1507.

HOWARD L. CORWIN

Critical Region

To test a statistical hypothesis, one must define a rule for determining when to "accept" or "reject" the specified hypothesis (*see* **Hypothesis Testing**). This rule is usually, but not always, based upon observed data. The set of observations that corresponds to the rejection of the specified hypothesis is said to be the *critical region* of the test. The steps for performing a test for a given statistical hypothesis are straightforward. First, one observes a *random sample* of data. Next, one determines whether this observed random sample is contained in a predefined critical region. This is usually done by calculating a statistic (e.g. the sample mean) based on the observed data and determining whether the statistic is contained within the critical region. If the sample (i.e. statistic) is in the critical region, then one rejects the hypothesis. The *size* of the critical region is also referred to as the size of the test (which is often denoted by α; *see* **Level of a Test**). The size of the test is the probability that the observed data would have fallen into the critical region if the hypothesis were true. It is important to note that there may be several possible critical regions for a fixed size of a test. For instance, for the normal distribution, if one chooses the size of the test equal to 0.05, then the following critical regions are all possible:

1. values of the statistic either larger than 1.96 or smaller than -1.96 (a two-tailed test)
2. values of the statistic larger than 1.645 (a one-tailed test; *see* **Alternative Hypothesis**)
3. values of the statistic larger than 1.75 or smaller than -2.326 (a two-tailed test with asymmetric cutpoints).

In all three examples, the critical region is different, and yet the size of the critical region (size of the test) is the same. Figure 1 shows three plots illustrating these three different critical regions.

Consider the following example. We wish to test the hypotheses as to whether the population mean, μ, for some normally distributed random variable is equal to μ_0 (some specified value) or whether it is larger than μ_0. We can write these hypotheses as $H_0 : \mu = \mu_0$ (the **null hypothesis**) and $H_a : \mu > \mu_0$ (the *alternative hypothesis*). Next, we define a rule for determining when to accept or reject the null hypothesis for a specified level of significance α. One rule may be that we reject the hypothesis if the mean of the observed data, $\overline{x}$, is larger than a specified value C. A common choice of C in this setting would be $C = \mu_0 + z_\alpha(\sigma/n^{1/2})$, where σ is the known standard deviation of the variable being studied, n is the sample size of the observed data, μ_0 is the hypothesized value of μ, and z_α is the value from a standard normal distribution which has α area above it and $1 - \alpha$ area below (i.e. for $\alpha = 0.05$, $z_\alpha = 1.645$ since the area above 1.645 in the standard normal distribution is 0.05 and the area below 1.645 is 0.95). One would then collect data, $X_1, \ldots, X_n$, and if the observed mean of these data, $\overline{x}$, was larger than C, then the null hypothesis would be rejected. For this example, the set of values of the test statistic $\overline{x}$ that leads to the rejection of the null hypothesis is the critical region.

RALPH B. D'AGOSTINO, JR

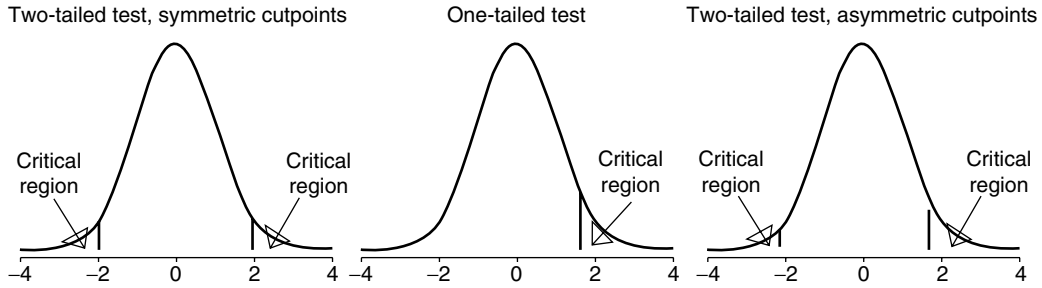

Figure 1 Critical regions for a two-tailed test with symmetric cutpoints, a one-tailed test, and a two-tailed test with asymmetric cutpoints

Cronbach's Alpha

Cronbach's alpha, or coefficient alpha, is widely used to assess the internal consistency or reliability of multiple-item instruments. A multiple-item instrument, also called a multiple-item scale, measuring an underlying construct, or latent variable, is internally consistent if its items are highly intercorrelated. Cronbach's alpha measures this internal consistency [1].

The relationship between the latent variable and the actual responses to the items is known as the measurement model. The classical measurement model assumptions are that measurement errors for each item are random, are not correlated with one another, and are not correlated with the true score, or latent variable. Two additional assumptions called the parallel test assumptions are that the latent variable affects all items equally and that there is equal measurement error in each item. Coefficient alpha is defined as the proportion of variance attributable to the true score of the latent variable [2].

An equivalent model is generated by assuming that items cannot be measured without error, and that the total variation in the items can be partitioned into the true, or common, variation and the error variation [4]. The true variation is the variation in latent score values across the subjects. Coefficient alpha can then be viewed as the ratio of the true variation to the total variation.

We now present the derivation of coefficient alpha. Consider a latent variable which is measured by k distinct items $x_1, x_2, \ldots, x_k$. The x_is are observed data measured on a sample of n subjects. The variance–**covariance matrix** of the k items is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_k^2 \end{bmatrix}.$$

We assume that each observed item, x_i, can be used to estimate the latent variable Y. A better estimate of the latent variable is produced by summing the responses over the set of observed items. Let Y denote the summed rating scale $Y = \sum_{i=1}^{k} x_i$. The total variation in Y is the sum of the elements of $\mathbf{\Sigma}$: $\text{var}(Y) = \sigma_Y^2 = \sum_{j=1}^{k} \sum_{i=1}^{k} \sigma_{ij}$, where $\sigma_{ii} = \sigma_i^2$. The unique variation is the sum of the diagonal elements of $\mathbf{\Sigma}$, or $\sum_{i=1}^{k} \sigma_i^2$. The ratio of the unique, or noncommon, variation to the total variation in Y is given by $\sum_{i=1}^{k} \sigma_i^2/\sigma_Y^2$. The proportion of common variation, or true score variation, in Y is defined as: $1 - \sum_{i=1}^{k} \sigma_i^2/\sigma_Y^2$. Coefficient alpha is based on this proportion of common or true score variation, and is adjusted to reduce dependency on the number of items, k, as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \left[\frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_Y^2}\right]\right).$$

As the proportion of true score variation accounted for by the observed items increases, the reliability increases. Higher reliability indicates less effect of random errors in the model.

Researchers interested in measuring an underlying construct or latent variable generally use either existing items (i.e. manifest or observed variables with known psychometric properties) or develop new items which relate to the underlying construct. For example, suppose a health maintenance organization (HMO) wishes to measure whether its patients are satisfied with their medical office visits. The underlying construct or latent variable is patient satisfaction. Suppose five items are created to measure patient satisfaction, one of which might be: "How satisfied are you with your physician's willingness to answer all of your questions during your office visits?" The response options for each item could be a **Likert scale**, as shown in Table 1.

Suppose the items are administered by questionnaire to a random sample of patients from the HMO. To evaluate whether a reliable, multiple-item satisfaction measure exists (i.e. whether the five items are consistently measuring the underlying construct, satisfaction) requires a series of, sometimes iterative, analyses. First, all of the items must be scored in the same direction (e.g. higher scores indicate more patient satisfaction for every item). Then the distributional properties of the items must be evaluated. Specifically, the means and standard deviations

Table 1

Not satisfied	Somewhat satisfied	Neither satisfied nor unsatisfied	Satisfied	Very satisfied
1	2	3	4	5

should be approximately equal (parallel tests assumption). If the distributional properties of the items are similar, then coefficient alpha is computed according to the formula shown above. The value of coefficient alpha is then evaluated (see guidelines below). If coefficient alpha is in the acceptable range, then a multiple-item scale is constructed. If coefficient alpha is not in the acceptable range, investigators should examine individual items carefully (see below and [3] for examples).

The theoretical range of coefficient alpha is 0 to 1. Suggested guidelines for interpreting coefficient alpha are: <0.60 unacceptable, 0.60–0.65 undesirable, 0.65–0.70 minimally acceptable, 0.70–0.80 respectable, 0.80–0.90 very good, and >0.90 consider shortening the scale by reducing the number of items [2].

Suppose, in our example, that the five items are highly consistent, with an observed coefficient alpha of 0.80. Here, each item has responses ranging from 1 to 5. The summed rating scale is computed by summing the five item responses for each patient. The theoretical minimum and maximum values for the satisfaction scale are 5 and 25, respectively.

Summed rating scales should not be created when items are not internally consistent (i.e. have a low alpha coefficient). If a set of items produces a low alpha coefficient, either there are too few items or the items have little in common. To assess the latter, the item–total correlations, which are the correlations between each individual item and the sum of the remaining items that constitute the scale, should be investigated. If a particular item has a low item–total correlation, it should be dropped from the scale and coefficient alpha should be recomputed.

Cronbach's alpha is also used for split-half **reliability** assessment. In split-half reliability assessment a set of k items is randomly split into equal halves of $k/2$ items each. Scores based on the two halves are computed (e.g. $Y_1 = \sum_{i=1}^{k/2} x_i$ and $Y_2 = \sum_{i=k/2+1}^{k} x_i$) and correlated. Let r_{12} denote the correlation between Y_1 and Y_2. There are many ways to split the items into the two halves. Cronbach's alpha (as defined above) is equivalent to the lower bound on the correlation between scores derived on sets of $k/2$ items (i.e. Cronbach's $\alpha < r_{12}$ over all possible r_{12}).

Alternative formulas for coefficient alpha are the Kuder Richardson Formula 20 (KR-20), which is primarily used to assess the reliability of measures based on dichotomous items, and the Spearman–Brown prophecy formula which is based on the average interitem **correlations**, as opposed to covariances. The respective formulas are:

$$\alpha = \frac{k}{k-1}\left(1 - \left[\frac{\sum pq}{\sigma_Y^2}\right]\right),$$

where $p = \Pr$ (affirmative response in each dichotomous item),

$$q = 1 - p,$$

and

$$\alpha = \frac{k\overline{r}}{1 + (k-1)\overline{r}},$$

where $\overline{r}$ = the average interitem correlation coefficient.

References

[1] Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* **16**, 297–334.

[2] DeVellis, R.F. (1991). *Scale Development: Theory and Applications*. Sage, Newbury Park.

[3] Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling*. SAS Institute Inc., Cary.

[4] Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Ed. McGraw-Hill, New York.

(*See also* **Principal Components Analysis**; **Psychometrics, Overview**)

KIMBERLY A. DUKES

Crossover Designs

Crossover trials are sometimes referred to as change-over or repeated measures designs, although the latter term is better reserved for the more general field of which crossovers form a part. They have been defined as follows: a crossover trial is one in which individual subjects are given sequences of treatments with the object of studying differences between individual treatments (or subsequences of treatments) [44].

Example 1

Graff-Lonnevig et al. [18] reported a crossover trial in asthma comparing single inhaled doses of 12 μg formoterol with a single inhaled dose of 200 μg salbutamol. Fourteen children with asthma were allocated at random to one of two sequences: formoterol followed by salbutamol or salbutamol followed by formoterol, and had their peak expiratory flow (PEF) measured at various times after treatment. One child (number 8) failed to complete both periods of treatment. PEF readings 8 h after treatment for the 13 other children for the two periods are given in Table 1, the last two columns of which need not concern us for the moment.

This is an example of the most studied type of crossover design, sometimes referred to as the two-period crossover and sometimes as the two-treatment, two-period crossover, but perhaps better designated, referring explicitly to the sequences employed, as the AB/BA design. (In this example we have A as formoterol and B as salbutamol or vice versa.) A schematic representation of such crossover trials is given in Figure 1.

Example 2

A crossover trial in migraine compared two doses (D1 and D2) of the potassium salt of diclofenac, a nonsteroidal anti-inflammatory drug, with placebo (P). Patients were allocated at random to one of six possible sequences of treatment: D1 D2 P, P D2 D1, D2 D1 P, D2 P D1, D1 P D2, and P D1 D2 [44].

In both of the above examples patients crossed over from one treatment to another, and this stratagem gives the design its name. In more complicated designs, however, not all patients cross over to all treatments and the essential feature is that which we have tried to capture in our definition. In both of the

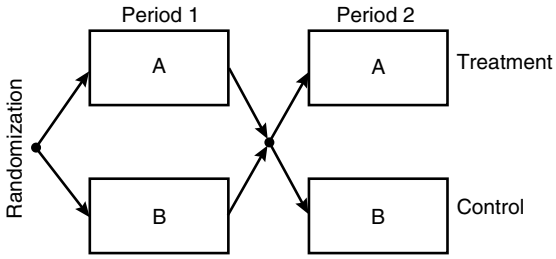

Figure 1 Schematic representation of an AB/BA crossover

Table 1 Data from Example 1 on PEF in l/min, 8 h after treatment, for a trial in asthmatic children

		PEF			
Sequence	Patient number	Period 1	Period 2	Basic estimator	Two period totals
Formoterol/Salbutamol	1	310	270	40	580
	4	310	260	50	570
	6	370	300	70	670
	7	410	390	20	800
	10	250	210	40	460
	11	380	350	30	730
	14	330	365	−35	695
Formoterol/Salbutamol	2	370	385	15	755
	3	310	400	90	710
	5	380	410	30	790
	9	290	320	30	610
	12	260	340	80	600
	13	90	220	130	310

designs above the number of treatments equals the number of periods, but **incomplete block designs** in which the number of periods is inferior to the number of treatments are not uncommon, as in Example 3.

Example 3

A parallel assay comparing beta-agonists in asthma was run in seven treatments (three doses of a reference formulation, three doses of an experimental formulation, and placebo) and five periods [48]. The trial was planned to have 126 patients allocated in equal numbers to 21 sequences. In the end, 161 patients were recruited in 15 centers and four countries with an average of about seven and a half patients per sequence but with at least six patients on every sequence.

Designs in which the number of periods exceeds the number of treatments have also been studied extensively, if rarely applied. An example is provided by Ebbut [6], and these designs are discussed below.

Crossover trials are commonly used in drug development for various indications where the disease is chronic and relatively stable – asthma, hypertension, sleep disturbance, angina, diabetes, migraine, rheumatism, and epilepsy are examples. They are not suitable for life-threatening diseases, for the obvious reason that the patient may die and be unavailable for further study and, conversely, they are generally unsuitable for conditions in which a permanent cure may be affected.

The main advantage of the crossover trial is its extreme efficiency. Because each patient forms his or her own control (a feature which makes such designs intuitively attractive to the physician), an important source of variability present in parallel group trials (*see* **Clinical Trials, Overview**), the variability between patients, is eliminated. As a consequence, fewer patients are generally needed to form an adequate conclusion. A further potential advantage is that individuals' reactions to treatment may be studied (although to do this adequately requires crossover trials in which patients are repeatedly allocated to the same treatments). Crossover trials place a heavier burden of participation on the individual patient, however, whose time in the trial will be much longer than for a comparable parallel group trial and, therefore, the danger that patients will drop out is increased (as for patient 8 in Example 1). The total time to conclude a crossover trial may be longer, although this is rarely the case, since the recruitment phase is what dominates most trials and since fewer patients are required. The analysis of crossover trials is also more complex and controversial than for parallel group trials and, as noted above, the design is suitable only for certain indications. A potential problem with crossover trials, which is dealt with in detail below, is that of carry-over.

As a consequence of these potential advantages and disadvantages, crossover trials are more extensively used in **Phase I** and **Phase II trials** than in the long-term therapeutic studies which characterize phase III. They are by far the most popular choice of design for **bioequivalence**. For the indications in which they may be used, they are generally much better at dose finding than are parallel group trials [49], the latter suffering from the fact that they require group averages to study effects which operate on the individual level. They are more likely to be used for single-dose pharmacodynamic studies of an explanatory nature (*see* **Pharmacokinetics and Pharmacodynamics**) than for multiple-dose therapeutic trials of a pragmatic nature. They are the design of choice for studying individual reactions to treatment [44]. Because of the potential problem of carry-over, they are generally viewed with suspicion by drug regulatory agencies. However, they are extremely popular in certain indications as independent physician-initiated trials, for the simple reason that they are often the only trials that have adequate **power** when run in a single center.

Carry-Over

A standard assumption in the design and analysis of experiments is that there is no interference between units or treatments given to units – for example, that crops growing in one part of a field and treated with a given fertilizer do not affect the growth of crops treated in another plot with an alternative fertilizer. In many medical investigations this assumption is often (and generally safely) ignored, but there are cases where it could conceivably break down. For example, if depressive patients in a trial are together on one ward, affecting the mental state of some may have an indirect effect on others. Similarly, a trial of prophylactic education of homosexual men regarding transmission of human immunodeficiency virus (HIV) may reduce the chance of infection

amongst those not selected to receive such education, either by reducing infection amongst potential sexual contacts or by an indirect spread of the relevant knowledge. In crossover trials, the experimental units are *episodes* of treatment rather than patients (*see* **Unit of Analysis**). If the treatment given in one period continues to affect the patient when subsequently treated, then this *carry-over* of the treatment effect is an example of interference between units. It can have serious consequences for the interpretation of treatment effects because we may imagine that we are studying the response to a single treatment, when in fact we are also observing a residual effect of a previous treatment. Carry-over is regarded by many commentators as being the outstanding problem of crossover trials and, indeed, having been evoked, it will continue to have a residual effect throughout the rest of this article.

Different types of carry-over can be envisaged. However, for reasons to be explained, we shall make the following restrictions on carry-over. First, we assume that it cannot disturb the conclusions of a trial unless there is a genuine difference between treatments, and secondly that a carry-over is a true residual effect of a treatment. These may seem to be an unrealistic restrictions. For example, we can surely envisage cases where otherwise similar but synergistic drugs (*see* **Synergy of Exposure Effects**) may have unequal persistence. In that case, where the shorter persisting drug is given second, it will benefit from an **interaction** which will be denied to the longer persisting drug when it is given second. The reason for ignoring such exotic cases is that similar problems can also be envisaged as affecting parallel group trials. Consider our Example 1. Salbutamol is a standard treatment for asthma developed in the 1960s, whereas formoterol is a newer treatment first registered in the UK in the 1990s and (at the time of writing) not yet available in the US. Most patients recruited to trials of formoterol had been taking salbutamol for many years. Thus a parallel group design would really be as illustrated in Figure 2. There are thus two ways at least in which carry-over could affect parallel group trials comparing formoterol to salbutamol. First, given some persistence of the effects of salbutamol beyond whatever wash-out period was instituted, there could be a salbutamol–formoterol synergy from which patients in the formoterol group could benefit but which those in the salbutamol group would be denied. Secondly, patients recruited could have tachyphilaxis to salbutamol and be no longer capable of showing the same effect with that drug. Both of these phenomena are extremely unlikely to be important, and since they would also be extraordinarily difficult to guard against, there is no point in organizing parallel group trials to deal with them. But, by the same token, we must not saddle crossover trials with these problems either. This is not to say that interactions between carry-over and treatments are necessarily unimportant (where carry-over is appreciable they may be), but that it is pointless to attempt to deal with the problem of carry-over for the case where there is no treatment effect at all. As a consequence, whereas carry-over may affect the power of a test of the effect of treatment, we shall assume that it cannot affect its size.

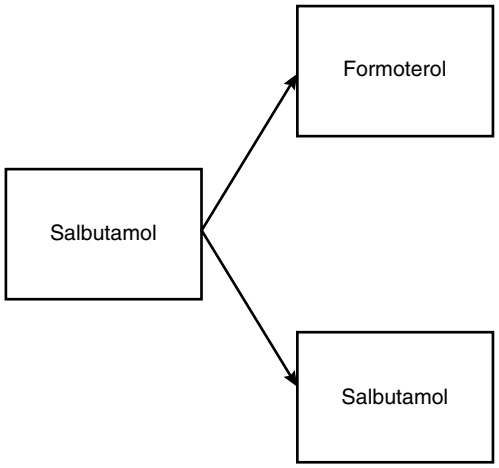

Figure 2 Illustration of the likely position in any parallel group trial in asthma comparing formoterol with salbutamol

Similarly, we shall not consider the related problem of period by treatment interaction. A comparable problem also affects parallel group trials. For example, three-month parallel group trials in asthma are commonly employed to support the registration of drugs, which may be then be taken by patients for a lifetime. This extrapolation requires an assumption about time-by-treatment interaction which, although not identical to period by treatment interaction (see below), is analogous. (The difference is essentially that in a parallel group trial a patient's time on the trial is, largely, time on the treatment.)

Periods

The word *period* is commonly used in the literature on crossover trials to describe the different occasions on which patients are treated. As we shall see, period effects are also commonly fitted in the models used to analyze crossovers. Two points are worth noting in this respect. First, the patients on clinical trials are rarely recruited at the same time, and crossover designs are no exceptions. Thus, one patient's period 2 may actually occur before another's period 1. Secondly, for multiple-dose crossover trials, in which the object is to study the steady-state effect of treatment, the trialist must determine the length of the period. The relevant primitive design constraint for such trials is likely to be the total length of time for which it is considered reasonable to have a patient on the trial. The number of periods is then a derived design constraint given the estimated time it takes for a patient to reach the steady state. This point has been misunderstood in much of the literature on optimal design of multiperiod crossovers, in which it has been implicitly assumed that the number of periods available is an absolute constraint, within which one must produce designs and associated analyses which guard against carry-over. In fact, by increasing the length of a period at the cost of reducing the number of periods, the trialist may reduce the risk of carry-over. The sort of assumptions necessary are no different in kind from the one commonly made that carry-over lasts for one period only. Obviously if this is true and the periods are made twice as long, carry-over can be eliminated completely.

Wash-Out

In designing crossover trials, precautions are commonly taken to attempt to eliminate carry-over. A period in which the residual effect of a treatment is presumed to disappear is known as a wash-out, although this is often a rather arbitrary label. In single-dose pharmacodynamic trials, the object of a crossover is to study onset of action as well as duration. The trialist has to determine how long after the previous treatment was given the next should be administered. The distinction between wash-out and treatment period is then essentially arbitrary. The treatment period is deemed to end once the trialist stops measuring the effect of a drug and the next treatment period starts once the next treatment is administered. The difference between the two is a wash-out, but essentially the trialist is faced with the task of determining the minimum *total* time interval between treatments. (In many trials the period varies from patient to patient, subject only to the constraint of a predetermined minimum, and the patient will take his or her regular treatment for at least some of the time between trial treatments. Both of these points are regularly overlooked.) For multiple-dose trials, where the object is to study the steady state, an active wash-out may be instituted: the effect of a previous treatment is presumed to wear off gradually during the period in which the following treatment is applied. For the purpose of analysis, measurements from the latter part of a period only are used. Thus the trialist's task is to determine the minimum total time between application of the first treatment and measurement of the effect of the second. A third case is where we have multiple-dose studies but wish to study onset of action as well. In this case there is little choice but to disrupt the rhythm of regular treatment and have a true wash-out in which no treatment is given at all. This is often considered to be a practical obstacle to running crossover trials, but it should be noted that it only arises from the need to study the onset of action and that, if this is the purpose, the need for a wash-out period will arise in a parallel group trial also, the difference being that the patient in a crossover trial will have at least two periods of wash-out rather than just one.

A Brief Historical Note

As with many experimental designs, crossover trials received an early application in agricultural research. Jones & Kenward [25] describe an experiment carried out in the nineteenth century over a long series of years at Rothamsted by John Lawes, in which each member of a pair of plots was treated either with minerals or ammonia. In clinical crossover trials, the main source of variability is usually patients rather than periods. In Lawes' experiment it would be years (owing to the weather) rather than plots. From one point of view, therefore, Lawes' experiment is an AB/BA crossover with plots taking the place reserved for periods in clinical crossovers and years being used for replication in the way that patients are. (If the plots were neighboring, the carry-over could occur both down the years and from plot to plot!)

An early crossover trial in medicine is the investigation published in 1905 by Cushny & Peebles [5] of the hypnotic effects of optical isomers. The data were used for illustrative purposes by Student (*see* **Gosset, William Sealy**) in his famous paper [11, 40, 47, 51]. Simpson's 1938 paper [50] provides an example of their use in studies of nutrition. Crossover trials are also covered in **Cochran** & **Cox's** famous book, although the applications considered are not medical [4]. Finney's 1956 paper [10] is important and considers the design of crossover trials for use in bioassay (*see* **Biological Assay, Overview**), a topic which had been considered with particular relevance to insulin assay earlier by **Irwin** [24] in 1937 and Fieller [9] in 1940, and even earlier by Marks in 1925 in the medical literature [35]. A burst of methodological investigation into the AB/BA crossover was initiated in the 1960s by the papers by Chassan [2] and Grizzle [22], the latter being extremely influential in its proposal of a two-stage approach to the analysis of such trials. This approach was endorsed with some hesitancy by an influential paper by Hills & Armitage [23], but was eventually shown to be potentially extremely misleading by Freeman [16], whose paper has set the agenda for current research into the AB/BA design. **Bayesian** approaches in bioequivalence were proposed by Selwyn et al. [43] in 1981, and an alternative method for crossover trials in general has been developed by Grieve [19–21, 41] in a series of papers starting in 1985. Koch [29] proposed a **nonparametric** approach for analysis in 1972. Early papers considering multiperiod designs are those of Simpson [50] (mentioned above) and **Yates** [54], also from 1938. Williams [53] in 1949 explicitly proposed designs for balancing the residual effects of treatment. Other influential papers will be mentioned in appropriate Sections below.

The AB/BA Crossover

The Basics of Analysis

This is the simplest of all crossover designs for the purpose of comparing two treatments, and is illustrated in Example 1. The fact that there is an extensive literature associated with such an apparently simple design may come as a surprise to those unfamiliar with the field. The reason has to do with carry-over. To discuss the effect of carry-over in detail it is necessary to introduce a model for the AB/BA design. However, since, given an assumption that carry-over has *not* taken place, a very simple analysis of such a design is possible, and which has some claims to being as good, if not better, than any other, it is perhaps worth postponing the model until this analysis has been explained. The method has been clearly described by Hills & Armitage [23].

The first step is to calculate for each patient what has been referred to as a *basic estimator* [44] or crossover difference [29]. This is the difference for a given patient between the two treatments and is what we should be constrained to use as our estimate of the treatment effect, τ, were we to have data only on the given patient. The basic estimator for each patient is reproduced in the last column of Table 1. It has been calculated as the reading under formoterol minus that under salbutamol. Thus, for patients of the first sequence it is the period 1 value minus the period 2 value, whereas for patients in the second sequence it is the converse. Once the basic estimators have been calculated, there are then two standard approaches to analysis. One is to ignore the distinctions between sequences, treat the values as a single sample of data, and calculate estimates, ***P* values**, and confidence intervals accordingly. A parametric approach uses the (matched pairs) t test (*see* **Student's *t* Distribution**), and the nonparametric equivalent is the **Wilcoxon signed-ranks test**, or even the **sign test**. For a **randomization**-based mode of inference (*see* **Randomization Tests**), these methods are compatible with a design in which patients have been allocated completely at random to the two sequences [44]. If the matched-pairs t test is applied to the basic estimators here, it will be found that the mean basic estimator is 45.4 l/min, its estimated standard error on 12 degrees of freedom is 11.3 l/min, the observed t ratio is 4.03, the critical values at the 5% level two-sided are ± 2.18, the 95% confidence limits are 21 l/min and 70 l/min, and the P value (two-tailed) is 0.0017.

The approach we shall concentrate on, however, *does* recognize the distinction between sequences. The first step is to calculate a separate mean of the basic estimators for each sequence. These two sequence means are 30.7 and 62.5. The difference between them is probably due to chance, but might conceivably be due to a period effect. For example, if there were a secular tendency for second period values to be higher, this would diminish basic estimators in the first sequence and increase them in the

second. This would have two consequences. First, conditional upon a given unbalanced allocation to the two sequences, a straightforward average of the basic estimators would be **biased**. (Of course, in a randomized design, over all possible allocations, it would not be biased.) A simple way of dealing with this is to block the design (*see* **Blocking**) so that there are equal numbers of patients per sequence, but this does not deal with the second difficulty, namely that under such circumstances the variance of the basic estimators calculated by treating them all as coming from a single sample will be inflated by a component due to the effect of the periods.

A simple solution which deals with both problems is to stratify on the sequences and take an unweighted average of the two sequence means. In this case, the unweighted average, which we denote $\hat{\tau}$, is 46.6 l/min. In general if the two sequences have n_1 and n_2 patients, respectively, then such a contrast will have a **variance** of $q\sigma^2/4$, where σ^2 is the variance of a basic estimator and $q = (1/n_1 + 1/n_2)$. The variance of a basic estimator may in turn be *estimated* using the standard approach of pooling within group sums of squares, familiar from the two-independent-samples t test, using

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

where S_1^2 and S_2^2 are the unbiased estimates of σ^2 from sequence groups 1 and 2, respectively. Putting this together, we may calculate a t statistic with indexed degrees of freedom as

$$t_{n_1+n_2-2} = \frac{\hat{\tau} - \tau}{(\hat{\sigma}/2)\sqrt{q}}. \qquad (1)$$

If we take τ as the value provided by some **null hypothesis** of interest (typically that $\tau = 0$), then we may use (1) in a significance or hypothesis test in the usual way (*see* **Hypothesis Testing**). Alternatively, by setting the left-hand side of (1) equal to some percentage point and treating τ as unknown, we may calculate a confidence limit for τ. For this example, $S_1^2 = 1086.9$, $S_2^2 = 1997.5$, and hence $\hat{\sigma} =$ 38.7 l/min. For $\tau = 0$, the t statistic is 46.6/10.8 = 4.3 and the P value is 0.0012. Alternately, since the critical values for a test at the 5% level (two-sided) on 11 degrees of freedom are $\pm$ 2.2, the 95% confidence limits are 23 l/min and 70 l/min.

An equivalent analysis to the above is to work with period differences rather than basic estimators. The difference between periods is then calculated in the same direction in each group so that the period differences in the second sequence group are simply the negative of the basic estimators. The semidifference between the two sequence means, rather than their average, now reflects the treatment effect. In fact, if all that is wanted is to test for the treatment effect, a two-sample t test may be used. Alternatively, if the semiperiod differences are calculated on the subjects in the first place, then a confidence interval may be calculated for τ using the two-sample t. The equivalent nonparametric approach is to use a **Wilcoxon–Mann–Whitney test** as proposed by Koch [29]. Alternatively, **Fisher's exact test** using the signs of the period differences or an adapted Brown–Mood median test may be used [44].

A Model for The AB/BA Design

The above analysis, most commentators would agree, adjusts the treatment effect adequately for any difference due to periods and also validly eliminates the effect of individual patients from the estimate of the standard error of the treatment effect. If, however, carry-over is present, then the estimator of τ is biased. This is not, necessarily, a compelling reason for abandoning this approach [44], but the possible bias has led statisticians to look at the problem of carry-over in depth. To follow these arguments we now introduce a model for carry-over as follows.

Let Y_{ijk} be the response of subject j, $j = 1, \ldots, n_1$ or n_2, of sequence $i, i = 1, 2$, in period $k, k = 1, 2$. (Thus both i and j are necessary to identify a given subject.) Then let

$$Y_{ijk} = \alpha_{ik} + s_{ij} + \varepsilon_{ijk}, \qquad (2)$$

where α_{ik} is an effect common to all subjects in a given sequence in a given period, s_{ij} is an effect due to the given subject which may, according to the circumstance, be treated as either fixed or random, and ε_{ijk} is a random error term. The ε_{ijk} are assumed to be independent with constant variance γ^2 and, where the s_{ij} are treated as random (*see* **Random Effects**), they are taken to be independently distributed both of each other and the ε_{ijk}, with constant variance ϕ^2. (Note that σ^2, as previously defined, is $2\gamma^2$.) This is a components of variance

(*see* **Variance Components**) approach which is commonly used. An alternative formulation of the random effects model [22], which is more general, is to model a combined error term ξ_{ijk} with a block diagonal form for variances and covariances so that **correlations** between observations on the same patient are equal to ρ and correlations of observations between different patients are zero. Such a formulation allows for the theoretically possible, but unlikely, case where the correlation between observations on the same patient is negative. If, however, the components of variance model is correct, then since $\xi_{ijk} = s_{ij} + \varepsilon_{ijk}$ we have $\rho = \phi^2/(\phi^2 + \gamma^2)$, which can never be negative. We continue with the components of variance approach.

The next step is to model the α_{ik} terms. A possible parameterization is given in Table 2, with treatment effect τ indexed by the treatment labels A and B, period effects, π indexed 1 and 2, and carry-over effects λ indexed by engendering and perturbed treatment and general level μ. The object of the crossover trial is then to make inferences about the contrast $\tau_A - \tau_B$, equivalent to τ in our previous formulation. Also given in Table 2 are four cell means, each of which is an unbiased estimate of the combination of parameters in the same cell. Three linear combinations of these cell means have been much studied in the literature and, together with their expectations, are given in Table 3. The names of the contrasts are as used by Freeman [16] and Senn [45].

The first of these, CROS (for crossover), is nothing less than the simple estimate, $\hat{\tau}$, of the treatment effect we encountered above. This can now be seen to be a biased estimator of τ, except where the differential carry-over effect, $\lambda = \lambda_{AB} - \lambda_{BA}$, is zero. This will most plausibly be the case where both λ_{AB} and λ_{BA} are zero, i.e. when the wash-out has been successful in eliminating carry-over. In other cases, the estimator is biased downward by half the differential carry-over (hereafter simply referred to as the carry-over). As a consequence, where this occurs, there will be some loss of power associated with the test of the treatment effect.

The second linear combination, SEQ (differences between sequences), has an expectation equal to the carry-over, and the third, PAR (so called because it is a between-patient contrast using first-period data only and hence of the sort commonly used in a parallel group trial), is an unbiased estimator of the treatment effect. The three linear combinations have been presented in terms of the four cell means given in Table 2. Each corresponds, however, to the difference between the two sequence groups of the mean over all the patients in the group of a simple summary statistic calculated for each patient. For PAR (L_p), this summary statistic is simply the first period value. For SEQ (L_s), it is the total over two periods, and for CROS (L_c), as already discussed, it is the semidifference between the first and second period values. Note that these statistics are related by the

Table 2 Model for the AB/BA design

		Period 1 ($k = 1$)	Period 2 ($k = 2$)
Sequence	AB ($i = 1$)	$\mu + \tau_A + \pi_1$ $\overline{Y}_{1.1}$	$\mu + \tau_B + \pi_2 + \lambda_{AB}$ $\overline{Y}_{1.2}$
	BA ($i = 2$)	$\mu + \tau_B + \pi_1$ $\overline{Y}_{2.1}$	$\mu + \tau_A + \pi_2 + \lambda_{BA}$ $\overline{Y}_{2.2}$

Table 3 Linear combinations of the cell means

Name	Label	Definition	Expectation
CROS	L_c	$[(\overline{Y}_{1.1} - \overline{Y}_{1.2}) + (\overline{Y}_{2.2} - \overline{Y}_{2.1})]/2$	$(\tau_A - \tau_B) - (\lambda_{AB} - \lambda_{BA})/2$ $= \tau - \lambda/2$
SEQ	L_s	$(\overline{Y}_{1.1} + \overline{Y}_{1.2}) - (\overline{Y}_{2.2} + \overline{Y}_{2.1})$	$(\lambda_{AB} - \lambda_{BA})$ $= \lambda$
PAR	L_p	$\overline{Y}_{1.1} - \overline{Y}_{2.1}$	$(\tau_A - \tau_B)$ $= \tau$

relationship $L_p = L_c + L_s/2$. Of the three summary statistics, only CROS eliminates the subject effect given by s_{ij} in (2). Therefore, if we are interested only in this particular contrast, it is irrelevant whether the subject effect is assumed to be random or fixed since it is eliminated anyway. For the other two contrasts this is not the case, and it becomes necessary to treat the subject effects as random. (For the discussion that follows, this assumption is now made.) Given this assumption, the two-independent-samples t test may be used in connection with SEQ and PAR. If this is done for Example 1, then the means (estimated standard errors) are $L_s = 14.4(80.4)$ l/min and $L_p = 53.8(45.3)$ l/min, and the t statistics are 0.18 and 1.19, respectively, both on 11 degrees of freedom. Thus, if we are prepared to use the potentially biased estimate of the treatment effect for this example, then it is significant, but if we insist on using the unbiased estimate of the treatment effect, then it is not significant. The inconsistency of such inferences is not unusual and indeed is to be expected. Crossover trials are designed to be used with far fewer patients than parallel group designs, and one should not be surprised if an estimate which uses the within-patient structure of a trial is significant whereas one which discards half the data and is between-patient is not. The question then arises: can a reasonable choice be made between the two? For many years the accepted wisdom for crossover trials was that it could be. We now examine this procedure.

The Two-Stage Procedure

In Example 1, the estimate of the carry-over effect, L_s, is not significant, and this suggests no particular evidence that carry-over has occurred. Why not, therefore, use the within-patient estimator of the treatment effect, L_c? On the other hand, if L_s had been significant, it might seem to be more appropriate to use L_p. Grizzle [22] proposed a general strategy for testing crossover trials as follows (see Figure 3). First perform a test for carry-over. If this is not significant, perform the within-patient test of the treatment effect. However, if the test for carry-over is significant, use the between-patient estimator. Because the power of the preliminary test for carry-over is typically low, Grizzle suggested using a higher (less stringent) nominal level of significance of 10%. This general procedure came to be known as the *two-stage procedure*, and for many years was the accepted way of analyzing crossover trials.

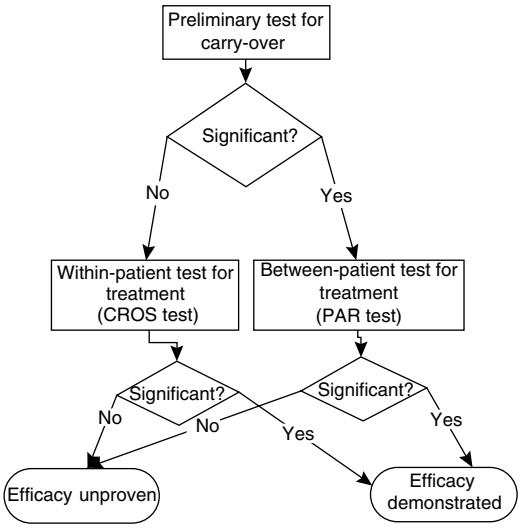

Figure 3 Schematic representation of the two-stage procedure

The fact that it now has few supporters amongst those researching in the field (for example, none of the three books devoted to crossover designs recommends it [25, 42, 44]) is due to the extremely important paper of Freeman's [16] which, analyzing the two-stage procedure as a whole, showed that it had a type I error rate of 7%–9.5% (depending on the correlation between successive measures, ρ for a claimed nominal level of 5%. (Note that, since we have claimed that carry-over cannot occur unless there is a treatment effect, a type I error can only be committed where there is no carry-over.) Freeman's result can be explained in a number of ways.

A simple explanation is that the test of carry-over, being a between-patient test, tends to show significance not only where there is genuine carry-over but also when, by chance, the patients in the two groups are very different. Under such circumstances, the last thing we should wish is to use a between-patient test of the effect of treatment, but this is precisely what the two-stage procedure leads us to do [44, 45]. An alternative explanation involves the variances of the contrasts. The argument for the two-stage procedure comes from considering the expectations of the contrasts, but considering their variances justifies the opposite procedure. As already discussed and as confirmed by the

main diagonal of the variance–**covariance matrix** given by Table 4, L_c has a much lower variance than that of L_p. Thus it is inevitable that they should often give very different answers. Furthermore, unless we should usually prefer the estimate given by L_c under the circumstance of their differing, there would be no point in using a crossover trial in the first place. However, $L_s = 2(L_p - L_c)$, so that, when the estimates differ, the contrast used for examining carry-over will tend to be large. Hence, based on the variances of the estimators, we should prefer CROS to PAR where SEQ is large, rather than vice versa.

A more formal explanation comes from considering the joint distribution of the three statistics. Table 4 gives not only the variances of the estimates but also their covariances. L_c and L_s are independent but L_p and L_s are highly correlated. The ratio of the covariance of L_p and L_s to the variance of L_s gives the **regression** of L_p on L_s, which is thus 1/2. In fact,

$$E[L_p|L_s] = \tau + \frac{L_s - \lambda}{2} \tag{3}$$

and the conditional variance is

$$\text{var}[L_p|L_s] = \frac{q\gamma^2}{2}. \tag{4}$$

Since L_c is independent of L_s, its expectation and variance are as given in Tables 3 and 4. We may note from (4), however, that conditional on L_s the variance of L_p is the same as that for L_c. This is, of course, an inevitable consequence of the fact that $L_p = L_c + L_s/2$. Furthermore, conditionally, L_p is not in general unbiased. Clearly, from (3) its bias is $(L_s - \lambda)/2$, and this is only zero where the carry-over is perfectly estimated. Of course, the expected value of $E[L_p|L_s]$ over the distribution of L_s is τ, but there are two objections to regarding this as having any relevance. The first is that in the two-stage procedure the use of L_p is *not* unconditional. Hence the unconditional expectation is irrelevant. Secondly, but for the possibility of carry-over, the value of L_s is a means of defining relevant subsets. Where such a relevant subset can be identified, one statistical viewpoint (that of Fisher [12] himself) is that the property of the set as a whole is irrelevant.

Table 4 Variance–covariance matrix for three contrasts

	CROS (L_c)	PAR (L_p)	SEQ (L_s)
CROS (L_c)	$q\gamma^2/2$		
PAR (L_p)	$q\gamma^2/2$	$q(\phi^2 + \gamma^2)$	
SEQ (L_s)	0	$q(2\phi^2 + \gamma^2)$	$q(4\phi^2 + 2\gamma^2)$

Whatever philosophical disagreements there may be about this point, however, the two-stage procedure as a whole has an inflated type I error rate because, whereas the conditional type I error rate of the CROS test (which will be used with probability 0.9) is 0.05 whatever the value of SEQ, that of the PAR test (which will be used with probability 0.1) is given by the integral over significant values of SEQ of the conditional type I error of PAR. This value lies between 0.25 and 0.5. The latter value arises if the within-patient variability is zero, so that $\gamma^2 = 0$. In that case there is a perfect correlation between L_p and L_s, the former being simply half the latter. Under such circumstances, the P value associated with PAR is the same as that with SEQ. Hence, since under the null hypothesis the P values have a **uniform distribution**, there is half a chance that a P value which is below 0.1 will be below 0.05. Hence if the P value for SEQ is significant at the 10% level, there is half a chance that the P value for PAR will be significant. Thus, the conditional type I error rate is 50%. Numerical integration produces the corresponding result for any combination of the values of γ^2 and ϕ^2, the other extreme being given where there is no between-patient error and $\phi^2 = 0$. This case gives a conditional type I error rate of 0.25. Putting these results together, the unconditional type I error rate for the two-stage procedure lies between $0.9 \times 0.05 + 0.1 \times 0.25 = 0.07$ and $0.9 \times 0.05 + 0.1 \times 0.5 = 0.095$. Further details are to be found in the paper by Freeman [16]. Although some commentators have since claimed that the two-stage analysis may be used after all [26], most regard it as no longer acceptable [42, 44, 45]. Tudor & Koch [52] have suggested performing the test for carry-over only if CROS is significant, and this, at least, would avoid the problem of the increase in type I error rate with the two-stage procedure. Note, however, that, when SEQ is significant, PAR and CROS will differ considerably, so that it is unwise to expect much comfort from a back-up test using PAR if the procedure is run this way round. Senn has shown how it is possible to correct the bias in the two-stage procedure by carrying out the PAR test (if used) at the 0.005 level rather than the 0.05 level but shows

that this has no power advantages over using CROS alone [46].

A Bayesian Approach

A very simple criticism of the two-stage analysis is possible from the Bayesian perspective (*see* **Bayesian Methods**). This is to note that the amount of information regarding carry-over is insufficient to resolve an initial genuine doubt as to its presence or not. Thus, first to pretest and then to behave either as if carry-over definitely is or is not present is incoherent. Grieve [19–21, 41], in a series of papers, has introduced an alternative Bayesian approach to the AB/BA design. This establishes the Bayesian posterior distribution of the treatment effect under the highly informative prior distribution that carry-over is zero (thus corresponding more or less to a CROS analysis) and also where the prior is uninformative (corresponding roughly to the PAR analysis). These posterior distributions can also be regarded as conditional distributions given particular models. However, the trialist can also express his prior belief in the validity of these two models, and this can be updated using evidence from the trial via a Bayes factor to produce posterior odds. The posterior odds can then be used to mix the two conditional posterior distributions to produce a single posterior. This technique has been extended by Grieve to cover the case with baselines also [20].

By permitting the adoption of intermediate positions, Grieve's approach is more flexible than the CROS analysis alone (which is a consistent Bayesian approach but corresponds to an extreme prior) whilst avoiding the pitfalls of the two-stage approach. However, it is not fully Bayesian since it does not model the dependence of belief in carry-over on belief in the effect of treatment (on which it must, in practice, depend strongly), and the appropriate prior weightings are a rather delicate matter. (It is necessary to give more prior weight to the case of no carry-over than would literally be believed to be true.)

Using Baseline Data

The difficulty in analyzing the AB/BA design is that there are not enough model **degrees of freedom** available to analyze the contrasts of interest whilst eliminating other parameters. One solution is to add information of a different type to the basic design. Such information may be provided by baseline information, i.e. measurements taken prior to treatment. As Ratkowsky et al. [42] point out, such measurements may be obtained before the first treatment period, at the end of the wash-out period which intervenes between the two treatment periods, and after wash-out at the end of the trial. The general scheme is illustrated in Figure 4. The third baseline value is most rarely taken and the first most commonly. The most common combination, however, especially in single-dose trials, is baselines 1 and 2. A cell mean parameterization corresponding to this latter case is given in Table 5.

Various authors have proposed various approaches to estimation of the treatment effect and its standard error [25, 39], depending upon which of the above baseline values are available and which of the following assumptions, if any, are made: (i) that the

Table 5 Cell mean expectations for baselines

		Period 1 ($k = 1$)	Period 2 ($k = 2$)
Sequence	AB ($i = 1$)	$\nu + \theta_1$ $\overline{X}_{1.1}$	$\nu + \theta_2 + \lambda_{A\_}$ $\overline{X}_{1.2}$
	BA ($i = 2$)	$\nu + \theta_1$ $\overline{X}_{2.1}$	$\nu + \theta_2 + \lambda_{B\_}$ $\overline{X}_{2.2}$

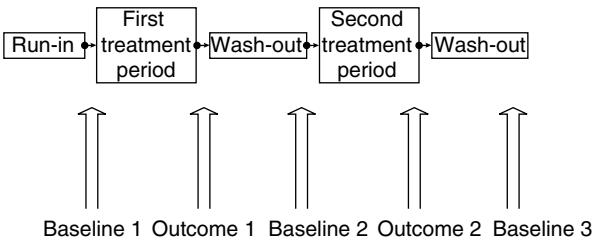

Figure 4 Schematic representation of outcome and baseline measurements in an AB/BA crossover trial

period effect applicable to a baseline measurement is the same as that which applies to the treatment which follows ($\theta_i = \pi_i$); (ii) that the effect of carry-over into a baseline value is the same as carry-over into a subsequent treatment ($\lambda_{A\_} = \lambda_{AB}$ and $\lambda_{B\_} = \lambda_{BA}$); and (iii) that baselines are measured on the same scale as outcomes and more particularly that the variances of outcomes and baselines are the same [var(X) = var(Y)]. (This latter assumption will not hold, for example, if some patients are placebo (*see* **Blinding or Masking**) responders and others are not.) Jones & Kenward [25] have pointed out, however, that assumptions (i) and (ii) are not necessarily reasonable and, indeed, one may also claim that assumption (iii) need not apply [44, 45].

In fact, it is probably fair to say that none of the solutions to the problem of carry-over by using baselines has achieved general assent. For example, if the carry-over into the baseline is assumed to be the same as that into the subsequent outcome, then the following within-patient estimator has expectation τ:

$$L_{bc} = \frac{\begin{bmatrix} (\overline{Y}_{11} - \overline{X}_{11}) - (\overline{Y}_{12} - \overline{X}_{12}) \\ +(\overline{Y}_{22} - \overline{X}_{22}) - (\overline{Y}_{21} - \overline{X}_{21}) \end{bmatrix}}{2}. \quad (5)$$

This is the same estimator as the conventional within-patient estimator CROS (L_c) associated with the design without baselines, but with the outcome values "corrected" by subtracting their respective baselines. However, unless the treatment period is short compared to the wash-out, then the condition that $\lambda_{A\_} = \lambda_{AB}$ and $\lambda_{B\_} = \lambda_{BA}$ is most unlikely to hold. Even then, the further assumption is required that carry-over of a treatment into a period where an active treatment is being given is the same (other things being equal) as carry-over into a period where no treatment is given. In fact, it is not hard to think of cases where either $\lambda_{A\_}$ or $\lambda_{B\_}$ would be large but (due to the longer time interval) λ_{AB} and λ_{BA} small. If that were the case, the bias in L_{bc} would be worse than the bias in L_c. For this reason Ratkowsky et al. [42] have proposed using the estimator

$$L_{bp} = (\overline{Y}_{11} - \overline{X}_{11}) - (\overline{Y}_{21} - \overline{X}_{21}), \quad (6)$$

but this is simply the standard PAR estimator, L_p, "corrected" for baselines. Hence, if this is used there is no point in carrying out a crossover trial unless the purpose is to study carry-over itself. (This would be an extremely wasteful use of the extra period.) Jones & Kenward [25] have proposed a complicated multistage testing process somewhat analogous to the Grizzle two-stage procedure, which uses the baselines to make the test for carry-over more powerful. This, however, may be criticized along rather similar lines to Freeman's criticism of the two-stage procedure and is not recommended.

In fact, considerations such as these suggest that, whilst there may be many cases where the AB/BA crossover may be safely analyzed by ignoring baselines altogether, there are not a few where attempting to use them would introduce problems with carry-over which could otherwise be avoided, and rather few where the problems with carry-over could be eliminated by resorting to baselines. This does not mean that baselines are useless, however. If patients are subjected to individual time trends, then trends in baseline values may be correlated with trends in outcome values. If carry-over is not a problem, then **analysis of covariance** using the baselines may bring a considerable reduction in variance. The estimator is of the form

$$L_{bc} = \frac{\begin{bmatrix} (\overline{Y}_{11} - \hat{\beta}\overline{X}_{11}) - (\overline{Y}_{12} - \hat{\beta}\overline{X}_{12}) \\ +(\overline{Y}_{22} - \hat{\beta}\overline{X}_{22}) - (\overline{Y}_{21} - \hat{\beta}\overline{X}_{21}) \end{bmatrix}}{2}, \quad (7)$$

where $\hat{\beta}$ is chosen so as to minimize the variance of L_{bc} and hence corresponds to the partial regression of outcomes on baselines. Details of the implicit error structure are covered by Jones & Kenward [25], the properties of this and other estimators are clearly reviewed by Koch [30], and further discussion and details of fitting are given by Senn [44, 45]. Although this estimator does require the assumption that neither baselines nor outcomes are subject to carry-over, it does not require the assumption that outcomes and baselines be measured on the same scale, nor that period effects for baselines are the same as for their corresponding outcomes.

Two-Treatment Designs in Three or More Periods

n-of-1 Trials

Adding further periods to the crossover design permits one to study further types of effect. One which is potentially important but which has received little attention is that of treatment by patient interaction.

The fact that a given patient receives the same treatment more than once may permit one to divide "within-patient" variation in the AB/BA design into two further sources: pure random variation from period to period and a personal response to treatment. The individual sequences which patients receive are sometimes referred to as "*n*-of-1 trials", and these are often performed with the object of analyzing each patient's results independently. However, if these trials are analyzed as a sequence using a random effects model, then the whole can be referred to as a multiperiod crossover design [44].

The Simple Carry-Over Model

By adding more periods to the AB/BA design, more model degrees of freedom are obtained and these may be used to estimate or eliminate the carry-over effect, provided that restrictive assumptions are made about its nature. The most popular set of assumptions are associated with the so-called simple carry-over model, namely, that carry-over will last for one period only and depends only on the engendering and not the perturbed treatment. Much investigation of optimality has been reported using this model [25, 28, 32, 36]. At one time this model appeared to have acquired an importance in the literature on the design of crossover trials which practical consideration did not justify; in more recent years it has come under heavy criticism [13, 14, 44]. For example, the model implies that the effect of a treatment reaches the steady state after two periods (and not after one or three periods). This, in turn, implies that the trialist has misjudged the length of a period, making it too short and that if, for example, a parallel group trial were used as an alternative, only the one-period direct effect of treatment would be captured, rather than the more relevant sum of the direct and the residual or carry-over effects. On the other hand, the much criticized AB/BA design, but with treatment periods twice as long, would capture the relevant treatment effect perfectly!

Dual Balanced Designs in Two Sequences

A dual sequence is obtained from another by interchanging the A and B treatment labels. So, for example, BA is the dual of AB and vice versa, and BAA is the dual of ABB, and so forth. Designs consisting only of such pairs and in which the two members of a pair appear equally are often referred to as dual-balanced. A particularly simple analysis is possible for designs consisting of a single balanced pair [25].

For example, a three-period design might allocate patients in equal numbers to the two sequences ABB and BAA. A four-period design might use the sequences ABBA and BAAB. Such designs can be analyzed using the two-sample t test as follows. Contrasts are found which, when compared between sequences, will eliminate subject, period and carry-over effects whilst estimating the treatment contrast of interest. For example, Table 6 shows cell means parameterizations for the AABB/BBAA design corresponding to four different models of carry-over. (The cells have been numbered so that

Table 6 Cell means representations of a two-treatment four-period design in two sequences for four different types of carry-over

Sequence and type of carry-over	Period			
	1	2	3	4
AABB	1	2	3	4
Simple	$\mu+\tau_A+\pi_1$	$\mu+\tau_A+\pi_2+\lambda_A$	$\mu+\tau_B+\pi_3+\lambda_A$	$\mu+\tau_B+\pi_4+\lambda_B$
Steady-state	$\mu+\tau_A+\pi_1$	$\mu+\tau_A+\pi_2$	$\mu+\tau_B+\pi_3+\lambda_{AB}$	$\mu+\tau_B+\pi_4$
Both	$\mu+\tau_A+\pi_1$	$\mu+\tau_A+\pi_2+\lambda_{AA}$	$\mu+\tau_B+\pi_3+\lambda_{AB}$	$\mu+\tau_B+\pi_4+\lambda_{BB}$
Neither	$\mu+\tau_A+\pi_1$	$\mu+\tau_A+\pi_2$	$\mu+\tau_B+\pi_3$	$\mu+\tau_B+\pi_4$
BBAA	5	6	7	8
Simple	$\mu+\tau_B+\pi_1$	$\mu+\tau_B+\pi_2+\lambda_B$	$\mu+\tau_A+\pi_3+\lambda_B$	$\mu+\tau_A+\pi_4+\lambda_A$
Steady-state	$\mu+\tau_B+\pi_1$	$\mu+\tau_B+\pi_2$	$\mu+\tau_A+\pi_3+\lambda_{BA}$	$\mu+\tau_A+\pi_4$
Both	$\mu+\tau_B+\pi_1$	$\mu+\tau_B+\pi_2+\lambda_{BB}$	$\mu+\tau_A+\pi_3+\lambda_{BA}$	$\mu+\tau_A+\pi_4+\lambda_{AA}$
Neither	$\mu+\tau_B+\pi_1$	$\mu+\tau_B+\pi_2$	$\mu+\tau_A+\pi_3$	$\mu+\tau_A+\pi_4$

they may be referred to subsequently.) If we let Y_{ijk} be the response for period k for subject j of sequence i and we calculate the linear combination $C_{ij} = \sum_{k=1}^{4} W_k Y_{ijk}$, where $w_1 = 6/20$, $w_2 = 4/20$, $w_3 = -7/20$, and $w_4 = -3/20$, then since $\sum_{k=1}^{4} W_k = 0$, C_{ij} has the subject effect eliminated from it. Furthermore, as may be seen by studying Table 6, if such a linear combination is calculated for a patient chosen at random from sequence 2 and subtracted from the corresponding contrast calculated from a patient in sequence 1, and simple carry-over applies, the expectation of the result is simply $\tau_A - \tau_B$. Thus a comparison of the means from the two sequences produces the contrast of interest and may be tested using the two-sample t test. A nonparametric approach would be to use the Wilcoxon–Mann–Whitney test [29].

The weakness of the simple carry-over model, however, is easily illustrated using this design, which is one of two supposedly optimal designs in four periods and two sequences. Suppose that B is a placebo but that A is an active treatment. The elimination of λ_A, the carry-over due to A, relies on the fact that 4/20 − 7/20 + 3/20 = 0, these being the weights which will be associated with cell means 2, 3, and 8. (Note that, in the construction of the final treatment estimate, weights for the cell means in the second sequence are the negative of those in the first.) However, **pharmacokinetic and pharmacodynamic** considerations might suggest that, if carry-over is important, it will have much more of an effect on cell mean 3, which measures the effect of a placebo after a double period of active treatment, than anywhere else in the design. Yet this is the mean which attracts the greatest weight. If the steady-state carry-over model applies instead, therefore, this scheme of weights will be seen to be quite undesirable.

With the steady-state carry-over model, the one-period assumption is retained, but it is assumed that a treatment will show no carry-over into itself. The relevant weights for this model are $w_1 = 1/4$, $w_2 = 1/4$, $w_3 = 0$, and $w_4 = -1/2$. On the other hand, to produce an estimator which guards against both kinds of carry-over, then $w_1 = 1$, $w_2 = -1/2$, $w_3 = 0$, and $w_4 = -1/2$. If the problem of carry-over is ignored altogether, then the weights are $w_1 = 1/4$, $w_2 = 1/4$, $w_3 = -1/4$, and $w_4 = -1/4$. Yet another possibility is to suppose that both kinds of carry-over apply but that we wish to estimate the total effect (direct + carry-over) of the treatments. Suitable weights would then be $w_1 = 0$, $w_2 = 1/2$, $w_3 = 0$, and $w_4 = -1/2$. However, since periods one and three are ignored altogether, the estimator is then identical to that which would be produced for an AB/BA crossover with periods twice as long. This illustrates the problem nicely, showing that, although more complicated designs appear to open up the possibility of eliminating carry-over, the reliance on assumptions is not easily banished.

In indexing the efficiency of the designs, it is often assumed that the within-patient errors are uncorrelated. Although some work has been done on the more general case where there is **autocorrelation**, it has been assumed either that there is no carry-over or only simple carry-over [31, 36]. Where standard assumptions of independence and homoscedasticity (*see* **Scedasticity**) apply, then the efficiency of a design/model combination is proportional to the sum of the squares of the weights associated with the cell means. Table 7 gives the variances of three possible designs in four periods. (These are three designs based on a single dual pair for which each patient receives each treatment twice. Such designs are sometimes referred to as being *uniform on the patients*. In the absence of carry-over they are more efficient than designs in which patients receive one treatment three times and the other once.) The Table gives the variance of the estimate for the treatment effect for each of four carry-over models as a ratio of the variance for the case where no carry-over is fitted.

As can be seen, if either the simple carry-over model is used, or the steady-state model is used, the first and third designs are the most efficient, the

Table 7 Variances of estimated treatment effects for three designs and four models as a ratio of the case where carry-over is not fitted

Design	Simple	Steady-state	Both	Neither	Cumulative
AABB/BBAA	1.1	1.5	6	1	2
ABAB/BABA	5.5	5.5	5.5	1	
ABBA/BAAB	1.1	2	6	1	

second being much more inefficient. However, if both forms of carry-over are to be eliminated, then the second design is most efficient. On the other hand, if the total or cumulative effect under the steady state is of interest, the first design is the only possibility.

More Complicated Designs

There is no need to restrict the treatment sequences to a single dual pair, and designs based on the use of four sequences can prove superior. As above, however, the optimality of a design depends on the carry-over model. In nearly all of the literature the simple carry-over model is assumed with no investigation of its suitability, and, indeed, in the purely medical literature, such designs are hardly ever applied. To drug developers, the steady-state model will seem at least as plausible as the simple model.

Where a more general model for the correlation structure applies, identifying "optimal" designs becomes quite complex. There have been a number of extensive investigations. However, quite apart from the difficulties with carry-over described above, there is the added problem that optimal designs depend on the correlation structure, and this will not be known in advance nor with any certainty even after running the trial. Matthews [37] covers the estimation of dispersion parameters.

Another paper of Matthews [38] gives a good general review of the field. Fleiss [13, 14] and Senn [44] have provided critical attacks on this general topic. Jones & Kenward [25] and Ratkowsky et al. [42] give further details of models and fitting. An alternate approach to modeling carry-over has come from the field of pharmacokinetic and pharmacodynamic (PK/PD) modeling, where residual effects and direct effects are treated in one comprehensive PK/PD model. Sheiner et al. [49] have used this to investigate the relative performance of crossover, dose-escalation, and parallel group studies in dose-finding.

Designs for Three or More Treatments in the Same Number of Periods

It is not uncommon to run crossover trials in three or more periods. Indeed, in drug development, contrary to popular opinion, except in the field of bioequivalence, such designs are more common than the AB/BA design. For example, they are very popular in certain fields as dose-finding studies, using designs comparing several doses to a placebo. A very common way of designing such trials is to choose sequences which, taken together, form one or more **Latin squares**. Example 2 described such a design, comparing two doses of diclofenac to placebo using six sequences, or two Latin squares. A four-treatments crossover trial comparing treatments A, B, C, and D might allocate patients in equal numbers to sequences such as

A C D B,

B D C A,

C B A D, and

D A B C.

Such a design is not only uniform on the periods (each treatment appears equally often in every period), as is any Latin square, but it is sometimes called *balanced*, since each treatment appears an equal number of times (in this case once) after every other. (This is, however, a rather confusing term since balance is used generally by medical statisticians in a different and less specialized sense, and is given a similar but nonetheless somewhat different meaning in describing **incomplete blocks**.) Such balanced Latin squares are known as Williams squares, and are more efficient if the simple carry-over model is being fitted [53]. However, this does not necessarily make them the best choice for dealing with carry-over (for similar reasons to those discussed above) [44].

Some alternative Latin squares may also be preferable for certain types of analysis. Consider, for example, the square

A B C D,

B A D C,

C D A B,

D C B A.

By ignoring treatments D and C and pairing off sequences 1 & 2 and then again 3 & 4, this design can be reduced to two AB/BA crossovers. On the other hand, for the purpose of comparing the treatments A and C, treatments B & D could be ignored and sequences 1 & 3, and 2 & 4 can be paired, and so on. This means that any technique which is available

for analyzing the AB/BA design (nonparametric, for **binary** data, and so forth) can be used to analyze contrasts for such a design. The results from the two subdesigns thus formed can be pooled using standard **meta-analytic** techniques.

Designs for Three or More Treatments in Fewer Periods

Occasionally it is of interest to study more treatments than can realistically be given to a single patient. Incomplete blocks designs are then an option, as discussed in Example 3. This design is uniform on the periods and balanced in the sense of incomplete blocks designs (i.e. each of the 21 possible pairs of treatments is given equally frequently to the patients). However, the design is not balanced in the carry-over sense since each treatment does not follow every other equally frequently. However, in this example, wash-out was adequate and carry-over was not adjusted for.

If, as was the case in this example, balancing for carry-over is ignored, then such designs are simply examples of incomplete blocks designs, about which there is an extensive literature. If it is desired to balance for simple carry-over, then added complications are involved. In principle, it is also possible to recover interblock information for such designs. (In fact even for complete blocks (*see* **Randomized Complete Block Designs**), if carry-over is fitted, since this induces some non**orthogonality**, there is some interblock information available [3].) In practice it seems common to ignore this refinement and, indeed, some see advantages in having pure within-patient estimators, as distributional assumptions regarding patient effects are then irrelevant.

Computer Analysis

With the help of modern statistical packages (*see* **Software, Biostatistical**), the analysis of continuous outcomes for crossover designs is relatively straightforward. An approach using SAS® is as follows. Suppose we have a design where a number of treatments A, B, C, D, etc. are being compared. First of all the outcome data (OUTCOME) are arranged into an $n \times k$ vector, where n is the number of patients and k is the number of periods. For each such observation the corresponding patient number (PATIENT), period (PERIOD), and treatment (TREAT) are recorded as categorical variables. The analysis via ordinary **least squares** can then be carried out with the help of proc glm® as follows:

```
proc glm;
class PATIENT PERIOD TREAT;
model OUTCOME = PATIENT PERIOD TREAT;
estimate "A−B" TREAT 1   −1    0   0 (etc.);
estimate "A−C" TREAT 1    0   −1   0 (etc.);
etc.
run;
```

If it is desired to fit simple carry-over, then this can be done by defining an additional variable CARRY which has one more level than the number of treatments. This is then coded A, B, etc., depending on the treatment given in the previous period. An extra code, say Z, can be used whenever the particular observation is at the beginning of a sequence. The "class" and "model" statements then need to include this term.

It should be noted, however, that for designs in more than two periods, the analysis using ordinary least squares produces residual degrees of freedom for error in excess of the number of patients. This is an indication that strong assumptions are involved and, in the presence (say) of patient by treatment interaction, this analysis may be invalid [40, 44].

Further SAS® code for various types of analysis is given by Ratkowsky et al. [42] and Senn [44], who also covers alternatives to ordinary least squares as well as nonparametric approaches.

Other Outcomes

This survey has been entirely in terms of continuous outcomes. These have a much greater relative importance for crossover trials than for parallel group trials, where, unlike for crossover trials (but see below), **survival**, for example, can be an important outcome. Nevertheless, there is a growing body of work on binary and other outcomes (*see* **Categorical Data Analysis**). All that will be attempted here is a very brief summary of the literature.

We have also concentrated on the analysis of continuous outcomes using the general linear model. In fact, various nonparametric approaches have been

introduced following Koch's original paper [29]. Various simple strategies are discussed by Senn [44], and an authoritative review listing all the major approaches is given by Tudor & Koch [52]. In general, these techniques are really partially parametric since various linear manipulations of the data have to be undertaken before proceeding to the final "nonparametric" test. Confidence limits are also available for these approaches.

A simple test which may be used for binary outcomes in the AB/BA design is the Mainland–Gart test [17, 34]. The most important modern work in this field has been by Jones & Kenward [25], who applied **loglinear models** and who have since discussed both marginal and subject-specific models in great detail [27]. Ezzet & Whitehead [7] have introduced an alternative random effects **proportional-odds model** for **ordered categorical** (and hence also **binary**) data. Again some simple techniques are described by Senn [44].

Occasionally, survival-type data are obtained from crossover trials. For example, patients may be asked to undertake an exercise test, and such observations can then be **censored**. A model has been proposed by France et al. [15]. An alternative approach is given by Feingold & Gillespie [8].

The crossover trial also provides the possibility of examining directly patient preferences for treatment. An important reference is Baskerville et al. [1], and the subject has recently been treated by Lindsey & Jones [33].

Further Reading

Three books devoted to crossover designs have already been referred to [25, 42, 44]. A complete issue of *Statistical Methods in Medical Research* (Vol. 3, no. 4, 1994) was devoted to crossover designs, and provides review articles on the AB/BA design [45], binary and categorical data [27], nonparametric methods [52], multiperiod crossovers [38], and Bayesian approaches [21].

References

[1] Baskerville, J.C., Toogood, J.H., Mazza, J. & Jennings, B. (1984). Clinical trials designed to evaluate therapeutic preferences, *Statistics in Medicine* **3**, 45–55.

[2] Chassan, J.B. (1964). On the analysis of simple cross-overs with unequal numbers of replicates, *Biometrics* **20**, 206–208.

[3] Chi, E.M. (1991). Recovery of interblock information in cross-over trials, *Statistics in Medicine* **10**, 1115–1122.

[4] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*. Wiley, New York.

[5] Cushny, A.R. & Peebles, A.R. (1905). The action of optical isomers. II. Hyoscines, *Journal of Physiology* **32**, 501–510.

[6] Ebbut, A.F. (1984). Three-period cross-over designs for two treatments, *Biometrics* **40**, 219–224.

[7] Ezzet, F. & Whitehead, J. (1991). A random effects model for ordinal responses from a cross-over trial, *Statistics in Medicine* **10**, 901–907.

[8] Feingold, M. & Gillespie, B.W. (1996). Cross-over trials with censored data, *Statistics in Medicine* **15**, 953–967.

[9] Fieller, M.A. (1940). The biological standardization of insulin, *Journal of the Royal Statistical, Society Supplement* **7**, 1–64.

[10] Finney, D.J. (1956). Cross-over designs in bioassay, *Proceedings of the Royal Society, Series B* **145**, 42–60.

[11] Fisher, R.A. (1925). *Statistical Methods for Research Workers*, reprinted in *Statistical Methods, Experimental Design and Scientific Inference*, H. Bennett, Ed. Oxford University Press, Oxford, 1990.

[12] Fisher R.A. (1956). *Statistical Methods and Scientific Inference*, reprinted in *Statistical Methods, Experimental Design and Scientific Inference*, H. Bennett, Ed. Oxford Scientific Publications, Oxford, 1990.

[13] Fleiss J.L. (1986). Letter to the editor, *Biometrics* **42**, 449–450.

[14] Fleiss, J.L. (1989). A critique of recent research on the two-treatment cross-over design, *Controlled Clinical Trials* **10**, 1121–1130.

[15] France, L.A., Lewis. J.A. & Kay, R. (1991). The analysis of failure time data in cross-over studies, *Statistics in Medicine* **10**, 1099–1113.

[16] Freeman, P.R. (1989). The performance of the two-stage analysis of two-treatment, two-period cross-over trials, *Statistics in Medicine* **8**, 1421–1432.

[17] Gart, J.J. (1969). An exact test for comparing matched proportions in cross-over designs, *Biometrika* **56**, 75–80.

[18] Graff-Lonnevig, V. & Browaldh, L. (1990). Twelve hours bronchodilating effect of inhaled formoterol in children with asthma: a double-blind cross-over study versus salbutamol, *Clinical and Experimental Allergy* **20**, 429–432.

[19] Grieve, A.P. (1985). A Bayesian analysis of the two-period cross-over trial, *Biometrics* **41**, 979–990.

[20] Grieve, A.P. (1994). Extending a Bayesian analysis of the two-period crossover to allow for baseline measurements, *Statistics in Medicine* **13**, 905–929.

[21] Grieve, A.P. (1994). Bayesian analyses of two-treatment crossover studies, *Statistical Methods in Medical Research* **3**, 407–429.

[22] Grizzle, J.E. (1965). The two-period change-over design and its use in clinical trials, *Biometrics* **21**, 467–480. (Corrigenda in Grizzle (1965) *Biometrics* **30**, 727 and Grieve, A. P. (1982) *Biometrics* **38**, 517.).

[23] Hills, M. & Armitage, P. (1979). The two-period cross-over trial, *British Journal of Clinical Pharmacology* **8**, 7–20.

[24] Irwin, J.O. (1937). Statistical method applied to biological assays, *Journal of the Royal Statistical, Society Supplement* **7**, 1–48.

[25] Jones, B.J. & Kenward, M.G. (1989). *Design and Analysis of Cross-Over Trials*. Chapman & Hall, London.

[26] Jones, B.J. & Lewis J. (1994). The case for cross-over trials in phase III, *Statistics in Medicine* **14**, 1025–1038.

[27] Kenward, M.G. & Jones, B. (1994). The analysis of binary and categorical data from cross-over trials, *Statistical Methods in Medical Research* **3**, 325–344.

[28] Kershner, R.P. & Federer, W.T. (1981). Two-treatment cross-over designs for estimating a variety of effects, *Journal of the American Statistical Association* **76**, 612–619.

[29] Koch, G.G. (1972). The use of nonparametric methods in the statistical analysis of the two-period change-over design, *Biometrics* **28**, 577–584.

[30] Koch, M.A. (1992). Precision and bias of baseline adjusted estimators in the 2×2 cross-over, in *American Statistical Association 1992 Proceedings of the Section on Biopharmaceuticals*. American Statistical Association, Alexandria, pp. 68–73.

[31] Kunert, J. (1985). Optimal repeated measurements designs for correlated observations and analysis by least squares, *Biometrika* **72**, 275–379.

[32] Lasker, E.M., Meisner, M. & Kushner, H.B. (1983). Optimal crossover designs in the presence of carryover effects, *Biometrics* **39**, 1089–1091.

[33] Lindsey, J.K. & Jones, B. (1996). A model for cross-over trials evaluating therapeutic preferences, *Statistics in Medicine* **15**, 443–447.

[34] Mainland, D. (1963). *Elementary Medical Statistics*. W.B. Saunders, Philadelphia.

[35] Marks, H.P. (1925). The biological assay of insulin preparations in comparison with a stable standard, *British Medical Journal*, December 12, 1102–1104.

[36] Matthews, J. (1989). Optimal cross-over designs for the comparison of two treatments in the presence of carry-over effects and autocorrelated errors, *Biometrika* **74**, 311–320.

[37] Matthews, J.N.S. (1989). Estimating dispersion parameters in the analysis of data from cross-over trials, *Biometrika* **76**, 239–244.

[38] Matthews, J. (1994). Multi-period cross-over trials, *Statistical Methods in Medical Research* **3**, 383–405.

[39] Patel, H.I. (1983). The use of baseline measurements in the two-period cross-over design, *Communication in Statistics – Theory and Methods* **12**, 2693–2721.

[40] Preece, D.A. (1982). T is for trouble (and textbooks): a critique of some examples of the paired-samples t-test, *Statistician* **31**, 169–195.

[41] Racine, A., Grieve, A.P., Flühler, H. & Smith, A.F.M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion), *Applied Statistics* **35**, 93–150.

[42] Ratkowsky, D.A., Evans, M.A. & Alldredge, J.R. (1993). *Cross-over Experiments, Design, Analysis and Application*. Marcel Dekker, New York.

[43] Selwyn, M.R., Dempster, A.P. & Hall, N.R. (1981). A Bayesian approach to bioequivalence for the 2×2 changeover, *Biometrics* **37**, 11–21.

[44] Senn, S.J. (1993). *Cross-Over Trials in Clinical Research*. Wiley, Chichester.

[45] Senn, S.J. (1994). The AB/BA crossover: past, present and future?, *Statistical Methods in Medical Research* **3**, 303–324.

[46] Senn, S.J. (1996). The AB/BA cross-over: How to perform the two stage analysis if you can't be persuaded that you shouldn't, in *Liber Amicorum Roel van Strik*, B. Hansen & M. de Ridder, Eds. Erasmus University, Rotterdam, pp. 93–100.

[47] Senn, S.J. & Richardson, W. (1994). The first t-test, *Statistics in Medicine* **13**, 785–803.

[48] Senn, S.J., Lillienthal, J., Patalano, F. & Till, D. (1996). An incomplete blocks cross-over in asthma: a case study in collaboration, in *Cross-Over Trials*, L. Hothorn, Ed. Fischer, Stuttgart.

[49] Sheiner, L.B., Hasimoto, Y. & Beal, S.L. (1991). A simulation study comparing studies for dose ranging, *Statistics in Medicine* **10**, 303–322.

[50] Simpson, T.W. (1938). Experimental methods and human nutrition, *Journal of the Royal Statistical Society, Series B* **5**, 46–69.

[51] Student (1908). The probable error of a mean, *Biometrika* **6**, 1–25.

[52] Tudor, G. & Koch, G.G. (1994). Review of nonparametric methods for the analysis of crossover studies, *Statistical Methods in Medical Research* **3**, 345–381.

[53] Williams, E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatments, *Australian Journal of Scientific Research* **2**, 149–168.

[54] Yates, F. (1938). The gain in efficiency resulting from the use of balanced designs, *Journal of the Royal Statistical Society, Series B* **5**, 70–74.

(*See also* **Clinical Trials, Overview**; **Covariate Imbalance, Adjustment for**)

STEPHEN SENN

Cross-sectional Study

A cross-sectional study is a study to estimate the distribution of a quantity of interest (or joint distribution of several quantities) in a **target population**, at a certain moment in time. Ideally, this is accomplished by measurements from a **random** or **stratified random sample** of the target population, although convenience samples may also be used. A cross-sectional study is characterized by the fact that only one set of observations is taken from each subject, as opposed to a longitudinal study in which study participants provide observations at more than one point in time (*see* **Cohort Study**; **Panel Study**). Even if repeated or serial cross-sectional studies are conducted in the same population, the same individuals generally will not be sampled again except by chance.

Cross-sectional studies often have a **binary** variable as the quantity of primary interest, such as estimating the **prevalence** of a certain disease, risk factor, or health behavior. Continuous variables may also be of interest, for example the subject's weight or blood cholesterol level, although such data are often grouped or categorized in epidemiologic studies. One use of a cross-sectional study is to determine the association between an outcome variable and some **explanatory variables**, for example to estimate the prevalence of a disease, perhaps as a function of some explanatory variables. An association between outcomes and explanatory variables may suggest causality, although a causal link usually cannot be established from a single cross-sectional survey (*see* **Causation**; **Hill's Criteria for Causality**), because such studies give no information on the temporal ordering of possibly causal events. A second use of cross-sectional studies is to monitor changes in a population over time using a series of cross-sectional surveys; for example, the Monitoring the Future surveys [7] track drug use by teenagers over time in this way. Sometimes a better case for a causal link can be made with serial cross-sectional data; for example, Pirkle et al. [9] analyzed blood lead measurements from the second and third National Health and Nutrition Examination Surveys (NHANES), conducted in 1976–1980 and 1988–1991, respectively, to document a decline in blood lead levels in the US population that resulted from the gradual removal, since 1976, of most lead from gasoline. A third use of cross-sectional studies is to make some **inference** about disease incidence; this is harder to achieve, although some techniques for estimating incidence from a cross-sectional survey are discussed below.

As the distribution that is being estimated may be changing over time, the ideal cross-sectional study would be conducted instantaneously. However, a real study typically requires some time to conduct. In practice, the primary requirement is that the target distribution changes negligibly over the course of the study. In some situations this requirement is easily met: for example, a study of the prevalence of carpal tunnel syndrome can safely be conducted over an entire year; a study of the prevalence of varicella (chicken-pox) probably should be conducted within a week or two. In some situations cross-sectional surveys are conducted over a long period as a surveillance system (*see* **Surveillance of Diseases**). For example, the **Centers for Disease Control and Prevention (CDC)**, use ongoing **telephone surveys** to assess the prevalence of several chronic disease risk factors and tracks these over time [11]. These studies should be distinguished from incidence studies in which all new occurrences of a disease arising in a certain population in a given period are recorded; a defining characteristic of a cross-sectional survey is that subjects are sampled solely on the basis of their membership in a target population, not on the basis of a change in status. This distinction is somewhat blurred, however, in cross-sectional surveys that also collect retrospective data on duration of disease, since this type of data would allow identification of new cases.

Although a cross-sectional study can only measure a distribution at a single time point, interest is often centered on dynamic or time-dependent quantities; disease incidence is often the quantity of interest. One circumstance in which disease incidence can be measured from a cross-sectional survey occurs when an ephemeral state associated with recent onset can be identified. If the **mean** duration, w, of this state is known, then the prevalence of individuals in the ephemeral state can be converted into disease incidence using the relation prevalence$/w =$ incidence. For a disease of short duration, the ephemeral state can be the entire course of the disease. For this method to be valid, the time-scale over which the disease incidence changes must be longer than w. For example, the serologic testing algorithm for recent HIV seroconversion [6] estimates the incidence of HIV(human immunodeficiency virus) infection in a

population by identifying the proportion of individuals who test positive for HIV on a standard(sensitive) screening assay but who will still test negative on a less sensitive assay. Weinstock et al. [13] used this approach to estimate HIV incidence among patients at clinics for sexually transmitted diseases. Note that although this method is sensitive to the value of w, it is possible to test for or estimate trends in incidence in serial cross-sectional surveys by comparing the prevalence of individuals in the ephemeral state over time. For such a comparison to provide valid inference on trends in incidence, it is not necessary to know w; we need know only that it is small compared with the time between the surveys [12].

The methods described above are appropriate for situations in which the incidence is changing over time. In many situations the population may be considered homogeneous over time; in these situations, interest may focus on the age of onset or on age-specific incidences. Keiding [8] discussed a variety of assumptions that allow estimation of these quantities from a cross-sectional survey, and also discussed methods to estimate disease prevalence from incidence data (*see* **Incidence–Prevalence Relationships**). Some of these methods require either retrospective data or external data such as age-specific mortality for the general population or for people with the specific disease of interest. Additional information about population dynamics can sometimes be obtained if two or more cross-sectional studies are conducted in the same population at different times. For example, it is possible to estimate the incidence of a disease by comparing two measurements of disease prevalence taken at different times and accounting for the aging of the population. By comparing the proportion of 15-year-olds with 19-year-olds who have a positive ppd (purified protein derivative) TB test with the proportion five years later among 20- to 24-year-olds, one could estimate rates of infection with *M. tuberculosis*, in essence treating these two groups as members of the same "pseudo-cohort". Techniques are available for converting these "cohort infection rates" to age-specific period rates [10]. However, the validity of this calculation depends on the assumption that there is no differential loss of people with disease (i.e. mortality or migration caused by the disease); otherwise, external data on the these effects are necessary. In addition, for different age groups to be treated as members of the same pseudo-cohort, the distribution of important **covariates** must be the same across age groups, a condition that may pose a special challenge for **observational studies**. For example, in a study of childbearing women, the women aged 15–20 years may have markedly different demographic characteristics than those aged 20–25 years.

Conclusions about incidence can sometimes be drawn by comparing crude (i.e. not age-specific) prevalences from successive cross-sectional surveys. For example, HIV prevalence among injection drug-users in Bangkok, Thailand, jumped from 1% at the start of 1988 to 32%–43% by August–September 1988 [14], implying a remarkable incidence of HIV infection over this period, as well as illustrating the usefulness of cross-sectional surveys as a surveillance tool. However, in less dramatic situations, trends in prevalence may be difficult to interpret because they are the net result of new cases and death or the loss of old cases. As with analyses of age pseudo-cohorts, knowledge of or assumptions on the nature of the death or loss of prevalent cases are required before inference on incidence can be made. An unchanging crude prevalence also does not necessarily imply steady-state conditions; Batter et al. [1] report an example where crude prevalence was steady over time but the distribution of cases by age had shifted.

The line between cross-sectional studies and a wide variety of **retrospective studies** is blurred when retrospective **longitudinal data** are collected in a cross-sectional survey. Examples of such data include age at menarche, the number of months breast-fed, prevalence of diarrhea in the last two weeks, or conditions surrounding the death of a child. The validity of such retrospective data is dependent on the respondent's ability to recall accurately the events of interest for the time-period considered. For example, to examine the association between short birth intervals and the survival of the subsequent child, it is possible to use birth history reports of mothers [5]. This same type of data can also be used to estimate fertility and mortality patterns many years before the study for analysis of long-term trends.

Several major differences between cross-sectional studies and longitudinal (or cohort) studies determine when either should be used. **Cohort studies** measure the effect of risk factors on disease incidence in a defined population, whereas cross-sectional studies measure the effects of risk factors on disease prevalence. As a result, differences may arise when the same **associations** are studied by the two methods. Factors that affect both disease incidence and

mortality after disease will show a different association when measured cross-sectionally than when measured longitudinally. For example, a risk factor that is associated with both increased incidence and increased mortality among individuals with disease will show a smaller association with disease prevalence in a cross-sectional survey than with disease incidence in a longitudinal study, because the persons with the risk factor will be less likely to survive until the time of the cross-sectional study (*see* **Case–Control Study, Prevalent**). Even factors unrelated to disease incidence may show an association with disease prevalence if they affect the survival of individuals with disease differentially. Thus, if disease etiology is of primary interest, cohort studies or incident **case–control studies** are usually preferable, whereas prevalence studies may provide more useful information for characterization of a population for public health purposes.

Cross-sectional studies sample prevalent, rather than incident, cases; the selection requirement of survival to the date at which the survey is conducted results in differences between the population of prevalent cases and the population obtained by following incident cases over time. The likelihood of being observed in a certain transient stage in a cross-sectional survey is proportional to the time spent in that stage (*see* **Length Bias**). As a result of heterogeneity in the course of disease, individuals with longer survival times are more likely to be sampled in a cross-sectional study than those with shorter survival times. Similarly, the population of prevalent cases who have a given characteristic may differ from the population of those who have ever developed that characteristic. A cross-sectional estimate of the relative prevalence of two types of cancer (one virulent, the other less so) would not be equal to the relative incidences, as the relative incidence of the virulent type would be greater than its relative prevalence. Although the results of the cross-sectional study may properly reflect the distribution of survival times or disease subtypes among those individuals currently living with disease, these quantities must be interpreted as instantaneous pictures of a dynamic, open population and not necessarily as reflective of some other population, such as those with incident disease (*see* **Biased Sampling of Cohorts**; **Screening Benefit, Evaluation of**).

Another distinction between cross-sectional and longitudinal studies is that a cross-sectional study or series of cross-sectional studies can only address aggregate changes in the population; unless the appropriate retrospective data are available, it is not possible to measure change at the individual level. For example, two cross-sectional surveys may find approximately the same proportion of respondents used a condom during their last sexual contact; to plan a public health campaign to increase condom usage, we may want to know additionally whether some respondents always use condoms and some never do, or if all individuals sometimes use condoms. Such data are most reliably obtained from a longitudinal study. In some cases, however, the closed nature of the longitudinal study is a disadvantage. For example, studies of HIV incidence conducted longitudinally often find a decreasing incidence of new HIV infections which may not represent the trend in the general population but instead represent a depletion of high-risk individuals in the cohort, as well as a change in behavior among study participants, who receive counseling on how to reduce their risk of acquiring HIV infection.

Cost is often a deciding factor that favors cross-sectional studies over longitudinal studies. This is especially true for studies of rare diseases, where very large cohorts or long follow-up may be required to observe enough cases to obtain statistically significant results. A variety of split-panel designs combine elements of both cross-sectional and longitudinal studies. Further discussion of these designs, as well as a general discussion of what types of questions can be answered with a cross-sectional study and what questions require a longitudinal study, can be found in Curtin & Feinleib [3] and Dwyer & Feinleib [4].

References

[1] Batter, V., Matela, B., Nsuami, M., Manzila, T., Kamenga, M., Behets, F., Ryder, R.W., Heyward, W.L., Karon, J.M. & St Louis, M.E. (1994). High HIV-1 incidence in young women masked by stable overall seroprevalence among childbearing women in Kinshasa, Zaire: estimating incidence from serial seroprevalence data, *Journal of Acquired Immune Deficiency Syndrome* **8**, 811–817.

[2] Brookmeyer, R. & Quinn, T.C. (1995). Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests, *American Journal of Epidemiology* **141**, 166–172.

[3] Curtin, L. & Feinleib, M. (1992). Considerations in the design of longitudinal surveys of health, in *Statistical*

Models for Longitudinal Studies of Health, J.H. Dwyer, M. Feinleib, P. Lippert & H. Hoffmeister, eds. Oxford University Press, New York.

[4] Dwyer, J. & Feinleib, M. (1992). Introduction to statistical models for longitudinal observation, in *Statistical Models for Longitudinal Studies of Health*, J.H. Dwyer, M Feinleib, P. Lippert & H. Hoffmeister, eds. Oxford University Press, New York.

[5] Hobcraft, J., McDonald, J. & Rutstein, S. (1985). Demographic determinants of infant and early child mortality, *Population Studies* **39**, 363–385.

[6] Janssen, R.S., Satten, G.A., Stramer, S., Rawal, B.D., O'Brien, T.R., Weiblen, B.J., Hecht, F.M., Jack, N., Cleghorn, F.R., Kahn, J.O., Chesney, M.A. & Busch, M.P., (1998). New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes, *Journal of the American Medical Association* **280**, 42–48.

[7] Johnston, L.D., O'Malley, P.M. & Bachman, J.G. (1996). National Survey Results On Drug Use From the Monitoring the Future Study, 1975–1995. Vol. 1: Secondary School Students, NIH Pub. No. 97–4139. National Institute on Drug Abuse, Rockville.

[8] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[9] Pirkle, J.L., Brody, D., Gunter, E.W., Paschal, D.C., Flegal, K.M. & Matte, T.D. (1994). The decline in blood lead levels in the United States: The National Health and Nutrition Examination Surveys, *Journal of the American Medical Association* **272**, 284–291.

[10] Preston, S. & Coale, A. (1982). Age structure, growth, attrition and accession: a new synthesis, *Population Index* **48**, 217–259.

[11] Remington, P.C., Smith, M.Y., Williamson, D.F., Anda, R.F., Gentry, E.M. & Hogelin, G.C. (1988). Design, characteristics and usefulness of state-based behavioural risk factor surveillance: 1981–1987, *Public Health Reports* **103**, 366–375.

[12] Satten, G.A., Janssen, R.A. & Stramer, S. et al. (2001). Development and validation of a serologic testing algorithm for recent HIV seroconversion, in *Quantitative Evaluation of HIV Prevention Programs*, E.H. Kaplan & R. Brookmeyer, eds. Yale University Press, New Haven.

[13] Weinstock, H., Dale, M., Gwinn, M., Satten, G.A., Kothe, D., Mei, J, Royalty, J., Linley, L., Fridlund, C., Parekh, B., Rawal, B.D., Busch, M.P., & Janssen, R.S. (2002). HIV seroincidence among patients at clinics for sexually transmitted diseases in nine cities in the United States, *Journal of Acquired Immune Deficiency Syndromes* **29**, 478–483.

[14] Weniger, B.G., Limpakarnjanarat, K., Ungchusak, K., Thanprasertsuk, S., Choopanya, K., Vanichseni, S., Uneklabh, T., Thongcharoen, P. & Wasi, C. (1991). The epidemiology of HIV infection and AIDS in Thailand, *Journal of Acquired Immune Deficiency Syndrome* **5**, Supplement 2, S71–S85.

GLEN A. SATTEN &
LAURENCE GRUMMER-STRAWN

Cross-validation

Cross-validation is one of several methods for error assessment of a statistical method. Alternatives include the **jackknife method**, the **bootstrap method**, Akaike's AIC (*see* **Akaike's Criteria**), and others.

The main ideas are simply illustrated by the statistical discrimination (also called classification) problem, discussed in detail in [5], for example. There one has, say, two populations, with "training samples" available from each, and it is desired to assign a new observation to one of the two populations. For a given discrimination scheme (which is based on the training data), the error rate, i.e. probability of misclassification, is a useful indicator of its performance (*see***Multivariate Classification Rules: Calibration and Discrimination**). Estimating the error rate by the proportion of misclassifications when the rule is used to classify the training data is usually inappropriate. This is because of an "optimistic bias", caused by the classification rule being fine-tuned for this particular realization of the data, which usually entails somewhat worse performance for a different realization from the same underlying distribution.

The idea behind the cross-validatory approach to this bias problem is to separate the data into two pieces, one of which is used to construct the classifier, and the other to "validate" or estimate the error rate of the classifier. Dependence on the particular dichotomy chosen is eliminated by averaging error rates over a large number of such dichotomies. The "cross" part of the name "cross-validation" comes from the fact that each data point is sometimes used for classifier construction, and at other times for validation.

A commonly used dichotomy is called "leave one out", where the validation set is a single observation, and the rest are used to construct the classification rule. In some situations, there is an advantage to using, say v, observations in the validation set (i.e. "leaving out v" in construction of the classifier). See Picard & Berk [9] for an interesting illustration of this point (in a different statistical setting). This form has been called "v-fold" cross-validation.

The ideas illustrated via the discrimination problem above apply to a wide variety of statistical problems. See [10] for a general formulation. See that paper and [1] for a good indication of the breadth of different contexts in which this principle applies, as well as for historical background. There was a period of very active research in the mid 1970s, but the method dates from well before.

An area where a large literature has developed around the cross-validation idea is in smoothing parameter (i.e. window width or bandwidth) selection for nonparametric curve estimation, i.e. smoothing methods. Rather different perceptions of the performance of cross-validatory methods have been developed, depending on the smoothing method used. Performance is viewed as "acceptable" by most of those who prefer smoothing splines (*see* **Spline Function**), and "unacceptable" by most of those who prefer kernel/local polynomial methods. The reason for this is unclear (because the essence of both smoothing methods is rather similar), but perhaps it is because spline researchers tend to work with "less noisy" data sets than kernel/local polynomial researchers.

In the context of smoothing splines, Wahba and co-workers have proposed, and demonstrated good properties of, a variation of cross-validation called "generalized cross-validation"; see [11], for example.

The deepest insights into the efficacy of cross-validation for smoothing parameter selection are available in the context of kernel density estimation; see [8] and references therein. Here the conclusion drawn by most researchers is that cross-validatory methods are "too variable", but this variability can be reduced to acceptable levels by alternative measures such as plug-in methods or the use of smoothed cross-validation. Similar ideas hold in kernel regression, as shown, for example, in [3, 6], and [7].

When data are dependent, cross-validation needs to be modified for use in smoothing parameter selection [4]. For deeper analysis, see [2].

References

[1] Bailey, R.A., Harding, S.A. & Smith, G.L. (1989). Cross-validation, in *Encyclopedia of Statistical Sciences*, Supplement Volume, S. Kotz, N.L.Johnson & C.B. Read, eds. Wiley, New York, pp. 39–44.

[2] Chu, C.K. & Marron, J.S. (1991). Comparison of two bandwidth selectors with dependent errors, *Annals of Statistics* **19**, 1906–1918.

[3] Gasser, T., Kneip, A. & Köhler, W. (1991). A flexible and fast method for automatic smoothing, *Journal of the American Statistical Association* **86**, 643–652.

[4] Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*,

Springer Lecture Notes in Statistics, Vol. 60. Springer-Verlag, Berlin.

[5] Hand, D.J. (1981). *Discrimination and Classification*. Wiley, Chichester.

[6] Härdle, W., Hall, P. & Marron, J.S. (1988). How far are automatically chosen regression smoothing parameter selectors from their optimum?, *Journal of the American Statistical Association* **83**, 86–101.

[7] Härdle, W., Hall, P. & Marron, J.S. (1992). Regression smoothing parameters that are not far from their optimum, *Journal of the American Statistical Association* **87**, 227–233.

[8] Jones, M.C., Marron, J.S. & Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association* **91**, 401–407.

[9] Picard, R.R. & Berk, K.N. (1990). Data splitting, *American Statistician* **44**, 140–147.

[10] Stone, M.C. (1978). Cross-validation: a review, *Mathematische Operationsforschungs, Series: Statistics* **9**, 127–139.

[11] Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Number 59. Society for Industrial and Applied Mathematics, Philadelphia.

(*See also* **Classification, Overview**)

J.S. MARRON

Crude Risk

Crude risk is the probability that an individual will develop a particular disease in a given time interval in the presence of other **competing risks** of death. For example, the probability that a 30-year-old woman will develop breast cancer between the ages of 30 and 60 is a crude risk. The crude risk is reduced by the fact that she may die of other diseases before she develops breast cancer. The term **absolute risk** is used synonymously with crude risk. Crude risk can be estimated without making special assumptions, such as the "independence" assumption used in competing risk analysis. Crude risk can be contrasted with the net risk in the theory of competing risks. Net risk refers to the probability of developing a particular disease if other competing risks are eliminated.

Crude risk is also used to describe the risk of disease in a heterogeneous population composed of different genders and age groups, for example. Crude risk is differentiated, in this context, from gender- and age-specific risks.

(*See also* **Aalen–Johansen Estimator**)

MITCHELL H. GAIL

Cumulative Hazard

The cumulative hazard on the interval $[0, t)$ is $\int_0^t \lambda(u)\,du$, where $\lambda(u)$ is the **hazard rate**. If $\lambda(u)$ is the hazard for total mortality, then the cumulative hazard is related to cumulative probability of death, $1 - \exp\left(-\int_0^t \lambda(u)\,du\right)$. If $\lambda(u)$ is a disease-specific **incidence rate** and if $\int_0^t \lambda(u)\,du$ is small, then the cumulative hazard approximates the "pure" probability of developing disease in the absence of other competing causes of death, provided that those other causes act independently of the cause of interest.

(*See also* **Competing Risks**; **Nelson–Aalen Estimator**; **Survival Analysis, Software**; **Survival Distributions and Their Characteristics**)

MITCHELL H. GAIL

Cumulative Incidence Rate

The cumulative **incidence rate** is a **cumulative hazard** and corresponds to the special case in which the hazard refers to the incidence rate for a specific disease. For small incidence rates, the cumulative incidence rate approximates the "pure" probability of developing the disease in the absence of competing causes of death (*see* **Competing Risks**) and should be distinguished from the crude probability of developing the disease in the presence of competing causes of death (*see* **Absolute Risk**; **Crude Risk**).

MITCHELL H. GAIL

Cumulative Incidence Ratio

The cumulative incidence ratio is the ratio of the **cumulative incidence** in an exposed cohort to that in an unexposed cohort over the same time period. The cumulative incidence ratio is the same as the **relative risk**.

(*See also* **Cohort Study**)

MITCHELL H. GAIL

Cumulative Incidence

Cumulative incidence is the proportion of individuals in a cohort initially free of a given disease who develop that disease in a defined age or time interval. Cumulative incidence is a **crude risk** and is synonymous with the terms cumulative risk, **risk**, and absolute risk. Sometimes cumulative incidence refers to the number, rather than the proportion, of individuals in a cohort initially free of a given disease who develop the disease in a given age or time interval.

(*See also* **Aalen–Johansen Estimator**)

MITCHELL H. GAIL

Cure Models

Most approaches to the analysis of survival, or time to event, data implicitly assume that, with sufficient follow-up, all subjects would experience the event of interest. However, there are situations when it is expected that a fraction of subjects will not experience the event. In the clinical setting, this often corresponds to the assumption that a fraction of patients treated for a disease will be cured of the disease under treatment, whereas the rest will experience a recurrence or adverse event of some sort.

Early work on models for data arising in such situations was done by Boag [2], Berkson & Gage [1], and Haybittle [12]. More recently, interest in such models has focused on the incorporation of **covariates**, or **explanatory variables**, into such models. The models can be characterized by defining a binary variable Y, where $Y = 1$ indicates that a subject will experience the event of interest and $Y = 0$ otherwise (*see* **Dummy Variables**). If $\mathbf{X}$ corresponds to a vector of explanatory variables, with x_i representing the observed values for the ith subject, then $p(x_i)$ can represent the probability of the event occurring for the ith subject, i.e. $\Pr(Y = 1|x_i)$, and $f(t|Y = 1, x_i)$ can represent the probability density function for the random variable T, which specifies the time of the event if it occurs. It is convenient also to specify the corresponding survivor function $S(t|Y = 1, x_i) = \Pr(T > t|Y = 1, x_i)$.

If the functions p and f have specified parametric forms, then **maximum likelihood** estimation of unknown parameters is possible. The **likelihood** is a product of contributions of the form $p(x_i)f(t|Y = 1, x_i)$ from individuals who experience the event, $1 - p(x_i) + p(x_i)S(t|Y = 1, x_i)$ from individuals observed to time t without experiencing the event, and, in some cases, $1 - p(x_i)$ if an individual is known not to have experienced the event. The latter contributions exist only if follow-up has continued past a time point before which it is known that the event must occur if it is to occur.

A variety of parametric forms have been considered. To incorporate covariates, the logistic model

$$p(x_i) = \exp(\alpha + x_i\boldsymbol{\beta})/[1 + \exp(\alpha + x_i\boldsymbol{\beta})],$$

where $\boldsymbol{\beta}$ is an appropriately defined vector of regression coefficients and α is a scalar location parameter, is convenient (*see* **Logistic Regression**). Farewell [4, 5] combined this with a *Weibull* regression model for f, but other choices are possible. For example, Larson & Dinse [14], in a **competing risk** framework, used a **proportional hazards** model with a step function for the baseline hazard, and Yamaguchi [22] used a class of **accelerated failure-time models**. Some specific results for **exponential** models were given by Ghitany and coauthors [9, 10].

In many situations, this mixture model approach to time-to-event data has a natural appeal. Some special considerations arise, however. There is an **identifiability** problem, since there can be a high **correlation** between the intercept term of the logistic model and any shape parameters in the model for f [6, 15]. This manifests itself in a nonquadratic and relatively flat **profile likelihood** function for α [6]. The extent of the problem depends on the generality of the model for f, but is minimized if there are a sizable number of censored observations at times well past the period when most events occur. A formal specification of this requirement was considered by Maller & Zhou [17]. Also, tests for the existence of a population with $Y = 0$, perhaps of particular interest when it would correspond to a "cured fraction" of patients or the presence of individuals in a population immune from some disease, involves nonstandard maximum likelihood theory as the null hypothesis lies on the boundary of the parameter space [23]. A conservative approach is to restrict the use of the models to situations in which there is external evidence for the existence of two populations. This is quite restrictive, however, and a general recommendation for caution in their use is perhaps sufficient to enable a wider application. For example, the models have been used to establish, in an informal manner, that there is little evidence for separate populations under plausible parametric assumptions [8, 19].

Recently, there has been interest in relaxing the parametric dependence of these models. Maller & Zhou [16] examined the estimator defined by the "flattening out" level of a **Kaplan–Meier** estimated survivor curve. This has been compared with parametric models [11]. Kuk & Chen [13] combined a logistic regression model for p with a proportional hazards model for f. Their approach estimates

the regression parameters of the proportional hazards model through an approximation to a marginal rank likelihood and then obtains an estimate of the baseline hazard function. Taylor [20] used an **EM algorithm** approach, as developed by Larson & Dinse [14], to replace the parametric event time distribution by a Kaplan–Meier type estimator. Although this work was not extended to include covariates in the event time distribution, it was shown to be quite efficient compared with a parametric alternative. Taylor [20] recommended that the survivor function S be forced to zero beyond the last event. Some such restriction is helpful to avoid nonidentifiability problems which will be potentially more acute with nonparametric approaches. It is not too restrictive when a mixture model is plausible and some follow-up is available at times in the upper tail of the event time distribution. Taylor [20] suggested that, if these conditions do not prevail, then the suitability of the model might be questioned in any event.

Comparisons of the use of a mixture model to the application of the commonly used relative risk regression model of Cox [3] have been made [5, 21]. Advantages associated with the mixture model relate to prediction [21] and to a simpler specification of covariate effects [5]. From an empirical point of view, the usual regression models for survival data will be useful even when a mixture model is appropriate. However, if only a subset of a population is expected to experience the event, the adoption of a formal mixture model may be more realistic and, through separate modeling of the probability of the event and the time to the event, more informative.

A comprehensive discussion of survival data with long-term survivors is provided by Maller & Zhou [18]. The general framework corresponding to "cure" models can also be applied to other types of data – for example, when noncure is evidenced by the observation of a nonzero count variable [7].

References

[1] Berkson, J. & Gage, R.P. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association* **47**, 501–515.

[2] Boag, J.W. (1948). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society, Series B* **11**, 15–44.

[3] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] Farewell, V.T. (1977). A model for a binary variable with time-censored observations, *Biometrika* **64**, 43–46.

[5] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* **38**, 1041–1046.

[6] Farewell, V.T. (1986). Mixture models in survival analysis: are they worth the risk?, *Canadian Journal of Statistics* **14**, 257–262.

[7] Farewell, V.T. & Sprott, D.A. (1988). The use of a mixture model in the analysis of count data, *Biometrics* **44**, 1191–1194.

[8] Farewell, V.T., Coates, R.A., Fanning, M.M., MacFadden, D.K., Read, S.E., Shepherd, F.A. & Struthers, C.A. (1992). The probability of progression to AIDS in a cohort of male sexual contacts of men with HIV disease, *International Journal of Epidemiology* **21**, 131–135.

[9] Ghitany, M.E. & Maller, R.A. (1992). Asymptotic results for exponential mixture models with long term survivors, *Statistics* **23**, 321–336.

[10] Ghitany, M.E., Maller, R.A. & Zhou. S. (1994). Exponential mixture models with long term survivors and covariates, *Journal of Multivariate Analysis* **49**, 218–241.

[11] Ghitany, M.E., Maller, R. & Zhou, S. (1995). Estimating the proportion of immunes in censored samples: a simulation study, *Statistics in Medicine* **14**, 39–49.

[12] Haybittle, J.L. (1965). A two parameter model for the survival curve of treated cancer patients, *Journal of the American Statistical Association* **53**, 16–26.

[13] Kuk, A.C. & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression, *Biometrika* **79**, 531–541.

[14] Larson, M.G. & Dinse, G.E. (1985). A mixture model for the regression analysis of competing risks data, *Applied Statistics* **34**, 201–211.

[15] Laska, E.M. & Meisner, M.J. (1992). Nonparametric estimation and testing in a cure model, *Biometrics* **48**, 1223–1234.

[16] Maller, R.A. & Zhou, S. (1992). Estimating the proportion of immunes in a censored sample, *Biometrika* **79**, 731–739.

[17] Maller, R.A. & Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data, *Journal of the American Statistical Association* **89**, 1499–1506.

[18] Maller, R.A. & Zhou, S. (1996). *Survival Analysis with Long Term Survivors*. Wiley, New York.

[19] Struthers, C.A. & Farewell, V.T. (1989). A mixture model for time to AIDS data with left truncation and an uncertain origin, *Biometrika* **76**, 814–817.

[20] Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture models, *Biometrics* **51**, 899–907.

[21] Taylor, J.M.G. & Kim, D.K. (1993). Statistical models for analysing time-to-occurrence data in radiobiology

and radiation oncology, *International Journal of Radiation Biology* **64**, 627–640.

[22] Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in Japan, *Journal of the American Statistical Association* **87**, 284–292.

[23] Zhou, S. & Maller, R.A. (1995). The likelihood ratio test for the presence of immunes in a censored sample, *Statistics* **27**, 181–201.

VERN T. FAREWELL

Cutler, Sidney Joshua

Born: April 13, 1917, in Odessa, Russia.
Died: October 21, 1993, in Silver Spring, Maryland.

Cutler immigrated to the US in 1923, graduated from City College of New York in 1938, received a master's degree in sociology from Columbia University in 1941, and a doctorate in epidemiology from the University of Pittsburgh in 1961. During World War II he served with the US Army in Europe.

A major part of Cutler's career as a biostatistician and epidemiologist, from 1948 to 1975, was with the National Cancer Institute, **National Institutes of Health** in Bethesda, Maryland. In the early 1950s Cutler authored a series of reports, including a monograph [5] on the 1947–48 incidence of and mortality from cancer in 10 metropolitan areas of the US. Twenty years later he was to direct a similar follow-up study. In 1954–55, nine years before the landmark US Surgeon General's Report *Smoking and Health* [7], Cutler published an overview of the strong epidemiologic evidence linking cigarette smoking to lung cancer [1] (*see* **Smoking and Health**), and, with Donald Loveland, an assessment of a smoker's lifetime probability of developing lung cancer, including the probable age of lung cancer onset [3].

In the late 1950s Cutler organized and became director of the End Results Evaluation program of the National Cancer Institute, a reporting system to which a number of cancer registries in the US contributed, on an ongoing basis, uniformly defined data on individual cases of cancer and their survival. The purpose of the program, later to become the SEER Program (Surveillance, Epidemiology, and End Results), was to develop data on population rates of incidence, mortality, and survival from cancer, that would allow comparisons of the results of various forms of therapy, of various regions of the country, and time trends. At a meeting in Bethesda in 1959, chaired by Michael Shimkin, representatives of six national cancer registration programs organized an international cooperative effort in the evaluation of end results of cancer therapy and in the investigation of epidemiologic questions on the effects of climate, diet, and other environmental factors on cancer incidence. Results of this effort were reported at a symposium in Norway in 1963 [6]. One of the findings from the international comparisons was that survival rates from mammary tumors were about the same in England and the US, although the method of treatment differed: in England the treatment was limited surgery, in the US it was radical surgery. The use of radical mastectomy decreased in the US subsequently.

As an aid in the analysis of survival data, Cutler developed the *relative survival rate*, a method still in use in the 1990s, as a way of correcting the survival rate for normal mortality [4]. In 1958 Cutler, with Fred Ederer, published a paper that explained in a form understandable by nonstatisticians how to describe the survival experience of cancer patients, how to include **censored data** in this description, and why it is important to include censored data [2]. The paper became a standard tool in **teaching statistics to medical students**. In the first 20 years after its publication, it was referred to in the medical literature 530 times, for which it won a citation award. In the 1960s Cutler, together with a number of other statisticians from the National Institutes of Health **Jerome Cornfield**, William Haenszel, Nathan Mantel, and **Marvin Schneiderman**), taught a course on current problems in public health at the University of Pittsburgh. During that time Cutler, together with Schneiderman and Samuel Greenhouse, another statistician from the National Institutes of Health, designed the first randomized trial of a screening agent (*see* **Screening Trials**), a study to determine whether periodic mammographic screening prevents death from breast cancer; the conduct of the trial was led by Sam Shapiro of the Health Insurance Plan of Greater New York.

Cutler was a Fellow of the **American Statistical Association**.

References

[1] Cutler, S.J. (1955). A review of the statistical evidence on the association between smoking and lung cancer, *Journal of the American Statistical Association* **50**, 267–282.

[2] Cutler, S.J. & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival, *Journal of Chronic Diseases* **8**, 699–712.

[3] Cutler, S.J. & Loveland, D.B. (1954). The risk of developing lung cancer and its relationship to smoking, *Journal of the National Cancer Institute* **15**, 201–211.

[4] Cutler, S.J., Griswold, M.H. & Eisenberg, H. (1957). An interpretation of survival rates: cancer of the breast, *Journal of the National Cancer Institute* **19**, 1107–1117.

[5] Dorn, H.F. & Cutler, S.J. (1959). *Morbidity from Cancer in the United States; Part I and Part II Combined*. US Department of Health, Education, and Welfare, Public Health Monograph 56. Government Printing Office, Washington.

[6] Shimkin, M.B. (1977). *Contrary to Nature*. DHEW Publication No. (NIH) 76–720. US Department of Health, Education, and Welfare, National Institutes of Health, Washington, p. 425.

[7] US Department of Health, Education, and Welfare (1964). *Smoking and Health*. Report of the Advisory Committee to the Surgeon General of the Public Health Service. US Government Printing Office, Washington.

FRED EDERER

Data Access, National and International

Data and information are fundamental ingredients of health situation analysis. A national health information system helps safeguard against deterioration of the health of the population, provides data for research to improve health, and provides information for management of the health care system. Additionally, the system supports and enhances understanding of how health and well-being impact on the economy and other social institutions. The data collected should be those which are needed for determining a population's health status, aiding medical research, preventing and controlling diseases, assisting in decision-making, framing health policies, organizing or reorganizing health services, and informing the population about their state of health. Indicators of mortality, natality, morbidity, health services, and resource availability are among those which are useful for this purpose. Additionally, special data needs have arisen in recent years due to the increased incidence of several diseases as well as the alarming appearance of new diseases such as Lyme disease and acquired immune deficiency syndrome (**AIDS**) resulting from human immunodeficiency virus (HIV). At the same time, interest in international comparisons of health data has increased (*see* **Mortality, International Comparisons**) and, therefore, gaining access to relevant and timely data has become an important health priority.

National Sources of Data

Vital statistics, hospital discharge data, and health manpower resource statistics are collected and disseminated in virtually all of the industrialized countries and many of the developing countries. Most countries collect and publish these data on an annual basis through government statistical agencies. Access to these data vary from country to country. Published annual or other periodic statistics reports are the most common form of dissemination; however, many countries have electronic files available for public use.

The US **National Center for Health Statistics (NCHS)** has developed an *International Health Data Reference Guide* [2] which provides information on the availability and sources of selected national vital statistics, hospital discharge data, health manpower resources, and population-based health survey statistics (*see* **Surveys, Health and Morbidity**). As of 1996, official agencies of 44 industrialized nations provided information about the availability of selected health data for their country. The names, addresses, and facsimile (fax) numbers of these agencies are listed in the *Guide* to facilitate requests for data. This publication is available upon request from the National Center for Health Statistics, 6525 Belcrest Road, Hyattsville, MD 20782, USA.

Countries publish health data in varying detail and scope, ranging from simple summary measures of births and deaths to complex tabulations and analyses of a wide range of health variables. In the US, for example, two major annual health statistics publications are prepared, in addition to many one-off or special topic reports:

1. *Health, US* – a comprehensive report on the health status of the nation. It presents national trend data health status and determinants, supply and utilization of health resources, health care resources, and health care expenditures [1].
2. *Vital Statistics of the US* – a compilation of mortality, natality, marriage, and divorce data with extensive demographic and geographic detail [4]. A monthly vital statistics report is available providing monthly and cumulative provisional data [3].

A few countries release data to the public through a combination of publications, public use electronic data files, and unpublished tabulations. An even smaller number of countries, including the US, make data available on floppy diskettes, CD-ROM, and the **Internet**.

International Sources of Data

In the early 1980s, the **World Health Organization (WHO)** launched the Global Strategy for Health for All by the Year 2000. The Strategy has the aim of making possible the attainment, by all citizens of the world by the year 2000, of a level of health that will permit them to lead a social and economically productive life. Through the need to change priorities of information support and the need to reflect progress in implementation as ascertained by

the monitoring and evaluation processes, the need for health data including international comparisons has increased considerably.

United Nations

The Statistical Division of the United Nations (UN) has supplied basic statistical data for demographers, economists, public-health workers, and sociologists for almost 50 years. A prime publication is the *Demographic Yearbook*, which is published annually [8].

The *Demographic Yearbook* is a comprehensive collection of international demographic statistics (*see* **Demography**) which features the results of population **censuses**. Through the cooperation of national statistical services, official demographic statistics are presented for about 233 countries or areas throughout the world. Tables are presented giving a world summary of basic demographic statistics, followed by tables presenting statistics on the size, distribution, and trends in population, natality, fetal mortality, **infant mortality**, **maternal mortality**, general mortality, marriages, and divorces.

The *Yearbook* is available in published format and on magnetic tape. A database known as the Demographic and Social Statistics Database containing data previously published in the *Yearbook* from 1950 is available on floppy disks for use on microcomputers. Contact may be made to the Director, Statistics Division, United Nations, New York, NY 10017, USA, for further information on purchasing data.

World Health Organization

The World Health Organization (WHO) is a specialized agency of the United Nations with primary responsibility for international health matters and public health. Through this organization, which was created in 1948, the health professionals of more than 180 countries exchange their knowledge and experience. As part of WHO's mandate to establish and maintain statistical services and to provide information in the field of health, they publish the *World Health Statistics Annual* [9].

The *Annual* provides a compilation of data reported by Member States on deaths by cause, age, and sex, detailed statistics on selected causes of death, information on cases of deaths from notifiable diseases, and other data of medical and public health interest. The *Annual* is available in published format, and the more detailed mortality databases are available electronically over the Internet. For further information, contact the World Health Organization, Geneva, Switzerland.

Pan American Health Organization

The Pan American Health Organization (PAHO) is a regional office of the World Health Organization that provides data and information about the countries of the Americas. An annual publication *Health Statistics from the Americas* complements the quadrennial publication of *Health Conditions in the Americas* [6, 7]. The *Health Statistics* publications present a mortality database, estimated sex–age-specific death rates by broad groups of causes, and historical summaries of reported cases of selected **communicable diseases**. The *Health Conditions* publication presents a regional overview of the health situation with an annex of health and development indicators and country reports summarizing some salient conditions and problems for each country.

These two sources of data are available in published format and electronically over the Internet. Contact may be made to the Pan American Health Organization, World Health Organization, 525 Twenty-third Street, NW, Washington, DC 20037, USA.

Organization for Economic Cooperation and Development

The Organization for Economic Cooperation and Development (OECD), Paris, France, has developed a set of health data files designed to facilitate macroeconomic analysis of health care systems in the 24 industrialized OECD member countries. The files entitled OECD HEALTH DATA cover data on expenditures, hospitalization, demography, **life expectancy**, death rates, socioeconomic environment, compensation of health care professionals, length of stays in hospitals by **diagnosis-related groups (DRG)**, frequency of selected medical procedures, and fee schedules [5].

The information in this data bank is available for analysis through a software package designed for use on microcomputers. The data are updated when national administrations release new statistics

or revise old ones, usually on an annual basis. Contact may be made to the Publications and Information Centres, Organization for Economic Cooperation and Development, 2 Rue André-Pascal, 75775 Paris Cedex 16, France.

While sources for health data for countries around the world are increasing, and interest grows in drawing conclusions about international differences, it is important to bear in mind that a number of factors must be carefully considered before deciding that health information from different countries is, in fact, comparable. Among these factors are:

1. completeness of the coverage and reliability of the data
2. lack of standardization of data collection methods and definitions of terms
3. base population differences, e.g. the noninstitutionalized population vs. the total resident population
4. coding and tabulation differences.

References

[1] National Center for Health Statistics. *Health, United States*. Public Health Service, Hyattsville (published annually).

[2] National Center for Health Statistics. *International Health Data Reference Guide*. Public Health Service, Hyattsville (published biennially).

[3] National Center for Health Statistics. *Monthly Vital Statistics Report*. Public Health Service, Hyattsville (published monthly).

[4] National Center for Health Statistics. *Vital Statistics of the United States, Vols. I, II*. Public Health Service, Hyattsville (published annually).

[5] Organization for Economic Cooperation and Development (1993).. *OECD Health Data: A Software Package for the International Comparison of Health Care Systems: Users' Manual*. Organization for Economic Cooperation and Development, Paris.

[6] Pan American Health Organization (1994). *Health Conditions in the Americas*. PAHO, Washington.

[7] Pan American Health Organization (1994). *Health Statistics from the Americas*. PAHO, Washington.

[8] United Nations (1996). *1994 Demographic Yearbook*. United Nations, New York.

[9] World Health Organization. *World Health Statistics Annual*. WHO, Geneva (published annually).

(*See also* **Administrative Databases**; **Health Services Data Sources in Canada**; **Health Services Data Sources in Europe**; **Health Services Data Sources in the US**)

JACQUELINE P. DAVIS

Data and Safety Monitoring

Phase III clinical trials (*see* **Clinical Trials, Overview**) usually have a (**DSMB**), with broad responsibility for monitoring the conduct of the trial. The responsibilities of the Board typically include insuring scientific integrity and patient safety, monitoring the occurrence of adverse events, and assessing efficacy. As the study progresses, the Board discusses these issues and makes recommendations to the study investigators regarding the conduct of the trial, possibly including a recommendation regarding early termination of the study.

To simplify the discussion of this monitoring activity we suppose that the trial is a randomized, double blind (*see* **Blinding or Masking**) study comparing an experimental treatment to either a placebo or a standard treatment. Initially, we also assume that efficacy is measured primarily by a single, univariate primary endpoint, allowing that numerous secondary measures may also be available (*see* **Oblimin Rotation**).

Monitoring the scientific integrity of the study involves reviewing information such as violations of the protocol, recommendations for modifying the protocol (*see* **Clinical Trials Protocols**), evaluations of data quality (for example, errors identified at data entry specific to each form, how many errors have been corrected and how many are outstanding) (*see* **Clinical Trials Audit and Quality Control**), recruitment, timeliness of follow-up, drop-outs and **Adjustment for Noncompliance**.

A primary responsibility of the DSMB is to insure that the trial is sufficiently safe to warrant continued participation of the patients. This activity normally involves a qualitative review of adverse events. In view of the varied and somewhat unpredictable nature of the adverse events that may ultimately become of concern, formal stopping rules may be less helpful for assessing safety than for assessing efficacy. However, one might form a combined safety endpoint, such as the occurrence of any major adverse event (suitably defined), and then evaluate this endpoint in the context of a stopping boundary. At the least, such an approach might provide the DSMB insights as to the frequency with which an observed imbalance between groups would occur by chance, and this in turn might aid in arriving at a decision whether or not to terminate the study.

Formal stopping rules have found particular usefulness in assessing efficacy. A large number of approaches have been proposed. We will focus here on the commonly used group sequential methods, all of which assume a classical statistical framework. **Decision theoretic** approaches are also discussed. Before describing the mechanics of implementing interim analyses for efficacy, it is important to consider the larger context in which these analyses occur. There are numerous considerations that affect the decision to terminate a study early, and the crossing of a statistical boundary is only one. Thus, the actual decision may not be the same as the decision suggested by the statistical test.

Reasons for desiring early termination of a trial are rather obvious. Certainly, if the evidence in favor of an experimental therapy is overwhelming, there is an ethical need to stop the trial and provide the drug to all study participants as well as other patients (*see* **Ethics of Randomized Trials**). There is also a clear financial incentive to stopping early under these circumstances, both in terms of the costs of the study as well as revenues which may accrue from sales of a new drug or medical device. Early termination may also release patients for other trials, especially if the reason for terminating early is convincing evidence that the drug is ineffective. Finally, early termination because of overwhelming evidence of treatment efficacy may facilitate initiation of additional follow-up studies to understand further the nature of the treatment effect.

Thus, there are important reasons to consider stopping early. However, there are also arguments against early termination. The consequent reduction in sample size may prevent a definitive evaluation of important secondary endpoints or evaluation of subgroups, analyses which might shed important light on the nature of the treatment effect (*see* **Treatment-covariate Interaction**). When the primary endpoint is survival, or time to a specified event (*see* **Survival Analysis, Overview**), a dramatic early effect may prove to be transitory, with the survival curves eventually coming together or even crossing, and early termination may cause this phenomenon to go unobserved.

As we will discuss in greater detail, early termination introduces a **bias** into conventional estimates of

the magnitude of the treatment effect. Although statistical techniques are available to adjust estimates, they are complex, thus compounding the difficulty of explaining these issues and adjustments to a nonstatistical audience.

Group Sequential Tests

The desirability of performing tests on accumulating data while a study is ongoing has long been recognized. Early work on sequential methods (*see* **Sequential Analysis**) focused on procedures which called for testing after each observation was realized, and sequential probability ratio tests (SPRT) were a cornerstone of this approach. It consists of computing the ratio of the **likelihood** functions under the **null** and **alternative hypotheses** (both specified as simple hypotheses) and comparing this to upper and lower boundaries. Depending on which boundary is crossed, one accepts the corresponding hypothesis. Fully sequential designs with a **binary** outcome were proposed by Bross [9] and Armitage [4].

Although a large body of theory relating to fully sequential designs has been developed, these methods are infrequently used. One limiting feature is that there is typically no upper bound on the number of observations that may be required to reach a decision. An additional problem is the logistical impracticality of testing after each observation.

These limitations led to consideration of group sequential tests, where tests are performed only periodically during the course of the study. The idea for this approach dates back to Armitage et al. [5] and Samuel-Cahn [36, 37]. A simple strategy for performing group sequential tests was proposed independently by Haybittle [21] and by Peto et al. [32]. The approach is to perform interim tests at a very stringent level of significance, and then test at the nominal level at the end of the study if the trial was not terminated early. For example, one might test at the 0.01 level if only one or two interim tests are planned, or at the 0.001 if more than two are planned. Although the overall probability of type I error (*see* **Hypothesis Testing**) will exceed the nominal level, the excess should be small. More recently, methods have been developed which provide accurate control over the type I error rate, and which give the investigator a wide choice of boundary shapes.

Some Commonly Used Procedures

In describing the most commonly used group sequential tests, we start with the simplest situation, in which **normally distributed** endpoints are observable immediately after treatment, $K - 1$ interim tests are planned, and n subjects are recruited into each arm of the study between successive tests. Let X_{ij} represent the observation on the ith subject receiving placebo therapy during the jth period of recruitment (between the $j-1$ and jth tests), $i = 1, \ldots, n$, $j = 1, \ldots, K$. It is assumed that the X variables are independently distributed with a common normal distribution having mean μ_P and known variance σ^2. Similarly, let $Y_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, K$, represent the observations on the experimentally treated patients, which are normally distributed with mean μ_E and variance σ^2.

For $k = 1, \ldots, K$, let T_k represent the usual t test statistic (*see* **Student's *t* Distribution**) based on the data available after the kth recruitment period:

$$T_k = \frac{\sum_{j=1}^{k} \left(\overline{Y}_{\cdot j} - \overline{X}_{\cdot j}\right) (n/2k)^{1/2}}{\sigma}.$$

The procedures described below can all be expressed in terms of boundaries depending on a set of constants determined by the choice of K and the overall type I error rate, $\alpha : C(1, \alpha), \ldots, C(K, \alpha)$. After collecting data up to the kth test ($k = 1, \ldots, K$) one computes T_k, and if the test statistic exceeds $C(k, \alpha)$ one concludes that the trial may be terminated with overall probability of type I error less than α.

Three of the most commonly used boundaries are listed below. For $k = 1, \ldots, K$:

1. Constant boundary (Pocock boundary [33]):

$$C(k, \alpha) = C_P(K, \alpha).$$

2. Monotone decreasing boundary (O'Brien and Fleming boundary [29]):

$$C(k, \alpha) = C_{OF}(K, \alpha) \left(\frac{K}{k}\right)^{1/2}.$$

3. Intermediate boundary (Fleming et al. [17]). Let π_k represent the probability of a type I error occurring at the kth test. $C_{FHO}(k, \alpha)$ is defined

such that $\pi_1 = \cdots = \pi_{K-1}$ and $\pi_1 + \cdots + \pi_K = \alpha$. Notice that the probability of a type I error occurring at each of the interim tests equals the nominal level of the first interim test.

As seen in Table 1, the Pocock (P) boundary offers the greatest opportunity for stopping at the first interim test. Conversely, it is the most stringent of the three boundaries at the final test if the study is not terminated early. This can result in an awkward situation for a DSMB, when the nominal P value at the end of the study is less than 0.05, but is not sufficiently small to achieve significance as defined by the group sequential boundary. For a fixed number of interim tests and equal sample sizes between each test, the P boundary gives the smallest average sample size among the three boundaries. However, for designs with equivalent maximal sample size, it gives the lowest **power**. Put differently, it requires the largest maximal sample size to achieve a specified power.

The O'Brien–Fleming (OF) boundary requires very strong evidence of an effect to terminate at the first interim test, whereas the criteria at the final test are rather close to those for a single sample design (that is, a design with no interim testing). This feature may be viewed as desirable in the sense that monitoring committees typically want very convincing evidence that a treatment effect is real before terminating a study very early (for all the reasons discussed previously), but desire a criterion close to the single sample test if early termination does not occur. However, the boundary at the first test is judged to be too extreme in some applications, particularly if the number of interim tests exceeds two. In this case, one may modify the decision rule to stop after the first test if $P < 0.001$. The effect of this modification on the overall type I error rate is negligible.

Table 1 Nominal P values for overall type I error of $0.05(\alpha = 0.05)$ Pocock, O'Brien–Fleming, and Fleming–Harrington–O'Brien boundaries

k	Pocock	O'Brien–Fleming[a]	Fleming–Harrington–O'Brien[b]
1	0.0294	0.0051	0.0150
2	0.0294	0.0415	0.0418
1	0.0221	0.0006 (0.001)	0.0050
2	0.0221	0.0151	0.0061
3	0.0221	0.0471	0.0459
1	0.0182	5×10^{-5} (0.001)	0.0067
2	0.0182	0.0039	0.0083
3	0.0182	0.0184	0.0103
4	0.0182	0.0412	0.0403
1	0.0158	5×10^{-6} (0.001)	0.0038
2	0.0158	0.0013	0.0048
3	0.0158	0.0085	0.0053
4	0.0158	0.0228	0.0064
5	0.0158	0.0417	0.0432

[a]Use of 0.001 for the first test is recommended for the OF boundary with $K > 2$.

[b]Letting π_1 represent the probability of type I error occurring at the kth test, $\pi_1 = \cdots = \pi_{K-1}$ (where π_1 is the tabled entry for $k = 1$) and $\pi_K = \alpha - (K-1)\pi_1$.

The boundary proposed by Fleming, Harrington & O'Brien (FHO) occupies middle ground between the P and OF boundaries, but is closer in spirit to the OF boundary. The distinguishing characteristic of the FHO boundary is that the probability of a type I error is held constant for each interim test. Thus, the boundary is determined by specifying the number of interim tests and the probability of a type I error occurring at the final test. In the examples shown, this error rate has been chosen to be close to the overall type I error rate ($\alpha = 0.05$).

A different type of approach to group sequential testing has been proposed by Whitehead [40] and Whitehead & Stratton [41], using a triangular boundary. Conceptually, one plots an estimate of the treatment effect on the vertical axis and a measure (V) of the information contained in Z on the horizontal axis. The null or alternative hypothesis is accepted depending on whether the lower or upper of two lines are crossed. Since the lines are constructed to cross eventually for suitably large V, the test terminates with probability one.

Although the original formulation of these group sequential designs assumed normal distributions with known variances, simulation studies have shown that they provide accurate control over the overall type I error rate for other types of data. Thus, they are commonly used when variances are unknown, for binary endpoints, and survival data. Theoretical work indicates that independent increments of information accumulating between successive tests is an important underlying consideration relating to the joint distribution of the K test statistics.

The Alpha-spending Function Approach

Formally, the Pocock and O'Brien–Fleming boundaries require two assumptions which may be violated

in practice. First, one assumes that the number of interim tests which will be conducted is specified in advance. This may be problematic if a decision is made to extend the trial due to slower than anticipated accrual. Notice that this assumption is not required for the FHO boundary, since one can adjust the level of the final test to allow for an increase in the number of interim tests conducted.

A second assumption is that an equal number of subjects are recruited between each test, or in the case of survival-type studies, that the number of events occurring between each test is constant. This assumption is typically unrealistic, since meetings of the DSMB are usually determined according to calendar times. **Simulation** studies [13, 29] indicate that the effects on size and power resulting from using unequally spaced tests is negligible.

Thus far, we have assumed that modifications to the monitoring plan are not motivated by accruing data. For example, it might be tempting to schedule more frequent tests if it appears that the stopping boundary is being approached. Proschan et al. [34] considered a variety of data-driven strategies and found that the overall type I error rate can be severely inflated. Thus, these types of monitoring strategies should be avoided.

In view of the considerations discussed thus far, it appears that the Pocock and O'Brien–Fleming boundaries are quite flexible, and can be used without modification in monitoring trials in a wide range of circumstances. However, it is of interest to consider versions of stopping boundaries which are a continuous function of the percentage of the study completed. This approach was proposed by Lan & DeMets [26], and is illustrated in Figure 1. In this example, an FHO boundary is specified as in Table 1, with two interim tests performed at equally spaced time points, with $\alpha(t_1) = 0.0050$ and $\alpha(t_2) = 0.0100$.

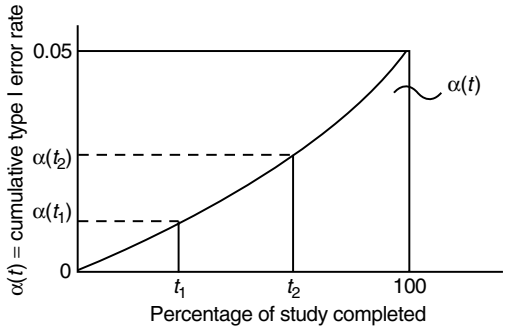

Figure 1 A spending function, $\alpha(t)$, expressing the cumulative type I error rate as a function of the percentage of the study completed

Notice that the times t_1 and t_2 need not have been specified in advance. For example, one might suppose that an FHO boundary had been selected, but the time of the first analysis (t_1) had not been planned. If t_1 corresponded to completion of 1/3 of the study as in the example, then $\alpha(t_1) = 0.0050$. However, if t_1 were some other time point, then the critical value would have been the point in the boundary corresponding to this time.

Continuous spending functions that correspond closely to the Pocock and O'Brien–Fleming boundaries have been identified by Lan & DeMets [26], and computing software has been developed in order to address the computational problems required to identify boundary points as a function of time.

Multiple Endpoints

Thus far, we have assumed that efficacy is defined by a single univariate endpoint. The situation actually encountered by a DSMB is often more complex. Of particular concern is the situation in which the primary endpoint is conceived as a single but multifaceted endpoint. For example, one might be evaluating the effects of an experimental therapy in improving the status of small nerve fibers in diabetic neuropathy. However, the status of small nerve fibers may be measured by performance on several neurologic tests, and it is the combined information from these tests which provides the definitive assessment of status (*see* **Multiplicity in Clinical Trials**).

Ideally, a clinically meaningful method for combining the various measures might be available. In this case, the resulting global score can be used, and the statistical issues associated with multiple endpoints do not arise. If such a score is not available, statistical algorithms for obtaining global scores can be used (see O'Brien [28]; and *see* **Multiple Endpoints, Multivariate Global Tests**).

A qualitatively different scenario which may be of concern to the DSMB is where multiple secondary endpoints, each of interest in its own right, are considered. The question may be asked if any of these endpoints have shown a response to therapy and, if so, which endpoints. Analyses directed at answering these questions may produce multiple statistical tests, in which case questions arise about what error rates

should be of interest (per comparison, experimentwise, or per experiment) and the best way to control them. An excellent review of these issues can be found in Hochberg & Tamhane [22]. The case against attempting to control experimentwise and per experiment error rates is discussed by Rothman [35].

It is perhaps unfortunate that discussion of testing multiple hypotheses tends to accept as a premise that one type of error rate is appropriate and the other not. An alternative approach would be to view the differing error rates as qualitatively different pieces of information, one or more of which might be helpful in any given instance.

Stochastic Curtailment

In comparing the different approaches for interim testing, it was observed that it is often desirable to stop only when the evidence for or against a treatment effect is overwhelming. Under these circumstances, the OF procedure has the desirable property that the probability of type I and type II errors are close to the error rates associated with a single sample test. Thus, one gains the opportunity to terminate early with little loss.

Stochastic curtailment follows this train of thought to its logical conclusion. One examines the data to see if the final result has already been determined with certainty [1, 2, 14, 19, 20, 27]. For example, if the data strongly suggest that treatment is efficacious, one can suppose that subsequent data will be the least favorable possible. If the **null hypothesis** would still be rejected under this assumption, then there would be no need to continue the trial, at least as far as the primary test for efficacy is concerned. Similarly, if the data indicate a lack of an effect, one could suppose that subsequent data will be the most favorable possible, and consider terminating the study if the null hypothesis would not be rejected under these circumstances.

Since it is unlikely that a trial will progress to the point that the outcome is completely determined, a DSMB is more typically concerned with the probability that a current trend might be reversed, so that statistical significance might be achieved despite an early negative trend, or vice versa. This approach is closely tied to the concept of conditional power. If the early trend is positive, one would compute the probability of achieving significance under the null hypothesis.

Conversely, if the trend is negative, one would compute the conditional power under a suitable alternative hypothesis. A natural alternative to consider is the one which was originally proposed in the study design and used for sample size calculations (*see* **Sample Size Determination for Clinical Trials**). This approach may be especially appropriate in studies in which the primary endpoint is time to an event, where one might suppose that lack of an effect initially may be due to a delayed onset of treatment effect.

In other applications, one might argue that, if the current data indicate that the originally hypothesized effect is implausible given the current data, one should consider an alternative which is both clinically meaningful and also plausible given the current data. Although some authors have proposed using the point estimate from current data as the alternative for computing conditional power, this reasoning seems somewhat circular.

The issues surrounding early stopping based on stochastic curtailment are similar in many respects to those described previously. In particular, any purely statistical algorithm will in practice be used by a DSMB as only part of the information to be considered in arriving at a decision whether or not to terminate the trial. However, a unique aspect of stochastic curtailment is that it typically does not allow one to say that statistical significance has occurred, or that it cannot occur. Rather, stochastic curtailment provides a mechanism for predicting whether or not significance will occur.

Point and Interval Estimation

When accumulating data are monitored during the course of a trial, one can expect that at times the results will appear overly encouraging. At other times, results may appear overly negative. When considering whether or not to terminate a trial early, a DSMB will need to account for this in evaluating the magnitude of treatment effects. Thus, there is a need to adjust point and interval estimates to account for interim monitoring, in much the same spirit that P values must be adjusted.

In formulating appropriate interval estimates, it is necessary to take into account that the sample space is two-dimensional, consisting of the number of tests undertaken ($k^* = 1, \ldots, K$) and the value of the test statistic at the last test (T^*). One must obtain

a univariate ordering of the sample space, and one way to do so is to order first according to k^*, then T^*. The ordering may be represented as a mapping from (k^*, T^*) to $Z(k^*, T^*)$. By inverting probability statements about $Z(k^*, T^*)$ and the parameter of interest (θ), one obtains confidence intervals for θ.

A detailed discussion of this approach appears in Tsiatis et al. [39]. A limitation is that the ordering of the sample space is somewhat arbitrary and may produce **confidence intervals** that may be counter-intuitive. For example, one would suppose that the need to adjust for overly optimistic data would be greatest for trials which terminate at the first interim test. However, using the ordering of the sample space described above, in this circumstance the adjusted confidence interval, is the same as the "naïve" single sample interval. This problem is discussed in Chang & O'Brien [11], who consider an ordering based on **maximum likelihood** criteria. However, there is no general agreement over the best ordering to use.

There is also the need to adjust point estimates of treatment effect. Methods for obtaining estimates which are unbiased in the usual sense are not generally available. An alternative approach is to use the confidence interval calculations described above, but to compute the 50% confidence interval, obtaining an estimate which may be described as "median unbiased" in the sense that the estimate will be greater than (and less than) the true value with probability 0.5.

The methods described above provide a means for adjusting point and interval estimates upon completion of a trial when the study has been terminated early. A conceptually different problem is to obtain repeated confidence intervals during the course of the study as an aid in judging the desirability of early stopping. This approach seems especially well suited to the way a DSMB works in practice, making assessments about the magnitude of effects which seem plausible, weighing this information with other considerations, and making a judgment about whether or not to continue.

Repeated confidence intervals are computed in a way to insure that all the intervals which may ultimately be computed will contain the true population parameter with a specified probability (e.g. 0.95). A method to accomplish this is to use the well known correspondence between interval estimation and hypothesis testing [23]. Since the probability statements pertain to all confidence intervals which may be computed, one could use the intersection of all intervals available at the time they are being considered. In practice, interest usually centers on the most recent interval.

One-sided Versus Two-sided Tests

Statisticians often disagree about whether tests should be one- or two-sided (*see* **Alternative Hypothesis**) [15, 16, 24, 25, 30, 31]. The issues involved are somewhat more complex in the context of monitoring a clinical trial. Careful consideration of these issues is worthwhile, because it provides important insights into the monitoring process itself. It is helpful first to consider the controversy in the more general context of **hypothesis testing**.

One argument is that a test should be two-sided if it is possible that an effect could go in either direction. Thus, if it is possible that a drug could be worse than placebo with respect to the primary endpoint, this view would imply that the test for efficacy should be two-sided. A second argument is that the test should be two-sided if the investigator would be interested in an effect in either direction. Since it is difficult to establish that a phenomenon is impossible (particularly when the effects of a drug are unknown), and since investigators are interested in any new information which might be gleaned from their study, these two arguments generally lead proponents to argue for two-sided tests.

A third point of view is that one must first determine what question the study is intended to answer. The translation of this question into corresponding null and alternative hypotheses then determines whether the test will be one-sided or two-sided. To illustrate, suppose the purpose of a study is to answer the question, "Is the experimental drug more efficacious than placebo?" The alternative hypothesis corresponding to this question is clearly one-sided, and the probability of a type I error (a claim of efficacy when in fact the drug is not more efficacious than placebo) is given by a one-sided ***P* value**. Although it might be possible that the drug effect is deleterious, and one would certainly be interested in such a finding, these are additional questions which, along with many others, would be an appropriate subject for secondary analyses. A qualitatively different situation occurs when one is comparing two competing therapies and the question motivating the study is, "Is one treatment superior to the other?"

This is a two-sided question, in which a difference in either direction will lead to an affirmative answer (and potentially a type I error).

These considerations become increasingly important in the context of monitoring a clinical trial. One might argue that, although the primary analysis at the end of the study addresses a one-sided question, the question during the monitoring phase is two-sided, since the DSMB will terminate the study if the evidence of a treatment effect is overwhelmingly positive or negative. Thus, one might argue for two-sided tests during the monitoring phase, but plan on a one-sided test at the final analysis for efficacy.

Another approach is for the DSMB to consider three distinct questions and target correspondingly distinct analyses towards each: (i) Is the drug efficacious? (ii) Is the drug safe? (iii) Is the probability of achieving statistical significance so remote that we might as well stop the trial now?

The first question is one-sided and might be addressed using a group sequential test or stochastic curtailment, using a one-sided test in either case. Similarly, the third question would be appropriately addressed using stochastic curtailment or conditional power based on a one-sided test.

The second question usually encompasses a wide range of potential hazards, and some of the dangers may only become apparent during the course of the trial. An adverse effect on the primary measure of efficacy may be only one of many such hazards, and is typically one of the least likely to materialize. In addition, the desirability of waiting until statistical significance is achieved, and the level of significance which is appropriate, may be quite different in assessing safety vs. efficacy. Thus, the implementation and interpretation of hypothesis testing in addressing safety may differ from assessments of efficacy.

Decision Theoretic Methods

Controling the overall type I error rate is a critical concern in monitoring clinical trials and in making judgments about the desirability of early termination. As indicated previously, this piece of information is combined with other factors, such as the estimated magnitude of the effect, in arriving at a decision. Although the **decision theoretic** approaches of the sort that we discuss next appear to be infrequently used by DSMBs, they might provide helpful additional information. In addition, **Bayesian methods** may be especially well suited to monitoring pilot studies in the drug development process prior to a large Phase III trial.

Maximizing the Number of Patients Receiving Better Treatment

Consider the total population of patients who will receive either the standard or experimental therapy in the future. One approach to the study design might be to attempt to maximize the number of patients who will ultimately receive the superior treatment. Specifically, let N represent the number of patients who might receive the new treatment. Among the number (n) who will be entered into the trial, let n_s and n_e represent the number who will be assigned to standard and experimental therapy, respectively. If we assume that the remaining $N - n$ will receive the drug selected at the end of the study and adopt a prior probability distribution that one treatment is superior to the other, then we can choose n_s and n_e to maximize the expected number of patients receiving superior treatment.

A fixed sample size test with equal numbers assigned to each treatment arm of the study, assuming that the study endpoint is dichotomous and immediately observable, was proposed by Canner [10]. The requirement of equal allocation between treatment arms was obviated in a method by Berry & Pearson [8]. Earlier work by Colton [12] considered procedures that are conceptually similar, using **minimax**, maximin, and Bayesian approaches focusing on the costs associated with a wrong treatment selection. Group sequential versions of these types of tests would be desirable for use by a DSMB.

A Bayesian Approach

In a sense, deciding that the evidence in favor of an experimental therapy is sufficient to justify terminating a trial and recommending the treatment for future patients is analogous to the decision a physician makes in choosing a new treatment for future patients [3, 6]. Furthermore, one might argue that, at least implicitly, physicians make these types of decisions based on a Bayesian probability; that "medical researchers ... act like Bayesians" [7].

This line of reasoning suggests a Bayesian decision rule rather than a classical group sequential

approach. A simple example of how such a rule might proceed is provided by Berger & Berry [6], who consider a randomized trial comparing experimental and standard therapies. They assume that patients are randomized in pairs and that a dichotomous response (success or failure) is immediately observable.

Let p represent the prior probability that there is no difference between drugs (H_0), and let θ represent the probability that E will be superior to S in a given pair, where the investigator must specify the constant p and the distribution for θ. In this case, we assume $p = 0.5$ and that θ follows a uniform distribution over the interval (0, 1).

Hypothetical data for the first 18 pairs are shown in Table 2, together with the posterior probability that E is superior to S. Based on these computations, it would appear that the evidence in favor of E was convincing well before the eighteenth pair.

It is of interest to consider how the monitoring of this trial might have proceeded using the classic group sequential approach, evaluating the data after every six pairs, supposing that a maximum of 18 pairs had been planned. The nominal P value for a one-sided test after 12 pairs is 0.019, slightly larger than the level required for the OF, but less than the level required by the P boundary.

Table 2 Results in 18 pairs of patients

Pair	Superior treatment	Cumulative preference for E	Posterior probability E better than S
1	E	1	0.750
2	S	0	0.500
3	E	1	0.687
4	E	2	0.812
5	E	3	0.891
6	S	2	0.773
7	E	3	0.855
8	E	4	0.910
9	E	5	0.945
10	E	6	0.967
11	E	7	0.981
12	E	8	0.989
13	E	9	0.994
14	S	8	0.982
15	E	9	0.989
16	E	10	0.994
17	E	11	0.996
18	E	12	0.998

Source: Berger & Berry [6].

This example provides some insights about the strengths and weaknesses of a Bayesian approach within the context of monitoring clinical trials. A subjective assessment may be especially useful for a pilot study in which the goal is to determine whether or not to proceed to the next step in drug development. On the other hand, for trials which are intended to provide a definitive answer about efficacy, the necessity to specify prior distributions may pose a difficulty for the Bayesian approach. Disagreement about priors may occur both within the DSMB as well as in the medical and scientific community.

Some ways to integrate Bayesian methods into a more traditional framework are: to compute the type I error rate for Bayesian stopping bounds; to adopt skeptical priors resulting in conservative stopping bounds; and to treat Bayesian analysis as a source of additional information complementing the information obtained from frequentist analyses. Excellent discussion of applied aspects of incorporating Bayesian methods in the monitoring of clinical trials is provided in Spiegelhalter et al. [38] and Freedman & Spiegelhalter [18].

References

[1] Alling, D.W. (1963). Early decision in the Wilcoxon two sample test, *Journal of the American Statistical Association* **58**, 713–720.

[2] Alling, D.W. (1966). Closed sequential tests for binomial probabilities, *Biometrika* **53**, 73–84.

[3] Anscombe, F.J. (1963). Sequential medical trials, *Journal of the American Statistical Association* **58**, 365–383.

[4] Armitage, P. (1960). *Sequential Medical Trials*. Thomas, Springfield.

[5] Armitage, P., McPherson, C.K. & Rowe, B.C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

[6] Berger, J.O. & Berry, D.A. (1985). Analyzing data: the great conditioning debate, unpublished manuscript.

[7] Berry, D.A. (1985). Interim analyses in clinical trials: classical vs. Bayesian approaches, *Statistics in Medicine* **4**, 521–526.

[8] Berry, D.A. & Pearson, L.M. (1985). Optimal designs for clinical trials with dichotomous responses, *Statistics in Medicine* **4**, 497–508.

[9] Bross, I. (1952). Sequential medical plans, *Biometrics* **8**, 188–205.

[10] Canner, P.L. (1970). Selecting one of two treatments when the responses are dichotomous, *Journal of the American Statistical Association* **65**, 293–306.

[11] Chang, M.N. & O'Brien, P.C. (1986). Confidence intervals following group sequential tests, *Controlled Clinical Trials* **7**, 18–26.
[12] Colton, T. (1963). A model for selecting one of two medical treatments, *Journal of the American Statistical Association* **58**, 388–400.
[13] DeMets, D.L. & Gail, M.H. (1985). Use of log rank tests and group sequential methods at fixed calendar times, *Biometrics* **41**, 1039–1044.
[14] DeMets, D.L. & Halperin, M. (1982). Early stopping in the two-sample problem for bounded random variables, *Controlled Clinical Trials* **3**, 1–11.
[15] Dubey, S.D. (1991). Some thoughts on the one-sided and two-sided tests, *Journal of Biopharmaceutical Statistics* **1**, 139–150.
[16] Fisher, L.D. (1991). The use of one-sided tests in drug trials: an FDA advisory committee member's perspective, *Journal of Biopharmaceutical Statistics* **1**, 151–156.
[17] Fleming, T.R., Harrington, D.P. & O'Brien, P.C. (1984). Designs for group sequential tests, *Controlled Clinical Trials* **5**, 348–361.
[18] Freedman, L.S. & Spiegelhalter, D.J. (1989). *Controlled Clinical Trials* **10**, 357–367.
[19] Halperin, M. & Ware, J.H. (1974). Early decision in a censored Wilcoxon two-sample test for accumulating survival data, *Journal of the American Statistical Association* **69**, 414–422.
[20] Halperin, M., Lan, K.K.G., Ware, J.H., Johnson, N.J. & DeMets, D.L. (1982). An aid to data monitoring in long-term clinical trials, *Controlled Clinical Trials* **3**, 311–323.
[21] Haybittle, J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment, *British Journal of Radiology* **44**, 793–797.
[22] Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.
[23] Jennison, C. & Turnbull, B.W. (1984). Repeated confidence intervals for group sequential clinical trials, *Controlled Clinical Trials* **5**, 33–45.
[24] Koch, G.G. (1991). One-sided and two-sided tests and *p* values, *Journal of Biopharmaceutical Statistics* **1**, 161–170.
[25] Koch, G. & Gillings, D. (1988). One-sided versus two-sided tests, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 218–222.
[26] Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.
[27] Lan, K.K.G., Simon, R. & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials, *Communication in Statistics, Sequential Analysis* **1**, 207–219.
[28] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1087.
[29] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.
[30] Overall, J.E. (1991). A comment concerning one-sided tests of significance in new drug applications, *Journal of Biopharmaceutical Statistics* **1**, 157–160.
[31] Peace, K.E. (1991). One-sided or two-sided *p* values: which most appropriately address the question of drug efficacy?, *Journal of Biopharmaceutical Statistics* **1**, 133–138.
[32] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I, Introduction and design, *British Journal of Cancer* **34**, 585–612.
[33] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.
[34] Proschan, M.A., Follmann, D.A. & Waclawiw, M.A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring, *Biometrics* **48**, 1131–1143.
[35] Rothman (1986). *Modern Epidemiology*, 1st Ed. Little, Brown & Company, Boston, pp. 147–150.
[36] Samuel-Cahn, E. (1974). Repeated significance test II, for hypotheses about the normal distribution, *Communications in Statistics* **3**, 711–733.
[37] Samuel-Cahn, E. (1974). Two kinds of repeated significance tests, and their application for the uniform distribution, *Communications in Statistics* **3**, 419–431.
[38] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society, Series A* **157**, 357–416.
[39] Tsiatis, A.A., Rosnar, G.L. & Mehta, C.R. (1984). Exact confidence intervals following a group sequential test, *Biometrics* **40**, 797–803.
[40] Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials*. Ellis Horwood, Chichester.
[41] Whitehead, J. & Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions, *Biometrics* **39**, 227–236.

Further Reading

Ellenberg, S.S., Fleming T.R. & DeMets D.L. (2002). *Data Monitoring Committees in Clinical Trials*. Wiley.

PETER C. O'BRIEN

Data Archives

Data archives are accessible and indexed compendia of data which can be accessed and utilized by researchers intending to perform secondary data analysis. Such archives have been in existence in the United States since the early 1960s and have spread internationally since then. An example is the Economic and Social Research Council (ESRC) data archive at the University of Essex, UK, established in 1967. Quantitative data from various governmental, administrative, and research sources are held in computer-readable forms. Qualitative data archives, including the ESRC QUALIDATA archive, are a more recent innovation.

Data archives preserve data against disposal or deterioration, provide indexing services, and can provide data in formats useful to secondary data analysts. There are several reasons why increasing accessibility of data, through the establishment of data archives, is beneficial to the research community [1]. Data collection is expensive, and the use of extant data to answer research questions which were not originally envisaged by the collectors of the data is an efficient use of resources. Study participants are protected from being overresearched by the multiple use of data. Archived data provide a rapid and inexpensive way of replicating the findings from other studies, and the interrogation of existing data at the time of the design of future studies is useful for the performance of **sample size determination**, evaluation of data collection instruments (*see* **Questionnaire Design**), and the exact formulation of research questions. Archived data can be linked to other data (*see* **Record Linkage**) if different sets of records on the same people exist, creating data sets which can be used to explore issues which cannot be examined in existing unlinked studies. The rapidly expanding field of **meta-analysis** also benefits from the ready availability of primary data. Statistical techniques can be applied to data which were collected when such methods of analysis had not been developed. Finally, the requirement to provide original data is one protection against the production of findings based on a particular statistical approach to the data – findings produced by what is commonly called "data torture" and the actual fraudulent invention of results.

Primary researchers may understandably view the provision of data to archives as threatening [1]. Guaranteed rights to initial publication, transfer of the costs of data preparation for archival storage to funding bodies or secondary analysts, and protection against the commercial exploitation of data by a recipient are all reasonable requests by primary data collectors. Criteria for the evaluation of secondary (including archived) data sources for use in epidemiologic research have been developed [2], and a formal appraisal of the value of secondary data analysis in relation to that of primary research would be valuable.

References

[1] Davey Smith, G. (1994). Increasing the accessibility of data (editorial), *British Medical Journal* **308**, 1519–1520.

[2] Sørensen, H.T., Sabroe, S. & Olsen, J. (1996). A framework for the evaluation of secondary data sources for epidemiological research, *International Journal of Epidemiology* **25**, 435–442.

GEORGE DAVEY SMITH

Data Management and Coordination

The term "data management" in **clinical trials** has become a very general term that covers the procedures both for the collection of data at clinical sites, and for the quality control of those data after they have been submitted to a central statistical or coordinating center (*see* **Clinical Trials Audit and Quality Control**). In both locations procedures should be established for managing the trial data, and steps taken to ensure that the quality of data is high throughout a trial. These steps and procedures are described in this article.

An individual who is responsible for the collection and quality control of data is known by various titles, including "data manager", "clinical research associate", "data coordinator" and "research assistant". For the sake of clarity, in this article the term "clinical research associate" (CRA) is used for an individual responsible for the abstraction of data and completion of forms at the participating institution, and the term "data manager" is used for an individual responsible for the quality control and computerization of the data at the statistical center. Usually, there is one clinician who has overall responsibility for the design and monitoring of the trial and, in this article, this person is referred to as the "study chair".

The importance of high-quality data cannot be overemphasized. The "end-product" of a clinical trial is the publication of its results in the scientific literature. To reach this point, trials have to be designed carefully, required data defined, forms developed, patients enrolled, data collected, data analyzed, and a manuscript prepared. The goal of the data management team is to collect complete and accurate data so that the results are correct. By "correct" results is meant true observations. Whether these are statistically or clinically significant is not the main concern of the data management team.

Trial Participants

Clinical trials can be single-institution or **multicenter trials**. A single-institution trial is conducted in one location with trial design, patient entry, data collection and analysis all being done at that institution. Multicenter trials are collaborations between investigators at multiple institutions. Study design is done as a team, patients are entered from all participating institutions and the data are usually sent to a central statistical center or coordinating center for quality control, computerization and analysis. Most large Phase III trials are done as multicenter trials, as few single institutions are able to accrue enough patients by themselves. In this article, the multicenter model is used for examples, but most of the discussion applies equally to small single-institution trials.

As well as the participating institutions and the statistical center, special reference centers (e.g. to review pathology slides, read x-rays and scans, or to do specialized laboratory testing not routinely available at the participating sites) may be used for specific trials. These reference centers usually generate data that become part of the database for the clinical trial, and therefore data management procedures need to be established to handle the collection and transfer of these data.

Another major participant in a trial is the sponsor. Most large trials have sponsors who provide funding or resources (such as drugs). The sponsor could be a government, a private agency that supports scientific research, or a company that manufacturers one of the treatment components. The sponsor may have special requirements for data management procedures, and it is important to discuss these during the design phase of the trial to be sure that they will be met.

Protocol and Forms Design

Data management input is important during all phases of a clinical trial, including the design phase. Data managers can contribute to the clarity, completeness, and consistency of a **clinical trial protocol** by providing feedback. In particular, they should review the eligibility section (*see* **Eligibility and Exclusion Criteria**), the section describing how patients will be entered, the treatment administration section, and the section describing the schedule for submitting required data forms. Data management review of each draft of the protocol can greatly improve the quality of the final protocol and make it easier to use. Reviewers should include both CRAs at sites that will enter patients and the data managers at the statistical center.

In parallel with the development of the protocol, there should be discussion about the data items

needed to meet the study objectives and monitor the progress of the trial (*see* **Data and Safety Monitoring**). The identification of the data items to be collected, and the subsequent design of the data collection instruments and the computer database (*see* **Database Systems**), are critical. These three activities are interrelated and are best performed in the order listed. The data collection instruments, whether paper forms or electronic screens, should be available prior to entry of the first patient. A trial should not be activated without data collection instruments.

Defining The Data Items

There are different types of data that may need to be collected, and it is important during the planning phase of a study to think through all the requirements for the trial. For example, besides the research data, it may be necessary to collect data to aid the administration of the trial and to document compliance with professional regulations and good clinical practice.

Identification Data. When forms are submitted they must be linked to the appropriate patient and also to the correct trial. Therefore, a form must have space for recording sufficient information for correct identification of the patient, the trial, and the local institution.

Research Data. The research data represent the information that is ultimately analyzed to address the study objectives. The required data should be identified during the protocol development phase, with input from all key members of the trial team, including the study chair statistician and data coordinator.

It is always tempting to collect data items "just in case" they turn out to be interesting when the data are analyzed. However, collecting large amounts of data on each individual can be detrimental to the study because, as the volume of data increases, the quality of the data can decrease. It is, therefore, important to limit data collection to those items that are truly necessary to answer the trial objectives and manage the trial. In assessing the data requirements, the team should distinguish between data that are needed for the clinical care of the patient and data that are needed to answer the research objectives. Data collection for the trial should be limited to those items related to the research objectives. In most trials, only a small fraction of the clinically relevant information is entered into the trial database.

Omissions at this stage will be very difficult to rectify once the trial has begun to accrue patients, as it is hard to collect data retrospectively. To help to ensure that all necessary data are collected, it is useful for the statistician and study chair to draft an analysis plan detailing the information to be included in the final report. Other members of the trial team can review this outline and provide input. Once this is done, it is easier to identify the required data items. At a minimum, the required data usually include key dates of events (such as date of entry to the trial), information about the treatment assigned and received, side effects of treatment, and the study endpoints (*see* **Outcome Measures in Clinical Trials**).

Administrative Data. It is usually necessary to collect administrative data to help with the management of the trial. An example is the recording of dates of dispatch of materials sent by the institutions to reference centers, and an inventory of the materials sent. The amount of administrative data depends on the size and complexity of the trial. In a small single-institution trial, much less information is needed than in a large multicenter trial, where data and materials are being shipped to various locations.

Regulatory Data. For some trials it may be necessary to collect documentation that shows compliance with local, national or international regulations (*see* **Drug Approval and Regulation**). This could include documentation of protocol approval (and periodic re-approval) by an ethics committee or institution review board prior to patient entry, consent of the patient prior to entry, and the professional qualifications of the personnel at a participating site. Normally, if the trial includes investigational treatments, then the statistical center needs to collect copies of these documents. For other trials, it may be sufficient to maintain a file of these at each participating site.

Design of Case Report Forms

Once the data items have been defined, case report forms (CRFs) should be developed. A CRF is a printed or electronic document that is designed to collect the required research, administrative, and regulatory data for a clinical trial. The measurement and

recording of the trial data are perhaps the most critical steps in the overall data management process, and it is, therefore, important that the CRFs be designed for clarity and ease of use. Forms should always be available before a trial is activated. Activating a trial without the CRFs available is likely to generate incomplete and inconsistent data; the urgency to activate a trial should, therefore, always be balanced by the need to have the forms in place. It is recommended that forms be piloted, to identify problems that can be corrected prior to starting the trial.

It is always useful to look at the data collected for other similar trials before designing new forms, and to take advantage of this prior experience. If there are existing forms that can be used for the new trial, it can eliminate much work. A book of more than 600 forms used in previous clinical trials is available [3] and may provide useful examples. If existing forms are used, it is important to check that these do, in fact, collect all the data needed for the new trial, and, conversely, do not collect data that are superfluous.

When designing forms for a trial, thought should be given to the following aspects.

Content and Organization of CRFs. The ultimate objective of the CRF is to collect the data needed to answer the trial's objectives. Once identified, it is necessary to decide how to organize the data items on the forms. It is not always best to minimize the number of CRFs by trying to fit as much as possible onto one form. It may be better to have more forms, each with a smaller amount of data. When designing CRFs, one should ask:

1. When will data be available?
2. Who will complete the forms?
3. Where will the data be collected?

As a first step, the *timing* of the collection of the different items should be established. For example, one should identify all of the data items that will be collected at the time that the patient is entered on the study. These normally include data on the patient's past medical history, data confirming the patient's eligibility, and results of baseline tests required by the protocol. Other relevant time points could be the different stages of the protocol treatment period, the end of treatment, and scheduled follow-up examinations after the treatment period. All of these are logical divisions and can help in deciding which data items belong on which forms.

As well as the timing of the data collection, it is also useful to identify *where* the data will be collected and *by whom*. These are also logical divisions that can help to decide which data should be collected on which form. For example, there may be baseline data that are gathered from the medical record by a CRA and other data that are completed by a medical specialist, such as a surgeon. Even though all the data are collected at the same time point, it is more efficient to have two different forms – one for the CRA and one for the surgeon. This allows each person to complete their part of the data collection in parallel, rather than one having to wait for the other to complete their part before passing the form on. Likewise, if some of the data are available in the cardiology department and some in the physical therapy department, two separate forms may work better.

Format of Questions and Coding Conventions. The goal of the CRFs is to collect complete and unambiguous data and to ensure standardization and consistency of data across participating clinics. The format of the CRFs should be designed with three functions in mind: (i) the completion of the form; (ii) the entry of the data onto computer; and (iii) the retrieval of data for analysis. The person completing the form should be able to answer the questions and record the answers in an efficient and effective way, minimizing the possibility of misinterpretation or transcription errors; the person entering the data onto computer should be able to transcribe values from the form to the keyboard with minimum effort in following the flow of responses and entering the data values; the person analyzing the data needs to be able to interface the data and the statistical **software** with minimal data conversion. Even if the data are not computerized, but are tabulated manually, it is important to design the forms with analysis in mind.

There are several ways to format questions on CRFs and there are conflicting ideas of the most effective format for collecting complete, accurate data. Because the goals of different trials and environments of data collection vary, it is recommended that the user develops forms in the format that best suits the research being done and the resources available. If several forms are developed for a trial, the most important criterion is to use a consistent format across forms, so that the users can become familiar with the format used. Page layouts should be similar

across forms, and the headers of the pages should be designed in the same way, collecting the same identification information. Coding conventions should also be consistent for all data items; for example, 1 = no, 2 = yes for all instances where "no" and "yes" are possible answers.

When designing the layout of the forms, the questions should be concise and unambiguous. The text should be contiguous to the box or space where the answer is to be written, and there should be adequate space for responding. Instructions should be clear and located next to the field to which they apply. Decisions need to be made about the codes to be used and the inclusion of values for "Unknown", "Other", "Not-Applicable" or "Not-Done" responses. It is often useful to leave clear space for the participants to enter comments using free text, as important information may be conveyed in this way.

Good forms design is a complex subject and cannot be discussed in detail here. Hosking et al. [1] summarize many of the issues involved in form design and their article is recommended for further reading (*see* **Questionnaire Design**).

Role of the CRA

The job description for a CRA varies from one institution to another, and can include many responsibilities, depending on the qualifications and training of the CRA. If the CRA is a nurse, the responsibilities can include patient care as well as those oriented towards data collection. The following responsibilities are usually part of the job description for a CRA.

Tracking Data Submission Requirements

At the participating institutions, particularly those participating in several trials simultaneously, it is important to develop systems to ensure submission of complete and accurate data according to the schedule defined in the protocol. Scheduling systems can be computer- or paper-based, and the system selected will depend on the local resources and skills, unless scheduling software is provided from a central office. A computer scheduling system usually requires the building of a database that contains information on each patient entered, information on the forms required for a trial, and the time frame for submission of the forms. Programs are needed to link these data and to generate calendars based on the patient's date of entry to the trial and any relevant events occurring during the trial. Forms can be required at fixed time frames (e.g. after each clinic visit), or after particular events (e.g. failure to respond, or toxicity). The system therefore requires the ongoing entry of relevant data to remind CRAs of forms submission requirements. Such a system normally requires programmer support in development and maintenance.

If computer support is not available at a location, a paper-based system can be developed, using tools such as a wall calendar or index cards. Entries are made for each patient by marking the date of entry on to the study. At each visit, or time of patient contact, the CRA completes and submits any required data. The CRA then calculates when the next form is due and, with the calendar system, makes an entry under that date indicating the form needed for that patient. The calendar needs to be checked regularly to see what forms are due each week or month. With the index card system, the CRA sets up a file with a section for every month or, if necessary, every week. There is a card for every patient, perhaps listing all the required forms. After a form is submitted for a patient, the date of submission is entered on the card to record that the data were sent; then the date that the next form is due is calculated, and the card is filed in the card system according to that date. The CRA pulls all cards in the relevant section each week or month, and makes sure that the required forms are submitted.

While a computer-based system allows more flexibility and eliminates the need for the CRA to calculate future due dates, a paper-based system can be equally effective, unless there are large numbers of patients to track or the trial is highly complex.

Data Recording

Another important responsibility of the CRA is to ensure the collection of complete and consistent data. This requires ensuring that all trial data are recorded in the patient's medical record and that all study parameters are followed. In a busy clinical environment, there may not always be sufficient time for a clinician or nurse to review a protocol and ensure that all tests are done and required data collected. Anything that a CRA can do to help with this process will improve the likelihood of complete data being recorded. The CRA can review the list of patients

attending clinic on a particular day, identify the protocol patients ahead of time, prepare a list of tests to be done at that visit, and place it in a prominent place in the patient's chart. This will help the clinical staff in ensuring protocol compliance.

If there are subjective data to be collected from the patient, or aspects of a physical examination (such as tumor measurements) that need to be recorded, the CRA can prepare a special internal data collection instrument for use by the clinical staff to ensure that they ask all the questions necessary. These data collection instruments can become part of the patient's trial record and used to complete the CRFs submitted to the statistical center, and are valuable for audit purposes.

Preparation for Audits

Many trials require complete or partial source verification by an outside organization, particularly if investigational treatments are involved, or if the data from the trial are likely to be used as part of a regulatory submission for approval of commercial use of the treatment. This monitoring is usually done by the sponsor or their designee. The primary goals of such a monitoring system are to verify that all regulatory requirements are being met, and that the data submitted to the statistical center are complete and can be substantiated by review of the original medical records of the patient. Maintaining complete and well-organized source documents can greatly facilitate this process. When an audit is scheduled, the CRA can assist the auditors by organizing the medical records and marking the relevant parts of the record. More details of the on-site monitoring process can be found in the article on **Clinical Trials Audit and Quality Control**.

Role of the Data Manager

The most important function of the data manager at the statistical center is quality control of the submitted data. Data need to be checked for completeness, clarity and consistency over time. If missing, unclear or inconsistent data are detected, the statistical center needs to query the responsible institution to resolve the issue and obtain the correct data values.

Quality control can be done by computer, visual review of the submitted data, or a combination of both of these. When planning a study, a quality control plan should be developed, defining the checks to be made. Once the trial is active and data are received and reviewed, the plan will probably need to be modified on an ongoing basis. It is important to follow the quality control plan and do consistent quality control checks on all data entered. Documentation should be maintained defining all the checks, and keeping a record of the type and timing of changes to the quality control plan.

Eligibility Check

There are several checks that should be done as part of the quality control process. A system should be developed to ensure that patients are registered in the trial prior to starting on protocol treatment. A check should be done to ensure that the patient is eligible for the trial. The institution can be solely responsible for ensuring that all eligibility criteria are met, or the statistical center can check eligibility by asking relevant questions at the time of registration (*see* **Eligibility and Exclusion Criteria**). It may also be important to check that all regulatory requirements have been met prior to entry, e.g. that informed consent has been given and that the protocol has ethics committee approval (*see* **Ethics of Randomized Trials**). Often, an eligibility checklist is prepared as part of the forms for a trial and is used by both the institution and the statistical center to confirm patient eligibility. The eligibility check should be repeated on review of the submitted data to ensure that the data given at the time of registration were accurate.

Logging Receipt of Forms

When data forms are received at the statistical center, they should first be checked to ensure that patient and trial identifiers are correct. For example, if patient initials are on the form, they can be checked with the initials given at the time of registration of the patient with that identifier, or the name of the institution can be checked. If there are any discrepancies, they will need to be checked with the institution that submitted the forms. The forms should then be logged in so that a record is kept of patient and trial identifiers, type of form and date of receipt. Depending on the size of the trial, this can be done manually or by computer. Bar code technology can be used to scan in this information. Logging receipt allows the statistical

center to know which forms have been received and which are overdue. It also allows tracking of timeliness of data submission.

Logical Checks

Submitted data need to be checked to ensure that the correct forms have been used, that patient identifiers are on each form, that the data are consistent over time and that the forms are complete. Logical checks can also be defined; for example, checks that dates are in logical sequence, or for consistency between associated data items. These types of checks can be done manually or by computer, and the statistical center should have a mechanism for sending queries back to the institutions when discrepancies are found.

Assessment of Compliance and Endpoints

There also need to be defined procedures for monitoring **compliance** to the protocol and evaluating the study endpoints for each patient, using the criteria specified in the protocol (*see* **Outcome Measures in Clinical Trials**). The design of the data collection forms and the level of data computerization determine whether these checks can be done by computer or need review by a data manager (or a combination of the two.) The checks usually require comparison of data over time or over different data forms. For example, in cancer trials, if there is a summary form that collects data on the best overall response to treatment and another form that collects tumor measurements over time, the measurements can be reviewed to confirm that the response assessment is correct. If the tumor measurements are not entered onto computer, then the check is done manually.

Clinical Review of Data

In many trials, data are also reviewed by the study chair or another designated clinician. This fulfills two primary purposes. First, it ensures that complex medical data are reviewed and assessed by someone with the appropriate training to detect any clinical nuances in the data. It also provides a quality control check on the assessments of the data manager at the statistical center. If clinical review is part of the trial procedures, then a system must be developed. This could involve copying all data as they are received and sending copies to the reviewer, or having the reviewer visit the statistical center on a regular basis to review the data on site. Normally, the results of the clinical review are compared with those of the data manager, and disagreements are discussed with the statistician in an attempt to reach consensus.

Coding Conventions

Conventions need to be developed and documented for dealing with problems such as **missing data** values, or tests not being done, with results consequently unavailable, or for flagging cases that still have unresolved questions. The statistician is normally closely involved in developing these rules and for ensuring that they are consistently applied.

Data Requests

It is important to ensure that data are collected and submitted in a timely way, and lists of overdue data should be sent to the institutions at frequent intervals. Often, bad news arrives early and the forms first received at the statistical center document study failures. Unless there is a balance to ensure that data are received on all cases according to the same schedule, there is a risk of overreacting to the bad news and drawing erroneous conclusions about the efficacy of the treatments under study.

Data Queries

A query to the institution can be generated at any time during the quality control process, when clarification is needed or missing data has been detected. Queries should be made in writing so that there is a record of both the query and the response. It is also advisable to keep track of dates that queries were sent so that the response can be tracked. When generating a query, it is important to provide sufficient information to the institution so that they understand the question. The patient and trial identifiers should be on the query, along with a clear statement of the question being asked. If the queries are generated by computer, it is important to avoid being cryptic in the text of the error message. Decisions need to be made about whether revised CRFs should be submitted, or whether the query can be answered with a note from the institution. If the latter is acceptable, the query

letter should have space for the institution to write a response and, when received at the statistical center, the returned query should be part of the patient's trial record.

Data Management Support for Analyses

When an interim or final analysis is being done, the data manager plays an important role. Normally, a cut-off date is selected, and all data that have been submitted by that date are quality controlled and entered into the trial database. It is important to try to recover as many responses to queries as possible. When the statisticians prepare for the analysis, they usually run programs to check the database, and inevitably detect further inconsistencies. It is the data manager's responsibility to resolve these discrepancies and make corrections to the database. For interim analyses, (*see* **Data and Safety Monitoring**) this is an ongoing process, and it is not necessary to have all issues resolved before the analysis is done. However, when the final analysis is being done in preparation for a manuscript, it is important to set enough lead time so that a final effort can be made to retrieve all missing data from the participants, to resolve all queries, to ensure that all follow-up is as up to date as possible, that clinical review (if being done) is complete, and that all database inconsistencies have been resolved. The data manager works closely with the statistician to ensure that this is done.

Computing Support

For all clinical trials except very small **Phase I** and **Phase II** studies, it is unlikely that data will be managed without the use of a computer for either quality control, data storage or statistical analysis. Decisions therefore need to be made about the computing system used.

Data Storage

Clinical trials data are usually managed by a database management system (DBMS) (*see* **Database Systems**). Choices are available for most types of computers. The most common type of DBMS in use for clinical trials is the relational database where data are stored in multiple tables that can be linked by the use of key fields [2]. Any database system is unlikely to provide all the functionality required for a clinical trials application, and it is usually necessary to develop application programs that meet the needs of the trial. Before embarking on the purchase of hardware or software, it is important that a detailed analysis of the requirements be done so that the system chosen is one that closely meets the needs of the trial. Mistakes can be costly, both in terms of purchases and in the time and effort expended to make the system work.

Whatever database management system is used, it is important that it provides an interface to the statistical software to be used. Many DBMS packages provide interfaces to the most commonly used statistical software packages, but if no interface is available, then one may need to be developed specifically. If the interface is developed locally, then it is important that it be rigorously tested to ensure that it produces complete and accurate data as input to the statistical software.

Data Entry

If data are collected on paper forms, they need to be entered onto the computer at the statistical center. Data entry is usually done by setting up computer screens that have a layout similar to the data forms being used. When data are keyed, range- and field-type checks are usually performed, and error messages appear if a value is out of range or of the wrong data type (e.g. alphabetic instead of numeric). Professional data entry operators are usually not trained to resolve these kinds of discrepancies, and are therefore required to mark the field that is causing the error and return the form to the data manager for resolution. The record is not entered into the database until the problem has been resolved. If a data manager is entering the data, he or she may be able to resolve the problem at the time of entry. Depending on the resources available, data may be entered and verified by double key entry. Data are keyed by one operator and then rekeyed by a second operator (or sometimes by the same operator after some specified time has elapsed). The two resulting files are compared and errors resolved. This type of system reduces the number of data entry errors but is considerably more expensive than single key entry.

Decisions need to be made about the software to be used for data entry. If the database for the trial

is being maintained on a mainframe computer or workstation, then there is usually an option to enter the data directly into the database, or to enter the data off-line (using a personal computer (PC)) and then transfer the data to the main computer for batch update. The decision depends on the programming resources available, as well as the types of computers. If a batch process is used, detailed checks can be done on the data prior to their entry into the trial database. If the data are entered directly into the database and checks are done while the data are being keyed, the data entry process will be slowed down as the operator deals with any error messages. With both systems, additional checks can be run against the entire database once the new data are in the database. An advantage of off-line data entry is that it can continue even if the database computer is unavailable.

Software Tools

Many software tools can be developed to assist with the management of data for a clinical trial. Examples include programs to generate reports on the progress of a study (e.g. accrual, eligibility rates, adverse events, etc.), programs to request missing or overdue data, programs to monitor the performance of participating sites, and programs for the use of the data manager to help with the data processing. The latter could include programs that allow easy inspection of data in the database, or programs that generate status reports of the data for a study. If scheduling software is not available at all sites, the statistical center could also generate patient calendars by computer to assist the institutions with scheduling visits and forms submission.

Distributed Computing

The model described so far is one where paper forms are completed at the institutions and sent to the statistical center where the data are entered into a computer. It is also possible to distribute all or part of the computing system to the participating sites. Again, the extent of distribution depends on the resources available, both in terms of computing capabilities and personnel support. The most common component of a system to be distributed to the sites is data entry, and this has been used successfully in several clinical trials. However, it may not be the best choice for all trials, and careful consideration should be given to the advantages and disadvantages within the context of a specific trial. The main aspects to consider, prior to deciding, are the volume and frequency of data to be collected, the number and stability of participating sites, the likelihood of frequent changes to the protocol or forms, the resources available at the sites and the statistical center, and the need for ready accessibility to data at the sites and the statistical center.

Volume

It is important to have accurate estimates of the volume of data to be collected at each site. Once the volume exceeds the capacity of the available hardware and personnel, the data will get backlogged and the system will be ineffective. If the volume and frequency of data exceed that which can be handled by a single person, then consideration will need to be given to either multiple PCs at the site or a computer environment with a multiuser operating system. There also need to be adequate personnel to deal with the volume of data entry.

Number of Sites

If the number of sites entering patients on the trial is large and subject to fluctuation over the life of the study, it is difficult to build and maintain a distributed system, and the costs may be prohibitively high. If hardware is purchased as part of the trial budget, as sites drop out and new ones join, the hardware needs to be transferred from one site to the other. Staff training to use the system also requires a substantial investment of time and effort. Distributed systems are more advisable in an environment where there are a relatively small number of sites which are likely to be participating throughout the life of a trial. In some environments (e.g. the cancer clinical trial cooperative groups), the large number of trials active at any point in time also makes distributed systems less viable.

Changes to Protocols and CRFs

If the protocol or forms are likely to change frequently during a study, the maintenance burden for a distributed system can be high, as it is essential

that any new software be installed simultaneously at all sites on the day of implementing the change. This requires rapid development and testing of the changes at the statistical center, and the ability to ensure that the new version of the software is implemented at each site. It is important that the changes should not be introduced until the software is ready, otherwise the data collection will be suspended pending introduction of the new software, and this could cause problems. The distributed system is more suited to an environment where there are unlikely to be major changes while the study is in progress.

Resources

Maintaining a distributed system is likely to require more resources than a paper-based system. Before deciding to implement a distributed system, it is important to ensure that there are sufficient resources for purchase and maintenance of hardware and any commercial software being used. Software development staff must be readily available at the statistical center and it is essential that there is ongoing user support to answer questions and deal with problems at the institutional sites. Without this kind of support, the sites are likely to get frustrated with the system. User manuals are also important, and need to be updated quickly as changes are made.

Accessibility to Data

While data entry is the most common function to be distributed, it is possible to pass responsibility for other aspects of a trial to the local site. The site can be responsible for maintaining their own trial database and for quality control of all data entered. In this model, it is important to set rules for the use of data in the local database so that an investigator does not release data prematurely. Because the data will be transferred eventually to the central trial database so that a full analysis can be done, it is important to decide which database is the official database, as changes are likely to be made to the data in both locations. It may also be necessary to define rules for concurrency of the two databases.

If it is important for the local sites to have immediate access to all their data, a distributed system may be the better choice. Examples of this are the use of a dynamic randomization balancing scheme that depends on data available only at the site, or where treatment decisions are based on recent data. It is possible to implement this kind of system in the central model, but it requires rapid transmission of data to the central database and the ability to access the database from the local sites. If there is a need to have data available rapidly on the central system, then distributed data entry may be the optimal solution.

Other Solutions

Optical scanning technology is not yet accurate enough to be a viable option for most clinical trials, although for small trials where all data can be recorded by marking special areas on the forms, it can be used. Error rates for character scanning are still higher than acceptable for a trial.

There is an increasing use of facsimile (fax) machines for submitting clinical trials data and several commercial software packages are now available for implementing this type of system. This provides a hybrid system with data being recorded on paper CRFs at the institution and then sent by fax to the statistical center. Software at the statistical center allows the data manager to review the CRFs and do quality control on-line. Images of the forms can be stored and, if necessary, retrieved and printed.

While a distributed system may seem appealing and, in this time of expanding network communications, the most logical model to use, it is essential that a detailed resource requirement analysis be done prior to deciding on such a system. Distributed systems are most effective in an environment where there are a small number of sites, a limited number of trials, and there is the likelihood that there will be few changes to protocol design or forms during the course of the study. Inadequate resources for personnel or equipment are likely to cause serious problems.

Summary

Good data management is essential to any clinical trial, no matter how large or small. It is important that extensive planning be done and all necessary systems be in place before a trial is activated. Both the CRAs at the participating sites and the data managers at the statistical center play an important role in all stages of the trial, including protocol development, forms design, system testing and implementation,

and in the collection of complete, consistent and objective data.

References

[1] Hosking, J., Newhouse, M., Bagniewska, A. & Hawkins, B. (1995). Data collection and transcription, *Controlled Clinical Trials* **16**, 66S–103S.

[2] McFadden, E., LoPresti, F., Bailey, L., Clarke, E. & Wilkins, P. (1995). Approaches to data management, *Controlled Clinical Trials* **16**, 30S–65S.

[3] Spilker, B. & Schoenfelder, J. (1991). *Data Collection Forms in Clinical Trials*. Raven Press, New York.

Bibliography

Controlled Clinical Trials – Special Edition on Data Management for Multicenter Studies: Methods and Guidelines, **16**, (1995). 2S.

McFadden, E.T. (1997). *Management of Data in Clinical Trials*. Wiley, New York.

Ronde, R.K., Varley, S.A. & Webb, C.F., eds (1993). *Clinical Data Management*. Wiley, New York.

ELEANOR T. MCFADDEN

Data Mining, Software Packages for

Data Mining and Biostatistics

The term **data mining**, often used in conjunction with Knowledge Discovery in Databases (KDD), refers to the identification – within a typically large database – of new, valid, and interesting patterns. The focus in data mining shifts away from that of statistical significance, since many effects might turn out to be significant solely because of the magnitude of the sample size. The techniques of data mining borrow from both traditional statistics and computer science, and include methods such as **exploratory data analysis** tools (suitable for large data sets) and predictive modeling tools such as **regression** analysis, **neural nets**, and decision trees (*see* **Computer-aided Diagnosis**). Exploratory methods include, for example, **cluster** and **principal component analyses**, as well as methods that amount to a combination of dimensionality reduction and clustering such as Kohonen maps (which are in fact a special case of a neural net). Techniques such as market basket or association analysis, in which association rules are identified ("those who buy cheese tend to also buy crackers", for example), are also widespread.

While data mining has become most popular in the context of, for example, database marketing, most of the methods under the data mining umbrella have been widely applied in biostatistics. We detail below which main applications have arisen recently.

Data Mining Techniques Commonly used in Biostatistics

Data mining is frequently mentioned in the context of pharmacovigilance. Reference [17] gives the results of a literature search on data mining, signal generation, knowledge discovery in relation to the detection of adverse drug events (ADE) or pharmacovigilance. Among the methods mentioned are predictive methods such as tree classifiers (*see* **Tree-structured Statistical Methods**) and regression models, and market basket analysis, also referred to as link analysis or association analysis. This latter method consists in identifying rules of the form "if x then y", such as, for example, "if drug A is taken, then event B is observed". Reference [15] presents a Bayesian data mining technique (**empirical Bayes**) to detect ADEs and provide details of the **algorithm**. Reference [3] also proposes Bayesian methods for pharmacovigilance signal detection.

Another major area where data mining is applied is the analysis of microarray data. Microarray data consist essentially of **gene expression** levels (one gene per row) for different samples (one sample per column). The Microarray Core Facility website at Dana Farber [10] describes some of the analysis techniques, mostly clustering, used in this context (to find similar gene expression patterns, for example). Hierarchical and K-means (nonhierarchical) clustering methods are described, as well as the Kohonen map algorithm, which in a nutshell, performs a reduction of dimensionality together with a clustering and produces most commonly a two-dimensional map where clusters can be visualized. These methods are also covered in the book [2] on data analysis tools for DNA microarrays.

We also note the application of principal component analysis to gene expression mapping problems [9]. Reference [14] discusses analysis, prediction and discovery in protein data, and [8] discusses association rules in the context of protein sequence patterns.

In the context of genetics, [7] discusses how the results of a market basket analysis, which can be cumbersome because of the large number of rules identified, can be made more useful to the scientist. Reference [13] describes how CART, Multiple Adaptive Regression Splines (MARS), Random Forests and MCMC (**Markov Chain Monte Carlo**) algorithms can be used, notably to help identify **interactions** in predictors of diseases.

We also mention an application of CART to the modeling of the occurrence of bad glycemic control in a diabetes data warehouse [1], and an application of neural nets to the prediction of **infant mortality** in India [16]. Exploratory **factor analysis** was used by [12] to summarize potential predictors of preterm birth in obstetrical patients. Reference [11] uses empirical Bayesian data mining in the detection of vaccine adverse event detection. Reference [5] applies text mining to the tracking of outbreaks of diseases.

We finally refer the reader to a presentation [6], which summarizes methods commonly used in mining health-related data.

A few Software Packages for Data Mining

In a recent article [4], the authors reviewed five software packages for data mining, in alphabetical order, Clementine (SPSS), Ghostminer, Quadstone, SAS Enterprise Miner, and XLMiner. We refer the reader to [4] for details and report here only a few features of particular interest to biostatisticians.

Clementine is a self-standing SPSS package; it can be run independently of SPSS. It covers all the standard data mining procedures and is relatively easy to install and to use. Ghostminer is a self-standing package with some advantages, as well as strong restrictions (see [4]). Ghostminer is easy to install and to use, and comes with a very good documentation, but is likely to be too restricted for the needs of most biostatisticians. Quadstone is a powerful self-standing package that can deal with very large data sets, but which uses methods that may not be standard to many biostatisticians. For example, there is no obvious way to perform a traditional **logistic regression** analysis in Quadstone. Quadstone comes with good documentation and is relatively easy to use once installed, but is very difficult to install (the installation is meant to be performed by an expert). SAS Enterprise Miner runs in conjunction with SAS, and acts as a special module with its own user interface. It is the most complete of all the reviewed packages, and is relatively easy to use. XLMiner is an Excel add-on (the student version was reviewed in [4]; a professional version is now available) with good capability to perform the most common data mining procedures, such as decision trees, logistic regression, and market basket analysis, for example. File sizes are of course restrained to the allowed Excel maximum. XLMiner is very easy to install and to use.

Most data mining packages, including the ones reviewed in [4], tend to target applications such as in database marketing, where focus is often on the *lift* a model can provide. In a nutshell, this means the following. Suppose an analyst is modeling who is likely to respond to an offer. A model is expected to provide a formula, or algorithm, to score the file, that is to assign a score to each observation as to the estimated probability of response. One then sorts the file from the most likely to the least likely to respond and divides the file into, for example, 10 deciles. The first decile should have a higher response rate (percent of responders) if the model is working well. The ratio of a decile's response rate to the overall response rate is often referred to as the *lift*. SAS Enterprise Miner, as well as Clementine and XLMiner provide lift charts.

Because most readers who use SAS might be familiar with the traditional SAS user interface, we reproduce here (Figure 1) a snapshot of an Enterprise Miner diagram, to give an idea of the different user interface.

We also reproduce typical market basket analysis output from SAS Enterprise Miner (Figure 2). The analysis was performed on the Bookbinders Club Case dataset from the Direct Marketing Educational Foundation.

The first rule indicates that people who bought children and art books also tended to buy geography books.

As for Kohonen maps, we refer the reader to the excellent packages provided by the Neural Networks Research Team at the Helsinki University of Technology (`http://www.cis.hut.fi/research/software.shtml`). Freely accessible libraries for the package R provide – among other tools, many of which will be familiar to readers – functions for decision trees, and some forms of the MARS algorithm, as well as functions that are more specialized to, for example, genetics (see `http://finzi.psych.`

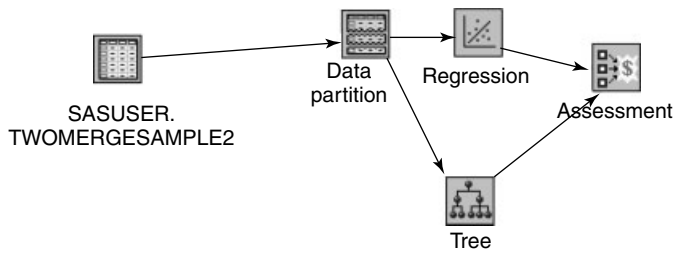

Figure 1 SAS EM diagram. Reproduced with permission from Haughton et al. (2003) [4]; Figure 15

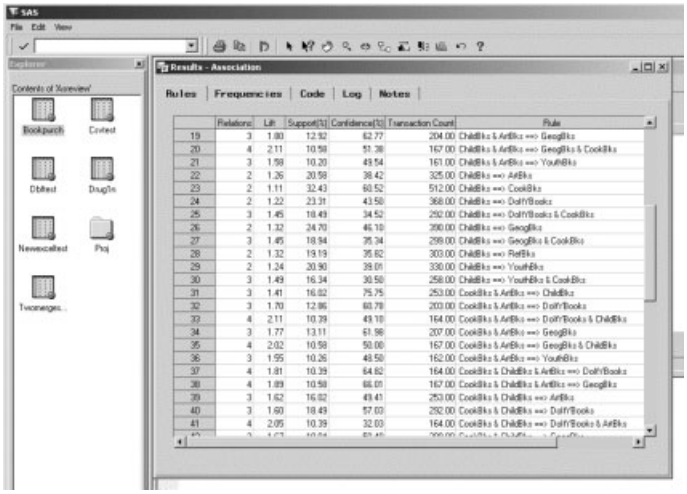

Figure 2 Association analysis results from SAS enterprise miner. Reproduced with permission from Haughton (2003) [4]; Figure 38

upenn.edu/R/library/, http://cran.us.r-project.org/src/contrib/PACKAGES.html).

References

[1] Breault, J.L. (2002). Data mining a diabetic data warehouse, *Artificial Intelligence in Medicine* **26**(1/2), 37–54.

[2] Draghici, S. (2003). *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC, Boca Raton, FL, USA.

[3] Gould, L. & Honig, P. (2004). *Perspectives on Automated Methods for Pharmacovigilance Signal Detection*, DIMACS Working Group on Data Mining and Epidemiology, http://dimacs.rutgers.edu/Workshops/WGDataMining/abstracts.html, accessed on April 20, 2004.

[4] Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N. & Topi, H. (2003). A review of software packages for data mining, *The American Statistician* **57**(4), 290–309.

[5] Hirschman, L. (2004). *Capture and Use of Free Text Information for Tracking Disease Outbreaks*, DIMACS Working Group on Data Mining and Epidemiology, http://dimacs.rutgers.edu/Workshops/WGDataMining/abstracts.html, accessed on April 20, 2004.

[6] Holmes, J.H. (2004). *Mining Health Related Data; Methods and Applications in Research, Public Health and Patient Care*, http://infranet.uwaterloo.ca/talks/2003-2004/2003-10-22/default.pdf, accessed on April 20, 2004.

[7] Imielinski, T. (2004). *Association Rule Mining of Biological Data Sets*, DIMACS Working Group on Data Mining and Epidemiology, http://dimacs.rutgers.edu/Workshops/WGDataMining/abstracts.html, accessed on April 20, 2004.

[8] Kam, H.J., Moon, H.S., Lee, D. & Lee, K.H. (2003). *Discovery of Association Patterns among Protein Sequence Motifs*. PAKDD 2003 Workshop on Biological Stat Mining, http://bi.snu.ac.kr/bdm2003/presentation.html, accessed on April 20, 2004.

[9] Maia, J.M. (2004). Gene expression mapping, *Sixth Annual Winter Workshop: Data Mining, Statistical Learning, and Bioinformatics, Department of Statistics*, University of Florida, www.stat.ufl.edu/symposium/2004/dmbio/Program.pdf, accessed on April 20, 2004.

[10] Microarray Core Facility, Dana Farber. (2004). http://chip.dfci.harvard.edu/stats/analysis.php, accessed on April 20, 2004.

[11] Niu, M.T., Erwin, D.E. & Braun, M.M. (2001). Data mining in the US vaccine adverse event reporting system (VAERS): early detection of intussuception and other events after rotavirus vaccination, *Vaccine* **19**(32), 4627–4634.

[12] Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L. & Hammond, W.E. (1997). Medical data mining: knowledge discovery in a clinical data warehouse, in *Proceedings of the American Medical Informatics Association Annual Fall Symposium*, Nashville, TN, USA, pp. 101–105.

[13] Ruczinski, I. (2004). Interactions and variable importance in genomic data, *Sixth Annual Winter Workshop: Data Mining, Statistical Learning, and Bioinformatics, Department of Statistics*, University of Florida,

www.stat.ufl.edu/symposium/2004/dmbio/Program.pdf, accessed on April 20, 2004.

[14] Schmidler, S. (2004). Statistical shape methods for data mining in protein structure databases, *Sixth Annual Winter Workshop: Data Mining, Statistical Learning, and Bioinformatics, Department of Statistics*, University of Florida, www.stat.ufl.edu/symposium/2004/dmbio/Program.pdf, accessed on April 20, 2004.

[15] Szarfman, A., Machado, S.G. & O'Neill, R.T. (2002). Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database, *Drug Safety* **25**(6), 381–392.

[16] Ventakesan, P. (2004). On the use of artificial neural networks for predicting infant mortality in epidemiological studies, *Sixth Annual Winter Workshop: Data Mining, Statistical Learning, and Bioinformatics, Department of Statistics*, University of Florida, www.stat.ufl.edu/symposium/2004/dmbio/Program.pdf, accessed on April 20, 2004.

[17] Wilson, A.M., Thabane, L. & Holbrook, A. (2003). Application of data mining techniques in pharmacovigilance, *British Journal of Clinical Pharmacology* **57**(2), 127–134.

Further Reading

Clementine. (2004). http://www.spss.com/clementine/, accessed on April 20, 2004.

Ghostminer. (2004). www.fqspl.com.pl, accessed on April 20, 2004.

Quadstone. (2004). www.quadstone.com, accessed April 20, 2004.

SAS Enterprise Miner. (2004). www.sas.com/technologies/analytics/datamining/miner/, accessed on April 20, 2004.

XLMiner. (2004). www.resample.com/xlminer/index.shtml, accessed on April 20, 2004.

DOMINIQUE HAUGHTON

Data Mining

Data mining is a new discipline, which has sprung up at the confluence of several other disciplines, stimulated chiefly by the growth of large **databases**. The basic motivating stimulus behind data mining is that these large databases contain information which is of value to the database owners, but this information is concealed within the mass of uninteresting data and has to be discovered. That is, one is seeking surprising, novel, or unexpected information and the aim is to extract this information. This means that the subject is closely allied to **exploratory data analysis**. However, issues arising from the sizes of the databases, as well as ideas and tools imported from other areas, mean that there is more to data mining than merely exploratory data analysis.

Perhaps, the main economic stimulus to the development of data mining tools and techniques has come from the commercial world: the promise of money to be made from data processing innovations is a familiar one, and huge commercial databases are now rapidly growing in size, as well as in number. However, there is also substantial scientific and medical interest: philosophers of science have remarked that advances and innovation often occur when a mismatch between the data and the predictions of a theory occurs, and nowadays to detect such mismatches often requires extensive analysis of large data sets. Examples of areas of scientific applications of data mining include astronomy [3] and molecular biology [8]. Genomics, proteomics, microarray data analysis, and bioinformatics, in general, are areas that are making extensive use of data mining tools (*see* **Bioinformatics**; **Genetic Markers**; **DNA Sequences**; **Gene Expression Analysis**).

Apart from the sizes of the data sets, one of the distinguishing features of data mining is that the data to which it is applied are often *secondary*. That is, the data will typically have been collected to answer some other question, or perhaps secondarily in the course of pursuing some other issue (medical records will have been collected for monitoring and treating patients, but can subsequently be analyzed *en masse* in the search for previously unsuspected relationships and causes of disease). Of course, there is no reason – apart from the expense of collecting large data sets – why data should not be collected specifically to answer a particular question, but then the analysis is a more standard statistical one.

The excitement of data mining is also partly a consequence of this secondary nature: it suggests that there is valuable information concealed within the data one already has, simply waiting for someone to tease it out. Unfortunately, the "simply" part of this exercise is rather misleading. Indeed, if it was simple, it would doubtless already have been done. One of the problems is that large data sets necessarily have a great deal of structure in them, but this structure has three major sources in addition to the target one of "important, *real*, undiscovered structure". These three sources are data contamination, chance occurrences of data, and structure which is already known to the database owner (or, if not explicitly articulated as known, sufficiently obvious once it has been pointed out to be of no genuine interest or value – such as the fact that married people come in pairs). The first and second of these are sufficiently important to warrant some discussion.

It is probably not too much of an exaggeration to say that all data sets are contaminated or distorted in some way, though with small data sets this may be difficult to detect. With large data sets, it means that the data miner may triumphantly return an unusual pattern which is simply an artifact of data collection, recording, or other inadequacies. Brunskill [1] describes errors that occurred in the coding of **birth weights**: 14 oz recorded as 14 lb, birth weights of one pound (1 lb) being read as 11 lb, and misplaced decimal points - for example, 510 gms recorded as 5100 gms. Note that all of these errors yield over-reporting of birth weights. A data mining exercise might therefore report an unusual excess of high birth weights (indeed, we might hope that a successful analysis *would* report this), but an excess that was of no real interest. Indeed, it seems that this may have happened [2, 7]. Digit preference is another cause of such curiosities, and is only detectable in large data sets. Wright and Bray [9] describe this occurring in measurements of nuchal translucency thickness, and Hand et al. [5] describe it in blood pressure measurements.

In statistics, one is often able to cope with data inadequacies by extending the model to cope with them. Thus, for example, distorted sampling may be allowed for by including a model for the case selection process (*see* **Selection Bias**), and incomplete vectors of measurements may be handled via

the **EM algorithm** (*see* **Missing Data**). However, such strategies can only be adopted if one has some awareness (and understanding) of the data contamination mechanism. In data mining, with secondary data, this is often not the case. Often, such problems are ignored – with obvious potential for misleading conclusions.

Statistics tackles the possibility of spurious (chance) patterns arising in the data using tools which estimate the probability of such structures arising by chance, merely as a consequence of random variation; that is, with **hypothesis** and significance tests. Unfortunately, with large data sets, and large sets of possible patterns being sought, the opportunity for the discovery of apparent structures is clearly great. This means that the statistical approach cannot be readily applied. Instead, data miners simply define score functions (for the "interestingness" or "unusualness" of a pattern), without any probability interpretation, and pass those patterns that show the largest such scores over to an expert for evaluation. This description reveals the *process* nature of data mining. Data mining is not a "one-off" exercise, to be done and finished with. Rather, it is an ongoing process: one examines a data set, identifies features of possible interest, discusses them with an expert, goes back to the data in the light of these discussions, and so on.

The score functions may be the same as the criteria used in statistical model fitting, without the probabilistic interpretation, or there may be other criteria. An illustration of the differences in perspective is given by **regression** analysis. A statistician may find the maximum likelihood estimates of the parameters, assuming a normal error distribution. In contrast, a data miner may adopt the sum of squared residuals as a score function to use in choosing the parameters (*see* **Least Squares**). Since **maximizing the likelihood** based on a **normal error distribution** leads to the sum of squares criterion, these two approaches yield the same result – but they start from different positions. The statistician has a formal model in mind, while the data miner is simply aiming to find a good description of the data. Of course, the distinction is not a rigid one – there is overlap between the two perspectives.

This example does show how central the concept of modeling is to the statistician. In contrast, data miners tend to place much more emphasis on **algorithms**. Given the essential role of computers in data mining, this algorithmic emphasis is perhaps not surprising. Moreover, when data sets are very large, the popular statistical algorithms may become impracticable (tools that make repeated passes through the data, for example, may be out of the question with a billion data points). Computers are, of course, also important for statistics, but many statistical techniques can be applied on small data sets without computers – many were originally developed that way. One consequence of the stress on algorithms is that it may be difficult to describe exactly what model is being fit to the data. This can have adverse consequences. For example, cluster analysis is widely used in data mining, but without careful thought about the nature of the procedure, it can be difficult to be clear about what sort of "clusters" are being found. Thus, compact structures may be appropriate in some situations (e.g. to produce compact summarizing descriptions, with the clusters being represented by "central" points), while in others, elongated shapes may be desirable, in which neighboring points in the same cluster are similar but distant ones are not. Without an awareness of the type of structure that the method reveals, inappropriate conclusions could be drawn: a species could be incorrectly partitioned on a dimension in which it has substantial variability (*see* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**; **Cluster Analysis, Variables**).

In the above, we have used phrases such as "model", "structure", and "pattern" without defining them. Hand et al. [4] define a model as a large scale summary of a set of data (i.e. as the standard statistical notion of a model), and a pattern as a small scale, local structure. A Box–Jenkins decomposition of a **time series** (*see* **ARMA and ARIMA Models**) is a model, whereas a conjunction of values, which occasionally repeats itself (e.g. a petit mal seizure in an EEG trace), is a pattern (*see* **Clinical Signals**). Models are the staples of statistics, but patterns are something with which it has generally not been concerned (there seem to be three main exceptions: the study of scan statistics, of spatial disease clusters (*see* **Clustering**), and of **outliers**, all concerned with local anomalies). An examination of the data mining literature shows that both models and patterns are important, but narrow views of data mining sometimes fail to recognize the diversity of the tools used. Thus, for example, it is sometimes claimed that data mining is merely the application of

recursive partitioning methods (e.g. tree classifiers; *see* **Tree-structured Statistical Methods**), but this is a parody of the breadth of the field. Likewise, the viewpoint sometimes proposed in the econometric literature, that data mining is merely an elaborate and extensive form of model search (*see* **Model, Choice of**), fails to recognize the various other kinds of data mining activities that go on.

A large number of different kinds of tools are used in data mining – reflecting the eclecticism of its origins. Some recent ones in pattern detection, culled from the data mining literature with no particular objective other than to indicate the diversity of different kinds of methods are: tools for characterizing, identifying, and locating patterns in multivariate response data; tools for detecting and identifying patterns in two dimensional displays (such as fingerprints); identifying sudden changes over time (as in patient monitoring); and identifying logical combinations of values that differ between groups. Some examples of important tools in model building in data mining (again chosen with no particular aim other than to illustrate the range of such methods) include: recursive partitioning, cluster analysis, regression modeling, segmentation of time series into a small number of segment types, techniques for condensing huge (tens of billions of data points) data sets into manageable summaries, and collaborative filtering, in which transactions are processed as they arrive, so that future transactions may be treated in a more appropriate manner.

Much statistical theory is aimed at producing valid inferences from a sample to some population (real or notional) from which the sample has been drawn. This might be so that one can make comparative statements about the populations, or for forecasting, or for other reasons. These methods are also appropriate in data mining, provided one has a sample and that it has been drawn in a probabilistic way (so that one knows the probability of each object appearing in the sample). Going further than this, in many data mining applications one has available data on the entire population (for example, all chemical molecules in a particular class) and then, in model building data mining applications, analyzing a sample from the data set may be a sensible way to proceed. In contrast, however, in a pattern detection exercise, it will typically be necessary to analyze the entire data set: if one is seeking those data points that are anomalous, there is no alternative to examining every data point.

It is clear that data mining will be of increasing importance as time progresses. However, the importance should not conceal the difficulties. Finding unsuspected structures in large data sets and identifying those that are due to phenomena of genuine interest and not merely arising from data contamination or due to chance is by no means a trivial exercise. Issues of theory, of data management, and of practice all arise. General descriptions of data mining include those of Fayyad et al. [4] and Hand, Mannila, and Smyth [6].

References

[1] Brunskill, A.J. (1990). Some sources of error in the coding of birth weight, *American Journal of Public Health* **80**, 72–73.

[2] David, R.J. (1980). The quality and completeness of birthweight and gestational age data in computerized birth files, *American Journal of Public Health* **70**, 964–973.

[3] Fayyad, U.M., Djorgovski, S.G. & Weir, N. (1996). Automating the analysis and cataloging of sky surveys, in *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, eds. AAAI Press, Menlo Park, pp. 471–493.

[4] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R., eds. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park.

[5] Hand, D.J., Blunt, G., Kelly, M.G. & Adams, N.M. (2000). Data mining for fun and profit, *Statistical Science* **15**, 111–131.

[6] Hand, D.J., Mannila, H. & Smyth, P. (2000). *Principles of Data Mining*. MIT Press.

[7] Neligan, G. (1965). A community study of the relationship between birth weight and gestational age, in *Gestational Age, Size and Maturity*, Vol. 19. Clinics in Developmental Medicine Spastics Society Medical Education Unit, pp.28–32.

[8] Su, S., Cook, D.J. & Holder, L.B. (1999). Knowledge discovery in molecular biology: identifying structural regularities in proteins, *Intelligent Data Analysis* **6**, 413–436.

[9] Wright, D.E. & Bray, I. (2003). A mixture model for rounded data, *The Statistician* **52**, 3–13.

DAVID J. HAND

Data Monitoring Committees

The randomized controlled clinical trial (*see* **Clinical Trials, Overview**) has become a standard as a research method during the past four decades to evaluate the risk–benefit ratio of novel or existing interventions or therapies. The results of a new or existing intervention are compared with a standard or a control, using clinically relevant outcome measures such as survival, morbidity, or quality of life. Most interventions or therapies should be evaluated by this rigorous comparative methodology before being widely accepted or used in practice. Such evaluation is often required for regulatory approval, especially for drugs and biologics.

For some trials the outcome of the disease or the risks of the therapy may be irreversible. Several issues in the design and conduct of a clinical trial were carefully considered by the Heart Special Projects Committee [13], commissioned by the National Heart Institute of the **National Institutes of Health** (NIH). The report of this committee is often referred to as the Greenberg Report. One of the recommended principles is that comparative trials should be carefully monitored for patient safety and for evidence of benefit. The process is often carried out by a committee, which we shall refer to as the Data Committee Monitoring (DMC). The principles used today in monitoring a randomized control clinical trial are largely influenced by the Greenberg Report. The experience with DMCs has been discussed from several perspectives at an NIH workshop and is described in [7]. Fleming & DeMets [8], DeMets [4], Pocock [20], Fleming [9], Armstrong & Furberg [1], and DeMets et al. [5] discuss various aspects of data monitoring and the role of the DMC. The Task Force of the Working Group on Arrhythmias of the European Society of Cardiology [22] discusses the causes, consequences, and control of early termination, including the role of the DMC. Ellenberg, Fleming, and DeMets provide a comprehensive discussion of DMC organization and activity in their text [8]. This article describes the rationale, responsibilities, and issues that involve the DMC for most trials.

Rationale for Data Monitoring

The goals of a randomized controlled clinical trial are to evaluate the effectiveness of a new intervention, and to assess its safety, so as to estimate the ratio of benefits to risks. One of the ethical principles of clinical trials is that they should continue no longer than necessary to meet the objectives stated in the trial protocol. This is especially true for trials with serious outcomes, such as mortality, morbidity requiring hospitalization, and irreversible adverse effects. Trials may be modified or terminated early if there is overwhelming evidence for a positive benefit to risk ratio, if the evidence strongly suggests harm or a negative benefit to risk ratio, or if the trial has no chance of resolving the primary objectives. Furthermore, if the observed data are not close to the design assumptions, then the trial may need to be modified, such as increasing the sample size, in order to preserve the integrity of the trial. If recruitment goals cannot be achieved in a reasonable time frame, then the viability of the trial must be reassessed. In some cases, logistical or data quality issues must be resolved or the credibility of the trial may be severely jeopardized. All of these aspects must be monitored carefully and considered in the continuation of any trial.

The decision to make protocol modifications, including early termination, is a complex process and several factors must be taken into consideration. This has been discussed by the Coronary Drug Project Research Group [3] and more recently by Fleming & DeMets [10]. Factors include the balance in risk factors between the intervention and control groups, potential biases in outcome ascertainment, compliance to intervention, consistency of results across primary and secondary outcome measures and across clinically relevant subgroups, consistency of results with external information, the effect of repeatedly testing a single outcome or testing multiple outcomes on false positive results, and the impact of early termination on trial participants and future users of the intervention. Examples of these complex decisions have been described for the Coronary Drug Project [3], the Beta Blocker Heart Attack Trial [6], the Cardiac Arrhythmia Suppression Trial [11], and the Physicians Health Study [2]. All of these trials involved early termination, either for an early benefit, an unexpected harmful effect, or for lack of an

effect in the primary outcome. In all cases the decision process was complex and difficult.

A particularly difficult issue is how much evidence indicating potential harm should be allowed to accumulate. In some cases, the choice may be between stopping a trial with a negative trend at the point when the trial has little or no chance to prove treatment beneficial, or continuing a trial with a negative trend until the evidence becomes convincing that the treatment is harmful. In the PROMISE trial [17], which evaluated the drug milrinone in congestive heart failure patients, the trial continued until a statistically significant harmful result was obtained. A similar experience occurred in the PROFILE trial [18], which involved the drug flosequinan in congestive heart failure. Part of the rationale for continuing to this level of evidence was that, unless shown to be convincingly harmful, other beneficial effects of each drug would encourage their continued use in a large patient population. In contrast to this experience, the CONSENSUS II trial [21] with the drug enalapril in congestive heart failure terminated with a negative trend before it became statistically significant. Here, the rationale was that the method of drug delivery would not be used unless it was beneficial. Once the point was reached where that outcome was highly unlikely, there was no reason to continue. In all three cases many factors had to be considered. Another difficult decision is whether the same degree of evidence is required to prove harm as is required for benefit. If the decision process is inherently asymmetric, then the statistical guidelines for data monitoring (*see* **Data and Safety Monitoring**) should also reflect that asymmetry.

While carefully monitoring outcome data in a clinical trial is often ethically mandated, the process of repeatedly examining data also increases the rate of a false positive result; that is, claiming that a difference between two interventions or treatments exists when in fact there is none. Typically, researchers set the false positive rate at 1% or 5% before conducting statistical tests and interpreting P values. However, if a particular outcome such as the primary outcome is tested five times using a P value of 0.05 each time as the criteria for significance, then the actual false positive rate is increased to almost 15%. This is clearly much higher than is scientifically acceptable and five interim analyses are not unusual during the course of a large multicenter trial. This issue is often referred to as repeated testing (*see* **Data and Safety Monitoring**). Another related issue is multiple testing, which refers to conducting statistical tests on multiple outcomes (*see* **Multiplicity in Clinical Trials**), and focusing attention on the one result which has a P value less than 0.05. Clearly, if 20 independent outcomes are statistically tested, then one will by chance alone have a P value less than 0.05. Thus, the ethical mandate of carefully monitoring the outcomes of a clinical trial must take into account the increased chance of falsely claiming a treatment benefit or harm due to the monitoring process. Statistical methods adjusting for the repeated testing and multiple testing have been developed and are discussed in [19, 16], and [14].

To evaluate these diverse factors thoroughly requires a great deal of expertise and experience in clinical trial design, biostatistics, epidemiology, and the subject-matter or disease process involved. No single individual is likely to possess such vast expertise. For this reason, the concept of a monitoring committee evolved, starting with the suggestions made in the Greenberg Report Heart Special Projects Committee [13] (*see* **Clinical Trials, Early Cancer and Heart Disease**). A recent report by the National Institutes of Health has reconfirmed that recommendation [15].

DMC Membership and Responsibility

Since several complex issues must be considered at each interim analysis, requiring expertise from several diverse but relevant disciplines, the membership of the DMC must reflect those disciplines in order to monitor the data and the safety of the patients. The disciplines typically included are the relevant clinical disciplines, laboratory expertise, epidemiology, biostatistics, clinical trials, and medical ethics. Often, three to five individuals are necessary to cover this broad range of expertise. Clinical trial experience by all members of the DMC is highly desirable, but prior experience of serving on a DMC by the DMC chair is essential. Appointment of members may be made by either the sponsor or the trial executive committee, but in either case, the appointments should be acceptable to both parties. The protocol (*see* **Clinical Trials Protocols**) should clearly specify the DMC appointment process.

The authority of this DMC is to review the accumulating data and make recommendations to either

the study chair or the sponsor, or both. While it is rare for a DMC recommendation not to be accepted and fully implemented, the DMC usually does not have the final decision-making authority. That final decision typically resides jointly with the sponsor and the trial executive committee. The trial protocol should carefully specify the lines of the DMC reporting so that any recommendations by the DMC can be properly received and rapidly taken into consideration by those with the final decision-making authority. However, regardless of the lines of reporting, both the trial investigators and the sponsor must be briefed as to the rationale for any DMC recommendation within a reasonably short period of time. Furthermore, the DMC should maintain the view in their deliberations and recommendations that they are primarily responsible to the trial participants, next to the investigators who are placing significant responsibility with them for their patients, then to the trial sponsor, and finally to the regulatory authorities. Any DMC should fully recognize those interests and take them into consideration.

The recommendation of the Greenberg Report was that a trial advisory committee such as the DMC should be independent of the trial. Members of the DMC should not be investigators entering patients or participants into the trial. Otherwise, ethical dilemmas arise as trends in data emerge which are not yet scientifically convincing but could disrupt clinical equipoise about the benefits of the treatment or intervention. In some trials the study chair has been allowed to be an ex-officio member of the DMC, to convey information about the trial to the DMC and to understand better the recommendations of the DMC. In such cases the study chair should not be entering or caring for patients in the trial.

Walters [23] writes that DMC independence is essential for a trial to achieve knowledge with maximum objectivity, respecting the contribution of the patients toward that goal. To be independent, the DMC members should be free of real conflicts of interest, especially since they are reviewing confidential and privileged information. Freedom from any conflict is probably not achievable if the DMC members are expected to have any expertise and experience with the goals of the trial. However, financial conflicts of DMC members should be avoided, including stock ownership and transactions, large consulting arrangements with the sponsor, or frequent speaking engagements on behalf of the intervention. DMC members should at least disclose any consulting or financial arrangements and sources of research funding. If conflicts are identified that could be perceived as serious and possibly damaging to the trial, then the DMC member should either remove the conflict or not continue to participate on the DMC. Neither a sponsor nor a member of a regulatory agency should be a member of the DMC since each has other specific responsibilities and interests and thus is not independent. In some cases, sponsors have been allowed to attend DMC meetings but their role must be limited.

Since the DMC has access to interim results, including primary outcome data (*see* **Outcome Measures in Clinical Trials**), absolute confidentiality is of utmost importance. Results of the trial should not be discussed beyond the DMC meetings and great care must be taken that interim reports are secure. This would include members from sponsoring agencies, should they be allowed to attend. Early in a trial, trends in accumulating data can be quite variable, and any release of early trends could be both misleading and damaging to the conduct of the trial. In some cancer trials, knowledge of interim data and emerging trends had the effect of hampering recruitment and definitive results were not achieved [12]. Of course, if early trends are overwhelming and scientifically convincing, then the DMC might very well recommend early termination. This was the case in the Cardiac Arrhythmia Suppression Trial [11]. However, the DMC must keep interim results confidential until the trial is terminated and the results are properly disseminated.

DMC Meetings

The meetings of the DMC must be structured so that all of the critical elements of a trial are properly reviewed and the patients' safety thoroughly examined. The DMC must meet often enough to carry out its responsibilities. For many trials, the DMC meetings are held at least once a year and may have at least three to five regularly scheduled meetings during the course of the trial. In general, meeting more often than after every 20% increment in patient recruitment or patient outcome is usually unnecessary, and does not lead to substantial gains in early termination. However, it is not unusual to hold an extra meeting if a decision to terminate is approaching. Other considerations such as slow patient recruitment or

unanticipated logistical problems may call for additional meetings of the DMC.

Each DMC meeting must evaluate patient recruitment progress, data quality, baseline characteristics, patient compliance, primary and secondary outcomes, adverse effects, and other safety measures. Interim reports must reflect all of those considerations and can be quite extensive, depending on the complexity of the trial. The DMC must have the authority to request any available data from the trial that is necessary to carry out its primary responsibilities. Reports should be provided through a confidential and secure process to the DMC prior to the meeting so that members have adequate time to review the analysis and identify concerns.

Attendance at the DMC meeting has been discussed by the proceedings edited by Ellenberg et al. [7]. The DMC for the National Institute of Health's AIDS Clinical Trial Group (ACTG) developed a meeting format that addresses most of the concerns of all interested parties [5]. The general format starts with an open session where all interested parties can attend and participate, is followed by a closed session where confidential data are reviewed and discussed by the DMC with the statistical analysis center that prepared the interim report, and concludes with an executive session for DMC members only. Following the executive session, the DMC chair may give the trial chair and sponsor a short briefing on DMC recommendations and any other concerns raised in the closed or executive session.

At the open session, sponsors, representatives of the investigators, and representatives from regulatory agencies may be present to discuss study progress, including recruitment, general data quality, logistical matters such as drug supply or shipment of laboratory specimens, and general compliance issues. Results of other new studies and their impact on the current trial may be discussed as well.

In the closed session, where confidential data are reviewed, the DMC and the trial statistician must be present. For many trials sponsored by the National Institutes of Health (NIH), NIH representatives are present during the closed session. The ACTG DMC did allow NIH representatives to attend, but industrial sponsors who often were also involved did not participate in the closed session. For totally industry-sponsored trials, the practice is not consistent. A few trials have had sponsor representatives present (e.g. PROMISE, PRAISE), but in general this is not recommended routine practice. If representatives of the sponsor do attend, they must abide by the same confidentiality as do members of the DMC, and must not interfere with the DMC deliberations or use the information to affect the trial, unless instructed by the DMC.

The executive session allows the DMC to deliberate and formulate their final recommendations without other influences. This format has been very effective for the ACTG DMC meetings, where several trials are considered during a session, and seems to be useful in many other settings as well.

Summary

The DMC in a clinical trial is central to reviewing accumulating evidence on patient safety and treatment benefit. The complexity of clinical trials makes the decision process to terminate or continue trials challenging and requires a DMC to have a diversity of expertise and experience. DMCs have been utilized in many clinical trials over the past three decades and their value to the clinical trial process is now well established.

References

[1] Armstrong, P.W. & Furberg, C.D. (1995). Clinical trial data and safety monitoring boards: the search for a constitution, *Circulation* **91**, 901–904.

[2] Cairns, J., Cohen, L., Colton, T., DeMets, D.L., Deykin, D., Friedman, L., Greenwald, P., Hutchison, G.B. & Rosner, B. (1991). Issues in the early termination of the aspirin component of the Physicians' Health Study. Data Monitoring Board of the Physicians' Health Study, *Annals of Epidemiology* **1**, 395–405.

[3] Coronary Drug Project Research Group (1981). Practical aspects of decision making in clinical trials: the Coronary Drug Project as a case study, *Controlled Clinical Trials* **1**, 363–376.

[4] DeMets, D.L. (1990). Data monitoring and sequential analysis – an academic perspective, *Journal of AIDS* **3**(Suppl 2), S124–S133.

[5] DeMets, D.L., Fleming, T.R., Whitley, R.J., Childress, J.F., Ellenberg, S.S., Foulkes, M., Mayer, K.H., O'Fallon, J., Pollard, R.B., Rahal, J.J., Sande, M., Straus, S., Walters, L. & Whitley-Williams, P. (1995). The data and safety board and Acquired Immune Deficiency Syndrome (AIDS) clinical trials, *Controlled Clinical Trials* **16**, 408–421.

[6] DeMets, D.L., Hardy, R., Friedman, L.M. & Lan, K.K.G. (1984). Statistical aspects of early termination in the

Beta-Blocker Heart Attack Trial, *Controlled Clinical Trials* **5**, 362–372.

[7] Ellenberg, S.S., Geller, N.L., Simon, R. & Yusuf, S., eds. (1993). Proceedings of "Practical Issues in Data Monitoring of Clinical Trials", Bethesda, Maryland, USA, January 27–28, 1992. *Statistics in Medicine* **12**, 415–616.

[8] Ellenberg, S., Fleming, T. & DeMets, D. (2002). *Data Monitoring in Clinical Trials: A Practical Perspective*. John Wiley & Sons, Ltd., West Sussex, England.

[9] Fleming, T.R. (1993). Data monitoring committees and capturing relevant information of high quality, *Statistics in Medicine* **12**, 565–570.

[10] Fleming, T.R. & DeMets, D.L. (1993). Monitoring of clinical trials: issues and experiences, *Controlled Clinical Trials* **14**, 183–297.

[11] Friedman, L., Bristow, J.D., Hallstrom, A., Schron, E., Proschan, M., Verter, J., DeMets, D.L., Fisch, G., Nies, A.S., Ruskin, J., Strauss, H. & Walters, L. (1993). Data monitoring in the Cardiac Arrhythmia Suppression trial, *Online Journal of Current Clinical Trials* **Doc 79**, CAST-I Online article Bristow.

[12] Green, S., Fleming, T. & O'Fallon, J. (1987). Policies for monitoring and interim reporting of results, *Journal of Clinical Oncology* **5**, 1477–1484.

[13] Heart Special Project Committee: Organization, Review, and Administration of Cooperative Studies (Greenberg Report) (1988). A report from the Heart Special Project Committee to the National Advisory Council, May 1967, *Controlled Clinical Trials* **9**, 137–148.

[14] Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.

[15] Interim Report of the NIH Director's Panel on Clinical Research (CRP), December 1996. `www.nih.gov/news/crp/index.html`.

[16] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[17] Packer, M., Carver, J.R., Rodeheffer, R.J., Ivanhoe, R.J., DiBianco, R., Zeldis, S.M., Hendrix, G.H., Bommer, W.J., Elkayam, U., Kukin, M.L., Mallis, G.I., Sollano, J.A., Shannon, J., Tandon, P.K. & DeMets, D.L. for the PROMISE Study Research Group (1992). Effect of oral milrinone on mortality in severe chronic heart failure, *New England Journal of Medicine* **325**, 1468–1475.

[18] Packer, M., Rouleau, J., Swedberg, K., Pitt, B., Fisher, L., Klepper, M., and the PROFILE investigators (1993). Effect of flosequinan on survival in chronic heart failure. Preliminary results of the PROFILE study, *Circulation* **S1411**, 1642.

[19] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.

[20] Pocock, S.J. (1993). Statistical and ethical issues in monitoring clinical trials, *Statistics in Medicine* **12**, 1459–1469.

[21] Swedberg, K., Held, P., Kjekhus, J., Rasmussen, K., Ryden, L. & Wedel, H. (1992). Effects of early administration of enalapril on mortality in patients with acute myocardial infarction – results of the Cooperative New Scandinavian Enalapril Survival Study II (Consensus II), *New England Journal of Medicine* **327**, 678–684.

[22] Task Force of the Working Group on Arrhythmias of the European Society of Cardiology: The Early Termination of Clinical Trials (1994). Causes, consequences, and control, *Circulation* **89**, 2892–2907.

[23] Walters, L. (1993). Data monitoring committees: the moral case for maximum feasible independence, *Statistics in Medicine* **12**, 575–580.

(*See also* **Ethics of Randomized Trials**)

DAVID L. DEMETS

Data Quality in Vital and Health Statistics

Data quality is not intrinsically interesting, but it is important. To quote **Greenwood**, writing about this in 1948,

> Most statistics can be used and are used for propaganda; hardly any are used more frequently for this purpose than medical statistics, so the student is once again urged to scrutinize the medical-statistical arguments of very important persons with great care and to verify the statistics used [7].

Analyses can only be as good as the data that underlie them: as the computing aphorism puts it, more succinctly if less elegantly than Greenwood's stricture, "Garbage in, garbage out".

To some extent, like beauty, quality is in the eye of the beholder – assessment of it is dependent on the purpose to which the data are to be put. Nevertheless, there are some general points that need to be considered when using routinely collected vital and health data (termed "routine data" hereafter, for brevity). Before discussing these, however, it is worth considering the particular ways in which data quality is of importance for routine data analyses.

Data Quality in Routine Statistics Compared with *ad hoc* Studies

Data quality is of course not solely an issue for routine data analyses – it is important too in *ad hoc* studies in which data are collected specifically for research. The issues present somewhat differently for routine data, however, for several reasons.

First, routine data sets are often extremely large – this is one of their attractions for scientific analysis, but also potentially one of their drawbacks. Files containing many thousands or even millions of events will inevitably have involved data collection by a very large number of people, often with less-close supervision, or at least less-uniform supervision, than can be achieved in smaller research studies.

Secondly, the data have often not been collected primarily for the purpose to which the research investigator may put them. They may have been collected to fulfill legal obligations, or to supply aggregated large-scale statistics for governmental or administrative purposes, and often it is only secondarily that they are utilized for scientific investigation. As a consequence, the data collection methods and quality controls have usually been organized by individuals who will not use the data for the purposes to which the statistician wishes to put them. In comparison, in an *ad hoc* study, the data collection has usually been targeted deliberately to collect the information the analyst requires. Because of this "second-hand" aspect of routine data, the assessment of quality by the user often has also to be a second-hand process. Information on quality may never have been collected, and often the details of how the data were collected, and with what constraints and instructions, may be known only to those working within the data-gathering organization, whose advice and knowledge may need to be sought.

Assessment of Data Quality

The following subsections discuss issues that need to be taken into account when assessing quality.

Quality of Information in the Underlying Data Sources, Especially with Respect to Diagnosis

Routine data can only be as good as the underlying information from which they are compiled. Thus, for instance, it is likely that advances in diagnostic methods have greatly increased the proportion of leukemias and myelomas that are diagnosed now compared with 50 years ago. Past routine data on these malignancies were highly incomplete, not because of inadequate data collection, but because the cases were often not diagnosed. Similarly, geographical differences in apparent incidence of conditions may reflect better diagnosis in one place than another (*see* **Geographic Patterns of Disease**; **Mortality, International Comparisons**; **Small Area Variation Analysis**).

Incompleteness and inaccuracy of diagnosis may arise at several stages in the process from disease incidence to diagnostic labelling, and each needs to be considered when deciding whether apparent variations in rates are an artefact. There may be variation in whether subjects with illness realize they are ill, and whether they present to a doctor for diagnosis, depending, for instance, on social, financial, and educational factors. If patients present to a doctor, the

diagnosis will depend on factors such as the propensity of the doctor to investigate, his or her diagnostic acumen, whether referral is made to a specialist, and the diagnostic methods and technologies employed, including, for fatal conditions, the extent to which autopsies are performed. Thus, for instance, incidence rates of prostatic cancer have been found to be much greater in Malmö, where "the autopsy service is superb", than in other cities in Sweden, but when cases found "accidentally" at autopsy were excluded, the incidence rates in Malmö and other cities were similar [20]. If screening for a disease is introduced (*see* **Screening, Overview**), then asymptomatic cases will be detected that either would never otherwise have been diagnosed, or would have been diagnosed at a later date after they had become symptomatic; the former detection would be expected to lead to a permanent artefactual increase in rates, the latter to a temporary increase.

For mortality, when a diagnosis has been reached, the quality of the eventual statistics will also depend on how well the person certifying the cause of death (*see* **Death Certification**) (usually a doctor, but not always – see below) knows the patient's past medical history, which diagnosis he or she considers to have caused death (*see* **Cause of Death, Underlying and Multiple**), and the way in which he or she completes the death certificate because the positioning of causes there can affect the underlying cause selected by the coding agency [8]. Studies requesting different practitioners to complete "dummy" death certificates for the same case history have been used to try to ascertain the extent to which, for instance, international differences in certification practice can affect apparent mortality rates [11], and similarly by requesting national statistical offices to code the certificates, to investigate how this coding affects rates [11, 17].

One method frequently used to check the quality of clinical or death certificate diagnostic information is to compare the diagnoses from this source with those reached by autopsy [6, 10]. An often-used general marker of quality of death certificate information in a population is the proportion of deaths recorded as due to senility: a high percentage of deaths certified to this cause suggests a poor quality of diagnostic information, at older ages at least. Quality of diagnostic data is also suggested by its source or basis – for instance, for mortality data, whether the diagnosis is supplied by a doctor or not, and for cancer registrations, the proportion of cases with histological verification, (although new diagnostic technologies, for instance, ultrasound imaging plus serum alpha-fetoprotein estimation for diagnosis of liver cancer [16], can sometimes lead to a reduced percentage histologically verified without reduced quality). For mortality, Alderson [1] gives tables of the extent to which deaths have been certified to ill-defined causes and the percentage of death certificates signed by doctors, in 31 countries (Japan and Western) since early in the twentieth century, and a description of the death registration system in each of these countries over time. For cancer registration (*see* **Disease Registers**), tables of the proportion of cases histologically verified, and other quality indicators – the proportion of cases registered from a death certificate only, the ratio of mortality to incidence, the percentage of cancers for which the primary site is unknown or ill-defined, and the proportion of cases with age unknown – for over 100 cancer registries worldwide, are given in *Cancer Incidence in Five Continents* [16].

As well as differences in diagnostic completeness and capability, routine data will also be affected by medical definitions of diseases, which may vary greatly by time or place, and can lead artefactually to apparent large differences in rates. Thus, for instance, great differences in diagnosis between countries have been shown for psychiatric conditions [3], and substantial apparent secular changes in bladder cancer rates can occur as a result of changes in pathological nomenclature for classifying papillomas (e.g. as "grade O carcinomas") [20].

The above discussion has related to the quality of diagnostic data, but the quality of the source data for other variables such as age, sex, and country of birth, which may be used in analysis, is also important and needs to be considered (see below).

Completeness of Data Collection

For legal reasons, certification of births and deaths is normally virtually complete in Western countries, although there can be exceptions, for instance during wartime [1]. Completeness may be deficient in a particular area within a country: for instance, a study in the West of Ireland, found that in 1966–69, 7.5% of deaths were not registered, and in 1974–77, 6.1% [13]. In the US, satisfactory levels of registration were achieved by different states in different years (the last, of the then-existent states, Texas, reached the level required to be admitted to the

"National Death Registration Area", for which the federal government publishes data, in 1933) [12]. One stratagem to overcome quality deficiencies in particular geographic areas is to analyze only a geographical subset of the overall data set (e.g. particular states), for which good quality data are available.

Registration of morbidity is more often incomplete, and when comparing morbidity rates between places or over time, one must consider whether differences in completeness may explain apparent differences in incidence. Multiple data sources (e.g. death certificates, hospital admissions lists, pathologists' reports) may be needed to gain a high level of completeness, and addition of a new data source, for instance adding death certificates as a source for cancer registration, can lead to an abrupt increase in recorded rates [22]. For cancer registrations [18] and infectious disease notifications [24], countries differ as to whether reporting is legally compulsory, but it is not clear that in practice this affects completeness. In several countries a fee is paid to the notifying doctor for each infectious disease notification [24], but again it is not clear that this provides high completeness.

Often, completeness is better for more-serious than for less-serious conditions, if only because the most serious lead to death. Thus, for instance, nonmelanoma skin cancers (which are rarely fatal) tend to be the worst registered of the common cancers, and notification of infections such as acute poliomyelitis and diphtheria is likely to be far more complete than that for dysentery [21]. For most purposes, substantially incomplete data are very difficult to interpret, especially because the missing data may be biased, but for certain uses they may be serviceable: for instance, if an infectious disease notification system is used primarily to detect epidemics, then it would not in principle invalidate its use if it was substantially incomplete, provided that the percentage incompleteness remained approximately constant over time. Nevertheless, greater confidence can be had in analyses based on reasonably complete data, because one can rarely be sure that the degree of incompleteness has remained unchanged; for example, for measles there is evidence that completeness can differ between epidemic and nonepidemic years, being greater during epidemics [5]. (Of course, if this was reliably the case, then it would actually improve detection of epidemics, although still diminishing the value of the data for scientific uses.)

Assessment of completeness is ideally carried out by comparison with a "**gold standard**" complete dataset collected by independent means, either by an *ad hoc* survey or in another routine data system – for instance, comparison of registration data with death certificates for anencephaly has been used to check completeness of congenital malformation notification [25]. Failing this, however, **capture–recapture** techniques can be used to assess completeness by comparing the data with those collected by other incomplete methods. For cancers, a frequently used but imperfect measure of completeness is the mortality to incidence ratio, comparing the numbers of cancers in mortality and cancer registration data for the same year(s): this can give an approximate guide to completeness, particularly for rapidly fatal cancers, but it is imperfect because it depends on comparable accuracy and precision in identification of cancer site and comparable definitions of place of residence in the two data sets, and on case-fatality and secular trends in incidence and mortality rates, as well as on completeness. It is also imperfect because death certificates are often used as a source of cancer registration, so that incidence and mortality datasets are not independent. Another often-used indicator of completeness of cancer registration is the percentage of cases registered from a death certificate only: a high percentage indicates likely incompleteness, since equivalent nonfatal cases would probably never be registered.

Duplication

Like incompleteness, duplication should not be a problem for mortality and births data in Western countries, but it can be a substantial one for morbidity data. To avoid duplication, the data collection agency must first link multiple notifications that refer to the same person (*see* **Record Linkage**), and then differentiate between genuine double occurrence of the disease or event of interest in the same individual – for instance, two primary cancers incident in the same person, or a particular infectious disease caught on more than one occasion – and inadvertent duplicate recording of the same morbid event, often because it has been reported from more than one source.

Much less tends to be published about the extent of duplication within routine datasets than about their completeness, and this makes it particularly difficult

for the user to be sure to what extent duplication has occurred. One issue that the user may need to clarify is the rules used by data coders to decide what should count as a duplicate – for instance, whether two primary cancers of the same site but different histologies, or two contralateral tumors in paired organs such as kidneys or testes, should count as one or two malignancies; the effect on recorded incidence rates can be appreciable. As another example, for tuberculosis statistics one needs to ascertain whether the data refer to new cases only or to all cases (new and relapses); for some countries the former are not readily available, so that international comparisons may need to use the latter [19]. On a broader question, the user will also need to note whether the dataset is based on persons, disease occurrences, or events.

Having determined the basis of the data set and the rules used for decisions on duplicates, the user needs to compare these with the purpose of the study. Thus, rates of cancer incidence in an analysis including second and subsequent primary cancers will be greater than rates restricted to first primaries; neither is of lower quality for the purpose for which it was intended, but either is inappropriate or of deficient quality if intended for the opposite purpose. Similarly, rates of hospital admission from a hospital in-patient data system, counting two admissions of the same person for the same disease incidence as two records, can give useful information for health care planning, but will generally be unsatisfactory for epidemiology.

Late Registrations, Alterations, and Deletions

Whereas birth and mortality data are normally collected within a few days of occurrence of the event, and will tend to produce complete statistics within a few months, complete collection of morbidity statistics may take several years. For instance, cancer registration inevitably takes a year or two to become reasonably complete: data must be obtained from several sources (for instance, clinical records, pathology records, and death certificates), cross-matched and duplicates eliminated. Furthermore, diagnostic confirmation of an initially suspected cancer may take weeks or months, and cancers initially identified by a registry at death may prove in retrospect to have been incident months or years earlier, and will then need to be registered as having occurred at that date of incidence. Because of these delays, plus the time taken to compile statistics and publish them, cancer registration statistics that are nominally for the same year of incidence may differ appreciably depending on how long after the incidence date they were published. If data are analyzed too soon after incidence, then apparent decreases in rates may prove to be artefacts. Similarly, assessments of completeness of cancer registration need to be conducted sufficient years after incidence, if they are not to confuse lack of timeliness with eventual incompleteness: an apparent recent decline in completeness may be a consequence of premature assessment.

As well as leading to late registration of an event not previously registered, late information may also alter a diagnosis or other variable already recorded – for instance, an autopsy may reveal that a cancer registered years earlier at initial diagnosis was in fact of a different site from that registered, or indeed was not a cancer at all.

Some routine data systems incorporate a specific facility to improve quality after initial data collection, by taking account of new information from subsequent diagnostic investigation: for instance, to amend a death certificate diagnosis on the basis of autopsy findings [23] or to correct (or delete) an infectious disease notification on the basis of laboratory reports [24].

Validity and Precision of Data Collection

Data extraction from original sources, often by clerks, needs to be conducted accurately and without, for instance, miscategorization of adjacent anatomical sites or similar sounding diseases (*see* **Misclassification Error**). Accuracy can be measured by comparison with data re-collected from original sources [2], or by searching the files for the frequency of impossible or unlikely values or combinations, suggestive of inaccuracy – for instance prostate operations on women. It should be noted, however, that this will only provide a proxy for general quality if the data-collecting agency has not already conducted range and consistency checks to rid the data of these particular errors, and that if they have done so, this will not in itself produce a completely "clean" dataset because it will leave all those errors that are not illogical or outside plausible ranges.

Less obviously, but also importantly, apparent rates of a disease will be affected by the extent to which information on diagnosis, even when correct, is sufficiently precise to identify that particular disease, rather than a more vague or general category that includes the disease. Thus, for instance, malignant melanoma of the conjunctiva is coded in the **International Classification of Diseases** to "malignant neoplasm of the conjunctiva" (ICD-9 code 190.3), but the same tumor described as a melanoma of the eye would be coded to a different 4-digit code (190.9), and if stated simply as "melanoma" would be coded to a different 3-digit category (172.9). In more extreme circumstances, if the tumor were simply known to be an eye disorder, unspecified, it would be coded in a different ICD chapter (379.9), and if stated as a death of unknown cause, in another chapter again (799.9). All of these codes are therefore locations where melanomas of the conjunctiva could be allocated, depending on the precision of data available, and the extent to which conjunctival melanoma statistics are valid depends on the extent to which such tumors are precisely specified and coded.

When procedures are introduced to improve precision of information, artefactual increases will occur in rates of precise diagnostic categories. Conversely, discontinuation of such procedures will tend to reduce apparent rates. For instance, in England and Wales from 1881 to 1992, "medical enquiries" were sent to certifying medical practitioners, requesting more precise diagnostic information when a death certificate diagnosis was deemed to be too vague. When this enquiry procedure has been discontinued, for instance, in 1981–82 due to a strike of Registrars, large changes in apparent rates of precise diagnostic categories occurred [23]. The data user can attempt to take account of such changes, first by asking the data collecting agency whether they have used such procedures and when these have changed; secondly by calculating and assessing rates for relevant imprecise ("dustbin") categories in parallel with consideration of the precise category under investigation; and thirdly, by tabulating data by single calendar year and looking for step-changes in rates, which are likely to indicate artefacts (of many types).

As well as the quality of diagnostic data, and of denominators (see below), it is also important to consider data quality, both in source material and in data collection, for variables by which **stratification** or adjustment (*see* **Standardization Methods**) will be made in analysis; for instance, age, sex, and occupation. The percentage of individuals with missing data for such variables (especially age and sex) can provide a useful overall quality indicator, and the extent to which digit preference is present for age (i.e. an excess of ages ending in 0 or 5) can indicate the quality of the age information.

Coding, Bridge-Coding, and Assignment of Underlying Cause

Interpretation of routine datasets is dependent on understanding of the coding system used. This may be an internationally accepted and accessible system, such as the **International Classification of Diseases** (ICD) [26], or, for instance often for operations or occupations, it may be a locally derived classification, which may be difficult to access outside the country. Even when the coding system is internationally agreed, there may be superimposed upon it local interpretations and deliberate deviations. Thus, in England and Wales, mortality coders use in addition to the ICD, a large locally derived manual with numerous instructions on actions to take for particular descriptions of disease for which the ICD does not give sufficient guidance, or the Registrar General has decided that the ICD should not be followed. Similarly, routine data agencies sometimes convert data coded under one revision of a coding system to another, and may not subsequently keep the originally coded data. Interpretation is then dependent on the methods for, and quality of, the conversion as well as the original coding. For instance, for disease coding, since there is not an internationally agreed ICD conversion system, local judgements will have been made, and these may be different from those that the investigator might have chosen.

For underlying cause mortality statistics, the statistical agency must, as well as coding diseases, select the "underlying cause" of death when more than one disease is mentioned on the death certificate (*see* **Cause of Death, Underlying and Multiple**). Again, although there are internationally agreed rules in the ICD for making this choice, local decisions are likely to have been superimposed, and occasionally the ICD rules may deliberately have been broken. For instance, in England and Wales from 1984 to 1992, Rule 3 of the ICD, concerning selection of the underlying cause of death, was deliberately set aside for

individuals with a "major cause" of death such as cancer in part II of the death certificate, but whose cause of death under Rule 3 would have been a "terminal event" such as heart failure or unspecified pneumonia. This change greatly increased apparent mortality from several conditions – for instance a 44% increase occurred in deaths from diabetes and a 22% increase in deaths from multiple sclerosis [15].

The above issues will be compounded when a data set has been formed by aggregation of records collected and coded by several different agencies, rather than only one: for instance, when data for a national morbidity registration system are collected and coded regionally, and then brought together to form a national data set. Artefacts may then occur from each coding source, and also secular discontinuities may occur when there is a change in the level of the hierarchy, e.g. regional vs. national, at which coding is undertaken.

Data Processing and Editing Errors

Clerical processing may produce individual errors or repeated ones (*see* **Data Management and Coordination**). Computers offer the opportunity to create large-scale errors in data at high speed. Often the data have been processed in batches through several stages, and errors may arise if a particular batch is incorrectly processed through a stage, or has a stage omitted. Thus, for instance, cancer registration data from registries using the International Classification of Diseases for Oncology (ICD-O) coding system will need to be re-coded if aggregate data from several registries are to be held and analyzed in ICD-9. Since most codes are the same in these two systems, it may not be immediately obvious if a batch of data has not been re-coded, but the presence of a code that exists in ICD-O but is impossible in ICD-9 (ICD code 169) should alert the user to a likely failure in re-coding.

Computer editing of data can also lead to artefacts. Thus, unknown to the user default values may have been substituted for missing information. In analyses of seasonality, for instance, such default coding for month and day of birth can lead to an apparent large peak of incidence in people born on June 30! Similarly, many editing systems do not allow entry of records with variables missing, and data coders or processors may then invent values, when these cannot be ascertained, to enable records to be entered into the system. One should be suspicious that this has occurred when large datasets, for instance on national mortality, are published with apparently no individuals with missing values.

Denominators

Rates calculated from routine data sources are dependent on the quality of denominators (*see* **Denominator Difficulties**), frequently but not always derived from the **census**, as well as on that of the numerators, to which more attention is often paid. Census data will tend to be reasonably accurate in Western countries, although for groups who are particularly difficult to count, e.g. vagrants, the very old, or immigrants, the quality of these data may be a more serious problem.

Inaccuracy may be a greater problem in population estimates for intercensal years, which need to take account of estimated migration, births and deaths since the last census count. Thus, for instance, Draper et al. [4] found apparent secular increases in childhood leukemia rates of 50% in non-Hispanic whites and 53% in blacks in Los Angeles based on population estimates relating to a previous census, but when the rates were recalculated with denominators that took account of a subsequent census, the increases became only 22% and 24%, respectively.

Numerator/Denominator Discrepancies

A greater problem than the completeness and accuracy of the denominator data is frequently their appropriateness and comparability with the numerator. When denominator data originate (as they usually do) from a different source to that used for the numerator, differences in quality, definitions, and data collection methods between these sources may lead to **bias**. For instance, a census might count university students at their term-time place of residence, but when ill they may return to their parental home for treatment, and hence rates for serious disease in young adults might be underestimated for university towns, especially those where a large proportion of the young population are students. Similarly, the occupational data recorded at the census for an individual will usually be that person's own statement of their occupation, whereas on their death certificate the occupational description will inevitably have been given by someone other than themselves, usually a relative. The relative may view the past career

of the deceased in a flattering light, such that electricians may become electrical engineers, and nursing assistants become nurses, or the relative may report a more prestigious occupation that the deceased left many years ago by early retirement (for instance, military officer or aircraft pilot), rather than a current, less prestigious occupation that might have been reported at the census [14]. In studies of disease risk by country of birth, the census report of an individual's birthplace may be different from that in a relative's statement, and discrepancies can also arise because these two variables are collected at different times, and boundaries of countries can change over time.

Special Issues for Analysis of Clusters and Other Rare Events

As well as the general considerations above, some further issues apply mainly, or with more force, to analyses of spatial or temporal clusters and other rare events (*see* **Clustering**). One is that rates of rare events may be far more disrupted by occasional, random errors or duplications in a dataset than are statistics for the data overall. For instance, a low level of duplication of records can produce an apparently highly significant cluster of cases in a particular small area which is due simply to records for one or two individuals being recorded two or three times, even though overall rates of disease in the dataset will have been little affected by the duplication rate. Similarly, low rates of random **misclassification** between a common and a rare category will have a far greater impact on the rate in the rare category than the common one. For instance, breast cancer is about 100 times more common in women than men. If there is a 1% random error rate in categorization of sex in cancer registration data, this will approximately double the apparent number of breast cancers in men but have a negligible effect on the number in women. Similarly, small random error rates in coding (or data entry) of diagnosis may greatly inflate apparent rates of rare diseases, while having little impact on rates of common diseases.

These considerations suggest that a higher level of data quality is needed for valid analysis of rare categories or clusters than for common categories, and that an effort should be made to verify the original individual records before coming to conclusions on the presence of clusters of small numbers of cases, or rates of rare diseases, within routine datasets.

Record linkage

Many of the most interesting uses of routine statistics relate to linkage between datasets – for instance, linking a births file with a childhood or adult morbidity file to determine whether prenatal risk factors are associated with risk of later disease (*see* **Record Linkage**). It should be noted that the validity of the analyses will be dependent on both the quality of the two datasets to be linked, and the quality of the linkage. Such problems will be cumulated where data are derived by successive linkage, in stages, through several data sets. For instance, in England and Wales notification of incident cancers in study cohorts can be obtained from the National Health Service Central Register (NHSCR); for childhood cancers it has been found that this is 12.5% incomplete, because of the cumulation of shortfalls of a few percent at each of the several stages at which the data are gained, transmitted or linked, starting from initial cancer registration by a regional cancer registry and ending with identification by NHSCR that the cancer occurred in a cohort member, and notification of this information to the investigator [9].

Conclusion

Most epidemiologists have had the disappointing experience of making an apparent discovery in a dataset, only to find on more careful examination that it was in fact due to a deficiency of quality in the data, or to a misunderstanding of the way in which the data were compiled. It is worth paying great heed to data quality, if only to try to ensure that as far as possible one's publications are not similarly in error. As Greenwood noted half a century ago in his discourse on this subject, "This may involve a little trouble – which is worth taking. One should *never* believe that a disease is becoming more, or less, deadly until all other explanations have been excluded" [7].

References

[1] Alderson, M. (1981). *International Mortality Statistics*. Macmillan, London.

[2] Brewster, D., Crichton. J. & Muir, C. (1994). How accurate are Scottish cancer registration data?, *British Journal of Cancer* **70**, 954–959.

[3] Cooper, J., Kendall, R., Gurland, B.J., Sharpe, L., Copeland, J.R.M. & Simon, R. (1972). *Psychiatric diagnoses in New York and London. A Comparative Study*

of Mental Health Admissions. *Maudsley Monograph* No. 20. Maudsley, London.

[4] Draper, G.J., Kroll, M.E. & Stiller, C.A. (1994). Childhood cancer, in *Trends in Cancer Incidence and Mortality*. R. Doll, J.F. Fraumeni Jr & C.S. Muir, eds. *Cancer Surveys*, Vol. 19/20. Cold Spring Harbor Laboratory Press, New York, pp. 493–517.

[5] Fine, P.E.M. & Clarkson, J.A. (1982). Measles in England and Wales – II: The impact of the measles vaccination programme on the distribution of immunity in the population, *International Journal of Epidemiology* **11**, 15–25.

[6] Goldman, L., Sayson, R., Robbins, S., Cohn, L.H., Bettmann, M. & Weisberg, M. (1983). The value of the autopsy in three medical eras, *New England Journal of Medicine* **308**, 1000–1005.

[7] Greenwood, M. (1948). The sources and nature of statistical information in special fields of statistics. Medical statistics. *Journal of the Royal Statistical Society, Series A* **111**, 230–234.

[8] Grulich, A.E., Swerdlow, A.J. dos Santos Silva, I. & Beral, V. (1995). Is the apparent rise in cancer mortality in the elderly real? Analysis of changes in certification and coding of cause of death in England and Wales, 1970–1990, *International Journal of Cancer* **63**, 164–168.

[9] Hawkins, M.M. & Swerdlow, A.J. (1992). Completeness of cancer and death follow-up obtained through the National Health Service Central Register for England and Wales, *British Journal of Cancer* **66**, 408–413.

[10] Heasman, M.A. & Lipworth, L. (1966). *Accuracy of Certification of Cause of Death*, *General Register Office SMPS* no. 20. HMSO, London.

[11] Kelson, M.C. & Heller, R.F. on behalf of the EEC Working Party (1983). The effect of death certification and coding practices on observed differences in respiratory disease mortality in 8 EEC countries, *Revue d'Épidémiologie et de Santé Publique* **31**, 423–432.

[12] MacMahon, B. & Pugh, T.F. (1970). *Epidemiology. Principles and Methods*. Little, Brown & Company, Boston.

[13] Medico-Social Research Board (1979). The registration and certification of deaths in the West of Ireland, in *Annual Report 1979*. Medico-Social Research Board, Dublin, pp. 13–17.

[14] Office of Population Censuses and Surveys (1978). *Occupational Mortality. The Registrar General's Decennial Supplement for England and Wales*, 1970–72. Series DS no. 1. HMSO, London.

[15] Office of Population Censuses and Surveys (1985). *Mortality Statistics, Cause, England and Wales 1984*. *Series DH2* no. 11. HMSO, London.

[16] Parkin, D.M. & Muir, C.S. (1992). Comparability and quality of data, in *Cancer Incidence in Five Continents*, Vol. VI, D.M. Parkin, C.S. Muir, S.L. Whelan, Y.T. Gao, J. Ferlay & J. Powell, eds. *IARC Scientific Publication* no 120. IARC Lyon, pp. 45–173.

[17] Percy, C. & Dolman, A. (1978). Comparison of the coding of death certificates related to cancer in seven countries, *Public Health Reports* **93**, 335–350.

[18] Powell, J. (1992). Techniques of registration, in *Cancer Incidence in Five Continents*. Vol. VI. D.M. Parkin, C.S. Muir, S.L. Whelan, Y.T. Gao, J. Ferlay & J. Powell, eds. *IARC Scientific Publication* no. 120. IARC, Lyon, pp. 3–24.

[19] Raviglione, M.C., Snider, D.E. Jr & Kochi, A. (1995). Global epidemiology of tuberculosis. Morbidity and mortality of a worldwide epidemic, *Journal of the American Medical Association* **273**, 220–226.

[20] Saxén, E.A. (1982). Trends: facts or fallacy, in *Trends in Cancer Incidence. Causes and Practical Implications*, K. Magnus, ed. Hemisphere, Washington, pp. 5–16.

[21] Stocks, P. (1949). *Sickness in the Population of England and Wales in 1944–1947*. *SMPS* no. 2. HMSO, London.

[22] Swerdlow, A.J. (1986). Cancer registration in England and Wales: some aspects relevant to interpretation of the data, *Journal of the Royal Statistical Society, Series A* **149**, 146–160.

[23] Swerdlow, A.J. (1989). Interpretation of England and Wales cancer mortality data: the effect of enquiries to certifiers for further information, *British Journal of Cancer* **59**, 787–791.

[24] Taylor, I. (1965). *The Notification of Infectious Diseases in Various Countries*. Public Health Papers 27. WHO, Geneva, pp. 17–68.

[25] Weatherall, J.A.C. (1969). An assessment of the efficiency of notification of congenital malformations, *Medical Officer* **121**, 65–68.

[26] World Health Organization (1977). *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death*. WHO, Geneva.

A.J. SWERDLOW

Database Systems

The word "database" is used here in two overlapping ways to refer to a collection of related data and its storage method, or to programming system software used to organize and integrate a collection of related data. It is management, access, and control (including efficiency and recoverability), not necessarily size, which distinguish a database from a data set [23]. There are other definitions of "database"; see, for example, [21]. The following concentrates on database software: choosing a database management system (DBMS), designing databases, and resources that might be helpful.

Defining a DBMS as "a collection of programs that enables users to create and maintain a database" [8], Date [6] identified the following benefits of using a proper DBMS. First, data can be shared securely. Secondly redundancy can be reduced, inconsistency avoided, and integrity maintained – that is, you have the data you intended to hold, with no repeats, unexplained absences or anomalies. Thirdly, data independence can be achieved – the user does not need to know how or where the physical data are stored, but instead identifies tables and variables by name. Finally, and best of all, (conflicting) requirements can be balanced – and standards can be enforced.

What a DBMS Stores

A DBMS can store data, which are usually alphanumeric, but increasingly include images, sound, and free-format text for multimedia application, as well as the following types of information:

1. "Business rules", namely database description tables containing extra information about the data not immediately apparent from the data – for instance, that a given field cannot be omitted, or that gender must be "M" or "F". A good relational DBMS (RDBMS) can enforce business rules for you (retrospectively) when required.
2. A data dictionary describing the data, and who has access to them, namely the authorization tables and the current access tables.
3. Audit and system performance information.

What a DBMS Does

For the user, a DBMS (and third-party tools that work with it) might offer: database design tools with a data definition language; database administration tools; a data manipulation language, or query language, to query and update information; data-entry screens and screen designers; interfaces to other applications or programming languages; and reporting and statistical summary tools.

Uses of Databases in Biostatistics

As well as all the typical accounting and administrative uses, databases are used in biostatistics for:

1. Administering studies – storing names and addresses of subject participants, generating letters and labels, allocating interviewer schedules, creating progress reports.
2. Collecting data – with answers or results typed straight into a laptop or notebook, or downloaded directly from a measuring instrument.
3. Managing data – data entry, storage, and secure access to all types of data.
4. Analyzing data – validation, description, modeling and analysis.
5. Documenting studies – even the study design can (and should) be stored in a database!

In fact, the limiting factors for how widely databases are used are more likely to be time, energy, and money than database shortcomings!

A Brief History of Databases

The history of databases parallels part of the history of computing. An overview of the history of computing can be found at `http://ei.cs.vt.edu/~history/index.html`. The earliest true DBMSs appeared in the 1960s. For example, in 1960 a new language, COBOL (for "Common Business Oriented Language"), had the novel approach of separating data description from programs – in a database – so data could be easily reused across applications. Programs could then refer to established variable names.

Early databases were built for speed, not flexibility. Database design was dependent on anticipated use. For example, in an (imaginary) early hospital

patient appointment system, you would have had a link from a patient to the point on a file where her appointments began, then all her appointments would be listed one after another. In this "hierarchical" database, some operations would be easy (finding all the appointments a patient had), some would be hard (adding a new appointment meant rewriting the whole file), and some almost impossible (finding all the patients with appointments falling on a certain day). To solve the latter two problems, more links (often called "pointers") were added to records. Then our imaginary patient would be linked to her first appointment, that to the next appointment, with her last appointment completing the circle and joining back to the patient. The name of the committee that made COBOL, CODASYL, is still used in discussions on such "network" DBMSs. The main drawback of CODASYL databases is that they are obscure to program – you need to know which way the links point. In 1997, experienced COBOL programmers are being called out of retirement to check code that might have the "millennium bug" – an artifact of treating all dates as belonging to the twentieth century.

In 1970, Codd [4] proposed a theoretical table-based data storage model, and a universal data manipulation language. His ideas (expressed concisely as a set of 12 rules in 1985 [5]) have yet to be fully realized, but his RDBMSs and Structured Query Language (SQL) are now common, developed by IBM's DB2, Ingres, and Oracle from the mid-1970s, to name a few of the industry leaders. In fact SQL is the standard for both the International Organization for Standardization (ISO) and the American National Standards Institute (ANSI) – see `http://www.jcc.com/sql_stnd.html` – so there is some cross-vendor portability. These fully implemented RDBMSs were (and are) hard to administer – Oracle, for instance, recommends identifying a database administrator (DBA) who should have 2 weeks' training before trying to install any software, with a week's extra training on "tuning". Even writing a data-entry screen is a professional job. It is also very expensive!

The 1980s saw an explosion in personal computers and simple, file-based databases. In these the data structure (as well as the number of records to expect and the length of each record) is held at the top of a data file. Joining data from different files is possible, but not intuitive – but as most people (biostatisticians included) still enter and look at data one table at a time, software like Ashton Tate's (now Borland's) "dBase" remains enduringly popular 10 years on.

For computer users, the 1990s have been characterized by the emergence of graphical user interfaces (GUIs). Accompanying this development has been the need for databases storing nonalphanumeric data – video, audio, images, free-format text – that is, "object" or "multimedia" DBMSs. GUIs also encourage a conceptual separation of the "client" (a program that shows data) and the "server" (a DBMS that manages it). Much effort has been put into hiding the connection between these components. Newer, mouse-driven databases, like Microsoft's "Access", can be self-contained, or can connect transparently as a client to SQL servers over a network, or even over the **Internet**. This "scalability" comes at a price – and setting up a client/server database is (currently) hard and expensive.

Choosing Software to Manage Data

For biostatisticians involved with drug trials (*see* **Pharmaceutical Industry, Statistics in**), the choice is straightforward – SAS (*see* **Software, Biostatistical**) is the *de facto* standard for statistics and data management (see [22] and `http://www.fda.gov/cdrh/ost/points.html`), combined optionally with an industry standard SQL server such as Oracle. However, these are too expensive (involving costs of the same order of magnitude as the servers on which they run) to be a default option outside the drug industry.

Many statisticians manage data in their statistics package of choice, be it (at increasing cost) EpiInfo, Stata, or SPSS (*see* **Software, Biostatistical**). Many health professionals keep data in spreadsheets like Microsoft's Excel (*see* **Spreadsheet**), but care needs to be taken to separate formatting and data.

In buying database software, the first consideration relates to the capabilities required of the database. Is a fully implemented DBMS (such as Oracle) necessary, or is a personal database (like Access or FoxPro) sufficient? Practical issues that suggest a fully implemented DBMS are:

1. More than one person needs to see the same data at the same time.

2. You have a large data set. The limiting factor is not the size of the file, but how quickly you can access it.
3. You have a complicated data set – some smaller databases are unable to make links on more than one variable, or links back to the same table.
4. You are storing unusual types of data (not text or numbers).
5. Your data must be securely managed, protecting confidential data through passworded access, and with a robust backup regime (*see* **Confidentiality and Computers**).

Of these, confidentiality and multiuser access are the two main reasons to centralize data, though limiting factors are that you need networked computers with a server (which rules out laptops, unless you enjoy configuring modems); financial resources; and that there is someone available to run a larger installation – the DBA. The aim is to use appropriate technology. In practice, we tend to differentiate small databases that require daily access by one person (like the administrative databases that are used to recruit and monitor study subjects) and larger multiuser databases that store study data sets.

Once the DBMS level has been decided, questions that help to differentiate products include:

1. How easily can I get data in and out? It should not require programming to upload existing data sets or download data for analysis!
2. Does it "talk to" my statistics package? Often part of the "data dictionary" (variable names and labels, missing value codes, groupings) is lost in transit.
3. What sort of access am I after? Most databases are optimized for quick look-ups and data entry ("transaction processing"), but sweeps through the data can be very slow.
4. Can it handle the throughput of data I have in mind? Uploading tables can be time-consuming.
5. Can I and do I have to program, or can I use a mouse-driven interface to create and manage data? A good package will allow both styles of access.
6. Do I need a new computer? Some laboratory databases are supplied on a "free" computer, usually a PC or Macintosh. For data work in the field a notebook might be necessary. And, a big data set might require more memory, or a faster disk.
7. How much will it really cost? Pricing policy is not always clear.
8. Most important of all, can I get help?

Choosing Between SQL Servers

DBMS Magazine publishes comparative reviews of the leading server-sized DBMSs periodically (see `http://www.dbms.mfi.com/9611d52.html`). Criteria used last time (November 1996) were adherence to the relational model and SQL standards, ability to store nonalphanumeric data (binary large objects or BLOBs), and availability of communication extras. The following "big six" industry leaders were compared: Oracle 7; Sybase; Informix-online; Microsoft SQL Server; IBM's DB2; and CA-OpenIngres.

Database Design

Usually we cannot "optimize" our data structures – they reflect a questionnaire or some other pre-given study data – but some effort is needed to "normalize" data. This is the process of squeezing repeating data out of the main data sets and into look-up tables and subtables. A simple example of a look-up would be giving doctors codes in a hospital database rather than repeating their names endlessly. More problematic is repeating subsets of data – so, in coding drugs, do we allot space for a fixed number (subscripting the variables drug 1 to drug n), or do we start a separate table for drugs, one line per patient per drug? Good database design practice suggests you do the latter, though this "entity-relationship modeling" suffers from diminishing returns. Fortunately products like Logic Work's ERWin data modeling toolset exist to help you – at a price!

A modern DBMS is often described by a task it might be good for: data warehousing, data mining, on-line analytical processing (OLAP), decision support, and management information systems (MIS). All of these mean that data are indexed and possibly stored for maximum efficiency of collation, not (as is the usual case) for speed of individual row selection ("transaction processing"). The jargon obscures a useful tip: that maybe we can design our databases to make analysis easy. In contrast to

the normalized model, the design of a data warehouse is one big table per topic, with only one level of look-ups around it. This star-shaped model (or "multidimensional" database) is far easier to analyze.

Database Interfaces

Many fourth-generation language (4GL) tools are available to build graphical interfaces to databases – newer, personal databases like Access and FileMaker include their own drag-and-drop "Form Wizards", but more scalable products include Powersoft's PowerBuilder, Oracle's PowerObjects, or Borland's Delphi for data-entry design, and Seagate's Crystal Reports for report generation. These are useful if you need a slick presentation for inexperienced computer users. Making pleasing interfaces is very time-consuming, and, as data entry is usually quicker without a mouse, these tools (if bought separately) are often a luxury.

Recent Developments

Relational databases are a mature technology in computer science, having remained remarkably stable over the last decade. Improvements have not changed the fundamental relational database management system (RDBMS) framework. Accordingly, the previous article (from the previous edition of this encyclopaedia) stands as a useful reference. This addendum notes changes in RDBMS usage, and describes the recent development of object-oriented techniques and multimedia databases.

Increments

Improvements to RDBMS have occurred in speed, storage capacity, network capability, and programming language interfaces. General changes in computing infrastructure, such as the continued growth of the internet, faster networks, faster processors, cheaper memory, and cheaper storage have encouraged these changes. Formerly, specialized areas such as data mining and distributed databases have become more widely available.

The open source movement has made this formerly expensive software free to all. The major free RDBMS are MySQL [10] and PostgreSQL [11]. Not free, but still a low cost option for many, is the Microsoft Access database, supplied as a standard component of Office Professional Edition [12]. Databases now form the back end to a multitude of websites, from the smallest personal sites to major online retailers such as Amazon.com [13].

Associated with the increase in usage, means of connecting to databases have also greatly multiplied. Middleware, which inserts a *database driver* between database and application, offers a wide variety of programmer interfaces, pre-built reporting and analysis packages, and graphical interfaces. These all rely on a standard low-level programmers' interface to the database system, such as Microsoft's ODBC [14] (Open Database Connectivity) and Sun's JDBC [15] (Java Database Connectivity). There is a plethora of middleware available for developers, supporting database access from many platforms, languages and across many network protocols; no major development project can afford to be without it [3, 20].

New Developments

Major new database techniques have arisen from use of the object-oriented programming paradigm [1]. In object-oriented programming, the focus is on considering a piece of information as a whole. For example, a "patient" can be considered as having attributes such as a name, address, condition, X-ray images, CAT scan images, and drug treatment history – all of which are stored together. There can also be subtypes of patient, where each type "inherits" the basic structure from the standard patient record, then adds specific information relevant to the group. Thus an "oncology patient" can have very different records to a "maternity patient", while keeping the basic patient data in common. This contrasts with the normal relational model, where information about a patient is scattered across many separate tables, and a set of complex Structured Query Language (SQL) queries may be required to collect everything known.

While this sounds admirable, it also has disadvantages [7]. The separation of data in the relational model is done to prevent repetition, and is an essential part of preserving data integrity. This must be handled in some other way in an object design. Also, the types of objects potentially in use are so varied that specialization remains necessary. There is thus no fully standard way of querying object databases;

each vendor supplies a different query language. A consortium of object database developers, the Object Data Management Group (ODMG) [24] did propose a standard (OQL) [2] in the early nineties, which was not much implemented.

OQL did have a major impact on the design of the 1999 release of ANSI SQL 3 ("SQL99") [16]. SQL99 is not a fully object-oriented query language; it is instead a superset of SQL92, with extensions for generic object structures. Any database implementing this is thus able to function as a standard RDBMS that uses the older SQL92 conventions.

Since SQL99, major vendors have supported steadily more object features in a more standard way. Even so, pure object-oriented database management systems (OODBMS) still tend to go their own way. Pure object-oriented database systems also remain rare, although they have found some success in niche markets. Object relational database management systems (ORDBMS), which implement most of SQL99, are being provided by major vendors including Oracle and IBM, and more recently Microsoft. These form the main stream of current database technology [19].

The ORDBMS is still a fully functional relational database, but has object capabilities added. Naturally, these additional capabilities vary across products. For example, all ORDBMS databases have internal table structures allowing inheritance, but many RDBMS have minor extensions such as support for storing data types other than text and numbers in table columns, with the minimal addition being the binary large object (BLOB) type. Major vendors mostly provide an object layer, in which the database designer specifies the object structure as it maps to the table structure. This relieves SQL users of the programming complexity needed to retrieve data, but hands that complexity over to the database designer. Normal SQL users can query the objects using a simple syntax, without needing to know the details of how the objects are stored.

An important feature of ORDBMS is its support of the notion of abstract data type (ADT). This feature allows database designers to specify objects with associated methods (functions). The database knows only the names of the methods, and that these take input and produce output of specific types. For example, a "photo" data type might have a method to return size in bytes, as an integer; another method to return file type, as a text string; a function that tells whether the picture is of a naked human being, as a Boolean; and another Boolean function that tells whether the picture is of a horse. The first two of these are simple; the latter two either rely on a human evaluation that is stored with the photo file, or on a complex function based on the colors and shapes in the image – indeed, a very interesting problem that no one has fully solved [9]. And yet the user can input queries without knowledge of any of these details.

Multimedia database systems currently tend to be ORDBMS, containing files with manually entered textual descriptions of the object contents. Multimedia systems that are purpose built for film and television may include a bundle of viewers, media files, and relational tables describing the data, together with ADT definitions specifying how data may be viewed and how it must be synchronized. While the common understanding of multimedia limits the binary files to image, audio or video, there is no real restriction. The same problems of storage and search of large sets of large binary data files arise whether one wishes to search television broadcast archives, fingerprints, MRI images, or X-rays [18].

While the size of the data can impose scalability problems for storage and retrieval times, the more difficult questions arise in identifying and searching items that are in some sense similar. Similar numerical and textual values can easily be retrieved with an SQL query, by approximate text matching or numeric ranges. But what counts as "similar" for an image, sound, or video? Classification of data by computational means continues to be a difficult problem, and often highly specific to a particular field. For example, the program that tells whether an image contains a naked human will not be able to tell if it contains a horse, nor whether the image is scanned from a photograph or a Rembrandt oil painting. While support for querying on spatial and temporal relations is currently being addressed with the development of a further SQL extension, SQL/MM [17], there are many problems that remain the subjects of active research.

Resources

Excellent current information is available from the World Wide Web. Some sites with good links, besides DBMS Magazine Online, include: Database Systems Laboratory, University of Massachusetts (`http://www-ccs.cs.umass.edu/db.html`)

which has an extensive collection of links to research and development sites (by specialty), journals, textbooks and manuals, newsgroups, and conferences, as well as to general sites and vendors' details; University of California at Berkeley DBMS Research Group, (`http://s2k-ftp.cs.berkeley.edu:8000/postgres/otherdbms.html`) which has links to vendors and standards (this group authored Postgres, the best server-sized free database); the Association for Computing Machinery's Special Interest Group on Management of Data Information Server, or ACM SIGMOD (`http://bunny.cs.uiuc.edu/`), with links to free software summaries; and Cetus Links: Object-Orientation/Databases (`http://www.rhein-neckar.de/~cetus/software.html`), for everything to do with object-oriented databases (the recent spate of "universal" databases are mixed object/relational databases).

Principal vendors (see the "big six" list above) have Web pages. For personal/desktop databases, the choice is large, but the best known are: Access, from Microsoft; FileMaker Pro, from Claris Software; Visual FoxPro, from Microsoft; and Visual dBase, from Borland International. All of these have Web pages. A "Catalog of Free Database Systems" can be found at `http://cuiwww.unige.ch/~scg/FreeDB/`.

References

[1] Booch, G. (1990). *Object-Oriented Design with Applications*, Benjamin/Cummings.

[2] Cattell, R.G.G. eds. (1994). *The Object Database Standard: ODMG-93*. Morgan Kaufmann Publ., (revised as ODMG 3.0, 2000).

[3] Charles, J. (1999). Middleware moves to the forefront, *IEEE Computer* **32**(5), 17–19.

[4] Codd, E.F. (1970). A relational model of data for large shared data banks, *Communications of the Association for Computing Machinery* **13**, 377–387.

[5] Codd, E.F. (1990). *The Relational Model for Database Management: Version 2*. Addison-Wesley, Reading.

[6] Date, C.J. (1986). *An Introduction to Database Systems*, Vol. 1, 4th Ed. Addison-Wesley, Reading.

[7] Date, C.J. & Darwen, H. (2000). *Foundation for Future Database Systems: The Third Manifesto*, 2nd Ed. Addison-Wesley.

[8] Elmasri, R. & Navathe, S.B. (1994). *Fundamentals of Database Systems*, 2nd Ed. Addison-Wesley, Reading.

[9] Forsyth, D.A. & Fleck, M.M. (1999). Automatic detection of human nudes, *International Journal of Computer Vision* **32**(1), 63–77.

[10] `http://www.mysql.com/`.

[11] `http://www.postgresql.org/`.

[12] `http://www.microsoft.com/office/access/default.asp`.

[13] `http://www.amazon.com`.

[14] `http://msdn.microsoft.com/library/default.asp?url=/library/en-us/odbc/htm/odp1.asp`.

[15] `http://java.sun.com/products/jdbc/`.

[16] International Organization for Standardization, Information Technology-*Database Language SQL*, Standard No. ISO/IEC 9075:1999, 1999.

[17] International Organization for Standardization, ISO/IEC Committee Draft, *SQL Multimedia and Application Packages*, ISO/IEC 13249-2:2002(E).

[18] Michael, M.D. Multimedia Database: Through the Looking Glass, Database Programming & Design, May 1997.

[19] Niccolai, J.G. IBM Steals Database Crown From Oracle, IDG News Service, May 07, 2002.

[20] Ritter, D. The Middleware Muddle, DBMS May 1998.

[21] Ullman, J.D. (1988). *Principles of Database and Knowledge-Based Systems*, Vol. 1. Freeman, New York.

[22] US Food and Drug Administration (1988). *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications*. US Department of Health and Human Services, Public Health Service, Food and Drug Administration, Washington, Appendix B.

[23] Westlake, A. (1993). Introduction, in *Relational DataBases*, A. Westlake, ed. Study Group on Computers in Survey Analysis, London, pp. 1–6.

[24] `www.odmg.org`.

N. WALKER & CATHERINE LAWRENCE

de Finetti, Bruno

Born: June 13, 1906, in Innsbruck, Austria.
Died: July 20, 1985, in Rome, Italy.

Although born in Austria, where his father was designing and building a railroad, de Finetti was Italian. He attended Milan University, where his work came to the attention of Corrado Gini in Rome, whom he briefly joined. He became an actuary in Trieste and held academic posts there and in Padova. Finally, he returned to Rome as a professor. Although he is mainly recognized today as a leading **probability theorist**, he was much concerned with practical applications and regarded probability as an essential tool in the business of life. He believed in an economic system that combined Pareto optimality with a notion of equity, far removed from the notion of greed that, in his view, permeated capitalism.

For de Finetti, the sole interpretation of probability was a number describing the belief of a person, conveniently referred to as "you", in the truth of a proposition. He coined the aphorism, "Probability does not exist"; meaning that it has no reality outside the individual's perception of the world. Probability describes a relationship between you and that world and is not solely of that world, as others contend. For any uncertain quantity X, he introduced the prevision $P(X)$, a number you use to replace X, in the sense that you would engage in *any* transaction that would yield you $s[X - P(X)]$, for sufficiently small s. For a proposition, $X = 1$ (true) or 0 (false), $P(X)$ is your probability that the proposition is true. The rules of probability easily follow in some beautiful, simple mathematics. Generally, prevision plays the role ordinarily occupied by expectation. Previsions that do not lose you money for sure, in a set of the above transactions, are said to be coherent. He established a basic theorem that shows how a set of previsions $\{P(X_i) : i = 1, 2 \ldots n\}$ impose constraints on a further, coherent prevision $P(X_{n+1})$. In this view, the scientific method results in a set of coherent judgments about the world.

In another theorem he showed how this personalistic approach included frequentist views of probability as a special case. If your previsions for a set of 0–1 quantities obey a condition he called **exchangeability**, then the theorem shows that the frequency of quantities that are one will tend to a limit, about which you have a prevision. Since most situations studied in modern statistics incorporate some form of exchangeability, they can be included in the personalistic view. The procedures it recommends are often different from those adopted by frequentists. His approach was Bayesian (*see* **Bayesian Methods**), but he differed from many Bayesians, like **Harold Jeffreys**, in refusing to admit that any value of a probability was more rational than another. You could coherently have one view, I another, and both be rational. Uncertainty is only removed, and agreement reached, by the accumulation of data.

He wrote extensively on the teaching of probability. He held that children should study uncertainty from an early age. In particular, they should be encouraged not to answer "yes" or "no" to a question, when in reality they were uncertain about it, but respond with a probability. He developed scoring rules to assess abilities from such responses. These have found use in medical diagnosis.

De Finetti's views are so original, and his style so parenthetical, that his writings are difficult to understand, even when he writes about statistical issues. His mathematics is always simple and he abhorred much technical writing. He does not use the term **random variable** for X above. There is nothing variable about it; it is a fixed quantity the value of which is uncertain for you. Those who have the patience to follow his writings find them immensely rewarding and correct. His first, important papers date from the early 1930s, but most people, especially those who do not read Italian, will find the most accessible approach to his work in the books from the 1970s [1, 2]. Of all statisticians of this century, **Fisher** and de Finetti are the true geniuses.

References

[1] De Finetti, B. (1972). *Probability, Induction and Statistics: The Art of Guessing*. Wiley, London.

[2] De Finetti, B. (1974/5). *Theory of Probability: A Clinical Introductory Treatment*, 2 Vols. Wiley, London (translation from the Italian).

DENNIS V. LINDLEY

de Moivre, Abraham

Born: May 26, 1667, in Vitry, France.
Died: November 27, 1754, in London.

Abraham De Moivre is celebrated by statisticians primarily for his derivation of the normal approximation to the **binomial distribution**, and for providing the first tabulation of the normal integral (*see* **Normal Distribution**).

He studied the humanities and mathematics at the Protestant University of Sedan, and later at the University of Saumur and the Sorbonne. After the repeal in 1685 of the Edict of Nantes, he suffered imprisonment, and to avoid further persecution as a Protestant he emigrated to London in 1688. There he worked as an itinerant tutor, and adviser to gamblers and brokers, and was elected to the Royal Society in 1697. His famous work, *The Doctrine of Chances: or a Method of Calculating the Probability of Events in Play*, appeared in successive editions in 1718, 1738, and 1756. His result on the normal approximation to the binomial appeared first in a separate paper in 1733, and used what is now known as Stirling's formula. He seems to have regarded the normal curve merely as a means of approximation, rather than as a distribution in its own right.

De Moivre published also a standard work on annuity theory, *Annuities upon Lives*, in 1718 and 1743 (*see* **Actuarial Methods**). As noted in [1], De Moivre's scope was limited; he excluded from consideration all the forms of insurance then practiced in London (fire, life, and maritime).

For fuller descriptions of De Moivre's life and works, see [2] and [3].

References

[1] Daston, L.J. (1987). The domestication of risk: mathematical probability and insurance 1650–1830, in *The Probabilistic Revolution*. Vol. 1. *Ideas in History*, L. Krüger, L.J. Daston & M. Heidelberger, eds. MIT Press, Cambridge, Mass., pp. 237–260.

[2] Seneta, E. (1982). De Moivre, Abraham, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.I. Johnson, eds. Wiley, New York, pp. 300–302.

[3] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press, Cambridge, Mass.

PETER ARMITAGE

Death Certification

The processes through which a civil registration and **vital statistics** system is informed of the facts about each death occurring within its coverage area may involve information supplied by several different informants. Relatives or friends may supply personal particulars of the deceased either directly to the civil registration authorities or indirectly through a funeral director or other intermediary who then gives the information to a civil registrar. However, the legal determination of the fact of death and the statement of the medical **causes of death** are usually the responsibility of an attending physician. In the absence of an attending physician or in the case of death resulting from actual or suspected violence (e.g. accident, suicide, homicide), a medical/legal officer usually investigates to determine the medical and legal facts of the case. In many civil registration systems, the medical/legal officer is known as a "coroner" or "medical examiner" whose specific responsibilities are prescribed by law. The physician or medical/legal officer is required to certify, to the best of his or her knowledge, that the death took place at the time and place specified and was due to the causes recorded on the death certificate. In some countries a shortage of trained medical personnel makes this process impossible or impractical to carry out, but in those vital statistics systems where deaths are attended just prior to death or reviewed after death by qualified medical practitioners, the **World Health Organization (WHO)** recommends a specific format and procedure for the certification of cause of death. The WHO recommendations are based on the concept that for each death occurring, one and only one "underlying cause of death" is to be determined, counted, and statistically analyzed (*see* **Cause of Death, Underlying and Multiple**). WHO defines the underlying cause of death as "(a) the disease or injury which initiated the train of morbid events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury". WHO further recommends the use of the International Form of Medical Certificate of Cause of Death (Figure 1), which is designed to facilitate the selection of the

Cause of death		Approximate interval between onset and death
I Disease or condition directly leading to death*	(a)	. . . . .
	due to (or as a consequence of)	
Antecedent causes Morbid conditions, if any, giving rise to the above cause, stating the underlying condition last	(b)	. . . . .
	due to (or as a consequence of)	
	(c)	. . . . .
	due to (or as a consequence of)	
	(d)	. . . . .
II Other significant conditions contributing to the death, but not related to the disease or condition causing it	. . . . .	. . . . .
	. . . . .	. . . . .

**This does not mean the mode of dying, e.g. heart failure, respiratory failure. It means the disease, injury, or complication that caused death.*

Figure 1 International form of medical certificate of cause of death

underlying cause of death based on the sequence of morbid events reported by the medical practitioner [1].

The medical certificate shown in Figure 1 provides a uniform format for the medical practitioner signing the death certificate to indicate which condition led directly to death and to report any antecedent conditions which gave rise to that condition. The medical certificate is designed to facilitate the selection of the underlying cause of death when two or more conditions are recorded. If only one condition is reported by the certifier, this single condition should be recorded on line (a) of Part I of the Medical Certificate of Cause of Death and is considered as the "originating antecedent cause". If there is more than one condition involved in the train of events leading to death, the direct cause is entered on line (a) and antecedent causes are entered on lines (b), and, if needed, (c) and (d). The lowest used line reflects the originating cause, and the causes entered on the other lines reflect, in sequence, the train of events leading to death. Therefore, in a properly completed Medical Certificate, the lowest used line in Part I is considered to be the originating antecedent cause. Usually, the originating antecedent cause corresponds to the underlying cause of death, the condition used for statistical tabulation and analysis. However, in some circumstances the originating antecedent cause may be superseded by a condition more suitable for use as the underlying cause of death. In cases where the certificate does not appear to be properly filled out (e.g. the reported sequence does not make medical sense, or the originating cause is a vague or nonspecific condition and there are other more specific conditions reported elsewhere on the certificate), there is a set of international rules promulgated by WHO for selecting an underlying cause of death. The selected underlying cause may then be modified by additional rules to make it more useful for statistical and epidemiologic purposes. These rules are particularly useful when it is impractical or impossible to query the medical practitioner who completed the certificate in order to obtain clarification or a more definitive description of the conditions leading to the death. While the rules may, in individual cases, appear to be arbitrary, they are intended to yield improved mortality statistics overall.

When determining the underlying cause of death from conditions recorded on a medical certificate, the international rules and guidelines first call for the application of the General Principle, which states, "... when more than one condition is entered on the certificate, the condition entered alone on the lowest used line of Part I should be selected only if it could have given rise to all of the conditions entered above it". If the General Principle does not apply, there are three Selection Rules that are to be applied sequentially until an originating antecedent cause is identified. However, that originating cause may not be the most useful and informative condition for statistical tabulation and analysis. For example, if senility or a generalized disease such as hypertension has been selected, this is less useful than if a reported manifestation of the aging process or of the hypertension had been chosen. Further, it might be necessary to modify the selected condition to conform with the requirements of the **International Classification of Diseases (ICD)**, either because a single code in the classification might represent two or more conditions that were both reported, or because the classification may give priority to a particular cause when it is reported with certain other conditions. Accordingly, there are six Modification Rules intended to improve the utility of mortality data. The Modification Rules are applied after the selection of the originating antecedent condition. Some Modification Rules require further application of the Selection Rules, resulting in an iterative process of selection, modification, and, if necessary, reselection before an underlying cause of death is determined [2].

In the case of perinatal deaths, WHO recommends a special Certificate of Cause of Perinatal Death. This certificate provides a section for the medical certifier to list diseases or conditions in the fetus or infant as well as maternal diseases or conditions which affected the fetus or infant. In the 10th Revision of the International Classification of Diseases (ICD-10), the perinatal period is defined as the period beginning at 22 completed weeks of gestation and ending at seven completed days after birth. WHO provides a special set of rules for the certification of deaths occurring during this period [3].

References

[1] World Health Organization (1967). *WHO Nomenclature Regulations*. World Health Organization, Geneva.

[2] World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems*, 10th rev., Vol. 2. World Health Organization, Geneva, pp. 30–88.

[3] World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems*, 10th rev., Vol. 2. World Health Organization, Geneva, pp. 89–96.

Death Indexes

An important aspect of follow-up studies (*see* **Cohort Study**) is the ability to determine which members of the original study cohort have been lost to follow-up because of death, and for many such studies, in addition to the fact of death, the **cause of death** is an essential piece of information. In places where a central register of all deaths is compiled (e.g. population register, or civil registration system) and is available for research use, the identification of individuals in a study who have died during some time period can be accomplished, provided that the necessary identifying information is available. However, even under relatively ideal circumstances, it is sometimes difficult to match study individuals against lists of deaths with 100% certainty that a correct match has been achieved. Most follow-up studies try to match on several variables (e.g. surname, given name, date of birth, mother's maiden name, etc.) and develop algorithms to establish "presumptive matches" (*see* **Matching, Probabilistic**).

While the fact of death is, in most jurisdictions, considered public information, the cause of death may in some places be considered confidential and not releasable to researchers without the expressed permission of a next of kin or legal representative of the deceased. Because of **confidentiality** provisions, some custodians of death files require study protocols to be reviewed for adequate privacy safeguards before authorizing the release of cause of death data.

In a few countries, the problem of adequate follow-up for deaths occurring amongst a cohort is further complicated by the existence of only decentralized death files. A notable example of this is the US where the primary responsibility for registration of vital events rests with the individual states. For many years it was necessary for "death clearance" of study cohorts for researchers to send their list of study participants to each of the more than 50 registration areas to determine if any of their subjects had died there during some stated period of time. This was a costly and time-consuming process and tended to stifle certain kinds of epidemiologic research. In addition, each state has its own laws and procedures regarding confidentiality and release of information. Therefore, in spite of the fact that there was a central statistical file for national **vital statistics** purposes located at the US **National Center for Health Statistics**, the states provided their data with the restriction that the Center not release individual record information without the consent of the states. After lengthy negotiations, an agreement between the states and the National Center for Health Statistics resulted in a US National Death Index which was designed to address these issues. To utilize this index, researchers submit their study protocol to a committee comprising selected state registration officials, federal officials, and representatives of the health research community. If this committee determines that the proposed study is bona fide research and not a commercial activity, and includes appropriate steps to protect confidential information and privacy, the protocol is approved and sent to the Director, National Center for Health Statistics, for final approval. Once the study has been approved, the National Center receives annual lists of study participants from researchers containing the required variables for computer matching against the statistical file of deaths (*see* **Record Linkage**). For each "presumptive match", the researcher receives information about the date, place of death, and death certificate registration number along with some details of the degree of agreement between the required variables. If the study requires cause of death information or other information from the death certificates, the researcher receives enough information to contact the appropriate State Registration Officials to request copies of the pertinent death certificates.

The US National Death Index has been in operation since data year 1979 and has significantly improved follow-up procedures for studies conducted in the US. It has reduced the necessary time and costs of efficient identification of deaths occurring in national follow-up study cohorts.

ROBERT A. ISRAEL

Decision Analysis in Diagnosis and Treatment Choice

Decision analysis is a quantitative method for identifying the optimal course of action among a well-defined set of alternatives under conditions of uncertainty [18] (*see* **Decision Theory**). The optimal course of action is defined as the one that maximizes (or minimizes) the expected value (*see* **Expectation**) of the outcome of interest. The application of decision-analytic methods to medicine is appealing because uncertainty is inherent in diagnosis and treatment choice. Consider the clinical setting where a physician must decide how to care for a patient when the true underlying disease state is rarely known with certainty. Although a particular disease will either be present or absent in any given patient, the physician must make decisions about use of diagnostic tests and treatments based on a subjective assessment of the underlying disease state (*see* **Computer-aided Diagnosis**). In this setting, decision analysis can identify the clinical approach that maximizes average survival, **life expectancy** or quality-adjusted life years (*see* **Quality of Life and Health Status**). More importantly, decision analysis can be used to highlight the critical factors in making decisions about patient care.

The application of decision analysis to clinical medicine was introduced more than 30 years ago [11–13]. Since then, its use in clinical medicine has grown [9, 10, 21], and the role of decision analysis in the economic evaluation of medical practices has also been firmly established [8, 24].

By identifying clinical approaches that maximize health outcomes, decision analysis can be used for guiding the care of individual patients or groups of patients. In the latter context, decision analysis can be useful for clinical guideline development and for informing health policy decision makers. When costs are considered as an end point, decision analysis can be used to identify the least costly course of action or for cost-effectiveness evaluation (*see* **Health Economics**).

The basic steps in decision analysis entail defining the decision problem, structuring the decision tree, assigning parameter values for both probabilities and outcome values to the tree, and analyzing the tree (Table 1). In the sections that follow, each step is described further in the context of a classic clinical decision problem involving a choice between no intervention, immediate testing, and immediate treatment in a patient suspected of having only one possible underlying disease. To demonstrate quantitative aspects of decision analysis, a more detailed clinical example is introduced in the section on analyzing the tree.

Table 1 Basic steps in decision analysis

1. Define the problem
   (a) Identify the decision maker and the objective.
   (b) Specify alternative actions and consequences.
2. Structure the decision tree
   (a) Represent alternative actions as branches emanating from a decision node.
   (b) Represent the temporal sequence of chance events, actions, and outcomes as subsequent chance nodes, decision nodes, and terminal nodes, respectively.
3. Assign values for each parameter in the decision tree
   (a) Assign appropriate probabilities based on their position in the tree.
   (b) Assign appropriate outcome value(s) at each terminal node.
4. Analyze the decision tree
   (a) Average-out and foldback the decision tree.
   (b) Conduct extensive sensitivity analyses.

Define the Problem

Identify Decision Maker(s) and the Objective

The first step in defining the decision problem is to identify the decision maker and to state clearly the decision maker's objective. This entails specifying the outcome of interest and whether it is to be maximized or minimized.

In the classic clinical decision problem, we assume that the physician is the decision maker and that his objective is to maximize survival. If the patient also wishes to maximize survival, then the physician and patient have a shared perspective. If the patient wishes to maximize a different end point, such as time without pain, then the result of the decision analysis may differ when done from the patient's perspective.

Specify Alternative Actions

Once the objective is defined, a complete set of alternative actions must be specified. The possible consequences and temporal sequence of each action must also be delineated.

Using the classic clinical decision problem as the paradigm, the alternative actions are no intervention, immediate testing, or immediate treatment [16]. For the testing alternative, a positive or negative test result is observed and a subsequent treatment decision must be made. Regardless of the actions taken, the end points of death or survival ultimately ensue. The probability of survival, however, varies according to the underlying disease state and the action that is taken.

Structure the Decision Tree

Decision trees are the basic structure underlying most applications of decision analyses in medicine. Other approaches to decision analysis, including the use of influence diagrams, are not discussed here [15]. Decision trees depict the temporal sequence of actions and consequences and are structured from left to right. They comprise nodes, the point from which branches emanate, and branches. Two node types – decision nodes and chance nodes – are always included in decision trees. Decision nodes are depicted by squares and indicate that a choice must be made. Branches emanating from the decision node specify the alternative actions being evaluated. Chance nodes are depicted by circles and indicate that one of several chance events or outcomes may occur. Branches emanating from chance nodes must represent the entire universe of events or outcomes being considered.

Consider the decision tree for the classic clinical decision problem introduced earlier (Figure 1). The tree begins with a decision node from which the alternative actions of "No Intervention", "Test", and "Treat" emanate as separate branches. Each of these branches is followed by chance nodes. The "No intervention" and "Treatment" branches are followed by chance nodes indicating the true underlying disease state. That is, disease may be present ("Disease") or absent ("No disease"). Following each disease state node is another chance node representing the survival outcome as either "Die" or "Survive". The "Test" branch is followed by a chance node indicating that the test result may be positive ("T+") or negative ("T−"). Following the test result node are additional decision nodes that represent the decision that must be made once test results are known. Decision nodes, such as this, that occur following chance nodes, are referred to as embedded decision nodes. (Note that, when embedded decision nodes are eliminated from decision trees and a set of actions contingent upon chance events is specified, then the tree is in strategic form. For example, if the strategy is to "Treat" following a positive test result and to "Do nothing" following a negative test result, then the decision nodes in Figure 1 following "T+" and "T−" would be removed to produce a tree in strategic form.) Regardless of test outcome and subsequent decisions, for the "Test" branch, the tree also includes the disease state and survival chance nodes. Branches at the far right of the decision tree (e.g. "Die" and "Survive" in Figure 1) are called *terminal nodes*.

Assign Parameter Values to the Tree

Assign Probabilities to the Tree

Once the decision tree has been structured, probabilities must be entered for branches emanating from each chance node. These probabilities may represent the frequency with which each chance event occurs. In the simple clinical example, the disease of interest will either be present (D+) or absent (D−) for each patient, and the frequency with which disease is present may be estimated based on observations in large populations of similar patients. Sometimes a clinical prediction rule can be used to estimate the frequency of disease [22] (*see* **Predictive Modeling of Prognosis**). Alternately, the probability that disease is present may reflect the **subjective probability** or strength of belief of the decision maker that disease is present.

Recall that the branches emanating from chance nodes represent the universe of possible (or modeled) events. Thus, by the summation principle of probabilities, regardless of the type of probability represented in the decision tree, the sum of probabilities for branches emanating from each chance node must equal 1.0. We designate the probability of an event as $\Pr(\cdot)$. In the hypothetical clinical example, the probability that disease is present is denoted as $\Pr(D+)$.

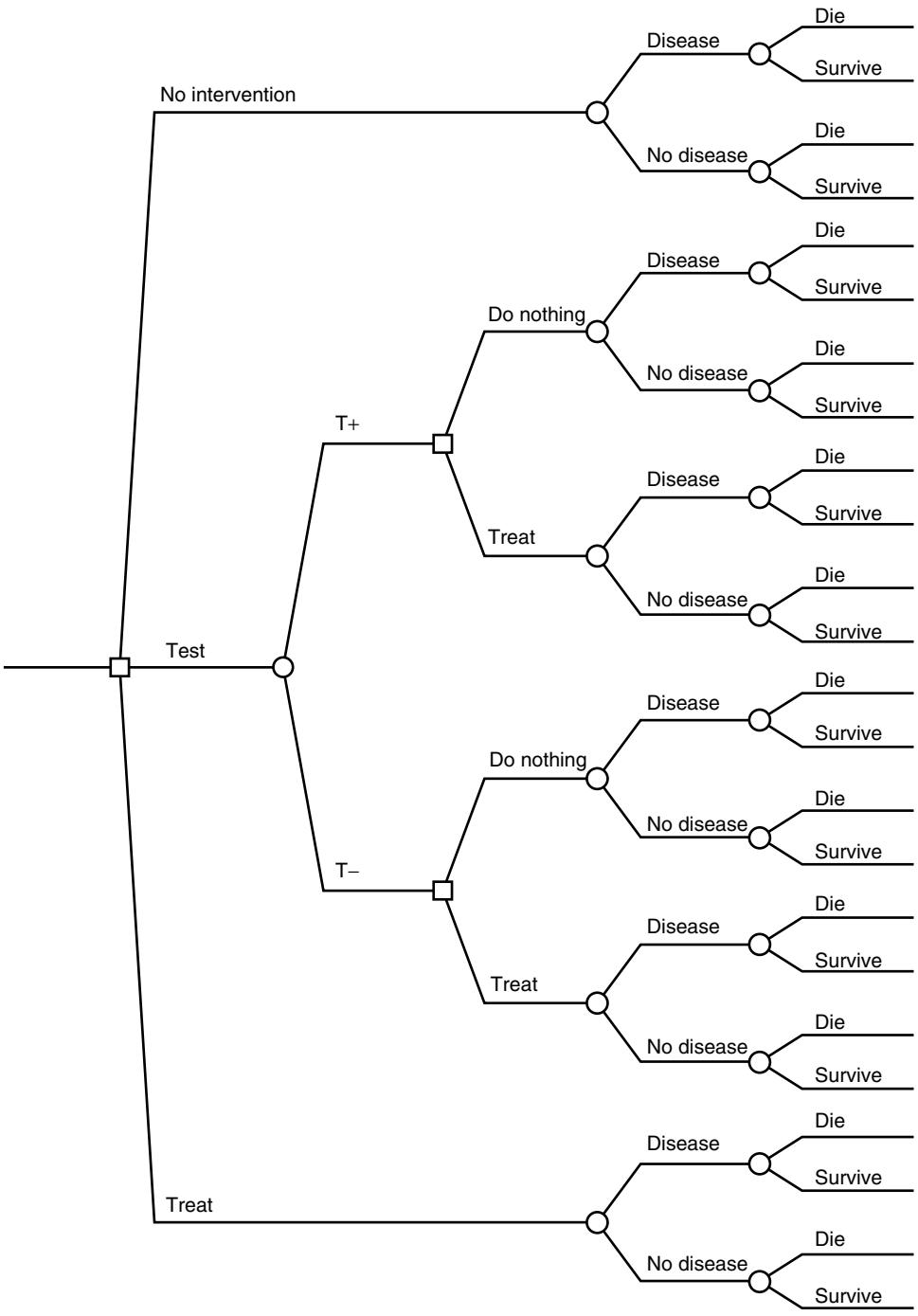

Figure 1 Decision tree for hypothetical clinical example. Decision nodes are depicted by squares (□) and chance nodes are depicted by circles (◯)

Tree Structure and Placement of Joint and Conditional Probabilities

When placing probabilities in a decision tree the correspondence between the position of each branch and the type of probability required is an important consideration. We refer to two general types of probabilities in decision trees – joint probabilities and **conditional probabilities**. Joint probabilities represent the chance that two or more events occur together. If we consider events A and B, then the probability of their joint occurrence is denoted as Pr(A,B). Conditional probabilities represent the chance that an event occurs given that another event is known to have preceded. The conditional probability that event A occurs given that event B has occurred is denoted as Pr(A|B). The relationship between joint and conditional probabilities is

$$\Pr(\mathrm{A}, \mathrm{B}) = \Pr(\mathrm{A}|\mathrm{B}) \Pr(\mathrm{B}).$$

When two events are independent, then

$$\Pr(\mathrm{A}|\mathrm{B}) = \Pr(\mathrm{A}),$$

and the joint probability of their occurrence simplifies to

$$\Pr(\mathrm{A}, \mathrm{B}) = \Pr(\mathrm{A}) \Pr(\mathrm{B}).$$

In our clinical example, the probabilities of disease following "No intervention" and "Treat" are straightforward and represent the **prevalence** or subjective opinion that disease is present, Pr(D+), among such patients. This reflects the assumption that the action taken will not affect the true disease state, though it may (and hopefully will) affect outcomes.

To demonstrate the correspondence between tree structure and use of joint and conditional probabilities, we consider a simplified portion of the "Test" branch in which "Treatment" follows a positive test and "No intervention" follows a negative test (Figure 2).

The four outcomes of test result and disease status may be modeled either as four branches emanating from the "Test" chance node [Figure 2(a)], or as a series of two **binary** chance nodes [Figure 2(b)]. Each terminal node is associated with a path through the decision tree. The path probability for each terminal node is obtained by multiplying the probabilities encountered along the path. In Figure 2, the path probabilities are shown at the end of each branch and correspond to the joint occurrence of true disease state and test result. Note that the path probabilities for each of the four possible outcomes are the same regardless of the structure chosen. What differs in Figures 2(a) and 2(b) is the type of probability used for each branch. In Figure 2(a), the joint probability of disease state and test outcome is modeled. In Figure 2(b), the conditional probability of disease represents the chance that disease occurs given the test result that was observed.

In general, the probability that a test result is positive, Pr(T+) and the complement, $\Pr(\mathrm{T}-) = 1 - \Pr(\mathrm{T}+)$ will depend on the **sensitivity**, $\Pr(\mathrm{T}+|\mathrm{D}+)$, and **specificity**, $\Pr(\mathrm{T}-|\mathrm{D}-)$, of the test and on the prevalence of disease, Pr(D+). Applying the summation principle to joint probabilities allows us to express the probability of a positive test as

$$\Pr(\mathrm{T}+) = \Pr(\mathrm{T}+, \mathrm{D}+) + \Pr(\mathrm{T}+, \mathrm{D}-).$$

The relationship between joint and conditional probabilities allows us to reexpress this as

$$\Pr(\mathrm{T}+) = \Pr(\mathrm{T}+|\mathrm{D}+) \Pr(\mathrm{D}+) + \Pr(\mathrm{T}+|\mathrm{D}-) \Pr(\mathrm{D}-).$$

Probability Revision: Bayes' Rule and Tree Inversion

The probability that disease is present following a test must be conditioned upon the test result. Thus, the post-test or posterior probabilities, $\Pr(\mathrm{D}+|\mathrm{T}+)$ and $\Pr(\mathrm{D}-|\mathrm{T}-)$, often referred to as the positive **predictive value** and negative predictive value, respectively, must be used in the decision tree as appropriate [Figure 2(b)]. Computation of these probabilities may be done using **Bayes' Theorem**, which is shown for the positive and negative predictive values as:

$$\Pr(\mathrm{D}+|\mathrm{T}+) = \frac{\Pr(\mathrm{T}+|\mathrm{D}+) \Pr(\mathrm{D}+)}{\Pr(\mathrm{T}+|\mathrm{D}+) \Pr(\mathrm{D}+) + \Pr(\mathrm{T}+|\mathrm{D}-) \Pr(\mathrm{D}-)},$$

$$\Pr(\mathrm{D}-|\mathrm{T}-) = \frac{\Pr(\mathrm{T}-|\mathrm{D}-) \Pr(\mathrm{D}-)}{\Pr(\mathrm{T}-|\mathrm{D}+) \Pr(\mathrm{D}+) + \Pr(\mathrm{T}-|\mathrm{D}-) \Pr(\mathrm{D}-)}.$$

An alternative to using Bayes' rule to compute post-test or posterior probabilities is first to structure a decision tree to accommodate the probabilities that are known. When structured with the true disease state preceding the test result [Figure 2(c)],

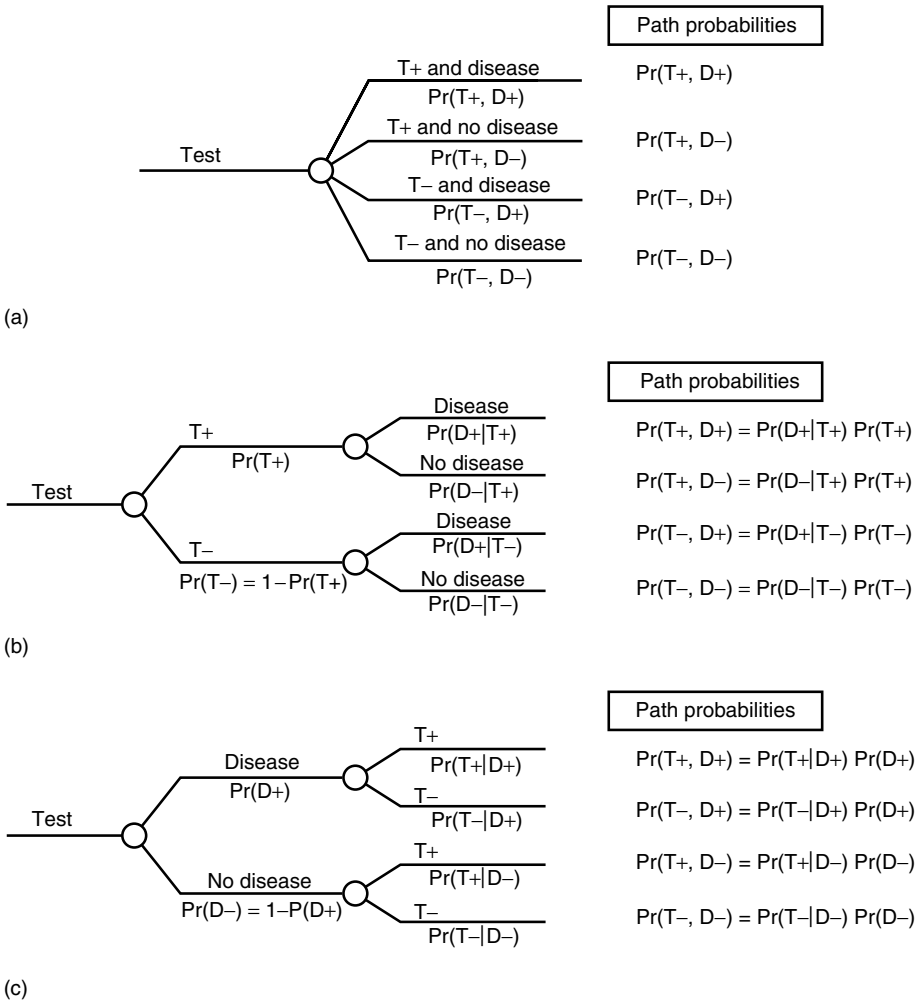

Figure 2 Modeling of simplified test branch and probabilities required: (a) joint outcomes of disease state and test result are modeled and joint probabilities are required; (b) test results and disease state are modeled sequentially and the probabilities of disease conditional on test results are required; (c) test branch modeled with disease prevalence and sensitivity and specificity to determine path probabilities

disease prevalence, test sensitivity, and specificity are entered directly as probabilities in the tree, and the path probabilities are computed. Because the overall probability of a series of chance events is independent of the order of the events, when the tree is inverted [Figure 2(b)] we know that the path probabilities are unchanged. Thus, we can solve for the unknown conditional post-test probabilities, i.e. $\Pr(D+|T+)$ and $\Pr(D-|T-)$, based on the known path probabilities in Figure 2C and known disease prevalence. This is best demonstrated with a numerical example.

Suppose that disease prevalence is 0.60, that test sensitivity is 0.55, and test specificity is 0.99. The tree structure that uses these probabilities directly [Figure 3(a)] models the true disease state as a chance node prior to the test result node. Using disease prevalence and test characteristics, the path probabilities are computed and the overall probability of a positive test is

$$\begin{aligned}\Pr(T+) &= \Pr(T+, D+) + \Pr(T+, D-) \\ &= 0.33 + 0.004 = 0.334.\end{aligned}$$

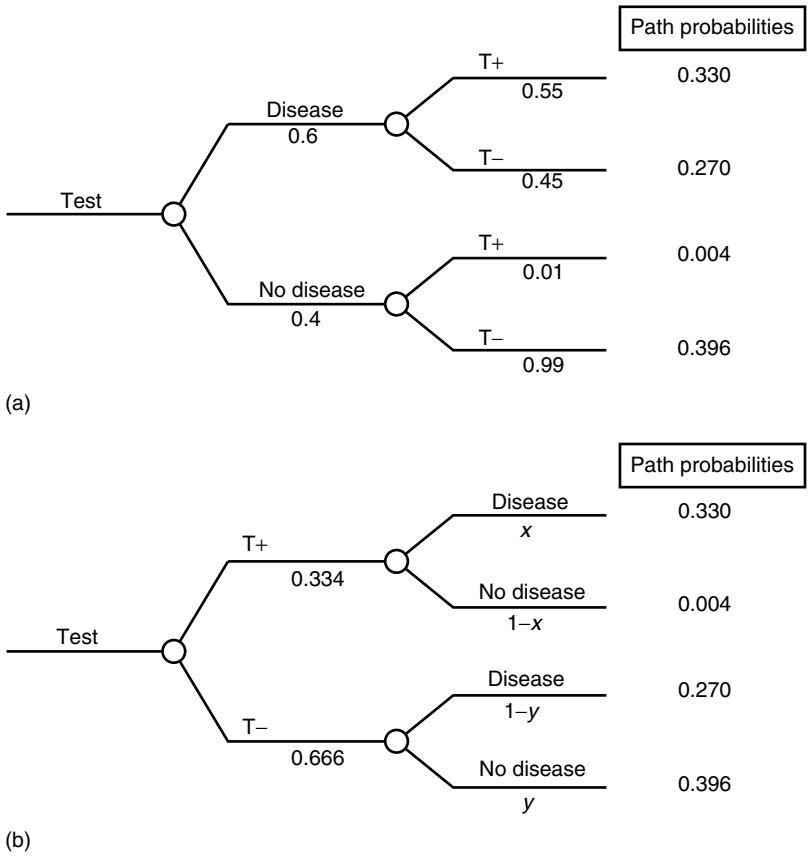

Figure 3 Example of probability revision using tree inversion: (a) test branch modeled with prevalence, sensitivity, and specificity to determine path probabilities; (b) inverted tree where unknown probabilities of disease conditional on test result are solved for based on path probabilities determined in (a)

By placing the path probabilities and probabilities of a positive test in the inverted tree [Figure 3(b)], the unknown positive and negative predictive values, denoted as x and y, respectively, are computed easily using the following equations:

$$0.334x = 0.33 \Rightarrow x = 0.988,$$

$$0.666y = 0.396 \Rightarrow y = 0.595.$$

Probability Revision with Multiple Disease and/or Multiple Test Result Categories

In many clinical applications it is necessary to model more than one disease. Both Bayes' rule and the tree inversion approach to probability revision are easily modified to reflect multiple disease categories, $D_i, i = 1, \ldots, I$. Likewise, when there are multiple test result categories, R_j, $j = 1, \ldots, J$, a general form of Bayes' rule is as follows:

$$\Pr(D_i|R_j) = \frac{\Pr(R_j|D_i)}{\sum_{i=1}^{I} \Pr(R_j|D_i)}$$

The threshold at which a test result is declared positive (T+), referred to as the positivity criterion, corresponds to a single set of test characteristics (i.e. sensitivity and specificity). Changes in the positivity criterion will alter both the sensitivity and specificity of the diagnostic test. To evaluate a diagnostic test over a range of performance, the **receiver operating characteristic (ROC) curve** [14] – a graph of the

true positive rate (sensitivity) against the false positive rate (1-specificity) as the positivity criterion is varied – is useful.

Assign Outcome Values to the Tree

Before analyzing the decision tree, at least one outcome value must be assigned to each terminal node. The relevant outcome and whether it should be maximized or minimized are usually determined when the decision problem is defined. In the classic clinical example introduced earlier, the outcomes of death and survival were modeled. In this setting, we could assign an outcome value of 0 to death and an outcome value of 1.0 to survival. The analysis of this tree would then produce estimates of expected survival.

Other common outcomes to model include life years [2, 3] and quality-adjusted life years (QALYs) [17]. When these outcomes are modeled, the analysis estimates life expectancy and quality-adjusted life expectancy, respectively. In analyses using QALYs, typically, the best possible outcome is assigned a value of 1.0 (e.g. perfect health) and the worst possible outcome (e.g. death) is assigned a value of 0. For intermediate health states formal utility assessment is undertaken to assign a value to the intermediate health states (*see* **Utility in Health Studies**).

When one alternative has probabilistic dominance over other alternatives, it is possible to avoid valuation of intermediate health states. To determine whether or not the principle of probabilistic dominance can simplify the outcome data required, we first order the outcomes from worst to best and index them with the subscript $i = 1, \ldots, I$. Let p_i represent the probability of obtaining outcome i. Then, if the following holds for all $J = 1, \ldots, (I - 1)$, alternative A is said to have probabilistic dominance over alternative B, and a decision can be made without formal valuation of the intermediate outcomes:

$$\sum_{i=1}^{J} p_i^{\text{Alternative A}} \leq \sum_{i=1}^{J} p_i^{\text{Alternative B}}.$$

For example, suppose that the alternatives "No intervention" and "Treatment" produce outcomes of death, partial paralysis, and full health with probabilities 0.18, 0.13, 0.65, and 0.2, 0.15, and 0.7, respectively. At first glance it would appear that if death is assigned a value of 0 and full health a value of 1.0, then a value would need to be obtained for the outcome "partial paralysis" before a decision could be made. Because we can rank order the outcomes from worst to best (e.g. death, partial paralysis, and full health) and because "No intervention" is preferred when death only is considered (0.18 vs. 0.2) and when partial paralysis or worse is considered (0.31 vs. 0.35), "No intervention" has probabilistic dominance over "Treatment" and it is unnecessary to assign a numeric value for the intermediate outcome of partial paralysis.

Often there is more than one end point of interest. For example, cost, life years, and quality-adjusted life years may all be relevant. By modeling each of these endpoints, the tradeoffs between increases in expected cost and increases in health can be quantified, and formal cost-effectiveness evaluation can be undertaken [8].

Analyze the Decision Tree

Clinical Example

To demonstrate the process of analyzing a decision tree, consider a hypothetical clinical scenario involving an intravenous drug-using patient who has **AIDS** and comes to the physician complaining about a persistent cough. The primary disease of concern in this setting is *pneumocystis carinii* pneumonia (PCP). Suppose that the physician must choose between no intervention, testing induced sputum (IS) for PCP using a toluidine blue stain, or immediate treatment with antibiotics for presumed PCP, and that the physician's objective is to maximize quality-adjusted life years (QALY). Assume the sensitivity of IS for PCP is 0.55 and the specificity is 0.99. Suppose that the prevalence of PCP among similar patients is 0.60 and treatment of patients with PCP will result in 1 quality-adjusted life year (QALY). Failure to treat PCP results in 0.5 QALY. Assume that no treatment in a patient without PCP results in 2 QALYs, but that treatment in a patient without PCP will result in only 1.5 QALYs. (Although this clinical example is derived from a published decision analysis [7], the QALYs used in this example are purely hypothetical.) The structure for this hypothetical clinical decision problem is shown in Figure 4.

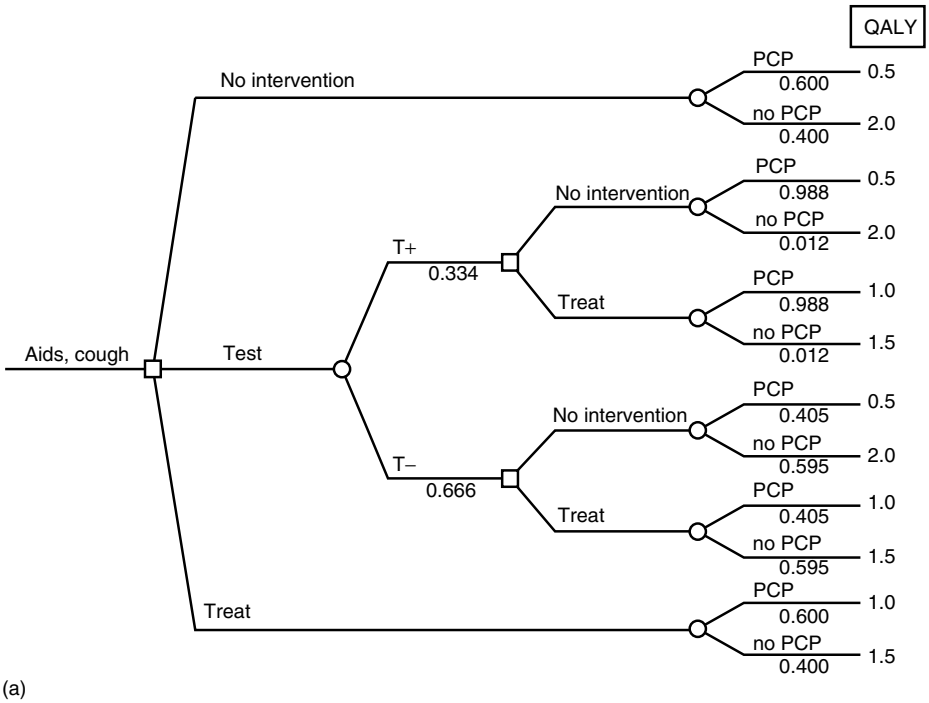

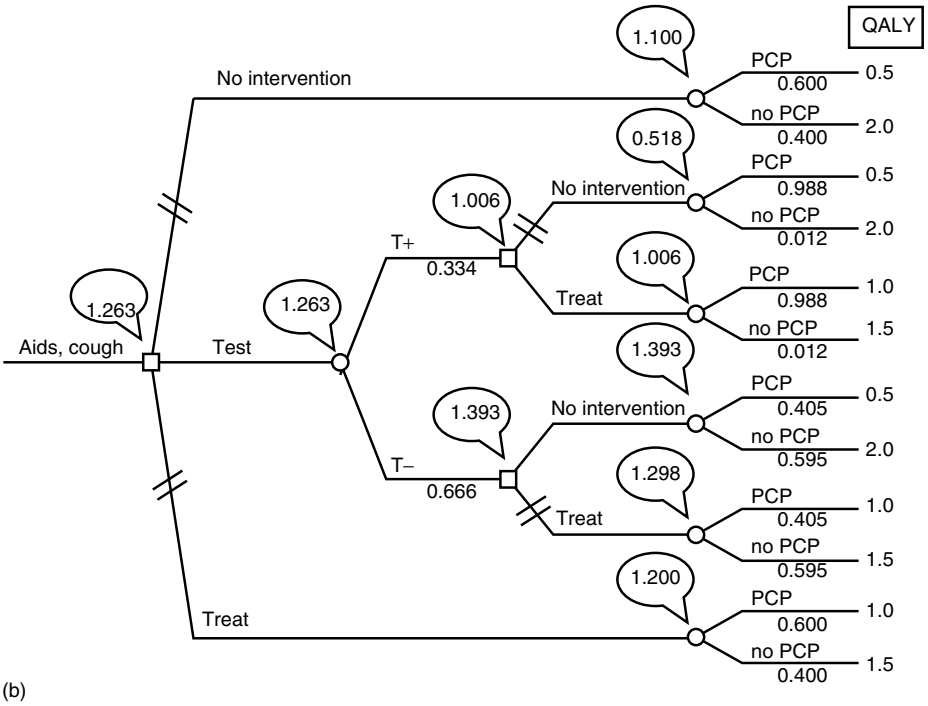

Figure 4 Decision tree and analyzed decision tree for clinical example: (a) decision tree for clinical example involving AIDS patient with persistent cough; (b) analysis of decision tree for clinical example involving AIDS patient with persistent cough

Averaging Out and Folding Back the Decision Tree: Baseline Evaluation

The decision tree is used to facilitate identifying the course of action that produces the best outcome "on average". Decision trees are evaluated by averaging out chance nodes as one moves from left to right and by folding back (i.e. eliminating) all but the branch with the best averaged-out value when a decision node is encountered.

To average out and fold back the tree in Figure 4, we begin at the far right of the tree and first average out the simple branches. The averaged-out value for the "No intervention" branch is 1.100 QALYs ($1.100 = 0.6 \times 0.5 + 0.4 \times 2.0$). The averaged-out value for the "Treat" branch is 1.200 QALYs ($1.200 = 0.6 \times 1.0 + 0.4 \times 1.5$).

For the more complex "Test" branch, the appropriate post-test probabilities are first computed. Based on the test characteristics and prevalence of PCP given above, the probability of a positive test is 0.334, the predictive value of a positive test [i.e. $\Pr(\text{PCP} + |\text{T}+)$] is 0.988, and the predictive value of a negative test [i.e. $\Pr(\text{PCP} - |\text{T}-)$] is 0.595.

Once these revised probabilities of disease are computed, we next focus on the decision following a positive test result. The averaged-out value for no intervention following a positive test result is 0.518 QALYs ($0.518 = 0.988 \times 0.5 + 0.012 \times 2.0$), and the averaged out value of treatment following a positive test result is 1.006 QALYs ($1.006 = 0.988 \times 1.0 + 0.012 \times 1.5$). Because there is a decision node here, we must fold back the tree pruning (denoted with a double slash mark) the option that is suboptimal. Here, we wish to maximize quality-adjusted life expectancy, and the option of "No intervention" following a positive test is pruned. Next, consider the decision following a negative test result. Averaging out the options of "No intervention" and "Treat" following the negative test yields expected values of 1.393 and 1.298, respectively. Thus, we fold back the tree by pruning the "No intervention" option. Now that the embedded decision nodes for the "Test" branch have been pruned, we are ready to average out the "Test" branch. The expected value of the "Test" branch is computed as $1.263 = 0.334 \times 1.006 + 0.666 \times 1.393$.

We are now ready to determine the initial action that results in the best life expectancy and compare the averaged out values for the three alternatives of "No intervention", "Test", and "Treat", which have expected values of 1.100, 1.263, and 1.200 QALYs, respectively. Based on this analysis, the optimal choice is to "Treat".

Expected Value of Clinical Information

The expected value of clinical information is defined as the difference in expected value for the outcome of interest *with* the information relative to the expected value of the best alternative *without* information. In our example of the patient with AIDS, the expected value with clinical information provided by the IS test is 0.063 QALYs and is computed as the difference in expected value between the "Test" and "Treat" branches ($0.063 = 1.263 - 1.200$). Note that, if the expected value of clinical information were negative, this would reflect a situation where the optimal action was not changed by the test result. When we averaged out and folded back the "Test" branch, our analysis resulted in "Treat" following a positive test result and "No intervention" following a negative test result as the optimal course of action. Thus, the optimal action varied according to the test result, and the expected value of clinical information was positive.

A related concept is the expected value of *perfect* information, which is defined as the difference in expected value for the outcome of interest *with* perfect information relative to the expected value of the best alternative *without* perfect information. In our example, if we knew that the patient had PCP we would elect to treat, and if we knew that the patient did not have PCP we would elect no intervention. Thus, with perfect information the expected value is 1.400 QALYs ($1.400 = 0.6 \times 1.0 + 0.4 \times 2.0$). Without a perfect test, the best alternative is the imperfect test, which has an expected value of 1.263 QALYs. Therefore, the expected value of perfect information is 0.137 QALYs ($0.137 = 1.400 - 1.263$).

The expected value of perfect information can be a useful filter for assessing the value of risky tests. In the AIDS example, the expected value of perfect information informs us about the maximum loss of QALYs that one should accept to distinguish PCP from other underlying causes of cough. A loss of 0.137 QALYs or more would be unacceptable even for a perfect test. To determine the probability of death, Pr(die), to which the loss of 0.137 QALYs corresponds, we solve the equation, $1.263 =$

$[1 - \Pr(\text{die})] \times 1.4$. This equation is obtained by setting the expected value of our best alternative, the imperfect test, equal to the expected value of obtaining perfect information among those who do not die from the risky perfect test. We find that a loss of 0.137 QALYs corresponds to accepting no greater probability of death than 0.098. Thus, any risky *imperfect* test must carry an even lower probability of death to be worthwhile.

Sensitivity Analysis

To assess the stability of the baseline results to each decision tree parameter (e.g. probabilities and outcome values), a series of analyses are undertaken in which model parameters are modified systematically over a range of reasonable values to assess whether or not the optimal choice varies. Such analyses are referred to as **sensitivity analyses** and they help identify the parameters that have the greatest influence on the results.

One-Way Sensitivity and Threshold Analyses

In one-way sensitivity analyses, a single parameter is varied over a reasonable range and the expected value of each action is recalculated. Graphical representations of one-way sensitivity analyses are useful for characterizing how the optimal action changes as a single parameter is varied. Such analyses are often valuable when debugging a decision tree.

In the example involving the AIDS patient suspected of having PCP, a graphical representation of the one-way sensitivity analysis for probability of PCP highlights the points at which the optimal action changes [Figure 5(a)]. The points at which the curves of expected value for each strategy cross are threshold values. In our example, Figure 5(a) identifies two thresholds. When the probability of PCP is below 0.018, "No intervention" is preferred; when the prevalence is greater than 0.688, "Treat" is preferred. Analyses that solve for threshold values directly among pairs of alternative actions are referred to as *threshold analyses*.

Multiway Sensitivity Analyses

To assess the impact of changes in multiple parameters simultaneously, higher-order sensitivity analyses are undertaken. Graphical representations of two-way sensitivity analyses are often useful for identifying ranges over which analysis results are stable. In the example involving the AIDS patient, we examined the impact of prevalence of PCP (baseline value = 0.6) and QALYs for patients with PCP and

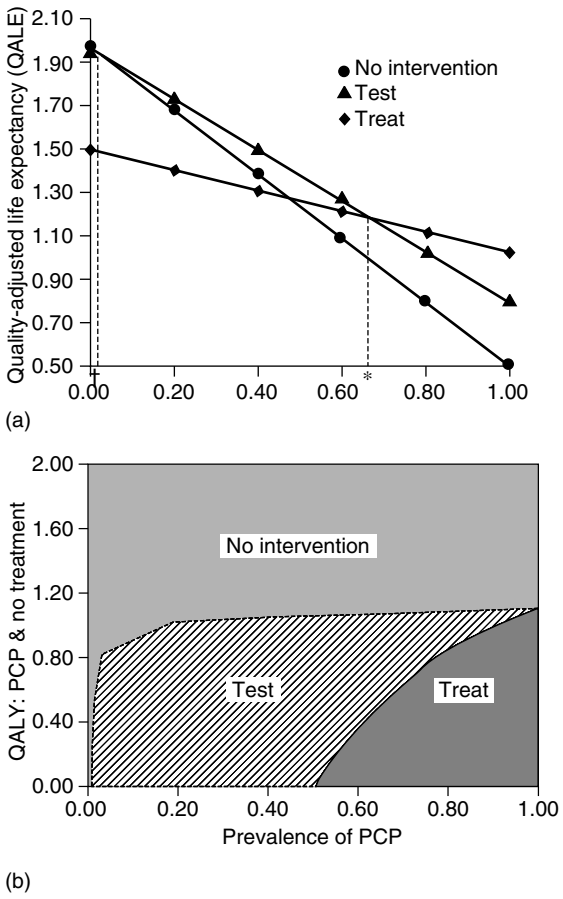

Figure 5 Results of sensitivity analyses for the clinical example involving AIDS patient with a persistent cough: (a) graphical representation of one-way sensitivity analysis. Prevalence of PCP is varied from 0 to 1.0 and the expected value of each strategy is graphed. Thresholds where the optimal action changes from no intervention to testing and from testing to treatment are identified as (+)0.018 and (∗)0.688, respectively; (b) graphical representation of two-way sensitivity analysis. The prevalence of PCP (baseline value, 0.6) and quality-adjusted life years (QALYs) for persons with PCP and no treatment (baseline value, 0.5 QALYs) are varied simultaneously. Regions where each action is preferred are labeled

no treatment (baseline value 0.5 QALYs) simultaneously. The corresponding graph highlights combinations of these two parameters for which each action is preferred [Figure 5(b)]. For high prevalence of PCP and low QALYs for untreated PCP, treatment is preferred. In contrast, for high QALYs for untreated PCP (higher than approximately 1.0 QALY), no intervention is preferred regardless of the underlying prevalence of PCP.

Probabilistic sensitivity analysis [6] is another form of multiway sensitivity analysis, which involves specification of probability distributions for model parameters and **Monte Carlo simulation**.

Extensions to Multiple or Repeated Diagnostic Tests

When **multiple diagnostic tests** are considered either simultaneously or in sequence, the decision tree becomes more complex. In particular, estimating the revised probability of disease based on either two repeated or two separate tests is challenging and often involves strong assumptions.

Conditional Independence

To evaluate the probability of disease on the basis of results from two or more tests requires either knowledge of the joint receiver operating characteristics of the tests or assumptions about the operating characteristics. Because the former is often not available, the assumption of independence of test results conditional on true disease state is often invoked. For results from two tests, denoted as R_1 and R_2, the conditional independence assumption is summarized as:

$$\Pr(R_1, R_2|D+) = \Pr(R_1|D+)\Pr(R_2|D+),$$

$$\Pr(R_1, R_2|D-) = \Pr(R_1|D-)\Pr(R_2|D-).$$

The extent to which such an assumption is reasonable will vary by disease area. To assess the validity or plausibility of the conditional independence assumption in the repeated test setting requires consideration of the nature of variation in test results.

Extensions to More Complex Decision Problems

To evaluate clinical management strategies involving a long time horizon, the use of decision trees becomes cumbersome. For example, consider evaluating whether or not women should be treated with hormone replacement therapy (HRT) at menopause and whether or not such treatment should depend on **screening** tests [23]. To model the effects of HRT accurately, the changing incidence of the multiple diseases that are affected by HRT (e.g. heart disease, breast cancer, and osteoporosis) must be modeled over the course of a woman's lifetime. A decision tree model for evaluation of HRT quickly becomes unwieldy. To evaluate diagnosis and treatment decisions involving long time horizons efficiently, which are common when chronic diseases are considered, **Markov** state-transition models are helpful [1, 20]. Markov state-transition models require that a set of health states be specified along with rules for transition between health states and corresponding transition probabilities, which depend on the current health state. The use of Markov models to evaluate diagnostic and treatment decisions in medicine is increasingly common. In another clinical example, a Markov state-transition model was used to evaluate the use of myocardial revascularization in patients with chronic stable angina [25].

Software for Analyzing Decision Trees

Software packages are available for analyzing decision trees and have been frequently used in decision analyses of diagnosis and treatment [4, 5, 19]. In addition to allowing for the analysis of simple decision trees, these software packages facilitate evaluation of more complex model structures, including recursive decision trees and Markov state-transition models.

References

[1] Beck, J.R. & Pauker, S.G. (1983). The Markov process in medical prognosis, *Medical Decision Making* **3**, 419–458.

[2] Beck, J.R., Kassirer, J.P. & Pauker, S.G. (1982). A convenient approximation of life expectancy (the "DEALE"). I. Validation of the method, *American Journal of Medicine* **73**, 883–888.

[3] Beck, J.R., Pauker, S.G. Gottlieb, J.E., Klein, K. & Kassirer, J.P. (1982). A convenient approximation of life expectancy (the "DEALE"). II. Use in medical decision-making, *American Journal of Medicine* **73**, 889–897.

[4] DATA 4.0 and DATA PRO for Healthcare (2001). TreeAge Software, Inc., 1075 Main Street, Williamstown.

[5] *Decision Maker 7.06* (1993). Pratt Medical Group, Boston.

[6] Doubilet, P., Begg, C.B., Weinstein, M.C., Braun, P. & McNeil, B.J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulations: a practical approach, *Medical Decision Making* **5**, 157–177.

[7] Freedberg, K.A., Tosteson, A.N.A., Cotton, D.J. & Goldman, L. (1992). Optimal management strategies for HIV-infected patients who present with cough or dyspnea: a cost-effectiveness analysis, *Journal of General Internal Medicine* **7**, 261–272.

[8] Gold, M.R., Siegel, J.E., Russell, L.B. & Weinstein, M.C. eds (1996). *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York.

[9] Hunink, M.G.M., Glasziou, P.P., Siegel, J.E., Weeks, J. Pliskin, J.S., Elstein, A.S., Weinstein, M.C. (2001). Decision making in health and medicine: Integrating evidence and valves.

[10] Kassirer, J.P., Moskowitz, A.J., Lau, J. & Pauker, S.G. (1987). Decision analysis: a progress report, *Annals of Internal Medicine* **106**, 275–291.

[11] Ledley, R.S. & Lusted, L.B. (1959). Reasoning foundations of medical diagnosis, *Science* **130**, 9–21.

[12] Lusted, L.B. (1968). *Introduction to Medical Decision Making*. Charles C. Thomas, Springfield.

[13] Lusted, L.B. (1971). Decision-making studies in patient management, *New England Journal of Medicine* **284**, 416–424.

[14] Metz, C.E. (1978). Basic principles of ROC analysis, *Seminars in Nuclear Medicine* **8**, 283–298.

[15] Oliver, R.M. & Smith, J.Q. eds (1990). *Influence Diagrams, Belief Nets and Decision Analysis*. Wiley, New York.

[16] Pauker, S.G. & Kassirer, J.P. (1980). The threshold approach to clinical decision making, *New England Journal of Medicine* **302**, 1109–1117.

[17] Pliskin, J.S., Shepard, D.S. & Weinstein, M.C. (1980). Utility functions for life years and health status, *Management Science* **28**, 206–224.

[18] Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Addison-Wesley, Reading.

[19] *SMLTREE* (1989). Hollenberg, New York.

[20] Sonnenberg, F.A. & Beck, J.R. (1993). Markov models in medical decision making: a practical guide, *Medical Decision Making* **13**, 322–338.

[21] Sox, H.C., Blatt, M.A., Higgins, M.C. & Marton, K.I. (1988). *Medical Decision Making*. Butterworths, Boston.

[22] Tosteson, A.N.A., Goldman, L., Udvarhelyi, I.S. & Lee, T.H. (1996). Cost-effectiveness of a coronary care unit versus an intermediate care unit for emergency department patients with chest pain, *Circulation* **94**, 143–150.

[23] Tosteson, A.N.A., Rosenthal, D.I., Melton, L.J. & Weinstein, M.C. (1990). Cost effectiveness of screening perimenopausal white women for osteoporosis: bone densitometry and hormone replacement therapy, *Annals of Internal Medicine* **113**, 594–603.

[24] Weinstein, M.C. & Stason, W.B. (1977). Foundations of cost-effectiveness analysis for health and medical practices, *New England Journal of Medicine* **296**, 716–721.

[25] Wong, J.B., Sonnenberg, F.A., Salem, D.N. & Pauker, S.G. (1990). Myocardial revascularization for chronic stable angina: analysis of the role of percutaneous transluminal coronary angioplasty based on data available in 1989, *Annals of Internal Medicine* **113**, 852–871.

(*See also* **Risk Assessment in Clinical Decision Making**; **Standard Gamble Technique**; **Time Trade-off Technique**)

ANNA N.A. TOSTESON

Decision Theory

The theory underlying methods for the selection of the best decision to be made in the setting of uncertainty is called statistical decision theory [2, 11, 12, 15, 17–19, 21]. While both frequentist and **Bayesian** approaches to decision problems exist, the Bayesian paradigm for gaining information from data provides the most coherent framework in which to make decisions in the setting of uncertainty [2]. This article gives a general overview of statistical decision theory, emphasizing the Bayesian approach, and illustrates some general classes of problems with examples. We begin by reviewing **Bayes' theorem** and defining some important expectations.

Updating Bayesian Probabilities

Within the Bayesian framework, the probability of any particular value of an unknown parameter being the true value is described by a probability density function (pdf). The unknown parameter, which may be a vector, is denoted by θ and the range of possible values is the parameter space Θ. Experimental data are often obtained in an effort to increase one's knowledge about the unknown parameter. Generally, one can make better decisions with more information about θ.

Before experimental data are available, the pdf describing our knowledge about θ is termed the **prior** pdf and is denoted $\pi(\theta)$. The information contained in the prior may come from previous experience in similar situations, from expert opinion, or from other sources [10]. Methods for determining prior pdfs will not be discussed here [10].

We assume that the probability of observing any particular set of experimental data, x, on the space $\mathcal{X}$, is given by a normalized pdf, denoted $f(x|\theta)$, which in turn is characterized by the unknown parameter θ. Then, given an observed set of data, x, the *posterior* pdf for θ, $\pi(\theta|x)$, is given by

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\displaystyle\int_{\Theta} f(x|\theta)\pi(\theta)\,\mathrm{d}\theta}.$$

This is the continuous form of Bayes theorem. Considered as a function of θ, $f(x|\theta)$ is called the **likelihood** function.

Decisions may be made before data from the experiment in question become available, on the basis of the prior $\pi(\theta)$, or after the experiment, on the basis of the posterior pdf $\pi(\theta|x)$. The pdf for θ at the time a decision must be made will be denoted $\pi^*(\theta)$ or π^*. Much of the notation used here has been adapted from Berger's excellent monograph [2].

Expectations

When determining the **expectation** of a function, one must distinguish the arguments that are random from those that are fixed. Three expectations will be important for the material that follows: (i) the expectation of a function when the parameter θ is fixed and the data x are random, having pdf $f(x|\theta)$; (ii) the expectation of a function when the data are fixed and θ is random, having pdf $\pi^*(\theta|x)$; and (iii) the expectation of a function when θ is random having pdf $\pi^*(\theta)$, *and* x is random for each θ, having pdf $f(x|\theta)$.

In the first case, in which θ is fixed and x is random, the expectation of $g(x)$ is written as $\mathrm{E}_{\theta}^{X} g(X)$. This is equivalent to $\int_{\mathcal{X}} g(x) f(x|\theta)\,\mathrm{d}x$. When x is fixed and θ is random, the expectation of $h(\theta)$ is written as $\mathrm{E}_{x}^{\pi^*} h(\theta)$. This is equivalent to $\int_{\Theta} h(\theta)\pi^*(\theta|x)\,\mathrm{d}\theta$. In the last case, in which both θ and x are random, the expectation of $j(\theta, x)$ is written as $\mathrm{E}^{\pi^*}[\mathrm{E}_{\theta}^{X} j(\theta, X)]$. This is equivalent to $\int_{\Theta}\int_{\mathcal{X}} j(\theta, x) f(x|\theta)\pi^*(\theta)\,\mathrm{d}x\,\mathrm{d}\theta$. If the function being considered is the loss function (see below), then these three expectations will be called the frequentist risk, the conditional Bayes risk, and the Bayes risk, respectively (see below) [2].

Elements of a Decision Problem

A decision problem consists of three parts: (i) the parameter space, Θ, in which the unknown parameter θ exists; (ii) the set of all possible actions $\mathcal{A}$, from which an action a is to be selected; and (iii) the loss function, $L(\theta, a)$ [2, 15].

Parameter Space or State of Nature

The parameter space, Θ, defines the possible values of θ. Usually the unknown parameter θ represents the true state of nature. In game theory, however,

θ may be the position or tactic of the opponent, and thus may be selected nonrandomly and with specific objectives in mind [15]. In a **classification** problem the parameter space is the list of possible classes. For example, if the objective of a decision problem is to determine the species of an animal on the basis of some measured characteristics, then the parameter space is the set of all possible species. In an **estimation** problem, θ is the parameter to be estimated.

A Set of Possible Actions or Decisions

The set of all possible actions or decisions is denoted $\mathcal{A}$. This set may be discrete, for example selecting a particular medical therapy, determining which disease a patient has, or deciding the number of patients to be enrolled in a clinical trial. An action space may also be continuous, as commonly occurs when the action is an estimate of an unknown parameter, such as the mean of a population. In problems of **experimental design**, $\mathcal{A}$ is the set of all possible study designs. The set of possible study designs may have both discrete elements, such as sample size, and continuous elements, such as a dosage level to be used, the duration of observation, or the parameter estimate to be given (see below).

The Loss Function

The **loss function**, $L(\theta, a)$, represents the loss associated with taking the action a when the true value of the unknown parameter is θ. The loss function must be bounded and defined on the space $\Theta \times \mathcal{A}$. In estimation problems the loss function usually includes a term related to the error of the estimate. In problems involving both estimation and the selection of a data collection or experimental strategy (e.g. deciding the number of subjects to be enrolled in a trial, or selecting an experimental design) the loss function should also include the cost of the data collection itself. The development of realistic loss functions for complex decision problems can be extremely difficult, and can require the consideration of economic, social, psychological, political, and other factors.

Utility is the negative of loss; it is the gain realized if the action a is taken when the true value of the unknown parameter is θ.

Example of Loss Functions. Consider a case in which a patient may have one of three diseases, labeled {A, B, C}, and may be given one of two drugs, labeled {D1, D2}. D1 is the better treatment for disease A, D2 is the better treatment for disease B, and neither drug is effective for Disease C. Table 1 shows a possible loss function. In addition to quantifying the chance of treatment failure with each of the drugs, this loss function should incorporate the monetary costs of the drugs, difficulty in taking the drugs, and the likelihood and seriousness of side-effects. The action that minimizes the expected loss will depend on the probability that the patient has each of the three diseases under consideration.

Table 1 An example of a loss function. Drug D1 is the best treatment for disease A and drug D2 is the best treatment for disease B. Both D1 and D2 are ineffective treatments for disease C

		Action	
		Give drug D1	Give drug D2
Actual disease present	Disease A	1	10
	Disease B	20	5
	Disease C	25	25

As another example, consider the situation in which one wishes to collect data in order to estimate the unknown mean, μ, of a normally distributed variable. There are two parts to the decision problem: (i) choosing the sample size, n, and (ii) estimating μ. In this case the loss function might be $L(\mu, a) = K(\mu - \mu^*)^2 + cn$, where $a = (\mu^*, n)$ incorporates both the estimate of μ to be given, denoted μ^*, and the sample size selected, n. The constant c is the cost per observation. The loss function thus incorporates both the error of the estimate ultimately given, and the cost of acquiring the data. This example is developed more fully later.

The Decision Rule

The decision rule, $\delta(x)$, is a function that maps the data, if any, into the possible set of actions – it is the action to be taken (or the estimate to be given) if the observed data are x. The space of all possible decision rules will be denoted $\mathcal{D}$. During the solution of a decision problem, a decision rule will be chosen to minimize the expected loss, where the expected loss is defined in some specific manner (see later).

A randomized decision rule has the characteristic that, for at least some data $x \in \mathcal{X}$, the action to be

taken is a random variable. Thus the data determine the *probability* of each action being taken, but more than one action may be possible for a given set of data.

Types of Loss and Risk

The term **risk** is used to denote measures of expected loss associated with a specific action or using a specific decision rule (Table 2). We will consider three types of risk: frequentist risk, conditional Bayes risk, and Bayes risk [2]. The definitions of types of risk and the related terminology vary from author to author.

Frequentist Risk. The frequentist risk of a decision rule is denoted $R(\theta, \delta(x))$ and is defined by

$$R(\theta, \delta(x)) = \mathrm{E}_{\theta}^{X}[L(\theta, \delta(X))] = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta)\, \mathrm{d}x.$$

The frequentist risk is the average loss incurred by using a decision rule $\delta(x)$ when the true value of the unknown parameter is θ. Sometimes this is called the *expected loss*.

In selecting a decision rule according to the **minimax** principle (see below), it is necessary to consider the maximum frequentist risk that might occur, considering all possible values of θ. This maximum frequentist risk is written

$$R_{\max}(\delta(x)) = \sup_{\theta \in \Theta} R(\theta, \delta(x)).$$

Conditional Bayes Risk. The conditional Bayes risk of an action or decision rule is the expectation of the loss incurred by using that action or decision rule, assuming the data are known. In determining the conditional Bayes risk, the data x are assumed fixed, so the action to be taken, $a_x = \delta(x)$, is also fixed if the decision rule is not randomized. This risk is denoted $\rho(\pi^*(\theta), a_x)$ and is defined by

$$\rho(\pi^*(\theta), a_x) = \mathrm{E}_{x}^{\pi^*} L(\theta, a_x) = \int_{\Theta} L(\theta, a_x)\pi^*(\theta|x)\mathrm{d}\theta.$$

The conditional Bayes risk is the expected loss associated with using a specific action, given the data. It is also called the *posterior expected loss*.

Consider the loss function in Table 1. Assume that on the basis of the available data, the current distribution is $\pi^*(\text{disease A}) = 0.3$, $\pi^*(\text{disease B}) = 0.2$, and $\pi^*(\text{disease C}) = 0.5$. The conditional Bayes risk for giving drug D1 is then

$$\rho(\pi^*, D1) = \sum_{\text{disease}=\{\text{A,B,C}\}} L(\text{disease, D1})\pi^*(\text{disease}) = 1 \times 0.3 + 20 \times 0.2 + 25 \times 0.5 = 16.8.$$

Similarly, for drug D2 the conditional Bayes risk is $\rho(\pi^*, D2) = 10 \times 0.3 + 5 \times 0.2 + 25 \times 0.5 = 16.5$. Thus, conditional on the available data, the optimal action is to give drug D2.

Bayes Risk. The Bayes risk of a decision rule is the expected loss, before the data are known, incurred by using the decision rule. Information about possible

Table 2 Definitions of risks used in the analysis of decision problems

Type of risk	Description	Symbol
Frequentist risk	The frequentist risk is the expected value of the loss function, $L(\theta, \delta(x))$, assuming that θ is fixed and the data x have distribution $f(x\|\theta)$.	$R(\theta, \delta(x))$
Conditional Bayes risk	The conditional Bayes risk is the expected loss (with respect to the current probability distribution for θ) after the data are known. It is the posterior expected loss.	$\rho(\pi^*(\theta), a_x)$
Bayes risk	The Bayes risk is the expected loss associated with using a particular decision rule, before the data are known. The expectation is taken with respect to the pdf for θ and the pdf for the data, given θ.	$r(\pi^*(\theta), \delta(x))$

values of θ enters through $\pi^*(\theta)$. This risk is denoted $r(\pi^*, \delta(x))$ and is defined by

$$\begin{aligned} r(\pi^*(\theta), \delta(x)) &= \mathrm{E}^{\pi^*}[R(\theta, \delta(x))] \\ &= \mathrm{E}^{\pi^*}[\mathrm{E}_\theta^X L(\theta, \delta(X))] \\ &= \int_\Theta \int_\mathcal{X} L(\theta, \delta(X)) f(x|\theta) \pi^*(\theta)\, \mathrm{d}x\, \mathrm{d}\theta. \end{aligned}$$

Minimization of the Bayes risk is an important criterion used to determine an optimal decision rule.

Criteria for Selecting Decision Rules

Different criteria can be used to judge which decision rules are "better" than others, or even optimal. The principal Bayesian criterion is minimizing the Bayes risk, $r(\pi^*, \delta(x))$, while an important frequentist criterion is minimizing the maximum risk, $R_{\max}(\delta(x))$.

Admissibility and Bayes Rules

A decision rule, $\delta_1(x)$, is *admissible* with respect to another decision rule, $\delta_2(x)$, if there is at least some value of θ for which the frequentist risk for $\delta_1(x)$, given by $R(\theta, \delta_1(x)) = \mathrm{E}_\theta^X L(\theta, \delta_1(X))$, is less than that for $\delta_2(\theta)$.

A decision rule, $\delta(x)$, is a "Bayes rule" if the decision rule minimizes the Bayes risk, $r(\pi^*(\theta), \delta(x))$; such a decision rule is denoted $\delta^*(x)$. In general, the Bayes rule will depend upon $\pi^*(\theta)$. The minimum risk achieved by a Bayes rule, for the particular $\pi^*(\theta)$, is denoted r^*. Any other decision rule which leads to the same Bayes risk is also a Bayes rule [2].

If a decision rule $\delta_1(x)$ is admissible with respect to all other decision rules in $\mathcal{D}$, meaning that there is at least one value of θ for which $R(\theta, \delta_1(x)) < R(\theta, \delta_i(x))$ for all other rules $\delta_i(x) \in \mathcal{D}$, $i \neq 1$, then $\delta_1(x)$ is a Bayes rule for some $\pi^*(\theta)$.

Minimax Rules

The definition of Bayes risk and the determination of a Bayes rule require specifying the pdf for the unknown parameter θ. In some cases one might desire to use a classical method to select the "best" decision rule, and avoid the use of the Bayesian pdf. One such method for rule selection seeks to minimize the maximum frequentist risk, $R_{\max}(\delta(x))$.

Recall that the maximum frequentist risk that might be incurred using a decision rule $\delta(x)$, as θ varies, is

$$R_{\max}(\delta(x)) = \sup_{\theta \in \Theta} R(\theta, \delta(x)).$$

According to the minimax principle, a decision rule $\delta_1(x)$ is preferred over a second rule $\delta_2(x)$ if

$$R_{\max}(\delta_1(x)) < R_{\max}(\delta_2(x)).$$

The value of θ that leads to the maximum value of $R(\theta, \delta_1(x))$ is not, in general, the same value of θ that leads to the maximum value of $R(\theta, \delta_2(x))$.

The best decision rule will be the one that minimizes the maximum frequentist risk; this is the minimax rule, denoted $\delta^{*\mathrm{M}}(x)$. The resulting minimax risk is given by

$$r^{*\mathrm{M}} = \inf_{\delta(x) \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta(x)).$$

The minimax rule is an appropriate method for selecting a decision rule if the goal is to minimize the "worst case" loss that might be incurred. This is a reasonable approach when the selection of θ is performed by an opponent who seeks to maximize your loss, for example when playing a game against an intelligent adversary [15].

Consider again the loss function shown in Table 1. Regardless of the drug given, the maximum loss of 25 occurs if the patient has disease C. Thus, both giving drug D1 and giving drug D2 are minimax rules and the minimax risk is 25. If the disease is selected by an intelligent adversary (say in biological warfare), then a minimax approach to treatment selection would be reasonable.

The Range of Decision Problems

In Bayesian decision theory the range of decisions that can be considered is very broad – limited only by the types of decision space that can be defined. Decision problems include simple classification, parameter estimation, optimal sample size determination, and experimental design [3–11, 14, 18, 24–26].

A distinction should be made between a decision problem and an estimation problem, in which the goal is to return the "best" (in some defined sense) estimate of an unknown parameter. Once a suitable

loss function, such as quadratic error loss, is defined, the goal of an estimation problem is still to minimize the Bayes risk. In an estimation problem, however, the decision rule is an estimator of the unknown parameter. A design problem is a decision problem in which the action space is the set of possible experimental designs. A special case is the problem of choosing a sample size: this is a decision problem in which the action space is $\{0, 1, 2, \ldots\}$.

Some examples of classes of Bayesian decision problems will be given in the following sections.

Example: Choosing a Treatment

Recall the loss function shown in Table 1 for the problem of selecting one of two treatments when the patient might have one of three diseases. Inspection of the loss structure shows that the optimal treatment will depend on the probabilities of the different diseases. Suppose we have a diagnostic test available, at a cost of one unit, the characteristics of which are shown in Table 3. We wish to decide whether to order the test and, if we use the test, what treatment to give for each possible result. As before, the probabilities of the three diseases before the test result is available are $\pi^*(\text{disease A}) = 0.3$, $\pi^*(\text{disease B}) = 0.2$, and $\pi^*(\text{disease C}) = 0.5$. Because the data space $\mathcal{X}$ consists of only two possible test results, we directly calculate the conditional Bayes risk for each test result, and for the strategy of not ordering the test and treating on the basis of prior information alone.

As shown above, if we do not order the test then the optimal action is to give drug D2 and the conditional Bayes risk is 16.5. If we order the test, and the result is negative, then the posterior probabilities for diseases A, B, and C are 0.056, 0.296, and 0.648, respectively. Using these posterior probabilities, the conditional Bayes risk associated with giving drug D1 is 22.2. The total cost, if one obtains a negative test and gives drug D1 anyway, is 22.2 plus the cost of the test, or 23.2. Similarly, the total cost of giving drug D2 after a negative test is 19.2. Thus, if the test is ordered and negative, then the optimal action is to give drug D2.

Table 3 Probabilities of a positive and negative test result, depending on the actual disease present, using a diagnostic test marketed as being useful for the identification of disease A

		Test result	
		Negative	Positive
Actual disease present	Disease A	0.1	0.9
	Disease B	0.8	0.2
	Disease C	0.7	0.3

If the test is positive, then the posterior probabilities of diseases A, B, and C are 0.587, 0.087, and 0.326, respectively. Now the cost of testing and giving drug D1 is 11.5 and the cost of testing and giving drug D2 is 15.5. Thus, if the test is ordered and positive, then the optimal action is to give drug D1. This is not surprising, since the test is meant to detect disease A and drug D1 is the better treatment for disease A.

To decide whether to order the test, one must calculate the prior probability that the test will be negative or positive. The prior (or "predictive") probability of a negative test result is $0.3 \times 0.1 + 0.2 \times 0.8 + 0.5 \times 0.7 = 0.54$. The prior probability of a positive test result is 0.46. Thus, the Bayes risk, if the test is obtained (and the optimal treatment is given for each test result) is $0.54 \times 19.2 + 0.46 \times 11.5 = 15.7$. Since the Bayes risk associated with ordering the test, 15.7, is less than the risk associated with not ordering the test, 16.5, the optimal decision is to order the test. If the test costs two units instead of one unit, however, then it would be better not to order the test and simply treat with drug D2 (*see* **Decision Analysis in Diagnosis and Treatment Choice**).

Example: Choosing a Sample Size

Assume one wishes to estimate the unknown mean, μ, of a normally distributed variable with a known standard deviation, σ. There are two parts to the decision problem, choosing the sample size, n, and estimating μ. We assume a loss function $L(\mu, a) = K(\mu - \mu^*)^2 + cn$, where $a = (\mu^*, n)$ includes both the estimate of μ to be given, μ^*, and the sample size to be used, n. The constant c is the cost per observation. The prior for μ is $\mu \sim \mathrm{N}(\mu_0, \sigma_0)$.

The observed data will be $X = \{X_1, X_2, \ldots, X_n\}$. Since the n observations are independent, $f(X|\mu)$ is given by

$$N_n(\mu, \sigma) = (2\pi\sigma^2)^{-(n/2)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right].$$

The Bayes risk after n measurements, using an estimator μ^* which will be a function of the prior

and the observed data, is

$$r(\mathrm{N}(\mu|\mu_0,\sigma_0),a) = \int_{-\infty}^{\infty}\int_{\mathcal{X}}[K(\mu-\mu^*)^2+cn] \times f(X|\mu)\mathrm{N}(\mu|\mu_0,\sigma_0)\,\mathrm{d}X\,\mathrm{d}\mu.$$

It can be shown that minimizing the Bayes risk is equivalent to minimizing the conditional Bayes risk (the posterior expected loss), for each value of X [2]. The posterior for μ is given by $\mathrm{N}(\mu_1,\sigma_1)$ where

$$\mu_1(X) = \left(\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right)\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{X}}{\sigma^2}\right)$$

and

$$\sigma_1^2 = \left(\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right).$$

Since σ^2 is known, $\overline{X}$ is a **sufficient statistic**.

The conditional Bayes risk is given by

$$\rho(\mathrm{N}(\mu|\mu_1,\sigma_1),a) = \left(\frac{1}{(2\pi\sigma_1^2)^{1/2}}\right) \times \int_{-\infty}^{\infty}[K(\mu-\mu^*)^2+cn] \times \exp\left[-\frac{(\mu-\mu_1(X))^2}{2\sigma_1^2}\right]\mathrm{d}\mu.$$

The conditional Bayes risk is minimized by expanding the loss function, differentiating by μ^*, and setting the result equal to zero. Thus,

$$0 = \frac{\mathrm{d}}{\mathrm{d}\mu^*}\int_{-\infty}^{\infty}[K\mu^2-2K\mu\mu^*+K(\mu^*)^2+cn] \times \exp\left[-\frac{(\mu-\mu_1(X))^2}{2\sigma_1^2}\right]\mathrm{d}\mu,$$

$$0 = 2K\int_{-\infty}^{\infty}[\mu^*-\mu]\exp\left[-\frac{(\mu-\mu_1(X))^2}{2\sigma_1^2}\right]\mathrm{d}\mu,$$

and

$$\left(\frac{\mu^*}{(2\pi\sigma_1^2)^{1/2}}\right)\int_{-\infty}^{\infty}\exp\left[-\frac{(\mu-\mu_1(X))^2}{2\sigma_1^2}\right]\mathrm{d}\mu = \left(\frac{1}{(2\pi\sigma_1^2)^{1/2}}\right)\int_{-\infty}^{\infty}\mu\exp\left[-\frac{(\mu-\mu_1(X))^2}{2\sigma_1^2}\right]\mathrm{d}\mu$$

and

$$\mu^* = \mathrm{E}^{N(\mu|\mu_1,\sigma_1)}[\mu] = \mu_1(X).$$

It is a general result that the best squared error loss estimator is the mean of the posterior distribution for the parameter [2, 15]. The Bayes conditional risk of this estimator is then given by

$$\rho(\mathrm{N}(\mu|\mu_1,\sigma_1),\mu_1(X)) = \left(\frac{1}{(2\pi\sigma_1^2)^{1/2}}\right)\int_{-\infty}^{\infty}[K(\mu-\mu_1(X))^2+cn] \times \exp\left[-\frac{(\mu-\mu_1(X))^2}{2\sigma_1^2}\right]\mathrm{d}\mu.$$

Performing the integration leads to

$$\rho(\mathrm{N}(\mu|\mu_1,\sigma_1),\mu_1(X)) = K\sigma_1^2+cn.$$

Thus, the Bayes conditional risk is proportional to the posterior variance plus the sampling cost [2].

To determine the sample size that minimizes the risk, one can differentiate the Bayes conditional risk by n and set the result equal to zero. Thus

$$0 = \frac{\mathrm{d}}{\mathrm{d}n}\left\{K\left(\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right)+nc\right\} = -K\left(\frac{\sigma^2\sigma_0^4}{(\sigma^2+n\sigma_0^2)^2}\right)+c.$$

The minimum conditional risk is given by setting n equal to

$$n^* = \sigma\left(\frac{K}{c}\right)^{1/2}-\frac{\sigma^2}{\sigma_0^2}.$$

This expression has the important characteristic that the optimal sample size depends on the ratio of the error loss to the sampling cost. Since n is not really a continuous variable, the two integer values of n closest to n^* must be checked to see which of the two gives the minimum.

This decision problem is simplified by the fact that the minimum conditional Bayes risk, which is proportional to the posterior variance, does not depend on the data, except through n. If the conditional Bayes risk depends on the data itself (as will occur with binomial sampling, for example), then the predictive distribution of the data is required to calculate the expected Bayes risk plus sampling costs (*see* **Sample Size Determination**).

Example: Sequential Stopping Rules

Bayesian decision theory can be used: (i) to decide the optimal estimate to be given once all data are

available; (ii) to decide the optimal sample size (see previous section); and (iii) if the measurements can be taken sequentially, to decide the optimal time to stop making observations. This last application is termed **sequential analysis** [1, 2, 8, 12, 15, 16].

In a sequential stopping problem, two sets of actions are available at each decision point: (i) the collection of more data; and (ii) stopping data collection and giving an estimate or making a decision on the basis of the data already collected. Many such decision points may exist. The total cost of the experiment includes contributions from both the terminal decision loss function (e.g. squared error loss of the final estimate), and the sampling cost.

The decision function for a sequential decision problem contains both a stopping rule, which determines which results will lead to termination of data collection, and a decision rule, which determines what action will be taken or estimate given once the experiment is terminated. The optimal decision and stopping rules minimize the Bayes risk, which must be calculated using the prior information *and* the predictive distribution of possible future data. The quantity of future data will depend, in turn, on the stopping rule, as the decision to stop the trial results in a truncation of the sequence of available data.

Once data collection has stopped, the optimal decision or estimate is determined by minimizing the conditional Bayes risk. Thus the sequential nature of the study does not complicate the determination of the decision rule.

The complexity of considering all possible future data can make the determination of optimal stopping rules difficult. The solution is greatly simplified if the posterior expected loss is independent of the observations, as this allows the direct determination of the value of n that minimizes the posterior expected loss. This simplification is illustrated by the example in the previous section. In that example the optimal sequential stopping rule is the same as the optimal fixed-size stopping rule.

When risks are finite, and the sampling cost is positive, then the optimal stopping rule will always require stopping after some maximum number of samples have been taken. The actual optimal stopping point will depend on the data. In this case the method of backward induction can be used to determine the optimal stopping rule. Because of the complex notation involved, the description of backward induction given here will be qualitative. The interested reader is referred to the following references for details and examples: [1, 2, 8, 9, 12, 15, 16].

In backward induction, one begins by determining the optimal decision and the associated risk for each result that might occur after the maximum data collection. At this "terminal" point, which we will call stage M, the option of continuing the trial does not exist and the determination of the optimal decision is made by minimizing the conditional Bayes risk. Next, for each possible data set available one experimental step earlier (stage $M-1$) one determines the expected cost of stopping the experiment, assuming the decision that minimizes the conditional Bayes risk is made. This expected "stopping" cost at stage $M-1$ is compared with the expected cost associated with continuing data collection until stage M. The expected cost associated with continuing data collection is calculated using the *predictive* distribution of future data, based on the data available at stage $M-1$. This is an example of "preposterior" analysis. The optimal action at stage $M-1$ is the action (stopping or continuing data collection) that minimizes the expected cost. Typically, there is some decrease in the expected decision loss associated with further data collection, but this gain may or may not offset the additional sampling cost.

In a similar manner, one can step backwards through each stage $\{M-2, M-3, \ldots, 2, 1, 0\}$ and, for each possible set of results at each stage, determine the optimal stopping and decision rule. This "backward induction" continues until stage 0 is reached, before any data collection. It is possible that the optimal action at this initial stage will be to not collect any data at all, and instead make a decision or estimate based solely on prior information. Thus, sequential analysis allows one to decide whether it is optimal to even conduct the experiment at all. An excellent example of the power of this approach is illustrated in [9].

Example: Sequential Allocation of Experiments (Bandits)

In most **clinical trials**, patients are randomized in a balanced fashion to the candidate therapies. The advantage of a balanced design is that it gives maximal information about the differences between therapies. When results of a trial are published they help to guide the treatment of patients who present thereafter. Patients in the trial are not ignored, and

data and safety monitoring boards are charged specifically with ensuring that patients in the trial are not exposed to undue risks. But effective treatment of patients in the trial is not a formal objective.

An alternative approach is to address explicitly the effective treatment of patients in the trial as well as those who present thereafter. The goal is to maximize overall effectiveness. Therapies – or arms – are assigned on the basis of accumulating results; that is, assignment is adaptive (*see* **Adaptive and Dynamic Methods of Treatment Assignment**). An arm that is performing well is more likely to be assigned than is a poor performer. Information gleaned during the trial about the relative effectiveness of the arms has value in that it improves therapy for patients entering later in the trial, and also for patients who come after the trial [4, 5, 7, 8].

The decision space is complicated. Its first component indicates the arm selected initially. Suppose that the first observation is X_1. The second component of the decision is the arm selected next, given X_1 and also given the first arm selected. The third component depends on X_1 and the second observation X_2 and on the corresponding arms selected. And so on.

Temporarily consider only the n patients in the trial and suppose that there are two available arms. Outcomes are dichotomous; arm 1 has success probability θ_1 and arm 2 has success probability θ_2. The goal is to maximize the expected number of successes among the n patients. (From the perspective of losses, this goal is the same as minimizing the expected number of failures.) Arm 1 is standard and has known success proportion θ_1. Arm 2 has unknown efficacy. Uncertainty about θ_2 is given in terms of a probability distribution F. To be specific, suppose that F is **uniform** on (0, 1).

If $n = 1$, then the decision space is simply the list of possible initial selections, $\{1, 2\}$, and the decision problem is easy. Choosing arm 1 has expected number of successes θ_1. Choosing arm 2 has conditional expected number of successes θ_2, and unconditional expected number of successes $\mathrm{E}(\theta_2) = \int_0^1 p F_i(\mathrm{d}p) = \int_0^1 p\,\mathrm{d}p = 1/2$.

Therefore arm 1 is optimal if $\theta_1 \geq \frac{1}{2}$ and arm 2 is optimal if $\theta_1 \leq \frac{1}{2}$. (Both arms – and any randomization between them – are optimal when $\theta_1 = \frac{1}{2}$.)

The problem is more complicated for $n \geq 2$. Consider $n = 2$. There are two initial choices and two choices depending on the result of the first observation. There are eight possible decisions, or sequences of decisions, called *strategies*. We can write a strategy as $\{a; a_{\mathrm{S}}, a_{\mathrm{F}}\}$, where a is the initial selection, a_{S} is the next selection should the first observation be a success, and a_{F} is the next selection should the first observations be a failure. To find the utility (the negative of the expected loss) of a strategy we need to know such quantities as the probability of a success on arm 2 after a success on arm 2 ($\mathrm{E}(\theta_2^2)/\mathrm{E}(\theta_2) = \frac{2}{3}$) and the probability of a success on arm 2 after a failure on arm 2 ($\mathrm{E}(\theta_2(1-\theta_2))/\mathrm{E}(1-\theta_2) = \frac{1}{3}$). The possible strategies and their utilities are given in Table 4.

Table 4 Possible strategies and the resulting expected number of successes

Strategy	Expected number of successes
$\{1; 1, 1\}$	$2\theta_1$
$\{1; 1, 2\}$	$\theta_1 + \theta_1^2 + (1-\theta_1)(1/2)$
$\{1; 2, 1\}$	$\theta_1 + \theta_1(1/2) + (1-\theta_1)\theta_1$
$\{1; 2, 2\}$	$\theta_1 + 1/2$
$\{2; 1, 1\}$	$1/2 + \theta_1$
$\{2; 1, 2\}$	$1/2 + (1/2)\theta_1 + (1/2)(1/3)$
$\{2; 2, 1\}$	$1/2 + (1/2)(2/3) + (1/2)\theta_1$
$\{2; 2, 2\}$	$2(1/2) = 1$

It is easy to check that only three of these utilities are candidates for the maximum, with the optimal strategy depending on θ_1. If $\theta_1 \geq \frac{5}{9}$, then $\{1; 1, 1\}$ is optimal; if $\frac{1}{3} \leq \theta_1 \leq \frac{5}{9}$, then $\{2; 2, 1\}$ is optimal; and if $\theta_1 \leq \frac{1}{3}$, then $\{2; 2, 2\}$ is optimal.

Enumeration of possible strategies is tedious for large n. Most strategies can be dropped from consideration on the basis of theoretical results [5, 7, 20]. For example, there is a break-even value of θ_1, say θ_1^*, such that arm 1 is optimal for $\theta_1 \geq \theta_1^*$. Also, one need consider only those strategies that continue to use arm 1 once it has been selected. But many strategies will still remain. Backward induction can be used to find an optimal strategy. Table 5 gives the expected proportion of successes for selected values of n and for fixed $\theta_1 = \frac{1}{2}$, using an optimal strategy. The asymptotic maximal expected proportion of successes is $\frac{5}{8}$, which is the expected value of $\max(\theta_1, \theta_2)$.

Both arms offer the chance of success on the current patient, but only sampling from arm 2 gives information that can help in choosing between the arms for treating later patients. Table 6 gives the break-even values θ_1^* for selected values of n.

Table 6 shows that information from arm 2 is more important for larger n. For example, if $\theta_1 = 0.75$, then

Table 5 The optimal expected proportion of successes for selected values of n, assuming $\theta_1 = \frac{1}{2}$. Optimal strategies were determined using backward induction

n	Proportion of successes
1	0.500
2	0.542
5	0.570
10	0.582
20	0.596
50	0.607
100	0.613
200	0.617
500	0.621
1000	0.622
10 000	0.6245

Table 6 The break-even values of θ_1^*, as a function of n

n	θ_1^*
1	0.500
2	0.556
5	0.636
10	0.698
20	0.758
50	0.826
100	0.869
200	0.902
500	0.935
1000	0.954
10 000	0.985

only using arm 1 would be an optimal strategy for $n = 10$, but it would be advisable to test arm 2 when $n = 100$ even though arm 1 has probability of 0.75 of being better than arm 2.

When there are several arms with unknown characteristics, the problem is even more complicated. Optimal strategies may well include selections of an arm that was used previously and set aside in favor of another arm because of inadequate performance. Methods and theory for solving such problems are described in [3, 7]. Optimal strategies are generally difficult to describe. Berry provides easy-to-use adaptive strategies that are not optimal and shows that they perform reasonably well [4].

Having the flexibility to choose an arm based on results of all previous patients is not common in clinical trials. Usually there are response delays. For example, when the end-point is survival, only partial information is available until the patient dies. Eick [13, 14] has shown how to handle such delays in the adaptive setting.

We have not yet addressed the matter of patients that are treated after the clinical trial. The patient horizon N is the number of patients who will be treated either in the trial or later with one of the therapies considered in the trial. Clearly, in almost every real situation N is unknown. Information about N can be included in the decision problem in the usual way, by incorporating uncertainty about N into a probability distribution. Alternatively, one can assume particular values of N and assess the sensitivity of the optimal strategy to values assumed. Suppose that the clinical trial allows for adaptive allocation and that later patients will be assigned the treatment that performs best in the clinical trial. Berry & Eick [6] considered the case of two arms in a trial with dichotomous response and showed how to incorporate all N patients into the decision problem. They compare a Bayes strategy assuming θ_1 and θ_2 independent and having uniform prior distributions on (0,1) with various other adaptive strategies and with balanced randomization. The Bayes strategy performs best on average, as it must, and it is robust in the sense that it outperforms the other strategies for essentially all (θ_1, θ_2).

References

[1] Anscombe, F.J. (1963). Sequential medical trials, *Journal of the American Statistical Association* **58**, 365–383.

[2] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed. Springer-Verlag, New York.

[3] Berry, D.A. (1972). A Bernoulli two-armed bandit, *Annals of Mathematical Statistics* **43**, 871–897.

[4] Berry, D.A. (1978). Modified two-armed bandit strategies for certain clinical trials, *Journal of the American Statistical Association* **73**, 339–345.

[5] Berry, D.A. (1985). One- and two-armed bandit problems, in *Encyclopedia of Statistical Sciences*, Vol. VI, S. Kotz, & N.L. Johnson, eds. Wiley, New York, pp. 418–422.

[6] Berry, D.A. (2001). Sequential Statistical Methods, in *International Encyclopedia of Social and Behavioral Sciences*, Vol. 20, N.J. Smelser & P.B. Baltes, eds. Elsevier, Oxford, pp. 13922–13927.

[7] Berry, D.A. & Eick, S.G. (1995). Adaptive assignment versus balanced randomization in clinical trials: a decision analysis, *Statistics in Medicine* **14**, 231–246.

[8] Berry, D.A. & Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman & Hall, London.

[9] Berry, D.A. & Ho, C.H. (1988). One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach, *Biometrics* **44**, 219–227.

[10] Berry, D.A., Mueller, P., Grieve, A.P., Smith, M., Parke, T., Blazek, R., Mitchard, N. & Krams, M. (2001). Adaptive Bayesian designs for dose-ranging drug trials, in *Case Studies in Bayesian Statistics V*, C. Gatsonis, B. Carlin & A. Carriquiry, eds. Springer-Verlag, New York, pp. 99–181.

[11] Berry, D.A. Wolff, M.C. & Sack, D. (1994). Decision making during a Phase III randomized controlled clinical trial, *Controlled Clinical Trials* **15**, 360–379.

[12] Chaloner, K. (1996). Elicitation of prior distributions, in *Bayesian Biostatistics*, D.A. Berry & D.K. Stangl, eds. Marcel Dekker, New York, pp. 141–156.

[13] Clemen, R.T. (1991). *Making Hard Decisions: An Introduction to Decision Analysis*. PWS-Kent, Boston.

[14] DeGroot, M.H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

[15] Eick, S.G. (1987). The two-armed bandit with delayed responses, *Annals of Statistics* **16**, 254–264.

[16] Eick, S.G. (1988). Gittins procedures for bandits with delayed responses, *Journal of the Royal Statistical Society, Series B* **50**, 125–132.

[17] Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, San Diego.

[18] Lewis, R.J. & Berry, D.A. (1994). Group sequential clinical trials: a classical evaluation of Bayesian decision-theoretic designs, *Journal of the American Statistical Association* **89**, 1528–1534.

[19] Lindley, D.V. (1986). *Making Decisions*. Wiley, London.

[20] Raiffa, H. (1968). *Decision Analysis*. Addison-Wesley, Reading.

[21] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*. MIT Press, Cambridge, Mass.

[22] Ross, S.M. (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press, New York.

[23] Smith, J.Q. (1988). *Decision Analysis*. Chapman & Hall, London.

[24] Stallard, N. (1998). Sample size determination for phase II clinical trials based on Bayesian decision theory, *Biometrics* **54**, 279–294.

[25] Stallard, N. (2003). Decision-theoretic designs for phase II clinical trials allowing for competing studies, *Biometrics* **59**, 402–409.

[26] Stallard, N., Thall, P.F., Whitehead, J. (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes, *Biometrics* **55**, 971–977.

(*See also* **Data and Safety Monitoring**; **Dynamic Allocation Index**)

ROGER J. LEWIS & DONALD A. BERRY

Degrees of Freedom

Different authors have variously described or attempted to define the term *degrees of freedom* as the effective number of independent observations; the number of free variables; the number of observations that are free to vary after applicable restrictions are imposed; the number of observations minus the number of restrictions; the number of observations minus the number of parameters that are estimated from the observations; or something similar. Although descriptions such as these may be appropriate in a number of contexts, they fail to cover certain situations (e.g. those involving noninteger degrees of freedom).

Thus, for full generality it seems more suitable to think of degrees of freedom simply as a *parameter* (ordinarily a *known* parameter) of the different distributions with which the term is associated. There are three such distributions: **chi-square** (χ^2), t (*see* **Student's *t* Distribution**), and $\boldsymbol{F}$. Although the **gamma distribution** is a generalization of the χ^2 distribution, and the **beta distribution** is obtainable from the F distribution through a simple transformation, the term *degrees of freedom* is not customarily used in connection with either the gamma or the beta distributions.

The usual abbreviation for degrees of freedom is df. The χ^2 distribution has just one df parameter, as does the t distribution. The F distribution, however, has two df parameters; if they are shown as "df $= 3, 29$", for example, then this means that there are 3 df for the numerator of the reported F statistic and 29 df for the denominator. Published works that report values of χ^2, t, or F should always indicate the associated df parameters so that readers will be able to make proper interpretations.

In general applications it is the *central* χ^2, t, and F distributions that one encounters. The more complicated, *noncentral* χ^2, t, and F distributions also have df parameters, however, in the same fashion as the central distributions (*see* **Noncentral *t* Distribution**). The noncentral distributions are used for **power** calculations.

Significance tests that involve the χ^2, t, and F distributions, with associated df, are available for numerous applications (*see* **Hypothesis Testing**). These include t tests for a mean and for the equality of two means (*see* **Student's *t* Statistics**), the χ^2 test for a variance, the F test for the equality of two variances, various t and F tests in **regression** and in **analysis of variance**, chi-square goodness-of-fit tests, **chi-square tests** for independence in **contingency tables**, and **likelihood ratio tests** that use χ^2. **Confidence intervals** and confidence regions that are related to these significance tests are also available in many cases.

Noninteger degrees of freedom arise when a variable, u, has a complicated distribution for which one seeks a simple approximation. Specifically, one tries to choose a constant c and a df parameter ν so that $y = cu$ follows approximately a χ^2 distribution with df $= \nu$. One would like to select ν and c so that y has the same mean and variance as χ^2 with df $= \nu$, that is, so that $\mathrm{E}(y) = \nu$ and $\mathrm{var}(y) = 2\nu$. This means choosing $\nu = 2[\mathrm{E}(u)]^2/\mathrm{var}(u)$ and $c = 2\mathrm{E}(u)/\mathrm{var}(u)$; but if $\mathrm{E}(u)$ and $\mathrm{var}(u)$ are unknown, then one has to replace them with estimates. Generally, ν will not be an integer. The technique just described is known as Satterthwaite's [3] approximation. Applications include those involving the **Behrens–Fisher problem**, variance components (see, for example, [1]), and contingency tables [2].

Before the advent of modern computers, an application with noninteger df required interpolation in a table of χ^2, t, or F. Today, however, this is no longer necessary, because values for noninteger df can be obtained through statistical **software** packages.

It may be useful to point out certain special relationships that pertain to degrees of freedom:

1. The t distribution with df $= 1$ is the same as the **Cauchy distribution** with median 0.
2. The χ^2 distribution with df $= 2$ is the same as the **exponential distribution** with mean 2.
3. A table of the t distribution typically has a line at the bottom that shows ∞ for df. For given significance levels (*see* **Level of a Test**), the values on this line are the same as those for a **standard normal deviate**. (This is because the distribution of t approaches the standard normal distribution as the df parameter approaches infinity.)
4. Similarly, tables of the F distribution typically have lines at the bottom that show ∞ for the denominator df. For given significance levels and for ν, ∞ as the df, the F values on these lines are the same as the values for $\chi^2(\nu)/\nu$ that appear in a table of χ^2 divided by its df. [Here the notation $\chi^2(\nu)$ refers to a χ^2 variable with df $= \nu$.]
5. The distribution of F with df $= 1, \nu$ is the same as the distribution of the square of a t variable

with df $= \nu$. In reporting results in a publication, however, it is better and more informative to show the value of $t(\nu)$ rather than the value of $F(1, \nu)$, because the former has a plus or minus sign that indicates the direction of the effect whereas the latter is always positive.

6. The distribution of χ^2 with df $= 1$ is the same as the distribution of the square of a standard normal deviate. In reporting results, though, it is more informative to show the standard normal deviate (with its plus or minus sign indicating direction) rather than $\chi^2(1)$.

References

[1] Anderson, R.L. & Bancroft, T.A. (1952). *Statistical Theory in Research*. McGraw-Hill, New York.

[2] Nass, C.A.G. (1959). The χ^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism, *Biometrika* **46**, 365–385.

[3] Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components, *Biometrics Bulletin* **2**, 110–114.

RICHARD F. POTTHOFF

Delayed Entry

Analysis of survival data with delayed entry has in principle been known for centuries, since any **life-table** construction involves following persons from an entrance age to an exit age and registering whether exit is due to death or end of observation for other reasons (**censoring**, in modern terminology). Kaplan & Meier [16] briefly mentioned the validity of their product-limit estimator (*see* **Kaplan–Meier Estimator**) also under delayed entry, and Cox & Oakes [5, Section 11.6] gave a brief, but highly informative survey. However, the topic is absent from the authoritative texts by Kalbfleisch & Prentice [15] and Fleming & Harrington [8].

I first introduce the two main approaches to studying delayed entry: *left* **truncation**, or complete observation of a conditional distribution (next section), and *left filtering*, or observing events only when an observation switch is "on" (following section). In spite of the very different conceptual foundations of the ideas of left truncation and left filtering, the modifications of many non- and semiparametric hazard-based survival analysis estimators to delayed entry situations are exactly the same, essentially consisting in modifying the **risk sets** to only include individuals that have entered. As a consequence, also the calculations are the same.

For both left truncation and left filtering, the possibility of defining a concept of *independent* delayed entry is discussed. Conditionally independent delayed entry given **covariates**, as well as the special role of the **Cox regression model** for studying delayed entry, are then outlined. Deviation from independent truncation may happen if there is association between truncation time and survival time not accounted for by observable covariates, as briefly surveyed in a later section.

I then discuss the special delayed-entry problems in the epidemiologic *prevalent cohort* study, in which a cross sectional sample of persons with a certain disease is followed up. At entry to the study the patients will already have had the disease for a certain time. When this current duration is known, analysis by delayed entry at sampling is possible and can be shown to be valid if the statistical model is rich enough. I also briefly comment on the associated **length-bias** problems, and discuss what to do if current duration is not observed (for fuller discussion, *see* **Biased Sampling of Cohorts**).

The next section gives a nonstandard example, employing delayed-entry methods to obtain a faster confirmatory test in a clinical trial with staggered entry. The final section makes some brief comments on the relation to left censoring, right truncation and the retro-hazard, and mentions as an example retrospectively collected time-to-pregnancy data.

Random Left Truncation

Let V and X be independent positive **random variables** with density functions $g(v)$ and $f(x)$, distribution functions $G(v)$ and $1 - S(x)$ and **hazards** $\gamma(v) = g(v)/[1 - G(v)]$ and $\varphi(x) = f(x)/S(x)$. We call $S(x)$ the survival function of X (*see* **Survival Distributions and Their Characteristics**). The random truncation model [19, 29, 33] considers n independent replications $(V_1^*, X_1^*), \ldots, (V_n^*, X_n^*)$ from the conditional distribution of (V, X) given $V < X$. An important property (*independent truncation*) of the random truncation model is that the *the hazard of X given $X > V = v$ at $x > v$ equals the hazard of X at x*: in heuristic notation, for $x > v$,

$$\begin{aligned}\frac{\Pr\{X = x | V = v, X > V\}}{\Pr\{X \geq x | V = v, X > V\}} &= \frac{\Pr\{X = x, V = v\}}{\Pr\{X \geq x, V = v\}} \\ &= \frac{\Pr\{X = x\}}{\Pr\{X \geq x\}} \\ &= \phi(x).\end{aligned}$$

This property is the key to the simple result that the **nonparametric maximum likelihood estimator** $S(x)$ is given by the product-limit estimator

$$\hat{S}(x) = \prod_{X_i \leq x} \left(1 - \frac{1}{Y(X_i)}\right),$$

where $Y(u)$ is the number at risk at u.

This result was briefly mentioned by Kaplan & Meier [16] under the interpretation of counting the delayed entrants as "negative losses" and further discussed and elaborated in the three references just mentioned.

An unsatisfactory aspect of the random truncation model is its formulation in terms of the "latent" random variables (V_i, X_i) which remain unobserved when $V_i > X_i$. It was pointed out by Wellek [32] and Tsai [25] that it is sufficient to require that the conditional density of (V, X) given $V < X$ may be

written as $f(x)g^*(v)$ for $v < x$. Indeed, the above calculation then yields (for $x > v$)

$$\frac{\Pr\{X = x | V = v, X > V\}}{\Pr\{X \geq x | V = v, X > V\}} = \frac{f(x)g^*(v)}{\int_x^\infty f(u)g^*(v)\,\mathrm{d}u} = \varphi(x).$$

The condition for independent truncation may therefore be formulated in terms only of the "observable" area $\{v < x\}$.

The above considerations rather immediately generalize to also accommodating right censoring (cf. [16, 25, 26], and [32]).

Tsai [25] derived a formal *test for independence* of truncation time and survival time, also valid under independent censoring. Tsai's test generalized Kendall's τ (based on the number of concordant and discordant pairs of observations (V_i, X_i), (V_j, X_j)) to a "conditional Kendall's τ" that is estimable from the conditional distribution of (V, X) given $(V \leq X)$ (*see* **Rank Correlation**). Kalbfleisch & Lawless [14] and Jones & Crowley [12] provided important additional discussion.

The Counting Process Approach and Filtering

The counting process approach to survival analysis allows an alternative approach to delayed entry and to the formalization of independence between survival and delay, directly generalizing the concept of *independent censoring*: see the article **Censored Data** or Andersen et al. [3, Chapter III].

For n independent, identically distributed uncensored survival times (random variables) $X_1, \ldots, X_n$ with survival function $S(x)$ and hazard $\varphi(x)$, define the *counting process*

$$N(t) = \sum_{i=1}^n N_i(t) = \sum_{i=1}^n I\{X_i \leq t\}.$$

With respect to the so-called self-exciting family of σ-algebras $(\mathcal{N}_t)$, $\mathcal{N}_t = \sigma\{N(u) : 0 \leq u \leq t\}$, $N(t)$ has the compensator

$$\int_0^t \varphi(u)Y(u)\,\mathrm{d}u,$$

with

$$Y(t) = \sum_{i=1}^n Y_i(t) = \sum_{i=1}^n I\{X_i \geq t\},$$

which means that the difference between $N(t)$ and the compensator is an $(\mathcal{N}_t)$-martingale. The integrand $\varphi(t)Y(t)$ is called the *intensity process*.

Now assume that observation is partially inhibited by some further noise, formalized by the concept of *filtering processes*, **stochastic processes** $C_i(t)$ on $[0, \infty)$ assuming the values 1 or 0 as observation is "on" or "off" and predictable with respect to an increasing family $(\mathcal{G}_t)$ of σ-algebras such that $\mathcal{N}_t \subseteq \mathcal{G}_t$ for all t. Then

$$N_i^c(t) = \int_0^t C_i(u)\,\mathrm{d}N_i(u) = I\{X_i \leq t, C_i(X_i-) = 1\}$$

is again a counting process, indicating whether the event has taken place before time t *and while observation is "on"*.

Independent filtering will be taken to mean that the intensity process of $N(t)$ is $\varphi(t)Y(t)$ also with respect to the larger family of σ-algebras $(\mathcal{G}_t)$ that includes information on the filtering. Intuitively, the intensity when individuals are observed should be the same as when they are not observed. Under this assumption, by stochastic integration theory

$$N_i^c(t) = \int_0^t C_i(u)\,\mathrm{d}N_i(u) = \int_0^t C_i(u)\varphi(u)Y_i(u)\,\mathrm{d}u + \int_0^t C_i(u)\,\mathrm{d}M_i(u),$$

where the last term is a $(\mathcal{G}_t)$-martingale, so that $N_i^c(t)$ has intensity process $\varphi(u)C_i(u)Y_i(u) = \varphi(u)Y_i^c(u)$ with

$$Y_i^c(u) = I\{C_i(u) = 1, X_i > u\};$$

that is, observation is on and the event has not happened. As before, we may aggregate: $N^c(t) = \sum N_i^c(t)$ has intensity process $\varphi(t)\sum Y_i^c(t)$, where $\sum Y_i^c(t)$ is the number at risk under the filtering.

Right censoring at U_i corresponds to $C_i(t) = I\{t \leq U_i\}$ and *left filtering and right censoring* corresponds to $C_i(t) = I\{V_i < t \leq U_i\}$.

With this definition of delayed entry, asymptotic distribution results are immediate and easily interpretable via increasing *number at risk*. This approach was advocated by Aalen [1, 2].

Kaplan & Meier [16] in their brief discussion of delayed entry took care to call this left *truncation*, in contrast to the then relatively newly specified concept of *censoring* (so termed by Hald [10, 11]). Still, the

spirit of Kaplan & Meier's idea seems closer to left filtering.

Consider now, for ease of exposition, a single individual. A minimal family $(\mathcal{G}_t)$ for independent left filtering only, with no right censoring, is then specified by

$$\mathcal{G}_t = \sigma(I\{V \le t\}, VI\{V \le t\}, I\{X \le t\}, XI\{X \le t\}).$$

The intensity process of $N_t = I\{X \le t\}$ with respect to $(\mathcal{G}_t)$ is $\Pr\{N_t - N_{t-} = 1|\mathcal{G}_{t-}\}$, and here are three cases. First, $\Pr\{N_t - N_{t-} = 1|X < t\} = 0$; secondly,

$$\begin{aligned}&\Pr\{N_t - N_{t-} = 1|X \ge t, V = v < t\}\\ &\quad = \Pr\{X = t|X \ge t, V = v < t\} = \varphi(t)\end{aligned}$$

as derived above; and thirdly,

$$\begin{aligned}&\Pr\{N_t - N_{t-} = 1|X \ge t, V \ge t\}\\ &\quad = \frac{\displaystyle\int_t^\infty f(t)g^*(v)\,\mathrm{d}v}{\displaystyle\int_t^\infty\int_t^\infty f(u)g^*(v)\,\mathrm{d}u\,\mathrm{d}v} = \varphi(t);\end{aligned}$$

so that altogether the intensity process

$$\begin{aligned}&\Pr\{N_t - N_{t-} = 1|\mathcal{G}_{t-}\} = \varphi(t)I\{X \ge t\}\\ &\quad = \Pr\{N_t - N_{t-} = 1|\mathcal{N}_{t-}\},\end{aligned}$$

as was to be proved. Hence it is seen that under the assumption that the conditional density of (V, X) given $V < X$ may be written as a product $g^*(v)f(x)$, left filtering is independent.

Delayed Entry and Covariates

If the lifetime X and entry time V are conditionally independent given a covariate Z (but possibly marginally dependent), then the filtering specified by delayed entry at V is still independent in the model for the *conditional intensity* given Z. For example, we then have, heuristically

$$\begin{aligned}&\frac{\Pr\{X = x|V = v, Z = z, X > V\}}{\Pr\{X \ge x|V = v, Z = z, X > V\}}\\ &\quad = \frac{\Pr\{X = x|Z = z\}}{\Pr\{X \ge x|Z = z\}}.\end{aligned}$$

This means that the conditional intensity given Z may be validly estimated also under "conditionally independent" delayed entry.

In particular, the Cox regression model specifies the dependence on time nonparametrically, and the estimation works with "numbers at risk" in the same way as for the fully nonparametric techniques. Delayed entry may therefore be handled – by truncation or filtering – as specified in the two previous sections. Cnaan & Ryan [4] compared this "correct" approach with other approaches and called attention to special problems with time-dependent covariates: see further Keiding & Knuiman [20] and the detailed discussion by Wang et al. [28].

Dependent Delayed Entry: Two Examples

Keiding [18] gave two examples where the assumption of independent delayed entry is violated.

In the first, assume that, given a random variable $Z = z$, V and X are independent, and for simplicity **exponentially distributed** with intensities γz and φz, respectively. The frailty Z is assumed to be an unobserved **gamma-distributed** random variable. A direct calculation then documents that the hazard of X given $X > V = v$ at $x > v$ differs from the marginal hazard of X at x. In this case, *unobserved heterogeneity* is the culprit.

The other example regarded *differential selection into observation* from two states with different hazards of death (such as a *healthy* state and a *diseased* state). A numerical example showed that the net effect of the differential selection can go either way depending on the concrete values of selection and death intensities.

The Prevalent Cohort Study

A cross sectional sample of patients suffering from a certain disease is taken at a particular calendar time t_V and followed up, usually for some fixed time interval.

Assume first that the age of onset of disease Y is known for each patient; let V be age at entry. The calendar time-, age- and duration-specific death intensity $\nu(t, a, d)$ may then be estimated by considering the patients as having delayed entry (with delay $Y - V$) at t_V (and possibly also right censoring at the end of the follow-up interval). Keiding [18] gave a formal calculation within the framework of illness–death processes (*see* **Stochastic Processes**) to show that this delayed entry is independent in the model for the $\nu(t, a, d)$. With $X =$ age at death,

T = time at death, $T_V = T - (X - V)$ time at entry, the hazard of X in the conditional distribution given $(Y = y, T = t, T_V = t_v, X > V)$ was shown to equal $\nu(t, x, x - y)$.

Keiding [18] also compared the delayed-entry estimator to two other possibilities: the *length-biased* estimator relevant if patients are counted from disease onset (see also Wang [27] and the article on **Length Bias**); and the *forward recurrence time estimator* [6] based on time from sampling to death only, not requiring knowledge of age at onset. The latter two possibilities require very strong stationarity assumptions and would rarely be justified in biomedical applications.

There is an extensive literature, primarily motivated by **AIDS**, on how to approach prevalent cohort studies when age at disease onset is unknown (for a survey, *see* **Biased Sampling of Cohorts**).

Example: Confirmatory Analysis of an Unexpected Finding at Interim Analysis

As an example, I quote an analysis regarding the prognostic significance of residual cancer tissue after diagnostic biopsy in breast cancer. The trials of the Danish Breast Cancer Cooperative Group (DBCG) were started during 1978 and, based upon the experience until December 31, 1981, a negative effect on recurrence-free survival of the presence of residual cancer tissue (RCT) after diagnostic biopsy was noted for premenopausal patients considered to be at high risk based on histological findings. The diagnostic biopsy is a tissue specimen that is examined by the pathologist for presence of malignant (i.e. cancerous) cells. If such cells were found, the whole breast was removed shortly thereafter, usually within 0.5–1.0 hours after the biopsy was taken, and only these patients are included in this study. It was considered unexpected, and not easily interpretable, to find a connection between the presence of cancer tissue in the biopsy cavity and recurrence-free survival, and a reanalysis on an independent set of data was therefore judged necessary before the finding could be considered an established fact within breast cancer prognostics. Ordinarily one would use patients accrued *after* January 1, 1982. Since accrual to this protocol was closed toward the end of 1982, only rather few patients were available, and even an average follow-up time of almost four years was not enough to reproduce the early finding. Alternatively, the recurrence-free survivors on January 1, 1982, might be included, counted with *delayed entry* with the duration obtained on that date: these patients are included with *left truncated* disease durations [21] (for full surgical discussion, see Watt-Boolsen et al. [31]).

Assuming that there are no hidden heterogeneities, the simple delayed entry correction provides a valid test based on the experience after January 1, 1982, conditional on survival until then for those accrued earlier (see Keiding et al. [21]). In fact, there were significant covariates (age at operation, pathoanatomical characteristics of the tumor; see Watt-Boolsen et al. [31]); in unpublished analyses by T. Bayer and N. Keiding these were included in the delayed entry approach through a Cox regression model, which did not change the qualitative conclusions.

Parner and Keiding [23] showed, under a broad class of model misspecifications, that the independence between interim and confirmatory analyses is robust.

Left Censoring, Right Truncation, Time-reversal, and the Retro-hazard

It is important to distinguish delayed entry (where the individual under study is only identified if the event (such as death) occurs after the entry time) from *left censoring*, where all individuals are counted, but for some of them it is only known that the event took place before a certain censoring time.

Technically, left censoring can sometimes be handled by the well-established hazard rate and risk set-based survival analysis techniques in *reverse time*; but examples of left censoring are rare in survival analysis – see, however, Ware & DeMets [30] and Andersen et al. [3, Examples I.3.7 and IV.3.5].

Right truncation can also often be handled by reversing time and focusing on the *retro-hazard* $\overline{\varphi}(x) = f(x)/F(x)$; see, for example, Lagakos et al. [22], Kalbfleisch & Lawless [13], Keiding & Gill [18], Keiding [17] and Gross & Huber-Carol [9]. Right truncation has become an important biostatistical tool in connection with **retrospective** epidemiologic observation plans particularly in studying AIDS, where only those individuals who contracted the disease before a certain calendar date are included. see Esbjerg et al. [7] for an application in neuroepidemiology (*see* **Truncated Survival**

Times). Also when retrospectively observing **time-to-pregnancy** it is quite common to observe only those who gave birth before study completion, necessitating correction for right truncation particularly in studying calendar time effects [24].

References

[1] Aalen, O.O. (1975). Statistical Inference for a Family of Counting Processes, *Ph.D. thesis*, University of California, Berkeley.

[2] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[3] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Cnaan, A. & Ryan, L. (1989). Survival analysis in natural history studies of disease, *Statistics in Medicine* **8**, 1255–1268.

[5] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[6] Denby, L. & Vardi, Y. (1986). The survival curve with decreasing density, *Technometrics* **28**, 359–367.

[7] Esbjerg, S. Keiding, N. & Koch-Henriksen, N. (1999). Reporting delay and corrected incidence of multiple sclerosis, *Statistics in Medicine* **18**, 1691–1706.

[8] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[9] Gross, S.T. & Huber-Carol, C. (1992). Regression models for truncated survival data, *Scandinavian Journal of Statistics* **19**, 193–213.

[10] Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Skandinavisk Aktuarietidskrift* **32**, 119–134.

[11] Hald, A. (1952). *Statistical Theory with Engineering Applications*. Wiley, New York.

[12] Jones, M.P. & Crowley, J. (1992). Nonparametric tests of the Markov model for survival data, *Biometrika* **79**, 513–522.

[13] Kalbfleisch, J.D. & Lawless, J.F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS, *Journal of the American Statistical Association* **84**, 360–372.

[14] Kalbfleisch, J.D. & Lawless, J.F. (1991). Regression models for right truncated data with applications to AIDS inoculation times and reporting lags, *Statistica Sinica* **1**, 19–32.

[15] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

[16] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[17] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[18] Keiding, N. (1992). Independent delayed entry, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Dordrecht, pp. 309–326.

[19] Keiding, N. & Gill, R.D. (1990). Random truncation models and Markov processes, *Annals of Statistics* **18**, 582–602.

[20] Keiding, N. & Knuiman, M.W. (1990). Letter to the editor on "Survival analysis in natural history studies of disease" by A. Cnaan and L. Ryan, *Statistics in Medicine* **9**, 1221–1222.

[21] Keiding, N., Bayer, T. & Watt-Boolsen, S. (1987). Confirmatory analysis of survival data using left truncation of the life times of primary survivors, *Statistics in Medicine* **6**, 939–944.

[22] Lagakos, S.W., Barraj, L.M. & DeGruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS, *Biometrika* **75**, 515–523.

[23] Parner, E. & Keiding, N. (2001). Misspecified proportional hazards models and confirmatory analysis of survival data, *Biometrika* **88**, 459–468.

[24] Scheike, T.H. & Jensen, T.K. (1997). A discrete survival model with random effects: an application to time to pregnancy, *Biometrics* **53**, 318–329.

[25] Tsai, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time, *Biometrika* **77**, 169–177.

[26] Tsai, W.-Y., Jewell, N.P. & Wang, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation, *Biometrika* **74**, 883–886.

[27] Wang, M.C. (1991). Nonparametric estimation from cross-sectional survival data, *Journal of the American Statistical Association* **86**, 130–143.

[28] Wang, M.-C., Brookmeyer, R. & Jewell, N.P. (1993). Statistical models for prevalent cohort data, *Biometrics* **49**, 1–11.

[29] Wang, M.-C., Jewell, N.P. & Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation, *Annals of Statistics* **14**, 1597–1605.

[30] Ware, J.H. & DeMets, D.L. (1976). Reanalysis of some baboon descent data, *Biometrics* **32**, 459–463.

[31] Watt-Boolsen, S., Ottesen, G., Andersen, J.A., Bayer, T., Jespersen, N.C.B., Keiding, N., Mouridsen, H.T., Dombernowsky, P. & Blichert-Toft, M. (1989). Significance of incisional biopsy in breast carcinoma: results from a clinical trial with intended excisional biopsy, *European Journal of Surgical Oncology* **15**, 33–37.

[32] Wellek, S. (1990). A nonparametric model for product-limit estimation under right censoring and left truncation, *Communications in Statistics – Stochastic Models* **6**, 561–592.

[33] Woodroofe, M. (1985). Estimating a distribution function with truncated data, *Annals of Statistics* **13**, 163–177; correction **15**, (1987). 883.

NIELS KEIDING

Delta Method

The delta method is really a theorem which states that a smooth function of an asymptotically normal estimator is also asymptotically normally distributed (*see* **Large-sample Theory**). The result is applied in numerous contexts for the computation of large-sample tests and confidence limits for nonlinear functions of parameters which have already been estimated. Typically, the method of **estimation** is a standard large-sample technique which cannot be directly applied to the problem of interest.

Let $\hat{\theta}_n$ denote a sequence of estimates of some parameter θ, such that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{L}} \text{N}[0, \sigma^2(\theta)] \tag{1}$$

(*see* **Convergence in Distribution and in Probability**). For example, $\hat{\theta}_n = \theta(X_1, X_2, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a random sample from a distribution F_θ, the simplest cases being $\hat{\theta}_n = \overline{X}_n$, the sample **mean**, with $\theta = \mu_x$, or $\hat{\theta}_n = s_n^2$, the sample **variance**, with $\theta = \sigma_x^2$. Let g be a function which is differentiable in a neighborhood of the true value θ with $g'(\theta) \neq 0$. Then the delta method states that

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{\text{L}} \text{N}\{0, \sigma^2(\theta)[g'(\theta)]^2\}. \tag{2}$$

The result follows from (1) using standard convergence theorems [14, 2c.4]. Briefly, substitution of $\hat{\theta}_n$ into a Taylor expansion for g about θ yields

$$g(\hat{\theta}_n) = g(\theta) + (\hat{\theta}_n - \theta)g'(\theta) + (\hat{\theta}_n - \theta)o_p(1),$$

where $o_p(1) \xrightarrow{\text{p}} 0$. Then

$$\begin{aligned}\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) - \sqrt{n}(\hat{\theta}_n - \theta)g'(\theta)\\ = \sqrt{n}(\hat{\theta}_n - \theta)o_p(1) = o_p(1).\end{aligned}$$

In fact, a slightly stronger result is needed, in which we assume that g' and σ^2 are continuous at θ, $\sigma(\hat{\theta}_n)|g'(\hat{\theta}_n)|$ is used to standardize (2), and convergence is to a standard normal distribution [14, 6a.2; 2, 12.1.2]. In practice, the **standard error** (se) of $\hat{\theta}_n$ is taken to be $\sigma/\sqrt{n}$, and the se of $g(\hat{\theta}_n)$ is $|g'(\theta)|\sigma(\theta)/\sqrt{n}$, with θ estimated by $\hat{\theta}_n$. Thus the delta method can also be viewed as a technique for approximating the mean and variance of a function of a random variable, $g(T)$ [10]. For this approximation to be valid we must have $\text{var}(T) = O(1/n)$. The Taylor expansion can also be used to provide a bias correction [8, 8.4(iii)].

In applications, a multivariate version of the theorem is typically needed. Suppose that $\hat{\theta}_n = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is an asymptotically [in the sense of (1)] multivariate normal random vector with asymptotic mean θ and variance matrix $\Sigma(\theta)$. Let $\mathbf{g}(\theta) = [g_1(\theta), \ldots, g_q(\theta)]'$ and $\partial\mathbf{g}/\partial\theta$ denote the $q \times k$ matrix of partial derivatives. Then, if for each i, $\partial g_i/\partial\theta_j \neq 0$ for some j, we have

$$\sqrt{n}[\mathbf{g}(\hat{\theta}_n) - \mathbf{g}(\theta)] \xrightarrow{\text{L}} \text{N}\left(\mathbf{0}, \frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}} \sum \frac{\partial\mathbf{g}'}{\partial\boldsymbol{\theta}}\right). \tag{3}$$

The proof involves either a multivariate Taylor expansion [1, Appendix C; [4], 14.6.3] or, alternatively, an application of (2) using an arbitrary linear combination of the coordinates of $\mathbf{g}(\hat{\theta}_n)$ and well-known characterizations of the **multivariate normal distribution** and of convergence in distribution of random vectors [14, 6a.2]. In most applications $q = 1$.

The delta method is closely related to the method of **maximum likelihood**. As is well known, if $\hat{\theta}$ is the maximum likelihood estimator (MLE) of θ and $\mathbf{g}$ is a one-to-one differentiable transformation, then $\hat{\phi} = \mathbf{g}(\hat{\theta})$ is the MLE of $\phi = \mathbf{g}(\theta)$. A Taylor expansion of the score vector (*see* **Likelihood**) may also be used to show that the **information matrix** becomes [6; [8], Exercise 4.15]

$$I(\phi) = \frac{\partial\theta'}{\partial\phi} I(\theta) \frac{\partial\theta}{\partial\phi}.$$

Thus the asymptotic distribution of $\hat{\phi}$ obtained from the delta method is the same as that of the MLE. Indeed, in single parameter **exponential families**, where there is a **sufficient statistic** for the natural parameter (such as the **multinomial distribution**) the method of maximum likelihood can be regarded as an application of the delta method using the transformation implicitly defined by the score equations [5, 3, 2, 12.2.1]

Two classes of applications may be distinguished. The first involves choosing a transformation $g(\theta)$ so that $\text{var}[g(\hat{\theta})]$ = constant, in which case g is known as a variance-stabilizing transformation. From the delta method,

$$g(x) = c \int \frac{\text{d}\theta}{\sigma(\theta)}.$$

Examples include the angular (arc sin square root) transformation for a binomial proportion, and Fisher's z transformation for the sample correlation coefficient, r, [14, 6g.4]. The multivariate delta method can be used to establish the asymptotic normality of r [15, Chapter 3]; a similar technique can be used to show the asymptotic normality of the sample variance, s^2. Another general class of examples is given by the ladder of powers, $\text{var}(X) = c^2\mu^{2\alpha}$ for $\alpha > 0$ (*see* **Power Transformations**). In this case

$$g(x) = \frac{1}{c(1-\alpha)}\mu^{1-\alpha},$$

with the understanding that $\alpha = 1$ (constant coefficient of variation) gives $g(x) = \log x$. Other special cases include the square root transformation ($\alpha = 1/2$), which is used with the **Poisson distribution**. Strictly speaking, if $X_n \sim \text{Poisson}\,(n\theta)$, then the delta method must be applied to the sequence $\hat{\theta}_n = X_n/n$ [4, Example 14.6-3; [15], p. 121].

The second class of applications involves nonlinear functions of parameters for which large-sample estimates can easily be obtained, often from the **central limit theorem**, or from a **generalized linear model**. The simplest example is provided by the asymptotic normality of the sample **standard deviation** $s = \sqrt{s^2}$. The standard example is the variance of a ratio:

$$\text{var}\left(\frac{T_1}{T_2}\right) \approx \left[\frac{\text{E}(T_1)}{\text{E}(T_2)}\right]^2 \left\{\frac{\text{var}(T_1)}{[\text{E}(T_1)]^2} - \frac{2\text{cov}(T_1, T_2)}{\text{E}(T_1)\text{E}(T_2)} + \frac{\text{var}(T_2)}{[\text{E}(T_2)]^2}\right\}.$$

An application is the estimation of the dose corresponding to a given frequency of response in the analysis of **quantal response** data in toxicology [13, 2.7.1]. Related applications include the variance of the log of the **relative risk**, $rr = p_1/p_2$, and the log **odds ratio** (the logarithms of these quantities being more nearly normally distributed) in epidemiology [11, 15.5]. For example,

$$\text{var}[\log(rr)] \approx \frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}.$$

Another interesting class of examples involves the calculation of large-sample standard deviations for various measures of **association** in two-way contingency tables. These have the form $\zeta = \nu(\pi_{ij})/\delta(\pi_{ij})$, where $\nu(\pi_{ij})$ and $\delta(\pi_{ij})$ are known functions of the population proportions [1, 10.3; [4], 11.3]. A special case is the measurement of **agreement** between two raters on a categorical scale. The standard measure of interrater reliability is the observed proportion of agreement corrected for chance, known as **kappa**. A large sample standard deviation may be computed using the delta method [4, 11.4; [9], 13.1].

In multinomial regression models, the delta method can be used to establish the asymptotic distribution of the predicted cell probabilities and residuals, typically standardized cell **residuals** [2, 12.2–12.3]. Cox & Ma [7] used a similar application to develop confidence bands for generalized **nonlinear regression** models. For extensions of the basic result ($g'(\theta) = 0$ and $\sigma_n \to 0$ instead of $\sigma/\sqrt{(n)} \to 0$), see [15, Chapter 3].

Lehman [15, Section 6.3] considers the extension to limit distributions of statistical functionals.

References

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

[2] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[3] Benichou, J. & Gail, M.H. (1989). A delta method for implicitly defined random variables, *American Statistician* **43**, 41–44.

[4] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[5] Cox, C. (1984). An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method, *American Statistician* **38**, 283–287.

[6] Cox, C. (1990). Fieller's theorem, the likelihood, and the delta method, *Biometrics* **46**, 709–718.

[7] Cox, C. & Ma, G. (1995). Asymptotic confidence bands for generalized nonlinear regression models, *Biometrics* **51**, 142–150.

[8] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[9] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[10] Johnson, N.L. & Kotz, S. (1988). *Encyclopedia of Statistical Sciences*, Vol. 8. Wiley, New York, pp. 646–647.

[11] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research*. Lifetime Learning Publications, Belmont.

[12] Lehman, E.L. (1999). *Elements of Large Sample Theory*. Springer-Verlag, New York.

[13] Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. Chapman & Hall, London.

[14] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. Wiley, New York.

[15] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

C. COX

Demography

Demography is the study of human populations with respect to their size, structure, and dynamics. For demographers, a population is a group of individuals that coexist at a point in time and share a defining characteristic such as residence in the same geographical area. The structure or composition of a population refers to the distribution of its members by age, sex, and other characteristics, such as place of residence and marital or health status. The age and sex structure of a population results from past trends in fertility, mortality, and migration. Thus, these processes comprise the components of demographic change. The age and sex structure of a population, in turn, affects birth rates, death rates, and rates of migration. Changes in status such as getting married or divorced interact with population structure in a similar way.

Some authorities reserve the term demography for the mathematical and statistical study of the interrelationships between population size and structure and the components of demographic change. According to this terminology, demography can be contrasted with population studies, which investigate the determinants and consequences of demographic phenomena drawing on the concepts and theories of disciplines such as the **social sciences**, health sciences, and history. Others encompass population studies within demography and use the term *formal demography* to distinguish the statistical core of the discipline. Demography (according to this wider definition) is a multidisciplinary field: subdisciplines such as economic demography, historical demography, anthropological demography, and mathematical demography exist. They differ not only in their subject of study but also in their theoretical orientation and methods.

The term demography has been ascribed to a Belgian statistician, Achille Guillard, who coined it in 1855. However, the origins of modern demography are usually traced back to **John Graunt**'s quantitative analyses of the "Bills of Mortality" published in 1662 [5]. The "Bills of Mortality" provided weekly lists of burials and baptisms in the parishes of London. Graunt used these data to examine the sex ratio at birth and to estimate the population of London. He showed that more deaths than births occurred in London, implying that the growth of the capital was due to in-migration from the countryside. He also estimated the proportion of births surviving to a range of ages, thereby developing the basic concept of the **life table**. Graunt's research prefigures modern applications of demographic science: information on fertility and mortality and population estimates for small areas (*see* **Small Area Estimation**) remain the fundamental results of demographic analysis required by those engaged in policy formulation and planning.

Demographic Data

In most developed countries, civil registration of births and deaths is the primary source of fertility and mortality data. Government agencies routinely collect demographic information when births and deaths are certified for administrative purposes (*see* **Vital Statistics, Overview**, Overview). The primary source of data on the size, structure and distribution of national populations is the population census. **Censuses** aim to enumerate the whole population of a defined geographical area. They collect individual-level data on the population's characteristics that refer to a single point in time. As well as collecting data on the size and composition of the population, most censuses also ask about moves in a fixed period of time before the enumeration. In countries where vital statistics data are incomplete, questions may also be asked about fertility and mortality. Countries that issue identity numbers and require their population to report their place of residence can maintain continuous population registers. In a few European countries these registers now fuse the functions of the registration system with those of the census.

The evolution of demographic analysis and of routine collection of data on populations by the government were interlinked. Standard demographic measures and techniques of analysis were developed largely for the study of vital statistics with census-based denominators. In recent decades, however, demographers have relied increasingly on survey data to supplement those from traditional sources (*see* **Surveys, Health and Morbidity**). In particular, in countries where registration of vital events is incomplete national sample surveys are the main source of vital statistics. One of the first subjects to be investigated in demographic surveys was family planning

(*see* **Reproduction**). Other early surveys collected women-based fertility histories to supplement the event-based data generated by birth registration. Fertility history and family planning data remain the focus of many demographic surveys, including the two major international programs of surveys conducted since the 1970s, namely the World Fertility Survey and the Demographic and Health Surveys.

Issues

Between the mid-nineteenth and mid-twentieth centuries the more developed regions of the world went through a *demographic transition* from a high-fertility, high-mortality, and low-growth demographic regime to a low-fertility, low-mortality, and low-growth demographic regime. As mortality tended to fall before fertility, this transition was marked by rapid population growth. Since 1945, a similar transition has begun in most less developed countries. As a result, the world's population has grown from about 2.5 billion to about 6 billion in the second half of the twentieth century. It is expected to grow to between 9 and 16 billion by 2100.

Efforts to understand the determinants of the transition of fertility and mortality to low levels are a central concern of demography. Many demographers now believe that explanations that focus on economic factors and the provision of health and family planning programs are inadequate and need to be supplemented by accounts that take into account the ideational and cultural determinants of demographic behavior.

Thomas Malthus was the first author to develop a systematic argument that high fertility leading to **population growth** could have adverse effects on economic welfare [7]. He argued that a growing population must eventually outstrip its subsistence base, bringing about rising mortality from famine, pestilence, and war. Although the past two centuries of human history have followed a very different path from that envisaged by Malthus, concern still exists about the impact of population growth on economic development and the environment. Today, however, economic demographers tend to be more sanguine about the consequences of population growth than those with a background in ecology [3].

Many demographic outcomes are of concern to policy makers and much demographic research has an avowedly applied intent. Demography bears on the efforts of international agencies and national governments to promote family planning and improve health in the developing world. In the developed world, population growth has slowed but low fertility and the reduction in death rates in old age are producing an increasingly aged population. Recent changes in patterns of marriage and divorce and of childbearing inside and outside marriage also have major implications for the family and public policy.

Demographic studies of mortality tend to focus on the analysis of routine data. Demographers' research into health and mortality cannot be distinguished clearly from that of epidemiologists. However, demographers tend to be concerned with the distribution of disease and premature death (*see* **Descriptive Epidemiology**) across social groups (*see* **Social Classifications**) and their implications for other aspects of social life, rather than with measuring risk factors for specific conditions.

Demographic Analysis

The aim of formal demographic analysis is to isolate the components of demographic patterns by dividing a population into relatively homogeneous subgroups. Analysis by age and sex has primacy over analysis by other compositional factors. Human biology causes the propensity to die and to give birth to be differentiated by age and sex everywhere. It imposes a degree of uniformity on age patterns of mortality and fertility in all human populations.

Classical demographic analysis is based on a fairly small set of measures and techniques. Most of these are also used in cognate disciplines. Calculation of **rates**, ratios, and proportions represent the basic way for controlling for population size. In demography, rates calculated for the whole population that make no allowance for the influence of population structure on the phenomenon of interest are referred to as *crude rates*. Examples are the crude birth rate and crude death rate (*see* **Vital Statistics, Overview**).

Calculation of age-specific rates and rates specific to other subgroups of the population allow the analyst to isolate the propensity to experience the event being studied from the influence of population structure. A range of methods of standardization are used to produce synthetic indices that summarize such specific rates (*see* **Standardization Methods**). The distinction between cohort analysis and **cross-sectional**

or period analysis is fundamental to demography. Demographers use the term *cohort* to refer to groups of individuals who experience a defining event at the same time. Examples include **birth cohorts** and marriage cohorts. Cohort analysis studies the subsequent experience of such groups. This contrasts with epidemiologic usage, which refers to all those eligible for recruitment into a longitudinal study as a **cohort**.

Period measures are often treated as referring to a synthetic or hypothetical cohort, so that summary indices can be calculated that indicate what would happen to a cohort that went through life experiencing the specific rates of the period under study. For example, the most widely used index of period fertility is the *total fertility rate*. This measures how many children women would bear on average if they went through life with the fertility of a specific period. It is calculated by summing the age-specific fertility rates of a particular year, usually for five-year age groups, over all ages at which women bear children. The total fertility rate is thus a form of directly standardized rate, calculated using a uniform age distribution as the standard.

Two basic aspects of any demographic process are its intensity, or quantum, and its timing, or tempo. The intensity of a nonrenewable event such as death or first marriage can be measured by the proportion of a cohort who eventually experience the event. Both the expected timing of any nonrenewable process and the distribution of times of its occurrence can be studied using **life table** methods. The intensity of a renewable process such as birth or disease incidence can be measured by the **mean** number of events per person, and their tempo by the characteristics of the distribution of the events in time. Renewable events can be categorized by the order of their occurrence, and events of a particular order can be analyzed as a nonrenewable process. For example, the proportion of women who have had a birth of order i that go on to bear a child of order $i + 1$ is known as a *parity progression ratio*.

Investigation of the determinants of fertility and mortality has been facilitated by making a distinction between proximate and distal determinants. The approach is most developed with respect to fertility. A proximate determinant is one that has a direct impact on the outcome of interest while a distal determinant can only affect the outcome via a proximate determinant. The proximate determinants of fertility are those factors that determine a woman's exposure to sexual intercourse, her probability of conceiving, and the probability that a pregnancy ends in a live birth. The strength of the approach is that only a few of the proximate determinants of individuals' fertility differ between populations in their impact at the aggregate level. Thus, the four main proximate determinants of fertility differences between groups and over time are the proportion of women in sexual unions, postpartum infecundity associated with breast-feeding, contraception, and abortion. Socioeconomic determinants of fertility must operate through these few proximate factors and a single characteristic may have countervailing effects on fertility via different proximate determinants.

Demographic Models

Analysis of data on actual populations is paralleled by mathematical models of the interrelationship between population size and structure and the components of demographic change. Stable population theory as developed by Lotka in the 1920s and 1930s demonstrates that any closed single-sex population subject to constant fertility and mortality rates converges on an unchanging age structure and a constant rate of growth. This stable outcome is independent of the initial age structure of the population. The special case of a stable population that is unchanging in size is termed a *stationary population*. Its age structure is a function of the life table. Recent developments, known as generalized stable population theory, demonstrate that the mathematics of stable populations can be extended to populations in which growth rates vary by age because of a history of fertility and mortality change and to populations subject to decrements other than mortality [8].

One crucial application of demography is to the forecasting of future population change. This is usually undertaken using cohort-component methods of population projection [2]. These methods provide a precise way of controlling for the influence of population structure and of working out the implications of any scenario postulated for future vital rates. Despite this, population forecasts have often proved wide of the mark. Fertility, mortality, and migration remain difficult to predict. Forecasts informed by a theoretical understanding of the determinants of these components of population change often perform little better than the simple extrapolation of past trends in vital rates.

The increasing availability of survey data and information technology that makes it practicable to undertake individual-level analysis of data on large samples, have facilitated convergence between demographic methods and other forms of statistical analysis. Thus, many of the developments in demographic analysis during the past few decades have been closely linked to those in statistical methods more generally. Demographers have both adopted and contributed to the development of methods such as event history analysis [10], the modeling of unobserved heterogeneity, and **random-effects** models [4]. Other fields of methodologic research in recent years include the extension of life table methods into multistate models that allow for increments as well as decrements from each state [6] and methods and models for the study of families and households [1]. (*see* **Multilevel Models**)

One particularly successful field has been the development of indirect methods for estimating vital rates in populations with limited and defective vital statistics [9]. Indirect methods use stable population theory and its extensions to describe the relationship between conventional indices of fertility, mortality, and migration and items of information that can be collected more reliably in single-round surveys and censuses in less developed countries. For example, it is possible to estimate life table indices of child mortality from data on the proportion of women's children ever-born who have died, tabulated by the age of the women concerned [9].

References

[1] Bongaarts, J., Burch, T. & Wachter, K., eds. (1987). *Family Demography: Methods and their Applications*. Clarendon Press, Oxford.

[2] Brass, W. (1974). Perspectives in population prediction, *Journal of the Royal Statistical Society, Series A*, **137**, 532–583.

[3] Cassen, R. (1994). Overview, in *Population and Development: Old Debates, New Conclusions*, R. Cassen and contributors. Transaction, Oxford.

[4] Goldstein, H. (1995). *Multilevel Statistical Models*. Edward Arnold, London.

[5] Graunt, J. (1964). *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*; London, 1662. Reprinted, with an introduction by B. Benjamin, in *Journal of the Institute of Actuaries* **90**, 1–61.

[6] Land, K.C. & Rogers, A., eds (1982). *Multidimensional Mathematical Demography*. Academic Press, New York.

[7] Malthus, T.R. (1970). *An Essay on the Principle of Population (London, 1798)*, A. Flew, ed. Penguin, Harmondsworth.

[8] Preston, S.H. & Coale, A.J. (1982). Age structure, growth, attrition, and accession: a new synthesis, *Population Index* **48**, 217–259.

[9] United Nations (1983). *Indirect Techniques for Demographic Estimation*. ST/ESA/Series A/81. United Nations, New York.

[10] Yamaguchi, K. (1991). *Event History Analysis*. Sage, London.

(*See also* **Actuarial Methods**)

IAN M. TIMÆUS

Dendrogram

The term "dendrogram" (Greek *dendron*, a tree) is used in **numerical taxonomy** for any graphical drawing or diagram giving a treelike description of a taxonomic system. More generally, a dendrogram is a two-dimensional diagram representing a tree of relationships, whatever their nature.

Some of the earliest examples of dendrograms are the customary phylogenetic trees used by systematists. It seems that the term "dendrogram" was first used by Mayr et al. [13] (see also [16]).

Depending on the nature of the relationships described by the diagram, the term "dendrogram" is sometimes replaced by another, such as phenogram or cladogram. The former is used for a dendrogram representing phenetic, and the latter for that representing cladistic relationships [3, 12].

The representation of a taxonomic system by a dendrogram is particularly suitable in connection with a cluster analysis applied to investigate the structure of the corresponding operational taxonomic units; that is, entities or individuals considered as the lowest-ranking taxa within the system, such as individual patients or case histories initially used to construct disease classifications [16]. This becomes apparent when it is desirable to interpret the results of the analysis in terms of a natural nonoverlapping taxonomic hierarchy. The usual basis for a cluster analysis is a resemblance (proximity) matrix, its rows and columns referring to the operational taxonomic units, and its entries being the estimates (measurements) of the resemblances between the corresponding units (*see* **Cluster Analysis of Subjects, Hierarchical Methods**; **Similarity, Dissimilarity, and Distance Measure**).

There are various ways of drawing a tree diagram to illustrate the fusions and partitions that have been made at each successive level of the cluster analysis applied. The early practice of drawing dendrograms tended to have the branches of the treelike diagram pointing upward or downward. But later, with ever-increasing numbers of operational taxonomic units, it has become more convenient to place the dendrograms, and particularly the phenograms, almost uniformly on their side, with branches running horizontally across the page. The abscissa is then scaled in the resemblance measure on which the clustering has been based, and the points of furcation between stems along the scale imply that the resemblance between two stems is at the similarity coefficient value shown on the abscissa. It should be realized, however, that the order in which the branches of a dendrogram are presented has no special significance, and can be changed within wide limits without actually changing the taxonomic relationships implied by the dendrogram. This multiplicity of ways in which the same relationships can be represented in a dendrogram may be regarded as a disadvantage. In fact, a dendrogram representation of a resemblance matrix is likely to be satisfactory only if the data are strongly clustered and have an hierarchical type of structure. Otherwise, the dendrogram can be very misleading. Several methods have been suggested to overcome this (see, for example, [14]).

Examples of dendrograms and methods of their presentation are described in many textbooks on numerical taxonomy and multivariate analysis (*see* **Cluster Analysis of Subjects, Hierarchical Methods**) (see also [6, 11, 14, 16], and [17]).

Two different clustering methods may lead to different dendrogram representations of the results, even if both methods are based on the same resemblance matrix. Among the various clustering methods, one is particularly relevant to a dendrogram representation. It is the **single linkage** cluster analysis, also known as the nearest-neighbor technique, introduced by Florek et al. [7, 8] and, independently, by McQuitty [10] and Sneath [15]. As shown by Gower & Ross [9], the most efficient procedure for the single linkage cluster analysis is based on producing the shortest dendrite (the minimum spanning tree; see [1]). In fact, the single linkage clusters can be obtained from the shortest dendrite by successively removing its edges, largest first, the second largest next, and so on. The shortest dendrite itself also gives an alternative graphical representation of the single linkage cluster analysis results (see, for example, [14]). This may appear more convenient than the usual application of a dendrogram, particularly when superimposing the dendrite on the operational taxonomic units scattered in an ordination plot of two or three dimensions obtained, for example, from the **principal components analysis** [16] or the canonical variate analysis [1].

Dendrograms can also be applied when using a cluster analysis in order to partition treatment means in the **analysis of variance**, as shown, for example, by Caliński & Corsten [2] (see also [4]).

Several statistical packages include computation procedures for displaying dendrograms: see Digby [5] in particular (*see* **Software, Biostatistical**).

References

[1] Caliński, T. (1982). Dendrites, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 302–305.

[2] Caliński, T. & Corsten, L.C.A. (1985). Clustering means in ANOVA by simultaneous testing, *Biometrics* **41**, 39–48.

[3] Camin, J.H. & Sokal, R.R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* **19**, 311–326.

[4] Corsten, L.C.A. & Denis, J.B. (1990). Structuring interaction in two-way tables by clustering, *Biometrics* **46**, 207–215.

[5] Digby, P.G.N. (1995). Procedure D DENDROGRAM, in *Genstat 5: Procedure Library Manual Release 3[3]*, R.W. Payne, G.M. Arnold & G.W. Morgan, eds. The Numerical Algorithms Group Ltd, Oxford, pp. 115–119.

[6] Everitt, B. (1974). *Cluster Analysis*. Wiley, New York.

[7] Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H. & Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini, *Colloquium Mathematicum* **2**, 282–285.

[8] Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H. & Zubrzycki, S. (1951). Taksonomia wrocławska, *Przegląd Antropologiczny* **17**, 193–211.

[9] Gower, J.C. & Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis, *Applied Statistics* **18**, 54–64.

[10] McQuitty, L.L. (1957). Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies, *Educational and Psychological Measurement* **17**, 207–229.

[11] Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, London.

[12] Mayr, E. (1965). Numerical phenetics and taxonomic theory, *Systematic Zoology* **14**, 73–97.

[13] Mayr, E., Linsley, E.G. & Usinger, R.L. (1953). *Methods and Principles of Systematic Zoology*. McGraw-Hill, New York.

[14] Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

[15] Sneath, P.H.A. (1957). The application of computers to taxonomy, *Journal of General Microbiology* **17**, 201–226.

[16] Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy*. W.H. Freeman, San Francisco.

[17] Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. W.H. Freeman, San Francisco.

(*See also* **Classification, Overview**; **Cluster Analysis, Variables**; **Multidimensional Scaling**; **Pattern Recognition**; **Projection Pursuit**)

TADEUSZ CALIŃSKI

Denominator Difficulties

Epidemiologists often report data in terms of **rates** and proportions. These measurements require denominators (defined by Last [3] as the lower portion of a fraction used to calculate a rate or ratio) as well as numerators (the upper portion of such a fraction [3]). A rate is a specific kind of fraction in which the numerator represents the frequency of occurrence of an event during a particular time period and the denominator represents the average population at, or exposed to, risk for occurrence of the event over the time period. Neglect of this "exposed to risk" term is one of the commonest errors made in the everyday use of statistics for calculating rates [2, pp. 238–247].

The essence of the denominator problem is that percentages, proportions, or ratios are misrepresented as rates [1, pp. 289–313; 2, pp. 238–247]. Methodological errors in regards to denominator data usually fall into one of three categories: (1) missing denominators; (2) wrong denominators; and (3) unknown denominators. *Missing denominators* are numbered among the more frequently found flaws in data analysis. When denominators are missing, for whatever reason, numerators may be found serving in their stead, and the proportions and percentages derived from these numerators are mistakenly interpreted as rates [1, pp. 289–313]. Confusion results when the misinformed reader tries to interpret these so-called rates; for example, proportional rates [2, pp. 248–257].

Suppose that physicians compared rates of a new infectious disease (ID) among diabetics and nondiabetics in their clinic and found that diabetics were ill more frequently with this disease. Based on these numerator data, they might wrongly conclude that morbidity rates from this ID were higher among diabetics. However, these data indicate only that diabetics account for the largest *proportion* of the cases. To compare *rates* of disease occurrence, the missing denominator data, representing the exposed population (the total numbers of diabetics and nondiabetics served by the clinic and "at risk" for the ID) must be supplied. If many more diabetics than nondiabetics attended the clinic, when one takes into consideration the magnitudes of the denominators, nondiabetics could have the higher morbidity rate for the ID despite having the lower number of actual cases.

The *wrong denominator* is likely to be used when the population exposed to risk cannot be estimated adequately. Suppose, for example, that a clinic reported rates of disease Z of four per 1000 visits among men but two per 1000 visits among women. Whereas disease Z may account for a greater *percentage*, (or, proportional rate) of visits by men than women, one cannot conclude that the **prevalence** *rate* for disease Z is twice as high in men than in women. These percentages are based on those who presented for evaluation and treatment, and may be specific for this clinic (sometimes called "treated prevalence"). The appropriate denominators for prevalence rates are based on the total numbers of "at risk" men and women in the community served by the clinic, rather than on the number of visits. The two sexes may have similar prevalence rates for disease Z, but because women may visit the clinic for all causes more often than men do, proportionately more men than women present with disease Z.

Another instance of the wrong denominator involves mortality (also, morbidity) rates. The mortality rate for a disease is calculated as the number of deaths due to the specific cause divided by the population at risk for the specific cause of death. Thus, the mortality rate is a measurement of the absolute risk for death from a specific cause (*see* **Vital Statistics, Overview**). If the number of deaths due to a specific cause is divided by the number of deaths due to all causes, the result is a proportion instead of a rate. **Proportional mortality** rates indicate the percentage of deaths due to a given cause in relation to other (or all) causes of death. It has also been argued that cause-specific mortality rates, themselves, are not the most accurate reflection of the **risk** of mortality, and that it would be more meaningful to use the risk of exposure to a specific cause as the denominator term [1, p. 277]. However, population data often are the only denominator data available in sufficient detail to allow rates to be compared.

The *unknown denominator* problem occurs when rates are derived from data that may have been distorted by **selection bias**. In other words, one may be misled by the cases one has at hand. For example, consider 100 depressed patients who initially responded to electroconvulsive therapy (ECT) and later experienced relapse. Prior to relapse and subsequent referral to a consulting psychiatrist, 50 were prescribed maintenance antidepressant therapy after ECT and 50 were not. The consultant may conclude

that maintenance drug treatment does not improve post-ECT one-year remission rates. However, maintenance therapy failures may represent a relatively small fraction of those so treated, compared with those who do not receive post-ECT antidepressant treatment, most of whom will relapse within one year. The fallacy is in calculating relapse rates without knowing the denominator; that is, deriving rates from the number referred for consultation instead of from the population at risk (the number who received each treatment regimen). Those responding to treatment without suffering a relapse will not be in this psychiatrist's series.

References

[1] Colton, T. (1974). *Statistics in Medicine*. Little, Brown & Company, Boston.

[2] Hill, A.B. & Hill, I.D. (1991). *Bradford Hill's Principles of Medical Statistics*, 12th Ed. Edward Arnold, London.

[3] Last, J.M., ed. (1995). *A Dictionary of Epidemiology*, 3rd Ed. Oxford University Press, New York.

HOWARD M. KRAVITZ

Density Estimation

Density estimation is the fitting of a **probability** density function (pdf), $f(x)$, to data. We have the choice of performing a parametric or nonparametric fit. In common usage, the phrase *density estimation* usually refers to the **nonparametric** methodology, which is the focus of this article. We introduce the classic nonparametric estimator, the histogram, and outline its theoretical properties as well as good practice (*see* **Frequency Distribution**). We demonstrate how to improve the histogram, leading to our discussion of popular kernel methods. We conclude with a bivariate example, a way of choosing smoothing parameters, and new directions that promise further improvements (*see* **Nonparametric Regression**).

Why choose nonparametric over parametric density estimation? Parametric density estimation requires both proper specification of the form of the underlying sampling density, $f_{\boldsymbol{\theta}}(x)$, and **estimation** of the parameter vector $\boldsymbol{\theta}$. Parametric modeling entails two risks of bias: in estimation of $\boldsymbol{\theta}$ and incorrect specification of $f_{\boldsymbol{\theta}}$ (*see* **Misspecification**). Nonparametric density estimation provides a consistent **algorithm** for nearly any continuous density and avoids the specification step. Although the cumulative distribution and probability density functions carry the same information, densities are more easily interpreted than distributions, especially in more than one dimension, so our focus on the density is appropriate.

Density estimation is broadly applicable for exploring data relationships, presenting data summaries, and constructing sophisticated nonparametric models of biostatistical data (*see* **Exploratory Data Analysis**). **Graphical** representation of data is a powerful tool for summarization. Three simple exploratory graphical summaries are the box-and-whiskers plot (or boxplot), the stem-and-leaf plot, and the histogram. Consider the cholesterol levels of 320 males with diagnosed coronary artery disease [21]. Figure 1 displays a boxplot of these data. The data appear symmetric with a few outliers. The various percentiles displayed in the boxplot do not hint of any unusual feature such as we see in Figure 2 in the right histogram, which shows mild evidence of bimodality (*see* **Mode**); however, even with 320 observations, the weight of evidence is probably not strong. Observe that the two histograms have the same bin width, but their meshes are shifted. The stem-and-leaf plot (not shown) indicates specific data values but otherwise has no frequency information beyond the histogram.

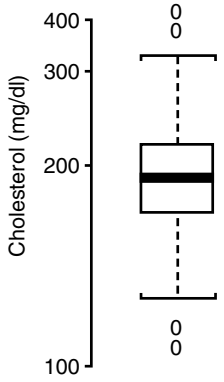

Figure 1 Boxplot of $\log_{10}$ cholesterol data ($n = 320$)

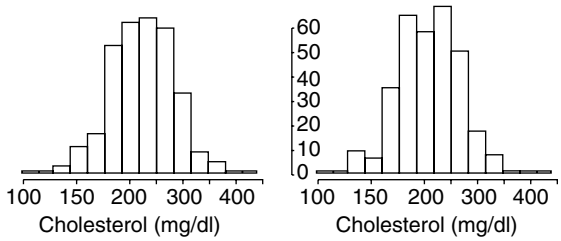

Figure 2 Two histograms of the cholesterol data with the same bin width but shifted meshes. The first appears Gaussian; the second appears bimodal

Histogram

A histogram is the simplest density estimator and is one example of a frequency curve, using tabulation of data in bins. Frequency curves have had an important role to play since their introduction by **John Graunt**, who searched for patterns of death in the Bills of Mortality collected during the London plague of the seventeenth century. Graunt performed a primitive **survival analysis** by grouping age of death in five-year-wide intervals. Here we highlight the theoretical properties of a histogram.

Given a sample $x_1, x_2, \ldots, x_n$ contained in an interval (a, b), the histogram is constructed over a partition $\{t_k\}$ of (a, b) into M intervals, $B_k =$

$[t_{k-1}, t_k)$, such that $a = t_0 < t_1 < \ldots < t_M = b$. Let the bin width of B_k be denoted by $h_k = t_k - t_{k-1}$. Let the bin count be denoted by ν_k, so, that $\sum_{k=1}^{M} \nu_k = n$. Then the histogram estimate of the density, $f(x)$, is given by

$$\hat{f}_{\mathrm{H}}(x) = \frac{\nu_k}{nh_k} = \frac{\nu_k}{n(t_k - t_{k-1})}, \quad x \in B_k,$$

and zero outside $[a, b)$. Observe that $\hat{f}_H$ satisfies both conditions of a density function as $\hat{f}_{\mathrm{H}} \geq 0$ and $\int \hat{f}_{\mathrm{H}} = 1$.

Usually, all the bin widths are chosen to be equal, $h_k = h$, as in the stem-and-leaf plot. While any choice of the bin width will produce an informative diagram, the notion of an optimal bin width has been studied extensively [6, 17]. At each point x, $\hat{f}_H(x)$ is generally biased, so that the **mean square error** (MSE) is an appropriate criterion. A Taylor's series analysis reveals that the sum of the pointwise variance and squared **bias** decomposition is given by

$$\mathrm{MSE}[\hat{f}_{\mathrm{H}}(x)] = \frac{f(x)}{nh} + \frac{1}{12}h^2 f'(x)^2.$$

(Terms omitted are of lower order n^{-1}.) If h is too large, then the histogram has too few bins and is "oversmoothed" – exhibiting low variance but high bias. However, if h is too small, then $\hat{f}_H(x)$ has too many bins and is "undersmoothed" – suffering high variance.

For an entire histogram, the pointwise mean squared error criterion may be integrated to give a global criterion, the integrated mean squared error (IMSE):

$$\begin{aligned}\mathrm{IMSE}(\hat{f}_{\mathrm{H}}) &= \int_{-\infty}^{\infty} \mathrm{MSE}[\hat{f}_H(x)]\,\mathrm{d}x \\ &= \frac{1}{nh} + \frac{1}{12}h^2 R(f'),\end{aligned}$$

where $R(f') = \int_{-\infty}^{\infty} f'(x)^2\,\mathrm{d}x$ is referred to as the "roughness" of the unknown density function. Ordinary calculus reveals the optimal bin width

$$h^*_{\mathrm{H}}(f) = 6^{1/3} R(f')^{1/3} n^{-1/3}.$$

The optimal IMSE decreases to zero at the rate $n^{-2/3}$, far short of the usual parametric rate of n^{-1}. Only one function of the unknown density is relevant to the optimal bin width.

Using a **normal** density, $\phi = N(\mu, \sigma^2)$, as a reference, $R(\phi') = 1/(4\pi^{1/2}\sigma^3)$, so that

$$h^*_{\mathrm{H}}(\phi) = 3.5\sigma n^{-1/3}.$$

In practice, the unknown standard deviation, σ, is replaced by the sample standard deviation or a **robust** estimate. This formula is more general and useful than its motivation might suggest. By considering *all* possible densities in $h^*_{\mathrm{H}}(f)$, it has been found that $h^*_{\mathrm{H}}(\phi)$ is within 7% of a theoretical upper bound [27]. Thus for real data, use of $h^*_{\mathrm{H}}(\phi)$ will almost always result in mild oversmoothing of the data. In no case should a wider bin width be selected. More refined choices will be discussed in the section Choosing Smoothing Parameters below.

Improvements on Histograms

Frequency Polygon

A continuous version of the histogram is the frequency polygon (FP), which is formed by interpolating the midpoints of a histogram. The theoretical properties of the frequency polygon are superior to the histogram. Scott [18] showed that its IMSE decreased at the much faster rate of $n^{-4/5}$, and that $h^*_{\mathrm{FP}}(\phi) = 2.15\sigma\, n^{-1/5}$. (This is within 8% of the oversmoothed upper bound bin width.) The optimal frequency polygon uses wider bins than the optimal histogram. (The optimal histogram requires more and narrower bins to try to approximate the density where the slope is greatest.)

Averaged Shifted Histogram (ASH)

Both the histogram and frequency polygon share a second design parameter that can have a large visual impact, especially for small sample sizes, as demonstrated in Figure 2. This parameter is the bin origin, t_0. For a fixed bin width, there is an unlimited number of possible choices for t_0 in the interval $(a - h, a]$. In many situations, t_0 may be viewed as a **nuisance parameter**. Scott [19] proposed averaging over shifted meshes to eliminate the bin edge effect. To be specific, form a finer (narrower) mesh of width $\delta = h/m$ for some positive integer m, and let B_k and ν_k refer to this new, finer set of bins. Then for $x \in B_k$, there are m different histograms with bin width $h = m\delta$ that cover (include) bin B_k. The bin counts for

these m shifted histograms are $(\nu_{k-m+1} + \cdots + \nu_k)$ to $(\nu_k + \cdots + \nu_{k+m-1})$. A little algebra reveals the mean or averaged shifted histogram (ASH) as

$$\hat{f}_A(x) = \frac{1}{m} \sum_{j=1-m}^{m-1} \frac{(m - |j|)\nu_{k+j}}{nh}$$

$$= \frac{1}{nh} \sum_{j=1-m}^{m-1} \left(1 - \frac{|j|}{m}\right) \nu_{k+j}, \quad x \in B_k. \quad (1)$$

Figure 3 displays an example for $m = 14$ for the cholesterol data as in Figure 2. The estimate is not only visually smoother and more appealing than the histogram, but also shares the improved theoretical properties of the frequency polygon, while, in fact, being about 20% more efficient. The two modes are more clearly uncovered.

Kernel Estimator

As $m \to \infty$, the ASH takes on an equivalent and widely studied form. Since ν_k is either 0 or 1 in the limit (excluding ties), the sum in (1) can be re-expressed as a sum over the n data points:

$$\hat{f}_K(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad (2)$$

with $K(t) = [1 - |t|]_+$, where $[x]_+$ is the positive part of x, or zero. This so-called kernel estimator was extensively studied by Rosenblatt [12] and Parzen [11] although first proposed in a technical report by Fix & Hodges [5]. Similar ideas were more developed in spectral density estimation at that time (*see* **Spectral Analysis**). The parameter h is no longer a bin width, *per se*, and is called a smoothing parameter.

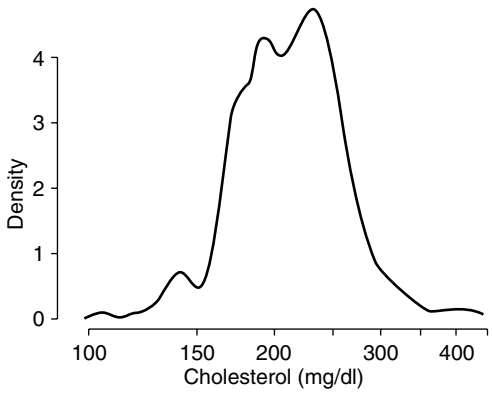

Figure 3 Weighted average of 14 shifted histograms of the cholesterol data. The weights $\{w_m(j)\}$ used were derived from the kernel $K(t) = 315/256(1 - t^2)^4$ according to (3)

The kernel density estimator turns out to be quite general. Any probability density that is square integrable can be selected for the kernel. The usual requirements are that $\int K = 1$ and $\int xK = 0$. Picking a symmetric probability density satisfies both. Even higher order rates of convergence such as $n^{-8/9}$ are possible if $\int x^2 K = 0$, but such kernels must take on negative values.

Apparently, the kernel estimator is a mixture of n densities, each centered on a data point. Computationally, kernel estimation can be quite expensive for large samples. Prebinning the data is an accepted technique for speeding the evaluation. This is equivalent to the ASH with weight function $w_m(j) = 1 - |j|/m$ in (1) replaced by

$$w_m(j) = \frac{mK(j/m)}{\sum_{i=1-m}^{m-1} K(i/m)} \quad (3)$$

for kernels defined on $[-1,1]$. **Fast Fourier transformations** may be used if the kernel is the Gaussian probability density function (pdf) [22].

There are many other techniques based on filtering or orthonormal estimation, but these may be shown to be equivalent to the use of a particular "equivalent" kernel function. Recent interest in the use of **wavelets** as an orthonormal basis illustrates the generality of kernel methodology (*see* **Orthogonality**).

Multivariate Density Estimation

The kernel estimator has a simple extension to two dimensions (and similarly for more dimensions) by using a bivariate probability density, $K(x, y)$, as the kernel:

$$\hat{f}_K(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right), \quad (4)$$

where each coordinate direction has its own smoothing parameter. Some authors [29] advocate the use of the **correlation** coefficient as an additional smoothing

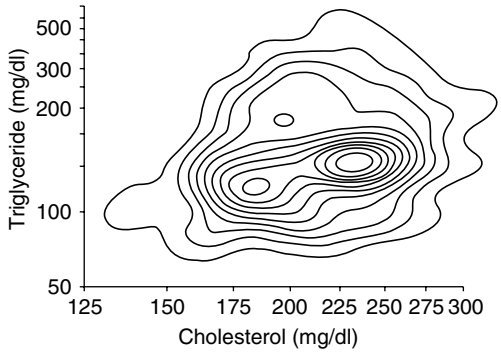

Figure 4 Bivariate averaged shifted histogram of cholesterol and triglyceride blood concentrations for 320 patients as in earlier figures. At least two patient clusters are suggested

parameter. Similarly, the averaged shifted histogram may be defined by averaging over histograms shifted along both the x and y axes. The latter is illustrated in Figure 4 on the coronary artery disease data. (The bivariate kernel was taken as the product of two univariate kernels used in Figure 3.) The bivariate ASH clearly reveals the nonnormality of the data, suggesting an extra mode or two much more strongly.

Examples in three and four dimensions, including visualization, with applications to **clustering, discrimination**, and **regression** are presented by Scott [20].

Choosing Smoothing Parameters

A good deal of research has appeared on improved and automatic algorithms for choosing h in (1), (2), and (4). Some focus on plug-in estimates of quantities such as $R(f')$, while others rely on modification of **maximum likelihood** or **information** measures.

We mention only one method, **least-squares cross-validation**, which if not the most **efficient** procedure, is clearly the most general and widely applicable. The IMSE criterion is the average of the integrated squared error (ISE) between $\hat{f}_h$ and f:

$$\begin{aligned}\mathrm{ISE}(h) &= \int [\hat{f}_h(x) - f(x)]^2 \, \mathrm{d}x \\ &= \int \hat{f}_h(x)^2 - 2\int \hat{f}_h(x) f(x) \, \mathrm{d}x \\ &\quad + \int f(x)^2 \, \mathrm{d}x \\ &= R(\hat{f}_h) - 2E\hat{f}_h(X) + R(f).\end{aligned}$$

Rudemo [13] and Bowman [1] observed that the third term, $R(f)$, is constant for all choices of h and can be ignored. The first integral is directly computable as h varies. Finally, the second term has an **unbiased** estimator:

$$-\frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(x_i),$$

where $\hat{f}_{h,-i}(x_i)$ is the density estimate based on the $n-1$ points with x_i omitted. For the histogram, the least-squares cross-validation functional is

$$\mathrm{CV}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2 h}\sum_k \nu_k^2.$$

The functional can be immediately extended to pick the bin origin, t_0, as well. The minimizer of $\mathrm{CV}(h)$ may be found by grid search or numerical methods (*see* **Optimization and Nonlinear Equations**). Several CV formulae for kernel estimates, including the multivariate case, are given in Sain et al. [14].

New Directions and Resources

Locally adaptive density estimation will play an increasingly important role now that that technology has begun to appear. Promising approaches include plug-in [16], log **spline** [9], wavelet [8], local likelihood [8, 10], and local CV bandwidths [15]. The reader is cautioned about overfitting difficulties, as many degrees of freedom are consumed during the estimation of the adaptive smoothing parameters (either explicitly or implicitly). A more conservative approach is to follow some of the simple transformation ideas of Wand et al. [30].

Historical information and greater detail are available in the following selection of monographs: Tapia & Thompson [25], Silverman [23], Devroye [2], Härdle [7], Scott [20], Tarter & Lock [26], Wand & Jones [29], Fan & Gijbels [4], and Simonoff [24]. These references also discuss the wide array of closely related nonparametric techniques including regression, spectral densities, **time series**, cluster analysis, discrimination, and survival and **hazard** estimation. **Software** is available in packages such as SAS, **S-PLUS**, and Systat, for example, or at **internet** sites such as statlib and netlib.

References

[1] Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* **71**, 353–360.

[2] Devroye, L. (1987). *A Course in Density Estimation*. Birkhäuser, Boston.

[3] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1996). Density estimation by wavelet thresholding, *Annals of Statistics* **24**, 508–539.

[4] Fan, J. & Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall, New York.

[5] Fix, E. & Hodges, J.L. (1951). Nonparametric discrimination: consistency properties, Reprinted by Silverman, B.W. & Jones, M.C. (1989). *International Statistical Review* **57**, 233–247.

[6] Freedman, D. & Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**, 453–476.

[7] Härdle, W. (1991). *Smoothing Techniques with Implementation in S*. Springer-Verlag, New York.

[8] Hjort, N.L. & Jones, M.C. (1996). Locally parametric nonparametric density estimation, *Annals of Statistics* **24**, 1619–1647.

[9] Kooperberg, C. & Stone, C.J. (1991). A study of logspline density estimation, *Computational Statistics and Data Analysis* **12**, 327–347.

[10] Loader, C.R. (1996). Local likelihood density estimation, *Annals of Statistics* **24**, 1602–1618.

[11] Parzen, E. (1962). On estimation of probability density function and mode, *Annals of Mathematical Statistics* **33**, 1065–1076.

[12] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* **27**, 832–837.

[13] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics* **9**, 65–78.

[14] Sain, S.R., Baggerly, K.A. & Scott, D.W. (1994). Cross-validation of multivariate densities, *Journal of the American Statistical Association* **89**, 807–817.

[15] Sain, S.R. & Scott, D.W. (1996). On locally adaptive density estimation, *Journal of the American Statistical Association* **91**, 1525–1534.

[16] Schucany, W.R. (1989). Locally optimal window widths for kernel density estimation with large samples, *Statistics and Probability Letters* **7**, 401–405.

[17] Scott, D.W. (1979). On optimal and data-based histograms, *Biometrika* **66**, 605–610.

[18] Scott, D.W. (1985). Frequency polygons: theory and application, *Journal of the American Statistical Association* **80**, 348–354.

[19] Scott, D.W. (1985). Averaged shifted histograms: effective nonparametric density estimators in several dimensions, *Annals of Statistics* **13**, 1024–1040.

[20] Scott, D.W. (1992). *Multivariate Density Estimation*. Wiley, New York.

[21] Scott, D.W., Gotto, A.M., Cole, J.S. & Gorry, G.A. (1978). Plasma lipids as collateral risk factors in coronary artery disease: a study of 371 males with chest pain, *Journal of Chronic Diseases* **31**, 337–345.

[22] Silverman, B.W. (1982). Algorithm AS176. Kernel density estimation using the fast Fourier transform, *Applied Statistics* **31**, 93–99.

[23] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[24] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.

[25] Tapia, R.A. & Thompson, J.R. (1978). *Nonparametric Probability Density Estimation*. Hopkins Press, Baltimore.

[26] Tarter, M.E. & Lock, M.D. (1993). *Model-Free Curve Estimation*. Chapman & Hall, New York.

[27] Terrell, G.R. & Scott, D.W. (1985). Oversmoothed nonparametric density estimates, *Journal of the American Statistical Association* **80**, 209–214.

[28] Wand, M.P. & Jones, M.C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation, *Journal of the American Statistical Association* **88**, 520–528.

[29] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

[30] Wand, M.P., Marron, J.S. & Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association* **86**, 343–361.

D.W. SCOTT

Density Sampling

Density sampling is a method of sampling **controls** in a **case–control study**. Controls are sampled from the population at risk at the times of incidence of each case or, as is more common in practice, over the period of accrual of the cases. Time-matched analysis of such case–control data yields **unbiased** estimates of the **relative hazard** (or **incidence density ratio**), even when the disease is common [1–4]. An advantage of density sampling is that it can reduce **bias** from secular changes in the **prevalence** of exposure during the course of the study [1].

References

[1] Greenland, S. & Thomas, D.C. (1982). On the need for the rare disease assumption in case–control studies, *American Journal of Epidemiology* **116**, 547–553.

[2] Miettinen, O.S. (1976). Estimability and estimation in case-reference studies, *American Journal of Epidemiology* **103**, 226–235.

[3] Prentice, R.L. & Breslow, N.E. (1978). Retrospective studies and failure time models, *Biometrika* **65**, 153–158.

[4] Sheehe, P.R. (1962). Dynamic risk analysis in retrospective matched pair studies of disease, *Biometrics* **18**, 323–341.

MITCHELL H. GAIL

Dermatology

Dermatology is the branch of medicine that is concerned with the physiology and pathology of the skin. It is estimated that one in three Americans has a skin condition serious enough to warrant a visit to a physician and that 10% of visits to physicians are at least in part for skin complaints [12]. The most common diseases seen by dermatologists include acne vulgaris, veruca vulgaris (warts), psoriasis, nonmelanoma skin cancer (basal and squamous cell skin cancer), dermatitis, dermatophytosis (fungal infections of the skin), and actinic keratoses [9]. Thus, commonly encountered skin diseases arise from a wide variety of etiologies. For example, warts and dermatophytosis are infections of the skin caused by human papilloma virus and fungi, respectively; nonmelanoma skin cancers are due to a combination of factors including genetic predisposition and exposure to ultraviolet light; and dermatitis is commonly caused by delayed type hypersensitivity (allergy) to a wide variety of chemicals.

Several statistical techniques and methodologies are heavily utilized in the study of skin diseases. Paramount among them are use of **case–control studies** and **multivariate analysis** to identify **risk** and **prognostic factors** in melanoma, development of tools to assess skin disease activity and response to treatment, and the use of **linkage analysis** to identify genetic defects responsible for hereditary skin diseases.

There were approximately 38 300 new cases of malignant melanoma and 7300 deaths attributed to it in the US in 1996. The **incidence** and mortality rates of melanoma have risen steadily since 1930. Finding ways to identify patients at highest risk and to identify alterable risk factors are important since early diagnosis and adequate surgical resection remain the best ways to treat melanoma. However, the development of melanoma is a complex trait that involves the interaction of genetic and environmental factors (*see* **Gene-environment Interaction**). Therefore, carefully conducted case–control studies have been essential to identify important risk factors for the development of melanoma. **Odds ratios** are highest for a family or personal history of atypical moles and melanoma, a family or personal history of melanoma, presence of atypical melanocytic nevi or many melanocytic nevi, a history of multiple sun burns, and intermittent, intense sun exposure [15, 16, 20]. Families that have a genetic susceptibility for developing melanoma have been identified [10, 14, 17]. These families have members who have multiple atypical nevi and have a lifetime risk of developing melanoma that approaches 100%. Many family members develop multiple melanomas.

Many methods have been advocated to predict the prognosis (*see* **Predictive Modeling of Prognosis**) of patients with melanoma. The tumor type, level of invasion, location, clinical features, thickness, cytologic characteristics and lymph node involvement have all been used and advocated as prognostic indicators. No single factor will predict accurately the outcome for all patients. Multivariate analyses of several large cohorts of melanoma patients were, therefore, essential in identifying the best indicators of prognosis in patients with melanoma. They have identified tumor thickness (from the top of the granular layer to the bottom of the tumor) as the best indicator of prognosis in patients with stage I melanoma (localized disease with no clinical or histopathologic lymph node involvement). The pivotal studies of Breslow and others have been confirmed by several groups and have established that the five-year survival of patients with tumor thickness of less than 0.75 mm, 0.76 mm–1.5 mm, 1.51 mm–3.99 mm and greater than 4 mm were 100%, 90%–94%, 76%–83%, and 40%, respectively [2, 4–6, 18]. Whereas the utility of tumor thickness as a prognostic indicator has its detractors [11] it remains the best and most widely used indicator.

Two principal methods are used to assess skin disease activity and to determine patient outcomes in dermatologic clinical trials [1, 3] (*see* **Clinical Trials, Overview**). The first involves examining patients before, during, and at the conclusion of treatment, and reporting how the patients appear at the various time points. The second involves determining the degree of improvement during treatment or during the observation period. An example of the first method was developed to assess the response of psoriasis to novel treatments. The psoriasis area and severity index (PASI) assigns numerical values to the amount of erythema, scaling, and degree of infiltration and multiplies them by the area of the body surface involved to formulate an "index" of the patient's condition [8]. The PASI ranges from 0 to 72. The major problem with indices is that they confound area of involvement with severity of disease. For

example, a patient with thick plaque-type psoriasis of the knees, elbows and scalp may have the same index as a patient with diffuse but minimal psoriasis of the trunk and arms [1, 3]. The second problem with indices is that they lend an undeserved air of precision to the analysis and presentation of data [1, 3]. For example, Tiling-Grosse & Rees [19] demonstrated that physicians and medical students were poor at estimating the area of skin disease and, therefore, some of the components that make up indices may be inaccurate. Finally, calculating the means, differences in means, and percentages of change in indices in response to treatment may not convey an accurate clinical picture of the changes that have occurred. The major limitations of the use of the second method (determining the degree of improvement) are that the categories of improvement are often not well defined and that the categories are often assumed to be, but are, in fact, not, additive [1, 3]. That is, 60%–80% improvement is often assumed to be twice as good as 20%–40%, although no such numerical relationship exists between these subjectively defined categories. To overcome some of the limitations of currently utilized methods, some groups are beginning to use **quality of life** assessment tools to assess skin disease activity [7].

As in other fields of medicine, clinical trials are the best sources for determining the best available treatments in dermatology. However, published clinical trials should be approached with a sense of skepticism. In many reviews of clinical trials, they often are found to be poorly performed or reported. Specific methods must be employed in the conduct of clinical trials to lead to valid conclusions. There are many practical and logistical difficulties inherent in conducting therapeutic trials in dermatology, including limited numbers of patients, difficulties in measuring the outcome of treatment (*see* **Outcome Measures in Clinical Trials**), and difficulties in **blinding**. Adherence to all recommended methods is not always possible. However, investigators must provide readers with adequate information about the methods employed in published clinical trials. Several features strengthen clinical trials and help validate their conclusions. These features include proper selection and allocation of patients, inclusion of an appropriate control group, **randomization**, prior selection of clinically and biologically important outcome variables, blinding of assessment when possible, consideration of patient **compliance** and dropout, and proper presentation and statistical analysis of results [3].

Linkage analyses (used in conjunction with positional cloning) have aided in the identification of the genetic causes of many skin diseases. Linkage analysis consists of finding a model M1 (of inheritance of a **gene** of interest and a phenotype of interest) that is much more likely to have produced the observed data than a null hypothesis M0 (in which the inheritance of an unrelated gene has no linkage to the phenotype) [13]. The evidence for the hypothesis is measured by the lod score [$\log_{10}$ (**likelihood ratio** of the hypothesis vs. the **null hypothesis**; where the likelihood ratio = Pr(data|M1/ Pr(data|M0)))]. Lod scores >3 are conventionally considered significant. Analyses are facilitated with the use of available computer programs (LIPED and LINKAGE) [13] (*see* **Software for Genetic Epidemiology**). Linkage analyses have aided in the identification of the genetic deficits that are responsible for the basal cell nevus syndrome, several forms of epidermolysis bullosa, several of the icthyoses, some cases of Waardenberg's syndrome, dyskeratosis congenita, and pachyonychia congenita. It has also identified possible genetic loci for psoriasis.

References

[1] Allen, A.M. (1980). Clinical trials in dermatology, Part 3: Measuring responses to treatment, *International Journal of Dermatology* **19**, 1–6.

[2] Balch, C.M., Murad, T.M., Soong, S.J., Ingalls, A.L., Halpern, N.B. & Maddox, W.A. (1978). A multifactorial analysis of melanoma: prognostic histopathological features comparing Clark's and Breslow's staging methods, *Annals of Surgery* **188**, 732–742.

[3] Bigby, M., & Gadenne, A.S. (1996). Understanding and evaluating clinical trials (review), *Journal of the American Academy of Dermatology* **34**, 555–590; quiz, 591–593.

[4] Breslow, A. (1970). Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma, *Annals of Surgery* **172**, 902–908.

[5] Day, C.L., Jr, Lew, R.A., Mihm, M.C., Jr, Harris, M.N., Kopf, A.W., Sober, A.J. & Fitzpatrick, T.B. (1981). The natural break points for primary-tumor thickness in clinical Stage I melanoma (letter), *New England Journal of Medicine* **305**, 1155.

[6] Eldh, J., Boeryd, B. & Peterson, L.E. (1978). Prognostic factors in cutaneous malignant melanoma in stage I. A clinical, morphological and multivariate analysis, *Scandinavian Journal of Plastic and Reconstructive Surgery* **12**, 243–255.

[7] Finlay, A.Y. & Coles, E.C. (1995). The effects of severe psoriasis on the quality of life of 369 patients, *British Journal of Dermatology* **132**, 236–244.

[8] Fredriksson, T. & Pettersson, U. (1978). Severe psoriasis – oral therapy with a new retinoid, *Dermatologica* **157**, 238–244.

[9] Friedman, G.D. (1994). Opportunities for research in dermatologic epidemiology, *Journal of Investigative Dermatology* **102**, 57S–58S.

[10] Goldstein, A.M. & Tucker, M.A. (1995). Genetic epidemiology of familial melanoma (review), *Dermatologic Clinics* **13**, 605–612.

[11] Green, M.S. & Ackerman, A.B. (1993). Thickness is not an accurate gauge of prognosis of primary cutaneous melanoma (review), *American Journal of Dermatopathology* **15**, 461–473.

[12] Johnson, M.-L. & Roberts, J. (1978). Skin conditions and related need for medical care among persons 1 to 74 years, United States, 1971–1974, in *Vital and Health Statistics, Series 1, No. 212. DHEW Publication (PHS) 79-1660*. US Government Printing Office, Washington.

[13] Lander, E.S. & Schork, N.J. (1994). Genetic dissection of complex traits, *Science* **265**, 2037–2048.

[14] Lucchina, L.C., Barnhill, R.L., Duke, D.M. & Sober, A.J. (1995). Familial cutaneous melanoma (review), *Melanoma Research* **5**, 413–418.

[15] Rigel, D.S. (1995). Identification of those at highest risk for development of malignant melanoma (review), *Advances in Dermatology* **10**, 151–70, discussion, 171.

[16] Slade, J., Salopek, T.G., Marghoob, A.A., Kopf, A.W. & Rigel, D.S. (1995). Risk of developing cutaneous malignant melanoma in atypical-mole syndrome: New York University experience and literature review (review), *Recent Results in Cancer Research* **139**, 87–104.

[17] Slade, J., Marghoob, A.A., Salopek, T.G., Rigel, D.S., Kopf, A.W. & Bart, R.S. (1995). Atypical mole syndrome: risk factor for cutaneous malignant melanoma and implications for management (review), *Journal of the American Academy of Dermatology* **32**, 479–494.

[18] Sober, A.J., Day, C.L., Jr, Fitzpatrick, T.B., Lew, R.A., Kopf, A.W. & Mihm, M.C., Jr (1983). Factors associated with death from melanoma from 2 to 5 years following diagnosis in clinical stage I patients, *Journal of Investigative Dermatology* **80**, 53S–55S.

[19] Tiling-Grosse, S. & Rees, J. (1992). Assessment of area of involvement in skin disease: a study using schematic figure outlines, *British Journal of Dermatology* **128**, 69–74.

[20] Whiteman, D. & Green, A. (1994). Melanoma and sunburn (review), *Cancer Causes and Control* **5**, 564–572.

MICHAEL BIGBY

Descriptive Epidemiology

Descriptive epidemiology is the study of the incidence and **prevalence** of diseases and associated mortality in populations. Unlike **analytic epidemiology**, descriptive epidemiologic studies usually do not rely on individual-level data, for example on exposures, disease outcome, and potential **confounders**. Instead, descriptive studies estimate the risk of disease in various groups defined by age, gender, and ethnicity (*see* **Ethnic Groups**), evaluate time trends in disease rates, identify geographic localization of populations with high rates of disease (*see* **Geographic Patterns of Disease**), and attempt to **correlate** disease rates in populations with features of the population, such as the average level of exposure to a potential carcinogen. Descriptive studies are used to determine the effectiveness of programs to control disease (*see* **Program Evaluation**), and descriptive studies are used to generate etiologic hypotheses that are then tested in analytic studies (*see* **Age–Period–Cohort Analysis**; **Correlational Study**; **Ecologic Study**).

MITCHELL H. GAIL

Design Effects

For an **estimator**, $\hat{x}$, under a particular sampling design, D, the *design effect* (denoted DEFF) is defined as the ratio of the **variance** of the estimator under the particular sampling design to its variance at equivalent sample size, n, under **simple random sampling** without replacement (*see* **Sampling With and Without Replacement**) [2–4]. That is:

$$\mathrm{DEFF}_{\mathrm{D}}(\hat{x}) = \frac{\mathrm{var}_{\mathrm{D}}(\hat{x})}{\mathrm{var}_{\mathrm{SRS}}(\hat{x})}, \tag{1}$$

where $\mathrm{DEFF}_{\mathrm{D}}(\hat{x})$ is the design effect for an estimate, $\hat{x}$, under sample design, D; $\mathrm{var}_{\mathrm{D}}(\hat{x})$ is the variance of $\hat{x}$ under the sample design, D; and $\mathrm{var}_{\mathrm{SRS}}(\hat{x})$ is the variance of $\hat{x}$ under simple random sampling without replacement.

The comparable ratios of **standard errors**, generally denoted DEFT, is referred to as the *design factor*.

A design effect greater than unity indicates that the particular sample design would yield an estimate having higher variance than what would be obtained from a sample of the same number, n, of units under simple random sampling. Conversely, a design effect less than unity would imply that the particular sampling design would result in the estimate having a lower variance than what would be obtained under simple random sampling. Design features such as **stratification** generally result in design effects less than unity, whereas **cluster sampling** in its many manifestations generally produces design effects greater than unity. Again, the above comparisons are made at equivalent n, whereas it may be more appropriate to make comparisons at equivalent cost (*see* **Multistage Sampling**).

The variance of a sample **mean** under simple random sampling is equal to $(S_x^2/n) \times (1 - f)$, where $S_x^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2/(N-1)$, N is the population size, $\overline{X}$ is the population mean, and f is the sampling fraction, n/N. Thus the design effect for a sample mean under a particular design, D, is given by

$$\mathrm{DEFF}_{\mathrm{D}}(\overline{x}) = \frac{\mathrm{var}_{\mathrm{D}}(\overline{x})}{(S_x^2/n) \times (1 - f)}. \tag{2}$$

From (2), we see that $\mathrm{var}_{\mathrm{D}}(\overline{x})$ can be put in the following form:

$$\mathrm{var}_{\mathrm{D}}(\overline{x}) = \frac{S_x^2(1 - f)}{n^*}, \tag{3}$$

where $n^* = n/\mathrm{DEFF}_{\mathrm{D}}(\overline{x})$ is known as the *effective sample size*. Since the numerator of (3) is that of the numerator for the variance of the estimator under simple random sampling, the effective sample size for the design tells us that a sample of n units under the particular design is equivalent to a sample of n^* units under simple random sampling. For example, if the design effect for a particular multistage sampling design is 2.8, then a sample of 100 units under this design is equivalent to a sample of 40 units under simple random sampling.

Design effects have been estimated numerically for many important surveys and for many variables (see, for example, Groves & Kahn [1]). These are very useful in the planning of sample surveys, particularly in determination of required sample sizes (*see* **Sample Size Determination**).

References

[1] Groves, R.M. & Kahn, R.L. (1979). *Surveys by Telephone. A National Comparison with Personal Interviews*. Academic Press, New York.

[2] Jolliffe, F.R. (1986). *Sample Design and Analysis*. Ellis Horwood, Chichester.

[3] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[4] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

PAUL S. LEVY

Detection Bias

Detection (also called diagnostic or unmasking) bias results from closer follow-up or more intense scrutiny of one comparison group than another. In a **case–control study**, the detection of a higher proportion of subclinical outcomes among the exposed leads to an overrepresentation of exposed cases relative to exposed controls in the study population. In a **cohort study**, as subjects are followed over time for the occurrence of a disease, subjects who develop unrecognized subclinical disease would be misclassified as nondiseased. If exposed subjects are under greater scrutiny than the unexposed, then they may be less likely to have such undiagnosed subclinical disease. This implies that detection bias can lead to **selection bias** in a case–control study, and can also be a source of differential misclassification in a follow-up study (*see* **Misclassification Error**).

HOLLY A. HILL & DAVID G. KLEINBAUM

Diagnosis Related Groups (DRGs): Measuring Hospital Case Mix

Introduction

The Diagnosis Related Group (DRG) system constitutes an approach to measuring hospital **case mix**, which may be understood as a "system for separating hospitalized patients into unique groups based on their diagnoses and procedures" [5]. Case-mix measurement has been identified by Hornbrook [9] as one of the three fundamental dimensions of hospital output, the other two being volume and quality. While volume is straightforward and refers to the total number of patients treated by the hospital, the definition of case mix and quality are more complex. Hornbrook defines quality as "the hospital's contribution to the successful outcome or resolution of patients' illnesses or health problems" [9, p. 295] and case mix as "the proportion of cases of each disease and health problem treated in the hospital" [9, p. 296].

While recognizing the importance of all dimensions of hospital output, in particular, the issue of **quality of care**, this paper will focus on one particular dimension, that is, the specification of hospital case mix and, in particular, the DRG case-mix classification system. The DRG system came to international prominence in 1983 when this was the chosen approach for case-mix adjustment within the Prospective Payment System introduced within the **Medicare** program by the US government. While DRGs have been shown to be amenable to a wide range of applications, the strength of the system in providing an accessible framework for the determination of resource requirements for different patient types has meant that it has been increasingly used within the United States and overseas by payment bodies interested in applying a case-mix adjustment within their reimbursement systems.

DRGs: Development and Construction of an Operational Case-mix Measure

If classes of patients that share common clinical attributes can be differentiated according to the "bundle" of services received as part of the therapeutic process, this framework constitutes the basis for a case-mix classification scheme that "provides a means for examining the products of the hospital, since patients within each class are expected to receive a similar product" [3]. The hospital product can therefore be defined by the development and application of a case-mix classification system consisting of discrete classes of patients exhibiting common clinical attributes and similar output utilization patterns.

The complexity of both illness and the therapeutic process means that, in turn, the development of a system for classifying case mix is a complicated undertaking. The development of the DRG patient classification system by the Health Systems Management Group at the Yale School of Organization and Management in the late 1960s was originally motivated by the need to develop operational techniques for utilization review. The importance of developing an explicit link between the clinical characteristics of patients and their use of hospital resources was recognized as an essential prerequisite to the evaluation of the appropriateness of service utilization within the hospital setting [4].

In developing a classification system for the definition of case types within the acute hospital setting, the following attributes were specified for the system by the Health Systems Management Group [3]:

1. The system must be interpretable medically, with subclasses of patients from homogeneous diagnostic categories;
2. Individual patient classes should be defined on variables commonly found on hospital abstract systems and relevant to output utilization;
3. The number of classes in the system must be manageable, mutually exclusive, and exhaustive;
4. The classes should be constituted by patients with similar expected measures of output utilization;
5. Class definitions should be comparable across different coding schemes.

Variable Specification and Measurement

The independent variables (*see* **Explanatory Variables**) used for the purpose of specifying a system to achieve these objectives were selected to be descriptive of the patient, of the patient's disease condition, and of the treatment process. In addition, it was considered essential that information relating to the

selected variables should be easily available on discharge abstract summaries if the resultant system was to be available for general application. The initial stages of the analyses identified a number of variables which, in descriptive studies of hospital activity, had been found to be associated with variations in length of stay and other resource use measures [3]. Ultimately, a set of independent variables were identified as representing the essential demographic and clinical attributes of inpatients. These variables include the following: primary diagnosis, secondary diagnoses, surgical procedures performed, age, sex, and discharge status.

For the specification of an accurate and acceptable measure of hospital case mix, a measure of hospital output had to be incorporated within the development process. To place the choice of output measure for the purpose of case-mix measurement in context, it may be useful at this point to consider the hierarchy of hospital output classification schemes constructed by Hornbrook [8]. This hierarchy follows the sequence of the medical care process and begins with iso-symptom groups, progressing through to iso-disease groups and iso-illness groups. When iso-illness groups are collapsed into classes that are homogeneous in terms of the level of resources used in treatment, iso-resource groups are produced. The DRG system fits into this category, as homogeneity with respect to clinical attributes is an essential prerequisite for class determination, with the additional expectation that resource use at the group level will also be relatively homogeneous.

For the development of the iso-resource groups, or DRGs, limitations on data availability meant that the options available for choosing an appropriate dependent variable (*see* **Response Variable**) were restricted. While costs may be a most desirable measure of output, accurate and comprehensive data on costs for a representative sample of hospitals are very difficult to obtain. Even where cost data are available, it can be very difficult to interpret because of variations in the method of collection and estimation. These data problems led to the Yale researchers choosing length of stay (LOS) as the measure of output to be used as the dependent variable [3]. Length of stay, as a measure of output, has the advantage of being standardized, reliable, and routinely available on discharge abstract summaries. In addition, length of stay and ancillary service use have been found to be significantly interrelated for a number of common medical and surgical conditions [6, 7].

The DRG Assignment Process

In developing a classification system with the required attributes, three key inputs were required: physician review, efficient information systems, and statistical **algorithms**. The objective of ensuring that the patient groups formed by the classification process were medically meaningful was the responsibility of panels of physicians established for this purpose. The basic framework for DRG assignment may be summarized as follows:

Step 1: Hospital discharges are partitioned into mutually exclusive and exhaustive primary diagnostic groupings called *Major Diagnostic Categories* (MDCs). The MDCs were specified under the following conditions [3]:

1. Major Diagnostic Categories must be consistent with regard to the anatomic, physiopathologic classification, or in the manner in which they are clinically managed;
2. Major Diagnostic Categories must have sufficient numbers of patients; and
3. Major Diagnostic Categories must cover all codes without overlap.

There are currently 25 MDCs [11]. This classification is primarily based on the organ system or the specialty that would usually provide patient care. There are a number of exceptions including MDC 12 (Diseases and Disorders of the Male Reproductive System) and MDC 13 (Diseases and Disorders of the Female Reproductive System), where urogenital conditions are split on the basis of the sex of the patient.

Step 2: Where relevant, discharges within the MDC are subdivided according to whether or not a surgical procedure was performed. For specific MDCs, there are some exceptions to this initial major procedure split, for example, MDC14 (pregnancy, child birth, and the puerperium) where the initial split is “delivery during this admission?”.

Step 3: Coming into this level, there are two groups within most MDCs – the medical group and the surgical group. During this stage, the medical patients

are further subdivided into categories based on their principal diagnosis. Surgical patients are categorized according to the procedures performed. The procedures, in turn, are ranked in terms of resource intensity. Surgical patients are categorized into subgroups on the basis of the most resource intensive procedure received, which is related to the primary diagnosis.

Step 4: The final stage in the classification involves the derivation of additional diagnostic or surgical subgroups based on age, specific secondary diagnoses, **comorbidities** or complications (CC), nonoperating room procedures, and discharge status where these variables have been found to be significant in clinical terms and have a significant effect on length of stay.

While this process of assignment represents the generic framework underlying the original development of the DRG system, this system has evolved into a number of manifestations in response to increasing demands on patient classification systems to address a wide range of requirements including (see [1])

- the comparison of hospitals across a range of resource and outcome measures,
- evaluation of differences in inpatient mortality rates,
- the implementation and support of critical pathways,
- facilitation of continuous quality improvement projects,
- support of internal management and planning systems,
- management of capitated payment systems, and so on.

Up to the development of the eighth version of the DRG system (in 1990), the principal diagnosis was the initial variable in DRG assignment. For the eighth and subsequent versions, the procedure performed, if any, is the initial step in DRG assignment. In particular, where liver, bone marrow, and lung transplants, together with tracheostomies are performed, the patients are assigned to the relevant DRGs independent of the MDC of the principal diagnosis (3M, 1998).

Experience with a wide range of DRG applications has now developed into an extensive literature and reflects the variety of such applications in different settings and different countries. One of the early examples of experimentation with DRGs on a European database involved analyses for selected high-volume pathologies on a multinational database including over 3.3 million cases from 12 countries [14]. Differences in length of stay both within and between countries were assessed for three alternative measures of hospital case mix applied to over 119 400 discharges for three surgical DRGs (DRG 39 lens procedures, DRG 198 cholecystectomy without common duct exploration (CDE), without complication/comorbidity (CC), and DRG 119 vein ligation and stripping) and two medical DRGs (DRG 294 diabetes age >35 years and DRG 122 circulatory disorders with acute myocardial infarction (AMI), without cardiovascular complications, discharged alive). The results showed that irrespective of the case-mix measure applied, substantial unexplained variation in hospital length of stay persisted leading to the conclusion that in addition to standardizing for case mix, future research on this issue should also focus on additional potentially influential factors, including health system characteristics, medical practice variation, and patient behavior and expectations.

DRG Grouper Development

Developments in the DRG system are a reflection of the evolution in potential applications as well as developments in expertise, information technology, and data systems. Six main categories of DRG system, as developed in the United States, may be identified [1]

- Medicare DRGs
- Refined DRGs (RDRGs)
- All Patient DRGs (AP-DRGs)
- Severity DRGs (SDRGs)
- All Patient Refined DRGs (APR-DRGs)
- International Refined DRGs (IR-DRGs).

The Medicare DRGs are the closest to the system originally developed in the 1980s and continue to constitute one of the most widely used groupers currently. Following the implementation of the Medicare prospective payment system in 1983, responsibility for the maintenance and modification of this system became the responsibility of the Health Care Financing Administration (HCFA). (HCFA is now known as the Centers for Medicare and Medicaid Services.)

There are currently over 500 groups within the Medicare DRG system that is updated annually.

Under a grant from HCFA to Yale University, the Refined Diagnosis Related Groups (RDRGs) were developed in the mid-1980s incorporating a revised specification of complications and comorbidities. Essentially, this system involved the categorization of the secondary diagnosis groups as moderate, major, or catastrophic [1]. All age and CC splits were eliminated and replaced by four subgroups for surgical patients (non-CC, moderate CC, major CC, and catastrophic CC) and three subgroups for medical patients (non-CC, moderate or major CC and catastrophic CC). The number of groups in the RDRGs approximates 1170 and as there is no single source for this system, versions produced by different vendors may differ.

During the late 1980s, a version of the HCFA DRGs was amended under contract with the National Association of Children's Hospitals and Related Institutions (NACHRI). A New York (NY) grouper was also developed around this time in response to legislative changes in the state of New York. These groupers form the basis for what have become known as the All Patient (AP) DRGs.

Given the origins of this system, it is not surprising that the AP-DRGs are more differentiated to reflect factors like birth weight that are considered significant in a neonatal and pediatric population. In addition, some of the advancements emerging from the development of the RDRGs, specifically, the designation of major CC splits, have also been incorporated into subsequent revisions to the AP-DRGs, which now number over 650 groups.

In the mid-1990s, a severity refined (SR) DRG system was developed following a reevaluation of the complications and comorbidities within the HCFA DRGs. When finalized, this system incorporated over 650 groups, though while published by HCFA in 1994, an implementation date was not established and the system has not subsequently been updated [1]. The All Patient Refined Diagnosis Related Groups (APR-DRGs) took as the starting point the AP-DRGs and focused on development to take account of severity of illness or risk of mortality. Within this system, a discharge is assigned three distinct descriptors: the base APR-DRGs, the severity-of-illness subgroup and the risk-of-mortality subgroup. Averill [1] notes with regard to APR-DRG assignment that "The most important component of determining the final patient subgroup is the recognition of the impact of interactions among secondary diagnoses" (p. 399). Each base APR-DRG may have four severity subgroups representing minor, moderate, major, or extreme severity of illness or risk of mortality amounting to a total of around 1530 APR-DRGs.

The system known as the International Refined Diagnosis Related Groups (IR-DRGs) had a somewhat different starting point to other such systems in that it was originally designed for use in the international health setting and, specifically, to be compatible with differences in coding schemes used in different health systems. The development of the IR-DRGs took as the starting point the AP-DRG concept, applied refinements to the base DRGs, and undertook a consolidation process for the complication and comorbidity splits (CC), the major complication and comorbidity splits (MCC), the age splits and the DRGs founded on the complicated Principal Diagnosis [12]. Unlike previous DRG systems in this series, which were developed from US databases, data from a number of European countries were used for the development of the IR-DRGs, which currently incorporate over 900 groups.

Conclusion

In addition to being applied extensively throughout North America, the case-mix systems described here are now also widely used internationally, particularly in Europe and Australia [13]. Many European countries employ some form of case-mix adjustment for hospital financing and/or management purposes, while local applications may include support for clinical budgeting, waiting-list reduction, increased productivity, and so on [10]. In Australia, states like Victoria and South Australia use case mix for hospital payment purposes in addition to supporting a range of management functions.

Outside of the United States, Australia is probably the country where the most extensive DRG-type development programme has been underway since the early 1990s. Following the initial release of the Australian National Diagnosis Related Groups (AN-DRGs) in 1993, this system was updated annually until 1996. The Australian Refined Diagnosis Related Groups (AR-DRGs) were then developed and a biennial update schedule put in place for subsequent revisions [2]. Since the mid-1990s, the total number of groups has remained stable at around 660 and there are currently 23 MDCs within the

AR-DRG system. While retaining the "DRG" label, it should be noted that the Australian system is now distinctly different from the US-developed systems described above. Specifically, the AR-DRGs are based on an Australian morbidity coding system (International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Australian Modification (ICD-10-AM)) and groups discharges based on data items including diagnoses (up to 30 per record), procedures (up to 30 per record), sex, age, mode of separation, length of stay, leave days, admission weight, mental health status, and same-day status [2]. As the Australian DRG system is copyrighted to the Commonwealth of Australia (Department of Health and Ageing), detailed information on the specifics of the system and updates applied can be found on the government's case mix internet site, `www.health.gov.au/casemix`. (*see* **International Classification of Diseases (ICD)**)

While the US-developed DRG systems are widely used internationally, the Australian DRG system is also being adopted for use in a number of countries. Specifically, the Australian DRG system is used in New Zealand, a number of Asian countries and the German government has adopted this system with some local modifications for use within the German health system. A number of European countries have also developed case-mix systems for use within national health systems. While systems like the Nord DRGs developed and used in the Nordic countries and the Groupes Homogenes de Malades used in France could be considered to fall within the DRG-type framework, other systems like the Leistungsorientierte Krankenanstaltenfinanzierung (LKF) in Austria, the Health Resource Groups in England, and the Diagnose Behandelings Combinatie (DBC) in the Netherlands pursue quite different approaches. Given the dynamic nature of **health service** development, case-mix measures generally and DRG-type systems specifically may be expected to continue to evolve in response to advancements in diagnoses, treatment innovations, and technological developments. Given the range of case-mix measures now applied internationally, it would also be expected that continuing diversification of case-mix-type systems like DRGs will become increasingly in evidence with improvements in the availability of high-quality activity and cost data systems, developments in the skills base, and better access to the appropriate information technology.

References

[1] Averill, R.F., Muldoon, J.H., Vertrees, J.C., Goldfield, N.I., Mullin, R.L., Fineran, E.C., Zhang, M.Z., Steinbeck, B. & Grant, T. (1999). The evolution of case mix measurement using diagnosis related groups, in *Physician Profiling and Risk Adjustment*, 2nd Ed., N. Goldfield, ed. Aspen Publication, Maryland, 391–454.

[2] Commonwealth of Australia (Department of Health and Ageing) (2002). Australian Refined Diagnosis Related Groups, Versions 5.0, Definitions Manual, Volume 1.

[3] Fetter, R.B., Shin, Y., Freeman, J.L., Averill, R.F. & Thompson, J.D. (1980). Case mix definition by diagnosis related groups, *Medical Care* **18**(Suppl. 2), 1–53.

[4] Fetter, R.B. ed. (1991). *DRGs Their Design and Development*. Health Administration Press, Ann Arbor.

[5] Fetter, R.B. (1998). Diagnosis related groups, in *Encyclopedia of Biostatistics*, P. Armitage, & T. Colton, eds. John Wiley & Sons, Chichester.

[6] Goldfarb, M.G., Hornbrook, M.C. & Craig, C.S. (1983). Determinants of hospital use: a cross- diagnostic analysis, *Medical Care* **XXI**(1).

[7] Hornbrook, M.C. & Goldfarb, M.G. (1981). Patterns of obstetrical care in hospitals, *Medical Care* **19**, 55.

[8] Hornbrook, M.C. (1982). Hospital case mix: its definition, measurement and use: part I. The conceptual framework, *Medical Care Review* **39**, 1–43.

[9] Hornbrook, M.C. (1985). Techniques for assessing hospital case-mix, *Annual Review of Public Health* **6**, 295–324.

[10] Langenbrunner, J., Orosz, E., Kutzin, J. & Wiley, M. (2004). Purchasing and paying providers, in *Purchasing to Improve Health System Performance*, J. Figueras, R. Robinson & E. Jakubowski, eds. Open University Press, forthcoming.

[11] 3M Diagnosis Related Groups Definitions Manual (1998). Version 16.0, Wallingford.

[12] 3M International Refined Diagnosis Related Groups (2002). Definitions Manual Version 1.2, Wallingford.

[13] Wiley, M.M. (1999). Development and localisation of casemix applications for inpatient hospital activity in EU member states, *Australian Health Review* **22**(2), 69–85.

[14] Wiley, M.M., Tomas, R. & Casas, M. (1999). A cross-national, case-mix analysis of hospital length of stay for selected pathologies, *European Journal of Public Health* **9**(2), 86–92.

MIRIAM M. WILEY

Diagnostic Test Accuracy

A diagnostic test is said to have high accuracy if it achieves a high overall proportion of correct diagnoses. The accuracy of a test reflects the **prevalence** of disease in the population being tested and the conditional **misclassification** rates for true cases and noncases of disease, or equivalently on the prevalence, and the test **sensitivity** and **specificity**. Because the clinical implications are quite different for a **false positive** diagnosis on a noncase compared to a **false negative** diagnosis for a disease case, diagnostic test performance is usually not summarized by an overall accuracy index, but by more detailed measures, such as sensitivity and specificity, false positive and negative rates, **predictive values**, or by its **receiver operating characteristic (ROC) curve**.

(*See also* **Gold Standard Test**; **Diagnostic Tests, Evaluation of**; **Diagnostic Tests, Multiple**)

STEPHEN D. WALTER

Diagnostic Test Evaluation Without a Gold Standard

Diagnostic tests are an important part of medical decision making. In daily clinical practice many tests are performed to obtain diagnoses. To interpret a test it is important to realize that a negative answer does not always mean that the disease is absent, because **false negative** results may occur. Also, a positive result does not always mean that disease is present. A finding usually associated with a disease sometimes occurs in patients who do not have the disease: a **false positive** result.

A perfect test is positive in all patients with the disease and negative in all patients who do not have the disease. Usually this test is referred to as the gold standard. After applying the **gold standard test** one knows which patients have the disease and which patients are free of it. However, most tests are imperfect. Measures to assess the performance of a diagnostic test are **sensitivity** and **specificity**. The sensitivity of a test is defined as the proportion of positive test results in those with the disease. The specificity is defined as the proportion of negative test results in those without the disease. To measure the sensitivity and specificity of a test for a disease, the test's results are compared with those on a gold standard test, as shown in Table 1. The sensitivity is the ratio of the number of patients with true positive tests and the number of diseased patients. The specificity is the ratio of the number of patients with false negative tests and the number of nondiseased patients.

Problems arise when the sensitivity and the specificity of the reference test are unknown. The **2 × 2 table** of test vs. reference test contains too little information to estimate all unknown parameters, even if the **prevalence** of the disease is known.

Hui & Walter [3] pointed out that the parameters can still be estimated for two tests if data can be collected from populations with different prevalences and it can be assumed that the test errors are conditionally independent given the disease status.

Notation

Let D stand for disease status; $D = 1$ if diseased and $D = 0$ if not. Let T_i stand for the result of test i, $i = 1$ or 2; $T_i = 1$ if the test is positive and $T_i = 0$ if the test is negative. The sensitivity of test i is denoted by SENS_i and the specificity by SPEC_i, i.e. $\text{SENS}_i = \Pr(T_i = 1|D = 1)$ and $\text{SPEC}_i = \Pr(T_i = 0|D = 0)$. Let there be G subpopulations (groups) indexed by g with prevalences $\pi_g = \Pr(D = 1|\text{ group } g)$. Under conditional independence of T_1 and T_2 given D, the probabilities in the $T_1 \times T_2$ contingency table in group g are given by

$$\begin{aligned}\Pr(T_1 = t_1, T_2 = t_2|\text{group } g) &= \pi_g \times \text{SENS}_1^{t_1} \times (1 - \text{SENS}_1)^{1-t_1} \\ &\quad \times \text{SENS}_2^{t_2} \times (1 - \text{SENS}_2)^{1-t_2} \\ &\quad + (1 - \pi_g) \times (1 - \text{SPEC}_1)^{t_1} \times \text{SPEC}_1^{1-t_1} \\ &\quad \times (1 - \text{SPEC}_2)^{t_2} \times \text{SPEC}_2^{1-t_2}. \qquad (1)\end{aligned}$$

Estimation

The n_g observations in group g follow a 4-nomial distribution (*see* **Multinomial Distribution**) with these probabilities for the four cells. The number of parameters is $G + 4$ (G prevalences, two sensitivities, and two specificities). The number of **degrees of freedom** is $3G$, so the minimal requirement for **identifiability** of the model is that $G \geq 2$. If at least two prevalences are different, identifiability is indeed obtained, provided that it is assumed that

Table 1 The relationship between the results of a test and gold standard

Results of test for disease under study	Results of gold standard		Total
	Disease present	Disease absent	
Positive	True positive	False positive	Positive tests
Negative	False negative	True negative	Negative tests
Total	Diseased patients	Nondiseased patients	

SENS + SPEC > 1. (See [3] for the invariance of the problem under "reflection with respect to 1/2".) The parameters can be obtained by direct **maximum likelihood** applied to the joint **likelihood** of the G tables, as proposed by Hui & Walter [3] for the case when $G = 2$.

In de Bock et al. it is shown that the estimation problem can be solved elegantly by the **EM algorithm** [2]. Multinomial distributions with amalgamated cells is one of the examples given in that paper. The EM algorithm considers the true disease status D of all individuals as the missing observation. If this information were available, the data for each group could be conveyed in the $2 \times 2 \times 2$ table of $T_1 \times T_2 \times D$.

The M-step of the EM algorithm estimates all parameters in a straightforward way. Let $X_{g,ijk}$ be the number in the ijk-cell of the gth $2 \times 2 \times 2$ table. The index i (0, 1) corresponds to T_1, index j to test T_2, and k to disease D. Let "+" stand for summation; then the estimates are given by

$$\hat{\pi}_g = \frac{X_{g,++1}}{X_{g,+++}}, \tag{2}$$

$$\widehat{\mathrm{SENS}}_1 = \sum_g X_{g,1+1} \Big/ \sum_g X_{g,++1},$$

i.e. $$\frac{\text{number positive on } T_1 \text{ and diseased}}{\text{number diseased}}, \tag{3}$$

and

$$\widehat{\mathrm{SPEC}}_1 = \sum_g X_{g,0+0} \Big/ \sum_g X_{g,++0},$$

i.e. $$\frac{\text{number negative on } T_1 \text{ and not diseased}}{\text{number not diseased}}, \tag{4}$$

and similarly for SENS_2 and SPEC_2.

In the E-step, the full table $X_{g,ijk}$ is reconstructed from the available table $X_{g,ij+}$ by

$$\hat{X}_{g,ijk} = \hat{P}(D = k | T_1 = i, T_2 = j, \text{ group} = g) \times X_{g,ij+}. \tag{5}$$

In the estimated probabilities $\hat{P}$ the parameter estimates from the previous step are used.

For example,

$$\hat{P}(D = 1 | T_1 = 1, T_2 = 1, \text{group} = g) = \frac{\hat{\pi}_g \times \widehat{\mathrm{SENS}}_1 \times \widehat{\mathrm{SENS}}_2}{\left\{ \begin{array}{c} \hat{\pi}_g \times \widehat{\mathrm{SENS}}_1 \times \widehat{\mathrm{SENS}}_2 + (1 - \hat{\pi}_g) \\ \times (1 - \widehat{\mathrm{SPEC}}_1) \times (1 - \widehat{\mathrm{SPEC}}_2) \end{array} \right\}}. \tag{6}$$

The EM algorithm converges in a slow but sure way to the ML estimator. The process can be stopped if the increase in total log likelihood is smaller than some prespecified ε. To compute the standard errors of the estimated parameter, second derivatives of log likelihood have to be used as in Hui & Walter [3]. The advantages of EM are that it is very easy to program, and that the estimates never exceed the boundaries of the parameter space.

Generalizations

De Bock et al. [1] generalized the situation sketched above to the case where there are more than two tests, and in each group test results are available for precisely two tests. The generalization of the EM algorithm is straightforward.

A second generalization can be made by relaxing the condition of conditional independence. The model can be extended by introducing **odds ratios** lambda (λ) in the conditional 2×2 tables to model the dependency. The simplest model is when λ is constant. An extension would be to have different λs for diseased ($D = 1$) and not diseased ($D = 0$). See LeCessie & van Houwelingen [4] for a general discussion on modeling dependence in 2×2 tables.

References

[1] de Bock, G.H., Houwing-Duistermaat, J.J., Springer, M.P., Kievit, J. & van Houwelingen, J.C. (1994). Sensitivity and specificity of diagnostics tests in acute maxillary sinusitis determined by maximum likelihood in the absence of an external standard, *Journal of Clinical Epidemiology* **47**, 1343–1352.

[2] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM-algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[3] Hui, S.L., & Walter, S.D. (1980). Estimating error rates of diagnostic tests, *Biometrics* **36**, 167–171.
[4] LeCessie, S. & van Houwelingen, J.C. (1994). Logistic regression for correlated binary data, *Applied Statistics* **43**, 95–108.

(*See also* **Diagnostic Tests, Evaluation of**; **Diagnostic Tests, Likelihood Ratio**; **Diagnostic Tests, Multiple**)

G.H. DE BOCK & J.C. VAN HOUWELINGEN

Diagnostic Tests, Evaluation of

The field of diagnostic medicine is complex. In part, this is due to the fact that the process of medical diagnosis is dynamic, and it is difficult to formulate straightforward scientific questions amenable to simple study designs. For example, in interpreting the result of an individual test the doctor must consider the context in which it is applied. Has it been selected to rule-in or rule-out a diagnosis? What other tests have already been performed and what were their results? What options are available for performing subsequent tests? What are the characteristics of the patient that might predispose to the diagnosis under consideration? The evaluation of a test in the context of other tests is addressed elsewhere (*see* **Diagnostic Tests, Multiple**). In this article, discussion is limited to evaluation studies of individual tests, or comparisons of two alternative tests. Furthermore, we consider only *diagnostic* tests, i.e. tests of symptomatic patients in which we wish to rule-in or rule-out a candidate diagnosis. This contrasts with *screening* tests, performed on asymptomatic normal subjects, e.g. the use of mammography on a population at risk of breast cancer (*see* **Screening Benefit, Evaluation of**; **Screening, Models of**; **Screening, Overview**).

Many diagnostic tests, especially radiologic and **psychometric** tests, are evaluated subjectively, leading typically to test results that are classified in ordinal categories which are defined verbally. This contrasts with tests which possess quantitative results, as is the case for most laboratory tests. In either case, a useful conceptual device is to consider the test as having an underlying continuous scale, which may be discretized into ordinal categories either by judgment or by arithmetic rounding. The underlying continuous scale provides a metric for trading off the two different kinds of errors of diagnosis, **false positives** and **false negatives**, and thus for establishing a scale on which to calibrate the results of alternate tests for the purposes of comparison. The notation defined below reflects this assumption. There are various outcomes that can in principle be used to evaluate and compare the utility of medical diagnostic tests. The ultimate outcome involves evaluating whether use of the test, and the subsequent impact on medical therapy, leads to improvements in the natural history of the disease, e.g. lower mortality from the disease. It is rare for tests to be evaluated against this standard. Even studies of the impact of individual diagnostic tests on patient management are unusual. Overwhelmingly, medical researchers have been satisfied with evaluating tests on the basis of measures of diagnostic accuracy. In the following we thus limit attention to the issue of diagnostic accuracy.

Measures of Accuracy

Consider a diagnostic test result denoted x, and let D be a **binary** indicator of the "true" disease status, where $D = 1$ represents disease and $D = 0$ represents absence of disease. Let $F_x(x) = \Pr(X \leq x|D = 1)$ be the distribution of the test result in diseased cases, and let $G_x(x) = \Pr(X \leq x|D = 0)$ be the corresponding distribution in "control" subjects, i.e. patients suspected of having the disease who are candidates for testing in the relevant medical context. The most commonly used measures of accuracy are based on a binary classification of the test result. Suppose that the classification point is x, i.e. the test is positive if $X > x$ and negative otherwise. Then the **sensitivity** of the test is defined to be the proportion of diseased patients who are classified as diseased, i.e. $1 - F_x(x)$. The **specificity** is the corresponding proportion of control patients who are classified as normal, i.e. $G_x(x)$ [25]. High values of the sensitivity and the specificity indicate an accurate test.

There are several other measures in common usage related to the sensitivity and specificity. The *false positive ratio* is the specificity subtracted from 1, i.e. $1 - G_x(x)$. The *false negative ratio* is the sensitivity subtracted from 1, i.e. $1 - F_x(x)$. "Prospective" measures of accuracy are defined in terms of the conditional probabilities that the patient is diseased given the test results [22]. Thus the **positive predictive value** is

$$\Pr(D = 1|X \geq x) = \frac{\pi(1 - F_x(x))}{\pi[1 - F_x(x)] + (1 - \pi)[1 - G_x(x)]},$$

where $\pi = \Pr(D = 1)$ is the "prior" probability of disease, i.e. the **prevalence** of disease in the population under study. Likewise the negative predictive

value is defined to be

$$\Pr(D = 0|X \leq x) = \frac{(1-\pi)G_x(x)}{(1-\pi)G_x(x) + \pi F_x(x)}.$$

In fact, the term "accuracy" is often used in medical circles to mean the overall relative frequency of correct diagnosis in a study. That is, if $A(x)$ is the accuracy, then

$$A(x) = \pi[1 - F_x(x)] + (1-\pi)G_x(x).$$

Finally, the **likelihood ratio** can be used to represent the extent to which the odds of disease is altered as a result of the test, via **Bayes' Theorem** [19] (*see* **Diagnostic Tests, Likelihood Ratio**). In the context of a binary test there are two likelihood ratios, corresponding to a negative and a positive test result, $F_x(x)/G_x(x)$ and $[1 - F_x(x)]/[1 - G_x(x)]$, respectively.

Clearly, all of these measures are limited by the fact that they correspond to a specific, and possibly arbitrary, classification point, x. For the likelihood ratio in particular, knowledge of the actual test result, X, leads obviously to a more appropriate factor for updating using Bayes Theorem, i.e. $f_x(X)/g_x(X)$, where $f_x(\cdot)$ and $g_x(\cdot)$ are the corresponding density functions of the test result. Likewise, changing the classification point will either increase the sensitivity at the expense of the specificity, or vice versa, with corresponding effects on the error rates and the predictive values. The arbitrariness of the classification point is especially a problem when diagnostic tests are being compared, or when the same test is used in different studies with different classification points since the classifications for the two tests are unlikely to be "calibrated" in practice (see later discussion). For this reason **receiver operating characteristic (ROC) curve** analysis has become a preferred method for evaluating and comparing tests. The ROC curve is a plot of $F_x(x)$ vs. $G_x(x)$. If the ROC plot lies along the 45° line, then the test is random and hence uninformative. The higher the curve lies above the 45° line, the more accurate is the test. Thus the area under the curve is often used as a measure of accuracy that does not require a specific classification point. The area, A, is given by

$$A = 1 - \int_0^1 F_x(x)\, \mathrm{d}G_x(x). \qquad (1)$$

The area can be interpreted as the probability that a randomly chosen diseased subject has a test result that is greater than that of a randomly chosen control subject [11].

Biases

Despite the availability of the various measures of accuracy described in the previous Section, diagnostic tests are characterized predominantly by their sensitivities and specificities in the literature. These are often reported on the basis of a **retrospective** analysis of a series of patients treated in a hospital or clinic, and may suffer from incomplete reporting of study details, especially the factors affecting selection of patients for inclusion in the analysis. However, regardless of the quality of the studies, it is empirically evident that the ranges of values reported for the sensitivity and specificity of any important diagnostic test are usually very wide. A typical example is presented in the **meta-analysis** of the use of myelography for the detection of lumbar disk herniation, where the sensitivity estimates ranged from 75% to 98%, and the specificity estimates ranged from 20% to 100% [14]. Wide variation in reported estimates is due to the fact that studies of diagnostic test accuracy are plagued by a number of common **biases** [3]. These have been studied extensively in recent years, and various methods have been proposed for providing bias corrections.

Perhaps the most important factor causing variation in reported estimates of sensitivities and specificities is the fact that the classification point for the test may differ dramatically from study to study. This is, of course, not really a bias, but merely a definitional problem, and one that can be resolved in the meta-analytic context by plotting the pairs of sensitivity/specificity estimates in an ROC format [13]. Usually this will demonstrate the fact that sensitivities and specificities are inversely related, due to the fact that the classification point varies from study to study. Ideally, the classification points used would be defined in the individual studies, but often this is not the case. Indeed, for subjectively interpreted tests, the classification point (or points) cannot be defined precisely, and can only be inferred empirically. If variation in the reported sensitivities and specificities is due only to variation in the classification points used, then the plotted values should lie on a single ROC curve, except for random variation in the estimates. However, a much wider scatter is common, and this can be due to a number of possible biases.

Perhaps the most common biases are those due to problems with the "gold standard" reference test (*see* **Gold Standard Test**). These fall into two major categories: the problem of verification bias in which only a selected subset of patients receive the reference test and where unverified patients are ignored, and the problem in which the reference test is recognized to be an imperfect standard.

Verification bias is an especially serious problem since it has a counterintuitive aspect (*see* **Bias in Observational Studies**). The bias is caused if the selection of patients to receive the reference test is influenced by the result of the test under investigation [18]. If the study is restricted to patients who receive the reference test, say biopsy proven cases, then the study is biased, yet many investigators will believe that such a restriction follows sound scientific practice. It is often impossible to design an **unbiased** study since application of the possibly invasive reference test may be unethical when the test under investigation is negative. That is, the risks or inconvenience of the reference test may be considered to be medically inappropriate in the absence of a positive test. The standard bias correction method is based on the assumption that selection of a patient for the reference test is a conscious decision, and must therefore be based on available clinical factors, such as the test result, x, and the results of other relevant tests or patient factors, denoted collectively by z [5]. Consequently, the predictive values, conditional on x and z, can be estimated without bias from the verified sample, denoted $v+$, i.e.

$$\Pr(D = 1|x, z, v+) = \Pr(D = 1|x, z),$$

and so unbiased estimates of the distributions $F_x(\cdot)$ and $G_x(\cdot)$ can be obtained by combining these unbiased predictive values with the distribution of the test result unconditional on disease status, denoted $h_{x|z}(x)$, estimated from all subjects, both verified and unverified, using

$$f_{x|z}(x) \propto h_{x|z}(x) \Pr(D = 1|x, z, v+)$$

and

$$g_{x|z}(x) \propto h_{x|z}(x) \Pr(D = 0|x, z, v+).$$

Clearly, such an approach requires that data be collected on the test results and **covariates** of all patients in the series on whom the test is applied, regardless of whether the reference test is performed subsequently.

All of the accuracy measures described earlier are defined in relation to a "gold standard" reference test. Inaccuracy in the reference test will invariably lead to bias in the estimated characteristics of the test under consideration. If conditional independence between the two tests can be assumed, then bias corrections are possible. However, this assumption is usually untenable since most tests of the same phenomenon are likely to be positively correlated, even after conditioning on true disease status [24]. In these circumstances, the effect of the bias will be to inflate the sensitivity and specificity estimates artificially.

In evaluating the reported accuracy of diagnostic tests there are a number of other issues that can adversely affect the validity of the estimates, or cause further between-study variation in accuracy measures. Frequently, a test may produce an uninterpretable result. For example, for abdominal examinations bowel gas may obscure the result of ultrasound [17]. Barium in the gastrointestinal tract may obscure the result of computed tomography. A needle aspirate for the diagnosis of hepatic cancer may produce fragments which are inadequate for histological examination [20]. These problems are frequently not reported, the uninterpretable tests being simply removed from the analysis [17]. For subjectively interpreted tests, interobserver variation (*see* **Observer Reliability and Agreement**) can have a substantial impact. This may be reflected by variation in the empirical classification points used, or by genuine variation in accuracy, or both, and only ROC analysis can resolve these issues. Finally, the accuracy of a test may change over time, due to improvements in the ability of the readers to make use of the technology, or due to technological enhancements. For example, it is widely accepted that the quality and resolution of mammograms has improved markedly during the three decades since they became available, and the published accuracy results reflect this trend [9].

Comparisons of Tests

Consider two diagnostic tests, with results denoted by x and y. Parameters of test y have corresponding notation to those of test x, as defined earlier. Comparison of the two tests on the basis of accuracy requires that we calibrate the classification points used, otherwise, for example, the sensitivity of test x may be

larger than that of test y merely because the classification rule was more strict for test x. ROC analysis is a natural way of calibrating the comparison, and this is the reason for its use as the definitive analytic tool.

Calibration of the comparison at a specific classification point is achieved by equating the marginal distributions of the two tests, where the prevalence of disease is standardized. Let the marginal distributions be defined by

$$M_x(x) = \pi F_x(x) + (1 - \pi)G_x(x),$$

and

$$M_y(y) = \pi F_y(y) + (1 - \pi)G_y(y). \qquad (2)$$

Let $x_z = M_x^{-1}(z)$ and $y_z = M_y^{-1}(z)$. Then x_z and y_z represent the classification points corresponding to the zth quantile of these marginal distributions, i.e. $M_x(x_z) = M_y(y_z)$, for all z. It is easily seen from Eq. (2) that if the sensitivity of test x is greater than the sensitivity of test y at this quantile, then the specificity of test x is also greater than the specificity of test y, i.e.

$$1 - F_x(x_z) > 1 - F_y(y_z) \Longleftrightarrow G_x(x_z) > G_y(y_z).$$

Thus the tests are fully equivalent if and only if $1 - F_x(x_z) = 1 - F_y(y_z)$ and $G_x(x_z) = G_y(y_z)$, for every value of z, i.e. throughout the entire ROC curve. The preceding theory relies on π being common to the two tests. In a comparison study this is necessarily true in the paired-sample design (*see* **Crossover Designs**), and indeed this design is the common design in comparison studies for reasons of efficiency [12].

The widely used methods developed for ROC analysis have mostly focused on comparing the areas under the ROC curves rather than comparisons at different calibrated quantiles. In fact, if one constructs the combined ranked sample of test results for a given test, i.e. combining diseased and normal subjects, and then evaluates the Wilcoxon statistic for comparing diseased and normal subjects (*see* **Wilcoxon–Mann–Whitney Test**), then this statistic is the area under the **nonparametric** trapezoidal ROC curve [2]. The asymptotic **variance** of this statistic is well known under the **null hypothesis** that the test is uninformative, i.e. the area is 0.5, and can be modified easily for the more common circumstance in which the area is substantially greater than 0.5 [11]. In the paired-sample setting, i.e. where both diagnostic tests are applied to each subject, the **correlation** between the test results within each subject must be taken into account, and methods have been developed specifically for this purpose [7].

Parametric methods are also widely used, primarily based on the binormal model. In this model it is assumed that if the distribution of tests results in normal (control) subjects is transformed to a standard **normal distribution**, then the same **transformation** on the diseased subjects will also lead to a normal distribution, with mean μ_x and variance σ_x^2, say, for test x [8]. In this case the area under the ROC curve is given by

$$A = \Phi\left(\frac{\mu_x}{(1 + \sigma_x^2)^{1/2}}\right),$$

where $\Phi(\cdot)$ is the standard normal distribution function. However, it is conventional to test for equivalence of the ROC curves by simultaneously testing that $\mu_x = \mu_y$ and $\sigma_x^2 = \sigma_y^2$, rather than simply testing for equality of the areas. In the paired sample setting the two test results are assumed to have corresponding **bivariate normal distributions** in the diseased and nondiseased populations, with correlation parameters to account for the within-patient dependencies [16]. Widely distributed noncommercial software is available for performing these analyses [15].

Finally, methods have recently been proposed for testing the equivalence of the two tests at all possible classification points in a nonparametric fashion, using **bootstrapping** techniques [6]. This can be accomplished for continuous paired data by permuting within pairs the ranks of the marginal rank order statistic, and obtaining a permutation test [23]. Such an analysis does not rely on the comparison of a summarized (parametric) measure of accuracy, such as the area or the binormal parameters as outlined above.

Study Design

In designing a study to evaluate or compare diagnostic tests, great care is necessary to ensure that the data are collected in a format suitable for resolving the problems and biases outlined in the previous Section. A source of detailed practical advice, with an emphasis on radiologic imaging studies, is the text by Swets & Pickett [21]. A recent trend has been the development of multi-institutional field studies,

which parallel the early development of **multicenter trials**, and which have provided guidance on the organizational and methodologic challenges of large-scale accuracy studies [10].

There are five general issues pertinent to comparative studies of diagnostic tests: representativeness of the sample; completeness of data reporting; recording of test results; mapping of test results to "truth" data; and control of the comparison. First, representativeness is especially important since the ease with which a patient can be diagnosed accurately varies widely from patient to patient, and so a nonrepresentative sample of patients could substantially bias the estimates of accuracy. Secondly, completeness of data recording and reporting is important in the context of verification bias and the problems of uninterpretable test results. In cases where verification bias might be a problem, the ideal study is one in which we can be sure that all patients are scheduled for the definitive reference test (e.g. surgery) prior to the conduct of the tests under evaluation, otherwise the **missing data** are likely to be selective, i.e. not missing at random. Thirdly, as we have seen, valid comparison of tests is only possible if a "calibrated" analysis is achievable, i.e. using ROC analysis. Therefore, the test data must be collected in sufficient detail to facilitate such an analysis. That is, binary reporting of test results (positive or negative) is inadequate, and a minimum requirement is several ordinal classifications. Fourthly, it is critical that the data from the experimental tests and from the reference test are recorded in a manner that permit meaningful correlation. In medical imaging studies (*see* **Image Analysis and Tomography**), this means that a precise anatomical mapping of the results is possible. For example, if the purpose of the study is not only to detect disease, but to localize it, each of the test results, including the reference test, needs to be recorded for each of the anatomic regions of interest. Thus, careful form design (*see* **Questionnaire Design**) and data collection is essential. Fifthly, if the tests under evaluation are interpreted subjectively, it is critical that evaluation of the second test is accomplished without knowledge of the first. Thus, **blinding** the test readers, or **randomization** of the test order, is necessary to prevent bias.

Finally, as is the case for all research studies, an adequate sample size is necessary to reduce statistical variation in the accuracy estimates to a level that permits meaningful interpretation of the data (*see* **Sample Size Determination**). Various methods for calculating study **power** are available [11, 15, 21].

Current Developments

There has been a substantial recent increase in research activity in the biostatistical literature on methods pertaining to the evaluation of diagnostic tests. The major themes of this work were summarized in a recent review article [4]. Even more recently, an issue of *Academic Radiology* was devoted to statistical developments in this field pertinent to medical imaging studies [1]. All of the articles addressed either one of two topics: **covariate** analysis of ROC curves, including models for accommodating **random effects**, such as test readers; and **meta-analysis of diagnostic tests**. Interest in covariate modeling stems from recognition of the fact that studies of the accuracy of diagnostic tests can be influenced by multiple factors. Meta-analysis is important in recognition of the fact that there exists a vast literature of published diagnostic accuracy studies, and we need methods that can synthesize these results reliably, in recognition of the limitations and biases that may be present in the individual studies. The field promises to continue to be an active area of biostatistical research in the foreseeable future.

References

[1] Advances in statistical methods for diagnostic radiology: a symposium (1995). *Academic Radiology* **2**, S1–S84.

[2] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical Psychology* **12**, 387–415.

[3] Begg, C.B. (1987). Biases in the assessment of diagnostic tests, *Statistics in Medicine* **6**, 411–423.

[4] Begg, C.B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's, *Statistics in Medicine* **10**, 1887–1895.

[5] Begg, C.B. & Greenes, R.A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias, *Biometrics* **39**, 207–215.

[6] Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests, *Statistics in Medicine* **13**, 499–508.

[7] DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* **44**, 837–846.

[8] Dorfman, D.D. & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: rating method data, *Journal of Mathematical Psychology* **6**, 487–496.

[9] Fletcher, S.W., Black, W., Harris, R., Rimer, B.K. & Shapiro, S. (1993). Report on the international workshop on screening for breast cancer, *Journal of the National Cancer Institute* **85**, 1644–1656.

[10] Gatsonis, C. & McNeil, B.J. (1990). Collaborative evaluations of diagnostic tests: experience of the Radiation Diagnostic Oncology Group, *Radiology* **175**, 571–575.

[11] Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic curve, *Radiology* **143**, 29–36.

[12] Hanley, J.A. & McNeil, B.J. (1983). A method of comparing the area under two ROC curves derived from the same cases, *Radiology* **148**, 839–843.

[13] Irwig, L., Tosteson, A.N.A., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T.C. & Mosteller, F. (1994). Guidelines for meta-analyses evaluating diagnostic tests, *Annals of Internal Medicine* **120**, 667–676.

[14] Kardaun, J.W. & Kardaun, O.J. (1990). Comparative diagnostic performance of three radiologic procedures for the detection of lumbar disc herniation, *Methods of Information in Medicine* **29**, 12–22.

[15] Metz, C.E. Fortran programs ROCFIT, CORROC, LABROC, CLABROC. Department of Radiology, University of Chicago, 5841 South Maryland Avenue.

[16] Metz, C.E., Wang, P.L. & Kronman, H.B. (1984). A new approach for testing the significance of differences between ROC curves for correlated data, in *Information Processing in Medical Imaging*, F. Deconick, ed. Nijhoff, The Hague, pp. 432–445.

[17] Poynard, T., Chaput, J.C. & Etienne, J.P. (1982). Relations between effectiveness of a diagnostic test, prevalence of the disease and percentages of uninterpretable results, *Medical Decision Making* **2**, 285–302.

[18] Ransohoff, D.F. & Feinstein, A.R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests, *New England Journal of Medicine* **299**, 926–930.

[19] Sackett, D.L., Haynes, R.B. & Tugwell, P. (1985). *Clinical Epidemiology: A Basic Science for Clinical Medicine*, Little Brown & Company, Boston.

[20] Schwerk, W.B., Durr, H.K. & Schmitz-Moorman, P. (1983). Ultrasound guided fine-needle biopsies in pancreatic and hepatic neoplasms, *Gastrointestinal Radiology* **8**, 219–229.

[21] Swets, J.A. & Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

[22] Vecchio, T.J. (1966). Predictive value of a single diagnostic test in an unselected population, *New England Journal of Medicine* **274**, 1171–1173.

[23] Venkatraman, E.S. & Begg, C.B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment, *Biometrika* **83**, 835–848.

[24] Walter, S.D. & Irwig, L.M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, *Journal of Clinical Epidemiology* **41**, 923–938.

[25] Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques, *Public Health Reports* **62**, 1432–1449.

(*See also* **Diagnostic Test Evaluation Without a Gold Standard**)

COLIN B. BEGG

Diagnostic Tests, Likelihood Ratio

The use of **sensitivity** and **specificity** to quantify the performance of a diagnostic test in relation to the true presence or absence of a specific disease is now well established in the medical literature. The emphasis on the formal evaluation of the information yielded by a diagnostic test and its incorporation into the diagnostic process has been an important theme in what is now called **clinical epidemiology** [4]. A renewed methodologic interest in the evaluation of diagnostic tests has led to alternative approaches being proposed, both to quantify test performance and to compare more easily the relative merits of competing tests [2]. One of these newer techniques is the use of **likelihood ratios** [1]. The objective of this article is to define the likelihood ratio in the context of a diagnostic test, to explain its use in the diagnostic process, and to contrast it with other techniques (*see* **Diagnostic Tests, Evaluation of**).

The Diagnostic Process

The term "diagnostic test" is used here to denote any item of diagnostic information derived from patient history, physical examination, biochemical or histologic examination of tissue, blood, or urine, or some sort of imaging. For example, a clinician might use the presence or absence of heart murmur to help diagnose a defective heart valve or a test that detects the presence of a normally intracellular enzyme in the blood of a suspected heart attack victim. In these situations the clinician would modify his/her degree of belief that the patient has the disease of interest on the basis of the test result. This might lead in turn to the immediate application of therapy, a decision to conduct further diagnostic testing which is usually more costly and/or more invasive (e.g. a coronary angiogram or ultrasound) but more definitive, or possibly to cease further workup on the grounds that the disease is unlikely to be present.

The result of the diagnostic test may be inherently dichotomous (e.g. the presence or absence of a physical sign), ordinal (e.g. the grade of murmur), or purely quantitative, as in the case of many laboratory tests. Despite this quantification, test results are often dichotomized into so-called positive and negative results based on some predetermined cutpoint. The choice of cutpoint may be fairly arbitrary (e.g. the 95th percentile for "normal" patients) or selected to yield the best discrimination between truly diseased and not diseased subgroups. Optimally chosen cutpoints would, in addition, take into account the "costs" associated with the consequences of subsequent clinical decisions and the true disease status.

Clearly, the dichotomization of a test result discards some of the diagnostic information but it allows the clinician to incorporate more easily the test information into the diagnostic process. Dichotomization allows the performance of the test to be quantified and the straightforward computation of post-test probability of disease by **Bayes' theorem** [3].

Sensitivity, Specificity, and Predictive Value

In the context of a dichotomous diagnostic test, sensitivity and specificity define the test's inherent ability to be positive when disease is truly present and negative when it is absent. In other words

$$\begin{aligned}\text{sensitivity} &= \Pr(\text{positive test} \mid \text{disease present})\\ &= 1-\beta,\\ \text{specificity} &= \Pr(\text{negative test} \mid \text{disease not present})\\ &= 1-\alpha.\end{aligned}$$

The complementary probabilities are analogous to the type I (α) and type II (β) errors in the context of **hypothesis testing**, hence the notation.

Suppose we have a population of patients in which a proportion, p, truly have a particular disease and the remainder, $1-p$, do not. In other words, the background **prevalence** of disease is p. If the diagnostic test was conducted on each member of the population, then the distribution of the test results that would occur is displayed in Table 1. The expression within each cell of the table represents the proportion of the population with a particular combination of test result and true disease status. Note that the tacit assumption here is that the sensitivity and specificity are known for the test *in this population*.

Faced with a diagnostic challenge for an individual patient in this population, the physician would start from the prior expectation, p, that the disease is present. This would then be modified in

Table 1 Test results in patients with and without disease

Diagnostic test result	Disease status		Predictive value (post-test probability)
	Present	Absent	
Positive	$p(1-\beta)$	$(1-p)\alpha$	$\frac{p(1-\beta)}{p(1-\beta)+(1-p)\alpha}$
Negative	$p\beta$	$(1-p)(1-\alpha)$	$\frac{p\beta}{p\beta+(1-p)(1-\alpha)}$

the light of the test result for the patient, increasing the probability of disease if the test was positive and reducing it if it was negative. Exactly how much the *prior probability* is modified in the light of the test result depends on the test's sensitivity and specificity. In the population as a whole, the proportion $p(1-\beta)+(1-p)\alpha$ would test positive; of this total, $p(1-\beta)$ would actually have the disease (true positives) and $(1-p)\alpha$ would not (false positives). Thus, given a positive test, a proportion $p(1-\beta)/[p(1-\beta)+(1-p)\alpha]$ would truly have the disease. Similarly, a proportion $p\beta/[p\beta+(1-p)(1-\alpha)]$ would have the disease in those who tested negative. These quantities are referred to either as post-test probabilities or **predictive values** and by inspection can be seen to result from a direct application of Bayes' theorem.

Post-Test Odds

Odds are an alternative way of expressing the likelihood of an event. Saying that an event has odds of one to three of occurring (i.e. an odds of 1/3) is equivalent to saying the probability is one quarter. Probabilities are thus converted to odds by the relationship

$$\text{odds} = \frac{\text{probability}}{1-\text{probability}},$$

and odds back to probability by

$$\text{probability} = \frac{\text{odds}}{1+\text{odds}}.$$

If, in the diagnostic situation above, we express the post-test chance of disease in terms of odds, then we produce the following expressions:

$$\begin{aligned}&\text{post-test odds(positive test)}\\ &= \frac{p(1-\beta)/[p(1-\beta)+(1-p)\alpha]}{(1-p)\alpha/[p(1-\beta)+(1-p)\alpha]}\\ &= \frac{p}{1-p}\times\frac{1-\beta}{\alpha}\\ &= \text{pre-test odds}\times\frac{\text{sensitivity}}{1-\text{specificity}}.\end{aligned}$$

The quantity sensitivity/(1 − specificity) is called the *likelihood ratio* (LR) for a positive test result. A similar calculation for the odds of disease given a negative test result leads to

$$\begin{aligned}&\text{post-test odds(negative test)}\\ &= \frac{p}{1-p}\times\frac{\beta}{1-\alpha}\\ &= \text{pre-test odds}\times\frac{1-\text{sensitivity}}{\text{specificity}}\end{aligned}$$

and the quantity (1 − sensitivity)/specificity is called the *likelihood ratio* for a negative test result.

A Numerical Example

The present-day horseshoe crab *(Limulus Polyphemus)* remains virtually unchanged from its primeval ancestors. Its primitive defenses include a blood-clotting mechanism designed to isolate and encapsulate certain types of bacteria infecting its blood stream. A purified extract of horseshoe crab blood forms the basis of the limulus lysate test for detecting the presence of gram-negative infections in humans. This test can yield a result in about one hour compared with two or three days for the definitive blood culture. In a recent study of febrile patients, the limulus test was found to have sensitivity 79% and specificity 96% in a population with a 4% prevalence of septicemia [6]. Thus, using the post-test probability expressions above, we have

$$\text{Pr(septicemia} \mid \text{positive test)}$$

$$= \frac{0.04 \times 0.79}{0.4 \times 0.79 + (1 - 0.04)(1 - 0.96)}$$
$$= 0.4514,$$

$$\Pr(\text{septicemia} \mid \text{negative test})$$
$$= \frac{0.04 \times (1 - 0.79)}{0.04 \times (1 - 0.79) + (1 - 0.04) \times 0.96}$$
$$= 0.0090.$$

In other words, a positive test would increase the probability of septicemia from 4% to 45%, whereas a negative test would reduce it to less than 1%. From the quoted sensitivity and specificity:

$$\text{LR(positive test)} = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{0.79}{1 - 0.96}$$
$$= 19.75,$$
$$\text{LR(negative test)} = \frac{1 - \text{sensitivity}}{\text{specificity}} = \frac{1 - 0.79}{0.96}$$
$$= 0.2188.$$

A pre-test probability of 0.04 corresponds to an odds of $0.04/(1 - 0.04) = 0.0417$. The post-test odds are then:

$$\text{post-test odds (positive test)} = 0.0417 \times 19.75$$
$$= 0.8236,$$
$$\text{post-test odds (negative test)} = 0.0417 \times 0.2188$$
$$= 0.0091.$$

Conversion from odds back to probability yields the same probabilities as above. Note that a positive test causes the odds to be multiplied by almost 20, whereas a negative test requires the odds to be reduced by a factor of about five.

Generalization of the LR

The initial objective of the diagnostic workup is to determine the probability that the patient has the disease in the light of the test result. While sensitivity and specificity are simple statistics that describe test performance, their combined influence on post-test probability is not obvious. By contrast, LR has a direct multiplicative effect on pre-test odds, making the impact of the additional diagnostic evidence provided by the test more apparent. The calculation of post-test odds can be done approximately using a little mental arithmetic. Alternatively, simple nomograms are available which convert pre-test to post-test probability via the appropriate LR [4].

The most important advantage of LR is that it can be generalized to handle ordinal (*see* **Ordered Categorical Data**) or purely quantitative tests [1]. While the computation of post-test probability via Bayes' theorem can incorporate a quantitative test result, the terms sensitivity and specificity can only be directly applied if the quantitative test is first dichotomized at some cutpoint into positive and negative categories. Other than for simplicity, there seems little justification for collapsing a quantitative test into this dichotomy. A patient whose test result was only just over the cutpoint for positivity would be assigned the same post-test probability as a patient whose test result was extremely elevated. At some stage the physician must make the decision whether the patient has, or does not have, the disease. However, this ultimate dichotomization should be based on actual post-test probability, not on the intermediate test result. This issue would be especially important if the current test was only one step in a more extensive sequential diagnostic workup.

For a continuous test result, X, the definition of LR is

$$\text{LR}(X) = \frac{\Pr(X \mid \text{patient diseased})}{\Pr(X \mid \text{patient not diseased})}.$$

These probabilities, and thus the LR, can be estimated empirically for an inherently ordinal test (e.g. the traditional +, ++, +++ grading of heart murmur) or, for a purely quantitative test, by dividing the test result into subranges. In either case, the post-test odds are computed as the product of pre-test odds and LR, but now the LR is computed for the patient's own test result, as opposed to an average LR for all test results above or below a cutpoint.

Empirical LR Estimates

Table 2 shows LR computed for various ranges of creatine kinase (CK) for patients with and without myocardial infarction (MI) from a study by Smith [5]. Now, by using the LR for the range of the test result obtained, one can compute a more specific post-test probability. Obviously, the more extreme CK value

Table 2 Creatine kinase (CK) in patients with and without myocardial infarction (MI)

CK range	Proportion of MI patients (A)	Proportion of non-MI patients (B)	LR (A/B)
≥280	97/230 0.4217	1/130 0.0077	54.8
200–279	37/230 0.1609	2/130 0.0154	10.5
120–199	51/230 0.2217	5/130 0.0385	5.8
40–119	43/230 0.1870	34/130 0.2613	0.71
<40	2/230 0.0087	88/130 0.6769	0.013

of ≥280 would lead to a higher post-test probability of MI compared with a relatively mildly elevated CK at say 150.

Purely Continuous LR

Although going some way to creating a LR for each level of the test result, in the example above we have had to combine patients within a range of CKs to provide enough data points to estimate the LR. Conceptually, the purely continuous test result situation is depicted in Figure 1. The individual test results for patients with and without the disease form two continuous *distributions* where the *heights* of the curves at any point are the relative frequencies of that test result for the populations of patients with and without the disease. By definition, the LR at any value of the test result, X, is the ratio of the relative frequency (i.e. probability density) of X in the diseased to nondiseased distributions. If the data for the test results from representative samples of diseased and nondiseased patients can be adequately described by some appropriate mathematical model, then the LR can in turn be described mathematically at each X. For example, if both samples of test results were **normally distributed** with different **means** and **variances** so that

$$X \sim \mathrm{N}(\mu_1, \sigma_1^2) \text{ for diseased patients,}$$

$$X \sim \mathrm{N}(\mu_2, \sigma_2^2) \text{ for nondiseased patients,}$$

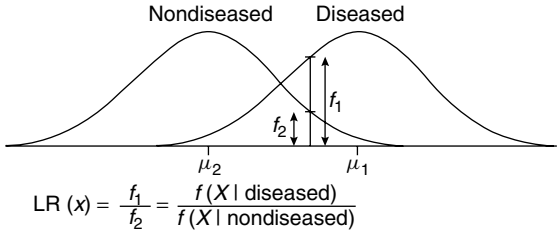

Figure 1 Likelihood ratio for a quantitative test

then the LR would be

$$\mathrm{LR}(X) = \frac{\left(2\pi\sigma_1^2\right)^{-1/2} \exp\left[-\dfrac{(X-\mu_1)^2}{2\sigma_1^2}\right]}{\left(2\pi\sigma_2^2\right)^{-1/2} \exp\left[-\dfrac{(X-\mu_2)^2}{2\sigma_2^2}\right]},$$

which simplifies a little to

$$\mathrm{LR}(X) = \frac{\sigma_2}{\sigma_1} \exp\left[-0.5(z_1^2 - z_2^2)\right],$$

where the z_1 and z_2 are the standardized deviates (*see* **Standard Normal Deviate**) of the test result X with respect to the diseased and nondiseased distributions, respectively.

The CK data from Smith are almost perfectly **lognormally distributed** with $\log_e$ CK having mean = 5.45 and SD = 0.737 for MI patients and mean = 3.19, SD = 1.030 for noninfarct patients. Substitution of these parameter estimates into the expression above leads to the continuous curve (Figure 2), which provides a value for LR corresponding to each individual value of the CK test.

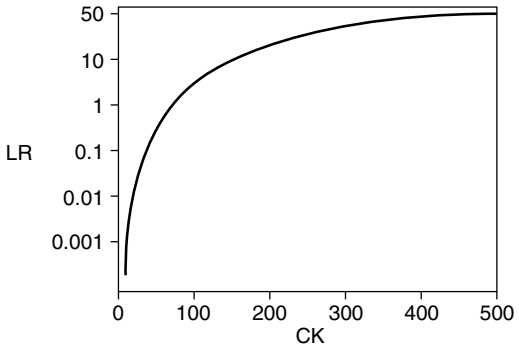

Figure 2 Continuous likelihood ratio curve

Summary

In its simplest form, LR is a useful measure of the diagnostic information conveyed by a diagnostic test. Compared with sensitivity and specificity, it offers a more interpretable measure of the impact of the test result on the probability of disease and also a simplification in the calculation, especially if one is prepared to think in terms of odds rather than probability. More importantly, it allows a natural extension to accommodate truly quantitative test results. The LR can reflect the diagnostic information at any level of test result. This leads to a post-test probability for the actual test result observed in the patient as opposed to a less specific post-probability computed from a test result which has been first dichotomized into positive and negative categories. Full utilization of the diagnostic information would require knowledge of the LR at each value of X and this may be quite feasible in many clinical situations.

References

[1] Albert, A. (1982). On the use and computation of likelihood ratios in clinical chemistry, *Clinical Chemistry* **28**, 1113–1119.

[2] Begg, C.C. (1991). Advances in statistical methodology for diagnostic medicine in the 1980s, *Statistics in Medicine* **10**, 1887–1895.

[3] Brown, B.W., Jr (1977). *Statistics – A Biomedical Introduction*. Wiley, New York, pp. 25–30.

[4] Sackett, D.L., Haynes, R.B. & Tugwell, P. (1985). *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Little, Brown & Company, Boston.

[5] Smith, A.F. (1967). Diagnostic value of serum creatine-kinase in a coronary care unit, *Lancet* **2**, 178–182.

[6] van Deventer, S.J., Büller, H.R., ten Cate, J.W., Sturk, A. & Pauw, W. (1988). Endotoxaemia: an early predictor of septicaemia in febrile patients, *Lancet* **1**, 605–609.

ROBIN S. ROBERTS

Diagnostic Tests, Multiple

In medical practice, diagnostic tests are rarely used in isolation. Typically, the evidence for forming a diagnosis comes from multiple sources, in the form of signs and symptoms and other patient characteristics, in addition to specific tests that are ordered to rule in or rule out a candidate diagnosis. Thus, the evidence from individual diagnostic tests must be used collectively in forming the diagnosis, and the statistical dependency of the information from these multiple sources becomes an important factor in assessing the weight of evidence. In this article we consider the two settings in which data from many diagnostic tests are relevant. In the first, we consider **discriminant analysis**, where one wishes to develop diagnostic rules on the basis of an available data set encompassing information on multiple tests. In the second setting we consider the problem of sequentially updating diagnostic probabilities as new tests are selected and performed in the course of making a diagnosis for an individual patient.

Discriminant Analysis

The general goal of discriminant analysis is to provide a statistical framework for characterizing two or more diagnostic categories on the basis of a set of diagnostic indicator variables (e.g. diagnostic tests), either to provide allocation rules, ideally with low error rates, or to provide realistic probabilities of the diagnostic categories for an individual (future) patient to facilitate decisions about medical management of the patient. Suppose that there is a baseline diagnostic category, denoted by zero, and an additional t diagnostic categories, where $z_d = 1$ is the patient belongs to the dth category, $d = 0, 1, \ldots, t$, and $z_d = 0$ otherwise. Let there be p diagnostic test variables, denoted $s = (s_1, \ldots, s_p)$. Let $\phi_d(\cdot)$ denote generically the probability of category d given the values of the variables in parentheses, and let $f_d(\cdot)$ denote the corresponding **sampling distribution** given the diagnostic category. As examples, $\phi_d(s) = \Pr(z_d = 1|s)$ is the probability that a patient with test vector s belongs to diagnostic category d, while $f_d(s_1|s_2, \ldots, s_p)$ is the conditional density function of s_1, given $s_2, \ldots, s_p$, and $z_d = 1$.

Discriminant analysis is a widely used statistical technique available in numerous commercial statistical packages (*see* **Software, Biostatistical**). The traditional formulation involves the assumption of **multivariate normal distributions** for $f_d(s)$. If a common **covariance matrix** across diagnostic categories can be assumed, then the **likelihood ratio** distinguishing any two diagnostic categories (*see* **Diagnostic Tests, Likelihood Ratio**) is a linear function of s, and so allocation of patients into diagnostic categories can be based on these linear functions. Diagnostic probabilities can then be obtained using **Bayes' theorem** and the **prevalences** or "prior" probabilities of the diagnoses. For a review of traditional approaches to discriminant analysis, see Lachenbruch [6].

An alternative "paradigm" is to model the diagnostic probabilities $\{\phi_d(s)\}$ directly, using models such as **logistic regression** [4]. The linear logistic model, for example, is consistent with the assumption of normal sampling distributions with equal covariance matrices, but is based on a more parsimonious and directly relevant parameterization than the modeling of the sampling distributions. Specifically, the simple linear logistic model is represented as follows:

$$\ln\left(\frac{\phi_d(s)}{\phi_0(s)}\right) = \beta_0 + \beta_1 s_1 + \cdots + \beta_p s_p, \qquad d = 1, \ldots, t.$$

By introducing additional terms the model can be made more flexible. For example, quadratic terms or **interaction** terms could be introduced to accommodate discrimination between multivariate normal populations with unequal covariances. Use of the predicted probabilities of disease categories obtained from a logistic regression on new patients requires the assumption that the relative frequencies of the diagnoses in the "training" data set, denoted $\pi_0, \pi_1, \ldots, \pi_t$, be representative. If one wants to alter these "prior" probabilities to $\pi_0^*, \pi_1^*, \ldots, \pi_t^*$, then the log odds comparing category d with the baseline would need to be changed by the addition of the factor $\log(\pi_d^*\pi_0/\pi_0^*\pi_d)$.

The literature on discriminant analysis encompasses numerous other modeling strategies, including formal **Bayesian** approaches that provide more realistic (i.e. more conservative) diagnostic predictions, kernel estimation techniques (*see* **Density Estimation**), and others. Recently, much attention has been directed toward computer-intensive techniques,

especially those based on classification and regression trees (CART) (*see* **Tree-structured Statistical Methods**), and also methods based on computerized **neural networks**. CART methods involve the partitioning of s into mutually exclusive subgroups of diagnostic test results that are diagnostically informative. It has been argued that this facilitates the interpretation of the rules for clinicians [2]. Neural networks involve the creation of a possibly hierarchical structure of test results, and these tend to be more heavily parameterized than conventional statistical models. For a review, see Cheng & Titterington [3].

Regardless of how the discriminant model is selected and applied, its validity can only be assessed definitively by evaluating its performance on a different data set from the one in which it was derived. In general, parametrically rich models will appear to perform relatively well when the error rates are evaluated on the data set on which the model was derived, and so the inevitably improved fit due to adding greater complexity to the model may be illusory. In the absence of a "test" data set to complement the "training" set, **cross-validatory** techniques can provide guidance on the degree of overfitting [9].

Sequential Testing

In medical practice a diagnosis is not usually reached either on the basis of a single definitive test or on a set of prescribed tests that could form the basis of a discriminant analysis, as outlined above. Tests may be applied in sequence, in an effort to ascertain a single diagnosis with high confidence. Doctors usually select the sequence of tests using medical judgment, and the level of confidence in the resulting diagnosis is assessed nonquantitatively. In recent years clinical epidemiologists have addressed the issue of how to evaluate these diagnostic probabilities quantitatively. Sackett et al. [8] provide examples of this process. The updating of diagnostic probabilities on the basis of the result of a new test, say s_k, given the information from previous tests, say $s_1, \ldots, s_{k-1}$, is the critical statistical problem in this setting.

Updating of diagnostic probabilities can be accomplished using **Bayes' theorem**, which is most conveniently expressed in its **odds ratio** form:

$$\frac{\phi_d(s_1, \ldots, s_k)}{\phi_0(s_1, \ldots, s_k)} = \frac{\phi_d(s_1, \ldots, s_{k-1})}{\phi_0(s_1, \ldots, s_{k-1})} \times \frac{f_d(s_k|s_1, \ldots, s_{k-1})}{f_0(s_k|s_1, \ldots, s_{k-1})}, \quad (1)$$

where $\phi_d(s_1, \ldots, s_{k-1})/\phi_0(s_1, \ldots, s_{k-1})$ is the "prior" odds before the test result s_k is obtained. Setting $\alpha_d(\cdot) = \phi_d(\cdot)/\phi_0(\cdot)$, $d = 1, \ldots, t$, the pairwise odds in (1) are easily converted to probabilities using

$$\phi_d(\cdot) = \frac{\alpha_d(\cdot)}{1 + \alpha_1(\cdot) + \cdots + \alpha_t(\cdot)}, \quad d = 1, \ldots, t.$$

In fact, if we consider, for expository purposes, that each of the tests has been administered sequentially, then it is convenient to express (1) in the form

$$\frac{\phi_d(s_1, \ldots, s_k)}{\phi_0(s_1, \ldots, s_k)} = \frac{\pi_d}{\pi_0} \frac{f_d(s_1)}{f_0(s_1)} \frac{f_d(s_2|s_1)}{f_0(s_2|s_1)} \cdots \times \frac{f_d(s_k|s_1, \ldots, s_{k-1})}{f_0(s_k|s_1, \ldots, s_{k-1})},$$

where each of the sequence of likelihood ratio terms represents the appropriate factor, at that stage, for updating the disease probabilities.

Sequential updating of probabilities using Bayes' theorem has become established as the theoretical paradigm for handling quantitative diagnostic information [8]. However, clinical epidemiologic texts rarely emphasize the critical requirement that the likelihood ratios should be conditioned on the data that have already been used to determine the current "prior" probabilities. In fact, the "current" prior should only be updated using the unconditional likelihood ratio $f_d(s_k)/f_0(s_k)$ if the test s_k is conditionally independent of the "tests" that have contributed to the current prior odds [1]. In practice, this is unlikely to be even approximately true, and so the use of unconditional likelihood ratios will tend to lead to overoptimistic, i.e. anti-conservative, diagnostic probabilities [5]. Moreover, in the absence of such conditional independence, the data requirements to provide the appropriate conditional probabilities for each step in a series of tests are formidable, and such data are generally not available [7].

The application of Bayes' theorem in updating diagnostic probabilities is most commonly utilized

in the construction of **decision analytic** models [10]. It is unlikely that many doctors utilize this formalized paradigm routinely in day-to-day clinical settings.

References

[1] Begg, C.B. (1994). Commentary on article by Miettinen and Caro, *Statistics in Medicine* **13**, 211–212.

[2] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.

[3] Cheng, B & Titterington, D.M. (1994). Neural networks: a review from a statistical perspective, *Statistical Science* **9**, 2–54.

[4] Dawid, A.P. (1976). Properties of diagnostic data distributions, *Biometrics* **32**, 647–658.

[5] Fryback, D.G. (1978). Bayes' theorem and conditional non-independence of data in medical diagnosis, *Computers in Biomedical Research* **11**, 423–434.

[6] Lachenbruch, P.A. (1982). Discriminant analysis, in *Encyclopedia of Statistical Sciences*, Vol. 2. S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 389–397.

[7] Miettinen, O.S. & Caro, J.J. (1994). Foundations of medical diagnosis: what actually are the parameters involved in Bayes' theorem?, *Statistics in Medicine* **13**, 201–209.

[8] Sackett, D.L., Haynes, R.B., Guyatt, G.H. & Tugwell, P. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Little, Brown, & Company, Boston.

[9] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society, Series B* **36**, 111–147.

[10] Weinstein, M.C. & Fineberg, H.V. (1980). *Clinical Decision Analysis*. Saunders, Philadelphia.

(*See also* **Diagnostic Tests, Evaluation of**)

COLIN B. BEGG

Diagnostics

Introduction

Diagnostics are methods for identifying and understanding differences between a model and the data to which it is fitted. This article is mainly concerned with the **linear regression** model, although the techniques of regression diagnostics, especially the study of the effect of the deletion of observations, are more widely applicable. Extensions to other models are described in the last section.

Some differences between the data and the model may be due to isolated observations: one, or a few, observations may be **outliers**, or may differ in some unexpected way from the rest of the data. Other differences may be systematic, for example, a term may be missing in a linear model. Systematic departures can often be detected by aggregate statistics, that is, quantities calculated over all the data, such as a t or F test in regression. But there is the important possibility that the evidence, for example, for an extra term, or a **transformation** of the response, may be being unduly influenced by a few observations. The main emphasis of diagnostics in statistical usage is on the effect of individual cases (observations y and the associated vectors of **explanatory variables** or carriers $\boldsymbol{x}$) on inferences about the model. These effects are customarily determined by deletion of individual cases. Exact formulae for the effects of deletion in regression mean that only one fit to the data yields the required diagnostic quantities. When exact formulae are not available, for example, for **Generalized Linear Models**, similar techniques yield useful approximations to the effect of deletion. In some statistical fields, such as econometrics, the term diagnostics is often taken to include aggregate statistics. An example is [Harvey 11, Section 5.4]. Here we discuss deletion diagnostics. Related material is included in the entries **Forward Search**, **Goodness of Fit** and **residuals**.

Outliers

An outlier is an isolated observation that does not agree with the model fitted to the majority of the data. The statistical problem is that fitting the model may disguise the presence of outliers. For simple samples, the effect of fitting is often not crucial. For example, a human **birth weight** recorded as 35 kg, perhaps due to multiplication of the weight by 10 on data entry, is clearly wrong, whatever model is fitted. The early chapters of Barnett and Lewis [5] describe the history of the definition of outliers and methods for their detection in univariate samples. For more complicated models, such as regression, large residuals indicate outliers, but outliers do not necessarily give rise to large **least-squares** residuals, especially if they occur at remote points in the factor space, which is at "leverage points". More formally, the least-squares estimate of the parameters in the linear regression model is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{1}$$

where $\boldsymbol{X}$ is the $n \times p$ **matrix** of carriers, that is, of explanatory variables and perhaps functions of them, such as quadratics and **interactions**. It is assumed that the additive errors of observation $\boldsymbol{\varepsilon}$ are independently distributed with constant variance σ^2. The least-squares residuals are then given by

$$\begin{aligned}\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}. \end{aligned} \tag{2}$$

In (2) $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\boldsymbol{H}$ is the "hat" matrix, so-called because $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. The diagonal elements h_i of $\boldsymbol{H}$ are such that $0 \le h_i \le 1$, with average value p/n. Observations with "large" values of h_i are said to be leverage points.

The residuals $\boldsymbol{e}$ are not independent, nor do they have the same variance since $\mathrm{var}(e_i) = \sigma^2(1 - h_i)$. The standardized residuals r_i are given by

$$r_i = \frac{e_i}{\sqrt{\{s^2(1 - h_i)\}}}, \tag{3}$$

where $s^2 = \sum e_i^2/(n - p)$ is used to estimate σ^2. The distribution of r_i^2 is a scaled **Beta distribution**.

The t test (*see* **Student's *t* Distribution**) for the hypothesis that observation i is an outlier is based on the deletion residual r_i^*, which is found by the deletion of case i or, equivalently, by fitting the mean shift outlier model

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{d}_i^T\phi, \tag{4}$$

where $\boldsymbol{d}_i$ is an $n \times 1$ vector of zeroes except for a 1 in the ith position. If $\hat{\boldsymbol{\beta}}_{(i)}$ is the least-squares estimate

of $\boldsymbol{\beta}$ when y_i is not used in fitting, the prediction at $\boldsymbol{x}_i$ is $\hat{y}_{(i)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$. The deletion residual is then

$$r_i^* = \frac{y_i - \hat{y}_{(i)}}{\text{s.e. } (y_i - \hat{y}_{(i)})} = \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}}{\text{s.e. } (y_i - \hat{y}_{(i)})}.$$
$$= \frac{e_i}{\sqrt{\{s_{(i)}^2 (1 - h_i)\}}}. \quad (5)$$

In (5) $s_{(i)}^2$ is the residual mean square estimate of σ^2 on $n - p - 1$ degrees of freedom after the deletion of case i. From the further results discussed in **residuals**, it follows that r_i^* can be calculated from the standardized residuals since

$$r_i^* = \frac{r_i}{\sqrt{\left\{\frac{(n - p - r_i^2)}{(n - p - 1)}\right\}}}. \quad (6)$$

In the absence of outliers, the deletion residuals have a t distribution on $n - p - 1$ degrees of freedom so that, unless the degrees of freedom are small, they will give an almost straight normal probability, or $Q-Q$, plot. If the straightness is in doubt, a simulation envelope can be generated of the type described in **residuals**.

As an example, we turn to the analysis of the data from Royston and Altman [14] on mandible length as a function of **gestational age** in 167 fetuses with ages from 12 to 33 weeks. The data are plotted in Figure 2 of **goodness of fit**. The analysis there and in **residuals** suggests that a transformation of y should be taken. Whether or not y is transformed, the skeleton **analyses of variance** in Table 1 indicate that a quadratic in age should be included (*see* **Polynomial Regression**). The F values for both linear and quadratic terms are higher for the transformed data than for the original, especially for the quadratic term. The F value for the cubic term is significant at the 5% level, but not at 1%. Given the number of observations that we have, we ignore the cubic term and work with a quadratic model in the gestational age, x, with log y as the response. Diagnostic scrutiny of both transformation and linear model in the articles on **fan plot** and **forward search** show that this choice does not depend on just a few observations. We also note that the doctors quoted by Royston and Altman felt that younger fetuses might differ systematically from those older than 28 weeks. These are cases 159 to 167.

Table 1 Skeleton analysis of variance for regression models and transformations fitted to data on mandible length

Response	y	log y
Source	F	F
Age	1468.7	1851.1
$(\text{Age})^2$	20.2	161.3
$(\text{Age})^3$	4.8	5.0

Figure 1 shows a normal $Q-Q$ plot of the deletion residuals for all cases. There is no evidence of any outliers. This plot is very different from Figure 3 of the article on **residuals** when a first-order model is fitted to the untransformed data. That $Q-Q$ plot showed three appreciable outliers, cases 149, 165, and 166, which lie well below the trend of the rest of the observations in the scatter plot of Figure 5. The logarithmic transformation of the data seems to reconcile these cases with the remaining data. However, their importance may be masked by the least-squares fit. The purpose of the methods described in this article is to assess the effect of such cases on the fitted model and conclusions drawn from it. Since the $Q-Q$ plot of Figure 1 is virtually straight, further analysis can proceed on the assumption that the residuals, and so the errors, are normally distributed.

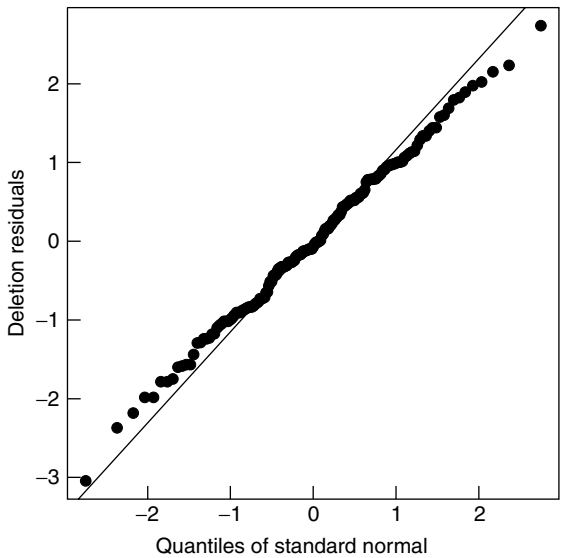

Figure 1 Mandible length data. Normal $Q-Q$ plot of deletion residuals r_i^* from a quadratic model in age and a logged response. The data seem well fitted by this model

Leverage

Several diagnostic measures are combinations of the least-squares residuals e_i and the leverage measures h_i. It is sometimes informative to look at the h_i on their own. Figure 2 is an index plot of h_i, that is, a plot against observation number, for the 167 observations on mandible length. Since the observations are in order of increasing y, they are pretty much in order of increasing x as well. It is clear from the figure, and expected, that observations at the extreme values of x have the highest leverages. For **multiple regression** data, the pattern is often less clear, although sometimes informative. It is however, the combination of y with x, which is important in making inferences about appropriate models.

Influence and Cook's Distance

The deletion residuals r_i^* provide a test for outliers. However, an outlier may or may not have an important effect on inferences drawn from the data. This information can be obtained by considering the change in the parameter estimates when case i is deleted, that is, the distance $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$, or components of this distance, preferably suitably scaled. Cook's distance [7] provides a measure of influence for all parameters based on the increase in the residual sum of squares for all the data when the deletion estimate $\hat{\boldsymbol{\beta}}_{(i)}$ replaces $\hat{\boldsymbol{\beta}}$. Cook's distance is defined as

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2}. \tag{7}$$

The use of standard deletion formulae (Cook and Weisberg [8, p. 210]; Atkinson [1, p. 19]) leads to the alternative form

$$D_i = \frac{e_i^2 h_i}{\{ps^2(1-h_i)^2\}} = \frac{r_i^2 h_i}{\{p(1-h_i)\}}. \tag{8}$$

For plotting purposes, Atkinson [1, p. 25] suggests the modified Cook Statistic C_i, which is a scaled square root of D_i with s^2 replaced by $s_{(i)}^2$. Thus,

$$C_i = \left(\frac{n-p}{p}\frac{h_i}{1-h_i}\right)^{\frac{1}{2}} |r_i^*|. \tag{9}$$

Figure 3 is an index plot of C_i for the 167 observations of the mandible length data. Six cases, 1, 7, 149, 163, 165, and 167 have large values of the modified Cook statistic, suggesting that they may be unduly influencing the parameters of the fitted model. Three of these are cases, which were originally regarded with some suspicion, three are not. To demonstrate

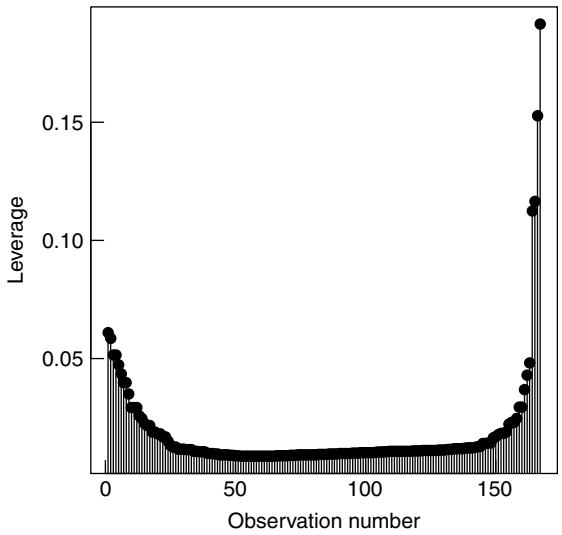

Figure 2 Mandible length data. Index plot of leverage measure h_i for all 167 cases

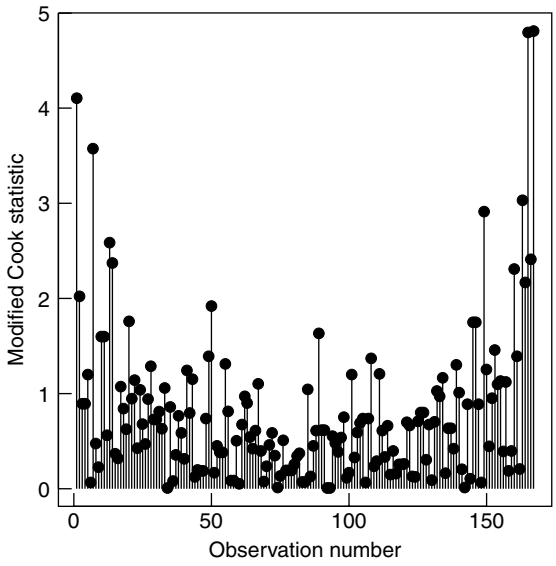

Figure 3 Mandible length data. Index plot of modified Cook's distance C_i for all 167 observations. Cases 1, 7, 149, 163, 165, and 167 appear highly influential

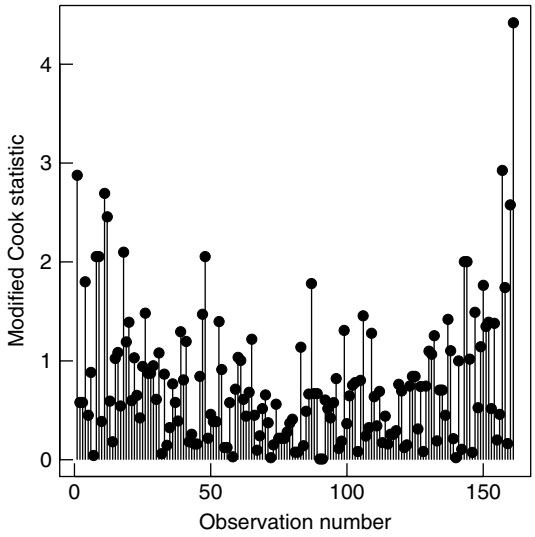

Figure 4 Mandible length data. Index plot of modified Cook's distance C_i after the deletion of six cases. Now original case 166 appears especially influential

what information can be obtained using diagnostic methods, we omit these six and refit to obtain the plot of Figure 4. Now original case 166 appears as especially influential.

As a result of this analysis, seven out of 167 cases have been identified as potentially outlying or otherwise different. It must be stressed that the purpose of diagnostic analyses is not to delete all cases with any egregious characteristics. Rather the purpose is to identify such cases, to ascertain their importance and, if they are crucial to an understanding of the data, to check for transcription errors, unsuspected variations in conditions of measurement or source of the data, or other explanations of apparent anomalies.

Multiple Deletion and very Robust Regression

The deletion diagnostics for individual case deletion can readily be extended to a subset of m observations with index I. However, such procedures are not as useful as the deletion of individual cases, due to the combinatorial explosion in the number of quantities to be considered. Even for the 167 cases of the mandible length data, there are 13 861 diagnostics from deletion of pairs. Often, information on diagnostic matters can be extracted by the repeated use of single deletion diagnostics. But sometimes this can fail, a condition known as "masking", caused for example, by the presence of several outliers at leverage points. In such conditions very **robust regression** can be used, which resists up to 50% of outliers in the data, and the results compared with those from least-squares analyses.

In least trimmed squares (*see* **Trimming and Winsorization**), the model is fitted by least squares to the q data points giving the smallest residual sum of squares. The greatest robustness against outliers is obtained by taking q as approximately half the data, that is,

$$q = \left[\frac{n}{2}\right] + \left[\frac{(p+1)}{2}\right], \qquad (10)$$

when allowance is made for fitting. Calculation of the least trimmed squares estimate involves repeated searches from random starting points. Figure 5 shows the fitted line obtained using the algorithm in the statistical package **S-PLUS**. Although log(length) was used as the response in all models fitted, the plot is on the original scale of the data. One consequence is that the outlying nature of cases 149, 165, and 166 on the original scale is apparent. The plot also shows

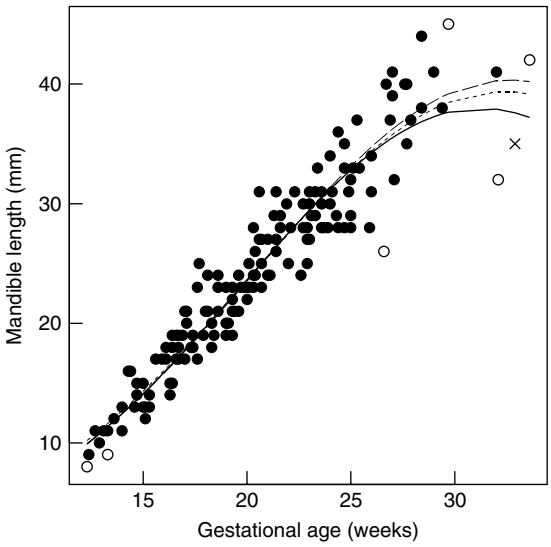

Figure 5 Mandible length data. Data with three fitted quadratic models with logged response: (a) ___ least squares fit to all data; (b) - - - - - - least squares fit when the seven marked cases are omitted; (c) . . . . . least trimmed squares. The very robust and diagnostic-led analyses give similar results. Case 166 is marked with a cross

why case 166, marked with a cross, is shown as influential in Figure 4 when cases 163, 165, and 167 have been deleted, but is not especially influential in the plot for all 167 cases in Figure 3.

Also shown is the fitted line obtained from the diagnostic procedure, which omitted seven cases. This plot is similar to that for the least trimmed squares, especially for the older cases, and distinct from the more curved least-squares fit to all the data. The plot confirms the doctors' suspicion that some readings at higher values of age might be atypical. However, after transforming the data to obtain a normal distribution of errors, it seems that some readings at low ages may also be atypical. Further, some of the readings at high ages are, despite the doctors' suspicions, informative about the general relationship between age and mandible length.

Extensions and Literature

Deletion diagnostics of the sort described here have received book length treatment by Belsley et al. [6], Cook and Weisberg [8], and by Atkinson [1]. They are also emphasized in the regression books of Weisberg [16] and of Ryan [15]. Added variable and constructed variable plots (*see* **Residuals**) indicate the contribution of individual cases to the evidence for inclusion of an explanatory variable, provided cases with high leverage are not important. If they are, deletion versions of the appropriate statistics are useful (Atkinson [2]). Interactive **graphics** allow exploration of multiple aspects of the effect of deletion and of the change of, for example, transformation parameters [9]. The diagnostic use of very robust regression is described by Rousseeuw and Leroy [13]. An example combining diagnostics and such regression is Atkinson [3]. The extension of diagnostics to **generalized linear** models is in Chapter 12 of McCullagh and Nelder [12]. A more detailed description is Atkinson [10]. Methods for determining the influence of several observations are described in the article on the **forward search** and, at greater length, in the book of Atkinson and Riani [4].

References

[1] Atkinson, A.C. (1985). *Plots, Transformations, and Regression*. Oxford University Press, Oxford.

[2] Atkinson, A.C. (1986). Diagnostic tests for transformations, *Technometrics* **28**, 29–37.

[3] Atkinson, A.C. (1986). Masking unmasked, *Biometrika* **73**, 533–41.

[4] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

[5] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Ed. Wiley, New York.

[6] Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics*. Wiley, New York.

[7] Cook, R.D. (1977). Detection of influential observations in linear regression, *Technometrics* **19**, 15–18.

[8] Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London.

[9] Cook, R.D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

[10] Davison, A.C. & Snell, E.J. (1991). Residuals and diagnostics, in *Statistical Theory and Modelling*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, pp. 83–106.

[11] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

[12] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[13] Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

[14] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[15] Ryan, T.P. (1997). *Modern Regression Methods*. Wiley, New York.

[16] Weisberg, S. (1985). *Applied Linear Regression*, 2nd Ed. Wiley, New York.

(*See also* **Model Checking**; **Model, Choice of**)

A.C. ATKINSON

Differential Error

Suppose a response variable Y has a conditional distribution $F(y|x)$ given a true exposure measurement $X = x$. Suppose that an error process yields Z instead of X. Then the error process is differential if $F(y|x, z) \neq F(y|x)$; namely, if it is not **nondifferential error**. Naive use of Z in place of X in the model $F(y|x)$ leads to estimates of **exposure effect** that can be **biased** in any direction (*see* **Bias in Observational Studies**; **Bias, Overview**; **Measurement Error in Epidemiologic Studies**; **Misclassification Error**; **Validity and Generalizability in Epidemiologic Studies**).

The term differential error can also be applied to errors in the outcome measure, Y. Suppose that one measures the error-prone version W of Y, instead of Y itself. Then the error process is differential if $F(w|x, y) \neq F(w|y)$. Such differential error can also result in bias in any direction if W is simply substituted for Y in the model $F(y|x)$.

MITCHELL H. GAIL

Diggle–Kenward Model for Dropouts

A typical longitudinal study design (*see* **Longitudinal Data Analysis, Overview**) specifies a time sequence of measurements on each of a number of subjects. A common feature of the resulting data is that some of the intended sequences of measurements are incomplete because the corresponding subjects withdraw prematurely from the study. We refer to this as attrition or *drop-out*.

Drop-outs can arise for many reasons. For example; the study protocol may specify the removal of subjects who show adverse side-effects of the treatment being administered; subjects may die for reasons related or unrelated to the context of the trial; subjects may be unwilling to continue because they perceive no benefit from further participation. Often, in practice, the exact causes of drop-outs are unknown. This raises a potential difficulty for the analysis of the resulting data, since the processes which govern the drop-out may be related to the measurement process which is the primary focus of the study. Figure 1, reproduced from Diggle et al. [3, Chapter 11] illustrates the problem. It shows a simulated data set in which, for each of 100 subjects, the mean response is constant over the intended duration of the study but the probability that a subject drops out at any time is inversely related to the value of their last recorded measurement. Also,

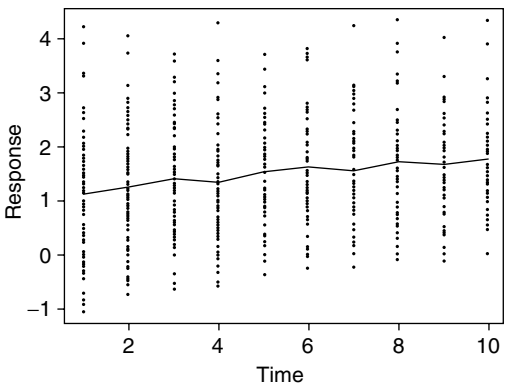

Figure 1 Simulation of a longitudinal data set in which the probability of drop-out is inversely related to the value of the last recorded measurement. Reproduced from [3] with permission from Oxford University Press

the sequence of measurements on any one subject is highly correlated. The result is that the lower-responding subjects are progressively removed from the study and the observed mean response amongst the nondrop-outs rises steadily over time. Suppose that each measurement represents a clinical response to treatment and that an increased response is beneficial. A naive analysis of the data might conclude that the treatment under investigation produces the desired rise in the mean response, whereas in reality no subject receives any clinical benefit from the treatment.

It is convenient to assume that the study protocol specifies a common set of measurement times for all subjects. Let $Y^* = (Y_1^*, \ldots, Y_n^*)$ then denote the intended sequence of measurements on a single subject, and D the subject's drop-out time, with the convention that an observed value $D = d$ means that the values of $Y_d^*, \ldots, Y_n^*$ are missing, whilst $D = n + 1$ signifies no drop-out. A statistical model for longitudinal data with drop-outs can then be thought of as a specification of the joint distribution of Y^* and D. Quite generally, we can write this joint distribution in two equivalent ways:

$$\begin{aligned} f^*(y, d) &= f^*(y)g(d|y) \\ &= g(d)f^*(y|d). \end{aligned} \tag{1}$$

Models derived from the first factorization are called *selection models* [4]. Those derived from the second factorization are called *pattern mixture models* [5]. One advantage of a selection model is that it includes, in the $f^*(y)$ term, a model for the study as designed. A counterbalancing advantage of a pattern mixture model is that it corresponds more directly to what is actually observed, namely the distributions of measurements within the subgroups defined by the different drop-out times.

Diggle & Kenward [2] introduce an explicit class of selection models for longitudinal data. They specify a multivariate Gaussian distribution for $f^*(y)$ and a logistic regression for $g(d|y)$. Explicitly, if $p_t(y)$ denotes the conditional probability of drop-out at time t, given $Y^* = y$, then

$$\text{logit}[p_t(y)] = \alpha_0 y_t + \sum_{k=1}^{r} \alpha_k y_{t-k}.$$

Molenberghs et al. [6] describe a model of the same kind, but for an ordered categorical response vector

Y^*, replacing the multivariate Gaussian specification of $f^*(y)$ by a model due to Dale [1].

From an inferential point of view, it is important to distinguish three subclasses of the Diggle–Kenward model, which correspond to the three kinds of missing data mechanism defined in a more general setting by Rubin [8]:

1. *completely random drop-outs* (CRDs): $p_t(y) = p_t$. The probability of drop-out at time t is independent of the measurement process Y^*
2. *random drop-outs* (RDs): $p_t(y) = p_t(y_1, \ldots, y_{t-1})$. The probability of drop-out at time t may depend on any or all of the observed measurement history $y_1, \ldots, y_{t-1}$, but cannot depend on the unobserved y_t
3. *informative drop-outs* (IDs): $p_t(y) = p_t(y_1, \ldots, y_t)$. The probability of drop-out at time t depends on the unobserved y_t.

For likelihood-based inference, it turns out that the important distinction is between IDs on the one hand and CRDs or RDs on the other. If we assume that the measurement model $f^*(y)$ and the drop-out model $g(d|y)$ are parameterized separately by θ and α, respectively, then under either CRDs or RDs, the log likelihood for θ and α separates into two terms, one involving θ only and, the other involving α only, whereas in the general ID case, the log likelihood is a sum of three terms, the third of which involves both θ and α. It follows that under CRD or RD assumptions, valid likelihood-based inferences about θ can be made without explicitly modeling the drop-out process. For this reason, CRD or RD mechanisms are sometimes collectively known as *ignorable* drop-outs. However, the example of Figure 1, which uses an RD model, makes the point that the practical interpretation of an analysis conducted on this basis still needs some care.

As pointed out in the discussion of Diggle & Kenward [2], the inferences made in the ID case are very sensitive to the underlying distributional assumptions. This is of particular concern because Molenberghs et al. [7] show that the RD hypothesis is not testable without additional assumption.

References

[1] Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses, *Biometrics* **42**, 909–917.

[2] Diggle, P.J. & Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion), *Applied Statistics* **43**, 49–93.

[3] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[4] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models, *Annals of Economic and Social Measurement* **5**, 475–492.

[5] Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association* **88**, 125–134.

[6] Molenberghs, G., Kenward, M.G. & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out, *Biometrika* **84**, 33–44.

[7] Molenberghs, G., Michiels, B. Kenward, M.G. & Diggle, P.J. (1998). Monotone missing data and pattern-mixture models, *Statistica Neerlandica* **52**, 153–161.

[8] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

(*See also* **Nonignorable Dropout in Longitudinal Studies**)

PETER J. DIGGLE

Dilution Method for Bacterial Density Estimation

There are many methods for measuring bacterial mass and concentration. For example, estimates of protein content, or of optical density, may be used as an index of bacterial mass, but such a measure would ignore heterogeneity of cell size and type and viability of cells. There are various methods for counting numbers of cells [4]. Cell numbers in a given suspension may be estimated by comparison with standard suspensions, but this method, depending on the parameter measured, may not distinguish between viable and nonviable cells. The number of viable bacteria in a suspension may be estimated in several ways. The most precise of these is a direct count of the number of viable bacteria, if these can be identified and counted; an indirect method for obtaining such a count is by determining the number of colony-forming units by counting the colonies produced when the suspension is cultured on a plate. When the number of (viable) bacteria (or other microorganisms) in a given volume of material cannot be directly counted, the dilution method may be used.

In this procedure a suspension of material is diluted to the point at which some samples contain such a small amount of the original suspension that they will contain no viable bacteria, whence dilution to extinction. A known volume of each dilution is inoculated into one or more experimental units, which in the classical method are tubes filled with culture medium. After incubation, the number of experimental units showing the presence or absence of the microorganism of interest is noted; if tubes filled with culture medium are used, the numbers of turbid and clear tubes at each dilution are noted. Experimental units showing a negative response are assumed not to have received even one organism.

If every viable microorganism gives a positive response, and if the organisms are distributed at random throughout the material being assessed, then the number of organisms in any volume of the suspension follows a **Poisson distribution**. These two assumptions, under which the frequency of negative responses, or clear tubes, p, is related to the mean number of microorganisms, m, by the relation $p = \mathrm{e}^{-m}$, provide the basis for estimation of the number of microorganisms. Fisher [2] noted that the ideal proportion of negative responses is just over one-fifth, and further noted that, even under the most favorable circumstances, 155 samples would be needed to reduce the standard error below 10%. Thus this method is rarely used if precise estimates are required.

The dilution method is frequently used when little is known about the likely concentration of bacteria and hence a large dilution factor, commonly tenfold, may be used. Tables of estimates of the "most probable number" are available (see, for example, [3, 4]) and, particularly where computational facilities are limited, experiments may be designed to conform to an arrangement for which the results have been tabulated. Tables of "acceptable results" have also been produced for some designs, based on the probabilities of various results; for example, under the assumptions above, it is virtually impossible for there to be no positives at the highest, and no negatives at the lowest, of three concentrations in an experiment using a tenfold dilution series [5]. For further details of analysis and design *see* **Serial Dilution Assay**.

Any method for the determination of bacterial mass or concentration should only be used after careful consideration of the relation of the measured quantity to the cells of interest together with all factors that affect that relation, since the validity of the determination depends on the validity of the assumed relation. Applications of modern technology have emphasized the heterogeneity of microbial populations, and have thus concentrated on counting and characterizing individual bacterial cells [1].

References

[1] Davey, H.M. & Kell, D.B. (1996). Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses, *Microbiological Reviews* **60**, 641–696.

[2] Fisher, R.A. (1951). *The Design of Experiments*, 6th Ed. Oliver & Boyd, London.

[3] Garthwright W.E. (1998). Appendix 2. Most probable number from serial dilutions. In Bacteriological Analytical Manual Online, U.S. Food and Drug Administration, http://vm.cfsan.fda.gov/-ebam/bam-a2.html

[4] Meynell, G.G. & Meynell, E. (1970). *Theory and Practice in Experimental Bacteriology*, 2nd Ed. Cambridge University Press, Cambridge.
[5] Taylor, J. (1962). The estimation of numbers of bacteria by tenfold dilution series, *Journal of Applied Bacteriology* **25**, 54–61.

(*See also* **Infectivity Titration**)

ROSE E. GAINES DAS

Direct and Indirect Effects

Of considerable interest in **epidemiology** and clinical medicine (*see* **Clinical Epidemiology**) is identifying the causal mechanism by which a risk exposure or treatment has its effect (*see* **Causal Direction, Determination**; **Causation**). One means of doing this is to consider the role of a putative intermediate variable M, shown in Figure 1, as the potentially mediating part of the **association** between the primary causal exposure E and outcome D.

If Pr(D|M,E) = Pr(D|E), then E is considered as having no direct effect on D; its effects are mediated or are indirect through M. The distinction between M being an intermediate variable or a **confounding** variable often depends upon theory or study design issues. For the association of E and D to be interpreted as one of indirect causation, there is an assumption that the association between E and M is one where E causes M.

In psychiatry, biometrical **genetics**, and **community medicine** the estimation of direct and indirect effects has been largely synonymous with the use of **path analysis** and path diagrams. Path diagrams display the directional (single-headed arrows) and **correlational** associations (lines with arrowheads at both ends) among a set of variables. Path analysis, first systematically developed by Sewall Wright [11], exploits linearity assumptions to allow the covariance between two variables on a path diagram to be decomposed into contributions arising from each legitimate path that connects them, with simple rules for determining the legitimate paths [4, 12] in the case of nonrecursive diagrams. If, to the traditional diagram, a residual variance is added to each variable in the form of a double-headed arrow running both from and to that variable, then the path tracing rules for legitimate paths can be reduced to the following: trace backward from a variable, change direction at a two-headed arrow, then trace forward. Multiplication of the (standardized) coefficients along each chain gives the expected (standardized) covariance for that path, and these may be summed to give the total, direct, and indirect expected covariances (or effects). Implicit in these diagrams is the fact that we are modeling these sets of variables as a set of simultaneous equations. These would now be commonly estimated in **software** for **structural equation modeling**.

Biostatistical applications often confront **categorical** outcomes and intermediate variables. Winship and Mare [10] elaborate path modeling in the case of **binary** variables where the simultaneous equations take the probit form (*see* **Quantal Response Models**) and where effects of each variable may arise from two sources: the effect of the observed binary value and the effect of the latent continuous (conditionally normal) variable that may underlie the observed binary variable (*see* **Latent Class Analysis**). This second effect arises where the binary variable reflects an underlying **propensity** that has been measured subject to error in categorical form. Both effects may be of interest, for example, psychiatrists might want to distinguish the effects of an actual episode of depression from the effects of a predisposition for depression, the latter reflecting genetics, among other things.

Where the linearity assumptions of path analysis are considered too strong or where the causal arguments are to be based in the paradigm of explicit counterfactuals, then the consideration and estimation of direct and indirect effects can be undertaken within the less parametric framework of directed acyclic graphs [3, 5, 6, 8]. In this framework, the direct effect of E on D controlling for M is a causal direct effect of E on D when the M value of everybody in the population is physically set to a predetermined value. Moreover, we may need to talk in terms of plural direct effects, one for each level of M. In this framework, G-estimation [7] (*see* **Structural Nested Failure Time Models**) or **marginal modeling** [9] may be the preferred method of estimation.

Whichever framework is used, the diagram or the equivalent algebraic representation of the full causal model plays a critical role in determining what is or is not included in the calculation of direct and indirect effects. The possible presence of other variables not shown in the diagram that could be influencing any or

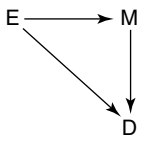

Figure 1 Direct and indirect effects of exposure E on disease D

all of E, M, and D should always be considered [1]. In addition, **measurement error** in the intermediate variable can attenuate the estimated indirect effects, resulting in artifactual residual direct effects. Many **longitudinal** studies of a repeat measure find, that even after controlling for a time 2 measure, a time 1 measure still predicts the time 3 measure but no longer does so once measurement error has been accounted for [2].

References

[1] Cole, S.R. & Hernan, M.A. (2002). Fallibility in estimating direct effects, *International Journal of Epidemiology* **31**, 163–165.

[2] Dunn, G., Everitt, B. & Pickles, A. (1993). *Modelling Covariances and Latent Variables using EQS*. Chapman & Hall London.

[3] Greenland, S., Pearl, J. & Robins, J.M. (1999). Causal diagrams for epidemiologic research, *Epidemiology* **10**, 37–48.

[4] Heise, D.R. (1975). *Causal Analysis*. Wiley-Interscience, New York.

[5] Maldonado, G. & Greenland, S. (2002). Estimating causal effects. *International Journal of Epidemiology* **31**, 422–429.

[6] Pearl, J. (1995). Causal diagrams for empirical research, *Biometrika* **82**, 669–710.

[7] Robins, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect, *Mathematical Modelling* **7**, 1393–1512.

[8] Robins, J.M. & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects, *Epidemiology* **3**, 143–155.

[9] Robins, J.M., Hernan, M.A. & Brumback, B. (2002). Marginal structural models and causal inference in epidemiology, *Epidemiology* **11**, 550–560.

[10] Winship, C. & Mare, R.D. (1983). Structural equations and path analysis for discrete data. *American Journal of Sociology* **89**, 54–110.

[11] Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.

[12] Wright, S. (1934). On the method of path coefficients. *Annals of Mathematical Statistics* **5**, 161–215.

ANDREW PICKLES

Discrete Survival-time Models

Most methods for analyzing survival-time data (*failure time* or *event history* data) are based on time as a continuously measured variate. A basic assumption for large parts of theory is that failure times are untied; see Andersen et al. [2]. In practice, there is always some smallest time unit, so that ties can occur (*see* **Tied Survival Times**). A moderate number of ties, while banned in theory, can be treated by appropriate modifications. If many ties occur, e.g. due to grouping in larger time units or intervals, or if time is truly discrete, then discrete survival or failure time models are more consistent with the data. Such situations arise in medical work when patients are followed up at fixed intervals like months, in certain biostatistical problems, for example human fertility studies and time to pregnancy [19], or in labor market studies where duration of unemployment is measured in weeks, at best, or in months. We review parametric models and outline recent nonparametric approaches. More details, in particular for parametric models, are given for example in Fahrmeir & Tutz [11, Chapter 9] and further references cited there.

Basic Concepts

Let time be divided into intervals $[a_0, a_1), [a_1, a_2), \ldots, [a_{q-1}, a_q), [a_q, \infty)$. Usually $a_0 = 0$ is assumed, and a_q denotes the final follow up. Identifying the discrete time index t with interval $[a_{t-1}, a_t)$, a discrete failure time T is considered, where $T = t$ denotes failure within the interval $t = [a_{t-1}, a_t)$. The basic quantity characterizing T is the *discrete hazard function*

$$\alpha(t) = \Pr(T = t | T \geq t), t = 1, \ldots, q, \qquad (1)$$

which is the conditional probability for the risk of failure in interval t given the interval is reached. The *discrete survivor function* for reaching interval t is

$$S(t) = \Pr(T \geq t) = \prod_{s=1}^{t-1} [1 - \alpha(s)], \qquad (2)$$

and the unconditional probability for failure at t is $\Pr(T = t) = \alpha(t)S(t)$.

For a homogeneous population, discrete failure time data are given by $(t_i, \delta_i), i = 1, \ldots, n$, where $t_i = \min(T_i, C_i)$ is the minimum of the survival time T_i and censoring time C_i, and δ_i is the indicator variable (*see* **Dummy Variables**) for failure ($\delta_i = 1$) or censoring ($\delta_i = 0$). In what follows we assume that censoring occurs at the end of the intervals, otherwise appropriate modifications have to be made. Simple estimates for $\alpha(t)$ are *crude death rates* $\hat{\alpha}(t) = d_t/n_t$, where n_t is the size of the population at risk and d_t the number of observed failures in $[a_{t-1}, a_t)$ (*see* **Vital Statistics, Overview**). The so-called *standard* **life table** *estimate* replaces n_t by $n_t - w_t/2$, where w_t is the number of censored observations in $[a_{t-1}, a_t)$, thereby assuming that censored observations are under risk for half the interval. In particular, for large t, where the size n_t of the risk set often becomes small, these estimates may be quite unsteady, and smoothing by one of the nonparametric methods outlined further below will be appropriate. This is illustrated in Figure 1, which shows crude

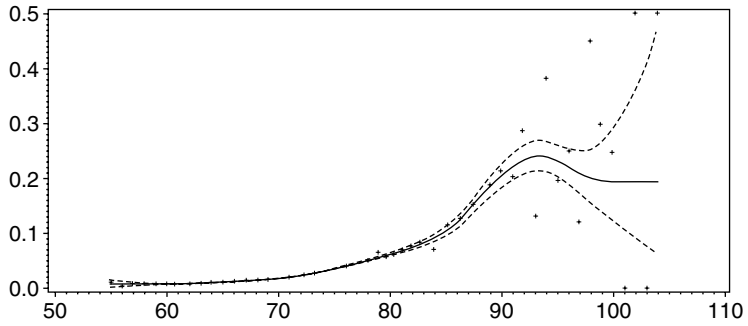

Figure 1 Posterior mean estimates (_____) and pointwise two standard deviation confidence bands (- - - - -) together with crude death rates (+)

death rates at age t in years for a population of retired American white females together with a smoothed estimate. A look at the data (Green & Silverman [13, p. 101]) shows that n_t becomes rather small for higher age t, resulting in unstable estimates towards the end of the observation period.

Discrete failure time data can also be described by a *discrete-time* **counting processes** $N_i(t), i = 1, \ldots, n$, defined by $N_i(0) = 0$ and

$$\Delta N_i(t) = N_i(t) - N_i(t-1) = \begin{cases} 1, & \text{if individual } i \text{ is at risk and fails at } t, \\ 0, & \text{else,} \end{cases}$$

for $t \geq 1$; see, for example, Arjas & Haara [3]. Thus, for every individual i under risk at t, the value $\Delta N_i(t)$ can be considered as the outcome of a **binary** experiment, with $\Pr(\Delta N_i(t) = 1) = \alpha(t)$. The sum $N(t) = \Sigma_i N_i(t)$ counts the number of observed failures up to t, and crude death rates can be derived as **nonparametric maximum likelihood** estimators, in analogy to the **Nelson–Aalen estimator** for continuous time.

In most studies a vector of possibly time-dependent basic or derived covariates $\mathbf{x}_{it}$ is observed in addition to failure times. The **time-dependent** components of $\mathbf{x}_{it}$ are assumed to be fixed within the interval t. Then the hazard function for survival time T_i of individual i will generally depend on covariates and is defined by

$$\alpha_i(t|\mathbf{x}_{it}^*) = \Pr(T_i = t | T_i \geq t, \mathbf{x}_{it}^*), \quad t = 1, \ldots, q, \tag{3}$$

where $\mathbf{x}_{it}^* = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it})$ denotes the history of covariates up to time t. Expressions for the survivor function (2) and for $\Pr(T = t)$ have to be modified accordingly. Also, the sequence of binary experiments for $\Delta N_i(t), t \geq 1$, will depend on $\mathbf{x}_{it}$. Unless a separate analysis for homogeneous subgroups can be carried out, it is natural to describe the dependence of **conditional probabilities** of failure by binary regression models. Let F_{t-} denote the history of events registered up to time t, but excluding the failure at t, and let $r_i(t)$ denote a risk indicator, with $r_i(t) = 1$ if individual i is at risk in interval t, and $r_i(t) = 0$ otherwise. Then it will be assumed that the conditional probability of failures can be expressed as

$$\Pr(\Delta N_i(t) = 1 | F_{t-}) = r_i(t)\alpha_i(t|\mathbf{x}_{it}^*),$$

with hazard functions linked to a time-varying predictor η_{it} by

$$\alpha_i(t|\mathbf{x}_{it}^*) = h(\eta_{it}) \tag{4}$$

through a suitable link function h, for example the logistic function (*see* **Logistic Regression**). The predictor η_{it} is modeled parametrically or nonparametrically as a function of time t and basic or derived covariates $\mathbf{x}_{it}$.

Parametric Models

This section deals mainly with parametric models (4), where the predictor has the common linear parametric form

$$\eta_{it} = z_{it}'\boldsymbol{\beta} \tag{5}$$

as in **generalized linear models**, with the design vector $\mathbf{z}_{it}$ formed from basic covariates. In many applications the linear predictor is chosen as

$$\eta_{it} = \beta_{0t} + \mathbf{x}_{it}'\boldsymbol{\beta}_x, \tag{6}$$

where $\boldsymbol{\beta}_x$ is a vector of covariate effects and $\beta_{0t}, t = 1, \ldots, q$, is a time-varying baseline effect. Model (6) can be written in the form (5) by defining $z_{it}' = (0, \ldots, 1, \ldots, 0, x_{it}')$ and $\boldsymbol{\beta}' = (\beta_{01}, \ldots, \beta_{0q}, \boldsymbol{\beta}_x')$. Other predictors are discussed further below.

Different discrete-time failure models are determined by choice of the link function h. Most common are discrete **proportional hazards** and logistic models.

The Discrete Proportional Hazards Model

Suppose that an underlying continuous failure time obeys a proportional hazard or **relative risk** model $\alpha_0(t)\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_x)$. If time T can only be observed as a discrete random variable, $T = t$ denoting failure in $[a_{t-1}, a_t)$, this yields the *discrete* proportional hazards model

$$\alpha(t|\mathbf{x}_{it}) = 1 - \exp[-\exp(\beta_{0t} + x_{it}'\boldsymbol{\beta}_x)], \tag{7}$$

with baseline effects

$$\beta_{0t} = \log \int_{a_{t-1}}^{a_t} \alpha_0(t)\,\mathrm{d}t$$

derived from the baseline function $\alpha_0(u)$ (see, for example, Kalbfleisch & Prentice [17]). An alternate

formulation of (7) is the complementary log–log model $\log\{-\log[1-\alpha(t|\mathbf{x}'_{it})]\} = \beta_{0t} + \mathbf{x}'_{it}\boldsymbol{\beta}_x$ (*see* **Quantal Response Models**). The parameter vector $\boldsymbol{\beta}_x$ is unchanged by the transition to the discrete version, so that the same analysis as with the proportional hazard model is possible as far as the influence of covariates is concerned. However, $\boldsymbol{\beta}_x$ and time-varying effects β_{0t} now have to be estimated jointly. If the number of intervals is large, then the dimension of $\beta_{01}, \ldots, \beta_{0q}$ may become dangerously high often even leading to the nonexistence of **maximum likelihood** estimates. Then more **parsimonious** parametric forms like polynomials $\beta_{0t} = \beta_0 + \cdots + \beta_k t^k$, piecewise constant effects, or regression **splines** with only a few cut points are preferable. Often, cubic-linear splines of the form $\beta_{0t} = \beta_0 + \beta_1 t + \beta_2(t-t_c)^2_- + \beta_3(t-t_c)^3_-$, are useful, where $(t-t_c)_- = \min(t-t_c, 0)$ and t_c is a cut-point. The baseline effect is cubic before t_c and linear after t_c. Such a simple spline model is more robust against few data at the end of the observation period than polynomials, and it is a smooth function as compared with piecewise constant modeling. Of course, other forms of regression splines may be considered. Also one may use the numerically more stable B-spline basis instead of the truncated power form, cf. Sleeper & Harrington [20] in a continuous-time setting. By appropriate definition of the design vector, regression spline models can also be written in linear parametric from (5).

The Logistic Model

An alternate model is the logistic model for the discrete hazard,

$$\alpha(t|\mathbf{x}_{it}) = \frac{\exp(\beta_{0t} + \mathbf{x}'_{it}\boldsymbol{\beta}_x)}{1 + \exp(\beta_{0t} + \mathbf{x}'_{it}\boldsymbol{\beta}_x)}, \tag{8}$$

considered by Thompson [21] and, in slightly different form, by Cox [5]. For short intervals this model becomes rather similar to the discrete proportional hazards model. An advantage of the logistic model is that the covariate effects $\boldsymbol{\beta}_x$ can be estimated semiparametrically, considering baseline effects β_{0t} as **nuisance parameters** and leaving them unspecified as in the continuous-time proportional hazards model; see Cox & Oakes [6].

Other discrete-time failure models result for other choices of h. Very flexible models are obtained if the link is an element of a parametric family of link functions. Examples are the model of Aranda–Ordaz (see Fahrmeir & Tutz [11, p. 318]) or the families considered by Czado [7].

Although choice of the link function is an important issue, we feel that careful modeling of the predictor is often even more essential. To simplify the discussion, we consider only two covariates, x and w, where x is a continuous variable like tumor size or hormone concentration and w is binary, indicating, for example, sex or treatment group.

Models with *time-varying effects* are obtained by assuming

$$\eta_{it} = \beta_{0t} + \beta_1 x_i + \beta_{2t} w_i, \tag{9}$$

where β_{2t} could be the time-varying effect of a therapy, possibly decreasing with time. Alternately, the term $\beta_{2t} w_i$ may be considered as a particular form of interaction between time t and the covariate w. The function β_{2t} may be modeled parametrically in the some way as the baseline effect β_{0t}. A more detailed discussion of parametric time-varying effects is in Yamaguchi [24]. If the simple linear form $\beta_1 x_i$ for the influence of x is too restrictive, then one may also try to replace it by a nonlinear smooth function $\beta_1(x)$ as in generalized additive models. As in Hastie & Tibshirani [15], one may go a step further and consider *varying coefficient models* of the form

$$\eta_{it} = \beta_{0t} + \beta_1(x_i) + \beta_2(x_i) w_i + \beta_{3t} w_i. \tag{10}$$

Here the smooth function β_2 may be viewed as an effect of w varying over x, or it is interpreted as an interaction term between the continuous covariate x and the binary covariate w. Without further prior knowledge it will often be difficult to specify certain parametric forms for the smooth functions in (9) and (10). Instead, it will be reasonable to explore patterns with nonparametric approaches outlined in the next section and to proceed then with a simpler parametric likelihood-based inference.

Likelihood Inference

Under appropriate conditions on censoring and covariate processes, the **likelihood** reduces to the common form known for binary regression models. Introducing the indicators

$$y_i = (y_{i1}, \ldots, y_{it}) = \begin{cases} (0, \ldots, 0), & \delta_i = 0, \\ (0, \ldots, 0, 1), & \delta_i = 1, \end{cases}$$

the log likelihood is proportional to

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{s=1}^{t_i} \{y_{is} \log \alpha_i(s|\mathbf{x}_{is}^*) + (1 - y_{is}) \times \log[1 - \alpha_i(s|\mathbf{x}_{is}^*)]\}.$$

Arjas & Haara [3] give a careful discussion of assumptions leading to $l(\boldsymbol{\beta})$ as a (partial) log likelihood (*see* **Partial Likelihood**). They will generally hold for noninformative random censoring and time-independent or external covariates, but can become critical for internal covariates. In particular, the likelihood is valid in the presence of ties, by making the weak assumption that failures at t are conditionally independent given covariates and past failures. By appropriate construction of the design vectors $\mathbf{z}_{it}$, the parameters β can then be estimated with **software** for binary regression models, and other tools of likelihood inference for these models may be adopted, see Fahrmeir & Tutz [11 Chapter, 9].

Nonparametric Approaches

Often, the common assumptions of linearity, additivity, and time-constancy of effects are definitely violated and the parametric specifications of more flexible models like (9) or (10) may be difficult. In this situation nonparametric approaches provide useful tools for detecting and exploring nonlinear or time-dependent effects. We outline the roughness penalty approach, leading to spline-type smoothing and related **Bayesian** nonparametric techniques. Other methods are based on discrete kernels (e.g. Fahrmeir & Tutz [11, Chapters 5 and 9] (*see* **Density Estimation**), or local likelihoods (Wu & Tuma [23], and in a continuous-time setting, Tutz [22]). Consider models like (9) or (10) with an unknown parameter vector $\boldsymbol{\beta}$ and unknown "smooth functions" $\beta_1, \beta_2, \ldots, \beta_q$ of time or continuous covariates. The roughness penalty approach maximizes a penalized log likelihood criterion

$$\text{pl}(\boldsymbol{\beta}, \beta_1, \ldots, \beta_p) = l(\boldsymbol{\beta}, \beta_1, \ldots, \beta_p) - \sum_{j=1}^{p} \lambda_j J(\beta_j),$$

where $J(\beta_j)$ are roughness penalties and λ_j are smoothing parameters. A simple roughness penalty for a time-varying effect $\beta_{jt}, t = 1, \ldots, q$, is

$$J(\beta_j) = \sum_{s=2}^{q} \frac{(\beta_{jt} - \beta_{j,t-1})^2}{a_t - a_{t-1}}. \tag{11}$$

The same form may be used for a function $\beta_j(x)$ of some continuous covariate x. Another common penalty is

$$J(\beta_j) = \int [\beta_j(x)'']^2 \, \mathrm{d}x$$

leading to cubic smoothing splines (see, for example, Green & Silverman [13]). Kiefer [18] proposes a discrete proportional hazards model with time-varying effects β_{jt} and penalty function (11). Dannegger et al. [8] use the roughness penalty approach to explore the nonlinear and time-varying effects of risk factors in a breast cancer study with monthly data. Related Bayesian nonparametric approaches put smoothness **priors** on β_{jt} or $\beta_j(x)$ and estimation is based on posteriors, given the data. If, for example, a random walk (*see* **Stochastic Processes**) of first order

$$\beta_{jt} = \beta_{j,t-1} + (a_t - a_{t-1})^{1/2} v_t, \quad v_t \sim \text{N}(0, 1/\lambda_j),$$

is taken as the smoothness prior for the sequence $\{\beta_{jt}\}$, then the posterior mode or MAP estimate is identical to the penalized likelihood estimate with penalty (11); see Fahrmeir [9]. Full posterior analyses can be carried out with **Markov chain Monte Carlo** (MCMC) techniques (see Fahrmeir & Knorr-Held [10]) and in the related context of generalized **additive models** (Biller & Fahrmeir [4]). Nonparametric methods are also useful for smoothing hazard functions in the absence of covariates. The smooth curve in Figure 1 is the posterior mean estimate obtained from a Bayesian MCMC approach. The corresponding cubic spline smoother is very close, see Green & Silverman [13].

Some Extensions

More Complex Discrete-Time Event History Data

Discrete failure time models can be extended in the same way as continuous-time models. Often one may distinguish between several types $R \in \{1, \ldots, m\}$ of failure or terminating events. For example, in a medical study there may be several causes of death, or in studies on unemployment duration one may consider

full-time and part-time jobs that end the unemployment duration. The basic quantities for *models with multiple modes of failure* are now cause-specific hazard functions,

$$\alpha_{ir}(t|\mathbf{x}_{it}^*) = \Pr(T = t, R = r|T \geq t, \mathbf{x}_{it}^*), \qquad (12)$$

i.e. conditional probabilities for failure of type r in interval t (*see* **Competing Risks**). Polytomous response models can be used for regression analysis of cause-specific hazard functions. A common candidate for unordered events is the **multinomial** logit model (e.g. Allison [1]) (*see* **Polytomous Data**). Other discrete choice models like a probit or a nested multinomial logit model [16] may also be considered. If events are ordered, then ordinal response models (e.g. [11]) are appropriate. Again parametric and nonparametric approaches are possible. Penalized likelihood and Bayesian smoothing techniques with models for time-varying effects are applied to unemployment durations in Fahrmeir & Wagenpfeil [12] and Fahrmeir & Knorr-Held [10].

Discrete failure time models can also be extended to general multiepisode–multistate models or, in counting process terminology, marked **point processes**. Hamerle [14] studies parametric regression analysis for such discrete event history data, but generally much less theoretical or applied work has been done here.

Unobserved Heterogeneity and Frailty Models

The above model specifications assume that individual heterogeneity can be described by observed variables. However, it is likely that not all relevant variables are included in a regression model. The conventional approach to account for neglected heterogeneity or **frailty** is to include individual-specific parameters into the predictor, i.e. modifying η_{it} to $\eta_{it}^h = \eta_{it} + \theta_i$, and to assume that the individual-specific parameters are independent, identically distributed (iid) random variables from a prior density function f, e.g. a normal density. Estimation can then be based on approaches for generalized mixed models with **random effects**, and recent MCMC methods seem particularly well suited. However, for single-spell failure time models the estimates can be very dependent on the choice of the prior specification. More experience is needed here. The problem becomes less severe with **repeated events**.

References

[1] Allison, P.D. (1982). Discrete-time methods for the analysis of event histories, in *Sociological Methodology*, Vol. 13, S. Leinhardt, ed. Jossey-Bass, San Francisco, pp. 61–98.

[2] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[3] Arjas, E. & Haara, P. (1987). A logistic regression model for hazard: asymptotic results, *Scandinavian Journal of Statistics* **14**, 1–18.

[4] Biller, C. & Fahrmeir, L. (1997). Bayesian spline-type smoothing in generalized regression models, *Computational Statistics* **12**, 135–151.

[5] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[6] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[7] Czado, C. (1992). *On Link Selection in Generalized Linear Models, Springer Lecture Notes in Statistics*, Vol. 78, Springer-Verlag, New York, pp. 60–65.

[8] Dannegger, F., Klinger, A., and Ulm, K. (1995). *Identification of Prognostic Factors with Censored Data*. Discussion Paper 11, Sonder forschungsbereich 386, Ludwig-Maximilians Universität, München.

[9] Fahrmeir, L. (1994). Dynamic modelling and penalized likelihood estimation for discrete time survival data, *Biometrika* **81**, 317–330.

[10] Fahrmeir, L. & Knorr-Held, L. (1997). Dynamic discrete-time duration models: estimation via Markov Chain Monte Carlo, in *Sociological Methodology*, Vol. 27, A. Raftery, ed. Blackwell, Oxford, pp. 417–452.

[11] Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

[12] Fahrmeir, L. & Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risk models, *Journal of the American Statistical Association* **91**, 1584–1594.

[13] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

[14] Hamerle, A. (1986). Regression analysis for discrete event history or failure time data, *Statistical Papers* **27**, 207–225.

[15] Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models, *Journal of the Royal Statistical Society, Series B* **55**, 757–796.

[16] Hill, D.H., Axinn, W.G. & Thornton, A. (1993). Competing hazards with shared unmeasured risk factors, in *Sociological Methodology*, Vol. 23, P.V. Marsden, ed. Blackwell, Oxford, pp. 245–277.

[17] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[18] Kiefer, N.M. (1990). Econometric methods for grouped duration data, in *Panel Data and Labor Market Studies*, J. Hartog, G. Ridder & J. Theeuwes, eds. Elsevier, Amsterdam, pp. 97–117.

[19] Scheike, T.H. & Jensen, T.K. (1997). A discrete survival model with random effects: an application to time to pregnancy, *Biometrics* **53**, 349–360.

[20] Sleeper, L.A. & Harrington, D.P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease, *Journal of the American Statistical Association* **85**, 941–949.

[21] Thompson Jr, W.A. (1977). On the treatment of grouped observations in life studies, *Biometrics* **33**, 463–470.

[22] Tutz, G. (1995). Dynamic modelling of discrete duration data: a local likelihood approach, *Report 95-15*. Institut für Quantitative Methoden, Tech. Univ. Berlin.

[23] Wu, L.L. & Tuma, N.B. (1991). Assessing bias and fit of global and local hazard models, *Sociological Methods and Research* **19**, 354–387.

[24] Yamaguchi, K. (1993). Modeling time-varying effects of covariates in event-history analysis using statistics from the saturated hazard rate model, in *Sociological Methodology*, Vol. 23, P.V. Marsden, ed. Blackwell, Oxford, pp. 279–317.

(*See also* **Survival Analysis, Overview**; **Survival Distributions and Their Characteristics**)

LUDWIG FAHRMEIR

Discriminant Analysis, Linear

Linear discriminant analysis is a statistical method for studying the differences between classes of objects. Broadly speaking, the method may be used with two, rather distinct, objectives in mind. The first is as a predictive tool (*see* **Prediction**). Here the aim is to formulate a rule which will permit objects to be classified into one of several predefined classes. The second is to help understanding. Here the aim is to build a model which helps us understand the structure in data (*see* **Model, Choice of**). We illustrate both of these uses below. Although different, both of these uses are based on the same underlying principles and both are essentially inferential: in the first case the inferences are from a sample of objects to new objects and in the second case from the sample to the underlying population of objects.

We begin with a sample of objects drawn from the population being studied. Usually this sample, often called a *design* or *development* sample, will be a simple random sample, though other sampling schemes can also be used. For each member of this sample we require that (i) we know its true class, and (ii) we know the values of a (fixed) set of variables describing that object. If our aim is prediction, then we will use this sample to construct a model which will allow us to predict the class membership of a new object from only its vector of measurements. If our aim is understanding, then we will use this sample to characterize the main ways in which the groups differ. Some examples will help to clarify these ideas.

1. Our aim is to construct a rule which will assist in diagnosing patients suffering from hepatitis as having either acute infectious hepatitis or hepatitis secondary to infectious mononucleosis. The information on which we must base our diagnosis are the values of two variables: the activity of lactate dehydrogenase isoenzyme-5 and the activity of lactate dehydrogenase isoenzyme-3 [20].
2. Is it possible to predict who is likely to suffer from osteoporosis in later life on the basis of the answers to the questions on a simple noninvasive screening questionnaire [5, 22]?
3. Can we identify which children will respond positively to a new behavioral treatment for enuresis, based on variables such as age, sex, whether or not the child was an only child, type of enuresis, and so on [19]?
4. Can we predict, without the need for surgery, for which patients prostate cancer will have spread to the surrounding lymph nodes [4]?
5. In what way do those infants at high risk of dying from Respiratory Distress Syndrome differ from those at low risk? Variables in the study included sex, responsiveness, gestational age, birthweight, etc. [36].

As far as the prediction problem is concerned, a legitimate question is: "Why bother?" If the true classes of the objects can be found – they were, after all, found for the objects in the design set – why not use the same method for new objects? The answer is indicated by some of the above examples. We might want a prognostic classification (*see* **Prognostic Factors for Survival**), so that we can undertake some treatment intervention, and we cannot wait to discover the true class.

Perhaps discovering the true class requires a postmortem examination. Perhaps it is very expensive to discover the true class so that a cheaper (if less accurate) method is needed. Maybe the procedure for determining the true class is very time-consuming – for example, growing a bacterial culture – and a quicker method of classifying an object is needed, and so on.

Since the objects in the design set have known class memberships, methods for tackling the above problems are sometimes described as methods of *supervised* classification (or supervised pattern recognition). This distinguishes them from *unsupervised* classification methods, where no class memberships are known a priori. The objective of the latter methods, also termed methods of *cluster analysis*, is to determine the class structures, not to model a preexisting such structure. We do not discuss the latter in this article. Details of such methods can be found in [1, 10, 17], and [41] (*see* **Classification, Overview**; **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**).

Many methods of supervised classification have been developed. They include linear discriminant analysis (which is the subject of this article), logistic discriminant analysis (*see* **Logistic Regression**), classification trees (*see* **Tree-structured Statistical Methods**), nearest neighbor methods, neural network

and other highly parameterized models, and innumerable variations on all of the above. Comparative descriptions and reviews of these methods may be found in [21, 37], and [23].

Linear discriminant analysis is the oldest of the methods, at least in terms of formal development. The ideas can be traced back at least as far as Fisher [11]. Although originally described for the case of only two classes, it has been extended to handle more than two classes. It is convenient for us, however, to introduce the ideas via the special case of only two classes, as we do in the next section.

The Two-Class Case

Suppose that we have a sample of objects from each of the two classes. For each of these objects we know its parent class and the values of a vector of measurements taken on that object. Suppose that our aim is the second of the two described in the opening paragraph – to gain an understanding of the nature of the differences between the classes, based on their measurements. To do this, we attempt to construct a measure of "classness", based on the measured variables, such that, for example, large values of this measure correspond to membership of one class and small values to membership of the other class. We can then tackle the first problem of the opening paragraph by imposing a threshold on this measure: those new objects whose "classness" value exceeds the threshold will be classified into one class and those whose value is less than the threshold will be classified into the other class.

Perhaps the most obvious initial approach is to look at each measured variable separately. However, this will not always be very helpful. Suppose, for simplicity (it permits us to draw diagrams), that only two variables, x_1 and x_2, are measured on each object, and that we have only four objects, two from each class, in the design set. Figure 1(a) shows these four objects plotted on a line showing their values for variable x_1 and Figure 1(b) shows them plotted against the x_2 variable. In each case, objects from one class are represented by circles and objects from the other class by crosses. No threshold on the x_1 variable permits good separation between the classes. Wherever we put the threshold, some design set objects lie on the "wrong" side of the line. The same applies to the x_2 variable. Now, however, consider Figure 2, which shows the design set points plotted in the two-dimensional space of the variables. If we project the points onto the direction of the dashed line, as shown in Figure 3, we see that we can easily produce a threshold which perfectly separates the two classes. The position on the dashed line represents "classness".

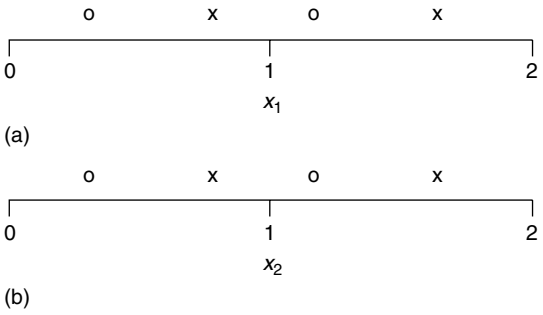

Figure 1 Four objects plotted according to their values on (a) the single variable x_1, and (b) the single variable x_2. In neither case can the two classes be well separated

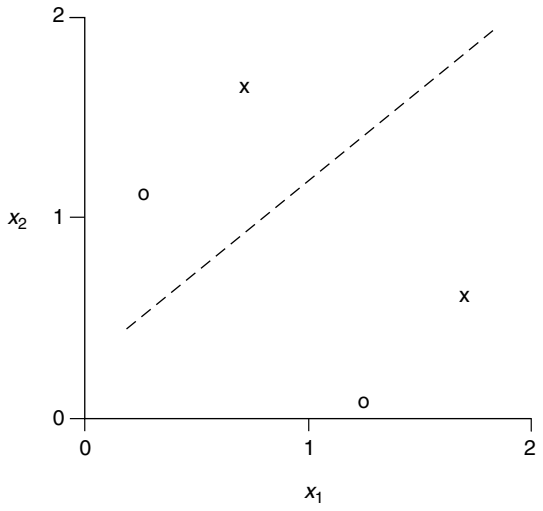

Figure 2 The four objects from Figure 1 plotted in the two-dimensional space of x_1 and x_2

Figure 3 The projection of the four objects onto the broken line in Figure 2 permits the two classes to be easily separated

This is the essence of linear discriminant analysis. We seek some direction in the space spanned by the measured variables such that the projections of the design set points onto that direction are well separated. This direction characterizes the difference between the classes, so that describing the direction permits us to tackle the second objective presented in the opening paragraph. Imposing a threshold on the direction permits us to tackle the first objective. In terms of the complete space, such a threshold corresponds to a surface (a line in the two-dimensional case). For the artificial example above, such a surface (line) is illustrated in Figure 4. This surface is called a *decision surface*.

Directions in a space defined by the measured variables correspond to linear combinations of the defining variables. So our objective is to find some linear combination which leads to good separation of the projections of the points in that direction. Denoting an arbitrary set of weights by $\mathbf{a} = (a_1, \ldots, a_p)$, the projection of a point $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, with p measurements x_{ij}, $j = 1, \ldots, p$, onto the direction defined by $\mathbf{a}$, is given by $y_i = \mathbf{a}'\mathbf{x}_i$. An obvious measure of the separation between the projections of the points in the two groups is then the separation between their means. However, for the same reason as with the two group t-test (*see* **Student's *t* Distribution**) Distribution, it makes more sense to take into account the within-group variability, and standardize using the (assumed common) within group standard deviation. That is, denoting the mean of the groups' projections by $\overline{y}^{(1)}$ and $\overline{y}^{(2)}$, and the standard deviation of the projections within each of the groups by $\mathrm{sd}(y)$, the

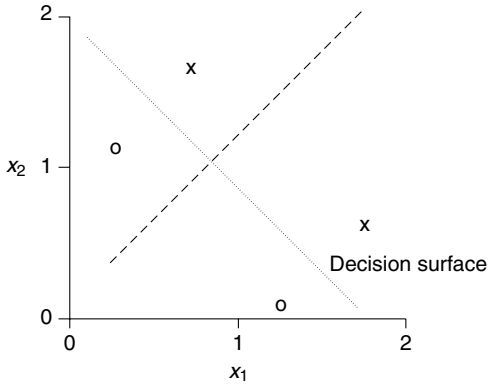

Figure 4 The dashed line shows the position of the decision surface separating the two classes

measure of separation is $(\overline{y}^{(1)} - \overline{y}^{(2)})/\mathrm{sd}(y)$. It will be more convenient to work with the square of this, which is $(\overline{y}^{(1)} - \overline{y}^{(2)})^2/\mathrm{var}(y)$. In terms of the original variables, this is $(\mathbf{a}'\overline{\boldsymbol{x}}^{(1)} - \mathbf{a}'\overline{\mathbf{x}}^{(2)})^2/\mathbf{a}'\mathbf{S}\mathbf{a}$, where $\overline{\mathbf{x}}^{(k)}$ is the mean vector for group k and $\mathbf{S}$ is the average of the sample covariance matrices of the two groups. We now need to find the vector $\mathbf{a}$ which maximizes this measure.

Differentiating the above measure with respect to $\mathbf{a}$ and equating to zero reveals that it is maximized by $\hat{\mathbf{a}} \propto \mathbf{S}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})$. The function $\hat{\mathbf{a}}'\mathbf{x}$ is then called a *linear discriminant function*. The components of the vector $\hat{\mathbf{a}}$ can thus be studied to understand how the two groups differ, what variables (when taken in combination with the others present) contribute most to the difference, and so on. To classify a new point we still have to decide on a threshold. The most natural way to choose a suitable threshold is revealed if we take a step back and attempt to tackle the classification problem directly.

Let the underlying distribution of points from class $k, k = 1, 2$, be $f(\mathbf{x}|k)$ and let the prior probability of belonging to class k be p_k, so that $p_1 + p_2 = 1$. Then a natural thing to do would be to classify a new point to class k if it seemed more likely to have come from class k. That is, we will estimate the probability that the new point came from classes 1 and 2 and assign it to that class with the highest estimated probability. Now, by **Bayes' theorem**, the probability of belonging to class k is $f(k|\mathbf{x}) = p_k f(\mathbf{x}|k)/f(\mathbf{x})$, where $f(\mathbf{x}) = p_1 f(\mathbf{x}|1) + p_2 f(\mathbf{x}|2)$ is the overall mixture distribution of the $\mathbf{x}$. Thus the solution is, replacing the unknown distributions by estimates, the class which maximizes $\hat{f}(k|\mathbf{x})$. With only two classes, this is equivalent to classifying into class 1 if the ratio $\hat{f}(1|\mathbf{x})/\hat{f}(2|\mathbf{x})$ exceeds 1 and to class 2 otherwise. Again using Bayes' theorem, this is equivalent to comparing

$$\frac{\hat{p}_1 \hat{f}(\mathbf{x}|1)/\hat{f}(\mathbf{x})}{\hat{p}_2 \hat{f}(\mathbf{x}|2)/\hat{f}(\mathbf{x})} = \frac{\hat{p}_1 \hat{f}(\mathbf{x}|1)}{\hat{p}_2 \hat{f}(\mathbf{x}|2)}$$

with 1.

So far we have not assumed any particular distributional form for the class conditional distributions $f(\mathbf{x}|k)$. Suppose we make the assumption, common in multivariate analysis, that they are **multivariate normal distributions**, and suppose we also assume that the **covariance matrices** of the two groups are equal. Then the above classification rule becomes:

compare

$$\frac{\hat{f}(1|\mathbf{x})}{\hat{f}(2|\mathbf{x})} = \frac{\dfrac{\hat{p}_1}{(2\pi)^{p/2}|\mathbf{S}|^{1/2}} \exp\left[-\dfrac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}^{(1)}\right)' \mathbf{S}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}^{(1)}\right)\right]}{\dfrac{\hat{p}_2}{(2\pi)^{p/2}|\mathbf{S}|^{1/2}} \exp\left[-\dfrac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}^{(2)}\right)' \mathbf{S}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}^{(2)}\right)\right]}$$

with 1. This can be considerably simplified if we take logarithms. It reduces to a comparison of

$$\mathbf{x}'\mathbf{S}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}) + \ln(\hat{p}_1/\hat{p}_2) - \tfrac{1}{2}\overline{\mathbf{x}}^{(1)'}\mathbf{S}^{-1}\overline{\boldsymbol{x}}^{(1)} + \tfrac{1}{2}\overline{\mathbf{x}}^{(2)'}\mathbf{S}^{-1}\overline{\boldsymbol{x}}^{(2)}$$

with 0. The first term in this expression is identical to the optimal $\hat{\mathbf{a}}'\mathbf{x}$ expression derived above. Moreover, the last three terms in this expression depend solely on the design set, and not on $\mathbf{x}$. Thus, approaching the problem from the perspective of classifying a point has led to the same solution as above: we must project the point to be classified onto the direction which we derived as the "best separating direction". The last three terms above are merely constants – they serve to define the threshold with which we compare $\mathbf{x}'\mathbf{S}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})$ to see if it is greater or less than 0. Put another way, the classification rule is: if $\mathbf{x}'\mathbf{S}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})$ is greater than $-\ln(\hat{p}_1/\hat{p}_2) + \frac{1}{2}\overline{\mathbf{x}}^{(1)'}\mathbf{S}^{-1}\overline{\mathbf{x}}^{(1)} - \frac{1}{2}\overline{\mathbf{x}}^{(2)'}\mathbf{S}^{-1}\overline{\mathbf{x}}^{(2)}$ assign the new object to class 1, otherwise to class 2. Ties can be settled arbitrarily.

The above derivation assumes that misclassifying a class 1 object as a class 2 object is equally as serious as the reverse. Although this is a common assumption (particularly in methodological studies) it is rarely realistic. More usually, one type of misclassification is more serious, or costly in some sense, than the other. Suppose, then, that the cost of misclassifying a class 1 object to class 2 is c_1 and the cost of misclassifying a class 2 object to class 1 is c_2. Our classification rule will be: if a point $\mathbf{x}$ falls in region Ω_1, classify it as belonging to class 1, and if it falls in region Ω_2 (which is the complement of Ω_1) classify it to class 2. Our aim is to choose these two regions such that the overall cost is minimized. This overall cost is

$$c_1 \int_{\Omega_2} f(1|\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} + c_2 \int_{\Omega_1} f(2|\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}.$$

To minimize this we must choose Ω_1 so that $\mathbf{x}$ is in Ω_1 whenever $c_2 f(2|\mathbf{x}) f(\mathbf{x}) < c_1 f(1|\mathbf{x}) f(\mathbf{x})$. That is, the overall cost will be minimized if we classify $\mathbf{x}$ as belonging to class 1 when $f(1|\mathbf{x})/f(2|\mathbf{x}) > c_2/c_1$ and to class 2 otherwise. Assuming multivariate normal distributions, and replacing them by their estimates, the optimal rule, in terms of minimal cost is then: if $\mathbf{x}'\mathbf{S}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})$ is greater than $\ln(c_2/c_1) - \ln(\hat{p}_1/\hat{p}_2) + \frac{1}{2}\overline{\mathbf{x}}^{(1)'}\mathbf{S}^{-1}\overline{\mathbf{x}}^{(1)} - \frac{1}{2}\overline{\mathbf{x}}^{(2)'}\mathbf{S}^{-1}\overline{\mathbf{x}}^{(2)}$ assign the new object to class 1, otherwise to class 2. Again ties can be settled arbitrarily. When $c_1 = c_2$ this reduces to the rule above, as it should.

Although the above was described in terms of multivariate normal distributions, exactly the same results apply to any ellipsoidal distribution – that is, any distribution defined in terms of the mean vector and covariance matrix.

For classification rule purposes, we can relax the restriction that the covariance matrices in the two groups are assumed equal. If we do this, the classification rule above becomes more complicated, including separate terms for the two covariance matrices, so that it takes a quadratic form. For purposes of understanding, introducing these extra complications is of limited value – it certainly makes interpretation more difficult. The reason for considering it when classification is the aim is that it might lead to more accurate results: the more flexible the model, the more accurately it can reflect the underlying structure of the populations. While this is true, it is not the only influence on accuracy. In particular, the parameters in the classification rule must be estimated from the design set, and the accuracy of the resulting rule depends on the accuracy with which they are estimated. When a linear decision surface is used, the decision surface involves only $p + 1$ parameters. However, when a quadratic surface is used, an extra $(p + 1)/2$ parameters need to be estimated. As a consequence, the resulting surface has a higher variance. To overcome this a larger design set is needed.

The linear discriminant function can also be derived by a **regression** argument. If we define a response variable (*see* **Dummy Variables**) taking one value for the members of group 1 and some other value for the members of group 2, then the regression function is proportional to $\mathbf{x}'\mathbf{S}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})$.

There is also a close link between linear discriminant analysis and logistic discriminant analysis. Whereas linear discriminant analysis can be viewed as starting with the class conditional distributions

$f(\mathbf{x}|j)$, and then using **Bayes' theorem** to derive estimates of the $f(j|\mathbf{x})$, logistic discriminant analysis goes straight for the latter. The basic model is

$$f(1|\mathbf{x}) = \exp\left(\alpha_0 + \sum_{k=1}^{p} \alpha_k x_k\right) \Bigg/ 1 + \exp\left(\alpha_0 + \sum_{k=1}^{p} \alpha_k x_k\right).$$

Using this model, the log **odds ratio** is

$$\ln[f(1|\mathbf{x})/f(2|\mathbf{x})] = \alpha_0 + \sum_{k=1}^{p} \alpha_k x_k,$$

that is, a linear function of the x_k. However, we have already seen that, if we assume multivariate normal distributions, then the linear discriminant analysis approach also yields a linear function for the log odds ratio. The difference between the two methods lies in the parameter estimation and the assumptions which are made: logistic discriminant analysis also yields a linear log odds ratio for other distributional forms.

More Than Two Classes

If we assume multivariate normal distributions (more generally, ellipsoidal distributions) then the above generalizes immediately to more than two classes. For a point $\mathbf{x}$ at which a classification is required, we want to find the class j which maximizes $\hat{f}(j|\mathbf{x})$, the estimated posterior probability of belonging to class j. That is,

$$\max_j \hat{f}(j|\mathbf{x}) = \max_j \frac{\hat{p}_j}{(2\pi)^{p/2}|\mathbf{S}_j|^{1/2}} \times \exp\left[-\frac{1}{2}\left(\mathbf{x} - \overline{\boldsymbol{x}}^{(j)}\right)' \mathbf{S}_j^{-1}\left(\mathbf{x} - \overline{\boldsymbol{x}}^{(j)}\right)\right].$$

Since log is a monotonic increasing function, we can, alternatively, seek the j that maximizes

$$\ln \hat{p}_j - \tfrac{1}{2}\ln|\mathbf{S}_j| - \tfrac{1}{2}\left(\mathbf{x} - \overline{\mathbf{x}}^{(j)}\right)' \mathbf{S}_j^{-1}\left(\mathbf{x} - \overline{\mathbf{x}}^{(j)}\right).$$

Also, if we again assume that the covariance matrices are equal, we need the j that maximizes

$$\ln \hat{p}_j - \tfrac{1}{2}\overline{\mathbf{x}}^{(j)'}\mathbf{S}^{-1}\overline{\mathbf{x}}^{(j)} + \mathbf{x}'\mathbf{S}^{-1}\overline{\mathbf{x}}^{(j)}.$$

Such functions, or equivalent variants of them, are often called *classification functions*.

Description requires identifying those dimensions of the spaced spanned by the measured variables which are most important in distinguishing between the classes. For the multiclass situation we can again generalize the approach of the preceding section. There we had two classes and we sought that direction in which the class means were most separated (standardized to take into account the within class variation in that direction). We can do the same sort of thing with more than two classes. In particular, we can seek that direction which maximizes the *variance* between the class means, standardized for the (assumed common) within class variance. Put another way, we seek that direction which gives the largest between-to-within variance ratio. Analogous to the preceding section, if the covariance matrix of the group means is $\mathbf{B}$ and the (common) covariance within the groups is $\mathbf{S}$, then we seek the direction $\mathbf{a}$ which maximizes $\mathbf{a}'\mathbf{B}\mathbf{a}/\mathbf{a}'\mathbf{S}\mathbf{a}$. In fact, a subtle point arises here. $\mathbf{B}$ can be estimated as an unweighted matrix, proportional to $\sum_j (\overline{\mathbf{x}}'_j - \overline{\mathbf{x}})'(\overline{\mathbf{x}}'_j - \overline{\mathbf{x}})$, where $\overline{\mathbf{x}}_j$ is the mean vector for class j and $\overline{\mathbf{x}}$ is the overall mean vector, or as $\sum_j n_j (\overline{\mathbf{x}}'_j - \overline{\mathbf{x}})'(\overline{\mathbf{x}}'_j - \overline{\mathbf{x}})$, where n_j is the number of objects in the sample which belong to class j. The latter is the more common; it places greater emphasis on separating the more important – larger – groups, and so is what is normally required.

The vector $\mathbf{a}$ which maximizes the ratio $\mathbf{a}'\mathbf{B}\mathbf{a}/\mathbf{a}'\mathbf{S}\mathbf{a}$ is given by the eigenvector corresponding to the maximal root of the equation $(\mathbf{B} - \lambda\mathbf{S})\mathbf{a} = 0$. In general, however, this equation will have more than one solution. This reflects the fact that, with more than two classes, more than one direction is needed to describe the differences between the mean vectors of the classes. With g classes and p variables, there will, in general, be $\min(p, g-1)$ solutions. (Special cases arise when the class means lie in a subspace of less than $\min(p, g-1)$ dimensions – if, for example, they were all in a straight line.) The **eigenvector** corresponding to the second solution is that direction which leads to the greatest value of the ratio, subject to this second eigenvector being orthogonal to the first. The third solution maximizes the ratio in the space orthogonal to the first two eigenvectors, and so on.

These "best separating" dimensions, or *canonical variates* (*see* **Canonical Correlation**), are often used to produce a graphical display of the distribution

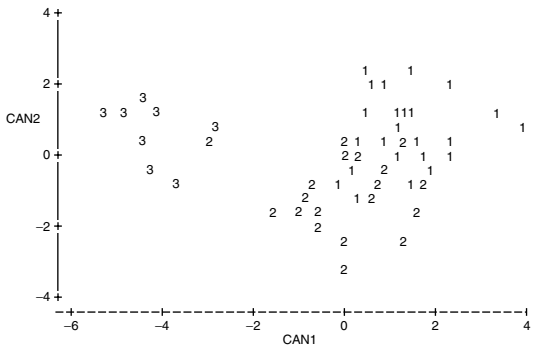

Figure 5 The samples of the three classes of kangaroo skulls, plotted in the space spanned by the first two canonical variates

Table 1 Total-sample standardized canonical coefficients

	CAN1	CAN2
BASLEN	0.61	−2.29
OCCLEN	−2.05	−1.42
NASLEN	3.03	1.93
NASWID	0.71	0.38
ZYGWID	−2.05	1.70
CREWID	−0.76	0.43
MANWID	0.36	1.80
ARAMHT	−0.83	−1.20

of the classes in multidimensional space. Figure 5 illustrates this for three species of kangaroos [2]: 1 = *M. giganteus*, 2 = *M.f. melanops*, 3 = *M.f.* fuliginosus. Eight variables were measured, these being the lengths, in tenths of a millimeter, of basilar length, occipitonasal length, nasal length, nasal width, zygomatic width, crest width, mandible depth, and ascending ramus height. The horizontal axis shows the first canonical variate and the vertical axis the second. The separation between the groups is quite clear.

Table 1 shows the coefficients of the canonical variates, standardized so that the canonical variates have zero mean and unit variance. From this it is possible to see which variables are important contributors to each of the canonical variates.

In the preceding section we showed how the linear discriminant function could be derived using **regression** analysis to predict an indicator variable which took different values for the two classes. This can be generalized to more than two classes: define a set of indicator variables to characterize the different groups and use **canonical correlation** analysis. The mathematics here is also equivalent to **multivariate analysis of variance**. The difference lies in what aspects of the results are the focus of interest. Typically in multivariate analysis of variance interest lies in testing particular between-group contrasts and particular combinations of variables. In discriminant analysis, however, interest lies either in describing the linear combinations which lead to overall differences between the groups (and concern is seldom with particular contrasts) or in formulating a rule permitting classification of future objects.

Assessing Classification Accuracy

A convenient way of summarizing the performance of a classification rule is by means of a *confusion matrix*. This is simply a cross-classification of the predicted class by the true class for a set of objects which the rule has classified. An example is given in Table 2

Table 2 Confusion matrix for classifying chromosomes into 10 classes

		True class									
		1	2	3	4	5	6	7	8	9	10
Predicted class	1	171	2	7	0	0	0	0	0	0	0
	2	3	177	0	0	0	0	0	0	0	0
	3	6	1	172	0	0	1	0	0	0	0
	4	0	0	0	344	16	0	0	0	0	0
	5	0	0	1	16	1379	2	2	0	0	0
	6	0	0	0	0	0	535	0	1	0	3
	7	0	0	0	0	0	0	157	16	7	0
	8	0	0	0	0	0	1	12	334	7	6
	9	0	0	0	0	0	0	9	5	343	3
	10	0	0	0	0	0	1	0	4	3	393

(from Tso & Graham [43]). This shows the results of classifying human chromosomes into ten groups (approximating what is known as the "Denver" classification). As is common with such matrices, the diagonal elements are the largest, signifying that most of the chromosomes are correctly classified. Such a matrix can be used to identify the classes which are most commonly confused.

The most popular single measure of performance is misclassification or error rate – the proportion of new objects which the rule will misclassify (*see* **Misclassification Error**). This is simply the proportion of the objects in the confusion matrix which lie off the leading diagonal. Of course, this measure does not take account of different severities of different types of misclassification, but these can be included with the aid of a cost matrix.

To obtain a reliable estimate of future performance, the confusion matrix (or any other measure of performance) must be computed from a data set other than that used to derive the classification rule. After all, the parameters of the rule will have been estimated to optimize, in some sense, performance on the design set. In the above, the $\bar{\mathbf{x}}^{(j)}$ and $\mathbf{S}$ were estimated from the design set, so that they will be particularly well matched to that data set. Any new data are unlikely to fit the model quite as well. Performance estimates based on the design set are known as *resubstitution* or *apparent* measures. Various approaches to estimating true future performance have been adopted. The most straightforward is to use an independent set of data, a *test set*, but this assumes that sufficient data are available. Alternative approaches involve repeatedly splitting the data into two parts, designing on one and testing on the other, and then averaging the results. They include the **jackknife method**, leave-one-out (*see* **Cross-validation**), and **bootstrap method**. Until recently the leave-one-out approach was the most popular. Here a single element is chosen for the test set and the classifier is built using the remaining $n-1$ design set elements. This is repeated for all n design set elements, and the predicted error rate of the rule based on all n is estimated as the proportion of the classifications which are correct. In general, the leave-one-out method is computationally intensive, since it requires n classification rules to be constructed. However, in the case of linear (and quadratic) discriminant analysis simple variants of the resubstitution estimator have been developed which yield the leave-one-out estimate without recalculating n times from scratch [31].

More recently bootstrap methods have gained in popularity. These involve adjusting the optimistic bias of a performance estimate based solely on the design data using estimates of that bias based on subsamples from the design set. There are several variants, but the so-called *632 estimator* [9] seems to be the most widely recommended. This can be approximated by a cross-validation approach based on using half the data for the design set and half for the test set, doing this repeatedly and averaging the results.

The jackknife procedure has a superficial similarity to leave-one-out, in that it requires n computations based on all but one of the data points, but in fact it is based on a different underlying principle.

The case of two classes is particularly important, especially in medical and epidemiological contexts where interest is often in whether or not a particular disease is present. Because of this, special measures of performance of classification rules have been developed for this case. Table 3 shows the confusion matrix for a generic two-class problem. Suppose that class 1 corresponds to "cases" and class 2 to "noncases". Then the **sensitivity** $(Se)^*$ of a rule is defined as $a/(a+c)$ and the **specificity** $(Sp)^*$ as $d/(b+d)$. Sometimes sensitivity and specificity are called *true positive rate* and *true negative rate*, respectively. Sensitivity and specificity define performance in terms of predicted classifications within each of the true classes. Complementary to this, defining performance as proportions correct within those predicted to belong to each class, we have the *positive predicted value* $(a/a+b)$ and the *negative predicted value* $(d/c+d)$ (*see* **Predictive Values**).

All of these measures, as well as the error rate, are dependent on a particular threshold having been chosen. If this is difficult to do (because, for example, precise costs cannot be determined) then one might prefer to examine performance over a range of

Table 3 Confusion matrix notation for the two-class case

		True class	
		1	2
Predicted class	1	a	b
	2	c	d

situations. **Receiver operating characteristic (ROC) curves** do this by plotting sensitivity on the vertical axis against 1-specificity on the horizontal axis (other equivalent variants are also sometimes used).

Classification is all very well, but sometimes more subtle insights into the performance of a rule are required. For example, one might want to know whether 80% of the objects to which the rule assigns a probability of 0.8 of belonging to class 1 really do belong to class 1. Measures of such things have gone under various names, including *reliability, calibration, validity*, and *imprecision*. Detailed discussion of these and the other performance issues discussed above are given in [23] (*see* **Multivariate Classification Rules: Calibration and Discrimination**).

Robustness

Linear discriminant analysis assumes that the covariance matrices of the classes are equal. If this is not the case, then the true decision surface is nonlinear, so that the linear discriminant analysis decision surface is necessarily biased in some parts of the measurement space. This is particularly important in a biostatistical context, where often the variables are categorical, and so cannot be supposed to follow a multivariate normal distribution. The problem is especially severe when the variables are simply binary – as is the case, for example, if they measure the presence or absence of symptoms. In this situation the covariance matrices are very unlikely to be equal if the mean vectors of the classes differ since, for Bernoulli variables, the means and variances are functionally related (*see* **Binary Data**).

Despite all this theoretical argument, practical studies have shown that the method often performs well. A partial explanation may be found in that fact that, as we have already pointed out in the context of the quadratic extension, the *bias* in the estimated decision surface may be more than compensated for by the reduction in variance which follows from the fact that fewer parameters need be estimated in the linear case. In general it seems that, if the true decision surface is approximately linear, the method will perform well, but this is clearly also a function of the number of variables involved, the sample size, and other aspects (*see* **Robustness**).

There have been many empirical and simulation studies comparing classification rules in general and linear discriminant analysis with other methods in particular. Examples include [16, 19, 35], and [42].

Choosing the Variables

Given a large set of potential discriminatory variables, interest often lies in identifying an effective subset (*see* **Variable Selection**). If the objective of the study is understanding, then there may be doubt about the relevance of all of the variables to the difference(s) between the groups. One might want to describe the differences in a concise and convenient summary. On the other hand, if the aim is to construct a classification rule, then one may want to identify an effective separating subset on practical grounds: the fewer variables which need to be measured on future objects the better, in terms of cost, speed, and so on. Moreover, an aspect of the relationship between bias and variance discussed above is that the more variables there are (relative to a fixed size design set) the more parameters there are to be estimated and the greater is the opportunity for overfitting the design set: the bias may be small but the variance may be large. This can be tackled by reducing the number of variables, so yielding better classification rules.

The obvious approach of choosing variables on the basis of the separation between groups using variables one at a time will generally not be very effective (recall Figures 1 and 2 and the associated discussion). What we want to know is which *set* of variables is effective, when they are taken *in combination*, not when examined individually. To answer this question we must, in principle, examine all possible subsets of variables. If there are p variables to choose from, then there are $2^p - 1$ possible subsets – potentially a vast number. When one considers that, for each such subset, a classification rule must be constructed and its performance assessed, the magnitude of the task becomes apparent. Since there are so many possible subsets of variables, a common strategy is to restrict the search through the space of subsets. The most popular way of restricting this search is to use *stepwise* methods. The basic idea is as follows, illustrated by a *forward* stepwise method. This begins by examining all the variables individually and selecting the best single one. Then each of the others is examined, one at a time, to see which of them, when used together with that already chosen, leads to the best results. The best is added to the one already chosen, to yield a pair. Then each

of the others is examined, to see which, when combined with that pair, yields the most effective triple, and so on. Backward stepwise methods work on a similar principle, but progressively eliminating variables according to which leads to least degradation in performance. More sophisticated versions can also be used, in which groups of variables, rather than single ones, are added or deleted at each step. Forward and backward methods can also be combined, for example, adding two variables and taking one away at each step. Forward methods have the advantage that they are generally less computationally expensive, but they may miss **interaction** detected by backward methods.

At each step of such a procedure it is necessary to decide which of the possible contenders is "the best". Clearly a performance criterion which is quick to evaluate is needed: constructing a classification rule and assessing its error rate may not be feasible. Multivariate analysis of variance constructs measures of the difference between the groups (based on comparing the variation between the mean vectors with the within-group variation, in the way discussed above), and these measures can be used. They include measures such as Wilks's lambda and the Pillai–Bartlett trace [25]. A slightly different variant is to find the subset of variables which maximizes **Mahalanobis distance** between the two closest groups. Other measures are possible and are implemented in various packages.

Often the variables are also examined individually, to ensure that they pass some minimum criterion for selection (in a forward procedure) or maximum criterion for exclusion (in a backward procedure). For example, in the former, if the extra separation produced by a possible new variable beyond that due to the variables already included is less than some threshold (sometimes called the *F-to-enter*), then the variable will not be considered (at this stage, at least). The *F-to-remove* serves a similar role for backward methods. Sometimes these are combined (for example checking if previously included variables have lost their separating power during the course of forward selection). A cautionary note is worth making here. Although these measures of (extra) separability are F statistics, they are the result of a sequence of steps in which "the best" has been selected. They therefore do not follow an $\boldsymbol{F}$ **distribution**.

Other approaches to variable selection include studying the canonical variates to see which variables do not contribute significantly to any, and examining "all" subsets though a search strategy such as branch and bound, which permits some subsets to be identified as suboptimal without explicitly testing them.

McKay & Campbell [32, 33], and Hand [18, Chapter 6] review variable selection methods in discriminant analysis.

Other Variants of Linear Discriminant Analysis

Linear discriminant analysis has been extended in many directions in an effort to produce more flexible models, less constrained by assumptions, and to produce more accurate classification rules. We have already referred to the extension to quadratic discriminant analysis which follows from relaxing the requirement that all the classes should have the same covariance matrix. Some other generalizations are as follows.

A compromise between the extremes of linear discriminant analysis and quadratic discriminant analysis is to require the covariance matrices of the classes to be the same apart from certain parameters. For example, in the *common principal components* model one assumes that they are the same apart from the variances, which may be proportional [12, 13].

Regularized discriminant analysis [14] sought to find an ideal compromise between the extra variability of quadratic discriminant analysis and the possible bias of linear discriminant analysis by adopting a weighted sum of the two. In particular, it estimates the covariance matrix of the jth class by $(1-\lambda)\hat{\boldsymbol{\Sigma}}_j + \gamma c_j \mathbf{I}$, where

$$\hat{\boldsymbol{\Sigma}}_j = \frac{(1-\lambda)(n_j-1)\mathbf{S}_i + \lambda(n-g)\mathbf{S}}{(1-\lambda)(n_j-1) + \lambda(n-g)}$$

and $c_j = (\text{trace } \hat{\boldsymbol{\Sigma}}_j)/p$. This thus shrinks the sample covariance matrix for class j towards the overall average within sample covariance matrix, $\mathbf{S}$, and then further shrinks the result towards the identity matrix (*see* **Shrinkage**).

Quadratic discriminant analysis relaxes the constraint of linear discriminant analysis that the covariance matrices should be equal. The penalty for the increased flexibility resulting from this relaxation is a danger of overfitting the design set. Regularized

discriminant analysis seeks to overcome this by averaging the quadratic and linear approaches. Another alternative is to try to model the covariance matrices, perhaps based on some knowledge of the processes underlying the data. For example, if the variables represent repeated measures, then likely structures for covariance matrices have been well explored (e.g. [7] and [24]). They are based on underlying mechanisms such as **random effects** and **serial correlation** between consecutive measurements. More generally one can model the covariance matrix in terms of postulated underlying relationships. In particular, one can hypothesize or identify likely conditional independence relationships: if two variables are conditionally independent given the others in the model, then the corresponding entry in the inverse of the covariance matrix is zero [6]. This leads to a covariance matrix in which fewer parameters need to be estimated, so reducing the overall variance of the final model.

Krzanowski ([28, 29], and other papers), in his *location model*, described extensions of linear discriminant analysis to the case when the variables are a mixture of continuous and categorical. In essence the model assumes that density functions of the classes are each multivariate normal in the space spanned by the continuous variables, but takes the mean vectors (and, perhaps, the covariance matrices) to differ between the cells induced by the cross-classification of the categorical variables.

Computation

Fisher's 1936 paper introducing linear discriminant analysis was written before the age of computer technology. Because of this, all the major packages (*see* **Software, Biostatistical**) include implementations of this procedure. Some examples are described below, but readers should be aware that software, above all else, evolves rapidly. New releases of programs appear regularly, including improvements and advances. In any case, we have simply attempted to indicate the flavor of the programs available and have not attempted a detailed specification of all of their features.

1. The discriminant procedure in SPSS/PC+ [40] provides a comprehensive routine for linear discriminant analysis. A choice of five criteria is given for choosing which variables should be entered in a stepwise variable selection process, including Wilks's lambda and maximizing the minimum pairwise separation between groups. Various options can be specified to control the selection process in stepwise procedures. The number of canonical variates computed can be specified [up to $\min(p, g-1)$] or alternatively a value for the cumulative percentage of the eigenvalues associated with the variates can be specified. Various diagnostics can be requested, such as Box's M test for equality of the covariance matrices. Classification can be on the basis of equal priors, priors proportional to the design set class sizes, or user-given values. A plot such as that in Figure 5 can be requested, as can various other plots and classification tables. Either quadratic or linear discriminant analysis can be selected for classifying objects.
2. SAS PROC DISCRIM [38] provides a variety of discriminant analysis procedures, including linear discriminant analysis and quadratic discriminant analysis but also including methods such as kernel and k-nearest-neighbor methods. Canonical variates are produced. Both the resubstitution and leave-one-out estimates of error rate can be requested, as can a smoothed estimate with lower variance than the leave-one-out method. Various graphical displays are available.
3. SAS PROC CANDISC [38] determines canonical variates, the scores of the cases on those variates, and performs univariate and multivariate analysis of variance.
4. SAS PROC STEPDISC [39] performs stepwise linear discriminant analysis using forward or backward methods, or a combination of the two.
5. BMDP 7M [3] performs stepwise discriminant analysis by either forward or backward methods. Important contrasts between the groups may be specified to guide the selection procedure. Resubstitution classification results may be requested, as may leave-one-out results. Prior probabilities can be specified and plots based on the first two canonical variates are available.

Conclusion

Linear discriminant analysis is just one of a large class of methods for performing supervised classification. It is the oldest and in some senses the simplest.

Other methods include:

1. logistic discriminant analysis, mentioned above
2. **nonparametric methods** such as k-nearest-neighbor and kernel methods (*see* **Density Estimation**). These apply ideas of local smoothing (*see* **Nonparametric Regression**), relating the predicted class of a new object to the classes of those objects which are most similar to it, where similarity is measured in terms of the variables describing the objects
3. recursive partitioning or **tree-structured statistical methods**, in which the space of the measured variables is sequentially split, to yield a partition within which each cell corresponds to a particular class
4. the feed-forward neural network and other flexible regression models, in which sophisticated combination and transformation procedures are applied to the raw measured variables to yield a predicted classification
5. expert systems, in which patterns of values in the measured variables and in derived variables are sequentially matched to stored patterns (*see* **Artificial Intelligence**).

The literature of the area is now huge. Books devoted to linear discriminant analysis include [31, 27], and [26]. More general works on supervised classification include [18, 23, 34], and [37]. Most books on multivariate statistics include sections on linear discriminant analysis; an excellent example is [30]. Books on statistical pattern recognition, such as [8] and [15], also often discuss linear discriminant analysis, though typically from a different perspective.

References

[1] Anderberg, M.R. (1973). *Cluster Analysis for Applications*. Academic Press, New York.

[2] Andrews, D.F. & Herzberg, A.M. (1985). *Data*. Springer-Verlag, New York.

[3] BMDP (1988). *BMDP Statistical Software Manual*, Vol. 1. University of California Press, Berkeley.

[4] Collett, D. (1991). *Modelling Binary Data*. Chapman & Hall, London.

[5] Cooper, C., Shah, S., Hand, D.J., Compston, J., Davie, M. & Woolf, A. (1991). Screening for vertebral osteoporosis using individual risk factors, *Osteoporosis International* **2**, 48–53.

[6] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies*. Chapman & Hall, London.

[7] Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*. Chapman & Hall, London.

[8] Devijver, P.A. & Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall, London.

[9] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation, *Journal of the American Statistical Association* **78**, 316–330.

[10] Everitt, B.S. (1974). *Cluster Analysis*. Heinemann, London.

[11] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–188.

[12] Flury, B.D. (1988). *Common Principal Components and Related Multivariate Models*. Wiley, New York.

[13] Flury, B.D. (1995). Developments in principal component analysis, in *Recent Advances in Descriptive Multivariate Analysis*, W.J. Krzanowski, ed. Clarendon Press, Oxford.

[14] Freidman, J.H. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association* **84**, 165–175.

[15] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, San Diego.

[16] Gilbert, E.S. (1968). On discrimination using qualitative variables, *Journal of the American Statistical Association* **63**, 1399–1412.

[17] Gordon, A.D. (1981). *Classification*. Chapman & Hall, London.

[18] Hand, D.J. (1981). *Discrimination and Classification*. Wiley, Chichester.

[19] Hand, D.J. (1983). A comparison of two methods of discriminant analysis applied to binary data, *Biometrics* **39**, 683–694.

[20] Hand, D.J. (1986). Pattern recognition, or how to tell it's one of those, in *The Fascination of Statistics*, R.J. Brook, G.C. Arnold, T.H. Hassard & R.M. Pringle, eds. Marcel Dekker, New York.

[21] Hand, D.J. (1992). Statistical methods in diagnosis, *Statistical Methods in Medical Research* **1**, 49–67.

[22] Hand, D.J. (1994). Recent results in pattern recognition theory and applications, *Current Topics in Pattern Recognition Research* **1**, 113–123.

[23] Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. Wiley, Chichester.

[24] Hand, D.J. & Crowder, M.J. (1996). *Practical Longitudinal Data Analysis*. Chapman & Hall, London.

[25] Hand, D.J. & Taylor, C.C. (1987). *Multivariate Analysis of Variance and Repeated Measures*. Chapman & Hall, London.

[26] Huberty, C.J. (1994). *Applied Discriminant Analysis*. Wiley, New York.

[27] Klecka, W.R. (1980). *Discriminant Analysis*. Sage, Beverly Hills.

[28] Krzanowski, W.J. (1976). Canonical representation of the location model for discrimination or classification, *Journal of the American Statistical Association* **71**, 845–848.

[29] Krzanowski, W.J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data, *Computers and Mathematics with Applications* **12a**, 179–185.

[30] Krzanowski, W.J. & Marriott, F.H.C. (1995). *Multivariate Analysis Part 2: Classification, Covariance Structures, and Repeated Measurements*. Arnold, London.

[31] Lachenbruch, P.A. (1975). *Discriminant Analysis*. Hafner, New York.

[32] McKay, R.J. & Campbell, N.A. (1982). Variable selection techniques in discriminant analysis I: description, *British Journal of Mathematical and Statistical Psychology* **35**, 1–29.

[33] McKay, R.J. & Campbell, N.A. (1982). Variable selection techniques in discriminant analysis II: allocation, *British Journal of Mathematical and Statistical Psychology* **35**, 30–41.

[34] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

[35] Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.

[36] Norušis, M.J. (1985). *SPSSX Advanced Statistics Guide*. McGraw-Hill, New York.

[37] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

[38] SAS (1989). *SAS/STAT User's Guide, Version 6*, 4th Ed., Vol. 1. SAS Institute Inc., Cary.

[39] SAS (1989). *SAS/STAT User's Guide, Version 6*, 4th Ed., Vol. 2. SAS Institute Inc., Cary.

[40] SPSS (1988). $SPSS/PC + Advanced\ Statistics\ 4.0$ *for the IBM PC/XT/AT and PS/2*. SPSS Inc., Chicago.

[41] Späth, H. (1985). *Cluster Dissection and Analysis*. Ellis Horwood, Chichester.

[42] Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F. & Gelpke, G.J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients (with discussion), *Journal of the Royal Statistical Society, Series A* **144**, 145–175.

[43] Tso, M.K.S. & Graham, J. (1983). The transportation algorithm as an aid to chromosome classification, *Pattern Recognition Letters* **1**, 489–496.

(*See also* **Multivariate Analysis, Overview**)

DAVID J. HAND

Discrimination and Clustering for Multivariate Time Series

New sophisticated instrumentation in biology and medicine will routinely result in massive **databases** that often are composed of many series that are measured over time. Examples are micro-arrays in gene studies (*see* **DNA Sequences**), and functional magnetic resonance images (fMRI) or EEG series in brain imaging studies (*see* **Image Analysis and Tomography**; **Clinical Signals**). Classic problems in analyzing such observed multivariate **time series** involve (1) the grouping or clustering of the time realizations into similar categories (*see* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**) and (2) the **classification** of new observed series, possibly belonging to one or more of the categories. Examples are clustering of gene patterns associated with a particular disease using micro-arrays, somatosensory discrimination using fMRI profiles, or detection of early onset Alzheimer's disease using multiple EEG sensors.

All of the above experiments receive data in a similar format, namely, as multivariate time series consisting of observations on a vector $\boldsymbol{y}_t = (y_{t1}, y_{t2}, \ldots, y_{tp})'$ observed over a number of time points, say $t = 1, 2, \ldots, n$. Typically, the number of time points, n, exceeds the number of components of the vector, p. For example, EEG measurements may contain thousands of observations in time measured at $p = 19$ channels monitoring different areas of the brain. Data from such studies are often divided a priori into a number of groups or populations, say $\Pi_1, \Pi_2, \ldots, \Pi_m$. A new measurement comes in and is to be classified into one of the groups using **discriminant analysis**. A more challenging problem materializes in *cluster analysis* when there are no a priori subgroups defined over the database. In that case, one wishes to determine the number of groups, m, and the group membership of experimentally observed series using one of a family of procedures for defining clusters.

Discrimination and clustering problems have, of course, been studied for conventional **multivariate** data and there exists a substantial literature devoted to discrimination and clustering of vector observations (for example, see [3, Chapters 11 and 12]; also **Multivariate classification rules: calibration and discrimination**). The components of the vector can be features extracted from the multivariate time profiles $\boldsymbol{y}_t$ measured by instrumentation like that mentioned above. In the physical and engineering sciences, there is a long history devoted to extracting features of time series that can be used for discrimination. For example, Shumway and Stoffer [6, Section 5.7] show an example that involves distinguishing series originating from earthquakes from those that might be nuclear explosions using various amplitude values extracted from the bivariate seismic recordings. The key to applying this *feature extraction* approach is the reduction of the vector time series to a small vector of discriminating features. Although the feature extraction approach can be moderately successful in the hands of a skilled analyst, the magnitude of the databases suggests that discrimination and clustering techniques based on using the complete waveforms are potentially more powerful.

For general discriminant analysis, one usually has a collection of p-dimensional vector time series representing each of m population groups, say $\boldsymbol{y}_{kti}, i = 1, 2, \ldots, n_k, k = 1, 2, \ldots, m$. The n_k $p \times 1$ vectors from group Π_k are used to find **maximum likelihood** estimators for the common parameters of the group, denoted generically here by $\hat{\Theta}_k$. These common parameters determine a log-**likelihood** function for each of the groups, say $\log L(\hat{\Theta}_k), k = 1, 2, \ldots, m$. We may also evaluate the log-likelihood for a new observation, say $\log L(\hat{\Theta})$ at the estimated parameters for that particular observation. The new observation is classified by choosing the population corresponding to the value of k leading to the minimum value of the log-likelihood ratio, namely $\log L(\hat{\Theta}) - \log L(\hat{\Theta}_k)$. This difference and various forms can be regarded as a rough measure of distance between the observed vector and the k^{th} group. Kakizawa et al. [5] show a number of different forms based on information theoretic measures of discrepancy.

Hierarchical cluster analysis assigns observations to clusters based on the same concept of distance. Since we have no a priori knowledge of either the number of clusters or the cluster composition, we begin with N clusters, each containing a single member, where N is the size of the database. Then, define the two closest members as a new single cluster, leading to a new partition into $N - 1$ clusters composed of $N - 1$ single member clusters and one cluster with

two members. Then, evaluate all between-cluster distances using the distance between the two closest elements of the cluster. Merge the two closest into a new cluster The procedure stops when every element of the database has been merged into a single cluster. An alternative *partitioned cluster analysis* approach begins with a set of clusters of a given size and adjusts the membership by evaluating sequentially the distance of each element in a cluster to every other cluster. If it is closer to a cluster to which it does not currently belong, it is moved to the closest neighboring cluster. When no more interchanges are indicated the current membership determines the cluster configuration for a given number of clusters.

There are two possible approaches to modeling the observed data leading to an evaluation of the log-likelihood function mentioned above. For lack of a better description, we divide the modeling approaches into those emphasizing *time domain* methods and those emphasizing *frequency domain* methods. Frequency domain methods (*see* **Spectral Analysis**) have the advantage that the representation of the data in terms of **stationary** or locally stationary processes agrees with physical intuition, suggesting that periodic phenomena can be modeled best in terms of cyclical behavior. Time domain methods (*see* **ARMA and ARIMA Models**) have the advantage that they can often be couched in terms of regression models for which there will be a large body of statistical software for computations. As a general rule in biostatistical applications, **longitudinal data** can often be fitted using time domain methods, whereas the frequency domain often is better suited for large data sets such are produced in fMRI and EEG analysis.

Time domain approaches usually attempt to model the vector series $\boldsymbol{y}_t$ as linear combinations of fixed **covariates** and **stochastic processes** and characterize population membership in terms of the parameters of the linear model. A number of examples of these kinds of representations for vector series can be found in [2] for fMRI series and in [6, Chapter 4], [4], or [1] for models that can be put into state-space form. In these models, one usually has an equation in mind for the observations that includes fixed covariates capable of modeling smooth behavior and a random series that induces a smoother stochastic component. Structural forms for these models in the multivariate case are given in the previously mentioned references. In general, there will be a parametric form for the process that will be typical of its population grouping, leading to a value of the log-likelihood $\log L(\hat{\Theta}_k), i = 1, 2, \ldots, m$ for each group. Computing the difference between the log-likelihood of the observation and that for the group leads to metrics for discriminant and cluster analysis in this case.

Frequency domain methods may be indicated when the stimuli are repetitive, as in many fMRI experiments, or when the response is expected to contain important information at given frequencies. For example, the alpha and beta frequencies for an EEG series are thought to be important components. In the frequency domain model, it is convenient to parameterize the Fourier transforms (*see* **Fast Fourier Transform (FFT)**) of the multivariate series as being approximately complex **multivariate normal** with covariance or spectral matrices that differ for the different population groups. In this case, we still evaluate a log likelihood, called the Whittle [7] likelihood, over frequencies of interest and follow the same general procedures as above for discriminant and cluster analysis. The population differences in this case are assumed to be characterized in terms of the spectra and coherences between the multivariate recordings. The data in this case are summarized by the *spectral matrix* of the vector. The likelihood classification and clustering then works best when using functionals of the estimated group k spectral matrix, $\hat{S}_k(f, t)$, and the estimated spectral matrix of the observation to be classified, $\hat{S}(f, t)$, where f denotes frequency in cycles per unit time and t denotes time. Examples using this approach for discriminating seismic recordings of earthquakes from those generated by presumed nuclear explosions are given in [6, Chapter 5, Section 5.7]. One can average over informative frequencies and over time windows where the data are locally stationary.

References

[1] Durbin, J. & Koopman, S.J. (2001). *Time Series Analysis by State-Space Methods*. Cambridge University Press, Cambridge.

[2] Fahrmeir, L. & Gössl, C. (2002). Semiparametric Bayesian models for human brain mapping, *Statistical Modeling* **2**, 235–249.

[3] Johnson, R.A. & Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*, 5th Ed. Prentice Hall, Englewood Cliffs.

[4] Jones, R.H. (1993). *Longitudinal Data With Serial Correlation: A State-Space Approach*. Chapman & Hall, London.

[5] Kakizawa, Y., Shumway, R.H. & Taniguchi, M. (1998). Discrimination and clustering for multivariate time series, *Journal of American Statistical Association* **93**, 438–340.

[6] Shumway, R.H. & Stoffer, D.S. (2000). *Time Series Analysis and Its Applications*. Springer-Verlag, New York.

[7] Whittle, P. (1961). Gaussian estimation in stationary time series, *Bull Int. Stat. Inst* **33**, 1–26.

ROBERT H. SHUMWAY

Disease Registers

Disease registers are an important tool for clinicians, epidemiologists, and health service planners. Their nature will vary according to the functions they are serving, but all relate to individuals with, or at high risk of, a specified chronic disease. Often registers are used for more than one purpose.

The following sections describe: examples and objectives of registers of different types; problems of case definition, **ascertainment**, and **biases**; validity checks; and possibilities created by **record linkage**.

Types of Disease Registers

Registers Contributing to the Organization and Quality of Clinical Care

Patient Registers Held by Clinicians. The simplest form of disease register is one set up by individual clinicians relating to their own patients. The conditions registered are usually those that require either regular maintenance therapy or **screening** for early signs of preventable complications. Such registers are essentially part of the normal process of clinical care. They are generally designed for **prevalent cases**, i.e. patients who are alive, have not moved away, and whose condition is clinically important.

One of the most common conditions for which disease registers are used is diabetes mellitus. This is typical in that most patients have an ongoing need of insulin or another prescribable drug. They are also at high risk of future complications, particularly problems of the feet, or eyes, and of the cardiovascular system. Many of these complications have been shown to be either preventable, or less serious if diagnosed and treated early.

In recent years the holding of disease registers has extended from one held by a clinician of his/her own patients, with a special interest in a particular condition, to the sharing of registers by groups of clinicians working in general or hospital practice. Moreover, for an increasing number of conditions, including diabetes [6] and coronary artery disease [11], there are now internationally shared registers, with all the necessary **confidentiality** constraints.

Registers of Relatives of Patients with Genetic Conditions. A more recent development is to extend registration to blood relatives of individuals known to have a serious genetic condition. An example of this is the condition of familial adenomatous polyposis. Persons with this condition have numbers of colonic polyps, initially benign but at high risk of becoming malignant.

One method of management is to screen teenage members of affected families and to remove the colon of those found to have polyps at the age of 18–20 as a prophylactic measure [2]. Another approach under investigation is to treat those at high risk with low-dose aspirin, which may be protective against malignancy. Implementation of such programs on a population scale, and their audit, is greatly assisted by the existence of registers of those at risk.

Registers of Individuals at Risk because of Hazardous Exposure. Where specific hazards are known to increase the risk of subsequent serious disorders, registration of those who have been exposed may be a useful clinical tool. This has been done for babies who were born extremely immature, or who have had severe asphyxial episodes. These babies are at high risk of neurologic damage which may not manifest itself for some years. Early diagnosis of sensory or neurologic problems in children on such registers allows prompt action, although it is mostly in visual and hearing disorders that it has been shown to be effective [5]. Parents of children at risk because of stormy births are normally aware of such risks and appreciate the **surveillance**.

Registers Held for the Implementation of Public Health Functions

Registers of serious common conditions may be held for the purpose of monitoring and improving the health of populations as opposed to that of individuals. They are usually held at the level of residents of an administrative area. Questions of exclusion or inclusion may arise when residents of one area are treated, move into, or are born or die, in another area.

In contrast to most registers held for clinical purposes which need only include *prevalent* cases, registers held for public health functions may need to include all **incident cases** regardless of severity or survival. Moreover, for most public health functions the measures used will be **incidence** or **prevalence rates** rather than absolute numbers of cases.

The calculation of rates implies that the number of individuals at risk is known. In some registers, for

instance those of congenital anomalies, this implies the inclusion of cases lost as late prenatal or postnatal death amongst the numerator and denominator. This is possible where there is a definitive diagnostic test which can be used prenatally or in the early postnatal period, e.g. the detection of a chromosomal anomaly, of a specific genetic defect, or ultrasound visualization of malformations visible during or after pregnancy [10].

The aims of such registers include:

1. the ascertainment of environmental hazards to health (*see* **Environmental Epidemiology**)
2. the provision of current and projected prevalence and severity data for health care planners (*see* **Health Services Research, Overview**)
3. the monitoring of **survival**, or **quality of life**, of affected individuals
4. the monitoring of the efficacy, implementation and acceptance of preventive measures.

Important examples are cancer, diabetes, ischemic heart disease, or congenital anomaly registers, which may be held at regional, national, or international levels.

Ascertainment of Environmental Hazards to Health. New environmental causes of ill-health may be suspected when trends in registration rates of specific conditions change over time, in different places or in persons of different characteristics, or occupations. New patterns of incidence, such as clusters over time and space, may also draw attention to possible causes (*see* **Clustering**). When the conditions concerned are rapidly lethal, or lethal prenatally, it is important to ascertain all incident cases as far as possible, as well as prevalent cases. Where evidence is found that there has been a real change in incidence, registered cases may act as a **sampling frame** to set up **case–control studies** to investigate possible causes.

Negative findings from such studies are as important as positive findings if they can rule out putative associations with environmental exposures.

Provision of Current and Projected Prevalence and Severity Data for Health Care Planners. Health care planners require information to allow them to project future needs of individuals with specific conditions, in terms of prevalence and severity. This requires good quality prevalent disease registers, which include clinical and survival data.

Monitoring of Outcome of Affected Individuals. Registers which provide information on survival, quality of life, and treatment given, are important sources of clinical audit, allowing the comparison of survival after different treatments or treatment in different places. Such audit will, however, also require basic demographic data such as age, sex, place of residence and, if possible, socioeconomic circumstances, which could **confound** comparisons of survival. Such comparisons, although not so rigorous as randomized controlled trials (*see* **Clinical Trials, Overview**), may point to differences that should be further explored.

Monitoring of the Efficacy, Implementation, and Acceptance of Preventive Measures. Preventive measures of disease may include primary prevention, namely the abolition of the cause. In the case of cancers or heart disease, these include smoking or alcohol abuse. Preventable serious congenital disorders include neural tube defects, in part preventable by periconceptional folic acid supplementation, and rubella embryopathy, preventable by preconceptional rubella immunization. Where registers exist of incident cases of such conditions, trends over time, place, or in different population groups will act as an audit of the extent to which the preventive action is being implemented.

In conditions where secondary preventive action may follow the screening out of asymptomatic or prenatal cases, the use of a disease register to monitor the prevalence of symptomatic cases, or births with specific congenital anomalies, will allow the auditing of the efficacy and effectiveness of specific screening programs. Examples are where cervical or breast cancer screening is on offer, and how this affects the mortality due to such cancers; or where prenatal screening programs are available, whether there is a change in ratio of legally terminated pregnancies with Down's syndrome or neural tube defects to registered affected births.

Registers Held for Research Purposes. Registers may be held purely for etiologic research. They typically require the inclusion of incident rather than prevalent cases. Their design and maintenance must

take account of, or may reveal, the **natural history** of the condition registered.

How the Natural History of a Disease may Affect Registration

Congenital Anomalies

Many congenital conditions which lead to permanent impairment are ascertainable and therefore registerable at birth, e.g. spina bifida. Others may not be visible or do not lead to symptoms until some time after birth. These include congenital heart defects, cerebral palsy, or mental retardation. For such conditions it is impossible to estimate true incidence rates since many affected children may have died before ascertainment, and only age-specific prevalence rates can be calculated. Other congenital conditions, e.g. gastrointestinal atresias, are curable by surgery shortly after birth. Such conditions may need to be considered in the ascertainment of incident cases, but not in ascertaining prevalent cases.

The advent of prenatal diagnosis of some conditions, often leading to termination of pregnancy, raises other questions. Had they not been prenatally diagnosed, many fetuses with conditions such as chromosomal anomalies would have been lost as spontaneous miscarriages, and the cause would not have been ascertained. This is an important point in registers of Down's syndrome, where in recent years in England and Wales about half of all affected pregnancies are diagnosed prenatally, leading to an apparent increase in incident cases, although the numbers of affected births are falling [1].

Acquired Diseases

Acquired chronic conditions may lead to permanent impairment which cannot be cured, or they may be "curable", at least in the sense of not recurring. The course of the disease may be variable, as in multiple sclerosis, with attacks and remissions, the patient sometimes having no symptoms or clinical signs in remission. Alternatively, in conditions such as ischemic heart disease, minor symptoms and signs of the disease may persist, but this may be punctuated by acute episodes of myocardial infarction. Tunstall–Pedoe [11] gives a full account of the methodologic problems raised in the registration of ischemic heart disease, stemming from the notification of heart attacks as acute episodes instead of as "abstractions from a chronic disease". He shows that such registration, which has been used for international comparisons, is the only way to measure the burden of chronic heart disease.

Case Definition

The nature and quality of a register is crucially dependent upon the ascertainment of individuals meeting a clear and unambiguous case definition. Case definition must include guidance on which cases should be included and which excluded, including the cutoff points in terms of level of severity or objective test results. Where relevant, it is helpful if registration forms include diagrams which indicate the parts of the body that are affected, or scales which indicate severity.

Sometimes researchers may choose to use a very broad definition on the assumption that they can select specific subgroups from the information requested.

One question that commonly arises is how to handle cases with multiple pathology, e.g. multiple apparently unrelated malformations or cancers. Particularly for research purposes, the setting up of a register must include a protocol which deals with these questions and the method of ascertainment to be used. For clinical registers held by one practitioner this may be less important, but as soon as clinical registers are shared with others (and this often implies a new use as a research tool also), such a protocol is equally important.

Ascertainment

The completeness of registers varies with the methods used for ascertainment and diagnosis.

Methods of Ascertainment

Ascertainment may depend on clinical presentation and the recognition of the defined condition, or it may be the result of a process of screening where this is clinically possible. Both the severity of the condition and the characteristics of the individuals ascertained will usually vary sharply depending on the methods

used. Examples are the marked differences between numbers of cases on registers of individuals with diabetes mellitus who presented for medical care, and those registers which resulted from population screening of urine and glucose tolerance testing [3]. The ease and completeness with which cases are found may be helped if the treatment is standard and unique, as in the case of insulin, where monitoring of prescriptions is a method of ascertainment of cases.

Methods of Diagnosis

Particularly where the register is shared with others, the diagnostic process must be based on a formal protocol. There are many different methods of diagnosis. For those conditions where a definitive diagnostic test is available e.g. an identifiable single **gene** defect or a chromosome anomaly, or the results of bacterial or viral culture, the easiest and most complete ascertainment may be obtained from laboratory results. Where diagnosis is largely based on clinical findings its success may depend on the personal acumen of the physician, but is usually backed up by objective blood, urine, or imaging investigations. Such methods may lead to full, or nearly full, ascertainment where the condition is such that self-referral is the rule. For lethal conditions clinical ascertainment can be backed up by searching for relevant details on **death certificates**.

Multiple Sources of Ascertainment

It is now increasingly common to use multiple sources of ascertainment. This can be particularly useful for chronic conditions of low lethality which may not always require medical care. For instance, individuals with conditions such as cerebral palsy or mental retardation may present to a variety of services – medical, paramedical (such as physiotherapy), educational, or social. Moreover, it has been shown that even the ascertainment of diabetes or cancer can be improved by the use of multiple sources. For diabetes, multiple sources that have been used include prescriptions, family practitioner registers, hospital diabetic clinic records, and, where relevant, health insurance data. For cancer registrations, sources include hospital records or death certificates with a mention of cancer, histology reports, and oncology clinic records.

Multiple ascertainment is designed to lead to duplication of notification, and the information gathered and the design of the register must be such that duplicates can be identified and eliminated.

Duplicate Notification. Duplicate notification will also occur where affected individuals already registered in one place move to another place which keeps a related register, and precautions must be taken to eliminate these. Clerical errors in recording dates, spelling mistakes in recording names, or name changes are all difficulties which must be taken account of in seeking for duplicates. Record linkage techniques can be used to check for duplicates, including phonetic name matching [7].

Case Identification

The degree to which registered cases need personal identification will vary according to the aims of the register. Clinical registers are often part of family practitioner or hospital records, and named identification is essential for their use.

Registers kept for public health or research purposes often do not need to be named except where a follow-up of registered individuals is planned. On the other hand the recording of some personal identifiers is essential, if only to allow the finding and elimination of duplicates, and to have such basic epidemiologic information as date of birth and sex.

If the aims of the register include an investigation of changes in incidence, prevalence, or survival, other dates must be collected, such as at first presentation and, where relevant, of death. Similarly, to seek for evidence of clustering, place of residence, and sometimes of birth or occupation, will be needed, usually recorded in the form of post- or zip-code data. When personal identifiers are kept there must be meticulous care to preserve patient confidentiality.

Record Linkage

The growth of computerized health information (*see* **Administrative Databases**) has led to opportunities to link records from different sources, and thus to enhance register information.

For instance, in the UK it is possible to access information from death registration. Given Ethics

Committee permission, bona fide researchers are permitted to arrange for the linkage of this information with their own register data, thus providing the necessary data to calculate the survival of the individuals on a register, and to find their causes of death. Similarly, linkage with nonconfidential items from birth records may provide valuable information linking birth events with health in later life [9].

Such linkage is an essential part of cancer registration in the UK, since recording of cancer as a **cause of death** is an important method of ascertainment for the regional cancer registers. Moreover, the linkage at national level of data from all regional cancer registers allows for identification and elimination of duplicate records due to patient movements [8].

Validation

An important aspect of maintaining the quality of disease registers is the validation of the data at regular intervals. Validation can include an assessment of completeness of registration and of data on each record, the success with which duplicates are eliminated, and most importantly the rigor to which the given case definition is adhered.

There is a growing literature on methods of examining the likely completeness of ascertainment. Where there are different but independent methods, **"capture–recapture"** techniques can be used [4].

The examination of the validity of case registration can be a difficult task. An area where this has received particular attention is in the **World Health Organization** MONICA study, which was the registration in a number of different countries of heart attacks. This is well described by Tunstall-Pedoe [11], who discusses problems arising from different standards of record keeping, the use of coding rules for clinical history, symptoms and diagnostic tests, and validity checks of the clinical data, coding, and laboratory or other tests.

Conclusions

Disease registers are becoming an increasingly powerful clinical and epidemiologic tool, particularly for international comparisons. Their use predicates clear aims, good design, coverage, complete and accurate recording of validated data, and rigorous methods of preserving confidentiality.

References

[1] Alberman, E.D., Mutton, D.E., Ide, R., Nicholson, A. & Bobrow, M. (1995). Down's syndrome births and pregnancy terminations in 1989–1993: preliminary findings, *British Journal of Obstetrics and Gynaecology* **102**, 445–447.

[2] Bulow, S., Bulow, C., Nielsen, T.F., Karlsen, L. & Moesgard, F. (1995). Centralized registration, prophylactic examination, and treatment results in improved prognosis in familial adenatomatous polyposis. Results from the Finnish Polyposis Register, *Scandinavian Journal of Gastroenterology* **30**, 989–993.

[3] Butterfield, J. (1964). Summary of results of the Bedford Diabetes Survey, *Proceedings of the Royal Society of Medicine* **57**, 196–200.

[4] Hook, E. & Regal, R.R. (1992). The value of capture-recapture methods even for apparent exhaustive surveys, *American Journal of Epidemiology* **135**, 1060–1067.

[5] Johnson, A. (1995). Use of registers in child health, *Archives of Disease in Childhood* **72**, 474–477.

[6] Krans, H.M.J., Porta, M., Keen, H. & Staehr Johansen, K. (1995). *Diabetes Care and Research in Europe: The St. Vincent Declaration Action Programme*. International Diabetes Federation European Region, World Health Organization, Regional Office for Europe, Copenhagen, Chapter 14.

[7] Langley, J.D. & Botha, J.L. (1994). Use of record linkage techniques to maintain the Leicestershire Diabetes Register. Computer Methods Programs, *Biomedicine* **41**, 287–295.

[8] Office of Population Censuses and Surveys (now Office of National Statistics) (1990). *Review of the National Cancer Registration System in England and Wales*. HMSO, London.

[9] Office of Population Censuses and Surveys (now Office of National Statistics) (1993). Uses of OPCS Records for Medical Research, *Occasional Paper 41*. HMSO, London.

[10] Office of Population Censuses and Surveys (now Office of National Statistics) (1995). The OPCS Monitoring Scheme for Congenital Malformations, *Occasional Paper 43*. HMSO, London.

[11] Tunstall-Pedoe, H. (1989). Diagnosis, measurement, and surveillance of coronary events, *International Journal of Epidemiology* **18**, Supplement 1, 169–173.

(*See also* **Birth Cohort Studies**; **Death Indexes**; **Follow-up, Active Versus Passive**; **Twin Registers**)

EVA ALBERMAN

Disease-marker Association

The primary aim of a disease–marker association study is to evaluate the potential role of a **gene**(s) in the expression of a measured trait. The gene may be measured directly at the DNA level (such as **DNA sequence** variation), at the level of the gene product (such as blood serum proteins and enzymes), or indirectly by DNA markers (such as restriction–fragment–length–**polymorphisms** or simple sequence repeats) that are on the same chromosome and physically close to, and associated with, the disease-causing gene. It is convenient to refer to any of these measured phenotypes of a gene as a **genetic marker**, which is defined as a genetically determined trait for which the relationship between **genotype** and phenotype is known. The DNA markers chosen for studies are often codominant, i.e. there is a one-to-one relationship between marker genotype and phenotype, allowing unambiguous determination of marker alleles. When a marker is not codominant, the error of classification of the genotype based on the marker phenotype should be considered in analyses, which requires knowing the distribution of the marker phenotype, given the marker genotype, as in a general **penetrance** function.

For a marker to be useful in an association study, it should have sufficient variation (i.e. **polymorphism**) in the population. Measures of polymorphism for a marker are (i) **heterozygosity**, the probability of having a heterozygous genotype and (ii) **polymorphism information content**, which was derived for linkage studies of a rare autosomal dominant disease using a codominant marker, and which corrects the heterozygosity for noninformative matings when parents and child all have the same heterozygous genotype.

Although the cause of association between genetic markers and disease cannot be determined by association studies, it is important to consider the possible causes of association so that potential biases can be evaluated. One most desirable cause is the direct effect of marker alleles on the trait phenotypes, such as in a candidate–gene study. A second, indirect, cause of association is **linkage disequilibrium** (LD), i.e. both linkage between disease and marker loci and nonrandom association of the alleles at these two loci on chromosomes in the population. This type of association is likely to occur if most diseased subjects inherited from a common ancestor a segment of chromosome containing the disease allele and marker allele (founder effect), and is most easily detected in a homogenous population, because the main factor influencing the association is recombination between the two loci. Recombination breaks association between the alleles at the two loci, so that after many recombinations, which accumulate over generations, the alleles at the two loci will be randomly associated on chromosomes in the population. However, this is a slow process when the chance of recombination, θ, is small. Denote by m_i a marker allele with frequency p_i, d_j an allele from the disease-causing locus with frequency q_j, and h_{ij} the frequency of a chromosome bearing alleles m_i and d_j (i.e. **haplotype**). The deviation of the haplotype frequency from random association (i.e. equilibrium) is $h_{ij} - p_i q_j$, which is expected to decrease by the factor $(1-\theta)^n$ after n generations, where θ is the recombination fraction (*see* **Linkage Analysis, Model-based**). Because of this, association studies based on LD are generally not sensitive to $\theta > 1\%$ (1 centimorgan).

A complication of associations caused by LD is that different mutations causing the same disease phenotype can arise on chromosomes that bear different marker alleles. This can cause the associated marker allele to differ across different populations, and can decrease LD in a population mixed with different mutations. To study this effect in detail, one could perform studies of haplotypes composed of multiple marker loci to determine whether particular haplotypes are associated with disease. However, determination of haplotypes may require much work (e.g. family studies), and there are statistical errors when inferring haplotypes. Note that if there are multiple genes over a short chromosomal region, each of which can be associated with disease, then a marker within that region can be associated with disease due to any one (or more) of those genes, making it difficult to determine the causative gene(s). Examples of this complexity occur in association studies of the major histocompatibility complex (*see* **HLA System**).

In addition to direct effects or LD causing associations, there are other less interesting causes of association that can exist at the population level, yet which can cause misleading interpretations. These are joint selection for both marker and disease locus

alleles, small population variation (random genetic drift (*see* **Population Genetics**)), and the structure of the population, which may include **inbreeding**, or stratification due to either **admixture** of different ethnic groups or nonrandom mating. If a population is composed of a recent admixture of different ethnic groups that have different frequencies of marker alleles, then any trait more frequent in an ethnic group will be positively associated with any marker allele that is more frequent in that group, even if these loci are not linked. This spurious association is an example of **confounding** due to ethnic background. Although linkage disequilibrium is often used to describe general associations between disease and marker alleles, it is better to use the term "allelic association", because of the other causes of association described above.

The traits evaluated in association studies are often based on affection status (diseased vs. normal), as in **case–control studies**. For this reason, case-control studies are discussed in detail, although other study designs may be useful: **cohort, cross-sectional**, admixed populations, case–parent, family-based controls, haplotype analyses, and pedigree studies. Furthermore, it is important to recognize that marker association studies are useful to address a number of scientific questions, such as prognosis of survival outcome of diseased subjects based on marker genotypes (using censored **survival analysis** methods), or the amount of variation of quantitative traits, among either diseased or normal subjects, explained by genetic markers (using **regression** analysis and **analysis of variance**).

Criteria for selection of diseased cases often include newly diagnosed **incident cases** (instead of **prevalent cases**, which are confounded by disease duration and survival), and ways to minimize phenocopies (e.g. strong family history, early onset, severe cases). To avoid confounding, it is best to match cases and controls by potential confounders such as age, sex, and ethnicity. Because ethnicity can be difficult to define and measure, and population structure may not be known, the choice of adequate controls can be quite difficult. Although relatives of cases may serve as convenient controls, especially for matching on ethnic background, the statistical dependence between cases and their relatives can lead to less **power** than a random sample of controls.

The comparison of marker phenotype frequencies, say G different types, between cases and controls can be performed using traditional methods for **case–control studies** [4, 27] (*see* **Analytic Epidemiology**): $2 \times G$ tables with **chi-square tests** or **exact tests** [35], cross-product **odds ratios** (for small samples it may be necessary to use Haldane's formula [10] by adding 0.5 to the cells of the table before computing odds ratios), **logistic regression** (unconditional or conditional for matched studies), and population **attributable risk** [1]. When the number of marker phenotypes (G) is large, it is common practice to compare the frequency of the presence of particular marker phenotypes, such as carriers of particular alleles, between cases and controls in multiple, **2 × 2 tables**, with correction for multiple testing (*see* **Multiple Comparisons**). Alternatively, when marker alleles can be unambiguously determined, as for codominant markers, power may be improved by reducing the large number of categories based on marker phenotypes to fewer categories based on marker alleles, say K distinguishable alleles. Allele frequencies can then be compared between cases and controls by constructing a $2 \times K$ **contingency table**, such that each person contributes two alleles, and then calculating a probability value based on either the large sample **chi-square distribution** of the Pearson chi-square statistic or **Monte Carlo** testing [30]. The chi-square statistic has $(K - 1)$ degrees of freedom. When it is plausible that a disease susceptibility allele is associated with only one of K marker alleles, although it is not known which allele is associated, power can be improved by use of a mixture **likelihood**, which uses the frequency of marker alleles to predict which allele is associated with disease, and creation of a **likelihood ratio** statistic [37].

The validity of Pearson's chi-square statistic for comparing allele frequencies requires independence of alleles in the general population. When randomly sampling cases and controls, genotypes are independent between people, but alleles within genotypes may not be independent, such as can occur when there is recent mixture of populations. Independence of alleles can be tested by comparing the observed genotype proportions to those expected when there is **Hardy–Weinberg Equilibrium** (HWE). This test should be performed only among the controls, because even if the general population is in HWE, the

expected marker genotype proportions among diseased cases can deviate from HWE when a true association exists, and the amount of deviation depends on the genetic mechanism. For example, if a marker allele is associated with a disease caused by a rare dominant disease susceptibility allele, then HWE is not expected to hold, yet for association with a recessive disease susceptibility allele, HWE may hold among the cases, but with a marker allele frequency greater than in the general population [21, 42].

If HWE does not hold among the controls, then the variances (and covariances) of allele frequencies should not be based on **binomial** variances (and covariances), because they do not account for the dependence of alleles. For example, let $n_{\mathrm{A}\cdot}$ and n_{AA} be the number of controls **heterozygous** and homozygous, respectively, for allele A, and let N be the total number of controls. The estimated relative frequency of allele A is $\hat{q}_{\mathrm{A}} = (n_{\mathrm{A}\cdot} + 2n_{\mathrm{AA}})/(2N)$, and the estimated frequency of AA homozygotes is $\hat{q}_{\mathrm{AA}} = n_{\mathrm{AA}}/N$. When there is deviation from HWE the estimated variance of $\hat{q}_A$ is $\mathrm{var}(\hat{q}_{\mathrm{A}}) = [\hat{q}_{\mathrm{A}}(1 - \hat{q}_{\mathrm{A}}) + (\hat{q}_{\mathrm{AA}} - \hat{q}_{\mathrm{A}}^2)]/(2N)$ [46], which deviates from the binomial variance because of the term $(\hat{q}_{\mathrm{AA}} - \hat{q}_{\mathrm{A}}^2)$. One can use corrected variances and covariances to compute a valid chi-square statistic for comparison of allele frequencies when HWE does not hold, although the usual Pearson chi-square statistic is robust (*see* **Robustness**) to moderate deviations from HWE.

In contrast, if the assumption of HWE is plausible, then analyses can be improved. For example, if the relationship between the marker genotype and phenotype is not one-to-one, then the assumption of HWE can be used to estimate allele frequencies by **maximum likelihood**; for example, using the **EM algorithm**; the HWE proportions give the relative probabilities of the different genotypes that correspond to the same phenotype. Furthermore, risk estimates can be improved with smaller variances [14]; multivariate statistical tests, which account for correlations among alleles from a single locus due to allele frequencies summing to one, as well as correlations among alleles from different loci, can be used as omnibus tests for association [32]; and one can fit genetic models to assess the effects of marker alleles on **relative risks** for disease [21].

When evaluating multiple alleles at a single marker locus, or multiple marker loci, one can use **logistic regression** and **loglinear models** [8, 45] to assess associations. Advantages of these regression models are that nongenetic covariates can be included (which allows estimation of marker relative risks adjusted for potential confounders), **interactions** of marker alleles on relative risk can be evaluated (say, additive vs. dominant effects of alleles on the log relative risk), and interactions between the marker alleles and nongenetic covariates can be assessed (*see* **Gene-environment Interaction**).

The statistical methods used to evaluate associations with marker alleles can also be used to evaluate associations with haplotypes created by multiple marker loci. Haplotypes can be inferred either by family studies, such as for cases, or statistically by using measures of population linkage disequilibrium to predict the most likely haplotypes [16], such as for controls. Some statistical difficulties are ambiguous haplotypes, error in haplotype prediction, ambiguity of disease genotype, and confounding from population structure. It is not unusual for the largest relative risks to occur with the smallest haplotype frequencies, resulting in large variances of risk estimates. Although it may be necessary to combine rare haplotypes to validate use of chi-square statistics, it may be better to use exact or simulated P values [30, 35]. A novel attempt to improve power in this situation is to use a "cladistic" analysis in which the evolutionary history of haplotypes (i.e. evolutionary tree) is first created, and then to perform nested analyses to determine which tree branches differ most between cases and controls [36].

With the availability of many genetic markers, one of the most challenging statistical issues is the choice of the level of statistical significance to maximize power yet minimize the chance of false positives. The success of an association study depends on the likelihood that the marker is involved (directly or indirectly via LD) in the disease process, so that the most meaningful association studies are those that evaluate markers with clear biological functions. When testing many markers, the prior probability that any one is associated with disease is often so low that one needs to be very conservative to avoid false-positive associations. To account for multiple comparisons with many alleles (or many haplotypes), the **Bonferroni** correction is often used, although one should consider the power of this method vs. omnibus multivariate methods. Also ***P* value** plots [28] and **empirical Bayes shrinkage estimates** [40, 41] may prove useful.

Because of the difficulty in defining an appropriate control group for association studies in heterogeneous populations, Falk & Rubenstein [7] proposed to measure the genetic marker on both the diseased cases and their parents in order to compare the frequencies of those alleles that were transmitted from parents to children with those that were not transmitted. For example, to test the association of allele A, vs. all other alleles combined into group B, in a sample of n cases ($2n$ parents), the alleles of each parent are classified as in Table 1, so that each parent contributes a count to this table. The genotype of each parent can be considered a matched pair of alleles, one transmitted and the other not, and the **McNemar test** for matched pairs, which does not require independence of parental alleles, is valid. This is also called the transmission/disequilibrium test, or TDT [34]. Note that only discordant pairs (*see* **Matching**), i.e. heterozygous parents, contribute to the TDT statistic. An alternative approach, which uses all parental alleles, is to ignore the matching in order to compare the frequency of allele A among the $2n$ transmitted alleles vs. the $2n$ nontransmitted alleles. This can be accomplished by rearranging the marginal totals of Table 1 into the cells of Table 2, and applying Pearson's chi-square statistic (also called the haplotype-based haplotype relative risk statistic, HHRR, for this type of analysis [38]). However, this method requires that parental alleles be independent in the population, which is not true for a stratified population [33]. The statistical properties of various methods of analysis for parental controls have been investigated [2, 6, 11, 12, 18–20, 25, 26, 31, 34, 38, 43, 44] but a general framework can be developed based on a conditional likelihood method [29]. To see this, note that by conditioning on the two alleles of the mother, m_1 and m_2, and the two alleles of the father, f_1 and f_2, there are four child genotypes that can be produced, $m_1 f_1$, $m_1 f_2$, $m_2 f_1$, and $m_2 f_2$. One of these four genotypes is that for the diseased case, and the remaining three can be considered matched hypothetical sib controls. This framework allows development of omnibus **score** statistics [24], as well as use of standard conditional logistic regression software (*see* **Software, Biostatistical**) to compute maximum likelihood estimates of allelic effects and to assess interactions between marker genotypes and environmental covariates [13, 23, 29, 39]. It is critical to recognize that these statistical methods are sensitive only to associations caused by both linkage disequilibrium and linkage.

Table 1 Matched analysis for transmitted and nontransmitted parental alleles: TDT $= (b-c)^2/(b+c)$

	Nontransmitted allele		
Transmitted allele	A	B	Total
A	a	b	w
B	c	d	x
Total	y	z	$2n$

Table 2 Nonmatched analysis for transmitted and nontransmitted parental alleles: HHRR $= (wz-xy)^2 \times 4n/[(w+y)(x+z)4n^2]$

	Allele type		
	A	B	Total
Transmitted alleles	w	x	$2n$
Nontransmitted alleles	y	z	$2n$
Total	$w+y$	$x+z$	$4n$

Association studies using pedigree data offer yet another useful design because members of the same pedigree are likely to have the same genetic etiology, which reduces etiologic heterogeneity, and extended pedigrees give more information about the genetic mechanisms underlying the trait than does a sample of unrelated persons. An important feature of pedigree data is that the statistical dependence of members in the same pedigree needs to be incorporated into analyses to obtain accurate estimates of the variances of parameter estimates. For a large number of independent pedigrees, the robust method of **generalized estimating equations** [15, 47] can be used. However, for one or a few pedigrees, the asymptotic results for generalized estimating equations are not likely to hold, so it is necessary to use a statistical model that incorporates familial correlations. For dichotomous traits, one can use either likelihood methods based on combined association, segregation (*see* **Segregation Analysis, Classical**), and linkage (*see* **Linkage Analysis, Model-based**) [17, 22], or a method called the Marker Association Segregation Chi-squares (MASC) [5], which fits models for the simultaneous segregation and association of marker alleles within pedigrees based on minimization of chi-squares. For continuous traits, methods based on the multivariate normal distribution with covariance matrices determined by the genetic relationships among pedigree members can be used [3, 9] (*see* **Genetic Correlations and Covariances**). It can sometimes be advantageous to confirm case-control population association studies with pedigree

studies in order to rule out spurious associations and to understand better the genetic mechanism causing the trait phenotype.

References

[1] Bengtsson, B.O. & Thomson, G. (1981). Measuring the strength of associations between HLA antigens and diseases, *Tissue Antigens* **18**, 356–363.

[2] Bickeböller, H. & Clerget-Darpoux, F. (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers, *Genetic Epidemiology* **12**, 865–870.

[3] Boerwinkle, E., Chakraborty, R. & Sing, C.F. (1986). The use of measured genotype information in the analysis of quantitative phenotypes in man, *Annals of Human Genetics* **50**, 181–194.

[4] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. International Agency for Research on Cancer, Lyon.

[5] Clerget-Darpoux, F. Babron, M.C., Prum, B., Lathrop, G.M., Deschamps, I. & Hors, J. (1988). A new method to test genetic models in HLA associated diseases: the MASC method, *Annals of Human Genetics* **52**, 247–258.

[6] Ewens, W.J. & Spielman, R.S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture, *American Journal of Human Genetics* **57**, 455–464.

[7] Falk, C.T. & Rubinstein, P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations, *Annals of Human Genetics* **51**, 227–233.

[8] Farewell, V.T. & Dahlberg, S. (1984). Some statistical methodology for the analysis of HLA data, *Biometrics* **40**, 547–560.

[9] George, V.T. & Elston, R.C. (1987). Testing the association between polymorphic markers and quantitative traits in pedigrees, *Genetic Epidemiology* **4**, 193–201.

[10] Haldane, J.B.S. (1955). The estimation and significance of the logarithm of a ratio of frequencies, *Annals of Human Genetics* **20**, 309–311.

[11] Jin, K., Speed, T.P., Klitz, W., & Thomson, G. (1994). Testing for segregation distortion in the HLA complex, *Biometrics* **50**, 1189–1198.

[12] Knapp, M., Seuchter, S.A. & Baur, M.P. (1993). The haplotype-relative-risk (HRR) method for analysis of association in nuclear families, *American Journal of Human Genetics* **52**, 1085–1093.

[13] Langholz, B., Tuomilehto-Wolf, E., Thomas, D., Pitkaniemi, J., & Tuomilehto, J. (1994). Variation in HLA-associated risks of childhood insulin dependent diabetes in the Finnish population: I. Allele effects at A, B, and DR Loci, *Genetic Epidemiology* **12**, 441–453.

[14] Lathrop, G.M. (1983). Estimating genotype relative risks, *Tissue Antigens* **22**, 160–166.

[15] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[16] Long, J.C., Williams, R.C. & Urbanek, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes, *American Journal of Human Genetics* **56**, 799–810.

[17] MacLean, C.J., Morton, N.E. & Yee, S. (1984). Combined analysis of genetic segregation and linkage under an oligogenic model, *Computers and Biomedical Research* **17**, 471–480.

[18] Ott, J. (1989). Statistical properties of the haplotype relative risk, *Genetic Epidemiology* **6**, 127–130.

[19] Parsian, A., Todd, R.D., Devor, E.J., O'Malley, K.L., Suarez, B.K., Reich, T. & Cloninger, C.R. (1991). Alcoholism and alleles of the human D_2 dopamine receptor locus, *Archives General Psychiatry* **48**, 655–663.

[20] Rice, J.P., Neuman, R.J., Hoshaw, S.L., Daw, E.W. & Gu, G. (1995). TDT tests with covariates and genome screens with MOD scores: their behavior on simulated data, *Genetic Epidemiology* **12**, 659–664.

[21] Risch, N. (1983). A general model for disease-marker association, *Annals of Human Genetics* **47**, 245–252.

[22] Risch, N. (1984). Segregation analysis incorporating linkage markers. I. Single-locus models with an application to Type I diabetes, *American Journal of Human Genetics* **36**, 363–386.

[23] Schaid, D.J. (1995). Relative-risk regression models using cases and their parents, *Genetic Epidemiology* **12**, 813–818.

[24] Schaid, D.J. (1996). General score tests for associations of genetic markers with disease using cases and their parents, *Genetic Epidemiology* **13**, 423–449.

[25] Schaid, D.J. & Sommer, S.S. (1993). Genotype relative risks: methods for design and analysis of candidate-gene association studies, *American Journal of Human Genetics* **53**, 1114–1126.

[26] Schaid, D.J. & Sommer, S.S. (1994). Comparison of statistics for candidate-gene association studies using cases and parents, *American Journal of Human Genetics* **55**, 402–409.

[27] Schlesselman, J.J. (1982). *Case-Control Studies*. Oxford University Press, New York.

[28] Schweder, T. & Spjotvoll, E. (1982). Plots of P-values to evaluate many tests simultaneously, *Biometrika* **69**, 493–502.

[29] Self, S.G., Longton, G., Kopecky, K.J. & Liang, K.Y. (1991). On estimating HLA/disease association with application to a study of aplastic anemia, *Biometrics* **47**, 53–61.

[30] Sham, P.C. & Curtis, D. (1995). Monte Carlo tests for associations between disease and alleles at highly polymorphic loci, *Annals of Human Genetics* **59**, 97–105.

[31] Sham, P.C. & Curtis, D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allelic marker loci, *Annals of Human Genetics* **59**, 323–336.

[32] Smouse, P.E. & Williams, R.C. (1982). Multivariate analysis of HLA-disease associations, *Biometrics* **38**, 757–768.

[33] Spielman, R.S. & Ewens, W.J. (1996). The TDT and other family-based tests for linkage disequilibrium and association, *American Journal of Human Genetics* **59**, 983–989.

[34] Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics* **52**, 506–516.

[35] StatXact Version 3 (1993). *Software for Exact Nonparametric Inference*. Cytel Software Corporation, Cambridge, Mass.

[36] Templeton, A.R. (1995). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E. locus, *Genetics* **140**, 403–409.

[37] Terwilliger, J.D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci, *American Journal of Human Genetics* **56**, 777–787.

[38] Terwilliger, J.D. & Ott, J. (1992). A haplotype-based "haplotype relative risk" approach to detecting allelic associations, *Human Heredity* **42**, 337–346.

[39] Thomas, D., Pitkaniemi, J., Langholz, B., Tuomilehto-Wolf, E., & Tuomilehto, J. (1994). Variation in HLA-associated risks of childhood insulin dependent diabetes in the Finnish population: II. Haplotype effects, *Genetic Epidemiology* **12**, 455–466.

[40] Thomas, D., Langholz, B., Clayton, D., Pitkaniemi, J., Tuomilehto-Wolf, E. & Tuomilehto, J. (1992). Empirical Bayes methods for testing associations with large numbers of candidate genes in the presence of environmental risk factors, with applications to HLA associations in IDDM, *Annals of Medicine* **24**, 387–392.

[41] Thomas, D.C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M. & Armstrong, B.G. (1985). The problem of multiple inference in studies designed to generate hypotheses, *American Journal of Epidemiology* **122**, 1080–1095.

[42] Thomson, G. (1995). HLA disease associations: models for the study of complex human genetic disorders, *Clinical Reviews in Clinical Laboratory Sciences* **32**, 183–219.

[43] Thomson, G. (1995). Analysis of complex human genetic traits: an ordered-notation method and new tests for model of inheritance, *American Journal of Human Genetics* **57**, 474–486.

[44] Thomson, G. (1995). Mapping disease genes: Family-based association studies, *American Journal of Human Genetics* **57**, 487–498.

[45] Tiret, L., Amouyel, P., Rakotovao, R., Cambian, F. & Ducimetière, P. (1991). Testing for association between disease and linked marker loci: a log-linear-model analysis, *American Journal of Human Genetics* **48**, 926–934.

[46] Weir, B.S. (1990). *Genetic Data Analysis*. Sinauer Associates, Sunderland, p. 34.

[47] Zeger, S.L. & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**, 121–130.

D. SCHAID

Distance Sampling

Distance sampling is the most widely used technique for estimating abundance of biological populations. It is used for a diverse range of populations: whales, dolphins, seals, apes, monkeys, deer, antelope, rabbits, seabirds, gamebirds, songbirds, fish, butterflies, trees, bird nests, animal burrows, animal carcasses, etc. There are several strategies available to wildlife managers and ecologists for assessing abundance [10]. The most obvious is to **census** (count) the population. If this is impracticable, sample counts might be made in randomly selected quadrats. Two forms of "quadrat" sampling are point counts, in which numbers of objects (usually birds or plants) in a circle about a point are counted, and strip transects, in which the observer travels along a line, counting all objects within a predetermined distance of the line. Both methods yield an estimated density (objects per unit area) simply by dividing the total count by the total area surveyed. For many populations, it is difficult to ensure that all objects within the circle or strip are detected and counted. Furthermore, for scarce species, the methods are wasteful, because detections of objects beyond the circle or strip boundary are ignored. If the radius of the circle or the width of the strip is made sufficiently small that detection of any object within the surveyed area is almost certain, then perhaps 50% or more of detections are outside the surveyed area, and so are ignored. In distance sampling, distances from the center line or point to detected objects are recorded, which allows us to estimate object density without having to assume that all objects within the surveyed area are counted.

The term "distance sampling" is used because the population can be regarded as the set of distances of objects from the line or point. We sample from this population of distances in a "size-biased" way, because smaller distances are more likely to be sampled. That is, objects closer to the line or point are more likely to be detected.

The two primary methods of distance sampling are line transect sampling, an extension of strip transect sampling in which line-to-object distances are sampled, and point transect sampling, an extension of point counts in which observer-to-object distances are sampled [4]. Related methods are cue counting [7], used on large whale populations, and trapping webs [1], which extend the applicability of distance sampling to small mammals and ground insect populations. The theory for these approaches is closely similar to that for point transects. Free software Distance [8] for analyzing distance sampling data is available from the web site `http://www.ruwpa.st-and.ac.uk/distance/`. Related techniques sometimes used by botanists to estimate densities (and sometimes also termed distance sampling) are nearest neighbor (*see* **Clustering**) and point-to-nearest object methods [5].

Methods for estimating wildlife abundance that do not involve distance sampling include **capture–recapture**, which is often more labor-intensive and more sensitive to failures of assumptions than distance sampling. However, it is applicable to some species that are not amenable to distance sampling methods, and can yield estimates of survival and recruitment rates, which distance sampling cannot do. Capture–recapture methods can be useful for populations that aggregate at some location each year, whereas distance sampling methods are more effective on dispersed populations. They should therefore be seen as different tools for different purposes. In fisheries applications, catch per unit effort, catch-at-age and catch-at-length are all commonly used to estimate abundance, as they require that the commercial catch is sampled, which is more cost-effective than sampling the living fish. Acoustic surveys of fish schools often provide data amenable to distance sampling methods. For difficult terrestrial species, abundance is often indexed using indirect methods, such as dung counts for deer or rabbits. To convert the index to an estimate of abundance, typically one or more rates must be estimated, such as deposition rate and decay rate for "standing crop" dung counts.

Line Transect Sampling

In line transect sampling, a series of straight lines is traversed by an observer. This may be achieved in various ways, depending on the study species. In terrestrial studies, these include walking, horseback, trail bike, all-terrain vehicle, fixed-wing aircraft, and helicopter. Transect surveys for aquatic environments can be conducted by divers with snorkels or scuba gear, from surface vessels ranging in size from small boats to large ships, from fixed-wing aircraft, helicopters or airships, from small submarines, or from sleds with mounted video units pulled underwater by a surface

vessel. In the case of large observation platforms, there is typically a team of observers.

Estimation

Perpendicular distances x are measured from the line to each detected object of interest (usually an animal, or "cluster" of animals). In practice, detection distances r and detection angles θ are often recorded, from which perpendicular distances are calculated as $x = r \sin\theta$ (Figure 1). Suppose k lines of lengths $l_1, \ldots, l_k$ (with $\sum l_j = L$) are positioned according to some randomized scheme, and n animals are detected, at perpendicular distances $x_1, \ldots, x_n$. Suppose that animals farther than some distance w from the line are not recorded. Then the surveyed area is $a = 2wL$, within which n animals are detected. However, not all animals within the surveyed area are detected. Let P_a be the probability that an animal within the surveyed area is detected, and suppose an estimate $\hat{P}_a$ is available. Then animal density D is estimated by

$$\hat{D} = \frac{n}{2wLP_a}. \tag{1}$$

To provide a framework for estimating P_a, we define the detection function $g(x)$ to be the probability that an animal at distance x from the line is detected, $0 \le x \le w$, and assume that $g(0) = 1$. That is, we are certain to detect an animal on the trackline. If we plot the recorded perpendicular distances in a histogram, then conceptually the problem is reduced to specifying a suitable model for $g(x)$, and fitting it to the perpendicular distance data. As shown in

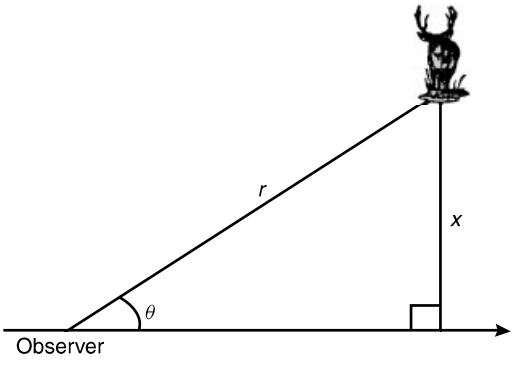

Figure 1 If sighting distance r and sighting angle θ are recorded, then perpendicular distance x of the animal from the line is found as $x = r \sin\theta$

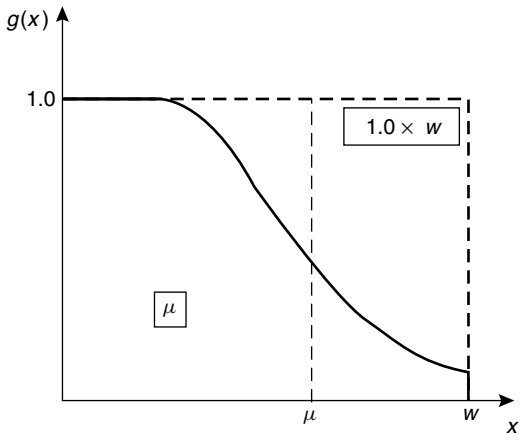

Figure 2 The area μ under the detection function $g(x)$, when expressed as a proportion of the area w of the rectangle, is the probability that an object within the surveyed area is detected; μ is also the effective strip width, and takes a value between 0 and w

Figure 2, if we define $\mu = \int_0^w g(x)\,dx$, then $P_a = \mu/w$. The parameter μ is called the effective strip (half-) width; it is the distance from the line for which as many animals are detected beyond μ as are missed within μ (Figure 2). Thus

$$\hat{D} = \frac{n}{a \times \hat{P}_a} = \frac{n}{2wL \times \hat{\mu}/w} = \frac{n}{2\hat{\mu}L}. \tag{2}$$

We now need an estimate $\hat{\mu}$ of μ. We can turn this into a more familiar estimation problem by noting that the probability density function of perpendicular distances to detected objects, denoted $f(x)$, is simply the detection function $g(x)$, rescaled so that it integrates to unity. That is, $f(x) = g(x)/\mu$. In particular, because we assume $g(0) = 1$, it follows that $f(0) = 1/\mu$ (Figure 3). Hence

$$\hat{D} = \frac{n}{2\hat{\mu}L} = \frac{n\hat{f}(0)}{2L}. \tag{3}$$

The problem is reduced to modeling the probability density function of perpendicular distances, and evaluating the fitted function at $x = 0$. The large literature for fitting density functions is now available to us. Distance uses the methods of Buckland [3], in which a parametric "key" function is selected and, if it fails to provide an adequate fit, polynomial or cosine series adjustments are added until the fit is judged to be satisfactory by one or more criteria.

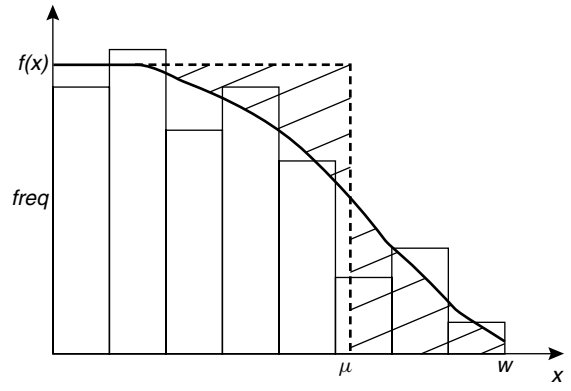

Figure 3 The probability density function of perpendicular distances, $f(x)$, plotted on a histogram of perpendicular distance frequencies (scaled so that the total area of histogram bars is unity). The area below the curve is unity by definition. Because the two shaded areas are equal in size, the area of the rectangle, $\mu \times f(0)$, is also unity. Hence $\mu = 1/f(0)$

Often, the perpendicular distances are recorded by distance category, so that each exact distance need not be measured, or data are grouped into distance categories before analysis. Standard **likelihood** methods for **multinomial** data are used to fit such "grouped" data.

Variance and Interval Estimation

The **variance** of $\hat{D}$ may be approximated using the **delta method**, assuming no **correlation** between n and $\hat{f}(0)$:

$$\hat{V}(\hat{D}) = \hat{D}^2 \left[\frac{\hat{V}(n)}{n^2} + \frac{\hat{V}[\hat{f}(0)]}{[\hat{f}(0)]^2} \right]. \qquad (4)$$

The variance of n is generally estimated from the sample variance in encounter rates, n_j/l_j, weighted by line lengths l_j. When $f(0)$ is estimated by **maximum likelihood**, its variance is estimated from the **information matrix**.

If we assume that $\hat{D}$ is **lognormally distributed**, approximate 95% **confidence limits** are given by $(\hat{D}/C, \hat{D} \times C)$ where

$$C = \exp\{1.96[\hat{V}(\log_e \hat{D})]^{0.5}\}, \qquad (5)$$

with

$$\hat{V}(\log_e \hat{D}) = \log_e \left[1 + \frac{\hat{V}(\hat{D})}{\hat{D}^2} \right]. \qquad (6)$$

Often, **bootstrap** variance and interval estimation (*see* **Estimation, Interval**) is preferred. Resamples are usually generated by **sampling with replacement** from the lines, so that independence between the lines is assumed, but independence between detections on the same line is not. If the model selection procedure is applied independently to each resample, the bootstrap variance includes a component due to model selection uncertainty.

Cluster Size Estimation

Animals often occur in groups, which we term "clusters". These may be flocks of birds, pods of whales, schools of fish, herds of antelope, etc. If one animal in a cluster is detected, then it is assumed that the whole cluster is detected, and the position of the cluster is recorded. Eq. (3) then gives an estimate of the density of clusters. To obtain the estimated density of individuals, we must multiply by an estimate of mean cluster size in the population, $E(s)$:

$$\hat{D} = \frac{n\hat{f}(0)\hat{E}(s)}{2L}. \qquad (7)$$

Probability of detection is often a function of cluster size, so that the sample of cluster sizes exhibits size bias. In the absence of size bias, we can take $\hat{E}(s) = \overline{s}$, the mean size of detected clusters. Several methods exist for estimating $E(s)$ in the presence of size bias [4]. One that works well in practice is to regress $\log s$ on $\hat{g}(x)$, the estimated probability of detection at distance x ignoring the effect of cluster size, and then predict $\log s$ when detection is certain, $\hat{g}(x) = 1$, as there can be no size bias in that circumstance. The prediction is back-transformed using a bias adjustment.

Assumptions

The physical setting for line transect sampling is idealized as below:

1. N objects are distributed through an area of size A according to some **stochastic process** with average rate parameter $D = N/A$.
2. Lines, placed according to some randomized design (*see* **Randomization**), are surveyed and a sample of n objects is detected.

It is not necessary that the objects be randomly (i.e. **Poisson**) distributed. Rather, it is critical that the line

or point be placed randomly with respect to the local distribution of objects. This ensures that objects in the surveyed strip are uniformly distributed with distance from the line. Thus if the strip has half-width w, animal-to-line distances available for detection are **uniformly distributed** between zero and w.

Three assumptions are essential for reliable estimation of density using standard line transect methods:

1. Objects directly on the line are always detected, $g(0) = 1$.
2. Objects are detected at their initial location, prior to any movement in response to the observer.
3. Distances are measured accurately (for ungrouped distance data), or objects are correctly allocated to distance interval (for grouped data).

A fourth assumption is made in many derivations of estimators and variances: whether an object is detected is independent of whether any other object is detected. Point estimates are **robust** to the assumption of independence, and robust variance estimates are obtained by taking the line to be the sampling unit, either by bootstrapping on lines, or by calculating a weighted sample variance of encounter rates by line.

It is also important that the detection function has a "shoulder"; that is, probability of detection remains at or close to one initially as distance from the line increases from zero. This is not an assumption, but a property that allows more reliable estimation of object density.

Point Transect Sampling

In point transect sampling, an observer visits a number of points, the locations of which are determined by some randomized design. The method is usually (but not exclusively) used for songbird populations, in which typically many species are recorded, and most detections are aural. By recording from points, the observer can concentrate on detecting the objects of interest, without having to navigate along a line, and without having to negotiate a randomly positioned line through possibly difficult terrain. The principal disadvantages are that detections made while traveling from one point to the next are not utilized, a problem especially for scarce species, and the method is unsuited to species that are generally detected by flushing them, or to species that typically change their location appreciably over the time period of a count (generally around 3–10 minutes).

Estimation

Detection distances r are measured from the point to each detected object of interest (usually a bird or a small group of birds). Suppose the design comprises k points, and distances $\leq w$ are recorded. Then the surveyed area is $a = k\pi w^2$, within which n objects are detected. As for line transect sampling, denote the probability that an object within the surveyed area is detected by P_a with estimate $\hat{P}_a$. Then we estimate object density D by

$$\hat{D} = \frac{n}{k\pi w^2 \hat{P}_a}. \tag{8}$$

We now define the detection function $g(r)$ to be the probability that an object at distance r from the point is detected, and we again assume that $g(0) = 1$. For line transects, the area of an incremental strip at distance x from the lines is $L\,\mathrm{d}x$, independent of x, which leads to the result that the probability density function of distances differs from the detection function only in scale. By contrast, an incremental annulus at distance r from a point has area $2\pi r\,\mathrm{d}r$, proportional to r, so that the probability density function of detection distances is $f(r) = 2\pi r g(r)/\nu$, where $\nu = 2\pi \int_0^w r g(r)\,\mathrm{d}r$. The respective shapes of the two functions are illustrated in Figure 4. If we define an effective radius ρ, analogous to the effective strip width of line transect sampling, then $\nu = \pi\rho^2$ is the effective area surveyed per point (Figure 5). Hence

$$\hat{D} = \frac{n}{a \times \hat{P}_a} = \frac{n}{k\pi w^2 \times \pi\hat{\rho}^2/\pi w^2} = \frac{n}{k\hat{\nu}}. \tag{9}$$

The area of the triangle in Figure 5 is $\rho^2 f'(0)/2$ where $f'(0)$ is the slope of $f(r)$ at $r = 0$. Since this is equal to the area under $f(r)$, which is unity, it follows that $\nu = \pi\rho^2 = 2\pi/f'(0)$, and

$$\hat{D} = \frac{n\hat{f}'(0)}{2\pi k}. \tag{10}$$

We therefore need to model the probability density function of detection distances, and evaluate the slope of the fitted function at $r = 0$. Distance does this using the same set of models for the detection function as for line transect sampling.

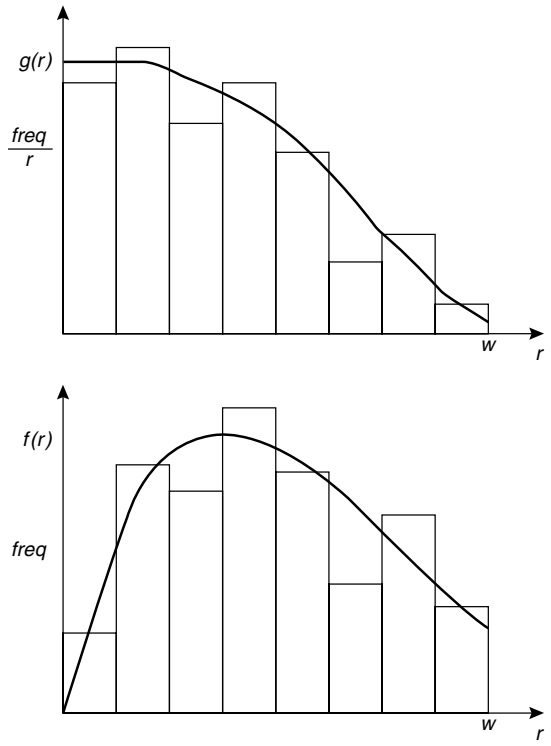

Figure 4 Histograms of detection distances from a point transect survey. In the upper plot, each histogram frequency has been scaled by dividing by the midpoint of the corresponding group interval. Also shown are the corresponding fits of the detection function [$g(r)$, upper plot] and the probability density function of detection distances [$f(r)$]

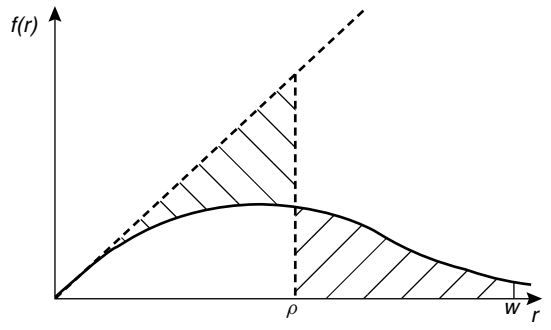

Figure 5 The probability density function of detection distances, $f(r)$. The area under the curve is unity by definition. Because the two shaded areas are equal in size, the area of the triangle, $\rho^2 \times f'(0)/2$, is also unity. Hence $\nu = \pi\rho^2 = 2\pi/f'(0)$

Variance and Interval Estimation

The methods for variance and interval estimation for line transect sampling apply also to point transects with minor modifications. In the case of the bootstrap, resampling is normally carried out by sampling with replacement from the points. However, point transect surveys are often designed by defining a series of lines, as if a line transect survey is to be carried out, then locating a series of points along each line. If the distance between neighboring points on the same line is smaller than the distance between neighboring points on different lines, then resampling should be carried out by sampling lines with replacement. If a line is selected, then all the points associated with that line are included in the bootstrap resample.

Assumptions

Assumptions are virtually unchanged from those given for line transect sampling. As there, the standard analyses are very robust to failure of the assumption of independent detections, but if objects occur in clusters, so that when one of the cluster is detected they all are, then the cluster is generally taken to be the object, and mean cluster size in the population is estimated using the same techniques as for line transect sampling. Point transect sampling is more subject to bias than line transect sampling when objects move through the area around a point. In principle, we try to obtain a snapshot, locating each object at the position it occupied when the count at that point started. However, the count is not instantaneous, because the observer needs time to detect all objects close to that point. If, during that time, movement brings new objects into the neighborhood of the point, then object density will be overestimated.

Current Research

Double platform methods are becoming commonplace in sightings surveys for whales. Observers search simultaneously from two platforms. This allows extension of the standard methods to the case that $g(0) < 1$, and also, given appropriate field methods, allows adjustment for responsive movement of animals prior to detection. There have been advances by several researchers recently in developing methodology for analyzing such data. Perhaps the most promising approach, based on **Horvitz–Thompson-type estimators** [2], has also

led to a unified approach to the analysis of line transect data, in which object density is estimated in a single step, which contrasts with the conventional strategy of independently estimating the three parameters: encounter rate, effective strip width, and mean cluster size.

Generally, probability of detection is a function of many factors other than distance of the object from the line or point. We have considered briefly one other factor, cluster size, because if we do not allow for size bias in detection, then our object density estimator will be biased. Other sources of heterogeneity contribute little to bias, provided $g(0) = 1$, but nevertheless, higher precision might be anticipated if additional **covariates** are recorded and their effects on $g(x)$ modeled. Ramsey et al. [9] modeled effective area surveyed as a function of covariates using **generalized linear modeling** methods. Again, a Horvitz–Thompson formulation provides a natural framework for estimation, if probability of detection is modeled as a function of relevant covariates.

The global spatial coordinates of detections are often recorded in distance sampling surveys so that spatial modeling of density surfaces is possible. Very little has been done on this problem (but see [6]); rather, current practice is simply to estimate average density by strata.

References

[1] Anderson, D.R., Burnham, K.P., White, G.C. & Otis, D.L. (1983). Density estimation of small-mammal populations using a trapping web and distance sampling methods, *Ecology* **64**, 674–680.

[2] Borchers, D.L. (1996). *Estimating abundance from line transect surveys when detection on the trackline is not certain*. PhD Thesis, University of Cape Town.

[3] Buckland, S.T. (1992). Fitting density functions using polynomials, *Applied Statistics* **41**, 63–76.

[4] Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L., (2001). *Introduction to Distance Sampling*. Oxford University Press, Oxford. (This book contains a bibliography of around 700 references on distance sampling.)

[5] Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

[6] Hedley, S.L., Buckland, S.T. & Borchers, D.L. (1999). Spatial model from line transect data, *Journal of Cetacean Research and Management* **1**, 255–264.

[7] Hiby, A.R. (1985). An approach to estimating population densities of great whales from sighting surveys, *IMA Journal of Mathematics Applied in Medicine and Biology* **2**, 201–220.

[8] Laake, J.L., Buckland, S.T., Anderson, D.R. & Burnham, K.P. (1993). *DISTANCE User's Guide, Version 2.0*. Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins.

[9] Ramsey, F.L., Wildman, V. & Engbring, J. (1987). Covariate adjustments to effective area in variable-area wildlife surveys, *Biometrics* **43**, 1–11.

[10] Seber, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*. Macmillan, New York.

STEPHEN T. BUCKLAND, DAVID R. ANDERSON, KENNETH P. BURNHAM & JEFFREY L. LAAKE

Distribution-free Methods for Longitudinal Data

Repeated measure designs are used in many areas of application and are especially common in biomedical research. The characteristic feature of such studies is that multiple measurements of a response variable are obtained from each independent experimental unit. The repeated measures may be obtained at a set of scheduled time points for each subject or experimental unit. In other applications the response from each experimental unit is measured under multiple conditions, rather than at multiple time points.

There are two main difficulties in the analysis of data from repeated measure designs. First, the analysis is complicated by the dependence among repeated observations made on the same experimental unit. Secondly, the investigator often cannot control the circumstances for obtaining measurements, so that the data may be unbalanced or partially incomplete.

Many approaches to the analysis of data from repeated measure designs have been studied; see, for example, the 1980 review and bibliography of parametric and nonparametric approaches by Koch et al. [28]. When the response variable is normally distributed, classical **multivariate analysis** techniques, repeated measures analysis of variance (*see* **Analysis of Variance for Longitudinal Data**), growth curve analysis (*see* **Nonlinear Growth Curve**), and mixed effects models can be used. The development of methods for the analysis of repeated measures categorical data, for binary, polytomous, and ordered categorical response variables, is also an important area of research (*see*, in particular, **Multivariate Methods for Binary Longitudinal Data**). In addition, recently developed **generalized estimating equations** approaches based on extensions of **generalized linear model** methodology can be applied to a wide variety of types of continuous and categorical response variables with marginal (univariate) distributions from the class of generalized linear models.

While all of the above methods require assumptions on either the joint or the marginal distributions of the response variable, there are at least three situations in which distribution-free methods may be useful (*see* **Nonparametric Methods**). First, when the response is continuous, the assumption of **multivariate normality** may not be reasonable, or the underlying distribution may be unknown. In this case, the use of standard parametric procedures may not be justified. Secondly, when the response is an ordered categorical variable with a large number of possible outcomes, the general categorical data methods may be inapplicable owing to sample size limitations. In addition, the restrictive proportional-odds assumption underlying some of the approaches for analyzing ordered categorical repeated measures may not be justified. Apart from these considerations, there are also situations in which it may be desirable to confirm the results of a parametric analysis using distribution-free methods.

Table 1 displays the general layout and notation for a repeated measures design with n subjects (experimental units) and t_i measurement times for the ith subject, $i = 1, \ldots, n$. The response from subject i at time j is y_{ij} and $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ is the corresponding $p \times 1$ vector of covariates. In general, the covariates can be a mixture of time-independent (between-subject) covariates and time-dependent (within-subject) covariates. Since values of the response variable and/or covariates might be

Table 1 Layout and notation for a general repeated measures design

Subject	Time point	Missing indicator	Response	Covariates		
1	1	δ_{11}	y_{11}	x_{111}	$\ldots$	x_{11_p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	δ_{1j}	y_{1j}	x_{1j1}	$\ldots$	x_{1jp}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_1	δ_{1t_1}	y_{1t_1}	x_{1t_11}	$\ldots$	x_{1t_1p}
$\ldots$						
i	1	δ_{i1}	y_{i1}	x_{i11}	$\ldots$	x_{i1_p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	δ_{ij}	y_{ij}	x_{ij1}	$\ldots$	x_{ijp}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_i	δ_{it_i}	y_{it_i}	x_{it_i1}	$\ldots$	x_{it_ip}
$\ldots$						
n	1	δ_{n1}	y_{n1}	x_{n11}	$\ldots$	x_{n1_p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	δ_{nj}	y_{nj}	x_{nj1}	$\ldots$	x_{njp}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_n	δ_{nt_n}	y_{nt_n}	x_{nt_n1}	$\ldots$	x_{nt_np}

missing, it may be convenient to define indicator variables δ_{ij} by

$$\delta_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ and } \mathbf{x}_{ij} \text{ are observed,} \\ 0, & \text{otherwise.} \end{cases}$$

One simplification of the general notation displayed in Table 1 is the case when every subject has the same fixed set of measurement times, so that $t_i = t$, for $i = 1, \ldots, n$. The situation is also simplified when there are no missing data. A special case which has been studied extensively from the distribution-free perspective is the multisample setting in which repeated measurements are obtained from samples from s subpopulations. The s groups may be defined by the s levels of a single covariate or by the cross-classification of several discrete covariates. In terms of the general notation, the s groups can thus be described in terms of $p = s - 1$ dichotomous, time-independent covariates. In this setting, the notation of Table 2 is useful, in which y_{hij} denotes the response at time j from subject i in group h, for $j = 1, \ldots, t$, $i = 1, \ldots, n_h$, and $h = 1, \ldots, s$.

In the case of repeated measurements obtained at t time points from each of n subjects from a single sample, the data can be displayed even more simply in an $n \times t$ matrix, as shown in Table 3. As before, missing value indicators can be defined by

$$\delta_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

Table 2 Layout and notation for a multisample repeated measures design

		Time point				
Group	Subject	1	...	j	...	t
1	1	y_{111}	...	y_{11j}	...	y_{11t}
	⋮	⋮	⋱	⋮	⋱	⋮
	i	y_{1i1}	...	y_{1ij}	...	y_{1it}
	⋮	⋮	⋱	⋮	⋱	⋮
	n_1	y_{1n_11}	...	y_{1n_1j}	...	y_{1n_1t}
h	1	y_{h11}	...	y_{h1j}	...	y_{h1t}
	⋮	⋮	⋱	⋮	⋱	⋮
	i	y_{hi1}	...	y_{hij}	...	y_{hit}
	⋮	⋮	⋱	⋮	⋱	⋮
	n_h	y_{hn_h1}	...	y_{hn_hj}	...	y_{hn_ht}
s	1	y_{s11}	...	y_{s1j}	...	y_{s1t}
	⋮	⋮	⋱	⋮	⋱	⋮
	i	y_{si1}	...	y_{sij}	...	y_{sit}
	⋮	⋮	⋱	⋮	⋱	⋮
	n_s	y_{sn_s1}	...	y_{sn_sj}	...	y_{sn_st}

Table 3 Layout and notation for a one-sample repeated measures design

	Time point				
Subject	1	...	j	...	t
1	y_{11}	...	y_{1j}	...	y_{1t}
⋮	⋮	⋱	⋮	⋱	⋮
i	y_{i1}	...	y_{ij}	...	y_{it}
⋮	⋮	⋱	⋮	⋱	⋮
n	y_{n1}	...	y_{nj}	...	y_{nt}

Univariate Methods

The simplest approach to repeated measures is to reduce the vector of responses from each subject or experimental unit to a single measurement. This avoids the issue of **serial correlation** among the repeated measures for each subject. Several authors [39, 35, 20, 16] refer to these types of methods as the "summary statistic approach" (*see* **Summary Measures Analysis of Longitudinal Data**). Crowder & Hand [11] and Diggle et al. [17] call such methods "response feature analysis" and "derived variable analysis", respectively. The univariate function of the repeated measures from each subject can then be analyzed using distribution-free methods.

For example, in the one-sample setting (Table 3), interest may focus on assessing the extent of association between the response variable and the repeated measures factor. If the Spearman **rank correlation** coefficient between the response variable and the repeated measures variable is used as the summary statistic for each subject, then the **sign test** or the **Wilcoxon signed-rank test** can be used to test if the median of the distribution of the summary statistic is equal to zero. In the multisample setting (Table 2), similar methods can be used. For example, the Mann–Whitney–Wilcoxon (if $s = 2$) or Kruskal–Wallis ($s > 2$) test (*see* **Nonparametric Methods**) can be used to assess if the distribution of the summary statistic is the same across the s groups.

While the summary statistic approach can be useful in certain situations, a shortcoming is that the results may be misleading if the selected univariate summary measure does not adequately describe each subject's data. Ghosh et al. [21] describe multivariate nonparametric methods based on the use of two or more summary statistics for each subject. This extension of the univariate summary statistic approach may be useful when multiple univariate statistics are necessary to adequately summarize each subject's data. Carr et al. [9] describe a different type of multivariate approach based on summary statistics. They consider the situation in which an ordered categorical or interval response variable is measured at multiple time points for each subject in two or more ordered groups. Rank measures of association between group and response are constructed at each time point; the estimated **covariance matrix** of these summary measures is then used to test hypotheses concerning the rank measures of association.

Multivariate Generalizations of Univariate Distribution-Free Methods

Standard asymptotically distribution-free tests for multivariate one-sample and multisample problems can also be used in the repeated measures setting. These rank-based methods are appropriate for samples from continuous multivariate distributions.

For the one-sample case with complete data, Hettmansperger [23, Chapter 6] and Puri & Sen [41, Chapter 4] study multivariate generalizations of the sign and Wilcoxon signed rank tests (*see* **Multivariate Median and Rank Sum Tests**). In the repeated measures setting of Table 3 with no missing data, let θ_t denote the median of the marginal distribution of the response at time t. By transforming each of the n t-component vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})'$ to a $(t-1)$-component vector of differences $\mathbf{y}_i^* = (y_{i1} - y_{i2}, \ldots, y_{i,t-1} - y_{it})'$, these methods can then be used to test the null hypothesis that $\theta_1 = \cdots = \theta_t$.

Hettmansperger [23] also considers the two-sample situation with complete data; the test statistic is a multivariate version of the Wilcoxon–Mann–Whitney test. Puri & Sen [41, Chapter 5] discuss multivariate generalizations of the Kruskal–Wallis [29] and Brown–Mood [8] tests for the multivariate multisample situation with complete data. On the basis of these results, Schwertman [43] gives a computer algorithm for two of these tests, the multivariate multisample rank test and the multivariate multisample median test. These methods can be applied to the repeated measures setting of Table 2. Let $F_h(\mathbf{u})$ denote the t-variate cumulative distribution function (cdf) in group h, for $h = 1, \ldots, s$, where $\mathbf{u} = (u_1, \ldots, u_t)'$. Assume that the cdfs F_h have a common unspecified form with possible differences in their location (or scale) parameters (*see* **Location–Scale Family**). For example, suppose that $F_h(\mathbf{u}) = F(\mathbf{u} + \boldsymbol{\Delta}_h)$, where $\boldsymbol{\Delta}_h = (\Delta_{h1}, \ldots, \Delta_{ht})'$. The null hypothesis of no difference among groups across all time points tests $\mathrm{H}_0 : \boldsymbol{\Delta}_1 = \ldots, \boldsymbol{\Delta}_s = (0, \ldots, 0)'$. The omnibus alternative hypothesis is that $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_s$ are not all equal. Schwertman [44] describes this approach in further detail and gives an example of its application to the analysis of repeated measurements.

Randomization Tests

In the one-sample repeated measures setting (Table 3), a **randomization test** based on the use of **Mantel–Haenszel methods** can be used to test the null hypothesis of no association between a repeated measurement factor and a response variable, adjusting for the effect of subject. The randomization model approach applies to categorical or continuous outcome variables y_{ij}, requires no distributional assumptions, and is useful in small samples. Landis et al. [32] give a general overview of the three types of Mantel–Haenszel statistics; Landis et al. [33] and Crowder & Hand [11, Section 8.6] describe the use of these procedures in analyzing repeated measures.

The basic idea underlying the Mantel–Haenszel randomization model approach to one-sample repeated measures is to restructure the $n \times t$ data matrix of Table 3 as follows. First, let c denote the number of distinct values of the response y_{ij}. If the response variable is categorical with a limited number of possible values, then c will be relatively small. At the other extreme, if each of the n subjects has a unique response at each time point, then $c = nt$. Now define indicator variables

$$n_{ijk} = \begin{cases} 1, & \text{if subject } i \text{ is classified in response category } k \text{ at time } j, \\ 0, & \text{otherwise,} \end{cases}$$

Table 4 Contingency table layout for subject i in the one-sample repeated measures design

	Response category			
Time point	1	...	c	Total
1	n_{i11}	...	n_{i1c}	n_{i1+}
⋮	⋮	⋱	⋮	⋮
t	n_{it1}	...	n_{itc}	n_{it+}
Total	n_{i+1}	...	n_{i+c}	n_i

for $i = 1, \ldots, n$, $j = 1, \ldots, t$, and $k = 1, \ldots, c$. The data from subject i can then be displayed in a $t \times c$ **contingency table**, as shown in Table 4. Thus, the data from a one-sample repeated measures study can be viewed as a set of n independent $t \times c$ contingency tables.

When the data are complete, i.e. the outcome variable is measured at every time point for each subject, the total sample size for each of the n tables is $n_i = t$ and every row marginal total n_{ij+} is equal to one. In this case, each row of Table 4 has exactly one n_{ijk} value equal to one and the remaining values are equal to zero. If, however, a particular subject has a missing response at one or more time points, then the corresponding row of the subject's table will have each n_{ijk} value, as well as the marginal total n_{ij+}, equal to zero. The total sample size n_i will then equal t minus the number of missing observations.

In the framework of Table 4, Mantel–Haenszel statistics can be used to test the null hypothesis of no association between the row dimension (time) and the column dimension (response), adjusted for subject. Under the assumption that the marginal totals $\{n_{ij+}\}$ and $\{n_{i+k}\}$ of each table are fixed, the null hypothesis is that, for each subject, the response variable is distributed at random with respect to the t time points. As discussed in Landis et al. [32], this null hypothesis is precisely the interchangeability hypothesis of Madansky [34]. In turn, the hypothesis of interchangeability implies marginal homogeneity in the distribution of the response variable across the t time points. Although the interchangeability hypothesis is a somewhat stronger condition than marginal homogeneity, the Mantel–Haenszel general association statistic [with $(t-1)(c-1)$df], mean score statistic (with $t-1$df), and correlation statistic (with 1 df) are directed at alternatives that correspond to various types of departures from marginal homogeneity.

Several common nonparametric test procedures are special cases of Mantel–Haenszel randomization model tests. These include the tests of Friedman [19], Durbin [18], Benard & van Elteren [6], and Page [37], as well as the aligned ranks test introduced by Hodges & Lehmann [24] and further studied by Koch & Sen [27]. Randomization tests for other types of repeated measures situations have also been studied. For example, Zerbe & Walker [52] and Zerbe [50, 51] developed randomization tests for the multisample situation of Table 2. These procedures can be used to test the equality of s mean growth curves over a specified time interval.

Other Methods

Asymptotically distribution-free analogs of parametric procedures for normally distributed outcomes have also been studied. Bhapkar [7] discusses nonparametric counterparts of **Hotelling's T^2** statistic and profile analysis (*see* **Longitudinal Data Analysis, Overview**). Sen [45] studies nonparametric analogs of the Potthoff & Roy [40] growth curve model.

Distribution-free methods for the two-sample case (Table 2 with $s = 2$) when the data are incomplete were studied by Wei & Lachin [49] and Wei & Johnson [48]. These approaches allow the missing value patterns in the two samples to be different, but require the assumption that the missing value mechanism is independent of the response. Wei & Lachin [49] study a family of asymptotically distribution-free tests for equality of two multivariate distributions. Although their methodology was motivated and developed for multivariate censored failure time data, an important application is to repeated measures with missing observations. The Wei–Lachin methodology is based on a random-censorship model and they focus on an omnibus test of equality vs. a general alternative. In contrast, Wei & Johnson [48] focus primarily on optimal methods of combining dependent tests and propose a class of two-sample nonparametric tests for incomplete repeated measures based on two-sample **U statistics**. Davis [12, 13] provides further discussion of these methods and a computer program. Lachin [31] proposes additional test statistics and provides estimators of the treatment difference, Palesch & Lachin [38] extend these methods to more than two groups, and Thall & Lachin [46], Davis & Wei [15], and

Davis [14] study related methods for special types of situations with incomplete data.

Another potential approach to the analysis of repeated measures when the underlying parametric assumptions are not satisfied is the **rank transform** method, which consists of replacing observations by their ranks and performing a standard parametric analysis on the ranks [10]. Unfortunately, the rank transform method has been shown to be inappropriate for many common hypotheses [2, 3]. Thompson [47] and Akritas & Arnold [4] provide valid asymptotic tests based on the rank transform for selected hypotheses of interest in several repeated measures models. Kepner & Robinson [25] consider the one-sample situation of Table 3 under the assumption that the repeated measurements y_{ij} from the ith subject are equally correlated. They show the relationships between the rank transform method and the rank tests of Agresti & Pendergast [1] and Koch [26] for testing the null hypothesis of no time effect, expressed as $H_0 : F(x_1, \ldots, x_t) = F(x_{\alpha(1)}, \ldots, x_{\alpha(t)})$, where $F(x_1, \ldots, x_t)$ is the t-variate distribution of the data vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $[\alpha(1), \ldots, \alpha(t)]$ is any permutation of the first t positive integers.

Müller [36], Diggle et al. [17, Chapter 3], and Kshirsagar & Smith [30, Chapter 10] discuss **nonparametric regression** methods for the analysis of repeated measurements, including kernel estimation, weighted local **least squares** estimation, and smoothing splines. Hart & Wehrly [22] study the theoretical properties of kernel regression estimation for repeated measures and show how the case of correlated errors changes the behavior of a kernel estimator; Altman [5] demonstrates that the standard techniques for bandwidth selection perform poorly when the errors are correlated. Raz [42] describes an analysis procedure for repeated measurements that combines nonparametric regression methods and the randomization tests of Zerbe [50].

Examples

Repeated Measures from a Single Population

As part of a protocol for the University of Iowa Mental Health Clinical Research Center, 44 schizophrenic patients participated in a four-week antipsychotic medication washout. The severity of extrapyramidal side effects was assessed just prior to discontinuation of antipsychotic medication and at weeks 1, 2, 3, and 4 during the washout period (*see* **Psychometrics, Overview**). Table 5 displays the resulting ratings on the Simpson Angus (SA) scale; a few missing values are denoted by a dash. The marginal distributions of the scores are clearly nonnormal.

Table 5 Simpson Angus ratings for 44 schizophrenic patients

Patient	Week 0	Week 1	Week 2	Week 3	Week 4
1	1	4	0	0	0
2	4	5	8	9	3
3	1	2	2	1	1
4	8	7	0	5	5
5	1	1	0	1	1
6	3	2	0	0	0
7	4	4	4	–	2
8	–	–	1	9	6
9	6	6	0	0	0
10	3	3	0	0	0
11	6	4	1	0	0
12	0	0	0	0	–
13	3	0	17	5	22
14	8	1	2	2	0
15	0	0	0	0	0
16	0	0	5	1	2
17	1	5	4	5	2
18	2	1	–	–	–
19	0	0	0	0	0
20	0	0	6	8	5
21	0	0	0	0	–
22	11	12	0	0	0
23	10	6	0	0	1
24	3	0	2	1	1
25	1	0	1	1	0
26	0	5	0	2	4
27	0	0	0	–	–
28	3	0	0	0	–
29	7	7	3	4	5
30	12	22	15	24	5
31	3	0	0	0	0
32	0	0	0	0	0
33	1	0	0	0	0
34	0	0	0	0	0
35	7	1	10	7	5
36	2	0	0	1	0
37	10	5	5	8	2
38	2	0	4	0	1
39	5	2	1	3	2
40	0	0	0	–	–
41	1	1	0	1	3
42	0	0	0	0	–
43	0	0	0	0	0
44	1	0	2	1	1

The summary statistic approach is one possible method of testing if there is an association between SA ratings and measurement week. When the Spearman rank correlation coefficient between SA rating and week is computed for each subject, the correlation coefficients range from -1 to 0.8. Of the 32 nonzero correlations, eight are positive and 24 are negative. On the basis of the sign test, the exact two-sided P value is 0.007. Using the Wilcoxon signed rank test, the sum of the ranks corresponding to positive correlations is 103 and the sum of the ranks of negative correlations is 425. The normal approximation to the distribution of the Wilcoxon statistic yields $P = 0.003$. Both tests indicate the tendency of scores to decrease over time.

The randomization model approach using the Mantel–Haenszel mean score and correlation statistics is also applicable. Using within-subject rank scores for the SA rating, the Mantel–Haenszel mean score χ^2 statistic is 13.674 with 4 df ($P = 0.008$); thus, there is substantial evidence that the distributions are not the same at the five measurement times. Using rank scores for the SA rating and the scores 0, 1, 2, 3, and 4 for the five measurement times, the Mantel–Haenszel correlation statistic is 10.375 with 1 df ($P = 0.001$). This result indicates that there is a consistent monotonic association between SA rating and week across subjects. Both of these methods use all available data from each subject.

Repeated Measures from Two Populations

Table 6 displays plasma inorganic phosphate measurements obtained from 13 control and 20 obese patients 0, 0.5, 1, 1.5, 2, and 3 hours after an oral glucose challenge [50]. The sample means are plotted in Figure 1. Since the relationship between plasma inorganic phosphate level and time is not monotonic, the univariate approach using slopes or correlation coefficients seems inappropriate. Zerbe [50] compared the two groups over the period from 0 to 3 hours using his randomization analysis of growth curves and reported a P value of 0.002.

The two groups can also be compared using the multivariate nonparametric tests of Puri & Sen [41]. Using the multivariate multisample rank sum test, the χ^2 statistic is 21.5 with 6 df ($P < 0.001$). The multivariate multisample median test gives a less significant result ($\chi^2 = 16.2$, df $= 6$, $P = 0.013$).

Table 6 Plasma inorganic phosphate levels in 13 control and 20 obese patients

		Hours after glucose challenge					
Group	Patient	0	0.5	1	1.5	2	3
Control	1	4.3	3.3	3.0	2.6	2.2	2.5
	2	3.7	2.6	2.6	1.9	2.9	3.2
	3	4.0	4.1	3.1	2.3	2.9	3.1
	4	3.6	3.0	2.2	2.8	2.9	3.9
	5	4.1	3.8	2.1	3.0	3.6	3.4
	6	3.8	2.2	2.0	2.6	3.8	3.6
	7	3.8	3.0	2.4	2.5	3.1	3.4
	8	4.4	3.9	2.8	2.1	3.6	3.8
	9	5.0	4.0	3.4	3.4	3.3	3.6
	10	3.7	3.1	2.9	2.2	1.5	2.3
	11	3.7	2.6	2.6	2.3	2.9	2.2
	12	4.4	3.7	3.1	3.2	3.7	4.3
	13	4.7	3.1	3.2	3.3	3.2	4.2
Obese	1	4.3	3.3	3.0	2.6	2.2	2.5
	2	5.0	4.9	4.1	3.7	3.7	4.1
	3	4.6	4.4	3.9	3.9	3.7	4.2
	4	4.3	3.9	3.1	3.1	3.1	3.1
	5	3.1	3.1	3.3	2.6	2.6	1.9
	6	4.8	5.0	2.9	2.8	2.2	3.1
	7	3.7	3.1	3.3	2.8	2.9	3.6
	8	5.4	4.7	3.9	4.1	2.8	3.7
	9	3.0	2.5	2.3	2.2	2.1	2.6
	10	4.9	5.0	4.1	3.7	3.7	4.1
	11	4.8	4.3	4.7	4.6	4.7	3.7
	12	4.4	4.2	4.2	3.4	3.5	3.4
	13	4.9	4.3	4.0	4.0	3.3	4.1
	14	5.1	4.1	4.6	4.1	3.4	4.2
	15	4.8	4.6	4.6	4.4	4.1	4.0
	16	4.2	3.5	3.8	3.6	3.3	3.1
	17	6.6	6.1	5.2	4.1	4.3	3.8
	18	3.6	3.4	3.1	2.8	2.1	2.4
	19	4.5	4.0	3.7	3.3	2.4	2.3
	20	4.6	4.4	3.8	3.8	3.8	3.6

Although the Wei–Lachin [49] and Wei–Johnson [48] procedures were developed for the two-sample case with incomplete data, these procedures can also be applied. The Wei–Lachin vector of test statistics at the six time points is $\mathbf{W}' = (-0.5539, -0.8862, -1.0761, -0.9390, -0.1319, -0.0633)$ with estimated **covariance matrix**

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.080180 & 0.047379 & 0.065353 & 0.062623 & 0.031898 & 0.042313 \\ 0.047379 & 0.088898 & 0.052950 & 0.028881 & 0.014827 & 0.013615 \\ 0.065353 & 0.052950 & 0.095713 & 0.062215 & 0.002469 & 0.021214 \\ 0.062623 & 0.028881 & 0.062215 & 0.093537 & 0.031306 & 0.045055 \\ 0.031898 & 0.014827 & 0.002469 & 0.031306 & 0.071744 & 0.049681 \\ 0.042313 & 0.013615 & 0.021214 & 0.045055 & 0.049681 & 0.075706 \end{pmatrix}$$

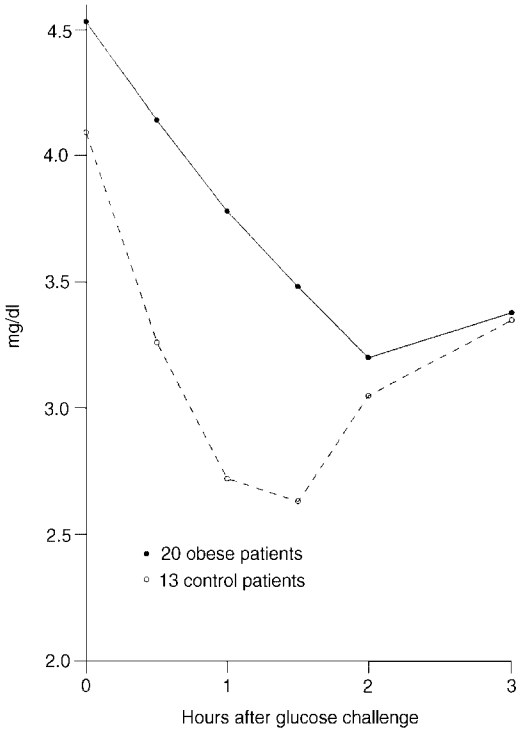

Figure 1 Mean plasma inorganic phosphate levels

The Wei–Lachin omnibus χ^2 statistic for testing equality of distributions is $\mathbf{W}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{W} = 20.9$ with 6 df ($P = 0.002$). The Wei–Johnson procedure using the "kernel" function

$$\phi(x, y) = \begin{cases} 1, & \text{if } x > y, \\ 0, & \text{if } x = y, \\ -1, & \text{if } x < y, \end{cases}$$

gives a vector $\mathbf{U}$ of test statistics equivalent (apart from a scale factor) to the Wei–Lachin $\boldsymbol{W}$, but uses a different estimator of the covariance matrix. Weighting each time point equally, the Wei–Johnson univariate statistic $\mathbf{c}'\mathbf{U}/(\mathbf{c}'\widehat{\mathbf{\Sigma}}_{\mathbf{U}}\mathbf{c})^{1/2}$, with $\mathbf{c}' = (1, \ldots, 1)$, is equal to -2.21. With reference to the standard normal distribution, the two-sided P value is 0.027.

References

[1] Agresti, A. & Pendergast, J. (1986). Comparing mean ranks for repeated measures data, *Communications in Statistics – Theory and Methods* **15**, 1417–1434.

[2] Akritas, M.G. (1991). Limitations of the rank transform procedure: a study of repeated measures designs, part I, *Journal of the American Statistical Association* **86**, 457–460.

[3] Akritas, M.G. (1993). Limitations of the rank transform procedure: a study of repeated measures designs, part II, *Statistics and Probability Letters* **17**, 149–156.

[4] Akritas, M.G. & Arnold, S.F. (1994). Fully nonparametric hypothesis for factorial designs, I: multivariate repeated measures designs, *Journal of the American Statistical Association* **89**, 336–343.

[5] Altman, N.S. (1990). Kernel smoothing of data with correlated errors, *Journal of the American Statistical Association* **85**, 749–759.

[6] Benard, A. & van Elteren, P. (1953). A generalization of the method of *m* rankings, *Proceedings Koninklijke Nederlands Akademie van Wetenschappen (A)* **56**, 358–369.

[7] Bhapkar, V.P. (1984). Univariate and multivariate multisample location and scale tests, in *Handbook of Statistics*, Vol. 4: *Nonparametric Methods*, P.R. Krishnaiah & P.K. Sen, eds. Elsevier, Amsterdam, pp. 31–62.

[8] Brown, G.W. & Mood, A.M. (1951). On median tests for linear hypotheses, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley.

[9] Carr, G.J., Hafner, K.B. & Koch, G.G. (1989). Analysis of rank measures of association for ordinal data from longitudinal studies, *Journal of the American Statistical Association* **84**, 797–804.

[10] Conover, W.J. & Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics, *American Statistician* **35**, 124–133.

[11] Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*. Chapman & Hall, London.

[12] Davis, C.S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials, *Statistics in Medicine* **10**, 1959–1980.

[13] Davis, C.S. (1994). A computer program for nonparametric analysis of incomplete repeated measures from two samples, *Computer Methods and Programs in Biomedicine* **42**, 39–52.

[14] Davis, C.S. (1996). Non-parametric methods for comparing multiple treatment groups to a control group, based on incomplete non-decreasing repeated measurements, *Statistics in Medicine* **15**, 2509–2521.

[15] Davis, C.S. & Wei, L.J. (1988). Nonparametric methods for analyzing incomplete nondecreasing repeated measurements, *Biometrics* **44**, 1005–1018.

[16] Dawson, J.D. (1994). Comparing treatment groups on the basis of slopes, areas-under-the-curve, and other summary measures, *Drug Information Journal* **28**, 723–732.

[17] Diggle, P.J., Liang, K.Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

[18] Durbin, J. (1951). Incomplete blocks in ranking experiments, *British Journal of Mathematical and Statistical Psychology* **4**, 85–90.

[19] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of

variance, *Journal of the American Statistical Association* **32**, 675–701.

[20] Frison, L. & Pocock, S.J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design, *Statistics in Medicine* **11**, 1685–1704.

[21] Ghosh, M., Grizzle, J.E. & Sen, P.K. (1973). Nonparametric methods in longitudinal studies, *Journal of the American Statistical Association* **68**, 29–36.

[22] Hart, J.D. & Wehrly, T.E. (1986). Kernel regression estimation using repeated measurements data, *Journal of the American Statistical Association* **81**, 1080–1088.

[23] Hettmansperger, T.R. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

[24] Hodges, J.L. & Lehmann, E.L. (1962). Rank methods for combination of independent experiments in analysis of variance, *Annals of Mathematical Statistics* **33**, 482–497.

[25] Kepner, J.L. & Robinson, D.H. (1988). Nonparametric methods for detecting treatment effects in repeated-measures designs, *Journal of the American Statistical Association* **83**, 456–461.

[26] Koch, G.G. (1969). Some aspects of the statistical analysis of "split plot" experiments in completely randomized layouts, *Journal of the American Statistical Association* **64**, 485–505.

[27] Koch, G.G. & Sen, P.K. (1968). Some aspects of the statistical analysis of the mixed model, *Biometrics* **24**, 27–48.

[28] Koch, G.G., Amara, I.A., Stokes, M.E. & Gillings, D.B. (1980). Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography, *International Statistical Review* **48**, 249–265.

[29] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583–621.

[30] Kshirsagar, A.M. & Smith, W.B. (1995). *Growth Curves*. Marcel Dekker, New York.

[31] Lachin, J.M. (1992). Some large-sample distribution-free estimators and tests for multivariate partially incomplete data from two populations, *Statistics in Medicine* **11**, 1151–1170.

[32] Landis, J.R., Heyman, E.R. & Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests, *International Statistical Review* **46**, 237–254.

[33] Landis, J.R., Miller, M.E., Davis, C.S. & Koch, G.G. (1988). Some general methods for the analysis of categorical data in longitudinal studies, *Statistics in Medicine* **7**, 109–137.

[34] Madansky, A. (1963). Test of homogeneity for correlated samples, *Journal of the American Statistical Association* **58**, 97–119.

[35] Matthews J.N.S., Altman, D.G., Campbell, M.J. & Royston, P. (1990). Analysis of serial measurements in medical research, *British Medical Journal* **300**, 230–235.

[36] Müller, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer-Verlag, Berlin.

[37] Page, E.B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks, *Journal of the American Statistical Association* **58**, 216–230.

[38] Palesch, Y.Y. & Lachin, J.M. (1994). Asymptotically distribution-free multivariate rank tests for multiple samples with partially incomplete observations, *Statistica Sinica* **4**, 373–387.

[39] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, New York.

[40] Potthoff, R. & Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika* **41**, 313–326.

[41] Puri, M.L. & Sen, P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[42] Raz, J. (1989). Analysis of repeated measurements using nonparametric smoothers and randomization tests, *Biometrics* **45**, 851–871.

[43] Schwertman, N.C. (1982). Algorithm AS 174: multivariate multisample non-parametric tests, *Applied Statistics* **31**, 80–85.

[44] Schwertman, N.C. (1985). Multivariate median and rank sum tests, in *Encyclopedia of Statistical Sciences Vol. 6*, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 85–88.

[45] Sen, P.K. (1984). Nonparametric procedures for some miscellaneous problems, in *Handbook of Statistics*, Vol. 4: *Nonparametric Methods*, P.R. Krishnaiah & P.K. Sen, eds. Elsevier, Amsterdam, pp. 699–739.

[46] Thall, P.F. & Lachin, J.M. (1988). Analysis of recurrent events: nonparametric methods for random-interval count data, *Journal of the American Statistical Association* **83**, 339–347.

[47] Thompson, G.L. (1991). A unified approach to rank tests for multivariate and repeated measures designs, *Journal of the American Statistical Association* **86**, 410–419.

[48] Wei, L.J. & Johnson, W.E. (1985). Combining dependent tests with incomplete repeated measurements, *Biometrika* **72**, 359–364.

[49] Wei, L.J. & Lachin, J.M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations, *Journal of the American Statistical Association* **79**, 653–661.

[50] Zerbe, G.O. (1979). Randomization analysis of the completely randomized design extended to growth and response curves, *Journal of the American Statistical Association* **74**, 215–221.

[51] Zerbe, G.O. (1979). Randomization analysis of randomized blocks extended to growth and response curves, *Communications in Statistics – Theory and Methods* **8**, 191–205.

[52] Zerbe, G.O. & Walker, S.H. (1977). A randomization test for comparison of groups of growth curves with different polynomial design matrices, *Biometrics* **33**, 653–657.

CHARLES S. DAVIS

DNA Sequences

The hereditary information of essentially all living organisms is carried by DNA molecules made up of complementary chains of nucleotides twisted around each other into a double-helical structure (Figure 1). The four possible nucleotides are the two purines, adenine (A) and guanine (G), and the two pyrimidines, cytosine (C), and thymine (T). As the first step in the synthesis of proteins, DNA is transcribed to RNA which is usually single-stranded and uses the nucleotide uracil (U) in place of thymine. The primary data on the composition of these molecules are then sequences of nucleotides presented in a four-letter alphabet (A, C, G, or T/U). The ability of molecular biologists to generate nucleotide sequences rapidly has created many new areas of biomedical research and fundamentally changed the focus of many others. The most important technology to facilitate rapid sequencing over the past decade has been the polymerase chain reaction (pcr) technique which was first described in the early 1970s [16] but not made practicable as a standard laboratory technique until the late 1980s [18]. In pcr amplification, two primers that flank the region to be sequenced are used to initiate the reaction. Alternating applications of heat which "melts" the DNA into a single stranded state, followed by slow cooling in the presence of synthetic nucleotides, allow for a heat-stable polymerase to catalyze the synthesis of new DNA. With each cycle, duplicates of the copies of the DNA bridging the two primers are created until they dominate the mixture and can be easily sequenced.

Databases with nucleotide sequences have grown exponentially in size – the most widely used database being GenBank, maintained by the National Center for Biotechnology Information of the US National Library of Medicine, which at this writing contains about a half-billion nucleotides from a half-million sequences. New technologies, such as the use of hybridization techniques to produce "DNA sequencing chips", promise to increase the rate that data are generated by another order of magnitude.

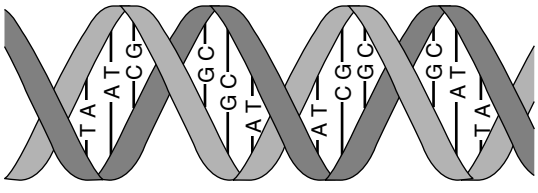

Figure 1 The DNA double helix. Each strand carries equivalent information because of the A:T, G:C pairing across strands

The statistical analysis of nucleotide sequences involves a collection of interrelated, and computationally intensive, problems which have been attacked somewhat in isolation. These include the problem of assembling a single sequence from many overlapping partial sequences, developing **stochastic models** of molecular evolution, aligning homologous combinations of sequences, and studying the common evolutionary history of a large group of sequences.

The Sequencing Chip

The process of assembling DNA sequences has led to many interesting problems in combinatorics [28]. Consider, for instance, the recent proposal to use hybridization techniques to find all distinct k-tuple fragments (i.e. subsequences of k nucleotides) that exist within the chain to be reconstructed. This requires the use of a "DNA sequencing chip" which is typically a matrix of 4^k probes [20]. For example, if $k = 5$ and the (unknown) sequence being constructed is TACGGAACGGAT, then the hybridized 5-tuples specified by the chip will be TACGG, ACGGA, CGGAA, GGAAC, GAACG, AACGG, and CGGAT. Note that the chip cannot tell us that the 5-tuple ACGGA appeared twice in the sequence. The problem is to reconstruct the full sequence from the observed k-tuples. To do this, a graph is formed with directed edges connecting two observed k-tuples if the first $k-1$ letters of one are equal to the last $k-1$ letters of the other. The possible full sequences are then represented by paths connecting all observed k-tuples. Reconstructing long sequences with many repetitive features is obviously a daunting task. However, the sequencing chip idea has recently become more practicable as a general tool with the current production of chips reading all 10-tuples in a sequence. Also, the specialized use of "designer chips" with a variety of different sized fragments geared toward a specific application is already a reality in some laboratories [11].

Models of Sequence Evolution

Stochastic models of molecular evolution typically assume that evolution proceeds independently at each nucleotide site via homogeneous **Poisson** substitution models. Specifically, assume that during a small increment of time $(t, t+h)$, (i) the **probability** that a substitution changes the nucleotide i into nucleotide j has probability $q_{ij}h + o(h)$, (ii) the probability that no substitution occurs at a site currently occupied by the nucleotide i is $1 - \sum_{j \neq i} q_{ij}h + o(h)$, and (iii) the probability that two or more substitutions occur is $o(h)$. This gives the framework for the models studied by Rodriquez et al. [21]. The specific form of the model is fixed by the instantaneous rate matrix, $\mathbf{Q}$, which has off-diagonal elements q_{ij} and diagonal elements $-\sum_{j \neq i} q_{ij}$ (i.e. each row sums to zero). The elements of the matrix $\mathbf{P} = \exp(t\mathbf{Q})(= \sum_{n=0}^{\infty}(\mathbf{Q}^n t^n / n!)$, where $\mathbf{Q}^0$ is the 4×4 identity matrix) then provide the transition probabilities for a single site:

$$P_{ij}(t) = \Pr[\text{nucleotide } j \text{ at time } t | \text{nucleotide } i \text{ at time } 0].$$

The simplest model of this type was introduced by Jukes & Cantor [14] and makes all of the $q_{ij}s (i \neq j)$ identical, say $= \alpha$. Under this model we can derive the explicit expression:

$$P_{ij}(t) = \begin{cases} \dfrac{1 + 3\exp(-4\alpha t)}{4}, & \text{if } i = j, \\ \dfrac{1 - \exp(-4\alpha t)}{4}, & \text{if } i \neq j. \end{cases}$$

If two sequences evolved independently for a time t from a common ancestor, then there is $2t$ time between them and $\theta = \Pr[\text{two sequences are different at a single site}] = 3[1 - \exp(-8\alpha t)]/4$. Also, the expected number of substitutions that occurred along the two branches over this time t is $6\alpha t = -3/4 \ln(1 - 4/3\theta)$. Finally, we may substitute $\hat{\theta}$, the proportion of sites that differ between the two sequences, into this last expression to obtain the Jukes–Cantor distance between the sequences. Jukes & Cantor's one-parameter model was extended by Kimura [15] to allow for a different rate when the substitution is between two purines or between two pyrimidines than when the substitution is from one group to the other. Felsenstein [5] extended the Jukes–Cantor model in a different way by taking $q_{ij} = \pi_j / 4\alpha$, where π_j is the stationary probability of observing nucleotide j at any site. The Kimura and Felsenstein models were combined into the five-parameter model investigated by Hasegawa et al. [12]. Distance measures between sequences can be based on any of these models using a procedure analogous to the one above based on the Jukes–Cantor model. Also, each of these widely used models is time reversible so that the probability of a change from nucleotide i to j is the same as the probability of a change from j to i [i.e. $\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$]. This property makes the direction of time irrelevant, which is helpful in **likelihood**-based phylogenetic analysis. Site independent substitution models with more than five parameters have generally been found to overfit the empirical data. To describe sequence evolution more accurately it is important to account for other complexities – the most discussed being: (i) allowance for the insertion or deletion of nucleotides, singly or in groups; (ii) allowance for time heterogeneity in rate parameters; (iii) allowance for different rates when the amino acid structure in protein coding sequences is changed by a substitution; (iv) allowance for **linkage** between sites; and (v) allowance for recombination events.

Aligning Sequences

Suppose that two sequences are thought to be homologous in some way, for example they might be for the same **gene** from two different species. In the pairwise alignment problem we wish to designate which are the corresponding sites in the two sequences to be related. As an illustration, consider the following two short sequence segments of the 28S rRNA gene from the human and the carp:

site	`123456789··········`
human	`CGGCAAGGCTTCCCTGCCGG`
carp	`CGGTCAAGCCTTCCCTCCGG`

The alignment implied by this way of arranging these two sequences requires at least seven substitutions to have occurred in their common history (the C/T change in site 4, the A/C change in site 5, etc.). However, if we use a model of evolution that also allows for insertions and deletions of nucleotides, the

following alignment becomes possible:

site	123456789·····
human	CGG-CAAGGCTTCCCTGCCGG
carp	CGGTCAAGCCTTCCCT-CCGG

This alignment requires two gaps (insertions or deletions denoted by a "-") but only one substitution (the G/C change at site 9). If insertions or deletions are common, then the second alignment should be preferred, but if they are rare, then gaps should be heavily penalized and the first alignment may be preferred. The choice of alignment then depends on a score function that specifies appropriate weights for different types of substitutions and gaps.

To give the problem a more mathematical framework, suppose the sequence **a**, which has nucleotide a_i at position i for $i = 1, \ldots, N$, is to be aligned with sequence **b**, which has nucleotide b_j at position j for $j = 1, \ldots, M$. The problem is to produce optimal sequences **a*** and **b***, each of length $K \geq \max\{N, M\}$, whose elements are either gaps or the original elements of **a** and **b**. A typical criterion for optimization is the function $\sum_{k=1}^{K} \delta(a_k^*, b_k^*)$. The score function δ may be a measure of similarity (the maximum is optimal) or of distance (the minimum is optimal) between the sequences.

Next, define $S(n, m)$ to be the optimal (say maximum) score aligning sequence **a** up to position n with sequence **b** up to position m. To find the optimal score, $S(N, M)$, for the global alignment of the two sequences, it is possible to use dynamic programming **algorithms** [1]. These are essentially based on the idea that the optimal alignment may be found through the appropriately initialized recursion:

$$\begin{aligned} S(n, m) = \max\{&S(n-1, m-1) + \delta(a_n, b_m), \\ &S(n-1, m) + \delta(a_n, \text{gap}), \\ &S(n, m-1) + \delta(\text{gap}, b_m)\}. \end{aligned}$$

Different dynamic programming algorithms have been proposed for different circumstances (see, for example, [19] for the 20-letter amino acid alphabet of protein sequences; [24] for the problem of finding local similarities). Typically, global alignment algorithms can be proven to produce an optimum in $O(K^2)$ time. Crucial to the implementation here is the decision of what scoring function, δ, is to be used. A simple choice is to take $\delta(x, y) = 1$ if $x = y \neq$ gap, α if x or $y =$ gap, and 0 otherwise. The choice of $\delta(x, y)$ can also be tied to a statistical model. In particular, take $\delta(x, y)$ to be the log of the probability that x is in sequence **a*** and y is in sequence **b*** at a particular site under a model of molecular evolution that assumes independence across sites. The resulting optimal alignment then maximizes the likelihood under this model (*see* **Maximum Likelihood**). Because the simultaneous insertion or deletion of several adjacent nucleotides is common in the evolution of some sequences, it is important to generalize the site independent models to allow for this possibility. In this regard, similarity-based score functions which are concave functions of the length of the gap can still be treated with efficient dynamic programming algorithms [17].

The problem of simultaneously aligning r sequences presents greater computational challenges since a dynamic programming algorithm leading to an exact solution takes $O((2K)^r)$ time. Suboptimal alignments are often found by various methods of merging all pairwise alignments. An alternative approach attempts to find the multiple alignment that is most compatible with a phylogenetic analysis [13, 29].

Instead of starting with sequences known to be homologous, sometimes a single target sequence is compared with a large library of sequences in a search for sequences that have a statistically significant degree of similarity with the target. This leads to a study of the probability distribution for independent sequences of alignment scores and of other measures of coincidence, such as the length of the longest exact match. Alignment procedures are also used extensively in sequence assembly procedures. For example, copying errors made during pcr can sometimes be eliminated by replicating the process several times and comparing the aligned results for differences. An excellent discussion of these problems and other computational, mathematical, and statistical aspects of alignment methodology may be found in [28].

Phylogenetic Analysis

Phylogenetic analysis tries to infer the evolutionary history that is most consistent with a set of observed nucleotide sequences. (Phylogenies are also commonly built from data on amino acids, from data on the allele's of specific genes, or from morphological data using procedures analogous to those described below.) Consider the aligned HIV-1

Table 1 Viral nucleotide sequences from five AIDS patients (84 bases in the TAT region)

Country	Sequence
Djibouti-1	CAGGAAGTCAGCCTAAAACTGCTTGTAATAAGTGTTATTGTAAAAAATGTAGCTATCATTGTCTAGTTTGCTTTCAGACAAAAG
Djibouti-2	**********C***A**
Zambia	*G**********CC**********G***T**********C*C***G*******TA*AC****
Brazil	**********AG****CC**T**C**********G***T***T******C*A******T**CACA*******
US	**********G**********CC**T**C**********G***T**CT******C*A******T**CACA*******
Uganda	**********C******C********T*G**********T**********A**********T**AC****
Cameroon	*T**G**CA****CC****CC******C***T**C*****C****G***CT**********-***TG**TG*****C*A*G***

Note: An asterisk (∗) designates the same nucleotide as the first patient from Djibouti.

sequences (from the TAT gene of the virus) of seven AIDS patients presented in Table 1.

We can see that the sequences from the two Djiboutian patients are very similar – indicating a recent common ancestry. However, the sequence from the Cameroonian patient is quite different from the others, indicating an earlier divergence. This evolutionary history is illustrated by means of a phylogenetic tree, such as the ones in Figure 2. The external nodes at the tips of the phylogenetic tree represent the current nucleotide sequences under consideration. The internal nodes represent ancestors of the current sequences. The trees in Figure 2 are both bifurcating (each node splitting into two branches) and show the same relative relationships among the seven sequences. In the rooted tree of Figure 2(a) evolutionary time proceeds from the root (the common ancestor of all nodes) at node 1 down to the current time. Whether rooted or unrooted, our interest lies in estimating key properties of the true underlying tree, especially (i) the branching structure or topology of the tree, and (ii) the lengths of the branches. The space of trees is very large; there are $\Pi_{k=2}^{n}(2k-3)$ possible distinct rooted topologies relating n sequences (when $n = 40$ this is already $> 10^{57}$).

The wide variety of methods used to estimate phylogenetic relationships generally fall into three categories: distance-matrix, **parsimony**, and maximum likelihood. Distance-matrix methods use pairwise distances to infer the tree. An early algorithm of this type is UPGMA, which is an unweighted recursive pairwise grouping method [25]. The neighbor-joining method [22] weights pairwise distances on the basis of their average distance from all other external nodes, while the Fitch–Margoliash procedure [9] minimizes a **chi-square**-like criterion measuring disagreement between the branch lengths of the tree

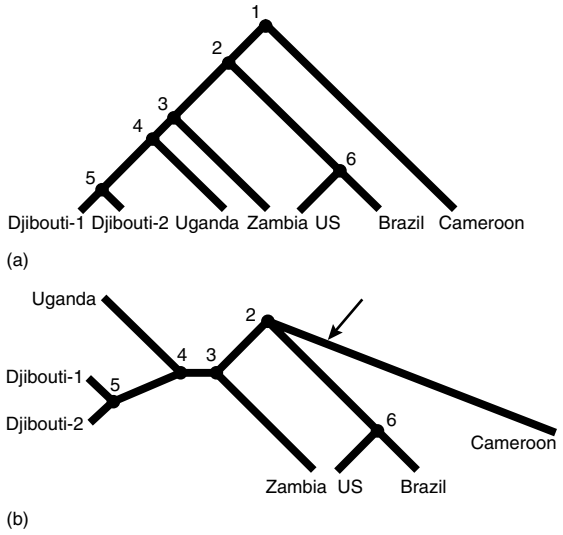

Figure 2 Two phylogenetic trees relating seven HIV-1 DNA sequences identified by country of origin: (a) a rooted tree; (b) an unrooted tree. The arrow indicates the hypothesized position of the root used to make tree (a)

and the corresponding observed distances. Distance-matrix methods are computationally very efficient, although important information may be lost in the reduction to pairwise distances.

Algorithms based on the parsimony concept attempt to minimize the number of evolutionary steps required to explain a given set of data. The idea was introduced for amino acid data by Eck & Dayoff [3] and for nucleotide sequences by Fitch [8]. As an example, the branching structure in Figure 2 is the most parsimonious for the HIV data in Table 1 – it explains the data using a total of 49 nucleotide base substitutions. Parsimony methods ignore the possibility of multiple substitutions at one site along the

same branch and have been shown to give inconsistent estimates of the true phylogeny [4]. However, parsimony methods that give weights to different types of events appear to improve the consistency and efficiency of the technique. Also, parsimony is the preferred method of analysis amongst researchers in cladistics, which concentrates on the branching relationships of species.

The maximum likelihood approach seeks to find the tree that maximizes the likelihood of the observed sequences under a particular model of the evolutionary process. Most algorithms require the stochastic model to be time-reversible and **Markovian** and to assume independence between sites. In this setting the log likelihood is summed across sites and the likelihood at an individual site is computed using Felsenstein's peeling algorithm [5]. In particular, define $L(i \text{at } j)$ to be the conditional likelihood of the subtree descending from node j given that nucleotide i is at j. The peeling algorithm is based on the observation that if node 1 is the parent of nodes 2 and 3 (with corresponding branch lengths t_{12} and t_{13}), then

$$L(i \text{at } 1) = \left\{ \sum_k P_{ik}(t_{12}) L(k \text{at } 2) \right\} \times \left\{ \sum_k P_{ik}(t_{13}) L(k \text{at } 3) \right\}.$$

Using this recursion, the likelihood is finally derived under the further assumption that the probability of each nucleotide being at the root node follows the stationary nucleotide distribution of the evolutionary model. Coincidentally, it turns out for the data of Table 1 that the maximum likelihood method using the Hasegawa–Kishino–Yano model gives the same topology as the parsimony method. The maximum likelihood method is **consistent** if the assumed model is true. However, it is very intensive computationally and when the number of sequences is larger than 20, we must resort to a heuristic search for the maximum.

To evaluate the variability in an estimated tree, Felsenstein [6] introduced the use of the **bootstrap method** to phylogenetic inference. Bootstrap samples are created by sampling the observed sites, with replacement, to create a set of pseudo-sequences of the same length as the original data. For each bootstrap sample, a bootstrap tree estimate is made using the specific phylogenetic construction method being studied. The sampling distribution of the estimated phylogeny is approximated by the empirical distribution of the bootstrap estimates. Care must be taken in the interpretation of the bootstrap results since no information is provided about whether the estimation technique being studied is itself biased.

Phylogenetic tree-building algorithms rely on many key assumptions and there have been a number of proposals to test their validity. Methods that assume rate homogeneity over the whole tree can examine this assumption with a relative rate test (e.g. [23]). Trees based on maximum likelihood are not robust to deviations in the assumption that all sites evolve at the same rate [10] and many proposals are available to deal with this problem (e.g. [27]). A phylogenetic tree cannot depict situations in which recombination events play an important role in the history of a group of homologous sequences. In this case a network, which allows for cycles amongst the nodes, may be appropriate [2]. The preponderance of work in the analysis of DNA sequences assumes that all the relevant information about the history and function of the molecule is contained in the sequence. Although higher-order structure probably plays an important functional role, the lack of data in this area makes the development of useful statistical procedures problematic.

Computer programs to carry out a wide variety of alignment and phylogenetic tree-building algorithms are generally available for little or no cost. A fairly comprehensive list of such programs is distributed as part of the documentation for the program PHYLIP [7]. A broad discussion of phylogenetic methods is provided by Swofford et al. [26].

References

[1] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton.

[2] Crandall, K.A. & Templeton, A.R. (1996). Applications of intraspecific phylogenetics, in *New Uses for New Phylogenies*, P.H. Harvey, A.J. Leigh Brown & J. Maynard Smith, eds. Oxford University Press, Oxford.

[3] Eck, R.V. & Dayhoff, M.O. eds. (1996). *Atlas of Protein Sequence and Structure 1966*. National Biomedical Research Foundation, Silver Springs.

[4] Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading, *Systematic Zoology* **27**, 401–410.

[5] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach, *Journal of Molecular Evolution* **17**, 368–376.

[6] Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap, *Evolution* **39**, 783–791.

[7] Felsenstein, J. (1996). *PHYLIP (Phylogeny Inference Package)*, homepage (`http://evolution.genetics.washington.edu/phylip.html`). Department of Genetics, University of Washington, Seattle.

[8] Fitch, W.M. (1971). Toward defining the course of evolution: minimal change for a specific tree topology, *Systematic Zoology* **20**, 406–416.

[9] Fitch, W.M. & Margoliash, E. (1967). Construction of phylogenetic trees, *Science* **155**, 279–284.

[10] Gaut, B.S. & Lewis, P.O. (1995). Success of maximum likelihood in the four taxon case, *Molecular Biology and Evolution* **12**, 152–162.

[11] Gibbs, W.W. (1996). New chip off the old block, *Scientific American* **275**, 42–44.

[12] Hasegawa, M., Kishino, H. & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution* **22**, 32–38.

[13] Hein, J. (1990). Unified approach to alignment and phylogenies, *Methods in Enzymology* **183**, 626–644.

[14] Jukes, T.H. & Cantor, C.R. (1969). Evolution of protein molecules, in *Mammalian Protein Metabolism*, H.N. Munro, ed. Academic Press, New York.

[15] Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences, *Journal of Molecular Evolution* **16**, 111–120.

[16] Kleppe, K.E., Ohtsuka, R., Molineux, I. & Khorana, H.G. (1971). Studies on polynucleotides XCVI. Repair replication of short synthetic DNAs as catalyzed by DNA polymerases, *Journal of Molecular Biology* **56**, 341–361.

[17] Miller, W. & Myers, E.W. (1988). Sequence comparison with concave weighting functions, *Bulletin of Mathematical Biology* **50**, 97–120.

[18] Mullis, K.B. & Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction, *Methods in Enzymology* **155**, 335–350.

[19] Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins, *Journal of Molecular Biology* **48**, 443–453.

[20] Pevzner, P.A. (1989). 1-tuple DNA sequencing: computer analysis, *Journal of Biomolecular Structure and Dynamics* **7**, 63–73.

[21] Rodriquez, F., Oliver, J.L. Marin, A. and Medina, J.R. (1990). The general stochastic model of nucleotide substitution, *Journal of Theoretical Biology* **142**, 485–501.

[22] Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology Evolution* **4**, 406–425.

[23] Sarach, V.M. & Wilson, A.C. (1973). Generation time and genomic evolution in primates, *Science* **179**, 1144–1147.

[24] Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular sequences, *Journal of Molecular Biology* **147**, 195–197.

[25] Sokal, R.R. & Michener, C. (1958). A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin* **28**, 1409–1438.

[26] Swofford, D.L., Olsen, G.J., Waddell, P.J. & Hillis, D.M. (1996). Phylogenetic inference, in *Molecular Systematics*, 2nd Ed., D.M. Hillis, C. Morits & B.K. Mable, eds. Sinauer Associates, Sunderland.

[27] Van de Peer, Y., Neefs, J.M., De Rijk, P. & Wachter, R.De (1993). Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock, *Journal of Molecular Evolution* **37**, 221–232.

[28] Waterman, M.S. (1995). *Introduction to Computational Biology*. Chapman & Hall, London.

[29] Wheeler, W.C. & Gladstein, D. (1994). MALIGN: a multiple sequence alignment program, *Journal of Heredity* **85**, 417–418.

DENNIS K. PEARL

Dorn, Harold Fred

Born: July 30, 1906, in Tompkins County, New York.
Died: May 9, 1963, in Bethesda, Maryland.

At the time of his death from cancer, Harold Dorn was Chief of the Biometrics Research Branch of the National Heart Institute of the **National Institutes of Health** (NIH). He was originally trained as a sociologist, receiving his Ph.D. from the University of Wisconsin in 1933. His interest in statistical methods was stimulated by his post-doctoral year in London where he attended, among others, the lectures of **Egon Pearson** and **R.A. Fisher**. He joined the US Public Health Service in 1936. Shortly thereafter he was assigned to the NIH where he spent his entire career as a medical statistician and epidemiologist. Dorn will be remembered for the major role he played in starting and developing the statistical program at the NIH. His primary interest, however, was the conduct of his own studies in the epidemiology of cancer. He was responsible for some early work in the association between smoking and lung cancer (*see* **Smoking and Health**) and, as a result of initiating a 10-city survey of cancer morbidity, was the first to collect a fairly large body of data on cancer incidence in the US. For these and other contributions that he made in medical statistics, epidemiology, and demography, he received a number of honors, among which were his designation as the Cutter Lecturer in Preventive Medicine at Harvard University in 1959 and his receiving the distinguished Service Award from the US Department of Health, Education and Welfare. He was a member of the National Committee on Vital and Health Statistics, a member of the **World Health Organization** (WHO) Expert Committee on Health Statistics, and a US representative from 1948 until his death to conferences on the revision of the International List of Diseases, Injuries and Causes (*see* **International Classification of Diseases (ICD)**). Dorn was a Fellow of the **American Statistical Association** (ASA), served on its Board of Directors and its Council, and at his death was Chairman of the Social Statistics Section.

(This account is based on the Dorn obituary written by **Jerome Cornfield** in the *American Statistician*, June 1963.)

SAMUEL W. GREENHOUSE

Dose-response in Pharmacoepidemiology

Since **pharmacoepidemiology** (PE) deals with the action of drugs within the population, the estimation of **dose–response** parameters is somewhat different than that used in either **biological assay** or **clinical trials**. As discussed below, these differences are primarily due to the controlled data collection available in bioassay and clinical trials which usually is not available in PE trials. However, the basic idea behind a dose–response curve is the same. As the amount of drug used increases, the number of subjects who will respond to the drug increases. This relationship produces an S-shaped or sigmoid curve when amount of drug is plotted on the horizontal axis and proportion of subjects responding is plotted on the vertical axis.

Such a dose–response curve is shown in Figure 1. The plotted points indicate the proportion of subjects (usually 3–6) who have responded to a particular dose of the drug being studied. The sigmoid curve has been drawn through this set of points. In practice, the parameters describing this curve would be estimated using **nonlinear regression** or by transforming the data (*see* **Transformations**) so that **linear regression** could be used [5]. The ends of the curve flatten out at the lower end where a certain amount of the drug has to be given to obtain any response, and at the upper end where increasing the amount of drug above a particular dose produces no increase in the effect that the drug produces.

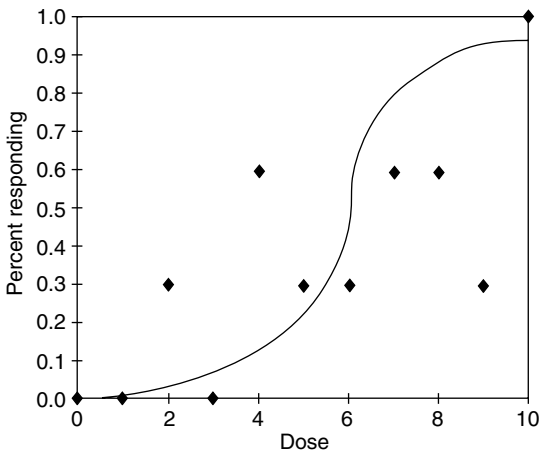

Figure 1 Plotted points and a sigmoid curve for a typical dose–response situation

Dose–response is evaluated with a pharmacodynamic (PD) model. That is, its results are measured in terms of an observable response by the subject. Underlying this response are **pharmacokinetic** (PK) actions of the drug within the body; for example, the time it takes the drug to clear from the body. PK information is thought to provide clues as to how a subject will respond to various doses of a drug, but it is not a perfect predictor of PD response.

The dose–response model in PE shares the S-shaped curve concept and the PD response with similar models in bioassay and clinical trials. In other ways it is quite different from these models. In PE the dose–response information comes exclusively from human subjects. Bioassay and clinical trials models are used in both animals and humans. Although cross species comparisons are difficult to make, the animal studies are used to complement human studies in bioassay and clinical trials.

In PE the dose–response model is best represented by two S-shaped curves. The first curve uses as response the medically intended purpose of the drug, i.e. its benefit. The second curve uses as response the negative or adverse outcomes that the drug produces. The curves in Figure 2 represent the ideal situation in which the adverse and beneficial effects

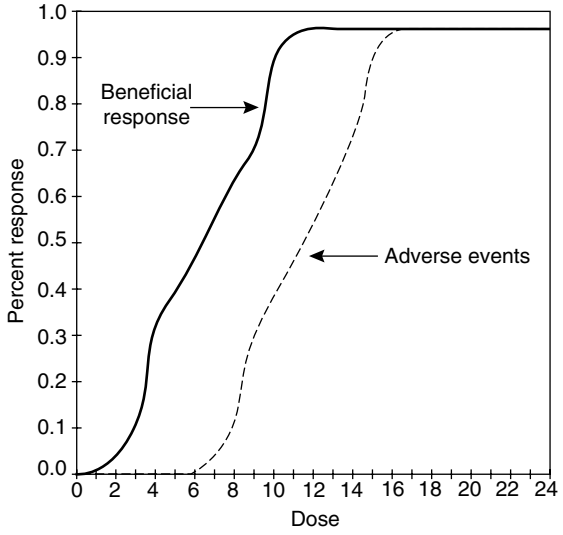

Figure 2 PE dose–response curves

are widely separated in dose and the curves are parallel to one another. In reality the two curves may overlap and cross one another or they may lie on top of one another. They do not have to be parallel, and their shape and relationship to one another may change with changing characteristics of the subjects being studied; for example, the age of the subjects usually affects this relationship. PE studies using dose–response models attempt to estimate parameters for both of these curves, and to estimate the effect that population characteristics have on these parameters and hence on the relationship between the curves.

Bioassay and clinical trials dose–response studies are controlled studies similar to the premarket studies conducted on drugs before they are allowed to be sold to the public. The number of subjects who are tested with a drug before it is put on the market is extremely small compared with the number of patients who will use the drug once it is generally available. For this reason the information about the relationship of the beneficial and adverse effects curves given in Figure 2 that is available from controlled (premarket) studies may not provide a complete picture of this relationship. One of the major research foci for PE is to determine the relationship of the two curves in Figure 2 in the general population. It is not uncommon that an effective dose for a drug can be lower than the doses used in premarket testing. It may also be necessary to increase the dosage for some subgroups in a population. Studies of the drug verapamil have shown that, for a given dose, elderly patients will have a blunted electrophysiologic response and a greater drop in blood pressure than younger patients given the same dose [15]. Using the recommended dose on the package insert for verapamil would therefore result in getting less electrophysiologic response than expected and a greater drop in blood pressure when the person being given the drug was older than the groups on which premarket studies had been done.

Bioassay and clinical trials dose–response experiments are controlled, prospectively planned, experiments. A great deal of the dose–response data in PE is currently gathered as retrospective information obtained from data sets that have been collected for other purposes. Some of the methodologic issues in PE dose–response research have to do with abstracting information from large computer databases (*see* **Administrative Databases**).

Since PE seeks to determine dose–response information for a population, it must deal with variables other than dose which can influence response. Decisions need to be made, often with limited information, about whether these variables are important enough to be studied separately. The alternative is to allow them to become one of the sources of variation for the models developed.

Finally, dose–response results in PE will almost surely be compared with data collected by pharmaceutical companies to determine adverse events, and with physicians' customary prescribing habits. While these results are not always in conflict, there are economic and political consequences that make the method of presenting these dose–response studies very important. The methods of PE are also being used to investigate drug prescribing practices among physicians (*see* **Drug Utilization Patterns**), the economic impact of altered dosing strategies, and the long-term adverse effects of drugs taken over extended periods of time (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**). These are areas of investigation that are difficult to incorporate into the classic bioassay or clinical trials structure.

As with any complicated problem, a good way to proceed is to divide the problem into manageable parts. I have chosen to divide this article into three sections, each dealing with an important aspect of dose–response in PE. First I will discuss the parameters that need to be estimated, next the current techniques for estimating these parameters and, finally, some of the computer software currently available for obtaining the estimates.

Dose–Response Variables

Responses can be divided into two categories – expectation and effect. There are two levels of expectation, anticipated and unanticipated, and two levels of effect, adverse and beneficial. We usually have some prior expectation about how a drug will affect the body both beneficially and adversely. These are the anticipated responses. They are the easiest to deal with, since we can include them in our advance planning. The unanticipated responses cannot be planned for in advance. Again, they can be either beneficial or adverse but they are determined by skillful observation of the effect of the drug on the human subject

during the course of the study. They are frequently detected by their increasing occurrence with increasing amounts of a drug.

The effect levels, adverse and beneficial, refer to PD responses as mentioned in the introduction. These can usually be clearly defined in the same way for all subjects. There are some situations, if for example a **quality of life** measure is used as the response, where some interpretation is necessary to decide what is beneficial and what is adverse.

While our basic dose–response model deals with PD responses, it seems reasonable to assume that the PD responses that we obtain from a drug depend to some extent upon how that drug acts in the body, that is, pharmacodynamics is some function of pharmacokinetics. Therefore the first set of responses that we want to estimate are not dose–response measures but the basic pharmacokinetic actions of a drug.

Pharmacokinetics (PK) has been defined as the "study of factors influencing the absorption, distribution, metabolism, and excretion of a drug" [7]. Pharmacodynamics is the study of the effect of a drug on the body [9], for example, whether or not the patient improves or has an adverse reaction when the drug is given. This effect is related to the drug's effective concentration **(bioavailability)** in the body. This concentration could be measured by the initial dose given, but because of the **confounding** effects caused by a subject's biochemistry other PK parameters may be better.

Several variables of drug administration influence the PK outcomes of any drug. Examples include the route of administration (oral, bolus injection or IV drip, etc.), the timing of administration (once per day, twice per day, etc.), the number of administrations, and the dose per administration. These variables, along with a subject's unique biochemistry, determine the bioavailability of a drug in the body. Aspects of this bioavailability are expressed by several PK parameters. The most common of these are the area under the curve (AUC), peak plasma concentration (C_p), and half-life ($T_{1/2}$). Depending upon the drug and the effect being examined, one of these parameters may be more important than the others in determining the PD response. The effect of some drugs is closely related to their maximum or peak concentration. Other drugs have effects related to their minimum concentration or trough. The PD response of others is most easily predicted by some integrated summation of bioavailability over time (AUC).

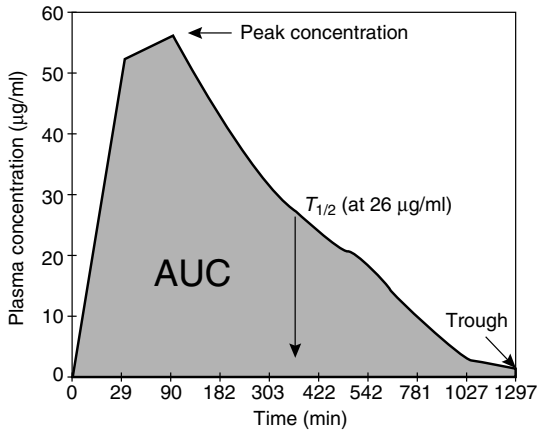

Figure 3 A pharmacokinetic curve for a single dose of cyclophosphamide

Figure 3 is a graphic representation of AUC, C_p, $T_{1/2}$, and trough for a single administration of the drug cyclophosphamide. From the time at which the drug is injected, 0, the plasma concentration rises until it reaches the peak, C_p. The time that it takes to reach half of C_p, about 362 minutes, is $T_{1/2}$, and the lowest value is the trough. The shaded area is the area under the curve, AUC.

In pharmacoepidemiologic dose–response (PDR) we determine the values of these PK parameters not just for a single individual but for a population of individuals. To the extent that PK values determine PD responses this would give us information about how the entire population would react to a drug. Obtaining estimates in this way is called population pharmacokinetic (PPK) modeling. DeVane et al. [4] give an example of this in their paper on alprazolam, a drug used for a number of psychiatric disorders. In that paper mixed-effect modeling was used to determine the mean and standard deviation for the clearance ($= \text{dose}/AUC$) for 94 psychiatric inpatients. The authors found an average clearance of 0.05 liters per hour per kg with a 95% confidence interval of 0.04–0.06. They also found that clearance was increased by 59% in women, and decreased by 26% in patients with multiple organ disease and 23% in patients over the age of 60. Increased clearance indicates that the drug is leaving the body more quickly and may not be in the plasma long enough to be fully effective. Decreased clearance indicates that the drug is remaining in the body longer and that may indicate a greater opportunity to cause an adverse reaction.

Subject factors such as age, gender, and co-morbid conditions can have considerable impact on a drug's effect (*see* **Co-morbidity**). The effect of these variables could be measured directly in planned experiments. What is more frequently done in PDR modeling is to use PPK models to determine how these subject specific variables effect PK parameter estimates and to impute from this what types of PD responses one would see in the population. This was the strategy used in the DeVane study mentioned above. The data for this study were not collected by taking multiple samples from a single patient and measuring the drug concentration at each time as in the linear kinetics example given below. Instead, two blood samples were drawn at random from each patient at some time during the course of their treatment. This method has developed because subject variables, and beneficial and adverse responses, are routinely recorded in medical charts and can be abstracted for the PPK models. To test the effect of these variables directly in planned experiments would be time consuming and expensive. The statistical methods necessary to obtain the PPK parameter estimates when data are collected in this fashion are discussed briefly below.

The final set of parameters to be estimated are the PD parameters which measure the subject's overall response to the drug. Beneficial parameters are typically proportions of subjects completely cured, partially cured, or relieved of symptoms. Adverse event parameters are proportions of patients experiencing these events. There are other measures of these responses that are continuous, e.g. time to remission, and are dealt with using other models such as **survival analysis**.

Both the PD and the PK parameters are needed here because PDR encompasses more than just the classic dose–response model. In its broadest form it can be seen as the reason for PE investigations: "The objective of many pharmacoepidemiologic investigations is to determine the incidence of an adverse event and identify risk factors associated with the event. These risk factors can include patient characteristics such as age, race, sex, and so forth, and drug characteristics such as the dose, the dosing regimen, and the level of systemic response" [7]. The PK parameters provide a broader spectrum of risk factors with which to evaluate a drug than just dose. Given, as above, that it is the availability of the drug to act in the body and not the dose administered that determines the response that we obtain, we should be able to take advantage of the PK parameters to develop active-concentration (bioavailable) response models and not just dose–response models.

It is not always easy to determine if a response can be directly attributed to the action of a drug. Determining this is difficult enough in a controlled experiment. Given the undesigned nature of the data collection for data used in PDR studies, it becomes even more difficult. Unanticipated responses are particularly hard to detect using only a classic dose–response design. The PK parameters can help to corroborate that a response is related, as one might expect, to one of the PD parameters. For example, chart notations of nausea that occurred within the population limits of a drug's peak concentration could more easily be called an adverse reaction than such notations that could not be linked to this PK measure.

What we know about dose–response from controlled experiments provides only a very small part of the response that we see in the larger population to which the drug is ultimately administered. Incorporating the PK parameters as part of PDR models allows us more latitude to explore the nuances of drug activity in this large population.

Using only the PD, **binary** outcome, of classic dose–response (dead/alive, improved/not improved) does not allow us to classify drugs with similar risk profiles into meaningful risk groups. Classifying drugs based upon how many and what types of responses were related to peak concentration, trough concentration or area under the curve will give us more knowledge about drug activity in general. It will also allow us to develop strategies for separating the beneficial dose–response curve and the adverse event curve more completely. Drugs where benefit depends upon the area under the curve, but where adverse events are only noted if a particular peak concentration is exceeded, can be administered to maximize benefit and minimize adverse results.

The final step in PDR modeling is to bring together all of the information available about all of the parameters mentioned above and form a coherent picture of the activity of a drug in the general population. This final model should allow us to predict the PD responses that will occur in any segment of the population as we manipulate the controllable aspects of drug administration in order to produce the most favorable PK responses.

Estimating the Parameters

Estimates for the values of the PK parameters are made using compartment modeling (*see* **Pharmacokinetics and Pharmacodynamics**), with experimental designs developed specifically for obtaining the PK values [14]. The definitions, figures, and procedures given below are based upon the simplest of these models, a single-compartment model with first-order linear kinetics.

In the simplest single compartment model the drug is almost immediately available from an outside source (bolus injection), resides only in one place in the body (the single compartment which is usually taken to be the plasma), and is then eliminated from the body. A representation of this model is shown in Figure 4.

To estimate the PK parameters, an injection of the drug is given at $t = 0$, where t is the time from injection. Blood is drawn from the subject at pre-specified intervals and the amount of drug still in the plasma is measured. Figure 5 indicates what the subsequent natural log (ln) by time concentration curve might look like for a first-order kinetics model, that is, the concentration curve follows an **exponential distribution** and the log of concentration by time curve will be a straight line.

Algebraically, our PK parameters are defined as follows. The concentration at time t, conc(t), is obtained from (2), although (3) is often easier. The concentration in the plasma immediately after the injection is conc(0). In this simple model

$$\text{conc}(0) = \frac{\text{dose injected}}{\text{volume of plasma}}. \tag{1}$$

Subsequent concentrations can be obtained from the following equations:

$$\text{conc}(t) = \text{conc}(0) \times \exp[-k_{10} \times t] \tag{2}$$

or

$$\ln[\text{conc}(t)] = \ln[\text{conc}(0)] - k_{10} \times t. \tag{3}$$

k_{10} is the constant that indicates how rapidly the drug is excreted from the body (plasma), compartment 1, to the outside world, compartment 0.

It is clear that what drives these relationships is the constant k_{10}. From the curve in Figure 5, k_{10} could be estimated by eye, but a more precise way is to perform a simple **least squares linear regression** on the data in this plot of (3). In this model conc(0) is equal to C_p, which is estimated as the intercept of the regression. The parameter k_{10} is then estimated as the slope of this regression. The half life of the concentration then becomes

$T_{1/2}$ = the time it takes to reduce the concentration

$$\text{by half} = -\frac{\ln(0.5)}{k_{10}}. \tag{4}$$

The value, $-\ln(0.5) = 0.693$. In Figure 5 $T_{1/2}$ can be obtained graphically by determining a 0.693 difference on the vertical axis and projecting this on to the time axis.

If we integrate the exponential function in (2) from 0 to infinity we find that the area under the curve is

$$AUC = \frac{\text{conc}(0)}{k_{10}}. \tag{5}$$

Since theoretically the logarithmic curve will never reach zero, the trough is taken as the lowest predicted value that is obtained before the next injection. When multiple injections are not given, the trough is reported as the lowest predicted value within the time of observation.

The values of these parameters for the data in Figure 5 are found as follows. We inject a bolus of drug at time zero. From blood drawn from the subject at times after $t = 0$, we obtain plasma concentrations. Logarithms of these concentrations are the plotted points in Figure 5. By regression, $C_p = 493$ ng/ml,

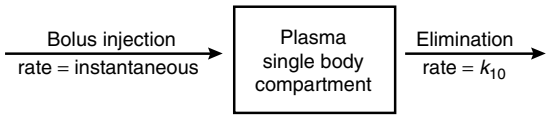

Figure 4 The single-compartment model

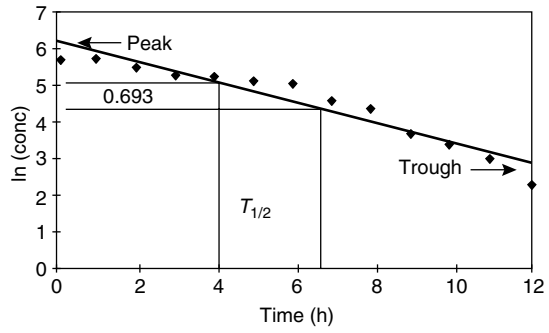

Figure 5 The ln(conc) vs. time curve

trough (at 12 h) = 18.18 ng/ml, and $k_{10} = 0.275$/h. From this we can obtain $T_{1/2} = 2.52$ h, which is also reflected in the drawing in Figure 5. AUC becomes 1127.27 ng h/ml. AUC has units which are difficult to interpret physiologically. It is primarily an intermediate variable which allows us to estimate the system clearance rate as

$$Cl_s = \frac{[\text{conc}(0) \times V_d]}{AUC}. \tag{6}$$

Here V_d is another intermediate variable called the volume of distribution. It is the proportion of drug that remains in the plasma and is not distributed to other body compartments. In this simple model, it equals 1. From the definitions it is clear that in this simple model Cl_s equals $V_d \times k_{10} = 0.275$ ml/h.

The PK models used in practice are much more complicated than those that we have presented here. They use both multiple administrations of a drug and multiple compartments within the body. However, the general approach to the estimation of the PK parameters remains the same.

The example of estimation of PK parameters given above is for data from a single subject. The pharmacoepidemiologic dose–response (PDR) is obtained from need the population values for these PK parameters. This means that we need information from more than one subject, and that we must consider variability in these values both within and between individuals in a population. In addition, we want our population model to provide information about how various subgroups react to the drug. Are older people more inclined to have adverse reactions because they have slower clearance? Is the drug more effective in people with **AIDS** because their condition allows it to reach a higher peak? Can we give a lower dose of the drug (preventing adverse events) because we find that in the general population the minimal effective dose is less than the dose used in the controlled clinical trials used to prove the drug's effectiveness?

Ideally, controlled, planned experiments to answer the above questions should be undertaken. Time and sample size constraints make this unrealistic in most cases. The strategy developed for PDR is to use information that is available in a patient's chart or in a medical database to try to obtain PK estimates that will provide answers to some of these questions [13]. This means that the data collected will be fragmentary, collected under a variety of conditions, unbalanced and subject to many sources of variation. Procedures for analyzing this type of data are currently being developed in the statistical literature. Yuh et al. [16] has an excellent bibliography on current work in this area. The development of standard methods for dealing with these data and appropriate software will greatly increase the effectiveness of PDR research.

As an example of PPK modeling, consider the curves in Figure 6. Here instead of having observations for a single individual for our simple, one-compartment model, we have results from ten individuals. This example is idealized, since PPK models seldom have enough information to estimate individual regression parameters for each individual. Typically, each individual has his or her own response to the drug. In this case, one subject has a response at odds with the other nine subjects. PPK models usually assume that responses of similar subjects are parallel. This nonparallel response should be investigated to see if this subject differs in some meaningful way from the rest of the group. If correction of initial data or adjustments for **confounding** factors is not appropriate, this subject's data should remain in the model. The effect of this decision will be to increase the **variance** of the parameter estimates obtained.

Where the PK example given above had one regression equation we now have ten equations of the following type:

$$\text{conc}(t)_i = \alpha_i \times \exp[-\beta_i \times t] \tag{7}$$

or

$$\ln[\text{conc}(t)_i] = \alpha_i - \beta_i \times t. \tag{8}$$

To emphasize the variability of different subjects to the drug, we have replaced conc(0) by α_i, which is the individual subject's intercept at the time of

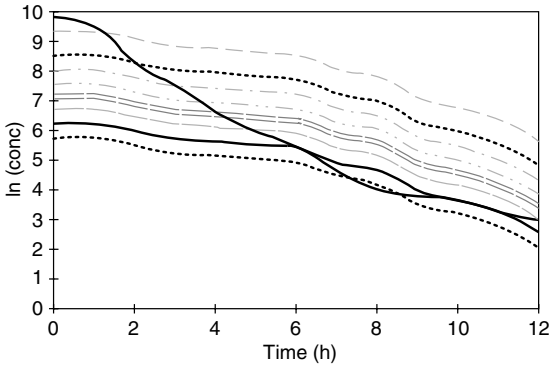

Figure 6 An example of population PK curves

injection. We have also replaced k_{10} by β_i, so that the discussion below is consistent with regression analysis models.

We usually assume that there is enough similarity in the response among individuals that the parameters α_i and β_i can be viewed as representative values from a distribution of such parameters, where the mean of the distribution is the population value α or β. They can then be written as

$$\alpha_i = \alpha + \xi_i, \qquad \beta_i = \beta + e_i. \tag{9}$$

In the simplest case, ξ_i and e_i are assumed to be independently distributed as $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively.

The regression equations required to obtain estimates for α_i and β_i are referred to in various parts of the statistical literature as mixed effects, **stochastic** parameter, or **empirical Bayes** models. Under the appropriate assumptions, these models will yield essentially the same results. Racine-Poon & Smith [12] has a comparison of the models mentioned above. While our simple model is linear using a logarithmic transformation, PK models are usually nonlinear. This increases the difficulty of obtaining the estimates.

Notice that the regression structure provides for much more model complexity than we are using here. In particular, we may want to incorporate some important **covariates** that have their own fixed effect or strata values. If our ten subjects were five males and five females, we might consider having a different α value for each gender, thus removing gender from the undifferentiated sources of variability. It is possible, however, that gender affects the rate of elimination and not the intercept of the concentration curve. In that case we would want a different β for each gender. It is obvious that models which incorporate subject specific covariates into the PK equations can become quite complex. Considerable thought and testing needs to go into the selection of the correct models (*see* **Model, Choice of**) The bibliography in Yuh et al. [16] contains references for the procedures currently being used to obtain estimates for the parameters in these models.

Naively, we might begin by ignoring the fact that these observations came from ten individuals. A single regression could then be performed and the results taken as the values for α and β. This approach ignores the complicated relations of inter- and intraperson variation. Its shortcomings are easily demonstrated if we consider how to deal with different numbers of observations for each subject.

A slightly better approach is first to estimate α_i and β_i from a single linear regression on the data of each individual. These individual estimates can then be combined in some fashion, usually a weighted mean, to produce population estimates of α and β. This approach allows for weighting for different numbers of observations per subject and it provides both individual and population estimates of the parameters. What is missing from this approach is a way of estimating simultaneously both the individual and the population parameters.

Simultaneous estimates of these parameters can proceed along the lines of **maximum likelihood** [10, 11] or **Bayesian** [12] estimation. While these simultaneous estimates are preferred, they too have a drawback. They require assumptions about the variance–covariance structure of these models (*see* **Covariance Matrix**) which are often difficult to make. They also require iterative solutions which may converge slowly to stable estimates. Finally they are currently being estimated using software which is under development. Nevertheless, the advantages which these estimates have in completely specifying the relationship between the population parameters and the individual parameters is thought to be worth the effort.

Computer Software

Several software packages are available for individual PK modeling. A recent paper [3] reviews a number of the programs available in terms of ease of use. In addition, procedures exist within large statistical packages such as SAS, **S-PLUS**, SPSS, and BMDP for mixed-effects analysis and **nonlinear regression** (*see* **Software, Biostatistical**). These procedures can be adapted to produce estimates for many of the individual PK models.

At this time, the only computer program that is specifically designed to deal with the type of data used in PDR is NONMEM [2], a program developed to obtain population estimates of PK parameters from available clinical data. While some studies show that NONMEM provides good estimates in data from Phase III trials [8], controversy still exists about its use in situations with more variability [12]. In addition, NONMEM provides only limited access to model the **variance components**, and it provides only

the population estimates α and β, and not α_i or β_i. It does, however, allow for some limited testing of parameters between different subgroups or strata (*see* **Stratification**) of the population.

Summary

Pharmacoepidemiologic dose–response (PDR) is closely related to the analysis of pharmacokinetic/pharmacodynamic data. In particular, the estimation of population pharmacokinetic parameters is the current analytic tool for PDR. The information that these estimates provide about the dose–response relationships in subgroups of the population is necessary for pharmacoepidemiologic investigation of risk factors associated with drug use. These estimates also provide information for improving the separation between dose–benefit and dose–toxicity curves in the population.

Estimation procedures and associated software for obtaining population pharmacokinetic parameters are still evolving. A bibliography for current estimation procedures is provided in [16] and software reviews and suggested improvements for current software are provided in [3] and [1]. While the procedures in use at this time are providing useful information, this area of pharmacoepidemiologic research will be greatly enhanced when standard methods for design and analysis, and readily available, user friendly, software have been developed.

References

[1] Aarons, L., Balant, L.P., Mentie, F., Morseli, P.L., Rowland, M., Steimer, J.L. & Vozel, S. (1994). Population approaches in drug development: report on an expert meeting to discuss population pharmacokinetics/pharmacodynamics software, *European Journal of Clinical Pharmacology* **46**, 389–391.

[2] Beal, S.L. & Sheiner, L.B. (1980). The NONMEM system, *American Statistician* **34**, 118–119.

[3] Buffington, D.E., Lampasona, V. & Chandler, M.H.H. (1993). Computers in pharmacokinetics. Choosing software for clinical decision making, *Clinical Pharmacokinetics* **25**, 205–216.

[4] DeVane, C.L., Grasela, T.H., Antal, E.J. & Miller, R.L. (1993). Evaluation of population pharmacokinetics in therapeutic trials, IV. Application to postmarketing surveillance, *Clinical Pharmacology and Therapeutics* **53**, 521–528.

[5] Finney, D.J. (1978). *Statistical Method in Biological Assay*, 3rd Ed. Griffin, London.

[6] Gibaldi, M. & Perrier, D. (1982). *Pharmacokinetics*. Marcel Dekker, New York.

[7] Grasela, T.H., Jr (1994). The role of therapeutic drug monitoring in pharmacoepidemiology, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, New York, pp. 413–430.

[8] Grasela, T.H., Jr, Antal, E.J., Townsend, R.J. & Smith, R.B. (1986). An evaluation of population pharmacokinetics in therapeutic trials, part I. Comparison of methodologies, *Clinical Pharmacology and Therapeutics* **39**, 605–612.

[9] Henry, D.A., Smith, A.J. & Hennessy, S. (1994). Basic principles of clinical pharmacology relevant in PE studies, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, New York, pp. 39–56.

[10] Laird, N.M. & Ware, J.H. (1982). Random effects models for longitudinal data, *Biometrics* **38**, 963–974.

[11] Lindstrom, M.J. & Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics* **46**, 673–687.

[12] Racine-Poon, A. & Smith, A.F.M. (1990). Population models, in *Statistical Methodology in the Pharmaceutical Sciences*, D.A. Berry, ed. Marcel Dekker, New York, pp. 146–147.

[13] Sheiner, L.B., Rosenberg, B. & Marathe, V.V. (1977). Estimation of population characteristics of pharmacokinetic parameters from routine clinical trials, *Journal of Pharmacokinetics and Biopharmaceutics* **5**, 445–479.

[14] Steimer, J., Ebelin, M. & Van Bree, J. (1993). Pharmacokinetic and pharmacodynamic data and models in clinical trials, *European Journal of Drug Metabolism and Pharmacokinetics* **18**, 67–76.

[15] Vestral, R., Wood, A. & Shand, D. (1979). Reduced beta adrenoceptor sensitivity in the elderly, *Clinical Pharmacology and Therapeutics* **26**, 181–186.

[16] Yuh, L., Beal, S., Davidian, M., Harrison, F., Hester, A., Kowalski, K., Vonesh, E. & Wolfinger, R. (1994). Population pharmacokinetics/pharmacodynamics methodology and applications: a bibliography, *Biometrics* **50**, 566–575.

(*See also* **Drug Interactions**; **Pharmacoepidemiology, Adverse and Beneficial Effects**; **Pharmacoepidemiology, Overview**)

J.R. MURPHY

Dose–Response Models in Risk Analysis

In biological and health sciences, a **dose–response** model is an expression that describes a measure of a biological or health effect as a mathematical function of a dose of a substance. In risk analysis (*see* **Risk Assessment**; **Risk Assessment for Environmental Chemicals**), the biological or health effect is generally a measure of an adverse outcome, such as the proportion of individuals with a disease or the magnitude of body weight loss. The primary goal is to estimate the proportion of individuals in a population that are expected to experience an adverse health effect when exposed to a specified biological insult, usually a chemical substance. The dose is the amount of the chemically active substance at the affected tissue site. Often the body converts a chemical into other forms (metabolites) that are the biologically active substances. Occasionally, elaborate differential equations based on physiology can be constructed to provide **pharmacokinetic models** to estimate the doses of the active chemical substances at the affected tissue sites. In the absence of this information, the amount of the active chemical is generally assumed to be proportional to the dose of a chemical that is administered by inhalation, dermal, or oral exposure. Since various species of animals tend to react similarly when a dose is administered on a body weight basis, dose often is expressed as milligrams (mg) of a substance per kilogram (kg) of body weight. For many biological effects, similar effects among species are noted when body weight is raised to a power; for example, 2/3 or 3/4. This often corresponds closely to a dose expressed as a concentration; for example, parts per million (ppm) in air, food, or water.

For risk analysis, dose–response data seldom are available for exposures of humans to specific toxic substances. Hence, most dose–response data are obtained from well-controlled animal **bioassays**. Typically, relatively small numbers of animals (generally 5–50) are exposed to a few dose levels (generally 3–5), and there is a group of control animals that are not exposed to the chemical substance under investigation. Generally, the purpose of the bioassay is to establish a dose–response function from which a relatively "safe" dose for human exposure can be estimated. Where the biological mechanism of toxicity is known, it may be possible to represent this situation with a biologically based dose–response model that will provide sufficiently precise estimates of risk at human exposure levels.

Generally, it is desired to limit risk levels (the proportion of adversely affected individuals) to less than 1 in 10 000. Hundreds of thousands of animals would be required to measure such levels of risk directly with precision. Since the resources to do this do not exist, relatively small numbers of animals are exposed to doses well above human exposure levels in order to elicit potential toxic effects. Then, it becomes necessary to extrapolate the results from high-dose experiments to low-dose human exposure levels. That is, the bioassay data are used only to obtain a dose–response model in the experimental dose range and some form of low-dose extrapolation is employed (*see* **Extrapolation, Low Dose**).

Various dose–response models have been employed for quantal (dichotomous) endpoints, where an animal is classified as either possessing an adverse effect (such as cancer or a birth defect) or considered normal. For quantal data, it has been common to use tolerance distributions to relate the probability of disease to dose (*see* **Quantal Response Models**). It is assumed that the probability that an animal will develop an adverse effect at a dose, d, is described by a specified distribution. For example, if the probability density function of tolerated doses is normal (Gaussian), then the probability of an effect at dose d is given by the cumulative normal distribution function (probit):

$$\Pr(d) = \Phi(\beta_0 + \beta_1 d),$$

where the regression coefficients, β_0 and β_1, are estimated from bioassay data [5]. More commonly, for biological endpoints the **lognormal** tolerance distribution provides a better description (log probit):

$$\Pr(d) = \Phi(\beta_0 + \beta_1 \ln d).$$

A second model used extensively for quantal responses is the logistic [2]:

$$\Pr(d) = \{1 + \exp[-(\beta_0 + \beta_1 \ln d)]\}^{-1}.$$

Chand & Hoel [4] showed that when the time-to-tumor distribution is a **Weibull distribution**, the

dose–response model follows an **extreme value** model:

$$\Pr(d) = 1 - \exp[-\exp(\beta_0 + \beta_1 \ln d)].$$

If k similar events are required to produce an adverse biological effect, the probability of this effect is given by the multi-hit model based on the **Poisson distribution**:

$$\Pr(d) = 1 - \sum_{i=0}^{k-1} \frac{(\lambda d)^i \exp(-\lambda d)}{i!}.$$

When λd is small, the multi-hit model is approximately

$$\Pr(d) = \beta d^k \quad \text{or} \quad \ln \Pr(d) = \ln \beta + k \ln d.$$

In the special case in which only one event is necessary to produce an adverse effect (for example, a single mutation in a cell), the resulting model is the one-hit model

$$\Pr(d) = 1 - \exp(-\lambda d).$$

When λd is small, the one-hit model is approximately linear:

$$\Pr(d) = \lambda d.$$

In the above models in which $\ln d$ is used, when $d = 0$ ($\ln 0 = -\infty$), the background rate is $\Pr(0) = 0$. This is often not the case. If a chemical acts independently of the background, the total (background plus induced) probability of an event is

$$\overset{*}{\Pr}(d) = \Pr(0) + [1 - \Pr(0)] \Pr(d).$$

If a chemical adds to the background dose (that is, the background rate is due to an effective dose d_0), then

$$\overset{*}{\Pr}(d) = \Pr(d + d_0).$$

The Armitage–Doll **multistage carcinogenesis model** [1] is widely used to represent the probability of cancer by a given age as a function of dose,

$$\overset{*}{\Pr}(d) = 1 - \exp\left[-\beta_0 \prod_{i=1}^{k} (\beta_{0i} + \beta_{1i} d)\right],$$

where it is assumed that a cell progresses through k stages leading to cancer and that the rate of change of the ith stage is $\beta_{0i} + \beta_{1i} d$, with $\beta_{0i} > 0$ and $\beta_{1i} \geq 0$ for $i = 1, 2, \ldots, k$. The parameter β_{0i} represents the spontaneous background rate and $\beta_{1i} d$ represents the increase in rate of the ith stage due to dose of a carcinogen; β_0 is constant greater than zero. The Armitage–Doll multistage model is generally expressed in a polynomial form,

$$\overset{*}{\Pr}(d) = 1 - \exp\left(-\sum_{i=0}^{k} \beta_i d^i\right),$$

where the constraints $\beta_i \geq 0$ are usually employed. The background tumor incidence is $\Pr^*(0) = [1 - \exp(-\beta_0)]$. In the special case in which the rate of only one stage is increased by a carcinogen, the one-hit model is obtained:

$$\overset{*}{\Pr}(d) = 1 - \exp[-(\beta_0 + \beta_i d)].$$

The Weibull model is a special case of the polynomial form:

$$\overset{*}{\Pr}(d) = 1 - \exp[-(\beta_0 + \beta_1 d^k)].$$

In studies of such duration where death occurs, death from causes other than the specific disease of interest can result in distorted dose–response curves. For example, suppose that mortality increases with increasing doses from causes other than the disease of interest. For a later occurring disease such as cancer, the incidence of cancer at high doses many decrease because the animals die from other causes before the disease is observed. Hence, tumor rates adjusted for intercurrent mortality are required to obtain unbiased dose–response functions. This is accomplished by adjusting the number of animals at risk. Kaplan & Meier [7] provide a nonparametric procedure for estimating the incidence of fatal tumors as a function of time (*see* **Kaplan–Meier Estimator**). Hoel & Walburg [6] present a method for estimating the prevalence of nonfatal (incidental) tumors as a function of time. Kodell et al. [8] combine these two procedures to provide nonparametric estimates for tumor onset. These procedures require assigning a tumor to the most likely category, fatal or nonfatal. All tumors observed as the result of a scheduled sacrifice of animals are considered incidental, even though they may generally progress to cause the death of an animal (*see* **Serial-sacrifice Experiments**).

Moolgavkar & Venzon [13] and Moolgavkar & Knudson [12] propose a **two-mutation carcinogenesis model** that includes proliferation of initiated cells.

In this model a normal cell may mutate to an initiated cell, an initiated cell may undergo clonal expansion, and an initiated cell may be transformed by a second mutation to a cancer cell. The probability that an individual or animal has a particular type of cancer by age t is given by

$$\Pr(t) = 1 - \left[\frac{2C\exp(-1/2(B+C)t)}{(B+C)\exp(-Ct) - (B-C)}\right]^{\nu/\beta}$$

where $C = [(\beta+\delta+\mu)^2 - 4\beta\delta]^{1/2}$, $B = \beta - \delta - \mu$ and the cell kinetic parameters ν, β, δ, and μ are defined as follows. It is assumed that the number of normal cells in a tissue that are at risk of producing initiated cells upon cell division remains relatively constant. Consequently, normal cells generate initiated cells by a **Poisson process** with constant intensity ν. An initiated cell living at time t may undergo cell division producing two initiated cells in the time interval $(t, t+\Delta t)$ with an approximate probability of $\beta\Delta t$, and die (either physiologically or pathologically) with an approximate probability of $\delta\Delta t$, where Δt is small. In this time interval, a living initiated cell may also divide into an initiated cell and a cancer cell with approximate probability of $\mu\Delta t$. New research in this area is increasing the number of stages, considering cell kinetics that change with age, incorporating DNA repair, considering specific events at the nucleotide level, and incorporating dose.

For continuous (nonquantal) data, a multitude of dose–response curves is used. It is desirable that the models reflect pharmacokinetics and physiology. For example, it might be hypothesized that limits are approached due to saturation of biological processes. Once a dose–response model is obtained to describe average levels, the distribution of individual levels about the average is needed for risk estimation. Some biological measures tend to follow a normal distribution (for example, body and organ weights). Frequently, biological measures appear to be lognormally distributed. In these cases, only the standard deviation is needed in addition to the average to describe the distribution of measures of biological effects. If an adverse level has been identified, then the probability of individuals being in the adverse range can be estimated as a function of dose. Where an adverse range is not specified, the probability of abnormal levels (for example, below the first percentile and/or above the 99th percentile of unexposed control individuals) can be estimated as a function of dose.

In reproductive and developmental studies, a chemical is administered to the father and/or mother before or during pregnancy and the results are evaluated in the fetuses or offspring. Hence the litter is the experimental unit. Common endpoints measured near the end of pregnancy are the number of implants per litter, the proportion of implants that result in resorptions or dead fetuses per litter, the proportion of specific types of malformations per litter among the live fetuses, and average fetal weight per litter.

One of the first models devised specifically for use with developmental data was provided by Rai & Van Ryzin [14]. Their model contained the feature that the probability of a malformed fetus is a function of litter size (s) as well as dose:

$$\overset{*}{\Pr}(d, s) = \{1 - \exp[-(\beta_0 + \beta_1 d)]\} \times \exp[-s(\alpha_0 + \alpha_1 d)],$$

where $\beta_0 > 0$, $\beta_1 > 0$, $\alpha_0 > 0$, and $(\alpha_0 + \alpha_1 d) > 0$. Rai & Van Ryzin [14] postulated that larger litters are an indication of healthier litters, so that the probability of a malformation should decrease with increasing litter sizes.

Kupper et al. [10] used a log-logistic dose–response model for litter-type data. A generalization of their model to include litter size s is

$$\overset{*}{\Pr}(d, s) = \beta_0 + \beta_1 s + \frac{1 - \beta_0 + \beta_1 s}{1 + \exp[\beta_2 + \beta_3 s - \beta_4 \log(d - d_0)]}$$

Kodell et al. [9] provide a model for the results of a bioassay with g dose groups (indexed by $i = 1, \ldots, g$), and n_i pregnant females in the ith group receiving a dose d_i. Let x_{ij} denote the number of affected fetuses out of the s_{ij} fetuses in the jth litter of the ith dose group. Kodell et al. [9] use the **beta-binomial distribution** for the x_{ij} given a fixed litter size s_{ij}. It is assumed that the occurrence of an affected fetus in a litter follows a binomial distribution with probability p_{ij}, while the p_{ij} within the ith dose group are distributed according to a beta distribution. The probability of an adverse effect for a fetus in a litter of size s_{ij} at dose d_i is

$$\Pr(d_i, s_{ij}) = 1 - \exp\{-[\beta_0 + \beta_1(s_{ij} - s)]\},$$

for d_i less than or equal to a threshold dose of d_0. For doses above the threshold dose,

$$\overset{*}{\Pr}(d_i, s_{ij}) = 1 - \exp\{-[\beta_0 + \beta_1(s_{ij} - s) + (\beta_2 + \beta_3(s_{ij} - s))(d_i - d_0)^k]\},$$

where s is the average litter size over all dose groups, with $\beta_0 \geq 0$, $\beta_2 \geq 0$, $d_0 \geq 0$, $k \geq 1$, $[\beta_0 + \beta_1(s_{ij} - s)] \geq 0$ for all s_{ij}, and $[\beta_2 + \beta_3(s_{ij} - s)] \geq 0$ for all s_{ij}. The parameters of this Weibull-type model can be estimated by the method of maximum likelihood.

If death and malformation occur independently, the number of dead, malformed, and normal fetuses in a litter can be modeled as a trinomial distribution (*see* **Multinomial Distribution**). In order to account for the **overdispersion** induced by litter effects, the **quasi-likelihood** method of McCullagh & Nelder [11] that inflates the standard multinomial variance can be used. Ryan [15] discusses techniques for estimating the various probabilities assuming a log-probit and logistic model to describe the probabilities as a function of dose. Catalano et al. [3] also include a continuous measurement (fetal weight) along with the quantal endpoints of death and malformation in a multivariate model.

The beta–Poisson model is frequently used for microbial risk assessment. The probability of infection is

$$\overset{*}{\Pr}(d) = 1 - (\beta_0 + \beta_1 d)^{-k},$$

where d is the dose of an organism (generally expressed as number per mass or volume). As k approaches infinity, the dose–response curve approaches the exponential.

References

[1] Armitage, P. & Doll, R. (1961). Stochastic models for carcinogenesis, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, L.M. LeCam & J. Neyman, eds. University of California Press, Berkeley.

[2] Berkson, J. (1994). Application of the logistic function to bioassay, *Journal of the American Statistical Association* **39**, 357–365.

[3] Catalano, P.J., Scharfstein, D.O., Ryan, L.M., Kimmel, C.A. & Kimmel, G.L. (1993). Statistical model for fetal death, fetal weight, and malformation, *Teratology* **47**, 281–290.

[4] Chand, N. & Hoel, D.G. (1974). A comparison of models for determining safe levels of environmental agents, in *Reliability and Biometry: Statistical Analysis of Lifelength*, F. Proschan & R. Serfling, eds. SIAM., Philadelphia.

[5] Finney, D.J. (1952). *Statistical Method in Biological Assay*. Hafner, New York.

[6] Hoel, D.G. & Walburg, H.E. (1972). Statistical analysis of survival experiments, *Journal of the National Cancer Institute* **49**, 361–372.

[7] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[8] Kodell, R.L., Shaw, G.W. & Johnson, A.M. (1982). Nonparametric joint estimators for disease resistance and survival functions in survival/sacrifice experiments, *Biometrics* **38**, 43–58.

[9] Kodell, R.L., Howe, R.B., Chen, J.J. & Gaylor, D.W. (1991). Mathematical modelling of reproductive and developmental toxic effects for quantitative risk assessment, *Risk Analysis* **11**, 583–590.

[10] Kupper, L., Portier, C., Hogan, M. & Yamamoto, E. (1986). The impact of litter effects on dose–response modeling in teratology, *Biometrics* **42**, 85–98.

[11] McCullagh, P. & Nelder, T.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[12] Moolgavkar, S.H. & Knudson, A. (1981). Mutation and cancer: a model for human carcinogenesis, *Journal of the National Cancer Institute* **66**, 1037–1052.

[13] Moolgavkar, S.H. & Venzon, D.J. (1979). Two-event models for carcinogenesis: incidence curves for childhood and adult tumors, *Mathematical Biosciences* **47**, 55–77.

[14] Rai, K. & Van Ryzin, J. (1985). A dose response model for teratological experiments involving quantal responses, *Biometrics* **41**, 1–9.

[15] Ryan, L. (1992). Quantitative risk assessment for developmental toxicity, *Biometrics* **48**, 163–174.

DAVID W. GAYLOR

Dose–Response

Dose–response refers to a relationship between an amount of exposure or treatment and the degree or probability of an outcome in an individual or population. The dose may represent the amount, duration or intensity of exposure or treatment, and the outcome may represent a favorable effect, such as lowering of elevated blood pressure, or an unfavorable effect, such as increased risk of developing cancer. For example, the risk of lung cancer is known to increase with the number of cigarettes smoked each day and with the duration of smoking. A monotonic relationship of increasing disease risk with increasing exposure is often taken as one indication of a causal relationship between exposure and risk (*see* **Hill's Criteria for Causality**).

(*See also* **Dose-response in Pharmacoepidemiology**; **Dose–Response Models in Risk Analysis**)

MITCHELL H. GAIL

Double Sampling

Double sampling procedures are widely used in sample surveys, in acceptance sampling, for quality control, and also for statistical tests of hypotheses (*see* **Hypothesis Testing**). Hill [36] presents a review of the literature on this method up to 1980, and Hewett & Spurrier [35] summarize the double sampling procedures for tests of hypotheses.

This article contains a review of the developments of this approach for finite population sampling, and related topics. In finite population sampling, double sampling is widely used for determining strata weights (*see* **Stratified Sampling**), estimating the mean of the auxiliary variable for **ratio and regression estimates**, finding sample sizes for comparing the means of domains or subgroups (*see* **Sample Size Determination**), estimating and comparing proportions and percentages from misclassified observations (*see* **Misclassification Error**), and similar purposes. Double sampling is also known as *two-phase sampling*; it should be distinguished from two-stage sampling, where a sample of clusters are selected at the first stage and a sample of elements are chosen at the second stage.

We first present some applications of the double sampling procedures, followed by stratification, ratio and regression methods, and other procedures that employ double sampling.

Selected Applications

Double sampling was employed by Hall [30] for estimating the number of transmission sources in a **Poisson process**, by Ahmed et al. [2] and Catchpole & Catchpole [11, 12] for biomass estimates, by Stockford & Page [96] and Potter et al. [60] for estimation related to the Vietnam service, by Crete et al. [21] for correcting helicopter counts of moose, by Eberhardt & Simmons [25] for estimating sizes of animal populations, by Fairley et al. [27] for estimation related to welfare programs, by Baker [5] for evaluating a new medical diagnostic test, by Heerschop & Lieftinck-Koeijers [34] for estimating employment rates, and by Oderwald [56], Oderwald & Jones [57], and Reich et al. [72] for estimation in forestry. Magden & Holstein [49] use the double sampling method for determining the sample size for estimation related to rare items.

Stratification

Consider a finite population of size N divided into L strata of sizes N_h, $h = 1, 2, \ldots, L$; $N = \sum N_h$. From samples of sizes n_h, $n = \sum n_h$, selected randomly without replacement (*see* **Sampling With and Without Replacement**) from the strata, the estimator for the population **mean** $\overline{Y}$ is

$$\overline{y}_{\text{st}} = \sum W_h \overline{y}_h,$$

where $W_h = N_h/N$ are the weights and $\overline{y}_h$ are the sample means.

In some situations, only the sample means are reported, but the weights are not available. In such cases, as recommended by Neyman [55], initially a large sample of size n' is selected without replacement from the N population units. With the observed sample size n'_h in the hth stratum, $n' = \sum n'_h$, W_h is estimated from $w_h = n'_h/n'$. At the second phase, a sample of size $n_h = \nu_h n'_h$, $0 \le \nu_h \le 1$, is selected without replacement from the n'_h units and the sample mean $\overline{y}_h$ is obtained. The population mean is now estimated from

$$\overline{y}_{\text{st(d)}} = \sum w_h \overline{y}_h.$$

The estimator of this **variance** and its variance are presented by Cochran [18, p. 333] and Rao [63].

The cost of selecting the samples at the two phases is considered to be of the form $C = c'n' + \sum c_h n_h$. Optimum n' and n_h for minimizing the **variance** of $\overline{y}_{\text{st(d)}} = \sum w_h \overline{y}_h$ for a given average cost or for minimizing the average cost for a given variance are presented by Cochran [18, p. 331].

An alternative method for determining n_h was suggested by Srinath [93] and Rao [63], and its consequences were discussed by the author, Rao [68]. Treder & Sedransk [105] also suggest a method of determining the sample sizes at the two phases. Instead of selecting n_h proportional to n'_h, Singh & Singh [81] suggest three alternatives: (i) select the n_h units *with* replacement from the n'_h units, (ii) select the n_h units *with* replacement, but consider only the distinct units, and (iii) for the sample size at the second phase, consider the minimum of $(n_h,\ n'_h)$.

Ratio Estimator

The ratio estimator for $\overline{Y}$ is

$$\widehat{\overline{Y}}_{\mathrm{R}} = \left(\frac{\overline{y}}{\overline{x}}\right)\overline{X},$$

where $(\overline{x}, \overline{y})$ are the means of a sample of n units selected randomly without replacement from the N units, and $\overline{X}$ is the population mean of an auxiliary variable. The double sampling procedure is implemented when $\overline{X}$ is unknown. In this situation, a large and inexpensive sample of n_1 units is first selected without replacement from the N units. The mean $\overline{x}_1$ of these n_1 units provides an **unbiased** estimator for $\overline{X}$. At the second phase, a sample of size n is selected without replacement from the n_1 units and the means $(\overline{x}, \overline{y})$ are obtained. The ratio estimator for $\overline{Y}$ is now

$$\widehat{\overline{Y}}_{\mathrm{Rd}} = \left(\frac{\overline{y}}{\overline{x}}\right)\overline{x}_1.$$

With a cost function of the form $C = c_1 n_1 + cn$, optimum values of n_1 and n can be determined to minimize the cost or variance.

Different procedures, including the **jackknife**, are available for reducing the **bias** of $\widehat{\overline{Y}}_{\mathrm{R}}$. Rao [67] investigates the merits of eight modifications and extensions of these procedures for estimating $\overline{Y}$ through the ratio method and the double sampling procedure.

In some situations, the second sample of n units is selected independently of the first sample. For such a case, Rao [65, 66] considers estimating $\overline{X}$ from $\overline{x}^* = a\overline{x}_1 + (1-a)\overline{x}$ or the mean $\overline{x}_\nu$ of the ν distinct units of the two samples, and compares

$$\widehat{\overline{Y}}^*_{\mathrm{Rd}} = \left(\frac{\overline{y}}{\overline{x}}\right)\overline{x}^*$$

and

$$\widehat{\overline{Y}}_{\mathrm{R}\nu} = \left(\frac{\overline{y}}{\overline{x}}\right)\overline{x}_\nu,$$

with $\widehat{\overline{Y}}_{\mathrm{Rd}}$; the constant a is chosen to minimize the variance of $\overline{x}'$.

With stratification, the combined (RC) and separate (RS) ratio estimators for $\overline{Y}$ are

$$\widehat{\overline{Y}}_{\mathrm{RC}} = \sum W_h \left(\frac{\overline{y}_h}{\overline{x}_h}\right)\overline{X}_h$$

and

$$\widehat{\overline{Y}}_{\mathrm{RS}} = \left(\frac{\overline{y}_{\mathrm{st}}}{\overline{x}_{\mathrm{st}}}\right)\overline{X},$$

respectively. In these expressions, $\overline{X}_h$ is the mean of the auxiliary variable of the N_h units of the hth stratum, $\overline{x}_h$ is the mean of a sample of n_h units selected randomly without replacement from the N_h units, and $\overline{x}_{\mathrm{st}} = \sum W_h \overline{x}_h$. If $\overline{X}_h$ and $\overline{X}$ are not known, then the above estimators can be obtained through the double sampling procedure. Tripathi & Bahl [107] extend such a procedure to the case of more than one auxiliary variable. Ratio estimation with stratification was also considered by Ige & Tripathi [39], and with **poststratification** by Sethi & Srivastava [79] and White [110].

For the chain ratio estimator of $\overline{Y}$, additional information on a variable which is inexpensive to measure but less correlated with y than x is utilized. Srivastava et al. [94, 95] and Singh & Singh [88] apply the double sampling procedure to this method of ratio estimation.

Regression Estimator

The **linear regression** estimator for $\overline{Y}$ is

$$\widehat{\overline{Y}}_{\mathrm{lr}} = \overline{y} + b(\overline{X} - \overline{x}),$$

where b is the regression coefficient obtained from the sample of size n selected randomly without replacement from the N units. As in the case of the ratio estimator, for the double sampling regression estimator the unknown $\overline{X}$ is replaced by the mean $\overline{x}_1$ of the first sample. Cochran [18, Chapter 12] presents the variance of this estimator, the estimator of variance, and the optimum sample sizes for minimizing the variance or the cost. A model-based estimate for the variance is presented by Dorfman [24]. Bellhouse & Joshi [7] examine the admissibility of the double sampling regression estimator.

Selection of the first and second samples independently was considered by Bose [9]. For this case, Rao [64] considers replacing $\overline{X}$ with $\overline{x}^*$ or $\overline{x}_\nu$. Tikkiwal [104] considers the second sample to be selected from the $(N - n_1)$ units.

For a large population, Han [31] first tests whether $\overline{X}$ is zero. If this hypothesis is rejected, then the above estimator can be used for estimating $\overline{Y}$; otherwise, $\overline{X}$ is set to zero. Esimai & Han [26] extend this

procedure to more than one auxiliary variable. Sisodia & Srivastava [91] consider an extension of this procedure by considering two specified alternative values for $\overline{X}$.

The regression method has been extended to more than one auxiliary variable, for example by Kiregyera [43] and Singh [83] for two auxiliary variables. Conniffe [19] considers the double sampling estimator with unequal regression coefficients. Matloff [52] employs the double sampling procedure for **nonlinear regression** and Baker et al. [6] for **logistic regression**. Tamhane [100] uses the double sampling procedure for the regression method and tests of hypothesis. Prediction through the double sampling ratio and regression methods was examined by Agrawal & Jain [1]. Singh & Singh [87] consider the double sampling chain regression estimator. Conniffe & Moran [20] use the double sampling regression estimator for comparative studies.

Multivariate ratio and regression estimators are formed by including information for more than one auxiliary variable. Tripathi [106] and Kiregyera [42] consider the multivariate chain ratio estimator with double sampling.

Further topics related to the regression or ratio methods appear in [6, 17, 53, 59, 88, 100, 108], and [109].

Analytic Surveys

In these types of surveys, importance is given to the determination of the sample sizes for estimating the differences of means of domains or subpopulations. Sedransk [75, 76] and Booth & Sedransk [8] present the double sampling procedures for finding the sample sizes for minimizing the variances or costs.

Nonresponse

When only n_1 of the n sampled units respond, Hansen & Hurwitz [32] suggested subsampling m of the $n_1 = n - n_2$ nonrespondents (*see* **Nonresponse**). The population mean is now estimated from

$$\widehat{\overline{Y}}_{\mathrm{H}} = \frac{(n_1\overline{y}_1 + n_2\overline{y}_{2m})}{n},$$

where $\overline{y}_1$ and $\overline{y}_{2m}$ are the means of the respondents and the subsampled units. If $\overline{X}$ is known, the ratio estimator for the mean is

$$\widehat{\overline{Y}}_{\mathrm{HR}} = \left(\frac{\widehat{\overline{Y}}_{\mathrm{H}}}{\widehat{\overline{X}}_{\mathrm{H}}}\right)\overline{X},$$

where $\widehat{\overline{X}}_{\mathrm{H}}$ for the auxiliary characteristic is defined analogous to $\widehat{\overline{Y}}_{\mathrm{H}}$. When $\overline{X}$ is unknown, Rao [69, 70] constructs estimators for $\overline{Y}$ through the ratio method and also extends them to the regression method.

Multistage Sampling

In two-stage sampling, cluster or primary sampling units (PSUs) are selected at the first stage and elements or secondary sampling units (SSUs) are chosen at the second stage from the units selected at the first stage (*see* **Cluster Sampling**; **Multistage Sampling**). **Stratification** may be implemented at either stage and ratio or regression methods of estimation may be employed. In multistage sampling, this procedure is extended to several stages. Double sampling for multistage sampling was considered by Robson [73], Robson & King [74], Garg & Pillai [28], and others.

Successive Sampling

For some types of survey, samples are selected at two or more periods of time. Stratification and ratio and regression methods of estimation are also used in these types of survey. Double sampling procedures for successive sampling are considered by Arnab & Okafor [4], Singh & Singh [82], Sen et al. [77], Sisodia [89], and Lamba & Singh [45].

Classification Errors

In some applications, errors made in observing or measuring a unit can lead to its classification into the wrong group. In medical diagnosis, these misclassifications are known as **false positives** and **false negatives**. For estimating **binomial** and **multinomial** proportions, Tenenbein [101–103] first considers inexpensive and less than perfect measurements on a sample of units and then expensive and accurate measurements on a subsample of them; the proportions are estimated from the observations of both the samples.

Nedelman [54] uses a double sampling method for estimating the **prevalence** of malaria from the inaccurate observations. Lie et al. [48] estimate the percentage of congenital malformations from double registrations. The double sampling procedure was considered by Hochberg [38] for analyzing misclassified categorical data, by Mak & Li [51] for estimating subgroup means, by Chernoff & Haitovsky [16] and DeWith [22] for comparing two binomial probabilities, and by Swensen [97] for estimating the change in a probability.

Jolayemi [40] considers the above type of estimation for multiple outcomes and misclassifications. Chen [15], Korn [44], and Palmgren [58] also consider it for misclassified categorical data.

Further Topics

Des Raj [23] considers double sampling for **sampling with probability proportionate to size (PPS)**. Bellhouse & Joshi [7] examine the double sampling and regression method of estimation for PPS sampling. Chaudhuri & Adhikary [13, 14] examine varying probability of selection with double sampling.

A double sampling procedure was suggested by J. N. K. Rao [62] for estimation when the **sampling frame** contains duplications and there is nonresponse in the selected sample. Hinkins & Scheuren [37] apply the "hot deck" imputation when nonresponse occurs for a double sampling procedure (*see* **Missing Data Estimation, "Hot Deck" and "Cold Deck"**).

The difference estimator is the regression estimator presented earlier with the regression coefficient set to unity. This method of estimation with double sampling was considered by Talukder [99] and Singh & Singh [87].

The product estimator for the mean is $\hat{\overline{Y}} = (\overline{xy})/\overline{X}$. For large samples, the **mean square error** of this estimator can be smaller than the variance of $\overline{y}$ if the **correlation** of x and y is negatively large. The ratio and product methods can be combined if some of the auxiliary variables are positively correlated with y and some are negatively correlated. For estimating the population mean of the auxiliary variables, double sampling procedures were considered by Ray & Singh [71], Sisodia & Dwivedi [90], and Singh [84, 85].

The jackknife method for reducing the biases of the double sampling ratio and product estimators was examined by Sengupta [78]. **Bayes** estimation with double sampling was considered by Chaudhuri & Adhikary [13, 14], Geng & Asano [29], and Smith [92]. Ahuja & Srivastava [3] consider the double sampling method for **systematic sampling**.

Double sampling for the measurement of process bias was considered by Lessler [46], for the response bias by Talukder [99], and for nonresponse at the second phase with stratification by Swensson [98]. Kawatheker [41] considers a modified ratio estimator with double sampling, and Li et al. [47] consider double sampling for estimating strata means with corrections based on the second phase sampling. Further results on double sampling appear in [10, 33, 50, 61, 80, 83, 108], and [109].

References

[1] Agrawal, M.C. & Jain, N. (1988). Predictive estimation in double sampling procedures, *American Statistician* **42**, 184–186.

[2] Ahmed, J.B., Charles, D. & Laycock, W.A. (1983). Comparison of techniques used for adjusting biomass estimates by double sampling, *Journal of Range Management* **36**, 217–221.

[3] Ahuja, D.L. & Srivastava, A.K. (1988). Sampling from two-dimensional populations spread over space and time, *Journal of the Indian Society of Agricultural Statistics* **40**, 83–95.

[4] Arnab, R. & Okafor, F.A. (1992). A note on double sampling over two occasions, *Pakistan Journal of Statistics, Series B* **8**, 9–18.

[5] Baker, S.G. (1991). Evaluating a new test using a reference test with estimated sensitivity and specificity, *Communications in Statistics–Theory and Methods* **20**, 2739–2752.

[6] Baker, S.G., Wax, Y. & Patterson, B.H. (1993). Regression analysis of grouped survival data. Informative censoring and double sampling, *Biometrics* **49**, 379–389.

[7] Bellhouse, D.R. & Joshi, V.M. (1984). On the admissibility of regression estimator, *Journal of the Royal Statistical Society, Series B* **46**, 268–269.

[8] Booth, G. & Sedransk, J. (1969). Planning some two-factor comparative surveys, *Journal of the American Statistical Association* **64**, 560–573.

[9] Bose, C. (1943). Note on the sampling errors in the method of double sampling, *Sankhyā* **6**, 330.

[10] Buonaccorsi, J.P. (1990). Double sampling for exact values in some multivariate measurement error problems, *Journal of the American Statistical Association* **85**, 1075–1082; **86**, 837.

[11] Catchpole, W.R. & Catchpole, E.A. (1991). Estimating biomass in a vegetation mosaic using double sampling with regression, *Australian Journal of Statistics* **33**, 279–289.

[12] Catchpole, W.R. & Catchpole, E.A. (1993). Stratified double sampling of patchy vegetation to estimate biomass, *Biometrics* **49**, 295–303.

[13] Chaudhuri, A. & Adhikary, A.K. (1983). On optimality of double sampling strategies with varying probabilities, *Journal of Statistical Planning and Inference* **8**, 257–265.

[14] Chaudhuri, A. & Adhikary, A.K. (1985). Some results on admissibility and uniform admissibility in double sampling, *Journal of Statistical Planning and Inference* **12**, 199–202.

[15] Chen, T.T. (1979). Log-linear models for categorical data with misclassification and double sampling, *Journal of the American Statistical Association* **74**, 481–488.

[16] Chernoff, H. & Haitovsky, Y. (1990). Locally optimal design for comparing two probabilities from binomial data subject to misclassification, *Biometrika* **77**, 797–805.

[17] Classon, D.L. & Southward, G.M. (1992). comparison of double sampling regression estimators, in *Proceedings of the 1990 Kansas State University Conference on Applied Statistics in Agriculture*, G.A. Milliken & J.R. Schwenke, eds. Kansas State University, Manhattan, pp. 257–264.

[18] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[19] Conniffe, D. (1975). Double sampling with regression-extension to the case of unequal regression coefficients, *Statistician* **24**, 259–266.

[20] Conniffe, D. & Moran, M.A. (1972). Double sampling with regression in comparative studies of carcass composition, *Biometrics* **28**, 1011–1023.

[21] Crete, M. Rivest, L.P., Jolicoeur, H. Brassard, J.M. & Messier, F. (1986). Predicting and correcting helicopter counts of moose with observations made from fixed-wing aircraft in Southern Quebec, *Journal of Applied Ecology* **23**, 751–761.

[22] De With, C. (1983). Two-stage plans for the testing of binomial parameters, *Controlled Clinical Trials* **4**, 215–226.

[23] Des Raj (1964). On double sampling for PPS estimation, *Annals of Mathematical Statistics* **35**, 900–902.

[24] Dorfman, A.H. (1994). A note on variance estimation for the regression estimator in double sampling, *Journal of the American Statistical Association* **89**, 137–140.

[25] Eberhardt, L.L. & Simmons, M.A. (1987). Calibrating population indices by double sampling, *Journal of Wildlife Management* **51**, 665–675.

[26] Esimai, G.O. & Han, C.-P. (1977). Double sampling in multi-auxiliary regression estimation based on conditional specification, in *American Statistical Association 1977 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 854–857.

[27] Fairley, W.B., Izenman, A.J. & Bagchi, P. (1990). Inference for welfare quality control programs, *Journal of the American Statistical Association* **85**, 874–890.

[28] Garg, R.C. & Pillai, S.S. (1975). Ratio-type estimators for two-stage designs, *Journal of the Indian Society of Agricultural Statistics* **27**, 37–48.

[29] Geng, Z. & Asano, C. (1989). Bayesian estimation methods for categorical data with misclassifications, *Communications in Statistics-Theory and Methods* **18**, 2935–2954.

[30] Hall, P. (1982). On Starr and Vardi's estimates of the number of transmission sources, *Journal of Applied Probability* **19**, 52–63.

[31] Han, C.-P. (1972). Double sampling with partial information on auxiliary variables, *Journal of the American Statistical Association* **68**, 914–918.

[32] Hansen, M.H. & Hurwitz, W.N. (1946). The problem of nonresponse in sample surveys, *Journal of the American Statistical Association* **41**, 517–529.

[33] Harishchandra, K. & Srivenkataramana, T. (1980). Conditional double sampling when the acceptance criterion is the variance, *Metron* **38**, 121–129.

[34] Heerschop, M.J. & Lieftinck-Koeijers, C.A.J. (1991). Registered unemployment in The Netherlands. Estimation of a dynamic population using retrospective information, *Statistician* **40**, 301–314.

[35] Hewett, J.E. & Spurrier, J.D. (1983). A survey of two stage tests of hypotheses, *Communications in Statistics-Theory and Methods* **12**, 2307–2325.

[36] Hill, I.D. (1982). Double sampling, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & M.L. Johnson, eds. Wiley, New York, pp. 419–423.

[37] Hinkins, S. & Scheuren, F. (1986). Hotdeck imputation procedure applied to a double sampling design, *Survey Methodology* **12**, 181–195.

[38] Hochberg, Y. (1977). On the use of double sampling schemes in categorical data with misclassification errors, *Journal of the American Statistical Association* **72**, 914–921.

[39] Ige, A. & Tripathi, T.P. (1987). On double sampling for stratification and use of auxiliary information, *Journal of the Indian Society of Agricultural Statistics* **39**, 191–201.

[40] Jolayemi, E.T. (1990). Relative frequency estimation in multiple outcome measurement with misclassifications, *Biometrical Journal* **32**, 707–711.

[41] Kawatheker, D.M. & Prabhu-Ajgaonkar, S.G. (1984). A modified ratio estimator based on the coefficient of variation in double sampling, *Journal of the Indian Society of Agricultural Statistics* **36**, 47–50.

[42] Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables, *Metrika* **27**, 217–223.

[43] Kiregyera, B. (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations, *Metrika* **31**, 215–226.

[44] Korn, E.L. (1982). The asymptotic efficiency of tests using misclassified data in contingency tables, *Biometrics* **38**, 445–450.

[45] Lamba, I.M.S. & Singh, P. (1993). Estimation of prevalence and relative risk in repetitive surveys, *Biometrical Journal* **35**, 877–892.

[46] Lessler, J.T. (1976). Survey designs which employ double sampling schemes for eliminating measurement process bias, in *American Statistical Association 1977 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 520–525.

[47] Li, H.G., Schreuder, H.T., Van Hooser, D.D. & Brink, G.E. (1992). Estimating strata means in double sampling with corrections based on second-phase sampling, *Biometrics* **48**, 189–199.

[48] Lie, R.T., Heuch, I. & Irgens, L.M. (1994). Maximum likelihood estimation of the proportion of congenital malformations using double registration systems, *Biometrics* **50**, 433–444.

[49] Magden, R.W. & Holstein, J.E. (1982). Determining sample size when searching for rare items, *IEEE Transactions on Reliability* **31**, 451–454.

[50] Majumdar, A. (1985). Systems of double sampling plans for fixed sample sizes, *Calcutta Statistical Association Bulletin* **34**, 233–236.

[51] Mak, T.K. & Li, W.K. (1988). A new method for estimating subgroup means under misclassification, *Biometrika* **75**, 105–111.

[52] Matloff, N.S. (1981). Use of regression functions for improved estimation of means, *Biometrika* **68**, 685–689.

[53] Mukherjee, R. & Chaudhuri, A. (1990). Asymptotic optimality of double sampling plans employing generalized regression estimators, *Journal of Statistical Planning and Inference* **26**, 173–183.

[54] Nedelman, J. (1988). The prevalence of malaria in Garki, Nigeria; Double sampling with a fallible expert, *Biometrics* **44**, 635–655.

[55] Neyman, J. (1938). Contributions to the theory of sampling human populations, *Journal of the American Statistical Association* **33**, 101–116.

[56] Oderwald, R.G. (1994). Stock and stand tables for point, double sampling with a ratio of means estimator, *Canadian Journal of Forest Research* **24**, 2350–2352.

[57] Oderwald, R.G. & Jones, E. (1992). Sample sizes for point, double sampling, *Canadian Journal of Forest Research* **26**, 980–983.

[58] Palmgren, J. (1987). Precision of double sampling estimators for comparing two probabilities, *Biometrika* **74**, 687–694.

[59] Patil, G.P., Sinha, A.K. & Taillie, C. (1993). Relative precision of ranked set sampling: a comparison with regression estimator, *Environmetrics* **4**, 399–412.

[60] Potter, F.J., Packer, L.E. & Batts, J.R. (1987). Double sampling for female Vietnam era veteran samples, in *American Statistical Association 1977 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 250–255.

[61] Prabhu-Ajagaonkar, S.G. (1975). The efficient use of supplementary information in double sampling procedures, *Sankhyā, Series C* **37**, 181–189.

[62] Rao, J.N.K. (1968). Some nonresponse sampling theory when the frame contains an unknown amount of duplication, *Journal of the American Statistical Association* **63**, 87–90.

[63] Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys, *Biometrika* **60**, 125–133, 669.

[64] Rao, P.S.R.S. (1972). On two phase regression estimator, *Sankhyā, Series A* **33**, 473–476.

[65] Rao, P.S.R.S. (1975). Hartley-Ross type estimators with two phase sampling, *Sankhyā, Series C* **37**, 140–146.

[66] Rao, P.S.R.S. (1975). On the two-phase ratio estimator in finite populations, *Journal of the American Statistical Association* **70**, 839–845.

[67] Rao, P.S.R.S. (1981). Efficiencies of nine two-phase ratio estimators for the mean, *Journal of the American Statistical Association* **76**, 434–442.

[68] Rao, P.S.R.S. (1983). Randomization approach, in *Incomplete Data in Sample Surveys, Theory and Bibliographies*, Vol. 2, W.G. Madow, I. Olkin & D.B. Rubin, eds. Academic Press, New York, pp. 97–105.

[69] Rao, P.S.R.S. (1986). Ratio estimators with subsampling the nonrespondents, *Survey Methodology* **12**, 217–230.

[70] Rao, P.S.R.S. (1990). Regression estimators with subsampling the nonrespondents, in *Data Quality Control Theory*, G. Liepins & V.R. Uppuluri, eds. Marcel Dekker, New York, pp. 191–208.

[71] Ray, S.K. & Singh, R.K. (1981). A product-type estimator in double sampling, *Biometrical Journal* **23**, 721–724.

[72] Reich, R.M., Bonham, C.D. & Remington, K.K. (1993). Double sampling revisited, *Journal of Range Management* **46**, 88–90.

[73] Robson, D.S. (1952). Multiple sampling of attributes, *Journal of the American Statistical Association* **47**, 203–215.

[74] Robson, D.S. & King, A.J. (1953). Double sampling and the Curtis impact survey, *Cornell University Agricultural Experimental Station Memorandum*, 231.

[75] Sedransk, J. (1965). A double sampling scheme for analytical surveys, *Journal of the American Statistical Association* **60**, 985–1004.

[76] Sedransk, J. (1967). Designing some multi-factor analytical studies, *Journal of the American Statistical Association* **62**, 1121–1139.

[77] Sen, A.R., Sellers, S. & Smith, G.E.J. (1975). The use of ratio estimate in successive sampling, *Biometrics* **31**, 673–683 and **33**, 767.

[78] Sengupta, S. (1981). Jack-knifing the ratio and the product estimation in double sampling, *Metrika* **28**, 245–256.

[79] Sethi, A.S. & Srivastava, A.K. (1987). Ratio estimator with post stratification design involving double sampling approach, *Journal of the Indian Society of Agricultural Statistics* **39**, 34–48.

[80] Shah, D.S. & Gupta, M.R. (1986). Comparison of double sampling estimators, *Metron* **44**, 417–420.

[81] Singh, B.D. & Singh, D. (1965). Some remarks on double sampling for stratification, *Biometrika* **52**, 587–590.

[82] Singh, D. & Singh, B.D. (1965). Double sampling for stratification on successive occasions, *Journal of the American Statistical Association* **60**, 784–792.

[83] Singh, R. (1984). Double sampling for two auxiliary variables. *Calcutta Statistical Association Bulletin*, **33**, 193–197.

[84] Singh, R.K. (1982). Generalized double sampling estimators for the ratio and product of population parameters, *Journal of the Indian Statistical Association* **20**, 39–49.

[85] Singh, R.K. (1983). A note on estimation in double sampling, *Journal of the Indian Society of Agricultural Statistics* **35**, 160–161.

[86] Singh, R.K. & Singh, G. (1983). Some double sampling estimators for population mean using auxiliary information, *Journal of Statistical Research* **17**, 7–19.

[87] Singh, S. & Singh, R. (1985). On random nonresponse in double sampling with difference estimator, *Communications in Statistics-Theory and Methods* **14**, 746–757.

[88] Singh, V.K. & Singh, G.N. (1991). Chain type regression estimators with two auxiliary variables under double sampling scheme, *Metron* **49**, 279–289.

[89] Sisodia, B.V.S. (1985). A note on successive sampling over two occasions, *Biometrical Journal* **27**, 97–100.

[90] Sisodia, B.V.S. & Dwivedi, V.K. (1982). On double sampling ratio-cum-product-type estimator with independent samples, *Biometrical Journal* **24**, 517–521.

[91] Sisodia, B.V.S. & Srivastava, A.K. (1982). Modifying regression estimators with preliminary test in double sampling, *Sankhyā, Series B* **44**, 295–303.

[92] Smith, P.J. (1990). Survey design optimization for catch curve analysis, *Journal of Statistical Planning and Inference* **26**, 277–290.

[93] Srinath, K.P. (1971). Multiphase sampling in nonresponse problems, *Journal of the American Statistical Association* **16**, 583–586.

[94] Srivastava, S.R., Khare, B.B. & Srivastava, S.R. (1990). A generalized chain ratio estimator for mean of finite population, *Journal of the Indian Society of Agricultural Statistics* **42**, 108–117.

[95] Srivastava, S.R., Srivastava, S.R. & Khare, B.B. (1989). Chain ratio type estimator for ratio of two population means using auxiliary characters, *Communications in Statistics-Theory and Methods* **18**, 3917–3926.

[96] Stockford, D.D. & Page, W.F. (1984). Double sampling and the misclassification of Vietnam service, in *American Statistical Association 1977 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 261–264.

[97] Swensen, A.R. (1988). Estimating change in a proportion by combining measurements from a true and a fallible classifier, *Scandinavian Journal of Statistics* **15**, 139–145.

[98] Swensson, B. (1983). On double sampling for stratification with non-response in the second phase, *Statistical Reviews* **21**, 111–116.

[99] Talukder, M.A.H. (1975). Response bias and difference estimate in double sampling, *Metrika* **22**, 65–76.

[100] Tamhane, A.C. (1978). Inference based on regression estimator in double sampling, *Biometrika* **65**, 419–428.

[101] Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications, *Journal of the American Statistical Association* **65**, 1350–1361.

[102] Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications: sample size determination, *Biometrics* **27**, 935–944.

[103] Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection, *Technometrics* **14**, 187–202.

[104] Tikkiwal, B.D. (1960). Classical regression and double sampling estimation, *Journal of the Royal Statistical Society, Series B* **22**, 131–138.

[105] Treder, R.P. & Sedransk, J. (1993). Double sampling for stratification, *Survey Methodology* **19**, 95–101.

[106] Tripathi, T.P. (1976). On double sampling for multivariate ratio and difference methods of estimation, *Journal of the Indian Society of Agricultural Statistics* **28**, 33–54.

[107] Tripathi, T.P. & Bahl, S. (1991). Estimation of mean using double sampling for stratification and multivariate auxiliary information, *Communications in Statistics-Theory and Methods* **20**, 2589–2602.

[108] Tripathi, T.P., Singh, H.P. & Upadhyaya, L.N. (1988). A generalized method of estimation in double sampling, *Journal of the Indian Statistical Association* **26**, 91–101.

[109] Tripathi, T.P., Singh, H.P. & Upadhyaya, L.N. (1989). Improved estimators for population mean based on double sampling, *Journal of the Indian Statistical Association* **27**, 89–99.

[110] White, D.B. (1990). Estimation using double sampling and dual stratification, *Survey Methodology* **16**, 105–116.

PODURI S.R.S. RAO

Drug Approval and Regulation

Before it can be marketed, any new medicinal product generally requires the regulatory approval of the appropriate governmental agency. In making the decision whether or not to license the product, the regulatory agency will normally consider evidence relating to the pharmaceutical quality of the product (its purity, and consistency), its efficacy (the extent to which it works, or achieves desired therapeutic effects), and its safety (the extent to which undesirable side effects attributable to the drug are absent). Some, but certainly not all, countries also require evidence that there is a clinical need for such a drug. The set of evidence considered will generally include results from preclinical toxicological studies (*see* **Preclinical Treatment Evaluation**), **pharmacokinetic and pharmacodynamic** studies, dose-finding studies *(see* **Phase I Trials**; **Phase II Trials**), Phase III comparative **clinical trials** of efficacy and safety, and sometimes from **bioequivalence** studies. Subsequently, data collected from **postmarketing surveillance** of adverse events may be reviewed at a later date. Much of what is described here with reference to licensing of medicinal products (drugs) also applies in parallel areas of approval and regulation of **medical devices**, dental and surgical materials, and veterinary products, although consideration of statistical issues is to date rather less well developed in some of these fields.

Regulatory Agencies and Statisticians

Statistical issues arise in many aspects of the process of drug approval and regulation, and statisticians are thus employed in, or used as advisors by, drug regulatory agencies to support that process. What may be more surprising is the relatively recent and haphazard provision for such involvement in some agencies. In the latter half of the twentieth century, the regulatory agencies that have had the greatest influence on pharmaceutical development are those of the United States, the European Union, and Japan. Most pharmaceutical companies have centered their regulatory strategies on satisfying the requirements of these authorities, knowing that relatively little extra work would generally be needed to satisfy the many other national authorities. However, this should not be taken to imply that there is an absence of scientific competence or statistical expertise in other agencies, many of which take strong independent positions on statistical issues when the need arises.

The regulatory agency in the United States is the **Food and Drug Administration** (FDA). Since the 1960s, this agency has employed large numbers of statisticians to review licensing applications. In addition, as part of the process of arriving at licensing decisions, the FDA has regularly elicited the opinions of experienced external statisticians, mainly academics, through their Advisory Committee Meetings.

The statistical resourcing of European agencies has been less generous and there is no consistent model [3]. However, starting in the 1990s, considerable improvements in this situation have been noted [4]. Since 1995, the coordinating and administrative body for the licensing of medicinal products in the European Union has been the European Medicines Evaluation Agency (EMEA). So far this organization has not employed statisticians. Indeed up to the present time, it has not employed any staff to carry out the work of scientific assessment of regulatory dossiers because, under the European procedures that were introduced in 1995, this activity is delegated to the national agencies of the individual Member States. Three national agencies (UK, Germany, and Sweden) have each employed small numbers of professional statisticians for 10 years or more, and their work has been mostly connected with the licensing of new products or modifications to existing licenses but has also extended into pharmacovigilance. All EU agencies, including the EMEA, have made some use of external expert statisticians for advice on statistical and methodological issues but the extent and manner of their use has varied considerably. Some Member States, such as France, have made extensive use of external statistical experts, involving them directly in assessment work, drafting guidelines, and other related activities. Other Member States, such as the Netherlands, have appointed statisticians to the national advisory committees that provide formal advice to aid regulatory decisions. These different ways of involving professional statisticians in the regulatory process are by no means exclusive: in addition to employing full-time professionals, the German agency also makes regular use of external expert statisticians in its assessment work; likewise, the UK agency had a long history of

appointing expert statisticians to its advisory bodies well before it took the step of recruiting full-time statistical staff.

In Japan, the licensing of medicinal products is carried out within the Ministry of Health, Labour and Welfare (MHLW). For several years, a small group of statisticians had been employed within this organization specifically for regulatory work. The MHLW also makes extensive use of academic consultants in providing regulatory advice and ensuring the appropriate conduct of regulatory clinical trials.

Statistical Assessment

In view of the considerable differences in the numbers of statisticians in regulatory bodies and in their manner of employment, it follows that statistical assessment procedures also vary markedly. In the United States, for example, the system ensures that FDA statisticians have access to the complete database of clinical trial and other data. They can thus carry out their own independent analyses of the key findings. These can then be considered in conjunction with the licensing applicant's own analyses by those charged with drawing scientific conclusions to support licensing decisions. This process has a tendency to polarize the analyses. The applicant's statisticians tend to emphasize the methods and aspects that support a more positive conclusion while the regulatory statisticians tend to be more conservative. Although there is clearly a danger that appreciable **bias** can be introduced by partisan presentations from either side, these are natural positions for the two groups to adopt and, by the end of the process, they usually lead to balanced decisions. This approach certainly minimizes the danger of assessors being unduly influenced by biased analyses from the applicant, and it also reduces the possibility of **fraud**.

In other agencies, such as those in the EU, independent analyses are rarely carried out. EU regulatory statisticians focus on verifying the validity and accuracy of the applicant's analyses and the appropriateness of the conclusions drawn from them. If they identify any apparent deficiencies, they may request the applicant to carry out further clarifying statistical work. Only rarely will they ask to be supplied with raw data in order to carry out their own analysis. In view of the limited resources available in the EU agencies, this approach is almost inevitable. However, it has the beneficial consequence that it is in the applicant's interest to submit a balanced analysis that discusses clearly all the pros and cons and presents well-justified conclusions. In this way, the applicant can try to avoid the delays that arise when further work is requested.

In the majority of regulatory agencies, the statistical resource is insufficient to assess all applications. This leads to another varying aspect of assessment, namely the question of which applications should be selected for assessment by a professional statistician. In some agencies, statistical review is invoked on grounds of perceived statistical complexity by nonstatistical reviewers. In others agencies, statistical assessment is routinely invoked for certain types of application or product, such as new chemical entities. Sometimes a senior statistician plays a part in selecting applications for statistical assessment. The different parts of an application also receive varying amounts of statistical attention. The most vigorously and routinely assessed part is the Phase III program of clinical trials and, in particular, the evidence of efficacy provided by these. Other issues, such as those associated with quality or safety, are less likely to be assessed by regulatory statisticians unless they are found to have a difficult methodological aspect.

The relative merits of the various assessment procedures described are a matter of continuing debate. However, the case for some form of statistical assessment of license applications is strong on grounds of both quality of the assessment made and efficiency with which it is made [5]. As noted above, drug license applications inevitably include a complex variety of statistical evidence. Critical professional assessment and review of this evidence is essential if correct licensing decisions are to be made efficiently. Methodological input to the regulatory process, leading to more rapid agreement on appropriate approaches to obtaining the required evidence of safety, quality, and efficacy, may have direct economic benefits to applicant companies in the short term, and thus to potential consumers of their products in the longer term. Erroneous decisions can be of two kinds: (i) inadequately stringent assessment may allow products with unsatisfactory risk/benefit ratios through to the market; or (ii) overcautious interpretation of evidence, perhaps in the light of statistical complexity, may lead to the failure of potentially valuable medicinal products to reach the market. In either case, the human and financial consequences

may be grave. For example, the first kind of error may result in patients being exposed unnecessarily to serious, possibly fatal, side effects of drugs; the second may prevent pharmaceutical companies from profiting from substantial investments in research and development but, more importantly, may prevent patients from receiving beneficial treatments.

License Application

A license application consists of a body of written evidence that supports a number of claims made on behalf of a medicinal product. At the base of the pyramid of evidence is usually a large number of pertinent studies, animal and human, *in vitro* and *in vivo*. The design and analysis of each of these individual studies are presented, and the conclusions that can be drawn from them are described. In addition, there may be summaries, and sometimes **meta-analyses** [2] that bring together the results of the separate studies to support specific licensing proposals.

The presentation of each study can and should involve statistical interpretation of evidence in support of the claims made. Summaries of the results of several studies should receive similar statistical attention. For the most part, this will involve straightforward application of standard methods. For example, we may expect or at least hope to see interval estimation (*see* **Confidence Intervals and Sets**) of the percentage of impurities in samples of the drug, of the response to the drug (on perhaps a **binary** success/failure scale, or as a continuous measure) in clinical trials, and of the **incidence** of adverse events following use of the drug in an identified cohort. We may also expect to find a discussion of the statistical precision and reliability of important estimates and sometimes an investigation of the **robustness** of the estimates to the analytical assumptions that have been made.

Unfortunately, at present some overemphasis on mechanical application of significance testing (*see* **Hypothesis Testing**) remains, and there are also other aspects of analysis where standardized approaches are overemphasized. Although mainstream applications of familiar statistical methods are often necessary, many of the more difficult issues require the intelligent application of a range of statistical approaches. Specification of "standard" approaches deemed acceptable by regulatory authorities are sometimes overzealously sought, perhaps especially by those lacking appropriate technical skills and experience, as a result of the commercial and health-related importance of the regulatory decisions that rest on them. There is often insufficient recognition that statistics is a science that requires interpretation and judgment; this in turn may require the evaluation of alternative approaches coupled with carefully argued justification of the conclusions finally reached.

The desire for the specification of approaches that are acceptable to regulatory authorities and the identification of areas where controversy remains has led to the development of regulatory guidelines in many areas of pharmaceutical science and, in particular, in pharmaceutical statistics. The development of statistical guidelines is fully described elsewhere, including the important ICH E9 **guideline** [1] that was developed as part of the international conference on harmonization (ICH) process. Statistical guidelines provide a consensus view of current methodology. However, they do not, and cannot, take away the need for professional statistical involvement in regulatory work, either in preparing applications or assessing them. Indeed, the ICH E9 guideline emphasizes the importance of the statistician's role. Difficult and contentious statistical issues often lie at the heart of the evidence provided by clinical trials and other experimental work. It is important that these are always investigated, interpreted, and described with care and understanding.

References

[1] International Conference on Harmonization E9 Expert Working Group (1999). Statistical principles for clinical trials: ICH harmonized tripartite guideline, *Statistics in Medicine* **18**, 1905–1942.

[2] Jones, D.R. & Lewis, J.A. (1992). Meta-analysis in the regulation of medicines, *Pharmaceutical Medicine* **6**, 193–205.

[3] Köpcke, W., Jones, D.R., Huitfeldt, B. & Schmidt, K. (1998). Statistics and statisticians in European drug regulatory agencies, *Drug Information Journal* **32**, 243–251.

[4] Lewis, J.A. (1996). Statistics and statisticians in the regulation of medicines, *Journal of the Royal Statistical Society, Series A* **159**, 359–365.

[5] Royal Statistical Society Working Party (1991). Statistics and statisticians in drug regulation in the United Kingdom, *Journal of the Royal Statistical Society, Series A* **154**, 413–419.

JOHN A. LEWIS & DAVID R. JONES

Drug Interactions

Interaction is a familiar term to most biostatisticians. When the effect of one factor differs across levels of a second factor, interaction between the factors is present. Drug disposition refers to the processes of how a drug is absorbed, distributed, metabolized (broken down), and excreted [6]. Variations in drug disposition and/or effect may result from interactions with, for example, diseases or genetic make-up. *Drug interactions* refers to the alteration of the disposition and/or effect of one drug, owing to the presence of a second drug.

What Causes Drug Interactions and Why are They Important?

Drug interactions arise from a myriad of complex physiologic conditions [5]. **Pharmacokinetics** refers to what the body does to the drug (processes of drug disposition), while pharmacodynamics refers to what the drug does to the body (the drug effect). Changes in the processes of drug disposition, known as the pharmacokinetic interaction, may take place when one drug's rate of elimination from the kidneys or liver is altered by a second drug. In such circumstances a drug can improperly accumulate in the body or be excreted too quickly. Another type of pharmacokinetic interaction can occur when specific enzymes that metabolize a drug become inhibited or induced by the presence of a second drug. A pharmacodynamic interaction refers to the alteration of the effects of one drug when given concurrently with another drug. The net result of a pharmacodynamic interaction may be an enhanced or diminished effect or the appearance of a new side-effect that was not seen with either drug alone.

Drug interactions may pose a dangerous threat to public health, especially when two commonly prescribed (and co-administered) drugs interact. A notable example is the gravely serious drug interaction between terfenadine (Seldane), a commonly prescribed anti-histamine, and ketoconazole, a popular anti-fungal drug [2, 4] When these drugs were taken simultaneously, unexpected life-threatening EKG changes (a syndrome known as Torsades de Pointes) and deaths occurred that were later attributed to an interference to the same key metabolizing enzymes shared by both drugs.

How Does Inter-Subject Variability Play a Role?

Studies that measure drug pharmacokinetics and/or pharmacodynamics are often challenged by substantial and unpredictable *inter-subject variability*. How the body processes a drug can differ greatly among subjects. This inherent variability in drug disposition is known as inter-subject pharmacokinetic variation. For a group of subjects given a fixed dose of a single drug, a large variation in serum drug levels (i.e. a coefficient of variation of 60% or greater) is commonly noted. Hence, for a two drug interaction study it is extremely difficult to partition the observed pharmacologic variation of one drug into underlying inter-subject pharmacokinetic variation vs. the variation due to the presence of a second drug. Moreover, identifying the sources of observed variability in drug effect, termed inter-subject pharmacodynamic variation, poses even greater difficulties. Suppose a target serum drug level can be achieved and maintained in a group of subjects. Even though the body's exposure to the drug is the same in all subjects, the variation in effect (e.g. lowered blood pressure) among subjects may be substantial. Introducing an additional source of variability, such as a second drug, further complicates the interpretation of inter-subject differences.

Which Study Design Addresses Inter-Subject Variability?

One design appropriate for testing drug interaction is a repeated measure design (*see* **Longitudinal Data Analysis, Overview**) [1]. This design, commonly called a crossover or **randomized complete blocks designs**, allocates all treatments to each subject, with an adequate "washout" period between treatments. Repeated measures denotes the serial measurements of drug disposition and/or effect after each treatment is administered. As each subject serves as his/her own control, all sources of variability among subjects are controlled. Only variation within subjects (the treatment effect) enters into the analysis. Typically, crossover studies designed to test for pharmacokinetic interaction enroll $10-25$ subjects.

For a two-drug interaction study of drugs A and B, each subject receives drug A, drug B, *and* a combination of drugs A and B. The order of the three treatments is often randomly assigned and balanced

so that the measurements are not **confounded** by treatment order. The model for a repeated measures design for a two-drug interaction study is

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \varepsilon_{ij},$$
$$i = 1, \ldots, n,\ j = 1, 2, 3,$$

where i denotes the subjects and j denotes the treatments (let $j = 1$ for drug A, $j = 2$ for drug B, and $j = 3$ for a combination of A and B). Y_{ij} denotes the measure of drug disposition or effect when the ith patient is given the jth treatment, $\mu_{..}$ denotes the overall outcome mean, ρ_i denotes the subject effect, τ_j denotes the treatment effect, and ε_{ij} denotes the error term. Individual subject effects are not of interest and only serve to reduce experimental error due to inherent inter-subject variability. Interaction is tested by a comparison between treatment means (analogous to a **paired *t* test**) and is performed by planned **contrasts**. For example, to test whether the effect of drug A is altered by drug B, the **null hypothesis** of no interaction is tested by

$$H_0 : \mu_{.j} - \mu_{.j'} = 0,$$

where, as noted above $j = 1$ and $j' = 3$. Similarly, to test whether the effect of drug B is altered by drug A, the null hypothesis of no interaction is tested by:

$$H_0 : \mu_{.j} - \mu_{.j'} = 0,$$

where $j = 2$ and $j' = 3$.

A recently published crossover study designed to test for the pharmacokinetic interaction between two agents, atovaquone and zidovudine, serves as an example [3]. Patients with human immunodeficiency virus (HIV) are at risk from adverse drug interactions because of the many drugs commonly prescribed to treat their disease and symptoms, such as *pneumocystis carinii* pneumonia (PCP). Atovaquone is an agent shown to be effective against PCP. Zidovudine is an anti-retroviral agent used as primary treatment for acquired immunodeficiency syndrome (AIDS). A high percentage of patients who receive treatment for PCP are also treated with anti-retroviral agents, so it is likely that these agents may be co-administered. A study was conducted to test whether the drugs could be co-administered without significant pharmacokinetic interaction. The treatment consisted of 26 consecutive days of therapy, defined by three dosing periods. Zidovudine was administered in the first dosing period (on days 1 and 2). Periods 2 and 3 consisted of 12-day intervals in which either atovaquone alone or atovaquone plus zidovudine was administered. The order of periods 2 and 3 was randomly assigned (*see* **Randomization**). Fourteen men with HIV enrolled on the study. Repeated measures analysis revealed that zidovudine and atovaquone could be co-administered without clinically significant pharmacokinetic interaction. Zidovudine had no effect on the disposition of atovaquone, while the systemic exposure of zidovudine was found to be increased by 33% after atovaquone administration.

References

[1] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York, Chapter 4, pp. 95–144.

[2] Honig, P.K., Wortham, D.C., Zamani, K., Conner, D.P., Mullin, J.C. & Cantilena, L.R. (1993). Terfenadine–Ketoconazole interaction: pharmacokinetic and electrocardiographic consequences, *Journal of the American Medical Association* **269**, 1513–1518.

[3] Lee, B.L., Tauber, M.G., Sadler, B., Goldstein, D. & Chambers, H.F. (1996). Atovaquone inhibits the glucuronidation and increases the plasma concentrations of zidovudine, *Clinical Pharmacology and Therapeutics* **59**, 14–21.

[4] Monahan, B.P., Ferguson, C.L., Killeavey, S., Lloyd, B.K., Troy, J. & Cantilena, L.R. (1990). Torsades de Pointes occurring in association with terfenadine use, *Journal of the American Medical Association* **264**, 2788–2790.

[5] Notari, R.E. (1987). *Biopharmaceutics and Clinical Pharmacokinetics: An Introduction*, 4th Ed. Marcel Dekker, New York, Chapter 8, pp. 354–369.

[6] Pratt, W.B. & Taylor, P. (1990). *Principles of Drug Action: The Basis of Pharmacology*, 3rd Ed. Churchill Livingstone, Edinburgh, Chapter 3, pp. 201–296.

(*See also* **Dose-response in Pharmacoepidemiology**; **Drug Approval and Regulation**; **Drug Utilization Patterns**; **Effect Modification**; **Interaction Model**; **Pharmacoepidemiology, Overview**; **Pharmacoepidemiology, Study Designs**; **Postmarketing Surveillance of New Drugs and Assessment of Risk**).

ROSEMARIE MICK

Drug Utilization Patterns

Exponential growth in prescription drug costs has raised questions about whether the expected cost-benefits (*see* **Health Economics**) of drug therapy are being realized in the population [61, 151]. At least three factors may dramatically alter the population's experience with a drug in comparison with effects observed in a **clinical trial**. First, most drug trials are carried out in populations of middle-aged adults, usually with a single health problem and who take few, if any, medications [52]. Evidence from these studies is generalized to the main users of prescription drugs, the elderly population, who receive 40% of prescription medication even though they comprise only 10 to 12% of the population in Western countries [46, 101, 133]. Seniors differ from middle-aged adults in several important respects. They are much more likely to have concurrent disease and use a number of medications, which may alter the absorption, distribution, and excretion of a drug, and influence the expected risks and benefits of therapy [13]. Secondly, drug use in the population often spans a broader set of indications than those evaluated in the context of a clinical trial. Thirdly, sample sizes in a clinical trial are usually estimated on the basis of the expected benefit of a drug (*see* **Sample Size Determination for Clinical Trials**). Therefore, clinical trials are usually underpowered (*see* **Power**) to detect clinically meaningful differences in less common, but clinically important, adverse outcomes. For example, **placebo**-controlled trials of the efficacy of aspirin in the prevention of myocardial infarction (MI) and stroke (*see* **Prevention Trials**) show a reduction in the occurrence of MI and stroke without evidence of adverse effects [22, 41, 149]. However, the size of the study population in each of the trials was insufficient to evaluate less common, but clinically important, adverse outcomes such as hemorrhagic stroke. When **meta-analysis** was used to pool results across trials, a twofold increase in the risk of hemorrhagic stroke was observed among aspirin users in comparison with placebo [90, 144].

To avoid such limitations, postmarket observational studies have been the primary means of providing information about the utilization of a drug in the population and its corresponding risks and benefits (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**). Skegg [130] discusses many methodological issues related to **observational studies** of postmarketing surveillance. Control of biases in selection, information, and **confounding** are the main challenges in observational studies (*see* **Bias in Observational Studies**). In this respect, the increasing use of population-based **administrative databases** to study prescription drug use in the population [7, 84, 112, 113, 125] and to examine the associations between prescription drug use and morbidity and mortality outcomes [49, 107, 108, 134] has come with its own advantages and particular challenges.

The Use of Administrative Databases in Drug Utilization Studies

To measure drug exposure, two types of administrative databases can be distinguished: (i) prescription claims databases (reimbursement claims for prescriptions dispensed to individuals in drug insurance plans) and (ii) pharmacy networks (pharmacy prescription drug profiles are linked by a computerized network in a region or district). Table 1 outlines the advantages and disadvantages of using these types of databases to ascertain drug exposure.

When a prescription database contains a unique identifier, such as a health insurance number, or nominal information, such as name, age, and birth date, there is the potential to link information describing an individual's exposure to a drug with other health care databases that contain information on morbidity and mortality *(see* **Record Linkage**). A description of the types of databases that are used, and the types and sources of information available, are summarized in Table 2.

Future Opportunities for Research through Point-of-care Data Collection from Electronic Health Records

In the next decade, we can expect to see a dramatic change in the data sources used for drug utilization research (*see* **Health Care Utilization Data**). The increasing use of information technologies to enable health care delivery are gradually replacing a predominantly paper-based system with integrated electronic health records. The integrated electronic health record can initially be expected to provide standardized, coded information on drugs prescribed and dispensed, laboratory, radiology and pathology results, operative reports, and active and past health

Table 1 Prescription databases: advantages and disadvantages

	Advantages	Disadvantages
Comprehensiveness	• All dispensed prescriptions (except exclusions) from all prescribers and all pharmacies documented, not just those known by the primary physician or reported by the patient [140] duration or prescribing	• Over-the-counter drugs excluded • Noninsured drugs excluded (for claims data only) • Borrowed drugs from friends excluded • Drugs during hospitalization may or may not be included • Documentation of prescription physician not always required for reimbursement
Drug information	• Data required for reimbursement is complete, including date dispensed, drug identification number (format, strength, drug), quantity, approximate duration, prescriber (claims data) • Completeness and accuracy of drug information data superior to that obtained from medical charts or by self-report	• Prescribed dosing not documented, which is a problem for drugs prescribed on an "as needed" basis, drugs prescribed on alternate day or graduated regimens (e.g. coumadin) • Substitution of one drug for another not recorded • Indication for prescription not recorded
Efficiency	• Data on prescription drug use can be retrieved at considerably less cost than retrieving the same information by primary data collection	• High-speed computer equipment and technical sophistication required to manage millions of prescription records and to produce usable measures for each patient
Population coverage	• A population-based sample can be readily assembled in many jurisdictions (exclusions noted) without concern for bias due to nonresponse or the significant cost of recruitment	• With claims data, subgroups in the population may not be covered by the drug insurance plan; also, inclusion in the drug insurance plan may be based on factors such as poor health or poverty, which may also influence the risk and benefit of a drug, and produce a biased estimate of its actual impact in the source population

problems. Eventually, with the increasing sophistication of the clinical interface, physicians and other health professionals will document history, lifestyle, and physical examination findings in electronic clinical notes. Hospitals have led the implementation of electronic health records, with the exception of a few countries such as Australia and the United Kingdom, where government initiatives have subsidized implementation in community-based care [12]. The transition to the electronic health records will provide the detailed clinical data that is needed to characterize study populations, and will enable new avenues of research, such as the effects of drugs on physiological parameters, the quantification of primary noncompliance with prescribed treatment and its predictors, and the treatment indications and outcomes associated with hospital-based drug and transfusion therapies. Surprisingly, only a few pioneering institutions [10, 11] have capitalized on the benefits of electronic health records for research; an issue that needs to be redressed through formal links between the health care information technology divisions and the research enterprise [89].

Studying the Effects of Drugs in the Population: Methodological Challenges

In many studies of drug utilization, the purpose of the study is to determine whether a class of drugs or a specific drug is associated with a greater risk of adverse outcomes than either other drugs or no drug. A number of methodological challenges are present in these types of observational studies, irrespective of the source of information: database or primary clinical data.

Assignment of Outcomes to Individual Drugs

Clinical Indication Bias. Patients who are selected for a specific drug treatment may differ considerably

Table 2 Databases available to assess morbidity and mortality outcomes

Database	Population	Source	Relevant data available
Hospitalization	Individuals discharged dead or alive from hospital	Abstraction and coding of the patient chart by medical records archivists [36, 111, 148]	• Cause of accidents leading to admission • Discharge diagnosis and 10–15 secondary diagnoses • Major treatments, surgery, and complications • Date of admission, discharge • Discharge destination • Death in hospital • Type of service delivered (e.g. fracture reduction of femur, carotid endarterectomy, pap smear)
Physician claims	Individuals receiving physician services in community and institutional settings	Reimbursement claims for physicians providing services on a fee-for-service basis	• Date and location of service and type of provider • Diagnosis (not an accurate indicator of reason for visit) • Primary cause of death
Vital statistics	Deaths occurring among national, regional residents	Death certificates completed by physicians, legal requirement for documentation	• Contributing causes of death • Name, birth date, sex, date of death

Table 3 The risk of adverse gastrointestinal events in a random database sample of 51 814 elderly in Quebec in 1990 [42]

Drug treatment group	Odds ratio	95% confidence interval
No NSAID prescription	1	–
NSAID alone	0.74	0.7, 0.8
NSAID + misoprostol	4.14	3.5, 5.4
NSAID + other GI drug prophylaxis	4.44	4.1, 4.8

in their risk of an adverse outcome from patients who are not treated. For example, misoprostol is a drug that has been shown in clinical trials to be efficacious as a prophylactic treatment for nonsteroidal (anti-inflammatory) drug (NSAID)-related ulcers [128]. In clinical practice, however, physicians tend to prescribe prophylactic gastrointestinal (GI) therapy to patients who are at higher risk of NSAID-related GI problems, or avoid prescribing NSAIDs altogether. As a result, as illustrated in Table 3, patients who receive prophylactic misoprostol or other GI-related therapy appear to have a greater risk of experiencing an adverse GI event, and persons receiving NSAIDs alone are at less risk of adverse GI events than the untreated population. These paradoxical findings dramatically illustrate the pitfalls of conducting observational studies on the risks and benefits of specific drug treatment without controlling for baseline differences, independent of the drug use, in risk of the outcome between patients started and not started on the drug.

In this example, the problem was addressed by using baseline preexposure information to measure a patient's risk of an adverse GI event prior to the initiation of NSAID therapy. Data from the prescription and physician claims databases and the hospitalization database were used to measure prior GI bleed, prior ulcer treatment, investigation and/or ulcer diagnosis, concurrent use of medications that would increase the risk of bleeding (e.g. coumadin), and other relevant comorbid conditions.

Propensity Scores. To reduce confounding by indication, it is essential to adjust the estimated drug

effect for the differences in the baseline characteristics of users versus nonusers. Adjustment through multivariable regression (*see* **Multivariate Multiple Regression**) becomes more popular in this context than alternative techniques such as **matching** or **stratification**. However, in some observational studies, the investigators' ability to adjust the drug effect for numerous **covariates** may be limited, for example, because of small number of outcomes, which would affect numerical stability of the estimates if the ratio of observed outcomes to the number of estimated regression parameters decreases below the critical range of 5 to 10. In such situations, the **propensity scores** approach may provide a feasible way to adjust for multiple potential confounders [114], provided these variables are available. A propensity score is an estimated conditional probability of a subject receiving a given drug, rather than an alternative drug or placebo, conditional on subject characteristics (covariates). The propensity scores are estimated using multiple **logistic regression**, with a binary indicator of the drug of interest (versus alternative treatment or placebo) as the dependent variable, and the relevant covariates as independent variables (*see* **Explanatory Variables**). Thus, the logic of the propensity score is, in fact, a weighted average of the original covariates, with weights corresponding to **maximum likelihood** estimates of the respective parameters of the multiple logistic regression model. Accordingly, including a *single* covariate, representing the estimated propensity score, in the model used to assess the effect of the drug on the outcome of interest, provides an approximate method of adjusting simultaneously for all potentially confounding subjects' characteristics that are related to the treatment choice [114]. D'Agostino [30] provides an excellent tutorial on the propensity scores methodology, together with interesting examples of its applications. An acknowledged limitation of the propensity score method is that it weights the importance of pretreatment patient characteristics by the strength of association with treatment assignment, and not the outcome [120]. This limitation may introduce residual confounding as equivalent propensity scores may be assigned to patients on the basis of different characteristics, both of equivalent importance in predicting probability of treatment assignment, but possibly only one being of importance in predicting the outcome.

Multiple Drug Use. In seniors, multiple drug use is the norm. Seniors fill an average of 29 prescriptions per year for seven different drugs [142]. Concurrent drug use creates a host of methodological problems. First, multiple drug exposure increases the risk of potential **drug interaction** (the effect of one drug is altered by the presence of a second drug, thereby increasing the risk of overdose toxicity or treatment failure) (see Figure 1) [137] and adverse drug events [51, 97] *(see* **Pharmacoepidemiology, Adverse and Beneficial Effects**).

Secondly, the assignment of an adverse outcome to a specific drug is complicated by the existence of a host of competing explanations, for example, other drugs and diseases [66]. Thirdly, methods of classifying multiple drug users into comparable groups with respect to the risk of adverse drug outcomes have not been developed. The simplest method has been to count the number of concurrent drugs taken. However, Granek [47] demonstrated that the risk of an injury with exposure to sedative hypnotics varied from 2.4 to 17.8 depending on the specific combination of drugs used. The methodological challenge in drug use research is to move beyond the simple solution of excluding multiple drug users to avoid confounding. Practicing physicians need to identify the subset of multiple drug users who are at greatest risk of adverse outcomes with a specific drug as

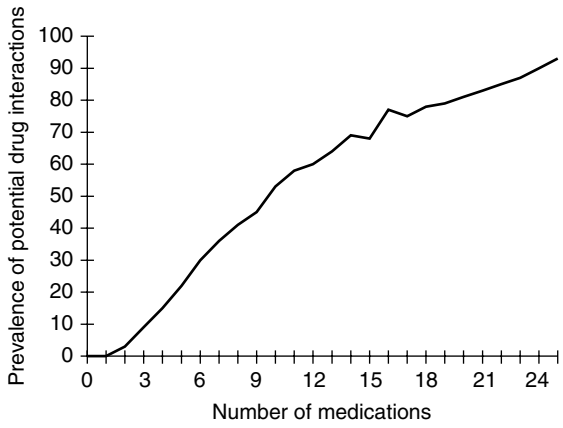

Figure 1 The association between the number of medications used by an individual patient and the prevalence of potential drug interactions. Reproduced from [137] by permission of Societé Française de Pharmacologie

well as the subset who will probably benefit. Population level databases [1, 36, 40, 46, 82, 110, 111, 149, 150] provide the opportunity to address some of these questions as a large number of individuals can be studied at a comparably low cost. Large sample sizes provide an opportunity to model simultaneously the effects of various medications and various diseases, thereby enabling the investigator to dissociate these effects as well as to test for clinically important interactions.

Prescriber Effects. Conventional multiple regression models rely on the assumption that outcomes of patients are independent [72]. Yet, there is substantial evidence that physicians vary considerably in their propensity to prescribe drugs, in their use of new drugs that come into the market, and in the appropriateness of their choice of drug treatment [32, 33, 38, 55, 92, 93, 126, 140]. Indeed, one recent study showed that physicians with lower scores on licensure examinations of clinical competence were not only more likely to prescribe inappropriate medications, but were also even less likely to demonstrate lower levels of quality of care in other domains. This suggests that physician competence may modify the effectiveness of prescribed therapy, and the probability of adverse outcomes [139]. Since most individual patients tend to be treated by one or two physicians [141], it is plausible that outcomes among patients of the same physician are correlated.

Comparison of the results of few recent studies that have provided relevant empirical evidence suggest that the strength of within-physician correlation depends strongly on the outcome. The reported values of the intraclass correlation coefficient (ICC) (*see* **Correlation**), measuring the proportion of the total **variance** in the outcomes of individual patients that can be attributed to physicians, ranged from ICC = 0.04 in a study of physician contributions to managed care pharmacy expenses [27] to ICC = 0.44 for the impact of unobserved prescriber characteristics on the occurrence of generic drug substitution [94]. Moreover, even after adjusting for patients' characteristics, a marked within-physician correlation has been also reported for various clinical outcomes of patients with rheumatoid arthritis, with ICC ranging from 0.16 to 0.25, depending on the outcome [29]. The same authors found a strong impact of individual rheumatologists on the decision to use prednisone and second-line agents. Ignoring such correlations may lead to underestimation of **standard errors** of regression coefficients [18] and inflated type I error rates in **hypothesis testing** [116]. Moreover, ignoring possibly systematic impact of skills, attitudes and/or beliefs of individual physicians on their prescribing choices and/or on the outcomes of their patients, may induce confounding bias if, for example, the choice of a given drug depends on physicians' characteristics that are also correlated with the outcome. Yet, Cowen and Strawderman [27] have recently pointed out that, in spite of the availability of more appropriate statistical models, most studies of the physician prescribing patterns continue to employ ordinary **least squares** regression (OLS), that relies on the independence assumption and ignores potential prescriber effects.

The methods able to account for within-cluster correlations of the outcomes include **generalized estimating equations** (GEE) [152], mixed models [70, 77], and **multilevel modeling** [16, 43, 44]. These methods handle both binary and continuous outcomes [76, 95, 152]. The choice of the specific method should depend on the scope of the analyses and on the plausibility of the underlying assumptions. In the simplest case, when the analyst is interested only in the effects of patient-level covariates, the standard "marginal" GEE model (sometimes referred to as GEE1 [85] may be used to correct the standard errors for the within-physician correlation [53]. In such analyses, the user may prefer to obtain a robust standard error which is not affected by **misspecification** of the structure of the **covariance matrix** of **residuals** (see below) [18]. Hanley et al. [53] provide an excellent nontechnical guide to applications of GEE in epidemiological and clinical studies with correlated outcomes, and identify some practically relevant limitations of this technique. Specifically, the GEE estimates of the regression coefficients for patients' characteristics, including treatment effects, have the "population average" interpretation, implying that they may be substantially different from the average within-physician effects that would be obtained, for example, from conventional matched analyses. This occurs because GEE yields a marginal version of the regression model, that is, the model that is estimated separately from the parameters of the covariance matrix [152]. Moreover, GEE does not provide an empirical criterion for the choice of an appropriate structure for the covariance of residuals [18] which, in our context, would represent deviations between observed outcomes of subsequent

patients of a given physician and the corresponding values predicted from the regression model [70, 146]. When correlations between such residuals are due, for example, to a systematic tendency of a given physician to chose a particular treatment more often than it could be expected based on his/her patients' characteristics, under the estimated regression model, the **exchangeable** covariance structure [142], sometimes referred to as compound symmetry structure [76], is a natural *a priori* choice as it assumes the expected correlation is the same for all pairs of patients of a given physician. However, in some cases the physicians' treatment preferences and the resulting prescribing patterns may change systematically over time [31]. In that case, the strength of between-patients correlations may decrease with increasing time elapsed between their respective prescriptions, making the autoregressive covariance structure, such as AR (1), a plausible alternative [146] (*see* **ARMA and ARIMA Models**).

An alternative approach relies on mixed models with **random effects** for individual physicians [70, 77]. Mixed models combine modeling of fixed effects of independent variables, representing their systematic impact on the outcome, as in the conventional multivariable regression, with random effects of individual physicians ("clusters"), that are not attributed to any observable covariates but are presumed to reflect their latent beliefs and attitudes. The random effects are assumed to be normally distributed (*see* **Normal Distribution**), with mean of 0 and variance that has to be estimated [27]. While the normality assumption may require some **transformation** of the dependent variable [15, 105], the above assumption avoids the need to specify explicitly the structure of the covariance of residuals [70]. Burton et al. [18] provide an excellent tutorial for researchers not familiar with a wide range of multilevel models for clustered and longitudinal data, including different versions of mixed models. Current research on prescribing patterns seems to favor the simplest version of the random effects model in that only intercepts are allowed to vary across physicians, while the associations between the patients' covariates and the outcome, such as treatment choice, are *a priori* assumed to be the same for all physicians [27, 29, 54, 94] The intercepts represent the individual physicians' systematic tendency to prescribe the drug of interest more or less often than it would be expected based on characteristics of their patients [27]. Cowen and Strawderman [27] explain well how to use random effects model with random intercepts in the research on prescribing patterns, and discuss its advantages, compared to either (i) naive OLS regression, or (ii) **fixed-effects** regression with $n-1$ **dummy variables** explicitly identifying each of n physicians. Specifically, random effects model increases efficiency of the analysis by replacing estimation of $n-1$ parameters by a single parameter representing the variance of random intercepts, helps separate the impact of individual physicians from the systematic effects of their characteristics, such as age or education, and avoids overestimating the impact of those physicians that contribute only relatively few patients. A further extension of the random-intercept model would include random slopes modeling [18], which would imply that the dependence of prescribing decisions on patient characteristics may vary across physicians. We found little evidence of the use of random slopes in the studies focusing on prescribing patterns, and the potential advantages of such more complex modeling remain to be investigated [27].

Finally, it should be noted that the mixed models represent a special case of the broader family of hierarchical multilevel models [16]. **Hierarchical models** are able to represent more than two levels of clustering, making it possible to account for clustering of patients within physicians' practices *and* for clustering of physicians within hospitals [17]. This allows for simultaneous modeling of patient-, physician- and hospital-level covariates, as well their interactions, while accounting for random effects of both physicians and hospitals [132]. For example, incorporation of cross-level interactions allows one to test if the impact of the patient health status on the choice on treatment varies between general practitioners and specialists. Finally, it is possible to quantify the proportion of the total variance in outcomes as these powerful models become increasingly available in user-friendly commercial **software** packages [16]. Raudenbush and Bryk [106] and Goldstein [45] present comprehensive descriptions of hierarchical multilevel models while Snijders et al. [132] and Burton et al. [18] provide user-oriented introductions. Burgess et al. [17] discuss many methodological issues relevant to implementation of hierarchical regression in the assessment of physicians' performance. Other examples of medical applications are found in [42] and [20].

Accounting for intercorrelations of either drug choices or potential clinical outcomes of medication use among patients of the same physician presents additional analytical challenges if the analysis has to rely on survival analytical methods such as the **Cox model**. Indeed, we are not aware of any commercial software package that would incorporate the Cox model analyses of multilevel data. Therefore, we tentatively propose two alternative approaches. Firstly, one may use a very recent method for incorporating random effects in Cox model, which relies on its affinity with **Poisson regression** [87]. An alternative approach would resemble the GEE methodology in that it yields "population average" point estimates of regression coefficients while attempting to correct the **confidence intervals** for the reduced amount of available information due to within-physician correlations of outcomes [53]. The proposed approach will adapt the method, developed by Abrahamowicz et al. [3], for **bootstrap** -based inference in complex nonlinear models involving nested or correlated data. The basic idea is to first estimate the conventional Cox model from the data pooled across all patients of all physicians, as for independent data. Next, the computer-intensive bootstrap procedure [35] is employed to simulate directly the effect of various sources of sampling error on the estimates. Specifically, the effects of (i) sampling physicians, and (ii) sampling patients within physicians' practices, are directly simulated by random resampling, with replacement, of (i) n original study physicians, and (ii) m_j patients original patients of each physician ($j = 1, \ldots, n$) sampled in step (i). To adapt this approach to our context, 1000 bootstrap samples should be generated by random resampling of the original data (Efron & Gong), and each sample should be independently analyzed with the Cox model. Finally, the 2.5th and 97.5th percentiles of the empirical distribution of the resulting 1000 Cox model-based **hazard ratio** (HR) estimates will provide the approximate bounds of the 95% confidence interval (CI) for the adjusted HR, and the **null hypothesis** of no effect of a given variable on hazard will be rejected at 0.05 level if the CI excludes 1.0 [3]. Whereas the procedure is computationally expensive, an exponential progress on the computational front makes it nowadays entirely feasible. However, further studies are necessary to assess the accuracy and relative efficiency of these alternative methods.

Measuring Drug Exposure

Compliance. To produce an accurate measure of the risk and benefit of drug treatment, patients need to be classified by the extent to which they were exposed to the respective drugs of interest. Database studies of drug utilization use information about dispensed prescriptions to determine whether an individual is likely exposed, information that tends to be superior to patient self-report or chart documentation of written prescriptions [50, 65, 71, 109, 134]. Although a dispensed prescription does not necessarily mean that a drug is taken, prescription refill rates have proven to be a valid means of measuring an individual's compliance with drugs that are used for chronic disease (e.g. hypertension) [69]. Figure 2 illustrates prescription refill rates for two hypnotic drugs. The beginning of each block is the date the drug was dispensed and the end is the date dispensed plus the recorded prescription duration (in days) (note that in some database files, duration is not specified, and **sensitivity analysis** [8, 39] is conducted to assess several duration time windows). Several approaches can be used to summarize these data, depending on whether the medication obtained by an early refill of a prescription (prescriptions 1 and 2) are added or not to the number of days of potential drug use.

Measuring Drug Dose. Figure 2 also provides information about drug dose, information that is ascertained from data on the quantity dispensed, the duration of the prescription, and the dose per unit of drug (e.g. 5 mg per pill). The first challenge is to decide how these data should be summarized to reflect the dose of drug taken in this 90-day window. For some drugs, such as lipid-reducing agents and prophylactic estrogens, the primary interest may be in the cumulative dose or average dose over time, while in others, such as sedative-hypnotics and nonsteroidal anti-inflammatory drugs, the starting dose, peak dose, or largest change in dose (such as precipitously stopping a high-dose, short half-life drug) may be the most important landmarks with respect to the risk of adverse events. The next challenge is to decide how overlaps of drugs in the same pharmacological class are to be handled, a situation that is relatively common in the elderly [136]. For example, it could be assumed that hypnotic B (prescription 3 in Figure 2) was substituted for hypnotic A (prescriptions 1 and 2), and hypnotic A

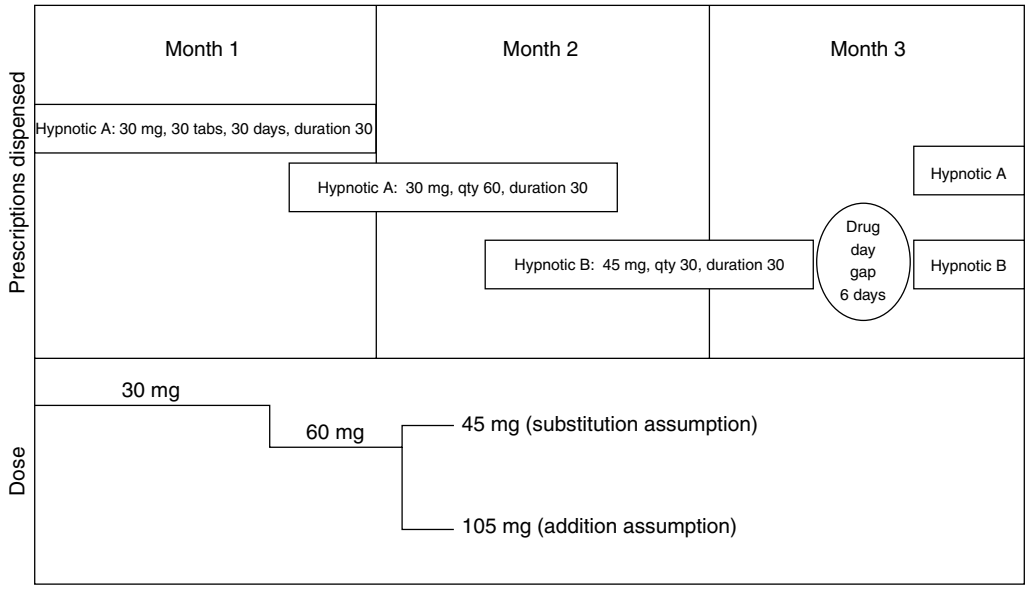

Figure 2 Summarizing drug use over time

was stopped. Using this assumption, daily hypnotic doses during the three-month time window vary from 30 mg per day to 60 mg per day (assuming **bioequivalence** for the sake of simplicity). However, hypnotics A and B were refilled in the subsequent month, suggesting that B was added – not substituted for – A. If it is assumed that both drugs were used concurrently, then the daily hypnotic dose varied from 30 mg per day to 105 mg per day. Generally, it is advisable to carry out the analysis under both types of assumptions as this may substantially alter the estimates of the association between drug dose and risks.

Patterns. Figure 2 also illustrates that drug use can fluctuate considerably, even over short periods of time. In the following section, we outline some methodological issues related to the description of variation over time in drug use and/or dose, and review some promising methods for analyzing such variations.

Describing Longitudinal Patterns. Many studies of drug utilization patterns are limited to simple descriptive *aggregate* statistics such as the **prevalence** of use and distribution of number of prescriptions [117], or temporal changes in either prevalence [115, 154] or the mean defined daily dose (DDD) [6] over time. Yet, a change in DDD is a compound of (i) the changes in prevalence of use, and (ii) changes in dose among users and, thus, may be difficult to interpret. This argument is supported by the results presented by van Hulten et al. [145], one of the very few among many studies of benzodiazepines use that assessed different aspects of longitudinal patterns of use. This study revealed several complexities, such as a nonmonotone change in the age-adjusted point prevalence rates over time, and the **interaction** between the effects of age and sex on the daily dose. Moreover, while the prevalence of use decreased considerably during the 10 years, the average number of prescriptions per user, and the relative proportions of incidental, regular and long-term users, remained quite stable. However, each of the three groups of users showed a different rate of decrease in the average dose per prescription [145]. Finally, the patterns of temporal changes in average dose for different benzodiazepines were quite inconsistent. On average, between 1983 and 1992 the dose increased for hypnotics and decreased for tranquilizers. Moreover, these trends differed substantially across individual compounds in each class of drugs. Among tranquilizers, the average daily dose increased for oxazepam, remained stable for diazepam, and decreased for lorazepam. Among hypnotics, the dose *increased threefold* for temazepam whereas flurazepam dose

showed a *threefold decrease* over the same period. Similarly, Hylan et al. [67] showed that both duration of initial prescription and the number of prescriptions depended on the specific antidepressant. These discrepancies emphasize the importance of separate assessments of each compound within the same class of drugs, and call for a careful modeling of different aspects of temporal patterns of exposure to medication.

Recently, Bartlett et al. [9] proposed some easy-to-implement methods for assessing different aspects of longitudinal patterns of drug use in *individual subjects*. Specifically, they suggested using subject-specific **Spearman rank correlation** between (a) the ordinal variable representing subsequent periods of uninterrupted medication use, and (b) period-specific dose, as a measure of individual's tendency to increase or decrease the daily dose over time. Next, they employed multiple **logistic regression** to identify subjects' characteristics associated with a particularly strong trend toward dose increase, operationally defined as the Spearman correlation above 0.89, corresponding to the 90th percentile of the sample distribution. The cut-off discriminates subjects with a very consistent, strong tendency to increase dose over time. Specifically, for subjects with three periods of use, it requires a systematic increase from each period to the next one. For subjects with four to seven periods of use, it identifies patterns of a strictly monotonic ($r = 1.0$) or reality monotonic (dose may increase or remain stable from interval i to $(i + 1)$ but never decreases) patterns of dose increase, with at least three different, gradually increasing dose levels. An alternative approach was also considered, based on GEE extension of the **multiple linear regression** [153], with the repeated-measures dependent variable defined as the subsequent values of daily dose, and independent variables including both subject baseline (fixed-in-time) characteristics and a time-dependent variable indicating subsequent periods of use. In addition, the GEE model included interaction terms between selected baseline variables and period of use, and the statistical significance of this interaction was tested to assess if the pattern of dose changes over time did depend on the corresponding characteristics [9]. These methods were then employed to demonstrate that longitudinal patterns of dose changes differed substantially across 11 different benzodiazepines.

Bartlett et al. [9] have also analyzed the patterns of switching from one to another benzodiazepine. They argued that a simple proportion of users of drug A who switch to drug B may not be a sufficiently sensitive measure of the popularity of drug B as a "second-line" treatment, because this proportion may largely reflect a current market share of drug B. Instead, they suggested that the proportion of subjects who had been switched from other drugs in the same class (numerator) among all new users of drug B (denominator) will capture the specific preference toward using drug B as the treatment of choice for switchers, over and above what could be explained by drug B's market share. This distinction becomes important, for example, in the context of comparing adverse effects of alternative products. Given that subjects who switch drugs may be expected to have worse outcomes [14] including poorer tolerance and lower probability of positive response [127], a comparison of outcomes between current users of different compounds may be biased if switching is not properly accounted for. Indeed, Bartlett et al. [9] reported that prevalent users of lorazepam were at higher risks of adverse events than prevalent users of several other benzodiazepines. Yet, Bartlett et al. [9] show that lorazepam was a particularly popular "second-line" choice for benzodiazepine switchers, who represented 22% of its new users, compared to 6 to 15% for nine other benzodiazepines. Thus, apparent higher risk among lorazepam users reported by Neutel et al. [96] could be an artifact, reflecting the worse outcomes for (more frequent) switchers rather than the impact of lorazepam *per se.* On the other hand, oxazepam had an almost as high proportion of switchers (20%) among its first users as lorazepam (22%). This would not be captured by the comparison of the proportions of switchers who had switched to oxazepam and lorazepam (17 versus 33%, respectively), which is largely affected by the much higher, and increasing over time, market share of lorazepam.

A different novel approach to assess the patterns was proposed by Coste et al. [25] who emphasize the multidimensionality of the drug prescribing phenomena, and propose the following empirical criteria for appropriateness of prescribing practice: placebo effect, novelty, "exoticism", misdosage, and drug–drug interactions.

Finally, it may be interesting to adapt methods employed to assess longitudinal patterns of changes

in different characteristics of individual subjects, not necessarily related to drug use, to the specific context of patterns of medication utilization. For example, methods proposed by [68, 123] may be employed to (i) identify different, statistically independent aspects of longitudinal changes in, for example, current dose, duration of uninterrupted drug use, and so on, and then (ii) perform **cluster analyses** to classify individual users into clusters of subjects with similar multivariate profiles of drug use.

Assessing Impact of Policy Changes. A related, methodologically challenging, issue of increasing societal importance concerns evaluating the impact of different changes in drug benefits policy on the longitudinal patterns of drug utilization. Schneeweiss et al. [122] and Tamblyn et al. [138] both used some form of interrupted **time series** analyses to assess to what extent the frequency and/or duration of drug use has changed after the policy implementation while accounting for the autoregressive structure of the data (*see* **ARMA and ARIMA Models**). A challenge in such analyses is to separate the impact of the policy change from the concurrent effect of secular trends in medication use. The strong secular trend toward increasing medication use, systematically observed in different populations *not* affected by any policy changes [130] may bias a "naive" pre-post comparison of the use before and after policy change. This may result, for example, in a spurious lack of difference if the reduction entailed by an increased co-payment is counterbalanced by a "natural" increase consistent with secular trends [138]. To avoid such biases, one may compare the actual use after the policy to the *expected* level of use, estimated by **extrapolating** the pre-policy trend over time [122, 138]. This approach may be further refined to discriminate between (i) the "main effect" of the policy change, modeled by a binary time-dependent policy indicator (assigned 0 before and 1 after the policy implementation) (*see* **Dummy Variables**), and (ii) the time-dependent time-by-policy interaction. The former would imply a constant-over-time difference between observed and expected utilization rates, corresponding to their initial drop followed by an increase with the slope on time similar to the pre-policy slope [122]. The latter would result in a post-policy slope being systematically different from the pre-policy slope, entailing a gradual increase in the discrepancy between the observed and expected utilization rates. In either case, subgroup analyses or interaction testing may be necessary to explore if the policy impact varies across different subpopulations. However, to avoid a serious increase of type I error rates, such analyses should be limited to a small number of *a priori* specified, clinically plausible potential **effect modifiers**. For example, one may expect that the impact of changes in the benefits will vary across socioeconomic strata, and will be smaller among users of essential, life-saving medications than among patients prescribed nonessential symptom-relief drugs [138].

Adequate representation of the pattern of within-patient changes in exposure levels and accurate modeling of such data present considerable challenges to biostatisticians. Whereas recent developments in flexible modeling of censored survival data offer some potentially useful tools to address some challenges, to date there is little evidence of their use in the empirical studies of drug effects. Moreover, some analytical problems occurring in modeling all potentially relevant aspects of the relationship between exposure and outcome remain to be addressed. In the next section, we review some promising new methods and suggest some directions for future research.

Modeling the Effects of Exposure to a Drug: Recent Developments in Survival Analysis and Future Challenges

Flexible Dose-response Modeling. In this section, we focus on the situation when an individual patient is assumed to be assigned a fixed dose of a drug, and the drug is prescribed once or at a constant dose throughout the follow-up period (*see* **Dose-response in Pharmacoepidemiology**). In most practical situations, there will be considerable between-patient variation in the duration of follow-up, and the first occurrence of an adverse reaction will often lead to a change in drug therapy or cessation of the drug. Both considerations suggest that time-to-event would be an appropriate outcome measure and, accordingly, that statistical methods for censored survival data should be used.

Medical applications of **survival analysis** are dominated by implementation of the **proportional hazards** model developed by Cox [28]. The theoretical elegance and versatility of the **Cox model** make it an extremely useful "first-line" tool for analyzing survival data in many practical contexts. However,

in some drug studies the assumptions underlying the model may be too restrictive to represent the **dose–response** relationship of interest. First, the conventional Cox model belongs to the parametric family of **general linear models** (GLM). The effect of a continuous **covariate** on the dependent variable, transformed by an appropriate link function (*see* **Generalized Linear Model**), is *a priori* assumed to be linear. In the context of the Cox model, this assumption implies that the logarithm of the **hazard** is a linear function of the covariate value. Yet, the assumption of the **loglinearity** of the dose–response relationship may be questionable in many studies of drug side-effects. *A priori* considerations suggest that, at least in some cases, an increase in dose over an interval of low-to-moderate doses may have very minor effects on the risk of an adverse reaction, whereas a further increase in the dose may result in a dramatic risk increase as the "tolerance threshold" will be exceeded. In such cases, the effect would be better described by an upward concave function, such as an exponential or quadratic curve. It would be advantageous to model the shape of this function in a flexible way. Moreover, it is possible that the "tolerance thresholds" differ systematically from one subgroup of patients to another. In this situation, the dose–response function may have several relatively flat portions interchanged with more abrupt increases at doses corresponding to subgroup-specific thresholds. To model such complex shapes of dose–response curves one may use some of the recently proposed generalizations of the Cox model that replace the loglinear HR by a more flexible function [34, 48, 56, 131]. Hastie & Tibshirani [56] discuss a number of **nonparametric** smoothers that can be used to provide a smooth estimate of the effect (*see* **Smoothing Hazard Rates**; **Smoothing Methods in Epidemiology**) while avoiding *a priori* assumptions about the functional form of the dose–response relationship.

The advantages of flexible nonparametric modeling of dose–response relationships have been demonstrated in many practical applications [2, 104, 105]; however, there is little evidence of the use of such methods in drug studies. This approach is of particular interest in observational studies of the effects of drug utilization where one may expect considerable between-patient variation in the dose. Computer **simulations** have demonstrated the ability of nonparametric methods to provide reasonably **unbiased** and stable estimates of a broad variety of functions, including curves with local plateaus and abrupt increase [56, 104]. These methods may be of interest in studying the effects of increasing dose on the risk of an adverse reaction.

Using nonparametric modeling in practice requires additional methodological decisions regarding the choice of a particular smoothing technique and the desirable complexity of the estimated model. Both decisions are far from trivial as the theoretical understanding of related phenomena and practical experience in this area have only recently begun to accumulate. On the basis of our experience in nonparametric modeling of biomedical data [2, 3, 105] and some computational considerations, we have a few suggestions. First, we suggest selecting *a priori* a specific type of smoother and **degrees of freedom**, otherwise statistical inference about the estimates becomes difficult [4]. Moreover, limited comparisons between methods suggest that the estimates are reasonably **robust** with respect to the smoothing technique chosen [103]. In our experience, the smoothing spline option incorporated in the **generalized additive models** (GAM) proposed by Hastie & Tibshirani [56] has provided numerically stable and clinically plausible estimates [2]. GAM also has two important practical features. First, the use of this methodology is largely facilitated by the fact that the GAM program is included in the **S-PLUS** commercial package and the monograph by GAM authors [56] is a useful guide on how to apply this powerful methodology in practice. Secondly, GAM allows the user to make an inference about the estimates. The simulations reported in the same monograph indicate that both type I error rates at conventional significance levels (0.05 or lower) and point-wise confidence intervals are accurate within a practically acceptable error margin [56].

Finally, flexibility of the GAM estimators can be used to determine if there is a threshold in the dose–response relationship, an issue of considerable importance in the studies of potential adverse effects of medications. Indeed, if one could establish that there is a threshold for the dose below which the risks of adverse effects are similar to those associated with a placebo, such a finding could suggest an "optimal" dose that would offer therapeutic benefits while avoiding negative side effects. In this context, methods proposed in [19, 75] for threshold detection are of interest, although their accuracy and efficiency remains to be systematically

evaluated through simulations. Moreover, statistical inference regarding both the existence of a threshold and the precision with which its location is estimated is complicated by the fact that the underlying model is nonlinear in its parameters (threshold location). Therefore, computer-intensive techniques such as **bootstrap** [35] may be applied to assess "honest" P values for testing the alternative hypothesis of a threshold-based dose–response curve against the null hypothesis of either no association or linear (no-threshold) dose–response, as well as to estimate empirical confidence intervals around the estimated threshold (*see* **Extrapolation, Low Dose**). To ensure the accuracy of bootstrap-based inference, it is essential to replicate the entire estimation and/or model selection process, including, for example, data-dependent selection of degrees of freedom in each individual bootstrap sample [3].

The fractional polynomials methodology may offer an interesting alternative to nonparametric GAM-based estimation of the dose–response relationships. A fractional polynomial model is estimated by first, selecting one or two best-fitting functions from an *a priori*-defined basis of a restricted number of fractional polynomials, that is, polynomials whose degrees are fractions such as, for example, 1/2 or 1/3 [118, 119]. Then, the standard multivariable regression methods (*see* **Multivariate Multiple Regression**) are employed to estimate the regression coefficients defining the "optimal" linear combination of the selected functions. Thus, fractional polynomials combine the **parsimony** and easy interpretation characteristic of parametric regression models with flexibility comparable to low-dimensionality nonparametric models [118]. The method is user-friendly and is implemented in the STATA computer package. In our opinion, further comparisons, involving both carefully designed simulations studies and analyses of empirical data, are necessary to assess relative advantages of GAM versus fractional polynomials, and their dependence on the sample size and/or complexity of the underlying "true" relationships.

Assessing and Modeling Time-dependence of the Drug Effects. Another important assumption underlying the Cox model [28] is that the ratio of hazards, corresponding to different covariate vectors, is constant over the entire follow-up period (i.e. that hazards are proportional). The proportional hazards (PH) assumption implies that the relative risks associated with different drug doses do not change over time. Although several tests of the PH assumption have been proposed in the statistical literature [86], their use in clinical and epidemiologic studies is limited and the PH model is typically accepted a priori as a valid model [5]. However, for many drugs, the effect of a constant dose of medication will change over time. For example, a review of findings on the effects of cholesterol lowering on coronary heart disease risk suggested that about five years of treatment is necessary to achieve the full benefits of lipid-lowering medication [79]. The presence of a lag in the effect of statins is also suggested by the results of a large clinical trial, namely, the 4S study where the **Kaplan–Meier** survival curves for the placebo and active treatment groups are identical for the first year of follow-up and then start to diverge gradually [143]. If the effect of a drug gradually increases with increasing treatment duration, then the PH model, in which these risks are *a priori* forced to be constant, will yield the hazard ratio estimate corresponding to the average-over-time **relative risks** [100]. Such an estimate will overestimate the early effect and underestimate the long-term effects of the treatment [105]. In other situations, the effect of a constant dose may decrease with increasing exposure duration. Analysis of data from a randomized trial of aspirin effectiveness in preventing cardiovascular events among asymptomatic patients with carotid bruits provides an example [26]. A constant dose of aspirin appeared to have a short-term protective effect that did not last beyond the first year. Interestingly, an *in vitro* study suggested a biological mechanism that might underlie such a gradual loss of aspirin efficacy with increasing exposure duration [26].

Moreover, the pattern of time-dependence may vary depending on the specific type of drug's effect. For example, tolerance to the depressant effects of benzodiazepines develops rapidly, in contrast to a gradual process of slowly increasing tolerance to their anxiolytic effect, with increasing duration of use [124].

To avoid such systematic biases, several authors have proposed flexible generalizations of Cox model [4, 48, 57, 63, 74]. In these models, the constant log hazard ratio, β, a parameter of the Cox model is replaced by a flexible function of follow-up time, $j\beta\ (t)$, and this function is estimated using various nonparametric methods. Abrahamowicz et al. [4] recently developed a regression **spline**-based model

for time-varying relative risks. In simulation studies, they demonstrated its ability to uncover a variety of patterns of hazard ratio changes over time. Moreover, by using simple properties of regression splines [105, 147] they were able to propose simple and relatively reasonable accurate inferences about their estimates. Clinical applications of such methods yielded new insights that could not be obtained with more conventional methods [26, 37, 48, 63, 83, 99, 100, 103].

It would be desirable to use similar methods in observational studies of drug utilization, as in many cases the relative risks of adverse reactions may change over time. For example, benzodiazepines are drugs that have a depressive effect on the central nervous system, with side effects of motor incoordination and diminished cognitive function. It is likely that the effect of benzodiazepine use on the risk of adverse events is not loglinear and is more consistent with a threshold model. However, such a threshold remains to be estimated. Moreover, the risk of injury resulting from the side effects of exposure is expected to be greatest when the drug is first started, but it will diminish over time, even at constant doses, because of metabolic adaptation. Sudden cessation of the drug (such as may occur during hospitalization), dose incrementation, or the addition of a second drug which potentiates the effect of benzodiazepines (drug interaction) may all contribute to an increase in the risk of an adverse event.

Thus, it is plausible that the effect of a constant dose of a drug may be at the same time nonloglinear and time-dependent. In these instances, simultaneous flexible modeling of both effects will be necessary to represent the relative risks of interest. Relatively little work has been done to date on hybrid models that could incorporate both effects simultaneously. Gray [48] discusses nonparametric modeling of either nonloglinear or time-varying effects of continuous risk factors, but does not address the issues specific to the situation when both effects are estimated for the same risk factor. Simultaneous modeling of both effects is possible using hazard regression (HARE) [74], a very versatile spline-based methodology. However, HARE relies on adaptive model selection, and none of the examples presented in the article shows the selection and simultaneous modeling of nonloglinear and timedependent effects for the same covariate. This may be partly due to potential **collinearity** between the nonloglinear and time-dependent effects of the same continuous predictor.

A hypothetical example will illustrate the problem. Assume, for example, that the relationship between a drug dose and the logarithm of the hazard is **exponential** rather than linear (i.e. changes in dose within the low-to-moderate dose range have very little effect on the risk, but increases in dose in the high-dose range result in dramatic increases in the risk of an adverse event). Assume, further, that this effect of a dose remains constant, which means that it does not change as a function of treatment duration. The correct modeling of this effect would involve flexible transformation of drug dose, X, but it would conform with the PH assumption that the hazard ratio is constant over time. Accordingly, the effect of drug dose should be represented by $\beta^* f(x)$, where β (constant over time) is the log hazard ratio and f is the flexible transformation (resulting from nonparametric modeling of the effect of dose at a fixed point in time). Yet the same data may be well represented by a loglinear, time-dependent model in which the effect of the dose will be modeled as, $\beta(t)^*x$, where $\beta(t)$ represents the slope of the linear effect of dose (x) at time t. This model is obviously inconsistent with the actual relationship between dose and risk. This incorrect alternative representation may fit the data quite well because the shape of the true "constant-in-time" risk function implies that most patients who are dispensed high doses will have adverse events quite early. Thus, the initial slope of the linear risk function, $\beta(t)^*x$ (for values of t close to zero) will be quite high to reflect the contrast between risks associated with low and high doses. However, as treatment duration increases, most of the individuals with high doses will be filtered out and the slope will be mostly determined by outcomes among those with low to moderate doses. As the dose has little effect in that interval, the slope of $\beta(t)$ will gradually decrease with increasing t, creating an apparent time-dependence of the dose effect. In the absence of *a priori* grounds to prefer a constant nonloglinear model over the time-dependent loglinear alternative, fitting such data may create **identifiability** problems. These theoretical concerns about identifiability have been confirmed in a **simulation** study [129]. The development of an appropriate methodology that is able to offer simultaneous modeling of various effects, as well as accurate inference about the estimates, is a major challenge for further development of survival analytic tools, useful for studying the effects of drug utilization.

Estimating Lags in the Effects of a Drug. Related to the issue of modeling time-dependent changes in the impact of a medication is the problem of assessing the latency or temporal lag between drug exposure and a subsequent change in risks. Indeed, both the therapeutic and adverse effects of many drugs may be observed only after a certain latency and/or may wane after another longer time interval. Conventional methods for assessing latency involve estimating alternative models with exposure lagged by a different time interval, and then selecting the best-fitting of this model as an "optimal" representation of the lag duration [91, 98, 121]. However, the underlying assumption that only exposure that occurred in a specific period is relevant limits the clinical plausibility of the results. An alternative approach is to include several period-specific exposure indicators, for example, representing NSAIDs use in subsequent five-year intervals such as 0 to 5 years ago, 5 to 10 years ago, 10 to 15 years ago, and so on in the same model, and to compare the respective relative risks estimates [21]. However, such an approach induces a risk of multi-collinearity if the exposure status of individual subjects changes little during their lifetime, and the implicit assumption that the impact of exposure is a discontinued step-function of the time since exposure may be questionable.

To avoid such difficulties, more recent methods rely on flexible modeling techniques to estimate smooth functions representing different aspects of the lagged exposure effect, based on different conceptual models. For example, Rachet et al. [102] estimate a smooth distribution of lag time, defined as the time elapsed between the beginning of exposure and the subsequent change in risks, assuming that the relative risks exposed/unexposed remain constant after that time. In contrast, Hauptmann et al. [60] assume that the effect of exposure changes gradually with increasing time since exposure and postulate that the etiologically relevant measure of cumulative exposure should be best calculated as a weighted function of exposure intensity in different time periods. They use cubic regression B-splines to estimate a smooth weight function directly from the data, and propose a **likelihood ratio test** to compare the resulting model with the simple unweighted measure of cumulative exposure. Other promising methods for assessing latency include the bilinear model [78] and the sliding time window approach of Hauptmann et al. [58].

Use of Time-dependent Covariates to Model Changes in Drug Use Over Time. The foregoing illustrates the complexity of analytic problems in studying the effect of drug doses that are fixed over time. It is clear that the same problems are considerably more challenging when modeling involves doses that change over time. In the conventional Cox model [28], dose changes can be represented by a **time-dependent covariate** [i.e. a covariate that is a function of time $X(t)$]. Dose changes also include situations in which cumulative dose rather than a constant daily dose is considered to be the most important determinant of the outcome. If the cumulative dose were treated as a fixed-in-time covariate in survival analysis, then the results will be considerably biased, and even paradoxical due to **length bias** [24]. Indeed, those having early adverse events would have much shorter exposure to the drug and, thus, a lower cumulative dose. Therefore, higher risks, corresponding to shorter time-to-event, would appear to be associated with a lower cumulative dose. Representing cumulative dose by a time-dependent covariate would eliminate the risk of such length-biased confounding (*see* **Screening Benefit, Evaluation of**).

In some applications where the dose changes over time, it may be preferable to separate information about these changes by using two time-dependent variables: one measuring current dose and the other indicating whether the dose had been increased in the last period. The former variable will be used to study the association between current dose and immediate risks and the latter to test the hypothesis that recent changes in dose increase the risks. However, simultaneous estimation of different time-dependent aspects of the same exposure may lead to multi-collinearity problems. A recent study of smoking exposure explores various models with time-dependent covariates that may be used to separate different interrelated aspects of the longitudinal exposure history, while avoiding multi-collinearity and ensuring interpretability of the estimates [81]. An alternative approach is to aggregate different aspects into a single compound exposure measure, representing clinically relevant overall index of cumulative exposure. For example, a one-compartment exponential elimination model (*see* **Compartment Models**) proposed by Hauptmann et al. [59] accounts simultaneously for the duration of use and medication dose, as well as for time elapsed since the end of exposure, while making specific assumptions about the

way they interact. However, further empirical and simulation studies are necessary to assess the performance of such models.

The incorporation of time-dependent covariates in flexible generalizations of the Cox model will be necessary to account for situations where (i) the dose may change over time, and (ii) the effect of a given dose (current or cumulative) may also change as a function of treatment duration. In principle, flexible modeling of time-dependent covariates does not generate major additional theoretical problems; however, very efficient **algorithms** are required to manage the increased computational burden. Indeed, Heinzl et al. [62] have proposed a regression spline model that incorporates time-varying effects of binary time-dependent covariates. This method allows for modeling the changes over time in the impact of the current drug use, so that, for example, one can test if the risk of adverse effects decrease with increasing duration of use, which may occur if the users gradually develop tolerance. Finally, Clarkson and Kooperberg [73] extend the HARE methodology to incorporate flexible modeling of both binary and continuous time-dependent covariates. Such a development would likely allow new insights into the mechanisms underlying the risks and benefits of drug use (*see* **Time-varying Treatment Effect**).

The aforementioned methods, developed in the context of environmental or occupational exposures, may offer new insights into the role of medication, especially in long-term observational studies of large cohorts with detailed data on the time and dose of a drug use that can be derived from administrative prescriptions databases. However, one common limitation of the administrative prescription databases is that exact timing of the actual drug exposure is unknown and has to be inferred from (i) the date when the prescription was filled, and (ii) its duration. Whereas it is reasonable to assume that the patient did start taking the drug at or soon after the date of its purchase, the actual end of the period of active use may be quite different from the date corresponding to the end of prescription. In such situations, some sensitivity analyses of the robustness of the results with respect to various assumptions about the plausible structure and magnitude of errors in the timing of exposure are recommended. This may involve simulating a large number of similar datasets, with randomly distributed errors in the "observed" duration of exposure, and directly assessing the changes in the parameters of interest. To this end, one may use a versatile permutational algorithm for generating censored time-to-events, described in [4] and validated in [88]. As described by Leffondré et al. [80], this algorithm can be adapted for generating events conditional on time-dependent covariates, that are necessary to represent temporal variation in drug use correctly. The analyst may also consider adapting here the simulation extrapolation method (SIMEX) (*see* **Measurement Error in Epidemiologic Studies**) for handling measurement errors in predictors [23, 135]. In some cases, where the expected variance of the measurement errors can be well approximated, for example, based on relevant previous publications, the SIMEX method can accurately estimate the underlying true relationship that would be observed with error-free variables [64].

References

[1] Aaronson, L.S. & Burman, M.E. (1994). Use of health records in research: reliability and validity issues, *Research in Nursing & Health* **17**, 67–73.

[2] Abrahamowicz, M., du Berger, R. & Grover, S.A. (1997). Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality, *American Journal of Epidemiology* **145**, 714–729.

[3] Abrahamowicz, M., Fortin, D.F., du Berger, R., Nayak, V., Veville, C. & Liang, M.H. (1998). The relationship between disease activity and expert physician's decision to start major treatment in active systemic lupus erythematosus: a decision aid for development of entry criteria for clinical trials, *The Journal of Rheumatology* **25**, 277–284.

[4] Abrahamowicz, M., MacKenzie, T. & Esdaile, J.M. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis, *Journal of the American Statistical Association* **91**(436), 1432–1439.

[5] Altman, D., De Stavola, B., Love, S., & Stepniewska, K. (1995). Review of survival analyses published in cancer journals, *British Journal of Cancer* **72**, 511–518.

[6] Anonymous (1994). *Anatomical Therapeutic Chemical (ATC) Classification Index Including Defined Daily Doses (DDDs) for Plain Substances. WHO Collaborating Center for Drug Statistics Methodology*, Oslo.

[7] Anderson, G.M., Kerluke, K.J., Pulcins, I.R., Hertzman, C. & Barer, M.L. (1993). Trends and determinants of prescription drug expenditures in the elderly: data from the British Columbia Pharmacare Program, *Inquiry* **30**, 199–207.

[8] Axelson, O. (1978). Aspects on confounding in occupational health epidemiology, *Scandinavian Journal of Work, Environment & Health* **4**, 85–89.

[9] Bartlett-Esquilant, G., Abrahamowicz, M., Tamblyn, R., Du Berger, R. & Capek, R. (2003). Longitudinal patterns of new benzodiazepine use in the elderly, *Pharmacoepidemiology and Drug Safety*; in press.

[10] Bates, D.W., Boyle, D.L., Vander, V.M., Schneider, J. & Leape, L. (1995). Relationship between medication errors and adverse drug events, *Journal of General Internal Medicine* **10**(4), 199–205.

[11] Bates, D.W., Cohen, M., Leape, L.L., Overhage, J.M., Shabot, M.M. & Sheridan, T. (2001). Reducing the frequency of errors in medicine using information technology, *Journal of the American Medical Informatics Association* **8**(4), 299–308.

[12] Bates, D.W., Ebell, M., Gotlieb, E., Zapp, J. & Mullins, H.C. (2003). A proposal for electronic medical records in U.S. primary care, *Journal of the American Medical Informatics Association* **10**(1), 1–10.

[13] Beers, M.H. & Ouslander, J.G. (1989). Risk factors in geriatric drug prescribing: a practical guide to avoiding problems, *Drugs* **37**, 105–112.

[14] Blais, L., Ernst, P. & Suissa, S. (1996). Confounding by indication and channelling over time: the risks of B-agonists, *American Journal of Epidemiology* **144**, 1161–1169.

[15] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of Royal Statistical Society Series B* **26**, 211–252.

[16] Bryk, A., Raudenbush, S. & Congdon, R. (1996). *Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*. SPSS Inc., Chicago.

[17] Burgess, J. & Christiansen, C. (2000). Michalak seal. Medical profiling improving standards and risk adjustments using hierarchical models, *Journal of Health Economics* **19**, 291–309.

[18] Burton, P., Gurrin, L. & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling, *Statistics in Medicine* **17**, 1261–1291.

[19] Cakmak, S., Burnett, R. & Krewski, D. (1999). Methods for detecting and estimating population threshold concentrations for air pollution-related mortality with exposure measurement error, *Risk Analysis* **19**, 487–496.

[20] Christiansen, C. & Morris, C. (1997). Improving the statistical approach to health care provider profiling, *Annals of Internal Medicine* **127**, 764–768.

[21] Collet, J-P., Sharpe, C., Belzile, E.B.J.-F., Hanley, J. & Abenhaim, L. (1999). Colorectal cancer prevention by non-steroidal anti-inflammatory drugs: effects of dosage and timing, *British Journal of Cancer* **81**(1), 62–68.

[22] Collins, R., MacMahon, S., Flather, M., Baigent, C., Remvig, C., Mortensen, S., Appleby, P., Godwin, J., Yusef, S. & Peto, R. (1996). Clinical effects of anticoagulant therapy in suspected acute myocardial infarction: systematic overview of randomized trials, *British Medical Journals* **313**(7058), 652–659.

[23] Cook, J. & Stefanski, L. (1994). A simulation extrapolation method for parametric measurement error models, *Journal of American Statistical Association* **89**, 1314–1328.

[24] Correa, J. & Wolfson, D. (1999). Length-bias: some characterizations and applications, *Journal of Statistical Computation and Simulation* **64**, 209–219.

[25] Coste, J. & Venot, A. (1999). An epidemiologic approach to drug prescribing quality assessment – a study in primary care practice in France, *Medical Care* **37**(12), 1294–1307.

[26] Côté, R., Battista, R., Abrahamowicz, M., Langlois, Y., Bourque, F. & Mackey, A., and the Asymptomatic Cervical Bruit Study Group. (1995). Lack of effect of aspirin in asymptomatic patients with carotid bruits and substantial carotid narrowing, *Annals of Internal Medicine* **123**, 649–655.

[27] Cowen, M. & Strawderman, R. (2002). Quantifying the physician contribution to managed care pharmacy expenses: a random effects approach, *Medical Care* **40**(8), 650–661.

[28] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of Royal Statistical Society Series B* **34**(2), 187–220.

[29] Criswell, L., Such, C., Neuhaus, J. & Yelin, E. (1997). Variation among rheumatologists in clinical outcomes and frequency of office visits for rheumatoid arthritis, *The Journal of Rheumatology* **24**(7), 1266–1271.

[30] D'Agostino, R.B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group, *Statistics in Medicine* **17**, 2265–2281.

[31] D'Aunno, T., Folz-Murphy, N. & Lin, X. (1999). Changes in methadone treatment practices: results from panel study 1988–1995, *The American Journal of Drug and Alcohol Abuse* **25**(4), 681–699.

[32] Davidson, W., Molloy, D.W. & Bedard, M. (1995). Physician characteristics and prescribing for elderly people in New Brunswick: relation to patient outcomes [see comments] [published erratum appears in *Can Med Assoc J*, **153**(2), 142]; *Canadian Medical Association Journal* **152**(8), 1227–1234.

[33] Davidson, W., Malloy, W., Somers, G. & Bédard, M. (1994). Relation between physician characteristics and prescribing for elderly people in New Brunswick, *Canadian Medical Association Journal* **150**(6), 917–921.

[34] Durrleman, S. & Simon, R. (1989). Flexible regression models with cubic splines, *Statistics in Medicine* **8**, 551–561.

[35] Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation, *The American Statistician* **37**, 36–48.

[36] Elliot, S., Fisher, M., Fredrick, S., Whaley, W., Krushat, M., Malenka, D., Fleming, C., Baron, J. & Hsai, D. (1992). The accuracy of medicare's hospital claims data: progress has been made, but problems

remain, *American Journal of Public Health* **82**(2), 243–248.

[37] Esdaile, J.M., Abrahamowicz, M., MacKenzie, T., Hayslett, P.J. & Kashgarian, M. (1994). The time-dependence of long-term prediction in lupus nephritis, *Arthritis & Rheumatism* **37**(3), 359–368.

[38] Ferguson, J.A. (1990). Patient age as a factor in drug prescribing practices, *Canadian Journal on Aging* **9**, 278–295.

[39] Gail, M.H., Wacholder, S. & Lubin, J.H. (1988). Indirect corrections for confounding under multiplicative and additive risk models, *American Journal of Industrial Medicine* **13**, 119–130.

[40] Goldberg, M.S., Carpenter, M., Theriault, G. & Fair, M. (1993). The accuracy of ascertaining vital status in a historical cohort study of synthetic textiles workers using computerized record linkage to the Canadian Mortality data base, *Canadian Journal of Public Health* **84**, 201–204.

[41] Goldman, S., Zadina, K., Copeland, J., Moritz, T. & Henderson, W. (1992). Aspirin in ischemic heart disease, *The New England Journal of Medicine* **327**(20), 1455–1456.

[42] Goldstein, H. & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance, *Journal of Royal Statistical Society, Series A* **159**, 385–443.

[43] Goldstein, H. (1986). Multilevel mixed linear modelling analysis using iterative generalized least squares, *Biometrika* **73**, 43–56.

[44] Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. Charles Griffin & Company, London.

[45] Goldstein, H. (1995). *Multilevel Statistical Models*. Edward Arnold, London.

[46] Gordon, M. (1987). Principles in prescribing for the older patient, *Drug Protocol* **2**, 15–24.

[47] Granek, B., Baker, S.P., Abbey, H., Robinson, B., Myers, A.H., Samkoff, J.S. & Klein, L. (1987). Medications and diagnoses in relation to falls in a long-term care facility, *Journal of the American Geriatrics Society* **35**(6), 503–511.

[48] Gray, R.J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis, *Journal of the American Statistical Association* **87**(420), 942–951.

[49] Guess, H.A., West, R., Strand, L.M., Helston, D., Lydick, E.G., Bergman, U. & Wol Ski, K. (1988). Fatal upper gastrointestinal hemorrhage or perforation among users and nonusers of nonsteroidal anti-inflammatory drugs in Saskatchewan, *Journal of Clinical Epidemiology* **41**(1), 35–45.

[50] Gurwich, E.L. (1983). Comparison of medication histories acquired by pharmacists and physicians, *American Journal of Hospital Pharmacy* **40**, 1541–1542.

[51] Gurwitz, J.H. & Avorn, J. (1991). The ambiguous relation between aging and adverse drug reactions, *Annals of Internal Medicine* **114**, 956–966.

[52] Gurwitz, J., Col, N.F. & Avorn, J. (1992). The exclusion of the elderly and women from clinical trials in acute myocardial infarction, *Journal of American Medical Association* **268**(11), 1417–1422.

[53] Hanley, J., Negassa, A., deB Edwardes, M. & Forrester, J. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation, *American Journal of Epidemiology* **157**(4), 364–375.

[54] Hansen, J., Olivarius, N., Siersma, V. & Andersen, J. (2003). Doctors' characteristics do not predict long-term glycaemic control in type 2 diabetic patients, *The British Journal of General Practice* **53**(486), 47–49.

[55] Hartzema, A.G. & Christensen, D.B. (1983). Nonmedical factors associated with the prescribing volume among family practitioners in an HMO, *Medical Care* **21**(10), 990–1000.

[56] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[57] Hastie, T.J. & Tibshirani, R.J. (1993). Varying-coefficient models (with discussion), *Journal of Royal Statistical Society, Series B* **55**(4), 757–796.

[58] Hauptmann, M., Lubin, J.H., Rosenberg, P., Wellmann, J. & Kreienbrock, L. (2000). The use of sliding time windows for the explanatory analysis of temporal effects of smoking histories on lung-cancer risk, *Statistics in Medicine* **19**, 2185–2194.

[59] Hauptmann, M., Pohlabein, H., Lubin, J., Jockel, K., Ahrens, W., Bruske-Hohlfeld, I. & Wichmann, H. (2002). The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies, *American Journal of Industrial Medicine* **41**(2), 89–97.

[60] Hauptmann, M., Wellmann, J., Lubin, J.H., Rosenberg, P. & Kreienbrock, L. (2000). Analysis of exposure-time response relationships using a spline weight function, *Biometrics* **56**(4), 1105–1108.

[61] Health Canada (1993). National Pharmaceutical Strategy Discussion Document. The National Pharmaceutical Strategy Office.

[62] Heinzl, H., Kaider, A. & Zlabinger, G. (1996). Assessing interactions of binary time-dependent covariates with time in cox proportional hazards regression models using cubic spline functions, *Statistics in Medicine* **15**, 2589–2601.

[63] Hess, K.R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions, *Statistics in Medicine* **13**, 1045–1062.

[64] Holcomb, J.P.J. (1999). Regression with covariates and outcome calculated from a common set of variables measured with error: estimation using the SIMEX methods, *Statistics in Medicine* **18**(21), 2847–2862.

[65] Honkanen, R., Ertama, L., Linnoila, M., Alha, A., Lukkari, I., Karlsson, M., Kiviluoto, O. & Puro, M. (1980). Role of drugs in traffic accidents, *British Medical Journal* **281**, 1309–1312.

[66] Hutchinson, T.A., Leventhal, J.M., Kraemer, M.S., Karch, F.E., Lipman, A.G. & Feinstein, A.R. (1979). An algorithm for the operational assessment of adverse drug reactions 2. Demonstration of reproducibility and validity, *Journal of American Medical Association* **242**(7), 633–638.

[67] Hylan, T., Crown, W., Meneades, L., Heiligenstein, J., Melfi, C., Croghan, T. & Buesching, D. (1999). SSRI antidepressant drug use patterns in the naturalistic setting: a multivariable analysis, *Medical Care* **37**(4 Suppl Lilly), AS36–AS44.

[68] Imhoff, M., Bauer, M., Gather, U. & Lohlein, D. (1998). Statistical pattern detection in univariate time series of intensive care on-line monitoring data, *Intensive Care Medicine* **24**, 1305–1314.

[69] Inui, T.S., Carter, W.B., Pecoraro, R.E., Pearlman, R.A. & Dohan, J.J. (1980). Variations in patient compliance with common long-term drugs, *Medical Care* **18**, 986–993.

[70] Jennrich, R.I. & Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics* **42**(4), 805–820.

[71] Kendrick, R. & Bayne, J.R.D. (1982). Compliance with prescribed medication by elderly patients, *Canadian Medical Association Journal* **127**, 961–962.

[72] Kleinbaum, D.G., Kupper, L.L. & Muller, K.E. (1988). *Applied Regression Analysis and other Multivariable Methods*, 2nd Ed. ISBN 0-87150-123-6, PWS-KENT Publishing Company, Boston.

[73] Kooperberg, C. & Clarkson, B. (1997). Hazard regression with interval-censored data, *Biometrics* **53**, 1485–1494.

[74] Kooperberg, C., Stone, C.J. & Truong, Y.K. (1995). Hazard regression, *Journal of the American Statistical Association* **90**(429), 78–94.

[75] Kuchenhoff, H. & Ulm, K. (1997). Comparison of statistical methods for assessing threshold limiting values in occupational epidemiology, *Computational Statistics* **12**, 249–264.

[76] Laird, N. (1992). Longitudinal studies with continuous responses, *Statistical Methods in Medical Research* **1**, 225–247.

[77] Laird, N.M. & Ware, J. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[78] Langholz, B., Thomas, D., Xiang, A. & Stram, D. (1999). Latency analysis in epidemiology studies of occupational exposures: application to the Colorado Plateau uranium miners cohort, *American Journal of Industrial Medicine* **35**(3), 246–256.

[79] Law, M.R., Wald, N.J. & Thompson, S.G. (1994). By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *British Medical Journals* **308**(6925), 367–372.

[80] Leffondré, K., Abrahamowicz, M. & Siemiatycki, J. (2003). Comparison of cox's model and logistic regression for case-control data with time-dependent covariates: a simulation study, *Statistics in Medicine* **22**(24), 3781–3794 .

[81] Leffondré, K., Abrahamowicz, M., Siemiatycki, J. & Rachet, B. (2002). Modeling smoking history: a comparison of different approaches, *American Journal of Epidemiology* **156**(9), 813–823.

[82] Lennard-Barrett, G.T. (1962). Dimensions of therapist response as causal factors in therapeutic change, *Psychology Monographs* **76**, 1–36.

[83] Lewis, R.F., Abrahamowicz, M., Côté, R. & Battista, R.N. (1997). Predictive power of duplex ultrasonography in asymptomatic carotid disease, *Annals of Internal Medicine* **127**, 13–20.

[84] Lexchin, J. (1992). Prescribing and drug costs in the province of Ontario, *International Journal of Health Services* **22**(3), 471–487.

[85] Liang, K.Y., Zeger, S.L. & Qaquish, B.F. (1992). Multivariate regression analyses for categorical data (with discussion), *Journal of the Royal Statistical Society* **B(54)**, 3–10.

[86] Lin, D.Y. & Wei, L.J. (1991). Goodness-of-fit tests for the general cox regression model, *Statistica Sinica* **1**, 1–17.

[87] Ma, R., Krewski, D. & Burnett, R.T. (2003). Random effects of cox models: a poison regression modeling, *Biometrika* **90**(1), 157–169.

[88] MacKenzie, T. & Abrahamowicz, M. (2002). Marginal and hazard ratio specific random data generation: applications to semi-parametric bootstrapping, *Statistical Computing* **12**(3), 245–252.

[89] Mandl, K.D. & Lee, T.H. (2002). Integrating medical informatics and health services research: the need for dual training at the clinical health systems and policy levels, *Journal of the American Medical Informatics Association* **9**(2), 127–132.

[90] Mayo, N.E., Levy, A.R. & Goldberg, M.S. (1991). Aspirin and hemorrhagic stroke, *Stroke* **22**(9), 1213–1214.

[91] Michelozzi, P., Forastiere, F., Fusco, D., Perucci, C., Ostro, B., Ancona, C. & Pallotti, G. (1998). Air pollution and daily mortality in Rome, Italy, *Occupational and Environmental Medicine* **55**(9), 605–610.

[92] Miles, D.L. (1977). Multiple prescriptions and drug appropriateness, *Health Services Research* **12**, 3–10.

[93] Mokkink, H., Smits, A., Grol, R., Meyboom, W. & van Son, J. (1990). *University of Nijmegen. Practice Performance and Quality of Care: Practice-Styles of Family Physician*. Can-Heal Publications, Ottawa, Book Number 0-921495-01-4.

[94] Mott, D. & Cline, R. (2002). Exploring generic drug use behavior: the role of prescribers and pharmacists in the opportunity for generic drug use and generic substitution, *Medical Care* **40**(8), 662–674.

[95] Neuhaus, J. (1992). Statistical methods for longitudinal and clustered designs with binary responses, *Statistical Methods in Medical Research* **1**, 249.

[96] Neutel, C.I., Hirdes, J.P., Maxwell, C.J. & Patten, S.B. (1996). New evidence on benzodiazepine use and falls: the time factor, *Age and Aging* **25**(273), 277.

[97] Nolan, L. & O'Malley, K. (1988). Prescribing for the elderly Part 1: sensitivity of the elderly to adverse drug reactions, *Journal of the American Geriatrics Society* **36**, 142–149.

[98] Pereira, L., Loomis, D., Conceicao, G., Braga, A., Arcas, R., Kishi, H., Singer, J., Bohm, G. & Saldiva, P. (1998). Association between air pollution and intrauterine mortality in Sao Paulo, Brazil, *Environmental Health Perspectives* **106**(6), 325–329.

[99] Perneger, V.P., Abrahamowicz, M., Bartlett, G. & Yerly, S. (2000). Time-dependence of survival predictions based on markers of HIV disease. Swiss HIV Cohort Study, *Journal of Investigative Medicine* **48**(3), 207–212.

[100] Quantin, C., Abrahamowicz, M., Moreau, T., Bartlett, G., MacKenzie, T., Tazi, M.A., Lalonde, L. & Faivre, J. (1999). Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models, *American Journal of Epidemiology* **150**(11), 1188–1200.

[101] Quinn, K., Baker, M.J. & Evans, B. (1992). A population-wide profile of prescription drug use in Saskatchewan, *Canadian Medical Association Journal* **146**, 2177–2186.

[102] Rachet, B., Abrahamowicz, M., Sasco, A.J. & Siemiatycki, J. (2003). Estimating the distribution of lag in the effect of the short-term exposures and interventions: adaptation of a non-parametric regression spline model, *Statistics in Medicine* **22**(14), 2335–2363.

[103] Rachet, B., Sasco, A.J., Abrahamowicz, M. & Benyamine, D. (1998). Prognostic factors in nasopharyngeal cancer: accounting for time-dependence of relative risks, *International Journal of Epidemiology* **27**, 772–780.

[104] Ramsey, J.O. & Abrahamowicz, M. (1989). Binomial regression with monotone splines: a psychometric application, *Journal of the American Statistical Association* **84**(408), 906–918.

[105] Ramsey, J.O. (1988). Monotone regression splines in action (with discussion), *Statistical Science* **3**, 425–461.

[106] Raudenbush, S. & Bryk, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Thousand Oaks, London.

[107] Ray, W.A., Fought, R.L. & Decker, M.D. (1992). Psychoactive drugs and the risk of injurious motor vehicle crashes in elderly drivers, *American Journal of Epidemiology* **136**(7), 873–883.

[108] Ray, M.A., Griffin, M.R. & Downey, W. (1989). Benzodiazepines of long and short elimination half-life and the risk of hip fracture, *Journal of American Medical Association* **262**, 3303–3307.

[109] Ray, W.A., Griffin, M.R., Downey, W. & Melton, L.J. (1989). Long-term use of thiazide diuretics and risk of hip fracture, *Lancet* **1**(8640), 687–690.

[110] Romano, P.S., Roos, L.L. & Jollis, J.G. (1993). Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives, *Journal of Clinical Epidemiology* **46**(10), 1075–1079.

[111] Roos, L.L., Nicol, J.P. & Cageorge, S.M. (1987). Using administrative data for longitudinal research: comparisons with primary data collection, *Journal of Chronic Diseases* **40**, 41–49.

[112] Roos, N.P. & Shapiro, E. (1981). The Manitoba longitudinal study on aging, *Medical Care* **19**(6), 644–657.

[113] Roos, N.P., Shapiro, E. & Tate, R. (1989). Does a small minority of elderly account for a majority of health care expenditures? A sixteen-year perspective, *The Milbank Quarterly* **67**(3–4), 347–369.

[114] Rosenbaum, P. & Rubin, D.B. (1983). The central role of the propensity score in non-experimental studies for causal effects, *Biometrika* **70**, 41–55.

[115] Rosholm, J., Gram, L., Isacsson, G., Hallas, J. & Bergman, U. (1997). Changes in the pattern of antidepressant use upon the introduction of the new antidepressants: a prescription database study, *European Journal of Clinical Pharmacology* **52**(3), 205–209.

[116] Rosner, B. & Milton, R. (1988). Significance testing for correlated binary outcome data, *Biometrics* **44**, 505–512.

[117] Rotmensch, H.H., Mendelevitch, I., Silverberg, D.S. & Liron, M. (1996). Prescribing pattern of antihypertensive drugs in the community, *Journal of Human Hypertension* **10**, S169–S172.

[118] Royston, P. (2000). A strategy for modelling the effect of a continuous covariate in medicine and epidemiology, *Statistics in Medicine* **19**, 1831–1847.

[119] Royston, P., Ambler, G. & Sauerbrei, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology, *International Journal of Epidemiology* **28**, 964–974.

[120] Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores, *Annals of Internal Medicine* **127**(8), 757–763.

[121] Savitz, D., Checkoway, H. & Loomis, D. (1998). Magnetic field exposure and neurodegenerative disease mortality among electric utility workers, *Epidemiology* **9**(4), 398–404.

[122] Schneeweiss, S., Maclure, M. & Soumerai, S. (2002). Prescription duration after drug copay changes in older people: methodological aspects, *Journal of the American Geriatrics Society* **50**(3), 521–525.

[123] Schulenberg, J. & Maggs, J. (2001). Moving targets: modeling developmental trajectories of adolescent alcohol misuse, individual and peer risk factors, and intervention effects, *Applied Developmental Science* **5**(4), 237–253.

[124] Shader, R.I. & Greenblatt, D.J. (1993). Use of benzodiazepines in anxiety disorders, *The New England Journal of Medicine* **328**, 1398–1405.

[125] Shapiro, E., Tate, R.B. & Roos, N.P. (1987). Do nursing homes reduce hospital use? *Medical Care* **25**(1), 1–8.

[126] Shorr, R.I., Bauwens, S.F. & Landefeld, C.S. (1990). Failure to limit quantities of benzodiazepine hypnotic drugs for outpatients: placing the elderly at risk, *The American Journal of Medicine* **89**, 725–732.

[127] Sift, R., van Staa, T.-P., Abenhaim, L. & Ebner, D. (1997). A study of the longitudinal utilization and switching-patterns of non-steroidal anti-inflammatory drugs using a pharmacy based approach, *Pharmacoepidemiology and Drug Safety* **6**, 263–268.

[128] Silverstone, F.E., Graham, D.Y., Senior, J.R., Davies, H.W., Struthers, B.J., Bittman, R.M. & Geis, S. (1996). Misoprostol reduces serious gastrointestinal complications in patients with rheumatoid arthritis receiving nonsteroidal anti-inflammatory drugs, *Annals of Internal Medicine* **123**, 241–249.

[129] MacKenzie, T. & Abrahamowicz, M. Simultaneous relaxation of proportional hazards and log-linearity, in *Annual Meeting of the Statistical Society of Canada*, Waterloo, June 2, 1996.

[130] Skegg, D.C.G. (2001). Evaluating the safety of medicines, with particular reference to contraception, *Statistics in Medicine* **20**, 3557–3569.

[131] Sleeper, L.A. & Harrington, D.P. (1990). Regression splines in the cox model with application to covariate effects in liver disease, *American Statistical Association* **85**, 941–949.

[132] Snijders, T., Boster, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publication, London.

[133] Sova, G. (1989). *Corpus Almanac and Canadian Sourcebook*, Vol. 1. Corpus Information Services, Don Mills.

[134] Spitzer, W.O., Suissa, S., Ernst, P., Horwitz, R.I., Habbick, B., Cockcroft, D. & Boivin, J. (1992). The use of B-Agonists and the risk of death and near death from asthma, *The New England Journal of Medicine* **326**, 501–506.

[135] Stefanski, L. & Cook, J. (1995). Simulation extrapolation: the measurement error jackknife, *Journal of American Statistical Association* **90**, 1247–1256.

[136] Strasen, L. (1988). Incorporating patient satisfaction standards into quality of care measures, *JONA* **18**(11), 5–6.

[137] Tamblyn, R. (1996). Medication use in seniors: challenges and solutions, *Therapie* **51**, 269–282.

[138] Tamblyn, R., Abrahamowicz, M., Brailovsky, C., Grand'Maison, P., Lescop, J., Norcini, J.J., Girard, N. & Haggerty, J. (1998). The association between licensing examination scores and resource use and quality of care in primary care practice, *Journal of American Medical Association* **280**(11), 989–996.

[139] Tamblyn, R., Abrahamowicz, M., Dauphinee, W.D., Hanley, J.A., Norcini, J., Girard, N., Grand'Maison, P. & Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care, *Journal of American Medical Association* **288**(23), 3019–3026.

[140] Tamblyn, R.M., Lavoie, G., Petrella, L. & Monette, J. (1995). The use of prescription claims databases in pharmacoepidemiological research: the accuracy and comprehensiveness of the prescription claims database in Quebec, *Journal of Clinical Epidemiology* **48**(8), 999–1009.

[141] Tamblyn, R.M., McLeod, P.J., Abrahamowicz, M. & Laprise, R. (1996). Do too many cooks spoil the broth? Multiple physician involvement in medical management and inappropriate prescribing in the elderly, *Canadian Medical Association Journal* **154**(8), 1177–1184.

[142] Tamblyn, R.M., McLeod, P., Abrahamowicz, M., Monette, J., Gayton, D., Berkson, L., Grad, R., Huang, A., Isaac, L., Schnarch, B. & Snell, L. (1994). Questionable prescribing for elderly patients in Quebec, *Canadian Medical Association Journal* **150**(11), 1801–1809.

[143] The Scandinavian Simvastatin Survival Study Group (1994). Randomized trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin survival study (4S), *Lancet* **334**(8934), 1383–1389.

[144] van Gijn, T. (1992). Aspirin: dose and indications in modern stroke prevention, *Neurological Clinics* **10**(1), 193–207.

[145] van Hulten, R., Leufkens, H. & Bakker, A. (1998). Usage patterns of benzodiazepines in a dutch community: a 10-year follow-up, *Pharmacy World & Science* **20**(2), 78–82.

[146] Waternaux, C., Laird, N.M. & Ware, J.M. (1989). Methods for analysis of longitudinal data: blood-lead concentration and cognitive development, *Journal of the American Statistical Association* **84**, 33–41.

[147] Wegman, E.J. & Wright, I.W. (1983). Splines in statistics, *Journal of the American Statistical Association* **78**, 351–365.

[148] Wennberg, J.E., Roos, N.P., Sola, L., Schori, A. & Jaffe, R. (1987). Use of claims data systems to evaluate health care outcomes mortality and reoperation following prostatectomy, *Journal of American Medical Association* **257**(7), 933–936.

[149] West, S.L., Savitz, D.A., Koch, G., Strom, B.L., Guess, H.A. & Hartzema, A. (1995). Recall accuracy for prescription medications: self report compared with database information, *American Journal of Epidemiology* **142**(10), 1103–1110.

[150] Wilkins, R. (1993). Use of postal codes and addresses in the analysis of health data, *Health Reports* **5**(2), 157–177.

[151] Zammit-Lucia, J. & Dasgupta, R. (1995). Reference Pricing The European Experience Health Policy Review, Paper No. 10. St. Mary's Hospital Medical School, London.

[152] Zeger, S.L. & Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**, 121–130.

[153] Zeger, S.L., Liang, K.Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

[154] Zuanetti, G., Latini, R., Avanzini, F., Franzosi, M., Maggioni, A., Colombo, F., Nicolis, E. & Mauri, F. (1996). Trends and determinants of calcium antagonist usage after acute myocardial infarction (the GISSI experience), *The American Journal of Cardiology* **78**(2), 153–157.

Further Reading

Von Korff, M., Wagner, E.H. & Saunders, K. (1992). A chronic disease score from automated pharmacy data, *Journal of Clinical Epidemiology* **45**(2), 197–203.

MICHAL ABRAHAMOWICZ & ROBYN TAMBLYN

Dummy Variables

Usually **explanatory variables** take on values that are measured or observed quantities (e.g. height, weight, age). Sometimes, however, they are categorical, or qualitative, and as such have no natural numeric values associated with them (e.g. political affiliation, sex, race). An individual is simply identified as belonging to one specific category out of a set of, say, $K \geq 2$ mutually exclusive categories or levels. To allow such variables to be included in statistical models, a set of so-called dummy variables can be defined to provide numerical representations for a categorical variable.

If there is an intercept term included in a regression model and if a K-level categorical variable is to be included in the model, then $K - 1$ dummy variables must be defined to represent this categorical variable. For example, if we want to include sex in a regression model containing an intercept, we need one dummy variable to represent sex in the model. The $K - 1$ dummy variables are usually chosen to be linearly independent.

The typical choice for the definition of a dummy variable is the use of an indicator variable (0, 1 representation) which indicates whether a particular observation belongs to a specific level of the categorical variable. If an intercept is included in a regression model, then indicators are derived for $K - 1$ of the K possible levels. The remaining level, not associated with an indicator, is termed the referent category or the baseline category. For example, suppose data are available for n individuals to be included in a **linear regression** model and suppose the relationship between sex of the individual and the response variable is of interest. An intercept term, $X_{0i} = 1$, for all $i = 1, \ldots, n$, is included in the design matrix (*see* **General Linear Model**) so we need one dummy variable to represent sex. If we let X_{1i} represent sex in a linear model, then we may choose to define the dummy variable as

$$X_{1i} = \begin{cases} 1, & \text{if individual } i \text{ is female,} \\ 0, & \text{otherwise.} \end{cases}$$

So, if Y_i is the response for individual i, then the model that would be investigated is

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_{1i}.$$

In this case the male category is the referent category. The choice of the referent category often depends on the study situation since specific choices may mean that regression coefficients have a more direct interpretation. This is particularly the case when comparisons relative to some control or unexposed group are of interest. In this case it is natural to have the unexposed category as the baseline.

The use of indicator dummy variables is also valuable if models for each level of a categorical variable are to be derived and compared. Kleinbaum et al. [1] present such an example where the relationship between systolic blood pressure (SBP) and age are compared for females and males. In the data which they present, there are 29 females and 40 males and therefore a total of 69 observations. Let Y_i represent SBP and let X_{1i} represent age for individual $i, i = 1, \ldots, 69$. Also, define

$$X_{2i} = \begin{cases} 1, & \text{if individual } i \text{ is female,} \\ 0, & \text{otherwise.} \end{cases}$$

Now, consider the model

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i}.$$

Substituting the appropriate X_{2i} value for males, the model is

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_{1i},$$

whereas for females the model is

$$\mathrm{E}(Y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i}.$$

By fitting one model, we can investigate whether the linear model for females has the same slope as the linear model for males ($\mathrm{H}_0 : \beta_3 = 0$) and/or whether the intercept for the model for females is the same as the intercept for the model for males ($\mathrm{H}_0 : \beta_2 = 0$). This example assumes a linear regression model. Note that identical arguments regarding tests for equal intercepts and slopes over varying levels of a qualitative variable can be made for other link functions in the **generalized linear models** framework.

So far, only a two-level categorical variable has been considered (i.e. sex). In [1], an example of a three-level categorical variable is presented. The categories are regions of the US and are listed as western, central, or eastern. One way of representing

this region variable is to define two dummy variables as

$$X_1 = \begin{cases} 1, & \text{if residence is western region,} \\ 0, & \text{otherwise,} \end{cases}$$

and

$$X_2 = \begin{cases} 1, & \text{if residence is central region,} \\ 0, & \text{otherwise.} \end{cases}$$

Note that the eastern region is the referent or baseline category in this representation.

While indicators have been used so far to define dummy variables, other representations can be used. For example, for sex one can use the dummy variable definition:

$$X_{\text{sex}} = \begin{cases} 1, & \text{if female,} \\ -1, & \text{if male.} \end{cases}$$

With this specification, the regression coefficient for X_{sex} provides the departure from an average outcome level (intercept) associated with being female or male. For the three-level region variable one could similarly define two dummy variables as:

$$X_{\text{region 1}} = \begin{cases} 1, & \text{if residence is western region,} \\ 0, & \text{if residence is central region,} \\ -1, & \text{if residence is eastern region.} \end{cases}$$

and

$$X_{\text{region 2}} = \begin{cases} 0, & \text{if residence is western region,} \\ 1, & \text{if residence is central region,} \\ -1, & \text{if residence is eastern region.} \end{cases}$$

Global tests concerning the categorical variable (i.e. testing the hypothesis that all regression coefficients corresponding to a set of dummy variables are equal to 0) will not depend on the choice of dummy variable definitions. Since the goal is to investigate a relationship between a response variable and a categorical variable, then this invariance to the choice of dummy variable representation is important. The interpretation of the coefficient for each dummy variable will, however, depend on the coding selected for a categorical variable.

Reference

[1] Kleinbaum, D.G., Kupper, L.L. & Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*, 2nd Ed. PWS-Kent, Boston.

G.A. DARLINGTON

Duration Dependence

A variety of techniques are available in survival analysis to assess the dependence on **covariates** of the time to some defined event, measured from a defined origin. However, in many problems more than one event may be measured on each subject. For example, a patient with Parkinson's disease is likely to require levodopa therapy within two years of diagnosis. The actual time to this event may be regarded as a measure, albeit imperfect, of the rate of progression of disability. At some later point in the disease process, side-effects of levodopa therapy, such as dyskinesias, may appear, or levodopa may lose its effectiveness in controlling symptoms, leading to the occurrence of end-of-dose "wearing-off" and the more extreme "on–off" phenomenon. Other milestones, such as "freezing", may occur before or after initiation of levodopa therapy. Ultimately, possibly after many years, the patient dies.

A simpler example is common in cancer studies. Patients, once diagnosed with a cancer, alternate between periods of remission and of relapse into active disease, and ultimately die.

There are many choices in modeling such relationships among the several outcomes, and the effect on these outcomes of covariates and treatments administered. First, one may use a single time origin for all types of failure, or reset the clock to zero when certain events occur. For example, in Parkinson's disease, time to freezing, which may occur before or after initiation of levodopa, is most naturally measured from the date of diagnosis. Time of initiation of levodopa may be a preferable origin for assessing time to levodopa-related side-effects, which by definition can occur only after initiation of levodopa. However, if primary interest centers on the occurrence of later events, modeling the full times to these events directly may lead to results that are more interpretable (cf. Parkinson Study Group [20]). The time origin for a randomized study is best taken at the date of randomization to preserve the **randomization** justification for the analysis.

One very simple model is that an individual moves among different states m (e.g. remission, relapse, death) according to a discrete epoch **Markov chain** and that the sojourn time spent in each state is **exponentially distributed**. Then the process $M(t)$ indicating the state at time t becomes a continuous time Markov chain, with death as an absorbing state. Remission and relapse may each be transient, as any number of transitions between these two states may occur before the ultimate transition to the third state. Lagakos et al. [12] allowed the sojourn times to have arbitrary distributions, leading to **semi-Markov** rather than **Markov processes**. It is also straightforward to allow the distribution of the sojourn times to vary with the number of previous transitions, so that successive periods of remission could become shorter, and periods of relapse could lengthen. Nonhomogeneous Markov processes may also be considered – see, for example, the discussion of "dynamic stratification" models below.

One typically observes a fairly short sequence of transitions on each of a large number of individuals. However, standard parametric or nonparametric procedures for estimation and inference may be applied. One may solve the inference problem for the full process by applying standard procedures to each of its probabilistic components – the transition matrix of the embedded discrete Markov chain is estimated by a discrete analog estimate, and the sojourn time distributions are estimated by the usual **Kaplan–Meier estimates** from the sojourn data for each state. These estimates have all the usual desirable features, including asymptotic normality and the availability of a simple variance estimate via Greenwood's formula (*see* **Aalen–Johansen Estimator**) for the sojourn time distributions. The estimates for the several components are asymptotically independent.

Inclusion of baseline covariates in the models via a proportional hazards assumption following Cox [5] is straightforward. Care must be taken with time-dependent covariates. Consider, for example, the modulated renewal processes suggested by Cox [6]. Here there is just a single type of event, which may occur many times, and the intensity of an event at time t is given by

$$h(t) = \exp\{\beta Z[t, \mathcal{H}(t)]\} h_0(U_t).$$

Here U_t is the backwards recurrence time – the time from the immediately preceding event – and $Z[t, \mathcal{H}(t)]$ is any function of the history $\mathcal{H}(t)$ of events experienced by that individual before time t. For example, $Z[t, \mathcal{H}(t)]$ could be the length of the immediately preceding interval, in which case the sequence of interevent times will follow a discrete time continuous state Markov chain. Cox suggested

that the usual **partial likelihood** approach be applied on the reordered time scale U_t. A difficulty pointed out by Oakes [16] and others is that reordering the time scale this way destroys the Markov property needed for the validity of this approach – Oakes [16] gave a simple example involving matched pairs (*see* **Matching**), where naïve use of this procedure would lead to inconsistent estimates (*see* **Consistent Estimator**). However, Oakes & Cui [19] showed that for a single long stationary sequence of events the standard asymptotic theory would hold for Cox's method, even though the standard martingale theory does not apply. Dabrowska et al. [7] allowed time-dependent covariates in a five-state Markov renewal model for bone marrow transplant data. By requiring that the covariates depend only on the backwards recurrence time to the previous event, they preserved the martingale structure of the process. Klein et al. [11] give a somewhat parallel application.

Prentice et al. [22] considered two models for the dependence of the intensity of an event at time t on the $\mathcal{H}(t)$, namely

$$h[t, \mathcal{H}(t)] = h_{0s}(t) \exp[\beta_s Z(t)] \tag{1}$$

and

$$h[t, \mathcal{H}(t)] = h_{0s}(U_t) \exp[\beta_s Z(t)], \tag{2}$$

where in each case the $h_{0s}(t)$ are arbitrary baseline functions – in the first case of the total time on study, and in the second case of the time U_t since the immediately preceding event. Here $s = s[\mathcal{H}(t)]$ is a stratum indicator, which may change over time for a given subject.

Gail et al. [8] also considered the model in (2). Andersen & Gill [2] discussed the special case of (1) with only a single stratum. There is no problem with applying standard partial likelihood approaches to (1). In (2) the partial likelihood can be applied so long as the $Z(t)$ are exogenous covariates (i.e. do not depend on the history of the process), or if the **stratification** is sufficiently fine to ensure that each individual may experience at most one failure in a given stratum. Voelkel & Crowley [25] give a careful discussion. In many examples $s = j$, which is one plus the number of previous failures experienced by that individual. Many of these models can now be seen as special cases of a general counting process formulation; see, for example, Andersen et al. [3, Sections IV.4 and X.1].

Direct modeling of the time to successive events from a common origin was introduced by Wei et al. [26] and has recently been popularized by Therneau [24]. They model the marginal (unconditional) intensity of the jth event at time t as

$$h_j(t) = h_{0j}(t) \exp[\beta_j Z_j(t)]. \tag{3}$$

Under this model separate partial likelihood estimates can be written down corresponding to the first, second, ..., jth, ... event occurring to each individual, and the usual asymptotic properties hold for each j separately. These estimates are also jointly asymptotically normal but, unlike in the model in (1), they are not independent. However, their correlation matrix can be consistently estimated and used to derive tests of hypotheses such as $\beta_1 = \beta_2 = \ldots = \beta_k$ and combined estimates of a common β. Hughes [10] examined the gain (or occasionally losses) in **power** from incorporating second events in the analysis of a randomized **clinical trial**, where covariate effects act on both first and second events through Wei et al.'s model.

Pepe & Cai [21], Oakes [18], and others have criticized the counterintuitive property of Wei et al.'s procedure that individuals are assumed to be at risk of the jth event before they have experienced the $(j-1)$th event. Lin [15] pointed out that this could be avoided by redefining risk set indicators appropriately, but this in effect converts the model to (1) and loses the "marginal" interpretation of the covariate effects. Pepe & Cai [21] proposed conditioning on the number $j-1$ of previous events, but not on their times of occurrence – a compromise between the fully conditional models of counting process theory and the marginal models of Wei et al. However, there is in general no simple relationship between this intensity and the corresponding marginal intensity.

For the model in (1), Pepe & Cai's partly conditional intensity becomes the full conditional intensity. This model, which was also discussed by Clayton [4], can be fitted by the dynamic stratification procedure in BMDP 2L (*see* **Software, Biostatistical**). Oakes [18] showed that the assumption of proportional hazards within a dynamic stratification model is not consistent with the assumption of proportional hazards within Wei et al.'s model.

Oakes [16] indicated how such models also arise naturally from a frailty (heterogeneity) interpretation.

Suppose that there is an unobserved individual-specific random variable W and that the full conditional intensity of an event at time t is just

$$h[t, \mathcal{H}(t), W] = Wb(t)$$

for some baseline intensity $b(t)$. Then the number of events in $(0, t)$ carries all predictive information from $\mathcal{H}(t)$ regarding W, so that the model of (1) applies (without covariates). An interesting special case, due in essence to Greenwood & Yule [9], is when the unobserved W has a **gamma distribution** (*See* **Accident Proneness**). In this case the various $h_j(t)$ turn out to be proportional to each other [though not to $b(t)$]. Covariates can then be included in a proportional hazards model for these conditional intensities and their coefficients estimated together with the index of the gamma frailty distribution from a single partial likelihood. See Oakes [17].

This model must be distinguished from one in which the covariates act on $b(t)$. Estimation procedures for this model have been proposed by Lawless [13, 14] and Self & Prentice [23]. Aalen & Husebye [1] allowed a frailty term in a simple renewal process, giving a conditional intensity of the form

$$h[t, \mathcal{H}(t), W] = Wh_0(U_t),$$

with a power law form for $h_0(t)$ leading to *Weibull* interevent times. These authors also make the very important point that the common practice of omitting the last incomplete interval from the analysis of a sequence of interevent intervals can lead to *biased* estimates, because $t - U_t$ is not a stopping time in terms of counting process theory (cf. Andersen et al. [2]).

References

[1] Aalen, O.O. & Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes, *Statistics in Medicine* **10**, 1227–1240.

[2] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[3] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Clayton, D.G. (1988). The analysis of event history data: a review of progress and outstanding problems, *Statistics in Medicine* **7**, 819–841.

[5] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[6] Cox, D.R. (1972). The statistical analysis of dependencies in point processes, in *Stochastic Point Processes*, P.A.W. Lewis, ed. Wiley, New York, pp. 55–66.

[7] Dabrowska, D.M., Sun, G.W. & Horowitz, M.H. (1992). Cox regression analysis in a Markov renewal model with application to the analysis of bone marrow transplant data, *Journal of the American Statistical Association* **89**, 876–877.

[8] Gail, M.H., Santner, T.J. & Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor, *Biometrics* **36**, 255–266.

[9] Greenwood, M. & Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease as repeated accidents, *Journal of the Royal Statistical Society* **83**, 255–279.

[10] Hughes, M.D. (1995). Power considerations for clinical trials using multivariate time-to-event data. *Statistics in Medicine* **16**, 865–882.

[11] Klein, J.P. Keiding, N. & Copelan, E.A. (1993). Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients, *Statistics in Medicine* **12**, 2315–2332.

[12] Lagakos, S.W., Sommer, C.J. & Zelen, M. (1978). Semi-Markov models for partially censored data, *Biometrika* **65**, 311–317.

[13] Lawless, J.F. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association* **82**, 808–815.

[14] Lawless, J.F. (1995). The analysis of recurrent events for multiple subjects, *Applied Statistics* **44**, 487–498.

[15] Lin, D.Y. (1995). Multivariate failure time data, in *Recent Advances in Clinical Trial Design and Analysis*, Thall, P.F., ed. Kluwer, Dordrecht, pp. 73–93.

[16] Oakes, D. (1981). Survival times: aspects of partial likelihood (with discussion), *International Statistical Review* **49**, 235–264.

[17] Oakes, D. (1992). Frailty models for multivariate event times, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Dordrecht, pp. 371–379.

[18] Oakes, D. (1997). Model-based and/or marginal analysis of multiple event-time data, in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D.Y. Lin & T.R. Fleming, eds. Springer-Verlag, New York, pp. 85–98.

[19] Oakes, D. & Cui, L. (1994). On semiparametric inference for modulated renewal processes, *Biometrika* **81**, 83–90.

[20] Parkinson Study Group (1996). Impact of deprenyl and tocopherol treatment on Parkinson's disease in DATATOP patients requiring levodopa, *Annals of Neurology* **39**, 29–36.

[21] Pepe, M.S. & Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time-dependent covariates, *Journal of the American Statistical Association* **88**, 811–820.

[22] Prentice, R.L., Williams, B.J. & Peterson, A.V. (1981). On the regression analysis of multivariate failure time data, *Biometrika* **68**, 373–379.

[23] Self, S.G. & Prentice, R.L. (1986). Incorporating random effects into multivariate relative risk regression models, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds. Wiley, New York, pp. 167–178.

[24] Therneau, T. (1997). Extending the Cox model, in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D.Y. Lin & T.R. Fleming, eds. Springer-Verlag, New York, pp. 51–84.

[25] Voelkel, J.G. & Crowley, J. (1984). Nonparametric inference for a class of semi-Markov processes with censored observations, *Annals of Statistics* **12**, 142–160.

[26] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling of marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

DAVID OAKES

Durbin–Watson Test

The standard linear multiple regression model assumes that the conditional expectation of a response variable Y is a linear function of k explanatory variables, $x_1, \ldots, x_k$. The "errors" are usually assumed to be independent $\mathrm{N}(\mu, \sigma^2)$ random variables. When a multiple regression model is fitted to time series data, it is often found that successive **residuals** through time are not independent but are (auto)correlated. There may, for example, be a run of positive residuals followed by a run of negative residuals, which corresponds to positive autocorrelation (*see* **Autocorrelation Function**). With correlated errors it may be possible to improve the model to get a better fit and better forecasts and the Durbin–Watson test is one way of checking for this.

Let $(y_t, x_{1t}, \ldots, x_{kt})$ denote the observed values of all the variables at time t for $t = 1, 2, \ldots, n$. Then the residual at time t is the difference between the observed value of y_t and the value predicted by the fitted regression model, and may be calculated as

$$\hat{z}_t = y_t - \hat{\beta}_1 x_{1t} - \cdots - \hat{\beta}_k x_{kt}. \tag{1}$$

The Durbin–Watson statistic is then given by

$$d = \frac{\sum_{t=2}^{n} (\hat{z}_t - \hat{z}_{t-1})^2}{\sum_{t=1}^{n} \hat{z}_t^2}. \tag{2}$$

The value of d is routinely calculated by most regression packages. It is unfortunate that many analysts misinterpret the result, as the value of d will be close to two, and not to zero, if the errors are independent. The sampling distribution of d under the null hypothesis of independence unfortunately depends on the value of k and on the x values as well as on n, so that it is not possible to give a single critical value for d. Instead, upper and lower critical values, say, d_{L} and d_{U}, have been tabulated for different values of k and n. If the observed value of d lies between d_{L} and d_{U}, then the test is annoyingly inconclusive.

Another difficulty with the Durbin–Watson test is that it is not strictly valid when the explanatory variables include lagged values of the response variable, as is often the case.

The time series analyst will generally be more familiar with checking for autocorrelation by looking at the autocorrelation function of the residuals. In particular, the first-order autocorrelation coefficient of the residuals measures the correlation between successive pairs of residuals and is given by

$$r_1 = \frac{\sum_{t=2}^{n} \hat{z}_t \hat{z}_{t-1}}{\sum_{t=1}^{n} \hat{z}_t^2}. \tag{3}$$

It can be shown that this statistic is related to the Durbin–Watson statistic d by

$$d \simeq 2(1 - r_1). \tag{4}$$

Positive autocorrelation ($r_1 > 0$) corresponds to a value of d less than two, and this is the more normal direction of departure from independence. To test the null hypothesis of independence against positive autocorrelation, we look up values of d_{L} and d_{U} depending on k and n and then, if $d < d_{\mathrm{L}}$, we reject H_0, while, if $d > d_{\mathrm{U}}$, we accept H_0. Otherwise, the test is inconclusive.

When k gets large and n small, the range of inconclusive values can get alarmingly wide, but for k small and n large, it can be much easier to carry out a test on the value of r_1 using the result that, for reasonably large n, r_1 is approximately $\mathrm{N}(0, 1/n)$ for random data. The Durbin–Watson test is usually carried out in a one-tailed form. For $k = 2$ and $n = 100$, say, the one-tailed 5% values of d_{L} and d_{U} are 1.63 and 1.72 when testing for positive autocorrelation, whereas the approximate one-tailed 5% value for r_1 is $1.64/\sqrt{n} = 0.16$. The latter corresponds to a d value of $2(1 - 0.16) = 1.68$, which is halfway between the d_{L} and d_{U} values. This writer's preference is generally to look at the autocorrelation function of the residuals and use an approximate test on r_1 rather than carry out the much more complicated Durbin–Watson test. However, with several explanatory variables, the latter may be advisable and significance points up to $k = 5$ and $n = 100$ are tabulated, for example, by Kendall & Ord [4], as well as in the three original papers by Durbin & Watson [1–3], after whom the test is named.

References

[1] Durbin, J. & Watson, G.S. (1950). Testing for serial correlation in least squares regression. I, *Biometrika* **37**, 409–428.

[2] Durbin, J. & Watson, G.S. (1951). Testing for serial correlation in least squares regression. II, *Biometrika* **38**, 159–178.

[3] Durbin, J. & Watson, G.S. (1971). Testing for serial correlation in least squares regression. III, *Biometrika* **58**, 1–19.

[4] Kendall, M. & Ord, J.K. (1990). *Time Series*, 3rd Ed. Edward Arnold, London.

(*See also* **Cox's Test of Randomness**; **Noise and White Noise**; **Randomness, Tests of**)

CHRIS CHATFIELD

Dynamic Allocation Index

Dynamic allocation indices (or *Gittins indices*) arise when it is necessary to optimize in a sequential manner the allocation of effort between a number of competing projects. The effort and the projects may take a variety of forms. Examples are: an industrial processor and jobs waiting to be processed; a server with a queue of customers; an industrial laboratory with research projects; any busy person with jobs to do; a stream of patients and alternative treatments; a searcher who may look in different places. In every case, effort is treated as being homogeneous, and the problem is to allocate it between the different projects so as to maximize the expected total reward which they yield. It is a sequential problem, as effort is allowed to be reallocated in a feedback manner, taking account of the pattern of rewards so far achieved. The choice at each stage is determined partly on the basis of maximizing the expected immediate rate of return, and partly by the need to reduce uncertainty, and thereby provide a basis for better choices later on. It is the tension between these two requirements that makes the decision problem both interesting and difficult. The reallocations are assumed to be costless, and to take a negligible time, since the alternative is to impose a traveling-salesman-like feature, thereby adding a serious further level of complication.

The techniques that come under the heading of dynamic programming have been devised for sequential optimization problems. The key idea is a recurrence equation relating the expected total reward (call this the *payoff*) at a given decision time to the distribution of its possible values at the next decision time. Sometimes this equation may be solved analytically. Otherwise a recursive numerical solution may, at any rate in principle, be carried out. This involves making an initial approximation to the payoff function, and then successive further approximations by substituting in the right-hand side of the recurrence equation. As Bellman [1], for many years the chief protagonist of this methodology, pointed out, using the recurrence equation involves less computing than a complete enumeration of all policies and their corresponding payoffs, but nonetheless soon runs into the sands of intractable storage and processing requirements as the number of variables on which the payoff function depends increases.

For the problem of allocating effort to projects, the number of variables is at least equal to the number of projects. An attractive idea, therefore, is to establish a priority index for each project, depending on its past history but not that of any other project, and to allocate effort at each decision time only to the project with the highest current index value. To calculate these indices it should be possible to calibrate a project in a given state against some set of standard projects with simple properties. If this could be done we should have a reasonable policy without having to deal with any function of the states of more than one project.

Gittins & Jones [6] showed that for exponentially discounted independent projects a policy of this form is actually optimal.

Since they may change as more effort is allocated, these priority indices may aptly be, and often are, termed *dynamic allocation indices*. The main methods available for determining the indices are by (i) interchange arguments, (ii) exploiting any special features of the bandit processes concerned, in particular those which lead to the optimality of myopic policies, (iii) calibration by reference to standard bandit processes, often involving iteration using the dynamic programming recurrence equation, and (iv) using the fact that a dynamic allocation index may be regarded as a maximized equivalent constant reward rate.

A detailed account of the theory, calculation, and application of Gittins indices is given in [4]. An important further contribution to the theory has been made by Bertsimas & Niño-Mora [3], who also give a useful review of the recent literature.

An important area of application is the selection of compounds for screening as potential drugs in pharmaceutical research. Here the competing projects are different families of compounds with different, and initially unknown, distributions of the relevant activity. More details are given in [2] and [5].

References

[1] Bellman, R.E. (1957). *Dynamic Programming*. Princeton University Press, Princeton.

[2] Bergman, S.W. & Gittins, J.C. (1985). *Statistical Methods for Pharmaceutical Research Planning*. Marcel Dekker, New York.

[3] Bertsimas, D. & Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems, *Mathematics of Operations Research* **21**, 257–306.

[4] Gittins, J.C. (1989). *Multiarmed Bandit Allocation Indices*. Wiley, Chichester.

[5] Gittins, J.C. (1997). *CPSDAI, An Introduction*, Technical Report. Department of Statistics, Oxford University, 1 South Parks Road, Oxford OX1 3 TG.

[6] Gittins, J.C. & Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments, in *Progress in Statistics: European Meeting of Statisticians, Budapest, 1972*, J. Gani, K. Sarkadi & I. Vincze, eds. North-Holland, Amsterdam, pp. 241–266.

(*See also* **Adaptive and Dynamic Methods of Treatment Assignment**; **Operations Research**)

J. GITTINS

Dynamic Population

A dynamic population is a population that gains and loses members, unlike a **fixed population**. A dynamic population is stable or in the steady state if the sizes of all subgroups (e.g. age and gender subgroups) remain constant. **Relative hazards** can be estimated in a dynamic population from **case–control studies** based on **density sampling**. The well-known relationship, disease **prevalence** = disease incidence × average disease duration, which holds when a dynamic population is stationary, requires modification for nonstationary dynamic populations (*see* **Incidence–Prevalence Relationships**).

MITCHELL H. GAIL

Dynamic Programming

The term *dynamic programming* was first coined by Richard Bellman as early as 1952. In 1957, his book on dynamic programming [2] provided the first comprehensive introduction to the mathematical theory of "multistage decision processes" or dynamic programming. The term *programming* describes an iterative mathematical approach to problem-solving, which became feasible with the advent of high-speed digital computers [8]. Dynamic programming is not to be confused, however, with *computer* programming, which is subsequently used to implement the mathematical technique. The term *dynamic* referred to the original application of the multistage decision approach to processes in which a time step played an important role. However, *dynamical systems* can be used to model any process in which the order of operations is crucial. In the alignment of molecular sequences, for example, (see the section on "Application to Molecular Sequence Analysis") each stage of the dynamic programming algorithm corresponds to a step along the sequence.

Dynamic programming is an approach to *optimization*, that is, the process of finding the "best solution" among a number of alternatives. This usually involves selecting from a large number of possible strategies, the strategy that will maximize (or minimize) the value of a function, which has been defined by a mathematical model. The dynamic programming method makes a decision at every stage based upon an assessment of the optimal strategy at that stage only. At the end of the process, the maximum value can be calculated by following the path (or course of action) consisting of these individual decisions.

The problems for which dynamic programming can be used are those that can be broken down into smaller iterative steps such that the overall maximum is reached by making the optimal choice at every step. The sum of these optimal choices is called an *optimal policy*. This was stated mathematically by Bellman in the *Principle of Optimality*:

> *An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

Applying the principle of optimality sequentially from the initial state to each subsequent state implies that the optimal policy will consist of the set of optimal decisions that were made at every stage of the process (The reader familiar with **Markov processes** will note some similarity here. An early generalization to the probabilistic decision-making (Markov) case appears in [6]. The connections between **hidden Markov models** and dynamic programming in molecular sequence analysis (see the section on "Applications") are well-documented in [4] and [3]).

Applications

Initially, dynamic programming was designed to solve problems that arose from calculus, such as time-dependent differential equations. Bellman noted in [2] that although it is possible in principle to find the exact solution to many differential equations using calculus alone, in complex situations involving many variables and multidimensional sets of equations, classical approaches to solving differential equations become infeasible. Referring to "the curse of dimensionality" faced by physicists attempting to solve complex systems of equations, Bellman introduced the idea of an "approximate (solution) in policy space". His more pragmatic approach was feasible due to the rapid development of high-speed computer technology.

A *dynamical system* is essentially a vector of values that describe a set of states. The components of this vector might be, for example, the positions in a queue. The queue might consist of patients who require **scheduling** in a medical clinic. The mathematical model would include factors such as the time required with different practitioners, the nature of the patient's condition, the schedule of each individual practitioner, and so on. The optimal schedule is found by making a sequence of decisions determined by the model.

Dynamic programming had its initial applications in physics, engineering, economics, and resource management. Problems arose, for example, in "the study of optimal inventory or stock control, input–output analysis of a complex of interdependent industries, in the scheduling of patients through a medical clinic, or the servicing of aircraft at an airfield, the study of logistics or investment policies. . .or in **sequential** testing. . ." [2].

But the most important breakthrough occurred in molecular biology where dynamic programming

greatly improved the computational accuracy and efficiency of molecular **sequence** alignment. Indeed, the new era of genetic research (*see* **Bioinformatics**) would have been impossible without the development of reliable computational and statistical methods with which to assemble and search the new **databases**. For publicly available genetic information, see, for example, the website of the National Center for Biotechnology Information (NCBI) at `www.ncbi.nlm.nih.gov`.

Applications to Molecular Sequence Analysis

The most innovative and successful application of dynamic programming to biology has been in the area of molecular sequence analysis, where it has provided an efficient method of finding the *optimal alignment* between two molecular (usually **DNA** or protein) sequences. At each step, the decision is made whether to align (i.e. match) the two next letters or not (in which case a gap is left in one or other of the sequences). At the end of this multistep process, we are able to read off the optimal alignment between the two sequences.

A single strand of DNA is essentially a string of *nucleotides* that contain the information required to build proteins. Protein sequences are strings of *amino acids* rather than nucleotides. The term *residue* refers to either. The residue is represented as a single letter from the four letter DNA alphabet {C,A,T,G} or the 20-letter amino acid alphabet.

Related genes from two different species will have similar DNA sequences. Throughout the evolution of DNA, not only do individual nucleotides change but also whole segments of DNA are inserted into or deleted from the sequence. As a result, if we want to assess the similarity between two possibly related segments of DNA or protein, we will need to count the number of unchanged (or similar) matching residues in some sections while allowing for the possibility of leaving "gaps" or unmatched residues in other sections. This process is called *aligning* the two sequences, and was originally done by eye.

Although no direct mention was made of Bellman's methods in [7], Needleman and Wunsch were the first to apply a dynamic programming method to protein sequence alignment in 1970. The gapped alignment of two sequences corresponds to an optimal path down a scoring **matrix** (see the section on "An Example: The Local Alignment of Two Sequences"). Adjustable *gap penalties* were introduced into the alignment process to discourage the opening of gaps without excluding them completely. The method produces a *global alignment*, which is the optimal alignment between the two sequences in their entirety.

The dynamic programming **algorithm** was modified by Smith and Waterman in 1981 [10]. Noting the existence of numerous noncoding regions or *introns* within the gene, Smith and Waterman altered the algorithm so that the optimal *local alignment* could be found. For local alignments, gaps at either side of the aligned sections were not penalized. The result of this modification was that shorter, biologically relevant regions of similarity (corresponding, for example, to coding regions or *exons*) could be detected efficiently.

It should be noted that in order to make use of the alignment scores obtained using the above methods, further statistical analysis is required; see, for example, [4].

An Example: the Local Alignment of Two Sequences

Say we have two sequences $A = a_1 a_2 \dots a_M$ and $B = b_1 b_2 \dots b_N$, where a_i and b_j are taken from the appropriate alphabet. In practice, when we compare two sequences we want to count not only when the two letters under consideration are the same but also when they are similar. Let $s(a, b)$ be the similarity score between letter a and letter b. (We usually store this information in a square *substitution matrix* $S = [s_{ij}]$ – not to be confused with the scoring matrix. How we create this substitution matrix is complex; see, for example, [3, 5]. In the simplest case, the substitution matrix is the identity matrix.)

We introduce a "gap penalty" g_k, which is some function of the gap-length, k. In the example given in [10], a gap penalty function of $g_k = 1 + \frac{1}{3}k$ was used. This is an example of the commonly used *affine* gap penalty [3]. A penalty of $1\frac{1}{3} = 1 + \frac{1}{3}$ is charged for *opening* a gap and a penalty of $\frac{1}{3}$ is charged for *extending* the gap by one. These values ($1\frac{1}{3}$ and $\frac{1}{3}$ in this case), referred to as *gap opening* and *gap extension* penalties, are set so that a biologically relevant (as opposed to a random) alignment will on average score well.

In Figure 1, we show the Smith–Waterman dynamic programming scoring matrix for the

		C	A	G	C	C	T	C	G	C	T	T	A	G
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
T	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	0.0	**1.0**	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	0.0	1.0	0.0	0.0	**2.0**	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	0.0	1.0	0.7	0.0	1.0	**3.0**	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	0.0	2.0	0.7	0.3	**1.7**	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
T	0.0	0.0	0.7	1.7	0.3	1.3	**2.7**	2.3	1.0	0.7	1.7	2.0	1.0	1.0
T	0.0	0.0	0.3	0.3	1.3	1.0	2.3	**2.3**	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	**3.3**	2.0	1.7	1.3	2.3	2.7
A	0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C	0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G	0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

Figure 1 Smith–Waterman dynamic programming scoring matrix for the DNA sequences CAGCCTCGCTTAG and AATGCCATTGACGG, which are the two example sequences given in [10]. The scoring scheme used is as follows: $s(a, b) = 1$ if $a = b$ and $s(a, b) = -1/3$ if $a \neq b$. The gap penalty function is $g_k = 1 + 1/3\ k$. The first row and first column of D are entered as zeros. The remaining values are then calculated using (1) starting from the top left corner and proceeding to the right and downwards until the matrix is complete. The optimal local alignment is then located by looking for the highest value in the matrix. Once this element is found, the other elements in the alignment are located by tracing back along the path. The traceback ceases when a value of zero is reached as this indicates the end of the optimal local alignment. The elements in the traceback path are shown in bold

DNA sequences AATGCCATTGACGG and CAGCCTCGCTTAG, which are the two example sequences given in [10]. The scoring scheme is as follows: $s(a, b) = 1$ if $a = b$ and $s(a, b) = -\frac{1}{3}$ if $a \neq b$, with the gap penalty function as in the last paragraph.

Starting the $(M + 1) \times (N + 1)$ scoring matrix $D = [d_{ij}]$ with a row of zeros and a column of zeros, the remaining values are then calculated using (1).

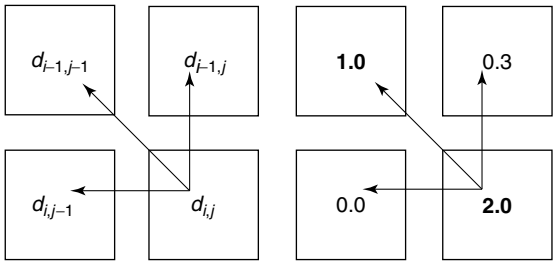

Figure 2 (a) A decision is made at each position i, j. The value of d_{ij} is chosen to be the maximum of three values (or zero) according to (1). The preferred path is recorded (as an arrow facing back towards the previous element) for the traceback procedure. (b) For example, the value of **2.0** in Figure 1 is obtained by taking the maximum value of $1 + 1 = 2$, corresponding to a match between C and C (which scores 1) being added onto the score of 1.0 in the cell diagonally upwards to the left

In Figure 2, the reader can see that at each position three paths are possible, noting that the convention is to draw the arrow backwards because we will want to trace our way back to get the final alignment. As in the global alignment, the optimal path proceeds either diagonally (which corresponds to a match between the ith and jth entry in the two sequences), horizontally or vertically (which corresponds to leaving a gap in the first or second sequence) at each step. At each step, the direction is stored so that we can *traceback* along the path. Starting from the top left corner and proceeding to the right and downwards, each cell is filled in until the matrix is complete.

$$d_{i,j} = \text{maximum} \begin{cases} d_{i-1,j-1} + s(a_i, b_j) \\ d_{i-1,j} - g_k \\ d_{i,j-1} - g_k \\ 0 \end{cases} \tag{1}$$

The optimal local alignment is then located by looking for the highest value in the matrix. Once this element is found, the other elements in the alignment are located using the *traceback* procedure, further details of which can be found in [10]. Essentially, the traceback follows back along the optimal path using the directions that were saved in every cell, although this process can be made more efficient [3]. The traceback ceases when a value of zero is

reached as this indicates the end of the optimal local alignment. The elements in the traceback path are shown in bold.

To read off the alignment in Figure 1, start with the boldface cell uppermost and to the left that contains a **1.0**. This corresponds to a match between G and G. Now go to the next boldface number, **2.0**. Since the path to that number is diagonal the next two letters, C and C, are aligned. As we progress down to **3.0**, the next step is also diagonal and we so match the next two letters, C and C. The next step of the path to **1.7** is vertical. This means a gap is left in the sequence written along the top of the matrix. (A horizontal step leaves a gap in the sequence written along the left of the matrix.) Proceeding in this way, the alignment can be read off the matrix.

The alignment obtained in the example is

G	*C*	*C*	*A*	*T*	*T*	*G*
G	*C*	*C*	-	*T*	*C*	*G*.

which has a Smith–Waterman score of 3.3, the final value in the path.

Computational Efficiency

Dynamic programming became popular due to its efficiency in searching for optimal solutions. In general terms, an exponential or factorial search space can often be reduced to a linear or quadratic time process using dynamic programming techniques.

If we were, for example, to search exhaustively through every possible sequence alignment in the above example of sequence alignment, looking for the highest scoring match according to our scoring scheme, we would require at least some multiple of $(N + M)!$ operations, that is $O((N + M)!)$ (*see* **Orders of Magnitude**). Using the dynamic programming algorithm, this is reduced to $O(MN)$, where M and N are the lengths of the sequences. This is calculated by noting that a finite number of operations must be made at each position of the matrix based upon (1) as illustrated in Figure 1. We refer to this machine complexity as $O(N^2)$ since $M \approx N$. The amount of storage space required is also $O(N^2)$. The algorithm can be made even more efficient by storing only two rows of the matrix at any one time; see, for example, [3].

Quite often in bioinformatics research, we are looking not for one optimal alignment but the highest scoring alignment between a nominated *query* sequence and a large database of sequences such as the human or mouse genome, for example. In this case, quadratic time algorithms are often prohibitively slow, even on the fastest computers. Faster alignment tools such as BLAST [1] and FASTA [9] are based upon a reduction of search space by first looking for high scoring sections. This is then followed by using dynamic programming to find the best alignment among a reduced number of the candidate sequences. Owing to the efficient shortcut made by BLAST, it is currently the most popular tool for searching the genome databases for the best match to a query sequence but it should be emphasized that BLAST itself is *not* a dynamic programming method, but used in conjunction with them.

A reliable Smith–Waterman dynamic programming algorithm, SSEARCH, can be accessed at the FASTA website: `http://alpha10.bioch.virginia.edu/fasta/`.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers E.W. & Lipman, D.J. (1990). Basic local alignment search tool, *Journal of Molecular Biology* **215**, 403–410.

[2] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton.

[3] Durbin, R., Eddy, S.R., Krough, A. & Mitchison, G. (1998). *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

[4] Ewens, W.J. & Grant, G.R. (2000). *Statistical Methods in Bioinformatics: An Introduction*. Springer, New York.

[5] Henikoff, S. & Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices, *Proteins* **17**, 49–61.

[6] Howard, R.A. (1960). *Dynamic Programming and Markov Processes*. John Wiley & Sons, New York.

[7] Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology* **48**, 443–453.

[8] Nemhauser, G.L. (1966). *Introduction to Dynamic Programming*. John Wiley & Sons, New York.

[9] Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison, *Proceedings of the National Academy of Science, USA* **85**, 2444–2448.

[10] Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences, *Journal of Molecular Biology* **147**, 195–197.

Further Reading

Waterman, M.S. (1995). *Introduction to Computational Biology*. Chapman & Hall, London, UK.

(*See also* **Dynamic Allocation Index**; **Operations Research**)

HILARY S. BOOTH

Ecologic Fallacy

The ecologic fallacy is the mistaken assumption that a statistical **association** observed between two ecologic (group-level) variables (*see* **Ecologic Study**) is equal to the association between the corresponding variables at the individual level in the same population. This assumption is often made implicitly or explicitly when using ecologic data to make **inferences** about the biologic (individual-level) effect of an exposure on the **risk** of a disease or other health outcome. Suppose, for example, we observe a positive ecologic association between exposure **prevalence** and the rate of a disease across many regions (groups). The magnitude and direction of the association between exposure status and disease risk within regions (at the individual level) could be different from the ecologic association, even if there is no error in measuring either ecologic variable. Just because the disease rate is higher in regions with a larger exposure prevalence does not mean that exposed individuals are at greater risk of disease than are unexposed individuals. It is possible that the risk is particularly high for unexposed individuals living in regions with a relatively high exposure prevalence. The underlying problem of the ecologic fallacy, therefore, is that each group is not entirely homogeneous with respect to the exposure. If every region were made up entirely of exposed individuals or unexposed individuals, then there would be no ecologic fallacy because information on the joint distribution of exposure and disease within groups would not be missing.

From a statistical perspective, the ecologic fallacy is due to *cross-level bias* in estimating the biologic effect of an exposure on disease risk on the basis of ecologic data. Thus, the fundamental problem of cross-level inference is not an all-or-none phenomenon, but rather a continuum of **systematic error** in effect estimation. In an ecologic analysis involving **simple linear regression**, cross-level bias arises when the disease rate in the unexposed (reference) population is correlated with exposure prevalence across groups or when the difference in rates between exposed and unexposed populations (biologic effect) varies across groups. (*see* **Ecologic Study** for a contemporary interpretation of "ecologic fallacy" and for a discussion of cross-level bias.)

HAL MORGENSTERN

Ecologic Study

An ecologic or aggregate study focuses on the comparison of groups, rather than individuals. The underlying reason for this focus is that individual-level data are missing on the joint distribution of at least two and perhaps all variables within each group; in this sense, an ecologic study is an "incomplete" design [48]. Ecologic studies have been conducted by social scientists for more than a century [18] and have been used extensively by epidemiologists in many research areas. Nevertheless, the distinction between individual-level and group-level (ecologic) studies and the inferential implications are far more complicated and subtle than they first appear. Before 1980, ecologic studies were usually presented in the first part of epidemiology textbooks as simple "descriptive" analyses in which disease rates are stratified by place or time to test hypotheses preliminarily; little attention was given to statistical methods or **inference** (for example [56]). The purpose of this article is to provide a methodologic overview of ecologic studies, which emphasizes study design, statistical methods, and causal inference. Although ecologic studies are easily and inexpensively conducted, the results are often difficult to interpret.

Concepts and Rationale

Before discussing the design and interpretation of ecologic studies, we must first define the concepts of ecologic measurement, analysis, and inference.

Levels of Measurement

The sources of data used in epidemiologic studies typically involve direct observations of individuals (e.g. age and blood pressure); they may also involve observations of groups, organizations, or places (e.g. social disorganization and air pollution). These observations are then organized to measure specific variables in the study population: individual-level variables are properties of individuals, and ecologic variables are properties of groups, organizations, or places. To be more specific, ecologic measures may be classified into three types:

1. *Aggregate measures* are summaries (e.g. **means** or proportions) of observations derived from individuals in each group, e.g. the proportion of smokers and **median** family income.
2. *Environmental measures* are physical characteristics of the place in which members of each group live or work, e.g. air-pollution level and hours of sunlight. Note that each environmental measure has an analog at the individual level, and these individual exposures (or doses) usually vary among members of each group (though they may remain unmeasured).
3. *Global measures* are attributes of groups, organizations, or places for which there is no distinct analog at the individual level (unlike aggregate and environmental measures), e.g. population density, level of social disorganization, the existence of a specific law, or type of health-care system.

Levels of Analysis

The **unit of analysis** is the common level for which the data on all variables are reduced and analyzed. In an *individual-level analysis*, a value for each variable is assigned to every subject in the study. It is possible, even common in **environmental epidemiology**, for one or more predictor variables to be ecologic measures. For example, the average pollution level of each county might be assigned to every subject who is a resident of that county.

In a *completely ecologic analysis*, all variables (exposure, disease, and **covariates**) are ecologic measures so that the unit of analysis is the group, e.g. region (*see* **Geographic Epidemiology**), worksite, school, health-care facility, demographic stratum, or time interval. Thus, within each group, we do not know the joint distribution of any combination of variables at the individual level (e.g. the frequencies of exposed cases, unexposed cases, exposed noncases, and unexposed noncases); all we know is the marginal distribution of each variable, e.g. the proportion exposed and the disease rate (i.e. the T frequencies in Figure 1).

In a *partially ecologic analysis* of three or more variables, we have additional information on certain joint distributions (the M, N, or A/B frequencies in Figure 1); but we still do not know the full joint distribution of all variables within each group (i.e. the ? cells in Figure 1 are missing). For example, in an ecologic study of cancer incidence by county, the joint distribution of age (a covariate) and disease

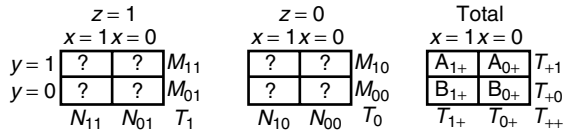

Figure 1 Joint distribution of exposure status ($x = 1$ vs. 0), disease status ($y = 1$ vs. 0), and covariate status ($z = 1$ vs. 0) in each group of a simple ecologic analysis: T frequencies are the only data available in a completely ecologic analysis of all three variables; M frequencies require additional data on the joint distribution of z and y within each group; N frequencies require additional data on the joint distribution of x and z within each group; A and B frequencies require additional data on the joint distribution of x and y within each group; and ? cells are always missing in an ecologic analysis

status within each county (the M frequencies in Figure 1) might be obtained from the **census** and a population tumor registry (*see* **Disease Registers**). From these sources, the investigator would be able to estimate age-specific cancer rates for each county.

Multilevel analysis is a special type of modeling technique that combines analyses conducted at two (or more) levels [7, 27, 100, 101]. For example, an individual-level analysis might be conducted in each group, followed by an ecologic analysis of all groups using the results from the individual-level analyses (*see* **Multilevel Models**). This approach will be described in a later section.

Levels of Inference

The underlying goal of a given epidemiologic study or analysis may be to make *biologic* (or biobehavioral) *inferences* about effects on individual *risks* or to make *ecologic inferences* about effects on group *rates* [62]. The target level of causal inference, however, does not always match the level of analysis. For example, the purpose of an ecologic analysis may be to make a biologic inference about the effect of a specific exposure on disease risk. As discussed later in this article, such *cross-level inferences* are particularly vulnerable to bias.

If the objective of a study is to estimate the *biologic effect* of wearing a motorcycle helmet on the risk of motorcycle-related mortality among motorcycle riders, the target level of causal inference is biologic. On the other hand, if the objective is to estimate the *ecologic effect* of helmet-use laws on the motorcycle-related mortality rate of riders in different states, the target level of causal inference is ecologic. Note that the magnitude of this ecologic effect depends not only on the biologic effect of helmet use, but also on the degree and pattern of compliance with the law in each state. Furthermore, the validity of the ecologic effect estimate depends on our ability to control for differences among states in the joint distribution of **confounders**, including individual-level variables such as age and amount of motorcycle riding.

We might also be interested in estimating the *contextual effect* of an ecologic exposure on individual risk, which is also a form of biologic inference [3, 92]. If the ecologic exposure is an aggregate measure, we would generally want to separate its effect from the effect of its individual-level analog. For example, we might estimate the contextual effect of living in a poor area on the risk of disease, controlling for individual poverty level [45]. Contextual effects can be profound in infectious-disease epidemiology, where the risk of disease depends on the **prevalence** of the disease in others with whom the individual has contact [50, 93] (*see* **Communicable Diseases**).

In evaluating motorcycle-helmet laws in the US, we would probably not expect a contextual *effect* of living in a state that mandates helmet use on the risk of motorcycle-related mortality in riders, controlling for individual helmet use. If a rider's helmet use does not change after the helmet law takes effect, we would not expect his risk of motorcycle-related mortality to change. Nevertheless, we might expect to observe a contextual *association* between the same variables after the law because of differential compliance with the law within states. That is, those riders who comply with the law, but who would not have worn helmets without the law, may be at lower risk than are riders who do not comply with the law. Consequently, the risk of motorcycle-related mortality among riders who do not wear helmets will be higher in states with the helmet law than in states without the law.

Rationale for Ecologic Studies

There are several reasons for the widespread use of ecologic studies in epidemiology, despite frequent cautions about their methodologic limitations:

1. *Low cost and convenience*. Ecologic studies are inexpensive and take little time because various

secondary data sources, each involving different information needed for the analysis, can easily be linked at the aggregate level. For example, data obtained from population registries, **vital statistics** records, large **sample surveys**, and the **census** are often linked at the state, county, or census-tract level.

2. *Measurement limitations of individual-level studies*. In environmental epidemiology and other research areas, we often cannot accurately measure relevant exposures or doses at the individual level for large numbers of subjects – at least not with available time and resources. Thus, the only practical way to measure the exposure may be ecologically [62, 63]. This advantage is especially true when investigating apparent clusters of disease in small areas [94] (*see* **Clustering**). Sometimes individual-level exposures, such as dietary factors, cannot be measured accurately because of substantial within-person variability; yet ecologic measures might accurately reflect group averages [41, 76].
3. *Design limitations of individual-level studies*. Individual-level studies may not be practical for estimating exposure effects if the exposure varies little within the study area. Ecologic studies covering a much wider area, however, might be able to achieve substantial variation in mean exposure across groups (for example [68, 72], and [81]).
4. *Interest in ecologic effects*. As noted above, the stated purpose of a study may be to assess an ecologic effect; i.e. the target level of inference may be ecologic rather than biologic – to understand differences in disease rates among populations [60, 81]. Ecologic effects are particularly relevant when evaluating the impacts of social processes or population interventions such as new programs, policies, or legislation. As discussed later in this article, however, an interest in ecologic effects does not necessarily obviate the need for individual-level data.
5. *Simplicity of analysis and presentation*. In large complex studies conducted at the individual level, it may be conceptually and statistically simpler to perform ecologic analyses and to present ecologic results than to do individual-level analyses. For example, data from large periodic surveys, such as the National Health Interview Survey, are often analyzed ecologically by treating some combination of year, region, and demographic group as the unit of analysis. As discussed later in this article, however, such simplicity of analysis and presentation often conceals methodologic problems.

Study Designs

In an ecologic study design, the planned unit of analysis is the group. Ecologic designs may be classified on two dimensions: the method of exposure measurement and the method of grouping [48, 62]. Regarding the first dimension, an ecologic design is called *exploratory* if there is no specific exposure of interest or the exposure of potential interest is not measured, and it is called *analytic* if the primary exposure variable is measured and included in the analysis. (This use of the term "analytic" is not to be confused with **analytic epidemiology**, which refers to **cohort** and **case–control** studies conducted at the individual level.) In practice, this dimension is a continuum, since most ecologic studies are not conducted to test a single hypothesis. Regarding the second dimension, the groups of an ecologic study may be identified by place (multiple-group design), by time (time-trend design), or by a combination of place and time (mixed design).

Multiple-Group Designs

Exploratory Study. In an exploratory multiple-group study, we compare the rate of disease among many regions during the same period. The purpose is to search for spatial patterns that might suggest an environmental etiology or more specific etiologic hypotheses. For example, the National Cancer Institute (NCI) mapped the age-adjusted cancer mortality rates in the US by county for the period 1950–69 [58]. For oral cancers, they found a striking difference in geographic patterns by sex: among men, the mortality rates were greatest in the urban Northeast; but among women, the rates were greatest in the Southeast. These findings led to the hypothesis that snuff dipping, which is common among rural southern women, is a risk factor for oral cancers [2]. The results of a subsequent case–control study supported this hypothesis [99].

Exploratory ecologic studies may also involve the comparison of rates between migrants and their offspring and residents of their countries of emigration and immigration [41, 56] (*see* **Migrant Studies**).

If the rates differ appreciably between the countries of emigration and immigration, migrant studies often yield results suggesting the influence of certain types of risk factors for the disease under study. For example, if US immigrants from Japan have rates of a disease similar to US whites but much lower than Japanese residents, the difference may be due to environmental or behavioral risk factors operating during adulthood. On the other hand, if US immigrants from Japan and their offspring have rates much lower than US whites but similar to Japanese residents, the difference may be due to genetic risk factors. Such interpretations, however, especially in the first instance, are often limited by differences between countries in the classification and detection of disease or cause of death.

In mapping studies (*see* **Mapping Disease Patterns**), such as the NCI investigation, a simple comparison of rates across regions is often complicated by two statistical problems. First, regions with smaller numbers of observed cases show greater variability in the estimated rate; thus, the most extreme rates tend to be observed for those regions with the fewest cases. Second, nearby regions tend to have more similar rates than do distant regions (i.e. **autocorrelation**) because unmeasured risk factors tend to cluster in space. Statistical methods for dealing with both problems have been developed by fitting an autoregressive spatial model to the data and using **empirical Bayes** techniques to estimate the smoothed rate for each region [12, 17, 61, 64] (*see* **Geographic Epidemiology**). The degree of spatial autocorrelation or clustering can be measured to reflect environmental effects on the rate of disease [96, 97]. The empirical Bayes approach can also be applied to data from analytic multiple-group studies (described below) by including covariates in the model (for example [11], and [15]).

Analytic Study. In an analytic multiple-group study, we assess the ecologic association between the average exposure level or prevalence and the rate of disease among many groups. This is the most common ecologic design; typically, the unit of analysis is a geopolitical region. For example, Hatch & Susser [38] examined the association between background gamma radiation and the incidence of childhood cancers between 1975 and 1985 in the region surrounding a nuclear plant. Average radiation levels for each of 69 tracts in the region were estimated from a 1976 areal survey. The authors found positive associations between radiation level and the incidence of leukemia (an expected finding) as well as solid tumors (an unexpected finding) (*see* **Leukemia Clusters**; **Radiation**).

Data analysis in this type of multiple-group study usually involves fitting a mathematical model to the data. Ordinary **least squares** procedures, however, may be inadequate because the groups typically vary in size and much of the unexplained variability in rates across groups cannot be attributed to sampling error alone. To address these concerns, Pocock et al. [69] proposed a linear model in which the unexplained variation is treated as **random effects**. Model parameters were estimated by an iteratively reweighted least squares procedure. A similar procedure was used by Breslow [6] to fit **loglinear models**. Prentice & Sheppard [73] proposed a linear **relative risk model**, which leads readily to the estimation of rate ratios (assuming the model is properly specified). Prentice & Thomas [77] considered an **exponential** relative risk model, which they argue may be more **parsimonious** than the linear-form model for specifying **covariates**. These methods can be applied to data aggregated by place and/or time (to be discussed below). Use of ecologic modeling to estimate exposure effects (rate ratios and differences) will be described in the next section.

Time-Trend Designs

Exploratory Study. An exploratory time-trend or time-series study involves a comparison of the disease rates over time in one geographically defined population. In addition to providing **graphical displays** of temporal trends, **time-series** data can also be used to **forecast** future rates and trends. This latter application, which is more common in the **social sciences** than in epidemiology, usually involves fitting autoregressive integrated moving average (ARIMA) models to the outcome data [39, 66] (*see* **ARMA and ARIMA Models**). The method of ARIMA modeling can also be extended to evaluate the impact of a population intervention [59], to estimate **associations** betweens two or more time-series variables [9, 66], and to estimate associations in a mixed ecologic design ([85]; see below).

A special type of exploratory time-trend analysis often used by epidemiologists is **age–period–cohort analysis**. This approach typically involves the

collection of retrospective data from a large population over a period of 20 or more years. Through graphical or tabular displays (for example [23] and [25]) or formal modeling techniques (for example [42] and [57]), the objective is to estimate the separate effects of three time-related variables on the rate of disease: age, period (calendar time), and birth cohort (year of birth). By describing the occurrence of disease in this way, the investigator attempts to gain insight about temporal trends, which might lead to new hypotheses.

Lee et al. [54] conducted an age–period–cohort analysis of melanoma mortality among white males in the US between 1951 and 1975. They concluded that the apparent increase in the melanoma mortality rate was due primarily to a cohort effect. That is, persons born in more recent years experienced throughout their lives a higher rate than did persons born earlier. In a subsequent paper, Lee [53] speculated that this cohort effect might reflect increases in sunlight exposure or sunburning during youth, which he hypothesized is a risk factor for melanoma.

From a purely statistical perspective, there is an inherent problem in making inferences from the results of age–period–cohort analyses because of the linear dependency among the three time-related variables [25, 26, 42]. Thus, we cannot allow the value of one variable to change when the values of the other two variables are held constant. As a result of this **identifiability** problem, each data set has alternative interpretations with respect to the combination of age, period, and cohort effects; there is no unique set of effect parameters when all three variables are modeled simultaneously. The only way to decide which interpretation should be accepted is to consider the findings in light of prior knowledge and, possibly, to constrain the model by ignoring one effect.

Analytic Study. In an analytic time-trend study, we assess the ecologic association between change in average exposure level or prevalence and change in disease rate in one geographically defined population. As with exploratory designs, this type of assessment can be done by simple graphical displays or by time-series regression modeling (for example [66]).

In their analytic time-trend study, Darby & Doll [16] examined the associations between average annual absorbed dose of radiation fallout from weapons testing and the incidence rate of childhood leukemia in three European countries between 1945 and 1985. Although the leukemia rate varied over time in each country, they found no convincing evidence that these changes were attributable to changes in fallout radiation.

Causal inference from analytic time-trend studies is often complicated by two problems. First, changes in disease classification and diagnostic criteria can produce distorted trends in the observed rate of disease, which can lead to substantial **bias** in estimating exposure effects. Second, there may be an appreciable induction/**latent period** between first exposure to a risk factor and disease detection. To deal with the latter issue in an ecologic time-trend study, the investigator can lag observations between average exposure and disease rate by a duration assumed to reflect the average induction/latent period of exposure-induced cases. There are two approaches for selecting the lag: (i) an a priori method based on knowledge of the disease; and (ii) empirical methods that maximize the observed association of interest or optimize the fit of the model that includes a lag parameter. Unfortunately, the first method is often problematic because adequate prior knowledge is lacking, and the second method can produce results that are biologically meaningless and very misleading [37].

Mixed Designs

Exploratory Study. The exploratory mixed design combines the basic features of the exploratory multiple-group study and the exploratory time-trend study. Time-series (ARIMA) modeling or age–period–cohort analysis can be used to describe or predict trends in the disease rate for multiple populations. For example, to test Lee's [53] hypothesis that changes in sunlight exposure during youth can explain the observed increase in melanoma mortality in the US, we might conduct an age–period–cohort analysis, stratifying on region according to approximate sunlight exposure (without measuring the exposure). Assuming the amount of sunlight in the regions has not changed differentially over the study period, we might expect the cohort effect described earlier to be stronger for sunnier regions.

Analytic Study. In an analytic mixed design, we assess the association between change in average exposure level or prevalence and change in disease

rate among many groups. Thus, the interpretation of estimated effects is enhanced because two types of comparisons are made simultaneously: change over time within groups and differences among groups. For example, Crawford et al. [14] evaluated the hypothesis that hard drinking water (i.e. water with a high concentration of calcium and magnesium) is a protective risk factor for cardiovascular disease (CVD) mortality. They compared the absolute change in CVD mortality rate between 1948 and 1964 in 83 British towns, by water-hardness change, age, and sex. In all sex–age groups, especially for men, the authors found an inverse association between trends in water hardness and CVD mortality. In middle-aged men, for example, the increase in CVD mortality was less in towns that made their water harder than in towns that made their water softer.

Effect Estimation

A major quantitative objective of most epidemiologic studies is to estimate the effect of one or more exposures on disease occurrence in a well-defined population at risk. A measure of effect in this context is not just any measure of association such as a **correlation** coefficient; rather, it reflects a particular causal parameter, i.e. a counterfactual contrast in disease occurrence [30, 33, 36, 63, 83]. In studies conducted at the individual level, effects are usually estimated by comparing the rate or risk of disease, in the form of a ratio or difference, for exposed and unexposed populations. In multiple-group ecologic studies, however, we cannot estimate effects directly in this way because of the missing information on the joint distribution within groups. Instead, we regress the group-specific disease rates, Y, on the group-specific exposure prevalences, X. (Note that throughout this article uppercase letters will be used to represent ecologic variables and their estimated regression coefficients; lowercase letters will be used to represent individual-level variables and their estimated regression coefficients.)

The most common model form for analyzing ecologic data is the linear model. Ordinary least-squares methods can be used to produce the following prediction equation: $\hat{Y} = B_0 + B_1 X$, where B_0 and B_1 are the estimated intercept and slope. An estimate of the biologic effect of the exposure (at the individual level) can be derived from the regression results [1, 28]. The predicted disease rate ($\hat{Y}_{x=1}$) in a group that is entirely exposed is $B_0 + B_1(1) = B_0 + B_1$, and the predicted rate ($\hat{Y}_{x=0}$) in a group that is entirely unexposed is $B_0 + B_1(0) = B_0$. Therefore, the estimated rate difference is $B_0 + B_1 - B_0 = B_1$, and the estimated rate ratio is $(B_0 + B_1)/B_0 = 1 + B_1/B_0$.

Alternatively, fitting a **loglinear (exponential) model** to the data yields the following prediction equation: $\ln[\hat{Y}] = B_0 + B_1 X$ or $\hat{Y} = \exp[B_0 + B_1 X]$. Applying the same method used above for linear models, the estimated rate ratio is $\hat{Y}_{x=1}/\hat{Y}_{x=0} = \exp[B_1]$.

Note that the ecologic method of effect estimation requires rate predictions be extrapolated to both extreme values of the exposure variable (i.e. $X = 0$ and 1), which are likely to lie well beyond the observed range of the data. It is not surprising, therefore, that different model forms (e.g. loglinear vs. linear) can lead to very different estimates of effect [31]. Fitting a linear model, in fact, may lead to negative, and thus meaningless, estimates of the rate ratio.

As an illustration of rate-ratio estimation in an ecologic study, consider Durkheim's [20] examination of religion and suicide in four groups of Prussian provinces between 1883 and 1890 (see Figure 2). The groups were formed by ranking 13 provinces according to the proportion (X) of the population that was Protestant. Using ordinary least-squares linear regression, we estimate the suicide rate ($\hat{Y}$, per 10^5/year) in each group to be $3.66 + 24.0(X)$. Therefore, the estimated rate ratio, comparing Protestants with other religions, is $1 + (24.0/3.66) = 7.6$. Note in Figure 2 that the fit of the linear model appears excellent ($R^2 = 0.97$). In general, however, ecologic tests of fit can be misleading about the underlying model at the individual level that generated the ecologic data [35].

Confounders and Effect Modifiers

There are two methods used to control for **confounders** in multiple-group ecologic analyses. The first is to treat ecologic measures of the confounders as covariates ($\mathbf{Z}$) in the model, e.g. percent male and percent white in each group. If the individual-level effects of the exposure and covariates are additive (i.e. if the disease rates follow a linear model), then the ecologic regression of Y on X and $\mathbf{Z}$ will also

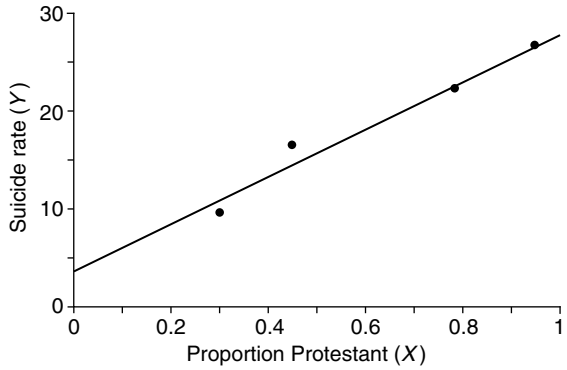

Figure 2 Suicide rate (Y, per 10^5/year) by proportion Protestant (X) for four groups of Prussian provinces, 1883–90. The four observed points (X, Y) are (0.30, 9.56), (0.45, 16.36), (0.785, 22.00), and (0.95, 26.46); the fitted line is based on unweighted least-squares regression. Adapted from Durkheim [20]

be linear with the same coefficients [31, 52] (*see* **Additive Model**). That is, the estimated coefficient for the exposure variable in a linear model can be interpreted as the rate difference adjusted for the covariates, provided the effects are truly additive and there are no other sources of bias. To estimate the adjusted rate ratio for the exposure effect, we must first specify values for all covariates ($\mathbf{Z}$) in the model, because the effects of X and $\mathbf{Z}$ are assumed to be additive – not multiplicative. Thus, the estimated rate ratio, conditional on covariate levels ($\mathbf{Z}$), is the predicted rate in a group that is entirely exposed ($\hat{Y}_{x=1|\mathbf{Z}}$) divided by the predicted rate in a group that is entirely unexposed ($\hat{Y}_{x=0|\mathbf{Z}}$).

Fitting an additive loglinear model to the ecologic data yields an estimate of the adjusted rate ratio that is independent of covariates – i.e. $\hat{Y}_{x=1|\mathbf{Z}}/\hat{Y}_{x=0|\mathbf{Z}} = \exp[B_1]$, where B_1 is the estimated coefficient for the exposure. Thus, the effects of X and $\mathbf{Z}$ are assumed to be multiplicative (*see* **Multiplicative Model**). Unfortunately, this ecologic estimate is a biased estimate of the individual-level rate ratio, even if the effects are multiplicative at the individual level and no other sources of bias are present [31, 79].

The second method used to control for confounders in ecologic analyses is rate standardization for these confounders, followed by regression of the standardized rates as the outcome variable (*see* **Standardization Methods**). Note that this method requires additional data on the joint distribution of the covariate and disease within each group (i.e. the M frequencies in Figure 1). Nevertheless, it cannot be expected to reduce bias unless all predictors in the model (X and $\mathbf{Z}$) are also mutually standardized for the same confounders [31, 34, 82]. Standardization of the exposure prevalences, for example, requires data on the joint distribution of the covariate and exposure within groups (i.e. the N frequencies in Figure 1); unfortunately, this information is not usually available in ecologic studies.

As in individual-level analyses, product terms (e.g. XZ) are often used in ecologic analyses to model **interaction** effects, i.e. to assess **effect modification**. In ecologic analyses, however, the product of X and Z (both group averages) is not, in general, equal to the average product of the exposure, x, and covariate, z, at the individual level within groups. Assuming a linear model, XZ will be equal to the mean xz in each group only if x and z are uncorrelated within groups [31]. Thus, as pointed out in the next section, interaction (nonadditive) effects at the individual level complicate the interpretation of ecologic results.

Methodologic Problems

Despite the many practical advantages of ecologic studies mentioned previously, there are several methodologic problems that may severely limit causal inference, especially biologic inference.

Ecologic Bias

The major limitation of ecologic analysis for making causal inferences is ecologic bias, which is the failure of expected ecologic effect estimates to reflect

the biologic effect at the individual level [22, 28, 34, 35, 62, 79]. In addition to the usual sources of bias that threaten individual-level analyses, the underlying problem of ecologic analyses for estimating biologic effects is heterogeneity of exposure level and covariate levels within groups. As noted earlier, this heterogeneity is not fully captured with ecologic data because of missing information on joint distributions (see Figure 1). Although researchers have long recognized the discrepancy between individual- and group-level associations (for example [24] and [91]), Robinson [80] was the first to describe mathematically how ecologic associations could differ from the corresponding associations at the individual level within groups of the same population. He expressed this relationship in terms of correlation coefficients, which was later extended by Duncan et al. [19] to regression coefficients in a linear model. The phenomenon became widely known as the **ecologic fallacy** [86], and researchers came to recognize that the magnitude of the ecologic bias may be severe in practice [13, 21, 79, 87, 89].

As an illustration of ecologic bias, consider again Durkheim's data on religion and suicide (Figure 2). The estimated rate ratio of 7.6 in the ecologic analysis may not mean that the suicide rate was nearly 8 times greater in Protestants than in non-Protestants. Rather, since none of the regions was entirely Protestant or non-Protestant, it may have been non-Protestants (primarily Catholics) who were committing suicide in predominantly Protestant provinces. It is certainly plausible that members of a religious minority might have been more likely to take their own lives than were members of the majority. The implication of this alternative explanation is that living in a predominantly Protestant area has a contextual effect on suicide risk among non-Protestants, i.e. there is an interaction effect at the individual level between religion and religious composition of one's area of residence.

Interestingly, Durkheim [20] compared the suicide rates (at the individual level) for Protestants, Catholics, and Jews living in Prussia. From his data, we find that the rate was about twice as great in Protestants as in other religious groups. Thus, there appears to be substantial ecologic bias (i.e. comparing rate-ratio estimates of about 2 vs. 8). Durkheim, however, failed to notice this quantitative difference because he did not actually estimate the magnitude of the effect in either analysis.

Greenland & Morgenstern [34] showed that ecologic bias can arise from three sources when using **simple linear regression** to estimate the crude exposure effect: the first may operate in any type of study; the latter two are unique to ecologic studies (i.e. *cross-level bias*) but are defined in terms of individual-level parameters.

1. *Within-group bias*: ecologic bias may result from bias within groups due to **confounding**, **selection** methods, or **misclassification**, even though within-group effects are not estimated. Thus, for example, if there is positive confounding of the crude effect parameter in every group, we would expect the crude ecologic estimate to be biased as well.
2. *Confounding by group*: ecologic bias may result if the background rate of disease in the unexposed population varies across groups, specifically, if there is a nonzero ecologic correlation between mean exposure level and the background rate.
3. *Effect modification by group* (on an additive scale): ecologic bias may also result if the rate difference for the exposure effect at the individual level varies across groups.

Confounding and effect modification by group (the sources of cross-level bias) can arise in three ways: (i) extraneous risk factors (confounders or modifiers) are differentially distributed across groups; (ii) the ecologic exposure variable has a contextual effect on risk separate from the biologic effect of its individual-level analog, e.g. living in a predominantly Protestant area vs. being Protestant (in the suicide example); or (iii) disease risk depends on the prevalence of that disease in other members of the group, which is true of many infectious diseases [50].

To appreciate the sources of cross-level bias, it is helpful to consider simple numerical illustrations involving both individual-level and ecologic analyses with the same population. The hypothetical example in Table 1 involves a dichotomous exposure, x, and three groups. At the individual level, both the rate difference and rate ratio vary somewhat across the groups, but the effect is positive in all groups; the crude and group-standardized rate ratio is 2.0. Fitting a linear model to the ecologic data, however, we find that the slope for the exposure variable, X, is negative and the rate ratio is 0.50, suggesting a protective

Table 1 Number of new cases, person-years (P-Y) of follow-up, and disease rate (Y, per 100 000/year), by group and exposure status (x) (top panel); summary parameters for each group (middle panel); and results of individual-level and ecologic analyses (bottom panel): hypothetical example of ecologic bias due to effect modification by group

Exposure status (x)	Group 1			Group 2			Group 3		
	Cases	P-Y	Rate	Cases	P-Y	Rate	Cases	P-Y	Rate
Exposed ($x = 1$)	20	7 000	286	20	10 000	200	20	13 000	154
Unexposed ($x = 0$)	13	13 000	100	10	10 000	100	7	7 000	100
Total	33	20 000	165	30	20 000	150	27	20 000	135
% exposed ($100X$)			35			50			65
Rate difference (per 10^5/year)			186			100			54
Rate ratio			2.9			2.0			1.5

Individual-level analysis:
Crude rate ratio[a] = 2.0
Adjusted rate ratio (SMR)[b] = 2.0

Ecologic analysis: Linear model
$\hat{Y} = 200 - 100X \quad (R^2 = 1)$
Rate ratio = 0.50

[a]Rate ratio for the total population, unadjusted for group.
[b]Rate ratio standardized for group, using the exposed population as the standard.

effect. The reason for such large ecologic bias is heterogeneity of the rate difference across groups (effect modification by group). In this example, there is no confounding by group because the unexposed rate is the same (100 per 10^5/year) in all three groups.

The example in Table 2 illustrates the conditions for no cross-level bias. First, group is not a modifier of the exposure effect at the individual level because the rate difference (100 per 10^5/year) is uniform across groups (even though the rate ratio varies). Second, group is not a confounder of the exposure effect because there is no ecologic correlation between the percent exposed ($100X$) and the unexposed rate. Thus, the individual-level and ecologic estimates of the rate ratio are the same (1.8) and unbiased, even though the R^2 for the fitted model is very low (0.029).

Unfortunately, the two conditions that produce cross-level bias cannot be checked with ecologic data because those conditions are defined in terms of individual-level associations. This inability to check the validity of ecologic results seriously limits biologic inference. Furthermore, the fit of the ecologic regression model, in general, gives no indication of the presence, direction, or magnitude of ecologic bias. Thus, a model with excellent fit may yield substantial bias, and one model with a better fit than another model may yield more bias. For example, there was substantial bias when fitting a linear model to Durkheim's suicide data in Figure 2, despite an excellent fitting model ($R^2 = 0.97$). Recall that the estimated rate ratio was 7.6, compared with a "true" rate ratio of approximately 2 (see section "Effect Estimation" above). If we fit a loglinear model to the same data, we get $\hat{Y} = \exp[1.974 + 1.418X]$ and $R^2 = 0.91$; therefore, the estimated rate ratio is $\exp[1.418] = 4.1$. Thus, the loglinear model produces less bias even though it has a smaller R^2 than does the linear model. In general, we cannot expect to reduce bias by using better-fitting models in ecologic analysis.

A potential strategy for reducing ecologic bias is to use smaller units in an ecologic study (e.g. counties instead of states) to make the groups more homogeneous with respect to the exposure. On the other hand, this strategy might not be feasible owing to the lack of available data aggregated at the same level, and it can lead to another problem: greater migration between groups (see the section "Other Problems", subsection "Migration Across Groups") [62, 95].

Problems of Confounder Control

As indicated in a previous section, covariates are included in ecologic analyses to control for confounding, but the conditions for a covariate being a confounder are different at the ecologic and individual levels [34, 35]. At the individual level, a risk factor must be associated with the exposure to be a confounder. In a multiple-group ecologic

Table 2 Number of new cases, person-years (P-Y) of follow-up, and disease rate (Y, per 100 000/year), by group and exposure status, x, (top panel); summary parameters for each group (middle panel); and results of individual-level and ecologic analyses (bottom panel): hypothetical example of no ecologic bias

Exposure status (x)	Group 1			Group 2			Group 3		
	Cases	P-Y	Rate	Cases	P-Y	Rate	Cases	P-Y	Rate
Exposed ($x = 1$)	16	8 000	200	30	10 000	300	24	12 000	200
Unexposed ($x = 0$)	12	12 000	100	20	10 000	200	8	8 000	100
Total	28	20 000	140	50	20 000	250	32	20 000	160
% exposed ($100X$)			40			50			60
Rate difference (per 10^5/year)			100			100			100
Rate ratio			2.0			1.5			2.0
Individual-level analysis:						*Ecologic analysis:* Linear model			
Crude rate ratio[a] = 1.8						$\hat{Y} = 133 + 100X$ ($R^2 = 0.029$)			
Adjusted rate ratio (SMR)[b] = 1.8						Rate ratio = 1.8			

[a]Rate ratio for the total population, unadjusted for group.
[b]Rate ratio standardized for group, using the exposed population as the standard.

study, in contrast, a risk factor may produce ecologic bias (i.e. it may be an ecologic confounder) even if it is unassociated with the exposure in every group, especially if the risk factor is ecologically associated with the exposure across groups [31, 34]. Conversely, a risk factor that is a confounder within groups may not produce ecologic bias if it is ecologically unassociated with the exposure across groups.

Control for confounders is more problematic in ecologic analyses than in individual-level analyses [31, 34, 35]. Even when all variables are accurately measured for all groups, adjustment for extraneous risk factors may not reduce the ecologic bias produced by these risk factors. In fact, it is possible for such ecologic adjustment to increase bias [34, 35].

It follows from the principles presented in the previous section that there will be no ecologic bias in a **multiple linear regression** analysis if all the following conditions are met:

1. There is no residual within-group bias in exposure effect in any group due to confounding by unmeasured risk factors, selection methods, or misclassification.
2. There is no ecologic correlation between the mean value of each predictor (exposure and covariate) and the background rate of disease in the joint reference (unexposed) level of all predictors (so that group does not confound the predictor effects).
3. The rate difference for each predictor is uniform across levels of the other predictors within groups (i.e. the effects are additive).
4. The rate difference for each predictor, conditional on other predictors in the model, is uniform across groups (i.e. group does not modify the effect of each predictor on the additive scale at the individual level).

These conditions are sufficient, but not necessary, for the ecologic estimate to be **unbiased**, i.e. there might be little or no bias even if none of these conditions is met. On the other hand, minor deviations from the latter three conditions can produce substantial cross-level bias [31]. Since the sufficient conditions for no cross-level bias cannot be checked with ecologic data alone, the unpredictable and potentially severe nature of such bias makes biologic inference from ecologic analyses particularly problematic.

The conditions for no cross-level bias with covariate adjustment are illustrated in the hypothetical example in Table 3. Both the exposure, x, and covariate, z, are dichotomous variables, and there are three groups. At the individual level, the covariate is not a confounder of the exposure effect because there is no exposure–covariate association within any of the groups. Thus, the crude and adjusted estimates of

Table 3 Number of new cases, person-years (P-Y) of follow-up, and disease rate (Y, per 100 000/year), by group, covariate status, z, and exposure status, x (top panel); summary parameters for each group (middle panel); and results of individual-level and ecologic analyses (bottom panel): hypothetical example of no ecologic bias; covariate is an ecologic confounder but not a within-group confounder

Covariate status (z)	Exposure status (x)	Group 1			Group 2			Group 3		
		Cases	P-Y	Rate	Cases	P-Y	Rate	Cases	P-Y	Rate
1	Exposed	18	3 000	600	24	4 000	600	24	4 000	600
	Unexposed	60	12 000	500	40	8 000	500	30	6 000	500
	Total	78	15 000	520	64	12 000	533	54	10 000	540
0	Exposed	4	2 000	200	8	4 000	200	12	6 000	200
	Unexposed	8	8 000	100	8	8 000	100	9	9 000	100
	Total	12	10 000	120	16	12 000	133	21	15 000	140
Total	Exposed	22	5 000	440	32	8 000	400	36	10 000	360
	Unexposed	68	20 000	340	48	16 000	300	39	15 000	260
	Total	90	25 000	360	80	24 000	333	75	25 000	300
% exposed ($100X$)				20			33			40
% with $z = 1$ ($100Z$)				60			50			40

Individual-level analysis:
Crude rate ratio[a] = 1.3
Adjusted rate ratio (SMR)[b] = 1.3

Ecologic analysis: Linear models
Crude: $\hat{Y} = 420 - 286X$ ($R^2 = 0.94$); rate ratio = 0.32
Adjusted: $\hat{Y} = 100 + 100X + 400Z$ ($R^2 = 1$); rate ratio[c] = 1.3

[a]Rate ratio for the total population, unadjusted for group or the covariate.
[b]Rate ratio standardized for group and the covariate, using the exposed population as the standard.
[c]Setting $Z = 0.50$ (the mean for all three groups).

the rate ratio are nearly the same (1.3). In the ecologic analysis, however, the covariate is a confounder because there is an inverse association between the exposure, X, and the covariate, Z, across groups. Thus, although the crude ecologic estimate of the rate ratio (0.32) is severely biased, the adjusted estimate (1.3) is unbiased. The reasons for no cross-level bias with covariate adjustment are: (i) the rate (100 per 10^5/year) in the joint reference group ($x = z = 0$) does not vary across groups, i.e. condition 2 is met; and (ii) the rate difference (100 per 10^5/year) is uniform within groups and across groups, i.e. conditions 3 and 4 are met.

The example in Table 4 illustrates cross-level bias when the **null hypothesis** is true. At the individual level, the covariate (z) is a strong confounder because it is a predictor of the disease in the unexposed population and it is associated with exposure status, x, within groups. Thus, the crude rate ratio (2.1) is biased. At the ecologic level, however, there is no association between the exposure, X, and the covariate, Z, so that the covariate is not an ecologic confounder. Nevertheless, both the crude and adjusted rate ratios (8.6) are strongly biased because the rate in the joint reference category ($x = z = 0$) is ecologically correlated with both the exposure, X, and the covariate, Z – i.e. condition 2 is not met.

Lack of additivity at the individual level (refer to condition 3) is common in epidemiology, but unmeasured modifiers do not bias results at the individual level if they are unrelated to the exposure [30]. Furthermore, interactions may be handled readily at the individual level by including product terms as predictors in the model (e.g. xz). In ecologic analyses, however, lack of additivity within groups is a source of ecologic bias, and this bias cannot be eliminated by the inclusion of product terms (e.g. XZ) unless the effects are exactly multiplicative and the two variables are uncorrelated within groups [78]. If x and z are correlated within groups, additional data on the x–z associations (the N frequencies in Figure 1) can be used to improve the ecologic estimate of each predictor effect controlling for the other [68, 76].

Another source of ecologic bias is **misspecification** of confounders [35]. Although this problem can

Table 4 Number of new cases, person-years (P-Y) of follow-up, and disease rate (Y, per 100 000/year), by group, covariate status, z, and exposure status, x (top panel); summary parameters for each group (middle panel); and results of individual-level and ecologic analyses (bottom panel): hypothetical example of ecologic bias due to confounding by group; covariate is a within-group confounder but not an ecologic confounder

Covariate status (z)	Exposure status (x)	Group 1			Group 2			Group 3		
		Cases	P-Y	Rate	Cases	P-Y	Rate	Cases	P-Y	Rate
1	Exposed	40	8 000	500	195	13 000	1500	140	14 000	1 000
	Unexposed	60	12 000	500	180	12 000	1500	60	6 000	1 000
	Total	100	20 000	500	375	25 000	1500	200	20 000	1 000
0	Exposed	2	2 000	100	6	2 000	300	12	6 000	200
	Unexposed	28	28 000	100	69	23 000	300	48	24 000	200
	Total	30	30 000	100	75	25 000	300	60	30 000	200
Total	Exposed	42	10 000	420	201	15 000	1340	152	20 000	760
	Unexposed	88	40 000	220	249	35 000	711	108	30 000	360
	Total	130	50 000	260	450	50 000	900	260	50 000	520
% exposed ($100X$)				20			30			40
% with $z = 1$ ($100Z$)				40			50			40

Individual-level analysis:
Crude rate ratio[a] = 2.1
Adjusted rate ratio (SMR)[b] = 1.0

Ecologic analysis: Linear models
Crude: $\hat{Y} = 170 + 1300X$ ($R^2 = 0.16$); rate ratio = 8.6
Adjusted: $\hat{Y} = -2040 + 1300X + 5100Z$ ($R^2 = 1$); rate ratio[c] = 8.6

[a]Rate ratio for the total population, unadjusted for group or the covariate.
[b]Rate ratio standardized for group and the covariate, using the exposed population as the standard; also the common rate ratio within each group.
[c]Setting $Z = 0.433$ (the mean for all three groups).

also arise in individual-level analyses, it is more difficult to avoid in ecologic analyses because the relevant confounder may be the distribution of covariate histories for all individuals within each group. In ecologic studies, therefore, adjustment for covariates derived from available data (e.g. proportion of current smokers) may be inadequate to control confounding. It is preferable, whenever possible, to control for more than a single summary measure of the covariate distribution (e.g. the proportions of the group in each of several smoking categories), provided the outcome rate is not standardized (see section "Effect Estimation" above). In addition, since it is usually necessary to control for several confounders (among which the effects may not be linear and additive), the best approach for reducing ecologic bias is to include covariates for categories of their joint distribution within groups. For example, to control ecologically for race and sex, the investigator might adjust for the proportions of white women, nonwhite men, and nonwhite women (treating white men as the referent), rather than the conventional approach of adjusting for the proportions of men (or women) and whites (or nonwhites).

Within-Group Misclassification

The principles of misclassification bias with which epidemiologists are familiar when interpreting the results of analyses conducted at the individual level do not apply to ecologic analyses. At the individual level, for example, nondifferential misclassification of exposure nearly always leads to **bias toward the null**. In multiple-group ecologic studies, however, this principle does not hold when the exposure variable is an aggregate measure. Brenner et al. [5] have shown that nondifferential misclassification of a dichotomous exposure within groups usually leads to bias away from the null and that the bias may be severe.

As an illustration of this distinct feature of ecologic analysis, consider the two-group example in Table 5, which contrasts analyses with correctly classified and misclassified exposure data at both the

Table 5 Number of new cases, person-years (P-Y) of follow-up, and disease rate (Y, per 100 000/year), by group, type of exposure classification (correct vs. misclassified[a]), and exposure status (top panel); % exposed by group (middle panel); and results of individual-level and ecologic analyses (bottom panel): hypothetical example of ecologic bias away from the null due to nondifferential exposure misclassification within groups

Exposure classification	Exposure status	Group 1			Group 2		
		Cases	P-Y	Rate	Cases	P-Y	Rate
Correctly classified	Exposed ($x = 1$)	50	20 000	250	100	40 000	250
	Unexposed ($x = 0$)	40	80 000	50	30	60 000	50
	Total	90	100 000	90	130	100 000	130
Misclassified[a]	Exposed ($x' = 1$)	49	26 000	188	93	42 000	221
	Unexposed ($x' = 0$)	41	74 000	55	37	58 000	64
	Total	90	100 000	90	130	100 000	130
% exposed – correctly classified ($100X$)				20			40
% exposed – misclassified ($100X'$)				26			42

Individual-level analysis:
Correct: rate ratio[b] = 5.0
Misclassified: rate ratio[c] = 3.4

Ecologic analysis: Linear models
Correct: $\hat{Y} = 50 + 200X$; rate ratio = 5.0
Misclassified: $\hat{Y} = 25 + 250X'$; rate ratio = 11.0

[a]Sensitivity = specificity = 0.9 for both cases and noncases (nondifferential misclassification).
[b]Common rate ratio within each group.
[c]Common rate ratio, using the Mantel-Haenszel method.

individual and ecologic levels. The **sensitivity** and **specificity** of exposure classification are assumed to be 0.9 for both cases and noncases in the population. The correct rate ratio at the individual level is 5.0; with nondifferential exposure misclassification, the observed rate ratio would be 3.4, which is biased toward the null. Although an ecologic analysis of the correctly classified data yields an unbiased estimate of the rate ratio (5.0), an analysis with misclassified data would yield an observed rate ratio of 11.0, which is strongly biased away from the null. To appreciate the direction of the misclassification bias in this ecologic analysis, notice that the difference in the percent exposed ($100X$) between the two groups decreases from $40\% - 20\% = 20\%$ to $42\% - 26\% = 16\%$ when the exposure is misclassified (see Table 5). Thus, the slope in the misclassified analysis increases from 200 to 250 per 10^5/year. In addition, the intercept decreases from 50 to 25 per 10^5/year. Each of these changes causes the observed rate ratio with the misclassified data to increase (away from the null).

It is possible to correct for nondifferential misclassification of a dichotomous exposure or disease in ecologic analyses, based on prior specifications of sensitivity and specificity [[4], Appendix 1; 32]. Suppose, for example, we wish to correct for nondifferential exposure misclassification when using simple linear regression (no covariates) to estimate the exposure effect. The corrected estimator of the rate ratio derived from the model results is $(B_0 + B_1\text{Se})/[B_0 + B_1(1 - \text{Sp})]$, where B_0 and B_1 are the estimated intercept and slope from the misclassified data, Se is the sensitivity of exposure classification, and Sp is the specificity. For example, the corrected rate-ratio estimate for the misclassified exposure data in Table 5 is $(25 + 250 \times 0.9)/(25 + 250 \times 0.1) = 5.0$, which is equal to the estimate based on Greenland & Brenner [32] also derived a corrected estimator for the variance of the estimated rate ratio.

In studies conducted at the individual level, misclassification of a covariate, if nondifferential with respect to both exposure and disease, will usually reduce our ability to control for that confounder [29, 84]. That is, adjustment will not completely eliminate the bias due to the confounder. In ecologic studies, however, nondifferential misclassification of a dichotomous confounder within groups does not affect our ability to control for that confounder, provided there is no cross-level bias [4].

If the outcome and all but one predictor (i.e. the exposure or a covariate) in a given analysis are measured at the individual level, then this partially ecologic analysis may also be regarded as nonecologic with the ecologic variable misclassified. Thus, the resulting bias may be understood in terms of misclassification bias operating at the individual level.

Other Problems

Lack of Adequate Data. Certain types of data, such as medical histories, may not be available in aggregate form; or available data may be too crude, incomplete, or unreliable, such as sales data for measuring behaviors [62, 95]. In addition, secondary sources of data from different administrative areas or from different periods may not be comparable. For example, disease rates may vary across countries because of differences in disease classification or case detection. Furthermore, since many ecologic analyses are based on mortality rather than incidence data, causal inference is further limited because mortality reflects the course of disease as well as its occurrence [48].

Temporal Ambiguity. In a well-designed cohort study of disease incidence, we can usually be confident that disease occurrence did not precede the exposure. In ecologic studies, however, use of incidence data provides no such assurance against this temporal ambiguity [62]. The problem is most troublesome when the disease can influence exposure status in individuals or when the disease rate can influence the mean exposure in groups (through the impact of population interventions designed to change exposure levels in areas with high disease rates).

The problem of temporal ambiguity in ecologic studies (especially time-trend studies) is further complicated by an unknown or variable induction and **latent periods** between exposure and disease detection [37, 95]. The investigator can only attempt to deal with this problem in the analysis by examining associations for which there is a specified lag between observations of average exposure and disease rate. Unfortunately, there may be little prior information about induction and latency on which to base the lag, or appropriate data may not be available to accommodate the desired lag.

Collinearity. Another problem with ecologic analyses is that certain predictors, such as sociodemographic and environmental factors, tend to be more highly correlated with each other than they are at the individual level [13, 87]. The implication of such **collinearities** is that it is very difficult to separate the effects of these variables statistically; analyses yield model coefficients with very large **variances** so that effect estimates may be highly unstable. In general, collinearity is most problematic in multiple-group ecologic analyses involving a small number of large, heterogeneous regions [19, 92].

Migration Across Groups. Migration of individuals into or out of the source population can produce **selection bias** in a study conducted at the individual level because migrants and nonmigrants may differ on both exposure prevalence and disease risk. Although it is clear that migration can also cause ecologic bias [49, 70], little is known about the magnitude of this bias or how it can be reduced in ecologic studies [63].

Ecologic Results and Epidemiologic Controversy

Contemporary epidemiologists take a conservative view of ecologic studies. Knowing that ecologic estimates of effect may be severely biased because of problems discussed in the previous section, epidemiologists tend to trust ecologic findings only if such findings agree with the results of other studies conducted at the individual level, particularly **case–control** and **cohort studies**. Nevertheless, this conservative view ignores the possibility that in certain situations ecologic results might be less biased than are results from case–control and cohort studies, which may, for example, involve appreciable error in measuring exposures (*see* **Measurement Error in Epidemiologic Studies**). Inconsistencies between the results of ecologic and other studies, therefore, can generate controversy about risk-factor effects and the potential for prevention.

One such controversy involving ecologic evidence concerns the possible effect of dietary fat on the risk of breast-cancer. In 1990, Prentice & Sheppard [74] reported results from three types of ecologic studies: (i) an international comparison of breast-cancer incidence during the period 1978–82 in 21 countries (analytic multiple-group design); (ii) a comparison of trends in breast-cancer incidence between

the 1960s and 1978–82 among 10 of the 21 countries in the previous analysis (analytic mixed design); and (iii) a comparison of breast-cancer incidence in US residents of Japanese descent vs. Japanese residents (exploratory multiple-group design). Using data from the 21-country study, the authors found that a 50% reduction in the supply (disappearance) of total fat is associated with a 60% reduction in the incidence of breast cancer among post-menopausal women (ages 55–69). The magnitude of this association did not change appreciably when controlling for gross national product, per capita supply of non-fat calories, and other ecologic variables available to the investigators. The results were similar when the exposure was measured as grams of fat per day and percent of calories from fat [75]. In addition, the association between fat and breast cancer observed in the 21-country analysis was consistent with the results of the two other ecologic analyses.

Despite the large effect and the consistency of these ecologic findings, causal inference is limited for several reasons (for example [40, 43, 75], and [98]): First, because of food wastage, nonhuman consumption, and poor reporting, the per capita *supply* of fat may not be proportional to the per capita *consumption* of fat across countries. Furthermore, the ecologic analyses were conducted within age–sex strata, but per capita fat supply was obtained only for the total population of each country. Second, there may have been systematic differences in breast-cancer detection across countries. Third, Prentice & Sheppard did not have data to control for certain breast-cancer risk factors, such as reproductive history and energy restriction or physical activity early in life. Fourth, the ecologic estimates of effect are susceptible to cross-level bias for other reasons discussed in the previous section.

Although Prentice & Sheppard [74] could not address the above limitations directly, they conducted additional analyses to demonstrate that their ecologic findings were consistent with the results of a pooled analysis of raw data from 12 case–control studies of fat and breast cancer [44]. Using the effect estimate from the 21-country study, they projected the rate ratios (**"relative risks"**) that would be expected in the pooled analysis of case–control studies, assuming random nondifferential error in measuring dietary fat – i.e. assuming the amount of measurement error does not depend on other variables in the analysis. To estimate the amount of measurement error, Prentice & Sheppard used the results of a **validation study** in which food-frequency data (the type used in the case–control studies) were compared with food-record data for the same subjects. They found, for example, that the projected rate ratio for the highest quintile of fat consumption vs. the lowest quintile was 1.46, compared with an observed rate ratio of 1.53 in the pooled analysis of case–control studies [74].

Reactions to the results of Prentice & Sheppard varied considerably. While Hiller & McMichael [40] called their work "a revitalization of ecological studies", Willett & Stampfer [98] maintained that "virtually all the analyses presented by Prentice and Sheppard are irrelevant to etiologic relationships between fat intake and risk of cancer".

One possible problem with the method of Prentice & Sheppard is their assumption that error in measuring fat intake is **nondifferential** with respect to disease status. Since dietary fat is measured after cases are detected in case–control studies, it is possible that cases were more likely to exaggerate their past consumption of fat or that controls were more likely to underestimate it; thus, the rate ratio would be positively biased. To address this concern, Hunter et al. [46] conducted a pooled analysis of raw data from seven cohort studies in which error in measuring fat intake at baseline would be expected to be non-differential with respect to subsequent disease status. The estimated rate ratio for the highest quintile of (energy-adjusted) fat intake versus the lowest quintile was 1.05, and this estimate did not change much when correcting for random nondifferential error in measuring fat intake. Thus, Hunter et al. [46] concluded that there was no evidence of a positive effect of dietary fat on breast-cancer risk. The implication is that effect estimates from the case–control studies were positively biased by differential recall of fat intake and/or selection methods (*see* **Recall Bias**) and that effect estimates from ecologic studies were also positively biased due to the problems mentioned above.

As coherent as these interpretations may appear, they still depend on rather strong assumptions about the error in measuring fat intake at the individual level. It is possible, for example, that the amount of measurement error depends on relevant variables other than disease status. This possibility was evaluated in a recent study by Prentice [71], who used the fat-effect estimate from the 21-country study to project the rate ratios expected in cohort studies under

more realistic assumptions of measurement error. To assess the amount of measurement error, Prentice used the results of another validation study in which food-frequency data were compared with 4-day food-record data at two times, one year apart. In this way, he allowed the amount of error in measuring fat intake to depend on body mass index; and he allowed for measurement errors for the two assessment instruments to be correlated. Under these conditions, he found that the projected rate ratio for the effect of total fat, comparing the highest and lowest quintiles, is approximately 1.1. Prentice concluded, therefore, that the results of Hunter et al.'s [46] pooled analysis of cohort studies is consistent with an effect of fat intake estimated from the international ecologic study.

It is not likely that these recent findings of Prentice will settle disagreements about the possible effect of dietary fat on the risk of breast cancer. Whether ecologic or nonecologic studies provide more accurate estimates of diet effects on cancer incidence remains controversial.

Multilevel Analyses and Designs

Knowing the severe methodologic limitations of ecologic analysis for making biologic inferences, many epidemiologists who report ecologic results argue that there can be no cross-level bias when their primary objective is to estimate an ecologic effect (for example [8, 10] and [88]). For example, we might want to estimate the ecologic effect (effectiveness) of state laws requiring smoke detectors by comparing the fire-related mortality rate in those states with the law vs. other states without the law [62]. Although this is a reasonable objective, the interpretation of observed ecologic effects is complicated by two related issues.

First, disease occurs in individuals; thus, the disease rate in a population is an aggregate, not a global, measure. Consequently, biologic inference may be implicit to the objectives of an ecologic study unless the underlying biologic and contextual effects are already known from previous research. Can smoke detectors placed appropriately in homes reduce the risk of fire-related mortality in those homes by providing an early warning of smoke? Does living in an area where most homes are properly equipped with smoke detectors reduce the risk of fire-related mortality in homes with and without smoke detectors? The first question refers to a possible biologic (biobehavioral) effect; the second question refers to a possible contextual effect. The ecologic effect of smoke-detector laws depends on these biologic and contextual effects as well as other factors, e.g. the level of enforcement, the quality of smoke-detector design and construction, the cost and availability of smoke detectors, and their proper placement, installation, operation, and maintenance. In an ecologic study without additional information, the ecologic effect is completely confounded with related biologic and contextual effects.

The second complicating issue in interpreting observed ecologic effects is the need to control for confounders measured at the individual level. Even if the exposure is a global measure, such as a law, groups are seldom completely homogeneous or comparable with respect to confounders. To make a valid comparison between states with and without smoke-detector laws, for example, we would need to control for differences among states in the joint distribution of extraneous risk factors, such as socioeconomic status of residents, firefighter availability and access, building design and construction (see also the earlier section "Problems of Confounder Control").

Perhaps the best solution to these problems is to incorporate both individual-level and ecologic measures in the same analysis. This approach might include different measures of the same factor; for example, each subject would be characterized by his or her own exposure level as well as the average exposure level for all members of the group to which he or she belongs (aggregate measure). Not only would this approach help to clarify the sources and magnitude of ecologic and cross-level bias, but it would also allow us to separate biologic, contextual, and ecologic effects. It is especially appropriate in social epidemiology, infectious-disease epidemiology, and the evaluation of population interventions.

There are various statistical methods for including both individual-level and ecologic measures in the same analysis; two will be discussed here. The first method, often called *contextual analysis* in the social sciences, is a simple extension of conventional **(generalized linear) modeling** such as multiple linear regression and logistic regression [3, 47]. The model, which is fit to the data at the individual level, includes both individual-level and ecologic predictors. For example, suppose we wanted to estimate the effect of

"herd immunity" on the risk of an infectious disease. The risk, y, of disease might be modeled as a function of the following linear component: $\beta_0 + \beta_1 x + \beta_2 X + \beta_3 x X$, where x is the individual's immunity status and X is the prevalence of immunity in the group to which that individual belongs [93]. Therefore, β_1 represents the biologic effect of individual immunity, β_2 represents the contextual effect of herd immunity, and β_3 represents the interaction effect, which allows the herd-immunity effect to depend on the individual's immune status. The interaction term is needed in this application, since we would expect no herd immunity effect among immune individuals. Note, however, that the interpretation of the interaction effect depends on the form of the model.

An important limitation of contextual analysis is that outcomes of individuals within groups are treated as independent. In practice, however, the outcome of an individual in one group often depends on the outcomes of other individuals in that group. Ignoring such within-group dependence ("clustering") generally results in estimated variances of contextual effects that are biased downward, making **confidence intervals** too narrow. To handle this problem of within-group dependence, we can add **random effects** to the conventional (contextual) model described above; this approach is called *mixed-effects modeling, multilevel modeling*, or *hierarchical regression* [7, 27, 100, 101]. Multilevel modeling is a powerful technique with many applications. It can be used to estimate contextual and ecologic effects and to derive improved (**empirical Bayes**) estimates of biologic effects. It can also be used to determine how much of the difference in outcome rates across groups (ecologic effect) can be explained by differences in the distribution of individual-level risk factors (biologic effects).

As an illustration, suppose that we want to estimate the biologic and contextual effects of income level on a continuous measure of health status (ignoring other potential confounders). Let y_{ij} = the health status of the ith individual living in the jth census tract (group), and x_{ij} = the annual income of the ith individual living in the jth census tract. At the first level of analysis, we model the individual's health status within each census tract as a function of income level – i.e.

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}, \qquad (1)$$

where ε_{ij} is the error term representing the unique (residual) effect associated with the ith individual in the jth census tract. At the second (ecologic) level, we model the census tract-specific intercepts (β_{0j}) and slopes, β_{1j}, from the first level as a function of average census-tract income, X_j – i.e.

$$\beta_{0j} = B_{00} + B_{01} X_j + E_{0j}, \qquad (2)$$

$$\beta_{1j} = B_{10} + B_{11} X_j + E_{1j}, \qquad (3)$$

where E_{0j} and E_{1j} are error terms representing the random effects associated with the jth census tract. The underlying assumption is that the census tract-specific regression parameters are **random samples** from a population of such parameters. By substituting (2) and (3) into (1), we obtain the following combined two-level model:

$$y_{ij} = B_{00} + B_{01} X_j + B_{10} x_{ij} + B_{11} x_{ij} X_j + E_{0j} + E_{1j} x_{ij} + \varepsilon_{ij}, \qquad (4)$$

where B_{10} represents the biologic effect of individual income on health status, B_{01} represents the contextual effect of average census-tract income on health status, and B_{11} represents the interaction effect of individual income and average census-tract income. Using an empirical Bayes procedure, we can also derive an improved estimate of the individual-income effect (β_{1j}) for each census tract. This is accomplished by computing a weighted average of the estimated slope for census tract j in level 1 (equation 1) and the predicted value of this slope using all census tracts in level 2 (equation 3).

Applying multilevel analysis to survey data collected in the UK, Humphreys & Carr-Hill [45] found that living in a poor area (electoral ward) had a detrimental effect on several health outcomes, controlling for socioeconomic status and other individual-level covariates. In a conventional ecologic analysis, the effects of living in a poor area and being poor (low socioeconomic status) would be confounded, and ecologic estimates of effect would be susceptible to cross-level bias.

Multilevel analysis can also be extended to more than two levels. For example, we might want to predict certain health outcomes in nursing-home residents as a function of characteristics of the residents (e.g. age and health status), their physicians (e.g. type of specialty and country of medical training), and the nursing homes (e.g. size and doctor-to-patient ratio).

In this type of analysis, residents are grouped by their physician (who might provide care to many residents in one home) and by their nursing-home affiliation.

The simplest design for generating multilevel analyses is a single survey of a population that is large and diverse enough so that multiple groups (e.g. counties or ethnic groups) can be defined for ecologic measurement and analysis. In addition to environmental and global variables for regions or organizations, ecologic measures are derived by aggregating all subjects in each group. An alternative, more efficient, approach is a *multilevel* or *hybrid design* in which a two-stage sampling scheme is used first to select groups (stage 1), followed by the selection of individuals within groups (stage 2) (for example [45] and [65]) (*see* **Multistage Sampling**). A hybrid design might involve conducting a conventional multiple-group ecologic study by linking different data sources, then obtaining supplemental data from individuals randomly sampled from each group (*see* **Record Linkage**). For example, by estimating the exposure–covariate association in each subsample, this approach can be used to improve the control of confounders in an ecologic analysis [65, 68, 73]. A variation of this hybrid design might involve a case–control study as the second stage. Cases would be identified in the first (ecologic) stage, and controls would be matched to cases on group affiliation and possibly other factors (*see* **Matching**).

Conclusions

There are several practical advantages of ecologic studies, which make them especially appealing for doing various types of epidemiologic research. Despite these advantages, however, ecologic analysis poses major problems of interpretation when making ecologic inferences and especially when making biologic inferences. From a methodologic perspective, it is best to have individual-level data on as many relevant nonglobal measures as possible. Just because the exposure variable is measured ecologically, for example, does not mean that other variables should be as well. The accuracy of effect estimates from ecologic studies can often be improved by obtaining additional data on the within-group associations between covariates, between the exposure and covariates, or between the disease and covariates.

Several epidemiologists have recently called for greater emphasis on understanding differences in health status between populations – a return to a public-health orientation, in contrast to the individual (reductionist) orientation of modern epidemiology [51, 55, 60, 67, 81, 90]. This recommendation represents an important challenge for the future of epidemiology, but it cannot be met simply by conducting ecologic studies; multiple levels of measurement and analysis are needed. Even when the purpose of the study is to estimate ecologic effects, individual-level information is often essential for drawing valid inferences about these effects. Thus, to address the underlying research questions, we typically would want to estimate and control for biologic and contextual effects, preferably using multilevel analysis. In contemporary epidemiology, the "ecologic fallacy" reflects the failure of the investigator to recognize the need for biologic inference and thus for individual-level data. This need arises even when the primary exposure of interest is an ecologic measure and the outcome of interest is the health status of entire populations.

Acknowledgments

Parts of this articles have been reproduced with permission from the *Annual Review of Public Health*, volume 16, pp. 61–81. © 1995 Annual Reviews, Inc.

References

[1] Beral, V., Chilvers, C. & Fraser, P. (1979). On the estimation of relative risk from vital statistical data, *Journal of Epidemiology and Community Health* **33**, 159–162.

[2] Blot, W.J. & Fraumeni, J.F., Jr (1977). Geographic patterns of oral cancer in the United States: etiologic implications, *Journal of Chronic Diseases* **30**, 745–757.

[3] Boyd, L.H., Jr & Iversen, G.R. (1979). *Contextual Analysis: Concepts and Statistical Techniques*. Wadsworth, Belmont.

[4] Brenner, H., Greenland, S. & Savitz, D.A. (1992). The effects of nondifferential confounder misclassification in ecologic studies, *Epidemiology* **3**, 456–459.

[5] Brenner, H., Savitz, D.A., Jöckel, K.-H. & Greenland, S. (1992). Effects of nondifferential exposure misclassification in ecologic studies, *American Journal of Epidemiology* **135**, 85–95.

[6] Breslow, N.E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics* **33**, 38–44.

[7] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Thousand Oaks.

[8] Casper, M., Wing, S., Strogatz, D., Davis, C.E. & Tyroler, H.A. (1992). Antihypertensive treatment and US trends in smoking mortality, 1962 to 1980, *American Journal of Public Health* **82**, 1600–1606.

[9] Catalano, R. & Serxner, S. (1987). Time series designs of potential interest to epidemiologists, *American Journal of Epidemiology* **126**, 724–731.

[10] Centerwall, B.S. (1989). Exposure to television as a risk factor for violence, *American Journal of Epidemiology* **129**, 643–652.

[11] Clayton, D.G., Bernardinelli, L. & Montomoli, C. (1993). Spatial correlation in ecological analysis, *International Journal of Epidemiology* **22**, 1193–1202.

[12] Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671–681.

[13] Connor, M.J. & Gillings, D. (1984). An empiric study of ecological inference, *American Journal of Public Health* **74**, 555–559.

[14] Crawford, M.D., Gardner, M.J. & Morris, J.N. (1971). Changes in water hardness and local death-rates, *Lancet* **2**, 327–329.

[15] Cressie, N. (1993). Regional mapping of incidence rates using spatial Bayesian models, *Medical Care* **31**, Supplement, YS60–YS65.

[16] Darby, S.C. & Doll, R. (1987). Fallout, radiation doses near Dounreay, and childhood leukaemia, *British Medical Journal* **294**, 603–607.

[17] Devine, O.J., Louis, T.A. & Halloran, M.E. (1994). Empirical Bayes methods for stabilizing incidence rates before mapping, *Epidemiology* **5**, 622–630.

[18] Dogan, M. & Rokkan, S. (1969). Introduction, in *Social Ecology*, M. Dogan & S. Rokkan, eds. MIT Press, Cambridge, Mass., pp. 1–15.

[19] Duncan, O.D., Cuzzort, R.P. & Duncan, B. (1961). *Statistical Geography: Problems in Analyzing Areal Data*. Greenwood Press, Westport, pp. 64–67.

[20] Durkheim, E. (1951). *Suicide: A Study in Sociology*. Free Press, New York, pp. 153–154.

[21] Feinleib, M. & Leaverton, P.E. (1984). Ecological fallacies in epidemiology, in *Health Information Systems*, P.E. Leaverton & L. Massé, eds. Praeger, New York, pp. 33–61.

[22] Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data, *American Sociological Review* **43**, 557–572.

[23] Frost, W.H. (1939). The age selection of mortality from tuberculosis in successive decades, *American Journal of Hygiene* **30**, 91–96.

[24] Gehlke, C.E. & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material, *Journal of the American Statistical Association* **29**, Supplement, 169–170.

[25] Glenn, N.D. (1977). *Cohort Analysis*, Series/No. 07-005. Sage, Thousand Oaks.

[26] Goldstein, H. (1979). Age, period and cohort effects – a confounded confusion, *Bias* **6**, 19–24.

[27] Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd Ed. Edward Arnold, London.

[28] Goodman, L.A. (1959). Some alternatives to ecological correlation, *American Journal of Sociology* **64**, 610–625.

[29] Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564–569.

[30] Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analysis, *American Journal of Epidemiology* **125**, 761–768.

[31] Greenland, S. (1992). Divergent biases in ecologic and individual-level studies, *Statistics in Medicine* **11**, 1209–1223.

[32] Greenland, S. & Brenner, H. (1993). Correcting for non-differential misclassification in ecologic analyses, *Applied Statistics* **42**, 117–126.

[33] Greenland, S., Maclure, M., Schlesselman, J.J., Poole, C. & Morgenstern, H. (1991). Standardized regression coefficients: a further critique and review of some alternatives, *Epidemiology* **2**, 387–392.

[34] Greenland, S. & Morgenstern, H. (1989). Ecological bias, confounding, and effect modification, *International Journal of Epidemiology* **18**, 269–274.

[35] Greenland, S. & Robins, J. (1994). Invited commentary: ecologic studies – biases, misconceptions, and counterexamples, *American Journal of Epidemiology* **139**, 747–760.

[36] Greenland, S., Schlesselman, J.J. & Criqui, M.H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect, *American Journal of Epidemiology* **123**, 203–208.

[37] Gruchow, H.W., Rimm, A.A. & Hoffman, R.G. (1983). Alcohol consumption and ischemic heart disease mortality: are time-series correlations meaningful?, *American Journal of Epidemiology* **118**, 641–650.

[38] Hatch, M. & Susser, M. (1990). Background gamma radiation and childhood cancers within ten miles of a US nuclear plant, *International Journal of Epidemiology* **19**, 546–552.

[39] Helfenstein, U. (1991). The use of transfer function models, intervention analysis and related time series methods in epidemiology, *International Journal of Epidemiology* **20**, 808–815.

[40] Hiller, J.E. & McMichael, A.J. (1990). Dietary fat and cancer: a comeback for ecological studies?, *Cancer Causes and Control* **1**, 101–102.

[41] Hiller, J.E. & McMichael, A.J. (1991). Ecological studies, in *Design Concepts in Nutritional Epidemiology*, B.M. Margetts & M. Nelson, eds. Oxford University Press, Oxford, pp. 323–353.

[42] Holford, T.R. (1991). Understanding the effects of age, period, and cohort on incidence and mortality rates, *Annual Review of Public Health* **12**, 425–457.

[43] Howe, G.R. (1990). Dietary fat and cancer, *Cancer Causes and Control* **1**, 99–100.

[44] Howe, G.R., Hirohata, T., Hislop, T.G., Iscovich, J.M., Yuan J.-M., Katsonyanni, K., Lubin, F., Marubini, E.,

Modan, B., Rohan, T., Toniolo, P. & Shunzhang, Y. (1990). Dietary factors and risk of breast cancer: combined analysis of 12 case–control studies, *Journal of the National Cancer Institute* **82**, 561–569.

[45] Humphreys, K. & Carr-Hill, R. (1991). Area variations in health outcomes: artefact or ecology, *International Journal of Epidemiology* **20**, 251–258.

[46] Hunter, D.J., Spiegelman, D., Adami, H.O., Beeson, L., van den Brandt, P.A., Folsom, A.R., Fraser, G.E., Goldbohm, A., Graham, S., Howe, G.R., Kushi, L.H., Marshall, J.R., McDermott, A., Miller, A.B., Speizer, F.E., Wolk, A., Yaun, S.-S. & Willett, W. (1996). Cohort studies of fat intake and the risk of breast cancer – a pooled analysis, *New England Journal of Medicine* **334**, 356–361.

[47] Iversen, G.R. (1991). *Contextual Analysis*. Sage, Thousand Oaks.

[48] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Van Nostrand Reinhold, New York, pp. 77–81, 130–134, 184–280.

[49] Kliewer, E.V. (1992). Influence of migrants on regional variations of stomach and colon cancer mortality in the western United States, *International Journal of Epidemiology* **21**, 442–449.

[50] Koopman, J.S. & Longini, I.M., Jr (1994). The ecological effects of individual exposures and nonlinear disease dynamics in populations, *American Journal of Public Health* **84**, 836–842.

[51] Krieger, N. (1994). Epidemiology and the web of causation: has anyone seen the spider? *American Journal of Epidemiology* **39**, 887–903.

[52] Langbein, L.I. & Lichtman, A.J. (1978). *Ecological Inference*, Series/No. 07–010. Sage, Thousand Oaks.

[53] Lee, J.A.H. (1982). Melanoma and exposure to sunlight, *Epidemiologic Reviews* **4**, 110–136.

[54] Lee, J.A.H., Petersen, G.R., Stevens, R.G. & Vesanen, K. (1979). The influence of age, year of birth, and date on mortality from malignant melanoma in the populations of England and Wales, Canada, and the white population of the United States, *American Journal of Epidemiology* **110**, 734–739.

[55] Link, B.G. & Phelan, J. (1995). Social conditions as fundamental causes of disease, *Journal of Health and Social Behavior* **5**((),extra issue), 80–94.

[56] MacMahon, B. & Pugh, T.F. (1970). *Epidemiology: Principles and Methods*. Little, Brown, & Company, Boston, pp. 137–198, 175–184.

[57] Mason, K.O., Mason, W., Winsborough, H.H. & Poole, W.K. (1973). Some methodological issues in the cohort analysis of archival data, *American Sociological Review* **38**, 242–258.

[58] Mason, T.J., McKay, F.W., Hoover, R., Blot, W.J. & Fraumeni, J.F., Jr. (1975). *Atlas of Cancer Mortality for US Counties: 1950–1969*, DHEW Publ. No. (NIH) 75–780. US Government Printing Office, Washington, pp. 36–37.

[59] McDowall, D., McCleary, R., Meidinger, E.E. & Hay, R.A., Jr (1980). *Interrupted Time Series Analysis*. Sage, Beverly Hills.

[60] McMichael, A.J. (1995). The health of persons, populations, and planets: epidemiology comes full circle, *Epidemiology* **6**, 633–636.

[61] Mollie, A. & Richardson, S. (1991). Empirical Bayes estimation of cancer mortality rates using spatial models, *Statistics in Medicine* **10**, 95–112.

[62] Morgenstern, H. (1982). Uses of ecologic analysis in epidemiologic research, *American Journal of Public Health* **72**, 1336–1344.

[63] Morgenstern, H. & Thomas, D. (1993). Principles of study design in environmental epidemiology, *Environmental Health Perspectives* **101**, Supplement 4, 23–38.

[64] Moulton, L.H., Foxman, B., Wolfe, R.A. & Port, F.K. (1994). Potential pitfalls in interpreting maps of stabilized rates, *Epidemiology* **5**, 297–301.

[65] Navidi, W., Thomas, D., Stram, D. & Peters, J. (1994).. Design and analysis of multilevel analytic studies with applications to a study of air pollution, *Environmental Health Perspectives* **102**, Supplement 8, 25–32.

[66] Ostrom, C.W., Jr (1990). *Time Series Analysis: Regression Techniques*, 2nd Ed. Sage, Newbury Park.

[67] Pearce, N. (1996). Traditional epidemiology, modern epidemiology, and public health, *American Journal of Public Health* **86**, 678–683.

[68] Plummer, M. & Clayton, D. (1996). Estimation of population exposure in ecological studies, *Journal of the Royal Statistical Society, Series B* **58**, 113–126.

[69] Pocock, S.J., Cook, D.G. & Beresford, S.A.A. (1981). Regression of area mortality rates on explanatory variables: what weighting is appropriate?, *Applied Statistics* **30**, 286–295.

[70] Polissar, L. (1980). The effect of migration on comparison of disease rates in geographic studies in the United States, *American Journal of Epidemiology* **111**, 175–182.

[71] Prentice, R.L. (1996). Measurement error and results from analytic epidemiology: dietary fat and breast cancer, *Journal of the National Cancer Institute* **88**, 1738–1747.

[72] Prentice, R.L., Kakar, F., Hursting, S., Sheppard, L., Klein, R. & Kushi, L.H. (1988). Aspects of the rationale for the Women's Health Trial, *Journal of the National Cancer Institute* **80**, 802–814.

[73] Prentice, R.L. & Sheppard, L. (1989). Validity of international, time trend, and migrant studies of dietary factors and disease risk, *Preventive Medicine* **18**, 167–179.

[74] Prentice, R.L. & Sheppard, L. (1990). Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption, *Cancer Causes and Control* **1**, 81–97.

[75] Prentice, R.L. & Sheppard, L. (1991). Dietary fat and cancer: rejoinder and discussion of research strategies, *Cancer Causes and Control* **2**, 53–58.

[76] Prentice, R.L. & Sheppard, L. (1995). Aggregate data studies of disease risk factors, *Biometrika* **82**, 113–125.

[77] Prentice, R.L. & Thomas, D. (1993). Methodologic research needs in environmental epidemiology: data analysis, *Environmental Health Perspectives* **101**, Supplement 4, 39–48.

[78] Richardson, S. & Hémon, D. (1990). Ecological bias and confounding (letter), *International Journal of Epidemiology* **19**, 764–766.

[79] Richardson, S., Stücher, I. & Hémon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations, *International Journal of Epidemiology* **16**, 111–120.

[80] Robinson, W.S. (1950). Ecological correlations and the behavior of individuals, *American Sociological Review* **15**, 351–357.

[81] Rose, G. (1985). Sick individuals and sick populations, *International Journal of Epidemiology* **14**, 32–38.

[82] Rosenbaum, P.R. & Rubin, D.B. (1984). Difficulties with regression analyses of age-adjusted rates, *Biometrics* **40**, 437–443.

[83] Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization, *Annals of Statistics* **6**, 34–58.

[84] Savitz, D.A. & Baron, A.E. (1989). Estimating and correcting for confounder misclassification, *American Journal of Epidemiology* **129**, 1062–1071.

[85] Sayrs, L.W. (1989). *Pooled Time Series Analysis*. Sage, Newbury Park.

[86] Selvin, H.C. (1958). Durkheim's *Suicide* and problems of empirical research, *American Journal of Sociology* **63**, 607–619.

[87] Stavraky, K.M. (1976). The role of ecologic analysis in studies of the etiology of disease: a discussion with reference to large bowel cancer, *Journal of Chronic Diseases* **29**, 435–444.

[88] Stewart, A.W., Kuulasmaa, K. & Beaglehole, R. (1994). Ecological analysis of the association between mortality and major risk factors of cardiovascular disease, *International Journal of Epidemiology* **23**, 505–516.

[89] Stidley, C. & Samet, J.M. (1994). Assessment of ecologic regression in the study of lung cancer and indoor radon, *American Journal of Epidemiology* **139**, 312–322.

[90] Susser, M. & Susser, E. (1996). Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology, *American Journal of Public Health* **86**, 674–677.

[91] Thorndike, E.L. (1939). On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them, *American Journal of Psychology* **52**, 122–124.

[92] Valkonen, T. (1969). Individual and structural effects in ecological research, in *Social Ecology*, M. Dogan & S. Rokkan, eds. MIT Press, Cambridge, Mass., pp. 53–68.

[93] Von Korff, M., Koepsell, T., Curry, S. & Diehr, P. (1992). Multilevel analysis in epidemiologic research on health behaviors and outcomes, *American Journal of Epidemiology* **135**, 1077–1082.

[94] Walter, S.D. (1991). The ecologic method in the study of environmental health. I. Overview of the method, *Environmental Health Perspectives* **94**, 61–65.

[95] Walter, S.D. (1991). The ecologic method in the study of environmental health. II. Methodologic issues and feasibility, *Environmental Health Perspectives* **94**, 67–73.

[96] Walter, S.D. (1992). The analysis of regional patterns in health data: I. Distributional considerations, *American Journal of Epidemiology* **136**, 730–741.

[97] Walter, S.D. (1992). The analysis of regional patterns in health data: II. The power to detect environmental effects, *American Journal of Epidemiology* **136**, 742–759.

[98] Willett, W.C. & Stampfer, M.J. (1990). Dietary fat and cancer: another view, *Cancer Causes and Control* **1**, 103–109.

[99] Winn, D.M., Blot, W.J., Shy, C.M., Pickle, L.W., Toledo, A. & Fraumeni, J.F., Jr (1981). Snuff dipping and oral cancer among women in the southern United States, *New England Journal of Medicine* **304**, 745–749.

[100] Wong, G.Y. & Mason, W.M. (1985). The hierarchical logistic regression model for multilevel analysis, *Journal of the American Statistical Association* **80**, 513–524.

[101] Wong, G.Y. & Mason, W.M. (1991). Contextually specific effects and other generalizations for the hierarchical linear model for comparative analysis, *Journal of the American Statistical Association* **86**, 487–503.

HAL MORGENSTERN

Econometric Methods in Health Services

Introduction

The purpose of this article is to provide a brief summary of prominent econometric applications in modern **health economics**, by providing a broad overview of problems to which econometric methods have been applied, without a detailed exposition of the underlying mathematics, which can be found elsewhere (see [20, 22]).

These methods typically are used to address problems in health, health care delivery, and health care costs from a cross-section of observations at a given time. Methods to analyze data generated from the same units of observation (typically individuals) over time also have been developed in many cases, but those are not presented here in detail. The methods discussed below have evolved to deal with four common, and sometimes interrelated problems with health economic data:

1. The inability to observe fully a variable of interest. For example, in some situations, the only available measure of the health of an individual is whether he or she died, was hospitalized, or incurred some other health outcome that can be measured only with an indicator variable (i.e. a one if the event occurred or a zero otherwise) (*see* **Dummy Variables**). Even when continuous measures are used, such as health expenditure data or measures of functional status from the SF-36, the data can be truncated. The research cannot detect functional status lower than the lowest score on the SF-36, and cannot observe negative values for health expenditures.
2. The high prevalence of **outliers**. This is especially salient in health care cost data. In virtually every health expenditure data set there are some very high users of medical care, resulting in skewed distributions of costs and other measures of utilization. The section on the two-part model and log **transformations** below discusses these issues.
3. The need to compare two groups that were not created through random assignment (*see* **Randomization**). For example, in comparing patients in a fee for service plan versus patients in a managed care organization, we account for the fact that patients self-select into these plans, instead of being randomly assigned. This self-selection may have implications for the underlying health or health behaviors of the two groups that may be difficult to observe. Thus, differences in outcomes between the two groups may reflect selection process rather than differences in the **quality of care** delivered to the two groups. Techniques of **instrumental variables** and sample selection address this type of problem.
4. The nonrandom sampling structure of many health databases. Many databases used by health economists were developed using nonrandom sampling techniques (*see* **Administrative Databases**). For example, some national databases randomly sample households but then survey several people in the household (e.g. National Health Interview Survey). Others randomly select a zip code or other geographic area and then sample persons within those zip codes to participate in the survey (e.g. Medicare Current Beneficiary Survey, Community Tracking Survey).

This chapter presents an overview of econometric methods that have been used in **health services research** and increasingly in biostatistics [14]. The statistical issues that econometric and biostatistical methods are designed to address are often the same, but different terminology is used to describe the same concept [21]. Table 1 below illustrates the common terms used to describe some of these concepts.

Econometric methods are useful in addressing a range of problems in health services and may be an alternative approach to analysis of health, health care, and health insurance problems addressed by biostatistical methods. A more detailed discussion of the methods reviewed here can be found in the chapter by Jones [18] or textbooks and published papers referenced herein.

The Estimation Problem

Suppose we have a dataset on the three health insurance plans, a fee-for-service (FFS) plan and two managed care (e.g. HMO) plans, offered to 1000 people working for 10 different employers. Also suppose that the dataset includes demographic information of

Table 1 Study design and statistical terms

Common term	Synonymous terms
Panel data study	Longitudinal or cohort study
Time series study	Longitudinal study
Cross-section, time series	Longitudinal study
Choice-based sampling	Case–control study
Dependent variable	Outcome, response, endogenous variable
Explanatory variable of interest	Dose, treatment, exposure, intervention, exogenous variable of interest, predictor variable
Explanatory variable	Confounder, independent variable, regressor, exogenous variable, covariate
Interaction	Effect modification
Parameter estimate	Beta, regression coefficient, treatment effect
Partitioned model	Stratified model
Multiple regression	Multivariate regression
Qualitative analysis	Categorical data analysis
Logit (or probit) model	Binomial logistic regression, logistic regression
Ordered logit regression	Ordinal logistic regression, ordinal log-linear regression
Multinomial logit regression	Polytomous logistic regression
Conditional logit regression	Conditional logistic regression, McFadden's logit
Survival analysis	Cox regression, hazard model, duration model failure-time analysis, event history analysis
Omitted variable	Unmeasured covariate, unmeasured confounder, unobservable
Sample selection bias	Censoring, selection bias, incidental truncation
Selection bias	Unmeasured confounding, omitted variable bias, confounding by indication or contraindication
Simultaneous equations	Multiple multivariate regression

Source: Maciejewski, Diehr, Smith and Hebert, 2002.

these 1000 people, key characteristics of each health plan (e.g. total premium, cost-sharing, benefits provided) in each of the 10 employers. In addition, we observe the total health care expenditures that the 1000 people incurred in their health plan over a one-year period and whether or not any employee died sometime during the year.

In the discussion of econometric methods in health services, there are various outcomes we will consider: (1) the dichotomous choice between the two managed care plans that each employee has to make, (2) the trichotomous choice between all three health plans, and (3) the total health care expenditures of each employee. If the outcome of interest is not continuously distributed, then methods to analyze these discrete dependent variables are necessary. An overview of these methods is discussed in the next section. The following section considers methods for analyzing continuous, nonnormally distributed outcomes, which is followed by a discussion of simultaneous equation methods. A section on methods for count data, such as utilization data (*see* **Health Care Utilization Data**), concludes the chapter.

Discrete Dependent Variables

Discrete dependent variable methods are designed for modeling dependent variables (*see* **Response Variable**) that take specific ordinal or nominal values as representations for a continuous "latent" variable (*see* **Path Analysis**) that is unobserved by the researcher [22]. A common outcome in health services research is whether someone dies or not in a study period. The researcher observes the discrete event of death (Y) but does not observe the latent or "true" propensity to die (Y^*). We would like to estimate whether some set of **explanatory variables** (X), such as age, gender, and race, are related to the propensity to die, such that

$$Y^* = XB + u, \tag{1}$$

where B is a vector of parameters of interest (*see* **Estimation**), and u is a **random error** term. We do this by assuming that the discrete outcome of death is observed if the propensity to die exceeds some threshold. So,

$$\begin{aligned} Y &= 0 \qquad \text{if } XB + u \leq 0 \\ Y &= 1 \qquad \text{if } XB + u > 0. \end{aligned} \tag{2}$$

We can estimate B in one of two basic ways. We can treat Y as if it were a continuous variable and estimate a linear probability model (LPM) (*see* **General Linear Model**) in which B is estimated by ordinary **least squares**. Alternatively, we can use **maximum likelihood** techniques to estimate B by expressing the probability that $Y = 1$ as:

$$\begin{aligned} P(Y = 1) &= P(XB + u > 0) \\ &= P(u > -XB) = 1 - F(-XB), \end{aligned} \tag{3}$$

where F is the cumulative distribution of u. This leads to the **likelihood** function:

$$\frac{\Pr(Y = 1)}{\Pr(Y = 0)} = \mathrm{e}^{X\beta}. \tag{4}$$

Assumptions about the cumulative distribution of u provide a means of estimating B. The probit model (*see* **Quantal Response Models**) assumes a standard **normal distribution**, while the logit model (*see* **Logistic Regression**) assumes a standard **logistic distribution**.

The logistic model is used most often in the health literature, but probit and logit models often give very similar results. An alternative is the (LPM, which has the benefit of not relying on untestable distributional assumptions regarding u, but is naturally heteroscedastic (*see* **Scedasticity**) and therefore inefficient. Moreover, predicted values of Y^* from the LPM can be greater than one or less than zero. The functional form of logit and probit models prevent predicted values from exceeding the [0,1] interval, but heteroscedasticity in the logit and probit models may be a problem and could lead to biased estimates of B.

The next types of discrete dependent variable methods are those that take on three or more discrete values, such as the decision by the 1000 employees in our example to enroll in the FFS plan or one of the two HMO plans. To explain this choice, we can include employees' demographic characteristics (e.g. "characteristics of the chooser") and/or information about the premium, cost-sharing, and benefits of each health plan (e.g. "characteristics of the choice"). With these dependent variables, the method to apply is dependent on two factors: whether there is a natural ranking or ordering to the values of the dependent variable and whether the independent variables are characteristics of the outcome or characteristics of the chooser. Qualitative dependent variables that can be ordered, such as levels of health (e.g. excellent, good, fair, poor) are estimated using ordered logit or ordered probit models (*see* **Ordered Categorical Data**). In this case, Y takes on different values depending on the range of the latent variable (See Figure 1).

From Figure 1 above, it is clear that Y^* crosses several thresholds indicated as c1 and c2. Specific values of Y are determined from the relation of specific values of Y^* with the various thresholds as indicated below:

$$\begin{aligned} Y &= 0 \qquad \text{if } Y^* < 0 \\ Y &= 1 \qquad \text{if c1} > Y^* > 0 \\ Y &= 2 \qquad \text{if c2} > Y^* > \text{c1} \\ Y &= 3 \qquad \text{if c2} < Y^*. \end{aligned} \tag{5}$$

If the dependent variable is unordered, such as patient choice of providers (e.g. primary care physician, specialist, chiropractor), and the independent variables (*see* **Explanatory Variables**) are characteristics of the choices, then a conditional logit is appropriate [22]. If the dependent variable is unordered and the independent variables are characteristics of the chooser, then a multinomial logit or multinomial probit is appropriate [13, 18, 22] (*see* **Polytomous Data**).

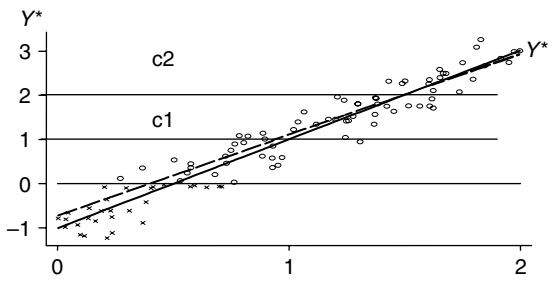

Figure 1 Classification of latent variable Y^* into discrete categories

Dowd et al. [8] describe an example of the multinomial logit model applied to health plan choice. To yield valid interpretations, the multinomial logit and conditional logit models must satisfy the assumption of independence of irrelevant alternatives (IIA), which states that the introduction or withdrawal of a choice will leave the probabilities of remaining choices unchanged. For example, if patients with lower back pain were allowed to choose either a physician or a chiropractor, and twice as many chose the physician, the IIA assumption states that if a new provider were introduced – for example, a physical therapist – the resulting distribution of patients across providers would still have twice as many patients choosing the physicians than the chiropractors.

This assumption often does not hold. If the IIA assumption is not met, nested logit models may be worth considering (*see* **Hierarchical Models**). In these models, choices are grouped into subgroups on the basis of some characteristic that is common to all choices in each subgroup. Feldman et al. [10] estimated a model of health plan choice using a nested logit model. The IIA assumptions held within the nests, but not between the nests. Additionally, recent computer processing advances have made estimation of multinomial and conditional probit models tenable, which do not require the IIA assumption to be met.

Limited Dependent Variables

Limited dependent variables are continuous variables from classical **linear regression** (6) that are **censored** or truncated for some reason [22]. This can be expressed as:

$$Y^* = \alpha + x\beta + e, \qquad \text{where } e \sim N(0, \sigma^2). \quad (6)$$

In classical linear regression, data for Y^* are available for all observations (e.g. individuals), but Y^* is never completely observed in limited dependent variable models. In censored data, Y^* is observed only if Y^* is greater (and/or lower) than some threshold, such as survival time or healthcare costs [22]. Otherwise, Y takes on the value of the threshold(s). For example, survival time can be censored if some people are still alive when the period of observation ends. This is censoring "from above". Health care expenditures often are censored "from below" because we cannot observe expenditures less than zero.

Truncated data are slightly different. For truncated data, we observe Y^* only for people for whom Y^* is greater than (or less than) a given threshold. We have no information on people for whom Y^* does not meet the threshold. For example, a sample of low-income families contains information on income and other variables only for families whose income is below a poverty threshold. No data exist for families above the income threshold.

Figure 2 demonstrates the consequences of censored data. The 'o's represent the true or uncensored data, the 'x's represent the data observed by the researcher, the solid line represents the regression line through the uncensored data, and the dotted line represents the regression line through the censored data points. The mass of observations at the censoring point – zero, in this example – forces the regression line to be more shallow that it should be. In general, the consequences of censored data is to bias the slope coefficients toward zero.

Figure 3 demonstrates the consequences of truncated data. The only data that are observable are those above the truncation point. This means that for

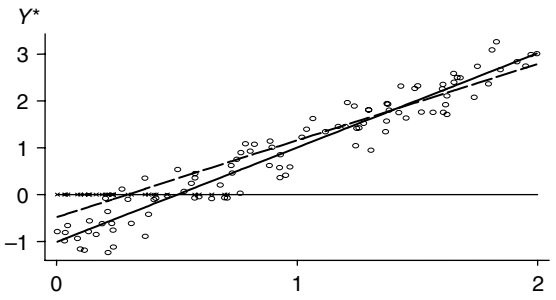

Figure 2 Censored data

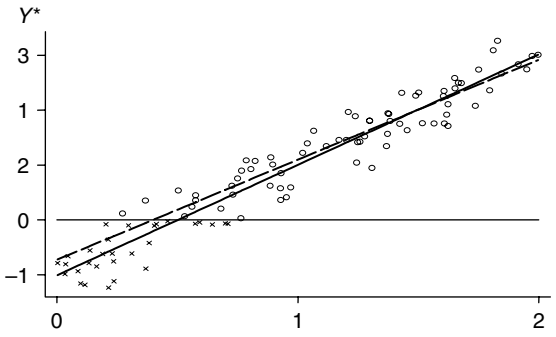

Figure 3 Truncated data

persons with low values of X, only those with unusually high values of Y given X are observed by the researcher. Again, this has the effect of biasing the slope coefficient toward zero.

Truncated data can be estimated using a Tobit model that estimates parameters based only on observations not at the limit(s) [30]. This model has been used to estimate such diverse outcomes as labor supply in depressed people [1], health related quality of life [2, 15], and utilization of health services [16].

Two-part models are a third type of limited dependent variable model that models a continuous variable outcome in which zero is a "true" value not due to censoring or truncation, unlike the Tobit model above. The most common application of two-part models is cost data, where individuals who never interacted with the health care system during the study period have zero costs [7, 23]. Two-part models address particular distributional properties of most cost data, namely, a large preponderance of zeros and tremendous skewness due to several high-cost outliers (*see* **Zero Padding**).

Two-part models estimate the continuous outcome (e.g. cost) in two parts, hence the name. In the first part, the probability of positive costs is estimated using a logit model. The second part estimates the level of positive costs only for those observations with positive costs. The second part may take a variety of distributions and variance structures. Normally distributed costs typically are modeled using untransformed costs, which is appealing for the ease of interpretability and generating predictions. If the positive observations are nonnormally distributed, then logarithmic or square root **transformations** are worth considering. Data that are log-transformed must be retransformed to enable interpretable estimates in the unlogged (or dollar) scale [9, 24]. If heteroscedasticity is a concern, retransformation must take account of the heteroscedasticity [23, 27, 32]. Alternatively, a **generalized linear model** with a **gamma** density and linear or log link may be worth exploring [3].

A fourth type of limited dependent variable model that is closely related to a Tobit model is the sample selection, or incidental truncation model. These models address a special case of truncation in which the process generating unobserved Y is not independent of the process generating specific values of observed Y [18]. For example, nonresponse to survey questions can be modeled using sample selection models. The process generating nonresponse is estimated as well as the process generating the observed values, which is the main equation of interest. These two processes can be modeled in two steps, which is called Heckman's two-step estimator [17]. Alternatively, the two processes can be modeled simultaneously with maximum likelihood, which yields more efficient estimates. The simultaneous model requires that the errors in the selection equation and the errors in the main equation of interest are distributed **bivariate normal**.

Sample selection models have been extended to include bivariate outcomes, such as whether or not someone chooses a zero-deductible plan. For example, if we wanted to use data from our example of the 1000 employers in which employees from 10 employers have a choice of three plans: an FFS plan, an HMO plan with a copayment for an office visit, and an HMO plan without a copayment. If we wanted to estimate whether an employee chooses the FFS plan or one of the HMO plans and then estimate whether the employee chooses the HMO plan with the copayment or the other HMO plan, bivariate probit models with sample selection are appropriate. For more information on these types of models, see papers by Van de Ven and Van Praag [31] and Meng and Schmidt [26].

Simultaneous Equations

For many health services problems, it is important to estimate the response of a variable to another *endogenous* factor determined by the model. For example, the *codetermination* of total health care expenditures and health plan choice must be taken into account in order to identify the effect of choosing an FFS plan on total healthcare expenditures. The steps are twofold; hence, the appropriate econometric procedure is termed two-stage least squares (2SLS). The reason for this two-stage procedure is that an ordinary least squares (OLS) estimate of the coefficient on FFS health plan in an (structural) equation would be biased by the correlation between the disturbance term and the actual health plan choice. A **structural equation** is one that lists all the variables associated with total health care expenditures, even if they are correlated with the disturbance term (e.g. endogenous). A reduced form equation excludes endogenous variables from the equation.

In the first step, one estimates a *reduced form* equation that expresses total health care expenditures

as determined by the type of health plan chosen by an employee and the cost-sharing and benefits of that plan. If health plan choice is endogenous and we want to know the effect of health plan type on total expenditures, we need to include the *predicted* value of total expenditures as a regressor in the second-stage equations.

"Identification" has a specific meaning in econometrics: it implies that the coefficients of the regressors in separate simultaneous structural equations can be uniquely determined (*see* **Identifiability**). The necessary ("order") condition for identification of a given equation requires that the number of variables whose coefficients are constrained to zero or restrictions on the value of some linear combination of coefficients should not be less than the number of endogenous variables in the equation. The rank condition, the second requirement that must be satisfied for identification, states that none of the structural equations in the system can be expressed as a linear combination of one or more of the other equations.

An alternative approach is the **instrumental variables** method that seeks to overcome the same simultaneity bias problem with OLS estimation. The instrumental variables approach also develops a "predictive" equation for the endogenous variable(s), the (causal) influence of which on other endogenous variables in the model is being estimated. As in the first stage of 2SLS, all the "predictors" (regressors) in the equation for the "instrument" (the name given to the endogenous variable whose causal influence on another variable is being estimated) are exogenous and uncorrelated with the error term in the equation of interest. Thus, *in principle,* the predicted value of the instrument will be independent of the disturbance term in the regression of the dependent variable of interest on that instrument.

The major difference between 2SLS and the instrumental variables (IV) method is that in IV applications the investigator does not necessarily develop the model as a simultaneous system of equations in which the reduced form solutions for all the endogenous variables in the system are determined by the same set of exogenous variables. A recent application of IV estimation in health economics was the study by Gaynor & Gertler [12] of physician compensation and physician production within medical groups. In this paper, the investigators acknowledged that physician production levels (e.g. productivity per hour worked and per physician) would likely be influenced by the method of physician compensation *and* would themselves affect the choice of compensation method by the medical group. They then developed instrumental variable estimates for choice of compensation using logistic regression and estimated the effect of compensation method on individual physician production using the instruments.

IV estimation also is often used to deal with **selection bias** issues in health services research applications. Selection bias is a common problem of **cohort studies** where the treatment (or intervention) is not randomly assigned, but rather is chosen by the parties being studied. A paper by McClellan et al. [25] dealt with this issue.

Given the importance of discerning causal effects *(see* **Causation**) in health economics, the need for econometric models that deal with two-way causation, and that eliminate (or at least mitigate) simultaneity bias in estimates of those causal effects, is apparent. Nelson & Startz [28] and Bound, Jaeger and Baker [4] have shown that the results of IV estimation can be highly misleading when the instrument is a poor one. This is often likely to be true in health services applications using administrative or health plan claims databases *(see* **Administrative Databases**) since these sources typically do not include information on key exogenous variables (e.g. household income, occupation, family size, and marital status) that might otherwise act as useful instruments.

Specifically, they show that, in cases where there is "feedback" between the dependent variable and an independent variable of interest (precisely the "two-way causation" that we discuss) *and* the instrument is a "poor one" (in the sense of low R^2 – in the IV estimating equation):

1. The probability limit of the estimated coefficient will approach "a value that is related to the amount of feedback, rather than to the true coefficient" [28]; and
2. even when the true coefficient is zero, the (spurious) level of significance of the IV estimate will increase with the amount of feedback.

They conclude, ironically, that in the very cases in which OLS is a poor estimator because of feedback, a poor instrument will be even worse. The analysis of Nelson and Startz [28] offers, therefore, an important cautionary note in the application of IV estimation to empirical health economics.

In **cross-sectional** analysis of a particular outcome, a different type of simultaneous equation may be useful. Switching regression methods are appropriate when the intercept *and* slope coefficients may be expected to vary from observation to observation. Typically, there are two regimes that may correspond to health states (e.g. healthy, sick) or policy environments. In **longitudinal data**, regimes typically correspond to time periods. Switching regressions are considered a type of simultaneous equation model, because two regressions are run simultaneously (one for each regime). The likelihood function for the switching regression model is listed below.

Gaynor [11] estimated a switching regression model of physician group practices, where demand for their services was constrained in one regime and demand was not constrained in another regime. Bretteville-Jensen [5] estimated a switching regression model of heroin consumption for men and women, where gender was the regime determination. Regimes can be exogenous variables, such as gender or age, or endogenous variables, such as a choice of provider. In the case of endogenous switching that is explicitly determined by the unit of observation, the choice of which regime to be in can be modeled explicitly. O'Donnell [29] models such an endogenous switching process and the regressions in each of the regimes in a model of disability benefits and labor participation by disabled people in the United Kingdom. Readers interested in these methods are directed to [19] for more information.

Count Data Variables

Count data variables are those that take on integer values [6, 18]. As in the two-part model discussed above, count data typically have a preponderance of zeros that leads OLS methods to be biased. Count data in health services commonly refers to events, such as number of physician visits, weeks of care, or length of stay.

Count data have most frequently been modeled as **Poisson processes** with a **Poisson distribution** where the expected number of events equals the variance of events [20]. In count data, the mean number of events may not always equal the variance of events. In cases where the mean number of events is greater than the variance, these data are defined as overdispersed. In the presence of overdispersion, count data can be modeled as **negative binomial distribution** that is a special case of the gamma distribution [13].

Finally, developments in these basic count models have addressed the possibility that count data will have a preponderance of zeros. Zero-inflated Poisson and zero-inflated negative binomial models have been developed to address this scenario [6]. These models allow explicit modeling of the zeros, akin to the Tobit model. Alternatively, there may be cases where there is some process that differentiates observations with values at zero and observations with positive values. In this case, it may be useful to model the zero observations in count data separately from the process determining the positive integer-valued observations in a hurdle model, which is akin to the two-part model [18, 27]. Readers interested in these methods are directed to [6] for more information. Panel data methods (*see* **Panel Study**) also are available and are discussed in [6].

Conclusion

This chapter summarizes some of the econometric methods currently used to address problems of health, health care, and health insurance. The focus was largely on cross-sectional methods using observational data, since most health services applications rely on nonexperimental settings. Application of these methods to longitudinal data is becoming more widespread.

Many statistical problems addressed in biostatistical methods have analogues in econometrics, as illustrated by Table 1. Survival and duration models have relatively recently been applied to econometric problems. Selection models (and other approaches to endogeneity) and count data models used to analyze utilization and costs may represent more unfamiliar ground for biostatisticians [14]. Most of the methods for analysis of cross-sectional models are available in SAS and STATA, while longitudinal models are increasingly available in these statistical packages (*see* **Software, Biostatistical**).

Econometric applications to health services are also expanding into semiparametric and nonparametric versions of the parametric methods presented here. Terminological differences aside, econometric methods represent another set of tools that biostatisticians should add to their toolbox, particularly if they expect to conduct analyses of nonexperimental data.

References

[1] Alexandre, P.K. & French, M.T. (2001). Labor supply of poor residents in Metropolitan Miami, Florida: the role of depression and the co-morbid effects of substance use, *The Journal of Mental Health Policy and Economics* **4**, 161–173.

[2] Austin, P.C. (2002). A comparison of methods for analyzing health-related quality-of-life measures, *Value in Health* **5**, 329–337.

[3] Blough, D.K., Madden, C.W. & Hornbrook, M.C. (1999). Modeling risk using generalized linear models, *Journal of Health Economics* **18**, 153–171.

[4] Bound, J., Jaeger, D.A. & Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association* **90**, 443–450.

[5] Bretteville-Jensen, A.L. (1999). Gender, heroin consumption and economic behaviour, *Health Economics* **8**, 379–389.

[6] Cameron, A.C. & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.

[7] Deb, P. & Trivedi, P.K. (2002). The structure of demand for health care: latent class versus two-part models, *Journal of Health Economics* **21**, 601–625.

[8] Dowd, B., Feldman, R., Cassou, S. & Finch, M. (1991). Health plan choice and the utilization of health care services, *Review of Economics and Statistics* **73**, 85–93.

[9] Duan, N. (1983). Smearing estimate: a nonparametric transformation, *Journal of the American Statistical Association* **78**, 605–610.

[10] Feldman, R., Finch, M., Dowd, B. & Cassou, S. (1989). The demand for employment-based health insurance plans, *Journal of Human Resources* **24**, 115–142.

[11] Gaynor, M. (1989). Competition within the firm: theory plus some evidence from medical group practice, *RAND Journal of Economics* **20**, 59–76.

[12] Gaynor, M. & Gertler, P. (1995). Moral hazard and risk spreading in partnerships, *RAND Journal of Economics* **26**, 591–613.

[13] Greene, W.H. (2000). *Econometric Analysis*, 4th Ed. Prentice Hall, Upper Saddle River.

[14] Greenland, S. (2000). An introduction to instrumental variables for epidemiologists, *International Journal of Epidemiology* **29**, 722–729.

[15] Hahl, J., Hamalainen, H., Sintonen, H., Simell, T., Arinen, S. & Simell, O. (2002). Health-related quality of life in type 1 diabetes without or with symptoms of long-term complications, *Quality of Life Research* **11**, 427–436.

[16] Hamilton, B.H. (1999). HMO selection and medicare costs: Bayesian MCMC estimation of a robust panel data tobit model with survival, *Health Economics* **8**, 403–414.

[17] Heckman, J.J. (1979). Sample selection bias as a specification error, *Econometrica* **47**, 153–161.

[18] Jones, A.M. (2000). Health econometrics, in *Handbook of Health Economics*, A.J. Culyer, & J.P. Newhouse, eds. Elsevier, New York, 265–344.

[19] Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. & Lee, T.-C. (1985). *The Theory and Practice of Econometrics*, 2nd Ed. John Wiley & Sons, New York.

[20] Kennedy, P. (1998). *A Guide to Econometrics*, 4th Ed. MIT Press, Cambridge.

[21] Maciejewski M.L., Diehr, P.D., Smith, M.A. & Hebert, P.L. (2002). Common methodological terms in health services research and their symptoms, *Medical Care* **40**, 477–484.

[22] Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

[23] Manning, W.G. (1998). The logged dependent variable, heteroskedasticity, and the retransformation problem, *Journal of Health Economics* **17**, 283–295.

[24] Manning, W.G. & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics* **20**, 461–494.

[25] McClellan, M., McNeil, B.J. & Newhouse, J.P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables, *JAMA* **272**, 859–866.

[26] Meng, C. & Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model, *International Economic Review* **26**, 71–86.

[27] Mullahy, J. (1997). Instrumental variable estimation of count data models: applications to models of cigarette smoking behavior, *Review of Economics and Statistics* **79**, 586–593.

[28] Nelson, C.R. & Startz, R. (1990). Some further results on the exact small sample properties of the instrumental variables estimator, *Econometrica* **58**, 967–976.

[29] O'Donnell, O. (1993). Income transfers and the labour market participation of disabled individuals in the UK, *Health Economics* **2**, 139–148.

[30] Tobin, J. (1958). Estimation of relationships for limited dependent variables, *Econometrica* **26**, 24–36.

[31] Van de Ven, W.P. & van Praag, B. (1981). The demand for deductibles in private health insurance, *Journal of Econometrics* **17**, 229–252.

[32] Welsh, A.H. & Zhou, X.H. (2002). *Estimating the Retransformed Mean in a Heteroscedastic Two-Part Model*, Working Paper.

MATTHEW L. MACIEJEWSKI, PAUL L. HEBERT, DOUGLAS A. CONRAD & SEAN D. SULLIVAN

Edgeworth Expansion

The expansions named for Edgeworth were developed by P.L. Chebyshev [6] and F.Y. **Edgeworth** [9–11]. Chebyshev was a Russian mathematician whose interest in the expansions derived from his earlier contributions to probability theory and to the theory of orthogonal functions (*see* **Orthogonality**). Edgeworth was a self-taught Irish mathematician, who trained as a linguist and lawyer but made his academic career as an economist in England. His contributions to the expansions that bear his name were motivated by a desire to explore, in a statistical setting, properties of probability distributions.

The Edgeworth expansion of one density, f say, with respect to another, ϕ, may be formally defined in terms of cumulants (*see* **Characteristic Function**). Specifically, writing D for the differential operator $\mathrm{d}/\mathrm{d}x$, it may be shown that

$$f(x) = \exp\left[\sum_{i=1}^{\infty} \left(\kappa_i^f - \kappa_i^\phi\right) \left\{\frac{(-D)^i}{i!}\right\}\right] \phi(x)$$
$$= \phi(x) - \left(\kappa_1^f - \kappa_1^\phi\right) \phi'(x) + \cdots, \qquad (1)$$

where κ_i^f and κ_i^ϕ denote the ith cumulants of the indicated densities, and, among other assumptions, it is supposed that κ_i^f and κ_i^ϕ become "close" sufficiently quickly as i increases. The latter condition is most likely to hold when the density f is converging to ϕ, in some sense (*see* **Convergence in Distribution and in Probability**). The context suggests the **central limit theorem**, where ϕ is the **standard Normal** density.

Indeed, the central limit theorem is the standard setting for studying Edgeworth expansions. In this case, assuming f and ϕ have been standardized for location and scale (so that the corresponding distributions have zero **mean** and unit **variance**), and noting that all cumulants of the standard Normal distribution vanish, we see that (1) simplifies to

$$f(x) = \exp\left\{-\kappa_3^f \left(\frac{D^3}{3!}\right) + \kappa_4^f \left(\frac{D^4}{4!}\right) + \cdots\right\} \phi(x). \qquad (2)$$

It is at this point that a connection is made to Chebyshev's interest in orthogonal functions. The Chebyshev–Hermite polynomials (*see* **Polynomial Approximation**), $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, and so on, sometimes written in the notation He_j rather than H_j, are orthogonal with respect to the Normal N(0, 1/2) density:

$$\int_{-\infty}^{\infty} H_i(x) H_j(x) \exp\left(-\frac{1}{2}x^2\right) \mathrm{d}x = 0, \qquad (3)$$

for $i \neq j$. The right-hand side of (2) may be formally expanded in terms of these functions, giving:

$$f(x) = \left\{1 + \frac{\kappa_3^f}{3!} H_3(x) + \frac{\kappa_4^f}{4!} H_4(x) + \frac{\kappa_5^f}{5!} H_5(x) + \frac{\kappa_6^f + 10\,(\kappa_3^f)^2}{6!} H_6(x) + \cdots\right\} \phi(x). \qquad (4)$$

The special case where f is the density of a sum of n independent and identically distributed **random variables** X_i, corrected for location and scale so that it has zero mean and unit variance, is of particular interest. Assume the summand distribution, after standardizing for location and scale, has ith cumulant λ_i. Then the ith cumulant, κ_i^f, of the distribution of the standardized sum is simply $n^{-(i-2)/2}\lambda_i$, for $i \geq 1$. Making this substitution into (4), we see that the expansion here assumes a particularly simple form, the coefficient of $H_i(x)$ now being of size $n^{-(i-2)/2}$. (The function f on the left-hand side of (4) is now the density of $n^{1/2}(\overline{X} - \mu)/\sigma$, where $\overline{X}$ denotes the mean of independent random variables $X_1, \ldots, X_n$, with $E(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$, and the ith cumulant of $(X_i - \mu)/\sigma$ equal to λ_i.)

In particular, noting that H_i is an odd or even polynomial according as i is odd or even, respectively, we see that we can write

$$f(x) = \left\{1 + n^{-1/2}\, p_1(x) + n^{-1}\, p_2(x) + \cdots\right\} \phi(x), \qquad (5)$$

where $p_i(x)$ is an odd or even polynomial according as i is odd or even. If we integrate (5) term by term, we obtain:

$$F(x) = \int_{-\infty}^{x} f(x)\ \mathrm{d}x = \Phi(x) + n^{-1/2}\, P_1(x)\, \phi(x) + n^{-1}\, P_2(x)\, \phi(x) + \cdots \qquad (6)$$

where P_i is an odd or even polynomial according as i is even or odd, respectively. (Therefore, parities

are reversed in passing from the density to the distribution expansion.)

The properties we have just described are common to a great many distributions, which are asymptotically Normal, not just to the distribution of a standardized sample mean. Details are given by Hall [14, Chapter 2]. Modulo regularity conditions, in particular, **moment** and continuity conditions on the distribution of the data, expansions (5) and (6), and the parity property of the associated polynomials p_i and P_i, apply when f and F are the density and distribution functions, respectively, of a statistic T, which can be expressed as a smooth function of a vector of means.

For example, they apply when $T = n^{1/2}\{\theta(\overline{Y}) - \theta(\nu)\}$ and

$$F(x) = \Pr\Big[n^{1/2}\{\theta(\overline{Y}) - \theta(\mu)\} \le \sigma x\Big], \qquad (7)$$

where $\overline{Y}$ denotes the average value of a sample of n random d-vectors with mean ν, θ is a smooth function of d variables, and $n^{-1}\sigma^2$ denotes the asymptotic variance of the statistic $\theta(\overline{Y})$. Examples include the case where $\theta(\overline{Y})$ is the sample variance, which is a function of two means – the mean of the X_i's and the mean of the X_i^2's. Therefore, in this case, $\overline{Y}$ is the mean of 2-vectors $Y_i = (X_i, X_i^2)$. Likewise, an empirical variance ratio (a function of four sample means), an empirical **correlation** coefficient (a function of five sample means), and so on, may be treated in this way. This approach uses the so-called "smooth function model" for developing valid Edgeworth expansions, introduced by Bhattacharya and Ghosh [2].

The expansion (6), of $F(x)$ defined by (7), continues to hold if σ in (7) is replaced by an estimator of σ, provided that estimator too can be expressed as a smooth function of means of independent random vectors. This is the so-called Studentized, or t, case (*see* **Studentization**). Again the parity properties hold, but the polynomial sequences $p_1, p_2, \ldots$ and $P_1, P_2, \ldots$ are different in the Studentized and non-Studentized settings. However, in either case, P_i is a polynomial of degree $3i - 1$, and the degree of p_i equals $3i - 2$.

Therefore, Edgeworth expansions for sample means, sample variances, variance ratios, the sample correlation coefficient, and related quantities can be derived in a unified way, and all have the same basic properties. There are of course many other statistics that admit Edgeworth expansions. Results there are usually derived using special properties of individual cases. Note, however, that the previously mentioned parity properties of polynomials will not necessarily hold.

The parity properties are important when using Edgeworth expansions to elucidate properties of two-sided **confidence intervals**. Indeed, note from (6) that, provided P_1 is an even polynomial, and P_2 is odd,

$$F(x) - F(-x) = 2\Phi(x) - 1 + 2n^{-1} P_2(x)\,\phi(x) + O\big(n^{-2}\big). \qquad (8)$$

This fortuitous cancellation of terms of size $n^{-1/2}$ can be used to show that, in cases where the parity properties hold, two-sided confidence intervals generally have coverage error no worse than $O(n^{-1})$, even though their one-sided counterparts may do no better than $O(n^{-1/2})$. (Coverage error is the difference between the true coverage of a confidence region and its nominal level, for example, 0.95.)

This type of application of Edgeworth expansions is the predominant one today; they are used mainly to explore properties of procedures that are not themselves based on those expansions. Modern methods for constructing confidence regions and **hypothesis tests** are often highly **computer-intensive**, and mathematical techniques based on Edgeworth expansions allow us to elucidate their behavior; see [14] for discussion of the **bootstrap** in this context.

In the past, however, Edgeworth expansions were sometimes relied upon to produce confidence intervals with good orders of coverage accuracy. Such an approach is not always particularly effective, since Edgeworth series, like those at (5) and (6), generally do not converge as infinite series. They are only "asymptotic" series, in the sense that the order of magnitude (as a function of n) of the remainder in a truncated form of either series equals the order of the first omitted term. Therefore, including more terms in the series can actually make the approximation less accurate, unless sample size, n, is sufficiently large.

Edgeworth expansions and Cornish–Fisher expansions [7, 12] are related, in that the latter are expansions of a **quantile** for a given probability level, while the former are expansions of a probability level for a given quantile. They are, therefore, essentially inverses of one another, and indeed either can be obtained by inverting the other; see [14, Chapter 2].

Gram–Charlier expansions (e.g. [15, p. 17ff] can be viewed as Edgeworth expansions with the terms rearranged in a different order. This operation degrades their convergence properties, however, and partly as a result, Gram–Charlier expansions are seldom studied today.

Surveys of Edgeworth expansions, and significant papers of historical interest in the development of statistical and econometric applications of the expansions, include those of Wallace [18], Bickel [4], Sargan [17], Albers, Bickel and van Zwet [1], Bhattacharya and Rao [3], Bhattacharya and Ghosh [2], Bowman, Beauchamp and Shenton [5], Phillips [16], Cressie [8] and Götze and Hipp [13].

References

[1] Albers, W., Bickel, P.J. & van Zwet, W.R. (1976). Asymptotic expansions for the power of distribution free tests in the one-sample problem, *Annals Statistics* **4**, 108–156. [Gives Edgeworth and related expansions for statistics related to the linear rank test].

[2] Bhattacharya, R.N. & Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion, *Annals of Statistics* **6**, 434–451. [Seminal paper on Edgeworth expansions for a general class of statistics].

[3] Bhattacharya, R.N. & Rao, R.R. (1976). *Normal Approximation and Asymptotic Expansion.*. Wiley, New York. [Detailed theoretical account of Edgeworth expansions for sums of independent vector-valued random variables].

[4] Bickel, P.J. (1974). Edgeworth expansions in non parametric statistics, *Annals of Statistics* **2**, 1–20. [Surveys Edgeworth expansions in nonparametric statistics].

[5] Bowman, K.O., Beauchamp, J.J. & Shenton, L.R. (1977). The distribution of the t-statistic under non-normality, *International Statistical Review* **45**, 233–242. [Classical account of Edgeworth expansions for Studentised mean, with references to historical literature].

[6] Chebyshev, P.L. (1890). Sur deux théorèmes relatifs aux probabilités, *Acta Mathematics* **14**, 305–315.

[7] Cornish, E.A. & Fisher, R.A. (1937). Moments and cumulants in the specification of distributions, *International Statistical Review* **5**, 307–322.

[8] Cressie, N. (1980). Relaxing assumptions in the one-sample t-test, *Australian Journal of Statistics* **22**, 143–153. [Survey of behaviour of Student's t statistic under non-Normality, with qualitative summary of the statistic's properties in this context].

[9] Edgeworth, F.Y. (1896). The asymmetrical probability curve, *Philos Magazine, 5th Series* **41**, 90–99.

[10] Edgeworth, F.Y. (1905). The law of error, *Proceedings of the Cambridge Philosophical Society* **20**, 26–65.

[11] Edgeworth, F.Y. (1907). On the representation of a stationary frequency by a series, *Journal of the Royal Statistical Society, Series A* **70**, 102–106.

[12] Fisher, R.A. & Cornish, E.A. (1960). The percentile points of distributions having known cumulants, *Technometrics* **2**, 209–226.

[13] Götze, F. & Hipp, C. (1983). Asymptotic expansions for sums of weakly dependent random vectors, *Z. Wahrscheinlichkeit Verw. Gebiete* **64**, 211–239. [Often-used theoretical properties of Edgeworth expansions for dependent variables].

[14] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. [Introduces Edgeworth expansion methods and uses them to explore properties of bootstrap confidence intervals and hypothesis tests].

[15] Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics. Continuous Univariate Distributions*, Vol. 1. Houghton Mifflin, Boston.

[16] Phillips, P.C.B. (1977). A general theorem in the theory of asymptotic expansions as approximations to the finite sample distributions of econometric estimators, *Econometrica* **45**, 1517–1534. [Account of Edgeworth expansions in the context of econometrics].

[17] Sargan, J.D. (1975). Gram-Charlier approximations applied to t ratios of k-class estimators, *Econometrica* **43**, 327–346. [Account of Gram-Charlier expansions for t ratios].

[18] Wallace, D.L. (1958). Asymptotic approximations to distributions, *Annals of Mathematical Statistics* **29**, 635–654. [Survey of asymptotic approximations to distributions, with some new results].

(*See also* **Large-sample Theory**; **Sampling Distributions**)

PETER HALL

Edgeworth, Francis Ysidro

Born: February 8, 1845, in Edgeworthstown, Ireland.
Died: February 13, 1926, in London, England.

Edgeworth, a rather reclusive figure, contributed importantly to the development of statistical theory at the end of the nineteenth century. After an Oxford degree in classics, he trained as a barrister while self-studying advanced mathematics. From 1880, he held university appointments in logic, and in economics and statistics, before accepting a chair in political economy at Oxford in 1891, which he occupied until retirement in 1922.

He published extensively in mathematical economics. His interest in probability and statistics, starting in the 1880s, was stimulated by contact with **Francis Galton** and later with **Karl Pearson**. His approach was usually, although not consistently, **Bayesian**, and included early advocacy of **maximum likelihood**. Many publications deal with the **normal distribution**, including the **bivariate** and **multivariate** forms, and with associated problems of **correlation** and **regression**. His work on skew distributions (*see* **Skewness**), involving the "Edgeworth expansion", brought him into conflict with Karl Pearson, whose own system of curves (*see* **Pearson Distributions**) had been criticized by Edgeworth.

For fuller accounts, see [1] and [2], and further references listed on p. 373 of [2].

References

[1] Stigler, S.M. (1978). Francis Ysidro Edgeworth, statistician (with discussion), *Journal of the Royal Statistical Society, Series A* **141**, 287–322.

[2] Stigler, S.M. (1986). *The History of Statistics: the Measurement of Uncertainty before 1900*, Belknap Press, Cambridge, Mass.

Effect Modification

The term *effect modification* is due to Miettinen [5], and is highly related to the concept of **interaction**. When we analyze the **association** of an exposure with disease incidence, an effect modifier is a variable over which the effect of exposure on disease risk varies. For example, we might use the **relative risk** to describe the association of cigarette smoking to lung cancer risk. If the relative risk associated with cigarette smoking is statistically different among asbestos-exposed subjects from that among subjects with no asbestos exposure, then we say that asbestos exposure modifies the effect of cigarette smoking on lung cancer risk, and we call asbestos exposure an effect modifier.

We study effect modification for a variety of reasons. We may be interested in effect modification for its public health implications: if the effects of certain modifiable risk factors for breast cancer are confined to a subgroup of women, for example, efforts to modify these risk factors or increase **screening** might be targeted only at subpopulations where the intervention will prevent the most women from developing advanced disease.

Or we may think that the joint biologic effect of two exposures may be either to inhibit or to enhance each others' individual effects, and we might expect this to cause effect modification. Siemiatycki & Thomas [7] and Thompson [8] have argued, however, that in most cases it is foolish to infer much about biological interaction from the pattern of disease rates in an epidemiologic study because so little is known about biologic mechanisms. The one exception to this is the case where an exposure decreases risk for one value of the effect modifier and increases risk for another value of the effect modifier. Thompson [8] has called this "crossover", and has argued that it is the one case where some form of biologic interaction may be inferred (*see* **Synergy of Exposure Effects**).

Relationship to Interaction

Effect modification is highly related to statistical interaction in **regression** models. In relative risk regression models, where regression coefficients for main effect exposure variables have the interpretation of log relative risks, a significant interaction between exposure and a second variable means that the second variable is an effect modifier (*see* **Relative Risk Modeling**). This is true because the model with interaction says the relative risk associated with exposure will be different depending on the value of the effect modifier.

However, the relationship between statistical interaction and effect modification depends on the correspondence between what measure we choose for the effect of exposure on disease risk and the form of the regression model we use to assess interaction. To see this, we examine three possible regression models and their corresponding measures of effect.

Multiplicative Models and Relative Risks

Logistic regression, Poisson regression with a log link function (*see* **Generalized Linear Model**), and multiplicative **Cox regression** are all examples of **multiplicative models** for which the relative risk is the implicit measure of effect [1, 2]. For example, if we applied the logistic regression model to data from a **case–control study** of smoking and asbestos exposure as risk factors for mesothelioma, then for

$$X_{\mathrm{A}} = \begin{cases} 0, & \text{no occupational asbestos exposure,} \\ 1, & \text{occupational asbestos exposure,} \end{cases}$$

$$X_{\mathrm{S}} = \begin{cases} 0, & \text{never smoked,} \\ 1, & \text{ever smoked,} \end{cases}$$

and $p =$ the probability of being a case in the case–control sample, a simple logistic model without interaction is

$$\text{logit } p = \ln \frac{p}{1-p} = \beta_0 + \beta_{\mathrm{A}} X_{\mathrm{A}} + \beta_{\mathrm{S}} X_{\mathrm{S}}.$$

In a **cohort study**, a similar model would hold with $p =$ the population probability of developing the disease during the study period.

This model implies that the **odds ratio** associated with ever having smoked, $e^{\beta_{\mathrm{S}}}$, is the same whether or not an individual has had occupational exposure to asbestos. To allow the odds ratios associated with smoking to differ according to whether or not the subject had been exposed to asbestos, we add the interaction term $\beta_{\mathrm{AS}} X_{\mathrm{A}} X_{\mathrm{S}}$ to the model. Thus, if we use the logistic regression model, the presence or absence of interaction corresponds to presence or absence of effect modification, where the measure of effect we use is the odds ratio, or the relative

risk, which it approximates. This correspondence also holds for the following Poisson regression model:

$$\ln \lambda = \beta_0 + \beta_A X_A + \beta_S X_S,$$

where λ is the disease incidence rate for a time interval/covariate combination (*see* **Time-dependent Covariate**), and for the multiplicative Cox regression model

$$\ln \lambda(t) = \ln \lambda_0(t) + \beta_0 + \beta_A X_A + \beta_S X_S,$$

where $\lambda(t)$ is the hazard function. If we are interested in whether the relative risk difference (see below) associated with smoking is different among asbestos-exposed individuals from what it is among non-asbestos-exposed individuals, the presence or absence of an interaction term in a multiplicative model will not give us this information. Instead, we need to look for interaction in an additive relative risk regression model.

Additive Relative Risk Models and Relative Risk Differences

Most investigators would agree that when they are considering public health implications of some exposure, additive measures of the effect of exposure on risk are more useful than multiplicative measures for identifying subgroups where interventions should be targeted. In addition, Rothman has argued that the additive scale is better for assessing whether there is biological interaction, and that **additive models** should be used instead of the multiplicative models when assessing whether there is effect modification [6]. However, this has been disputed by Siemiatycki & Thomas [7] and Thompson [8] among others, who show that, depending on the biologic model for how exposure affects disease risk, biologic interaction may or may not manifest itself as a statistical interaction on either the additive or multiplicative scales.

In a case–control study, the additive measure of effect that can be estimated is the additive relative risk. For a case–control study of lung cancer like the one described above, the additive relative risk is defined as follows. Let $p(0, 0)$ be the probability of being a case among those exposed to neither smoking nor asbestos, and $p(X_A, X_S)$ the probability of being a case among those with asbestos and smoking exposure given by X_A and X_S. Then the relative risk difference comparing combinations of exposure (X_A, X_S) and (X'_A, X'_S) is

$$RRD = \frac{p(X_A, X_S) - p(X'_A, X'_S)}{p(0, 0)}.$$

If this is the measure of the effect of exposure (X_A, X_S) compared with exposure (X'_A, X'_S), then we say that asbestos exposure modifies the effect of smoking if the relative risk difference associated with smoking is different among those occupationally exposed to asbestos from what it is among those without occupational asbestos exposure. Whether or not this type of effect modification exists depends on whether or not there is an interaction term in the additive relative risk regression model:

$$\frac{p}{1-p} = e^{\beta_0}(1 + \beta_A X_A + \beta_S X_S).$$

Similar correspondences hold for the additive relative risk versions of the Poisson regression model

$$\lambda = e^{\beta_0}(1 + \beta_A X_A + \beta_S X_S),$$

and the Cox regression model

$$\lambda(t) = \lambda_0(t)(1 + \beta_A X_A + \beta_S X_S).$$

See Breslow & Day [2] for more details.

Additive Risk Models and Risk Differences

The relative risk difference measures the difference in risk of disease for different exposure combinations relative to the disease risk in a baseline group where all the covariates have the value zero. Using data from a cohort study, it is also possible to estimate absolute differences in the risk of disease or the disease **incidence rate**. Letting $p(x_A, x_S)$ denote the disease risk or probability of developing disease during the study period for those with asbestos exposure given by X_A and smoking exposure given by X_S, then the risk difference comparing combinations of exposure (X_A, X_S) and (X'_A, X'_S) is

$$RD = p(X_A, X_S) - p(X'_A, X'_S).$$

If the risk difference is the measure of the effect of exposure (X_A, X_S) compared with (X'_A, X_S), then we say asbestos exposure modifies the effect of smoking if the risk difference associated with smoking is different among those occupationally exposed to

Table 1 History of having given birth, family history of breast cancer, and breast cancer risk, from Colditz et al. [3]

	No family history				Family history			
	Cases	Person-years	Relative risk	95% CI	Cases	Person-years	Relative risk	95% CI
Never given birth	150	69 666	1.0[a]	–	16	5 816	1.0[a]	–
Ever given birth	1788	994 628	0.83	(0.71, 0.99)	5816	78 559	1.4	(0.83, 2.3)

[a]Reference group

asbestos from what it is among those without occupational asbestos exposure. Whether or not this type of effect modification exists depends on whether or not there is an interaction term in the additive risk model:

$$p = \beta_0 + \beta_A X_A + \beta_S X_S.$$

Similar correspondences hold for additive versions of the Poisson regression model

$$\lambda = \beta_0 + \beta_A X_A + \beta_S X_S,$$

and incidence rate regression models

$$\lambda(t) = \lambda_0(t) + \beta_A X_A + \beta_S X_S.$$

See Lin & Ying [4] for inference under the additive incidence rate model.

Example

Colditz et al. [3] studied how a variety of known risk factors for breast cancer were modified by family history of breast cancer. Data abstracted from the article are given in Table 1, broken down by whether the woman had ever given birth and whether she had a family history of breast cancer in a mother or sister. Crude relative risks based on a Poisson regression model are also given in Table 1.

From these results we see that without adjustment for other factors, among women without a history of breast cancer in their mother or sisters, the birth of a child appears to confer protection from breast cancer. However, among women with a history of breast cancer in the mother or a sister, the birth of a child is associated with, if anything, an increase in risk. If these differences are also seen in other studies, they might argue that screening schedules should be the most frequent in parous women with a family history of breast cancer. If the relative risk associated with parity among women with a family history were statistically different from one, these data would satisfy Thompson's [8] criteria for crossover, from which some interaction in the biological mechanisms might be inferred.

References

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. I: *The Analysis of Case–Control Studies*. Oxford University Press, Oxford.

[2] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*. Oxford University Press, Oxford.

[3] Colditz, G.A., Rosner, B.A. & Speizer, F.E. (1996). Risk factors for breast cancer according to family history of breast cancer, *Journal of the National Cancer Institute* **88**, 365–371.

[4] Lin, D.Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61–71.

[5] Miettinen, O. (1974). Confounding and effect modification, *American Journal of Epidemiology* **100**, 350–353.

[6] Rothman, K.J. (1976). Causes, *American Journal of Epidemiology* **104**, 587–593.

[7] Siemiatycki, J. & Thomas, D.C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis, *International Journal of Epidemiology* **10**, 383–387.

[8] Thompson, W.D. (1991). Effect modification and the limits of biological inference from epidemiologic data, *Journal of Clinical Epidemiology* **44**, 221–232.

BARBARA MCKNIGHT

Egret

Egret is a commercial software produced by Cytel Software Corporation. It was originally developed as a DOS-based software but a Windows version has been available since 1999, which makes Egret fairly easy to use. It is devoted to the analysis of epidemiologic and biomedical studies. It allows users to perform some data editing. Its main strength is its fairly extensive analytical capabilities. They include **linear regression**, unconditional and conditional **logistic regression** (with the useful option of including **random effect** terms), **Poisson regression**, and semiparametric and parametric **survival analysis**. Moreover, Egret offers additive and multiplicative versions of logistic and Poisson regression models (*see* **Relative Risk Modeling**), fairly extensive regression **diagnostics**, and **goodness-of-fit** procedures for all models, some exact or quasi-exact tests (*see* **Exact Inference for Categorical Data**) and procedures for the analysis of **contingency tables**, and has some graphical capabilities. Besides Egret, Egret-SIZ is a separate DOS-based program also produced by Cytel Software, which performs advanced **sample size** calculations for unconditional and conditional logistic regression, Poisson regression, and **Cox regression**, and offers the option of performing **power** calculations based on **Monte-Carlo** simulations.

(*See also* **Software, Biostatistical**; **Software, Epidemiological**; **Survival Analysis, Software**)

JACQUES BENICHOU

Eigenvalue

Eigenvalues (also known as eigenroots or latent roots) are a feature of square matrices. The definition and calculation of eigenvalues involves determinants. The eigenvalues of square matrix $\mathbf{A}$, of order n, are the n solutions for λ to what is called the *characteristic equation* of $\mathbf{A}$, namely,

$$|\mathbf{A} - \lambda\mathbf{I}| = 0, \tag{1}$$

i.e. the determinant of $\mathbf{A} - \lambda\mathbf{I}$ is equated to zero.

The nature of the determinant of a matrix (*see* **Matrix Algebra**) is such that for $\mathbf{A}$ of order n (1) is a polynomial equation of order n, thus having n solutions for λ. Those solutions are the eigenvalues of $\mathbf{A}$. As an example, for

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix},$$

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| &= \begin{vmatrix} 3-\lambda & 1 \\ 2 & 4-\lambda \end{vmatrix} \\ &= (3-\lambda)(4-\lambda) - 2 \\ &= \lambda^2 - 7\lambda + 10, \end{aligned}$$

and so the characteristic equation is

$$\lambda^2 - 7\lambda + 10 = 0,$$

which has solutions $\lambda = 2$ and 5. Thus 2 and 5 are the eigenvalues of $\mathbf{A}$.

General Properties

1. An eigenvalue, through being a solution of a polynomial equation, can be positive, negative, zero, real, or complex.
2. Eigenvalues of a matrix need not all be different. If λ is a root m times, then it is said to be a multiple root with multiplicity m.
3. For a scalar c, an eigenvalue of $c\mathbf{A}$ is c (eigenvalue of $\mathbf{A}$).
4. When λ is an eigenvalue of $\mathbf{A}$, λ^r is an eigenvalue of $\mathbf{A}^r$, for r being zero or a positive integer. This extends to a polynomial function of $\mathbf{A}$: $p(\mathbf{A})$ has $p(\lambda)$ as an eigenvalue.
5. The sum of all n eigenvalues of $\mathbf{A}$ of order n equals the trace of $\mathbf{A}$, the sum of its diagonal elements.
6. The product of all n eigenvalues of $\mathbf{A}$ is the determinant of $\mathbf{A}$.

Special Cases

1. Nonsingular matrices: all eigenvalues are nonzero.
2. Inverse matrices: $\mathbf{A}^{-1}$ has eigenvalue $1/\lambda$ where λ is an eigenvalue of $\mathbf{A}$.
3. Positive (semi)definite matrices: all eigenvalues are positive (zero or positive).
4. Symmetric matrices: all eigenvalues are real and the number of nonzero eigenvalues equals the rank of the matrix.
5. Orthogonal matrices: eigenvalues come in pairs λ and $1/\lambda$, with one value being ± 1 when the matrix is of odd order.
6. Idempotent matrices: all eigenvalues are $+1$ or zero; the number that is $+1$ is the rank of the matrix.

There are numerous uses of eigenvalues in **multivariate analysis** (see, for example, [1, Section 11.2]). One is in **principal components analysis** applied to a vector $\mathbf{X}$ of random variables. The linear combinations of these variables known as *principal components*, are $\boldsymbol{\beta}_r'\mathbf{X}$, where $\boldsymbol{\beta}_r'$ is the **eigenvector** corresponding to an eigenvalue λ_r of the **variance–covariance matrix** of the variable $\mathbf{X}$. Then the variance of $\boldsymbol{\beta}_r'$ is λ_r, and so one can rank the principal components according to the size of their variances.

A second use of eigenvalues is in **linear discriminant analysis** where one wants to classify observational units on the basis of a vector of variables, $\mathbf{X}$, say, measured on each unit. This is done using some $\mathbf{t}'\mathbf{X}$, often through maximizing the ratio of two quadratic forms, $\mathbf{t}'\mathbf{A}\mathbf{t}$ and $\mathbf{t}'\mathbf{B}\mathbf{t}$, say. This is achieved by choosing λ and $\mathbf{t}$ so that $(\mathbf{B} - \lambda\mathbf{A})\mathbf{t} = \mathbf{0}$, or equivalently $(\mathbf{A}^{-1}\mathbf{B} - \lambda\mathbf{I})\mathbf{t} = \mathbf{0}$. Thus, λ is the eigenvalue of $\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{t}$ a corresponding eigenvector.

Other uses of eigenvalues involve what is known as Wilks's test criterion for testing a linear hypothesis in the multivariate linear model. It consists of a ratio of the form of $\det \mathbf{A}/\det(\mathbf{A} + \mathbf{B})$, where $\mathbf{A}$ and $\mathbf{B}$ are matrices of sums of squares and products: $\det(\mathbf{A})$ represents the determinant of $\mathbf{A}$, which is the product of all the eigenvalues of $\mathbf{A}$ (see [3, Chapter 8]). Some alternatives to this are Hotelling's [2] trace

of $\mathbf{BA}^{-1}$, which is the sum of the eigenvalues of $\mathbf{BA}^{-1}$, or Pillai's [4] trace of $\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}$, being the sum of the eigenvalues of $\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}$, or Roy's [5] criterion of the largest eigenvalue of $\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}$.

References

[1] Anderson, R.L. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] Hotelling, H. (1951). A generalized t-test and measure of multivariate dispersion, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 23–41.

[3] Kshirsagar, A.M. (1972). *Multivariate Analysis*. Marcel Dekker, New York.

[4] Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis, *Annals of Mathematical Statistics* **26**, 117–121.

[5] Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

SHAYLE R. SEARLE

Eigenvector

Eigenvectors and eigenvalues (or eigenroots) are features of any square matrix. For square $\mathbf{A}$, of order n, its **eigenvalues** are the n solutions for λ to what is known as the characteristic equation:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0. \tag{1}$$

For each solution λ_i there is always a vector, $\mathbf{u}_i$, say, such that

$$\mathbf{A}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \quad \text{or equivalently } (\mathbf{A} - \lambda_i\mathbf{I})\mathbf{u}_i = \mathbf{0}. \tag{2}$$

The vector $\mathbf{u}_i$ is called an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ_i. Sometimes "characteristic" (or even, old-fashionably, "latent") is used in place of "eigen".

Example

For

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix},$$

(1) simplifies to $\lambda^2 - 7\lambda + 10 = 0$, giving $\lambda_1 = 2$ and $\lambda_2 = 5$ as eigenvalues. The second equation in (2) is satisfied as follows:

$$\begin{bmatrix} 3-2 & 1 \\ 2 & 4-2 \end{bmatrix}\begin{bmatrix} a \\ -a \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and}$$

$$\begin{bmatrix} 3-5 & 1 \\ 2 & 4-5 \end{bmatrix}\begin{bmatrix} b \\ 2b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{3}$$

Thus $\mathbf{u}_1 = [a \quad -a]'$ is the eigenvector corresponding to $\lambda_1 = 2$, and $\mathbf{u}_2 = [b \quad 2b]'$ corresponds to $\lambda_2 = 5$.

General Properties

For scalar c, an eigenvector of $\mathbf{A}$ is also an eigenvector of $c\mathbf{A}$. This is so because $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$ implies that $c\mathbf{A}\mathbf{u} = c\lambda\mathbf{u}$, i.e. $(c\mathbf{A})\mathbf{u} = (c\lambda)\mathbf{u}$. The latter is also $\mathbf{A}(c\mathbf{u}) = \lambda(c\mathbf{u})$, showing that if $\mathbf{u}$ is an eigenvector of $\mathbf{A}$ so is $c\mathbf{u}$. This is evident in (3), where a and b can be any scalars.

There is also the simple algebraic result that $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$ gives $\mathbf{A}^2\mathbf{u} = \mathbf{A}(\mathbf{A}\mathbf{u}) = \mathbf{A}(\lambda\mathbf{u}) = \lambda\mathbf{A}\mathbf{u} = \lambda(\lambda\mathbf{u}) = \lambda^2\mathbf{u}$. Thus $\mathbf{u}$ as an eigenvector of $\mathbf{A}$ is also an eigenvector of $\mathbf{A}^2$. This extends to $\mathbf{u}$ being an eigenvector of any integer power of $\mathbf{A}$ (and negative powers for nonsingular $\mathbf{A}$).

Because every eigenvalue λ_i has a corresponding eigenvector $\mathbf{u}_i$,

$$\mathbf{A}[\mathbf{u}_1 \quad \mathbf{u}_2 \ldots \mathbf{u}_i \ldots \mathbf{u}_n] = [\lambda_1\mathbf{u}_1 \quad \lambda_2\mathbf{u}_2 \ldots \lambda_i\mathbf{u}_i \ldots \lambda_n\mathbf{u}_n] \tag{4}$$

and, on defining $\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2 \ldots \mathbf{u}_i \ldots \mathbf{u}_n]$ and $\mathbf{D}$ as the diagonal matrix of the λs, (4) is

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{D}. \tag{5}$$

Calculation

For eigenvalue λ_i, the matrix $\mathbf{A} - \lambda_i\mathbf{I}$ is always singular. The theory of solving linear equations then yields a solution for $\mathbf{u}_i$ to (2) as $[(\mathbf{A} - \lambda_i\mathbf{I})^-(\mathbf{A} - \lambda_i\mathbf{I}) - \mathbf{I}]z$ for $(\mathbf{A} - \lambda_i\mathbf{I})^-$ being a generalized inverse (*see* **Matrix Algebra**) of $\mathbf{A} - \lambda_i\mathbf{I}$, and $\mathbf{z}$ being an arbitrary vector of order n.

Multiple Eigenvalues

Since (1) is a polynomial equation of order n it has n solutions for λ, which need not be all different. Suppose λ_t is a root m_t times, for $t = 1, \ldots, k$, for $\lambda_1 \ldots \lambda_k$ being all different. Then m_t is called the multiplicity of λ_t, and $\sum_{t=1}^{k} m_t = n$. When $(\mathbf{A} - \lambda_t\mathbf{I})$ has rank $n - m_t$ one can always find m_t linearly independent eigenvectors corresponding to λ_t. When this rank property holds for all $t = 1, \ldots, k$ (and it always holds whenever $m_t = 1$), then all eigenvectors are linearly independent and $\mathbf{U}$ is nonsingular.

Nonsymmetric Matrices

For nonsymmetric $\mathbf{A}$ it is the preceding rank condition (known as the diagonability theorem, e.g. [1, p. 305], which determines whether $\mathbf{U}$ is nonsingular or not. When it is nonsingular, (5) yields $\mathbf{D} = \mathbf{U}^{-1}\mathbf{A}\mathbf{U}$ and $\mathbf{D}$ is known as the canonical form under similarity, or equivalently as the similar canonical form. Likewise $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1}$ and $\mathbf{A}^r = \mathbf{U}\mathbf{D}^r\mathbf{U}^{-1}$.

Symmetric Matrices

For symmetric $\mathbf{A}$:

1. All λ_i and $\mathbf{u}_i$ are real.
2. $\mathbf{U}$ is always nonsingular.

3. Eigenvectors are pairwise orthogonal: $\mathbf{u}_i'\mathbf{u}_j = 0$ for $i \neq j$.
4. Each $\mathbf{u}_i$ can be standardized to be a unit vector $\mathbf{v}_i = \mathbf{u}_i/(\mathbf{u}_i'\mathbf{u}_i)^{1/2}$, so that $\mathbf{v}_i'\mathbf{v}_i = 1$.
5. Replacing each $\mathbf{u}_i$ in $\mathbf{U}$ by $\mathbf{v}_i$ makes $\mathbf{U}$ orthogonal: $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}$.
6. $\mathbf{D} = \mathbf{U}'\mathbf{A}\mathbf{U}$ is called the canonical form under orthogonal similarity; $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}' = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i'$, the latter being known as the spectral decomposition of $\mathbf{A}$.

These properties are important to statistics wherein symmetric matrices occur in a variety of situations, e.g. dispersion matrices, and $\mathbf{X}'\mathbf{X}$ in linear models.

Reference

[1] Searle, S.R. (1982). *Matrix Algebra Useful For Statistics*. Wiley, New York.

SHAYLE R. SEARLE

Eligibility and Exclusion Criteria

The choice of eligibility criteria in a **clinical trial** can increase or decrease the magnitude of between-patient variation, which will in turn decrease or increase the statistical **power** of the trial for a given sample size. Theoretically, the more homogeneous the trial population, the greater is the power of the trial, but the more limited is the ability to generalize the results to a broad population. Thus, the choice of eligibility criteria can profoundly influence both the results and the interpretation of the trial. Besides controlling variation, the Institute of Medicine (IOM) Committee on the Ethical and Legal Issues Relating to the Inclusion of Women in Clinical Studies [16] discusses four other issues related to the choice of trial population; namely, disease stage, clinical contraindications, regulatory or ethical restrictions, and compliance considerations. We will discuss these and the related issues of explanatory vs. pragmatic trials, screening and recruitment processes, and the impact of eligibility criteria on the generalizability of trial results. Other factors influencing the selection of patients, such as factors in the selection of institutions in **multicenter studies** and physician preferences are discussed elsewhere [2, 22].

Explanatory vs. Pragmatic Trials

The objectives of a trial affect the appropriate eligibility criteria [20]. If the trial is designed to estimate the biological effect of a treatment (explanatory trial), then the eligibility criteria should be chosen to minimize the impact of extraneous variation, as in early investigations of protease inhibitors against human immunodeficiency virus (HIV) infection [18]. If, however, the trial is designed to estimate the effectiveness of a treatment in a target population (pragmatic trial), then the eligibility criteria should be chosen to allow valid **inferences** to that population. For example, the Hypertension Prevention Trial (HPT) was aimed at normotensive individuals 25–49 years old with diastolic blood pressure between 78 mm Hg and 90 mm Hg, and these were the main eligibility criteria [3]. Choosing the narrow eligibility criteria often appropriate for an explanatory trial can make it difficult to apply the results to a broader population [11]. Yusuf [23], moreover, argues that a truly homogeneous cohort cannot be constituted because even apparently similar individuals can have very different outcomes. The consensus is that most Phase III randomized trials should be regarded as pragmatic.

The Uncertainty Principle

Byar et al. [4] describe the simplest possible form of eligibility criteria for a trial, in which patients are eligible provided the treating physician and the patient have "substantial uncertainty" as to which of the treatment options is better. This definition, known as the *uncertainty principle*, incorporates all factors that contraindicate one or more of the treatment options including stage of disease, co-existing disease, and patients' preferences (*see* **Ethics of Randomized Trials**). However, it also largely devolves definition of eligibility to the individual physicians participating in the trial. The consequent lack of control and strict definition of the cohort of patients entering the trial has been unattractive to some investigators.

Control of Variation vs. Ease of Recruitment

The debate over the uncertainty principle highlights the tension between two different ways of improving the precision of the estimated effect of treatment in a randomized trial. By using very strict eligibility criteria we seek to reduce between-patient variation in clinical outcomes, leading to improved precision of the treatment difference estimate. By using very flexible eligibility criteria (as with the uncertainty principle), we seek to allow a wider entry to the trial, thereby increasing the number of eligible patients and usually the precision of the treatment difference estimate. The question is, therefore, do we try to control variation and accept the smaller size of the sample, or do we try to increase the sample size and accept a wider between-patient variation? While this debate continues, the general consensus among clinical trial statisticians is that it is generally difficult to control between-patient variation successfully because often we do not know the important determinants of prognosis. Therefore, attempts to use very strict eligibility criteria are less successful than attempts to gain precision by entering very large numbers of patients into

trials [23]. However, if there are categories of patients who are considered very unlikely to benefit from the treatment, it is clearly conceivable to exclude them from the trial (see later sections on Stage of Disease and Clinical Contraindications).

Begg [2] criticizes the common practice of introducing a long list of eligibility criteria in clinical trials, particularly in the treatment of cancer. Such an approach greatly increases the difficulty of recruiting patients in large numbers. In examining such lists it is often found that many of the criteria are of questionable importance and do not relate directly to the safety of the patient or to the lack of benefit to be derived from the treatment.

Issues in the Screening and Recruitment Process

Establishing eligibility often involves a screening process. Examples include choosing individuals for a heart disease trial with ejection fraction between 0.35 and 0.8 and a specific number of ectopic beats, or choosing HIV-infected individuals with slowly rather than rapidly progressing disease [15].

Some eligibility criteria may be implicit in this process. For example, the recruitment method may require the patients to be accessible by telephone contact or to be able to read and write in English, such as trials in which the initial contact is via a prepaid postal response card. Multiple "baseline" visits that are sometimes used in the screening process can provide multiple opportunities for exclusion, e.g. the Coronary Primary Prevention Trial used five baseline visits and the HPT used three baseline visits. Thus, those ultimately enrolled may affect the recruitment and screening mechanisms and resources for multiple participant contacts as much as the protocol-specific eligibility criteria.

The impact of eligibility criteria and of recruitment procedures on the overall cost of the trial has rarely been investigated. Borhani [3] indicated that the ordering of the application of eligibility criteria can substantially affect costs. These costs are also sensitive to the cutoffs applied to continuous responses, e.g. diastolic blood pressure, high density lipoprotein cholesterol, coronary ejection fraction, and T-cell lymphocyte counts.

As mentioned earlier, eligibility criteria can have a strong impact on the ease of recruitment. For example, the need to enroll newly diagnosed or previously untreated patients can severely restrict the ability to recruit. The need to enroll rapidly after a stroke, myocardial infarction, head trauma, or exposure to infectious agent can lead to difficulties. If the condition renders the patient unconscious for some period of time, or the patient lives far from the treatment center, or is unaware that an infection, stroke, infarction, or other event has occurred, it is less likely that they will be available for enrollment. Similarly, Carew [5] suggests that recruitment be enhanced by broad eligibility criteria, allowing potentially more sites and more individuals to participate.

Stage of Disease

Often the stage of disease strongly affects the outcome of treatment, and is a primary source of variation. Eligibility is often restricted to the stages of disease most appropriately managed by the treatment.

For many diseases, classification or staging systems have been developed to aid clinical management. Eligibility criteria involving stage of disease are best defined using an established classification system that is in wide use. Examples of such classification systems include the coronary functional class [7], Dukes' colon cancer staging system, and the **World Health Organization** staging system for HIV infection [26].

Clinical Contraindications

Exclusions arising because one of the treatments is clearly contraindicated are common [14]. For example, 18% of those screened for the Beta Blocker Heart Attack Trial were excluded due to contraindications to the administration of propranolol [12]. Since these prior conditions would preclude use of some of the treatments in a trial, the trial results could not apply to individuals with those conditions. Some argue that contraindications should be clearly delineated in the protocol to avoid investigator or regional differences in their use.

Compliance Considerations

A run-in (or qualification) period is sometimes built into the trial design so as to identify potential noncompliers and exclude them from enrollment. This reduces the dilution of treatment differences that noncompliance introduces. In some studies, this period

can also be used to eliminate placebo (*see* **Blinding or Masking**) responders. In these cases, the determination of noncompliance becomes one of the **outcome measures of the trial** (*see* **Compliance Assessment in Clinical Trials**).

Regulatory or Ethical Considerations

Various demographically or otherwise defined populations have been excluded from clinical trials in the past. For example, in trials of heart disease prevention, women have been excluded as their incidence of heart disease is lower than in men and their inclusion would have required a larger sample size. Similarly, minority groups have sometimes had little or no representation because no special efforts had been made to include them. Recent changes in US regulations have required special justification for the exclusion of women, minorities, or the elderly from **National Institutes of Health** sponsored trials. The scientific argument for including these groups is that it provides a more solid basis for **extrapolating** the results of the trial to the general population [6, 8, 10, 13, 17, 19, 24, 25] (*see* **Validity and Generalizability in Epidemiologic Studies**). There will usually be inadequate statistical power for detecting different effects in subpopulations, but sometimes **meta-analysis** of several studies may be able to detect such differences.

Implementing the Eligibility Criteria

The characterization of the **target population** and baseline homogeneity can be subverted by deviations during the conduct of the trial from the protocol specified eligibility criteria. If extensive, these can adversely affect the assumptions underlying analyses and the interpretation of the results. Thus, monitoring the determination of eligibility criteria during the conduct of the trial is an important component of the implementation of the trial. Often, the office that conducts the **randomized treatment assignment** checks the eligibility criteria before enrolling the patient. Finkelstein & Green [9] discuss the exclusion from analysis of individuals found to be ineligible after enrollment in the trial.

Generalization of Results to Broader Populations

Treatment trials (or prevention trials) are usually conducted on samples of convenience, enrolling participants who present at specific hospitals or clinical sites. Therefore, the population to whom the trial results apply is generally not well defined. External validity – the ability to generalize from the trial to some broader population – is the ultimate goal of any trial. Adequately randomized trials can be assumed to produce valid results for the specific group of individuals enrolled, i.e. internal validity; the difficulties arise in extending the inference beyond that limited cohort. Since complete enumeration of the target population is rarely possible, inferences from studies are based on substantive judgment. A strong argument that is often used is that treatment *differences* in outcome are generally less variable among different patient populations than the outcomes themselves [23].

Following publication, critics questioned the generalizability of the results of the International Cooperative Trial of Extracranial–Intracranial (EC/IC) Arterial Anastomosis to evaluate the effect of the EC/IC procedure on the risk of ischemic stroke. The results showed a lack of benefit that surprised many in the surgical profession. It became clear that many of the eligible patients at the participating clinical sites did not enter the trial, while those enrolled in the trial were considered to have poorer risk and some argued that they were less likely to benefit from surgery [1, 21]. The ensuing controversy slowed acceptance of the trial results by the surgical community, although eventually they had a profound effect on the frequency with which EC/IC was performed.

Conclusions

The goals and objectives of the trial, the intended target population, and the anticipated inferences from the trial results should all be carefully specified from the outset. If that is done, then the appropriate choice of eligibility criteria usually becomes clearer. Experience has shown that simplifying eligibility criteria generally enhances recruitment, allows a wider participation, and gives greater justification for generalizing the results to a broader population.

References

[1] Barnett, H.J.M., Sackett, D., Taylor, D.W., Haynes, B., Peerless, S.J., Meissner, I., Hachinski, V. & Fox, A. (1987). Are the results of the extracranial–intracranial

bypass trial generalizable?, *New England Journal of Medicine* **316**, 820–824.

[2] Begg, C.B. (1988). Selection of patients for clinical trials, *Seminars in Oncology* **15**, 434–440.

[3] Borhani, N.O., Tonascia, J., Schlundt, D.G., Prineas, R.J. & Jefferys, J.L. (1989). Recruitment in the Hypertension Prevention Trial, *Controlled Clinical Trials* **10**, 30S–39S.

[4] Byar, D.P., Schoenfeld, D.A. & Green, S.B. (1990). Design considerations for AIDS trials, *New England Journal of Medicine* **323**, 1343–1348.

[5] Carew, B.D., Ahn, S.A., Boichot, H.D., Diesenfeldt, B.J., Dolan, N.A., Edens, T.R., Weiner, D.H. & Probstfield, J.L. (1992). Recruitment strategies in the Studies of Left Ventricular Dysfunction (SOLVD), *Controlled Clinical Trials* **13**, 325–338.

[6] Cotton P. (1990). Is there still too much extrapolation from data on middle-aged white men?, *Journal of the American Medical Association* **263**, 1049–1050.

[7] Criteria Committee of NYHA (1964). *Diseases of the Heart and Blood Vessels: Nomenclature and Criteria for Diagnosis*, 6th Ed. Little, Brown & Company, Boston.

[8] El-Sadr, W. & Capps, L. (1992). Special communication: the challenge of minority recruitment in clinical trials for AIDS, *Journal of the American Medical Association* **267**, 954–957.

[9] Finkelstein, D.M. & Green, S.B. (1995). Issues in analysis of AIDS clinical trials, in *AIDS Clinical Trials*, D.M. Finkelstein & D.A. Schonfeld, eds. Wiley–Liss, New York, pp. 243–256.

[10] Freedman L.S., Simon, R., Foulkes, M.A., Friedman, L., Geller, N.L., Gordon, D.J. & Mowery, R. (1995). Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993 – the perspective of NIH clinical trialists, *Controlled Clinical Trials* **16**, 277–285.

[11] Gail, M.H. (1985). Eligibility exclusions, losses to follow-up, removal of randomized patients, and uncounted events in cancer clinical trials, *Cancer Treatment Reports* **69**, 1107–1112.

[12] Goldstein, S., Byington, R. & the BHAT Research Group (1987). The Beta Blocker Heart Attack Trial: recruitment experience, *Controlled Clinical Trials* **8**, 79S–85S.

[13] Gurwitz, J.H., Col, N.F. & Avorn, J. (1992). The exclusion of the elderly and women from clinical trials in acute myocardial infarction, *Journal of the American Medical Association* **268**, 1417–1422.

[14] Harrison K., Veahov, D., Jones, K., Charron, K. & Clements, M.L. (1995). Medical eligibility, comprehension of the consent process, and retention of injection drug users recruited for an HIV vaccine trial, *Journal of Acquired Immune Deficiency Syndrome* **10**, 386–390.

[15] Haynes, B.F., Panteleo, G. & Fauci, A.S. (1996). Toward an understanding of the correlates of protective immunity to HIV infection, *Science* **271**, 324–328.

[16] IOM Committee on the Ethical and Legal Issues Relating to the Inclusion of Women in Clinical Studies, (1996). *Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies*, A.C. Mastroianni, R. Faden & D. Federman, eds. National Academy Press, Washington.

[17] Lagakos, S., Fischl, M.A., Stein, D.S., Lim, L. & Vollerding, P. (1991). Effects of zidovudine therapy in minority and other subpopulations with early HIV infection, *Journal of the American Medical Association* **266**, 2709–2712.

[18] Markowitz, M., Mo, H., Kempf, D.J., Norbeck, D.W., Bhat, T.N., Erickson, J.W. & Ho, D.D. (1996). Triple therapy with AZT, 3TC, and ritonavir in 12 subjects newly infected with HIV-1, Eleventh International Conference on AIDS, Abstract Th.B. 933.

[19] Patterson, W.B. & Emanuel, E.J. (1995). The eligibility of women for clinical research trials, *Journal of Clinical Oncology* **13**, 293–299.

[20] Schwartz, D., Flamant, R. & Lellouch, J. (1980). *Clinical Trials*. Academic Press, New York.

[21] Sundt, T.M. (1987). Was the international randomized trial of extracranial–intracranial arterial bypass representative of the population at risk?, *New England Journal of Medicine* **316**, 814–816.

[22] Taylor, K.M., Margolese, R.G. & Soskolne, C.L. (1984). Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer, *New England Journal of Medicine* **310**, 1363–1367.

[23] Yusuf, S., Held, P., Teo, K.K. & Toretsky, E.R. (1990). Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria, *Statistics in Medicine* **9**, 73–86.

[24] Yusuf, S. & Furberg, C.D. (1991). Are we biased in our approach to treating elderly patients with heart disease?, *American Journal of Cardiology* **68**, 954–956.

[25] Wenger, N.K. (1992). Exclusion of the elderly and women from coronary trials: is their quality of care compromised?, *Journal of the American Medical Association* **268**, 1460–1461.

[26] World Health Organization (1990). Acquired immune deficiency syndrome (AIDS): interim proposal for a WHO staging system for HIV infection and disease, *Weekly Epidemiology Record* **65**, 221–228.

(*See also* **Intention to Treat Analysis**)

MARY A. FOULKES

Eligibility Restriction

Eligibility restriction, or simply restriction, is a design strategy used to control for a potential **confounder**. For example, if gender is a potential confounder in a study of heart disease and diet, then one might choose to restrict the study to females. This design would eliminate gender as a potential confounder, but it would yield no direct information on the effect of diet on heart disease risk in males (*see* **Matching**).

Eligibility restrictions are also used in **clinical trials** for various other purposes, such as eliminating individuals not thought to benefit from the treatments under study or eliminating subjects not thought to be healthy enough to comply with protocol requirements (*see* **Eligibility and Exclusion Criteria**).

MITCHELL H. GAIL

Elston–Stewart Algorithm

Define a pedigree of size n to be a set of n related persons such that for each person in the set either neither parent or both parents are also in the set. Suppose that the first n_1 of these persons have no parents in the pedigree, and call them founders; and that the last $n - n_1$ of these persons have both parents in the pedigree, and call these nonfounders. Let g_i denote the **genotype**, and y_i the phenotype, of individual i. Assume that, conditional on the genotypes g_i, all the phenotypes y_i are independent. Then the joint probability of the g_i and y_i for all the pedigree members can be expressed as

$$\prod_{i=1}^{n_1} \Pr(g_i) \prod_{i=n_1+1}^{n_2} \Pr\left(g_i | g_{i_M}, g_{i_F}\right) \prod_{i=1}^{n} \Pr(y_i | g_i), \quad (1)$$

where $\Pr(g_i)$ is the population probability of founder i's genotype, $\Pr(g_i | g_{i_M}, g_{i_F})$ is the probability of nonfounder i's genotype conditional on the genotypes of the mother and father of i (**genetic transition probability**), and $\Pr(y_i | g_i)$ is the probability of i's phenotype conditional on genotype (**penetrance** function). Any of these probabilities may be a function of model parameters, and as a function of these the probability (1) is a joint **likelihood**. However, the genotypes are unobserved latent variables, so that the likelihood of interest, given only the observations y_i, is

$$\sum_{g_1} \sum_{g_2} \cdots \sum_{g_n} \prod_{i=1}^{n_1} \Pr(g_i) \prod_{i=n_1+1}^{n_2} \Pr\left(g_i | g_{i_M}, g_{i_F}\right) \times \prod_{i=1}^{n} \Pr(y_i | g_i). \quad (2)$$

Elston & Stewart [2] derived this likelihood by a different argument, specifically for a pedigree of simple structure – i.e. a pedigree in which there are no loops and each pedigree member traces back to the same single set of ancestral parents. This results in a formulation in which each summation sign in (2) is pushed as far to the right as possible, suggesting a recursive algorithm for its calculation that decreases the amount of computing time necessary. In particular, the amount of computation in this Elston–Stewart algorithm increases linearly with the size of the pedigree, but exponentially with the number of loci in the genotype g. The algorithm was soon adapted to pedigrees of complex structure [1, 6], but at increased computational cost. Later, with the advent of **multipoint linkage analysis**, a completely different algorithm for calculating a pedigree likelihood was proposed by Lander & Green [5] (see also [4]); the computational time for this algorithm, for which pedigree loops are no impediment, increases exponentially with the size of the pedigree but linearly with the number of loci. These two algorithms are widely used in the analysis of pedigree data.

The likelihood for a pedigree under **polygenic inheritance** can be expressed in a form analogous to (2): each summation is changed to an integration and each probability function becomes a normal density function (*see* **Normal Distribution**). Elston & Stewart [2] similarly expressed this likelihood with each integration sign pushed as far to the right as possible, and gave an algorithm to perform each integral in sequence analytically (see also [3]). In this way the likelihood under polygenic inheritance for a simple pedigree of size n, which can also be expressed as an n-variate normal density (*see* **Multivariate Normality, Tests of**), can be evaluated without needing inversion of an $n \times n$ symmetric matrix.

References

[1] Cannings, C., Thompson, E.A. & Skolnick, M.H. (1978). Probability functions on complex pedigrees, *Advances in Applied Probability* **10**, 26–61.

[2] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[3] Elston, R.C., George, V.T. & Severtson, F. (1992). The Elston–Stewart algorithm for continuous genotypes and environmental factors, *Human Heredity* **42**, 16–27.

[4] Kruglyak, L., Daly, J.H. & Lander, E.S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping, *American Journal of Human Genetics* **56**, 519–527.

[5] Lander, E.S. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans, *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2363–2367.

[6] Lange, K. & Elston, R.C. (1975). Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees, *Human Heredity* **25**, 95–105.

ROBERT C. ELSTON

EM Algorithm

In the practice of biostatistics, one is often faced with the problem of estimation with incomplete or missing data. Some common examples include **grouped**, **censored** or **truncated data**, and multivariate data with some individuals having missing responses (*see* **Missing Data**). Many simple techniques have been developed for this problem, which are based on a "filling in" or "successive substitution" algorithm. The idea is based on the intuitive notion that (i) if we had the values of the missing observations we could estimate the parameters in the "standard" way, often without iteration, and (ii) if we knew the parameters, we could "fill in" the missing observations by setting them equal to their expected values under the model. This suggests estimation can proceed iteratively, alternating between computing expectations for the incomplete data and estimating parameters with "complete data".

The EM Algorithm [5, 11, 17, 18, 30, 47, 57] is a general computational method for calculating **maximum likelihood** estimates with incomplete data. Its implementation capitalizes on the intuitive notion behind "filling in" algorithms. The name EM was introduced by Dempster et al. [11], hereinafter referred to as DLR; its name is derived from the two steps required at each iteration: an E-step for computing the expectation of the missing data and an M-step for computing the maximum likelihood estimates of the parameters assuming complete data. Not all "filling in" algorithms are versions of the EM algorithm; both the E-step and the M-step must be specified with reference to an appropriate formulation of the incomplete data and its likelihood in order for the resulting parameter estimates to be ML.

Before describing the EM algorithm and presenting some of its properties, we present two simple examples which illustrate many of the features of the EM. The first is a very simple version of the gene counting algorithm in genetics, one of the earliest uses of the EM algorithm. The second example is two-way **analysis of variance** (ANOVA) with missing cells. In this case, a naive "filling in" algorithm is not EM, but a simple adjustment can be made which leads to the EM, using the general theory presented in the following section.

Example 1: Gene Counting

Suppose we wish to estimate the frequency of an allele A at a gene (*see* **Gene Frequency Estimation**), based on a random sample of N individuals. Each person has two alleles; we assume **Hardy–Weinberg equilibrium** and random mating, so that if we could observe the value of each allele directly, the desired probability estimate would be the observed proportion of A alleles out of $2N$ independent alleles. To formalize this idea, we introduce some notation. Let p_{A} be the proportion of A alleles, n_{AA} be the number of individuals who are AA, and likewise for n_{Aa} and n_{aa}, so that $n_{\mathrm{AA}} + n_{\mathrm{Aa}} + n_{\mathrm{aa}} = N$ and $\hat{p}_{\mathrm{A}} = (2n_{\mathrm{AA}} + n_{\mathrm{Aa}})/2N$. Here, lower-case "a" is used to denote an allele which is not A.

Depending upon what data are available on the N individuals, the value of an allele may not be directly observable. However, p_{A} can still be estimated by assuming a specific genetic model. For illustration, we will assume we observe a recessive trait with $Y = 1$ if a person is AA, and $Y = 0$ otherwise. Here we know an individual has two A alleles if $Y = 1$, but if $Y = 0$ we only know they have less than two A alleles. Denote by n_1 and n_0 the number of individuals with $Y = 1$ and $Y = 0$, respectively. Notice that $n_1 = n_{\mathrm{AA}}$ and $n_0 = n_{\mathrm{Aa}} + n_{\mathrm{aa}}$. Given n_0 and n_1 and a provisional estimate of p_{A}, say $\tilde{p}_{\mathrm{A}}$, we can calculate the expected number of A alleles as

$$\mathrm{E}(2n_{\mathrm{AA}} + n_{\mathrm{Aa}}) = 2n_1 + \mathrm{E}(n_{\mathrm{Aa}}|n_0, n_1, \tilde{p}_{\mathrm{a}}).$$

The last term can be easily calculated assuming Hardy–Weinberg equilibrium as

$$\begin{aligned}\tilde{n}_{\mathrm{Aa}} &= \mathrm{E}(n_{\mathrm{Aa}}|n_0, n_1, \tilde{p}_{\mathrm{A}}) = n_0 \Pr(\mathrm{Aa}|\mathrm{Aa \text{ or } aa}) \\ &= \frac{n_0 2\tilde{p}_{\mathrm{A}}(1-\tilde{p}_{\mathrm{A}})}{2\tilde{p}_{\mathrm{A}}(1-\tilde{p}_{\mathrm{A}}) + (1-\tilde{p}_{\mathrm{A}})^2} \\ &= \frac{n_0 2\tilde{p}_{\mathrm{A}}}{1+\tilde{p}_{\mathrm{A}}}. \qquad (1)\end{aligned}$$

Now, treating $\tilde{n}_{\mathrm{Aa}}$ as if it were actually the count of Aa individuals, we get our updated estimate of p_{A} as

$$\begin{aligned}p_{\mathrm{A}}^{\mathrm{new}} &= \frac{(2n_{\mathrm{AA}} + \tilde{n}_{\mathrm{Aa}})}{2N} \qquad (2) \\ &= \frac{2n_1 + 2n_0\tilde{p}_{\mathrm{A}}/(1+\tilde{p}_{\mathrm{A}})}{2N}. \qquad (3)\end{aligned}$$

In this example, iterating between (1) and (2) to solve for p_A will yield the ML estimate. Note that setting $\tilde{p}_A = p_A^{new} = \hat{p}_A$ in (3) yields an equation for $\hat{p}_A$ which can be solved to yield

$$\hat{p}_A = \frac{\sqrt{n_1}}{N}.$$

This estimate can be derived directly by noting that under Hardy–Weinberg equilibrium $\Pr(Y = 1) = p_A^2$: showing that it is a fixed point of the EM algorithm shows that $\hat{p}_A$ is ML. Although no iteration is required in this simple example, with three or more alleles iterative computations are required and the EM is often used. Ceppellini et al. [8], present the gene counting algorithm for the general case, and also consider estimation of allele frequencies when the data consist of random samples of families. Hartl & Clark [16] and Lange [29] give examples implementing the EM for estimating the allele frequencies which control the ABO **blood groups**.

Example 2: ANOVA with Missing Cells

Suppose we have the usual two-way layout with one observation per cell, $x_{ij}, i = 1, \ldots, I, j = 1, \ldots, J$. We assume the standard model (*see* **Analysis of Variance**):

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where $\sum \alpha_i = \sum \beta_j = 0$ and the e_{ij}s are taken to be independent with zero mean and variance σ^2. With complete data, the standard estimates for μ, α_i, β_j and σ^2 are

$$\begin{aligned} \mu &= \overline{x}_{..}, \\ \alpha_i &= (\overline{x}_{i.} - \overline{x}_{..}), \\ \beta_j &= (\overline{x}_{.j} - \overline{x}_{..}), \end{aligned} \tag{4}$$

$$\sigma^2 = \frac{\sum_{ij}(x_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}_{..})^2}{(I-1)(J-1)}. \tag{5}$$

If observations are missing for some cells, one can still obtain **least squares** estimates for the parameters by setting up a design matrix (which is no longer **orthogonal**), and using any standard regression package (*see* **Software, Biostatistical**). Prior to the easy availability of regression packages, a popular alternative to the required matrix inversion was to use "filling in" algorithms. Several versions were proposed; Little & Rubin [35] give a rationale for using "filling in" algorithms in this setting and review the literature. The version we describe is due to Healy & Westmacott [19] and is also given in Little & Rubin [35].

One can start the algorithm with any set of initial values for the parameters. One could, for example, use (4) and (5) where the means and sum of squares are based only on those cells where x_{ij} is observed. Having provisional estimates, any missing cell, say x_{lk}, is filled in by

$$x_{lk} = \mathrm{E}(x_{lk}|\tilde{\mu}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) = \tilde{\mu} + \tilde{\alpha}_l + \tilde{\beta}_k. \tag{6}$$

Since (4) and (6) do not involve σ^2, we may iterate between (4) and (6) to obtain the least squares estimates of $\mu, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$. That is, given the current estimates, fill in for any missing cells using (6) and then use (4) to reestimate $\mu, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$, treating the $\tilde{x}_{lk}$ as observed data.

Using "filling in" algorithms to obtain estimates of variance components is more complicated. A naive approach would be to continue to use (6) to fill in for the missing values and to use (5) for estimating σ^2. But since $(\tilde{x}_{lk} - \tilde{\mu} - \tilde{\alpha}_l - \tilde{\beta}_k)^2 = 0$ for all missing cells, this would be equivalent to ignoring the missing cells in computing the sums of squares and dividing by an inflated number of degrees of freedom. Healy & Westmacott [19] suggested just this, but replacing $(I-1)(J-1)$ with the correct number of degrees of freedom, $(I-1)(J-1) - m$, where m denotes the number of missing cells.

Alternatively, notice that for least squares estimates, (5) can be written

$$(I-1)(J-1)\hat{\sigma}^2 = \sum_{ij} x_{ij}^2 - \sum_{ij}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)^2. \tag{7}$$

This suggests filling in $\mathrm{E}(x_{lk}^2|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})$ in (7) for the missing x_{lk}. Since

$$\mathrm{E}(x_{lk}^2|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) = \sigma^2 + (\mu + \alpha_l + \beta_k)^2,$$

this would lead to the following iterative equation for σ^2:

$$\sigma_{new}^2 = \frac{\sum_{ij}(x_{ij} - \tilde{\mu} - \tilde{\alpha}_i - \tilde{\beta}_j)^2 + m\tilde{\sigma}^2}{(I-1)(J-1)}. \tag{8}$$

Since $(\mu, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ do not depend upon σ^2, (8) can be solved directly for $\hat{\sigma}^2$ to yield

$$\hat{\sigma}^2 = \frac{\sum_{ij}(x_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2}{(I-1)(J-1) - m}, \tag{9}$$

yielding the Healy & Westmacott estimator. This is not ML under the assumption of normality for the error terms because the maximum likelihood estimate of σ^2 is the error sums of squares divided by the sample size rather than the number of degrees of freedom. Substituting IJ for $(I-1)(J-1)$ in (8) and (9) yields the EM equations and the ML estimator for σ^2.

These simple examples illustrate several features of the algorithm. First is the simple and intuitive nature of the algorithm. This is particularly true for the gene counting algorithm. For the second example, filling in $\mathrm{E}(x_{lk}^2)$ for the missing cells in estimating the variance is not so obvious. The general principle, discussed in the next section is that one must "fill in" by computing expectations of the **sufficient statistics** (in the case of **exponential families**). If all sufficient statistics are linear in the data, as they are in standard multinomial problems, then filling in by computing expectations of the data is appropriate. Second is the simple nature of each step of the algorithm. Although it can be implemented in a wide variety of situations, it is easiest to implement when maximum likelihood with complete data has closed form solutions and when the required expectations can be computed in closed form. Third is the fact that the EM is not always the best computational procedure; in both of our simple examples, the existence of a closed form solution and widespread availability of computing resources make the EM obsolete in these cases. Even in this setting, characterizing an estimate as the fixed point of the EM can be an easy method for deriving formulas for maximum likelihood estimates. We now turn to a discussion of the general theory underlying the EM.

Maximum Likelihood Estimation with Incomplete Data: The EM Algorithm

We discuss here the problem of maximum likelihood with incomplete data in complete generality, but will refer back to our two examples to make ideas concrete. We denote the complete data vector by $\boldsymbol{x}$ and its associated density by $f(\mathbf{x}|\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ denotes an r-vector of parameters. Here $\mathbf{x}$ could be an n-vector of independent scalar observations as in Example 2 where $\mathbf{x} = (x_{11}, \ldots, x_{IJ})^{\mathrm{T}}$. In Example 1, there are many choices for $\mathbf{x}$; it could be the $2N$ vector of allele values, the N vector of genotypes for each subject, or the counts $n_{\mathrm{AA}}, n_{\mathrm{Aa}}$, and n_{aa}. Any choice leads to equivalent results, but potentially different implementations of the algorithm. We will take $\mathbf{x}$ to be the $2N$ vector of allele values. In many settings, each sampling unit will contribute a vector of observations to the complete data vector.

Denote the observed data by $\mathbf{y}$. There are two sample spaces, X and Y, and the data vectors $\mathbf{x}$ and $\mathbf{y}$ define a many-to-one mapping from X to Y. For example 2, X is R^{IJ} and Y is R^{IJ-m}; $\mathbf{y}$ is the $IJ-m$ vector of observed x_{ij}s. For Example 1, X is the space of $2N$ vectors whose components are binary, x_{i1} and x_{i2} being indicator vectors for the ith subject's two alleles, i.e. $x_{ij} = 1$ if A, and 0 otherwise. We shall take Y to be the space of N vectors whose components are also binary, $y_i = 1$ if a person is AA, $y_i = 0$ otherwise.

By definition, the density of the observed data $\mathbf{y}$ can be written as

$$g(\mathbf{y}|\boldsymbol{\phi}) = \int_{X(\mathbf{y})} f(\mathbf{x}|\boldsymbol{\phi})\, \mathrm{d}\mathbf{x}, \tag{10}$$

where $X(\mathbf{y})$ denotes the subset of X where $\mathbf{x}$ must lie, having observed $\mathbf{y}$. For example 2, $X(\mathbf{y})$ is just R^m and (10) simply integrates out the missing x_{ij}s, leaving the normal density of the observed data vector. For Example 2, we have $y_i = x_{i1}x_{i2}$; hence the integral is a summation over all values of x_{i1} and x_{i2} which yield y_i. This implies that each y_i is binary with probability density

$$\begin{aligned} g(y_i) &= (p_{\mathrm{A}}^2)^{y_i}(2p_{\mathrm{A}}(1-p_{\mathrm{A}}) + (1-p_{\mathrm{A}})^2)^{1-y_i} \\ &= \sum p_{\mathrm{A}}^{x_{i1}} p_{\mathrm{A}}^{x_{i2}} (1-p_{\mathrm{A}})^{1-x_{i1}} (1-p_{\mathrm{A}})^{1-x_{i2}}, \end{aligned}$$

where summation is over all (x_{i1}, x_{i2}) such that $y_i = x_{i1}x_{i2}$.

Notice that this formulation is quite general and does not apply merely to missing data in the usual setting where one or more subjects are missing some or all of their observations. It also applies to convolutions, where each component of y is the sum of several components of x, and to latent variable problems (*see* **Random Coefficient Repeated Measures Model**), **random effects** models and **mixtures**

(*see* **Method of Moments**), where inherently unobservable variables are considered to be part of the complete data.

Since $\mathbf{y}$ is completely determined by $\mathbf{x}$, their joint density is simply $f(\mathbf{x}|\boldsymbol{\phi})$. Thus, by definition,

$$f(\mathbf{x}|\boldsymbol{\phi}) = k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})g(\mathbf{y}|\boldsymbol{\phi}),$$

where $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$ denotes the conditional density of $\mathbf{x}$ given $\mathbf{y}$. Taking logs yields

$$l(\boldsymbol{\phi}; \mathbf{x}) = l(\boldsymbol{\phi}; \mathbf{y}) + l(\boldsymbol{\phi}; \mathbf{x}|\mathbf{y}), \tag{11}$$

where $l(\boldsymbol{\phi}; \cdot)$ denotes the log likelihood associated with the density of $(\cdot|\boldsymbol{\phi})$. To obtain the ML estimate of $\boldsymbol{\phi}$ we maximize $l(\boldsymbol{\phi}; \mathbf{y})$, or equivalently $l(\boldsymbol{\phi}; \mathbf{x}) - l(\boldsymbol{\phi}; \mathbf{x}|\mathbf{y})$. Taking the expectation of both sides of (11) with respect to $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}')$, we get

$$Q(\boldsymbol{\phi}|\boldsymbol{\phi}') = L(\boldsymbol{\phi}) + H(\boldsymbol{\phi}|\boldsymbol{\phi}'), \tag{12}$$

where

$$L(\boldsymbol{\phi}) \equiv l(\boldsymbol{\phi}; \mathbf{y}),$$

$$Q(\boldsymbol{\phi}|\boldsymbol{\phi}') = \mathrm{E}(\log f(\mathbf{x}|\boldsymbol{\phi})|\mathbf{y}, \boldsymbol{\phi}'),$$

and

$$H(\boldsymbol{\phi}|\boldsymbol{\phi}') = \mathrm{E}(\log k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})|\mathbf{y}, \boldsymbol{\phi}').$$

As a consequence of the **information** inequality [30], $H(\boldsymbol{\phi}|\boldsymbol{\phi}')$ is maximized as a function of $\boldsymbol{\phi}$ by setting $\boldsymbol{\phi} = \boldsymbol{\phi}'$ for any $\boldsymbol{\phi}'$. Furthermore, the ML estimator, $\hat{\boldsymbol{\phi}}$, maximizes $L(\boldsymbol{\phi})$ by definition. Thus $L(\boldsymbol{\phi})$ and $H(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}})$ are both maximized when $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$, and it follows that $Q(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}})$ is maximized by setting $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$. This fact both provides a characterization of the MLE in terms of the complete data log likelihood, and also suggests a computing **algorithm**. The algorithm in its general form is as follows. Given the current estimate of $\boldsymbol{\phi}$ at the pth iteration, say $\boldsymbol{\phi}^{(p)}$:
E-step: Compute

$$\mathrm{E}[\log f(\mathbf{x}|\boldsymbol{\phi})|\mathbf{y}, \boldsymbol{\phi}^{(p)}] = Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)}).$$

M-step: Maximize $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)}$ as a function of $\boldsymbol{\phi}$ to obtain $\boldsymbol{\phi}^{(p+1)}$. It follows immediately that the MLE is a fixed point of the EM algorithm, since $\hat{\boldsymbol{\phi}}$ will maximize $Q(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}})$.

Although this general formulation can be successfully implemented in many examples, it is easiest to implement the EM when the complete data density has a regular **exponential family** form. In this case, ignoring any functions of $\mathbf{x}$ alone, we can write

$$\log f(\mathbf{x}|\boldsymbol{\phi}) = \boldsymbol{\phi}^{\mathrm{T}}\mathbf{t} - \log a(\boldsymbol{\phi}),$$

where we now assume $\boldsymbol{\phi}$ denotes an r-vector of the canonical parameters and $\mathbf{t}$ is an r-vector of sufficient statistics which are functions of $\mathbf{x}$. Since $\log f(\mathbf{x}|\boldsymbol{\phi})$ is linear in the sufficient statistics, the E-step is easily implemented by setting:
E-step: $\mathbf{t}^{(p)} = \mathrm{E}(\mathbf{t}|\mathbf{y}; \boldsymbol{\phi}^{p})$.

For the M-step we maximize $\boldsymbol{\phi}^{\mathrm{T}}\mathbf{t}^{(p)} - \log a(\boldsymbol{\phi})$, treating $\mathbf{t}^{(p)}$ as if it were the sufficient statistic based on completely observed data. For regular exponential families, $\partial \log a(\boldsymbol{\phi})/\partial\boldsymbol{\phi} = \mathrm{E}(\mathbf{t}|\boldsymbol{\phi})$, hence the likelihood equations are obtained by setting $\mathbf{t}$ equal to its expected value, or solving

$$\mathrm{E}(\mathbf{t}|\boldsymbol{\phi}) - \mathbf{t} = 0.$$

Thus, regarding $\mathbf{t}^{(p)}$ as data, and maximizing $\boldsymbol{\phi}^{T}\mathbf{t}^{(p)} - \log a(\boldsymbol{\phi})$ to find $\boldsymbol{\phi}^{p+1}$ leads to:
M-step: Solve

$$\mathrm{E}(\mathbf{t}|\boldsymbol{\phi}^{(p+1)}) - \mathbf{t}^{(p)} = 0.$$

At convergence $\boldsymbol{\phi}^{(p+1)} = \boldsymbol{\phi}^{(p)}$, and we have

$$\mathrm{E}(\mathbf{t}|\hat{\boldsymbol{\phi}}) - \mathrm{E}(\mathbf{t}|\mathbf{y}, \hat{\boldsymbol{\phi}}) = 0.$$

This striking representation for the likelihood equations provides an easy derivation in settings where the complete data can be chosen to have an exponential family distribution. The recipe is: (i) find the vector of sufficient statistics t, assuming complete data (ii) find an expression for its expectation, $\mathrm{E}(\mathbf{t}|\boldsymbol{\phi})$, as a function of $\boldsymbol{\phi}$ with respect to the complete data, (iii) find an expression for the expectation given the observed data $\mathbf{y}$, $\mathrm{E}(\mathbf{t}|\mathbf{y}, \boldsymbol{\phi})$, and (iv) equate the two expectations to solve for $\boldsymbol{\phi}$. Since the representation for the complete data is not unique, there may be several choices for $\mathbf{x}$ and thus $\mathbf{t}$; some choices may lead to an easier solution of the M- and E-steps than others. In multivariate settings, such as multinomial or normal, the conditional distribution of the complete data conditional on observed data will often have the same form as the complete data, making the conditional expectations easy to compute.

Having given a general characterization of the EM, we now return to our two examples. The properties of the algorithm and other applications are discussed in the following sections.

Example 1

Since we chose $\mathbf{x}$ to be the $2N$ vector of indicators for each allele, our complete data likelihood is binomial $(p_A, 2N)$; hence the sufficient statistic is the number of A alleles, or $t = \sum_i (x_{i1} + x_{i2})$. The M-step is $\tilde{p_A} = t/2N$ and given $\tilde{p}_A$ the E-step computes $E(\mathbf{t}|\mathbf{y}, \tilde{p}_A)$. Since t is linear in $x_{i1} + x_{i2}$, its expectation can be computed by considering

$$E(x_{i1} + x_{i2}|y_i, \tilde{p}_A) = \begin{cases} 2, & \text{if } y_i = 1, \\ 2\tilde{p}_A/(1 + \tilde{p}_A), & \text{if } y_i = 0. \end{cases} \tag{13}$$

Thus, $E(\mathbf{t}|\mathbf{y}, \tilde{p}_A) = 2n_1 + 2n_0\tilde{p}_A/(1 + \tilde{p}_A)$, as before. Note that (13) is based on $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$; in this example, and in many others, it is not necessary to calculate this density in complete generality, but only to be able to take conditional moments of x_{ij}.

Example 2

Assuming the error terms are independently and normally distributed gives an exponential family form for $f(\mathbf{x}|\boldsymbol{\phi})$. Two choices are possible for dealing with the parameter space. We can take $\boldsymbol{\phi}$ to consist of μ, σ^2, the $I\alpha_i$s, the $J\beta_j$s and impose constraints, or take only $(I-1)\alpha_i$s and $(J-1)\beta_j$s. We choose the former because it yields a simpler set of sufficient statistics and illustrates the implementation of the EM with constraints; either choice gives equivalent results.

When $\boldsymbol{\phi}$ is the vector of $(I + J + 2)$ parameters, the sufficient statistics are easily found to be

$$\begin{aligned} t_1 &= x_{++}, \\ \mathbf{t}_2^T &= (x_{1+}, \ldots, x_{I+}), \\ \mathbf{t}_3^T &= (x_{+1}, \ldots, x_{+J}), \\ t_4 &= \sum\sum x_{ij}^2. \end{aligned} \tag{14}$$

Here a $+$ replacing a subscript indicates summation over that index. The complete data log likelihood can be maximized by setting $\mathbf{t}^T = (t_1, t_2^T, t_3^T, t_4)$ equal to its expected value, subject to the constraints $\sum \alpha_i = 0$ and $\sum \beta_j = 0$. Assuming these constraints hold, the expectations of $\mathbf{t}$ are:

$$\begin{aligned} E(t_1) &= IJ\mu, \\ E(\mathbf{t}_2^T) &= J(\mu + \alpha_1, \ldots, \mu + \alpha_I), \\ E(\mathbf{t}_3^T) &= I(\mu + \beta_1, \ldots, \mu + \beta_J), \\ E(t_4) &= IJ\sigma^2 + \sum\sum(\mu + \alpha_i + \beta_j)^2. \end{aligned} \tag{15}$$

Equating (14) and (15) and solving for the parameters gives the well-known estimates (apart from the denominator of σ^2):

$$\begin{aligned} \hat{\mu} &= \overline{x}_{++}, \\ \hat{\alpha}_i &= \overline{x}_{i+} - \overline{x}_{++}, \\ \hat{\beta}_j &= \overline{x}_{+j} - \overline{x}_{++}, \\ \hat{\sigma}^2 &= \frac{1}{IJ}\sum(x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot J} + \overline{x}_{\cdot\cdot})^2. \end{aligned} \tag{16}$$

If cells are missing, then the E-step is easy, since $\mathbf{t}$ is just summations of x_{ij} and x_{ij}^2 over individuals. Explicit derivation of the sufficient statistics shows why $E(x_{lk})$ is substituted into the formulas for $(\hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and $E(x_{lk}^2)$ is substituted into the formula for $\hat{\sigma}^2$. In this case the conditional expectations are particularly easy since the x_{ij} are all independent. This implies $f(x_{lk}|\mathbf{y}, \boldsymbol{\phi}) = f(x_{lk}|\boldsymbol{\phi}) \sim N(\mu + \alpha_l + \beta_k, \sigma^2)$, i.e. the distribution of the unobserved xs is independent of the observed ys, but it does depend upon the unknown parameters.

Notice that it is not actually necessary to identify the vector of canonical parameters, $\boldsymbol{\phi}$. It is only necessary to identify the vector of sufficient statistics $\mathbf{t}$ and to be able to express $E(\mathbf{t}|\boldsymbol{\phi})$ as a function of the parameters of interest. In our Example 2 it was particularly easy to solve the complete data maximization problem subject to the constraints. In other settings Lagrange multipliers can be used.

The use of the term EM *algorithm* is sometimes criticized because it is actually only a prescription for an algorithm. The exact formulas for its implementation will vary in each application. In many examples the E- and M-steps will be obvious and easily implemented. In other cases, either the E- or the M-step will be difficult; later we will consider extensions of the EM designed to deal with several problems which arise in practice. First we will consider the properties of the EM.

Properties of the EM Algorithm

As previously mentioned, the MLE of ϕ is a fixed point of the EM. In addition, the EM is numerically very stable and each iteration increases the likelihood. This fact is easily proved. By definition of the algorithm,

$$Q(\boldsymbol{\phi}^{(p+1)}|\boldsymbol{\phi}^{(p)}) \geq Q(\boldsymbol{\phi}^{(p)}|\boldsymbol{\phi}^{(p)}), \qquad (17)$$

and, as previously noted, the information inequality implies

$$H(\boldsymbol{\phi}^{(p+1)}|\boldsymbol{\phi}^{(p)}) \leq H(\boldsymbol{\phi}^{(p)}|\boldsymbol{\phi}^{(p)}).$$

Thus

$$\begin{aligned} L(\boldsymbol{\phi}^{(p+1)}) &= Q(\boldsymbol{\phi}^{(p+1)}|\boldsymbol{\phi}^{(p)}) - H(\boldsymbol{\phi}^{(p+1)}|\boldsymbol{\phi}^{(p)}) \\ &\geq Q(\boldsymbol{\phi}^{(p)}|\boldsymbol{\phi}^{(p)}) - H(\boldsymbol{\phi}^{(p)}|\boldsymbol{\phi}^{(p)}) = L(\boldsymbol{\phi}^{(p)}). \end{aligned}$$

See also Baum et al. [5]; DLR, and Lange [30].

The EM algorithm naturally incorporates parameter bounds and constraints whenever the complete data come from an exponential family. For example, with a complete data sample from the **multivariate normal**, the sample variance–**covariance matrix** is the MLE; it is also nonnegative definite. With missing data, using the EM to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ means that, provided the initial value for $\boldsymbol{\Sigma}$ is nonnegative definite, each iterate will remain so. Another common example is estimates of probabilities which are equal, with complete data, to x/n, say, for a count $0 < x < n$. When x and possibly also n are incompletely observed, the probability estimate will remain between zero and one with the EM, since $0 < \mathrm{E}(x) < \mathrm{E}(n)$. Parameter bounds such as these will not hold in general for other iterative algorithms.

The EM is not guaranteed to converge, even to a local maximum, except in special circumstances. The original proof of convergence given in DLR is flawed. Wu [63] and Boyles [6] have studied convergence. Two general results given by Wu [63] are (1) if the complete data density is a curved exponential family with a compact parameter space, then all limit points of any EM sequence are stationary points of the likelihood and (2) if the likelihood function is unimodal and the first derivative of the Q function with respect to $\boldsymbol{\phi}$ is continuous in both $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$, then the EM converges to a unique maximum. In most instances it will be difficult to show that an incomplete data likelihood is unimodal; thus case 1 is of more interest. Since convergence guarantees only a stationary point, and not even a local maximum, much less a global one, it is sometimes useful to use several starting values for the algorithm. Multiple solutions are a feature of many incomplete data problems, especially boundary solutions. See Baker et al. [3] and Baker & Laird [4].

Even though the EM algorithm may not provide the most efficient computational approach, it often provides an easy way to characterize the derivatives of the log-likelihood, and thus expressions for the ML score equations and observed information matrix, in terms of the complete data log likelihood. In addition, because of its stability and simplicity, the EM algorithm can be easily programmed in many cases; its property of increasing the likelihood can be useful in debugging programs. These facts make it worthwhile to represent data as incomplete in cases where it is possible to do so, while choosing the complete data to correspond to an easily handled case.

As an optimization procedure, the EM enjoys many advantages over its main competitor, Newton–Raphson. Although more iterations are generally required, each iteration may be faster and easier to program. It is less sensitive to poor starting values, it automatically bounds parameter values in their proper space and it is easier to implement with many parameters.

The EM algorithm is a method for computing ML estimates which does not require second derivatives. This is attractive when second derivatives are difficult to evaluate and often makes it easy to program the equations, but it can also mean slow convergence and the asymptotic variance–covariance matrix is not available as a byproduct of the computations. Enhancements to the basic algorithm to deal with these and other issues are discussed in the next section.

Enhancements and Modifications to the EM Algorithm

Because of its popularity, many enhancements and modifications have been suggested for the EM algorithm. If the M-step and/or the E-step are difficult, i.e. the E-step requires numerical integration or the M-step requires numerical methods to maximize the Q-function, then it may be infeasible to implement the EM easily. In their original paper, DLR

proposed a Generalized EM (GEM) designed to deal with the case where maximization is difficult at the M-step. A GEM algorithm is an EM, with the M-step changed to

M-step: Choose $\boldsymbol{\phi}^{p+1}$ so that

$$Q(\boldsymbol{\phi}^{(p+1)}|\boldsymbol{\phi}^{(p)}) \geq Q(\boldsymbol{\phi}^{(p)}|\boldsymbol{\phi}^{(p)}).$$

Clearly the GEM retains many of the properties of the EM and will be easier to implement in some cases.

Various other proposals have been suggested for dealing with a Q-function which is difficult to maximize. One obvious approach to maximizing the Q-function at each iteration is to use some rapidly converging algorithm, such as Newton–Raphson, and only take one iteration away from $\boldsymbol{\phi}^{(p)}$ at each M-step. Providing the single iteration increases the likelihood, this would provide an instance of a GEM. This approach has been formalized by Lange [27] who describes a gradient algorithm locally equivalent to EM. Meng & Rubin [45] propose using a series of "conditional" maximization steps (hence the name ECM), or cyclic coordinate ascent to maximize $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)})$. Liu & Rubin [36] extend this idea with the ECME algorithm, which allows some components of $\boldsymbol{\phi}^{(p+1)}$ to be chosen to maximize the observed data log likelihood rather than $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)})$. Green [14] proposes an extension of the EM for the Bayesian setting, where addition of the log of the prior to the Q function makes its maximization intractable.

There have been fewer proposals for how to deal with intractable E-steps arising because numerical methods are required to evaluate the expectations. One approach is to approximate $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}^{(p)})$ by a distribution which makes the E-step expectations easy to compute. This approach was used by Laird [23] and by Stiratelli et al. [56] in estimating **variance components** in a random effects model with binary data. They assumed that the conditional distribution of the missing random effects was approximately normal, given an individual's observed data; this makes it easy to compute the required expectations, but the approximation fails if individuals have only a few data points. More recently, Steele [55] used Laplace's method to obtain an analytic approximation for the E-step which performs well in simulations.

Other ways for dealing with this problem are based on **Monte Carlo methods** or data augmentation. Tanner & Wong's [58] data augmentation method for iteratively computing the entire posterior density of $\boldsymbol{\phi}$ with missing data is similar in spirit to the EM algorithm. Wei & Tanner [62] introduce the Monte Carlo EM (MCEM) for use when the E-step is difficult to implement. If one can generate a sample $\mathbf{x}^1, \ldots, \mathbf{x}^m$ from $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}^{(p)})$, Q can be approximated by

$$Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)}) = \frac{1}{m}\sum_{j=1}^{m} \log f(\mathbf{x}^j|\boldsymbol{\phi}).$$

Although this avoids integrals, one must be able to generate data from $k(\cdot|\cdot)$ and maximizing the Q function may be more difficult because of the mixing. Meng & Schilling [46] describe the use of Gibbs sampling (*see* **Markov Chain Monte Carlo**) to carry out a MCEM for item factor models.

The convergence of the EM can be very slow and several authors have suggested methods for speeding convergence. Louis [37], Laird et al. [24], and Lindstrom & Bates [34] suggested Aitken acceleration; its use has met with mixed success. Jamshidian & Jennrich [20] apply generalized conjugate gradient methods to accelerate the EM and term the resulting algorithm the AEM (accelerated EM). Lange [28] suggests a version of quasi-Newton which uses EM type ideas to approximate the Hessian used by the Newton–Raphson algorithm to maximize the observed data log likelihood. An alternate method [2] is to start with the EM and switch to Newton–Raphson. This is advantageous because just when the EM slows down, near a maximum, the Newton–Raphson works best.

Because the convergence rate of the EM is determined by the amount of missing information, if one can find ways to specify the complete data so as to decrease the amount of "missing data", one should increase the rate of convergence. Meng & van Dyk [42] give a general approach for choosing the complete data to minimize the fraction of missing data with applications to fitting **Student's *t* distribution**, random effects models, and image reconstruction (*see* **Image Analysis and Tomography**). The approach is like the EM itself in that there is only a general methodology and the implementation must be worked out independently for each case.

Besides slow convergence, the absence of second derivative computations means that asymptotic standard errors require additional computation. As noted below, there have been a considerable number of proposals for approximating the asymptotic variance–covariance matrix. However, in practice it is often simplest to calculate the observed information

matrix numerically and invert it. Because derivatives of Q with respect to its left variable coincide with the score, one can compute second derivatives by taking finite differences of the exact first derivatives of Q. This approach will be attractive whenever the first derivative of Q is easily computed. In early work, Hartley & Hocking [18] present a general method for computing the observed information in an incomplete data setting which is based on creating a system of simultaneous equations from successive iterations of the EM algorithm, one for each parameter to be estimated.

Many recent procedures have been proposed for computing the second derivatives of $L(\boldsymbol{\phi})$ which capitalize on the representation of the data as incomplete. By rearranging (12), differentiating twice under the integral signs and evaluating at $\boldsymbol{\phi}' = \boldsymbol{\phi}$, we can derive an expression for the observed information, $\mathbf{I}_{\mathbf{y}}$, as

$$\mathbf{I}_{\mathbf{y}} = \mathbf{I}_{\mathbf{x}c} - \mathbf{F}_{\mathbf{x}|\mathbf{y}}. \tag{18}$$

Here $\mathbf{F}_{\mathbf{x}|\mathbf{y}}$ is the expected or Fisher information in $\mathbf{x}|\mathbf{y}$:

$$\mathbf{F}_{\mathbf{x}|\mathbf{y}} = \mathrm{E}\left[\frac{\partial^2 \log k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^{\mathrm{T}}}\middle|\mathbf{y}, \boldsymbol{\phi}\right],$$

where expectation is with respect to $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$ and $\mathbf{I}_{\mathbf{x}c}$ is the expected value of the information in $\mathbf{x}$, conditioned on $\mathbf{y}$:

$$\mathbf{I}_{\mathbf{x}c} = \mathrm{E}\left[\frac{\partial^2 \log f(\mathbf{x}|\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^{\mathrm{T}}}\middle|\mathbf{y}, \boldsymbol{\phi}\right].$$

Eq. (18) has the appealing interpretation as

observed data information
= complete data information
− missing data information,

which Orchard & Woodbury [47] termed the *missing information principle*. Taking expectations of both sides of (18) with respect to $g(\mathbf{y}|\boldsymbol{\phi})$ yields

$$\mathbf{F}_{\mathbf{y}} = \mathbf{F}_{\mathbf{x}} - \text{ave } \mathbf{F}_{\mathbf{x}|\mathbf{y}},$$

where ave denotes expectation over $\mathbf{g}(\mathbf{y}|\boldsymbol{\phi})$. When the complete data have the exponential family form, expression (18) simplifies to

$$\mathbf{I}_{\mathbf{y}} = \mathrm{var}(\mathbf{t}|\boldsymbol{\phi}) - \mathrm{var}(\mathbf{t}|\boldsymbol{\phi}, \mathbf{y}),$$

and the expected information in $\mathbf{y}$ is

$$\mathbf{F}_{\mathbf{y}} = \mathrm{var}[\mathrm{E}(\mathbf{t}|\mathbf{y}, \boldsymbol{\phi})].$$

Louis [37] shows that $\mathbf{F}_{\mathbf{x}|\mathbf{y}}$ can also be expressed in terms of the derivatives of $\log f(\mathbf{x}|\boldsymbol{\phi})$:

$$\mathbf{F}_{\mathbf{x}|\mathbf{y}} = \mathrm{E}[\mathbf{S}(\mathbf{x}; \boldsymbol{\phi})\mathbf{S}(\mathbf{x}; \boldsymbol{\phi})^{\mathrm{T}}] - [\mathrm{E}\mathbf{S}(\mathbf{x}; \boldsymbol{\phi})][\mathrm{E}\mathbf{S}(\mathbf{x}; \boldsymbol{\phi})]^{\mathrm{T}}, \tag{19}$$

where $\mathbf{S}(\mathbf{x}; \boldsymbol{\phi})$ is the derivative of $\log f(\mathbf{x}|\boldsymbol{\phi})$ and expectations are with respect to $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$. Under regularity conditions on the Q function, it will be maximized by setting its derivatives to zero; hence the second term in (19) vanishes at $\hat{\boldsymbol{\phi}}$. Assuming $\mathbf{x}$ consists of n independent vectors, say $\mathbf{x}_i$, and $\mathbf{y}_i(\mathbf{x}) = \mathbf{y}_i(\mathbf{x}_i)$, a simple expression can be derived for $\mathbf{I}_{\mathbf{y}}$ using expressions for the first and second derivatives of $\log f(\mathbf{x}_i|\boldsymbol{\phi})$ [37]. For this same setting of independent observations, Meilijson [41] and Redner & Walker [50] propose the use of the empirical Fisher information to estimate the asymptotic variance:

$$\hat{\mathbf{F}}_{\mathbf{y}} = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i, \boldsymbol{\phi})\mathbf{s}^{\mathrm{T}}(\mathbf{y}_i, \boldsymbol{\phi}) - \frac{1}{n^2}\mathbf{S}(\mathbf{y}; \boldsymbol{\phi})\mathbf{S}^{\mathrm{T}}(\mathbf{y}; \boldsymbol{\phi}), \tag{20}$$

where $\mathbf{s}(\mathbf{y}_i, \boldsymbol{\phi})$ is the derivative of $\log g(\mathbf{y}_i|\boldsymbol{\phi})$, and $\mathbf{s}(\mathbf{y}_i; \boldsymbol{\phi})$ is the sum of the $\mathbf{s}(\mathbf{y}_i; \boldsymbol{\phi})$, hence the derivative of the observed data log likelihood. At the maximum, $\mathbf{S}(\mathbf{y}; \hat{\boldsymbol{\phi}}) = 0$. As shown in Louis [37],

$$\mathbf{s}(\mathbf{y}_i; \boldsymbol{\phi}) = \mathrm{E}[\mathbf{s}(\mathbf{x}_i; \boldsymbol{\phi})|\mathbf{y}_i, \boldsymbol{\phi}], \tag{21}$$

where $\mathbf{s}_i(\mathbf{x}_i, \boldsymbol{\phi})$ is the derivative of $\log f(\mathbf{x}_i|\boldsymbol{\phi})$. Thus $\mathbf{s}_i(\mathbf{y}_i, \boldsymbol{\phi})$ is available from the E-step computations. Meilijson [41] attributes (21) to Fisher [12]; (20) and (21) combine to give a method for estimating the asymptotic variance with independent data vectors which is very easy to implement in many settings, but is not a fully efficient estimate.

Meng & Rubin [44] propose a method for computing the asymptotic variance which is based on a reexpression of (18) which represents $\mathbf{I}_{\mathbf{y}}$ in terms of $\mathbf{I}_{\mathbf{x}c}$ and the "fraction of missing information" matrix given by

$$\mathbf{DM} = \mathbf{F}_{\mathbf{x}|\mathbf{y}}\mathbf{I}_{\mathbf{x}c}^{-1}.$$

Their method, like Louis's [37], requires the code for the complete data asymptotic variance–covariance matrix. It does not require independently distributed data, and uses numerical differentiation only to approximate the "fraction of missing information" matrix.

Baker [1, 2] has developed a general method for computing the observed information matrix when using the EM with **categorical data**. It is used

when the incomplete data is a vector of cell counts which can be expressed as a linear function of complete data cell counts. The basic approach is to express the vector of expected cell counts in terms of matrix functions; the variance–covariance matrix can be easily obtained as a function of these matrices and the vector of expected cell counts. Baker [1] also gives a review of methods for computing the information matrix with incomplete categorical data.

Applications

Perhaps the most attractive feature of the EM algorithm is the very wide range of problems which can be characterized as incomplete data problems. Meng & Pedlow [43] conducted a bibliographic search of the EM literature from 1977 through 1991. They found over 1000 articles from almost 300 statistical and nonstatistical journals that contain material on the EM. Many of these articles describe applications in medicine, genetics, engineering, psychology, animal breeding and economics, to mention a few. Clearly, the range of application is broad. In their original paper, DLR described applications in missing data, grouping, censoring and truncation, finite mixtures, variance components, hyperparameter estimation, iteratively reweighted least squares (*see* **Generalized Linear Model**) and **factor analysis**. We refer the reader to that paper for detailed discussion of those applications, concentrating here on more recent work and applications in genetics.

Indirect Measurement Problems, Including Image Reconstruction

An application of the EM which has met with much success is its use in image reconstruction problems, including positron emission tomography [31, 52], transmission tomography [32], **stereology** [53], and particle size reconstruction via diffusion batteries [38]. These problems are also sometimes called indirect measurement problems. A related problem is back projection, or using prevalence data to estimate the distribution of incident cases of a disease which has a long and variable incubation period [39, 40] (*see* **Back-calculation**).

The general problem is to estimate a distribution, usually of particle sizes or intensities, when the particles undergo some known thinning process before being observed. The approach is to discretize the original distribution, into bins or pixels, and estimate the density in each bin, say μ_j. The complete data can be represented as independent **Poisson** counts; with complete data the density in each bin is estimated by the count in the bin divided by the total count. Because of the thinning process, the observed data are also Poisson counts, with a mean which can be expressed as an integral, or in the discretized version, as a linear combination of the μ_j and known coefficients which characterize the thinning process. Given the observed data, the μ_j, and the characteristics of the thinning process, it is easy to compute expectations of the complete data and implement the EM. Transmission tomography is somewhat more complex, since here the logarithms of the mean of the observed Poisson counts are linear in the μ_j.

Because the problem can sometimes be formulated as a standard regression problem, except that the responses are counts, and the unknown coefficients are constrained to be positive, ordinary least squares, or nonnegative least squares have been used in some cases [38], but the results are not very satisfactory. The EM is attractive in this setting because the nonnegativity constraints are automatically satisfied (as they would be if complete data were observed), the algorithm can converge to a boundary maximum (some $\hat{\mu}_j = 0$), and is computationally feasible in the tomography applications where the number of parameters range in the thousands. In this case, and others, the maximization is often ill-conditioned, and a penalty function (*see* **Penalized Maximum Likelihood**) or a Bayes **prior** have been used both to enhance the quality of the reconstructed image and eliminate the ill-conditioning [13, 32, 39, 53]. In fact, Lange & Fessler [32] are able to establish global convergence for the EM when the log posterior is maximized.

A related example concerns estimation of the distribution of infectivity (*see* **Infectivity Titration**), as measured by infectious units in a fixed volume of blood, in a sample of AIDS patients [64]. Here, infectivity cannot be directly measured, but must be assessed using blood from uninfected donors and **serial dilution assays**. The complete data are the infectivity levels for each individual and with complete data the empirical cumulative distribution function (cdf) could be used to estimate the distribution nonparametrically. The observed data are a vector of

binary observations for each individual, each component indicating whether or not the individual's blood infected the donor blood at a given concentration. The EM can be used to recover the empirical cdf of infectivity, estimating both the support points and the probability mass at each point. The general method extends easily to other serial dilution assays and for use with parametric assumptions for the underlying distribution.

Molecular Biology

Alignment and restoration of **DNA sequences** are another area where the EM algorithm has proven useful. With the advent of the Human Genome Project and the goal of sequencing the entire human genome, methods for large-scale DNA sequencing are important. Sequencing means determining the values (A, C, G, or T) of a sequence of nucleotides in a strand of DNA. In large-scale DNA sequencing, the goal is to determine the sequence of large segments of DNA made up of many thousands of bases. Because the current technology can only sequence relatively short fragments ($<$1000 bases) using size separation on electrophoretic gels, the process of sequencing involves breaking the original large fragment into multiple overlapping smaller fragments, this process being repeated sequentially. To obtain the DNA sequence of the original segment, one must reconstruct the order of the smaller fragments, using the information in the overlap. Depending upon the strategy used, the location of the fragments may not be known, but must be deduced from determining the overlap with other fragments. Churchill & Waterman [10] and Thorne & Churchill [60] describe EM methods for the restoration of sequences; Lawrence & Reilly [33] and Cardon & Stormo [7] discuss the use of the EM for the related problem of finding common protein binding sites in a series of unaligned fragments. Churchill [9] gives an overview of statistical issues in DNA sequencing.

Genetics

The use of the EM algorithm is quite natural in genetics because the genetic data one would like to use for inference about genetic models and parameters are typically not directly observable, but only indirectly observable by measured traits, referred to as phenotypes. The gene counting algorithm used to estimate allele frequencies is one of the earliest uses of the EM algorithm [8, 54]. These authors also noted its application to other areas of genetics, including **segregation** and **linkage** analysis.

The general idea behind segregation analysis is to test the fit of a specific genetic model to phenotypic data from families. Smith [54] shows how the EM can be used to estimate a 'segregation ratio' and thus test whether it differs from the ratio specified by the genetic model. The specific example he considers is an assumed recessive model for a common trait with disease allele G. Individuals with genotypes GG are affected, and those with Gg or gg are not. We draw random samples of families with one affected parent, one unaffected parent, and at least one affected child. The unaffected parent can be assumed to be Gg, since they are unaffected, but they have an affected child. The segregation ratio is the ratio of the number of affected to unaffected children in families of a given mating type; it should be 1:1 in the given model since a child always gets a G gene from the affected parent, and gets a G gene from the unaffected parent with probability 1/2. The difficulty is that some Gg $\times$ GG matings will have no affected children, hence be absent from our sample (*see* **Ascertainment**). To get a proper test, the segregation ratio must be estimated. Smith [54] does this by treating the "lost" families as missing data and using the EM to estimate the number of unobserved unaffected cases. Weeks & Lange [61] give a general treatment of this problem.

Segregation analysis is somewhat more complicated with quantitative phenotypes, largely because realistic genetic models are more complicated and involve numerous parameters which need to be estimated. Ott [49] shows how the EM may be used to obtain ML estimates of the parameters in **polygenic** and **mixed models**. Ott [48] also shows how the Mendelian transmission probabilities,

$$p_1 = \Pr(\text{child inherits g}|\text{parent gg}) = 1,$$

$$p_2 = \Pr(\text{child inherits g}|\text{parent Gg}) = 1/2,$$

$$p_3 = \Pr(\text{child inherits g}|\text{parent GG}) = 0,$$

can be tested by assuming general values for (p_1, p_2, p_3), writing down the likelihood based on samples of families, and using EM to estimate the transmission probabilities.

Linkage analysis is another area where the EM algorithm is used. In the typical linkage study, one

seeks to locate the position of a disease gene by correlating inheritance of the disease with inheritance of a "marker" whose position in the genome is known, at least approximately. A marker is a segment of DNA whose alleles can be observed directly; usually markers are highly polymorphic, meaning they have multiple alleles with nontrivial frequencies. In the formation of gametes, parental chromosomes may break and recombine, so that the chromosome inherited by a child may consist of a combination of disease and marker alleles that did not exist in the parent; this is called a recombination event. If the disease and marker alleles inherited by the child are identical to those located on one of the parental chromosomes, then no recombination event has occurred. The distance between two locations on a chromosome is measured statistically by the recombination fraction, which gives the probability that a recombination occurs between the marker and the disease gene during the formation of gametes. The larger the recombination fraction, the farther apart the marker and disease gene are. Even if one could directly observe both marker and disease alleles for parents and children, it might still not be possible to tell if a recombination has occurred because one can only infer from genotypes which allele is inherited from which parental chromosome. Thus if there is duplication of allele values for the parental marker and disease genes, there may be several possible transmission patterns (and hence possible recombinations) consistent with a child's genetic values. When one can only observe marker alleles directly, and only disease phenotypes, there are additional missing data with some mating types.

In this setting, the missing data which would make the estimation of the recombination fraction trivial are an indicator for each child's data, telling us whether or not a recombination occurred between the two locations of interest. Thus, unlike gene counting and segregation analysis, the values of the alleles are not used directly; the relevant information is knowing the parental source of DNA at each child's location so that recombination can be determined unambiguously.

In plant and animal genetics, crosses can be arranged between parents with known genetic combinations (called haplotypes), so that recombinants and nonrecombinants can be directly observed from the offspring of phenotypes. Smith [54] shows how the EM can be used with "repulsion single backcross" data, where even though parent haplotype data are known, it is not possible to count recombinants directly. He also developed a general framework for linkage and applied it to estimate the recombination fraction between the genes for color blindness and muscular dystrophy, both conditions assumed to be sex-linked recessives. Ott [48] gives a very general treatment of using the EM in linkage analysis, considering the multiparameter situation where the recombination fraction might depend upon age, sex, etc. and also permitting the estimation of additional parameters in the genetic model.

Thompson [59] discusses the utility of a simultaneous analysis of three or more markers as opposed to considering a series of pairwise analyses. She proposes the use of the EM to estimate the recombination fractions and also Sundberg's [57] formulas for curvature with incomplete data, to compute the information gain in a joint marker linkage analysis. Lander & Green [26] present another approach to estimating recombination fractions between multiple markers using the EM with data from three-generation pedigrees, assuming that the order of the markers is known. Here the complete data are the inheritance vectors at each location, telling which founder alleles are inherited by each person at each marker location. With this information it is easy to tell where recombinations have occurred, and estimate the recombination fractions between each location (M-step). Having the recombination fractions, the observed marker data at each location, and knowing the pedigree relationships, one needs to compute the expected number of recombinations in each interval (E-step). They give three alternative methods for carrying out the E-step which are useful in different settings. Using the EM in this setting permits the simultaneous analysis of many more markers than previously possible, using large numbers of small pedigrees. In related work, Lander & Botstein [25] propose the use of the EM for linkage studies involving multiple markers and quantitative outcomes. Guo & Thompson [15] suggest using a Monte Carlo EM for a combined linkage and segregation analysis, again with quantitative outcomes.

An alternate method for testing linkage is to count the number of alleles which are identical-by-descent (IBD) in pairs of affected relatives, and compare the observed to that expected based on the relationship alone. An allele shared IBD in two relatives means they share two copies of a single

gene by inheritance; for example a parent and his or her offspring share exactly one allele IBD at each location in the genome provided there is no inbreeding. They might also share other alleles by chance, called IBS for identity-by-state, if parents share the same type of alleles at some locations. It is not always possible to determine IBD status, either because of missing relative information or duplicate versions of the same allele in relatives. To solve this problem, Risch [51] suggested a simple EM algorithm for estimating the proportions shared IBD based on the observed data at a single locus. This approach was extended by Kruglyak & Lander [22] to incorporate information from markers at other locations by calculating the distribution of the inheritance vectors to determine IBD status. They also extend this basic approach to carry out a maximum likelihood analysis of allele sharing when the disease outcome is measured quantitatively. These approaches were extended to incorporate parametric methods of linkage analysis with multiple marker locations by Kruglyak et al. [21]. Other applications of the EM in genetics, and a review, are given in Weeks & Lange [61].

Acknowledgments

The author thanks Stuart Baker and Ken Lange for helpful comments.

References

[1] Baker, S.G. (1992). A simple method for computing the observed information matrix when using the EM algorithm with categorical data, *Journal of Computational and Graphical Statistics* **1**, 63–76.

[2] Baker, S.G. (1994). Composite linear models for incomplete multinomial data, *Statistics in Medicine* **13**, 609–622.

[3] Baker, S.G., Freedman, L.S. & Parmar, M.K.B. (1991). Using replicate observations in observer agreement studies with binary assessments, *Biometrics* **47**, 1327–1338.

[4] Baker, S.G. & Laird, N.M. (1988). Regression analysis for catagorical variables with outcome subject to nonignorable nonresponse, *Journal of the American Statistical Association* **83**, 62–69.

[5] Baum, L.E., Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics* **41**, 164–171.

[6] Boyles, R.A. (1983). On the convergence of the EM algorithm, *Journal of the Royal Statistical Society, Series B* **45**, 47–50.

[7] Cardon, L.R. & Stormo, G.D. (1992). Expectation maximization algorithm for identifying protein binding sites with variable lengths from unaligned DNA fragments, *Journal of Molecular Biology* **223**, 159–170.

[8] Ceppellini, R., Siniscalco, M. & Smith, C.A.B. (1955). The estimation of gene frequencies in a random-mating population, *Annals of Human Genetics* **20**, 97–115.

[9] Churchill, G.A. (1995). Accurate restoration of DNA sequences, in *Case Studies in Bayesian Statistics*, Vol. 2, C. Gatsaris, J.S. Hodges, R.E. Kass & N.D. Singpurwalla eds. Springer-Verlag, New York, pp. 90–148.

[10] Churchill, G.A. & Waterman, M.S. (1992). The accuracy of DNA sequences: estimating sequence quality, *Genomics* **14**, 89–98.

[11] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[12] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

[13] Geman, S. & McClure, D. (1985). Bayesian image analysis: an application to single photon emission tomography, in *American Statistical Association 1985 Proceedings of the Section on Statistical Computing*. American Statistical Association, Alexandria, pp. 12–18.

[14] Green, P.J. (1990). On use of the EM algorithm for penalized likelihood estimation, *Journal of the Royal Statistical Society, Series B*, **52**, 443–452.

[15] Guo, S.W. & Thompson, E.A. (1992). A Monte Carlo method for combined segregation and linkage analysis, *American Journal of Human Genetics*, **51**, 1111–1126.

[16] Hartl, D.L. & Clark, A.G. (1989). *Principles of Population Genetics*. Sinauer Associates, Sunderland.

[17] Hartley, H.O. (1958). Maximum likelihood estimation from incomplete data, *Biometrics* **14**, 174–194.

[18] Hartley, H.O. & Hocking, R.R. (1971). The analysis of incomplete data, *Biometrics* **27**, 783–808.

[19] Healy, M. & Westmacott, M. (1956). Missing values in experiments analysed on automatic computers, *Applied Statistics* **5**, 203–206.

[20] Jamshidian, M. & Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm, *Journal of the American Statistical Association* **88**, 221–228.

[21] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.

[22] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.

[23] Laird, N.M. (1978). Empirical Bayes methods for two-way tables, *Biometrika* **65**, 581–590.

[24] Laird, N.M., Lange, N. & Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm, *Journal of the American Statistical Association* **82**, 97–105.

[25] Lander, E.S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* **121**, 185–199.

[26] Lander, E.S. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans, *Proceedings of the National Academy of Science, USA* **84**, 2363–2367.

[27] Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm, *Journal of the Royal Statistical Society, Series B* **57**, 425–437.

[28] Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm, *Statistica Sinica* **5**, 1–18.

[29] Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Springer-Verlag, New York.

[30] Lange, K. (1998). *Numerical Analysis for Statisticians*, to appear.

[31] Lange, K. & Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography, *Journal of Computer Assisted Tomography* **8**, 306–316.

[32] Lange, K. & Fessler, J.A. (1995). Globally convergent algorithms for maximum a posteriori transmission tomography, *IEEE Transactions of Image Processing* **4**, 1430–1438.

[33] Lawrence, C.E. & Reilly, A.A. (1990). An Expectation Maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, *Proteins* **7**, 41.

[34] Lindstrom, M.J. & Bates, D.M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data, *Journal of the American Statistical Association* **83**, 1014–1022.

[35] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[36] Liu, C. & Rubin, D.B. (1995). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence, *Biometrika* **81**, 633–648.

[37] Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society, Series B* **44**, 226–233.

[38] Maher, E. & Laird, N.M. (1985). Reconstruction of particle size distributions from diffusion battery data using the EM algorithm, *Journal of Aerosol Sciences* **16**, 557–570.

[39] Marschner, I.C. (1995). Computation of age-specific HIV incidence estimates using the EM algorithm, *Journal of Statistical Computation and Simulation* **53**, 299–312.

[40] Marschner, I.C. (1996). Fitting a multiplicative incidence model to age- and time-specific prevalence data, *Biometrics* **52**, 492–499.

[41] Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms, *Journal of the Royal Statistical Society, Series B* **51**, 127–138.

[42] Meng, X.-L. & van Dyk, D. (1997). The EM algorithm: an old folk song sung to a fast new tune, *Journal of the Royal Statistical Society, Series B* **59**, 511–568.

[43] Meng, X.-L. & Pedlow, S. (1992). EM: a bibliographic review with missing articles, in *American Statistical Association 1992 Proceedings of the Section on Statistical Computing*. American Statistical Association, Alexandria, pp. 24–27.

[44] Meng, X.-L. & Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices – the SEM algorithm, *Journal of the American Statistical Association* **86**, 899–909.

[45] Meng, X.-L. & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika* **80**, 267–278.

[46] Meng, X.-L. & Schilling, S. (1996). Fitting full-information item factors models and an empirical investigation of bridge sampling, *Journal of the American Statistical Association* **91**, 1254–1267.

[47] Orchard, T. & Woodbury, M.A. (1972). A missing information principle: theory and applications, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 697–715.

[48] Ott, J. (1977). Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis, *Annals of Human Genetics* **40**, 443–454.

[49] Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees, *American Journal of Human Genetics* **31**, 161–175.

[50] Redner, R.A. & Walker, H.F. (1984). Mixture densities, maximum likelihood, and the EM algorithm, *SIAM Review* **26**, 195–239.

[51] Risch, N. (1990). Linkage strategies for genetically complex traits, III. The effect of marker polymorphism on analysis of affected relative pairs, *American Journal of Human Genetics* **46**, 242–253.

[52] Shepp, L.A. & Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography, *IEEE Transactions of Image Processing* **1**, 113–121.

[53] Silverman, B.W., Jones, M.C., Wilson, J.D. & Nychka, D.W. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography, *Journal of the Royal Statistical Society, Series B* **52**, 271–324.

[54] Smith, C.A.B. (1957). Counting methods in genetical statistics, *Annals of Human Genetics* **21**, 254–276.

[55] Steele, B.M. (1996). A modified EM algorithm for estimation in generalized mixed models, *Biometrics* **52**, 1295–1310.

[56] Stiratelli, R., Laird, N.M. & Ware, J.H. (1985). Random effects models for serial observations with dichotomous response, *Biometrics* **40**, 961–972.

[57] Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family, *Scandinavian Journal of Statistics* **1**, 49–58.

[58] Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association* **82**, 528–550.

[59] Thompson, E.A. (1984). Information gain in joint linkage analysis, *IMA Journal of Mathematics Applied in Medicine and Biology* **1**, 31–49.

[60] Thorne, J.L. & Churchill, G.A. (1995). Estimation and reliability of molecular sequence alignments, *Biometrics* **51**, 100–113.

[61] Weeks, D.E. & Lange, K. (1989). Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis, *IMA Journal of Mathematics Applied in Medicine and Biology* **6**, 209–232.

[62] Wei, G.C.C. & Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association* **85**, 699–704.

[63] Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Annals of Statistics* **11**, 95–103.

[64] Zackin, R., De Gruttola, V. & Laird, N.M. (1996). Mixed effects models for estimating the effect of antiviral therapy on the burden of the human immunodeficiency virus, *Journal of the American Statistical Association* **91**, 52–61.

NAN M. LAIRD

Empirical Bayes

"Empirical Bayes" is the term Herbert Robbins coined in his 1955 paper [26] to describe how, if certain **estimation** situations are encountered many times, the data from all these situations can be used in combination to construct estimates for each individual case that approach the greater accuracy of a Bayes estimator (*see* **Bayesian Methods**). This is without knowing the Bayesian's **prior distribution**. Since then, the concept of empirical Bayes has been widened periodically so that today the term is used by authors in biostatistics [5] and in other sciences to embrace an array of models and methods that include, for example, **hierarchical models**, **random effects** models, **linear mixed effect models for longitudinal data**, and **multilevel models**. Empirical Bayes derived from the frequency perspective, emphasizing asymptotic evaluations of procedures. Today's **inferences** often are developed by using Bayesian methods, although less emphasis is often placed on evaluating new procedures.

Empirical Bayes research concerns analyses of observed data $\boldsymbol{y}$ that follow a two-level (hierarchical) model with known density function f at level 1:

$$Y|\theta \sim f(y|\theta) \tag{1}$$

Here, θ is a vector of random effects. At level 2, with α, an unknown parameter vector (hyperparameter) that governs the possible distributions of θ, and with G, having density $g_\alpha(\theta)$, modeled as a known (distribution) function

$$\theta \sim g_\alpha(\theta). \tag{2}$$

Robbins' empirical Bayes referred to this framework with θ a k-dimensional vector of unknowns $\theta = (\theta_1, \ldots \theta_k)$ and $\boldsymbol{y} = (y_1, \ldots y_k)$ independent and with k going to infinity. In (2), G corresponds to letting g be a one-dimensional completely unspecified density for θ_1, and then letting all the other parameters $\theta_2 \ldots \theta_k$ be independently identically distributed (i.i.d.) with the same g. In the context of (2), the parameter α is identified with g here, and the key assumption for repeated problems is that among all possible distributions on k-dimensions, the θ are i.i.d.

$$Y_i|\theta_i \sim f(y_i|\theta_i) \text{ indep} \quad i = 1, \ldots, k. \tag{3}$$

$$\theta_i \sim g(\theta_i) \text{ indep} \quad i = 1, \ldots, k. \tag{4}$$

This setup allows for building up of information about this unknown g.

A scientist who knows $g(\,)$ would use it to calculate the Bayes estimate for each θ, for example, the posterior mean

$$\hat{\theta}_{i,g}(y_i) = \mathrm{E}(\theta_i|y_i, g) = \frac{\int \theta f(y_i|\theta) g(\theta)\, \mathrm{d}\theta}{\int f(y_i|\theta) g(\theta)\, \mathrm{d}\theta}. \tag{5}$$

Robbins

Robbins and his direct successors focused on the construction of estimates for each θ_i that asymptotically, as $k \to \infty$, perform as well as the Bayes rule in (5), as summarized in Maritz and Lwin [24]. Robbins [26] made independence assumptions of (3) and (4) in his initial example, with y_i having a **Poisson distribution** with mean θ_i, and θ_i having an unknown distribution, as in (4). Such a model could apply to evaluations of medical units if, for example, y_i were the number of patients who experienced bad outcomes (mortality, or postoperative infection, assuming these are rare events) at treatment center i ($i = 1, \ldots k$), and then θ_i would be the expected number of bad outcomes. Under these assumptions, the marginal distributions of the observed y_i values are i.i.d. and depend on the unknown g (*see* **Marginal Models**). Then, it is easily seen for any component i with outcome Y and $\mathrm{E}(Y) = \theta$ in this Poisson setup, that Bayes' formula gives $\mathrm{E}(\theta|Y = y) = (1 + y) * P(Y = y + 1)/P(Y = y)$. Robbins used this fact and **consistency** of the sample distribution function based on the marginal distribution of $(y_1, \ldots y_k)$ to approximate the Bayes estimate of θ_i in (5) by his "empirical Bayes" estimate

$$\hat{\theta}_i = (1 + y_i)\frac{N_{1+y_i}}{N_{y_i}}. \tag{6}$$

Here, N_{y_i} is the number of units (treatment centers, here) that observed exactly y_i Poisson events, and N_y/k is a consistent estimate of $P(Y = y)$ as $k \to \infty$.

This two-level Poisson model (3) and (4) applies in practice if the expected number of bad outcomes at center i are i.i.d. and the number of patients exposed is the same in every center. (This is rare, in practice.) Robbins' estimate (6) of the Bayes rule is consistent so that, as $k \to \infty$, it improves as an approximation to the Bayes estimate (5) for most centers if k is sufficiently large.

Unfortunately, no matter how many treatment centers, k, are considered, some centers always will be estimated badly by (6). To see this, let $y_{\max}$ denote the maximum number of bad outcomes in any center. This is a finite number for any k (but it goes to infinity as $k \to \infty$). By definition of $y_{\max}$, $N(1 + y_{\max}) = 0$, so (6) estimates the corresponding θ_i to be 0. We see that the treatment center with the worst outcome record is estimated by (6) to be the best!

Better empirical Bayes estimates than (6) were developed by Robbins' immediate successors for this Poisson situation and other distributions, and Maritz' book [24] summarizes many of these investigations. While those advances improved the estimation techniques, they still required $k \to \infty$, and the models considered (3) and (4) often were prohibitively restrictive for real applications, which require much less symmetry at both levels. For example, treatment centers almost always treat different numbers of patients, so the level-1 distribution f in (3) must depend on i. Also, in real problems, the identically distributed assumption for the $\{\theta_i\}$ values in (4) becomes harder to meet, especially if k must be large. And, methods were needed that work for small and moderate values of k, for example, for a moderate number of treatment centers.

Parametric Empirical Bayes

Stein first proved in 1955 [27] that the means of k independent **Normal distributions**, in the setup of (3) (but not introducing or considering (4)), could be estimated better, for sums of squared error **loss functions** if k is at least 3. A specific, simple estimator was introduced by James and Stein [18]. Efron and Morris [10–13] added (4) and reinterpreted Stein's estimator as an empirical Bayes estimator of the posterior mean of θ_i, given the data. They allowed for different distributions f in (3), replacing f by f_i, so that sample sizes (or "exposures", in the Poisson case) could differ across units. Since the level-2 distributions were Normal, for example, although with unknown moments, this was a "parametric empirical Bayes". In this Normal case, (3) and (4) become

$$Y_i|\theta_i \sim N\left(\theta_i, V_i = \frac{\sigma^2}{n_i}\right) \text{ indep } i = 1, \ldots, k \quad (7)$$

$$\theta_i \sim N(\mu_i, \tau^2) \text{ indep } i = 1, \ldots, k. \quad (8)$$

Then the posterior distribution of θ_i, assuming $\alpha = (\beta_0, \beta_1, \tau)$ in this case, is known, with $B_i = V_i/(\tau^2 + V_i)$, $V_i \equiv \sigma^2/n_i$, takes the form

$$\mathrm{E}\theta_i = \mu_i = \beta_0 + \beta_1 x_i, \quad (9)$$

with x_i a **covariate** (more generally x_i could be a vector).

The y_i's, are observed, and the n_i values are known. Usually σ, needed in V_i, is known or is so accurately estimated that it can be assumed known. Then V_i is known, and unknown hyperparameters $\alpha = (\beta_0, \beta_1, \tau)$ are the α in (2) and must be estimated (with errors, but that diminish as k increases). The terms μ_i and τ are of central interest in random effects research, but the emphasis in empirical Bayes focuses on making inferences about the many values of θ_i.

Efron and Morris, in the 1970s, [10–13] developed parametric empirical Bayes that were useful in practice for moderate values of k, and with varying V_i values, but these papers were mainly about point estimation of the parameters θ_i. Parametric empirical Bayes estimators took the form, for example, [11]

$$\theta_i|Y_i, \alpha \sim N((1 - B_i)y_i + B_i\mu_i, V_i(1 - B_i)). \quad (10)$$

$$B_i = \frac{V_i}{V_i + \tau^2}, \quad (11)$$

The B_i in (10) are "**shrinkage** factors", in the sense that the estimate $\hat{\theta}_i$ is shrunken away from y_i by the fraction B_i toward the mean μ_i. Thus B_i represents the fraction of **regression toward the mean**, μ_i. Their developments were that **shrinkage estimates** of the θ's emerging from this two-level structure had lower risk (often for mean squared error) for each i. Risks, as a function of α, are computed by averaging over the data and assuming the level-2 distribution (8) holds while not knowing the hyperparameter vector α (= β and τ, here). In keeping with the evaluation requirements of the empirical Bayes perspective, various resulting shrunken estimators were shown to improve uniformly (every component i, all α) on the best procedures that do not combine data for fixed k, including for fairly small k. Other developments included letting each y_i be multivariate, and including regressions at level 2, so that the θ_i need not be identically distributed. That is, the distribution in (8) extends to let μ_i depend on i (so α is $(\tau, \beta_0, \ldots, \beta_r)$, with a constant term and with r observed predictors).

Empirical Bayes estimates of θ_i then typically mimic the posterior mean in (10) using estimates of B_i and μ_i, so that

$$\hat{\theta}_i = (1 - \hat{B}_i)y_i + \hat{B}_i\hat{\mu}_i. \quad (12)$$

Making Full Inferences for Each Unit

Until the 1980s, research from the empirical Bayes perspective said little about interval estimates. Actual applications generally require interval estimates, and empirical Bayes methods would be of limited interest without them. If α in (2) or (8) were known exactly, or if a hyperprior distribution were known for α, then posterior probability (Bayes) intervals provide such intervals.

Large-sample approaches are unreliable for empirical Bayes interval estimates. To illustrate, an often-applied approach to deriving interval estimates in the Normal case has been to find the **maximum likelihood** estimate (MLE) of α (e.g. of μ and τ in (8)) and then to plug those values into (9) to obtain an estimated variance for θ_i. However, the MLE of the shrinkage factor B_i is biased toward overshrinkage, and noticeably so for small values of k and for B_i not near 0. Worse yet, in some data sets (even with large k), the MLE of τ^2 can equal 0. Then, the MLE estimates B_i as 1 and the posterior variance $V_i(1 - B_i)$ in (10) as 0. This is likely if the true τ is near 0, even for large k. Then, the corresponding MLE interval estimates for θ_i are given zero width, and they cannot possibly cover the true value. Despite this, rules based on the MLE or on other "plug-in" estimates for τ used in $V_i(1 - B_i)$ continue to be proposed and used, even though they give intervals that are much too small.

Conditional probability and Bayesian reasoning can guide construction of procedures meant to have good frequency properties. Bayes rules based on an uninformative hyperprior distribution on α show how to account for the added variability due to not knowing the level-2 mean parameters and the variance τ^2 by identifying additional terms needed in the posterior variance of θ_i. That approach was used [25] to construct interval estimates in the Normal case. Standards for empirical Bayes inference are defined in [25], requiring that procedures meet those standards in repeated sampling for every value of the hyperparameter vector, α. One then seeks prior distributions on α that lead to empirical Bayes inferences that also meet these standards. For Normal problems (7) to (12), choosing a **uniform distribution** on τ^2 (and flat distributions on μ, or on β_0, β_1 in (8) or (11)) has produced interval estimates that approximately meet or exceed the nominally claimed coverages. Certain other distributions, including some uniform shrinkage priors [7] also meet these standards. Such intervals then cover all θ_i values appropriately in settings covered by (7) to (12). Evaluations of these procedures, which do not allow knowledge of α, must be checked to hold for every possible value of $\alpha = (\beta, \tau^2)$, and for various values of k, including fairly small k. These empirical Bayes evaluations depend on the distributional assumptions (e.g. Normal, independence) made at both levels of the model. If one does not allow averages over the level-2 distributions of θ_i (for each fixed value of α) then no shrinkage interval estimate can achieve the empirical Bayes nominal coverage standard for every i.

Bayes, Hierarchical Bayes, and Other Approaches

Bayesian ideas for carefully chosen prior distributions on the hyperparameter have been so successful, even in the frequency sense, that in recent years, there have been a variety of such approaches that address the two-level model (1) and (2) in many forms, for example [19], [23]. That includes the approach of Lindley and Smith [21], which deals with models similar to (7) and (8) from the Bayesian perspective. More recently, powerful computation coupled with **Markov chain Monte Carlo** (MCMC) sampling tools has made Bayesian computations feasible. The texts by Carlin and Lewis [5] and by Gelman et al. [15], as well as Greenland's article [16], are excellent sources for these developments. Other frequency approaches include analysis of the two-level model (1) and (2) with random and mixed effect models, repeated measurement models, models for **longitudinal data**, **generalized linear mixed models**, **random coefficient** models, models for latent variables (*see* **Latent Class Analysis**), graphical models, **hidden Markov models**, **variance components**, and so on. Some of these topics are covered in this encyclopedia and probably all could be models for biostatistical applications. The term empirical Bayes may not often appear, even when interest is on estimating multiple values θ_i, but such methods could be considered under this term.

Empirical Bayes Evaluation

A key concern of empirical Bayes has always been the attempt to evaluate the performance of such procedures over a range of incidences. That needs to be done more often when methods are being proposed under these other names. In particular, Bayesian methods depend on choices of a hyperprior distribution on α, and such evaluations can identify which hyperpriors are likely to produce rules that can be broadly applied. Asymptotic methods often are used by frequentists to evaluate estimates, and many methods such as **generalized estimating** equations (GEE), MLE, penalized **quasi-likelihood** (PQL), and **overdispersion** approaches are used.

Such methods, justified for large samples (large k), may or may not work for a small k. For example, let us return to the original empirical Bayes–Poisson setting of Robbins, (3), but now with parametric assumptions for $g(\theta)$ in (4). Christiansen and Morris [7] studied several such methods, including when the hyperparameters α for the prior **gamma distributions** are estimated by MLE or by standard GLM overdispersion methods, and the Poisson parameters θ_i are assumed to follow their estimated posterior distribution (via plugging in $\hat{\alpha}$ for α). When $k = 15$, their examples show that resulting nominal 95% **confidence intervals** cover only about 70% of the time for the MLE, and 80% of the time for GLM. However, accurate interval estimates are derived in the paper [7] that do meet their nominal 95% coverages.

Small-sample empirical Bayes methods can be derived, often drawing on parametric models at level 2, and often relying on Bayesian methods or Bayesian heuristics. With sufficient analysis and **simulation**, some of these can be shown to meet their nominal risk and coverage claims.

Applications: Empirical Bayes and Gene Expression Data

Analyzing DNA microarray data and other gene expression data leads to testing many hypotheses, each meant to measure whether a gene is an effective marker (*see* **DNA Sequences**). There may be thousands of such tests made, one for each gene in the array. That raises the **multiple comparisons** problem, in that many "statistically significant" genes will found, even when none are true markers. Benjamini and Hochberg [3] introduced the false discovery rate (FDR) to help measure and control the seriousness of this problem in such applications. Several authors, including Efron and Tibshirani [9, 14], and Kendziorski, Newton, Lan, and Gould [20] have shown that empirical Bayes, and parametric empirical Bayes modeling and analyses offer a powerful way to clarify and attack these multiplicity issues, and produce methods closely related to the FDR.

To see how these tests are related to empirical Bayes, Efron and Tibshirani suppose there are k statistical tests made independently (*see* **Hypothesis Testing**). More complicated assumptions are considered, but in simpler settings the ith case might be based on a test statistic y_i that is $N(0, 1)$ under the **null hypothesis**, but when the site is a genuine marker, has mean θ_i that differs from 0. Then, a model like (7) and (8) is assumed, but with a distribution function $G(\theta_i)$ that is more general than Normal at level 2:

$$\theta_i \sim G() \text{ indep}, \; i = 1, 2, \ldots, k. \qquad (13)$$

The problem then is to learn about the conditional distribution of each θ_i value, given all the data $(y_1, \ldots, y_k)$. For hypothesis testing, G in (13) would give substantial positive probability p_0 to θ_i being 0. Since the same G is assumed for every gene, the observed marginal distribution of y_i is available to learn about the unknown G. If G were known, then Bayes rule could be used for inferences about θ_i, and, of course, it would depend only on y_i. That is, the role of the other $k - 1$ values y_j is only to help learn about G. Efron and Tibshirani study this model and its inferences without making parametric assumptions about G.

In a related study of breast cancer data, Kendziorski et al. [20] observe positive intensities y_i, and so their parametric empirical Bayes model starts with Gamma distributions at level 1 and then specifies conjugate Gamma distributions at level 2. Another model in [20] considers **lognormal** distributions at level 1, in which case, their log transformations and Normal distributions at level 2 provide a structure much like (7) and (8). For each gene, results are the posterior odds that different rat strains have differential expressions for that gene. The authors use simulations to verify that their procedures are reliable for the empirical Bayes model assumed.

Empirical Bayes applications abound in the literature. Some examples appear in [1, 2, 4, 6, 8, 17, 22, 28].

References

[1] Beckett, L.A. & Tancredi, D.J. (2000). Parametric empirical Bayes estimates of disease prevalence using stratified samples from community populations, *Statistics in Medicine* **19**(5), 681–695.

[2] Bedrick, E.J. & Hill, J.R. (1999). Properties and applications of the generalized likelihood as a summary function for prediction problems, *Scandinavian Journal of Statistics* **26**, 593–609.

[3] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B, Methodological* **57**, 289–300.

[4] Burridge, J. (1981). Empirical Bayes analysis of survival time data, *Journal of the Royal Statistical Society, Series B, Methodological* **43**, 65–75.

[5] Carlin, B.P. & Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, Boca Raton, FL.

[6] Chattopadhyay, M., Lahiri, P., Larsen, M. & Reimnitz, J. (1999). Composite estimation of drug prevalence for sub-state areas, *Survey Methodology* **25**, 81–86.

[7] Christiansen, C.L. & Morris, C.N. (1997). Hierarchical Poisson regression modeling, *Journal of the American Statistical Association* **92**, 618–632.

[8] Dagne, G.A., Howe, G.W., Brown, C.H. & Muthen, B.O. (2002). Hierarchical modeling of sequential behavioral data: an empirical Bayesian approach, *Psychological Methods* **7**(2), 262–280.

[9] Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* **96**, 1151–1160.

[10] Efron, B. & Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach, *Journal of the American Statistical Association* **68**, 117–130.

[11] Efron, B. & Morris, C. (1975). Data analysis using Stein's estimator and its generalization, *Journal of the American Statistical Association* **70**, 311–319.

[12] Efron, B. & Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators – Part II: the empirical Bayes case, *Journal of the American Statistical Association* **67**, 130–139.

[13] Efron, B. & Morris, C. (1977). Stein's paradox in statistics, *Scientific American* **236**(5), 119–127.

[14] Efron, B. & Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays, *Genetic Epidemiology* **23**(1), 70–86.

[15] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2003). *Bayesian Data Analysis*, 2nd Ed. Chapman & Hall, New York.

[16] Greenland, S. (2000). Principles of multilevel modelling [review], *International Journal of Epidemiology* **29**(1), 158–167.

[17] Hoef, J.M.V. (1996). Parametric empirical Bayes methods for ecological applications, *Ecological Applications* **6**, 1047–1055.

[18] James, W. & Stein, C. (1961). Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1*, University of California Press, Berkeley, pp. 361–379.

[19] Kass, R.E. & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models), *Journal of the American Statistical Association* **84**, 717–726.

[20] Kendziorski, C.M., Newton, M.A., Lan, H. & Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles, *Statistics in Medicine* **22**, 3899–3914.

[21] Lindley, D.V. & Smith, A.F.M. (1972). Bayes estimates for the linear model, *Journal of the Royal Statistical Society. Series B, Methodological* **34**(1), 1–41.

[22] Louis, T.A. & Shen, W. (1999). Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks, *Statistics in Medicine* **18**, 2493–2505.

[23] Lu, W. (1999). The efficiency of the method of moments estimates for hyperparameters in the empirical Bayes binomial model, *Computational Statistics* **14**, 263–276.

[24] Maritz, J.S. & Lwin, T. (1989). *Empirical Bayes Methods*, 2nd Ed. Chapman & Hall, London.

[25] Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications, *Journal of the American Statistical Association* **78**, 47–55.

[26] Robbins, H. (1955). An empirical Bayes approach to statistics, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, Calif, pp. 157–163.

[27] Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 197–206.

[28] Stern, H.S. & Cressie, N. (2000). Posterior predictive model checks for disease mapping models, *Statistics in Medicine* **19**(17–18), 2377–2397.

CARL N. MORRIS & CINDY L. CHRISTIANSEN

Endocrinology

There are two systems that we commonly envisage to control the functioning of animals: the nervous system, and the endocrine, or hormonal, system. A simple model would view the endocrine system as consisting of several well-defined ductless glands (pituitary, thyroid, parathyroid, adrenals, gonads, and pancreas) that secrete chemical messengers, called hormones, directly into the blood stream. The hormones then travel, via the circulatory system, to certain target cells that then respond in specific ways. By contrast, the faster-acting nervous system functions by transmitting electrical impulses down specially constructed nerve cells.

We can, then, define the subject of endocrinology as the study of all aspects of the endocrine system. These include: the physiology of the glands; the action and effect of the hormones; and the genetic sequencing (*see* **DNA Sequences**) of the **genes** responsible for the hormones. Since many metabolic diseases and impairments result from some deficiency in the endocrine system, endocrinology has always been considered under the remit of clinical medicine, and, to a large extent, still is. However, endocrinology is now beginning increasingly to come under the remit of molecular biology (*see* **Molecular Epidemiology**).

Most of the "true" endocrine glands were discovered in antiquity by such early workers as Aristotle and Galen. The last endocrine gland discovered was the parathyroid in 1891 by Gley. The anatomy of "true" endocrine glands has, therefore, essentially long been completed. The bulk of research is currently focused on trying to determine the effect and method of action of hormones, and optimal therapeutic treatments.

The first real scientific experiment was conducted by Berthold in the mid-nineteenth century. Berthold demonstrated that by transplanting testes from roosters into previously castrated roosters (heterotransplantation) he could maintain the male characteristic of these roosters; in particular, crowing. Yet, even to this day, this type of *in vivo* experiment, where a gland is surgically removed (a process known as ablation), forms the basis of much physiological endocrine research. An excellent account of the history of endocrinology can be found in [6].

One key feature often encountered in endocrinology is the negative feedback mechanism. Here, high levels of a particular hormone within the circulatory system inhibit secretion of more hormone. This usually results in a **dose–response** curve that peaks at a single maximum, and then declines with additional stimulation [1]. It should also be noted that any individual hormone may have a variety of effects.

All of this makes for a complex picture of the endocrine system. This is even more the case when it is considered that many of the current models are very experimental [7] and current research is constantly updating them. Fortunately, the system is not as difficult to control as it might first appear, since there is often a natural hierarchy of hormones within any endocrine system.

Experimental Methods Used

The experimental process in endocrinology typically follows a step-by-step methodology. Usually, different teams of researchers work on each step in the research process. The entire process is typically motivated from a clinical aspect, through a metabolic disease, with a measurable response.

The first step in endocrinology is to isolate the gland responsible. This is usually done through ablation. Next, it is necessary to demonstrate that this is the gland *directly* responsible for the metabolic response of interest. This is usually done using *in vitro* experimental techniques, or transplanting the gland to another site in the body (homotransplantation). The third stage is then to isolate the hormone in question. This is usually demonstrated by ablation of the gland, and hormone replacement therapy to remove the clinical symptoms. Typically, a dose−response curve is constructed at this point. Having isolated the hormone, molecular biology techniques are then used to sequence the DNA responsible for the hormone, with the aim of using recombinant techniques to synthesize it. Since the use of such techniques is relatively new, with the hormone insulin being the first such product [4], well over half of the published work in endocrinology in the last decade has taken place in this area. Finally, therapeutic trials (*see* **Clinical Trials, Overview**) are undertaken to discover the optimal treatment regimes for those who suffer from the metabolic condition.

There is a great deal of literature on therapeutic trials, mainly within the medical journals. This is

because the selection of a particular treatment regime usually needs very careful consideration. The reason for this difficulty lies in medicine's inability to match exactly the body's own homeostatic regulation system. So that, although taking one tablet per day of hormone will give a correct **mean** normal serum level of hormone, it will lead to wide fluctuations in hormone level, which can cause complications. This is particularly important as a single hormone can have several functions. For example, oestrogen replacement in post-menopausal women has been linked to an increased risk of breast cancer. Therefore, determining the correct therapeutic regime can require much investigation, often over many years. Finally, the treatment regime must be tailored to the individual. For example, in diabetes it is necessary not only to tailor treatment, with insulin, around the physical characteristics of the patient, but also around their lifestyle.

The consequences of such a step-by-step approach are that at each stage the **experimental design** rests on the conclusions of the previous stage. However, these assumptions must be interpreted from a variety of sources. Endocrinology is, therefore, clearly an area of research requiring good review articles, and many endocrinology journals do include a forum for such articles; for example, Campbell & Scanes [2]. Yet such review articles rarely employ a systematic review, or the use of meta-analysis techniques (*see* **Meta-analysis of Clinical Trials**).

Statistical Methods Used

Much endocrinology is investigative and, because of this, some researchers, particularly in the past, have employed only descriptive statistics. Yet most of the main hormones are now probably known and so the remaining investigations need to be increasingly sophisticated. Therefore, most studies are now accompanied by some form of **hypothesis testing**, or **confidence intervals** and, indeed, this is often now expected by the relevant subject journals; for example, the *Journal of Endocrinology*. However, in general, the coverage of statistics within endocrinology journals involves routine techniques.

The analysis of an ablation experiment is usually fairly simple and typically uses routine methods. A more complex experiment may involve monitoring the effect of hormone replacement, or ablation, over a time course. This is becoming increasingly popular with the advent of slow-release hormonal implants and more sophisticated monitoring systems. The analysis involved for this type of experiment is almost without exception carried out as a **multivariate analysis of variance** repeated measures analysis (*see* **Longitudinal Data Analysis, Overview**).

Another type of analysis often conducted is the calculation of a dose–response curve. It is often difficult to fit a prior functional form to the dose–response curve, since the particular characteristics may vary considerably, and, indeed, the appropriate dose range may be known only very vaguely. For example, in two articles on diabetes research, two growth curves were postulated. The first article, [8] plotted empirically the means of the data, with confidence intervals, and suggested a bi-modal response resulting from a complex negative feedback mechanism. The second [5] postulated that there was no negative feedback mechanism involved and fitted parametric hyperbolic responses to the data (*see* **Nonlinear Growth Curve**).

Future Developments

The use of modeling (*see* **Model, Choice of**), or **multivariate multiple regression** techniques should, and almost certainly will, become more popular because the administration of complex schemes of hormone replacement and monitoring is now beginning to become practical. This is necessary if for no other reason than that the large amount of data generated will be difficult to handle using the current visual inspection techniques. In addition, such monitoring techniques will allow for the adjustment of responses by **covariates**, which is rarely done at present.

Once a fairly complete understanding has been gained of the entire system being studied it should then be possible to use the techniques of **pharmacokinetics**, such as compartmental modeling, to try to gain some deeper model-based description of the system. Unfortunately, at present the best understood endocrine system is the glucose regulation system. However, even here knowledge is not yet sufficient to undertake such an analysis competently, although such analyses have been attempted. Yet, as endocrinology progresses, a more detailed knowledge of the endocrine system can only go toward providing a higher confidence in such models, and increase their usage.

For the future, the development of mainstream techniques, hopefully incorporated within statistical **software**, for the analysis of repeated measures categorical data and ranked data (*see* **Ranks**) is one area of applied, and theoretical, statistical research that would benefit researchers in this field. This is because there are many measurements of disease that can only be measured somewhat indirectly; in particular, behavioral response. Such a response can often only be based on a ranked, or an **ordered categorical** scale. For example, if modeling dominance or aggression, which are known responses of androgen hormones, a ranking procedure, or ordered scale, may be the only feasible outcome measure [3].

References

[1] Binkley, S.A. (1995). *Endocrinology*. Harper Collins, New York.

[2] Campbell, R.M. & Scanes, C.G. (1995). Endocrine peptides "moonlighting" as immune modulators: roles for somatostatin and GH-releasing factor, *Endocrinology* **147**, 383–396.

[3] Cashdan, E. (1995). Hormones, sex and status in women, *Hormone Behaviour* **29**, 354–366.

[4] Itakura, K., Hirose, T., Crea, R., Riggs, A., Heyneker, H., Bolivar, F. & Boyer, H. (1977). Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin, *Science* **198**, 1056–1062.

[5] Kahn, S.E., Prigeon, R.L., McCulloch, D.K., Boyko, E.J., Bergman, R.N., Schwartz, M.W. & Neifing, J.L. (1993). Quantification of the relationship between insulin sensitivity and beta-cell function in human subjects. Evidence for a hyperbolic function, *Diabetes* **42**, 1663–1672.

[6] McCann, S.M. (1988). *Endocrinology: People and Ideas*. American Physiological Society, Maryland.

[7] McLachlan, R.I., Wreford, N.G., O'Donnel, L., de Kretser, D.M. & Robertson, D.M. (1996). The endocrine regulation of spermatogenesis: independent roles for testosterone and FSH, *Endocrinology* **148**, 1–9.

[8] Teruya, M., Takei, S., Forrest, L.E., Grunewald, A., Chan, E.C. & Charles, M.A. (1993). Pancreatic islet function in nondiabetic and diabetic BB rats, *Diabetes* **42**, 1310–1317.

P. YOUNG

Environmental Epidemiology

Environmental epidemiology encompasses a wide array of topics related to the study and evaluation of the determinants of diseases in human populations. The term "environmental" as used here is general in scope, and pertains to all aspects of our environment that may influence disease **risk**. Environmental exposures include substances that might affect our immediate surroundings, such as pollutants in the air we breathe or water we drink, as well as factors related to our occupation, recreation, and lifestyle that might influence probabilities of disease occurrence. Thus, for example, the dominant environmental determinant of lung cancer in nearly all populations around the world is cigarette smoking, while diet and nutrition contribute to risk of several other cancers and cardiovascular and other diseases. Since molecular and genetic traits may predispose individuals to the adverse effects of exposure to specific environmental factors, environmental epidemiology can be considered to include the entire spectrum of research into the etiology and prevention of human diseases.

In this article we describe methods for and examples of epidemiologic studies of the environmental determinants of chronic diseases, the major killers in human populations today. Emphasis is placed upon epidemiologic and statistical methodologic tools for the detection and evaluation of risks associated with environmental exposures. Acute diseases are not considered, although some of the basic techniques for assessing environmental factors have arisen from principles developed in tracking outbreaks of infectious diseases.

The article is divided according to method of study of environmental factors in disease risk. First are **descriptive epidemiologic** studies. These investigations, which study patterns of disease in general populations and their **correlations** with environmental indices of the populations, are useful primarily for generating hypotheses about disease etiology. More important are **case–control** and **cohort studies**, which are **analytic epidemiologic** studies that evaluate risk of disease in individuals characterized by presence and level of exposure to environmental variables of interest. These **observational** (nonexperimental) epidemiologic studies form the basis for most of what is known about the causes of chronic diseases. Finally, are randomized trials (*see* **Clinical Trials, Overview**), whereby agents or procedures that are thought to have potential for reduction in disease risk are evaluated experimentally with random assignment of individuals to various exposure groups (*see* **Randomization**; **Randomized Treatment Assignment**). The credibility of evidence from these clinical intervention trials is usually higher than from observational studies, but for practical reasons only a limited number of such trials have been undertaken.

Providing Clues to Environmental Factors

Clues to environmental causes of disease are often uncovered by examining patterns of disease mortality and incidence. For extremely rare diseases, even occurrence of a few cases within a short period of time and at a particular space can raise suspicion. Such "clusters" of disease (*see* **Clustering**), often detected by alert clinicians, can in some situations lead to the eventual discovery of the causal agents. For instance, the development of hepatic angiosarcoma among three workers in a single manufacturing plant in Kentucky led to the identification of vinyl chloride as the likely causal agent [24], and the observation of vaginal adenocarcinoma in several young women in Boston was traced to synthetic estrogens taken by their mothers during pregnancy [38]. The large majority of clusters of a few cases of a disease, however, have proven to be uninformative with respect to discovering or evaluating a causal agent. Heath [36], for example, described investigations by the **Centers for Disease Control** of clusters of childhood cancer in a number of communities in the US, none of which conclusively linked the leukemia or other cancer cases with environmental exposures. **Leukemia clusters** around nuclear power facilities have also failed to be causally related to **radiation** exposures from the plants [46]. Similarly, clusters of birth defects and other abnormalities have been assessed among residents near hazardous waste sites with potential for exposure to solvents, metals, and other compounds, but firm conclusions have been difficult to achieve [55, 65]. Although some clusters may be due to environmental determinants, it seems likely that many apparent clusters have resulted from **selection bias**, limitations of geographic boundaries or time periods, or chance [46, 56].

The play of chance is sometimes underestimated in assessment of clusters of small numbers of disease events, especially when the clusters occur in one of many arbitrarily and narrowly defined space-time units. Occurrence of a cluster of two events when 0.2 are "expected", for example, can result in a "significant" ($P = 0.02$ for a one-sided test under **Poisson** assumption) excess. If the boundaries of the cluster were drawn specifically to encompass the cases, however, the statistical significance loses its nominal meaning. Furthermore, if the cluster was in one of many time–space units evaluated, the **multiple comparisons** may generate at least one with a "significant" excess. Sometimes investigators are tempted to include the index cluster (that is, the cases that generated the cluster) in determining whether the disease excess occurs in other time–space units, but this fundamental violation of principles of independence invalidates such an evaluation. Finally, the existence of a cluster *per se* of a small number of disease events in a time–space unit conveys no information about the causes of the cluster.

Geographic clustering can also occur for more common chronic diseases, and be based on fairly large numbers of disease events. Excess occurrences of more common diseases are not so obvious to the practicing physician, but broad clustering may be uncovered through a systematic monitoring of disease morbidity and mortality (*see* **Surveillance of Diseases**). These clusters, typically based on large enough numbers of cases for the calculation of stable rates across time and space, may more often prove to be useful in generating productive leads to disease **causation**. Primary examples are the clusterings of high rates of certain diseases in contiguous areas seen in national atlases, which depict the distribution of mortality rates across small geographic units, such as counties in the US [47, 62] (*see* **Geographic Epidemiology**; **Mapping Disease Patterns**).

Geographic variation suggestive of environmental determinants can be particularly useful for leads to cancer studies. The US cancer maps have shown distinctive patterns of clearly nonrandom distributions of various cancers [47, 62]. Sharply elevated rates of oral cancer mortality among women, for example, have clustered in the southeastern part of the country. The finding led to several hypotheses, including one concerning occupational exposures in the textile industry, an industry employing large numbers of southern women, and one concerning smokeless tobacco, used by some women in rural areas of the South [8]. Subsequent analytic epidemiologic studies generated by the patterns seen in the cancer maps identified the use of oral snuff as the key risk factor and the cause of the large majority of cheek and gum cancers, tumors occurring where the tobacco powder was typically placed [70].

Mortality records have generally been the primary source of health data for generating clues to environmental factors for chronic diseases. In most countries of the world, systematic recording of all deaths is conducted by local governments (*see* **Death Certification**) and used for the compilation of national death rates by **causes of death**. Using population estimates generated from the national **censuses** as denominators, mortality **rates** by age, sex, and race can be computed for deaths due to various causes across time and for various geographic units. The ascertainment of deaths is nearly 100% complete in most populations, but inaccurate determination of the cause of death and certain other limitations may affect routinely collected mortality data. Cancer deaths are generally properly identified on death certificates, although the accuracy and completeness in recording cancer deaths vary by the type of cancer [59], but **misclassification** can be problematic for some other causes of death. Changes in recording practices and the coding of cause of death (*see* **International Classification of Diseases (ICD)**) also may contribute to the apparent changes in secular trends of cause-specific mortality [33]. Mortality data are of limited use for diseases with low fatality, and disease-specific death rates can be influenced by nonenvironmental factors such as improved survival due to changes in treatment modalities or early detection (*see* **Vital Statistics, Overview**).

Some limitations that may affect mortality patterns can be circumvented by using incidence data. Registries sometimes exist for various diseases, most notably cancer, with population-based registries in many parts of the world [58] (*see* **Disease Registers**). In the US, registries participating in the Surveillance, Epidemiology, and End Results (SEER) Program supported by the National Cancer Institute have been collecting diagnostic, treatment, and survival information on newly diagnosed cancer cases in about 10% of the population across the country since 1973 [43]. These data have been used for monitoring cancer incidence trends and patterns of cancer occurrence by demographic and geographic subgroups [27]. For

instance, it has been observed recently that adenocarcinomas of the esophagus and gastric cardia are among the cancers with the most rapid rise in incidence during the past two decades in the US, particularly among white men [12]. This striking trend has led to the generation of multiple hypotheses about environmental factors. One hypothesis is that the increasing use of exposures that promote reflux, particularly obesity and pharmaceutical agents that relax the lower esophageal sphincter, may contribute to the rising incidence trends [20, 69].

Registries of incident cases of nonmalignant diseases are more limited in number and tend not be national in scope. If the population base from which the cases arise is well defined, however, these too can provide the basis for the calculation of rates and trends which may trigger hypotheses about environmental causes. Thus, for example, in Sweden and Denmark, all hospitalizations are registered, so that national patterns of various diseases requiring hospitalizations can be routinely monitored [1]. Illnesses not resulting in hospitalization will be missed, but the incidence and **prevalence** of serious conditions can be ascertained for the entire country.

Information on births is usually routinely collected in populations throughout the world. In addition to sociodemographic variables such as maternal age, race and occupation, information on **birthweight**, Apgar score and method of delivery is often collected. In the US birth certificates were standardized nationally in 1989 to include a checkbox for congenital abnormalities and medical risk factors, including tobacco and alcohol use, medical history, and prenatal care [19, 71]. The birth certificate data therefore can be used not only for monitoring disease occurrence among newborns [49], but also as a research tool to identify potential risk factors for these newborn illnesses, such as maternal sociodemographic characteristics in relation to congenital syphilis [26] and maternal smoking, ethnicity, and birthweight in relation to sudden infant death syndrome [45].

Systematically ascertained data on exposure to environmental agents are less readily available than data on measures of disease mortality or incidence. Thus, for example, the prevalence of cigarette smoking – the single most important environmental cause of disease in the US – is not known for counties across the country. National probability sample surveys, such as the National Health Interview Surveys conducted periodically beginning in 1960 and the series of National Health and Nutrition Examination Surveys (NHANES) starting from NHANESI in 1971–75 to the NHANESIII in 1988–94 have estimated smoking prevalences by broad, but not small-area, geographic regions. These **National Center for Health Statistics (NCHS)** surveys can be useful for monitoring national estimates of prevalence of a number of environmental exposures classified by demographic subgroups and over time [23, 42, 63]. The exposures include not only tobacco consumption, but also diet and nutrition, medical variables, occupation, and other characteristics measured over time and by geographic areas.

Census data often provide a rich source of exposure data for generating clues to environmental risk factors. In addition to population counts by age, sex, and race, the census yields information on income, education, urbanization, occupation, and other factors for various geographic units, the smallest in the US being census tracts and postal zip codes. Usually this information is provided every 10 years. In some countries special censuses of manufacturing provide detailed industrial data at the small-area level, enabling calculation of indices of the percent of the population employed in hundreds of industrial categories. Although small-area data are generally not routinely available on average levels of general population exposures to chemical or physical agents, some registries of environmental exposures (e.g. to radon) exist in selected areas. In Sweden, a number of registries of exposure to chemical substances have been established for **record linkage** with national registries of cancer and mortality [1].

Statistical analyses of correlations of mortality or other aggregate health data with measures of average environmental exposures for the populations can help generate and refine hypotheses about disease causation. The correlations can assess not only concordance across geographic areas but also across time, and can be useful in helping to refute as well as refine hypotheses. Thus, rising then recently declining trends in lung cancer mortality correlate well with the rise and decline in smoking prevalence among American men [11, 31]. However, mortality rates of all cancers combined in the US have been relatively steady since the 1930s once lung cancer is removed [3, 21, 43], a pattern not consistent with the theory that increases in environmental exposures from pesticides or other chemicals are causing large-scale

increases in cancer. Rising incidence of breast and prostate cancers, the major cancers among women and men, respectively, has been reported recently, but the increases seem related to changes in diagnostic techniques rather than environmental influences [27]. Furthermore, complete explanations for the rising incidence of non-Hodgkin's lymphoma are not clear. Geographic correlations have indicated that rates of lymphoma were highest in the north central part of the country, and have led to case–control studies evaluating workplace and environmental exposures associated with farming and other occupations [72] (*see* **Occupational Epidemiology**). Similarly, rates of lung cancer in US counties among workers employed in the chemical, petroleum, paper/pulp, and shipbuilding industries, adjusted for urbanization and other demographic factors (but not cigarette smoking), suggested that exposures associated with these industries may be contributing to lung cancer in the affected counties [9]. As described later, some of these hypotheses have been confirmed and some dismissed through subsequent analytic epidemiologic studies.

One of the reasons the descriptive studies linking rates of disease and exposure for groups may generate spurious leads is that correlation can occur among group averages in the absence of associations between disease and exposure at the individual level, the so-called "**ecologic fallacy**" (*see* **Ecologic Study**). Thus, studies of individual patients and their exposures typically are required to evaluate adequately hypotheses about the environmental determinants of disease. Such investigations are described in the next section.

Testing Hypotheses About Environmental Risk Factors

There are many environmental substances for which sufficient exposure is known to increase risk of certain chronic diseases. Foremost is cigarette smoking, believed to result in premature death in nearly half of all individuals who smoke and to account for more than one in every six deaths in the US [68] (*see* **Smoking and Health**). Table 1 lists a number of substances besides cigarette smoke which have been classified as causes of cancer in humans [16, 40]. Many are drugs used to treat some cancers that can subsequently increase risk of other cancers,

Table 1 Agents classified by the International Agency for Research on Cancer as carcinogenic to humans

Aflatoxins
Aluminum production
4-Aminobiphenyl
Analgesic mixtures containing phenacetin
Arsenic and arsenic compounds
Asbestos
Auramine, manufacture of
Azathioprine
Benzene
Benzidine
Beryllium and beryllium compounds
Betel quid with tobacco
N, *N*-Bis(2-chloroethyl)-2-naphthylamine (Chlornaphazine) Bis(chloromethyl)ether and chloromethyl methyl ether (technical-grade)
Boot and shoe manufacture and repair
1, 4-Butanediol dimethanesulphonate (Myleran)
Cadmium and cadmium compounds
Chlorambucil
1-(2-Chloroethyl)-3-(4-methylcyclohexyl)-1-nitrosourea (Methyl-CCNU)
Chromium compounds, hexavalent
Chronic infection with hepatitis B virus
Chronic infection with hepatitis C virus
Coal gasification
Coal-tar pitches
Coal-tars
Coke production
Cyclophosphamide
Diethylstilbestrol
Erionite
Estrogen replacement therapy
Estrogens, nonsteroidal
Estrogens, steroidal
Ethylene oxide
Furniture and cabinet making Hematite mining, underground, with exposure to radon
Human papilloma virus
Infection with schistosoma hematobium
Iron and steel founding Isopropyl alcohol manufacture, strong-acid process
Magenta, manufacture of
Melphalan 8-Methoxypsoralen (Methoxsalen) plus ultraviolet radiation
Mineral oils, untreated and mildly treated MOPP (combined therapy with nitrogen mustard, vincristine, procarbazine, and prednisone) and other combined chemotherapy including alkylating agents
Mustard gas (sulfur mustard)
2-Naphthylamine
Nickel and nickel compounds
Oral contraceptives, combined
Oral contraceptives, sequential

Table 1 (*continued*)

Radon
Rubber industry
Shale-oils
Solar radiation
Soots Strong inorganic acid mists containing sulfuric acid
Talc containing asbestiform fibers
Tobacco products, smokeless
Tobacco smoke
Treosulphan
Vinyl chloride
Wood dust

but also included are environmental substances to which certain occupational groups or certain segments of the general population may be exposed. Almost all of these substances have been identified by means of epidemiologic studies, although supporting evidence from experimental studies in animals generally exists. There are other compounds for which carcinogenicity is suspected because the compounds (often at very high doses) have induced tumors in one or more species of nonhuman animals (*see* **Tumor Incidence Experiments**), but evidence from environmental epidemiology is required before a substance can be considered a known human carcinogen.

Determining whether an environmental exposure has caused an increased risk of disease is generally not easy. A series of criteria need be satisfied before an **association** between the exposure and the disease can be considered causal in nature (*see* **Hill's Criteria for Causality**), and evidence regarding whether they are met is not always clear. The primary tools for making such an assessment are epidemiologic case–control and cohort studies; their utility in environmental epidemiology is described below.

Case–Control Studies

The most common epidemiologic study design for evaluating the environmental determinants of chronic diseases is the case–control study. This approach provides the advantage of the ability to assemble relatively large numbers of patients with the disease of interest (often not feasible via cohort studies) whose exposure histories can then be ascertained. These histories are then compared with those obtained in a similar manner from appropriately selected **controls**, and **odds ratios** calculated as the measure of association between the environmental exposure of interest and the risk of the disease. The case–control approach typically enables the collection of information not only on the key exposure, but also on other factors that may influence risk, so that **confounding** by these other disease determinants can be controlled.

Methods for the appropriate selection of controls, a key concern in case–control studies, are described in detail in the article on **case–control studies**. The essential feature is that the controls be selected from the same base population (study base) from which the cases arise. Thus, if cases arise, say, from among patients with lung function impairment detected by screening employees in a particular industrial facility, whereas controls are selected from the general population of the area where the facility is located, it is possible that case–control differences could be influenced by nonenvironmental determinants that led to employment in the facility. Some of these confounding factors (e.g. age, sex, education) might be controlled for in the statistical analysis, but differences in unmeasured factors may exist.

This fundamental requirement of the same study base for cases and controls is sometimes violated in studies of environmental risk factors because the underlying population from which the cases arose is not always well defined. For example, in studies of rare diseases, specialty treatment centers are often sought for the ascertainment of cases. Selecting controls from other patients is generally advantageous when cases are restricted to a particular one or several hospitals, but in this situation the case patients come not only from the surrounding areas, but also from far distances to receive the specialized care. Patients admitted to the same facility for other conditions might not have the same referral patterns, and thus selecting the most appropriate controls is problematic.

The method of ascertainment of information should also be similar for cases and controls. Suppose the cases with lung function abnormalities mentioned above and the controls were both drawn from employees of similar characteristics (except for their disease) at the industrial facility. Suppose also that information on exposure to silica, the environmental factor of interest, and on presence of concomitant silicosis, was obtained for the cases as part of their evaluation of lung abnormalities. Then if information

for the controls came not from a review of personnel or radiographic information as for the cases, but from questionnaires about silica exposures and about diagnoses of silicosis, differences between cases and controls could be due to the way the information was ascertained rather than to the presence of silica exposure or silicosis *per se* (*see* **Bias in Case–Control Studies**).

Issues regarding **selection**, information or other biases must be considered in all case–control studies. When evaluating certain environmental exposures, additional concerns need be addressed. One is **recall bias**. Individuals afflicted with a disease, particularly if life threatening in themselves or in a close relative, often tend to wonder about what may have caused the illness, or may have been prompted to examine past events during medical work-up and treatment. Sometimes this involves attempts at reconstructing events that may have taken place many years before, and can involve a process of speculating as to possible critical initiating "environmental" events. Controls, on the other hand, will typically have not gone through such prompting or soul searching. Thus, when cases and controls are interviewed, the responses may be different in part because the cases may have thought more about their illness (*see* **Bias in Observational Studies**).

Such recall biases can be mitigated in part by asking structured and specific as opposed to open-ended questions (*see* **Interviewing Techniques**; **Questionnaire Design**). At one time it was thought that exposure to chicken pox and the Varicella virus during pregnancy might increase the risk of cancer in the offspring. The suggestion came from a case–control study in which mothers were asked to list illnesses occurring during pregnancy, and mothers of the cancer patients more often listed chicken pox than mothers of the controls [7]. When prenatal medical records were examined for mention of chicken pox, however, no case–control differences were found. Furthermore, when the questionnaire for the mothers was changed to ask specially about chicken pox during pregnancy, again no case–control difference was apparent [13]. The main effect of medical record review and specific questioning was to raise the reported prevalence of chicken pox among the controls to match that among the cases. The controls had underreported the infection when asked the nonspecific open-ended question about illnesses during pregnancy, probably because they had not thought about and recalled antecedent conditions to the extent of the mothers whose children had developed cancer.

Despite the potential problems with the case–control approach, if conducted properly these studies can provide crucial information about environmental determinants of disease. Case–control studies are often the most appropriate mechanism to test hypotheses generated by the ecologic studies of grouped cancer rates and their correlates described in the previous section. Thus, the US cancer maps spawned a series of case–control studies in areas of the country where rates of particular cancers were elevated. The case–control studies obtained detailed information on the lifestyle, occupational, medical, and other characteristics of the subjects. These studies determined, for example, that although cigarette smoking was the dominant risk factor for lung cancer, employment in shipyards during World War II (and presumed exposure to asbestos) contributed to the clustering of excess rates of this cancer in the 1960s–1980s in southern coastal areas [10, 14]. Other case–control studies showed that use of herbicides in farming was associated with the higher rates of non-Hodgkin's lymphoma in the plains states [66], factors associated with northern European ancestry contributed to the clustering of excess kidney cancer mortality in north central states [52], and use of moonshine whiskies was largely responsible for the excess of esophageal cancer among black men in coastal South Carolina [18].

Case–control studies have also helped elucidate environmental factors for other diseases, especially when used to test etiologic hypotheses. The studies also have been used for hypothesis generation. An advantage of the case–control approach is the ability to look at many different antecedent exposures simultaneously. Often multiple associations are examined in case–control studies with a tendency to report in separate articles those associations based on odds ratios whose **confidence limits** exclude the value 1.0. The inherent **multiple comparison** problem is sometimes masked by the splitting of findings into multiple publications, especially when *ex post facto* explanations of the findings emerge as if they were a priori hypotheses, which can lead to undue emphasis of their importance. The case–control study, however has proven to be the key tool of environmental epidemiology, and its strengths generally outweigh

its disadvantages. Some of the problems that beset case–control studies of environmental factors can be overcome in cohort studies, as described below.

Cohort Studies

Cohort studies involve the identification of individuals characterized by exposure status and followed for the occurrence of disease after exposure. The studies tend to be much larger, and often more expensive than case–control studies, because sizable numbers of participants must be enrolled and followed to generate sufficient numbers of cases for meaningful analysis.

Cohort studies are especially useful in tracking occupational groups exposed to chemical or physical substances hypothesized to increase risk of cancer or other diseases. The hypotheses about potential risks can best be tested among groups of people with the widest range of exposure to the substance, typically workers involved in the manufacture or use of the agent. Of the over 50 compounds or processes classified as capable of causing cancer in humans following sufficient exposure (Table 1), more than half are found in occupational settings. Hence studies, typically cohort studies, among occupational groups have provided key evidence regarding whether human exposure might increase risk of cancer.

Occupational cohort studies also are of direct relevance to assessing effects of general environmental exposures. If no excess risk is seen among workers handling or otherwise heavily exposed to the agents, then it is highly unlikely that off-site exposures, generally at much lower doses, would increase risk. If an occupational excess is found, detailed study would be undertaken to characterize the risk as a function of level, timing and duration of exposure. The resultant **dose–response** trends would then be informative in predicting risks at low environmental levels. Hence, much of the information on potential effects of general environmental exposures arises from occupational studies, typically cohort studies (*see* **Occupational Mortality**).

In evaluating environmental agents with cohort studies, a key element is the measurement and classification of exposure status of cohort members. One of the strengths of a cohort study, besides its ability to ascertain a broad spectrum of health outcomes, is its ability to characterize all participants by exposure level prior to disease outcome. In principle, a prospective cohort study (where current cohort members are followed forward in time) should generally be able to obtain more reliable exposure data than a case–control study. Special problems may arise in **historical cohort studies** (where cohort members identified in the past are traced to the present) and in difficult settings where there may be imprecision of exposure assessment, including misclassification of categorical levels of exposure. In occupational studies of potentially hazardous substances, especially retrospective or historical cohort surveys where rosters of past employees are assembled, complete knowledge of levels and duration of exposure to individual workers is seldom known, even for relatively heavily studied substances. For example, asbestos is a well-known carcinogen, substantially increasing risk of lung cancer and mesothelioma when exposure levels are sufficiently high, but in nearly 50 cohort studies of various groups of asbestos-exposed workers, estimates of cumulative exposures to individual workers are available in less than a dozen, and even in those the individual estimates are based on rough approximations of presumed average airborne asbestos exposure concentrations associated with specific jobs across broad time periods [39]. Thus, some level of misclassification of exposure is bound to occur. For dichotomous categorization of exposure, if misclassification is random (nondifferential), then *on average* the **relative risks** associated with the exposure will be dampened and pulled towards the null (*see* **Bias Toward the Null**), although this is not necessarily the case when multiple categories are involved (*see* **Measurement Error in Epidemiologic Studies**). Of course, in any particular investigation, chance errors in classification could result in exaggerated as well as attenuated **relative risk** estimates.

One cohort study with relatively precise exposure estimates is the study of survivors of the atomic bombs of Hiroshima and Nagasaki. Since the early 1950s, a cohort of nearly 100 000 individuals has been tracked for mortality, and subsets have been tracked for other health outcomes [41]. Each cohort member at enrollment into the study was questioned about his/her whereabouts at the time of explosion. The event was so traumatic that nearly all individuals could recall exactly where they were and even what position they were standing in when the blast occurred. The radiation from the bombs was released almost instantaneously, so

that exposure occurred within seconds (there was little radioactive fallout). Experimental models had demonstrated that levels of gamma and neutron radiation declined exponentially as distance from the hypocenter increased and provided the basis, after taking into account shielding from metal, wooden, brick, and other structures, for estimates of radiation received for almost all cohort members. Subsequent statistical analyses have shown that rates of mortality from leukemia, breast, and several other cancers, but not nonmalignant disease, varied in proportion to radiation dose [66], and have provided a valuable base of information for the establishment of radiation safety standards worldwide.

Cohort studies of nonoccupational population groups exposed to environmental chemicals or other pollutants seldom are able to measure exposure very precisely. For example, in 1976 an explosion in a plant near Seveso, Italy, resulted in 2,3,7,8-tetrachlorodibenzo-*p*-dioxin contamination in surrounding neighborhoods. Chloracne and other reversible acute effects of exposure were observed in some residents closest to the plant, and a long-term monitoring for mortality, cancer, and other health outcomes was established. Cohorts of over 50 000 residents, classified in residential zones according to degree of potential for dioxin exposure, have been followed since. Results have been mixed, with little departure from expectations in mortality, suggested excesses of certain types of cancer, and no evidence of birth abnormalities following the contamination [4–6, 48, 60]. There have been difficulties in estimating exposure, but average exposure levels may have been below the limits of epidemiology as a tool for detecting and quantifying risks of chronic diseases in this population [17].

Other prospective cohort studies also may enable precise exposure classification for exposures occurring at the start of or during follow-up. In the **Framingham** [25, 30] and other cohorts where heart disease was the primary endpoint, measurement of blood pressure, serologic indicators, and other exposure markers could be assessed using best available methodology, and participants could be classified by baseline levels of these variables. Many of the markers are measured with error, sometimes with the particular measured value being a realization from an underlying probabilistic distribution with a large **variance**, so that even in these investigations perturbations in exposure classification can occur. Nevertheless, cohort studies have provided key information on the environmental determinants, including lifestyle factors, and disease risk. In the Framingham study, follow-up of approximately 5000 residents beginning in 1948 has demonstrated the predictive value of serum cholesterol (originally total and subsequently LDL and HDL fractions), hypertension, smoking, and dietary factors in cardiovascular disease risk [32, 64]. Other cohorts established from the 1950s classified individuals by tobacco smoking status. Reports from the US Surgeon General in a series of comprehensive US governmental monographs which unequivocally declared that cigarette smoking increases risk of lung cancer, relied heavily on results from cohorts of British physicians [28], American Cancer Society volunteers [34], US veterans [51], and other groups in reaching this conclusion.

One of the problems facing case–control studies, the potential for recall bias associated with the differential recollection of events because of the disease, is eliminated in cohort studies since the environmental exposure is determined prior to and independent of the disease occurrence.

Cohort studies also are less likely than case–control studies to be affected with selection or information biases, because the study base can often be unambiguously defined and data on all cohort members may be more readily collected in a standardized fashion (*see* **Bias in Cohort Studies**). Potential study base problems can arise, however, in situations where the cohort consists of exposed individuals whose disease experience is compared with that of an external population, for example with national rates of disease. Such comparisons are common in occupational and other cohort studies, but the resulting indices of risk (e.g. relative and **absolute risks**, standardized incidence or mortality ratios; *see* **Standardization Methods**) can be influenced by differences between the cohort and external populations other than the exposure itself. Among the 50 cohort studies of asbestos-exposed workers, for example, only a minority involved comparisons of disease rates among heavy vs. light vs. nonexposed workers, but instead usually compared the overall occupational group vs. national or local populations [39]. It is known that employed populations tend to have a somewhat more favorable mortality experience than the general population because ill and less fit individuals are less likely

to be employed (see discussion of the "healthy worker effect" in **Occupational Epidemiology**). On the other hand, some groups, especially of blue collar workers, may have higher prevalences of cigarette smokers (this is the case for heavily asbestos exposed insulation workers [35]) or have other attributes which could increase risk relative to general population norms. This problem is mitigated by use of internal comparisons by level of exposure, but even here nonexposure-related differences could confound comparisons. Thus, for example, neurologic and behavioral differences of production workers in the same facility could be related to social or other traits rather than to exposure to solvents or other chemicals.

Hence, cohort studies, like case–control studies, must take care to control for confounding in evaluations of potential adverse effects of environmental exposures. Cohort studies, however, are often at a disadvantage compared to case–control studies with respect to control for confounding. With their smaller size, case–control studies usually seek information on all known or suspected risk factors for the disease being studied. Cohort studies, on the other hand, typically with multiple disease endpoints and large number of participants, generally do not have this luxury and must limit the amount of information obtained per subject. A solution to this problem is offered by conducting case–control studies nested within cohort studies (*see* **Case–Control Study, Nested**). This approach enables detailed exposure and confounding variable assessment in samples of cohort members rather than in the entire cohort. In a cohort study of over 35 000 workers in mines and factories in Southern China, for example, broad classification of exposure to silica was obtained for all cohort members, but detailed occupational exposure, smoking, and other histories were obtained only for the nearly 300 persons with lung cancer plus about twice as many matched controls [29, 50]. The study found mixed results, suggesting a small increase in lung cancer among those with silicosis but not with silica exposure *per se*.

Cohort studies, particularly those involving a large number of lifestyle variables, can also suffer from the tendency to report on one variable at a time in a publication, as mentioned earlier when discussing case–control studies. Thus, the **multiple comparison** problem can arise in cohort studies when results are divided and described according to the least publishable unit.

Together, cohort and case–control studies provide the basis for most of what is known about the environmental determinants of human illness. Nevertheless, because of methodologic limitations these nonexperimental studies often fall short of providing sufficient evidence to determine whether exposure to a particular environmental agent can increase risk of disease. The most definitive epidemiologic evidence for determining a cause-and-effect relation can come from an experimental trial, whereby random assignment of individuals to exposed and unexposed groups mitigates against the biases that can afflict observational studies (*see* **Bias in Observational Studies**; **Bias, Overview**). Such trials are described below.

Randomized Trials

Trials involving exposure of individuals to environmental substances are ethical only when there is sufficient suspicion that the substance may lower risk of disease (*see* **Ethics of Randomized Trials**). Trials involving the evaluation of new drugs, for example, have been common and have provided the mechanism for the discovery and/or confirmation of the effectiveness of various treatments for human illness. The principles of these clinical trials also apply to the evaluation of a variety of agents that offer potential for the prevention of disease.

A number of randomized **prevention trials** have been launched, many within the past decade or so. The largest investigations have involved nutritional interventions, randomly assigning participants into groups receiving vs. not receiving certain vitamins, minerals or dietary modifications. Follow-up has typically been concurrent with the intervention, with cancer and heart disease generally the primary endpoints. In some instances the randomization unit is not the individual, but rather a group, for example, as in the one trial where communities were randomized to receive intense vs. routine educational programs aimed at smoking cessation [22] (*see* **Group-randomization Designs**). In this trial the direct endpoint was not a health outcome, but rather an exposure (cigarette smoking), which if reduced would lower subsequent disease risk. Some of the key trials in cardiovascular disease research also have involved interventions to lower exposure markers. The Multiple Risk Factor Intervention Trial, a large trial enrolling nearly 13 000 males age 35–57 at high

risk of heart disease, sought to alter several exposure (e.g. smoking), biomarker (e.g. serum cholesterol), and precursor conditions (e.g. hypertension) that increase risk of cardiovascular disease [53, 54]. Thus, intervention trials can be used to provide experimental tests both of whether increasing exposure to an environmental agent thought to reduce disease risk or decreasing exposure to an agent thought to be hazardous results in a lower risk.

The primary advantage of clinical/intervention trials over observational studies arises from randomization. The random assignment of persons to treatment groups tends to reduce differences between the groups with respect to all variables except the intervention or variables correlated with the intervention. Thus two of the main afflictions of case–control and cohort studies, namely bias and confounding, are removed. Chance is still an issue, but by choosing a sufficiently large study size for the trial, the effects of random errors can be minimized. One of the largest intervention trials involved random assignment, within strata defined by age and sex, of nearly 30 000 individuals in Linxian, China, into one of eight treatment groups [15]. The groups were defined by a one-half replicate of a 2^4 factorial experimental design (*see* **Fractional Factorial Designs**), whereby four types of vitamin/mineral supplements were being assessed as potential inhibitors of esophageal and stomach cancer in a population with one of the world's highest rates of these cancers. Cigarette smoking status, a risk factor for these cancers, was not matched for in the design, but the randomization accomplished this task. After randomization, the prevalence of smoking across the eight intervention groups varied by less than 1% [44]. Similarly, other measured differences across the treatments were all uniformly small, providing confidence that unmeasured variables were also likely to be evenly distributed by treatment, and thus bias and confounding were unlikely to affect study results.

The large vitamin/mineral intervention trials thus far have shown mixed results for the effects of supplementation on subsequent cancer or heart disease risk, despite the consistent demonstration from both case–control and cohort studies of lowered risks among persons with high intakes of foods (especially fruits and vegetables) with high contents of carotenoids, vitamin C, and other nutrients. The Linxian trial [15] found a small (13%) but significant reduction in cancer mortality following 5 years of supplements with a combination of beta carotene, vitamin E, and selenium. Large trials in Finland [2] and the US [57], however, found significantly increased, rather than decreased, risks of lung cancer among smokers supplemented with beta carotene or beta carotene plus retinol, respectively, while a 12-year follow-up of among 22 000 US physicians, few of whom smoked, found no effect of beta carotene supplementation on cancer or heart disease [37].

The beta carotene results from these trials provide a stark reminder of the limitations of the observational (case–control and cohort) studies. The vast majority of observational studies have shown lowered cancer risks, with reductions typically of 30%–50% among heavy compared with light consumers of foods rich in beta carotene [67]. Limited evidence of cancer inhibition from beta carotene studies in experimental animals provided a biologic basis for the hypothesis, and even a potential mechanism of action, in particular beta carotene's ability to quench singlet oxygen radicals. Publication in 1981 of a prominent review article [61] helped to stimulate enthusiasm for the beta carotene hypothesis and to lead to the incorporation of beta carotene in several randomized trials. The results of these trials now indicate that the enthusiasm may have been misplaced, and that correlates of beta carotene rather than beta carotene *per se* may have been responsible for the reduced risks associated with intake of beta carotene-containing foods seen in the case–control and cohort studies. This unfolding of events suggests that caution be applied in the interpretation of case–control or cohort studies linking various environmental exposures with disease risk, and heightens the necessity for careful assessment of bias, confounding, and chance before etiologic interpretations are offered, especially since few environmental exposures will be able to be evaluated via randomized intervention trials.

References

[1] Adami, H.-O. (1996). Sweden: a paradise for epidemiologists?, *Lancet* **347**, 588–589.

[2] Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers, *New England Journal of Medicine* **330**, 1029–1035.

[3] American Cancer Society (1995). *Cancer Facts and Figures, 1995*. American Cancer Society, Atlanta.

[4] Bertazzi, P.A., Pesatori, A.C., Consonni, D., Tironi, A., Landi, M.T. & Zocchetti, C. (1993). Cancer incidence in a population accidentally exposed to 2,3,7,8-tetrachlorodibenzo-para-dioxin, *Epidemiology* **4**, 398–406.

[5] Bertazzi, P.A., Zocchetti, C., Pesatori, A.C., Guercilina, S., Consuinni, D., Tironi, A. & Landi, M.T. (1992). Mortality of a young population after accidental exposure to 2,3,7,8-tetracholorodibenzodioxin, *International Journal of Epidemiology* **21**, 118–123.

[6] Bertazzi, P.A., Zocchetti, C., Pesatori, A.C., Guercilena, S, Sanarico, M. & Radice, L. (1989). Ten-year mortality study of the population involved in the Seveso incident in 1976, *American Journal of Epidemiology* **129**, 1187–1200.

[7] Bithell, J.F., Draper, G. & Gerbach, P. (1973). Association between malignant disease in children and maternal virus infections, *British Medical Journal* **2**, 706–710.

[8] Blot, W.J. & Fraumeni, J.F., Jr (1977). Geographic patterns of oral cancer in the United States: etiologic implications, *Journal of Chronic Diseases* **30**, 745–757.

[9] Blot, W.J. & Fraumeni, J.F. Jr (1979). Studies of respiratory cancer in high risk communities, *Journal of Occupational Medicine* **21**, 276–278.

[10] Blot, W.J. & Fraumeni, J.F., Jr (1981). Cancer among shipyard workers, in *Banbury Report*, Vol. 9. Cold Spring Harbor Laboratory, New York, pp. 37–50.

[11] Blot, W.J. & Fraumeni, J.F., Jr (1996). Cancers of the lung and pleura, in *Cancer Epidemiology and Prevention*, 2nd Ed., D. Schottenfeld & J. Fraumeni, eds. Oxford University Press, New York, pp. 637–665.

[12] Blot, W.J., Devesa, S.S., Kneller, R.W. & Fraumeni, J.F., Jr (1991). Rising incidence of adenocarcinoma of the esophagus and gastric cardia, *Journal of the American Medical Association* **265**, 1287–1289.

[13] Blot W.J., Draper, G., Kinlen, L. & Kinnier-Wilson, M. (1980). Childhood cancer in relation to prenatal exposure to chicken pox, *British Journal of Cancer* **42**, 342–344.

[14] Blot, W.J., Harrington, J.M., Toledo, A., Hoover, R., Heath, C.W., Jr & Fraumeni, J.F., Jr (1978). Lung cancer after employment in shipyards during World War II, *New England Journal of Medicine* **299**, 620–624.

[15] Blot, W.J., Li, J.-Y., Taylor, P.R., Guo, W., Dawsey, S., Wang, G.-Q., Yang, C.S., Zheng, S.-F., Gail, M., Lit, G.-Y., Yu, Y., Liu, B.-Q., Tangrea, J., Sun, Y.-H., Lin, F., Fraumeni, J.F., Jr, Zhang, Y.-H. & Li, B. (1993). Nutrition intervention trials in Linxian, China: supplementation with specific vitamin/mineral combination, cancer incidence, and disease-specific mortality in the general population, *Journal of the National Cancer Institute* **85**, 1483–1492.

[16] Boffetta, P., Kogevinas, M., Simonato, L., Wilbourn, J. & Saracci, R. (1995). Current perspectives on occupational cancer risks, *International Journal of Occupational and Environmental Health* **1**, 315–325.

[17] Boroush, M. & Gough, M. (1994). Can cohort studies detect any human cancers that may result from exposure to dioxin? Maybe, *Regulatory Toxicology and Pharmacology* **20**, 198–210.

[18] Brown, L.M., Blot, W.J., Schuman, S.H., Smith, V.M., Ershow, A.G., Marks, R.D. & Fraumeni, J.F., Jr (1988). Environmental factors and high risk of esophageal cancer among men in coastal South Carolina, *Journal of the National Cancer Institute* **80**, 1620–1625.

[19] Buescher, P.A., Taylor, K.P., Davis, M.H. & Bowling, J.M. (1993). The quality of the new birth certificate data: a validation study in North Carolina, *American Journal of Public Health* **83**, 1163–1165.

[20] Chow, W.H., Finkle, W.D., McLaughlin, J.K., Frankl, H., Ziel, H.K. & Fraumeni, J.F., Jr (1995). The relation of gastroesophageal reflux disease and its treatment to adenocarcinomas of the esophagus and gastric cardia, *Journal of the American Medical Association* **274**, 474–477.

[21] Cole, P. & Rodu, B. (1996). Declining cancer mortality in the United States, *Cancer* **78**, 2045–2048.

[22] COMMIT Research Group (1995). Community Interventional Trial for smoking cessation (COMMIT): I. Cohort results from a four-year community intervention, *American Journal of Public Health* **85**, 183–192.

[23] Cooper, R.S., Liao, Y. & Rotimi, C. (1996). Is hypertension more severe among U.S. blacks, or is severe hypertension more common?, *Annals of Epidemiology* **6**, 173–180.

[24] Creech, J.L. & Johnson, M.N. (1974). Angiosarcoma of the liver in the manufacture of polyvinyl chloride, *Journal of Occupational Medicine* **16**, 150.

[25] Dawber, T.R., Meadors, G.F. & Moore, F.E. (1951). Epidemiological approaches to heart disease: the Framingham study, *American Journal of Public Health* **41**, 279–286.

[26] Desenclos, J.C., Scaggs, M. & Wroten, J.E. (1992). Characteristics of mothers of live infants with congenital syphilis in Florida, 1987–1989, *American Journal of Epidemiology* **136**, 657–661.

[27] Devesa, S.S., Blot, W.J., Stone, B.J., Miller, B.A., Tarone, R.E. & Fraumeni, J.F., Jr (1995). Recent cancer trends in the United States, *Journal of the National Cancer Institute* **87**, 175–182.

[28] Doll, R., Peto, R., Wheatly, K., Gray, R. & Sutherland, I. (1994). Mortality in relation to smoking: 40 years' observations on male British doctors, *British Medical Journal* **309**, 901–911.

[29] Dosemeci, M., Chen, J.Q., Hearl, F.J., Wu, Z., McCawley, M.A., Chen, R.A., McLaughlin, J.K., Peng, K., Cheng, A.L., Rexing, S.H. & Blot, W.J. (1993). Estimating historical exposure to silica for mine and pottery workers in the People's Republic of China, *American Journal of Industrial Medicine* **24**, 55–66.

[30] Feinlieb, M. (1985). The Framingham Study: sample selection, follow-up, and methods of analyses, in *Selection, Follow-up, and Analysis in Prospective Studies: A Workshop*, NIH Publication No. 85–2713, L. Garfinkel, O. Ochs & M. Mushinski, eds. US Department of Health and Human Services, National Institutes of Health, Bethesda, pp. 59–64.

[31] Fiore, M.C., Novotny, T.E., Pierce, J.P., Hatziandreu, E.J., Patel, K.M. & Davis, R.M. (1989). Trends in cigarette smoking in the United States: the changing influence of gender and race, *Journal of the American Medical Association* **261**, 49–55.

[32] Gordon, T., Castelli, W., Hjortland, M.C., Rannel, W.B. & Dawber, T.R. (1977). High density lipoprotein as a protective factor against coronary heart disease, *American Journal of Medicine* **62**, 707–714.

[33] Grulich, A.E., Swerdlow, A.J., Dos Santos Silva, I. & Beral, V. (1995). Is the apparent rise in cancer mortality in the elderly real? Analysis of changes in certification and coding of cause of death in England and Wales, 1970–1990, *International Journal of Cancer* **63**, 164–168.

[34] Hammond, E.C. (1996). Smoking in relation to the death rates of one million men and women, *National Cancer Institute Monograph* **19**, 127–204.

[35] Hammond, E.C., Selikoff, I.J. & Seidman, H. (1979). Asbestos exposure, cigarette smoking, and death rates, *Annals of the New York Academy of Science* **330**, 473–490.

[36] Heath, C.W. (1988). Investigation of cancer case clusters: possibilities and limitations, in *Unusual Occurrences as Clues to Cancer Etiology*, R.W. Miller, et al., eds. Japan Scientific Press, Tokyo, pp. 27–38.

[37] Hennekens, C.H., Buring, J.E., Manson, J.E., Stampfer, M., Rosner, B., Cook, N.R., Belanger, C., LaMotte, F., Gaziano, J.M., Ridker, P.M., Willett, W. & Peto, R. (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease, *New England Journal of Medicine* **334**, 1145–1149.

[38] Herbst, A.L., Ulfelder, H. & Poskanzer, D.C. (1971). Adenocarcinoma of the vagina: association of maternal stilbesterol therapy with tumor appearance in young women, *New England Journal of Medicine* **284**, 878–881.

[39] Hughes, J.M. & Weill, H. (1994). Asbestos and manmade fibers, in *Epidemiology of Lung Cancer*, J.M. Samet, ed. Marcel Dekker, New York, pp. 185–205.

[40] International Agency for Cancer Research (1987). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Overall Evaluations of Carcinogenicity: An Updating of IARC Monographs Vols. 1 to 42*. IARC, Lyon.

[41] Jablon, S. (1985). Selection, followup and analysis in the Atomic Bomb Casualty Commission Study, *National Cancer Institute Monograph* **67**, 53–58.

[42] Johnson, C.L., Rifkind, B.M., Sempos, C.T., Carroll, M.D., Bachorik, P.S., Briefel, R.R., Gordon, D.J., Burt, V.L., Brown, C.D., Lippel, K. & Cleeman, J.I. (1993). Declining serum total cholesterol levels among US adults. The National Health and Nutrition Examination Surveys, *Journal of the American Medical Association* **269**, 3002–3008.

[43] Kosary, C.L., Ries, L.A.G., Miller, B.A., Hankey, B.F., Harras, A. & Edwards, B.K., eds (1995). *SEER Cancer Statistics Review, 1973–1992: Tables and Graphs*, NIH Publication No. 96–2789. National Cancer Institute, Bethesda.

[44] Li, B., Taylor, P.R., Li, J.-Y., Dawsey, S.M., Wang, W., Tangrea, J.A., Lin, B.-Q., Ershow, A.G., Zheng, S.-F., Fraumeni, J.F., Jr., Yang, Q., Yu, Y., Sun, Y., Li, G., Zhang, D., Greenwald, P., Lian, G.-T., Yang, C.S. & Blot, W.J. (1993). Linxian nutrition intervention trials: design, methods, participant characteristics, and compliance, *Annals of Epidemiology* **3**, 577–585.

[45] Li, D.K. & Daling, J.R. (1991). Maternal smoking, low birth weight, and ethnicity in relation to sudden infant death syndrome, *American Journal of Epidemiology* **134**, 958–964.

[46] MacMahon, B. (1992). Leukemia clusters around nuclear facilities in Britain, *Cancer Causes and Control* **3**, 283–288.

[47] Mason, T.J., McKay, F.W., Hoover, R., Blot, W.J. & Fraumeni, J.F., Jr (1975). *Atlas of Cancer Mortality for U.S. Counties 1950–1969*, DHEW Publication No. (NIH) 75–780. US Department of Health Education and Welfare, National Institutes of Health, Bethesda.

[48] Mastroiacovo, P., Spagnolog, A., Marni, E., Meazza, L., Bertollini, R., Segni, G. & Burgna-Pignatti, C. (1998). Birth defects in the Seveso area after TCDD contamination, *Journal of the American Medical Association* **259**, 1668–1672.

[49] Mathis, M.P., Lavoie, M., Hadley, C. & Toomey, K.V. (1995). Birth certificates as a source for fetal alcohol syndrome case ascertainment – Georgia, 1989–1992, *Morbidity and Mortality Weekly Reports* **44**, 251–253.

[50] McLaughlin, J.K., Chen, J.Q., Dosemeci, M., Chen, R.A., Rexing, S.H., Wu, Z., Hearl, F.J., McCawley, M.A. & Blot, W.J. (1992). A nested case-control study of lung cancer among silica exposed workers in China, *British Journal of Industrial Medicine* **49**, 167–171.

[51] McLaughlin, J.K., Hrubec, Z., Blot, W.J. & Fraumeni, J.F., Jr (1995). Smoking and cancer mortality among U.S. veterans: a 26-year followup, *International Journal of Cancer* **60**, 190–193.

[52] McLaughlin, J.K., Mandel, J.S., Blot, W.J., Schuman, L.M., Mehl, E.S. & Fraumeni, J.F., Jr (1984). A population-based case-control study of renal cell carcinoma, *Journal of the National Cancer Institute* **72**, 275–284.

[53] Multiple Risk Factor Intervention Trial Group (1977). Statistical design considerations in the NHLI Multiple Risk Factor Intervention Trial (MRFIT), *Journal of Chronic Diseases* **30**, 261–275.

[54] Multiple Risk Factor Intervention Trial Group (1982). Multiple risk factors intervention trial: risk factor changes and mortality results, *Journal of the American Medical Association* **248**, 1465–1477.

[55] National Research Council (1991). *Environmental Epidemiology*, Vol. 1: *Public Health and Hazardous Wastes*. National Academy Press, Washington.

[56] Olsen, S.F., Martuzzi, M. & Elliott, P. (1996). Cluster analysis and disease mapping – Why, when, and how?, *British Medical Journal* **313**, 863–866.

[57] Omenn, G.S., Goodman, G.E., Thornquist, M.D., Blames, J., Cullen, M.R., Glass, A., Keogh, J.P., Meyskens, F.L., Valanis, B., Williams, J.H., Barnhart, S. & Hammar, S. (1996). Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease, *New England Journal of Medicine* **334**, 1150–1155.

[58] Parkin, D.M., Muir, C.S., Whelan, S.L., Gao, Y.-T., Ferlay, J. & Powell, J., eds (1992). *Cancer Incidence in Five Continents*, Vol. VI. IARC Scientific Publications, Lyon, France.

[59] Percy, C.L., Miller, B.A. & Ries, L.A.G. (1990). Effect of changes in cancer classification and the accuracy of cancer death certificates on trends in cancer mortality, *Annals of the New York Academy of Science* **609**, 87–97.

[60] Pesatori, A.C., Consonni, D., Tironi, A., Zocchetti, C., Fini, A. & Bertazzi, P.A. (1993). Cancer in a young population in a dioxin-contaminated area, *International Journal of Epidemiology* **22**, 1010–1013.

[61] Peto, R., Doll, R., Buckley, J.D. & Sporn, M.B. (1981). Can dietary beta-carotene materially reduce human cancer rates?, *Nature* **290**, 201–208.

[62] Pickle, L.W., Mason, T.J., Howard, N., Hoover, R. & Fraumeni, J.F., Jr (1987). *Atlas of U.S. Cancer Mortality Among Whites: 1950–1980*, DHHS Publication No. (NIH) 87–2900. US Department of Health and Human Services, National Institutes of Health, Bethesda.

[63] Pirkle, J.L., Flegal, K.M., Bernert, J.T., Brody, D.J., Etzel, R.A. & Maurer, K.R. (1996). Exposure of the US population to environmental tobacco smoke. The Third National Health and Nutrition Examination Survey, 1988 to 1991, *Journal of the American Medical Association* **275**, 1233–1240.

[64] Posner, B.M., Franz, M.M., Quatromoni, P.A., Gagnon, D.R., Sytkowski, P.A., D'Agostino, R.B. & Cupples, A. (1995). Secular trends in diet and risk factors for cardiovascular disease: The Framingham Study, *Journal of the American Dietetic Association* **95**, 171–179.

[65] Sever, L.E. (1995). Epidemiologic aspects of environmental hazards to reproduction, in *Introduction to Environmental Epidemiology*, E. Talbott & G. Graun, eds. CRC Press, Boca Raton, pp. 81–98.

[66] Shimizu, T., Kato, H. & Schull, W.J. (1990). Studies of the mortality of A-bomb survivors, *Radiation Research* **130**, 249–266.

[67] Steinmetz, K.A. & Potter, J.D. (1991). Vegetables, fruit and cancer. I. Epidemiology, *Cancer Causes and Control* **2**, 325–357.

[68] Surgeon General (1989). *Reducing the Health Consequences of Smoking: 25 Years of Progress*. US Department of Health and Human Services, Office of Smoking and Health, Rockville.

[69] Wang, H.H., Hsieh, C. & Antonioli, D.A. (1994). Rising incidence rate of esophageal adenocarcinoma and use of pharmaceutical agents that relax the lower esophageal sphincter, *Cancer Causes and Control* **5**, 573–578.

[70] Winn, D.M., Blot, W.J., Shy, C.M. & Fraumeni, J.F., Jr (1981). Snuff dipping, oral cancer and dentures, *New England Journal of Medicine* **305**, 230–231.

[71] Woolbright, L.A. & Harshbarger, D.S. (1995). The revised standard certificate of live birth: analysis of medical risk factor data from birth certificates in Alabama, 1988–92, *Public Health Reports* **110**, 59–63.

[72] Zahm, S.H., Weisenburger, D.D., Babbitt, P.A., Saal, R.C., Vaught, J.B., Cantor, K.P. & Blair, A. (1990). A case-control study of non-Hodgkin's lymphoma and the herbicide 2, 4-dichlorophenoxyacetic acid (2,4-D) in eastern Nebraska, *Epidemiology* **1**, 349–356.

WILLIAM J. BLOT, WONG-HO CHOW &
JOSEPH K. MCLAUGHLIN

Epi Info

Epi Info is a free software originally developed by the US **Centers for Disease Control** (CDC) and further developed from a collaboration between the CDC and the **World Health Organization**. It was originally developed as a DOS-software but a Windows version has been available since 2002. Epi Info is devoted to the design and analysis of epidemiologic studies. It has several functions for performing basic **sample size** calculations, developing a study **questionnaire** and creating a **database**, helping with data entry and checking (*see* **Data Management and Coordination**), performing statistical analysis, and displaying study data on maps (*see* **Statistical Map**) and graphs. Whereas its capabilities are extensive regarding questionnaire development, data entry, and checking, its analytical capabilities are more limited. Complex survey analysis (*see* **Sample Surveys in the Health Sciences**) and **logistic regression** can be performed in Epi Info 2002 for Windows. The software can be downloaded from the Epi Info web site (`http://www.cdc.gov/epiinfo/`). Epi Info is a widely known and used software worldwide. According to information on the web site, over 280 000 downloads of Epi Info had been documented in 2001 (one million according to an update in 2003) and the manual had been translated from English into 13 additional languages.

(*See also* **Software, Biostatistical**; **Software, Epidemiological**)

JACQUES BENICHOU

Epicure

Epicure is a DOS-based commercial software, produced by HiroSoft International Corporation (`http://www.hirosoft.com`). It is devoted to the analysis of epidemiological and biomedical studies. It is structured as a package of five programs, GMBO, PECAN, PEANUTS, DATAB, and AMFIT. It allows users to perform some data editing (*see* **Data Management and Coordination**; **Clinical Trials Audit and Quality Control**). Its main strength is its fairly extensive analytical capabilities. They include **linear** and **nonlinear regression**, unconditional and conditional **logistic regression**, **Poisson regression**, and semiparametric and parametric **survival analysis**, through the use of the programs, GMBO, PECAN, PEANUTS, and AMFIT. Moreover, Epicure allows additive and multiplicative versions of regression model forms (*see* **Relative Risk Modeling**) and a number of unique alternative and generalized forms, as well as an easy creation of **person-year** tables through the use of the program DATAB, and a fairly advanced standardized mortality ratio analysis (*see* **Standardization Methods**) with the program AMFIT. Being a DOS-based program, Epicure may have a slightly steeper learning curve than Windows-based programs, although it is not particularly difficult to use. No **sample size** calculations are available, although the developers of Epicure have produced a free, separate program, "Power", that can be downloaded from the website of the US National Cancer Institute's Division of Cancer Epidemiology and Genetics (`http://dceg.cancer.gov`) and can be easily used for tests of **interaction**, a rare feature in sample size calculation programs.

(*See also* **Software, Biostatistical**; **Software, Epidemiological**; **Survival Analysis, Software**)

JACQUES BENICHOU

Epidemic Curve

An epidemic curve is a plot of time trends in the occurrence of a disease or other health-related event for a defined population and time period. The epidemic curve can help to demonstrate that the events are in excess of what would be expected based on past experience [4]. Time intervals are indicated on the x-axis and the event rate is shown on the y-axis. The event rate may measure the number of cases per unit time (e.g. cases per day or year), or it may express the numbers of events relative to the number in the study population (cases per 100 000 person–years). The latter **incidence rates** may be age-adjusted (*see* **Standardization Methods**).

Historically, the epidemic curve has been widely used by infectious disease epidemiologists to document the scope and duration of an epidemic, to help determine the source of the infection and the modes of transmission or exposure, and to glean information about the **incubation period** of the disease [6]. Epidemic curves are used in other settings to document the scope of public health problems.

Example 1 An estimated 224 000 persons nationwide became ill with salmonellosis during 1994 after eating a nationally distributed brand of ice cream that was contaminated with *Salmonella enteritidis* [5]. Epidemic curves for Minnesota, an epicenter of the outbreak, and for the entire US helped to determine that the outbreak occurred during September and October of that year as a result of contamination that occurred between mid-August and mid-October.

Example 2 The number of **AIDS** cases in the US exhibited exponential growth during the early 1980s, followed by a slowing of the rate of increase beginning in mid-1987 [2]. This pattern likely reflects a decline in the number of new HIV infections compared with peak infection rates in the mid-1980s.

Example 3 Age-adjusted lung cancer death rates per 100 000 men in the US climbed 15-fold between 1930 and 1990 [7], a result of trends in tobacco consumption during earlier decades [3] (*see* **Smoking and Health**). The epidemic curve is expected to decline in the US as a result of smoking cessation [3]; but epidemiologists predict a rising epidemic curve for tobacco-related deaths in the next century among Chinese men, as a result of their recent increase in cigarette smoking [8].

While it is reasonable to speak of an "epidemic" of violent death or of specific neoplastic diseases (Example 3), analysis of the epidemic curve is currently most refined in the field of infectious disease epidemiology.

In a *common source outbreak*, susceptibles are exposed to a pathogen about the same point in time (Example 1). In this type of outbreak the resulting epidemic curve tends to be relatively short and sharp, as the cases distribute according to the incubation period of the disease. With food-borne outbreaks, a point source of exposure is sometimes identified from the case reports and the pathogen may be recognized from the observed incubation period. In a *propagated* or *progressive outbreak*, cases result from person-to-person transmission, often yielding a broader epidemic curve (Example 2).

Statistically, if the infection curve is known, one can estimate the distribution of the incubation period from the observed epidemic curve [1]. Conversely, if the incubation period is known, one can estimate the infection curve from the observed epidemic curve (*see* **Back-calculation**). To avoid bias, it is essential that the epidemic curve be constructed using a well-defined case definition and that surveillance for the event is as complete and consistent as possible. In practice, the case definition may include both clinical and epidemiologic criteria [9].

References

[1] Bacchetti, P. & Moss, A.J. (1989). Incubation period of AIDS in San Francisco, *Nature* **338**, 251–253.

[2] Brookmeyer, R. (1991). Reconstruction and future trends of the AIDS epidemic in the United States, *Science* **253**, 37–42.

[3] Devesa, S.S., Blot, W.J. & Fraumeni, J.F. Jr (1989). Declining lung cancer rates among young men and women in the United States: a cohort analysis, *Journal of the National Cancer Institute* **81**, 1568–1571.

[4] Evans, A.S. (1989). Epidemiologic concepts and methods, in *Viral Infections of Humans: Epidemiology and Control*, A.S. Evans, ed., 3rd Ed. Plenum, New York.

[5] Hennessy, T.W., Hedberg, C.W., Slutsker, L., White, K.E., Besser-Wick, J.M., Moen, M.E., Feldman, J., Coleman, W.W., Edmonson, L.M., MacDonald, K.L. & Osterholm, M.T. (1996). A national outbreak of *Salmonella enteritidis* infection from ice cream, *New England Journal of Medicine* **334**, 1281–1286.

[6] Kelsey, J.L., Thompson, W.D. & Evans, A.S. (1986). *Methods in Observational Epidemiology*. Oxford University Press, New York, Chapter 9, pp. 212–253.

[7] Parker, S.L., Tong, T., Bolden, S. & Wingo, P.A. (1996). Cancer statistics, 1997, *CA-A Cancer Journal for Clinicians* **47**, 5–27; see Figure 4.

[8] Peto, R., Chen, Z. & Boreham, J. (1996). Tobacco-the growing epidemic in China, *Journal of the American Medical Association* **275**, 1683–1684.

[9] Sharrar, R.G. (1992). General principles of epidemiology, in *Preventive Medicine and Public Health*, B.J. Cassens, ed., 2nd Ed. Williams & Wilkins, Baltimore, pp. 1–21.

(*See also* **Communicable Diseases**; **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**; **Epidemic Thresholds**; **Infectious Disease Models**)

PHILIP S. ROSENBERG

Epidemic Models, Control

The main reason for studying mathematical models of disease spread must surely be the hope that improved understanding of the transmission processes may lead to more effective control strategies.

In the literature on mathematical modeling (whether deterministic or stochastic) of infectious disease spread, there have been essentially two approaches to epidemic control (*see* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**). These are (a) to intervene in such a way as to reduce the basic reproduction ratio R_0 to below (*see* **Reproduction Number**) 1; and (b) to define explicit costs of intervention and of infection and choose the intervention policy which minimizes the expected total cost. In either case, there are a variety of possible forms of intervention that may be considered, including vaccination of susceptible individuals, isolation of infective individuals from the susceptible population, or interventions (such as public education campaigns) aimed at reducing the rate of contact between infectives and susceptibles.

A third approach that has been suggested more recently is that of *contact tracing*. This may be regarded as a variant of (a) above, in which individuals to be vaccinated are chosen not at random, but rather by tracing potential infectious contacts made by known infected individuals.

Threshold-based Intervention

Much work has concentrated upon vaccination strategies aimed at reducing R_0 to below 1. The motivation for this is provided by threshold theorems such as that of Kermack and McKendrick [8], which tell us that if $R_0 < 1$, then a major epidemic is not possible; only a minor outbreak can occur (*see* **Epidemic Thresholds**). To be more specific, consider a particular model for disease spread. The most commonly used continuous time stochastic model, sometimes referred to as the general stochastic **SIR** (Susceptible – Infective – Removed) epidemic, is defined as follows.

For $t \geq 0$, denote by $X(t)$, $Y(t)$, $Z(t)$ the numbers of susceptible, infective, and removed individuals, respectively, in the population at time t. The population is supposed closed, so that total population size $X(t) + Y(t) + Z(t)$ remains constant over time. Thus, the process is completely described by $\{(X(t), Y(t)) : t \geq 0\}$, which is taken to be a continuous time **Markov chain** with infinitesimal transition probabilities

$$\begin{aligned}\Pr(X(t+\delta t) = x-1, Y(t+\delta t) = y+1 \mid X(t) \\ = x, Y(t) = y) = \beta x y \delta t + o(\delta t), \\ \Pr(X(t+\delta t) = x, Y(t+\delta t) = y-1 \mid X(t) \\ = x, Y(t) = y) = \gamma y \delta t + o(\delta t), \qquad (1)\end{aligned}$$

where β is the infection rate parameter and γ the removal rate parameter.

For the general SIR model, the basic reproduction ratio is given by $R_0 = \beta X(0)/\gamma$, so that (by the threshold theorem) major outbreaks are only possible if $X(0) > \gamma/\beta$. If $R_0 > 1$, then in order to prevent a major outbreak, we must vaccinate at least a proportion $\theta = 1 - (1/R_0)$ of the susceptible population, so that the vaccinated population has reproduction ratio $R_0' \leq \beta(1-\theta)X(0)/\gamma = 1$.

For more complicated models, such as those featuring nonhomogeneous contact structure between individuals, reducing R_0 to below 1 can be a more challenging problem. It may not be merely a matter of calculating how many susceptibles to vaccinate, but also of deciding which particular individuals should be vaccinated. For instance, Ball et al. [4] define (*inter alia*) a stochastic model for the spread of infection through a population consisting of a large number of households, the rate of infectious contact being greater between individuals within the same household than between individuals in distinct households. In [4], it is conjectured (and proved in certain particular cases) that the policy that requires the least number of vaccinations in order to reduce the value of its threshold parameter $\tilde{R}_0$ to below 1 is the *equalizing strategy* of vaccinating in such a way as to leave the remaining numbers of susceptibles in each household as nearly equal as possible.

Contact Tracing

Usually, vaccination models suppose that the individuals to be vaccinated will be chosen at random, whether from the entire population or from particular groups within the population. An alternative approach is via *contact tracing*. One now attempts to identify infectious contacts of each known infected index case, targeting the vaccination effort toward

such contacted individuals. Such an approach has been suggested for STDs, tuberculosis, and infections spread by needle sharing ([12] and references therein). The advantage is that vaccinations are targeted at those most in need. The disadvantage is that tracing such individuals may be costly and time-consuming, so that the total number of vaccinated individuals at any time will be smaller than under a mass vaccination strategy; furthermore, vaccinations might be better directed to protect those who have not yet been contacted.

There has been particular interest recently in whether mass vaccination or targeted vaccination based on contact tracing would be more effective in combating a bioterrorist smallpox attack. The possibility of combining the two approaches has also been considered, either with targeted vaccination in the early stages of an epidemic, switching to mass vaccination if the number infected grows beyond a certain point, or alternatively, mass vaccination applied to a certain proportion of the population prior to any outbreak, with further targeted vaccinations once an outbreak begins.

The conclusion as to which method is more efficient is found to depend upon the particular modeling assumptions, such as whether a continuous population or discrete individual model is used, how one allows for the time taken to carry out vaccinations, and whether the population is assumed to mix homogeneously or in some more structured way ([9] and references therein). The importance of realistic models, and of considering a variety of models capturing different aspects of the reality, is thereby vividly illustrated.

Control Theoretic Approaches

For practical implementation, the approach of vaccinating susceptibles so as to reduce R_0 to below 1 has the great advantage of simplicity. However, such an approach does not make explicit the costs of infection or of intervention. One effectively assumes that prevention of a major epidemic is worth any cost, whereas intervention that does not prevent a major epidemic is worthless, even if the spread of infection is reduced. Thus, for instance, a policy of isolating some infectives from the susceptible population, which clearly will tend to decrease the spread of infection, nevertheless, has no effect upon the value of R_0.

A more thoroughgoing approach to epidemic control can be provided through mathematical control theory. When adopting such an approach, it is necessary to set out explicitly the costs of intervention and of infection. One then aims to find the intervention policy which minimizes the expected total cost. A comprehensive review of applications of control theory to infectious disease models up to 1977 is given in [14].

Isolation of Infectives

To give a specific example, we return to the general stochastic SIR epidemic described above. For this model, Abakuks [2] considered a form of intervention, which allows for the instantaneous isolation of any number of infective individuals from the susceptible population. The cost of disease spread was taken to be proportional to the number of individuals infected during the epidemic, while the cost of intervention was taken proportional to the number of infectives artificially isolated.

Suppose that at time t the epidemic is in state $(X(t), Y(t)) = (x, y)$. Then one must decide whether to isolate a single infective, or do nothing until the next natural transition occurs. (Isolation of several infectives can be achieved by repeatedly choosing to isolate single infectives.) Take the cost of an individual being infected to be the unit of cost; denote by L the cost of isolating one infective, and denote by $V(x, y)$ the expected total future cost of adopting an optimal policy. Then for $x, y \geq 1$, the expected future cost of waiting until after the next natural transition, and adopting an optimal policy from then on, is given by

$$W(x, y) = \frac{\beta x}{\beta x + \gamma}(1 + V(x - 1, y + 1)) + \frac{\gamma}{\beta x + \gamma} V(x, y - 1). \quad (2)$$

The expected future cost of isolating an infective immediately, then adopting an optimal policy, is $L + V(x, y - 1)$. The optimal policy is to take whichever of the above two actions results in the smaller expected cost, and so the expected cost of such a policy is

$$V(x, y) = \min\{W(x, y), L + V(x, y - 1)\} \quad (x, y \geq 1). \quad (3)$$

Equations (2) and (3) are the Bellman optimality equations for this system (*see* **Dynamic Programming**). Together with the boundary conditions $V(x, 0) = 0$ for $x \geq 0$ and $V(0, y) = 0$ for $y \geq 0$, the Bellman equations may be recursively solved to determine the cost function $V(x, y)$ and the optimal policy itself. From (2) and (3), it was shown in [2] that the optimal isolation policy takes the following simple form. For each $x \geq 0$, there exists an integer $s(x) \geq 0$ such that the optimal policy is to isolate *all* infectives if $1 \leq y \leq s(x)$, but to isolate none otherwise. Furthermore, $s(x + 1) \geq s(x)$ for $x \geq 0$. So if an epidemic starts from a state (x, y) with $y \leq s(x)$, then the optimal policy immediately isolates all infectives, thereby terminating the epidemic; on the other hand, if $y > s(x)$, then the epidemic is allowed to proceed unimpeded until the first time that $Y(t) \leq s\,(X(t))$, at which point the epidemic is terminated by the isolation of all infectives.

The intervention considered in [2] is of *impulse* type, meaning that intervention causes an instantaneous change in the state of the system. An alternative to this is intervention to change the *rate* at which transitions occur. For instance, for the general stochastic SIR epidemic Wickwire [13] considers intervention to increase the removal rate parameter at time t from γ to $\gamma + v(t)$ $(0 \leq v(t) \leq 1)$. The cost of infection is taken equal to the time spent in an infectious state by all individuals during the course of the epidemic, $\int_0^\infty Y(s)\,\mathrm{d}s$, while the cost of intervention is taken to be proportional to $\int_0^\infty v(s)Y(s)\,\mathrm{d}s$, so that the total expected cost is

$$E\left[\int_0^\infty (1 + hv(s))Y(s)\,\mathrm{d}s\right] \tag{4}$$

for some constant h.

The expected cost $V(x, y)$ of an optimal policy starting from state (x, y) then satisfies the Bellman equations

$$V(x, y) = \min_{v \in [0,1]} \frac{\begin{array}{c}(1 + hv) + \beta x V(x - 1, y + 1) \\ +(\gamma + v)V(x, y - 1)\end{array}}{\beta x + \gamma + v} \qquad (x, y \geq 1). \tag{5}$$

An analysis of (5) allows Wickwire [13] to show that the form of the optimal policy is as follows:

If $h\gamma \leq 1$ then take $v = 1$ for all (x, y) with $x, y \geq 1$.

If $h\gamma > 1$ then for each $x \geq 1$ there exists an integer $\tilde{s}(x) \geq 0$ such if $1 \leq y < \tilde{s}(x)$ then take $v = 1$, if $y \geq \tilde{s}(x)$ take $v = 0$.

The optimal policy is thus of *bang-bang* type; that is, the value of v is always taken to be either 0 or 1, never any intermediate value. Notice that this policy is of the same form as the optimal impulse control policy of [2]. In each case, if the number of infectives y lies above some boundary then we intervene to isolate infectives as rapidly as possible; if the value of y lies below the boundary, we do not intervene.

In the examples above, the spread of infection is reduced by isolation of infective individuals from the susceptible population. One could alternatively intervene by immunizing susceptible individuals, or by reducing the rate of contact between susceptible and infective individuals.

Immunization

Policies which allow the instantaneous immunization of susceptible individuals are studied in [1]. Equations corresponding to (2) and (3) can be written down and solved recursively just as for isolation policies, but the structure of the optimal policy proves somewhat more complicated. For this reason, in [3] attention is restricted to total immunization policies, in which the only possible intervention is to immunize all x susceptibles simultaneously, at cost $A + Kx$. In this case it is found that for each $x \geq 0$ there exists an integer $t(x)$, $0 \leq t(x) \leq \infty$, such that the optimal policy is to immunize all susceptibles if $y > t(x)$, but not otherwise.

Alternatively, one may allow for immunizations at a finite rate. Some results for this problem are outlined in [14]. The transition rates of the general stochastic SIR epidemic are now modified by allowing additional transitions $(x, y) \to (x - 1, y)$ to occur at rate $u(t)$, where the value of $u(t)$ is to be chosen in $0 \leq u(t) \leq 1$. Higher values of $u(t)$ incur higher costs, but reduce the spread of infection. A cost function similar to (4) can be written down, and Bellman equations corresponding to (5) derived. The structure of the optimal policy is found to be of similar form to the impulse control of [1].

Deterministic Models

Although our discussion so far has been based on stochastic models, similar considerations apply when

dealing with deterministic models. As an illustration, consider the general deterministic SIR epidemic. This is described by the differential equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\beta xy, \frac{\mathrm{d}y}{\mathrm{d}t} = \beta xy - \gamma y, \tag{6}$$

where $x(t)$, $y(t)$ denote, respectively, the numbers of susceptible and infective individuals in the population at time t. We can, for instance, control the spread of infection by increasing the removal rate parameter from γ to $\gamma + v(t)$, where $0 \leq v(t) \leq 1$. This control problem has been studied in [13], where the total cost was taken to be $\int_0^\infty (1 + kv(s))y(s)\,\mathrm{d}s$. As for the corresponding stochastic problem, it is found that the optimal control is of bang-bang type; that is, $u(t)$ should always be taken to be either 0 or 1.

Extensions

In the above examples, quite simple cost functions have been used, as much for mathematical convenience as from considerations of realism. Consider first the cost due to infection. In [1, 2, 3], this was taken to be proportional to the number of individuals becoming infected, whereas in [13], it was taken to be proportional to the total time spent in an infectious state by all individuals during the epidemic. (If we assume that infectious individuals are those displaying symptoms and unable to work, then this latter cost function has an economic interpretation as the number of hours work lost due to the infection.) More generally, the cost due to infection could be some more complicated function of the number of individuals to become infected and of the time spent in the infectious state. Other possible contributions to the cost function include the total duration of the epidemic and the maximum number of infectives to be present at any time during the epidemic.

Secondly, consider the cost of intervention. For impulse control policies, this can most simply be taken proportional to the number of individuals isolated, or vaccinated, but in general, could be some more complicated function of the number of interventions. When we intervene by altering the rates of transitions, for instance, by increasing the removal rate parameter from γ to $\gamma + v(t)$, then the cost due to intervention could reasonably be taken proportional to either $\int_0^\infty v(s)\,\mathrm{d}s$, or $\int_0^\infty v(s)Y(s)\,\mathrm{d}s$. More generally, the cost due to intervention could be taken as $\int_0^\infty f(v(s))\,\mathrm{d}s$ or $\int_0^\infty f(v(s))Y(s)\,\mathrm{d}s$ for some nondecreasing function $f(v)$.

In practical applications, it is clearly important to ensure that the cost function used reflects the reality of the situation. This may well be difficult to achieve. Quantification of the cost of intervention, while not entirely trivial, seems relatively straightforward; quantification of the cost due to infection seems much more problematic, particularly for serious diseases. However, for nonserious infections it seems reasonable to express the cost of infection in terms of the number of working hours lost. For infections of livestock, the cost of any infection can be taken to be the monetary cost to the farmers involved. A further difficulty is that if the cost function is not sufficiently simple, then the mathematics required in order to evaluate an optimal policy analytically may well prove intractable.

For more complicated models than the general stochastic SIR epidemic, finding an optimal policy can be considerably more difficult. However, some progress has been made for various specific models. For instance, for models of SIR type, with infection rate $\beta(x, y)$ in place of βxy and removal rate $\gamma(x, y)$ in place of γy, Clancy [6] found conditions on the functions $\beta(x, y)$ and $\gamma(x, y)$ under which the results of [2], [3] for the general stochastic SIR epidemic still apply. Kyriakidis [10], [11] considered adapting the results of [2], [3] to a model for two competing diseases. Greenhalgh [7] investigated isolation policies for a model in a heterogeneous population, in which the contact rate between a pair of individuals depends upon the particular individuals involved.

For practical implementation, there remain substantial difficulties with the use of optimal control theory. First, one must specify an appropriate cost function; then the optimal policy must be found, which is often a nontrivial problem; finally, implementation of the optimal policy may be difficult, since one must keep track of the numbers of susceptible and infected individuals in the population at all times. Even if taking the simpler approach of reducing R_0 to below 1, practical problems remain. One must take account of the fact that vaccines are generally not 100% effective. The true values of parameters such as β and γ are not known, but must be estimated from data. In a control theoretic context, the issue of estimating parameter values has been addressed in [5].

For the time being, and for more serious infections whose cost is not easily quantified, the more simple-minded strategy of aiming to reduce R_0 to below 1 appears to remain the more practical approach. However, for nonserious infections or for diseases of livestock, when the cost of infection can reasonably be measured in economic terms, the control theoretic approach should ultimately lead to control policies which explicitly take into account the relative costs of infection and of intervention. The mathematics involved in evaluating such optimal policies for realistic models for specific diseases is, however, quite involved, and much remains to be done.

References

[1] Abakuks, A. (1972). Some Optimal Isolation and Immunisation Policies for Epidemics. D. Phil. Thesis, University of Sussex.

[2] Abakuks, A. (1973). An optimal isolation policy for an epidemic, *Journal of Applied Probability* **10**, 247–262.

[3] Abakuks, A. (1974). Optimal immunisation policies for epidemics, *Advances in Applied Probability* **6**, 494–511.

[4] Ball, F.G., Mollison, D. & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing, *Annals of Applied Probability* **7**, 46–89.

[5] Cai, H. & Luo, X. (1994). Stochastic control of an epidemic process, *International Journal of Systems Science* **25**, 821–828.

[6] Clancy, D. (1999). Optimal intervention for epidemic models with general infection and removal rate functions, *Journal of Mathematical Biology* **39**, 309–331.

[7] Greenhalgh, D. (1988). Some results on optimal control applied to epidemics, *Mathematical Biosciences* **88**, 125–158.

[8] Kermack, W.O. & McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London, Series A* **115**, 700–721.

[9] Koopman, J. (2002). Controlling smallpox, *Science* **298**, 1342–1344.

[10] Kyriakidis, E.G. (1995). Optimal control of two competing diseases or species, *The Mathematical Scientist* **21**, 56–66.

[11] Kyriakidis, E.G. (1999). Optimal isolation policies for controlling two competing diseases, *The Mathematical Scientist* **24**, 56–67.

[12] Müller, J., Kretzschmar, M. & Dietz, K. (2000). Contact tracing in stochastic and deterministic epidemic models, *Mathematical Biosciences* **164**, 39–64.

[13] Wickwire, K. (1975). Optimal isolation policies for deterministic and stochastic epidemics, *Mathematical Biosciences* **26**, 325–346.

[14] Wickwire, K. (1977). Mathematical models for the control of pests and infectious diseases: a survey, *Theoretical Population Biology* **11**, 182–238.

(*See also* **Decision Theory**; **Vaccine Studies**)

DAMIAN CLANCY

Epidemic Models, Deterministic

In this article we review some of the approaches, both standard and recent, to the theory of deterministic epidemic models. By deterministic we mean that one considers populations consisting of sufficiently many well-mixed individuals. At the level of the individuals, however, we do consider processes such as infection and the individual course of infection to be stochastic events. What separates the deterministic models from the approach taken in stochastic epidemic models (*see* **Epidemic Models, Stochastic**) is that we invoke a **law of large numbers** argument to lift these individual stochasticities to population determinism.

There are many ways in which an article like this could be structured. We have chosen to use four concrete infections as stepping-stones to review current issues in epidemic modeling and the types of deterministic model employed and questions studied. These infections are measles, malaria, helminths, and HIV, respectively, reflecting the order in which they were important historically in shaping epidemic theory. Of course, many of the models discussed under these headings are used in a great variety of settings and there is a certain degree of arbitrariness in the place in which they occur here. We do not intend to provide detailed mathematical results, but on the contrary stress the evolution of the subject and the practical problems faced. Our main aim is to point the interested reader to more advanced literature.

The current issues in epidemic modeling – both stochastic and deterministic – are illustrated in the recent review books by Mollison [42], Isham & Medley [33] and Grenfell & Dobson [20]. In these most of the issues touched on below are discussed and reviewed extensively. For a review of the recent mathematical deterministic theory, see [13].

Measles

As in stochastic epidemic modeling, the diseases that sparked the development of modern deterministic theory are the childhood infections, most notably measles. This arises predominantly from their large public health importance in the late nineteenth and early twentieth centuries. In late nineteenth century England a sophisticated system of **vital statistics** had been initiated by **William Farr**, and data series became available that were both reliable enough and long enough to generate hypotheses about the possible mechanisms underlying epidemic spread. It should be noted that the germ theory of infection became firmly established only after the 1880s. Germ theory is the notion that certain diseases are caused by living organisms multiplying within the host and capable of being transmitted between hosts.

The most striking aspect of measles epidemics, i.e. their regular cyclic behavior, was noticed first by Arthur Ransome around 1880. Speculation about the underlying cause centered around the availability of sufficiently many susceptible individuals of the right age-class in close enough proximity to each other. Early shimmers of the modern notion of critical community size for sustaining endemic measles [4, 5] were also present. Two factors that commonly occur in many current models to investigate epidemic spread of measles and other infections are the importance of age-structure and of periodicity in contacts. The age and school season were recognized as important as early as 1896 [25].

Against this background it was William Hamer who published a discrete time epidemic "model" for the transmission of measles [26]. The mathematical rendering was later given in [52]. Hamer's observation can be reformulated as stating that the incidence of new cases in a time interval is proportional to the product SI of the (spatial) density S of susceptibles and the (spatial) density I of infectives in the population. This assumption of mass action – in analogy to its origin in chemical reaction kinetics – is fundamental to the modern theory of deterministic epidemic modeling (*see* **Random Mixing**). When densities are constant and one would like to express the incidence in terms of numbers of individuals, the incidence is proportional to SI/N, where N is the total population size [9].

The popularity of mass action is explained because of its mathematical convenience and the fact that at low densities it is a reasonable approximation of a much more complex contact process. At higher densities it grossly overestimates contact opportunities. As a side remark we note that Frost and Reed in 1928 recognized for measles that although multiple contacts of infectives with the same susceptible can occur, this susceptible can become infected

only once. This led them to develop the stochastic Reed–Frost model that does not have this particular problem (*see* **Chain Binomial Model**). Incidentally, this problem with mass action had already been solved by P.D. En'ko long before it was introduced. He studied a model for measles transmission that is very similar to the Reed–Frost approach (see [14]). For a comparison between all approaches, see Dietz & Schenzle [16].

Currently, expressions like $\beta C(N)SI/N$ are frequently used for the incidence, where $\beta C(N)$ gives the successful contacts per unit of time for a given infective. A fraction S/N of these contacts will be with a susceptible in a homogeneously mixing population. The function $C(N)$ is taken to saturate with increasing density, reflecting the fact that contacts take time and moreover satiation occurs (think of sexual contacts and insect vectors taking blood meals; see, for example, [28], [53]). A standard compartment model for infection transmission with a **latent period** and lasting immunity (a so-called SEIR model) is

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \beta C(N)\frac{SI}{N} - \mu S, \tag{1}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta C(N)\frac{SI}{N} - (\sigma + \mu)E, \tag{2}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - (\gamma + \mu)I. \tag{3}$$

An equation for R is superfluous here since $N = S + E + I + R$ is constant in this model. A very large number of variants of this model – usually denoted by strings of S, E, I, and R – are studied in the literature for a large number of different infections. In most cases this entails analysis of the dynamic behavior with practical applications to problems concerning control of the infection (see [31] for a brief review of this area). Examples of important extensions are introducing an additional death rate due to the disease, loss of immunity with re-entering of the S-class, a birth rate directly into the infective class (vertical transmission) (see [7] for all of these), density-dependent demography, and of course heterogeneity which we will discuss later. In [34] a comprehensive treatment of the mathematical theory of compartmental systems is given. Whether certain possible additions matter depends partly on the time scale of the phenomenon one is interested in. For example, on the time scale of individual epidemic outbreaks, the population can often be regarded as being in a demographic steady state (see [12]), and the inflow of new susceptibles can often be neglected in a short enough period (closed population). Useful results include the so-called final size equations for closed populations. These relate the initial size S_0 of the susceptible population to the basic **reproduction number** R_0 (see below) and the size $S(\infty)$ of the susceptible population that remains after the epidemic has come to an end. Both S_0 and $S(\infty)$ might be observable in practical situations or estimated from population data. Estimates of R_0 can then be obtained from a final-size relation. In the case of system (1)–(3), with $C(N) = N$ (mass action) and disregarding latency (i.e. $\sigma \to \infty$), we obtain

$$\ln\frac{S(\infty)}{S_0} = R_0\left[\frac{S(\infty)}{S_0} - 1\right]. \tag{4}$$

One of the many present-day approaches towards understanding the dynamics of measles epidemics (see [6] for a review) is to include a periodic forcing $\beta(t) = b_0[1 + b_1 cos(2\pi t)]$ in the contact rate of compartmental models to mimic the increase in successful contact opportunities during school seasons (see [21]). The dynamics of these models have been extensively studied (see, for example, [49], [51]) and compared with data. Detecting nonlinearity and **chaos** from stochastic effects in data of recurrent epidemics is an important theme in this area [17].

In contrast to seasonal contact rates – that are studied in only a few other settings (see malaria section below) – the incorporation of age structure is relevant to almost all important human infections (see [3], [7]). Age-structured models come in discrete age-class systems of differential equations and continuous age systems of partial differential equations [54]:

$$\frac{\partial S}{\partial t} + \frac{\partial S}{\partial a} = -\mu S - \lambda S, \tag{5}$$

$$\frac{\partial I}{\partial t} + \frac{\partial I}{\partial a} = -\mu I + \lambda S - \gamma I, \tag{6}$$

$$\frac{\partial R}{\partial t} + \frac{\partial R}{\partial a} = -\mu R + \gamma I, \tag{7}$$

with

$$S(t,0) = \int_0^\infty b(a)N(t,a)\ \mathrm{d}a,\ \ I(t,0) = R(t,0) = 0, \tag{8}$$

and appropriate initial conditions. The so-called *force of infection* $\lambda(t)$ is defined by

$$\lambda(t) = \int_0^\infty k(a')I(t, a')\,\mathrm{d}a'. \tag{9}$$

Often the force of infection depends on the age a of the susceptible with the kernel $k(a')$ replaced by $k(a, a')$ to reflect the different mixing patterns of individuals of different ages. Extensions of models like this have been compared with data for measles, notably by Schenzle [50], and showed remarkable accuracy even in following the shift in trend after vaccination was implemented. Mathematical theory centers around dynamic behavior and stability criteria for endemic steady states for various variants of this model (see [32]).

One of the applied issues in studying age-structured models is to predict the effects of changes in vaccination strategies [18] (*see* **Vaccine Studies**). For childhood infections not all strategies of vaccinating children at various ages can result in eradication of the infectious agent from the population. For this to occur the basic reproduction number R_0 has to be less than 1 in a population where the distribution with respect to age and vaccination status is in a demographic steady state. The basic reproduction number is one of the central notions of epidemic theory. It is defined as the expected number of secondary cases caused by a single infective in a susceptible population that is in a demographic steady state. For modern theory of calculating R_0 for infections in heterogeneous populations (see [11], [27]). Consider system (5)–(9). Let the stable age distribution be given by $S(a) = S_0 e^{-ra}\mathcal{F}_d(a)$, where r is the intrinsic **population growth** rate and $\mathcal{F}_d$ is the survival function. One can show that

$$R_0 = S_0 \int_0^\infty k(a)\mathrm{e}^{-ra}\mathcal{F}_d(a)\,\mathrm{d}a. \tag{10}$$

In the article on **Reproduction Number** a relation is given for the fraction of the population that needs to be vaccinated in order to assure eradication:

$$\nu > 1 - \frac{1}{R_0}. \tag{11}$$

If we vaccinate a fraction ν at birth, then we have a susceptible population of size $S_0(1-\nu)$ and the relation is the same as above with R_0 given by (10).

Optimal vaccination policies can also take economic considerations into account using similar methods (see, for example, [24], [43]).

These models have shown that some vaccination strategies in use can have unforeseen detrimental effects in a population (see [3], for exposition and review). If fewer susceptibles are around in certain age-classes as a result of vaccination, then the average age at acquiring infection will rise. For infections such as rubella, complications can arise when the infection is contracted for the first time during pregnancy. This leads to the situation where even though rubella prevalence can decrease due to vaccination, the number of serious complications can increase as a result. It is in this type of application for childhood infections that age-structured models and their analysis have provided important epidemiologic insight that might otherwise have been difficult to obtain.

Malaria

Independently from Hamer, Ronald Ross in 1911 introduced the mass action idea in continuous time in his study of the transmission of malaria [46]. Ross's work in subsequent years (see [47]) qualify him as the true founding father of modern epidemic theory. It was partly under his influence that **Anderson McKendrick** started his own studies into the mathematical modeling of epidemic phenomena, initially also in the context of malaria and other tropical infections. His series of papers with Kermack from 1927 onwards, see below, is regarded as the foundation upon which much of modern theory rests. The papers have recently been reprinted [36].

One of the distinguishing characteristics of malaria is that the protozoan parasite is indirectly transmitted between humans by mosquitoes. Several important human infections depend on similar vectors for their transmission. For modeling this brings about a new problem in that the population dynamics of the vector have to be described. A simple model capturing the essentials is (see [3])

$$\frac{\mathrm{d}x}{\mathrm{d}t} = ab\frac{M}{N}y(1-x) - \gamma x, \tag{12}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = acx(1-y) - \mu y, \tag{13}$$

where x and y are the fractions of infected humans and mosquitoes, respectively, and where M/N is the

number of (female) mosquitoes per human host in an infection free steady state. Common extensions in a similar vein are seasonality in vector emergence and the incorporation of heterogeneity in the human population. The basic reproduction number for the above model is given by

$$R_0 = \frac{M}{N}\frac{a^2bc}{\gamma\mu}. \tag{14}$$

Since the work of Ross and notably Macdonald [38], eradication campaigns have been aimed at controlling the mosquito population strongly enough to achieve $R_0 < 1$. With the emergence of both vector resistance to chemical control and parasite resistance to drug treatment malaria is regaining its strength as arguably the most severe infectious disease of man.

There has been much debate over the possibility of eradicating malaria by vaccination. We have seen that the minimum proportion of hosts that must be vaccinated to prevent a disease from establishing, or to eradicate it when present in the population, is $1 - 1/R_0$. Estimates of $R_0 > 80$ based on the age at first infection would then imply that eradication by vaccine would be a hopeless task. However, Gupta & Day [22] have suggested that malaria could be composed of several strains, each of which confers strain-specific lifelong immunity. This observation is consistent with sero prevalence data obtained from the Gambia. The age at first infection would then be

$$A = \frac{L}{\sum_j R_0^j}, \tag{15}$$

where the R_0^j are the basic reproduction numbers of the individual strains, and L is the life expectancy of the host. The proportion that must be vaccinated is then $1 - 1/\max(R_0^j)$. As max (R_0^j) could be in the range 5–10 this analysis implies that the eradication of malaria by vaccination could be a feasible proposition. See Saul [48] for a critique of this theory.

One of the most important outstanding issues in malaria and in many other infections is to understand the phenomenon of acquired immunity. Here, an individual's immune level to disease rises with frequent reinfection. It is unclear how this should be modeled mathematically. An infection pressure in the vector population gives rise to a distribution of initial doses of infection that a human receives. The infectious output towards biting mosquitoes of a given infected human is the result of a complex battle with the immune system. This output distribution then feeds back into the population. Currently the word immunoepidemiology is used to signify an area of modeling that tries to link both immunologic processes within individual hosts with epidemiologic processes of transmission between hosts. Given the implications that this interaction between immunity and epidemiology has for control strategies that affect the build-up of natural immunity, this area is likely to see much activity by epidemic modelers in the near future.

Helminths

Chronologically speaking, the tropical helminth infections such as schistosomiasis are the next step in the genesis of epidemic theory. Early work by Kostitzin in 1934 was followed 30 years later by Macdonald's study of schistosomiasis [39] and a flourishing of activity in the 1970s and 1980s (see [3] for a review).

A major difference between microparasites and macroparasites is that the former reproduce rapidly within the host, whereas the latter reproduce by releasing infective stages into the environment, which eventually complete a life cycle and (re)infect hosts. Hence, for infections caused by parasitic helminths the compartment models that classify a host as susceptible, infectious, etc. become inappropriate, and a model that allows multiple infections is required. A second problem then presents itself. The notion that the pool of susceptibles diminishes during the course of an epidemic does not necessarily hold, differential equation models no longer have a negative feedback mechanism that is automatically incorporated, and careful attention must be paid to the mechanisms that regulate the parasite population. Early models for parasites of wild animals [1] included increased mortality of the host due to parasitic infection, therefore heavily parasitized hosts had a short life expectancy, and upon dying removed large numbers of parasites from the system. For many helminth infections of humans this would not be the case, and cognisance must be taken of regulatory mechanisms such as acquired immunity.

A simple model for the dynamics of a parasitic helminth in a population of constant size would be

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mu(Q(M) - 1)M, \tag{16}$$

where M is the mean number of parasites per host, μ is the loss rate of parasites from the system and $Q(M)$ is the ratio of parasite transmission rate to loss rate. This model could stand as a prototype for many similar formulations to be found in [3] and other sources, and would be appropriate where host immunity was a function of current mean parasite burden. Hence Q is a positive nonincreasing function of M, at steady state $Q(M) = 1$, and the parasite population can persist whenever $Q(0) > 1$.

The number $Q(0)$ is the basic reproduction number (ratio, quotient) for the parasite population. It may be defined as the expected number of offspring of a typical parasite that reach reproductive maturity, in a completely susceptible host population [2]. Hence, whereas for microparasites the reproduction number is defined in terms of secondary infections of hosts, for macroparasites it is defined in terms of the parasite population dynamics. This definition has been formalized in Heesterbeek & Roberts [29].

The model presented above incorporates simplistic expressions for parasite transmission and host immunity. Many helminth parasites have complicated development stages outside the definitive host that must be modeled explicitly. For example, cestode parasites are tapeworms, with an obligatory intermediate host and transmission maintained by a carnivore–herbivore relationship (see [45]). The dominant feedback is provided by immunity to the larval stage acquired by the intermediate host. For trematodes, such as the parasites that cause schistosomiasis, there is an obligatory two host cycle (for example human/snail) with a free-living stage, and immunity to superinfection is acquired by the definitive host.

Immunity acquired against adult helminth parasites may be stimulated by larvae (hence challenge) or adult parasites, and may act against larvae by protecting the host from further infection, against adult parasites by increasing their rate of mortality, or against continued transmission by reducing their egg output. A theoretical framework for these mechanisms has been developed by Woolhouse [55]. Essentially, the model presented above was extended to include the age structure of the host to obtain

$$\frac{\partial M}{\partial t} + \frac{\partial M}{\partial a} = \mu(Q(M) - 1)M(a, t), \qquad (17)$$

where, for fixed time t, $M(a, t)$ is the density of mean parasite burden over host age a. Woolhouse [55] remarks that the function Q (in our notation) "represents the entire process of transmission between hosts, incorporating the dynamics of any free-living or vector-born stages, host exposure and innate susceptibility to infection". The inclusion of an age structure in the model was motivated by the fact that host–parasite data are often presented via age–intensity curves, and age-structured models may be used to analyze these. Assuming that the epidemiology of the parasite has remained constant for some time we set $\partial M/\partial t = 0$ and obtain predicted age–intensity relationships.

Acquired immunity may be included in a variety of ways, for example if Q were also a function of past worm burden we could have $Q(M, H)$ with $\partial Q/\partial H < 0$ and, when $\partial M/\partial t = 0$, write

$$\frac{\mathrm{d}H}{\mathrm{d}a} = M - \sigma H, \qquad (18)$$

where σ is the rate of fade of immunity. Woolhouse [56] used this framework to compare the age–intensity and age–egg output curves generated by four assumed mechanisms with data from studies of schistosomiasis infections. See [8] for the incorporation in a more comprehensive model for schistosomiasis control. For a review of the epidemiology and modeling of schistosomiasis, see [57].

Since parasite numbers within hosts are frequently too low to warrant a deterministic description at that level, models have been developed that allow for stochasticity within individuals but continue to treat the host population deterministically with an age-structured model (see [23], [37]). In addition, these models incorporate the fact that as a rule there is large variation in individual parasite burden. This implies that a description using only mean burdens is often less appropriate.

HIV/AIDS

The infection that has sparked off a tremendous increase in epidemic modeling activity is undoubtedly HIV, starting with seminal papers by Anderson and May (e.g. [40]) (*see* **AIDS and HIV**). The effect has been that in the past 10 years more different infections of humans and animals have been studied with more realistic models than ever before. Progress of the whole area of epidemic modeling is no longer attached to specific classes of infections as it was

in the early days. There is now much progress not only on the applied front, but also in mathematical advances necessary to cope with the more involved models that aim to take relevant heterogeneity in the population into account. One of the reasons could be that sexually transmitted diseases, and certainly AIDS, call for the incorporation of much structure combined. Examples are age structure for differences in infectivity, susceptibility, and certainly also contact structure and discrete characteristics such as sex, sexual preferences, and sexual activity. Two complications are particularly important: varying infectivity as a function of time elapsed since infection and long-lasting partnerships. The first complication is relevant to almost all infectious diseases, the second is related to sexual transmission. We deal with both below. One of the key notions to come out of HIV modeling is that of a core-group of infecteds. This is a small group that is very active in contacts and can keep the epidemic going in a much larger group where the internal contacts alone cannot sustain it [35] (*see* **Partner Study**).

It was realized early on in the modeling of HIV that humans generally form (sexual) partnerships that last longer than individual sexual contacts. Let us assume monogamous partnerships. During such a partnership two susceptible partners are not at risk to infection, and an infective can only cause a single new case until the partnership dissolves. This has consequences for the spread of infection. This observation has given rise to pair formation models where the formation and breaking up of partnerships is explicitly taken into account [15]. These models differentiate in the most basic case between single susceptibles and infectives and three types of pairs. If one lets partnership duration tend to zero while keeping the infection potential during a partnership constant, then one obtains the mass action models we discussed previously. In recent years progress has turned to more complicated contact structures such as circles of friends [10] and beyond deterministic theory to random graphs that reflect an underlying dynamic social contact structure in a population.

We now turn to variable infectivity. Chronologically this important aspect of deterministic epidemic theory was introduced by Kermack and McKendrick in 1927 after McKendrick's ventures into epidemiology had broadened beyond tropical infections. Their integral equation model incorporates – in modern notation – a function $A(\tau)$ to describe the average infectivity of an individual at a time τ since this individual became infected. For childhood infections, influenza, and other infections with long-lasting immunity $A(\tau)$ is a one-humped curve with narrow support. Usually the function does not rise away from zero immediately since many infections have a nonnegligible latency period between infection and becoming infectious. The latency period should not be confused with the **incubation period** which separates infection from the occurrence of symptoms. We consider incubation periods below. For HIV, the function $A(\tau)$ is typically a two-humped function. There is a peak in the early months of infection, followed by a long period of very low but probably nonzero infectivity (typically lasting several years), and ending in a second rise to high levels as the infection progresses to full-blown AIDS and ultimately death.

One striking feature of infection with the HIV virus is the long (>8 year) incubation period of AIDS. A proposed explanation is that the virus mutates within the host, with each strain stimulating both strain-specific and nonspecific immune responses; and although each strain is able to multiply in the presence of specific responses, it is regulated by the combination of specific and nonspecific responses. A simple mathematical model then shows that initially the individual virus strains are suppressed at a low level by the immune system of the host, but eventually the number of strains exceeds a "viral diversity threshold". When this occurs the nonspecific responses can no longer contain the virus population, and all viral strains are able to multiply [44]. Here a mathematical model has generated a hypothesis for the within-host dynamics of a disease that has yet to be tested against observation.

Now imagine an infection that results in complete immunity or death, in a population that is closed (i.e. inflow of new susceptibles is negligible on the time scale of the epidemic), where contacts are described by mass action. Let $S(t)$ be the density of susceptibles in the population at time t. Assume that a single infection triggers an autonomous process within the host. This allows an age representation for the infectivity. We can describe the infection process by the following integral equation:

$$\frac{\mathrm{d}S}{\mathrm{d}t}(t) = S(t)\int_0^\infty A(\tau)\frac{\mathrm{d}S}{\mathrm{d}t}(t-\tau)\,\mathrm{d}\tau. \qquad (19)$$

The incidence of new infecteds $i(t,0) = -\,\mathrm{d}S/\,\mathrm{d}t(t)$ and we can reformulate the above in terms of the incidence $i(t,\tau)$ as

$$i(t,0) = S(t)\int_0^\infty A(\tau)i(t-\tau,0)\,\mathrm{d}\tau, \qquad (20)$$

where the latter integral is the force of infection. This equation can be understood by noting that the infected individuals of infection age τ that are infecting susceptibles at time t are doing so with infectivity $A(\tau)$. These infecteds are precisely those who became infected at time $t-\tau$. Only if we choose $C(N) = N$ (mass action) and the unrealistic $A(\tau) = \beta\exp(-\gamma\tau)$, and if we neglect latency, can we reduce the above to the system of ordinary differential equations (1)–(3) by calculating $I(t) = \int_{-\infty}^{t}\exp[-\gamma(t-\tau)](\,\mathrm{d}S/\,\mathrm{d}t)(\tau)\,\mathrm{d}\tau$ and differentiating. Kermack and McKendrick already showed that the disease will spread in the population of constant density S_0 if and only if $R_0 > 1$ with

$$R_0 = S_0\int_0^\infty A(\tau)\,\mathrm{d}\tau. \qquad (21)$$

This basic result of deterministic epidemic theory can be extended to populations with arbitrary heterogeneity. Consider the individuals labeled with a variable ξ, say, taking values in some state space $\Omega \subset \mathbb{R}^m$. Now, both S and A can depend on the type of individual. The general integral equation formulation for a closed population is

$$i(t,\xi) = S(t,\xi)\int_\Omega\int_0^\infty A(\tau,\xi,\eta)i(t-\tau,\eta)\,\mathrm{d}\tau\,\mathrm{d}\eta. \qquad (22)$$

The compartmental ordinary differential equation models, with and without heterogeneity, and the age-structured model above are special cases of this equation for specific choices of A. One can show that R_0 generalizes naturally to the spectral radius of the so-called *next generation operator* associated with equation (22) [11, 13, 30].

To show the connection with age-structured models let $\Omega = [0,\infty)$ and $\xi = a$ and disregard demography. Under the condition

$$A(\tau,a,b) = k(a,b+\tau)\exp\left[-\int_b^{b+\tau}\gamma(c)\,\mathrm{d}c\right] \qquad (23)$$

define

$$I(t,a) = \int_0^a i(t-\tau,a-\tau)\exp\left[-\int_{a-\tau}^a \gamma(c)\,\mathrm{d}c\right]\mathrm{d}\tau. \qquad (24)$$

Differentiating I leads to the system (5)–(6) (with $\mu = 0$).

In the HIV/AIDS context questions relate again to evaluating effects of control measures (including behavior change). In an age-structured setting the possible demographic impact of HIV in developing countries is an important issue [41]. In Heesterbeek & Dietz [27] it is shown how models with continuous age structure in the form (22) relate to models with discrete age-classes and the so-called WAIFW matrices (Who Acquires Infection From Whom). These matrices are studied as an approach to link theory to population data [3, 19]. One can also give a final size relation for (22) in a closed population.

The modeling process within the deterministic frame of model (22) has shifted to specifying the infectivity kernel A. The modeling depends of course on the type of question to be studied. Often the modeling will make use of **stochastic processes** – notably **Markov processes** – to describe change in an individual's set of characteristics. Many models are in some sense special cases of (22). In a way, these models have already made a choice for A, as we have seen in the age-structured and compartmental model for measles. The modeling of A for more realistic (measurable) situations combining heterogeneity (in individual traits, individual behavior and space) and immunologic and evolutionary processes is one of the major challenges in the near future. This modeling will draw on deterministic but increasingly also on stochastic techniques and theory. It is likely that the days are numbered for realistic progress in theory that is only deterministic. The theory has reached a stage where progress on current issues of epidemiologic importance (e.g. understanding persistence and critical community size) can only be achieved if deterministic and stochastic theory go hand in hand.

References

[1] Anderson, R.M. & May, R.M. (1978). Regulation and stability of host-parasite population interactions. 1. Regulatory processes, *Journal of Animal Ecology* **47**, 219–247.

[2] Anderson, R.M. & May, R.M. (1982). *Population Biology of Infectious Diseases*. Springer-Verlag, Berlin.

[3] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[4] Bartlett, M.S. (1957). Measles periodicity and community size, *Journal of the Royal Statistical Society, Series A* **120**, 48–70.

[5] Bartlett, M.S. (1960). *Stochastic Population Models in Ecology and Epidemiology*. Methuen, London.

[6] Bolker, B. & Grenfell, B.T. (1995). Space, persistence and the dynamics of measles, *Philosophical Transactions of the Royal Society, Series B* **348**, 309–320.

[7] Busenberg, S. & Cooke, K. (1993). *Vertically Transmitted Diseases, Models and Dynamics*. Springer-Verlag, Berlin.

[8] Chan, M.S., Guyatt, H.L., Bundy, D.A.P., Booth, M., Fulford, A.J.C. & Medley, G.F. (1995). The development of an age structured model for schistosomiasis transmission dynamics and control and its validation for *Schistosoma mansoni*, *Epidemiology and Infection* **115**, 325–344.

[9] De Jong, M.C.M., Diekmann, O. & Heesterbeek, J.A.P. (1995). How does transmission of infection depend on population size?, in *Epidemic Models, their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge, pp. 84–94.

[10] Diekmann, O., De Jong, M.C.M. & Metz, J.A.J. (1998). A deterministic epidemic model taking account of repeated contacts between the same individuals, *Journal of Applied Probability* **35**, 448–462.

[11] Diekmann, O., Heesterbeek, J.A.P. & Metz, J.A.J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations, *Journal of Mathematical Biology* **28**, 365–382.

[12] Diekmann, O., Heesterbeek, J.A.P. & Metz, J.A.J. (1995). The legacy of Kermack and McKendrick, in *Epidemic Models, their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge, pp. 95–115.

[13] Diekmann, O. & Heesterbeek, J.A.P. (2000). *Mathematical Epidemiology of Infectious Diseases: model building, analysis and interpretation*. John Wiley & Sons, Chichester.

[14] Dietz, K. (1988). The first epidemic model: A historical note on P.D. En'ko, *Australian Journal of Statistics* **30A**, 56–65.

[15] Dietz, K. & Hadeler, K.P. (1988). Epidemiological models for sexually transmitted diseases, *Journal of Mathematical Biology* **26**, 1–25.

[16] Dietz, K. & Schenzle, D. (1985). Mathematical models for infectious disease statistics, in *A Celebration of Statistics: the ISI Centenary Volume*, A.C. Atkinson, & S.E. Feinberg, eds. Springer-Verlag New York, pp. 167–204.

[17] Ellner, S., Gallant, R. & Theiler, J. (1995). Detecting nonlinearity and chaos in epidemic data, in *Epidemic Models, their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge, pp. 229–247.

[18] Greenhalgh, D. (1990). Vaccination campaigns for common childhood diseases, *Mathematical Biosciences* **100**, 201–240.

[19] Greenhalgh, D. & Dietz, K. (1994). Some bounds on estimates for reproductive ratios derived from the age-specific force of infection, *Mathematical Biosciences* **124**, 9–57.

[20] Grenfell, B.T. & Dobson, A.P., eds (1995). *Ecology of Infectious Diseases in Natural Populations*, Cambridge University Press, Cambridge.

[21] Grenfell, B.T., Bolker, B. & Kleczkowski, A. (1995). Seasonality, demography and the dynamics of measles in developed countries, in *Epidemic Models, their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge.

[22] Gupta, S. & Day, K.P. (1994). A strain theory of malaria transmission, *Parasitology Today* **10**, 476–481.

[23] Hadeler, K.P. & Dietz, K. (1984). Population dynamics of killing parasites which reproduce in the host, *Journal of Mathematical Biology* **21**, 45–65.

[24] Hadeler, K.P. & Müller, J. (1996). Optimal vaccination patterns in age-structured populations, in *Models for Infectious Human Diseases, Their Structure and Relation to Data*. V. Isham & G. Medley, eds. Cambridge University Press, Cambridge, pp. 90–104.

[25] Hamer, W.H. (1896). Age-incidence in relation with cycles of disease-prevalence, *Transactions of the Epidemiological Society of London* **XVI**(1896–97), 64–77.

[26] Hamer, W.H. (1906). Epidemic disease in England, the evidence of variability and of persistency of type, *Lancet* **i**, 733–739.

[27] Heesterbeek, J.A.P. & Dietz, K. (1996). The concept of R_0 in epidemic models, *Statistica Neerlandica* **50**, 89–110.

[28] Heesterbeek, J.A.P. & Metz, J.A.J. (1993). The saturating contact rate in marriage and epidemic models, *Journal of Mathematical Biology* **31**, 529–539.

[29] Heesterbeek, J.A.P. & Roberts, M.G. (1995). Threshold quantities for helminth infections, *Journal of Mathematical Biology* **33**, 415–434.

[30] Heesterbeek J.A.P. (2002). A brief history of R_0 and a recipe for its calculation. *Acta Biotheoretica* **50**, 189–204.

[31] Hethcote, H.W. (1994). A thousand and one epidemic models, in *Frontiers in Theoretical Biology*, S.A. Levin, ed. *Lecture Notes in Biomathematics*, Vol. 100, Springer-Verlag, New York, pp. 504–515.

[32] Inaba, H. (1990). Thresholds and stability results for an age-structured epidemic model, *Journal of Mathematical Biology* **28**, 411–434.

[33] Isham, V. & Medley, G., eds (1996). *Models for Infectious Human Diseases, Their Structure and Relation to Data*. Cambridge University Press, Cambridge.

[34] Jacquez, J.A. (1997). *Compartmental Models*. BioMedware, Ann Arbor.

[35] Jacquez, J., Simon, C. & Koopman, J. (1995). Core groups and the R_0's for subgroups in heterogeneous SIS and SI models, in *Epidemic Models, their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge, pp. 279–301.

[36] Kermack, W.O. & McKendrick, A.G. (1927). Contributions to the mathematical theory of epidemics, part I, *Proceedings of the Royal Society, Series A* **115**, 700–721. Reprinted (along with parts II and III), *Bulletin of Mathematical Biology* **53**(1991) 33–55.

[37] Kretzschmar, M. (1989). A renewal equation with a birth-death process as a model for parasitic infections, *Journal of Mathematical Biology* **27**, 191–221.

[38] Macdonald, G. (1957). *The Epidemiology and Control of Malaria*. Oxford University Press, Oxford.

[39] Macdonald, G. (1965). The dynamics of helminth infections, with special reference to schistosomes, *Transactions of the Royal Society for Tropical Medicine* **59**, 489–506.

[40] May, R.M. & Anderson, R.M. (1988). The transmission dynamics of human immunodeficiency virus (HIV), *Philosophical Transactions of the Royal Society, Series B* **321**, 565–607.

[41] May, R.M., Anderson, R.M. & McLean, A. (1988). Possible demographic consequences of HIV/AIDS epidemics: I. Assuming HIV infection always leads to AIDS, *Mathematical Biosciences* **90**, 475–505.

[42] Mollison, D. (ed.) (1995). *Epidemic Models, their Structure and Relation to Data*. Cambridge University Press, Cambridge.

[43] Müller, J. (1994). *Optimal Vaccination Patterns in Age-structured Populations*. PhD Thesis, University of Tübingen.

[44] Nowak, M.A. & May, R.M. (1991). Mathematical biology of HIV infections: antigenic variation and diversity threshold, *Mathematical Biosciences* **106**, 1–21.

[45] Roberts, M.G. (1994). Modelling of parasitic populations: Cestodes, *Veterinary Parasitology* **54**, 145–160.

[46] Ross, R. (1911). *The Prevention of Malaria*, 2nd Ed. John Murray, London.

[47] Ross, R. & Hudson, H.P. (1917). An application of the theory of probabilities to the study of a priori pathometry, part III *Proceedings of the Royal Society of London, Series A* **43**, 225–240.

[48] Saul, A. (1996). Transmission dynamics of Plasmodium falciparam, *Parasitology Today* **12**, 74–79.

[49] Schaffer, W.M. (1985). Can nonlinear dynamics elucidate mechanisms in ecology and epidemiology?, *IMA Journal of Mathematics Applied to Medicine and Biology* **2**, 221–252.

[50] Schenzle, D. (1984). An age-structured model of pre- and post vaccination measles transmission, *IMA Journal of Mathematics Applied to Medicine and Biology* **1**, 169–191.

[51] Schwarz, I.B. (1985). Multiple stable recurrent outbreaks and predictability in seasonally forced nonlinear epidemic models, *Journal of Mathematical Biology* **21**, 347–361.

[52] Soper, M.A. (1929). The interpretation of periodicity in disease prevalence, *Journal of the Royal Statistical Society, Series A* **92**, 34–61.

[53] Thieme, H. (1992). Epidemic and demographic interaction in the spread of potentially fatal diseases in growing populations, *Mathematical Biosciences* **111**, 99–130.

[54] Webb, G.F. (1985). *Theory of Nonlinear Age-dependent Population Dynamics*. Marcel Dekker, New York.

[55] Woolhouse, M.E.J. (1992). A theoretical framework for the immunoepidemiology of helminth infection. *Parasite Immunology* **14**, 563–578.

[56] Woolhouse, M.E.J. (1994). Immunoepidemiology of human schistosomes: taking the theory into the field, *Parasitology Today* **10**, 196–202.

[57] Woolhouse, M.E.J. (1994). Epidemiology of human Schistosomes, in *Parasitic and Infectious Diseases: Epidemiology and Ecology*. M.E. Scott & G. Smith, eds. Academic Press, San Diego, pp. 197–217.

(*See also* **Epidemic Models, Spatial**; **Epidemic Thresholds**; **Infectious Disease Models**; **Mathematical Biology, Overview**)

J.A.P. HEESTERBEEK & M.G. ROBERTS

Epidemic Models, Inference

Introduction

Statistical inference for epidemics is most often based on stochastic epidemic models (*see* **Epidemic Models, Stochastic**). A special property of such models is that individuals are dependent in that the chance of getting infected depends on whether or not other individuals are infected. When making inference, another complicating property is that most often the underlying epidemic process is only partially observed. It is very rare that information about who infected whom is available. The most common type of data actually consists of only knowing who was infected and who was not, that is, having no information about the time evolution of the spread. This type of data is called *final size data*.

In the present overview, we present inference procedures for what is known as the *general epidemic model*, which assumes a homogeneous community, and a model for a structured community (*see* **Epidemic Models, Structured Population**) in which the community is partitioned into households. Which inference procedure to use depends on the underlying model, but also on the type of available data. Below, both **maximum likelihood** and martingale methods are used on the general epidemic model, depending on the type of data. Further, in a separate section, **Markov chain Monte Carlo** (MCMC) methods for more complex models, having other structured communities or partial observations, are discussed.

Outbreak in a Homogeneous Community

Below, we present inference procedures for the general epidemic model. It assumes a community of homogeneous individuals that mixes uniformly. One way to relax the assumption of homogeneity is to allow for different types of individual, where different types may have different susceptibility, infectivity, and/or mixing patterns. Inference procedures for such extended models can for example be found in [10], where inference for a multitype epidemic in a closed community is considered, or Farrington et al. [18], who consider estimation procedures for an endemic situation where types corresponds to age-cohorts.

The general epidemic is an SIR model (*see* **SIR Epidemic Models**) for a closed community. Let $S(t)$, $I(t)$, and $R(t)$ respectively denote the number of susceptible, infectious, and removed (= recovered and immune) individuals, at time t, and let n denote the community size. One way to define the general epidemic is by specifying the intensities for the two **counting processes** $N(t) = n - S(t)$ (the number of individuals who have been infected) and $R(t)$: given the process at time t, the rate of new infections (the intensity for $N(t)$) is $\lambda_N(t) = \beta\overline{S}(t)I(t)$, where $\overline{S}(t) = S(t)/n$, and the rate of removals (the intensity for $R(t)$) is $\lambda_R(t) = \gamma I(t)$; see [2] for theory on counting processes. The parameter β is hence the rate at which an infectious individual has contact with other individuals, so $\beta\overline{S}(t)$ is the rate at which he or she infects other individuals since only susceptible individuals can be infected. The parameter γ is the recovery rate of infectious individuals, so $1/\gamma$ is the average length of the infectious period.

Complete Data

First, we sketch how to perform inference assuming the epidemic process is observed continuously – so-called complete data. If $(S(u), I(u), R(u))$, or equivalently $(N(u), R(u))$ is observed continuously up to time t, then the log-likelihood is given by

$$\begin{aligned}\ell(\beta,\gamma) = &\int_0^t [\log(\beta\overline{S}(u-)I(u-))\,\mathrm{d}N(u)\\ &- \beta\overline{S}(u)I(u)\,\mathrm{d}u]\\ &+ \int_0^t [\log(\gamma R(u-))\,\mathrm{d}R(u) - \gamma R(u)\,\mathrm{d}u].\end{aligned} \tag{1}$$

The first term of each integral above is actually a sum. The counting process $N(u)$ increase 1 unit at a time making $\mathrm{d}N(u) = 1$ at these time instants and $\mathrm{d}N(u) = 0$ otherwise. The first term of the first integral is hence the sum of $\log(\beta\overline{S}(u-)I(u-))$ evaluated at these time instants, and similarly, for the first term of the second integral.

From this, the maximum likelihood estimates can be derived and shown to equal:

$$\hat{\beta}_{\mathrm{ML}} = \frac{N(t)}{\int_0^t \overline{S}(u)I(u)\,\mathrm{d}u}, \tag{2}$$

$$\hat{\gamma}_{\mathrm{ML}} = \frac{R(t)}{\int_0^t I(u)\,\mathrm{d}u}. \tag{3}$$

Standard errors can also be derived using large population results from the general epidemic (e.g. [8]). The most important parameter, the basic **reproduction number** R, for the general epidemic is given by $R = \beta/\gamma$, so the maximum likelihood estimator of R, given complete data, is

$$\hat{R}_{\mathrm{ML}} = \frac{\hat{\beta}_{\mathrm{ML}}}{\hat{\gamma}_{\mathrm{ML}}} = \frac{N(t)\int_0^t I(u)\,\mathrm{d}u}{R(t)\int_0^t \overline{S}(u)I(u)\,\mathrm{d}u}. \tag{4}$$

The critical vaccination coverage v^*, the community proportion necessary to vaccinate in order to obtain herd immunity assuming a 100% effective vaccine, is given by $v^* = 1 - 1/R$ (*see* **Epidemic Thresholds**). Accordingly v^* is estimated by

$$\hat{v}^*_{\mathrm{ML}} = 1 - \frac{1}{\hat{R}_{\mathrm{ML}}}. \tag{5}$$

Standard errors for $\hat{R}_{\mathrm{ML}}$ and $\hat{v}^*_{\mathrm{ML}}$ can be obtained using the **delta method**.

Final Size Data

As mentioned in the introduction, the most common type of data is final size data in which only the final state of the outbreak is observed, that is, how many were infected and how many were not. It is not possible to estimate β and γ separately for such data, since both parameters are related to time, and final size data contains no information about the time evolution of the epidemic. In fact, the log-likelihood in (1) is not observable for final size data. Instead, we use that $M_1 = N(t) - \int_0^t \beta\overline{S}(u)I(u)\,\mathrm{d}u$ and $M_2 = R(t) - \int_0^t \gamma I(u)\,\mathrm{d}u$ are martingales (*see* **Counting Process Methods in Survival Analysis**; also [2] for the underlying theory). From M_1 and M_2, we can form a new martingale such that the unobservable quantities of M_1 and M_2 cancel out. It turns out that the "right" martingale is

$$\begin{aligned} M(t) &= \int_0^t \frac{1}{\overline{S}(u-)}\,\mathrm{d}M_1(t) - \frac{\beta}{\gamma}M_2(t) \\ &= \int_0^t \frac{1}{\overline{S}(u-)}\,\mathrm{d}N(t) - \frac{\beta}{\gamma}R(t) \end{aligned} \tag{6}$$

$$\begin{aligned} &= \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{S(t)+1} \\ &\quad - \frac{\beta}{\gamma}R(t) \approx n\log\left(\frac{n}{S(t)}\right) - \frac{\beta}{\gamma}R(t). \end{aligned} \tag{7}$$

The second equality relies on the assumption that initially one individual was infectious and the rest were susceptible, that is, $(S(0), I(0), R(0)) = (n-1, 1, 0)$. At the end of the epidemic ($t = \tau$) there are no infectious individuals present, so $R(\tau) = n - S(\tau)$ and $M(\tau) \approx -n\log(1-\tilde{p}) - n(\beta/\gamma)\tilde{p}$, where $\tilde{p} = R(\tau)/n$ is the observed final proportion infected. Since M is a zero mean martingale, we can apply the method of moments to get an estimate of $R = \beta/\gamma$ from final size data:

$$\begin{aligned} \hat{R}_{\mathrm{FSD}} &= \frac{\left(\frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{n-R(\tau)+1}\right)}{R(\tau)} \\ &\approx \frac{-\log(1-\tilde{p})}{\tilde{p}}. \end{aligned} \tag{8}$$

This is the same estimator as if estimation would be based on the deterministic limit of the general epidemic (*see* **Epidemic Models, Deterministic**) where the final proportion infected p is known to solve the equation $1 - p = \exp(-Rp)$. However, in the stochastic setting we can also obtain standard errors for the estimator using martingale theory (e.g. [24]):

$$\begin{aligned} s.e.(\hat{R}_{\mathrm{FSD}}) = \Bigg[&\frac{1}{(n-1)^2} + \frac{1}{(n-2)^2} \\ &+ \cdots + \frac{1}{(n-R(\tau)+1)^2} + \frac{\hat{R}^2_{\mathrm{FSD}}}{n}\tilde{p}\Bigg]^{1/2} \Big/ \tilde{p}. \end{aligned} \tag{9}$$

The critical vaccination coverage $v^* = 1 - 1/R$ is of course estimated by $\hat{v}^*_{\mathrm{FSD}} = 1 - 1/\hat{R}_{\mathrm{FSD}}$ from final size data. Standard errors can as before be obtained by applying the delta method.

The maximum likelihood estimate of R, and hence also of v^*, given final size data can in principle be derived using formulae for the final size distribution (e.g. [5, pp. 93, 94]). However, these formulae quickly become cumbersome for large communities, making such inference computationally involved and numerically unstable.

Outbreak in a Community of Households

We now present inference procedures in a different setting where individuals reside in households and where it is believed that infection rates are much higher between individuals of the same households than between individuals of different households. We do this for a fairly simple model originating from Longini and Koopman [20], where households are treated as if they were independent. Since then these ideas have been refined in several ways, for example, by allowing individuals of different types and/or treating a fully stochastic model where households are dependent (e.g. [1], [6], [11] and [21]). The key idea in the Longini–Koopman model [20] is to treat the probability of getting infected from outside the household during the course of the epidemic as a parameter. In reality, this probability depends on the number of individuals who get infected and is hence a stochastic quantity, but the simplifying assumption reduces computational complexities tremendously. Further, by estimating the parameter it will be close to its "correct" value.

A Simple Household Model

Individuals reside in households. An individual who gets infected has infectious contacts with other individuals in the household independently and with equal probability $p_W = 1 - q_W$. Additionally, each individual receives an infectious contact from outside the household with probability $p_B = 1 - q_B$ (the indices stand for within and between households). Individuals who receive at least one infectious contact from infected household members, or from outside the household, get infected. Only those who escape infectious contacts both from within and outside the household avoid getting infected during the epidemic outbreak. Let $p_h(j); j = 0, \ldots, h$ denote the probability that j individuals get infected in a household having h (initially susceptible) individuals. Then these probabilities can be derived recursively from the following equations:

$$p_h(j) = \binom{h}{j} q_W^{j(h-j)} q_B^{h-j} - \sum_{r=0}^{j-1} \binom{h-r}{j-r} \times p_h(r) q_W^{(h-j)(j-r)} \qquad j = 0, \ldots, h, \quad (10)$$

(e.g. [1]). For example, $p_h(0) = q_B^h$ and $p_h(1) = \binom{h}{1}(1 - q_B)q_B^{h-1}q_W^{h-1}$, which can easily be explained. No one gets infected if everyone escapes infection from outside. One individual gets infected if 1 out of h gets infected from outside, and the remaining $h - 1$ individuals escape infection both from outside and from the infected household member. The probabilities quickly become complicated as the requested number of infected increases, but for households smaller than say 5 or even 10, formulae for them can be obtained using a computer algebra package.

Inference for the Simple Household Model

Inference is quite straightforward once the relevant $p_h(j)$'s have been calculated, since households were assumed independent. Let $\{n_h(j)\}$ denote the collected data, where $n_h(j)$ denotes the observed number of households of size h in which j individuals got infected during the epidemic. Then the log-likelihood for the data is simply

$$\ell(q_W, q_B) = \sum_{h,j} n_h(j) \log(p_h(j)), \qquad (11)$$

where the dependence on the parameters is implicit from the definition of $\{p_h(j)\}$ in (10). The parameters are simply estimated by maximizing the log-likelihood with respect to q_W and q_B. Because households are assumed independent, standard large population theory is applicable when the number of households is large, and the maximum likelihood estimators are **consistent**. Standard errors for the estimates can be obtained from the observed **information matrix** by differentiating the log-likelihood twice (e.g. [13]).

As the model is defined, there is no basic reproduction number R, because households behave independently. In Ball et al. [6], a related fully stochastic model is considered, enabling estimation of the basic reproduction number R.

Inference Using MCMC Methods

In previous sections, we have mainly treated models and data for which it was possible to derive expressions for outcome probabilities. In more realistic (i.e. complex) settings, this may not be practically possible. Often, the detail in the data does not allow

for straightforward estimation of parameters. Then some **missing data** method can sometimes be helpful. There are a few examples where the **EM algorithm** can be helpful (see e.g. [4]), but here we focus on MCMC methods (e.g. [17]). This methodology has been successfully applied in a few situations but its real breakthrough in epidemic inference still lies ahead.

The main idea of MCMC analysis in epidemic inference is to explore the outcome space of unobserved (latent) variables for which inference procedures would have been much easier, had these variables been observed (*see* **Instrumental Variables**). Most often, uninformative priors are used for model parameters, but in specific cases prior knowledge can of course be incorporated into informative **prior distributions**. Below, we list some inference problems where MCMC methods have been applied, and refer to listed references for details.

Inference is nontrivial even for the general epidemic model when the removal times, but not the infection times, are observed. Such data is quite common since the removal time of an individual is approximately the same as detection time, which is quite often known. The reason for the complication is that the likelihood then has to be integrated over all possible infection times, a time-consuming task even for very small community sizes. In O'Neill and Roberts [23], this problem is analyzed using MCMC methods in which the Markov chain explores the space of possible infection times. (A different approach, using martingales, is performed in [9].)

Also for household data, detection times but not infection times may sometimes be available. For a model allowing a fairly general distribution for the infectious period, perhaps preceded by a latency period, inference is complicated even for households of size two and when treated as independent. In O'Neill et al. [22], such data is analyzed using MCMC methods, where the unobserved infection times and latency periods are explored in the Markov chain.

It is of course hard to include all heterogeneities into a model. For example, to determine all social connections between individuals in a community is impossible. A way out of this problem is to model unknown social structures by introducing unobserved random social contacts. In [12], a first step in this direction was taken by modeling the social structure using a random graph, and assuming that transmission may only occur between neighboring individuals of the graph. Inference is performed without assuming any information about the social graph, and the Markov chain explores the possible graphs, where detection times close in time increase the probability of a social link between the corresponding pair of individuals.

Concluding Remarks

The emphasis of this article has been on inference procedures for epidemic models in general, rather than on models for specific diseases. The methods are suited for diseases in which transmission occurs by person-to-person contact, and not for vector-borne diseases like malaria or **infectious diseases** caused by contaminated water or food like salmonella. Examples of such diseases are childhood diseases like measles and mumps, smallpox, HIV (although heterogenous structures tend to be very complex here; *see* **AIDS and HIV**), influenza, and common cold.

We have described inference procedures for a few stochastic epidemic models. In many applications, the underlying setting is too complicated to enable inference from stochastic models, for example when long term endemic situations are considered and the community changes dynamically, or when there are too many types of heterogeneities. Then data can be calibrated to deterministic models thus giving parameter estimates. A thorough treatment of many such situations is given in [3] (*see also* **Epidemic Models, Deterministic**). Inference using stochastic models, as opposed to deterministic, has the advantage that it provides uncertainty estimates of parameters. Stochastic models are also better suited for situations where small social units, such as households, play an important role in the disease spread. In this case deterministic models, relying on large population results, may give misleading results. Deterministic models on the other hand, have the clear advantage of being simpler to analyze, thus permitting more complex models to be used.

The practical problem to estimate the effect of a vaccine against an infectious disease, the vaccine efficacy, is not treated in the present article (*see* **Vaccine Studies**). Clearly, this is an important inferential problem within infectious disease epidemiology, but it is left out from the presentation as epidemic models play a minor role in such analyses. Estimation

procedures for such problems can, for example, be found in [19] and [15] and the references therein.

For more detailed presentations on statistical inference for epidemic models, we recommend the monographs by Becker [7] and Andersson and Britton [4], and the survey paper [8]. More on epidemic models in general can be found in [5], [3], [14], and [16].

References

[1] Addy, C.L., Longini, I.M. & Haber, M. (1991). A generalized stochastic model for the analysis of infectious disease final size data, *Biometrics* **47**, 961–974.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[3] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans; Dynamic and Control*. Oxford University Press, Oxford.

[4] Andersson, H. & Britton, T. (2000). *Stochastic Models and Their Statistical Analysis*, *Springer Lecture Notes in Statistics*, Vol. 151. Springer, New York.

[5] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

[6] Ball, F., Mollison, D. & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing, *Ann. Applied Probability* **7**, 46–89.

[7] Becker, N.G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

[8] Becker, N.G. & Britton, T. (1999). Statistical studies of infectious disease incidence, *Journal of Royal Statistical Society. B* **61**, 287–307.

[9] Becker, N.G. & Hasofer, A.M. (1997). Estimation in epidemics with incomplete observations, *Journal of the Royal Statistical Society B* **59**, 415–429.

[10] Britton, T. (2001). Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage, *Journal of the Royal Statistical Society B* **63**, 705–715.

[11] Britton, T. & Becker, N.G. (2000). Estimating the immunity coverage to prevent epidemics in a community of households, *Biostatistics* **1**, 389–402.

[12] Britton, T. & O'Neill, P.D. (2002). Bayesian inference for stochastic epidemics in populations with random social structure, *Scandinavian Journal of Statistics* **29**, 375–390.

[13] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[14] Daley, D.J. & Gani, J. (1999). *Epidemic Modeling: an Introduction*. Cambridge University Press, Cambridge.

[15] Datta, S., Halloran M.E. & Longini I.M. (1999). Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual or household, *Biometrics* **55**, 792–798.

[16] Diekmann, O. & Heesterbeek, J.A.P. (2000). *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Wiley, Chichester.

[17] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[18] Farrington, C.P., Kanaan, M.N. & Gay, N.J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data, *Applied Statistics* **50**, 251–292.

[19] Halloran, M.E., Haber, M. & Longini, I.M. (1992). Interpretation and estimation of vaccine efficacy under heterogeneity, *American Journal of Epidemiology* **136**, 328–343.

[20] Longini, I.M. & Koopman, J.S. (1982). Household and community transmission parameters from final distributions of infections in households, *Biometrics* **38**, 115–126.

[21] Lyne, O.D. & Ball, F.G. (1999). Parameter estimation for SIR epidemics in households. Bull. Int. Statist. Inst. 52nd Session, Contributed Papers, Vol. LVIII, Book 2, p 251.

[22] O'Neill, P., Balding, D., Becker, N.G., Eerola, M. & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods, *Applied Statistics* **49**, 517–542.

[23] O'Neill, P. & Roberts, G. (1999). Bayesian inference for partially observed stochastic epidemics, *Journal of the Royal Statistical Society A* **162**, 121–129.

[24] Rida, W.N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model, *Journal of the Royal Statistical Society B* **53**, 269–283.

TOM BRITTON

Epidemic Models, Multi-strain

Multi-strain epidemics typically refer to the simultaneous development of infectious processes in a population, caused by different strains of a pathogenic agent, where the fact of having been infected by one or more of the agents may modify the sensitivity to infection by the other agents. In general, as in the single strain situation, the focus may be either on the endemic or on the epidemic timescale. Typical "endemic" questions are – which strains will dominate in the long run, whether sustained oscillations are possible (*see* **Epidemic Models, Recurrent**), or whether new strains will be able to gain a foothold in the population, in the presence of already established strains. In the epidemic perspective, the main question usually is what happens upon introduction of the diseases in a human population and what composition outbreaks will have, in terms of the different strains.

The endemic aspect is usually dealt with in deterministic, differential equation, models based on the classic **SIR** (susceptible $\rightarrow$ infective $\rightarrow$ removed) model, but with an I and R class for each one of the strains and demographic dynamics (*see* **Epidemic Models, Deterministic**). The description of the effects of previous infections on the sensitivity to further infections may be complicated (the phenomenon is often referred to as cross-immunity, in the literature), and the behavior of the resulting models also seems to be quite complicated (for some recent studies, see e.g. [2, 6]).

We will explain the epidemic case in some detail, starting with two mutually exclusive strains that confer immunity after infection. This is the only situation for which some definite results seem to be available (see [5, 7]). The evolution of the two infections in a population may be modeled by a system of differential equations, in the same style as the classical SIR epidemic. The population size is assumed to remain constant (closed population), $s(t)$ represents the proportion of susceptible individuals, $i_1(t)$ and $i_2(t)$ the proportions currently infected and infective with each one of the two strains, and $r_1(t)$ and $r_2(t)$, the respective proportions of removed individuals, that is, immune to further infection after having been infected by one of the two strains. Infection with one strain is assumed to confer immunity to further infection from the same strain, but also from the other strain. Each infection lasts, on average, $1/\mu_i$ time units ($i = 1,2$) and each infected is assumed to cause β_i ($i = 1,2$) new, secondary cases per time unit, in a completely susceptible population. The equations representing the evolution are (it is customary to write the involved functions without the time argument, since time is not explicitly used in the description; a so called autonomous system)

$$\begin{aligned} s' &= -(\beta_1 i_1 + \beta_2 i_2)s \\ i_1' &= \beta_1 i_1 s - \mu_1 i_1 \\ i_2' &= \beta_2 i_2 s - \mu_2 i_2 \\ r_1' &= \mu_1 i_1 \\ r_2' &= \mu_2 i_2 \end{aligned} \tag{1}$$

with initial conditions $s(0) = 1 - \varepsilon_1 - \varepsilon_2$, $i_1(0) = \varepsilon_1$, $i_2(0) = \varepsilon_2$, $r_1(0) = r_2(0) = 0$, where ε_1 and ε_2 are assumed to be small (a few initial infectives enter a wholly susceptible population).

This system does not admit explicit solutions, but its properties may be studied in a qualitative way. Eventually, the numbers of infective will tend to zero and the spread will stop. This will leave the population with a proportion $s(\infty)$ of individuals not having been infected during the epidemic and proportions $r_1(\infty)$ and $r_2(\infty)$ having had, respectively, disease 1 or 2. The question whether one or both diseases may give rise to large outbreaks is answered by considering the reproductive values $\theta_i = \beta_i/\mu_i$ ($i = 1,2$) of the two diseases (these are the same as the basic **reproduction number** R_0 in models of one disease) and the **epidemic threshold** conditions $\theta_i = 1$ ($i = 1,2$). If both parameters are below or on threshold, both diseases die out quickly after introduction and no large outbreaks may arise. If one is above and the other one is below threshold, there will be a large outbreak of the first one and few cases of the other one. The large outbreak will have affected a proportion $r(\infty)$ of the population, where $r(\infty)$ is the largest solution of the final size equation (see [4] for this and many other useful results on epidemic models)

$$r = 1 - \exp(-\theta r) \tag{2}$$

(with the appropriate θ-parameter and assuming that ε_1 and ε_2 are negligible, i.e. essentially zero), just as in the case with one disease in a closed population.

When both parameters are above threshold, both diseases start to spread, and using the same technique as used to derive the one-strain final size equation, one finds that the final proportions of the two diseases satisfy the equation

$$r_1(\infty) + r_2(\infty) = 1 - \exp(-\theta_1 r_1(\infty) - \theta_2 r_2(\infty)) \tag{3}$$

This equation does not have a single solution; in fact, if one thinks of the two axes in the plane as representing possible $(r_1(\infty), r_2(\infty))$ values, the equation describes a curve connecting the two solutions describing one-strain epidemics only (the other component is set to zero and the equation reduces to the one-strain case) on the two axes in the first quadrant. The equation also admits (0,0) as a solution, representing small outbreaks of both diseases. In order to determine a specific point in the plane, a second equation would be needed, but in their paper [5], Kendall and Saunders conclude that no such relation has been found.

Some recent results [7] on the asymptotics (essentially large population size) of the corresponding stochastic model (*see* **Epidemic Models, Stochastic**) cast some light on the problem. In essence, depending on parameter values, there may be no single final point, as expected from one-strain theory, but stochastic variability at the start of the spread (in deterministic theory represented by the pair (r_1, r_2) still in, or very close to, the origin) may yield "unpredictable", stochastic final size proportions distributed along the whole curve. The results are also interesting because they show that the classical parameters (θ_1, θ_2) are not sufficient to describe the final size of the spread in the population. In addition to these parameters, two new parameters $\zeta_i = \beta_i - \mu_i$ $(i = 1,2)$, the "initial speeds of spread", and the numbers of initial infectives need to be considered. In order to understand the results, it is useful to recall some classical results on the time course of a stochastic epidemic in large population, first strictly given by Barbour [3] (see also [1] for stochastic epidemic models in general). The stochastic model itself uses the same terminology and parameters as the deterministic model (1) in the two-strain case, with a corresponding natural reformulation in terms of just s, i and r in the one-strain case. However, individuals are thought of as remaining infected for a random, **exponentially distributed**, time with average $1/\mu$ and the term *bis* in the (1) is seen as the intensity with which a susceptible becomes infected, adopting a standard **Markov process** formulation. It is also more natural in the stochastic model, to have the numbers of susceptible, infective, and removed individuals as basic quantities, described as integers. The basic difference between a deterministic and a stochastic one-strain model is that, if $\theta \leq 1$, both models predict, in large populations, only small outbreaks $(r(\infty) \approx 0)$, while, if $\theta > 1$, the deterministic model predicts that there will always be a large outbreak with $r(\infty)$ close to the solution of (2), while the stochastic model asserts that the outbreak may still be small (with a certain probability that can be determined from the parameters) due to chance, but if it becomes large, $r(\infty)$ will again be close to the solution of (2). Associated with these possibilities, there is a description of the time course. If the outbreak is small, the disease dies out rather quickly, and the distribution of this time is independent of population size. If the outbreak is large, the numbers of infectives (and removed) grow exponentially, as a **branching process**, for a time that is proportional to the logarithm of population size, until the epidemic reaches a size of the same order of magnitude as the population size, then moves quickly, essentially following the deterministic differential equations, through the culmen of the epidemic, then again, now being very close to the final size, spend a time proportional to the logarithm of population size until effective die-out of the disease is achieved.

The asymptotic behavior of the two-strain model can now be explained in the following way, omitting some details. First, if one or both the θ-parameters are below threshold, the two epidemics (one for each strain) will behave as independent one-strain epidemics, according to the behavior explained above. If both parameters are above threshold, chance may still lead to a small outbreak of one of the strains; in that case, the other strain may give rise to either a small or a large outbreak, as in a one-strain model. These three possibilities (both small or one small and one large) have well-defined probabilities and correspond to the solutions (0,0) and the two points on the two axes of the (3). Finally, if both epidemics start growing, the fastest one (with the largest ζ-parameter), which we assume is denoted as type 1, completes its course as if it were alone in the population, reaching its final proportion $r_1(\infty)$, which is the solution of (2) with the parameter θ_1. The slower strain will

then achieve a final proportion corresponding to a redefined θ-parameter equal to $(1 - r_1(\infty))\theta_2$ (representing the effective secondary cases that can be achieved in the susceptible population left untouched by the first strain). This point is a point on the curve defined by (3). Finally, if initial speeds are equal, both epidemics reach the same order of magnitude as the total population at the same time, but the initial proportions of infectives have now been modified by stochastic effects coming from the exponential growth phase. With these new, random, proportions as starting points, the two epidemics follow the corresponding trajectory of the differential equations (1) to a corresponding point on the final curve (3) (there is a one-to-one mapping between starting points $\neq (0, 0)$ and points on the curve). Thus, the random distribution of the "population-size" starting points maps to a random distribution on the curve. This distribution does not have an explicit expression, but can be computed numerically. A notable case is when, in addition to equal speeds, the two strains also have equal θ-parameters. Then, the curve (3) is a straight line and, if one initial infective of each strain is assumed, the above described distribution becomes uniform on this line.

There are two main directions in which the model could be made more general. One is allowing more than two strains. As long as each one confers immunity to all the others, the results should be in line with those for two strains, only more complicated to state. A more significant generalization would be to consider other forms of interaction than complete exclusion, that is, allowing susceptibility to one strain to depend in a general way on the previous infections, maybe only decreasing it slightly or even increasing it. At the present time, there seem to be no results about such epidemic models.

References

[1] Andersson, H. & Britton, T. (2000). Stochastic epidemic models and their statistical analysis, *Lecture Notes in Statistics 151*. Springer-Verlag, New York.

[2] Andreassen, V., Lin, J. & Levin, S.A. (1997). The dynamics of cocirculating influenza strains conferring partial cross-immunity, *Journal of Mathematical Biology* **35**, 825–842.

[3] Barbour, A. (1975). The duration of the closed stochastic epidemic, *Biometrika* **62**, 477–482.

[4] Diekmann, O. & Heesterbeek, J.A.P. (2000). Mathematical epidemiology of infectious diseases, *Model Building, Analysis and Interpretation*. John Wiley & Sons, Chichester.

[5] Kendall, W.S. & Saunders, I.W. (1983). Epidemics in competition II: the general epidemic, *Journal of the Royal Statistical Society, Series B* **45**, 238–244.

[6] Pugliese, A. (2002). On the evolutionary coexistence of parasite strains, *Mathematical Biosciences* **177**(178), 355–375.

[7] Svensson, Å. & Scalia Tomba G. (2001). Competing Epidemics in Closed Populations. Research Report 2001:8, Department of Mathematical Statistics, University of Stockholm, Sweden.

GIANPAOLO SCALIA-TOMBA

Epidemic Models, Recurrent

Observations of the number of childhood infections, such as measles, are available over long time periods from a large number of cities in developed countries. Data from the time before large-scale vaccination was introduced show two interesting properties. One is recurrence of infection outbreaks, and the other one is spontaneous disappearance of the infection in small populations. It is a classical challenge in mathematical epidemiology to understand the mechanisms that cause these phenomena. A description of the work in this area will follow the two lines formed by the competing branches of deterministic and stochastic modeling (*see* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**). A major finding is that the deterministic model is a poor approximation of the stochastic one. All models for recurrent epidemics are based on the idea that an inflow of new susceptibles is necessary before a new outbreak can occur. This goes back to the model formulated by Soper [38], based on ideas by Hamer [15]. It means that the model accounts for both epidemic and demographic forces.

The Classic Endemic Model – Deterministic Version

The deterministic model given by the following system of differential equations is denoted the classic endemic model by [18]. We use this term in a broader sense, since we shall study both deterministic and stochastic versions of the same model.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \frac{\beta}{N} SI - \mu S, \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\beta}{N} SI - (\gamma + \mu) I, \tag{2}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mu R. \tag{3}$$

Here, S, I, R denote the number of individuals that are susceptible, infective, and removed, respectively. The total population size $S(t) + I(t) + R(t) = N$ is constant, provided $S(0) + I(0) + R(0) = N$. Four parameters are used, namely, the contact rate β, the death rate μ, the recovery rate γ, and the total population size N. This model has, with some variation in the parameterization, been studied by a number of authors, including [1, 11, 16, 17, 24, 25, 26] (*see* **SIR Epidemic Models**).

The model is two-dimensional, since R does not affect S and I. The parameter space can be simplified by the following reparameterization:

$$R_0 = \frac{\beta}{\gamma + \mu},$$
$$\alpha = \frac{\gamma + \mu}{\mu}. \tag{4}$$

Here, R_0 is the basic reproduction ratio (*see* **Reproduction Number**), and α is the ratio of life length to time infected. Note that the parameter α is large for the recurrent epidemic models. After the reparameterization, we still have four parameters. By using the parameter classification introduced in [32], we find that two of the parameters, R_0 and α, are essential, while the remaining ones, N and μ, are innocent. (A parameter is innocent if it can be eliminated by rescaling of state variables or of time). The results for the deterministic version of the model can be described with reference to the two-dimensional space of essential parameters (R_0, α), modulo rescaling if necessary.

We single out three properties of major interest, namely, the threshold (*see* **Epidemic Thresholds**), the recurrence, and the extinction.

With regard to the first of these properties, the deterministic model has a threshold at $R_0 = 1$; see [18]. This means that if R_0 is less than or equal to one, then any infection will ultimately disappear, while an endemic infection level will establish itself at

$$\overline{I} = \frac{R_0 - 1}{\alpha R_0} N \tag{5}$$

if $R_0 > 1$ and $I(0) > 0$.

The recurring outbreaks of infection are reflected in the deterministic model solutions as damped oscillations about the endemic infection level if $R_0 > 1$. The quasi-period of these oscillations is found from the **eigenvalues** of the **matrix** determined by linearization of the system of differential equations about the critical point corresponding to the endemic infection level. It is approximately given by (for

large α)

$$T_0 \approx \frac{2\pi}{\mu}\frac{1}{\sqrt{\alpha(R_0-1)}}. \tag{6}$$

The tendency to oscillate is a realistic feature, while the damping is not.

The third item of interest, the extinction of the infection for $R_0 > 1$, cannot be explained by the deterministic model.

Mainly for historical reasons, we describe also the deterministic model due to Hamer and Soper:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \frac{\beta}{N}SI,$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\beta}{N}SI - \gamma I. \tag{7}$$

This model allows for inflow of new susceptibles, but it does not allow for death of susceptible or infected individuals. It has no threshold, but it is quite similar to the classic endemic model with regard to recurrence and extinction. The model is further discussed by [5] and [30].

The Classic Endemic Model – Stochastic Version

The stochastic version of the model is a so-called **Markov** population process. This means that it is a **Markov chain** with continuous time and discrete state space, where only transitions to nearest neighbors are possible. The size of any population can only take nonnegative integer values. This is clearly more realistic than the continuous state space used for the deterministic model. The stochastic version of the model takes the form of a bivariate Markov chain $\{S(t), I(t)\}$, as described in detail by [30].

The deterministic version of the model can be derived as an approximation of the stochastic one by first scaling the state variables S and I with N and then letting N approach infinity. This explains why N, which is essential for the stochastic version, becomes innocent as we go to the deterministic version. The stochastic model has a reputation of being difficult to analyze. **Monte Carlo simulation** is therefore a useful method.

The states $(s, 0)$, where the number of infected individuals is equal to zero, correspond to absence of infection and form what is called an *absorbing class* for the **stochastic process**. Absorption can be interpreted as extinction of the infection. The stochastic process of concern here will reach the absorbing class in finite time. After that, it will be confined to this class forever. The remaining states (s, i), where $i \geq 1$, are transient. It turns out that there exists an important distribution, called the *quasi-stationary distribution*, which is a stationary distribution, conditional on nonextinction. The concept of quasi-stationarity for continuous-time Markov chains was introduced by [10]. It is supported on the transient states. It is a useful approximation of the distribution of states of the process before extinction. The long-term behavior of the state of the process can be described by two quantities: the quasi-stationary distribution and the time to extinction. (There is also a stationary distribution, without any conditioning. It is mathematically easier to deal with than the quasi-stationary distribution, but it is much less informative. It is supported on the absorbing class, and deals therefore only with susceptible individuals. It is completely independent of the parameter R_0.)

The powerful threshold result for the deterministic model is based on bifurcation for the system of differential equations (1)–(3). An extension of this result to the stochastic setting is highly desirable. It can be achieved by first noting that the deterministic threshold can be described as a partition of the parameter space where R_0 takes its values into two subsets $0 < R_0 < 1$ and $R_0 > 1$, where phase portraits for the differential equations differ qualitatively. The counterpart for the stochastic model is an identification of three subsets of the parameter space, where R_0 and N take values, in which the quasi-stationary distribution and the time to extinction differ qualitatively. Exact expressions for these quantities are not available. (And if they were, they would not be useful in this connection, since they would not show qualitative differences in different parameter regions.) Progress in this situation is made by seeking asymptotic approximations. It turns out that the asymptotic approximations for both of these quantities show qualitative differences in different parameter regions, exactly as required for an extension to the stochastic setting of the threshold result for the deterministic model.

The approach for a qualitative analysis of a stochastic model outlined in the preceding paragraph can be followed for univariate models, as exemplified by the Verhulst logistic model analyzed by [31]. There are still mathematical difficulties to overcome

for multivariate models, as exemplified by the one treated here.

The stochastic version of the Hamer–Soper model was introduced and analyzed by Bartlett in some classical papers [6, 7, 8]. Bartlett's main conclusion was that both the recurrence and the extinction phenomena, which were not well modeled by the deterministic Hamer–Soper model, could be explained with the aid of the stochastic feature.

For the recurrence, Bartlett claimed that the damping in the deterministic model was offset by random variability. This result was supported in his 1956 paper by Monte Carlo simulations. Bartlett's recurrence result was strengthened by [33]. This paper uses the important mathematical result that the deterministic model solution is an approximation of the expectation over a large number of realizations of the stochastic model with the same initial point. It is important to realize that this expectation does not necessarily share properties with the individual realizations. This is, in particular, true for the troublesome damping associated with the deterministic model solutions. Individual stochastic model realizations that avoid extinction show recurrent outbreaks with stochastically varying periods and amplitudes, and essentially without damping. One consequence of this stochastic variability is that early outbreaks are closely synchronized, but that the synchronization is weaker for later outbreaks. This implies that the expectation over a large number of such realizations gives damped oscillations. The damping associated with the deterministic model is thus explained as a measure of the stochastic variability in periods and amplitudes.

With regard to extinction, Bartlett derived an analytic approximation of the expected time to extinction that showed it to be an increasing function of the population size N. Improvements of these analytic results are given by [30]. Bartlett showed also that this qualitative result was supported by data from several cities in England and the United States: Fade-out was observed in small cities, but not in large ones. He therefore introduced the concept "Critical Community Size" (CCS) as "the size for which measles is as likely as not to fade out after a major epidemic." This constitutes a threshold result for the stochastic model, as discussed by [29, 30]. It generalizes the deterministic model threshold $R_0 = 1$ by giving a threshold value for R_0 as a function of N that approaches the value one as $N \to \infty$. The inverse of this function gives the critical community size as a function of R_0.

A crude approximation of the expected time to extinction is derived in [30]. It takes the form

$$E(\tau_Q) \approx \frac{\rho}{\mu R_0} \frac{\Phi(\rho)}{\varphi(\rho)}, \tag{8}$$

where

$$\rho = \frac{\sqrt{(R_0 - 1)N}}{\alpha}, \tag{9}$$

and where Φ and φ denote the **normal distribution** function and the normal density function, respectively. Putting the right-hand side of (8) equal to a constant defines R_0 as a function of N for fixed values of μ and α. This function is an approximation of the persistence threshold, since the latter is defined by the requirement that the expected time to extinction is constant. By putting this constant equal to K/μ and using the approximation (8) for $E(\tau_Q)$, we can derive the following crude approximations of the persistence threshold value in the case where average life time is $1/\mu = 70$ years, $K/\mu = 3$ years, and $\alpha = 1 + \gamma/\mu = 3500$: If $N = 10^4$, then $\rho \approx 2.86$, and the threshold value for R_0 is approximately 10 000. If $N = 10^5$, then $\rho \approx 1.59$, and the threshold value for R_0 is approximately 310. In these cases, therefore, the deterministic threshold value $R_0 = 1$ is a poor approximation of the persistence threshold.

We summarize by noting that the deterministic model is an unacceptable approximation of the stochastic one (unless the population size is really huge), with respect to all three of the major indicators, namely, threshold, recurrence, and extinction.

Variations and Extensions of the Classic Endemic Model

The phenomenon of recurring epidemic outbreaks has aroused a large interest among deterministic modelers. As described above, the deterministic version of the classic endemic model (or the rather similar Hamer–Soper model) is unrealistic since it predicts *damped* oscillations. A broad search for alternative mechanisms that predict *undamped* oscillations has therefore taken place. A review of the work in this area up until 1989 is given by Hethcote and Levin in [19]. Most of this work has taken the form of suggesting a variation or extension of the deterministic model. It has tended to disregard the explanation

put forward by Bartlett, and supported by the later work by Nåsell, that undamped oscillations can be explained by the stochastic version of the model, without any additional hypotheses. The review in [19] deals with five different types of variations or extensions of the deterministic version of the classic endemic model, namely, (1) Models with periodic coefficients, (2) Models with delays in the removed class, (3) Models with nonlinear incidence, (4) Models with variable population size, (5) Models with age structure. For all of these cases, deterministic models with undamped periodic solutions exist.

Among the models with periodic coefficients, particular attention has been paid to an SEIR model, with a state E (for exposed, meaning infected but not yet infective) inserted between the states S and I. The contact rate was assumed to be periodic, with a period of one year reflecting the periodicity in contact caused by the aggregation of children in schools (*see* **Seasonal Time Series**). The resulting seasonally forced SEIR model can exhibit **chaos**, as was reported in a series of papers, [4, 9, 12, 27, 34–36]. Arguments similar to those in the preceding section can be given to argue that the deterministic SEIR model with constant infection rate is a poor approximation of the corresponding fully stochastic model. The prospects that periodicity in the contact rate would improve this approximation are slim. The chaos phenomenon was studied in a very concrete way by [33], using the stochastic version of the SEIR model with periodic forcing. It was found with parameter values typical for measles that the minimum "number" of infected individuals between outbreaks could go down to the order of 10^{-11} in a population of 1 million individuals. (This low value resembles the "atto-fox" discovered by [28] in a model for rabies. One atto-fox equals 10^{-18} foxes.) The conclusion reaffirmed an easy extension of our previous finding, namely, that the deterministic approximation of the stochastic model is unacceptable. A direct consequence is that the chaos phenomenon is driven by an unacceptable mathematical approximation.

In search for additional realism, Schenzle established an SEIR model in [37] that combined age structure and seasonality in transmission. It explains prevaccination data on measles in England and Wales very closely. It is referred to as the RAS (Realistic Age-Structured) model by [14]. The model formulation is deterministic, and evaluations are numerical, but the author notes that it is straightforward to formulate and simulate the corresponding stochastic formulation.

Schenzle's work has stimulated a study of deterministic models with age dependence, but without seasonal variation in the contact pattern. One goal of these studies has been to search for conditions that lead to undamped oscillations. Some positive answers, but involving rather extreme conditions, are given by [3, 39].

An additional possible reason for periodic solutions in deterministic models was suggested by Feng and Thieme in [13]. They study an SIQR model, where a state Q (for quarantine) is inserted between the states I and R. They argue that infected individuals stay at home after they develop disease symptoms. Their isolation (quarantine) reduces their ability to infect others. Analysis of the resulting ODE model shows that periodic solutions are possible for a range of lengths of isolation period. Further extensions of these results are given by [20].

Keeling and Grenfell in [21] showed that the stochastic version of the Schenzle model, formulated in the natural way as a Markov population process, predicted much less persistence than is observed. They point to one consequence of the Markov population model, namely, that the distribution of waiting time in each state is **exponential**. This is quite unrealistic. They claim that considerable improvement in the prediction of persistence can be achieved by assuming a more realistic distribution of the duration of infection. Investigations along the same line were undertaken by Lloyd in [22]. The question of what influence a more realistic distribution of infectious period has on the persistence is far from settled: Some of the conclusions reached by Lloyd are opposite to those of Keeling and Grenfell. Further studies in the same direction have been undertaken by Andersson and Britton, [2].

Spatial heterogeneity is believed to play an important role in the persistence of recurrent epidemics, with asynchrony between subpopulations allowing global persistence, even if the infection dies out locally (*see* **Epidemic Models, Spatial**). A study of such questions, accounting for both stochastic and deterministic aspects, is contained in [23].

References

[1] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, New York, Tokyo.

[2] Andersson, H. & Britton, T. (2000). Stochastic epidemics in dynamic populations: Quasi-stationarity and extinction, *Journal of Mathematical Biology* **41**, 559–580.

[3] Andreasen, V. (1995). Instability in an SIR-model with age-dependent susceptibility, in *Mathematical Population Dynamics, Vol. One: Theory of Epidemics*, O. Arino, D. Axelrod, M. Kimmel & M. Langlais, eds. Wuerz Publication, Winnipeg, pp. 3–14.

[4] Aron, J.L. & Schwartz, I.B. (1984). Seasonality and period-doubling bifurcations in an epidemic model, *Journal Theoretical Biology* **110**, 665–679.

[5] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

[6] Bartlett, M.S. (1956). Deterministic and stochastic models for recurrent epidemics, in *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. University of California Press, Berkeley, pp. 81–109.

[7] Bartlett, M.S. (1957). Measles periodicity and community size, *Journal of the Royal Statistical Society. Series A* **120**, 48–70.

[8] Bartlett, M.S. (1960). *Stochastic Population Models in Ecology and Epidemiology*. Methuen, London.

[9] Billings, L. & Schwartz, I.B. (2002). Exciting chaos with noise: unexpected dynamics in epidemic outbreaks, *Journal of Mathematical Biology* **44**, 31–48.

[10] Darroch, J.N. & Seneta, E. (1967). On quasi-stationary distributions in absorbing continuous-time finite Markov chains, *Journal of Applied Probability* **4**, 192–196.

[11] Dietz, K. (1975). Transmission and control of arbovirus diseases, in *Epidemiology*, D. Ludwig & K.L. Cooke, eds. Society for Industrial and Applied Mathematics, Philadelphia, 104–121.

[12] Engbert, R. & Drepper, F.R. (1994). Chance and chaos in population biology – models of recurrent epidemics and food chain dynamics, *Chaos Solitons Fractals* **4**, 1147–1169.

[13] Feng, Z. & Thieme, H.R. (1995). Recurrent outbreaks of childhood diseases revisited: the impact of isolation, *Mathematical Biosciences* **128**, 93–130.

[14] Grenfell, B.T., Bolker, B. & Kleczkowski, A. (1995). Seasonality, demography and the dynamics of measles in developed countries, in *Epidemic Models: Their Structure and Relation to Data*, D. Mollison, ed. Publications of the Newton Institute, Cambridge University press, Cambridge, 248–268.

[15] Hamer, W.H. (1906). Epidemic disease in England – the evidence of variability and of persistence of type, *Lancet* **1**, 733–739.

[16] Hethcote, H.W. (1974). Asymptotic behavior and stability in epidemic models, in *Lecture Notes in Biomathematics*, Vol. 2. Springer-Verlag, Berlin.

[17] Hethcote, H.W. (1976). Qualitative analysis of communicable disease models, *Mathematical Biosciences* **28**, 335–356.

[18] Hethcote, H.W. (2000). The mathematics of infectious diseases, *SIAM Review* **42**(4), 599–653.

[19] Hethcote, H.W. & Levin, S.A. (1989). Periodicity in epidemiological models, in *Biomathematics: Applied Mathematical Ecology*, Vol. 18, S.A. Levin, T.G. Hallam, & L.J. Gross, eds. Springer-Verlag, Berlin, 193–211.

[20] Hethcote, H., Zhien, M. & Shengbing, L. (2002). Effects of quarantine in six endemic models for infectious diseases, *Mathematical Biosciences* **180**, 141–160.

[21] Keeling, M.J. & Grenfell, B.T. (1997). Disease extinction and community size: modelling of the persistence of measles, *Science* **275**, 65–67.

[22] Lloyd, A.L. (2001). Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics, *Theoretical Population Biology* **60**, 59–71.

[23] Lloyd, A.L. & May, R.M. (1996). Spatial heterogeneity in epidemic models, *Journal of Theoretical Biology* **179**, 1–11.

[24] Lotka, A.J. (1923). Martini's equations for the epidemiology of immunising diseases, *Nature* **111**, 633–634.

[25] Lotka, A.J. (1956). *Elements of Mathematical Biology*. Dover Publications, New York.

[26] Martini, E. (1921). *Berechnungen und Beobachtungen sur Epidemiologie and Bekämpfung der Malaria*. Gente, Hamburg.

[27] May, R.M. (1995). Necessity and chance: Deterministic chaos in ecology and evolution, *Bulletin of American Mathematical Society* **32**, 291–308.

[28] Mollison, D. (1991). The dependence of epidemic and population velocities on basic parameters, *Mathematical Biosciences* **107**, 255–287.

[29] Nåsell, I. (1995). The threshold concept in stochastic epidemic and endemic models, in *Epidemic Models: Their Structure and Relation to Data*, D. Mollison, ed. Publications of the Newton Institute, Cambridge University Press, Cambridge, 71–83.

[30] Nåsell, I. (1999). On the time to extinction in recurrent epidemics, *Journal of Royal Statistical Society Series B* **61**, 309–330.

[31] Nåsell, I. (2001). Extinction and quasi-stationarity in the Verhulst logistic model, *Journal of Theoretical Biology* **211**, 11–27.

[32] Nåsell, I. (2002). Endemicity, persistence, and quasi-stationarity, in *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction*, *The IMA Volumes in Mathematics and its Applications*, Vol. 125, C. Carlos-Castillo, S. Blower, P. van den Driessche, D. Kirschner & A.-A. Yakubu, eds. Springer-Verlag, New York, 199–227.

[33] Nåsell, I. (2002). Measles outbreaks are not chaotic, in *Mathematical approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory*, *The IMA Volumes in Mathematics and its Applications*, Vol. 126, C. Carlos-Castillo, S. Blower, P. van den Driessche, D. Kirschner & A.-A. Yakubu, eds. Springer-Verlag, New York, 85–114.

[34] Olsen, L.F., Truty, G.L. & Schaffer, W.M. (1988). Oscillations and chaos in epidemics: A nonlinear study

of six childhood diseases in Copenhagen, Denmark, *Theoretical Population Biology* **33**, 344–370.

[35] Schaffer, W.M. (1985). Can nonlinear dynamics elucidate mechanisms in ecology and epidemiology? *IMA Journal of Mathematics Applied in Medicine and Biology* **2**, 221–252.

[36] Schaffer, W.M. & Kot, M. (1985). Nearly one dimensional dynamics in an epidemic, *Journal of Theoretical Biology* **112**, 403–427.

[37] Schenzle, D. (1984). An age-structured model of pre- and post-vaccination measles transmission, *IMA Journal of Mathematics Applied in Medicine and Biology* **1**, 169–191.

[38] Soper, H.E. (1929). The interpretation of periodicity in disease prevalence (with discussion), *Journal of the Royal Statistical Society. Series A* **92**, 34–73.

[39] Thieme, H. (1991). Stability change of the endemic equilibrium in age-structured models for the spread of SIR type infectious diseases, in *Differential Equations Models in Biology, Epidemiology and Ecology*, *Lecture Notes in Biomathematics*, Vol. 92, S. Busenberg & M. Martelli, eds. Springer Verlag, New York, pp. 139–158.

Further Reading

Keeling, M.J. & Grenfell, B.T. (1998). Effect of variability in infection period on the persistence and spatial spread of infectious diseases, *Mathematical Biosciences* **147**, 207–226.

INGEMAR NÅSELL

Epidemic Models, Sensitivity Analysis

Recent years have seen the use of increasingly complex models for epidemic dynamics. For example, the emergence of the AIDS epidemic has resulted in the development of models including a variety of features [2, 3, 5–9, 13–15, 19, 27, 28] (*see* **AIDS and HIV**).

All these developments are necessary to increase our understanding of the mechanics of transmission. For example, again referring to the AIDS epidemic, certain features have had to be included to explain such things as:

1. the variability in infectiousness observed in **partner studies**;
2. the long duration of the **incubation period**;
3. the pattern of early growth of the epidemic; and
4. the differences between high- and low-risk subgroups of the population.

These complex models have developed out of a range of much simpler models. When we consider simple models with relatively few parameters it is a straightforward task to analyze and compare models: for example: To which changes in parameters is the model sensitive? What ranges of epidemic dynamics are possible? Are two models similar?

Consider, for example, the two models, A and B, described in Figure 1. There is only one difference between these models: where model A has two parameters c (the contact rate) and p, model B has a single parameter β. It is clear that if β is equal to cp and if all other parameters are equal, then the two models will give precisely the same dynamics. So, in one sense, model A contains a parameter that is redundant. The reason for why we might use the formulation described in model A is that it gives us, in a simplistic way, a more detailed understanding of the mechanics of transmission.

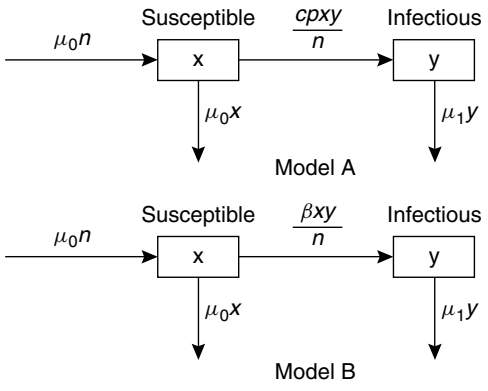

Figure 1 Two simple models for epidemic spread

This sort of reasoning has been the driving force behind the development of the complex models for AIDS described above. However, an important aspect of the development of more complex models is the need to identify whether or not the addition of an additional parameter makes a substantial difference to the dynamics of an epidemic model. The example which compares models A and B described above is a special case: the extra parameter in model A has absolutely no effect on the dynamics. In other cases the situation may not be so clear-cut. In most cases the addition of an extra parameter will add to the variety of outcomes which may be observed, but this may range from a substantial to a very insignificant change.

The reduction in the number of parameters between model A and model B is similar to the approach described by Nåsell [29]. Nåsell, who considers equilibrium incidence of malaria, uses dimensional analysis to strip out a number of unnecessary parameters.

The question of how much influence individual parameters have on certain quantities of interest was first considered in a systematic fashion by Bailey & Duppenthaler [4]. They presented a detailed approach to sensitivity analysis by considering the full set of basic parameters. By incorporating the uncertainty associated with each parameter they assessed which parameters have the most influence on the level of incidence in the equilibrium state. The use of random input parameters to investigate sensitivity is a method which has rarely been used since. The most significant development along these lines is the method described by Blower & Dowlatabadi [6]. Their method is described in more detail in a later section and conforms to the conventional notion of a **sensitivity analysis**.

Sometimes it is possible for us to reparameterize a model to give us more obvious insight into its

workings. Let us look again at the dynamics of model B:

$$\frac{dx}{dt} = \mu_0 n - \mu_0 x - \frac{\beta x y}{n},$$
$$\frac{dy}{dt} = \frac{\beta x y}{n} - \mu_1 y.$$

Suppose the value of β is known to be equal to 0.1. What can be said about the epidemic? The answer is, of course, "very little": the epidemic could die out or it could take off and be quite substantial. We can only say something about the dynamics when we know the values of some of the other parameters.

Suppose, instead, that we define

$\tau_0 = \frac{1}{\mu_0}$, the mean lifetime in the absence of infection,
$\tau_1 = \frac{1}{\mu_1}$, the mean duration of infectiousness,
$R_0 = \beta\tau_1$, the basic reproductive ratio: the mean number of secondary cases of infection caused by one primary infective in the early stages of the epidemic (*see* **Reproduction Number**), and
$\theta = \beta - \mu_1 = \frac{R_0 - 1}{\tau_1}$, the initial growth rate.

Then

$$\frac{dx}{dt} = \frac{n}{\tau_0} - \frac{x}{\tau_0} - \frac{R_0}{\tau_1}\frac{x}{n}y,$$
$$\frac{dy}{dt} = \frac{R_0}{\tau_1}\frac{x}{n}y - \frac{y}{\tau_1}.$$

Using this formulation, if we know the value of R_0, then we immediately know something about how the epidemic will progress: there will be an epidemic if and only if $R_0 > 1$ and the disease will become endemic again only if $R_0 > 1$. If we also know the value of τ_1, then we begin to get a picture of how quickly the epidemic will progress: the higher the value of τ_1 the slower the progress. Furthermore we can derive the value of θ from R_0 and τ_1 and this tells us about the early dynamics (*see* **Epidemic Thresholds**).

A conventional sensitivity analysis is carried out without such a reparameterization. As a consequence, a typical conclusion is that all parameters have a relatively significant impact on the dynamics of the epidemic. Blower & Dowlatabadi [6] developed this by using methods which allow us to make statements along the lines of "parameter θ_i explains 25% of the variation in the total number of cases".

We later describe the method of primary components [9, 10]. This is a method that formalizes the process of reparameterization. In particular, the aim is to find a parameterization in which there is a small number of parameters (or *primary components*) which between them explain a very large part of the dynamics of the epidemic. Put another way: suppose the primary components are fixed. The remaining *secondary parameters* may be uncertain or random, but, within the possible range of values that they can take, the dynamics of the epidemic model are largely unchanged. The approach is similar to a **principal components analysis**. The important distinction is that the primary components have a simple interpretation (for example, the basic reproductive ratio R_0) and, therefore, lose the optimality of principal components.

There are three differences from the Blower–Dowlatabadi approach. First, the method of primary components considers the whole epidemic curve or some subsection of it rather than a single outcome. This is the aim of many sensitivity analyses, and the later section describes how this can be done in a systematic way and with reference to past data. Secondly, by reparameterizing the model, a small number of inputs (the primary components) explain a much larger part of the variance in the output. Thirdly, the method is easily applied to model fitting and **projection** of an epidemic curve.

Reasons for Carrying Out a Sensitivity Analysis

Before thinking about why a sensitivity analysis is necessary and desirable it is important to consider why we are modeling in the first place. There are various reasons:

1. to explain the observed pattern of epidemic spread;
2. to explain the mechanics of transmission;
3. to predict
   - the total numbers of cases
   - the evolution of the epidemic curve

- the severity of spread into different subgroups
- the chance that the epidemic dies out;

4. to make provision for future healthcare;
5. to investigate and devise strategies for epidemic reduction or eradication through

- vaccination (*see* **Vaccine Studies**)
- treatment of infected individuals
- education;

6. to determine modes of behavior

- thresholds
- endemicity
- stability.

Let us consider a general model. Let

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ the vector of input parameters,

$\mathbf{y}(t) = y(t, \boldsymbol{\theta})$, the vector of output variables of interest.

For example, $\mathbf{y}(t)$ might include numbers of susceptible, infectious, removed, immune or dead (**prevalence** curves) or the rates of new infection, recovery or death (**incidence** curves). $\mathbf{y}(t)$ might be deterministic or stochastic, certain components may be incompletely observed, or not observable at all, and most components will be subject to some degree of **measurement error**.

The purpose of a sensitivity analysis is to assess how sensitive $\mathbf{y}(t)$ (or some of its components) is to changes in the values of the various components of the parameter vector $\boldsymbol{\theta}$. The reasons behind such an analysis are partially motivated by the reasons for modeling in the first place.

1. to consider the effects of uncertain parameter values;
2. to consider the sensitivity of the outcome of modeling to the choice of model;
3. to assess uncertainty in the future;
4. to assess the likely effectiveness of specified control strategies;
5. to identify which parameters we should obtain improved estimates of in order to reduce future uncertainty by the maximum amount;
6. to gauge how much information about the underlying model can be obtained from an observed incidence curve;
7. to assess the importance of certain components or structures in a model;
8. to assess which parameter combinations are consistent with existing data;
9. to be aware of the possible existence of bifurcations within the likely range of parameter values.

This list does not intend to be comprehensive, but it does indicate the wide variety of questions that need to be considered.

The driving force behind the need for sensitivity analyses are the first two items: parameter and model uncertainty. If both of these were known, then the task of an epidemic modeler would be considerably easier, but also rather boring! However, even if parameter values are known for the past, changes can happen in the future as a result of control strategies, changes in the underlying population, or mutations of the disease. The effects of these sorts of changes are rather harder to predict, but sensitivity analysis is still desirable since we can get some sort of a feel for the magnitude of the effect.

It is important that a sensitivity analysis should be as comprehensive as possible in terms of how parameter values are varied. Bailey & Duppenthaler [4] warn against the risks of focusing on a subset of the parameter set: "Examination of small numbers of supposedly important parameters may be substantially affected by unconscious bias".

Sensitivity Analysis for Single Quantities

The approach described here follows that of Blower & Dowlatabadi [6]. For the basic form of analysis we make the following assumptions:

1. we are interested in a single quantity (for example, cumulative cases of AIDS over the next 30 years or the equilibrium incidence of new infection);
2. the input parameters are independent random variables;
3. we do not use past incidence or prevalence data.

Assumptions 2 and 3 imply that we can use only data from secondary studies which, for example, follow the progress of specific individuals. Each of these assumptions can be relaxed: a point that will be discussed later in this section.

Let $y = y(\theta)$ be the quantity of interest and $f(\boldsymbol{\theta})$ be the probability density function for $\boldsymbol{\theta}$. Under

assumption 2:

$$f(\boldsymbol{\theta}) = f_1(\theta_1) f_2(\theta_2) \cdots f_k(\theta_k).$$

We are interested in the unconditional distribution of $y(\boldsymbol{\theta})$ and also in the degree of **correlation** between $y(\boldsymbol{\theta})$ and each of the inputs $\theta_1, \ldots, \theta_k$.

Let $\boldsymbol{\theta}^*$ be the mean value of $\boldsymbol{\theta}$ given the density function $f(\boldsymbol{\theta})$. Then

$$y(\boldsymbol{\theta}) = y(\boldsymbol{\theta}^*) + \mathbf{h}^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(|\boldsymbol{\theta} - \boldsymbol{\theta}^*|)$$

where $\mathbf{h} = \mathbf{h}(\boldsymbol{\theta}^*)$ is the response vector.

If the range of values for $\boldsymbol{\theta}$ around $\boldsymbol{\theta}^*$ is relatively small, and if there are no bifurcation points nearby, then $\hat{y}(\boldsymbol{\theta}) = y(\boldsymbol{\theta}^*) + \mathbf{h}^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ gives a good approximation to $y(\boldsymbol{\theta})$. We then have (in a similar fashion to Bailey & Duppenthaler [4]):

$$\begin{aligned}
\mathrm{E}[\hat{y}(\boldsymbol{\theta})] &= y(\boldsymbol{\theta}^*), \\
\mathrm{var}[\hat{y}(\boldsymbol{\theta})] &= \sum_{i=1}^{k} h_i^2 \mathrm{var}(\theta_i), \\
\mathrm{cov}[\hat{y}(\boldsymbol{\theta}), \theta_i] &= h_i \mathrm{var}(\theta_i) \\
\Rightarrow \rho_i &= \mathrm{cov}[\hat{y}(\boldsymbol{\theta}), \theta_i] \\
&= \frac{h_i \mathrm{var}(\theta_i)^{1/2}}{\left[\sum_{j=1}^{k} h_j^2 \mathrm{var}(\theta_j)\right]^{1/2}}.
\end{aligned}$$

Under this assumption of linearity, the correlation coefficient, ρ_i, is a measure of the sensitivity of $y(\boldsymbol{\theta})$ to changes in θ_i. The closer ρ_i is to 1 or -1 the more sensitive $y(\boldsymbol{\theta})$ is to changes in θ_i. Note that this measure explicitly accounts for the level of uncertainty in the input parameters. Now the ρ_i indicate the levels of sensitivity relative to other input parameters. Absolute levels of sensitivity can be gauged by combining this information with the variance of $y(\boldsymbol{\theta})$.

The response vector, $\mathbf{h}(\boldsymbol{\theta}^*)$ can be estimated by carrying out only $k+1$ **simulations** of the model: one with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and one for each $i = 1, \ldots, k$ with $\theta_i = \theta_i^* + \varepsilon$ and $\theta_j = \theta_j^*$ for $j \neq i$ (where ε is small).

Latin Hypercube Sampling and Partial Rank Correlation Coefficients

Now $y(\boldsymbol{\theta})$ may in fact be sufficiently nonlinear within the range of values of $\boldsymbol{\theta}$ (possibly with a bifurcation point) so that $\hat{y}(\boldsymbol{\theta})$ gives a poor approximation to $y(\boldsymbol{\theta})$ except near to $\boldsymbol{\theta}^*$. The method described above still gives a good guide to the sensitivity of $y(\boldsymbol{\theta})$ to changes in θ_i. However, if we wish to be more precise, then the method of *partial rank correlation coefficients* (as described by Blower & Dowlatabadi [6]) is appropriate.

A complete, deterministic analysis would need to cover all possible values of $\boldsymbol{\theta}$. This is only feasible if the model has very few parameters and if $y(\boldsymbol{\theta})$ can be computed relatively quickly. Often there are 10 or more parameters and $y(\boldsymbol{\theta})$ can be computationally very expensive, so that another approach is necessary.

A simple approach is to use simple **random sampling**. This takes N independent and identically distributed realizations of $\boldsymbol{\theta}$. By analyzing correlations between each parameter $\theta_1, \ldots, \theta_k$ and $y(\boldsymbol{\theta})$ we can assess the level of sensitivity of $y(\boldsymbol{\theta})$ to each input parameter. The approach described by Blower & Dowlatabadi [6] is similar but they make use of Latin hypercube sampling to choose the N parameter sets.

The Latin hypercube sampling technique was first described by McKay et al. [24]. The method can significantly reduce the variance of various estimates relating to the distribution of $y(\boldsymbol{\theta})$ (for example, its mean and variance) and its relationship with each of the input parameters. Put another way, Latin hypercube sampling requires a smaller number of simulations of $y(\boldsymbol{\theta})$ to match the variance of various estimates under simple random sampling.

Some of the mathematics behind the technique were tightened up by Stein [34] who, in particular, proved that the variance of estimates under Latin hypercube sampling is lower for large N in all cases. In the worst cases the advantage in using Latin hypercube sampling might be small, but in practice the difference between the two methods can be very significant.

Owen [30] and Park [31] have both devised **algorithms** for choosing Latin hypercube designs which optimize certain criteria. The technique proceeds as follows. We wish to generate N values for $\boldsymbol{\theta}$. For each $i = 1, \ldots, k$ let θ_{i0} be the left-hand end of the range of θ_i and θ_{iN} be the right-hand end. We also define $F_i(x) = \int_{-\infty}^{x} f_i(u)\,\mathrm{d}u$ to be the marginal cumulative distribution function for θ_i. Thus, $F_i(\theta_{i0}) = 0$ and $F_i(\theta_{iN}) = 1$. We also define $\theta_{i0} \leq \theta_{i1} \leq \cdots \leq \theta_{iN}$ such that $\Pr(\theta_{ij} < \theta_i \leq \theta_{ij+1}) = F_i(\theta_{ij+1}) - F_i(\theta_{ij}) = 1/N$. Let $P_1 = (P_{11}, \ldots, P_{1N})$, $P_2, \ldots, P_k$ be k independent random

permutations of $(1, 2, \ldots, N)$. For $i = 1, \ldots, k$ and $j = 1, \ldots, N$ let ξ_{ij} be independent and identically distributed random variables with a **uniform distribution** on [0, 1] and let

$$Z_{ij} = F_i^{-1}\left[\frac{(P_{ij} - 1 + \xi_{ij})}{N}\right].$$

For each i and j the unconditional distribution for Z_{ij} is the same as the marginal distribution for θ_i. We also have, for each i, exactly one Z_{ij}, $j = 1, \ldots, N$, in each of the intervals $(\theta_{i0}, \theta_{i1}], \ldots, (\theta_{iN-1}, \theta_{iN}]$. This gives a rather more uniform look to a Latin hypercube sampling scheme than a typical random sample.

For each j let $\mathbf{Z}_j = (Z_{1j}, Z_{2j}, \ldots, Z_{kj})$ and $y_j = y(\mathbf{Z}_j)$. We can immediately now look at the estimated or empirical distribution for $y(\boldsymbol{\theta})$. The empirical distribution derived from $y_1, \ldots, y_N$ will tend to that of $y(\boldsymbol{\theta})$ as N gets large. Furthermore, for a given N, the arguments of McKay [24] and Stein [34] indicate that this empirical distribution is likely to be more accurate than that using simple random sampling.

If the output variable is significantly nonlinear within the normal range of values for the input parameters, then a conventional analysis of correlation will tend to understate the importance of those input parameters to which $y(\boldsymbol{\theta})$ is most nonlinear. To avoid this problem, Iman et al. [18] suggest making use of the ranks of the input parameters and of the outputs rather than their absolute values.

This approach was used by Blower & Dowlatabadi [6] in their analysis of a model for the spread of AIDS. They suggest that partial rank correlation coefficients should be used to investigate the sensitivity of $y(\boldsymbol{\theta})$ to changes in each of the input parameters.

The method proceeds as follows. Recall that each of the P_i is a random permutation of $1, 2, \ldots, N$. Thus P_{ij} is the **rank** of Z_{ij} in the set $\{Z_{i1}, \ldots, Z_{iN}\}$. Let $\mathbf{R} = (r_{ij})$ be a $(k + 1) \times N$ matrix with $r_{ij} = P_{ij}$ for $1 \le i \le k$ and $1 \le j \le N$, and $r_{k+1,j}$ be equal to the rank of $y_j = y(\mathbf{Z}_j)$ in the set $\{y_1, \ldots, y_N\}$. Let $\mathbf{C} = (\rho_{ij})$ be the symmetric $(k + 1) \times (k + 1)$ matrix defined by

$$\rho_{ij} = \frac{\sum_{t=1}^{N}(r_{it} - \mu)(r_{jt} - \mu)}{\left[\sum_{t=1}^{N}(r_{it} - \mu)^2 \sum_{s=1}^{N}(r_{jt} - \mu)^2\right]^{1/2}},$$

where $\mu = (N + 1)/2$ is the mean rank. The matrix $\mathbf{C}$ is the matrix of **rank correlation** coefficients. Another matrix, $\mathbf{B}$, is defined as the inverse of $\mathbf{C}$, i.e. $\mathbf{B} = \mathbf{C}^{-1}$. The partial rank correlation coefficient between parameters i and j [or between parameter i and $y(\theta)$ if $j = k + 1$] is then defined as (see [12] and [20])

$$\gamma_{ij} = \frac{-b_{ij}}{(b_{ii}b_{jj})^{1/2}}.$$

$\gamma_{i,k+1}$ is the correlation between the rank of θ_i and the rank of $y(\boldsymbol{\theta})$ given that all other input parameters are fixed. The larger the value of $\gamma_{i,k+1}$, the more sensitive $y(\boldsymbol{\theta})$ is to changes in θ_i.

There exist tests of significance of the hypothesis that θ_i and $y(\boldsymbol{\theta})$ are independent (for example, see [12]). This independence, of course, is very rarely the case. With a limited sample size, however, some of the $\gamma_{i,k+1}$ might not be significantly different from 0. On the other hand, if we take a large enough sample size, then we will eventually find that all of the $\gamma_{i,k+1}$ are significantly different from 0.

Both the rank correlation coefficients, $\rho_{i,k+1}$, and the partial rank correlation coefficients, $\gamma_{i,k+1}$, give us information about the sensitivity of $y(\boldsymbol{\theta})$ to changes in parameter i. If the input parameters are independent, then the ρ_{ij} will tell us as much as the γ_{ij}. However, if $y(\boldsymbol{\theta})$ gets closer to being a deterministic function of $\boldsymbol{\theta}$, then the $\gamma_{i,k+1}$ will all get closer to 1 while the ρ_{ij} converge to values that still reflect, more obviously, the relative importance of each input parameter. Blower & Dowlatabadi [6] use the transformation

$$t_i = \gamma_{i,k+1}\left(\frac{N - 2}{1 - \gamma_{i,k+1}}\right)^{1/2}.$$

The t_i, relative to one another, give a much better measure of the relative importance of each parameter. In this respect the values of the t_i relative to one another give a much better measure of the relative importance of each parameter.

If the input parameters are not independent, then this will tend to distort the ρ_{ij} and the γ_{ij}. For example, take model A in Figure 1. Suppose that the two infection parameters, c and p, are correlated in such a way that $cp = \beta + \varepsilon$, where β is constant and ε has zero mean and a relatively low variance. Then we will find that the partial rank correlation coefficients are very high while the rank correlation coefficients are very low.

The use of both sensitivity measures is discussed by Iman & Helton [17]. They note that while they will

give qualitatively similar results, they may indicate significantly different levels of sensitivity to certain input parameters.

Relaxation of Assumptions

Often we are interested in a projection of the epidemic curve rather than a single quantity at a single point in time: that is, $y(t, \boldsymbol{\theta})$ for $t > T_0$, the time of the last observation. Clearly it is not sensible to repeat the exercise, say m times (at times $T_0 < t_1 < \cdots < t_m$), in succession. Instead, the exercise should be run m times in parallel, to take advantage of the fact that $y(t_1, \boldsymbol{\theta}), \ldots, y(t_m, \boldsymbol{\theta})$ are all directly connected. Thus, for each $\mathbf{Z}_j$, $j = 1, \ldots, N$, we generate a full epidemic up to time t_m and extract values for $y(t_1, \mathbf{Z}_j), \ldots, y(t_m, \mathbf{Z}_j)$. This allows the construction of some sort of confidence band for the epidemic. However, in the sensitivity analysis, we can still only talk about the sensitivity of $y(t_j, \boldsymbol{\theta})$ to changes in parameter i. We are not able easily to talk about the sensitivity of the whole curve to changes in parameter i.

The input parameters do not need to be independent, although random but dependent input parameters can be more difficult to generate. Blower & Dowlatabadi [6] give a simple example of how this can be done: they specify that θ_2 has a triangular distribution with minimum value θ_1 and maximum 1. Stein [34] describes a more general method for generating dependent input parameters within a Latin hypercube sampling framework.

Besides using secondary data to help specify the distributions for the input parameters it is desirable to make use also of the observed epidemic curve itself. This can be a lengthy process since it involves first, numerical evaluation of, for example, the **maximum likelihood** or **Bayesian** estimators; and secondly, derivation of a suitable approximate distribution around this central estimate. Taking account of the observed epidemic curve is not easy, therefore, within the framework described in this section.

Since it is essential that we should take account of the observed epidemic curve, the method of primary components was devised.

Primary Component Analysis

This section describes an alternative method of sensitivity analysis which considers the whole epidemic curve, and which is appropriate for model fitting and projection. The method seeks to reparameterize the model leaving us with a small number of *primary components* which dictate epidemic dynamics. A common example of a primary component is the basic reproductive ratio, R_0, as in the development of model B at the beginning of the article.

The concept of such a set of primary components has been described by Mollison [25] and Cairns [9, 10]. Typically, a complex model will have many parameters but perhaps only three or four primary components. Knowledge of the values of the primary components will provide enough information to describe the epidemic curve to within a very high degree of accuracy.

The central idea behind this section is the notion that, in many cases, it is sufficient to estimate the primary components of a model to be able to obtain an adequate fit of past data and an adequate projection of the future course of an epidemic.

An important problem is that of how to fit a model to primary data (the epidemic curve) and secondary data (data gathered indirectly through medical and other studies related to the epidemic). Full maximum likelihood or a full Bayesian analysis would take all this data together resulting in a rather complex and difficult-to-evaluate function. First, maximization would take a long time. Secondly, the **likelihood** surface often will be very flat in some dimensions, meaning that the estimates of some parameters are subject to large standard errors.

The primary component method offers two principal advantages over full maximum likelihood methods:

1. estimation of the small number of primary components is much easier and faster than estimation of the full set of basic parameters; and
2. primary components can still be estimated reliably when secondary parameters are only backed up by unreliable or statistically unsound secondary data.

The fundamental requirements for a set of primary components are as follows:

1. A primary component is a function of the basic epidemic parameters which dictates epidemic dynamics.
2. Each primary component should be simple to interpret. The reason for this is that it makes it

much easier to transmit our conclusions from an analysis to nonexperts or laymen. Often these are the people who will make decisions about, for example, future provision of healthcare facilities. If results can be presented in a way that is simple to interpret, then it is more likely that an epidemic modeler will be in a position to influence what decisions are made.

3. The set of primary components should be as small as possible. This ties in with an overall objective of minimizing the time spent in fitting a model to a set of data: "Why estimate 10 parameters when 3 will do?" However, this can conflict with the previous requirement. It may be that a set of primary components could be reduced further but at the cost of leaving a set in which one or more of the primary components no longer has a simple interpretation. In such circumstances it may be undesirable to proceed with this final shrinkage of the set in order to retain the ease of interpretation.

It is also desirable (but not essential) that each primary component affects either short- or long-term dynamics, but not both. This is a useful criterion from the point of view that it eases comparison of epidemic curves generated by different parameter sets and it can speed up the process of estimation of parameter values.

The requirement that primary components should be easy to interpret provides the main distinction over the more rigorously founded theory of principal component analysis. This would produce a set of **orthogonal** components in order of their magnitude of influence on the likelihood function (and hence on the dynamics of the epidemic). While this would provide a theoretically optimal set of primary components, the result would be both at the expense of interpretability and also of restricting the effect of individual components to either short- or long-term dynamics. It is no coincidence, however, that a set of primary components will be closely aligned with the principal components.

Complementing the set of primary components we have the set of *secondary parameters*. The characteristics of this set are that provided the primary components remain fixed, then we can vary the remaining set of secondary parameters without significantly altering the dynamics of the epidemic. The range of values tested is designed to be consistent with the various sources of secondary data. In terms of model fitting this means that it is not of significance whether we optimize over the full set of parameters or just over the set of primary components in combination with a *realistic* rather than optimal set of secondary parameter values.

If, however, we do find that we can vary the secondary parameters within a realistic range and significantly alter the dynamics, then this indicates that the set of primary components is in some way incomplete or inadequate.

Cairns [9, 10] describes how primary components can be identified by subjective means:

1. Identify a potential first primary component (or the first two, say). Typically this might be the basic reproductive ratio, R_0, or the initial exponential growth rate of the epidemic. Let this set be denoted by $\boldsymbol{\theta}_P$.
2. Subject to $\boldsymbol{\theta}_P$ being fixed, vary the remaining parameters, $\boldsymbol{\theta}_S$, within a realistic range around a central value, $\boldsymbol{\theta}_0$ (without specific reference to probability distributions).
3. If there is no significant variation, then the set of primary components, $\boldsymbol{\theta}_P$, is complete.
4. Otherwise we need either to choose an alternative set of primary components of the same size or to increase the size of the set by one. Let $\boldsymbol{\theta}_P$ be the new, altered set of primary components and then return to step 2.

An example of this process applied to a simple model for the spread of HIV and AIDS can be found in [10]. Cairns [10] also discusses how the number of components depends on whether we consider just the incidence of new cases of AIDS or include the numbers of new infections in addition.

An objective means of carrying out a primary component analysis of an epidemic model was first proposed by Cairns (in discussion of Mollison et al. [26]). Hearne [16] has described the use of a similar objective function in the sensitivity analysis of a systems dynamics problem. This will be developed in the next section.

A Measure of Sensitivity

Previous work (for example, [9] and [10]) has relied on a degree of subjectivity when a decision must be made as to whether or not two epidemic curves are significantly different.

Consider Figure 2. It is clear that there is a much better match between curves (a) and (b) than curves (a) and (c). In this example the distinction is clear, but often this is not the case and this section aims to aid this process.

Let us consider the problem of modeling the AIDS epidemic. Suppose we assume that observed cases of AIDS occur as a **Poisson process** with an intensity following the deterministic curve (this is only valid when the population is large) then the log likelihood will be

$$l(\mathbf{A}, \boldsymbol{\theta}) = \sum_{t=1}^{T} [A_t \log(a_t) - a_t] + \text{constant},$$

where A_t is the observed incidence of AIDS in time period t and $a_t = a_t(\boldsymbol{\theta})$ is the predicted incidence based on the parameter values $\boldsymbol{\theta}$.

If the aim of modeling is at some stage to fit the model to a set of data $\{A_t\}$, then it seems appropriate to design a sensitivity function that reflects the characteristics of the log likelihood. An appropriate function is thus

$$l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \int_0^T [a_t(\boldsymbol{\theta}_0) \log a_t(\boldsymbol{\theta}) - a_t(\boldsymbol{\theta})] \, \mathrm{d}t,$$

where $\boldsymbol{\theta}_0$ is the central or "true" parameter set, and $a_t(\boldsymbol{\theta}_0)$ and $a_t(\boldsymbol{\theta})$ are the epidemic curves generated by $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, respectively. $\boldsymbol{\theta}_0$ is chosen to reflect existing primary and secondary data, a task that becomes easier as we reparameterize the model.

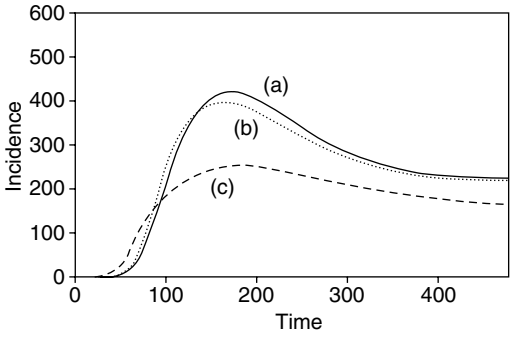

Figure 2 Three epidemic curves: (a) and (b) are better matched in some sense than (a) and (c).

It may be appropriate to consider the discrete form of the sensitivity function. This is

$$l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \sum_{t=1}^{T} [a_t(\boldsymbol{\theta}_0) \log a_t(\boldsymbol{\theta}) - a_t(\boldsymbol{\theta})].$$

In what follows we consider the continuous and discrete forms interchangeably.

A further alternative is to note that $-2 \times$ (log likelihood) is asymptotically equivalent to the **chi-square statistic** (*see* **Likelihood Ratio Tests**). The equivalent sensitivity function would therefore be

$$c(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{[a_t(\boldsymbol{\theta}_0) - a_t(\boldsymbol{\theta})]^2}{a_t(\boldsymbol{\theta})},$$
$$\approx 2[l(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - l(\boldsymbol{\theta}_0, \boldsymbol{\theta})].$$

For a given value of t note that $a_t(\boldsymbol{\theta}_0) \log a_t(\boldsymbol{\theta}) - a_t(\boldsymbol{\theta})$ is maximized when $a_t(\boldsymbol{\theta}) = a_t(\boldsymbol{\theta}_0)$ (for example, if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$). Hence

$$l_{\max} = \sup_{\boldsymbol{\theta}} l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = l(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0).$$

Furthermore, by Taylor's expansion we have

$$l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = l_{\max} - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o(|\boldsymbol{\theta} - \boldsymbol{\theta}_0|^2), \quad (1)$$

where $\mathbf{H} = -D^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}_0, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ (the matrix of second derivatives evaluated at $\boldsymbol{\theta}_0$) is a positive semidefinite $k \times k$ matrix and k is the number of basic parameters in the model.

If $l(\cdot)$ is thought of as a likelihood function, then $\mathbf{H}$ is an **information matrix** in the normal statistical sense. The **eigenvalues** of $\mathbf{H}$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and the corresponding **eigenvectors** are $\mathbf{e}_1, \ldots, \mathbf{e}_k$ with $\mathbf{e}_i^{\mathrm{T}} \mathbf{e}_j = \delta_{ij}$. A large eigenvalue, λ_i, means that $l(\cdot)$ is more sensitive to changes in the parameter values in line with $\mathbf{e}_i$ (at least in absolute terms). This also means that $\mathbf{e}_1$ gives us the most information about the shape of the epidemic curve, $\mathbf{e}_2$ the second most information, etc.

In theory, we should select primary components that match exactly the principal eigenvectors. However, the resulting components would not have a simple interpretation, and would violate the second requirement for a primary component. We therefore investigate different potential sets of primary components which are not necessarily optimal in the

sense of providing maximum information but which are, nevertheless, closely aligned with the principal eigenvectors.

The following sections develop the sensitivity function from a statistical point of view, but the first step is always to reparameterize the model in such a way that we can divide the set of parameters $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_{\mathrm{P}}$, the p potential primary components, and $\boldsymbol{\theta}_{\mathrm{S}}$, the remaining $s = k - p$ secondary parameters.

Model Fitting

Before carrying out a projection of an epidemic and a sensitivity analysis of this projection with respect to variation of the input parameters, it is necessary to fit the model to what data we have available. We then have an approximate joint distribution for the primary components, $\boldsymbol{\theta}_{\mathrm{P}}$, and the secondary parameters, $\boldsymbol{\theta}_{\mathrm{S}}$, derived from the existing data. This can be used to generate random input parameter values, each producing a different projection of the epidemic. A sensitivity function for the projected part of the epidemic curve, of the form described in the previous subsection, can then be analyzed for sensitivity to changes in each of the input primary components or secondary parameters.

To aid comprehension, the remainder of this section will continue to consider a model for the spread of HIV and AIDS.

Two types of data are assumed to be available: (primary) AIDS incidence data; and (secondary) data from other studies which give us information about the secondary parameters and perhaps also about the primary components. Within the field of AIDS modeling, an example of such secondary data is that which relates to the distribution of the infectious period (for example, [21–23]).

These data can be treated in a number of different ways. One of these is to pool the data and maximize the likelihood over the full set of basic parameters. When the model is complex this may be a very lengthy or perhaps even impossible task if the data are not suitably detailed.

Here we describe an alternative approach.

1. Use the secondary data to estimate the secondary parameters. These estimates will then be used in the second stage.
2. Use the primary data and maximize their likelihood over the set of primary components only.

A modeler needs to be satisfied, however, that the use of fixed rather than uncertain secondary parameter values does not produce a significantly narrower range of fits and projections.

The estimation procedure will inevitably result in some loss of accuracy. However, it is intended that if the set of primary components has been chosen carefully, then this loss of accuracy will be minimal. Conversely, if the loss of accuracy is significant, then the set of primary components should be considered as being inappropriate or inadequate in some way.

The major advantage is that by breaking down the process of estimation we can make the task much simpler and faster without having a significant loss of accuracy.

We consider four different estimators for $\boldsymbol{\theta}$:

1. $\hat{\boldsymbol{\theta}}$: the maximum likelihood estimator (MLE) based on primary data alone;
2. $\tilde{\boldsymbol{\theta}}$: the MLE based on secondary data alone;
3. $\dot{\boldsymbol{\theta}}$: the MLE based on pooled primary and secondary data; and
4. $\overline{\boldsymbol{\theta}}$: the primary component maximum likelihood estimator (described below).

Let $l(\mathbf{A};\boldsymbol{\theta})$ be the likelihood function given only the primary AIDS incidence data $\{A_t\}$. Then

$$l(\mathbf{A};\boldsymbol{\theta}) \approx l_{\max} - \tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{H}_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

where $l_{\max} = l(\mathbf{A};\hat{\boldsymbol{\theta}})$ is the maximum likelihood, and $\mathbf{H}_p$ is the information matrix for the primary data as defined in (1).

Suppose, also, that secondary data have been collected and that we wish to fix the secondary parameters at $\boldsymbol{\theta}_{\mathrm{S}} = \tilde{\boldsymbol{\theta}}_{\mathrm{S}}$. Estimation of the primary components, $\boldsymbol{\theta}_{\mathrm{P}}$, is then reduced to the problem

$$\text{minimize } f(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{H}_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

$$\text{subject to } \boldsymbol{\theta}_{\mathrm{S}} = \tilde{\boldsymbol{\theta}}_{\mathrm{S}}.$$

Given that $\boldsymbol{\theta}_{\mathrm{P}}$ is $p \times 1$ and $\boldsymbol{\theta}_{\mathrm{S}}$ is $s \times 1$, we write

$$\mathbf{H}_p = \begin{pmatrix} \mathbf{H}_{p11} & \mathbf{H}_{p12} \\ \mathbf{H}_{p21} & \mathbf{H}_{p22} \end{pmatrix},$$

where $\mathbf{H}_{p11}$ is a $p \times p$ matrix, $\mathbf{H}_{p12} = \mathbf{H}_{p21}^{\mathrm{T}}$ is $p \times s$, and $\mathbf{H}_{p22}$ is $s \times s$.

The minimization with the constraint is then equivalent to the unconstrained problem:

$$\text{minimize } f(\boldsymbol{\theta}_{\mathrm{P}}) = (\boldsymbol{\theta}_{\mathrm{P}} - \overline{\boldsymbol{\theta}}_{\mathrm{P}})^{\mathrm{T}}\mathbf{H}_{p11}(\boldsymbol{\theta}_{\mathrm{P}} - \overline{\boldsymbol{\theta}}_{\mathrm{P}}) + C,$$

where

$$\overline{\boldsymbol{\theta}}_{\mathrm{P}} = \hat{\boldsymbol{\theta}}_{\mathrm{P}} - \mathbf{H}_{p11}^{-1}\mathbf{H}_{p12}(\tilde{\boldsymbol{\theta}}_{\mathrm{S}} - \hat{\boldsymbol{\theta}}_{\mathrm{S}})$$

and

$$C = (\tilde{\boldsymbol{\theta}}_{\mathrm{S}} - \hat{\boldsymbol{\theta}}_{\mathrm{S}})^{\mathrm{T}}[\mathbf{H}_{p22} - \mathbf{H}_{p21}\mathbf{H}_{p11}^{-1}\mathbf{H}_{p12}](\tilde{\boldsymbol{\theta}}_{\mathrm{S}} - \hat{\boldsymbol{\theta}}_{\mathrm{S}}).$$

It is clear that this is minimized when $\boldsymbol{\theta}_{\mathrm{P}} = \overline{\boldsymbol{\theta}}_{\mathrm{P}}$.

Note that $C = f(\overline{\boldsymbol{\theta}}_{\mathrm{P}}, \tilde{\boldsymbol{\theta}}_{\mathrm{S}}) \geq 0$ since $\mathbf{H}_p$ is positive semidefinite. Also, since $\tilde{\boldsymbol{\theta}}_{\mathrm{S}}$ is a random variable depending on the secondary data, C is a random variable.

We define $\overline{\boldsymbol{\theta}}_{\mathrm{S}} = \tilde{\boldsymbol{\theta}}_{\mathrm{S}}$ and call $\overline{\boldsymbol{\theta}} = (\overline{\boldsymbol{\theta}}_{\mathrm{P}}, \overline{\boldsymbol{\theta}}_{\mathrm{S}})$ the primary component maximum likelihood estimate (PCMLE). We can then write.

$$\overline{\boldsymbol{\theta}} = \mathbf{E}_p\hat{\boldsymbol{\theta}} + \mathbf{E}_s\tilde{\boldsymbol{\theta}},$$

where $$\mathbf{E}_p = \begin{pmatrix} \mathbf{I}_p & \mathbf{H}_{p11}^{-1}\mathbf{H}_{p12} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and $$\mathbf{E}_s = \begin{pmatrix} \mathbf{0} & -\mathbf{H}_{p11}^{-1}\mathbf{H}_{p12} \\ \mathbf{0} & \mathbf{I}_s \end{pmatrix}.$$

We are concerned with the accuracy of the PCMLE, $\overline{\boldsymbol{\theta}}$, relative to the full MLE, $\dot{\boldsymbol{\theta}}$, and to do this we need to know about $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$.

With a large population, it is well known that $\hat{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}(\boldsymbol{\theta}_0, \mathbf{A}_p^{-2})$, where $\mathbf{A}_p^2 = \mathbf{H}_p$. Similarly, we have $\tilde{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}(\boldsymbol{\theta}_0, \mathbf{A}_s^{-2})$, where $\mathbf{A}_s^2 = \mathbf{H}_s$ and $\mathbf{H}_s$ is the information matrix for the secondary data.

Assuming that these distributions apply, we have the following distributional results:

$$\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \dot{\sim} \mathrm{N}(\mathbf{0}, \mathbf{A}_b^{-2}), \tag{2}$$

where

$$\mathbf{A}_b^{-2} = \mathbf{E}_p\mathbf{H}_p^{-1}\mathbf{E}_p^{\mathrm{T}} + \mathbf{E}_s\mathbf{H}_s^{-1}\mathbf{E}_s^{\mathrm{T}},$$

and

$$(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{H}_p(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx X_p + X_s, \tag{3}$$

where $$X_p \sim \chi_p^2$$

and $$X_s = \sum_i \lambda_i' Y_i^2,$$

and where $\mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{p+s})$ and $\lambda_1' \geq \lambda_2' \geq \cdots \geq \lambda_s' \geq 0 = \lambda_{s+1}' = \cdots = \lambda_{s+p}'$ are the eigenvalues of the matrix $\mathbf{A}_s^{-1}\mathbf{E}_s^{\mathrm{T}}\mathbf{H}_p\mathbf{E}_s\mathbf{A}_s^{-1}$.

Eq. (2) tells us about parameter estimates; (3) tells us about how well we have estimated the true underlying epidemic curve up to time T. We anticipate that if the set of primary components has been well chosen, then the "error" term, X_s, will be small relative to X_p (that is, $\lambda_1', \ldots, \lambda_s' \ll 1$).

To facilitate comparison with the results of full maximum likelihood described below, we define $\lambda = (\lambda_1, \ldots, \lambda_{p+s})$, where $\lambda_1 = \cdots = \lambda_p = 1$ and $\lambda_{p+k} = \lambda_k'$ for $k = 1, \ldots, s$.

Suppose, on the other hand, we consider the full MLE, $\dot{\boldsymbol{\theta}}$. Using a similar normal approximation we have:

$$\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \dot{\sim} \mathrm{N}(\mathbf{0}, \mathbf{A}_f^{-2}), \tag{4}$$

where $$\mathbf{A}_f^2 = \mathbf{H}_f$$

and $$\mathbf{H}_f = \mathbf{H}_p + \mathbf{H}_s,$$

and $$(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{H}_p(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx \sum_i \nu_i \mathbf{Y}_i^2 \tag{5}$$

where $\mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{p+s})$ and $1 \geq \nu_1 \geq \cdots \geq \nu_{p+s} \geq 0$ are the eigenvalues of the matrix $\mathbf{A}_f^{-1}\mathbf{H}_p\mathbf{A}_f^{-1}$. If the primary component maximum likelihood method is appropriate, then we should find that of the eigenvalues, ν_i, of $\mathbf{A}_f^{-1}\mathbf{H}_p\mathbf{A}_f^{-1}$, p will be close to 1, and the remaining s will be close to 0, so that

$$(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{H}_p(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \dot{\sim} \chi_p^2.$$

In (2) and (3) we also found that

$$(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{H}_p(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \dot{\sim} \chi_p^2;$$

that is, the accuracy of the model fitting will be improved only marginally if the full process of maximum likelihood is performed instead of primary component maximum likelihood.

Departures from these approximations may occur if:

1. The secondary data contain significant information about some of the primary components. That is, if $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$ and $\mathbf{H}_p\mathbf{v} = \gamma\mathbf{v}$, where γ is large (so $\mathbf{v}$ is predominantly aligned within the space of primary components) and if $(\mathbf{v}^{\mathrm{T}}\mathbf{H}_s\mathbf{v})/\gamma$ is *not* close to zero, then one or more of $\nu_1, \ldots, \nu_p$ will be significantly less than 1. Hence full maximum likelihood will produce a better estimate of the true underlying epidemic curve.

2. The secondary data contain little information about some of the secondary parameters. That is, if $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$ and $\mathbf{H}_p\mathbf{v} = \gamma\mathbf{v}$ where γ is small (so $\mathbf{v}$ is predominantly aligned within the space of secondary parameters) and if $(\mathbf{v}^{\mathrm{T}}\mathbf{H}_s\mathbf{v})/\gamma$ is not very large, then one or more of the $\nu_{p+1}, \ldots, \nu_{p+s}$ will be significantly greater than zero. This will have a similar effect on the primary component estimate by making the error term, X_s, more significant. However, whereas the ν_i will always be bounded above by 1, no such bound exists for the eigenvalues used in X_s. For a given set of primary components this problem is more likely to occur as the size of the population increases. This indicates that, when one of the λ_i exceeds, say, 0.5 or 1, the set of primary components should be reviewed and perhaps enlarged.

The process described here is similar in some respects to the process of model selection (for example, *see* **Akaike's Criteria** [1, 32, 33]). In simple terms, the model selection process keeps adding in extra parameters until the fit of the next model is no longer a significant improvement on the previous, simpler model. This matches the process described here of increasing the set of primary components until the dynamics of the model are no longer sensitive (in a significant way) to changes in the remaining (secondary) parameters.

Projection

A great variety of problems exist here, so that it is only possible to discuss a small but typical subset. Our starting point is that the projected epidemic curve for $t > T$ is $a_t(\dot{\boldsymbol{\theta}})$, using full maximum likelihood, or $a_t(\overline{\boldsymbol{\theta}})$, using primary component maximum likelihood.

We are concerned with the following questions:

1. Given the past data, to what input parameters is the projected epidemic curve sensitive?
2. Are the ranges of projections (confidence bands) for the full MLE and PCMLE approaches similar?

We can also consider how the level of uncertainty in the future can be reduced in the most effective way by carrying out further secondary studies. The framework described here will identify whether or not a proposed study will have the effect of reducing this uncertainty by a significant margin.

Projection of the Epidemic Curve. Suppose we wish to consider the accuracy of the projected curve between times T_0 and T_1 (commonly T, the end of the period of observation, and T_0 will coincide). Again we may use the sensitivity function

$$l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \int_{T_0}^{T_1} \left[a_t(\boldsymbol{\theta}_0) \log a_t(\boldsymbol{\theta}) - a_t(\boldsymbol{\theta})\right] \mathrm{d}t$$
$$\approx l_{\max} - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{H}_q(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where $a_t(\boldsymbol{\theta}_0)$ is the true future incidence rate and $a_t(\boldsymbol{\theta})$ is the projection based on the (PC)MLE, $\boldsymbol{\theta}$.

It is appropriate, first, to carry out a preliminary analysis similar to that described in an earlier section. The starting point will be the set of primary components, $\boldsymbol{\theta}_{\mathrm{P}}$, relevant for the observed epidemic curve up to time T. This set may already, in effect, fix dynamics between times T_0 and T_1 (that is, the projected curve is not sensitive to changes in $\boldsymbol{\theta}_{\mathrm{S}}$). If this is the case, then there is no need to proceed any further and it should be found that the observed epidemic curve will provide enough information to permit an accurate projection. On the other hand, the dynamics between T_0 and T_1 may be sensitive to changes in $\boldsymbol{\theta}_{\mathrm{S}}$. If this is the case, then $\boldsymbol{\theta}_{\mathrm{P}}$ should be enlarged in such a way that it fully describes the dynamics both up to time T and between T_0 and T_1. (For convenience we will call the revised set $(\boldsymbol{\theta}_{\mathrm{P}}, \boldsymbol{\theta}_{\mathrm{Q}})$, where $\boldsymbol{\theta}_{\mathrm{Q}}$ is the set of additional primary components). In these circumstances the observed epidemic curve will not contain much (if any) information about $\boldsymbol{\theta}_{\mathrm{Q}}$. $\boldsymbol{\theta}_{\mathrm{Q}}$ is, however, required for an accurate projection between T_0 and T_1. If we cannot get good estimates of all of $\boldsymbol{\theta}_{\mathrm{Q}}$, then potentially there will be considerable uncertainty in the projected curve. In particular, the projected curve:

1. will be sensitive to changes in $\boldsymbol{\theta}_Q$;
2. will not be sensitive to changes in the remaining reduced secondary parameter set, $\boldsymbol{\theta}_{\mathrm{R}} = \boldsymbol{\theta}_{\mathrm{S}} \backslash \boldsymbol{\theta}_{\mathrm{Q}}$; and
3. may be sensitive to changes in the original set of primary components, $\boldsymbol{\theta}_{\mathrm{P}}$.

The sensitivity to changes in $\boldsymbol{\theta}_Q$ could be reduced if the existing secondary data contain enough information to get a good estimate of $\boldsymbol{\theta}_Q$.

This heuristic approach can be placed in a more rigorous framework in the following way.

Recall that $\overline{\boldsymbol{\theta}} \approx \mathbf{A}_b^{-1}\mathbf{Z}$, where $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{p+s})$ and $\mathbf{A}_b^{-2} = \mathbf{E}_p\mathbf{H}_p^{-1}\mathbf{E}_p^{\mathrm{T}} + \mathbf{E}_s\mathbf{H}_s^{-1}\mathbf{E}_s^{\mathrm{T}}$ [Eq. (2)]. Hence,

$$2(l(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - l(\boldsymbol{\theta}_0, \overline{\boldsymbol{\theta}})) \approx \mathbf{Z}^{\mathrm{T}}\mathbf{A}_b^{-1}\mathbf{H}_q\mathbf{A}_b^{-1}\mathbf{Z} = \sum_i \psi_i Y_i^2,$$

where $\psi_1 \geq \psi_2 \geq \cdots \geq \psi_{p+s} \geq 0$ are the eigenvalues of $\mathbf{A}_b^{-1}\mathbf{H}_q\mathbf{A}_b^{-1}$ and $\mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{p+s})$.

The accuracy of the projection is therefore determined by the magnitudes of the ψ_i. Clearly, the accuracy depends on how well the primary components of the observed epidemic curve relate to the primary components of the projected curve. For example, if the primary data contain good information about all the primary components of the projected curve, then the projection will be accurate. On the other hand, if the primary data contain relatively little information about one or more of the primary components of the projected curve, then the projection may be quite inaccurate (this is a problem with long-term projection based on a limited amount of early incidence data – see [9]).

If accuracy of projection is our objective, then it may be appropriate to choose a number and set of primary components which will be estimated using the observed incidence data which minimizes the eigenvalues of $\mathbf{A}_b^{-1}\mathbf{H}_q\mathbf{A}_b^{-1}$.

Similarly, recall that $\dot{\boldsymbol{\theta}} \approx \mathbf{A}_f^{-1}\mathbf{Z}$, where $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{p+s})$. Hence

$$2[l(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - l(\boldsymbol{\theta}_0, \dot{\boldsymbol{\theta}})] \approx \mathbf{Z}^{\mathrm{T}}\mathbf{A}_f^{-1}\mathbf{H}_q\mathbf{A}_f^{-1}\mathbf{Z} = \sum_i \xi_i Y_i^2$$

where $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_{p+s} \geq 0$ and $\mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{p+s})$.

Because $\mathbf{H}_q$ is quite independent of $\mathbf{H}_p$ and $\mathbf{H}_s$, it is impossible to make any theoretical remarks on the relative accuracy of the two methods of projection. This can only be done by considering specific examples.

Sensitivity Analysis of the Projected Epidemic Curve. From the earlier discussion, it is necessary only to carry out a sensitivity analysis with reference to the set of primary components $(\boldsymbol{\theta}_P, \boldsymbol{\theta}_Q)$. Such an analysis could be carried out in the same way as in the section "Sensitivity Analysis for Single Quantities". One immediate advantage of having first carried out a reparameterization to separate out the primary components is that we now find that the small set of primary components will explain a very large part of the variation in dynamics. The usual form of sensitivity analysis would have found that knowledge of a much larger number of the basic parameters would be required to get the same level of accuracy.

This reduction in the number of significant components can be put to advantage. Instead of carrying out a sensitivity analysis of the full parameter set it is only necessary to consider the primary components. The preceding analysis took a quadratic approximation around the PCMLE. It may be felt, however, that this approximation may not be adequate within the range of values for $(\boldsymbol{\theta}_P, \boldsymbol{\theta}_Q)$. Typically, the number of primary components will be as small as three or four. This allows quite a wide range of combinations to be investigated, far more than would be possible to consider if the full parameter set was being investigated.

This can be done in a number of ways:

1. Latin hypercube sampling can be employed, followed by an analysis of the rank and partial rank correlation coefficients.
2. A regular lattice framework around the PCMLE, as described by Cox & Medley [11] and Cairns [9], can be used, taking into account the relative likelihood of each point on the lattice.
3. Or a quadratic approximation around the PCMLE can be assumed.

The small number of primary components, in particular, allows investigation of interactions between different combinations of components and over the full range. Method 1 can be employed in an investigation of the sensitivity of single quantities of interest to changes in the primary components. Methods 2 and 3 can be applied to the sensitivity function defined in the previous subsection. If there are only two or three primary components, then it is possible to present the results of the sensitivity analysis graphically, for example using contour plots. Methods 2 and 3 are also appropriate for the construction of **confidence intervals** [9–11].

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*. Akademia Kiado, Budapest, pp. 267–281.

[2] Anderson, R.M., Blythe, S.P., Gupta, S. & Konings, E. (1989). The transmission dynamics of the human immunodeficiency virus type 1 in the male homosexual community in the United Kingdom: the influence of changes in sexual behaviour, *Philosophical Transactions of the Royal Society of London, Series B* **325**, 45–98.

[3] Bailey, N.T.J. (1993). An improved hybrid HIV/AIDS model geared to specific public health data and decision making, *Mathematical Biosciences* **117**, 221–237.

[4] Bailey, N.T.J. & Duppenthaler, J. (1980). Sensitivity analysis in the modelling of infectious disease dynamics, *Journal of Mathematical Biology* **10**, 113–131.

[5] Becker, N.G. & Egerton, L.R. (1993). A transmission model for HIV infection with application to the Australian epidemic, *Mathematical Biosciences* **119**, 205–224.

[6] Blower, S.M. & Dowlatabadi, H. (1994). Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example, *International Statistical Review* **62**, 229–243.

[7] Blower, S.M., Hartel, D., Dowlatabadi, H., Anderson, R.M. & May, R.M. (1991). Sex, drugs and HIV: a mathematical model for New York City, *Philosophical Transactions of the Royal Society of London, Series B* **331**, 171–187.

[8] Cairns, A.J.G. (1990). Epidemics in heterogeneous populations. II: Non-exponential incubation periods and variable infectiousness, *IMA Journal of Mathematics Applied in Medicine and Biology* **7**, 219–230.

[9] Cairns, A.J.G. (1991). Model fitting and projection of the AIDS epidemic, *Mathematical Biosciences* **107**, 451–489.

[10] Cairns, A.J.G. (1995). Primary components of epidemic models, in *Epidemic Models: Their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge.

[11] Cox, D.R. & Medley, G.F. (1989). A process of events with notification delay and the forecasting of AIDS, *Philosophical Transactions of the Royal Society of London, Series B* **325**, 135–145.

[12] Daniel, W.W. (1990). *Applied Nonparametric Statistics*. PWS-Kent, Boston.

[13] Diekmann, O., Dietz, K. & Heesterbeek, J.A.P. (1991). The basic reproductive ratio for sexually transmitted diseases: 1. theoretical considerations, *Mathematical Biosciences* **107**, 325–339.

[14] Dietz, K. (1988). On the transmission dynamics of HIV, *Mathematical Biosciences* **90**, 397–414.

[15] Dietz, K., Heesterbeek, J.A.P. & Tudor, D.W. (1993). The basic reproductive ratio for sexually transmitted diseases. Part 2: Effects of variable HIV infectivity, *Mathematical Biosciences* **117**, 35–47.

[16] Hearne, O. (1987). An approach to resolving the parameter sensitivity problem in system dynamics methodology, *Applied Mathematical Modelling* **11**, 315–318.

[17] Iman, R.L. & Helton, J.C. (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models, *Risk Analysis* **8**, 71–90.

[18] Iman, R.L., Helton, J.C. & Campbell, J.E. (1981). An approach to sensitivity analysis of computer models: Part I – Introduction, input variable selection and preliminary variable assessment, *Journal of Quality Technology* **13**, 174–183.

[19] Jacquez, J.A., Simon, C.P., Koopman, J., Sattenspiel, L. & Perry, T. (1988). Modelling and analyzing HIV transmission: the effect of contact patterns, *Mathematical Biosciences* **92**, 119–199.

[20] Kendall, M., Stuart, A. & Ord, J.K. (1991). *The Advanced Theory of Statistics*, Vol. II. Hafner, New York.

[21] Longini, I.M., Byers, R.H., Hessol, N.A. & Tan, W.Y. (1992). Estimating the stage-specific numbers of HIV infection using a Markov model and back calculation, *Statistics in Medicine* **11**, 831–843.

[22] Longini, I.M., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F. & Hethcote, H.W. (1989). Statistical analysis of the stages of HIV infection using a Markov model, *Statistics in Medicine* **8**, 831–843.

[23] Longini, I.M., Clark, W.S., Gardner, L.I. & Brundage, J.F. (1991). The dynamics of $CD4^+$ T-Lymphocyte decline in HIV infected individuals: a Markov modelling approach, *Journal of AIDS* **4**, 1141–1147.

[24] McKay, M.D., Conover, W.J. & Beckman, R.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics* **21**, 239–245.

[25] Mollison, D. (1984). Simplifying simple epidemic models, *Nature* **310**, 224–225.

[26] Mollison, D., Isham, V. & Grenfell, B. (1994). Epidemics: models and data (with discussion), *Journal of the Royal Statistical Society, Series A* **157**, 115–149.

[27] Morris, M. (1993). Telling tails explain the discrepancy in sexual partner reports, *Nature* **365**, 437–440.

[28] Morris, M. (1993). Data driven network models for the spread of disease, in *Epidemic Models: Their Structure and Relation To Data*, D. Mollison, ed. Cambridge University Press, Cambridge.

[29] Nåsell, I. (1985). *Hybrid Models of Tropical Infections*, *Lecture Notes in Biomathematics* Vol. 59. Springer-Verlag, Berlin.

[30] Owen, A.B. (1994). Controlling correlations in Latin Hypercube samples, *Journal of the American Statistical Association* **89**, 1517–1522.

[31] Park, J.-S. (1994). Optimal Latin Hypercube designs for computer experiments, *Journal of Statistical Planning and Inference* **39**, 95–111.

[32] Schwartz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

[33] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika* **63**, 117–126.

[34] Stein, M. (1987). Large sample properties of simulations using Latin Hypercube sampling, *Technometrics* **28**, 143–151.

(*See also* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**; **Model, Choice of**; **Monte Carlo Methods**)

ANDREW J.G. CAIRNS

Epidemic Models, Spatial

If we are to gain proper understanding of the dispersal and control of diseases such as malaria, rabies, and **AIDS**, then we have to recognize that they develop within a truly spatial framework. The common assumption that individuals mix homogeneously over the whole region available to them stems mainly from mathematical convenience (*see* **Random Mixing**); in real life, we have to accept that both individuals and disease often develop within separate subregions. Classic examples of such spatial catastrophes include: 25 million deaths in fourteenth century Europe from Black Death out of a population of 100 million; the Aztecs lost half their population of 3.5 million from smallpox; around 20 million died in the world influenza pandemic in 1919; whilst millions of people are believed to be currently affected by HIV/AIDS. A particularly interesting case is the spread of one of the world's greatest cholera pandemics, the El Tor strain. It was first identified outside Mecca in 1905, and was later recognized in the 1930s as being endemic in the Celebes. Little was heard of it until 1961, when it suddenly exploded out of the Celebes, reaching India in 1964, and advancing into central Africa, Russia, and Europe by the early 1970s. The total burden of misery and suffering that results from such disease is clearly immense, and any understanding that modeling techniques can bring to alleviate this terrible state of affairs *has* to invoke spatial transmission properties.

Disease is spread through two different mechanisms. First, infected individuals may *migrate* to a different location, thereby infecting susceptibles at this new site. Migration patterns can be truly local (spread of HIV in "shooting galleries"), mid-range (sexual transmission between neighboring cities), or global (spread of human disease through intercontinental travel). Second, the disease itself may spread through *cross-infection*, either locally (between neighboring trees) or globally (aerosol dispersal of plant disease). Some situations may involve both mechanisms, such as the UK outbreaks of foot-and-mouth disease. Hengeveld's account [6] of documented invasion scenarios contains many varied examples, including cholera in North America, stripe rust in wheat, the expansion of cattle egret in North and South America, and rabies in Central Europe.

If migration or cross-infection is highly localized, then infectives/infection may *diffuse* over a continuous region. In contrast, if it results in substantive changes in location, then we either have a *spatial jump* process (plants infected by windblown spores), or a *stepping-stone* process if infection can only occur at specific sites (influenza epidemics in Icelandic coastal settlements).

Given that many populations develop within reasonably well-defined subregions, the stepping-stone approach is a sensible one to consider first. We envisage the process as being spatially distributed amongst n sites, with migration and/or cross-infection being allowed between them. This may involve nearest neighbors, all sites with a common transmission rate, or all sites but with the transmission rate changing with intersite distance (called the *contact distribution*). Such migration scenarios were first posed by Kimura [8] in a genetics context, but substantive theoretical development really began following Bailey's simple birth–death–migration process [1]. In this model, the population develops on an infinite set of colonies (thereby avoiding edge-effect problems), all individuals undergo a simple birth–death process with rates λ and μ, respectively (*see* **Stochastic Processes**), and individuals in colony i can migrate at rate ν_1, ν_2 to the two nearest neighbors $i+1$, $i-1$. For the equivalent general epidemic process, with $X_i(t)$ susceptibles and $Y_i(t)$ infectives in colony i at time t, the infective population at i increases at rate $\beta X_i(t)Y_i(t)$. In the opening stages, $\beta X_i(t) \simeq \beta X_i(0) = \lambda$ (say), so there the two processes are roughly equivalent. Unfortunately, even Bailey's process teeters on the edge of mathematical tractability, so the prospects for making substantial theoretical progress with more complicated spatial epidemic processes are remote. Replacing migration with cross-infection (at rate α_1, α_2) makes this situation even worse, since the infective population birth rate changes to $X_i(t)[\beta Y_i(t) + \alpha_1 Y_{i-1}(t) + \alpha_2 Y_{i+1}(t)]$.

Consider, for example, the recent (nonspatial) upsurge of interest in modeling the population dynamics of the AIDS epidemic. Much of the mathematical development is deterministic (*see* **Epidemic Models, Deterministic**), though this does facilitate the allowance of many sources of change [7]. One surprisingly tractable nonlinear model is that of Ball & O'Neill [2], and to place this within a spatial nearest-neighbor setting, let $x_i(t)$, $y_i(t)$, and $z_i(t)$ denote the

number of susceptible, HIV-infected, and removed (i.e. full-blown AIDS or dead) individuals at site i. Then allowing for the migration of infectives gives rise to the deterministic representation

$$\begin{aligned}\frac{dx_i}{dt} &= \frac{-\beta x_i y_i}{x_i + y_i},\\ \frac{dy_i}{dt} &= \frac{\beta x_i y_i}{x_i + y_i} - (\nu_1 + \nu_2)y_i + \nu_1 y_{i-1} + \nu_2 y_{i+1},\\ \frac{dz_i}{dt} &= \gamma y_i. \end{aligned} \tag{1}$$

This situation is in marked contrast with the spatial general epidemic model with cross-infection, with

$$\begin{aligned}\frac{dx_i}{dt} &= -x_i[\beta y_i + \alpha_1 y_{i-1} + \alpha_2 y_{i+1}],\\ \frac{dy_i}{dt} &= x_i[\beta y_i + \alpha_1 y_{i-1} + \alpha_2 y_{i+1}] - \gamma y_i,\\ \frac{dz_i}{dt} &= \gamma y_i. \end{aligned} \tag{2}$$

Such equations are easily modified to enable general migration at rate ν_{ij} from site i to site j, and cross-infection at rate α_{ij} between infectives in site i and susceptibles at site j. Exact solution is usually not possible, though approximate results may be obtained using careful linearization procedures: for numerical solutions use MATLAB, and so on. Often, we are interested in qualitative, rather than quantitative, behavior, and visual inspection of graphical output over a range of parameter settings is usually sufficient to highlight the most important aspects of the process.

Although the propagation of an epidemic through towns or villages is easily visualized in terms of a stepping-stone process, for disease dispersal in animals or plants, a diffusion model may be more appropriate. Near the wavefront itself, the number of susceptibles may be assumed to be fairly constant, and so there the process reduces to a simple birth–death process amenable to Skellam's diffusion approach [21]. On describing the infective density at position (u, v) by **Brownian motion** with zero drift and displacement variances $\mathrm{var}[u(1)] = \mathrm{var}[v(1)] = D^2$, we have the polar normal probability density function (pdf) (*see* **Bivariate Normal Distribution**)

$$\phi(r, \theta; t) = (2\pi D^2 t)^{-1} r \exp\left[\frac{-r^2}{2D^2 t}\right]. \tag{3}$$

Since there is no drift, this pdf spreads out in ever-expanding circles, and for an infective population of final size N, the radial velocity $R(t)/t$ is $D\{[2\ln(N)]/t\}^{1/2}$, which decreases as $t^{-1/2}$. For a long timescale, say, several decades, which is the case for fox rabies in Europe and the El Tor cholera strain, we might assume exponential growth at rate ψ, whence N is replaced by $N\exp(\psi t)$ and the velocity now remains constant at $D\{[2\psi\ln(N)]\}^{1/2}$. The combination of population growth and diffusion is essential if spatial expansion is not to fade out.

The diffusion approach involves a poor Taylor series expansion, and so the two scenarios can give rise to substantially different results. For example, with Bailey's birth–death process, the wavefront velocities (for $\lambda > \mu$) are the solutions to the equation [13]

$$\begin{aligned}\nu_1 + \nu_2 + \mu - \lambda &= (c^2 + 4\nu_1\nu_2)^{1/2}\\ &\quad - c\ln\left\{\frac{[c + (c^2 + 4\nu_1\nu_2)^{1/2}]}{(2\nu_1)}\right\}, \end{aligned} \tag{4}$$

while the equivalent diffusion velocities take the much simpler form

$$c_{\mathrm{diff}} = (\nu_1 - \nu_2) \pm \{2(\lambda - \mu)(\nu_1 + \nu_2)\}^{1/2}. \tag{5}$$

These two results are compatible only if $\lambda - \mu \ll \nu_1 + \nu_2$.

Mollison [9] argues strongly that when considering the velocity of spread, one should lean heavily towards using basic linear deterministic models, claiming that their assumptions are relatively transparent, they are easy to analyze, yet they generally give the same velocity as more complex linear stochastic and nonlinear deterministic models. Their relative simplicity allows more freedom to choose a biologically/epidemiologically realistic model, and hence, greatly facilitates examination of the dependence of conclusions on model components. Note, however, that such linear models provide only an upper bound for the velocity of more realistic stochastic nonlinear models. Further, both deterministic and stochastic linear models are usually completely unsuitable for modeling complex features such as the transition to endemicity and endemic patterns. Nonlinear deterministic models may provide useful information regarding the transition to endemicity but they are usually wholly inadequate for fluctuations about an endemic state.

Many useful conclusions from models for spatial spread are sensitive to the assumptions made in formulating and fitting them, and incorporating realistic epidemiological parameters will make exact theoretical analysis impossible to achieve. Such parameters can be framed in terms of the following concepts. The *basic reproductive ratio* R_0 is the mean number of contacts made by an infective, and this plays a crucial role in determining whether an epidemic outbreak can occur (*see* **Reproduction Number**); the *carrying capacity* K, which enters via R_0, denotes the maximum population density. The time T of a typical infection relative to that of its parent infective is called the *generation gap*, and its relative location in space X, the *dispersal distance*; whilst the distribution of T itself is called the *reproduction kernel* and that of X, the *dispersal* or *contact* distribution. The *wavefront velocity* c can then be expressed as a function of $R_0\beta(x, t)$, where $\beta(x, t)$ is a probability kernel describing the joint distribution of X and T; see [9] for details.

Note that when determining *population size*, linearization is a highly suspect technique, since different nonlinear models can have the same linearization (e.g. epidemics with (i) removals and (ii) recovery); though it is strongly conjectured that nonlinear differential equations for population spread will always have the same *velocity* as their linear approximation.

Given that substantial behavioral differences can occur between deterministic and stochastic analyses of the same process, ideally, a deterministic approach should always be performed in parallel with a stochastic analysis. Unfortunately, even the simplest stochastic spatial scenario of a two-site birth–death–migration process produces intractable mathematics. Some degree of success is possible using approximation techniques, such as regarding $\{x_i(t), y_i(t), z_i(t)\}$ as a **multivariate normal distribution** with **moments** obtained from the cumulant equations by replacing third- and higher-order cumulants by zero. Though a far more powerful way of using such moment closure is to evaluate cumulants up to the third- or fourth-order, and then use these in the multivariate saddlepoint approximation, thereby determining a much more realistic approximating probability density function [17]. Any awkward algebraic manipulation may be easily overcome through the use of a **computer algebra** package, whilst direct numerical computation of the original population probability equations presents another option.

The problem with probability "solutions" is that they usually convey information only on population values at a fixed time t. What we really require is the full history of process development. **Simulation** provides the answer, for given the rapidly expanding nature of affordable computer power, moments and probabilities may be obtained using standard **Monte Carlo** procedures. Detailed examples of how to construct simulation code for space–time stochastic models are contained in [14], and these are easily modified to cope with any spatial epidemic construction. No matter how complicated, a process can always be described as a series of events $E_1, E_2, \ldots$ occurring at times $t_1, t_2, \ldots$. First, detail all possible infection, removal, migration, and cross-infection changes. Then, in essence:

1. evaluate the corresponding rates $r_1, r_2, \ldots$ and put $R = r_1 + r_2 + \cdots$;
2. generate two **uniform** U(0,1) random variables U_1 and U_2;
3. select the jth event if $r_1 + \cdots + r_{j-1} \leq U_1 R < r_1 + \cdots + r_j$;
4. evaluate the interevent time $s = -\ln(U_2)/R$;
5. update population sizes and time $t \to t + s$, and return to 1.

Figure 1 shows two simulations of a two-colony Ball & O'Neill process under both migration and cross-infection regimes. At time $t = 0$ there are 100 susceptibles in each colony, with one infective in colony 1 and none in colony 2. For illustration, only one-way spatial rates are used, namely, from colony 1 to 2. Thus, an epidemic in colony 2 has to be kick-started from colony 1 before all the infectives there have been removed. Whilst the deterministic and stochastic developments for cross-infection are broadly comparable, under migration, substantial time-shift differences occur between them, especially in colony 2. Though rough agreement between stochastic and deterministic realizations will usually occur, the problem is one of consistency. Unlike cross-infection, with migration, total colony sizes are not fixed, so individual sites may pass through their threshold population values and thereby undergo considerable behavioral change.

Such differences can become even more marked when the system comprises three or more sites, and susceptibles may both migrate and give birth. For with appropriate parameter values, susceptibles can move ahead of epidemic flare-ups and grow to

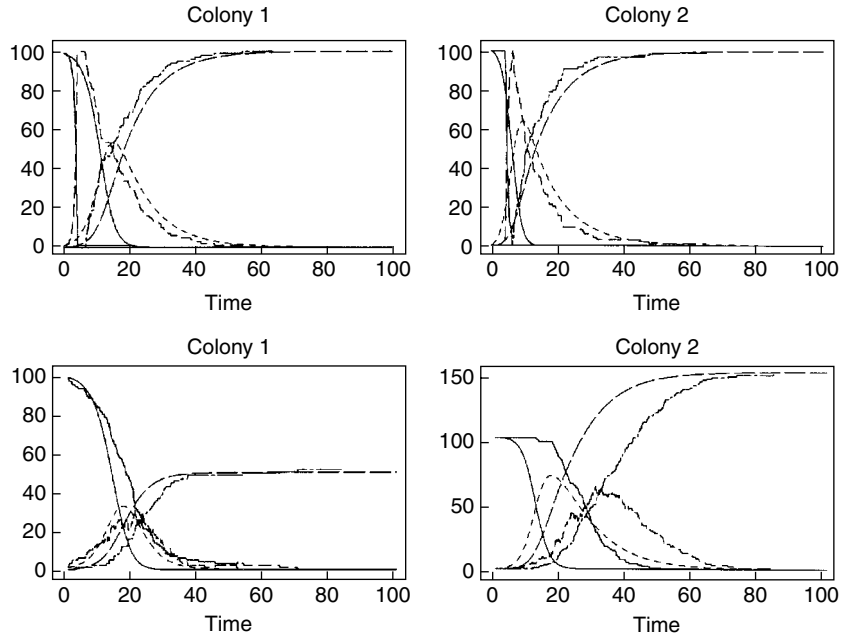

Figure 1 Deterministic (smooth) and stochastic (rough) realizations of a two-colony Ball & O'Neill model under cross-infection (upper) and migration (lower) showing the number of susceptibles (————), infectives (- - - - -) and removals (— — — —): parameter values are $\beta = 0.5$, $\gamma = 0.1$, $\nu_1 = 0.1$, $\nu_2 = 0$, $\alpha_1 = 0.01$, $\alpha_2 = 0$ (produced by Ian Hirsch)

above the local threshold population value before either a migrating infective or cross-infection starts a fresh epidemic outbreak (*see* **Epidemic Thresholds**). Persistence occurs through a *stochastic dynamic*: it is precisely the ability of susceptibles to be constantly on the move recolonizing empty sites, and infectives to pursue them, that keeps the whole process alive. In such situations, we have to rely on simulating individual stochastic realizations. For even if exact probability expressions could be constructed, they would tell us little, being an average over all possible realizations. Moreover, if the behavioral variability between realizations is considerable, then even using a basic deterministic approach can be risky, especially when it relates to epidemic control (*see* **Epidemic Models, Control**). Mollison [9] provides a striking example of this, in which he challenges Murray et al.'s deterministic study [12] of how fox rabies might invade a new country: they predict a roughly circular expanding wave of advance, followed after a quiet phase of about seven years by another wave originating from the same starting point. First, European evidence suggests that after a short while, the rabies invasion could break back across the devastated territory immediately behind it and induce an epidemic equilibrium there. Second, the later wave is an artifact of modeling population size as continuous, rather than discrete. For the model has fox density declining not to zero, but to 10^{-18} of a fox per square kilometer, and this "atto-fox" restarts the epidemic wave as soon as the susceptible population has grown sufficiently large. Though such numerical nonsense may be easily eliminated by replacing any population size below a given cut-off value by zero, the discrepancies between the overall predictions and reality are a serious cause for concern, and highlight the danger in using deterministic models at very low levels of infection prevalence.

The mathematics surrounding spatial stochastic processes is notoriously difficult, and where deterministic solutions can be of considerable help is in determining *qualitative* behavior when there exists an underlying endemic equilibrium level $\{X^*, Y^*\}$ of susceptibles and infectives. In a brilliant pioneering paper, Turing [23] developed elegant deterministic solutions that predict the types of behavior likely to

be encountered when N colonies lie on a ring. In general, let $f(X_i, Y_i)$ and $g(X_i, Y_i)$ denote the rates of change at colony i in susceptibles, $X_i(t)$, and infectives, $Y_i(t)$, respectively. Then if susceptibles and infectives migrate to neighboring sites at rates μ and ν,

$$\frac{dX_i}{dt} = f(X_i, Y_i) + \mu(X_{i+1} - 2X_i + X_{i-1}),$$
$$\frac{dY_i}{dt} = g(X_i, Y_i) + \nu(Y_{i+1} - 2Y_i + Y_{i-1}). \quad (6)$$

On considering local departures from equilibrium by writing $X_i(t) = X^* + x_i(t)$ and $Y_i(t) = Y^* + y_i(t)$, the functions f and g may be approximated by linear forms in x_i and y_i [14, 16]. The resulting equations are amenable to Laplace transform solution, whilst adding white noise (*see* **Noise and White Noise**) to the linearized deterministic equations allows the construction of second-order moments and **spectra** [15]. Cross-infection may be treated similarly. Turing's aim was to examine whether it is feasible to generate *spatially* stable waves, and his idea is simple but profound. For, if in the absence of diffusion, X_i and Y_i tend to a *linearly stable* uniform state, then under certain conditions, spatially inhomogeneous patterns can evolve through *diffusion-driven instability*. Since diffusion is usually considered to be a stabilizing process, care is clearly needed when "guessing" how nonspatial models will behave when they are placed in a spatial environment. Furthermore, the behavior of nonlinear stochastic models can change radically with dimension, as even the number of local sites affected by the migration or cross-infection contact distribution increases markedly as the dimension increases.

Whilst so far we have considered population *numbers* of infected, susceptible, immune, recovered, and so on individuals, for processes that develop over a grid, it is worthwhile highlighting the close link with *percolation processes*. For a wealth of asymptotic theory has been developed (see references in [5]), which can be carried across directly to epidemic scenarios. Here, the information is essentially qualitative, rather than quantitative, with each site being in (say) one of three states, namely, immune, healthy, or infected. Note the close interpretation here with models for "forest fires", which have the equivalent states burned, live, and on fire. Typically, an infected individual emits germs according to a **Poisson process**, which then move to one of the four nearest neighbors chosen at random. If a germ goes to a healthy site, then that site becomes infected and immediately starts to emit more germs, staying infected for a random time with known distribution function until it recovers and is immune to further infection. Questions of interest revolve around the set of sites that will ever become infected if initially the origin is infected and all other sites are healthy. Though this structure lends itself to substantial mathematical analysis, to study time-dependent behavior, we have to revert to using simulation. The advantage of this latter approach is that there is no need to make unrealistic assumptions in order to achieve mathematical tractability, and that with a little practice, computer codes can be developed extremely quickly. A prime example relates to the 2001 UK foot-and-mouth epidemic, whose aftermath left heated discussion over the control policies employed. A simple QBASIC program with good screen graphics output can be developed almost instantaneously to show (for example) an array of farms where each site is either healthy, infected, burned, or culled [18]. Everyday, each healthy site next to an infected site becomes infected itself with probability q; healthy sites neighboring an infected site are culled with probability p; whilst infected sites are burned (i.e. become removed) with probability r. Simulation experiments quickly reveal not only threshold values of p and r for fixed q, above which the disease soon stops but below which the infection keeps on advancing, but also the existence of "creep" in which slow advance relentlessly continues in spite of the process appearing to be under control. This latter behavior was observed for real in parts of the United Kingdom. Had such qualitative features been known at the start of the outbreak, far better control strategies could have been developed, especially since the position, size, and network connections of all farms are held on GIS (geographic information system) **databases**, thereby enabling this simple grid-based simulation exercise to be extended to the UK itself through the development of a more refined space–time structure.

Although the spread of infectives/infection through local migration/contact is commonplace, propagation will often occur between nonnearest colonies. Provided the colonies lie on a regular grid, such as a Turing ring, spatial measures of autocorrelation and frequency may be obtained by using **time-series** techniques [19]. However, sites will often not be regularly spaced: for example, cities, towns, and villages

connected by air, road, and rail; and we need to use weighted measures based upon local population size, area of location, extent of links with other areas, and so on. [4]. We therefore have a space–time bivariate marked point process $\{(X_{u,v}, Y_{u,v}); (u, v) \in R\}$ with association between the locations (u, v) in a region R, local epidemic reactions at each location, and spatial epidemic migration/infection between different locations. The study of such complexity is still in its infancy (see [20] for a single-"species" discussion), and stochastic modeling has to proceed through simulation. Appropriate measures of spatial correlation that are applicable to both marks (X, Y) and points (u, v) can be found in Stoyan and Stoyan's excellent overview [22].

For the purpose of illustration, we have concentrated on purely spatially homogeneous scenarios. However, recent interest in AIDS has stimulated much progress in diverse areas of epidemic modeling, particularly with regard to the treatment of heterogeneity, both between individuals and in mixing of subgroups of the population. The study of epidemics is an exciting, active, and rapidly expanding field, and the review papers of Mollison et al. [11] and Bolker et al. [3] provide excellent starting points for investigating the dynamics of diseases in human, animal, marine, and plant populations. Key theoretical issues are addressed in [10]. Moreover, improved computer technology has led to the availability of better databases and computationally intensive methods in the analysis of data: it has also allowed the simulation of more detailed and realistic models. We can therefore now tackle major challenges to our understanding of spatial epidemics, including the effects of: heterogeneity due to differences between both individuals and mixing; the dependence of persistence on chaotic behavior and spatial patchiness (*see* **Chaos Theory**); nonstationarity due to weather, demographic variables, and evolution; varying migration and cross-infection scenarios; and boundary edge-effects.

References

[1] Bailey, N.T.J. (1968). Stochastic birth, death and migration processes for spatially distributed populations, *Biometrika* **55**, 189–198.

[2] Ball, F. & O'Neill, P. (1993). A modification of the general epidemic motivated by AIDS modelling, *Advances in Applied Probability* **25**, 39–62.

[3] Bolker, B.M., Altmann, M., Ball, F., Barlow, N.D., Bowers, R.G., Dobson, A.P., Elkington, J.S., Garnett, G.P., Gilligan, C.A., Hassell, M.P., Isham, V., Jacquez, J.A., Kleczkowski, A., Levin, S.A., May, R.M., Metz, J.A.J., Mollison, D., Morris, M., Real, L.A., Sattenspiel, L., Swinton, J., White, P. & Williams, B.G. (1995). Group report: spatial dynamics of infectious diseases in natural population, in *Ecology of Infectious Diseases in Natural Populations*, B.T. Grenfell & A.P. Dobson, eds. Cambridge University Press, Cambridge, pp. 399–420.

[4] Cliff, A.D. & Ord, J.K. (1981). *Spatial Processes: Models and Applications*. Pion, London.

[5] Cox, J.T. & Durrett, R. (1988). Limit theorems for the spread of epidemics and forest fires, *Stochastic Processes and Their Applications* **30**, 171–191.

[6] Hengeveld, R. (1989). *Dynamics of Biological Populations*. Chapman & Hall, London.

[7] Isham, V. (1988). Mathematical modelling of the transmission dynamics of HIV infection and AIDS: a review, *Journal of the Royal Statistical Society, Series A* **151**, 5–30.

[8] Kimura, M. (1953). Stepping stone model of population, *Annual Report of the National Institute of Genetics, Japan* **3**, 62–63.

[9] Mollison, D. (1991). Dependence of epidemic and population velocities on basic parameters, *Mathematical Biosciences* **107**, 255–287.

[10] Mollison, D. ed. (1995). *Epidemic Models: Their Structure and Relation to Data*. Cambridge University Press, Cambridge.

[11] Mollison, D., Isham, V. & Grenfell, B. (1994). Epidemics: Models and data, *Journal of the Royal Statistical Society, Series A* **157**, 115–149.

[12] Murray, J.D., Stanley, E.A. & Brown, D.L. (1986). On the spatial spread of rabies among foxes, *Proceedings of the Royal Society of London, Series B* **229**, 111–150.

[13] Renshaw, E. (1977). Velocities of propagation for stepping-stone models of population growth, *Journal of Applied Probability* **14**, 591–597.

[14] Renshaw, E. (1991). *Modelling Biological Populations in Space and Time*. Cambridge University Press, Cambridge.

[15] Renshaw, E. (1994a). The linear spatial-temporal interaction process and its relation to $1/\omega$-noise, *Journal of the Royal Statistical Society, Series B* **56**, 75–91.

[16] Renshaw, E. (1994b). Non-linear waves on the Turing ring, *Mathematical Scientist* **19**, 22–46.

[17] Renshaw, E. (2000). Applying the saddlepoint approximation to bivariate stochastic processes, *Mathematical Biosciences* **168**, 57–75.

[18] Renshaw, E. (2001). From killer bees to nematodes: a stochastic modeller's paradise, in *Spatial Modelling Theme Conference of the Royal Statistical Society*. University of Glasgow, UK, p. 20.

[19] Renshaw, E. & Ford, E.D. (1983). The interpretation of process from pattern using two-dimensional spectral analysis: methods and problems of interpretation, *Applied Statistics* **32**, 51–63.

[20] Renshaw, E. & Särkkä, A. (2001). Gibbs point processes for studying the development of spatial-temporal stochastic processes, *Computational Statistics and Data Analysis* **36**, 85–105.

[21] Skellam, J.G. (1951). Random dispersal in theoretical populations, *Biometrika* **38**, 196–218.

[22] Stoyan, D. & Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics*. Wiley, New York.

[23] Turing, A.M. (1952). The chemical basis of morphogenesis, *Philosophical Transactions of the Royal Society of London, Series B* **237**, 37–72.

(*See also* **Epidemic Models, Stochastic**; **Infectious Disease Models**; **Migration Processes**)

ERIC RENSHAW

Epidemic Models, Stochastic

Although the development of an epidemic in a population susceptible to disease is a **stochastic** (random) **process**, it can often be described with reasonable accuracy by a deterministic model, provided the initial number of susceptibles is sufficiently large (*see* **Epidemic Models, Deterministic**). The deterministic results for the numbers of susceptibles and infectives in the population at time $t \geq 0$, are taken to represent the equivalent means of the more accurate stochastic process.

When the initial number of susceptibles is moderate or small, as in a school or a household, the deterministic model is inadequate and it becomes necessary to rely on a stochastic model. Such a model can be constructed in discrete time $t = 0, 1, 2, \ldots$, with the unit being the latent period of infection, or possibly a day or a week. It may also be constructed in continuous time $t \geq 0$; in both cases the models are usually Markovian (*see* **Markov Chains**; **Markov Processes**). Most of the models assume homogeneous mixing (law of mass action), which states that the probability of a susceptible becoming infected is proportional to the number of possible contacts between susceptibles and infectives, or the product of these in the population at the instant of infection (*see* **Random Mixing**).

Explicit solutions can be found for some of the main stochastic models in use, but, however intractable the problem may be analytically, one can always describe the development of an epidemic by **simulation** methods. In this article we give a brief outline of the most common discrete- and continuous-time stochastic epidemic models.

Stochastic Models in Discrete Time

The most commonly used models in discrete time are the **chain binomial models**; these are due to Reed & Frost (see [1]), and Greenwood [12]. Since the Greenwood model is simpler, we consider it first.

The Greenwood Chain Binomial

In this model, we assume that at time $t = 0$, there are $X(0) = n$ susceptibles subject to an infection which is not dependent on the existing number of infectives in the population. We follow the progress of the epidemic at times $t = 1, 2, \ldots$, the epochs at which the number of infectives and surviving susceptibles are recorded. Suppose that the probability of instantaneous infection of a susceptible at time $t = 0$ is $p < 1$. If each susceptible is infected independently, then the distribution of the remaining susceptibles at time $t = 1$ will be **binomial**, with the probability $q = 1 - p$ of noninfection, so that

$$\Pr\{X(1) = x_1 | X(0) = n\} = \binom{n}{x_1} q^{x_1} p^{n-x_1}, \quad x_1 = 0, 1, 2, \ldots, n. \quad (1)$$

The infectives $Y(1) = n - x_1$ are now removed, and the infection process is repeated for $t = 1, 2, \ldots, T$ until either no further infectives are produced at T, $Y(T) = 0$, or all the susceptibles have been infected, $X(T) = 0$. The infection process then ceases, and T is referred to as the duration of the epidemic. The evolution of the epidemic is dictated by the sequence of binomial distributions, whence the name "chain binomial" for the model. Gani & Jerwood [10] noted that the process $\{X(t); t = 0, 1, \ldots, T\}$ was in fact a simple Markov chain with transition probability matrix

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & \cdot & \cdots & n \end{array} \\ \begin{bmatrix} 1 & 0 & \cdot & \cdots & 0 \\ p & q & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ p^n & \binom{n}{1} p^{n-1} q & \cdot & \cdots & q^n \end{bmatrix} \end{array}. \quad (2)$$

This formulation allows one to carry out simple calculations on the probabilities of such quantities as the number of infectives generated up to time t, or the duration of the epidemic, within the Markov chain framework. For example, the probability of the duration T of the epidemic is given by

$$\Pr\{t = T\} = E' \begin{bmatrix} 0 & \cdot & \cdots & 0 \\ p & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ p^n & \binom{n}{1} p^{n-1} q & \cdots & 0 \end{bmatrix}^{T-1} \begin{bmatrix} 1 \\ q \\ \vdots \\ q^n \end{bmatrix}, \quad (3)$$

where the $1 \times (n+1)$ row vector $E' = \{0 \ldots 0\ 1\}$ indicates that $X(0) = n$, the $(n+1) \times 1$ column vector $\{1\ q \ldots q^n\}'$ gives the probabilities that $Y(t) = 0$ at any time t, and the central matrix is the **P** of (2) with its diagonal elements replaced by zeros. This matrix **geometric distribution** states that $X(t)$ circulates among the states $0, 1, 2, \ldots, n$ for the first $T-1$ epochs, before $Y(T) = 0$ at T.

The Reed–Frost Chain Binomial

In this slightly more complex model, we assume that at time $t = 0$ there are $X(0) = n$ susceptibles and $Y(0) = y_0$ infectives. Infection is now dependent on the *number* of infectives y_0. The probability of contact of each susceptible with an infective is $p < 1$, with a contact resulting in infection; $q = 1 - p$ is the probability of no contact. If each infective is independent of all other infectives, then the probability of at least one infectious contact will be $(1 - q^{y_0})$. If the susceptibles are also independent, then at time $t = 1$, with $x_1 + y_1 = n$,

$$\Pr\{X(1) = x_1, Y(1) = y_1 | X(0) = n, Y(0) = y_0\} = \binom{n}{x_1} (q^{y_0})^{x_1} (1 - q^{y_0})^{y_1}. \quad (4)$$

Note that if q^{y_0} is replaced by q, then (4) reduces to the Greenwood formula (1).

We see that we now have a bivariate Markov chain $\{X(t), Y(t); t = 0, 1, 2, \ldots\}$ which provides us with standard methods for calculating probabilities related to the epidemic. We also remark that for small values of p, the probability of at least one infective contact is $1 - q^{y_0} \sim p y_0$, so that the mean number of new infectives is $p y_0 n$, as for the law of mass action. At $t = 1$, the infectives y_1 are again removed, but not before they can infect the remaining susceptibles, which they are assumed to do instantaneously. The process is now repeated for $t = 2, 3, \ldots, T$ until either $Y(T) = 0$, $X(T) > 0$, or $X(T) = 0$, when the epidemic terminates.

An example for $X(0) = 3$, with $Y(0) = 1, 2$, or 3 may help to visualize the Markov chain more clearly. The transition probability matrix **P** takes the form shown at the top of the opposite page or, in abridged form

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{Q} & \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{Q}^2 & \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \\ \mathbf{Q}^3 & \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} \end{bmatrix}, \quad (5)$$

One can carry out Markov chain calculations with **P** in much the same way as for the Greenwood model. The structure of larger matrices for $n > 3$ is similar, with probabilities of the form (4).

In the present example, the duration T of the epidemic with $X(0) = 3$ and $Y(0) = 1$, will have the distribution

$$\Pr\{t = T\} = E' \times \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ 0 & \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \\ 0 & \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} \end{bmatrix}^{T-1} \begin{bmatrix} \mathbf{I} \\ \mathbf{Q} \\ \mathbf{Q}^2 \\ \mathbf{Q}^3 \end{bmatrix}. \quad (6)$$

where $E' = \{0000,\ 0001,\ 0000,\ 0000\}$.
Here, the initial 1×16 row vector records the values $X(0) = 3$ and $Y(0) = 1$ at $t = 0$, the final 16×4 matrix gives the probabilities that $Y(t)$ is zero at any time t, and the central matrix indicates that $X(t)$ circulates among the states 0, 1, 2, 3, for $T - 1$ epochs before $Y(T) = 0$ at time T.

While the structure of the bivariate Markov chain in the Reed–Frost model is more complex than that of the simple Markov chain in the Greenwood model, both follow the same basic principles. It should be pointed out that for $X(0) = 2$ and $Y(0) = 1$, the two models yield exactly the same probabilities, but this is not the case for larger values of $X(0)$ or $Y(0)$. For further details of these models, the reader is referred to Bailey's treatise [4] and Daley and Gani's monograph [6]. It may be worth mentioning that the models can be modified to allow for immigration into and emigration out of the population subject to infection; for such an example on the spread of HIV (*see* **AIDS and HIV**) among intravenous drug users, see [11].

Stochastic Models in Continuous Time

There are many continuous time models in use, of which three are the most common. The first is the simple epidemic (SI model) in which the population is subdivided into two categories, susceptibles (S) and infectives (I). This is not entirely realistic, but may hold approximately over a short period of time. The second is the general epidemic (SIR model) where there are three categories, susceptibles (S), infectives (I) and removals (R), that is individuals who have recovered and are immune, or who have

$$
\begin{array}{cc|cccc|cccc|cccc|cccc}
 & Y(t+1)= & \multicolumn{4}{c|}{0} & \multicolumn{4}{c|}{1} & \multicolumn{4}{c|}{2} & \multicolumn{4}{c}{3} \\
 & X(t+1)= & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 \\
X(t)= & 0 & 1 & & & & & & & & & & & & & & & \\
 & 1 & & 1 & & & & & & & & & & & & & & \\
Y(t)=0 & 2 & & & 1 & & & & & & & & & & & & & \\
 & 3 & & & & 1 & & & & & & & & & & & & \\
\hline
 & 0 & 1 & & & & 0 & & & & 0 & & & & 0 & & & \\
 & 1 & & q & & & p & 0 & & & 0 & 0 & & & 0 & 0 & & \\
1 & 2 & & & q^2 & & 0 & 2pq & 0 & & p^2 & 0 & 0 & & 0 & 0 & 0 & \\
 & 3 & & & & q^3 & 0 & 0 & 3pq^2 & 0 & 0 & 3p^2q & 0 & 0 & p^3 & 0 & 0 & 0 \\
\hline
 & 0 & 1 & & & & 0 & & & & 0 & & & & 0 & & & \\
 & 1 & & q^2 & & & 1-q^2 & 0 & & & 0 & 0 & & & 0 & 0 & & \\
2 & 2 & & & q^4 & & 0 & 2(1-q^2)q^2 & 0 & & (1-q^2)^2 & 0 & 0 & & 0 & 0 & 0 & \\
 & 3 & & & & q^6 & 0 & 0 & 3(1-q^2)q^4 & 0 & 0 & 3(1-q^2)^2q^2 & 0 & 0 & (1-q^2)^3 & 0 & 0 & 0 \\
\hline
 & 0 & 1 & & & & 0 & & & & 0 & & & & & & & \\
 & 1 & & q^3 & & & 1-q^3 & 0 & & & 0 & 0 & & & & 0 & & \\
3 & 2 & & & q^6 & & 0 & 2(1-q^3)q^3 & 0 & & (1-q^3)^2 & 0 & 0 & & & 0 & 0 & \\
 & 3 & & & & q^9 & 0 & 0 & 3(1-q^3)q^6 & 0 & 0 & 3(1-q^3)^2q^3 & 0 & 0 & (1-q^3)^3 & 0 & 0 & 0
\end{array},
$$

died from the disease. This is a more realistic model for a population of fixed size. The third is the carrier-borne epidemic (CSR model) consisting of infective carriers (C) of the disease who may not know that they are infectious and are gradually dying off, and a separate category of susceptibles (S) subject to infection by the carriers, who are then removed (R) directly after they become infected. We consider each of these in turn.

The Simple SI Epidemic

In this model, we shall consider an initial population consisting of $X(0) = n$ susceptibles, and $Y(0) = 1$ infective for simplicity, subject to homogeneous mixing. At any time $t > 0$, we assume that in any time interval $(t, t + \delta t)$, the infinitesimal transition probability of a further infection is given by

$$
\begin{aligned}
&\Pr\{X(t+\delta t) = x-1,\ Y(t+\delta t) = y+1 | X(t) = x,\\
&\quad Y(t) = y\} = \beta x y \delta t + o(\delta t),
\end{aligned}
\tag{7}
$$

where $y = n + 1 - x$, and β is the infection rate. Note that the infectives remain infectious for all time, and $X(t) + Y(t) = X(0) + Y(0) = n + 1$, so that we need keep track of only one quantity, say $X(t)$ at $t \geq 0$. The equation (7) indicates that $\{X(t); t \geq 0\}$ is a Markov chain in continuous time, namely a death process with the state-dependent death parameter

$$\mu_x = \beta x(n + 1 - x).$$

It is readily shown that the state probabilities $p_x(t) = \Pr\{X(t) = x | X(0) = n\}$ satisfy the system of Kolmogorov forward differential equations

$$
\begin{aligned}
\frac{\mathrm{d}p_x}{\mathrm{d}t} &= \beta(x+1)(n-x)p_{x+1} - \beta x(n+1-x)p_x,\\
&\qquad 0 \leq x \leq n-1,\\
\frac{\mathrm{d}p_n}{\mathrm{d}t} &= -\beta n p_n,
\end{aligned}
\tag{8}
$$

subject to the initial condition $p_n(0) = 1$. Bailey [3] has shown that an explicit solution of these equations is possible by solving the partial differential equation for the **moment generating function** (or the probability **generating function**) of the process, in terms of **hypergeometric** functions, but these prove rather difficult to handle.

A simpler method relies on the use of the Laplace transforms

$$p_x^*(s) = \int_0^\infty \exp(-st)\, p_x(t)\, \mathrm{d}t, \quad \mathrm{Re}(s) > 0,$$

of the state probabilities. From (8), it is easily seen that

$$p_n^*(s) = \frac{1}{s + \beta n}, \tag{9}$$

or $p_n(t) = \exp(-\beta nt)$, and

$$p_x^*(s) = \frac{\beta(x+1)(n-x)}{s+\beta x(n+1-x)} p_{x+1}^*(s), \quad 0 \le x \le n-1. \tag{10}$$

Thus, in principle, the transform $p_x^*(s)$ can be found as

$$p_x^*(s) = \frac{n!(n-x)!}{x!}\beta^{n-x} \prod_{j=x}^{n} \frac{1}{s+\beta j(n+1-j)}, \tag{11}$$

so that $p_x(t)$ may be derived explicitly. Unfortunately, the values $s+\beta j(n+1-j)$ are repeated for $j=n$ and $j=1$, $j=n-1$ and $j=2$, and so on, with the result that $p_x(t)$ is rather more complicated than a simple sum of exponentials.

The problem can be overcome by an approximation which replaces the integer n by the number $N=n+e$, where $e>0$ is some small positive quantity. Then, the Laplace transforms $p_x^*(s)$ of (11) are replaced by the approximate $q_x^*(s)$ of the form

$$q_x^*(s) = \frac{n!(n-x)!}{x!}\beta^{n-x} \prod_{j=x}^{n} \frac{1}{s+\beta j(N+1-j)}, \tag{12}$$

where the values $s+\beta j(N+1-j)$ are now distinct for all values of j. It follows that $q_x(t)$ is a sum of exponentials of the form

$$q_x(t) = \sum_{j=x}^{n} c_{xj} \exp[-\beta j(N+1-j)t], \tag{13}$$

for which the coefficients c_{xj} can be readily evaluated (see [4]). Letting $e \to 0$ in (13) allows us to derive the exact values $p_x(t)$.

Since the random intervals between each infection have negative **exponential** density functions

$$\beta j(n+1-j)\exp[-\beta j(n+1-j)t], \quad 1 \le j \le n,$$

the mean duration of the epidemic has the form

$$\mathrm{E}(T) = \sum_{j=1}^{n} \frac{1}{\beta j(n+1-j)}. \tag{14}$$

This can be approximated by

$$\frac{1}{\beta(n+1)} \int_1^n \left(\frac{1}{x} + \frac{1}{n+1-x}\right) \mathrm{d}x = \frac{2\ln n}{\beta(n+1)}. \tag{15}$$

Kendall [16] has obtained the elegant result that for large values of n, the distribution of $W = (n+1)T - 2\ln n$ can be approximated explicitly by a modified Bessel function of the second kind.

The General SIR Epidemic

This is possibly the most frequently used continuous-time epidemic model; it was foreshadowed in a paper by **McKendrick** [18] and analyzed in more detail by Bartlett [5]. Here the closed population is subdivided into three categories: susceptibles (S), infectives (I) and removals (R), with their initial values being respectively $X(0)=n$, $Y(0)=1$ for simplicity, and $Z(0)=0$, the total population remaining fixed at $n+1$.

We assume that in any time interval $(t, t+\delta t)$, the infinitesimal transition probabilities of the process are given by the probability of a further infection

$$\Pr\{X(t+\delta t) = x-1,\ Y(t+\delta t) = y+1 | X(t)=x,\ Y(t)=y\} = \beta xy\delta t + o(\delta t), \tag{16}$$

precisely as for the SI epidemic in (7), and the probability of a removal

$$\Pr\{X(t+\delta t) = x,\ Y(t+\delta t) = y-1 | X(t)=x,\ Y(t)=y\} = \gamma y\delta t + o(\delta t). \tag{17}$$

These hold for all $0 \le x \le n$, $0 \le y \le n+1-x$, with β as the infection rate and γ as the removal rate, except when the values of $X(t)$ or $Y(t)$ are outside their permissible ranges. Note that $x+y \le n+1$ for all $t \ge 0$; we do not need to keep track of the value of $Z(t)$, since $X(t)+Y(t)+Z(t) = n+1$ for all $t \ge 0$.

In this model, $\{X(t), Y(t); t \ge 0\}$ is a bivariate Markov chain in continuous time. $X(t)$ is a death process with parameter $\mu_{xy} = \beta xy$, dependent on both the number of susceptibles x and the number of infectives y, while $Y(t)$ is a birth and death process with birth parameter $\mu_{xy} = \beta xy$ and death parameter γy (*see* **Stochastic Processes**).

The state probabilities $p_{xy}(t) = \Pr\{X(t) = x, Y(t) = y | X(0) = n, Y(0) = 1\}$ satisfy the forward Kolmogorov differential equations

$$\frac{dp_{xy}}{dt} = \beta(x+1)(y-1)p_{x+1,y-1} - (\beta x + \gamma)yp_{xy} + \gamma(y+1)p_{x,y+1}, \quad 0 \le x \le n, \quad 0 \le y \le n+1-x, \quad (18)$$

with $p_{xy}(t) = 0$ when x or y is outside the permissible range, and $p_{n1}(0) = 1$. Gani [9] was able to obtain an explicit solution for the Laplace transforms of the $p_{xy}(t)$ based on a matrix formulation of the problem, but this is rather complicated, and a simpler approach such as that of Griffiths et al. [13] may prove more suitable in practice.

Quantities of interest are the final size of the epidemic, apart from the initial $Y(0) = 1$, and its distribution for varying parameters β and γ. Bailey [4] lists the probabilities of this final size for $X(0) = 1, 2, 3, 4, 5$, and $Y(0) = 1$, and exhibits graphs for the cases $X(0) = 10, 20, 40$ for increasing values of $\rho = \gamma/\beta$. But perhaps the most illuminating result about the final size of the epidemic is the Threshold Theorem of Whittle [24] (*see* **Epidemic Thresholds**). This is obtained by bounding the stochastic process for the number $Y(t)$ of infectives above and below by birth and death processes with a simpler birth parameter than the actual value βxy. We shall simply quote Whittle's results, which the reader can study in greater depth by reference to his paper.

Assuming an intensity i for the epidemic, so that the final number of infectives other than the initial $Y(0) = 1$ is ni, and writing $\rho = \gamma/\beta$ and

$$\pi_i = \sum_{w=0}^{ni} P_w,$$

where the $P_w = \Pr\{X(\infty) = n - w\}$, $0 \le w \le n$, are the probabilities of a final size w of the epidemic, Whittle [24] proves that for large n

$$\begin{aligned} &\frac{\rho}{n} \le \pi_i \le \frac{\rho}{n(1-i)}, && \text{for } \rho < n(1-i), \\ &\frac{\rho}{n} \le \pi_i \le 1, && \text{for } n(1-i) \le \rho < n, \\ &\pi_i = 1, && \text{for } n \le \rho. \end{aligned} \quad (19)$$

This may be interpreted as stating that if $\rho \ge n$, then there is a zero probability that the epidemic exceeds any preassigned intensity i, while if $\rho < n$, then the probability of an epidemic is approximately $1 - \rho/n$ for small i. Similar results hold for the case where $Y(0) = a > 1$ with ρ/n and $\rho/n(1-i)$ now raised to the power a in the inequalities (19). This threshold theorem is the stochastic analog of Kermack & McKendrick's threshold theorem [17] for the deterministic general epidemic.

The Carrier-Borne CSR Epidemic

In this model, the carriers $U(t)$, $t \ge 0$, form a separate category, with an initial number $U(0) = b \ge 1$. The process $\{U(t); t \ge 0\}$ is a pure death process with parameter $\mu_u = \mu u$, such that the infinitesimal probability of a carrier dying in $(t, t + \delta t)$ is

$$\Pr\{U(t+\delta t) = u - 1 | U(t) = u\} = \mu u \delta t + o(\delta t), \quad 1 \le u \le b,$$

independent of the number of susceptibles in the population. The state probabilities of this process at any time $t \ge 0$ are known to be of the binomial form

$$\Pr\{U(t) = u | U(0) = b\} = \binom{b}{u} [\exp(-\mu u t)][1 - \exp(-\mu t)]^{b-u}, \quad 0 \le u \le b. \quad (20)$$

The susceptibles $X(t)$ are infected by homogeneous mixing with the carriers $U(t)$, and the infinitesimal probability of such an infection in any interval $(t, t + \delta t)$ when $U(t) = u \ge 1$, is

$$\Pr\{X(t+\delta t) = x - 1, U(t+\delta t) = u | X(t) = x, U(t) = u\} = \beta x u \delta t + o(\delta t), \quad 1 \le x \le n,$$

where β is the infection parameter. After becoming infected, a susceptible is removed directly from the population.

The process $\{X(t), U(t); t \ge 0\}$ is a bivariate Markov chain in continuous time, in which $U(t)$ is itself an independent Markov chain which influences the process $X(t)$. If we denote the state probabilities at

time $t \geq 0$ by

$$p_{xu}(t) = \Pr\{X(t) = x, U(t) = u | X(0) = n, \quad U(0) = b\}, \quad 0 \leq u \leq b, \quad 0 \leq x \leq n,$$

we can derive the forward Kolmogorov differential equations of the process as

$$\frac{\mathrm{d}p_{xu}}{\mathrm{d}t} = \beta(x+1)up_{x+1,u} - (\beta x + \mu)up_{xu} + \mu(u+1)p_{x,u+1}, \tag{21}$$

with $p_{xu}(t) = 0$ when x or u are outside their permissible ranges, and $p_{nb}(0) = 1$.

The model was originally formulated by Weiss [23], and solved by him, Dietz [7] and Downton [8]. A straightforward method of solution, also outlined in Bailey [4], involves the derivation of the partial differential equation of the probability generating function obtained from (21). Its solution is found by the method of separation of variables. An alternative method relies on the more general approach presented by Puri [21], and outlined in Daley and Gani [6]. The probabilities $p_{xu}(t)$ are found explicitly as

$$p_{xu}(t) = \binom{n}{x}\binom{b}{u}\sum_{j=x}^{n}(-1)^{j-x}\binom{n-x}{j-x} \times \left(\frac{\mu}{\mu+j\beta}\right)^{b-u}(\exp -u(\mu+j\beta)t) \times [1-\exp -(\mu+j\beta)t]^{b-u}, \tag{22}$$

with the expectation of $X(t)$ given by

$$\mathrm{E}[X(t)] = n\left(\frac{\mu+\beta\exp -(\mu+\beta)t}{\mu+\beta}\right)^{b}. \tag{23}$$

The distribution of the duration time T of the epidemic, which ends when either $U(t) = 0$ or $X(t) = 0$, can also be derived explicitly. The model can be made more complex by making the parameters time-dependent, and also allowing emigration and immigration of both the susceptibles and carriers.

Concluding Remarks

A very large number of stochastic models have been developed for a variety of diseases, including most recently AIDS. These include spatial models for the geographic spread of infections (*see* **Epidemic Models, Spatial**), models for parasitic or host–vector diseases such as malaria and schistosomiasis, and models for sexually transmitted diseases. While each disease may require a slightly different model in order to approximate realism, the principles used in constructing them are similar to those displayed in the small range of examples above.

There is a wealth of recent literature on stochastic epidemic research, and the reader may wish to refer to the recent review paper by Mollison et al. [20], the book of papers on AIDS epidemiology edited by Jewell et al. [15], or the books on more general epidemic models edited by Mollison [19] and Isham & Medley [2, 14, 22].

References

[1] Abbey, H. (1952). An examination of the Reed–Frost theory of epidemics, *Human Biology* **24**, 201–233.

[2] Andersson, H. & Britton, T. (2000). Stochastic Epidemic Models and their Statistical Analysis. Springer Lecture Notes in Statistics 151, New York.

[3] Bailey, N.T.J. (1963). The simple stochastic epidemic: a complete solution in terms of known functions, *Biometrika* **50**, 235–240.

[4] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases*. Griffin, London.

[5] Bartlett, M.S. (1949). Some evolutionary stochastic processes, *Journal of the Royal Statistical Society, Series B* **11**, 211–229.

[6] Daley, D.J. & Gani, J. (1999). *Epidermic Modelling: An Introduction*. Cambridge University Press, Cambridge.

[7] Dietz, K. (1966). On the model of Weiss for the spread of epidemics by carriers, *Journal of Applied Probability* **3**, 375–382.

[8] Downton, F. (1967). Epidemics with carriers: a note on a paper of Dietz. *Journal of Applied Probability* **4**, 264–270.

[9] Gani, J. (1967). On the general stochastic epidemic, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. University of California Press, Berkeley, pp. 271–279.

[10] Gani, J. & Jerwood, D. (1971). Markov chain methods in chain binomial epidemic models, *Biometrics* **27**, 591–604.

[11] Gani, J. & Yakowitz, S. (1993). Modelling the spread of HIV among intravenous drug users, *IMA Journal of Mathematics Applied in Medicine and Biology* **10**, 51–65.

[12] Greenwood, M. (1931). On the statistical measure of infectiousness, *Journal of Hygiene* **31**, 336–351.

[13] Griffiths, J.D., Smedley, J.K. & Weale, T.G. (1987). Terminal distributions along a "Knight's Line" for

a stochastic epidemic, *IMA Journal of Mathematics Applied in Medicine and Biology* **4**, 69–79.

[14] Isham, V. & Grenfell, B.T., eds (1996). *Models for Infectious Human Diseases: Their Structure and Relation to Data*. Cambridge University Press, Cambridge.

[15] Jewell, N.P., Dietz, K. & Farewell, V.T., eds (1992). *AIDS Epidemiology: Methodological Issues*. Birkhauser, Boston.

[16] Kendall, D.G. (1957). La propagation d'une épidémie ou d'un bruit dans une population limitée, *Publications de l'Institut de Statistique de l'Université de Paris* **6**, 307–311.

[17] Kermack, W.O. & McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics I, *Proceedings of the Royal Society, Series A* **115**, 700–721.

[18] McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society* **44**, 98–130.

[19] Mollison, D. (ed.) (1995). *Epidemic Models: Their Structure and Relation to Data*. Cambridge University Press, Cambridge.

[20] Mollison, D., Isham, V. & Grenfell, B.T. (1994). Epidemics: models and data (with discussion), *Journal of the Royal Statistical Society, Series A* **157**, 115–149.

[21] Puri, P.S. (1975). A linear birth and death process under the influence of another process, *Journal of Applied Probability* **12**, 1–17.

[22] Tan Wai-Yuan. (2000). *Stochastic Modeling of AIDS Epidemiology and HIV Pathogenesis*. World Scientific, Singapore.

[23] Weiss, G.H. (1965). On the spread of epidemics by carriers, *Biometrics* **21**, 481–490.

[24] Whittle, P. (1955). The outcome of a stochastic epidemic – a note on Bailey's paper, *Biometrika* **42**, 116–122.

(*See also* **Infectious Disease Models**)

J. GANI

Epidemic Models, Structured Population

The use of mathematical models to describe the spread of **infectious disease** has become a topic of increasing importance in recent years. Carefully formulated models play a vital role in helping to understand the dynamics of disease outbreaks that have already occurred, and in analyzing the effectiveness of potential strategies to deal with future outbreaks. Diseases such as HIV/**AIDS**, **BSE/CJD** and Foot-and-mouth disease have all received considerable modeling attention. However, in order to correctly describe real-life disease dynamics, it quickly becomes apparent that models need to capture something of the mixing behavior of individuals within the at-risk population. For example, it is rarely realistic to assume that an entire population mixes homogeneously, that is, that each currently susceptible individual is equally likely to be infected by any currently infective individual in the population. Predictions based on models defined using unrealistic assumptions should be viewed with considerable caution. Motivated by such concerns, a number of models have been developed that attempt to take into account some elements of structure within the population of interest. In the following, we review some of these models and draw attention to their salient features.

Independent-households Model

For many infectious diseases, the mode of transmission involves the sort of close contact between individuals that would be facilitated by those individuals living or working together in a shared environment. It is therefore natural to consider models in which individuals who commonly spend time together have a different (usually greater) chance of transmitting the disease to one another than individuals who seldom meet. Longini and Koopman [11] describe the following model, in which the population is divided into households, and infections can arise either from infected household members, or from the community at large.

Consider a population of N individuals that is partitioned into households, which need not necessarily be all of the same size. Individuals are assumed to be equally susceptible to contracting the disease in question. Every individual in the population has a fixed probability, q_c, of avoiding infection from outside their household. This probability is assumed to be once-and-for-all, rather than per unit time, and can be thought of as the probability of not acquiring the disease from the community at large for the duration of the epidemic. The fates of different individuals with respect to this community-acquired infection are assumed to be independent. Thus, the number of individuals who are infected from the community has a **binomial distribution** with parameters N and $1 - q_c$.

Within a household initially comprising n susceptible individuals, the disease spreads as follows. First, suppose that a number $Y_0 = a$ of the individuals become infected from the community. If $a = 0$, the household epidemic is over, with no infections occurring. Otherwise, each infective has a probability $1 - q_h$ of transmitting the disease to each of the still-susceptible household members, independently, during their infectious period. Thus, to remain susceptible, a susceptible individual must avoid infection from all a infectives, which occurs with probability q_h^a. Thus, the number of newly created infectives, Y_1 say, has a binomial distribution with parameters $n - a$ and $1 - q_h^a$. Each of these new infectives then has a probability $1 - q_h$ of infecting any one of the remaining $n - Y_0 - Y_1$ susceptibles as before, and so on. The household epidemic continues in this way until zero new infections occur. The within-household epidemic is thus described by a Reed–Frost model [4, Chapter 1] with a initial infectives, $n - a$ susceptibles, and infection probability $1 - q_h$ (*see* **Chain Binomial Model**).

We mention three important features of the above model. First, there are two parameters, q_c and q_h, which control, respectively, community-acquired and within-household-acquired infections. Thus, the model allows consideration of two different routes of infection, and can, for example, be used to determine which route is most important in disease spread in specific applications. Second, the fates of different households are independent of each other, in the sense that the number infected in one household is independent of the number infected in another. Third, the model as described above is essentially not temporal, in the sense that it only describes final outcomes (infected or not), rather than the times at which events occur.

The probability mass function of the final number infected (i.e. the *final size* of the epidemic) in a given

household can be obtained as follows. Let π_{jk} denote the probability that the epidemic produces a total of j infective individuals in a household initially containing k susceptibles. There are $\binom{k}{j}$ ways to select the j individuals who become infected. For the subepidemic propagated among these j individuals, the probability that all become infected is π_{jj}. Additionally, each of the remaining $k-j$ individuals must avoid infection both from the community, and also from the j infected individuals, the probability of these avoidances being $q_c^{k-j} q_h^{(k-j)j}$. We thus obtain that for $k \geq 1$,

$$\pi_{jk} = \binom{k}{j} \pi_{jj} q_c^{k-j} q_h^{(k-j)j}, \quad 0 \leq j < k, \qquad (1)$$

where

$$\pi_{kk} = 1 - \sum_{j=0}^{k-1} \pi_{jk}, \quad \text{and } \pi_{00} = 1. \qquad (2)$$

Equations (1) and (2) can be used to recursively determine the π_{jk}'s. It is also possible to find a closed-form expression for π_{jk} in terms of certain nonstandard polynomials; see [14]. Finally, since households are independent, the probability of observing a particular set of final outcomes in the community of households is obtained by multiplying the relevant π_{jk} values together. Viewed as a function of the model parameters, this probability is simply the **likelihood**, which provides a basis for statistical inference procedures.

The basic model above can be extended in a number of ways, of which we now mention a few. The within-household epidemic can be generalized by removing the independence assumption between the fates of individuals who may be contacted by a given infective 1. This situation arises naturally when one assumes that an infective individual can infect others during an infectious period of random length T. Suppose that, during their infectious period, the individual makes contact with each susceptible in the household at times given by the points of a **Poisson process** of rate β. It is assumed that the Poisson processes corresponding to different susceptible-infective pairs are mutually independent. Then, the probability that a given set of $k-j$ susceptibles escape infection from a single infective during his or her infectious period is simply $\mathrm{E}_T[\exp(-\beta T(k-j))] = \phi(k-j)$, say. Within the nontemporal framework of the model, we can without loss of generality set $\beta = 1$, so that ϕ then depends only upon the parameters of T. The corresponding equation to (1) is obtained by replacing the avoidance probability q_h^{k-j} by $\phi(k-j)$. Note that the original model is obtained when T is constant.

An alternative generalization of the basic model is to introduce some kind of heterogeneity into the population. O'Neill et al. [14] consider a model in which each individual in the population independently has immunity to the disease with some fixed probability v. This could represent, for example, immunity acquired via vaccination, or by previous exposure to the disease. This essentially corresponds to a **random-effects** model for individuals. Alternatively, individuals could be categorized according to some known quantity (e.g. age, or a physiological measure), and then assumed to be susceptible to the disease in a manner that depends on their category. In this setting, extra model parameters are introduced, corresponding to the different categories [1, 12].

Models with Two Levels of Mixing

In the independent-households model described above, the probability of acquiring infection from outside the household is not dynamically altered by the numbers infected within each household. This slightly unrealistic assumption can be overcome by allowing person-to-person transmission both within a household, and in the entire population, at two (possibly) different rates. Ball et al. [8] define the following model. Consider a population of N individuals that is partitioned into households. Initially, all but a small number of individuals are susceptible, and the rest infective. An individual j who becomes infected remains so for a period of time $T^{(j)}$, where $T^{(j)}$ is distributed according to some prespecified nonnegative random variable T. The infectious periods of different individuals are assumed to be mutually independent. During their infectious period, individual j makes contact independently with each susceptible in the population at times given by the points of a Poisson process of rate λ_G/N. The first such contact results in the immediate infection of the susceptible. Simultaneously, and independently, j also has infectious contacts with each member of his or her household according to a Poisson process of rate λ_L. Thus λ_L and λ_G determine, respectively, the

local (i.e. within-household) and global (i.e. with any population member) rates of infection. At the end of its infectious period, j can make no further contacts, and moreover is assumed to be immune to reinfection. The epidemic continues until there are no more infectives remaining in the population. Note that this model is temporal, because all events are explicitly modeled with respect to time.

Note that the scaling factor of N in the global contact rate λ_G/N appears in order to maintain a realistic model as N becomes large. Specifically, as N increases so the Poisson rate of global contacts made by an individual, namely, $N(\lambda_G/N) = \lambda_G$, remains the same. Conversely, it is natural to suppose that as N increases, individual household sizes should remain roughly constant, so that it is the number of households that increases rather than their size. Thus the local infection rate parameter, λ_L, need not be scaled. Note also that the fact that household sizes remain locally "small" as N increases mean that a corresponding deterministic model (*see* **Epidemic Models, Deterministic**), defined by a set of differential equations describing the epidemic within each household, is inappropriate as an approximation of the stochastic model (*see* **Epidemic Models, Stochastic**), and, in particular, has different **epidemic threshold** behavior (see [8] for details).

As might be expected, the interdependence of households within the two-level mixing model complicates any analysis considerably. For example, the exact final size distribution is essentially intractable for any realistic population, because evaluation implicitly involves summing over all possible paths of infection. However, by appealing to asymptotic considerations as the number of households tends to infinity, it is possible to make analytic progress (see [8]). This is essentially due to the fact that, in an infinite population, households contained in any specified finite set are mutually independent of one another. There are two key results. First, a threshold theorem can be obtained, which indicates that the early stages of an epidemic can be approximated by a suitable **branching process**. Moreover, there is a threshold parameter, R_T, such that, in the limit as the number of households tends to infinity, an epidemic of infinite final size can occur with positive probability if and only if $R_T > 1$. Second, there exists a **central limit theorem** for the final size and final severity, where the latter represents the total amount of person-time units of infection. This result essentially says that, when the number of households is large, and conditional upon a major outbreak occurring, the joint distribution of final size and severity can be approximated by a **bivariate normal distribution** whose mean vector and **covariance matrix** depend on λ_L and λ_G. Methods of both classical and **Bayesian** statistical inference for this model have also been considered; see [5] and [13], respectively.

The notion of two levels of mixing, namely, local and global, can be extended to more general models. Examples include the great circle model ([7, 8]) in which individuals are situated on the circumference of a circle, and in which local contacts occur with adjacent or other nearby neighbors, while global contacts are chosen uniformly at random from all individuals in the population. Such a model can be motivated by disease spread among certain kinds of spatially structured populations, such as animals living in lines of pens, or animals that inhabit adjacent stretches of a coastline. Models of this kind are closely related to so-called small-world models [15, 16], which have attracted considerable attention within the Social-Sciences and Physics literatures. Another generalization of the household model is to allow populations to be partitioned in more than one way, for example, by household and school/workplace [6, 9]. Finally, it is also possible to formulate models with three or more levels of mixing, for example, representing within-household, within-city, and between-city mixing.

Epidemics on Graphs

Given a population of individuals, it is mathematically natural to represent their contact structure by means of a graph. Specifically, suppose that each vertex in a graph corresponds to an individual, and each edge, joining two vertices, corresponds to some form of social contact, which is sufficient to allow the possibility of disease transmission between two individuals. Given such a graph, an epidemic process can be defined as follows. A small number of vertices are initially infectious, the rest susceptible. If a vertex j becomes infected, it remains so for a period of time given by some random variable $T^{(j)}$, which is distributed according to some prespecified nonnegative **random variable** T. During its infectious lifetime, a vertex makes infectious contacts with each of its

neighbors in the graph (i.e. those vertices to which j is joined by an edge) at times given by the points of a Poisson process of rate λ. Any such contact results in the immediate infection of the neighbor, provided that the vertex has not been previously infected. At the end of its infectious period, a vertex becomes immune to further reinfection, and essentially plays no further part in the epidemic. All infectious periods and Poisson processes are assumed to be mutually independent. The epidemic continues to spread until there are no more infectives remaining.

The above formulation is clearly very general, and contains several special cases of importance.

Complete Graph

The complete graph (in which every pair of vertices is connected) corresponds to a homogeneously mixing population, and the model reduces to a standard **SIR** (Susceptible-Infected-Removed) model (see [4, Chapter 2].)

Lattice Model

Suppose that vertices are situated at all points (x, y) in the plane, where both x and y are integers, and that the vertex at (x, y) is connected to the four nearest vertices in the plane according to Euclidean distance. Suppose further that the infectious period T is constant, so that $\theta = 1 - \exp(-\lambda T)$ is the probability that a given vertex infects any of its susceptible neighbors, independently of the fate of its other neighbors. This model is equivalent to a percolation model on the two-dimensional integer lattice, in which each "infected" edge corresponds to an open channel, and vertices are either "wet" (no longer susceptible) or "dry" (susceptible). Suppose that, initially, precisely one site is wet. It is well-known from percolation theory that, for $\theta > \frac{1}{2}$, there is a positive probability of an infinite wet connected cluster. Conversely, for $\theta < \frac{1}{2}$, all wet clusters are finite in size with probability one. In terms of the epidemic, this translates to a threshold result stating that infinitely many vertices become infected with positive probability if $\lambda > T^{-1} \log 2$, and with probability zero if $\lambda < T^{-1} \log 2$.

Random Graphs

In some contexts, the exact contact structure of a population is unknown, and thus it is reasonable to also attempt to model this structure stochastically. Andersson [3] considers various cases of this kind, including a simple Bernoulli structure (described below), graphs with predetermined vertex degree sequences, and dynamic graphs in which social structure changes over time (see also [2]). The extent to which such models can be analyzed mathematically varies considerably, depending upon the exact structure considered.

As an example, consider the case where the contact structure is described by a Bernoulli random graph. Specifically, each vertex has a fixed probability p of sharing an edge with any other given vertex, independently of all other vertex pairs. Thus the number of edges emanating from a single vertex has a Binomial distribution with parameters $n - 1$ and p, where n is the total number of vertices in the graph. Setting $p = \mu/n$, so that the average of this Binomial distribution is bounded as $n \to \infty$, and assuming that the infectious period T is **exponentially distributed** with mean γ^{-1}, it can be shown that the threshold parameter of this model is

$$R_0 = \frac{\mu\lambda}{\lambda + \gamma}, \tag{3}$$

see [10]. The quantity R_0 should be interpreted as meaning that, in an infinite population, an epidemic can infect infinitely many individuals with positive probability if and only if $R_0 > 1$.

The temporal behavior of this stochastic model is not straightforward to analyze in detail, although under suitable assumptions the model behaves as a limiting deterministic model as $n \to \infty$. For this deterministic model, it is possible to (numerically) find time-dependent trajectories for the numbers of infective and susceptible individuals at a given time, and to analytically determine the final outcome [3]. Regarding inference reference, [10] contains methods of Bayesian inference, using **Markov chain Monte Carlo** techniques, for the stochastic model.

References

[1] Addy, C.L., Longini, I.M. & Haber, M. (1991). A generalized stochastic model for the analysis of infectious disease final size data, *Biometrics* **47**, 961–974.

[2] Altmann, M. (1995). Susceptible-Infected-Removed epidemic models with dynamic partnerships, *Journal of Mathematical Biology* **33**, 661–675.

[3] Andersson, H. (1999). Epidemic models and social networks, *The Mathematical Scientist* **24**, 128–147.

[4] Andersson, H. & Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*, *Lecture Notes in Statistics* 151, Springer-Verlag, New York.

[5] Ball, F.G. & Lyne, O.D. Statistical inference for epidemics among a population of households. In preparation.

[6] Ball, F.G. & Neal, P. (2002). A general model for stochastic SIR epidemics with two levels of mixing, *Mathematical Biosciences* **180**, 73–102.

[7] Ball, F.G. & Neal, P. (2003). The Great Circle Epidemic Model, Stochastic Processes and Their Applications 107, 233–268.

[8] Ball, F.G., Mollison, D. & Scalia-Tomba, G.-P. (1997). Epidemic models with two levels of mixing, *Annals of Applied Probability* **7**, 46–89.

[9] Becker, N.G. & Dietz, K. (1995). The effect of the household distribution on transmission and control of highly infectious diseases, *Mathematical Biosciences* **127**, 207–219.

[10] Britton, T. & O'Neill, P.D. (2002). Bayesian inference for stochastic epidemics in populations with random social structure, *Scandinavian Journal of Statistics* **29**, 375–390.

[11] Longini, I.M. & Koopman, J.S. (1982). Household and community transmission parameters from final distributions of infections in households, *Biometrics* **38**, 115–126.

[12] Longini, I.M., Koopman, J.S., Haber, M. & Cotsonis, G.A. (1988). Statistical inference for infectious diseases: Risk-specific household and community transmission parameters, *American Journal of Epidemiology* **128**,(4), 845–859.

[13] O'Neill, P.D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods, *Mathematical Biosciences* **180**, 103–114.

[14] O'Neill, P.D., Balding, D.J., Becker, N.G., Eerola, M. & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods, *Applied Statistics* **49**, 517–542.

[15] Watts, D.J. (1999). *Small Worlds*. Princeton University Press, Princeton.

[16] Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks, *Nature* **393**, 440–442.

(*See also* **Epidemic Models, Spatial**)

P.D. O'NEILL

Epidemic Thresholds

The practically most important result to come out of the mathematical theory of epidemics is the threshold theorem, which broadly states that an epidemic can only become established in a population if the initial susceptible population size is larger than some critical value, which depends on the parameters governing the spread of disease. The threshold theorem is important because it immediately tells us what proportion of susceptibles need to be vaccinated in order to prevent an epidemic occurring.

Two broad classes of epidemic models are considered in this article. The majority of the article is devoted to *closed population* epidemic models, which assume that the timescale of the epidemic is sufficiently short so that demographic changes in the population can be ignored. The models considered are of the SIR (susceptible $\rightarrow$ infective $\rightarrow$ removed) type, in which a susceptible individual becomes infected by having "adequate contact" with an infective. It then remains infectious for a while before being removed, by either death, the termination of its infectious period or public health measures. Removed individuals are assumed to be immune to further infection and thus play no further role in the epidemic. The threshold behaviors of a homogeneously mixing deterministic model and its stochastic counterpart, the so-called deterministic and stochastic general epidemics, are considered first, before moving on to more general stochastic models incorporating, for example, more realistic infection mechanisms, heterogeneous populations, and spatial effects. The article closes with a brief description of the threshold behavior of *open population* models, which incorporate demographic effects.

Closed Population Epidemics

General Deterministic Epidemic

The general deterministic epidemic is defined by the following system of differential equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\beta xy, \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = \beta xy - \gamma y, \qquad \frac{\mathrm{d}z}{\mathrm{d}t} = \gamma y, \tag{1}$$

where $x(t)$, $y(t)$, and $z(t)$ denote, respectively, the numbers of susceptible, infectious, and removed individuals at time t, and the parameters β and γ are known as the infection and removal rates (*see* **Epidemic Models, Deterministic**). The model assumes a homogeneously mixing population, with adequate contacts between two given individuals occurring at rate β. Thus if there are x susceptibles and y infectives at time t, there are xy possible contacts, each occurring at rate β, that will result in a new infection occurring; hence the term βxy in (1). The model also assumes that infectious individuals are each removed at rate γ, giving rise to the term γy in (1).

Suppose that at time $t = 0$ there are a infectives, n susceptibles, and no removed cases. It follows from the second formula in (1) that, provided that $y > 0$, $\mathrm{d}y/\mathrm{d}t > 0$ if and only if $x > \rho$, where $\rho = \gamma/\beta$.

Thus a build-up of infection will occur in the population if and only if $n > \rho$. This is part of the celebrated threshold theorem of Kermack & McKendrick [17].

General Stochastic Epidemic

The general stochastic epidemic is obtained by replacing the infinitesimal transition rates governing (1) by infinitesimal transition probabilities (*see* **Epidemic Models, Stochastic**). For $t \geq 0$, let $X(t)$, $Y(t)$, and $Z(t)$ be, respectively, the numbers of infective, susceptible, and removed individuals at time t. Suppose that $(X(0), Y(0), Z(0)) = (n, a, 0)$, so that $X(t) + Y(t) + Z(t) = n + a \ \ (t \geq 0)$. Then the epidemic is completely specified by $\{(X(t), Y(t)); t \geq 0\}$, which is a continuous time **Markov chain** with infinitesimal transition probabilities

$$\Pr\{(X(t+h), Y(t+h)) = (i-1, j+1) | (X(t), Y(t)) = (i, j)\} = \beta ijh + o(h)$$

for an infection, and

$$\Pr\{(X(t+h), Y(t+h)) = (i, j-1) | (X(t), Y(t)) = (i, j)\} = \gamma j + o(h)$$

for a removal.

The epidemic terminates as soon as the number of infectives becomes zero. Let $T = n - X(\infty)$ be the total size of the epidemic; that is, the number of initial susceptibles that are ultimately infected. Note that rescaling the time axis so that the infection rate β is one shows that the distribution of T depends on β and γ only through $\rho = \gamma/\beta$. A system of linear equations governing the distribution of T is given

in Bailey [3, p. 94], in which diagrams illustrating the distribution for various values of n and ρ are also given [3, pp. 98–99]. When $n < \rho$ the distribution of T is unimodal, with the mode at some small argument value (often zero), while if $n > \rho$ the distribution of T is bimodal, with a second mode at a large argument value. Thus again there is a threshold at $n \approx \rho$, although, because of the presence of chance effects, the change in behavior at the threshold is less sharp than in the deterministic model. Also, the value of n at which the distribution of T changes from being unimodal to bimodal is slightly larger than ρ [20].

To understand its threshold behavior it is fruitful to give a more detailed, but equivalent, description of the general stochastic epidemic. The assumptions underlying the general stochastic epidemic are consistent with a model in which infectives behave independently, making contacts at the points of a **Poisson process** with rate $n\beta$ throughout an infectious period that follows a negative **exponential distribution** with mean γ^{-1}. For each contact, the individual contacted is chosen independently and uniformly from the n initial susceptibles. If a contacted individual is susceptible then it becomes infected; otherwise, nothing happens. Clearly, the rate at which an infective is removed is γ, and if there are x susceptibles and y infectives at time t the rate at which infectious contacts are being made is $yn\beta$. However, the probability that a given contact is with a susceptible is x/n. Hence, the rate at which new infections occur is $yn\beta \times x/n = \beta xy$, as required by the general stochastic epidemic.

Note that if all the contacts in the above epidemic were to result in the spread of infection, then the process of infectives would follow a birth-and-death process (*see* **Stochastic Processes**) (see for example, [12, pp. 270–273]) with birth rate $n\beta$ and death rate γ. Of course, it is unlikely that all the contacts made in the epidemic are with susceptibles, so the birth-and-death process is usually only an upper bound to the process of infectives. However, if n is large the probability of contacting a previously contacted individual will be small, particularly in the early stages of the epidemic. Thus, for large n, the early stage of the epidemic is well approximated by the above birth-and-death process and occurrence of a minor/major epidemic may be associated with extinction/nonextinction of the birth-and-death process. Hence, by standard results for birth-and-death processes, the probability that a major epidemic occurs is given by

$$p_{\mathrm{MAJ}} = \begin{cases} 0, & \text{if } n\beta \le \gamma, \\ 1 - (\gamma/n\beta)^a, & \text{if } n\beta > \gamma, \end{cases}$$

so major epidemics can occur only if $n > \rho$.

The above threshold behavior can be made mathematically precise in several ways: see, for example, Whittle [23], who gave the first stochastic epidemic threshold theorem, Williams [24] and Ball [4].

Although the deterministic and stochastic general epidemics both have the same threshold value of $n = \rho$, the interpretation of the threshold behavior is quite different in the two models. In the deterministic model, if $n \le \rho$ ($n > \rho$) minor (major) epidemics will always occur. In the stochastic model, for large n, if $n \le \rho$ minor epidemics always occur, while if $n > \rho$ a major epidemic occurs with a probability lying strictly between 0 and 1.

R_0 and Vaccination Strategies

A unifying concept in the analysis of the threshold behavior of epidemic models is the **reproduction number** (or ratio) R_0 of the epidemic, which is usually defined informally as the expected number of infectious contacts made by a typical infective during its entire infectious period, in a population consisting of susceptibles only (see, for example, [13, 14]). The difficulty in applying this definition for complex models is in determining what is a typical infective. Diekmann et al. [10] show that, for a very broad class of deterministic models, R_0 is given by the maximal **eigenvalue** of an appropriate "next generation" linear operator, thus providing a formal definition. However, in the general stochastic epidemic, it is clear that a typical infective makes infectious contacts at the points of a Poisson process with rate $n\beta$ throughout an infectious period that follows a negative exponential distribution with mean γ^{-1}. Thus, $R_0 = n\beta/\gamma$ and from the previous section major epidemics can only occur if and only if $R_0 > 1$.

Now consider a general epidemic that is above threshold and suppose that a proportion θ of initial susceptibles are vaccinated against the disease being modeled. After vaccination, the initial number of susceptibles is reduced to $n' = (1 - \theta)n$ and hence R_0 is reduced to $R_0' = (1 - \theta)R_0$. Thus major epidemics

will be prevented if $R_0' \leq 1$; that is if

$$\theta \geq 1 - \frac{1}{R_0}.$$

This formula, which gives the critical level of vaccination coverage to prevent an epidemic occurring, holds quite generally for single population epidemic models. It was given first by Smith [22].

General Single Population Epidemic

The approximation of the process of infectives by a birth-and-death process holds for a very wide class of epidemic models, although the approximating process is generally a branching process. Now consider an epidemic, initiated by a infectives among n susceptibles, in which infectious individuals have independent and identically distributed life histories, $H = (T_I, \eta)$, where T_I is the time elapsing between an individual's infection and its eventual removal or death, and η is a point process of times, relative to an individual's infection, at which infectious contacts are made. As before, for each contact, the individual contacted is chosen independently and uniformly from the n initial susceptibles and an infection occurs only if the contacted individual is still susceptible.

If the initial number of susceptibles n is large, the process of infectives can be approximated by a **branching process** (corresponding to the case in which all contacts result in new infections), in which a typical individual lives until age T_I and reproduces at ages according to η. Moreover, the approximation can be made precise in the limit as $n \to \infty$ (see [6]). Let R be the number of contacts made by a typical infective in the epidemic model, let $R_0 = \mathrm{E}(R)$, and let $f(s) = \mathrm{E}(s^R)$ be the probability **generating function** of R. Then, by standard branching process theory (see, for example, [16]), a major epidemic occurs with nonzero probability if and only if $R_0 > 1$ and the probability of a major epidemic is $1 - p^a$, where p is the smallest solution of $f(s) = s$ in [0, 1].

A few examples illustrate the generality of the model:

1. Suppose that η is a Poisson process with rate β. Then R follows a Poisson distribution with random mean βT_I, so $R_0 = \beta \mathrm{E}(T_I)$. Note that if T_I follows a negative exponential distribution with mean γ^{-1}, then the general stochastic epidemic is obtained.
2. In most, if not all, real-life epidemics the infectious period of an infective is preceded by a **latent period** during which a recently infected individual is unable to infect other susceptibles. Let T_L and T_I be random variables describing the lengths of typical latent and infectious periods, respectively. Suppose that η is a Poisson process with rate $\beta(t)$, where

$$\beta(t) = \begin{cases} \beta, & \text{if } T_L < t < T_L + T_I, \\ 0, & \text{otherwise.} \end{cases}$$

   Then, again, $R_0 = \beta \mathrm{E}(T_I)$. Note that the introduction of a latent period does not change R_0.
3. Suppose that η is a Poisson process with random rate $\Lambda(t)$ $(0 \leq t < \infty)$. Then R is **Poisson** with random mean $\int_0^\infty \Lambda(t)\, \mathrm{d}t$, so $R_0 = \int_0^\infty \mathrm{E}[\Lambda(t)]\, \mathrm{d}t$. Such a model might be appropriate for the spread of **AIDS**, as it is known that the infectiousness of an infective varies considerably throughout the long **incubation period** (see, for example, [1]).

General Multipopulation Epidemic

Now consider the spread of an epidemic among a population that is partitioned into m groups, labeled $1, 2, \ldots, m$, with group i consisting initially of a_i infectives and n_i susceptibles. The partitioning of the population into groups could reflect important heterogeneities (such as owing to age, sex, and genotype), geographic location, or a multispecies population, as in host–vector epidemics such as malaria. Infectious individuals have independent life histories, with life histories of infectives in the same group being identically distributed. For $i = 1, 2, \ldots, m$, the life history of a typical group i infective is $H_i = (T_I^{(i)}, \eta_{i1}, \eta_{i2}, \ldots, \eta_{im})$, where $T_I^{(i)}$ denotes the infectious period and, for $j = 1, 2, \ldots, m$, η_{ij} is a point process governing times when infectious contacts are made with group j individuals. For each contact, the individual contacted is chosen independently and uniformly from the initial susceptibles in the contacted group.

If the initial numbers of susceptibles in every group are all large, then the process of infectives approximately follows a multitype branching process, and the approximation can be made precise

in the limit as $n_i \to \infty$ $(i = 1, 2, \ldots, m)$ [5]. For $i, j = 1, 2, \ldots, m$, let R_{ij} be the total number of group j contacts made by a typical group i infective throughout its infectious period. Let $\mathbf{M} = (m_{ij})$ be the $m \times m$ matrix with elements $m_{ij} = \mathrm{E}(R_{ij})$ and let R_0 be the eigenvalue of $\mathbf{M}$ having maximum modulus. Then, by standard multitype branching process theory [18], subject to mild regularity conditions, a major epidemic occurs with nonzero probability if and only if $R_0 > 1$.

The models of the last two sections can be extended to allow for some or all previously infected individuals to become susceptible again, either immediately following their infectious period or at some later time. The process of infectives for such models can be sandwiched between that of the corresponding SIR model and its branching process approximation, so the two models have identical threshold behavior. Models in which all infectives become susceptible immediately following their infectious period are known as SIS (susceptible $\to$ infective $\to$ susceptible) models.

Deterministic versions of the models of the last two sections can usually be written down, although they will often involve a continuous partitioning of the population; for example, to incorporate nonexponential infectious periods. The framework of Diekmann et al. [10] can be used to determine R_0 for such a deterministic model. The value of R_0 will be the same as for the corresponding stochastic model.

Structured Populations

The threshold behavior of the above multipopulation epidemic assumes that all the group sizes are large. Although this may be reasonable in some practical situations, in others it clearly is not. Two such cases are now outlined.

Epidemics Among Households

Consider the spread of an epidemic among a population consisting of m households, each of size n. Suppose that infectives have independent and identically distributed life histories, $H = (T_I, \eta_L, \eta_G)$, where T_I denotes the infectious period of a typical infective, and η_L and η_G are point processes governing times at which *local* and *global* contacts are made, respectively. Each local (global) contact of a given infective is with an individual chosen independently and uniformly from the $n(nm)$ initial individuals in its household (the population).

Becker & Dietz [9] consider the case of highly infectious diseases, and assume that if one individual in a household becomes infected then the whole household becomes infected. Let $\tilde{R}$ be the total number of global contacts emanating from a typical infectious household. Then, provided that the number of households, m is large, the process of infected households can be approximated by a branching process with offspring distribution the distribution of $\tilde{R}$. Thus a major epidemic (one affecting a large number of households) can only occur if $\tilde{R}_0 = \mathrm{E}(\tilde{R}) > 1$. Note that under the above "highly infectious" assumption $\tilde{R}_0 = nR_0$, where now R_0 is the expected number of global contacts made by a typical infective. In general, $\tilde{R}_0 = \mu R_0$, where μ is the expected total size (including the initial infective) of a single household epidemic initiated by one infective, in which global infections are ignored; see Ball et al. [7], where extensions – for example, to unequal household sizes – are discussed.

Spatial Epidemics

A spatial model is often appropriate for plant diseases and also for animal diseases, such as fox rabies (*see* **Epidemic Models, Spatial**). The simplest spatial models usually assume that individuals are located one to each point of a regular lattice and that successive contacts of an infective are with individuals at locations (relative to the infective) chosen independently from a *contact distribution* (see, for example, [19]). The threshold behavior of such epidemics is usually obtained by taking the lattice to be infinite and determining conditions under which a finite initial number of infectives can give rise to an infinite epidemic.

Let R_0 be the expected number of contacts made by a typical infective. For one-dimensional lattices, the epidemic goes extinct with probability one, so no threshold exists. For two-dimensional lattices, there is a critical value of R_0, R_0^{CRIT} say, such that the probability of an infinite epidemic is zero if $R_0 < R_0^{\mathrm{CRIT}}$ and strictly positive if $R_0 > R_0^{\mathrm{CRIT}}$. The existence of R_0^{CRIT} is usually shown by comparing the epidemic with an appropriate percolation process. The value of R_0^{CRIT} depends on the contact distribution. It is

known only in a few very special cases. However, $R_0^{\mathrm{CRIT}} > 1$ and for nearest-neighbor infection models simulations show that $R_0^{\mathrm{CRIT}} \approx 2$.

Note that for both epidemics among a community of households and spatial epidemics, the deterministic and stochastic models will have *different* threshold values, essentially because a deterministic model implicitly assumes that all the group sizes are large. For stochastic models, the threshold values of SIR and corresponding SIS models are now different.

Open Population Epidemics

Consider the general deterministic epidemic, and suppose that susceptibles are recruited into the population at rate ν and all individuals die from natural causes at rate μ. Then the differential equations in (1) become

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -\beta xy - \mu x + \nu, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= \beta xy - (\gamma + \mu)y, \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= \gamma y - \mu z. \end{aligned} \tag{2}$$

Setting $\mathrm{d}x/\mathrm{d}t = 0$ and $\beta = 0$ shows that the disease-free equilibrium population size is $x_0 = \nu/\mu$. Thus the expected number of infectious contacts made by an infective in an otherwise susceptible population is given by $R_0 = \beta\nu/(\gamma + \mu)\mu$, since now infectives are effectively removed at rate $\gamma + \mu$.

Setting $\mathrm{d}x/\mathrm{d}t = \mathrm{d}y/\mathrm{d}t = 0$ in (2) gives the equilibrium numbers of susceptibles and infectives. When $R_0 \leq 1$, the only equilibrium point is the *disease-free* one $(x^*, y^*) = (\nu/\mu, 0)$. Moreover, this equilibrium is globally asymptotically stable, in the sense that $(x(t), y(t)) \to (x^*, y^*)$ as $t \to \infty$, irrespective of the initial values $(x(0), y(0))$. When $R_0 > 1$, there is a second *endemic* equilibrium point $(x^*, y^*) = (\beta^{-1}(\gamma + \mu), (\gamma + \mu)^{-1}\nu - \beta^{-1}\mu)$, and this too is globally asymptotically stable (unless, of course, $y(0) = 0$); see, for example, Hethcote [15] for details. Thus, if $R_0 \leq 1$ the disease cannot become established in the population, while if $R_0 > 1$ it will become established and remain endemic. A similar conclusion holds for a very broad range of open population deterministic epidemic models, including multipopulation models.

The stochastic version of the above model is far more difficult to analyze. Suppose that the disease is introduced into a susceptible population, which is at its disease-free equilibrium level $x_0 = \nu/\mu$. Then, provided that x_0 is sufficiently large, the early stages of the epidemic can still be approximated by a birth-and-death process. Hence, the epidemic will only have a nonzero probability of taking off if $R_0 > 1$. However, even if it does take off, the epidemic will ultimately go extinct with probability one (cf. [21]), although it may take a very long time to do so. Thus, for practical purposes, endemic behavior is possible. However, simulations and observed data on epidemics show that long-term persistence of infection can only occur if the population is larger than some critical level. This has a long history, going back to the pioneering work of Bartlett [8]. The problem of determining the critical community size, for endemic outbreaks to occur, in terms of the parameters of the underlying model still awaits a satisfactory solution [11]. See Andersson and Bitton [2], pp73–77 for a brief discussion of this stochastic model.

References

[1] Anderson, R.M. (1988). The epidemiology of HIV infection: variable incubation plus infectious periods and heterogeneity in sexual activity, *Journal of the Royal Statistical Society, Series A* **151**, 66–93.

[2] Andersson, H. & Bitton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*, *Lecture Notes in Statistics 151*. Springer, New York.

[3] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

[4] Ball, F.G. (1983). The threshold behaviour of epidemic models, *Journal of Applied Probability* **20**, 227–241.

[5] Ball, F.G. (1997). The threshold behaviour of stochastic epidemics, in *Proceedings of the Fourth International Conference on Mathematical Population Dynamics*, to appear.

[6] Ball, F.G. & Donnelly, P.J. (1995). Strong approximations for epidemic models, *Stochastic Processes and Their Applications* **55**, 1–21.

[7] Ball, F.G., Mollison, D. & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing, *Annals of Applied Probability* **7**, 46–89.

[8] Bartlett, M.S. (1956). Deterministic and stochastic models for recurrent epidemics, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* Vol. 4. University of California Press, Berkeley, pp. 81–109.

[9] Becker, N.G. & Dietz, K. (1995). The effect of the household distribution on transmission and control of

highly infectious diseases, *Mathematical Biosciences* **127**, 207–219.

[10] Diekmann, O., Heesterbeek, J.A.P. & Metz, J.A.J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations, *Journal of Mathematical Biology* **28**, 365–382.

[11] Dietz, K. (1995). Some problems in the theory of infectious disease transmission and control, in *Epidemic Models: Their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge, pp. 3–16.

[12] Grimmett, G.R. & Strizaker, D.R. (2001). *Probability and Random Processes*, 3rd Ed. University Press, Oxford.

[13] Heesterbeek, J.A.P. (2002). A brief history of R_0 and a recipe for its calculation, *Acta Biotheretica* **50**, 189–204.

[14] Heesterbeek, J.A.P. & Dietz, K. (1996). The concept of R_0 in epidemic theory, *Statistica Neerlandica* **50**, 89–110.

[15] Hethcote, H.W. (1976). Qualitative analyses of communicable disease models, *Mathematical Biosciences* **28**, 335–356.

[16] Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley, Chichester.

[17] Kermack, W.O. & McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London, Series A* **115**, 700–721.

[18] Mode, C.J. (1971). *Multitype Branching Processes: Theory and Application*. Elsevier, New York.

[19] Mollison, D. (1977). Spatial contact models for ecological and epidemic spread, *Journal of the Royal Statistical Society, Series B* **39**, 283–326.

[20] Nåsell, I. (1995). The threshold concept in stochastic epidemic and endemic models, in *Epidemic Models: Their Structure and Relation to Data*, D. Mollison, ed. Cambridge University Press, Cambridge pp. 71–83.

[21] Ridler-Rowe, C.J. (1967). On a stochastic model of an epidemic, *Journal of Applied Probability* **4**, 19–33.

[22] Smith, C.E.G. (1964). Factors in the transmission of virus infections from animal to man, Scientific Basis of Medicine Annual Review 1964, 125–150.

[23] Whittle, P. (1955). The outcome of a stochastic epidemic – a note on Bailey's paper, *Biometrika* **42**, 116–122.

[24] Williams, T. (1971). An algebraic proof of the threshold theorem for the general stochastic epidemic, *Advances in Applied Probability* **3**, 223.

FRANK BALL

Epidemiology as Legal Evidence

In tort cases concerned with diseases resulting from exposure to a toxic chemical or drug, epidemiologic studies are used to assist courts in determining whether the disease of a particular person, typically the plaintiff, was a result of his or her exposure. This may seem puzzling to scientists, because whenever there is a natural or background rate of an illness one cannot be certain that its manifestation in a specific individual who was exposed to a toxic agent actually arose from that exposure. Indeed, the probability of causation in a specific individual is nonidentifiable [19]. The standard of proof courts utilize in civil cases, however, is the preponderance of the evidence or the "more likely than not" criterion. Thus, scientific evidence that a particular agent can cause a specific disease or set of related diseases in the general population supports an individual's claim that his or her disease came from their exposure. Conversely, scientific studies indicating no increased risk of a specific disease amongst exposed individuals are relied on by defendants, typically producers of the chemical or drug, to support the safety of their product. Similar questions of **causation** arise in cases alleging harm from exposure to hazardous wastes, although the issue in these cases is often whether the exposure was sufficient in magnitude and duration to cause the disease (*see* **Risk Assessment for Environmental Chemicals**).

Epidemiologic studies are also used to determine eligibility for Workers' Compensation, where the issue is whether the employee's disease arose from exposure to an agent in the course of employment [6, p. 831], in regulatory hearings to determine safe exposure levels in the workplace (*see* **Occupational Health and Medicine**), and have even been submitted as evidence in criminal cases [5, p. 153]. We emphasize scientific evidence in tort law, which includes product liability and mass chemical exposure cases, because it is the major area of the law utilizing epidemiologic studies as evidence.

Tort Law

Tort law generally concerns suits for wrongful injury that do not arise from a contract between the parties. Thus, remedies to compensate for injuries from a wide variety of accidents resulting from someone's negligence, e.g. professional malpractice, assault and battery, environmentally induced injury, and **fraud** can be obtained by a successful plaintiff. Product liability is a special area of tort law dealing with the obligations of manufacturers of products to consumers who may suffer personal injury arising from the use of the product.

In any tort claim the plaintiff needs to establish a *prima facie* case by showing that the defendant has a legal duty of care due to the plaintiff and that the defendant breached that duty. In addition, a plaintiff needs to show that (i) she suffered an injury and that the defendant's failure to fulfill its duty of care was the (ii) factual and (iii) legal cause of the injury in question. The law also recognizes defenses that relieve the defendant of liability. The two most prominent ones in tort suits are contributory negligence by the plaintiff and statutes of limitations, which bar suits that are brought after a specified period of time has elapsed from either the time of the injury or the time when the relationship between the injury and the use of the product was known to the plaintiff [8]. In some jurisdictions, especially in Europe [14, p. 834], if the injury results from a defect arising from the product's compliance with a mandatory legal provision at the time it was put on the market, then the manufacturer is not liable. There are substantial differences between jurisdictions as to whether a plaintiff's contributory negligence totally absolves the defendant from liability, reduces it in proportion to the relative fault of the parties, or has no effect on the liability of a defendant whose contribution to the injury was small. In the US, the plaintiff's fault is rarely a complete bar to recovery when the defendant's negligence had a significant role. Similarly, the effective starting date of the limitations period varies among nations and among the states in the US.

When reading actual legal cases that rely on scientific evidence one needs to be aware of the relevant legal rules. For example, although the epidemiologic evidence linking the appearance of a rare form of vaginal cancer in a young woman to her mother's use of diethylstilbestrol during pregnancy is quite strong [1], some states barred plaintiffs from suing because the statute of limitations had expired. Since the cancers were recognized only when the young women passed puberty, typically in the late teens or early twenties, a number of injured women could not

receive compensation. Other states, however, interpreted the limitations period as beginning at the time the plaintiff should have been aware of the connection. In Europe, the European Economic Community (EEC) directive of 1985 provides for a 10-year statute of limitations and allows plaintiffs to file claims within three years after discovering the relationship. Markesinis [14] summarizes the directive and the relevant English and German laws.

Epidemiologic evidence is most useful in resolving the issue of cause in fact, i.e. whether exposure to the product made by the manufacturer, or chemicals spilled onto one's land by a nearby company, can cause the injury suffered by the plaintiff. An alternate formulation of the factual cause issue is whether exposure increases the probability of contracting the disease in question to an appreciable degree. **Case–control studies** were used for this purpose in the litigation surrounding Rely and other highly absorbent tampons [6, p. 840]. Within a year or two after these products were introduced, the incidence of TSS (toxic shock syndrome) amongst women who were menstruating at the time of the illness began to rise sharply. Several studies, cited in Gastwirth [6, p. 918], indicated that the estimated **relative risk** of contracting the disease for users of these tampons was at least 10, which was statistically significant (*see* **Hypothesis Testing**).

In light of the sharp decline in the incidence of TSS after the major brand, Rely, was taken off the market, the causal relationship seems well established and plaintiffs successfully used the studies to establish that their disease was most likely a result of using the product. When only one case–control study, however, indicates an association between exposure and a disease, courts are less receptive. Inskip [12] describes the problems that arose in a British case concerning radiation exposure of workers and leukemia in their children (*see* **Leukemia Clusters**).

There is a rough rule relating the magnitude of the relative risk, R, of a disease related to exposure and the legal standard of preponderance of the evidence, i.e. at least half of the cases occurring amongst individuals exposed to the product in question should be attributable to exposure. As the **attributable risk** is $(R-1)/R$, this is equivalent to requiring a relative risk of at least 2.0. While a substantial literature discusses this requirement (see [20, pp. 1050–1054], [6, Chapters 13 and 14, 24], and [10, pp. 167–170] for discussion and references), courts have been reluctant to adopt it formally, since it would allow the public to be exposed to agents with relative risks just below 2.0 without recourse. The lowest value of R accepted by a court the writer has seen is 1.5, in a case concerning the health effects of asbestos exposure. Courts usually require that the estimated R be statistically significantly greater than 1.0 and have required a **confidence interval** for R but also consider the role of other error rates [10, pp. 153–154]. When a decision must be based on sparse evidence, courts implicitly consider the **power** of a test and may not strictly adhere to significance at the 0.05 level.

The relative risk estimated from typical case–control studies is taken as an average for the overall population. Courts also consider the special circumstances of individual cases and have combined knowledge of the prior health of a plaintiff, the time sequence of the relevant events, the time and duration of exposure, as well as the **latent period** of the disease, with epidemiologic evidence to decide whether or not exposure was the legal cause of a particular plaintiff's disease.

So far, our discussion has dealt with the criteria for factual causality where an injury has already occurred. In some cases concerning exposure to a toxic chemical, plaintiffs have asked for medical monitoring, such as periodic individual exams or a follow-up study. As this is a new development, a specific minimal value of R has not been established.

In product liability law, a subclass of tort, in addition to negligence claims, sometimes one can assert that the manufacturer is subject to strict liability [15, 18]. In strict liability the test is whether the product is unreasonably dangerous, not whether the manufacturer exercised appropriate care in producing the product. Epidemiologic studies indicating a substantial increased risk of a disease can be used to demonstrate that the product is "unreasonably dangerous" from the viewpoint of the consumer.

Some product liability cases concern the manufacturer's duty to warn of dangers that were either known to the manufacturer or could reasonably have been foreseen at the time the product was marketed. In the US, producers are also expected to keep abreast of developments after the product is sold and to issue

a warning and possibly recall the product if **post-marketing** studies show an increased risk of serious disease or injury.

One rationale underlying the duty to warn is informed consent [17, p. 209] (*see* **Ethics of Randomized Trials**). Because asbestos was linked to lung cancer by a major study [23] published in the 1960s, the plaintiff in *Borel vs. Fibreboard Paper Products Corp.*, 493 F. 2d 1076 (5th Cir. 1973) prevailed on his warning claim. The opinion observed that a duty to warn arises whenever a reasonable person would want to be informed of the risk in order to decide whether to be exposed to it.

The time when the risk is known or knowable to the manufacturer is relevant. In *Young vs. Key Pharmaceuticals*, 922 P. 2d 59 (Wash. 1996) the plaintiff alleged that the defendant should have warned about the risk of seizure from the drug given him for asthma. The firm argued that the studies existing in 1979, when the child was injured, were not clinically reliable. Even though subsequent research confirmed those early studies that suggested an increased risk, the court found that the defendant did not have a duty to warn in 1979.

The reverse situation may have occurred in the *Wells* case, 788 F. 2d 741 (11th Cir. 1986). At the time the mother of the plaintiff used the spermicide made by the defendant, two studies had shown an increased risk of limb defects and the court found the firm liable for failing to warn. Subsequent studies, which still may not be definitive, did not confirm the earlier ones, and in a later case the defendant was found not to be liable. While this seems inconsistent from a scientific point of view, from a legal perspective both decisions may be reasonable because the information available at the two times differed.

Government Regulation

Epidemiologic studies are used by regulatory agencies such as the **Food and Drug Administration (FDA)** and Occupational Safety and Health Administration (OSHA) to get manufacturers to recall harmful products or give an appropriate warning. Indeed, the manufacturer of Rely tampons recalled the product after the fourth case–control study linked it to toxic TSS. More recently, case–control studies supported a warning campaign.

In 1982, after a fourth study indicated an association between aspirin use and Reye's syndrome, the FDA proposed a warning label on aspirin containers. The industry challenged the original studies, and the Office of Management and Budget (OMB) asked the FDA [16, 22] to wait for another study. The industry suggested that caretakers of cases would be under stress and might guess aspirin, especially if they had heard of an association, so two new control groups (children hospitalized for other reasons and children who went to an emergency room) were included in the follow-up study [25]. The **odds ratios** (OR) for cases compared with each of these two control groups were about 50, far exceeding those of the school (OR = 9.5) and neighborhood controls (OR = 12.6).

In late 1984 the government, aware of these results, asked for a voluntary warning campaign; a warning was mandatory as of June 1986. The following are the Reye's syndrome cases and fatalities from 1978 to 1989: 1978 (236, 68); 1979 (389, 124); 1980 (555, 128); 1981 (297, 89); 1982 (213, 75); 1983 (198, 61); 1984 (204, 53); 1985 (93, 29); 1986 (101, 27); 1987 (36, 10); 1988 (25, 11); 1989 (25, 11). The cases are graphed in Figure 1. Notice the sharp

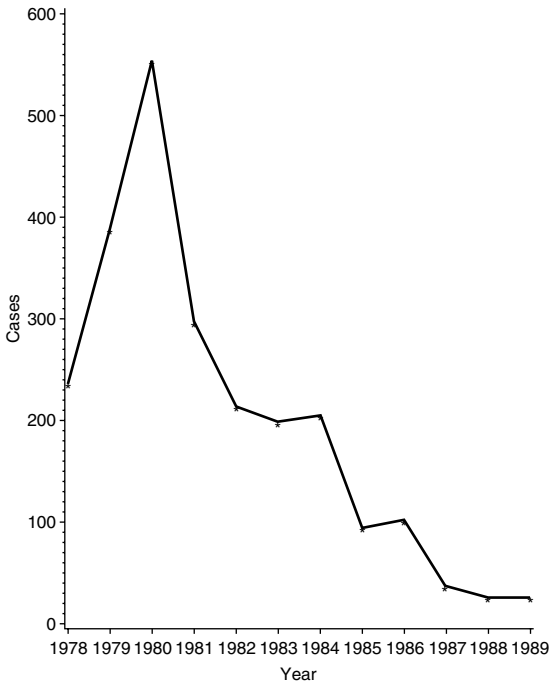

Figure 1 The number of cases of Reye's syndrome for the years 1978–1989

decline between 1983–84 and 1985–86, reflecting the effect of the warning campaign.

Criteria for Admissibility of Studies as Evidence

Courts are concerned with the reliability of scientific evidence, especially as it is believed that lay people may give substantial weight to scientific evidence. In the US, the *Daubert* decision, 113 US 2786 (1993), set forth criteria that courts may use to screen scientific evidence before it goes to a jury. The case concerned whether a drug, Bendectin, prescribed for morning sickness caused birth defects, especially in the limbs. Related cases and the studies are described at length in Green [9]. The *Daubert* decision replaced the *Frye* 293 F. 1013 (DC Cir. 1923) standard, which stated that the methodology used by an expert should be "generally accepted" in the field by the criteria in the Federal Rules of Evidence. The court gave the trial judge a gatekeeping role to ensure that scientific evidence is reliable. Now judges may examine the methodology used and inquire as to whether experts are basing their testimony on peer reviewed studies and methods of analysis before admitting the evidence at trial.

The US Supreme Court decision in *Daubert* remanded the case for reconsideration under the new guidelines for scientific evidence. The lower court, 43 F. 3d (9th Cir. 1995), decided that the expert's testimony did not satisfy the *Daubert* guidelines for admissibility in part because the plaintiff's expert never submitted the **meta-analysis** of several studies, which was claimed to indicate an increased relative risk, for peer review. Similarly, in *Rosen vs. Ciba-Geigy*, 78 F. 3d 316 (7th Cir. 1996), the court excluded expert testimony that a man's smoking while wearing a nicotine patch for three days caused a heart attack. The appeals court said that the expert's opinion lacked the scientific support required by *Daubert* because no study supported the alleged link between short-term use of the patch and heart disease caused by a sudden nicotine overdose. The *Rosen* opinion notes that the trial judge is not to do science but to ensure that when scientists testify in court they adhere to the same standards of intellectual rigor they use in their professional work. If they do so and their evidence is relevant to an issue in the case, then their testimony is admissible, even though the methods used are not yet accepted as canonical in their branch of science.

In two opinions that followed *Daubert, Joiner vs. General Electric*, 522 U.S. 136 (1997) and *Kumho Tire Co. v. Carmichael*, 119 S.Ct. 1167 (1999) the Court expanded the trial judge's role in screening expert testimony for reliability. Now, testimony relying on studies from social science and technical or engineering experience will be subject to review by the judge before the expert is allowed to testify. The *Kumho* opinion noted that the factors mentioned in *Daubert* (e.g. whether the theory or technique on which the testimony is based has been tested, whether it has been subject to peer review and publication, the known or potential error rate) were only a guideline rather than criteria to be strictly applied to prospective expert testimony. In particular, the circumstances of the particular case will have a major role. Commentators [2, 3, 11, 21]) have discussed its implications as well as cases where the circuit courts (covering different regions of the U.S.) have disagreed in their evaluations of similar evidence. Fienberg et al. [4] and Loue [13] discuss the reviewing process, noting some important factors for judges to consider.

Courts have reached different conclusions concerning the admissibility of the method of differential diagnosis, where medical experts conclude that a disease was caused by a particular exposure by eliminating other potential causes. The cases concerning the drug Parlodel and its relationship to stroke, discussed in [7] illustrate the problem. After studies showed that the drug could cause ischemic strokes, some plaintiffs offered expert testimony that these studies showed the drug could cause hemorrhagic strokes too. The Rider v. Sandoz, 295 F. 3d 1194 (11th Cir. 2002) opinion upheld a lower court's rejection of this extrapolation. At the same time it cited favorably Globetti v. Sandoz (111 F. Supp. N.D. Ala 2001) which admitted testimony based on differential diagnosis and also stated that epidemiologic studies are not an absolute requirement. While no human studies had been carried out, animal studies had indicated a risk. The expert was allowed to utilize this information in a differential diagnosis. Thus, the trial judge's assessment of the care and thoroughness with which a differential diagnosis or other scientific study has been carried out by a prospective expert as well as whether the expert has considered all other relevant evidence that is available at the time will be

a major factor in deciding whether the testimony is admissible.

References

[1] Apfel, R.J. & Fisher, S.M. (1984). *To Do No Harm; DES and the Dilemmas of Modern Medicine*. Yale University Press, New Haven.

[2] Berger, M.A. (2000). The Supreme Court's trilogy on the admissibility of expert testimony in *Reference Manual on Scientific Evidence*, Federal Judicial Center, Washington, pp. 1–38.

[3] Faigman, D.L. (2000). The law's scientific revolution: Reflections and ruminations on the law's use of experts in year seven of the revolution. *Washington and Lee Law Review*, **57**, 661–684.

[4] Fienberg, S.E., Krislov, S.H. and Straf, M.L. (1995). Understanding and evaluating scientific evidence in litigation. *Jurimetrics*, **36**, 1–32.

[5] Finkelstein, M.O. & Levin, B. (1990). *Statistics for Lawyers*. Springer-Verlag, New York.

[6] Gastwirth, J.L. (1988). *Statistical Reasoning in Law and Public Policy*. Academic Press, San Diego.

[7] Gastwirth, J.L. (2003). The need for careful evaluation of epidemiologic evidence in product liability cases: A reexamination of *Wells v. Ortho* and *Key Pharmaceuticals* (to appear).

[8] Green, M.D. (1988). The paradox of statutes of limitations in toxic substances litigation, *California Law Review* **76**, 965–1014.

[9] Green, M.D. (1996). *Bendectin and Birth Defects*. University of Pennsylvania Press, Philadelphia.

[10] Green, M.D., Freedman, D.M. & Gordis, L. (2000). Reference guide on epidemiology, in *Reference Manual on Scientific Evidence*. Federal Judicial Center, Washington, pp. 122–178.

[11] Hall, M.A. (1999). Applying *Daubert* to medical causation testimony by clinical physicians. *Toxics Law Reporter*, **14**, 543–552.

[12] Inskip, H.M. (1996). Reay and Hope *versus* British Nuclear Fuels plc: issues faced when a research project formed the basis of litigation, *Journal of the Royal Statistical Society, Series A* **159**, 41–47.

[13] Loue, S. (2000). Epidemiolgical causation in the legal context: Substance and procedures in *Statistical Science in the Courtroom* (Ed.Gastwirth, J.L.), 263–280.

[14] Markesinis, B. (1994). *German Tort Law*, 3rd Ed. Clarendon Press, Oxford.

[15] Markesinis, B. & Deakin, S.F. (1994). *Tort Law*, 3rd Ed. Clarendon Press, Oxford.

[16] Novick, J. (1987). Use of epidemiological studies to prove legal causation: asprin and Reye's syndrome, a case in point, *Tort and Insurance Law Journal* **23**, 536–557.

[17] Phillips, J.J. (1988). *Products Liability*, 3rd Ed. West, St Paul.

[18] Robertson, D.W., Powers, W. Jr & Anderson, D.A. (1988). *Cases and Materials on Torts*. West, St Paul.

[19] Robins, J. & Greenland, S. (1989). The probability of causation under a stochastic model for individual risk *Biometrics* **45**, 1125–1138.

[20] Rubinfeld, D.L. (1985). Econometrics in the courtroom, *Columbia Law Review* **85**, 1048–1097.

[21] Sacks, M.J. (2000). Banishing *Ipse Dixit*: The impact of *Kumho Tire* on forensic identification science. *Washington and Lee Law Review*, **57**, 879–900.

[22] Schwartz, T.M. (1988). The role of federal safety regulations in products liability actions, *Vanderbilt Law Review* **41**, 1121–1169.

[23] Selikoff, I.J., Hammond, E.C. & Churg, J. (1964). Asbestos exposure, smoking and neoplasia, *Journal of the American Medical Association* **188**, 22–26.

[24] Thompson, M.M. (1992). Causal inference in epidemiology: implications for toxic tort litigation, *North Carolina Law Review* **71**, 247–291.

[25] US Public Health Service (1985). Public health service study on Reye's syndrome and medications, *New England Journal of Medicine* **313**, 847–849.

(*See also* **Drug Approval and Regulation**; **Pharmacoepidemiology, Adverse and Beneficial Effects**)

JOSEPH L. GASTWIRTH

Epidemiology, Overview

Epidemiology and biostatistics together constitute the quantitative foundation for public health and clinical research. Epidemiology has been variably defined [20, 25], but all definitions have as essential components the collection and use of data from populations or groups. Epidemiology might be viewed as formulating study designs to provide unbiased evidence for testing hypotheses by applying methods for gathering and using data from populations or groups of people. The domain of epidemiology includes both observation and experiment, although, ethically, the study of injurious factors is limited to observation. The consequences of exposure to potentially injurious agents can only be assessed by comparing disease risks in persons exposed through natural circumstances, including personal choice, with disease risks in those not exposed. Potentially beneficial agents, like chemopreventive micronutrients, might be evaluated using the same observational approaches or, in a clinical trial, by randomly assigning participants to the agent to be tested or to a placebo or other comparison therapy.

The principles of epidemiologic research are not unique to epidemiology and, in fact, permeate other branches of science concerned with human health and well-being: health services research, psychology, sociology and anthropology. Nor can a sharp point of demarcation be drawn between biostatistics and epidemiology. The most basic distinction places statistical aspects of design and data analysis in the domain of biostatistics and overall design and data collection in epidemiology, but the conducting of contemporary epidemiologic research needs integrated efforts from biostatisticians and epidemiologists, and often from clinicians and basic scientists. In addition, since the findings of much epidemiologic research often have immediate applications to clinical and public health policy, considerable media and public attention is directed to epidemiologic studies, frequently before they have been replicated and their results confirmed.

This article provides an overview of the field of epidemiology, setting a context for the more specific articles in this volume. It addresses the history of epidemiology, the purposes of epidemiologic research, the pathways for using epidemiologic evidence to further public health, and the current scope of the field, which is increasingly fragmented into specific areas of inquiry. The other articles in this book provide detailed reviews of different study designs, analytic methods and specifically focused areas of epidemiology.

History of Epidemiology

The beginning of contemporary epidemiology is often dated to the mid-twentieth century, when many large-scale studies were initiated to assess the causes of the shifting pattern of disease in the developed world observed during the preceding decades: rising mortality from seemingly new chronic diseases, like coronary heart disease and lung cancer, even as mortality from infectious diseases declined [39]. The many landmark studies on this theme that gave rise to current approaches are well known to biostatisticians and epidemiologists alike; for example, the Framingham Heart Study initiated in 1949 [11], and the British physicians' study initiated in 1951 [13]. In addition to these cohort studies, case–control studies were also carried out to characterize more quickly and efficiently the causes of the emerging chronic diseases. For example, the first convincing evidence on smoking and lung cancer was derived from case–control studies reported in the 1940s and early 1950s [52]. Cohort studies were also initiated to characterize the consequences of unique exposures, like the study of Japanese atomic-bomb survivors, which still continues today [43].

The origins of epidemiology, however, can be traced back centuries. Society has continually attempted to find the causes of epidemics of disease, whether the plague centuries ago or the sudden appearance of acquired immune deficiency syndrome, (AIDS), only two decades ago [46]. The search for causes, discussed in the third section, is intrinsically linked to the search for cures and avenues for prevention, and now as in the past, epidemiologic evidence remains central to the development of policies to protect and improve the public's health. An epidemiologic perspective is also central to the provision of care for individual patients who need to be cared for in a population context that recognizes the many factors determining their health and disease status. While "evidence-based" medicine and clinical epidemiology have been only recently touted [18, 40], the role of quantitative inference in clinical

medicine was recognized in the nineteenth century by the French physician, Pierre Louis [28, 46].

One element of epidemiology is the description of the occurrence of disease, generally by person, place and time. The counting of disease events can be traced to Graunt who published his book, *Natural and Political Observations Made Upon the Bills Of Mortality*, in 1662 [6]. In this volume, he analyzed the bills of mortality for London, which included the weekly numbers of deaths and their causes, and the numbers of children christened. From these data, he inferred a lifetable for survival in London at the time. His acquaintance, Sir William Petty, also saw the relevance of counting to medicine and he too attempted to estimate life expectancy at birth [6, 46]. Thirty years after the publication of Graunt's book, Edmund Halley, better known for the comet bearing his name, described a lifetable for Breslau and in doing so he showed a clear understanding of population dynamics.

In the nineteenth century, major developments again took place in London. Wlliam Farr advanced counting to a new level through his work at the General Register Office [15, 46]. Farr held responsibility for health statistics for England and Wales and in that capacity he systematically collected and analyzed data, developing new methods and showing the insights into population health that could be gained from valid descriptive data (*see* **Vital Statistics, Overview**). Farr's contemporary, John Snow, undertook investigations of cholera epidemics in London and also practiced anesthesia, giving chloroform to Queen Victoria for childbirth [44]. Snow's investigations of cholera in London, undertaken by the newly founded London Epidemiological Society, led to the determination that cholera was transmitted via contaminated drinking water. Proof of this hypothesis prompted preventive interventions, including the recommendation to remove the handle of the Broad Street pump, a source of contaminated drinking water.

The further rise of epidemiology to the modern era was based in a scientific framework grounded in the emerging recognition of the role of microorganisms in causing disease and the rise of infectious disease epidemiology. In fact, the first principles for evaluating research findings for evidence of causality are often attributed to Robert Koch, although he had benefited from his teacher, Jacob Henle [14]. Koch applied these principles in his identification of the tubercle bacillus as the cause of tuberculosis.

Epidemiology has also used experimental methods. Early eighteenth-century examples include Lind's small trial of fresh fruit to prevent scurvy and Jenner's experimental use of cowpox vaccination to prevent smallpox. Early in the twentieth century, Goldberger conducted experiments that showed pellagra to result from a dietary deficiency, subsequently shown to be a lack of niacin [2, 32]. The contemporary clinical trial originated in the twentieth century as the concept of randomization of participants was introduced and the power of the design was shown in studies of streptomycin for tuberculosis and of vaccination for polio, for example [27, 31].

The first academic department of epidemiology was founded at the Johns Hopkins University School of Hygiene and Public Health in Baltimore in 1919 with the appointment of Wade Hampton Frost [16]. Frost combined interests in infectious diseases and research methods [30] and he saw the relevance of epidemiology to solving problems in public health. His department and its problem-oriented teaching methods became a model for institutions worldwide. Now, schools of public health throughout the world grant master's and doctoral degrees in epidemiology, as do some medical schools.

By the mid-twentieth century, the stage was set for the rise of modern epidemiology: academic departments were established and the new epidemics of coronary heart disease, chronic lung disease, and cancer motivated new research approaches that could address multicaused diseases with lengthy incubation periods and long natural histories. The prospective cohort study was initially the central design for investigating these diseases. Prospectively conducted cohort studies afforded the opportunity to collect data to test multiple hypotheses concerning disease etiology and the strength of this design was quickly shown by the success of the Framingham and other studies in identifying causes of heart disease and through the rapid confirmation that smoking caused lung cancer and other diseases by the studies of British physicians and other groups, including the one million persons enrolled in the American Cancer Society's Cancer Prevention Study (CPS) [41].

The Framingham study is still considered a model for community-based research. Dawber [11] has

chronicled the origins of the study, which was implemented in the late 1940s to address the rising occurrence of cardiovascular disease. The long-term success of the study can be attributed to the selection of a small and cooperative community, sustained support from the National Institutes of Health (NIH), and to the prescience of the original investigators who established rigorous and standardized protocols for data collection. Data were collected relevant to testing the principal extant hypotheses concerning etiology, which were listed at the study's beginning. As a result, much of our initial understanding of risk factors for cardiovascular diseases was based on evidence from this study. Supplementary studies of other diseases capitalized on the opportunity afforded by having the Framingham population under follow-up, and offspring of the original cohort have now been enrolled in a new cohort study that should be informative on familial factors affecting cardiovascular disease risk. The longitudinal data on multiple risk factors necessitated methodologic advances, since appropriate multivariate methods had not been available. For example, Gordon and colleagues [21] described application of discriminant analysis in a 1959 paper.

The cohort design (*see* **Cohort Study**) remains central to observational epidemiologic research, although elaborations of the design have been made to enhance feasibility while reducing costs. For example, large cohorts, like the Nurses' Health Study participants, have been followed primarily by using mailed questionnaires and matching against central registries to determine vital status. It has even been possible to obtain biologic specimens, including toenails for trace metal analysis and blood for DNA, using this approach. In the US, the NIH has taken the lead in establishing multicenter prospective cohort studies, particularly in the area of cardiovascular disease – for example, the Atherosclerosis Risk in Communities (ARIC) study, the Community Heart Study (CHS), and the Strong Heart Study of heart disease in Native Americans. These multisite studies gain external validity by drawing participants from communities across the US. Data collection is standardized and data are accumulated, evaluated and managed at central coordinating centers.

Opportunities for data linkage have now facilitated the conduct of cohort studies. Using record linkage approaches, lists of exposed individuals can be matched for outcome against death indexes and disease registries (*see* **Record Linkage**). Pioneering cohort studies based on this approach were conducted in Canada, where a mortality register of deaths back to 1950 has been available for matching and establishing vital status and cause of death [36]. The National Health and Nutrition Examination Survey (NHANES) conducted by the National Center for Health Statistics has been given a longitudinal component by linkage against death certificates and additional follow-up data collection [10].

The **case–control** design, discussed in detail elsewhere in this volume, is the other principal observational design for testing hypotheses and has also evolved over the same 50 years. Inherent limitations of this design have led some epidemiologists to consider it inferior to the cohort study [2, 17]. Information obtained by interview from cases and controls may be affected by bias; differential bias across cases and controls may create confusing patterns of association. The results of case–control studies conducted among persons selected through a particular institution, e.g. a hospital or clinic, may also be subject to selection bias [4]. Control selection may also be problematic [46] and design principles and feasibility may be in conflict.

There is now substantial understanding of these problems, however, and a methodologic foundation for the case–control design has been firmly established [1]. Cornfield [9] proposed the odds ratio as an estimate of the relative risk for case–control data and Mantel & Haenszel [29] described methods of stratified analyses in their 1959 paper: "Statistical aspects of the analysis of data from retrospective studies of disease". Miettinen further elaborated the underlying principles and analytic methods [33, 34] as did others [7, 42]. More generally, the links between cohort and case–control designs were noted and case-based sampling designs for cohort studies were proposed that unified the two approaches [26, 38].

The case–control design has proved effective for identifying strong causes of disease, such as smoking and lung cancer, diethylstilbestrol and adenocarcinoma of the vagina, and vinyl chloride and angiosarcoma of the liver. The design has been widely applied in research on the etiology of cancer, primarily by using population-based registries to identify cases and sampling to select representative controls. This design is conceptually equivalent to a case–control study nested within a cohort representing all residents of the registry's catchment

area. One landmark study of this design addressed artificial sweeteners and bladder cancer; the study included 3010 cases and 5783 controls [23]. The case–control approach has now also been applied to assess screening and the risks and benefits of therapy.

The randomized trial, an experimental design, is widely held to be the gold standard of population studies. While experimental designs have been used to test therapeutic interventions for several centuries, the modern clinical trial originated in the twentieth century [27]. The use of randomization was advocated in the late 1940s by Austin Bradford Hill and applied in two seminal trials: a test of the pertussis vaccine and an assessment of the efficacy of streptomycin in treating tuberculosis [12]. The clinical trial has since rapidly evolved to include variations in the design and the development of multicenter approaches that make extremely large trials possible.

The core element is the random assignment of subjects to different therapeutic or preventive options. While randomization does not guarantee comparability of the study groups, it does eliminate the potential bias that may result from an investigator's preconceptions. Randomization makes it impossible to predict the assignment of the next person enrolled in the study. While in observational studies we will often match on variables that are known to influence outcomes, the advantage of randomization over matching is that randomization increases the likelihood of comparability of the groups even for factors that influence prognosis but of which we may be unaware or may be unable to measure. Ideally, randomized trials are conducted "blindly" – that is, the subject is unaware of which regimen he is receiving, and the physician or other health care provider does not know to which therapy the individual has been assigned. This is often accomplished by using a placebo, an inert material that looks and tastes like the active drug. At times, however, blinding may be difficult or impossible to implement, a problem that is most significant when the outcome being studied is a subjective one such as pain. In recent years considerable attention has focused on ethical issues pertaining to the use of placebos since using placebos may often involve not offering a currently available agent that is at least partially effective.

The randomized trial has most often been applied to clinical therapies, but has found increasing value for studying the benefits of community-wide interventions with public health measures.

The Specialization of Epidemiology

With the increasing complexity of epidemiologic research, epidemiologists and their areas of inquiry have become increasingly focused and specific. The bifurcation of the field into "infectious disease epidemiology" and "chronic disease epidemiology" no longer holds. The field has become multidimensional with cells defined by disease (e.g. cancer or heart disease), exposure (e.g. environment or nutrition), methods (e.g. genetic or molecular) and problem domain (clinical or outcome). Increasingly, genetics overlays all lines of inquiry, particularly those directed at disease etiology.

The core methods and principles are comparable across these areas of epidemiology, but each has its own special aspects. Of course, in each area there is a specific biomedical substrate, reflecting the exposures and outcomes of interest and the underlying biological phenomena. Additionally, methods for exposure and outcome assessment may be specific to an area. In studying occupation and health there are specific measurement methods for characterizing workplace exposures and the job itself may be used as an exposure surrogate, sometimes with the application of a job-by-exposure matrix [8]. Studies in the domain of genetic epidemiology use specific study designs, often family-based, and analytic methods that characterize patterns of association of disease risk with genetic markers. Other articles in this volume cover clinical epidemiology, **environmental epidemiology**, genetic epidemiology, **nutritional epidemiology**, **occupational epidemiology**, **pharmacoepidemiology** and **risk assessment**.

Epidemiology, Policy and Public Health

The direct linkage of epidemiologic evidence to making policy intended to advance public health is widely acknowledged. Almost universally, epidemiologists tell the story of John Snow and the Broad Street pump as an illustration of the immediacy of observational findings for solving public health problems. This example is particularly compelling because Snow demonstrated the waterborne transmission of cholera before there was knowledge of the existence of the *Vibrio cholerae* organism. There are numerous other examples, also considered as triumphs of epidemiologic inquiry: establishing cigarette smoking as a

cause of lung cancer and other diseases, identifying powerful and remediable causes of cancer like asbestos exposure and diethylstilbestrol administration during pregnancy, and the characterization of risk factors for AIDS.

As a core discipline of biomedical research, epidemiology is not unique in generating evidence relevant to policy. After all, the ultimate goal of all biomedical research is to advance the health of people. Epidemiology, as a scientific method applied directly in the population context, brings evidence that directly bears on the health of the population and this direct linkage is what distinguishes epidemiology from other branches of biomedical research. As a consequence, epidemiologic findings generally have immediate relevance to setting policies pertinent to health and this relevance often gives prominence to epidemiologic evidence in the diverse processes by which policies are made. This prominence has occasioned targeted review and criticism of specific epidemiologic findings and of epidemiology generally. As epidemiologic research has addressed increasingly complex questions concerning the causes of disease, the risks of environmental factors and the benefits of interventions, the resulting evidence may be subject to uncertainties that cloud decision-making, leading some to question the utility of epidemiologic data.

The community of epidemiologic researchers is divided in its view of epidemiology and policy. At one extreme, some would consider epidemiology as being no different from other branches of science where the rationale for research is often given as advancing knowledge; at the other, epidemiologic research would be construed as justified only if the evidence were to be relevant to advancing public health. Epidemiologists are similarly divided in their view of the role of epidemiologists in policy-making processes. Some eschew such involvement and one respected journal, *Epidemiology*, does not allow authors to offer policy recommendations. Others have called for renewed activism by epidemiologists and engagement with the sweeping social problems that underlie many of the increased risks that epidemiologists have elegantly and repetitively described [37, 45]. Even as debate continues, the use of epidemiology for policy purposes is burgeoning with the rise of the outcomes movement and calls for evidence-based medicine, and the need to apply the explosively expanding knowledge of the human genome in clinical and population contexts.

The paths and processes leading from hypothesis to policy are diverse and often lengthy and ill-defined (see Figure 1 and Table 1). In the area of infectious diseases, findings may lead quickly to action; John Snow acted immediately in response to his own findings on the waterborne transmission of cholera. Continuing in this tradition, investigators addressing infectious disease problems make policy recommendations more often than investigators working in other areas [24]. For some areas of inquiry, evidence

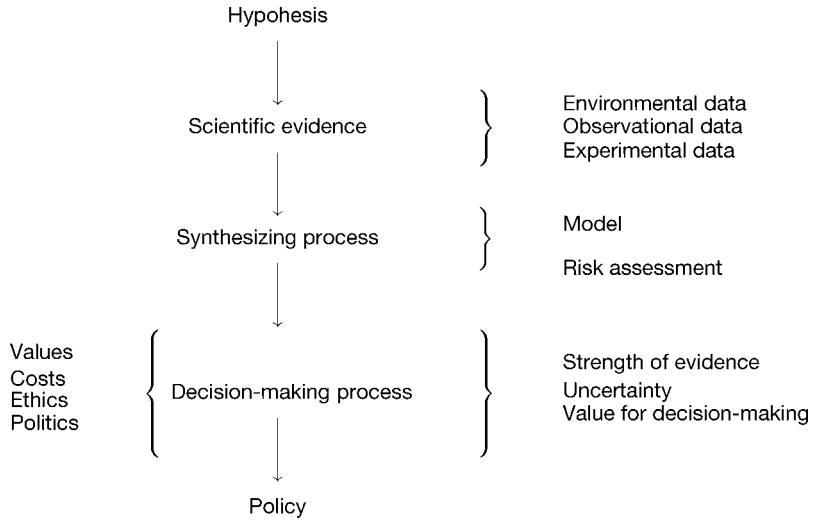

Figure 1 Science/policy interface

Table 1 Some pathways and examples for translation of epidemiologic evidence into policy

Regulatory
• Occupational health and safety
• Environmental quality
• Drug safety
Public health recommendations
• Vaccination
• Diet
• Smoking
Legal system
• Causation of injury
Health care delivery
• Practice guidelines
• Outcome assessment

Table 2 Some processes for translation of epidemiologic evidence into policy

Application of causal criteria
Expert opinion
Consensus methods
Committee review
Quantitative synthesis
Risk assessment
Jury evaluation

may accumulate slowly, e.g. diet and cancer, and only reach a level of certainty sufficient for policy-making after decades of research. Of course, research and policy-making are interactive and iterative, and policies may change as evidence evolves.

Some of the routes for translating epidemiologic and other data into policy are listed in Table 1. They range from formal and structured, as in the requirements of specific regulations, to informal and unstructured, as in the choices that individuals take for their own lifestyles. For example, the 1996 draft cancer policy guidelines of the US Environmental Protection Agency [51] offer instruction for evaluating and interpreting epidemiologic data. Criteria for causality have been rigorously applied in the reports of the Surgeon General on smoking and health [49, 50]. Gail [19] traced the application of these criteria to the evidence on smoking and lung cancer and showed their utility for organizing the relevant lines of evidence and making certain that alternatives to the causal hypothesis could be satisfactorily addressed. Specific actions may be invoked if the evidence reaches a threshold of certainty, e.g. a causal association is found or a target level of risk is reached. Embedded within these translation routes are processes for identifying and evaluating the relevant evidence (Table 2).

New tools for conducting epidemiologic research, together with the increasing capacity to manage and analyze large databases, have made epidemiologic evidence more informative for answering policy-maker's questions. Large administrative databases, such as the Health Care Financing Administration's Medicare files, can be explored to test hypotheses with immediate policy relevance – outcome of myocardial infarction in relation to hospital volume [47] and patterns of care by race and gender [22, 35], for example. Increasingly powerful multivariable methods for data analysis can detect policy-relevant patterns of association with the confidence that the associations are not spurious, while new models for longitudinal data analysis facilitate the capacity to describe disease and its development in time [48].

For many policy issues, the evidence comes from numerous and sometimes heterogeneous studies. Synthesis of such data for policy purposes has often been accomplished by expert review and consensus, tabular summary, or the application of criteria for causality. These processes have proved effective, particularly for strong associations, but uncertainties in the evidence have undermined conclusions, particularly if conclusions weighted by policy are reached. An example is the epidemiologic evidence on passive smoking, which has been the scientific basis for programs to reduce smoking in public places and repeatedly questioned by the tobacco industry and its consultant scientists. Combining evidence from multiple studies, whether experimental or observational, has proved to be an efficacious approach for synthesis. This combination can be accomplished by meta-analysis, combining summary estimates from individual studies, and pooled analysis, analyzing data jointly from individual participants in multiple studies (*see* **Meta-analysis in Epidemiology**). While the use of meta-analysis has been questioned [3], properly conducted meta-analyses have yielded useful and sometimes unexpected findings [5]. Pooled analysis is a more powerful approach, offering the possibility of controlling, confounding and exploring effect modification at the individual level, but requiring the effort of creating the pooled data set for analysis. The array of alternative approaches for synthesis, ranging from expert opinion to quantitative summary,

has not been rigorously evaluated, but more recent approaches, involving a systematic evaluation and quantitative summary of data, seem preferable.

Summary

The twentieth century has seen a remarkable evolution of epidemiology and also of biostatistics, the companion quantitative science of public health and medicine. Epidemiology has moved from being a problem-solving approach used in the field to a core scientific method of biomedical research. As scientific questions around the public's health have become more complex, the field of epidemiology has itself become more complex with speciation into subareas defined by exposures, outcomes, and the methods and domains of inquiry. Evidence forthcoming from epidemiologic research is given weight in policy development for health care and public health, attesting to the immediacy and relevance of epidemiologic data.

References

[1] Armenian, H. & Lilienfeld, D.E. (1994). Overview and historical perspective, *Epidemiologic Reviews* **16**, 1–5.

[2] Austin, H., Hill, H.A., Flanders, W.D. & Greenberg, R.S. (1994). Limitations in the application of case-control methodology, *Epidemiologic Reviews* **16**, 65–76.

[3] Bailar, J.C., III. (1997). The promise and problems of meta-analysis, *New England Journal of Medicine* **337**, 559–561.

[4] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data, *Biometrics* **2**, 47–53.

[5] Berlin, J.A. & Colditz, G.A. (1999). The role of meta-analysis in the regulatory process for foods, drugs, and devices, *Journal of the American Medical Association* **281**, 841–844.

[6] Bernstein, P.L. (1996). *Against the Gods. The Remarkable Story of Risk*. Wiley, New York.

[7] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. International Agency for Research on Cancer, Lyon.

[8] Checkoway, H., Pearce, N.E. & Crawford, D.J. (1989). *Research Methods in Occupational Epidemiology*. Oxford University Press, New York.

[9] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix, *Journal of the National Cancer Institute* 1269–1275.

[10] Cox, C.S., Rothwell, S.T. & Madans, J.H. et al. (1992). Plan and operation of the NHANES I Epidemiologic Follow-up Study, 1987. National Center for Health Statistics, *Vital and Health Statistics* **27**, 1–190.

[11] Dawber, T.R. (1980). *The Framingham Study. The Epidemiology of Atherosclerotic Disease*. Harvard University Press, Cambridge, Mass.

[12] Doll, R. (1992). Sir Austin Bradford Hill and the progress of medical science, *British Medical Journal* **305**, 1521–1526.

[13] Doll, R. & Hill, A.B. (1954). The mortality of doctors in relation to their smoking habits. A preliminary report, *British Medical Journal* **1**, 1451–1455.

[14] Evans, A.S. (1993). *Causation and Disease: A Chronological Journey*. Plenum Medical Books, New York.

[15] Eyler, J.M. (1978). e conceptual origins of William Farr's epidemiology: numerical methods and social thought in the 1830s, in *Times, Places, and Persons. Aspects of the History of Epidemiology*, A.M. Lilienfeld, ed. Johns Hopkins University Press, Baltimore, pp. 1–21.

[16] Fee, E. (1987). *Disease and Discovery. A History of the Johns Hopkins School of Hygiene and Public Health*. Johns Hopkins University Press, Baltimore.

[17] Feinstein, A.R. (1973). Clinical biostatistics. XX. The epidemiologic trohoc, the ablative risk ratio, and "retrospective" research, *Clinical Pharmacology and Therapeutics* **14**, 291–307.

[18] Fletcher, R.H. (1996). *Clinical Epidemiology: The Essentials*. Williams & Wilkins, Baltimore.

[19] Gail, M.H. (1996). Statistics in action, *Journal of the American Statistical Association* **91**, 1–13.

[20] Gordis, L. (1996). *Epidemiology*. W.B. Saunders, Philadelphia.

[21] Gordon, T., Moore, F.E., Shurtleff, D. & Dawber, T.R. (1959). Some methodologic problems in the long-term study of cardiovascular disease: observations on the Framingham study, *Journal of Chronic Diseases* **10**, 186–206.

[22] Gornick, M.E., Eggers, P.W., Reilly, T.W., Mentnech, R.M., Fitterman, L.K., Kucken, L.E. & Vladeck, B.C. (1996). Effects of race and income on mortality and use of services among Medicare beneficiaries, *New England Journal of Medicine* **335**, 791.

[23] Hoover, R.N. & Strasser, P.H. (1980). Artificial sweeteners and human bladder cancer preliminary results, *Lancet* **1**, 837–840.

[24] Jackson, L.W., Lee, N.L. & Samet, J.M. (1999). Frequency of policy recommendations in epidemiologic publications, *American Journal of Public Health* **89**, 1206–1211.

[25] Last, J.M. (1995). *A Dictionary of Epidemiology*. Oxford University Press, New York.

[26] Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977). Methods of cohort analysis: appraisal by application to asbestos mining, *Journal of the Royal Statistical Society* **140**, 469–491.

[27] Lilienfeld, A.M. (1982). *Ceteris paribus*: the evolution of the clinical trial, *Bulletin of the History of Medicine* **56**, 1–18.

[28] Lilienfeld, A.M. & Lilienfeld, D.E. (1980). *Foundations of Epidemiology*, 2nd Ed. Oxford University Press, New York.

[29] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[30] Maxcy, K.F. (1941). *The Papers of Wade Hampton Frost. A Contribution to the Epidemiological Method*. Commonwealth Fund, New York.

[31] Meinert, C.L. & Tonascia, S. (1986). *Controlled Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

[32] Middleton, J. (1999). The blues and pellagra: a public health detective story, *British Medical Journal* **319**, 1209.

[33] Miettinen, O.S. (1970). Matching and design efficiency in retrospective studies, *American Journal of Epidemiology* **91**, 111–118.

[34] Miettinen, O.S. (1985). The "case-control" study: valid selection of subjects, *Journal of Chronic Diseases* **38**, 543–548.

[35] Mustard, C.A., Kaufert, P., Kozyrskyj, A. & Mayer, T. (1998). Sex differences in the use of health care services, *New England Journal of Medicine* **338**, 1678–1683.

[36] Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford University Press, Oxford.

[37] Pearce, N. (1996). Traditional epidemiology, modern epidemiology, and public health, *American Journal of Public Health* **86**, 678–683.

[38] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[39] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*. Lippincott-Raven, Philadelphia.

[40] Sackett, D.L., Richardson, W.S., Rosenberg, W. & Haynes, R.B. (1997). *Evidence-Based Medicine. How to Practice and Teach EBM*. Churchill Livingstone, New York.

[41] Samet, J.M. & Muñoz, A. (1998). Evolution of the cohort study, *Epidemiologic Reviews* **20**, 1–14.

[42] Schlesselman, J.J. (1982). *Case Control Studies*. Oxford University Press, New York.

[43] Schull, W.J. (1997). Brain damage among individuals exposed prenatally to ionizing radiation: a 1993 review, *Stem Cells* **15**, 129–133.

[44] Shephard, D.A.E. (1995). *John Snow. Anesthetist to a Queen and Epidemiologist to a Nation. A Biography*. York Point Publishing, Cornwall, Prince Edward Island, Canada.

[45] Shy, C.M. (1997). The failure of academic epidemiology: witness for the prosecution, *American Journal of Epidemiology* **145**, 479–484.

[46] Stolley, P.L. & Lasky, T. (1995). *Investigating Disease Patterns: The Science of Epidemiology*. W.H. Freeman, San Francisco.

[47] Thiemann, D.R., Coresh, J., Oetgen, W.J. & Powe, N.R. (1999). The association between hospital volume and survival after acute myocardial infarction in elderly patients, *New England Journal of Medicine* **340**, 1640–1648.

[48] Thomas, D. (1998). New techniques for the analysis of cohort studies, *Epidemiologic Reviews* **20**, 122–134.

[49] US Department of Health and Human Services (USDHHS) (1989). *Reducing the Health Consequences of Smoking. 25 Years of Progress. A Report of the Surgeon General.* US Government Printing Office, Washington.

[50] US Department of Health Education and Welfare (DHEW) (1964). *Smoking and Health. Report of the Advisory Committee to the Surgeon General*, DHEW Publication No. (PHS) 1103. US Government Printing Office, Washington.

[51] US Environmental Protection Agency (EPA) (1996). *Proposed Guidelines for Carcinogen Risk Assessment*, EPA/600/P-92/003C. Office of Research and Development, Washington.

[52] White, C. (1990). Research on smoking and lung cancer: a landmark in the history of chronic disease epidemiology, *Yale Journal of Biology and Medicine* **63**, 29–46.

JONATHAN M. SAMET & LEON GORDIS

Epilepsy

Approximately 22 of every 1000 people will suffer a seizure (convulsion or fit) sometime during their lives and, of these, 17 will have a recurrence within a comparatively short period (about one year) and be diagnosed with epilepsy; about 10 per 1000 are currently prescribed anti-epileptic drugs (AEDs), and about 5 per 1000 have "active" epilepsy (a seizure within the previous two years). The last is equivalent to a point prevalence for epilepsy of 500 per 100 000 population. With an annual incidence of 40 per 100 000 there are about 65 new cases each day in a population of 60 million.

The epilepsies are one of the oldest documented groups of diseases; they were described and classified in a neo-Babylonian stone "textbook" of medicine (700–600 BC), and effective treatment was claimed by Hippocrates (460–377 BC) who advocated "use of drastic measures including drugs"; various forms of epilepsy are featured in the Bible. Despite this long history, "epilepsy" is poorly understood, a cause being clearly identified in only 50% of cases. This is changing following the introduction of noninvasive, high-definition, computerized scanning techniques, such as magnetic resonance imaging, which can localize lesions deep within the brain. Epilepsy affects all age groups, being particularly associated with birth trauma and infection in the very young, head injury in teenagers and young adults, alcohol in middle age, and vascular disease, cerebral tumors, and pneumonia in the elderly; seizures may be induced by lack of sleep, stress, the menses (catamenial epilepsy), and in some people who are photosensitive, by watching television and playing video games. In some patients seizures may only occur during specific periods (whilst asleep or on wakening), in others randomly.

Towards the end of the nineteenth century Gowers (1845–1915), the first neurologist to study epilepsy scientifically and who initiated the modern era of epilepsy epidemiology, listed about 50 extracts from plants and other chemical compounds (some of them poisons) which had been used during the previous 1400 years to treat epilepsy. Earlier, in 1860, Sieveking had wryly commented that "there is scarcely a substance in the world capable of passing through the gullet of man that has not at one time or another enjoyed the reputation of being antiepileptic". The modern era of pharmacological treatment started with the introduction of bromides (which are quite toxic) by Locock in the 1860s, and culminated by 1976 in the licensing of 37 major drugs under 480 proprietary names worldwide; (see [14]). Today most patients are maintained on one of four AEDs: phenobarbitone (discovered in 1912), phenytoin (1938), carbamazepine (1974), and sodium valproate (1978). During the 10 years following the introduction of sodium valproate few new AEDs were discovered, a situation which has reversed since the late 1980s with 15–20 new agents now under intense investigation, and several already licensed.

Seizures

Most epileptic seizures are spontaneous (unprovoked), short-lived (rarely exceeding 15 minutes in duration), and self-limiting (thus terminate without intervention); however, they are recurrent in some patients and may re-occur over many years either intermittently or in bursts (clusters). An epileptic seizure is thus a transient disturbance of brain function resulting from repeated simultaneous firing (paroysmal discharge) of nerve cells sometimes limited within a specific region of the brain (for example, the temporal lobe or motor cortex), sometimes spreading from a specific region (focus) to the whole brain, and sometimes engulfing the whole brain from onset. This gives rise to a natural subdivision into partial seizures (sometimes secondarily generalized) and primary generalized seizures; partial seizures may or may not be associated with loss of consciousness, leading to further subdivision. In practice there are several different types of both partial and generalized seizures and these have been delineated in a classification system devised by the International League Against Epilepsy (ILAE). Besides a classification of seizure types the ILAE has also devised a classification of the epilepsies (the pathology that underlines the seizures themselves); the former is simpler and ubiquitous (almost mandatory) in studies of epilepsy, the latter is much more complex and, consequently, of less practical use. Both systems require an electroencephalogram (EEG) for precise application. (The EEG is a multi-electrode device which, from the surface of the scalp, records electrical activity integrated over a few million nerve cells situated within different regions of the brain and outputs

multiwave (one from each region, often about 16) continuous traces; these are examined for several specific abnormal patterns characterizing high frequency neuronal discharges.) The majority of patients suffer just one type of seizure, the commonest being primary generalized tonic–clonic and complex partial seizures. With the former the subject loses consciousness, falls to the ground, stiffens (tonic phase), and then suffers violent jerking (clonic phase) of the whole body, which eventually subsides; the subject then regains consciousness but may be confused and tired, and consequently sleeps. With partial complex seizures the subject engages in complicated circular movements of the arms and/or legs and/or fidgety movements with the fingers. A minority of patients manifest seizures of two, or much more rarely, three types. It is important to stress that not all seizures are associated with epilepsy. For example, children under six years of age may suffer a (febrile) convulsion as a result of a high temperature induced by fever; such seizures are not epileptic. However, the induction of a series of epileptic seizures by electrical stimulation of the brain is a therapeutic maneuver for the treatment of some psychiatric disorders [electroconvulsive therapy (ECT)]. Furthermore, some patients may suffer from an illusion that they have had seizures and provide a credible description of them – a psychiatric disorder categorized as "pseudoseizures".

Treatment and Prognosis

In clinical practice patients who experience a seizure are often first seen by a community physician (general practitioner), who will refer most to a neurologist for further investigation, and may, dependent on circumstances, initiate treatment with an AED. Since referral to a specialist often takes about three months, the neurologist will have a longer clinical history upon which to base a firm diagnosis and consequently better grounds for deciding in consultation with the patient whether or not to initiate treatment. In the UK and elsewhere in Europe, drug treatment will start with a comparatively low dose of one of the three or four first line AEDs, with dose escalation only on inadequate seizure control to the point where the patient experiences toxic and unacceptable side-effects. At this point another AED may either be added or substituted. Treatment is sometimes monitored by assaying serum concentrations of the AED as a guide to whether it is actually being taken (many patients – perhaps 40% – may be "noncompliant") and sometimes as a guide to dose or drug changes. Following **observational studies** serum drug concentrations have been grouped into three broad strata: subtherapeutic (inadequate seizure control); therapeutic (seizure control without unacceptable side-effects), and toxic (unacceptable side-effects of the drug). In the US patients may be treated more aggressively, being started on higher doses of AEDs with the objective of quickly achieving optimal (i.e. therapeutic) serum concentrations. However, many patients do remain seizure-free once low dose AEDs are started. In developing areas of the world epileptic seizures are often not treated with AEDs, first because they are unlikely to be available and secondly because many are far too expensive for patients to afford. (Net costs (2003) NHS prescribing (UK) for minimum recommended dosage (cost per year); phenobarbitone (£10); phenytoin (£20); carbamazepine (£77); sodium valproate (£108); and two newer drugs –lamotrigine (£420) and vigabatrin (£655)). Local customs and medical practices may result in patients being treated by traditional methods.

In any group representative of newly diagnosed patients about 75% will become seizure-free for long periods (in excess of two years) within five years of diagnosis, about 10% will continue to experience an occasional seizure or cluster of seizures, and the remaining 15%, despite (multi-) AED therapy (polytherapy), will develop chronic epilepsy experiencing seizures at least once every six weeks and perhaps up to several hundred per day. Although most people with epilepsy reside in the community, some with chronic epilepsy will have an occasional stay in a specialist institution where (re-)establishment of seizure control may be attempted; the more extreme chronic patients will reside there permanently. Some patients for whom AED treatment has failed will request and be assessed for neurosurgery, where part of the brain containing the epileptogenic focus will be removed. Optimistic claims, sometimes startling, of the successful outcome of "epilepsy surgery" have led to increased demands for these procedures over the past decade in Europe and the US though it should be noted that the efficacy of this form of **surgery** has only recently been established in a randomized **clinical trial** (RCT) [40].

Even today, in developed countries, epilepsy is a socially stigmatizing disease which can lead to unemployment and difficulty in finding work, debarment from certain occupations (such as working within the vicinity of potentially dangerous machinery or driving a public service or heavy goods vehicle), restricted social networks, and genuine fear and anxiety. In primitive communities it is still regarded as "possession by a devil or spirits" and treated by traditional methods ("native medicine" (*see* **Alternative Medicine**)). People in the UK who report the occurrence of a seizure (as required by law) to the Driving Vehicles Licensing Authority will have their driving licenses withdrawn until they have demonstrated complete seizure control (usually freedom from seizures of all types for a period of at least one year); those who reduce their doses of AEDs (perhaps because of side-effects or pregnancy) are advised to stop driving for several months. However, driving regulations vary widely from one country to another, and in the US from state to state; most require a period of two years free of seizures, others impose a lifetime ban.

Early Clinical Trials

There have been hundreds, perhaps thousands of studies assessing the efficacy of AEDs mainly through observing numbers of seizures; indeed one of the best known and most quoted references in epilepsy is a monograph by Coatsworth [7] that may claim to be one of the first published comprehensive overviews of treatment. Not only does it predate the first formal **meta-analysis** in epilepsy by 35 years, but it demonstrates some appreciation of exclusion bias for, as Penry states in the Foreword "In keeping with the general philosophy of thorough documentation for better evaluation, a bibliography of publications judged unworthy of profile has been included in addition to the bibliography of profiled articles". Coatsworth garnered articles (either clinical trials which utilize prospectively some form of **experimental design**, or case reports (*see* **Case Series, Case Reports**) of the results of drug treatment retrospectively without initial design) from 64 different journals and published over the period from 1920 to mid-1970. Of the 110 clinical trials, 43% had fewer than 50 patients, 27% more than 100; almost one-half (47%) did not report duration of the study; crucially, three were "multiple group **crossovers**", presumably explaining the only two that were double-blind and one that was single-blind, as well as the three that were randomized! For of these 110 "trials", 106 were single-group studies, with no information about patient evaluation in 74. In Coatsworth's summary:

> the average reported clinical trial may be characterized as a study of one drug given over a variable period to a group of 20 to 29 outpatients of differing seizure types. No controls are used, and the drug is varied in dosage by the needs of the patient. Seizure counts, types of seizure, and side effects are the data collected by an unreported evaluator using the clinical examination and laboratory data as his observational methods. The patients are evaluated before the trial and irregularly during the trial. The results of treatment are reported by the percentage of patients improved. In those studies with fair to good results, the investigator's opinion is that this drug is a valuable addition to the present regimen of antiepileptics.

This situation is apparent even today, for in the early stages of drug development the single-group "before and after" study is used as a screening mechanism for further investigation and, though important for initial safety testing, the results of such studies are sometimes presented with inappropriate conclusions. Patients who suffer seizures (despite AED treatment) are observed over a baseline period (not necessarily of fixed duration) and those with a seizure frequency above a stipulated threshold are all given the new putative AED for a set "test" period. The changes in seizure frequencies between baseline and test periods are then summarized and any reduction interpreted as a demonstration of efficacy. That such reductions may be explained by **regression to the mean**, even in the absence of any treatment effect, is gradually becoming better understood. Indeed, it has been demonstrated, in the form of regression towards the **median**, by Spilker & Segreti [37] who abstracted data from published epilepsy trials that included at least 10 patients and where the baseline and placebo periods were of equal length.

Crossover Trials

The next stage of development (Phase II and early Phase III) consists of two-treatment, two-period **crossover trials** in patients with "drug-resistant" epilepsy. Here again patients are observed over a baseline period (perhaps 8–12 weeks) and those

taking (specific) AEDs and experiencing seizures of predefined types and in excess of a stipulated threshold are randomized to the two sequences of "add-on" treatment with the new agent or "add-on" placebo; treatment periods frequently last up to 12 weeks with an intervening washout period of four weeks. Occasionally the new agent will be substituted for, rather than added to, existing AEDs, in which case the first four weeks of each treatment period may be used for tapering-off the original AEDs while starting the new. The primary responses are seizure counts over the three periods and side-effects over the two treatment periods. Since the variation in seizure counts between patients may be almost two orders of magnitude greater than variation within patients, these designs are much more efficient than the parallel-group design (*see* **Clinical Trials, Overview**). Unfortunately, much of this gain in efficiency is sacrificed in analysis by massive data reduction. Seizure counts are not **normally distributed**, and rather than attempt any form of data **transformation** to achieve this, or the application of any statistical method (such as **Poisson regression** or a mixture model) to allow for it, triallists resort to reporting and comparing the percentages of patients who achieve at least a (rather arbitrary) 50% reduction in seizure frequency by comparison with the baseline period.

Trials of this type are expected to demonstrate an improvement in seizure control on the new agent over that achieved with placebo; however, since there is no active control it is possible that these trials may miss potentially useful agents since the recruitment of "drug-resistant" patients may result in a sample that does not respond to either treatment; this cannot be gauged without the external validity supplied through the inclusion of an internal "active" control. For this and other reasons use of the crossover trial in epilepsy has declined and in the development of some of the newer AEDs abandoned altogether in favor of parallel group studies.

Parallel Group Trials

Traditionally, once a series of crossover trials had been successfully completed in patients who suffered more serious epilepsy, attention switched to the routine clinical treatment of newly diagnosed patients in the community. Such (Phase III) trials are parallel group studies which incorporate an active control; although perhaps originally designed to demonstrate an important clinical advantage of a new AED over established AEDs, experience suggests that this is in fact extremely difficult to achieve, and in practice such studies are now more frequently designed as **equivalence trials** or noninferiority trials. It is of interest that many of the established AEDs, particularly the older ones, have never been subjected to placebo-controlled trials, and it is now regarded as unethical to treat newly diagnosed patients with placebo. The majority of patients in the community experience few seizures and consequently seizure counts are not a useful outcome measure, often being too sparse for sensible analysis. Attention focuses instead on **survival-type analyses** of events associated with seizure remission (that is complete absence of seizures) as much as seizure recurrence.

The interval from **randomization** to first seizure recurrence is popular because it provides an estimate of the percentage of patients who remain seizure-free. However, it has the disadvantage of concentrating upon events that occur early in the follow-up period and thereby makes no allowance for dose adjustment. Alternatives are interval from randomization to first seizure after an initial window of (say) two months during which dose titration may take place to establish seizure control, interval from randomization to a particular type of seizure [e.g. major (tonic–clonic) seizure], or interval from randomization to the nth seizure ($n > 1$), or the nth day on which a seizure occurs. (Interestingly, Shofer & Temkin [36] investigated the **power** of analyses based on time to the nth seizure by comparison with seizure frequency in a **simulation** study of crossover trials where seizure frequencies were quite high and assumed to follow a **negative binomial distribution**. They found that while statistical tests based on seizure frequency itself exhibited the highest power, tests on time to the twelfth seizure for a sample size of 50 approached the power of tests on seizure frequency with a sample size of 20).

However, since the main purpose of AEDs is to eliminate seizures altogether through the establishment of long-term control, analyses now mainly focus on the interval from randomization to the achievement of a defined period of remission (complete absence of seizures of any type); popular choices are six months, one, two, and five years. In epidemiologic studies, and increasingly in clinical trials where some

measure of long-term outcome is required, the proportion of patients in terminal remission at different stages of follow-up is also reported; for example, the proportion of patients free from seizures during the two years immediately before five-year follow-up.

Modified Designs

Problems with the active equivalence design clinical trial in epilepsy have been succinctly summarized by Leber [21] and Gram [9], and these, combined with the ethical problems of withholding effective treatment (*see* **Ethics of Randomized Trials**) have led to suggestions for alternate designs. Pledger & Kramer [29] discuss two of these, including the active low-dose control where the aim is to establish a treatment difference, rather than equivalence, between the randomized groups. Another option is the design proposed by Amery & Dony [3] where all eligible patients are first treated for a set period with the drug under investigation; those who do not show a beneficial response are not followed further while those who do are randomized either to continue treatment or to have the investigational drug substituted by a placebo.

Quality of Life

People with epilepsy suffer social stigmatization as a consequence both of seizures and of taking AEDs; one result of this has been an awareness of the need to study not just the recurrence of seizures but their severity and their social consequences. Several groups have developed scales for assessing seizure severity and for disease-specific **quality of life**; these are now used frequently in clinical trials [39].

International League Against Epilepsy (ILAE)

The ILAE is a confederation of the world's leading experts in epilepsy and allied specialties that has been responsible for setting standards through Guidelines produced by Commissions set up to advise on specific issues. Apart from the classification systems mentioned above these include, among other Guidelines, those for Clinical Evaluation of Antiepileptic Drugs [12, 13], for Therapeutic Monitoring of Antiepileptic Drugs [16], for Antiepileptic Drug Trials in Children [17], and for Epidemiologic Studies on Epilepsy [15, 18]. ILAE has also produced an International Glossary of Antiepileptic Drugs [14], and commentaries on the economic burden of epilepsy [19], and genetic epilepsies [20].

Journals

There are four journals devoted entirely to epilepsy, although papers in this specialty are published regularly in general medical and general neurology journals. The four are *Epilepsia* (official journal of the ILAE; started 1960), *Epilepsy Research* (1987), *Seizure* (journal of the British Epilepsy Association; 1992), and *Journal of Epilepsy* (1988). The standard of statistical presentation in these journals is comparable with most other medical specialties and rarely ventures farther than the use of routine summary measures and associated significance tests; **actuarial methods** first appeared about 20 years ago. In some ways this is surprising since epilepsy is a heterogeneous, recurrent condition in which seizures may not be accurately recorded, and which would appear to offer an ideal opportunity for the application of a range of sophisticated, contemporary statistical techniques such as **multilevel**, **random effects** models, **frailty** models, and stochastic models (*see* **Stochastic Processes**) including **measurement error**; all with recurrence.

However, there are a few exceptions; for example, Hopkins et al. [10] and Milton et al. [28] examine seizure occurrence patterns using **Poisson processes**; the former, as well as Albert [1] look at **Markov processes**. Racine-Poon & Dubois [30] use a **hierarchical random-effects model** to predict maximum plasma carbamazepine concentrations in individual patients. Thall & Vail [38] and Breslow and Clayton [5] present reanalyses of an earlier trial using **generalized linear mixed models** that account for **overdispersion**, heteroscedasticity (*see* **Scedasticity**), and dependence (*see* **Statistical Dependence and Independence**).

Landmark Studies

With many hundreds of papers on epilepsy published over more than a century it is difficult, if not unfair, to identify any as being of such outstanding merit that methodologically they are vastly superior to all

others. While there have been many comparatively useless investigations, there have also been many very fine studies that demonstrate the best principles of design, conduct, and analysis. Those identified below should therefore be viewed as a sample of the better studies, not the only ones.

The first RCTs in epilepsy were published in 1956 [31, 35] but it was another 30 years before comparative trials in the community recruited in excess of 100 patients per treatment [25], and more than another 10 years before the first meta-analysis [24]. The importance of overviews for emerged since the establishment of the Cochrane Epilepsy Group in Liverpool, UK, in 1996 with its database of over 350 trials involving more than 3500 patients. (*see* **Cochrane Collaboration**). Meta-analyses from this Group have been used to illustrate methods for extracting estimates of hazard ratios from published trials [42], and methods for evaluating putative interactions between epilepsy type and treatment outcome [41], as well as to discuss difficulties in the interpretation of the summary statistic, number needed to treat ([11, 22]).

The largest RCT in epilepsy was the Medical Research Council Anti-epileptic Drug Withdrawal Study [26] which recruited over 1000 patients with a history of epilepsy who had been seizure-free for more than two years, and compared the risks of seizure recurrence in those randomized to withdraw AEDs with those randomized to continue; it also provided a predictive model for the risks of seizure recurrence under the two treatment policies that is now used in counseling patients [27].

While the first randomised trial of epilepsy surgery has been mentioned above [40], another innovative trial also published in 2001, was that of treatment of patients suffering prolonged seizures (*status epilepticus*) and unconscious, with lorazepam, diazepam, or placebo. Single drug packs were allocated randomly to ambulances, and trial entry was authorized by paramedical staff using radio contact to a physician at the base hospital under a waiver of informed consent [23]. The trial demonstrated the safety and effectiveness of the two active drugs compared with placebo [2].

The methodologic problems of epidemiologic follow-up studies in epilepsy have been summarized by Sander & Shorvon [33]; besides describing the difficulties of case **ascertainment**, and the problems of diagnosis, classification, and **selection bias**, they include an appendix of prevalence and incidence studies. Three groups have presented a series of papers based on follow-up of extensive cohorts of patients. The first in Minnesota, USA, is a retrospective study of 618 patients diagnosed with epilepsy at the Mayo Clinic between 1935 and 1974 (later extended to 1984) [4]; they report on the remission of seizures, prevalence, incidence, and mortality, as well as the risk of recurrence after an initial unprovoked seizure, a subject which has caused considerable debate over many years as a result of diagnostic and verification problems. The second [8] identified 122 patients with epilepsy out of a population of 6000 from a single general practice in the UK; they reported prevalence and outcome. The third, also in the UK, is the National General Practice Study of Epilepsy [34] in which 1195 patients considered by their general practitioners to have a possible diagnosis of epilepsy or febrile convulsions were followed up prospectively; the study reported risks of seizure recurrence and remission, and also examined mortality.

Recent information about epilepsy and antiepileptic drugs can be found in Browne & Holmes [6], and Sabers & Gram [32]. In addition, there are several very useful websites:

American Epilepsy Society: `//www.aesnet.org/`
Cochrane Epilepsy Group: `//www.liv.ac.uk/epilepsy/`
Epilepsy Action: `//www.epilepsy.org.uk/`
European Epilepsy Academy: `//www.epilepsy-academy.org/`
International League Against Epilepsy: `//www.ilea-epilepsy.org/`
National Society for Epilepsy: `//www.epilepsynse.org.uk/`

References

[1] Albert, P.S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics* **47**, 1371–1381.

[2] Alldredge, B.K., Gelb, A.M., Isaacs, S.M., Corry, M.D., Allen, F., Ulrich, S., Gottwald, M.D., O'Neil, N., Neuhaus, J.M., Segal, M.R., Lowenstein, D.H. (2001). A comparison of lorazepam, diazepam, and placebo for the treatment of out-of-hospital status epilepticus. *New England Journal of Medicine* **345**, 631–637.

[3] Amery, W. & Dony, J. (1975). A clinical trial design avoiding undue placebo treatment, *Journal of Clinical Pharmacology* **15**, 674–679.

[4] Annegers, J.F., Hauser, W.A. & Elveback, L.R. (1979). Remission of seizures and relapse in patients with epilepsy, *Epilepsia* **20**, 729–737.

[5] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

[6] Browne, T.R. & Holmes, G.L. (2001). Epilepsy. *New England Journal of Medicine* **344**, 1145–1151.

[7] Coatsworth, J.J. (1971). *Studies on the Clinical Efficacy of Marketed Antiepileptic Drugs*. National Institute of Neurological Diseases and Stroke (NINDS) Monograph No 12, DHEW Publication No.(NIH) 73-51.

[8] Goodridge, D.M.G. & Shorvon, S.D. (1983). Epileptic seizures in a population of 6000, *British Medical Journal* **287**, 641–647.

[9] Gram, L. (1997). Antiepileptic drug monotherapy designs in clinical trials. *Epilepsia* **38**,(Suppl. 5), S14–S16.

[10] Hopkins, A., Davies, P. & Dobson, C. (1985). Mathematical models of patterns of seizures: their use in the evaluation of drugs, *Archives of Neurology* **42**, 463–467.

[11] Hutton, J.L. (2000). Number needed to treat: properties and problems. *Journal of the Royal Statistical Society, Series A*, **163**, 403–419.

[12] International League Against Epilepsy (1973). First Commission on Antiepileptic Drugs: Principles for clinical testing of antiepileptic drugs. *Epilepsia* **14**, 451–458.

[13] International League Against Epilepsy (1989). Guidelines for Clinical Evaluation of Antiepileptic Drugs, *Epilepsia* **30**, 400–408.

[14] International League Against Epilepsy (1992). International Glossary of Antiepileptic Drugs. *Epilepsia* **33**,(Suppl. 2).

[15] International League Against Epilepsy (1993). Guidelines for Epidemiologic Studies on Epilepsy, *Epilepsia* **34**, 592–596.

[16] International League Against Epilepsy (1993). Guidelines for Therapeutic Monitoring of Antiepileptic Drugs. *Epilepsia* **34**, 585–587.

[17] International League Against Epilepsy (1994). Guidelines for Antiepileptic Drug Trials in Children, *Epilepsia* **35**, 94–100.

[18] International League Against Epilepsy (1997). ILAE Commission Report. The epidemiology of the epilepsies: future directions. *Epilepsia* **38**, 614–618.

[19] International League Against Epilepsy (2002). ILAE Commission on the burden of epilepsy, Subcommission on the economic burden of epilepsy: final report 1998–2001. *Epilepsia* **43**, 668–673.

[20] International League Against Epilepsy (2002). ILAE Genetics Commission Conference Report: molecular analysis of complex genetic epilepsies. *Epilepsia* **43**, 1262–1267.

[21] Leber, P.D. (1989). Hazards of inference: the active control investigation, *Epilepsia* **30**(Supplement 1), S57–S63.

[22] Lesaffre, E. & Pledger, G. (1999). A note on the number needed to treat. *Controlled Clinical Trials* **20**, 439–447.

[23] Lowenstein, D.H., Alldredge, B.K., Allen, F., Neuhaus, J., Corry, M., Gottwald, M., O'Neil, N., Ulrich, S., Isaacs, S.M., Gelb, A. (2001). The prehospital treatment of status epilepticus (PHTSE) study: design and methodology. *Controlled Clinical Trials* **22**, 290–309.

[24] Marson, A.G., Kadir, Z.A. & Chadwick, D.W. (1996). New antiepileptic drugs: a systematic review of their efficacy and tolerability, *British Medical Journal* **313**, 1169–1174.

[25] Mattson, R.H., Cramer, J.A., Collins, J.F., Smith, P.B., Delgado-Escueta, A.V., Browne, T.R., Williamson, P.D., Treiman, D.M., McNamara, J., McCutchen, C.B. et al. (1985). Comparison of carbamazepine, phenobarital, phenytoin, and primidone in partial and secondarily generalized tonic-clonic seizures, *New England Journal of Medicine* **313**, 145–151.

[26] Medical Research Council Antiepileptic Drug Withdrawal Study Group (1991). Randomized study of antiepileptic drug withdrawal in patients in remission, *Lancet* **337**, 1175–1180.

[27] Medical Research Council Antiepileptic Drug Withdrawal Study Group (1993). Prognostic index for recurrence of seizures after remission of epilepsy. *British Medical Journal* **306**, 1374–1378.

[28] Milton, J.G., Gotman, J., Remillard, G.M., Andermann, F. (1987). Timing of seizure recurrence in adult epileptic patients: a statistical analysis. *Epilepsia* **28**, 471–478.

[29] Pledger, G.W. & Kramer, L.D. (1991). Clinical trials of investigational antiepileptic drugs: monotherapy designs, *Epilepsia* **32**, 716–721.

[30] Racine-Poon, A. & Dubois, J.P. (1989). Predicting the range of plasma carbamazepine concentrations in patients with epilepsy, *Statistics in Medicine* **8**, 1327–1337.

[31] Rettig, J.H. (1956). Chlorpromazine and mysoline in the control of convulsive epilepsy in mentally deficient patients. *Journal of Nervous and Mental Disorders* **124**, 607–611.

[32] Sabers, A. & Gram, L. (2000). Newer anticonvulsants: comparative review of drug interactions and adverse effects. *Drugs* **60**, 23–33.

[33] Sander, J.W.A.S. & Shorvon, S.D. (1987). Incidence and prevalence studies in epilepsy and their methodological problems: a review, *Journal of Neurology, Neurosurgery and Psychiatry* **50**, 829–839.

[34] Sander, J.W.A.S., Hart, Y.M., Johnson, A.L. & Shorvon, S.D. (1990). National General Practice Study of Epilepsy, *Lancet* **336**, 1267–1274.

[35] Schwade, E.D., Richards, R.K., Everett, G.M. (1956). Peganone, a new anticonvulsant drug. *Diseases of the Nervous System* **17**, 155–158.

[36] Shofer, J.B. & Temkin, N.R. (1986). Comparison of alternative outcome measures for antiepileptic drug trials, *Archives of Neurology* **43**, 877–881.

[37] Spilker, B. & Segreti, A. (1984). Validation of the phenomenon of regression of seizure frequency in epilepsy, *Epilepsia* **25**, 443–449.

[38] Thall, P.F. & Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**, 657–671.

[39] Trimble, M.R. & Dodson, W.E., eds (1994). *Epilepsy and Quality of Life*. Raven Press, New York.

[40] Wiebe, S., Blume, W.T., Girvin, J.P., Eliasziw, M., for the Effectiveness and Efficiency of Surgery for Temporal Lobe Epilepsy Study Group (2001). A randomised, controlled trial of surgery for temporal-lobe epilepsy. *New England Journal of Medicine* **345**, 311–318.

[41] Williamson, P.R., Clough, H.E., Hutton, J.L., Marson, A.G., Chadwick, D.W. (2002). Statistical issues in the assessment of the evidence for an interaction between factors in epilepsy trials. *Statistics in Medicine* **21**, 2613–2622.

[42] Williamson, P.R., Tudur Smith, C., Hutton, J.L., Marson, A.G. (2002). Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine* **21**, 3337–3351.

ANTHONY L. JOHNSON

Equivalence Trials

Frequently, a **clinical trial** is designed to evaluate whether an experimental treatment (e.g. a therapy, preventive agent, or medical device) is sufficiently similar to an accepted or "standard" treatment, by some measure of treatment effect, to justify its use. The experimental treatment is often expected to be equal in effect to the standard, but not superior to it. Hence, such a study is commonly called an *equivalence* trial [5, 7]. The term *similarity* trial [5] has also been suggested, since we cannot actually show treatments to be equivalent but may be able to show that they are sufficiently similar by an appropriate criterion. Furthermore, investigators might not expect the treatments to have exactly equal effects, but they might nevertheless wish to demonstrate that the difference in effects is acceptable, considering the benefits of the experimental treatment, such as fewer side effects, greater convenience of use, or lower cost. In this article, *equivalence* and *similarity* are used interchangeably. Other terms that have been suggested include *active control* [19], *positive control* [19, 24], *δ-equivalence* [22], and *δ-no-worse-than* [22], where δ denotes the difference in the outcome measure between two treatments.

The Question Under Study

Formally, we design the trial to show that the experimental and standard treatments are similar with respect to a suitable parameter Θ, which might represent, for example, a difference or ratio of proportions, a difference or ratio of means, or a ratio of event rates of hazards. Assume Θ is defined so that positive values (or values greater than one, in the case of a ratio) indicate superiority of the standard treatment. Letting subscripts E and S refer to experimental and standard, respectively, Θ might be the difference $p_E - p_S$ in probabilities of disease, or the ratio μ_S/μ_E of geometric means of protective serum antibody levels induced by a vaccine. Because of the variability inherent in biologic experimentation, we cannot demonstrate exact equivalence statistically, even if the treatments are identical. We can, however, estimate Θ as precisely as desired from a large enough study.

Generally, in a study of therapeutic or preventive agents where the **outcome** is a clinical measure of treatment effect such as death or onset of disease (or a **surrogate** thought to be related to the clinical outcome), an equivalence or similarity trial will be designed to show that the standard treatment is not superior by a prespecified quantity Θ_0 or more. Thus, the question of primary interest in such a trial is usually one-sided (*see* **Alternative Hypothesis**); we place no restriction on the possible degree of superiority of the experimental treatment. (Note that the standard can be superior to the experimental treatment by an amount less than Θ_0 and the experimental treatment still be considered acceptable. If this is not the case, the trial is not a similarity trial, and it should be designed to show superiority of the experimental treatment.) For example, if Θ represents a difference $p_E - p_S$ in probabilities of disease and it is considered sufficient to demonstrate that the probability of disease using experimental treatment is not more than 0.15 greater than the probability using standard treatment, then $\Theta_0 = 0.15$.

Choice of a meaningful value for Θ_0 is crucial, since it defines levels of similarity sufficient to justify use of the experimental treatment. Θ_0 must be considered reasonable by clinicians (*see* **Clinical Significance Versus Statistical Significance**) and must be less than the corresponding value for placebo compared to standard treatment, if that is known [24]. Thus, the sample size required (*see* **Sample Size Determination for Clinical Trials**) may be much larger than that for a trial comparing the experimental treatment to placebo [15]. The choice of Θ_0 depends on the seriousness of the primary clinical outcome, as well as the relative advantages of the treatments in considerations extraneous to the primary outcome [7]. Careful choice of design parameters is not, however, unique to a similarity trial; in a trial designed to show superiority, it is important to select a meaningful and realistic value for the minimum effect to be detected [20].

Confidence Intervals and Hypothesis Tests

There are advantages to considering design and analysis of an equivalence trial in terms of **confidence intervals** rather than **hypothesis testing** [7, 15, 18, 25]. The confidence interval approach makes the desired outcome of the trial clear and avoids the mistake of choosing an inappropriate hypothesis (see below). Figure 1 depicts a $100\ (1 - 2\alpha)\%$ confidence interval for Θ around a point estimate $\widehat{\Theta}$. We assume

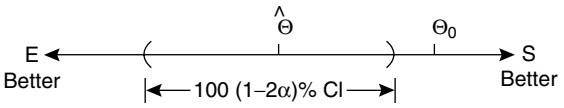

Figure 1 $100(1-2\alpha)\%$ confidence interval for Θ around a point estimate $\widehat{\Theta}$. An upper limit less than Θ_0 allows a conclusion of similarity, or "equivalence". E: experimental treatment; S: standard treatment

equal probability α in each tail. For a specified value of α (e.g. 0.05, 0.025, 0.01), we consider the treatments similar if the upper confidence limit is less than the prespecified quantity Θ_0. Thus, for example, the conventional α of 0.05 corresponds to a two-sided 90% confidence interval. Only a one-sided 100 $(1-\alpha)\%$ interval is really needed. It is appropriate, however, to report a two-sided interval with confidence coefficient $1-2\alpha$, since there is usually interest in the lower limit also. Note, in particular, that a lower limit >0, which indicates superiority of the standard, is consistent with similarity or "equivalence" according to our definition as long as the upper limit is less than Θ_0.

In hypothesis-testing terms, the appropriate null hypothesis is that the standard is superior to the experimental treatment by at least Θ_0; i.e. we test the null hypothesis $H_0 : \Theta \geq \Theta_0$. Rejection of H_0 at significance level α in favor of the one-sided alternative that $\Theta < \Theta_0$ is then sufficient to show similarity. Though such a procedure is correct and is equivalent to the corresponding confidence interval procedure, it encourages consideration of the situation as one of decision making, whereas it is usually one of estimation. A test tells us the strength of the evidence against a specific hypothesis, whereas a confidence interval tells us what values of Θ are consistent with the data.

Design of a similarity or equivalence trial has often been based on the **null hypothesis** $H_0 : \Theta \leq 0$ (or $H_0 : \Theta \leq 1$, for a ratio), as though the purpose of the trial were to demonstrate superiority of the standard treatment. Failure to reject the null hypothesis would then lead to a conclusion of equivalence. Such an approach is *not* appropriate [1, 3, 6, 7, 13, 18, 19, 25, 26], for several reasons. First, it contorts the logic of hypothesis testing; as has long been recognized, we cannot prove statistical hypotheses [9], so we design a study to *reject*, not accept, the null hypothesis [3], i.e. we measure the strength of the evidence *against*, not for, the null hypothesis [1, 2]. Secondly, if a trial has insufficient sample size, large variability, or some other defect due to poor design or conduct, it may fail to reject the null hypothesis of no difference, even if an important difference exists [3, 6, 7, 15, 18, 19, 24–26]. Thus, an inadequately designed or conducted trial may give the desired result, whereas a well designed and well conducted one would not. Furthermore, in a trial designed with high **power**, we may reject the null hypothesis of no difference and thus fail to conclude equivalence, even if the actual magnitude of the difference is unimportant [1, 7, 25, 26]. Hence, the hypothesis of no difference is an inappropriate basis for designing an equivalence trial, and the ***P* value** against that hypothesis is irrelevant to a conclusion of similarity. The size of trial so obtained may be larger or smaller than that based upon a correct approach (see section on sample size).

One-Sided and Two-Sided Questions

As already noted, in an equivalence trial with a clinical outcome, the question of primary interest is usually one-sided, namely can we show that the experimental treatment is not worse than the standard treatment by as much as Θ_0 [3, 18]? We do not wish to show that $\Theta > \Theta'$ for some Θ'. (Note that demonstrating superiority of the experimental treatment – i.e. showing $\Theta < 0$ for a difference or $\Theta < 1$ for a ratio – does not imply a test in the "other" direction from Θ_0, but corresponds to $\Theta_0 = 0$ or 1; if we wish to show superiority of the experimental treatment, then by definition we do not have a similarity trial.)

If we wish to demonstrate that the effects of two treatments do not differ much in either direction, we are interested in whether $\Theta_{01} < \Theta < \Theta_{02}$ for some Θ_{01} and Θ_{02} with $\Theta_{01} < \Theta_{02}$. To show this, the confidence interval for Θ should lie entirely between Θ_{01} and Θ_{02}. Such a situation is two-sided. An important class of two-sided trial is a **bioequivalence** or **bioavailability** trial [1, 13, 25, 26], in which it is desired to show that the biological activity or availability of two treatments is similar in both directions. Comparison of lots of a vaccine might also involve a two-sided question. Consideration of the corresponding hypothesis tests demonstrates, however, that the two-sided equivalence setting is different from the more familiar two-sided test situation (Figure 2(a)). The latter involves a null hypothesis that Θ takes a

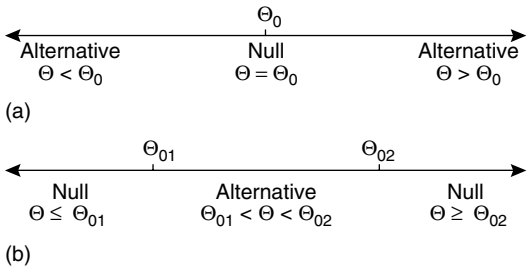

Figure 2 Null and alternative hypotheses for different types of two-sided tests: (a) two-sided alternative; (b) two-sided null

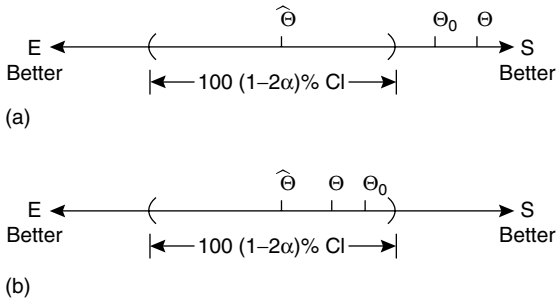

Figure 3 (a) Type I and (b) type II errors in an equivalence or similarity trial. E: experimental treatment; S: standard treatment. Θ denotes the true value, $\widehat{\Theta}$ the point estimate, and Θ_0 the criterion for concluding similarity

single value Θ_0 and an **alternative hypothesis** that Θ differs from Θ_0 in either direction; thus, the *alternative* is "two-sided", since its values lie on both sides of the null value. In the two-sided equivalence situation [Figure 2(b)], however, it is the *null* hypothesis (i.e. $\Theta \leq \Theta_{01}$ or $\Theta \geq \Theta_{02}$) that is two-sided, because it includes values on both sides of the alternative values [1, 23]. This difference has implications for sample size calculation, as indicated below.

Here, we will continue to assume a one-sided situation unless it is specifically stated otherwise.

Type I and Type II Errors

In an equivalence trial, we make a type I error if we falsely conclude similarity; i.e. if we obtain an upper confidence limit $< \Theta_0$ when the true value of Θ is $\geq \Theta_0$ [Figure 3(a)] or, in hypothesis-testing terms, if we reject $H_0 : \Theta \geq \Theta_0$ when H_0 is true. When we base the conclusion on the upper limit of a two-sided $100(1-2\alpha)\%$ confidence interval or on the corresponding one-sided test, the probability of a type I error is $\leq \alpha$.

A type II error is a failure to conclude similarity when the treatments are similar; i.e. obtaining an upper confidence limit $\geq \Theta_0$ when the true value of Θ is $< \Theta_0$ [Figure 3(b)] or failing to reject H_0 when H_0 is false. We denote the probability of a type II error for some specific value of $\Theta < \Theta_0$ by β; $1-\beta$ is then the power of the test or confidence interval procedure for that value of Θ.

It is, of course, desirable to keep both α and β small. However, as is usually the case in clinical trials, it is generally more important to keep α small.

Sample Size

Sample size formulations are available for various types of comparative measures in one-sided similarity trials – e.g. a difference in **normally distributed** means, a difference or ratio of proportions, and a ratio of event rates or hazards. In the descriptions of specific formulations below, Θ refers to the value of the difference or ratio for which we calculate power and Θ_0, as before, denotes the value we wish to rule out. We assume $\Theta < \Theta_0$. The sample size will also depend on the type I and type II error probabilities α and β and, in comparisons of proportions, the assumed true values of the proportions.

For testing the hypothesis of no difference between means from two normal distributions with equal and known variances, we have

$$N = \sigma^2 \left[\frac{1}{k} + \frac{1}{1-k} \right] \frac{(z_\alpha + z_\beta)^2}{\delta^2},$$

where N is the total number of individuals in two groups, σ^2 is the common variance, k and $(1-k)$ are the proportions of the total sample in the two groups, z_α and z_β are upper quantiles of the normal distribution corresponding to α and β, and δ is the minimum difference to be detected [16]. The only change necessary for a similarity trial is to replace δ with $\Theta_0 - \Theta$. Thus, for $\Theta = 0$, we obtain the same sample size for a similarity trial as for the conventional hypothesis of no difference with Θ_0 as the minimum difference to be detected.

If the comparison is based on a ratio of geometric means, as in comparing serum antibody levels, we can take logarithms of the original observations and

proceed as above, if the logarithms are approximately normally distributed.

Various methods have been suggested for analyzing differences [3, 6, 12, 18] and ratios [4, 11] of proportions in an equivalence trial. A **likelihood** score approach has been suggested as the best method overall in these settings [4, 8, 11, 21]. Sample size formulas based on score statistics have been derived for analysis of both differences [8, 21] and ratios [8]. If the null difference is 0 and the null ratio is 1, the two likelihood score sample size formulations of Farrington & Manning [8] are equivalent, and both reduce to a familiar formula for comparing proportions [16].

Table 1 shows sample sizes for some trials in which the outcome is an adverse event, such as the development of disease or a reaction to a treatment or vaccination. Total sample sizes for two equal-sized groups, calculated by the likelihood score method, are given for comparisons based on a difference in proportions ($p_E - p_S$) and on a ratio of proportions (p_E/p_S) [8]. In these examples, the ratio comparison requires larger sample sizes than the difference comparison. Nevertheless, for small proportions, the ratio may be the preferred measure, since it may be more stable than the difference over a variety of settings. Table 1 also shows the sample size obtained from the inappropriate approach of testing the hypothesis of no difference; it can be either larger or smaller than the sample size obtained from the correct calculation. In addition, if the probability of disease is assumed to be higher with the experimental treatment than with the standard treatment, then the sample size required can be much greater than if the probabilities are assumed equal.

Table 1 Total sample size in two equal-sized groups for comparing proportions: $\alpha = 0.05$, $\beta = 0.10$

			Total sample size[a]		
p_E	p_S	p_{E0}^b	Difference		Ratio
0.01	0.01	0.015	14 100	(17 000)[c]	21 100
0.10	0.10	0.200	332	(434)	684
0.40	0.40	0.600	202	(212)	324
0.50	0.40	0.600	826	(212)	1 330
0.60	0.60	0.700	820	(776)	972

[a]Calculated by method of likelihood scores [8].

[b]p_{E0} refers to the value of p_E under the null hypothesis that $p_E - p_S \geq \Theta_0$ (or $p_E/p_S \geq \Theta_0$), where $\Theta_0 = p_{E0} - p_S$ (or p_{E0}/p_S).

[c]Sample size for testing hypothesis of no difference [16], where now p_{E0} is assumed to be the true value of p_E.

If the outcome of interest is an event rate or time to an event (*see* **Survival Analysis, Overview**), one may wish to show similarity via the ratio of event rates or ratio of hazards. Fleming [10] has given a sample size formula for the familiar two-sided situation of a single null value to be ruled out (*not* for a two-sided equivalence trial; see below). The sample size for a one-sided similarity trial can be obtained from Fleming's formula by substituting α for $\alpha/2$. A **Poisson distribution** assumption can be employed for analyzing the ratio of sufficiently small event rates [4].

For a two-sided equivalence trial, where we want to show that $\Theta_{01} < \Theta < \Theta_{02}$, a 100 $(1 - 2\alpha)$% confidence interval with equal tail probabilities provides a test of the hypothesis H: $\Theta \leq \Theta_{01}$ or $\Theta \geq \Theta_{02}$ with significance level $\leq \alpha$ (not 2α) [13, 23]. The power, however, will be lower than for either of the corresponding one-sided trials with the same numbers of individuals. In the symmetric situation where $\Theta_{01} = -\Theta_{02}$ and the assumed value of Θ is 0 (or, for ratios, $\Theta_{01} = 1/\Theta_{02}$ and the assumed value is 1), the sample size for type I error probability α and power $1 - \beta$ in two equal-sized groups is obtained from the formula for the corresponding one-sided problem, with the same z_α but with $z_{\beta/2}$ substituted for z_β [8]. This is a consequence of the null hypothesis, not the alternative, being two-sided; in the more familiar two-sided situation of a single-valued null hypothesis, we would substitute $\alpha/2$ for α, but keep β the same. Sample size formulations have also been given for bioequivalence studies designed as two-period **crossover** trials [1, 26].

Interim Analysis

In monitoring accumulating data from an equivalence trial (*see* **Data and Safety Monitoring**) for possible early stopping, familiar methods can be applied – for example, a confidence interval procedure, corresponding to a group sequential test, requiring an upper limit $< \Theta_0$ before considering termination [14]. Durrleman & Simon [7] have suggested a related procedure. Alternatively, if a trend develops in favor of the experimental treatment, it might be desirable to continue the trial unless the stronger result is obtained that the experimental treatment

is clearly superior to the standard; that is, unless the upper limit of the interim confidence interval is <0.

Special Considerations

Several authors have pointed out problems associated with equivalence trials and requirements for a conclusion of effectiveness of an experimental treatment on the basis of similarity to a standard treatment [15, 19, 24]. Assuming there is no placebo control in the trial, for ethical (*see* **Ethics of Randomized Trials**) or other reasons, the conclusion must rely in part on historical data. From earlier trials, it must be clear that the standard is superior to placebo. It must also be clear, from the equivalence trial itself, not only that the standard and experimental treatments are similar to each other, but also that both are more effective than placebo would have been in that trial. Thus, the equivalence trial should include similar patients and employ similar procedures to those in the trials that previously showed the standard to be effective. The trial should also be large enough to support a conclusion that both treatments are superior to placebo.

Failure to conduct an equivalence trial rigorously according to protocol may obscure important differences between the treatments and contribute to a conclusion of similarity [15, 19, 24]. Thus, the incentive to adhere strictly to the protocol (*see* **Clinical Trials Protocols**), to obtain the desired result, is not present. Appropriate monitoring of protocol adherence is, therefore, especially important in this type of trial.

Most statisticians agree that an **intention-to-treat analysis** is essential when reporting the results of a randomized clinical trial. Such an analysis includes all randomized individuals, in the groups to which they were originally assigned, regardless of the treatment actually received and their level of **compliance** with the treatment regimen. In a trial designed to show superiority, this approach provides a valid test of the null hypothesis of interest, that the experimental treatment has no benefit. That is, the probability of a type I error does not exceed the nominal value. Such may not be the case, however, in an equivalence or similarity trial. To the extent that individuals receive the wrong treatment or are otherwise noncompliant, an intention-to-treat analysis tends to make the treatments appear more similar in effect than they are. An equivalence trial can thus have a type I error probability larger than the nominal value; a conclusion of similarity may therefore require support from other appropriate analyses [17, 19].

Clearly, there are special issues, as noted above, associated with equivalence trials. Nevertheless, such trials – well designed, conducted, and analyzed – are essential in many areas of medicine for properly evaluating new treatments.

References

[1] Anderson, S. & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials, *Communications in Statistics – Theory and Methods* **12**, 2663–2692.

[2] Armitage, P. & Berry, G. (1994). *Statistical Methods in Medical Research*, 3rd Ed. Blackwell Science, Oxford.

[3] Blackwelder, W.C. (1982). "Proving the null hypothesis" in clinical trials, *Controlled Clinical Trials* **3**, 345–353.

[4] Blackwelder, W.C. (1993). Sample size and power for prospective analysis of relative risk, *Statistics in Medicine* **12**, 691–698.

[5] Blackwelder, W. (1995). Similarity/equivalence trials for combination vaccines, *Annals of the New York Academy of Sciences* **754**, 321–328.

[6] Dunnett, C.W. & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables, *Biometrics* **33**, 593–602.

[7] Durrleman, S. & Simon, R. (1990). Planning and monitoring of equivalence studies, *Biometrics* **46**, 329–336.

[8] Farrington, C.P. & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, *Statistics in Medicine* **9**, 1447–1454.

[9] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd. London.

[10] Fleming, T.R. (1990). Evaluation of active control trials in AIDS, *Journal of Acquired Immune Deficiency Syndrome* **3**, Supplement 2, S82–S87.

[11] Gart, J.J. & Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness, *Biometrics* **44**, 323–338.

[12] Gart, J.J. & Nam, J. (1990). Approximate interval estimation of the difference in binomial parameters: correction for skewness and extension to multiple tables, *Biometrics* **46**, 637–643.

[13] Hauck, W.W. & Anderson, S. (1992). Types of bioequivalence and related statistical considerations, *International Journal of Clinical Pharmacology, Therapy and Toxicology* **30**, 181–187.

[14] Jennison, C. & Turnbull, B.W. (1993). Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses, *Biometrics* **49**, 31–43.

[15] Jones, B., Jarvis, P., Lewis, J.A. & Ebbutt, A.F. (1996). Trials to assess equivalence: the importance of rigorous methods, *British Medical Journal* **313**, 36–39.

[16] Lachin, J.M. (1981). Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials* **2**, 93–113.

[17] Lewis, J.A. & Machin, D. (1993). Intention to treat – who should use ITT?, *British Journal of Cancer* **68**, 647–650.

[18] Makuch, R. & Simon, R. (1978). Sample size requirements for evaluating a conservative therapy, *Cancer Treatment Reports* **62**, 1037–1040.

[19] Makuch, R.W., Pledger, G., Hall, D.B., Johnson, M.F., Herson, J. & Hsu, J.-P. (1990). Active control equivalence studies, in *Statistical Issues in Drug Research and Development*, K.E. Peace, ed. Marcel Dekker, New York, pp. 225–262.

[20] Meinert, C.L. (1986). *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

[21] Nam, J. (1995). Sample size determination in stratified trials to establish the equivalence of two treatments, *Statistics in Medicine* **14**, 2037–2049.

[22] Ng, T.-H. (1993). A specification of treatment difference in the design of clinical trials with active controls, *Drug Information Journal* **27**, 705–719.

[23] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.

[24] Temple, R. (1996). Problems in interpreting active control equivalence trials, *Accountability in Research* **4**, 267–275.

[25] Westlake, W.J. (1979). Statistical aspects of comparative bioavailability trials, *Biometrics* **35**, 273–280.

[26] Westlake, W.J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations, in *Biopharmaceutical Statistics for Drug Development*, K.E. Peace, ed. Marcel Dekker, New York, pp. 329–352.

WILLIAM C. BLACKWELDER

Errors in the Measurement of Covariates

Problems involving **covariate** measurement error arise frequently in health research. In retrospective epidemiological studies, for example, it may be difficult or impossible to accurately assess the level of exposure to potential **risk factors** for cancer such as radiation [20], herbicides [25], or dietary fat [22] (*see* **Case–Control Study**). In prospective studies, risk factors of interest may be difficult to observe because of physical location or cost (*see* **Cohort Study**). For example, the degree of narrowing of coronary arteries may reflect risk of heart failure, but physicians may measure the degree of narrowing in carotid arteries instead because of the less invasive nature of this method of assessment. In other settings, the risk factor may be an average value of a quantity over time and any practical way of measuring such a quantity necessarily features measurement error. This was the case in the Framingham Heart Study [15] where one of the risk factors of interest for coronary artery disease was average daily systolic blood pressure. In some settings, interest may lie in assessing associations between **categorical** exposure variables and disease status. Some exposure variables are inherently categorical (e.g. genetic classifications), but categorical covariates also arise when cutpoints are specified for continuous covariates such as systolic blood pressure (SBP) (e.g. SBP <140, $140 \leq$ SBP < 160, $160 \leq$ SBP) (*see* **Categorizing Continuous Variables**). In the latter case, measurement error in the continuous scale results in incorrect category assignments.

When mismeasured covariates are continuous, by convention the problems are said to involve covariate measurement error, whereas mismeasured categorical covariates are said to be misclassified (*see* **Misclassification Error**). In both of these settings, models based on covariates measured with error generally produce biased estimates of parameters characterizing the association between the covariates and the outcome of interest (*see* **Unbiasedness**). A textbook treatment of measurement-error problems is given by Fuller [9] for **multiple linear regression** models and by Carroll et al. [5] for **generalized linear models**.

A comprehensive review of the issues and methodologic advances pertaining to covariate and response measurement error is found in the entry **Measurement Error in Epidemiologic Studies**. Statistical methods for dealing with misclassified categorical variables are reviewed in the entry **Misclassification Error**. In this article, we provide a brief survey of strategies for dealing with mismeasured or misclassified covariates in the context of regression models. An illustrative application is presented using data from a case–control study [3], and the impact of misspecifying the nature of the misclassification distribution is discussed.

Survey of Methods

Let Y be a response variable, $\boldsymbol{Z}$ the vector of covariates free of error, $\boldsymbol{X}$ be the covariates subject to measurement error, and $\boldsymbol{W}$ be the result of attempting to observe $\boldsymbol{X}$ in the presence of measurement error (i.e. $\boldsymbol{W}$ is the mismeasured version of $\boldsymbol{X}$). Suppose that the distribution of Y is dependent on $(\boldsymbol{X}, \boldsymbol{Z})$ through a model denoted by $h(Y, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_x, \boldsymbol{\beta}_z)$, where the dependence of Y on $\boldsymbol{X}$ and $\boldsymbol{Z}$ is reflected by a linear predictor $\boldsymbol{X}'\boldsymbol{\beta}_x + \boldsymbol{Z}'\boldsymbol{\beta}_z$ and additional **nuisance parameters** are suppressed for convenience. Primary interest lies in the estimation of the vector of regression coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_x', \boldsymbol{\beta}_z')'$. Naive use of $\boldsymbol{W}$ in place of $\boldsymbol{X}$ could introduce a considerable bias in the estimation of $\boldsymbol{\beta}_x$, and in many cases $\boldsymbol{\beta}_z$. Mechanisms inducing measurement error can lead to differential or nondifferential measurement error. If the distribution of Y given $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ depends only on $(\boldsymbol{X}, \boldsymbol{Z})$, then the measurement-error process is called *nondifferential*, but otherwise it is called *differential*. Nondifferential measurement error can be equivalently stated such that the **conditional** distribution of $\boldsymbol{W}$ given $(\boldsymbol{X}, \boldsymbol{Z}, Y)$ depends only on $(\boldsymbol{X}, \boldsymbol{Z})$. In this case, it is clear that the term "nondifferential" means that the error distribution is not different for those with different values of Y.

Corrections for measurement error typically rely on there being some data available with which to estimate parameters of the error distribution. This may be, for example, a data set of replicated measurements $\boldsymbol{W}$ for a particular value of $\boldsymbol{X}$ to facilitate estimation of **variance components** of the error distribution in the case of continuous covariates. In other contexts, a validation data set may be available with which to estimate misclassification rates

for categorical covariates. Further remarks on the different types of supplementary data including replication, **validation**, and instrumental data are found in **Measurement Error in Epidemiologic Studies**.

Approximate Methods of Correction

A variety of approaches have been proposed to correct for the bias induced by measurement error (e.g. [1, 30, 31, 41]). Two widely used approaches are *regression* **calibration** and **simulation** *extrapolation* [5]. Except for some special models such as linear and loglinear regression models, these approaches only yield approximately **consistent** estimators; the appeal of these approaches lies in their simplicity and ease of implementation.

If $\mu_{xz} = E(Y|\boldsymbol{X}, \boldsymbol{Z})$, then given a link function $g(.)$, a regression model may be formed by specifying

$$g(\mu_{xz}) = \boldsymbol{X}'\boldsymbol{\beta}_{\boldsymbol{x}} + \boldsymbol{Z}'\boldsymbol{\beta}_{\boldsymbol{z}}. \tag{1}$$

It is not possible to estimate $\boldsymbol{\beta}$ from this model using standard methods since $\boldsymbol{X}$ is not available. Given a validation data set, however, an approximate version of the model may be obtained by replacing $\boldsymbol{X}$ with an estimate of its conditional expectation given $\boldsymbol{W}$ and $\boldsymbol{Z}$, which we denote by $m(\boldsymbol{W}, \boldsymbol{Z}; \boldsymbol{\theta})$. That is, if $\hat{\boldsymbol{\theta}}$ denotes an estimate of $\boldsymbol{\theta}$ and $\hat{m}(\boldsymbol{W}, \boldsymbol{Z}; \hat{\boldsymbol{\theta}})$ denotes the estimate of the conditional expectation, the mean specification

$$g(\mu_{xz}) \approx \hat{m}(\mathbf{W}, \boldsymbol{Z}; \hat{\boldsymbol{\theta}})'\boldsymbol{\beta}_x + \boldsymbol{Z}'\boldsymbol{\beta}_z \tag{2}$$

may be used to estimate $\boldsymbol{\beta}$. If only replication data are available, alternative estimates of $\boldsymbol{X}$ may be used in place of $\boldsymbol{X}$, such as the mean value over all replications. **Instrumental** data provide an alternative approach for inserting an estimate of $\boldsymbol{X}$ based on suitable regression models.

The approach of regression calibration was first suggested by Prentice [21] for the **proportional hazard** models in **survival analysis**. A general form of regression calibration was suggested by Carroll & Stefanski [7] and Gleser [10]. Further developments for generalized linear models were discussed in [2, 23], and [24]. More recently, Wang et al. [38] studied this approach in the context of the **Cox model** with **missing** or mismeasured covariate data, where the missing data are estimated on the basis of a validation data set. The method performs surprisingly well given its simplicity, though the estimates of regression parameters may not be consistent in general.

Simulation extrapolation is an attractive simple approach for reducing bias due to measurement error. Estimates based on simulation extrapolation are obtained by adding additional measurement error to the data in a resampling-like step, establishing a trend of measurement error-induced bias as a function of the variance of the added measurement error, and extrapolating back to the case of no measurement error. This approach, proposed by Cook & Stefanski [8], is well suited to additive or multiplicative measurement errors and leads to improved estimates subject to correct model specification [5]. Refinements have been developed in [4] and [34], where the theoretical justification to this approach and the asymptotic distribution of the simulation extrapolation estimates have been investigated.

Methods Based on Estimating Equations

Unbiased estimating equations (*see* **Generalized Estimating Equations**) provide consistent estimates for parameters of interest and it is natural to consider ways to construct unbiased **estimating functions** for problems involving measurement error. The various forms of estimating functions may be formulated under the assumptions of either a so-called functional error model for which the covariates $\boldsymbol{X}$ are treated as fixed constants, or a structural error model for which $\boldsymbol{X}$ are treated as random variables arising from a particular distribution. Two widely used approaches are based on conditional and corrected score functions.

Suppose that given $(\boldsymbol{X}, \boldsymbol{Z})$, Y arises from a distribution in the **exponential family** and (1) holds. When the surrogate $\boldsymbol{W}$ is assumed to have a **normal distribution** with, for example, $\boldsymbol{W} \sim N(\boldsymbol{X}, \Sigma)$, a generalized linear measurement-error model may be formulated by combining these two distributions. Stefanski & Carroll [32] specifically discussed the conditional score method in this context, where the estimating functions are obtained by conditioning on sufficient statistics for some important models such as linear, **logistic**, loglinear, and the inverse-gamma.

Nakamura [17] proposed the use of corrected score functions, illustrating this method with applications to a number of practical models (e.g. Gaussian and **Poisson**), when the measurement error is additive with a distribution. A corrected score function is a

function of the observed data $(\boldsymbol{W}, \boldsymbol{Z}; Y)$ such that its expectation with respect to the conditional distribution of $\boldsymbol{W}|\boldsymbol{X}$ is equal to the score function based on the distribution of $(\boldsymbol{X}, \boldsymbol{Z}; Y)$. More specifically, if $S(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{Z}, Y)$ is the score function from the true model of Y and $(\boldsymbol{X}, \boldsymbol{Z})$, any function $S^*(\boldsymbol{\beta}; \boldsymbol{W}, \boldsymbol{Z}, Y)$ is called *a corrected score function* if

$$\mathrm{E}_{\boldsymbol{W}|\boldsymbol{X}}(S^*(\boldsymbol{\beta}; \boldsymbol{W}, \boldsymbol{Z}, Y)) = S(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{Z}, Y). \quad (3)$$

Estimation of $\boldsymbol{\beta}$ may proceed on the basis of the equation

$$S^*(\boldsymbol{\beta}; \boldsymbol{W}, \boldsymbol{Z}, Y) = 0. \quad (4)$$

The corrected score function approach is a functional method in the sense of [5] and therefore does not require specification of a distribution for $\boldsymbol{X}$. One drawback of this approach, however, is that the corrected score function does not always exist and it is not generally easy to obtain when it does exist. Novick & Stefanski [18] discussed using **Monte Carlo** simulation to obtain corrected score functions for many problems, and this approach is related to the approach of simulation extrapolation.

Likelihood and Pseudo-likelihood Methods

The methods described thus far are often applied to problems with continuous covariates subject to error where few distributional assumptions are required. **Likelihood** based analyses provide an alternative approach for many quite general problems including those with misclassified covariates. Let $P_{Y|\boldsymbol{X},\boldsymbol{Z}}(y|\boldsymbol{x}, z; \boldsymbol{\beta})$ be the model of interest where the error-prone true covariate $\boldsymbol{X}$ is unobserved for some subjects and $\boldsymbol{\beta}$ is the parameter of primary interest. Then $P_{Y,\boldsymbol{W}|\boldsymbol{Z}}(y, \boldsymbol{w}|z; \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\lambda})$ is the model for the observed data. With the nondifferential measurement error, we may construct the likelihood from the following factorization

$$P_{Y,\boldsymbol{W}|\boldsymbol{Z}}(y, \boldsymbol{w}|z; \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\lambda}) = \int P_{Y|\boldsymbol{X},\boldsymbol{Z}}(y|\boldsymbol{x}, z; \boldsymbol{\beta}) \times P_{\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{w}|\boldsymbol{x}, z; \boldsymbol{\delta}) P_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|z; \boldsymbol{\lambda})\, \mathrm{d}\mu(\boldsymbol{x}), \quad (5)$$

where $\mathrm{d}\mu(\boldsymbol{x})$ indicates that the integrals are sums if $\boldsymbol{X}$ is discrete and integrals if $\boldsymbol{X}$ is continuous, and $P_{\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{w}|\boldsymbol{x}, z; \boldsymbol{\delta})$ and $P_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|z; \boldsymbol{\lambda})$ represent the conditional distributions of $\boldsymbol{W}$ given $\boldsymbol{X}$ and $\boldsymbol{Z}$ and of $\boldsymbol{X}$ given $\boldsymbol{Z}$, respectively.

Likelihood approaches to measurement error have received comparatively less attention in the literature, perhaps due to its computationally demanding nature and a possible lack of **robustness** [29]. However, likelihood methods may be more flexible, efficient, or reliable for dealing with some problems involving measurement error [29]. For example, **likelihood ratio** inferences can be considerably better than those based on **bootstrap** or normal approximations (e.g. [5, 29]). Stefanski & Carroll [33] compared estimates obtained from the conditional score function approach and **maximum likelihood** for logistic regression models and found that the latter is more efficient when the measurement error is "large" or the logistic coefficient is "large". Schafer & Purdy [29] studied likelihood analysis of normal linear regression models with mismeasured covariates using the **EM algorithm**.

When a direct likelihood approach is computationally demanding, one may use an approximate but simpler version to replace the likelihood function and proceed with the so-called **pseudo-likelihood** approach [3]. Hanfelt & Liang [11] proposed an approximate approach for generalized linear models with measurement error in covariates. Other work that explores this approach for measurement error, includes [6, 12, 26, 27, 28, 37, 40].

Semiparametric and Nonparametric Methods

One may classify approaches for measurement error problems as being parametric, **semiparametric**, or **nonparametric** by considering the assumptions regarding the distribution of the measurement error. Unlike the approaches described above, one may not need to assume a parametric family for the error distribution, but rather can make simple conditional moment assumptions. Pepe & Fleming [19] used the empirical estimation of the likelihood to deal with the mismeasured covariate problem with validation data when measurement error is described nonparametrically. Stefanski et al. [35] proposed a semiparametric correction for bias caused by measurement error. Jiang et al. [14] discussed semiparametric regression models with random effects and measurement errors. The error distribution is not fully specified except for assumptions on the moment **generating function** of the error. Kulich & Lin [16] developed a class of estimating functions for the regression parameters for the

additive hazards models with covariates subject to measurement error. Tsiatis & Davidian [36] developed a semiparametric method for estimation in the context of the proportional hazards model with **time-dependent covariates** measured with error. Huang & Wang [13] avoid distributional assumption and proposed a nonparametric approach to deal with the Cox model when repeated measurements on the covariates are available.

An Application Involving Misclassification

Misclassified Exposure Variables in Case–Control Studies

We now consider an illustrative example based on a case–control study examining the association between invasive cervical cancer and exposure to herpes simplex virus type 2 (HSV-2) [3]. Exposure to HSV-2 was assessed both by a refined western blot procedure and by a less accurate western blot procedure for cases ($Y = 1$) and controls ($Y = 0$). Primary interest is in evaluating the relationship between Y and the result of the refined western blot test (X), but this test result is only directly observed for less than 6% of the subjects. Data based on the less accurate standard western blot test (W) are available for all subjects. Data reported in [3] are reproduced in Table 1.

If there are n subjects under study, let $Y_i = 1$ if subject i is a case and $Y_i = 0$ otherwise, $i = 1, 2, \ldots, n$. Let X_i and W_i be the result of the refined and standard western blot tests respectively for subject i, and let $V_i = I(X_i \text{ is observed}), i = 1, 2, \ldots, n$. Then $\mathcal{V} = \{i : V_i = 1\}$ is the index set for subjects in the validation sample and $\overline{\mathcal{V}} = \{i : V_i = 0\}$ is the index set of subjects in the nonvalidation sample. The triple $\{Y_i, X_i, W_i\}$ is therefore available if $i \in \mathcal{V}$ but only $\{Y_i, W_i\}$ is observed if $i \in \overline{\mathcal{V}}$. Let $n_{vy} = \sum_{i=1}^{n} I(V_i = v, Y_i = y)$ and $n_v = \sum_{i=1}^{n} I(V_i = v)$, so, for example, n_{11} is the number of cases and n_1 is the total number of subjects in the validation data set.

Table 1 Validation and nonvalidation data from cervical cancer case–control study reported in [4]

	Y	X	W	Frequency
Validation data	1	0	0	13
	1	0	1	3
	1	1	0	5
	1	1	1	18
	0	0	0	33
	0	0	1	11
	0	1	0	16
	0	1	1	16
Nonvalidation data	1		0	318
	1		1	375
	0		0	701
	0		1	535

Consider a prospective logistic model

$$P(Y_i = 1|X_i = x, V_i = v) = \frac{\exp(\beta_0^{(v)} + \beta_1 x)}{1 + \exp(\beta_0^{(v)} + \beta_1 x)}, \tag{6}$$

where $\beta_0^{(0)}$ and $\beta_0^{(1)}$ are the intercepts corresponding to the nonvalidation and validation data sets respectively, and β_1 characterizes the increase in risk of cervical cancer with exposure to HSV-2 according to the refined western blot test. Let $\beta = (\beta_0^{(0)}, \beta_0^{(1)}, \beta_1)'$. Note that

$$\beta_0^{(1)} = \beta_0^{(0)} + \log\left(\frac{P(V_i = 1|Y_i = 1)/P(V_i = 0|Y_i = 1)}{P(V_i = 1|Y_i = 0)/P(V_i = 0|Y_i = 0)}\right), \tag{7}$$

and if $\pi_v = P(Y_i = 1|V_i = v)$ then the second term on the right-hand side can be reexpressed as $\log(\pi_1(1 - \pi_0))/(\pi_0(1 - \pi_1))$ and since $\hat{\pi}_v = n_{v1}/n_v$, it is estimated by $\log(n_{11}n_{00}/n_{10}n_{01})$.

The objective here is to estimate β_1 based on the full data set, taking into account the fact that W is a misclassified assessment of X. Inferences on β_1 may be carried out by considering the likelihood

$$\begin{aligned} L = \prod_{i \in \mathcal{V}} & [P(X_i = x_i|Y_i = y_i, V_i = 1) \\ & \cdot P(W_i = w_i|X_i = x_i, Y_i = y_i, V_i = 1)] \\ & \cdot \prod_{i \in \overline{\mathcal{V}}} \left[\sum_{x_i} P(X_i = x_i|Y_i = y_i, V_i = 0) \right. \\ & \left. \cdot P(W_i = w_i|X_i = x_i, Y_i = y_i, V_i = 0) \right], \end{aligned} \tag{8}$$

where we note that

$$P(X_i = 1|Y_i = y, V_i = v) = \frac{P(Y_i = y|X_i = 1, V_i = v) \cdot P(X_i = 1|V_i = v)}{\pi_v}.$$

The conditional probabilities $\delta_{xy} = P(W_i = 1|X_i = x, Y_i = y, V_i = v)$ characterize the misclassification distribution and we let $\delta = (\delta_{00}, \delta_{01}, \delta_{10}, \delta_{11})'$. The assumption of nondifferential misclassification is imposed when we constrain $\delta_{xy} = \delta_x$, $x, y = 0, 1$.

The likelihood in (8) is a function of $\theta = (\beta', \delta', \lambda')'$, where λ is a vector of additional nuisance parameters required to evaluate $P(X_i = x|V_i = v)$ and $\pi_v = P(Y_i = y|V_i = v)$. We may replace π_v by the empirical estimate n_{v1}/n_v, $v = 0, 1$, and λ may simply comprise the parameters $\lambda_1 = P(X_i = 1|V_i = 1)$ and $\lambda_2 = P(X_i = 1|V_i = 0)$.

Table 2 summarizes parameter estimates and standard errors obtained from likelihood analyses of the validation data alone and the full data set [3]. For both data sets, the likelihood in (8) is constructed under the assumption of differential and nondifferential misclassification. The estimate of the **odds ratio** characterizing the increase in risk of cervical cancer with exposure to HSV-2 based on the full data set is 1.84 (95% CI (0.93, 3.65)) under the assumption of differential misclassification and 2.61 (95% CI (1.64, 4.15)) under the assumption of nondifferential misclassification. These estimates can be contrasted with the findings from a naive analysis of the full data set in which W_i is used in place of X_i, where we obtain an odds ratio estimate of 1.57 (95% CI (1.31, 1.89)). Analyses based on the validation data alone give unbiased estimators of β_1, but with somewhat lower precision.

Misspecification of the Misclassification Distribution

Differential misclassification is a concern in case–control studies since disease status is known at the start of the study and cases and controls can therefore be treated differently. If the true error distribution features differential misclassification, but one assumes $\delta_{xy} = \delta_x$, then the estimator maximizing (8) will not be consistent for β. Some authors have empirically investigated the impact of assuming nondifferential misclassification when it is in fact differential (e.g. [3]) and a detailed study can be carried out using the theory of **misspecified** models [39].

Let $S(\theta)$ be the estimating function obtained from the assumed likelihood (8), and let $\hat{\theta}$ be the solution to $S(\theta) = 0$. If (8) is constructed assuming nondifferential misclassification, the asymptotic bias in $\hat{\beta}_1$ under differential misclassification can be evaluated as follows. Let $E_T(\cdot)$ denote the expectation operator with respect to the true joint distribution involving differential misclassification, given by

$$\prod_{i \in \mathcal{V}} P(W_i = w_i, X_i = x_i|Y_i = y_i, V_i = 1) \cdot \prod_{i \in \overline{\mathcal{V}}} P(W_i = w_i|Y_i = y_i, V_i = 0). \quad (9)$$

This joint distribution may be reexpressed as

$$\prod_{i \in \mathcal{V}} [P(X_i = x_i|Y_i = y_i, V_i = 1) \cdot P(W_i = w_i|X_i = x_i, Y_i = y_i, V_i = 1)] \cdot \prod_{i \in \overline{\mathcal{V}}} \Big[\sum x_i P(X_i = x_i|Y_i = y_i, V_i = 0) \cdot P(W_i = w_i|X_i = x_i, Y_i = y_i, V_i = 0) \Big]. \quad (10)$$

Table 2 Parameter estimates and standard errors arising from likelihood analyses of cervical cancer data [4]

	Validation data set				Entire data set			
	Differential		Nondifferential		Differential		Nondifferential	
Parameter	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
β_1	0.681	0.400	0.681	0.400	0.609	0.350	0.958	0.237
δ_{00}	0.250	0.065	–	–	0.311	0.055	–	–
δ_{01}	0.188	0.098	–	–	0.189	0.085	–	–
δ_{10}	0.500	0.088	–	–	0.578	0.0657	–	–
δ_{11}	0.783	0.086	–	–	0.784	0.068	–	–
δ_0	–	–	0.223	0.055	–	–	0.257	0.043
δ_1	–	–	0.618	0.064	–	–	0.679	0.041

Let $E_T(S(\theta))$ denote the expectation of the estimating function and let θ^* denote the value of θ solving $E_T(S(\theta)) = 0$. Then θ^* represents the value to which the naive estimators of θ converge, so $\beta_1^* - \beta_1$ represents the asymptotic bias in the estimator of β_1 when nondifferential misclassification is incorrectly assumed.

To illustrate, consider the same model adopted in the preceding example and take the design features and true parameter values to be roughly comparable to those reported. Specifically, we set $\beta_0^{(0)} = -1$, $\beta_1 = \log(2)$, $n_{00} = 1600$, $n_{01} = 800$, $n_{10} = 200$, and $n_{11} = 100$. We let $\delta_{00} = 0.3$, $\delta_{10} = 0.6$, and let $\psi_v = \delta_{v1}(1 - \delta_{v0})/(\delta_{v0}(1 - \delta_{v1}))$, $v = 0, 1$. If $\psi_0 = \psi_1 = 1$, then the misclassification is nondifferential, but not otherwise. Figures 1 and 2 illustrate how the asymptotic percent relative bias in β_1 varies as a function of $(\psi_0, \psi_1)'$, where attention is restricted to

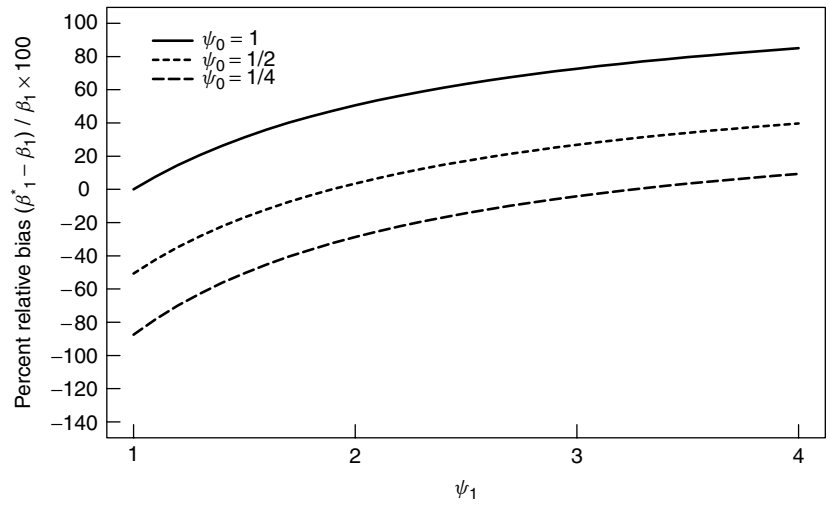

Figure 1 Asymptotic percent relative bias in regression coefficient $(\beta_1^* - \beta_1)/\beta_1 \times 100$ as a function of ψ_0 for specified values of ψ_1

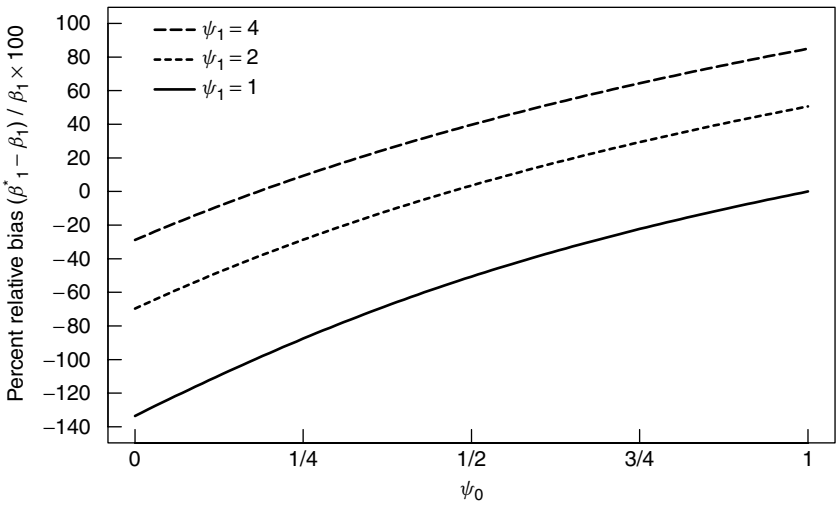

Figure 2 Asymptotic percent relative bias in regression coefficient $(\beta_1^* - \beta_1)/\beta_1 \times 100$ as a function of ψ_1 for specified values of ψ_0

values of ψ_v, $v = 0, 1$ such that the respective misclassification rates are no larger than those implied by δ_{00} and δ_{10}. Note that the resulting bias may be positive or negative in sign, reflecting the fact that bias away from the null value may arise under nondifferential misclassification. Moreover, seemingly small degrees of differential misclassification can lead to appreciable bias.

References

[1] Amemiya, Y. & Fuller, W.A. (1988). Estimation for the nonlinear functional relationship, *Annals of Statistics* **16**, 147–160.

[2] Armstrong, B. (1985). Measurement error in generalized linear models, *Communications in Statistics, part B - Simulation and Computation* **14**, 529–544.

[3] Carroll, R.J., Gail, M.H. & Lubin, J.H. (1993). Case-control studies with errors in covariates, *Journal of the American Statistical Association* **88**, 185–199.

[4] Carroll, R.J., Küchenhoff, H., Lombard, F. & Stefanski, L.A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models, *Journal of the American Statistical Association* **91**, 242–250.

[5] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement error in Nonlinear Models*. Chapman & Hall, London.

[6] Carroll, R.J., Spiegelman, C., Lan, K.K., Bailey, K.T. & Abbott, R.D. (1984). On errors-in-variables for binary regression models, *Biometrika* **71**, 19–26.

[7] Carroll, R.J. & Stefanski, L.A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors, *Journal of the American Statistical Association* **85**, 652–663.

[8] Cook, J. & Stefanski, L.A. (1994). A simulation extrapolation method for parametric measurement error models, *Journal of the American Statistical Association* **89**, 464–467.

[9] Fuller, W.A. (1987). *Measurement Error Models*. Wiley, New York.

[10] Gleser, L.J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models, in *Statistical Analysis of Measurement Error Models and Application*, P.J. Brown & W.A. Fuller, eds. American Mathematical Society, Providence.

[11] Hanfelt, J.J. & Liang, K.-Y. (1997). Approximate likelihood for generalized linear errors-in-variables models, *Journal of Royal Statistical Society, Ser. B* **59**, 627–637.

[12] Hu, P., Tsiatis, A.A. & Davidian, M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error, *Biometrics* **54**, 1407–1419.

[13] Huang, Y. & Wang, C.Y. (2000). Cox regression with accurate covariates unascertainable: a nonparametric-correction approach, *Journal of the American Statistical Association* **95**, 1209–1219.

[14] Jiang, W., Turnbull, B.W. & Clark, L.C. (1999). Semiparametric regression models for repeated events with random effects and measurement error, *Journal of the American Statistical Association* **94**, 111–124.

[15] Kannel, W.B., Neaton, J.D., Wentworth, D., Thomas, H.E., Stamler, J., Hulley, S.B. & Kjelsberg, M.O. (1986). Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for MRFIT, *American Heart Journal* **112**, 825–836.

[16] Kulich, M. & Lin, D.Y. (2000). Additive hazards regression with covariate measurement error, *Journal of the American Statistical Association* **95**, 238–248.

[17] Nakamura, T. (1990). Corrected score functions for errors-in-variables models: methodology and application to generalized linear models, *Biometrika* **77**, 127–137.

[18] Novick, S.J. & Stefanski, L.A. (2002). Corrected score estimation via complex variable simulation extrapolation, *Journal of the American Statistical Association* **97**, 472–481.

[19] Pepe, M.S. & Fleming, T.R. (1991). A nonparametric method for dealing with mismeasured covariate data, *Journal of the American Statistical Association* **86**, 108–113.

[20] Pierce, D.A., Stram, D.O., Vaeth, M. & Schafer, D. (1992). Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors, *Journal of the American Statistical Association* **87**, 351–359.

[21] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[22] Prentice, R.L. (1996). Dietary fat and breast cancer: measurement error and results from analytic epidemiology, *Journal of the National Cancer Institute* **88**, 1738–1747.

[23] Rosner, B., Spiegelman, D. & Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error, *American Journal of Epidemiology* **132**, 734–745.

[24] Rosner, B., Willett, W.C. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error, *Statistics in Medicine* **8**, 1051–1070.

[25] Rudemo, M., Ruppert, D. & Streibig, J.C. (1989). Random effect models in nonlinear regression with applications to bioassay, *Biometrics* **45**, 349–362.

[26] Satten, G.A. & Kupper, L.L. (1993). Inferences about exposure-disease association using probability of exposure information, *Journal of the American Statistical Association* **88**, 200–208.

[27] Schafer, D. (1987). Covariate measurement error in generalized linear models, *Biometrika* **74**, 385–391.

[28] Schafer, D. (1993). Likelihood analysis for probit regression with measurement errors, *Biometrika* **80**, 899–904.

[29] Schafer, D.W. & Purdy, K.G. (1996). Likelihood analysis for errors-in-variables regression with replicate measurements, *Biometrika* **83**, 813–824.

[30] Stefanski, L.A. (1985). The effects of measurement error on parameter estimation, *Biometrika* **72**, 583–592.

[31] Stefanski, L.A. & Carroll, R.J. (1985). Covariate measurement error in logistic regression, *Annals of Statistics* **13**, 1335–1351.

[32] Stefanski, L.A. & Carroll, R.J. (1987). Conditional scores and optimal scores in generalized linear measurement error models, *Biometrika* **74**, 703–716.

[33] Stefanski, L.A. & Carroll, R.J. (1990). Structural logistic regression measurement models, in *Proceedings of the Conference on measurement Error Models*, P.J. Brown & W.A. Fuller, eds. Wiley, New York.

[34] Stefanski, L.A. & Cook, J. (1995). Simulation extrapolation: the measurement error jackknife, *Journal of the American Statistical Association* **90**, 1247–1256.

[35] Stefanski, L.A., Knickerbocker, R. & Carroll, R.J. (1994). A semiparametric correction for attenuation, *Journal of the American Statistical Association* **89**, 1366–1373.

[36] Tsiatis, A.A. & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error, *Biometrika* **88**, 447–458.

[37] Wang, N., Carroll, R.J. & Liang, K.Y. (1996). Quasi-likelihood and variance functions in measurement error models with replicates, *Biometrics* **52**, 401–411.

[38] Wang, C.Y., Hsu, L., Feng, Z.D. & Prentice, R.L. (1997). Regression calibration in failure time regression, *Biometrics* **53**, 131–145.

[39] White, H.A. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1–25.

[40] Whittemore, A.S. & Gong, G. (1991). Poisson regression with misclassified counts: application to cervical cancer mortality rates, *Applied Statistics* **40**, 81–93.

[41] Whittemore, A.S. & Keller, J.B. (1988). Approximations for regression with covariate measurement error, *Journal of the American Statistical Association* **83**, 1057–1066.

(*See also* **Measurement Error in Survival Analysis**)

GRACE Y. YI & RICHARD J. COOK

Errors in Variables

Errors in variables may occur for a variety of reasons; for example, because of limitations of a measuring instrument (e.g. weight) or because of random fluctuations in a physiologic process (e.g. blood pressure). Consider a simple **linear regression** model in which a dependent variable Y is related to an **explanatory variable** X according to $Y = \alpha + \beta X + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Observation errors in Y are usually taken to be part of the error specification in the model, while the explanatory variable or covariate, X, is usually assumed to be fixed. Concern is therefore primarily with errors in the measurement of X. Here, we shall refer to the value of the explanatory variable, if it could be measured without error, as the "true covariate" and to the actual value of the variable as the "observed covariate".

If the estimated regression parameters are to be used only for prediction and if the true value of the explanatory variable will never be available, then the usual regression model is appropriate [7, 8]. However, errors in variables can be a problem when the primary interest is in the estimation of the relationship between the dependent variable and the true covariate values.

In uncontrolled experiments, in which the independent variable is allowed to vary freely, the observed covariate is usually assumed to be measured as the true covariate plus an error term. Such errors in measurement cause the estimate of the slope parameter to be biased [3, 7]. The estimate of the slope obtained from the observed covariate, β^*, is related to the slope appropriate for the true covariate, β, according to $\beta^* = \beta\sigma_x^2/(\sigma_x^2 + \sigma_\varepsilon^2)$, where σ_x^2 is the variance of the true covariate in the population and σ_ε^2 is the variance of the errors. In controlled experiments in which the dependent variable is measured at prespecified levels of the explanatory variable, the estimated coefficients are found to be unbiased. This is true because on average, over repeated experiments, the value of the explanatory variable is correct in controlled experiments. The controlled experiment, in which the true covariate is randomly distributed about the observed covariate, has become known as the "Berkson model" in the literature [3].

Generalized Linear Models

A number of strategies for revising estimates of coefficients of mismeasured covariates in a **generalized linear model** have been suggested. Armstrong [1] describes the use of an iteratively reweighted least squares algorithm (*see* **Generalized Linear Model**) to find maximum **quasi-likelihood** estimates, Stefanski [15] describes methods for reducing bias in M-estimates (*see* **Robustness**), and Schafer [14] suggests using the **EM algorithm** and treating the mismeasured covariates as missing variables. Methods of correcting the score function (*see* **Likelihood**) and estimating its efficiency have also been suggested for use when covariates are measured with errors in generalized linear models [9, 17]. It has been found that errors in **confounding** covariates included in the model may lead to incorrect conclusions when testing for treatment effects in unbalanced nonrandomized studies [5] (*see* **Measurement Error in Epidemiologic Studies**). The severity of this problem depends on the size of the error, the degree to which the study is unbalanced, and the strength of the relationship between the confounder and the dependent variable. When confounders are measured with error in randomized studies, treatment effects are underestimated [5]. Pepe & Fleming [10] discuss the use of **nonparametric methods** to estimate coefficients of covariates measured with error when both the true and the observed data are available for a subsample of the population.

Regression Models for Binary Data

Error in the measurement of the explanatory variable in **logistic regression** causes estimates of the **odds ratio** to be biased toward one in most situations [13, 15, 16]. When the majority of the cases have very high or very low risk, however, the odds ratios may be biased away from one [15]. If more than one variable is included in the model, and one or more of these is measured with error, any of the effects may be under- or overestimated, even for variables measured without error [13]. Estimates of regression coefficients from probit models (*see* **Quantal Response Models**), for covariates measured with error according to the Berkson model, have also been shown to be biased [4]. The use of validation information to correct estimates of relative risk has been suggested for probit regression with binary or ordinal

outcomes [18] and for logistic regression when more than one covariate is measured with error [13]. Methods of improving estimates of relative risks have been suggested for situations in which the errors in the covariates are assumed to be normally distributed [6, 15]. Errors in covariates have also been shown to bias the asymptotic levels of test statistics related to treatment in nonrandomized studies when the covariates are unbalanced between treatment groups [15].

Regression Models for Survival Data

Error in the measurement of controlled covariates has been shown to bias regression coefficient estimates in the Cox **relative risk** regression model [12] (*see* **Cox Regression Model**; **Measurement Error in Survival Analysis**). In this model, explanatory variables may depend on time. When the true covariate, $Z(t)$, can be assumed to be normally distributed about the observed covariate, $X(t)$, with variance $h[X(t)]\sigma^2$, the estimated relative risk is $\mathrm{E}\{\exp[\beta Z(t)]\} = \exp\{\beta X(t) + \beta^2 h[X(t)]\sigma^2/2\}$. Error in the measurement of uncontrolled covariates also causes estimates of the relative risk from this model to be biased [11]. The coefficient of the observed covariate, β^*, is related to the coefficient of the true covariate, β, according to $\beta^* = \beta i(\beta)/(i(\beta) + \sigma^2)$, where $i(\beta)$ is the second derivative of the log likelihood and σ^2 is the variance of the errors. The amount of bias in the coefficient of the variable measured with error has been shown to be increased by the presence of a confounder, even when the confounder is measured without error [2]. When only the confounder is measured with error, then one may not fully adjust for it and the coefficient of the covariate of interest will still be biased due to confounding. If both the confounder and the covariate of interest are measured with error, and if the errors in these covariates are correlated, the relative risk may be biased toward or away from one.

References

[1] Armstrong, B. (1985). Measurement error in the generalized linear models, *Communications in Statistics – Simulation and Computation* **14**, 529–544.

[2] Armstrong, B. (1990). The effects of measurement errors on relative risk regressions, *American Journal of Epidemiology* **132**, 1176–1184.

[3] Berkson, J. (1950). Are there two regressions?, *Journal of the American Statistical Association* **78**, 90–98.

[4] Burr, D. (1988). On errors-in-variables in binary regression – Berkson case, *Journal of the American Statistical Association* **83**, 739–743.

[5] Carroll, R.J. (1989). Covariance analysis in generalized linear measurement error models, *Statistics in Medicine* **8**, 1075–1093.

[6] Carroll, R.J., Spiegelman, C.H., Lan, K.K.G., Bailey, K.T. & Abbott, R.D. (1984). On errors-in-variables for binary regression models, *Biometrika* **71**, 19–25.

[7] Madansky, A. (1959). The fitting of straight lines when both variables are subject to error, *Journal of the American Statistical Association* **54**, 173–205.

[8] Miller, R.G. (1986). *Beyond ANOVA, Basics of Applied Statistics*. Wiley, New York.

[9] Nakamura, T. (1990). Corrected score function for errors-in-variables models: methodology and application to generalized linear models, *Biometrika* **77**, 127–137.

[10] Pepe, M.S. & Fleming, T.R. (1991). A non-parametric method for dealing with mismeasured covariate data, *Journal of the American Statistical Association* **86**, 108–113.

[11] Pepe, M.S., Self, S.G. & Prentice, R.L. (1989). Further results on covariate measurement errors in cohort studies with time to response data, *Statistics in Medicine* **8**, 1167–1178.

[12] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[13] Rosner, B., Spiegelman, D. & Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error, *American Journal of Epidemiology* **132**, 734–745.

[14] Schafer, D.W. (1987). Covariate measurement error in generalized linear models, *Biometrika* **74**, 385–391.

[15] Stefanski, L.A. (1985). The effects of measurement error on parameter estimation, *Biometrika* **72**, 583–592.

[16] Stefanski, L.A. & Carroll, R.J. (1985). Covariate measurement error in logistic regression, *Annals of Statistics* **13**, 1335–1351.

[17] Tosteson, T.D. & Tsiatis, A.A. (1988). The asymptotic relative efficiency of score tests in the generalized linear model with surrogate covariate, *Biometrika* **75**, 507–514.

[18] Tosteson, T.D., Stefanski, L.A. & Schafer, D.W. (1989). A measurement-error model for binary and ordinal regression, *Statistics in Medicine* **8**, 1139–1147.

J.M. RABOUD

Establishment Surveys With Population Survey–Generated Sampling Frames

Introduction

Though virtually all establishment surveys use stand-alone establishment **sampling frames**, defined as frames that list all establishments in the universe, the stand-alone frame is not a requirement for unbiased **estimation**. When stand-alone frames contain establishment size measures, the Hansen–Hurwitz (HH) pps estimator is typically used to estimate the volume of transactions between establishments and households (*see* **Sampling With Probability Proportional to Size**). This article features a **network sampling** (NS) version of the HH estimator. Because the NS estimator does not use a stand-alone frame, it is a potential competitor of the HH estimator, particularly when stand-alone frames with good quality establishment coverage and size measures are unavailable or are relatively expensive to construct and maintain. The NS estimator uses a population-survey-generated establishment sampling frame – a sample sampling frame that lists the households enumerated in a population **sample survey**, the establishments with whom the survey households have transactions, and the number of transactions each establishment has with survey households. The population-survey-generated frame is constructed with information that is collected from household respondents in the population sample survey.

The survey design of the Consumer Price Index (CPI) is a rare and notable example of an establishment survey design that uses population-survey-generated sampling frames [1]. CPI pricing surveys use population-survey-generated sampling frames that contain the business establishments, which sell commodities to households in the CPI continuing point of purchase surveys, a population sample survey.

This article features a two-stage unbiased NS estimator of X, the sum of the x-variate over transactions between establishments and populations residing in households. The NS estimator and its first- and second-stage **variance components** are presented in the section "NS Estimation". Clusters of establishments with whom survey households have transactions are the first-stage sampling units, and the transactions that those establishments have with all households are the second-stage selection units. The effect of alternative second-stage sampling procedures on the stage 2 **variance** of the NS estimator is also presented in the section "NS Estimation". The well-known two-stage pps HH estimator and variance are presented in the section "HH Estimation" (*see* **Multistage Sampling**). The section "Relative Precision of the HH and NS Estimators" compares the precision of the two estimators in single- and two-stage sample designs under conditions yielding roughly the same expected sample sizes in both types of establishment surveys. Recent research on population-survey-generated sampling frames is cited in the section "Research on Population-survey-generated Frames", and some potential benefits and outstanding challenges of research on population-survey-generating frames are summarized in the final section.

Research on Population-survey-generated Frames

Research on designing establishment sample surveys using population-survey-generated sampling frames has a relatively short history and is in an early developmental stage. The research was motivated by a recommendation of a Panel of the Committee on National Statistics [11] convened by the **National Center for Health Statistics** (NCHS) to review its plans to restructure the national health care provider surveys [3]. The Panel proposed using listings of health care providers servicing households enumerated in the National Health Interview Survey (NHIS) as the sampling frames of health care provider surveys. (The NHIS [2] is a continuing survey of the civilian noninstitutional population.) In view of the difficulties of constructing and maintaining good quality stand-alone frames, especially during periods of rapid institutional structural changes like the ones now occurring in the nation's health care delivery system, the Panel proposed that NHIS-generated establishment frames might provide better quality frames at lower frame construction and maintenance costs.

NCHS [4] presents rough comparisons of the precision of estimates of the volume of dental services between two-stage sample surveys using a stand-alone sampling frame of dentists and an NHIS-generated sample frame of dentists. Difficulties initially encountered in obtaining closed formulas of the variances of establishment survey estimates using population-survey-generated sampling frames were overcome by developing two-stage network sampling estimators [6, 10]). Sirken, Shimizu, and Judkins [9] and Shimizu and Sirken [5] determined the Hansen–Hurwitz versions of the NS estimator and variance presented in the section "NS Estimation". Sirken [7] determined the differences in the sampling efficiencies of the two estimators presented in the section "Relative Precision of the HH and NS Estimators". Sirken and Shimizu [8] determined the Horvitz–Thompson version of the two-stage NS estimator, but it is not included in this article.

Notation

A universe of R establishments has M transactions with a population of N households.

Let M_{ij} denote the number of transactions which household i $(i = 1, 2, \ldots, N)$ has with establishment j $(j = 1, 2, \ldots, R)$, where $M_{ij} \geq 0$. Then

$M_j = \sum_{i=1}^{N} M_{ij}$ = the number of transactions of establishment j with N households and

$M = \sum_{j=1}^{R} M_j$ = the total number of transactions between R establishments and N households.

$\overline{M} = M/N$ = the average number of transactions per household.

Let X_{jk} denote the value of the x-variate for transaction k $(k = 1, 2, \ldots, M_j)$ of establishment j $(j = 1, 2, \ldots, R)$. Then,

$X_j = \sum_{k=1}^{M_j} X_{jk}$ = the sum of the x-variate over the M_j transactions of establishment j,

$X = \sum_{j=1}^{R} X_j$ = the sum of the x-variate over the transactions of R establishments, and

$\overline{X}_j = X_j/M_j$ = the average value of the x-variate over the M_j transactions of establishment j.

NS Estimation

From the perspective of network sampling, a two-stage establishment survey using a population-survey-generated frame is viewed as a two-stage population sample survey in which clusters of establishments having transactions with households $(i = 1, 2, \ldots, N)$ are first-stage selection units, and the M_j transactions of establishment j $(j = 1, 2, \ldots, R)$ that has transactions with household i, $M_{ij} > 0$, are second-stage selection units.

This article assumes that a sample of n households is selected by **simple random sample** (srs) with probabilities $\pi_i = \pi = n/N$ $(i = 1, 2, \ldots, n)$ and with replacement (*see* **Sampling With and Without Replacement**). (However, the NS estimator can be applied to more complex population survey designs than are considered in this article.) Respondents at the N households identify establishments with whom they have transactions and report the number, M_{ij} $(i = 1, 2, \ldots, n)$ $(j = 1, \ldots, R)$, of their transactions with each establishment. (The population-survey-generated frame is constructed with the information provided by household respondents). In the survey conducted with the r_{NS} establishments reported by the N sample households, establishment j $(j = 1, 2, \ldots, r_{\mathrm{NS}})$ reports the values of the x-variate for a sample of $m_j = c_{\mathrm{NS}} \sum_i^n M_{ij}$ from its M_j transactions, where c_{NS} is a positive integer. Two sampling procedures for selecting m_j of the M_j transactions from establishment j $(j = 1, 2, \ldots, r_{\mathrm{NS}})$ are presented and compared:

Procedure (α): Samples of $m_{ij} = c_{\mathrm{NS}} M_{ij}$ transactions are independently drawn from the M_j transactions of establishment j by srs without replacement for each household i $(i = 1, \ldots, n)$ that has transactions with establishment j;

Procedure (β): A single sample of $m_j = c_{\mathrm{NS}} \sum_i^n M_{ij}$ transactions is drawn from the M_j transactions of establishment j by srs without replacement.

The **unbiased** NS estimator of X using stage 2 sampling procedure α is

$$X'_{\mathrm{NS}} = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{R} M_{ij} \overline{X}'_j(i), \qquad (1)$$

where

$$\overline{X}'_j(i) = \frac{1}{c_{\mathrm{NS}} M_{ij}} \sum_{k=1}^{c_{\mathrm{NS}} M_{ij}} X_{jk} \tag{2}$$

is an unbiased estimate of $\overline{X}_j$ based on the sample of $c_{\mathrm{NS}} M_{ij}$ transactions of establishment j selected for household i. Because households are selected with replacement, the NS estimator counts the quantity $\sum_j^R M_{ij} \overline{X}'_j(i)$ every time the same household i ($i = 1, 2, \ldots, n$) is selected in the sample. Because the same establishment has transactions with multiple households, the NS estimator also counts $M_{ij} \overline{X}'_j(i)$ every time establishment j has transactions with a different sample household.

The variance of X'_{NS} is

$$\mathrm{Var}(X'_{\mathrm{NS}}) = \frac{N^2}{n} \sigma^2_{\mathrm{NS1}} + \frac{N}{n c_{\mathrm{NS}}} \sum_{i=1}^{N} \sum_{j=1}^{R} M_{ij} \times \left(1 - \frac{c_{\mathrm{NS}} M_{ij}}{M_j}\right) \sigma_j^2, \tag{3}$$

where the first and second terms respectively on the right side of (3) are the first- and second-stage variance components, and

$$\sigma^2_{\mathrm{NS1}} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{R} M_{ij} \overline{X}_j - \frac{X}{N} \right)^2 \tag{4}$$

is the between-household population variance and

$$\sigma_j^2 = \frac{1}{M_j} \sum_{k=1}^{M_j} (X_{jk} - \overline{X}_j)^2 \tag{5}$$

is the within-establishment population variance for establishment j ($j = 1, 2, \ldots, R$).

An unbiased estimate of the variance of the NS estimator in (3) is

$$\hat{\sigma}^2_{\mathrm{NS}} = \frac{N^2}{n(n-1)} \sum_{i=1}^{n} \left(\sum_{j=1}^{R} M_{ij} \overline{X}'_j(i) - \frac{X'}{N} \right)^2 + \frac{N}{n c_{\mathrm{NS}}} \sum_{i=1}^{n} \sum_{j=1}^{R} M_{ij} \left(1 - \frac{c_{\mathrm{NS}} M_{ij}}{M'_j}\right) \hat{\sigma}_j^2, \tag{6}$$

where

$$\hat{\sigma}_j^2 = \frac{1}{m_j - 1} \sum_{k=1}^{m_j} (X_{jk} - \overline{X}'_j)^2 \tag{7}$$

is an unbiased estimate of σ_j^2 and

$$\overline{X}'_j = \frac{1}{m_j} \sum_{k=1}^{m_j} X_{jk} \tag{8}$$

is the unbiased estimate of the transaction mean $\overline{X}_j$ for establishment j.

The second term on the right side of (3) represents the stage 2 variance using sampling procedure α. The stage 2 variance using sampling procedure β is

$$\left(\frac{N}{n}\right)^2 \mathrm{E} \left[\sum_{i-1}^{n} \sum_{j=1}^{R} M_{ij}^2 \left(\frac{1}{m_j} - \frac{1}{M_j} \right) \sigma_j^2 \right]. \tag{9}$$

Using the inequality $\sum_i M_{ij}^2 \le \left(\sum_i M_{ij}\right)^2$, it is demonstrated below that the stage 2 variance based on sampling procedure α is the upper bound of the stage 2 variance based on sampling procedure β.

$$\begin{aligned}
&\left(\frac{N}{n}\right)^2 \mathrm{E} \left[\sum_{i-1}^{n} \sum_{j=1}^{R} M_{ij}^2 \left(\frac{1}{m_j} - \frac{1}{M_j} \right) \sigma_j^2 \right] \\
&= \left(\frac{N}{n}\right)^2 \mathrm{E} \left[\sum_{j=1}^{R} \left(\frac{\sum_{i=1}^{n} M_{ij}^2}{c_{\mathrm{NS}} \sum_{i=1}^{n} M_{ij}} - \frac{\sum_{i=1}^{n} M_{ij}^2}{M_j} \right) \sigma_j^2 \right] \\
&\le \left(\frac{N}{n}\right)^2 \mathrm{E} \left[\sum_{j=1}^{R} \left(\frac{\left(\sum_{i=1}^{n} M_{ij}\right)^2}{c_{\mathrm{NS}} \sum_{i=1}^{n} M_{ij}} - \frac{\sum_{i=1}^{n} M_{ij}^2}{M_j} \right) \sigma_j^2 \right] \\
&= \left(\frac{N}{n}\right)^2 \sum_{j=1}^{R} \mathrm{E} \left(\frac{\sum_{i=1}^{n} M_{ij}}{c_{\mathrm{NS}}} - \frac{\sum_{i=1}^{n} M_{ij}^2}{M_j} \right) \sigma_j^2 \\
&= \frac{N}{n c_{\mathrm{NS}}} \sum_{i=1}^{N} \sum_{j=1}^{R} M_{ij} \left(1 - \frac{c_{\mathrm{NS}} M_{ij}}{M_j}\right) \sigma_j^2.
\end{aligned} \tag{10}$$

Thus, when feasible, it is more efficient to use sampling procedure β rather than procedure α at the second stage.

HH Estimation

A two-stage establishment sample survey using a stand-alone frame is conducted to estimate X. The stand-alone frame lists R establishments and the number of their respective transactions M_j $(j = 1, 2, \ldots, R)$. At stage 1, a sample of r_{HH} establishments j $(j = 1, \ldots, r_{\mathrm{HH}})$ is selected by pps with replacement, and at stage 2, a fixed size sample of c_{HH} transactions is independently selected from the M_j transactions of establishment j $(j = 1, 2, \ldots, r_{\mathrm{HH}})$ by srs without replacement.

The unbiased HH estimator of X is

$$X'_{\mathrm{HH}} = \frac{M}{r_{\mathrm{HH}}} \sum_{j=1}^{r} \overline{X}'_j, \tag{11}$$

where $\overline{X}'_j = \sum_{k=1}^{c_{\mathrm{HH}}} X_{jk}/c_{\mathrm{HH}}$ is an unbiased estimate of $\overline{X}_j = X_j/M_j$ $(j = 1, 2, \ldots, R)$. The variance of X'_{HH} is

$$\mathrm{Var}(X'_{\mathrm{HH}}) = \frac{M^2}{r_{\mathrm{HH}}}\sigma^2_{\mathrm{HH1}} + \frac{M}{r_{\mathrm{HH}}c_{\mathrm{HH}}} \times \sum_{j=1}^{R} (M_j - c_{\mathrm{HH}})\sigma_j^2, \tag{12}$$

where the first and second terms respectively on the right side of (12) are the first- and second-stage variance components, and

$$\sigma^2_{\mathrm{HH1}} = \frac{1}{M}\sum_{j=1}^{R} M_j \left(\overline{X}_j - \frac{X}{M}\right)^2 \tag{13}$$

is the between-establishment variance, and σ_j^2 is the within-establishment population variance for establishment j $(j = 1, 2, \ldots, R)$ defined in (5).

Relative Precision of the HH and NS Estimators

Let

$$m_{\mathrm{HH}} = c_{\mathrm{HH}} r_{\mathrm{HH}} = \text{the transaction sample size of the HH estimator of } X$$

and

$$m_{\mathrm{NS}} = c_{\mathrm{NS}} \sum_{i=1}^{n}\sum_{j=1}^{R} M_{ij} = \text{the transaction sample size of the NS estimator of } X.$$

Set $c_{\mathrm{NS}} = c_{\mathrm{HH}} = c$ and $r_{\mathrm{HH}} = \sum_{i=1}^{n}\sum_{j=1}^{R} M_{ij}$ and it follows that HH and NS transaction sample sizes are equivalent, $m_{\mathrm{HH}} = m_{\mathrm{NS}}$, and $\mathrm{E}(m_{\mathrm{HH}}) = \mathrm{E}(m_{\mathrm{NS}}) = cn\overline{M}$. (From the viewpoint of survey costs, it is worthy to note that for a fixed transaction sample size, the expected number of distinct establishments is smaller in NS than in HH because multiple households have transactions with the same establishment and often households have multiple transactions with the same establishment.) Under these simplifying conditions, the difference between the variances of the NS and HH estimators of X given in (3) and (12), respectively, can be written as

$$\mathrm{Var}(X'_{\mathrm{NS}}) - \mathrm{Var}(X'_{\mathrm{HH}}) = \frac{N^2}{n}(\sigma^2_{\mathrm{NS}} - \overline{M}\sigma^2_{\mathrm{HH1}}) - \frac{N}{nc}\sum_{j=1}^{R}\rho_j(\alpha)\sigma_j^2, \tag{14}$$

where the first term and second terms, respectively, on the right side of (14) are the differences between the stage 1 and stage 2 variance components of the NS and HH estimators, and

$$\rho_j(\alpha) = (M_j - c) - \left(\frac{1}{M_j}\right)\sum_{i=1}^{N} M_{ij}(M_j - cM_{ij}) = \left(\frac{c}{M_j}\right)\sum_{i=1}^{N} M_{ij}(M_{ij} - 1) \geq 0 \tag{15}$$

is the difference between the HH and NS second-stage finite population corrections for establishment j if the NS estimator is based on sampling procedure α. If the NS estimator is based on sampling procedure β, then $\rho_j(\beta) \geq \rho_j(\alpha)$.

Under the conditions of equivalent transaction sample sizes specified earlier, the NS and HH estimators are equally efficient in stage 1 and stage 2 components, if and only if, the M transactions are distributed over N households such that every household i $(i = 1, 2, \ldots, N)$ has a single transaction. Under these conditions, the NS and HH estimators are equivalent. The direction and magnitude of differences between variances of the NS and HH estimators will vary from survey to survey depending on the kinds and extent of clustering of transactions within households. For example, the first-stage variance component is likely to be less for the HH estimator than for the NS estimator if $\overline{M} = M/N$ is a small fraction and/or if households have multiple transactions

with the same establishments. On the other hand, as evident from (15), the second-stage variance component is less for the NS estimator than for the HH estimator if any of the establishments have multiple transactions with the same households, and the magnitude of their difference depends on the extent of households having multiple transactions with the same establishments.

Summary Remarks

Unbiased estimation can be obtained in establishment surveys that use population survey-generated frames, that is, sample sampling frames that list only the establishments that have transactions with households in a population sample survey. Clearly, population-survey-generated establishment frames definitely deserve consideration whenever stand-alone establishment frames of good quality are unavailable or prohibitively expensive to construct and maintain. However, even when good stand-alone frames are available, population-survey-generated frames may be competitive.

For example, this article compares the sampling variances in two-stage establishment sample survey designs that use a HH estimator and an NS estimator that depend on flawless sampling frames. The HH estimator depends on a stand-alone sampling frame with parametric measures of establishment size, and the NS estimator depends on a population-survey-generated sampling frame showing the total number of transactions of each establishment with survey households. When the transaction sample sizes of the NS estimator and the HH estimator are roughly the same, neither estimator is necessarily more efficient than the other, and the direction and magnitude of the difference between their variances depends on several parameters whose values are likely to vary considerably from survey to survey and between population domains in the same survey. Empirical studies are needed to estimate the values of the parameters.

Empirical studies are also needed to determine the cost and error effects of three major assumptions of the theoretical findings presented in this article: (1) the stand-alone frame and the population-survey-generated frame are flawless in coverage and size measures, (2) the population survey, which yields the population-survey-generated sampling frame is based on an srs design, and (3) the simplifying conditions specified to yield equivalent HH and NS transaction sample sizes. Fortunately, the estimation procedures are **robust** and are applicable to more complex survey designs and less stringent survey conditions than are considered in the article.

References

[1] Leaver, S. & Valliant, R. (1995). Statistical problems in estimating the U.S. consumer price index, in *Business Survey Methods*, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott eds John Wiley & Sons, New York, pp. 543–566.

[2] Massey, J.T., Moore, T.F., Parsons, V.L. & Tadros, W. (1989). Design and Estimation for the National Health Interview Survey, 1985–94. National Center for Health Statistics, *Vital Health Statistics* **2**(110).

[3] McLemore, T. (2000). Health care establishment surveys of the National Center for Health Statistics, in *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, VA, pp. 1181–1186.

[4] National Center for Health Statistics. (1999). National Health Interview Survey: research for the 1995–2004 redesign, *Vital Health Statistics* **2**(126), 76–80.

[5] Shimizu, I. & Sirken, M. (1998). More on population based establishment surveys, in *Proceedings of the Survey Research Section*, American Statistical Association, Alexandria, VA, pp. 7–12.

[6] Sirken, M.G. (1998). Network sampling, in *Encyclopedia of Biostatistics*, P.A. Armitage & T.D. Colton, eds John Wiley & Sons, Chichester, U.K, 2977–2986.

[7] Sirken, M. (2002). Design effects of sampling frames in establishments survey, *Survey Methodology* **28**(2), 183–190.

[8] Sirken, M. & Shimizu, I. (1999). Population-based establishment sample surveys: the Horvitz-Thompson estimator, *Survey Methodology*, **25**(2), 187–191.

[9] Sirken, M., Shimizu, I., Judkins, D. (1995). The population-based establishment surveys, in *Proceedings of the Survey Research Section*, American Statistical Association, Alexandria, VA, pp. 470–473.

[10] Thompson, S. (1992). *Sampling*. John Wiley & Sons, New York, 117–118.

[11] Wunderlich, G.S. ed., (1992). *Toward a National Health Care Survey: At Data System for the 21st Century*. National Research Council and Institute of Medicine National Academy Press, Washington, D.C.

MONROE SIRKEN & IRIS SHIMIZU

Estimating Functions

In common with other areas of statistics, a major activity in biostatistics consists of constructing probabilistic models. These models can be classified very broadly as (i) parametric models and (ii) semiparametric models. In (i) the distributions are specified up to a number of unknown *parameters*, some of which are of scientific interest; the others are commonly called **nuisance parameters**. In (ii) the parameters of scientific interest are modeled directly thus eliminating generally, though not altogether, the effects of the nuisance parameters. For estimation of the unknown parameters two distinct methodologies have been in practice for a long time. The method of **maximum likelihood** was put forward with a statistical justification by Fisher [13], for parametric models in (i) above; the method, however, was also known to **Gauss** a century before. The method of **least squares** was formulated by Legendre [28] and was statistically justified by Gauss [15, 16]. As illustrated below, both methods, maximum likelihood and least squares, have their strengths and weaknesses. It is further demonstrated how the method of *estimating functions* put forward by one of the present authors, Godambe [17], has in recent years provided a unification and extension of the two historical methods of maximum likelihood and least squares. The estimating function methodology eliminates the weaknesses and combines the strengths, of both the maximum likelihood and the least squares methods. This methodology, because of its inbuilt flexibility, provides a versatile tool for applications in diverse areas including biostatistics. The topics particularly covered here are **case–control studies**, prospective and **retrospective** sampling, **overdispersion**, **longitudinal data**, and the like.

The Methods of Maximum Likelihood (ML) and Least Squares (LS)

We assume the **random variate y** has a density function $f(\mathbf{y};\theta)$. The function f is completely specified up to the unknown parameter, θ, assumed for simplicity to be a scalar. The maximum likelihood method (ML) consists of estimating the unknown parameter θ on the basis of the response, $\mathbf{y}$, by the value $\hat{\theta}(\mathbf{y})$ which, for the fixed $\mathbf{y}$, maximizes $f(\mathbf{y};\theta)$ for all variations of θ. Granting regularity conditions the score function (*see* **Likelihood**) is defined as $\partial \log f/\partial\theta$ and the ML estimate, $\hat{\theta}(\mathbf{y})$, is given by solving for θ the *ML equation*

$$\frac{\partial \log f(\mathbf{y};\theta)}{\partial\theta} = 0. \tag{1}$$

If the random variate $\mathbf{y}$ consists of n components $\mathbf{y} = (y_1, \ldots, y_n)$, then under very general conditions, as $n \to \infty$, i.e. asymptotically (*see* **Large-sample Theory**) the solution of the ML equation (1) tends to be an unbiased minimum variance estimate of θ, with the variance $\to 0$; Fisher [13]. That simply means that among the **consistent** estimates of θ, the ML estimate has an asymptotically smallest variance. This is the *strong* property of ML estimation. But now suppose the density f of the random variate $\mathbf{y}$ is specified by two parameters, θ_1 and θ_2, $f = f(\mathbf{y};\theta_1,\theta_2)$, θ_1 being the parameter of interest and θ_2 the nuisance parameter. Now given the response $\mathbf{y}$, what is the ML estimate of θ_1 ignoring θ_2? This question has no answer. Suppose we estimate both parameters θ_1 and θ_2 by the ML method; $(\hat{\theta}_1, \hat{\theta}_2)$. That is, the density $f(\mathbf{y};\theta_1,\theta_2)$ is maximized for the joint variation of (θ_1,θ_2) at $\theta_1 = \hat{\theta}_1(\mathbf{y})$ and $\theta_2 = \hat{\theta}_2(\mathbf{y})$. If now, as before, $\mathbf{y} = (y_1, \ldots, y_n)$ and if the dimensionality of the nuisance parameter θ_2 increases with n, then $\hat{\theta}_1(\mathbf{y})$ may tend to a false value of θ_1 as $n \to \infty$ (*see* **Convergence in Distribution and in Probability**). Suppose, for example, $y_i = (y_{i1}, y_{i2})$, $i = 1, \ldots, n$, and that all ys are distributed independently and normally with the means $\mathrm{E}(y_{i1}) = \mathrm{E}(y_{i2}) = \theta_{2i}$, $i = 1, \ldots, n$, $\theta_2 = (\theta_{21}, \ldots, \theta_{2n})$ and with a common variance θ_1. In this case the ML estimate

$$\hat{\theta}_1(\mathbf{y}) = \frac{1}{4}\frac{\sum_1^n (y_{i1} - y_{i2})^2}{n} \longrightarrow \frac{1}{2}\theta_1.$$

That is, the ML estimate $\hat{\theta}_1$ is inconsistent [40]. This then is a major weakness of ML estimation.

As noted above, the ML method of estimation presupposes a fully parametric model $f(\mathbf{y};\theta)$. However, the least squares (LS) method assumes only a semiparametric model: the mean values of the responses are modeled as functions of the parameter of interest θ. Let $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)$ be the vector of independent random variates with mean values $\mathrm{E}(y_i) = \alpha_i(\theta)$ and variances $\mathrm{E}(y_i - \alpha_i)^2 = \sigma_i^2$, $i = 1, \ldots, n$, where the α_i are some specified functions of the parameter $\theta;\theta$, as before, is assumed a scalar. The

variances σ_i^2 are assumed to be known constants. As usual, the value of the parameter θ is unknown. The LS method consists of estimating θ, on the basis of the response, $\mathbf{y}$, by the value $\hat{\theta}(\mathbf{y})$, which minimizes for the fixed $\mathbf{y}$, $\sum_1^n[(y_i - \alpha_i)^2/\sigma_i^2]$ for all variations of θ. Thus the LS estimate, $\hat{\theta}(\mathbf{y})$, is obtained by solving for θ the *LS equation*

$$\sum_{i=1}^{n}[y_i - \alpha_i(\theta)]\frac{\partial\alpha_i(\theta)/\partial\theta}{\sigma_i^2} = 0. \tag{2}$$

For linear functions α_i, this method of estimation, as previously indicated, was proposed by Legendre [28]. The estimate $\hat{\theta}(\mathbf{y})$, so obtained, was shown by Gauss [16] to have minimum variance in the class of all linear **unbiased** estimates of θ. Usually this is known as the Gauss–Markoff theorem. Thus the Gauss–Markoff theorem gives a *strong* finite sample property of LS estimation. Even if the functions α_i are not linear, granting some regularity conditions, the LS estimate, though no longer unbiased, is *consistent*. The LS method, however, breaks down when the variances σ_i^2 are not "known constants" but are "known (specified) functions" of θ; $\mathrm{E}(y_i - \alpha_i)^2 = \sigma_i^2(\theta)$, $i = 1, \ldots, n$. Now the *LS equation* (2) is to be replaced by

$$\left\{\sum_{i=1}^{n}[y_i - \alpha_i(\theta)]\frac{\partial\alpha_i(\theta)/\partial\theta}{\sigma_i^2(\theta)}\right\} + \left\{2\sum_{i=1}^{n}[y_i - \alpha_i(\theta)]^2 \times \frac{1}{\sigma_i^3(\theta)} \cdot \frac{\partial\sigma_i(\theta)}{\partial\theta}\right\} = 0. \tag{3}$$

Here the second term on the left-hand side of (3), for large samples, is of the order $\sum_1^n \partial \log \sigma_i(\theta)/\partial\theta$. Hence it is easy to see that the solution of (3), unlike that of (2), would provide an *inconsistent* estimate of θ. This then is a major *weakness* of the LS method.

The Method of Estimating Functions (EF)

The two methods of estimation, ML and LS, are characterized by the "estimating equations", (1) and (2), respectively. More generally, we define an *estimating function* as a function of the random variate $\mathbf{y}$ and the parameter of interest θ, $g(\mathbf{y}, \theta)$ say. For every observation $\mathbf{y}$, the solution of the estimating equation $g(\mathbf{y}, \theta) = 0$, namely $\hat{\theta}$, provides an estimate of θ. Thus in (1) and (2) the estimating functions are

$$g_1(\mathbf{y}, \theta) = \frac{\partial \log f(\mathbf{y}; \theta)}{\partial\theta} \quad \text{and}$$

$$g_2(\mathbf{y}, \theta) = \sum_{i=1}^{n}[y_i - \alpha_i(\theta)]\frac{\partial\alpha_i(\theta)}{\partial\theta}\frac{1}{\sigma_i^2}, \tag{4}$$

respectively. Unlike the two methods of estimation, ML and LS, which emphasize the properties of the "estimate", i.e. the solution of the equation $g(\mathbf{y}, \theta) = 0$, the EF method emphasizes the properties of the estimating function g itself. For instance, it is called unbiased if, for all θ, the expectation

$$\mathrm{E}g(\mathbf{y}, \theta) = 0. \tag{5}$$

Thus, although the estimates (solutions) $\hat{\theta}$ obtained from (1) and (2) are generally biased, the corresponding estimating functions g_1 and g_2 in (4) are unbiased: $\mathrm{E}(g_1) = 0$ and $\mathrm{E}(g_2) = 0$. Interestingly, the inconsistency of the solution of the estimating equation, (3), is due to the fact that the corresponding estimating function, namely

$$g_3 = g_2 + \left\{2\sum_{i=1}^{n}(y_i - \alpha_i)^2\frac{1}{\sigma_i^3}\frac{\partial\sigma_i}{\partial\theta}\right\}, \tag{6}$$

unlike g_2, is *not* unbiased.

In EF methodology we deal with EFs g which are unbiased, i.e. they satisfy the property (5). To compare the efficiencies of various unbiased estimating functions g, the *variance* of g, $\mathrm{E}(g^2)$, is not a useful criterion, for trivially $\mathrm{E}(g^2) \equiv 0$ for the unbiased estimating function g which is identically 0 for all values of $\mathbf{y}$ and θ. This difficulty is overcome by using the standardized version of g, namely

$$g_s = \frac{g}{\mathrm{E}\left(\frac{\partial g}{\partial\theta}\right)}. \tag{7}$$

Note that the standardized versions of two unbiased estimating functions, namely g and $k(\theta)g$, $k(\theta)$ being a constant depending on θ, are *identical*. This is necessary for the estimating functions g and $k(\theta)g$ to have the same inferential content. Although the standardization given by (7) is somewhat arbitrary, it has proved to be very versatile, with applications in various fields of statistics. This is illustrated below. Initially, the standardization (7) provides a definition of an *optimal* estimating function.

Definition. For a given model, let $\mathcal{G}$ be an arbitrary class of unbiased estimating functions g, $\mathcal{G} = \{g\}$. In $\mathcal{G}$, the estimating function g^* is said to be optimal if for any other $g \in \mathcal{G}$,

$$\mathrm{E}(g_{\mathrm{s}}^*)^2 \leq \mathrm{E}(g_{\mathrm{s}})^2, \tag{8}$$

for all θ.

Now consider the parametric model $f(\mathbf{y};\theta)$ mentioned previously in relation to ML estimation. Furthermore, in the above definition, let $\mathcal{G}_0$ be the class of all unbiased estimating functions satisfying some regularity conditions. Then in $\mathcal{G}_0$ the optimal estimating function g^* is given by the score function (SF)

$$g^* = \frac{\partial \log f(\mathbf{y};\theta)}{\partial \theta} \tag{9}$$

(see [17]). This is a finite sample optimality of ML estimation in contrast to its asymptotic property [13] mentioned earlier. Actually, the latter follows from the former. Again, as before, we consider the nuisance parameter situation. In the model $f(\mathbf{y};\theta), \theta = (\theta_1, \theta_2)$, let θ_1 be the parameter of interest and θ_2 the nuisance parameter. To estimate θ_1, ignoring θ_2, we consider the class $\mathcal{G}_{\mathrm{c}}$ of unbiased estimating functions g which depend on θ only through θ_1. Now suppose the model f admits a complete **sufficient statistic** $\boldsymbol{t}$ (independent of θ_1) for θ_2 and $f(\mathbf{y}|t;\theta_1)$ denotes the density of $\mathbf{y}$ conditional on t. Then granting some regularity conditions, the optimal estimating function in $\mathcal{G}_{\mathrm{c}}$ is given by the conditional score function

$$g^* = \frac{\partial \log f(\mathbf{y}|t;\theta_1)}{\partial \theta_1}; \tag{10}$$

the definition of "optimality" used being the same as before [18]. The solution of the estimating equation $g^* = 0$, in (10), unlike that of the ML equations, (1), is a consistent estimate generally for problems of the Neyman–Scott [40] type. Thus, EF theory corrects the ML estimation for its major weakness. A similar correction to LS estimation is provided by the EF theory. This is shown in what follows.

As previously indicated, LS estimation presupposes a semiparametric model. We have a random vector of observations, $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)$, with independent components. The mean values, $\mathrm{E}(y_i) = \alpha_i(\theta)$, and the variances, $\mathrm{E}(y_i - \alpha_i(\theta))^2 = \sigma_i^2(\theta), i = 1, \ldots, n; \alpha_i$, and σ_i are some specified functions of the parameter of interest θ. It was noted earlier that LS estimation works well when α_i are linear functions of θ and σ_i are independent of θ. In this case the estimation derived from the LS equation (2) is supported by the Gauss–Markoff theorem. However, when σ_i depends on θ, the LS equation (3) leads to an inconsistent estimate. This inconsistency is due to the fact that the EF g_3 in (6) is *not* unbiased. To correct this situation the EF theory here starts with the *elementary* unbiased EFs $h_i = y_i - \alpha_i(\theta)$ and $\mathrm{E}(h_i) = 0, i = 1, \ldots, n$. Now the linear estimates of the Gauss–Markoff theorem are replaced by the class $\mathcal{G}_l$ of *linear* unbiased EFs,

$$g = \sum_{i=1}^{n} h_i a_i, \tag{11}$$

where a_i can be arbitrary functions of θ. Note that g is linear in h_i and is unbiased, $\mathrm{E}(g) = 0$. Furthermore, it can be shown that in this class $\mathcal{G}_l$ of linear unbiased EFs the "optimum" g^* satisfying the inequality (8) is given by

$$g^* = \sum_{i=1}^{n} [y_i - \alpha_i(\theta)] \frac{\partial \alpha_i(\theta)/\partial \theta}{\sigma_i^2(\theta)}. \tag{12}$$

Note that when the variances σ_i^2 are constants independent of θ, the estimating function g_2 in (4) provided by the LS method is identical to g^* in (12). However, when σ_i^2 depends on $\theta, \sigma_i^2 = \sigma_i^2(\theta)$, g^* in (12) is different from the EF g_3 in (6), provided by the LS method. Not only does the equation $g^* = 0$, unlike the LS equation $g_3 = 0$, provide a consistent solution, but in fact $g^* = 0$ is the ML equation (g^* being the score function) for the **exponential family** of distributions [3, 45]. Actually, since Wedderburn [45], $g^* = 0$ is called the **quasi-likelihood** equation [34] (*see* **Generalized Linear Model**). The optimality of the estimating function g^* in (12) was established in a wider setting of stochastic processes by Godambe [19]; following that reference we call g^* the *quasi-score function*.

The quasi-score function is a *synthesis*, provided by the EF theory, of the two traditionally distinct methods of estimation namely ML and LS. For, note that the EFs g^* in (9), (10), and (12) all satisfy the same criterion of optimality, namely (8); only the competing classes in the three cases $\mathcal{G}_0, \mathcal{G}_{\mathrm{c}}$, and $\mathcal{G}_l$, respectively, are different. These classes are derived from the underlying models. The quasi-score function, that is the optimum EF g^* in (12), has a wider domain of application than either the ML or

the LS methods. The domain of application of g^* is vastly enhanced further by letting the elementary estimating functions h_i in (11) be arbitrary functions of $\mathbf{y}$ and θ, which are *conditionally* unbiased, i.e. unbiased conditional on some partition (σ-field) of the sample space. For, in that case the constants $a(\theta)$ in (11) can be replaced by any functions of $\mathbf{y}$ and θ which are measurable with respect to the σ-field generated by the partition. This provides an enlarged class $\mathcal{G}_{\text{lc}}, \mathcal{G}_{\text{lc}} \supset \mathcal{G}_{\text{l}}$ of the competing estimating functions. The optimal estimating function g^* in $\mathcal{G}_{\text{lc}}$, i.e. the one satisfying condition (8), was obtained by Godambe & Thompson [23]. This covers areas of **stochastic processes** underlying many biostatistical applications.

From the above discussion it should be clear that the optimal EF, g^*, not only provides a point estimate through the equation $g^* = 0$, but it also provides a substitute for a score function (1), in a semiparametric model. Let g be any unbiased EF, and furthermore let the semiparametric model be a union of families of parametric distributions. Then, very generally, we have:

1. $\mathrm{E}(g^* - \mathrm{SF})^2 \leq \mathrm{E}(g - \mathrm{SF})^2$.
2. **correlation** $(g^*, \mathrm{SF}) \geq \mathrm{correlation}(g, \mathrm{SF})$, where the score function, SF, and the expectation, E, correspond to the underlying parametric family of distributions.
3. Moreover, the **confidence intervals** obtained by inverting the distribution of g^* are asymptotically shorter than the corresponding ones based on g [21].

Now for a parametric family of distributions the optimal EF, g^*, is given by the score function as in (9). Also, according to the just stated property, 3, the confidence intervals based on the score function are asymptotically shortest [48].

One important implication of point 3 above, for all statistical (including biostatistical) practice, is that the confidence intervals based on the inversion of the distribution of the optimal EF g^* are preferable to those based on any (unbiased minimum variance) estimate. Even operationally, the distribution of the estimate often is far less tractable than that of an estimating function consisting of independent components. This is often the case in biostatistical applications, as will be seen subsequently.

Case–Control Studies

Case–control or retrospective studies are of great importance in both biostatistics and the social sciences where they are referred to as choice-based sampling studies. In such a study comparisons are made between individuals who have a particular disease or condition, the cases, and individuals who do not have the disease, the controls. In the simplest case, where a single binary risk factor or exposure variable is of interest, one might sample cases from a population having the disease and also controls from the disease-free population. One then ascertains retrospectively whether or not the sampled individuals have been exposed to the risk factor. In the simplest situation this reduces to a comparison of two **binomial distributions** with the **odds ratio** $\theta = [p_1(1 - p_2)/p_2(1 - p_1)]$ the parameter of interest, this being a measure of association between disease and exposure. Here $p_i, i = 1, 2$, is the probability of exposure for cases and controls, respectively. The nuisance parameter complementary to θ may be taken as p_2 and one is interested in an optimal estimating function for θ free of the nuisance parameter p_2. If $\mathbf{y} = (y_1, y_2)$ are the observed number of exposed individuals in the samples of cases and controls, respectively, then $t = y_1 + y_2$ is a complete sufficient statistic for the nuisance parameter p_2. Hence by a result of Godambe [18], mentioned in (10) above, the optimal estimating function for θ is the conditional score $\partial \log f(\mathbf{y}|t;\theta)/\partial\theta$. In this case

$$\frac{\partial \log f(\mathbf{y}|t;\theta)}{\partial\theta} = y_1 - \mathrm{E}(y_1|t;\theta), \qquad (13)$$

which is free of p_2 and clearly unbiased. We note also that while conditioning is important in eliminating the nuisance parameter, the result of Godambe [18] implies unconditional global optimality for this estimating function. We note also that this argument in terms of estimating functions is related to Fisher's use of conditional likelihood for the **2 × 2 table** (*see* **Fisher's Exact Test**).

In practice, there will be a number of **confounding** variables or factors that may conceal or exaggerate the true effect of the risk factor of immediate interest. A common strategy is to stratify the cases and controls into a number of strata $i, i = 1, \ldots, N$, say, on the basis of such confounding variables (*see* **Stratification**). If the odds ratios $\theta = [p_{i1}(1 - p_{i2})/p_{i2}(1 - p_{i1})], i = 1, \ldots, N$, are the same for each stratum,

then we have N 2×2 tables with p_{i1} and p_{i2} being the exposure probabilities for cases and controls in the ith stratum. Here we have a single parameter θ of interest and N nuisance parameters $p_{i2}, i = 1, \ldots, N$. The difficulties involved in maximum likelihood estimation with many nuisance parameters, as mentioned earlier, are well known since Neyman & Scott [40]. This is a classic Neyman–Scott type situation. However, the estimating function approach avoids the inconsistency of the maximum likelihood estimate of θ by conditioning on $t_i = y_{i1} + y_{i2}$, where y_{i1} and y_{i2} are the number of exposed individuals for the cases and controls, respectively, in the ith stratum. The conditioning statistic, t_i, is again complete and sufficient for p_{i2} and the optimal estimating function for θ is obtained by summing terms like (13) for each stratum. Yanagimoto [51] considers combinations of unbiased estimating functions from different strata which, although suboptimal, may be simpler computationally than (13). Yanagimoto suggests the following unbiased estimating equation for stratum i:

$$g_i(\mathbf{y}_i; \theta) = y_{i1}(n_{i2} - y_{i2}) - \theta y_{i2}(n_{i1} - y_{i1}),$$

where $\mathbf{y}_i = (y_{i1}, y_{i2})$ and n_{i1} and n_{i2} are the sample sizes for cases and controls, respectively. The optimal weighted combination of these, minimizing Godambe's criterion (8), is

$$g(\mathbf{y}; \theta) = \sum_{i=1}^{N} w_i g_i(\mathbf{y}_i; \theta),$$

where $w_i = w_i(\theta) = \mathrm{E}(\partial g_i / \partial \theta)\mathrm{var}^{-1}(g_i)$. For $\theta = 1$, Yanagimoto finds $w_i(1) = (n_i + m_i)^{-1}$, yielding the celebrated **Mantel–Haenszel** estimator [32].

The optimality of the conditional score function exemplified above for the binomial distribution applies more generally to canonical exponential families. For example, mortality rates are often modeled under **Poisson** assumptions, and the parameter of interest, the standardized mortality rate θ, may be taken as a ratio of Poisson means for exposed vs. nonexposed. Again, $t_i = y_{i1} + y_{i2}$, the total number of deaths in each stratum, is a complete sufficient statistic for the nuisance stratum parameter and derivation of the optimal estimating function parallels the binomial case. Computationally the situation is simpler in that the conditional score involves the binomial distribution of y_{i1}, given $y_{i1} + y_{i2}$, while (13) above involves the noncentral **hypergeometric**.

Prospective and Retrospective Sampling

Frequently the analysis of case–control studies proceeds by ignoring retrospective sampling and assuming that the data arose prospectively. Prentice & Pyke [42] provided one justification for this by showing that the resulting estimators of the **logistic regression** coefficients are consistent and that the usual standard errors are asymptotically correct. In this respect the advantages of the estimating functions approach have been recently demonstrated by Carroll et al. [6] as follows.

Simple ideas from estimating function theory may be used to generalize the above result of Prentice & Pyke to a variety of other analyses and sampling schemes. For example, the logistic model may be replaced by **robust** models; **measurement error** models may be accommodated; **missing data** patterns of general types, as well as stratified studies, are encompassed. The multiplicative model of Weinberg & Wacholder [46], and other models not necessarily of the logistic form, are also included. In the more general situation just mentioned a somewhat weaker result than that of Prentice & Pyke holds. If the case–control sampling scheme is ignored and asymptotic standard errors are derived as if the study were prospective, then the standard errors are, in general, at worst asymptotically conservative. However, the asymptotic theory based on estimating functions enables one to identify a simple sufficient condition which ensures that prospectively derived standard errors are, in fact, asymptotically correct. This condition can be shown to apply to a variety of examples.

The simple case of the classical logistic model exemplifies the essence of the argument. The estimating equations for prospective sampling are given by

$$\mathbf{0} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \end{pmatrix} [D_i - H(\theta_0^* + \theta_1 X_i)], \qquad (14)$$

where D_i is the binary variable for the ith individual, X_i the covariate, and $H(\cdot)$ the usual logistic distribution function. This may be written as

$$\mathbf{0} = \sum_{i=1}^{n} \psi_i(D_i, X_i, \boldsymbol{\theta}^*) = \boldsymbol{\Psi}, \qquad (15)$$

where $\boldsymbol{\theta}^* = (\theta_0^*, \theta_1)'$ is the unknown parameter. Prospectively, the right-hand side of (15) is an unbiased estimating function. Furthermore,

$$\mathrm{E}\psi_i(D_i, X_i, \boldsymbol{\theta}^*) = \mathbf{0} \tag{16}$$

for each i, so that each component is also an unbiased estimating function in the prospective sampling scheme. The asymptotic standard error of the solution to (15) is, by the usual Taylor series argument,

$$\mathbf{B}^{-1}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}^{-1}(\boldsymbol{\theta}^*),$$

where

$$\mathbf{B}(\boldsymbol{\theta}^*) = n^{-1}\mathrm{E}\left(\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\theta}^*}\right)$$

and

$$\mathbf{A}(\boldsymbol{\theta}^*) = n^{-1}\mathrm{cov}(\boldsymbol{\Psi}).$$

Moreover, because of (16):

$$\begin{aligned}\mathbf{A}(\boldsymbol{\theta}^*) &= n^{-1}\sum_{i=1}^{n}\mathrm{E}[\psi_i(D_i, X_i, \boldsymbol{\theta}^*)\psi_i'(D_i, X_i, \boldsymbol{\theta}^*)] \\ &= \mathbf{C}(\boldsymbol{\theta}^*), \text{ say.}\end{aligned}$$

Turning now to the retrospective sampling scheme, with $\boldsymbol{\theta}^*$ replaced by $\boldsymbol{\theta} = (\theta_0, \theta_1)'$, where $\theta_0 = \theta_0^* + \log(n_1/n_0) - \log[\Pr(D=1)/\Pr(D=0)]$, where n_i is the sample size for individuals with $D = i, i = 0$ or 1, and $\Pr(D=1)$ is the unknown prospective probability of success. Consider (15) with $\boldsymbol{\theta}^*$ replaced by $\boldsymbol{\theta}$. While $\boldsymbol{\Psi}$ remains unbiased [42], the component estimating functions $\psi_i(D_i, X_i, \boldsymbol{\theta})$ are, in general, not unbiased. Instead of $\mathbf{A}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})$, we now have, by an elementary calculation,

$$\mathbf{A}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta}) - \mathbf{D}(\boldsymbol{\theta}),$$

with

$$\mathbf{D}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathrm{E}[\psi_i(D_i, X_i, \boldsymbol{\theta})]\mathrm{E}[\psi_i'(D_i, X_i, \boldsymbol{\theta})]$$

a positive semi-definite matrix. The positive semi-definiteness of $\mathbf{D}(\boldsymbol{\theta})$ implies that the use of the prospective formula $\mathbf{B}^{-1}(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\mathbf{B}^{-1}(\boldsymbol{\theta})$ will, in general, produce inflated standard errors when applied to data collected retrospectively.

This simple argument can be generalized to a wide variety of models and measurement error schemes. A key assumption is that a general version of (15), which is prospectively unbiased, is also retrospectively unbiased. As an informal argument justifying this assertion note that prospective unbiasedness implies $\mathrm{E}\{\psi(D, \mathbf{X}, \boldsymbol{\theta})|\mathbf{X}\} = 0$, for the classical model, i.e.

$$\sum_{d=0}^{1}\psi(d, x, \theta)H^d(x, \theta)[1 - H(x, \theta)]^{1-d} = 0. \tag{17}$$

Similarly, conditioning on the disease outcome variable D, retrospective unbiasedness implies $\mathrm{E}\{\psi(D, \mathbf{X}, \boldsymbol{\theta})|D\} = 0$. Suppose we define

$$k_d = \int \psi(d, x, \theta)H^d(x, \theta)[1 - H(x, \theta)]^{1-d}\tau(x)\,\mathrm{d}x,$$

where $\tau(x)$ is the marginal density of x induced by the case–control sampling scheme. Then it follows from (17) that $k_0 + k_1 = 0$, implying retrospective unbiasedness. Note that k_1 equals the first column of $-\mathrm{E}(\partial\Psi/\partial\theta)$, which is the condition that prospectively derived standard errors are retrospectively asymptotically correct. These results for the classical model can be generalized to a variety of more complex schemes.

Many other applications of estimating functions in biostatistics are similar to this one in that the primary focus is on the generation of point estimates together with asymptotic standard errors. However, a consequence of the optimality of an estimating function is that confidence intervals based directly on it are asymptotically the shortest, as pointed out earlier. Hence, interval estimation based on inversion of appropriately standardized optimal estimating functions should be superior to those based on estimates and standard errors, *even if* these latter are obtained from optimal estimating functions. Special cases of this phenomenon have been demonstrated, e.g. Boos [2] or Sprott & Viveros [43].

In the works of Prentice & Pyke [42] and Carroll et al. [6] discussed above, and others, the parameters estimated belong to a *model* and the sample is supposed to be drawn from the *hypothetical population* generated by the model. However, generally in biostatistical applications the "sample" is drawn from a finite *survey population* consisting of individuals or units. This survey population is supposed to be a random sample from the hypothetical population mentioned above. In many situations the statistician is interested not only in estimating the parameters of

the hypothetical population (model), but also in estimating the survey population parameter *related* to or *induced* by the model [22]. For instance, if the model parameter β represents the regression coefficient in the hypothetical population, then the induced survey population parameter β_s is the regression coefficient in the finite survey population. The estimating function theory provides optimal estimation of both parameters β and β_s *jointly*, on the basis of a sample drawn from the survey population with a specified sampling design. Utilizing this theory, Godambe & Vijayan [24] have obtained, among other things, optimal estimation for the logistic model based on a "sample" drawn with response-dependent retrospective sampling from a survey population. These results, as a special case, become identical to those of the earlier authors when the "sample" is the same size as the "survey population". The authors also provide a proof of the often conjectured large-sample equality of "prospective" and "retrospective" scores for the parameter of interest.

Modeling Overdispersed Data

Data involving counts or proportions occur frequently in medical applications. For example, disease incidence data or mortality statistics usually involve data in the form of counts. In **teratology** the proportion of malformed fetuses in litters, for which the mother has been subjected to a given dose of teratogen, provides data used to model the probability of malformation as a function of dose. The simplest models for such data, namely the Poisson and binomial, respectively, are usually inadequate in that they predict less variation than is exhibited in such data. One source of such **overdispersion** or extravariation is lack of independence. For example, in spatial mortality statistics, counts in contiguous areas will tend to be correlated (*see* **Geographic Epidemiology**), while in teratology the binary responses of members of the same litter tend to be correlated, the so-called litter effect (*see* **Preclinical Treatment Evaluation**). Estimation of regression parameters in the mean of the response variable often will be of primary interest, although in studies of disease association within families, correlation or association parameters will be of substantive interest. In the former case the parameters in the link function for the mean will be the parameters of interest, whereas correlation or association effects will be described by nuisance parameters. Optimal estimating functions for the parameters of interest in the presence of nuisance parameters are therefore of some interest. However, estimation of the nuisance parameters is important in that standard errors and the validity of associated tests and confidence intervals for the parameters of interest will depend on these so-called nuisance parameters.

Consider the generalized linear models with, for example, Poisson errors for counts and binomial errors for proportions. Here often the systematic part of the model is assumed to be "correct" with an appropriate choice of a link function and covariates. Yet the variance of replicate or near replicate observations is considerably greater than that suggested by such simple exponential family models. As Cox [7] points out, maximum likelihood estimates of regression parameters will not be seriously in error if one ignores this so-called overdispersion. However, associated standard errors, and hence tests and confidence intervals based on them, which ignore such overdispersion can be very misleading.

Overdispersion is frequently modeled using a fully parametric mixed Poisson or mixed binomial model. In this approach the means associated with each Poisson count or binomial proportion are assumed to be random variables with specified parametric distributions. Estimation of the parameters in these mixed distributions proceeds by full maximum likelihood. Manton et al. [33], Hinde [26], and Lawless [27] for count data and Crowder [8], Williams [49, 50] for proportions, provide examples of this. Manton et al., for example, deal with an application to modeling spatial variability in lung cancer mortality rates. The mixing distributions most frequently used (i.e. **gamma** for the Poisson (*see* **Contagious Distributions**) and beta for the binomial (*see* **Beta-binomial Distribution**)) are selected for their mathematical tractability rather than scientific plausibility. Sometimes, if the actual mechanism leading to the overdispersion is known, then such fully parametric modeling may be entirely convincing, but in practice this is the exception rather than the rule.

To avoid this criticism, models that specify only second-order properties of the mixing distribution, or equivalently second-order properties of the observed counts or proportions, may be formulated. This leads to a more robust approach based on quasi-likelihood. Many *ad hoc* techniques used historically in the analysis of counts or proportions may now be regarded

as instances of quasi-likelihood. For example, the use of a multiplicative variance inflation factor in probit analysis [[12], Chapter 4] (*see* **Quantal Response Models**), or the treatment of continuous responses by Fisher [14] with variance proportional to the mean as essentially having a Poisson likelihood, can be regarded as primitive instances of quasi-likelihood.

Estimation of Dispersion Parameters

Major interest usually focuses on the regression parameters, the dispersion parameters being of importance only to the extent that they reflect the precision of the regression parameter estimates. For this reason, relatively *ad hoc* methods are used to estimate the dispersion parameters. Sometimes, however, the dispersion parameters are of interest in their own right. Furthermore, in **multivariate multiple regression** problems, association or correlation parameters are also of interest, e.g. in **segregation analysis** in genetics; see Whittemore & Gong [47] or Zhao [52]. Also, Zhao & Prentice [53] note the importance of simultaneous estimation of mean, variance, and covariance parameters in a variety of biostatistical applications. The theory of optimal estimating functions treats all parameters on the same logical footing and shows the way to a formal theory of joint estimation of both regression and dispersion parameters. Less formal approaches have been developed by Nelder & Pregibon [39] and Davidian & Carroll [9]. Wedderburn's quasi-likelihood was originally designed for parameters in the link function. Nelder & Pregibon [39] attempt to obtain a function of both mean and dispersion parameters which has the properties of a log likelihood for both sets of parameters. However, consistent estimation of parameters in the variance function is not achieved in general owing to the lack of the unbiasedness property of the underlying estimating functions. Clearly, then, their extended quasi-score cannot be identified with an optimal estimating function in the sense of this article. The same terminology as used by Nelder & Pregibon is adopted by Godambe & Thompson [23], but in their case the identification of an extended quasi-score with an optimal estimating function is preserved – at the expense of specification of higher-order moments.

More formally, let us assume that we have independent observations, $y_i, i = 1, \ldots, n$, with

$$\mathrm{E}(y_i) = \mu_i, \qquad \mu_i = \mu_i(\beta),$$

$$\mathrm{var}(y_i) = V(\mu_i; \lambda).$$

Standard multiplicative overdispersion, or more complicated structural variance parameters, as in McCullagh & Nelder [[34], Chapter 10], can be accommodated in this formulation. The common approach to this is to note that if the variance parameter λ is known, then the Wedderburn equations,

$$U(y; \beta, \lambda) = \sum \frac{y_i - \mu_i}{V(\mu_i; \lambda)} \frac{\partial \mu_i}{\partial \beta} = 0, \qquad (18)$$

are optimal linear estimating equations for β. A variety of methods for estimating the variance parameters λ have been proposed to supplement (18). Examples of such methods are:

1. moment methods (e.g. [4, 27], and [35]);
2. extended quasi-likelihood [39]; and
3. **pseudo-likelihood** [9].

These methods do not lead jointly to optimal estimating equations in the sense of this article. However, methods based strictly on optimal estimating functions may well be less robust than these methods.

To review them briefly, method (1) involves supplementing (18) with the moment equation for λ, given by

$$\sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)^2}{V(\mu_i; \lambda)} - 1 \right] = 0, \qquad (19)$$

or the bias-corrected version,

$$\sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)^2}{V(\mu_i; \lambda)} - \frac{n-p}{n} \right] = 0, \qquad (20)$$

where p is the dimension of β. Then (19) or (20) and (18) would be solved jointly, possibly in a doubly iterative fashion. The above is for a scalar λ. In the case where λ is a vector, additional quadratic forms could be equated to their expected values to obtain additional equations, as in McCullagh & Nelder [34, Chapter 10].

For method (2), the extended quasi-likelihood of Nelder & Pregibon is given by

$$Q^{+}(\beta, \lambda) = -\frac{1}{2} \sum_{i=1}^{n} D(y_i, \mu_i, \lambda) - \frac{1}{2} \sum_{i=1}^{n} \log[2\pi V(y_i; \lambda)], \qquad (21)$$

where

$$D(y, \mu, \lambda) = -2 \int_y^{\mu} \frac{y - t}{V(t; \lambda)} \, \mathrm{d}t.$$

Q^+ is then maximized jointly in β and λ.

Finally, in the pseudo-likelihood method of Davidian & Carroll, one fixes β at a preliminary estimate, $\beta = \hat{\beta}$ say, and maximizes the normal theory likelihood in λ:

$$l_{pl}(\hat{\beta}, \lambda) = -\sum_{i=1}^{n} \log V(\mu_i; \lambda) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i; \lambda)}. \tag{22}$$

An updated estimate of β can be obtained by successive iterations.

The methods of pseudo-likelihood and extended quasi-likelihood correspond to taking the Pearson chi-square statistic and the deviance, respectively, as response variables in the analysis of dispersion (*see* **Chi-square Tests**). Davidian & Carroll [9] prefer pseudo-likelihood, based on asymptotic considerations. However, recent work by Nelder & Lee [38] suggests that extended quasi-likelihood performs better in finite sampling. In the analysis of Taguchi-type experiments, for example, dispersion parameters must frequently be estimated on the basis of rather limited data, so that finite sample performance may be of some practical importance. A similar concern applies to many biostatistical applications.

Godambe & Thompson [23] obtain jointly optimal estimating equations as follows. Let γ_{1i} and γ_{2i}, respectively, denote the **skewness** and **kurtosis** of the ith observation. Also, let $\Delta_i = \gamma_{2i} + 2 - \gamma_{1i}^2$. Define the elementary orthogonal estimating functions for $i = 1, \ldots, n$ as

$$h_{1i} = y_i - \mu_i, h_{2i} = (y_i - \mu_i)^2 - V(\mu_i; \lambda) - \gamma_{1i}[V(\mu_i; \lambda)]^{1/2}(y_i - \mu_i).$$

Then, the optimal combination of h_{1i} and h_{2i} is given by Godambe & Thompson [23] as

$$\sum_{i=1}^{n} \left[\frac{h_{1i} \partial \mu_i / \partial \beta_j}{V(\mu_i; \lambda)} - h_{2i} \times \frac{[V(\mu_i; \lambda)^{1/2} \gamma_{1i} - \partial V / \partial \mu_i] \partial \mu_i / \partial \beta_j}{V(\mu_i; \lambda) \Delta_i} \right] = 0, \tag{23}$$

$j = 1, \ldots, p$, and

$$\sum_{i=1}^{n} h_{2i} \frac{\partial V / \partial \lambda_j}{V^2(\mu_i; \lambda) \Delta_i} = 0, \tag{24}$$

$j = 1, \ldots, k$, where k is the dimension of λ. We remark that within the exponential family, the second term on the left-hand side of (23) vanishes. Thus, the optimum linear and quadratic equations for β coincide in this case.

Godambe & Thompson [23] refer to these as extended quasi-likelihood equations. They are clearly not, however, the same as the equations derived from (21). The adjective "extended" is unnecessary in a sense here, as pointed out by Heyde [25], because if one identifies the quasi-score with the optimal estimating function according to criterion (8), then these are simply the quasi-score equations derived from the class of combinations of the elementary functions h_{1i} and h_{2i}. The other methods are not quasi-score functions in this sense. Nelder [36] has explored the consequences of replacing h_{2i} above by an approximately unbiased elementary estimating function based on the deviance component, d_i, of the ith observation. Under certain approximating assumptions, he shows that deviance-based versions of (23) and (24) are the same as those obtained using (21). Nelder [37], in discussion of Desmond [10], points out that the use of deviance **residuals** absolves the statistician of the need to make assumptions about the form of third and fourth moments (at least to first order). He predicts that with further refinement this may lead to good approximations to jointly optimum estimating equations which do not depend on assumptions about higher-order cumulants.

Other Applications

The aim of this article has been twofold. First, to point out that the methodology of estimating functions produces a unification and extensions of approaches to statistical modeling and inference, both parametric and semi-parametric. Secondly, to illustrate, via selected examples, how this methodology applies in biostatistics. We have limited our discussion here to case–control studies, prospective and retrospective sampling, and overdispersion. There are many other applications not discussed here, in particular to **longitudinal studies** which we outline briefly

in the following paragraphs. Since this is discussed elsewhere, we simply draw the reader's attention to some recent pertinent work.

Many studies in the biomedical sciences involve data which are longitudinal in nature. In such studies a response variable, frequently a health indicator, is measured repeatedly in time for the individuals in the study, together with covariate vectors, including treatment covariates which may influence the response. Frequently, more than one response variable, which may be discrete or continuous or a combination of both, is measured leading to a **multivariate** response vector. Liang et al. [31] give several examples from public health research. Such data structures lead naturally to problems involving the estimation of mean and association parameters for discrete and continuous multivariate regression analysis. There has been a great deal of development in this area recently for non-Gaussian responses, especially as it applies to binary longitudinal data (*see* **Multivariate Methods for Binary Longitudinal Data**). Also, likelihood approaches to multivariate regression modeling of normal data, including longitudinal studies, have been much studied, e.g. Ware [44]. These methods exploit the tractability of the multivariate normal distribution to provide full maximum likelihood solutions. Likelihood analysis for non-Gaussian data is considerably less tractable. Prentice [41], in the context of binary data, emphasizes the difficulties in obtaining computationally simple likelihood analyses for such longitudinal data, except for special cases.

The estimating function methodology is particularly attractive here for two reasons. First, in lieu of specifying a complete likelihood we can simply model the first- and second-order moments of the response in terms of mean, dispersion, and association parameters i.e. construct a semiparametric model. Secondly, this method is more robust to incorrect model assumptions about higher-order moments, e.g. Liang & Rathouz [29]. We do not describe this work here, since an excellent exposition is given in Diggle et al. [11] (*see* **Generalized Estimating Equations**). However, for a more expansive discussion of the connections of GEE methods with the concepts of quasi-score and optimal estimating functions, see [10] and [30].

In the regression models treated in this article we have assumed that covariates are measured without error. However, measurement error in the covariates is often an important consideration in the biomedical sciences, since ignoring it can produce serious bias in the estimation of covariate effects (*see* **Errors in Variables**). Carroll et al. [5] consider a variety of methods for dealing with measurement error in nonlinear models, and in particular make use of estimating functions in their development.

Further examples of areas of application in biostatistics, e.g. **survival analysis**, may be found in Godambe [20], which is a good general reference for both theoretical developments and applied investigations. Additionally, [1] contains contributions from leading researchers in the field.

References

[1] Basawa, I., Godambe, V.P. & Taylor, R.L., eds (1997). *Selected Proceedings of a Symposium on Estimating Functions*, University of Georgia, March 1996. IMS Lecture Notes Series, to appear.

[2] Boos, D.D. (1980). A new method for constructing approximate confidence intervals from *M* estimates, *Journal of the American Statistical Association* **75**, 142–145.

[3] Bradley, E.L. (1973). The equivalence of maximum likelihood and weighted least squares in the exponential family, *Journal of the American Statistical Association* **68**, 199–200.

[4] Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models, *Journal of the American Statistical Association* **85**, 565–571.

[5] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

[6] Carroll, R.J., Wang, S. & Wang, C.Y. (1995). Prospective analysis of case–control studies, *Journal of the American Statistical Association* **90**, 157–169.

[7] Cox, D.R. (1983). Some remarks on overdispersion, *Biometrika* **70**, 269–274.

[8] Crowder, M.J. (1978). Beta-binomial ANOVA for proportions, *Applied Statistics* **27**, 34–37.

[9] Davidian, M. & Carroll, R.J. (1988). A note on extended quasi-likelihood, *Journal of the Royal Statistical Society, Series B* **50**, 74–82.

[10] Desmond, A.F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling (with discussion), *Journal of Statistical Planning and Inference* **60**, 77–121.

[11] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

[12] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, Cambridge.

[13] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–706.

[14] Fisher, R.A. (1949). A biological assay of tuberculins, *Biometrics* **5**, 300–316.

[15] Gauss, C.F. (1809). Theoria motus corporum coelestum, *Werke* **7**(.), Translated into English by C.H. Davis (1963). Dover, New York.

[16] Gauss, C.F. (1823). Combinationes erroribus minimis obnoxiae. Parts 1, 2 and Supplement, *Werke* **4**, 1–108.

[17] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation, *Annals of Mathematical Statistics* **31**, 1208–1212.

[18] Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations, *Biometrika* **63**, 277–284.

[19] Godambe, V.P. (1985). The foundations of finite sample estimation in stochastic processes, *Biometrika* **72**, 419–428.

[20] Godambe, V.P., ed. (1991). *Estimating Functions*. Oxford University Press, Oxford.

[21] Godambe, V.P. & Heyde, C.C. (1987). Quasi-likelihood and optimal estimation, *International Statistical Review* **55**, 231–244.

[22] Godambe, V.P. & Thompson, M.E. (1986). Parameters of superpopulation and survey populations: their relationships and estimation, *International Statistical Review* **54**, 127–138.

[23] Godambe, V.P. & Thompson, M.E. (1989). An extension of quasi-likelihood (with discussion), *Journal of Statistical Planning and Inference* **22**, 137–172.

[24] Godambe, V.P. & Vijayan, K. (1996). Optimal estimation for response-dependent retrospective sampling, *Journal of the American Statistical Association* **91**, 1724–1734.

[25] Heyde, C.C. (1989). In discussion of Godambe, V.P. and Thompson, M.E., *Journal of Statistical Planning and Inference* **22**, 137–172.

[26] Hinde, J. (1982). Compound Poisson regression models, in *Proceedings of the International Conference on Generalized Linear Models*, R. Gilchrist, ed. Springer-Verlag, Berlin, pp. 109–112.

[27] Lawless, J.F. (1987). Negative binomial and mixed Poisson regression, *Canadian Journal of Statistics* **15**, 209–225.

[28] Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Courcier, Paris.

[29] Liang, K.-Y. & Rathouz, P.J. (1997). In discussion of Desmond, A.F., *Journal of Statistical Planning and Inference* **60**, 77–121.

[30] Liang, K.-Y. & Zeger, S.L. (1996). Inference based on estimating functions in the presence of nuisance parameters (with discussion), *Statistical Science* **11**, 158–199.

[31] Liang, K.-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[32] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[33] Manton, K.G., Woodbury, M.A. & Stallard, E. (1981). A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties, *Biometrics* **37**, 257–269.

[34] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[35] Moore, D.F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions, *Biometrika* **23**, 583–588.

[36] Nelder, J.A. (1991). Quasi-likelihood and optimum estimating functions, Paper presented at *Symposium on Recent Concepts in Statistical Inference*, University of Waterloo.

[37] Nelder, J.A. (1997). In discussion of Desmond, A.F., *Journal of Statistical Planning and Inference* **60**, 77–121.

[38] Nelder, J.A. & Lee, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons, *Journal of the Royal Statistical Society, Series B* **54**, 273–284.

[39] Nelder, J.A. & Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika* **74**, 221–232.

[40] Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations, *Econometrica* **16**, 1–32.

[41] Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**, 1033–1048.

[42] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case–control studies, *Biometrika* **66**, 403–411.

[43] Sprott, D.A. & Viveros, R. (1984). The interpretation of maximum likelihood, *Canadian Journal of Statistics* **12**, 27–38.

[44] Ware, J.H. (1985). Linear models for the analysis of longitudinal studies, *American Statistician* **39**, 95–101.

[45] Wedderburn, R.W.M. (1974). Quasi-likelihood, generalized linear models and the Gauss–Newton method, *Biometrika* **61**, 439–447.

[46] Weinberg, C.R. & Wacholder, S. (1993). Prospective analysis of case–control data under general multiplicative intercept risk models, *Biometrika* **80**, 461–465.

[47] Whittemore, A.S. & Gong, G. (1994). Segregation analysis of case–control data using generalized estimating equations, *Biometrics* **50**, 1073–1087.

[48] Wilks, S.S. (1939). Shortest average confidence intervals from large samples, *Annals of Mathematical Statistics* **9**, 166–175.

[49] Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics* **31**, 949–952.

[50] Williams, D.A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics* **31**, 144–148.

[51] Yanagimoto, T. (1989). Combining moment estimates of a parameter common through strata, *Journal of Statistical Planning and Inference* **25**, 187–198.

[52] Zhao, L.P. (1994). Segregation analysis of human pedigrees using estimating equations, *Biometrika* **81**, 197–209.

[53] Zhao, L.P. & Prentice, R.L. (1991). Use of a quadratic exponential model to generate estimating equations for means, variances and covariances, in *Estimating Functions* V.P. Godambe, ed. Clarendon Press, Oxford, pp. 103–117.

A.F. DESMOND & V.P. GODAMBE

Estimation, Interval

As the phrase implies, *interval estimation* concerns the use of available data from a study to construct an interval estimator (often called a **confidence interval**) that is used to make a statistical **inference** about the true (but unknown) value of a parameter θ of interest (*see* **Estimation**). More specifically, an exact $100(1-\alpha)\%$ confidence interval is defined in terms of two **random variables**, a lower limit L and an upper limit U, such that $\Pr[L < \theta < U] = 1-\alpha$, where $0 < \alpha < 1$ and where α is typically chosen to take values such as 0.10, 0.05, 0.02, and 0.01. The exact confidence level $(1-\alpha)$ associated with the interval estimator (L, U) can be interpreted in two ways. In an infinite number of repetitions of the study leading to an infinite number of such $100(1-\alpha)\%$ confidence intervals, an exact proportion $(1-\alpha)$ of all such intervals, the *confidence coefficient*, will enclose the true value of the parameter θ. Equivalently, before the study is conducted and hence before any data are collected, the probability is exactly $(1-\alpha)$ that any random interval (L, U) will enclose the true value of the parameter θ. For any particular study, once the data are collected and the confidence interval is computed using the available data, the actual probability is either 0 or 1 that this observed interval [called the realization of the interval estimator (L, U)] actually encloses the true value of θ, and it is not known which of these two values (0 or 1) is correct. For example, if 1.25 and 3.80 are the observed (or realized) values of L and U for a given set of data, a statement like "$\Pr(1.25 < \theta < 3.80) = 0.95$" is statistically incorrect. Also, since it is L and U that are the random quantities (and not the parameter θ), it is inappropriate terminology to say that "θ falls inside the confidence interval". Indeed, the true unknown value of θ is fixed, and it is the interval estimator (L, U) that either encloses or does not enclose the true value of θ.

To consider a simple example, suppose that $Y_1, Y_2, \ldots, Y_n$ constitute a random sample of size n from an $\mathrm{N}(\mu, \sigma^2)$ population, and we wish to use these n observations to construct an exact $100(1-\alpha)\%$ confidence interval for the parameter μ. With $\overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, then the random variable $T_{n-1} = \sqrt{n}(\overline{Y} - \mu)/S$ has exactly a **Student's *t*-distribution** with $n-1$ degrees of freedom (df), namely $T_{n-1} \sim t_{n-1}$. Hence, if $t_{n-1,1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile point of Student's t-distribution with $n-1$ df, then

$$\begin{aligned} 1-\alpha &= \Pr(-t_{n-1,1-\alpha/2} < T_{n-1} < t_{n-1,1-\alpha/2}) \\ &= \Pr(-t_{n-1,1-\alpha/2} < \sqrt{n}(\overline{Y} - \mu)/S \\ &\quad < t_{n-1,1-\alpha/2}) \\ &= \Pr(\overline{Y} - t_{n-1,1-\alpha/2}S/\sqrt{n} < \mu \\ &\quad < \overline{Y} + t_{n-1,1-\alpha/2}S/\sqrt{n}), \end{aligned}$$

so that $L = \overline{Y} - t_{n-1,1-\alpha/2}S/\sqrt{n}$ and $U = \overline{Y} + t_{n-1,1-\alpha/2}S/\sqrt{n}$. Thus, based on the assumption that we have a random sample of size n from an $N(\mu, \sigma^2)$ population, then $\overline{Y} \pm t_{n-1,1-\alpha/2}S/\sqrt{n}$ is an exact $100(1-\alpha)\%$ confidence interval for μ. If the stated assumption is true, then this is the best $100(1-\alpha)\%$ confidence interval for μ in the sense that it has the shortest expected width.

Making the fully parametric assumption that the available data consist of independent random samples from normal populations allows for the construction of exact confidence intervals in more general settings. In particular, consider using data on n independent subjects to fit the **multiple linear regression** model

$$\mathrm{E}(Y_i | x_{i1}, x_{i2}, \ldots, x_{ik}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} = \mathbf{x}_i'\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \tag{1}$$

where the (univariate) response for the ith subject is Y_i, where the covariate vector for the ith subject is $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$, and where the vector of unknown regression coefficients is $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k)$. Under the classical assumptions that the **conditional distribution** of Y_i given $x_{i1}, x_{i2}, \ldots, x_{ik}$ is normal with conditional mean given by (1) and with conditional (homogeneous) variance σ^2, then the **maximum likelihood** (and unweighted **least squares**) unbiased estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$ of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, where $\mathbf{X}' = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ and where the row vector of responses is $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_n)$. Under the given assumptions, for $j = 0, 1, 2, \ldots, k$, it follows that $\hat{\beta}_j \sim \mathrm{N}(\beta_j, v_{jj}\sigma^2)$, where v_{jj} is the jth diagonal element of the matrix $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$. With

the unbiased estimator of σ^2 being $\hat{\sigma}^2 = (n - k - 1)^{-1}(\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y})$, the standardized random variable $(\hat{\beta}_j - \beta_j)/(v_{jj})^{1/2}\hat{\sigma}$ has exactly a t_{n-k-1} distribution. Thus, it follows directly that the exact $100(1-\alpha)\%$ confidence interval for β_j has the specific form $\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2}(v_{jj})^{1/2}\hat{\sigma}$.

However, in contrast to the classical multiple linear regression scenario described above, there are many realistic and important multivariable modeling applications in biostatistics where assumptions like normality of the response variable, homogeneous variance, and even independence among responses are not justified. For example, if the response variable Y_i is dichotomous (e.g. $Y_i = 1$ if subject i has a certain disease, and $Y_i = 0$ if not), then the distribution of Y_i is point-binomial with mean $\mathrm{E}(Y_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \Pr(Y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ik}) = \pi_i$ and with variance $\mathrm{var}(Y_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \pi_i(1 - \pi_i)$, so that the variance of Y_i is not the same for all i (i.e. is not homogeneous across subjects) since the mean of Y_i varies with i. In this situation, a useful and popular maximum likelihood approach for modeling π_i as a function of covariates $x_{i1}, x_{i2}, \ldots, x_{ik}$ is **logistic regression**, where we consider the logistic regression model

$$\mathrm{logit}(\pi_i) = \ln \frac{\pi_i}{(1 - \pi_i)} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}. \qquad (2)$$

In contrast to model (1), model (2) describes logit $[\mathrm{E}(Y_i|x_{i1}, x_{i2}, \ldots, x_{ik})]$ as a linear function of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$. The theoretical aspects and practical applications of logistic regression methods have been described in several textbooks (e.g. see Breslow & Day [1], Hosmer & Lemeshow [5], and Kleinbaum et al. [6]).

More generally, the logistic regression model is one example of a very broad family of **generalized linear models** for describing discrete and continuous outcome data [7]. The generalized linear model (GLM) family includes, as some special cases, multiple linear regression and **analysis of variance** models under normality, logistic regression and **Poisson regression** models, probit models (*see* **Quantal Response Models**), **multinomial** response models for categorical outcomes, and some commonly used models for **survival data**.

Maximum likelihood methods, which are optimal asymptotically under certain regularity conditions, are typically used to fit generalized linear models; thus, associated interval estimation procedures are generally valid only for large samples. To discuss the use of maximum likelihood-based interval estimation methods, suppose that $L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})$ is a **likelihood** function, where $\mathbf{Y}$ and $\mathbf{X}$ are as defined earlier and where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ is a vector of unknown parameters. For example, $L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})$ would be a product of appropriately defined normal distributions for the multiple linear regression example considered earlier, where $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \sigma^2)$ and $p = k + 2$. And $L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})$, with $\boldsymbol{\theta} = \boldsymbol{\beta}$ and $p = k + 1$, would be a product of point-binomial distributions for the logistic regression situation previously discussed. More generally, $L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta})$ could be a **hypergeometric** distribution-based likelihood appropriate for conditional logistic regression [6, Chapter 20], or it could represent a **partial likelihood** appropriate for the **Cox regression model** in **survival analysis** [7, Chapter 13].

The maximum likelihood estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)'$ of $\boldsymbol{\theta}$ is the vector solution to the set of p maximum likelihood (or score) equations $\partial \ln L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})/\partial\theta_j = 0$, $j = 1, 2, \ldots, p$. In most generalized linear model situations (one notable exception being the multiple linear regression example considered earlier), these maximum likelihood equations are nonlinear in the elements of $\boldsymbol{\theta}$, and so they must be solved by iteratively reweighted least squares. Iteratively reweighted least squares methods produce the large-sample estimated variance–covariance matrix of $\hat{\boldsymbol{\theta}}$ as the $(p \times p)$ matrix $\hat{\mathbf{V}} = [\mathbf{I}(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}})]^{-1}$, where $\mathbf{I}(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}})$ is the observed information matrix with (j, j')th element defined as $-\partial^2 \ln L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})/\partial\theta_j \partial\theta_{j'}$, evaluated at $\hat{\boldsymbol{\theta}}$. Since the estimated variance of $\hat{\theta}_j$ is $\hat{v}_{jj}$, the (j, j)th element of $\hat{\mathbf{V}}$, it follows from maximum likelihood theory that, for large samples, the quantity $(\hat{\theta}_j - \theta_j)/\sqrt{\hat{v}}_{jj}$ is approximately distributed as a standard normal random variable if the assumed statistical model is valid. Hence, an approximate $100(1-\alpha)\%$ large-sample (normal approximation-based) confidence interval for the parameter θ_j is $\hat{\theta}_j \pm Z_{1-\alpha/2}\hat{v}_{jj}$.

It is also possible to construct a confidence interval for θ_j using the principles underlying **likelihood ratio tests**. In particular, let $\boldsymbol{\psi}_{\theta_j} = (\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)'$ be the $(p-1) \times 1$ column vector of all parameters in $\boldsymbol{\theta}$ except for θ_j. Then, if $\hat{\boldsymbol{\psi}}_{\theta_j}$ is the maximum likelihood estimator of $\boldsymbol{\psi}_{\theta_j}$ with θ_j fixed

(so that $\hat{\boldsymbol{\psi}}_{\theta_j}$ will, in general, be a function of θ_j), then the log likelihood ratio statistic

$$2 \ln L(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}) - 2 \ln L(\mathbf{Y}, \mathbf{X}; \theta_j, \hat{\boldsymbol{\psi}}_{\theta_j})$$

has an approximate χ_1^2 distribution (**chi-square distribution** with one **degree of freedom**) for large samples if the assumed model is valid. Then, the set of all values of θ_j satisfying

$$2 \ln L(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}) - 2 \ln L(\mathbf{Y}, \mathbf{X}; \theta_j, \hat{\boldsymbol{\psi}}_{\theta_j}) \leq \chi^2_{1,1-\alpha}$$

constitutes an approximate large-sample $100(1 - \alpha)\%$ confidence interval for θ_j. It is important to note that such likelihood ratio-based confidence intervals often perform better for small samples than do the normal approximation-based intervals discussed earlier. The partially maximized log likelihood $\ln L(\mathbf{Y}, \mathbf{X}; \theta_j, \hat{\boldsymbol{\psi}}_{\theta_j})$ is called the **profile log likelihood** for θ_j. All of these methods can be generalized to produce confidence sets for a vector of parameters (e.g. [7]).

There are several other general methods for constructing interval estimators of unknown parameters. For example, for correlated response data as encountered in **longitudinal studies** where the same response variable is measured more than once on each subject, **generalized estimating equations** (GEE) methods based on **quasi-likelihood** theory can be used to develop interval estimators that involve the use of a **robust** estimator of the variance–covariance matrix of the parameter estimates (e.g. [2]). There are numerous **nonparametric methods** available for developing interval estimators (e.g. [4] and [9]), and there are so-called exact confidence interval methods based on the use of certain network algorithms (e.g. [8]) (*see* **Exact Inference for Categorical Data**). And, last but not least, **computer-intensive bootstrap** interval estimation methods [3] have become very popular in recent years.

In most practical situations it is possible to consider the use of several different types of interval estimators, some involving more assumptions than others. It is generally good advice to use various alternative interval estimation methods in such situations, these methods hopefully ranging from being fully parametric to being fully nonparametric in nature. Only when global statistical conclusions about the parameters of interest vary distinctly from method to method would one have to be seriously concerned about the appropriateness of any assumptions that have been made. When in doubt, it is best to use interval estimation methods that possess good statistical properties not depending directly on the validity of unverifiable assumptions.

References

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1. *The Analysis of Case–Control Studies*. International Agency for Research on Cancer, Lyon.

[2] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, London.

[3] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[4] Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods*. Wiley, New York.

[5] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[6] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont.

[7] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, New York.

[8] Mehta, C.R., Patel, N.R. & Gray, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables, *Journal of the American Statistical Association* **80**, 969–973.

[9] Puri, M.L. & Sen, P.K. (1985). *Nonparametric Methods in General Linear Models*. Wiley, New York.

(*See also* **Bayesian Methods**; **Fiducial Probability**)

LAWRENCE L. KUPPER

Estimation

The object of estimation is to use available data to estimate or guess the values of unknown quantities. The unknown quantities, which are called *parameters*, may be familiar population quantities such as the population **mean** μ, population proportion p, population **variance** σ^2, and population **median** ν. In other situations the parameters are part of more elaborate statistical models, such as the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in a **linear regression** model

$$Y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \varepsilon,$$

which relates a response variable Y to **covariates** or **explanatory variables**, $x_1, x_2, \ldots, x_p$.

For instance, we may be interested in the mean diastolic blood pressure $\mu = \mathrm{E}(Y)$ of a target population, in which case the sample mean $\overline{Y}$ is a familiar estimator, or we may model the blood pressure as a function of the predictor $x =$ age, and a random error ε, in a linear model such as $Y = \beta_0 + \beta_1 x + \varepsilon$. In this case natural estimators of the parameters β_0 and β_1 are the **least squares** estimators

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x}) Y_i}{\sum (x_i - \overline{x})^2}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}.$$

The field of biostatistics is, to a large extent, concerned with developing estimators for parameters in different types of medical and public health studies and to give measures of the accuracy of these estimates. Another important concern is to select **efficient**, or, if possible, optimal estimators of a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ from classes of reasonable estimators.

Let $P = P(\mathbf{y})$ denote the probability distribution of the data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. In the discrete case, P is the probability function of $\mathbf{Y}$, and in the continuous case it is the distribution function of $\mathbf{Y}$. The postulated relationship between P and $\boldsymbol{\theta}$ is crucial for estimation. Two commonly used approaches are as follows:

1. In *parametric* models, a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ determines P. Thus, suppose $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$; then, if the distribution of the response Y is **normal**, $\boldsymbol{\theta}$ determines P, that is, $P(\mathbf{y}) = P(\mathbf{y}, \boldsymbol{\theta})$.
2. In the $\boldsymbol{\theta} = (\mu, \sigma^2)$ example, if it is not possible to assume normality or any other specified distribution for $\mathbf{Y}$, then $\boldsymbol{\theta}$ does not determine P. In such cases it is often useful to express $\boldsymbol{\theta}$ as a function of P, that is, $\boldsymbol{\theta} = \boldsymbol{\theta}(P) = [\mu(P), \sigma^2(P)]$. In general, such representations are not unique. For a given $\boldsymbol{\theta}$ in a study we will see in what follows that it is possible to have $\boldsymbol{\theta} = h_1(P)$ and $\boldsymbol{\theta} = h_2(P)$ for different h_1 and h_2. Another parameterization in the (μ, σ^2) example consists of introducing the parameter $\boldsymbol{\theta} = (\mu, \sigma^2, F_\varepsilon)$, where F_ε is the unknown distribution of $\varepsilon = Y - \mu$. Under general conditions this $\boldsymbol{\theta}$ determines P. In such cases, where at least one of the unknowns in θ is real and at least one is a function, the model is called *semiparametric*. In the simple linear model for blood pressure vs. age, if ε is normal $(0, \sigma^2)$ and the equation $Y = \beta_0 + \beta_1 x + \varepsilon$ is assumed, then the distribution of Y is determined by $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ and the model is parametric. However, we can define β_0 and β_1 to be the coefficients in the best linear predictor of Y, that is, they minimize $\mathrm{E}[Y - (b_0 + b_1 X)]^2$. In this case $\beta_1 = \mathrm{cov}(X, Y)/\mathrm{var}(X)$, $\beta_0 = \mathrm{E}(Y) - \beta_1 \mathrm{E}(X)$, $\sigma^2 = \mathrm{var}[Y - (\beta_0 + \beta_1 X)]$, and $(\beta_0, \beta_1, \sigma^2)$ does not determine the distribution F of (X, Y). This second approach has the advantage that it is not necessary to assume a linear model. Instead, the goal is to estimate the coefficients of the best *linear* predictor $\beta_0 + \beta_1 X$ of Y. In this case β_0, β_1, and σ^2 are functions of the distribution P of (X, Y). In some cases it is useful to consider the semiparametric model where the full parameter is $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2, F_X, F_{\varepsilon,x})$. Here F_X is the distribution of X and $F_{\varepsilon,x}$ is the **conditional distribution** of $\varepsilon = Y - (\beta_0 + \beta_1 X)$ given $X = x$. Under general conditions, this expanded $\boldsymbol{\theta}$ determines P.

Note that the notational scheme $\boldsymbol{\theta} = \boldsymbol{\theta}(P)$ also works in the parametric case; however, when appropriate, parametric modeling with $P = P_{\boldsymbol{\theta}}$ often leads to simple analysis and efficient estimators. Parametric models are often natural for studies where the distributions are binomial, **multinomial**, **hypergeometric**, **negative binomial** or, more generally, can be derived

by probabilistic calculations from an experimental situation. The semiparametric approach is appropriate when population means, regression coefficients, variances, and other population quantities are of interest but it is not possible to assume normality of distributions or linearity of mean relationships.

The next sections discuss some of the standard methods of estimation including some of the theory. Books that treat these topics in detail include Bickel & Doksum [1], Lehmann [6], and Bickel et al. [2]. Books that focus on estimation in biostatistics include Kalbfleisch & Prentice [4], Lawless [5], and Cox & Oakes [3].

Methods of Estimation

We consider several techniques for estimating a vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ or, more generally, a vector $[q_1(\boldsymbol{\theta}), \ldots, q_r(\boldsymbol{\theta})]$ of functions of $\boldsymbol{\theta}, r \leq m$. The first three classes of estimates are for parameters that can be expressed as $\boldsymbol{\theta} = \boldsymbol{\theta}(P)$. The fourth and fifth are parametric with $P = P_{\boldsymbol{\theta}}$.

The Frequency Plug-in Method

Suppose we obtain a sample of n independent identically distributed responses and classify each response into one of d distinct categories. Let p_j denote the probability of the jth category and let N_j denote the number of responses that fall in the jth category. Then the distribution of $N_1, \ldots, N_d$ is multinomial $(n, p_1, \ldots, p_d)$ and $p_j = \mathrm{E}(N_j)/n$. The frequency plug-in estimator $\hat{p}_j$ of p_j replaces the unknown frequency $\mathrm{E}(N_j)$ with the observed frequency N_j, that is, $\hat{p}_j = N_j/n$. For functions $\theta_s = h_s(p_1, \ldots, p_d)$ the plug-in estimators are $\hat{\theta}_s = h_s(\hat{p}_1, \ldots, \hat{p}_d), s = 1, \ldots, m$.

Examples are provided by multiway **contingency tables**. For instance, using different subscripting, suppose N_{ijk} denotes the number of high blood pressure patients in a study that fall in the category (i, j, k) based on the following three classifications:

1. treatment membership ($i = 0$, control; $i = 1$, treatment),
2. reduction in blood pressure ($j = 0$, zero or negative; $j = 1$, moderate; $j = 2$, high),
3. age group ($k = 0$, <30; $k = 1$, 30–40; $k = 2$, 41–50; $k = 3$, 51–60; $k = 4$, >60).

Let p_{ijk} be the probability of the category (i, j, k), then $\hat{p}_{ijk} = N_{ijk}/n$. Moreover, the contrast parameters

$$\theta_k = \sum_{j=1}^{2} j(p_{1jk} - p_{ojk}), \quad k = 0, 1, \ldots, 4,$$

which measure the expected improvement in blood pressure score due to the treatment for age group k, has the plug-in estimators

$$\hat{\theta}_k = \sum_{j=1}^{2} j(\hat{p}_{1jk} - \hat{p}_{ojk}), \quad k = 0, 1, \ldots, 4.$$

Returning to the general notation $p_j, N_j, j = 1, \ldots, d$, consider the case where $p_j = p_j(\boldsymbol{\theta})$ depends on some parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m), m \leq d$. For example, suppose θ denotes the probability of alleles A_1 at a certain locus and suppose A_2 has frequency $(1 - \theta)$. In the **Hardy–Weinberg** model the three genotypes $A_1\ A_1$, $A_1\ A_2$, and $A_2\ A_2$ have probabilities

$$p_1 = \theta^2, \qquad p_2 = 2\theta(1 - \theta), \qquad p_3 = (1 - \theta)^2.$$

In this case, θ can be written as $\theta = \sqrt{(p_1)}$, or $\theta = 1 - \sqrt{(p_3)}$, or $\theta = p_1 + \frac{1}{2}p_2$. Thus plug-in estimators are not unique and three plug-in estimators are $\sqrt{(\hat{p}_1)}, 1 - \sqrt{(\hat{p}_3)}$, and $\hat{p}_1 + \frac{1}{2}\hat{p}_2$. In general, plug-in estimators of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m), m \leq k$, are obtained by solving the equations $p_j(\boldsymbol{\theta}) = \hat{p}_j, j = 1, \ldots, k$, for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Let the solution be $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$, then the plug-in estimator of $(q_1(\boldsymbol{\theta}), \ldots, q_r(\boldsymbol{\theta}))$ is $(q_1(\hat{\boldsymbol{\theta}}), \ldots, q_r(\hat{\boldsymbol{\theta}}))$.

The Method of Moments

Another plug-in method is the **method of moments**. For a vector $\mathbf{X} = (X_1, \ldots, X_k)$, of observations, let the moments be

$$m_{jkrs} = \mathrm{E}(X_r^j X_s^k), \quad j \geq 0, k \geq 0, r, s = 1, \ldots, k.$$

For independent identically distributed $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik}), i = 1, \ldots, n$, we define the empirical or sample moment to be

$$\hat{m}_{jkrs} = \frac{1}{n} \sum_{i=1}^{n} X_{ir}^j X_{is}^k, \quad j \geq 0, k \geq 0, r, s = 1, \ldots, k.$$

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ can be expressed as a function of the moments, then the method of moment estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by replacing m_{jkrs} by $\hat{m}_{jkrs}$. For instance, when $k = 1$, the method of moments estimator of $\boldsymbol{\theta} = (\mu, \sigma^2) = [\mathrm{E}(X), \mathrm{E}(X^2) - \mu^2]$ is $[\overline{X}, (1/n)\sum X_i^2 - \overline{X}^2]$, where $\overline{X} = n^{-1}\sum X_i$. In the example where β_0 and β_1 are the parameters of the best linear predictor of X and Y,

$$\beta_1 = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)} = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)}{\mathrm{E}(X^2) - \mathrm{E}^2(X)},$$
$$\beta_0 = \mathrm{E}(Y) - \beta_1 \mathrm{E}(X).$$

Thus the method of moment estimators are

$$\hat{\beta}_1 = \frac{n^{-1}\sum X_i Y_i - \overline{X}\,\overline{Y}}{n^{-1}\sum X_i^2 - (\overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

For a parametric example consider a study where the survival time T is modeled to have a **gamma distribution** with density

$$\left[\frac{\lambda^\alpha}{\Gamma(\alpha)}\right] t^{\alpha-1} \exp\{-\lambda t\}, \quad t > 0,\ \alpha > 0,\ \lambda > 0.$$

In this case $\theta = (\alpha, \lambda)$, $\mu_1 = \mathrm{E}(T) = \alpha/\lambda$, and $\mu_2 = \mathrm{E}(T^2) = \alpha(1 + \alpha)/\lambda^2$. Solving for θ gives

$$\alpha = \left(\frac{\mu_1}{\sigma}\right)^2, \quad \hat{\alpha} = \left(\frac{\overline{X}}{\hat{\sigma}}\right)^2,$$
$$\lambda = \frac{\mu_1}{\sigma^2}, \qquad \hat{\lambda} = \frac{\overline{X}}{\hat{\sigma}^2},$$

where $\sigma^2 = \mu_2 - \mu_1^2$ and $\hat{\sigma}^2 = n^{-1}\Sigma X_i^2 - \overline{X}^2$. In this example the method of moment estimator is not unique. We can express θ as a function of μ_1 and $\mu_3 = \mathrm{E}(T^3)$ and obtain a method of moment estimator based on $\hat{\mu}_1$ and $\hat{\mu}_3$.

Weighted and Generalized Least Squares Estimators

Suppose Y is a response such as blood pressure whose distribution depends on the levels $x_1, \ldots, x_p$ of p predictor variables such as $X_1 =$ level of treatment, $X_2 =$ age, $X_3 =$ dietary salt intake, etc. Suppose $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is the vector of coefficients of the best linear predictor, that is, $\boldsymbol{\beta}$ minimizes

$$\mathrm{E}\left[\left(Y - \sum_{j=0}^{p} b_j X_j\right)^2 w(\mathbf{X}, Y)\right],$$

where $X_0 = 1$, $\mathbf{X} = (X_0, X_1, \ldots, X_p)$, and $w(\mathbf{X}, Y)$ is a given weight function exemplified in what follows. Here $\boldsymbol{\beta} = \boldsymbol{\beta}(P)$ depends on the probability distribution P of $(\mathbf{X}, Y)$. To estimate $\boldsymbol{\beta}(P)$ we replace P by the empirical probability distribution P_n which gives probability n^{-1} to each observed sample point $(x_{i1}, \ldots, x_{id}, y_i) = (\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$. This leads us to seek the minimizer $\hat{\boldsymbol{\beta}}$ of

$$\sum \left(y_i - \sum_{j=0}^{p} b_j x_{ij}\right)^2 w_i,$$

where $w_i = w(x_i, y_i)$ and $x_{i0} = 1$. The solution is $\hat{\boldsymbol{\beta}} = (\mathbf{X}_\mathrm{D}^\mathrm{T}\mathbf{W}\mathbf{X}_D)^{-1}\mathbf{X}_\mathrm{D}^\mathrm{T}\mathbf{W}_\mathbf{y}$, where $\mathbf{X}_\mathrm{D}$ is the $n \times (p + 1)$ design matrix (x_{ij}), $\mathbf{y} = (y_1, \ldots, y_n)^\mathrm{T}$, $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ and $\mathbf{X}_\mathrm{D}$ and $\mathbf{W}$ are assumed to have ranks $p + 1$ and n, respectively. This $\hat{\boldsymbol{\beta}}$ is called the *weighted least squares estimator*.

The weights w_i are determined by the application. For example, consider a study where a health indicator such as CD4 blood cell count is taken at time points $t_0 < t_1 < \cdots < t_k$ for HIV infected subjects. After adjusting for changes in blood cell counting technology, the mean of the fourth root of such blood cell counts decreases linearly over time [7]. That is, the slope is constant. For the ith subject the responses are the local slopes

$$Y_{ij} = \frac{\{CD4 \text{ at time } t_{i,j}\}^{1/4} - \{CD4 \text{ at time } t_{i,j-1}\}^{1/4}}{t_{i,j} - t_{i,j-1}},$$
$$j = 1, \ldots, k_i, i = 1, \ldots, m,$$

where $k_i =$ number of measurements for subject i. The predictors considered are $x_{i1} =$ age and $x_{i2} = \{\text{CD8 at first visit time} = t_{i,0}\}^{1/4}$, where CD8 is another blood cell count health indicator. The shorter the time interval length $\Delta_{ij} = t_{i,j} - t_{i,j-1}$ is, the more variable Y_{ij} is. The standard deviation of Y_{ij} is close to being proportional to Δ_{ij}^{-1}; thus a reasonable weighted least squares estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ is the minimizer of

$$\sum_{i=1}^{m}\sum_{j=1}^{k_i}[y_{ij} - (b_0 + b_1 x_{i1} + b_2 x_{i2})]^2 \Delta_{ij}.$$

Returning to the general design matrix notation, the *generalized least squares estimator* is defined as

the minimizer of

$$(\mathbf{Y} - \mathbf{b}\mathbf{X}_D)^{\mathrm{T}}\mathbf{W}(\mathbf{Y} - \mathbf{b}\mathbf{X}_D)$$

where $\mathbf{b} = (b_0, b_1, \ldots, b_p)^{\mathrm{T}}$ and $\mathbf{W}$ is an $n \times n$ matrix not necessarily diagonal. Again, it is assumed that $\mathbf{X}_D$ has rank $p + 1$ and $\mathbf{W}$ has rank n, in which case the estimator is given by the same formula as before.

Maximum Likelihood Estimators

The idea of **maximum likelihood** is to find the value $\hat{\boldsymbol{\theta}}(\mathbf{y})$ of $\boldsymbol{\theta}$ which is "most likely" to have produced the data $\mathbf{y} = (y_1, \ldots, y_n)$. These estimators $\hat{\boldsymbol{\theta}}(\mathbf{y})$ are defined for *regular parametric models*, that is, for models where in the discrete case the set of $\mathbf{y}$, where the probability function $p(\mathbf{y}, \boldsymbol{\theta}) = \Pr(\mathbf{Y} = \mathbf{y})$, is nonzero does not depend on $\boldsymbol{\theta}$, and in the continuous case $\mathbf{Y}$ has a density $p(\mathbf{y}, \boldsymbol{\theta})$. It is also assumed that $p(\mathbf{y}, \boldsymbol{\theta}) < \infty$ for all $\mathbf{y} \in R^n$ and all $\boldsymbol{\theta}$ in Θ, where the parameter set Θ of possible $\boldsymbol{\theta}$ is a subset of R^k for some k. If $\mathbf{y} = (y_1, \ldots, y_n)$ are data values obtained in a study, the **likelihood** function $L_{\mathbf{y}}(\boldsymbol{\theta})$ is defined as the function of $\boldsymbol{\theta}$ given by $L_{\mathbf{y}}(\boldsymbol{\theta}) = p(\mathbf{y}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. A *maximum likelihood estimator* (mle) of $\boldsymbol{\theta}$ is a value $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ which satisfies $L_{\mathbf{y}}(\hat{\boldsymbol{\theta}}) = \max\{L_{\mathbf{y}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. The maximum likelihood estimator of $[q_1(\boldsymbol{\theta}), \ldots, q_r(\boldsymbol{\theta})]$ is defined as $[q_1(\hat{\boldsymbol{\theta}}), \ldots, q_r(\hat{\boldsymbol{\theta}})]$. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained as the solution to the *likelihood equations*

$$\frac{\partial}{\partial \theta_j} l_{\mathbf{y}}(\boldsymbol{\theta}) = 0, \quad j = 1, \ldots, m,$$

where $l_{\mathbf{y}}(\boldsymbol{\theta}) = \log L_{\mathbf{y}}(\boldsymbol{\theta})$.

For example, if $Y_1, \ldots, Y_n$ are independent, identically distributed with normal (μ, σ^2) distribution, then

$$l_{\mathbf{y}}(\boldsymbol{\theta}) = -\tfrac{1}{2}n \log(2\pi) - n \log \sigma - \tfrac{1}{2}\sigma^{-2} \sum (y_i - \mu)^2$$

and the maximum likelihood estimators are easily shown to be $\hat{\mu} = \overline{y}$ and $\hat{\sigma}^2 = n^{-1} \sum (y_i - \overline{y})^2$. In the trinomial Hardy–Weinberg example with $p_1 = \theta^2$, $p_2 = 2\theta(1 - \theta)$, $p_3 = (1 - \theta)^2$, we have

$$L_{\mathbf{n}}(\theta) = p(\mathbf{n}, \theta) = P(\mathbf{N} = \mathbf{n}, \theta) = \frac{n!}{n_1! n_2! n_3!} \theta^{2n_1} [2\theta(1 - \theta)]^{n_2} (1 - \theta)^{2n_3},$$

where $n = n_1 + n_2 + n_3$. Solving the likelihood equation $l'_{\mathbf{n}}(\boldsymbol{\theta}) = 0$, we find the mle $\hat{\theta} = \hat{p}_1 + \frac{1}{2}\hat{p}_2$, which we recognize as the third plug-in method estimator in the section on the Frequency Plug-in Method above.

Bayesian Estimators

In **Bayesian** models a distribution $\pi(\theta)$, called a **prior distribution**, is introduced for the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. In the discrete case, $\pi(\boldsymbol{\theta})$ is the probability function of $\boldsymbol{\theta}$ and in the continuous case it is the density. The probability function (discrete case) or density (continuous case) of the data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ now represents the conditional distribution of $\mathbf{Y}$ given $\boldsymbol{\theta}$, and is written $p(\mathbf{y}|\boldsymbol{\theta})$. The conditional distribution of $\boldsymbol{\theta}$ given the data $\mathbf{y}$ is called the *posterior distribution* of θ and by **Bayes' Theorem**, it is given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = c\pi(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}),$$

where $c^{-1} = \sum_k \pi(k) p(\mathbf{y}|k)$ in the discrete case and $c^{-1} = \int \pi(t) p(\mathbf{y}|t)\, \mathrm{d}t$ in the continuous case. One approach to Bayesian estimation consists of selecting the value $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ that makes the observed $\mathbf{y}$ most "probable" according to the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, that is, it is a value of $\boldsymbol{\theta}$ that maximizes $\pi(\boldsymbol{\theta}|\mathbf{y})$. Another approach is to use the value $\boldsymbol{\theta}_B$ that minimizes the posterior **mean square error**. That is,

$$\boldsymbol{\theta}_B = \arg\min \mathrm{E}[(\boldsymbol{\theta}_B - \boldsymbol{\theta})^2|\mathbf{y}] = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) = \text{the posterior mean.}$$

Unbiased Estimation. Residual Sum of Squares

The error in using the observable $\hat{\theta}$ to estimate the unknown θ is $\hat{\varepsilon} = \hat{\theta} - \theta$. In many cases it is possible to adjust estimators so that their long-run average error is zero, that is, so that $\mathrm{E}(\hat{\varepsilon}) = 0$. Such estimators are called **unbiased**. For example, consider the parameter $\sigma^2 = \mathrm{var}(Y)$. Suppose $Y_1, \ldots, Y_n$ are independent identically distributed, then we arrived earlier at the estimator $\hat{\sigma}^2 = n^{-1} \sum (Y_i - \overline{Y})^2$. With a little algebra it can be shown that $\mathrm{E}(\hat{\sigma}^2) = [(n - 1)/n]\sigma^2$, and $\hat{\sigma}^2$ is not unbiased. There is a debate as to whether unbiased estimation is desirable. However, a nearly universal tradition has developed

where variances are unbiasedly estimated. In our case, this amounts to adjusting $\hat{\sigma}^2$ by multiplying it by $(n-1)/n$. That is, σ^2 is unbiasedly estimated by

$$S^2 = (n-1)^{-1} \sum (Y_i - \overline{Y})^2.$$

As another example, consider the linear model

$$Y_i = \beta_0 + \sum x_{ij}\beta_j + \varepsilon_i, \qquad \mathrm{E}(\varepsilon_i) = 0,$$

$$\sigma^2 = \mathrm{var}(Y_i) = \mathrm{var}(\varepsilon_i).$$

Here $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed. When the design matrix $\mathbf{X}_\mathrm{D} = (x_{ij})$ has rank $p+1$, the unbiased estimator of σ^2 is $RSS/[n-(p+1)]$, where RSS = residual sum of squares is defined by

$$\mathrm{RSS} = \sum_i \left[Y_i - \sum_j x_{ij}\hat{\beta}_j \right]^2$$

(*see* **Multiple Linear Regression**). Here $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are the unweighted [$\mathbf{W}$ = weight matrix = diag(1, $\ldots$, 1)] least squares estimators of $\beta_0, \ldots, \beta_p$.

Standard Errors

It is crucial to provide a measure of the accuracy of estimators. The *error* in using $\hat{\theta}$ to estimate θ is $\hat{\varepsilon} = \hat{\theta} - \theta$. Suppose that $\hat{\varepsilon}$ has an expected value close to zero in the sense that $\sqrt{n}\mathrm{E}(\hat{\varepsilon}) \to 0$ as $n \to \infty$. Then a measure of how close the distribution of the error is concentrated near zero is given by the standard deviation $\sigma_0 = \mathrm{sd}(\hat{\varepsilon}) = \mathrm{sd}(\hat{\theta})$ of $\hat{\varepsilon}$. In the parametric case σ_0 is a function $\sigma_0(\theta)$ of θ and in the general case it is a function $\sigma_0(P)$ of the probability distribution P. Thus we can use the estimation methods provided earlier to estimate σ_0. Such an estimate $\hat{\sigma}_0$ of σ_0 is called the **standard error** of $\hat{\theta}$, and it is written as $\mathrm{se}(\hat{\theta})$.

For example, if $Y_1, \ldots, Y_n$ are independent identically distributed, then for estimating the mean μ, $\mathrm{sd}(\overline{Y}) = \sigma/\sqrt{n}$ and $\mathrm{se}(\overline{Y}) = S/\sqrt{n}$, where $S^2 = (n-1)^{-1}\Sigma(Y-\overline{Y})^2$. In the trinomial Hardy–Weinberg case, $\hat{\theta} = \hat{p}_1 + \frac{1}{2}\hat{p}_2$ and it can be shown that $\mathrm{var}(\hat{\theta}) = (2n)^{-1}\theta(1-\theta)$. It follows that $\mathrm{se}(\hat{\theta}) = [(2n)^{-1}\hat{\theta}(1-\hat{\theta})]^{1/2}$.

In the linear model $Y_i = \beta_0 + \Sigma x_{ij}\beta_j + \varepsilon_i$, with $\sigma^2 = \mathrm{var}(\varepsilon_i)$, the covariance matrix of the (unweighted) least squares estimator $\hat{\boldsymbol{\beta}}$ is $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}_\mathrm{D}^\mathrm{T}\mathbf{X}_D)^{-1}\sigma^2$. The standard error of $\hat{\beta}_j$ is then $\sqrt{v_j}S_p$, where v_j is the jth diagonal element of $(\mathbf{X}_\mathrm{D}^\mathrm{T}\mathbf{X}_\mathrm{D})^{-1}$ and $S_p^2 = \mathrm{RSS}/n - (p+1)$ is the unbiased estimator of σ^2.

Comparison of Estimators: Efficiency

For any given unknown $q(\boldsymbol{\theta})$ there are many possible estimators. The preferable estimator is the one that makes the most efficient use of the data; that is, the estimator $T(\mathbf{Y})$ of $q(\boldsymbol{\theta})$ should be such that the distribution of the error $\hat{\varepsilon} = T(\mathbf{Y}) - q(\boldsymbol{\theta})$ is as closely as possible concentrated near zero. There are many ways of making this idea precise. The most common is to try to minimize the long-run squared error, that is, to minimize the *mean squared error (mse)*:

$$M(T;\boldsymbol{\theta}) = \mathrm{E}[T(\mathbf{Y}) - q(\boldsymbol{\theta})]^2.$$

The mse can be decomposed into a "systematic error" represented by the square of the bias $B(T;\boldsymbol{\theta}) = \mathrm{E}[T(\mathbf{Y}) - q(\boldsymbol{\theta})]$ and the intrinsic variability represented by the variance $V(T;\boldsymbol{\theta}) = \mathrm{var}(T(\mathbf{Y}))$. Thus

$$M(T;\boldsymbol{\theta}) = B^2(T;\boldsymbol{\theta}) + V(T;\boldsymbol{\theta}).$$

For instance, suppose $Y_1, \ldots, Y_n$ are independent identically distributed random variables that represent the survival times of n patients in a medical study. If the distribution is **exponential**, we can write the survival function $S(y) = P(Y > y)$ as $\exp\{-y/\theta\}$, where θ is the mean survival time $\mathrm{E}(Y)$. It is easy to see that the mle of θ is $\overline{Y} = n^{-1}\Sigma Y_i$. Thus the mle of $q(\theta) = \exp\{-y/\theta\}$ is

$$T(\mathbf{Y}) = q(\overline{Y}) = \exp\left\{\frac{-y}{\overline{Y}}\right\}.$$

Since $2\sum_1^n Y_i/\theta$ has a **chi-square distribution** with $2n$ **degrees of freedom** (χ^2_{2n}), **numerical integration** will yield the systematic error (bias) and variance of $T(\mathbf{Y})$. On the other hand, good approximations can be obtained from the expansion

$$q(\hat{\theta}) \cong q(\theta) + q'(\theta)[\hat{\theta} - \theta] + \tfrac{1}{2}q''(\theta)[\hat{\theta} - \theta]^2$$

and the first four moments of the χ^2_{2n} distribution. Such computations yield insight into the components of the mse of $T(\mathbf{Y})$. However, it is useful to have general results to check whether a given estimator

can be improved upon. One such result for parametric models is the **information** *inequality*

$$\text{var}[T(\mathbf{Y})] \geq \frac{[\psi'(\theta)]^2}{I(\theta)}, \tag{1}$$

where $\psi(\theta) = \text{E}[T(\mathbf{Y})]$ and

$$I(\theta) = \text{E}\left\{\left[\frac{\partial}{\partial\theta}\log p(\mathbf{Y},\theta)\right]^2\right\} \tag{2}$$

is the *Fisher information* of the model with probability function (discrete case) or density function (continuous case) $p(\mathbf{Y},\theta)$. If $Y_1,\ldots,Y_n$ are independent, identically distributed, then $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the expression (2) with $\mathbf{Y}$ replaced by Y_1. In this case it can be shown (e.g. [6]) under general conditions that if $\hat{\theta}$ is the mle, then $n^{1/2}[q(\hat{\theta}) - q(\theta)]$ is asymptotically normal with mean zero and variance $[q'(\theta)]^2/nI(\theta)$. That is, in the limiting distribution, $\sqrt{n}$ (bias) is zero and n (var) reaches the lower bound in (1). (However, it should be remembered that the variance in the limiting distribution is not always equal to the limit of the variance.) This analysis is used to conclude that in the preceding sense, the mle $q(\hat{\theta})$ is an asymptotically optimal (most *efficient*) estimator of $q(\theta)$. (There are other asymptotically optimal estimators, e.g. Bayes estimators.) Note that from the above it can be concluded that the standard error of $q(\hat{\theta})$ is $|q'(\hat{\theta})|/[nI(\hat{\theta})]^{1/2}$. The above considerations can be extended to the case of multiple parameters $\theta_1,\ldots,\theta_m$, (e.g. [6]) and to semiparametric models, (e.g. [2]).

References

[1] Bickel, P.J. & Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall, New Jersey.

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[3] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[4] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[5] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

[6] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[7] Normand, S.-L. & Doksum, K.A. (1997). Gaussian models for degradation processes. Part II: Analysis of biomarker data, in *Lifetime Data Analysis*, to appear.

(*See also* **Asymptotic Relative Efficiency (ARE)**; **Confidence Intervals and Sets**; **Estimating Functions**; **Estimation, Interval**; **Generalized Estimating Equations**; **Generalized Maximum Likelihood**; **Inference**; **Sufficiency**)

KJELL A. DOKSUM

Ethics of Randomized Trials

Ethical concerns about human experimentation have been of concern for thousands of years [3, 16]. However, since the Nazi atrocities [12] in the name of medical research, there has been a resurgence of ethical concerns. This has resulted in enactment of codes, guidelines, and laws, including the Nuremberg Code [15], the Declaration of Helsinki [5], and Federal Regulations [7]. In addition, there has been ongoing concern for the ethical dimensions of the patient and physician relationship; as evidenced, for example, by the Hippocratic oath [14]. The ethics of randomized clinical trials (RCTs) deal with a subset of concerns about human experimentation. This article addresses both general and specific ethical concerns.

Philosophical Basis

There are no universally accepted ethical principles to guide experimentation and RCTs in humans, neither through religious revelation and dogma nor through any philosophical system. However, all thoughtful commentators agree that there is a need for guidance. Even in more recent times, there have been examples of behavior that most commentators consider unethical. For example, the Tuskegee study [8] withheld the newly discovered cure for syphilis and continued to observe the natural history of syphilis on humans without obtaining their consent.

In addition to religious revelation, there are two primary approaches to human ethics. During the eighteenth and nineteenth centuries, David Hume, Jeremy Bentham, and John Stuart Mill, among others, developed Utilitarian philosophy as a method of determining appropriate standards for human behavior [2]. This philosophy is sometimes summarized as obtaining the greatest good for the greatest number, or maximizing the sum of human happiness. Utilitarian philosophies advocate acts that result in increasing some measure of utility for the entire human population. At one end of the Utilitarian spectrum each act is assessed for the results for the entire human population (possibly including those yet unborn). Other Utilitarian philosophers argue that it is impracticable to evaluate each act; principles that approximately result in optimal utility should be used. Under most Utilitarian philosophies, RCTs result in good for the greater number of individuals and are justified.

Other, deontologic ethical systems advocate general principles that are appropriate for human behavior. For example, Beauchamp & Childress [2] present four principles for biomedical ethics. The principle of *autonomy* states that there is inherent value in allowing each individual to act as an independent being. The principle of *beneficence* states that one should act for the good of others. The principle of *nonmalfeasance* states that one should avoid doing harm to others. The principle of *justice* states that one should act in a just manner. In the clinical trial area, Pappworth [13] suggests the principle of equality: the investigator should be willing to participate or have family members participate under similar conditions.

Specific codes for the physician have a long history, including the Hippocratic oath [14]. Current important codes for the medical community include the Nuremberg code [15] and the Declaration of Helsinki of the World Medical Congress [5]. Both agree that subjects should be informed of experimentation, that the physician's first responsibility is to her/his patient, and that experiments must not expose subjects to undue risk that is not commensurate with the gain from the experiment. These principles may also be argued from the ethical principles given above [2]. Specific possible implications and dilemmas of current ethical standards are now considered.

Informed Consent

Informed consent was first introduced as a term in case law in 1957 [6], but was emphasized (without the terminology) in the Nuremberg Code [15]. The roots go back at least as far as 1900, when the Prussian minister for religious, educational, and medical affairs issued a directive that the subject had to give "unambiguous consent" after a "proper explanation of the possible negative consequences . . ." [19]. In most countries, informed consent is now a matter of law. In practice, the issue of informed consent is a difficult one, since communication about complex scientific, biologic, medical material is problematic even in the best of circumstances. There are exceptions to requiring informed consent of the experimental subject. In some countries, for minors and individuals who cannot reasonably give an informed consent, those with power of attorney are allowed to give consent. In

other circumstances, where time and situation do not permit informed consent (e.g. during cardiac arrest) separate rules have been developed in some countries to allow research (including RCTs) after appropriate review. Zelen [20] has proposed a design for a standard therapy against a new therapeutic arm. He argues that it is ethical to randomize before talking to the subjects; only those randomized to the new therapy need to give informed consent, since the others would have received the standard therapy if there had been no trial. The ethics of the design *vis-à-vis* informed consent have been considered questionable.

Review of Ethics

A second procedure for protecting patient and subject safety is to require an independent body to review and approve a study before it can begin. In the United States, this independent body is an institutional review board (IRB). For trials under the purview of the US **Food and Drug Administration** (FDA), such boards must satisfy a number of federal regulations (that carry the force of the enabling law [7]). In addition, experiments involving new drugs, biologics (that is, drugs that are compounds naturally occurring in the human body), and devices must have the experimental protocol approved by the FDA as well as an IRB before implementation. In approving experiments with new compounds, both law and the Declaration of Helsinki [5] mandate that appropriate **preclinical** (basic science and animal) studies be performed before beginning human experimentation.

Physician–Scientist Conflict

The most difficult and extensively argued ethical debate about clinical trials and medical experimentation is the real or perceived conflict between the physician's duty to put the patient's welfare and treatment first and the scientist's desire to obtain adequate and well-controlled experimental data. In a RCT the physician has delegated to a statistical, or random, assignment his role of collaboratively deciding with the patient the best diagnostic or therapeutic approach (*see* **Randomization**; **Randomized Treatment Assignment**). To some this is viewed as a potentially morally unacceptable role for the physician; for example, Hellman & Hellman [11] mention that the conflicting physician roles arise "from the classic conflict between rights-based moral theories and utilitarian ones". Some argue that randomization is appropriate if the physician is in *equipoise*; that is, the physician truly has no preference as to the best treatment. It has also been argued that it is enough that there be "clinical equipoise" [10]; that is, "if there is genuine uncertainty within the expert medical community – not necessarily on the part of the individual investigator – about the preferred medical treatment". A large group of **AIDS** researchers has advocated the use of "the uncertainty principle" [4]:

> Patients and physicians can be encouraged to participate in randomized clinical trials by adhering to the uncertainty principle and understanding the limited circumstances in which randomization may not be appropriate. By the uncertainty principle, we mean the principle that a trial should be open only to patients for whom the choice of a treatment remains substantially uncertain.

In addition, if a treatment is known but the consequences of no treatment are reversible and moderate – for example, a headache, with the possibility of rescue medication – one can argue under certain circumstances that the risk–benefit ratio still allows a randomized trial.

Approximate equipoise may be considered a reasonable requirement to *begin* a trial. However, as any small amount of additional evidence becomes available an exact individual equipoise would be lost. Thus complete equipoise cannot reasonably be maintained throughout a RCT, even though it holds at the beginning of the trial. For this reason, RCT investigators must address the potential conflict between a health researcher's or physician's responsibilities to her or his patient and the physician's responsibilities as a scientist. If a physician has an overriding obligation in every circumstance to recommend the "best therapy", no matter how lacking or slender the scientific evidence, then most RCTs are probably unethical. Arguably, such a stance is not good for society as a whole; but proponents of this view argue that ethical considerations of the patient–physician relationship, or implied contract, must take precedence. On the other hand, if, lacking a reasonable standard of evidence about the risk–benefit ratio of a proposed therapy, it is reasonable to perform human experimentation, then many RCTs are appropriate and ethical. Trials may then continue until a reasonable standard of evidence shows that one of the assigned treatments is efficacious or harmful. Of course, such

trials must have appropriate informed consent. Where informed consent is inappropriate (e.g. with some mental states) or impossible (e.g. in a subject undergoing a cardiac arrest), there is an even greater need for independent review of its ethics. Other RCTs may be appropriate when the risk to the subjects is small compared to the potential gain from the experiment.

Many countries have decided *de facto* that RCTs are ethical if the objective level of scientific evidence about a potential therapy is lacking. In these countries drug, biologic, or device laws restrict the physician's choices until a new therapeutic or diagnostic method has adequate scientific evaluation. Depending upon one's focus, the medical declarations seem to favor either RCTs or the primacy of the physician–patient relationship. For example, the widely quoted and used Declaration of Helsinki [5] notes:

> The Declaration of Geneva of the World Medical Association binds the physicians with the words. "The health of my patient will be my first consideration." The International Code of Medical Ethics declares that, "A physician shall act only in the patient's interest when providing medical care which might have the effect of weakening the physical and mental condition of the patient".

Yet the Declaration of Helsinki also goes on to give principles for medical experimentation and notes that the research "cannot be legitimately carried out unless the importance of the objective is in proportion to the inherent risk to the subject". The tension between what has been called the "clinical imperative" and research seems unlikely to be totally resolvable. For this reason, safeguards about independent review of research proposals and informed consent must be taken most seriously. All individuals, including statisticians, involved in such clinical research have an obligation to insure an ethical RCT (to the extent that they may reasonably be expected to understand the ethical issues).

The theme of the clinical imperative versus the need of society for reasonable evidence of efficacy and safety for medical diagnostic and therapeutic methods has analogs in other areas. Similar balancing of needs between the individual and society exist for eminent domain (the government's right to take property for public use, usually with compensation), mandatory immunization, mandatory service in the armed services, restrictions on smoking, reportable diseases, quarantine, and so on. However, in medical research the issue is more difficult, because of the dual role and commitment of the physician/scientist.

Placebo Treatment Arms

Another related major area of ethical concern in clinical trials is the use of placebo treatment arms (*see* **Blinding or Masking**). Some commentators assert that when there is a known therapy of value, it is unethical to use a placebo (see, for example, [17]). Many scientists involved in RCTs feel that placebos may ethically be used when any adverse outcomes are reversible and the potential gain from the experiment justifies the risk or discomfort to the experimental subjects. For example, trials of analgesics for headache pain or upset stomach and antihypertensive drug trials often use placebos with rescue allowed or required for severe or protracted pain or blood pressures above fixed levels. Because of the many well-known difficulties with active control trials, it has been suggested that "if it is ethical to use a placebo it is unethical not to use one" [9]. There is a consensus that placebos are not ethical in most circumstances when there is a known beneficial treatment that prevents a serious irreversible adverse outcome: in this case active control trials are used.

Trial Monitoring

There is an ethical consensus that trials with a serious, irreversible endpoint must not continue when one treatment of a trial is shown to be superior. The Nuremberg Code [15] states:

> During the course of the experiment the scientist in charge must be prepared to terminate the experiment at any stage, if he has probable cause to believe, in the exercise of the good faith, superior skill and careful judgment required of him that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.

The Declaration of Helsinki [5] states that "Physicians should cease any investigation if the hazards are found to outweigh the potential benefits". In RCT practice, this means that trials with serious, irreversible endpoints must stop when one therapy is "proven" to be superior. This had led to boards or committees, often called Data and Safety Monitoring Boards (DSMBs), that evaluate the study data as the

accumulate. As a result there is a substantial literature on stopping rules for clinical trials (*see* **Data and Safety Monitoring**). Because investigators delegate their ethical responsibility to monitor accumulating data for patient harm or benefit, the ethical responsibility and pressure is assumed by such boards. By current guidelines, the DSMBs consider a therapy not proven to be efficacious or harmful when the trial data are very close to reaching a stopping criteria but have not reached a stopping criteria, but consider the result as decisive with a small amount of additional data that are sufficient to satisfy such a criterion (with all other factors supporting this interpretation). It should be emphasized that the statistical stopping rules are only guidelines; a DSMB must consider all relevant information (including, but not restricted to, other trials of the same treatment or related treatments, biologic reasoning and plausibility, and other possible explanations, such as imbalance between treatment arms, unblinding, and so on).

Experimental Design

The Nuremberg Code [15] asserts that "The experiment should be such as to yield fruitful results for the good of society ...". It has been argued that this implies that an experiment must have a sound **experimental design**, including appropriate statistical analysis plans and adequate statistical power. Others feel that pilot studies do not need to have enough statistical power to meet the usual standards of proof under important **alternative hypothesis**. As a member of the research team in a RCT, the biostatistician should insure an appropriate experimental design.

Professional Conduct

Both the **American Statistical Association** [1] and the **Royal Statistical Society** [18] publish guidelines for the practice of statistics. Both guidelines address the need to respect privacy, although the majority of the points made refer to more general statistical practice. The Declaration of Helsinki [5] states that "In publication of the results of his or her research, the physician is obliged to preserve the accuracy of the results". All of the ethical strictures on the collection, analysis and reporting of data [1, 18] hold for RCT data (*see* **Data Management and Coordination**). This is particularly important, since there are often powerful scientific career or financial pressures on biostatisticians involved in the conduct, analysis, presentation, and review of RCT results.

References

[1] American Statistical Association (1995). *Ethical Guidelines for Statistical Practice*. American Statistical Association. Alexandria.

[2] Beauchamp, T.L. & Childress, J.F. (1989). *Principles of Biomedical Ethics*. Oxford University Press, Oxford.

[3] Bull, J.P. (1959). The historical development of clinical therapeutic trials, *Journal of Chronic Diseases* **10**, 218–248.

[4] Byar, D.P., Schoenfeld, D.A., Green, S.B. et al. (1990). Design considerations for AIDS trials, *New England Journal of Medicine* **323**, 1343–1347.

[5] Declaration of Helsinki: Recommendations Guiding Physicians in Biomedical Research Involving Human Subjects (1983). As amended in October 1983.

[6] Faden, R.R. & Beauchamp, T.L. (1986). *A History and Theory of Informed Consent*. Oxford University Press, New York, p. 87.

[7] Federal Regulations (1996). 21 CFR, Chapter 1.

[8] Final Report of the Tuskegee Syphilis Study Ad Hoc Advisory Council (1973). US Public Health Service, Washington, DC (pp. 5–15 given in reference [16]).

[9] Fisher, L.D. (1996). Personal communication.

[10] Freedman, B. (1987). Equipoise and the ethics of clinical research, *New England Journal of Medicine* **317**, 141–145.

[11] Hellman, S. & Hellman, D.S. (1991). Of mice but not men: problems of the randomized clinical trial, *New England Journal of Medicine* **324**, 1585–1589.

[12] Lifton, R.J. (1986). *The Nazi Doctors*. Basic Books, New York.

[13] Pappworth, H.M. (1967). *Human Guinea Pigs: Experimentation on Man*. Oxford University Press, Oxford, pp. 185–212.

[14] Hippocratic oath (1977),. in S.J. Reiser, A.J. Dyck & W.J. Curran, eds. (1977) *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*. MIT Press, Cambridge, Mass, p. 5.

[15] Nuremberg Code (1949). In *Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law*, No. 10, Vol. 2. Government Printing Office, Washington, DC, pp. 181–182.

[16] Reiser, S.J., Dyck, A.J. & Curran, W.J., eds (1977). *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*. MIT Press, Cambridge, Mass.

[17] Rothman, K.J. & Michels, K.B. (1994). The continuing unethical use of placebo controls, *New England Journal of Medicine* **331**, 394–398.

[18] Royal Statistical Society (1995). *The Royal Statistical Society Code of Conduct*. Royal Statistical Society, London.

[19] Vollman, J. & Winau, R. (1996). Informed consent in human experimentation before the Nuremberg code, *British Medical Journal* **313**, 1445–1447.

[20] Zelen, M. (1970). A new design for randomized clinical trials, *New England Journal of Medicine* **300**, 1242–1245.

(*See also* **Clinical Trials Audit and Quality Control**; **Clinical Trials, Overview**)

LLOYD D. FISHER

Ethnic Groups

The word "ethnic" is defined in *Webster's Unabridged Dictionary* as "designating or of any of the basic divisions or groups of mankind, as distinguished by customs, characteristics, language, etc.". Technically, the word "racial" has a more biological or genetic definition. However, by custom, these terms have come to be used interchangeably. Since the issues related to study design and data analysis are the same for either "ethnicity" or "race", this distinction is not important for the present discussion. In most contexts, an individual's ethnicity is either self-declared or assigned on the basis of nativity, ancestral origin, or some other observable characteristic.

Ethnicity is an important variable in human studies because of the substantial differences in risk among ethnic groups for many chronic diseases, both infectious (e.g. hepatitis B, AIDS) and noninfectious (e.g. heart disease, cancer). These ethnic disparities may be explained by differences in the prevalence, intensity or type of exposures to environmental or behavioral factors, such as cigarette smoking or diet, that are causally related to the disease. They may also be due to certain inherited genetic variations (*see* **Genetic Epidemiology**) acting as modifiers of these associations (e.g. in genes involved in metabolic activation or detoxification of drugs and carcinogens), or they may result from an increased prevalence of a major susceptibility gene in a particular ethnic group (e.g. the breast cancer 1 gene in Ashkenazi Jews). Therefore, ethnicity may be considered as either a **confounder** or as an **effect modifier** in research studies.

However, ethnicity is generally not a simple variable to manage. It is usually classified as a **nominal** variable, although division into degrees of adherence to cultural traditions or extent of genetic admixture is sometimes possible. Data sources often agglomerate distinct ethnic groups into assemblages that are not meaningful for a particular study, yet cannot be disaggregated. For example, the US census combines Asians with Pacific Islanders, each of which is itself a heterogeneous mix of genetic and cultural groups. Assessment of ethnicity can lead to substantial **misclassification**, because individuals of mixed ancestry may only partially identify their origins, or, if the characterization is based on appearance, it may be misassigned. A further problem when using existing data sources in an analysis is lack of comparability in definition between two or more sources of ethnic information. For example, the computation of cancer incidence rates utilizes data from tumor registries (*see* **Disease Registers**) for the numerators and **census** data for the denominators; the basis for classifying individuals in these two sources often varies.

Despite the difficulties in accurately assigning ethnicity, research studies can often be enhanced by including different ethnic groups. For example, by doing so, the extent of variation in exposure variables is generally increased, and this can be helpful to the research in appropriately combined analyses. An examination of similarities and differences across ethnic groups may also offer new clues to etiology, and may help to elucidate the role of genetic factors in disease occurrence. Furthermore, ethnic-specific analyses are sometimes important for public health, economic, or other nonbiological reasons.

The way ethnicity is dealt with in study design and data analysis depends on whether it is considered as a confounder or an effect modifier, and on the sample composition. To begin with, if few in number, subjects of some ethnic backgrounds may be excluded altogether. Sample size requirements for the ethnic groups studied will differ depending on whether ethnicity is considered as an adjustment variable or an interaction variable. Also at the design stage, a matched **case–control** design or oversampling may be used to ensure balanced representation of smaller ethnic groups, a desired feature even if ethnicity is merely an adjustment variable. If an unmatched design is preferred, or in a prospective study, ethnicity is usually adjusted for in the analysis, either by **stratification** or by introducing dummy variables in a **multivariate analysis**. However, the danger of **overmatching** or overadjusting must be considered when ethnicity is thought to be on the same etiologic pathway as the dependent variable. An unmatched design must also be used if a residual association with ethnicity is to be tested for, after adjustment for other independent risk factors.

In summary, ethnicity is an important variable for classifying individuals in biomedical research. It is often an indicator of group exposure or genetic differences that may be important in disease causation. It is also a major confounding factor, which can be adjusted for either in the design or analysis stage of a study.

L.N. KOLONEL & L. LE MARCHAND

Eugenics

The study of human biology as an intellectual and academic discipline expanded vigorously during the late nineteenth century, largely in response to the publication of Charles Darwin's *On the Origin of Species by Means of Natural Selection* in 1859 and *The Descent of Man* in 1871. Two significant consequences of the introduction of evolutionary theory were the birth of the Eugenics Movement and the emergence of the concept of social Darwinism. The Eugenics Movement was founded by **Francis Galton**, **Karl Pearson**, and their colleagues at the Galton Laboratory for National Eugenics in London. These investigators concentrated their inquiries on the perpetuation of traits that they held to be preferable in the human species. These traits, under the idea of social Darwinism, included personal qualities of "talent" and "natural ability" observed more frequently in classes of persons regarded as socially "fit". Excluded were the many undesirable traits that abounded among lower social classes that were regarded as socially "unfit". Galton coined the term "eugenics" in 1883, deriving it from Greek roots meaning "good in birth", and intending the word to describe a scientific inquiry into describing and improving the genetic endowment of man [13].

The theory and practice of eugenics quickly expanded to the US. In 1910 the leading American eugenicist, Charles B. Davenport, founded the Eugenics Record Office (ERO) at Cold Spring Harbor, New York, and installed Harry H. Laughlin as the director. The ERO became the center of eugenic research in the US and was generously funded from numerous sources, including the fortunes of the Rockefeller, Harriman, and Carnegie families [13]. American eugenic research followed two distinct but related lines of inquiry: first, the documentation of the transmission of biological and social traits in succeeding generations of particular families, and, secondly, the development of social and scientific policies and legislation that would support biological solutions to the unwanted perpetuation of undesirable human traits.

The activities of both the British and the American schools of eugenics proceeded in tandem with the development of eugenic theory in Germany. The unique nature of German eugenics as a social force rested in large measure on the position of most eugenicists as members of the respected and revered university-based academic elite. Although many supporters of German eugenics were not affiliated with university communities [15], the strength of the movement was solidly established in the German universities – a fact that afforded credibility and acceptability to professional writings and public policies about the inequalities of man [10].

As the study of eugenics expanded, several lines of inquiry were pursued at different times, in various locations. These more circumscribed areas of research included, early on, intense investigations of the physical traits of man, accompanied by interpretations about the relationship between physical attributes and social qualities. Somewhat later, numerous social traits received close attention and analysis, as did mental traits, including intelligence, mental retardation, and several types of mental illness. As the Eugenics Movement reached its peak, underlying theories and historical policies became the justification for "directed medical killing" [14], barbarous medical experimentation [17], and racial genocide [16].

Physical Traits

Curiosity about hereditary physical differences among human beings has been documented since the writings of the ancient Greeks, including Hippocrates, who noted, for example, the familial nature of baldness, blue eyes, and "long" heads [20]. As scientists raced to investigate the implications of social Darwinism in the last decades of the nineteenth century, they amassed enormous sets of data on the size of the brain in humans and primates and were able to conclude that the larger brains "in men than women, in eminent men than in men of mediocre talent", reflected the superior intelligence of talented men. The practice of weighing brains evolved in the late nineteenth century into more sophisticated measurements of myriad body parts as the science of **anthropometry** emerged. Calipers became standard investigative tools as investigators sought to correlate physical characteristics with social traits, and anthropometry enjoyed a period of dubious glory in the late nineteenth century with the studies of the Italian anthropologist Cesare Lombroso, who concluded that "[c]riminals are the apes in our midst, marked by anatomical stigmata of atavism", such that criminality and guilt could be reasonably predicted from

physical measurements in certain groups including the handicapped and Gypsies [11]. These early methods found new applications in the 1930s and 1940s when the Nazis used anthropometry to support racial classifications and segregation.

Social Traits

The leap from anthropometry and criminality to a plethora of other social characteristics was a short one. At the time when Lombroso was measuring the bodies of criminals and promoting the elimination of these persons from society, eugenicists in the Galton Laboratory were noting the concentration of talent among the educated classes of the British population. Decades before the rediscovery of **Mendel's laws** of single gene inheritance, Galton, in 1869, published his concepts of eugenics and the origins of "natural ability" under the title, *Hereditary Genius: An Inquiry into Its Laws and Consequences* [13]. British investigators noted two centuries of numerous blood relatives in the ranks of jurists and statesmen, artists and musicians, politicians and military leaders, and scientists and humanists. These observations were used to support conclusions about the hereditary superiority of these more accomplished classes of human beings.

While the Galton Laboratory collected data on those who excelled, eugenicists on both sides of the Atlantic were documenting the familial nature of a number of less attractive social traits in the human population. The most notable account of social pathology was the study of seven generations of the Juke family of New York, published by Richard Dugdale in 1877, and confirmed in a follow-up study some 40 years later [8]. The Jukes were documented to be a family of criminals, prostitutes, and social misfits, characterized by personal traits of "feeblemindedness, indolence, licentiousness, and dishonesty". Similar studies of other social traits expanded the reach of genetics and social Darwinism to alcoholism, pauperism, insanity, and communicable social diseases. The possibility that traits of social degeneracy could be related to the dismal living circumstances of legions of families in the early twentieth century was either unrecognized or ignored as the proponents of eugenic theory looked to future control of the propagation of these unfortunate human groups.

From an analytic point of view, the major difference between British and American approaches was the biometric approach of the British in their analysis of metric traits, contrasted to the intense compulsion of the Americans to document a monogenic etiology of familial social traits. Of importance to the followers of Galton were principles of **correlation** and **regression to the mean**, both of which added strength to ideas of hereditary social degeneracy and decline. American eugenics, however, found support in suggestions that complex traits, or qualities, could be determined by simple modes of single gene inheritance. Thus, for example, the tendency of seafaring fathers to have seafaring sons was attributed to a single gene for a trait that was designated thalassophilia, as was nomadism, or the impulse to wander, among the Comanches, the Gypsies, and the Huns.

Mental Traits

In addition to physical and social traits, the familial nature of mental traits was also scrutinized by researchers in eugenics as well as by investigators in psychology and education. Early measures of intelligence were developed by Alfred Binet, a French psychologist, and later refined by British and American investigators, to produce the intelligence quotient (IQ), a measure initially assumed to represent an innate, immutable quality of individual mental acuity. Early studies of intelligence in twins, with application of methods of factor analysis, added credence to claims of the genetic determination of intelligence, an idea supported by the widely read, but fraudulent, reports of Cyril Burt, who claimed to have studied 53 twin pairs who in fact had never existed. Significant applications of intelligence testing in the American military suggested that most white American males were of very low intelligence, and tests used in American immigration procedures purported to prove that immigrants from southern Europe and eastern Mediterranean countries were of consistently lower intelligence than immigrants from northern Europe. More recent studies continue to support claims of substantial genetic differences among races in the development of intelligence [12].

Tests of mental capacity were also used to define the mental status of persons who were mentally retarded, and individual scores were used to assign

the retarded to groups of morons, imbeciles, or idiots, depending on the degree of mental deficiency. (This language was later replaced by less pejorative terms of mild, severe, and profound retardation.) Retarded children born to parents of limited mental capacity lent credence to claims of the predominantly hereditary nature of mental deficiency, although some studies continued to indicate the significant role of environmental influences in the development of mental abilities. Complex and conflicting evidence for monogenic, polygenic, and environmental contributions to the phenotypes of mental retardation led to exhaustive studies by Lionel Penrose in the 1930s into the sources of retardation among the patients at the Royal Eastern Counties' Institution in Colchester, England. Penrose confirmed the hereditary nature of some monogenic diseases, including phenylketonuria, congenital hypothyroidism, Huntington's disease and others. He also noted the similarities among patients with Down's syndrome and the increased frequency of these children born to older mothers. But, he noted, most patients appeared to suffer from retardation that arose in combinations of environmental, pathological, and genetic factors, and he cautioned against facile labeling of patients without firm justification.

In contrast to the diverse etiologies of mental retardation, mental illness includes a number of organic pathological processes that underlie such diseases as schizophrenia, depression, and manic-depression. While early family studies of these illnesses noted significant clustering in certain families, suggesting single gene inheritance, later introduction of sound epidemiologic principles and correction for ascertainment bias have reduced the figures for disease frequency and recurrence risks and have implied multifactorial etiologies for these disorders [13]. Mental illnesses have also been the subject of twin studies (*see* **Twin Analysis**) that have generally verified a significant genetic contribution to disease phenotypes, as well as a greater risk to closely related persons who share similar environments [20].

Providing care for the mentally retarded and mentally ill has progressed hesitantly over much of Western history. In ancient times, the burden of caring for defective infants was often eliminated when the infants were exposed and abandoned. In later centuries, placing the backward child in the care of an incompetent nurse often hastened the end of the child's existence [7]. Society later assumed the responsibility of caring for these unfortunate persons, although early institutional care consisted of little more than providing separate warehouses for male and female patients. As the Eugenics Movement gathered momentum, however, proponents of improving the genetic endowment of man advocated more permanent methods of genetic control among the mentally disabled, particularly surgical sterilization of both men and women whose continuing fertility could result in further generations of mentally deficient persons. These suggestions became the subject of legislation in some 28 American states before 1931, and tens of thousands of mental patients were forcibly sterilized under these statutes [18]. The propriety of sterilization programs for eugenic purposes was reinforced in 1927 when the United States Supreme Court upheld a Virginia sterilization statute by noting that "three generations of imbeciles are enough" [5]. These programs continued into the 1960s, long after the Supreme Court reexamined the issue of reproductive rights in 1942 and found the right to procreate to be a fundamental right, deserving of the strictest constitutional protection [19]. Finally, during the years of legislatively sanctioned sterilization programs, voluntary sterilization was generally encouraged among families in the general public who had mentally retarded or otherwise severely deficient children or were at risk of having such children in the future.

Racial and Ethnic Traits

From its inception the Eugenics Movement was dedicated to the proposition of ranking and valuing human groups on the basis of racial and ethnic characteristics, although distinctions among racial, ethnic, and even national origins were often indistinct. Pearson, for example, noted that Jewish youth in London's East End were quite as intelligent as Gentiles but tended to be physically inferior and somewhat dirtier. Davenport regarded the Poles, the Irish, the Italians, and the "Hebrews" as distinct racial groups. He further concluded that the Poles were "independent and self-reliant though clannish", that the Italians were prone to "crimes of personal violence", and that the Hebrews were "intermediate between the slovenly Serbians and Greeks and the tidy Swedes, Germans, and Bohemians" though given to "thieving". These types of quasi-cultural distinctions were ultimately compiled, along with exhaustive biological

traits, by the German eugenicists Erwin Baur, Eugen Fischer, and Fritz Lenz, who noted, for example, that the "Mediterranean" race was typified by "narrow, long skulls, like the Nordic, but smaller and somewhat wider and steeper forehead", very dark brown to black hair and eyes, and significantly brownish skin; furthermore, this group had "neither the quiet industry of the Mongoloid race nor the initiative and energy of the Nordic people". With respect to Germanic peoples, these authors further noted that the "sturdy, blond race" of Westfalen and Schwaben could be distinguished by their clumsiness, stubbornness, rigidity, and sedentary nature, compared with the "slim, blond race" of more Nordic descent that was more inclined toward thought, discovery, nature, and beauty of form [3]. Both American and European eugenicists consistently agreed that the Negro race represented the nether end of the human racial scale, while the white, Anglo-Saxon, Protestant population of northern Europe represented the fortunate epitome of human evolution. These and thousands of other distinctions, delineated in professional publications and in the popular press, gave substance to human differences and were the foundation for further judgments about the relative value of various groups in the human "hierarchy" [13].

Racial and ethnic distinctions, whether based in fact or fiction, became the foundation for vigorous campaigns to limit immigration, particularly from southern and eastern Europe into the US. Davenport was concerned that the influx of immigrants from these areas, their tendency to have large families, and their subsequent marriage into the American population, would result in making the American population "darker in pigmentation, smaller in stature, more mercurial ... more given to crimes of larceny, kidnapping, assault, murder, rape, and sex-immorality". So fervent was this fear of racial decline that a strong lobby, led by Laughlin, developed for the single purpose of restricting immigration into the US by establishing quotas for immigration from individual European countries. The immigration lobby succeeded in introducing federal legislation that resulted in the Immigration Act of 1924, overwhelmingly passed by Congress and signed by President Calvin Coolidge, who was already on record as noting that "America must be kept American. Biological laws show ... that Nordics deteriorate when mixed with other races" [13] (*see* **Admixture in Human Populations**).

Excesses of National Socialism in Germany

The development of eugenics in Germany during the 1930s and 1940s was the result of a confluence of social and political forces that had been developing for several decades. On the practical side, for example, the German economy had never recovered from the crippling penalties imposed at the end of World War I, and German eugenicists were careful to emphasize, at every opportunity, the interminable and substantial costs of housing and caring for persons with inherited or acquired disabilities and diseases. On the theoretical and intellectual side, from a negative perspective, the university community read and absorbed the early writings of Adolf Jost, who argued that the German "Volk", as a racial–cultural entity, should have the right to kill some individuals in order to preserve the health of the body of the whole Volk. Somewhat later, Karl Binding and Alfred Hoche, a jurist and a psychiatrist, argued that the state had a right, and perhaps even a duty, to end the lives of people whose "lives were not worthy of life" [4]. Finally, among the innumerable racist and eugenic arguments included in *Mein Kampf* were declarations that the fiscal woes of Germany could, in large measure, be ascribed to the Jews, and that a person's right to life ends when he is no longer capable of fighting for his own health. From a positive perspective, both academic eugenicists as well as the Nazis argued the physical and mental superiority of the Nordic, or "Aryan", race and proposed programs that would encourage qualified individuals and couples to produce children who would be a credit to their German Vaterland.

The academic thrust of eugenics was well established by the time the Nazis assumed power in 1933. Thereafter, the government implemented programs that for the first time legally separated undesirable persons from the body of society through restrictive social legislation about marriage and occupational licenses. Separations from society continued to escalate through legislation and clandestine directives until, in the world of medicine, the mentally retarded and the mentally ill who were cared for in public institutions were targeted for destruction. These patients had such severe disabilities that they were not expected to return to productive lives in German society: these were the "lebensunwertig", the lives "not worthy of life" [1]. As these groups disappeared, the next step was the

annihilation of institutionalized Jews who, though mentally ill, had favorable prognoses for recovery. Medical and anthropological interests soon reached into the concentration camps that filled rapidly with the unwanted in German society, including political detainees, Jews, Gypsies, homosexuals, convicted criminals, and, later, prisoners of war. These vast assemblies provided material for anatomy collections and descriptive studies at major universities and for numerous medical experiments, including investigations of sterilization methods, infectious diseases, the effects of poisons, and the effects of high altitude and extreme cold. The meticulous system of documentation in the German bureaucracy later provided exhaustive information about the nature and extent of Nazi biomedical research and experimentation [6].

The atrocities committed during the Third Reich were the subject of criminal prosecution during the Nuremberg trials, and many of the more prominent leaders of National Socialism were sentenced to death or to long prison terms. While most of the current community of bioethics and medical research continues to acknowledge revulsion at the horrors of Nazi research and experimentation, professionals in many disciplines remain divided over the issue of using various sets of Nazi data in contemporary studies. One conflict that may never enjoy a resolution is whether using these data, on the one hand, gives tacit approval to the barbarous circumstances under which the data were collected, or, on the other hand, gives acknowledgement and value to the lives of the subjects who perished in the experiments [2, 9].

Contemporary Perspective

The Eugenics Movement has a long record of intellectual and racist elitism that in recent decades has lost much of its earlier support. The Movement was founded and nourished by intellectual leaders in Europe and the US, and it enjoyed strong support from persons of political, financial, and social advantage. Early contributions and support from such disciplines as biometrics, genetics, anthropology, evolutionary biology, medicine, psychology, and sociology provided credence to concepts that eventually became major contributing factors to vast programs of social and racial genocide. As the world discovered the unspeakable truth of the Holocaust, the word "eugenics" acquired a tone of terror in the interrelationships of man, a tone that constantly implied the obscene violation of human rights. Only recently, however, have researchers in genetics and bioethics begun to distinguish between eugenics imposed by government and eugenics practiced by individual choice. The former was the source of forced sterilizations, legislation designed to protect "racial purity", programs of "directed medical killing", and ultimately the practice of mass murder. The latter, however, is an integral part of contemporary programs of nondirective **genetic counseling** that are dedicated to moral and legal principles of personal autonomy and individual choice in issues of human reproduction. Eugenics has come full circle, from genuine, though misguided, concern about the health of the human race, through brutal programs of cleansing society of its undesirable elements, to a new concern for human health that allows personal choice about bearing children with genetic impairments.

References

[1] Aly, G., Chroust, P. & Pross, C. (1994). *Cleansing the Fatherland: Nazi Medicine and Racial Hygiene*. The Johns Hopkins University Press, Baltimore.

[2] Annas, G.J. & Grodin, M.A. (1992). *The Nazi Doctors and the Nuremberg Code: Human Rights and Human Experimentation*. Oxford University Press, New York.

[3] Baur, E., Fischer, E. & Lenz, F. (1927). *Menschliche Erblichkeitslehre und Rassenhygiene*, 3. Auflage, in 2 volumes, A.F. Lehmanns Verlag, München.

[4] Binding, K. & Hoche, A. (1920). *Permitting the Destruction of Unworthy Life: Its Extent and Form (Die Freigabe der Vernichtung lebensunwerten Lebens, Ihr Mass und ihre Form)*. Meiner, Leipzig. Translated by Wright, W.E., Derr, P.G. & Salamon, R. (1991). *Issues in Law and Medicine* **8**, 231–265.

[5] Buck v. Bell. 274 U.S. 200 (1927).

[6] Caplan, A.L., ed. (1992). *When Medicine Went Mad: Bioethics and the Holocaust*. Humana Press, Totowa.

[7] de Mause, L. (1974). *The History of Childhood*. Harper & Row, New York.

[8] Estabrook, A.H. (1916). *The Jukes in 1915*. Carnegie Institution of Washington, Washington.

[9] Faden, R.R. & Beauchamp, T.L. (1986). *A History and Theory of Informed Consent*. Oxford University Press, New York.

[10] Friedlander, H. (1995). *The Origins of Nazi Genocide: From Euthanasia to the Final Solution*. The University of North Carolina Press, Chapel Hill.

[11] Gould, S.J. (1981). *The Mismeasure of Man*. Norton, New York.

[12] Herrnstein, R.J. & Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. The Free Press, New York.

[13] Kevles, D.J. (1985). *In the Name of Eugenics: Genetics and the Uses of Human Heredity*. Alfred A. Knopf, New York.

[14] Lifton, R.J. (1986). *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. Basic Books, New York.

[15] MacIntyre, B. (1992). *Forgotten Fatherland: The Search for Elisabeth Nietzsche*. Farrar Straus Giroux, New York.

[16] Müller-Hill, B. (1988). *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies, and Others Germany 1933–1945*. Oxford University Press, New York.

[17] Proctor, R.N. (1988). *Racial Hygiene: Medicine under the Nazis*. Harvard University Press, Cambridge, Mass.

[18] Reilly, P.R. (1991). *The Surgical Solution: A History of Involuntary Sterilization in the United States*. Johns Hopkins University Press, Baltimore.

[19] Skinner v. Oklahoma 316 U.S. 535 (1942).

[20] Vogel, F. & Motulsky, A.G. (1986). *Human Genetics: Problems and Approaches*. 2nd Ed. Springer-Verlag, Berlin.

MARY Z. PELIAS

European Federation of Statisticians in the Pharmaceutical Industry (EFPSI)

The Federation is open to nationally constituted groups of statisticians who are working in or for the pharmaceutical industry (see www.EFSPI.ORG). Thus, EFSPI is engaged in statistical aspects of research, development, production, and **post-marketing surveillance** of drugs, biologics, and medical devices. The objectives of the Federation are (according to the EFSPI constitution):

1. to promote professional standards of statistics and the standing of the statistical profession in matters pertinent to the European pharmaceutical industry
2. to offer a collective expert input on statistical matters to national and international authorities and organizations.
3. to exchange information on and to harmonize attitudes to the practice of statistics in the European pharmaceutical industry and within the member groups

The Federation was constituted and officially launched in August 1992. There are now eleven member groups representing:

Belgium – BVS/SBS	(Belgishe Verening voor Statistiek/Societe Belge de Statistique) – Biostatistics Section
Denmark – DSBS	(Dansk Selskab for Biofarmaceutisk Statistik)
Finland – SSL	(Statistikot Suomen Lääketeollisuudessa)
France – SFdS	Societé Francaise de Statistique – Biopharmacy and Health Group
Germany – APF	(Arbeitsgruppe Pharmazeutische Forschung at German region of International Biometric Society)
Italy – BIAS	(Biometristi dell'Industria-Farmaceutica Associati)
Netherlands – PSDM	(Workgroup Pharmaceutische Statistiek en Datamanagement van Verenining voor Statistiek – Biometrische Sectie)
Spain – ABC if	(Asociacion de Biometria Clinica para la Investigacion Farmaceutia.
Sweden – FMS	(Foreningen for Medicinsk Statistik)
Switzerland – BBS	(Basler Biometrische Sektion of International Biometric Society)
UK – PSI	(Statisticians in the Pharmaceutical Industry)

Thus, 11 European countries, with a total membership of over 2000, statisticians are represented via EFSPI and can be reached by mailings organized by the respective national groups. Links to the individual member organizations can be found via EFSPIs website at www.EFSPI.ORG.

EFSPI has established a number of special interest groups (including one in nonclinical statistics). EFSPI working parties are set up when appropriate to consider specific topical issues. The question of certification of statisticians and the interpretation of the term "qualified and experienced statistician", a description used in regulatory guidelines, have been considered. The initial results of the working party were published in 1999 in *Drug Information Journal* [1].

The main activities of the Federation, however, have been concerned with guidelines for, and other regulatory aspects (*see* **Drug Approval and Regulation**) of, statistical issues in **clinical trials**. In recent years, this has involved EFSPI input to a number of guidelines including those issued by the Committee for Proprietary Medicinal Products (CPMP) of the European Community and by the **Food and Drug Administration (FDA)** in the US. These have included guidance documents on good clinical practice [2] and on biostatistical methodology in clinical trials, in general or within specific therapeutic areas. Examples are guidance documents on "Bioavailability and Bioequivalence" and on Points to Consider documents on more specific statistical topics such as for example, "Superiority, Noninferiority and Equivalence", "Validity and Interpretation of Meta-Analysis, and one Pivotal Study", "Missing Data and

Multiplicity Issues in Clinical Trials" and "Baseline Covariates" [3] and [4]. In addition, input has been made to guidelines developed under the International Conference on Harmonization (ICH), which has included guidelines on the structure and content of clinical trial reports, on statistical principles for clinical trials, and on the choice of control group. In each case, review comments have been harmonized between the member EFSPI groups and sent to the regulatory agencies and industry associations as a joint statement.

Activities of EFSPI have been presented at the **International Statistical Institute's** (ISI) 49th Session in Florence and 52nd in Helsinki and at Drug Information Association (DIA) meetings, since 1993 in both clinical and nonclinical areas. In collaboration with DIA, EFSPI contributes to suggestions of topics for the clinical and nonclinical DIA workshops.

Cooperation has been achieved with a number of other organizations, including:

1. the EFPIA (European Federation of Pharmaceutical Industries Association), concerning European regulatory issues
2. the FDA and the PhRMA (Pharmaceutical Research and Manufacturers Association) biostatistics subsection, in the US, concerning regulatory harmonization
3. the DIA concerning scientific publications and conferences
4. the ISCB (**International Society for Clinical Biostatistics**).

EFSPI now has a website `www.EFSPI.ORG`.

References

[1] EFSPI working group (1999). Qualified Statisticians in the Pharmaceutical Industry. *Drug Information Journal*, **33**, 407–415.

[2] Zipfel, A. (1995). A European Concept for Good Statistical Practices in Global Drug Development. Report of a DIA/EFSPI workshop. *Drug Information Journal*, **29**, 471–510.

[3] Huitfeldt, B., Danielson, L., Ebbutt A. & Schmidt, K. (2001). Choice of Control in Clinical Trials – Issues and Implications of ICH-E10. On behalf of EFSPI. *Drug Information Journal*, **35**, 1147–1156.

[4] Roes, K.R. (2004). Dynamic allocation as a balancing act. *Pharmaceutical Statistics*.

MICK GODLEY & MERETE JØRGENSEN

European Organization for Research and Treatment of Cancer (EORTC)

The main aim of the European Organisation for Research and Treatment of Cancer (EORTC), which is based in Brussels, is to conduct, develop, coordinate, and stimulate multidisciplinary research in Europe on the experimental and clinical bases of cancer treatment, with the final goal being to develop optimal treatment strategies in order to improve the standards of cancer treatment. It is a scientifically driven organization consisting of a unique pan-European network involving more than 2000 clinical investigators and scientists in some 300 hospitals, laboratories, and research institutions in 31 different countries. In 2003, approximately 6000 new patients were entered into the 130 trials handled at the EORTC Data Center, with more than 800 patients being entered in intergroup trials carried out with other regional, national, or international research groups.

The EORTC was founded in 1962, as an international nonprofit organization under Belgian law, by workers in the main cancer research institutes of the European Common Market countries and Switzerland. In order to provide a central coordination of its **clinical trials**, an EORTC Coordinating Office was established in 1969. Through the efforts of Henri Tagnon and Marvin Zelen and the support of the US National Cancer Institute (*see* **National Institutes of Health (NIH)**), this was transformed into the EORTC Data Center, which was formally launched in January 1974. Located at the Institut Jules Bordet in Brussels, Belgium, it was set up with the help and expertise of a number of American statisticians, including Al Bartolucci, Jim Williams, and Steve George. The last two served as the initial (temporary) directors and principal statisticians. Upon the departure of Steve George in the summer of 1975, Maurice Staquet took over as director, and its first permanent full-time statistician, Richard Sylvester, joined the staff. In 1979, he was named Assistant Director for Biostatistics, a post that he still holds. In 1984, members of the EORTC Data Center edited *Cancer Clinical Trials: Methods and Practice* [1], a very successful book dealing with the methodology of designing, conducting, and analyzing cancer clinical trials.

In 1990, the EORTC Data Center moved to the Brussels campus of the medical school of the Université Catholique de Louvain. A year later in 1991, Françoise Meunier was appointed as Director of the EORTC Central Office/Data Center. In 1995, she was appointed as Director General of the EORTC and Patrick Therasse was appointed as Director of the EORTC Data Center.

Since its inception, the primary role of the EORTC Data Center has been to provide state-of-the-art statistical, medical, **data management**, **quality control**, and computer expertise to some 20 EORTC Clinical Research Groups for the **phase I**, **phase II** and large, **multicenter** phase III clinical trials that they carry out. It has also assumed a major role in **teaching statistics** and the methodology of cancer clinical trials, and has developed Fellowship Programs for the training of statisticians and other scientists.

As of January 2004, there were nine biostatisticians working at the Data Center. Traditionally, its statistical research activities have been applied in nature. In recent years, it has greatly expanded its statistical activities through the development of an applied statistical research program dealing with problems in treatment **outcome research** and **frailty** models, designs for phase II/III trials, nonproportional hazard models (*see* **Survival Analysis, Overview**), design and analysis of **Quality of life** quality of life studies, independent data monitoring committees and interim analyses (*see* **Data Monitoring Committees**), and the design and analysis of biomarker studies (*see* **Molecular Epidemiology**). Statistical methodology and results are discussed in its twice-monthly Stats Club meetings.

Reference

[1] Buyse, M.E., Staquet, M.J. & Sylvester, R.J. eds. (1984). *Cancer Clinical Trials: Methods and Practice*. Oxford University Press, Oxford.

RICHARD SYLVESTER

Event History Analysis

Introduction

Event history analysis deals with data obtained by observing individuals over time, focusing on events occurring for the individuals. Thus, typical outcome data consist of *times of occurrence* of events and of *types of events* that occurred. Frequently, an event may be considered as a *transition* from one state to another and, therefore, *multistate models* will often provide a relevant modeling framework for event history data. Multistate models are discussed from several points of view in the books by Andersen et al. [10], Blossfeld & Rohwer [18], Courgeau & Lelièvre [26] and Hougaard [45, Chapters 5-6] and Kalbfleisch & Prentice [48, Chapter 8]; see [14, 22, 44] for recent survey papers.

Survival Data

The simplest multistate model is the two-state model for **survival** data with one transient state "0: alive" and one absorbing state "1: dead"; see Figure 1. In general, an *absorbing* state is a state from which further transitions cannot occur, while a *transient* state is a state that is not absorbing. The observation for a given individual will here in the simplest form, consist of a random variable, say T, representing the time from a given origin (time 0) to the occurrence of the event "death". The distribution of T may be characterized by the probability distribution function $F(t) = \text{Prob}(T \leq t)$ or, equivalently, by the survival distribution function $S(t) = 1 - F(t) = \text{Prob}(T > t)$. It is seen that $S(t)$ and $F(t)$, respectively, correspond to the probabilities of being in state 0 or 1 at time t. If every individual is assumed to be in state 0 at time 0, then $F(t)$ is also the *transition probability* from state 0 to state 1 for the time interval from 0 to t. In continuous time, the distribution of T may also be characterized by the **hazard rate** *function*

$$\alpha(t) = \frac{-\,\mathrm{d}\log S(t)}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{\text{Prob}(T \leq t + \Delta t \mid T \geq t)}{\Delta t}, \quad (1)$$

that is,

$$S(t) = \exp\left(-\int_0^t \alpha(u)\,\mathrm{d}u\right) \quad (2)$$

(*see* **Survival Distributions and Their Characteristics**).

Thus, $\alpha(\cdot)$ is the *transition intensity* from state 0 to state 1, that is, the instantaneous probability per time unit of going from state 0 to state 1.

In general, event history analysis deals with inference for transition intensities and transition probabilities in multistate models. This includes **estimation** and **hypothesis tests** for these quantities and analysis of **regression** models where these quantities are related to (possibly time-dependent) **explanatory variables** observed for the individuals under study. Most frequently, multistate models are defined by their transition intensities from which transition probabilities may or may not be derived depending on the modeling assumptions. This latter activity is some times denoted *"survival synthesis"*.

A typical feature of event history analysis is the inability to observe complete event histories, for example, by the end of the observation period all individuals under study may not have reached an absorbing state. In survival analysis, this would correspond to individuals still being alive by the end of the study, and this kind of *incomplete observation* is known as *right* **censoring**. Furthermore, all individuals may not have been observed from the same time origin. This kind of incomplete observation where individuals are only observed conditionally on not having reached an absorbing state by the time of initiation of the study is known as *left*-**truncation**. Restricting attention to right censoring, a crucial problem is whether the available incomplete data enables one to make a valid inference on parameters in the multistate model for the complete data. The condition for this is known as *independent* right censoring (*see* **Censored Data**).

Multistate Models

In this section, we will present a number of different multistate models. We will begin with a few heuristic

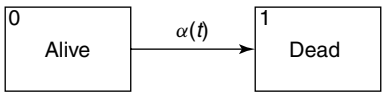

Figure 1 The two-state model for survival data

definitions that may all be made rigorous within the framework of so-called *marked point processes* [10, Chapter III; 16].

A *multistate process* is a **stochastic process** $(X(t), t \in \mathcal{T})$ with a finite *state space* $\mathcal{S} = \{1, \ldots, p\}$ and with right-continuous sample paths: $X(t+) = X(t)$. Here, $\mathcal{T} = [0, \tau]$ or $[0, \tau)$ with $\tau \leq +\infty$. The process has *initial distribution* $\pi_h(0) = \text{Prob}(X(0) = h)$, $h \in \mathcal{S}$. A multistate process $X(\cdot)$ generates a *history* $\mathcal{X}_t$ (a σ-algebra) consisting of the observation of the process in the interval $[0, t]$. Relative to this history, we may define *transition probabilities* by

$$P_{hj}(s, t) = \text{Prob}(X(t) = j \mid X(s) = h, \mathcal{X}_{s-}) \quad (3)$$

for $h, j \in \mathcal{S}, s, t \in \mathcal{T}, s \leq t$ and *transition intensities* by the derivatives

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t, t + \Delta t)}{\Delta t} \quad (4)$$

which we shall assume exist. Some transition intensities may be 0 for all t. Graphically, multistate models may be illustrated using diagrams with boxes representing the states and with arrows between the states representing the possible transitions, that is, the nonzero transition intensities [10, Section I.3; 43]. A state $h \in \mathcal{S}$ is *absorbing* if for all $t \in \mathcal{T}, j \in \mathcal{S}, j \neq h, \alpha_{hj}(t) = 0$; otherwise h is *transient*. The *state probabilities* $\pi_h(t) = \text{Prob}(X(t) = h)$ are given by

$$\pi_h(t) = \sum_{j \in \mathcal{S}} \pi_j(0) P_{jh}(0, t). \quad (5)$$

Notice that the $P_{hj}(\cdot, \cdot)$ and thereby the $\alpha_{hj}(\cdot)$ depend on both the probability measure Prob and on the history though this dependence has been suppressed in the notation. If $\alpha_{hj}(t)$ only depends on the history via the state $h = X(t)$ occupied at t, then the process is **Markovian**. Sometimes, one is interested in considering an extended history that also includes observed **covariates**. If only time-fixed covariates Z are studied, then the observed history is $\mathcal{F}_t = \mathcal{X}_t \vee \mathcal{Z}_0$, whereas **time-dependent covariates** $Z(t)$ give rise to an extended history of the form $\mathcal{F}_t = \mathcal{X}_t \vee \mathcal{Z}_t$ where, in both cases, $\mathcal{Z}_t$ is the history generated by the covariates in $[0, t]$. We shall here focus on the *purely endogenous* case where $\mathcal{Z}_t \subset \mathcal{X}_t \vee \mathcal{Z}_0$; that is, the covariates are either all time-fixed or the random development of the time-dependent covariates is fully specified by the history of the process itself (see, however, the section " Partial model Specification" for cases with time-dependent covariates that are not endogenous).

The Two-state Model for Survival Data

This model, illustrated in Figure 1, has $p = 2$ states and only one possible transition from state 0 to state 1. The corresponding transition intensity $\alpha_{01}(t)$ is given by the hazard rate function $\alpha(t)$, while $\alpha_{10}(t) = 0$ for all t, that is, state 1 is absorbing. The initial distribution is degenerate in 0: $\pi_0(0) = 1$ and the process is Markovian. Covariates may be entered into the model using a regression model for $\alpha(\cdot)$.

The Competing Risks Model

This model (*see* **Competing Risks**) has one transient state "0: alive" and a number, k, of absorbing states, state $h, h = 1, \ldots, k$ corresponding to "death from cause h". Thus, there are $p = k + 1$ states. The model is illustrated for $k = 2$ in Figure 2.

The transition intensities $\alpha_{0h}(t)$ for $h = 1, \ldots, k$ are given by the cause-specific hazard functions:

$$\alpha_h(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(\textit{Dead from cause } h \textit{ by } t + \Delta t \mid T \geq t)}{\Delta t} \quad (6)$$

where T is the survival time. The initial distribution is degenerate in 0, the only transient state of the model, that is, $\alpha_{hj}(t) = 0$ for all $h \neq 0$ and all j.

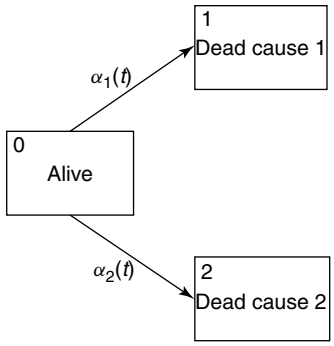

Figure 2 Competing risks model for mortality from 2 causes

The transition probabilities are given by the survival function

$$P_{00}(0,t) = S(t) = \text{Prob}(T > t) = \exp\left(-\int_0^t \sum_{h=1}^{k} \alpha_h(u)\,\mathrm{d}u\right), \quad (7)$$

and the cumulative incidence functions

$$P_{0h}(0,t) = \int_0^t S(u-)\alpha_h(u)\,\mathrm{d}u, \quad h = 1,\ldots,k. \quad (8)$$

Like the simple two-state model ($k = 1$), the competing risks model is Markovian and covariates may be included into the model via regression models for the cause-specific hazards. This model was studied by Andersen et al. [9].

The Illness–death Model

This model (*see* **Aalen–Johansen Estimator**) is illustrated in Figure 3. Often, the time t is the age of the individual, and usually individuals will be assumed to be in state 0 at $t = 0$.

However, individuals will not always be observed from $t = 0$ as shall be further discussed in the sections "Counting Process Representation, Likelihood" and "Statistical Model Specification". The mortality $\alpha_{12}(t)$ of the diseased (the *lethality*) may sometimes depend on duration d since entry to state 1 in addition to the dependence on "age" t. (Notice that, despite the notation, $\alpha_{12}(t)$ then depends on the *random* time of the most recent transition into 1.) If $\alpha_{12}(t)$ does not depend on d the process is Markovian, otherwise it is a **semi-Markov process**, an example of a purely endogenous process.

In Figure 3, we have indicated the possibility of *reversibility*: the transition back from state 1 to 0 is possible. It will turn out that the simple unidirectional model where $\alpha_{10}(t) = 0$ is rather easier to analyze statistically. Thus, the transition probabilities in this model have simple explicit expressions:

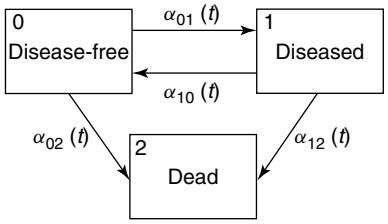

Figure 3 The illness–death or disability model

$$P_{00}(s,t) = \exp\left(-\int_s^t (\alpha_{02}(u) + \alpha_{01}(u))\,\mathrm{d}u\right) \quad (9)$$

and (in the Markovian case)

$$P_{01}(s,t) = \left(\int_s^t P_{00}(s,u-)\alpha_{01}(u)P_{11}(u,t)\,\mathrm{d}u\right) \quad (10)$$

where

$$P_{11}(s,t) = \exp\left(-\int_s^t \alpha_{12}(u)\,\mathrm{d}u\right).$$

More generally, the lethality $\alpha_{12}(\cdot)$ may depend on both age and duration. If we then define

$$\alpha_{12}(t,d) = \lim_{\Delta t \to 0} \frac{\text{Prob}\begin{pmatrix} X(t+\Delta t) = 2 | X(t) = 1, \\ 0 \to 1 \text{ transition at } t - \mathrm{d} \end{pmatrix}}{\Delta t},$$

$P_{11}(u,t)$ in (10) should be replaced by $\exp(-\int_u^t \alpha_{12}(s, s-u)\,\mathrm{d}s)$. The illness–death model is one of the most important multistate models and it was discussed in early papers by Fix & Neyman [31] and Sverdrup [71].

Repeated Events

If interest focuses on repeated occurrences of a given event, for example, hospital admissions, childbirths, infections and so on. then a model as illustrated in Figure 4 (where transitions to an absorbing "Dead" state have been omitted) may be considered. In applications of such a model, an interesting functional is often the expected number of occurrences of the event over the time interval $[0, t]$; see [24] and **Repeated events**.

Interaction Between Life History Events

This Markov model, illustrated in Figure 5, describes the joint behavior of two life events A and B; if $\alpha_{0B} = \alpha_{A,AB}$ but $\alpha_{0A} \neq \alpha_{B,AB}$, A is called *locally dependent on* B but B is not locally dependent on A. The temporal order of events allows for this

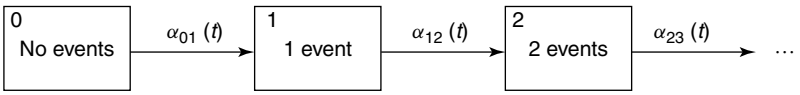

Figure 4 A model for repeated events

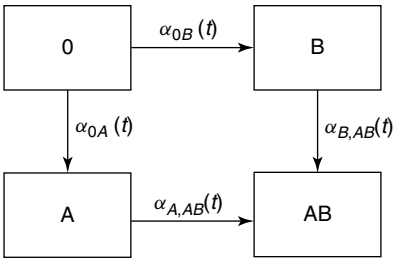

Figure 5 Interaction between life history events

asymmetric concept of dependence, which yields more information for drawing causal inference (*see* **Causal Direction, Determination**; **Causation**) than the standard symmetric association concepts from **cross-sectional studies**. Similar duration dependence as in the illness–death process might be added. A model of this type was discussed by Aalen et al. [6] for a study of interaction between menopause and a certain chronic skin disease; see also [18, 26].

Bone Marrow Transplantation

A model combining most of the above features has been studied in detail (e.g. [56, 58, 60]) as describing some of the possible states of a leukemia patient following bone marrow **transplantation**; see Figure 6.

Patients have been given various kinds of therapy to temporarily keep the disease down, they are said to be in *remission*. In our context these patients are followed since bone marrow transplantation ($t = 0$), initially considered in state 0. Two different types of complications are considered: acute graft-versus-host disease (A), chronic graft-versus-host disease (C), and a special state AC is defined for those patients acquiring both A and C. Patients are followed until relapse of the leukemia (R) or death (D) while still in remission. Relapsed patients are not followed further in this context. If all transition rates depend only on time t since transplantation, we have again a Markov process, but various kinds of duration dependence (semi-Markov process models) may also be relevant.

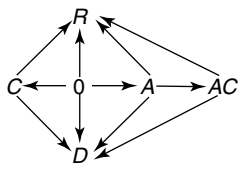

Figure 6 A model for events following bone marrow transplantation: Acute and chronic graft-versus-host disease, relapse and death

Counting Process Representation, Likelihood

Assume that multistate processes $X_i(t)$, such as described in the section "Multistate Models", are observed over intervals $[0, \tau_i]$ for individuals $i = 1, \ldots, n$. Assume, first that τ_i is a fixed (i.e. non-random) time of termination of observation for individual i. Random right-censoring (see the section "Survival Data") and delayed entry are treated below. Since $X_i(t)$ is constant between transitions, it is equivalent to record $X_i(0)$ and the *counting processes*

$$N_{hj}^i(t) = \#\,(\text{direct transitions } h \to j \text{ in } [0,t] \text{ for } i), \tag{11}$$

described by the times T_{hj}^{ik} of these transitions, where

$$0 < T_{hj}^{i1} < \cdots < T_{hj}^{iN_{hj}^i(\tau_i)} \leq \tau_i \tag{12}$$

(*see* **Counting Process Methods in Survival Analysis**).

Let $N_{hj}(t) = \sum_{i=1}^n N_{hj}^i(t)$. It will turn out to be useful to also introduce $Y_h^i(t) = I\{X_i(t-) = h\}$ and

$$Y_h(t) = \#\,(\text{individuals "at risk" in state } h \text{ at time } t-) = \sum_{i=1}^n Y_h^i(t).$$

Note that, for $t > \tau_i$, $N_{hj}^i(t) = N_{hj}^i(\tau_i)$, and $Y_h^i(t) = 0$; these can thus be considered to be defined on $(0, \infty)$.

For individual i, denote the initial distribution $(\pi_h^i(0))$, the density of time-fixed covariates $f(Z_i)$, and the transition intensities $\alpha_{hj}^i(t)$, then the **likelihood** is (see [10, Section III.2])

$$\prod_{i=1}^{n} f(Z_i)\pi_{X_i(0)} \prod_{h\neq j} \prod_{k=1}^{N_{hj}^i(\tau_i)} \alpha_{hj}^i(T_{hj}^{ik}) \times \exp\left(-\int_0^{\tau_i} \alpha_{hj}^i(t) Y_h^i(t)\, \mathrm{d}t\right). \quad (13)$$

It is very common to condition on Z_i and on the initial values $X_i(0)$ (the distribution of which may often be degenerate anyway), and consequently omit the factors $f(Z_i)\pi_{X_i(0)}$ from the likelihood. We shall do so without further comment in the sequel.

Recall from above that the notation $\alpha_{hj}^i(t)$ represents possible dependence of the transition intensity on the whole history $\mathcal{X}_t^i$ of the process. Thus, $\alpha_{hj}^i(t)$ may well contain covariates and other random elements, as already exemplified.

Two patterns of incomplete observations are particularly easily tractable, because they lead to only minor modification of this likelihood: **Delayed entry** where individual i enters at some time V_i, and *right-censoring* where nothing is known about i after some time U_i. Both V_i and U_i may be random although either only dependent on the previous history of the process or independent of the process (see e.g. [10, Chapter 3], for precise specification of this and further discussion). The reason for the particular tractability of these mechanisms is that the "at risk" indicator $Y_h^i(t) = I\{X_i(t-) = h\}$ in the likelihood just needs to be amended to

$$Y_h^i(t) = I\{X_i(t-) = h,\ V_i < t \leq U_i\}. \quad (14)$$

(*see* **Counting Process Methods in Survival Analysis**).

Statistical Model Specification

As indicated in the introduction, the first purpose of event history analysis is to gain insight into the dynamics of the processes by quantifying *transition intensities* and perhaps assessing their dependence on covariates, possibly using various stratifications. Sometimes, additional functionals are useful, particularly various types of *transition probabilities* obtained by integrating certain functions of the transition intensities. A final purpose may be **prediction**, both as illustration of the dynamics and for concrete practical purposes.

Markov Processes

The most important class of models is the (continuous time) *Markov process* $X(t)$ on the finite state space $\mathcal{S} = \{1, \ldots, p\}$, where the dependence of $\alpha_{hj}^i(t)$ on the history $\mathcal{X}_t$, introduced at the beginning of the section "Multistate Models", is only via the current state of $X(t)$ (and possibly via time-fixed covariates). Statistical models are usually obtained by specifying the class of transition intensities $(\alpha_{hj}^i(t))$ for each individual i.

Parametric Models for Transition Intensities. The simplest class of models is obtained by keeping the transition rates *constant*: $\alpha_{hj}^i(t) = \alpha_{hj}^i$. *Piecewise constant* intensities

$$\alpha_{hj}^i(t) = \alpha_{hj}^{i(l)}, \quad t_{l-1}^{hj} < t \leq t_l^{hj}, \quad \text{all } t_0 = 0, \quad (15)$$

form the next step up and this choice is of widespread use, particularly in large studies in econometrics, epidemiology, sociology and demography [10, Section VI.1; 21, 43, 62].

Transition probabilities for the constant and piecewise constant Markov process models are explicit functions of the transition intensities (e.g. [20]), allowing direct "plug-in" **maximum likelihood** estimation, as well as calculation of standard error estimates via the **delta method**.

Although the piecewise constant model is often sufficient to describe the dependence of intensities on time, other possibilities exist. Certain mathematical functions of time may generate the model, such as the Gompertz–Makeham model for mortality (*see* **Aging Models**)

$$\alpha(t) = \alpha + \beta\gamma^t \quad (16)$$

but except for mortality studies in **actuarial** and some **demographic** contexts, such parametric models are used rather little. One reason for this may be the powerful development of methodology for "**nonparametric**" statistical inference, where $\alpha_{hj}(t)$ is left unspecified (*see* **Semiparametric Regression**).

Freely Varying ("Nonparametric") Transition Intensities. Assume first that the transition intensities are the same for all individuals but that they

are allowed to vary freely with time: $\alpha_{hj}^i(t) = \alpha_{hj}(t)$. Statistical inference is then conveniently phrased in terms of the counting process approach pioneered by Aalen [1, 2]; see Andersen et al. [10] for a detailed exposition. Estimators (which may be given a **nonparametric maximum likelihood** interpretation) of the integrated intensities

$$A_{hj}(t) = \int_0^t \alpha_{hj}(u)\,\mathrm{d}u \tag{17}$$

are obtained as the *Nelson–Aalen estimators* (*see* **Nelson–Aalen Estimator**).

An elaborate mathematical theory based on stochastic integrals and martingales is available to study exact and asymptotic properties of these estimators (*see* **Counting Process Methods in Survival Analysis**).

When estimates are desired of the transition intensities $\alpha_{hj}(t)$ themselves rather than their integrals, *smoothing* techniques are necessary (e.g. [10, Section IV.2]) (*see* **Smoothing Hazard Rates**).

An important feature of the nonparametric approach is its elegant generalization (due to Aalen and Johansen [7]) to estimating *transition probabilities*. The basic tool is the (matrix) *product-integral* (*see* **Product-integration**).

Define $\alpha_{hh}(t) = -\sum_{j \neq h} \alpha_{hj}(t)$ and the intensity matrix function $A(t) = ((\alpha_{hj}(t)))$; then the matrix $P(s,t) = ((P_{hj}(s,t)))$ of transition probabilities

$$P_{hj}(s,t) = \mathrm{Prob}(X_i(t) = j | X_i(s) = h) \tag{18}$$

is given by

$$P(s,t) = \Pi_s^t (I + A(\,\mathrm{d}u)). \tag{19}$$

The *Aalen–Johansen* estimator of $P(s,t)$ is obtained by plugging the matrix of Nelson–Aalen estimators $((\hat{A}_{hj}(t)))$ into this formula:

$$\hat{P}(s,t) = \Pi_s^t (I + \hat{A}(\,\mathrm{d}u)). \tag{20}$$

(*see* **Aalen–Johansen Estimator**).

For the simple two-state model for survival data, $\hat{P}_{00}(0,t)$ reduces to the classical **Kaplan–Meier** [49] **estimator** $\hat{S}(t) = \prod_{T_i \leq t}(1 - \mathrm{d}N_{01}(T_i)/Y_0(T_i))$ of the survival function $S(t)$.

As documented in detail by Anderson et al. [10, Section IV.4], there is a well-developed theory, again based on stochastic integrals and martingales, about the asymptotic properties of the Aalen–Johansen estimator.

Markov Regression Models. For Markov models with several states, there will often be too little empirical basis for estimating freely varying transition intensities between all states for all subgroups, so that more **parsimonious** regression models are required. The most frequently used regression models in event history analysis have a multiplicative structure with a baseline $h \to j$ transition intensity $\alpha_{hj0}(t)$, assumed common for all individuals. For an individual, i, with time-fixed covariates $Z_i = (Z_{im})$, the transition intensity is then modeled as

$$\alpha_{hj}^i(t) = \alpha_{hj0}(t) \exp\left(\beta_{hj}' Z_i\right), \tag{21}$$

where the effect of a covariate Z_{im} is described by factors of proportionality $\exp(\beta_{hjm})$. In (21), the baseline hazard may be completely unspecified as in the Cox [27] **proportional hazards** model for survival data (*see* **Cox Regression Model**), or it may be assumed to be piecewise constant leading to **Poisson regression** models. In both cases, inference may be based on the likelihood (13), which for the Cox model leads to the so-called Cox's **partial likelihood** [10, Section VII.2; 28]. The choice between Cox and Poisson models is frequently a matter of convenience, though the latter may be advantageous in large studies where a sufficiency reduction of data into tables of event counts and **person-years** within groups of (categorical) covariates is feasible (e.g. [21, Chapter 31]). In contrast, application of the Cox model requires one data record per individual for each transition.

In (21), the notation suggests that separate baseline hazards and regression coefficients are assumed for each possible transition. If that is the case, then the parameters may be estimated by fitting separate Cox or Poisson models for each transition. However, more parsimonious models may be obtained by assuming some baseline transition intensities proportional (e.g. [56, 58]) or by assuming some covariates to have the same effect on several transitions (e.g. [10, p. 494]). Also, models in which the proportional hazards assumption is relaxed may be considered. In the Poisson case, this is simply an interaction between time and the covariate giving rise to nonproportionality whereas, for the Cox model, the less restrictive model is known as the *stratified* Cox model.

Andersen and Keiding [14] described how such flexible Cox models may be formulated in a way

that shows how standard computer **software** may be applied. In a similar way, Poisson regression models may be analyzed using standard **generalized linear models** software (e.g. [68, 72]).

Another regression model for survival data that readily extends to multistate models is Aalen's [3, 4] nonparametric additive model:

$$\alpha_{hji}(t) = \alpha_{hj0}(t) + \beta_{hj}^{'}(t)Z_i \qquad (22)$$

(see also [10, Section VII.4]; **Aalen's additive regression model**). In this model, both the baseline transition intensities $\alpha_{hj0}(t)$ and the regression functions $\beta_{hjm}(t)$ are left unspecified and nonparametric estimates may be obtained using a generalized **least squares** procedure. Aalen et al. [5] presented a review of this model and its use in multistate models.

"Survival synthesis", that is, combination of the regression estimates for the transition intensities into transition or state probability estimates, may be performed using the product-integral as described by Andersen et al. ([10, Section VII.2.3]; see also [19]). However, except for the simple case of survival data no standard software exists for these computations.

Beyond Markov Processes

The most important deviations from the Markov property in practice are various kinds of **duration dependence**, where transition intensities depend on other time origins than $t = 0$, typically the time at entry to the present state. There are two main approaches to handling these.

As long as transition intensities depend only on one time origin each (e.g. all intensities depend only on duration in the present state), a model for the multistate process may be obtained by combining independent submodels for each transition intensity. These may, in turn, be modeled as constant or piecewise constant or by non- or semiparametric models, and as long as there is a unidirectional flow in the model, transition probabilities are still straightforward explicit functionals that may be estimated by plugging in the intensity estimates. Variance calculations may, however, become less direct.

More elaborate models will include several time origins (such as both age, disease duration, calendar time), often in piecewise constant intensity (Poisson) models or semiparametric regression models such as the Cox model.

In the Poisson models, the various time variables all enter the models as explanatory factors in a symmetrical way (and also symmetrical with respect to the other covariates of the model). However, in Cox models, one of the time variables must be chosen as the "baseline" time variable while the others may be included as time-dependent covariates. The choice of baseline time variable may be governed by several considerations. First, the effect of the baseline time variable is given by the unspecified baseline hazard and, therefore, no regression coefficients are estimated for this variable. Thus, a time variable whose effect is of particular interest may not be the obvious choice as the baseline time variable. On the other hand, if a time variable is suspected to have an irregular effect that may not be easy to model parametrically via a time-dependent covariate, then this time variable may conveniently be chosen as the baseline time variable (e.g. [21, Chapter 31]).

Hypothetical Calculations in Multistate Models

As mentioned earlier, there is often considerable interest in studying the consequences of the estimated transition intensities by calculating summary measures such as transition probabilities. When a full model has been estimated, this can be done not only for the model observed "in this world". Rather, the consequences of an assumed (or fitted) multistate model may be usefully further illustrated by calculating transition probabilities in hypothetical models obtained by changing some of the parameters. An elaborate example of this was given by Keiding et al. [56] for the bone marrow transplantation context; see also [60]. Similar calculations have in fact been performed in the competing risks model ever since the first discussion by Bernoulli [17] of the effect on population mortality of removing smallpox through **vaccination**. The interpretational justification of such calculations was discussed by Gail [34], Prentice et al. [69], Kalbfleisch and Prentice [48, Chapter 7] and Andersen et al. [9].

Partial Model Specification

We have so far assumed that the multistate model was completely specified through statistical modeling of all transition intensities and a specific probability mechanism for the combination of these into transition probabilities. In a series of papers, [25, 65,

66, 67], Pepe and her colleagues have developed estimates of certain functionals in multistate models without assuming a full probability structure. One example is the *prevalence* of a transient condition indicated by state c defined as

$$\frac{P_{0c}(0,t)}{\sum_{j\in\mathcal{T}} P_{0j}(0,t)},$$

where $\mathcal{T}$ is the set of transient states and 0 is a fixed "initial" state. The idea is to estimate numerator and denominator separately by simple linear combinations of Kaplan–Meier estimates. Easily applicable variance estimates are then available, which in one recent application [59], showed that the precision of the Pepe approach was close to the more elaborate (and restrictive) complete Markov model.

Datta and Satten [29], Satten and Datta [70], and Glidden [37] studied the product-integral of the Nelson–Aalen estimator and showed that, also for non-Markovian processes, this combined with the initial distribution $\pi_h(0)$ provides consistent and asymptotically normal estimators for the state probabilities $\pi_h(t)$ (see the section, "Multistate Models").

Andersen et al. [15] showed how regression models for transition probabilities $P_{hj}(s,t)$ or state probabilities $\pi_h(t)$ may be obtained directly in multistate Markov models using **jackknife** pseudo-observations. In fact, their approach may be extended to state probabilities in non-Markovian models using the results of Glidden [37].

Another example of partial model specification occurs when the model contains time-dependent covariates that are not purely endogenous. In fact, for time-fixed covariates, we just (see the section "Multistate Models") conditioned on their observed values without specifying their distribution $f(Z_i)$, but for time-dependent covariates, such a conditioning is more tricky. Formally, the likelihood will contain factors for the stochastic development of $Z_i(t)$, given the history $\mathcal{F}_{t-} = \mathcal{X}_{t-} \vee \mathcal{Z}_{t-}$ and the likelihood (2) is not the full likelihood but only a partial likelihood for the parameters for the transition intensities $\alpha^i_{hj}(t)$. This means that inference for the transition *intensities* may be based on this partial likelihood, whereas the transition *probabilities* will typically depend also on the parameters in the model for $Z_i(t)$.

Thus, if the model contains time-dependent covariates that are not purely endogenous, then transition probabilities cannot be estimated using only a partial model specification. A joint model for the multistate process $X_i(t)$ and the time-dependent covariates $Z_i(t)$ is needed. When $Z_i(t)$ only takes a finite number of values, this joint model could, again, be a multistate model where $Z_i(t)$ is now endogenous (e.g. [8, 13]). Examples of more general joint models were presented by Wulfsohn & Tsiatis [73] and Henderson et al. [40] (see **Joint Modeling of Longitudinal and Event Time Data**).

Kalbfleisch and Prentice [48, Chapter 9] studied intensities obtained from conditioning on a smaller history than that generated by the process itself, particularly within the repeated events framework.

Observational Patterns

As emphasized in the introduction, event history data are rarely observed completely. Some patterns of incomplete observation are more easily handled than others, and this final section aims at introducing some of the more important classes. As mentioned above (see the section "Counting Processes Representation, Likelihood"), independent delayed entry and right censoring only modify the likelihood slightly and the statistical methods then all go through.

Interval Censoring

An important class of incomplete observational patterns consists in the times of some (but often not all) transitions not being known exactly but only up to an **interval**, for example, between visits to a clinic or between censuses. (The *number* of transitions is usually assumed to be known precisely.) A classical approach in demography [41] is to approximate the "exposure"

$$S^i_h = \int_{V_i}^{U_i\wedge\tau_i} Y^i_h(t)\,\mathrm{d}t; \tag{23}$$

another to impute values in the observation interval for the time at risk. Systematic studies of nonparametric maximum likelihood estimation under interval censoring of the healthy $\rightarrow$ diseased transition in the unidirectional illness–death model are by Frydman [32, 33], Joly et al. [46], and Gaüzère [35]. Kay [50] and Andersen et al. [12] exemplified interval-censored observation in the reversible illness–death model, using piecewise constant intensity models. A

particular example is **panel data**; see for example, [36] and the references therein. Interval censoring in multistate models was discussed by Commenges [23], (*see* **Interval Censoring**).

Conditioning in Multistate Models

Many observational patterns in event history analysis may be described by **conditional** distributions in simpler models, which often describe "direct" observations that are practically unobtainable. A prime example is left truncation; another, *right truncation* with widespread use in studies of AIDS patients whose development is often observed conditional on having contracted the disease before the study entry. For general discussions, see [47, 54, 39]. A more elaborate application of such retrospective observational plans obtained by conditioning in an underlying "prospective" Markov process model was documented by Aalen et al. [6] for the four-state interaction of life history events example described above (Figure 6). These authors relied heavily on the concise but important general framework of Hoem [42]. We shall here briefly outline how this methodology works for a simple example of retrospective incidence estimation, obtained from the Markov illness–death model without recovery illustrated in Figure 4. Assume that we study a random sample of individuals alive at same fixed age u; for those who had by then contracted the disease the age at which this happened is recorded. The observed multistate model has state space $\mathcal{K} = \{0, 1\}$ and transition probabilities

$$Q_{hj}(s,t) = \text{Prob}\{X(t) = j \mid X(s) = h, X(u) \neq 2\} \tag{24}$$

for $h, j \in \{0, 1\}$ and $0 < s < t < u$. We get

$$Q_{hj}(s,t) = P_{hj}(s,t)\frac{P_{j\mathcal{K}}(t,u)}{P_{h\mathcal{K}}(t,u)}, \tag{25}$$

where $P_{hj}(s,t)$ are the transition probabilities in the original illness–death process and $P_{h\mathcal{K}} = P_{h0} + P_{h1}$. Hoem [42] used the term *purged* for the conditional Markov process on $\mathcal{K}$ with transition probabilities $Q_{hj}(s,t)$. The transition intensity of the purged process is

$$\lambda_{01}(t) = \alpha_{01}(t)\frac{P_{11}(t,u)}{P_{1\mathcal{K}}(t,u)} \tag{26}$$

and it may be proved that if the mortality of the diseased is never smaller than that of the healthy, that is, $\alpha_{02}(t) \leq \alpha_{12}(t)$ for all $t \leq u$, then $\lambda_{01}(t) \leq \alpha_{01}(t)$, with equality if and only if $\alpha_{02}(t) \equiv \alpha_{12}(t)$. This documents the intuitively obvious result, that the retrospective study will underestimate the disease incidence because of *survivor selection*. Andersen and Green [11] used such methodology to study robustness of diabetes incidence estimates in a situation where diabetics were only observed conditionally on not emigrating before a certain age.

The Prevalent Cohort Study. An important sampling frame for the illness–death model without recovery is the *prevalent cohort study* in which a cross-sectional sample of diseased is taken at a fixed calendar time; *see* **Biased Sampling of Cohorts**. Keiding [52], (cf. Lund [63]), discussed the conditions for correct inference on mortality $\alpha_{12}(t)$ based on follow-up of the diseased, studied under left truncation, and compared to inference based on the **length-biased** durations, which include the retrospective time from disease onset, as well as to the *forward recurrence time* from sampling to death, assuming **stationarity**.

Retrospective estimation of incidence based on the disease onset information of the survivors and independent lethality information was exemplified by Keiding et al. [55] (cf. Ogata et al. [64]).

Some Other Partial Information Designs

Sometimes interval-censoring is extreme – in a cross-sectional study, it is for all individuals where it is only known whether or not an event has happened at age of sampling. Such *current-status data* were discussed in detail by Diamond & McDonald [30], Keiding [51] and Keiding et al. [53]; and there is an elaborate recent mathematical–statistical development in this area; see, for example, [38] and Lin et al. [61] (*see* **Interval Censoring**). For *time to pregnancy* data, Keiding et al. [57] proposed using the *current duration* elapsed so far; under suitable stationarity conditions, this distribution can be considered a backward recurrence time; (*see* **Time to Pregnancy**).

References

[1] Aalen, O.O. (1975). *Statistical Inference for a Family of Counting Processes*. PhD Thesis, University of California, Berkeley.

[2] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[3] Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes, *Springer Lecture Notes on Statistics* **2**, 1–25; Klonecki, W., Kozek, A. & Rosiński, J. eds. *Mathematical Statistics and Probability Theory*. Springer, Heidelberg.

[4] Aalen, O.O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**, 907–925.

[5] Aalen, O.O., Borgan, O. & Fekjær, H. (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach, *Biometrics* **57**, 993–1001.

[6] Aalen, O.O., Borgan, O., Keiding, N. & Thormann, J. (1980). Interaction between life history events: nonparametric analysis of prospective and retrospective data in the presence of censoring, *Scandinavian Journal of Statistics* **7**, 161–171.

[7] Aalen, O.O. & Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics* **5**, 141–150.

[8] Andersen, P.K. (1986). Time-dependent covariates and Markov processes, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds. John Wiley and Sons, New York, pp. 82–103.

[9] Andersen, P.K., Abildstrom, S. & Rosthøj, S. (2002). Competing risks as a multi-state model, *Statistical Methods in Medical Research* **11**, 203–215.

[10] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer, New York.

[11] Andersen, P.K. & Green, A. (1985). Evaluation of estimation bias in an illness-death-emigration model, *Scandinavian Journal of Statistics* **12**, 63–68.

[12] Andersen, P.K., Hansen, L.S. & Keiding, N. (1991a). Assessing the influence of reversible disease indicators on survival, *Statistics in Medicine* **10**, 1061–1067.

[13] Andersen, P.K., Hansen, L.S. & Keiding, N. (1991b). Non- and semi-parametric estimation of transition probabilities from censored observations of a non-homogeneous Markov process, *Scandinavian Journal of Statistics* **18**, 153–167.

[14] Andersen, P.K. & Keiding, N. (2002). Multi-state models for event history analysis, *Statistical Methods in Medical Research* **11**, 91–115.

[15] Andersen, P.K., Klein, J.P. & Rosthøj, S. (2003). Generalized linear models for correlated pseudo-observations, with applications to multi-state models, *Biometrika* **90**, 15–27.

[16] Arjas, E. & Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates, *Scandinavian Journal of Statistics* **11**, 193–209.

[17] Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour le prévenir, *Mémoires de Mathématique et de Physique de l'Académie Royale des Sciences*, Paris, 1–45.

[18] Blossfeld, H. & Rohwer, G. (1995). *Techniques of Event History Modeling*. Lawrence Erlbaum, New Jersey.

[19] Borgan, Ø. (2002). Estimation of covariate-dependent Markov transition probabilities from nested case-control data, *Statistical Methods in Medical Research* **11**, 183–202.

[20] Chiang, C.L. (1968). *Introduction to Stochastic Processes in Biostatistics*. John Wiley & Sons, New York.

[21] Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

[22] Commenges, D. (1999). Multi-state models in epidemiology, *Lifetime Data Analysis* **5**, 315–327.

[23] Commenges, D. (2002). Inference for multi-state models from interval-censored data, *Statistical Methods in Medical Research* **11**, 167–182.

[24] Cook, R.J. & Lawless, J.F. (2002). Analysis of repeated events, *Statistical Methods in Medical Research* **11**, 141–166.

[25] Couper, D. & Pepe, M.S. (1997). Modelling prevalence of a chronic condition: chronic graft-versus-host disease after bone marrow transplantation, *Statistics in Medicine* **16**, 1551–1571.

[26] Courgeau, D. & Lelièvre, E. (1992). *Event History Analysis in Demography*. Clarendon, Oxford.

[27] Cox, D.R. (1972a). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society Series B* **34**, 187–220.

[28] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[29] Datta, S. & Satten, G.A. (2001). Validity of the Aalen Johansen estimators of stage occupation probabilities and Nelson Aalen integrated transition hazards for non-Markov models, *Statistics and Probability Letters* **55**, 403–411.

[30] Diamond, I.D. & McDonald, J.W. (1992). Analysis of current-status data, in *Demographic Applications of Event History Analysis*, J. Trussel, R. Hankinson & J. Tilton, eds. Clarendon Press, Oxford, 231–252.

[31] Fix, E. & Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients, *Human Biology* **23**, 205–241.

[32] Frydman, H. (1992). A non-parametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS, *Journal of the Royal Statistical Society, Series B* **54**, 853–866.

[33] Frydman, H. (1995). Semiparametric estimation in a three-state duration dependent Markov model from interval-censored observations with application to AIDS data, *Biometrics* **51**, 502–511.

[34] Gail, M. (1975). A review and critique of some models used in competing risk analysis, *Biometrics* **31**, 209–222.

[35] Gaüzère, F. (2000). Approche Non-paramétrique pour un Modèle 3 états avec Censures par Intervalles – Application à la Dépendance. PhD Thesis, Université Victor Segalen, Bordeaux 2.

[36] Gentleman, R.C., Lawless, J.F., Lindsay, J.C. & Yan, P. (1994). Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease, *Statistics in Medicine* **13**, 805–821.

[37] Glidden, D.V. (2002). Robust inference for event probabilities with non-Markov event data, *Biometrics* **58**, 361–368.

[38] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

[39] Gross, S.T. & Huber-Carol, C. (1992). Regression models for truncated survival data, *Scandinavian Journal of Statistics* **19**, 193–213.

[40] Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data, *Biostatistics* **1**, 465–480.

[41] Hoem, J.M. (1969a). Fertility rates and reproduction rates in a probabilistic setting, *Biométrie-Praximétrie* **10**, 38–66. Correction, 11 (1970), p. 20.

[42] Hoem, J.M. (1969b). Purged and partial Markov chains, *Skandinavisk Aktuarietidskrift* **52**, 147–155.

[43] Hoem, J.M. (1976). The statistical theory of demographic rates. A review of current developments (with discussion), *Scandinavian Journal of Statistics* **3**, 169–185.

[44] Hougaard, P. (1999). Multi-state models: a review, *Lifetime Data Analysis* **5**, 239–264.

[45] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

[46] Joly, P. & Commenges, D. (1999). A penalized approach for a progressive three-state model with censored and truncated data: application to AIDS, *Biometrics* **55**, 887–890.

[47] Kalbfleisch, J.D. & Lawless, J.F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS, *Journal of the American Statistical Association* **84**, 360–372.

[48] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. Wiley, New York.

[49] Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[50] Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies, *Biometrics* **42**, 855–865.

[51] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[52] Keiding, N. (1992). Independent delayed entry, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Dordrecht, 309–326.

[53] Keiding, N., Begtrup, K., Scheike, T.H. & Hasibeder, G. (1996). Estimation from current-status data in continuous time, *Lifetime Data Analysis* **2**, 119–129.

[54] Keiding, N. & Gill, R.D. (1990). Random truncation models and Markov processes, *Annals of Statistics* **18**, 582–602.

[55] Keiding, N., Holst, C. & Green, A. (1989). Retrospective estimation of diabetes incidence from information in a current prevalent population and historical mortality, *American Journal of Epidemiology* **130**, 588–600.

[56] Keiding, N., Klein, J.P. & Horowitz, M.M. (2001). Multistate models and outcome prediction in bone marrow transplantation, *Statistics in Medicine* **20**, 1871–1885.

[57] Keiding, N., Kvist, K., Hartvig, H., Tvede, M. & Juul, S. (2002). Estimating time to pregnancy from current durations in a cross-sectional sample, *Biostatistics* **3**, 565–578.

[58] Klein, J.P., Keiding, N. & Copelan, E.A. (1993). Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients, *Statistics in Medicine* **12**, 2315–2332.

[59] Klein, J.P., Keiding, N., Shu, Y., Szydlo, R.M. & Goldman, J.M. (2000). Summary curves for patients transplanted for chronic myeloid leukemia salvaged by a donor lymphocyte infusion: the current leukemia free survival curve, *British Journal of Haematology* **109**, 148–152.

[60] Klein, J.P. & Shu, Y. (2002). Multi-state models for bone marrow transplantation studies, *Statistical Methods in Medical Research* **11**, 117–139.

[61] Lin, D.Y., Oakes, D. & Ying, Z. (1998). Additive hazards regression with current status data, *Biometrika* **85**, 289–298.

[62] Lindsay, J.C. & Ryan, L.M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments, *Applied Statistics* **42**, 283–300.

[63] Lund, J. (2000). Sampling bias in population studies - how to use the Lexis diagram, *Scandinavian Journal of Statistics* **27**, 589–604.

[64] Ogata, Y., Katsura, K., Keiding, N., Holst, C. & Green, A. (2000). Empirical Bayes age-period-cohort analysis of retrospective incidence data, *Scandinavian Journal of Statistics* **27**, 415–432.

[65] Pepe, M.S. (1991). Inference for events with dependent risks in multiple endpoint studies, *Journal of the American Statistical Association* **86**, 770–778.

[66] Pepe, M.S., Longton, G. & Thornquist, M. (1991). A qualifier Q for the survival function to describe the prevalence of a transient condition, *Statistics in Medicine* **10**, 413–421.

[67] Pepe, M.S. & Mori, M. (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* **12**, 737–751.

[68] Pierce, D.A. & Preston, D.L. (1993). Joint analysis of site-specific cancer risks for the atomic bomb survivors, *Radiation Research* **134**, 134–142.

[69] Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Flournoy, N., Farewell, V.T. & Breslow, N. (1978). The analysis of failure time data in the presence of competing risks, *Biometrics* **34**, 541–554.

[70] Satten, G.A. & Datta, S. (2002). Marginal estimation for multi-state models: waiting time distributions and

competing risks analyses, *Statistics in Medicine* **21**, 3–19.

[71] Sverdrup, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health, *Skandinavisk Aktuarietidskrift* **48**, 184–211.

[72] Wohlfahrt, J., Andersen, P.K. & Melbye, M. (1999). Multivariate competing risks, *Statistics in Medicine* **18**, 1023–1030.

[73] Wulfsohn, M.S. & Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error, *Biometrics* **53**, 330–339.

Further Reading

Cohen, J.E. (1972). When does a leaky compartment model appear to have no leaks? *Theoretical Population Biology* **3**, 404–405.

Cox, D.R. (1972b). The statistical analysis of dependencies in point processes, in *Stochastic Point Processes*, P.A.W. Lewis, ed. John Wiley and Sons, New York, pp. 55–66.

(*See also* **Longitudinal Data Analysis, Overview**; **Model, Choice of**; **Transition Models for Longitudinal Data**).

PER KRAGH ANDERSEN & NIELS KEIDING

Event-related Potential

Event-related potentials (ERPs) are brain electrical potentials used to study sensory and cognitive processing [8]. In a typical ERP experiment, a sensory stimulus is presented to a human subject or experimental animal, and the brain electrical activity following the stimulus is recorded by electrodes on the scalp or implanted in the skull or brain. Commonly used stimuli include tones presented to one or both ears, and symbols presented to a specified part of the visual field. ERPs allow millisecond time resolution of brain electrical activity, so they can be used to study the fast time course of sensory and cognitive processing. Brain imaging techniques, such as functional magnetic resonance imaging (fMRI), yield much better spatial resolution, but they have poor time resolution measured in seconds or minutes.

Vaughan [9] proposed the term "event-related potential" to include potentials related to either cognitive or sensory events. The more specific term "evoked potential" refers to responses directly related to sensory processing of stimuli. Data containing both evoked potentials and cognitive ERPs are shown in Figure 1. These data were collected in an auditory oddball experimental design. In this design, the subject is presented with a series of stimuli, each of which is one of two different tones. The two tones are presented in a random sequence, but one tone ("standard") occurs with probability 0.8, while the other ("rare target") occurs with probability 0.2. The subject is instructed to count the number of occurrences of the rare target tone. The recorded **time series** show a large negative trough near 100 milliseconds ("N100"), which occurs equally in response to either the standard or the rare target tone. In contrast, the positive wave occurring approximately 300 milliseconds after the stimulus ("P300") has a larger amplitude in response to the rare target tone, which is more salient to the subject's task. Based on results such as these, investigators have concluded that the N100 is an evoked potential reflecting sensory processing of the stimulus, while the P300 is an ERP related to cognitive processing.

The characteristic peaks and troughs that appear in plots such as Figure 1 are called "components". Many components occurring within the first 100 milliseconds after the stimulus can be attributed to activity in localized anatomic structures along known sensory pathways [2]. Later components appear to arise from more dispersed cortical sources [2].

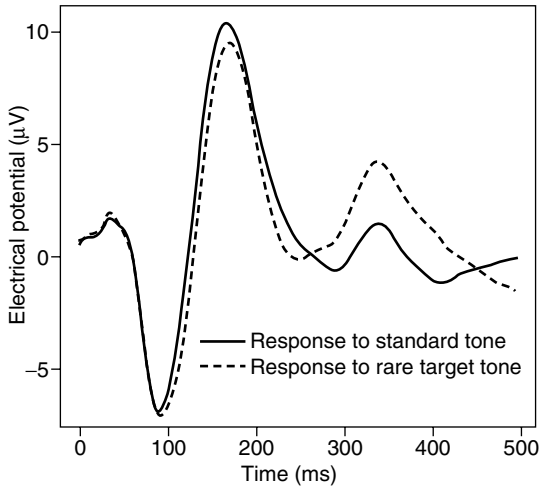

Figure 1 Human auditory event-related potentials. The response is plotted against time from stimulus presentation. Each time series is the average of ERPs acquired from 10 subjects, each of which is the average of responses to many stimulus presentations. The peak between 300 ms and 400 ms has much greater amplitude in response to the target stimulus, which indicates that it is related to cognitive processing. The large trough near 100 ms is an evoked potential reflecting sensory processing

Statistical analysis of ERPs is challenging, since typical ERP data sets consist of a large number of time series, which vary among individuals, recording channels, and **explanatory variables** such as stimulus type, drug condition, and disease state. Investigators commonly wish to decompose these many time series into meaningful component waveforms, estimate the effect of the explanatory variables on the amplitude and latency (time since stimulus presentation) of each component, and estimate the locations of the brain sources of the components.

The simplest and most commonly used method for analyzing ERP data is to measure the amplitudes and latencies of various peaks and troughs in the recorded time series, and then apply **analysis of variance** and other traditional statistical methods to these derived measures. This approach is not always satisfactory, since two or more brain processes may be simultaneously active, resulting in superposition of components

in the recorded time series. Furthermore, identification of relevant peaks can be highly subjective when noisy recordings are analyzed, and analysis of peaks does not provide estimates of the component waveforms themselves.

Donchin [3] proposed analyzing ERPs using **principal component analysis** (PCA) by treating each time point as a separate "variable" and each time series from a particular individual, combination of explanatory variables, and recording channel as a multivariate "observation". Investigators typically apply **varimax rotation**, interpret the **factor loadings** (rotated eigenvectors) as ERP components, and use analysis of variance to test for effects of explanatory variables on the **factor scores**. Several authors have criticized this use of PCA. The factor scores are necessarily uncorrelated, which can lead to misleading inference, and the choice of rotation sometimes seems arbitrary.

Implicit in the use of PCA is a bilinear statistical model in which the expected responses are represented by unknown linear combinations of unknown component time series. Möcks [4] noted that while bilinear models are not identifiable due to rotation, multilinear models are identifiable up to scaling, and he proposed multilinear models for ERPs.

Brillinger [1] and several other statisticians have suggested analyzing ERPs by modeling the Fourier coefficients of the data. This approach is particularly useful when the explanatory variables are assumed to change the amplitude and latency of some underlying common response, as in the following time-domain model:

$$Y_{jt} = \beta_j \gamma(t + \tau_j) + \varepsilon_{jt},$$
$$j = 1, \ldots, m; \quad t = 1, \ldots, n, \qquad (1)$$

where Y_{jt} is the recorded potential at time t under the jth combination of explanatory variables, β_j is the amplitude, τ_j is the latency, and $\gamma(t)$ is an underlying mean function, with appropriate constraints to make the model identifiable. The noise vector $(\varepsilon_{j1}, \ldots, \varepsilon_{jn})$ is assumed to be generated by a stationary, mixing random process. This simple model assumes a single component and applies to data from a single recording channel and a single subject, but more general frequency domain models have been proposed.

Applying the discrete Fourier transform to (1) yields an approximate frequency domain model in which the latency effects τ_j appear as phase shifts. When n is large, the errors in the frequency domain model are approximately independent and normally distributed with variances proportional to the **spectral** power of the noise process. This representation leads to inference based on an approximate **multivariate normal** likelihood.

The frequency domain model can be generalized to more than one ERP component (assuming that constraints are introduced to ensure identifiability), but it will be less useful when each component is localized in time, since such components will not be parsimoniously represented by their Fourier coefficients. If the Fourier transform is replaced by the wavelet transform (or one of the many related transforms), then localization in both time and frequency is achieved, but some of the advantages of the frequency domain approach are lost.

Estimation of the location of the brain sources of ERPs is of great scientific interest, but poses a variety of difficult problems in biophysical modeling. A statistical approach was proposed by Raz et al. [7], who used a relatively simple biophysical model of the head as part of a frequency domain method for source localization.

When a time series of brain electrical potentials is acquired following an experimentally controlled stimulus, the response to the stimulus is small and is usually obscured by unrelated brain activity. Investigators almost always average the responses to many stimulus presentations ("single trials") to enhance the signal-to-noise ratio and obtain an estimate of the ERPs, which are the brain potentials that are time-locked to the stimulus. The average recorded potential, however, may be a poor estimator of the expected brain response ("signal") if the brain responds differently to different stimulus presentations; that is, if the signal is heterogeneous. Several statisticians have considered the problem of estimating heterogeneous signals and testing the null hypothesis of signal homogeneity [5, 6].

ERP data continue to present challenges for statisticians. Open problems include: constructing statistical models that realistically account for the variability among subjects, recording channels, levels of explanatory variables, and single trials; further advancing a statistical perspective on the difficult problem of source localization; and developing statistical methods for relating ERPs to brain images.

References

[1] Brillinger, D.R. (1981). Some aspects of the analysis of evoked response experiments, in *Statistics and Related Topics*, M. Csörgo, D.A. Dawson, J.N.K. Rao & A.K.Md.E. Saleh, eds. North-Holland, Amsterdam, pp. 155–169.

[2] Buchwald, J.S. (1989). Comparisons of sensory and cognitive brain potentials in the human and in an animal model, in *Springer Series in Brain Dynamics*, Vol. 2, E. Başar & T.H. Bullock, eds. Springer-Verlag, Berlin, pp. 242–257.

[3] Donchin, E. (1966). A multivariate approach to the analysis of average evoked potentials, *IEEE Transactions on Biomedical Engineering* **13**, 131–139.

[4] Möcks, J. (1988). Topographic components model for event-related potentials and some biophysical considerations, *IEEE Transactions on Biomedical Engineering* **35**, 482–484.

[5] Möcks, J., Pham, D.T. & Gasser, T. (1984). Testing for homogeneity of noisy signals evoked by repeated stimuli, *Annals of Statistics* **12**, 193–209.

[6] Raz, J. & Fein, G. (1992). Testing for heterogeneity of evoked potential signals using an approximation to an exact permutation test, *Biometrics* **48**, 1069–1080.

[7] Raz, J., Turetsky, B. & Fein, G. (1992). Frequency domain estimation of the parameters of human brain electrical dipoles, *Journal of the American Statistical Association* **87**, 69–77.

[8] Regan, D. (1989). *Human Brain Electrophysiology*. Elsevier, New York.

[9] Vaughan, H. (1969). The relationship of brain activity to scalp recording of event-related potentials, in *Averaged Evoked Potentials*, E. Donchin & D.B. Lindsley, eds. NASA, Washington, pp. 45–94.

(*See also* **Clinical Signals**; **Stimulus–Response Studies**)

JONATHAN RAZ & BRUCE TURETSKY

Evidence-based Medicine

Evidence-based medicine has been defined by its proponents as the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients [10] (*see* **Decision Analysis in Diagnosis and Treatment Choice**). In this definition, the practice of evidence-based medicine means integrating individual clinical expertise with a critical appraisal of the best available external clinical evidence from systematic research. By individual clinical expertise is meant the proficiency and judgment that individual clinicians acquire through clinical experience and clinical practice. Increased expertise is reflected in many ways, but especially in more effective and efficient diagnosis and in the more thoughtful identification and compassionate use of individual patients' predicaments, rights, and preferences in making clinical decisions about their care. By best available external clinical evidence is meant clinically relevant research, often from the basic sciences of medicine, but especially from patient-centered clinical research into the accuracy and precision of **diagnostic tests** (including the clinical examination), the power of **prognostic factors**, and the efficacy and safety of therapeutic, rehabilitative, and preventive regimens [4].

The practice of evidence-based medicine is a process of lifelong, self-directed learning in which caring for one's own patients creates the need for clinically important information about diagnosis, prognosis, therapy, and other clinical and health care issues, and in which its practitioners:

1. Convert these information needs into answerable questions.
2. Track down, with maximum efficiency, the best evidence with which to answer them (and making increasing use of secondary sources of the best evidence). Examples of such secondary sources are the Cochrane Library and journals of critically appraised clinical articles such as *ACP Journal Club* and *Evidence-Based Medicine*.
3. Critically appraise that evidence for its validity (closeness to the truth) and usefulness (clinical applicability).
4. Integrate the appraisal with clinical expertise and apply the results in clinical practice.
5. Evaluate one's own performance.

Evidence-based medicine is one of several disciplines that has evolved from **clinical epidemiology** and critical appraisal. Parallel developments, still with the individual patient as the focus of attention, are occurring in other clinical disciplines (evidence-based surgery, evidence-based nursing, evidence-based dentistry, etc.). Other evidence-based disciplines consider the community as the focus of attention rather than the individual patient (evidence-based public health), or add an explicit economic element and seek to purchase or provide that mix of health care that will maximize some group or public benefit (evidence-based purchasing).

Recent audits in the front lines of clinical care have documented that some inpatient clinical teams in general medicine [3], psychiatry [5], and surgery (P. McCulloch, personal communication) have provided evidence-based care to the vast majority of their patients. Such studies show that busy clinicians who devote their scarce reading time to selective, efficient, patient-driven searching, appraisal, and incorporation of the best available evidence can practice evidence-based medicine.

Common misconceptions about evidence-based medicine include the concern that it might degenerate into "cookbook" medicine. However, because it requires a bottom-up approach that integrates the best external evidence with individual clinical expertise and patient choice, it cannot result in slavish, cookbook approaches to individual patient care. External clinical evidence can inform, but can never replace, individual clinical expertise, and it is this expertise that decides whether the external evidence applies to the individual patient at all and, if so, how it should be integrated into a clinical decision. Similarly, any external guideline must be integrated with individual clinical expertise in deciding whether and how it matches the patient's clinical state, predicament, and preferences, and thus whether it should be applied. Clinicians who fear top-down cookbooks will find the advocates of evidence-based medicine joining them at the barricades.

Others fear that evidence-based medicine will be hijacked by purchasers and managers to cut the costs of health care. This would not only be a misuse of evidence-based medicine but suggests a fundamental misunderstanding of its financial consequences. Doctors practicing evidence-based medicine will identify and apply the most efficacious interventions to maximize the quality and quantity of life for individual

patients; this may raise rather than lower the cost of their care.

Finally, in terms of study designs, evidence-based medicine is not restricted to randomized trials (*see* **Clinical Trials, Overview**) and **meta-analyses**. It involves tracking down the best external evidence with which to answer our clinical questions. To find out about the accuracy of a diagnostic test (*see* **Diagnostic Test Accuracy**), its practitioners seek **likelihood ratios, sensitivities**, and **specificities** derived from proper **cross-sectional studies** of patients clinically suspected of harboring the relevant disorder, not a randomized trial. For a question about prognosis, they search for multivariate prediction rules (*see* **Multivariate Multiple Regression**) generated from proper follow-up studies (*see* **Cohort Study**) of patients assembled at a uniform, early point in the clinical course of their disease; sometimes the evidence will come from the basic sciences such as genetics or immunology. It is when asking questions about therapy that the practitioners of evidence-based medicine avoid the nonexperimental approaches (*see* **Observational Study**), since these routinely lead to **false positive** conclusions about efficacy. Because the randomized trial, and especially the systematic review of several randomized trials [2], is so much more likely to inform clinicians and so much less likely to mislead them, it has become the **gold standard** for judging whether a treatment does more good than harm. Clinically useful measures of the effects of treatment are sought, such as the **number needed to treat** to prevent one additional event (the reciprocal of the absolute risk reduction).

Despite its ancient origins, evidence-based medicine remains a relatively young discipline whose positive impacts [1] are just beginning to be validated, and it will continue to evolve. This evolution will be enhanced as several undergraduate, postgraduate, and continuing medical education programs adopt and adapt it to their students' needs.

For more reading, users can refer to any of the growing number of texts on this subject [6, 8, 9] or examine it on the **Internet** [7].

References

[1] Bennett, K.J., Sackett, D.L., Haynes, R.B. & Neufeld, V.R. (1987). A controlled trial of teaching critical appraisal of the clinical literature to medical students, *Journal of the American Medical Association* **257**, 2451–2454.

[2] Cook, D.J., Sackett, D.L. & Spitzer, W.O. (1995). Methodologic guidelines for systematic reviews of randomized trials in health care from the Potsdam Consultation on Meta-Analysis, *Journal of Clinical Epidemiology* **48**, 167–171.

[3] Ellis, J., Mulligan, I., Rowe, J. & Sackett, D.L. (1995). Inpatient general medicine is evidence based, *Lancet* **346**, 407–410.

[4] Evidence-Based Medicine Working Group (1992). Evidence-based medicine: A new approach to teaching the practice of medicine, *Journal of the American Medical Association* **268**, 2420–2425.

[5] Geddes, J.R., Game, D., Jenkins, N.E., Peterson, L.A., Pottinger, G.R. & Sackett, D.L. (1996). In-patient psychiatric care is evidence-based, *Proceedings of the Royal College of Psychiatrists Winter Meeting*, Stratford, UK, January 23–25.

[6] Gray, J.A.M. (1997). *Evidence-Based Health Care*. Churchill-Livingstone, London.

[7] `Http://cebm.jr2.ox.ac.uk/ or http://hiru.mcmaster.ca/`.

[8] Ridsdale, L. (1995). *Evidence-Based General Practice*. W.B. Saunders, London.

[9] Sackett, D.L., Richardson, S.W., Rosenberg, W.R. & Haynes, R.B. (1997). *Evidence-Based Medicine; How to Practice and Teach EBM*. Churchill-Livingstone, London.

[10] Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M. & Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't, *British Medical Journal* **312**, 71–72.

DAVID L. SACKETT

Exact Inference for Categorical Data

Modern statistical methods rely heavily on **nonparametric** techniques for comparing two or more populations. These techniques generate ***P* values** without making any distributional assumptions about the populations being compared. They rely, however, on asymptotic theory that is valid only if the sample sizes are reasonably large and well balanced across the populations (*see* **Large-sample Theory**). For small, sparse, skewed, or heavily tied data, the asymptotic theory may not be valid; see [5] for some empirical results, and [42] for a more theoretical discussion.

One way to make valid statistical inferences in the presence of small, sparse, or unbalanced data is to compute exact P values and **confidence intervals**, based on the permutational distribution of the test statistic (*see* **Randomization Tests**). This approach was first proposed by R. A. **Fisher** [18] and has been used extensively for the single **2 x 2** contingency table (*see* **Fisher's Exact Test**). Previously, exact tests were rarely attempted for tables of higher dimension than 2×2, primarily because of the formidable computing problems involved in their execution. In recent years, however, the easy availability of immense quantities of computing power combined with many new, fast, and efficient **algorithms** for exact permutational inference have revolutionized our thinking about what is computationally feasible. Problems that would previously have taken several hours or even days to solve now take only a few minutes. Exact **inference** is now a practical proposition and has been incorporated into standard statistical **software** packages.

In the present paper, we present a unified framework for exact inference, anchored in the permutation principle. We demonstrate that, for a very broad class of nonparametric problems, such inference can be accomplished by permuting the entries in a contingency table subject to fixed margins. Exact and **Monte Carlo** algorithms for solving these permutation problems are referenced. We then apply these algorithms to several data sets. Both exact and asymptotic P values are computed for these data so that one may assess the accuracy of the asymptotic methods. Finally, we discuss the availability of software and cite an **internet** resource for performing exact permutational inference.

Exact Permutation Tests for $r \times c$ Contingency Tables

For a broad class of statistical tests, the data can be represented in the form of the $r \times c$ **contingency table** $\boldsymbol{x}$ displayed in Table 1.

The entry in each cell of this $r \times c$ table is the number of subjects falling in the corresponding row and column classifications. The row and column classifications may be based on either **nominal** or *ordered* variables (*see* **Ordered Categorical Data**; **Measurement Scale**). Nominal variables take on values that cannot be positioned in any natural order. An example of a nominal variable is profession – Medicine, Law, Business. In some statistical packages, nominal variables are also referred to as *class* variables, or *unordered* variables. Ordered variables take on values that can be ordered in a natural way. An example of an ordered variable is Drug Dose – Low, Medium, High. Ordered variables may of course assume numerical values as well (for example, the number of cigarettes smoked per day).

The Exact Permutation Distribution of $\boldsymbol{x}$

The exact probability distribution of $\boldsymbol{x}$ depends on the sampling scheme that was used to generate $\boldsymbol{x}$. When both the row and column classifications are categorical, Agresti [1] lists three sampling schemes that could give rise to $\boldsymbol{x}$; full **multinomial** sampling, product multinomial sampling, and **Poisson** sampling. Under all three schemes, the probability distribution of $\boldsymbol{x}$ contains unknown parameters, π_{ij}, relating to the individual cells of the $r \times c$ table (*see* **Loglinear Model**). For instance, for full multinomial sampling, π_{ij} denotes the probability of classification in row

Table 1 Layout for a generic $r \times c$ contingency table

Rows	Col_1	Col_2	...	Col_c	Row_total
Row_1	x_{11}	x_{12}	...	x_{1c}	m_1
Row_2	x_{21}	x_{22}	...	x_{2c}	m_2
⋮	⋮	⋮	...	⋮	⋮
Row_r	x_{r1}	x_{r2}	...	x_{rc}	m_r
Col_tot	n_1	n_2	...	n_c	N

i and column j, whereas for product multinomial sampling π_{ij} denotes the **conditional probability** of falling in column j given that the subject belongs to row i.

Consider the **null hypothesis** of no row by column **interaction**. Since statistical inference is based on the distribution of $\boldsymbol{x}$ under the null hypothesis of no row by column interaction, the number of unknown parameters is reduced (π_{ij} being replaced by $\pi_{i.}\pi_{.j}$ or $\pi_{.j}$ depending on the sampling scheme) but not eliminated. Asymptotic inference relies on estimating these unknown parameters by **maximum likelihood** and related methods. The key to exact permutational inference is getting rid of all **nuisance parameters** from the probability distribution of $\boldsymbol{x}$. This is accomplished by restricting the sample space to the set of all $r \times c$ contingency tables that have the same marginal sums as the observed table $\boldsymbol{x}$. Specifically, define the reference set

$$\Gamma = \left\{ \boldsymbol{y} : \boldsymbol{y} \text{ is } r \times c; \sum_{j=1}^{c} y_{ij} = m_i; \sum_{i=1}^{r} y_{ij} = n_j; \text{ for all } i, j \right\}. \quad (1)$$

Then one can show that, under the null hypothesis of no row by column interaction, the probability of observing $\boldsymbol{x}$ conditional on $\boldsymbol{x} \in \Gamma$ is of the **hypergeometric** form

$$\Pr(\boldsymbol{x} \mid \boldsymbol{x} \in \Gamma) \equiv P(\boldsymbol{x}) = \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{n_j! m_i!}{N! x_{ij}!}. \quad (2)$$

Equation (2), which is free of all unknown parameters, holds for categorical data whether the sampling scheme used to generate $\boldsymbol{x}$ is full multinomial, product multinomial, or Poisson. (See, for example, [2].)

Since (2) contains no unknown parameters, exact inference is possible. The nuisance parameters were, however, eliminated by conditioning on the margins of the observed contingency table. Now some of these margins were not fixed when the data were gathered. Thus, it is reasonable to question the appropriateness of fixing them for purposes of inference. The justification for conditioning at inference time on margins that were not naturally fixed at data sampling time has a long history. R.A. Fisher [18] first proposed this idea for exact inference on a single 2×2 contingency table. At various times since then prominent statisticians have commented on this approach. The principles most cited for conditioning are the **sufficiency** *principle*, the **ancillarity** *principle*, and the **randomization** *principle* (*see* **Conditionality Principle**). An informal intuitive explanation of these three principles is provided below.

Sufficiency Principle. The margins of the contingency table are sufficient statistics for unknown nuisance parameters. Thus, conditioning on them affords a convenient way to eliminate nuisance parameters from the **likelihood** function. For example, if the data are generated by product multinomial sampling, the row margins, m_i, would ordinarily be considered fixed but the column margins, n_j, would be considered random variables. The null hypothesis of interest states that $\pi_{ij} = \pi_{.j}$ for all i. Thus, the probability of $\boldsymbol{x}$ depends on c unknown nuisance parameters, $(\pi_{.1}, \pi_{.2}, \ldots \pi_{.c})$ even under the null hypothesis. By the sufficiency principle, these nuisance parameters are eliminated if we condition on $(n_1, n_2, \ldots n_c)$, their **sufficient statistics**. It follows that by restricting our attention to $r \times c$ tables in Γ, we are implicitly conditioning on $(n_1, n_2, \ldots n_c)$, since the other set of margins, $(m_1, m_2, \ldots m_r)$, are fixed naturally by the sampling scheme. Similar sufficiency arguments can be made for full multinomial and Poisson sampling.

Ancillarity Principle. The principle underlying hypothesis testing is to compare what was actually observed with what could have been observed in hypothetical repetitions of the original experiment, under the null hypothesis. In these hypothetical repetitions, it is a good idea to keep all experimental conditions unrelated to the null hypothesis unchanged as far as possible. The margins of the contingency table are representative of nuisance parameters whose values do not provide any information about the null hypothesis of interest. In this sense, they are ancillary statistics. Fixing them in hypothetical repetitions is the nearest we can get to fixing the values of the nuisance parameters themselves in hypothetical repetitions, since the latter are unknown.

Randomization Principle. The case for conditioning is especially persuasive if the r rows of the contingency tables represent r different treatments,

with m_i subjects being assigned to treatment i by a randomization mechanism. Each subject provides a multinomial response that falls into one of the c columns. Thus, n_j represents the total number of responses of the jth type. Now, under the null hypothesis, the r treatments are equally effective. Therefore, the response that a patient provides is the same, regardless of the treatment to which that patient is randomized. Thus, the value of n_j is predetermined and may be regarded as fixed. The statistical significance of the observed outcome is judged relative to its permutational distribution in hypothetical repetitions of the randomization rule for assigning patients to treatments.

An excellent exposition of the conditional viewpoint is available in [52]. For a theoretical justification of the sufficiency and ancillarity principles, refer to [16, 43]. For a detailed exposition of the randomization principle, highlighting its applicability to a broad range of problems, refer to [17]. Throughout the present paper, we shall adopt the conditional approach. It provides us with a unified way to perform exact inference and thereby compute accurate P values and confidence intervals for $r \times c$ contingency tables (*see* **Stratification**), stratified 2×2 contingency tables, stratified $2 \times c$ contingency tables, and **logistic regression**.

Exact P Values

Having assigned an exact probability $P(\boldsymbol{y})$ to each $\boldsymbol{y} \in \Gamma$, the next step is to order each contingency table in Γ by a test statistic or "discrepancy measure" that quantifies the extent to which that table deviates from the null hypothesis of no row by column interaction. Let us denote the test statistic by a real valued function $D : \Gamma \longrightarrow \mathcal{R}$ mapping $r \times c$ tables from Γ onto the real line $\mathcal{R}$. The functional form of D for some important nonparametric tests is specified in the next section.

The P value is defined as the sum of null probabilities of all the tables in Γ, which are at least as extreme as the observed table, $\boldsymbol{x}$, with respect to D. In particular, if $\boldsymbol{x}$ is the observed $r \times c$ table, the exact P value is

$$p = \sum_{D(\boldsymbol{y}) \geq D(\boldsymbol{x})} P(\boldsymbol{y}) = \Pr\{D(\boldsymbol{y}) \geq D(\boldsymbol{x})\}. \quad (3)$$

Classical nonparametric methods rely on the large-sample distribution of D to estimate p. For $r \times c$ tables with large cell counts and the usual forms for the function D, it is possible to show that D converges in distribution to a **chi-square distribution** with appropriate **degrees of freedom**. Thus, p is usually estimated by $\tilde{p}$, the chi-square tail area to the right of $D(\boldsymbol{x})$. Modern algorithmic techniques have made it possible to compute p directly instead of relying on $\tilde{p}$, its asymptotic approximation. This is achieved by powerful recursive algorithms that are capable of generating the actual permutation distribution of D instead of relying on its asymptotic chi-square approximation. We shall see later that p and $\tilde{p}$ can differ considerably for contingency tables with small cell counts.

The main advantage of using p rather than $\tilde{p}$ is that it is guaranteed to bound the type 1 error rate of the **hypothesis testing** procedure to any desired **level**. Moreover, this guarantee is provided unconditionally even though each P value, p, is calculated conditionally by restricting attention to a specific reference set Γ. To see this, let

$$\mathcal{S}(\Gamma) = \Pr(p \leq \alpha | \Gamma). \quad (4)$$

That is, $\mathcal{S}(\Gamma)$ is the conditional type 1 error rate of a level-α hypothesis testing procedure in which you repeatedly generate $r \times c$ tables from the same reference set, Γ, under the null hypothesis, and reject whenever $p \leq \alpha$. Under the null hypothesis, $\mathcal{S}(\Gamma) \leq \alpha$. Now the unconditional type 1 error rate, where Γ may be different each time you execute the test, is

$$\mathcal{S} = \sum \mathcal{S}(\Gamma) \Pr(\Gamma), \quad (5)$$

the sum being taken over all possible reference sets, Γ. Notice that (5) is a weighted sum of terms of the form $\mathcal{S}(\Gamma)$, where each such term is less than or equal to α, the weights, $\Pr(\Gamma)$, are positive, and they sum to 1. Thus,

$$\mathcal{S} \leq \alpha.$$

That is, the guaranteed protection against the type 1 error of an exact conditional hypothesis test also applies unconditionally. Note, however, that this guarantee does not hold if you use $\tilde{p}$ rather than p in the decision to reject the null hypothesis, since $\Pr(\tilde{p} \leq 0.05 | \Gamma) \leq \alpha$ holds only asymptotically.

Choosing the Test Statistic

As stated previously, the reference set Γ is ordered by the test statistic D. Here, we define D for three

important classes of problems; tests on unordered $r \times c$ contingency tables, tests on singly ordered $r \times c$ contingency tables and tests on doubly ordered $r \times c$ contingency tables.

When both the row and column classifications of the table are nominal the table is said to be unordered and the Fisher, Pearson, and **likelihood ratio test** statistics are the most appropriate. Tests based on these three statistics are known as omnibus tests for they are powerful against any general alternative to the null hypothesis of no row by column interaction (*see* **Chi-square Tests**).

Fisher's exact test orders each table, $\boldsymbol{y} \in \Gamma$, in proportion to its hypergeometric probability, $P(\boldsymbol{y})$, given by (2). Fisher [18] originally proposed this test for the single 2×2 contingency table. The idea was extended to tables of higher dimension by Freeman and Halton [19]. Thus, this test is also referred to as the Freeman–Halton test. Asymptotically, under the null hypothesis of no row by column interaction, $-2 \log \gamma P(\boldsymbol{y})$ has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom, where γ is a normalizing constant [31].

The Pearson test orders the tables in Γ according to their Pearson chi-squared statistics. Thus, for each $\boldsymbol{y} \in \Gamma$ the test statistic is

$$D(\boldsymbol{y}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - m_i n_j / N)^2}{m_i n_j / N}. \qquad (6)$$

Asymptotically, under the null hypothesis of no row by column interaction, the Pearson statistic has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

The Likelihood Ratio test orders the tables in Γ according to the likelihood ratio statistic. Specifically, for each $\boldsymbol{y} \in \Gamma$ the test statistic is

$$D(\boldsymbol{y}) = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij} \log \left(\frac{y_{ij}}{m_i n_j / N} \right). \qquad (7)$$

In many textbooks, this statistic is denoted by G^2. Asymptotically, under the null hypothesis of no row by column interaction, $D(\boldsymbol{y})$ has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

When there is a natural ordering of the columns of the $r \times c$ table, but the row classifications are based on nominal categories, appropriate tests are the Kruskal–Wallis test [25] (*see* **Nonparametric Methods**), and its generalization, the one-way **analysis of variance** (ANOVA) test [37]. For example, suppose that the r rows represent r different drug therapies, and the c columns represent c distinct ordered responses (such as, no response, mild response, moderate response, severe response, etc.). One is interested in testing the null hypothesis that the r drugs have the same multinomial response rates. The Kruskal–Wallis and generalized one-way ANOVA tests are more powerful than the Fisher, Pearson, or likelihood Ratio tests for testing this null hypothesis against ordered alternatives which imply that some of these r drugs are more responsive than others. These tests take advantage of the natural ordering of the columns by assigning a rank or column **score** to all the observations in a column. The test statistic is obtained as a quadratic function of an r-dimensional vector whose components are formed by summing the column scores of the observations in each of the r rows and standardizing each sum. For the Kruskal–Wallis test, the observations in a column are assigned their midrank and the special case, $r = 2$, yields the **Wilcoxon–Mann–Whitney** rank-sum test. For the generalized one-way ANOVA test, any monotone scores may be assigned. By suitable choice of these scores, one can construct a large number of tests, including the **normal scores**, exponential scores (*see* **Order Statistics**), and **logrank tests** as special cases. The test statistics for all these tests are given in Chapter 18 of the **StatXact** User Manual [48]. Asymptotically, they are all distributed as chi-square, with $(r-1)$ degrees of freedom under the null hypothesis of no row by column interaction.

When the $r \times c$ contingency table has a natural ordering along both its rows and its columns, the Jonckheere–Terpstra test [24] and the linear-by-linear association test [3] have more power than the Kruskal–Wallis test or the various $(r-1)$ degree of freedom generalized ANOVA tests. For example, suppose the r rows represent r distinct drug therapies at progressively increasing doses and the c columns represent c ordered responses. Now one would be interested in detecting alternatives to the null hypothesis in which drugs administered at larger doses are more responsive than drugs administered at smaller doses. The Jonckheere–Terpstra and linear-by-linear association test statistics cater explicitly to such alternatives for they are better able to pick up departures from the null hypothesis in which the response distribution shifts progressively toward the

right as we move down the rows of the contingency table. The Jonckheere–Terpstra statistic is the normalized sum of $r(r-1)/2$ Wilcoxon rank-sum statistics formed by taking all possible pairs of rows from the r rows of the observed $r \times c$ contingency table and computing a Wilcoxon rank-sum statistic for each resulting $2 \times c$ contingency table. The linear-by-linear association statistic is obtained by standardizing $\sum_{i,j} u_i v_j y_{ij}$, where the u_i's are arbitrary row scores and the v_j's are arbitrary column scores. The row scores often represent progressively increasing doses of a treatment, while the column scores often represent progressively increasing levels of response to treatment. If the u_i's and v_j's represent the original raw data, the linear-by-linear test is a test of significance for Pearson's **correlation** coefficient. However, if the raw data are replaced by Wilcoxon midrank scores, we have a test of **Spearman's rank** correlation coefficient. Refer to Chapter 22 of the StatXact User Manual [48] for the precise functional forms of the Jonckheere–Terpstra and the linear-by-linear test statistics. Under the null hypothesis of no row by column interaction these test statistics are normally distributed. The special case, $r = 2$, yields the family of two-sample linear **rank** tests. For these tests, row scores are irrelevant but a large number of different column scores, covering most of the important nonparametric tests, are listed in Chapters 7, 17 and 18 in the StatXact User Manual [48]; see also **Linear Rank Tests in Survival Analysis**; **Nonparametric Methods**; **Rank Correlation**; **Isotonic Inference**.

Extension to Continuous Data

The methods described above extend naturally to continuous data. In principle, such data can also be represented as contingency tables but the columns of these tables will sum to 1. Thus, these methods provide a unified approach to handling nonparametric data both for the categorical case and the more traditional continuous case. For example, consider the two-sample problem involving continuous data displayed in Table 2. The two groups are "males" and "females". The continuous variable being compared in the two groups is "monthly income".

These data can be represented by the 2×8 contingency table, displayed as Table 3, which may then be permuted in the usual way for exact inference.

The same idea extends to continuous K-sample data with or without stratification, and with or without censoring.

Stratified 2 × 2 Contingency Tables

A very important class of exact nonparametric tests and confidence intervals is defined on data in the form of s 2×2 contingency tables. The ith such table is displayed in Table 4 below.

We may regard the two columns of each table as arising from two independent **binomial distributions**. Specifically, let (x_{i1}, x_{i2}) represent the number of successes in (n_{i1}, n_{i2}) Bernoulli trials, with respective success probabilities (π_{i1}, π_{i2}). The **odds ratio** for the ith table is defined as

$$\Psi_i = \left(\frac{\pi_{i2}}{1-\pi_{i2}}\right) \Big/ \left(\frac{\pi_{i1}}{1-\pi_{i1}}\right). \tag{8}$$

Stratified 2×2 contingency tables arise commonly in prospective studies with binary end points as well as in retrospective **case–control studies**. Thus, although we have specified that the two columns of the 2×2 table represent two independent binomial distributions, this is just a matter of notational convenience. We could equivalently assume that the two rows represent the disease status (present or absent) and the two columns represent the exposure status (not exposed or exposed) in the ith of s matched sets.

Table 2 Two-sample continuous data represented the traditional way

M	M	M	M	F	F	F	F
2010	3100	2555	2095	1990	2122	1875	2550

Table 3 Two-sample continuous data represented as a 2×8 contingency table

Rows	Col_1	Col_2	Col_3	Col_4	Col_5	Col_6	Col_7	Col_8	Row_total
Male	0	0	1	1	0	0	1	1	4
Female	1	1	0	0	1	1	0	0	4
Col_tot	1	1	1	1	1	1	1	1	8
Col_score	1875	1990	2010	2095	2122	2550	2555	3100	

Table 4 Layout for the ith of s 2×2 contingency tables

Rows	Col_1	Col_2	Row_total
Row_1	x_{i1}	x_{i2}	m_{i1}
Row_2	x'_{i1}	x'_{i2}	m_{i2}
Col_tot	n_{i1}	n_{i2}	N_i

We shall be interested in deriving an exact test for the null hypothesis that

$$\Psi_1 = \Psi_2 = \cdots = \Psi_s = \Psi. \tag{9}$$

This is known as the homogeneity test. Next, under the assumption of homogeneity, we shall be interested in computing an exact **confidence interval** for the common odds ratio, Ψ, and in testing that it equals 1.

Homogeneity Test

Let $\boldsymbol{x}$ denote the observed collection of s 2×2 contingency tables, where the ith table in this collection is displayed in Table 4, and define

$$t = x_{11} + x_{21} + \cdots + x_{s1}. \tag{10}$$

Let Ω denote a reference set of collections of s 2×2 contingency tables whose margins are fixed at the values that were actually observed:

$$\Omega = \left\{ \boldsymbol{y}\colon \begin{array}{ll} y_{i1} + y_{i2} = m_{i1}; & y'_{i1} + y'_{i2} = m_{i2} \\ y_{i1} + y'_{i1} = n_{i1}; & y_{i2} + y'_{i2} = n_{i2}. \end{array} \right\} \tag{11}$$

Define the more restricted reference set

$$\Omega_t = \{\boldsymbol{y} \in \Omega\colon y_{11} + y_{21} + \cdots + y_{s1} = t\}. \tag{12}$$

Zelen [53] has shown that under the null hypothesis of homogeneity (9)

$$\Pr(\boldsymbol{x}|\boldsymbol{x} \in \Omega_t) = \frac{\prod_{i=1}^{s} \prod_{j=1}^{2} \binom{n_{ij}}{x_{ij}}}{\sum_{\boldsymbol{y} \in \Omega_t} \prod_{i=1}^{s} \prod_{j=1}^{2} \binom{n_{ij}}{y_{ij}}}. \tag{13}$$

An exact test for the homogeneity of odds ratios can thus be constructed by ordering all elements $\boldsymbol{y} \in \Omega_t$ according to the test statistic

$$D(\boldsymbol{y}) = -\log \Pr(\boldsymbol{y}|\boldsymbol{y} \in \Omega_t) \tag{14}$$

and computing the exact P value

$$p = \sum_{D(\boldsymbol{y}) \geq D(\boldsymbol{x})} \Pr(\boldsymbol{y}|\boldsymbol{y} \in \Omega_t). \tag{15}$$

This test is known as Zelen's exact test. A statistic proposed by Breslow and Day [12] is approximately distributed as chi-square with $(s-1)$ degrees of freedom under the null hypothesis (*see* **Breslow–Day Test**).

Common Odds Ratio Inference

Exact inference about Ψ, the common odds ratio, is based on the conditional distribution of

$$T = y_{11} + y_{21} + \cdots + y_{s1} \tag{16}$$

given $\boldsymbol{y} \in \Omega$. It is shown in [32] that

$$\Pr(T = t|\boldsymbol{y} \in \Omega) = \frac{C_t \Psi^t}{\sum_u C_u \Psi^u}, \tag{17}$$

where

$$C_t = \sum_{\boldsymbol{y} \in \Omega_t} \prod_{i=1}^{s} \prod_{j=1}^{2} \binom{n_{ij}}{y_{ij}}, \tag{18}$$

and the denominator of (17) is simply the normalizing constant obtained by summing over all possible values of u in the range $t_{\min} \leq u \leq t_{\max}$.

To test the null hypothesis that $\Psi = 1$ and to compute an exact confidence interval for this common odds ratio, we need the coefficients C_t for all possible values of t. Network algorithms for this and related computations are described in [32]. Once these coefficients have been obtained, the conditional distribution of t for any value of Ψ can be generated by (17) and hypothesis tests and confidence intervals may thereby be obtained as shown in the above references.

Asymptotic inference for Ψ is usually based on the popular **Mantel–Haenszel** [28] method.

Extension to Stratified $2 \times c$ Contingency Tables

In this section, we discuss inference on stratified $2 \times c$ tables, where the ith of s such tables is displayed as Table 5.

This collection of s $2 \times c$ tables, denoted by $\boldsymbol{x}$, can accommodate two situations; two multinomial populations, and c binomial populations. For both cases,

Table 5 Layout for the ith of s $2 \times c$ contingency tables

Rows	Col_1	Col_2	...	Col_c	Row_total
Row_1	x_{i1}	x_{i2}	...	x_{ic}	m_{i1}
Row_2	x'_{i1}	x'_{i2}	...	x'_{ic}	m_{i2}
Col_tot	n_{i1}	n_{i2}	...	n_{ic}	N_i
Col_score	v_{i1}	v_{i2}	...	v_{ic}	

we assume that data are stratified into s independent strata. Inference is conditional on ordering all three-way collections of s $2 \times c$ tables in the conditional reference set

$$\Lambda = \left\{ \mathbf{y}\colon\ y_{ij} + y'_{ij} = n_{ij}, \forall ij; \right.$$
$$\left. \sum_{j=1}^{c} y_{ij} = m_{i1},\ \sum_{j=1}^{c} y'_{ij} = m_{i2}, \forall i \right\} \tag{19}$$

according to some discrepancy measure $D(\mathbf{y})$. We shall be concerned in this section with the special case where the c columns of each $2 \times c$ contingency table have a natural ordering. In this case, an appropriate (unstandardized) discrepancy measure is the linear rank test statistic

$$t(\mathbf{y}) = \sum_{i=1}^{s} \sum_{j=1}^{c} v_{ij} y_{ij} \tag{20}$$

where the v_{ij}'s are arbitrary column scores.

Two Multinomial Populations. The two rows of stratum i represent two independent multinomial populations. Each observation falls into one of c ordinal response categories. Thus, x_{ij} is the number of stratum-i observations, out of a total of m_{i1}, falling into ordered category j for population 1, and x'_{ij} is the number of stratum-i observations out of a total of m_{i2} falling into ordered category j for population 2. The Wilcoxon rank-sum test, the Normal scores test, the Savage test, and the logrank test are examples of tests that are applicable to such data. The v_{ij} scores for these tests are defined in StatXact-3 [48, Chapter 15].

Several Binomial Populations. The c columns of stratum i represent c independent binomial populations with row 1 representing successes and row 2 representing failures. For population j in stratum i, there are x_{ij} successes and x'_{ij} failures in n_{ij} independent Bernoulli trials. The Cochran–Armitage Trend test and the Permutation test with arbitrary scores are applicable to such data, and determine whether the success rates of the c populations are the same, as against the alternative that they follow an increasing or decreasing trend (*see* **Trend Test for Counts and Proportions**). The scores, $v_{i1}, v_{i2}, \ldots v_{ic}$ typically represent doses, or levels of exposure, affecting the success rates of the c binomial populations. Often one uses the equally spaced scores $v_{ij} = j$ for all i.

We shall assume that there exists no three-factor interaction between rows, columns, and strata, although an analogue of Zelen's test does exist for assessing homogeneity in $2 \times c$ tables. This is an extension of the exact test of homogeneity for 2×2 tables that we will discuss further on. Given that there is no three-factor interaction, we are interested in testing the null hypothesis that the row and column classifications in each stratum are independent. This is known as the hypothesis of conditional independence. One can show that, for both the two multinomial and the c binomial settings under the null hypothesis of conditional independence, the probability of observing $\mathbf{y}$ given $\mathbf{y} \in \Lambda$ is

$$\Pr(\mathbf{y}|\mathbf{y} \in \Lambda) = \frac{\prod_{i=1}^{s} \prod_{j=1}^{c} \binom{n_{ij}}{y_{ij}}}{\prod_{i=1}^{s} \binom{N_i}{m_{i1}}}. \tag{21}$$

The exact one-sided P value for testing the null hypothesis of conditional independence is therefore

$$p_1 = \sum_{t(\mathbf{y}) \geq t(\mathbf{x})} \Pr(\mathbf{y}|\mathbf{y} \in \Lambda). \tag{22}$$

The exact two-sided P value is defined by reflecting the observed value of the test statistic an equal distant away from its mean in the opposite tail; see StatXact [48] for details.

Test for Interaction in 2 × c Tables

We wish to test whether the set of odds ratios (1) describing association between a binomial response and exposure, or (2) whether the set of odds ratios describing association between an ordered multinomial response and a dichotomous **covariate**, vary

across strata. This represents a generalization of the Breslow–Day and Zelen tests for homogeneity already described for the stratified 2×2 setting. These tests can be extended naturally regardless of the underlying sampling mechanism. However, as the derivation of these extension to $2 \times c$ tables differs for multinomial versus binomial sampling, we describe the underlying formulation for binomial data. A derivation for multinomial data is given in StatXact [48].

As before, the data are in the form of s $2 \times c$ contingency tables consisting of two rows, c columns, and s strata. Let π_{jk} be the "success" probability associated with population j in stratum k. We begin with a general logistic model for the binomial response probabilities. Specifically, consider the logistic regression model

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_k + (\beta + \lambda_k)w_j. \tag{23}$$

Identifiability of model parameters requires a constraint such as $\sum_{k=1}^{s} \lambda_k = 0$ or $\lambda_1 = 0$. Denote the set of $s(c-1)$ odds ratios that describe associations in the $s \times 2 \times c$ table by

$$\Psi_{jk} = \frac{\pi_{jk}(1-\pi_{1k})}{\pi_{1k}(1-\pi_{jk})}, \tag{24}$$

$k = 1, \ldots, s$ and $j = 2, 3, \ldots, c$. The model yields the odds ratio model

$$\log \Psi_{jk} = (\beta + \lambda_k)w_j, \tag{25}$$

so that the stratum specific sets of $(c-1)$ odds ratios describing association between rows and columns are allowed to vary across strata. Given model (23) and the **identifiability** constraint $\lambda_1 = 0$, the null hypothesis of no interaction across strata is

$$H_0 : \lambda_2 = \cdots = \lambda_s = 0. \tag{26}$$

An asymptotic test can be derived using likelihood-based methods. For binomial data, such a test compares the logistic model (23) that includes the appropriate interaction terms to the logistic model that does not.

Whether the data arise from binomial or multinomial populations, an exact test of interaction is derived by considering a restricted conditional reference set $\Lambda^I = \{\mathbf{y} \in \Lambda : \sum_i \sum_j w_{ij} y_{ij} = t\}$, where Λ is the conditional reference set given in (19), and $t = \sum_i \sum_j w_{ij} x_{ij}$. This additional constraint ensures the elimination under H_0 of all nuisance parameters contained in either the logistic model for stratified binomial populations or the loglinear model for stratified multinomial populations. A probability length test can therefore be carried out by ordering the tables in Λ^I according to their hypergeometric probabilities under the null hypothesis of no interaction. This framework provides a generalization of Zelen's test to stratified $2 \times c$ tables. Further details are given in StatXact [48].

Stratified $2 \times c$ Contingency Tables with Clustered Data

When the data arise from cluster-correlated binomial populations, the dependence among observations within clusters leads to what is known as "extrabinomial variation" or **"overdispersion"** (e.g. see [38], Chapter 6). An investigator must account for this extra variability to obtain an accurate P value. If the researcher naively treats observations within a cluster as independent – for example, by collapsing over the clusters and using the standard trend test already discussed – the most common result is an anticonservative P value, or a P value that is inaccurately smaller due to an artificial inflation of sample size. The question, however, is fundamentally the same as the one addressed in introducing the trend test for several ordered binomial populations described previously: is there an increasing (or decreasing) average success rate for increasing levels of the risk factor under investigation? Many well-known methods may be employed in evaluating this hypothesis for clustered binomial data, including **random-effects** models and **marginal models**. However, these methods are justified by large-sample distributional approximations, and they may fare poorly with samples that are small or sparse.

The formulation is slightly different from that shown in Table 5. In this case, in the ith stratum, there are c_i clusters, with all members of any one cluster exposed to some distinct level of the **risk factor** of interest. The exposure level for the jth cluster in stratum i is quantified by the "score" w_{ij}. The data can hence be represented in stratified form as a collection of s tables, where the ith of these tables is $2 \times c_i$. Note that j is always indexed over $j = 1, \ldots, c_i$, and i is always indexed as $i =$

Table 6 Layout for the ith of s $2 \times c_i$ contingency tables comprised of cluster-correlated binary populations

	Stratum k				
Rows	Col_1	Col_2	...	Col_c_k	Row-total
Row_1	y_{1k}	y_{2k}	...	$y_{c_k k}$	m_k
Row_2	y'_{1k}	y'_{2k}	...	$y'_{c_k k}$	m'_k
Col-total	n_{1k}	n_{2k}	...	$n_{c_k k}$	N_k
Col-score	w_{1k}	w_{2k}	...	$w_{c_k k}$	

$1, \ldots, s$. The ith such table is displayed in Table 6. For unstratified data, $s = 1$.

The tabular representation of stratified *uncorrelated* binomial populations shown in Table 5 has an equal number of populations across strata, where the number of populations is determined by the number of exposure levels. In the case of *correlated* binomial data, however, the number of populations for a given stratum is equal to the number of clusters within that stratum. Hence, the number of populations may differ from stratum to stratum.

In addition, the conditional inference for clustered binomial data varies slightly from that described for uncorrelated data. First, for notational convenience, let

$$u = \sum_{i=1}^{s} \sum_{j=1}^{c_i} y_{ij} y'_{ij}. \tag{27}$$

Analogous to (19), we define the reference set Λ^C as all possible three-way collections of $2 \times c_i$ contingency tables – for $i = 1, \ldots, s$ – whose row and column margins are fixed at the corresponding values of the observed three-way collection of tables, **x**, displayed above, with one additional constraint:

$$\Lambda^C = \left\{ \mathbf{y}\text{: } \mathbf{y} \text{ is } 2 \times c_i \ \forall i;\ y_{ij} + y'_{ij} = n_{ij},\ \forall\ i, j; \right.$$
$$\left. \sum_{i=1}^{c_i} y_{ij} = m_i, \sum_{i=1}^{s} \sum_{j=1}^{c_i} y_{ij} y'_{ij} = u \right\}. \tag{28}$$

The primary difference between the Λ defined in (19) for independent data and Λ^C defined here is the additional constraint that $\sum_{i=1}^{s} \sum_{j=1}^{c_i} y_{ij} y'_{ij} = u$. This conditions out the nuisance overdispersion effect (or the effect due to extra variation) induced by dependence among observations within clusters. Corcoran et al. [14], show that this conditioning leads to a distribution of **Y** that is known, thereby making exact inference possible even where the binomial populations are clustered. Under the null hypothesis of no row and column interaction, the probability of observing any specific $\mathbf{y} \in \Lambda^C$ is

$$\Pr(\mathbf{Y} = \mathbf{y}) = P(\mathbf{y}) = \prod_{i=1}^{s} \frac{m_i! m'_i! \Pi_{j=1}^{c_i} n_{ij}!}{N_i! \Pi_{j=1}^{c_i} y_{ij}! y'_{ij}!}. \tag{29}$$

The linear rank test statistic for clustered binomial data has a form similar to (20):

$$T = \sum_{i=1}^{s} \sum_{j=1}^{c_i} w_{ij} Y_{ij}, \tag{30}$$

where the w_{ij} are chosen to accurately reflect the effect of exposure on the average response probability. As in the uncorrelated case, the observed statistic $t^* = \sum_{i=1}^{s} \sum_{j=1}^{c_i} w_{ij} y_{ij}$ will be used to test the null hypothesis that there is no association between the two rows and c_i columns of each of the s strata. This inference is based on assessing how extreme the observed statistic t^* is relative to other values of t that could have been observed under the null hypothesis of interest. Analogous to the uncorrelated case, in making this assessment, it is convenient to restrict attention to all possible values of $\mathbf{y} \in \Lambda^C$. With the possible values of the test statistic arising from Λ^C ordered, and their exact probabilities determined from (29), the one-sided P value is computed in a manner identical to that shown in (22), and the exact two-sided P value is defined by reflecting the observed value of the test statistic an equal distant away from its mean in the opposite tail; see StatXact [48] for details.

Calculating Exact Power and Sample Size

For unstratified 2×2 and $2 \times c$ contingency tables, we can compute the **power** of the exact test versus a specific **alternative hypothesis**. However, this is a somewhat more computationally challenging problem in comparison to exact conditional inference, as power must be calculated unconditionally.

For example, suppose that an investigator wishes to compare a new drug to a standard therapy, and hence plans an experiment where 50 subjects will be randomized to each of the two treatment groups (*see* **Clinical Trials, Overview**). Note that before the

trial is complete the investigator does not know how many of the 100 subjects will exhibit a response. Once the trial is finished, there are legitimate reasons to condition on the total number of responses (e.g. the ancillarity principle, the sufficiency principle, and the randomization principle) in order to compare the success probabilities of the two regimens. However, when assessing the power of the study *a priori* against a specific alternative, the investigator must consider all 101 possible outcomes: there may be any number of total responses between 0 and 100.

We illustrate first by describing exact unconditional power computations for comparing two binomial proportions, and then for comparing several ordered binomial proportions against an ordered alternative. Such calculations are also available for two ordered multinomial populations, although we do not describe this derivation here (see [23]).

Exact Power when Comparing two Binomial Populations

Consider first the problem of sampling from two independent binomial populations, where π_j is the binomial probability, n_j is the sample size, and x_j is the binomial response of population j, $j = 1, 2$. The observed data may thus be represented as a single 2×2 contingency table, an example of which is shown in Table 4.

We wish to compute the power of tests of the null hypothesis

$$H_0 : \pi_1 = \pi_2 \equiv \pi \tag{31}$$

versus the two-sided alternative hypothesis

$$H_1 : \pi_1 \neq \pi_2 \tag{32}$$

at fixed sample sizes n_1 and n_2.

As discussed previously, the exact probability of $\boldsymbol{x}$ under H_0, conditional on $x_1 + x_2 = m$, is given by

$$\Pr(\boldsymbol{x}|m, H_0) = \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{N}{m}}. \tag{33}$$

Notice that (33) does not depend on the common null response probability π. Thus, this probability need not be specified for purposes of calculating power. The two response probabilities π_1 and π_2 are, however, needed to evaluate the probability of $\boldsymbol{x}$ under H_1.

We will only be concerned here with two-sided tests. To test (31) versus (32), we will consider Fisher's exact test, Pearson's exact test, and the likelihood ratio exact test. For notational convenience, we will denote all three of these statistics by the symbol T.

Our goal is to compute the exact power of two-sided level-α tests based on the statistic T. Recalling that the sample sizes n_1 and n_2 are already fixed, let

$$\Gamma_m = \{\boldsymbol{x} : \; x_1 + x_2 = m\} \tag{34}$$

and define its **critical region**

$$\Gamma_m(t) = \{\boldsymbol{x} \in \Gamma_m : \; T \geq t\}. \tag{35}$$

The exact null distribution of T may then be obtained by evaluating

$$\Pr(T \geq t|m, H_0) = \sum_{\boldsymbol{x} \in \Gamma_m(t)} \left[\frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{N}{m}} \right], \tag{36}$$

for each possible value of t.

Let α be the maximum allowable type-1 error and $t_\alpha(m)$ be the smallest possible cut-off such that

$$\Pr(T \geq t_\alpha(m)|m, H_0) \leq \alpha. \tag{37}$$

The *conditional* power is defined as

$$\Pr(T \geq t_\alpha(m)|m, H_1) = \sum_{\boldsymbol{x} \in \Gamma_m(t_\alpha(m))} \times \left[\frac{\prod_{j=1}^{2} \binom{n_j}{x_j} \pi_j^{x_j} (1 - \pi_j)^{n_j - x_j}}{\sum_{\boldsymbol{x} \in \Gamma_m} \prod_{j=1}^{2} \binom{n_j}{x_j} \pi_j^{x_j} (1 - \pi_j)^{n_j - x_j}} \right]. \tag{38}$$

Denote this conditional power by $\beta(m)$. Finally, the *unconditional* power is defined as

$$\beta = \sum_{m=0}^{N} \beta(m) P(m) \tag{39}$$

where

$$P(m) = \Pr(x_1 + x_2 = m|H_1), \tag{40}$$

a convolution of two binomials under H_1. It is relatively straightforward to compute (39) as only 2×2 tables are involved.

Example Suppose $n_1 = n_2 = 5$. The exact distribution of Pearson's statistic, conditional on $x_1 + x_2 = 5$, is given in Table 7 under both the null hypothesis ($\pi_1 = \pi_2 \equiv \pi$), and under the alternative hypothesis ($\pi_1 = 0.2, \pi_2 = 0.8$). Note that the null distribution does not depend on the common value of π, which need not be specified.

From these distributions, it is clear that $t_{0.05}(5) = 10$ so that $\beta(5) = 0.336$. That is, the exact conditional power of Pearson's test conducted at the 5% significance level, given $m = 5$, is 34%. Unconditional power is a weighted sum of conditional powers, one for each value of m, as computed by (39). Table 8 displays each possible value of m, its weight, and the corresponding conditional power. Values of m whose contributions to total power are less than 0.00001 have been ignored in this table as they will not affect the first three decimal digits of the answer.

The exact unconditional power is thus

$$\beta = \sum_{m=0}^{9} \beta(m) P(m) = 0.376 \tag{41}$$

Table 7 Exact conditional distributions under null and alternative hypotheses for comparing two independent binomial populations

t	$\Pr(T \geq t \mid 5, H_0)$	$\Pr(T \geq t \mid 5, H_1)$
0.4	1.000	1.000
3.6	0.206	0.861
10	0.007	0.336

Table 8 Conditional power for possible values of m

m	$\beta(m)$	$P(m)$
0	0.533	0.212
1	0.533	0.212
2	0.533	0.212
3	0.533	0.212
4	0.642	0.302
5	0.671	0.237
6	0.713	0.111
7	0.563	0.033
8	0.408	0.006
9	0.533	0.212
10	0.181	0.001

This idea has been extended to power and sample size computations for two ordered multinomial populations by Hilton and Mehta [23], and to several ordered binomial populations by Mehta et al. [35]. Corcoran [15] extend this latter work to compare the exact operating characteristics of the exact trend test versus the asymptotic trend test (*see* **Trend Test for Counts and Proportions**). Further details can be found in StatXact [48].

Unconditional Exact Inference for 2 × 2 Tables

To this point, we have based our inference on conditioning to eliminate nuisance parameters under the null hypothesis. It is possible also, though computationally much more challenging, to eliminate nuisance parameters unconditionally. Because of the relatively greater computational complexity required for unconditional exact inference, such tools are not yet widely available aside from those used to analyze 2×2 tables. We describe one such method here.

We are interested in comparing two population proportions. The starting point is the 2×2 contingency table, $\boldsymbol{x}$, displayed as Table 4 with $s = 1$. As we consider only a single stratum, for convenience, we will drop the subscript i. Without loss of generality, suppose that the row variable specifies the "outcome" of interest, so that we may consider the column totals n_1 and n_2 fixed by design. Then x_1 and x_2 are realizations, respectively, of Binomial(n_1, π_1) and Binomial(n_2, π_2) random variables, where π_1 and π_2 individually represent the outcome probabilities for each population. We wish to test

$$H_0: \pi_1 = \pi_2 = \pi, \tag{42}$$

against two-sided alternatives of the form

$$H_2: \pi_1 \neq \pi_2. \tag{43}$$

Recall that this table was created by taking n_j independent Bernoulli samples from population j, and observing x_j successes, $j = 1, 2$. The unconditional probability of observing $\boldsymbol{x}$ under H_0 is $f_0(\boldsymbol{x})$, specified by (17) with $\Psi = 1$. In order to compute an exact P value, we need to specify a reference set of 2×2 contingency tables and sum the probabilities of tables that are at least as extreme as $\boldsymbol{x}$ in it. For conditional inference, we used the reference set Ω in which both the row and column sums of the 2×2

tables were fixed at their observed values. Unconditional inference uses a larger reference set of 2×2 contingency tables in which only the column sums, or the binomial sample sizes, are fixed. The row sums are treated as random variables. Denote this reference set by

$$\Omega^* = \{\boldsymbol{y}\colon y_j + y_j = n_j,\ j = 1, 2\}, \qquad (44)$$

and order each table $\boldsymbol{y} \in \Omega^*$ according to the test statistic

$$D(\boldsymbol{y}) = \frac{\hat{\pi}_2 - \hat{\pi}_1}{\sqrt{\left(\frac{y_1 + y_2}{N}\right)\left(\frac{y_1 + y_2}{N}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad (45)$$

where $\hat{\pi}_j = y_j/n_j$, $j = 1, 2$. If $y_1 = y_2 = 0$, or $y_1 = y_2 = 0$, set $D(\boldsymbol{y}) = 0$. The denominator of (45) is the **standard error** of the observed difference of binomial proportions under the null hypothesis. Therefore, the statistic $D(\boldsymbol{y})$ has a mean of 0 and variance of one under H_0. A large positive value for the observed statistic $D(\boldsymbol{x})$ furnishes evidence against H_1, while a large negative value furnishes evidence against H_1'.

The exact P value is the sum of probabilities of all tables $\boldsymbol{y} \in \Omega^*$ that are more extreme than the observed table $\boldsymbol{x}$ with respect to the test statistic (45). The trouble is that each such extreme table has a probability $f_0(\boldsymbol{y})$, which, by (17), depends on the unknown nuisance parameter, π. In our previous discussion of conditional inference, we were able to eliminate the nuisance parameter by conditioning on its sufficient statistic, m_1. But we can no longer do so because we have specified Ω^* rather than Ω to be the reference set. For exact unconditional inference, we utilize a different argument to eliminate π. We consider all possible values of π in its range and select that value, which produces the largest P value. This produces a conservative answer so that no matter what the true value of π might be, the type-1 error of the test cannot exceed its nominal significance level.

The main advantage of using Ω^* is that it is larger than Ω. Consequently, the distribution of any test statistic usually has more support points if it is defined on Ω^* rather than on Ω. This reduces conservatism, since it is possible to construct exact hypothesis tests whose true significance levels come closer to their nominal significance levels under Ω^* than they do under Ω. The main disadvantage is that the nuisance parameter, π, can only be eliminated by considering all possibilities in its range and catering to the worst case. This increases the conservatism of the hypothesis test. There is thus a trade-off between the advantage gained by enriching the reference set and the disadvantage of catering to the worst case for the nuisance parameter. For a single 2×2 contingency table, Mehta and Hilton (1993) [29] have shown that, on balance, the gain in power from using Ω^* outweighs the loss in power because of catering to the worst case. They go on to show, however, that this advantage quickly evaporates as the dimensions of the table increase from 2×2 to 2×3.

We compute the P value in two stages. At the first stage, we express the P value as a function of π. Then, at the second stage, we obtain the supremum of this function over all values of $\pi \in (0, 1)$. We use this supremum as the P value. Since the P value based on the actual value of π can never exceed the supremum over all possible values of π, this procedure guarantees that the type-1 error will always be preserved. In effect, we compute a conservative P value that will preserve the desired type-1 error rate no matter what the true value of π might be, since it is designed to cater for the worst case.

This test is known as Barnard's test, named after George Barnard who first proposed it as an alternative to Fisher's exact test amidst some controversy. (See [7, 8, 9, 10, 52] for interesting philosophical discussions and references). Its critical values for exact testing of H_0 have been tabulated by Suissa and Shuster (1985) [50]. A proposed restriction by Berger and Boos [11] adds stability and reduces the conservatism of the procedure.

Computational Issues

Computing quantities such as (3) is nontrivial. When analyzing an $r \times c$ table, for instance, the size of the reference set grows exponentially so that explicit enumeration of all the tables in Γ soon becomes computationally infeasible. For example, the reference set of all 5×6 tables with row sums of (7, 7, 12, 4, 4) and column sums of (4, 5, 6, 5, 7, 7) contains 1.6 billion tables. Yet, the tables in this reference set are all rather sparse and unlikely to yield accurate P values based on large-sample theory. Network algorithms have been developed by Mehta, Patel, and

coworkers, [30–32, 34, 36] to enumerate the tables in Γ implicitly. In these algorithms, the reference set is represented by a network of nodes and arcs. A sequence of connected arcs from the starting to the terminal node constitutes a path through the network. Each such path represents one and only one table in Γ. The length of a path equals the value of the test statistic for the table to which that path corresponds. The probability of the path equals the probability of the corresponding table. Thus, the problem of computing an exact P value is equivalent to the problem of identifying paths whose lengths equal or exceed a specified value, and summing the probabilities of all these paths. This can be accomplished by well-known **operations research** techniques such as backward induction and forward probing through the network. These methods are very efficient and make it feasible to compute exact P values.

Alternate approaches are provided by Pagano and Halvorsen [39], Pagano and Tritchler [40], Streitberg and Rohmel [49], Baglivo, Olivier and Pagano [6], Vollset, Hirji, and Elashoff [51], and Cheung and Klotz [13]. Sometimes a data set is too large even for implicit enumeration, yet it is sufficiently sparse that the asymptotic results are suspect. For such situations, a Monte Carlo estimate and associated 99% confidence interval for the exact P value may be obtained. In the Monte Carlo method, tables are sampled from Γ in proportion to their hypergeometric probabilities (2), and a count is kept of all the sampled tables that are more extreme than the observed table. For details, refer to [4, 33, 41] and [45].

Analysis of Data Sets

In this section, we will illustrate the techniques developed in the previous sections with some data analyses. Each example will highlight the different conclusions one might draw if an asymptotic analysis were performed instead of an exact analysis. A large number of additional examples are available at the Cytel web site `http://www.cytel.com`. All results were obtained by the StatXact-3 software package [48].

An Unordered Contingency Table

Data were obtained on the location of oral lesions, in house to house surveys in three geographic regions of rural India, by Gupta, Mehta, and Pindborg [22]. Consider a hypothetical subset of these data displayed by Table 9 as a 9×3 contingency table in which the counts are the number of patients with oral lesions per site and geographic region.

The question of interest is whether the distribution of the site of the oral lesion is significantly different in the three geographic regions. The row and column classifications for this 9×3 table are clearly unordered, making it an appropriate data set for either the Fisher, Pearson, or likelihood ratio tests. The exact and asymptotic P values are displayed in Table 10. There are striking differences between the two methods. The exact analysis suggests that the row and column classifications are dependent, but the asymptotic analysis fails to show this.

A Singly Ordered Contingency Table

The tumor regression rates of five chemotherapy regimens, Cytoxan (CTX) alone, Cyclohexylchloroethyl nitrosourea (CCNU) alone, Methotrexate (MTX) alone, CTX + MTX, and CTX + CCNU + MTX were compared in a small clinical trial. Tumor regression was measured on a three-point scale: no response, partial response, or complete response. The results are tabulated in Table 11.

Table 9 Oral lesions data

Site of Lesion	Kerala	Gujarat	Andhra
Labial mucosa	0	1	0
Buccal mucosa	8	1	8
Commissure	0	1	0
Gingiva	0	1	0
Hard palate	0	1	0
Soft palate	0	1	0
Tongue	0	1	0
Floor of mouth	1	0	1
Alveolar ridge	1	0	1

Table 10 Exact and asymptotic P values for oral lesions data

	Three tests of independence		
Type of inference	Pearson	Fisher	Likelihood ratio
Value of $D(\boldsymbol{x})$	22.1	19.72	23.3
Asymptotic P value	0.1400	0.2331	0.1060
Exact P value	0.0269	0.0101	0.0356

Table 11 Chemotherapy pilot study data

Chemo	No resp.	Partial resp.	Complete resp.
CTX	2	0	0
CCNU	1	1	0
MTX	3	0	0
CTX + CCNU	2	2	0
CTX + CCNU + MTX	1	1	4

Small pilot studies like this one are frequently conducted as a preliminary to planning a large-scale randomized clinical trial. For such data, the Kruskal–Wallis test may be used to determine whether or not the five drug regimens are significantly different with respect to their tumor regression rates. The observed value of the Kruskal–Wallis statistic for this table is 8.682. Referring this value to a chi-square distribution with four degrees of freedom yields an asymptotic P value of 0.0695, which is not significant at the 0.05 level. However, on the basis of the permutation distribution of the Kruskal–Wallis statistic, the exact P value is 0.039, which is statistically significant.

Analysis of Stratified 2 × 2 Contingency Tables

Testing the Homogeneity of Odds Ratios. The binary response data tabulated in Table 12 compare a new drug with a control drug at 22 hospital sites.

The data can be thought of as twenty-two 2×2 contingency tables, one for each site. If you examine the 2×2 tables carefully, you notice that site 15 appears to be different from the others. Whereas all the other sites have a low response rate for both the new drug and the control drug, the response rate of the control drug is 79% at site 15. The Homogeneity test can tell you whether the observed difference at site 15 is a real difference or whether it is just a chance fluctuation due to a small sample. Because of the sparseness in the data, the asymptotic (Breslow–Day) statistic might not yield an accurate P value. The exact (Zelen) test is preferred. The exact P value is 0.0135. Thus, we reject the null hypothesis that there is a common odds ratio across the 22 sites. The data strongly suggest that the odds ratio at site 15 is different from the other odds ratios. The asymptotic (Breslow–Day) P value is much larger (0.0785) and is only marginally significant.

Table 12 Site by treatment interaction data

Test site	New drug		Control drug		Test site	New drug		Control drug	
	Resp	No	Resp	No		Resp	No	Resp	No
1	0	15	0	15	12	0	12	1	11
2	0	39	6	32	13	0	24	5	19
3	1	20	3	18	14	2	10	2	11
4	1	14	2	15	15	0	14	11	3
5	1	20	2	19	16	0	53	4	48
6	0	12	2	10	17	0	20	0	20
7	3	49	10	42	18	0	21	0	21
8	0	19	2	17	19	1	50	1	48
9	1	14	0	15	20	0	13	1	13
10	2	26	2	27	21	0	13	1	13
11	0	19	2	18	22	0	21	0	21

Estimating the Common Odds Ratio

The court case of Hogan versus Pierce [20] involved the minority hiring data displayed in Table 13.

The most notable feature of these data is that at each hiring opportunity not a single black was hired, whereas small numbers of whites were hired. This makes it impossible to use the usual large-sample maximum likelihood or Mantel–Haenszel [28] methods for estimating the odds of being hired for whites relative to blacks. These methods simply fail to converge. Only the exact method provides a valid answer and it shows that the odds of being hired for a white relative to a black are no lower than 2.3 to 1, with 95% confidence.

Table 13 Minority hiring data

Date of hire	Whites		Blacks	
	Hired	Not	Hired	Not
7/74	4	16	0	7
8/74	4	13	0	7
9/74	2	13	0	8
4/75	1	17	0	8
5/75	1	17	0	8
10/75	1	29	0	10
11/75	2	29	0	10
2/76	1	30	0	10
3/76	1	30	0	10
11/77	1	33	0	13

Test of Trend in Stratified 2 × c Contingency Tables

The data for this example were provided by the US Food and Drug Administration (FDA). Animals were treated with four dose levels of a carcinogen and then observed (at necropsy) for the presence or absence of a tumor type. The data were stratified by survival time (in weeks) into the four time-intervals 0 to 50, 51 to 80, 81 to 104, and terminal sacrifice. Since there were no tumors found in the first time-interval, this stratum may be excluded from data entry. The data for the remaining three strata are displayed in Table 14.

We use the stratified Cochran–Armitage trend test (Breslow and Day [12], p. 148) to determine if there is a **dose-response** relationship between the level of carcinogen and the presence of tumors. The test statistic is defined by (20), where v_{ij} is the dose-level of carcinogen and y_{ij} is the number of animals with tumors, at the jth dose level in the ith stratum. The results are tabulated in Table 15.

Table 14 FDA Animal toxicology data

Stratum 1: 51–80 weeks of survival					
	Dose of carcinogen				
Disease status	None	1 unit	5 units	50 units	Total
Tumor present	0	0	0	1	1
Tumor absent	7	10	6	8	31
Stratum 2: 81–104 weeks of survival					
	Dose of carcinogen				
Disease status	None	1 unit	5 units	50 units	Total
Tumor present	0	1	0	1	2
Tumor absent	11	9	13	14	47
Stratum 3: Sacrificed at end of 104 weeks					
	Dose of carcinogen				
Disease status	None	1 unit	5 units	50 units	Total
Tumor present	1	1	1	2	5
Tumor absent	29	26	28	20	103

Table 15 One and two-sided P values for FDA data

P values	One-sided	Two-sided	Double one-sided
Exact	0.0651	0.0769	0.1302
Asymptotic	0.0410	0.0820	0.0820

There are large differences between the exact and asymptotic one-sided P values, and they lead to different conclusions about the significance of the dose-response relationship. They also show that the usual practice of doubling the one-sided P value is unnecessarily conservative with asymmetric distributions. But the most interesting finding of all is that the distribution of the linear rank statistic (20) has multiple towers. A normal approximation would be seriously misleading. This is shown in Figure 1.

Software and Related Resources for Exact Inference

We have presented the essential idea behind exact permutational inference, described one numerical algorithm, referenced others, and shown through several examples that exact inference is a valuable supplement to corresponding asymptotic methods.

Software support for these methods is available in many standard packages including StatXact [48], LogXact [26], SPSS Exact Tests [47], and SAS Version 9 [44]. A brief description of the StatXact and LogXact software packages is given elsewhere (*see* **StatXact**).

Some of the newer textbooks on nonparametric methods, for example, [1, 17, 21, 27, 46] devote considerable space to exact and Monte Carlo methods of inference for **categorical data**. A useful survey paper, in which a unified treatment of exact inference for categorical data is presented through the **loglinear model**, was recently published by Agresti

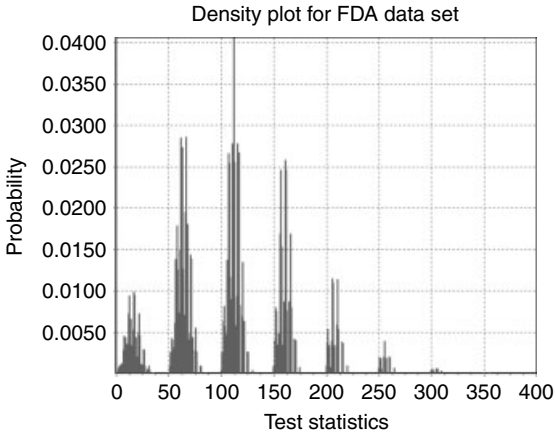

Figure 1 Distribution of trend test statistic for FDA data

[2]. A complete collection of references to statistical methodology, numerical algorithms, commercial software, shareware, and textbooks on exact permutational inference can be obtained by visiting the ExactStats worldwide web site on the Internet. The address is `http://jiscmail.ac.uk/lists/exact-stats.html`.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, New York.

[2] Agresti, A. (1992). A survey of exact inference for contingency tables (with discussion), *Statistical Science* **7**(1), 131–177.

[3] Agresti, A., Mehta, C.R. & Patel, N.R. (1990). Exact inference for contingency tables with ordered categories, *Journal of the American Statistical Association* **85**(410), 453–458.

[4] Agresti, A., Wackerly, D. & Boyett, J.M. (1979). Exact conditional tests for cross-classifications, *Psychometrika* **44**, 75–83.

[5] Agresti, A. & Yang, M. (1987). An empirical investigation of some effects of sparseness in contingency tables, *Communications in Statistics* **5**, 9–21.

[6] Baglivo, J., Olivier, D. & Pagano, M. (1988). Methods for the analysis of contingency tables with large and small cell counts, *Journal of the American Statistical Association* **83**, 1006–1013.

[7] Barnard, G.A. (1945). A new test for 2×2 tables, *Nature* **156**, 177.

[8] Barnard, G.A. (1947). Significance tests for 2×2 tables, *Biometrika* **34**, 123–138.

[9] Barnard, G.A. (1949). Statistical inference, *Journal of the Royal Statistical Society Series B* **11**, 115–139.

[10] Barnard, G.A. (1989). On alleged gains in power from lower p-values, *Statistics in Medicine* **8**, 1469–1477.

[11] Berger, R.L., Boss, D.D. (1994). P Values Maximized Over a Confidence Set for the Nuisance Parameter, *Journal of the American Statistical Association* **89**, 1012–1016.

[12] Breslow, N.E. & Day, N.E. (1980). *The analysis of case-control studies*, IARC Scientific Publications No. 32, Lyon, France.

[13] Cheung, Y.K. & Klotz, J.H. (1997). The Mann Whitney Wilcoxon distribution using linked lists, *Statistica Sinica* **7**(3), 805–813.

[14] Corcoran, C., Ryan, L., Senchaudhuri, P., Mehta, C., Patel, N. & Molenberghs, G. (2001). An exact trend test for correlated data, *Biometrics* **57**, 931–948.

[15] Corcoran, C., Mehta, C.R., Patel, N., Senchaudhuri, P. (1999). Power comparisons for tests of trend in dose-response studies, *Statistics in Medicine* **19**, 3037–3050.

[16] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[17] Edgington, E.S. (1995). *Randomization Tests*, 3rd Ed. Marcel Dekker, New York.

[18] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

[19] Freeman, G.H. & Halton, J.H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance, *Biometrika* **38**, 141–149.

[20] Gastwirth, J.L. (1984). Combined tests of Significance in EEO cases, *Industrial and Labor Relations Review* **38**(1).

[21] Good, P. (1993). *Permutation Tests*. Springer-Verlag, New York.

[22] Gupta, P.C., Mehta, F.R. & Pindborg, J. (1980). *Community Dentistry and Oral Epidemiology* **8**, 287–333.

[23] Hilton, J., Mehta, C.R. (1993). Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* **49**, 609–616.

[24] Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods*. John Wiley, New York.

[25] Landis, R., Heyman, E.R. & Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests, *International Statistical Review* **46**, 237–254.

[26] LogXact Version 5 for Windows. (2004). *Software for Exact Logistic Regression, featuring Cytel Studio*. Cytel Software Corporation, Cambridge, MA.

[27] Manly, B.F.J. (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman & Hall, London.

[28] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[29] Mehta, C.R. & Hilton, J.F. (1993). Exact power of conditional and unconditional tests: going beyond the 2×2 contingency table, *American Statistician* **47**, 91–98.

[30] Mehta, C.R. & Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables, *Journal of the American Statistical Association* **78**(382), 427–434.

[31] Mehta, C.R. & Patel, N.R. (1986). A hybrid algorithm for Fisher's exact test on unordered $r \times c$ contingency tables, *Communications in Statistics* **15**(2), 387–403.

[32] Mehta, C.R., Patel, N.R. & Gray, R. (1985). On computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables, *Journal of the American Statistical Association* **80**(392), 969–973.

[33] Mehta, C.R., Patel, N.R. & Senchaudhuri, P. (1988). Importance sampling for estimating exact probabilities in permutational inference, *Journal of the American Statistical Association* **83**(404), 999–1005.

[34] Mehta, C.R., Patel, N.R. & Senchaudhuri, P. (1992). Exact stratified linear rank tests for ordered categorical and binary data, *Journal of Computational and Graphical Statistics* **1**, 21–40.

[35] Mehta, C.R., Patel, N.R. & Senchaudhuri, P. (1998). Exact Power and Sample-Size Computations for the Cochran-Armitage Trend Test, *Biometrics* **54**, 1615–1621.

[36] Mehta, C.R., Patel, N.R. & Tsiatis, A.A. (1984). Exact significance testing to establish treatment equivalence for ordered categorical data, *Biometrics* **40**, 819–825.
[37] Miller, R.G. (1981). *Simultaneous Statistical Inference*. Springer-Verlag, New York.
[38] Morgan, B.J.T. (1992). *The Analysis of Quantal Response Data*. Chapman & Hall, London.
[39] Pagano, M. & Halvorsen, K. (1981). An algorithm for finding exact significance levels of $r \times c$ contingency tables, *Journal of the American Statistical Association* **76**, 931–934.
[40] Pagano, M. & Tritchler, D. (1983). On obtaining permutation distributions in polynomial time, *Journal of the American Statistical Association* **78**, 435–441.
[41] Patefield, W.M. (1981). An efficient method of generating $r \times c$ tables with given row and column totals. (Algorithm AS 159), *Applied Statistics* **30**, 91–97.
[42] Read, R.C. & Cressie, N.A. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.
[43] Reid, N. (1995). The roles of conditioning in inference (with discussion), *Statistical Science* **10**(2), 138–157.
[44] SAS Version 9 (2004). SAS Institute Inc., Cary, NC.
[45] Senchaudhuri, P., Mehta, C.R. & Patel, N.R. (1995). Estimating exact p-values by the method of control variates, or Monte Carlo rescue, *Journal of the American Statistical Association* **90**(430), 640–648.
[46] Sprent, P. (1993). *Applied Nonparametric Statistical Methods*, 2nd Ed. Chapman & Hall, London.
[47] SPSS Exact Tests for Windows. (1995). SPSS Inc., Chicago.
[48] StatXact Version 6 for Windows. (2004). *Software for Exact Nonparametric Inference, featuring Cytel Studio*. Cytel Software Corporation, Cambridge, MA.
[49] Streitberg, B. & Rohmel, R. (1986). Exact distributions for permutation and rank tests, *Statistical Software Newsletter* **12**, 10–17.
[50] Suissa, S. Shuster, J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial, *Journal of the Royal Statistical Society Series A* **148**, 317–327.
[51] Vollset, S.E., Hirji, K.F. & Elashoff, R.M. (1991). Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables, *Journal of the American Statistical Association* **86**, 404–409.
[52] Yates, F. (1984). Test of significance for 2×2 contingency tables, *Journal of the Royal Statistical Society Series A* **147**, 426–463.
[53] Zelen, M. (1971). The analysis of several 2×2 contingency tables, *Biometrika* **58**(1), 129–137.

(*See also* **Software, Biostatistical**)

CHRISTOPHER D. CORCORAN,
PRALAY SENCHAUDHURI, CYRUS R. MEHTA &
NITIN R. PATEL

Excess Mortality

In the modeling of mortality in clinical or epidemiologic studies, it is sometimes relevant to use the mortality of the general population as a reference for comparisons. Rather than establishing a reference sample of the general population from which the study sample is drawn, one usually relies on published life tables and includes the population mortality rate as a known function in the model for the survival times in the study group. Two classes of hazard rate models have been studied in some detail: *multiplicative hazard rate* models, in which the mortality of the study group is described by multiplying the reference rate by some parameter, *the relative mortality*, which may further depend on specific risk factors or follow-up time; and **additive hazard rate** *models*, in which the reference rate is modified by adding a parameter, *the excess mortality*, which again may depend on specific risk factors or follow-up time. Multiplicative models are related to calculation of *standardized mortality ratios* (SMR), a technique that has been employed by epidemiologists for many years (*see* **Standardization Methods**). However, additive models may be viewed as the theoretical basis for calculation of *relative survival* and the *corrected survival curve*.

In a simple additive hazard rate model, the mortality rate $\lambda_i(t)$ of an individual $i, i = 1, \ldots, n$, in the sample satisfies

$$\lambda_i(t) = \mu_i(t) + \gamma(t),$$

where $\mu_i(t)$ is the *known* population rate at time t for an individual of the same sex and born in the same year as individual i. The excess mortality $\gamma(t)$ is assumed common for all individuals in the sample. The integrated excess mortality is defined as

$$\Gamma(t) = \int_0^t \gamma(u)\,\mathrm{d}u.$$

Unlike the multiplicative model, which is mainly a descriptive means for relating the observed mortality in a sample to population mortality rates, the additive model may be given an interpretation in a **competing risk** framework when the excess mortality rate is positive. If the sample consists of individuals suffering from a given disease, one may consider using the population mortality rate for all other causes of deaths as the known rate $\mu_i(t)$ and the excess rate $\gamma(t)$ will then represent mortality due to the disease. In this situation, the model may therefore permit estimation of *cause-specific mortality* without relying on information about cause of death.

For a sample of individuals $i = 1, \ldots, n$, let x_i denote the time of entry and $X_i \geq x_t$ the survival time (in which case $D_i = 1$) or **censoring** time (in which case $D_i = 0$). In a clinical setting the time t would usually be time since treatment and x_i is typically zero, but the model could also be used with age as the underlying time scale and then left truncation, i.e. $x_i > 0$, will often be present. Define $N(t)$ to be the observed number of deaths in $[0, t]$ and let $Y(t)$ denote the number at risk at time t:

$$Y(t) = \sum_{i=1}^{n} Y_i(t) = \sum_{i=1}^{n} I(x_i < t \leq X_i).$$

Following Andersen & Væth [2], the integrated excess mortality may be estimated by

$$\hat{\Gamma}(t) = \sum_{X_i \leq t} \frac{D_i}{Y(X_i)} - \int_0^t \mu^*(u)\,\mathrm{d}u.$$

The first term is the ordinary **Nelson–Aalen** estimate, and the second term is the integral of the average population mortality rate, $\mu^*(u)$, defined for each u as the average of the population mortality rates corresponding to the individuals at risk at time u

$$\mu^*(u) = \frac{1}{Y(u)} \sum_{i=1}^{n} \mu_i(u) Y_t(u).$$

The variance of $\hat{\Gamma}(t)$ can be estimated by

$$\sum_{X_i \leq t} \frac{D_i}{[Y(X_i)]^2}$$

An estimate of the excess mortality rate $\gamma(t)$ can be obtained by kernel smoothing techniques (see, for example, Andersen et al. [3, Section IV.4.2]) (*see* **Smoothing Hazard Rates**). The survival function

$$S^*(t) = \exp\left(-\int_0^t \mu^*(u)\,\mathrm{d}u\right)$$

derived from the hazard rate $\mu^*(t)$ may be viewed as a continuous time generalization of the so-called Ederer Method II for calculation of the *expected survival curve*, which, unlike Ederer Method I, adjusts for deaths and censoring during follow-up (see [6]

or [10]). Furthermore, an estimate of the "survival" function

$$\exp\left(-\int_0^t \gamma(u)\,\mathrm{d}u\right)$$

for the excess mortality is obtained as the *relative* survival function $\hat{S}(t)/S^*(t)$.

A *parametric* version of the simple additive hazard rate model above has been studied by Buckley [4], who considered **maximum likelihood** estimation in a model *with piecewise constant excess mortality rate*. An iterative procedure is required to solve the likelihood equations, but simple, explicit moment estimates are also available (*see* **Method of Moments**). For the special case with a constant excess mortality rate γ for all $t \in [0, \tau]$, $\tau < \infty$ denoting an upper limit for the observed survival times, one may show [2, 4] that the maximum likelihood estimate is the solution to

$$\sum_{i=1}^{n} \frac{D_i}{\hat{\gamma} + \mu_i(X_i)} = T(\tau),$$

where

$$T(\tau) = \int_0^\tau Y(u)\,\mathrm{d}u = \sum_{i=1}^{n}(X_i - x_i)$$

is the total number of *person–years at risk* during follow-up (i.e. the **total time on test**). The moment estimate is simply

$$\tilde{\gamma} = \frac{N(\tau) - \mathrm{E}(\tau)}{T(\tau)},$$

where

$$\mathrm{E}(\tau) = \sum_{i=1}^{n} \int_0^\tau \mu_i(u) Y_i(u)\,\mathrm{d}u$$

may be interpreted as the expected number of deaths during follow-up, and the moment estimate is therefore the excess number of deaths divided by the total time at risk. The variance of the estimate $\tilde{\gamma}$ can be estimated by

$$\frac{N(\tau)}{[T(\tau)]^2}.$$

Within the framework of parametric models, standard large-sample methods provide **goodness of fit** tests for the constant excess mortality model relative to a piecewise constant excess mortality. Alternatively, generalized *total time on test procedures* are available for assessing the goodness of fit of a constant excess mortality rate (see [2] and [3, Section VI.3.2–3]).

Both parametric and nonparametric regression models generalizing the simple, additive excess mortality model have been developed. The parametric models for the excess mortality rate include, among others, **loglinear regression** models studied by Pocock et al. [8] and Hakulinen & Tenkanen [7], and a linear regression model considered by Campbell [5]. A semiparametric **proportional hazards** regression model for the excess rate has been proposed and studied by Sasieni [9], and Zahl [11] has introduced a linear nonparametric regression model for the excess rate, generalizing Aalen's linear hazard regression model [1].

References

[1] Aalen, O.O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**, 907–925.

[2] Andersen, P.K. & Væth, M. (1989). Simple parametric and nonparametric models for excess and relative mortality, *Biometrics* **45**, 523–535.

[3] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Buckley, J.D. (1984). Additive and multiplicative models for relative survival rates, *Biometrics* **40**, 51–62.

[5] Campbell, M.J. (1985). Multiplicative and additive models with external controls in a cohort study of cancer mortality, *Statistics in Medicine* **4**, 353–360.

[6] Ederer, F., Axtell, L.M. & Cutler, S.J. (1961). The relative survival rate: a statistical methodology, *National Cancer Institute Monographs* **6**, 101–121.

[7] Hakulinen, T. & Tenkanen, L. (1987). Regression analysis of relative survival rates, *Applied Statistics* **36**, 309–317.

[8] Pocock, S.J., Gore, S.M. & Kerr, G.R. (1982). Long term survival analysis: the curability of breast cancer, *Statistics in Medicine* **1**, 93–104.

[9] Sasieni, P.D. (1996). Proportional excess hazards, *Biometrika* **83**, 127–141.

[10] Zahl, P.H. (1995). A proportional regression model for 20 year survival of colon cancer in Norway, *Statistics in Medicine* **14**, 1249–1261.

[11] Zahl, P.H. (1996). A linear non-parametric regression model for the excess intensity, *Scandinavian Journal of Statistics* **23**, 353–364.

(*See also* **Survival Analysis, Overview**)

MICHAEL VÆTH

Excess Relative Risk

Relative risks are commonly used to describe the relationship between the **risks** or rates in different populations. For populations with risks R_0 and R_1, the relative risk is $RR = R_1/R_0$. In many situations, it is useful to describe R_1 as $R_0 + E$, where E is the excess risk. In this case we have $RR = (R_0 + E)/R_0 = 1 + ERR$, where *ERR* represents the excess relative risk. The most commonly used approach to modeling relative risks involves **loglinear models** of the form $RR = e^{\beta z}$. But it is often useful, especially in **dose–response** analyses, to model the *ERR* directly.

When exposures vary over a broad range, simple *ERR* models (e.g. linear in dose) can provide a clearer, and in many cases better, description of the exposure effect on the risk than loglinear risk models. Ratios of *ERR*s are more appropriate than ratios of *RR*s as a summary of the impact of exposure. **Effect modification** and analyses of the joint effects of multiple exposures (*see* **Synergy of Exposure Effects**) are often expressed more naturally in terms of effects on the *ERR* rather than as **interactions** in **multiplicative** relative risk models. For additional details about specific *ERR* models and issues related to the use of these models (*see* **Parametric Models in Survival Analysis**; **Poisson Regression in Epidemiology**).

DALE L. PRESTON

Excess Risk

The excess risk or rate in a population is the difference between the **risk** (rate) R_1 for a population exposed to some risk factor (e.g. radiation or smoking) and the risk R_0 in an otherwise identical population without the exposure. In simple terms the excess risk E is defined as $E = R_1 - R_0$. The excess risk is closely related to the **attributable risk** $AR = E/R_1$ and the **excess relative risk** $ERR = E/R_0$.

Since excess risk models involve the sum of background and excess risks, they are intrinsically additive, and it is common, though potentially confusing, to refer to them as **additive models**. The development of adequate excess risk models generally requires that the risk be modeled as a sum of nonlinear functions describing the background and excess risks. Models for both the background and excess risks often involve multiplicative functions of risk-modifying factors. In contrast to **relative risk models**, for which it is possible to use **semiparametric Cox regression models**, fitting excess risk models generally requires explicit parametric modeling of R_0 or specification of R_0 with external rates.

It is both feasible and useful, however, to model background rates directly for problems involving either excess or relative risk by using modern statistical methods, such as **Poisson regression**. (*See* **Parametric Models in Survival Analysis** for additional information on the modeling of excess risks.)

Relative risk models have come to dominate discussion of risk in epidemiologic studies. However, description in terms of excess risks and rates is important for: understanding the impact of an exposure on risk in a population; developing exposure standards to limit risks to the general public or to special groups; and developing and assessing mechanistic models of the effect of exposure on risk.

Historically, attention has focused on the comparison of simple (usually time-constant) excess and relative risk models [3]. However, when one makes use of more general classes of excess and relative risk models, it is best to view these models as complementary, rather than competing, descriptions of risk.

Excess risk and attributable risk can also be estimated from **population-based case–control studies** [1, 2].

References

[1] Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case–control studies: a review, *Statistics in Medicine* **10**, 1753–1773.

[2] Benichou, J. & Wacholder, S. (1994). A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control studies, *Statistics in Medicine* **13**, 651–661.

[3] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*. Vol. II. The Design and Analysis of Cohort Studies. IARC Scientific Publication No. 82, Oxford University Press, New York.

DALE L. PRESTON

Exchangeability

In probability theory, the **random variables** $Y_1, \ldots, Y_N$ are said to be *exchangeable* (or *permutable* or *symmetric*) if their joint distribution $F(y_1, \ldots, y_N)$ is symmetric; that is, if F is invariant under permutation of its arguments, so that

$$F(z_1, \ldots, z_N) = F(y_1, \ldots, y_N)$$

whenever $z_1, \ldots, z_N$ is a permutation of $y_1, \ldots, y_N$. There is a related epidemiologic usage which is described in the article on **confounding**. In many ways, sequences of exchangeable random variables play a role in subjective **Bayesian** theory analogous to that played by independent identically distributed (iid) sequences in classical frequentist theory. In particular, the assumption that a sequence of random variables is exchangeable allows the development of inductive statistical procedures for inference from observed to unobserved members of the sequence [1–3, 5, 6, 9].

Exchangeable random variables are identically distributed, and iid variables are exchangeable. Now suppose that $Y_1, \ldots, Y_N$ are iid given an unknown parameter θ that indexes their joint distribution (*see* **Identifiability**). Such variables will not be unconditionally independent when θ is a random variable, but will be exchangeable. Consider, for example, the case in which $Y_1, \ldots, Y_N$ have a joint density. The unconditional density of $Y_1, \ldots, Y_N$ will be

$$\begin{aligned} f(y_1, \ldots, y_N) &= \int_\theta f(y_1, \ldots, y_N|\theta)\ \mathrm{d}F(\theta) \\ &= \int \prod_i f(y_i|\theta)\ \mathrm{d}F(\theta). \end{aligned}$$

Exchangeability of $Y_1, \ldots, Y_N$ follows from the identity of the marginal densities in the product. However, given that these densities depend on θ, the integral and product cannot be interchanged, so that $f(y_1, \ldots, y_N) \neq \prod_i f(y_i)$. We thus have that a mixture of iid sequences is an exchangeable sequence, but not iid except in trivial cases.

One consequence of this result is that the usual procedures for generating a sequence $Y_1, \ldots, Y_N$ of iid random variables for inference on an unknown parameter (such as Bernoulli trials of **binary data** with unknown success probability) generate only an exchangeable sequence when the parameter is generated randomly and the sequence is considered unconditionally. From a Bayesian perspective, this means that, when your uncertainty about the parameter is integrated with your uncertainty about the realizations of $Y_1, \ldots, Y_N$, the latter are (for you) exchangeable but dependent. This subjective dependence is immediately clear if you consider (say) tossing a coin $N = 99$ times, with Y_i the indicator of heads on toss i. Starting from a **uniform prior** for the chance of heads, you should have $\Pr(Y_{99} = 1) = 1/2$ before seeing any toss, but

$$\Pr\left(Y_{99} = 1 \left| \sum_{i=1}^{98} Y_i = 98 \right.\right) = 0.99$$

after seeing the first 98 tosses come up heads [8].

A generalization important for statistical modeling is *partial* or *conditional* exchangeability [2, 3]. For example, suppose that the sequence $Y_1, \ldots, Y_N$ is partitioned into disjoint subsequences. Then the sequence is said to be partially exchangeable given the partition if each subsequence can be permuted without changing the joint distribution. If the Y_i represent survival times within a cohort of male stroke patients, then a judgment of unconditional exchangeability of the Y_i would be unreasonable if the patient ages were known, because age is a known predictor of survival time. Nonetheless, one might regard the survival times as partially exchangeable, given age, if no further prognostically relevant partitioning was possible based on the available data.

While exchangeability is weaker than iid, de Finetti [[3], Chapter 11] proved that finite subsequences of an infinite exchangeable sequence of Bernoulli (binary) variates must have representations as mixtures of iid Bernoulli sequences – a partial converse of the fact that any mixture of iid sequences is an exchangeable sequence. More precisely, suppose that $Y_1, Y_2, \ldots$ is an infinite sequence of exchangeable Bernoulli variates (that is, every finite subsequence of the sequence is exchangeable), and that θ is the limit of $(Y_1 + \cdots + Y_n)/n$ as n goes to infinity. De Finetti showed that there exists a distribution function $P(\theta)$ for θ such that, for all n,

$$\begin{aligned} \Pr(Y_1 = y_1, \ldots, Y_n = y_n) &\equiv \Pr(y_1, \ldots, y_n) \\ &= \int_0^1 \theta^s (1-\theta)^{n-s}\ \mathrm{d}P(\theta), \end{aligned} \tag{1}$$

where $s = y_1 + \cdots + y_n$. Many Bayesian statisticians find this theorem helpful, because it partially specifies the form of the predictive probability $\Pr(y_1, \ldots, y_n)$ when $Y_1, \ldots, Y_n$ can be considered a subsequence of an infinite exchangeable sequence.

In the representation shown in (1), $P(\theta)$ is recognizable as the *prior distribution* for θ, a distribution that may be developed from what is known about θ before the Y_i are observed. As noted in [7], however, the strength of the theorem's conclusion is easy to overstate: it does *not* imply that all binary data must be analyzed using the representation shown in (1); it merely says that if you judge $Y_1, Y_2, \ldots$ to be an exchangeable sequence, then there is a $P(\theta)$ that allows you to use (1) to specify $\Pr(y_1, \ldots, y_n)$.

Finite versions of the theorem [4] show that, if $Y_1, \ldots, Y_n$ is the start of an exchangeable Bernoulli sequence $Y_1, \ldots, Y_N$ and n/N is small enough, then $\Pr(y_1, \ldots, y_n)$ may be approximately expressed as in (1), with the approximation improving as n/N approaches zero. There are further generalizations to exchangeable sequences of **polytomous** variates, as well as exchangeable sequences of continuous variates [4]. The latter generalization requires a prior distribution on the space of continuous distributions, however, which can be much harder to specify than a prior for a vector of **multinomial** parameters, and which may lead to intractable computational problems [5].

References

[1] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, New York.

[2] De Finetti, B. (1937). Foresight: its logical laws, its subjective sources, reprinted in *Studies in Subjective Probability*, H.E. Kyburg & H.E. Smokler, eds. Wiley, New York, 1964.

[3] De Finetti, B. (1974). *Theory of Probability* (two vols). Wiley, New York.

[4] Diaconis, P. & Freedman, D. (1980). Finite exchangeable sequences, *Annals of Probability* **8**, 745–764.

[5] Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion), *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

[6] Draper, D., Hodges, J., Mallows, C. & Pregibon, D. (1993). Exchangeability and data analysis (with discussion), *Journal of the Royal Statistical Society, Series A* **196**, 9–37.

[7] Freedman, D.A. (1995). Some issues in the foundations of statistics (with discussion), *Foundations of Science* **1**, 19–83.

[8] Good, I.J. (1983). *Good Thinking*. University of Minnesota Press, Minneapolis.

[9] Lindley, D.V. & Novick, M.R. (1981). The role of exchangeability in inference, *Annals of Statistics* **9**, 45–58.

(*See also* **Foundations of Probability**; **Subjective Probability**)

SANDER GREENLAND & DAVID DRAPER

Exclusion Mapping

Exclusion mapping is the identification of the location of a disease **gene** by excluding other possible genomic regions; it relies on the fact that informative meioses can provide evidence for or against linkage (*see* **Linkage Analysis, Model-based**; **Linkage Analysis, Model-free**). If a disease locus exists, then it must be somewhere in the genome. Therefore, if the disease locus is not linked to any of the regions that have already been considered, then it must be present elsewhere in the genome. By using the existing linkage information to define regions of exclusion, efforts can be focused on the remaining regions of the genome that are more likely to contain the gene. Of course, the strength of the evidence for or against linkage in a particular region must be taken into account, since a locus may be present in a region even if it is not detected (e.g. if the sample size is too small). Therefore, in order to create an exclusion map, one must define a measure of the strength of the evidence for or against linkage and a criterion that represents at most a small **likelihood** that the gene is contained in the region. This is done using the lod score.

Consider first the case in which there is a single family or a collection of families but no locus heterogeneity (i.e. a single disease locus predisposes to the disease). All informative meioses provide evidence for or against linkage to a particular location. The *absence* of recombination between a **marker** and the phenotype provides evidence *for* linkage to a particular region, while the *presence* of recombination consistent with a rate expected by chance (1/2 of meioses) provides evidence *against* linkage in that region. The lod score provides us with a yardstick to measure the strength of the evidence for or against linkage in the region. In the context of testing a single marker, the classic limit for exclusion is a lod score of -2 [9], which represents a $100:1$ **likelihood ratio** against linkage in the region. This criterion is quite stringent, and generally results in a small region of exclusion around a marker locus. Larger regions of the genome can be excluded by testing multiple markers that are located close together. In this case, a lod score of -2 is still considered sufficient evidence against linkage, despite the fact that multiple tests are performed.

In the presence of locus heterogeneity (*see* **Genetic Heterogeneity**) and more than one sampled family, exclusion mapping results are more difficult to interpret. The study sample may include some families that are linked to the region and some families that are not linked to the region. The lod score will reflect the proportion of linked families. Therefore, even if we find strong evidence against linkage in a region, a small proportion of the families may be, in fact, linked to the region. Since this is the case for **complex diseases**, evidence for linkage to a region is often more convincing than evidence against linkage to a region, and exclusion mapping is not helpful. Methods that allow for locus heterogeneity within the study sample, such as estimating the proportion of linked families in addition to the recombination fraction [10], may provide a way to use exclusion mapping when locus heterogeneity is present.

Programs that provide a visual representation of the excluded and nonexcluded regions include: Exclude [3], Lodview [5], and Mapmaker/Sibs [7]. In the Exclude program, the probability of containing the gene is computed for each chromosomal region. An equal prior probability of containing the gene for each chromosomal region is assumed. The posterior probability computed by this program incorporates the fact that if a gene is excluded from one region, then the probability must increase in the remaining regions. The Lodview program produces a graphical view to observe regions with a lod score less than -2 from data generated by programs such as Linkage [8]. The method implemented in Mapmaker/Sibs identifies regions of exclusion depending on particular model parameters such as λ_s, the locus-specific sibling recurrence **risk ratio**. Regions are excluded for a gene that confers susceptibility with λ_s greater than some pre-specified value. However, it may be difficult for a user to identify which regions have been excluded for the phenotype of interest because of the difficulty in accurately estimating λ_s [4].

Genes for several diseases have been mapped using exclusion mapping, including Marfan syndrome and Rett syndrome. Marfan syndrome is a connective tissue disorder that is inherited in an autosomal dominant fashion (*see* **Segregation Analysis, Classical**). Blanton et al. [2] produced an exclusion map for Marfan syndrome based on linkage data for 75 marker loci from nine contributing laboratories. Eleven **candidate** regions were suggested

from this analysis, including chromosome 15. Kainulainen et al. [6] found significant evidence for linkage (lod score = 3.92) to a region on chromosome 15 using eight Finnish families with Marfan syndrome, which was later found to include the fibrillin gene. Rett syndrome is a progressive neurodevelopmental disorder that causes mental retardation. Because this disorder is found almost exclusively in females, it was suggested that Rett syndrome is caused by an **X-linked** dominant mutation that is lethal for hemizygous males. Rett syndrome families were used to exclude regions of the X chromosome and to map the locus to the Xq28 region. Amir et al. [1] identified mutations in the MECP2 gene in this region, and found this gene to be responsible for causing Rett syndrome.

References

[1] Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U. & Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2, *Nature Genetics* **23**, 185–188.

[2] Blanton, S.H., Sarfarazi, M., Eiberg, H., de Groote, J., Farndon, P.A., Kilpatrick, M.W., Child, A.H., Pope, F.M., Peltonen, L., Francomano, C.A., Boileau, C., Keston, M. & Tsipouras, P. (1990). An exclusion map of Marfan syndrome, *Journal of Medical Genetics* **27**, 73–77.

[3] Edwards, J.H. (1987). Exclusion mapping, *Journal of Medical Genetics* **24**, 539–543.

[4] Guo, S.-W. (1998). Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting, *American Journal of Human Genetics* **63**, 252–258.

[5] Hildebrandt, F., Pohlmann, A. & Omran, H. (1993). Lodview: a computer program for the graphical evaluation of lod score results in exclusion mapping of human disease genes, *Computers and Biomedical Research* **26**, 592–599.

[6] Kainulainen, K., Pulkkinen, L., Savolainen, A., Kaitila, I. & Peltonen, L. (1990). Location on chromosome 15 of the gene defect causing Marfan syndrome, *New England Journal of Medicine* **323**, 935–939.

[7] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.

[8] Lathrop, G.M., Laloue, J.M., Julier, C. & Ott, J. (1984). Strategies for multilocus linkage analysis in humans, *Proceedings of the National Academy of Sciences* **81**, 3443–3446.

[9] Morton, N.E. (1956). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–317.

[10] Ott, J. (1986). Linkage probability and its approximate confidence interval under possible heterogeneity, *Genetic Epidemiology* Supplement **1**, 251–257.

KATRINA A.B. GODDARD

Expectation

The *expectation* or *expected value* of a random variable Y, denoted $\mathrm{E}(Y)$, is its **mean**. When Y has density or probability mass function $f(y)$, $\mathrm{E}(Y) = \int f(y)\,\mathrm{d}y$ or $\sum yf(y)$, respectively. A rigorous treatment [1, 2] first defines $\mathrm{E}(Y)$ for simple **random variables** (discrete random variables taking only finitely many values), then extends the definition to arbitrary nonnegative random variables, and then to arbitrary random variables. For a simple random variable Y taking values $y_1, \ldots, y_k$ with respective probabilities $p_1, \ldots, p_k$, $\mathrm{E}(Y)$ is defined as $\sum_{i=1}^{k} y_i p_i$. Any nonnegative random variable Y may be written as a limit $\lim_{n\to\infty} Y_n$ of increasing simple random variables Y_n; $\mathrm{E}(Y)$ is defined as $\lim_{n\to\infty} \mathrm{E}(Y_n)$. Any random variable Y may be written as $Y_+ - Y_-$, where $Y_+ = YI(Y \geq 0)$ and $Y_- = -YI(Y < 0)$ are nonnegative; $\mathrm{E}(Y)$ is defined by $\mathrm{E}(Y_+) - \mathrm{E}(Y_-)$ provided at least one of these is finite. If exactly one is finite, the expectation is $+\infty$ or $-\infty$ depending on whether the infinite term is $\mathrm{E}(Y_+)$ or $\mathrm{E}(Y_-)$. If neither is finite, the expectation does not exist. For example, the mean does not exist for the **Cauchy** density $f(y) = \{\pi(1 + y^2)\}^{-1}$, $-\infty < y < \infty$.

The term "expectation" is a misnomer; a Bernoulli random variable (*see* **Binary Data**) Y with parameter 3/4 has expectation 3/4, though $Y = 3/4$ would be most unexpected since $Y = 0$ or 1. Still, $\mathrm{E}(Y)$ provides the best prediction of Y in that it minimizes the **mean squared error** $\mathrm{E}\{(Y - a)^2\}$ over all constants a, assuming $\mathrm{E}(Y^2) < \infty$. It is a measure of the center of a distribution. Because $(Y - a)^2$ is very sensitive to extreme Y values, $\mathrm{E}(Y)$ is pulled toward them for **skewed** distributions. For such distributions, the **median** is a better measure of central tendency than $\mathrm{E}(Y)$; it minimizes $\mathrm{E}(|Y - a|)$ over a, and $|Y - a|$ is less sensitive to extremes than is $(Y - a)^2$.

The expectation of a random variable Y may also be thought of as its long-run average, assuming $\mathrm{E}(|Y|) < \infty$. If one continually and independently repeats the experiment that generated Y, the average value $\overline{Y} = (1/n)\sum_{i=1}^{n} Y_i$ will be very close to $\mathrm{E}(Y)$ by the strong **law of large numbers**.

Useful Properties of Expectation

E1. If $\mathrm{E}(|Y|) < \infty$, then $|\mathrm{E}(Y)| \leq \mathrm{E}(|Y|)$.

E2. $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$ for real numbers a and b, provided the right side is not of the form $+\infty - \infty$ or $-\infty + \infty$.

E3. Jensen's inequality: If $\psi(y)$ is convex and $\mathrm{E}(Y)$ and $\mathrm{E}\{\psi(Y)\}$ are both finite, then $\mathrm{E}\{\psi(Y)\} \geq \psi\{\mathrm{E}(Y)\}$.

E4. Monotone convergence theorem: If $Y_n \uparrow Y$ and $\mathrm{E}(Y_n) > -\infty$ for some n, then $\mathrm{E}(Y_n) \uparrow \mathrm{E}(Y)$.

E5. Dominated convergence theorem: If $Y_n \to Y$ in probability and $|Y_n| \leq Z, \mathrm{E}(Z) < \infty$, then $\mathrm{E}(Y_n) \to \mathrm{E}(Y)$ (*see* **Convergence in Distribution and in Probability**).

E6. If $Y_1, Y_2, \ldots$ are nonnegative or $\sum_{i=1}^{\infty} \mathrm{E}(|Y_i|) < \infty$, then $\mathrm{E}\left(\sum_{i=1}^{\infty} Y_i\right) = \sum_{i=1}^{\infty} \mathrm{E}(Y_i)$.

E7. If Y is nonnegative, then $\mathrm{E}(Y) = \int_0^{\infty} P(Y > y)\,\mathrm{d}y$, whether finite or not.

E8. If $\mathrm{E}(Y^2) < \infty$, then $a = \mathrm{E}(Y)$ minimizes $\mathrm{E}(Y - a)^2$ over all $a \in \Re$.

Expectation can also be taken conditioned on events, random variables, or arbitrary sigma fields (*see* **Conditional Probability**). Assume that $\mathrm{E}(|Y|) < \infty$. If A is an event of nonzero probability, the *conditional expectation of Y given A*, denoted $\mathrm{E}(Y|A)$, is defined as $\mathrm{E}\{YI(A)\}/P(A)$, where $I(A)$ is the indicator of event A. For example, let (X, Y) have joint probability mass function $f(x, y)$, and let $A = \{X = x\}$. Then $\mathrm{E}\{YI(X = x)\} = \sum_y yf(x, y)$ and

$$\psi(x) = \mathrm{E}(Y|X = x) = \sum_y \frac{yf(x, y)}{g(x)}, \qquad (1)$$

where $g(x)$ is the **marginal probability** mass function of X. In other words, we can take the expectation of Y with respect to the conditional probability mass function $h(y|x) = f(x, y)/g(x)$. Similarly, if (X, Y) has a density $f(x, y)$, we could, by analogy, integrate over the conditional density $h(y|x) = f(x, y)/g(x)$:

$$\psi(x) = \mathrm{E}(Y|X = x) = \int \frac{yf(x, y)\,\mathrm{d}y}{g(x)} \qquad (2)$$

when $g(x) \neq 0$. But what if (X, Y) does not have a density or probability mass function? We need a more general definition of conditional expectation given a random variable. Consider the discrete setting and replace x with X in $\psi(x) : \psi(X) = \sum_y yf(X, y)/g(X)$. For any Borel set B of x points

with nonzero probability,

$$\begin{aligned}\mathrm{E}\{\psi(X)I(X \in B)\} &= \sum_{x\in B}\sum_{y}\left\{\frac{yf(x,y)}{g(x)}\right\}g(x)\\ &= \sum_{y}\sum_{x\in B} yf(x,y)\\ &= \mathrm{E}\{YI(X \in B)\}. \qquad (3)\end{aligned}$$

In other words, $\psi(X)$ has the same average value as Y over any X set of nonzero probability. The same thing happens in the continuous situation with sums replaced by integrals. Note that $\mathrm{E}\{\psi(X)I(X \in B)\} = \mathrm{E}\{YI(X \in B)\}$ also holds when $P(X \in B) = 0$ because both sides are 0. This is the generalization we are looking for. For any random vector $\underline{X} = (X_1, \ldots, X_n)$, *the conditional expectation of* Y *given* $\underline{X}$, *denoted* $\mathrm{E}(Y|\underline{X})$, is any (Borel) function $\psi(\underline{X})$ such that

$$\mathrm{E}\{\psi(\underline{X})I(\underline{X} \in B)\} = \mathrm{E}\{YI(\underline{X} \in B)\} \qquad (4)$$

for all n-dimensional Borel sets B.

Strictly speaking, we call $\psi(\underline{X})$ a version of $\mathrm{E}(Y|\underline{X})$ because more than one function satisfies (4). For example, if Y has a density, then $Z_1 = Y$ is a version of $\mathrm{E}(Y|Y)$, but so is $Z_2 = YI(Y \neq 0) + 10I(Y = 0)$. Though Z_1 and Z_2 are not identical, $P(Z_1 \neq Z_2) = P(Y = 0) = 0$. It is clear that by replacing 10 with any other number we can create infinitely many versions. That two versions Z_1 and Z_2 of $\mathrm{E}(Y|\underline{X})$ differ only on a set of probability, 0 is not unique to this example. It always holds.

When (X, Y) has a probability mass function or density $f(x, y)$, one version of $\mathrm{E}(Y|X)$ is (1) or (2). Indeed, we have already seen that (1) implies (4). Sometimes $\mathrm{E}(Y|X)$ is clear without resorting to conditional densities. For example, suppose we draw a **random sample** $X_1, \ldots, X_n$ from a population with expectation μ, and select one of $\{X_1, \ldots, X_n\}$ at random. The randomly selected value, Y, has the same unconditional distribution as an X_i. But conditioned on $X_1, \ldots, X_n$, $Y = X_i$ with probability $1/n$, $i = 1, \ldots n$. Thus, the conditional expectation of Y given $X_1, \ldots, X_n$ is $\sum_{i=1}^{n} X_i(1/n) = \overline{X}$.

Here is an example showing why $\mathrm{E}(Y|X = x)$ should not be thought of as conditioning on the event $A = \{X = x\}$ when A has probability 0. Suppose we want to compare the difference in sample proportions, $\hat{p}_1 - \hat{p}_0$, to the **relative risk**, $\hat{p}_1/\hat{p}_0$ from two independent random samples. Of specific interest is how likely various values of the difference in sample proportions are given the relative risk. Because $\hat{p}_0$ and $\hat{p}_1$ behave asymptotically like independent normal random variables (X, Y) with respective means p_0 and p_1 and respective variances $p_0(1 - p_0)/n$ and $p_1(1 - p_1)/n$, the **delta method** implies that the asymptotic joint distribution of $(\hat{p}_1 - \hat{p}_0, \hat{p}_1/\hat{p}_0)$ is that of $(Y - X, Y/X)$. It seems, therefore, that the conditional distribution of $(\hat{p}_1 - \hat{p}_0|\hat{p}_1/\hat{p}_0)$ must be that of $(Y - X|Y/X)$. To avoid resorting to the use of transformations and Jacobians, write the event $\hat{p}_1/\hat{p}_0 = \lambda$ as $\hat{p}_1 - \lambda\hat{p}_0 = 0$. It is tempting to say that the conditional distribution of $(Y - X|Y/X = \lambda)$ must be that of $(Y - X|Y - \lambda X = 0)$. The latter distribution is easy to compute: $(Y - X, Y - \lambda X)$ is **bivariate normal**, so the conditional distribution of $(Y - X|Y - \lambda X = 0)$ is univariate normal.

A surprising mistake in the above argument is concluding that because the events $\{Y/X = \lambda\}$ and $\{Y - \lambda X = 0\}$ are the same (barring $X = 0$), the conditional distribution of $(Y - X|Y/X = \lambda)$ is that of $(Y - X|Y - \lambda X = 0)$. In fact, Proschan and Presnell [5] show using Jacobians that the former conditional distribution is not normal. This example shows the danger of viewing conditional expectations given a random variable as conditional expectations given an event of probability 0. Incidentally, another flaw in the above argument is the assumption that because the asymptotic joint distribution of $(\hat{p}_1 - \hat{p}_0, \hat{p}_1/\hat{p}_0)$ is that of $(Y - X, Y/X)$, the conditional distribution of $(\hat{p}_1 - \hat{p}_0|\hat{p}_1/\hat{p}_0)$ must be that of $(Y - X|Y/X)$. In general, convergence of (U_n, V_n) to (U, V) in distribution does not imply convergence of the distribution of $(V_n|U_n)$ to that of $(V|U)$ (the converse is also not true; weak convergence of marginal and conditional distributions does not necessarily imply weak convergence of joint distributions; see [7]). Other interesting anomalies in conditional expectation are given in [4–6].

Conditional expectation may be viewed geometrically as a projection [3], Chapter 8. Equation (4) says that $\mathrm{E}\{(Y - Z)I(\underline{X} \in B)\} = 0$ for every n-dimensional Borel set B, where $Z = \mathrm{E}(Y|\underline{X})$. It follows that for any simple random variable W that is a function of $\underline{X}$, $\mathrm{E}\{(Y - Z)W\} = 0$. Because simple random variables are the building blocks used to generate any random variable, it can be shown that $\mathrm{E}\{(Y - Z)W\} = 0$ for any W that is a function of $\underline{X}$ such that $\mathrm{E}\{(Y - Z)W\}$ exists and is finite. To ensure $\mathrm{E}\{|(Y - Z)W|\} < \infty$, assume $\mathrm{E}(Y^2) < \infty$

and consider those W that are Borel functions of $\underline{X}$ with finite variance, $L_2(\underline{X}) = \{W : W = f(\underline{X}),$ f[is a Borel function, $\mathrm{E}(W^2) < \infty\}$. Then $\mathrm{E}\{(Y - Z)W\} = 0$ for all $W \in L_2(\underline{X})$.

When $\mathrm{E}(UV) = 0$ for random variables U and V, we can think of U and V as being **orthogonal** and write $U \perp V$. The analogy with n-dimensional vectors $\underline{u}$ and $\underline{v}$ is that $\underline{u}$ and $\underline{v}$ are orthogonal $\Leftrightarrow \sum u_i v_i = 0 \Leftrightarrow \mathrm{E}(UV) = 0$, where (U, V) is a random draw from a population with values $\{(u_1, v_1), \ldots, (u_n, v_n)\}$. The condition $\mathrm{E}\{(Y - Z)W\} = 0$ means that $Y - Z \perp W$. Picture the random variable Y as a vector in the plane. The set $L_2(\underline{X})$ is a linear subspace, which we can picture as a line. Then $Z = \mathrm{E}(Y|\underline{X})$ is the projection of Y onto $L_2(\underline{X})$ (Figure 1). This geometric perspective makes clear certain properties of conditional expectation. For example, if $L_2(\underline{X}_1) \subseteq L_2(\underline{X}_2)$, then $\mathrm{E}\{\mathrm{E}(Y|\underline{X}_2)|\underline{X}_1\} = \mathrm{E}(Y|\underline{X}_1)$ with probability 1. To see this, let $W \in L_2(\underline{X}_1)$ and set $Z_1 = \mathrm{E}(Y|\underline{X}_1)$ and $Z_2 = \mathrm{E}(Y|\underline{X}_2)$. Then $\mathrm{E}\{(Z_2 - Z_1)W\} = \mathrm{E}[\{-(Y - Z_2) + Y - Z_1\}W] = 0$ because $Y - Z_2 \perp L_2(\underline{X}_2) \supseteq L_2(\underline{X}_1)$ and $Y - Z_1 \perp L_2(\underline{X}_1)$. Because $\mathrm{E}\{(Z_2 - Z_1)W\} = 0$ for every $W \in L_2(\underline{X}_1)$, Z_1 is a version of $\mathrm{E}(Z_2|Z_1)$. Though this geometric argument implicitly assumes $\mathrm{E}(Y^2) < \infty$, the result holds whenever $\mathrm{E}(|Y|) < \infty$.

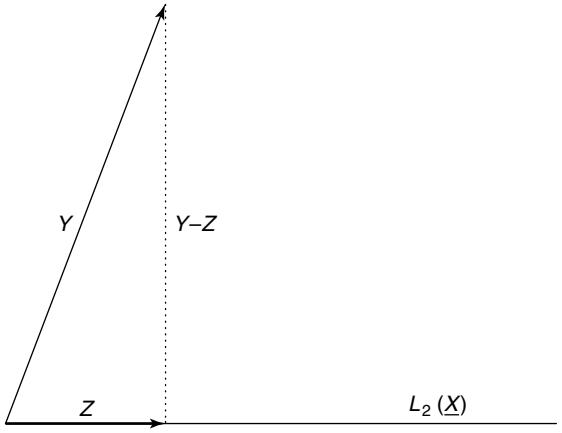

Figure 1 Conditional expectation as a projection. Think of the random variable Y as a two-dimensional vector, and the set $L_2(\underline{X}) = \{W : W = f(\underline{X}) : \mathrm{E}(W^2) < \infty\}$ of potential predictors of Y based on covariates $\underline{X}$ as vectors on a line. Then $Z = E(Y|X)$ is the projection of Y onto $L_2(\underline{X})$. That is, $Y - Z \perp W$ for every $W \in L_2(\underline{X})$

The geometric perspective is useful for understanding another interesting fact about conditional expectation. We have noted that when $\mathrm{E}(|Y|) < \infty$ and in the absence of any other information, $\mathrm{E}(Y)$ is the **prediction** of Y that minimizes the mean squared error $\mathrm{E}\{(Y - a)^2\}$ over all a. But now suppose one is allowed to predict Y after observing random variables $X_1, \ldots, X_n$. Among all predictors $W = f(X_1, \ldots, X_n)$ with finite variance, $Z = \mathrm{E}(Y|X_1, \ldots, X_n)$ minimizes the mean squared error $\mathrm{E}\{(Y - W)^2\}$. This is readily apparent from Figure 1 and the fact that $\mathrm{E}\{(Y - W)\}^2$ is the L_2 distance between Y and W. More formally, $\mathrm{E}\{(Y - W)^2\} = \mathrm{E}\{(Y - Z + Z - W)^2\}$. The cross product term $2\mathrm{E}\{(Y - Z)(Z - W)\}$ is 0 because $Y - Z \perp L_2(\underline{X})$ and $Z - W \in L_2(\underline{X})$, so $\mathrm{E}\{(Y - W)^2\} = \mathrm{E}\{(Y - Z)^2\} + \mathrm{E}\{(Z - W)^2\} \geq \mathrm{E}\{(Y - Z)^2\}$.

The definition of conditional expectation can be generalized. Notice that condition (4) involves X only through the collection of sets $\{\underline{X} \in B, B$ Borel$\}$ it generates, known as the sigma field generated by $\underline{X}$ and denoted $\sigma(X)$. For example, suppose X is the indicator that a patient has a history of hypertension, and Y is his systolic blood pressure on a given day. Further, suppose that $\mathrm{E}(Y|X = 0) = 120$ and $\mathrm{E}(Y|X = 1) = 150$. Even though $\mathrm{E}(Y|X = 0) \neq \mathrm{E}(Y|1 - X = 0)$ and $\mathrm{E}(Y|X = 1) \neq \mathrm{E}(Y|1 - X = 1)$, the random variables $\mathrm{E}(Y|X)$ and $\mathrm{E}(Y|1 - X)$ are the same, namely, 120 on the set where $X = 0$ and 150 on the set where $X = 1$. Similarly, $\mathrm{E}(Y|\underline{X}) = \mathrm{E}\{Y|t(\underline{X})\}$ for any invertible transformation $t(\underline{X})$. Because $\mathrm{E}(Y|X)$ depends only on the sigma field $\Im = \sigma(X)$, some authors prefer the notation $E(Y|\Im)$. It is helpful to think in terms of sigma fields rather than random vectors. Sigma field $\Im_2$ contains more information than sigma field $\Im_1$ if $\Im_1 \subseteq \Im_2$. For example, $\Im_1 = \sigma(|X|) \subseteq \sigma(X) = \Im_2$. Knowing which events in $\Im_2$ occurred tells us which events in $\Im_1$ occurred, but not vice-versa. Similarly, $\Im_1 = \sigma(X_1) \subseteq \sigma(X_1, X_2) = \Im_2$. Conditional expectation can be defined for arbitrary sigma fields (not just those generated by a random vector); $\mathrm{E}(Y|\Im)$ is any $\Im$-measurable random variable Z (meaning that $\{\omega : Z(\omega) \in B\} \in \Im$ for every Borel set B) such that $\mathrm{E}\{ZI(A)\} = \mathrm{E}\{YI(A)\}$ for all $A \in \Im$. Define $L_2(\Im)$ as the collection of $\Im$-measurable random variables W with $\mathrm{E}(W^2) < \infty$. The above results can be generalized as follows.

Useful Properties of Conditional Expectation

C1–C7: same as E1–E7 with expectation replaced by conditional expectation and the phrase "with probability 1" added.

C8. Among all functions $W \in L_2(\Im)$, $Z = \mathrm{E}\{Y|\Im\}$ minimizes the conditional and unconditional mean squared errors: $\mathrm{E}\{(Y-W)^2|\Im\}$ and $\mathrm{E}\{(Y-W)^2\}$, assuming $\mathrm{E}(Y^2) < \infty$.

C9. If W is $\Im$-measurable and $\mathrm{E}(|W|) < \infty$, $\mathrm{E}(|Y|) < \infty$, then $\mathrm{E}(WY|\Im) = W\mathrm{E}(Y|\Im)$ with probability 1.

C10. If $\mathrm{E}(|Y|) < \infty$ and $\Im_1 \subseteq \Im_2$, $\mathrm{E}\{\mathrm{E}(Y|\Im_2)|\Im_1\} = \mathrm{E}(Y|\Im_1)$ with probability 1.

C11. If $\mathrm{E}(|Y|) < \infty$, $\mathrm{E}\{\mathrm{E}(Y|\Im)\} = \mathrm{E}(Y)$.

References

[1] Billingsley, P. (1979). *Probability and Measure*. Wiley, New York.

[2] Chung, K. (1974). *A Course In Probability Theory*. Academic Press, New York.

[3] Karr, A. (1993). *Probability*. Springer-Verlag, New York.

[4] Perlman, M. & Wichura, M. (1975). A note on substitution in conditional distribution, *The Annals of Statistics* **31**, 1175–1179.

[5] Proschan, M. & Presnell, B. (1998). Expect the unexpected from conditional expectation, *The American Statistician* **52**, 248–252.

[6] Rao, M. (1988). Paradoxes in conditional probability, *Journal of Multivariate Analysis* **27**, 434–446.

[7] Sethuraman, J. (1961). Some limit theorems for joint distributions, *Sankhya A*, **23**, 379–386.

MICHAEL A. PROSCHAN

Expected Number of Deaths

Several procedures are available if the mortality in a study group is to be compared with the mortality of the general population from which the sample is drawn. In a clinical setting calculation of an *expected survival curve* is often performed, and in epidemiologic studies of geographic or occupational variations of mortality the observed number of deaths is usually compared with *the expected number of deaths* based on published mortality rates for the general population. Calculation of expected number of deaths is an integral part of the **standardization** of vital rates, which is one of the oldest statistical techniques; see Keiding [10] for a review of early applications of the method. Standardization of rates is closely related to **multiplicative** *hazard rate models* and the use of *relative mortality* to describe deviations from the expected mortality (see, for example, Breslow [5], Breslow & Day [6] or Hoem [9]).

Two methods have been developed for the calculation of expected number of deaths. The classical approach, *the person-years method*, is derived as a sum of products of age- and sex-specific rates and the corresponding time at risk, whereas *the prospective method* involves calculation of a sum of conditional survival probabilities. For a further description the following setup is introduced.

Consider a sample of n independent individuals. Individual i is followed from time u_i to time t_i if death does not occur prior to t_i. Let T_i denote the time of death and define $X_i = \min(t_i, T_i)$. The mortality rate at time t, assuming that the individual is subject to the same risks as a person from the external reference population having the same demographic description, is denoted $\lambda_i(t)$. The corresponding survival function is denoted $S_i(t)$. The reference population is usually taken to be the general population, and the mortality rate and the survival function may be determined from published data taking into account the sex, date of birth, and age at entry of the person.

Set D_i to 1 if death occurs during follow-up and to 0 otherwise, i.e. $D_i = I(u_i < T_i \leq t_i)$. The observed number of deaths is then $D = \sum D_i$. Introduce

$$A_i = \int_{u_i}^{X_i} \lambda_i(s)\, \mathrm{d}s,$$

and let $A = \sum A_i$ denote the *total exposure to death*. The probability of dying during follow-up is obtained as

$$p_i = \frac{S_i(u_i) - S(t_i)}{S_i(u_i)}.$$

It is easily seen that $\mathrm{E}(D_i | T_i > u_i) = p_i$. Moreover, one may show (see, for example Breslow [4] or Berry [2]) that also $\mathrm{E}(A_i | T_i > u_i) = p_i$.

These results suggest two different ways of calculating the expected number of deaths, $\mathrm{E}(D)$, on the assumptions that mortality in the study group is identical to that of the reference population.

The prospective method [8, 11] utilizes the relationship $\mathrm{E}(D) = \sum p_i$ directly; the expected number of deaths is simply obtained as $\sum p_i$. Note, however, that the potential follow-up time, t_i, must be known for *all* individuals in order to compute the expected number of deaths by this method. Such knowledge is not available for many **censoring** schemes, and this requirement therefore severely limits the applicability of the prospective method.

Deviations from the expected number of deaths may be assessed by computing

$$X^2_{\mathrm{P}} = \frac{\left(D - \sum p_i\right)^2}{V},$$

where $V = \sum \mathrm{var}(D_i) = \sum p_i(1 - p_i)$. The distribution of the test statistic is approximately a **chi-square distribution** on one **degree of freedom**.

The **person-years** method (see, for example, Case & Lea [7] or Berry [2]) relies on the relationship $\mathrm{E}(A) = \sum p_i$, which shows that A is an unbiased estimate of the expected number of deaths on the hypothesis that the mortality in the study group is identical to that of the reference population. The total exposure to death, A, is therefore used as *an estimate* of the expected number of deaths. Usually the distinction between "total exposure to death" and "expected number of deaths" is not done and the random variable A simply denotes the expected number of deaths. Note, however, that for extended follow-up of old individuals the contribution A_i to the total exposure to death may exceed 1. Consequently, one may encounter situations where the expected number of deaths is larger than the number of individuals in the sample (see Smith [12] for one such example) suggesting that this terminology is misleading.

Note, moreover, that A is not an unbiased estimate of the expected number of deaths if the mortality in the sample differs from that of the reference population. The size of the bias has been studied by Keiding & Væth [11] within the framework of the multiplicative hazard rate model.

With this method the hypothesis of no difference between observed and expected number of deaths may be tested by computing

$$X^2_{\mathrm{PY}} = \frac{(D-A)^2}{A}.$$

On the null hypothesis the distribution of the test statistic is approximately a χ^2 distribution on one degree of freedom. Calculations of **asymptotic relative efficiency** by Anderson & Anderson [1] indicate that the person-years method is more efficient than the prospective method for **proportional hazards** alternatives and that the efficiency gain increases as the proportion of individuals dying during follow-up goes up.

The person-year method is easily adapted to studies of cause-specific mortality (*see* **Competing Risks**). The total mortality rate in the reference population is simply replaced by the relevant cause-specific mortality rate and deaths from other causes are treated as censoring. This solution is not applicable for the prospective method as it would require knowledge of when an individual who dies of the cause in question would have died from one of the other causes. One may instead extend the above model by introducing cause-specific mortality probabilities for each individual taking all causes of mortality into account. Interpretation of deviations from the expected number of deaths is, however, complicated by the fact that excess death for one cause will necessarily imply a deficit for some of the other causes [12].

In application of the person-years method one usually assumes that the mortality rate $\lambda_i(t)$ is constant in one-year (or five-year) intervals, and the contribution A_i is then the sum of products of the age-specific rates and the time spent in the corresponding age category during follow-up. Interchanging the order of summation one obtains the standard formula for the expected number of deaths, $A = \sum \lambda_{\mathrm{as}} Y_{\mathrm{as}}$, where λ_{as} is the sex- and age-specific mortality rate of the reference population and Y_{as} is the total person-years at risk in the corresponding sex and age category. From this formulation it is seen that D/A is the *standardized mortality ratio* (see, for example, Breslow and Day [6, Chapter 2]), which is used extensively in the analysis of epidemiologic data on occupational hazards for comparing mortality in a study population with mortality in a standard population (*see* **Standardization Methods**).

The classical approach of indirect standardization and calculation of a standardized mortality ratio (SMR) follows from a simple, multiplicative model relating the mortality rate $\lambda_i(t)$ of an individual in the study population to the known mortality rate $\lambda_i^*(t)$ for the reference population

$$\lambda_i = \theta \lambda_i^*(t).$$

For this model one may show [3] that the standardized mortality ratio, D/A, is the **maximum likelihood** estimate of the relative mortality θ and that the test statistic, X^2_{PY}, given above is a score test (*see* **Likelihood**) for testing the hypothesis $\theta = 1$. The simple, multiplicative model has been developed further to deal with regression problems using **proportional hazards** regression models of the form

$$\lambda_i(t; \mathbf{z}_i) = \exp(\beta' \mathbf{z}_i)\lambda_i^*(t),$$

where $\mathbf{z}_i$ is a vector of independent variables and β the corresponding vector of regression parameters. With grouped data such models are often referred to as **Poisson regression** models (see, for example, Berry [2] or Breslow & Day [6, Chapter 4]).

References

[1] Anderson, J.R. & Anderson, K.M. (1984). Letter to the editor re Hartz et al. (1983), *Statistics in Medicine* **4**, 107–108.

[2] Berry, G. (1983). The analysis of mortality by the subject-year method, *Biometrics* **39**, 173–184.

[3] Breslow, N.E. (1975). Analysis of survival data under the proportional hazards model, *International Statistical Review* **43**, 45–58.

[4] Breslow, N.E. (1978). The proportional hazards model: applications in epidemiology, *Communications in Statistics-Theory and Methods* **7**, 315–332.

[5] Breslow, N.E. (1985). Cohort analysis in epidemiology, in *A Celebration of Statistics: The ISI Centenary Volume*, A.C. Atkinson & S.E. Fienberg, eds. Springer-Verlag, Heidelberg, pp. 109–143.

[6] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*. Vol. II. *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[7] Case, R.M. & Lea, A.J. (1955). Mustard gas poisoning, chronic bronchitis and lung cancer, *British Journal of Preventive and Social Medicine* **9**, 62–72.

[8] Hartz, A.J., Giefer, E.E. & Hoffmann, R.G. (1983). A comparison of two methods for calculating expected mortality, *Statistics in Medicine* **2**, 381–386.

[9] Hoem, J.M. (1987). Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review, *International Statistical Review* **55**, 119–152.

[10] Keiding, N. (1987). The method of expected number of deaths, 1786-1886-1986, *International Statistical Review* **55**, 1–20.

[11] Keiding, N. & Væth, M. (1986). Calculating expected mortality, *Statistics in Medicine* **5**, 327–334.

[12] Smith, P.G. (1984). Letter to the editor re Hartz et al. (1983), *Statistics in Medicine* **3**, 301.

(*See also* **Cohort Study**; **Occupational Epidemiology**; **Survival Analysis, Overview**)

MICHAEL VÆTH

Expected Survival Curve

Expected survival curves aim at calculating how an actual group of patients would have survived under "standard" or "historical" conditions, (cf. the general discussion in **Bias from Historical Controls**).

The standard or historical conditions are available as a survival curve $S_i(t)$ for each patient i individually, perhaps based on estimates in a regression model for **survival** data. The *direct adjusted survival curve* or *corrected group prognostic curve* $\overline{S}(t)$ was discussed by Makuch [14], Chang et al. [3], Gail and Byar [5], Markus et al. [15] and Thomsen et al. [18] as

$$\overline{S}(t) = \frac{1}{n}\sum S_i(t). \tag{1}$$

An important objection to the use of the direct adjusted survival function is that is does not take the realized **censoring pattern** into account – on the contrary, it depends strongly on an assumption of independent censoring and involves an averaging operation across the censoring pattern. Invoking each patient's *potential follow-up time* Bonsel et al. [2] proposed what in continuous time would amount to the following estimator. Let $0 < f_1 < \cdots < f_n$ be the potential follow-up times for the n patients, and define iteratively, for $f_j < t \le f_{j+1}$, *Bonsel's estimator*

$$S_B(t) = S_B(f_j)\frac{\sum_{j+1}^{n} S_i(t)}{\sum_{j+1}^{n} S_i(f_j)}. \tag{2}$$

A slightly different definition was proposed by Væth [20], (cf. [9, 12, 19)].

If the historical mortality is given as a **Cox regression model** so that patient i has **hazard** $\lambda_i(t) = \lambda_0(t)e^{\beta' z_i}$, it has been fairly common to calculate the average **covariate** $\overline{z} = \sum z_i/n$ and use $S(t, \overline{z})$ as expected survival curve. This *average-covariate approach* is clearly suboptimal, as explained in [3, 6, 17, 18].

As discussed in the article **expected number of deaths**, it is often unrealistic to assume the potential follow-up times known (possible occurrence of ordinary loss to follow-up gives unknown potential follow-up times), and it may be preferable to base the calculation of the *expected survival curve*, S^*, on exposing each study individual to his/her standard mortality rate over the actually experienced period at risk. One such proposal [18] generalized the classical calculation of expected number of deaths as follows. Assume that the historical mortality is given as a Cox regression model so that patient i has hazard $\lambda_i(t) = \lambda_0(t)e^{\beta' z_i}$. Let $Y_i(t) = I\{\text{patient } i \text{ still at risk at time } t\}$, $Y(t) = \sum Y_i(t)$. Then under the historical hypothesis on the mortality, the average hazard of the patients still under observation at time t would be

$$\lambda^*(t) = \frac{\sum Y_i(t)\lambda_0(t)\exp(\beta' z_i)}{Y(t)}. \tag{3}$$

Defining the cumulative hazard as $\Lambda^*(t) = \int_{u=0}^{t} \lambda^*(u)\,du$, the survival function $S^*(t) = \exp\{-\Lambda^*(t)\}$ is a continuous-time version of the "expected survival rate" [4] or "expected survival curve" [8]. Andersen and Væth [1] pointed out that it has the following desirable property: let the standard **Nelson–Aalen estimator** of $\Lambda(t)$ be defined by

$$\hat{\Lambda}(t) = \sum_{T_j \le t} \frac{1}{Y(T_j)}, \tag{4}$$

where $T_1 < T_2 < \cdots$ are the times of (observed) deaths. Then $\Lambda^*(t) - \hat{\Lambda}(t)$ has expectation zero, under the **null hypothesis** that patient i has hazard $\lambda_i(t)$. Therefore, $S^*(t)$ represents, under the null hypothesis, an *expected survival curve*. Outside of the null hypothesis, it is not so easy to interpret $S^*(t)$, containing as it does information on study group mortality, through $Y_i(t)$, as well as standard mortality, through $\lambda_i(t)$. $S^*(t)$ therefore cannot be recommended as an expected survival function, despite its wide use for this purpose.

$S^*(t)$ is, however, useful for inference on **excess mortality**, as follows. In the particular case where the study group has an *additive* excess mortality $\alpha(t)$ over the standard, that is, patient i has hazard $\alpha(t) + \lambda_i(t)$, one may obtain an estimate of the corresponding "survival" function

$$\exp\left[-\int_0^t \alpha(u)\,du\right] \tag{5}$$

as the so-called *relative* survival function $\hat{S}(t)/S^*(t)$. Andersen and Væth [1] showed that results on **unbiasedness**, **consistency**, and asymptotic **normality** are available. Thomsen et al. [18, 19]

compared this "expected survival curve" to the previously discussed estimators.

These matters were discussed earlier by Hakulinen [7] who considered three estimators of "the expected survival rate". He divided the study population into homogeneous subgroups and considered a weighted average of the group-specific survival functions (his formula (2.2)) and the survival functions corresponding the two different weighted averages of the standard mortality rates (his formulas (2.3) and (2.4)). When each subgroup consists of a single patient, his formula (2.2) equals the direct adjusted survival curve and his formula (2.3) equals S^*. Finally, the mortality rate in his formula (2.4) equals the mortality rate corresponding to Bonsel's estimator; see [17] for a related idea.

Nielsen's Asymptotic Results

Nielsen [16] considered the observed survival curve, $\hat{S}$, estimated by the **Kaplan–Meier** method, and estimated survival curves based on the Cox regression model. He showed the following asymptotic results under the null hypothesis of standard mortality, where $\| \cdot \|$ is $\sup_{\tau \leq t} | \cdot |$,

1. Under standard boundedness regularity conditions, and assuming conditional independence between survival time and censoring time given covariates, $\| S^* - \hat{S} \| \xrightarrow{P} 0$ and $\| S_B - \hat{S} \| \xrightarrow{P} 0$.
2. Assume, in addition that the censoring times are identically distributed and marginally independent of covariates (and hence survival times), then $\| \overline{S} - \hat{S} \| \xrightarrow{P} 0$.

Nielsen also proved asymptotic normality and gave martingale-based test statistics for the historical hypothesis (*see* **Counting Process Methods in Survival Analysis**).

Thus, the use of the direct adjusted survival curve requires marginally independent censoring, but no information about actual survival or potential follow-up is needed. In addition, it has the simple interpretation as the expected survival curve of the study population under standard mortality in the absence of censoring. The use of Bonsel's method requires conditional independence and information on all potential follow-up times. S^* requires neither restrictive assumptions on the censoring pattern nor any knowledge of the potential follow-up times, only information about the actual time at risk is needed. However, the latter depends on both the actual survival of the study group and the standard mortality rates and, therefore, has no interpretation outside the null hypothesis.

Individual Comparison of Prognosis for New versus Old Treatment

Keiding et al. [10] extended the above framework, to also include a fitted regression model for the study patients. Under the Cox model for the historical hypothesis, the survival probability of patient i would be (in obvious notation)

$$S_H^i(t) = \exp\left[-\Lambda_H(t)\exp(\beta_H' z_i)\right], \qquad (6)$$

while under the Cox model for current treatment, the survival probability would be

$$S_T^i(t) = \exp\left[-\Lambda_T(t)\exp(\beta_T' z_i)\right]. \qquad (7)$$

Since the prognostic indices $\beta_H' z_i$ and $\beta_T' z_i$ as well as the underlying intensities $\Lambda_H(t)$ and $\Lambda_T(t)$ will usually be rather different, the relative survival at time t

$$\gamma_i(t) = \frac{S_H^i(t)}{S_C^i(t)} \qquad (8)$$

will usually depend strongly on patient i and time t. Keiding et al. [10], Fig. 5) proposed using $\gamma_i(t)$ as one aid in deciding on the best treatment for each individual patient and developed a diagram to illustrate the rather dramatic variation of $\hat{\gamma}_i(t)$ when comparing **transplantation** to conservative treatment in a set of Nordic primary biliary cirrhosis patients; (*see* **Prognostic Factors for Survival**).

Difficulties in Applying Non- or Semiparametric Models for the Historical Controls

Two largely unnoticed difficulties in applying the Cox regression model and similar non- or **semiparametric** models as historical controls concern the *interpretation of the time variable* in the underlying intensity and the use of **time-dependent covariates**.

The first point is most easily explained through specific reference to the transplantation example hinted at above. The model based on the historical

data specifies that the death intensity at duration t after entrance into the clinical trials is $\lambda_0(t)\exp(\beta' z)$. This is to be compared with survival *since transplantation* for the study patients. Patients would usually be assumed to be at a more advanced stage of disease at transplantation than at entry into a trial, so that the conventional choice of $t = 0$ at transplantation (that is, underlying intensity $\lambda_0(t)$ at time t since transplantation) is ill motivated. Only if $\lambda_0(t)$ is constant over t does the conventional application seem justified, and one ought indeed to always postulate this and (if possible) refit this parametric model to the historical data before the comparison; see [13] for a general discussion of comparing survival after transplantation with conservative treatment.

The other difficulty is related to the use of time-dependent covariates, in particular, when a natural time origin is available prior to the entry on study of the study patients. **Delayed entry** (**left truncation**) methods are then appropriate, but as pointed out by Keiding and Knuiman [11], it is then usually impossible to incorporate time-dependent covariates into the Cox regression model.

References

[1] Andersen, P.K. & Væth, M. (1989). Simple parametric and nonparametric models for excess and relative mortality, *Biometrics* **45**, 523–535.

[2] Bonsel, G.J., Klompmaker, I.J., van't Veer, F., Habbema, J.D.F. & Slooff, M.J.H. (1990). Use of prognostic models for assessment of value of liver transplantation in primary biliary cirrhosis, *Lancet* **335**, 493–497.

[3] Chang, I.M., Gelman, R. & Pagano, M. (1982). Corrected group prognostic curves and summary statistics, *Journal of Chronic Diseases* **35**, 668–674.

[4] Ederer, F., Axtell, L.M. & Cutler, S.J. (1961). The relative survival rate: a statistical methodology, *National Cancer Institute Monograph* **6**, 101–121.

[5] Gail, M.H. & Byar, D.B. (1986). Variance calculations for direct adjusted survival curves with applications to testing for no treatment effect, *Biometrical Journal* **28**, 587–599.

[6] Ghali, W.A., Quan, H., Brant, R., van Melle, G., Norris, C.M., Faris, P.D., Galbraith, P.D. & Knudtson, M.L. (2001). Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models, *JAMA* **286**, 1494–1497.

[7] Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal, *Biometrics* **38**, 933–942.

[8] Hill, C., Laplanche, A. & Rezvani, A. (1985). Comparison of the mortality of a cohort with the mortality of a reference population in a prognostic study, *Statistics in Medicine* **4**, 295–302.

[9] Keiding, N. (1995). Historical controls and modern survival analysis, *Lifetime Data Analysis* **1**, 19–25.

[10] Keiding, S., Ericzon, B.-G., Eriksson, S., Flatmark, A., Höckerstedt, K., Isoniemi, H., Karlberg, I., Keiding, N., Olsson, R., Samela, K., Schrumpf, E. & Söderman, C. (1990). Survival after liver transplantation of patients with PBC in the Nordic countries: comparison to expected survival from another series of transplantations and from an international trial of medical treatment, *Scandinavian Journal of Gastroenterology* **25**, 11–18.

[11] Keiding, N. & Knuiman, M.W. (1990). Letter to the editor on 'Survival analysis in natural history studies of disease' by A. Cnaan and L. Ryan, *Statistics in Medicine* **9**, 1221–1222.

[12] Keiding, N. & Thomsen, B.L. (1999). Survival curves, Bonsel and Væth estimators of, *Encyclopedia of Statistical Sciences Update 3*. Wiley, New York, pp. 228–230.

[13] Klein, J.P. & Zhang, M.-J. (1996). Statistical challenges in comparing chemotherapy and bone marrow transplantation as a treatment for leukemia, in *Lifetime Data: Models in Reliability and Survival Analysis*, N.P. Jewell, A.C. Kimber, M.-L.T. Lee & G.A. Whitmore, eds. Kluwer, Dordrecht, pp. 175–185.

[14] Makuch, R.W. (1982). Adjusted survival curve estimation using covariates, *Journal of Chronic Diseases* **3**, 437–443.

[15] Markus, B.H., Dickson, E.R., Grambsch, P.M., Fleming, T.R., Mazzaferro, V., Klintmalm, B.G.B., Wiesner, R.H., van Thiel, D.H. & Starzl, T.E. (1989). Efficacy of liver transplantation in patients with primary biliary cirrhosis, *New England Journal of Medicine* **320**, 1709–1713.

[16] Nielsen, B. (1997). Expected survival in the Cox model, *Scandinavian Journal of Statistics* **24**, 275–287; Addendum **26**, 1999, 159.

[17] Nieto, F.J. & Coresh, J. (1996). Adjusting survival curves for confounders: A review and a new method, *American Journal of Epidemiology* **143**, 1059–1068.

[18] Thomsen, B.L., Keiding, N. & Altman, D.G. (1991). A note on the calculation of expected survival, *Statistics in Medicine* **10**, 733–738.

[19] Thomsen, B.L., Keiding, N. & Altman, D.G. (1992). Reply to a letter to the editor, cf. Thomsen et al. (1991), *Statistics in Medicine* **11**, 1528–1529.

[20] Væth, M. (1992). Letter to the editor re Thomsen et al. (1991), *Statistics in Medicine* **11**, 1527–1528.

(*See also* **Marginal Models**; **Marginal Models for Multivariate Survival Data**)

NIELS KEIDING

Experimental Design

R.A. Fisher's 1935 text, *The Design of Experiments* [4], unified applied statistics and still shapes the subject today. Fisher had to define and defend his notion of a scientific experiment, of statistical **inference**, and of probability, and then lay out a set of practical procedures for the novice. In the Preface to his text, he proclaims:

> A clear grasp of simple and standardized statistical procedures will ... go far to elucidate the principle of experimentation; but these procedures are themselves only the means to a more important end. Their part is to satisfy the requirements of sound and intelligible experimental design, and to supply the machinery for unambiguous interpretation.

Fisher used the word "experiment" with special force. The investigator was obliged to plan and to control the experiment, leaving only one aspect to chance, the random assignment of the treatments to the study subjects or objects (*see* **Randomized Treatment Assignment**). **Randomization** legitimized the statistical inference by mechanically imposing a probability distribution on the set of potential observations. Using a prespecified type I error, one tested the **null hypothesis** with a set of reproducible arithmetic procedures (*see* **Hypothesis Testing**).

When **William Cochran** and **Gertrude Cox** wrote their 1950 text, *Experimental Designs* [2], nearly all statisticians followed these protocols and most preferred to analyze laboratory-like experiments, such as the first real-data example in the Cochran & Cox text, "to measure the effectiveness of 4 soil fumigants in keeping down the number of eelworms in the soil". Typical experimental units were "a plot of land, a patient in a hospital, or a lump of dough, or it may be a group of pigs in a pen, or a batch of seed" (*see* **Unit of Analysis**).

The term *treatments* has a positive connotation. Treatments cure patients, increase crop yield, and improve the quality of industrial products. Other terms are deliberately bland, such as "blocks, plots, classes, units, and replicates". The experimental layout resembles a rectangle of land, partitioned into blocks or plots that contain units. A typical design allocates treatments in either a regular or a random manner to units within plots or within blocks. Units receiving the same treatment are "replicates".

In Fisher's time "design" connoted choice. One reviewed various shapes and sizes for blocks and plots and ways of making regular or random assignments of units. Then one chose a design to maximize the chance of detecting at a prespecified type I error level a factor producing a pattern of wide dispersion among a set of mean treatment effects. As the novelty of choice has worn off, "design" has come to mean a precise description of the statistical model. Searle et al. [6] have provided a brief, readable, and insightful history that covers this subject in the context of **variance components**.

Today, investigators often conduct experiments in which they cannot entirely control the experimental conditions. Accordingly, the theory and methods have extended to deal with these irregular conditions.

This article gives a narrow overview of the basic theory and methods of classical design of experiments and then briefly discusses some of the major recent extensions. After introducing some terminology, this article addresses: (i) the mathematical core of the subject, the decomposition of the sum of squared deviations, and the associated computational worktable; (ii) the link between the **analysis of variance** (ANOVA) and **linear regression**; (iii) the design of medical experiments; (iv) **fixed** and **random effects**; and (v) balance, **missing data** and **robustness**.

Some Terminology for the Design of Experiments

Fisher had provided a unified theory, a choice of many designs, and a worktable (the ANOVA table) that allowed a novice to carry out tests of significance. He moved study design beyond the adequacy of sample size (statistical **power**). Designs could include several interacting factors such as the factor "fertilizers" with four levels (F, G, H, K) and the factor "fumigants" with two levels (a, b) (*see* **Factorial Experiments**). The investigator could choose to have all possible combinations (Fa, Fb, Ga, Gb, Ha, Hb, Ka, Kb) or a balanced subset such as (Fa, Gb, Ha, Kb) (*see* **Fractional Factorial Designs**). Combinations introduced **"interactions"** between a type of fertilizer and a type of fumigant.

A **balanced incomplete blocks (bib) design** is suited to small block sizes, n units per block, and many treatments, $I > n$; more treatments than one could fit into a block. A typical bib design places n of

the I distinct treatments in each block and has overall balance among treatments. For example, with each column representing a block of size 4, and cyclically listing the seven treatments (ABCDEFG), a layout of a 7-block design is:

A	E	B	F	C	G	D
B	F	C	G	D	A	E
C	G	D	A	E	B	F
D	A	E	B	F	C	G

Crossed and nested designs represent extremes in balanced designs that have two (or more) distinct treatment factors. For example, let factor 1 with levels (ABCD) be completely crossed with factor 2 with levels (efgh). That is, each unit has a pair of treatments, one from factor 1 and one from factor 2. A layout for this crossed design is:

A		B		C		D	
e	f	e	f	e	f	e	f
g	h	g	h	g	h	g	h

A layout for a completely nested design is:

A		B		C		D	
e	e	f	f	g	g	h	h
e	e	f	f	g	g	h	h

A more precise notation such as "Ae Af Ag Ah" lists the explicit pairs with the uppercase treatment level first and the lowercase treatment level second, whereas the notation above shows how the second pair-member in the crossed design varies more than in a nested design.

When the number of levels of factor 1 (blocks) equals the number of levels of factor 2 (treatments ABC) and columns and rows are physically present, as in an agricultural field, then one can simultaneously balance treatments within both columns and within rows by means of a **Latin square design**. For $n = 3$, with columns representing blocks, the layout of a Latin square design is

A	B	C
B	C	A
C	A	B

Lattice designs have a block size n and a treatment factor with $I = n^2$ levels. A lattice design assigns a subset of distinct treatments to each block and each pair of levels appears equally often within the blocks. For example, with 12 blocks, block size = 3, and $I = 9$ treatments (abcdefghi), a layout for a lattice design with columns denoting blocks is:

a	d	g	a	b	c	a	b	c	a	b	c
b	e	h	d	e	f	e	f	d	f	d	e
c	f	i	g	h	i	i	g	h	h	i	g

"Classical" design of experiments contains many other topics (*see* **Magic Square Designs**; **Orthogonal Designs**; **Partially Balanced Incomplete Block Design**; **Youden Squares and Row–Column Designs**).

The Decomposition of the Sum of Squared Deviations

Squared Deviations

Under the **null hypothesis**, all the levels or categories of treatment have the same effect on a continuous outcome, Y. The vague **alternative hypothesis** merely posits that treatment effects differ. The method of analysis, the analysis of variance (ANOVA), does not reveal which treatment appears to work best. For ANOVA, one computes the grand mean of Y and the set of treatment means. If the values in this set substantially deviate from the grand mean, then the factor "treatments" is significant. The individual treatments become anonymous when their deviations from the grand mean enter into a single summary statistic, the sum of squared deviations from the grand mean.

The ANOVA table is a worktable for computing the test of a null hypothesis. The user provides "sums of squares" based on a "decomposition" of the total sum of squared deviations.

As an example, the $N = 8$ observations, Y_{ij}, shown in Table 1, displayed in two rows with $J = 4$ per row, have a grand mean of $\overline{Y}_{..} = Y_{..}/N = 32/8 = 4$. Table 1 also displays the deviations from the mean, $Y_{ij} - \overline{Y}_{..}$. The small sample size that makes it easy

Table 1 Notation for one-way ANOVA with observed data, Y_{ij}

Row	Y_{ij}				Total	Mean	$Y_{ij} - \overline{Y}_{..}$			
1	2	4	4	2	$Y_{1.} = 12$	$\overline{Y}_{1.} = 3$	−2	0	0	−2
2	7	4	6	3	$Y_{2.} = 20$	$\overline{Y}_{2.} = 5$	3	0	2	−1
					$\overline{Y_{..} = 32}$	$\overline{\overline{Y}_{..} = 4}$				

to verify the calculations would, in practice, call for a cautious interpretation of results.

The *total sum of squared deviations*, the *total sum of squares*, or *total deviations* is

$$\sum_{ij}(Y_{ij} - \overline{Y}_{..})^2 = 4 + 0 + 0 + 4 + 9 + 0 + 4 + 1 = 22.$$

The *within sum of squared deviations*, the *within sum of squares*, or *within deviations*, is

$$\sum_{ij}(Y_{ij} - \overline{Y}_{i.})^2 = 1 + 1 + 1 + 1 + 4 + 1 + 1 + 4 = 14.$$

The Decomposition of the Total Sum of Squared Deviations

For the one-way ANOVA model, the decomposition formula is

$$\sum_{ij}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{ij}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{ij}(\overline{Y}_{i.} - \overline{Y}_{..})^2,$$

$$\text{total deviations} = \text{within deviations} + \text{between deviations}.$$

With J observations per row, the between deviations or the sum of squared deviations between the row mean and the grand mean is

$$\sum_{ij}(\overline{Y}_{i.} - \overline{Y}_{..})^2 = J\sum_{i}(\overline{Y}_{i.} - \overline{Y}_{..})^2,$$

which in Table 1 is $4[(3-4)^2 + (5-4)^2] = 8$.

The general formula allows row i to have n_i columns. When all $n_i = J$, the ANOVA design has "balanced" replications; otherwise replications are "unbalanced".

To prove the decomposition formula, one shows that a cross-term equals zero, thereby linking the formula to the geometry of **least squares**. Specifically, rewrite the deviation $Y_{ij} - \overline{Y}_{...}$ as

$$(Y_{ij} - \overline{Y}_{i.}) + (\overline{Y}_{i.} - \overline{Y}_{..}).$$

Now, the squared deviation has two squared terms and a cross-term. Summing over all i and j yields the decomposition formula when the cross-term drops out; that is,

$$\sum_{ij}(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..}) = \sum_{i}(\overline{Y}_{i.} - \overline{Y}_{..})\sum_{j}(Y_{ij} - \overline{Y}_{i.}),$$

$$= \sum_{i}(\overline{Y}_{i.} - \overline{Y}_{..})0 = 0.$$

The vector of within deviations is **orthogonal** to the vector of between deviations because their inner product is zero. Hence, the decomposition formula illustrates the Theorem of Pythagoras by forming a right-angled triangle from the three vectors of deviations with the vector of total deviations as the hypotenuse.

Other terms are used in place of "between" and "within". If the model for a one-way ANOVA holds, then the data in each row cluster about the row mean. Such a model "explains" between deviations, while within deviations remain as unexplained errors. Hence, instead of the terms "between" and "within", some authors use "model" and "error" or "explained" and "unexplained".

The ANOVA Table

The null hypothesis that the row means are equal is tested by comparing the between-mean square to the within-mean square. A balanced one-way ANOVA (I rows and J columns with $N = IJ$ observations) has the ANOVA table shown in Table 2.

Within a row of the ANOVA table, the mean square equals the sum of squares divided by the **degrees of freedom**. For the data from Table 1, the ANOVA table values are shown in Table 3.

The sum of squares column contains the values for the terms in the decomposition. The degrees of freedom column entries total $N - 1$, one less than the sample size.

Table 2 Balanced one-way ANOVA

Source	Sum of squares	df	Mean square	F-ratio
Between (model)	Between deviations	$I - 1$	Between-mean square	$\frac{\text{Between-mean square}}{\text{Within-mean square}}$
Within (error)	Within deviations	$I(J-1)$	Within-mean square	
Total	Total deviations	$N - 1$		

Table 3 ANOVA table values from data in Table 1

Source	Sum of squares	df	Mean square	F-ratio
Between	8	1	$8/1 = 8.0$	$8.0/2.3 = 3.4$
Within	14	6	$14/6 = 2.3$	
Total	22	7		

The underlying normality assumptions (*see* **Normal Distribution**) imply that the between deviations and the within deviations have independent **chi-squared distributions** with the respective degrees of freedom given in the ANOVA table. The observed F ratio follows an ***F* distribution**.

For Table 1 the row means would differ significantly if the observed F ratio of 3.4 exceeded v, the value of the F distribution with one and six degrees of freedom for which $\Pr(F \geq v) = 5\%$. But the F ratio of 3.4 is less than $v = 5.99$ and the null hypothesis cannot be rejected.

The phrase "degrees of freedom" has a useful interpretation. Typically it arises when discussing the sample **variance**,

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{(n-1)}.$$

The sum has n terms, yet is divided by $n - 1$. This suggests that a degree of freedom is lost because the parameter, the population mean, μ, is estimated by the sample mean, $\overline{x}$. The first column of the ANOVA table identifies a source or a factor associated with each row. The number of degrees of freedom is roughly the number of terms in the sum of squares minus the number of parameters associated with the row source.

The Link Between the Analysis of Variance and Linear Regression

The framework of linear regression helps to specify the model and the null hypothesis (or hypotheses). But the ANOVA model has many simple representations as a regression model. For example, let each row in Table 1 represent a treatment applied to four subjects. The one-way ANOVA model corresponds to a linear regression model with intercept α, namely

$$y = \alpha + \beta t_1 + \text{error},$$

where the zero–one variable $t_1 = 1$ for treatment 1 and $t_1 = 0$ for treatment 2. One might choose this model if treatment 1 were the experimental treatment and if treatment 2 were the control group.

The one-way ANOVA model also corresponds to a no-intercept model

$$y = \beta_1 t_1 + \beta_2 t_2 + \text{error},$$

where t_1 and t_2 are the respective zero–one variables for treatments 1 and 2. Note that in this model if $t_1 = 0$ then $t_2 = 1$ and vice versa. One might choose this model if both treatments were active treatments. One can show that $\alpha = \beta_1 + \beta_2$ and that $\beta = \beta_1$, indicating the sense in which the models are equivalent. The test that the row effects are equal in the ANOVA model, the test that $\beta = 0$ in the intercept model, and the test that $\beta_1 = \beta_2$ in the no-intercept model are all the same test. In each of these three tests under the given constraint the null hypothesis expected value of y does not vary with treatment.

In fact, the ANOVA model is equivalent to an infinite class of equivalent regression models, a source of confusion until one decides to choose one convenient model and ignore all the others.

When adopting the intercept model, one treatment category is viewed as "baseline" or "control" while all other categories of treatment are viewed as "active" compared with the common baseline treatment. A one-way ANOVA model with three treatments, t_1, t_2, and t_3, has the corresponding no-intercept regression model

$$y = \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \text{error},$$

and lumps all treatments under one global null hypothesis that $\beta_1 = \beta_2 = \beta_3$. In this model if $t_1 = 0$ then $t_2 = t_3 = 0$; if $t_2 = 0$ then $t_1 = t_3 = 0$, and if $t_3 = 0$ then $t_1 = t_2 = 0$. The regression model permits one to break up the global ANOVA test into separate tests for each β coefficient.

ANOVA Equations

ANOVA models simultaneously test all levels of a factor within a row of the ANOVA table. The general procedure for testing each of several factors is as follows: decompose the total deviations into one line of (between) deviations for each factor followed by a final line of within deviations. For each factor, the F ratio is the mean square for a factor (factor deviations/factor degrees of freedom) divided by the within-mean square.

Regression equations list all parameters, while ANOVA equations list only the factors. For example, for the one-way model with I levels of treatment and J replications per treatment, one form of the ANOVA model equation is

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where Y_{ij} is replication j of treatment i, μ_i is the mean for treatment i, and ε_{ij} are mutually independent normally distributed random errors each with mean zero and variance σ^2. Another form is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where α_i is the difference between the mean for treatment i and the grand mean μ. Note that μ is not the intercept. The ANOVA table yields only tests of the factors. Thus, a variety of post hoc tests, known as "**multiple comparison** tests", typically compare the many pairs of mean values after carrying out the ANOVA F tests. These *post hoc* tests more closely resemble the many tests in a regression model, one for each β coefficient.

The General Linear Model

The mathematical theory behind the design of experiments provides precision and places ANOVA within the general theory of linear regression, "the **general linear model**". Few statisticians can resist this efficient but abstract view of the design of experiments. Hence, most modern texts and articles use the formalisms and notation of the general linear model. It succinctly expresses the decomposition formula and relates it to the geometry of least squares. The vector of estimates of the regression β coefficients are obtained by applying a "projection" matrix that maps from the vector space of observations (total deviations) into the subspace of the model (between deviations). The orthogonal complement to the model subspace is the error subspace (within deviations), thereby completing the right-angled triangle. The notation streamlines the expression of the null hypotheses and the development of the distribution of test statistics associated with these hypotheses. Under normality assumptions, summary statistics based on orthogonal vectors are statistically independent.

The design matrix, $\mathbf{X}$, appears in the regression model equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \text{error},$$

where the observation vector $\mathbf{Y}$ is an $N \times 1$ vector, $\mathbf{X}$ is an $N \times k$ matrix, and the vector of regression parameters $\boldsymbol{\beta}$ is a $k \times 1$ vector. The design is specified by the design matrix $\mathbf{X}$ and by the joint distribution of the error terms. The investigator then gathers the data, $\mathbf{Y}$, estimates the unknown vector of parameters, $\boldsymbol{\beta}$, and tests the null hypothesis under the specified distribution.

For the one-way ANOVA model with $I = 2$ rows and $J = 4$ columns as in Table 1, one can write the transpose of the vector $\mathbf{Y}$ as

$$\mathbf{Y}^T = (Y_{11}, Y_{12}, Y_{13}, Y_{14}, Y_{21}, Y_{22}, Y_{23}, Y_{24}).$$

Now, consider again the intercept and no-intercept regression models discussed at the beginning of this section, associating the subscript "1" with "intercept" and "0" with "no-intercept". Then for the intercept-regression model with $\boldsymbol{\beta}^T = (\alpha, \beta)$, and for the no-intercept regression model with $\boldsymbol{\beta}^T = (\beta_1, \beta_2)$ the respective design matrices, $\mathbf{X}_1$ and $\mathbf{X}_0$, are

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Note that the second column of $\mathbf{X}_1$ is the same as the first column of $\mathbf{X}_0$, whence $\beta = \beta_1$. If one adds the first column of $\mathbf{X}_0$ to the second column of $\mathbf{X}_0$, one obtains a column of 1s which is the same as the first column of $\mathbf{X}_1$, whence $\alpha = \beta_1 + \beta_2$. Then $\mathbf{X}_1$ and the transformed $\mathbf{X}_0$ have the same pair of column vectors, but in reverse order. It follows that the columns of each matrix span the same two-dimensional space.

In general, given a regression model of the form, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \text{error}$, with an $N \times k$ design matrix $\mathbf{X}$ of rank $k \leq N$, let the columns of $\mathbf{X}$ form a basis for k-dimensional subspace E_k of the N-dimensional vector space E. Then the set of all possible design matrices for the ANOVA model corresponding to $\mathbf{X}$ is the set of all possible bases for E_k. This characterizes all regression models that yield the same ANOVA model.

In practice, one specifies a convenient basis for E_k; that is, $\mathbf{X}$ is fixed by choosing a convenient form of the regression model such as the model with an

intercept and a baseline category for each categorical variable.

The Design of Medical Experiments

Currently ANOVA is applied to many contexts that little resemble a laboratory setting. For example, to compare medical treatments one might enroll patients within several hospitals. Ideally, to achieve the same level of control as in an agricultural study, all patients might be brought to a large rectangular room with rows and columns of beds organized in blocks or plots. Setting aside one block of beds for each hospital, treatment might be allocated in regular patterns within each block (*see* **Blocking**).

Obviously, few medical studies can arrange for such convenient designs, but balance and varying adjacent treatments remain important methods of controlling error. For example, one does not want to give treatment A to the first 20 patients that enroll in a study and treatment B to the next 20 patients that enroll. This borrows from agricultural designs; if the left end of a field is the more fertile, then

A B A B A B
B A B A B A

is preferable to

A A A B B B
A A A B B B

Superimposing a structural factor, such as the 2×2 blocks in the first pattern, controls error. Thus, the design for a study that enrolls 40 patients might divide the sequence into 10 blocks of size 4 and within each block randomly assign two patients to treatment A and two patients to treatment B. Unlike the regular pattern in the agricultural design, the medical study randomly assigns treatments within blocks to prevent an investigator from anticipating what treatment the next patient might receive.

Factorial Designs

Factorial designs have one or more factors. For example, patients with systemic lupus erythematosus (SLE) at four hospitals received either at-home counselling or in-hospital counselling. They were stratified by the severity of disease; those without and with system damage (usually cardiac or kidney failure). The factors are treatments, hospitals, and severity of disease. A chronic disease, SLE "flares" take the form of fatigue, rash, and other such symptoms. Experts believe that deteriorating patients have more flares. Counselling promotes taking medications regularly and thereby reduces the number of flares.

Factors often divide into treatment factors that have scientific import and structural factors, like blocks, that pattern the allocation of treatments. The factor, "hospitals", may be structural, may have scientific import (if testing the hypothesis that teaching hospitals do no better treating SLE than other hospitals), or fall into the gray area between "treatment" and "structure", into a third category, "adjustment factors", that include "**confounders**", "**nuisance parameters**", and the type of factor that converts an ANOVA into an **analysis of covariance (ANCOVA)**. The underlying linear regression model absorbs adjustment factors in two ways, as extra terms each with a β coefficient or as a set of **interaction** terms that replace a single term as in the extension form an ANOVA to an ANCOVA model.

The lack of control over hospital protocols, physician practices, counselling practices, patient behavior, and the accuracy of medical charts, combined with uncertainty about how the disease progresses, raise many design issues. Factorial designs capture some of these extra sources of variation by introducing interaction terms. The ANOVA model can include two-way interactions (hospitals and treatments, treatments and disease severity, and hospitals and disease severity) and possibly three-way interactions.

This approach may "overfit" the model. A few ANOVA factors may add so many insignificant factors and interactions that the error (within deviations) shrinks, the F ratios inflate, and null hypotheses are falsely rejected (type I error). Suppose the study includes four hospitals, three levels of severity of disease, and two treatments. Then, while the ANOVA equation has seven terms (three main factors, three two-way and one three-way interaction), the corresponding regression equation has 24 β coefficients including one intercept, six main effect, 11 two-way effect, and six three-way effect parameters.

This calculation also gives the formula for the degrees of freedom in the corresponding ANOVA table. A one-factor design with R rows has R parameters; either a grand mean and $R-1$ of the rows of the factor as in the intercept regression model, or R rows as in the no-intercept model. A two-factor

model with R rows and C columns has RC interaction terms. The mean of column 1 is the grand mean of the R interaction terms with column 1. Thus, after fixing all the row and column means, there remain $(R-1)(C-1)$ two-way interaction terms to determine. Note that $(R-1)(C-1)+(R-1)+(C-1)+1=RC$. Analogously, a three-factor model with R rows, C columns, and S slices partitions the RCS parameters into $(R-1)(C-1)(S-1)$ three-way interactions, $(R-1)(C-1)+(R-1)(S-1)+(C-1)(S-1)$ two-way interactions, $R-1+C-1+S-1$ main effect parameters, and one grand mean.

One must decide which interactions to include, how to model the adjustment factors, in what order to test the hypotheses, and whether to simplify the model after failing to reject a null hypothesis. For instance, should the model drop a "useless" factor? If such a factor and all its interactions are insignificant, should one pool its associated sums of squares with the error sum of squares in subsequent hypothesis tests?

Classical design of experiments [2] contains testing protocols that control for type I error and that specify how to update the model after an insignificant test result. But precisely when to apply such protocols depends on the underlying biology. The biological model dictates which factors belong in the model.

Medical studies often uncover new factors. In the SLE example, the frequency of flares may depend on physicians' practice styles, a factor not considered so far in the model. Previous data or expert opinion may support inclusion of the factor "physicians" in the SLE ANOVA model. One or two physicians may stand out as specialists. Should the model identify them as separate factors, or merely include them, anonymously, among all the physicians?

Cochran & Cox [2] addressed the same issue by extending the eelworm-data model to account for other soil factors. In Chapter 3 of their text, they use the analysis of covariance to add to the ANOVA model an extra **covariate**; namely, the level of eelworm infestation before the experiment began. Taking samples before fumigants were applied determined the initial level of infestation. Without such data the analyst would have had to assume equal initial levels of infestation.

Who recognized the need to gather such data? R.A. Fisher had designed the eelworm experiment conducted in 1935 by the Rothamsted Experimental Station. A major figure in genetics, Fisher knew the pertinent biology and agronomy.

Fixed and Random Effects

Random Effects

The agronomist of Fisher's era fixed the levels of fertilizer and then controlled the allocation of these treatments. The term "**fixed effects**" refers to fixing the levels of a factor. Strictly speaking, "**random effects**" ought to imply making a random selection of levels from an extensive list or "population" of possible levels. Why would anyone make such an odd selection? Typically, in **randomized complete blocks designs**, one randomly assigns treatments to blocks of subjects (the levels of a structural factor) because one has no other choice. For example, in the SLE example, only one form of counselling treatment might be feasible at each hospital because patients counselled at home would "contaminate" the treatments by speaking with those given in-hospital counselling. Many surveys save money by using **cluster sampling**, and hence introduce random effects into the analysis. An interviewer goes to a home and interviews all three adults in this household, rather than traveling to three separate homes and interviewing only one adult in each home, thereby introducing the random factor "household". Individuals who are interviewed several times (e.g. repeated measures) or the aliquots of an individual laboratory specimen introduce the random factor "individual".

In the SLE example, the randomized block design assigns the same treatment to the entire block of patients at each hospital. In contrast, a "completely randomized design" would randomize each subject separately. From what population are the blocks randomly selected? Were the hospitals drawn from a long list of all US hospitals, or from a "convenient" shortlist of hospitals inclined to participate?

Many convenience samples of blocks are treated as "random effects" because this lack of control more closely resembles "random effects" than "fixed effects" and because the former provides more conservative inference than the latter. For example, some forms of **meta-analysis** [5] based on published articles use a random effects model, without any data on the rest of the population; that is, unpublished articles. However, in the SLE example, if one regards the

factor "hospital" as a fixed effect, then critics might argue that a significant treatment difference would only occur at the study hospitals. Hence, regarding the factor "hospital" as a random effect adds credence to the assertion that treatment effects generalize to other hospitals.

Computation Using the ANOVA Table

The ANOVA table and all the foregoing theory assume fixed effects or fixed factor level models. Fisher also used the ANOVA table as a worktable for the random effects model (all factors random) and the mixed model (some factors fixed and some factors random). But each row in the ANOVA table for a random factor now contributes to the error sum of squares and the F ratio becomes much harder to compute. This means that the one-way fixed effects model has only one source of error (within) deviations, while the one-way random effects model has two sources of error, a source within treatments and a source between treatments. Fisher adopted a reasonable, but imperfect, strategy to separate the model deviations from the error deviations. Today, high-speed computing makes feasible other more rigorous solutions [6]. In the fixed effects model, both the between-mean square and the within-mean square have the same expected value, σ^2, the error variance. In the random effects model, these expected mean squares are expressions linear in σ^2 and in the variance of the randomly selected levels of the between factor, σ_{A}^2. Fisher set these two expressions equal to the observed values in the mean square column of the ANOVA table, solved the two simultaneous equations in two unknowns, σ_{A}^2 and σ^2, and set the F ratio equal to the ratio of these solutions.

For the balanced one-way random effects model with $N = IJ$ observations, J replicates of each level $i = 1, 2, \ldots I$, the expected values of the between-mean square, B, and of the within-mean square W, are, respectively, $J\sigma_{\mathrm{A}}^2 + \sigma^2$ and σ^2. It follows that the estimate of σ_{A}^2 is $(B - W)/J$, a negative quantity whenever B is less than W. A negative value undermines the hypothesis test. Mixed or random models with two or more factors generate sets of three or more simultaneous linear equations with solutions that can be negative and that seldom have simple closed forms.

The Abstract Viewpoint

The one-way random effects model has the ANOVA equation

$$Y_{ij} = m_i + \varepsilon_{ij},$$

where the m_i are mutually independent normally distributed random variables with mean μ and variance σ_{A}^2, the ε_{ij} are mutually independent normally distributed random errors each with mean zero and variance σ^2, and each of the m_i is independent of each and every ε_{ij}. Note that the fixed effects model is an extreme case of the random effects model because m_i is fixed when $\sigma_{\mathrm{A}}^2 = 0$. Also, note that when $j = j'$ the variance, $\mathrm{var}(Y_{ij})$, equals the covariance, $\mathrm{cov}(Y_{ij}, Y_{ij},) = \sigma_{\mathrm{A}}^2 + \sigma^2$, but when $j \neq j'$, $\mathrm{cov}(Y_{ij}, Y_{ij},) = \sigma_{\mathrm{A}}^2$. It follows that, when σ_{A}^2 is positive and when $j \neq j'$, the random variables Y_{ij} and $Y_{ij'}$, are dependent and have a positive **correlation** $\sigma_{\mathrm{A}}^2/(\sigma_{\mathrm{A}}^2 + \sigma^2)$ that Fisher named the "intraclass correlation coefficient". In contrast, in the fixed effects model, all the random variables, Y_{ij}, are independent.

The general linear model transforms the dependent normal random variables, Y_{ij}, into independent normal random variables, Z_{ij}. Then, within the framework of the fixed effects models, one obtains from the Z_{ij}, the sum of squares, degrees of freedom, and distribution theory needed for the Y ANOVA table.

This provides formulas and distributions for the simpler mixed models and a theoretical foothold for justifying approximations when the estimation scheme breaks down (e.g. a negative estimate of a variance). However, the abstract approach merely asserts that a factor is "random" without asking what mechanism has generated the associated normal errors.

Balance, Missing Values, and Robustness

Balance and Missing Values

Having "balance" greatly simplifies the theory, the formulas, and the ANOVA table. Generally, when the number of replicates per treatment level varies (imbalance), the F ratio may no longer follow an F distribution and closed formulas and computation become much more complicated. Hence, having a few missing values from a balanced design often calls for restoring the balance by imputing the missing values.

Having many missing values or major imbalance remains a thorny issue. While imputation may restore balance and vastly simplify theory and computation, the investigator still has to justify the assumptions behind the imputation. However, the credibility of computer-intensive solutions (on the basis of **maximum likelihood** or **Bayesian methods**) often depend on the correctness of the model and of the error distributions [6].

Robustness

Until recent advances removed some barriers, limits on computation allowed only simple patterns of correlation among normally distributed errors, such as equal correlation among all observations in a block of units. The body of theoretical results associated with **generalized estimating equation (GEE)** models [3] and **restricted maximum likelihood (REML)** methods [6] addressed the problem of an excess number of parameters destabilizing the parameter estimates. For example, in a one-way ANOVA design with $N = IJ$ observations, J replications for each of I subjects, the maximum number of $I(I-1)/2$ correlations among the I subjects exceeds N whenever $J < (I-1)/2$. Advances, such as the GEE models, have produced **robust** estimates of treatment effects; robust against a wide variety of possible correlation patterns among the error terms. As embodied in the REML solution, one integrates the full **likelihood** function over the possible correlation patterns to obtain estimates from a **marginal likelihood** function containing only the treatment parameters.

Normality assumptions are another source of concern. While recent advances allow other distributional structures for error, notably the **binomial** and **Poisson** families for discrete data, the **multivariate normal distribution** dovetails with the theory of least squares. **Simulation** methods such as the **bootstrap** and well-chosen approximate **nonparametric methods** offer effective validation of results based on normal assumptions.

Errors in variables refers to errors in the measurement of the values in the design matrix, $\mathbf{X}$, that represents the factor levels, as opposed to the error terms associated with the outcome measure, $\mathbf{Y}$, as in the regression model matrix equation, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} +$ error. For example, a medical study might require a standard form on each patient at the time of diagnosis, but rely on a medical chart review for patients diagnosed before the study began. The quality of medical chart data varies enormously and is seldom as accurate as a standard form. To distinguish these sources let $\mathbf{X}$ denote the data from the standard form and let $\mathbf{W}$ denote the same "surrogate" data abstracted from medical charts.

Solutions for the **errors in variables** problem [1] call for assumptions about the models relating the random variables $\mathbf{W}$, $\mathbf{X}$, and error distribution. These models extend to include other sets of covariates besides $\mathbf{X}$. Carroll et al. [1] have provided robust solutions by extending their results to **nonlinear regression** models and introducing functional modeling that makes minimal assumptions about the distribution of $\mathbf{X}$.

Previous discussion has touched on the lack of randomness when using convenience samples, as in the SLE example where the study hospitals were those inclined to join and in meta-analyses of published articles. One can emulate a randomized study by the methods of **quasi-experimental design** or, more broadly, the methods of choosing controls for a **case–control study**. All such methods try to balance known measurable confounding factors, or at least produce a conservative **bias** in known but hard-to-measure factors. But only randomization can balance unknown factors.

References

[1] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

[2] Cochran, W.G. & Cox, G. (1950). *Experimental Designs*. Wiley, New York.

[3] Diggle, P.J., Liang, K. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

[4] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[5] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando.

[6] Searle, S.R., Casella, G. & McCullogh, C.E. (1992). *Variance Components*. Wiley, New York.

ROBERT LEW

Experimental Study

An experimental study is a study in which conditions are controlled and manipulated by the experimenter. For example, in a comparative **clinical trial** the method of assigning treatments to subjects is determined by the investigator. Often, **randomized treatment assignment** is employed to assure that the innumerable potential **confounding** factors not controlled by the **experimental design** have similar distributions in the various treatment groups. Special design features are used to improve the efficiency of an experimental study, including **stratification**, **matching**, and **factorial experiments** to study several treatments simultaneously. Experimental studies afford good opportunities to reduce the possibility of confounding, to obtain good measurements on various factors that might influence outcomes (*see* **Covariate**; **Effect Modification**), to avoid **biases** in measuring outcomes, and to limit the obfuscating impact of other controllable factors that influence outcomes.

An experimental study is distinguished from an **observational study** in which the investigator does not control the treatment or exposure assignment, nor many other aspects of the process under study. An observational study is thus more subject to problems of confounding, **measurement error**, and bias than an experimental study, but many of these issues must also be carefully considered and controlled in the design, conduct, and interpretation of experimental studies.

(*See also* **Bias from Nonresponse**; **Bias in Observational Studies**)

MITCHELL H. GAIL

Experiment-wise Error Rate

Medical experiments are frequently designed to perform **multiple comparisons**. For example, a two-treatment cancer trial can require comparisons on the following outcome variables (called endpoints): survival, quality of life, and reduction in weight. Similarly, three pairwise comparisons are possible on a three-treatment trial. Alternatively, patients can be **stratified** and subgroup analysis performed on each stratum. Also, repeated significance testing is possible at different stages of a lengthy trial. Let us suppose that L comparisons are to be made, resulting in M' incorrect decisions. Then, the experiment-wise error rate is defined as $\Pr(M' \geq 1)$. Usually, only false positives (type I errors) are of interest, resulting in the typical definition of the experiment wise error rate as $\Pr(M \geq 1)$, where $M \leq M'$ is the number of type I errors committed (*see* **Hypothesis Testing**). The experiment-wise error rate is frequently called the *family-wise error rate*, although the term family can be more restrictive than experiment, as an experiment can consist of several families of comparisons. Ignoring such differences, the experiment-wise error rate, $\Pr(M \geq 1)$, is denoted by FWE.

A companion measure is the *per-experiment error rate*, defined as $\mathrm{E}(M)$. Clearly, $\Pr(M \geq 1) \leq \mathrm{E}(M)$. For $L = 1$ and significance level α, $\Pr(M \geq 1) = \Pr(M = 1) = \mathrm{E}(M) = \alpha$. More recently, a less stringent measure of error rate, called the *false discovery rate*, has received attention; see, for example, [1, 2].

Consider L independent comparisons, each at size α. Assume that the **null hypothesis** for each comparison is true. Then $\Pr(M \geq 1) = 1 - (1 - \alpha)^L$. If, for example, $\alpha = 0.05$ and $L = 6$, then $\Pr(M \geq 1) = 0.26$, which is far greater than the nominal $\alpha = 0.05$. Medical researchers frequently ignore this multiplicity effect and base their conclusions on multiple individual unadjusted comparisons resulting in the over-reporting of false positive treatment differences [7]. On the conservative extreme, the **Bonferroni inequality**, which uses α/L for each comparison, protects the FWE in the *strong* sense, but at the cost of considerable loss of **power**. Such protection can frequently be achieved more efficiently, for example, by using Tukey's T procedure for pairwise comparisons in a one-way **analysis of variance** (ANOVA). Protection in the *strong* sense means that $\Pr(M \geq 1) \leq \alpha$, even when some of the individual null hypotheses are not true. Similarly, protection in the *weak* sense guarantees that $\Pr(M \geq 1) \leq \alpha$ when the L individual null hypotheses are true. The difference between these two concepts can be illustrated with one-way ANOVA. For that case, the *least significant difference* (LSD) procedure recommends performing first an overall F test of size α, and then making individual comparisons, again of size α, only when the overall test is significant. Although this procedure protects the FWE weakly, it can result in a large FWE value when, for example, one of the treatments differs from the others. This difference leads to a significant overall finding, resulting in a loss of protection on the individual comparisons.

The approach used to handle errors in a multiple comparison experiment depends on the type of experiment being performed. Protection of the FWE in ANOVA procedures has been carefully studied [4]. Group sequential methods based on alpha-spending functions (*see* **Data and Safety Monitoring**) have been devised for adjusting for multiple looks [5]. The handling of *multiple endpoints* and multiple treatments are less developed areas at this stage. Multiple endpoints can sometimes be combined into a single measure, such as the *Karnofsky* performance score in gastroenterology. Alternatively, some endpoints can be viewed as primary, while others are secondary, with control of the FWE only for the primary endpoints. For a more technical discussion of treating multiple endpoints in an experiment, see [6]. Multiple treatments can be compared using the Bonferroni inequality, although less stringent methods are often possible. As an example, the comparison of two different treatments with a control can be viewed as two separate experiments, each at size α, as this is the way in which they would have been viewed had they been performed by different experimenters. An excellent source for discussion of the FWE and multiple comparisons procedures is [3]. In summary, care needs to be exercised during the planning of an experiment to ensure protection of the experiment-wise error rate.

References

[1] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Ser. B* **57**, 289–300.

[2] Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. (2001). Empirical Bayes analysis of microarray experiment, *Journal of the American Statistical Association* **96**(456), 1151–1160.

[3] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[4] Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, New York.

[5] Kim, K. & DeMets, D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function, *Biometrika* **74**, 149–154.

[6] Pocock, S.J., Geller, N.L. & Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics* **43**, 487–498.

[7] Pocock, S.J., Hughes, M.D. & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials: a survey of three medical journals, *New England Journal of Medicine* **317**, 426–432.

(*See also* **Simultaneous Inference**; **Studentized Range**)

BORIS IGLEWICZ

Expert Witness, Statistician as

Until fairly recently, the applications of inferential statistics (*see* **Inference**) in legal proceedings have been minor and limited. With the advent of civil rights legislation, however, the courts have embraced statistical inference with enthusiasm. The needs of the courts are not well matched with the usual practice of statistics, and this mismatch has serious adverse consequences for both fields. The various sources of difficulty are outlined, and tentative proposals for their amelioration are put forward.

Although the field of statistics can find its origins in matters pertaining to society and its governance, statistics as a formal discipline has only recently received special recognition in legal proceedings. To be sure, statistics in the sense of numerical summaries are pervasive – in legal settings as in many others. But statistical inference based on probability models is another matter, and in that respect statistics has had only a minor and restricted role in the law.

The Howland will case of 1867 [13], in which Benjamin Peirce undertook a statistical analysis of handwriting, is a case in point. The analysis was ingenious, and might even have been persuasive, but the court in that instance found a technical excuse to put it aside. From time to time, most notably in the Collins case a century later [16], statistical analyses of identification evidence have come before the courts, and generally the courts have rejected them, except in rather special cases of genetic evidence of paternity (*see* **Paternity Testing**) and of fingerprint evidence.

Following the passage of the Civil Rights Act of 1964, however, the courts have looked to statistical analysis to decide on the substantiality of evidence of illegal discrimination, and by now the statistical expert witness is definitely in the big time.

It might be thought that this is cause for celebration within our profession. Inference is our field, of course, and what could be more appropriate than a long overdue recognition by the courts of our special expertise. However, there is room for second thoughts as well, when we pause to consider the consequences for other professions that have come to occupy a similar role. The situation of **psychiatry** (to choose a not-at-all random example) is notorious. In the case of the most recent would-be presidential assassin, John Hinckley, neither the prosecution nor the defense had any difficulty in finding distinguished psychiatrists, academics, and others to testify that Hinckley was or was not legally sane at the time he fired the shots. Indeed, it is not stretching matters to say that the courts and the bar, and even the public at large, have come to hold the profession of psychiatry in considerable contempt – as a clan of hired guns, available for a price to whichever side first knocks on the door. That this perception is not altogether fair is beside the point. The statisticians may have cause for congratulation in this new-found status – they also have cause for worry.

Indeed, psychiatry is not alone in its notoriety. The evident ease with which experts in almost any field can be found to testify in support of either side of a case has led to an aphorism in the law that has a familiar ring to statisticians: there are three kinds of liars – liars, damned liars, and expert witnesses. (The origins of this aphorism are uncertain, but it appears in various forms in legal writing during the past century.) Statistics has had a hard time establishing its credibility as a scientific domain, and the credit that it now has may well be threatened by our new-found prominence.

The views expressed here are idiosyncratic, and the interested reader may wish to consult additional sources dealing with the interaction between statistics and law. In particular, the collections edited by Peterson [17], Monahan & Walker [15], and DeGroot et al. [4] present much relevant material and a number of alternate views.

The remainder of this article is divided into four sections: the first reviews the different domains in which statistical testimony is sought; the next discusses the environment in which such testimony is given and contrasts that environment with the quite different system prevalent in Europe. The manifold corrupting influences that lead to the unsavory reputation of expert witnesses in American courts are then reviewed, and I close with a very modest proposal for reform.

Domains of Application of Statistics in Law

I start by distinguishing particular domains in which statistical expertise is called upon.

Scientific Sampling

The simplest and, in many ways, the most satisfactory application of statistics in legal proceedings, is the use of scientific sampling methods (*see* **Probability Sampling**). In this area, W.E. Deming has been the preeminent pioneer [5], and he has taken the pains to lay out a clear recipe for satisfactory performance. This consists largely of eschewing any responsibility for choice of population to be sampled or for the evaluation of sampled units. His advice, which I believe to be eminently sensible, is that the sampling expert limit himself to testimony about the inference from the sample to the population, when the same evaluation process is used for both. Deming emphasizes that although the sampling expert may have become familiar with the substantive field and may have given good advice about other aspects of the study, his *professional* expertise is limited, and he should testify only within that area.

Following in Deming's footsteps, on a number of occasions I have assisted in the sampling of railroad traffic, in connection with studies of the effects of a merger between two railroads or of the effects on railroad A of the abandonment of certain lines by railroad B. I have presented such work before administrative law judges of the Interstate Commerce Commission and have often been cross examined thereon. However, despite strong controversy and aggressive examination of management personnel, the sampling testimony has generally been accepted with minimum fuss, and the cross examination has ordinarily consisted solely of emphasizing the limits of my responsibility.

Sampling testimony is not always so cut and dried, however. An early case is that of *Sears Roebuck & Co. vs. City of Inglewood* [19], in which the sales to nonresidents had been erroneously subject to tax and Sears was seeking recovery. The expert retained by Sears sampled 33 days from the 826 business days in the period at issue, and all sales slips from each of those 33 days were examined and assessed. The estimated overpayment was $27 000, subject to a **standard error** of $2000. No quarrel with the method of sampling was made, but the judge in the case was uneasy about this unfamiliar technique. He ruled that no recovery could be made for individual sales that had not themselves been individually examined, and Sears had to go back and look at each sales slip. The result is a choice teaching example, of course, because it is one of those exceedingly rare cases in which a well-drawn sample is followed by a complete **census**, illustrating the ultimate validation of the statistician's art. In this case the complete count was surprisingly close to the estimate – a deviation of less than $300.

However, even in the clean world of scientific sampling, a witness may find himself in difficulty. He may be asked to comment on the sample drawn by the opposite party – perhaps by a nonstatistician – and it may turn out to have been a **systematic sample**, without **randomization** of any kind. And here a prudent witness has a problem. The failure to randomize opens the way to possible **biases**, but as all experienced in sampling are aware, for a great many **sampling frames** (i.e. those with very little internal structure) the bias in the estimate and even in the calculated standard error is not likely to be large. Should one testify that the job was not competently done and the results should therefore not be given credence? (Counsel for one's own side would believe such testimony is entirely proper and the least that is owed him.) Or should one testify that, although the sample does not adhere to the canons, it is not likely that the result is for that reason wide of the mark? One will be tempted to add that when incompetence is manifest in the visible part of the operation, it is suspect in that which is less visible, and therefore the result should be received with caution. Should one yield to that particular temptation, however, one is most likely to be cut off by an objection from the opposing attorney, protesting testimony that is "mere speculation". Since the court is only interested in the result of one's judgment about design, one is likely either to overstate the objections to the systematic sample or so understate them as to make one's client wonder why he was put to the trouble of drawing a random sample in his case. (And what should one say when the sample one is called upon to criticize was drawn by Dr Deming or Professor **W.G. Cochran**, who – in the exercise of professional judgment – decided that the fuss required to randomize was not worth the trouble in the case at hand? Indeed, Cochran was fond of telling of the occasion on which he was called on to carry out a sampling study of, I believe, a class of retail stores, and he instructed that the sample consist of every tenth establishment of that type listed in the Yellow Pages. The judge, he said, welcomed his expert testimony as a learning experience and remarked, after Cochran

had been sworn, "I am glad to hear and to learn from Professor Cochran about this scientific sampling business, because I know virtually nothing about it. In fact, about the only thing I *do* know is that you should not just start at the beginning and take every 10th one after that".) I confess that, not being Deming or Cochran, I make it a point when drawing a sample to be sure that the design has as much internal credibility as I can give it, and as little dependence on the quality of my own judgment as I can manage.

Paternity and Fingerprints

In the sampling domain, statistical inference works well because we impose the probability model directly on the situation – through randomization – and our testimony has both the appearance and the substance of relative objectivity. We can feel rather sanguine about our contributions to legal proceedings in this domain.

We have somewhat less security when we turn to certain areas of identification evidence. I refer to blood tests for assessing evidence of paternity and to fingerprint evidence. In the former, at any rate, there is a probability element introduced by Mendelian genetics (*see* **Mendel's Laws**), and the statistical expert may have a real contribution to make. Unfortunately, the ultimate probability calculation depends on population **gene frequencies**; even where these are known for the population at large, it is often some subset of the population that is at issue, for which the frequency is not well established, and the expert finds himself on doubly uncertain ground. There are controversies aplenty in this domain, but this is not where most of the action lies.

Observational Data

The broad, almost limitless, domain in which the courts have come more and more to look to statisticians for guidance is in the analysis of **observational data**.

Consider, for example, the association between cigarette smoking and the subsequent development of lung cancer (*see* **Smoking and Health**). First identified as an incidental and highly uncertain association, the accumulated evidence today appears overwhelming, although there are no clinical experiments to support it and only indirect support from cellular biology. The primary evidence is indeed statistical, and it *is* convincing, but not as a result of conventional significance testing (*see* **Hypothesis Testing**). Rather, it is the robustness of that association over time, place, and population that is convincing. For the most part, probability-based statistical testing is irrelevant to the strength of our conviction.

I do not quite share David Freedman's hardline position against formal statistical inference for observational data. As explained in a superb elementary textbook [9], Freedman regards probability-based testing in a situation without a plausible probability model to be at best irrelevant and more likely misleading. I think, in contrast, that such testing serves a useful purpose as a benchmark. If the observed **association** would not be counted statistically significant had it arisen from a randomized study, it could not be counted as persuasive, when even that foundation was lacking. If the observed association is highly statistically significant, however, the extent of its persuasiveness depends on many uncertain judgments about background factors, and its persuasive value is not at all reflected in the significance level itself.

However, the principles of statistical inference relevant to the courts are not the province of the statistics profession alone.

The Courts, Civil Rights, and Statistics

The Supreme Court has canonized formal statistical inference in a series of decisions, beginning with a jury discrimination case, *Castaneda vs. Partida* [3], decided in March 1977. Having noted that the population of Hidalgo County was 79% Spanish surnamed, but that the jury panels selected in accordance with the prevailing Texas "key man" system averaged only 39% Spanish surnamed (i.e. 339 of 870 jurors), the Supreme Court itself – or more likely one of the Justices' law clerks – calculated the familiar critical ratio (*see* **Normal Scores**) according to the **binomial distribution**; that is, the difference (39% minus 79%, or 40%) divided by the standard error $((pq)^{1/2}/n$, which works out to be 1.5%), obtaining a critical ratio of 29. The Court then commented, "as a general rule for such large samples, if the difference between the expected value and the observed number is greater than 2 or 3 standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist".

Formal significance testing next appears in an employment discrimination case, *Hazelwood School*

District vs. United States [14], decided three months later. In that case, the proportion of qualified teachers in St Louis County (excluding the city of St Louis) who were black was estimated to be 6%, and during the two year period at issue, only 15 of 405, or 4%, were black. The *Hazelwood* court now says, "A precise method of measuring the significance of such statistical disparities was explained in '*Castaneda v. Partida*' ...", and the opinion goes on to paraphrase the earlier two or three standard deviation rule, but with a slight shift; that is, "... if the difference exceeds 2 or 3 standard deviations, then the hypothesis that teachers were hired *without regard to race* would be suspect" (emphasis added). The reference to randomness is now absent, as is the social scientist. Since it is self-evident that the process of selection is not – nor is it desirable that it be – random, it is far from clear why either the social scientist or the Supreme Court should look upon a standard based on randomness as appropriate to assess the likelihood of purposeful discrimination. To be sure, there was much other evidence in the case, showing explicit discrimination at earlier dates, but the preceding quotation is the only place in the opinion where the relevance of the statistical significance test is in any way explained. Nonetheless, in *Hazelwood* the court went further, in an obscure remark that pointed clearly to the preeminence of statistics. It said, "Where gross statistical disparities can be shown, they alone may in a proper case constitute prima facie proof of a pattern or practice of discrimination". Thus, in the space of a less than half a year, the Supreme Court had moved from the traditional legal disdain for statistical proof to a strong endorsement of it as being capable, on its own, of establishing a prima facie case against a defendant. (It is sometimes argued that the use of doubtful evidence to support a prima facie – that is, preliminary – finding is a matter of small legal consequence. Such a finding merely shifts the burden of proof from the plaintiff to the defendant. In fact, however, there is nothing at all "mere" about this shifting of the burden, since the difficulty of proving oneself innocent of discrimination turns out to be great indeed.)

The accelerating role of statisticians in employment discrimination cases arises from a combination of the statistical significance testing endorsed in *Hazelwood* with an earlier decision, *Griggs vs. Duke Power Company* [11], in 1971. In *Griggs*, it was found that the requirements of a high school diploma and a certain score on a standardized IQ test for employment in such jobs as maintenance and laboratory work operated to exclude black applicants far more frequently than they did to exclude white applicants. The court concluded that, in the absence of direct evidence that these criteria related to improved performance on the job, the "adverse impact" of those requirements constituted a violation of Title VII, even though there may have been no intent to discriminate on the grounds of race. (To be sure, there was plenty of evidence of intent to discriminate in the *Griggs* case, most especially in the facts that the power company had explicitly excluded blacks prior to passage of the Civil Rights Act, and that it had put in the new requirements at the same time that the jobs were first made available to blacks. The principle established in *Griggs* clearly put the issue of intent aside, however, and the doctrine has been widely applied by lower courts in cases in which there was little or no evidence of invidious intent. Once again, proof of "job relatedness" or, as the Supreme Court says, "business necessity", has proved generally elusive, and a great many employment standards and admissions tests have been found to be in violation of the law for lack of such proof.)

The criterion seems reasonable enough until we are faced, as was the Supreme Court, with a case such as *Washington vs. Davis* [20]. In this case, Walter Washington, mayor of Washington, D.C., and his police chief had set about to recruit blacks into the D.C. police force, with an aggressive campaign to encourage black applications. The campaign was successful and many blacks were recruited, but among the newly encouraged applicants the written test "operated", in the language of the *Griggs* decision, "to disqualify Negroes at a substantially higher rate than White applicants". Thus, affirmative action, clearly intended to recruit blacks, fell foul of the adverse impact principle developed in *Griggs*. The trial court dismissed the charge, but the appeals court reversed, citing *Griggs*. The Supreme Court sidestepped the issue. It supported the district program, but it did so on a technicality that did not require it to comment on the general validity of the *Griggs* principle. In subsequent cases, the *Griggs* principle has continued to guide the lower courts.

It must be acknowledged that there has been some slackening of the tide in Title VII enforcement in the last two or three years, especially with regard

to race and sex discrimination. Under the Reagan administration there was more emphasis on freeing business from government interference and less on righting the wrongs of the oppressed. In response, the courts appear to have given somewhat less weight than before to purely statistical evidence.

The Position of the Statistical Expert

The result of the preceding and related decisions has been to place the statistical expert witness in a most unaccustomed and exalted position. Despite the moderate recent decline in enforcement, the role of the statistical expert remains critical in the cases that are brought. Lawyers gaze with awe as he examines the entrails of complex **multiple regression** computer output, and they await breathlessly his conclusion that the coefficient of the variable designating sex is indeed more than twice the standard error. Similar attention attends his calculation of continuity-corrected 2×2 chi-squares to see whether they are larger or smaller than 3.84 (*see* **Chi-square Tests**). Indeed, the case may be won or lost largely on these outcomes.

That this position is a false one, none can doubt. Certainly the statistical experts know it, and most of them say so to some extent – or at least they assert that the meaningfulness of their numerical results depends on a number of assumptions that they are unable to verify. The courts, however, are not engaged in academic exercises and, having urgent need to come to some conclusion, turn to the Supreme Court instead of to the witness's cautional phrases for guidance. Those opinions have, intentionally, a somewhat Delphic quality. They tell us that "gross disparity" in pass rates is evidence of illegal discrimination, and they also tell us that the hypothesis of random selection is made to appear doubtful when a difference is larger than two or three standard errors. They do not quite say that statistical significance at the 5% level constitutes gross disparity, but that is how the lower courts read them. The statistical experts cannot help but find this heady stuff, and we should not be surprised to find ourselves speaking with far more assurance about our conclusions than an objective appraisal of the evidence might warrant.

Thus we are led to the unedifying spectacle of two well-qualified statistical witnesses providing analyses that they interpret oppositely, each supporting the interest of the party who introduces him. Other categories of expert witnesses have been there before us, of course – the psychiatrists, medical internist and surgeons, and structural engineers, among others. The courts have urgent need for the assistance of these experts, but they seem uncommonly ill served by them. The point was made clearly in an editorial in the *British Medical Journal* [2]:

> Medical evidence delivered in our courts of law has of late become a public scandal and a professional dishonour. The Bar delights to sneer at and ridicule it; the judge on the bench solemnly rebukes it; and the public stand by in amazement; and honourably minded members of our profession are ashamed of it.

This was printed more than a century ago, but little has changed in the intervening years. Statisticians have escaped comparable condemnation because we have been, until recently, too unimportant in the courts to be noticed, not because of any higher ethical standard of our profession.

One cannot help speculating on the possibility of improvement. In fact, I believe that some of the difficulty is structural, and that there are ways in which we could function usefully in legal settings without so large a sacrifice of professional integrity. To this end, I now discuss the players in the game and the key influences on them. Among the players or participants in the legal ballet, I distinguish three: (i) the courts themselves, most especially the Supreme Court, who together with the Congress set the rules by which the system operates; (ii) the lawyers – collectively, the Bar – who primarily control the direction of play within those rules: and (iii) the expert witnesses on whose performance the integrity of the enterprise ultimately depends.

The Courts and the Expert

The obvious objective of the courts in respect to expert testimony is to optimize the search for truth. The courts would like to get the most well-qualified expert, keep him in a situation in which he can devote his best efforts to analyzing the evidence, and have him testify in an atmosphere free of coercion or bias. The courts also want to be sure that the expert is adequately examined to test and verify his qualifications, the adequacy of his preparation, and his objectivity.

To this end, the courts in Germany and France arrange matters very differently from the English

and American courts. In cases in which experts are needed, they are in the first instance appointed by and responsible to the court and not to either party: they are first examined by one of the judges and also cross examined by him. Attorneys for the plaintiff and defendant may also cross examine, but the proceedings are not generally adversarial as are our own, and the appearance of neutrality, at least, is the rule. Thus the continental system seeks the best witnesses and seeks to put them in a neutral setting, primarily by putting the major responsibilities in the hands of the judges.

The Anglo-American system, in contrast, is based on the proposition that truth is most likely to emerge through the best efforts of adversaries. No point in favor of the defendant will be overlooked or undervalued, it is thought, if responsibility for bringing it out is assigned to the defendant's advocate.

Nonetheless, whatever the merits of the adversary system may be in general, it is well recognized that it wreaks havoc with expert testimony, and proposals for reform appear regularly. Chief among them is to borrow from the continental system and to have the primary expert witnesses appointed by and responsible to the court. This reform was vigorously advocated by the past century's revered commentator on legal evidence, John Wigmore [21], and model codes have been proposed to this end. Indeed, Rule 706 of the Federal Rules of Evidence provides explicitly for court-appointed experts. Regardless of the merits, in practice this power is used extremely sparingly. (There may be some cases of court-appointed statisticians in Title VII cases, but I have not heard of any.) One can conceive of many reasons for the ineffectiveness of these "reforms", not least the vulnerability to criticism of a judge who appoints an expert later shown to be inadequate, but it is enough for my purposes to observe that such reforms have not taken hold in this country, and that they do not seem likely to become influential in the near future.

The Bar and the Expert

The position of legal counsel, although in principle identical to that of the judge, is in fact quite different. Having committed himself to the adversary system as the best method of reaching a just conclusion, the lawyer for the plaintiff now accepts his position in the system, that of advocate, and leaves to the court the responsibility for discerning the path of justice. To him, the expert is simply one of the elements that he must fit into place to make the most effective case. To be sure, any lawyer of competence recognizes that it is usually favorable to his case for the witness to appear to be dispassionate and objective. The best lawyers recognize that a witness will make the best appearance of objectivity if he feels that he is indeed free to go where his research and reflection lead him. This is not to say that these excellent advocates are really in the market for unbiased witnesses who may testify to their side's disadvantage.

John C. Shepherd of St Louis, a distinguished trial lawyer who was president of the American Bar Association in 1984–1985, spoke to a conference for lawyers on relations with the expert witness, and this [18, pp. 21–22] is what he said:

> Many people are convinced that the expert who really persuades a jury is the independent, objective, nonarticulate type ... I disagree. I would go into a lawsuit with an objective, uncommitted, independent expert about as willingly as I would occupy a foxhole with a couple of non-combatant soldiers.
>
> If you find the expert you choose is independent and not firmly committed to your theory of the case, be cautious about putting him on the stand. You cannot be sure of his answers on cross-examination. When I put an expert on the stand, he is going to know which side we are on.
>
> The trial lawyer must make of the expert a convincing, persuasive witness. The lawyer deals in words, and he knows how to put the package together to impress the jury favorably. It is his job to instruct the expert, an exercise requiring great tact and firm conviction.

Keep in mind that the lawyer does not need to make bricks without straw. It is perfectly proper for him to consult a great many potential witnesses but to bring to court only that one whose honest convictions fit well with the lawyer's needs. The phenomenon of "shopping for witnesses" is well recognized by the courts, and it contributes to the wary attitude that they have about experts in general. The shopping is done by the lawyers, however, and is thus not subject to exposure in the actual testimony.

Corrupting Influences

As we have just seen, the professional integrity of the expert witness and, through him, of the profession that he represents, is not well protected by the courts

and hardly at all by counsel. But before we assume too readily that simple morality and personal ethics will be an adequate substitute, we should reflect for a bit on what I call, for lack of a more delicate phrase, *corrupting influences*. Some are inherent in the nature of the situation, and others are special to the adversary situation.

First, there is the fact that the expert witness is playing someone else's game and, inevitably, has to accept the rules as he finds them. His instructor in these matters is, of course, his client's counsel, and the witness is ill-equipped to resist the role of adversary when his lawyer thrusts it upon him. But even supposing that the lawyer is less demanding than Shepherd, the expert is beset with temptations.

General. Among the most difficult of the corrupting influences to deal with is what I call *aggrandizement*. In Title VII cases (i.e. those dealing with employment discrimination), the Supreme Court has placed the statistician in the key role. Long ignored and treated with contempt in literature and in the courts, the statistician has been elevated to Olympian levels. Thus the *Hazelwood* court, quoting its remark in an earlier case, commented [12]:

> We also noted that statistics can be an important source of proof in employment discrimination cases, since, "absent explanation, it is ordinarily to be expected that nondiscriminating hiring practices will in time result in the work force more or less representative of the racial and ethnic composition of the population in the community from which employees are hired." Evidence of long lasting and gross disparity between the composition of the work force and that of the general population that may be significant even though paragraph 703 (j) makes clear that a work force need not mirror the general population.

Taken together with the court's embrace of statistical significance testing, the statistician is here given a virtual license for intellectual robbery. Indeed, not only the court but a large contingent of fellow academics (economists numerous among them) give strong endorsement to the particularly magical properties of multiple regression analysis. (Two articles in the *Columbia Law Review* – Fisher [8] and Finkelstein [7] – are noteworthy in this regard.) All in all, the statistician is strongly tempted to give the definitive rather than a qualified answer to the key questions. He will be tempted to ignore or to minimize those qualifications that he might emphasize in a more academic setting, he may fail to emphasize the existence of schools of thought other than his own, and he may lay claim to overly broad scope for the inferences he draws.

Adversarial. The adversary system adds a host of additional influences, some quite direct, but others indirect:

1. *Bribery*. The witness is paid by his client and, as often noted, he who pays the piper feels a right to call the tune. To be sure, all the client is entitled to is an honest report of the expert's best effort, but an expert who habitually finds evidence against his client will not be much sought after.
2. *Flattery*. Some, of course, are not bought by money or the prospect of future money: either they already have enough of it or they are sufficiently on guard against that particular type of seduction. Other corruptions await them.
   I well recall an occasion on which I was asked to consult in a case at a time that was not especially convenient. I explained that I really could not participate on this occasion. The lawyer, with whom I had worked before and for whom I had a great deal of respect, pled the sorry state of statistical testimony in the courts in general, and in the instant case in particular. He read from the transcript some particularly egregious quotes from the statistical expert for the other side, and he urged the importance for the future of statistics in the domain of public affairs of having corrective testimony. That being a viewpoint I could only share, and tacitly mindful of our shared opinion that I was the ideal candidate to champion the honor of the profession, I reluctantly agreed to testify. Imagine my chagrin when, at a later date, I read some other remarks of the trial lawyer, John Shepherd, whom I quoted earlier. He advises on "Approaching the Expert" as follows [18, p. 19]:

   > Almost every one who considers the subject of experts in court will start with the same thought: The first thing you need to get along with your expert witness is money. But the hiring and successful use of an expert may not be that easy – a lot of good experts are rich. Although you will eventually be talking about money with your expert, it is wiser to begin on another tack. Tell your expert how justice will be served if he will testify on your

> side of the case. Remind him that the unfortunate situation in our courts today can be improved if we have people of his caliber to help in the administration of justice. That ploy will impress even the rich expert.

3. *Co-option*. To be sure, effective as this ploy may be, it does not in itself lead the expert away from his duty. It establishes an aura of objectivity and mutual respect, however, which may make the expert especially vulnerable to another inevitably corrupting aspect of the adversary system: that is, the simple fact that the expert's introduction to the case comes from the client's counsel and will inevitably tend to appear in the light most favorable to the client. He will be introduced to the principals – perhaps a plaintiff, movingly indignant about years of abuse and low pay, perhaps a defendant who truly believes that his cause is just and is worried sick about the distraction of his institutional resources from their proper role into the defense against a baseless charge. This goes along with co-option into advocacy arising when one is asked to review the other side's testimony, point out flaws therein, and assist in the development of effective cross examination.
   F. Downton of the University of Birmingham has written cogently about this latter difficulty, in a symposium on statistics and the law [6]. Downton had been consulting with the police on games of chance, because the law prescribed strict rules for games that, if violated, would allow the police to close the clubs. Since the clubs were widely regarded as dens of iniquity, this was clearly a public service. Downton wrote [6, p. 171]:

   > As in any other consulting situation, a certain amount of identification with the aims of the client is inevitable; it is fortunate that probability and statistics are basically mathematical in content, since the constraints of mathematics act as a brake on overenthusiasm. It cannot, however, be denied that a conscious change of attitude was needed to effect the change-over from helpful consultant to objective expert witness. . .. This ambiguity of roles did create a conflict, which presumably can only be resolved by individual witnesses in their own way.

4. *Gladiatorial Role*. The adversarial environment works against objectivity in yet other ways. The object of cross examination is not only to expose weaknesses in the expert's analysis but, if possible, to discredit the witness and the weight that should be given to his testimony generally. Thus the cross examiner may, by adroit framing of questions, force the witness into complex explanations and apparent contradictions. Feeling his credibility slipping away, such a witness may be less likely to give a full and frank answer to a later question that might fairly expose a fact or conclusion operating in favor of the other side. The expert no longer views his interrogator as a fellow searcher for truth, but as an adversary against whom he must defend.
5. *Personal Views*. My final source of corruption is perhaps the most difficult to deal with, and that is the problem of strongly held personal views. Surely there are many cases in which the expert is a priori indifferent between the claims of the contestants, but in other areas, particularly in the great domain opened up by Title VII, there are few of us without strong opinions. In the matter of a contest between a chemical waste disposal company and the residents of a new Love Canal, for example, I would be reluctant to testify on behalf of the company. It might well be, for example, that the evidence of adverse health effects caused by carelessly buried wastes is really nonexistent. Feeling as strongly as I do, however, that such careless behavior is reprehensible and deserving of punishment, I should not like to assist the company's case. I have no problem reconciling my preferences and my professional responsibilities in this case. I am, and should be, free to accept an engagement or not, for whatever personal reason, and reasons of this kind are at least as good as most others.

My problem comes on the other side. Suppose that I should be an expert retained by the residents affected by the dump. I find that, in respect of total mortality, there is no evidence of an effect, but in the matter of childhood leukemias there is an **excess mortality** amounting to 1.8 standard errors greater than the rate in some **control** group. If I ignore the fact that I am reporting on leukemia because it is the disease category showing the largest difference, and if I adopt the conventional 5% significance level as a standard, and if I urge the relevance

here of a one-sided significance test (*see* **Alternative Hypothesis**), I may be able to strike a blow for truth and justice, and it would no doubt be tempting to do so. But to paraphrase a major figure in the Watergate investigation, "I could do that, but it would be wrong". I really do not think one-sided 5% level deviations provide convincing evidence one way or the other, and – whatever one's views on that – I expect that most statisticians would agree with me that it is misleading to the point of dishonesty to quote an unadjusted significance level (*see* **Level of a Test**) when I have chosen to present the most extreme of a number of alternative measures.

Perhaps the point can be brought home most forcefully by addressing an even touchier example. There are many of us who view the legacy of slavery as our most appalling and pressing social problem, and the effort to explain the low status of the oppressed on the grounds of inherent inferiority as an intolerable offense. Indeed, although the possibility of *some* average difference in intellectual capacity among different groups can never be ruled out, the evidence appears clear that whatever differences there might be in *average* innate ability, they are quite small compared with the variation between individuals. The effects ascribed to race in **regression** analyses of school child performance, after adjustment for age, years of schooling, mother's socioeconomic status, and the like, are readily explainable as attenuation and other distracting effects that afflict regression analyses generally, and they need not be interpreted as reflecting a real difference due to race.

At the same time, we observe the past and present systematic discrimination against blacks in many areas of employment. Such discrimination has many forms, but its pervasiveness, except where sharply controlled by law, is hardly in doubt. Being confident, then, that a charge of race discrimination likely corresponds to the existence of actual discrimination, what are we to say about a multiple regression in which a salary difference unfavorable to blacks emerges as significant even after adjustment for age, years of schooling, mother's socioeconomic status, and the like? The problems of attenuation apply with equal force, but we may now be reluctant to dismiss the evidence of bias in pay. This time we may believe that there really is discrimination in the system, but it is by no means clear why we should, as statisticians, take different positions in the two situations.

Ways to Defend the Integrity of Statistical Testimony

With the variety of assaults on the credibility of expert statistical testimony, I turn again to the question of what possible defensive measures could be implemented. A change to the apparently more neutral continental system is the one answer that has come from the courts, but there seems little likelihood of its adoption. There have been proposals that experts who testify falsely should be punished for perjury, as is an ordinary witness who testifies falsely about an event. This, too, seems far-fetched, since the essential nature of expert testimony is that it is largely a matter of informed opinion.

Professional Codes

There seems to be only one other direction in which to turn, and I have only slender hopes for it. Seeing that neither the bench nor the bar will help us, the only alternative is to help ourselves; that is, to develop limited codes of ethical behavior in the context of legal proceedings that may help to ameliorate the worst excesses.

I come to this conclusion reluctantly, because I have little taste for collective moral instruction and little confidence in its efficacy in general. And yet one cannot deny that codes of ethics for judges, while they do not eliminate venality, are good to have. Violation of these rules can and does lead to discipline, on occasion, as in the case of a distinguished Supreme Court Justice a few years ago, and reminders such as that help to keep others on the right path. Similarly, codes of medical ethics dealing with the proper relationship between physician and patient serve a useful purpose.

A quarter of a century ago, Gibbons [10] reviewed our society's efforts in the direction of developing such codes, and she clearly laid out some of the problems that such codes might help to solve. Evidently, there was some movement in that direction in the early 1950s, but momentum was lost, and nothing came of it. The issue of ethical codes continues to elicit debate, a recent instance being the report of the Ad Hoc Committee on Professional Ethics [1]. I see no sign, however, that this or any other code is likely to be adopted as a guide by any of our major professional organizations. Indeed, although discussion of codes of ethics for statisticians continues, I know of

only one instance in which such a discussion has had any noticeable practical effect.

The exception is an interesting one, and it may be instructive. As I mentioned earlier, in the context of sample survey design and analysis, W. Edwards Deming established a code that he provided to clients, explaining the reach and the limitation of his methods. The code is notable for its careful restriction of the role of the statistician. In effect, Deming acknowledges that the statistical consultant may come to have a good deal of knowledge about the subject matter under study, and that this knowledge may help him to design an effective sample. He makes clear, however, that responsibility for the choice of population to be sampled (*see* **Target Population**), and for the processing of each sampled unit, belongs to the client and not to the sampling consultant. The consultant undertakes to say only that, had the entire population been processed in the same way that the sampled elements were processed, the sample estimate for the population would be found to be close to the population value, subject to error limits that can be given in the usual probability sense. It might seem that the scope of the sampling expert's testimony is so narrow according to this code that his contribution will have little weight in the proceedings. In fact, of course, the contrary is true. By not reaching beyond well-stated boundaries, the testimony of sampling experts has achieved an enviable level of credibility.

Deming's code, effective as it is in a specific context, gives only a little guidance for the expert testifying in a Title VII case. I submit that a proper code for the latter expert should copy Deming by being specific to the situation and rather restrictive as to the scope of the testimony. I do not think it will pay to start with an ethical code trying to embrace all statistical activities. Let me try to clarify my proposal by being specific. (Here I borrow from Deming where I can.)

I suggest that a statistician asked to testify in court should require that he be given access to all data thought by the client to be relevant, and to all previous analyses of that data, and he should demand a commitment on the part of the client to a "good faith" effort to supply whatever other data the statistician may judge relevant.

He should advise the client that in his professional role he will remain neutral between the parties (and he should pray for strength when he does this, for he will need it). He undertakes to provide his best effort to analyze the data in ways that seem to him pertinent, and he undertakes further to provide a written report. His report, if it is to be used, must be taken in its entirety.

When testifying, the expert will explain the limitations of his techniques, as seen by a professional statistician, regardless of any statistical principles anointed by the Supreme Court. He will explain the variety of schools of thought within the profession and his place among them.

Doubtless there are a number of other principles to be enunciated, but this is not the time or place for full details. Some will think that the principles given are simple and obvious, but they can be assured that they are not obvious to most lawyers. Many lawyers, for example, think it proper to select for attention the principles laid down by the high court as a basis for expert testimony. (I cannot help but wonder if, should a court declare that $\pi = 3$, these lawyers would insist that we accept that too.) Gratuitous testimony about limitations will be especially unwelcome (and, in my opinion, especially necessary). The point is that the expert should be much more his own man and much less the puppet of his client's counsel than is typically the case today.

Consequences

The consequences of such a code, should we adopt and use it, are substantial and not entirely welcome. I do not believe it would help us much in protecting against the influence of our own strongly held social views or against the biases that arise because we are oriented and informed by just one of the adversaries. Nor would it keep us from reflexive defensiveness under hostile cross examination.

Adherence to such a code is likely to result in a reduction of the pivotal role that statistical analysis has come to play in discrimination law, and we may see the resulting gap filled, by others whose competence and good will we question even more than our own. It is conceivable that a more modest posture might lead the courts to seek greater clarity by adopting the reforms, if such they be, of the continental system with court-appointed experts. It is certain, I think, that adherence to such a code would improve the credibility of statistical witnesses.

Acknowledgments

This article is a modified version of a previously published paper: "Damned Liars and Expert Witnesses", *Journal of the American Statistical Association* **81** (1986) 269–276.

References

[1] Ad Hoc Committee on Professional Ethics (1983). Ethical guidelines for statistical practice, *American Statistician* **37**, 1–20.
[2] *British Medical Journal* (1863). Medical evidence in courts of law (editorial), 2 May, 456–457.
[3] *Castaneda vs. Partida* (1977). 430 US 482.
[4] DeGroot, M., Fienberg, S. & Kadane, I.B., eds. (1997). *Statistics and the Law*. Wiley, New York, to appear.
[5] Deming, W.E. (1954). On the presentation of the results of sample surveys as legal evidence, *Journal of the American Statistical Association* **49**, 814–825.
[6] Downton, F. (1977). Experience as an expert witness in gambling cases, *Statistician* **26**, 163–172.
[7] Finkelstein, M.O. (1980). The judicial reception of multiple regression studies in race and sex discrimination cases, *Columbia Law Review* **80**, 737–754.
[8] Fisher, F. (1980). Multiple regression in legal proceedings, *Columbia Law Review* **80**, 702–736.
[9] Freedman, D., Pisani, R. & Purves, R. (1978). *Statistics*. Norton, New York.
[10] Gibbons, J.D. (1973). A question of ethics, *American Statistician* **27**, 72–76.
[11] *Griggs vs. Duke Power Company* (1971). 401 US 424.
[12] *Hazelwood School District vs. United States* (1977). 433 US 299.
[13] Meier, P. & Zabell, S. (1980). Benjamin Peirce and the Howland will, *Journal of the American Statistical Association* **75**, 497–506.
[14] Meier, P., Sacks, J. & Zabell, S. (1984). What happened in Hazelwood: statistics, employment discrimination, and the 80% rule, *American Bar Foundation Research Journal*, **Winter**, 139–186.
[15] Monahan, J. & Walker, L. (1985). *Social Sciences in Law: Cases and Materials*. Foundation Press, Mineola.
[16] *People vs. Collins* (1968). 68 Cal. 2d 319.66 Cal. Rptr. 497.
[17] Peterson, D.W. (1983). Statistical inference in litigation, *Law and Contemporary Problems* **46**, 1–303.
[18] Shepherd, J.C. (1973). Relations with the expert witness, in *Experts in Litigation*, G.W. Holmes, ed. Institute of Continuing Legal Education, Ann Arbor.
[19] Sprowls, R.C. (1957). The admissibility of sample data into a court of law: case history, *UCLA Law Review* **54**, 222–232.
[20] *Washington vs. Davis* (1976). 433 US 229.
[21] Wigmore, J.H. (1940). *Evidence in Trials at Common Law*. Little, Brown, & Company, Boston.

PAUL MEIER

Explained Variation Measures in Survival Analysis

When fitting a **regression** model to survival data, one is frequently faced with a situation in which accurate **predictions** for individual survival cannot be derived from the model even though the regression coefficients are highly significant and the model fits the data well. For example, consider a population where the five-year survival rate drops from 60 to 30% depending on whether a certain **risk factor** is present or absent (*see* **Prognostic Factors for Survival**). Suppose that the risk factor is present in 50% of the population, so that, overall, 45% will survive beyond year five. Although the risk factor has an important impact on survival, it is hard to use it for predictions of whether an individual will be alive after five years. In particular, knowledge of the risk factor does not gain much precision for predictions compared to predictions based on the overall rate of survival. Thus, the predictive power of the risk factor is low. Still, in a sample of sufficient size, a model that accounts adequately for the risk factor will fit the data well and yield a significant regression coefficient. Measures for the proportion of explained variation similar to the *coefficient of determination* used in **multiple linear regression**, R^2, can be helpful in order to quantify predictive power and to separate it conceptually from statistical significance and **goodness of fit** [5].

The first measures of explained variation suitable for survival data appeared around the beginning of the 1990s, and there is as yet no commonly agreed choice [10]. Two proposals are described here in some detail, with more references toward the end of the article.

Korn and Simon [4] took an approach based on **loss functions** to measure the explained variation. Let T be a survival time distributed according to a *survival curve* $S(t) = P(T > t)$ (*see* **Survival Distributions and Their Characteristics**), and let $L(t, p)$ be the loss incurred when survival time p is predicted for an individual with actual outcome t. An optimal predictor $\tilde{p}$ for T is one that minimizes the expected loss $\int_0^\infty L(t, p)\, \mathrm{d}[1 - S(t)]$ over p (*see* **Decision Theory**). Let $R(S)$ be the expected loss (or risk) associated with $\tilde{p}$. Then, on the basis of a sample of size n in which x_i is a **covariate** for individual i and $\hat{S}(.|x_i)$ is an estimate of the survival curve for individual i, the variation explained by the model $\hat{S}(.|x_i), i = 1, \ldots, n$, may be defined as

$$\text{explained variation} = 1 - \frac{\frac{1}{n}\sum R(\hat{S}(.|x_i))}{R(\overline{S})}, \qquad (1)$$

where $\overline{S} = \frac{1}{n}\sum \hat{S}(.|x_i)$ estimates the marginal distribution of survival based on the mean of the estimated conditional survival curves given the covariate. With squared error loss, $(t - p)^2$, the optimal predictor $\tilde{p}$ is the mean, and (1) is a model-based estimator of the proportional reduction of variance achieved when the marginal variance of the survival outcome is compared to the mean conditional variance, given the covariate. Other loss functions [4] include absolute error, $|t - p|$, squared and absolute error loss on log scale, $(\log t - \log p)^2$ and $|\log t - \log p|$, and squared error loss censored at some specified time t_0: This means that a prediction $p = t_0$ is considered successful even if the actual survival t is greater than t_0, and incurs no loss. Thus, compared with simple squared error loss, $\mathrm{var}(\tilde{T})$ is used in place of $\mathrm{var}(T)$, where $\tilde{T} = \min(T, t_0)$. Finally, if all that matters is prediction of whether an individual will have died by time t_0, binary squared prediction error loss, $[I(T > t_0) - p]^2$, with I the indicator function, may be considered. The resulting risk is the **binomial** variance of the survival status at t_0, $S(t_0)[1 - S(t_0)]$. In the introductory example, this gives an explained variation of $1 - \left\{\frac{1}{2}(0.6 \times 0.4 + 0.3 \times 0.7)\right\}/(0.45 \times 0.55) = 0.09$ at $t_0 = 5$ years.

Note that, in this proposal, **censoring** is adequately dealt with in two ways. The choice of a censored version of the loss function allows one to view (1) as an estimator of a population parameter where the range of data to be collected in a sample can be accounted for. For example, in a study that is planned to be analyzed at a median follow-up time of five years, squared error loss censored at year five may be an appropriate choice. On the other hand, if some individuals are censored prior to year five because of limited follow-up, the estimate given in (1) can still be calculated, since estimated survival curves for those individuals will usually be available beyond the time of censoring. If censoring is dealt with adequately in the estimation of survival curves in the sense that it introduces no bias, then so it is in the estimation of explained variation. However, a drawback of the measure is that when misspecified models

are used to estimate survival curves, the explained variation shown in (1) may be misleading: It incorporates no comparison of observed versus predicted survival and thus measures only the variation that the model itself appears to explain.

Graf et al. [2] proposed a measure of explained variation, where observed and predicted values of the survival status are contrasted explicitly. It is based on binary squared prediction error loss. For a sample of size n, let t_i denote the time individual i was under observation, and let the censoring indicator δ_i equal 1 if individual i was observed to fail at t_i, 0 if it was censored at t_i. Then let

$$\text{explained variation} = 1 - \frac{\sum_{i=1}^{n} w_i [I(t_i > t_0) - \hat{S}(t_0|x_i)]^2}{\sum_{i=1}^{n} w_i [I(t_i > t_0) - \hat{S}(t_0)]^2}, \qquad (2)$$

where the weights

$$w_i = \begin{cases} \dfrac{\delta_i}{n\hat{G}(t_i)} & \text{if } t_i \le t_0 \\ \dfrac{1}{n\hat{G}(t_0)} & \text{if } t_i > t_0 \end{cases}$$

incorporate the **Kaplan–Meier** estimator $\hat{G}$ of the censoring or potential follow-up distribution [9] calculated by exchanging the role of censored and uncensored observations, that is, with censoring indicator $1 - \delta_i$. $\hat{S}$ denotes the usual Kaplan–Meier estimator of the entire sample, which is used to estimate the marginal failure time distribution $S(.)$. Under suitable regularity conditions, this measure produces a consistent estimator of the population explained variation

$$1 - \frac{\mathrm{E}_X\{\int_0^\infty [I(t > t_0) - \hat{S}(t_0|X)]^2 \, \mathrm{d}[1 - S(t|X)]\}}{\int_0^\infty [I(t > t_0) - S(t_0)]^2 \, \mathrm{d}[1 - S(t)]}, \qquad (3)$$

in which the sample means given in (1) and (2) are replaced by the expected risks of the covariate-based model and the marginal distribution of failure time. This will hold even if censoring is present and if a misspecified model $\hat{S}(.|X)$ is used instead of the optimal predictor. The measure can be adapted for other loss functions.

Two other explained variation measures were proposed specifically for the context of **proportional hazards** models (*see* **Cox Regression Model**), and censoring is adequately dealt with in both. Schemper and Henderson [8] constructed a measure in which the model-based and marginal variance of the binary survival status $I(T > t)$ at time t is averaged across a time-interval $[0, t_0]$, weighted by the marginal survival distribution. O'Quigley and Xu [7] suggested a measure that compares mean squared Schoenfeld residuals (*see* **Residuals for Survival Analysis**) under a proportional hazards model to the null model of no covariate effect. Although their measure has a range of desirable properties, it actually measures the predictability of a covariate from a given failure time and thus does not directly aim at quantifying the ability of the model to predict time to failure from a given covariate.

Other proposals to modify R^2 in a way suitable for survival analysis rest upon interpretations of R^2 different from the proportion of explained variation. The relation of R^2 to the **likelihood ratio** for a model with covariate against the null of no covariate effect is explored in [6, 11, 12]. Approaches related to the interpretation of R^2 as (squared) correlation between observed and predicted values are taken in [1, 3, 4].

References

[1] Akazawa, K. (1997). Measures of explained variation for a regression model used in survival analysis, *Journal of Medical Systems* **21**, 229–238.

[2] Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data, *Statistics in Medicine* **18**, 2529–2545.

[3] Harrell, F.E. Jr., Lee, K.L., Califf, R.M., Pryor, D.B. & Rosati, R.A. (1984). Regression modelling strategies for improved prognostic prediction, *Statistics in Medicine* **3**, 143–152.

[4] Korn, E.L. & Simon, R. (1990). Measures of explained variation for survival data, *Statistics in Medicine* **9**, 487–503.

[5] Korn, E.L. & Simon, R. (1991). Explained residual variation, explained risk and goodness of fit, *American Statistician* **45**, 201–206.

[6] Nagelkerke, N.J.D. (1991). A note on the general definition of the coefficient of determination, *Biometrika* **78**, 691–692.

[7] O'Quigley, J. & Xu, R. (2001). Explained variation in proportional hazards regression, in *Handbook of*

Statistics in Clinical Oncology, J. Crowley, ed., Marcel Dekker, Inc., New York, pp. 397–409.

[8] Schemper, M. & Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression, *Biometrics* **56**, 249–255.

[9] Schemper, M. & Smith, T.L. (1996). A note on quantifying follow-up in studies of failure time, *Controlled Clinical Trials* **17**, 343–346.

[10] Schemper, M. & Stare, J. (1996). 'Explained variation in survival analysis', *Statistics in Medicine* **15**, 1999–2012.

[11] Verweij, P.J.M. & Van Houwelingen, H.C. (1993). Cross-validation in survival analysis, *Statistics in Medicine* **12**, 2305–2314.

[12] Xu, R. & O'Quigley, J. (1999). A R^2 type measure of dependence for proportional hazards models, *Nonparametric Statistics* **12**, 83–107.

(*See also* **Survival Analysis, Overview**)

ERIKA GRAF

Explanatory Variables

In many studies, multiple measurements are available for each individual. One variable may be considered to be the response of interest or the outcome (*see* **Response Variable**). A model may be developed to investigate the form of the relationship between the outcome and the remaining variables, this usually being some form of **regression** model. These variables are then often termed the explanatory variables. Thus, explanatory variables are distinguished from the response or outcome. They are denoted as explanatory variables because the model will investigate how they explain the outcome. Other terms used for explanatory variables are independent variables, regressor variables, predictor variables, and, in some situations, **covariates**. The term "explanatory variables" is somewhat more generic than the other alternatives, which may or may not be appropriate in different settings. For example, the use of the term "predictor variable" presumes that **prediction** is the goal of model building, and this may not be applicable. Use of the term "independent variable" is not recommended since the dependence between the response variable and explanatory variables is the goal of modeling and since independence among explanatory variables often cannot be assumed.

Various ways of classifying explanatory variables are described in Cox & Snell [1]. In particular, an explanatory variable can be quantitative or qualitative. If it is quantitative, then the way in which it is included in the model must be investigated. One question to consider is whether the explanatory variables enter the model in an **additive** or **multiplicative** manner. Another question is whether the original explanatory variable or some **transformation** of it should be used. A related question is whether polynomial terms should be considered.

If an explanatory variable is qualitative, then **dummy variables** must be defined to represent the explanatory variable. The choice of dummy variable definitions depends on the study situation. In a regression model, the significance of the relationship between the outcome and the qualitative explanatory variable can then be assessed by testing the hypothesis that all regression coefficients for the dummy variables defined from this explanatory variable are equal to 0.

Explanatory variables can be fixed by the design of the experiment, representing factors such as levels of treatment. Explanatory variables can also simply be measurements for an individual obtained without any experimental control, as is generally the case in observational studies.

Reference

[1] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London.

(*See also* **Collinearity**; **Linear Regression, Simple**; **Multiple Linear Regression**)

G.A. DARLINGTON

Exploratory Data Analysis

Statisticians, as well as others who apply statistical methods to data, have often made preliminary examinations of data in order to explore their behavior. In this sense, exploratory data analysis has long been a part of statistical practice. Since about 1970, "exploratory data analysis" has most often meant the attitude, approach, and techniques developed, primarily by John W. Tukey, for flexible probing of data.

Broad Phases of Data Analysis

One description of the general steps and operations that make up practical data analysis identifies two broad phases: an exploratory phase and a confirmatory phase. Exploratory data analysis is concerned with isolating patterns and features of the data and with revealing these forcefully to the analyst. It often provides the first contact with the data, preceding any firm choice of models for either structural or stochastic components (*see* **Model, Choice of**); but it also serves to uncover unexpected departures from familiar models (*see* **Diagnostics**). An important element of the exploratory approach is flexibility, both in tailoring the analysis to the structure of the data and in responding to patterns that successive steps of analysis uncover.

Confirmatory data analysis concentrates on assessing the reproducibility of the observed patterns or effects. Its role is closer to that of traditional statistical **inference** in providing statements of significance (*see* **Hypothesis Testing**) and **confidence;** but the confirmatory phase also encompasses (among others) the step of incorporating information from an analysis of another, closely related, body of data and the step of validating a result by collecting and analyzing new data.

In brief, exploratory data analysis emphasizes flexible searching for clues and evidence, whereas confirmatory data analysis stresses evaluating the available evidence. The rest of this article describes the basic concepts of exploratory data analysis and illustrates some simple techniques.

Four Themes

Throughout exploratory data analysis, four main themes appear and often combine. These are resistance, residuals, re-expression, and display.

Resistance

Resistance is a matter of insensitivity to misbehavior in data. More formally, an analysis or summary is *resistant* if an arbitrary change in any small part of the data produces only a small change in the analysis or summary. This attention to resistance reflects an understanding that "good" data seldom contain less than about 5% gross errors, and protection against the adverse effects of these should always be available.

It is worthwhile to distinguish between resistance and the related notion of **robustness**. Robustness generally implies insensitivity to departures from assumptions surrounding an underlying probabilistic model. (Some discussions regard resistance as one aspect of "qualitative robustness".)

In summarizing the location of a sample, the **median** is highly resistant. (In terms of **efficiency**, it is not so highly robust because other estimators achieve greater efficiency across a broader range of distributions.) By contrast, the **mean** is highly nonresistant. A number of exploratory techniques for more-structured forms of data provide resistance because they are based on the median.

Residuals

Residuals are what remain of the data after a summary or fitted model has been subtracted out, according to the schematic equation

$$\text{residual} = \text{data} - \text{fit}.$$

For example, if the data are the pairs (x_i, y_i) and the fit is the line $\hat{y}_i = a + bx_i$, then the residuals are $r_i = y_i - \hat{y}_i$.

Exploratory data analysis takes the attitude that an analysis of a set of data is not complete without a careful examination of the residuals. This emphasis builds on the tendency of resistant analyses to provide a clear separation between dominant behavior and unusual behavior in the data (*see* **Outliers**). When the bulk of the data follows a consistent pattern, that pattern determines a resistant fit. The residuals then contain any drastic departures from the pattern, as well as the customary chance fluctuations. Unusual residuals suggest a need to check on the circumstances surrounding those observations. As in more-traditional practice, the residuals can warn of systematic difficulties with the data – curvature, nonadditivity, and nonconstancy of variability (*see* **Scedasticity**).

Re-expression

Re-expression involves the question of what scale would help to simplify the analysis of the data (*see* **Measurement Scale**). Exploratory data analysis emphasizes the benefits of considering, at an early stage, whether the scale in which the data are originally expressed is satisfactory. If not, a re-expression into another scale may help to promote symmetry, constancy of variability, straightness of relationship, or additivity of effect, depending on the structure of the data.

The re-expressions most often used in exploratory data analysis come from the family of functions known as **power transformations**, which take y into y^p (almost always with a simple value of p such as $\frac{1}{2}, -1$, or 2), together with the logarithm (which, for data-analysis purposes, fits into the power family at $p = 0$). For example, one investigation of the relationship between gasoline mileage and the characteristics of automobiles gained substantially from re-expressing mileage in a reciprocal scale and working with gallons per 100 miles instead of miles per gallon.

Such changes of scale are not solely a specialized concern of data analysis; they arise in everyday experience. For example, in the Richter scale for intensity of earthquakes, the pH scale for acidity, and the average speeds in an auto race, the numbers reported have already been re-expressed from the scale in which the basic data were collected. Hoaglin [1] discusses these and other examples.

Displays

Displays meet the analyst's need to see behavior – of data, of fits, of diagnostic measures, and of residuals – and thus to grasp the unexpected features as well as the familiar regularities (*see* **Graphical Displays**).

A major contribution of the developments associated with exploratory data analysis has been the emphasis on visual displays and the variety of new graphical techniques. The example below includes one of these, the *schematic plot* or *boxplot*.

Example

An example illustrates the themes of exploratory data analysis, with particular emphasis on re-expression and display.

Table 1 Concentration of nicotine (nanograms per milliliter) in the urine of a group of nonsmokers exposed to a smoke-filled room (NS1), two groups of nonsmokers with customary exposure to smoke (NS2 and NS3), and a group of smokers (S)

NS1	NS2	NS3	S
13	0.8	0	104
33	0.8	0.2	109
35	1.2	0.2	128
45	3.3	2.0	312
45	3.5	2.2	375
61	6.2	4.2	802
92	6.3	5.2	833
93	8.0	9.0	937
98	8.6	12.0	1006
157	10.3	12.0	1049
208	11.3	19.3	1629
	21.0	23.5	1788
	28.6	26.0	1808
	64.3		1967
			1990
			2073
			2609
			2732

Source: Russell & Feyerabend [5].

In research that bears on the effects of passive smoking, Russell & Feyerabend [5] studied the effect on nonsmokers of exposure to tobacco smoke. For four groups of volunteers they measured the concentration of nicotine in each subject's urine during the early afternoon. The first group of nonsmokers (NS1) spent an average of 78 minutes seated among smokers in an unventilated smoke-filled room. The second and third groups of nonsmokers (NS2 and NS3) were not subject to any special conditions and simply received their usual exposure to smoke in the workplace. The fourth group, the smokers (S), was included for comparison. For the four groups Table 1 shows the urine nicotine levels (in nanograms per milliliter).

In this scale the data do not yield an effective display. A straightforward plot that includes the data from all the smokers would show almost no detail in the other three groups. This is one of the difficulties that re-expression aims to remedy. For these data, re-expression in a logarithmic scale leads to a much more effective analysis. The results are easy to interpret, because an additive difference in the log scale corresponds to a ratio in the original scale. For example, for logarithms to base 10, the medians

of NS1 and S are 1.8 and 3.0, respectively. The difference, 1.2, corresponds to a ratio of 16, similar to what one would get by comparing observations near the middle of the two groups.

To facilitate comparisons among the four groups, Figure 1 displays their boxplots. In this graphical summary of the data the box extends from the lower hinge (an approximate quartile; *see* **Quantiles**) to the upper hinge and has a line across it at the median. The dashed lines show the extent of the data, except for observations that are apparent strays (defined according to a rule of thumb based on the hinges). Such strays appear individually – in order to focus attention on them – as the lowest observation in NS3 does in Figure 1. The general intent is to indicate the median, outline the middle half of the data, and show the **range**, with more detail at the ends if needed.

Figure 1 shows that the passive smokers (NS1) have relatively little overlap with the smokers (S) above or with the nonsmokers (NS2 and NS3) below. As mentioned previously, the passive smokers' urine nicotine levels are typically about 1/16 of the smokers'. However, the difference of about 1 on the log scale between NS1 and either NS2 or NS3 means that the passive smoking produced about a tenfold increase in urine nicotine. These differences in level are the main story; and it is easy to focus on them because the variation within the four groups is quite similar, whether measured by the length of the boxes or the extent of the dashed lines. Some members of NS3 apparently were more successful in avoiding exposure to smoke, as indicated by the longer dashed line at the lower end of that boxplot and, especially, by the one stray observation. This subject had no detectable nicotine in her urine. As a mathematical function, the logarithm transforms 0 into $-\infty$, but this result does not affect the median or the lower hinge for the group.

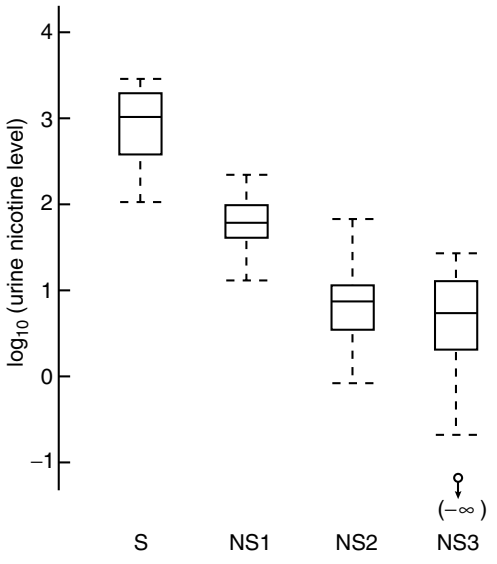

Figure 1 Parallel boxplots for the logarithm (base 10) of nicotine concentration in the urine of the smokers (S) and the three groups of nonsmokers (NS1, NS2, and NS3). The subjects in NS1 were exposed to the smoke-filled room

Literature

The first published presentation of exploratory data analysis was the preliminary edition of the book by John W. Tukey [6]; the 1977 edition [9] represents the definitive account of the subject. A paper by Tukey [8] describes and illustrates a number of the most important displays.

The book by Mosteller & Tukey [4] contains substantial discussions of exploratory attitudes and techniques.

The two volumes edited by Hoaglin et al. [2, 3] provide conceptual, logical, and mathematical support for a number of exploratory techniques. They also explain and illustrate the connections of those techniques to more-conventional techniques and to classical statistical theory.

Broader discussions of the roles of exploratory and confirmatory data analysis in scientific inquiry appear in Tukey [7, 10].

References

[1] Hoaglin, D.C. (1988). Transformations in everyday experience, *Chance* **1**, 40–45.

[2] Hoaglin, D.C., Mosteller, F. & Tukey, J.W., eds (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

[3] Hoaglin, D.C., Mosteller, F. & Tukey, J.W., eds (1985). *Exploring Data Tables, Trends, and Shapes*. Wiley, New York.

[4] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading.

[5] Russell, M.A.H. & Feyerabend, C. (1975). Blood and urinary nicotine in nonsmokers, *Lancet* **i**, 179–181.

[6] Tukey, J.W. (1970–1971). *Exploratory Data Analysis*, Limited Preliminary Ed. Addison-Wesley, Reading. (Available from University Microfilms.).

[7] Tukey, J.W. (1972). Data analysis, computation, and mathematics, *Quarterly of Applied Mathematics* **30**, 51–65.

[8] Tukey, J.W. (1972). Some graphic and semigraphic displays, in *Statistical Papers in Honor of George W. Snedecor*, T.A. Bancroft, ed. Iowa State University Press, Ames, pp. 293–316.

[9] Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.

[10] Tukey, J.W. (1980). We need both exploratory and confirmatory, *American Statistician* **34**, 23–25.

(*See also* **Graphical Presentation of Longitudinal Data**; **Multivariate Graphics**; **Transformations**)

DAVID C. HOAGLIN

Exponential Distribution

The simplest distribution used to model survival data is the exponential distribution (*see* **Parametric Models in Survival Analysis**). This model has survival function $S(t) = e^{-\lambda t}$ and density function $f(t) = \lambda e^{-\lambda t}$. The exponential distribution is characterized by a constant **hazard** rate, $h(t) = \lambda$. This constant hazard rate implies a lack of aging and leads to several other equivalent characterizations of the exponential distribution.

The first characterization is referred to as the lack of memory property. If T has an exponential distribution, then it follows that

$$\Pr(T \geq t + x | T \geq t) = \Pr(T \geq x).$$

This means that the chance that an individual of age t survives an additional x years is the same as a newborn surviving to age x. Because of this distributional property, it follows that $E(T - t | T \geq t) = E(T) = 1/\lambda$, i.e. the mean residual life is constant. Because the time until the future occurrence of an event does not depend upon past history, this property is called the "no-aging" property or the "old as good as new" property of the exponential.

The second characterization of the exponential distribution is as the distribution of the interarrival times of a **Poisson process**. If $N(t)$ is a Poisson process with intensity rate λ, then the times between successive occurrences of the process are independent exponential random variables with hazard function, λ. Also, the length of the interval from some fixed time to the next occurrence of the Poisson process has an exponential distribution with rate λ.

For an exponential random variable the mean is λ^{-1} and the variance is λ^{-2}, so the coefficient of variation is 1. The median time to the event is $(\log 2)/\lambda$. Based on independent, possibly right-**censored** survival times $X_1, \ldots, X_n$ from the exponential distribution with hazard λ, the **maximum likelihood** estimator for λ is the "occurrence/exposure rate"

$$\hat{\lambda} = \frac{\sum D_i}{\sum X_i},$$

where D_i is an indicator for failure of individual i. For large n, $\hat{\lambda}$ is approximately normally distributed with mean λ and a variance which may be estimated by $\sum D_i / (\sum X_i)^2$.

The exponential model may be checked using the **Nelson–Aalen estimator**, which, under the model, will approximate a straight line with slope λ. Also, **total time on test** (see [1]) techniques are designed especially for testing for exponentiality.

Regression analysis of exponentially distributed survival times was pioneered by Feigl & Zelen [2], who studied uncensored data, and generalized to right-censoring by Zippin & Armitage [3].

A more general form of the exponential distribution is the two-parameter exponential distribution with survival function

$$S(t) = \begin{cases} 1, & \text{if } t < \mu, \\ \exp[-\lambda(t - \mu)], & \text{if } t \geq \mu. \end{cases}$$

Here the parameter μ is a threshold or guarantee parameter. This distribution still has a constant hazard rate and the lack of memory property.

The exponential distribution has found limited use in biomedical applications since its lack of aging property is too restrictive for most problems. It is useful, however, as a special case of both the **Weibull distribution**, the **gamma distribution**, and the distribution with a piecewise constant hazard function (*see* **Grouped Survival Times**). Furthermore, the model is frequently used for **sample size determination** and in **simulation** studies.

References

[1] Barlow, R.E. & Campo, R. (1975). Total time on text processes and applications to failure time data analysis, in *Reliability Fault Tree Analysis*, R.E. Barlow, J.B. Fussell & N.D. Singpurwalla, eds. SIAM, Philadelphia, pp. 451–481.

[2] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[3] Zippin, C. & Armitage, P. (1966). Use of concomitant variables and incomplete survival information with estimation of an exponential survival parameter, *Biometrics* **22**, 655–672.

JOHN P. KLEIN, PER KRAGH ANDERSEN & NIELS KEIDING

Exponential Family

Many of the families of probability distributions arising in biostatistics are of exponential form, e.g. **normal**, **exponential**, **gamma**, **beta, binomial**, **multinomial**, **hypergeometric**, **Poisson**, etc. The class of models is theoretically important because in multiparameter settings exact frequency methods of **inference** exist only for certain exponential family settings and for another general class called transformation models, e.g. **location-scale** models. Much of the underpinnings of general theory of inference arise from consideration of exponential families. Moreover, adequate understanding of some of the most important aspects of biostatistics, e.g. matched case–control studies (*see* **Matching**), requires a grounding in theory of inference for exponential families.

Basic Definitions and Examples

For a collection of data $\mathbf{y}$, an exponential family is a class of distributions with densities (meant to include probability mass functions for discrete data) of the form

$$f(\mathbf{y};\boldsymbol{\theta}) = m(\mathbf{y}) \exp\left[\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) s_j(\mathbf{y}) - d(\boldsymbol{\theta})\right], \quad (1)$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is typically a vector parameter. The density is often zero outside some specified set, which must not depend on $\boldsymbol{\theta}$. Very often $\mathbf{y}$ consists of independent *but not identically* distributed observations $y_1, y_2, \ldots, y_n$, which themselves have exponential family distributions of simpler form. The primary aims involve comparison of the distributions of the y_i or studying their relation to **covariates** of interest. Most intermediate-level statistical theory texts focus on the case of identically distributed observations, but this fails to meet practical needs or to raise generally important issues discussed here. The texts by Cox & Hinkley [4], Cox & Snell [6], and Lehmann [12] give treatments closer to the following, with the latter giving mathematical details avoided here. Further mathematical treatment and some interesting general inferential aspects are given in Barndorff–Nielsen [1].

For example, basic modeling concepts of **least squares** regression can be extended to **likelihood** analysis where the component observations y_i have exponential family densities of the form

$$f(y_i;\omega) = a_i(y_i) \exp[\omega y_i - b_i(\omega)] \quad (2)$$

by modeling the parameter for y_i as $\omega_i = \mathbf{x}_i'\boldsymbol{\theta}$, where $\mathbf{x}_i$ is an associated vector of covariables. The joint density can be written in the form of (1), with $c_j(\boldsymbol{\theta}) = \theta_j$, the coordinates of $\boldsymbol{\theta}$, and $s_j(\mathbf{y}) = \sum_i x_{ij} y_i$. An important example is **logistic regression**, where the component observations y_i have **binomial** distributions with probabilities of success π_i which are to be modeled in terms of covariables. The densities can be expressed in the form of (2) as

$$f(y_i;\pi_i) = a_i(y_i) \exp\left\{\log\left[\frac{\pi_i}{(1-\pi_i)}\right] y_i - b_i(\pi_i)\right\},$$

and the logistic model relates the π_i to covariate vectors $\mathbf{x}_i$ in a manner such that $\log[\pi_i/(1-\pi_i)] = \mathbf{x}_i'\boldsymbol{\beta}$, where in more conventional notation we write $\boldsymbol{\beta}$ in place of $\boldsymbol{\theta}$.

Often a single parameter ω is not adequate for modeling the component distributions, and this is generalized to models with a vector parameter $\boldsymbol{\omega}$,

$$f(y_i;\boldsymbol{\omega}) = a_i(y_i) \exp\left[\sum_{j=1}^{l} \omega_j g_j(y_i) - b_i(\boldsymbol{\omega})\right], \quad (3)$$

also leading to the form of (1) for the joint density for some $k \geq l$. An instance of this with $l = 2$ arises in **gamma** regression with scale parameters σ_i and unknown common shape parameter ν, where the densities can be expressed in the form of (3) as

$$f(y_i;\sigma_i,\nu) = a_i(y_i) \exp\left[\left(\frac{-1}{\sigma_i}\right) y_i + \nu \ln(y_i) - b_i(\sigma_i,\nu)\right].$$

A regression model which is theoretically tractable, but not always the most practically useful, takes $1/\sigma_i = \mathbf{x}_i'\boldsymbol{\beta}$. Then, in (1), the $c_j(\boldsymbol{\theta})$ consist of the coordinates of $\boldsymbol{\beta}$ supplemented by ν, and the $s_j(\mathbf{y})$ consist of the quantities $-\sum_i x_{ij} y_i$ supplemented by $\sum_i \ln(y_i)$.

Both of the above examples are **generalized linear models** [10, 14]. Examples not of this form, in

particular where the data **y** do not consist of independent observations, include Gaussian mixed models such as **randomized block designs**, Gaussian **multivariate** and **time series** models, and **multinomial** or **hypergeometric** models.

The likelihood function based on (1) depends on the data only through the collection of statistics $[s_j(\mathbf{y}); j = 1, \ldots, k]$, and thus such a reduction of the data is a **sufficient statistic** – one of the key aspects of exponential families. Ordinarily there are various choices for how (1) is expressed for a given setting, and henceforth we assume that this is in terms of the smallest possible choice of k, which is called the *order* of the family. The collection $[s_j(\mathbf{y}), j = 1, \ldots, k]$ is then minimally sufficient. It is the classical Fisher–Darmois–Koopman–Pitman result that, in general, under regularity conditions, when data **y** consist of n independent observations, a sufficient statistic of fixed dimension k for all n can exist only when the component observations y_i have exponential family distributions as indicated by (3). Even then, when the y_i are not identically distributed, this occurs only under special conditions discussed below. It is assumed that the coordinates of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ can vary independently of one another, i.e. that the dimension of the space $\boldsymbol{\Theta}$ is really $\dim \boldsymbol{\theta}$. Since the distribution of **y** is determined by $[c_j(\boldsymbol{\theta}), j = 1, \ldots, k]$, unless $k \geq \dim \boldsymbol{\theta}$ the parameter $\boldsymbol{\theta}$ will contain redundancies resulting in lack of identifiability, as in one-way classifications in parameterization $\mu + \tau_j$. We assume here that these redundancies have been eliminated. The present discussion is not mathematically rigorous, but is intended to convey the essential ideas as simply as possible.

Canonical Parameters; Regular and Curved Families

It is useful to consider a reparameterization, taking $[\gamma_j = c_j(\boldsymbol{\theta}); j = 1, \ldots, k]$, and thus to express the model in (1) as

$$f(\mathbf{y}; \boldsymbol{\gamma}) = m(\mathbf{y}) \exp\left[\sum_j \gamma_j s_j(\mathbf{y}) - K(\boldsymbol{\gamma})\right].$$

The γ_j are called the *canonical parameters* of the exponential family, and the statistics $s_j(\mathbf{y})$ the *canonical sufficient statistics*. Note that these are only specified up to linear transformations; any full-rank linear transformation of the collection (γ_j) is another set of canonical parameters, with correspondingly transformed canonical sufficient statistics. The density of **s** takes similar form,

$$f(\mathbf{s}; \boldsymbol{\gamma}) = n(\mathbf{s}) \exp[\boldsymbol{\gamma}'\mathbf{s} - K(\boldsymbol{\gamma})], \tag{4}$$

for a function $n(\mathbf{s})$ which in the discrete case is the sum of $m(\mathbf{y})$ for **y** values mapping into **s**. Another useful parameterization utilized below is in terms of the vector $\boldsymbol{\mu} = \mathrm{E}_\gamma(\mathbf{s})$, which is referred to as the *mean parameter*.

The function $K(\boldsymbol{\gamma})$ is essentially the cumulant generating function (*see* **Characteristic Function**) of the statistic **s**. That is, derivatives of $K(\boldsymbol{\gamma})$ yield moments of **s** in a useful form: $\dot{K}(\boldsymbol{\gamma}) = \mathrm{E}_\gamma(\mathbf{S})$, $\ddot{K}(\boldsymbol{\gamma}) = \mathrm{var}_\gamma(\mathbf{S})$, $\dddot{K} = \mathrm{skew}_\gamma(\mathbf{S})$, and so forth. Here, and in the following, overdots on functions denote derivatives, with respect to parameters if there are additional arguments. As seen above, $\boldsymbol{s}$ is often the sum of independent observations, and in any case normal approximations for its distribution apply as the information in the data becomes large.

Recall that the order of the family is the minimal value of k which can be used in (1) – that is, the minimal dimension of $\boldsymbol{\gamma}$ in the above expressions. There are then two major cases: where $k = \dim \boldsymbol{\theta}$ and where $k > \dim \boldsymbol{\theta}$. If $k = \dim \boldsymbol{\theta}$, as in the two examples above, the family is said to be *regular*, which simplifies inferential issues. In particular, the maximum likelihood estimator of $\boldsymbol{\theta}$ is a one-to-one function of **s**, and hence is a sufficient statistic. In addition, **s** is complete; that is, there is at most one unbiased estimator based on **s** of any given parametric function $\psi(\boldsymbol{\theta})$ [12, Section 4.3]. If $k > \dim \boldsymbol{\theta}$, then the parameters $[\gamma_j(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}]$ are functionally related, jointly tracing out a curved space of dimension $\dim \boldsymbol{\theta}$ within k-dimensional Euclidean space $\mathcal{R}^k$, and thus the model is called a *curved* exponential family [7, 8]. The **maximum likelihood** estimator of $\boldsymbol{\theta}$ is then not a sufficient statistic, being of dimension less than that of the minimal sufficient statistic. Ideal inference is severely hampered by the need to utilize the information not contained in the maximum likelihood estimator of $\boldsymbol{\theta}$, and approximate methods, which may be very good, are always employed.

The above examples are readily modified to yield curved exponential families, by modeling quantities other than the canonical parameter as linear in $\boldsymbol{\beta}$. That is, in the binomial example, if $\pi_i = g(\mathbf{x}_i'\boldsymbol{\beta})$

for any function $g(\cdot)$ other than that corresponding to $\log[\pi_i/(1-\pi_i)] = \mathbf{x}_i'\boldsymbol{\beta}$, then the family is curved unless the covariates are very special, i.e. taking on only $\dim \boldsymbol{\beta}$ or fewer values. Without special covariate structure the minimal value of k is the sample size n. In the gamma example, if some function other than $1/\sigma_i$ is modeled as $\mathbf{x}_i'\boldsymbol{\beta}$, then the family is similarly curved. In that setting it is often more natural and practically useful to take $\ln \sigma_i = \mathbf{x}_i'\boldsymbol{\beta}$. Although these are practically important instances, curved exponential families arise in many other ways. An interesting and more tractable example arises in identically distributed negative exponential response times which are **censored** if they exceed a fixed time C. The minimal sufficient statistic is (T, r), where T is the total time on test and r is the number of failures; and the maximum likelihood estimator of the scale parameter is T/r. Another often-considered example arises from identically distributed Gaussian observations with known coefficient of variation.

Basic Issues of Estimation

For inferential purposes it will be best to consider inference about some underlying parameter $\boldsymbol{\theta}$, rather than $\boldsymbol{\gamma}$, with the understanding that in the regular case one may sometimes simply want to consider $\boldsymbol{\theta} = \boldsymbol{\gamma}$. The log likelihood function for $\boldsymbol{\theta}$ in terms of $\mathbf{s}$ is

$$l(\boldsymbol{\theta}; \mathbf{s}) = \boldsymbol{\gamma}(\boldsymbol{\theta})'\mathbf{s} - K[\boldsymbol{\gamma}(\boldsymbol{\theta})]. \tag{5}$$

For regular families where the coordinates of $\boldsymbol{\gamma}$ vary freely, the maximum likelihood equations are $\mathrm{E}_\theta(\mathbf{S}) = \mathbf{s}$; that is, the maximum likelihood estimator of the mean parameter $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \mathbf{s}$. This can be seen by reparameterizing in terms of $\boldsymbol{\gamma}$ and differentiating (5), using the fact that $\dot{K}(\boldsymbol{\gamma}) = \boldsymbol{\mu}$. For curved families $\boldsymbol{\mu}(\boldsymbol{\theta})$ does not vary freely over $\mathcal{R}^k$ and the equations $\hat{\boldsymbol{\mu}} = \mathbf{s}$ generally have no solution. In this case the maximum likelihood estimator $\hat{\boldsymbol{\mu}}$ is a weighted least squares approximate solution to these equations, as described below.

In either case the expected **information** for $\boldsymbol{\theta}$, i.e. $-\mathrm{E}[\ddot{l}(\boldsymbol{\theta}; \mathbf{S})] = \mathrm{var}[\dot{l}(\boldsymbol{\theta}, \mathbf{S})]$, is $i(\boldsymbol{\theta}) = (\partial\boldsymbol{\gamma}/\partial\boldsymbol{\theta})'\ddot{K}(\boldsymbol{\gamma})(\partial\boldsymbol{\gamma}/\partial\boldsymbol{\theta})$, or simply $\ddot{K}(\boldsymbol{\gamma})$ for regular families when $\boldsymbol{\theta} = \boldsymbol{\gamma}$. The observed information given data $\mathbf{s}$ is $j(\hat{\boldsymbol{\theta}}) = -\ddot{l}(\hat{\boldsymbol{\theta}}; \mathbf{s})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator. For regular exponential families, $i(\hat{\boldsymbol{\theta}}) = j(\hat{\boldsymbol{\theta}})$, and for curved exponential families these always differ except for samples with $\mathbf{s} = \hat{\boldsymbol{\mu}}$. In fact, the standard deviation of the ratio $j(\hat{\boldsymbol{\theta}})/i(\hat{\boldsymbol{\theta}})$ of these information measures can provide a useful measure of "how curved" the family is [7, 8]. When the information is substantial the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is approximately normally distributed with mean $\boldsymbol{\theta}$ and variance $[i(\boldsymbol{\theta})]^{-1}$. For curved families a dominant part of the information not contained in the maximum likelihood estimator can often be utilized by replacing this by $[j(\hat{\boldsymbol{\theta}})]^{-1}$ [9].

When for regular families the maximum likelihood equations cannot be solved in closed form, or when a least squares solution is called for because the model is curved, numerical methods can often be organized most simply in terms of iterative nonlinear weighted least squares (*see* **Generalized Linear Model**). Often this is for convenience based on a version of (5), where $k = \dim \mathbf{s}$ is not chosen minimally, and in particular may be where $\mathbf{s}$ there is taken as the original data $\mathbf{y}$. Departing momentarily from the assumption that k has necessarily been chosen minimally, the following holds for any choice of $k \geq \dim \boldsymbol{\theta}$, and for either regular or curved models. Let $\mathcal{M}$ represent the space of values of the mean parameter $[\mu(\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}]$, a space of dimension $p = \dim \boldsymbol{\theta}$ within $\mathcal{R}^k$, which is typically curved even for regular families if k is not minimal. Differentiating (5) and using the relation that $\dot{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \ddot{K}(\boldsymbol{\gamma})\dot{\boldsymbol{\gamma}}(\boldsymbol{\theta})$, it is seen that the vector $(\mathbf{s} - \hat{\boldsymbol{\mu}})$ is orthogonal to $\mathcal{M}$ at the point $\hat{\boldsymbol{\mu}}$ in the sense that

$$(\mathbf{s} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{T}} = \mathbf{0},$$

where $\hat{\boldsymbol{\Sigma}} = \mathrm{var}_{\hat{\theta}}\mathbf{S}$ and the columns of $\hat{\mathbf{T}}$ span the tangent space to $\mathcal{M}$ at $\hat{\boldsymbol{\mu}}$; for example, $\hat{\mathbf{T}} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$. This may be solved by iterative nonlinear least squares, with the step from an iterative value $\tilde{\boldsymbol{\mu}}$ given by

$$(\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}) = [\tilde{\mathbf{T}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{T}}]^{-1}\tilde{\mathbf{T}}'\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{s} - \tilde{\boldsymbol{\mu}}).$$

This is the Newton–Raphson method if the model is regular, and the Fisher scoring method if the family is curved (*see* **Optimization and Nonlinear Equations**).

Exact Frequency Inference

Exact **inference** for multiparameter exponential families is a fundamental aspect of theory of inference. Consider inference about a scalar-valued

parametric function $\psi = \psi(\boldsymbol{\theta})$ in terms of a significance test of $\psi = \psi_0$ against the alternative $\psi > \psi_0$ (*see* **Hypothesis Testing**). This can then be inverted to yield a lower *confidence limit* given by the smallest ψ-value which would not be rejected at the specified level as a hypothesized value. In our view confidence intervals are best determined by individually determined lower and upper confidence limits. It is useful in theoretical development to consider a reparameterization from $\boldsymbol{\theta}$ to $(\psi, \boldsymbol{\nu})$, where $\boldsymbol{\nu}$ is referred to as a **nuisance parameter**.

The fundamental condition imposed in deriving an inference about ψ is that operating characteristics under the hypothesis of the inferential procedure – that is the size of a test (*see* **Level of a Test**), the ***P* value**, or the coverage probability of a **confidence interval** – should be independent of $\boldsymbol{\nu}$. Exactness is achieved only when the model is regular and ψ is a canonical parameter, i.e. a linear function of any specific choice $\boldsymbol{\gamma}$ of canonical parameters. In this case $\boldsymbol{\nu}$ can be chosen as a complementary set of canonical parameters, and the density of a sufficient statistic can be expressed as

$$f(t, \mathbf{s}; \psi, \boldsymbol{\nu}) = n(t, \mathbf{s}) \exp[\psi t + \boldsymbol{\nu}'\mathbf{s} - K(\psi, \boldsymbol{\nu})]. \tag{6}$$

All inferences which are independent of $\boldsymbol{\nu}$ in the sense indicated above may be based on the conditional distribution of T given $\mathbf{S} = \mathbf{s}$; note that $\mathbf{s}$ is the complete sufficient statistic for $\boldsymbol{\nu}$ in the subfamily, where ψ is fixed at any value [12, Chapter 4; 4, Chapter 5] (*see* **Conditionality Principle**). This conditional distribution depends only on ψ and belongs to a related exponential family of the form

$$f(t|\mathbf{S} = \mathbf{s}; \psi) = n(t, \mathbf{s}) \exp[\psi t - K^*(\psi; \mathbf{s})] \tag{7}$$

for a function $K^*(\psi; \mathbf{s})$, which is, of course, the cumulant generating function of the conditional distribution. The P value for testing $\psi = \psi_0$ against the alternative $\psi > \psi_0$ is $\Pr[T \geq t_{\text{obs}}|\mathbf{S} = \mathbf{s}_{\text{obs}}; \psi_0]$, where $t_{\text{obs}}, \mathbf{s}_{\text{obs}}$ are the observed data.

The conditional distribution involved in this is ordinarily fairly intractable. However, for a variety of settings involving continuous data, in particular Gaussian and gamma models, this P value can be reexpressed in terms of an unconditional probability. The standard t tests for means and regression coefficients for Gaussian data, the standard F *tests* for variances of Gaussian data, and F tests for negative exponential and gamma data can be derived in this manner [12, Chapter 5].

For discrete data this conditional P value can often be computed by enumeration of the sample space for either t or the original data $\mathbf{y}$, within the set where $\mathbf{S} = \mathbf{s}_{\text{obs}}$. Aside from possibly extensive calculations, this is straightforward since the conditional density is simply proportional to $n(t, \mathbf{s}) \exp(\psi_0 t)$, and if necessary the first term can be computed as indicated following (4) by enumerating samples $\mathbf{y}$ mapping into $(t, \mathbf{s})$. In particular, these calculations do not require the function $K^*(\psi; \mathbf{s})$, which may be difficult to calculate. This method leads, for example, to **Fisher's exact test** for comparison of two binomial samples [4, Section 5.2; 6, Section 2.3]. Much of the methodology for matched case–control studies (*see* **Matched Analysis**) is based on this theory. A simple version of this pertains to independent pairs of binomial observations y_{1i}, y_{2i} with probabilities π_{1i}, π_{2i} modeled as $\log[\pi_{1i}/(1 - \pi_{1i})] = \nu_i$ and $\log[\pi_{2i}/(1 - \pi_{2i})] = \nu_i + \psi$. Inference about the common log **odds ratio** ψ is made by conditioning on the sufficient statistics $(y_{\cdot i})$ for the collection of nuisance parameters (ν_i) [6, Section 2.4; 3, Chapter 7].

Sophisticated enumeration **algorithms** have been developed, providing for feasible computation of P values for various settings involving single and multiple **contingency tables**, logistic regression, and related problems (for example, see LogXact [13]; *see* **Exact Inference for Categorical Data**; **StatXact**). However, these methods for discrete data are, in general, an application of theory which applies exactly only for continuous data, and when the discreteness of the distribution of $T|\mathbf{S} = \mathbf{s}$ is substantial the operating characteristics of the procedures are often quite poor. That is, the unconditional size of tests, the unconditional distribution of P values, and even the conditional level of confidence intervals, can be far from the nominal levels (see, for example, [21] and [11]). In our view [16], it is generally best to "smooth the data" by using, without continuity corrections, the best of the asymptotic methods discussed below.

The above theory can also be applied indirectly when $\psi(\boldsymbol{\theta})$ is the ratio of canonical parameters, and this need arises frequently – for example, in testing a Gaussian mean. If $\psi = \gamma_i/\gamma_j$, then testing $\psi = \psi_0$ is equivalent to testing $\gamma_i - \psi_0\gamma_j = 0$, which reduces the problem to a hypothesis for a canonical

parameter. Of course, if ψ is a monotonic function of a canonical parameter, then the theory applies simply by transforming the hypotheses to one involving the canonical parameter. These extensions exhaust the situations where there is even in principle an exact inference for a single parametric function in multiparameter exponential families. Extensions of the exact theory to settings where $\dim \boldsymbol{\psi} > 1$ are rather limited, consisting mainly of the standard F tests for Gaussian regression. There is virtually never an exact inferential method for curved exponential families.

Approximate Methods of Inference

There are several approximate methods of inference, whose definition and most basic properties are not really special to the exponential family setting [4]. However, this setting provides a good opportunity to compare and evaluate them, since they can be referred to exact methods. We continue to write $\psi = \psi(\boldsymbol{\theta})$ as the interest parameter, but for the methods of this section it is not required that this be a canonical parameter. It is assumed in the following that ψ-values of special interest, either maximum likelihood estimators or hypothesized values, are not on the boundary of the parameter space.

The direct use of approximate normality of the maximum likelihood estimator $\hat{\psi}$, often called the Wald method (*see* **Likelihood**), utilizes an approximate standard error computed from $[i(\hat{\boldsymbol{\theta}})]^{-1}$ or $[j(\hat{\boldsymbol{\theta}})]^{-1}$ by using the **delta method** (if necessary). Some caution is required in using this method. If $\rho = g(\psi)$ is a monotonic function of ψ, then ideally most inferences should be essentially independent of the choice of which parameter is used; this is called invariance. For example, a confidence interval for ρ should consist of the mapping of a confidence interval for ψ under the transformation $g(\psi)$. This does not obtain when using the Wald method, and of course the normal approximation will be better for some representations of the parameter than for others. When the information is modest, one should seldom rely on the Wald method without some consideration of choosing a suitable parameterization. The advantage of the method is the simplicity and transparency of reporting a point estimate and its approximate standard error.

The following alternative methods are presented in terms of significance tests for hypothesized values of ψ, which represent a rather incomplete inference without attention to estimation. However, confidence intervals for ψ may serve to rectify this even better than maximum likelihood estimators. These can be obtained from the following approximate methods, in the same manner as indicated above for exact tests, as the set of ψ-values which would not be "rejected" as hypothesized values.

The score test method (*see* **Likelihood**), sometimes called Rao's method, has useful characteristics. Writing (5) in terms of ψ and an arbitrary choice of nuisance parameter $\boldsymbol{\nu}$ as

$$l(\psi, \boldsymbol{\nu}; \mathbf{s}) = \boldsymbol{\gamma}(\psi, \boldsymbol{\nu})'\mathbf{s} - K[\boldsymbol{\gamma}(\psi, \boldsymbol{\nu})], \qquad (8)$$

the score test of $\psi = \psi_0$ is based on the derivative of (8) with respect to ψ, evaluated at ψ_0. Large values of this, in absolute value, indicate evidence against the hypothesis, without the need to compute $\hat{\psi}$ where this derivative is zero. None of the methods in this section depends on the particular choice of $\boldsymbol{\nu}$. In contrast to the Wald method, both the score test and the likelihood ratio method discussed below are invariant under monotonic reparameterizations $\rho = g(\psi)$. For the score test the derivative presented below would then simply be multiplied by the constant $1/\dot{g}(\psi_0)$, which has no effect on the ultimate test.

More precisely, the derivative of (8) is evaluated at ψ_0 and $\hat{\boldsymbol{\nu}}_0$, where the latter is the maximum likelihood estimator of $\boldsymbol{\nu}$ under the hypothesis. When ψ is a canonical parameter this is given by $t - \mathrm{E}_{\psi_0, \hat{\nu}_0}(T)$ in terms of the notation for (6). This special case gives a clear view of the fundamental nature of the score statistic, whose more general form is given by

$$[\partial \boldsymbol{\gamma}(\psi_0, \hat{\boldsymbol{\nu}}_0)/\partial \psi]'[\mathbf{s} - \mathrm{E}_{\psi_0, \hat{\nu}_0}(\mathbf{S})].$$

The expectation of this statistic under the hypothesis is approximately zero, and its approximate variance is the adjusted information for ψ allowing for estimation of $\boldsymbol{\nu}$. This adjusted information is the reciprocal of the $\mathrm{var}(\hat{\psi})$ element in the inverse information matrix $[i(\psi_0, \hat{\boldsymbol{\nu}}_0)]^{-1}$. The score test statistic consists of the log likelihood derivative divided by its asymptotic standard deviation, with the P value for alternatives $\psi > \psi_0$ being the probability that a standard normal variate is as large as the observed value of this statistic. The most important characteristic of the score test is that one needs only to compute the maximum likelihood estimator under the

hypothesis, $\hat{\boldsymbol{\nu}}_0$, rather than the full maximum likelihood estimator $(\hat{\psi}, \hat{\boldsymbol{\nu}})$, which may simplify computations substantially. For example, consider the logistic regression example above when x_i consists of a constant term and a single covariable, and testing whether the regression coefficient of the covariable is zero. Writing m_i for the binomial sample sizes, the numerator of the score test is simply $\sum x_i(y_i - \hat{\pi})$, where $\hat{\pi} = \sum y_i / \sum m_i$; its asymptotic variance is given by $\hat{\pi}(1-\hat{\pi}) \sum m_i (x_i - \overline{x}_w)^2$, where $\overline{x}_w$ is the average of the x_i using weights m_i.

Generally, the most reliable inference is based on the asymptotic distribution of the generalized likelihood ratio (*see* **Likelihood Ratio Tests**), which is sometimes referred to as Wilks' method. Denoting the likelihood function by $L(\boldsymbol{\theta}; \mathbf{y})$, this likelihood ratio is defined as

$$W(\psi_0, \mathbf{y}) = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{y}) / \max_{\boldsymbol{\theta}:\psi(\boldsymbol{\theta})=\psi_0} L(\boldsymbol{\theta}; \mathbf{y}),$$

which will tend to take large values when the data do not support the hypothesis. In terms of a reparameterization to $(\psi, \boldsymbol{\nu})$ as above, computing the constrained maximum of the denominator amounts to fitting the nuisance parameter $\boldsymbol{\nu}$ by maximum likelihood when $\psi = \psi_0$. When the hypothesis is true, the sampling distribution of $2\ln[W(\psi_0; \mathbf{Y})]$ is approximately **chi-square** on 1 **degree of freedom** (df). Moreover, the distribution of

$$z(\psi_0, \mathbf{Y}) = \text{sign}(\hat{\psi} - \psi_0)\{2\ln[W(\psi_0; \mathbf{Y})]\}^{1/2} \quad (9)$$

is approximately **standard normal**, and can be used for directional inference. A P value for alternatives $\psi > \psi_0$ is taken as the chance that a standard normal variate is as large as $z(\psi_0, \mathbf{y}_{\text{obs}})$. This normal approximation is the best of the three discussed in this section, and is ordinarily adequate in practice unless $\dim \boldsymbol{\nu}$ is moderately large in relation to the available information. Improvements for that case are discussed in the following section.

Inversion of tests based on (9) to obtain confidence limits is often best carried out in terms of the **profile likelihood** function

$$L_p(\psi; \mathbf{y}) \propto \max_{\boldsymbol{\theta}:\psi(\boldsymbol{\theta})=\psi} L(\boldsymbol{\theta}; \mathbf{y}).$$

In practice one typically chooses a suitable grid of ψ-values and computes $L_p(\psi; \mathbf{y})$ by numerical methods or otherwise for each value in the grid. A plot of the result is useful, and a confidence interval corresponding to the likelihood ratio test can be taken as those ψ-values for which the ratio of $L_p(\psi; \mathbf{y})$ to its maximum value is at least a threshold value. For a $100\alpha\%$ level confidence interval this threshold value is taken as $\exp(-\frac{1}{2}\chi^2_{1-\alpha})$. Although this may be a computationally intensive method, with modern computing capabilities it is quite feasible once one organizes the calculations, or uses software with such procedures incorporated. It is noteworthy that the information computed from differentiation of the log profile likelihood is the same as the adjusted information for ψ as defined earlier.

All of these approximate methods generalize naturally to the case that $\dim \psi > 1$ in terms of chi-square tests. For tests based directly on the maximum likelihood estimator, and for score tests, a quadratic form is computed using the multivariate statistic and its asymptotic variance matrix. For the likelihood ratio test, the approximate distribution of $2\ln[W(\psi_0; \mathbf{Y})]$ is chi-square on $\dim \psi$ df.

Improved Approximations Based on Higher-Order Asymptotics

For regular families more accurate approximations to the distribution of $\mathbf{s}$ may be obtained from so-called saddlepoint approximations, which might better be referred to as likelihood ratio approximations. These are readily computed from $K(\boldsymbol{\gamma})$ and the maximum likelihood estimator $\hat{\boldsymbol{\gamma}}$ corresponding to the value of $\mathbf{s}$ at which the approximation is desired. In terms of the density of $\mathbf{s}$ this is discussed by Barndorff–Nielsen & Cox [2], and approximations to the distribution function are reviewed by Pierce & Peters [15]. These approximations are both very accurate and theoretically intriguing due to their connections with likelihood ratio concepts. The text by Severini [17] gives a thorough treatment of the following, in the general setting not restricted to exponential families.

Improvements on the above classical approximations, based on these methods, are relatively simple to implement for regular exponential families, and are particularly useful when $\dim \boldsymbol{\nu}$ is moderately large in relation to the information available. The Cox–Reid [5] adjusted profile likelihood function

approximates more closely the conditional likelihood function based on (7), and when ψ and $\boldsymbol{\nu}$ are canonical parameters it takes the form

$$L_{\mathrm{ap}}(\psi;\mathbf{y}) \propto L_{\mathrm{p}}(\psi;\mathbf{y})|\partial^2 K(\psi, \hat{\boldsymbol{\nu}}_\psi)/(\partial \boldsymbol{\nu})^2|^{1/2},$$

where $\hat{\boldsymbol{\nu}}_\psi$ denotes the maximum likelihood estimator when ψ is held fixed (see also Pierce & Peters [15]). This also applies without modification when $\dim \psi > 1$. With some modification it applies to noncanonical parameters and also outside of exponential family settings. The matrix whose determinant is involved is the expected information for $\boldsymbol{\nu}$ when ψ is fixed, and is available when one uses the iterative method indicated above for computing $\hat{\boldsymbol{\nu}}_\psi$. The maximizing value of L_{ap}, and confidence intervals computed from it as indicated above in terms of L_{p}, have reduced bias when $\dim \boldsymbol{\nu}$ is moderately large.

For inference about canonical parameters when $\dim \psi = 1$ it is possible through these approximations to modify the likelihood ratio statistic, (9), to obtain an improved estimate of $\Pr(T \geq t_{\mathrm{obs}}|\mathbf{S} = \mathbf{s}_{\mathrm{obs}}; \psi_0)$. This makes more accurate allowance both for the conditioning which eliminates the nuisance parameter, and for nonnormality of the conditional distribution involved [15]. This modification is simple to compute, and again the quantities required are ordinarily byproducts of the fitting process required for computing (9).

Bayesian Inference

Bayesian inference is also particularly tractable for exponential families. Consider the form (4) of an exponential family, and a family of prior distributions for $\boldsymbol{\gamma}$, with parameters $\boldsymbol{\lambda}$ and κ, of the form

$$\pi(\boldsymbol{\gamma};\boldsymbol{\lambda},\kappa) \propto \exp\{\kappa[\boldsymbol{\gamma}'\boldsymbol{\lambda} - K(\boldsymbol{\gamma})]\}.$$

This is referred to as a *conjugate family* of **prior distributions**, attractive in that the posterior distribution based on data with sufficient statistic $\boldsymbol{s}$ following (4) is

$$\pi(\boldsymbol{\gamma}|\mathbf{s};\boldsymbol{\lambda},\kappa) \propto \exp\left\{(\kappa+1)\left[\boldsymbol{\gamma}'\frac{(\mathbf{s}+\kappa\boldsymbol{\lambda})}{(\kappa+1)} - K(\boldsymbol{\gamma})\right]\right\}$$
$$\propto \pi(\boldsymbol{\gamma};\boldsymbol{\lambda}^*,\kappa^*),$$

which is an updated member of the conjugate family with parameters $\boldsymbol{\lambda}^* = (\mathbf{s}+\kappa\boldsymbol{\lambda})/(\kappa+1)$ and $\kappa^* = \kappa + 1$. Not only is this highly tractable but it provides an interpretation of the prior information specified by $(\boldsymbol{\lambda}, \kappa)$ as equivalent to hypothetical data with sufficient statistic $\boldsymbol{\lambda}$ but having weight κ times that of $\mathbf{s}$, combined with an initially locally uniform prior distribution. Asymptotic methods for Bayesian inference are closely related to those discussed above, both in terms of the classical results and the higher-order asymptotics [18–20].

References

[1] Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

[2] Barndorff-Nielsen, O.E. & Cox, D.R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion), *Journal of the Royal Statistical Society, Series B* **41**, 279–312.

[3] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. 1. *The Analysis of Case–Control Studies*. International Agency for Research on Cancer, Lyon.

[4] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[5] Cox, D.R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion), *Journal of the Royal Statistical Society, Series B* **49**, 1–39.

[6] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London.

[7] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second-order efficiency), *Annals of Statistics* **3**, 1189–1217.

[8] Efron, B. (1978). The geometry of exponential families, *Annals of Statistics* **6**, 362–376.

[9] Efron, B. & Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator, *Biometrika* **65**, 457–481.

[10] Firth, D. (1991). Generalized linear models, in *Statistical Theory and Modelling*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, Chapter 3.

[11] Haviland, M.G. (1990). Yates' correction for continuity and the analysis of 2×2 contingency tables, *Statistics in Medicine* **9**, 363–367.

[12] Lehmann, E. (1986). *Testing Statistical Hypotheses*, 2nd Ed. Wiley, New York.

[13] LogXact (1992). *A Software Package for Exact and Asymptotic Logistic Regression, Version 1.0*. Cytel Software, Cambridge, Mass.

[14] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall. London.

[15] Pierce, D.A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 701–737.

[16] Pierce, D.A. and Peters, D. (1999). Improving on exact tests by approximate conditioning, *Biometrika* **86**, 265–280.

[17] Severini, T.A. (2000). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.

[18] Sweeting, T.J. (1995). A framework for Bayesian and likelihood approximations in statistics, *Biometrika* **82**, 1–24.

[19] Sweeting, T.J. (1995). A Bayesian approach to approximate conditional inference, *Biometrika* **82**, 25–36.

[20] Tierney, L. & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**, 82–86.

[21] Upton, G.J.G. (1982). A comparison of alternative tests for the 2×2 comparative trial, *Journal of the Royal Statistical Society, Series A* **145**, 86–105.

(*See also* **Large-sample Theory**)

DONALD A. PIERCE

Exposure Effect

An exposure effect is a quantitative measure of the impact of exposure on an outcome measure. Estimates of exposure effects are derived by contrasting the outcomes in an exposed population with outcomes in an unexposed population or in a population with a different level of exposure. **Relative risk, excess risk**, and **relative odds** are examples of exposure effects used in connection with dichotomous outcomes, whereas **relative hazards** are used to characterize exposure effects when the outcome is time-to-response or the event rate per **person–year** exposure time. **Mean** differences are often used to characterize exposure effects for quantitative outcomes.

When **regression** models are used, exposure effects correspond to model parameters. For example, consider a model with **multiple linear regression** $\beta_0 + \beta_1 E + \beta_2 X$, where E indicates a level of exposure and X some other factors influencing outcome. The exposure effect, β_1, measures the effect on outcome of a unit increase in exposure, with other factors, X, held constant. If E only takes on values 1 for exposed and 0 for unexposed, β_1 is the adjusted mean difference between exposed and unexposed; and if, in addition, the outcome is dichotomous, β_1 is the adjusted risk difference. If the effect of exposure depends on levels of another factor, say X_1, an **interaction** term involving $X_1 E$ is needed in the previous regression. Then no single number characterizes the exposure effect, and **effect modification** is said to occur.

In the theory of **causation**, each individual or study unit is hypothesized to have two responses, the response if exposed and the response if unexposed. Only one such response is observed on each individual, but hypothesized individual-level effects can be defined, such as the difference in responses the individual would have if exposed and if unexposed. In the context of this "counterfactual" theory of causal effects, the exposure effects described in the previous paragraph can be regarded as summary measures of individual-level causal exposure effects.

(*See also* **Cox Regression Model**; **Generalized Linear Model**; **Logistic Regression**; **Poisson Regression**; **Relative Risk Modeling**)

MITCHELL H. GAIL

Extrapolation, Low Dose

At high doses, most substances are toxic to humans. Thus, it is imperative to establish conditions for the use of toxic substances that are relatively safe. Because of its devastating effects, cancer is a major concern. Considerable research is focused on identifying chemical substances (carcinogens) that cause cancer. Occasionally, it is possible to identify carcinogens in human populations by epidemiologic studies. Since humans are exposed to many substances, generally at unknown dose levels, it is difficult to quantify the risk (probability) of cancer from human studies. Cancer risks are generally estimated from **bioassays** conducted in laboratory animals (*see* **Tumor Incidence Experiments**). Animals are generally exposed at a few dose levels, and unexposed control animals are used, for a lifetime (from weaning up to two years in rodents). Chemicals are administered in the diet, drinking water, subcutaneously, via gavage, or by inhalation. Usually fewer than 100 animals are used per dose level. Thus, high doses must be used to elicit potential cancer effects at incidence rates high enough to be detected and measured in animal bioassays. However, humans are usually exposed to much lower levels of toxic substances. This requires extrapolation (estimation) of cancer incidence (risk) at doses well below the experimental dose range.

The dose of interest is the active dose at the target tissue. Many chemicals are metabolized by the body to an active or inactive state. Physiologically based **pharmacokinetic** studies are often conducted to study the absorption, distribution, and excretion of chemicals in the body. In some cases it is possible to measure the dose of the active chemical at the target tissue site. Generally, this information is not available and it is assumed that the target tissue dose is proportional to the administered dose. The nonlinearity of dose–response curves can be due, in part or wholly, to the pharmacokinetics of a chemical (*see* **Dose–Response Models in Risk Analysis**). In general, it is not known whether a test species is more or less sensitive than humans to a carcinogen. In the absence of such information, the US Environmental Protection Agency [22] proposes equal sensitivity across species when dose is expressed as a ratio of body weight to the 3/4 power. This is nearly equivalent to expressing dose as concentration in the diet or drinking water.

Risk is generally expressed as the probability that an individual will develop cancer by some age, usually lifetime, when exposed to a specified dose of a carcinogen (*see* **Risk Assessment**). For a population of individuals exposed to different doses, the expected number of cancer cases is

$$\text{expected number} = \sum_i N_i \Pr(d_i),$$

where N_i is the number of individuals exposed at dose d_i and $\Pr(d_i)$ is the probability (proportion) developing cancer at that dose. The average risk for the population is

$$\Pr = \frac{\sum_i N_i \Pr(d_i)}{\sum N_i}.$$

If the dose–response is linear, then the population risk is simply the slope (risk per unit dose) times the average dose level in the population.

One of the earliest procedures for low-dose cancer risk estimation was proposed by Mantel & Bryan [19]. They noted for many chemicals that the doses that produced cancer appeared to be **lognormally** distributed. That is, plots of tumor incidence as probits vs. log dose were approximately linear (*see* **Quantal Response Models**). Mantel & Bryan [19] observed that the probits versus log dose slopes were steep and generally much greater than one. They proposed extrapolating from the low end of the observable experimental dose–response to lower doses using a conservative shallow slope of one. Mantel & Bryan [19] postulated that this procedure should overestimate the true cancer incidence at low doses and provide conservative overestimates at doses with low levels of risk.

Theoretically, one molecule of a genotoxic carcinogen can interact with the DNA of a cell and initiate a carcinogenic process. In such cases there is no threshold dose below which cancer does not occur; that is, there is a positive slope as the dose approaches zero. Furthermore, if the chemical under study augments an existing carcinogenic process that is already producing tumors, then no threshold dose for the added chemical exists, as the endogenous dose already exceeds the threshold dose, if one exists. Hence, there is an increase (positive slope) for tumor

incidence at low doses of the administered chemical with additivity to the background carcinogenic process [6, 20].

Since the slope of the log probit procedure of Mantel & Bryan [19] approaches zero at low doses, that procedure is not compatible with the nonthreshold, low-dose positive slope expected for genotoxic chemicals or with additivity to background. For any dose–response that is curving upward in the low-dose region, linear extrapolation to zero excess risk at zero dose from the low end of the dose–response curve provides an overestimate of the risk at low doses. Since the shape of the dose–response curve cannot be determined at low incidence rates with the numbers of animals typically used, Gaylor & Kodell [10] proposed linear extrapolation to zero from an upper **confidence** limit on the estimate of risk at the lowest experimental dose, to provide a conservative procedure for low-dose extrapolation. Since the accuracy of the estimate at the lowest experimental dose could be poor, Farmer et al. [7] modified the procedure to extrapolation from a dose with a minimum excess risk of 1% or the lowest experimental dose, whichever was larger. The choice of the form of the dose–response model generally makes little difference above this point, as long as an adequate fit to the data is obtained.

Where a chemical operates independent of any background carcinogenic process, the total risk can be expressed as

$$\overset{*}{\Pr}(d) = \Pr(0) + [1 - \Pr(0)]\Pr(d),$$

where the excess risk due to the chemical of interest is

$$\Pr(d) = \frac{\Pr^*(d) - \Pr(0)}{1 - \Pr(0)}.$$

This is also known as Abbott's correction (*see* **Quantal Response Models**). In general, the total risk can be expressed as

$$\overset{*}{\Pr}(d) = \Pr(0) + \Pr(d),$$

where the excess risk due to the chemical under study is

$$\Pr(d) = \overset{*}{\Pr}(d) - \Pr(0).$$

The **multistage carcinogenesis model** [2] is used widely to describe cancer risk by a specific age as a function of dose:

$$\overset{*}{\Pr}(d) = 1 - \exp\left[-\prod_{i=0}^{k}(\beta_{0i} + \beta_{1i}d)\right],$$

where $\beta_{0i} > 0$ is the spontaneous rate for the ith stage and $\beta_{1i}d \geq 0$ is the increase in the rate of the ith stage due to dose d. For fitting dose–response data, the model is written in a generalized form [6]:

$$\overset{*}{\Pr}(d) = 1 - \exp\left[-\sum_{i=0}^{k}(\beta_i d^i)\right],$$

where $\beta_i \geq 0$. For small d, the risk is approximately

$$\overset{*}{\Pr}(d) = 1 - \exp[-(\beta_0 + \beta_1 d)],$$

which is approximately linear at low doses:

$$\overset{*}{\Pr}(d) = \beta_0 + \beta_1 d.$$

Since low-dose linearity cannot be excluded, the upper confidence limit for the generalized multistage model becomes linear at low doses. An upper limit of excess risk at low doses is provided by

$$\Pr(d) = q_1^* d,$$

where q_1^* is an upper limit on the estimate of β_1 also referred to as the low-dose slope (risk per unit dose) or carcinogenic potency. The above process has been used widely by US regulatory agencies to provide conservative overestimates of cancer risks at low doses. Conversely, doses can be estimated that correspond to specified levels of allowable risk; for example, risks of less than one in 100 000.

In addition to upper limit estimates of risk, a best (central) estimate can be obtained over the range where the dose–response is linear. **Maximum likelihood** estimates of the linear term in the generalized multistage model are quite unstable with sample sizes commonly used, and a discontinuity occurs where the estimate changes from zero to a positive value. Gaylor et al. [14] provide a stable estimate of the low-dose slope for the linear dose–response region

$$\hat{\Pr}(d) = \hat{\beta}_1 d,$$

where $\hat{\beta}_1 = 0.01/\mathrm{ED}_{01}$ and ED_{01} is the estimate of the dose producing an excess tumor incidence of 1%.

An upper limit can be estimated by

$$\mathrm{UL}(\hat{\beta}_1) = \frac{0.01}{\mathrm{LED}_{01}},$$

where LED_{01} is a lower confidence limit on a dose producing an excess risk of 1%.

Krewski et al. [17] proposed a model-free approach to low-dose extrapolation without making an assumption other than low-dose linearity. For each of the dose groups an upper confidence limit (L_i) on the tumor incidence is calculated. A lower confidence limit (L_0) is calculated for the tumor incidence of the control group. An upper limit for the low-dose slope for the ith dose group is

$$\hat{\beta}_i = \frac{(U_i - L_0)}{d_i},$$

where d_i is the dose level. The minimum of these values is then used for low-dose excess risk of cancer:

$$\hat{\mathrm{Pr}}(d) = (\min \hat{\beta}_i)d.$$

Because the minimum of k slopes is selected, the individual confidence limits must be adjusted so that the overall confidence level is maintained. Using the **Bonferroni inequality**, the individual confidence limits are set at the $\{1 - [0.05/(k+1)]\} \times 100\%$ level to maintain an overall confidence level of 95%. In general, upper limits of estimates of low-dose risk based on this model-free approach and the multistage model are comparable. For convex dose–response curves with low background tumor rates, upper limits of risk estimates based on the model-free procedure are typically twofold or more higher than those based on the multistage model.

One measure of carcinogenic potency is the TD_{50} (the daily dose that causes a tumor in 50% of the exposed animals that otherwise would not develop that type of tumor in a standard lifetime, generally two years for rodents). Variation in the TD_{50} appears to be approximately described by a lognormal distribution. Gaylor et al. [12] show that, for near-replicate bioassays, approximately 95% of the TD_{50}s are within a factor of four of their mean. Among strains within species and among species, approximately 95% of the TD_{50}s are within a factor of their means of 11 and 32, respectively. For a select group of 20 chemicals that have been shown to be carcinogenic in both humans and animals, the overall variability in the TD_{25}s is about a factor of 110 [1]. Based on these 20 chemicals, about 2/3 of the time the potency of human carcinogens is within a factor of ten of the potency in animals at doses high enough to produce a measurable incidence of cancer. This does not include the uncertainty in the shape of the dose–response relationship at lower doses.

The maximum tolerable dose (MTD) is used as the highest dose for chronic bioassays conducted by the US National Toxicology Program. The purpose of using the MTD is to maximize the probability of detecting carcinogenicity. The optimal experimental design for estimating cancer potency depends on the shape of the dose–response relationship. For typical shapes of dose–response curves and the limited number of animals generally used, Portier & Hoel [21] and Gaylor et al. [13] concluded that a good all-purpose design for both testing and estimation is to allocate equal numbers of animals at the MTD, MTD/2 and MTD/4 or MTD, MTD/3 and MTD/9, plus unexposed control animals.

Human exposure to carcinogens often is via exposure to mixtures of carcinogens. At low doses well below saturation of physiologic or metabolic processes, the total risk may be estimated by the sum of the risks of the individual components. In such cases an upper limit of the total risk can be approximated by $U = (\sum U_i^2)^{1/2}$ where U_i is the upper limit for the ith component [8]. For the two-stage clonal expansion model of carcinogenesis (the Moolgavkar–Venson–Knudson model), the age-specific **relative risk** is **multiplicative** for simultaneous exposure to an initiator and a completer or to an initiator and a promoter. At low levels of risk, additive and multiplicative risks of mixtures are nearly identical. Kodell et al. [16] show that simultaneous exposure to two promoters results in supra-multiplicative relative risks.

In most bioassays, animals are exposed to chemicals for a lifetime (generally two years in rodents). Exposures of humans to carcinogens often are intermittent. Based on the multistage model of carcinogenesis, Kodell et al. [15] show that using the average daily lifetime dose to estimate cancer risk is not likely to underestimate the risk by more than a factor of six. Using the **two-mutation carcinogenesis model** (the Moolgavkar–Venson–Knudson model), utilizing the average daily lifetime dose is not likely to underestimate the risk by more than a factor of ten [4].

Several researchers have noted a strong correlation between the MTD and cancer potency. This is due in

part to the limited range of doses used in bioassays and the limited range of tumor incidence observable for animal carcinogens with relatively small numbers of animals per dose group. Gaylor & Gold [9] show that a quick estimate of potential cancer potency can be obtained from a 90-day MTD without conducting a two-year bioassay. Based on a survey of chemicals shown to be carcinogenic by the US National Toxicology Program, the TD_{01} is generally within a factor of ten of the MTD/74.

As discussed earlier, linear extrapolation is used for low-dose risk estimation for genotoxic carcinogens. Also, low-dose linearity is expected when a substance augments a carcinogenic process (additivity to background) that is already producing tumors due to endogenous or other exogenous exposure. For other conditions, it is generally postulated that adverse health effects do not occur at low doses, where the body is capable of excreting or detoxifying small amounts of potentially toxic substances. That is, it is often postulated that there are threshold doses below which no toxicity occurs. However, this becomes a rather tenuous assumption for a heterogeneous population. The recent US Environmental Protection Agency [23] carcinogen risk assessment guidelines propose estimation of the ED_{10} (the dose producing an excess 10% tumor incidence), being a dose that is in or near the experimental dose range and can generally be determined with adequate precision. A lower confidence limit LED_{10} is used to account for experimental variation. The LED_{10} is then used as a point of departure for low-dose extrapolation. Where linearity is expected, low-dose risk estimation is obtained by linear extrapolation to zero. That is, the probability of cancer produced by a dose d is $\Pr(d) = (0.10/\mathrm{LED}_{10})d$, where $(0.10/\mathrm{LED}_{10})$ is the low-dose slope (carcinogenic potency). Where a nonlinear dose–response is expected, biologically based dose–response models may be used to estimate risk or the LED_{10} may be divided by a series of safety (uncertainty) factors to arrive at a dose with an acceptably low level of risk, and perhaps zero risk.

Noncancer Endpoints

For biological endpoints other than cancer, quantitative risk estimation generally is not used. Here it is assumed that excretion and/or detoxification at low doses will eliminate any adverse health effects. That is, it is assumed that there is a threshold dose below which no adverse health effects occur. This may be a tenuous assumption for heterogeneous populations or where a chemical exposure may contribute to an ongoing process that already results in adverse effects in unexposed individuals. In this case, additivity of a dose to an endogenous/exogenous dose that already surpasses a threshold dose will result in increased risk.

The general safety assessment procedure for noncancer endpoints is to divide the no observed adverse effect level (NOAEL) by safety (uncertainty) factors to obtain an acceptable daily intake (ADI) or reference dose (RfD)

$$\mathrm{ADI} = \mathrm{RfD} = \frac{\mathrm{NOAEL}}{(\mathrm{U}_1 * \mathrm{U}_2 * * *)}.$$

A safety (uncertainty) factor of ten is generally used for extrapolation from animals to humans. Another factor of ten is generally used for sensitive individuals in a population. Another factor of up to ten may be used for shortcomings in the available data. If all the doses in a bioassay produce an adverse effect, a NOAEL is not present. In this case, the lowest observed adverse effect level (LOAEL) is used, and an additional factor up to ten is introduced to account for the ratio of the LOAEL to the NOAEL. Barnes & Dourson [3] state that: "RfD (reference dose) is an estimate (with uncertainty spanning perhaps an order of magnitude) of a daily exposure to a human population (including sensitive subgroups) that is likely to be without appreciable risk of deleterious effects during a lifetime". Biologically based dose–response models have been developed sparingly to estimate risk below the NOAEL or LOAEL.

The process utilizing the NOAEL is limited to the dose levels used in an experiment, and does not make use of the dose–response information. Poorer experiments with higher NOAELs are unjustly rewarded with higher ADIs or RfDs. To overcome these shortcomings, Crump [5] proposed a benchmark dose (BMD) approach. Here the dose–response data are used to estimate a low level of excess risk; for example, 1%–10%. This dose is generally in or near the experimental dose range and is directly estimable without extrapolation. To account for experimental variation, a lower confidence limit on the BMD is used. This rewards better experiments with tighter limits with a higher ADI or RfD. This limit on the BMD can then be used as a substitute for the

NOAEL in calculating an ADI or RfD. This procedure eliminates the use of NOAELs with potentially high levels of risk, perhaps greater than 20%, for quantal data [18].

Often, continuous (nonquantal) data are obtained for noncancer endpoints; for example, clinical chemistry, hematology, and body and organ weights. For such cases, there generally is no sharp demarcation between normal and adverse levels of an effect. Gaylor & Slikker [11] proposed that the distribution of the results in unexposed control animals be used to establish an abnormal range; for example, below the 1st percentile or above the 99th percentile (*see* **Quantiles**). From the distribution of levels in dosed animals – for example, normal or lognormal – it is possible to estimate the proportion in the abnormal range as a function of dose. This procedure can also be used to estimate benchmark doses for low-dose safety assessment.

References

[1] Allen, B.C., Crump, K.S. & Shipp, A.M. (1988). Correlation between carcinogenic potency of chemicals in animals and humans, *Risk Analysis* **8**, 531–544.

[2] Armitage, P. & Doll, R. (1961). Stochastic models for carcinogenesis, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, L.M. LeCam & J. Neyman, eds. University of California Press, Berkeley.

[3] Barnes, D.G. & Dourson, M. (1988). Reference dose (RFD): description and use in health risk assessments, *Regulatory Toxicology and Pharmacology* **8**, 471–486.

[4] Chen, J.J., Kodell, R.L. & Gaylor, D.W. (1988). Using the biological two-stage model to assess risk from short-term exposures, *Risk Analysis* **8**, 223–230.

[5] Crump, K.S. (1984). A new method for determining allowable daily intakes, *Fundamental and Applied Toxicology* **4**, 854–871.

[6] Crump, K.S., Hoel, D.G., Langley, C.H. & Peto, R. (1976). Fundamental carcinogenic processes and their implications for low dose risk assessment, *Cancer Research* **36**, 2973–2979.

[7] Farmer, J.H., Kodell, R.L. & Gaylor, D.W. (1982). Estimating and extrapolating of tumor probabilities from a mouse bioassay with survival/sacrifice components, *Risk Analysis* **2**, 27–34.

[8] Gaylor, D.W. & Chen, J.J. (1996). A simple upper limit for the sum of the risks of the components in a mixture, *Risk Analysis* **16**, 395–398.

[9] Gaylor, D.W. & Gold, L.S. (1995). Quick estimate of the regulatory virtually safe dose based on the maximum tolerated dose for rodent bioassays, *Regulatory Toxicology and Pharmacology* **22**, 57–63.

[10] Gaylor, D.W. & Kodell, R.L. (1980). Linear interpolation algorithm for low dose risk assessment of toxic substances, *Journal of Environmental Pathology and Toxicology* **4**, 305–312.

[11] Gaylor, D.W. & Slikker, W. Jr (1990). Risk assessment for neurotoxic effects, *Neurotoxicology* **11**, 211–218.

[12] Gaylor, D.W., Chen, J.J. & Sheehan, D.M. (1993). Uncertainty in cancer risk estimates, *Risk Analysis* **13**, 149–154.

[13] Gaylor, D.W., Kodell, R.L. & Chen, J.J. (1985). Experimental design of bioassays for screening and low dose extrapolation, *Risk Analysis* **5**, 9–16.

[14] Gaylor, D.W., Kodell, R.L., Chen, J.J., Springer, J.A., Lorentzen, R.J. & Scheuplein, R.J. (1994). Point estimates of cancer risk at low doses, *Risk Analysis* **14**, 843–850.

[15] Kodell, R.L., Gaylor, D.W. & Chen J.J. (1987). Using average lifetime dose rate for intermittent exposures to carcinogens, *Risk Analysis* **7**, 339–345.

[16] Kodell, R.L., Krewski, D. & Zielinski, J.M. (1991). Additive and multiplicative relative risk in the two-stage clonal expansion model of carcinogenesis, *Risk Analysis* **11**, 483–490.

[17] Krewski, D., Gaylor, D.W. & Szyszkowics, M. (1991). A model-free approach to low-dose extrapolation, *Environmental Health Perspectives* **90**, 279–285.

[18] Leisenring, W. & Ryan, L. (1992). Statistical properties of the NOAEL, *Regulatory Toxicology and Pharmacology* **15**, 161–171.

[19] Mantel, N. & Bryan, W.R. (1961). "Safety" testing of carcinogenic agents, *Journal of the National Cancer Institute* **27**, 455–470.

[20] Peto, R. (1978). Draft report: carcinogenic effects of chronic exposure to very low levels of toxic substances, *Environmental Health Perspectives* **22**, 155–161.

[21] Portier, C. & Hoel, D.G. (1983). Optimal design of the chronic animal bioassay, *Journal of Toxicology and Environmental Health* **12**, 1–19.

[22] US Environmental Protection Agency (1992). A cross-species scaling factor for carcinogen risk assessment based on equivalence of $mg/kg^{3/4}$ / day, *Federal Register* **57**, 24152–24173.

[23] US Environmental Protection Agency (1996). Proposed guidelines for carcinogen risk assessment: notice, *Federal Register* **61**, 17960–18011.

(*See also* **Animal Screening Systems**; **Serial-sacrifice Experiments**)

DAVID W. GAYLOR

Extrapolation

As used here, the term, *extrapolation*, refers to a projection made at a point that is beyond the range of the data used to estimate the parameters of the statistical model on which the projection is based. For example, data from a hypothetical **observational study** among persons 20–64 years of age may have been used to derive a **simple linear regression** model that states that the expected maximum heart rate, y, for a person x years of age follows the relationship given by

$$y = 210 - 0.5x.$$

If this model, derived from subjects 20–64 years of age, were applied to teenagers or to very elderly persons, then it may be that the resulting extrapolation does very poorly. While the relationship may be linear within the range of the data used in the estimation of parameters, the same linear model may not fit outside this range.

In spite of the dangers involved in extrapolation, policy decisions on the health risks of certain environmental exposures are sometimes based on evidence found in studies in which the exposures are much higher than those found in even the most highly polluted environments (*see* **Risk Assessment for Environmental Chemicals**). For example, human studies on the existence and strength of the putative relationship between asbestos and lung cancer are invariably based on epidemiological investigations conducted in occupational settings, where the exposure to asbestos among workers (*see* **Occupational Epidemiology**) is orders of magnitude higher than what would be found in the environment. Likewise, in studying the association with cancer of variables such as the intake of micronutrients, food additives, pharmaceutical drugs, etc. evidence is often presented on the findings of relationships in animal experiments (*see* **Tumor Incidence Experiments**). In these experiments, however, it is frequently the case that the animals are tested not only at levels of the particular substance that are much higher than those experienced by humans, but also under conditions that cannot be duplicated in humans. An example of this is the decision made by the Food and Drug Administration in the late 1960s to ban the food additive *cyclomate*, because a Canadian Study conducted at very high levels of cyclomate seemed to indicate an association with bladder cancer.

In **time series** studies, models are constructed showing the relationship of a variable with time and the parameters fit on the basis of available current and past data. The main objective of these studies, however, is to extrapolate from these models into the future.

In conclusion, extrapolation is risky but sometimes necessary, and investigators should be aware that they are extrapolating and should exert extreme caution in the interpretation of their findings.

PAUL S. LEVY

Extreme Values

Extreme-value statistics, or statistics of extremes, is concerned with the occurrence and sizes of rare events, be they larger or smaller than usual. Examples are the ages of the oldest members of a population, highest annual tides [5] or rainfall [7] or wind speeds [8], athletics records [33], and the time to failure of a system with many components. This last example makes a connection with *reliability theory* and **survival analysis**, to which statistics of extremes is connected, though it has a somewhat different emphasis. Primarily under the impetus of problems in environmental science and engineering, statistics of extremes has developed very rapidly in the two decades. Coles [2] gives an excellent introduction to the subject, and the more mathematical book-length treatment [12] is oriented toward applications in finance. A recent overview is provided by the edited volume [14], while [31] and [23] give other accounts.

The discussion below concerns high extremes – maxima – but in applications, minima can be dealt with by reversing the signs of the observations and applying procedures for maxima.

Maxima

The commonest approach to extremes is through sample maxima. A key result in this context is the Extremal Types Theorem [15–17], which addresses the following question: given a sample of independent identically distributed **random variables** $X_1, \ldots, X_k$, what are the possible limiting distributions of $M_k = a_k[\max(X_1, \ldots, X_k) - b_k]$ as $k \to \infty$? The answer the theorem gives is that if a nondegenerate limiting cumulative distribution function (cdf) exists for some sequences of constants a_k and b_k, it must fall into one of the following three classes

$$\begin{aligned}
&\text{I: } F(x) = \exp[-e^{-x}], \quad -\infty < x < \infty,\\
&\text{II: } F(x) = \begin{cases} 0, & x \le 0,\\ \exp(-x^{-\alpha}), & x > 0, \alpha > 0,\end{cases}\\
&\text{III: } F(x) = \begin{cases} \exp[-(-x)^{\alpha}], & x < 0, \alpha > 0,\\ 1, & x \ge 0.\end{cases}
\end{aligned} \tag{1}$$

These distributions are known collectively as the extreme-value distributions, with types I, II, and III known as the Gumbel, Fréchet, and Weibull types (*see* **Parametric Models in Survival Analysis**). The more usual form for the **Weibull distribution** arises as a limit for minima rather than for maxima, and is given by the type III class with x replaced by $-x$. The importance of the Extremal Types Theorem is that it guarantees that if a limit exists for maxima, it must have one of the specified forms. Consequently, much of the older literature [18, 30] on extreme values focuses on fitting these distributions to sample maxima. The modern approach is to combine them into the generalized extreme-value distribution (GEV) with cdf

$$H(y) = \exp\left\{-\left[1 + \xi\left(\frac{y-\eta}{\tau}\right)\right]^{-1/\xi}\right\}, \quad -\infty < \eta, \xi < \infty, \tau > 0, \tag{2}$$

defined for values of y for which $1 + \xi(y - \eta)/\tau > 0$. Apart from a location and scale change parameterized by η and τ, this gives the type II and III classes for $\xi > 0$ and $\xi < 0$, and the type I arises in the limit as $\xi \to 0$. Consequently, the shape parameter ξ plays a key role, with $\xi > 0$ giving distributions with heavy upper tails and $\xi < 0$ giving distributions with a finite upper endpoint, while the Gumbel distribution has cdf

$$H(y) = \exp\left\{-\exp\left[-\left(\frac{y-\eta}{\tau}\right)\right]\right\}, \quad -\infty < y < \infty, \tag{3}$$

and lies between the two.

A typical hydrological application of the GEV is to fit it to a sample of annual maxima of daily river levels and to use the fitted distribution to estimate the $1/p$-year return level, that is, the river level exceeded once on average every $1/p$ years; here $0 < p < 1$. The quantity $1/p$ is known as the return period and is important in engineering design. The usual return level estimate is the $1 - p$ **quantile** of the GEV,

$$y_{1-p} = \eta - \frac{\tau}{\xi}\left\{1 - [-\ln(1-p)]^{-\xi}\right\}, \tag{4}$$

with parameters replaced by estimates. Two concerns in practice are that inference is often required for a return period longer than the amount of data available and that the fitted GEV is very sensitive to the values of the most extreme observations; these difficulties are inherent in the subject.

Many methods [22] have been proposed for fitting extreme-value distributions to an independent identically distributed sample of maxima, $Y_1, \ldots, Y_n$. Most attention has been focused on the Gumbel case, for which a *probability plot* of sample **order statistics** $Y_{(1)} \leq \cdots \leq Y_{(n)}$ is a valuable tool. This plots $Y_{(r)}$ against the quantiles $-\ln\{-\ln[r/(n+1)]\}$, $r = 1, \ldots, n$, and is useful for detecting **outliers** and assessing the fit of the distribution, in addition to providing graphical estimates of η and τ from the intercept and slope of the graph. Other methods of fitting include the use of **moments** (*see* **Method of Moments**) – which can be highly inefficient – and of probability-weighted moments [19], but these approaches are hard to extend to the more complicated models needed when data are **censored**, for example. **Bayesian** modeling is discussed by Coles and Powell [3] and Coles and Tawn [6].

Usually, **maximum likelihood** estimates of the GEV parameters can be obtained numerically and their **standard errors** calculated from the observed **information matrix**, though convergence problems can arise. There are theoretical difficulties when $\xi < -\frac{1}{2}$, in which case the usual properties of maximum likelihood estimates do not apply [34]. **Confidence intervals** for the estimated return level $\hat{y}_{1-p}$ can be obtained from its **profile loglikelihood**, or by applying the **delta method** to get standard errors for $\ln \hat{y}_{1-p}$. Similar techniques can be applied to more complex situations, for example, where η and τ depend on **explanatory variables**.

Point Process Characterization

A serious objection to the approach sketched above is that the use of maxima alone is wasteful of data: most of the information in the sample is ignored. This has led to other approaches, based on the following characterization. Let $X_1, \ldots, X_{nk}$ be a set of nk independent identically distributed random variables, and consider the pattern in the plane with points at (x, y) coordinates $(j/(nk+1), a_k(X_j - b_k))$, $j = 1, \ldots, nk$. Then, provided a_k and b_k are chosen in such a way that a limiting distribution for M_k exists as $k \to \infty$ with n fixed, the pattern of points above a sufficiently large threshold t will converge to a nonhomogeneous **Poisson process** with the properties that: (a) the numbers of points in nonoverlapping regions of the set $(0, 1) \times (t, \infty)$ are independent; and (b) if $u > t$, the probability that there are no points in the rectangular region $(x_1, x_2) \times (u, \infty)$ in the (x, y)-plane can be written as

$$\exp\left[-n(x_2 - x_1)\left(1 + \xi\frac{u-\eta}{\tau}\right)^{-1/\xi}\right]. \quad (5)$$

This characterization can be used to derive a variety of limiting results for maxima. The simplest is given by noting that the rescaled maximum of k observations, M_k, is less than $y > t$ only if there are no points in the set $(0, 1/n) \times (y, \infty)$, in which case (5) immediately gives (2). For a second result, the characterization shows that if N observations, $y_1, \ldots, y_N$, exceed a threshold $u > t$ over a period of n years, their joint probability density function (pdf) is

$$\exp\left[-n\left(1 + \xi\frac{u-\eta}{\tau}\right)^{-1/\xi}\right] \times \prod_{j=1}^{N} \frac{1}{\tau}\left(1 + \xi\frac{y_j - \eta}{\tau}\right)^{-1/\xi-1},$$

which can be used as a **likelihood** for η, τ, and ξ. Maximum likelihood inference can be performed numerically for this point process model and regression models based on it [36].

One practical matter is the choice of level t above which this characterization can be used. Too high a value for t will result in loss of information about the process of extremes, while too low a value will lead to **bias** because the point process model applies only asymptotically for high thresholds and will not fit the data adequately, if t is too low. The value of t is usually chosen empirically, by calculating the quantities of interest for a number of thresholds and choosing the lowest above which the results are relatively stable.

Further uses of this characterization are outlined below.

Resnick [32] discusses related point process representations for extremes.

r-largest Extremes

Clearly, the r largest observations among $X_1, \ldots, X_k$ will contain more information about the extremes than the maximum alone. Let us denote the r

largest observations by $M_k^1 \geq \cdots \geq M_k^r$. Then on setting $u = M_k^r$ in the point process characterization described above, we see that the asymptotic joint pdf of $M_k^1, \ldots, M_k^r$ at $m_k^1, \ldots, m_k^r$ is

$$\exp\left[-\left(1+\xi\frac{m_k^r-\eta}{\tau}\right)^{-1/\xi}\right] \times \prod_{j=1}^{r}\frac{1}{\tau}\left(1+\xi\frac{m_k^j-\eta}{\tau}\right)^{-1/\xi-1},$$

which can be used to form a likelihood for the parameters. In applications, r is chosen in order to trade off the increased precision from large values of r against the bias incurred if r is too large; a typical choice is to take r in the range from 5 to 10 values per year of data. Once again, likelihood inference for more complicated situations can be based on this model; see Smith [35] and Tawn [37].

Threshold Methods

A further approach is based on the idea of modeling exceedances of the data over a high threshold. The characterization sketched above can be used to show that the cdf of the amount by which an observation exceeds a high threshold t, given that it has done so, is

$$\begin{aligned}\text{pr}(X \leq t+y \mid X > t) &= G(y)\\ &= 1-\left(1+\xi\frac{y}{\tau}\right)^{-1/\xi}, \quad y>0; \end{aligned} \tag{6}$$

this is called the generalized **Pareto distribution** (*see* **Parametric Models in Survival Analysis**). As $\xi \to 0$, $G(y)$ becomes the **exponential distribution** with mean τ, which here occupies the same central role as the Gumbel distribution for maxima. Notice that the conditioning in (6) removes dependence on the location parameter η. Equation (6) can be used as the basis for a likelihood for τ and ξ, and its properties also lead to procedures for choosing the threshold t, for example, by taking the lowest threshold above which the *mean residual life plot* of the exceedances is straight (*see* **Life Expectancy**).

Davison and Smith [9] give an extensive discussion of this model.

Dependence

In practice, extreme values generally arise from series of dependent observations, rather than from independent data, and this would seem to limit the usefulness of the results described above. Problems might potentially be raised either by long-range or short-range dependence of the series.

An extensive mathematical theory summarized in Leadbetter et al. [25] and Leadbetter and Rootzén [26] shows that provided there is no long-range dependence between extremes, the same limiting results apply for maxima of dependent series as for independent series. Independence of widely separated extremes seems reasonable in most applications, but they almost always display short-range dependence, in which clusters of extremes occur together.

Suppose that a stationary series $X_1, \ldots, X_k$ with no long-range dependence of extremes has short-range dependence, which leads to extremes occurring in clusters with mean size $1/\theta$, where $0 \leq \theta \leq 1$; θ is called the extremal index of the process. The notion of a cluster here is deliberately vague. Let $X_1^*, \ldots, X_k^*$ be a sequence of independent variables with the same marginal distribution as the X_j. Then a key result [24] is that the rescaled maximum $M_k = a_k[\max(X_1, \ldots, X_k) - b_k]$ has a nondegenerate limiting distribution $H(y)$ if and only if $M_k^* = a_k[\max(X_1^*, \ldots, X_k^*) - b_k]$ has a nondegenerate limiting distribution $H^*(y)$, and $H(y) = [H^*(y)]^\theta$. The importance of this is to show that the type of limit distribution is unaffected by short-term dependence, as although the values of η and τ for $H(y)$ and $[H^*(y)]^\theta$ will differ, they have the same value of ξ. The usual solution to **clustering** therefore is to identify clusters, fit the point process model to their maxima, and then to adjust the estimated values of η and τ to allow for the clustering through an estimated extremal index. Ferro and Segers [13] give a recent discussion of estimation of θ.

Multivariate Extremes

Multivariate extremes arise when there is interest in rare values of two or more different series. Suppose, for example, that there is a series of bivariate observations (X_{1j}, X_{2j}), $j = 1, \ldots, k$, and that interest is focused on the componentwise maxima $M_{1k} = \max(X_{11}, \ldots, X_{1k})$ and $M_{2k} = \max(X_{21}, \ldots, X_{2k})$. The analogue of the Extremal Types Theorem is then

to seek sequences of constants a_{1k}, b_{1k}, a_{2k} and b_{2k} such that

$$\text{pr}\,[a_{1k}(M_{1k} - b_{1k}) \leq x_1, a_{2k}(M_{2k} - b_{2k}) \leq x_2] \quad (7)$$

converges to a nondegenerate limiting distribution. If they exist, the marginal limiting distributions for the rescaled versions of M_{1k} and M_{2k} must be of form (2). The joint distribution can then be of a wide range of possible forms, subject to mild conditions. Suppose that the probability integral transform is used to transform each of the rescaled maxima into a minimum, with a unit exponential distribution. Then the possible joint limiting distributions of maxima correspond to a class of bivariate exponential distributions for minima Z_1 and Z_2, whose joint survivor function $S(z_1, z_2) = \text{pr}(Z_1 > z_1, Z_2 > z_2)$ can be written as

$$S(z_1, z_2) = \exp\left[-(z_1 + z_2)A\left(\frac{z_2}{z_1 + z_2}\right)\right], \quad (8)$$

where the dependence function $A(w)$ is a convex function on $w \in [0, 1]$, with its graph lying entirely in the triangle with vertices $(0, 1)$, $(1, 1)$ and $\left(\frac{1}{2}, \frac{1}{2}\right)$. An example of such a function is $A(w) = [w^{1/\alpha} + (1 - w)^{1/\alpha}]^{\alpha}$, with parameter α such that $0 \leq \alpha \leq 1$ [20], though numerous other functions have been proposed [23, Chapter 3]. The usual approach to estimation of such a model based on a sample of pairs (Y_{1j}, Y_{2j}), $j = 1, \ldots, n$, is to maximize the likelihood for the parameters of the dependence function and both marginal distributions.

Ledford and Tawn [27–29] have investigated the behavior of joint extremes when the underlying variables are close to asymptotic independence.

For problems with several maxima, a somewhat different approach is better [11], though the basic ideas are similar; the extension to extremal processes of maxima [1, 7, 10] is more complicated. There is a related approach based on threshold analyses of the marginal series [4, 21].

Acknowledgment

S. G. Coles commented helpfully on a draft.

References

[1] Coles, S.G. (1993). Regional modelling of extreme storms via max-stable processes, *Journal of the Royal Statistical Society, Series B* **55**, 797–816.

[2] Coles, S.G. (2001). *An Introduction to the Statistical Modeling of Extreme Values*. Springer, New York.

[3] Coles, S.G. & Powell, E.A. (1996). Bayesian methods in extreme value modelling: A review and new developments, *International Statistical Review* **64**, 119–136.

[4] Coles, S.G. & Tawn, J.A. (1991). Modelling extreme multivariate events, *Journal of the Royal Statistical Society, Series B* **53**, 377–392.

[5] Coles, S.G. & Tawn, J.A. (1994). Statistical methods for multivariate extremes: An application to structural design (with Discussion), *Applied Statistics* **43**, 1–48.

[6] Coles, S.G. & Tawn, J.A. (1996a). A Bayesian analysis of extreme rainfall data, *Applied Statistics* **45**, 463–478.

[7] Coles, S.G. & Tawn, J.A. (1996b). Modelling extremes of the areal rainfall process, *Journal of the Royal Statistical Society, Series B* **58**, 329–347.

[8] Coles, S.G. & Walshaw, D. (1994). Directional modelling of extreme wind speeds, *Applied Statistics* **43**, 139–157.

[9] Davison, A.C. & Smith, R.L. (1990). Models for exceedances over high thresholds (with discussion), *Journal of the Royal Statistical Society, Series B* **52**, 393–442.

[10] de Haan, L. (1984). A spectral representation for max-stable processes, *Annals of Probability* **12**, 1194–1204.

[11] de Haan, L. & Resnick, S.I. (1977). Limit theory for multivariate sample extremes, *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* **40**, 317–337.

[12] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

[13] Ferro, C.A.T. & Segers, J. (2003). Inference for clusters of extreme values, *Journal of the Royal Statistical Society, Series B* **65**, 545–556.

[14] Finkenstädt, B. & Rootzén, H. eds. (2004). *Extreme Values in Finance, Telecommunications, and the Environment*. Chapman & Hall/CRC, New York.

[15] Fisher, R.A. & Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society* **24**, 180–190.

[16] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd Ed. Krieger, Melbourne, Fl.

[17] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics* **44**, 423–453.

[18] Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

[19] Hosking, J.R.M., Wallis, J.R. & Wood, E.F. (1985). Estimation of the Generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics* **27**, 251–261.

[20] Hougaard, P. (1986). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678.

[21] Joe, H., Smith, R.L. & Weissman, I. (1992). Bivariate threshold methods for extremes, *Journal of the Royal Statistical Society, Series B* **54**, 171–183.

[22] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). *Continuous Univariate Distributions*, 2nd Ed., Vol. I. Wiley, New York.

[23] Kotz, S. & Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.

[24] Leadbetter, M.R. (1983). Extremes and local dependence in stationary sequences, *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* **65**, 291–306.

[25] Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

[26] Leadbetter, M.R. & Rootzén, H. (1988). Extremal theory for stochastic processes, *Annals of Probability* **16**, 431–478.

[27] Ledford, A.W. & Tawn, J.A. (1996). Statistics for near independence in multivariate extreme values, *Biometrika* **83**, 169–187.

[28] Ledford, A.W. & Tawn, J.A. (1997). Modelling dependence within joint tail regions, *Journal of the Royal Statistical Society, Series B* **59**, 475–499.

[29] Ledford, A.W. & Tawn, J.A. (2003). Diagnostics for dependence within time series extremes, *Journal of the Royal Statistical Society, Series B* **65**, 521–543.

[30] Mann, N.R. & Singpurwalla, N.D. (1982). Extreme-value distributions, in *Encyclopedia of Statistical Sciences*, Vol. 2, S., Kotz & N.L., Johnson, eds. Wiley, New York.

[31] Reiss, R.-D. & Thomas, M. (1997). *Statistical Analysis of Extreme Values*. Birkhäuser, Basel.

[32] Resnick, S.I. (1987). *Extreme Values, Point Processes and Regular Variation*. Springer, New York.

[33] Robinson, M.E. & Tawn, J.A. (1995). Statistics for exceptional athletics records, *Applied Statistics* **44**, 499–511.

[34] Smith, R.L. (1985). Maximum likelihood estimation in a class of non-regular cases, *Biometrika* **72**, 67–92.

[35] Smith, R.L. (1986). Extreme value theory based on the r largest annual events, *Journal of Hydrology* **86**, 27–43.

[36] Smith, R.L. (1989). Extreme value analysis of environmental time series: An example based on ozone data (with discussion), *Statistical Science* **4**, 367–393.

[37] Tawn, J.A. (1988). An extreme value theory model for dependent observations, *Journal of Hydrology* **101**, 227–250.

A.C. DAVISON

F Distributions

This is a statistical distribution with considerable practical importance. Suppose that X_1^2 and X_2^2 have independent **chi-square distributions** with m and n degrees of freedom, respectively. The ratio of these two χ^2 distributed variables, each divided by its degrees of freedom, has an F distribution. The F distribution is sometimes called the variance ratio distribution, or Snedecor's F distribution, and is related to Fisher's z distribution. A formal theorem which defines the F distribution is as follows.

Theorem. Let χ_1^2 have a distribution with m degrees of freedom; and let χ_2^2 have an independent χ^2 distribution with n degrees of freedom. Then the random variable

$$F = \frac{(\chi_1^2/m)}{(\chi_2^2/n)}$$

has an F distribution with m and n degrees of freedom. Such an F distribution is described by the density function

$$f(x) = \frac{m^{m/2}n^{n/2}}{B(m/2, n/2)} \frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}}, \quad x \geq 0.$$

The expression $B(m/2, n/2)$ represents a beta function evaluated at $m/2$ and $n/2$. The density function for a typical F distribution with degrees of freedom $m = 5$ and $n = 40$ is displayed in Figure 1.

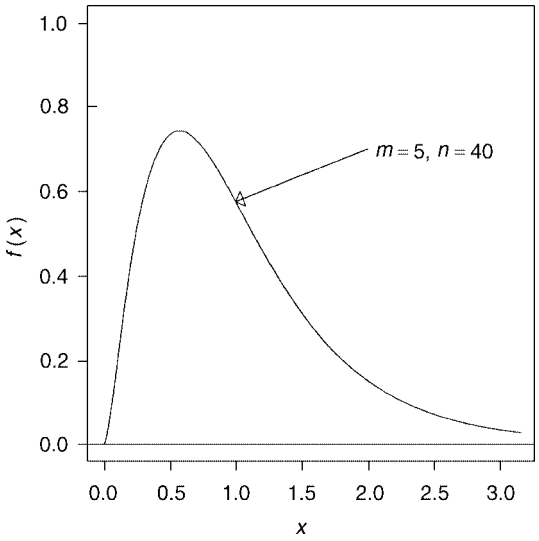

Figure 1 A typical F distribution: degrees of freedom $m = 5$ and $n = 40$

Properties

Like most density functions, the density $f(x)$ defines the **moments** of the F distribution. The expectation is

$$\mathrm{E}(F) = \frac{n}{n-2}, \quad n > 2,$$

with associated variance

$$\mathrm{var}(F) = \frac{2n^2(m+n-2)}{m(n-1)^2(n-4)}, \quad n > 4.$$

For an F distribution with $m = 5$ and $n = 40$ degrees of freedom, the expectation is $\mathrm{E}(F) = 1.053$ with $\mathrm{var}(F) = 0.529$. Values of the parameters m and n define a family of F distributions. Examples are illustrated in Figure 2 ($m = \{3, 5, \text{ and } 10\}$ and $n = 40$) and Figure 3 ($m = 10$ and $n = \{5, 10, \text{ and } 20\}$). The shape of the F distribution is called quasi-symmetric, since

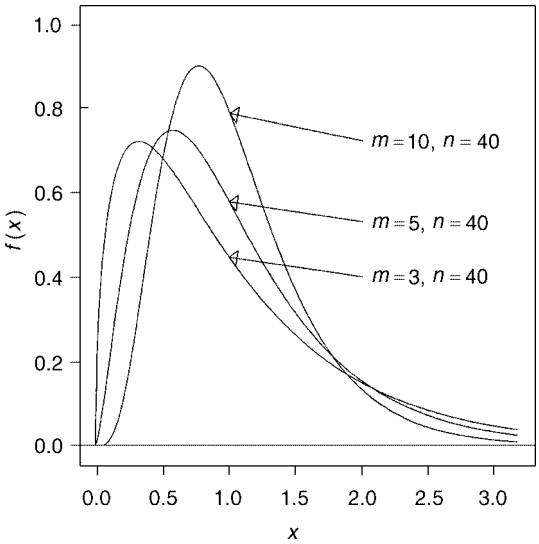

Figure 2 A family of three F distributions where $m = 3, 5$, and 10 for $n = 40$

$$F_\alpha[m, n] = \frac{1}{F_{1-\alpha}[n, m]},$$

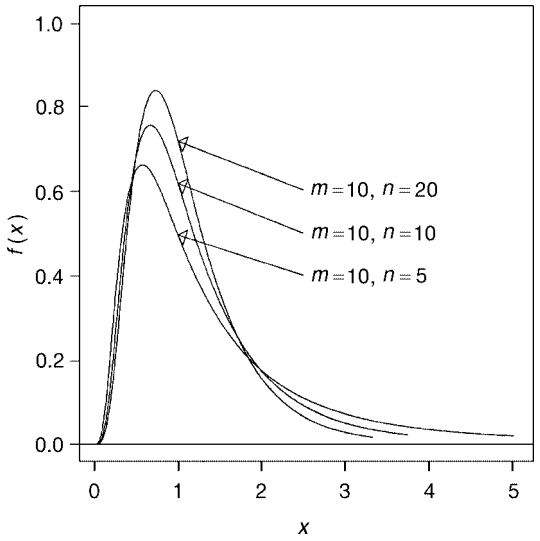

Figure 3 A family of three F distributions for $m = 10$, where $n = 5$, 10, and 20

where $F_\alpha[m, n]$ is the α-level **quantile** of the F distribution (i.e. $\Pr(F \leq F_\alpha[m, n]) = \alpha$). In more concrete terms, $\Pr(F \leq 0.75) = 0.409$ when $m = 5$ and $n = 40$ and $\Pr(F \leq 1/0.75) = \Pr(F \leq 1.333) = 0.591 = 1 - 0.409$ when $m = 40$ and $n = 5$. This quasi-symmetry property is displayed in Figure 4.

The probabilities associated with an F value (a random value with an F distribution) can be found in tables (e.g. [3]) for a variety of degrees of freedom, particularly when the m and n are small. For F values not in tables, a number of approximations exist [1]. A remarkably accurate ($n > 10$ or so) but relatively simple approximation based on the Wilson–Hilferty approximation [2] to the χ^2 distribution produces a value z from the standard normal distribution given by

$$z = \frac{\left[\left(1 - \frac{2}{9n}\right)x^{1/3} - \left(1 - \frac{2}{9m}\right)\right]}{\left(\frac{2}{9n}x^{2/3} + \frac{2}{9m}\right)^{1/2}}$$

where $\Pr(F \leq x) \approx \Pr(Z \leq z)$. For example, $\Pr(F \leq 0.75) \approx \Pr(Z \leq -0.235)$, where $z = -0.235$ is calculated from the Wilson–Hilferty expression when $m = 5$ and $n = 40$. Specifically, $\Pr(Z \leq -0.235) = 0.407$ from a standard normal distribution [3]. The exact value is 0.409, as before.

Relationships to Other Statistical Distributions

Several important statistical distributions are related to the F distribution. The random variable mF has a

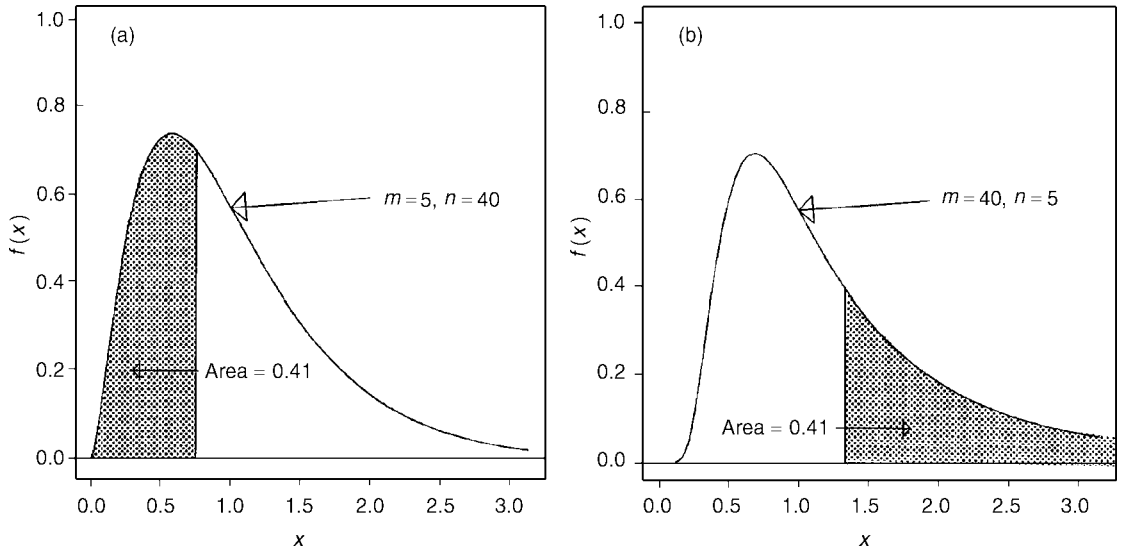

Figure 4 An example of the quasi-symmetry property of the F distribution (degrees of freedom $m = 5$ and $n = 40$). (a) df = 5, 40; (b) df = 40, 5

χ^2 distribution with m degrees of freedom when the variable F has an F distribution with m and infinite (∞) degrees of freedom. Similarly, the random variable $\sqrt{F}$ has a **Student's *t* distribution** with n degrees of freedom when F has an F distribution with 1 and n degrees of freedom. If k is an integer ($0 \leq k \leq N$), then

$$\Pr\left[F \geq \frac{(1-p)n}{pm}\right] = \Pr(B \geq k)$$

and B has a **binomial distribution** with parameters N and p when F has an F distribution with $m = 2(N-k+1)$ and $n = 2k$ degrees of freedom. The F distribution can be viewed as a special case of the β distribution. When the value F has an F distribution with m and n degrees of freedom, then the random variable $Y = mF/(n+mF)$ has a **beta distribution** with shape parameters $m/2$ and $n/2$. Also, Fisher's z distribution is simply another version of the F distribution, where $Z = (1/2)\log(F)$.

Applications of the *F* distribution

The most fundamental application of the F distribution comes from exploring differences between two estimated variances. If observations represented by y_{ij} are sampled independently from two normally distributed populations, the estimated variances are

$$S_1^2 = \frac{\sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2}{n_1-1} \quad \text{and} \quad S_2^2 = \frac{\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2}{n_2-1},$$

where n_i represents the number of observations sampled from each population ($i = 1, 2$). The ratio $F = S_1^2/S_2^2$ has an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom when both sampled populations have the same variance. Large or small values of the test statistic F are unlikely when the two population variances are equal. The F distribution, therefore, makes it possible to assign formal significance probabilities to the likelihood associated with differences observed between two estimated variances using an F ratio (*see* **Hypothesis Testing**).

The primary importance of the F distribution derives from its application to a large number of statistical tests involving the comparison of nested analytic models. A general approach to testing hypotheses involves collecting a random sample of n observations (represented by y_i) and contrasting the "fit" of the sampled data to two models developed under differing conditions. The first set of conditions (H_1) is evaluated by

$$SS_1 = \sum_{i=1}^{n}(y_i - [\text{mean estimated under specified conditions}])^2,$$

with associated degrees of freedom df_1. The degrees of freedom equal the number of sampled observations minus the number of independent estimates necessary to establish the components of the analytic model. The quantity SS_1 measures the goodness of fit of the data under the conditions specified by H_1 and is referred to as a residual sum of squares. A more restricted model (nested) is then postulated (H_0, called the null hypothesis) and a second residual sum of squares calculated, where

$$SS_0 = \sum_{i=1}^{n}(y_i - [\text{mean estimated under more restricted conditions}])^2,$$

with associated degrees of freedom df_0. One model is nested within the other when the second model is a special case of the first (*see* **Hierarchical Models**). Frequently, a nested model is created by deleting terms (H_0) from a more extensive model (H_1). The property that H_0 is a restricted case of H_1 guarantees that SS_0 will be equal to or larger than SS_1 and df_0 will be larger that df_1. The difference $SS_0 - SS_1$ reflects the impact of the additional restrictions which created the second model. Furthermore, if the y_i values are sampled from normal distributions with the same variances, then the random variable

$$F = \frac{(SS_0 - SS_1)/(df_0 - df_1)}{SS_1/df_1}$$

has an F distribution with $df_0 - df_1$ and df_1 degrees of freedom when the restrictions have only random effects on the sampled values y_i. Again, this makes it possible to use the F distribution to assign formal probabilities to observed test statistics. The comparison of residual sums of squares produces a systematic approach to statistical testing called the **analysis of variance**.

A test of two simple hypotheses concerning the mean value illustrates the contrasting of two nested models using an F test. The question addressed

is whether the sampled data are consistent with a population mean value of $\mu = 2[H_0]$ or whether there is evidence that the population mean value is not $\mu = 2[H_1]$. A small data set consisting of $n = 20$ observations independently sampled from a normal distribution is

$$y = \{0.4,\ 1.4,\ 2.0,\ 0.4,\ 1.4,\ 0.9,\ 2.5,\ 2.6,\ 0.5,\ 1.6,\ 0.7,\ 1.0,\ 2.4,\ 2.9,\ 2.1,\ 0.9,\ 0.4,\ 1.2,\ 2.7,\ \text{and}\ 2.3\}.$$

The residual sum of squares assessing the conjecture that the population mean is not equal to $2[H_1 : \mu \neq 2]$ is

$$SS_1 = \sum_{i=1}^{20} (y_i - \overline{y})^2 = 14.026,$$

where

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 1.515.$$

For the more restricted conditions where the population mean is postulated to be $\mu = 2[H_0 : \mu = 2]$, the second residual sum of squares is

$$SS_0 = \sum_{i=1}^{20} (y_i - 2)^2 = 18.730.$$

An F statistic ($df_1 = n - 1 = 19$ and $df_0 = n - 0 = 20$) allows a significance probability to be calculated reflecting on the likelihood that the difference $SS_0 - SS_1 = 4.705$ arose by chance when the mean of the sampled population is in fact $\mu = 2(H_0$ is true). The F statistic is $F = 4.705/0.738 = 6.373$ and the corresponding significance probability is 0.021. Using an F test (analysis of variance) approach, the data indicate that it is not likely that the 20 observations are a random sample from a population with a mean value of $\mu = 2$. This same comparison can be made with Student's t test ($\overline{y} = 1.515$ vs. $\mu = 2$). However, the pattern of comparing nested models with an F ratio is the basis of a large variety of statistical testing procedures.

References

[1] Johnson, N.L. & Kotz, S. (1970). *Continuous Univariate Distributions*. Houghton Mifflin, New York.
[2] Kendall, M.G. & Stuart, A. (1976). *The Advanced Theory of Statistics*, 3rd Ed. Hafner, New York.
[3] Owen, D.B. (1962). *Handbook of Statistical Tables*. Addison-Wesley, Reading.

S. SELVIN

Factor Analysis, Confirmatory

The purpose of a **factor analysis** is to study the intercorrelations among p observed variables by postulating a set of common **factors**. Ideally, the number of common factors, say m, is less than the number of the observed variables, p. In a common-factor analysis model, each observed variable is written as a weighted sum of m common-factor scores and one unique-factor score. The collection of these equations for all p variables is called a *factor pattern*. It is given as follows:

$$\begin{aligned} X_1 &= a_1 A + b_1 B + \cdots + m_1 M + u_1 U_1, \\ X_2 &= a_2 A + b_2 B + \cdots + m_2 M + u_2 U_2, \\ &\vdots \qquad\qquad \vdots \\ X_p &= a_p A + b_p B + \cdots + m_p M + u_p U_p, \end{aligned}$$

where $X_1, X_2, \ldots, X_p$ represent the standardized measurement of the p observed variables, $A, B, \ldots, M$ represent the standardized scores in the uncorrelated common factors, $a_i, b_i, \ldots, m_i, i = 1, \ldots, p$, are the *common factor loadings*, $U_1, U_2, \ldots, U_p$ represent the standardized scores on the p unique factors, and $u_1, u_2, \ldots, u_p$ are the *unique factor loadings*. In the above model the m latent factors and the p unique factors are uncorrelated.

Through this linear model, we attempt to find a small number of underlying factors, which contain all the essential information about the **correlations** among the observed variables. In general, there are two steps to perform this analysis: (i) initial factoring of the data determining the number of salient factors to be retained (*see* **Battery Reduction**); and (ii) **rotating** the factors to obtain unique and interpretable results. There are various procedures available to carry out each of the three steps. Discussion on those procedures can be found in the articles on **Factor Analysis**, **Principal Components Analysis**, **Rotation of Axes**, and articles on different rotation methods. In the above we use factor analysis to explore the underlying construct of the data and to find and interpret the factors through the estimated factor matrices.

Sometimes, an investigator may have a hypothesized factor pattern or she may have obtained a factor matrix from a previous study which contains

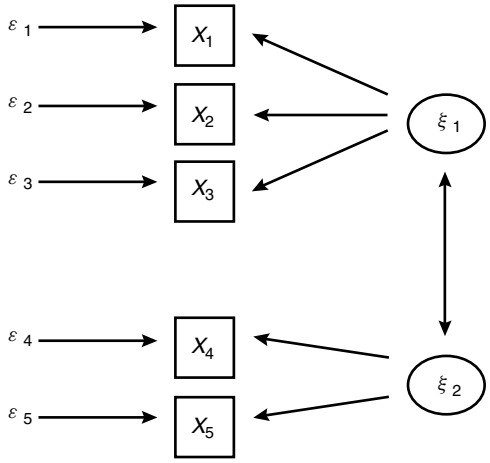

Figure 1 Path diagram for confirmatory factor analysis

the relationships between the variables and the factors. Instead of exploring the underlying construct of new data, the investigator might wish to estimate and test the fit of the data to the hypothesized factor model or to the model obtained from the previous study. In other words, she wants to perform an analysis to find out how the estimated factor matrix from the data matches with the hypothesized factor matrix. We call this analysis a *confirmatory factor analysis*. Figure 1 shows an example of a path diagram for a confirmatory factor analysis. It shows the paths that link the observed variables X_i with the factors ξ_j and the error terms ε_i involved in the system. In the Figure, the observed variables are enclosed in boxes, the factors are enclosed in circles, and ε_i represent the error terms. The single-headed arrow indicates the influence of the factors and error terms on the observed variables. The double-headed arrow indicates that the factors are correlated. The mathematical model for this confirmatory factor analysis is

$$\begin{aligned} X_1 &= a_1 \xi_1 + \varepsilon_1, \\ X_2 &= a_2 \xi_1 + \varepsilon_2, \\ X_3 &= a_3 \xi_1 + \varepsilon_3, \\ X_4 &= b_1 \xi_2 + \varepsilon_4, \\ X_5 &= b_2 \xi_2 + \varepsilon_5. \end{aligned} \tag{1}$$

Krzanowski [3] presented the **Procrustes** analysis, which is appropriate to use in a confirmatory factor analysis. The procedure involves standardization of

matrices in terms of their relative positions, orientations, and scales and also computation of the test statistic based on the standardized matrices. Lawley & Maxwell [4] have outlined some **least squares** methods for rotating to factor structures that come as close as possible to an a priori pattern of ones and zeros. Jöreskog & Sörbom [2] developed the **LISREL** package which consists of computer programs for such confirmatory factor analysis. There is also a computer package in SAS (*see* **Software, Biostatistical**) that can perform a confirmatory analysis [5]. The procedure is known as PROC CALIS. Hatcher [1] presented a detailed step-by-step approach on how to perform the confirmatory factor analysis using PROC CALIS and how to interpret the output obtained from the program.

References

[1] Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling*. SAS Institute Inc., Cary.

[2] Jöreskog, K.G. & Sörbom, D. (1982). *LISREL V – Estimation of Linear Structural Equations by Maximum Likelihood Methods*. National Educational Resources, Chicago.

[3] Krzanowski, W.J. (1988). *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, New York.

[4] Lawley, D.N. & Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*. Butterworth & Co., London.

[5] SAS Institute Inc. (1989). *SAS/STAT® User's Guide, Version 6*, 4th Ed, Vol. 2. SAS Institute Inc., Cary.

(*See also* **Factor Loading Matrix**; **Oblimin Rotation**; **Oblique Rotation**; **Optres Rotation**; **Orthoblique Rotation**; **Orthogonal Rotation**; **Structural Equation Models**)

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Factor Analysis, Overview

Suppose that we have measurements on a moderate or large number of variables. The central idea of factor analysis is that a smaller number of unobservable "factors" underlie the measured variables. More specifically, each measured variable can be written as a linear function of the factors, apart from a residual, or specific, factor.

Spearman [19] is generally acknowledged to be the "inventor" of factor analysis. However, as Bartholomew [2] points out, Spearman's original paper has little in common with the later developments which led to factor analysis as it is currently known. Spearman's "two-factor theory" assumed that scores of individuals on a number of tests could be decomposed into a "general" factor, common to all variables, measuring general intelligence, and a "specific" factor which was different for each variable. Bartholomew [2] describes Spearman's contribution more fully.

In current factor analysis terminology, Spearman's theory involves only a single (common) factor, whereas most factor analyses involve multiple factors. A key early reference to the idea of multiple factors is Thurstone [20]. The "factors" which were sought and modeled in much of the development of factor analysis were psychological factors. For example, Yule et al. [22] measured the scores, between 0 and 20, for 150 children aged $4\frac{1}{2}$–6 years, on ten subtests of the Wechsler Pre-School and Primary Scale of Intelligence. Five of the tests were "verbal" tests and five were "performance" tests so the expectation might be that the ten scores could be largely explained by two underlying factors, measuring "verbal" and "performance" abilities, respectively. This is, to some extent, true but as we shall see later when we look in detail at this example, this is not the whole story. Although factor analysis was developed largely as a tool for psychometricians (*see* **Psychometrics, Overview**), it has increasingly found use elsewhere in biostatistics. It has the potential to model and explain successfully a data set whenever the measured (continuous) variables can be linearly related to a smaller number of unobservable factors.

The Factor Model

Suppose that $\mathbf{x}$ is a vector of p variables (measurements, test scores) $x_1, x_2, \ldots, x_p$. The most usual model for factor analysis is, in matrix form,

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e}, \tag{1}$$

where a typical equation in the system given by (1) is

$$x_i = \mu_i + \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{im} f_m + e_i, \quad i = 1, 2, \ldots, p. \tag{2}$$

In (1), $\boldsymbol{\mu} = \mathrm{E}(\mathbf{x})$; in (2) $f_1, f_2, \ldots, f_m$ are unobservable common factors comprising the vector $\mathbf{f}$, and e_i is a residual or specific factor for variable x_i, $i = 1, 2, \ldots, p$. (1) and (2) imply that the measured variables can be expressed as linear combinations of the common factors, apart from a factor specific to each variable.

Certain assumptions are usually associated with the factor model, namely

$$\mathrm{E}[\mathbf{f}] = \mathrm{E}[\mathbf{e}] = \mathbf{0},$$

$$\mathrm{E}[\mathbf{f}\mathbf{e}'] = \mathbf{0} \text{ (matrix of zeros)},$$

$$\mathrm{E}[\mathbf{e}\mathbf{e}'] = \boldsymbol{\Psi} \text{ (diagonal)},$$

$$\mathrm{E}[\mathbf{f}\mathbf{f}'] = \mathbf{I}_m \text{ (identity matrix)}.$$

The assumptions concerning the expectations of $\mathbf{f}$ and $\mathbf{e}$ are conventional, convenient, and lose no generality. The first two covariance assumptions are simply stating that the elements of $\mathbf{e}$ are specific to each x_i (hence uncorrelated), and that these specific factors are uncorrelated with any common factors. The final assumption is not always made. It states that the common factors are uncorrelated and is a common and convenient assumption. However, some factor analysts argue that in reality common factors are often correlated (oblique) and so relax the assumption. For the moment we impose all the assumptions above. An additional assumption is sometimes made, namely that the x_is are standardized, so $\mathrm{E}[\mathbf{x}\mathbf{x}'] = \mathbf{I}_p$. We shall not use this assumption at present.

The factor model is one type of latent variable model in which a set of measured variables is "explained" by postulating the existence of unobserved "latent" variables. Factor analysis is a simple case in which both the measurements **(x)** and the latent variables **(f)** are continuous and relationships are linear. Bartholomew and Knott [3] describe models in which either $\mathbf{x}$ or $\mathbf{f}$ define categories. Some recently developed techniques such as **structural equations models** [4] and the **neural network** [5] also involve latent variables.

Estimation of the Factor Model

In the factor model there are a number of parameters which need to be estimated. The vector $\boldsymbol{\mu}$ is usually estimated by the vector $\bar{\mathbf{x}}$ of sample means for $x_1, x_2, \ldots, x_p$, but estimation of the elements of the matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ is more complex. The matrix $\boldsymbol{\Lambda}$ comprises so-called *loadings* (*see* **Factor Loading Matrix**), which describe how the variables are related to the common factors, while $\boldsymbol{\Psi}$ tells us how much of the variation in each x_i cannot be explained by the common factors. To understand some of the subtleties in estimating $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, we find the **covariance matrix** of each side of (1) which, using our assumptions, gives

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{3}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the vector of variables $\mathbf{x}$.

In practice $\boldsymbol{\Sigma}$ is unknown, but a sample covariance matrix, $\mathbf{S}$, is available, and fitting a factor model can be viewed as finding $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ which satisfy

$$\mathbf{S} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$$

as closely as possible. If the x_is are standardized, then $\mathbf{S}$ is replaced with the sample correlation matrix $\mathbf{R}$.

There is indeterminacy in finding $\boldsymbol{\Lambda}$, for suppose that $\mathbf{T}$ is an orthogonal matrix and that $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}$. Then

$$\begin{aligned}\boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*'} &= (\boldsymbol{\Lambda}\mathbf{T})(\boldsymbol{\Lambda}\mathbf{T})' = \boldsymbol{\Lambda}(\mathbf{T}\mathbf{T}')\boldsymbol{\Lambda}' \\ &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}',\end{aligned}$$

so $\boldsymbol{\Lambda}^*$ is as acceptable as $\boldsymbol{\Lambda}$ in fitting the model. This multiplication of $\boldsymbol{\Lambda}$ by an orthogonal matrix is known as *rotation* and, having found an initial solution, it is used to "simplify" the elements of $\boldsymbol{\Lambda}^*$ as much as possible. Simplification can be even greater if $\boldsymbol{\Lambda}$ is multiplied by a nonorthogonal matrix, leading to oblique, rather than orthogonal factors. In this article we concentrate on how to find an initial solution. Details of rotation in factor analysis are given elsewhere (*see*, in particular, **Orthogonal Rotation**; **Oblique Rotation**; **Oblimin Rotation**).

We describe three commonly used methods for finding initial estimates. A number of other methods have been proposed over the years, though few are widely used.

Maximum Likelihood Estimation

Maximum likelihood is perhaps the most "respectable" method of **estimation**, statistically speaking, but it makes the strong assumption that $\mathbf{f}$ and $\mathbf{e}$, and hence $\mathbf{x}$, have a **multivariate normal distribution**. We can then write down the **likelihood** function for $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\mu}$ as

$$\begin{aligned}L(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\mu}; \mathbf{x}) &= (2\pi)^{-p/2}|\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}|^{-1/2} \\ &\quad \times \exp\left[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}(\mathbf{x} - \boldsymbol{\mu})\right].\end{aligned}$$

Maximizing this with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ gives $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, and the following equations for $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ (for details of the derivation, see [3]):

$$\begin{aligned}&\hat{\boldsymbol{\Psi}}^{-1/2}\mathbf{S}\hat{\boldsymbol{\Psi}}^{-1/2}(\hat{\boldsymbol{\Psi}}^{-1/2}\hat{\boldsymbol{\Lambda}}) \\ &\quad = (\hat{\boldsymbol{\Psi}}^{-1/2}\hat{\boldsymbol{\Lambda}})(\mathbf{I} + \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}}),\end{aligned} \tag{4}$$

$$\hat{\boldsymbol{\Psi}} = \operatorname{diag}(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' - \mathbf{S}). \tag{5}$$

Eqs. (4) and (5) must be solved iteratively. For fixed $\hat{\boldsymbol{\Lambda}}$ an estimate $\hat{\boldsymbol{\Psi}}$ is found from (5) and substituted into (4). The eigenequation (4) is then solved for $\hat{\boldsymbol{\Psi}}^{-1/2}\hat{\boldsymbol{\Lambda}}$, and hence for $\hat{\boldsymbol{\Lambda}}$. The new estimate $\hat{\boldsymbol{\Lambda}}$ is substituted into (5), and so on until convergence.

Example

Table 1 lists the variables for the example from Yule et al. [22] introduced earlier, together with estimates of the loadings λ_{ij} obtained from maximum likelihood estimation, assuming that the number of common factors, m, is 2. Also given are maximum likelihood estimates of the specific variances, ψ_i, subtracted from 1. ψ_i is the amount of $\operatorname{var}(x_i)$ unaccounted for by the common factors, so $\operatorname{var}(x_i) - \psi_i$, known as the **communality** for variable i, is the variance in variable i explained by the common factors. Recall that often, as in Table 1, factor analysis is done using variables standardized to have variance 1, so that the communality is $1 - \psi_i$. In this case, the covariance matrices $\boldsymbol{\Sigma}$ and $\mathbf{S}$ become correlation matrices.

The results given in Table 1 were produced for the correlation matrix using Minitab, which has a fairly limited range of options for factor analysis (*see* **Software, Biostatistical**). Most standard statistical packages will have some factor analysis procedures, and in some cases an extensive selection of estimation and rotation techniques is available. It can be

Table 1 Maximum likelihood factor analysis for data from [22]: two-factor solution

	Unrotated loadings		Rotated loadings		
Variable	Factor 1	Factor 2	Factor 1	Factor 2	Communality
Information (x_1)	0.703	0.363	0.757	0.229	0.626
Vocabulary (x_2)	0.729	0.300	0.732	0.294	0.622
Arithmetic (x_3)	0.723	0.071	0.568	0.453	0.528
Similarities (x_4)	0.587	0.131	0.512	0.316	0.362
Comprehension (x_5)	0.710	0.307	0.723	0.275	0.598
Animal house (x_6)	0.528	−0.181	0.252	0.497	0.311
Picture completion (x_7)	0.663	−0.195	0.340	0.602	0.478
Mazes (x_8)	0.616	−0.377	0.179	0.700	0.522
Geometric design (x_9)	0.458	−0.115	0.249	0.402	0.223
Block design (x_{10})	0.777	−0.368	0.300	0.805	0.739

seen from Table 1 that the first unrotated factor has nontrivial loadings for all 10 variables, so that it could be interpreted as a general factor, measuring overall ability in the tests. The second unrotated factor contributes mainly to the first, second and fifth variables in a positive sense, and negatively to the eighth and tenth variables.

The communalities show that over 45% of the variability in the first, second, third, fifth, seventh, eighth, and tenth variables is accounted for by the two common factors, but that variables 4, 6, and 9 are rather poorly explained by these two factors.

Also given in Table 1 are the factor loadings after rotation. Different methods of rotation will not be discussed here. (*See* the section on Rotation in **Principal Components Analysis**.) For illustration Table 1 shows the results for the normal **varimax rotation**, which is often the default method in computer software. Different **orthogonal rotation** methods will often give similar results. The effect of rotation is to replace the general factor and the contrast between variables, which we had before rotation, by two factors involving subsets of the variables. The first factor contributes much more strongly to the first five variables than to the last five, reflecting the grouping of the 10 variables into two sets of five, which was noted earlier. The second factor is less clear; it has large loadings for variables 7, 8, and 10 and moderate loadings for variables 3, 6, and 9. The aim of rotation is to simplify the loadings and to produce a **simple structure**. This has been less successful in this example than in many, with substantial contributions from more than one factor to the same variable (x_3). Communalities are unchanged by orthogonal rotation.

The maximum likelihood approach to estimation seems rather restrictive because it assumes that all the variables in the factor model have multivariate normal distributions. In fact, the same solution can be obtained without making this distributional assumption. If the factor model in (1) is valid, together with its usual assumptions concerning means and variances, then the partial correlations between the elements of $\mathbf{x}$, given the values of $\mathbf{f}$, will all be zero. In attempting to fit the model we might therefore try to make all such partial correlations small. We can construct a $p \times p$ matrix of these partial correlations, the determinant of which will be maximized when all off-diagonal elements are zero. If this determinant is expressed in terms of the unknown parameters λ_{ij} and ψ_i, and then maximized, the resulting estimates of λ_{ij} and ψ_i are identical to the maximum likelihood estimates; for more details see [17, Section 8.8]. It can be argued that the use of linear models and correlations goes some way to an implicit assumption of multivariate normality, but the alternative derivation does away with the explicit assumption, and means that the maximum likelihood estimates are relevant in a rather broader range of problems.

Estimation Using Principal Components

This is the most common method of estimating initial factor loadings – it is often the default in computer software. Taking a pragmatic point of view, it is relatively simple to implement and often works well, giving similar results to other estimation techniques. However, its use has been responsible for a great deal of confusion between factor analysis on the one hand and **principal components analysis** on the other [11,

13]. It has led to the common misconception that principal components analysis is simply a special case of factor analysis, whereas they are really quite distinct techniques. We give here a brief introduction to principal components analysis, returning to a more detailed comparison with factor analysis later in the article.

Principal components analysis, like factor analysis, is a dimension-reducing technique, but it does not postulate any underlying unmeasurable factors. It simply finds linear functions of the measured variables $y_1 = \boldsymbol{\alpha}_1'\mathbf{x}$, $y_2 = \boldsymbol{\alpha}_2'\mathbf{x}, \ldots, y_p = \boldsymbol{\alpha}_p'\mathbf{x}$, which successively have maximum variance, subject to each y_i being uncorrelated with earlier y_js. The matrix $\mathbf{A}$, the columns of which are $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$, contains elements which are used to express the derived variables $\mathbf{y}$ in terms of the measured variables $\mathbf{x}$. The properties of principal components analysis mean that $\mathbf{A}$ is orthogonal, and we can also use its elements to express the measured variables $\mathbf{x}$ in terms of the derived variables $\mathbf{y}$. This is beginning to look rather like factor analysis, so the first m columns of $\mathbf{A}$ are often used as an initial estimate of $\boldsymbol{\Lambda}$ in an m-factor model. Table 2 gives the same information as Table 1, except that principal components loadings are presented instead of maximum likelihood loadings. It can be seen that, although there are differences in detail, the general patterns of loadings are very similar in Tables 1 and 2. The communalities are larger in Table 2 than in Table 1 for nine of the 10 variables. This is not unexpected because the definition of principal components analysis implies that it will maximize the sum over the variables of the communalities for any value of m.

After rotation, the general pattern of the loadings in Table 2 is again very similar to that in Table 1. It is quite often the case that the choice of an initial factor solution (and the choice of rotation method, provided we stick to orthogonal rotation) has little effect on the solution. The choice of the number of factors, which is discussed below, is often much more important in determining what the factors "look like".

Principal Factor Solutions (Common Factor Analysis)

Historically, various types of principal factor solution have been widely used; we describe them briefly here. Principal factor solutions recognize that principal components analysis is not really designed to fit the factor model in (1), and attempt to modify it in an *ad hoc*, but appropriate, way. Principal components analysis is based on an eigenanalysis of the covariance matrix $\mathbf{S}$ (*see* **Eigenvalue**; **Eigenvector**), and because principal components maximize variances they concentrate on "fitting" the diagonal elements of $\mathbf{S}$. Factor analysis, on the other hand, is mainly concerned with fitting the off-diagonal elements of $\mathbf{S}$ using common factors. As noted earlier, the variables $\mathbf{x}$ are often standardized, so that $\mathbf{S}$ is a correlation matrix. In this case, common factor analysis concentrates on explaining the correlations between variables using common factors, whilst allowing some of the (unit) variance of each variable to be unique to that variable. What principal factor analysis does is to replace principal components analysis's eigenanalysis of $\mathbf{S}$ by a similar analysis of $\mathbf{S} - \hat{\boldsymbol{\Psi}}$, where $\hat{\boldsymbol{\Psi}}$ is a diagonal matrix containing estimates of ψ_i. The diagonal elements of $\mathbf{S} - \hat{\boldsymbol{\Psi}}$, $s_{ii} - \hat{\psi}_i$, are estimates of the

Table 2 Two-factor solution based on principal components for data from [22]

	Unrotated loadings		Rotated loadings		
Variable	Factor 1	Factor 2	Factor 1	Factor 2	Communality
x_1	0.732	0.413	0.816	0.201	0.706
x_2	0.746	0.393	0.813	0.225	0.711
x_3	0.762	0.109	0.630	0.443	0.593
x_4	0.647	0.253	0.645	0.259	0.483
x_5	0.739	0.343	0.773	0.256	0.663
x_6	0.587	−0.260	0.250	0.591	0.412
x_7	0.706	−0.284	0.320	0.690	0.579
x_8	0.648	−0.539	0.103	0.837	0.711
x_9	0.511	−0.230	0.215	0.518	0.314
x_{10}	0.780	−0.353	0.327	0.791	0.733

communalities in the factor model, and these are the parts of the variances of $\mathbf{x}$ which are explainable by the common factors.

The different varieties of principal factor analysis arise for two reasons. The first is that $\hat{\boldsymbol{\Psi}}$ can be defined in several ways. A popular choice where $\mathbf{S}$ is a correlation matrix is to estimate the communality $(1 - \psi_i)$ by the square of the multiple correlation between x_i and the other $(p - 1)$ x variables. The second reason is that the eigenanalysis may be done just once to get estimates of $\hat{\boldsymbol{\Lambda}}$, given $\hat{\boldsymbol{\Psi}}$, or we can iterate. Iteration involves estimating the communalities by $\sum_{j=1}^{m} \hat{\lambda}_{ij}^2$, and hence ψ_i by $1 - \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$, given some estimate of the λ_{ij}s. An eigenanalysis is then done using the new estimates of ψ_i, leading to new estimates of $\boldsymbol{\Lambda}$, and so on until convergence of $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ see [8, Section 5.2] and [9, Section 6.3] for further discussion of communality estimation.

Rotation of Factors

Factor analysis without rotation is not really factor analysis at all. We have seen in the example above that rotation aims to simplify the factor loadings in the sense that each loading should ideally be close to, or far from, zero, so that it is clear whether or not a given factor has an effect on a particular variable. Medium-sized, equivocal, loadings are to be avoided. It was noted above that orthogonal rotation is achieved by postmultiplying the original matrix of loadings $\hat{\boldsymbol{\Lambda}}$ by an orthogonal matrix $\mathbf{T}$, to give $\hat{\boldsymbol{\Lambda}}^* = \hat{\boldsymbol{\Lambda}}\mathbf{T}$, where $\mathbf{T}$ is chosen so that $\hat{\boldsymbol{\Lambda}}^*$ has a **simple structure**. Various criteria exist for achieving this simplification. Usually there are only small differences between results for different forms of orthogonal rotation, but oblique rotation may provide somewhat different, and simpler, structure. Oblique rotation also transforms $\hat{\boldsymbol{\Lambda}}$ to $\hat{\boldsymbol{\Lambda}}^* = \hat{\boldsymbol{\Lambda}}\mathbf{T}$, but here $\mathbf{T}$ is no longer constrained to be orthogonal, and so does not strictly give a "rotation" of rigid axes. If $\boldsymbol{\Phi} = (\mathbf{T}'\mathbf{T})^{-1}$ is the matrix of correlations between factors after rotation, then $\boldsymbol{\Phi}$ is the identity matrix for orthogonal rotation, but is only restricted to have unit elements on the diagonal for oblique rotation, with no restrictions on off-diagonal elements; see [12] for further details.

Estimation of Factor Scores

Factor analysis may stop with the estimation and interpretation of (rotated) factor loadings and communalities. Sometimes, however, it is desirable to "estimate" the values or *scores* of each individual observation on the factors (*see* **Factor Scores**). As an illustration, computing such scores for the 150 children in our example will rank the children with respect to the "ability factors" which the analysis has uncovered. Plotting the scores might reveal groups or clusters of children with similar abilities, or individuals who differ from the bulk of the data set with respect to these factors. Finally, the scores might be used in further analyses, such as **regression** or **linear discriminant analysis**.

In (1) the variables are expressed in terms of the factors, whereas in computing the scores we require the relationship to be in the opposite direction. As Bartholomew [1] has pointed out, "estimation" is really the wrong word here. The factor scores are **random variables**, not unknown parameters, so **prediction** is a better description of what is being attempted.

If we make the normality, and other, assumptions used in deriving maximum likelihood estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, the conditional distributional of $\mathbf{f}$, given $\mathbf{x}$, can be found. It is the multivariate normal distribution $\mathrm{N}[\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1}]$. One plausible way of calculating factor scores is to use a sample version of the mean of this distribution:

$$\hat{\mathbf{f}} = \hat{\boldsymbol{\Lambda}}'\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}). \tag{6}$$

An alternative approach to the prediction of factor scores makes no explicit distributional assumptions, but attempts to find a linear function of $\mathbf{x}$ which minimizes the variance of the prediction error for each factor. This leads to (see [1, Section 3.5])

$$\hat{\mathbf{f}} = (\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}})^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x} - \overline{\mathbf{x}}). \tag{7}$$

Note that (6) can also be written

$$\hat{\mathbf{f}} = [\mathbf{I} + (\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}})]^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x} - \overline{\mathbf{x}}),$$

which is rather similar to (7) in its general form. Other methods are discussed in [3]. At times the factor scores are simplified further to produce a **cluster score**. This involves obtaining groups (or clusters) of variables and generating linear functions for each group (*see* **Cluster Analysis, Variables**; [8, Chapter 14]).

Choice of Number of Factors

The decision on how many factors, m, underlie the data can be a crucial one in the analysis. Consider

Table 3 Three- and four-factor solutions, based on principal components and varimax rotation, for data from [22]

	Three-factor solution			
Variable	Factor 1	Factor 2	Factor 3	Communality
x_1	0.815	0.131	0.182	0.714
x_2	0.805	0.254	0.069	0.717
x_3	0.623	0.335	0.308	0.596
x_4	0.632	0.347	0.001	0.520
x_5	0.775	0.113	0.292	0.698
x_6	0.216	0.818	−0.064	0.720
x_7	0.296	0.714	0.225	0.648
x_8	0.094	0.535	0.672	0.747
x_9	0.231	−0.014	0.847	0.771
x_{10}	0.310	0.646	0.470	0.734

	Four-factor solution				
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Communality
x_1	0.799	0.128	0.202	0.150	0.719
x_2	0.810	0.227	0.211	−0.032	0.753
x_3	0.587	0.330	0.291	0.242	0.596
x_4	0.423	−0.001	0.758	0.223	0.804
x_5	0.819	0.233	0.028	0.166	0.753
x_6	0.076	0.501	0.718	−0.099	0.783
x_7	0.335	0.751	0.213	−0.080	0.728
x_8	0.168	0.803	−0.029	0.352	0.797
x_9	0.174	0.192	0.078	0.911	0.903
x_{10}	0.281	0.682	0.328	0.288	0.734

Table 3, which shows rotated factors for our example, based on a principal component initial solution and varimax rotation, for $m = 3$ and $m = 4$ factors.

As m varies, so does the factor structure. For $m = 2$ (Table 2), the two factors corresponded roughly to the first and last five of the 10 variables. When $m = 3$, the association of the first five variables with a single factor (factor 1) becomes stronger. Factors 2 and 3 are mainly associated with (x_6, x_7, x_{10}) and (x_8, x_9), respectively, although x_8 and x_{10} have nontrivial contributions from both factors 2 and 3.

For $m = 4$, the four factors are associated chiefly with (x_1, x_2, x_5), (x_7, x_8, x_{10}), (x_4, x_6), and (x_9), respectively. The remaining variable, x_3, has nontrivial contributions from all four factors.

The choice of m can be made in a number of ways. One possibility is to examine solutions for several values of m, as we have done, and decide subjectively which gives the most clear-cut structure. With underfactoring (too few factors) the retained factors contain too many high loadings, and are not as "simple" as they might be. Conversely, overfactoring (too many factors) leads to factors which are fragmented and difficult to interpret meaningfully. It is arguable that, on this basis, $m = 3$ is the best choice in our example.

We could also use communalities to decide on m. If prior values are available for communalities, not necessarily the same for all variables, we can choose m to be the minimum number of common factors for which the estimated communalities are no less than the specified values for all variables. In our example, for $m = 2$, communalities for three of the variables are less than 0.5, which is perhaps rather low, although unavoidable when looking for common factors if any of the variables is largely independent of all others. Communalities for two of these three variables are increased substantially if m is increased to 3, and for $m = 4$ only x_3 has a communality less than 0.7.

Another approach to choosing m is to examine the eigenvalues of the correlation matrix $\mathbf{S}$. A well-established, but subjective, way of proceeding is to plot the eigenvalues $l_k, k = 1, 2, \ldots, p$, against

k. Joining these points gives the so-called *scree graph* due to Cattell [6]. On the graph the slope generally becomes less steep as k increases, and we look for an "elbow" where the decrease in slope is substantial. If the elbow occurs at $k = m + 1$, we choose to retain $m = k - 1$ factors. Individual eigenvalues can also be used to choose m. The most common rule (Kaiser's rule [14]) is to keep as many factors as there are eigenvalues greater than 1. The reasoning behind this rule is that individual standardized variables each have variance 1, and for a factor to be worth keeping it must explain more variation than this. The validity of this rule is blurred by the fact that the eigenvalues represent variances of principal components, not factors. Also, after rotation variation becomes more evenly distributed among factors than before rotation. Hence, each of a set of rotated factors may account for more variance than an individual variable even if unrotated factors corresponding to eigenvalues less than 1 have been retained.

In our example two eigenvalues exceed 1, but the third is only just below this threshold. The scree graph is unhelpful – it suggests taking $m = 1$.

A final type of approach to the choice of m is based on **hypothesis testing**. The scree graph is an informal way of looking for eigenvalues which are significantly larger than the remainder, the latter corresponding to "noise" rather than "real factors". More formally, if we make normality assumptions we can construct a **likelihood ratio test** of the null hypothesis that the covariance matrix $\mathbf{\Sigma}$ has the form given in (3) for a specified value of m, against the alternative of an unrestricted $\mathbf{\Sigma}$. The χ^2 approximation associated with this test statistic is often valid even when multivariate normality does not hold [3]. To decide on a suitable value for m we could conduct the test successively for $m = 1, 2, \ldots$ until we first fail to reject the null hypothesis, although the nonindependence of such a sequence of tests makes it difficult to assess the overall significance level of such a procedure. In addition, these procedures are influenced by the total sample size N. There is a tendency for underfactoring to occur for small N, with overfactoring for large N [8, p. 301].

This section is by no means exhaustive – for further discussion and methods, see [3, paragraphs 3.8, 3.13, Chapters 4 and 5], [8, Section 5.5], [9, Chapter 8], [13, Section 6.1] and [14, Section 2.4]. Although [13] is concerned with principal components analysis, many of the techniques described are more relevant to factor analysis.

Comparison with Principal Components Analysis

It was noted above that there is much confusion between principal components analysis and factor analysis. A number of differences between the two techniques have already been mentioned. Here we reiterate these briefly, and discuss some others.

1. Factor analysis postulates a model for the data – principal components analysis does not.
2. Factor analysis concentrates on explaining *covariance or correlation* by means of a few common factors. Principal components analysis is concerned mainly with explaining *variance*.
3. If the number of retained principal components is increased from m to $m + 1$, the first m principal components are unchanged. For rotated factors we have seen in our example that there can be substantial changes in *all* factors if m is changed.
4. Because principal components are defined as linear functions of $\mathbf{x}$, computation of principal component "scores" is unambiguous, unlike factor analysis.
5. Principal components analysis, like factor analysis, can be performed on standardized or on unstandardized variables, corresponding to an eigenanalysis of the correlation or covariance matrix, respectively. For principal components analysis the results of the two analyses are different, and one cannot be derived directly from the other. However, for maximum likelihood factor analysis the results are invariant in the sense that those based on the covariance matrix are equivalent to those derived from the correlation matrix.

Second-order Factor Analysis

When oblique rotation is used in factor analysis, the resulting factors are correlated. The correlation matrix between these (first-order) factors can be used as the input to a factor analysis, in the same manner as the correlation matrix for the original measured variables. This leads to second-order factor analysis (*see* **Factor Analysis, Second-order**).

Confirmatory Factor Analysis

What has been discussed above is exploratory in nature. There is no preconception of what form the factors will take – we let the data tell us this. Although this is what is usually meant by "factor analysis", it may not always be what is required. Sometimes there is some psychological or other theory which leads to a specified factor structure, in the sense that some loadings are known to be zero in this structure. **Confirmatory factor analysis** is concerned with estimating and testing the fit of such models. The technique is a special case of structural equations models, which have a large literature – see, for example, [4].

Further Reading

The multiple-factor approach to psychological data was first described at length in 1947 by Thurstone [21]. A number of substantial texts appeared in the next 40 years; the traditional, psychologist's, approach to the subject is well documented in [7]. Other traditionally based references are [8, 9], and [10]. Lawley & Maxwell [15] gave the first extensive account of the statistical, likelihood-based, approach, and a more recent account is in [3]. Chapter 17 of [11], and Chapter 13 of [15] provide good reviews, together with many references, of various aspects of factor analysis which are noted only briefly in this entry. Lewis-Beck [16] packages together five short monographs on factor analysis and related methods: the first two are readable introductory texts on exploratory factor analysis; the third covers principal components analysis; and the last two discuss confirmatory factor analysis and structural equations models.

References

[1] Bartholomew, D.J. (1987). *Latent Variable Models and Factor Analysis*. Griffin, London.

[2] Bartholomew, D.J. (1995). Spearman and the origin and development of factor analysis, *British Journal of Mathematical and Statistical Psychology* **48**, 211–220.

[3] Bartholomew, D. & Knott, M. (1995). *Latent Variable Models and Factor Analysis*, 2nd Ed. Arnold, London.

[4] Bentler, P.M. & Stein, J.A. (1992). Structural equation models in medical research, *Statistical Methods in Medical Research* **1**, 159–181.

[5] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

[6] Cattell, R.B. (1966). The scree test for the number of factors, *Journal of Multivariate Behavioral Research* **1**, 245–276.

[7] Cattell, R.B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. Plenum Press, New York.

[8] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[9] Gorsuch, R.L. (1983). *Factor Analysis*, 2nd Ed. Lawrence Erlbaum, Hillsdale.

[10] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd rev. Ed. University of Chicago Press, Chicago.

[11] Jackson, J.E. (1991). *A User's Guide to Principal Components*. Wiley, New York.

[12] Jennrich, R.I. & Sampson, P.F. (1966). Rotation for simple loadings, *Psychometrika* **31**, 313–323.

[13] Jolliffe, I.T. (2002). *Principal Component Analysis*, 2nd Ed. Springer, New York.

[14] Kaiser, H.F. (1960). The application of electronic computers to factor analysis, *Educational and Psychological Measurement* **20**, 141–151.

[15] Lawley, D.N. & Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*, 2nd Ed, 1971. Butterworths, London.

[16] Lewis-Beck, M.S. (1994). *Factor Analysis and Related Techniques*. Sage, London.

[17] Morrison, D.F. (1976). *Multivariate Statistical Methods*, 2nd Ed. McGraw-Hill, Tokyo.

[18] Rencher, A.C. (1995). *Methods of Multivariate Analysis*. Wiley, New York.

[19] Spearman, C. (1904). "General intelligence" objectively determined and measured, *American Journal of Psychology* **5**, 201–293.

[20] Thurstone, L.L. (1931). Multiple factor analysis, *Psychological Review* **38**, 406–427.

[21] Thurstone, L.L. (1947). *Multiple-Factor Analysis*. University of Chicago Press, Chicago.

[22] Yule, W., Berger, M., Butler, S., Newham, V. & Tizard, J. (1969). The WPPSI: an empirical evaluation with a British sample, *British Journal of Educational Psychology* **39**, 1–13.

(*See also* **Classification, Overview**; **Matrix Algebra**; **Multidimensional Scaling**)

IAN T. JOLLIFFE

Factor Analysis, Second-order

In **factor analysis** we factor a reduced **correlation** matrix $\mathbf{R}^*$ to obtain an initial factor matrix $\mathbf{F}$, which in turn is often *rotated* by means of a transformation matrix $\mathbf{\Lambda}$ to produce a rotated factor matrix $\mathbf{V}$. $\mathbf{R}^*$ is obtained by replacing the estimated **communalities** on the diagonal of the original correlation matrix $\mathbf{R}$. The elements of $\mathbf{F}$ and $\mathbf{V}$ are called loadings (*see* **Factor Loading Matrix**). They are usually correlations of the original variables and the factors. From the high loadings, we interpret the factors of $\mathbf{V}$ in terms of the meanings of the functions measured by the original variables. This analysis is a *first-order factor analysis*. If an **oblique rotation** is used to transform $\mathbf{F}$ to $\mathbf{V}$, the factors of $\mathbf{V}$ are correlated in general. With the correlation of these factors, we may perform a *second-order factor analysis* to push the interpretations a step further.

In second-order factor analysis we first obtain the second-order correlation matrix $\mathbf{R}_s$, which is given by

$$\mathbf{R}_s = \mathbf{T}'\mathbf{T},$$

where $\mathbf{T} = (\mathbf{\Lambda}')^{-1}\mathbf{D}$ is the primary structure transformation matrix. $\mathbf{\Lambda}$ is the matrix that is used to transform the initial factor matrix $\mathbf{F}$ to the first-order factor matrix $\mathbf{V}$. $\mathbf{D}$ is the diagonal matrix that normalizes the columns of $(\mathbf{\Lambda}')^{-1}$. The matrix $\mathbf{R}_s$ contains the correlations among the primary axes of the first-order factor analysis. From $\mathbf{R}_s$, we obtain the reduced second-order correlation matrix $\mathbf{R}_s^*$ (with estimated communalities on the diagonal). We can then factor this $\mathbf{R}_s^*$ to obtain $\mathbf{F}_s$, the second-order initial factor matrix. We can next rotate $\mathbf{F}_s$ to obtain $\mathbf{V}_s$, the second-order rotated factor matrix. Again, the factors of $\mathbf{V}_s$ can be interpreted from the high loadings in terms of the meanings of the functions represented by the first-order factors.

We interpret the first-order factors directly in terms of the original variables. The original variables may have errors of measurement as well as specific factors. These error factors and specific factors lump together in the unique factors and lie outside the first-order common-factor space in which the first-order factors (or factors of $\mathbf{V}$) are defined. Therefore, when we interpret the second-order factors directly in terms of these first-order factors, any error factors existing in this analysis can be due only to imperfect first-order analysis. Usually, the error factors in the second-order analysis are negligible and we can conclude that the second-order unique factors are entirely specific factors which lie outside the second-order common-factor space.

Two First-order Factors

In general, we do not perform a second-order factor analysis when there are only two first-order factors because there is usually very little information this analysis can add to the first-order analysis.

Three First-order Factors

In the case of having three first-order factors, there are three linearly independent correlations in the second-order correlation matrix $\mathbf{R}_s$, namely, r_{12}, r_{13}, and r_{23}. Using the theorems on triads, if these correlations can be accounted for by a factor pattern with one general factor and three unique factors, then all the triads computed from these correlations must be positive or zero and not greater than unity (see [1, Chapter 1, Section 1.5.5] for more related theorems on triads). On the Basis of the theorems, we can perform the second-order analysis by first computing the triads:

$$t_{123} = \frac{r_{12}r_{13}}{r_{23}},$$

$$t_{213} = \frac{r_{12}r_{23}}{r_{13}},$$

$$t_{312} = \frac{r_{13}r_{23}}{r_{12}}.$$

If all the computed triads are at least 0.001 and not greater than 0.999, then these triads correspond to the communalities of the second-order factors. They are usually denoted by h_j^2, where $j = 1, 2, \ldots, m$ and m is the number of the first-order factors. The second-order general-factor loadings, g_j, are the square roots of the communalities, and the specific-factor loadings are $s_j = (1 - h_j^2)^{1/2}$ for $j = 1, 2, 3$. For each factor j, $g_j + s_j = 1$.

To make this second-order analysis possible, we not only need to have all the triads to be within the specified range, we also need to have a very

good first-order analysis with additional refinement of the **simple structure** on the reference vectors (*see* **Primary Factors**). (See [1, Chapter 10] for various refinement transformations.)

Four First-order Factors

In the case of having four first-order factors, there are six second-order correlations, namely r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, and r_{34}. If these correlations can be accounted for by a factor pattern with one general factor and four unique factors, we may proceed to calculate the 12 triads and perform the triad analysis as follows:

$$h_1^2 = \frac{(t_{123} + t_{124} + t_{134})}{3} = \frac{(r_{12}r_{13}/r_{23} + r_{12}r_{14}/r_{24} + r_{13}r_{14}/r_{34})}{3},$$

$$h_2^2 = \frac{(t_{213} + t_{214} + t_{234})}{3} = \frac{(r_{12}r_{23}/r_{13} + r_{12}r_{24}/r_{14} + r_{23}r_{24}/r_{34})}{3},$$

$$h_3^2 = \frac{(t_{312} + t_{314} + t_{324})}{3} = \frac{(r_{13}r_{23}/r_{12} + r_{13}r_{34}/r_{14} + r_{23}r_{34}/r_{24})}{3},$$

$$h_4^2 = \frac{(t_{412} + t_{413} + t_{423})}{3} = \frac{(r_{14}r_{24}/r_{12} + r_{14}r_{34}/r_{13} + r_{24}r_{34}/r_{23})}{3}.$$

Then we can obtain the general factor loadings, $g_j = \sqrt{h_j^2}$, and the specific factor loadings, $s_j = (1 - h_j^2)^{1/2}$ for $j = 1, 2, 3, 4$.

To find out if one general factor is sufficient to account for the second-order correlations, we have to consider the standard errors of the tetrads. The tetrads are computed as:

$$t_{1234} = r_{12}r_{34} - r_{13}r_{24},$$

$$t_{1243} = r_{12}r_{34} - r_{14}r_{23},$$

$$t_{1342} = r_{13}r_{24} - r_{14}r_{23},$$

where $t_{1234} + t_{1342} = t_{1243}$ and a conservative approximation to their standard errors is given by $1/(N-1)^{1/2}$, where N is the total sample size. If the absolute values of all three tetrads are less than $1/(N-1)^{1/2}$, we can tentatively assume that there is only one second-order general factor and proceed to calculate the triads and the factor loadings as described above. If there are any triads that are lower than $-1/(N-1)^{1/2}$ or higher than $1 + 1/(N-1)^{1/2}$ and also if there are any h_j^2 that are less than 0.001 or greater than 0.999, then we have to consider that only one second-order general factor is insufficient.

If one second-order general factor is not sufficient, then we have to use a more complicated procedure to postulate two second-order factors. For postulating one second-order general factor and one second-order group factor, we can use the triads analysis given in [1, Chapter 1, Sections 1.5.10 and 1.6.2–1.6.7]. For postulating two second-order general factors, we can employ the principal-axes method and factor the second-order correlation matrix twice to make the sum of the final communalities close to the sum of the estimated communalities for the second analysis, and then to rotate the second-order factor matrix obliquely.

Five or More First-order Factors

In the case of having five or more first-order factors, we can use principal axes to factor the reduced second-order correlation matrix $\mathbf{R}_s^*$. If there appears to be only one salient second-order factor, we will refactor using the squares of the first-factor loadings from the first factoring as the communality estimates. The first factor of the second factoring is then the second-order general factor, with loadings g_j, $j = 1, 2, \ldots, m$, and m is the number of salient first-order factors. The second-order specific factor loadings are $s_j = (1 - g_j^2)^{1/2}$, and for each factor, $g_j^2 + s_j^2 = 1$. The second-order matrix $\mathbf{G}$ is then formed as

$$\begin{bmatrix} g_1 & h_1^2 & s_1 \\ g_2 & h_2^2 & s_2 \\ \vdots & \vdots & \vdots \\ g_m & h_m^2 & s_m \end{bmatrix}.$$

When there is more than one salient second-order factor, with the number less than half of the first-order factors, we refactor using the row sums of squares of the first $\mathbf{G}$ as the estimated communalities. The resulting second $\mathbf{G}$ may be rotated to obtain a second-order reference-structure $\mathbf{V}$ matrix. Interpretation of

this matrix must be based on quite clear interpretation of the meanings of the first-order factors.

In second-order analysis, one is likely to obtain communality greater than unity. This is called a Heywood case. If this happens, we may try increasing or decreasing the number of second-order factors by one.

Reference

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Factor Loading Matrix

A factor loading matrix is a matrix of coefficients (or weights) for a set of linear equations relating p observed variables to m factors (or components) (*see* **Factor Analysis, Overview**; **Principal Components Analysis**). Sometimes, it is referred to as a *factor pattern*. The rows of the matrix correspond to the observed variables and the columns correspond to the factors. Consider the hypothetical factor pattern matrix, **F** obtained from a common factor model (see Table 1). The a_i and b_i are the coefficient from the following set of specification equations:

$$X_1 = a_1 A + b_1 B + u_1 U_1,$$
$$X_2 = a_2 A + b_2 B + u_2 U_2,$$
$$X_3 = a_3 A + b_3 B + u_3 U_3,$$
$$X_4 = a_4 A + b_4 B + u_4 U_4,$$
$$X_5 = a_5 A + b_5 B + u_5 U_5,$$

where X_i is the standardized score of the original variable i, a_i and b_i are the coefficients for the common factors A and B of variable i, u_i is the coefficient for the unique factor U_i of variable i. Under this model, we usually call **F** the initial common-factor loadings matrix. Each pair of these loadings (or coefficients) can be expressed as the coordinates of a point in a plane. Figure 1(a) presents a plot of the factor pattern loadings for a variable X with respect to the initial unrotated factor axes a and b. The a coordinate is measured parallel to the axis a from the axis b and the b coordinate is measured parallel to the axis b from the axis a.

The term "factor loading matrix" is also used to refer to the matrix that contains the **correlations** between the variables and the factors. This matrix is then called a *factor structure*. It is obtained by postmultiplying the initial factor matrix, **F**, with a matrix the columns of which represent the coordinates of a set of unit axes. These axes can be a set of primary axes or reference vectors (*see* **Primary Factors**). The row vectors of the resulting factor structure are the correlations between the variables and the factors defined by these axes. In geometrical terms, these structure loadings can be thought of as the projections of the variable point i on the axes that define the factors. They are measured as the distances from

Table 1 Factor pattern matrix, **F**

		Factor	
		a	b
Variable	1	a_1	b_1
	2	a_2	b_2
	3	a_3	b_3
	4	a_4	b_4
	5	a_5	b_5

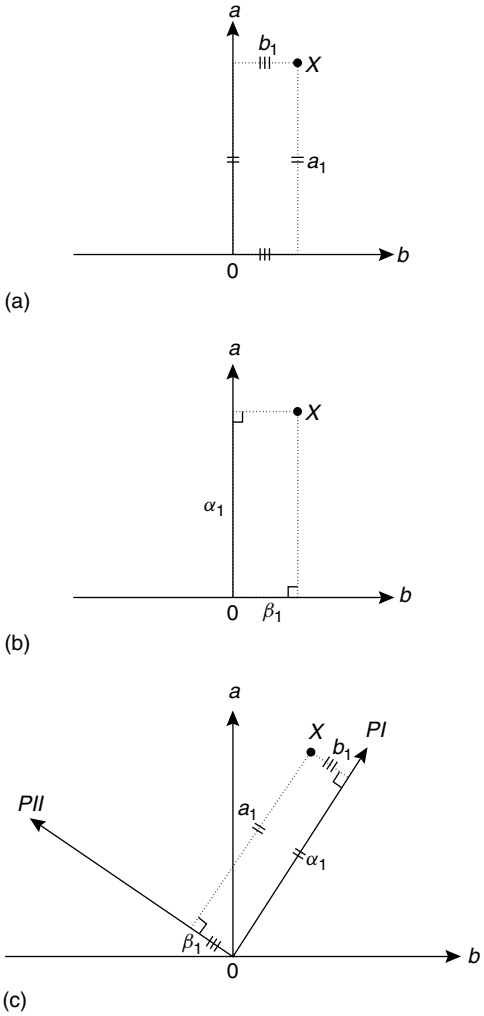

Figure 1 (a) Initial pattern loadings with respect to the initial axes a and b. (b) Initial structure loadings with respect to the initial axes a and b. (c) Orthogonal rotated pattern and structure loadings with respect to the primary axes PI and PII

the origin to the intersection points of the axes and the perpendicular lines dropping from the variable point i. A plot of the structure loadings for a variable X with respect to the initial unrotated axes a and b is presented in Figure 1(b). The pattern and structure loadings are identical in Figures 1(a) and 1(b). In fact, these two types of loadings coincide for any orthogonal rotation of the factor axes. A plot illustrating both the pattern and structure of X with respect to a pair of **orthogonal** primary axes PI and PII is shown in Figure 1(c). As long as we assume that the factors are uncorrelated (orthogonal), there is no confusion in the use of the term "factor loading matrix", because both the pattern and structure matrices provide the same set of loadings.

However, the pattern and structure loadings are not identical after an **oblique rotation**. In this situation, we need to have an explicit designation to indicate if the factor loading matrix is a pattern or a structure. Figure 2 presents plots of the pattern and structure loadings of a variable X with respect to two oblique-rotated primary axes PI and PII. The dotted lines show the projections of the point X on the primary axes.

In Figure 2(a), the vectors $0c$ and $0d$ are the parallel projections on the primary axes PI and PII. They are the **primary factor** pattern loadings, which are the regression coefficients of the variables on the factors. The primary pattern loadings are interpretable as measures of the contribution of each factor to the variances of the variables. The formula to obtain this primary factor pattern matrix $\mathbf{P}$ is given by $\mathbf{P} = \mathbf{F}\mathbf{\Lambda}\mathbf{D}^{-1}$, where $\mathbf{F}$ is the initial factor matrix, $\mathbf{\Lambda}$ is the matrix the column vectors of which represent the coordinates of the reference axis, and $\mathbf{D}$ is the matrix of correlations between the reference vectors and the primary axes. (See Cureton & D'Agostino [1, Chapter 6] for the derivations of each transformation matrix. Also see below for clarification of reference vectors.) This primary factor pattern matrix is useful in the interpretation of the factors because the variables that are highly loaded on one factor usually have low loadings on other factors. We can identify a cluster of similar variables in terms of their coefficients. Unlike correlation coefficients, these coefficients do not lie within ± 1. It is possible for these coefficients to be greater than 1 even when the variables are standardized.

The vectors $0e$ and $0f$ of Figure 2(b) are the perpendicular projections on the primary axes PI and PII. They are the primary structure loadings, which are the correlation coefficients between the variable X and the oblique primary factors if X is a standardized variable. This structure matrix, $\mathbf{S}$, on the primary factors is defined as $\mathbf{S} = \mathbf{FT} = \mathbf{F}(\mathbf{\Lambda}')^{-1}\mathbf{D}$. This matrix need not conform to the **simple structure** principle, so it may not be useful for interpretation. However, it may be used in the determination of **factor scores**.

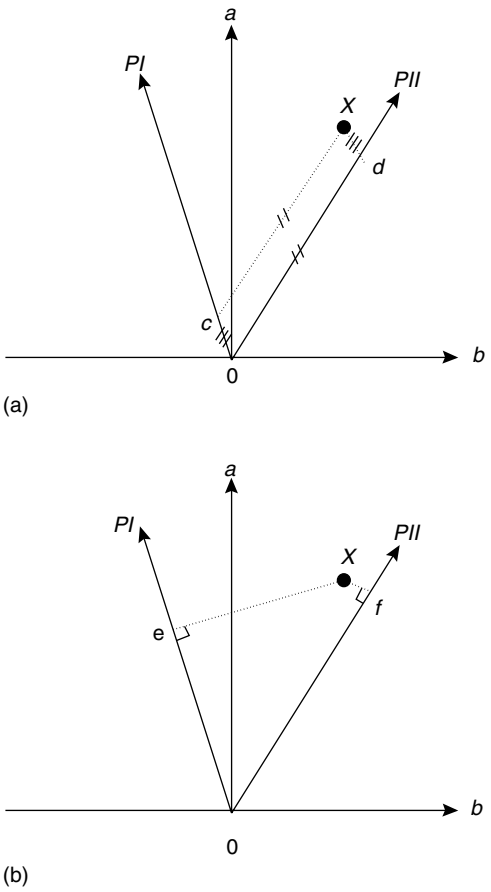

Figure 2 (a) Oblique-rotated pattern loadings with respect to the primary axes PI and PII. (b) Oblique-rotated structure loadings with respect to the primary axes PI and PII

With an oblique rotation, an alternative set of axes may be employed. They are the reference vector axes, which are defined as axes normal to planes or hyperplanes, depending on the number of factors involved, and they usually set a boundary for the primary factor axes unless the angle between the primary factor axes is obtuse (i.e. a rare case). In

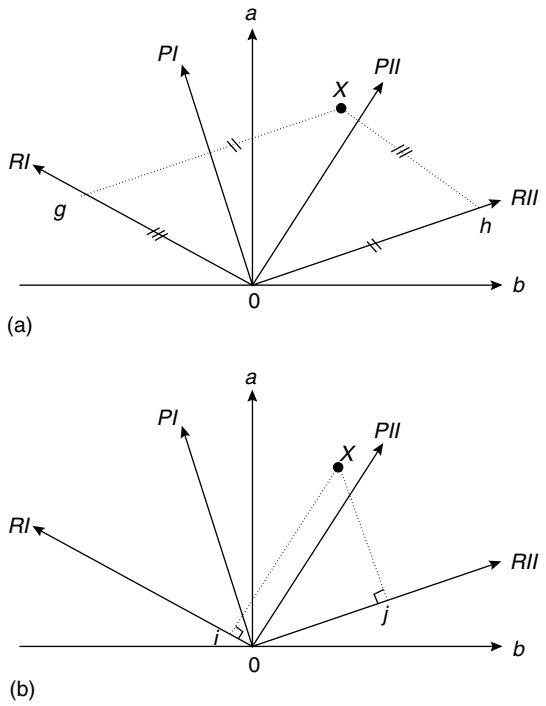

Figure 3 (a) Oblique-rotated pattern loadings with respect to the reference axes RI and RII. (b) Oblique-rotated structure loadings with respect to the reference axes RI and RII

the two-factor case, the reference axes are simply the axes that are orthogonal to the primary factor axes. Plots of the pattern and structure loadings of a variable X with respect to the reference axes RI and RII are shown in Figure 3. The dotted lines show the projections of the point X on the reference axes. Note that since RI is the reference vector for PI, it is correlated with PI and uncorrelated or orthogonal to PII. Similarly, RII is correlated with PII and uncorrelated with PI.

The vectors $0g$ and $0h$ of Figure 3(a) are the parallel projections on the reference vectors RI and RII. They are the reference vector pattern loadings. Figure 3(a) shows that these loadings are in fact proportional to the primary factor structure loadings. The complete pattern matrix on these reference vectors is given by $\mathbf{W} = \mathbf{F}(\mathbf{\Lambda}')^{-1}$. This matrix is seldom used for interpretation.

The vectors $0i$ and $0j$ of Figure 3(b) are the perpendicular projections on the reference vectors RI and RII. They are the reference vector structure loadings. Figure 3(b) shows that these loadings are proportional to the primary factor pattern loadings. These structure loadings usually reflect a simple structure solution to a higher degree than do the loadings on the primary scale [2]. The values of these structure loadings are constrained to lie within ± 1. $\mathbf{V} = \mathbf{F}\mathbf{\Lambda}$ gives the structure matrix on the reference vectors. It is useful for interpretation of factors.

For simplicity and visual convenience, we have focused the discussion of the factor loading matrices in terms of a two-dimensional model throughout this article. The reader should refer to Cureton & D'Agostino [1] for further discussion using models with higher dimension.

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: an Applied Approach*. Lawrence Erlbaum, Hillsdale.

[2] Thurstone, L.L. (1947). *Multiple Factor Analysis*. University of Chicago Press, Chicago.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Factor Scores

After a set of interpretable **factor loadings** is determined, the next possible step is to compute the factor scores. Exact factor scores can be obtained for a principal components analysis model. These scores are usually referred to as component scores. Methods for computing these component scores are given in **Principal Components Analysis**. In this article we present primarily methods that compute factor scores in the common factor analysis model. Under this model, it is not possible to obtain the exact factor scores because we have more unknowns (m common factors and p unique factors) than the observed variables (of size p). We can only estimate the scores on the m common factors in this case. There are various methods that can be employed to obtain the estimates, but the most common method to estimate the factor scores is least squares regression.

The **linear regression** of any factor score on the p standardized original variables is expressed as follows:

$$\mathbf{Z} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\eta}, \tag{1}$$

where $\mathbf{Z}$ is the $m \times N$ matrix of true (unknown) standardized factor scores, $\mathbf{X}$ is the $p \times N$ matrix of standardized original variables, $\boldsymbol{\beta}$ is the $m \times p$ matrix of true regression coefficients, and $\boldsymbol{\eta}$ is the $m \times N$ matrix of residuals. N is the sample size. Each row of $\mathbf{Z}$ of (1) represents one primary factor, and we can solve the scores on each factor separately by minimizing for that row the sum of squares of the residuals. The resulting least squares estimates for the factor scores are given by

$$\hat{\mathbf{Z}} = \boldsymbol{\Delta}\boldsymbol{\Lambda}^{-1}\mathbf{F}'\mathbf{R}^{-1}\mathbf{X}, \tag{2}$$

where $\boldsymbol{\Delta}$ is the $m \times m$ diagonal matrix that normalizes the columns of $(\boldsymbol{\Lambda}^{-1})'$, $\boldsymbol{\Lambda}$ is the $m \times m$ transformation matrix, $\mathbf{F}$ is the $p \times m$ initial factor loading matrix, $\mathbf{R}$ is the $p \times p$ correlation matrix of the original variables, and $\mathbf{X}$ is the matrix of standardized original variables. An alternative formula to (2) is given by

$$\hat{\mathbf{Z}} = \mathbf{R}_s\mathbf{F}'\mathbf{R}^{-1}\mathbf{X}. \tag{3}$$

This formula can be used if $\mathbf{P}$ the primary pattern matrix and $\mathbf{R}_s$ the second-order correlation matrix are available (*see* **Factor Analysis, Second-order**). We can also get the standardized weights for these factor scores from the SAS software package PROC FACTOR [3]. An example of these standardized weights is given in Table 1. These weights are obtained from performing a factor analysis with varimax rotation on the **Framingham** depression data (*see* **Principal Components Analysis** for data description). For a complete discussion on this regression method, readers can refer to [1] for details.

Table 1 Factor scoring coefficients for retained components for Framingham depression data
Rotation method: varimax
Standardized scoring coefficients

	Factor 1	Factor 2	Factor 3
EFFORT	−0.043	0.321	0.034
RESTLESS	−0.006	0.137	0.005
DEPRESS	0.275	0.048	0.059
HAPPY	0.195	0.131	−0.084
LONELY	0.222	−0.055	0.014
UNFRIEND	−0.029	−0.018	0.349
ENJOYLIF	0.148	0.012	−0.042
FELTSAD	0.314	−0.073	0.024
DISLIKED	−0.053	0.007	0.356
GETGOING	−0.062	0.326	0.050

Even though $\hat{\mathbf{Z}}$ contains good estimates for the true common factor scores, they are still subject to error. A measure of the deviation of the estimates from the true scores is the multiple **correlation** of the estimated factor scores with the p variables of the data. These squared multiple correlations $\mathbf{R}_j^2$, $j = 1, \ldots, m$, can be obtained as the diagonal elements of

$$(\boldsymbol{\Delta}\boldsymbol{\Lambda}^{-1}\mathbf{F}')\mathbf{R}^{-1}(\boldsymbol{\Delta}\boldsymbol{\Lambda}^{-1}\mathbf{F}')'.$$

$\mathbf{R}_j^2$ is shown to be the variance of the estimated factor scores. Since the variance is not equal to unity, these estimates leave parts of the total variance (the total unique variance) unaccounted for. The standard error of estimate for a set of primary factor scores can be computed as

$$\text{se}_j = (1 - \mathbf{R}_j^2)^{1/2}.$$

The higher the $\mathbf{R}_j$, the lower the standard error of estimate. This method maximizes the validity of the factor scores (i.e. it gives the highest $\mathbf{R}_j$).

The drawback of this method is that the regression estimates are not univocal: the correlations of these estimated factor scores are not the same as the correlations between the primary factors. In particular, these scores are correlated even when the primary factors are not. A number of other estimation procedures have been proposed to improve on the criteria of univocality and **orthogonality** of these regression estimates (see [2, Chapter 16]), but none of these methods has the level of validity provided by the regression method.

In general, there is limited value in the use of factor scores, for even though the factors are interpretable, the estimated factor scores are not well determined. Cureton & D'Agostino [1] suggest the use of the **cluster scores** as the more stable score to be used as a summary measure.

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[2] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd Rev. Ed. University of Chicago Press, Chicago.

[3] SAS Institute, Inc. (1990). *SAS/STAT User's Guide, Release 6.04*, 4th Ed. SAS Inc., Cary.

RALPH B. D'AGOSTINO, SR & HEIDY K. RUSSELL

Factor

Factor analysis is a set of procedures that use mathematical models to explain the interrelationships of a set of manifest (i.e. observed) variables by a smaller number of underlying *factors* that cannot be observed or measured directly. These underlying factors are sometimes known as the latent variables. In 1927, Spearman [4] developed the common factor analysis model for two factors and Thurstone [5] later extended the model to multiple factors. The Spearman–Thurstone approach expresses a variable as a weighted sum of some unknown *common factors* and a *unique factor*. A *factor pattern* is the set of equations relating the measurements on p variables to the m postulated common factors ($m < p$) and p unique factors. A general factor pattern is given as follows:

$$\begin{aligned} X_1 &= a_1A + b_1B + \cdots + m_1M + u_1U_1, \\ X_2 &= a_2A + b_2B + \cdots + m_2M + u_2U_2, \\ &\vdots \qquad\qquad\qquad \vdots \\ X_p &= a_pA + b_pB + \cdots + m_pM + u_pU_p, \end{aligned} \tag{1}$$

where: $X_1, X_2, \ldots, X_p$ represent measurements (usually the standardized measurements) of the p manifest variables; $A, B, \ldots, M$ represent the standard scores on the m common factors; $a_i, b_i, \ldots, m_i$, $i = 1, \ldots, p$, are the weights or *common* **factor loadings** of the p variables on the m common factors; $U_1, U_2, \ldots, U_p$ represent the standard scores on the p unique factors; and $u_1, u_2, \ldots, u_p$ are the weights or *unique factor loadings* of the p variables on the p unique factors. In many settings the common factor loadings are also the correlations of the manifest variables and the common factors.

The common factors are interpreted with reference to the p manifest variables. The unique factor in each variable is merely whatever part of that variable is uncorrelated with all the other variables, including its error of measurement. The p unique factors are always taken to be uncorrelated with one another and with the common factors. These common factors may or may not be uncorrelated.

The factors or latent variables are not uniquely determined by the intercorrelations among the manifest variables. The factors are only hypothetical constructs that are postulated in order to arrive at an explanation of the intercorrelations of the original set of variables. Once the number of factors m and the final factor pattern model in (1) is determined, the factor complexity is used to interpret the factors. Factor *complexity* refers to the number of manifest variables that have moderate or high factor loadings (or weights) on a factor. What an investigator deems moderate or high depends on the assessment of error in their data, and the overall intercorrelation between the manifest variables and the findings of other similar studies. The various approaches of factor analysis can generate different configurations of factor complexity. So, it is important to decide upon the appropriate procedure for performing the factor analysis. Standard procedures are common factor analysis [2] and maximum likelihood factor analysis [1, 3]. Other procedures are also well developed [1, 3] (also *see* **Factor Analysis, Overview**). Within the concept of factor complexity, different types of factors can be identified. If a common factor has moderate or high loadings on all p manifest variables, it is called a *general factor*. If it has moderate or high loadings on only two or more, but not all, of the variables, then it is called a *group factor*. If a group factor has moderate or high loadings on only two variables, it is termed a *doublet factor*. If the loadings of the group factor are such that there are both high positive and negative loadings, it is called a *bipolar factor*. These different types of factors are illustrated in Table 1.

Table 1

	Factor			
Variable	1	2	3	4
1	+	0	+	0
2	+	+	+	0
3	+	0	0	0
4	+	0	+	0
5	+	+	0	0
6	+	+	0	0
7	+	0	−	0
8	+	0	−	0
9	+	0	0	+
10	+	+	0	+
	general		bipolar	doublet
		group		

Note: + = high or moderate positive loadings.
− = high or moderate negative loadings.
0 = near-zero loadings.

In Table 1, factor 1 is a general factor, while factors 2, 3, and 4 are group factors. Among the three group factors, factor 3 is a bipolar factor and factor 4 is a doublet factor.

One traditional aim in factor analysis is to obtain a **simple structure**, where the m factors are group factors, each highly or moderately correlated to only approximately m manifest variables.

References

[1] Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley, New York.

[2] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[3] Reyment, R.A. & Jöreskog, K.G. (1993). *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, Cambridge.

[4] Spearman, C. (1927). *The Abilities of Man*. Macmillan, New York.

[5] Thurstone, L.L. (1935). *Multiple-Factor Analysis*. University of Chicago Press, Chicago.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Factorial Designs in Clinical Trials

Factorial experiments test the effect of more than one treatment (factor) using a design that permits an assessment of **interactions** between the treatments. A treatment could be either a single therapy or a combination of interventions. The essential feature of factorial designs is that treatments are varied systematically (i.e. some groups receive more than one treatment), and the experimental groups are arranged in a way that permits testing whether or not the treatments interact with one another.

The technique of varying more than one treatment in a single study has been used widely in agriculture and industry based on work by **Fisher** [10, 11] and **Yates** [33]. Influential discussions of factorial experiments were given by Cox [8] and **Snedecor** & **Cochran** [28]. Factorial designs have been used relatively infrequently in medical trials, except recently in disease prevention studies (*see* **Prevention Trials**). The discussion here will be restricted to randomized factorial **clinical trials**.

Factorial designs offer certain advantages over conventional comparative designs, even those employing more than two treatment groups. The factorial structure permits certain comparisons to be made that cannot be achieved by any other design. In some circumstances, two treatments can be tested in a factorial trial using the same number of subjects ordinarily used to test one treatment. However, the limitations of factorial designs must be understood before deciding whether or not they are appropriate for a particular therapeutic question. Additional discussions of factorial designs in clinical trials can be found in **Byar** & Piantadosi [6] and Byar et al. [7]. For a discussion of such designs related to cardiology trials, particularly in the context of the ISIS-4 trial [17], see Lubsen & Pocock [20]. This article is based on a recent chapter discussing factorial designs in medical studies given by Piantadosi [24].

Basic Features of Factorial Designs

The simplest factorial design has two treatments (A and B) and four treatment groups (Table 1). There might be n patients entered into each of the four treatment groups for a total sample size of $4n$ and a balanced design. One group receives neither A nor B, a second receives both A and B, and the other two groups receive one of A or B. This is called a 2×2 (two by two) factorial design. The design generates enough information to test the effects of A alone, B alone, and A plus B.

The 2×2 design generalizes to higher order factorials. For example, a design studying three treatments, A, B, and C, is the $2 \times 2 \times 2$. Possible treatment groups for this design are shown in Table 2. The total sample size is $8n$ if all treatment groups have n subjects.

These examples highlight some of the prerequisites necessary for, and restrictions on, using a factorial trial.

First, the treatments must be amenable to being administered in combination without changing dosage in the presence of each other. For example, in Table 1, we would not want to reduce the dose of A in the lower right cell where B is present. This requirement implies that the side effects of the treatments cannot be cumulative to the point where the combination is impossible to administer.

Secondly, it must be ethically acceptable to withhold the individual treatments, or administer them at lower doses as the case may be (*see* **Ethics of Randomized Trials**). In some situations, this means

Table 1 Treatment groups and sample sizes in a 2×2 balanced factorial design

	B		
A	No	Yes	Total
No	n	n	$2n$
Yes	n	n	$2n$
Total	$2n$	$2n$	$4n$

Table 2 Treatment groups in a balanced $2 \times 2 \times 2$ factorial design.

	Treatments			Sample
Group	A	B	C	size
1	No	No	No	n
2	Yes	No	No	n
3	No	Yes	No	n
4	No	No	Yes	n
5	Yes	Yes	No	n
6	No	Yes	Yes	n
7	Yes	No	Yes	n
8	Yes	Yes	Yes	n

having a no-treatment or placebo group in the trial. In other cases A and B may be administered in addition to a "standard" so that all groups receive some treatment.

Thirdly, we must be genuinely interested in learning about treatment combinations; otherwise, some of the treatment groups might be unnecessary. Alternately, to use the design to achieve greater efficiency in studying two or more treatments, we must know that some interactions do not exist.

Fourthly, the therapeutic questions must be chosen appropriately. We would not use a factorial design to test treatments that have exactly the same mechanisms of action (e.g. two ACE inhibitors for high blood pressure) because either would answer the question. Treatments acting through different mechanisms would be more appropriate for a factorial design (e.g. radiotherapy and chemotherapy for tumors). In some prevention factorial trials, the treatments tested also target different diseases.

Efficiency

Factorial designs offer certain very important efficiencies or advantages when they are applicable. Consider the 2×2 design and the estimates of treatment effects that would result using an **additive model** for analysis (Table 3). Assume that the responses are group averages of some **normally distributed** response denoted by $\overline{Y}$. The subscripts on $\overline{Y}$ indicate which treatment group it represents. Half the patients receive one of the treatments (this is also true in higher order designs). For a moment, further assume that the effect of A is not influenced by the presence of B.

There are two estimates of the effect of treatment A compared to placebo in the design, $\overline{Y}_A - \overline{Y}_0$ and $Y_{AB} - \overline{Y}_B$. If B does not modify the effect of A, the two estimates can be combined (averaged) to estimate the overall effect of $A(\beta_A)$,

$$\beta_A = \frac{(\overline{Y}_A - \overline{Y}_0) + (\overline{Y}_{AB} - \overline{Y}_B)}{2}. \qquad (1)$$

Table 3 Treatment effects from a 2×2 factorial design

	B	
A	No	Yes
No	$\overline{Y}_0$	$\overline{Y}_B$
Yes	$\overline{Y}_A$	$\overline{Y}_{AB}$

Similarly,

$$\beta_B = \frac{(\overline{Y}_B - \overline{Y}_0) + (\overline{Y}_{AB} - \overline{Y}_A)}{2}. \qquad (2)$$

Thus, in the absence of interactions (i.e. the effect of A is the same with or without B, and vice versa), the design permits the full sample size to be used to estimate two treatment effects.

Now suppose that each patient's response has a **variance** σ^2 that is the same in all treatment groups. We can calculate the variance of β_A to be

$$\text{var}(\beta_A) = \frac{1}{4} \times \frac{4\sigma^2}{n} = \frac{\sigma^2}{n}.$$

This is the same variance that would result if A were tested against placebo in a single two-armed comparative trial with $2n$ patients in each treatment group. Similarly,

$$\text{var}(\beta_B) = \frac{\sigma^2}{n}.$$

However, if we tested A and B in separate trials, we would require $4n$ subjects in each trial or a total of $8n$ patients to have the same precision. Thus, in the absence of interactions, factorial designs estimate main effects efficiently. In fact, tests of both A and B can be conducted in a single factorial trial with the same precision as two single-factor trials using twice the sample size.

Interactions

The effect of A might be influenced by the presence of B (or vice versa). In other words, there might be a *treatment interaction*. Some of the efficiencies just discussed will be lost. However, factorial designs are even more relevant when interactions are possible. Factorial designs are the only type of trial design that permits study of treatment interactions. This is because the design has treatment groups with all possible combinations of treatments, allowing the responses to be compared directly.

Consider again the two estimates of A in the 2×2 design, one in the presence of B and the other in the absence of B. The definition of an interaction is that the effect of A in the absence of B is different from the effect of A in the presence of B. This can be estimated by comparing

$$\beta_{AB} = (\overline{Y}_A - \overline{Y}_0) - (\overline{Y}_{AB} - \overline{Y}_B) \qquad (3)$$

to zero. If β_{AB} is near zero, we would conclude that no interaction is present. It is straightforward to verify that $\beta_{AB} = \beta_{BA}$. When there is an AB interaction present, we must modify our interpretation of the main effects. For example, the estimates of the main effects of A and B [(1) and (2)] assumed no interaction was present. We may choose to think of an overall effect of A, but recognize that the magnitude (and possibly the direction) of the effect depends on B. In the absence of the other treatment, we could estimate the main effects using

$$\beta'_A = (\overline{Y}_A - \overline{Y}_0) \tag{4}$$

and

$$\beta'_B = (\overline{Y}_B - \overline{Y}_0). \tag{5}$$

In the $2 \times 2 \times 2$ design, there are three main effects and four interactions possible, all of which can be tested by the design. Following the notation above, the effects are

$$\beta_A = \tfrac{1}{4}[(\overline{Y}_A - \overline{Y}_0) + (\overline{Y}_{AB} - \overline{Y}_B) + (\overline{Y}_{AC} - \overline{Y}_C) + (\overline{Y}_{ABC} - \overline{Y}_{BC})], \tag{6}$$

for treatment A,

$$\beta_{AB} = \tfrac{1}{2}\{[(\overline{Y}_A - \overline{Y}_0) - (\overline{Y}_{AB} - \overline{Y}_B)] + [(\overline{Y}_{AC} - \overline{Y}_C) - (\overline{Y}_{ABC} - \overline{Y}_{BC})]\}, \tag{7}$$

for the AB interaction, and

$$\beta_{ABC} = [(\overline{Y}_A - \overline{Y}_0) - (\overline{Y}_{AB} - \overline{Y}_B) - (\overline{Y}_{AC} - \overline{Y}_C) - (\overline{Y}_{ABC} - \overline{Y}_{BC})] \tag{8}$$

for the ABC interaction.

When certain interactions are present, we may require an alternative estimator for β_A or β_{BA} (or for other effects). Suppose that there is evidence of an ABC interaction. Then, instead of β_A, one possible estimator of the main effect of A is

$$\beta'_A = \tfrac{1}{2}[(\overline{Y}_A - \overline{Y}_0) + (\overline{Y}_{AB} - \overline{Y}_B)],$$

which does not use β_{ABC}. Other estimators of the main effect of A are possible. Similarly, the AB interaction could be tested by

$$\beta'_{AB} = (\overline{Y}_A - \overline{Y}_0) - (\overline{Y}_{AB} - \overline{Y}_B),$$

for the same reason. Thus, when treatment interactions are present, we must modify our estimates of main effects and lower order interactions, losing some efficiency.

Scale of Measurement

In the examples just given, the treatment effects and interactions have been assumed to exist on an additive scale. This is reflected in the use of sums and differences in the formulas for estimation. Other scales of measurement may be useful. As an example, consider the response data in Table 4, where the effect of Treatment A is to increase the baseline response by 10 units. The same is true of B and there is no interaction between the treatments on this scale because the joint effect of A and B is to increase the response by 20 units.

In contrast, in Table 5 are shown data in which the effects of both treatments are to multiply the baseline response by 3.0. Hence, the combined effect of A and B is a nine fold increase which is greater than the joint treatment effect for the additive case. If the analysis model were **multiplicative**, then Table 4 would show an interaction, whereas if the analysis model were additive, then Table 5 would show an interaction. Thus, to discuss interactions, we must establish the scale of measurement.

Main Effects and Interactions

In the presence of an interaction in the 2×2 design, one cannot speak simply about an overall, or main,

Table 4 Response data from a factorial trial showing no interaction on an additive scale

	B	
A	No	Yes
No	5	15
Yes	15	25

Table 5 Response data from a factorial trial showing no interaction on a multiplicative scale

	B	
A	No	Yes
No	5	15
Yes	15	45

effect of either treatment. This is because the effect of A is different depending on the presence or absence of B. In the presence of a small interaction, where all patients benefit from A regardless of the use of B, we might observe that the magnitude of the "overall" effect of A is of some size and that therapeutic decisions are unaffected by the presence of an interaction. This is called "quantitative" interaction, so-named because it does not affect the direction of the treatment effect. For large quantitative interactions, it may not be sensible to talk about overall effects.

In contrast, if the presence of B reverses the effect of A, then the interaction is "qualitative", and treatment decisions may need to be modified. Here, we cannot talk about an overall effect of A, because it could be positive in the presence of B, negative in the absence of B, and could yield an average effect near zero (*see* **Interaction in Factorial Experiments**; **Treatment-covariate Interaction**).

Analysis

Motivation for the estimators given above can be obtained using **general linear models**. There has been little theoretic work on analyses using other models. One exception is the work by Slud [27] describing approaches to factorial trials with survival outcomes (*see* **Survival Analysis, Overview**). Suppose we have conducted a 2×2 factorial experiment with group sizes given by Table 1. We can estimate the AB interaction effect using the linear model

$$\mathrm{E}\{Y\} = \beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_A X_B, \quad (9)$$

where the Xs are indicator variables for the treatment groups and β_{AB} is the interaction effect. The design matrix has dimension $4n \times 4$ and is

$$\mathbf{X}' = \begin{bmatrix} 1 & \dots & 1 & \dots & 1 & \dots & 1 & \dots \\ 0 & \dots & 1 & \dots & 0 & \dots & 1 & \dots \\ 0 & \dots & 0 & \dots & 1 & \dots & 1 & \dots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 & \dots \end{bmatrix},$$

where there are four blocks of n identical rows representing each treatment group and the columns represent effects for the intercept, treatment A, treatment B, and both treatments, respectively. The vector of responses has dimension $4n \times 1$ and is

$$\mathbf{Y}' = \{Y_{01}, \dots, Y_{A1}, \dots, Y_{B1}, \dots, Y_{AB1}, \dots\}.$$

The ordinary least squares solution for the model (9) is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The **covariance matrix** is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, where the variance of each observation is σ^2.

We have

$$\mathbf{X}'\mathbf{X} = n \times \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & 2 & 1 & 1 \\ 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n} \times \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 2 & 1 & -2 \\ -1 & 1 & 2 & -2 \\ 1 & -2 & -2 & 4 \end{bmatrix},$$

and

$$\mathbf{X}'\mathbf{Y} = n \times \begin{bmatrix} \overline{Y}_0 + \overline{Y}_A + \overline{Y}_B + \overline{Y}_{AB} \\ \overline{Y}_A + \overline{Y}_{AB} \\ \overline{Y}_B + \overline{Y}_{AB} \\ \overline{Y}_{AB} \end{bmatrix},$$

where $\overline{Y}_i$ denotes the average response in the ith group. Then,

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \overline{Y}_0 \\ -\overline{Y}_0 + \overline{Y}_A \\ -\overline{Y}_0 + \overline{Y}_B \\ \overline{Y}_0 - \overline{Y}_A - \overline{Y}_B + \overline{Y}_{AB} \end{bmatrix}, \quad (10)$$

which corresponds to the estimators given above in (3)–(5). However, if we assume no interaction, then the β_{AB} effect is removed from the model, and we obtain the estimator

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \frac{3}{4}\overline{Y}_0 + \frac{1}{4}\overline{Y}_A + \frac{1}{4}\overline{Y}_B - \frac{1}{4}\overline{Y}_{AB} \\ -\frac{1}{2}\overline{Y}_0 + \frac{1}{2}\overline{Y}_A - \frac{1}{2}\overline{Y}_B + \frac{1}{2}\overline{Y}_{AB} \\ -\frac{1}{2}\overline{Y}_0 - \frac{1}{2}\overline{Y}_A + \frac{1}{2}\overline{Y}_B + \frac{1}{2}\overline{Y}_{AB} \end{bmatrix}.$$

The main effects for A and B are as given above in (1) and (2).

The covariance matrices for these estimators are

$$\widehat{\mathrm{cov}}\{\hat{\boldsymbol{\beta}}\} = \frac{\sigma^2}{n} \times \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 2 & 1 & -2 \\ -1 & 1 & 2 & -2 \\ 1 & -2 & -2 & 4 \end{bmatrix}$$

and

$$\widehat{\mathrm{cov}}\{\hat{\boldsymbol{\beta}}^*\} = \frac{\sigma^2}{n} \times \begin{bmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix}.$$

In the absence of an interaction, the main effects of A and B are estimated independently and with higher precision than when an interaction is present. The interaction effect is relatively imprecisely estimated, indicating that larger sample sizes are required to have a high power to detect such effects.

Examples

Several clinical trials conducted in recent years have used factorial designs. A sample of such studies is shown in Table 6. One important study using a 2×2 factorial design is the Physicians' Health Study [16, 30]. This trial has been conducted, in 22 000 physicians in the US and was designed to test the effects of (i) aspirin on reducing cardiovascular mortality and (ii) β-carotene on reducing cancer incidence. The trial is noteworthy in several ways, including its test of two interventions in unrelated diseases, use of physicians as subjects to report outcomes reliably, relatively low cost, and an all-male (high risk) study population. This last characteristic has led to some unwarranted criticism.

In January 1988 the aspirin component of the Physicians' Health Study was discontinued, because evidence demonstrated convincingly that it was associated with lower rates of myocardial infarction [20]. The question concerning the effect of β-carotene on cancer remains open and will be addressed by continuation of the trial. In the likely absence of an interaction between aspirin and β-carotene, the second major question of the trial will be unaffected by the closure of the aspirin component.

Another noteworthy example of a 2×2 factorial design is the α-tocopherol β-carotene Lung Cancer Prevention Trial conducted in 29 133 male smokers in Finland between 1987 and 1994 [3, 15]. In this study, lung cancer incidence is the sole outcome. It was thought possible that lung cancer incidence could be reduced by either or both interventions. When the intervention was completed in 1994, there were 876 new cases of lung cancer in the study population during the trial. Alpha-tocopherol was not associated with a reduction in the **risk** of cancer. Surprisingly, β-carotene was associated with a statistically significantly *increased* incidence of lung cancer [4]. There was no evidence of a treatment interaction. The unexpected findings of this study have been supported by the recent results of another large trial of carotene and retinol [32].

The Fourth International Study of Infarct Survival (ISIS-4) was a $2 \times 2 \times 2$ factorial trial assessing the efficacy of oral captopril, oral mononitrate, and intravenous magnesium sulfate in 58 050 patients with suspected myocardial infarction [12, 17]. No significant interactions among the treatments were observed and each main effect comparison was based

Table 6 Some recent randomized clinical trials using factorial designs

Trial	Design	Reference
Physicians' Health Study	2×2	Hennekens & Eberlein [16]
ATBC Prevention Trial	2×2	Heinonen et al. [15]
Desipramine	2×2	Max et al. [22]
ACAPS	2×2	ACAPS Group [1]
Linxian Nutrition Trial	2^4	Li et al. [19]
Retinitis pigmentosa	2×2	Berson et al. [5]
Linxian Cataract Trial		Sperduto et al. [29]
Tocopherol/deprenyl	2×2	Parkinson Study Group [23]
Womens' Health Initiative	2^3	Assaf & Carleton [2]
Polyp Prevention Trial	2×2	Greenberg et al. [14]
Cancer/eye disease	2×2	Green et al. [13]
Cilazapril/hydrochlorothiazide	4×3	Pordy [25]
Nebivolol	4×3	Lacourciere et al. [18]
Endophthalmitis vitrectomy study	2×2	Endophthalmitis Vitrectomy Study Group [9]
Bicalutamide/flutamide	2×2	Schellhammer et al. [26]
ISIS-4	2^3	ISIS-4 Collaborative Group [17]

Source: adapted from Piantadosi [16].

on approximately 29 000 treated vs. 29 000 control patients. Captopril was associated with a small but statistically significant reduction in five-week mortality. The difference in mortality was 7.19% vs. 7.69% (143 events out of 4319), illustrating the ability of large studies to detect potentially important treatment effects even when they are small in relative magnitude. Mononitrate and magnesium therapy did not significantly reduce five-week mortality.

Similar Designs

Fractional and Partial Factorial Designs

Fractional factorial designs are those which omit certain treatment groups by design. A careful analysis of the objectives of an experiment, its efficiency, and the effects that it can estimate may justify not using some groups. Because many cells contribute to the estimate of any effect, a design may achieve its intended purpose without some of the cells.

In the 2×2 design, all treatment groups must be present to permit estimating the interaction between A and B. However, for higher order designs, if some interactions are thought biologically not to exist, omitting certain treatment combinations from the design will still permit estimates of other effects of interest. For example, in the $2 \times 2 \times 2$ design, if the interaction between A, B, and C is thought not to exist, omitting that treatment cell from the design will still permit estimation of all the main effects. The efficiency will be somewhat reduced, however. Similarly, the two-way interactions can still be estimated without $\overline{Y}_{ABC}$. This can be verified from the formulas above.

More generally, fractional high-order designs will produce a situation termed "aliasing", in which the estimates of certain effects are algebraically identical to completely different effects. If both effects are biologically possible, the design will not be able to reveal which effect is being estimated. Naturally, this is undesirable unless additional information is available to the investigator to indicate that some aliased effects are zero. This can be used to advantage in improving efficiency and one must be careful in deciding which cells to exclude. See Cox [8] or Mason & Gunst [21] for a discussion of this topic.

The Women's Health Initiative clinical trial is a $2 \times 2 \times 2$ partial factorial design studying the effects of hormone replacement, dietary fat reduction, and calcium and vitamin D on coronary disease, breast cancer, and osteoporosis [2]. All eight combinations of treatments are given, but participants may opt to join one, two, or all three of the randomized components. The study is expected to accrue over 64 000 patients and is projected to finish in the year 2007. The dietary component of the study will randomize 48 000 women using a 3:2 allocation ratio in favor of the control arm and nine years of follow-up. Such a large and complex trial presents logistical difficulties, questions about adherence, and sensitivity of the intended power to assumptions that can only roughly be validated.

Incomplete Factorial Designs

When treatment groups are dropped out of factorial designs without yielding a fractional replication, the resulting trials have been termed "incomplete factorial designs" [7]. In incomplete designs, cells are not missing by design intent, but because some treatment combinations may be infeasible. For example, in a 2×2 design, it may not be ethically possible to use a placebo group. In this case, one would not be able to estimate the AB interaction. In other circumstances, unwanted aliasing may occur, or the efficiency of the design to estimate main effects may be greatly reduced. In some cases, estimators of treatment and interaction effects are biased, but there may be reasons to use a design that retains as much of the factorial structure as possible. For example, they may be the only way in which to estimate certain interactions.

References

[1] ACAPS Group (1992). Rationale and design for the Asymptomatic Carotid Artery Plaque Study (ACAPS), *Controlled Clinical Trials* **13**, 293–314.

[2] Assaf, A.R. & Carleton, R.A. (1994). The Women's Health Initiative clinical trial and observational study: history and overview, *Rhode Island Medicine* **77**, 424–427.

[3] ATBC Cancer Prevention Study Group (1994). The alpha-tocopherol beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance, *Annals of Epidemiology* **4**, 1–9.

[4] ATBC Cancer Prevention Study Group (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers, *New England Journal of Medicine* **330**, 1029–1034.

[5] Berson, E.L., Rosner, B., Sandberg, M.A., Hayes, K.C., Nicholson, B.W., Weigel-DiFranco, C. & Willett, W. (1993). A randomized trial of vitamin A and vitamin E supplementation for retinitis pigmentosa, *Archives of Ophthalmology* **111**, 761–772.

[6] Byar, D.P. & Piantadosi, S. (1985). Factorial designs for randomized clinical trials, *Cancer Treatment Reports* **69**, 1055–1063.

[7] Byar, D.P., Herzberg, A.M. & Tan, W.-Y. (1993). Incomplete factorial designs for randomized clinical trials, *Statistics in Medicine* **12**, 1629–1641.

[8] Cox, D.R. (1958). *Planning of Experiments*. Wiley, New York.

[9] Endophthalmitis Vitrectomy Study Group. Results of the Endophthalmitis Vitrectomy Study (1995). A randomized trial of immediate vitrectomy and of intravenous antibiotics for the treatment of postoperative bacterial endophthalmitis, *Archives of Ophthalmology* **113**, 1479–1496.

[10] Fisher, R.A. (1935). *The Design of Experiments*. Collier Macmillan, London.

[11] Fisher, R.A. (1960). *The Design of Experiments*, 8th Ed. Hafner, New York.

[12] Flather, M., Pipilis, A., Collins, R. et al. (1994). Randomized controlled trial of oral captopril, of oral isosorbide mononitrate and of intravenous magnesium sulphate started early in acute myocardial infarction: safety and haemodynamic effects, *European Heart Journal* **15**, 608–619.

[13] Green, A., Battistutta, D., Hart, V., Leslie, D., Marks, G., Williams, G., Gaffney, P., Parsons, P., Hirst, L., Frost, C. et al. (1994). The Nambour Skin Cancer and Actinic Eye Disease Prevention Trial: design and baseline characteristics of participants, *Controlled Clinical Trials* **15**, 512–522.

[14] Greenberg, E.R., Baron, J.A., Tosteson, T.D., Freeman, D.H., Jr, Beck, G.J., Bond, J.H., Colacchio, T.A., Coller, J.A., Frankl, H.D., Haile, R.W., Mandel, R.W., Nierenberg, J.S., Rothstein, D.W., Richard, S., Dale, C., Stevens, M.M., Summers, R.W. & van Stolk, R.U. (1994). A clinical trial of antioxidant vitamins to prevent colorectal adenoma. Polyp Prevention Study Group, *New England Journal of Medicine* **331**, 141–147.

[15] Heinonen, O.P., Virtamo, J., Albanes, D. et al. (1987). Beta carotene, alpha-tocopherol lung cancer intervention trial in Finland, in *Proceedings of the XI Scientific Meeting of the International Epidemiologic Association*, Helsinki, August, 1987. Pharmy, Helsinki.

[16] Hennekens, C.H. & Eberlein, K. (1985). A randomized trial of aspirin and beta-carotene among U.S. physicians, *Preventive Medicine* **14**, 165–168.

[17] ISIS-4 Collaborative Group (1995). ISIS-4: a randomized factorial trial assessing early captopril, oral mononitrate, and intravenous magnesium-sulphate in 58 050 patients with suspected acute myocardial infarction, *Lancet* **345**, 669–685.

[18] Lacourciere, Y., Lefebvre, J., Poirier, L., Archambault, F. & Arnott, W. (1994). Treatment of ambulatory hypertensives with nebivolol or hydrochlorothiazide alone and in combination. A randomized double-blind, placebo-controlled, factorial-design trial, *American Journal of Hypertension* **7**, 137–145.

[19] Li, B., Taylor, P.R., Li, J.Y., Dawsey, S.M., Wang, W., Tangrea, J.A., Liu, B.Q., Ershow, A.G., Zheng, S.F., Fraumeni, J.F., Jr et al. (1993). Linxian nutrition intervention trials. Design, methods, participant characteristics, and compliance, *Annals of Epidemiology* **3**, 577–585.

[20] Lubsen, J. & Pocock, S.J. (1994). Factorial trials in cardiology (editorial), *European Heart Journal* **15**, 585–588.

[21] Mason, R.L. & Gunst, R.L. (1989). *Statistical Design and Analysis of Experiments*. Wiley, New York.

[22] Max, M.B., Zeigler, D., Shoaf, S.E., Craig, E., Benjamin, J., Li, S.H., Buzzanell, C., Perez, M. & Ghosh, B.C. (1992). Effects of a single oral dose of desipramine on postoperative morphine analgesia, *Journal of Pain & Symptom Management* **7**, 454–462.

[23] Parkinson Study Group (1993). Effects of tocopherol and deprenyl on the progression of disability in early Parkinson's disease, *New England Journal of Medicine* **328**, 176–183.

[24] Piantadosi, S. (1997). Factorial designs, in *Clinical Trials: a Methodologic Perspective*. Wiley, New York. *See* Chapter 15.

[25] Pordy, R.C. (1994). Cilazapril plus hydrochlorothiazide: improved efficacy without reduced safety in mild to moderate hypertension. A double-blind placebo-controlled multi-center study of factorial design, *Cardiology* **85**, 311–322.

[26] Schellhammer, P., Shariff, R., Block, N., Soloway, M., Venner, P., Patterson, A.L., Sarosdy, M., Vogelzang, N., Jones, J. & Kiovenbag, G. (1995). A controlled trial of bicalutamide versus flutamide, each in combination with lutenizing hormone-releasing hormone analogue therapy, in patients with advanced prostate cancer. Casodex Combination Study Group, *Urology* **45**, 745–752.

[27] Slud, E.V. (1994). Analysis of factorial survival experiments, *Biometrics* **50**, 25–38.

[28] Snedecor, G.W. & Cochran, W.G. (1980). *Statistical Methods*, 7th Ed. The Iowa State University Press, Ames.

[29] Sperduto, R.D., Hu, T.S., Milton, R.C., Zhao, J.L., Everett, D.F., Cheng, Q.F., Blot, W.J., Bing, L., Taylor, P.R., Li, J.Y. et al. (1993). The Linxian cataract studies. Two nutrition intervention trials, *Archives of Ophthalmology* **111**, 1246–1253.

[30] Stampfer, M.J., Buring, J.E., Willett, W. et al. (1985). The 2×2 factorial design: its application to a randomized trial of aspirin and carotene in U.S. physicians, *Statistics in Medicine* **4**, 111–116.

[31] Steering Committee of the Physicians' Health Study Research Group (1989). Final report on the aspirin

component of the ongoing physicians' health study. *New England Journal of Medicine* **321**, 129–135.

[32] Thornquist, M.D., Owenn, G.S., Goodman, G.E. et al. (1993). Statistical design and monitoring of the carotene and retinol efficacy trial (CARET), *Controlled Clinical Trials* **14**, 308–324.

[33] Yates, F. (1935). Complex experiments (with discussion), *Journal of the Royal Statistical Society, Series B* **2**, 181–247.

STEVEN PIANTADOSI

Factorial Experiments

Experimental designs in which the treatments can be classified by the levels of two or more factors are called factorial experiments. For example, consider a study conducted to compare the effectiveness of four *treatments* in the reduction of blood cholesterol. The treatments were: diet with a palm oil supplement, diet with a rice-bran oil supplement, diet with both the supplements, and diet with neither supplement. The four treatments could be grouped into two *factors*, namely, palm oil (factor A) and rice-bran oil (factor B) each having two *levels*, namely the presence or absence of the respective diet supplement. The various combinations of the factor levels comprise the four treatments.

In the factorial experiment the treatments are randomly assigned to the study subjects (*see* **Randomized Treatment Assignment**) and an *outcome* variable (*see* **Outcome Measures in Clinical Trials**), say the reduction in cholesterol after 8 weeks, is observed. Table 1 presents hypothetical data on the reduction in cholesterol from the experiment. In this example each treatment was *replicated* on two individuals (*experimental units*). The number of replicates on each treatment need not be the same. If the number of replicates is the same for all the treatments, then the experiment is called a *balanced* experiment; if not, then the experiment is called an *unbalanced* experiment.

The design of factorial experiment described here is known as a 2×2 or a 2^2 factorial design. Factorial experiments could have several factors and the number of levels for the factors could vary. For instance, in the cholesterol reduction example, if we include exercise as another factor (factor C) at two levels (whether or not an individual is prescribed an exercises regimen), then the experiment would be 2^3 factorial. That is, it is a factorial experiment with three factors where each factor has two levels. In general, a factorial experiment that has k factors, each having two levels, is called a 2^k factorial design. An example of a more general factorial design would be a $4 \times 5 \times 3$ design that has three factors, where the first factor has four levels, the second has five levels and the third has three levels. The number of experimental units (or *runs*) needed for this design would be $4 \times 5 \times 3 = 60$ per replicate for a balanced design.

Table 1 Hypothetical data from a 2×2 factorial experiment on cholesterol reduction

Factor A: palm oil	Factor B: rice-bran oil	
	Present	Absent
Present	25	11
	20	13
	Mean = 22.5	Mean = 12
Absent	18	−5
	23	4
	Mean = 20.5	Mean = −0.5

Analysis of Factorial Experiments

The purpose of a factorial experiment is to examine whether the factors have significant *effects* on the outcome being measured. In the cholesterol reduction example the purpose would be to test the hypotheses (*see* **Hypothesis Testing**) that: (i) palm oil supplement has an influence on the reduction of cholesterol, and (ii) rice-bran oil supplement has an influence on the reduction of cholesterol. To test these hypotheses, estimates of the effects of each supplement, called the *main* effects, are calculated.

The main effect for palm oil, for example, would be the average of the difference between the two levels of palm oil at each level of the rice-bran oil. That is,

$$\frac{(22.5 - 20.5) + (12.0 - (-0.5))}{2} = 7.25.$$

Similarly, the main effect for the rice-bran oil is

$$\frac{(22.5 - 12) + (20.5 - (-0.5))}{2} = 15.25.$$

These estimates are interpreted as follows: the palm oil supplement, on average, will reduce cholesterol by 7.25 units, while the rice-bran oil will reduce it by 15.25 units. However, notice that the effect of palm oil is much higher $[12 - (-0.5) = 12.5]$ when the rice-bran oil is absent than when it is present (2.0). Viewed in another way, the effect of rice-bran oil is much higher (21.0) when the palm oil is absent than when it is present (10.5). This differential effect could be shown by a plot of the cell means, as shown in Figure 1.

Notice that the line corresponding to the case where the rice-bran oil is present is not parallel

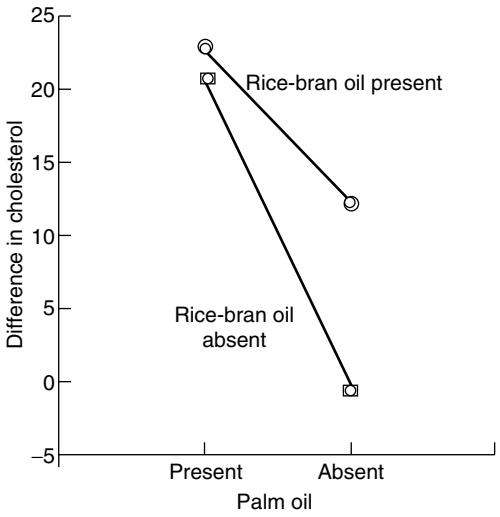

Figure 1 Interaction between the supplements

to the line corresponding to the case where rice-bran oil is absent. This differential effect is called an *interaction* between the factors A and B (*see* **Interaction in Factorial Experiments**). When the interaction effect is statistically significant the main effects are not *additive*, and hence it is not meaningful to interpret them. In that situation the effect of each factor should be interpreted at each level of the other factor. The interaction effect between any two factors is measured by the difference between the average effects of one factor at the two levels of the other factor. For the cholesterol reduction example the interaction effect between palm oil and rice-bran oil is estimated as

$$\frac{(22.5 - 12) - (20.5 - (-0.5))}{2} = 5.25.$$

To test formally whether the effects in a 2^k factorial experiment are statistically significant, an estimate of the **standard deviation** of the effects is needed. Since there are two replicate measurements at each treatment combination, an estimate of the standard deviation could be obtained in this situation. Assuming the **variances** of the four treatments to be equal, an estimate of the variance could be obtained by pooling the variances of each run. Since there are only two replicates in the cholesterol example, the variances for each run could be computed by the square of the difference between the observations divided by 2. Thus the estimate of the pooled standard deviation is

$$s = \left[\frac{(25-20)^2 + (11-13)^2 + (18-23)^2 + (-5-4)^2}{2 \times 4} \right]^{1/2} = 4.12.$$

The **degrees of freedom** for the pooled standard deviation equal the number of runs. Then, assuming the reduction in cholesterol is **normally distributed**, for any given effect, the statistic

$$F = \frac{\text{effect}^2}{s^2/4},$$

which follows an $\boldsymbol{F_{1,4}}$ **distribution** can be used to test the hypothesis that the effect is zero. Thus an **analysis of variance** (ANOVA) table for the example could be constructed (Table 2). The observed F values in the table should be compared with the $F_{1,4}$ distribution to obtain the $\boldsymbol{P}$ **values**.

The **estimation** of the effects in higher-order factorial designs is performed employing similar concepts. Several **algorithms** for facilitating calculations are available (*see* **Yates's Algorithm**). A widely used algorithm that uses **contrasts** is described here. Consider a 2^3 factorial design. Suppose the levels of the factors are denoted by a + (present, high, etc.) or a − (absent, low, etc.) and let y denote the outcome. Then, a table of contrast coefficients is constructed (Table 3).

The first three columns in Table 3 simply represent the treatment combination for the eight runs. For example, if exercise is included as the third factor C in the cholesterol reduction experiment, the combination −,−,− corresponding to the A, B, and C columns, would represent the run in which both palm oil and rice-bran oil supplements were absent in the data and the subject did not exercise. Then the signs in the AB column are obtained by multiplying the A

Table 2 ANOVA for the experiment on cholesterol reduction

Effect	Sum of squares	df	Mean sum of squares	F
Main A	52.56	1	52.56	12.40
Main B	232.56	1	232.56	54.85
Interaction	33.06	1	33.06	7.80
Error	16.97	4	4.24	
Total	45.22	7		

Table 3 Contrast table for calculating effects in a 2^3 factorial design

A	B	C	AB	AC	BC	ABC	Mean	Observation
−	−	−	+	+	+	−	+	y_1
+	−	−	−	−	+	+	+	y_2
−	+	−	−	+	−	+	+	y_3
+	+	−	+	−	−	−	+	y_4
−	−	+	+	−	−	+	+	y_5
+	−	+	−	+	−	−	+	y_6
−	+	+	−	−	+	−	+	y_7
+	+	+	+	+	+	+	+	y_8

column by the B column (and defining $- \times - = +$ and $- \times + = -$) and the AC column is obtained by multiplying the A column by C column, and so on. *The effects are then obtained by summing the observations with the respective signs attached and dividing by 4. (The divisor for the mean column however is 8.)* Thus, for instance the effect for the BC interaction will be estimated by

$$\frac{(y_1 + y_2 - y_3 - y_4 - y_5 - y_6 + y_7 + y_8)}{4}.$$

The same method could be extended to higher-order factorial designs by filling the first k columns by the 2^k treatment combinations and then by multiplying the appropriate columns for interactions. The divisor for each effect in a 2^k factorial design is 2^{k-1}.

Since one of the main advantages of factorial design, especially when there are a large number of factors, is that it is inexpensive, it is primarily employed for conducting pilot studies. Therefore, adequate replicate measurements are seldom available to estimate the **standard error** of the effects with reasonable **power**. Consequently, formal testing of the effects using F tests as described above becomes infeasible without making additional assumptions. For instance, if the higher-order interactions could be assumed to be zero, the corresponding effects could be treated as random variability. Then, using this as an estimate of the variance, formal F tests as above could be obtained. An alternative method for identifying the significant effects is performed using normal probability plots.

Under the assumption that the observations are normally distributed with equal variance, the effects, being linear combinations of the observations, would also be normally distributed with mean zero (under the **null hypotheses**) and equal variance. Therefore, a normal probability plot of the effects should fall on a straight line that has a slope of 1 and passes through the origin. The effects that fall away from this line could then be deemed important.

An example of a normal probability plot for the 15 effects from a 2^4 factorial design is presented in Figure 2. Since in Figure 2 the main effects of A, B, and D, and the interaction between B and D, do fall relatively farther than the other effects from the straight line passing through the origin, they are deemed important.

Fractional Factorial

When the number of factors in a factorial design is large, the number of runs needed may be too large to perform within a reasonable time or with available resources. In such situations a **fractional factorial** design is utilized. In this design, a fraction of the 2^k runs is selected systematically so that all the main effects can be estimated, while compromising some higher-order interactions.

Blocking and Confounding in Factorial Designs

In all factorial experiments, it is desirable to perform the various runs on homogeneous subjects.

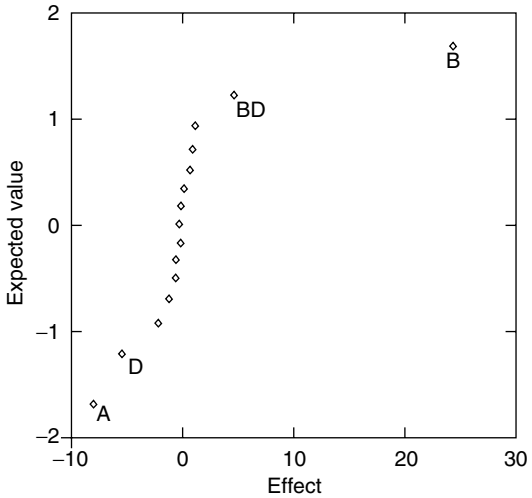

Figure 2 Normal probability plot for a 2^4 factorial experiment. Reproduced from Box et al. [2], Figure 10.9(a), p. 332, by permission of John Wiley & Sons

However, in some experiments the number of homogeneous runs that can be performed together may be restricted. In these circumstances the number of runs is often split into smaller equal-sized homogeneous groups. This procedure is called **blocking**. Suppose in the 2^3 factorial design for the cholesterol reduction experiment (including exercise as factor C) only eight individuals were available, of which four are males and four are females. Considering that the males and females may respond to diet and exercise differently, the eight runs randomly assigned could not be assumed to come from homogeneous individuals. Therefore blocking, in this case by gender, is necessary.

Blocking poses a peculiar problem. How to split the eight runs into four for males and four for females? If four of the runs that include the palm oil supplement are all given to the males and the other four that do not are given to the females, it would be impossible to distinguish the effect of palm oil supplement from the effect of gender. Therefore, the choice of the effects has to be made judiciously. Whichever way the block assignments are made, one of the effects will be indistinguishable from the effect of gender. This phenomenon is called **confounding**. In other words, the block (or gender) effect is confounded with the main effect due to palm oil.

The highest-order interaction is often the choice for confounding where blocks are used. In the cholesterol example the ABC interaction would be the best choice. In other words, assign the runs to the blocks according to the ABC column in Table 3. Therefore, the males will get the first, fourth, sixth, and seventh runs, while the females will get the second, third, fifth, and eighth runs. When more than two blocks are required, more than one effect will be confounded with the block effect. The blocking scheme for these situations is much more complex. The book by Box et al. [2] describes different schemes for constructing the blocks and also provides tables that suggest blocking assignments.

ANOVA for General Factorial Experiments

Typical steps in the analysis of a general factorial (or a multifactor) experiment include: (i) formulating a model; (ii) estimating the effects; (iii) testing the effects; and (iv) model **diagnostics**. For an $a \times b$ two-factor experiment the model could be written

$$\begin{aligned}\text{outcome} = {} & \text{overall average} + \text{effect due to factor A} \\ & + \text{effect due to factor B} \\ & + \text{effect due to AB interaction} \\ & + \text{random error.}\end{aligned}$$

In mathematical notation the model is written

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

where Y_{ijk} is the observation corresponding to the kth individual receiving the ith level of factor A and the jth level of factor B. The parameter μ corresponds to the overall **mean** and the parameters α_i, β_j, and $(\alpha\beta)_{ij}$ correspond to the main effects of A and B and the interaction effect, respectively. The last term in the model, ε_{ij}, represents the **random error**. The assumptions of the model are that the errors are independent and identically distributed as normal with mean zero and equal variance and the effects summed over the respective levels of each factor add up to zero (*see* **Additive Model**).

The estimates of the parameters are obtained by a **least squares** method that minimizes the sum of the squared errors. This method of estimation requires the equality of the error variance but the normality is not required. However, under normality, the least square estimates coincide with the **maximum likelihood** estimates.

The distribution of the estimates of the various parameters, under the assumption of normality of errors, can be shown to be normal as well and hence, using appropriate quadratic forms, F tests for testing the hypothesis on each effect can be derived. In the case of a balanced design the F tests can be derived as ratios of the respective mean sums of square deviations.

As in all statistical methods, the assumptions of the model must be verified by appropriate diagnostic methods. In the factorial model, the assumptions of normality, equal variance, and independence must be checked. A normal probability plot of the residuals and scatter plots of the residuals vs. specific variables are useful for diagnostics. Books by Cook & Weisberg [3] or Box & Draper [1] are wonderful sources for a more complete understanding of this topic.

Most computer packages carry the procedures that can perform the analyses of factorial experiments.

The SAS package [6] has PROC ANOVA for balanced design and PROC GLM and PROC MIXED for complex factorials that may have unbalanced or missing data. SPSS [7] and MINITAB [5] can also perform these analysis (*see* **Software, Biostatistical**).

An Historical Note on Factorial Experiments

The father of statistics is also the father of the factorial experiment. The term "factorial experiment" itself was coined by **Fisher** [4]. Prior to his introduction of the term these experiments were exclusively known as "complex experiments". **Yates** continued to use the term complex experiments during the earlier years of his work on this topic, but later he also referred to them as factorial experiments. Factorial experiments were primarily used for agricultural experiments. According to Yates [8] the use of factorial experiments (complex experiments) dates back to 1843.

The major extensions and developments of Fisher's presentations of the concepts on factorial experiments were bestowed to the field of statistics by Yates. Other major contributors to the field include **Cochran**, Finney, Kempthorne, Rao, and **Snedecor**.

References

[1] Box, G.E.P. & Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.

[2] Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters*. Wiley, New York.

[3] Cook, R.D. & Weisberg, S. (1982). *Residual and Influence in Regression*. Chapman & Hall, London.

[4] Fisher, R.A. (1926). The arrangement of field experiments, *Journal of Ministry for Agriculture* **33**, 503–513.

[5] MINITAB (1989). *MINITAB Reference Manual, Release 7*. Minitab Inc., State College.

[6] SAS (1987). *SAS User's Guide: Statistics, Version 6*. SAS Institute, Cary.

[7] SPSS (1986). *SPSS User's Guide*, 2nd Ed. SPSS, Chicago.

[8] Yates, F. (1970). *Experimental Design: Selected Papers of Frank Yates, CBE, FRS*. Hafner (Macmillan), New York.

Bibliography

Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York.

Fisher, R.A. (1966). *The Design of Experiments*, 8th Ed. Hafner, New York.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. & Kutner, M.H. (1995). *Applied Linear Statistical Models*, 4th Ed. Irwin, Homewood.

Snedecor, G.W. & Cochran, W.G. (1980). *Statistical Methods*, 7th Ed. Iowa State University Press, Ames.

V. RAMAKRISHNAN

False Negative Rate

The false negative rate of a **diagnostic** or **screening** test is conventionally taken to be the probability that a true case of disease is given an incorrect, negative result; in other words, the false negative rate is Pr(negative test result | disease). In the table in the article **Sensitivity**, the false negative rate is $c/(a+c)$. The false negative rate is the complement of sensitivity, which is $a/(a+c)$.

As has been pointed out [1, 2], the name implies some ambiguity about the appropriate denominator for the false negative rate. The definition given here uses as denominator the number of true disease cases, $a+c$. Elsewhere, one may encounter this rate based on a denominator $c+d$, i.e. the total number of negative tests.

References

[1] Altman, D.G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall, London, Chapter 14.

[2] Sackett, D.L., Haynes, R.B., Guyatt, G.H. & Tugwell, P. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*, 2nd Ed. Little, Brown & Company, Boston, Chapter 4.

(*See also* **Gold Standard Test**)

STEPHEN D. WALTER

False Positive Rate

The false positive rate of a **diagnostic** or **screening** test is conventionally taken to be the probability that a noncase of disease is given an incorrect, positive result, or Pr(positive test result | no disease). In the table in the article on **Sensitivity**, the false positive rate is $b/(b+d)$. Its complement, $d/(b+d)$, is known as the **specificity** of the test.

As with the definition of **false negative rate**, there is some ambiguity about the appropriate denominator. The definition invoked here uses as denominator for the false positive rate the number of true noncases of disease, $b+d$. Elsewhere, one may encounter this rate based on a denominator $a+b$, i.e. the total number of positive tests.

(*See also* **Gold Standard Test**)

STEPHEN D. WALTER

Familial Correlations

In a lecture at the Royal Institution on February 9, 1877, **Francis Galton** [13] described reversion as "the tendency of the ideal mean filial type to depart from the parental type, reverting to what may be roughly and perhaps fairly described as the average ancestral type". Having found it difficult to obtain human data for two generations, Galton reported the analysis of data from carefully selected sweet pea seeds. Galton introduced the first letter of the word "reversion" as a symbol for representing a numerical measure of what he [14] later termed **regression**. Galton had in essence defined the interclass **correlation** coefficient. With the assistance of Hamilton Dickson of Cambridge University, in 1886 Galton [15] reported the discovery of the **bivariate normal distribution** in which the correlation coefficient is expressed as a parameter. In this study of family likeness in stature, he set about the calculation of coefficients of correlation among various pairs of relatives: parent and offspring, brothers, fathers and sons, uncles and nephews, grandparents and grandsons. In considering the correlations among brothers, Galton [15] had in essence defined the intraclass correlation coefficient as the simple correlation over all possible pairs of brothers. The term *co-relation* or *correlation* does not appear until 1889, when Galton [15] defined it in the context of heredity to describe the degree of likeness among family members or so-called *relations*. Galton [16] used the term "partial co-relation" in the path towards the development of the multiple correlation coefficient.

In the process of defining and providing estimates of familial correlations, Galton formulated regression analysis and the **multivariate normal distribution** – two major elements not just of quantitative genetics but also of mathematical statistics.

In 1896, **Karl Pearson** [26] proposed the well-known product-moment estimators of the intraclass and interclass correlation coefficients by replacing the median and probable error (or interquartile range) by the mean and standard deviation, respectively, in Galton's estimators – which incidentally are more robust.

Similarly, Pearson [26] proposed the product-moment estimator of the parent–offspring interclass correlation coefficient as the simple correlation coefficient computed in a sample of k families over all possible pairs of observations in the ith family formed from the parent's value y_i and the offspring's values $(x_{i1}, x_{i2}, \ldots, x_{in_i})$:

$$r_{\mathrm{ms}} = \frac{\sum_{i=1}^{k}(y_i - \overline{y}_n)\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_n)}{\left[\sum_{i=1}^{k} n_i (y_i - \overline{y}_n)^2\right]^{1/2}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_n)^2\right]^{1/2}}, \tag{1}$$

where the sample means $\overline{y}_n = (n_1 y_1 + n_2 y_2 + \cdots + n_k y_k)/N$ and $\overline{x}_n = (n_1 \overline{x}_1 + n_2 \overline{x}_2 + \cdots + n_k \overline{x}_k)/N$ for $N = n_1 + n_2 + \cdots + n_k$ and $\overline{x}_i = (x_{i1} + x_{i2} + \cdots + x_{in_i})/n_i$.

Without explicitly stating the formula, Pearson [26] proposed the product-moment estimator of the sibling intraclass correlation coefficient as the simple correlation coefficient computed in a sample of k families over all possible pairs of observations on siblings in each family:

$$r_p = \frac{\sum_{i=1}^{k}\sum_{\substack{j=1 \\ m=1 \\ j\neq m}}^{k}(x_{ij} - \overline{x}_p)(x_{im} - \overline{x}_p)}{\sum_{i=1}^{k}(n_i - 1)\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_p)^2}, \tag{2}$$

where $\overline{x}_{\mathrm{p}} = \sum_{i=1}^{k} n_i (n_i - 1)\,\overline{x}_i / \sum_{i=1}^{k} n_i (n_i - 1)$. Some researchers substitute $\overline{x}_n$ for $\overline{x}_p$ in the above definition for consistency with the previous definition for the pairwise estimator of the interclass correlation – see Konishi [22].

In 1913, Harris [17] simplified the laborious hand calculation of the product-moment estimator r_{p} of the intraclass correlation by considering in the balanced case a decomposition of the pairwise sum of squares which avoids the computation of the cross product in the numerator. This decomposition in the unbalanced case produces

$$r_{\mathrm{p}} = \frac{\sum_{i=1}^{k} n_i (n_i - 1)(\overline{x}_i - \overline{x}_{\mathrm{p}})^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2}{\sum_{i=1}^{k}(n_i - 1)\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{\mathrm{p}})^2}. \tag{3}$$

In 1915, Fisher [9] derived the **sampling distribution** of the simple (interclass) correlation coefficient and developed the z-transformation as a variance stabilizing **transformation** (*see* **Delta Method**).

In 1918, Fisher [10] studied the impact of dominance, epistasis (*see* **Genotype**), **linkage**, and **assortative mating** on familial correlations under Mendelian inheritance (*see* **Mendel's Laws**) and introduced a new method: the **analysis of variance**. Yet again, a third major element of quantitative genetics and also of mathematical statistics is formulated in the context of familial correlations. Fisher [10] is noted for both laying the foundations of biometrical genetics and reconciling this theory with Mendelian genetics.

In 1921, Fisher [11] derived the exact sampling distribution of the product-moment estimator of the intraclass correlation for the balanced case under normality and developed the z-transformation for this estimator. He also showed that the range of values of the intraclass correlation is given by $[-1/(n-1), 1)$ for families of size n, in contrast to the range of $(-1, 1)$ for the simple correlation coefficient. Fisher [12] later argued that "there is probably nothing in the production of a leaf or a child which necessitates that the number in such a family should be less than any number however great, and in the absence of such a necessary restriction we cannot expect to find negative correlations within such families". This restriction has been accepted by some researchers who have adopted estimators of the intraclass correlation that are truncated at zero. But Fisher [12] also noted in his example of card games that where the number of suits is limited to four, the correlation between the number of cards in different suits in the same hand may have negative values down to $-1/3$, and so some researchers do not use truncated estimators.

In 1925, Fisher [12] proposed an estimator of the intraclass correlation based on a ratio of variance components from the one-way analysis of variance (ANOVA) in the balanced case under the assumption of normality. In so doing, Fisher [12] broadened the scope of application of the intraclass correlation which led to its application in **sensitivity analysis** in the design of experiments, intracluster variation in sample surveys and reliability theory in psychology. From Fisher [12], the statistical model for the jth member (or sibling) of the ith group (or family) is

$$x_{ij} = \mu + a_i + e_{ij},$$

where μ is the mean of the observations in the population, the group effects $\{a_i\}$ are normal and identically distributed with mean 0 and variance σ_a^2, the random errors $\{e_{ij}\}$ are normal and identically distributed with mean 0 and variance σ_e^2, and the $\{a_i\}$ and $\{e_{ij}\}$ are independent. The intraclass correlation coefficient parameter is defined to be $\rho_s = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. Furthermore, under Fisher's model [12], it can be shown that $\mathrm{corr}(x_{ij}, x_{ik}) = \rho_s$ for $j \neq k$.

Without explicitly stating the formula, Fisher [12] defined the ANOVA estimator of the intraclass correlation coefficient in the balanced case to be

$$\tilde{\rho}_s = \frac{MSB - MSW}{MSB + (k-1)MSW}, \tag{4}$$

where the mean sum of squares between families is $MSB = \sum_i^k n(\overline{x}_i - \overline{x}_n)^2/(k-1)$ and the mean sum of squares within families is $MSW = \sum_i^k \sum_j^n (x_{ij} - \overline{x}_i)^2/(N-k)$. Note the minimum value that Fisher's ANOVA estimator $\tilde{\rho}_s$ can attain is $-1/(k-1)$, as is the case for Pearson's product-moment estimator r_{p}, which is at odds with the definition of the parameter ρ which takes values on the interval [0, 1). So some researchers use a version of $\tilde{\rho}_s$ truncated at zero.

Fisher [12] commented in comparison that Pearson's product-moment estimator ought to be considered "slightly defective" in that one estimator of a component of variance (*see* **Variance Components**) implicit in the product-moment estimator is biased. The introduction by Fisher [12] of the ANOVA estimator of the intraclass correlation has had the effect of dividing researchers developing and using intraclass correlations into two factions: one faction using estimators based upon the analysis of variance; and the other faction using estimators based upon product moments, that is, those who use the ML method.

Estimation of the Intraclass Correlation Coefficient

Wald [38] and Bhargava [3] independently developed an interval estimator based on a statistic known to follow the ***F* distribution** exactly even in the unbalanced case for **normally distributed** data. By choosing

$$u_i = n_i \frac{1 - \rho_{\mathrm{s}}}{n_i - (n_i - 1)(1 - \rho_{\mathrm{s}})}$$

as a weight associated with the ith family and defining $\overline{x}_u = \sum_i u_i \overline{x}_i / \sum_i u_i$, they showed that the endpoints of a $1-\alpha$ **confidence interval** can be found by solving the nonlinear equation

$$\frac{N-k}{k-1} \frac{\sum_i w_i (\overline{x}_i - \overline{x}_u)^2}{\sum_i \sum_j (x_{ij} - \overline{x}_i)^2} = F_{\gamma,k-1,N-k} \qquad (5)$$

for values of γ equal to $1-\alpha/2$ and $\alpha/2$. By setting $\gamma = 0.5$, a median unbiased estimator of the intraclass correlation can be obtained. The point and interval estimates can be found by iterative algorithms such as the bisection method or the secant method.

Fieller & Smith [8] generalized the ANOVA estimator to the unbalanced case. Smith [30] provided a further generalization of this estimator by allowing arbitrary weights other than the number of offspring when computing the between-family sum of squares. Smith [30] also derived the asymptotic variance of this new generalized weighted ANOVA estimator. Karlin et al. [18] introduced the use of arbitrary weights in a generalization of the product-moment estimator. Eliasziw & Donner [5] provided the asymptotic variance for the generalized weighted product-moment estimator following the method of Smith [30]. These results can be summarized as follows [19]. Associate weights v_i, α_i, and β_i with the ith family. Define $\overline{x}_v = \sum_i v_i x_i / \sum_i v_i$, $SS_i = \sum_j (x_{ij} - \overline{x}_i)^2$, and the following components of variance:

$$SX_v = \sum_i v_i (\overline{x}_i - \overline{x}_v)^2,$$

$$SE_\alpha = \sum \alpha_i SS_i,$$

$$SE_\beta = \sum \beta_i SS_i.$$

Depending on the choice of weights,

$$\tilde{r}_v = \frac{SX_v - SE_\alpha}{SX_v + SE_\beta} \qquad (6)$$

can represent any one of the previously mentioned product-moment or ANOVA point estimators of the intraclass correlation coefficient. Selecting $v_i = n_i(n_i - 1)$ yields the pairwise weights implicitly conceived by Pearson [26]. Selecting $v_i = n_i$ yields the sibship- or group-size weights implicitly conceived by Fisher [11]. Selecting $v_i = 1$ yields the uniform weights explicitly conceived by Smith. Upon selecting v_i, if one sets $\alpha_i = v_i/[n_i(n_i-1)]$ and $\beta_i = v_i/n_i$, the product-moment estimators of Karlin et al. [18] are obtained. Upon selecting v_i, defining $V = \sum_i v_i$, $v_c = \sum_i (v_i - v_i^2/V)/n_i$, and $v_o = (V - \sum_i v_i^2/V)/v_c$, and setting $\alpha_i = v_c/(N-k)$ and $\beta_i = v_c(v_o - 1)/(N-k)$, the ANOVA estimators of Smith [30] are obtained.

Monte Carlo simulations [19, 20] have shown that the efficiency of $\overline{r}_v$ is a function of the parameter value, with some choices of weights being better than others given prior information about the parameter value. The pairwise weights are best for $\rho_s < 0.2$, the uniform weights for $\rho_s > 0.8$, and the group-size weights for the intermediate range. As far as the choice of product moment vs. ANOVA is concerned, the product moment is preferred for pairwise weights and the ANOVA for group-size and uniform weights.

Typically, prior information about the intraclass correlation parameter is unavailable. So following the approach of Srivastava [34] and introducing the weights w_i, κ_i, and λ_i with $W = \sum_i w_i$, $\overline{x}_w = \sum w_i \overline{x}_i / W$, $w_c = \sum_i (w_i - w_i^2/W)/n_i$, and $w_o = (W - \Sigma_i w_i^2/W)/w_c$ for a second estimator

$$\tilde{r}_w = \frac{SX_w - SE_\kappa}{SX_w + SE_\lambda}, \qquad (7)$$

a combination estimator

$$\tilde{\rho}_{vw} = \frac{\tilde{\rho}_v}{1 + \tilde{\rho}_v - \tilde{\rho}_w} \qquad (8)$$

is defined with asymptotic variance given by

$$\begin{aligned} av(\tilde{\rho}_{vw}) = {} & (1 - \tilde{\rho}_{vw})^2 av(\tilde{\rho}_v) \\ & + 2\tilde{\rho}_{vw}(1 - \tilde{\rho}_{vw}) ac(\tilde{\rho}_v, \tilde{\rho}_w) \\ & + \tilde{\rho}_{vw}^2 av(\tilde{\rho}_v). \end{aligned} \qquad (9)$$

The asymptotic variance of either $\tilde{\rho}_v$ or $\tilde{\rho}_w$ and the asymptotic variance of these two estimators can be determined from

$$\begin{aligned} ac(\tilde{r}_v, \tilde{r}_w) = \frac{2(1-\rho)^2}{\psi_v \psi_w} \Bigg\{ & D_{vw} + \sum_i (n_i - 1)\alpha_i \kappa_i \\ & + \sum_i (n_i - 1)[(\alpha_i \lambda_i + \beta_i \kappa_i)\rho + \beta_i \lambda_i \rho^2] \Bigg\}, \end{aligned} \qquad (10)$$

where

$$D_{vw} = \frac{\mathrm{cov}(SX_v, SX_w)}{2\sigma^4} = \sum_{i=1}^{k}\sum_{j=1}^{k} v_i w_j \phi_{vw}^2(i,j),$$

$$\phi_{vw}(i,j) = \delta_{ij}\tau_i - \frac{w_i}{W}\tau_i - \frac{v_j}{V}\tau_j + \sum_l \frac{v_l w_l}{VW}\tau_l,$$

$$\tau_i = \rho + \frac{1-\rho}{n_i},$$

$$\psi_v = v_c + \sum_i (n_i - 1)\beta_i + \left[(v_o - 1)v_c - \sum_i (n_i - 1)\beta_i\right]\rho,$$

$$\psi_w = w_c + \sum_i (n_i - 1)\lambda_i + \left[(w_o - 1)w_c - \sum_i (n_i - 1)\lambda_i\right]\rho$$

with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$ by the approach of Smith [30]. A suitable estimator of ρ when estimating the asymptotic variance of either a product-moment or ANOVA estimator is the estimator itself, whereas a suitable estimate of ρ when estimating the asymptotic variance of the combination estimator is the combination estimator.

Choosing the pairwise weights for v_i and uniform weights for w_i in the combination estimator results in a point estimator that is nearly fully efficient, in comparison with the maximum likelihood estimator, for the full range of the intraclass correlation coefficient [20]. An appreciation of this result can be gained from the equation

$$\tilde{\rho}_{vw} = (1 - \tilde{\rho}_{vw})\tilde{\rho}_v + \tilde{\rho}_{vw}\tilde{\rho}_w,$$

which also defines the estimator $\tilde{\rho}_{vw}$. Note that $\tilde{\rho}_{vw}$ shrinks towards $\tilde{\rho}_v$ for small ρ_s and towards $\tilde{\rho}_w$ for large ρ_s. So it is reasonable to expect that $\tilde{\rho}_{vw}$ will perform well for small ρ_s if $\tilde{\rho}_v$ is efficient for small ρ_s – with a similar consideration for large ρ_s if $\tilde{\rho}_w$ is efficient for large ρ_s.

Estimation of the Parent–Offspring Interclass Correlation Coefficient

The parent–offspring interclass estimator assumes only one parent and a variable number of offspring per family.

Karlin et al. [18] developed a generalized weighted version of the pairwise product-moment estimator. Srivastava & Keen [35] developed a generalized weighted version of the ANOVA estimator given by Srivastava [33] based on the uniform weighting scheme. Srivastava & Keen also derived the asymptotic variance of the weighted ANOVA estimator. The asymptotic variance of the product-moment estimator is given by Eliasziw & Donner [5].

Using the notation above for the sibling intraclass correlation and letting y_i denote the parent's value in the ith family, the weighted product moment and a version of the weighted ANOVA estimators of the interclass correlation can be expressed as

$$\tilde{r}_{ms} = \frac{\sum_i v_i (y_i - \overline{y}_v)(\overline{x}_i - \overline{x}_v)}{\left[\sum_i v_i (y_i - \overline{y}_v)^2\right]^{1/2}} \times \frac{1}{\left[\sum_i v_i (\overline{x}_i - \overline{x}_v)^2 + \sum_i \beta_i SS_i\right]^{1/2}}, \quad (11)$$

where $\overline{y}_v = (v_1 y_1 + v_2 y_2 + \cdots + v_k y_k)/V$. Upon selecting v_i, if one sets $\beta_i = v_i/n_i$, then the product-moment estimators are obtained, or if one sets $\beta_i = v_c(v_o - 1)/(N - k)$, then versions of the ANOVA estimators are obtained.

The choices of weights that have been studied in Monte Carlo simulations [5, 35] are the sibship-size weights ($v_i = n_i$) and the uniform weights ($v_i = 1$), with the recommendation of using sibship-size weights when the sibling intraclass correlation $\rho_s < 0.3$ and uniform weights otherwise.

When the sibship-size weights are used, the resulting product-moment and ANOVA parent–offspring interclass correlation estimators have been referred to as pairwise estimators. It must be understood that the pairs are taken with respect to parent and offspring.

A number of other parent–offspring estimators have appeared in the literature. On the basis of Monte

Carlo simulations, Rosner et al. [29] did not recommend the sib-mean and random-sib mean estimators – recommending instead the pairwise and ensemble estimators. (The ensemble estimator is obtained by setting $v_i = 1$ and $\beta_i = (k-1)/(kn_i)$ in the formula for $\tilde{r}_{ms}$.) However, Srivastava & Keen [35] established theoretically, under normality for the uniform ANOVA estimator $\tilde{\rho}_{ms,u}$ and the ensemble estimator $\tilde{\rho}_{ms,e}$, that $av(\tilde{\rho}_{ms,u}) \leq av(\tilde{\rho}_{ms,e})$. In Monte Carlo comparisons of the asymptotic variance formulas for $\tilde{\rho}_{ms,u}$ and $\tilde{\rho}_{ms,e}$, Eliasziw & Donner [5] reached the same conclusion.

The exact sampling distribution of the uniform ANOVA estimator of the interclass correlation is given by Velu & Rao [37].

In parallel with the intraclass correlation, Eliasziw & Donner [6] showed for the interclass correlation that the pairwise product-moment estimator has a smaller asymptotic variance than the pairwise ANOVA estimator.

Maximum Likelihood Estimation of Familial Correlations

Elston [7] derived expressions for the **maximum likelihood** estimators of correlations in nuclear families for the balanced case, together with expressions for the asymptotic variances of these estimators. From this flows the justification of the pairwise estimators of sibling and parent–offspring correlations in the general unbalanced situation, reducing to the maximum likelihood estimators in the balanced case.

In the unbalanced case, closed-form expressions are not available for correlations in nuclear families and it is necessary to resort to iterative numerical algorithms [2, 27, 28, 31, 32, 36, 39] to find these estimates. Nevertheless, closed-form expressions are available for the asymptotic covariances of all parameters (including sibling intraclass correlation) in data consisting of siblings only [4] and for all parameters (including brother intraclass correlation, sister intraclass correlation and brother–sister interclass correlation) in data consisting of brother and sisters only [21]. Of course, the parent–offspring situation can be regarded as a special case of the brother–sister situation.

Estimation of Multivariate Correlations in Extended Families

Suppose that there are q classes of relatives to be considered and a random sample of k independent families is available. Suppose the jth family has n_{ij} observations for the ith class with $k_{ij} > 0$ for at least one class i so the jth family is nonempty. Here the groups may represent a three-generation family or even an extended pedigree structure. Assume the vector of observations for a member of the ith class of relatives has length p_i so the variables measured on members of each class may not be identical either in quality or quantity. Let $\mathbf{x}_{ijs}$ denote the random vector of p_i observations for the sth member in the ith class of relatives in the jth family. The first and second moment structure is defined by

$$\begin{aligned}
\boldsymbol{\mu}_i &= \mathrm{E}(\mathbf{x}_{ijs}),\\
\mathbf{A}_i &= (A_{i\alpha\beta}) = \mathrm{cov}(\mathbf{x}_{ijs}),\\
\mathbf{B}_i &= (B_{i\alpha\beta}) = \mathrm{cov}(\mathbf{x}_{ijs}, \mathbf{x}_{ijt}), \quad s \neq t,\\
\mathbf{C}_{im} &= (C_{im\alpha\beta}) = \mathrm{cov}(\mathbf{x}_{ijs}, \mathbf{x}_{mjt}), \quad i \neq m,
\end{aligned}$$

for $1 \leq i \leq q$ and $1 \leq m \leq q$. The mean vector $\boldsymbol{\mu}_i$ is of length p_i, the $p_i \times p_i$ matrix $\mathbf{A}_i$ is the class covariance matrix of the ith class, the $p_i \times p_i$ matrix $\mathbf{B}_i$ is the intraclass covariance matrix of the ith class, and the $p_i \times p_m$ matrix $\mathbf{C}_{im}$ is the interclass **covariance matrix** between classes. Note that the class and intraclass covariance matrices are symmetric. Although $\mathbf{C}_{im}$ is not symmetric, it is easily verified that $\mathbf{C}_{im} = \mathbf{C}'_{mi}$.

Upon defining the diagonal matrix composed of class variances

$$\mathbf{D}_i = \begin{pmatrix}
A_{i11} & 0 & 0 & \cdots & 0 & \\
0 & A_{i22} & 0 & \cdots & 0 & \\
0 & 0 & A_{i33} & \cdots & 0 & \\
\vdots & \vdots & \vdots & \ddots & \vdots & \\
0 & 0 & 0 & \cdots & & A_{ip_ip_i}
\end{pmatrix},$$

the correlation structure for the model is

$$\begin{aligned}
\boldsymbol{\Lambda}_i &= \mathrm{corr}(\mathbf{x}_{ijs}) = \mathbf{D}_i^{-1/2}\mathbf{A}_i\mathbf{D}_i^{-1/2},\\
\boldsymbol{\Omega}_i &= \mathrm{corr}(\mathbf{x}_{ijs}, \mathbf{x}_{ijt}) = \mathbf{D}_i^{-1/2}\mathbf{B}_i\mathbf{D}_i^{-1/2}, \quad s \neq t,\\
\boldsymbol{\Phi}_{im} &= \mathrm{corr}(\mathbf{x}_{ijs}, \mathbf{x}_{mjt}) = \mathbf{D}_i^{-1/2}\mathbf{C}_{im}\mathbf{D}_m^{-1/2},
\end{aligned}$$

provided $i \neq m$, where $\boldsymbol{\Lambda}_i$ is the class correlation matrix for class i, $\boldsymbol{\Omega}_i$ is the intraclass correlation

matrix for class i, and $\mathbf{\Phi}_{im}$ is the interclass correlation coefficient between classes i and m. The parameters in this model can be estimated in the case of multivariate normality by any number of numerical methods currently available, including the Newton–Raphson method [27, 36]. Standard errors are found by inverting the Hessian matrix (of second derivatives) of the likelihood function evaluated at the maximum likelihood estimates.

Konishi & Khatri [23] have provided generalized noniterative estimators of correlation matrices in the multivariate parent–offspring case based on the product-moment and ANOVA approaches including the uniformly–weighted ANOVA estimators introduced by Srivastava et al. [36].

Konishi et al. [25] discussed the use of **canonical correlations** to find suitable measures for the degree of resemblance between parent and offspring. Konishi & Rao [24] discussed the application of **principal components analysis** for multivariate familial data. Both discussions use the generalized estimators of Konishi & Khatri [23].

Other Estimation Issues

Although the preceding discussion is with respect to continuous random variables, with the exception of the maximum likelihood estimators, the preceding estimators and their standard errors are valid without modification for **binary** random variables and, in the multivariate setting, for combinations of continuous and binary random variables.

With regard to dropping the requirement for the assumption of normality, a discussion of robust M-estimation of the intraclass correlation is given by Bansal & Bhandary [1]. Alternatively, one can return to the early work of Galton, and replace means by **medians** and standard deviation by probable errors (interquartile ranges) in the product-moment estimators.

References

[1] Bansal, N.K. & Bhandary, M. (1994). Robust M-estimation of the intraclass correlation coefficient, *Australian Journal of Statistics* **36**, 287–301.

[2] Bener, A. & Huda, S. (1987). Maximum likelihood estimation of components of variance and correlations in the analysis of family data, *Annals of Human Genetics* **51**, 259–264.

[3] Bhargava, R.P. (1946). Tests of significance for intraclass correlation when family sizes are not equal, *Sankhyā*, **7**, 435–438.

[4] Donner, A. & Koval, J.J. (1980). The large sample variance of an intraclass correlation, *Biometrika* **67**, 719–722.

[5] Eliasziw, M. & Donner, A. (1991). Generalized estimators of familial correlations, *Annals of Human Genetics* **55**, 77–90.

[6] Eliasziw, M. & Donner, A. (1995). Evaluation of a proposed modification to the pairwise estimator of a parent–offspring correlation, *Sankhyā, Series B* **57**, 151–157.

[7] Elston, R.C. (1975). On the correlation between correlations, *Biometrika* **62**, 133–140.

[8] Fieller, E.C. & Smith, C.A.B. (1951). Note on the analysis of variance and intraclass correlation, *Annals of Eugenics* **16**, 97–105.

[9] Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika* **10**, 507–521.

[10] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[11] Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample, *Metron* **1**, 3–32.

[12] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[13] Galton, F. (1877). Typical laws of heredity, *Proceedings of the Royal Institution* **8**, 282–301.

[14] Galton, F. (1885). Presidential Address, Section H, Anthropology, *British Association Reports* **55**, 1206–1214.

[15] Galton, F. (1886). Family likeness in stature, *Proceedings of the Royal Society* **40**, 42–73.

[16] Galton, F. (1889). Co-relations and their measurement, chiefly from anthropometric data, *Proceedings of the Royal Society* **45**, 135–145.

[17] Harris, J.A. (1913). On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large, *Biometrika* **9**, 446–472.

[18] Karlin, S., Cameron, E.C. & Williams, P.T. (1981). Sibling and parent–offspring correlation estimation with variable family size, *Proceedings of the National Academy of Science* **78**, 2664–2668.

[19] Keen, K.J. (1993). Estimating the correlation among siblings, *Annals of Human Genetics* **57**, 297–305.

[20] Keen, K.J. (1996). Limiting the effects of single member families in the estimation of the intraclass coefficient, *Biometrics* **52**, 823–832.

[21] Keen, K.J. & Srivastava, M.S. (1991). The asymptotic variance of the interclass correlation coefficient, *Biometrika* **78**, 225–228.

[22] Konishi, S. (1985). Asymptotic properties of estimators of interclass correlation from familial data, *Annals of*

the Institute of Statistical Mathematics, Series A **34**, 505–551.

[23] Konishi, S. & Khatri, C.G. (1990). Inferences on interclass and intraclass correlations in multivariate familial data, *Annals of the Institute of Statistical Mathematics, Series A* **34**, 561–580.

[24] Konishi, S. & Rao, C.R. (1992). Principal component analysis for multivariate familial data, *Biometrika* **79**, 631–641.

[25] Konishi, S., Khatri, C.G. & Rao, C.R. (1991). Inferences on multivariate measures of interclass and intraclass correlations in familial data, *Journal of the Royal Statistical Society, Series B* **53**, 649–659.

[26] Pearson, K. (1896). Mathematical contributions to the theory of evolution – III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society, Series A* **187**, 253–318.

[27] Rao, D.C., Vogler, G.P., McGue, M. & Russell, J.M. (1987). Maximum-likelihood estimation of familial correlations from multivariate quantitative data on pedigrees: a general method and examples, *American Journal of Human Genetics* **41**, 1104–1116.

[28] Rosner, B. (1979). Maximum likelihood estimation of interclass correlations, *Biometrika* **66**, 533–538.

[29] Rosner, B., Donner, A. & Hennekens, C.H. (1977). Estimation of interclass correlation from familial data, *Applied Statistics* **26**, 179–187.

[30] Smith, C.A.B. (1957). On the estimation of intraclass correlations, *Annals of Human Genetics* **21**, 363–373.

[31] Smith, C.A.B. (1980). Estimating genetic correlations, *Annals of Human Genetics* **43**, 265–284.

[32] Smith, C.A.B. (1980). Further remarks on estimating genetic correlations, *Annals of Human Genetics* **44**, 95–105.

[33] Srivastava, M.S. (1984). Estimation of interclass correlations in familial data, *Biometrika* **71**, 177–185.

[34] Srivastava, M.S. (1993). Estimation of the intraclass correlation coefficient, *Annals of Human Genetics* **57**, 159–165.

[35] Srivastava, M.S. & Keen, K.J. (1988). Estimation of the interclass correlation coefficient, *Biometrika* **75**, 731–739.

[36] Srivastava, M.S., Keen, K.J. & Katapa, R.S. (1988). Estimation of interclass and intraclass correlations in multivariate familial data, *Biometrics* **44**, 141–150.

[37] Velu, R. & Rao, M.B. (1990). Estimation of parent–offspring correlation, *Biometrika* **77**, 557–562.

[38] Wald, A. (1940). A note on the analysis of variance with unequal class frequencies, *Annals of Mathematical Statistics* **11**, 96–100.

[39] Wette, R., McGue, M.K., Rao, D.C. & Cloninger, C.R. (1988). Maximum likelihood estimators, *Biometrics* **44**, 717–725.

(*See also* **Genetic Correlations and Covariances**; **Path Analysis**)

K.J. KEEN

Family History Validation

A fundamental source of data for any **genetic epidemiologic** study is the family history of the research subject [10, 12, 19, 28]. The essential family history consists of two components: the pedigree structure and the specification of those members of the pedigree who have the trait(s) or disease(s) of investigative interest. Increasingly, the family history also includes information on relatives, such as age of disease onset or diagnosis, or environmental exposures such as tobacco and alcohol use. The family history is obtained by direct interview of or questionnaire administration to an informant, usually the proband or index case (*see* **Ascertainment**). If the proband is deceased, a surrogate or proxy (spouse or other next of kin) is asked to provide the information.

The quality of the family history thus obtained can vary according to the amount of information sought, the way in which the information is collected, and who is providing the information. Depending upon the nature of the study, it also follows that if the family history information sought is detailed or needs to be of high quality, then the process to obtain such data can be time consuming for both the informant and the investigator [17]. A key methodologic issue in this regard is validation of the completeness or accuracy of the information provided by the informant. In turn, what is learned from empirical studies of family history validation can inform the study design and analytic strategies to compensate for missing or inaccurate data.

Validation of Reported Family History

There are two major methods by which a reported family history can be validated:

1. Direct query or health assessment (by interview or questionnaire) of the family members about whom the informant has provided information. A variation of this approach is to interview as many members of the kindred as feasible and compare reported family histories.
2. Obtain documentation of disease in family members by medical records, death certificates, or population-based disease registries.

Table 1 Assessment of agreement of informants' reported family history

	Family member's self-report or medical documentation that family member is:	
Informant reports that family member is:	Affected	Unaffected
Affected	a	b
Unaffected	c	d

When such information has been obtained, family histories are validated by designating as the "gold standard" the self-report given by the relative (Method 1) or the medical documentation (Method 2). The agreement or concordance of the informant's reported family history is then assessed by **sensitivity** $= a/(a+b)$ and/or **specificity** $= d/(c+d)$, according to Table 1.

In general, use of Method 1 can provide fairly complete data to estimate both sensitivity and specificity, and Method 2 yields good data for estimating sensitivity, but less so for specificity – except perhaps when population-based disease registries are employed for validation. The quality of sensitivity and specificity estimates are thus dependent upon the quality and completeness of the "gold standard" data.

Family History Validation Studies

There have been numerous studies that performed validation of reported family histories. The choice of validation method and the accuracy of reported history are often dependent upon the disease under study. Studies of psychiatric disorders [5, 13, 16, 26] and one study of aneurysms [8] validated family history by Method 1, interviewing or directly assessing relatives in the kindreds. These studies found that reported family histories were of variable quality, and that specificity was generally high, but sensitivity was modest.

In contrast, studies of cancer and heart disease have often utilized Method 2, obtaining medical records or pathology reports and death certificates [1–4, 6, 14, 17, 18, 23, 24]. Large-scale, population-based family history validation studies are possible because of the existence of regional or national tumor registries [11]. In general, these studies have found that sensitivity of reporting cancer family history is

high, greater than 70%. As mentioned above, given the type of information used to assess accuracy, specificity often cannot be properly estimated, unless documentation is exhaustive.

An interesting variation on family history validation was conducted by Glanz et al. [7] in which first-degree relatives of patients who were known to have colorectal cancer were asked to complete a family history survey. (The relatives were recruited under the auspices of a different research question.) All respondents were known beforehand to have a relative with colorectal cancer, yet the investigators found that family history of this cancer was underreported by 25.4%.

Factors that Influence Accuracy of Reported Family History

Many of the above studies also investigated the factors that affected the accuracy of family history of disease. It is clear that the type of disease under investigation greatly influences accuracy; how visible the condition is, and the degree of stigma that is associated with the condition (both socially and in the family) have an impact. Personal characteristics of the respondent, including age, gender, affection status, comorbid conditions, education level, and knowledge and concern about the condition can influence accuracy. Finally, degree of relationship between the respondent and the relative being reported on affects accuracy: it has been confirmed in a number of studies that the more distant the relationship, the less accurate the history.

In the context of **case–control studies**, when the validity of reported exposures – such as tobacco or alcohol consumption by family members – has been investigated, the accuracy has been reported to be reasonably high, ranging from 60% to 90% agreement by next of kin or close proxy respondents [9, 15, 20]. However, reports about all relatives across kindreds have not been systematically examined.

Analytic Implications

The results of family history validation studies to date suggest that family histories of some diseases can often be considered reliable, particularly of major cancers or common/familiar conditions among first-degree relatives. However, these studies also suggest that recall **bias** and incomplete data collection can occur, and that the method of querying, demographic and clinical characteristics of the informant, and psychosocial factors may influence the accuracy of the reported history.

The consequence of an inaccurate family history is misclassification or missing data, which can severely influence the results of the analysis [25, 27]. Saito et al. [21, 22] showed the bias that can occur in the estimation of **odds ratios** for family history of stroke and diabetes when the gender and age of the relatives is ignored. Silberberg et al. [24] found that classification of "Don't know" responses by informants with respect to family history of cancer or heart disease about second-degree relatives could result in lower sensitivity. In conclusion, an awareness of these issues should help to guide the design, analysis, and imputation of missing data in studies that use family histories.

References

[1] Airewele, G., Adatto, P., Cunningham, J., Mastromarino, C., Spencer, C., Sharp, M., Sigurdson, A. & Bondy, M. (1998). Family history of cancer in patients with glioma – a validation study of accuracy, *Journal of the National Cancer Institute* **90**, 543–544.

[2] Aitken, J., Bain, C., Ward, M., Siskind, V. & MacLennan, R. (1995). How accurate is self-reported family history of colorectal cancer?, *American Journal of Epidemiology* **141**, 863–871.

[3] Anton-Culver, H., Kurosaki, T., Taylor, T.H., Gildea, M., Brunner, D. & Bringman, D. (1996). Validation of family history of breast cancer and identification of the BRCA1 and other syndromes using a population-based cancer registry, *Genetic Epidemiology* **13**, 193–205.

[4] Bondy, M.L., Strom, S.S., Colopy, M.W., Brown, B.W. & Strong, L.C. (1994). Accuracy of family history obtained through interviews with relatives of patients with childhood sarcoma, *Journal of Clinical Epidemiology* **47**, 89–96.

[5] Davies, N.J., Sham, P.C., Gilvarry, C., Jones, P.B. & Murray, R.M. (1997). Comparison of the family history with the family study method: report from the Camberwell Collaborative Psychosis Study, *American Journal of Medical Genetics* **74**, 12–17.

[6] Douglas, F.S., O'Dair, L.C., Robinson, M., Evans, D.G.R. & Lynch, S.A. (1999). The accuracy of diagnoses as reported in families with cancer: a retrospective study, *Journal of Medical Genetics* **36**, 309–12.

[7] Glanz, K., Grove J., Le Marchand, L. & Gotay, C. (1999). Underreporting of family history of colon cancer: correlates and implications, *Cancer Epidemiology, Biomarkers and Prevention* **8**, 635–639.

[8] Greebe, P., Bromberg, J.E.C., Rinkel, G.J.E., Algra, A. & van Gijn, J. (1997). Family history of subarachnoid haemorrhage: supplemental value of scrutinising all relatives, *Journal of Neurology, Neurosurgery and Psychiatry* **62**, 273–275.

[9] Herrmann, N. (1985). Retrospective information from questionnaires. I. Comparability of primary respondents and their next-of-kin, *American Journal of Epidemiology* **121**, 937–947.

[10] Hunt, S.C., Williams, R.R. & Barlow, G.K. (1986). A comparison of positive family history definitions for defining risk of future disease, *Journal of Chronic Diseases* **39**, 809–821.

[11] Kerber, R.A. & Slattery, M.L. (1997). Comparison of self-reported and database-linked family history of cancer data in a case-control study, *American Journal of Epidemiology* **146**, 244–248.

[12] Khoury, M.J., Beaty, T.H. & Cohen, B.H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press, New York.

[13] Kosten, T.A., Anton, S.F. & Rounsaville, B.J. (1992). Ascertaining psychiatric diagnoses with the family history method in a substance abuse population, *Journal of Psychiatric Research* **26**, 135–147.

[14] Love, R.R., Evans, A.M. & Josten, D.M. (1985). The accuracy of patient reports of a family history of cancer, *Journal of Chronic Diseases* **38**, 289–293.

[15] McLaughlin, J.K., Mandel, J.S., Mehl, E.S. & Blot, W.J. (1990). Comparison of next-of-kin with self-respondents regarding questions on cigarette, coffee and alcohol consumption, *Epidemiology* **1**, 408–412.

[16] Mendlewicz, J., Fleiss, J.L., Cataldo, M. & Rainer, J.D. (1975). Accuracy of the family history method in affective illness. Comparison with direct interviews in family studies, *Archives of General Psychiatry* **32**, 309–314.

[17] Novakovic, B., Goldstein, A.M. & Tucker, A.M. (1996). Validation of family history in deceased family members, *Journal of the National Cancer Institute* **88**, 1492–1493.

[18] Parent, M.-E., Ghadirian, P., Lacroix, A. & Perret, C. (1997). The reliability of recollections of family history: implications for the medical provider, *Journal of Cancer Education* **12**, 114–120.

[19] Phillips, P.H., Linet, M.S. & Harris, E.L. (1991). Assessment of family history information in case-control cancer studies, *American Journal of Epidemiology* **133**, 757–765.

[20] Rocca, W.A., Fratiglioni L., Bracco, L., Pedone, D., Groppi, C. & Schoenberg, B.S. (1986). The use of surrogate respondents to obtain questionnaire data in case-control studies of neurologic diseases, *Journal of Chronic Diseases* **39**, 907–912.

[21] Saito, T., Furukawa, T., Nanri, S. & Saito, I. (1999). Potential errors resulting from sex and age difference in assessing family history of diabetes, *Preventive Medicine* **28**, 33–39.

[22] Saito, T., Nanri, S., Saito, I. & Furukawa, T. (2000). Importance of sex and age factor in assessing family history of stroke, *Journal of Epidemiology* **10**, 328–334.

[23] Sijmons, R.H., Boonstra, A.E., Reefhuis, J., Hordijk-Hos, J.M., de Walle, H.E.K., Oosterwijk, J.C. & Cornel, M.C. (2000). Accuracy of family history of cancer: clinical genetic implications, *European Journal of Human Genetics* **8**, 181–186.

[24] Silberberg, J., Wlodarczyk, J., Hensley, M., Ray, C., Alexander, H., Basta, M. & Hughes, J. (1994). Accuracy of reported family history of heart disease: the impact of "don't know" responses, *Australian and New Zealand Journal of Medicine* **24**, 386–389.

[25] Szatmari, P. & Jones, M.B. (1999). Effects of misclassification on estimates of relative risk in family history studies, *Genetic Epidemiology* **16**, 368–381.

[26] Thompson, W.D., Orvaschel, H., Prusoff, B.A. & Kidd, K.K. (1982). An evaluation of the family history method for ascertaining psychiatric disorders, *Archives of General Psychiatry* **39**, 53–58.

[27] Tsuang, M.T., Lyons, M.J. & Faraone, S.V. (1987). Problems of diagnoses in family studies, *Journal of Psychiatric Research* **21**, 391–399.

[28] Williams, R.R., Hunt, S.C., Barlow, G.K., Chamberlain, R.M., Weinberg, A.D., Cooper, H.P., Carbonari, J.P. & Gotto, A.M. Jr (1988). Health family trees: a tool for finding and helping young family members of coronary and cancer prone pedigrees in Texas and Utah, *American Journal of Public Health* **78**, 1283–1286.

GLORIA M. PETERSEN

Family-based Association for Quantitative Traits

Allelic **association** tests have been introduced to analyze the relationship between allelic variability and individual traits (*see* **Gene**). As with ordinary association tests, population **admixture** is a potential source of confounding. Below we describe family-based approaches to adjusting for potential confounding due to population admixture when analyzing a quantitative (continuous) trait. These family-based association tests often aim to extend the transmission/disequilibrium test (TDT) [13, 15] from a dichotomous (affectation status) trait to a quantitative trait (*see* **Family-based Case–Control Studies**).

The approaches can be differentiated by the kind of assumptions they make on the distribution of the quantitative trait (e.g. normal distribution), the ascertainment of the sample (e.g. random sample), and the kind of data that can be evaluated (e.g. only family trios). Most of the approaches test for **linkage** in the presence of association but there also exist tests for association in the presence of linkage.

First we review papers that introduce tests that make assumptions on the distribution of the quantitative trait given the **genotypes**. Later we will turn to nonparametric tests that do not make assumptions on the distribution of the quantitative trait.

Allison [2] developed five test statistics for samples of independent nuclear-family trios. The first test (TDT_{Q1}) is a t-test that assumes random sampling of the family data. The other four tests ($TDT_{Q2}-TDT_{Q5}$) can be used both for random and extreme sampling. Three of the tests (TDT_{Q1}, TDT_{Q3}, TDT_{Q5}) are ordinary least squares approaches, which assume that the residuals are independent and normally distributed; the tests tend to be quite robust to nonnormality because the central limit theorem ensures that the coefficient estimators are asymptotically normally distributed. TDT_{Q5} was found to be more powerful than the other tests under a variety of genetic models. It is an F-test that compares the fit of two regression models. For the three informative mating types (Aa × AA, Aa × Aa or Aa × aa), the first model regresses the offspring phenotype value on the parental mating types. The second model regresses the offspring phenotypes on both the parental mating types and the offspring genotypes.

For data with siblings and no parents, Allison et al. [3] proposed two tests: a mixed effect model and a permutation test, which does not make assumptions on the distribution of the quantitative trait. Simulation studies have shown that the permutation test is more powerful than the mixed effect model test for additive or nearly additive quantitative trait loci.

Xiong et al. [18] developed an approach that is similar to the one by Allison [2]. The test assumes that parental genotype information is available and that the phenotype is normally distributed or that the sample size is large. The test, which allows for more than one child per family, compares the average trait values of offspring inheriting one allele vs. the other from **heterozygous** parents. George et al. [6] proposed a **regression**-based TDT method, which regresses the trait on the parental transmission of a marker allele. Similarly, Zhu & Elston [22] developed a TDT method for quantitative traits by defining a linear transformation to condition out founder information. Both methods allow one to analyze pedigree data since they do not assume independence of observations. Both tests assume normally distributed **residuals** and random sampling. Simulation studies comparing the two regression methods have been described in [21].

Yang et al. [19] defined a different regression-based approach to adjusting for confounding due to population admixture. Their approach involves augmenting linear regression models with additional regressors that are defined through family genotype data. This ensures that the estimates of the regression coefficients that parameterize the influence of allelic variability on the trait are **unbiased**. The basic assumption that underlies the approach is that the offspring genotypes and the residual trait value are independent conditional on the parental genotypes. The validity of the approach has only been derived for settings where individuals are sampled at random. The approach can be extended to general pedigrees and missing parental genotype information.

Clayton & Jones [4] use a **likelihood** framework to derive a quantitative trait TDT-type test for trio families. The authors assume that the quantitative trait for a given individual has a normal distribution conditional on the individual's genotype. Furthermore, they assume that on some arbitrary scale, the genotype effects on the trait mean may be decomposed into sums of **haplotype** effects: if for genotype (i, j) the corresponding mean trait value is denoted $\mu_{(i,j)}$,

then they assume the existence of a function $h(\cdot)$ such that $h[\mu_{(i,j)}] = h(\mu) + \beta_i + \beta_j$. The probability of the offspring genotype data is computed conditional on the parental genotypes (to protect against population stratification effects) and on the offspring trait values (to allow one to use data irrespective of the ascertainment scheme). Finally, the authors derive a score statistic that is similar to the distribution-free test introduced by Rabinowitz [10] (see below).

For normally distributed traits and complete parental genotypes, Van den Oord [16] describes a framework for identifying quantitative trait loci in association studies using structural equation modeling. Fulker et al. [5] introduce a **variance component** model for a combined quantitative trait locus (QTL) linkage and association analysis. The authors simultaneously model the means and covariances of sibling pairs. The linkage test is based on differences in covariances according to the identity-by-descent (ibd) status at the candidate locus. The model for the means is partitioned into between- and within-pairs (i.e. inter- and intra-sibship) components; an association test based on the within-pair component is robust to population stratification. The variance component model is not directly applicable to either samples selected for extreme trait values or nonmoral quantitative phenotypes. Abecasis et al. [1] and Sham et al. [12] extend this variance component approach to nuclear families of any size and to arbitrary sibships, respectively. For randomly ascertained sibships, Sham et al. [12] found that the power of their association test is related to the QTL **heritability** and the square of the **linkage disequilibrium** measure, while the power of the linkage method is related to the square of the QTL heritability.

We now discuss approaches that make fewer or no assumptions on the distribution of the quantitative trait. Rabinowitz [10] introduced an approach that does not make any distributional assumptions on the quantitative trait values. The essence of the approach is to start with a statistic for association between a **marker** and phenotypes and then to use parental information to modify the statistic to avoid the possibility of spurious association induced by population admixture. Because no assumption is made about the distribution of the trait values under this approach, the tests are valid for any type of sampling schemes based on the phenotypes of the individuals. Denote by n the number of families and by n_i the number of offspring in the ith family. Denote the trait value of the jth offspring in the ith family by $Q_{i,j}$. Following the notation outlined in Zhao [20], define the following index function $Y^{\mathrm{m}}_{i,j} = 1/2$ or $(-1/2)$ if the mother in the ith family is heterozygous and transmits the A (or a) allele to the jth offspring, and $Y^{\mathrm{m}}_{i,j} = 0$ if the mother is homozygous; similarly, define $Y^{\mathrm{f}}_{i,j}$ for the father. Under the **null hypothesis** of no linkage between the marker locus and the quantitative trait loci, the trait value and the index functions $Y^{\mathrm{m}}_{i,j}$ and $Y^{\mathrm{f}}_{i,j}$ are conditionally independent given the parental alleles. Thus, for any constant c, conditional on the trait values and the parental genotypes

$$s(c) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (Q_{i,j} - c)(Y^{\mathrm{f}}_{i,j} + Y^{\mathrm{m}}_{i,j}) \qquad (1)$$

has mean 0. The test statistic proposed by Rabinowitz takes the form $s(c)/\sigma(c)$, where $\sigma(c)$ is an estimate of the conditional variance of $s(c)$. Rabinowitz [10] suggests using the trait average of all the children in all the families to replace c. This approach was generalized by several authors to include families with missing parental information; see Sun et al. [14], Rabinowitz & Laird [11], and Horvath et al. [7].

Waldman et al. [17] propose a **logistic-regression**-based extension of the TDT that also allows one to examine the relation between a marker and one or more continuous or categorical explanatory variables. The method models allelic transmission as the dependent variable and the quantitative and categorical explanatory variables as the independent variables.

Apart from those described in [6, 21] and [22], the tests described thus far test for linkage. However, there are situations when it may be advantageous to test for association in the presence of linkage, e.g. if a chromosomal region is shown to be linked to a trait, association tests may be useful for further localization of a susceptibility locus. For this setting, Monks & Kaplan [9] develop three family-based tests that avoid finding spurious association due to population stratification. The first test (T_{QP}), which adapts the test by Rabinowitz [10] to this new setting, uses genotype information of the parents and all of their children. The second test, T_{QS}, ignores parental information and uses only sibship information. Finally, the third test, T_{QPS}, is a combination of the previous two test types. These tests

make no assumption on the distribution of the quantitative trait and allow one to use information from all available offspring. Lake et al. [8] have generalized the method described in Rabinowitz & Laird [11] to allow tests of association of a quantitative trait in the presence of linkage between the marker and the trait gene, which permits arbitrary family configurations.

References

[1] Abecasis, G.R., Cardon, L.R. & Cookson, W.O.C. (2000). A general test of association for quantitative traits in nuclear families, *American Journal of Human Genetics* **66**, 279–292.

[2] Allison, D.B. (1997). Transmission-disequilibrium tests for quantitative traits, *American Journal of Human Genetics* **60**, 676–690.

[3] Allison, D.B., Heo, M., Kaplan, N. & Martin, E.R. (1999). Sibling-based tests of linkage and association for quantitative traits, *American Journal of Human Genetics* **64**, 1754–1763.

[4] Clayton, D. & Jones, H. (1999). Transmission/disequilibrium tests for extended marker haplotypes, *American Journal of Human Genetics* **65**, 1161–1169.

[5] Fulker, D.W., Cherny, S.S., Sham, P.C. & Hewitt, J.K. (1999). Combined linkage and association sib-pair analysis for quantitative traits, *American Journal of Human Genetics* **64**, 259–267.

[6] George, V., Tiwari, H.K., Zhu, X. & Elston, R.C. (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression, *American Journal of Human Genetics* **65**, 236–245.

[7] Horvath, S., Xu, X. & Laird, N.M. (2001). The family based association test method: strategies for studying general genotype–phenotype associations, *European Journal of Human Genetics* **9**, 301–306.

[8] Lake, S., Blacker, D. & Laird, N.M. (2000). Family-based tests of association in the presence of linkage, *American Journal of Human Genetics* **67**, 1515–1525.

[9] Monks, S. & Kaplan, N.L. (2000). Removing the sampling restrictions from family-based tests of association for a quantitative trait locus, *American Journal of Human Genetics* **66**, 576–592.

[10] Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci, *Human Heredity* **47**, 342–350.

[11] Rabinowitz, D. & Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information, *Human Heredity* **50**, 211–223.

[12] Sham, P.C., Cherny, S.S., Purcell, S. & Hewitt, J.K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-component models, for sibship data, *American Journal of Human Genetics* **66**, 1616–1630.

[13] Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus, *American Journal of Human Genetics* **52**, 506–516.

[14] Sun, F.Z., Flanders, W.D., Yang, Q.H. & Zhao, H.Y. (2000). Transmission/disequilibrium tests for quantitative traits, *Annals of Human Genetics* **64**, 555–565.

[15] Terwilliger, J. & Ott, J. (1992). A haplotype based "haplotype relative risk" approach to detecting allelic associations, *Human Heredity* **42**, 337–346.

[16] Van den Oord, E.J. (2000). Framework for identifying quantitative trait loci in association studies using structural equation modeling, *Genetic Epidemiology* **18**, 341–359.

[17] Waldman, I.D., Robinson, B.F. & Rowe, D.S. (1999). A logistic regression based extension of the TDT for continuous and categorical traits, *Annals of Human Genetics* **63**, 329–340.

[18] Xiong, M.M., Krushkal, J. & Boerwinkle, E. (1998). TDT statistics for mapping quantitative trait loci, *Annals of Human Genetics* **62**, 431–452.

[19] Yang, Q., Rabinowitz, D., Isasi, C. & Shea, S. (2000). Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits, *Human Heredity* **50**, 227–233.

[20] Zhao, H. (2000). Family-based association studies, *Statistical Methods in Medical Research* **9**, 536–587.

[21] Zhu, X. & Elston, R.C. (2000). Power comparison of regression methods to test quantitative traits for association and linkage, *Genetic Epidemiology* **18**, 322–330.

[22] Zhu, X. & Elston, R.C. (2001). Transmission/disequilibrium tests for quantitative traits, *Genetic Epidemiology* **20**, 57–74.

STEVE HORVATH & NAN M. LAIRD

Family-based Case–Control Studies

Case–control studies are used often in epidemiologic studies to investigate the **association** between disease and one or more risk factors. With increasing frequency, the set of risk factors being considered includes **genotypes** at one or more susceptibility, candidate, or **marker** loci. The goals of association studies will differ, depending on the state of knowledge about a given disease. For example, once a susceptibility locus has been identified, the goals include estimating the **relative risk** and **penetrance** associated with specific **mutations**, and testing for **interaction** with environmental exposures or other genes [14] (*see* **Disease-marker Association**; **Gene-environment Interaction**). If a candidate locus has been identified on the basis of a biological hypothesis that relates gene function to phenotypic expression, then the primary goal is testing the **null hypothesis** of no association between the locus and disease. Finally, multiple tests of association with finely spaced markers in a chromosomal region that has been previously established to contain loci linked to the disease (*see* **Linkage Analysis, Model-based**) may be used in the hope of detecting **linkage disequilibrium** with a disease locus.

For diseases, **candidate genes** may be part of a larger hypothesized disease pathway that includes other genes and/or environmental exposures. For example, Gilliland et al. [13] have hypothesized that asthma and other respiratory phenotypes are related to air pollution through a pathway of oxidative stress that includes several genetic loci (*see* **Causation**). Even for BRCA1, a major susceptibility locus for breast cancer that substantially increases risk by itself, there is evidence for some effect modification by use of oral contraceptives [56]. It is therefore important in studies of candidate genes to consider not only the main effect of the gene but also its interactive role with other genes and/or environmental factors.

Traditional unmatched or matched case–control studies [1] may not be optimal for the study of genes. A potential problem is that estimates of genetic effect are subject to **confounding** when cases and controls differ in their ethnic backgrounds. This phenomenon, also known as population stratification **bias** [2, 23], can occur when both disease risk and genetic mutation frequencies vary among ethnic groups (*see* **Bias in Case–Control Studies**). If all or part of the disease-risk variation is due to factors other than the candidate gene (e.g. environmental exposures, a second gene), and those other risk factors also vary among ethnic groups, then a spurious association with the candidate gene may occur simply due to the indirect correlation of its distribution with the other risk factor(s). A classic example of such confounding is the reported association between the Gm locus and non insulin-dependent diabetes in American Indians that disappeared when the analysis was restricted to full-heritage Pima–Papago Indians [17]. To avoid the problem of population stratification bias, one can attempt to match cases to controls on ethnic background (*see* **Matched Analysis**; **Matching**). However, determination of ethnicity in a large-scale epidemiologic study is difficult, especially with the great diversity in cultural backgrounds that exists in the urban areas where studies are most likely to be conducted. It should be noted here that the degree to which population stratification can cause spurious findings for a candidate gene is the subject of current debate [57]. There has also been interest in using genetic markers to adjust for population stratification [32, 33, 37].

In this article we review family-based case–control designs that have as one of their primary features freedom from population stratification bias. These include the *case–sibling* and *case–parent* designs. We discuss methods for analyzing these designs that provide parameter estimates and hypothesis tests for genetic main effects and for gene–environment ($G \times E$) or gene–gene ($G \times G$) interactions (*see* **Multilocus (Gene × Gene Interaction)**). We provide comparisons among the case–sib, case–parent, and standard case–control designs of the sample size requirements for testing these hypotheses. Finally, we review alternative study designs that make use of family data.

Case–Sibling Design

In this design, one matches each case to one or more unaffected siblings. For complex diseases with variable age at onset (*see* **Age-of-onset Estimation**), controls should be sampled from the "risk set" consisting of those siblings who were disease-free at the age the case became affected (the *index age*) [12, 61,

65]. A sibling who is disease-free at the index age but is known to later develop the disease should not be eliminated from consideration as a control. In fact, general elimination of such siblings will result in a genetic effect estimate that is biased away from the null [26]. Data on known environmental risk factors *at the index age* should be collected for both cases and controls.

If only recent incident cases are included, then the age-matching requirement restricts control selection to older siblings. This could lead to confounding of the effects of environmental exposures that have secular trends or birth-order effects [12, 61, 65]. For example, older siblings may be less likely in general to take up smoking [47]. While this phenomenon will probably have little effect on an estimate of genetic risk, it should be considered as a potential source of **bias** in estimates of environmental and G × E interaction effects.

Some cases may not have any controls who have attained the index age, effectively excluding them from the analysis. This can be corrected in principle by constructing a **likelihood** involving the probability that younger controls remain disease-free up to the index age. However, inclusion of such controls can still pose problems if time-dependent covariates are involved. The validity of the case–sibling design also depends on the assumption that the ability to recruit cases and controls is not differentially related to their genotype, conditional on parental genotypes [61].

From a practical standpoint, the use of sibling controls may offer several nonstatistical advantages over population controls. The occurrence of disease in the case may make his or her relatives easier to recruit than an unrelated subject from the general population. In addition to reducing cost, this may improve data quality, since family members of controls may be more careful filling out risk-factor questionnaires. Researchers can also cross-validate questionnaire information related to family-specific variables that has been obtained from the case and sibling (*see* **Family History Validation**). The availability of family-based cancer registries can also make finding sibling controls much less expensive than finding controls from the general population. These resources can be used to identify case–sib pairs that are potentially most informative for testing genetic effects; for example, pairs that have a parent affected with the disease of interest [12]. However, care must be taken in using such a restricted sampling design, since bias in parameter estimates can result [16].

Analysis

Standard methods for the analysis of matched case–control data can be applied to the case–sibling design [1]. These include **McNemar's** and Mantel–Haenszel chi-squared tests and the associated estimates of the **odds ratio** [20, 48]. More generally, conditional **logistic regression** can be used to simultaneously model genetic and environmental main effects, as well as G × E interaction. If we let β_g, β_e, and β_{ge} denote parameters for the effect of a gene (G), environmental factor (E), and G × E interaction, respectively, then the conditional likelihood for a sample of N case–sibling sets has the form:

$$L(\beta_g, \beta_e, \beta_{ge}) = \prod_{i=1}^{N} \frac{\exp(\beta_g G_{i1} + \beta_e E_{i1} + \beta_{ge} G_{i1} E_{i1})}{\sum_{j \in M(i)} \exp(\beta_g G_{ij} + \beta_e E_{ij} + \beta_{ge} G_{ij} E_{ij})}. \quad (1)$$

The index 1 refers to the case, and the set $M(i)$ includes the case and all controls from family i. If controls are matched to the case's age and selected according to the principles of risk set sampling as described above, the quantities $R_g = \exp(\beta_g)$, $R_e = \exp(\beta_e)$, and $R_{ge} = \exp(\beta_{ge})$ can be interpreted as the corresponding hazard-rate ratios. If age of onset is not a factor (e.g. the disease is a birth defect), then these quantities represent odds ratios. The genetic **covariate** G is coded according to the assumed susceptibility of each genotype. For example, for a diallelic locus, $G(\mathrm{AA}) = 1$, $G(\mathrm{Aa}) = \delta$, and $G(\mathrm{aa}) = 0$. If the gene is assumed to be dominant, then $\delta = 1$; if recessive, then $\delta = 0$; if log-additive, then $\delta = 0.5$. Alternatively, G could be coded using two dummy variables to allow for separate estimation of the heterozygote (Aa) and homozygote (AA) effects relative to baseline (aa). Any of these coding schemes can be generalized to a locus with more than two alleles [38].

Score, Wald and **likelihood ratio tests** based on (1) can be formed as usual to test hypotheses about main or interactive effects [5]. If there are more than two subjects per family, however, then these

tests will only be valid if outcomes are conditionally independent within families. This will not be the case if the locus under study is simply a marker that is linked to the disease-predisposing locus. A number of different methods have recently been proposed to test for association in the presence of linkage [6, 15, 18, 19, 29, 35, 46].

Case–Parent Design

In this design, no actual controls are selected. Instead, genotypic data are obtained on the parents of the case, and the genotype transmitted to the case is compared with the three genotypes (*pseudo-siblings*) that were not transmitted to the case. For example, if the father's genotype is Aa, the mother's Aa, and the case's AA, then the pseudo-sibling genotypes are Aa (paternal A, maternal a), aA, and aa. The validity of this approach depends on the assumption that parental alleles are transmitted with equal and independent probability in the population. This assumption would fail if, for example, inheriting a certain genotype led to fetal death. Validity also depends on the assumptions that the ability to recruit a case is independent of the genotype given the parental genotypes, and that the genotyped "parents" are in fact the case's biological parents [61].

As with the case–sibling design, parents are more likely to be willing to participate than a population control, and the design will take advantage of information from a family-based registry. The disease status of the parents is not required in this design, nor is any information on their environmental exposures. In practice, the utility of this design is limited to disorders that occur at young enough ages that parents of the cases are likely to be alive.

Analysis

Conditional logistic regression for 1 : 3 matched sets provides a flexible framework for analyzing case–parent data [11, 38–41, 43]. The likelihood including both a genetic main effect and a G × E interaction has the form:

$$L(\beta_g, \beta_{ge}) = \prod_{i=1}^{N} \frac{\exp(\beta_g G_{i1} + \beta_{ge} G_{i1} E_{i1})}{\sum_{G_{ij}|G_{i\mathbf{P}}} \exp(\beta_g G_{ij} + \beta_{ge} G_{ij} E_{i1})}, \quad (2)$$

where again the index 1 refers to the case. The summation in the denominator of (2) is over the four possible genotypes that could be transmitted to an offspring given parental genotypes $G_{i\mathbf{P}}$. As above, the covariate G can be coded to reflect assumptions about the relationship among alleles [38]. The quantities $R_g = \exp(\beta_g)$ and $R_{ge} = \exp(\beta_{ge})$ can be interpreted as relative risks. Estimation of a main environmental effect is not possible since the three pseudo-siblings are perfectly matched to the case except for genotype. Valid estimation of the interaction effect β_{ge} requires that G and E are independently distributed in the population. However, even if independence holds, G×E interactions can be difficult to interpret in this design – absent knowledge of the main effect of exposure [61].

Under a log additive genetic model without G × E interaction, the **maximum likelihood** estimate $\hat{\beta}_g$ and the score test based on (2) are equivalent to the McNemar log odds ratio and chi-square test from the diallelic transmission-disequilibrium test (TDT) [49]. The TDT treats each parent as an independent matched pair of alleles, and compares the alleles transmitted from heterozygous parents of cases to those not transmitted. Schaid & Sommer [42] present the often-confusing history of this statistic and compare it with earlier variants, such as the *haplotype relative risk* [8, 36, 52].

Again, if more than one case per family is included in the analysis, then the usual tests based on (2) will only be valid if outcomes are conditionally independent within families. Multiple cases can be included if the conditioning event in (2) is restricted to genotypes that have the same identity-by-descent status as the cases' observed genotypes [24, 28, 45] or if an empirical estimate of the variance of the score (treating nuclear families as independent) is used [18, 22]. The Pedigree Disequilibrium Test [29, 30] also accommodates multiple-case families.

If genotype information for only one parent is available, then restricting the TDT to parent–child pairs where transmission is unambiguous (e.g. heterozygous parent; homozygous offspring) can induce bias [7]. Several techniques have been introduced in the last few years which remain valid when only one parent is available [34, 51, 54, 59, 63]. Another analysis approach to case–parent data based on Poisson regression instead of (2) can test for **parent-of-origin** effects as well as G × E interaction [55, 60, 62, 64].

Sample Size Considerations

We compare sample size requirements between the case–sibling and case–parent designs, and also compare each of these with the requirements for the standard matched **case–control** design (*see* **Matching**). In the latter design, controls are matched to cases on one or more factors (e.g. ethnicity, age), but are assumed to be genetically unrelated to cases. The conditional likelihood in (1) for the case–sibling design can also be used to estimate hazard-rate ratios or odds ratios for this design. For each design, we compute the minimum number of matched sets required to provide 80% power for testing genetic main effects, G × E interaction, and G × G interaction. Calculations were carried out using the program QUANTO, freely available at `http://hydra.usc.edu/GxE`. Details on the methods used for computing sample size are included in the program documentation, or in Gauderman [11].

In addition to the magnitudes of the relative risks R_g, R_e, and R_{ge}, sample size requirements depend on the prevalence of genetically susceptible individuals in the population [$\Pr(G = 1)$]. We consider four types of genes: rare dominant, common dominant, rare recessive, and common recessive (*see* **Mendel's Laws**; **Segregation Analysis, Classical**). A rare gene is defined as one for which 1% of the population is genetically susceptible, i.e. $\Pr(G = 1) = 0.01$, while for a common gene we assume $\Pr(G = 1) = 0.20$.

Table 1 shows the number of matched sets required to detect a genetic main effect for varying values of the true genetic **relative risk**, $R_g = \exp(\beta_g)$. For rare genes, sample size requirements are prohibitively large for all three designs, unless the genetic relative risk is large. For a more common gene, relative risks in the range of 1.5 to 2.0 can be detected with attainable sample sizes. In all situations, the case–sibling design requires the largest sample size, approximately 1.5 to 2 times greater than the case–control design. This reduced efficiency is due to **overmatching** of cases and their siblings on the genotype of interest [65]. The case–parent design, on the other hand, is nearly equivalent in efficiency to the case–control design for a dominant gene but more efficient for a recessive gene. The reason for the increased efficiency in the recessive situation is based on the fact that here the probability that a parent has the Aa genotype is relatively large. Examination of the likelihood in (2) shows that only parents with the Aa genotype are informative for testing genetic main effects. Thus, for a recessive gene, use of the case–parent design enriches the sample for informative parent-to-case transmissions.

To compare sample sizes needed to detect a G × E interaction, we assume $R_g = R_e = 1$, i.e. that risk is only increased in subjects who are both exposed and genetically susceptible. We also assume that the prevalence of exposure ($E = 1$) is 0.3, and

Table 1 Number (N) of matched sets required for 80% power to detect a genetic main effect with true relative risk R_g

Proportion susceptible [$\Pr(G = 1)$]	Mode of inheritance	R_g	Case–control, N	Case–sibling		Case–parent	
				N	(Ratio)[a]	N	(Ratio)[a]
0.01	Dominant	1.5	7914	15808	(0.50)	7905	(1.00)
		3.0	773	1544	(0.50)	772	(1.00)
		5.0	283	566	(0.50)	283	(1.00)
	Recessive	1.5	7914	11234	(0.70)	5449	(1.45)
		3.0	773	1097	(0.70)	505	(1.53)
		5.0	283	402	(0.70)	178	(1.59)
0.20	Dominant	1.5	536	1040	(0.51)	521	(1.03)
		3.0	66	127	(0.52)	64	(1.03)
		5.0	30	59	(0.51)	30	(1.00)
	Recessive	1.5	536	901	(0.59)	443	(1.21)
		3.0	66	110	(0.60)	53	(1.25)
		5.0	30	51	(0.59)	24	(1.25)

[a]Compared with the case–control design; ratios above (below) 1.0 indicate greater (lesser) efficiency.
Assumptions: 0.05 significance level and two-sided alternative hypothesis.

Table 2 Number (N) of matched sets required for 80% power to detect a G × E interaction with true relative risk ratio R_{ge}

Proportion susceptible [Pr($G = 1$)]	Mode of inheritance	R_{ge}	Case–control, N	Case–sibling		Case–parent	
				N	(Ratio)[a]	N	(Ratio)[a]
0.01	Dominant	2.0	12699	11901	(1.07)	12669	(1.00)
		3.0	4585	4072	(1.13)	4573	(1.00)
		5.0	1949	1584	(1.23)	1945	(1.00)
	Recessive	2.0	12700	12215	(1.04)	8372	(1.52)
		3.0	4586	4267	(1.07)	2881	(1.59)
		5.0	1949	1715	(1.14)	1151	(1.69)
0.20	Dominant	2.0	849	821	(1.03)	809	(1.05)
		3.0	326	308	(1.06)	309	(1.06)
		5.0	152	141	(1.08)	145	(1.05)
	Recessive	2.0	851	828	(1.03)	669	(1.27)
		3.0	327	312	(1.05)	247	(1.32)
		5.0	153	143	(1.07)	112	(1.37)

[a]Compared with the case–control design; ratios above (below) 1.0 indicate greater (lesser) efficiency.
Assumptions: Exposure prevalence 0.30, $R_g = 1$, $R_e = 1$, 0.05 significance level and two-sided alternative hypothesis.

Table 3 Number (N) of matched sets for 80% power to detect a gene–gene interaction with magnitude $R_{gh} = 3.0$

Proportion susceptible		Mode of inheritance		Case–control, N	Case–sibling		Case–parent	
Pr($G = 1$)	Pr($H = 1$)	G	H		N	(Ratio)[a]	N	(Ratio)[a]
0.01	0.20	Dom	Dom	5617	7001	(0.80)	3157	(1.78)
		Dom	Rec	5617	6613	(0.85)	2906	(1.93)
		Rec	Dom	5617	6355	(0.88)	2685	(2.09)
		Rec	Rec	5617	6163	(0.91)	2568	(2.18)
0.20	0.20	Dom	Dom	400	498	(0.80)	231	(1.73)
		Dom	Rec	400	475	(0.84)	213	(1.88)
		Rec	Rec	400	458	(0.87)	201	(1.99)

[a]Compared with the case–control design; ratios above (below) 1.0 indicate greater (lesser) efficiency.
Assumptions: $R_g = 1$, $R_h = 1$, 0.05 significance level and two-sided alternative hypothesis.

that there is no **correlation** in exposure between siblings. Table 2 shows the required sample size to detect a G × E interaction for varying the magnitude of R_{ge}. Again, for a rare gene and moderate G × E effect, sample sizes are prohibitive. However, the case–sibling and case–parent designs are always more powerful than the case–control design. In this context, siblings' concordance on genotype makes them more informative for testing for G × E interaction (see Table 1 in [12]). However, this increase in efficiency disappears as the at-risk allele becomes more common, or with increasing correlation in exposure between siblings.

The likelihoods in (1) and (2) can be modified to test for the effect of a second measured locus simply by replacing E by a second genetic effect H, which can be coded according to assumptions about inheritance at the second locus. If, for example, both loci are diallelic and recessive, the interaction term G × H will only be 1 if the case is a homozygote carrier at both loci. Table 3 shows the required sample sizes to detect a three-fold gene–gene interaction effect ($R_{gh} = 3$) when H is a common gene and $R_g = R_h = 1$. The case–sibling design is least efficient. The case–parent design, on the other hand, provides large efficiency gains over the case–control design, requiring approximately half the matched sets. Again, this is because the case–parent design enriches the sample for informative parent-to-case transmissions.

Discussion

Family based case–control studies offer an attractive alternative to population-based case–control designs using unrelated controls. Their primary advantage is that they overcome the problem of population stratification that can lead to spurious associations. For the case–sibling design, this protection from bias comes at the price of reduced statistical efficiency for some tests due to overmatching on genotype. For the case–parent design, one can generally expect increased efficiency relative to the case–control design, particularly for tests of gene–gene interaction. One should keep in mind that the case–parent design will probably be more expensive per matched set, as DNA needs to be collected and processed for three subjects rather than for two subjects in the other designs. Finally, family designs offer nonstatistical advantages, such as improved cooperation and reduced cost. These must be weighed against any losses in sample size from cases who do not have a suitable family control and the potential selection bias if such losses are nondifferential (*see* **Bias in Case–Control Studies**; **Validity and Generalizability in Epidemiologic Studies**).

Statistical methodology for the analysis of family-based association studies remains an active area of research. The basic methods described above do not utilize all aspects of family data. For example, if phenotype and genotype information are available on parents *and* siblings of a case, the analytic approaches described above will discard information. Several new approaches make use of cases, parents, and siblings in a unified framework – and have also been extended to utilize data from extended pedigrees [21, 22, 25, 27, 34, 54, 63]. Many of these approaches also allow analysis of a continuous phenotype, in addition to binary and censored-age-of-onset traits. Also proposed are methods that will make use of relatives of cases whose phenotypes are known, but who have not been genotyped. The "**kin–cohort**" and "genotyped proband" designs are examples of these approaches, both of which have been used to estimate age-specific penetrance of BRCA1 and BRCA2 among Ashkenazi Jews [9, 10, 31, 50, 58].

Finally, we mention extensions of family designs that use haplotypes from multiple tightly linked markers instead of analyzing individual loci separately (*see* **Haplotype Analysis**). This approach is becoming possible with the increasing density of available markers, and should improve our ability to localize a disease-causing gene. If the haplotype phase could be unambiguously determined for all subjects, then a set of markers could be analyzed as one highly polymorphic locus. Unfortunately, this phase often cannot be uniquely determined. New methods to handle this situation have been proposed [3, 4, 44, 66], and this will continue to be an active area of future research. Furthermore, the number of possible haplotypes can create **multiple comparisons** problems. Without a plausible biologic model for grouping haplotypes together, one approach to this problem is first to use statistical tools to cluster haplotypes and then estimate relative risk parameters for each cluster [44, 53] (*see* **Cladistic Analysis**). In general, the utility of family-based study designs relative to other alternatives should be re-evaluated in the context of haplotype-oriented studies.

Acknowledgments

This work was supported in part by NIEH grant ES10421 and NCI grant CA52862

References

[1] Breslow, N.E. & Day, N.E. (1980). Statistical methods in cancer research: I. The analysis of case–control studies, in W. Davis, ed., *Statistical Methods in Cancer Research*, Vol. 32. IARC Scientific Publications, Lyon.

[2] Caparaso, N., Rothman, N. & Wacholder, W. (1999). Case-control studies of common alleles and environmental factors, *Monograph of the National Cancer Institute* **26**, 25–30.

[3] Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission, *American Journal of Human Genetics* **65**, 1170–1177.

[4] Clayton, D. & Jones, H. (1999). Transmission/disequilibrium tests for extended marker haplotypes, *American Journal of Human Genetics* **65**, 1161–1169.

[5] Cox, D. & Hinkley, D. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[6] Curtis, D. (1997). Use of siblings as controls in case–control association studies, *Annals of Human Genetics* **61**, 319–333.

[7] Curtis, D. & Sham, P. (1995). A note on the application of the transmission disequilibrium test when a parent is missing, *American Journal of Human Genetics* **56**, 811–812.

[8] Falk, C. & Rubenstein, P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations, *Annals of Human Genetics* **51**, 227–233.

[9] Gail, M., Pee, D. & Carroll, R. (1999). Kin–cohort designs for gene characterization, *Monograph of the National Cancer Institute* **26**, 55–60.

[10] Gail, M., Pee, D., Benichou, J. & Carroll, R. (1999). Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case–control, and genotype-proband designs, *Genetic Epidemiology* **16**, 15–39.

[11] Gauderman, W. (2002). Sample size requirements for matched case–control studies of gene-environment interaction, *Statistics in Medicine*, **21**.

[12] Gauderman, W., Witte, J. & Thomas, D. (1999). Family-based association studies, *Monograph of the National Cancer Institute* **26**, 31–37.

[13] Gilliland, F., McConnell, R., Peters, J. & Gong, H. Jr (1999). A theoretical basis for investigating ambient air pollution and children's respiratory health, *Environmental Health Perspective* **107**, 403–407.

[14] Goldstein, A. & Andrieu, N. (1999). Detection of interaction involving identified genes: available study designs, *Monograph of the National Cancer Institute* **26**, 49–54.

[15] Horvath, S. & Laird, N. (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data, *American Journal of Human Genetics* **63**, 1886–1897.

[16] Hsu, L., Zhao, L. & Aragaki, C. (1998). A note on a conditional-likelihood approach for family-based association studies of candidate genes, *Human Heredity* **50**, 194–200.

[17] Knowler, W., Williams, R., Pettitt, D. & Steinberg, A. (1988). Gm3,5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture, *American Journal of Human Genetics* **43**, 520–526.

[18] Kraft, P. (2001). A robust score test for linkage disequilibrium in general pedigrees, *Genetic Epidemiology* **21**, S447–S452.

[19] Kraft, P. & Siegmund, K. (2000). Testing linkage disequilibrium in sibships using conditional logistic regression with robust variance estimators, *Genetic Epidemiology* **19**, 257.

[20] Laird, N., Blacker, D. & Wilcox, M. (1998). The sib transmission/disequilibrium test is a Mantel-Haenszel test, *American Journal of Human Genetics* **63**, 1915.

[21] Laird, N., Horvath, S. & Xu, X. (2000). Implementing a unified approach to family-based tests of association, *Genetic Epidemiology* **19**, Supplement 1, S36–S42.

[22] Lake, S., Blacker, D. & Laird, N. (2000). Family-based tests of association in the presence of linkage, *American Journal of Human Genetics* **67**, 1515–1525.

[23] Lander, E. & Schork, N. (1994). Genetic dissection of complex traits, *Science* **265**, 2037–2048.

[24] Lazzeroni, L. & Lange, K. (1998). A conditional inference framework for extending the transmission/disequilibruim test, *Human Heredity* **48**, 67–81.

[25] Li, H. & Fan, J. (2000). A general test of association for complex diseases with variable age at onset, *Genetic Epidemiology* **19**, S43–S49.

[26] Lubin, J. & Gail, M. (1984). Biased selection of controls for case–control analysis of cohort studies, *Biometrics* **40**, 63–75.

[27] Lunetta, K., Faraone, S., Biederman, J. & Laird, N. (2000). Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions, *American Journal of Human Genetics* **66**, 605–614.

[28] Martin, E., Kaplan, N. & Weir, B. (1997). Tests for linkage and association in nuclear families, *American Journal of Human Genetics* **61**, 439–448.

[29] Martin, E., Monks, S. & Kaplan, N. (1999). A weighted sibship disequilibrium test for linkage and association in discordant sibships, *American Journal of Human Genetics* **65**, A434.

[30] Martin, E., Monks, S., Warren, L. & Kaplan, N. (2000). A test for linkage and association in general pedigrees: The Pedigree Disequilibrium Test, *American Journal of Human Genetics* **67**, 146–154.

[31] Moore, D., Chatterjee, N., Pee, D. & Gail, M. (2001). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin–cohort study, *Genetic Epidemiology* **20**, 210–227.

[32] Pritchard, J., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data, *Genetics* **155**, 945–959.

[33] Pritchard, J., Stephens, M., Rosenberg, N. & Donnelly, P. (2000). Association mapping in structured populations, *American Journal of Human Genetics* **67**, 170–181.

[34] Rabinowitz, D. & Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information, *Human Heredity* **50**, 211–223.

[35] Rieger, R., Kaplan, N. & Weinberg, C. (2001). Efficient use of siblings in testing for linkage and association, *Genetic Epidemiology* **20**, 175–191.

[36] Rubinstein, P., Walker, M. & Carpenter, C. (1981). Genetics of HLA disease association: the use of the haplotype relative risk (HRR) and the "Haplo-Delta" (DH) estimates in juvenile diabetes from three racial groups, *Human Immunology* **3**, 384.

[37] Satten, G., Flanders, W. & Yang, Q. (2001). Accounting for unmeasured population substructure in case–control studies of genetic association using a novel latent-class model, *American Journal of Human Genetics* **68**, 466–477.

[38] Schaid, D. (1996). General score tests for associations of genetic markers with disease using cases and their parents, *Genetic Epidemiology* **13**, 423–449.

[39] Schaid, D. (1999). Case–parents design for gene-environment interaction, *Genetic Epidemiology* **16**, 261–273.

[40] Schaid, D. (1999). Likelihoods and TDT for the case-parents design, *Genetic Epidemiology* **16**, 250–260.

[41] Schaid, D. & Sommer, S. (1993). Genotype relative risks: methods for design and analysis of candidate-gene association studies, *American Journal of Human Genetics* **53**, 1114–1126.

[42] Schaid, D. & Sommer, S. (1994). Comparison of statistics for candidate-gene association studies, *American Journal of Human Genetics* **55**, 402–409.

[43] Self, S., Longton, G., Kopecky, K. & Liang, K. (1991). On estimating HLA/disease association with application to a study of aplastic anemia, *Biometrics* **47**, 53–61.

[44] Seltman, H., Roeder, K. & Devlin, B. (2001). Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes, *American Journal of Human Genetics* **68**, 1250–1263.

[45] Siegmund, K. & Gauderman, W. (2001). Association tests in nuclear families, *Human Heredity*, in press.

[46] Siegmund, K., Langholz, B., Kraft, P. & Thomas, D. (2000). Testing linkage disequilibrium in sibships, *American Journal of Human Genetics* **67**, 244–248.

[47] Simon, W. (1973). Ordinal position of birth in the family constellation and adult smoking behavior, *Journal of Social Psychology* **90**, 157–158.

[48] Spielman, R. & Ewens, W. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test, *American Journal of Human Genetics* **62**, 450–458.

[49] Spielman, R., McGinnis, R. & Ewens, W. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics* **52**, 506–516.

[50] Struewing, J., Hartge, P., Wacholder, S., Baker, S., Berlin, M., McAdams, M., Timmerman, M., et al. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews, *New England Journal of Medicine* **336**, 1401–1408.

[51] Sun, F., Flanders, W., Yang, Q. & Khoury, M. (1999). Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT, *American Journal of Epidemiology* **150**, 97–104.

[52] Terwilliger, J. & Ott, J. (1992). A haplotype based haplotype relative risk approach to detecting allelic associations, *Human Heredity* **42**, 337–346.

[53] Thomas, D., Morrison, J. & Clayton, D. (2001). Bayes estimates of haplotype effects, *Genetic Epidemiology* **21**, S712–S717.

[54] Tu, I., Balise, R. & Whittemore, A. (2000). Detection of disease genes by use of family data. II. Application to nuclear families, *American Journal of Human Genetics* **66**, 1341–1350.

[55] Umbach, D. & Weinberg, C. (2000). The use of case–parent triads to study the joint effects of genotype and exposure, *American Journal of Human Genetics* **66**, 251–261.

[56] Ursin, G., Henderson, B., Haile, R., Zhou, N., Diep, A. & Bernstein, L. (1997). Is oral contraceptive use more common in women with BRCA1/BRCA2 mutations than in other women with breast cancer? *Cancer Research* **57**, 3678–3681.

[57] Wacholder, S., Rothman, N. & Caporaso, N. (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias, *Journal of the National Cancer Institute* **92**, 1151–1158.

[58] Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. & Tucker, M. (1998). The kin cohort study for estimating penetrance, *American Journal of Epidemiology* **148**, 623–630.

[59] Weinberg, C. (1999). Allowing for missing parents in genetic studies of case–parent triads, *American Journal of Human Genetics* **64**, 1186–1193.

[60] Weinberg, C. (1999). Methods for detection of parent-of-origin effects in genetic studies of case–parent triads, *American Journal of Human Genetics* **65**, 229–235.

[61] Weinberg, C. & Umbach, D. (2000). Choosing a retrospective design to assess joint genetic and environmental contributions to risk, *American Journal of Epidemiology* **152**, 197–203.

[62] Weinberg, C., Wilcox, A. & Lie, R. (1998). A log linear approach to case–parent data: assessing effects of disease genes that act directly or through maternal effects, and may be subject to parental imprinting, *American Journal of Human Genetics* **62**, 969–978.

[63] Whittemore, A. & Tu, I. (2000). Detection of disease genes by use of family data. I. Likelihood-based theory, *American Journal of Human Genetics* **66**, 1328–1340.

[64] Wilcox, A., Weinberg, C. & Lie, R. (1998). Distinguishing the effects of maternal and offspring genes through studies of "case parent triads", *American Journal of Epidemiology* **148**, 893–901.

[65] Witte, J.S., Gauderman, W.J. & Thomas, D.C. (1999). Asymptotic bias and efficiency in case–control studies of candidate genes and gene–environment interactions: basic family designs, *American Journal of Epidemiology* **148**, 693–705.

[66] Zhao, H., Zhang, S., Merikangas, K., Trixler, M., Wildenauer, D., Sun, F. & Kidd, K. (2000). Transmission/disequilibrium tests using tightly linked markers, *American Journal of Human Genetics* **67**, 936–946.

(*See also* **Bias in Case–Control Studies**; **Disease-marker Association**; **Linkage Disequilibrium**)

W. JAMES GAUDERMAN & PETER KRAFT

Fan Plot

Introduction

The fan plot is a **graphical** procedure for determining the effect of one or more observations on the **transformation** parameter λ in the Box and Cox family of power transformations of the **response** in **regression**. Such transformations, for example, from $\boldsymbol{y}$ to $\log \boldsymbol{y}$, are often important for ensuring that the assumptions behind **least squares** are satisfied and that therefore, efficient use is made of data (*see* **Power Transformations**). The fan plot is based on a **forward search** through the data to fit subsets of increasing numbers of observations, with any **outliers** being included toward the end of the search. The plot monitors the behavior of the approximate score test for five different transformations and reveals whether the evidence for a transformation depends on a few observations or is, preferably, spread throughout the data.

Interest is in transformation of the response $\boldsymbol{y}$ in the **multiple regression** model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

$\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters and it is assumed that the additive errors of observation ε are independently distributed with constant variance σ^2. Also in (1) $\boldsymbol{X}$ is the $n \times p$ **matrix** of carriers, that is, of **explanatory variables** and perhaps functions of them, such as quadratics and interaction terms. To obtain the approximate score test we add a "constructed variable" (*see* **Residuals**) to the regression model and obtain the augmented model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}\gamma + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{w}$ is $n \times 1$ and γ is a scalar parameter. The approximate score test is the **Student *t*-test** t_γ for testing that γ in (2) equals zero. The constructed variable for the transformation is derived in the next section. Testing that $\gamma = 0$ is testing that there is no evidence for any transformation of the response.

A Score Test for Transformations

The analysis of the data on mandible length in the article on **residuals** shows appreciable evidence not only of the normality of the residuals (*see* **Normality, Tests of**, Figure 3) but also of increasing variance with fitted value, Figure 1. Often, normality and constant variance can be achieved by fitting the regression model not to $\boldsymbol{y}$ but to a function of $\boldsymbol{y}$; Figure 1 of the article on **diagnostics** shows the beneficial effect of the transformation to $\log(\boldsymbol{y})$ combined with quadratic regression (*see* **Polynomial Regression**) on the residuals from the mandible length data. The appropriate transformation frequently, but, as will be seen later, not always, also leads to a simple linear model, without quadratic or **interaction** terms.

The logarithmic transformation is one special case of the normalized power transformation [4]

$$z(\lambda) = \begin{cases} \dfrac{\boldsymbol{y}^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y} \log \boldsymbol{y} & \lambda = 0, \end{cases} \tag{3}$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i / n)$. For inference about the transformation parameter λ, Box and Cox suggest **likelihood ratio tests**. A computationally simpler alternative test is the approximate score statistic (*see* **Likelihood**) derived by Taylor series expansion of (3) as

$$\begin{aligned} z(\lambda) &\doteq z(\lambda_0) + (\lambda - \lambda_0) \left. \frac{\partial z(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0} \\ &= z(\lambda_0) + (\lambda - \lambda_0)\boldsymbol{w}(\lambda_0). \end{aligned} \tag{4}$$

In (4), $\boldsymbol{w}(\lambda_0)$ is the "constructed variable" for the transformation and can be treated as is the extra-explanatory variable in (2). To test the transformation $\lambda = \lambda_0$ the response $\boldsymbol{y}$ is transformed to $\boldsymbol{z}(\lambda_0)$ in (3). The approximate score statistic, $T_p(\lambda_0)$, is then the t statistic t_γ for regression of the transformed response on $\boldsymbol{w}(\lambda_0)$ in (2). Details of the constructed variables are in the article on **residuals**.

The Fan Plot

In the **forward search**, the p parameters of the regression model (1) are estimated by least squares applied to a carefully chosen subset of m observations. We start the search with m small, usually p or perhaps $p + 1$, and randomly select 1000 subsamples. The initial subset provides the least median of squares estimator, that is it minimizes the median squared

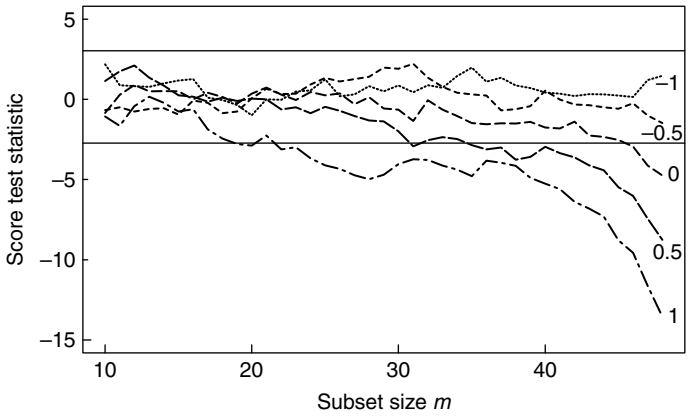

Figure 1 Poisson data: fan plot–forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost: both $\lambda = -1$ and $\lambda = -0.5$ are acceptable. There is no evidence of any outliers or influential observations

residual [5]. We then order the residuals and augment the subset.

When m observations are used in fitting, the optimum subset yields n residuals $\boldsymbol{e}(m^*)$. We order the squared residuals $\boldsymbol{e}^2(m^*)$ and take the observations corresponding to the $m + 1$ smallest as the new subset. Usually, this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave. Owing to the form of the search, outliers, if any, tend to enter as m approaches n.

We combine calculation of the test statistic $T_p(\lambda_0)$ with the forward search. Since observations that are outlying on one scale may not be outlying for a different transformation, we conduct several searches for different values of λ_0. In most applications, including the examples here, we use five searches for the values $\lambda = -1$, -0.5, 0, 0.5, and 1. If there are outliers for a particular λ, they will enter the search last and influence the value of the test statistic.

As a first example, we use the Poisson Data from Box and Cox [4], partly analyzed in the article on **residuals**. These data are well behaved: there are no outliers or influential observations that cannot be reconciled with the greater part of the data by a suitable transformation. Our fan plot clearly indicates the reciprocal transformation. We then consider a series of modifications of the data in which an increasing number of outliers is introduced. The fan plot reveals the structure in all instances.

The data are the times to death of animals in a 3×4 **factorial experiment** with four observations at each factor combination. All our analyses use an additive model, that is, without interactions, so that $p = 6$, the model used by Box and Cox when finding the reciprocal transformation. The implication is that the model should be additive in death rate, not in time to death.

The fan plot of the values of the approximate score statistic $T_p(\lambda)$ for the five searches as the subset size m increases is given in Figure 1 and shows that the reciprocal transformation is acceptable as is the inverse square root transformation ($\lambda = -0.5$). The horizontal lines are at ± 2.58, corresponding to 1% significance, assuming the statistics have a **standard normal** distribution. The results of Atkinson and Riani [3] show that this is a good working approximation.

Initially, for small subset sizes, there is no evidence against any transformation. During the whole forward search, there is never any evidence against either $\lambda = -1$ or $\lambda = -0.5$ (for all the data $\hat{\lambda} = -0.75$). The log transformation is also acceptable until the last four observations are included by the forward search. These are some of the largest observations, which will be informative about the need to transform. Evidence that some transformation is needed is spread throughout the data, less than half of the observations being sufficient to reject the hypothesis that $\lambda = 1$. There are no jumps in this curve, just an increase in evidence against $\lambda = 1$ as each observation is introduced into the subset. The relative smoothness of the curves reflects the lack of outliers and exceptionally influential cases and

the general shape of the plot justifies the name of "fan plot".

For the introduction of a single outlier into the Poisson data, we follow Andrews [1] and change observation 8, one of the readings for Poison II, group A, from 0.23 to 0.13. This is not one of the larger observations, so the change does not create an outlier in the scale of the original data. The effect on the estimated transformation of all the data is, however, to replace the reciprocal with the logarithmic transformation: $\hat{\lambda} = -0.15$. And, indeed, the fan plot of the score statistics from the forward searches in Figure 2 shows that, at the end of the forward search, the final acceptable value of λ is 0, with -0.5 on the boundary of the acceptance region.

Figure 2 clearly reveals the altered observation and the differing effect it has on the five searches. Initially, the curves are the same as those of Figure 1. But for $\lambda = 1$, there is a jump due to the introduction of the outlier when $m = 41$, which provides evidence for higher values of λ. For other values of λ, the outlier is included further on in the search. When $\lambda = 0.5$, the outlier comes in at $m = 46$, giving a jump to the score statistic in favor of this value of λ. For the other values of λ, the outlier is the last value to be included. Inclusion of the outlier has the largest effect on the inverse transformation. It is clear from the figure how this one observation is causing an appreciable change in the evidence for a transformation.

We now further modify the Poisson data; in addition to the previous modification, we also change observation 38 (Poison I, group D) from 0.71 to 0.14. This creates an example of masking, in which one outlier hides the effect of another, so that neither is evident when using the methods for the deletion of single observations described in the article on **diagnostics**.

The effect of the two outliers is clearly seen in the fan plot, Figure 3. Here, only $\lambda = 0$ is acceptable at the end of the search. The plot also reveals the differing effect the two altered observations have on the five searches. Initially, the curves are again similar to those of the original data shown in Figure 1. The difference is greatest for $\lambda = -1$ where addition of the two outliers at the end of the search causes the statistic to jump from an acceptable 1.08 to 10.11. The effect is similar, although smaller, for $\lambda = -0.5$. It is most interesting, however, for the log transformation. Toward the end of the search this statistic is trending downwards, below the acceptable region. But addition of the last two observations causes a jump in the value of the statistic to a nonsignificant value. The incorrect log transformation is now acceptable.

For these three values of λ, the outliers are the last two observations to be included in the search. They were created by introducing values that are too near zero when compared with the model fitted to the rest of the data. For the log transformation, and more so for the reciprocal, such values become extreme and so have an appreciable effect on the fitted model. For the other two values of λ, the outliers are included earlier in the search. The effect is most clearly seen when $\lambda = 1$; the outliers come in at $m = 40$ and 46, giving upward jumps to the score statistic in favor of

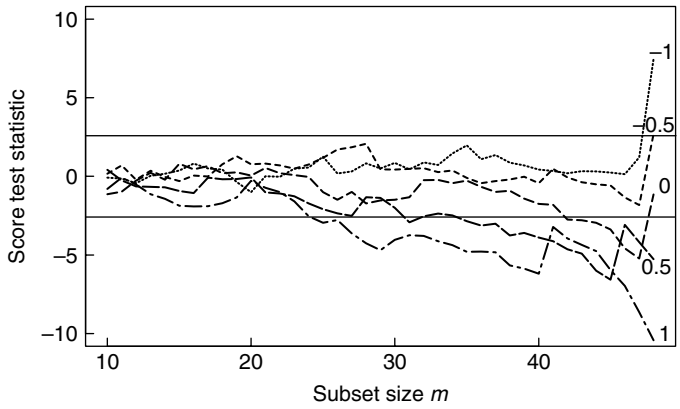

Figure 2 Modified Poisson data: fan plot–forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost: the effect of the outlier is evident in making $\lambda = 0$ appear acceptable at the end of the search

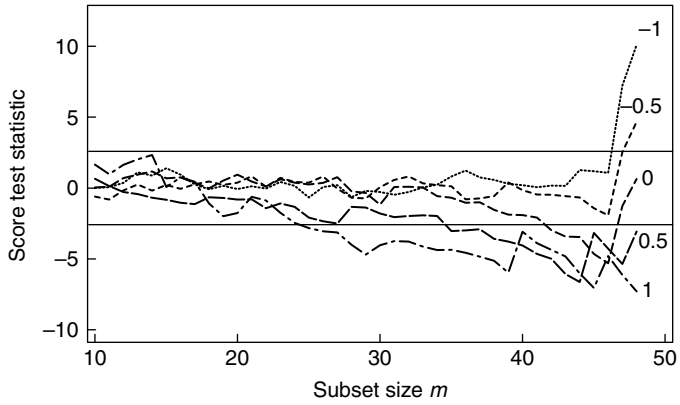

Figure 3 Doubly modified Poisson data: fan plot–forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost; the effect of the two outliers is clear

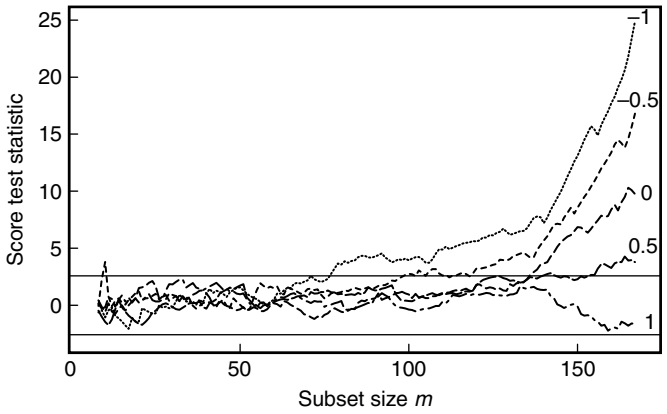

Figure 4 Mandible length data: fan plot–forward plot of $T_p(\lambda)$ for the five transformations of the data when the regression is on age

this value of λ. For the remaining value of 0.5, one of the outliers is the last value to be included.

These three plots exhibit the main features of the fan plot. Further analyses of the examples and comparison with other procedures are in Atkinson and Riani [2, Sections 4.4 and 4.7]. One conclusion is that alternative diagnostic procedures, such as the constructed variable plot in Figure 5 of **residuals**, can fail in the presence of masking and multiple outliers.

Mandible Length Data

The preceding examples calibrate the properties of the fan plot. We now use it to analyze transformations of the mandible length data.

The plot of the residuals of the untransformed data after regression on **gestational age**, for example, Figure 3 of **residuals**, showed three negative outliers as well as many smaller residuals lying outside the simulation envelope. In contrast, the residuals after regression of $\log \boldsymbol{y}$ on a quadratic in age, Figure 1 of **diagnostics**, are much more nearly normal. Is the evidence for this transformation largely dependent on the outlying observations? How is it affected by the linear model?

We start with just simple regression. Figure 4 is a fan plot for the five transformations of the data when the regression is on age. There is no evidence for a transformation – all values except $\lambda = 1$ are rejected by the end of the search. The

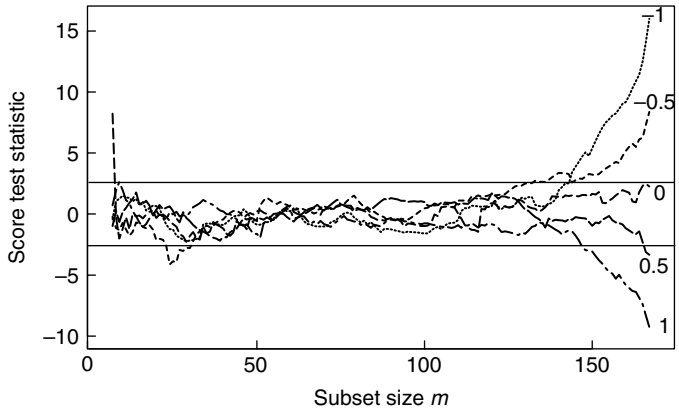

Figure 5 Mandible length data: fan plot–forward plot of $T_p(\lambda)$ for the regression of log $\boldsymbol{y}$ on a quadratic in age

statistic for this value remains within the bounds of ± 2.58 throughout the search. Although the values are toward the lower boundary at the end of the search, there is no obvious evidence of the effect of the three outlying observations, of the kind seen in Figure 3. Such jumps in the curve of the statistic are most in evidence for the reciprocal transformation $\lambda = -1$, where the observations giving negative residuals on the untransformed scale are even more extreme after transformation.

Although there is no evidence for transformation when regression is on age, we know from Table 1 of **diagnostics** that the quadratic term in this regression is significant. The final plot, Figure 5, is therefore the fan plot for the regression of log $\boldsymbol{y}$ on a quadratic in age. It shows that, for this more complicated model with an extra term, the log transformation is the only one that is acceptable. Although the last three observations to enter the search increase the value of the statistic, it does not change dramatically. There are no jumps in the other curves of the kind visible for $\lambda = -1$ in Figure 4.

The general conclusion is that the logarithmic transformation with a quadratic model is to be preferred to simple regression and no transformation. As the forward plots of t statistics for regression coefficients in Figure 4 of the article on the **forward search** show, this conclusion is supported by all the data and is in agreement with the $Q-Q$ plots of residuals mentioned above. An interesting feature of the analysis is that transformation has strengthened the evidence for a more complicated regression model. Often transformations result in a simpler model, but here there is a conflict between the linearity of the plot of $\boldsymbol{y}$ against $\boldsymbol{x}$ and the increasing variance with $\boldsymbol{y}$ evident in Figure 2 of the article on **Goodness of Fit**. This conflict was a reason for the fractional polynomial models used by Royston and Altman [6]. An alternative analysis is to keep the simple linear model, but to transform both sides of the model to obtain errors with constant variance (*see* **Power Transformations**). The forward search for this transformation is illustrated in [2, Section 4.12].

References

[1] Andrews, D.F. (1971). A note on the selection of data transformations, *Biometrika* **58**, 249–254.

[2] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

[3] Atkinson, A.C. & Riani, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression, *Chemometrics and Intelligent Laboratory Systems* **60**, 87–100.

[4] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* **26**, 211–246.

[5] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association* **79**, 871–880.

[6] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

(*See also* **Model Checking**; **Model, Choice of**)

A.C. ATKINSON & MARCO RIANI

Farr, William

Born: November 30, 1807, in Shropshire, UK.
Died: April 14, 1883, UK.

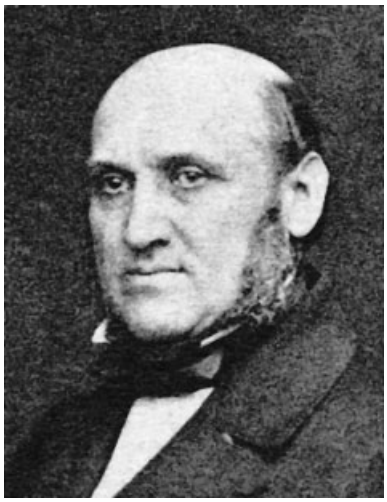

Reproduced by permission of the Royal Statistical Society

William Farr must be counted as the founder of epidemiology in its modern form [4]. He studied many aspects of the distribution and determinants of health disorders in populations, and the application of the studies to the prevention and control of disease. His publications have been grouped under six headings: population, marriages, births, deaths, **life tables**, and miscellaneous [4]. He developed an interest in accurate mortality statistics by age and sex and changed the British system of death registration into an instrument for measuring the sanitary condition of the country. His studies on mortality differences among different occupations helped in understanding industrial hazards [2].

Farr's work had a great impact and influence on many pioneers in the health field. Among them are Edwin Chadwick and John Simon, who were the driving forces of the new movement for sanitary reform, and the mathematicians **Karl Pearson**, Ronald Ross, and **John Brownlee**, who improved on his method of description. Farr anticipated the germ theory of disease by almost 20 years before John Snow's classic studies of cholera, and he had an extensive and productive relationship with **Florence Nightingale** in her campaign for improving social health and hygiene [1].

Coming from a very poor family, William Farr was apprenticed at age eight to Joseph Pryce of Dorrington, near Shrewsbury, who enabled him to pursue his elementary studies. Farr was a self-taught mathematician and an accomplished linguist who learned Latin, French, Italian, and also studied Hebrew. Recognizing Farr's interest in learning and his insatiable desire for knowledge, Pryce encouraged him to enter a profession. In May 1826 he began his medical education. In November 1828, a legacy of £500, inherited on Pryce's death, made his trip to Paris possible. The Paris school was then the leader in clinical medicine, hygiene, and medical statistics, and there he learned the statistical methods applied by **Pierre C. A. Louis**. In March, 1832, he passed the examination of the Society of Apothecaries – the only examination he ever took [3]. A year later, he married Miss Langford and tried to set up a practice at 8 Grafton Street, Fitzroy Square, London. After the death of his wife, he married Miss Whittall in 1841 and they had eight children, of whom four daughters and one son survived.

In 1830, several of his articles on hygiene and **vital statistics** appeared in the *Lancet*, and in 1835 he served as a medical editor and began to study vital statistics. He also edited his own journal, *British Annals of Medicine, Pharmacy, Vital Statistics, and General Science*, which lasted only from January to August 1837. He wrote six major articles for this journal; four on vital statistics, and two on medical reform. He clearly emphasized that medicine was both a science and a social institution [1]. In his many editorials in the *British Annals of Medicine* he insisted on the restructuring of the medical profession and the potential value of medicine to society, especially in matters of prevention and hygiene.

Farr accounted for the importance of demographic data in determining mortality patterns. He constantly strove to improve the quality and the extent of data collected and devised standard classifications for diseases and for causes of deaths [2]. He examined secular changes in mortality, specific causes of death, deaths by area, season of year, residence, occupation, and by marital status. He observed the association of mortality with the density of the population, including other factors such as water and air pollution. He also made the **census** serve not only as a denominator, but as a vehicle for national surveys of the **prevalence**

of blindness and deafness. He developed descriptive statistics for institutions, thus illuminating the characteristics and the ill-effects of hospitals, workhouses, asylums for the insane, and prisons. Furthermore, Farr envisaged a national system of morbidity statistics that embraced hospitals and other services.

Farr's medical training influenced his statistical career. When he began to consider social problems and public health, his medical perspective remained and tempered both his statistical approach and his ideas of reform. He worked on data analysis, on the procedures of tabulation, comparison, and **inference**. He measured the death-toll by defining standard rates for comparison, invented the standardized mortality rate (*see* **Standardization Methods**), and compared areas and occupations by means of summary statistics unconfounded by demographic differences such as sex and age (*see* **Occupational Mortality**). He used **life tables** to estimate the effects of prevention on the expectation of life. He was the first to describe **epidemic curves** and to use models for **prediction**. He defined the relationship between death, health, and energy of body and mind, between death, birth, and marriage, and he also connected the effect of literacy on the **quality of life**.

In the years 1838–1839, he derived the general "law of epidemics" from the 15-month epidemic of smallpox. In 1839, under the influence of Sir James Clark, he was appointed as the Compiler of Abstracts in the General Register Office, then as superintendent of the Statistical Department. He wrote in his first letter to the Registrar General that "diseases are more easily prevented than cured, and the first step to their prevention is the discovery of their exciting causes" [4]. For the next 40 years, he devoted himself to the task of creating and developing a national system of vital statistics, that was used not only in England but in all the civilized countries. He blamed the appalling number of maternal deaths (*see* **Maternal Mortality**) to the failure of the Royal Colleges to train doctors and midwives in obstetrics. In the cholera epidemic of 1854, he admonished the water companies for supplying water that was dangerously contaminated.

Among the honorary degrees and distinctions that were presented to him are the Honorary M.D. Degrees by the New York University in 1847, and by Trinity College, Dublin, in 1857. Also in 1857, he was elected Honorary Fellow of the Royal Medical and Chirurgical Society. In 1867, he became an Honorary Fellow of King and Queen's College of Physicians, Dublin. In 1868, he became a member of the important Joint Committee on State Medicine of the British Medical Association and the Social Science Association. By 1869, his interest had grown in medical statistics, and social problems – public health in particular – and he was made president of the Section of State Medicine by the British Medical Association. Between 1853 and 1876, he took an active part and interest in the International Statistical Congresses, and between 1876 and 1881 he was a member of the Anthropometric Committee of the British Association [4].

In his later career, Farr was a member of the Scientific Committee appointed by the General Board of Health to investigate the cholera epidemic of 1853–1854. In 1880, William Farr resigned his post at the General Register Office, and retired completely from public life. After Farr's retirement, a resolution was passed by the council of the British Medical Association [3]:

> The Gold Medal of the Association be awarded by the Committee of Council of the British Medical Association to William Farr, M.D., F.R.S., D.C.I., C.B., as an expression of their high appreciation of his long, unwearied, and successful labours in behalf of statistical and sanitary science; as a recognition of the light he has thrown upon many physiological and pathological problems, and on account of the extraordinary services his work has rendered to the advancement of the health of the nation.

William Farr was a man of undoubted genius.

References

[1] Eyler, J.M. (1979). *Victorian Social Medicine: The Ideas and Methods of William Farr*. Johns Hopkins University Press, Baltimore.

[2] Grebenik, E. (1968). Vital statistics, in *International Encyclopedia of the Social Sciences*, Vol. 16, D.L. Sills, ed. Macmillan & Free Press, New York, pp. 340–343.

[3] Greenwood, Major (1936). *The Medical Dictator and Other Biographical Studies*. Williams & Norgate, London.

[4] Susser, M. & Adelstein, A. (1975). Introduction, in *Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr*. Scarecrow Press, Metuchen.

LINA ASMAR

Fast Fourier Transform (FFT)

The discrete Fourier transform (DFT) of the time domain sequence $\boldsymbol{g} = (g_0, \ldots, g_{N-1})^\top$,

$$G_n = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} g_t \mathrm{e}^{-i2\pi nt/N}, \tag{1}$$

gives us the discrete-frequency representation of $\boldsymbol{g}$. G_n is often called the "Fourier coefficient" or "Harmonic" associated with frequency $f = n/N$. If t is measured in seconds, then f is in Hertz (or cycles per second). However, there is no reason why t has to be measured in time – it could, for example, be distance. The Fourier transform of the data, whatever the units of measurement, tells us about the periodic structure of the data. Calculation of the DFT is a critical step in many procedures, such as spectral density estimation, which describes how the variance of the process is distributed amongst the frequencies (*see* **Spectral Analysis**). Spectral density estimation finds widespread use across many disciplines including the analysis of electrocardiology (ECG), electroencephalography (EEG) data, and medical **image analysis** (*see* **Clinical Signals**).

The calculation of the N Fourier coefficients can be phrased in terms of **matrix algebra**,

$$\boldsymbol{G} = F_N \boldsymbol{g}, \tag{2}$$

where $\boldsymbol{G}$ and $\boldsymbol{g}$ are $N \times 1$ vectors given by $\boldsymbol{G} = (G_0, \ldots, G_{N-1})^\top$, $\boldsymbol{g} = (g_0, \ldots, g_{N-1})^\top$ and the $N \times N$ matrix F_N is given by,

$$F_N = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \ldots & & & & \ldots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)^2} \end{bmatrix}, \tag{3}$$

where $w = \mathrm{e}^{-i2\pi/N}$. This matrix is full, in that none of the elements are zero, this implies that direct calculation of the DFT is of order N^2, and as such would quickly become unfeasible for the majority of datasets, particular high-resolution images.

The fast calculation of the DFT is possible via a family of fast Fourier transform (FFT) **algorithms**. While the FFT has been attributed originally to **Gauss** (see [3]), it has been popularized by the Cooley–Tukey [1] algorithm. When N is a power of 2, the matrix can be factorized into the multiplication of three matrices as follows:

$$F_N = \begin{bmatrix} I_{N/2} & D_{N/2} \\ I_{N/2} & -D_{N/2} \end{bmatrix} \begin{bmatrix} F_{N/2} & 0_{N/2} \\ 0_{N/2} & F_{N/2} \end{bmatrix} \Pi_N, \tag{4}$$

where $I_{N/2}$ is the identity matrix, $0_{N/2}$ is a matrix of zeroes, and $D_{N/2}$ is a diagonal matrix with diagonal $(1, w, w^2, \ldots, w^{N/2})$. The second matrix is only half-full compared to F_N, which implies less arithmetic. Π_N is a permutation matrix that converts $(g_0, g_1, \ldots, g_{N-1})^\top$ into $(g_0, g_2, \ldots, g_{N-2}, g_1, g_3, \ldots, g_{N-1})^\top$. The effect of this factorization is to split the data into an *even* and an *odd* vector and calculate the DFT of these half-length vectors, the first matrix then recombines this vector to produce the right answer. This process can be repeated for the matrix containing $F_{N/2}$, until the matrix is reduced to the multiplication of very sparse matrices. The order of arithmetic calculations is required to evaluate the DFT drops from N^2 to $N \log_2 N$; this decrease in computational need becomes more pronounced the larger N becomes. For an excellent tutorial of FFTs, see [2].

The FFT can be used as a vital step in determining underlying periodicities in data. Given data $(X_0, \ldots X_{N-1})^\top$, the simplest spectral density estimator is the periodogram,

$$I(f_j) = \frac{1}{N} \left| \sum_{t=0}^{N-1} X_t \mathrm{e}^{-i2\pi f_j t} \right|^2, \quad f_j = \frac{j}{N}, \quad j = 0, \ldots, \frac{N}{2}. \tag{5}$$

and so utilizes the DFT of the data sequence.

Figure 1 shows some simulated data with an underlying annual and daily periodicity and the associated periodogram. While the annual cycle is evident in the time domain plot, the higher frequency daily periodicity is harder to distinguish. The peaks in the spectrum clearly show both inherent periodicities in the data – namely, at frequencies $1/365$ (annual) and $1/7$ (weekly).

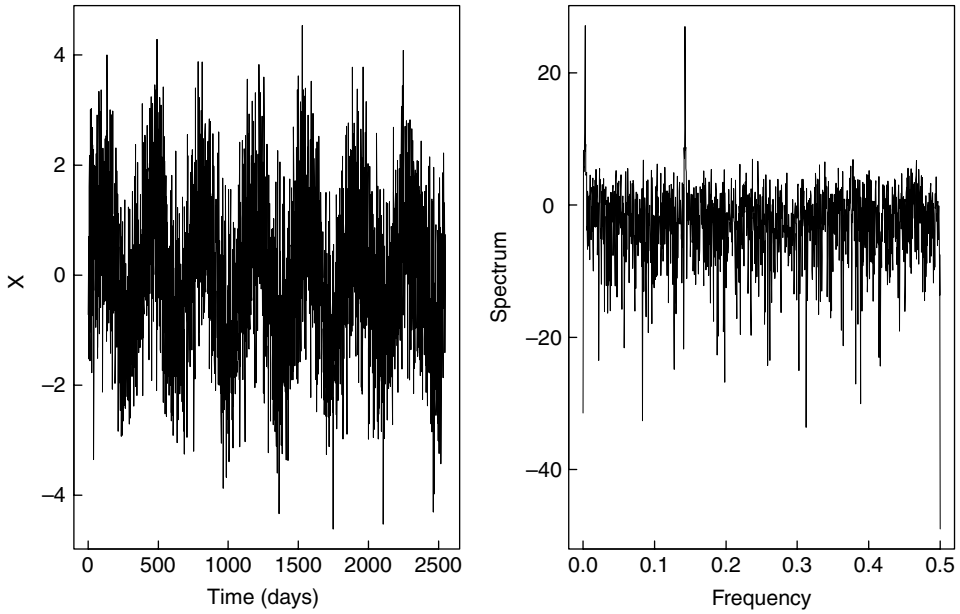

Figure 1 (a) Simulated data and (b) associated periodogram

References

[1] Cooley, J.W. & Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series, *Mathematics and Computation* **19**, 297–301.

[2] Duhamel, P. & Vetterli, M. (1990). Fast Fourier transforms: a tutorial review and a state of the art, *Signal Processing* **19**, 259–299.

[3] Heideman, M.T., Johnson, D.H. & Burrus, C.S. (1984). Gauss and the history of the fast Fourier transform, *IEEE ASSP Magazine* **1**(4), 14–21.

(*See also* **Time Series**)

E.J. McCoy

Fiducial Probability

What is fiducial probability? Introduced in a brief article [5] for the Cambridge Philosophical Society on "inverse probability", that is inference from sample to population, **R.A. Fisher** proposed fiducial probability as his alternative to **Bayesian** posterior probability. Consider a continuous univariate distribution parameterized by a parameter θ. Let $F(x, \theta)$ denote the one-parameter cumulative distribution function (cdf) for the **random variable** x, with density $f(x, \theta) = \partial F/\partial x$ (with respect to the Lebesgue measure). With $\text{fid}(\theta|x)$ denoting the fiducial probability density for θ given x, Fisher [5, p. 534] defined it as

$$\text{fid}(\theta|x) \propto -\frac{\partial F}{\partial \theta}. \tag{1}$$

Example 1 [7, p. 346]

Let $x_i \sim \text{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, be independent, identically distributed (iid) normal variates, where μ and σ^2 are the unknown population parameters. The sample statistics $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$ and $s^2 = \{1/(n-1)\}\sum_{i=1}^{n}(\overline{x} - x_i)^2$ are jointly **sufficient** for the two population parameters. Note that s^2 has a distribution that depends only on σ^2. Specifically, $(n-1)s^2/\sigma^2$ follows the **chi-square distribution** on $n-1$ **degrees of freedom**. Applying (1) to this distribution we obtain a fiducial probability for σ^2 that is inverse-χ^2, based on the sample variance s^2.

Following the 1930 publication, during the remaining 32 years of his life, through two books [[23], Chapter 10; [24], Section III.3] and numerous articles [6–15, 17–22, 25], Fisher steadfastly held to the idea captured in (1). However, his reasoning behind (1) evolved quite noticeably and contributed to the enigmatic quality of fiducial probability. If we distinguish between fiducial probability, (1), and the reasoning leading to it, which we may call "fiducial inverse inference", then there is little wonder that Fisher caused such puzzles with his novel idea.

There were confusions between fiducial inverse inference and, on the one hand, Bayesian posterior probability, such as **H. Jeffreys** advocated [27, 28, Section 7.1] at about the same time. Fisher [6] acknowledged that, at least sometimes, there were straightforward Bayesian models, i.e. "**prior densities**" $p(\theta)$, which resulted in the same inverse probabilities through **Bayes' theorem** [$p(\theta|x) \propto f(x, \theta)p(\theta)$] as his $\text{fid}(\theta|x)$ gave directly. Also, there were confusions between fiducial probability and the rival theory of **confidence intervals**, developed by Neyman [33] only a few years later. (Fisher's contribution to the discussion of Neyman's 1934 paper [33] only added to such confusions.) Here is an elementary case illustrating how all three theories might look alike.

Example 2

Consider x **uniformly distributed** on the interval $[0, \theta]$. That is, let $x \sim \text{U}[0, \theta], \theta > 0$. Thus, $f(x, \theta) = 1/\theta$ for $0 \le x \le \theta$, and $f(x, \theta) = 0$ otherwise, and $F(x, \theta) = x/\theta$ for $0 \le x \le \theta$, $F(x, \theta) = 0$ for $x \le 0$, and $F(x, \theta) = 1$ for $x \ge \theta$. Evidently, $\text{fid}(\theta|x) \propto -\partial F/\partial\theta = 1/\theta^2$ for $x \le \theta$, and $\text{fid}(\theta|x) = 0$ otherwise. [The normalizing constant here equals x, so that $\text{fid}(\theta|x) = x/\theta^2$.] Evidently, this is also a Bayesian posterior density for θ given x, based on the "improper" prior density $p(\theta) = 1/\theta$. This is nothing less than Jeffreys' recommended prior for a scale parameter, which θ is. Also, since the **likelihood** function [$f(x, \theta)$ taken as a function of θ] is monotone decreasing in θ for $x \le \theta$, by a well-known result of Neyman, there is a system of "best" confidence intervals for this problem. These are obtained by inverting on the family of uniformly **most powerful** (UMP) **likelihood ratio tests** for the different possible values of the null hypothesis $\theta = \theta_0$ against the two-sided alternative $\theta \ne \theta_0$. All three approaches produce the same numerical values for interval estimates. For instance, each assigns, respectively, a fiducial or posterior or confidence level of $1 - \alpha$ to the interval $[x \le \theta \le x/\alpha], 0 < \alpha \le 1$, e.g. the level 0.5 is assigned to the interval estimate $x \le \theta \le 2x$.

Fisher generalized his 1930 fiducial reasoning by focusing fiducial inference on pivotal quantities. Let θ be a real-valued parameter, let t be a statistic of the data, and let h be some real-valued function of them both. Call $v = h(\theta, t)$ a *pivotal* quantity if the distribution of v, $F(v)$, does not depend upon the parameter θ. Then pivotal reasoning in support of fiducial probability is just to say that $F(v) = F(v|t)$; that is, the **information** contained in t is irrelevant to v. From Example 1, let $t = (n-1)s^2$ and then $v = t/\sigma^2$ is pivotal, having a χ^2 distribution on

$n - 1$ degrees of freedom. Fiducial probability for the parameter is obtained by inverting on the distribution for the pivotal variable given the data, assuming the data are irrelevant to that pivotal.

When do pivotal variables exist? In the one-dimensional, continuous case they always do, as the cdf $F(x, \theta)$ is itself a pivotal! The cdf, F, is uniformly distributed on the closed unit interval, independent of $\theta : F \sim U[0, 1]$. If we invert on F as a pivotal, the resulting fiducial probability is Fisher's [5] $\text{fid}(\theta|x) \propto -\partial F/\partial\theta$, as given by (1). In Example 2, the pivotal $F = x/\theta$ illustrates this technique. (See Fraser [26] for important extensions of fiducial pivotal reasoning with group structures.)

Fiducial Probability and Confidence Intervals

Evidently, such pivotal reasoning conforms to confidence interval theory. However, not all pivotals support a fiducial probability, as the following illustrates.

Example 2 (Continued)

Suppose, as before, that $x \sim U[0, \theta]$, $\theta > 0$. However, let the parameter space be **truncated** with a known upper bound, $\theta \leq k$. Since the "best" confidence intervals for this problem are based on inverting UMP tests, they do not change with the upper bound, they merely get truncated. That is, the "best" system of confidence intervals for this problem assigns a confidence level of $1 - \alpha$ to the interval $[x \leq \theta \leq \min(x/\alpha, k)]$, $0 < \alpha \leq 1$. However, this system results in interval estimates that, at less than 100% confidence, cover the full parameter space consistent with the data. For example, if $k = 20$ and $\alpha = 0.1$, then this confidence interval system covers all possible parameter values at a 0.90 level whenever $x \geq 2$. Thus, inverting on a pivotal does not always produce a fiducial probability, since it does not always produce a probability, though it supports estimation by confidence intervals.

In the light of the phenomenon just illustrated, to permit the assumption that the data are irrelevant to the distribution of the pivotal, Fisher restricted fiducial reasoning with pivotals to those that have the same range and which are one-to-one for all possible data (and which are based on a sufficient statistic for the parameter of interest; see Tukey [39]). A practical case in point arose with interval estimates for the ratio of two normal means, where the pivotal reasoning supporting confidence intervals is invalid for fiducial probability, since those interval estimates cover the full parameter space at less than 100% confidence: the so-called "Creasy–Fieller" problem (*see* **Fieller's Theorem**). (See Fisher [14].) However, the difference between fiducial probability and confidence interval statements became evident 20 years earlier, in connection with Fisher's treatment of the so-called **"Behrens–Fisher" problem**: inference about the difference in two normal means from independent populations with unknown variance ratios. To appreciate that development, we need to consider more carefully the relationship between fiducial probability and Bayesian posterior probability.

According to Fisher, fiducial probability is special only by its genesis, not by its content. (He says this in numerous places, e.g. [24, p. 59].) That is, whatever we call fiducial probability must satisfy the mathematical calculus of **probability**. The uniqueness of fiducial probability is, supposedly, that it provides statements of inverse probability without admitting into the inference any (unwarranted) "prior" probability for statistical hypotheses, i.e. without relying on Bayes' theorem to derive inverse probability from a likelihood and prior probability. More accurately, fiducial inference attempts to derive inverse probability in the absence of statistically based prior probability. As Fisher expresses it [24, p. 59], a precondition for fiducial inference is that there is insufficient background knowledge to determine an initial (or "prior") value for probability about unknown parameters by direct inference using, say, a hyperpopulation.

Fisher's claim that fiducial probability is probability becomes the basis for its use in Bayes' theorem to solve other forms of statistical inference. In what follows we illustrate three such applications: inverse inference with data of two "kinds", inverse inference involving **nuisance parameters** – multiparameter fiducial inference, and predictive inference.

Data of Two "Kinds"

Suppose datum x admits fiducial inference about parameter θ, but that (independent) datum y (where y's distribution also depends only on θ) does not

allow fiducial inference. For instance, θ may be continuous though y is discrete and there is no acceptable pivotal connecting y and θ. Bayes' theorem yields: $p(\theta|x, y) \propto p(y|\theta)p(\theta|x)$. Fisher relies on fiducial inference to derive the inverse probability term "$p(\theta|x)$" and uses it in Bayes' theorem in this way, as illustrated in the next example.

Example 3 [24, Section 5.6]

Let x be **exponential**, with $F(x, \theta) = 1 - \exp(-x\theta)$ for $0 < \theta$, $0 \leq x$. Let y be a binomial count of a successes and b failures out of n independent trials, each trial with a chance of success $\rho = \exp(-c\theta)$. Then, given x, there is an inverse fiducial density $\text{fid}(\theta|x) = x\exp(-x\theta)d\theta$. Let $\lambda = x/c$. Transformed to express inverse probability for ρ, $\text{fid}(\rho|x) = \lambda\rho^{\lambda-1}\,\text{d}\rho$. But $p(y|\rho) \propto \rho^a(1-\rho)^b$. Hence, with the fiducial probability serving as a "prior" for ρ in Bayes' theorem, $p(\rho|x, y) \propto \rho^{a+\lambda-1}(1-\rho)^b\,\text{d}\rho$.

Fiducial Inference With Nuisance Factors – the "Step-by-step" Argument [24, Section 6.12]

Suppose δ, the parameter of interest, is bound to a nuisance parameter ζ, $p(\text{data}|\delta, \zeta)$ depends on ζ. That is, there is no satisfactory pivotal connecting (a sufficient summary of) the data with δ alone. Instead, let the likelihood factor in two components, for example,

$$p(g, h|\delta, \zeta) = p(g|\delta, \zeta, h)p(h|\zeta),$$

where (g, h) are a jointly sufficient reduction of the data with respect to the two parameters. Suppose the second factor, $p(h|\zeta)$, supports fiducial inference to yield $p(\zeta|g, h) = \text{fid}(\zeta|h)$. This corresponds to the claim that, in the absence of knowledge of δ, h summarizes all the relevant evidence about ζ. (The claim makes sense, I believe, only in connection with the step-by-step method, which affords a Bayesian check for the coherence of the claim. It is used in Example 1, for instance, to say that s^2 contains all the relevant information about σ^2 in the absence of knowledge of μ.) Last, suppose the first factor supports fiducial inference for δ from g, given ζ and h, $\text{fid}(\delta|\zeta, g, h)$. Then these terms may be combined using Bayes' theorem to yield

$$p(\delta|g, h) = \int_\zeta p(\delta|\zeta, g, h)p(\zeta|g, h)\,\text{d}\zeta.$$

This is Fisher's "step-by-step" method for solving the infamous Behrens–Fisher problem.

Example 4: The Behrens–Fisher Problem

Let x_i be iid $\text{N}(\mu_x, \sigma_x^2)$, $i = 1, \ldots, n$. Likewise, let y_i be iid $\text{N}(\mu_y, \sigma_y^2)$, $i = 1, \ldots, n$. All four parameters are unknown, but we are interested in the difference in means: $\delta = \mu_x - \mu_y$. The variances, σ_x^2 and σ_y^2, are nuisance factors. Define $\zeta = (\sigma_x^2/\sigma_y^2)$, the population variance ratio, and let $z = s_x^2/s_y^2$, the sample variance ratio. Last, define the quantity

$$d' = \frac{\overline{x} - \overline{y}}{\left\{\left[s_x^2 + \left(s_y^2\right)\zeta\right]\left[\dfrac{(\zeta+1)}{2\zeta}\right]\right\}^{0.5}}.$$

Then, given ζ, there is a simple fiducial inference from d' to δ, yielding: $p(\delta|d', \zeta)$, as $p(d'|\delta, \zeta)$ is a **Student's *t*** (with $2n - 2$ df), centered on the parameter of interest, δ. Fisher uses a fiducial inference from z to ζ, yielding $p(\zeta|z)$, as $p(z|\zeta)$ has Fisher's ***F* distribution**. (Here is where Fisher assumes z is sufficient for ζ in the absence of knowledge of δ.) Then these fiducial probabilities are combined using Bayes' theorem:

$$p(\delta|\text{data}) = \int_\zeta p(\delta|\,\text{d}', \zeta)p(\zeta|z)\,\text{d}\zeta.$$

It is important to understand that there can be no (exact) confidence intervals or "direct" fiducial argument obtained by inverting on a pivotal, duplicating this inference about δ [32]. That is, to appreciate the Bayesian aspects of the Behrens–Fisher solution, where Bayes' theorem is used to integrate out the nuisance parameter ζ, let us contrast it with the "step-by-step" fiducial method for inference about an unknown normal mean, μ, when σ is a nuisance parameter.

Example 5: Student's t-distribution as a Fiducial Probability

Let x_i be iid $\text{N}(\mu, \sigma^2)$, with both parameters unknown, but with μ, alone, the parameter of interest.

The two sample statistics $\overline{x}$ and s^2, are jointly sufficient for the two parameters. Recall that the likelihood for the data factors:

$$p(\overline{x}, s^2|\mu, \sigma^2) = p(\overline{x}|\mu, \sigma^2)p(s^2|\sigma^2).$$

The second term, $p(s^2|\sigma^2)$, supports fiducial inference about the nuisance factor σ^2, given s^2, as in Example 1. The first term, $p(\overline{x}|\mu, \sigma^2)$, supports fiducial inference about the parameter of interest, μ, given $\overline{x}$ and σ^2. Fiducially, $p(\mu|\overline{x}, \sigma^2)$ is normal $\mathrm{N}(\overline{x}, \sigma^2/n)$. These fiducial probabilities may be used in Bayes' theorem to solve for the marginal, inverse probability for the parameter of interest, just as in the Behrens–Fisher problem:

$$p(\mu|\overline{x}, s^2) = \int_\sigma p(\mu|\overline{x}, \sigma^2)p(\sigma^2|s^2)\ \mathrm{d}\sigma.$$

This yields the familiar Student's t distribution ($n - 1$ df) as a fiducial probability for μ. However, unlike the Behrens–Fisher distribution for δ, the t distribution may also be derived in a one-step, "direct" argument using the pivotal variable: $t = \mathrm{N}^{1/2}(\mu - \overline{x})/s$, which has Student's t distribution (with $n - 1$ df).

Alas, there is no guarantee that such "direct" fiducial reasoning coheres with what the "step-by-step" methods yields. The difficulty is illustrated in the next example.

Example 5 (Continued): The Buehler–Feddersen Problem

Let $n = 2$, so we have two (iid) observations from $\mathrm{N}(\mu, \sigma^2)$. Trivially, there is the direct probability,

$$p(x_{\min} \leq \mu \leq x_{\max}|\mu, \sigma^2) = 0.5, \qquad (2)$$

for each pair (μ, σ^2). Likewise, the fiducial "marginal" t probability (1 df) satisfies,

$$p(x_{\min} \leq \mu \leq x_{\max}|x_1, x_2) = 0.5, \qquad (3)$$

for all samples (x_1, x_2). Define the statistic $u = |x_1 + x_2|/|x_1 - x_2|$. Then, as Buehler & Feddersen [1] proved, within a year of Fisher's death, for each pair (μ, σ^2),

$$p(x_{\min} \leq \mu \leq x_{\max}|\mu, \sigma^2, u \leq 1.5) > 0.518. \qquad (4)$$

If the observed sample satisfies $u \leq 1.5$, then does not the inequality (4) give relevant information that conflicts with the statement (2), thus precluding (3)? Given $u \leq 1.5$, is not the fiducial step invalid for the t pivotal? Is the evidence $u \leq 1.5$ relevant to inference about the pivotal quantity: $p(t) \neq p(t|u \leq 1.5)$? The question remains open. For some recent discussion of this apparent conflict, see Seidenfeld [37, 38] and Zabell [40].

Fiducial Prediction

Prediction of independent observations offers a third variety of fiducial inference using fiducial probabilities in Bayes' theorem. Suppose that the joint likelihood for our data factors: $p(x_1, x_2|\mu) = p(x_1|\mu)p(x_2|\mu)$. Bayes' theorem leads to the result:

$$p(x_2|x_1) \propto \int_\mu p(x_2|\mu)p(\mu|x_1)\ \mathrm{d}\mu.$$

We can use the fiducial probability $\mathrm{fid}(\mu|x_1) = p(\mu|x_1)$ in this simple consequence of Bayes' theorem to derive a fiducial prediction for x_2 given x_1. Fisher [19] gives this analysis for a case of normal prediction when both μ and σ^2 are unknown. That problem involves the joint fiducial posterior for (μ, σ^2) given an observed sample, which then is integrated out to yield the fiducial prediction for a second, independent sample from the same population.

Non-Bayesian Aspects of Fiducial Inference

The fiducial argument displays its non-Bayesian character through reliance on the sample space of possible observations to locate its Bayesian model. That is, the prior in a Bayes' model for fiducial inference may depend upon which component of the likelihood is used to drive the fiducial argument.

Example 6: Inconsistent Fiducial Inferences Using Bayes' Theorem

Let $x \sim \mathrm{N}(\mu, 1)$ and, independently, let $y \sim \mathrm{N}(\nu, 1)$, where $\mu = \nu^3$. Such a variety of data might arise by using different measurement techniques for the same (theoretical) unknown parameter. However, because μ and ν are not linearly related, there is no real-valued sufficient statistic for the pair (x, y) – they

are minimally sufficient by themselves. The joint likelihood factors:

$$p(x, y|\mu) = p(x|\mu)p(y|\mu),$$

so there is the opportunity for using Bayes' theorem with a fiducial probability based on (either) one of these factors:

$$p(\mu|x, y) \propto p(x|\mu)\ \mathrm{fid}(\mu|y)$$

and

$$p(\mu|x, y) \propto p(y|\mu)\ \mathrm{fid}(\mu|x).$$

However, contrary to Bayes' theorem, the fiducial probability, $p(\mu|x, y)$, depends upon which factor of the likelihood is used for fiducial inference. This is readily understood in terms of Jeffreys' Bayesian model for the fiducial argument. When we create $p(\mu|y)$ by fiducial reasoning, we use the pivotal $y - \nu$ whose Bayesian model requires a uniform ("improper") prior over ν. When, instead, we create $p(\mu|x)$ by fiducial reasoning, we use the pivotal $x - \mu$ whose Bayesian model requires a uniform ("improper") prior over μ. Because μ and ν are nonlinear transformations of the same quantity, it is impossible to have a uniform distribution simultaneously over both. (See Lindley [31] for an important discussion of conditions when fiducial and Bayesian posterior probability agree. Fisher's reply [19] is disappointing, by comparison.) In his 1957 paper, Fisher [16] proposes a modified fiducial argument with inequalities in place of equalities of probabilities, e.g. fiducial conclusions of the form $p(\theta \geq 0) > 0.5$ to replace statements like $p(\theta \geq 0) = 0.5$. This idea relates to current research using sets of probabilities, rather than a single probability, to represent an inductive conclusion. Can ignorance be depicted by a large set of prior probabilities? Explicit connection of this approach with fiducial inference is found in Dempster's [3] work and in Kyburg's [29, 30] novel theory. Perhaps it is premature to say we have seen the end of the fiducial idea!

In a 1963 conference on fiducial probability, Savage wrote [36, p. 926]: "The aim of fiducial probability ... seems to be what I term 'making the Bayesian omelet without breaking the Bayesian eggs'." In that sense, fiducial probability is impossible. As with many great intellectual contributions, what is of lasting value is what we learn trying to understand Fisher's insights on fiducial probability. (See Edwards [4] for much more on this theme.) His solution to the Behrens–Fisher problem, for example, was a brilliant treatment of nuisance parameters using Bayes' theorem. In this sense, "... the fiducial argument is 'learning from Fisher"' [36, p. 926]. Thus interpreted, it certainly remains a valuable addition to the statistical lore.

References

[1] Buehler, R.J. & Feddersen, A.P. (1963). Note on a conditional property of Student's t, *Annals of Mathematical Statistics* **34**, 1098–1100.

[2] Dempster, A.P. (1963). On direct probabilities, *Journal of the Royal Statistical Society, Series B* **35**, 100–110.

[3] Dempster, A.P. (1966). New methods for reasoning towards posterior distributions based on sample data, *Annals of Mathematical Statistics* **37**, 355–374.

[4] Edwards, A.W.F. (1995). Fiducial inference and the fundamental theorem of natural selection, *Biometrics* **51**, 799–809.

[5] Fisher, R.A. (1930). Inverse probability, *Proceedings of the Cambridge Philosophical Society* **26**, 528–535.

[6] Fisher, R.A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters, *Proceedings of the Royal Society of London, Series A* **139**, 343–348.

[7] Fisher, R.A. (1935). The fiducial argument in statistical inference, *Annals of Eugenics* **6**, 391–398.

[8] Fisher, R.A. (1936). Uncertain inference, *Proceedings of the American Academy of Arts and Sciences* **71**, 245–258.

[9] Fisher, R.A. (1939). The comparison of samples with possibly unequal variances, *Annals of Eugenics* **9**, 174–180.

[10] Fisher, R.A. (1941). The asymptotic approach to Behrens's integral, *Annals of Eugenics* **11**, 141–172.

[11] Fisher, R.A. (1945). The logical inversion of the notion of the random variable, *Sankhyā* **7**, 129–132.

[12] Fisher, R.A. (1948). Conclusions fiduciaires, *Annales de l'Institut Henri Poincaré* **10**, 191–213.

[13] Fisher, R.A. (1951). Statistics, in *Scientific Thought in the Twentieth Century*, A.E. Heath, ed. Watts, London, pp. 31–55.

[14] Fisher, R.A. (1954). Contribution to a discussion of a paper on interval estimation by M.A. Creasy, *Journal of the Royal Statistical Society, Series B* **16**, 212–213.

[15] Fisher, R.A. (1955). Statistical methods and scientific induction, *Journal of the Royal Statistical Society, Series B* **17**, 69–78.

[16] Fisher, R.A. (1957). The underworld of probability, *Sankhyā* **18**, 201–210.

[17] Fisher, R.A. (1958). The nature of probability, *Centennial Review* **2**, 261–274.

[18] Fisher, R.A. (1959). Mathematical probability in the natural sciences, *Technometrics* **1**, 21–29.
[19] Fisher, R.A. (1960). On some extensions of Bayesian inference proposed by Mr. Lindley, *Journal of the Royal Statistical Society, Series B* **22**, 299–301.
[20] Fisher, R.A. (1961). Sampling the reference set, *Sankhyā* **23**, 3–8.
[21] Fisher, R.A. (1961). The weighted mean of two normal samples with unknown variance ratio, *Sankhyā* **23**, 103–114.
[22] Fisher, R.A. (1962). Some examples of Bayes' method of the experimental determination of probabilities a priori, *Journal of the Royal Statistical Society, Series B* **24**, 118–124.
[23] Fisher, R.A. (1971). *The Design of Experiments*, 8th Ed. Hafner, New York.
[24] Fisher, R.A. (1973). *Statistical Methods and Scientific Inference*, 3rd Ed. Hafner, New York.
[25] Fisher, R.A. & Cornish, E.A. (1960). The percentile points of distributions having known cumulants, *Technometrics* **2**, 209–225.
[26] Fraser, D.A.S. (1961). The fiducial method and invariance, *Biometrika* **48**, 261–280.
[27] Jeffreys, H. (1932). On the theory of errors and least squares, *Proceedings of the Royal Society of London, Series A* **138**, 48–55.
[28] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Oxford University Press, Oxford.
[29] Kyburg, H.E. (1961). *Probability and Logic of Rational Belief*. Wesleyan University, Middletown.
[30] Kyburg, H.E. (1974). *Logical Foundations of Statistical Inference*. Reidel, Boston.
[31] Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem, *Journal of the Royal Statistical Society, Series B* **20**, 102–107.
[32] Linnik, Yu.V. (1963). On the Behrens-Fisher problem, *Bulletin of the International Statistical Institute* **40**, 833–841.
[33] Neyman, J. (1934). On the two different aspects of the representative method, *Journal of the Royal Statistical Society, Series B* **97**, 558–625.
[34] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London, Series A* **236**, 333–380.
[35] Neyman, J. (1941). Fiducial argument and the theory of confidence intervals, *Biometrika* **32**, 128–150.
[36] Savage, L.J. (1963). Discussion, *Bulletin of the International Statistical Institute* **40**, 925–927.
[37] Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference*. Reidel, Dordrecht.
[38] Seidenfeld, T. (1992). R.A. Fisher's fiducial argument and Bayes' theorem, *Statistical Science* **7**, 358–368.
[39] Tukey, J.W. (1957). Some examples with fiducial relevance, *Annals of Mathematical Statistics* **28**, 687–695.
[40] Zabell, S.L. (1992). R.A. Fisher and the fiducial argument, *Statistical Science* **7**, 369–387.

(*See also* **Estimation, Interval**)

TEDDY SEIDENFELD

Fieller's Theorem

Fieller's theorem is used in finding a **confidence set** for a ratio of parameters, $\rho = \theta_1/\theta_2$. This problem arises in a variety of biostatistical problems including inverse dose estimation in quantal bioassay (*see* **Quantal Response Models**) [estimation of the LD_{50} (or **median effective dose**) is a special case], estimation of relative potency in **slope-ratio** and **parallel-line** bioassays (see [2]), and the assessment of **bioequivalence**. Other applications include inverse prediction in linear **calibration** and estimation of the point of intersection of two **linear regressions** or the point of extremum in a quadratic regression (*see* **Polynomial Regression**).

In general there are two statistics, $\hat{\theta}_1$ and $\hat{\theta}_2$, which estimate θ_1 and θ_2, respectively. It is assumed that $(\hat{\theta}_1, \hat{\theta}_2)$ follows either exactly or approximately a bivariate normal distribution with mean (θ_1, θ_2) with $\sigma_{11} = \text{var}(\hat{\theta}_1)$, $\sigma_{22} = \text{var}(\hat{\theta}_2)$, and $\sigma_{12} = \text{cov}(\hat{\theta}_1, \hat{\theta}_2)$. An estimate of the covariance matrix is available with $\hat{\sigma}_{ij}$ denoting the estimate of σ_{ij}. In the original, exact form of Fieller's theorem [1] $(\hat{\theta}_1, \hat{\theta}_2)$ is exactly normally distributed with $\sigma_{ij} = \sigma^2 v_{ij}$, where the v_{ij} are known constants, and there is a $\hat{\sigma}$, independent of $(\hat{\theta}_1, \hat{\theta}_2)$, such that $d\hat{\sigma}^2/\sigma^2$ follows a **chi-square distribution** with d **degrees of freedom**. This leads to

$$H(\rho) = \frac{(\hat{\theta}_1 - \rho\hat{\theta}_2)}{(\hat{\sigma}_{11} - 2\rho\hat{\sigma}_{12} + \rho^2\hat{\sigma}_{22})^{1/2}}$$

following exactly a **Student's *t* distribution** with d degrees of freedom. This arises when estimating a ratio of linear combinations in a normal linear model; see [5]. In other contexts $(\hat{\theta}_1, \hat{\theta}_2)$ is only approximately normal and the $\hat{\sigma}_{ij}$s are **consistent estimators** of the σ_{ij}s leading to $H(\rho)$ being approximately t distributed with d degrees of freedom; where $d = \infty$ designates a standard normal distribution.

With $t_{1-\alpha/2}(d)$ denoting the $100(1-\alpha/2)$th percentile of the t distribution with d degrees of freedom,

$$P[H(\rho)^2 \leq t_{1-\alpha/2}(d)^2] = 1 - \alpha. \qquad (1)$$

This holds exactly or approximately according to whether $H(\rho)$ follows exactly or approximately a t distribution, and the resulting confidence set is exact or approximate accordingly. Eq. (1) can be rewritten as $P(Q(\rho) \leq 0) = 1 - \alpha$, where $Q(\rho) = f_0 - 2f_1\rho + f_2\rho^2$ is a quadratic function of ρ, with $f_0 = \hat{\theta}_1^2 - t_{1-\alpha/2}(d)^2\hat{\sigma}_{11}$, $f_1 = \hat{\theta}_1\hat{\theta}_2 - t_{1-\alpha/2}(d)^2\hat{\sigma}_{12}$, and $f_2 = \hat{\theta}_2^2 - t_{1-\alpha/2}(d)^2\hat{\sigma}_{22}$. A confidence set for ρ with confidence coefficient $1 - \alpha$ is given by the set of values c satisfying $Q(c) \leq 0$. Defining $D = f_1^2 - f_0 f_2$, $r_1 = (f_1 - D^{1/2})/f_2$, and $r_2 = (f_1 + D^{1/2})/f_2$, the confidence set for ρ is:

Case 1. A finite interval $[r_1, r_2]$, if $D \geq 0$ and $f_2 \geq 0$.

Case 2. The complement of a finite interval, $(-\infty, r_2] \cup [r_1, \infty)$, if $D \geq 0$ and $f_2 < 0$.

Case 3. $(-\infty, \infty)$ if $D < 0$ and $f_2 < 0$.

It is known that $D < 0$ and $f_2 \geq 0$ cannot occur together; see [4]. Hence we get a finite interval if and only if $f_2 \geq 0$ or equivalently $|\hat{\theta}_2/\hat{\sigma}_{22}^{1/2}| \geq t_{1-\alpha/2}(d)$, which means rejecting $H_0 : \theta_2 = 0$ (*see* **Hypothesis Testing**). When $f_2 < 0$ we do not reject $H_0 : \theta_2 = 0$ and the confidence set is infinite (Cases 2 and 3). Note that if $\theta_2 = 0$, then ρ is ill defined. While such confidence sets have often been dismissed as uninformative or worse (Miller [3] calls them "absurdities") they can have a reasonable interpretation. Fortunately, a finite interval usually results in practice.

One alternative confidence interval uses

$$\hat{\rho} \pm t_{1-\alpha/2}(d)\left[\frac{\hat{\sigma}_{11} - 2\hat{\rho}\hat{\sigma}_{12} + \hat{\rho}^2\hat{\sigma}_{22}}{\hat{\theta}^2}\right]^{1/2},$$

where $\hat{\rho} = \hat{\theta}_1/\hat{\theta}_2$. This results from a **delta method** approximation to the variance of $\hat{\rho}$. While this interval (which is close to $[r_1, r_2]$ from Fieller's theorem for many data sets) is sometimes suitable it can perform badly in terms of achieving the desired confidence coefficient of $1 - \alpha$.

References

[1] Fieller, E.C. (1940). The biological standardization of insulin, *Royal Statistical Society* **7**, Supplement, 1–64.

[2] Finney, D.J. (1978). *Statistical Methods in Biological Assay*, 3rd Ed. Macmillan, New York.

[3] Miller, R.J. Jr (1986). *Beyond Anova, Basics of Applied Statistics*. Wiley, New York.

[4] Steffens, F.E. (1971). On confidence sets for the ratio of two normal means, *South African Statistical Journal* **5**, 105–113.

[5] Zerbe, G.O. (1978). On Fieller's theorem and the general linear model, *American Statistician* **32**, 103–105.

(*See also* **Biological Assay, Overview**)

JOHN P. BUONACCORSI

Finite Population Correction

In **simple random sampling** without replacement of n units from a population of N units (*see* **Sampling With and Without Replacement**), the **variance** of an **estimate** differs from that which would have been obtained under simple random sampling *with* replacement by a factor known as the *finite population correction* (fpc). This factor is given by

$$\text{fpc} = \frac{N-n}{N-1}.$$

For example, the variance of a sample **mean**, $\overline{x}$, of a variable, χ, is given under simple random sampling without replacement by the expression:

$$\text{var}(\overline{x}) = \frac{\sigma_x^2}{n}\left(\frac{N-n}{N-1}\right),$$

where σ_x^2 is the variance of the distribution of x. Under simple random sampling with replacement, the variance would be equal to σ_x^2/n without the fpc.

The effect of the fpc on the **standard error** of an estimate is very small in surveys where the sampling fraction, n/N, is low (e.g. below 10%). For sampling fractions that are not small, however, it could produce a sizable reduction in the standard error of an estimate and should not be ignored in the construction of **confidence intervals** or tests of hypotheses (*see* **Hypothesis Testing**). High sampling fractions are most likely to occur in practice when oversampling of relatively small population groups is carried out.

For an *estimated* variance under simple random sampling without replacement, the finite population correction is given by the expression, $(N-n)/N$, and the estimated variance of an estimated sample mean given by:

$$\widehat{\text{var}}(\overline{x}) = \frac{s_x^2}{n}\left(\frac{N-n}{N}\right),$$

where s_x^2 is the sum of the squared deviations about the sample mean divided by $n-1$. The derivation of this is shown in most sampling texts (e.g. Levy & Lemeshow [1]).

Some form of an fpc appears for designs other than simple random sampling that involve sampling without replacement (e.g. **stratified sampling**, single-stage **cluster sampling**, **multistage sampling**). For example, in single-stage cluster sampling, the fpc has the form, $(M-m)/(M-1)$, where M is the number of clusters in the population, and m is the number of clusters in the sample.

Reference

[1] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

PAUL S. LEVY

Fisher Lectures

The major statistical legacy of **R.A. Fisher** is marked by three series of lectures: in North America, in the United Kingdom, and in Australia.

North America

The R.A. Fisher Lectureship was established in 1963 by the **Committee of Presidents of Statistical Societies** (COPSS) to honor both the contributions of Sir Ronald Aylmer Fisher and the work of a present-day statistician for their advancement of statistical theory and applications. The list of Fisher lectures well reflects the prestige that COPSS and its constituent societies place on this lectureship. Biometrics and biostatistics have repeatedly been prominent in the lectures themselves in large part because of Fisher's pivotal contributions to these areas, directly and indirectly. The Fisher Lectureship is a very high recognition of meritorious achievement and scholarship in statistical science and recognizes the highly significant impact of statistical methods on scientific investigations.

The Fisher lecture is intended to be broadbased and emphasizes those aspects of statistics and probability, which bear close relationship to the scientific collection and interpretation of data, areas in which Fisher made outstanding contributions. The lecturer is expected to prepare a manuscript based on the appropriate lecture and to submit it to one of the COPSS society journals.

Fisher Lecturers

1964 Maurice S. Bartlett, "R.A. Fisher and the last fifty years of statistical methodology"
1965 Oscar Kempthorne, "Some aspects of experimental inference"
1967 John W. Tukey, "Some perspectives in data analysis"
1968 Leo A. Goodman, "The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries"
1970 Leonard J. Savage, "On rereading R.A. Fisher"
1971 Cuthbert Daniel, "One-at-a-time plans"
1972 William G. Cochran, "Experiments for nonlinear functions"
1973 Jerome Cornfield, "On making sense of data"
1974 George E.P. Box, "Science and statistics"
1975 Herman Chernoff, "Identifying an unknown member of a large population"
1976 George A. Barnard, "Robustness and the logic of pivotal inference"
1977 R.C. Bose, "R.A. Fisher's contribution to multivariate analysis and design of experiments"
1978 William H. Kruskal, "Statistics in society: problems unsolved and unformulated"
1979 C.R. Rao, "Fisher efficiency and estimation of several parameters"
1982 F.J. Anscombe, "How much to look at the data"
1983 I.R. Savage, "Nonparametric statistics and a microcosm"
1985 T.W. Anderson, "R.A. Fisher and multivariate analysis"
1986 David H. Blackwell, "Likelihood and sufficiency"
1987 Frederick Mosteller, "Methods for studying coincidences"
1988 Erich L. Lehmann, "Model specification: Fisher's views and some later strategies"
1989 Sir David R. Cox, "Probability models: their role in statistical analysis"
1990 Donald A.S. Fraser, "Statistical inference: likelihood to significance"
1991 David R. Brillinger, "Nerve cell spike train data analysis: a progression of technique"
1992 Paul Meier, "The scope of general estimation"
1993 Herbert E. Robbins, "N and n - sequential choice between two treatments"
1994 Elizabeth A. Thompson, "Likelihood and linkage: from Fisher to the future"
1995 Norman E. Breslow, "Statistics in epidemiology: the case-control study"
1996 Bradley Efron, "R.A. Fisher in the 21st Century"
1997 Colin L. Mallows, "The Zeroth Problem"
1998 Arthur Dempster, "Logistic statistics: modeling and inference"
1999 Jack D. Kalbfleisch, "Estimating functions and the bootstrap"
2000 Ingram Olkin, "R.A. Fisher and the combining of evidence"
2001 James O. Berger "Could Fisher, Jeffreys, and Neyman have agreed on testing?"

2002 Raymond J. Carroll, "Variability is not always a nuisance parameter"
2003 Adrian F.M. Smith, "On rereading L. J. Savage rereading R. A. Fisher"
2004 Donald B. Rubin, "Causal inference using potential outcomes: design, modelling, decisions"

United Kingdom

The series of Fisher Memorial Lectures in Great Britain is the responsibility of a Committee established in 1965 by representatives of the Royal Society of London, the **Royal Statistical Society**, the Genetical Society, and the British Region of the **International Biometric Society**. The Committee are the trustees of a fund whose purpose is to provide for lectures to be given "on any subject or field of science or learning associated with the name of the late Sir Ronald Aylmer Fisher, namely, the application of mathematics to biology", any residue to be used "to further knowledge and research in any of the said fields".

1966 F. Yates, "Computers, the second revolution in statistics"
1968 R.R. Race, "Blood groups in human genetics"
1969 E.A. Cornish, "Developments from the Fisher–Cornish expansions"
1970 K. Mather, "On biometrical genetics"
1972 G.A. Barnard, "Statistical inference in its historical development"
1974 L.L. Cavalli-Sforza, "Cultural versus biological evolution"
1977 R. Hide, "Motions in planetary fluids"
1978 D.J. Finney, "Bioassay and the practice of statistical inference"
1981 J. Maynard Smith, "The evolution of the sex ratio"
1981 J.H. Bennett, "R.A. Fisher and the genetical theory of natural selection"
1983 S. Karlin, "Kin selection and altruism"
1984 D.R. Cox, "Regression and the design of experiments"
1986 S.M. Stigler, "Francis Galton and the unravelling of the normal world"
1988 G.E.P. Box, "Scientific method in quality and productivity improvement"
1990 Sir Walter Bodmer, "Genetic sequences"
1992 D.V. Lindley, "Statistics of the market place"
1993 A.J. Jeffreys, "Molecular sleuthing: the story of genetic fingerprinting"
1994 A.W.F. Edwards, "Fiducial inference and the fundamental theorem of natural selection"
1995 M.J.R. Healy, "The life and work of Frank Yates"
1996 J.A. Nelder, "Computers: the continuing revolution in statistics"
1998 Sir John Kingman, "Mathematics of genetic diversity: before and after DNA"
2000 B. Efron, "The essential Fisher"
2001 Sir Richard Doll, "Proof of causality: Deductions from epidemiological evidence"
2002 Oliver Mayo, "The realisation of Fisher's research programme"
2003 Warren Ewens, "Statistics and the transition from genetics to genomics"
2004 Adrian F.M. Smith, "Towards an evidence-based society: the role of statistical thinking"

Australia

With the impending centenary of Fisher's birth in 1990, the following lectures were given at La Trobe University, Victoria, supported by the Department of Genetics and Human Variation:

1986 Bryan C. Clarke, "The Selective theory of molecular evolution"
1987 R.J. Berry, "Mice and the mess of molecular evolution"

In 1989, Professor P.A. Parsons decided to endow the continuation of this series at the University of Adelaide, where Fisher spent the last two years of his life. The lectures are specified to be "in a field to which Sir Ronald Fisher contributed namely, Genetics, Evolutionary Biology or Statistics". The lectures so far have been given by:

1990 R.S. Holmes, "Ultraviolet light and the cornea: genetic and biochemical aspects of ultra-violet radiation photoreception"
1992 A.W.F. Edwards, "Mendel, Galton, Fisher"
1995 P.A. Parsons, "Conservation strategies: adaptation to stress versus the preservation of genetic diversity"
1995 J.H. Bennett, "Fisher's work on inheritance in the tetrasomic wild plant *Lythrum salicaria* 1935–1959"

1996	Sir Gustav Nossal, "Genetics and vaccination: theoretical and practical aspects"	DOUGLAS G. ALTMAN
1997	Sean B. Carroll, "Living in the past: Xox genes and the evolution of animal body patterns"	

Fisher, Ronald Aylmer

Born: February 17, 1890, in East Finchley, London, UK.
Died: July 29, 1962, in Adelaide, Australia.

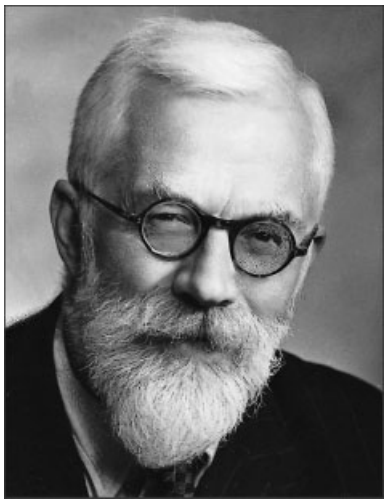

Reproduced by permission of the Royal Statistical Society

Ronald Aylmer Fisher achieved original scientific research of such diversity that the integrity of his approach is masked. Born into the era of Darwin's evolutionary theory and Maxwell's theory of gases, he sought to recognize the logical consequences of an indeterministic world, the certainties of which were essentially statistical. His interests were those of **Karl Pearson**, who dominated the fields of evolution, biometry, and statistics during his youth, but his perspective was very different. His ability to perceive remote logical connections of observation and argument gave his conceptions at once universal scope and coherent unity, so that he was little influenced by current scientific vogue at any period of his life.

Fisher was the seventh and youngest child of George Fisher, fine arts auctioneer in the West End, and Katie, daughter of Thomas Heath, solicitor of the City of London. His ancestors had showed no strong scientific bent, but his uncle, Arthur Fisher, was a Cambridge Wrangler.

Already, in childhood, Fisher met the misfortune of poor eyesight and the eye strain that was always to limit his private reading, and he learned to listen while others read aloud to him. His general intelligence and mathematical precocity were apparent early. From Mr Greville's school in Hampstead he went on to Stanmore in 1900, and entered Harrow School in 1904 with a scholarship in mathematics. In his second year there he won the Neeld Medal in mathematical competition with the whole school. To avoid eye strain he received tuition in mathematics under G.H.P. Mayo without pencil, paper, or other visual aids. Choosing spherical trigonometry for the subject of these tutorials, he developed a strong geometrical sense that was greatly to influence his later work. Fisher's interest in natural history was reflected in the books chosen for special school prizes at Harrow, culminating in his last year, in the choice of the complete works of Charles Darwin, in 13 volumes. In 1909 he won a scholarship in mathematics to Cambridge University. In 1912 he graduated from Cambridge as a Wrangler and, awarded a studentship for one year in the Cavendish Laboratory, studied the theory of errors under F.J.M. Stratton and statistical mechanics and quantum theory under J. Jeans.

In April 1912, Fisher's first paper [1] was published, in which the method of **maximum likelihood** was introduced (although not yet by that name). As a result, that summer Fisher wrote to **W.S. Gosset** ("Student") questioning his divisor $(n-1)$ in the formula for the **standard deviation**. He then reformulated the problem in an entirely different and equally original way in terms of the configuration of the sample in n-dimensional space, and showed that the use of the sample **mean** instead of the population mean was equivalent to reducing the dimensionality of the sample space by one; in this way he recognized the concept of what he later called **degrees of freedom**. Moreover, the geometrical formulation immediately yielded **Student's *t* distribution**, which Gosset had derived empirically, and in September Fisher sent Gosset the mathematical proof. This was included in Fisher's paper when, two years later, using the geometrical representation, he derived the general **sampling distribution** of the **correlation** coefficient [2].

Fisher's mathematical abilities were directed into statistical research by his interest in evolutionary theory, especially as it affected man. This interest, already developing at Harrow, resulted in the formation of the Cambridge University Eugenics Society in the spring of 1911, at Fisher's instigation. He served on its Council even while he was chairman

of the undergraduate committee, and he was the main speaker at the second annual meeting of the Society. While famous scientists wrangled about the validity either of evolutionary or of genetic theory, Fisher accepted both as mutually supportive. While the applicability of genetic principles to the continuous variables in man was disputed on biometric grounds, Fisher assumed that the observed variations were produced genetically, and in 1916 [3] justified this view by biometric argument.

In its application to man, selection theory raised not only scientific but practical problems. The birth rate showed a steep and regular decline relative to increased social status. This implied the existence throughout society of selection against every quality likely to bring social success (*see* **Adverse Selection**). Fisher believed, therefore, that it must result in a constant attrition of the good qualities of the population, such as no civilization could long withstand. He considered it important to establish the scientific theory on a firm quantitative basis through statistical and genetic research, and, more urgently, to publicize the scientific evidence so that measures could be taken to annul the self-destructive fertility trend.

Fisher accepted at once J.A. Cobb's suggestion in 1913 that the cause of the dysgenic selection lay in the economic advantage enjoyed at every level of society by the children of small families over those from larger families. Later, he proposed and urged the adoption of various schemes to spread the financial burden of parenthood, so that those who performed similar work should enjoy a similar standard of living, irrespective of the number of their children. In this he was not successful, and the family allowance scheme adopted in Great Britain after World War II disappointed him.

To further these aims, on leaving college he began work with the Eugenics Education Society of London, which was to continue for 20 years. From 1914 he was a regular book reviewer for the *Eugenics Review*; in 1920 he became business secretary and in 1930 vice-president of the Society; and he pursued related research throughout this period. Major Leonard Darwin, Charles Darwin's fourth son and president of the Society from 1912 to 1929, became a dear and revered friend, a constant encouragement and support while Fisher was struggling for recognition, and a stimulus to him in the quantitative research that resulted in *The Genetical Theory of Natural Selection* [12].

In 1913 Fisher took a statistical job with the Mercantile and General Investment Company in the City of London. He trained with the Territorial Army and, on the outbreak of war in August 1914, volunteered for military service. Deeply disappointed by his rejection due to poor eyesight, he served his country for the next five years by teaching high school physics and mathematics. While he found teaching unattractive, farming appealed to him both as a service to the nation and as the one life in which a numerous family might have advantages. When, in 1917, he married Ruth Eileen, daughter of Dr Henry Grattan Guinness (head of the Regions Beyond Missionary Union at the time of his death in 1915), Fisher rented a cottage and smallholding from which he could bicycle to school, and with Eileen and her sister began subsistence farming, selling the excess of dairy and pork products to supply needs for which the family could not be self-sufficient. Their evening hours were reserved for reading aloud, principally in the history of earlier civilizations.

In these years, Fisher's statistical work brought him to the notice of Karl Pearson. In 1915, Pearson published in ***Biometrika*** [2] Fisher's article on the general sampling distribution of the correlation coefficient, and went on to have the ordinates of the error distribution of estimated correlations calculated in his department. The resulting cooperative study was published in 1917, together with a criticism of Fisher's paper not previously communicated to its author. Pearson had not understood the method of maximum likelihood that Fisher had used, and condemned it as being inverse, or **Bayesian**, inference, which Fisher had deliberately avoided. Fisher, then unknown, was hurt by Pearson's high-handedness and lack of understanding, which eventually led to their violent confrontation. Meanwhile, Pearson ignored Fisher's proposal to assess the significance of correlations by considering not the correlation r itself but a remarkable **transformation** $z = \frac{1}{2}\ln[(1+r)/(1-r)]$ that reduced the highly **skewed** distributions with unequal **variances** to distributions which, to a close approximation, are **normal** with constant variance.

Fisher's paper on the correlation between relatives on the supposition of Mendelian inheritance [3] (*see* **Mendel's Laws**), submitted to the Royal Society in 1916, had to be withdrawn in view of the referees' comments (by Pearson and R.C. Punnett). Knowing that Pearson disagreed with his conclusions, Fisher had hoped that his new method, using **analysis of**

variance components, might have been persuasive. In this paper the subject and methodology of biometric genetics were created (*see* **Human Genetics, Overview**). These facts influenced Fisher's decision, in 1919, not to accept Pearson's guarded invitation to apply for a post in his department at University College London.

In September 1919, Fisher started work in a new, at first temporary, post as statistician at Rothamsted Experimental Station, where agricultural research had been in progress since 1843. He quickly became established in this work. He began with a study of historical data from one of the long-term experiments, with wheat on a field known as Broadbalk, but soon moved on to consider data obtained in current field trials for which he developed the analysis of variance. These studies brought out the inadequacies of the arrangement of the experiments themselves, and so led to the evolution of the science of **experimental design**. As Fisher worked with experimenters using successively improved designs, there emerged the principles of **randomization**, adequate replication, **blocking** and **confounding**, and **randomized blocks**, **Latin squares**, **factorial** arrangements, and other designs of unprecedented efficiency. The statistical methods were incorporated in successive editions of *Statistical Methods for Research Workers* (1925) [8], and an 11-page paper on the arrangement of field experiments [9] was expanded into a book, *The Design of Experiments* (1935) [14]. These volumes were supplemented by *Statistical Tables for Biological, Agricultural and Medical Research* (1938) [18], co-authored with **Frank Yates**.

Following up his work on the distribution of the correlation coefficient, Fisher derived the sampling distributions of other statistics in common use, including the variance ratio (called ***F*** in his honor by **G.W. Snedecor**) and the multiple correlation coefficient (*see* **Multiple Linear Regression**). Using geometric representations, he solved, for normally distributed errors, all the distribution problems for the **general linear model**, both when the **null hypothesis** is true and when an **alternative hypothesis** is true [7, 10].

Concurrently, the theory of **estimation** was developed in two fundamental papers in 1922 [5] and 1925 [6]. Fisher was primarily concerned with the small samples of observations available from scientific experiments, and was careful to draw a sharp distinction between sample statistics (estimates) and population values (parameters to be estimated). In the method of maximum likelihood he had found a general method of estimation that had many advantages. It not only provided a method by which to calculate unique numerical estimates for any problem that could be precisely stated, but also indicated what mathematical function of the observations ought to be used to estimate the parameter. His first application of the method of maximum likelihood was in 1922, to the estimation of genetic **linkage** in an example with no fewer than seven parameters.

In 1920, Fisher had compared two different estimators of the standard deviation σ of a normal distribution [4], showing that the sample standard deviation s was not only better but uniquely best, because the distribution of any other measure of spread conditional on s does not involve the parameter σ of interest. Once s is known, therefore, no other estimate gives any further information about σ. Fisher called this quality of s **sufficiency**. This finding led to his introduction of the concept of the amount of **information** in the sample, and to the criteria of the **consistency** and **efficiency** of estimators measured against the yardstick of available information. He exploited the **asymptotic relative efficiency** of the method of maximum likelihood in 1922, and, extending consideration to small samples in 1925, observed that small-sample sufficiency, when not directly available, was obtainable via **ancillary statistics** derived from the **likelihood** function.

Thus, seven years after moving to Rothamsted, Fisher had elucidated the underlying theory and provided the statistical methods that research workers urgently needed to deal with the ubiquitous variation encountered in biological experimentation. Thereafter, he continued to produce a succession of original researches on a wide variety of statistical topics. For example, he initiated nonlinear design, invented k-statistics, and explored **extreme-value** distributions, harmonic analysis, **multivariate analysis** and the **discriminant** function, the **analysis of covariance**, and new elaborations of experimental design and of sample surveys.

In developing his theory of estimation Fisher explored the type of uncertainty expressible precisely in terms of the likelihood, and his ideas on the subject never ceased to evolve. From the beginning he distinguished likelihood from mathematical probability, the highest form of scientific **inference**, which

he considered appropriate only for a restricted type of uncertainty. He accepted classical probability theory, of course, and used **Bayes' theorem** in cases in which there was an observational basis for making probability statements in advance about the population in question, as was often the case in genetics; furthermore, he proposed the **fiducial** argument as leading to true probability statements, at least in one common class of cases.

Fisher introduced the fiducial argument in 1930 [11]. In preparing a paper on the general sampling distribution of the multiple correlation coefficient in 1928, he noticed that in the test of significance the relationship between the estimate and the parameter was of a type that he later characterized as pivotal. He argued that if the one quantity were fixed, then the distribution of the other was determined; consequently, once the observations fixed the value of the observed statistic, the whole distribution of the unknown parameter was determined. Thus, in cases in which a pivotal relationship existed, true probability statements concerning continuous parameters could be inferred from the data provided that exhaustive (fully informative) estimators were available [13].

Controversy arose immediately. Fisher had proposed the fiducial argument as an alternative to the Bayesian argument of inverse probability, which he condemned in all cases in which no objective prior probability could be stated. While **H. Jeffreys** led the debate on behalf of the less restrictive use of inverse probability, in 1934 **J. Neyman** developed an approach to the theory of estimation by deliberately omitting from Fisher's fiducial theory the requirement for exhaustive estimation, thus inaugurating the theory of **confidence intervals**. In some instances this led to numerical results that were different from Fisher's. For many years the debate focused on the case of estimating the difference between two normally distributed populations with unknown variances not assumed to be equal (Behrens's test; *see* **Behrens–Fisher Problem**). Later difficulties with the fiducial argument arose in cases of multivariate estimation because of the nonuniqueness of the pivotals. Fisher did not achieve clarification of the criteria for selection among such alternative pivotals; he was working on the problem at the end of his life.

In proposing the fiducial argument in 1930, Fisher highlighted the main issues of scientific inference and compelled a more critical appreciation of the assumptions made, and of their consequences, in the various approaches to the problem (*see* **Inference, Foundations of**). In reviewing the subject in *Statistical Methods and Scientific Inference* (1956) [17], he distinguished the conditions in which he believed significance tests (*see* **Hypothesis Testing**), likelihood statements, and probability statements each had an appropriate and useful role to play in scientific inference.

In his genetic studies, having demonstrated the consonance of continuous variation in man with Mendelian principles, and having thereby achieved the fusion of biometry and genetics [3], in the 1920s Fisher tackled the problems of natural selection, expressed in terms of **population genetics**, culminating in *The Genetical Theory of Natural Selection* in 1930 [12], which heralded the neo-Darwinian revolution. The book was dictated to his wife during evenings at home; for a while it took the place of the reading and conversation that ranged from all of the classics of English literature to the newest archaeological research.

At home, too, was Fisher's growing family. His oldest son George was born in 1919; then followed a daughter who died in infancy, a second son, and in the end six younger daughters. True to his **eugenic** ideal, Fisher invested in the future, living simply under conditions of great financial stringency while the children were reared. He was an affectionate father, especially with George, who was soon old enough to join him in such activities as looking after genetic mouse stocks. Wherever possible, he brought the children into his activities and answered their questions seriously, with sometimes brilliant simplicity; he promoted family activities and family traditions. Domestic government was definitely patriarchal, and he punished larger offences against household rules, although with distaste. For as long as possible, the children were taught at home, for he trusted in their innate curiosity and initiative in exploring their world rather than in any imposed instruction. In fact, he treated his children like his students, as autonomous individuals from the beginning, assuming that they would act and think on their own responsibility, even when doing so involved danger or adult disapproval.

In 1929, Fisher was elected a Fellow of the Royal Society, as a mathematician. The influence of his statistical work was spreading, and he was already concerned that statistics should be taught as a practical art employing mathematical reasoning, not as self-contained mathematical theory. In 1933 Karl Pearson

retired and his department at University College, London was split; **E.S. Pearson** succeeded his father as head of the statistics department, and Fisher succeeded as Galton Professor of Eugenics, housed in the same building. For both men, it was an awkward situation. While others gave their interpretation of Fisher's ideas in the statistical department, he offered a course on the philosophy of experimentation in his own. After J. Neyman joined the statistics department in 1934, relations between the new departments deteriorated and controversy followed.

Fisher continued both statistical and genetic research. In 1931, and again in 1936, he was visiting professor for the summer sessions at Iowa State University at Ames, Iowa, at the invitation of G.W. Snedecor, director of the Statistical Laboratory. In 1937–1938 he spent six weeks as the guest of **P.C. Mahalanobis**, director of the Indian Statistical Institute in Calcutta. In his own department, where Karl Pearson had used only biometric and genealogic methods, Fisher quickly introduced genetics. Work with mouse stocks moved from the attic at his home, was expanded, and experimental programs were initiated on a variety of animal and plant species; for example, to study the problematic tristyly in *Lythrum salicaria*. Fisher was very eager also to initiate research in human genetics.

Sponsored by the Rockefeller Foundation, in 1935 he was able to set up a small unit for human serologic research under G.L. Taylor, who was joined by R.R. Race in 1937. His hopes that a linkage map of man could be built up using **blood groups** as **genetic markers** with a view to locating disease loci were only to be realized after his death. In 1943, he interpreted the bewildering results obtained with the new *Rh* blood groups in terms of three closely linked loci and correctly predicted the discovery of two new antibodies and an eighth allele. Fisher's enthusiasm for blood-group **polymorphisms** continued to the end of his life, and he did much to encourage studies of associations between blood groups and disease.

In 1927, Fisher proposed a way of measuring selective intensities on genetic polymorphisms occurring in wild populations, by a combination of laboratory breeding and field observation, and by this method later demonstrated very high rates of selective predation on grouse locusts. E.B. Ford was one of the few biologists who believed in natural selection at the time, and in 1928 he planned a long-term investigation of selection in the field, based on Fisher's method. To the end of his life, Fisher was closely associated with Ford in this work, which involved development of **capture–recapture** techniques and of sophisticated new methods of statistical analysis. The results were full of interest, and wholly justified their faith in the evolutionary efficacy of natural selection alone.

Fisher took a continuing delight in computation, introducing an electric calculating machine to Rothamsted and taking pleasure in the neat devices by which he could reduce the labor of computation. He showed little enthusiasm for electronic computers; yet in 1950 he was the first person to publish results in a biological context obtained from the new machines [16].

Forcibly evacuated from London on the outbreak of war in 1939, Fisher's department moved to Rothamsted and, finding no work as a unit, gradually dispersed; Fisher himself could find no work of national utility. In 1943 he was elected Arthur Balfour Professor of Genetics at Cambridge, which carried with it a professorial residence. Lacking other accommodation, he moved his staff and genetic stocks into the residence, leaving his family in Harpenden. Estranged from his wife, separated from home, and deeply grieved by the death in December 1943 of his son George on active service with the Royal Air Force, Fisher found companionship with his fellows at Caius College, and with the serologic unit (evacuated to Cambridge for war work with the Blood Transfusion Service), which planned to join his new department after the war.

There was little support after the war for earlier plans to build up an adequate genetics department. No bid was made to keep the serologic unit. No departmental building was erected. Support for the research in bacterial genetics, initiated in 1948 under L.L. Cavalli (Cavalli-Sforza), was withdrawn in 1950 just as Cavalli's discovery of the first *Hfr* strain of *Escherichia coli* heralded the remarkable discoveries soon to follow in bacterial and viral genetics. Fisher cultivated his garden, continued his research, published *The Theory of Inbreeding* (1949) [15] following his lectures on this topic, and built a group of good quantitative geneticists. He attempted to increase the usefulness of the university diploma in mathematical statistics by requiring all candidates to gain experience of statistical applications by doing research in a scientific department. Speaking as founding president of the **International Biometric**

Society (1947), as president of the **Royal Statistical Society**, and as a member or as president of the **International Statistical Institute**, he pointed out how mathematical statistics itself owes its origin and continuing growth to the consideration of scientific data rather than of theoretic problems.

In 1957, Fisher became involved in the controversy over the interpretation of the association between **smoking** and lung cancer, believing that the inference of **causation** was premature and likely to inhibit the further research that he felt was necessary.

Fisher's own interests extended to the work of scientists in many fields. He was a fascinating conversationalist at any time – original, thoughtful, erudite, witty, and irreverent; with the younger men his genuine interest and ability to listen, combined with his quickness to perceive the implications of their research, were irresistible. He encouraged, and contributed to, the new study of geomagnetism under S.K. Runcorn, a fellow of Gonville and Caius College, of which he was president during the period 1957–1960.

Fisher received many honors and awards: the Weldon Memorial Medal (1928), the Guy Medal of the Royal Statistical Society in gold (1947), and three medals of the Royal Society, the Royal Medal (1938), the Darwin Medal (1948), and the Copley Medal (1956); he was created Knight Bachelor by Queen Elizabeth in 1952.

After retirement in 1957, Sir Ronald Fisher traveled widely, joining E.A. Cornish in 1959 as honorary research fellow of the C.S.I.R.O. Division of Mathematical Statistics in Adelaide, South Australia. He died in Adelaide on July 29, 1962. His ashes are interred in St Peter's Cathedral, beneath a plaque in a side aisle.

References

[1] Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves, *Messenger of Mathematics* **41**, 155–160.

[2] Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika* **10**, 507–521.

[3] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[4] Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error, *Monthly Notices of the Royal Astronomical Society* **80**, 758–770.

[5] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, Series A* **222**, 309–368.

[6] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

[7] Fisher, R.A. (1925). Applications of "Student's" distribution, *Metron* **5**, 90–104.

[8] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[9] Fisher, R.A. (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture of Great Britain* **33**, 503–513.

[10] Fisher, R.A. (1928). The general sampling distribution of the multiple correlation coefficient, *Proceedings of the Royal Society, Series A* **121**, 654–673.

[11] Fisher, R.A. (1930). Inverse probability, *Proceedings of the Cambridge Philosophical Society* **26**, 528–535.

[12] Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

[13] Fisher, R.A. (1934). Two new properties of mathematical likelihood, *Proceedings of the Royal Society, Series A* **144**, 285–307.

[14] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[15] Fisher, R.A. (1949). *Theory of Inbreeding*. Oliver & Boyd, Edinburgh.

[16] Fisher, R.A. (1950). Gene frequencies in a cline determined by selection and diffusion, *Biometrics* **6**, 353–361.

[17] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Oliver & Boyd, Edinburgh.

[18] Fisher, R.A. & Yates, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver & Boyd, Edinburgh.

Bibliography

Bennett, J.H., ed. (1983). *Natural Selection, Heredity, and Eugenics. Including Selected Correspondence of R.A. Fisher with Leonard Darwin and Others*. Clarendon Press, Oxford.

Bennett, J.H., ed. (1990). *Statistical Inference and Analysis: Selected Correspondence of R.A. Fisher*. Clarendon Press, Oxford.

Box, J.Fisher (1978). *R.A. Fisher: The Life of a Scientist*. Wiley, New York.

Fisher, R.A. (1950). *Contributions to Mathematical Statistics*. Wiley, New York.

Fisher, R.A. (1971–74). *Collected Papers of R.A. Fisher*, 5 vols, J.H. Bennett, ed. University of Adelaide, Adelaide, Australia.

JOAN FISHER BOX & A.W.F. EDWARDS

Fisher's Exact Test

Fisher's exact test can be used to assess the significance of a difference between the proportions in two groups (*see* **Exact Inference for Categorical Data** for a broader context). The test was first described in independently written articles by Irwin [14] and Yates [25]. Yates used the test primarily to assess the accuracy of his correction factor to the χ^2 test, and attributed the key distributional result underlying the exact test to R.A. Fisher. Fisher successfully promoted the test in 1935, presenting two applications, one to an observational study on criminal behavior patterns [8], and another to an artificial example of a controlled experiment on taste discrimination [9].

Typical recent applications are to the results of simple experiments comparing a treatment with a control. The design must be completely randomized, and each experimental unit must yield one of two possible outcomes (like success or failure). Consider, for example, the study reported by Hall et al. [13]. This was a randomized, double-blind, placebo-controlled study on the effect of ribavirin aerosol therapy on a viral infection (RSV) of the lower respiratory tract of infants. After five days of treatment, each infant was examined for the continued presence of viral shedding in nasal secretions.

There were 26 patients in the randomized trial. For illustrative purposes, the following discussion focuses on hypothetical results from a smaller set of only eight patients. Also, a patient showing no signs of viral shedding in nasal secretions will be said to have recovered. Consider, then, the "results" displayed in Table 1.

All three recoveries were in the treatment group. For a frequency table based on only four treatments and four control subjects, the evidence could hardly be more convincing, but is it statistically significant? Had the experiment included more patients, an approximate P value could have been obtained using the standard χ^2 test (*see* **Chi-square Tests**). But we cannot trust the accuracy of this approximation when it is based on observations on so few patients. Fisher's exact test provides a way around this difficulty.

The reasoning behind the test is as follows. Suppose that the treatment was totally ineffectual, and that each patient's recovery over the subsequent five days was unaffected by whether the treatment were applied or not.

Precisely three patients recovered. If the treatment was ineffectual, then these three, and only these three, individuals would have recovered regardless of whether they were assigned to the treatment or control group. The fact that all three did indeed appear in the treatment group would then have been just a coincidence whose probability could be calculated as follows.

When four out of the eight subjects were randomly chosen for the treatment group, the chance that all three of those destined to recover should end up in the treatment group is given by the **hypergeometric distribution** as

$$\frac{({}_3C_3)({}_5C_1)}{{}_8C_4} = 0.071.$$

This is the standard ***P* value** for Fisher's exact test of the **null hypothesis** of no treatment effect against the one-sided alternative that the treatment has a positive benefit.

Consider the more general setting, as portrayed in Table 2.

The P value for testing the null hypothesis that the treatment has no impact vs. the one-sided alternative that it has a positive value is

$$P = \sum_{y=a}^{\min(n,S)} \frac{({}_SC_y)({}_FC_{n-y})}{{}_NC_n}. \qquad (1)$$

For a two-sided alternative there is no universally accepted definition. The two most common

Table 1 Results from a small, comparative experiment

	Recovered	Not recovered	Totals
Treatment	3	1	4
Control	0	4	4
Totals	3	5	8

Table 2 Notation for a 2×2 frequency table of outcomes from a comparative experiment

	Recovered	Not recovered	Totals
Treatment	a	b	n
Control	c	d	m
Totals	S	F	N

approaches are (i) to double the one-sided P value, or (ii) to extend the above sum over the other tail of the distribution, including all those terms which are less than or equal to the probability for the observed table. The latter strategy is deployed by the major statistical packages, BMDP (Two-Way Tables in [5]), JMP (Contingency Table Analysis in Fit Y by X in [21]), SAS (FREQ procedure in [20]), S-PLUS (function, fisher.test in [17]), SPSS (Crosstabs in [23]), StatXact [6], and Systat (Tables in [24]). Gibbons & Pratt [12] discuss possible alternatives.

The test can be extended to an $r \times c$ **contingency table**, as proposed by Freeman & Halton [11]. It is also used on $r \times 2$ tables for **multiple comparisons**, with the usual controversy over adjustments for simultaneous inferences on a single data set (see [22] and references therein).

Applicability and Power

A major advantage of Fisher's exact test is that it can be justified solely through the randomization in the experiment. The user need not assume that all patients in each group have the same recovery probability, nor that patient recoveries occur independently. The patients could, for example, go to one of four clinics, with two patients per clinic. Two patients attending the same clinic might well have experienced delayed recovery from having contacted the same subsidiary infection in their clinic, but the above argument would still be valid as long as *individuals* were randomly selected without restriction from the group of eight for assignment to the treatment vs. control groups.

If, however, the randomization was applied at the clinic level, with two of the four clinics selected for assignment to the treatment group, then the test would be invalid. Compared with the hypergeometric distribution, the data would most likely be overdispersed (*see* **Overdispersion**).

Similarly, if the randomization was restricted by blocking with respect to clinic, the pair of individuals from each clinic being randomly split between the treatment and control groups, then the test would again be invalid. These alternative designs certainly have their place, particularly in larger experiments with more subjects, but the results would have to be analyzed with another test.

The example also illustrates a major weakness of Fisher's exact test. The evidence for a table based on only four subjects in each of two groups could hardly have been more favorable to the alternative. Yet the P value still exceeds 5%, and most observers would rate the evidence as not statistically significant. It is in general difficult to obtain a statistically significant P value with Fisher's exact test, and the test therefore has low power. The most effective way to increase the power may well be to take quantitative measurements. Suppose, for instance, that all four patients who received the treatment showed reduced nasal shedding of the virus. By quantifying this evidence, and subjecting the quantitative measurements to a test of significance, the experimenter could, in many instances, generate a more powerful test.

One could also, of course, consider running the study on a larger group of patients.

Competing Binomial-model Test

It is also possible to obtain greater power by analyzing the above table with another statistical model. The most commonly used competitor involves assuming that the numbers of recovered patients in each group are independently binomially distributed (*see* **Binomial Distribution**). The test was mentioned by Irwin [14], and promoted by Barnard [2]. Although he soon withdrew his support [3], it has since become a popular alternative. Its increased power has been amply demonstrated by D'Agostino et al. [7] and others. For the above table, the P value is 0.035 vs. the 0.071 for Fisher's exact test. The P value based on this binomial model is typically smaller than the one generated by Fisher's exact test. The main reason for the difference is that the standard definition of the P value contains the probability of the observed table, and this probability is higher for Fisher's exact test than for the binomial model [1, 10, 18]. Thus the null hypothesis is more frequently rejected, and the binomial-model test is more powerful. This test is available in **StatXact** [6].

However, the increased power comes at a cost. To justify the binomial model, one must either assume that all patients within each group have the same recovery probability, or envisage that the patients were randomly sampled from some larger group. The trial must also have been conducted so as to ensure that patient deaths occur independently. They cannot, for example, attend four clinics, with two patients per clinic.

There is another, more subtle problem with the binomial model. Simple calculations show that had fewer than three or more than five patients recovered, then neither P value could possibly have been significant. This puts the researcher in an awkward quandary. For example, had only two patients recovered after 5 days, the researcher would have had an incentive either to present the results after more than five days of treatment when at least one more patient had recovered, or to incorporate more patients into the experiment. One does not win accolades for announcing the results of experiments that are not only statistically insignificant, but also apparently barely capable of ever producing significant results.

These are important complications when it comes to interpreting the results of these sorts of small experiments. Suppose, for example, that in the above experiment the researcher was to have adjusted the five-day reporting time, if necessary, so as to guarantee between three and five recoveries. Then the binomial P value would be invalid. The probability of obtaining a table at least as favorable to the treatment as the above one can be shown to be 0.056, not 0.035, as generated by the standard binomial model.

The Mid-*P* Value

The P value of 0.071 generated by Fisher's exact test is still large compared with the 0.056 figure produced by this modified binomial model. There is yet another alternative with important theoretical and practical advantages (see, for example, [16, 4, 1], and [19]). This is the mid-P value, first introduced in 1949 by Lancaster [15]. In place of the standard definition,

$$P \text{ value} = \Pr(\text{evidence at least as favorable to } H_a \text{ as observed } |H_0),$$

they propose the alternative,

$$\text{mid-}P \text{ value} = \Pr(\text{evidence more favorable to } H_a \text{ as observed } |H_0) + \tfrac{1}{2}\Pr(\text{evidence equally favorable to } H_a \text{ as observed } |H_0).$$

Table 3 summarizes the possible P values for the above example. This table illustrates that the mid-P has the potential to provide a smaller, more significant-looking P value, and to reduce the discrepancy between P values generated by competing models. However, by using a smaller P value, one may reject a valid null hypothesis too frequently. Fortunately, amongst other desirable attributes of the mid-P, its routine use does indeed control a quantity closely related to the type I error rate (see [19], and references therein). The computer package, StatXact [6] facilitates the calculation of the mid-P by providing the probability of the observed table along with the standard P value.

Table 3 Comparison of P values for the data in Table 1

	Fisher's exact test	Binomial model	Modified binomial model
Standard P value	7.1%	3.5%	5.6%
Mid-P value	3.6%	2.0%	3.0%

Conclusion

Fisher's exact test provides a widely applicable way to assess the results of simple randomized experiments leading to 2×2 contingency tables. But it has low power, especially when the standard P value is used. The power can be increased considerably through (i) using the mid-P value, or (ii) carefully constructing a test based at least in part on a binomial model. Further power increases can be generated through (iii) taking quantitative measurements on each subject, or (iv) running the trial with a larger number of patients.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley–Interscience, New York.

[2] Barnard, G.A. (1945). A new test for 2×2 tables, *Nature* **156**, 177.

[3] Barnard, G.A. (1949). Statistical inference, *Journal of the Royal Statistical Society, Series B* **11**, 115–139.

[4] Barnard, G.A. (1989). On alleged gains in power from lower p-values, *Statistics in Medicine* **8**, 1469–1477.

[5] BMDP Statistical Software, Inc. (1990). *BMDP Statistical Software Manual: To Accompany the 1990 Software Release*. University of California Press, Berkeley.

[6] Cytel Software Corporation (1995). *StatXact-3 for Windows*. Cytel Software Corporation, Cambridge, Mass.

[7] D'Agostino, R.B., Chase, W. & Belanger, A. (1988). The appropriateness of some common procedures for

testing the equality of two independent binomial populations, *American Statistician* **42**, 198–202.

[8] Fisher, R.A. (1935). The logic of inductive inference, *Journal of the Royal Statistical Society, Series A* **98**, 39–84.

[9] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[10] Franck, W.E. (1986). P-values for discrete test statistics, *Biometrical Journal* **4**, 403–406.

[11] Freeman, G.H. & Halton, J.H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance, *Biometrika* **38**, 141–149.

[12] Gibbons, J.D. & Pratt, J.W. (1975). P-values: interpretation and methodology, *American Statistician* **29**, 20–25.

[13] Hall, C.B., McBride, J.T., Gala, C.L., Hildreth, S.W. & Schnabel, K.C. (1985). Ribavirin treatment of respiratory syncytial viral infection in infants with underlying cardiopulmonary disease, *Journal of the American Medical Association* **254**, 3047–3051.

[14] Irwin, J.O. (1935). Tests of significance for differences between percentages based on small numbers, *Metron* **12**, 83–94.

[15] Lancaster, H.O. (1949). The combination of probabilities arising from data in discrete distributions, *Biometrika* **36**, 370–382.

[16] Lancaster, H.O. (1961). Significance tests in discrete distributions, *Journal of the American Statistical Association* **56**, 223–234.

[17] MathSoft, Inc. (1993). *S-PLUS Reference Manual, Version 3.2*. MathSoft, Inc., Seattle.

[18] Routledge, R.D. (1992). Resolving the conflict over Fisher's exact test, *Canadian Journal of Statistics* **20**, 201–209.

[19] Routledge, R.D. (1994). Practicing safe statistics with the mid-p, *Canadian Journal of Statistics* **22**, 103–110.

[20] SAS Institute, Inc. (1989). *SAS/STAT User's Guide, Version 6, 4th Ed., Vol. 1*. SAS Institute Inc., Cary.

[21] SAS Institute, Inc. (1995). *JMP Statistics and Graphics Guide, Version 3.1*. SAS Institute Inc., Cary.

[22] Savitz, D.A. & Olshan, A.F. (1995). Multiple comparisons and related issues in the interpretation of epidemiological data, *American Journal of Epidemiology* **142**, 904–908.

[23] SPSS, Inc. (1991). *SPSS Statistical Algorithms*, 2nd Ed. SPSS Inc., Chicago.

[24] SYSTAT, Inc. (1992). *SYSTAT for Windows: Statistics, Version 5*. SYSTAT, Inc., Evanston.

[25] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society, Supplement* **1**, 217–235.

(*See also* **Two-by-Two Table**; **Yates's Continuity Correction**)

RICK ROUTLEDGE

Fixed Effects

Consider an **explanatory variable** which takes on k possible values in a particular data set and which is to be related to a **response variable** via a **regression** model. Assume that some function of the response variable is related to the linear predictor $\mu + \alpha_i, \ \ i = 1, k-1$; or, equivalently, $\alpha_i, i = 1, k$, where i indexes the possible values of the explanatory variable (*see* **Dummy Variables**). Examples of such explanatory variables are an indicator for clinics in a multiclinic study, school classrooms in a study of school children, different studies in a **meta-analysis**, and blocking factors in **experimental design**. Other explanatory variables may be included in the linear predictor. For example, a single additional variable could be added to define a linear predictor $\mu + \alpha_i + \beta X, \ \ i = 1, k-1$.

If the regression analysis focuses on the estimates of the α_is only for values of the explanatory variable in the data set, then the α_is are referred to as fixed effects. This contrasts with the assumption that the α_is represent a **random sample** from a distribution of effects associated with a wider range of possibilities for the explanatory variable, in which case the α_is are referred to as **random effects**. For general discussions of the distinction and its effect on methods of analysis *see* **Analysis of Variance**.

VERN T. FAREWELL

Fixed Population

A fixed population or fixed cohort is a group of individuals defined by a common fixed characteristic, such as all men in the US born in 1941. Membership in a fixed population does not change over time by immigration or emigration, unlike a **dynamic population**, although members of a fixed cohort may experience the health event under study, or may die or be lost to follow-up.

MITCHELL H. GAIL

Fix–Neyman Process

Fix & Neyman [2] introduced a stochastic model to describe recovery, relapse, death and loss of patients in medical follow-up studies of cancer patients. The model is useful also in other areas of research including compartmental analysis, **survival analysis**, and reliability studies. Some examples of applications of the process are given in Table 1.

The Fix–Neyman process has two transient states, S_1 and S_2, denoting health and illness of patients, and two absorbing states, R_1 and R_2, for causes of death. We shall consider it as a special case of a finite **Markov process** [1] with two states. Generally, the states in a Markov process are communicative: any state in the process can be reached from any other state in the process. This means that we shall consider only the two health states S_1 and S_2 in the process, and derive formulas related to the death states R_1 and R_2 through their relationship with the health states. Since a death state is an absorbing state, the number of death states does not influence the complexity of the process.

Fix & Neyman regarded state S_1 as a health state, and S_2 as an illness state, so that a transition from S_1 to S_2 means onset of illness or a relapse, and a transition from S_2 and S_1 means recovery. In the following discussion, we shall consider illness as a state of health, and allow transitions $S_1 \to S_2$ and $S_2 \to S_1$ to take place without specific designation.

Consider, then, a system consisting of two health states, S_1 and S_2, and two death states, R_1 and R_2. Transitions are possible between the two health states S_1 and S_2 and from either one of the health states to a death state. Each death state is an absorbing state: once an individual enters it, he will remain there for ever. Figure 1 is a graphic description of the transition process; the arrows indicate the directions a transition may take place.

Table 1 Examples of Applications of the Fix–Neyman Process

	S_1	S_2
A person is	employed	unemployed
An elevator is	occupied	unoccupied
A telephone line is	engaged	free
A nuclear particle counting instrument is	free	dead
An automobile is	working	out of order

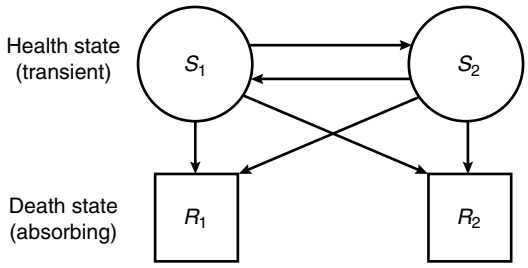

Figure 1 Transitions in the Fix–Neyman process

Transition Probabilities and Intensity Functions

For a time interval (τ, t), we define the *health transition probabilities*

$$P_{\alpha\beta}(\tau, t) = \Pr\{\text{an individual in state } S_\alpha \text{ at } \tau \text{ will be in state } S_\beta \text{ at } t\}, \quad \alpha, \beta = 1, 2, \qquad (1)$$

and the *death transition probabilities*

$$Q_{\alpha\delta}(\tau, t) = \Pr\{\text{an individual in state } S_\alpha \text{ at } \tau \text{ will be in state } R_\delta \text{ at } t\}, \quad \alpha = 1, 2, \delta = 1, 2.$$

At time $t = \tau$, these probabilities will have obvious specific values:

$$P_{\alpha\beta}(\tau, \tau) = \begin{cases} 1, & \beta = \alpha, \\ 0, & \beta \neq \alpha, \alpha, \beta = 1, 2, \end{cases}$$
$$Q_{\alpha\delta}(\tau, \tau) = 0, \quad \alpha = 1, 2, \delta = 1, 2.$$

We assume independence between the transitions in (1). Consider two contiguous time intervals (τ, ξ) and (ξ, t) and two events:

$$A = \{\text{an individual in state } \alpha \text{ at time } \tau \text{ will be in state } \gamma \text{ at time } \xi\}$$

and

$$B = \{\text{an individual in state } \gamma \text{ at time } \xi \text{ will be in state } \beta \text{ at time } t\}.$$

If the events A and B are assumed independent, then we have an important set of Chapman–Kolmogorov equations:

$$P_{\alpha\beta}(\tau, t) = \sum_{\gamma} P_{\alpha\gamma}(\tau, \xi) P_{\gamma\beta}(\xi, t). \tag{2}$$

Transition Intensity Functions

Transition from one health state S_α to another health state S_β is governed by the health intensity function, or transition rate, $\nu_{\alpha\beta}$, and transition from a health state S_α to a death state R_δ is governed by the death intensity function $\mu_{\alpha\delta}$, with the additional definition

$$\nu_{\alpha\alpha} = -\left[\nu_{\alpha\beta} + \sum_{\delta=1}^{2} \mu_{\alpha\delta}\right], \quad \alpha \neq \beta, \alpha, \beta = 1, 2, \tag{3}$$

where $\nu_{\alpha\alpha} < 0, \alpha = 1, 2$. Since these intensity functions are independent of time, the process is time homogeneous.

Differential Equations and Transition Probabilities

For the interval $(\tau, t + \Delta)$ and the two contiguous intervals (τ, t) and $(t, t + \Delta)$, the Chapman–Kolmogorov equations are

$$\begin{aligned} P_{\alpha\alpha}(\tau, t + \Delta) &= P_{\alpha\alpha}(\tau, t) P_{\alpha\alpha}(t, t + \Delta) \\ &\quad + P_{\alpha\beta}(\tau, t) P_{\beta\alpha}(t, t + \Delta), \\ P_{\alpha\beta}(\tau, t + \Delta) &= P_{\alpha\alpha}(\tau, t) P_{\alpha\beta}(t, t + \Delta) \\ &\quad + P_{\alpha\beta}(\tau, t) P_{\beta\beta}(t, t + \Delta). \end{aligned}$$

These equations, as $\Delta \to 0$, lead to the following differential equations:

$$\begin{aligned} \frac{\partial}{\partial t} P_{\alpha\alpha}(\tau, t) &= P_{\alpha\alpha}(\tau, t)\nu_{\alpha\alpha} + P_{\alpha\beta}(\tau, t)\nu_{\beta\alpha}, \\ \frac{\partial}{\partial t} P_{\alpha\beta}(\tau, t) &= P_{\alpha\alpha}(\tau, t)\nu_{\alpha\beta} + P_{\alpha\beta}(\tau, t)\nu_{\beta\beta}, \\ &\quad \alpha \neq \beta, \alpha, \beta = 1, 2. \end{aligned} \tag{4}$$

This is a system of linear, homogeneous, first-order ordinary differential equations with constant coefficients. (Note that, since τ is fixed, the resemblance of (4) to a set of partial differential equations is only formal.) The solution is

$$P_{\alpha\alpha}(\tau, t) = \sum_{i=1}^{2} \frac{\rho_i - \nu_{\beta\beta}}{\rho_i - \rho_j} \exp[\rho_i(t - \tau)]$$

and

$$\begin{aligned} P_{\alpha\beta}(\tau, t) &= \sum_{i=1}^{2} \frac{\nu_{\alpha\beta}}{\rho_i - \rho_j} \exp[\rho_i(t - \tau)], \\ &\quad j \neq i, \ \alpha \neq \beta, \ j, \alpha, \beta = 1, 2, \end{aligned} \tag{5}$$

where

$$\rho_1 = \tfrac{1}{2}\{\nu_{11} + \nu_{22} + [(\nu_{11} - \nu_{22})^2 + 4\nu_{12}\nu_{21}]^{1/2}\},$$

and

$$\rho_2 = \tfrac{1}{2}\{\nu_{11} + \nu_{22} - [(\nu_{11} - \nu_{22})^2 + 4\nu_{12}\nu_{21}]^{1/2}\}$$

are both negative.

The probabilities in (5) depend only on the difference $t - \tau$ but not on τ and t separately; thus the process is *homogeneous with respect to time*, as pointed out earlier. We shall therefore let $\tau = 0$ and let t be the interval length, and write

$$\begin{aligned} P_{\alpha\alpha}(0, t) &= \sum_{i=1}^{2} \frac{\rho_i - \nu_{\beta\beta}}{\rho_i - \rho_j} \exp(\rho_i t), \\ P_{\alpha\beta}(0, t) &= \sum_{i=1}^{2} \frac{\nu_{\alpha\beta}}{\rho_i - \rho_j} \exp(\rho_i t), \quad j \neq i, \\ &\quad \alpha \neq \beta, \ j, \alpha, \beta = 1, 2. \end{aligned} \tag{6}$$

Death Transition Probabilities

The death transition probability $Q_{\alpha\delta}(0, t)$ has a definite relation with the health transition probabilities:

$$Q_{\alpha\delta}(0, t) = \int_0^t P_{\alpha\alpha}(0, \tau)\mu_{\alpha\delta}\, d\tau + \int_0^t P_{\alpha\beta}(0, \tau)\mu_{\beta\delta}\, d\tau. \tag{7}$$

Substitution of (6) into (7), and integration of the resulting expression, gives the formula for the death transition probability:

$$Q_{\alpha\delta}(0, t) = \sum_{i=1}^{2} \frac{\exp(\rho_i t) - 1}{\rho_i(\rho_i - \rho_j)}[(\rho_i - \nu_{\beta\beta})\mu_{\alpha\delta}$$

$$+ \nu_{\alpha\beta}\mu_{\beta\delta}],\ j \neq i,\ \alpha \neq \beta,$$
$$j, \alpha, \beta, \delta = 1, 2. \tag{8}$$

Chapman–Kolmogorov Equation

The Fix–Neyman process described in this article is a Markov process in the sense that the transitions an individual might make in the future are independent of the transitions made in the past. An important consequence of this Markovian property is the Chapman–Kolmogorov equations in (2). Since this process is homogeneous with respect to time, we may rewrite (2) as follows:

$$P_{\alpha\alpha}(0, t) = P_{\alpha\alpha}(0, \tau)P_{\alpha\alpha}(\tau, t) + P_{\alpha\beta}(0, \tau)P_{\beta\alpha}(\tau, t)$$

and

$$P_{\alpha\beta}(0, t) = P_{\alpha\alpha}(0, \tau)P_{\alpha\beta}(\tau, t) + P_{\alpha\beta}(0, \tau)P_{\beta\beta}(\tau, t),$$

for $0 \leq \tau \leq t, \alpha \neq \beta, \alpha, \beta = 1, 2$.

Chapman–Kolmogorov-type equations can be established also for the transition probabilities leading to death:

$$Q_{\alpha\delta}(0, t) = Q_{\alpha\delta}(0, \tau) + P_{\alpha\alpha}(0, \tau)Q_{\alpha\delta}(\tau, t) + P_{\alpha\beta}(0, \tau)Q_{\beta\delta}(\tau, t),$$
$$\alpha \neq \beta,\ \alpha, \beta, \delta = 1, 2.$$

These equations may be verified from (6) and (8).

Expected Duration of Stay

In a study of human health, we may wish to estimate the length of time a person is expected to be healthy. In other studies we may inquire about the length of time an automobile is expected to be in working condition, a person is expected to be employed, or a telephone line is expected to be busy. These inquiries lead to an important concept in the Fix–Neyman process: What is the expected duration of stay in each of the states S_1, S_2, R_1, and R_2 within a time period of length t? This duration depends on the initial state and the corresponding transition probability. For an individual in state S_α at time $t = 0$, let

$$e_{\alpha\beta}(t) = \text{the expected duration of stay in } S_\beta \text{ in the interval } (0, t), \quad \beta = 1, 2,$$

and

$$\varepsilon_{\alpha\delta}(t) = \text{the expected duration of stay in } R_\delta \text{ in the interval } (0, t), \quad \delta = 1, 2.$$

These quantities are related to the transition probabilities by the following equations:

$$e_{\alpha\beta}(t) = \int_0^t P_{\alpha\beta}(0, \tau)\, d\tau$$

and

$$\varepsilon_{\alpha\delta}(t) = \int_0^t Q_{\alpha\delta}(0, \tau)\, d\tau.$$

Substitution from (6) and (8) gives the explicit formulas:

$$e_{\alpha\alpha}(t) = \sum_{i=1}^{2} \frac{\rho_i - \nu_{\beta\beta}}{\rho_i(\rho_i - \rho_j)}(\exp(\rho_i t) - 1)$$

and

$$e_{\alpha\beta}(t) = \sum_{i=1}^{2} \frac{\nu_{\alpha\beta}}{\rho_i(\rho_i - \rho_j)}(\exp(\rho_i t) - 1)$$

for $\alpha \neq \beta, \alpha, \beta = 1, 2$, and

$$\varepsilon_{\alpha\delta}(t) = \sum_{i=1}^{2} \left\{ \frac{1}{\rho_i}[\exp(\rho_i t) - 1] - t \right\} \times \frac{(\rho_i - \nu_{\beta\beta})\mu_{\alpha\delta} + \nu_{\alpha\beta}\mu_{\beta\delta}}{\rho_i(\rho_i - \rho_j)},$$

for $\alpha \neq \beta, \alpha, \beta, \delta = 1, 2$. The sum of the expected durations of stay over all the states is equal to the entire length of the interval:

$$e_{\alpha 1}(t) + e_{\alpha 2}(t) + \varepsilon_{\alpha 1}(t) + \varepsilon_{\alpha 2}(t) = t, \quad \alpha = 1, 2.$$

Limiting Probabilities

Since both ρ_1 and ρ_2 are negative, each of the health transition probabilities in (6) approaches zero as $t \to \infty$:

$$\lim_{t\to 0} P_{\alpha\alpha}(0, t) = \lim_{t\to\infty} \sum_{i=1}^{2} \frac{\rho_i - \nu_{\beta\beta}}{\rho_i - \rho_j} \exp(\rho_i t) = 0,$$

$$\lim_{t\to 0} P_{\alpha\beta}(0, t) = \lim_{t\to\infty} \sum_{i=1}^{2} \frac{\nu_{\alpha\beta}}{\rho_i - \rho_j} \exp(\rho_i t) = 0,$$
$$j \neq i,\ \alpha \neq \beta,\ j, \alpha, \beta = 1, 2,$$

and each of the death transition probabilities in (8) approaches a constant:

$$\lim_{t\to\infty} Q_{\alpha\delta}(0,t) = \lim_{t\to\infty} \sum_{i=1}^{2} \frac{\exp(\rho_i t) - 1}{\rho_i(\rho_i - \rho_j)} \times [(\rho_i - \nu_{\beta\beta})\mu_{\alpha\delta} + \nu_{\alpha\beta}\mu_{\beta\delta}] = \sum_{i=1}^{2} \frac{-[(\rho_i - \nu_{\beta\beta})\mu_{\alpha\delta} + \nu_{\alpha\beta}\mu_{\beta\delta}]}{\rho_i(\rho_i - \rho_j)},$$
$$\alpha \neq \beta,\ \alpha, \beta, \delta = 1, 2.$$

where

$$\sum_{\delta=1}^{2} Q_{\alpha\delta}(0, \infty) = 1.$$

A Time-dependent Fix–Neyman Process

In the above discussion the intensity functions $\nu_{\alpha\beta}$ and $\mu_{\alpha\delta}$ were assumed to be independent of time. If they are replaced by time-dependent functions $\nu_{\alpha\beta}\theta(\xi)$ and $\mu_{\alpha\delta}\theta(\xi)$, with (3) and ρ_1 and ρ_2 unchanged, the formulas for the transition probabilities become

$$P_{\alpha\alpha}(\tau, t) = \sum_{i=1}^{2} \frac{\rho_i - \nu_{\beta\beta}}{\rho_i - \rho_j} \exp\left[\rho_i \int_\tau^t \theta(\xi)\, d\xi\right],$$

$$P_{\alpha\beta}(\tau, t) = \sum_{i=1}^{2} \frac{\nu_{\alpha\beta}}{\rho_i - \rho_j} \exp\left[\rho_i \int_\tau^t \theta(\xi)\, d\xi\right],$$
$$j \neq i,\ \alpha \neq \beta,\ j, \alpha, \beta = 1, 2,$$

and

$$Q_{\alpha\delta}(\tau, t) = \sum_{i=1}^{2} \frac{\exp\left[\rho_i \int_\tau^t \theta(\xi)\, d\xi\right] - 1}{\rho_i(\rho_i - \rho_j)} \times [(\rho_i - \nu_{\beta\beta})\mu_{\alpha\delta} + \nu_{\alpha\beta}\mu_{\beta\delta}]$$

for $\alpha \neq \beta, \alpha, \beta, \delta = 1, 2$.

References

[1] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Krieger, New York.

[2] Fix, E. & Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients, *Human Biology* **23**, 205–241.

CHIN LONG CHIANG

Fleiss, Joseph L.

Born: November 13, 1937, in Brooklyn, New York.
Died: June 12, 2003, in Ridgewood, New Jersey.

Joseph L. Fleiss was a statistician whose writings influenced thousands of biomedical researchers in fields ranging from **psychiatry** to dentistry. His most influential contributions were in the analysis of binary and **categorical data**, statistical analysis of diagnostic reliability (*see* **Diagnostic Tests, Evaluation of**), and the design and analysis of **clinical trials**. He wrote more than 200 scientific and statistical papers, as well as two books that are considered classics in biostatistics.

Fleiss received an A.B. cum laude from Columbia College in 1959. While still in college, he worked as a statistical clerk with the Biometrics Research Unit at the New York State Psychiatric Institute (NYSPI). After attending the program in Biostatistics at the University of Minnesota in 1960, he returned to the Columbia School of Public Health where he earned his M.S. degree in Biostatistics in 1961. He received his Ph.D. from Columbia's Department of Mathematical Statistics in 1967, writing a dissertation on "Analysis of variance in assessing errors in interview data". This work was motivated by applied problems that he encountered in the Biometrics Unit, and the progress he made guaranteed his position at NYSPI for many years to come. He worked full time at NYSPI as a Research Scientist and Biostatistician until 1975 and remained affiliated with the institute until 1986.

In 1975, Fleiss was recruited by Columbia University to be Professor and Head of the Division of Biostatistics at the School of Public Health, taking over a program that had been established by John W. Fertig. During Fleiss's tenure as Head of the Biostatistics division, the program grew in stature and size. He established a Ph.D. program in 1977. Over the next 15 years, he recruited a first-class faculty to train new doctoral and master's degree students, generate independent research in biostatistics, and support clinical research initiatives throughout the health sciences at Columbia. Even as the department thrived, Fleiss's health went into a steep decline. A particularly disabling form of Parkinson's disease led to his stepping down as Biostatistics Head in 1992, at the age of 55. He continued to think about biostatistics problems as long as he could but was unable to work beyond the age of 58. His last statistical paper, a commentary on **meta-analysis**, was published in 1995.

Even though Fleiss's career was tragically cut short by illness, his 30 years of productivity had an enormous impact. His first book, *Statistical Methods for Rates and Proportions* [4], addressed the apparently simple issue of using proportions to summarize counts and frequencies. Although this appeared to be a specialized topic, the book attracted a wide readership with its many engaging examples and a thorough, but accessible, discussion of esoteric statistical principles. The book was released in a second edition in 1981. It was particularly influential in the fields of **epidemiology** and psychiatry, where it was used to define acceptable statistical practice for many scientific journals. After Fleiss's death in 2003, a third edition was released by Wiley, through the efforts of his Columbia colleagues Bruce Levin and Myunghee Cho Paik [14].

Fleiss's second book, *The Design and Analysis of Clinical Experiments* [7], was equally influential but with a completely different group of medical researchers. This book reviewed the principles of **experimental design** for experiments on people. Unlike rats in basic science studies, people present experimenters with all sorts of difficulties. They refuse to participate after **randomization**, they drop out, they fail to take the treatment, and

they sometimes misrepresent outcomes. Fleiss's text addressed these issues directly. Instead of simply dwelling on the details of statistical analysis of data already in hand, Fleiss challenged biostatisticians and medical researchers to think carefully about how to plan and interpret studies that used humans as the experimental subjects. This book was so influential that the publisher, John Wiley and Sons, reissued it in the Wiley Classics Library series.

Fleiss began his career in psychiatric research and won acclaim in this field for his many valued publications. His work with colleagues at NYSPI is widely cited as instrumental in the study of measurement and reliability of psychiatric diagnoses [2, 20–23]. A paper with Endicott, Spitzer, and Cohen [2] on the Global Assessment Scale has been especially influential, with over 2000 citations in the literature as of January 2004.

Dentistry also benefited from Fleiss's contributions. In 1983, he was recruited as chairman of the Task Force on Design and Analysis, an independent, not-for-profit organization of biostatisticians and clinical scientists working in dental and craniofacial research. Under Fleiss's leadership, the Task Force promoted the development and use of creative strategies in the conduct and analysis of dental research studies, with major funding provided by commercial sponsors. Over the years, the Task Force has played a key role in the development of basic study designs for oral health trials and techniques for measuring periodontal disease. Fleiss coauthored a number of published manuscripts on the design and analysis of dental research data [1, 10, 15, 16], as well as two Task Force–sponsored guideline papers on conducting clinical trials of treatments for gingivitis and periodontitis [17, 18].

Fleiss was also recognized as a leader in the field of randomized controlled trials. He served as the senior statistical consultant on several important, major randomized trials in **cardiology** and **neurology**. He also produced reader-friendly manuscripts on the history, design, and analysis of clinical trials [6, 8, 9, 11].

Fleiss was a prolific researcher, publishing over 200 journal articles in his curtailed career, as well as numerous book chapters, book reviews, and editorial letters. A number of his articles on statistical methods continue to be influential today. Among his most cited articles are a paper with Cohen and Everitt on large sample standard errors for **kappa** [13], with 458 citations; a paper on measuring nominal scale agreement [3], with 682 citations; another paper with Cohen on kappa [12], with 326 citations; a paper on measuring agreement between two judges (*see* **Agreement, Measurement of**) [5], with 220 citations; and a very frequently referenced paper with Shrout on the use of intraclass correlations in assessing reliability (*see* **Correlation**) [19], with almost 2500 citations (January 2004 data).

Recognized as a leading biostatistician, Fleiss was asked to participate in a number of scientific panels, workshops, and review groups for the **Food and Drug Administration** and the **National Institutes of Health**. He served as a reviewer for many scientific journals and as an associate editor for **Biometrics** (a leading journal in statistical methods for medical, public health, and biological research) from 1975 to 1984. In 1986, Fleiss was elected President of the Eastern North American Region of the **International Biometric Society**.

Fleiss received many honors for his work bridging statistics and medical science. He was elected Fellow of the **American Statistical Association** and won the Mortimer **Spiegelman** Health Statistics Award from the **American Public Health Association** in 1973. In 1988, the Statistics Section of the American Public Health Association (1988) presented him with their Recognition Award. The Departments of Epidemiology and Biostatistics at Harvard University honored him with a Lifetime Contribution Award in 1998.

Fleiss married Isabel Bogorad, whom he met on Columbia University's Morningside campus. They are survived by their three children, Arthur, Elizabeth, and Deborah, and six grandchildren (Amir, Yamit, Eden, Shana, Sarah, and Jesse).

References

[1] Chilton, N.W. & Fleiss, J.L. (1986). Design and analysis of plaque and gingivitis clinical trials, *Journal of Clinical Periodontology* **13**, 400–410.

[2] Endicott, J., Spitzer, R.L., Fleiss, J.L. & Cohen, J.L. (1976). The global assessment scale: a procedure for measuring overall severity of psychiatric disturbance, *Archives of General Psychiatry* **33**, 766–771.

[3] Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin* **76**, 378–382.

[4] Fleiss, J.L. (1973). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York.

[5] Fleiss, J.L. (1975). Measuring agreement between 2 judges on presence of absence of a trait, *Biometrics* **31**, 651–659.

[6] Fleiss, J.L. (1982). Multicenter clinical trials: Bradford Hill's contributions and some subsequent developments, *Statistics in Medicine* **1**, 353–359.

[7] Fleiss, J.L. (1986a). *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, New York.

[8] Fleiss, J.L. (1986b). Analysis of data from multiclinic clinical trials, *Controlled Clinical Trials* **7**, 267–275.

[9] Fleiss, J.L. (1989). A critique of recent research on the two-treatment crossover design, *Controlled Clinical Trials* **10**, 237–243.

[10] Fleiss, J.L. (1992). General design issues in efficacy, equivalency and superiority trials, *Journal of Periodontal Research* **27**, 306–313.

[11] Fleiss, J.L., Bigger, J.T., McDermott, J., Miller, J.P., Moon, T., Moss, A.J., Oakes, D., Rolnitzky, L.M. & Therneau, T.M. (1990). Nonfatal myocardial infarction is, by itself, an inappropriate endpoint in clinical trials in cardiology, *Circulation* **81**, 684–685.

[12] Fleiss, J.L. & Cohen, J.L. (1973). The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability, *Educational and Psychological Measurement* **33**, 613–619.

[13] Fleiss, J.L., Cohen, J.L. & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa, *Psychological Bulletin* **72**, 323–327.

[14] Fleiss, J.L., Levin, B. & Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*, 3rd Ed. John Wiley & Sons, New York.

[15] Fleiss, J.L., Mann, J., Paik, M., Goultchin, J. & Chilton, N.W. (1991). A study of inter- and intra-examiner reliability of pocket depth and attachment level, *Journal of Periodontal Research* **26**, 122–128.

[16] Fleiss, J.L., Park, M.H. & Chilton, N.W. (1987). Within-mouth correlations and reliabilities for probing depth and attachment level, *Journal of Periodontology* **58**, 460–463.

[17] Imrey, P., Chilton, N.W., Pihlstrom, B.L., Proskin, H.M., Kingman, A., Listgarten, M.A., Zimmerman, S.O., Ciancio, C.G., Cohen, M.E., D'Agostino, R.B., Fischman, S.L., Fleiss, J.L., Gunsolley, J.C., Kent, R.L., Jr., Killoy, W.J., Laster, L.L., Marks, R.G. & Varma, A.O. (1994a). Recommended revisions to American dental association guidelines for acceptance of chemotherapeutic products for gingivitis control, *Journal of Periodontal Research* **29**, 299–304.

[18] Imrey, P., Chilton, N.W., Pihlstrom, B.L., Proskin, H.M., Kingman, A., Listgarten, M.A., Zimmerman, S.O., Ciancio, S.G., Cohen, M.E., D'Agostino, R.B., Fischman, S.L., Fleiss, J.L., Gunsolley, J.C., Kent, R.L., Killoy, W.J., Laster, L.L., Marks, R.G. & Varma, A.O. (1994b). Proposed guidelines for the American dental association acceptance of products for professional, non-surgical treatment of adult periodontitis, *Journal of Periodontal Research* **29**, 348–360.

[19] Shrout, P.E. & Fleiss, J.L. (1979). Intra-class correlations: uses in assessing rater reliability, *Psychological Bulletin* **86**, 420–428.

[20] Shrout, P.E., Spitzer, R.L. & Fleiss, J.L. (1987). Quantification of agreement in psychiatric diagnosis revisited, *Archives of General Psychiatry* **44**, 172–177.

[21] Spitzer, R.L., Cohen, J.L., Fleiss, J.L. & Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis: a new approach, *Archives of General Psychiatry* **17**, 83–87.

[22] Spitzer, R.L., Endicott, J., Fleiss, J.L. & Cohen, J.L. (1970). Psychiatric status schedule: a technique for evaluating psychopathology and impairment in role functioning, *Archives of General Psychiatry* **23**, 41–55.

[23] Spitzer, R.L. & Fleiss, J.L. (1974). A re-analysis of the reliability of psychiatric diagnosis, *British Journal of Psychiatry* **125**, 341–347.

PATRICK E. SHROUT & MELISSA D. BEGG

Floating Point Arithmetic

Floating point number representations are a response to the demand to store and manipulate in a computer, using a fixed number of digits, numbers which range widely in magnitude. Most formats that are in common use store the rough equivalent of either 7 or 15 decimal digits. The consequent finite precision of floating point arithmetic has large implications for practical computation, of a kind that we will explore in this article.

Floating point contrasts with the *fixed point* representations that are common on metering devices. For example, a car's dashboard indicators may represent distance to the nearest tenth of a kilometer or of a mile, i.e. with the decimal point fixed one place from the right. In addition to storing the digits of the number, a floating point representation stores an indicator of the position of a varying or *floating* point.

Floating Point Representation

For illustration, assume a computer representation which retains the seven most significant decimal digits in any calculated result. While modern computers typically use a base of 2 or 16, exactly the same principles apply.

On such a computer, 12024 may appear as 0.12024×10^5. The *fraction* $(= 0.12024)$ and *exponent* $(= 5)$ are stored as separate items:

exponent			fraction							
+	0	5	+	1	2	0	2	4	0	0

Note that the number is *normalized*, i.e. stored so that the leading digit in the fraction is nonzero. The number of digits for the fraction determines the *precision* of the representation, while the number of digits for the exponent determines the range of numbers that can be represented.

The square of 12024, calculated in a double-width register, is 144576576. Storage back in a seven-digit register gives

exponent			fraction							
+	0	7	+	1	4	4	5	7	6	6

The relative error from rounding to seven decimal digit precision is largest when a number of the form $0.10000004999\ldots \times 10^z$ is rounded to 0.1×10^z. Thus, the relative error is never more than $u = 0.5 \times 10^{-6}$. The number u, known as the *unit roundoff*, measures the relative *precision* of the numerical representation. Machine epsilon (*macheps*), which is the distance between 1.0 and the next largest exact representation, is twice the unit roundoff.

Observe the contrast between *precision* and *accuracy*. The limited *precision* of the arithmetic restricts the *accuracy* that is possible in any actual calculation. The example that now follows demonstrates how failure to take account of the finite precision of floating point numbers may lead to a catastrophic loss of accuracy.

The Sum of Squares About the Mean

Sums of squares and products about the mean appear in the calculation of variances, covariances, correlations, and a variety of multivariate calculations. We consider the calculation of sums of squares about the mean, i.e. of

$$S = \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{1}$$

The alternative expression,

$$S = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}, \tag{2}$$

allows, in principle, a one-pass calculation of $\sum_{i=1}^{n} (x_i - \overline{x})^2$. With $x_1 = 4007, x_2 = 4008, x_3 = 4009$, on a computer which rounds the result of each pairwise calculation to the seven most significant decimal digits before proceeding, it yields $16056050 + 16064060 + 161072080 - 144576600/3 = 48192190 - 48192200 = -10$.

Table 1 gives the steps at which accuracy is lost in this calculation.

The final subtraction in (2) made obvious a loss of accuracy that had occurred earlier in the formation of each of the quantities $\sum_{i=1}^{n} x_i^2$ and $(\sum_{i=1}^{n} x_i)^2/n$. More generally, the use of the algorithm that is based on (2) for a set of numbers where the mean is $\overline{x}$ and the standard deviation is s can be expected to

Table 1

	Exact	Machine representation
x_1^2	16056049	0.1605605×10^8
x_2^2	16064064	0.1606406×10^8
x_2^3	16072081	0.1607208×10^8
$x_1^2 + x_2^2 + x_3^2$	–	0.4819219×10^8
$(x_1 + x_2 + x_3)^2$	144576576	0.1445766×10^9
$(x_1 + x_2 + x_3)^2/3$	–	0.4819220×10^8

lose around $2 \log_{10} \overline{x}/s$ decimal digits of precision [6, p.12] in the calculated result.

A simple *one-pass* algorithm that is vastly preferable to (2) uses

$$S = \sum_{i=1}^{n} (x_i - x_1)^2 - \left[\sum_{i=1}^{n} (x_i - x_1)\right]^2 \Big/ n.$$

Updating Algorithms

Even better is the use of a formula that updates the mean at each step. It illustrates the

$$\text{new value} = \text{old value} + \text{correction}$$

approach which is often a good starting point for the design of a stable algorithm, i.e. an algorithm that avoids unnecessary loss of accuracy. We include also the extension that allows the calculation of cross-products. For this, we define $\overline{x}^{(k)} = k^{-1} \sum_{i=1}^{k} x_i$, with $\overline{y}^{(k)}$ defined similarly, and

$$s_{xx}^{(k)} = \sum_{i=1}^{k} (x_i - \overline{x}^{(k)})^2,$$

$$s_{xy}^{(k)} = \sum_{i=1}^{k} (x_i - \overline{x}^{(k)})(y_i - \overline{y}^{(k)}).$$

Then

$$\overline{x}^{(k)} = \overline{x}^{(k-1)} + k^{-1}(x_k - \overline{x}^{(k-1)}),$$

$$s_{xx}^{(k)} = s_{xx}^{(k-1)} + (x_k - \overline{x}^{(k-1)})(x_k - \overline{x}^{(k)}),$$

$$s_{xy}^{(k)} = s_{xy}^{(k-1)} + (x_k - \overline{x}^{(k-1)})(y_k - \overline{y}^{(k)})$$

(see [1] and [6, p. 13]).

Implications for Regression Calculations

The least squares regression problem (*see* **Linear Regression, Simple**) determines **b** such that $\mathbf{(y - Xb)'(y - Xb) = ||y - Xb||}_2$ is a minimum. Algebraically, it is equivalent to solving the normal equations

$$\mathbf{X'Xb = X'y}.$$

Whenever the mean of one or more columns of **X** is large relative to its standard deviation, the computed versions of $\mathbf{X'X}$ and $\mathbf{X'y}$ may be inaccurate representations of the data in **X** and **y**.

Floating Point Number Systems

A floating point number system is characterized by the base β, the number of digits t in that base, and the exponent range. Thus, numbers take the form $\pm m \times \beta^{\pm z}$. In the example just given we had a base $\beta = 10$, with $t = 7$ digits stored. Some Hewlett-Packard calculators had $\beta = 10$, $t = 12$, and were able to represent numbers ranging from approximately 10^{-499} to just under 10^{500} with a relative error of no more than $0.5 \times \beta^{1-t} = 0.5 \times 10^{-11}$. The use of base 10 does not take advantage of the simplicity of the numeric representation that is available when the base is a power of 2, typically 2 itself or $2^4 = 16$. While it avoids conversion of numbers between bases at input and output, it slows arithmetic calculations.

The IEEE Standard

The use of the base $\beta = 2$ is now almost universal for computers, with $\beta = 16$ on high-end IBM machines the main exception. The Institute of Electrical and Electronic Engineers (IEEE) standard 754 [4], developed by a working group over several years, has become the accepted international standard for floating point arithmetic. This specifies $\beta = 2$, $t = 24$ binary digits ($\approx$ 7 decimal digits) for single precision, and $t = 53$ ($\approx$ 15 decimal digits) for double precision. Table 2 gives details of selected parameters of IEEE single- and double-precision floating point arithmetic. Most major computer manufacturers now supply processors that implement these standards.

Overflow, i.e. a number that is too large to be represented, or division of a nonzero number by zero, both lead to $\pm\infty$. The operations 0/0, $0 \times$

Table 2 Selected parameters of IEEE single and double-precision floating point arithmetic

	Single-precision	Double-precision
Maximum floating point number	$2^{128} = 3.403 \times 10^{38}$	$2^{1024} = 1.798 \times 10^{308}$
Minimum floating point number[a]	$2^{-126} = 1.755 \times 10^{-38}$	$2^{-1022} = 2.225 \times 10^{-308}$
Unit roundoff (u)	$2^{-24} = 5.96 \times 10^{-8}$	$2^{-53} = 1.110 \times 10^{-16}$
Machine epsilon	$2^{-23} = 1.192 \times 10^{-7}$	$2^{-52} = 2.220 \times 10^{-16}$

[a]This is the smallest number that is stored to full precision.

∞, and $\sqrt{-1}$, all lead to NaN (not a number). Graceful underflow ensures that, whenever possible, leading zeros in the fraction (= *mantissa*) supplement the exponent. Otherwise underflow gives a result of zero.

There are, in addition, IEEE single- and double-extended number formats. The double-extended format must ensure a relative error of no more than 5.42×10^{-20} for numbers in a range of at least $10^{\pm 4932}$. A double-extended IEEE format is supported by the Intel 8087 chip and its successors through to the Pentium and beyond, and by the Motorola 68000 series chips used in Macintosh computers. There are as yet few software systems that take advantage of the double-extended format.

The default rounding mode for IEEE arithmetic is rounding to the nearest representable number, so that unit roundoff is $u = 0.5 \times \beta^{1-t}$, i.e. 5.96×10^{-8} for single-precision and 1.11×10^{-16} for double-precision arithmetic. The standard requires that all arithmetic operations are to be performed as if they were first calculated to infinite precision and then *rounded* to the specified precision. For comments on how this is achieved, see [2] and [3].

The importance of the IEEE standard is that results are accurate and predictable, providing a sound basis on which to build reliable numerical software. Even where IEEE arithmetic is available, compilers do not necessarily allow access to all features. For a detailed discussion of IEEE arithmetic, and comments on the aberrant arithmetics still found on some computers, see [3]. There is a public domain computer program, due in the first place to W. Kahan, which checks out in detail the arithmetic of the computer on which it runs. BASIC, C, FORTRAN, Modula and Pascal versions are available; go to `http://netlib.bell-labs.com/netlib` and search for *paranoia*. In **S-PLUS** the list `.Machine` holds settings of a number of machine arithmetic parameters.

Scaling to Avoid Overflow

Computations can often be reorganized so that the risk of overflow becomes, instead, the possibility of a harmless form of underflow. Consider $a = 10^{20}$, $b = 1$, and assume single-precision IEEE arithmetic. Then evaluation of $a^2 + b^2$ as a first step in the calculation of $r = (a^2 + b^2)^{1/2}$ will lead to overflow. Thus r will be set to $(\infty)^{1/2} = \infty$, with implications for all subsequent calculations that involve r. This is avoided by setting $d = \max\{|a|, |b|\}$ and calculating

$$r = d\left[\left(\frac{a}{d}\right)^2 + \left(\frac{b}{d}\right)^2\right]^{1/2} = 10^{20}(1 + 10^{-40})^{1/2}.$$

Underflow in the attempt to calculate 10^{-40}, so that the result is set to zero, is harmless.

Extreme Fitted Proportions in Logistic and Related Models

A general principle is that one should avoid, or seek a detour around, the computation of intermediate quantities that are incapable of accurate representation. Consider the following logit model for binomial data (*see* **Logistic Regression**):

$$\log \frac{\pi_i}{1 - \pi_i} = \eta_i = \mathbf{x}_i' \boldsymbol{\beta}, \tag{3}$$

where π is the expected value of a binomial proportion p, and $\mathbf{x}_i'$ is the ith row of the model matrix $\mathbf{X}$. This is a Nelder & Wedderburn style **Generalized Linear Model** [7], with logit *link* and binomial *error*.

We consider implications from the finite precision of floating point arithmetic when one or more estimated π_i is very close to one. (Almost inevitably the observed p_i is then one.) Dose–mortality studies with insects or other pests or pathogens, where the highest dose may be designed to achieve a very high

level or mortality, provide examples. Note in passing that when $n_i\pi_i \approx n_i$ for one or more i, the asymptotic theory for the distribution of both deviance and Pearson **chi-square distribution** breaks down.

An iteratively reweighted least squares *scoring* algorithm [7, pp. 40–43] fits this model by solving, successively for $k = 1, 2, \ldots$, the following regression equation:

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta}^{(k)} = \mathbf{X}'\mathbf{W}\mathbf{u}^{(k)}. \tag{4}$$

Here $\mathbf{W}$ has diagonal elements

$$w_i = (\mathrm{var}[p_i])^{-1}\left(\frac{\mathrm{d}\pi_i}{\mathrm{d}\eta_i}\right)^2 = \mathrm{diag}[n_i\pi_i(1-\pi_i)],$$

and $\mathbf{u}^{(k)}$ has elements

$$\begin{aligned} u_i^{(k)} &= \eta_i^{(k-1)} + \frac{\mathrm{d}\eta_i}{\mathrm{d}\pi_i}(p_i - \pi_i^{(k-1)}) \\ &= \eta_i^{(k-1)} + \frac{p_i - \pi_i}{\pi_i(1-\pi_i)}. \end{aligned}$$

Expressions involving η_i and π_i are evaluated at their current estimates $\eta_i^{(k-1)}$ and $\pi_i^{(k-1)}$.

If $\exp -\eta_i$ is less than u = unit roundoff, then

$$\pi_i = \frac{\exp\eta_i}{1+\exp\eta_i} \tag{5}$$

will be calculated as 1. In fact, to retain around d significant decimal digits, $\exp -\eta_i$ must be at least $10^d u$. It is important to calculate $1-\pi_i$, not by subtracting π_i from 1, but as

$$1 - \pi_i = \frac{1}{1+\exp\eta_i}. \tag{6}$$

This allows accurate calculation of $\pi_i(1-\pi_i)$ provided that overflow does not occur in the evaluation of $\exp\eta_i$. This has the much less restrictive requirement that $1-\pi_i > 10^{-308}$ approximately, in order to calculate the result to d significant digits in IEEE double precision arithmetic. A further refinement is to rewrite (5) and (6) in terms of $\exp -(\eta_i/2)$ and $\exp(\eta_i/2)$.

If in place of the model (3), one has the complementary log–log model (*see* **Quantal Response Models**)

$$\log[-\log(1-\pi_i)] = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{7}$$

then in (4)

$$\begin{aligned} \mathbf{W} &= \mathrm{diag}\{\pi_i^{-1}(1-\pi_i)[\log(1-\pi_i)]^2\}, \\ u_i^{(k)} &= \eta_i^{(k-1)} - \frac{p_i - \pi_i^{(k-1)}}{(1-\pi_i)\log(1-\pi_i)}. \end{aligned}$$

Here $1-\pi_i$ should be evaluated, using the value of η_i from the previous iterate, as $\exp(-\mathrm{e}^{\eta_i})$.

Even these steps may not be adequate. For the model (7), consider the data $x = 0, 2, 3, 4, 6, 12$; $p = 0.450, 0.837, 0.977, 0.998, 1.0, 1.0$; with $n_i = 1000$, $i = 1, 2, \ldots, 6$. Then $\hat{\eta}_i = 0.5420 + 0.5974x$. For $x_6 = 12$, $\hat{\eta}_6 = 6.627$, so that in IEEE double-precision arithmetic $\exp(-\mathrm{e}^{\eta_6})$ underflows to 0. The crossover is at $\log[-\log(2.25 \times 10^{-308})] \approx 6.563$. Once the current estimate of η_6 exceeds this, it will be necessary to set w_6 and u_6 to their limiting values as $\pi_6 \to 1$, i.e. $w_6 = 0$ and $u_6 = \eta_6$. Similar issues arise when $\hat{\pi}_i \simeq 0$ for one or more i. The aim is to ensure that small and mostly meaningless differences of any $\hat{\pi}_i$ from one or from zero do not prevent calculations from running to completion.

Rounding Errors and Error Analysis

Rounding errors are likely to occur in any extended sequence of calculations that involve floating point numbers. This happens even with carefully designed **algorithms**. A key result for IEEE arithmetic is that

$$\mathrm{fl}(x \text{ op } y) = (x \text{ op } y)(1+\delta), \quad |\delta| \le u$$

where u is the unit roundoff, op is any of the four arithmetic operations $+, -, \times, \div$ [3, p.44], and $\mathrm{fl}(x \text{ op } y)$ denotes the result of the floating point computation of x op y. This forms the basis for error analysis for IEEE arithmetic. It may seem surprising that the result applies to $x - y$ when x and y have the same sign. As happened in the example that followed (2) above, in the calculation of x and y any serious loss of accuracy occurs before the subtraction.

For the two-pass calculation of $S = \sum_{i=1}^{n}(x_i - \overline{x})^2$ that is based on (1), one can show [3, p.38] that, neglecting terms in u^2 or higher powers, the relative error in the calculated value of S is no more than $(n+3)u$, where u is the unit roundoff. This bound is unlikely to be attained in practice. A more realistic bound may be $c(n)^{1/2}u$, where a conservative choice for c might be 1.5. While the error may be of no consequence when n is small, it may become serious

for single-precision calculations with, for example, $n = 10^9$.

The bound $(n + 3)u$ on the error in the computed result is an example of a forward error analysis. It contrasts with backward error analysis which sets bounds on the perturbations needed in the input values so that an exact computation would yield the computed result. For further examples of such analyses, see [3, 8] and [9].

Condition Numbers and Ill-conditioning

Ill-conditioning, i.e. a large condition number, implies that changes that are of the order of the relative precision of representation of x will lead to large relative perturbations in y, making a highly accurate result impossible.

Let $y = f(x)$ be the result of an exact calculation with input value x. A *condition number* relates changes in the elements of y, measured in a manner that is convenient or appropriate, to changes in the elements of x. The calculation of $y = \log x$ provides a simple example. The approximate change, Δy, which results from a small change Δx in x, is $\Delta y \approx f'(x)\Delta x$. Thus

$$\frac{\Delta y}{y} \approx \frac{\Delta x}{x}\frac{1}{\log x}.$$

Any change in x is magnified by a condition number $1/\log x$, which becomes large when x is close to 1. Small changes in x will then lead to large relative perturbations in y, making a highly accurate result impossible. A stable algorithm carries out a computation to the precision allowed by the condition number. It will give an accurate answer to a well-conditioned problem, and do as well as can be expected for an ill-conditioned problem.

Observe that the definition of a condition number is entirely a matter of algebra. An approximate estimate of the relative change in y that results from a relative change $\Delta x/x$ is usually adequate. These ideas have wide application to error analysis in **matrix computations, optimization, numerical integration**, and the numerical solution of differential equations (*see* **Numerical Analysis**).

Condition numbers for vectors or matrices are usually expressed in terms of vector or matrix norms. These are further discussed in the article on **Matrix Computations**.

Further Reading

Good general references for this article are [2, 3, 8–10].

References

[1] Chan, T.F., Golub, G.H. & LeVeque, R.J. (1983). Algorithms for computing the sample variance: analysis and recommendations, *American Statistician* **371**, 242–247.

[2] Goldberg, D. (1991). What every computer scientist should know about floating point arithmetic, *ACM Computing Surveys* **23**, 5–48.

[3] Higham, N.J. (1996). *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia.

[4] IEEE (1985). *IEEE Standard for Binary Floating Point Arithmetic, ANSI/IEEE Standard 754–1985*. Institute of Electrical and Electronic Engineers, New York. Reprinted in *SIGPLAN Notices* **22** 9–25.

[5] IEEE (1987). *A Radix-Independent Standard for Floating Point Arithmetic, ANSI/IEEE Standard 854–1987*. Institute of Electrical and Electronic Engineers, New York.

[6] Maindonald, J.H. 1984. *Statistical Computation*. Wiley, New York.

[7] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[8] Stewart, G.W. (1996). *Afternotes on Numerical Analysis*. SIAM, Philadelphia.

[9] Stoer, J. & Bulirsch, R. (1992). *Introduction to Numerical Analysis*, 2nd Ed. Springer-Verlag, New York.

[10] Thisted, R.A. 1988. *Elements of Statistical Computation. Numerical Computation*. Chapman & Hall, New York.

[11] Wallis, P.J.L., ed. (1990). *Improving Floating Point Programming*. Wiley, New York.

JOHN H. MAINDONALD

Follow-up, Active Versus Passive

Disease registers are records of all cases of a disease that have occurred in a known population, and follow-up is the essential process whereby they are kept up-to-date, the most important item being vital status (whether alive or dead).

Cancer registries are the most well-established disease registers, and for them, recording vital status is essential. All registries do this, the means being either active or passive follow-up. The most commonly used is the passive method where registries depend on the national death registration system (*see* **Death Certification**) to inform them about the death of any cancer patient resident in the area covered by the registry. The alternative is active follow-up, where, in addition to passive notification of death, the registry reviews the vital status of all registered patients not known to be dead at regular intervals, say five years, until death.

If the death and cancer registration systems were both perfect, the results of active or passive follow-up should be the same. However, the number of deaths to be matched with cancer registrations is huge. In England and Wales, for example, the list is reduced by notifying cancer registries only of (i) those deaths of residents in their area in which cancer appeared on the death certificate, and (ii) deaths of patients already known to be on a cancer register. Deaths may be missed for a variety of reasons; patient migration after diagnosis, clerical error and computer matching failure, for example. This results in an optimistic view of survival and "immortal" patients who only come to light when they apparently achieve an impossible age.

Active follow-up, if done properly and patients can be traced, should avoid these problems. In addition to passive follow-up, clerks find out from hospitals or from the patient's own doctor or from a central register if one exists, when they were last known to be alive or when they died.

Use of an active follow-up system produces estimates of survival that depend wholly on patients whose vital status is known; after the date they were last known to be alive no assumptions are made about their status and they do not contribute to the estimate of survival (the process of censoring; *see* **Censored Data**). While this method is **unbiased**, there is a loss of precision in proportion to the number of patients lost to follow-up. Estimates of survival at a time after completion of follow-up mean that all living patients will be censored, but since passive notification of death continues, a very pessimistic estimate of survival, depending only on the patients known to have died at the time of the estimate, will result. To avoid this, survival estimates must refer to a time at or before the time of follow-up; if a later date is needed, then the assumptions and methods of passive follow-up must be used.

Active follow-up is more expensive and gives a less **biased** and more truthful estimate of survival, but whether it is worth doing depends on the loss of accuracy and timeliness resulting from a dependence on passive follow-up, and how important this is thought to be.

T. DAVIES & S. GODWARD

Food and Drug Administration (FDA)

The Food and Drug Administration (FDA) is a federal regulatory agency with oversight of a wide range of consumer products: food; medical treatments, devices, and diagnostics (human and veterinary); vaccines; blood and blood products; cosmetics; and other related products. Approximately one-quarter of consumer dollars are spent on products regulated by the FDA. The agency headquarters are in Rockville, Maryland, but field offices are located across the United States and its territories, and an important research unit of the FDA, the National Center for Toxicological Research, is located in Arkansas. More than 9000 individuals are employed by the FDA, including many laboratory and clinical scientists, statisticians, epidemiologists, computer scientists, engineers, lawyers, and others. The FDA is part of the Department of Health and Human Services in the Executive Branch of the federal government.

Although some aspects of federal oversight of health issues can be traced back to the nineteenth century, the first major step in regulating the use of medical products was the passage of the Biologics Control Act of 1902. This legislation, resulting from the deaths of 13 children who received a diphtheria antitoxin that had been contaminated with tetanus, provided for federal oversight of the production of vaccines, serums, toxins, antitoxins, and related products. A few years later, in 1906, the Federal Food and Drugs Act prohibited the sale of adulterated and/or misbranded food and drugs. Additional legislation further clarified and extended regulatory authorities during the early part of the twentieth century, but the next major step in protecting consumers from unsafe products was the passage of the Federal Food, Drug and Cosmetic Act in 1938. As with the Biologics Control Act, the FD&C act was motivated by tragedy; 107 deaths resulted from an elixir that had been manufactured using a toxic substitute for the alcohol that was a routine component of all "elixirs". This Act greatly expanded the regulatory authority of the growing agency, which had been officially designated as the Food and Drug Administration in 1930. It added cosmetics and **medical devices** to the scope of regulated products, authorized factory inspections, required drugs to be labeled for safe use, enhanced enforcement capabilities, and, perhaps most notably, established a drug approval process so that manufacturers would now have to demonstrate a drug was safe before it could be marketed. In 1962, the Act was significantly amended. New provisions included the requirement that drugs be shown to be effective as well as safe before marketing approval could be granted (*see* **Drug Approval and Regulation**). Other additions included the requirement that patients involved in **clinical trials** provide informed consent, the shifting of authority over drug advertising from the Federal Trade Commission to the FDA, and the establishment of good manufacturing practices for the drug industry.

Changes in FDA structure and function continued into the latter part of the twentieth century. In 1971, FDA assumed responsibility for regulating products emitting radiation. In 1972, the regulation of biological products, which had been a function of the National Institutes of Health, was transferred to the FDA, and in 1976, the Medical Device Amendments provided increased and more specific authority for the regulation of medical devices. The 1992 institution of user fees – fees paid by pharmaceutical companies at the time a marketing application was submitted – resulted in an increased workforce in the areas of drugs and biologics and consequent acceleration of product review. User fees were extended to medical devices in 2003.

FDA consists of six centers, each focused on a particular area of FDA responsibility, plus the field offices (Office of Regulatory Affairs) that handle special investigations, consumer complaints, and inspections. The basic structure is shown in Figure 1. The National Center for Toxicological Research (NCTR) is a basic research facility; the other five centers are focused on particular areas within FDA's jurisdiction: drugs, biologics, medical devices/radiology, foods, and veterinary products.

Center for Biologics Evaluation and Research (CBER)

CBER has responsibility for ensuring the safety, purity, potency, and efficacy of biological products used to treat, prevent, or diagnose disease and for ensuring the safety of the nation's blood supply. (Biological products are substances derived from living organisms or produced by biotechnological processes.) This Center regulates vaccines, allergenic

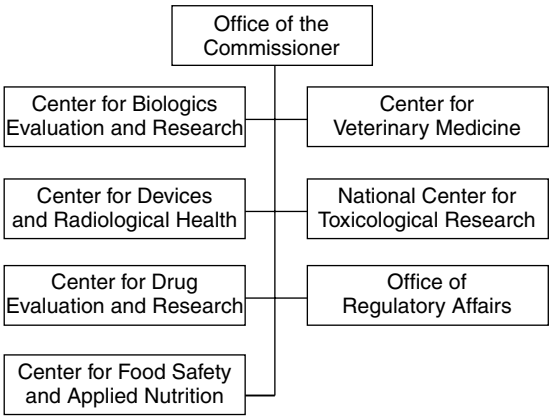

Figure 1 The structure of FDA

extracts, blood and blood products, test kits for donor blood, cellular therapies, tissues, and gene therapies. Biological therapeutics – cytokines such as interferons and interleukins, monoclonal antibodies, and other therapeutic proteins – have also been regulated by CBER, but in 2003, authority for these products was transferred to the Center for Drug Evaluation and Research.

Center for Devices and Radiological Health (CDRH)

CDRH regulates medical devices, ranging from the simplest items such as tongue depressors to pacemakers and cardiac stents, diagnostic imaging devices such as X-ray and mammography machines, computerized axial tomography (CAT) scanners and magnetic resonance imaging (MRI) devices, contact lenses, and computer software used in the diagnosis and treatment of disease. Also regulated in this Center are products that emit radiation that could potentially impact on health, for example, cellular phones, microwave ovens, lasers, and sun lamps.

Center for Drug Evaluation and Research (CDER)

CDER is responsible for drug products, including over-the-counter (nonprescription) drugs and generic drugs. The Center is organized according to disease areas, so that different divisions address **oncology** drugs, cardiovascular and renal drugs (*see* **Cardiology and Cardiovascular Disease**), anti-infectives, antivirals, anti-inflammatory drugs, neuropharmacological drugs, and so on. CDER regulates the vast majority of new medications, including over-the-counter and generic drugs, and plays the lead role within the FDA in regard to policy development relating to drug development and evaluation.

Center for Food Safety and Applied Nutrition (CFSAN)

CFSAN oversees safety of domestic and imported food, with the exception of meat, poultry, and eggs, which are regulated by the Department of Agriculture. Its regulatory oversight also extends to infant formula, dietary supplements, food additives, and cosmetics. CFSAN is directly responsible for safety aspects of food manufacturing, processing, and storage during the distribution process, including development of nutritional labeling, and for developing the guidance and model standards used by states for oversight of restaurants, grocery stores, and other food outlets.

Center for Veterinary Medicine (CVM)

CVM is responsible for ensuring the safety and effectiveness of drugs and food additives (including those derived from biotechnology techniques) approved for use in animals, including companion animals (dogs, cats, and horses), food animals (poultry, cattle, swine), and other species such as honeybees, wildlife, and zoo animals. When drugs are approved for use in food animals, the determination of safety also includes a demonstration that any food derived from these animals treated according to label directions is safe to consume.

National Center for Toxicological Research (NCTR)

NCTR studies issues related to FDA's regulatory needs, particularly with regard to understanding the potential toxicity of FDA-regulated products. Scientists at NCTR collaborate with scientists in other FDA centers as well as in other parts of the government and in academic institutions to better understand the biology underlying safety issues affecting

foods and drugs and to develop methods to improve **risk assessment** in a variety of areas such as food safety and human susceptibility to adverse effects of pharmaceuticals.

Office of Regulatory Affairs (ORA)

ORA is the enforcement arm of the FDA. Most ORA staff work in more than 150 field offices across the United States and in Puerto Rico. To ensure that FDA-regulated products intended for use in the United States meet required standards and that clinical studies of such products are conducted appropriately, ORA staff inspect manufacturing and warehouse facilities in the United States as well as abroad, inspect clinical research sites, and test samples in ORA laboratories. In cases of suspected criminal misconduct relating to these products, ORA takes the lead in determining appropriate legal actions.

To accomplish their respective missions, each center has many components. These may include units focusing on clinical science, laboratory science, statistics, epidemiology, compliance and enforcement, policy development, press and legislative liaison, legal issues, administrative issues (personnel, budget, etc.), staff training, information technology, international activities, and other areas. The Office of the Commissioner coordinates many initiatives in these areas, particularly when more than one center is involved, and oversees all center activities.

The Regulatory Process

The FDA derives its authority from legislation, going back to the FD&C and Biologics Control Acts; amendments to these acts; and additional legislation designed to clarify regulatory authority in specific areas. Regulations specifying required standards for all aspects of product development are developed on the basis of this legislation and are published in the *Federal Register* when issued. Generally, regulations are issued in draft form, with opportunity for public comment to be submitted and considered before a final version of the regulation is issued. Volume 21 of the *Code of Federal Regulations,* published each year by the Government Printing Office, includes all current regulations pertaining to FDA authorities.

In addition to regulations, the FDA issues many guidance documents. These documents are developed to clarify the intent of the regulations and/or to present methodological approaches that are recommended (but not necessarily required) by the agency in specific contexts. Guidance documents are also published in the *Federal Register* and are also usually revised on the basis of public comment, but no compendium of current guidance documents is published yearly, as is done for regulations. All operative regulations and guidance documents are available for reading or printing from the website of the FDA center(s) issuing the document.

Statistics at the FDA

Each FDA component except for the Office of Regulatory Affairs and the Office of the Commissioner has its own statistical group whose work focuses on the particular responsibilities of the center in which it resides. For example, the statisticians in the Center for Food Safety and Applied Nutrition deal substantially with survey and other observational data and risk assessments related to possible food contaminants (*see* **Nutritional Epidemiology**); statisticians at the National Center for Toxicological Research work in the area of **bioassay**; while the statisticians in the centers that evaluate investigational drugs, biologics, and medical devices review data from clinical trials, testing these new products, as well as the study designs that are proposed for such tests. Statistical aspects of postmarketing safety **surveillance**, and risk assessment more generally, are of increasing interest in all centers. FDA statisticians are also often involved in the development of new regulations and guidance documents (*see* **Guidelines On Statistical Methods in Clinical Trials**) that focus on statistical issues or for which statistical considerations are important and frequently publish on methodological issues arising from their review work.

In 2003, there were more than 150 statisticians at the FDA. All FDA statisticians are automatically granted membership in the Food and Drug Administration Statistical Association (FDASA), established in 1995 to provide a focus for cross-center statistical activities. The FDASA cosponsors an annual meeting with the Biopharmaceutical Section of the **American Statistical Association** and plans a yearlong seminar program for FDA statisticians.

In addition to statisticians employed by the FDA, a large number of statisticians from academic institutions and other organizations contribute statistical expertise to FDA decision making, primarily by serving on FDA Advisory Committees. These committees are established by each center to provide expert advice and comment on product submissions, potential risks of regulated products, and other issues. Most Advisory Committees include at least one statistician. Statisticians external to the FDA may also be asked to assist with the review of submissions or development of policy documents that include discussion of statistical issues.

Relevant Websites

`www.fda.gov`: Main FDA website; includes links to all Center websites and general information about the FDA.

`www.gpo.gov/nara/cfr/index.html`: Code of Federal Regulations

`www.gpo.gov/su_docs/aces/aces140.html`: Federal Register

SUSAN ELLENBERG

Forecasting

Many data sets in biostatistics arise naturally as **time series**, meaning a set of data collected sequentially through time. Examples include (i) an electrocardiogram trace (ECG), (ii) the number of cases of measles in successive months in a particular country, and (iii) the size of an insect colony on successive days. The analysis of time series data poses special problems because successive observations are usually not independent but are correlated through time. This phenomenon is called *autocorrelation* (*see* **Autocorrelation Function**).

The first step in any time series analysis or forecasting activity is to plot the data against time in a graph called a *time plot*. This sounds a simple task, but, in fact, it may not be easy to choose appropriate scales to present the data. Figure 1 shows two plots of a famous time series called the lynx data. Figure 1(b) has a more compressed vertical scale and may initially seem a less natural way to present the data. However, it does, in fact, allow the viewer to see that the series typically falls faster than it rises. This nonlinear feature cannot be seen in Figure 1(a). An alternative possibility is to plot a transformation of the data such as logarithms. More generally, attention to detail is needed to label the scales carefully, to give the graph a clear self-explanatory title, and so on. Only then will it be possible to interpret the graph safely. Indeed, a good time plot should enable the analyst to get a good idea of the properties of

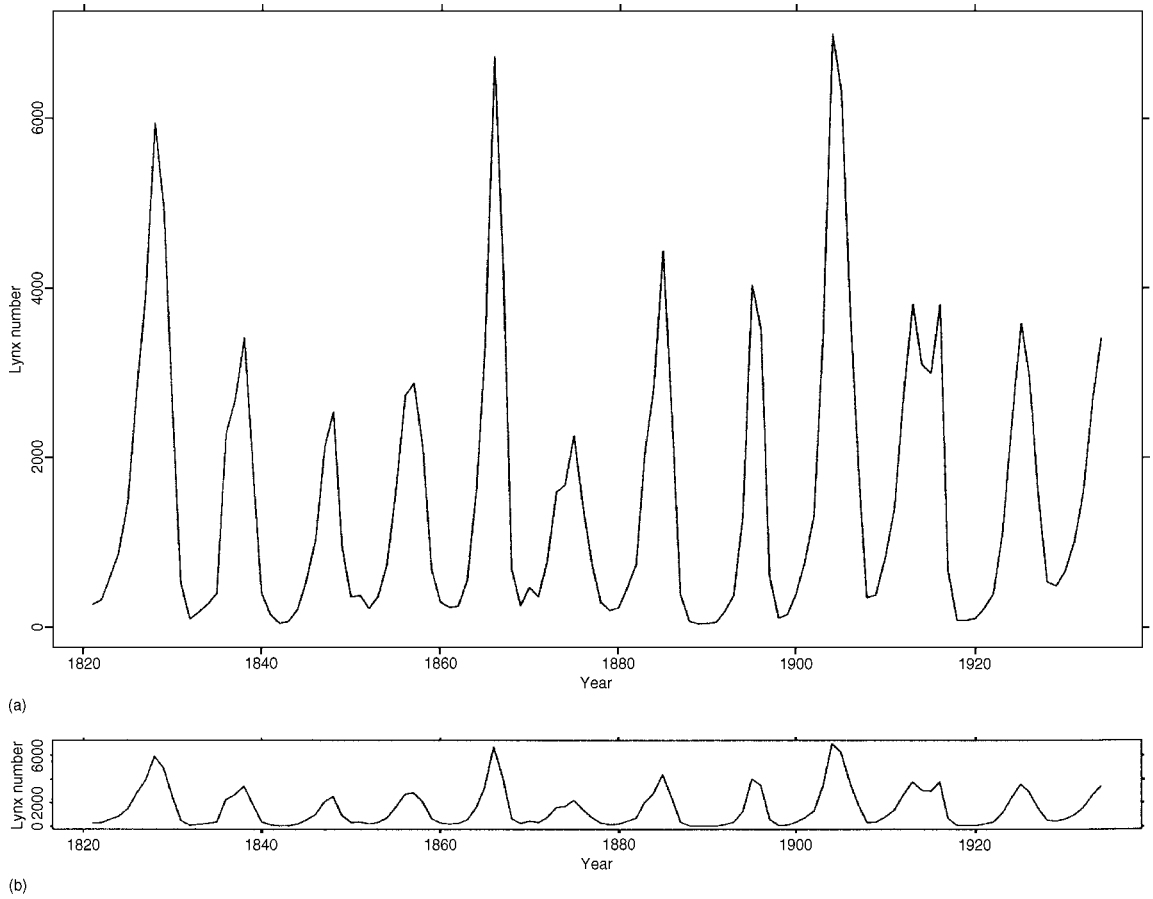

Figure 1 Two time plots of the annual numbers of Canadian lynx trapped in the Mackenzie River district of North-West Canada over the period 1821–1934. Part (b) has a compressed vertical axis that enables the rise and fall of the graph to be more easily assessed. Source: Records of the Hudson Bay Company

the time series, as well as identifying any **outliers** or discontinuities.

Finding an appropriate time series model for a given time series is a complex process that involves looking at a time plot of the data and also at the autocorrelations of the series at different lags. The reader is referred to Diggle [4] for a biostatistical introduction and to Chatfield [3] and Wei [6] for more general introductions. These three books also include introductions to the special time series activity of forecasting, which is the topic of this article (*see* **Spectral Analysis**; **Variogram**).

Denote an observed time series by $x_1, x_2, \ldots, x_n$ and suppose we wish to predict a value at some time in the future, say, x_{n+k}. Here, the integer k is called the *lead time* and the forecast of x_{n+k} made at time n is denoted by $\hat{x}_n(k)$. Most authors use the terms "forecast" and "prediction" interchangeably.

Many different forecasting methods are available and this article can only give a flavor of them. The choice of method depends partly on *clarifying the objectives* of a particular study and finding out exactly how a forecast will be used. The choice also depends on the skill of the analyst; for example, in deciding whether to use a univariate projection forecast, which only uses past values of the given variable, or a multivariate forecast, which incorporates the effects of one or more explanatory variables. This article restricts attention to univariate forecasts, apart from noting that multiple regression can be used to produce multivariate forecasts, but has difficulty in coping with correlated errors and can give spuriously high values of R^2.

Some methods are specifically designed to cope with two sources of variation called *trend* and *seasonality* (*see* **Seasonal Time Series**). Trend may loosely be described as a long-term change in the underlying mean level, while seasonality describes cyclic variation (*see* **Circadian Variation**) that might take place over a period of one day (diurnal variation), one week or one year. For example, the number of (human) deaths is typically higher in winter than summer. Figure 1 shows little evidence of trend in the number of lynx caught, but shows clear evidence of cyclic variation with a period of between nine and 10 years. However, as the period does not appear to be fixed, the variation would generally be described as cyclic rather than seasonal, and is more difficult to model and forecast.

An obvious procedure for forecasting a time series that shows an approximately linear trend is to fit a simple linear regression on time; namely, $x_t = \beta_0 + \beta_1 t + \varepsilon_t$, and substitute the required future value of t – see, for example, [4, Section 7.1]. This procedure emphasizes that forecasting is **extrapolation** in that it involves using the model outside the range of data to which it has been fitted. If the trend changes, forecasts will have poor accuracy. In any case, the trend is unlikely to be *exactly* linear, and a more modern approach is to assume *local linearity* and update forecasts by a procedure such as Kalman filtering and smoothing. Details will not be given here, but we introduce a very simple example of a Kalman filter called *exponential smoothing*, which is more typical of modern time series forecasting methods than linear regression on time.

A general form of linear forecast is to take a linear combination of the observed values; namely,

$$\hat{x}_n(1) = \sum_{j=0}^{n-1} w_j x_{n-j}, \tag{1}$$

where the weights $\{w_j\}$ need to be determined. It is natural to give more weight to more recent observations and to choose a set of weights that sum to unity. One suitable set of weights is given by

$$w_j = \alpha(1-\alpha)^j, \tag{2}$$

where α is a constant such that $0 < \alpha < 1$. These weights do sum to unity as $n \to \infty$ and decay geometrically. However, the real virtue of choosing geometric weights is that (1) incorporating (2) can be rewritten in the recursive form

$$\hat{x}_n(1) = \alpha x_n + (1-\alpha)\hat{x}_{n-1}(1). \tag{3}$$

Thus, as each new observation becomes available, forecasts can readily be updated. The smoothing parameter, α, can be chosen so as to optimize the one-step-ahead forecasts for the data we already have. A higher value of α gives more weight to more recent observations.

Exponential smoothing can readily be generalized to cope with a linear trend (called Holt's linear trend method) and also with seasonality (when it is generally called the Holt–Winters method).

Another class of models that is often used to produce forecasts is that of *autoregressive* (AR) models. (*see* **ARMA and ARIMA Models**). A time series,

X_t, is said to follow an autoregressive process of order p, denoted AR(p), if X_t is a linear function of the preceding p values in the time series, together with an "error" term, ε_t, that is usually assumed to be from an independent sequence of $N(0, \sigma^2)$ variables. Thus, an AR(1) process may be written

$$X_t = \alpha X_{t-1} + \varepsilon_t, \tag{4}$$

where the parameter α has to be chosen such that $|\alpha| < 1$ in order for the process to be stationary. It turns out that AR models can generally be fitted by methods similar to those used for ordinary regression models, except that the explanatory variables are now lagged values of the given variable. Forecasts may be computed in an intuitively obvious way. For example, the AR(1) model gives a one-step-ahead forecast at time n equal to $\hat{x}_n(1) = \alpha x_n$. AR models can be generalized to include lagged values of the "error" terms (called *moving average* terms – a rather misleading terminology) and can also be applied to nonstationary data by suitably *differencing* the data. For example, first-order differencing gives the series $(x_2 - x_1), (x_3 - x_2), \ldots, (x_n - x_{n-1})$ and will remove a linear trend. The large class of models formed in the above way is called the *autoregressive integrated moving average* (ARIMA) class and forms the basis of the *Box–Jenkins* forecasting approach – see [1].

The final approach mentioned here is based on the *state-space* or **structural time series model** – see, for example, [5]. A simple example is given by the so-called *steady* model for which

$$X_t = \mu_t + n_t, \tag{5}$$

where μ_t denotes the unobservable current level and n_t denotes the error term. It is further assumed that μ_t is itself a random variable that evolves through time according to a random walk so that

$$\mu_t = \mu_{t-1} + w_t, \tag{6}$$

where w_t denotes a second error variable independent of n_t. The (unobservable) variable μ_t is called a state variable and the current estimate of it is provided by the updating procedure called the Kalman filter. Given an estimate of μ_t, forecasts of X_t may readily be computed.

Choosing the most appropriate forecasting method for a particular problem is not easy and the reader is referred to [3, Section 5.4].

This article has concentrated on the computation of *point* forecasts. In practice, it is often desirable to calculate an *interval* forecast. The term *prediction interval* is used to describe an interval within which a future value is expected to occur with a specified probability. Some forecasting methods lend themselves more easily than others to the computation of the relevant standard errors that enable prediction intervals to be calculated. However, the possible presence of model uncertainty (the model may change in the future or may have been wrongly identified) means that prediction intervals are typically too narrow. A review of methods for computing prediction intervals is given in [2].

References

[1] Box, G.E.P., Jenkins, G.M. & Reinsel, G.C. (1994). *Time Series Analysis*, 3rd Ed. Prentice Hall, Englewood Cliffs.

[2] Chatfield, C. (1993). Calculating interval forecasts (with discussion), *Journal of Business and Economic Statistics* **11**, 121–144.

[3] Chatfield, C. (1996). *The Analysis of Time Series*, 5th Ed. Chapman & Hall, London.

[4] Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

[5] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

[6] Wei, W.W.S. (1990). *Time Series Analysis*. Addison-Wesley, Redwood City.

(*See also* **Prediction**)

CHRIS CHATFIELD

Forensic Medicine

There are several problems in forensic medicine for which a statistical approach is particularly apt. These include, in particular, estimation of the post-mortem interval, or time since death, the estimation of the age at death, determination of sex from skeletal remains and, amongst the living, the estimation of the quantity of alcohol consumed, as well as issues of **paternity**. Simple summary statistics are used often; for example, in recording variations in relative frequencies amongst **genetic marker** systems in different populations. This issue is extended to a general discussion of problems of forensic identification, with particular reference to DNA profiling, by Dawid & Mortera [2]. Many standard statistical techniques, such as **regression**, are used. Only recently have other techniques, e.g. kernel **density estimation** and **Bayesian methods** been suggested.

Post-mortem Interval

Accurate estimation of the post-mortem interval (PMI) is of obvious importance in the resolution of an investigation involving a corpse. The most common approach is to study factors, such as the temperature of various parts of the body, which vary with PMI, and to determine a suitable relationship between these factors and temperature. In 1962 Marshall & Hoare [9] published the following formula modeling rectal body temperature:

$$\frac{T_{\mathrm{r}} - T_{\mathrm{a}}}{T_0 - T_{\mathrm{a}}} = A\exp(Bt) + (1 - A)\exp\left(\frac{AB}{A-1} \times t\right), \tag{1}$$

where T_{r} denotes rectal temperature at any time, T_{a} denotes ambient temperature, T_0 denotes rectal temperature at death $(t = 0)$, A is a constant that expresses the relative duration of a post-mortem temperature plateau phase, B is a constant that describes the cooling rate for as long as there is a difference between the ambient temperature and that of the body, and t is the time of death. Note, however, that it is a mathematical formula. While no attempt appears to have been made to model the errors implicit in the estimation of the parameters, there have been many empirical studies to determine the magnitude of the errors. Correction factors have been introduced to allow for different environmental factors, for example. Sometimes nomograms are used that relate rectal temperature, ambient temperature and body weight to time since death. The Marshal–Hoare formula measures time since death in the early post-mortem period (i.e. in hours). For longer periods of time, measurement of post-mortem enzyme activity may be used [5].

Age at Death

Gustafson [6] determined age at death on the basis of a regression of adult human age on morphological changes of six characteristics in the structure of teeth. This was based on applying normal **linear regression** techniques to ordinal and categorical data. Gustafson claimed an error of about three to four years, though later estimates of about seven years or even 16 years have been determined. Various Bayesian approaches that account for the data structure and provide results with mean absolute deviations of four to six years are advocated by Lucy et al. [7] and kernel density methods are described by Aykroyd et al. [1].

Sex Determination

Linear **discriminant analysis** is used to aid the determination of the sex of skeletal remains. The high accuracies of discrimination obtained have their basis in the unique form of sexual dimorphism exhibited by the adult human pelvis. One recent study [8] derived a score function, using discriminant analysis, from 122 adults of known sex and applied this to 230 other adults of known sex with 100% correct classification.

Blood Alcohol Measurements

The amount (A) of alcohol consumed based on the blood alcohol concentration (C_t) is calculated using Widmark's [12] formula:

$$A = r \times p \times [C_t + (\beta \times t)], \tag{2}$$

where r is the ratio of the total body ethanol concentration to the blood ethanol concentration, p is the body weight, and β is the ethanol elimination rate constant. Note that r varies between males and females. Various empirical studies have investigated the relationship between predicted and actual

concentrations. The formula is also used for breath alcohol concentration by the substitution of its value for C_t in (2). This introduces another source of error, generally leading to a reduction in the estimated amount of alcohol consumed [4].

Inverse Prediction

Notice that (1) gives an equation for determining rectal temperature from time since death and that (2) gives an equation for determining the amount of alcohol consumed from a blood alcohol concentration. In both cases, the inverse prediction is required. This has been discussed briefly by Aykroyd et al. [1], where age at death is regressed on a score determined from six dental indicators of Gustafson [6].

Paternity

In a paternity case, a male is alleged by the mother of a child to be the father of the child. The truth of the allegation can be partially tested by calculating a so-called "probability of nonpaternity" or "probability of exclusion" (Q, say) in a specific genetic system. The **genotypes** of the mother and child provide information about the true father in that males with certain **genes** are excluded from fatherhood of the child.

Consider a co-dominant system where all genotypes are detectable (in contrast to a dominant/recessive system in which only phenotypes are detectable). Let

$$p_1, p_2, \ldots, p_k, \quad \left(\sum_{i=1}^{k} p_i = 1.0\right)$$

represent the **gene frequencies** associated with a co-dominant system with k alleles, then

$$Q = \sum_{i=1}^{k}[p_i(1-p_i)]^2 + \sum_{j=1}^{k-1}\sum_{i=j+1}^{k} p_i p_j\{[1-p_i]^3 + [1-p_j]^3 + [p_i+p_j][1-(p_i+p_j)]^2\},$$

where the assumption is made that all individuals involved in the paternity case come from a large random mating population at equilibrium [10].

Consider now several loci and let Q_l be the probability of exclusion at locus l. The overall probability of exclusion (i.e. the probability the system will exclude a falsely accused male in a paternity action), Q, follows from being able to exclude the alleged father from at least one locus. Thus, if the loci are independent [11],

$$Q = 1 - \prod_{l}(1 - Q_l).$$

A related approach expresses the probability that the alleged father is the true father (F), given the evidence ($E_1, E_2, \ldots, E_n$) of n phenotypic systems, as follows:

$$\Pr(F|E_1, E_2, \ldots, E_n) = \left\{1 + \frac{\Pr(\overline{F})}{\Pr(F)}\prod_{i=1}^{n}\frac{\Pr(E_i|\overline{F})}{\Pr(E_i|F)}\right\}^{-1},$$

where $\overline{F}$ is the event that the alleged father is not the true father. A particular example of this approach with $\Pr(F) = \Pr(\overline{F})$ is described by Essen-Möller [3].

References

[1] Aykroyd, R.G., Lucy, D. & Pollard, A.M. (1996). Statistical methods for the estimation of human age at death, *Report No. STAT 96/08*. Department of Statistics, University of Leeds.

[2] Dawid, A.P. & Mortera, J. (1996). Coherent analysis of forensic identification, *Journal of the Royal Statistical Society, Series B* **58**, 425–443.

[3] Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis: Theoretische Grundlagen (The evidential value of similarity as proof of paternity, fundamental principles), *Mitteilungen der Anthropologischen Gesellschaft in Wein* **68**, 9–53.

[4] Friel, P.N., Logan, P.K. & Baer, J. (1995). An evaluation of the reliability of Widmark calculations based on breath alcohol measurements, *Journal of Forensic Sciences* **40**, 91–94.

[5] Gos, T. & Raszeja, S. (1993). Postmortem activity of lactate and malate dehydrogenase in human liver in relation to time after death, *International Journal of Legal Medicine* **106**, 25–29.

[6] Gustafson, G. (1950). Age determination on teeth, *Journal of the American Dental Association* **41**, 45–54.

[7] Lucy, D., Aykroyd, R.G., Pollard, A.M. & Solheim, T. (1996). A Bayesian approach to adult human age estimation from dental observations by Johanson's age changes, *Journal of Forensic Sciences* **41**, 189–194.

[8] Luo, Y.-C. (1995). Sex determination from the pubis by discriminant analysis, *Forensic Science International* **74**, 89–98.

[9] Marshall, T.K. & Hoare, F.E. (1962). Estimating the time of death. The rectal cooling after death and its mathematical expression, *Journal of Forensic Sciences* **7**, 56–81.

[10] Selvin, S. (1980). Probability of nonpaternity determined by multiple allele codominant systems, *American Journal of Human Genetics* **32**, 276–278.

[11] Weir, B.S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Sunderland.

[12] Widmark, E.M.P. (1932). *Principles and Applications of Medicolegal Alcohol Determination*. Biomedical Publications, Davis, 1981.

C.G.G. AITKEN

Forward Search

Introduction

The forward search is a powerful robust statistical method for exploring the relationship between data and fitted models. It is a development of the methods described in the articles on **Residuals** and **Diagnostics** that aids the discovery of clusters of observations and previously unidentified important subsets of the data as well as revealing any groups of outliers.

In this article, we give examples of the use of the forward search for **regression** and **generalized linear models**. These applications, together with the material on **transformations** of data covered in the article **Fan Plot** and the extension to **nonlinear regression**, are described by Atkinson and Riani [1]. We also give an example involving **multivariate** data, a topic extensively covered in [3].

Regression and Residuals

The forward search orders the observations by closeness to the assumed model, starting from a small subset of the data and increasing the number of observations m used for fitting the model. **Outliers** and small unidentified subsets of observations enter at the end of the search.

We write the multiple regression model as

$$ym = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters, and it is assumed that the additive errors of observation ε are independently distributed with constant variance σ^2. Also in (1), $\boldsymbol{X}$ is the $n \times p$ **matrix** of carriers, that is, of **explanatory variables** and perhaps functions of them, such as quadratics and interaction terms.

It is helpful to list the various stages of the forward search.

1. Notation. The vector of p parameters $\boldsymbol{\beta}$ is estimated by **least squares** applied to subsets of the observations. For an arbitrary subset of m observations, the estimate is denoted $\hat{\boldsymbol{\beta}}(m)$. For a subset $S^*(m)$ of size m chosen by the forward search, the estimate is written $\hat{\boldsymbol{\beta}}(m^*)$.

2. Starting the Search. The search starts from a small subset of size m_0; usually $m_0 = p$ or perhaps $p + 1$. To find the starting subset $S^*(m_0)$, we randomly select 1000 subsamples of size m_0. The initial subset $S^*(m_0)$ provides the least **median** of squares estimator $\hat{\boldsymbol{\beta}}(m_0^*)$, that is, it minimizes the median squared residual (Rousseeuw [5]) of the observations over the 1000 samples.

3. Moving Forward in the Search. When the m observations constituting $S^*(m)$ are used in fitting, the fitted values from the estimate $\hat{\boldsymbol{\beta}}(m^*)$ yield n least-squares residuals $\boldsymbol{e}(m^*)$. We order the squared residuals $\boldsymbol{e}^2(m^*)$ and take the observations corresponding to the $m + 1$ smallest as the new subset $S^*(m + 1)$. Usually, this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave. This may also happen at the beginning of the search, where $S^*(m_0)$ is chosen to minimize the median squared residual, not to find the subset yielding the $m_0 + 1$ smallest squared residuals. Because of this very robust starting point and the form of the search, outliers, if any, tend to enter as m approaches n.

4. Monitoring the Search. If any quantity is of interest when it is calculated for the complete set of n observations, we can monitor its evolution during the forward search. In our example, we first look at a forward plot of the residuals $\boldsymbol{e}(m^*)$, scaled by the final estimate of σ. Examples of forward plots of other quantities of interest, such as estimates of the parameters $\boldsymbol{\beta}$ and σ^2 are given by Atkinson and Riani.

The analysis of the data on mandible length [6] in the article on **Goodness of Fit** using simple regression shows appreciable evidence of nonnormality of the residuals. The normal plot of the least-squares residuals in Figure 3 of the article "Goodness of Fit" shows three large negative residuals and two further residuals that are also rather large.

This structure is apparent in the forward plot of the residuals in Figure 1. Units 165, 166, and 149 have large negative residuals throughout the search. Units 146 and 167 also have appreciable negative residuals for much of the search. Working backwards, the last units to join the search are these five, in order 165, 166, 149, 146, and 167. These are the five negative residuals visible in the $Q-Q$ plot in **Goodness of Fit**, which is of the unscaled version of the residuals at the end of the search in Figure 1. The forward search shows that, in this example, the residual plot,

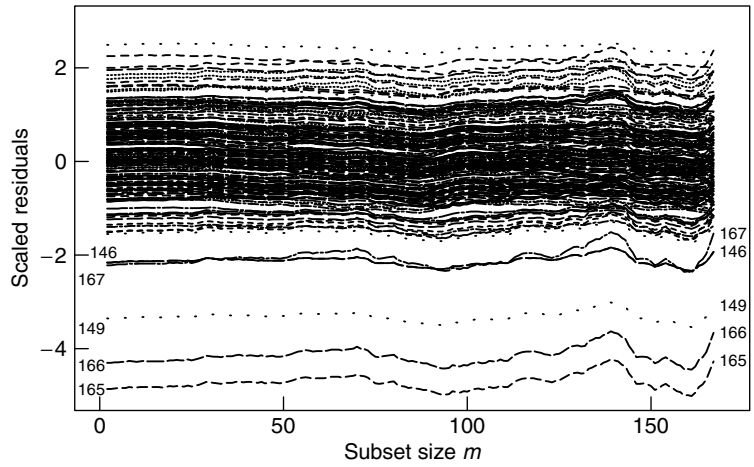

Figure 1 Mandible length data, first-order model: forward plot of scaled residuals. There are five large negative residuals for much of the search, but those for units 146 and 167 are masked at the end of the search

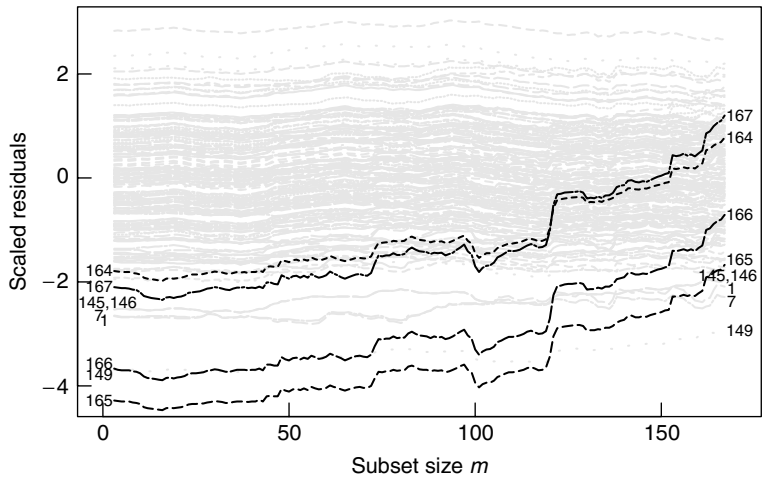

Figure 2 Mandible length data, second-order model, logged response: forward plot of scaled residuals. Four units, 164 to 167, behave differently from the rest, which have an approximately normal distribution

when all observations are fitted, identifies most of the structure of the residuals. The values of $\boldsymbol{x}$ and $\boldsymbol{y}$ for these units is clear from the scatter plot of Figure 5 of the article on **Diagnostics**. If the last three units, which are shown as open circles or crosses in the plot, are excluded, the straight line fitted to the data becomes such that unit 167 has an appreciable negative residual. As the forward plot shows, this residual is reduced when the last three units enter the subset. There is therefore some masking of the outlying nature of this unit.

The units with large residuals identified in this analysis are not all of those plotted with open circles in Figure 5 of the article **Diagnostics**. One reason is that these were identified as being influential observations, rather than having large residuals. A second reason is that that analysis was for a logged response with a second-order model. Figure 2 shows the forward plot of the residuals from this model.

Four units are highlighted in Figure 2. If we ignore them, the forward plot of the residuals is virtually symmetrical throughout the search, with no other

appreciable outliers. The most negative residuals are those for units 149, 7, 1, 145, and 146. But these values do not change much during the search and, as the $Q-Q$ plot in Figure 1 of the article on Diagnostics shows, these are not particularly extreme values when compared with **order statistics** from a normal distribution. The four highlighted units in the figure are units 164 to 167. They are highlighted because their behavior is very different. Initially, they all have large negative residuals, but by the end of the search, the residuals are all appreciably smaller, two having become positive. These units are those for the four oldest fetuses. It seems as if the model toward the end of the search may be being altered by their presence and so produces small residuals. Certainly, this would not be surprising as such extreme points in $\boldsymbol{X}$ space will be leverage points, a property amplified by fitting a quadratic model. Figure 2 of the article on **Diagnostics** shows how extreme these leverage values are. A question we then have to consider is how the evidence for a quadratic model depends on these four units.

Forward Added Variable T Test

If the fitted model and data agree, the parameter estimates should be reasonably constant throughout the forward search. These estimates are **orthogonal** to the residuals used to order the entry of units into the subset $S^*(m)$. The same is not true of the estimate of σ^2, which, being the sum of squared residuals, increases during the search as increasingly outlying observations are included in the subset. As a result, the t tests (*see* **Student's t Distribution**) for the parameters in the linear model decrease dramatically during the forward search. We describe here an alternative form of search that provides information on the inferential effect of the units on the estimated linear model.

If the standard regression model (1) is rewritten as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{Q\theta} + \boldsymbol{w}\gamma + \boldsymbol{\varepsilon}, \qquad (2)$$

$\boldsymbol{Q}$ is the $n \times p-1$ matrix of carriers obtained by deleting the column $\boldsymbol{w}$ from $\boldsymbol{X}$. At the end of the search, the t test for the column of $\boldsymbol{X}$ corresponding to $\boldsymbol{w}$ from **multiple regression** on $\boldsymbol{X}$ is identically the added variable test described immediately after equation (11) in the article on **Residuals**. This is found by first regressing $\boldsymbol{y}$ and $\boldsymbol{w}$ on $\boldsymbol{Q}$ and then testing the regression through the origin of the resulting residuals of $\boldsymbol{y}$ on those of $\boldsymbol{w}$.

We adapt the added variable test to the forward search by dropping each column of $\boldsymbol{X}$ in turn to create $p-1$ vectors $\boldsymbol{w}$. We then use regression on each $\boldsymbol{Q}$ to provide a forward search from which $\boldsymbol{w}$ is excluded. We monitor the behavior of the added variable test for each $\boldsymbol{w}$, thus obtaining $p-1$ plots of t statistics from $p-1$ different forward searches: $p-1$ because we are not usually interested in testing hypotheses about the value of the constant in the regression model. Because we exclude $\boldsymbol{w}$ from the search, the t test for $\boldsymbol{w}$ has the correct distribution and increases during the search rather than decreasing. The details are in [2].

We start, in Figure 3, with a forward plot of the added variable t test for regression of untransformed mandible length on **gestational age**. The plot shows a steady upward trend to a very significant value of 35.90. There is no sign of the importance of individual observations such as the units giving large residuals in Figure 1; evidence for the regression is spread throughout the data.

Figure 4, for regression of log $\boldsymbol{y}$ on a quadratic in age is similarly well behaved. The value of t_1, the t test for regression on age, rises steadily to 18.08, while that for t_2 for the quadratic term decreases to -12.55. The leverage points 164 to 167, which are such a notable feature of Figure 2, do not enter at the end of either of the added variable searches on which the plots in Figure 4 are based. The plot shows no evidence that these four units are responsible for the quadratic term in the model. Despite the appearance of Figure 2, the evidence of curvature

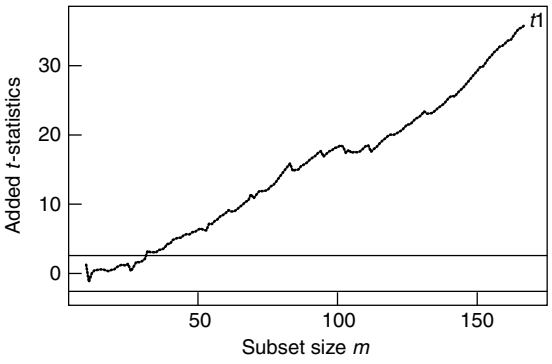

Figure 3 Mandible length data, first-order model: forward plot of added variable t test t_1 for regression on age. Evidence for the regression is spread throughout the data

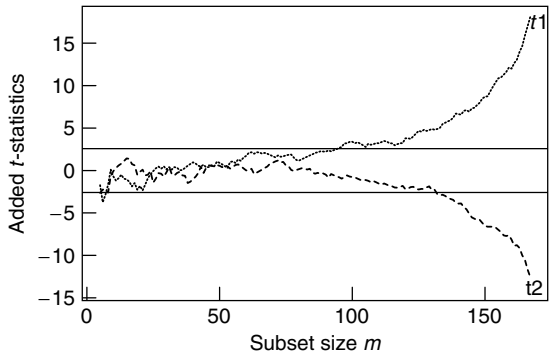

Figure 4 Mandible length data, second-order model, logged response: forward plot of added variable t tests t_1 and t_2 for regression on age and its square. Evidence for the regression is again spread throughout the data

in the relationship with a logged response is spread throughout the data.

Our analysis thus shows that taking a logged response combined with a quadratic model produces residuals, which have an approximately normal distribution, with four leverage points, the residuals for which change appreciably during the search. These four units are not influential for the choice of terms in the linear model. However, they might be influential for the choice of the transformation. But the forward plot of the test for transformation in Figure 5 of the article on the **Fan Plot** shows that this is not the case. Thus, these procedures provide no evidence for the suggestion mentioned by Royston and Altman that the fetuses with an age greater than 28 weeks were different from the younger ones.

Generalized Linear Models

The structure provided by the theory of generalized linear models allows us to apply the forward search to, particularly, **gamma**, **Poisson**, and **binomial** data in a manner analogous to that used for multiple linear regression. Chapter 6 of Atkinson and Riani [1] contains theory and examples.

In generalized linear models, we have a response $\boldsymbol{y}$, a vector of linear predictors with elements $\boldsymbol{\eta} = \boldsymbol{x}^T\boldsymbol{\beta}$, and a link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ connecting the two. In the article on **Residuals**, the deviance D, the analogue of the residual sum of squares in regression, was written as

$$D = \sum_{i=1}^{n} d_i^2, \tag{3}$$

where d_i^2 is the contribution of the ith unit to the total deviance. The deviance residual was then defined as

$$r_{Di} = d_i \text{ sign } (y_i - \hat{\mu}_i). \tag{4}$$

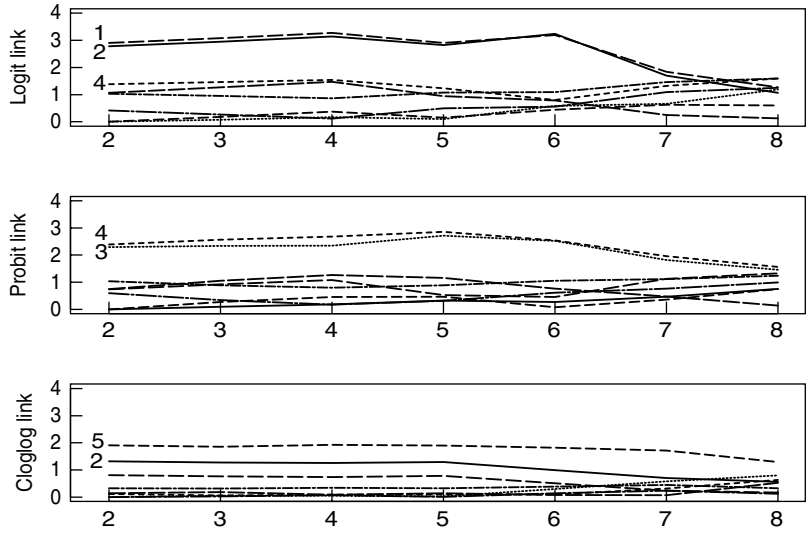

Figure 5 Bliss's beetle data: absolute values of deviance residuals as the subset size increases: (a) logit, (b) probit and (c) complementary log-log links

To extend the forward search to generalized linear models, we replace the least-squares residuals e_i with the deviance residuals r_{Di}. Then, as before, when m observations are used in fitting, the optimum subset $S^*(m)$ yields n deviance residuals $\boldsymbol{r}_D(m^*)$. We order the squared residuals $\boldsymbol{r}_D^2(m^*)$ and take the observations corresponding to the $m+1$ smallest as the new subset $S^*(m+1)$.

For the regression models in the previous sections, we looked at forward plots of residuals and of t tests for components of the linear predictor. As well as problems about individual outliers and the correct form of the linear predictor, there is also a need in generalized linear models to specify the correct form of link function. In the articles on **Goodness of Fit** and **Residuals**, analyses are given of Bliss's beetle data. These are binomial data in which the probability of success θ_i at dose level x_i is modeled by the link function $g(\theta_i) = \eta_i$. The analysis used the **logistic** link

$$g(\theta) = \log \frac{\theta}{1-\theta}. \tag{5}$$

There was evidence that this link was not satisfactory for these data. Alternative links are the probit

$$g(\theta) = \Phi^{-1}(\theta), \tag{6}$$

where Φ is the cdf of the standard normal distribution, and the complementary log–log link

$$g(\theta) = \log\{-\log(1-\theta)\}. \tag{7}$$

(*see* **Quantal Response Models**).

We explore these three possible link functions by looking at forward plots of absolute deviance residuals, which will indicate whether the unsatisfactory nature of the logistic link was caused by a few outliers or whether there is a systematic lack of fit. Figure 5 shows plots of absolute deviance residuals from forward searches for three models in which the explanatory variable is log(dose) and the three links are the logit, probit and complementary log–log. The observations are numbered from the lowest dose level to the highest. For the logit link observations, 1 and 2 are the last two to be included in the forward search. The crossing of the lines at the end of the plot in the top panel of Figure 5 shows that the inclusion of observations 1 and 2 seems noticeably to affect the ordering of the residuals. With the probit link units 3 and 4 (the last two to be included) seem to be different from the rest of the data: they are badly predicted by models in which they are not included. However, the residuals from the forward search with the complementary log–log link in the bottom panel of the figure show no such behavior; all residuals are smaller than two throughout, and relatively constant. Since the scale parameter is not estimated, it is possible to make such absolute comparisons of the residuals across different models, even if they come from different link families.

The conclusion from Figure 5 is that the complementary log–log link is satisfactory and that the other two are not. This conclusion is not dependent on a few observations, but is spread throughout the data. To sharpen and quantify this general impression based on forward plots of residuals, we now consider the goodness of link test, introduced in the article on **Goodness of Fit**. This provides a test for the adequacy of each link from the t test for the inclusion of the constructed variable $\hat{\boldsymbol{\eta}}^2$ in the linear predictor. The constructed variable plot in Figure 8 of the article on **Goodness of Fit** indicates rejection of the logistic link when all observations are used in fitting. We use forward plots of the test statistics to test three links and to see whether the conclusions are based on all observations.

Figure 6 shows a forward plot of the goodness of link test, the order of introduction of the observations, as in Figure 5, being different for the three links. For the logit and probit links, these plots show evidence of lack of fit at the 5% level, which is indicated by the statistic going outside the bounds in the plot. Although, it is inclusion of the last two observations

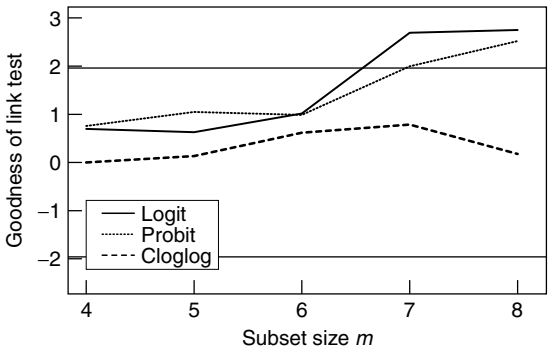

Figure 6 Bliss's beetle data: forward plot of the goodness of link test. Only the complementary log-log link is satisfactory

that causes the values of the statistic to become significant, it is clear from the steady upward trend of the plots that lack of fit is due to all observations. The plot for the complementary log–log link shows no evidence of any departure from this model. This plot also shows that unit 5, which is the one with the biggest residual for the complementary log–log link and the last to be included in this forward search, has no effect on the t value for the goodness of link test.

This analysis shows that, of the three links considered, only the complementary log–log link is satisfactory. The plot of fitted values for the logistic link in Figure 6 of the article on **Residuals** relates this finding to individual observations. The fitted dose response curve for this symmetrical link fits badly in the center of the experimental region, whereas, as Figure 6.36 of Atkinson and Riani [1] shows, the asymmetric complementary log–log link provides an appreciably better fit over the whole range of $\boldsymbol{x}$ values.

Multivariate Data

With multivariate observations, we replace the squared residuals used in the forward search for regression and generalized linear models with the squared **Mahalanobis distances**

$$d_i^2(m^*) = \{\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}(m^*)\}^T \hat{\boldsymbol{\Sigma}}^{-1}(m^*)\{\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}(m^*)\}, \tag{8}$$

where $\hat{\boldsymbol{\mu}}(m^*)$ and $\hat{\boldsymbol{\Sigma}}(m^*)$ are estimates of the mean and **covariance matrix** of the observations based on the subset $S^*(m)$. These distances are used for ordering the observations and for determining how we move forward in the search. We use the robust bivariate boxplots of Zani et al. [7] to determine an initial subset, which is not outlying in any two-dimensional plot of the data. The content of the contours is adjusted to give an initial subset of the required size. Once we have some idea of the structure of the data, we start the search with subsets that seem potentially interesting.

As an example with some expected and some unexpected structure, we look at readings on six dimensions of 200 Swiss bank notes, 100 of which may be genuine, and 100 forged. All notes have been withdrawn from circulation and classified by an expert, so some of the notes in either group may have been misclassified. Also, the forged notes may not form a homogeneous group. For example, there may be more than one forger at work. The data, and a reproduction of the bank note, are given by Flury and Riedwyl [4, pp. 4–8].

Figure 7 is a forward plot of Mahalanobis distances scaled by the estimate of $\boldsymbol{\Sigma}$ at the end of the search. The search starts with 20 observations on notes believed genuine. In the first part of the search, up to $m = 93$, the observations seem to fall into two groups. One has small distances and is composed of observations within or shortly to join the subset. Above these there are some outliers and then, higher still, a concentrated band of outliers, all of which are behaving similarly. The plot clearly shows the difference between the genuine notes and the forgeries. Toward the end of the search, there is evidence that the group of forgeries is not homogeneous.

The structure of the group of forgeries is also readily revealed by the forward search. Figure 8 is a forward plot of the scaled Mahalanobis distances just for the forgeries. In the center of the plot, around $m = 70$, this shows a clear structure of a central group, one outlier from that group and a second group of 15 outliers. As successive units from this cluster enter after $m = 85$, they become less remote and the distances decrease. By the end of the search there is appreciable masking, so that the group of 15 observations is no longer clear from the plot of the Mahalanobis distances. Under such

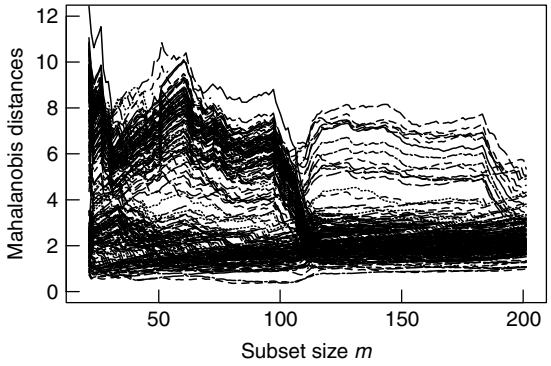

Figure 7 Swiss Banknote Data, all 200 observations: forward plot of scaled Mahalanobis distances starting with 20 notes believed to be genuine. The two groups are clear, but a third group seems to appear toward the end of the search

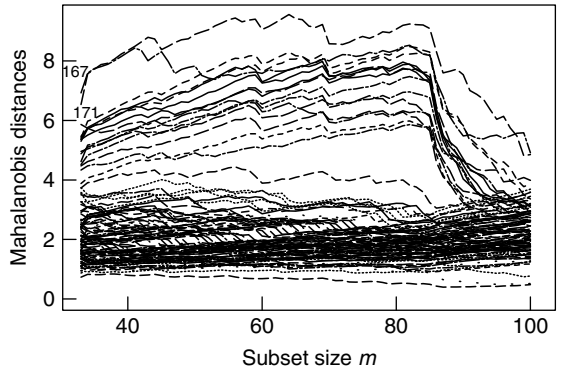

Figure 8 Swiss Banknote Data, 100 notes classified as forgeries: forward plot of scaled Mahalanobis distances. Toward the end of the search, there seems to be a group of 15 observations and a further single outlier

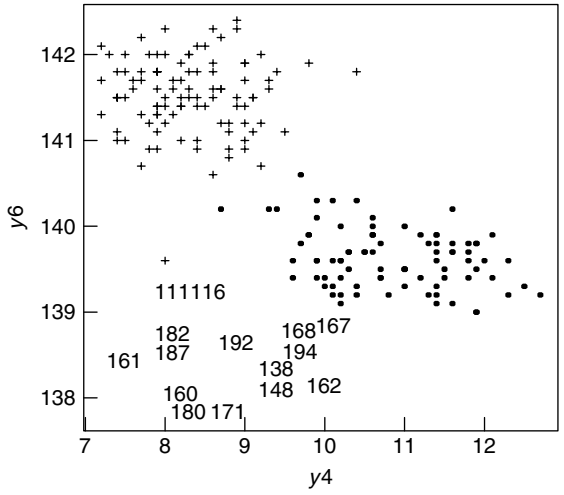

Figure 9 Swiss Banknote Data: scatterplot of y_6 against y_4. The "genuine" notes are marked with crosses; the labeled units are the last 15 to enter the search

conditions, the deletion methods described in the article on **Diagnostics** are likely to fail to reveal the structure.

In this example, the forward search clearly indicates not only the presence of two groups of notes, but also an unexpected subset of 15 observations, showing that the group of forgeries is not homogeneous but consists of two subgroups. Once attention has been drawn to the existence of this structure, it is possible to find it in the data. Figure 9 is one of the 15 different panels of the scatterplot matrix for these six dimensional data and by far the most revealing. The last 15 observations to enter the subset are numbered: the other forgeries are shown by filled circles and the "genuine" notes by crosses. It seems that one genuine note has been misclassified.

The entries in this article show various ways in which the forward search can elucidate the structure of data and, in the case of the third example, reveal unexpected subsets. A fuller analysis of the data on Swiss banknotes, together with numerous other applications of the forward search to multivariate data, are described in [3].

References

[1] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

[2] Atkinson, A.C. & Riani, M. (2002). Forward search added variable t tests and the effect of masked outliers on model selection, *Biometrika* **89**, 939–946.

[3] Atkinson, A.C., Riani, M. & Cerioli, A. (2003). *The Forward Search in Multivariate Data Analysis*. Springer–Verlag, New York. (In preparation).

[4] Flury, B. & Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Chapman and Hall, London.

[5] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association* **79**, 871–880.

[6] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[7] Zani, S., Riani, M. & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection, *Computational Statistics and Data Analysis* **28**, 257–270.

(*See also* **Model Checking**; **Model, Choice of**)

A.C. ATKINSON & MARCO RIANI

Foundations of Probability

Probability obeys three, basic laws (given below) about which there is little disagreement; and what there is has little effect on their mathematical consequences or on practice. Any quantity obeying these laws is termed a probability and, in contrast to the mathematics, there is considerable disagreement about the interpretation of the quantity. Furthermore, the disagreements over interpretation have important, practical consequences in that an **Inference** made from a data set can vary profoundly as a result of distinct views about probability. Probability is a subject where the foundations really matter. This article begins on the solid ground of the three laws, and then passes to the shifting sands of interpretation.

Probability is a numerical measure of uncertainty about an event. Uncertainty depends on the knowledge available at the time the uncertainty is being quantified. Probability therefore depends on *two* arguments, the uncertain event, A, under consideration and the truth of an event, B, describing the knowledge then possessed. It is written $\Pr(A|B)$ and reads "the probability of A, given B", the vertical line separating the uncertain and given events. [Many writers introduce it as $\Pr(A)$, omitting reference to the second argument, but experience shows that this can be a cause of practical confusion. They then term $\Pr(A|B)$ a **conditional probability**]. The three, basic laws, holding for a suitable collection of events, usually members of a Borel field, are

1. *Convexity*. For all $A, C, 0 \leq \Pr(A|C) \leq 1$ and $\Pr(C|C) = 1$.
2. *Addition*. If the events of a sequence $A_1, A_2, \ldots$ are exclusive, then
$$\Pr(\cup_i A_i|C) = \sum_i \Pr(A_i|C).$$
3. *Multiplication*.
$$\Pr(A \cap B|C) = \Pr(A|C)\Pr(B|A \cap C).$$

(*see* **Axioms of Probability**). Events are exclusive if the truth of any one precludes any of the others being true. $\cup_i A_i$ denotes the union of the events; namely, the event that is true if and only if any one of the individual events is true. In the case of two events, the union is written $A \cup B$. $A \cap B$ is the event that is only true if A and B are both true.

The whole of the rich calculus of probabilities flows from these three laws, or axioms. Yet the laws are merely simple expressions about proportions, as is seen by consideration of an urn containing a number, N, of balls, identical except that they are either white or black, and simultaneously either plain or spotted. Let one ball be drawn at random from the urn, event C above. Let A be the event that the ball is white, then $\Pr(A|C)$ is interpreted as the proportion of white balls w/N. It therefore lies between 0 and 1. Also $\Pr(C|C) = 1$ since some ball is certain to be drawn. This is convexity. To demonstrate additivity, let B be the event that the withdrawn ball is spotted and suppose that there are no white, spotted balls in the urn. This ensures that A and B are exclusive, and that $A \cup B$ is the event that the withdrawn ball is either white or spotted. Then,

$$\begin{aligned}\Pr(A \cup B|C) &= \frac{w+s}{N} = \frac{w}{N} + \frac{s}{N} \\ &= \Pr(A|C) + \Pr(B|C),\end{aligned}$$

where s is the number of spotted balls. For multiplication, suppose that the exclusive condition is removed, so that there are balls that are both white and spotted, r in number, and $A \cap B$ is the withdrawal of such a ball. Then,

$$\begin{aligned}\Pr(A \cap B|C) &= \frac{r}{N} = \left(\frac{w}{N}\right) \cdot \left(\frac{r}{w}\right) \\ &= \Pr(A|C)\Pr(B|A \cap C).\end{aligned}$$

That $r/w = \Pr(B|A \cap C)$ follows, since, in the probability, A, being a white ball, is known to be true and the only uncertainty in B is the additional requirement of it being spotted. It is, therefore, the proportion of spotted balls amongst the white.

Mathematically, probability is very simple, namely just a proportion. It is remarkable that such rich, mathematical consequences follow from such elementary ideas. In performing numerical calculations with probabilities, it is often useful and sensible to think of them as proportions. The addition law has been illustrated for only two events. Using more than two colours of balls, it easily extends to any finite number of events. The only disagreement about the laws of any significance is whether the addition law extends to an enumerable infinity of events.

The minority who say it does not, call probability *finitely-additive*. Otherwise, it is *sigma-additive*, although, this being the majority view, the adjective is usually omitted. Occasionally, the convexity law is extended by replacing $\Pr(C|C) = 1$ by $\Pr(A|C) = 1$ if and only if C logically implies A. This conveniently excludes some zero probabilities, since, equally, $\Pr(A|C) = 0$ if and only if C logically implies the falsity of A. The mathematical foundations of probability are simple, precise and essentially agreed. We now turn to interpretations, of which there are three principal ones.

Interpretations

The Classical View

The mathematical treatment of probability began with games of chance, like the rolling of dice. Here, there are often a number, N, of exclusive possibilities each of which has the same uncertainty. For example, with a single die, $N = 6$ and, if it is well-made and sensibly thrown, each of the six faces has the same chance of occurring. The illustration above, of an urn with a ball drawn at random, means that each ball has the same chance N^{-1} of being drawn. Exactly as in the discussion above, this leads to other probabilities and to the three laws based on the equiprobable cases. In games of chance, the laws may then be used as the basis of mathematical calculations to obtain values of interest, such as the probability of winning a game. Though influential in the early development of the subject, and still valuable in calculations, the classical view fails because it is seldom applicable. Thus, if actuaries want to find the probability of death within a period, there is no obvious set of cases having the same uncertainty.

The Frequentist View

This is currently the most popular interpretation of probability and that used in most biostatistical studies. It finds expression in many forms. The one that is most nearly related to the classical view is to think in terms of a population, usually infinite and often conceptual, where a proportion p of members have a property, A. In a change of language, p is the frequency of A in the population and $\Pr(A) = p$ is the probability that a member of the population has property A. [Here is an example where the second argument of probability is omitted. In this view, $\Pr(A|B)$ is defined, at least when $\Pr(B) \neq 0$, by the multiplication law, as $\Pr(A \cap B)/\Pr(B)$.] Thus, the actuary can think of a population, say, of white females in their 50s in a country, and think of the proportion dying within a year.

Another expression of the frequentist view is illustrated by the repeated tossing, under similar circumstances, of a drawing pin (American: thumb tack), which may either fall with the point up, U, or down, D. Experience shows that the frequency of Us in a long sequence of similar tosses appears to settle down to a limit. This limit is interpreted as $\Pr(U)$ for a single toss. It is not difficult to see that such an interpretation will satisfy the three laws. The concept is of considerable value in science because a good experiment can be thought of as a member of a sequence of similar experiments and the identification of probability with frequency makes sense. Indeed, the ability to repeat a phenomenon under controlled conditions is a hallmark of science. In the usual model for statistical inference, in which data x has probability (or probability density) $f(x|\theta)$ dependent on a parameter θ, the interpretation is the frequency of occurrence of x were the parameter to have the value θ, exhibiting the conditional form used in the formulation of the axioms above.

The classical exposition of the foundations of the frequentist view is by von Mises [13], though the important ideas of **Fisher** [5] have had more influence. For example, we have seen that probability requires the *knowledge* of a population with known frequencies. Fisher pointed out that it is also necessary to have *ignorance*, expressed by our inability to recognize any subpopulation within which different frequencies obtain. This observation has become important in restricting the sample space from all values of a **random variable** X to the subset of those values having a statistic with the value observed in the data. For instance, Fisher claimed that in the analysis of a **contingency table**, the population should be restricted from all tables of a given total size to tables having the same margins as that observed. The idea of recognizable subsets or subpopulations is important in applied work.

Although useful and widely applied, the frequency view has some limitations, the most serious of which is that it typically does not apply to uncertainty about the parameter in $f(x|\theta)$. θ is an uncertain number,

but it is not usually natural to think of it as a member of a population, or as having a frequency in a series. Consequently, in the frequentist position, it is not ordinarily sensible to refer to the probability of θ, given x, despite the fact that the value of the parameter, given the data, may be the principal interest in inference. These, and other difficulties, have largely been overcome by extensive developments this century. For instance, **confidence limits** for θ, based on x, provide an appealing substitute for $\Pr(\theta|x)$. Significance levels for a hypothesis H (*see* **Hypothesis Testing**; ***P* Value**) replace $\Pr(\sim H|x)$, where $\sim H$ is the complement of H.

The Bayesian View

This interpretation is fundamentally different from the other two. It starts with a person, or subject, conveniently referred to as "you". You are uncertain about the truth of an event, A, but have knowledge summarized by B. Then $\Pr(A|B)$ is a numerical measure of your belief that A is true, given B. The key concepts here are "you" and "belief". This interpretation of probability is personalistic, or subjective, and expresses the opinions of a person, or subject, about the uncertain aspects of the world as seen by that person. For the not very sound reason that **Bayes' theorem** plays a more important role when probability is interpreted as belief, rather than frequency, this attitude is called **Bayesian**. It has an immediate advantage over the frequentist view because it is always applicable. In particular, it makes sense to talk about $\Pr(\theta|x)$ or the probability that global warming is taking place. There are, however, some considerable difficulties associated with it.

The first problem is why your beliefs should be capable of being expressed by numbers and, if they are, why they should obey the three laws of probability. This has been answered by several, varied lines of argument that demonstrate that the numeracy and the laws follow from other, simpler axioms. These demonstrations will be discussed later. For the moment, it suffices to remark that there is substantial, logical support for the Bayesian position.

A second problem is that, even if belief can be equated with probability, how are the numbers to be obtained? In the frequency view, data on observable frequencies in finite series are available and only require the conceptual passage to the limit. The probability of death is measured by the frequency of death, or at least by suitable, actuarial treatment of observed frequencies. How are you to measure your belief that you will die within the year? Or, even harder, your belief that the political party you support will win at the next election? This difficulty has not been overcome satisfactorily, although, in many cases studied in frequentist statistics, it presents no serious problem. A valuable, recent reference is [16].

It is often said that a third difficulty with the Bayesian view is its subjectivity. A great strength of science is its claimed objectivity, so that the Bayesian position is often thought inappropriate for scientific inference. There are two responses. In the first it is held that two people with the same information, B in our notation, would, if logical, agree on $\Pr(A|B)$, so that any differences between persons are due either to different information or to false logic. One approach is to introduce situations, B_0, in which probabilities, given B_0, are agreed and then to calculate the general values by Bayes' theorem. B_0 is commonly thought of as a position of ignorance. Whilst substantial progress has been made, difficulties remain. This is sometimes called a "necessary" view: for stated A and B, $\Pr(A|B)$ necessarily follows. The second response is to deny that all of science is objective. Scientists differ strongly in their beliefs concerning global warming. The early stages of a scientific study, which is where inference may be useful, is subjective. Objectivity occurs when enough data have been accumulated and beliefs converge. The sun will rise tomorrow. Additionally, we notice that much of "objective" science is based on probability, or what we will later call "chance". Quantum mechanics and genetics are two examples. A modern, subjective approach is provided in [8].

These are the three, principal, foundational aspects of probability. In the remainder of this article, the frequentist and Bayesian views are explored in greater depth because their practical differences are of importance to biostatisticians.

Justification for the Laws

Ever since the earliest days of probability, the laws have been discussed. They were finally put into a precise, mathematical form in 1933 by **Kolmogorov** [7], essentially that given above. This formulation has provided the basis for the enormous advance in probability, as distinct from scientific inference,

since then. Probability is a field in which the foundations are the laws, and problems of interpretation are often not needed or, if they are, can comfortably be accommodated within the frequentist view. The situation is different in statistics. At about the same time as Kolmogorov, Ramsey [10] took a very different approach. He was concerned with someone, "you", having to choose amongst a number of actions in a world where uncertainty was present and relevant. He asked what principles you might use to decide on one action in preference to others. He developed a number of basic principles that he felt should obviously obtain in decision-making and used these as axioms to construct a calculus. Here is an example of such an axiom. Suppose event A is the only uncertainty present and that, were A to be true, you would prefer action a_1 to action a_2. Suppose that this preference persists were A to be false. Then the axiom says that you would still prefer a_1 to a_2 when A is uncertain for you. This is a form of the "sure-thing" principle; sure, because the status of A is irrelevant.

The calculus of decision making that Ramsey developed had to encompass the uncertainties present, described by beliefs about relevant events. To produce a single number as a measure of uncertainty, he introduced the idea of an ethically, neutral event of probability 1/2: one for which you did not care whether it was true or not, and where the uncertainties of truth and falsity were equal. The toss of a coin that you judge to be fair is an example. This provided a standard by which other uncertainties could be compared. He showed that the axioms implied that beliefs should obey exactly the three laws listed above. In other words, he provided a justification for the Bayesian position by showing that beliefs had to be expressed through probabilities. Ramsey's axioms are at a more fundamental level than Kolmogorov's. The latter's laws are the former's theorems.

It was a startling advance but lay essentially unappreciated until **Savage** [11] independently presented a similar development that attained the level of modern, rigorous mathematics. By then, two other approaches had appeared. In 1939, **Jeffreys** [6], a geophysicist, was concerned with inference in the handling of scientific data and developed axioms for reasonable inferential procedures. Again, the principal deduction from his axioms was that scientific beliefs must obey the probability laws. Unlike other writers, Jeffreys went on to develop practical procedures for treating scientific data. In particular, he developed an original method of testing a hypothesis, H, that addressed $\Pr(H|\text{data})$ directly, rather than through a significance level. Partly because Jeffreys was a physical scientist, and partly because Fisher, working in biological fields, was successfully originating frequentist methods, these Bayesian ideas have had less impact in biostatistics than in other fields.

Meanwhile, two other approaches were being developed by **de Finetti** [4]. Suppose you were asked to describe your belief in the truth of an uncertain event, A, by a number x. Suppose, further, you were told that, were you to state x, you would be given a penalty score $(x-1)^2$ if A were subsequently shown to be true, and x^2 if false. What value of x would you choose? Extend this idea to several events, each with its associated value and consequent score. Finally, the scores are to be added. What properties would the chosen values possess? In some simple and beautiful mathematics, de Finetti showed that your numbers must obey the laws, although, in the case of the addition law, it need only hold for a finite number of events – the finite additivity already mentioned. The procedure for calculating the penalty is called a scoring rule. That above is the quadratic scoring rule. It has subsequently been shown that the result holds for any, reasonable rule.

De Finetti also introduced a second method based on bets concerning an uncertain event, A. Suppose that you were required to post **odds** x against A, in the sense that you would accept a stake s on A, and be prepared to pay out xs (and return the stake) were A subsequently shown to be true, retaining the stake if false. Furthermore, unlike most real betting situations, you were also prepared to have the stake s placed on $\sim A$, returning $x^{-1}s$ if $\sim A$ were true; odds against $\sim A$ being the inverse of odds against A. Suppose that you did this for several events, each with its associated odds. Then he showed that unless the values $(1+x)^{-1}$ obeyed the laws of probability, again with the finite restriction, a person could make a Dutch book against you; that is, the person could place a series of bets that would result in your experiencing sure loss, whatever the truth or falsity of the events. Again, the laws of probability are an inevitable consequence of reasonable requirements.

There have been numerous extensions of these ideas and the position today is that there are many and varied ways, either based on decision-making under uncertainty, or directly on belief as a primitive notion,

that lead to the use of the probability calculus in the expression of beliefs about uncertain events. These all provide a justification for the Bayesian view. It is noteworthy that adherents of the frequentist position have not been prepared to state which of the axioms they object to. This they should do, since implicit acceptance of them implies the Bayesian paradigm. There have been cogent objections raised, but these lead to attitudes that are far removed from those adopted by frequentists. For example, one approach leads to upper and lower probabilities, employing two numbers to describe uncertainty for an event instead of the single probability. These two values obey laws similar in character to, and extensions of, those listed above. Walley [15] is an excellent reference. Shafer [12] has introduced another variant, in "belief functions".

Inference and Action

The approach to uncertainty, leading to probability, used by Ramsey, Savage and some others, is indirect in that it investigates, and treats as fundamental, decision-making, rather than belief (*see* **Decision Theory**). An advantage is that it yields two other dividends besides probability. It demonstrates that the outcome to any action needs to be described by a number describing its worth to you. This is your **utility** for that outcome. Furthermore, the optimum act is that which **maximizes your expected utility**, MEU. The expectation is calculated by reference to the probabilities. Thus, if an act can lead to one of n exclusive and exhaustive outcomes with utilities $u_1, u_2, \ldots, u_n$, having probabilities $p_1, p_2, \ldots, p_n$, the expected utility of that act is $\sum u_i p_i$. Finally, that act is taken of maximum expected utility.

In a modified form, this conclusion has been accepted by the frequentist school. This stems from the work of **Wald** [14]. He used loss in place of utility, so minimizing expected loss instead of MEU, but the change is of no serious content. Wald started from the frequentist approach and adopted loss as a primitive idea in extension of the earlier ideas of losses in connection with errors of the two kinds in hypothesis-testing. Loss, unlike utility, was not deduced from more basic requirements, as with Ramsey. Wald proved a basic, general theorem that essentially showed that only decisions that were MEU could be sensible. Against any other rule, a Dutch book could be made. Wald called such solutions to a decision problem, Bayes' solutions. Because he had no way of determining the probabilities (p_i above), which did not have a meaning within the frequentist paradigm, they were merely positive numbers, adding to one. Wald was unable to recommend a unique decision rule, but only to say that a sensible rule must be a Bayes' solution. He did advocate a rule, **minimax**, which was soon shown to be unsound. Today, frequentists often choose a Bayes' solution, justifying the p's by frequency considerations of how the rule would behave in conceptual repetitions.

The decision-oriented work is important for inference because it supplies perhaps the best answer to the question: what is the purpose of inference? Why do we test hypotheses and estimate parameters? Some might reply that it is carried out in pursuit of an understanding of the world and that it is part of knowledge for its own sake. But knowledge, however "pure" scientists like to think of it, is used to select actions. Knowledge about DNA is used to change plants. The studies we have mentioned show that MEU is the proper method to select an action. It follows, therefore, that our pure knowledge should be stated in a form appropriate to MEU. In particular, scientific inference should be so appropriate. MEU requires, in estimation, probability statements about parameters, $\Pr(\theta|\text{data})$, probabilities that are only available in the Bayesian approach. Confidence intervals are not in a form suitable for MEU. Therefore, according to followers of Savage, they are unsatisfactory as a form of statistical inference. Similarly, in hypothesis-testing, $\Pr(H|\text{data})$ fits with MEU, whereas a significance level does not.

Exchangeability

It often happens that probability as belief differs numerically from probability as frequency. To take a simple example, return to the tossing of a drawing pin mentioned above. After you have tossed it a few times, you will have a belief about the uncertain event of it falling point uppermost at the next toss. This may differ from the limiting frequency of such falls, a value that may be unknown to you. The biometrician, W. F. R. Weldon, had probability 1/6 that a die his research assistant was tossing would show six. After many tosses, the frequency was found to exceed that value. There is, however,

a relationship between the two values of belief and frequency expressed in a result usually ascribed to de Finetti. Consider a sequence of uncertain quantities (random variables is the frequentist terminology) $X_1, X_2, \ldots, X_n$. With the pin or the die, $X_i = 0$ or 1 according to the result of the ith toss. There are many situations in which your beliefs about the sequence are such that they are invariant under permutation of the suffixes. Thus, $\Pr(X_1 = 0, X_2 = 1, X_3 = 1) = \Pr(X_3 = 0, X_1 = 1, X_2 = 1)$, etc. If this is true for all n, the sequence is said to be exchangeable; any X_i can be exchanged for any other as far as your beliefs are concerned. With exchangeable, binary X_i, your probability for any finite sequence of 1's and 0's depends only on r and $n - r$, the numbers of 1's and 0's, respectively, in the sequence. De Finetti's result says that for a binary, exchangeable sequence, the probability of any sequence with r 1's and $n - r$ 0's is

$$\int_0^1 x^r (1 - x)^{n-r} \, \mathrm{d}F(x)$$

for some distribution function $F(\cdot)$ on [0, 1]. Furthermore, r/n tends, with probability one, to a limit, θ say, as $n \to \infty$. There is a similar result for general, exchangeable sequences.

The integral effectively says that you can express your beliefs about the sequence by supposing that there exists a value x such that, given x, the X's are independent and identically distributed with $\Pr(X_i = 1|x) = x$, in other words, a Bernoulli sequence (*see* **Binary Data**). Your belief about x is described by the function $F(\cdot)$. Using the limit result, you, as a Bayesian, can act like a frequentist, having a Bernoulli sequence with parameter θ, but, unlike a frequentist, having a probability distribution for θ. Notice that although to a frequentist θ is a probability, indeed, it is the defining expression; to a Bayesian it is not usually a belief. It would describe a belief were it known, which, being a limiting frequency, it usually is not, as with Weldon. Bayesians often describe θ as a *chance*, to emphasize the distinction. Without the distinction, $\Pr(\theta)$ appears as a probability of a probability, which is nonsense.

Since most situations studied in statistical inference are either based on exchangeable sequences (as with a random sample from a population) or on sequences that are modified from an underlying exchangeable one (as in an autoregressive process (*see* **ARMA and ARIMA Models**), $X_{n+1} = \alpha X_n + \varepsilon_n$ where the ε's are exchangeable), this result is of wide applicability in providing a link between frequentist and Bayesian ideas. It is also important in physics, where the behavior of a set of particles is often judged exchangeable. The probabilities that physicists use are chances in the above terminology, not beliefs.

The Likelihood Principle

It therefore happens that a Bayesian and a frequentist will use the same model of independent and identically distributed random variables, given a parameter θ, but the former will add a probability distribution for θ. Foundationally, the two viewpoints come close together as a result of de Finetti's theorem. However, there is something that pulls them apart and can make their results differ. Repeating what was said above, the foundations of probability really do matter. What separates them are their attitudes towards the likelihood principle.

Let data x be obtained as a result of observing a random variable X having density $f(x|\theta)$ for each parameter value θ. This is the part common to the two schools. As a function of θ for fixed x, $f(x|\theta)$ is called the likelihood of θ (at x). Let $p(\theta)$ be the density for θ adopted by you, as a Bayesian, expressing your beliefs about the parameter based on general knowledge of the situation, but excluding the data. This is usually termed the prior (to x) distribution. The data will change your belief about θ to

$$p(\theta|x) \propto f(x|\theta) p(\theta)$$

by Bayes' theorem. This is the posterior (to x) distribution of θ, given x. The important point here is that the posterior belief about θ depends on X, the quantity observed, only through x, the observed value. In particular, if two values, x and y, have the same likelihood, $f(x|\theta) = f(y|\theta)$ for all θ, then the inferences about θ from x and from y are the same. That statement constitutes a form of the likelihood principle. Bayesians respect and obey the principle in inferences from data, as the above development shows. (The principle does not apply in experimental design or generally in pre-data analysis.) A good reference is [2].

The frequentist does not obey the principle and therein lies a major, practical difference between the two paradigms. For example, a frequentist may use

an unbiased (*see* **Unbiasedness**) estimate $t(x)$ of θ; that is, one satisfying

$$\int t(x) f(x|\theta)\, \mathrm{d}x = \theta,$$

for all θ, where the integral is over the sample space, the range of X. This concept uses, in the integration, values of X other than the observed value x, thereby violating the principle. Often the space is restricted, as we saw above with Fisher's ideas, by confining the range to values of x having common elements with that observed, as with the margins of a contingency table. The key point is that the frequentist requires a space of x-values; the Bayesian does not. By the likelihood principle, the latter restricts the space to one value; that observed. The contrast between the two views is clarified by a slight exaggeration: Bayesians operate in the space of the parameter θ; frequentists in the sample space of X. For example, in contrast to the frequentist's unbiased estimate, the Bayesian might use the posterior mean

$$t(x) = \int \theta p(\theta|x)\, \mathrm{d}\theta,$$

with integration over the whole parameter space.

Here is an example that illustrates the difference. Suppose a drawing pin is tossed a number of times under conditions that lead both the frequentist and the Bayesian to think the sequence exchangeable and therefore, by de Finetti, Bernoulli. Consider two scenarios:

1. An integer n is selected. The pin is tossed n times and is observed to fall point uppermost on r occasions.
2. An integer r is selected. The pin is tossed until is falls uppermost for the rth time. This takes n tosses.

Suppose the values (r, n) are the same in the two scenarios. Then, for a given sequence; that is, including the order of the results, the likelihood for both scenarios is $\theta^r(1-\theta)^{n-r}$ for parameter (chance) θ. The Bayesian will make the same inference in both cases; namely,

$$p(\theta|r, n) \propto \theta^r(1-\theta)^{n-r} p(\theta).$$

However, the frequentist will use the **binomial distribution** ${}^nC_r\theta^r(1-\theta)^{n-r}$ in 1 and the negative binomial ${}^{n-1}C_{r-1}\theta^r(1-\theta)^{n-r}$ in 2. In 1, the unbiased estimate of θ is r/n; in 2, it is $(r-1)/(n-1)$.

The violation of the likelihood principle by frequentists is surprising, since the principle follows logically from the **conditionality principle** and **sufficiency**, usually accepted by frequentists, see [3].

Hypothesis Testing

A common activity in science is that of testing a hypothesis. A null hypothesis is erected as an "Aunt Sally" (American: "Straw man") and data are collected in an attempt to destroy it. Repeated failure to do this leads to its acceptance as part of scientific knowledge. The frequentist approach to testing uses the concept of a significance level. The Bayesian, following Jeffreys, addresses the problem directly through the probability of the null hypothesis, given the data. To illustrate the substantial differences between the approaches, take the case where exchangeable observations $x_1, x_2, \ldots, x_n$ are, given θ and σ^2, normally distributed with mean θ and variance σ^2. Here σ^2 is known but θ is not and the null hypothesis is that $\theta = 0$. Although somewhat specialized, many testing situations resemble it approximately, at least in large samples, and the conclusions derived from it apply rather generally. A more detailed treatment is provided by [1].

The frequentist will use the statistic $t = n^{1/2}\overline{x}/\sigma$, where $\overline{x}$ is the sample mean. On the null hypothesis, this is standard normal with zero mean and unit variance. The significance level of a two-sided test is $P = 2\Phi(-t)$, where Φ is the distribution function of a **standard normal deviate**. Thus, if $t = 1.96$, $P = 0.05$.

The Bayesian will require a prior distribution for θ when $\theta \neq 0$. Suppose this is $\mathrm{N}(0, \tau^2)$. The mean is reasonably at the null value and τ measures the spread of plausible alternatives about the null value. The next stage is to calculate by Bayes' theorem the posterior odds on $\theta = 0$. Odds are chosen in preference to probabilities since the theorem is simpler in terms of them. Under the null, $\overline{x}$, which is sufficient, is $N(0, \sigma^2/n)$. Under the alternative, it is $N(0, \sigma^2/n + \tau^2)$. To pass from prior to posterior odds it is necessary to multiply by the likelihood ratio; namely, the ratio of the probability of $\overline{x}$ when $\theta = 0$, to that when $\theta \neq 0$. This is

$$\frac{n^{1/2}\sigma^{-1} \times \exp[-n\overline{x}^2/2\sigma^2]}{(\sigma^2/n + \tau^2)^{-1/2} \times \exp[-x^2/2(\sigma^2/n + \tau^2)]},$$

which simplifies to

$$(1 + \rho^{-2})^{1/2} \times \exp[-t^2/2(1 - p^2)],$$

where $\rho = \sigma/\tau n^{1/2}$. This is well approximated by its form for a large sample size when $\rho \to 0$;

$$n^{1/2}\tau/\sigma \times \exp(-t^2/2).$$

The posterior odds will therefore be small if, like the P value of the frequentist, $|t|$ is sufficiently large. Both schools therefore agree that $t = n^{1/2}\overline{x}/\sigma$ is the right statistic and that numerically large values cast doubt on the null. But how large does it need to be for significance? The frequentist says larger in modulus than a value that does not depend on n, but only on P. Thus, if $P = 0.05$, it needs to exceed in modulus 1.96 for all n. The Bayesian treatment says it has to be large enough for the posterior odds on $\theta = 0$ to be sufficiently small. Taking logarithms of the approximate value above, this amounts to saying

$$t^2 > \log n + 2\log\left(\frac{\tau}{\sigma}\right) + K, \qquad (*)$$

where K is a constant depending on the critical value selected for the odds. Thus, to compare with the frequentist $P = 0.05$, odds of 19 to 1 against might be selected, when $K = 2\log 19 = 5.89$.

A striking difference between the two schools is now revealed. The frequentist demands for significance that t^2 is greater than a fixed value, whereas the Bayesian demands that it exceeds the value in $(*)$. The latter depends on what alternatives were reasonable a priori as expressed by τ/σ. But, more importantly, it depends on the sample size, unlike the frequentist's value. Either for sufficiently large τ or for a sufficiently large sample, the Bayesian would not declare significance, whereas the frequentist would. It is easier for the frequentist to obtain significance than it is for the Bayesian. The inclusion of the term in $\log n$ in the Bayesian approach, but not in the frequentist, makes for substantial operational differences between the two schools. This conclusion applies not just in this little problem but more widely; for example, in model selection, the logarithm acting like an automatic Occam's razor and deterring the introduction of extra parameters (*see* **Parsimony**).

Fuzzy Logic

Probability is a way of handling uncertainty. It is not the only way that has been suggested. An alternative approach is by fuzzy logic. It is now possible to buy electronic equipment "designed by fuzzy logic". The laws of this logic concern uncertain events, as does probability, but are not based on addition and multiplication, but on minimization and maximization. This results in considerable operational differences between the two methods. One feature that distinguishes fuzzy logic from probability is that the former has no axiomatic justification. The minimization law has not been proved from other, more primitive, assumptions on the lines followed by Ramsey, de Finetti and others. Rather, it has been invented as simple and apparently sensible. A good discussion of the foundational position and its practical effect is given in [9], with discussion.

References

[1] Berger, J.O. & Delampady, M. (1987). Testing precise hypotheses (with discussion), *Statistical Science* **2**, 317–352.

[2] Berger, J.O. & Wolpert, R.L. (1988). *The Likelihood Principle*. Monograph Series, Vol. 6. IMS, Hayward.

[3] Birnboum, A. (1962). On the foundations of statistical inference, *Journal of American Statistical Association* **57**, 269–306.

[4] de Finetti, B. (1974–1975). *Theory of Probability*. 2 vols. (translated from the Italian). Wiley, London.

[5] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Oliver & Boyd, Edinburgh.

[6] Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford.

[7] Kolmogorov, A.N. (1956). *Foundations of the Theory of Probability* (translated from the German). Chelsea, New York.

[8] Lad, F. (1996). *Operational Subjective Statistical Methods*. Wiley, New York.

[9] Laviolette, M., Seaman, J.W. Jr., Barrett, J.D. & Woodall, W.H. (1995). A probabilistic and statistical view of fuzzy logic (with discussion), *Technometrics* **37**, 249–292.

[10] Ramsey, F.P. (1926). Truth and probability, in *The Foundations of Mathematics and Other Logical Essays*. Kegan, Paul, Trench, Trubner, London 1931.

[11] Savage, L.J. (1954). *The Foundations of Statistics*. Wiley, New York.

[12] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

[13] von Mises, R. (1964). *Mathematical Theory of Probability and Statistics*. Academic Press, New York.

[14] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.
[15] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, London.
[16] Wright, G. & Ayton, P. (1994). *Subjective Probability*. Wiley, Chichester.

(*See also* **Likelihood**; **Subjective Probability**)

DENNIS V. LINDLEY

Founder Effect

The "establishment of a new population by a few original founders (in an extreme case, by a single fertilized female) that carry only a small fraction of the total genetic variation of the parental population" was termed the *founder principle* by Mayr [3]. The effect on the descendant population resulting from the small number of **genes** brought by the founders is called the founder effect. The effects are many. First, there is a great reduction in genetic variability compared with the parental population. If **heterozygosity**, averaged over a large number of loci, is used as a measure of genetic variability, then there is a great reduction in heterozygosity. Secondly, there is a high probability of extinction of the descendant population. Thirdly, the effect of random genetic drift is very pronounced, resulting in large fluctuations of allele frequencies in the initial generations. Fourthly, the levels of **inbreeding** are high in the initial generations, which adds to loss of heterozygosity. Compared with the parental population, there is not only a great reduction in average heterozygosity (calculated from allele frequencies at several neutral loci; *see* **Gene Frequency Estimation**) in the small, newly founded population, but the average heterozygosity remains reduced for a very long time even after the descendant population recovers its original parental population size [4]. The frequencies of recessive disease alleles (*see* **Genotype**) and neutral alleles are particularly affected. Some disease alleles are lost through drift, and some alleles that may have been rare in the parental population are pushed to high frequencies. For example, three specific **mutations** in BRCA1 and BRCA2 genes that are responsible for breast cancer are more common in the Ashkenazi Jewish population, whose ancestry can be traced to a small group of founders from central and eastern Europe. Studies indicate that there is a markedly increased prevalence of two mutations in the BRCA1 gene, 185delAG and 5382insC, and one mutation, 6174delT, in the BRCA2 gene, among Ashkenazi Jews than among their ancestral populations, which has been attributed to the founder effect [5–7]. These effects and features are being fruitfully exploited for understanding the architectures of genetic diseases.

Consider a large population, from which a group of individuals moved out and founded a new population. In the large parental population, except for newly arisen ones, disease genes occur on many haplotype backgrounds (*see* **Haplotype Analysis**). That is, if one considers a number of marker loci in the region flanking the disease locus, then the combinations of alleles at these **marker** loci on chromosomes that carry the disease-causing allele are usually many. However, in the newly founded population, because the number of founding individuals is small, the haplotypes of chromosomes carrying the disease-causing allele are much fewer. Thus, individuals who are affected will show a much greater sharing of marker alleles in the genomic region surrounding the disease locus in the descendant population than in the parental population. Therefore, by examining the sharing of haplotypes among affected individuals, one can map more easily the disease locus in the descendant population than in the parental population. This enhances the power of mapping genes in a newly founded population with a small number of founders. This feature has been very successfully exploited to map a gene for a rare autosomal recessive disease, namely benign recurrent intrahepatic cholestasis [2]. This gene was mapped to chromosome 18 by an analysis of shared genomic segments in only four affected individuals belonging to an isolated fishing community, of several thousand individuals, in The Netherlands.

In newly founded populations, dominant disease alleles are lost much more quickly than recessive disease alleles, unless the reduction in fitness due to the dominant disease is negligible. In other words, recessive disease alleles tend to persist longer in small populations compared with dominant disease alleles. However, due to the founder effect, and subsequent genetic drift, different newly founded populations often have different sets of recessive diseases. Therefore, such populations are very useful in mapping genes underlying various recessive diseases.

Unless the initial number of founders is very small, alleles and haplotypes that are common in the parental population are seldom completely absent in a descendant population [8]. Many common diseases are currently thought to be due to actions of common alleles at multiple loci. Furthermore, genotypes at the loci that underlie a common disease also are known to interact with environmental (including cultural) factors (*see* **Gene-environment Interaction**). Populations that are founded by a small number

of individuals often share environmental and cultural factors. Thus, such populations are generally environmentally and culturally homogeneous. This is a nongenetic founder effect, which is currently being exploited by geneticists to map genes underlying common and **complex diseases**. Examples of populations, which are currently being investigated for such purposes, are the Amish of the US and the Bedouins of west Asia. While founder effects can be fruitfully exploited for mapping genes, it is becoming increasingly clear that there are other crucial determinants of the success of such endeavors. Haplotype sharing among affected individuals provides clues to the locations of disease genes. However, when there are multiple genes underlying a disease, such as a common, complex disease, it is harder to discern haplotype sharing. If a disease allele is old, then the region around the disease locus in which marker alleles will be shared among affected individuals is expected to be short. Thus, gene mapping by examination of haplotype sharing is more efficient in newly founded populations. There is currently a major interest in estimating the nature and extent of **linkage disequilibrium** in isolated populations with different founding and demographic histories. The data being generated through individual efforts of researchers and the Single Nucleotide Polymorphism (SNP) Consortium will be of great value in judging the usefulness of population isolates that exhibit strong founder effects [1].

References

[1] Collins, F.S., Guyer, M.S. & Chakravarti, A. (1997). Variations of a theme: cataloguing human DNA sequence variation, *Science* **278**, 1580–1581.

[2] Howen, R.H. et al. (1994). Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis, *Nature Genetics* **8**, 380–386.

[3] Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press, New York, p. 237.

[4] Nei, M., Maruyama, T. & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations, *Evolution* **29**, 1–10.

[5] Oddoux, C., Struewing, J.P., Clayton, C.M. et al. (1996). The carrier frequency of the BRCA2 6174delT mutation among Ashkenazi Jewish individuals is approximately 1%, *Nature Genetics* **14**, 188–190.

[6] Roa, B.B., Boyd, A.A.,Volcik, K. et al. (1996). Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2, *Nature Genetics* **14**, 185–187.

[7] Struewing, J.P., Abeliovich, D., Peretz, T. et al. (1996). The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals, *Nature Genetics* **11**, 198–200.

[8] Terwilliger, J.D., Zollner, S., Laan, M. & Paabo, S. (1998). Mapping in small populations with no demographic expansion, *Human Heredity* **48**, 138–154.

(*See also* **Genetic Correlations and Covariances**; **Genetic Counseling**; **Inbreeding**)

PARTHA P. MAJUMDER

Fractional Factorial Designs

Fractional factorial designs are designs in which only a fraction of the total number of possible factor combinations is used (*see* **Factorial Experiments**). Fractional factorials were introduced by **Fisher** [5] in 1942 for use in agricultural experiments. Unfortunately, these powerful designs were not widely used in medicine or the health sciences until much later. In the biopharmaceutical arena, fractional factorial designs are widely used in exploratory studies of new organic synthesis routes and pharmaceutical formulation development. In addition, fractional factorials are the basis of *matrix designs* used in drug stability studies (see [13]).

Fractional factorials are a very efficient methodology for reducing the number of experimental runs necessary for exploratory studies. When one is interested in many factors, use of a full design would result in a prohibitively large number of runs. For example, suppose one wants to include eight factors in a two-level design; a full factorial would require $2^8 = 256$ runs! Experience has shown that it is reasonable to assume that three-way and higher-order **interactions** are negligible, and hence it is indeed wasteful to use a full factorial for a moderate to large number of factors. With eight factors, there are only eight main effects and 28 two-way interactions, which requires 37 runs not 256 runs.

In this article we will see how to select fractions of full designs so that effects of interest are estimated as efficiently as possible. The focus of the discussion will be directed toward two-level designs; however, other fractional designs are discussed and references are given for the interested reader.

Construction of Fractional Factorials

The construction of general fractional factorials is a challenging combinatorial problem. There is no unique method for construction, and given the characterization of the factorial, many methods have been developed. A variety of construction methods are summarized in Raktoe et al. [15].

Fractional factorials are characterized by the number of levels $r_1, \ldots, r_k$ specified for a set of k factors. Designs are *symmetric* if all factors have the same number of levels, $r_i = r, i = 1, \ldots, k$, and are said to be *mixed* otherwise. *Prime powered* designs have $r_i = p^m$, where p is a prime and m is a positive integer. Construction methods based on group theory and finite geometry have been developed for prime powered symmetric fractional designs. Details are given in Raghavarao [14]. The special subclass of two-level designs is expounded upon in the next section.

An Algorithm for Standard Two-level Designs

A two-level fractional factorial design with k variables is denoted as a 2^{k-p} design, where $1/2^p$ is the size of the fraction and 2^{k-p} is the number of runs. It is easy to construct these designs with a specialized representation of the design and model matrices. The design matrix of a factorial design defines the settings of individual factors over the series of experiments. The model matrix is an extension of the design matrix that includes all terms in the model: main effects and interactions. The first column of the model matrix is a column of 1s, the next k columns comprise the design matrix, and the remaining columns depict any interactions of interest.

We begin with a simple example for three factors. This is for illustration of principles only. The model matrix for a full 2^3 design is given in Table 1. We use both the ± 1 notation and the lowercase letter notation to denote experimental combinations.

Suppose we wish to select four of these eight runs to estimate the main effects of the three factors. Then we need a 2^{3-1} design. The one rule we must always obey is to select runs that result in an **orthogonal design**. An orthogonal design is one in which the sums of all pairwise column products equals zero. This ensures that we have an **optimal design**. One orthogonal run selection is c, b, a, and abc, which

Table 1 A full 2^3 design

Experiment		I	A	B	C	AB	AC	BC	ABC
1	(1)	1	−1	−1	−1	1	1	1	−1
2	c	1	−1	−1	1	1	−1	−1	1
3	b	1	−1	1	−1	−1	1	−1	1
4	bc	1	−1	1	1	−1	−1	1	−1
5	a	1	1	−1	−1	−1	−1	1	1
6	ac	1	1	−1	1	−1	1	−1	−1
7	ab	1	1	1	−1	1	−1	−1	−1
8	abc	1	1	1	1	1	1	1	1

results in the following fractional design matrix:

Run	A	B	C
c	-1	-1	1
b	-1	1	-1
a	1	-1	-1
abc	1	1	1

What are the consequences of use of the four-run vs. the eight-run design? The answer is that less information is available. The reader may verify that, in the fractional design, the interaction effects are such that $BC = A$, $AC = B$, $AB = C$, and $ABC = I$. This phenomenon is called **confounding** and it means that the interactions are not separable from the main effects. Unseparable effects are called *alias sets*. The four orthogonal estimable effects in the fractional design,

$$\delta_1 = \frac{(-c - b + a + abc)}{2},$$
$$\delta_2 = \frac{(-c + b - a + abc)}{2},$$
$$\delta_3 = \frac{(c - b - a + abc)}{2},$$
$$\delta_4 = \frac{(c + b + a + abc)}{4},$$

actually estimate $A + BC$, $B + AC$, $C + AB$, and $I + ABC$, respectively. Thus, if δ_1 is large, then we do not know if it is due to the main effect of A or the interaction BC or both. This is the price we pay. However, this is not serious because designs may be augmented with additional experiments to gather more information if necessary. Even in situations where there is some confounding with two-way interactions, one can intelligently guess which effects are likely to be contributors because experience has shown that real interactions are unlikely when main effects are not present. Suppose, in the above example, only δ_1 and δ_2 are large; then one can postulate that main effects A and B are probably present rather than interactions BC and AC since δ_3 was small. However, confirmatory experiments should always be run!

The general **algorithm** for construction of 2^{k-p} designs is based upon defining design *generators*. Design generators generate the design and identify associated alias sets. We begin with an example of six factors in 16 runs. This is a 2^{6-2} design. The process begins with a full 2^4 design in which columns $A-D$ are designated as usual. We complete the design matrix by assigning two added factors E and F with $E = BCD$ and $F = ACD$ to produce the design matrix in Table 2. All of the resulting confounding relationships are obtained with use of the defining relationships $E = BCD$ and $F = ACD$. The other confounding relationships are identified by use of a convenient multiplicative relationship, $A \odot B = AB$ and $A \odot A = I$. The $\odot$ binary multiplication of uppercase letters corresponds to columnwise multiplication of $\pm$ 1s in the design matrix. We begin by multiplying both sides of defining relationships by the letter on the left. The defining relationships produce $E \odot E = BCD \odot E$ or $I = BCDE$ and $F \odot F = ACD \odot F$ or $I = ACDF$. Both of these imply $I \odot I = BCDE \odot ACDF$ or $I = ABEF$. We now have the design generator sequence:

$$I = BCDE = ACDF = ABEF.$$

A design generator sequence has 2^p effects, where p is the degree of fractionation. The complete alias structure of the design is comprised of 2^{k-p} cosets with p aliased effects in each. A 2^{6-2} design has 16 cosets with four aliases in each. For example, the second coset (take the first coset as the design generator sequence) is $A = ABCDE = CDF = BEF$. The complete alias structure is:

$$I = BCDE = ACDF = ABEF,$$

Table 2 Design matrix for a 2^{6-2} design

Experiment		A	B	C	D	E	F
1	$abcdef$	1	1	1	1	1	1
2	abc	1	1	1	-1	-1	-1
3	abd	1	1	-1	1	-1	-1
4	$abef$	1	1	-1	-1	1	1
5	$acdf$	1	-1	1	1	-1	1
6	ace	1	-1	1	-1	1	-1
7	ade	1	-1	-1	1	1	-1
8	af	1	-1	-1	-1	-1	1
9	$bcde$	-1	1	1	1	1	-1
10	bcf	-1	1	1	-1	-1	1
11	bdf	-1	1	-1	1	-1	1
12	be	-1	1	-1	-1	1	-1
13	cd	-1	-1	1	1	-1	-1
14	cef	-1	-1	1	-1	1	1
15	def	-1	-1	-1	1	1	1
16	(1)	-1	-1	-1	-1	-1	-1

$$A = ABCDE = CDF = BEF,$$
$$B = ABCDF = CDE = AEF,$$
$$C = ABCEF = BDE = ADF,$$
$$D = ABDEF = BCE = ACF,$$
$$E = ACDEF = BCD = ABF,$$
$$F = BCDEF = ACD = ABE,$$
$$AB = ACDE = BCDF = EF,$$
$$AC = ABDE = BCEF = DF,$$
$$AD = ABCE = BDEF = CF,$$
$$AE = ABCD = CDEF = BF,$$
$$AF = ABCDEF = CD = BE,$$
$$BC = ABDF = ACEF = DE,$$
$$BD = ABCF = ADEF = CE,$$
$$ABC = ADE = BDF = CEF,$$
$$ABD = ACE = BCF = DEF.$$

Upon assuming that three-way and higher-order interactions are negligible, we have "clean" main effects and aliasing of two-way interactions with each other. This is a *Resolution IV* design.

Design resolution is categorized as follows:

1. *Resolution III*: main effects are aliased with two-way interactions.
2. *Resolution IV*: main effects are aliased with three-way interactions and two-way interactions are aliased with each other.
3. *Resolution V*: main effects are aliased with four-way interactions and two-way interactions are aliased with three-way interactions.

It is not necessary to generate the entire alias structure to ascertain the resolution of a design because resolution is simply the shortest word length in the generator sequence.

Before proceeding to an example, it is important to note that one should be very careful in the selection of the original defining relationships. It is quite possible to try different defining relationships in the 2^{6-2} design. What about $E = ABCD$ and $F = ACD$ rather than $E = BCD$ and $F = ACD$? The implications of the new assignment with $E = ABCD$ are the design generator sequence $I = ABCDE = ACDF = BDE$, which produces a less efficient Resolution III design.

Tables that provide the highest resolution fractional designs for a given number of runs and factors are available. Extensive tables are listed in the National Bureau of Standards [12] and a shorter list for up to seven factors in eight runs and 15 factors in 16 runs appears in Lochner & Matar [9]. There are also statistical software packages that will generate 2^{k-p} designs. Two relatively inexpensive Windows software packages, ECHIP and SAS JMP, generate the designs, provide the alias structure, and include analysis options. Any statistical package that performs **regression** analysis may used for actual analysis if the user specifies only one member per alias set (*see* **Software, Biostatistical**).

An Example in Preformulation

A 2^{5-1} design was employed to determine the effect multiple components of excipients exerted upon drug stability. Four excipient components and the amount of water added to a wet granulation technique comprised the factors of interest. The example appears in [8]. The first four factors were designated to be filler (lactose, mannitol), lubricant (stearic acid, magnesium stearate), disintegrant (starch, Avicel), and binder (PVP, gelatine). The fifth factor was amount of water added (0, 3%w/w). Samples were prepared and stored at 50°C for 4 weeks to determine the effects of the factors on drug stability. The design and data are given in Table 3, where -1 and 1 correspond to the first and second of the pairs specified in the description of the factors.

Leuenberger & Becher [8] constructed the design by hand with defining relationship $E = ABCD$, that is, the designated level changes for the amount of water added was determined by the four-way interaction of the remaining factors. A Yates analysis scheme (*see* **Yates's Algorithm**) was used to calculate effects.

For the sake of illustration, an example of analysis by SAS JMP is included in the present material. First, the design was created by choosing the Design Experiment Option in the Tables menu of JMP. The two-level designs option was selected to create the Resolution V 16-run design. The design was augmented with the data and we proceeded with the analysis. These data have 15 available **degrees of freedom**, all of which are used to estimate the five

Table 3 Design and results of a drug stability study

Sample	Filler (A)	Lubricant (B)	Disintegrant (C)	Binder (D)	Water (E)	Yield
1	−1	−1	−1	−1	1	59.6
2	1	−1	−1	−1	−1	86.4
3	−1	1	−1	−1	−1	95.0
4	1	1	−1	−1	1	97.0
5	−1	−1	1	−1	−1	83.4
6	1	−1	1	−1	1	53.8
7	−1	1	1	−1	1	93.7
8	1	1	1	−1	−1	99.7
9	−1	−1	−1	1	−1	54.1
10	1	−1	−1	1	1	45.8
11	−1	1	−1	1	1	92.8
12	1	1	−1	1	−1	96.1
13	−1	−1	1	1	1	53.6
14	1	−1	1	1	−1	64.7
15	−1	1	1	1	−1	94.0
16	1	1	1	1	1	96.3

main effects and the 10 two-way interactions. For this reason, we examine the relative size of the effects through the normal plot option (see [1]). The plot is depicted in Figure 1, where sizable effects are labeled. It is evident that lubricant, binder, and water appear to be significant. This is confirmed through the results of an **analysis of variance** on the reduced three-factor model, which is given in Table 4. The cube plot in Figure 2 is a nice graphical depiction of the results.

In summary, stability of drug product is sensitive to moisture, magnesium stearate is clearly superior to stearic acid, and PVP is somewhat preferable to gelatine. Fillers and disintegrants do not affect

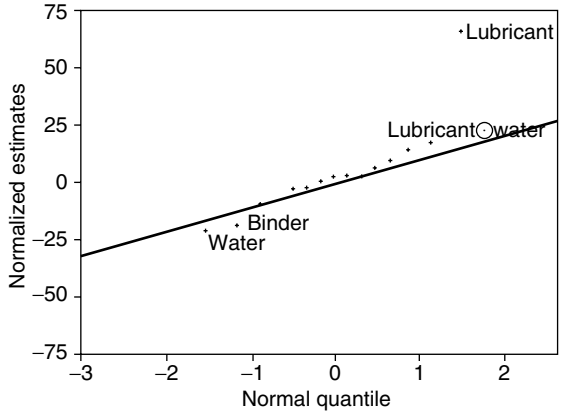

Figure 1 Normal plot of effects

Table 4 Analysis of variance table

Source	df	Sum of squares	F ratio	Pr > F
Lubricant	1	4329.6400	190.0910	<0.0001
Binder	1	316.8400	13.9107	0.0047
Water	1	408.0400	17.9148	0.0022
Lubricant⊙binder	1	216.0900	9.4873	0.0131
Lubricant⊙water	1	313.2900	13.7549	0.0049
Binder⊙water	1	100.0000	4.3905	0.0656

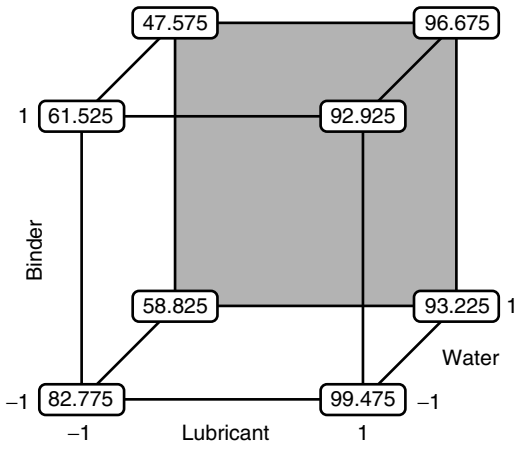

Figure 2 Cube plot of significant factors

the stability and hence the formulator may choose among the four combinations of these two components. Finally, the binder effect depends heavily on

the lubricant used. Use of PVP is clearly indicated in the presence of stearic acid, but either PVP or gelatine provide desirable results in the presence of magnesium stearate. The investigators chose to recommend lactose, mannitol, starch, Avicel, PVP, and magnesium stearate as suitable excipients.

Augmenting Designs to Separate Confounded Effects

Of course, one can visualize instances in which it is impossible to single out aliases that are the real contributors. One of the very important advantages of using a well-designed fractional factorial is that an initial design may be subsequently augmented with additional runs, if necessary, to separate effects without starting over. For example, suppose one alias set of two-way interactions is significant and the aim is to separate the confounded effects in this set. An economic route for separation of m effects is to augment the experiment with a minimum number of $(m-1)$ *complementary runs*. Most software packages that generate designs include an option to augment the design. Details on selection of runs for augmentation are given in [16].

If one wishes to separate some effects in all alias sets, a *foldover design* should be considered. Foldovers are most commonly used to separate two-way interactions from main effects in Resolution III designs. A foldover design is the complementary fraction of the original design. Foldovers are easily generated by simply reversing the signs of the main effects in the initial design.

The alias structure of the combined design follows from examination of the alias structures in complementary fractions. There are 2^p fractions of size 2^{k-p} available from a 2^k design. These fractions may be aligned into complementary pairs. To illustrate, consider a 2^{5-2} design generated with $D = AB$ and $E = AC$, and consequently, $I = ABD = ACE = BCDE$. Three additional fractions may be generated with $D = -AB$ and $E = AC$, $D = AB$ and $E = -AC$, and $D = -AB$ and $E = -AC$. The complete set of four fractions has defining relationships:

$$I = ABD = ACE = BCDE, \quad (1)$$

$$I = -ABD = ACE = -BCDE, \quad (2)$$

$$I = ABD = -ACE = -BCDE, \quad (3)$$

$$I = -ABD = -ACE = BCDE. \quad (4)$$

The generator sequence of a foldover design is such that the sign of words with an odd number of letters is reversed and the sign of words with an even number is the same. Thus, designs (1) and (4) and (2) and (3) are complementary. The contrasts in the A columns of designs (1) and (4), say δ_{A1} and δ_{A2}, estimate $A + BD + CE$ and $A - BD - CE$, respectively. Thus, $(\delta_{A1} + \delta_{A2})/2$ and $(\delta_{A1} - \delta_{A2})/2$ respectively estimate A and $BD + CE$. All other main effects are also separated from their aliased two-way interactions.

Other Fractional Factorial Designs

So far the discussion has been of two-level designs with the number of runs equal to a power of two. There are situations for which this is not the most efficient choice. This is particularly true when one has interest in only main effects such as in **screening** experiments with a large number of factors. For example, suppose one has 10 factors and wishes to detect which subset of these factors is really influential in the outcome of interest. The minimum number of runs necessary for the experiment is 11. The smallest standard design one could use would be a 2^{10-6} design with 16 runs. There are alternative *saturated* designs that may be used. A design for k factors is saturated if the number of runs is $N = k + 1$. Two types of saturated designs are Plackett–Burman and simplex designs.

Plackett–Burman Designs

Plackett–Burman designs are saturated designs for k two-level factors with $N = k + 1$, where N must be a power of four (*see* **Response Surface Methodology**). These designs are equivalent to the standard 2^{k-p} designs when N is a power of two. Construction of Plackett–Burman designs is based on the theory developed around *Hadamard* matrices, which are simply orthogonal matrices with all elements being equal to ± 1s (see [6]). Plackett–Burman designs may be folded over, but the confounding relationships are not easy to obtain. Details are discussed in [4].

Simplex Designs

Simplex designs are also saturated, but N is not restricted to be a multiple of four as in the Plackett–Burman designs. However, the simplex design

only applies to quantitative factors. These designs are often used as sequential searching algorithms to locate optimal domains for a full experimental study. A simplex is a k-dimensional figure constructed so that each vertex has the same Euclidean distance to all other vertices. In two dimensions, a simplex is simply an equilateral triangle. Natural design points are generated with specialized scaled factors that satisfy the equidistant vertices restriction. If we let x_{ij} denote the level of the ith scaled factor on the jth run, then the level of the ith natural factor on the jth run is $u_{ij} = u_i + \delta_i x_{ij}$, where u_i is a selected initial value and δ_i is a selected step size. A four-factor simplex design from Carlson [3] is given in Table 5. The background and application of simplex designs is thoroughly discussed in [2] and [11] (*see* **Simplex Models**).

D-optimal Designs

In general, **optimal designs** are designs for which **variances** of estimated effects are as small as possible. The variance of the vector of estimated effects is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, where $\mathbf{X}$ is the model matrix of the design and σ^2 is the experimental error. Thus, variances may be controlled through selection of $\mathbf{X}$. One mathematical criterion is to maximize the determinant of $\mathbf{X}'\mathbf{X}$, which in turn minimizes the determinant of the inverse. This criterion is deemed D-optimality. Other optimality criteria exist, such as A-, E-, and G-optimality, but in most cases all criteria produce equivalent designs. Kiefer [7] is an excellent source for a rigorous understanding of optimality.

All of the designs we have discussed thus far are D-optimal and can be constructed by hand. In general, D-optimal designs must be constructed by special computer algorithms. In particular, the exchange algorithm developed by Mitchell [10] is widely used in software packages. To obtain a D-optimal design, the number of runs desired, as well as the terms one wishes to estimate, must be specified. The algorithm initially randomly selects a set of N runs. Then an iterative procedure sequentially adds and deletes runs until the determinant of $\mathbf{X}'\mathbf{X}$ is maximized.

A general D-optimal approach is necessary if: (i) some combinations of factors are not reasonable (these may be excluded in algorithmic design searches); (ii) a saturated design with some interactions is desired; (iii) one would like to add complementary runs; and (iv) factors have mixed levels. It must be stressed that the designs obtained are only optimal if the model specified is accurate. The Windows packages ECHIP and SAS JMP use the D-optimality criterion for some choices of designs. A very good software package for choice of optimality criteria, choice of algorithms, and multiple choices of designs is available in PROC OPTEX of the SAS/QC module.

Table 5 A four-factor simplex design

Experiment	Addition time	Reaction time	Ratio of reagents	Amount of catalyst
1	60	180	3.00	0.400
2	32	167	2.78	0.378
3	54	125	2.78	0.378
4	54	167	2.07	0.378
5	54	167	2.78	0.300

References

[1] Benski, H.C. (1989). Use of a normality test to identify significant effects in factorial designs, *Journal of Quality Technology* **21**, 174–178.

[2] Burton, K.W. & Nickless, G. (1987). Optimization via simplex: Part I. Background, definitions, and a simple application, *Chemometrics and Intelligent Laboratory Systems* **1**, 135–149.

[3] Carlson, R. (1992). *Design and Optimization in Organic Synthesis*. Elsevier, New York, pp. 225–248.

[4] Draper, N.R. (1985). Plackett-Burman designs, in *Encyclopedia of Statistical Sciences*, Vol. 6. S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 754–758.

[5] Fisher, R.A. (1942). The theory of confounding in factorial experiments in relation to the theory of groups, *Annals of Eugenics* **11**, 341–353.

[6] Hedayat, A. & Wallis, W.D. (1978). Hadamard matrices and their applications, *Annals of Statistics* **6**, 1184–1238.

[7] Kiefer, J. (1975). Optimal designs: variation in structure and performance under change of criterion, *Biometrika* **62**, 277–288.

[8] Leuenberger, H. & Becher, W. (1975). A factorial design for compatibility studies in preformulation work, *Pharmaceutica Acta Helvetica* **50**, 88–91.

[9] Lochner, R.H. & Matar, J.E. (1990). *Designing for Quality*. ASQC, Milwaukee, pp. 77–111.

[10] Mitchell, T. (1974). An algorithm for the construction of D-optimal experimental designs, *Technometrics* **16**, 203–210.

[11] Morgan, E., Burton, K.W. & Nickless, G. (1990). Optimization using the modified simplex method, *Chemometrics and Intelligent Laboratory Systems* **7**, 209–222.

[12] National Bureau of Standards (1961). Fractional factorials at two or three levels, *Applied Mathematics Series*, No. 58. Government Printing Office, Washington.

[13] Nordbrook, E. (1992). Statistical comparison of stability study designs, *Journal of Biopharmaceutical Statistics* **2**, 91–113.

[14] Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. Wiley, New York, pp. 272–290.

[15] Raktoe, B.L., Hedayat, A. & Federer, W.T. (1981). *Factorial Designs*. Wiley, New York, pp. 167–204.

[16] Saha, G.M., Raktoe, B.L. & Pesotan, H. (1982). On the problem of augmented fractional factorial designs, *Communications in Statistics, Series A* **11**, 2731–2745.

Bibliography

Box, E.P. & Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.

Box, E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters*. Wiley, New York.

Dey, A. (1985). *Orthogonal Fractional Factorial Designs*. Wiley, New York.

Federer, W.T. & Raktoe, B.L. (1983). Fractional factorial designs, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 189–196.

Gunst, R.F. & Mason, R.L. (1991). *How to Construct Fractional Factorial Experiments*. ASQC Press, Milwaukee.

Khuri, A.I. & Cornell, J.A. (1987). *Response Surfaces*. Marcel Dekker, New York.

FRANCES P. STEWART

Frailty

Introduction

Conceptually, the frailty models are similar to the mixed models, so that conditional on some random variable (which in **survival** data is the term denoted frailty), the observations are independent. Unconditionally, that is, when the frailty is integrated out, the observations are dependent. Thus the frailty generates the dependence between the times. A review is in [13]. Frailty models have been used for univariate data to extend parametric models and to understand the effect of heterogeneity (not accounting for important **covariates**), but they are more interesting in the multivariate case, where the frailty approach makes a completely new class of models.

Basically, there are four types of **multivariate survival** data where frailty models are relevant. The first type is the time to some event (e.g. death) for several individuals, related by family membership (*see* **Familial Correlations**), marriage, exposure to some agent, and so on. Secondly, there are failures of several similar physically related components, like right/left eye or right/left kidney on a single individual. Thirdly, there are recurrent events (*see* **Repeated Events**), where the same event, like myocardial infarction (*see* **Cardiology and Cardiovascular Disease**), **epileptic** seizure, childbirth, or car **accident**, can happen several times for an individual. The fourth type is a **repeated measurements** type, typically the result of a designed experiment, where the time to the same event is studied on multiple occasions for the same individual. There are two further types of multivariate survival data where these models are slightly less relevant. First, there is the study of different events on a single individual, like the times to complication for the eyes, nerves, and kidneys for a diabetic patient. Finally, there are **competing risks** (data on cause of death), where the frailty models are relevant probability models, but where the basic parameters cannot be identified, making them less relevant for statistical inference [13].

These models may be applied in biostatistics to study dependence in lifetimes as well as time to onset of specific diseases in order to evaluate the **clustering** within families. Also, they can be used to quantify the subject differences for recurrent events and may be applied in **actuarial** science to describe a joint insurance for a married couple or for updating the risk of car accidents on the basis of the experience that accumulates for each driver as time goes by. The aim of using frailty models can either be to study the dependence *per se* or to account for the dependence in the evaluation of the effect of explanatory factors (and its uncertainty). Using a model with dependence, with as well as without a given covariate (say, describing a specific gene in a **twin** study), allows for discussing the extent to which the covariate "explains" the dependence seen between the times.

The term frailty was first introduced for univariate data by Vaupel et al. [32] to illustrate the consequences of a lifetime being generated from several sources of variation. However, frailty models have actually been used earlier for recurrent events [9] (*see* **Accident Proneness**) and for describing the dependence between lifetimes of several individuals [6]. To illustrate the idea, consider family data as a standard setup. Suppose there are n independent families, indexed by i. Let $j = 1, \ldots, k$ denote the number of members within families. The number of members could vary between families, without making the problem more complicated. The frailty, say Y_i, is specific to the family. The key assumption is that the lifetimes $(T_{i1}, \ldots, T_{ik})$ are conditionally independent given the family's frailty. Technically, this is obtained by assuming that the **hazard** is

$$Y_i \lambda(t),$$

where t denotes age, and $\lambda(t)$ is a function describing the age dependence. This can be generalized to include known covariates, say a p-vector z_{ij} for individual (i, j), giving a conditional hazard function of $Y_i \exp(\beta' z_{ij})\lambda(t)$. By assigning some distribution to Y_i and integrating it out, we have created a multivariate survival model with dependence between the coordinates.

The frailty is an unobservable quantity. For small groups, we can only obtain limited information on the individual value by observing the time of death or event respectively (*see* **Censored Data**). But we can evaluate the frailty variability by studying a population with many groups.

While most of the literature on frailty models is of the "common frailty" (or "shared frailty") type described above, where all members of a group have the same (constant) value Y_i of the frailty, this may not fully capture the complexity of relationship in all cases. There are a number of ways where the

models can be extended to allow for more general dependence structures; see the section on "Extension to Multivariate Frailty".

Comparison to the Variance Components Model

One can ask why the **normal distribution** models are not applied. There are many simple results and a broad experience with these models. However, they are not well-suited for survival data for three reasons.

1. Data are often censored, meaning that for some observations, it is only known that they exceed some given values, for example, the lifetime of a person alive today is only known to be at least his current age. This makes it difficult to apply a standard **variance components** model, because multivariate distribution functions are needed for the calculations, and simple formulas for them are not available in the normal case.
2. For survival data, it also makes sense to condition on the history up to a certain time point, giving **truncated** data, and this is not well-suited to the normal distribution.
3. The normal distribution gives a very bad fit to survival times, as lifetimes of human beings are left-**skewed** positive variables with a very high variation, and the normal distribution is symmetric and not concentrated on the positive numbers. The standard approach to get positive variables is to apply a normal distribution after a logarithmic **transformation**, but that makes the original variable right skewed, rather than left skewed.

Four other aspects, however, make the analysis of **random effects** more complicated for survival data.

1. The normal distributions satisfy very simple mixture properties, as the sum of two independent normally distributed variables is again normally distributed. The normal distributions are the only distributions with finite variance satisfying this property together with the property that any such distribution can be transformed into any other normal distribution by a linear transformation. The mixture results behind other random effects models necessarily are more complicated. In the frailty model case, both the **gamma** distributions and the positive stable distributions have interesting theoretical properties.
2. Balanced variance components model can be analyzed very simply, by means of the **analysis of variance** (ANOVA) decomposition of the sum of squares, but due to censoring, survival data are rarely balanced, and thus simple analyses are seldom possible.
3. Whereas rather general dependence structures are possible for variance components models, both including many different components and including general linear functions of a few components (**random coefficient** regression models), similar general models are not yet available for survival data; see the section "Extension to Multivariate Frailty".
4. In the normal case, it is simple and natural to evaluate the dependence by means of either the values of the variances or by the **correlation** coefficients of the observations. It is more difficult to quantify the dependence in the survival case.

Distributional Assumptions

Various choices are possible for the distribution of the frailty term. Most applications use a gamma distribution, with density $f(y) = \theta^\delta y^{\delta-1} \exp(-\theta y)/\Gamma(\delta)$. In most models, the scale parameter is unidentifiable, and therefore it is necessary to let $\delta = \theta$ during estimation, giving a mean of 1 and a variance of $1/\theta$ for Y. Let D_{ij} be an indicator of death of individual (i, j). Then, the observed (marginal; integrating over the frailty) likelihood (neglecting index i) is

$$\begin{aligned}
&\int f(y) \Pr(D_1, \ldots, D_k, T_1, \ldots, T_k | y)\, \mathrm{d}y \\
&= \int \frac{\theta^\theta y^{\theta-1} \exp(-\theta y)}{\Gamma(\theta)} \\
&\qquad \times \prod_{j=1}^{k} [y\lambda_j(t_j)]^{D_j} \exp\left[-y \int_0^{t_j} \lambda_j(s)\, \mathrm{d}s\right] \mathrm{d}y \\
&= \frac{\theta^\theta}{\Gamma(\theta)} \prod_{j=1}^{k} \lambda_j(t_j)^{D_j} \int y^{\theta-1+D.} \exp[-y(\theta + \Lambda.)]\, \mathrm{d}y \\
&= \frac{\theta^\theta \Gamma(\theta + D.)}{(\theta + \Lambda.)^{\theta+D.} \Gamma(\theta)} \prod_{j=1}^{k} \lambda_j(t_j)^{D_j}, \qquad (1)
\end{aligned}$$

where $D_\cdot = \Sigma_j D_j$ and $\Lambda_\cdot = \sum_{j=1}^{k} \int_0^{t_j} \lambda_j(s)\,\mathrm{d}s$. The gamma distribution has the further advantage that the conditional distribution of Y, given the survival experience in the family – the integrand in the penultimate expression above – is also gamma, with the shape parameter increased by the number of deaths in the family, that is, with parameters $(\theta + D_\cdot, \theta + \Lambda_\cdot)$ instead of (θ, θ). In a similar manner, the joint survival function can be derived as

$$S(t_1, \ldots, t_k) = \left[\sum_{j=1}^{k} S_j^{1-\theta}(t_j) - (k-1) \right]^{-1/(\theta-1)}, \tag{2}$$

where $S_j(t)$ is the marginal survival function for individual j. Using the marginal survivor functions offers an alternative parameterization of the model as further discussed in the section, "Conditional and Marginal Models".

A positive stable distribution of Y has other nice probabilistic properties [12]. It has one parameter (the stability index) α, $\alpha \in (0, 1]$, where $\alpha = 1$ corresponds to independence and α near 0 corresponds to maximal dependence. If $\lambda(t)$ corresponds to a **Weibull distribution** of shape parameter γ, the unconditional distribution of the lifetime is also Weibull, but of shape $\alpha\gamma$. This result is probably the closest we can come to the variance components model, where the normal distribution appears on all stages of the model. The change from γ to $\alpha\gamma$ corresponds to increased variability. If there are covariates in a **proportional hazards** model, and Y follows a positive stable distribution, the unconditional distributions also show proportional hazards (unlike the gamma model), but the regression coefficients are changed from β to $\alpha\beta$.

Basically, any other distribution on the positive numbers can be applied, but the probability results are not equally simple. The multivariate distribution can be simply formulated by means of the derivatives of the Laplace transform $L(s) = \mathrm{E}\exp(-sY)$. The general density is $(-1)^{D_\cdot} L^{(D_\cdot)}(\Lambda_\cdot) \prod_{j=1}^{k} \lambda_j(t_j)^{D_j}$, where $L^{(p)}(s)$ is the p-th derivative of $L(s)$. The gamma distributions and the positive stable distributions can be unified in a three parameter family; see [11, 31]. The **inverse Gaussian distributions** are also included in this family. The family is called the power variance function (PVF) model because it is characterized by the variance being a power function of the mean, when considered as a natural **exponential family**. The positive stable frailty distributions lead to high dependence initially, whereas the gamma distributions lead to high late dependence [13]. The inverse Gaussian distributions are intermediate. From a practical point of view, it may be difficult to discriminate between the various distribution families because data may be insufficient to estimate two dependence parameters, but from a theoretical point of view, it is interesting to study the different nice properties that the families offer. Another argument for using the more general model is that (particularly with heavily censored data) the estimated degree of dependence is sensitive to the model assumed and thus the use of one of the simpler models may suggest a too precise value for the dependence; instead the two-parameter model may better capture the high variation due to the limited knowledge that follows from the censoring.

Furthermore, **lognormal distributions** have been suggested for the frailty; this allows the use of **restricted maximum likelihood** (REML)-like procedures [21]. The inverse Gaussian and the lognormal distributions are reasonably similar to each other. Oakes [25] reviews various frailty distributions.

Univariate Models

Initially, the frailty models were used to illustrate consequences of hidden heterogeneity, that is, **risk factors** being unknown or unobserved for independent (i.e. univariate) data. In this model, the total variation of a lifetime is split into the effect of known covariates, unknown covariates, and randomness. The effect of unknown covariates is modeled by a proportional hazards frailty model, and the randomness is modeled by the hazard function. One consequence is that even though individual hazard functions increase, the hazard function in the population can decrease due to differences between the individual hazard functions. With a gamma frailty distribution, a conditional proportional hazards model, using a known explanatory variable z, is marginally no longer of the proportional hazards form. Instead, the marginal hazard for a single individual j is

$$\mu(t, \mathbf{z}_j) = \frac{\lambda(t)\exp(\boldsymbol{\beta}'\mathbf{z}_j)}{1 + \theta\Lambda(t)\exp(\boldsymbol{\beta}'\mathbf{z}_j)}, \tag{3}$$

the denominator reflecting the "survival of the fittest" effect, that is, the differential survival implies removal

of the high-risk subjects over time. Although in the presence of known covariates, it is, in principle, possible to estimate the frailty variance using only survival-time data on independent individuals, this estimator depends critically on the proportionality of the hazards conditional on the frailty, an assumption that is unlikely to be strictly true in practice and is inherently untestable. It can be discussed as to the extent to which it should be possible to discriminate between unknown factors and plain randomness. This is analogical to the normal distribution variance components models, where it is not possible to discriminate between variation between and within individuals, unless there are several observations for each individual, with or without known covariates. The positive stable Weibull model is a similar model for survival data. In this model, it is not possible with univariate data to separately determine the influence of unknown covariates and of plain randomness. Aalen [2] reviews univariate frailty models.

Conditional and Marginal Models

Instead of having the hazard in the conditional distribution $\lambda(t)$ as starting point, one may use the hazard in the marginal distribution, say $\mu(t)$, as basis (*see* **Marginal Models for Multivariate Survival Data**). This function has the advantage that it is directly observable, and therefore can be estimated in several different ways. This is particularly relevant when the dependence is not the object of study but a nuisance that has to be accounted for. Such models are considered with the assumption that the marginal distributions are of a specific form, possibly after a transformation. If the marginals are **uniform** (0,1), the multivariate distribution is said to be of **copula** form [8]. In this way, one can separate the inference into that regarding the marginal distribution and that regarding the dependence structure. Liang et al. [20] compare the conditional and marginal approaches. The positive stable frailty distribution is the only one where the two approaches are identical for the proportional hazards model.

Accelerated Failure Time Models

In the case of Weibull models, the frailty models can also be given an **accelerated failure time** formulation, so that the lifetime has a multiplicative expression as $Y^{-1/\gamma}T_0$, where T_0 has a Weibull distribution of shape γ. With covariates, the formula is $Y^{-1/\gamma}T_0\exp(-\eta' z)$, where the relation to the coefficients in the hazards formulation is $\eta = -\beta/\gamma$. In the positive stable frailty model, this implies that the hazard regression parameter β differs between the conditional and the marginal distribution, but the accelerated failure time regression parameters η are the same, and thus they make a better parameterization [16].

Quantifying Dependence

A key problem for quantifying dependence is that standard nonparametric survival analysis assigns no relevance to the variance of the lifetime, making it irrelevant to evaluate correlations and variance components. Some authors have used the variance of the frailty term as a measure of that variance component, but this is inconvenient, as can be illustrated in the Weibull model of shape γ, say with hazard $Y\gamma t^{\gamma-1}$. Conditional on Y, this has mean of $Y^{-1/\gamma}\Gamma(1+1/\gamma)$, say $c_1(\gamma)Y^{-1/\gamma}$ and variance $Y^{-2/\gamma}[\Gamma(1+2/\gamma)-\Gamma(1+1/\gamma)^2]$, say c_2 $Y^{-2/\gamma}$ giving the unconditional variance $\mathrm{Var}(T) = c_1^2\ \mathrm{Var}(Y^{-1/\gamma}) + c_2\mathrm{E}(\mathrm{Y}^{-2/\gamma})$, and a correlation of $\{c_1^2\mathrm{Var}(Y^{-1/\gamma})\}/\mathrm{Var}(T)$ between two individuals with common Y. This shows that the inverse **moments** of Y are more relevant than the ordinary moments. Much simpler formulas are obtained for the logarithm to the time, where the conditional variance is $\pi^2/(6\gamma^2)$, the unconditional variance is $\{\mathrm{Var}(\log Y)+\pi^2/6)\}/\gamma^2$, and the correlation is $\mathrm{Var}(\log Y)/\{\mathrm{Var}(\log Y)+\pi^2/6)\}$, independently of γ, and this may not be well described by the moments on the ordinary scale.

For the nonparametric models, it is more relevant to consider measures that are unchanged by transformations of the time axis, for example, Kendall's coefficient of concordance τ (*see* **Rank Correlation**). This is defined as $\mathrm{Esign}\{(T_{11}\text{-}\mathrm{T}_{21})(T_{12}\text{-}\mathrm{T}_{22})\}$, where (T_{11},T_{12}) and (T_{21},T_{22}) are the lifetimes for two pairs from the distribution. Another possibility is the grade correlation (the theoretical measure underlying the **Spearman** correlation), which has the advantage of an interpretation similar to the standard product moment correlation.

Estimation

The first estimation method for multivariate data with covariates was suggested by Clayton and Cuzick [7], but most applications have rather used an **EM algorithm** [17]. It is also possible to maximize the observed **nonparametric likelihood** function, that is, where the frailties have been integrated out, but including a parameter for each time of event [13]. This method has the advantage of directly giving a variance estimate for all parameters [4]. Instead of using the conditional hazards, one may use the marginal hazards for this evaluation, which has the advantage that the dependence parameters and the hazard parameters are closer to be stochastically independent. An alternative is a **penalized likelihood** approach [29].

The gamma and lognormal shared frailty models can be fitted by means of **S-Plus** [29]. There is no other commercially available software that handles frailty models with nonparametric hazard functions.

Asymptotics

The statistical inference has been performed by doing standard calculations, that is, using **maximum likelihood** estimation and using normal distributions for the estimates, with the variance evaluated as the inverse of (minus) the second derivative of the log likelihood function, the so-called observed **information**. For **parametric models**, this is easily justified. For the bivariate positive stable Weibull model also, the Fisher (expected) information has been calculated for uncensored data [26]. A similar evaluation for the gamma frailty model was made by Bjarnason and Hougaard [5].

For non- and **semiparametric** models, the standard approach also works, although it has been more difficult to prove that it does. For the gamma frailty model with nonparametric hazard, Murphy [22] has found the asymptotic distribution of the estimators and a **consistent estimator** of the asymptotic variance. These results were generalized by Parner [27]. Murphy and van der Vaart [23] show that using the observed nonparametric likelihood as a standard likelihood is correct for testing and for evaluating the variance of the dependence parameter as well as for the explanatory factors.

Extension to Multivariate Frailty

The assumption of the frailty being common for the individuals in the family (the shared frailty model) has been criticized for not being sufficiently flexible, particularly for large pedigrees. One extension is to exchange the scalar Y_i with a multivariate vector $(Y_{i1}, \ldots, Y_{ik})$ with each component connected to one observation. One such trivariate nested model based on the positive stable distribution was suggested by Hougaard [12]. A bivariate model based on dependent gamma variables was suggested by Aalen [1], but this model is complicated to handle, except for some special cases. The simplest model in which exact calculations can be done is to let the multivariate frailty vector be a linear function of independent terms, for example, the so-called correlated frailty model, $Y_{ij} = X_{i0} + X_{ij}$, with all X variables independent [33]. In this case, the Laplace transforms are easily calculated. In the bivariate case, this is not a major extension because the individual terms only modify the marginal distribution and not the dependence structure. It can, however, be useful for combining groups with different degree of dependence into one analysis, for example, monozygotic and dizygotic twins. In the multivariate case, this can be made a real generalization. Korsgaard and Andersen [18] describe an extension of this approach to pedigree data, where each possible line of descent has an associated gamma frailty component. Unfortunately, this approach leads to the number of frailty components growing exponentially with family size.

The multivariate lognormal distribution may, after a few approximations, be used for more complex pedigree data and may thus be much easier to handle than an additive frailty model [19].

Another generalization is to let the frailty change over time, as a **stochastic process** or as piecewise constant values. This allows a dependence, which is more concentrated in time than the shared frailty models, and this may be relevant. As an example, there could be cause-specific frailty terms, like one for accidents and one for heart disease. As accidents have a high influence at young ages, and heart disease at older ages, this apparently leads to a short-term dependence in total mortality.

Applications to Multivariate Data

Hougaard [13] gives a list of references to applications, of which most are to family data.

Guo [10] and Klein [17] study mortality of general families, Nielsen et al. [24] study mortality of adoptive children and their relation to the lifetimes of the biological and adoptive parents, and Hougaard et al. [14] study the dependence in the lifetimes of twins. Thomas et al. [30] study breast cancer concordance in twins using the shared gamma frailty model. Yashin and Iachine [33] study the lifetimes of twins, by means of the correlated gamma frailty model. Pickles et al. [28] study other times than lifetimes, and consider several of the extended models. Aalen et al. [3] have a dental application to the lifetimes of amalgam fillings for a number of individuals.

For recurrent events, Hougaard et al. [15] considered various frailty models for counts of epileptic seizures.

References

[1] Aalen, O.O. (1987). Mixing distributions on a Markov Chain, *Scandinavian Journal of Statistics* **14**, 281–289.

[2] Aalen, O.O. (1994). Effects of frailty in survival analysis, *Statistical Methods in Medical Research* **3**, 227–243.

[3] Aalen, O.O., Bjertness, E. & Sønju, T. (1995). Analysis of dependent survival data applied to lifetimes of amalgam fillings, *Statistics in Medicine* **14**, 1819–1829.

[4] Andersen, P.K., Klein, J.P., Knudsen, K.M. & Palacios, R.T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties, *Biometrics* **53**, 1475–1484.

[5] Bjarnason, H. & Hougaard, P. (2000). Fisher information for two gamma frailty bivariate Weibull models, *Lifetime Data Analysis* **6**, 59–71.

[6] Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65**, 141–151.

[7] Clayton, D. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion), *Journal of the Royal Statistical Society, Series A* **148**, 82–117.

[8] Genest, C. & MacKay, J. (1986). Copules archimediennes et familles de lois bidimensionnelles dont les marges sont donnees, *Canadian Journal of Statistics* **14**, 145–159.

[9] Greenwood, M. & Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society* **83**, 255–279.

[10] Guo, G. (1993). Use of sibling data to estimate family mortality effects in Guatemala, *Demography* **30**, 15–32.

[11] Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions, *Biometrika* **73**, 387–396. (Correction **75**, 395).

[12] Hougaard, P. (1986b). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678. (Correction, **75**, 395).

[13] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.

[14] Hougaard, P., Harvald, B. & Holm, N.V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881–1930, *Journal of the American Statistical Association* **87**, 17–24.

[15] Hougaard, P., Lee, M.-L.T. & Whitmore, G.A. (1997). Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes, *Biometrics* **53**, 1225–1238.

[16] Hougaard, P., Myglegaard, P. & Borch-Johnsen, K. (1994). Heterogeneity models of disease susceptibility, with application to diabetic nephropathy, *Biometrics* **50**, 1178–1188.

[17] Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm, *Biometrics* **48**, 795–806.

[18] Korsgaard, I.R. & Andersen, A.H. (1998). The additive genetic frailty model, *Scandinavian Journal of Statistics* **25**, 255–269.

[19] Korsgaard, I.R., Madsen, P. & Jensen, J. (1998). Bayesian inference in the semiparametric log normal frailty model using Gibbs sampling, *Genetics Selection Evolution* **30**, 241–256.

[20] Liang, K.-Y., Self, S.G., Bandeen-Roche, K.J. & Zeger, S.L. (1995). Some recent developments for regression analysis of multivariate failure time data, *Lifetime Data Analysis* **1**, 403–415.

[21] McGilchrist, C.A. (1993). REML estimation for survival models with frailty, *Biometrics* **49**, 221–225.

[22] Murphy, S.A. (1995). Asymptotic theory for the frailty model, *Annals of Statistics* **23**, 182–198.

[23] Murphy, S.A. & van der Vaart, A.W. (2000). On profile likelihood, *Journal of the American Statistical Association* **95**, 449–485.

[24] Nielsen, G.G., Gill, R.D., Andersen, P.K. & Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics* **19**, 25–43.

[25] Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**, 487–493.

[26] Oakes, D. & Manatunga, A.K. (1992). Fisher information for a bivariate extreme value distribution, *Biometrika* **79**, 827–832.

[27] Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model, *Annals of Statistics* **26**, 183–214.

[28] Pickles, A., Crouchley, R., Simonoff, E., Eaves, L., Meyer, J., Rutter, M., Hewitt, J. & Silberg, J. (1994). Survival models for development genetic data: Age of onset of puberty and antisocial behaviour in twins, *Genetic Epidemiology* **11**, 155–170.

[29] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

[30] Thomas, D.C., Langholz, B., Mack, W. & Floderus, B. (1990). Bivariate survival models for analysis of genetic and environmental effects in twins, *Genetic Epidemiology* **7**, 121–135.

[31] Tweedie, M.C.K. (1984). An index which distinguishes between some important exponential families, in *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, J.K. Ghosh & J. Roy, eds. Indian Statistical Institute, Calcutta, pp. 579–604.

[32] Vaupel, J.W., Manton, K.G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty of the dynamics of mortality, *Demography* **16**, 439–454.

[33] Yashin, A.I. & Iachine, I. (1995). How long can humans live? Lower bound for biological limit of human longevity calculated from Danish twin data using correlated frailty model, *Mechanisms of Ageing and Development* **80**, 147–169.

(*See also* **Multilevel Models**; **Overdispersion**)

PHILIP HOUGAARD

Framingham Study

When the Framingham Study began, the physical and intellectual apparatus available to assemble and use the data collected was, by today's standards, primitive. This article describes some of the statistical issues and how they were addressed. Topics discussed are: the study design, data handling problems, the development and uses of **logistic regression**, using repeated characterization, subsampling problems, and the injection of new questions into the study.

Study Design

The idea of the Framingham Study is, in retrospect, a fairly straightforward one [4]. A defined population, namely a **random sample** of the population of Framingham, Massachusetts, would be enlisted, examined, and at periodic intervals re-examined, to determine what changes had occurred and how these changes were related to pre-existing characteristics (*see* **Cohort Study**). The primary changes of interest were the development of the first clinical evidence of hypertensive disease, coronary heart disease (CHD), and stroke, but once the study population was under observation, any number of changes and clinical events could be examined.

A number of papers have reported the difficulties encountered in realizing this design [6, 11–13, 14]. This is not the place to recapitulate this story but one issue warrants restating. When the first examination began, the study design was not in place and even as the design developed the details changed. What started out as a purely volunteer population in 1948 was altered during recruitment into a random sample of the Framingham population aged 30–59 as of 1950, and then altered again to a mixed random sample and volunteer group. The purposes of the study were also redefined, so that the study ended up as a relatively clear-cut prospective study of cardiovascular disease.

However, what is obvious a posteriori is not always so clear a priori. For example, if one is interested in the development of hypertensive disease, why include persons with already evident hypertensive disease? If one is interested in the development of coronary heart disease, why include persons with already present coronary heart disease? The answer to the first question is that hypertension may be a factor in the development of coronary heart disease or stroke. Moreover, hypertension is, by definition, the upper end of the blood pressure distribution, so that omitting persons with hypertension amounted to truncating the blood pressure distribution – a statistical nuisance. The answer to the second question is that coronary heart disease may be a factor in the development of stroke and may influence blood pressure levels, and that neither hypertension nor coronary heart disease is necessarily well defined or static. Fortunately, the logical error implicit in excluding persons with pre-existing disease was quickly realized and corrected before the study was well under way.

After specifying the population to study and the diseases to investigate, the next question was: which characteristics should be investigated as possible factors in the development of these diseases? It is difficult to remember that at the time the study began it was not at all clear that cigarette smoking was one of those factors. In fact, the study began without obtaining smoking histories. Only part way through the first examination was this deficiency repaired. Moreover, the standard technique for measuring the concentration of serum cholesterol, the Abell–Kendall method, became available only part way through the initial examination, when it was hastily substituted for a much inferior method. Thus, the beginning of the study was characterized by patching and improvization. While this story has been detailed before, it is well to remember these facts since each of them entailed problems in assembling and analyzing the study results. These have been identified and discussed in a number of early reports [6, 11–13, 15, 16, 18].

Data Handling

It is difficult now to imagine the problems involved in handling large bodies of data when the Framingham Study began. And it was literally "handling". Data were entered on IBM punch cards, which were typically run through a counting sorter or a tabulator, two ingenious pieces of machinery which occasionally would mutilate the punch cards. The trick was to design a coding and punching scheme to put as much information on one card as possible. This entailed a high level of alertness in tabulating the information. It also meant that when the computer became available, transferring the information to tape taxed both

the ingenuity and patience of those responsible. Only gradually did the investigators become comfortable with the computer and only by slow stages were the older techniques for tabulating the data relinquished.

The other major tabulating accessory was the 3×5 card. This was used to enter information about **incident cases**. By using all four corners and the middle, and both sides of the card, a surprising amount of information could be entered. This may seem primitive, but any time one wanted to retabulate, the means to do so were right on one's desk. For those who have prompt turnaround time on their computer this may seem a trivial matter, but immediate computer access is a relatively modern phenomenon and for a long time was simply not a consideration.

Ultimately, the investigators were simply forced to move their data on to the computer, since the counting sorter and tabulator were no longer maintained properly and were finally phased out. The transitional period was painful in the extreme. When the initial computer system was replaced by a better but noncompatible system, the changeover was again painful.

These details may seem trivial but they are at the heart of analytic productivity. It is often assumed that the key to analysis is primarily statistical technique. It is not. The essential key is in the organization of data files and in the maintenance of data quality. Gradually the Framingham investigators developed files that were more and more accessible and more and more trustworthy. Such matters are now managed by software of varying sophistication and which is more or less simple to use (*see* **Software, Biostatistical**). That was not always true. At the beginning of the study it was necessary to cobble together everything necessary to assemble the study files and fit them to the needs of the study (*see* **Data Management and Coordination**).

The Introduction of Logistic Regression

Initially, a major restriction in analyzing the study results was the slow accumulation of new clinical events, but this was compounded by an inability to deal efficiently with the analysis. Suppose one wanted to consider two variables, serum cholesterol and blood pressure. Both of these were continuous variables but they could be dealt with categorically by arbitrarily dividing them into, say, high and not-high. One could, then, examine rates in the four cells obtained by cross-classifying the independent variables. Clearly, as one increased the number of classes and the number of variables – that is, as one got closer to the original data in its full detail and complexity – cross-classification was no longer of much use.

The original publication of the six-year follow-up data dealt with the **multivariate** problem by categorizing events according to whether persons were high on one independent variable, high on two, or high on three [4]. As one went up that scale, the incidence of CHD increased. However, this did not leave one with a measure of the contribution of each of the specific variables. Faced with this problem the investigators appealed to a procedure developed by **Jerome Cornfield**, which used continuous variables as continuous variables and addressed the multivariate question in terms of specific independent variables. The idea he had was the following [2].

Suppose one had two **normal** populations with different **means** but the same **variance**, say a population of cases and a population of noncases. Then it could be shown that the mixture of these two populations led to a logistic regression. In effect this transformed a classification approach, historically dealt with by **discriminant analysis**, into a **regression** approach. If one considered two **bivariate normal distributions** (the variables originally used were serum cholesterol and blood pressure) one could extend this concept to the bivariate case. It was demonstrated that the procedure was quite **robust**, at least where the variables were continuous. Ultimately, Truett et al. [24] extended the model to deal with any number of variables, although robustness in this case was less assured. Logistic regression has since become a mainstay of analysis in prospective studies generally; and while there may be misgivings about some of the uses to which it has been put [7], there is no question it has proved quite useful.

The Cornfield approach to logistic regression assumed, among other things, equal variances in the cases and noncases. Manifestly, that could not be true for dichotomous variables, since a difference in means implied a difference in variances. This difficulty was noted by **Max Halperin** and others, who showed that in the case of dichotomous variables one could sometimes arrive at absurd results using the Cornfield model [19].

In the meantime, another method for estimating the parameters of the logistic regression was

being developed by Walker & Duncan [25]. The study statisticians replaced the estimating procedure suggested by Walker & Duncan with a **maximum likelihood** approach and incorporated this modified Walker–Duncan procedure into the statistical armamentarium of the study. The estimating procedure, however, did not lead to an explicit solution and required successive iterations. The trick was to develop a method that minimized the number of iterations and protected against divergence. Over time the study statisticians were able to devise and refine their own computing program to achieve these *desiderata*. At the same time, the program gave associated statistics to assist in interpreting the results.

There was a problem, however. At the time this iterative computation became available, computer capacity was limited and computer costs per unit of work were quite high. And so the study statisticians looked for a method to estimate the parameters of the logistic regression by maximum likelihood that would minimize costs. When the problem was put to Nathan Mantel he came up with a sampling procedure which led to **unbiased** estimates using all the cases, but only a sample of noncases [22]. The method was ingenious but there was inevitably an increase in sampling error compared with using the full data set. Clearly, if the variables involved were truly normally distributed and the variances for cases and noncases not too dissimilar, then the Cornfield estimation, which was explicit and cheap to compute, was manifestly simpler and better. Before long, computer capacity increased and computer costs decreased, so that ingenuity was no longer necessary. Mantel's procedure, however, led to applications in **case–control studies**. It could be argued that case–control studies are the wrong place for such a procedure; but if the cases and the controls come from the same population, then it is clearly justifiable to treat them by this method, and if they do not come from the same population, then no analytic method is quite safe (*see* **Case–Control Study, Nested**).

One of the hazards in using logistic regression is that the user may confuse it with **linear regression**. In particular, there has been a tendency to ask and answer the question: what proportion of the variance in the dependent variable is accounted for by the independent variables, i.e. the known risk factors? Presumably, it was felt that the larger the proportion explained the more we understand about the causes of the disease – a rather unscientific notion. However, since the dependent variable was either a case or a noncase, it should be obvious that the analogy with **multivariate normal** theory might be hazardous. As Max Halperin pointed out, this is not an easily addressed question; in fact, it does not ordinarily allow for a precise answer [14].

Replication

A key element of the scientific method is replication. In epidemiology one would like to find that similar relations hold in different populations. Logistic regression was used to explore this issue in two ways. One was to apply the parameter estimates derived from the Framingham Study to a number of other populations included in the American Heart Association Pooling Project [23]. It was shown that the estimates derived from the Framingham Study closely predicted the actual incidence of coronary heart disease in all but one of these populations. The Framingham Study parameter estimates were also applied to the CHD experience found in a Yugoslav population. While the estimates for the parameters associated with the independent variables were quite similar for the two groups, the estimates for the constant term were not. Thus, the Framingham Study estimates grossly overstated the CHD incidence actually observed in Yugoslavia [20]. A similar approach was used in comparing data from Honolulu and Puerto Rico with data from Framingham [10].

Prediction

The preceding observations bear on another use made of observations from the Framingham Study. If the findings from Framingham with respect to the relationship between blood pressure, serum cholesterol, and cigarette smoking are generalizable to other US populations, then it made sense to estimate the **risk** of developing CHD as a guide to preventive programs, not only identifying who was at high risk but what their actual risk was. For that purpose the Framingham investigators constructed a set of CHD risk tables for the American Heart Association to use in their public health programs [17]. These tables emphasized two things: first, that the disease, at least so far as **prediction** was concerned, was a multifactorial disease; secondly, that persons were more or less at risk and that it was conceptually wrong to

assert that a person was either a sure case-to-be or was sure never to develop the disease. It was also a fact that for CHD, contrary, say, to lung cancer, there was no really strong predictor, but there were a number of weak predictors. Hence, the more pertinent information one had about the individual the more precisely one could estimate the risk of developing CHD. While this way of looking at disease prediction has not always prevailed in discussions of the etiology of CHD, it seems to have permeated the thinking about this disease.

Using Repeated Characterization

The study design called for repeated examinations of the Framingham cohort (*see* **Longitudinal Data Analysis, Overview**). The primary reason for that was to determine disease incidence and progression, and also to identify changes in factors thought to predispose to disease. Ultimately the interval between examinations came to be two years with a window on either side. Thus, each person returning for examination had a series of observations. How should those repeated observations be used? The method to use depended, of course, on the question to be addressed and the assumptions one was prepared to make.

Take a simple question: What was the incidence of coronary heart disease for each age–sex group? If the cohort of 5209 persons were divided, say, into five-year age groups by sex, then the incidence for each of these groups in, say, two years would be so small that the calculated rates would be subject to very large sampling variability. However, at each examination the entire cohort was, on average, two years older than it was on the previous examination. This meant that any five-year age group would be depleted of some persons (who had moved into the next age bracket) and replenished by other persons recruited from the next younger age group until the entire cohort moved out of the age range. Thus, if one redefined the cohort at each examination in terms of age and disease status, one could simply pool the observations from successive examinations to obtain average incidence rates over the period of observation. This method was extended to other characteristics that were repeatedly observed, e.g. serum cholesterol concentrations and blood pressure. Thus, it was possible to calculate the conditional incidence in a very simple fashion using all of the available data from the successive examinations.

The major assumption in this procedure was that successive subgroups of the total population had the same conditional incidence rate. This is probably true only in an approximate sense. While it was possible to test that assumption, the small sample sizes meant that the **sensitivity** of the test was obviously very low. In some sense what was being presented was an average picture for the various generations included in the cohort.

Obviously this is not the only feasible approach. It was possible to characterize the cohort at entry and follow them for some specified period. That would tell us the **predictive value** of, say, a serum cholesterol observation for the defined period of follow-up. It was also possible to use repeated measurements taking account of the fact that they were made on the same persons [26]. That would obviously give somewhat different conclusions than the pooling procedure. It would also lead to highly complex computation and to necessarily strong assumptions respecting **missing data**. One would hope that no matter how the data were looked at, the subject-matter conclusions would remain consistent. In general, that seemed to be the case, but since each analytic technique really implies a different question, consistency is not necessarily easy to parse.

Subsampling: The Selection Trap

Early in the Framingham Study a sample of the cohort was interviewed to determine what they usually ate, the primary interest being in the relationship between what they ate and their serum cholesterol levels. Since each interview took an experienced dietitian a long time, it was impracticable to interview the entire cohort. A sample was chosen consisting of a random sample supplemented by all persons with either very high or very low serum cholesterol levels. This seemed to be (and in this case, was) an efficient way to address the question at issue.

It was not recognized, however, that linear regressions of serum cholesterol level on various dietary components calculated by the usual **least squares** method would yield estimates that were **biased**, in some instance grossly biased. This, of course, was a consequence of selection on the dependent variable. Dietary data, because of the

intercorrelations among the various elements of the diet, are exceedingly difficult to interpret under the best of circumstances [9] and the **sampling frame** used only added to the difficulty.

The sampling problem was addressed by DeMets & Halperin [5]. They showed that unbiased regressions could be estimated from such data, provided that the necessary supplementary information had been collected, which fortunately had been done. Their solution has proved highly useful to persons concerned with sampling procedures.

Pouring New Wine into Old Bottles

One of the really strong points of a long-running prospective study in which the population is subject to continuing recharacterization is the opportunity to examine new questions as new ideas and new techniques arise. Formally, of course, one is supposed to initiate a study with a well-defined set of questions and methods for addressing them. However, ideas change, new methods become available, and some flexibility, if cautiously administered, is all to the good. There were many instances where the Framingham Study introduced new questions and new methods into their protocol.

From the beginning, the Framingham Study cohort was used to study a variety of chronic diseases and noncardiovascular conditions. The study staff has been receptive to proposals by investigators whose interests lay outside of cardiovascular disease but were anxious to avail themselves of the opportunities for investigations within the study cohort. In the more than 1000 publications based on the Framingham study cohort are papers on a variety of subjects, e.g. nontoxic thyroid nodules, gall bladder disease, eye disease, osteoarthritis, and cancer. But the major innovations over time were in the area of cardiovascular disease. One example may be cited.

The classical lipid measurement included in the Framingham Study was total serum cholesterol, although other lipid measurements were also included. However, serum cholesterol is not a single substance but includes components carried in lipoproteins of varying density. When it was discovered in a cooperative case–control study in which the Framingham Study participated that there was a strong negative relationship between serum cholesterol in the high density lipoproteins and coronary heart disease [1], the Framingham Study was able almost immediately to demonstrate that this was equally true prospectively [8]. In fact, this negative relationship had been demonstrated almost as early as the beginning of the Framingham Study, but discovery and acceptance are two different things. The Framingham investigators were aware of these early findings but, like others, chose to ignore them. However, the confluence of replication in the case–control studies, confirmation in a prospective study, and the application of multivariate analysis techniques which had become a standard in analyzing Framingham Study data, made it clear that this negative relationship between the concentration of HDL cholesterol and CHD was not some kind of artifact. Even with the great weight of evidence that accumulated so quickly, there was considerable covert resistance to accepting this new-old finding, which paralleled the resistance to the initial reports. Resistance was overcome, however, in part because of the authority that the Framingham Study had by then come to command.

Extensions of the Original Study

The Framingham Study has grown dramatically over the years. In 1971 the Framingham Offspring Study began consisting of children of the original cohort and the children's spouses. This consisted of 2489 males and 2646 females with examinations conducted every four years. In addition to all the examination components of the original cohort, new technologies such as carotid ultrasound, heart and brain MRIs and echocardiograms are part of the evaluations. Further, the study often leads the field in collecting new and novel risk factors such as homeostatic (clotting) factors including fibrinogen, inflammatory markers, and infection marker.

In 2002, a sample of over 3500 third generation children (Gen3 Study) and their spouses began. The Framingham Study is unique as a genetics study. Not only is there the risk factor collection (phenotypes) but there is also DNA and cell lines on many of the three cohorts (Original Cohort, Offspring Study, and Gen3 Study). The original study consisted mainly of whites of European extraction. A new substudy, the Omni Study, consisting on nonwhites is now an integral component of the study. Further, a number of ancillary studies evaluating stroke, dementia, arthritis, hearing, vision, sleep apnea, cancer, and nutrition

are major components of the study. Initially, the Framingham Study was the Framingham Heart Study.

Study Validation

The Framingham Study is an epidemiological study consisting mainly of whites of European extraction. The generalization of its results is always a major concern. Other epidemiological studies and clinical trials have repeatedly verified and validated its results (for example, [10]). Recently, a major validation study showed again the validation and transportability of its results to other populations [3]. Its results with minor calibration adjustments are valid even for the entire country of China [21].

Conclusion

The unique thing about the Framingham Study is its continuity and productivity. The continuity arises from its sponsorship since 1948, with one break, by the **National Institutes of Health**. We owe that to the persuasiveness of the first statistician involved in the study, Felix Moore, who cajoled the then National Heart Institute to embrace this study, reluctantly it must be said, and who transformed it into a clearly defined prospective study and set it on its course of methodical data collection. The productivity was a function of the close collaboration between the statisticians and the medical staff but was made possible and fruitful by careful handling and organization of the data files and continuing search for appropriate statistical techniques. It is sometimes forgotten that many epidemiologic studies have failed to realize their potential because the data that got into the medical folders remained in the medical folders. Providing adequate staff and resources for exploring the data is as important as conceiving and operating a prospective study. In this respect, the Framingham Study has been singularly fortunate for the able medical and statistical investigators that have worked together over the years. Both were necessary. Fortunately, both were available.

Closing Comment

The Framingham Study was a bold innovation in epidemiology. It remains unique among epidemiological studies. We can no longer perform such a study. Today we treat blood pressures as low as systolic blood pressure of 140 and diastolic pressures of 85. We treat total cholesterols of 200. These are correct decisions based, in part, on the results of the Framingham Study. Still they render impossible the study of natural history. The Framingham Study is unique in that is has true natural history of the development of coronary and cardiovascular disease. It has this over generations. It is truly unique and precious.

References

[1] Castelli, W.P., Doyle, J., Gordon, T., Hames, C., Hjortland, M.C., Hulley, S.B., Kagan, A. & Zukel, W.J. (1977). HDL cholesterol and other lipids in coronary heart disease: the Cooperative Lipoprotein Phenotyping Study, *Circulation* **55**, 767–772.

[2] Cornfield, J., Gordon, T. & Smith, W.S. (1961). Quantal response curves for experimentally uncontrolled variables, *Bulletin of the International Statistical Institute* **38**, 97–115.

[3] D'Agostino, R.B., Grundy, S., Sullivan, L. & Wilson, P. (2001). Validation of the Framingham coronary heart disease prediction score: results of a multivariate ethnic groups investigation, *Journal of the American Medical Association* **296**, 180–187.

[4] Dawber, T.R., Moore, F.E. & Mann, G.V. (1957). Coronary heart disease in the Framingham Study, *American Journal of Public Health* **47**, 4–24.

[5] DeMets, D. & Halperin, M. (1967). Estimation of a simple regression coefficient in samples arising from a sub-sampling procedure, *Biometrics* **33**, 47–56.

[6] Gordon, T. (1968). The Framingham Study: follow-up to the eighth examination, in *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*, Section 2, W.B. Kannel & T. Gordon, eds. National Heart Institute, Bethesda.

[7] Gordon, T. (1974). Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies, *Journal of Chronic Diseases* **27**, 97–102.

[8] Gordon, T., Castelli, W.P., Hjortland, M.C., Kannel, W.B. & Dawber, T.R. (1977). High density lipoprotein as a protective factor against coronary heart disease: the Framingham Study, *American Journal of Medicine* **62**, 707–714.

[9] Gordon, T., Fisher, M. & Rifkind, B.M. (1984). Some difficulties inherent in the interpretation of dietary data from free-living populations, *Journal of Clinical Nutrition* **39**, 152–156.

[10] Gordon, T., Garcia-Palmieri, M.R., Kagan, A., Kannel, W.B. & Schiffman, J. (1974). Differences in coronary heart disease in Framingham, Honolulu and Puerto Rico, *Journal of Chronic Diseases* **27**, 329–344.

[11] Gordon, T. & Kannel, W.B. (1968). The Framingham Study: introduction and general background, in *The*

Framingham Study: An Epidemiological Investigation of Cardiovascular Disease, Section 1, W.B. Kannel & T. Gordon, eds. National Heart Institute, Bethesda.

[12] Gordon, T. & Kannel, W.B. (1970). The Framingham, Massachusetts Study twenty years later, in *The Community as an Epidemiologic Laboratory: A Casebook of Community Studies*, I.J. Kessler & M.L. Levin, eds. Johns Hopkins Press, Baltimore, pp. 123–146.

[13] Gordon, T. & Kannel, W.B. (1972). The prospective study of cardiovascular disease, in *Trends in Epidemiology: Application to Health Service Research and Training*, G.T. Steward, ed. Thomas, Springfield, pp. 189–211.

[14] Gordon, T., Kannel, W.B. & Halperin, M. (1979). Predictability of coronary heart disease, *Journal of Chronic Diseases* **32**, 483–491.

[15] Gordon, T., Moore, F.E., Shurtleff, D. & Dawber, T.R. (1959). Some methodological problems in the long-term study of cardiovascular disease: observations on the Framingham Study, *Journal of Chronic Diseases* **10**, 186–206.

[16] Gordon, T. & Shurtleff, D. (1973). Means at each examination and inter-examination variation of specified characteristics: Framingham Study, Exam 1 to Exam 10, in *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*, Section 29, W.B. Kannel & T. Gordon, eds. DHEW Publication No. 74–478 (NIH), US Government Printing Office, Washington.

[17] Gordon, T., Sorlie, P. & Kannel, W.B. (1971). Coronary heart disease, atherothrombotic brain infarction, intermittent claudication – multivariate analysis of factors related to their incidence, in *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*, Section 27, W.B. Kannel & T. Gordon, eds. US Government Printing Office, Washington.

[18] Gordon, T. & Verter, J. (1969). Serum cholesterol, systolic blood pressure and Framingham relative weight as discriminators of cardiovascular disease, in *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*, Section 23, W.B. Kannel & T. Gordon, eds. US Government Printing Office, Washington.

[19] Halperin, M., Blackwelder, W. & Verter, J. (1971). Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches, *Journal of the American Statistical Association* **66**, 587–589.

[20] Kozarevic, D., Pirc, B., Racic, Z., Dawber, T.R., Gordon, T. & Zukel, W.J. (1976). The Yugoslavia Cardiovascular Disease Study: 2. Factors in the incidence of coronary heart disease, *American Journal of Epidemiology* **104**, 133–140.

[21] Liu, J., Hong, Y., D'Agostino, R.B., Wu, Z., Wang, W., Sun, J., Wilson, P.W.F., Kannel, W.B., Zhao, D. Predictive value for the Chinese population of the Framingham CHD Risk Assessment Tool compared with the Chinese Multi-provincial Cohort Study. *Journal of the American medical Association* 2004. **291**(21), 2591–2599.

[22] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics* **29**, 479–486.

[23] McGee, D. & Gordon, T. (1976). e results of the Framingham Study applied to four other U.S.-based epidemiologic studies of cardiovascular disease, in *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*, Section 31, W.B. Kannel & T. Gordon, eds. DHEW publication No. 76–1083 (NIH), US Government Printing Office, Washington.

[24] Truett, J., Cornfield, J. & Kannel, W.B. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham, *Journal of Chronic Diseases* **20**, 511–520.

[25] Walker, S. & Duncan, D. (1967). Estimation of the probability of an event as a function of several variables, *Biometrika* **54**, 167–179.

[26] Wu, M. & Ware, J. (1979). On the use of repeated measurements in regression analysis with dichotomous responses, *Biometrics* **35**, 513–521.

TAVIA GORDON & RALPH B. D'AGOSTINO, SR

Fraud in Clinical Trials

Scientific research has a long history of fraud [4, 10, 15]. Over 150 years ago, Charles Babbage, the far-seeing inventor of the calculating machine, established a catalog of data manipulations, which he called *trimming* (reducing the variance of the data while preserving their mean by deleting extreme observations), *cooking* (reporting only selected observations: "If a hundred observations are made, the cook must be very unhappy if he cannot pick out fifteen or twenty which will do for serving up") and *forging* (inventing data). Allegations of data tampering have been made against Ptolemy, Galileo, Newton, Dalton, Mendel, and Burt, to name just a few. R.A. Fisher's reanalysis of Gregor Mendel's data on peas is a celebrated example of the use of statistical methods to reveal abnormalities: the agreement between the observed frequencies of certain traits of the peas with the theory was too good to be plausible, which suggested that Mendel or one of his assistants had either manipulated the observations or reported only those results most closely matching theoretical expectations. The fraud perpetrated by Cyril Burt was far worse, since it seems to have involved the complete fabrication of data on identical twins supposedly separated at birth. Here again the fraud was discovered because of numeric anomalies. The number of identical twins reported by Burt (53 pairs) was too large to be plausible and, while the number of pairs increased from less than 20 to more than 50 in a series of Burt's papers, the average correlation of IQ measurements between pairs remained unchanged to the third decimal place! In recent years, scientific journals as well as the lay press have extensively debated over some cases of fraud in clinical trials, thereby fueling suspicions that fraud is a major problem in clinical research [2, 13].

Definitions

Fraud comes in many guises, including some that are well-intended. The boundary between fraud and simple carelessness is often fuzzy, although the former is characterized by a *deliberate* attempt to deceive [10]. The deliberate character of fraud may be very hard to prove in the absence of positive external evidence or confession. Data discrepancies expected as part of the research process, such as transcription errors between the source documents and the case report forms, may potentially be regarded as fraud if they occur in some systematic way or with abnormally high frequency, two circumstances that require a statistical assessment. In many cases, statistical evidence is however likely to reveal misunderstandings and unintentional errors rather than fraud [1].

In the US, the term "fraud" implies injury or damage to victims, therefore the term "misconduct" might be preferred. However "misconduct" also includes practices that fall beyond the scope of this chapter, such as plagiarism, conflicts of interest, misuse of funds, and other questionable research practices. In the UK, a Joint Consensus Conference on Misconduct in Biomedical Research held in October 1999 defined research misconduct as "behaviour by a researcher, *intentional or not*, that falls short of good ethical and scientific standards" [7]. We shall use the term "fraud" specifically to refer to *data fabrication* (making up data values), and *data falsification* (changing or eliminating data values). We are aware that this use of the word is at once far more restrictive than is implied in normal conversation, and less specific than in legal texts, but we prefer the use of this single word for brevity.

Prevalence of Fraud in Clinical Trials

Scientific fraud (in the limited sense of data fabrication or falsification) is, in all likelihood, a rare phenomenon, although other misconduct may well be common [15]. Many authors also believe that fraud is also uncommon in clinical trials [4]. A few cases were uncovered, but they attracted so much media attention that the uncritical observer may have been misled into thinking that the problem was far worse than it actually is. While there may be substantial bias in estimating the actual number of cases of fraud (because of those cases that remain unnoticed or unreported), all systematic investigations carried out to uncover fraudulent data found that the proportion of investigators who had actually committed fraud was less than 1% (Table 1).

Even though serious deficiencies were found in 11% of the audits performed by the US Food and Drug Administration (FDA) between 1977 and 1988, the "for-cause" investigations that followed revealed

Table 1 Incidence of fraud found in random audits performed by several groups

Group	Period	Number of audits	Incidence of fraud
FDA [23]	1977–1988	1955	Incidence of fraud not reported *Note*: 11% of audits showed "serious deficiencies"; 4% required a "for-cause" investigation
South West Oncology Group (SWOG) [20]	1983–1990	1751	No case of fraud = 0%
Pharmaceutical company [18]		1000	Four of 1000 = 0.4% *Note*: fraud affected 438 (6%) of 7000 patients
CALGB [19]	1982–1992	691	Two of 691 = 0.3%
Pharmaceutical company [12]	1990–1994	234	One of 234 = 0.4%

in most cases sloppiness or incompetence rather than fraud. Hence all available estimates point to a low prevalence of fraud. Against these reassuring statistics lingers the possibility that a large number of cases of fraud may have remained completely unnoticed, and that reported cases only constitute the tip of the iceberg. Although this situation remains hypothetical, there have been reports of fraud being detected and then covered up in trials sponsored by pharmaceutical companies as well as in those performed in academic settings. Additionally, audits may be ineffective at detecting fraud, as demonstrated by the fact that an investigational center that passed Cancer and Leukemia Group B (CALGB) audits was subsequently found to have problems [24]. In a recent cross-sectional survey of biostatisticians who were members of the International Society for Clinical Biostatistics in 1998, more than half of the respondents (51%) stated that they knew of a project in which fraud had occurred in the previous 10 years, while almost one-third (30%) of them had been engaged in a project in which fraud took place or was about to take place [16]. All in all we lack reliable data to estimate the true prevalence of fraud, and further prospective investigations in this matter would be very valuable.

Perception of Fraud

Shapiro claims that "scientific misconduct is common enough in investigational drug trials to be a continuing public concern" [15]. Quantitatively, the opposite seems true: fraud in clinical trials is probably rather uncommon and, as we discuss below, it is often inconsequential in multicenter settings. Fraud, however, is so much at variance with the ethics of scientific research that *any* amount of it is deemed utterly unacceptable. The concern with fraud in clinical research may in fact be due to common practices that are scientifically unacceptable, such as data dredging, *post hoc* analyses, selective reporting of the most "interesting" results, nonpublication of negative findings, etc. Although such practices do not constitute fraud in the narrow sense used in this article, they may profoundly bias the results of a study as well as their interpretation. The public may in fact be far more misguided by studies that are poorly designed, wrongly analyzed and inappropriately reported, or not reported at all, than by fraud [1]. Systematic reviews of randomized trials are very important for clinical practice and policy; when important data are not reported this can lead to bias in the overall conclusions.

Intent of Fraud

The major difference between fraud and mere error lies in the "intention to cheat" that defines fraud [10]. This difference must however be qualified by the nature of the intent, as illustrated by the following examples. Consider, first, the case of data falsification. Suppose that at some time point the diastolic blood pressure of a patient is read as equal to 96 mm Hg. If the value is reported as being equal to 100 mm Hg, the discrepancy between the value read and the value reported would constitute a case of data falsification. However, the physician who read the diastolic blood pressure may have reported 100 mm

Hg for simplicity, in recognition of the fact that blood pressure varies substantially in the same patient and that the measurement error is of the order of 5 mm Hg anyway. The reporting of a round number that closely approximates the truth would not *per se* be wrong. If, however, a value of 100 mm Hg made the patient eligible for the trial while 96 mm Hg did not, then the biased reporting would be cause for concern. However, such bias prior to randomization will not affect the overall results of a trial in terms of whether a treatment is efficacious (its internal validity). A much more serious problem arises if this biased reporting took place in a nonblinded trial in order to make the control group worse. Then the charge of fraud and the need for corrective action would be more than justified. The same arguments hold true for data fabrication. Suppose that the level of neutrophils, a required laboratory examination, were truly missing at the last visit of a patient in a certain trial. Any reported value would therefore have had to be fabricated, perhaps by simply carrying over the value of the previous visit. If this had been done in order to avoid a query from the data management center for a safety variable of secondary interest, there would be less cause for concern than if neutrophils constituted the primary endpoint of the trial.

The most serious cases of fraud are those in which there is an expectation of gain in terms of prestige, advancement or money. These cases may involve fabricating complete patients or tampering with the data in order to obtain a desirable result. These cases may also be the easiest to detect statistically, especially in **multicenter** studies, as will be discussed below.

Data Items Frequently Affected by Fraud

Some data items collected in clinical trials seem to be more prone to error and/or fraud than others, as follows.

1. *Eligibility criteria*: data may be "pushed" a little to make a patient eligible for the trial when in fact that patient does not strictly meet the criteria. Many such examples of fraud may have occurred because eligibility criteria were excessively restrictive and widening entry standards is often a good solution (*see* **Eligibility and Exclusion Criteria**).
2. *Repeated measurements*: when the same measurements are requested repeatedly over time (such as, for instance, a battery of laboratory examinations), data may be "propagated" from the previous visit if the measurements are missing for a particular visit. Such imputation of missing values should be reported clearly. Imputation methods may be appropriate at the time of analyzing the data, not at the time of making the observations.
3. *Adverse events*: adverse events are likely to be underreported by some investigators (although such underreporting may reveal a lack of interest or differences in interpretation rather than fraud).
4. *Compliance data*: these data are notoriously unreliable if they are based on the number of medications returned ("pill counts"). Whenever compliance information is deemed important, it is advisable to use objective measurements based on blood or urine tests (*see* **Compliance Assessment in Clinical Trials**).
5. *Patient diaries*: a number of cases of data fabrication have been detected through the color and texture of the pen supposedly used on successive days by the patient, the patient's handwriting, etc. The reliability of information collected in patient diaries can often be questioned.

Preventing Fraud by Making Trials Simple

Some types of fraud could be prevented through a drastic simplification of randomized clinical trials. Two measures might be particularly effective in this respect: a simplification of the eligibility criteria and a reduction in the amount of data requested. These two measures are feasible and desirable in a surprisingly large number of clinical trials, and neither jeopardizes the validity of the trial results. For example, very detailed measurement of large amounts of biochemical tests in late-phase trials is usually unnecessary. In trials requiring prolonged observation of each patient, the follow-up can generally be kept as simple and no more frequent than in routine clinical practice. Simplicity is essential in trials conducted with a public health intent, especially when these are large, but it can be justified even in trials conducted as part of a new drug development program, for there is no regulatory requirement that pivotal trials for drug approval be especially complex. However, the risk of failing

to get approval for a new drug along with the fear of potential litigation may dominate all other considerations of cost or efficiency, and as a result clinical trials may end up being excessively complex. The growing number of regulations governing the conduct of clinical trials, even with approved drugs, may also have the unintended consequence of making trials ever more complex. As mentioned above, such complexity may be counterproductive and may create a situation where an investigator is tempted to fabricate data.

Preventing Fraud by Allowing Missing Data

There is obviously no excuse for making up data, but the temptation will be great for investigators to find ways of avoiding long lists of queries in trials conducted in a fastidious way. It is the responsibility of a competent trial organization to make sure that investigators are not submitted to excessive requests for data clarification. **Missing data** occur in the real world, and thus they should generally be tolerated in clinical trials (except, of course, for the primary endpoint of the trial). While complete data are undoubtedly better than missing data, attempting to collect too much data, and repeatedly demanding complete data on all patients, may be conducive to fraud, even though it does not exonerate the trial participants if they commit it!

Detecting Fraud through Intensive On-site Monitoring

The traditional approach to detect fraud has involved monitoring visits to the clinical centers participating in the trial (*see* **Clinical Trials Audit and Quality Control**). Some such on-site monitoring may be needed and useful, for many types of fraud would remain completely undiscovered if it were not for the careful checks carried out during these visits. However, on-site monitoring is labor intensive and hence expensive, and it too may fail to pick up fraudulent data. Moreover, the law of diminishing returns indicates that it is not **cost-effective** to demand 100% verification of all source data. The approach used for quality control in industrial or laboratory settings can be used so that the monitoring activities are limited to some random selection of the data, with the possible exception of data pertaining to the primary endpoint (*see* **Outcome Measures in Clinical Trials**) of the trial. The random selection can be done at the level of the investigators, the patients or the data items themselves. With such a random sampling scheme, one can estimate the overall data error rate with prespecified precision, and increase the amount of on-site monitoring if the observed rate exceeds some upper limit. Another approach consists of visiting only the centers in which problems, errors or fraud are suspected. Such "for-cause" audits have confirmed major cases of fraud both in multicenter trials and in single institution trials [6, 22]. The major International Conference on Harmonization (ICH) **guidelines** on Good Clinical Practice (GCP) emphasize that a sampling scheme may be appropriate and these can be acceptable to regulatory authorities. It is a major misunderstanding that GCP requires 100% source data verification.

Detecting Fraud through Statistical Checks

Most data entry and data management software used for clinical trials perform basic checks, such as range and consistency checks, but more extensive data checks typically occur at the end of the study along with other statistical analyses, far too late for corrective action. Batteries of checks using standard statistical techniques could also be used early on in the course of a trial without large increases in costs, and could save considerable time if problems are detected and corrected early (*see* **Data Management and Coordination**).

The principles involved in uncovering fraud through statistical techniques rest on the difficulty of fabricating plausible data, particularly in high dimensions [17]. Univariate observations can always be fabricated to fall close to the mean, although preserving their variance is more of a challenge to the inexperienced. Even the astute cheater who takes care to preserve both the mean and the variance may be tripped up by examination of the kurtosis of the distribution. Multivariate observations must in addition be consistent with the correlation structure between their individual components. In general, when data are fabricated to pass certain statistical tests, they are likely to fail on others; Haldane referred to this as "second order" faking [17].

Another way of checking fabricated data is based on the fact that humans are poor random number

generators. Even informed people seem unable to generate long sequences of numbers that pass simple tests for randomness. Digit preference, especially terminal digit preference, or an excess of round numbers may easily reveal data fabrication. Benford's law may also be used to check the randomness of the first digit of all real numbers reported by a single individual (or a single center). This law establishes that the probability of the first significant digit being equal to D ($D = 1, \ldots, 9$) is approximately given by a logarithmic distribution [11]:

$$\Pr(D) \approx \log(D + 1) - \log(D).$$

Hence the frequency of 1's as the first digit should be as high as 30%, the frequency of 2's as the first digit should be close to 18%, while that of 9's should be lower than 5%, a result that runs against intuition. More sophisticated techniques are available to check the randomness of digits in a sequence of data values.

Statistical approaches may also take full advantage of the highly structured nature of clinical trials, which are prospective studies, entirely specified in a written **protocol** and data collection instrument (the "case report form"), usually involving several centers and, when comparative, a randomly assigned treatment. Comparing each center or treatment with the others in terms of the distribution of some variables, either taken in isolation (univariate approach) or jointly (multivariate approach), can detect unusual patterns in the data. Comparisons between centers are particularly informative if there are more than a few observations per center (in which case fraud in any one center may have a sizeable impact on the overall result). Such comparisons are useful with different types of fraud; for instance, the presence of outliers or the consistency in the effect of treatment may reveal fraud aimed at exaggerating the effect, while the presence of "inliers" or underdispersion in the data may reveal invented cases.

Univariate Methods

Beyond range checks and missing data checks, which are performed as part of routine data management, one can use other univariate statistical techniques to inspect the data (Table 2).

Statistical checks may reveal unusual data patterns that are often the mark of fraud (Table 3). Invented or manipulated data tend to have too little variance, no outliers or an abnormally flat distribution [8]. Their distribution may be too close to a simple but

Table 2 Some statistical techniques that may be used to uncover fraud

One variable at a time	Descriptive statistics Box and whisker plot Frequency histogram Stem and leaf plots Tests for slippage
Several variables at a time	Cross-tabulation/scatterplot Correlation/regression Cook's distance Mahalanobis' distance Cluster analysis Discriminant analysis Chernoff faces Star (needle, spike) plots Hotelling's T^2 Tests for treatment contrasts
Repeated measurements	Autocorrelations Profiles Polynomial contrasts Runs tests
Calendar time	Residual plots Cusum Control charts

Table 3 Some patterns that may reveal fraud in clinical trial data

One variable at a time	Digit preference Round number preference Too few or too many outliers Too little or too much variance Strange peaks Data too skewed
Several variables at a time	Multivariate inliers Multivariate outliers Leverage Too weak or too strong correlation
Repeated measurements	Interpolation Duplicates Invented patterns
Calendar time	Breach of randomization Days of week (Sundays or holidays) Implausible accrual Time trends

implausible model, such as a Normal distribution with round numbers for the mean and standard deviation.

Since fraud usually occurs in a single center (except in the unlikely situation of a coordinated fraud across several centers), statistical checks must be performed within each center as well as overall. A comparison of the results reported by different centers may reveal too little variability in one or more centers as compared with the overall variability. Perfect compliance with the protocol, for instance, may be the mark of fraud. Such a comparison may also reveal "slippage" of one or more centers, the null hypothesis being that the means of the variable of interest are equal, but for random fluctuations, to the overall mean [5]. These tests are not informative if there are many centers and few patients per center; on the other hand, grouping small centers could mask a problem in any one of them and is therefore not generally advisable.

Multivariate Methods

Data management usually includes logical checks to ensure the consistency between the values of two or more variables. Multivariate statistical techniques offer more checking possibilities, but they are seldom used in clinical trials, if at all. Multivariate statistical methods include correlations between several patient-related variables as well as comparisons between the randomized groups (Table 2). Simple two-way cross-tabulations or scatter plots for various pairs of variables can be compared across centers, and unusual patterns investigated further. Outlying observations, or outlying groups of observations coming from the same center, can be detected more effectively in multidimensional space than in a single dimension. Moreover, in multidimensional space, "inliers" can be detected through the use of the Mahalanobis' distance just as well as outliers: inliers have an abnormally low Mahalanobis' distance (they fall too close to the multivariate mean), while outliers have an abnormally high Mahalanobis' distance (they fall too far from the multivariate mean) (Table 3). The Mahalanobis' distance is computed by standardizing the variables of interest (subtracting the mean and dividing by the standard deviation), and summing the squares of these standardized variables. The sum approximately follows a χ^2 distribution with N degrees of freedom, if N variables are considered. The detection of inliers may be more useful to detect fraud than the detection of outliers, because fabricated data will tend not to contain outliers which are at higher risk of being detected than are values close to the (multivariate) mean. This method can also be used to see if the N variables of interest are too close to each other for some pair(s) of individuals, in which case one of the individuals in the pair may be a (slightly modified) copy of the other. Robust methods such as using ranks in place of the observations are advisable for the detection of outliers because these can create severe departures from multivariate normality.

Repeated Measures

When, as is often the case, some variables are measured repeatedly over the course of the trial on the same patient, these measures lend themselves well to a variety of checks (Table 2). Here, again, an insufficient variability over time may reveal the propagation of previous values rather than genuine observations (Table 3). Sometimes the fraud involves a mechanism or computer algorithm for making up data.

Calendar Time

In any trial with prolonged patient entry and follow-up, one can use calendar time to perform additional checks on the data (Table 2). Simple checks can be performed on the day of the week, as certain events or examinations are unlikely to have taken place on a Sunday. Time intervals between successive visits and the number of visits per unit time provide further opportunities for checking the plausibility of a sequence of events (Table 3). A comparison of treatment groups by week or month of randomization can reveal suspect periods during which all treatments were not allocated with equal probability. Perfect compliance with the protocol in terms of dates may be a marker, as noted above, but should not be taken on its own as necessarily misconduct. More advanced checks may sometimes be performed, such as the variance of observations over time. An excellent example is provided by an animal study in which the variability of the heart rates of dogs treated consecutively showed far too little variance initially, leading to a strong suspicion of data fabrication in the early stages of the study [3].

Impact of Fraud on the Results of Clinical Trials

The highly publicized case of fraud in the **National Surgical Adjuvant Breast and Bowel Project** (NSABP) provides a framework to examine the impact of such fraud on the results of clinical trials [9]. Briefly, one of the investigators in breast cancer trials systematically altered some baseline patient data so that these patients became eligible for entry into the trials. The data subject to falsification were the dates of surgery and biopsy or estrogen receptor values. For example, in one study, the delay between the surgery and randomization had been set to a maximum of 30 days by the trial protocol, and dates were falsified for a few patients in whom this limit had been exceeded. The fraud clearly was not aimed at distorting the results of the trials one way or another (it could only have done so had the treatment effect been substantially different among the wrongly entered patients than among the others). As a matter of fact, a careful reanalysis of NSABP trials without the data from the fraudulent institution confirmed that the trial outcomes had not been materially affected by the fraud [9]. In another large trial in stroke published recently, all data from one center suspected of fraud were excluded from the analysis, again with negligible impact on the study results. Yet this center had contributed 452 (6.4%) of the 7054 patients randomized in the study, and statistical analysis of the variability of their data supported the belief that no real patients had actually been studied in this center [21]!

Fraud is unlikely to affect the results of a trial if any of the following conditions hold:

- the fraud is limited to one or a few investigators (perhaps one center in a multicenter setting) and/or to a few data items, provided that there are many investigators or centers;
- the fraud bears on secondary variables that have little or no effect on the primary endpoint of the trial;
- the fraud affects all treatment groups equally, and hence does not bias the results of the trial. Fraud committed without regard to the treatment assignments (e.g. prior to randomization or in double-blind trials) generates noise but no bias.

At least one of these conditions frequently holds, and therefore fraud should not be expected to have a major impact on the results of multicenter clinical trials. As a matter of fact, a search of Medline from 1966 to 1997 revealed that 235 articles had been retracted, 86 of which were deemed to be due to misconduct [4]. These numbers do not bear specifically on clinical trial reports, but they are quite small compared with the total number of articles published during the same period.

One caveat is that where an increase in noise occurs, this can make dissimilar treatments appear similar. With a trend towards using **equivalence** or noninferiority trials for licensing purposes this is of concern and could result in ineffective medicines being licensed.

Actions in Cases of Fraud

When fraud is suspected in a center, all analyses can be repeated with and without that center, in order to assess the sensitivity of the trial results to the fraud. Although fraudulent data would in general have to be excluded from the main analysis of the trial, other validated data from the same center might well be kept in the analysis. If a **Data and Safety Monitoring Board** oversees the trial, then it seems appropriate to leave such decisions to their discretion. Biostatistical methods can only point at problems; further investigations and hard evidence are needed to confirm fraud.

Conclusion

Randomized clinical trials constitute, by design, the most reliable type of medical experiment. Their data can be verified using statistical techniques that take advantage of their highly structured nature. Their results are generally robust to occasional cases of data falsification and fabrication at some participating centers. As George put it, "the methodology of clinical trials *de facto* provides a measure of protection against deliberate deception that is generally unappreciated by those not familiar with the methodology" [10]. Claims to the contrary notwithstanding, we are not aware of quantitative evidence that fraud is common in clinical trials. However, fraud is a cause for concern regardless of its prevalence or consequences because the "habit of truth" is the cardinal value in scientific endeavors. Fraud must be fought, but attempts to impose more bureaucracy and heavier

monitoring on clinical trials is the wrong answer to a problem that can be overrated in the media. Fraud can largely be prevented through proper design of the trial protocol and case report form, and detected by statistical procedures and computerized checks that make use of the unique structure of clinical trial data [14].

Acknowledgments

This article is based largely on a paper with further references published on behalf of the Subcommittee on Fraud of the International Society for Clinical Biostatistics [4].

References

[1] Andersen, B. (1990). *Methodological Errors in Medical Research*. Blackwell Scientific Publications, Oxford.

[2] Angell, M. & Kassirer, J.P. (1994). Setting the record straight in the breast-cancer trials, *New England Journal of Medicine* **330**, 1448–1450.

[3] Bailey, K.R. (1991). Detecting fabrication of data in a multicenter collaborative animal study, *Controlled Clinical Trials* **12**, 741–752.

[4] Buyse, M., George, S.L., Evans, S., Geller, N., Ranstam, J., Scherrer, B., Lesaffre, E., Murray, G., Edler, L., Hutton, J., Colton, T., Lachenbruch, P., Verma, B. for the ISCB Subcommittee on Fraud (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials, *Statistics in Medicine* **18**, 3435–3452.

[5] Canner, P.L., Huang, Y.B. & Meinert, C.L. (1981). On the detection of outlier clinics in medical and surgical trials: I. Practical considerations, *Controlled Clinical Trials* **2**, 231–240.

[6] Christian, M.C., McCabe, M.S., Korn, E.L., Abrams, J.S., Kaplan, R.S. & Friedman M.A. (1995). The National Cancer Institute Audit of the National Surgical Adjuvant Breast and Bowel Project Protocol B-06, *New England Journal of Medicine* **333**, 1469–1474.

[7] Christie, B. (1999). Panel needed to combat research fraud, *British Medical Journal* **319**, 1222.

[8] Evans, S. (1998). Fraud and misconduct in medical science, in *Encyclopedia of Biostatistics*, Vol. 2, P. Armitage & T. Colton, eds. Wiley, Chichester, pp. 1583–1588.

[9] Fisher, B., Anderson, S., Redmond, C.K., Wolmark, N., Wickerham, D.L. & Cronin, W.M. (1995). Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer, *New England Journal of Medicine* **333**, 1456–1461.

[10] George, S.L. (1997). Perspectives on scientific misconduct and fraud in clinical trials, *Chance* **10**, 3–5.

[11] Hill, T.P. (1999). The difficulty of faking data, *Chance* **12**, 27–31.

[12] Hone, J. (1993). Combating fraud and misconduct in medical research, *Scrip Magazine* **March**, 14–15.

[13] Horton, R. (2000). After Bezwoda, *Lancet* **355**, 942–943.

[14] Knatterud, G.L., Rockhold, F.W., George, S.L., Barton, F.B., Davis, C.E., Fairweather, W.R., Honohan, T., Mowery, R. & O'Neill, R.T. (1998). Guidelines for quality assurance procedures for multicenter trials: a position paper, *Controlled Clinical Trials* **19**, 477–493.

[15] Lock, S. & Wells, F., eds (1996). *Fraud and Misconduct in Medical Research*, 2nd Ed. BMJ Publishing Group, London.

[16] Ranstam, J., Buyse, M., George, S.L., Evans, S., Geller, N., Scherrer, B., Lesaffre, E., Murray, G., Edler, L., Hutton, J., Colton, T. & Lachenbruch, P. for the ISCB Subcommittee on Fraud (2000). The biostatistician's view of fraud in medical research, *Controlled Clinical Trials* **21**, 415–420.

[17] Rao, C.R. (1989). *Statistics and Truth*. International Cooperative Publishing House, Burtonsville.

[18] Schmidt, J., Gertzen, H., Aschenbrenner, K.M. & Ryholt-Jensen, S. (1995). Detecting fraud using auditing and biometrical methods, *Applied Clinical Trials* **May**, 40–50.

[19] Shapiro, M.F. & Charrow, R.P. (1989). The role of data audits in detecting scientific misconduct: results of the FDA program, *Journal of the American Medical Association* **261**, 2505–2511.

[20] Sunderland, M., Kuebler, S., Weiss, G. & Coltman, C. (1990). Compliance with protocol: quality assurance data from the Southwest Oncology Group, *Proceedings of the American Society of Clinical Oncology* **9**, (Abstract 299).

[21] The ESPS2 Group (1997). European Stroke Prevention Study 2. Efficacy and safety data, *Journal of Neurological Sciences* **151**, Supplement, S1–S77.

[22] Weiss, R.B., Rifkin, R.M., Stewart, F.M., Theriault, R.L., Williams, L.A., Herman, A.A. & Beveridge, R.A. (2000). High-dose chemotherapy for high-risk primary breast cancer: an on-site review of the Bezwoda study, *Lancet* **355**, 999–1003.

[23] Weiss, R.B., Vogelzang, N.J., Peterson, B.A., Panasci, L.C., Carpenter, J.C., Gavigan, M., Sartell, K., Frei, E. & McIntyre, O.R. (1993). A successful system of scientific data audits for clinical trials, *Journal of the American Medical Association* **270**, 459–464.

[24] Wood, W.C. (1994). Audit of Cancer and Leukemia Group B, *New England Journal of Medicine* **331**, 279.

MARC BUYSE & STEPHEN J.W. EVANS

Frequency Distribution

A frequency distribution is a method of summarizing a set of numerical data in tabular or graphical form. For example, one might summarize the height of individuals, family incomes, or chemical concentrations in blood samples. A frequency distribution is constructed from k nonoverlapping class intervals, usually of equal length. Let $a_0, a_1, \ldots, a_k$ denote the class boundaries. A value, x, is in the jth class interval if $a_{j-1} < x \leq a_j$. A frequency distribution reports the frequency or count of data values falling into each class interval. Let the jth frequency, f_j, denote the number of data values falling into the jth class. The above description can be summarized in tabular form (Table 1).

If numerical observations are from a completely censused population, then the result of the above procedure is a population frequency distribution. When the observations are from a sample of the population, as is usually the case, Table 1 is called a sample frequency distribution.

The procedure can be illustrated with weights of 92 Pennsylvania State University students as reported in Ryan & Joiner [2] (Table 2).

Care and judgment must be exercised in selecting the number of classes and the class boundaries. Whenever possible it is important to keep the class intervals the same length: this maintains the close relationship between the sample frequency table and the underlying population density function. Also, class intervals should be written unambiguously so that each data value falls in exactly one interval. For example, if in the student weight data frequency table the weight classes had been described as 140–150 and 150–160, then the 10 students who self reported their weights as 150 pounds could not be assigned unambiguously to a weight class. In the above frequency distribution the intervals are unambiguous, and also boundaries are kept away from weights that are multiples of 5, i.e. those weights where self-reporting biases are most likely.

Before the advent of modern computing, frequency tables were often used to condense the data before doing time-consuming arithmetic computations: for example, approximate means and variances can be quickly calculated from frequency distributions. Today, however, frequency distributions are primarily used to display and understand data.

Frequency distributions are usually **graphically displayed** as *histograms*. In a histogram each class interval is represented by a vertical bar whose base is the class interval and whose height is the number of observations in the class interval. Figure 1 shows

Table 1 Summary of frequency distribution

Class interval	Class frequency
$a_0 < x \leq a_1$	f_1
$a_1 < x \leq a_2$	f_2
$\cdots$	$\cdots$
$a_{j-1} < x \leq a_j$	f_j
$\cdots$	$\cdots$
$a_{k-1} < x \leq a_k$	f_k
Sum	n

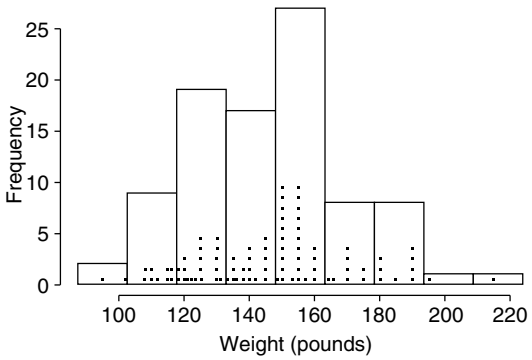

Figure 1 Histogram for student weight data

Table 2 Pennsylvania State University student weight data

Weight in pounds	Frequency	Weight in pounds	Frequency
$87.5 < x \leq 102.5$	2	$162.5 < x \leq 177.5$	8
$102.5 < x \leq 117.5$	9	$177.5 < x \leq 192.5$	8
$117.5 < x \leq 132.5$	19	$192.5 < x \leq 207.5$	1
$132.5 < x \leq 147.5$	17	$207.5 < x \leq 222.5$	1
$147.5 < x \leq 162.5$	27	Sum	$n = 92$

the histogram for the student weights, along with a dot plot representing each of the 92 weights. For unequally spaced intervals the histogram should be modified so that the area of each bar is proportional to the frequency for that class interval.

There are other ways of presenting the same summary information contained in a frequency distribution. The relative frequency distribution reports the proportion of observations falling in each class interval. And the cumulative frequency table reports the number of observations falling in or below each class interval. Kendall et al. [1] give a detailed review of frequency distributions and their extensions.

References

[1] Kendall, M.G., Stuart, A., Ord, J.K. & Ord, K. (1994). *The Advanced Theory of Statistics*. Vol. 1. *Distribution Theory*, 6th Ed. Wiley, New York.

[2] Ryan, B.F. & Joiner, B.L. (1994). *Minitab Handbook*, 3rd Ed. Duxbury Press, North Scituate.

W. SMITH

Frequency Matching

Frequency matching, also known as category matching, is a sampling design used in **case–control studies** that yields **controls** with the same distribution over categories defining levels of potential **confounders** as is observed in the cases. For example, suppose that cases are classified into 20 categories defined by gender and by ten ten-year age intervals, and that the distribution of cases in these categories is observed. A frequency-matched sample of controls with this same distribution could be obtained by sampling as controls a constant multiple of the number of cases in each category. Sometimes, for convenience, one obtains frequency-matched controls that conform to the expected distribution of cases rather than wait to match the on actual distribution of cases.

Frequency matching is also used in **cohort studies** to assure that the control cohort has the same distribution over categorical levels of potential confounders as the exposed cohort.

(*See also* **Matching**)

MITCHELL H. GAIL

Galton, Francis

Born: February 16, 1822, in Birmingham, UK.
Died: January 17, 1911, in Haslemere, UK.

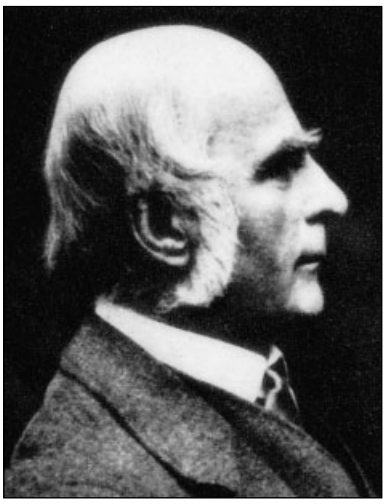

Francis Galton was the founder of biometry. The origins of many statistical procedures can be seen in his pioneering efforts. A Victorian polymath, although no mathematician, he influenced the development of mathematical statistics through **Karl Pearson** (who wrote a massive life of him in three volumes), **F. Y. Edgeworth** and, at one remove, **R. A. Fisher**. He was a half-cousin of Charles Darwin (they shared a grandfather in Erasmus Darwin) and throughout his life he enjoyed both private means and a wide circle of intellectual relatives and friends.

A precocious child, he went up to Trinity College, Cambridge, in 1840 to study mathematics and then medicine, but suffered a breakdown which forced him to abandon an honors mathematics degree. Although he almost completed his medical studies he disliked the idea of practising, and with his father's death in 1844 he inherited sufficient wealth to remove the need. In 1853 he married Louisa Butler, daughter of the Dean of Peterborough, whose brother Montagu was later to be Master of Trinity.

Between 1844 and 1853 Galton indulged in African travels, sailing up the Nile and subsequently undertaking extensive explorations in South West Africa, which earned him the 1854 Founder's Medal of the Royal Geographical Society, on whose Council he served for many years. He used his experiences as the basis for a book *The Art of Travel* [1], published in 1855, and the following year he was elected a Fellow of the Royal Society.

In 1861, Galton's statistical inclination showed as he gathered meteorological information from all over Europe from which, on plotting the barometric pressure, he discovered – and named – the "anticyclone". His statistical interests were channeled into questions of heredity by the publication, in 1859, of Charles Darwin's *On the Origin of Species*, coupled very probably with the realization that he and Louisa were likely to remain childless. The result was Galton's first important book, *Hereditary Genius* [2] (*see* **Human Genetics, Overview**).

From the publication of *Hereditary Genius* in 1869 to his death in 1911, Galton occupied himself principally with statistical questions bearing on heredity. In 1874, with the Reverend H. W. Watson, he wrote a paper [7] on the extinction of families, which is regarded as the origin of the statistical theory of **branching processes**. The following year he discovered that a normal mixture of **normal distributions** is itself a normal distribution, about which S. M. Stigler has written "Galton's conceptual use of the result was new and ingenious and represents the most important step in perhaps the single major breakthrough in statistics in the last half of the nineteenth century". The breakthrough continued with Galton's introduction of **regression** in 1877 in connection with the analysis of a genetics experiment that he had planned which was the continuous analog of Mendel's 1865 discrete experiments, and which laid the foundations of biometric genetics (*see* **Genetic Epidemiology**; **Mendel's Laws**).

In 1885, Galton presented his explanation of regression in terms of the geometry of the **bivariate normal distribution**, for which J. Hamilton Dickson of Cambridge provided the mathematics. All of these advances were brought together early in 1889 with the publication of *Natural Inheritance* [3]. After the book had gone to press, Galton realized that if in his regression diagram the two variates were scaled so as to have the same probable error on the paper, then both regression lines would have the same slope (each with respect to its proper axis) which was therefore suitable as an "index of co-relation", or "**correlation** coefficient", as it soon became. In 1899, he invented normal probability paper, a natural extension of his earlier use of the cumulative normal probability

distribution, which he termed an "ogive". In 1901, Galton assisted Pearson and his colleague W. F. R. Weldon in the launch of the journal **Biometrika**, and contributed a short introduction to *Biometry* for the first number.

In 1883, Galton had coined the word "**eugenics**", but it was not until the early years of the twentieth century that he was able to promote his ideas. In 1904 he gave University College money to establish a "Eugenics Records Office", with a Research Fellow and an assistant, later to become the "Eugenics Laboratory" under Pearson. In his will he left University College £45 000 to endow a Chair of Eugenics, initially called the Galton Professorship of Eugenics. The first two holders were Pearson and Fisher (ultimately the Laboratory became the Galton Laboratory and the Professorship one of Human Genetics). In 1907 a "Eugenics Education Society" was formed, and Galton soon agreed to become its Honorary President. The Society flourished, especially under the later presidency of Leonard Darwin, becoming the Eugenics Society (and now the Galton Institute).

Galton was a prolific writer on many subjects, numbering 17 books amongst his 300 publications. Of those not yet mentioned, his pioneering books on fingerprints and their uses (1892 [4] and 1895 [5]) may be recalled.

In 1902 Galton was awarded the Darwin Medal of the Royal Society, and was particularly delighted to be elected an Honorary Fellow of his old Cambridge college, Trinity. Further honors came in extreme old age: he was knighted in June 1909 and received the Copley Medal, the Royal Society's highest award, in October 1910 a few months before he died at the age of 88.

Although Galton was very much a product of his times and his social class, his views were by no means typical. He thought, as did others, that the explorer H. M. Stanley had treated Africans badly; he went out of his way, in *Hereditary Genius*, to express his "grief" that paucity of data prevented him from discussing the influence of mothers on their offspring, which he knew to be as important as that of fathers; and in respect of eugenics he wrote at the end of his autobiography [6]: "Man is gifted with pity and other kindly feelings; he also has the power of preventing many kinds of suffering. I conceive it to be within his province to replace Natural Selection by other processes that are more merciful and not less effective".

References

[1] Galton, F. (1855). *The Art of Travel*. Murray, London.

[2] Galton, F. (1869). *Hereditary Genius*. Macmillan, London.

[3] Galton, F. (1889). *Natural Inheritance*. Macmillan, London.

[4] Galton, F. (1892). *Finger Prints*. Macmillan, London.

[5] Galton, F. (1895). *Finger Print Directories*. Macmillan, London.

[6] Galton, F. (1908). *Memories of My Life*. Methuen, London.

[7] Galton, F. & Watson, H.W. (1874). On the probability of the extinction of families, *Journal of the Anthropological Institute* **4**, 138–144.

Further Reading

Forrest, D.W. (1974). *Francis Galton: The Life and Work of a Victorian Genius*. Elek, London.

Gillham, N.W. (2001). *The Life of Sir Francis Galton*, Oxford University Press.

Keynes, M., ed. (1993). *Sir Francis Galton, FRS: the Legacy of His Ideas*. Macmillan, London.

Pearson, K. (1914–1930). *The Life, Letters and Labours of Francis Galton*, 3 vols. Cambridge University Press, Cambridge.

A.W.F. EDWARDS

Galton–Watson Process

The Galton–Watson process should more accurately be called the Bienaymé–Galton–Watson process; or, better, just the simple **branching process**. Descendants of a single ancestor at generation zero at any given subsequent generation produce offspring, independently of each other and of individuals in preceding generations. The probability distribution of the number of offspring of any one individual is identical with that of the initial ancestor. Denoting by $p_r, r = 0, 1, 2, \ldots$, the probability that any one individual has r offspring, and by Z_n the number of individuals at generation n, we have $p_r = \Pr(Z_1 = r)$. If we write $F(s) = \sum_{r=0}^{\infty} p_r s^r$, $0 \leq s \leq 1$, for the probability generating function (pgf) of Z_1 (*see* **Generating Functions**), then a crucial property is that the pgf of Z_n, $F_n(s)$, is in fact the nth functional iterate of $F(s)$ (i.e. $F_n(s) = F(F_{n-1}(s))$, where $F_0(s) = s$). This is reflected in the fact that if the mean number of offspring per individual is denoted by m (i.e. $m = \sum_{r=0}^{\infty} r p_r = F'(1)$), then the mean number of individuals at time n is m^n (i.e. $\mathrm{E}Z_n = m^n$). If $m > 1$, we thus have "exponential" (Malthusian) average growth with n. However, even in this "supercritical" case, extinction may occur with positive probability. The Criticality Theorem for the process, if the trivial case $p_1 = 1$ is excluded, asserts that if $m \leq 1$ extinction occurs with probability $q = 1$; but if $m > 1$ the probability q of ultimate extinction is the unique root of the equation $F(x) = x$ in the interval $0 \leq x < 1$.

The process and the Criticality Theorem are important because of the breadth of practical applicability [2, 4]. The individuals in the **stochastic process** $\{Z_n\}$ may be (as in the original application to the problem of extinction of surnames) direct male descendants of a single male ancestor; or carriers of copies of a mutant **gene**, electrons in an electron multiplier, neutrons in a nuclear chain reaction, branch units in a polymer molecule [3], or branches emanating from a point of propagation in crack growth. Even though the independence assumptions will tend to break down in practice if numbers become large, particularly in biological applications, the value of q calculated under these assumptions in the case $m > 1$ will nevertheless often provide a good approximation [4]. The structure of the process is easily generalized to several types of individual (the multitype process, with an accompanying Criticality Theorem) which is of very great applicability; for example, in **population genetics** and polymer chemistry. Another important direction of generalization is to permit immigration (e.g. recurrent mutation), in which case when $m < 1$ an equilibrium may result due to a balance between the immigrants and tendency to extinction.

For the simple branching process, I.J. Bienaymé [1] gave a completely correct statement of the Criticality Theorem in 1845. This passed unnoticed until recent decades [5, 7]. The usual designation for the process originates from the partly correct statement of the Criticality Theorem in 1873–1874 by **F. Galton** and H.W. Watson. Kendall [6] gives a vivid history from this starting point.

References

[1] Bienaymé, I.J. (1845). De la loi de multiplication et de la durée des familles, *Société Philomatique de Paris-Extraits, Ser. 5*, pp. 37–39 (also in L'Institut, Paris **589**, 131–132; and reprinted in [7]).

[2] Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, 3rd Ed, Vol. 1. Wiley, New York; see especially Sections XII.4 and XII.5.

[3] Flory, P.J. (1953). *Principles of Polymer Chemistry*. Cornell University Press, Ithaca, New York, Chapter IX, especially pp. 352–353.

[4] Harris, T.E. (1963). *The Theory of Branching Processes*. Springer-Verlag, Berlin.

[5] Heyde, C.C. & Seneta, E. (1972). The simple branching process, a turning point test and a fundamental inequality: a historical note on I.J. Bienaymé, *Biometrika* **59**, 680–683.

[6] Kendall, D.G. (1966). Branching processes since 1873, *Journal of the London Mathematical Society* **41**, 385–406.

[7] Kendall, D.G. (1975). The genealogy of genealogy: branching processes before (and after) 1873, *Bulletin of the London Mathematical Society*, **7**, 225–253.

Bibliography

Athreya, K.B. & Ney, P. (1972). *Branching Processes*. Springer-Verlag, Berlin.

Heyde, C.C. & Seneta, E. (1977). *I.J. Bienaymé: Statistical Theory Anticipated*. Springer-Verlag, New York.

Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley, New York.

Mode, C.J. (1971). *Multitype Branching Processes*. Elsevier, New York.

Sevastyanov, B.A. (1971). *Vetviashchiesia Protsessi (Branching Processes)*. Nauka, Moscow (in Russian).

E. SENETA

Gamma Distribution

The gamma density in its general form is

$$g(x; s, a, \rho) = \frac{(x-s)^{\rho-1}}{a^{\rho}\Gamma(\rho)} \exp\left(\frac{s-x}{a}\right)$$
$$(s < x < \infty; a, \rho > 0).$$

It is a Pearson type III density, and includes the χ^2 density (*see* **Chi-square Distribution**); the latter refers to a **"goodness-of-fit"** test based on the sum of squared normed deviations. This in turn relates to the distribution of the sum of squares of **normal** deviates. The distribution may be reverse J-shaped if $0 < \rho < 1$. The basic parameters (s, a, ρ) are associated with origin, scale, and shape; the **skewness** is $\alpha_3 = \sqrt{\beta_1} = \mu_3/\mu_2^{3/2} = 2/\sqrt{\rho}$, and **kurtosis**, $\beta_2 = \mu_4/\mu_2^2 = 3 + 6/\rho$. For normality these two moment parameters take the values 0 and 3, respectively, so that ρ has to be large to decrease the skewness to insignificance. It is well known that the gamma distribution approaches normality very slowly (*see* **Convergence in Distribution and in Probability**). A recent study by Revfeim [7], using a **transformation**, and manipulation of series, relates the distribution and its inverse to the corresponding normal functions.

If we choose $g(\cdot)$ as a suitable model to account for experimental data with some success, then sample size is important. Applications of the gamma distribution are numerous and cover many applied sciences. The reader may refer to the bibliography in Johnson et al. [5], which includes some 300 references. Historically, we may glance at *Karl Pearson's Early Statistical Papers* [6], illustrations relating to the state of the art towards the end of the nineteenth century. Examples are:

1. Range of the barometer; p. 80;
2. Professor Weldon's crab measurements; p. 82;
3. Heights for 25 878 recruits in the US Army (1875); p. 83;
4. Length–breadth index of 900 Bavarian skulls; p. 86;
5. The distribution of 8689 cases of enteric fever received into the Metropolitan Asylums Board Fever Hospitals (1871–1893); p. 88.

Examples of the gamma density are given in Figure 1.

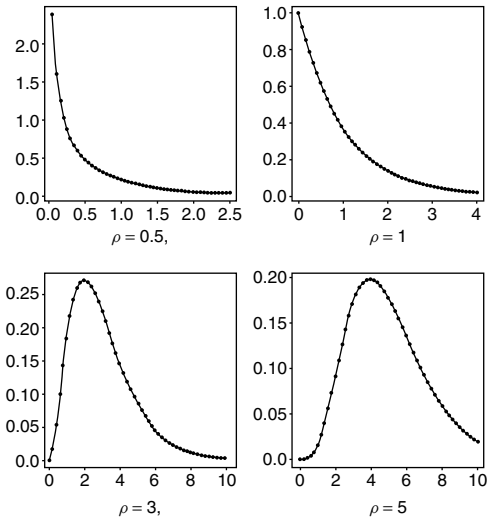

Figure 1 Illustrations of the densities $g(x; 0, 1, \rho)$. Note that for $a > 0$, $g(x; s, a, \rho)$ tends to ∞ as $x \to s$ when $0 < \rho < 1$, but equals unity when $x = s$, and $\rho = 1$

Histograms should be unimodal (*see* **Frequency Distribution**); high density near the origin may lead to estimation problems especially in the estimation of s. A brief account of **simulation** studies for the **maximum likelihood** estimators [1] shows that the distribution of s may be U-shaped if ρ is small (less than unity). If, on the other hand, ρ is large then the distribution may have a large variance. Moreover, in this case it was found that **confidence intervals** for ρ would require sample sizes exceeding 500 to control their variance. Cohen & Whitten [3] have considered modified maximum likelihood estimators including the use of the smallest sample value (*see* **Order Statistics**) to estimate s, and with special attention to cases when ρ is small.

In sampling studies of the estimators (s, a, ρ), given a gamma density (s, a, ρ) with specified sample size, among the great variety of possible forms samples with negative skewness may arise. For example, in sampling from (0, 1, 6) and a sample of 50, there were 20 000 valid solutions with 2157 failures including 432 cases of negative skewness. Similarly for $\rho = 10$, and $n = 50$, there were 182 cases of negative skewness out of some 20 000 trials.

A new parametric form due to Cheng and Traylor [2] avoids the negative skewness syndrome with μ, σ, and λ relating to origin, scale, and shape, the

density being

$$g(x;\mu,\sigma,\lambda) = \frac{1}{\sigma\lambda\Gamma(\lambda^2)}\left\{\lambda^{-2}\left[1+\frac{\lambda(\lambda-\mu)}{\sigma}\right]\right\}^{\lambda^2-1} \times \exp\left\{-\frac{1}{\lambda^2}\left[1+\frac{\lambda(x-\mu)}{\sigma}\right]\right\}$$

$$(\sigma > 0; \lambda \neq 0 \text{ and } 1+\frac{\lambda(x-\mu)}{\sigma} > 0).$$

In our notation

$$\rho = \frac{1}{\lambda^2}, \qquad a = \sigma|\lambda|, \qquad s = \mu - \sigma\lambda^{-1}.$$

Moreover, $\sqrt{\beta_1} = 2/\sqrt{\rho}$ = skewness.

If $\lambda > 0$, then the distribution has positive skewness; if $\lambda < 0$, then the distribution is valid with negative skewness. Hirose [4] has a program for maximum likelihood estimation using a predictor–corrector algorithm. Asymptotic covariances using the Hessian matrix will be available, but caution in finite sample interpretation is advised. Let it be said that in finite samples there is no panacea in maximum likelihood estimation procedures.

In the present context it would be remiss to ignore the link between the gamma distribution and the gamma function. Euler in the eighteenth century used the unnormed gamma density as the integrand and integrated over the range $(0, \infty)$ to define the gamma function. Of course, at that time there was no connection with statistics. This definition, interpreted for complex variables, is to be found in modern textbooks. After Euler, Weierstrass introduced an infinite product formula. Inevitably new forms arose for products of gamma functions such as $\Gamma(x)$, $\Gamma\left(x+\frac{1}{2}\right)$, and $\Gamma(x)\Gamma(1-x)$, as well as the question of approximations for large variables. Stirling's formula (eighteenth century)

$$\ln[\Gamma(x)] = \left(x-\tfrac{1}{2}\right)\ln(x) - x + \tfrac{1}{2}\ln(x\pi) + J(x) \quad (x \to 0, |\arg x| < \pi)$$

plays a role in many asymptotic structures, including, for example, the transition to normality of the **binomial distribution**. The residue $J(x)$ is a divergent power series in odd powers of $1/z$, the coefficients being Bernoulli numbers; these diverge ultimately. A partial sum of the series has an integral form which may be bounded and shown to be valid. Stieltjes [8] produced a convergent continued fraction for the residue and this gives monotonic increasing, and decreasing bounds when the variable is real and positive.

References

[1] Bowman, K.O., Shenton, L.R. & Karlof, C. (1995). Estimation problems associated with the three parameter gamma distribution, *Communications in Statistics – Theory and Methods* **24**, 1355–1376.

[2] Cheng, R.C.H. & Traylor, L. (1995). Non-regular maximum likelihood problems, *Journal of the Royal Statistical Society, Series B* **57**, 3–44.

[3] Cohen, A.C. & Whitten, B.J. (1982). Modified moment and modified maximum likelihood estimators for parameters of the three-parameter gamma distribution, *Communications in Statistics – Simulation and Computation*, **11**, 197–216.

[4] Hirose, H. (1995). Maximum likelihood parameter estimation in the three-parameter gamma distribution, *Computational Statistics and Data Analysis* **20**, 343–354.

[5] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd Ed. Wiley, New York.

[6] Pearson, Karl (1948). *Karl Pearson's Early Statistical Papers*. Cambridge University Press, Cambridge.

[7] Revfeim, K.J.A. (1991). Approximation for the cumulative and inverse gamma distribution, *Statistica Neerlandica* **45**, 327–331.

[8] Stieltjes, T.J. (1918). *Oeuvres Completes*, Tome 2. P. Noordhoff, Groningen.

K.O. BOWMAN & L.R. SHENTON

Gardner, Martin John

Born: July 25, 1940, in Essex, UK.
Died: January 22, 1993, in Southampton, UK.

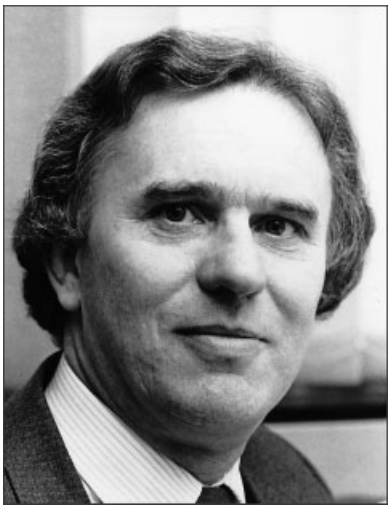

Although Martin Gardner achieved fame for his work at Sellafield, which showed that paternal exposure to ionizing radiation was linked to leukemia in childhood [3] (*see* **Leukemia Clusters**), this discovery came towards the end of a varied and highly productive career in medical statistics. After receiving a first class degree in mathematics he joined the Medical Research Council Social Medicine Unit at the London School of Hygiene and Tropical Medicine. It was there that he became interested in the problem of why the rates of common diseases vary so much between different parts of Britain – an interest to which he returned repeatedly throughout his career. He carried out a series of novel analyses on the geographical links between socioeconomic conditions and coronary heart disease (*see* **Geographic Patterns of Disease**; **Geographic Epidemiology**). Later, at the MRC Environmental Epidemiology Unit in Southampton, he produced two uniquely detailed atlases of mortality in England and Wales [4, 5].

In 1971 Gardner moved to Southampton as a senior lecturer. One of his hallmarks in those early years was his ability to form creative partnerships with doctors. This arose from his facility in grasping medical problems and reducing them to their essence, and from his personality. He was a modest and immensely likeable man, and his promotions to a Readership followed by a personal chair in 1985 were widely welcomed.

He was a gifted teacher of statistical methods. He once gave a lecture on **life tables** that evoked spontaneous applause. Only seldom do lectures on statistics evoke such a response from doctors. He was statistical advisor to the *British Medical Journal* for many years, a position of which he was justly proud. He campaigned vigorously to improve the quality of statistics in medical journals (*see* **Statistical Review for Medical Journals**) and helped to draw up guidelines for statistical analysis. He coordinated the development of statistical checklists for referees, and the production of a series of papers on **confidence intervals**. These were brought together in *Statistics with Confidence* [2]. When refereeing papers Gardner was positive, searching out strengths and scientific importance and not dwelling unduly on weaknesses and errors.

In 1978 Gardner and Donald Acheson published their report on the ill effects of asbestos on health [1]. It was a tribute to Gardner's balanced approach to difficult and contentious issues – a balance that made him much in demand to serve on committees. He rapidly became an expert on control limits for industrial hazards. He was a frequent visitor to the US and through this and his work with the **International Agency for Research Against Cancer** established an international reputation.

Gardner gave time to the scouts, schools, and other local activities. As a schoolboy he had been an exceptional athlete, and sport was a major hobby throughout his life. He and his wife, Linda, whom he met at Berkeley, had three children.

In 1983 Gardner was asked to join the committee, chaired by Sir Douglas Black, examining the excess of childhood leukemia around the nuclear plant at Sellafield in Cumbria. Over the next few years he devised and carried out three major studies together with colleagues in the Environmental Epidemiology Unit. The "Gardner Report" [3], the result of fastidious research, aroused immense public interest and scientific controversy. He handled the media with great skill and addressed the scientific debate with balance and good humor. His career ended all too soon, but on a high note.

References

[1] Acheson, E.D. & Gardner, M.J. (1979). The ill-effects of asbestos upon health in man, in *Asbestos*, Vol. 2, *Final Report of the Advisory Committee on Asbestos*. HMSO, London, pp. 7–83.

[2] Gardner, M.J. & Altman, D.G., eds (1989). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. British Medical Journal, London.

[3] Gardner, M.J., Snee, M.P., Hall, A.J., Powell, C.A., Downes, S. & Terrell, J.D. (1990). Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria, *British Medical Journal* **300**, 423–429.

[4] Gardner, M.J., Winter, P.D. & Barker, D.J.P. (1984). *Atlas of Mortality from Selected Diseases in England and Wales, 1968–1978*. Wiley, Chichester.

[5] Gardner, M.J., Winter, P.D., Taylor, C.P. & Acheson, E.D. (1983). *Atlas of Cancer Mortality in England and Wales, 1968–1978*. Wiley, Chichester.

DAVID BARKER

Gastroenterology

The specialty of gastroenterology concerns diseases of the digestive tract from the esophagus to the rectum. The biliary tract and pancreas are included, but diabetes is regarded as a subspeciality of **endocrinology**. Often the definition of gastroenterology is widened to include the liver also (*see* **Hepatology**). At one time the internal organs constituting the gastrointestinal tract were directly accessible only at surgery. Nowadays, less invasive investigative techniques include contrast radiography and endoscopy for direct visualization and biopsy. Surgical maneuvers may be effected via endoscopic or minimal-access "keyhole" routes.

Disease taxonomy, though not as problematic as in **psychiatry**, is nevertheless not entirely straightforward. Quite exhaustive investigations fail to demonstrate an organic cause in many cases of abdominal pain. Thus nonspecific acute abdominal pain and irritable bowel syndrome are recognized diagnoses, albeit defined by absence rather than presence of specific diagnostic features.

One important area of ambiguity relates to inflammatory bowel disease (or nonspecific colitis). Crohn et al. [3] described a regional ileitis with granulomatous pathology, yet which in many respects closely resembled the already well-established diagnosis of ulcerative colitis (or idiopathic proctocolitis). Differential diagnosis between ulcerative colitis and Crohn's disease is sometimes clear-cut but is often so problematic that many have debated whether they should be regarded as different diseases. Jones et al. [11] sought to resolve this issue by applying **numerical taxonomy** to a series of patients with nonspecific colitis. Two main clusters emerged, that could be identified with proctocolitis and Crohn's disease, but the latter category was relatively heterogeneous, with several subclusters apparent.

Subsequently Harries et al. [8] pointed out that ulcerative colitis is less frequent in current smokers than in lifelong nonsmokers or ex-smokers; the latter often develop the disease within a year or so of quitting smoking. This observation, contrasting with Crohn's disease, which is slightly commoner in smokers, suggests differences in etiology. Ulcerative colitis shares its nonsmoking epidemiology with a few other conditions, notably the related condition of oral aphthous ulceration, and Alzheimer's and Parkinson's diseases. The possibility of therapeutic use of nicotine is limited by the propensity to cause side-effects. It is thus a particularly appealing option for ulcerative colitis, for which topical administration is feasible.

It was traditionally held as axiomatic that no microorganism could thrive in the highly acidic environment of the stomach. This view was overturned when Warren [17] demonstrated the existence of a hitherto unidentified strain of curved bacilli on the gastric epithelium in patients with chronic gastritis. This organism, now known as *Helicobacter pylori*, is regarded as playing a major role in the etiology of peptic ulceration, and perhaps gastric cancer also. Prospects for identifying and eradicating the organism from symptomatic patients are regarded as good [9]. Guidelines for **clinical trials** of eradicative regimes are given in [18].

Population gastroscopic screening for malignancy is practiced in Japan, where the incidence is much higher than in the West. Evidence from large population-based randomized trials [7, 13, 14] indicates a moderate benefit from **screening** for colorectal cancer using fecal occult blood tests – colonoscopy is also being developed as a screening method. As in other screening contexts, optimizing **sensitivity** and **specificity** is critical, both by avoidance of factors (here, dietary ones) that lead to a false result, and also by choice of test. Studies such as that by Hope et al. [10] point to the great need for methods to compare two tests for both sensitivity and specificity simultaneously. A suitable approach is now available [15].

Gastroenterology has been a major area for development of knowledge-based differential diagnosis systems. de Dombal et al. [5] described a computerized system to assist in the diagnosis of acute abdominal pain (*see* **Computer-aided Diagnosis**), which was able to outperform experienced clinicians. However, one of the greatest benefits was an educative one: as they used the system, clinicians directed their attention towards the factors that were most discriminatory, and their performance improved [6]. Knill-Jones et al. [12] described a model for diagnosis of jaundice. Clamp et al. [2] produced a scoring system for the differential diagnosis of inflammatory bowel disease. Spiegelhalter & Knill-Jones [16] outlined many of the statistical issues involved in scoring systems such as GLADYS (Glasgow Dyspepsia System), which grew out of de Dombal's work. An

excellent review of progress, including a balanced view of the usefulness to clinical practice, is given by de Dombal [4].

Gastroenterological journals are similar to other speciality journals in their statistical content. As an example of process quality improvement, *Gut* has had a statistical advisor for several years, and now has an identified team of some 20 statistical referees. Clinical trials will be required to conform to the CONSORT guidelines with effect from January 1998 [1].

References

[1] Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. & Stroup, D.F. (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement, *Journal of the American Medical Association* **276**, 637–639.

[2] Clamp, S.E., Myren, J., Bouchier, I.A.D., Watkinson, G. & de Dombal, F.T. (1982). Diagnosis of inflammatory bowel disease-international multicentre scoring system, *British Medical Journal* **284**, 91–95.

[3] Crohn, B.B., Ginzburg, L. & Oppenheimer, G.D. (1932). Regional ileitis. A pathological and clinical entity, *Journal of the American Medical Association* **99**, 1323–1329.

[4] de Dombal, F.T. (1987). Back to the future – or forward to the past, *Gut* **28**, 373–376.

[5] de Dombal, F.T., Leaper, D.J., Staniland, J.R., McCann, A.P. & Horrocks, J.C. (1972). Computer-aided diagnosis of acute abdominal pain, *British Medical Journal* **2**, 9–13.

[6] de Dombal, F.T., Leaper, D.J., Horrocks, J.C., Staniland, J.R. & McCann, A.P. (1974). Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians, *British Medical Journal* **1**, 376–380.

[7] Hardcastle, J.D., Chamberlain, J.O., Robinson, M.H.E., Moss, S.M., Amar, S.S., Balfour, T.W., James, P.D. & Mongham, C.M. (1996). Randomised controlled trial of fecal occult blood screening for colorectal cancer, *Lancet* **348**, 1472–1477.

[8] Harries, A.D., Baird, A. & Rhodes, J. (1982). Non-smoking: a feature of ulcerative colitis, *British Medical Journal* **284**, 706.

[9] Harris, A. & Miesiewicz, J.J. (1996). Treating *helicobacter pylori* – the best is yet to come?, *Gut* **39**, 781–783.

[10] Hope, R.L., Chu, G., Hope, A.H., Newcombe, R.G., Gillespie, P.E. & Williams, S.J. (1996). Comparison of three faecal occult blood tests in the detection of colorectal neoplasia, *Gut* **39**, 722–725.

[11] Jones, J.H., Lennard-Jones, J.E., Morson, B.C., Chapman, M., Sackin, M.J., Sneath, P.H.A., Spicer, C.C. & Card, W.I. (1973). Numerical taxonomy and discriminant analysis applied to non-specific colitis, *Quarterly Journal of Medicine* **42**, 715–732.

[12] Knill-Jones, R.P., Stern, R.B., Girmes, D.H., Maxwell, J.D., Thompson, R.P.H. & Williams, R. (1973). Use of sequential Bayesian model in diagnosis of jaundice by computer, *British Medical Journal* **1**, 530–533.

[13] Kronborg, O, Fenger, C., Olsen, J., Jørgensen, O.D. & Søndergaard, O. (1996). Randomised study of screening for colorectal cancer by screening for faecal occult blood, *Lancet* **348**, 1467–1471.

[14] Mandel, J.S., Bond, J.H., Church, T.R., Snover, D.C., Bradley, G.M., Schuman, L.M., and Ederer, F. (1993). Reducing mortality from colorectal cancer by screening for faecal occult blood, *New England Journal of Medicine* **328**, 1365–1371.

[15] Newcombe, R.G. (1996). Simultaneous comparison of sensitivity and specificity of two tests. A straightforward graphical approach, Presented at Royal Statistical Society conference, Guildford, September.

[16] Spiegelhalter, D. & Knill-Jones, R.P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroenterology, *Journal of the Royal Statistical Society, Series A* **147**, 35–76.

[17] Warren, J.R. (1983). Unidentified curved bacilli on gastric epithelium in active chronic gastritis, *Lancet* **1**, 1273.

[18] Working Party of the European *Helicobacter pylori* Study Group (Working Party coordinators: F. Megraud, C. O'Morain & P. Malfertheiner) (1997). Guidelines for clinical trials in *Helicobacter pylori* infection, *Gut* **41**, Supplement 2.

R.G. NEWCOMBE

Gauss, Carl Friedrich

Born: April 30, 1777, in Brunswick, Germany.
Died: February 23, 1855, in Göttingen, Germany.

Described in [1] as "one of the greatest scientific virtuosos of all time", Gauss came from a poor and semiliterate family, and quickly revealed himself as a highly intelligent child and a calculating prodigy. He was supported from 1792 by a stipend from the Duke of Brunswick, which enabled him to study in Brunswick, and later at Göttingen and Helmstedt. During this period he produced a stream of original mathematical and astronomical work, and in 1807 he became director of the Göttingen observatory, where he remained for the rest of his life.

May [1] lists the subjects in which Gauss worked as follows: arithmetic, number theory, algebra, analysis, geometry, probability and statistics, astronomy, geodesy, geomagnetism, mechanics, dioptrics, and physics. In all these he had an enormous international reputation. He was something of a lone worker, with no major mathematical collaborators and little personal contact with other mathematicians. He collaborated to a greater extent in applied work, particularly with an experimental physicist, Wilhelm Weber.

Gauss's importance in statistics rests on his work in the theory of **least squares**, and the central role played in it by the **normal (or Gaussian) distribution**. Least squares had been proposed by A.M. Legendre in 1805 as an intuitively satisfactory way of combining observations. Gauss's account, published in 1809, underpinned the "principle" by a "theory". Using what we should now describe as a **multiple linear regression** model, and a **Bayesian** formulation with **uniform prior** distributions for the parameters, he chose the mode of the posterior distribution (or, equivalently, the **maximum likelihood** (ML) solution). He considered that, for a single sample from a distribution, the arithmetic **mean** was a reasonable estimator, and showed that this implied the use of the normal distribution, which in turn led to the least squares solution to the ML equations. He showed also that, for general error distributions, the least squares solution minimized **mean square error** amongst all linear estimators (the result usually called the Gauss–Markov Theorem). For fuller accounts, see [2] and [3].

May's memoir [1] provides a broad coverage of Gauss's scientific work, and an extensive bibliography.

References

[1] May, K.O. (1981). Gauss, Carl Friedrich, in *Dictionary of Scientific Biography*, Vol. 5, C.C. Gillespie, ed. Scribner, New York, pp. 298–315.

[2] Sprott, D.A. (1983). Gauss, Carl Friedrich, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 305–309.

[3] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press, Cambridge, Mass.

PETER ARMITAGE

Gavarret, Louis–Denis–Jules

Born: January 28, 1809, in Astaffort, France.
Died: August 21, 1890, in Château de Valmont, France.

Jules Gavarret was born into a middle-class family and initially planned on a military career. To pursue this end, he enrolled in the École Polytechnique in 1829 and was named lieutenant of artillery in 1831; however, he resigned his military position in 1833 to pursue a career in medicine. Subsequently, Gavarret began an extensive collaboration with the eminent Parisian clinician Gabriel Andral with whom he conducted several investigations into the composition of blood. In 1843, Gavarret became a doctor by defending a thesis entitled "De l'Emphysème des Poumons et ses Rapports Avec les Différentes Maladies du Coeur et des Bronches" and ascended to the chair of medical physics in the Paris Faculty of Medicine. During his ensuing 44 year career in academic medicine, Gavarret wrote several more studies that attempted to expand the understanding of how physical principles (heat, electricity, etc.) could be used to explain the functioning of the human body; he retired in 1887 after being awarded numerous honors for his scientific accomplishments.

While still a student of Andral, Gavarret wrote the book that would be his principal legacy to the field of biostatistics: *Principes Généaux de Statistique Médicale, ou Développement des Règles Qui Doivent Présider à Son Emploi* (1840). In this book, Gavarret responded to a debate that had occurred at the Paris Academy of Medicine in 1837 over **Pierre Charles Alexandre Louis's** "numerical method". According to Louis, one should record the numbers of patients who died and recovered from each disease in the hospital wards as well as the types of treatments that each patient received. From these numerical results one could compute the percentage of patients who died after receiving a particular treatment and the percentage of patients who died after not receiving the treatment. If the former percentage was higher than the latter (as was the case for the then-common practice of bloodletting), the procedure should be suspended since it was not truly efficacious. In critiquing Louis's numerical method, Gavarret drew on the probabilistically based work of the contemporary mathematician **Siméon Denis Poisson** to show that the percentages could vary between "limits of oscillation" which depended on the number of cases observed. By applying Poisson's idea to Louis's various numerical conclusions, Gavarret determined, for example, that the average mortality rate from typhoid fever could vary between 26% and 49% on the basis of the 140 cases observed.

Throughout the latter half of the nineteenth century Gavarret's probabilistic analysis of medical statistics was discussed in treatises produced in America, Great Britain, France, and Germany. Although some German commentaries attempted to modify Gavarret's formulas to make them more useful in practice, most accounts did little more than repeat Gavarret's examples verbatim and highlight the "novelty" of applying the "calculus of probabilities" to medical statistics. As a result, few nineteenth-century biostatisticians saw Gavarret's probabilistic concerns as central to their work; nevertheless, Gavarret's approach is still important historically since it foreshadows the probabilistically informed concerns that have become so common within contemporary biostatistical research.

J. ROSSER MATTHEWS

Geisser, Seymour

Born: October 5, 1929 in the Bronx, New York.
Died: March 11, 2004 in St. Paul, Minnesota.

Seymour Geisser was a leading proponent of the importance of **prediction** in the practice of statistics as well as a leading exponent of **multivariate analysis**. He was also a prominent administrator and consultant in the fields of statistics and biostatistics.

Seymour's parents emigrated from Poland to New York City in the early 1920s where they became garment workers. He was a student at Lafayette High School in Brooklyn and received his undergraduate degree in mathematics from the City College of New York in 1950. "It was quite a chore to get up to City College from where I lived", said Geisser, as he remembered his undergraduate days. "City College was up on Convent Avenue and 137th Street and I lived down in Bensonhurst [Brooklyn]. It took almost two hours going and two hours coming back. I spent a lot of time sleeping on the subway."

His interest in statistics arose from conversations with his cousins. His cousin and his cousin's wife (Leon and Dorothy Gilford) were statisticians who assured him that the field provided good opportunities for employment. Dorothy Gilford worked for the Census Bureau and had been trained by **Harold Hotelling** at Columbia University. They suggested that Seymour apply for graduate school at Hotelling's new institution, the University of North Carolina. "When I left City College to go to Chapel Hill, I thought I was entering a country club," reflected Geisser. "It was such a pretty campus." Many of his fellow students went on to prominent careers in statistics. **Ralph Bradley**, Sudish Ghurye, Ingram (Red) Olkin, Milton Terry, Sutton Munroe were already students there. Others, like Don Burkholder, Ted Colton, Fred Descloux, Ed Gehan, Shanti Gupta, Jack Hall, T.V. Narayana, Bill Howe, Jim Pacheres, K.D. Ramachandran, Bill Thompson, John Wilkinson, and Marvin Zelen were contemporaries.

In graduate school, Seymour selected Harold Hotelling as his major professor and wrote a Master's thesis on computing roots and characteristic vectors of matrices (*see* **Matrix Algebra**). His Ph.D. thesis was on the mean square successive difference in statistics. During the summers of 1952 and 1953, while still a graduate student, he worked at the Aberdeen Proving Ground in Maryland. This is where he formulated his Ph.D. thesis problem, which was an extension of the work von Neumann had done there. His Masters and Ph.D. degrees were conferred in 1952 and 1955, respectively.

His first position after graduate school was as an assistant to Herman Chernoff at the National Bureau of Standards (NBS). He worked at a branch of the Operations Research Office located in the Army War College. Others who worked for NBS at the time were Jack Youden, Churchill Eisenhart, Marvin Zelen, Bill Connor, Bill Clatworthy, and Norman Severo. He then joined the Commissioned Officers Corps of the US Public Health Service, with the understanding that he would be assigned to the National Institute of Mental Health, working under **Sam Greenhouse**. **Jerry Cornfield**, **Max Halperin**, **Nathan Mantel**, and **Marvin Schneiderman** also worked at the **National Institutes of Health (NIH)**. Seymour spoke fondly of that time. "We used to eat lunch together and talk about everything from history, to statistics, to religion, to politics. The table talk was really interesting." In those conversations, Jerry Cornfield provided Seymour with his first introduction to **Bayesian** ideas.

His first Bayesian work was in multivariate analysis. In the early sixties, Seymour wanted to see what would happen if he considered the usual multivariate problems from the Bayesian point of view. After looking at **multivariate analysis of variance**, he turned to **classification** and discrimination. Geisser said "It dawned on me that with Bayesian Theory you didn't have to make a [strict] separation for linear discrimination. For example, everything on one side was guilty and the other side innocent, if you like. [A Bayesian] could find the probability of each individual being one or the other. It was a much finer distinction than using, say, the usual Fisher linear discriminant. That really swung me to the Bayesian approach." (*see* **Discriminant Analysis, Linear**).

His appointments at the National Bureau of Standards and the National Institute of Mental Health were from 1955 to 1961. From 1961 until 1965, he was Chief of the Biometry Section at the National Institute of Arthritis and Metabolic Diseases. At this time he also began teaching. From 1960 to 1965, he was a professorial lecturer at George Washington University, teaching in the Statistics Graduate Program at night. Sam Greenhouse and Solomon Kullback were also on the faculty.

In 1965, he became the founding Chair of the Department of Statistics at the State University of New York (SUNY), Buffalo, remaining in that capacity until 1970. During his tenure there, the faculty included Norman Severo, Bill Clatworthy, Marvin Zelen, Manny Parzen, Charles Mode, Jack Kalbfleisch, Ross Prentice, and Peter Enis, who was Seymour's first Ph.D. student at George Washington. When asked about his efforts to build the SUNY, Buffalo department, he said, "Building a group is difficult. There are lots of ups and downs. When you have money, so does everybody else. So you are competing for some very good people. That becomes a difficult chore. And when you don't have money, you can't hire anybody. So a lot of time is spent haggling and fighting with deans about lines, space, money: the usual trinity."

During the 1960s, Seymour developed a strong interest in prediction. This started with his work on classification, which is essentially a prediction problem. His thinking during this period is summarized in [8]. From this point on, his professional efforts were largely driven by his strong belief that the majority of statistical endeavor should be focused on prediction of observables rather than on estimation of unobservable parameters. "It always seemed to me," said Geisser, "that prediction was critical to modern science. There are really two parts, especially for statistics. There is description; that is, you are trying to describe and model some sort of process. [The model] will never be true and, essentially, you introduce lots of artifacts... Prediction is the one thing you can really talk about because what you predict will either happen or not happen and you will know exactly where you stand... Science changes when predictions do not come true." The majority of his work on prediction is summarized in Geisser [12].

In 1971, he became the founding Director of the School of Statistics at the University of Minnesota, remaining in that position until 2001. In the early years, the faculty included Don Berry, Kit [Chris] Bingham, Bob Buehler, Dennis Cook, Somesh Das Gupta, Joe [Morris L.] Eaton, Steven Fienberg, Cliff Hildreth, David Hinkley, F. Kinley Larntz, Bernie Lindgren, David Lane, Frank Martin, Michael Perlman, Milton Sobel, and Sandy Weisberg. During his tenure at Minnesota, there was an emphasis on the foundations of statistics. Geisser commented, "We brought in of a lot of people who were interested in foundations. We had a lot of seminars on it and there was a lot of interesting work that was done on foundations at that time. That was, in a sense, more interesting than methodology. [Without foundations, statistics is] just a trite engineering problem." In this regard, the School held a series of lectures on **R.A. Fisher's** contributions, which led to the monograph, *R. A. Fisher: An Appreciation*, in 1980 [2]. Seymour taught a graduate course called "*Statistical Inference*" for many years. In that course, he discussed the various modes of statistical **inference** and gave thought provoking lectures about the relative advantages and disadvantages of these modes. Recently, he compiled his lecture notes into a manuscript entitled *Modes of Parametric Statistical Inference* [13].

In the late 1980s and early 1990s, Seymour developed an interest in forensic DNA profiling. He was involved as an expert in over 100 litigations involving murder, rape, paternity, and other issues (*see* **Expert Witness, Statistician as**). His experiences in dealing with the FBI throughout these litigations are catalogued in the article "Statistics, litigation and conduct unbecoming," published in [4]. His purpose in these litigations was to point out that statistical calculations displayed in court should be valid. It was his contention that the methods then being used by the prosecution in DNA cases were flawed.

In discussing the people who most influenced his career, in addition to Harold Hotelling and Jerry Cornfield, he mentioned **George Barnard**. "I was particularly influenced by George Barnard. I always read his papers. He had a great way of writing. Excellent prose. And he was trained in philosophy, in logic, at Cambridge." Seymour's favorite statistics books were Fisher's, *Statistical Method in Scientific Inference* [3] and Cramer's, *Mathematical Methods of Statistics* [1]. Seymour chose Cramer for the mathematics of statistics and Fisher for the philosophical underpinnings of statistics. Late in life he said, "I still read those books. There always seems to be something in there I missed the first time, the second time, the third time..."

Seymour authored or coauthored 176 scientific articles, discussions, book reviews, and books over his career. One of his articles, Greenhouse and Geisser [14], is a citation classic. He pioneered several important areas of statistical endeavor. He and Mervyn Stone simultaneously and independently invented the now popular method for validating statistical models called **cross-validation**. Geisser [9] developed the equivalent method of "predictive

sample reuse." He also pioneered the areas of **Bayesian Multivariate Analysis** and discrimination [5, 6, 7, 17], Bayesian diagnostics for statistical prediction and estimation models [15, 16], Bayesian interim analysis [10] and testing for **Hardy–Weinberg equilibrium** using forensic DNA data [11].

Seymour served on many committees of the NIH, **Food and Drug Administration**, National Institute of Statistical Science, and National Research Council. In addition, he was a National Science Foundation Lecturer in Statistics from 1966 to 1969, a member of the National Research Council Committee on National Statistics from 1984 to 1987, Chair of the National Academy of Sciences panel on Occupational Safety and Health Statistics from 1986 to 1987. He delivered the **American Statistical Association** President's Invited Address in 1991.

He held numerous visiting professorships including the University of Tel Aviv, 1971; Stanford University, 1976, 1977, 1988; Harvard University, 1981; the University of Chicago, 1985; the University of Warwick (England), 1986. He was the Lady Davis Visiting Professor, Hebrew University of Jerusalem, 1991, 1994, 1999, and the Schor Scholar, Merck Research Laboratories (2002–2003). He was a Fellow of the Institute of Mathematical Statistics and the American Statistical Association.

Two special conferences were convened to honor his contributions to statistics. The first was organized by Jack Lee and held at the National Chiao Tung University of Taiwan in December of 1995. The second was organized by Glen Meeden and held at the University of Minnesota in May of 2002. In conjunction with the former conference, a special volume entitled *Modeling and Prediction: Honoring Seymour Geisser,* edited by Lee et al., was published in 1996 [18].

Seymour had an avid interest in history, archeology, particularly biblical archaeology, and religion, philosophy, and literature. He was a prolific reader of novels. He studied Latin, French, and German. He enjoyed traveling, especially to wildlife preserves and national parks.

He married his first wife while a graduate student in Chapel Hill. They had four children: Mindy, Dan, Georgia, and Adam. He met his second wife, Anne, while visiting Harvard University. They were married for 22 years. His brother, Martin, was a high school teacher and counselor. He had five grandchildren, Emma, Liam, and triplets Joshua, Eden, and Rachel.

The Department of Statistics at the University of Minnesota has established the Geisser Lectureship in Statistics. Each year, starting in the fall of 2005, an individual will be named the Seymour Geisser Lecturer for that year and will be invited to give a special lecture. Individuals will be selected on the basis of excellence in statistical endeavor and their corresponding contributions to science, both statistical and otherwise. For more information, visit the University of Minnesota Department of Statistics web page, `www.stat.umn.edu`.

References

[1] Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[2] Fienberg, S.E. & Hinkley, D.V. (1980). *R. A. Fisher: An Appreciation*. Springer-Verlag, New York.

[3] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Oliver & Boyd, Edinburgh.

[4] Gastwirth, J.L. (2000). *Statistical Science in the Courtroom*. Springer-Verlag, New York, 71–86.

[5] Geisser, S. & Cornfield, J. (1963). Posterior distributions for multivariate normal parameters, *Journal of the Royal Statistical Society B* **25**, 368–376.

[6] Geisser, S. (1964). Posterior odds for multivariate normal classification, *Journal of the Royal Statistical Society B* **1**, 69–76.

[7] Geisser, S. (1965). Bayesian estimation in multivariate analysis, *Annals of Mathematical Statistics* **56**, 150–159.

[8] Geisser, S. (1971). The inferential use of predictive distributions, in *Foundations of Statistical Inference*, V.P. Godambe & D.A. Sprott, eds. Holt, Rinehart, and Winston, Toronto, 456–469.

[9] Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association* **70**, 320–328.

[10] Geisser, S. (1992). On the curtailment of sampling. *Canadian Journal of Statistics*, **20**, 297–309.

[11] Geisser, S. & Johnson, W.O. (1992). Testing Hardy-Weinberg equilibrium on allelic data from VNTR loci, *American Journal of Human Genetics* **51**, 1084–1089.

[12] Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall, New York.

[13] Geisser, S. (2005). *Modes of Parametric Statistical Inference*. John Wiley & Sons, pp. 224.

[14] Greenhouse, S.W. & Geisser, S. (1959). On methods in the analysis of profile data, *Psychometrika* **24**, 95–112.

[15] Johnson, W.O. & Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis, *Journal of the American Statistical Association* **78**, 137–144.

[16] Johnson, W.O. & Geisser, S. (1985). Estimative influence functions in the multivariate general model, *Journal of Statistical Planning and Inference* **11**, 33–56.

[17] Lee, J.C. & Geisser, S. (1972). Growth curve prediction, *Sankhya* **A**, 393–412.
[18] Lee, J.C., Johnson, W.O. & Zellner, A. (1996). *Modeling and Prediction: Honoring Seymour Geisser*. Springer-Verlag, New York.

RONALD CHRISTENSEN & WESLEY JOHNSON

Gene Conversion

Gene conversion is said to occur when one of the alleles, "A" or "a", at a locus, is converted to the other allele during the process of replication. In diploid species, if the paternal **genotype** at a locus is "AA" and the maternal genotype is "aa", then all offspring are anticipated to have genotype "Aa" under normal circumstances. Occasionally, offspring from this mating are observed to have either genotype "AA" or genotype "aa". This implies that one allele has been replaced or converted by the other into a form like itself. Gene conversion frequently occurs in association with recombination in the flanking sequences, and the prevailing hypothesis is that gene conversion happens as a consequence of recombination and erroneous mismatch repair in the small segment of heteroduplex **DNA** that contains one strand (allele) from each parental molecule. Gene conversion drives the evolution of some genes. Gene conversion events are difficult to detect in species such as humans, where all the products of meiosis cannot be observed, but are easily identified in fungi such as yeast or *Neurospora*, where all the four products of meiosis (*see* **Linkage Analysis, Model-based**) are contained in a 4- or 8-spore ascus.

S. IYENGAR

Gene Expression Analysis

A (protein coding) **gene** is determined to be *expressed* in a cell or group of cells when its transcribed messenger RNA (mRNA) or the resulting protein product is detected (*see* **DNA Sequences**). There is a wide variety of techniques for determining and quantifying gene expression, and many of these have substantial statistical components to them. In this article we review some of the statistical models and methods used in analyzing gene expression data, focusing entirely on approaches quantifying mRNA. Before doing so, we present a small sample of the extensive biological and technological background to gene expression analysis.

Why do we measure gene expression? The most common experiment is *comparative*: we want to compare the mRNA levels of one or more genes in cells from different sources. Comparisons of interest include tumor vs. normal cells, cells from a specific organ in a mutant or genetically modified organism vs. cells from the same organ in a normal organism of the same strain, and cells before and after an intervention such as a drug treatment. Another important class are the time-course experiments, where cells are sampled at different times, e.g. after the administration of a drug, or as the cell cycle or development proceeds, and interest is in temporal patterns of gene expression. Yet other experiments focus on spatial patterns of gene expression. There are many other kinds of gene expression experiments, essentially as many as there are organisms, cell types, and conditions of biological interest.

How do we measure gene expression? As stated above, there are many techniques for doing so, but most rely on DNA–RNA or DNA–DNA *hybridization*. This is the biophysical process through which single-stranded DNA or RNA molecules find and base-pair with their complementary sequences amidst a complex mixture of many molecules of the same kind. The terminology we adopt names the sequence representing a gene of interest, the *probe*, while the pool within which a complementary copy of the probe is sought is named the *target* DNA or RNA. Other terminologies are the reverse of ours.

On what scale do we measure gene expression? Much of the recent interest by statisticians in this area stems from the availability of data sets giving expression measurements on tens of thousands of genes, so-called *microarray* gene expression data. However, nylon membrane filters with thousands of genes spotted on them have been around for over a decade, and smaller-scale quantitative expression data for much longer. We begin with a discussion of the first and simplest method of quantifying RNA, as many of the features of the high-throughput methods are already present here.

Low-throughput Methods

Quantitative Northern Blots

Isolated mRNA is separated according to size by electrophoresis, and transferred by blotting to an immobilizing matrix such as a nylon membrane. A labeled DNA probe is incubated with the blot under conditions that promote annealing, and the probe will then bind to the RNA molecules on the blot complementary to it. This is the hybridization reaction. The result is then imaged, either directly [e.g. by laser scanning or the use of a charge-couple device (CCD) camera], or indirectly, by exposing an X-ray film to the blot.

The amount of RNA can be quantified by measuring the intensity of the signal in the image in regions corresponding to the probe of interest. Usually control RNA is measured at the same time, typically a gene that is thought to be expressed at a more or less constant level, (a "housekeeping" gene), and the expression level of the gene of interest is then given relative to the control gene.

Although this technique has been in use for over 20 years, it has attracted little attention from statisticians. In part this is because low-throughput assays with simple read-outs are usually seen as outside the domain of statistical analysis, apart from simple matters such as analyzing replicate data. This attitude changes when the assay becomes high-throughput, or when much more data are collected on a given unit. These considerations lead naturally into our next topic, which is an important development of the northern blot.

Quantitative PCR, Including Real-time PCR

The polymerase chain reaction (PCR) can be used to estimate the concentration of a particular target RNA relative to a standard. Standards are control sequences

(such as housekeeping genes) that are present in the same preparation of RNA as the target sequence. Quantification is achieved by amplifying the target RNA (and the reference RNA) to a more readily detectable quantity, and by comparing the amount of amplified product generated by the standard and the target sequences.

The method works well if the amplified products are measured during the exponential phase of the chain reaction, if the reference and target sequences are present in approximately equal concentrations, and if they amplify with equal efficiency. More accurate variants involve adding reference molecules in known amounts to a series of amplification reactions.

A more recent technique for quantitating RNA is real-time PCR (RT-PCR) [14], where the target and reference sequences are amplified and detected in the same instrument, and the endpoint is when the reported fluorescence passes a fixed threshold above baseline. There are a number of different protocols, including TaqMan, and a number of different instruments for carrying out this assay. Details can be found in the technical notes from PE Applied Biosystems [24, 25] and Roche Molecular Biochemicals [27]. Rather more statistical research has been devoted to improving quantification methods for RT-PCR; see, for example, [26], but there are still many issues remaining. This is a fertile area for biostatisticians.

High-throughput Methods

Serial Analysis of Gene Expression

Serial analysis of gene expression (SAGE) is a method for comprehensive analysis of gene expression patterns. It is the main quantitative approach to gene expression not based upon hybridization. Three principles underlie the SAGE methodology: (a) a short sequence tag (10–14 bp) contains sufficient information to uniquely identify an mRNA transcript, provided that the tag is obtained from a unique position within each transcript; (b) sequence tags can be linked together to form long serial molecules that can be cloned and sequenced; and (c) a count of the number of times a particular tag is observed provides the expression level of the corresponding transcript.

A typical SAGE experiment would involve two sources of mRNA, say tumor and the corresponding normal tissue. For each source a set (called a library) of (say) 50 000 tags would be derived using the SAGE protocol. In these two libraries there might be 20 000 distinct (termed unique) tags observed, and for each unique tag, the frequency with which that tag appeared in each library could be calculated. The data for this comparative experiment are then two lists of counts, one for each unique tag observed.

The first question a biologist asks here is: Which tags are significantly differentially represented in the two libraries? For any given tag, say tag i, the natural **null hypothesis** here is H_i : the proportions of tag i in the two libraries coincide. Rejection of this null hypothesis leads to the conclusion that the gene corresponding to tag i is differentially expressed between the two sources of RNA.

Making an independence assumption that might be difficult to verify, one current approach to this question starts with the observation that under H_i, the number of times tag i appears in library 1, say, given the total number across the two libraries, is binomial with $p = 1/2$. This is the basis of a test of H_i, and when this is done for all $i = 1, \ldots, 20\,000$, a **Bonferroni** adjustment can be used. The test just described is one of a number in use [2, 21]. There is a range of outstanding questions with these data including: dealing with sequencing errors, which might be of the order of 1%–3% per base in the tags; considering the independence assumption leading to the binomial model; and seeking a valid multiple testing correction less conservative than Bonferroni. The difficulty is that because of the co-expression of genes, different tag counts in a library cannot be regarded as independent. However, the extent to which this matters is not yet clear. When more SAGE libraries accumulate in a given context, questions will undoubtedly arise that lead naturally to classification and cluster analyses; see below in the context of microarray data. As with the technologies outlined above, there seem to be many opportunities for biostatistical research involving SAGE data. A general source on this topic is `http://www.sagenet.org`.

Array-based Approaches

The principal class of high-throughput methods for quantifying gene expression are those based on microarrays, although the term "macroarray" is also used. Broadly speaking there are three basic microarray technologies: nylon membrane arrays, spotted

arrays, and high-density oligonucleotide arrays. The special supplement of *Nature* [23] provides a good overview of the production and utilization of the last two technologies.

We explain each briefly before turning to statistics. There we will attempt to discuss the issues in a general way when applicable to two or more of these technologies, and leave the reader to consult the references for material on topics rather more specific to the different technologies.

Different Array Technologies

Nylon Membrane Filters

This is the oldest array technology, but one that is still widely used around the world. A typical filter microarray has 5000 *complementary DNA* (cDNA) clones, 600–2400 bases in length, spotted in a grid on the membrane. Radio-labeled target cDNA derived from the mRNA of interest is hybridized to the array, and the filter is then exposed to X-ray film and the film imaged. The resulting digital image constitutes the raw data from the experiment.

A very high-density variant of the traditional filter-based microarray is the oligonucleotide filter array, which can have 50 000 spots consisting of pools of 10-mers [22].

Spotted cDNA Microarrays

Introduced in [29], a typical spotted array consists of 5000–20 000 cDNA probes of length 600–2400 bp placed in a regular pattern on a glass microscope slide. The main advantage of the nonporous glass support is that it facilitates miniaturization and the use of fluorescence (rather than radiolabel) based detection. Essentially all spotted arrays use two sources of mRNA. They are converted to cDNA and at the same time labeled with one of two fluorophores having different emission spectra following laser excitation. The labeled cDNAs are mixed in equal quantities and competitively hybridized to the spots on the slide. Following hybridization, laser excitation stimulates the spots to fluoresce, and the photons emitted are collected using band-pass filters tuned to each of the two fluorophores. These are then amplified, converted to digital form, and presented as two digital images of the slide, each quantifying the amount of cDNA on the spots labeled by one of the two fluorophores. These two digital images are the raw data of a spotted microarray experiment. In an obvious sense, each spotted array experiment may be regarded as several thousand paired comparisons.

A variant of the spotted arrays uses as probes long (60–75 bp) oligonucleotides representing part of a gene or expressed sequence tag [15]. These are put onto the glass using an ink-jet printer device, and generally lead to higher quality data. As with the original spotted arrays, a two-color system is used, although the technology may well be good enough to provide reliable single-color quantification.

High-density Oligonucleotide arrays

A quite different technology can be used to place up to 250 000 short (25 bp) oligonucleotide probe pairs on a small glass chip, with 16 or 20 of these probe pairs representing a part or all of a single gene, see [12] and [20]. Each probe pair consists of a perfect match (PM) probe, and a mismatch (MM) probe, the latter being the same as the former apart from a single nucleotide change (A $\leftrightarrow$ G or C $\leftrightarrow$ T) in the middle (13th) position. A tagged target cRNA sample hybridizes with the complementary oligonucleotides on the chip, and detection is via laser excitation followed by the collection of fluorescence emission, as with spotted arrays. As with the approaches already discussed, the image is the starting point of analysis.

Statistical Issues

Design of Experiments

The careful design of microarray experiments is in its infancy. Most work to date concerns spotted array experiments, which require more care by virtue of the paired nature of each experiment. Also, many users of spotted arrays construct the arrays themselves, whereas filter arrays and high-density oligonucleotide arrays tend to be bought "off the shelf". In spotted array experiments, there are two main aspects to the design question: (a) the design of the array itself, i.e. deciding which cDNA probe sequences to print on the slide, whether to use replicated spots and control sequences, and how many and where these should be

printed on the slide; and (b) the allocation of mRNA target samples to the slides, i.e. deciding how the mRNA samples should be paired for hybridization, the dye assignments, and the type and number of replicates.

Proper experimental design is needed to ensure that questions of interest can be answered and that this can be done accurately and efficiently given experimental constraints, such as cost of reagents and availability of mRNA. Designs specifically suited for the question of interest and judicious pairing of mRNA samples for hybridization can greatly improve the efficiency of microarray experiments by ensuring the precise measurement of relevant effects. A number of statisticians have been involved in these questions, but there is little literature so far. For some initial work in this area, see [16]. We can expect much more published research on this topic in the near future.

Image Analysis

As explained above, the "raw data" arising from all microarray technologies are images: of labeled probes on a nylon filter, a glass slide, or a glass chip. There seems little doubt that the results of downstream analyses can be appreciably influenced by the initial image analysis, though few studies of this topic exist at present; see [35] for one such.

Three broad analysis issues can be identified with microarray images, although not all approaches proceed in this way: finding the probe centers (registration); partitioning the pixels in the image into probe and nonprobe regions (segmentation); and assigning summary values to probe intensity and background (quantification). Rather than assign pixels to probe and nonprobe categories, some approaches (especially with nylon filters) use parametric, semiparametric, or nonparametric modeling to determine probe intensity. Once summary values of probe intensities are calculated, there remains the question of combining these to measure absolute or relative gene expression. With nylon filter and spotted arrays, intensity is usually the difference between foreground and background values, and ratios of these quantities are the main vehicle for later analysis. In general there are many ways of carrying out the image analysis, and several commercial and freely available packages for doing so, see [7] for nylon filter arrays [35], and references therein for spotted arrays, and [28] for high-density oligonucleotide arrays. Brandle et al. [5] is a good overall reference, and other articles in that volume can be consulted on this topic, as well as Buhler et al. [6] and Wang [32].

In the case of high-density oligonucleotide arrays, the image analysis does not result in expression values, but in PM and MM probe intensity values. One further analytical step is necessary with this technology before we have a gene (or probe set) expression value: the 16 or 20 PM and MM pairs must be summarized. This is not entirely straightforward and research on it is continuing, but see [18] and [19] for the most thorough published discussion to date.

Preprocessing Tasks: Normalization

As indicated earlier, the most common gene expression experiment is the comparative one. With nylon filter arrays this leads us to compare the images from two hybridizations onto copies of the same basic filter. Sometimes this is done by stripping the results of a first hybridization and re-using the filter, but more commonly a new filter is used. Because the nylon substrate is not solid, there may be warping, and this can make registration across different filters a challenging problem. When this is adequately addressed, interest focuses on comparing the two expression levels for each of the genes spotted onto the array. An entirely analogous situation arises when we have reduced the two images of a single spotted array or two high-density oligonucleotide array experiments to lists of gene expression values. We are back to the same (biologist's) question that we met with SAGE data: Which genes seem to be significantly differentially expressed between the two mRNA sources?

Before we can address this question in the microarray context, however, there is usually a need for normalization. This is a generic term describing the identification and removal of systematic sources of variation, other than differential expression, from the measured gene expression values. Systematic effects can come from different labeling efficiencies, different scanning parameters, and a variety of other causes, see [30] for a good list. These effects can be related to intensity, location on the filter, slide or chip, and other features of the process such as reagent batch and laboratory conditions. The need for normalization can be seen most clearly in

experiments involving two identical mRNA samples hybridized to different membranes or chips, or on the same glass slide, as long as the results are appropriately visualized.

Pairs of gene expression values, say from a treated (T) and a control (C) source, are usually displayed by plotting the $\log_2$ or $\log_{10}$ intensities against one another, e.g. $\log_2 T$ vs. $\log_2 C$. Such plots give an unrealistic sense of concordance between the two sets of intensities and can mask important features of the data. It is better to plot $M = \log_2 T/C$ vs. $A = \log_2 \sqrt{TC}$, which amounts to a rotation of the previous plot and a rescaling of the axes. Assuming, as is almost always the case, that we expect the majority of genes to be expressed at about the same level in both cell samples, regardless of overall intensity, the M vs. A plot should be scattered around the horizontal (A-) axis, in a more or less symmetrical manner, and the histogram of M values should be centered around zero. This is rarely found to be the case.

A standard normalization for nylon filter and spotted array data is to shift the log ratios so that their mean or median is zero. Frequently there is a strong enough intensity dependence that a smoothing of M values along the A-axis defines a better, A-dependent centering. Spatial effects require a modified solution, and there are yet other effects that need to be dealt with from time to time. For a discussion of these issues in the context of spotted arrays, see [34], while [30] is also of interest. Normalization is also relevant to the high-density oligonucleotide technology, but is less well discussed and somewhat more complex, see [18] and [19] for some comments.

Comparative Analyses

Once the log ratios of intensities have been normalized, interest focuses on those that seem to be genuinely different from zero, i.e. that correspond to genes that are differentially expressed. There is no reliable method of assigning statistical significance to log ratios from unreplicated experiments, although a number of model-based approaches claiming to do this can be found in the literature, see [9] for a discussion of this issue in the context of replicated spotted microarrays. For a single comparison, the best approach is probably to apply a careful normalization to the log ratios, rank them, and construct a normal $Q-Q$ plot of them. Typically the plot will not be linear, but an examination of the extremes in conjunction with the M vs. A plot can give a good sense of the outlier log ratios. It is also advisable to carry out a quality examination of the spots corresponding to extreme log ratios. Exactly where to draw the line with ranked log ratios, when determining putatively differentially expressed genes, will depend on a variety of factors, such as the shape of the $Q-Q$ plot, the level of false positive and false negative rates deemed acceptable, and the nature and number of follow-up experiments envisaged. No simple guidelines seem possible, and no formal statistical approach seems available which deals with the question. The situation is different when there are replicate pairs of filters, slides, or chips. We broaden the context somewhat to discuss the issue of multiple testing (*see* **Multiple Comparisons**) more generally.

Multiple Testing

The identification of differentially expressed genes, i.e. genes whose expression levels are associated with a response or covariate of interest, is but one of the testing problems that arise with microarray data. The covariates could be either polytomous, e.g. treatment/control status, cell type, drug type, or continuous, e.g. dose of a drug, time, and the responses could be, for example, censored survival times or other clinical outcomes. The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates. As a typical microarray experiment measures expression levels for several thousands of genes simultaneously, we are faced with an extreme multiple testing problem. Special problems arising from the multiplicity aspect include defining an appropriate type I error rate (i.e. false positive rate) and devising powerful multiple testing procedures that control this error rate and account for the joint distribution of the gene expression levels.

A number of recent papers have addressed the question of multiple testing in the context of microarray experiments [10, 13, 31]. However, the proposed

solutions were not cast in the standard statistical framework and do not provide adequate type I error rate control. When going from single to multiple hypothesis testing, several definitions of the type I error rate are possible and include: the per-comparison error rate (PCER), defined as the expected value of (number of type I errors/number of hypotheses); the family-wise error rate (FWER), defined as the probability of at least one type I error; and the false discovery rate (FDR), or expected proportion of type I errors among the rejected hypotheses. In general, for a given multiple testing procedure, PCER $\leq$ FWER and FDR $\leq$ FWER, one should thus decide on an appropriate error rate to control for the problem under consideration. It is important to note that the expectations and probabilities above are conditional on assumptions concerning which hypotheses are true, i.e. on which genes are differentially expressed. A fundamental, yet often ignored distinction in multiple testing is that between strong and weak control of the type I error rate. Strong control refers to control of the type I error rate under *any* combination of true and false hypotheses, i.e. for any combination of differentially and constantly expressed genes. In contrast, weak control refers to control of the type I error rate only when *none* of the genes is differentially expressed, i.e. under the complete null hypothesis that all the null hypotheses are true. In general, weak control without any other safeguards is unsatisfactory. In the microarray setting, where it is very unlikely that none of the genes is differentially expressed, it seems particularly important to have strong control of the type I error rate.

Adjusted ***P* values** provide useful and flexible summaries of the strength of the evidence in favor of differential expression. The adjusted P value for a particular gene reflects the overall false positive error rate for the family of hypotheses when genes with smaller P values are declared differentially expressed. Adjusted P values may also be used to summarize and compare the results from different multiple testing procedures.

In their 1993 book, Westfall & Young [33] proposed resampling-based P value adjustment procedures that are highly relevant in the context of microarray experiments. In particular, these authors defined adjusted P values for multiple testing procedures that control the FWER and take into account the dependence structure between test statistics (their $\min P$ and $\max T$ adjusted P values). In Dudoit et al. [9] these ideas are applied in the context of microarray data. It is clear that this area is undergoing rapid development.

Classification and Clustering

Microarray experiments have revived interest in both cluster and **discriminant analysis** by raising new methodological and computational challenges. In discriminant analysis, also called supervised learning or class prediction, we might have observations on tumor mRNA samples known to belong to prespecified classes, and the task is to build predictors for allocating new observations to these classes. By contrast, in cluster analysis, also called unsupervised learning or class discovery, the classes are unknown a priori and the task is to determine these classes from the data themselves, i.e. to determine the number of classes and assign each observation to one of these classes. Either experiments or genes or both can be clustered, and the commonest approach uses hierarchical procedures based on correlation as a measure of dissimilarity. Clustering of this kind is currently the most popular way of analyzing gene expression data, undoubtedly because of the power of the technique to group co-expressed genes and hence shed light on the function of uncharacterized genes. For some examples see [3, 11] and [1].

The ability successfully to distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel genomic approach to cancer classification. There are already many papers on this topic, and almost every technique from the field of machine learning has already been applied to this problem. How do they compare? Are there advantages to the more recent or more elaborate classification techniques? While it is not possible to give a single long-term answer to this question, it is possible to obtain some insights. The study by Dudoit et al. [8] compared a number of familiar methods for classifying tumors based on gene expression data, including nearest neighbor classifiers, linear discriminant analysis, and classification trees. Two recent machine learning devices known as bagging and boosting were also considered. The discrimination methods were all applied to data sets from three recently published cancer gene expression studies, and the main conclusion, for these data sets, was that simple classifiers

such as diagonal linear discriminant analysis and nearest neighbors performed remarkably well compared with more elaborate ones such as aggregated classification trees. These conclusions may change as the size of data sets grows.

Other Topics

When expression levels are measured for thousands of genes in time and space, a challenging problem is to discover and recognize reproducible temporal expression patterns, including those not previously known. Networks of interacting genes that might suggest new biochemical or signaling pathways are of particular interest. Current approaches to this class of questions with microarray data are rather *ad hoc*, usually involving one- or two-dimensional clustering methods. These methods, typified by Eisen's "heat diagrams" [11], rearrange the order of genes and experiments to map the data onto a plane in a more visually compelling way. The hope is that visual examination of the resulting image will identify patterns to which explanations can be attached. Other researchers rely on multi-dimensional scaling, which uses distances between genes or arrays to produce a scatter plot in the plane for subsequent visual examination.

More systematic approaches are clearly needed, with [17] being one such. This are will undoubtedly attract a great deal of biostatistical research in coming years.

References

[1] Alizadeh, A.A., Eisen, M.B., Davis, R.E. et al. (2000). Different types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**, 503–511.

[2] Audic, S. & Claverie, J.-M. (1997). The significance of digital gene expression profiles, *Genome Research* **7**, 986–995.

[3] Bassett, D.E., Jr, Eisen, M.B. & Boguski, M.S. (1999). Gene expression informatics – its all in your mind, *Nature Genetics* **21**, Supplement, 51–55.

[4] Bittner, M.L., Chen, Y., Dorsel, A.N. & Dougherty, E.R., eds (2001). *Microarrays: Optical Technologies and Informatics. Progress in Biomedical Optics and Imaging*, Vol. 2, Proceedings of SPIE, Vol. 4266. The International Society for Optical Engineering, Bellingham.

[5] Brandle, N., Bischof, H. & Lapp, H. (2001). generic and robust approach for the analysis of spot array images, in *Microarrays: Optical Technologies and Informatics. Progress in Biomedical Optics and Imaging*, Vol. 2, M.L. Bittner et al., eds. The International Society for Optical Engineering, Bellingham, pp. 1–12.

[6] Buhler, J., Ideker, T. & Haynor, D. (n.d.). Dapple: improved techniques for finding spots of DNA microarrays. Technical Report, Department of Molecular Biotechnology, University of Washington, Seattle.

[7] Carlisle, A.J., Prabhu, V.V., Elkahloun, A., Hudson, J., Trent, J.M., Linehan, W.M., Williams, E.D., Emmert-Buck, M.R., Liotta, L.A., Munson, P.J. & Krizman, D.B. (2000). Development of a prostate cDNA microarray and statistical gene expression analysis package, *Molecular Carcinogenesis* **28**, 12–22.

[8] Dudoit, S., Fridlyand, J. & Speed, T. (2001). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, in press.

[9] Dudoit, S., Yang, Y.H., Callow, M.J. & Speed, T.P. (2001). Statistical methods for identifying genes with differential expression in replicated microarray experiments, *Statistica Sinica*, in press.

[10] Efron, B., Tibshirani, R., Goss, V. & Chu, G. (2000). Microarrays and their use in a comparative experiment. Technical Report, Stanford University Department of Statistics.

[11] Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences* **95**, 14863–14868.

[12] Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. & Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis, *Science* **251**, 767–673.

[13] Golub, T.R., Slonom, D.K., Tamayo, P. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression profiling, *Science* **286**, 531–537.

[14] Higuchi, R., Fockler, C., Dolinger, G. & Watson, R. (1993). Kinetic PCR: real time monitoring of DNA amplification reactions, *Biotechnology* **11**, 1026–1030.

[15] Hughes, T.R., Mao, M., Jones, A.R. et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer, *Nature Biotechnology* **19**, 342–347.

[16] Kerr, M.K. & Churchill, G.A. (2001). Experimental design got gene expression microarrays, *Biostatistics* **2**, 183–201.

[17] Kim, S., Dougherty, E.R., Chen, Y et al. (2000). Multivariate measurement of gene expression relationships, *Genomics* **67**, 201–209.

[18] Li, C. & Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proceedings of the National Academy of Sciences* **98**, 31–36.

[19] Li, C. & Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology* **2**, 32.1–32.11.

[20] Lockhart, D.J., Dong, H.L., Byrne, M.C. et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology* **14**, 1675–1680.

[21] Man, M.Z., Wang, X. & Wang, Y. (2000). POWER_SAGE: comparing statistical tests for SAGE experiments, *Bioinformatics* **16**, 953–959.

[22] Meier-Ewert, S., Lange, J., Gerst, H. et al. (1998). Comparative gene expression profiling by oligonucleotide fingerprinting, *Nucleic Acids Research* **26**, 2216–2223.

[23] *Nature* (1999). The Chipping forecast, Supplement to *Nature Genetics* **21**. See also: `http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v21/n1s/index.html`

[24] PE Applied Biosystems (1997). User Bulletin No. 2 36 pp.

[25] PE Applied Biosystems (1998). User Bulletin No. 5 20 pp.

[26] Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR, *Nucleic Acids Research* **29**, 2002–2007.

[27] Roche Molecular Biochemicals (2000). Technical Note No. LC/10.

[28] Schadt, E., Li, C., Su, C. & Wong, W.H. (2000). Analyzing high-density oligonucleotide gene expression data, *Journal of Cellular Biochemistry* **80**, 192–202.

[29] Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270**, 467–470.

[30] Schuchhardt, J., Beule, D., Mailk, A., Wolski, E., Eickhoff, H., Lehrach, H. & Herzel, H. (2000). Normalization strategies for cDNA microarrays, *Nucleic Acids Research* **28**, E47.

[31] Tusher, V.G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to transcriptional response to ionizing radiation, *Proceedings of the National Academy of Sciences* **98**, 5116–5121.

[32] Wang, X., Ghosh, S. & Gou, S.-W. (2001). Quantitative quality control in microarray image processing and data acquisition, *Nucleic Acids Research* **29**, E75.

[33] Westfall, P.H. & Young, S.S. (1993). *Resampling-Based Multiple Testing: examples and Methods for P-Value Adjustment*. Wiley, New York.

[34] Yang, Y.H., Dudoit, S., Luu, P. & Speed, T.P. (2001). Normalization for cDNA microarray data, in *Microarrays: Optical Technologies and Informatics. Progress in Biomedical Optics and Imaging*, Vol. 2, M.L. Bittner et al., eds. The International Society for Optical Engineering, Bellingham, pp. 141–152.

[35] Yang, Y.H., Buckley, M.J., Dudoit, S. & Speed, T.P. (2001). Image processing on cDNA microarray data, *Journal of Computational and Graphical Statistics*, in press.

TERRY P. SPEED

Gene Frequency Estimation

Gene frequency estimation refers to the estimation of population frequencies of alleles at a given genetic locus or to the estimation of the population frequencies of **haplotypes**. A haplotype, derived from the term "haploid genotype", refers to the particular set of alleles present at a series of linked loci on a chromosome, i.e. alleles at loci that are present relatively close together on a continuous strand of DNA. Each human being has two haplotypes for any given series of linked autosomal loci; one inherited maternally, and the other, paternally. The haplotype transmitted to an offspring of that individual may be identical to one of these two, or may reflect reshufflings due to recombination events. Allele and haplotype frequencies may be estimated from various types of human data, including information from pedigrees, parent–child dyads, or individuals; however, estimation based upon random samples of individuals is emphasized in this article.

Good estimates of gene frequencies are needed for a variety of purposes: they are essential to the generation of valid risk estimates in the context of **genetic counseling**, and form the basis for population descriptions and comparisons and for decisions about genetic screening of populations. Because of their impact upon **power** considerations, sound estimates of gene frequencies are also helpful in planning genetic studies. Marker allele frequencies are required for the application of certain types of **linkage analysis**, such as the Haseman–Elston model-free approach to sib-pair analysis in the absence of parental marker information [7], and the affected-pedigree-member method of Weeks & Lange [15]. Beyond the specifics of gene frequency estimation methodology outlined here, it must always be borne in mind that the sample from which inferences are made must be drawn from the appropriate population according to the needs of the analysis, and must be of a size to achieve the requisite precision. Care must also be taken that purportedly random samples have not been collected in such a way that it is likely that family members have been included.

Allele Frequency Estimation

When all allelic variants corresponding to a particular locus are codominantly expressed, **genotypes** can be fully discerned on the basis of phenotype, and the estimation of allele frequencies becomes a simple matter of "gene counting". Each individual bears two alleles at a particular locus, so that persons expressing only a single allele are known to be homozygous, possessing two copies of that particular allelic variant (*see* **Heterozygosity**). If a random sample of N individuals is collected, and there are X_i copies of allele A_i among the $2N$ alleles, then the estimator of the allele frequency q_i is $X_i/2N$, the **maximum likelihood** estimator. The **variance** is obtained using the standard **binomial** expression, i.e. $q_i(1-q_i)/2N$. However, if dominance relationships govern the phenotypic expression of genotypes at a given locus (for example, one allele masks the expression of another) the situation becomes more complicated. By way of illustration, suppose there are a number of codominantly expressed allelic variants at a given autosomal locus, and also a null allele that results in no detectable product. Then, individuals who are homozygous for a particular codominant allele cannot be distinguished phenotypically from those heterozygotes who have one copy of that allele and a copy of the null allele – both express only a single allele at the phenotypic level. The "gene counting" approach cannot be used in such instances unless haplotype analysis has been used to clarify exactly which alleles are present at all loci concerned; for example, by examination of transmission among family members [16] (*see* **Genetic Transition Probabilities**).

Whatever the nature of these relationships, a maximum likelihood approach may be taken. Suppose that a random sample of individuals is taken from the relevant population, and that there are K distinguishable phenotypes associated with the locus of interest, indexed by $k = 1, 2, \ldots, K$. Generally, an assumption must be made in order to set up the correspondences between genotype and phenotype, expressed via the allele frequencies; most typically, the **Hardy–Weinberg equilibrium** assumption is made so that the population frequency Φ_k of each phenotype can be expressed as a function of the allele frequencies. On the basis of the observation of n_k individuals with the kth phenotype, the **likelihood** of the sample is given by

$$L \propto \prod_{k=1}^{K} \Phi_k^{n_k}$$

and the ln likelihood as

$$\ln L = \sum_{k=1}^{K} n_k \ln(\Phi_k) + \text{a constant.}$$

For example, consider a locus with a codominant alleles $A_1, A_2, \ldots, A_i, \ldots, A_a$ plus a null allele A_x, which cannot be detected, with population allele frequencies $q_1, q_2, \ldots, q_a, q_x$. Then, there are $\binom{k}{2}$ phenotypes of the type $A_i\ A_j$, which occur with population frequency $2q_iq_j$, and k phenotypes of the type that express only a single allele A_i, which may be either A_i homozygotes or heterozygotes involving a null allele and occur with population frequency $(q_i^2 + 2q_iq_x)$, and a phenotype expressing no allele A_x, which occurs with population frequency q_x^2. If we define the sample counts of these phenotypes analogously, we obtain the log likelihood

$$\ln L = \sum\sum_{i<j} n_{ij} \ln(2q_iq_j) + \sum_i n_i \ln(q_i^2 + 2q_iq_x) + n_x \ln(q_x^2).$$

The estimates of the allele frequencies are obtained using standard iterative methods based upon the maximum likelihood approach, and estimates of the **standard errors** can be derived numerically via the Fisher **information** using standard methods. Excellent initial estimates can be obtained using the method of Bernstein [1]; these estimators are also based upon Hardy–Weinberg assumptions. If f_i denotes the proportion of individuals expressing the ith allele, then the Bernstein estimator is $\tilde{q}_i = 1 - (1 - f_i)^{1/2}$ (since under Hardy–Weinberg expectations this is $1 - [(1 - q_i)^2]^{1/2} = q_i$; q_x is generally estimated by subtraction [8]; and additional refinements of the Bernstein method exist [1, 8].

It may be noted that this same maximum likelihood approach is applicable even in the presence of more complex dominance relationships, and the Bernstein estimators generally provide reasonable initial values for iteration. Furthermore, another source of so-called "null" or "blank" alleles is the technical inability to detect particular alleles. For example, in the case of the **HLA system**, specificities were originally detected serologically, utilizing sera from multiparous women; particular alleles could therefore be designated as "blanks" not because they produced no product, but because the reagents that would permit their detection were not yet available. This led to a decreasing frequency of the blank allele as the entire constellation of alleles at a particular locus eventually came to be identifiable. The subsequent application of molecular techniques revealed that at least some serologically determined "alleles" could be further subdivided.

An alternative approach to allele frequency estimation in the presence of dominance relationships, which also yields the maximum likelihood estimate, employs the **EM algorithm** [3]. From this perspective, the genotypes of individuals may not be fully observable, but can be estimated under a specific genetic model with suitable assumptions, utilizing phenotypic data, which can be observed. Given phenotypic counts, and a provisional estimate of the allele frequencies, the expected numbers of genotypes and so of alleles can be obtained (generally under Hardy–Weinberg assumptions), and the gene counting approach used to update allele frequency estimates. Iteration continues in this manner until convergence; this approach is often referred to as "iterative gene counting" for obvious reasons.

Haplotype Frequency Estimation

If haplotyping has been carried out, then all haplotypes in the sample are known, and haplotype frequency estimation is accomplished in a straightforward manner by a process analogous to the gene counting procedure described previously: the number of haplotypes of a given type are counted and standard binomial theory applied to obtain the point estimates and their estimated standard errors [5]. However, in the absence of such information, haplotype frequencies are more typically estimated from observed phenotypic frequency data, based upon a random sample of individuals. Consider estimation of a particular combination of alleles A_i at the A locus and B_j at the B locus. Under equilibrium conditions, the frequency of the haplotype A_iB_j would be equal to the product of their two allele frequencies, p_i and q_j, respectively; however, this is not true in general, and this displacement from equilibrium is measured by the coefficient of disequilibrium $\Delta_{ij} = h_{ij} - p_iq_j$, where h_{ij} is the population frequency of the A_iB_j haplotype. Alternatively, the observation may be made that

$$h_{ij} = p_iq_j, +\Delta_{ij}.$$

This is the basis of one approach to haplotype frequency estimation [2, 11, 12] in which phenotypic data are first used to estimate the coefficient of disequilibrium under Hardy–Weinberg assumptions, as follows:

$$\Delta_{ij} = \sqrt{f}_{--} - [(f_{+-} + f_{--})(f_{-+} + f_{--})]^{1/2},$$

where f_{--} denotes the proportion of persons with neither the A_i nor the B_j allele, f_{+-} denotes the proportion of persons with the A_i but not the B_j allele, and f_{-+} denotes the proportion of persons with the B_j but not the A_i allele. The relevant allele frequencies, p_i and q_j, are then estimated using the phenotypic data as previously described, and the estimated quantities are then substituted into the equation $h_{ij} = p_i q_j + \Delta_{ij}$ to obtain the estimate of h_{ij}. An alternative iterative approach that provides maximum likelihood estimates [17] can also be employed.

It is important to note that the Hardy–Weinberg assumptions are critical to the estimation of haplotype frequencies from phenotypic data in the situations described. The utilization of such methods in an inappropriate context, such as an inbred population (*see* **Inbreeding**), can lead to invalid and misleading results. Kostyu et al. [10] demonstrated the importance of haplotyping in inbred populations and the dangers of inappropriate application of Hardy–Weinberg assumptions in conjunction with phenotypic data, which could seriously impair inferences about the nature of **linkage disequilibrium** and thereby produce erroneous haplotype frequency estimates. Moreover, it should be noted that very large sample sizes are required to achieve reasonable levels of power to detect deviations from Hardy–Weinberg proportions due to inbreeding in human populations [4] even on the basis of fully codominant systems; therefore, it may be unwise to assume that failure to reject Hardy–Weinberg assumptions for such loci implies that they may be safely made in general, particularly when the presence of inbreeding is known or suspected. In studies of inbred populations, the investigator may attempt to study the entire population. Failing that, it may be advisable to strive for both breadth and depth in sampling on the basis of social and anthropological considerations; for example, sampling from a large number of demographic units, such as colonies, and securing information from a large proportion of the members of each such unit. Other approaches have included restriction of the sample to persons of reproductive age and to considerations of persons of a single gender, such as females in a patrilocal system (e.g. where the woman relocates to the colony of her husband) [9, 13].

Although this treatment has emphasized inference based upon random samples from the population of interest, gene frequency estimates may be obtained from other types of data, although appropriate statistical methods must be employed. Allele frequency estimates may be obtained from family data [3, 6, 14, 16] and are typically estimated jointly with other genetic parameters; for example, in the context of **segregation analysis** of pedigrees [6]. However, such estimates are derived mainly from the information provided by persons marrying into the families in the sample; if the sample consists of many small pedigrees or nuclear families, then one would expect better estimates of allele frequencies than if a few large pedigrees were being analyzed. Furthermore, if the pedigrees were brought into the sample because of the presence of one or more affected individuals, the allele frequency estimates may be affected if the correction for **bias** due to **ascertainment** is not accomplished completely.

References

[1] Bernstein, F. (1930). Fortgesetzte untersuchungen aus der theorie der blutgruppen, *Z Inducktiven Abstammungs-Vererbungslehre* **56**, 233–273.

[2] Cavalli-Sforza, L.L. & Bodmer, W.F. (1971). *The Genetics of Human Populations*. Freeman, San Francisco.

[3] Ceppellini, R., Siniscalo, M. & Smith, C.A.B. (1955). The estimation of gene frequencies in a random-mating population, *Annals of Human Genetics* **20**, 97–115.

[4] Chakraborty, R. & Zhong, Y. (1994). Statistical power of an exact test of Hardy-Weinberg proportions of genotypic data at a multiallelic locus, *Human Heredity* **44**, 1–9.

[5] Dausset, J., Legrand, L., Lepage, V., Contu, L., Marcelli-Barge, A., Wildloecher, I., Benjam, A., Meo, T. & Degos, L. (1978). A haplotype study of HLA complex with special reference to the HLA-DR series and to Bf.C2 and glyoxalase I polymorphisms, *Tissue Antigens* **12**, 297–307.

[6] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[7] Haseman, J.R. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[8] Heuther, C.A. & Murphy, E.A. (1980). Reduction of bias in estimating the frequency of recessive genes, *American Journal of Human Genetics* **32**, 212–222.

[9] Kostyu, D.D., Ober, C.L., Dawson, D.V., Ghanayem, M., Elias, S. & Martin, A.O. (1989). Genetic analysis of HLA in the US Schmiedenleut Hutterites, *American Journal of Human Genetics* **45**, 261–269.

[10] Kostyu, D., Dawson, D., Ciftan, E., Stewart, A., Lewis, D., Parc, F., Laigret, J., McCollum, R. & Amos, B. (1984). HLA in two islands of French Polynesia, *Tissue Antigens* **23**, 217–228.

[11] Mattiuz, P.L., Ihde, D., Piazza, R., Ceppellini, R. & Bodmer, W.F. (1970). New approaches to the population genetic and segregation analysis of the HLA system, in *Histocompatibility Testing 1970*, P. Terasaki, ed. Williams & Wilkins, Baltimore, pp. 197–203.

[12] Mendell, N.R. & Ward, F. (1984). Statistical methods in human genetics and immunology, in *Mathematical Methods in Medicine*, D. Ingram & R.F. Block, eds. Wiley, New York, Chapter 4, pp. 161–223.

[13] Morgan, K., Holmes, T.M., Schlaut, J., Marchuk, L., Kovithavongs, T., Pazderka, F. & Dossetor, J.B. (1980). Genetic variability of HLA in the Dariusleut Hutterites. A comparative genetic analysis of the Hutterites, the Amish, and other selected Caucasian populations, *American Journal of Human Genetics* **26**, 489–503.

[14] Morton, N.E. & MacLean, C.J. (1974). Analysis of family resemblance. III. Complex segregation analysis, *American Journal of Human Genetics* **26**, 489–503.

[15] Weeks, D.E. & Lange, K. (1988). The affected-pedigree-member method of linkage analysis, *American Journal of Human Genetics* **42**, 315–326.

[16] Weeks, D.E., Sobel, E., O'Connell, J.R. & Lange, K. (1995). Computer programs for multilocus haplotyping of general pedigrees, *American Journal of Human Genetics* **56**, 1506–1507.

[17] Yasuda, N. & Tsuji, K. (1975). A counting method of maximum likelihood for estimating haplotype frequency in the HLA-A system, *Japanese Journal of Human Genetics* **20**, 1–15.

DEBORAH V. DAWSON

Gene

The word *gene* has been, and still is, used with various meanings. The genetic material of higher organisms is DNA (deoxyribonucleic acid), a constituent of the 23 pairs of chromosomes a person has. Each DNA molecule has a sequence of many bases along it (*see* **DNA Sequences**), most of which have no known function. About 3% of the bases code for various products, either polypeptide chains or molecules of ribonucleic acid, and a commonly accepted definition of a gene is a combination of DNA segments that together code for one of these functional units. Genes thus occur in pairs at locations, or loci, along the pairs of chromosomes. Different genes that can occupy the same locus are allelic, or alleles, so that there may be many alleles at a given locus in the population – but only two in any one person. Genes that occupy different loci are nonallelic. A common definition of alleles is given as "different forms of the same gene", so that the term "gene" then means the general type of DNA that occurs at a particular locus, as opposed to an allele, which is a specific sequence of DNA that occurs at the locus. It is unfortunate that the word gene is used with either of these two distinct meanings, perhaps better differentiated as "locus" and "allele". A rough analogy can be made with the distinction between "estimator" and "estimate", the latter being a specific instance of the former. English-speaking writers, earlier writers, and population geneticists have tended to use the word gene with the specific ("allele") meaning, whereas non-English writers, more recent writers, and biochemical geneticists tend to use it with the more general ("locus") meaning. (But biochemical geneticists talk of "cloning genes", meaning alleles.) A further disparity occurs when human geneticists speak of there being "two genes at each locus", whereas Drosophila geneticists speak of "two loci for each gene". Just as statisticians are rarely confused when the word estimate is used when estimator might be better (as in the distribution of an estimate), so are geneticists rarely confused by the word gene, whose exact meaning is always clear (to them) by the context.

ROBERT C. ELSTON

Gene-environment Interaction

The interdependent action of **genes** and environment in disease causality is measured by gene–environment interaction. When interaction exists, the combined action of genes and environment can increase or decrease disease risk beyond that due to purely genetic and purely environmental actions. While Haldane [3] first considered gene–environment interaction over half a century ago, evidence continues to emerge that most diseases do not result from entirely genetic or entirely environmental factors, but rather from a complicated interaction of these factors [7]. In fact, Khoury et al. [5] assert that the basis of **genetic epidemiology** comes from the evolving recognition that gene–environment interactions contribute to the etiology of most diseases.

Types of causal interaction include synergism, whereby both factors are needed for disease to occur, and antagonism, where each factor results in disease only when the other is absent [6]. The relation between the recessive gene for Phenylketonuria (PKU) and dietary phenylalanine in mental retardation provides an example of gene–environment interaction. PKU is a genetic disorder in which phenylalanine metabolism is blocked. The interaction between PKU and blood phenylalanine levels produces an increased risk of mental retardation beyond that resulting from either factor alone [10].

We focus here on statistical **interaction**, which does not necessarily imply interaction on the biological or mechanistic level. Statistical interaction is commonly measured by departures from additivity of effects on the chosen outcome scale. Following [8], let R_{ij} be the average risk when $G = i$ and $E = j$, where $G \in (g, \overline{g})$, $E \in (e, \overline{e})$, and $g(\overline{g})$ indicates the presence (absence) of the genetic factor of interest and $e(\overline{e})$ indicates the presence (absence) of the environmental factor of interest. Then one can express additivity on the risk ratio scale (*see* **Relative Risk**) as

$$\frac{R_{ge}}{R_{\overline{ge}}} = \frac{R_{g\overline{e}}}{R_{\overline{ge}}} + \frac{R_{\overline{g}e}}{R_{\overline{ge}}} - \frac{R_{\overline{ge}}}{R_{\overline{ge}}}. \quad (1)$$

That is to say, the risk ratio due to the combined action of the genetic and environmental effects is simply equal to the sum of the risk ratios for the genetic effect and for the environmental effect (minus the null effect). Departures from (1) indicate statistical interaction on an additive risk ratio scale. Lack of departure from (1), however, implies the existence of statistical interaction on a **multiplicative** risk scale [8]. One can express multiplicativity of the risk ratios as

$$\frac{R_{ge}}{R_{\overline{ge}}} = \frac{R_{g\overline{e}}}{R_{\overline{ge}}} \times \frac{R_{\overline{g}e}}{R_{\overline{ge}}}. \quad (2)$$

That is to say, the risk ratio due to the combined action of the genetic and environmental effects is equal to the product of the risk ratios for the genetic effect and for the environmental effect. If the genetic and environmental exposures of interest are associated with disease (i.e. $R_{g\overline{e}}/R_{\overline{ge}} > 1$ and $R_{\overline{g}e}/R_{\overline{ge}} > 1$), then (1) and (2) can never both be true. Specifically, if (1) is true, then the only way (2) can also be true is if $R_{g\overline{e}}/R_{\overline{ge}} = 1$ or $R_{\overline{g}e}/R_{\overline{ge}} = 1$. Clearly, the opposite is also true (i.e. if both g and e have effects and (2) is true, then (1) cannot be true). Hence, the presence or absence of gene–environment interaction depends on the scale used to measure effects, and statistical interaction on the risk scale implies no interaction on the log risk scale and vice versa [8].

One can use stratified analysis (*see* **Stratification**) to assess gene–environment interaction. Cross-classifying the genetic and environmental factors by disease status and applying statistical techniques such as Woolf's test of homogeneity [11] can provide estimates and tests for interaction. Evaluating multiple potential gene–environment interactions simply requires further stratification, but this can lead to problems in sparse data. Instead, one can use a regression approach to assessing interaction. This entails entering into a **regression** model terms for the genetic and environmental factors, plus a product term for the interaction. For example, suppose that in a **case–control study** one collects genetic data x_g, environmental data x_e, and data on disease status. Then one can estimate the coefficients corresponding to the effects of these exposures on disease by fitting a **logistic regression** model,

$$\Pr(\text{disease} \mid x_g, x_e) = \frac{\exp(\beta_0 + \beta_g x_g + \beta_e x_e + \beta_{g\cdot e} x_g x_e)}{1 + \exp(\beta_0 + \beta_g x_g + \beta_e x_e + \beta_{g\cdot e} x_g x_e)}, \quad (3)$$

to the data. Here, β_0 is an intercept term, β_g is the main effect due to genes, and β_e is the main effect of environment. The coefficient $\beta_{g\cdot e}$ of the product $x_g \times x_e$ estimates the gene–environment interaction on the logit scale. One can evaluate this interaction using the likelihood ratio test of a logistic model without the $x_g \times x_e$ product vs. (1) [1]. When $\beta_{g\cdot e} \neq 0$, a departure from odds-ratio multiplicativity (i.e. interaction on the logit scale) exists. If no interaction exists, this model implies that the odds ratio for each factor (genes or environment) is constant across levels of the other factor.

One can evaluate the numerous potential gene–environment (or gene–gene and environment–environment) interactions expected in multifactorial diseases by incorporating additional product terms (including triple products and more complex combinations) into a regression model. The phrase "*n*-order interactions" refers to interactions where the '*n*' is one less than the number of factors involved. However, interpretation of coefficients from models with many interaction terms can be quite complicated. Finally, one should be aware that detecting gene–environment interactions can require substantially larger sample sizes than are necessary for detecting genetic or environmental effects alone [2, 4], and that gene–environment interactions can be **confounded** by **dose–response** relations [9].

References

[1] Breslow, N.E. & Day, N.E. (1982). *Statistical Methods in Cancer Research: The Analysis of Case-Control Studies*, IARC Scientific Publications, No. 32. Oxford University Press, Oxford.

[2] Greenland, S. (1983). Tests for interaction in epidemiologic studies: a review and a study of power, *Statistics in Medicine* **2**, 243–251.

[3] Haldane, J.B.S. (1946). The interaction of nature and nurture, *Annals of Eugenics* **13**, 197–205.

[4] Hwang, S.-J., Beaty, T.H., Liang, K.Y., Coresh, J. & Khoury, M.J. (1994). Minimum sample size estimation to detect gene-environment interaction in case-control designs, *American Journal of Epidemiology* **140**, 1029–1037.

[5] Khoury, M.J., Beaty, T.H. & Cohen, B.H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press, New York, pp. 13, 126.

[6] Miettinen, O.S. (1982). Causal and preventive interdependence: elementary principles, *Scandinavian Journal of Work, Environment, and Health* **8**, 159–168.

[7] Ottman, R. (1995). Gene-environment interaction and public health, *American Journal of Human Genetics* **56**, 821–823.

[8] Rothman, K.J. & Greenland, S. (1997). *Modern Epidemiology*, 2nd Ed. Lippincott-Raven, Philadelphia.

[9] Thomas, D.C. (1981). *Are Dose-Response, Synergy, and Latency Confounded?* American Statistical Association, Alexandria.

[10] Tourian, A. & Sidbury, J.B. (1983). Phenylketonuria and hypophenylalaninemia, in *The Metabolic Basis of Inherited Disease*, 5th Ed., J.B. Stanbury, J.B. Wyngaarden, D.S. Fredrickson, J.L. Goldstein & M.S. Brown, eds. McGraw-Hill, New York, pp. 270–286.

[11] Woolf, B. (1954). On estimating the relation between blood group and disease, *Annals of Human Genetics* **19**, 251–253.

(*See also* **Disease-marker Association**; **Interaction Model**)

JOHN S. WITTE

General Linear Model

The term "general linear model" refers to a specific model formulation which relates a **response variable**, Y, to a set of **explanatory variables**, $X_1, \ldots, X_k$. The basic relation is written as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e,$$

where y is the observed value of Y corresponding to an observed set of values for the explanatory variables, $x_1, \ldots, x_k$, and the βs are regression coefficients, which are usually to be estimated. It is further assumed that

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

and therefore, that $\mathrm{E}(e) = 0$. Note that the linearity of the model relates to the occurrence of the βs and that the x values may represent nonlinear functions of other variables, polynomial terms and standard **transformations**, such as logarithms, being commonly used.

The general linear model is usually written in vector notation as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e},$$

where $\mathbf{y}$ represents a column vector of n Y values, $\mathbf{X}$ is a $n \times K$ matrix with rows corresponding to sets of X values, β is a column vector of regression coefficients, and $\mathbf{e}$ is a column vector of values usually called residual or error terms. The term "design matrix" is often used for $\mathbf{X}$. The variance–**covariance matrix** of $\mathbf{e}$, and therefore of $\mathbf{y}$, is nonnegative definite, and is usually denoted by $\mathbf{V} = \mathrm{E}[\mathbf{y} - \mathrm{E}(\mathbf{y})][\mathbf{y} - \mathrm{E}(\mathbf{y})]'$. A common assumption is that all e values have the same variance, σ^2, and that the covariance of all pairs of e values is zero. In this case, $\mathbf{V} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

Estimation of β does not require further distributional assumptions (*see* **Least Squares**) but it is customary to assume that the es are normally distributed to allow **hypothesis testing** and **interval estimation**. Some authors include the assumption of normal distributions as part of the general linear model formulation.

Many standard statistical models fall into the class of general linear models. These include the models used in **linear regression (simple), multiple linear regression, analysis of variance**, and **analysis of covariance**. See the articles on these topics for a description of inference procedures for the general linear model.

The general linear model is a special case of a **generalized linear model**, a term used to refer to a regression model that relates a function of the mean of a response variable to a linear function of explanatory variables.

VERN T. FAREWELL

General Practice

General or family practice – more generically, primary care – is defined as "the provision of integrated, accessible health care services by clinicians who are accountable for addressing a large majority of personal health care needs, developing a sustained partnership with patients and practicing in the context of a family and community" [28].

Strictly speaking, *general or family practice* is the medical element of a much wider set of health care services delivered in community settings. The generic term for such services is *primary care*, which is characterized by the fact that patients usually attend for care on their own initiative. General practice is therefore distinguished from secondary (hospital) and tertiary (specialist center) care in that it is immediately available to the population, without the need for referral from other agencies. Although such referrals can and do take place, in the UK about a half of all consultations with doctors in general practice are patient-initiated [34].

It is of course true that the precise nature of effective access to such services (and in particular the funding of it (*see* **Health Care Financing**)) varies from one health care system to another. There are nevertheless a number of common features which are pertinent to the biostatistical issues which arise in the specialty. The most important of these is that the medical conditions which are seen most commonly by family doctors are, quite literally, those that occur commonly in the community (*see* **Community Medicine**). Examples include acute upper respiratory tract infection, low back pain, accidents, and problems associated with the ageing process such as osteoarthritis. In general practice, the old adage of "when you hear the sound of hooves, think of horses not zebras" is particularly apt. The reverse of this coin is that conditions which to specialists appear quite often, such as cancer of the lung or breast and even acute myocardial infarction, are relatively rare in general practice.

The common presenting conditions also include those of a chronic, multifactorial nature. The psychological and social aspects of general practice thus require a multidisciplinary approach to care. They have also led to an emphasis on both organizational issues and on care of the family, rather than just purely medical treatment by doctors for individual patients. Concomitantly, research in general practice is necessarily multidisciplinary, incorporating qualitative as well as quantitative methodologies.

Biostatistical Issues in General Practice

Clearly, many of the biostatistical issues which arise in general practice are the same as in many if not all such areas of application. Even some of these common issues take on a particular flavor, though. For example, consider the importance of estimating the magnitude of differences between groups of subjects using **confidence intervals** rather than just relying on statistical significance (*see* **Hypothesis Testing**) [46]. With the large numbers of individuals often available for community-based studies, confidence intervals are necessary to portray possibly excessive precision (for example, where effects of negligible clinical importance are nevertheless statistically significant; (*see* **Clinical Significance Versus Statistical Significance**) as well as to highlight inadequate **power** when sample sizes are too small. Another example is that the wide variety of sources of (systematic and random) differences between subjects means that the advantage in comparative studies of random allocation (*see* **Randomization**) over nonrandom allocation followed by statistical adjustment for **confounding** effects, is arguably particularly marked in a community setting.

The methodologies discussed in detail here, though, will be a selection of those which have a specific relevance to the specialty, although few if any of them are solely applicable to general practice. First, there is the issue of ascertaining the basic pattern of demand and symptomatology in general practice, and variation in referral to secondary care. Next there are methodological aspects of **clinical epidemiology** – such as **screening** and **diagnostic tests** – which have particular relevance. Lastly, issues in the design and analysis of randomized controlled trials (*see* **Clinical Trials, Overview**) in general practice will be discussed.

Observational Studies

Routine Statistics

The key difficulties are in ascertaining the correct numerators and **denominators** for assessing patterns of demand and levels of morbidity in general

practice [32, 34, 35]. A particular problem has been obtaining valid and reliable estimates of the population at risk. For example, Morrell et al. [34] applied corrections to their observed denominators by using simple **regression** techniques to adjust for changes in registrations, including those who were at risk but not registered with the study practice. The ongoing process of computerization of registers, both practice- and community-based, ameliorates this problem but does not obviate the need for such statistical considerations.

The difficulties are compounded in developing countries, where two-fifths of the population have been estimated to be outside the health care system [10]. In this context, methods based on **cluster sampling** for health **surveys** have been extensively developed and widely used [3, 10, 31].

More recent developments in the use of routine statistics in general practice have included investigations of referrals to (usually expensive) secondary services. These studies have often been based on the assumption of the **Poisson distribution**, both in the context of statistical models for observed data or computer **simulation** techniques [2, 15, 33].

Cluster Sampling

This is a special case of **multistage sampling** [1, 11], where *all* the second stage units (for example, individuals) are selected within the first stage units sampled (such as households). The biostatistical issues relate to the two concepts of *accuracy* (systematic **bias**) and *precision* (sampling error). **Unbiased** estimates are as usual obtained by *equal probability of selection methods* (epsem), which for multistage sampling depends on whether fixed numbers or fixed proportions of second stage units are sampled. For the latter (which is the case in cluster sampling where the sampling fraction is 100% for the second stage units), **simple random sampling** or **stratified sampling** of first stage units is required [1, 37].

The statistical theory of sampling errors in such designs is well established [11, 30]. Whether for sample size planning or for estimating precision and confidence intervals, at issue is the calculation of the **design effect** (or *inflation factor* or *Kish design effect*) to allow for the inefficiency of this design compared with single stage random sampling. There are two general approaches: one requiring an estimate of the **variance** between the clusters [14, 31]; the other requiring an estimate of the *intracluster* (more generally, *intraclass*) **correlation** *coefficient* [3, 22]; for more details, see the section below on Cluster Randomization. In either case, although there is no practical constraint on an estimate of the design effect, in general the consequence is to increase standard errors.

Clinical Epidemiology

Screening and Diagnostic Tests

An example of a methodologic aspect of clinical epidemiology which is particularly relevant to general practice is the effect of the **prevalence** of the condition on the standard performance statistics for screening and diagnostic tests [25, 42]. In particular, the *positive* **predictive value** (probability of disease among subjects with a positive test result) falls as the underlying prevalence decreases, even if the **sensitivity** (ability of the test to detect the disease) and **specificity** (ability of the test to exclude the disease) remain unaltered. For example, a test with 70% sensitivity and 90% specificity for prostatic cancer gave a positive predictive value of 93% for a group of patients with a particular clinical symptom. For patients in a general practice population, however, the same test yielded a positive predictive value below 0.5% [25]. This fundamental point is not altered by employing more sophisticated statistical measures such as **likelihood ratios** for positive and negative test results (used in conjunction with **Bayes' Theorem** to update pre-test to post-test probabilities of disease) and **receiver operating characteristic (ROC) curves** (plots of sensitivity against 1-specificity).

In summary, diagnostic tests which are highly useful in a secondary care environment may well be of no practical value in general practice, where the test conditions are likely to be relatively rare. This is one reason why screening in a general population is often difficult to justify [38].

Agreement

The issue of *interobserver* **agreement** often arises in research in general practice, particularly in the context of **health status** measurement [47]. For continuous variables, simple correlation coefficients are inappropriate as measures of agreement; rather, the

analysis should be based on paired differences [5]. Tests and confidence intervals, for example on (paired) mean differences, assess the "bias" of one assessment relative to the other. Assuming a **normal distribution** for the differences, 1.96 **standard deviations** (of the differences) either side of the mean difference forms a 95% reference range for the paired differences, or *limits of agreement* [5]. A plot of each paired difference against the mean of the respective two measurements portrays the agreement graphically, and may suggest that a logarithmic **transformation** would be worthwhile.

For noncontinuous variables, the most widely used measure of agreement is the **kappa statistic** κ [12]. Specifically, kappa measures chance-corrected agreement, with the option of introducing weights for different levels of disagreement where there are more than two ordered categories [1, 13, 24]. If quadratic weights are employed (that is, weights proportional to the square of the discrepancy on the original scale), then, apart from terms in 1/(sample size), *weighted kappa* is equivalent to the intraclass correlation coefficient [1, 47].

For a **binary** variable, unweighted kappa is equivalent to the intraclass correlation coefficient using scores of 0 and 1 [22, 47]. **McNemar's test** for paired proportions may be used to ascertain whether there is evidence that the discrepancy is in a particular direction, but this is difficult to interpret without also bearing in mind the extent of disagreement overall.

Randomized Controlled Trials

Design

The key biostatistical features of designing randomized controlled trials (RCTs) in general practice are that they should be essentially *pragmatic* rather than *explanatory* [43], and that the unit of randomization is often not the individual but a group of individuals (*see* **Group-randomization Designs**) [21]. In addition, by the nature of the context, selection criteria for entry into such trials should be inclusive rather than exclusive (*see* **Eligibility and Exclusion Criteria**). In more general terms, all aspects of their design should be realistic, recognizing and allowing for the limitations in respect of, for instance, **blinding** and contamination.

Pragmatic RCTs are where the interventions are designed to be as close as possible to the situations in which they would be applied [43]. In this way, the *efficacy* of an intervention (performance under ideal circumstances) which RCTs evaluate is brought as close as possible to *effectiveness* (performance under everyday circumstances) (*see* **Pharmacoepidemiology, Adverse and Beneficial Effects**) [4, 38]. The primary data analyses should be on an **intention-to-treat** basis, comparing the groups as they were randomized. Comparing subjects in terms of intervention actually received (the explanatory approach) will in general be biased since it does not correspond to the random allocation; nevertheless, secondary analyses of this kind will often be worthwhile. Even in the primary analysis the impact of missing outcome data should not be ignored [27], and different assumptions regarding missing values may well need to be the subject of sensitive analyses [41].

The second key feature is allocation of groups rather than individual subjects – that is, *cluster randomization*. As indicated in the section above on Cluster Sampling, the effect of this is to reduce the efficiency of the trial; all other things being equal, then, the sample sizes of such trials will need to be larger. **Sample size determination** for cluster-randomized trials has received considerable attention in epidemiological and statistical journals [14, 16, 22, 45], although it has until recently been relatively neglected in the applied primary care literature [7–9]. In such applications, the cluster is usually the general practice, though it may be individual practitioners. The design is considered as an option because of the likelihood of contamination between the interventions if they are conducted within the same practice.

There has been debate about the importance of removing the effect of contamination in the study design [49]. Cluster randomized trials are, however, likely to remain a reasonable option and the UK Medical Research Council is currently developing guidelines regarding their design, conduct and presentation. Certainly, leaving adjustments for contamination to the analysis has dangers in terms of the ability to estimate and control for it adequately; moreover, any such adjustment for extraneous variables in the primary analysis of the trial runs the risk of interfering with the central principles of an intention-to-treat analysis [27].

As noted above, the two general approaches to sample size planning require prior estimation of either the *intercluster variation* [14, 31] or the *intracluster correlation coefficient* [3, 22]. While the latter is

at least potentially more generalizable, particularly for small clusters, these methods are related [22]. In practice, though, the choice will often depend on the type of prior information available. Basically, each obtains an *inflation factor* as the inverse of the *relative efficiency* of cluster randomization compared with individual randomization. The sample size requirement calculated using standard methods for individual randomization is then multiplied by this inflation factor [16].

The first approach, described by Cornfield [14] for comparing a binary outcome between two interventions, obtains the inflation factor as $\mathrm{IF} = (c\sigma^2)/[p(1-p)]$, where c is the mean cluster size, and p and σ^2 are the mean and (population) variance, respectively, of the cluster-specific proportions. The second, described by Donner et al. [22] for continuous and binary outcome variables, calculates $\mathrm{IF} = 1 + (c-1)\rho$, where c is as before and ρ is the intracluster correlation coefficient [16, 22, 30]. For binary variables, the latter is equivalent to the kappa statistic, κ, described above [12, 24].

From the second approach it is clear that if $\kappa = 0$ (the case of "mavericks" when individuals within clusters are no more similar to one another than individuals randomly selected from the population), then the inflation factor is unity – the cluster design introduces no additional inefficiency over individual randomization. At the other extreme, $\kappa = 1$ (the case of "clones" within clusters) and $\mathrm{IF} = c$; the sample size is then effectively the number of clusters.

In the context of binary outcomes in general practice, the values of κ observed are quite small – for instance, of the order of 0.05 for controlled hypertension [23]. Given the proportional relationship of the **design effect** with cluster size, however, apparently small clustering effects in terms of κ can translate into large design effects when the number of individuals per practice is large. To offset this inefficiency, such designs should almost always be accompanied by features that improve efficiency, such as **stratification** and applications of **factorial** and **Latin square designs** [1]. In fact, the efficiency benefits of stratification are in general even more marked for cluster randomization than for individual randomization [30]. In addition, stratified randomization to ensure balance in respect of key variables is likely to be essential in the (common) situation where the number of clusters is relatively small.

Analysis

It is essential that the statistical analysis of data from RCTs does at least at some point take into account any complex design features such as stratification and cluster randomization [14, 16, 21, 50]. While a simple (unadjusted) analysis may be helpful initially, the effects of adjustments for such characteristics should at some point be ascertained. For instance, having stratified for a variable in the design, only conditioning on this factor in the analysis will achieve both unbiased estimates and maximize efficiency [44].

This can be achieved by the use of either corrections to basic analytic procedures, or **generalized linear models**, with for example fixed effects terms for any stratification variables [17–20]. For variables representing clusters which have been randomized – practices, for example – one approach is to represent these by **random effects** terms in an appropriate model [18]. Whatever the magnitude of the effect of these corrections, the key trial comparisons derived from relatively complex models can be presented (in terms of statistical significance and confidence intervals) just as simply as the corresponding results from unweighted analyses such as t tests (*see* **Student's *t* Statistics**).

The most general approach to all the above analytic issues, though, is by **multilevel modeling** [18, 26, 40]. Such models are, of course, not restricted to studies (experimental or observational) where the clustering is explicitly taken into account in the design. As mentioned earlier, general practice is implicitly characterized by a variety of sources of variation at a number of levels.

Anticipated Developments and Unresolved Problems

The unresolved problems generally take one of two forms. The first is the ascertainment of the magnitude of the benefit which will accrue in general practice research by the application of the more complex of the methods described above. To achieve this, both methodologic work and applied research is needed – for instance, only rarely in the literature are values of intracluster correlation coefficients quoted [123], although this situation is improving [50]. In addition, research is needed into the circumstances in which either Cornfeld's or Donner's approaches to sample size determination is to be preferred, and also into circumstances in which one or both corrections appear

not to be applicable – for example, when an estimated inflation factor is less than unity. Further investigation is also required into whether there are any patterns in the number and nature of the major sources of variation in suitable multilevel models in general practice.

The second general challenge is for research into the relative values of different study designs in evaluating interventions in general practice [4]. This includes the value of large amounts of routinely collected data, particularly as computer technology in general practice becomes more widespread and better developed for this purpose. Clearly, though, the issues of validity and reliability remain even with computerization. Moreover, with observational data (particularly in a context such as general practice with many sources of potentially major variation), comparability in terms of **case mix** will continue to be difficult to establish.

Overall, randomized controlled trials in general practice have met many practical problems – for example, recruitment of practitioners and subjects, objections to **randomization**, and poor subsequent **compliance** [29, 48]. This has led many to question their central role as the **gold standard** for evaluating interventions [39], calling for a role for observational studies [4] and the *patient preference trial*, where patients who express a strong preference are given their choice and only the remainder are randomized [6, 29]. A further possibility is the *comprehensive cohort*, where randomized subjects are nested within a wider cohort of individuals who for one reason or another were not included in the trial [36]. To date, all these alternatives remain either relatively little developed, or controversial, or both. Whilst further methodologic and applied research in general practice will no doubt continue on these and similar approaches, the central role of the pragmatic randomized controlled trial will undoubtedly continue in general practice research.

Acknowledgments

I am grateful to all my colleagues who gave me helpful comments and suggestions, in particular Debbie Sharp, Ian Harvey, Max Bachmann, and Chris Watkins.

References

[1] Armitage, P. & Berry, G. (1994). *Statistical Methods in Medical Research*, 3rd Ed. Blackwell Science, Oxford.

[2] Bachmann, M.O. & Bevan, G. (1996). Determining the size of a total purchasing site to manage the financial risks of rare costly referrals: computer simulation model, *British Medical Journal* **313**, 1054–1057.

[3] Bennett, S., Woods, T., Liyanage, W.M. & Smith, D.L. (1991). A simplified general method for cluster-sample surveys of health in developing countries, *World Health Statistics Quarterly* **44**, 98–106.

[4] Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care, *British Medical Journal* **312**, 1215–1218.

[5] Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **i**, 307–310.

[6] Brewin, C.R. & Bradley, C. (1989). Patient preferences and randomised clinical trials, *British Medical Journal* **299**, 313–315.

[7] Butler, C. & Bachmann, M. (1996). Design and analysis of studies evaluating smoking cessation interventions where effects vary between practices or practitioners, *Family Practice* **13**, 402–407.

[8] Campbell, M.J. (2000). Cluster randomized trials in general family practice. *Statistical Methods in Medical Research* **9**, 81–94.

[9] Campbell, M.K., Mollison, J., Steen, N., Grimshaw, J.M. & Eccles, M. (2000). Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice* **17**, 192–196.

[10] Carlson, B.A. (1985). The potential of national household survey programmes for monitoring and evaluating primary health care in developing countries, *World Health Statistics Quarterly* **38**, 38–64.

[11] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[12] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[13] Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213–220.

[14] Cornfield, J. (1978). Randomization by group: a formal analysis, *American Journal of Epidemiology* **108**, 100–102.

[15] Crump, B.J., Cubbon, J.E., Drummond, M.F., Hawkes, R.A. & Marchment, M.D. (1991). Fundholding in general practice and financial risk, *British Medical Journal* **302**, 1582–1584.

[16] Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials – a review, *Statistics in Medicine* **3**, 199–214.

[17] Donner, A. (1989). Statistical methods in ophthalmology: an adjusted chi-square approach, *Biometrics* **45**, 605–611.

[18] Donner, A. (1998). Some aspects of the design and analysis of cluster randomisation trials. *Applied Statistics* **47**, 95–113.

[19] Donner, A. & Klar, N. (1993). Confidence interval construction for effect measures arising from cluster

randomization trials, *Journal of Clinical Epidemiology* **46**, 123–131.

[20] Donner, A. & Klar, N. (1994). Methods for comparing event rates in intervention studies when the unit of allocation is a cluster, *American Journal of Epidemiology* **140**, 279–289.

[21] Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomisation Trials in Health Research*. Arnold, London.

[22] Donner, A., Birkett, N. & Buck, C. (1981). Randomization by cluster: sample size requirements and analysis, *American Journal of Epidemiology* **114**, 906–914.

[23] Fahey, T.P. & Peters, T.J. (1996). What constitutes controlled hypertension? Patient based comparison of hypertension guidelines, *British Medical Journal* **313**, 93–96.

[24] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[25] Fletcher, R.H., Fletcher, S.W. & Wagner, E.H. (1996). *Clinical Epidemiology: The Essentials*, 3rd Ed. Williams & Wilkins, Baltimore.

[26] Goldstein, H. (1995). *Multilevel Statistical Models*. Edward Arnold, London.

[27] Hollis, S. & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal* **319**, 670–674.

[28] Institute of Medicine (1994). *Defining Primary Care: An Interim Report*. M. Donaldson, K. Yordy & N. Vanselow, eds. National Academy Press, Washington.

[29] King, M., Broster, G., Lloyd, M. & Horder, J. (1994). Controlled trials in the evaluation of counselling in general practice, *British Journal of General Practice* **44**, 229–232.

[30] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[31] Lemeshow, S. & Robinson, D. (1985). Surveys to measure programme coverage and impact: a review of the methodology used by the expanded programme on immunization, *World Health Statistics Quarterly* **38**, 65–75.

[32] McCormick, A., Fleming, D. & Charlton, J. (1995). *Morbidity Statistics from General Practice: Fourth National Study 1991–1992*. Office of Population Censuses and Surveys, Series MB5, No. 3. HMSO, London.

[33] Moore, A.T. & Roland, M.O. (1989). How much variation in referral rates among general practitioners is due to chance?, *British Medical Journal* **298**, 500–502.

[34] Morrell, D.C., Gage, H.G. & Robinson, N.A. (1970). Patterns of demand in general practice, *Journal of the Royal College of General Practitioners* **19**, 331–342.

[35] Morrell, D.C., Gage, H.G. & Robinson, N.A. (1971). Symptoms in general practice, *Journal of the Royal College of General Practitioners* **21**, 32–43.

[36] Olschewski, M., Schumacher, M. & Davis, K. (1992). Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design, *Controlled Clinical Trials* **13**, 226–239.

[37] Peters, T.J. & Eachus, J.I. (1995). Achieving equal probability of selection under various random sampling strategies, *Paediatric and Perinatal Epidemiology* **9**, 219–224.

[38] Peters, T.J., Wildschut, H.I.J. & Weiner, C.P. (1996). Epidemiologic considerations in screening, in *When to Screen in Obstetrics and Gynecology*, H.I.J. Wildschut, C.P. Weiner & T.J. Peters, eds. Saunders, London, pp. 1–12.

[39] Pringle, M. & Churchill, R. (1995). Randomised controlled trials in general practice: gold standard or fool's gold?, *British Medical Journal* **311**, 1382–1383.

[40] Rice, N. & Leyland, A. (1996). Multilevel models: applications to health data, *Journal of Health Services Research and Policy* **1**, 154–164.

[41] Richards, S.H., Bankhead, C., Peters, T.J., Austoker, J., Hobbs, F.D.R., Brown, J., Tydeman, C., Roberts, L., Formby, J., Redman, V., Wilson, S. & Sharp, D.J. (2001). A cluster randomised controlled trial comparing the effectiveness of two interventions in primary care aimed at improving attendance for breast screening. *Journal of Medical Screening* **8**, 91–98.

[42] Sackett, D.L., Haynes, R.B., Guyatt, G.H. & Tugwell, P. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*, 2nd Ed. Little, Brown & Company, Boston.

[43] Schwartz, D. & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials, *Journal of Chronic Diseases* **20**, 637–648.

[44] Senn, S. (1997). *Statistical Issues in Drug Development*. Wiley, Chichester.

[45] Shipley, M.J., Smith, P.G. & Dramaix, M. (1989). Calculation of power for matched pair studies when randomization is by group, *International Journal of Epidemiology* **18**, 457–461.

[46] Sterne, J.A.C. & Davey Smith, G. (2001). Sifting the evidence- what's wrong with significance tests? *British Medical Journal* **322**, 226–231.

[47] Streiner, D.L. & Norman, G.R. (1989). *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford University Press, Oxford.

[48] Tognoni, G., Alli, C., Avanzini, F., Bettelli, G., Colombo, F., Corso, R., Marchioli, R. & Zussino, A. (1991). Randomised clinical trials in general practice: lessons from a failure, *British Medical Journal* **303**, 969–971.

[49] Torgerson, D.J. (2001). Contamination in trials: is cluster randomisation the answer? *British Medical Journal* **322**, 355–357.

[50] Ukoumunne, O.C., Gulliford, M.C., Chinn, S., Sterne, J.A.C. & Burney, P.G.J. (1999). Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment* **3**(5).

T.J. PETERS

Generalized Additive Model

In the statistical analysis of **clinical trials** and **observational studies**, the identification and adjustment for **prognostic factors** is an important component. Valid comparisons of different treatments requires the appropriate adjustment for relevant prognostic factors. The failure to consider important prognostic variables, particularly in observational studies, can lead to errors in estimating treatment differences. In addition, incorrect modeling of prognostic factors can result in the failure to identify nonlinear trends or threshold effects on survival.

This article describes flexible statistical methods that may be used to identify and characterize the effect of potential prognostic factors on an outcome variable. These methods are called "generalized additive models", and extend the traditional **general linear model**. They can be applied in any setting in which a linear or **generalized linear model** is typically used. These settings include standard continuous response **regression**, **categorical** or **ordered categorical** response data, count data, **survival** data and **time series**.

One of the most commonly used statistical models in medical research is the **logistic regression** model for **binary data**. We use it here as a specific illustration of a generalized additive mode. Logistic regression (and many other techniques) model the effects of prognostic factors x_j in terms of a linear predictor of the form $\sum x_j\beta_j$, where the β_j are parameters. The generalized additive model replaces $\sum x_j\beta_j$ with $\sum f_j(x_j)$, where f_j is a unspecified ("nonparametric") function. This function is estimated in a flexible manner using a scatterplot smoother (*see* **Graphical Displays**). The estimated function $\hat{f}_j(x_j)$ can reveal possible nonlinearities in the effect of x_j.

We first give some background on the methodology, and then discuss the details of the logistic regression model and its generalization. Some related developments are discussed in the last section.

Smoothing Methods and Generalized Additive Models

The building block of the generalized additive model algorithm is the scatterplot smoother. We will first describe scatterplot smoothing in a simple setting, and then indicate how it is used in generalized additive modeling.

Suppose that we have a scatterplot of points (x_i, y_i) such as that shown in Figure 1. Here y is a **response** or outcome variable, and x is a prognostic factor. We wish to fit a smooth curve $f(x)$ that summarizes the dependence of y on x. If we were to find the curve that simply minimizes $\sum[y_i - f(x_i)]^2$, the result would be an interpolating curve that would not be smooth at all.

The cubic **spline** smoother imposes smoothness on $f(x)$. We seek the function $f(x)$ that minimizes

$$\sum[y_i - f(x_i)]^2 + \lambda \int f''(x)^2 \, \mathrm{d}x. \qquad (1)$$

Notice that $\int f''(x)^2$ measures the "wiggliness" of the function f: linear fs have $\int f''(x)^2 = 0$, while nonlinear fs produce values greater than zero. λ is a nonnegative smoothing parameter that must be chosen by the data analyst. It governs the tradeoff between the **goodness of fit** to the data (as measured by $\sum[y_i - f(x_i)]^2$) and wiggliness of the function. Larger values of λ force f to be smoother.

For any value of λ, the solution to (1) is a cubic spline; that is, a piecewise cubic polynomial with pieces joined at the unique observed values of x in the dataset. Fast and stable numerical procedures are available for computation of the fitted curve. The right panel of Figure 1 shows a cubic spline fit to the data.

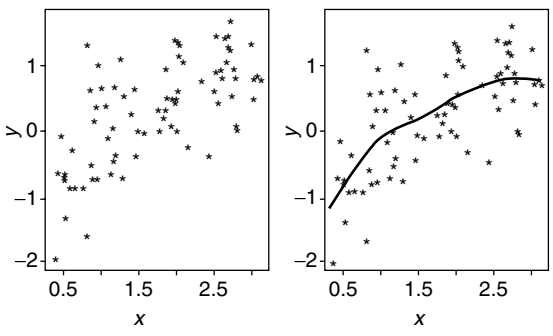

Figure 1 The left panel shows a fictitious scatterplot of an outcome measure y plotted against a prognostic factor x. In the right panel, a scatterplot smoother has been added to describe the trend of y on x

What value of λ did we use in Figure 1? In fact, it is not convenient to express the desired smoothness of f in terms of λ, as the meaning of λ depends on the units of the prognostic factor x. Instead, it is possible to define an "effective number of parameters" or "**degrees of freedom**" of a cubic spline smoother, and then use a numerical search to determine the value of λ to yield this number. In Figure 1 we chose the effective number of parameters to be 5. Roughly speaking, this means that the complexity of the curve is about the same as a **polynomial regression** of degree 4. However, the cubic spline smoother "spreads out" its parameters in a more even manner, and hence is much more flexible than a polynomial regression. Note that the degrees of freedom of a smoother need not be an integer.

The above discussion tells how to fit a curve to a single prognostic factor. With multiple prognostic factors, if x_{ij} denotes the value of the jth prognostic factor for the ith observation, we fit the additive model

$$\hat{y}_i \approx \sum_j f_j(x_{ij}). \tag{2}$$

A criterion such as (1) can be specified for this problem, and a simple iterative procedure exists for estimating the f_js. We apply a cubic spline smoother to the outcome $y_i - \sum_{j \neq k} \hat{f}_j(x_{ij})$ as a function of x_{ik}, for each prognostic factor in turn. The process continues until the estimates $\hat{f}_k$ stabilize. This procedure is known as "backfitting", and the resulting fit is analogous to a multiple regression for linear models.

When generalized additive models are fit to binary response data (and in many other settings), the appropriate error criterion is a penalized log likelihood or a penalized log partial-likelihood (*see* **Penalized Maximum Likelihood**). To maximize it, the backfitting procedure is used in conjunction with a **maximum likelihood** or maximum **partial likelihood** algorithm. The usual Newton–Raphson routine (*see* **Optimization and Nonlinear Equations**) for maximizing log likelihoods in these models can be cast in an IRLS (iteratively reweighted least squares) form (*see* **Generalized Linear Model**). This involves a repeated weighted linear regression of a constructed response variable on the covariates: each regression yields a new value of the parameter estimates which give a new constructed variable, and the process is iterated. In the generalized additive model, the weighted linear regression is simply replaced by a weighted backfitting algorithm. Details can be found in [7, Chapter 6].

The Generalized Additive Logistic Model

Generalized additive models can be used in virtually any setting in which linear models are used. The basic idea is to replace $\sum x_{ij}\beta_j$, the linear component of the model with an additive component $\sum f_j(x_{ij})$.

In the logistic regression model the outcome y_i is 0 or 1, with 1 indicating an event (such as death or relapse of a disease) and 0 indicating no event. We wish to model $p(y_i|x_{i1}, x_{i2}, \ldots, x_{ip})$, the probability of an event given prognostic factors $x_{i1}, x_{i2}, \ldots x_{ip}$. The linear logistic model assumes that the log odds are linear:

$$\log \frac{p(y_i|x_{i1}, \ldots, x_{ip})}{1 - p(y_i|x_{i1}, \ldots, x_{ip})} = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p. \tag{3}$$

The generalized additive logistic model assumes instead that

$$\log \frac{p(y_i|x_{i1}, \ldots, x_{ip})}{1 - p(y_i|x_{i1}, \ldots, x_{ip})} = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}). \tag{4}$$

The functions $f_1, f_2, \ldots, f_p$ are estimated by an **algorithm** like the one described earlier.

To illustrate this, we describe a study on the survival of children after cardiac surgery for heart defects [13]. The data were collected during the period 1983–1988. A pre-operation warm-blood cardioplegia procedure, thought to improve chances for survival, was introduced in February 1988. This was not used on all of the children after February 1988, only on those for which it was thought appropriate and only by surgeons who chose to use the new procedure. The main question is whether the introduction of the warming procedure improved survival; the importance of risk factors age, weight, and diagnostic category is also of interest.

If the warming procedure was given in a randomized manner, we could simply focus on the post-February 1988 data and compare the survival of those who received the new procedure to those who did not. However, allocation was not random, so we can only

try to assess the effectiveness of the warming procedure as it was applied. For this analysis, we use all of the data (1983–1988). To adjust for changes that might have occurred over the five-year period, we include the data of the operation as a covariate. However, operation date is strongly confounded with the warming operation and thus a general nonparametric fit for date of operation might unduly remove some of the effect attributable to the warming procedure. To avoid this, we allow only a linear effect for operation date. Hence we must assume that any time trend is either a consistently increasing or decreasing trend.

We fit a generalized additive logistic model to the binary response death, with smooth terms for age and weight, a linear term for operation date, a categorical variable for diagnosis, and a binary variable for the warming operation. All the smooth terms are fitted with four degrees of freedom.

The resulting curves for age and weight are shown in Figure 2. As one would expect, the highest risk is for the lighter babies, with a decreasing risk over 3 kg. Somewhat surprisingly, there seems to be a low risk age around 200 days, with higher risk for younger and older children. Note that the numerical algorithm is not able to achieve exactly four degrees of freedom for the age and weight terms, but 3.80 and 3.86 degrees of freedom, respectively.

An analysis of deviance (*see* **Generalized Linear Model**) can be carried out for inference from a generalized additive model, analogous to that done for generalized linear models. The only new twist is estimation of the degrees of freedom or effective number of parameters of the fitted model, which was discussed in the previous section. This analysis shows that the warming procedure is strongly beneficial to survival. There are strong differences in the diagnosis categories, while the estimated effect of operation date is not large.

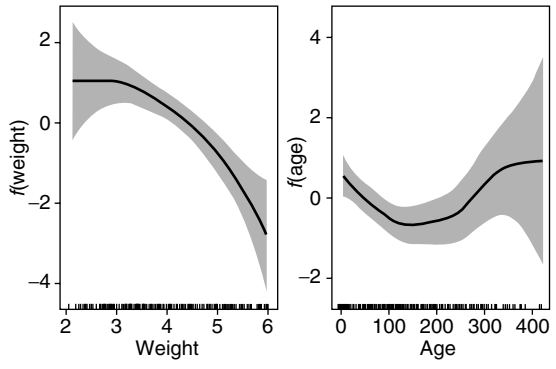

Figure 2 Estimated functions for weight and age for warm cardioplegia data. The shaded region represents twice the pointwise asymptotic standard errors of the estimated curve

Since a logistic regression is additive on the logit scale but not on the probability scale, a plot of the fitted probabilities is often informative. Figure 3 shows the fitted probabilities broken down by age and diagnosis, and is a concise summary of the findings of this study. The beneficial effect of the treatment at the lower weights is evident. As with all nonrandomized studies, the results here should be interpreted with caution. In particular, one must insure that the children were not chosen for the warming operation based on their prognosis. To investigate this, we perform a second analysis in which a **dummy variable** (say, period), corresponding to before vs. after February 1988, is inserted in place of the dummy variable for the warming operation. The purpose of this is to investigate whether the overall treatment strategy improved after February 1988. If this turns out not to be the case, it will imply that warming was used only for patients with a good prognosis, who would have survived anyway. A linear adjustment for operation date is included as before. The results are qualitatively very similar to the first analysis: age and weight are significant, with effects similar to those in Figure 2; diagnosis is significant, while operation date (linear effect) is not. Period is highly significant. Hence there seems to be a significant overall improvement in survival after February 1988. For more details, see [13].

Discussion

The nonlinear modeling procedures described here are useful for two reasons. First, they help to prevent model misspecification, which can lead to incorrect conclusions regarding treatment efficacy. Secondly, they provide information about the relationship between prognostic factors and disease risk that is not revealed by the use of standard modeling techniques. Linearity always remains a special case, and thus simple linear relationships can be easily confirmed with flexible modeling of covariate effects.

The most comprehensive source for generalized additive models is [7], from which the example was

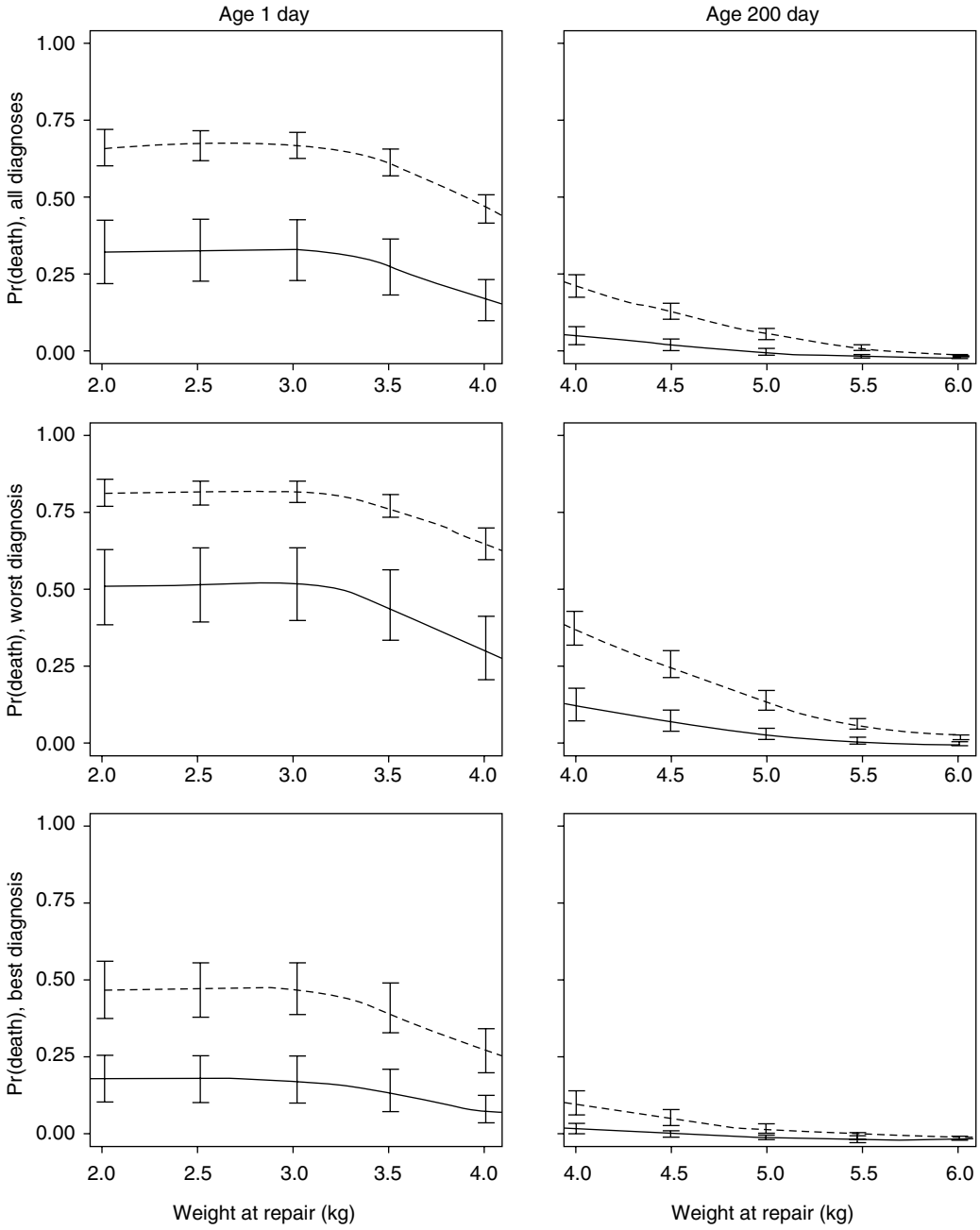

Figure 3 Estimated probabilities for warm cardioplegia data, conditioned on two ages (columns) and three diagnostic classes (rows). The broken line is standard treatment; the solid line is warm cardioplegia. Bars indicate ± 1

taken. A detailed example of the use of generalized additive models in the **proportional hazards** setting is given in [10]. Other medical applications are discussed in [6] and [9]. Penalization and spline models in a variety of settings are discussed in [5], and [12] is a good source for the mathematical background of spline models. See also [3] for an exposition of modern developments in statistics (including generalized additive models), for a nonmathematical audience.

There has been some recent related work in this area. A different method for flexible hazard modeling is described in [11] and a generalization of additive modeling that finds **interactions** among prognostic factors is proposed in [4]. Of particular interest in the **proportional hazards** setting is the *varying coefficient* model [8] (*see* **Semiparametric Regression**), in which the parameter effects can change with other factors such as time. The model has the form

$$h(t|x_{i1}, \ldots, x_{ip}) = h_0(t) \exp \sum_{j=1}^{p} \beta_j(t) x_{ij}. \qquad (5)$$

The parameter functions $\beta_j(t)$ are estimated by scatterplot smoothers in a similar fashion to the methods described earlier. This gives a useful way of modeling departures from the proportional hazards assumption by estimating the way in which the parameters β_j change with time.

Software for fitting generalized additive models is available in the **S/SPLUS** statistical environment [1, 2], in a FORTRAN program called gamfit available at statlib (in general/gamfit at the ftp site `lib.stat.cmu.edu`) and also in the GAIM package for MS-DOS computers, available from the authors.

Acknowledgment

Trevor Hastie was partially supported by grant DMS-9504495 from the National Science Foundation, and grant ROI-CA-72028-01 from the National Institutes of Health. Robert Tibshirani was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Becker, R., Chambers, J. & Wilks, A. (1988). *The New S Language*. Wadsworth International Group, Pacific Grove.

[2] Chambers, J. & Hastie, T. (1991). *Statistical Models in S*. Wadsworth/Brooks Cole, Pacific Grove.

[3] Efron, B. & Tibshirani, R. (1991). Statistical analysis in the computer age, *Science* **253**, 390–395.

[4] Friedman, J. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**, 1–141.

[5] Green, P. & Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman & Hall, London.

[6] Hastie, T. & Herman, A. (1990). An analysis of gestational age, neonatal size and neonatal death using nonparametric logistic regression, *Journal of Clinical Epidemiology* **43**, 1179–1190.

[7] Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[8] Hastie, T. & Tibshirani, R. (1997). Discriminant analysis by Gaussian mixtures, to appear.

[9] Hastie, T., Botha, J. & Schnitzler, C. (1989). Regression with an ordered categorical response, *Statistics in Medicine* **8**, 785–794.

[10] Hastie, T., Sleeper, L. & Tibshirani, R. (1992). Flexible covariate effects in the proportional hazards model, *Breast Cancer Research and Treatment* **22**, 241–250.

[11] Kooperberg, C., Stone, C. & Truong, Y. (1993). Hazard Regression, *Technical Report*. Department of Statistics, University of California, Berkeley.

[12] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

[13] Williams, W., Rebeyka, I., Tibshirani, R., Coles, J., Lightfoot, N., Freedom, R. & Trusler, G. (1990). Warm induction cardioplegia in the infant: a technique to avoid rapid cooling myocardial contracture, *Journal of Thoracic and Cardiovascular Surgery* **100**, 896–901.

(*See also* **Nonparametric Regression**)

TREVOR HASTIE & R. TIBSHIRANI

Generalized Estimating Equations

It is common practice in statistical modeling, as epitomized in the **generalized linear model**, to divide variation in some measure into a systematic part and a random part. The validity of classical **maximum likelihood** inference for such models depends upon the correct choice of model for both parts, including a specific choice of distribution for the random part. Typically, substantive theory guides the construction of the model for the systematic part (*see* **Model, Choice of**). For the random part, although some guidance may be obtained from knowledge of the distributional consequences of a simple and plausible **stochastic process** that might be responsible for this variation, and from experience gained with other comparable data, we rarely have complete confidence in our choice of distribution. This is of especial concern where samples are too small to allow scope for detailed checking of the distributional assumptions, where a **multivariate distribution** must be specified for the random part, and where the available forms of multivariate distribution lack flexibility. This describes well the circumstances that we face with many longitudinal datasets, and where it is often helpful to use "generalized estimating equations" (GEEs) [11]. The GEE approach, while making weaker distributional assumptions than those required for a fully parametric **likelihood**-based model, maintains the properties of **consistency** and asymptotic normality of parameter estimates [8] (*see* **Large-sample Theory**). The description of GEE methods presented here focuses on their use with longitudinal data. However, these methods are also applicable to other forms of multivariate or correlated response data including, for example, studies involving clustered or multistage sampling and the joint analysis of several response features.

Tables 1 and 2 provide simple examples of the kind of longitudinal data at issue. Table 1, from [4], concerns 56 patients from one center (center 1) in a **multicenter** treatment trial for a respiratory disease, with a **binary** self-rated response variable measured at a baseline and on four subsequent occasions. The effects of interest are those for the *time-constant* or *between-subjects* variables for treatment (active

Table 1 Repeated measures data on respiratory disease

ID	1	2	3	4	0	*t*	*s*	Age
1	0	0	0	0	0	0	0	46
2	0	0	0	0	0	0	0	28
3	1	1	1	1	1	1	0	23
4	1	1	1	1	0	0	0	44
5	1	1	1	1	1	0	1	13
6	0	0	0	0	0	1	0	34
7	0	1	0	1	1	0	0	43
8	0	0	0	0	0	1	0	28
9	1	1	1	1	1	1	0	31
10	1	0	1	1	0	0	0	37
11	1	1	1	1	1	1	0	30
12	0	1	1	1	0	1	0	14
13	1	1	0	0	0	0	0	23
14	0	0	0	0	0	0	0	30
15	1	1	1	1	1	0	0	20
16	0	0	0	0	1	1	0	22
17	0	0	0	0	0	0	0	25
18	0	0	1	1	1	1	1	47
19	0	0	0	0	0	0	1	31
20	1	1	0	1	0	1	0	20
21	0	1	0	1	0	1	0	26
22	1	1	1	1	1	1	0	46
23	1	1	1	1	1	1	0	32
24	0	1	0	0	0	1	0	48
25	0	0	0	0	0	0	1	35
26	0	0	0	0	0	1	0	26
27	1	1	0	1	1	0	0	23
28	0	1	1	0	0	0	1	36
29	0	1	1	0	0	0	0	19
30	0	0	0	0	0	1	0	28
31	0	0	0	0	0	0	0	37
32	0	1	1	1	1	1	0	23
33	1	1	1	1	0	1	0	30
34	0	0	1	1	0	0	0	15
35	0	0	0	1	0	1	0	26
36	0	0	0	0	0	0	1	45
37	0	0	1	0	0	1	0	31
38	0	0	0	0	0	1	0	50
39	0	0	0	0	0	0	0	28
40	0	0	0	0	0	0	0	26
41	0	0	0	0	1	0	0	14
42	0	0	1	0	0	1	0	31
43	1	1	1	1	1	0	0	13
44	0	0	0	0	0	0	0	27
45	0	1	0	1	1	0	0	26
46	0	0	0	0	0	0	0	49
47	0	0	0	0	0	0	0	63
48	1	1	1	1	1	1	0	57
49	1	1	1	1	1	0	0	27
50	0	0	1	1	1	1	0	22
51	0	0	1	1	1	1	0	15

Table 1 (*continued*)

ID	1	2	3	4	0	*t*	*s*	Age
52	0	0	0	1	0	0	0	43
53	0	0	0	1	0	1	1	32
54	1	1	1	1	0	1	0	11
55	1	1	1	1	1	0	0	24
56	0	1	1	0	1	1	0	25

Adapted from [4].

vs. placebo), and age of the subject and the *time-varying* or *within-subjects* effect of time. Table 2, from [34], gives data of a very similar structure, but here concerned with a 59 patient treatment trial (0 = placebo, 1 = progabide) for epilepsy, where the response measure is a count reflecting the number of seizures within four successive intervals of 2 weeks. In both datasets a pretreatment baseline measure of the response was also taken. In models fitted later in this article this baseline measure has been included as an additional **covariate**, in the style of **analysis of covariance** (*see* **Baseline Adjustment in Longitudinal Studies**).

The **multivariate normal distribution** provides the basis of many methods for analyzing longitudinal continuous responses. Since the normal distribution can be linked to a binary response through its cumulative distribution function, the probit, it is not surprising that methods based on the multivariate probit likelihood have been proposed for multivariate binary data, such as that of Table 1 (*see* **Quantal Response Models**). However, the probit does not have a closed form, and requires the use of moment-based approximations typically requiring large samples (e.g. [25]) or computationally intensive numerical quadrature (e.g. [9]). The probit scale also lacks some of the desirable properties of the log **odds** scale, for example the ability to estimate the same effect from both retrospective **case–control studies** and prospective **cohort studies**. However, multivariate generalizations of the logistic distribution lack the flexibility of the multivariate normal.

For the **multinomial distribution** the same parameters occur in both the first and higher-order **moments** of the distribution, and no model exists that can simultaneously produce simple expressions, in terms of model parameters, for the joint, marginal, and conditional distributions [15]. Similar considerations often apply with other types of data with which we are faced. Thus, an important

Table 2 Treatment trial of epilepsy

ID	Y1	Y2	Y3	Y4	Treatment	Baseline	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
206	11	0	0	5	0	50	26
210	0	0	3	3	0	18	28
213	37	29	28	29	0	111	31
215	3	5	2	5	0	18	32
217	3	0	6	7	0	20	21
219	3	4	3	4	0	12	29
220	3	4	3	4	0	9	21
222	2	3	3	5	0	17	32
226	8	12	2	8	0	28	25
227	18	24	76	25	0	55	30
230	2	1	2	1	0	9	40
234	3	1	4	2	0	10	19
238	13	15	13	12	0	47	22
101	11	14	9	8	1	76	18
102	8	7	9	4	1	38	32
103	0	4	3	0	1	19	20
108	3	6	1	3	1	10	30
110	2	6	7	4	1	19	18
111	4	3	1	3	1	24	24
112	22	17	19	16	1	31	30
113	5	4	7	4	1	14	35
117	2	4	0	4	1	11	27
121	3	7	7	7	1	67	20
122	4	18	2	5	1	41	22
124	2	1	1	0	1	7	28
128	0	2	4	0	1	22	23
129	5	4	0	3	1	13	40
137	11	14	25	15	1	46	33
139	10	5	3	8	1	36	21
143	19	7	6	7	1	38	35
147	1	1	2	3	1	7	25
203	6	10	8	8	1	36	26
204	2	1	0	0	1	11	25
207	102	65	72	63	1	151	22
208	4	3	2	4	1	22	32

(*continued overleaf*)

Table 2 (*continued*)

ID	Y1	Y2	Y3	Y4	Treatment	Baseline	Age
209	8	6	5	7	1	41	25
211	1	3	1	5	1	32	35
214	18	11	28	13	1	56	21
218	6	3	4	0	1	24	41
221	3	5	4	3	1	16	32
225	1	23	19	8	1	22	26
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	2	2	1	12	37

Reproduced from [34] by permission of the publisher.

preliminary consideration before embarking on longitudinal analysis of discrete data is to decide which effects should be directly parameterized in the model, and which represented only by more complex functions of those parameters or treated as a nuisance. This is the rationale underlying the GEE approach, which for the most part is concerned with how to make inference about the marginal regression parameters of a generalized linear model (GLM), and to a lesser extent with the association among responses.

Generalized Linear Models

GLMs cover a range of models in common use in medical statistics, including Gaussian **linear regression**, **logistic regression**, and **loglinear models** for count data. As described in **Generalized Linear Model**, the general class of model requires the specification of a link function that relates the mean response to a vector of covariates $\mathbf{X}_i$ and a variance function that relates the variance of the response to the mean. Then, for any response with a distribution that is a member of the **exponential family**, the likelihood L_i can be expressed as $L_i = f(y_i) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}$, where in canonical form $\theta_i = \eta_i = \mathbf{X}_i\boldsymbol{\beta}$, $\mathrm{E}(y_i) = \mu_i = b'(\theta_i)$, and $\mathrm{var}(y_i) = b''(\theta_i)a(\phi)$. In the case of binary logistic regression $b(\theta_i) = \log(1 + \mathrm{e}^{\theta_i})$ and $\phi = 1$, while **Poisson regression** for a count response has $b(\theta_i) = \mathrm{e}^{\theta_i}$ again with $\phi = 1$.

For any members of this wide class of models, maximum likelihood estimates of the regression coefficients $\boldsymbol{\beta}$ can be obtained by an iterative weighted least squares solution (*see* **Generalized Linear Model**) to the score equations (the derivatives of L_i with respect to the regression parameters)

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} v_i^{-1}\{Y_i - \mu_i(\boldsymbol{\beta})\} = 0, \quad (1)$$

where $\{v_i^{-1}\}$ are weights derived from the variance function $v_i = \mathrm{var}(Y_i)$. The large sample **covariance matrix** of the parameter estimates is given by the inverse of the Hessian $\mathbf{H}_1(\boldsymbol{\beta})$

$$\hat{\mathbf{H}}_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} v_i^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right). \quad (2)$$

Huberized/Sandwich Estimator of a Sample Covariance Matrix

The validity of the covariance matrix based on (2) depends upon the correctness of the specification of the variance function. An alternative estimator for the covariance matrix that provides a consistent estimate even when the specification of the variance function is incorrect is the so-called "sandwich" estimator $\mathbf{H}_1^{-1}(\boldsymbol{\beta})\mathbf{H}_2(\boldsymbol{\beta})\mathbf{H}_1^{-1}(\boldsymbol{\beta})$, where

$$\hat{\mathbf{H}}_2(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} v_i^{-1}\{Y_i - \mu_i(\boldsymbol{\beta})\} \times \{y_i - \mu_i(\boldsymbol{\beta})\}^{\mathrm{T}} v_i^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right). \quad (3)$$

The "bread" of the sandwich is the standard covariance matrix estimator and the sandwich is the cross-product of the empirical scores [13, 37, 38]. This is also known as the **robust** or "heteroscedastic consistent" covariance estimator, or the "variance correction".

The sandwich estimator plays an important role in GEE methodology, but its appropriateness should not be accepted uncritically. For comparison purposes Drum & McCullagh [6] consider the simple case of a difference of means test for two independent samples, for which the conventional estimate of the variance of the group difference would be $s^2(1/n_1 + 1/n_2)$, whereas the sandwich estimate is $s_1^2[(n_1 - 1)/n_1^2] + s_2^2[(n_2 - 1)/n_2^2]$. These correspond to the alternative forms of the two-sample t test that use pooled or separate variance estimates (*see* **Student's *t* Statistics**). Where one or both samples is small, the loss of power in using the latter form is substantial, and should not be accepted lightly. With medium to large samples the loss of power is likely to be slight [1].

Bias in this estimator for unbalanced designs is also considered by Chesher & Jewitt [3]. Such considerations have prompted some criticism of the use of the label "robust" for the sandwich estimator.

Independence Working Models

Combining generalized linear models and the robust estimator for var($\boldsymbol{\beta}$) provides a very simple means of analyzing repeated measures data. Individuals now provide a response vector $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iT})$ and the score equations take the form

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \{y_i - \mu_i(\boldsymbol{\beta})\} = \mathbf{0}, \qquad (4)$$

where $\mathbf{D}$ is a vector of partial derivatives $\partial\mu/\partial\beta$ and $\mathbf{V}_i$ is a covariance matrix for the Ys. The solution to these equations provides consistent estimates of $\boldsymbol{\beta}$ even where the $\mathbf{V}_i$ matrix has been specified incorrectly. Appropriate standard errors can be obtained for such a misspecified model by using the sandwich estimator. If, for estimation purposes, V_i is specified to by $\boldsymbol{\Delta}_i = \mathrm{diag}[\mathrm{var}(Y_{i1}), \ldots, \mathrm{var}(Y_{iT})]$, in other words the responses are naively assumed to be independent (an *independence working model* or IWM), then estimation can proceed using standard GLM software with each subject contributing as many records as repeated measures. The operational simplicity of this approach is hard to exaggerate. Of course, these estimates are not **efficient**, but the loss of efficiency is often not great [10].

Table 3 gives the parameter estimates from a logistic regression fitted to the respiratory disease data of Table 1, in which each subject contributes a set of four records. Three forms of **standard error** are shown together with their associated test statistic: (i) the classical standard error that is based on the false assumption that the four records within a set are independent; (ii) the "robust/sandwich" standard error; and (iii) one based on **bootstrap** resampling of subjects. The bootstrap estimate is always closer to the robust/sandwich estimate than the classical estimate, which, for between-subjects effects, is much too small. However, in the case of the within-subjects effect for the time trend, it is the robust and bootstrap estimates that are smaller. Assuming independence is not always anticonservative.

Table 3 Logistic regression estimates for respiratory data

	Independence working model			Random effects model		
Regressor	Estimate	Standard error Classical (Robust) [Bootstrap[a]]	z test	Estimate	Standard error Classical	z test
Time	−0.131	0.148 (0.132) [0.122]	−0.89 −0.99	−0.174	0.172	−1.01
Treatment	0.938	0.337 (0.445) [0.456]	2.78 2.11	1.220	0.596	2.05
Age	−0.034	0.016 (0.019) [0.025]	−2.13 −1.79	−0.041	0.027	−1.52
Baseline	2.770	0.392 (0.506) [0.615]	7.07 5.47	3.758	0.767	4.90
Constant	−0.141	0.639 (0.706) [0.711]		−0.389	1.030	
Scale parameter				1.480	0.381	

[a]From 500 replicates.

Marginal vs. Conditional Estimation

Before describing potentially more efficient use of generalized estimating equations for longitudinal data it is worth considering the interpretation of the parameters that we have just obtained. We have estimated a **cross-sectional model** simultaneously to each of the four measurement occasions. The fit is to the four univariate margins and not to any joint distribution of responses as such. Moreover, unlike in a **random effects** model for repeated data there is no *subject-specific* effect. Estimates for the effects of covariates, whether time-dependent or not, are thus not conditional upon any such subject-specific effect, but instead relate to effects averaged over individuals.

The two alternative measures of the effect of a covariate are the same in the case of linear link models with additive subject-specific effects and log link models with multiplicative subject effects. For a logistic link, as with the other link functions, the average effect coincides with the subject-specific effect only for no effect or no subject-specific variance. Where subject-specific effects are substantial, Figure 1 illustrates how the marginal response curve is obtained from averaging the subject-specific curves which, if each was logistic, results not only in a marginal curve of smaller slope but one that is not even logistic in form. Zeger et al. [41] discuss the relationship among these different estimates of the regression coefficients. The broad consensus is that average effects are often of interest from a public health point of view, but may not be so pertinent where interest lies in scientific investigation of the individual-level process or in individual-level prediction. It is also of interest that **marginal models** rarely provide descriptions of the data that fully specify a possible data-generating mechanism as would be required, for example, to specify a **Monte Carlo** simulation.

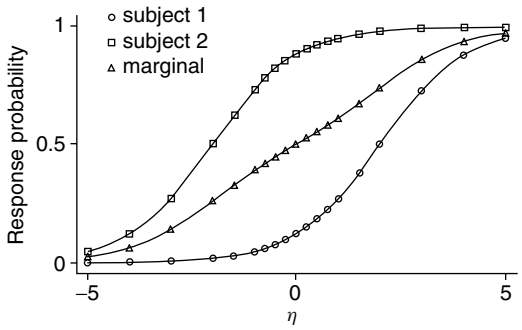

Figure 1 Logistic subject-specific and marginal effects

For comparison, Table 3 presents the estimates and standard errors from a random effects logistic model [24] applied to the respiratory data. As expected, these subject-specific regression coefficients are larger than the corresponding marginal parameters, but so too are the standard errors, resulting in comparable test statistics.

Quasi-likelihood

Wedderburn [35] pointed out that the GLM score equations (1) could be solved for any choice of link and variance function even where the integral of the score equations – a likelihood-type function given the name **quasi-likelihood** – did not actually correspond to a member of the exponential family nor even to a known parametric distribution. McCullagh [22] showed that the regression estimates obtained from solving such *quasi-score functions* were approximately normal with mean $\boldsymbol{\beta}$ and variance still given by (2).

Medical statisticians frequently encounter counts and proportions data where perhaps, through geographic, social, or genetic proximity, the several units contributing to each response are not independent. Such data then typically possess greater variance than expected under the ordinary GLM. In the case of count data, in place of a GLM specified with $\mathrm{var}(Y_i) = \mathrm{E}(Y_i)$ as appropriate to **Poisson** distributed counts, a model in which $\mathrm{var}(Y_i) = \phi \mathrm{E}(Y_i)$ might be estimated with $\phi > 1$ to account for the extra variance or **overdispersion**. Although in this instance this does not change the estimated regression coefficients, the estimated parameter variances increase by the factor ϕ.

Simple moment estimators for this scale parameter ϕ were proposed [23] on the basis of the Pearson **residuals** $\hat{r}_i = \{y_i - b'(\hat{\theta}_i)\}\{b''(\hat{\theta}_i)\}^{-1/2}$, of the form $\phi = \sum_i \hat{r}_i^2/(N-p)$. Firth [7] examined the relative efficiency of such quasi-likelihood estimators concluding that greater efficiency can only be obtained by making assumptions about higher-order **moments**.

Wei & Stram [36] consider a direct application of this kind of approach to longitudinal data in which a separate occasion-specific quasi-likelihood equation is estimated for each time point after which a special parameter covariance matrix covering all

the parameters is estimated and used as the basis of subsequent tests. This represents a special case of the IWM approach described previously (see [4]).

Table 4 presents the results from fitting overdispersed Poisson regression models to the epilepsy data using the IWM approach. For illustrative purposes, the fitted model includes simple main effects and a common pattern of overdispersion, though Thall & Vail [34] find some evidence for a more complicated structure. The estimated dispersion parameter based on the Pearson residuals is 5.1, showing that the variability in these data is well in excess of those for Poisson counts, for which it would be 1. Again the standard errors derived using the sandwich estimator are larger than the naive estimates for the between-subjects effects but is smaller for the within-subjects effect of time. We have again provided estimates from a random effects model for comparison. This model assumes a **gamma distributed** random effect uncorrelated with included regressors, acting multiplicatively on the Poisson rate parameter. Unlike the logistic model there is no consistent increase in the parameter estimates over those from the marginal model.

Generalized Estimating Equations

Liang & Zeger [16] and Zeger & Liang [40] extended the quasi-likelihood approach to consider a multivariate mean vector and suggested that improved efficiency could be obtained by simultaneously estimating parameters in the covariance matrix of the response vector. This corresponds to taking the estimating equations (3) and, where previously we have considered using a diagonal matrix as the working matrix for $\mathbf{V}_i$, now using something which begins to approximate the off-diagonal covariance. An obvious form is

$$\mathbf{V}_i = \frac{\boldsymbol{\Delta}_i^{1/2}\mathbf{R}_i(\alpha)\boldsymbol{\Delta}_i^{1/2}}{\phi},$$

where $\mathbf{R}_i(\alpha)$ is a $T \times T$ "working **correlation** matrix" with parameter vector α. Typical possibilities for $R_i(\alpha)$ are:

1. An identity matrix – equivalent to the IWM approach of the previous section.
2. An **exchangeable** correlation matrix with a single parameter, similar to that which underlies repeated measures analysis of variance in which $\text{corr}(Y_{ij}, Y_{ik}) = \alpha$, $j \neq k$.
3. An *AR-1 autoregressive* correlation matrix, also with a single parameter but in which $\text{corr}(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$, $j \neq k$ (*see* **ARMA and ARIMA Models**).
4. An unstructured correlation matrix with $T(T-1)/2$ parameters, similar to that underlying **multivariate analysis of variance**, in which $\text{corr}(Y_{ij}, Y_{ik}) = \alpha_{jk}$.

If V_i is correctly specified, then the covariance matrix for the fitted regression coefficients could be obtained

Table 4 Overdispersed Poisson regression estimates for the epilepsy data

	Independence working model			Random effects model		
Regressor	Estimate	Standard error Classical (Robust)	z test	Estimate	Standard error Classical (Robust)	z test
Time	−0.059	0.046 (0.035)	−1.28 −1.68	−0.059	0.019 (0.030)	−3.13 −1.96
Treatment	−0.154	0.108 (0.171)	−1.42 −0.90	−0.197	0.169 (0.166)	−1.17 −1.18
Age	−0.023	0.009 (0.012)	−2.56 −1.92	−0.016	0.014 (0.010)	−1.19 −1.61
Baseline $\times 10^{-1}$	0.227	0.011 (0.012)	20.6 18.9	0.279	0.030 0.026	8.97 10.70
Constant	0.712	0.326 (0.349)		0.698		
Frailty variance				0.366		

from the inverse of the **information matrix** $\mathbf{H}_1(\beta)$:

$$\sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i.$$

However, although we should have regard to possible loss of **power**, it is in general safer to use a robust sandwich variance estimator with a construction that parallels that already described.

In the first applications of non-IWM GEE estimation, Liang & Zeger [16] followed the pattern set in quasi-likelihood estimation of using moment estimators (*see* **Method of Moments**) for both the scale parameter, ϕ, and for the correlations, α. With observation-specific Pearson residuals, the scale parameter $\hat{\phi}$ is first estimated by $1/\{\sum_i \sum_j \hat{r}_{ij}^2/(N-p)\}$ and then $\hat{\alpha}$ is estimated by $\hat{\phi} \sum_i \sum_{k>j} \hat{r}_{ij}\hat{r}_{ik}/\{\sum_i [n_i(n_i-1)/2] - p\}$ for the exchangeable model, or by the regression slope of $\log(\hat{r}_{ij}\hat{r}_{ik})$ on $\log(|j-k|)$ for the AR-1 model. In the case of the unstructured or saturated form of the covariance matrix for balanced data the (jk)th element can be estimated by $\hat{\phi} \sum_i \hat{r}_{ij}\hat{r}_{ik}/(N-p)$. This last form is typically only practicable with large samples or with few observation occasions. Whichever form is chosen, the estimation process consists of an iteration between iterative weighted least squares estimation of the regression parameters for a given estimate of the association parameters followed by recalculation of the estimates of the association parameters based on the residuals from the current estimates of the regression parameters.

Of course the correlation matrix need not be parameterized directly. Indeed, since multivariate discrete distributions give rise to complex constraints on the feasible correlation matrices there may be some advantage in not doing so. In the case of binary data Lipsitz et al. [19] suggested using the pairwise marginal **odds ratios** γ_{ijk} for $j = 1, T-1$ and $k = j+1, T$ where

$$\gamma_{ijk} = \frac{\mathrm{E}(Y_{ij}Y_{ik})\mathrm{E}[(1-Y_{ij})(1-Y_{ik})]}{\mathrm{E}[Y_{ij}(1-Y_{ik})]\mathrm{E}[(1-Y_{ij})Y_{ik}]}. \qquad (5)$$

Although these estimators for the association parameters α may suffice where scientific interest lies in the estimation of the regression coefficients, they are not especially good estimators of the association parameters themselves. Moreover, it is sometimes appropriate to consider the association, for example as characterized by the correlation matrix, as depending explicitly on covariates or experimental or sampling design variables. Prentice [26] proposed extending the **estimating equation** approach to the estimation of the association parameters. A second set of equations is estimated for which the "observed data" are the cross-products over measurements of the time-specific residuals $s_{ijk} = r_{ij}r_{ik}$. As was the case for the score equations for the regression coefficients, this involves a choice of some "working covariance matrix". In simple cases the estimates obtained correspond to the simple moment estimates already described.

These two sets of estimating equations can be combined [28, 42] as follows:

$$\sum_i \begin{pmatrix} \dfrac{\partial \mu_i}{\partial \beta}, & \dfrac{\partial \mu_i}{\partial \alpha} \\ \dfrac{\partial \sigma_i}{\partial \beta}, & \dfrac{\partial \sigma_i}{\partial \alpha} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathrm{cov}(y_i), & \mathrm{cov}(y_i, s_i) \\ \mathrm{cov}(y_i, s_i), & \mathrm{cov}(s_i) \end{pmatrix}^{-1} \times \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix} = \mathbf{0},$$

where $\sigma_{ijk} = \mathrm{E}(s_{ijk})$. Liang et al. [17] present a similar set of estimating equations for the mean and marginal odds ratios. In either case, the score functions for β and α are solved together using a modified Fisher scoring algorithm (*see* **Generalized Linear Model**). If the equations for α and β are assumed orthogonal, then the off-diagonal elements of the first two matrices (for derivatives and weights, respectively) are assumed zero and these are referred to as *first-order estimating equations* and the estimation method as GEE1. If these matrix elements are nonzero and the two sets of equations are solved jointly, then this corresponds to a *second-order estimating equations* approach or GEE2. GEE1 gives consistent estimates of β when the model for α is misspecified. GEE2 does not share this "robustness" but can give more efficient estimates. In practice (e.g. [17]), there is little or no gain in efficiency for β in the use of GEE2, but there can be substantial gains for the association parameters α.

As with the estimation equation for β, these estimating equations for α now involve the specification of a "working" covariance matrix for $\mathrm{cov}(s_i)$ (for GEE1 and GEE2) and $\mathrm{cov}(y_i, s_i)$ (for GEE2 only). These involve third and fourth moments of the data about which most applied statisticians will have little intuitive grasp and which may be subject to complex constraints. (In theory, the estimating equation approach could be extended further to estimate these

higher-order moments as well.) In addition, for large numbers of repeated measures the size of this covariance matrix can make it difficult to invert.

In the discussion of Liang et al. [17], Prentice & Pepe [27] suggested a reparameterization in terms of conditional residuals, which, tending to be less correlated than unconditional residuals, might provide greater efficiency. Carey et al. [2] (but see also [5]) proposed a simple iterative schema involving conditional residuals for the special case of odds-ratio association models with multivariate binary data. They suggested estimating a common odds ratio $\alpha_{ijk} = \alpha$ [see (4) above] from the following logistic regression:

$$\text{logit } \Pr(Y_{ij} = 1 | Y_{ik} = y_{ik}) = \alpha y_{ik} + \log\left(\frac{u_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right),$$

where $\mu_{ijk} = \Pr(Y_{ij} = Y_{ik} = 1)$ and the second term on the right-hand side is an *offset*. Since μ_{ijk} and hence the offset depends upon the values of α and β, iteration is required between the offset logistic regression for α and the GEE logistic regression for β, giving rise to the name *alternating logistic regression* (ALR). The approach can be extended to allow variation in the degree of association with covariates by replacing y_{ik} in the above equation by $x_{ijk} y_{ik}$ with a corresponding change in α to a vector of regression-type coefficients.

This last approach exploits the robustness properties of GEE1 estimation with respect to the β regression parameters but typically achieves almost as good efficiency in the estimation of the odds-ratio association parameters α as the much more complex GEE2 method, and at the same time substantially reduces the matrix inversion problem. Lipsitz & Fitzmaurice [18] consider the conditional residual approach where the association is modeled using correlations. They report similarly good performance for autoregressive correlation structures, and more generally where there are missing data.

Recently Wild & Yee [39] have adapted the GEE approach to allow for the fitting of **generalized additive models** (GAMs rather than GLMs) of the sort described by Hastie & Tibshirani [12].

In practice, the GEE estimation algorithms described above do not always converge. Lipsitz et al. [20] considered restricting the estimation process to one step. Starting from an IWM solution, the association parameters are estimated using the IWM residuals and the regression coefficients are re-estimated once more. They concluded that where the fully iterated method failed to converge, the one-step estimator could well be adequate. They found no discernible differences in power or bias among IWM logistic regression, one-step or iterated GEE1. In addition, efficiency was comparable except where the correlation over time was high and the variable time-varying.

Comparative Results from GEE Models

Table 5 presents results of logistic regressions with exchangeable, AR-1, unstructured, and ALR estimates, with classical and robust standard errors. Table 6 gives the lower triangle of the estimated correlation matrices from each of these models. The estimated correlations from the unstructured model look to be closer to those of an exchangeable structure than the autoregressive structure, but regression estimates and standard errors from the three models are all very similar, as also are the classical and robust standard errors from within the same model (where any differences might have suggested possible model misspecification).

Table 7 presents results from the epilepsy data with exchangeable, overdispersed exchangeable, AR-1, and unstructured models, again with standard and robust standard errors. As expected, allowing for overdispersion in the exchangeable model does not result in different parameter estimates nor in different estimates of the robust standard errors. Differences do arise in the case of the classical standard errors, reflecting the fact that for these data allowing for the correlation alone is not sufficient to avoid a misspecified model, a fact indicated by the substantial differences between classical and robust standard errors. Once allowance is also made for overdispersion the two forms of standard error are much more comparable. Inspection of the estimated correlation matrix under the unstructured model shown in Table 8 might suggest the autoregressive structure as the more appropriate of the structured models.

Missing Data, Weighted Estimating Equations, and Complex Sampling Designs

In general, under the assumption of data *missing completely at random* (MCAR) [21], marginal model

Table 5 Logistic regression estimates for respiratory data

			Classical		Robust	
Regressor	Model	Estimate	Standard error	z test	Standard error	z test
Time	Exch	−0.131	0.129	−1.01	0.132	−0.99
	AR-1	−0.148	0.152	−0.97	0.133	−1.11
	Unstr	−0.169	0.131	−1.29	0.140	−1.21
	ALR	−0.118			0.131	−0.90
Treatment	Exch	0.924	0.438	2.11	0.444	2.08
	AR-1	0.894	0.395	2.26	0.447	2.00
	Unstr	0.939	0.439	2.14	0.450	2.09
	ALR	0.934			0.459	2.04
Age	Exch	−0.034	0.020	−1.67	0.019	−1.76
	AR-1	−0.034	0.018	−1.85	0.019	−1.77
	Unstr	−0.034	0.020	−1.67	0.019	−1.77
	ALR	−0.034			0.019	−1.82
Baseline	Exch	2.739	0.507	5.41	0.504	5.43
	AR-1	2.739	0.455	6.02	0.501	5.47
	Unstr	2.881	0.521	5.53	0.506	5.70
	ALR	2.777			0.513	5.42
Constant	Exch	−0.150	0.751		0.704	
	AR-1	−0.113	0.720		0.685	
	Unstr	−0.088	0.745		0.745	
	ALR	−0.237			0.582	

Table 6 Estimated correlation matrices for the respiratory data

Exchangeable (and ALR) models	AR-1 model	Unstructured
1	1	1
0.23 (0.18) 1	0.23 1	0.13 1
0.23 (0.18) 0.23 (0.18) 1	0.05 0.23 1	0.18 0.19 1
0.23 (0.18) 0.23 (0.18) 0.23 (0.18) 1	0.01 0.05 0.23 1	0.21 0.35 0.34 1

specification together with an appropriate parameterization of the covariance matrix requires the occasion-wise subsets of data to be complete for the modeling of each response measurement, but does not require complete data across all measurement occasions (*see* **Nonignorable Dropout in Longitudinal Studies**). Moreover, GEE methods typically provide estimates of parameters whose interpretation is not dependent upon the pattern of MCAR missing data. However, the occurrence of missing data complicates the estimation of the totally unspecified correlation matrix since the estimate obtained from the nonmissing data is not guaranteed to be positive definite.

Robins et al. [30] describe how weighting the estimating equations by the weights given by the inverse of the response probabilities can extend the missing data properties of GEE estimation to the case of data *missing at random* (MAR). This approach also allows the application of the GEE methodology to multiphase designs and other designs involving the use of **surrogate** measurement. **Multistage sampling** designs are considered by Qaqish & Liang [29].

Specialized methods are required for the application of GEE methodology in the presence of *nonignorable* missing data.

Discussion

The field of GEE continues to be one of rapid development, particularly in respect of efficient estimation of the structure of association among observations.

Table 7 GEE Poisson regression estimates for the epilepsy data

			Classical		Robust	
Regressor	Model	Estimate	Standard error	z test	Standard error	z test
Time	Exch	−0.059	0.016	−3.76	0.035	−1.67
	Exch[a]	−0.059	0.035	−1.67	0.035	−1.67
	AR-1	−0.064	0.020	−3.20	0.034	−1.87
	Unstr	−0.052	0.018	−2.87	0.043	−1.21
Treatment	Exch	−0.154	0.071	−2.10	0.173	−0.89
	Exch[a]	−0.154	0.161	−0.88	0.161	−0.93
	AR-1	−0.165	0.068	−2.41	0.162	−1.02
	Unstr	−0.148	0.067	−2.20	0.132	−1.12
Age	Exch	0.023	0.006	3.92	0.012	1.93
	Exch[a]	0.023	0.014	1.73	0.012	1.96
	AR-1	0.026	0.006	4.53	0.012	2.17
	Unstr	0.024	0.006	4.19	0.012	1.92
Baseline	Exch	0.228	0.008	30.1	0.012	18.2
×10	Exch[a]	0.228	0.017	13.3	0.013	18.1
	AR-1	0.232	0.007	32.01	0.013	18.5
	Unstr	0.228	0.007	31.9	0.012	19.3
Constant	Exch	0.712	0.205		0.352	
	Exch[a]	0.712	0.464		0.358	
	AR-1	0.597	0.199		0.354	
	Unstr	0.625	0.195		0.381	

[a]Allowing for overdispersion.

This will further enhance the value of these methods for longitudinal data. As with other model fitting methods for longitudinal data, special care needs to be taken in the interpretation of results where predictor variables include those that may be endogenous. This can include baseline measures of the type used in the illustrations.

Programs for GEE model fitting are available for **software** platforms such as **S-PLUS** [33], with a suite of such programs and data exploration and management tools specifically for longitudinal data being provided by OSWALD [31]. With the more recent availability of basic GEE methods in more popular packages (e.g. the SAS macros of [14]; Stata V.5.0 [32]) widespread application of these methods is assured.

Table 8 Estimated correlation matrix for the epilepsy data

Unstructured model			
1			
0.24	1		
0.42	0.68	1	
0.21	0.29	0.59	1

Acknowledgments

This work was partially supported by grant H519255031 funded under the ALCD program of the Economic and Social Research Council.

References

[1] Breslow, N.E. (1990). Tests of hypotheses in overdispersion regression and other quasi-likelihood models, *Journal of the American Statistical Association* **85**, 565–571.

[2] Carey, V., Zeger, S.L. & Diggle, P. (1993). Modeling multivariate binary data with alternating logistic regressions, *Biometrika* **80**, 517–526.

[3] Chesher, A. & Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator, *Econometrica* **55**, 1217–1222.

[4] Davis, C.S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials, *Statistics in Medicine* **8**, 1959–1980.

[5] Diggle, P. (1992). Discussion of Liang, K-Y., Zeger, S.L. & Qaqish, B. (1992) "Multivariate regression analyses for categorical data" (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[6] Drum, M. & McCullagh, P. (1993). Comment on Fitzmaurice, G.M., Laird, N. and Rotnitzky, A. "Regression

models for discrete longitudinal responses", *Statistical Science* **8**, 284–309.

[7] Firth, D. (1987). On the efficiency of quasi-likelihood estimation, *Biometrika* **74**, 233–246.

[8] Fitzmaurice, G.M., Laird, N. & Rotnitzky, A. (1993). Regression models for discrete longitudinal responses, *Statistical Science* **8**, 284–309.

[9] Gibbons, R.D. & Hedeker, D.R. (1992). Full-information item bi-factor analysis, *Psychometrika* **57**, 423–436.

[10] Glonek, G.F.V. & McCullagh, R. (1995). Multivariate logistic models, *Journal of the Royal Statistical Society, Series B* **57**, 533–546.

[11] Godambe, V.P. (1960). An optimal property of regular maximum likelihood estimation, *Annals of Mathematical Statistics* **31**, 1209–1211.

[12] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[13] Huber, P. (1967). The behaviour of maximum likelihood estimators under nonstandard conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 221–233.

[14] Karim, M.R. (1989). *Technical Report 674*. Department of Biostatistics, Johns Hopkins University.

[15] Kenward, M.G. & Jones, B. (1992). Alternative approaches to the analysis of binary and categorical repeated measurements, *Journal of Biopharmacological Statistics* **2**, 137–170.

[16] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[17] Liang, K.-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[18] Lipsitz, S.R. & Fitzmaurice, G.M. (1996). Estimating equations for measures of association between repeated binary responses, *Biometrics* **52**, 903–912.

[19] Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J. & Laird, N.M. (1994). Performance of generalized estimating equations in practical situations, *Biometrics* **50**, 270–278.

[20] Lipsitz, S.R., Laird, M. & Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, *Biometrika* **78**, 153–160.

[21] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[22] McCullagh, P. (1983). Quasilikelihood functions, *Annals of Statistics* **11**, 59–67.

[23] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[24] Mauritsen, R.H. (1984). Logistic Regression With Random Effects, *Unpublished Ph.D. thesis*. Department of Biostatistics, University of Washington.

[25] Muthen, B.O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators, *Psychological Medicine* **49**, 115–132.

[26] Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**, 1033–1048.

[27] Prentice, R.L. & Pepe, M.S. (1992). Contribution to the discussion of Liang, K-Y., Zeger, S.L. and Qaqish, B. (1992) "Multivariate regression analyses for categorical data" (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[28] Prentice, R.L. & Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics* **47**, 825–883.

[29] Qaqish, B.F. & Liang, K.-Y. (1992). Marginal models for correlated binary responses with multiple classes and multiple levels of nesting, *Biometrics* **48**, 939–950.

[30] Robins, J., Rotnitzky, A. & Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**, 106–121.

[31] Smith, D.M. & Diggle, P.J. (1994). *OSWALD Version 2: Object Oriented Software for the Analysis of Longitudinal Data in S*. Department of Mathematics and Statistics, University of Lancaster, England.

[32] StataCorp (1997). *Stata Statistical Software: Release 5.0*. Stata Corporation, College Station.

[33] SSI (1988). *S-Plus User Manual*. Statistical Sciences Inc., Cary.

[34] Thall, P.F. & Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**, 657–671.

[35] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss–Newton method, *Biometrika* **61**, 439–447.

[36] Wei, L.J. & Stram, D.O. (1988). Analyzing repeated measurements with possibly missing observations by modelling marginal distributions, *Statistics in Medicine* **7**, 139–148.

[37] White, H. (1980). A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**, 817–850.

[38] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1–25.

[39] Wild, C.J. & Yee, T.W. (1996). Additive extensions to generalized estimating equations methods, *Journal of the Royal Statistical Society, Series B* **58**, 711–725.

[40] Zeger, S.L. & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**, 121–130.

[41] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

[42] Zhao, L.P. & Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika* **77**, 642–648.

ANDREW PICKLES

Generalized Linear Mixed Models

Introduction

Generalized linear mixed models (GLMMs) are an extension of the class of **generalized linear models** in which **random effects** are added to the linear predictor. This modification extends the broad class of generalized linear models to accommodate **correlation** via random effects, while retaining the ability to model nonnormal distributions and allowing nonlinear models of specific form. The class of GLMMs includes the special cases of **linear mixed models**, **random coefficient models**, random effects **logistic regression**, and random effects **Poisson regression**, to name a few.

The incorporation of random effects is a natural way to model or accommodate correlation in the context of a nonlinear model for nonnormal data. It generates a rich class of correlated data models that would be difficult to specify directly. Readily available, flexible, **multivariate distributions** analogous to the **multivariate normal distribution** do not exist for most nonnormally distributed data.

Inferences for these models can be of the usual variety, that is, modeling the effect of predictors on the mean, in which case the random effects and correlation are "nuisance" features of the model. In other situations, however, both estimation and testing of the **variances** of the random effects, as well as prediction of the realized values of the random effects, may be of interest (*see* **Variance Components**).

We will illustrate several of our points using as an example the **longitudinal** study of physicians and their patients described by Korff et al. [8]. This study classified 44 primary care physicians in a large HMO according to their practice styles in treating back **pain** management (low, moderate, or high frequency of prescription of pain medication and bed rest), and followed an average of 24 patients per physician for 2 years (1 month, 1 year, and 2 year follow-ups) after the index visit. Outcome variables included functional measures (e.g. Did you experience moderate to severe activity limitation?), patient satisfaction (e.g. "After your visit with the doctor, did you fully understand how to take care of your back problem?"), and cost.

Generalized Linear Mixed Models: A Definition

Generalized linear mixed models constitute a class of models for describing the stochastic relationship of an n-dimensional outcome vector $\boldsymbol{Y}$ to an $(n \times p)$-dimensional matrix of **covariates** $\boldsymbol{X}$, with rows $\boldsymbol{x}_i'$.

The construction of generalized linear mixed models begins with the specification of a generalized linear model conditional on a vector $\boldsymbol{u}$ of random effects. That is, given a vector $\boldsymbol{u}$ (often with components specific to a subject or cluster), the conditional density of Y_i is of the **exponential family** form $f(y_i|\boldsymbol{u}) = \exp[\{y_i\theta_i - b(\theta_i)\}\phi + c(y_i, \phi)]$, where b and c are functions of known form. In addition, one assumes that $E(Y_i|\ \boldsymbol{u}, \boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u})$, where $\boldsymbol{z}_i$ is a specified vector of covariates, analogous to $\boldsymbol{x}_i$. Given $\boldsymbol{u}$, the model additionally assumes that the responses Y_i are independent. The function g links the linear predictor to the expected value of the response. The model further assumes that the random effects $\boldsymbol{u}$ follow a distribution G, typically (but not necessarily) multivariate normal with mean vector $\mathbf{0}$ and **covariance matrix** $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a vector of (co)variance parameters, for example, variances and correlation coefficients.

Thus, the model assumes that the linear predictor consists of two portions: the fixed effects portion $\boldsymbol{x}_i'\boldsymbol{\beta}$, and the random effects portion $\boldsymbol{z}_i'\boldsymbol{u}$, for which a distribution is assigned to $\boldsymbol{u}$. Just as with linear mixed models, the assumption of a distribution for the random effects induces correlations among observations. Finally, the assumptions underlying GLMMs specify the multivariate distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, so that one can base inference with these models on **likelihood** methods.

The specification of covariate effects conditional on random effects determines the interpretation of the fixed effects parameters $\boldsymbol{\beta}$. For example, considering the activity limitation outcome in the back pain study, GLMMs would measure how the risk of activity limitation in a particular patient of a particular physician changes over time, and how that change relates to the practice style of the physician. Using GLMMs, one can also directly relate changes in **explanatory variables** within an individual subject to changes in the expected value of the subject's response.

To be more specific, we consider the simple and common case in which the data are correlated in clusters, where $i = 1, \ldots, m$ indexes the clusters (e.g.

patients) while $j = 1, \ldots, n_i$ indexes units within clusters (e.g. different time points) (*see* **Cluster Sampling**). In a GLMM with a single fixed effect, the parameter β measures the change in the conditional expectation of Y corresponding to a unit increase in the covariate within the ith cluster,

$$\beta = g[E(Y_{ij}|x_{ij} + 1, \boldsymbol{u}_i)] - g[E(Y_{ij}|x_{ij}, \boldsymbol{u}_i)], \quad (1)$$

where $\boldsymbol{u}_i$ is a cluster-specific vector of random effects. This contrasts with **marginal models**, where one specifies the marginal or population-averaged (PA) distribution of the response of the jth unit in the ith cluster, integrated over the distribution of $\boldsymbol{u}$, together with some working (hypothesized) covariance structure for the n_i responses in the ith cluster to account for intraclass correlation. For the single fixed effect example, marginal models measure the change in the marginal expectation $\mathrm{E}(Y_{ij}|x_{ij})$, unconditional on the random effects, associated with change in the covariate. Such covariate effects are exactly those one would estimate with a single response per subject; the cluster structure thus plays no role in the interpretation of the model regression coefficients.

In addition to estimates of the effects of covariates on the expected value of the response, GLMMs can provide estimates of the dependence of responses within clusters, such as subjects. Measures such as the intraclass correlation coefficient, $\mathrm{corr}(Y_{ij}, Y_{ij'})$, depend on the random effects distribution G, along with its parameters $\boldsymbol{\gamma}$; using estimates of $\boldsymbol{\gamma}$ one can construct estimates of intraclass correlation.

As well as modeling within-cluster response dependence and estimates of its magnitude, GLMMs allow consideration of the individual random effects, which themselves may be of interest. For example, in the back pain study, we would include random effects to describe both the physician and patient effects on each of the outcomes. We might be interested in obtaining predicted values for the random effects of each physician to help indicate which physicians had better outcomes and/or lower costs, after adjusting for fixed effects of model covariates. The random effects in a GLMM are best predicted by their conditional expectations given the data, $E(\boldsymbol{u}|\boldsymbol{Y})$. However, this expectation is unknown since it depends on the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Hence, one typically estimates $E(\boldsymbol{u}|\boldsymbol{Y})$ using estimates of these parameters. The estimated values of $E(\boldsymbol{u}|\boldsymbol{Y})$ are *shrinkage* estimates and "borrow strength" across the data set in order to improve estimates of individual random effects, especially when data incorporating a particular random effect are sparse (*see* **Shrinkage Estimation**).

Inference and Estimation

Maximum likelihood (ML), or variants such as **restricted maximum likelihood** (REML), are standard methods of estimation for linear mixed models and generalized linear models (e.g. logistic regression). Evaluation of the likelihood and hence likelihood inference with GLMMs is computationally difficult, however, because the random effects on which the likelihood is conditioned must be integrated out of the distribution prior to maximization as a function of the fixed effects. Although several useful computational methods currently exist, the development of new methods for GLMMs continues to be an active research area.

To illustrate the inherent complexity, consider a general mixed logistic regression model for binary data. The marginal likelihood takes the form

$$\int \cdots \int \exp\left\{\sum_i Y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u})\right\} \times \prod_i \{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u})\}^{-1} \, \mathrm{d}G(\boldsymbol{u}), \quad (2)$$

where G is the distribution function of the random effects and the integration is of a dimension equal to the dimension of $\boldsymbol{u}$. For most choices of G, the integral cannot be evaluated in closed form although, for simple cases like random intercept and/or random slope models, (1) reduces to a product of lower-dimensional integrals amenable to numerical integration. **Numerical integration** becomes inaccurate for three or more dimensions, but simulation-based methods such as **Markov Chain Monte Carlo** [3], in particular, Gibbs sampling, and Monte Carlo **EM** [13] have proven useful in these settings. Such methods can also handle complications such as crossed random effects.

If ML estimation is feasible, then the usual inferential methods are available. In particular,

- ML estimators are asymptotically normal, with standard errors available from second derivatives of the log likelihood.

- One can carry out hypothesis tests using **likelihood ratio**, score, or Wald procedures.
- One can calculate best-predicted values as expected values of the random effects conditional on the data, substituting ML or REML estimates for unknown parameters. Typically, one cannot evaluate the conditional expected values in closed form, so these calculations involve numerical integration.
- One can test whether variances of random effects are zero using the likelihood ratio statistic. As with linear mixed models, the asymptotic null distribution involves a mixture of chi-square distributions rather than the null **chi-square distribution** usual in fixed effects models [15].

For the case of random intercepts only, several authors have proposed a **semiparametric** mixed model approach that jointly estimates the regression parameters and the (**nonparametric**) mixing distribution G. These methods are given in several papers, including [1, 2, 9, 10, 11]. This approach provides consistent estimation of the effects of all covariates and of G under conditions of **identifiability** [7].

Although GLMMs require specification of the random effects distribution G, several studies of the performance of logistic models with random intercepts show that estimates of the fixed effects parameters $\boldsymbol{\beta}$ are robust to **misspecification** of G. For example, using both approximations and **simulations**, Neuhaus et al. [14] showed that incorrectly assuming that G was Gaussian produced fixed effects covariate effect estimates with very little bias (*see* **Unbiasedness**). Heagerty and Kurland [5] corroborated these findings but pointed out that more severe model misspecifications, such as the failure to model interactions of covariates and random effects, could yield biased estimates of covariate effects.

Contrasting the Marginal and Conditional Approaches

We return to the special case of correlated clusters and the notation used previously to describe them. Many investigators have considered alternative methods for clustered data that focus on models for the marginal expectation of the response,

$$E(Y_{ij}|\,\boldsymbol{x}_{ij}) = \int y \left[\int \cdots \int f(y|\boldsymbol{x}_{ij}, \boldsymbol{u}_i)\, \mathrm{d}G(\boldsymbol{u}_i) \right] \mathrm{d}y, \quad (3)$$

where $\boldsymbol{x}_{ij}$ is the vector of covariate values associated with Y_{ij}. Some approaches, such as **Generalized Estimating Equations** (GEEs), do this without fully specifying the functional form of the joint distribution of the responses $Y_{i1}, \ldots, Y_{in_i}$ within the ith cluster.

GLMM and marginal models are similar in that both parameterize the mean and covariance matrices of correlated groups of observations, and both base inferences on **marginal likelihoods** or marginal **quasi-likelihoods** of the observed data. However, the implications of modeling the response distribution conditional on random effects, as do GLMMs, rather than averaged over random effects, as do marginal models, are profound. With a nonlinear link function, the impact of a conditional main effect in the linear predictor varies, on the scale of $\mathrm{E}(Y)$, with the values of accompanying fixed and random effects. Hence, in contrast with the linear case, when averaging over the random effect distribution, the mean linear predictor does not correspond (transform to) to the marginal conditional expectation $\mathrm{E}_{\boldsymbol{u}_i}[\mathrm{E}(Y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{u}_i)]$. Similarly, linear predictors based on within-group covariate means do not transform to means of the response within the corresponding groups. Hence, as stated earlier, marginal approaches measure conceptually and numerically different covariate effects than do GLMMs. In some cases, scientific interest and inferential goals of a problem lead one clearly to the marginal or conditional specification of a model; in other cases, the appropriate direction is less clear. Since the distinctions between marginal models and GLMMs are commonly blurred in practice, it is useful to contrast them further.

Marginal models are most helpful when correlation among observations cannot be ignored, but neither the nature of the clustering that generates such correlations, nor the individual clusters themselves, are otherwise of particular scientific interest. This characterization often applies when public health impact is the focus of an investigation. From that perspective, there may be little concern with either intracluster factors or with predictions about the specific units, such as families observed cross-sectionally or individuals observed over time, that generate the

measurement clusters. In the back pain study mentioned earlier, in which practice style was constant within subjects over time, marginal modeling would be natural to study a presumed homogeneous effect of practice style across subjects and times, and any population time trend. (However, it would not be useful for quoting the **odds** of reduction over time in, say, the risk of activity limitation for an individual patient.)

In such circumstances, for example, where longitudinal observation is used for logistical reasons or statistical efficiency, and correlations are nuisance parameters in analysis, then, there is technical advantage in bypassing a conditional model (*see* **Efficiency and Efficient Estimators**). For example, GLMMs that mistakenly assume random effects that are homoscedastic can produce biased estimators [5, 6]. In contrast, estimates of marginal model parameters obtained using GEE are **consistent**, even if the association structure is misspecified.

In other circumstances, however, marginal models may not measure covariate effects of primary scientific interest, for example, in longitudinal studies in which explanatory variables change over time *within a subject* and interest is in how *individuals respond* to such changes. In the back pain study, investigators were interested specifically in assessing patterns of change over time in individual subjects. In such situations, GLMMs are more ambitious than marginal models in attempting to (i) parse, using explicit random effects and predicted values for them, sources of variation that produce correlated observations, and to (ii) portray and predict, for instance, shapes of individual longitudinal disease trajectories, vulnerabilities of individual litters to teratogenesis, breeding values of individual bulls, and susceptibilities of individual families to inheritable diseases. The price of this ambition is paid in more stringent assumptions, and in greater complexity of the model fitting process and computations.

If one is willing to pay such a price, then GLMMs may be used to make inferences about marginal distributions, even when purely marginal methods would be adequate. Marginal modeling, however, without such assumptions *does not* allow conditional inference, for example, about longitudinal trajectories typical of individual subjects. More specifically,

- Variations of **Simpson's paradox** and the **ecologic fallacy** may apply at several levels. The marginal distribution may be of a different form from any conditional distribution and, indeed, the form of a conditional mixing distribution required for common marginal models may be quite unusual [17]. In extreme cases, features in every conditional model may be absent in the marginal model, for example, the marginal effect averaged across 2×2 conditional tables might be opposite in direction to those in each conditional table. More commonly, however, population average effects may simply understate the strengths of effects on individuals [18].
- GEE for marginal models may not estimate the variance–covariance structure efficiently and does not allow, without further assumptions, prediction of random effects. However, see [4, 16, 18] for developments in these directions.

For a more detailed critique of marginal modeling, see [12].

Summary

During the past decade, GLMMs have become an important statistical tool and now see heavy use for modeling correlated, nonnormally distributed data. Software for fitting GLMMs is starting to mature, and much experience has contributed to better appreciation of both the utility and pitfalls of the currently available techniques. GLMMs are a natural modeling approach for longitudinal data when changes within subjects are of interest.

References

[1] Butler, S.M. & Louis, T.A. (1992). Random effects models with non-parametric priors, *Statistics in Medicine* **11**, 1981–2000. Disc:2017–2023.

[2] Follmann, D. & Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing, *Journal of the American Statistical Association* **87**, 295–300.

[3] Gilks, W.R., Richardson, S. & Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[4] Heagerty, P. (1999). Marginally specified logistic-normal models for longitudinal binary data, *Biometrics* **55**, 688–698.

[5] Heagerty, P. & Kurland, B. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models, *Biometrika* **88**, 973–986.

[6] Heagerty, P.J. & Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors), *Statistical Science* **15**(1), 1–26.

[7] Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics* **27**, 887–906.

[8] Korff, M., Barlow, W., Cherkin, D. & Deyo, R. (1994). Effects of practice style in managing back pain, *Annals of Internal Medicine* **121**, 187–195.

[9] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixture distribution, *Journal of the American Statistical Association* **73**, 805–811.

[10] Lesperance, M. & Kalbfleisch, J. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution, *Journal of the American Statistical Association* **87**, 120–126.

[11] Lindsay, B., Clogg, C. & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis, *Journal of the American Statistical Association* **86**, 96–107.

[12] Lindsey, J.K. & Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials, *Statistics in Medicine* **17**, 447–469.

[13] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92**, 162–170.

[14] Neuhaus, J.M., Hauck, W.W. & Kalbfleisch, J.D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models, *Biometrika* **79**, 755–762.

[15] Self, S. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association* **82**, 605–610.

[16] Waclawiw, M.A. & Liang, K.-Y. (1993). Prediction of random effects in the generalized linear model, *Journal of the American Statistical Association* **88**, 171–178.

[17] Wang, Z. & Louis, T. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function, *Biometrika 90* 765–775.

[18] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach (Corr: V45 p347), *Biometrics* **44**, 1049–1060.

CHARLES E. MCCULLOCH &
JOHN M. NEUHAUS

Generalized Linear Model

The term *generalized linear model* was first introduced in a landmark paper by Nelder & Wedderburn [23], in which a wide range of seemingly disparate problems of statistical modeling and inference (**analysis of variance** (ANOVA), **analysis of covariance** (ANCOVA), Gaussian, binomial, and **Poisson regression**, and so on) were cast in an elegant unifying framework. The flexibility and power of the generalized linear model are perhaps best illustrated by initial consideration of the simple Gaussian linear model (*see* **Linear Regression, Simple**). Let Y_i be a random **response variable** and let x_i denote an **explanatory variable**. In the Gaussian linear model, we assume that

$$Y_i = \beta_0 + \beta_1 x_i + \sigma E_i, \quad i = 1, 2, \ldots, n,$$

where $E_1, E_2, \ldots, E_n$ are errors that are independently and identically distributed standard **normal** random variables. An equivalent way of writing the model is as $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, n$, where $Y_1, Y_2, \ldots, Y_n$ are independently, but not identically, distributed and $\mu_i = \beta_0 + \beta_1 x_i$, $i = 1, 2, \ldots, n$. The objective of this model is to use the explanatory variable to characterize the variation in the mean of the response distribution across observational units, and hence to learn about the relationship between the explanatory and response variables.

Frequently, interest lies in the formulation of **regression** models for responses with other continuous or discrete distributions. In such settings, while the objective is typically to model the mean of the distribution, it must be modeled indirectly via the use of parameter transformations. In a generalized linear model this is done through the introduction of a link function $g(\cdot)$ and the model assumption, which, in the case of a single explanatory variable, takes the form

$$g(\mu_i) = \beta_0 + \beta_1 x_i.$$

The error distribution must also be generalized, usually in a way that complements the choice of link function. This leads to a very broad class of regression models.

This unifying theory has impacted the way such statistical methods are taught, has provided greater insight into the connections between various statistical procedures, and has led to considerable further research. McCullagh & Nelder [22] provide a comprehensive lucid account of the theory and applications which involve generalized linear models. Dobson [12] serves as an excellent introduction to the topic. In the following exposition, we draw primarily on the former.

The Generalized Linear Model (GLM)

The Exponential Family

Let Y be a random response variable of interest, and let y denote its corresponding realized value. The random variable Y has a distribution in the **exponential family** if

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi) + c(y, \phi)}\right\}, \qquad (1)$$

where $f_Y(y; \theta, \phi)$ is a probability density or mass function, if Y is a continuous or discrete random variable respectively, $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are known functions, and ϕ is known. The parameter θ is often referred to as the canonical parameter or natural parameter, and ϕ is called the dispersion parameter. When ϕ is unknown, it is considered a **nuisance parameter** and (1) may or may not be a member of the exponential family. For simplicity, we will initially assume that ϕ is known, and we will subsequently denote its specified value as ϕ_0.

The function $a(\phi_0)$ typically takes the form ϕ_0/w, where w is a fixed weight that may vary from observation to observation. The importance of the weight function arises in the analysis of grouped data, where it is often the group size. The function $b(\cdot)$ is termed the cumulant function, since it plays a central role in the determination of the cumulants (*see* **Characteristic Function**), and $c(\cdot, \cdot)$ is an arbitrary function of (y, ϕ_0), and possibly the weight w.

With $\phi = \phi_0$, the log likelihood arising from (1) takes the form

$$l(\theta; y) = \frac{\theta y - b(\theta)}{a(\phi_0)} + c(y, \phi_0). \qquad (2)$$

Since $\mathrm{E}\{\partial l/\partial\theta\} = 0$ it follows that $\mathrm{E}\{Y\} = \mu = \partial b(\theta)/\partial\theta = b'(\theta)$. Furthermore, since $\mathrm{E}\{\partial^2 l/\partial\theta^2\} + \mathrm{E}\{(\partial l/\partial\theta)^2\} = 0$, it follows that $\mathrm{var}\{Y\} = V = a(\phi_0)b''(\theta)$. To make the dependence of the variance V on μ explicit, we will sometimes write it as

$\text{var}\{Y\} = a(\phi_0)v(\mu)$, where $v(\cdot)$ is called the variance function.

Formulation of a Regression Model

Let $Y_1, Y_2, \ldots, Y_n$ be a set of random response variables in which Y_i, the response for the ith unit, has a distribution governed by

$$f_{Y_i}(y_i; \theta_i, \phi_0) = \exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a_i(\phi_0)} + c_i(y_i, \phi_0)\right\}. \tag{3}$$

In (3), $a_i(\cdot)$ and $c_i(\cdot, \cdot)$ are all assumed to have the same functional form across responses; the subscripts are introduced only to accommodate varying weights across responses (e.g. $a_i(\phi_0) = \phi_0/w_i$). The cumulant function is assumed to be common for all observations, and the subscript on θ accommodates different expectations for the different responses. The dispersion parameter is considered to be fixed and common.

We then let $y_1, y_2, \ldots, y_n$ represent a sample of size n arising from the corresponding n random variables, which we write more compactly in vector form as $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$. Let $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ denote a $p \times 1$ vector of explanatory variates arising from the same unit on which y_i is observed, with $x_{1i} = 1$, and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ be a $p \times 1$ vector of regression coefficients. For convenience we let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$ denote the $n \times p$ matrix of covariate vectors. The inner product $\mathbf{x}_i'\boldsymbol{\beta} = \eta_i$ consists of a linear combination of the regression parameters and hence is termed the linear predictor for the ith observation. Thus $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)'$ is an $n \times 1$ vector of linear predictors. The objective in forming a generalized linear model is to relate these linear predictors to the corresponding means, and to model the variation in the mean from one observation to the next, using the explanatory variates and the regression parameters. For estimability, interpretability, and to serve the central purpose of data reduction, $p < n$.

The link function is a monotonic differentiable function that typically maps the parameter space for the mean $\mu_i = b'(\theta_i)$ on to the real line. The role of the link function is to make explicit the nature of the relationship between the linear predictor and the mean. That is, we typically select a link function $g(\cdot)$ and let

$$g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

A wide variety of link functions are often suitable, but several criteria are available to guide their selection. First, while it is not essential, link functions that map on to the real line are generally preferred to avoid numeric difficulties in estimation that arise since the linear predictor is unconstrained. Secondly, certain link functions lead to regression parameters that have attractive properties in terms of their interpretation. Thirdly, for each member of the exponential family there exists a so-called canonical link function that leads to inference for $\boldsymbol{\beta}$ based solely on **sufficient statistics**. Canonical link functions arise by equating the canonical parameter to the linear predictor; substitution of $\theta_i = \eta_i$ into a **likelihood** based on (3) gives $\mathbf{X}'\mathbf{y}$ as a sufficient statistic for $\boldsymbol{\beta}$. Fourthly, using procedures for model **diagnostics** one can consider the **goodness of fit** of various link functions to guide the choice.

Examples

Before proceeding, it is perhaps worthwhile to consider some particular members of the exponential family with a view to gaining better insight into the current model formulation. In each of the following cases, we begin with a probability density or mass function in its usual form, rewrite it in the form of (1), identify the canonical and dispersion parameters as well as the functions $a(\cdot), b(\cdot)$, and $c(\cdot, \cdot)$, and finally consider possible link functions. For convenience, we drop the subscript i on the response variable and the vector of explanatory variables.

Example 1: the GLM for Poisson Responses

Suppose that Y is a **Poisson** random variable with mean $\mu > 0$. Then

$$\Pr(Y = y; \mu) = f_Y(y; \mu) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots .$$

This can be rewritten as

$$f_Y(y; \theta, \phi_0) = \exp[(\theta y - e^{\theta}) - \log(y!)],$$

where $\theta = \log \mu$, $a(\phi_0) = 1$, $b(\theta) = e^{\theta}$, and $c(y, \phi_0) = -\log(y!)$. The dispersion parameter ϕ_0 can in fact

be taken as one, and since there are no weights involved, we obtain the particularly simple form of $a(\cdot)$. Note that $\mathrm{E}\{Y\} = \mu = b'(\theta) = \mathrm{e}^{\theta}$ and $\mathrm{var}\{Y\} = a(\phi_0)b''(\theta) = v(\mu) = \mathrm{e}^{\theta}$. Thus $v(\mu)$ is an identity function and the mean equals the variance, as one would expect for a Poisson random variable.

The log link is the canonical link and is the standard link function for **Poisson regression**. It generates the familiar **loglinear model**

$$\log(\mu) = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Note that covariate effects that are additive on the log scale have multiplicative effects on the mean of the distribution. However, any link functions that map the positive real line on to the entire real line are also possible.

Example 2: the GLM for Binomial Responses

Suppose that $mY \sim \mathrm{bin}(m, \pi)$ with $\mathrm{E}\{Y\} = \mu = \pi$. Then

$$\begin{aligned}\Pr(Y = y; \pi) &= f_Y(y; \pi) \\ &= \binom{m}{my} \pi^{my}(1-\pi)^{m-my}, \\ & \qquad y = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1.\end{aligned}$$

This can be rewritten as

$$f_Y(y; \theta, \phi_0) = \exp\left[\frac{\theta y - \log(1+\mathrm{e}^{\theta})}{m^{-1}} + \log\binom{m}{my}\right],$$

where $\theta = \log(\pi/(1-\pi))$, $a(\phi_0) = 1/m$, $b(\theta) = \log(1+\mathrm{e}^{\theta})$, and

$$c(y, \phi_0) = \log\binom{m}{my}.$$

Again since the variance is a function solely of the mean, the dispersion parameter may be set to one. The function $a(\phi_0) = 1/m$ then takes the familiar form, with the number of Bernoulli trials forming the **binomial** sample serving as the weight. Note that $\mathrm{E}\{Y\} = \mu = b'(\theta) = \exp\{\theta\}/(1+\exp\{\theta\}) = \pi$ and $\mathrm{var}\{Y\} = a(\phi_0)b''(\theta) = a(\phi_0)v(\mu) = \exp\{\theta\}/(m(1+\exp\{\theta\})^2) = \pi(1-\pi)/m$, as one would expect.

The canonical link leads to the specification of a **logistic regression** model, which is perhaps the most widely used for binomial data. Reasons include those discussed earlier in the context of canonical links, the fact that the regression coefficients may be interpreted as log **odds ratios**, and the resulting attractive features for the regression analysis of retrospective data [22] (*see* **Retrospective Study**). We obtain

$$\log\left[\frac{\pi}{(1-\pi)}\right] = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

In particular contexts, other link functions are routinely adopted, the basic requirement being that they map the interval [0, 1] on to the entire real line. For example, in **dose–response** studies the probit link, given by $g(\pi) = \Phi^{-1}(\pi)$, where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal random variable, is commonly used [14]. In applications involving discrete time **survival data** the complementary log–log link ($g(\pi) = \log(-\log(\pi))$) has connections to **proportional hazards** models, and so is also frequently adopted [22] (*see* **Quantal Response Models**).

Example 3: the GLM for Gaussian Responses

Suppose that $Y \sim \mathrm{N}(\tau, \sigma^2)$ with σ^2 known, giving the usual Gaussian probability density function

$$f_Y(y; \tau, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left[\frac{-(y-\tau)^2}{2\sigma^2}\right], \quad -\infty < y < \infty.$$

This can be rewritten as

$$f_Y(y; \theta, \phi_0) = \exp\left\{\left(\frac{\theta y - \theta^2}{2}\right)\Big/\phi_0 - \frac{1}{2}\left[\frac{y^2}{\phi_0} + \log(2\pi\phi_0)\right]\right\},$$

where $\theta = \tau$, $\phi_0 = \sigma^2$, $b(\theta) = \theta^2/2$, $a(\phi_0) = \phi_0$, and $c(y, \phi_0) = -\frac{1}{2}$ $[y^2/\phi_0 + \log(2\pi\phi_0)]$. Here the mean and variance are functionally independent and hence the variance function can be taken as $v(\mu) = 1$. The weights are all taken to be unity.

Note that $\mathrm{E}\{Y\} = \mu = b'(\theta) = \theta = \tau$ and $\mathrm{var}\{Y\} = a(\phi_0)b''(\theta) = a(\phi_0)v(\mu) = \phi_0 = \sigma^2$.

The canonical link function is the identity link function, giving $g(\mu) = \mu = \eta$, or

$$\mu = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

which is the deterministic part of the usual **multiple linear regression** model.

Other examples of distributions within the exponential family include the **gamma** and **inverse Gaussian** distributions. Details regarding the formulation of related regression models are provided in McCullagh & Nelder [22].

Estimation

The Fisher Scoring Algorithm

We now consider maximum likelihood as a method for obtaining an estimate of the regression parameter vector $\boldsymbol{\beta}$. In the following derivations, it will be useful to bear in mind the role of the following mappings:

$$\theta \xrightarrow{b'(\cdot)} \mu \xrightarrow{g(\cdot)} \eta.$$

We now reintroduce the subscripts to distinguish between observations within the sample. Consider a log likelihood arising from n observations, each with a probability distribution governed by the density or mass function of the form given in (3) with specified ϕ_0. Letting $l_i(\theta_i; y_i) = \log[f_{Y_i}(y_i; \theta_i, \phi_0)]$, we obtain

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} l_i(\theta_i; y_i) = \left\{ \sum_{i=1}^{n} [\theta_i y_i - b(\theta_i)] \big/ a_i(\phi_0) + \sum_{i=1}^{n} c_i(y_i, \phi_0) \right\}.$$

Given a covariate vector and upon specification of a link function, a transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\beta}$ is defined indicating that the log likelihood can be written in terms of $\boldsymbol{\beta}$ as $l(\boldsymbol{\beta}; \mathbf{y})$. The **maximum likelihood** estimate is the solution to a system of p score equations of the form $\partial l / \partial \boldsymbol{\beta} = \mathbf{0}$. One strategy for solving this system is to employ the Fisher scoring methods, which can be thought of as a standard Newton–Raphson search in which the matrix of second derivatives is replaced by its expectation (*see* **Optimization and Nonlinear Equations**).

Specifically, note that by the chain rule the contribution to $U_j(\boldsymbol{\beta}) = \partial l / \partial \beta_j$ from the ith subject may be written as

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}, \tag{4}$$

where

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi_0)}, \qquad \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)}, \quad \text{and} \quad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Substituting into (4), we obtain

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - b'(\theta_i)}{a_i(\phi_0)} \frac{1}{b''(\theta_i)} \frac{\partial \mu_i}{\partial \beta_j} = x_{ij}.$$

Noting that $\mu_i = b'(\theta_i)$ and $V_i = a_i(\phi_0)b''(\theta_i)$, we can write

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}. \tag{5}$$

The complete score vector is given by $\mathbf{U}(\boldsymbol{\beta}) = (U_1(\boldsymbol{\beta}), \ldots, U_p(\boldsymbol{\beta}))'$.

The contribution from the ith subject to the (j, k)th element of the observed **information matrix** $\mathbf{I}(\boldsymbol{\beta})$ is given by $\partial^2 l_i / \partial \beta_j \partial \beta_k$ and may be written

$$[y_i - b'(\theta_i)] \frac{\partial}{\partial \beta_k} \left[\frac{1}{a_i(\phi_0)b''(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \right] + \left[\frac{1}{a_i(\phi_0)b''(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \right] \frac{\partial}{\partial \beta_k} [y_i - b'(\theta_i)]. \tag{6}$$

The use of the expected (Fisher) information obviates the need to derive an expression for the derivative in the first term, since $\mathrm{E}\{y_i - b'(\theta_i)\} = 0$. Therefore, if $[\mathcal{I}(\boldsymbol{\beta})] = \mathrm{E}\{[\mathbf{I}(\boldsymbol{\beta})]\}$ is the expected information matrix, the (j, k)th entry is given by

$$[\mathcal{I}(\boldsymbol{\beta})]_{jk} = \sum_{i=1}^{n} \frac{1}{a_i(\phi_0)b''(\theta_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 x_{ij} x_{ik}. \tag{7}$$

The Fisher scoring method involves iterating according to

$$\tilde{\boldsymbol{\beta}}^{(h)} = \tilde{\boldsymbol{\beta}}^{(h-1)} + [\mathcal{I}(\tilde{\boldsymbol{\beta}}^{(h-1)})]^{-1} \mathbf{U}(\tilde{\boldsymbol{\beta}}^{(h-1)}), \tag{8}$$

where $\tilde{\boldsymbol{\beta}}^{(h)}$ denotes the value of $\boldsymbol{\beta}$ after h steps. Typically, iteration is continued until the distance between successive estimates becomes less than some tolerance level.

Note that when the canonical link is used, $\partial\mu_i/\partial\eta_i = \partial b'(\theta_i)/\partial\theta_i = b''(\theta_i)$, indicating that the first term in (6) vanishes even for the observed information matrix. Therefore, since the second term in (6) is not random, for canonical links the Fisher scoring procedure coincides with a standard Newton–Raphson search.

Iteratively Reweighted Least Squares

The Fisher scoring method is sometimes referred to as "iteratively reweighted least squares" and it is of interest to consider the reason for this. Multiplying both sides of (8) by $\mathcal{I}(\tilde{\boldsymbol{\beta}}^{(h-1)})$ gives

$$[\mathcal{I}(\tilde{\boldsymbol{\beta}}^{(h-1)})]\tilde{\boldsymbol{\beta}}^{(h)} = [\mathcal{I}(\tilde{\boldsymbol{\beta}}^{(h-1)})]\tilde{\boldsymbol{\beta}}^{(h-1)} + \mathbf{U}(\tilde{\boldsymbol{\beta}}^{(h-1)}). \tag{9}$$

The expected information matrix given by (7) can be expressed as

$$\mathcal{I} = \mathbf{X}'\mathbf{W}\mathbf{X},$$

where $\mathbf{X}$ is the matrix of covariates as before, and $\mathbf{W}$ is an $n \times n$ diagonal matrix with (j, j)th entry

$$[\mathbf{W}]_{jj} = \frac{(\partial\mu_j/\partial\eta_j)^2}{V_j}. \tag{10}$$

Also note that the score vector $\mathbf{U}$ may be rewritten using this notation as $\mathbf{X}'\mathbf{W}\mathbf{D}$, where $\mathbf{D} = (D_1, \ldots, D_n)'$ with $D_j = (y_j - \mu_j)\partial\eta_j/\partial\mu_j$. Thus, if $\mathbf{W}^{(h-1)}$ is the matrix formed by (10) evaluated at $\tilde{\boldsymbol{\beta}}^{(h-1)}$, (9) may be rewritten as

$$\begin{aligned}\mathbf{X}'\mathbf{W}^{(h-1)}\mathbf{X}\tilde{\boldsymbol{\beta}}^{(h)} &= \mathbf{X}'\mathbf{W}^{(h-1)}\mathbf{X}\tilde{\boldsymbol{\beta}}^{(h-1)} \\ &\quad + \mathbf{X}'\mathbf{W}^{(h-1)}\mathbf{D}^{(h-1)} \\ &= \mathbf{X}'\mathbf{W}^{(h-1)}\mathbf{Z}^{(h-1)},\end{aligned}$$

where $\mathbf{D}^{(h-1)}$ again indicates evaluation at $\boldsymbol{\beta}^{(h-1)}$, and $\mathbf{Z}^{(h-1)} = (Z_1^{(h-1)}, \ldots, Z_n^{(h-1)})'$ with

$$Z_j^{(h-1)} = \mathbf{x}_j'\tilde{\boldsymbol{\beta}}^{(h-1)} + D_j^{(h-1)}.$$

Therefore the iterative step given by (8) can equivalently be written as

$$\mathbf{X}'\mathbf{W}^{(h-1)}\mathbf{X}\tilde{\boldsymbol{\beta}}^{(h)} = \mathbf{X}'\mathbf{W}^{(h-1)}\mathbf{Z}^{(h-1)}. \tag{11}$$

Because the expression above is reminiscent of the equations of estimation for weighted least squares, the resulting algorithm is sometimes referred to as "iteratively reweighted least squares".

Alternative Methods of Estimation

For hierarchical **loglinear models**, **iterative proportional fitting** is another approach for maximum likelihood estimation [10]. In this approach, the parameter estimates are iteratively modified until the fitted values for the sufficient statistics are as close as possible to the observed values. See Agresti [1] or Bishop et al. [3] for details.

Other methods such as minimum chi-square, modified minimum chi-square, and minimum discrimination algorithms [1] are feasible, but do not in general provide maximum likelihood estimates, and so are not commonly used (*see* **Ban Estimates**).

Assessing Model Fit

A central question in fitting generalized linear models relates to quality of fit. That is, it is important that the model adequately reflect the variation and trends in the data before much stock is placed in the parameter estimates and related inferences. Issues pertaining to model fit may be broadly classified as based on the assessment of a particular model, or the assessment of one model relative to another.

Omnibus Goodness-of-fit Statistics

A natural basis on which to judge the quality of fit of a model is how closely the model-based estimates for the expected values, or fitted values, approximate the observed data. These fitted values are typically computed as

$$\hat{\mu}_i = \mu_i(\hat{\boldsymbol{\beta}}) = g^{-1}(\hat{\eta}_i),$$

where $\hat{\eta}_i = x_i'\hat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$. To facilitate such an assessment it is instructive to conceptualize a so-called saturated model in which the observed data are reproduced exactly by the fitted values. Such a model can be obtained by allowing the number of parameters in the linear predictor to equal the number of observations and by judicious choice of the covariate vector. Note that while such a model

is attractive in that it reproduces the data, it does not serve the purpose of data reduction, and so is not particularly useful apart from as a benchmark for model assessment.

Let $\boldsymbol{\beta}^{\text{sat}}$ denote the $n \times 1$ vector of regression coefficients arising from the saturated model, and $\boldsymbol{\beta}^{\text{red}}$ denote the $p \times 1$ regression parameter of our reduced model ($p < n$). Again, we let $\hat{\boldsymbol{\beta}}^{\text{sat}}$ and $\hat{\boldsymbol{\beta}}^{\text{red}}$ represent the corresponding maximum likelihood estimates. Since the **likelihood ratio test** statistic

$$g_{\text{D}} = 2[l(\hat{\boldsymbol{\beta}}^{\text{sat}}) - l(\hat{\boldsymbol{\beta}}^{\text{red}})]$$

measures the "distance" or deviance between the saturated and reduced models on the likelihood metric, and because it plays an important role in what follows, it is referred to as the scaled deviance statistic". To be specific, we let G_{D} and g_{D} denote the random variable and realized value of the scaled deviance respectively. The (unscaled) deviance statistic simply refers to $\phi_0 g_{\text{D}}$.

An alternative intuitive omnibus measure for the quality of fit is the Pearson **chi-square statistic**, which takes the form

$$g_{\text{P}} = \sum_{i=1}^{n} \left[\frac{(y_i - \hat{\mu}_i)^2}{\hat{V}_i} \right] = \sum_{i=1}^{n} \left[\frac{(y_i - \hat{\mu}_i)^2}{a_i(\phi_0)v(\hat{\mu}_i)} \right].$$

A variety of other types of goodness of fit statistics are available. In fact, Cressie & Read [5] show that both the scaled deviance and Pearson chi-square statistics are members of a broader class of measures in the power divergence family. Here we restrict consideration to the scaled deviance and Pearson chi-square statistics.

For the Gaussian linear model with a known variance, g_{D} and g_{P} are the error sum of squares divided by the variance parameter, and so, subject to correct model specification are exactly chi-square distributed on $n - p$ degrees of freedom, In this context, values of g_{D} or g_{P} near $n - p$ are typically thought to represent an adequate fit of the model to the data, with large values suggesting that a substantial reduction in the quality of fit is incurred in using the reduced model vs. the saturated model. This judgment may be formalized and made somewhat more rigorous by exploiting the relevant **chi-square distribution** on $n - p$ **degrees of freedom**. That is, one can test with size α the **null hypothesis** of an adequate fit with the scaled deviance and claim an inadequate fit if $\Pr(T > g_{\text{D}}) < \alpha$, where T is a generic $\chi^2(n - p)$ random variable. Usually, however, in the Gaussian linear model, the dispersion parameter ϕ is unknown. It can be eliminated from the goodness-of-fit assessment by the use of the standard $\boldsymbol{F}$ **test** [12].

There is considerable debate over the role of the scaled deviance and Pearson statistics in the assessment of the overall fit of a non-Gaussian linear model. A key issue of debate is the adequacy of the approximation involved in adopting a chi-square reference distribution for these quantities. For concreteness and convenience, we illustrate the following points in the context of binomial data and subsequently make remarks specific to other distributions in the exponential family. Let the data consist of pairs $(y_i, \mathbf{x}_i)$, where $Y_i \sim$ bin $(m_i, \pi_i), i = 1, \ldots, n$ with $\log[\pi_i/(1 - \pi_i)] = \mathbf{x}_i'\boldsymbol{\beta}$. Then $\text{E}(Y_i) = \mu_i = m_i\pi_i$, which is estimated as $\hat{\mu}_i = m_i\hat{\pi}_i = m_i \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})/[1 + \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})]$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate. If interest lies in the asymptotic behavior of a function of the estimated means, it is important to recognize that there are two ways in which the total amount of data may increase. The chi-square approximation of the deviance statistic under the binomial distribution is based on the behavior as m_i (and hence μ_i) tends to infinity, $i = 1, \ldots, n$, with the total number of binomial samples, n, fixed. In this case G_{D} is approximately independent of $\hat{\pi}_i, i = 1, \ldots, n$, and hence is a reasonable measure of goodness of fit. In contrast, McCullagh & Nelder [22] demonstrate that, as $n \to \infty$, if $m_i\pi_i(1 - \pi_i)$ is bounded, not only is the chi-square approximation to the distribution of G_{D} poor, but G_{D} is not independent of $\hat{\boldsymbol{\beta}}$. This latter point suggests that larger values of G_{D} need not reflect poor fit, but could simply arise due to a particular value of $\hat{\boldsymbol{\beta}}$. If $\hat{\mu}_i$ or $m_i - \hat{\mu}_i, i = 1, \ldots, n$ are relatively small, then one is faced with so-called sparse data. The Pearson statistic shares the degeneracy properties of the deviance statistic arising with very sparse data, which were considered by McCullagh & Nelder [22]. Thus caution is warranted for the use of either measure in these settings. Further studies of the asymptotic behavior of G_{D} and G_{P} for multinomial data, reviewed by Cressie & Read [6], suggest that the same considerations apply in this context. For Poisson data, the sparse data setting arises when $\hat{\mu}_i = \exp(\hat{\eta}_i)$ are small for some $i, i = 1, \ldots, n$. Hence the asymptotic chi-square approximations for Poisson models improve as $\mu_i \to \infty$ for each $i, i = 1, \ldots, n$.

A second consideration relates to the interpretation and the particular type of lack of fit that these statistics are designed to detect. Pierce & Schafer [24] suggest that in general the deviance provides a more meaningful measure of lack of fit than the Pearson statistic, and that given the inadequacy of the chi-square approximation for the distribution of G_D, efforts to provide a better approximation to its distribution, or to develop modifications to the deviance that maintain its attractive features, would be worthwhile. In particular, computer-intensive exact methods can be utilized to enumerate the distribution of G_D or G_P under an assumed model [6]. Alternatively, McCullagh [20, 21] suggests that the effect of the estimates on the distribution of G_D and G_P may be addressed by relying on estimates of the conditional **moments** of a goodness-of-fit statistic. Thus, rather than relying on approximate chi-square distributions, moments of the distribution of G_D, conditional on the sufficient statistics for $\boldsymbol{\beta}$, should be derived. Having derived the conditional means and variances, a normal approximation can then be employed to compute approximate significance levels for testing goodness of fit. To date, this proposal has received relatively little attention. See Firth [16] and the article on **Goodness of Fit** for further details.

Model Selection

Since parsimony is a major consideration in model selection, given a model with q covariates, it is often desirable to test whether a model with fewer covariates (say, $p < q$) would suffice. Let $\boldsymbol{\beta}_q = (\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_q)'$ and $\boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p)'$ and suppose that interest lies in testing H_0: $\beta_{p+1} = \beta_{p+2} = \cdots = \beta_q = 0$. It is natural to carry out such a test by computing the corresponding likelihood ratio statistic $2[l(\hat{\boldsymbol{\beta}}_q) - l(\hat{\boldsymbol{\beta}}_p)]$, which is approximately chi-square distributed on $q - p$ degrees of freedom. Note that this may equivalently be thought of as the difference in the scaled deviance statistics for the model with q and p regression parameters.

For this reason one can think of tests of this sort as being directed at comparing the quality of fit of the reduced and fuller models and doing so based on the change in deviance. Specifically, we do this by examining whether the simpler model significantly reduces the quality of fit relative to the full model. If $G_D^{(q)}$ represents the deviance of a model with an $q \times 1$ vector $\boldsymbol{\beta}$, and if

$$\Delta G_D^{(q,p)} = G_D^{(q)} - G_D^{(p)},$$

then we reject $H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_q = 0$ with size α if

$$\Pr(T > \Delta g_D^{(q,p)}) < \alpha,$$

where T is a generic $\chi^2(q - p)$ random variable and $\Delta g_D^{(q,p)}$ represents the realized change in deviance.

We remark that while the asymptotic chi-square approximation for the distribution of the scaled deviance may be questionable in many practical situations, the validity of the approximation for the comparison of nested unsaturated models, as described here, is typically quite good.

Residual Analyses

Residual analyses provide another means by which to investigate issues pertaining to quality of fit. We consider three types of residuals here, and cite Pierce & Schafer [24] for further details. The first two types are closely related to the scaled deviance and Pearson statistics previously discussed.

Note that the scaled deviance may be written as

$$g_D = \sum_{i=1}^{n} 2[l_i(\hat{\boldsymbol{\beta}}^{\text{sat}}) - l_i(\hat{\boldsymbol{\beta}}^{\text{red}})] = \sum_{i=1}^{n} d_i,$$

where d_i is the contribution from the ith subject to the overall statistic. The (scaled) deviance residual is taken as the signed square root of d_i and is denoted

$$r_{D_i} = (-1)^{I(y_i - \hat{\mu}_i < 0)}(d_i)^{1/2},$$

where $I(A)$ is an indicator function (*see* **Dummy Variables**) taking the value 1 if A is true and zero otherwise. These residuals have a limiting (i.e. $\mu_i \to \infty$) standard normal distribution, but the approximation for modest sample sizes can be improved by adopting an adjusted deviance residual of the form

$$r_{AD_i} = r_{D_i} + \delta(\hat{\boldsymbol{\beta}}),$$

where $\delta(\hat{\boldsymbol{\beta}}) = E[(Y_i - \mu_i)^3/6\hat{V}_i^{3/2}]$ [24].

The second type of residual is the Pearson residual, which takes the form

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{[V_i(\hat{\mu}_i)]^{1/2}}.$$

It is the signed square root of the contribution to g_P from the ith observation, and also has an asymptotically standard normal distribution.

Another class of residuals may be formed by considering **transformations** to the data such that the distribution of the residuals, linear in the transformed data, is more closely represented by the standard normal distribution. These transformations are generally derived to reduce the skewness of residuals taking the form

$$r_{A_i} = \frac{T(y_i) - \hat{E}[T(Y_i)]}{\{\text{var}[T(Y_i)]\}^{1/2}}.$$

Wedderburn is credited [2] with showing that, for likelihoods in the exponential family under a generalized linear model, the appropriate transformation $T(\cdot)$ is given by

$$T(\cdot) = \int \frac{d\mu}{v^{1/3}(\mu)}.$$

The particular form of such residuals will clearly depend on the family of distributions, and **numerical integration** may be required to find the appropriate form of $T(\cdot)$. Having obtained, analytically or numerically, the form of $T(\cdot)$, the approximate mean and variance terms necessary to compute r_{A_i} may be approximated by Taylor series expansion.

For discrete distributions, Pierce & Schafer [24] recommend adopting continuity corrected versions of the above residuals by replacing y_i with $y_i - 1/2$ if $y_i > \mu_i$ or with $y_i + 1/2$ if $y_i < \mu_i$. A detailed study suggests that the adjusted deviance residual, with continuity correction applied as appropriate, appears to be the most attractive since its sampling distribution is well approximated by the standard normal distribution, at least for binomial and Poisson data [24].

For further definitions of residuals, and a discussion of useful diagnostic procedures based on residuals, see the articles on **Residuals** and **Diagnostics**.

Assessment of the Link Function

When several link functions are available for use, it is desirable to consider the implications of link function **misspecification**, and to explore which link function provides the best fit to the data. On the former point, there exists a general theory of misspecified models [32]. With regard to generalized linear models, Fahrmeier & Tutz [13] make pertinent comments, and further related references are contained in this text.

The objective of discriminating between several link functions can be facilitated by casting these link functions into a broader parametric family. Thus, having specified a distribution from the exponential family and a set of covariates, a model may be specified with a parametric link $g(\mu; \boldsymbol{\gamma}) = \eta$, where $\boldsymbol{\gamma}$ is an $r \times 1$ parameter vector indexing members of the family of link functions. Typically, $r = 2$ forms a sufficiently rich class that it may contain relevant links, and even this has been argued by some to require a great deal of data to enable identification of suitable links. Specific two-parameter families have been proposed for various members of the exponential family, but applications to binomial data have received the most attention [8, 27, 29]. Prentice [26] and Pregibon [25] present quite general approaches.

There are several options available in terms of analyses involving such families. Perhaps the most obvious is to maximize the enriched log likelihood $l(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$ with respect to all parameters. It may be desirable, however, to identify which of the more commonly used link functions are supported by the data. In this case, examination of the **profile likelihood** of $\boldsymbol{\gamma}$ (in which $\boldsymbol{\beta}$ is a nuisance parameter) provides insight into the most plausible values of $\boldsymbol{\gamma}$. If $\boldsymbol{\gamma}_{10}$ and $\boldsymbol{\gamma}_{20}$ are specific values of $\boldsymbol{\gamma}$ corresponding to two commonly used links, likelihood ratio tests can in principle be performed to help determine which is more appropriate. An alternative strategy is to fit a model with a specific link given by $\boldsymbol{\gamma}_0$, and to carry out a score test of the hypothesis H_0: $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ [27, 29] (*see* **Likelihood**).

Pregibon [25] describes the following convenient way in which to carry out approximate tests for the adequacy of a given link. If $g(\mu; \boldsymbol{\gamma}_0)$ is the specified link and $g(\mu; \boldsymbol{\gamma})$ is the correct link, note that by a first-order Taylor series expansion one can write

$$g(\mu; \boldsymbol{\gamma}_A) \simeq g(\mu; \boldsymbol{\gamma}_0) + \boldsymbol{\gamma}_*' \mathbf{D}(\mu; \boldsymbol{\gamma}_0),$$

where $\boldsymbol{\gamma}_* = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)$ and $\mathbf{D}(\mu; \boldsymbol{\gamma}) = \partial g(\mu; \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$. Given the fitted value $\hat{\mu}$ obtained under the specified link, one can then evaluate $\mathbf{D}(\hat{\mu}; \boldsymbol{\gamma}_0)$. The approach then involves thinking of $\mathbf{D}(\hat{\mu}; \boldsymbol{\gamma}_0)$ as a vector of supplementary covariates to be added to the linear predictor. Fitting a revised model with $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ and

$$\mathbf{x} = [x_1, x_2, \ldots, x_p, D_1(\hat{\mu}; \boldsymbol{\gamma}_0), \ldots, D_r(\hat{\mu}; \boldsymbol{\gamma}_0)]',$$

and testing the significance of $\boldsymbol{\gamma}_*$ via the change in deviance, is an approximate test of $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0$.

Pregibon [25] indicates that this method of testing may be thought of as equivalent to a test based on the first iteration of the method of Fisher scoring for the full likelihood approach with starting values given by $\hat{\boldsymbol{\beta}}_0$ (the MLE under the null model) and $\boldsymbol{\gamma}_0$.

Inference

Interval Estimation and Hypothesis Testing

Upon obtaining the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, it is natural to consider **hypothesis testing** and **interval estimation**. Here we discuss various options for interval estimation. Related procedures for hypothesis tests follow directly.

We consider the multiparameter case and, with only a slight loss of generality, suppose that interest lies in the first $r < p$ elements of $\boldsymbol{\beta}$. Full generality may be immediately achieved by reordering the elements of $\boldsymbol{\beta}$ such that the r parameters of interest are listed first, and by rearranging the information matrix conformably with this parameter vector. Let $\boldsymbol{\beta}_{(1)} = (\beta_1, \beta_2, \ldots, \beta_r)'$, $\boldsymbol{\beta}_{(2)} = (\beta_{r+1}, \beta_{r+2}, \ldots, \beta_p)'$, and hence $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. Let $\mathcal{I}_{(1,1)}$ and $\mathcal{I}_{(2,2)}$ denote the $r \times r$ and $(p-r) \times (p-r)$ submatrices of $\mathcal{I}$ corresponding to $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ respectively, and let $\mathcal{I}_{(1,2)}$ and $\mathcal{I}_{(2,1)}$ denote the upper and lower off-diagonal submatrices, respectively, giving

$$\mathcal{I} = \begin{bmatrix} \mathcal{I}_{(1,1)} & \mathcal{I}_{(1,2)} \\ \mathcal{I}_{(2,1)} & \mathcal{I}_{(2,2)} \end{bmatrix}.$$

Furthermore, let

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{(1,1)} & \boldsymbol{\Sigma}_{(1,2)} \\ \boldsymbol{\Sigma}_{(2,1)} & \boldsymbol{\Sigma}_{(2,2)} \end{bmatrix}$$

denote the inverse of the expected information matrix partitioned conformably with $\mathcal{I}$.

From standard likelihood theory, $(\hat{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{(1)})' [\boldsymbol{\Sigma}_{(1,1)}]^{-1} (\hat{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{(1)})$ is an approximate $\chi^2(r)$ pivotal quantity (*see* **Fiducial Probability**) if $\boldsymbol{\Sigma}_{(1,1)}$ is evaluated at $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$. Hence one can construct an approximate joint **confidence** region for $\boldsymbol{\beta}_{(1)}$ with simultaneous coverage probability $100(1-\alpha)\%$, by finding all values of $\boldsymbol{\beta}_{(1)}$ for which

$$(\hat{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{(1)})' [\boldsymbol{\Sigma}_{(1,1)}]^{-1} (\hat{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{(1)}) < \chi^2_{1-\alpha}(r), \tag{12}$$

where $\boldsymbol{\Sigma}_{(1,1)}$ is evaluated at $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$.

As an alternative, score-based confidence intervals may be derived as follows. Let $\hat{\boldsymbol{\beta}}_{(2|1)}$ denote the profile likelihood estimate of $\boldsymbol{\beta}_{(2)}$ for specified $\boldsymbol{\beta}_{(1)}$. Furthermore, let $\mathbf{U}_{(1)}(\boldsymbol{\beta}_{(1)}, \hat{\boldsymbol{\beta}}_{(2|1)})$ denote the first r elements of $\mathbf{U}$ evaluated at $\boldsymbol{\beta}_{(1)}$ and $\hat{\boldsymbol{\beta}}_{(2|1)}$. Then an approximate $100(1-\alpha)\%$ score based joint confidence region for $\boldsymbol{\beta}_{(1)}$ is constructed by finding all values of $\boldsymbol{\beta}_{(1)}$ such that

$$\mathbf{U}_{(1)}'(\boldsymbol{\beta}_{(1)}, \hat{\boldsymbol{\beta}}_{(2|1)}) [\boldsymbol{\Sigma}_{(1,1)} - \boldsymbol{\Sigma}_{(1,2)} \boldsymbol{\Sigma}_{(2,2)}^{-1} \boldsymbol{\Sigma}_{(2,1)}]^{-1} \times \mathbf{U}_{(1)}(\boldsymbol{\beta}_{(1)}, \hat{\boldsymbol{\beta}}_{(2|1)}) < \chi^2_{1-\alpha}(r), \tag{13}$$

where all elements of $\boldsymbol{\Sigma}$ are evaluated at $(\boldsymbol{\beta}_{(1)}, \hat{\boldsymbol{\beta}}_{(2|1)})$.

Finally, an approximate joint likelihood based confidence region with simultaneous coverage probability $100(1-\alpha)\%$ is given by all values of $\boldsymbol{\beta}_{(1)}$ for which

$$2[l(\hat{\boldsymbol{\beta}}_{(1)}, \hat{\boldsymbol{\beta}}_{(2)}), -l(\boldsymbol{\beta}_{(1)}, \hat{\boldsymbol{\beta}}_{(2|1)})] < \chi^2_{1-\alpha}(r). \tag{14}$$

Under fairly standard regularity conditions, the asymptotic approximations involved for Wald and score-based intervals are $O_p(n^{-1/2})$, while those for likelihood ratio-based intervals are $O(n^{-1})$. High-order improvements in the former are available via the use of corrections from Edgeworth expansions (*see* **Skewness**), but the performance of likelihood ratio-based intervals is generally quite good even for relatively small samples [22].

Let $\boldsymbol{\beta}_{(10)}$ denote a specific value of $\boldsymbol{\beta}_{(1)}$. Approximate tests of the hypothesis H_0: $\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(10)}$ can be carried out by evaluating the approximate pivotal quantities in (12)–(14) at the null value and computing the corresponding significance level. Specifically, Wald, score, and likelihood ratio based P values are derived by computing

$$P_W = \Pr(T > (\hat{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{(10)})' [\boldsymbol{\Sigma}_{(1,1)}]^{-1} \times (\hat{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{(10)})),$$

$$P_S = \Pr(T > \mathbf{U}_{(1)}'(\boldsymbol{\beta}_{(10)}, \hat{\boldsymbol{\beta}}_{(2|1)}) [\boldsymbol{\Sigma}_{(1,1)} - \boldsymbol{\Sigma}_{(1,2)} \times \boldsymbol{\Sigma}_{(2,2)}^{-1} \boldsymbol{\Sigma}_{(2|1)}]^{-1} \mathbf{U}_{(1)}(\boldsymbol{\beta}_{(10)}, \hat{\boldsymbol{\beta}}_{(2|1)})),$$

and

$$p_L = \Pr(T > 2[l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}_{(10)}, \hat{\boldsymbol{\beta}}_{(2|1)})]),$$

where $\hat{\boldsymbol{\beta}}_{(2|1)}$ now represents the maximum profile likelihood estimate of $\boldsymbol{\beta}_{(2)}$ at $\boldsymbol{\beta}_{(10)}$, and T is a generic $\chi^2(r)$ random variable.

The Dispersion Parameter

Implication of an Unknown Dispersion Parameter

Note that, thus far, we have assumed that the dispersion parameter ϕ was a known fixed parameter, and hence that the likelihood was a function only of the regression parameters in the linear predictor. As indicated earlier, this is entirely appropriate for binomial and Poisson responses with no **overdispersion** when the dispersion parameter is defined to be one. For other distributions in the exponential family, however, this parameter will also require estimation.

It would be natural to wonder how this might impact the methods of estimation and inference considered thus far. Fortunately, both the score vector and information matrix are scaled by ϕ and hence the Fisher scoring algorithm is unaffected by unknown ϕ; the maximum likelihood estimate for $\boldsymbol{\beta}$ arising from (8) is unaffected by unknown ϕ.

In terms of the interval estimates arising from (12)–(14), if an estimate $\tilde{\phi}$ is obtained, suitable adjustments are made by replacing ϕ_0 in $\mathcal{I}(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ with $\tilde{\phi}$. Asymptotically, the methods for interval estimation and testing previously described will maintain their frequency properties.

Estimation of the Dispersion Parameter

A convenient, consistent and approximately unbiased estimate of ϕ is obtained by noting that if

$$G_{\mathrm{P}} = \sum_{i=1}^{n} \left[\frac{(y_i - \hat{\mu}_i)^2}{\hat{V}_i} \right]$$

is approximately chi-square distributed on $n - p$ degrees of freedom, then $\mathrm{E}(G_{\mathrm{P}}) = n - p$. However if $V_i = \phi v(\mu_i)/w_i$, a moment-type estimate (*see* **Method of Moments**) can be obtained by substituting this expression into G_{P}, equating this to its expected value, and solving for ϕ to give

$$\tilde{\phi} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i) w_i^{-1} (n - p)}. \qquad (15)$$

Note, for example, that in linear regression models with Gaussian residuals, (15) gives $\tilde{\phi}$ as the usual sample variance estimate based on the **mean square error**.

Overdispersed Data

One might elect to estimate a dispersion parameter even when one is not part of the specified distribution function. Distributions for which the variance is functionally determined by the mean (and possibly a weight) are potentially too restrictive for some applications. The binomial and Poisson distributions are two such distributions in which, even after control of all appropriate covariates, deviance statistics, residual plots, and other diagnostic procedures may demonstrate a nonnegligible lack of fit. If this apparent lack of fit is not limited to a small fraction of exceptional observations but, rather, appears to be a general inadequacy of the model, it is common to describe this feature as **overdispersion**, meaning that there is greater variability in the data than that expected based on the model.

In such situations, it is common to generalize the standard variance function $v(\mu) = \mu(1 - \mu)$ for binomial data and $v(\mu) = \mu$ for Poisson data) by the introduction of a dispersion parameter. The variance function is extended in the usual way, taking the form $V = \phi v(\mu)/w$.

The revised expected information matrix becomes $\mathcal{I}/\phi$, leading to information-based large-sample variance estimates inflated by a factor ϕ. This in turn leads to tests that are appropriately more conservative and confidence intervals for the regression coefficients that are correspondingly wider. The dispersion parameter is estimated again as in (15).

Alternative strategies for accommodating overdispersion are increasingly common. Mixed models may be formulated in which a latent **random effect** may be thought of as mimicking the explanatory role of one or more missing covariates that would explain the outlying observations. Such mixed models have seen a great deal of application in clustered/longitudinal data where, conditional on the random effect, the responses follow a distribution from the exponential family. The resulting marginal likelihoods sometimes have a closed form (as in a Poisson regression model with canonical link and a random intercept, $\exp(\beta_1)$, following a **gamma distribution**), but generally do not. See Fahrmeier & Tutz [13] for a good discussion of modeling, estimation, and inference issues related to mixed generalized linear models.

Applications

Example 4: Beetle Mortality Data

Bliss [4] provides data from a toxicology study in which beetles were assigned to one of eight groups and subsequently exposed to a specified dose of carbon disulfide. The response of interest relates to survival over a five hour exposure period, and can be summarized at the group level simply as the fraction of the entire sample of beetles at that dose group that survived. The dose, x_i, is measured as the logarithm (base 10) of the concentration of carbon disulfide (mg/l). If m_i is the number of beetles assigned to the ith dose group, then we assume that Y_i has a binomial probability mass function as in Example 2. The data are summarized in Table 1.

The fitted values which arise from fitting regression models with three different link functions are also given in Table 1, the corresponding maximum likelihood estimates are given in Table 2, and the corresponding Pearson residual plots are given in Figure 1. The deviance residuals give a similar plot. Note that the fitted values and residual plots indicate considerable variation in the quality of fit for the different link functions. In particular, since the residuals are much closer to zero on average, it appears that the model with the complementary log–log link fits the data better than either the logistic or probit models (*see* **Quantal Response Models**). This statement is not contradicted by the scaled deviance and Pearson chi-square statistics, suggesting that there is no need to accommodate overdispersion. On this basis it is most reasonable to base inference on the complementary log–log link model.

We can therefore claim that there is a very highly significant dose–response effect. An approximate Wald-based 95% confidence interval for β_2 is (18.52, 25, 56).

Example 5: Cellular Differentiation Data

Consider an *in vitro* biomedical study with the objective of investigating the tendency for two agents [tumor necrosis factor (TNF) and interferon (INF)] to induce cellular differentiation, and their tendency to act in a synergistic manner. In a study reported by Trinchieri et al. [30], cells were grouped and received one of 16 combination doses of TNF and IFN according to a two-way **factorial design**. We take as the response of interest

Table 1 Data and fitted values for beetle mortality data

Observation	Dose, x_i	Number, m_i	Dead, y_i	Fitted value, $\hat{\mu}_i$		
				Logit	Probit	CLL
1	1.691	59	6	3.457	3.358	5.589
2	1.724	60	13	9.842	10.722	11.281
3	1.755	62	18	22.451	23.482	20.954
4	1.784	56	28	33.898	33.816	30.369
5	1.811	63	52	50.096	49.615	47.776
6	1.837	59	53	53.291	53.319	54.143
7	1.861	62	61	59.222	59.665	61.113
8	1.884	60	60	58.743	59.228	59.947

Source: Bliss [4].

Table 2 Maximum likelihood estimates for beetle mortality data

Parameter	Logit		Probit		CLL	
	$\hat{\beta}_j$	$\{[\Sigma(\hat{\boldsymbol{\beta}})]_{jj}\}^{1/2}$	$\hat{\beta}_j$	$\{[\Sigma(\hat{\boldsymbol{\beta}})]_{jj}\}^{1/2}$	$\hat{\beta}_j$	$\{[\Sigma(\hat{\boldsymbol{\beta}})]_{jj}\}^{1/2}$
Intercept, β_1	−60.72	5.174	−34.93	2.648	−39.57	3.237
Log (dose), β_2	34.27	2.908	19.73	1.487	22.04	1.797
Scaled deviance, g_D		11.232		10.120		3.446
Pearson's χ^2, g_P		10.005		9.514		3.292

Source: Bliss [4].

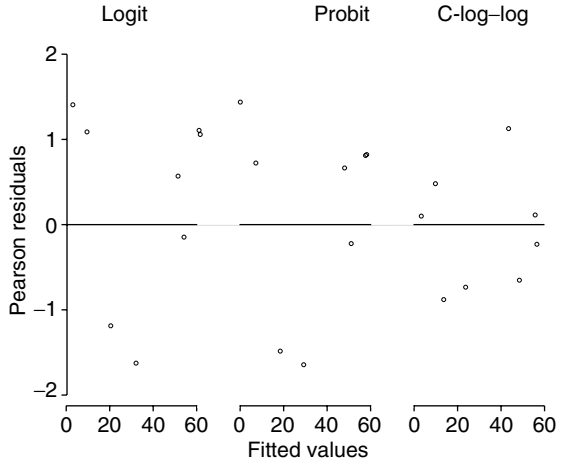

Figure 1 Residual plots for binary models with various link functions

Y_i, the number of cells showing evidence of cellular differentiation after exposure, and model this according to a Poisson process, as in Fahrmeier & Tutz [13].

Since a central question relates to the possible **synergism** of the two agents, we fit a Poisson regression model with the two main effects and a first-order interaction. The results are given in Table 3 and appear to indicate synergism. Examination of various types of residuals (plots not shown here) reveal serious lack of fit of this model to the data. Since no other covariates are available, we elect to introduce a dispersion parameter into the model. Using the moment estimate for ϕ given by (15), we obtain $\tilde{\phi} = 140.82/(16-4) = 11.735$. The revised large sample standard errors for the regression coefficients are then computed from the revised large sample covariance matrix $(11.735)^{1/2} \times [\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})]^{1/2}$ and are also given in Table 3. An approximate 95% confidence interval for β_4 is given by $(-1.472 \times 10^{-4}, 3.383 \times 10^{-5})$, and suggests that there remains little evidence of a synergistic effect after addressing the lack of fit of the Poisson model via the introduction of the additional dispersion parameter.

Quasi-likelihood

Wedderburn [31] used the term **quasi-likelihood** to describe objective functions that can be used to generate estimates of a linear regression model in a somewhat more general context than has been discussed thus far. The approach is based on the fact that the score vectors and information matrices arising from a generalized linear model from the exponential family rely solely on the first and second moments of the assumed distributions. This suggests more general models may be formulated subject to specification of mean and variance functions (and a dispersion parameter as appropriate).

To this end, let $y_1, y_2, \ldots, y_n$ denote a sample of observations, where associated with y_i is a $p \times 1$ vector of explanatory covariates $\mathbf{x}_i$. Let $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = V_i = \phi v(\mu_i)/w_i$ and $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$ as before, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters and $g(\cdot)$ is a link function. The generalization originates from the fact that $v(\mu_i)$ can be a somewhat arbitrary variance function and is not necessarily determined by a particular distribution in the exponential family.

Equations of the form

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi w_i^{-1} v(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j = 1, 2, \ldots, p,$$

may be solved to obtain a consistent estimate of $\boldsymbol{\beta}$. This solution may be aided as before, by a modified

Table 3 Maximum likelihood estimates for cellular differentiation data

Parameter	Poisson		Overdispersed Poisson	
	$\hat{\beta}_j$	$\{[\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})]_{jj}\}^{1/2}$	$\hat{\beta}_j$	$\{\tilde{\phi}[\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})]_{jj}\}^{1/2}$
Intercept, β_1	3.436	6.377E−2	3.436	2.184E−01
TNF (U/ml), β_2	1.553E−2	8.308E−4	1.553E−2	2.846E−3
IFN (U/ml), β_3	8.946E−3	9.668E−4	8.946E−3	3.312E−3
TNF∗IFN, β_4	−5.670E−5	1.348E−5	−5.670E−5	4.619E−5
Scaled deviance, g_D	142.39			
Pearson's χ^2, g_P	140.82			

Source: Trinchieri et al. [30].

Newton–Raphson procedure in which the variance of $\mathbf{U}(\boldsymbol{\beta})$ is estimated by the expected value of $\mathrm{E}\{\partial\mathbf{U}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\}$ which is a $p \times p$ matrix with (j,k)th element

$$\mathcal{I}_{jk}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial\mu_i/\partial\beta_j \partial\mu_i/\partial\beta_k}{\phi w_i^{-1} v(\mu_i)}.$$

Specifically,

$$\hat{\boldsymbol{\beta}}^{(h)} = \hat{\boldsymbol{\beta}}^{(h-1)} + [\mathcal{I}(\hat{\boldsymbol{\beta}}^{(h-1)})]^{-1}\mathbf{U}(\hat{\boldsymbol{\beta}}^{(h-1)}).$$

Again, by equating the Pearson statistic to its degrees of freedom, the dispersion parameter may be estimated as

$$\tilde{\phi} = \sum \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i) w_i^{-1}(n-p)}, \tag{16}$$

and the large sample covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by $[\mathcal{I}(\hat{\boldsymbol{\beta}})/\tilde{\phi}]^{-1} = \tilde{\phi}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$.

The term "quasi-likelihood" comes from the fact that $\mathbf{U}(\boldsymbol{\beta})$ behave in many respects like score vectors from a bona fide likelihood function from the exponential family, and thus may be thought of as quasi-score equations. Furthermore, one can construct a **quasi-likelihood function** $Q(\boldsymbol{\beta}, \phi; y)$ by integrating

$$Q(\boldsymbol{\beta}, \phi; y) = \int_y^{\mu} \frac{(y-t)}{\phi w_i^{-1} v(\mu)} \frac{\mathrm{d}\mu}{\mathrm{d}\beta}\, \mathrm{d}t.$$

This can in turn be used to define quasi-likelihood ratio statistics which behave similarly to genuine likelihood ratio statistics, although the distributional approximations are of lower order [19].

A variety of other quasi-likelihood approaches have been proposed for use when the variance function does not have the form $\mathrm{var}(Y_i) = \phi v(\mu)/w_i$. Scenarios for which this is not a viable assumption include most mixed models, autoregressive models (*see* **ARMA and ARIMA Models**), and very general **multivariate** responses. Liang & Zeger [18] adopted a quasi-likelihood approach for estimation in the context of longitudinal/clustered data and coined the term **generalized estimating equation**. In this context the variance functions are arranged in matrix form and, with arbitrary specification, the estimating equations yield consistent estimators for the regression parameters. Various estimates for the covariance parameters under particular formulations may be defined based on Pearson-type residuals.

Crowder [7], Firth [15], and Godambe & Thompson [17] discuss quadratic estimating equations in which improved efficiency of estimation can be achieved by specification of higher-order moments of the response (*see* **Estimating Functions**). Prentice & Zhao [28] motivate joint estimating equations for mean and covariance parameters from quadratic exponential models. For further comments, see Dean [9] and Zhao & Prentice [33]. For textbook treatments of the subject, refer to Fahrmeier & Tutz [13] and Diggle et al. [11].

Software

There are now numerous statistical **software** packages available for fitting generalized linear models. Fahrmeier & Tutz [13] contains an annotated appendix devoted to the discussion of software packages and so serves as a useful reference.

The package GLM is specialized for fitting generalized linear models and there is now a "glm" function in **S-PLUS**. SAS has various procedures available for fitting particular types of generalized linear model, and a generic procedure is currently under development. SPSS and BMDP also have routines that facilitate fitting many types of linear models from the exponential family, but no generic procedures are currently available.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

[3] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Multivariate Categorical Analysis*. Massachusetts Institute of Technology, Cambridge, Mass.

[4] Bliss, C.J. (1935). The calculation of the dosage-mortality curve, *Annals of Applied Biology* **22**, 134–167.

[5] Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

[6] Cressie, N. & Read, T.R.C. (1989). Pearson's χ^2 and the loglikelihood ratio statistics G^2: a comparative review, *International Statistical Review* **57**, 19–43.

[7] Crowder, M. (1987). On linear and quadratic estimating functions, *Biometrika* **74**, 591–597.

[8] Czado, C. (1994). Parametric link modification of both tails in binary regression, *Statistical Papers* **35**, 189–201.

[9] Dean, C.B. (1991). Estimating equations for mixed Poisson models, in *Estimating Functions*, V.P. Godambe, ed. Oxford Science Publications, Toronto.

[10] Deming, W.E. & Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics* **11**, 427–444.

[11] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

[12] Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*. Chapman & Hall, London.

[13] Fahrmeier, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, London.

[14] Finney, D.J. (1971). *Probit Analysis*. Cambridge University Press, London.

[15] Firth, D. (1987). On the efficiency of quasi-likelihood estimation, *Biometrika* **74**, 233–245.

[16] Firth, D. (1991). Generalized linear models, in *Statistical Theory and Modelling*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London.

[17] Godambe, V.P. & Thompson, M.E. (1989). An extension of quasi-likelihood estimation, *Journal of Statistical Planning and Inference* **22**, 137–152.

[18] Liang, K.-Y. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[19] McCullagh, P. (1983). Quasi-likelihood functions, *Annals of Statistics* **11**, 59–67.

[20] McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models, *International Statistical Review* **53**, 61–67.

[21] McCullagh, P. (1986). The conditional distribution of goodness of fit statistics for discrete data, *Journal of the American Statistical Association* **81**, 104–107.

[22] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[23] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

[24] Pierce, D.A. & Schafer, D.W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association* **81**, 977–986.

[25] Pregibon, D. (1980). Goodness of link tests for generalized linear models, *Applied Statistics* **29**, 15–24.

[26] Prentice, R.L. (1975). Discrimination among some parametric models, *Biometrika* **64**, 607–614.

[27] Prentice, R.L. (1976). A generalization of the Probit and Logit methods for dose response curves, *Biometrics* **32**, 761–768.

[28] Prentice, R.L. & Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics* **47**, 825–840.

[29] Stukel, T. (1988). Generalized logistic models, *Journal of the American Statistical Association* **83**, 426–431.

[30] Trinchieri, G., Kobayashi, M., Rosen, M., Loudon, R., Murphy, M. & Perussia, B. (1986). Tumor necrosis factor and lymphotoxin induce differentiation of human myeloid cell lines in synergy with immune interferon, *Journal of Experimental Medicine* **164**, 1206–1225.

[31] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**, 439–447.

[32] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1–25.

[33] Zhao, L.P. & Prentice, R. (1991). Use of a quadratic exponential model to generate estimating equations for means, variances, and covariances, in *Estimating Functions* V.P. Godambe, ed. Oxford Science Publications, Toronto.

(*See also* **General Linear Model**)

RICHARD J. COOK

Generalized Linear Models for Longitudinal Data

Generalized linear models [8] have unified **regression** analysis for discrete and continuous, *independent* responses. In **longitudinal** studies, however, we observe repeated observations on each of many independent persons. It is likely that repeated responses for the same person are **autocorrelated** with one another. This correlation must be taken into account to draw valid and efficient inferences about parameters of scientific interest.

With a single observation on each subject, the only available target of estimation is $E(Y)$, the marginal mean or cross-sectional average value among persons with the same value of x. For example, in a **cross-sectional study** of alcohol use, this might be the prevalence of reported use in the population sample under study. With repeated observations on each person, there are additional possible targets, including the conditional mean given past responses, or the conditional mean given underlying latent variables (*see* **Random Coefficient Repeated Measures Model**). Approaches to longitudinal data analysis can be distinguished by their target of estimation. Different targets also correspond to different assumptions about the source of autocorrelation.

It is possible to formulate linear models for correlated responses so that the interpretation of regression parameters is insensitive to the particular target or model for **correlation**. With nonlinear models such as **logistic regression**, distinct targets have regression parameters with distinct interpretations. This article contrasts three approaches to longitudinal data analysis: **marginal**, **random effects**, and **transition** models, each with its own target of estimation and implied autocorrelation structure.

To illustrate the ideas, we focus on data from a Johns Hopkins Prevention Research Center (PRC), randomized community trial [5] in which first grade youths in Baltimore City received in 1986 either a behavioral or reading intervention and were visited at least once a year to monitor their mental health development. We consider data on reported alcohol use and level of psychiatric distress in a subset of 692 youths who had complete data for the years 1991–1994. In 1991 the youths were between the ages of 10 and 14. A question we address is whether youths with higher levels of distress as measured by a 14-item questionnaire are more likely to self-report having ever used alcohol.

Psychiatric distress was measured by administering 14 items asking the youth whether or not he: worries a lot, is afraid a lot, has trouble sleeping, worries that bad things will happen, is sad, has nothing that makes him/her happy, is afraid to go outside, wants to hurt himself, worries parents will never come back, is tired all the time, and does not feel like eating. Each item is rated on a four-point scale with higher score indicating greater psychiatric distress. An age-standardized mean score was used as the predictor variable in this analysis. The response variable is a self-report of whether or not the youth had ever consumed alcohol. Note that in the absence of reporting errors, having reported use in one year would determine the outcome in subsequent years. But this is not the case here because of the substantial inconsistencies in the reported records.

Approaches to Modeling

In this section we consider model formulation and interpretation for the three kinds of generalized linear model: marginal, random effects, and transition models, each with a unique target of estimation. To illustrate the differences between the three approaches, we focus in each case on the problem of relating a **binary** response, Y, such as alcohol use, and a single explanatory variable, x, for example psychiatric distress score. The standard generalized linear model for this problem when the responses are mutually independent is the logistic linear regression model,

$$\text{logit } \Pr(Y = 1) = \beta_0 + \beta_1 x, \qquad (1)$$

where $\text{logit } \Pr(Y = 1) = \log[\Pr(Y = 1)/\Pr(Y = 0)]$ is also called the log **odds**.

For longitudinal data, repeated observations on the same subject are typically correlated. The three kinds of model discussed below differ in their target of estimation and in the way they introduce correlation structure, and this has implications for the correct interpretation of the regression parameter β_1.

Notation

We use Y_{ij} to denote the jth of n_i responses on the ith of m subjects. Each Y_{ij} is associated with a unique time t_{ij} at which the response is measured. Associated with each response Y_{ij} is a vector $\mathbf{x}'_{ij} = (x_{ij1}, \ldots, x_{ijp})$ of explanatory variables. We write $\mu_{ij} = \mathrm{E}(Y_{ij})$ and $V_{ij} = \mathrm{var}(Y_{ij})$.

A classical generalized linear model [8] assumes that the complete set of responses Y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ are mutually independent, with means and variances determined as follows:

$$h(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}, \text{ for some known } \textit{link function } h(\cdot);$$
$$V_{ij} = \phi V(\mu_{ij}), \quad \text{for some known } \textit{variance function } V(\cdot).$$

We call the above a *cross-sectional* generalized linear model, to distinguish it from a *longitudinal* generalized linear model, in which we retain this basic structure but relax the independence assumption.

Marginal Models

In a marginal model the target of estimation is the population-average or cross-sectional mean response, μ_{ij}. We model the relationship of this marginal mean and the explanatory variables $\mathbf{x}_{ij}$ separately from the within-subject correlation. Specifically, a marginal model makes the following assumptions:

$$h(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta},$$
$$V_{ij} = \phi V(\mu_{ij}),$$
$$\mathrm{corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha}), \quad \text{where } \rho(\cdot) \text{ is a known function.}$$

The first two of these assumptions are exactly the assumptions made in a cross-sectional generalized linear model. It follows that the marginal regression coefficients, $\boldsymbol{\beta}$, have the same interpretation as coefficients from a cross-sectional analysis.

In the PRC study, we express the log odds of alcohol use as a function of age and psychiatric distress score. From (1) we see that the regression parameter β for distress score represents the change in the log odds of reporting having ever used alcohol per unit increase in the explanatory variable x. By construction, this change is averaged over the whole population. A marginal regression model does not address questions concerning heterogeneity between subjects. Nor, in the longitudinal setting, does it address questions concerning the possible effect of a subject's previous responses on their current response.

Random Effects Models

In a random effects generalized linear model, the target of estimation is the mean of Y_{ij}, conditionally on the values of unobserved (latent) **random variables**, $\mathbf{U}_i$, specific to person i. For example, when the outcome is a binary indicator of alcohol use, the latent variable might represent the youth's predisposition to use and/or report the use of alcohol. Specifically, for each i let $\mathbf{U}_i$ denote a vector of random variables of dimension q, representing the ith subject, and let $\mathbf{d}_{ij}$ denote an associated vector of q explanatory variables. The $\mathbf{U}_i$ are assumed to be mutually independent with a common underlying **multivariate distribution**, usually multivariate Gaussian (*see* **Multivariate Normal Distribution**), and the assumptions of the cross-sectional generalized linear model are modified to

$$h(\mu^{\mathrm{c}}_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}^* + \mathbf{d}'_{ij}\mathbf{U}_i,$$
$$V^{\mathrm{c}}_{ij} = \phi V(\mu^{\mathrm{c}}_{ij}),$$

where μ^{c}_{ij} and V^{c}_{ij} denote the conditional mean and variance of Y_{ij}, given $\mathbf{U}_i$. We use $\boldsymbol{\beta}^*$ rather than $\boldsymbol{\beta}$ to emphasize that the substantive meaning of the regression parameter is different from that of $\boldsymbol{\beta}$ in a marginal model. The random vector $\mathbf{U}_i$ represents a set of unobserved, or latent, characteristics of the ith subject which influence the mean response; for example, if $\mathbf{d}'_{ij} = (1, t_{ij})$, then the elements of $\mathbf{U}_i$ correspond to the intercept and slope of a subject-specific time trend in the mean response.

In the specific case of a simple logistic regression and scalar Gaussian U_i, the random effects GLM reduces to

$$\text{logit } \Pr(Y_{ij} = 1|U_i) = \beta^*_0 + \beta^*_1 x_{ij} + \alpha U_i, \qquad (2)$$

where $U_i \sim \mathrm{N}(0,1)$. Note that the restriction to a **standard normal** distribution for U_i implies no loss of generality, as any other mean and/or variance of U_i could be absorbed into the model parameters β^*_0 and α. In (2), the regression parameter β^*_1 again represents a change in the log odds per unit change in x, but this is now conditional on the subject's own value of

U_i. It is instructive to derive the marginal properties of the random effects model (2). This requires us to integrate out the dependence on the unobserved $\mathbf{U}_i$. For example, the unconditional mean response is

$$\Pr(Y_{ij}=1) = \int \Pr(Y_{ij}=1|u) f(u)\ \mathrm{d}u$$
$$= \int \frac{\exp(\beta_0^* + \beta_1^* x_{ij} + \alpha u)}{1+\exp(\beta_0^* + \beta_1^* x_{ij} + \alpha u)} f(u)\ \mathrm{d}u,$$

where $f(\cdot)$ is the standard Gaussian density function. This integral is not easily expressible in closed form, but a good approximation is available. Zeger et al. [15] show that for the model (2),

$$\text{logit}\ P(Y_{ij}=1) \approx (c^2\alpha^2+1)^{-1/2}(\beta_0^* + \beta_1^* x_{ij}), \tag{3}$$

where $c = 16(3)^{1/2}/(15\pi)$, from which it follows that

$$\boldsymbol{\beta} \approx (c^2\alpha^2+1)^{-1/2}\boldsymbol{\beta}^*, \tag{4}$$

where $c^2 \approx 0.346$. This **shrinkage** effect is also easily demonstrated by **simulation**. Figure 1 illustrates a simulation of the model (2) when $\beta_0^* = -1$, $\beta_1^* = 1$, and $\alpha = 1.5$. The solid lines show $\Pr(Y_{ij}=1|U_i)$ as functions of x for each of 25 subjects whilst the dotted line shows $\Pr(Y_{ij}=1)$, calculated as the average of all 25 subject-specific functions. The dotted line, which is in effect what we would be estimating in a marginal model, is very well approximated by a linear logistic, but with the regression parameter β_1 substantially smaller than β_1^*, as predicted by (4). Note that α is a measure of the degree of heterogeneity between subjects, because the subject-specific intercepts are $\beta_0 + \alpha U_i; i = 1, \ldots, m$, and the U_i have a standard Gaussian distribution.

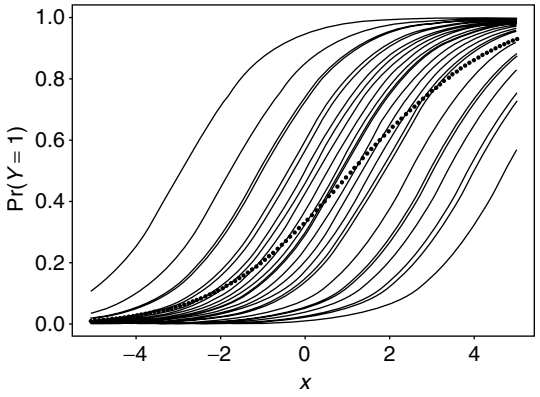

Figure 1 Simulation of the probability of a positive response in a random intercept model logit $\Pr(Y_{ij}=1|U_i) = -1.0 + x_{ij} + 1.5U_i$, where U_i is a standard normal random variable. The dotted line is the average over all 25 subjects

Incidentally, if we replace the logit link in (3) by the probit we can derive an exact expression for the shrinkage of the regression parameter. We now have

$$\Pr(Y_{ij}=1) = \int \Phi(\beta_0^* + U_i + \beta_1^* x_{ij}) f(u)\ \mathrm{d}u, \tag{5}$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function. Using the threshold interpretation of the probit model and the property that the sum of two Gaussian random variables is itself Gaussian, (5) becomes

$$\Pr(Y_{ij}=1) = \Phi[(1+\alpha^2)^{-1/2}(\beta_0^* + \beta_1^* x_{ij})]. \tag{6}$$

In particular, the marginal regression parameter in this random effects model is $\beta_1 = (1+\alpha^2)^{-1/2}\beta_1^*$.

Transition Models

In a transition GLM, the target of estimation is the conditional mean at a fixed time given the history of responses to that point. Hence we model the mean and variance of Y_{ij}, conditionally on past responses $Y_{i,j-k}$, for $k \geq 1$. For example, we might replace the assumptions of a cross-sectional GLM by

$$h(\mu_{ij}^{\mathrm{t}}) = \mathbf{x}_{ij}'\boldsymbol{\beta}^{**} + \sum_{k=1}^{r} \alpha_k y_{i,j-k},$$
$$V_{ij}^{\mathrm{t}} = \phi V(\mu_{ij}^{\mathrm{t}}),$$

where now μ_{ij}^{t} and V_{ij}^{t} are the expectation and variance of Y_{ij} conditional on all $Y_{i,j-k}$ for $k \geq 1$, and the notation $\boldsymbol{\beta}^{**}$ emphasizes that the regression parameters again differ in their substantive meaning from the regression parameters in the analogous marginal or random effects models. The integer r is called the *order* of the model. Note that, strictly, the above assumptions do not determine the joint distribution of $Y_{i1}, \ldots, Y_{in}$, but only the conditional distribution of $Y_{i,r+1}, \ldots, Y_{in}$ given $Y_{i1}, \ldots, Y_{ir}$. This reduces the practical usefulness of transition models when n, the number of observations per subject, is small.

A simple example of a transition logistic regression GLM is the first-order model in which

$$\text{logit } \Pr(Y_{ij} = 1 | Y_{i,j-k} : k \geq 1) = \beta_0^{**} + \beta_1^{**} x_{ij} + \alpha Y_{i,j-1}. \quad (7)$$

The transition model (7) is superficially similar to the random effects model (2), in the sense that both include a stochastic term in the linear predictor for the conditional mean response. However, the effect of the stochastic term is somewhat different for two reasons: first, the "random intercept" $\beta_0^{**} + \alpha Y_{i,j-1}$ takes one of only two possible values according to whether $Y_{i,j-1} = 0$ or 1; secondly, and perhaps more importantly, the random intercept for a given subject changes over time and has a reinforcing effect over time because for $\alpha > 0$, a realized value of $Y_{ij} = 1$ increases the conditional probability that $Y_{i,j+1}$ will also equal 1. Note also that, in contrast to the analogous random effects model (2), the physical meaning of the parameter α depends on the time separation between $t_{i,j-1}$ and t_{ij}. The model as defined therefore makes no sense either for data collected at irregularly spaced times, or for data in which the set of measurement times is not common to all subjects.

Simulation again provides a convenient way to illustrate the kind of relationships that can arise between the transition regression parameter β_1^{**} in (7) and the corresponding marginal regression parameter. Figure 2 shows a simulation of the model (7) in which $\beta_0^{**} = -1, \beta_1^{**} = 1, \alpha = 1.5$, and measurements are taken on each of 1100 subjects at times $t_j = j - 6; j = 1, \ldots, 11$. Part (a) uses $x_{ij} = t_j$ for each subject to represent an increasing time trend in the probability of a positive response, whereas part (b) uses $x_{ij} = x_i$ to correspond to a time-independent explanatory variable, and with 100 of the 1100 subjects assigned to each of 11 equally spaced values of x_i to span the range -5 to 5. The solid lines show the two conditional probabilities $\Pr(Y_{ij} = 1)$ given $Y_{i,j-1} = 0$ and given $Y_{i,j-1} = 1$, each as a function of x. The dots show the observed proportions of positive responses amongst the 100 responses associated with each of the 11 values of x. In both cases this marginal proportion increases more rapidly with x than do either of the two conditional proportions, that is, the marginal regression parameter of (7) is $\beta_1 > \beta_1^{**}$. In these simulations we generated initial observations Y_{i0} at time $t_0 = -6$ by sampling from

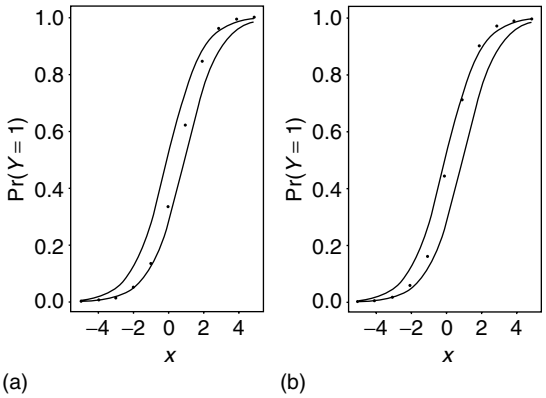

Figure 2 Simulation of the probability of a positive response in a transition logistic regression model $\text{logit } \Pr(Y_{ij} = 1 | Y_{i,j-k} : k \geq 1) = -1.0 + x_{ij} + 1.5 Y_{i,j-1}$, $j = 1, \ldots, 11, i = 1, \ldots, 1100$. Part (a) $x_{ij} = j - 6$ for each subject; Part (b) uses $x_{ij} = x_i$, to correspond to a time-independent explanatory variable with 100 of the 1100 subjects assigned to each of 11 equally spaced values of x_i to span the range -5 to 5. The solid lines show the two conditional probabilities $\Pr(Y_{ij} = 1)$ given $Y_{i,j-1} = 0$ and given $Y_{i,j-1} = 1$, each as a function of x. The dots show the observed marginal proportions of positive responses amongst the 100 responses associated with each of the 11 values of x

independent Bernoulli distributions with

$$\text{logit } \Pr(Y_{i0} = 1) = \beta_0^{**} + \beta_1^{**}[(x_{i1} - (x_{i2} - x_{i1})] + 0.5\alpha,$$

which is equivalent to extrapolating the time trend in the explanatory variable for each subject and taking a notional value of 0.5 for a fictitious observation at time $t_{-1} = -7$.

Inference

With random effects and transitional extensions of the GLM, it is possible to estimate unknown parameters using traditional **maximum likelihood** methods.

For random effects models, the likelihood of the data, expressed as a function of the unknown parameters, is given by

$$L(\boldsymbol{\beta}^*, \boldsymbol{\alpha}; \mathbf{y}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} | \mathbf{u}) f(\mathbf{u}; \boldsymbol{\alpha}) \, \mathrm{d}\mathbf{u}, \quad (8)$$

where $\boldsymbol{\alpha}$ represents the parameters of the random effects distribution. The likelihood is the integral over the unobserved random effects of the joint distribution of the data and the random effects. Except in the special case of a Gaussian linear model, **numerical integration** techniques are usually necessary to evaluate the likelihood (8).

Transition models can also be fitted using a version of maximum likelihood. The joint distribution of the responses $Y_{i1}, \ldots, Y_{in_i}$ can be written in the form

$$\begin{aligned} f(y_{i1}, \ldots, y_{in_i}) = {} & f(y_{in_i}|y_{i,n_i-1}, \ldots, y_{i1}) \\ & \times f(y_{i,n_i-1}|y_{i,n_i-2}, \ldots, y_{i1}) \ldots \\ & f(y_{i2}|y_{i1}) f(y_{i1}). \end{aligned} \tag{9}$$

In a first-order Markov model (*see* **Markov Chains**)

$$f(y_{ij}|y_{i,j-1}, \ldots, y_{i1}; \boldsymbol{\beta}^{**}, \boldsymbol{\alpha}) = f(y_{ij}|y_{i,j-1}; \boldsymbol{\beta}^{**}, \boldsymbol{\alpha})$$

so the **likelihood** contribution from person i simplifies to

$$\begin{aligned} & f(y_{i1}, \ldots, y_{in_i}; \boldsymbol{\beta}^{**}, \boldsymbol{\alpha}) \\ & \quad = f(y_{i1}; \boldsymbol{\beta}^{**}, \boldsymbol{\alpha}) \prod_{j=2}^{n_i} f(y_{ij}|y_{i,j-1}; \boldsymbol{\beta}^{**}, \boldsymbol{\alpha}). \end{aligned} \tag{10}$$

One difficulty that arises with (10) is that the marginal distribution of Y_{i1} often cannot be determined from the conditional distributions $f(y_{ij}|y_{i,j-1})$ without additional assumptions. A simple alternative is then to maximize the *conditional likelihood* of $Y_{i2}, \ldots, Y_{in_i}$ given Y_{i1}, which is obtained by omitting $f(y_{i1})$ from the equation above (*see* **Conditionality Principle**). Conditional maximum likelihood estimates can be found using standard GLM **software**, treating functions of the previous responses as explanatory variables. Inferences conditional on Y_{i1} are less efficient than maximum likelihood estimators but are all that is available without additional assumptions about $f(Y_{i1})$. The need to condition on the initial response from each subject makes it clear why these models are of limited value for short series, and the problem is exacerbated for transition models of higher order.

In the marginal model described above, we need only specify the first two **moments** of the responses for each person. With Gaussian data, the first two moments fully determine the likelihood, but this is not the case for GLM models in general. Hence, to use likelihood-based inference, additional assumptions about higher order moments must also be made. Examples for binary data are given by Prentice & Zhao [10], Fitzmaurice & Laird [2], and Liang et al. [7].

Even if additional assumptions are made, the likelihood is often intractable and involves many nuisance parameters in addition to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which must be estimated. For this reason, in applications for which the marginal regression parameters address the questions of primary scientific inference, a better approach may be to use **generalized estimating equations** or GEE. This is a multivariate analog of **quasi-likelihood**, with the same feature that it leads to **consistent** inferences about mean responses without requiring specific assumptions to be made about second and higher-order moments. Here, we give only a brief outline. For more detailed accounts, see Liang & Zeger [6], Zeger & Liang [14] and Prentice [9].

In the absence of a convenient likelihood function, the GEE method estimates $\boldsymbol{\beta}$ by solving a multivariate analog of the quasi-score function [12]:

$$\mathbf{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \mathrm{var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{11}$$

In the multivariate case there is the additional complication that $\mathbf{S}_\beta$ depends on $\boldsymbol{\alpha}$ as well as on $\boldsymbol{\beta}$ since $\mathrm{var}(\mathbf{Y}_i) = \mathrm{var}(\mathbf{Y}_i; \boldsymbol{\beta}, \boldsymbol{\alpha})$. This can be overcome by replacing $\boldsymbol{\alpha}$ in the equation above by an $m^{1/2}$-consistent estimate, $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$. Liang et al. [7] and Gourieroux et al. [3] show that the solution of the resulting equation is asymptotically as **efficient** as if $\boldsymbol{\alpha}$ were known.

The correlation parameters $\boldsymbol{\alpha}$ may be estimated in a similar fashion by simultaneously solving $\mathbf{S}_\beta = \mathbf{0}$ and

$$\mathbf{S}_\alpha(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\alpha}} \right)' \mathbf{H}_i^{-1} (\mathbf{W}_i - \boldsymbol{\eta}_i) = \mathbf{0}, \tag{12}$$

where $\mathbf{W}_i = (Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i,n_i-1}Y_{in_i}, Y_{i1}^2, Y_{i2}^2, \ldots, Y_{in_i}^2)'$, the set of all products of pairs of responses and squared responses, and $\boldsymbol{\eta}_i = \mathrm{E}(\mathbf{W}_i; \boldsymbol{\beta}, \boldsymbol{\alpha})$ [9].

The choice of the weight matrices, $\mathbf{H}_i$, in (12) will affect the efficiency of the resulting estimators, and good choices are problem-dependent. When the

parameters $\boldsymbol{\alpha}$ are not of direct interest, we can use simple models for the within-subject correlation leading to a *working variance matrix*, $\mathbf{W}_i$. Substituting this for var(Y_i), gives the estimating equations

$$\mathbf{S}_\beta(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0. \quad (13)$$

The solution, $\hat{\boldsymbol{\beta}}$, of (14) is asymptotically Gaussian [6], with variance consistently estimated by

$$\left(\sum_{i=1}^{m} \mathbf{D}_i' \mathbf{W}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^{m} \mathbf{D}_i' \mathbf{W}_i^{-1} \mathbf{V}_{0i} \mathbf{W}_i^{-1} \mathbf{D}_i \right) \times \left(\sum_{i=1}^{m} \mathbf{D}_i' \mathbf{W}_i^{-1} \mathbf{D}_i \right)^{-1} \quad (14)$$

evaluated at $\hat{\boldsymbol{\beta}}$, where

$$\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

and

$$\mathbf{V}_{0i} = (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'.$$

This *empirical* variance estimate [4, 13] is consistent as the number of individuals contributing to each element of the matrix goes to infinity. For example, in an **analysis of variance** problem, the number of units in each treatment group must get large.

GEE estimators enjoy two properties. First, $\hat{\boldsymbol{\beta}}$ is nearly efficient relative to the maximum likelihood estimates of $\boldsymbol{\beta}$ in many practical situations provided that the working variance matrices, $\mathbf{W}_i$, are reasonable approximations to var($\mathbf{Y}_i$) (e.g. [6] and [7]). In fact, GEE is equivalent to maximum likelihood for multivariate Gaussian data and for binary data from a **loglinear model** when var($\mathbf{Y}_i$) is correctly specified [2]. Secondly, $\hat{\boldsymbol{\beta}}$ is consistent as $m \to \infty$, even if the covariance structure of $\mathbf{Y}_i$ is incorrectly specified. When marginal regression coefficients are the scientific focus, it may be reasonable to sacrifice a small amount of efficiency in return for robustness against possible misspecification of the second and higher moment structure.

The robustness of the inferences about $\boldsymbol{\beta}$ can be checked in particular applications by fitting a final model using different covariance assumptions and comparing the two sets of estimates and their robust standard errors. If these differ substantially, then a more careful treatment of the covariance model may be necessary [1].

Example

To illustrate the main ideas above, we use each of the three approaches to describe a different aspect of the dependence of alcohol use on the reported level of psychiatric distress. In the marginal approach, the target of estimation is the prevalence (or log odds) of reporting ever having used alcohol and its dependence on psychiatric distress score and time. In a marginal analysis, we must also specify a model for the association among the four repeated responses for each youth. After preliminary analyses, we assume that the association, measured by the log odds ratio for Y_{ij} and Y_{ik}, is a linear function of the difference between and the mean of the two observation times. Hence, the model specification is

$$\text{logit } \Pr(Y_{ij} = 1) = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij},$$
$$\log \text{OR}(Y_{ij}, Y_{ik}) = \alpha_0 + \alpha_1 |t_{ij} - t_{ik}| + \frac{\alpha_2 (t_{ij} + t_{ik})}{2},$$

where x_{ij} is the psychiatric distress score for youth i at visit j.

The parameter estimates, model-based and empirical standard errors, obtained using GEE, are shown in Table 1. There is evidence of increasing prevalence of reporting ever having used alcohol in older children and significantly higher reported use among children with higher distress. The population odds of use is estimated to be 19% [= exp(5 × 0.0345) − 1] higher

Table 1 Results from marginal logistic regression analysis of alcohol use data

Variable	Estimate	Standard errors	
		Model-based	Empirical
Mean model			
Intercept (β_0)	−0.243	0.0702	0.0701
Time (β_1)	0.228	0.0293	0.0294
MH score (β_2)	0.0345	0.00723	0.00707
Association model			
Intercept (α_0)	1.66		0.328
Time lag (α_1)	−0.225		0.0823
Average time (α_2)	0.118		0.111

in one subgroup with psychiatric distress score that is five points higher than another (95% CI: 11%, 28%). The odds ratio between repeated observations on a subject decreases with increasing lag from 4.3 for observations one year apart to 2.8 at three years apart. At a fixed lag, the estimated odds ratio increases toward the later years of the study, but the evidence for this increase is weak.

A random effects model was fit to these same data to estimate a youth's predisposition of reporting ever having used alcohol and to examine the dependence of this risk on psychiatric distress. The approximate likelihood approach of Stiratelli et al. [11] was used for estimation. The results for the following random intercept logistic model are shown in Table 2:

$$\text{logit } \Pr(Y_{ij} = 1|U_i) = \beta_0^* + \beta_1^* t_{ij} + \beta_2^* x_{ij} + \alpha U_i.$$

In this model it is assumed that the random effects U_i are an independent sample from a Gaussian distribution with mean 0 and variance 1 and that given U_i, the repeated binary responses for youth i are independent of one another. Given the cumulative nature of the outcome, this assumption is unlikely to be valid.

Using the random intercept model, we estimate that a youth's odds of reporting ever having used alcohol increases by 27% $[= \exp(5 \times 0.0473) - 1]$ if his distress score increases by five points (95% CI: 16%, 38%). The standard deviation, α, of the random intercept is estimated to be 1.64 so that roughly 95% of youths would have log odds of reporting ever having used alcohol within $\pm$ 3.3 of the mean value. Note this is an extreme level of heterogeneity, indicating strong association among repeated observations on each youth. This also reflects, if less directly, the consistency of reports across time by a youth.

The coefficients for the random effects model are larger than those from the marginal model owing to the attenuation that results from averaging personal risks of reported use to obtain prevalences. Note that the degree of attenuation is similar for all coefficients and is roughly equal to the value given in (3).

Finally, the following transition model can be estimated from these data:

$$\text{logit } \Pr(Y_{ij} = 1|Y_{i,j-1}) = \beta_0^{**} + \beta_1^{**} t_{ij} + \beta_2^{**} x_{ij} + \alpha_1 Y_{i,j-1} + \alpha_2 Y_{i,j-1} x_{ij}.$$

Note that we have allowed the log odds of reporting ever having used alcohol in one year to depend on whether or not the youth reported use in the prior year and have included an interaction between psychiatric distress score and prior report. In this way we estimate a separate effect of psychiatric distress on a new report of alcohol use given no reported use at the prior visit β_2^{**}, and on the confirmation of a prior report of ever using alcohol, $\beta_2^{**} + \alpha_2$. The results are shown in Table 3.

The strongest predictor of reported use at a given visit is having reported use at the prior visit with an odds ratio of 5.5 (95% CI: 4.5, 6.7). This indicates a reasonably high degree of consistency in repeated assessments of ever use. Psychiatric distress score is positively associated with the outcome among those with no prior reported use, the odds being 25% higher for children whose score is five points higher (95% CI: 12%, 38%). The indication from these data is that the effect of psychiatric distress is smaller among those who reported having used alcohol at the prior visit; a five point higher distress score is associated with only a 10% higher odds of consistently reporting use in this case.

Acknowledgment

The data for our example were gathered with a National Institute on Drug Abuse research grant (DA04392), for which James C. Anthony is Principal Investigator. The authors also gratefully acknowledge support from the National Institute of Mental Health grants MH38725-11 (Dr P. Leaf, PI) and MH56639 (Dr S. Zeger).

Table 2 Results of random intercept model for the alcohol use data

Variable	Estimate	Standard error
Intercept (β_0^*)	−0.363	0.0907
Time (β_1^*)	0.322	0.0350
MH score (β_2^*)	0.0473	0.0090
α	1.64	

Table 3 Results for transition model fit to the alcohol use data

Variable	Estimate	Standard error
Intercept (β_0^{**})	−1.11	0.136
Time (β_1^{**})	0.245	0.0599
MH score (β_2^{**})	0.0442	0.0599
Prior use (α_1)	1.71	0.0979
$Y_{i,j-1}^*$MH score (α_2)	−0.0247	0.0164

References

[1] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[2] Fitzmaurice, G.M. & Laird, N.M. (1993). Regression models for discrete longitudinal responses, *Statistical Science* **8**, 601–612.

[3] Gourieroux, C., Monfort, A. & Trognon, A. (1984). Psuedo-maximum likelihood methods: theory, *Econometrica* **52**, 681–700.

[4] Huber, P.J. (1967). The behaviour of maximum likelihood estimators under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, L.M. LeCam & J. Neyman, eds. University of California Press, Berkeley, pp. 221–233.

[5] Kellam, S.G., Rebok, G.W., Ialongo, N.S. & Mayer, L. (1994). The course and malleability of aggressive behavior from early first grade to middle school: results of an epidemiologically-based preventive trial, *Journal of Child Psychology and Psychiatry and Allied Disciplines* **35**, 983.

[6] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[7] Liang, K.-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[8] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

[9] Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**, 1033–1048.

[10] Prentice, R.L. & Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics* **47**, 825–839.

[11] Stiratelli, R., Laird, N. & Ware, J.H. (1984). Random effects models for serial observations with binary responses, *Biometrics* **40**, 961–971.

[12] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gaussian method, *Biometrika* **61**, 439–447.

[13] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrics* **50**, 1–25.

[14] Zeger, S.L. & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**, 121–130.

[15] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

SCOTT L. ZEGER, PETER J. DIGGLE & W. HUANG

Generalized Maximum Likelihood

Many interpretations of the phrase "generalized maximum likelihood" are possible. A few such, which are to be found in the literature, are listed here, in no particular order. These methods have been studied by many authors, but here, for brevity, we quote only one or two references for each from which other work can be traced. Also, discussion will be restricted to the fully parametric case, so that semiparametric methods such as Cox's **partial likelihood** for the **proportional hazards** model (*see* **Cox Regression Model**), and nonparametric estimators such as that of **Kaplan and Meier** for the survivor function, are not covered. Throughout, "**maximum likelihood**" is abbreviated to ML.

Maximum Probability Estimators

Weiss & Wolfowitz [12] drew attention to the overly simplistic, in their view, assumptions made in order to construct "regular" likelihood theory. In particular, they were concerned with the common restriction to estimators which have asymptotic normal distributions (*see* **Large-sample Theory**).

Let the data be D_n, where n represents the sample size or some similar quantity, and let $f_n(\cdot;\theta)$ be the density or probability function of D_n depending on a parameter θ. The ordinary ML estimator is found by maximizing $f_n(D_n;\theta)$ over θ. It is assumed that a normalizing sequence $k_n \to \infty$ for the family $f_n(D_n;\theta)$ can be found somehow such that: there exists an estimator T_n and $m(\theta) > 0$ such that

$$\liminf_{n\to\infty} \Pr_{\theta}[k_n|T_n - \theta| < m(\theta)] > 1 - \varepsilon,$$

for every $\varepsilon > 0$. In addition, it is assumed that if k_n is replaced by a sequence k_n' tending to infinity faster than k_n, $k_n'|T_n - \theta|$ is not bounded in this way for any T_n, so the above probability then tends to zero instead of being close to 1. Thus, k_n is the maximal rate, and T_n is a **consistent estimator** tending to θ at this rate. In the "regular" likelihood case, k_n can be taken as $n^{1/2}$ (or as $3n^{1/2}$ or $n^{1/2} + \log n$, etc.) and T_n as $\hat{\theta}_n$, the ML estimator (or as $\hat{\theta}_n + 3/n^2$, etc.).

A maximum probability (MP) estimator is defined as the θ value that maximizes

$$I_n(\theta) = \int_{\theta - r/k_n}^{\theta + r/k_n} f_n(D_n;t)\,\mathrm{d}t,$$

r being some specified positive constant. As $n \to \infty$, the interval $(\theta - r/k_n, \theta + r/k_n)$ shrinks towards its midpoint θ. The MP estimator is chosen so that the average value of $f_n(D_n;\theta)$ over this interval; namely, $I_n(\theta) \div (2r/k_n)$, is maximized. When $f_n(D_n;t)$ is continuous in t, $I_n(\theta) \div (2r/k_n) \to f_n(D_n;\theta)$ as $r/k_n \to 0$, and then MP estimation becomes equivalent to ML estimation. In this sense, MP estimation is a kind of smoothed version of ML estimation, and is asymptotically equivalent to it in the regular case. Weiss & Wolfowitz actually defined MP estimation more generally with the integral over a shrinking set $\{t: k_n(t - \theta) \in R\}$, where R is a specified bounded subset of the real line.

The MP estimate has an optimality property: roughly speaking, $k_n(T_n - \theta_0)$ is asymptotically smallest in probability when T_n is the MP estimator, θ_0 being the true parameter. This maximum efficiency of the MP estimator was explained by Weiss & Wolfowitz by interpreting it as being asymptotically equivalent to a certain Bayes estimator with respect to a **prior distribution** uniform on $(\theta - r/k_n, \theta + r/k_n)$. Various other technical properties and extensions, including that to vector parameters, were discussed by Weiss & Wolfowitz. The main thrust is that the MP estimator is not restricted to the regular cases which support the standard theory of ML estimation. Some further results are given in [1, Section 3.3].

Maximum Probability of Spacings

Let $x_1, \ldots, x_n$ be an ordered random sample from a continuous univariate distribution with distribution function $F(x;\boldsymbol{\theta})$ and density $f(x;\theta)$ on (α_1, α_2). The end points α_1 and α_2 may be known or unknown; in the latter case they are included as components of the parameter vector θ.

Define $d_i(\theta) = F(x_i;\theta) - F(x_{i-1};\theta)$ for $i = 1, \ldots, n+1$, taking x_0 as α_1 and x_{n+1} as α_2. Under θ, the $d_i(\theta)$ are the "uniform spacings" derived from the x-sample, with $\sum d_i(\theta) = 1$. The maximum product of spacings (MPS) method of Cheng & Amin [3]

chooses as the estimator of θ that value which maximizes

$$G(\theta) = \prod_{i=1}^{n+1} d_i(\theta).$$

Note that $G(\theta) \leq (n+1)^{-(n+1)}$, this maximum possible value being attained when the $d_i(\theta)$s are all equal to $(n+1)^{-1}$. The MPS estimator is thus the θ value which makes the sample spacings most nearly uniform.

Cheng & Amin pointed out that, in "regular" situations, the remainder term $r(x_i, x_{i-1}; \theta)$ in

$$d_i(\theta) = f(x_i; \theta)(x_i - x_{i-1}) + r(x_i, x_{i-1}; \theta)$$

becomes negligible as $n \to \infty$, and $(x_i - x_{i-1})$ does not depend on θ. Then, maximizing $\prod d_i(\theta)$ is asymptotically equivalent to maximizing $\prod f(x_i; \theta)$, the likelihood function, and so MPS estimation is asymptotically equivalent to ML estimation. In nonregular situations this equivalence can break down, and Cheng & Amin gave examples in which MPS yields better estimates than ML, the latter method sometimes failing altogether.

Corrected Likelihood

Let $x_1, \ldots, x_n$ be a random sample from a continuous univariate distribution with distribution function $F(x; \theta)$ and density $f(x; \theta)$, and let

$$H(\theta) = \prod_{i=1}^{n} [F(x_i + h_i; \theta) - F(x_i; \theta)],$$

where the h_i are small positive quantities. This would be the **likelihood** function appropriate to **grouped data** in which the ith observation is only known to lie in the interval $(x_i, x_i + h_i]$; it has been argued that, because all measurement is of limited accuracy, this form of likelihood is more realistic in practice. Cheng & Iles [4] described the standard approach in which one makes the approximation

$$H(\theta) \doteqdot \prod_{i=1}^{n} [f(x_i; \theta) h_i],$$

where $f(x; \theta)$ is the density function, and then ignores the h_i since they do not involve θ. Thus, the standard likelihood function is defined as $L(\theta) = \prod f(x_i; \theta)$.

Consider, as an example, the **Weibull distribution** function $F(x; \theta) = 1 - \exp[-(x - \alpha)^\beta]$ on (α, ∞) with $\theta = (\alpha, \beta)$, α being the lower endpoint and β the shape parameter. It is known that when $\beta < 1$ there is no consistent solution of the likelihood equation $\partial \log L(\theta)/\partial\theta = 0$, so the usual ML theory fails. Cheng & Iles identified the source of the problem as follows. For $\beta < 1$, $L(\theta)$ is an increasing function of α, so $\hat{\alpha} = \min(x_1, \ldots, x_n) = x_m$, say. The contribution to $H(\theta)$ from x_m when $\theta = (\hat{\alpha}, \beta)$ is

$$F(x_m + h_m; \theta) - F(x_m; \theta) = 1 - \exp(-h_m^\beta) \doteqdot h_m^\beta.$$

In contrast, the corresponding contribution in the approximated version is

$$f(x_m; \theta) h_m = h_m \beta (x - x_m)^{\beta - 1} \exp[-(x - x_m)^\beta] = \infty.$$

Their suggestion was to use $H(\theta)$ as the "corrected" likelihood. In regular cases, this will be equivalent to using $L(\theta)$. In the Weibull example, only h_m will have any effect, and Cheng & Iles make some suggestions for its choice.

General Estimating Functions

Perhaps the most general definition of an estimator is simply that it is some function of the data, and that of a generalized ML estimator is that it results from maximizing some function of the data, this function deputizing for the likelihood. In the above, this function is $I_n(\theta)$ for MP estimation, $G(\theta)$ for MPS estimation, and $H(\theta)$ for "corrected likelihood" estimation. These three methods assume that the parametric distributional form for the data is known, and the generalization focuses upon extending the theory beyond the usual regularity restrictions. Other suggestions in the literature are more concerned with **robustness**: the deputizing function is chosen to avoid critical dependence on the parametric form adopted for the likelihood. Such proposals include: **least squares** [7]; conditional least squares [10]; Gaussian estimation [6, 13]; M-estimators [8]; maximum **quasi-likelihood** [11]; and minimum chi-square [2]. General asymptotic theory for estimating functions has been given by Huber [9] and Crowder [5].

References

[1] Akahira, M. & Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts of Higher Order Asymptotic Efficiency*. Springer-Verlag, New York.

[2] Berkson, J. (1980). Minimum chi-square, not maximum likelihood!, *Annals of Statistics* **8**, 457–487.

[3] Cheng, R.C.H. & Amin, N.A.K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin, *Journal of the Royal Statistical Society, Series B* **45**, 394–403.

[4] Cheng, R.C.H. & Iles, T.C. (1987). Corrected maximum likelihood in non-regular problems, *Journal of the Royal Statistical Society, Series B* **49**, 95–101.

[5] Crowder, M.J. (1986). Consistency and inconsistency of estimating equations, *Econometric Theory* **2**, 305–330.

[6] Crowder, M.J. (1987). On linear and quadratic estimating functions, *Biometrika* **74**, 591–597.

[7] Gauss, C.F. (1809). *Theoria motus corporum coelestum in sectionibus conicisolem ambientium*. Pathes & Besser, Hamburg.

[8] Huber, P.J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**, 73–101.

[9] Huber, P.J. (1967). The behavior of maximum likelihood estimators under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L.M. LeCam & J. Neyman, eds. University of California Press, Berkeley.

[10] Klimko, L.A. & Nelson, P.I. (1978). On conditional least squares estimation for stochastic processes, *Annals of Statistics* **6**, 629–642.

[11] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method, *Biometrika* **61**, 439–447.

[12] Weiss, L. & Wolfowitz, J. (1974). *Maximum Probability Estimators and Related Topics*. Springer-Verlag, New York.

[13] Whittle, P. (1961). Gaussian estimation in stationary time series, *Bulletin of the International Statistical Institute* **39**, 1–26.

(*See also* **Generalized Estimating Equations**)

M.J. CROWDER

Generating Functions

Biostatistical information usually comes in the form of a sample of observed values, e.g. serum cholesterol levels in a group of 30 students. Often a hypothesis is constructed concerning the distribution of the values in the population from which the sample (group) is drawn. This may be specified by a formula for its probability density function (pdf) for a continuous distribution, or by its probability mass function (pmf) for a discrete distribution. Pdfs and pmfs are, however, often less informative than the **moments** of a distribution. Moments can be summarized by a **moment generating function**.

A generating function is similar to an alphabetical list – it is a single piece of documentation about a sequence of items. When a moment generating function for a distribution is expanded into a series, it gives the values of all the moments as the coefficients of successive terms in the series. The first term gives the mean, the second tells us about the variability, the third about the **skewness**, etc.

Another type of sequence that occurs in biostatistics is a sequence of probabilities, for instance the probabilities of $0, 1, 2, \ldots$ spina bifida births in a region per month. The coefficients in a probability generating function (pgf) provide a sequence of probabilities.

Suppose that $a_0, a_1, a_2, \ldots$ is an infinite sequence of real numbers all of which are finite. Then the power series

$$H(s) = a_0 + a_1 s + a_2 s^2 + \cdots = \sum_{r=0}^{\infty} a_r s^r$$

turns the sequence into the function $H(s)$ [given a finite sequence, $a_0, a_1, \ldots, a_n$, then the corresponding function is $H(s) = \sum_{r=0}^{n} a_r s^r$]. $H(s)$ is called the generating function of the sequence and s is called the generating variable (s is a mathematical artifact – it has no statistical interpretation and is not a random variable). When the sequence is infinite, the restriction that all the a_r are finite ensures that $H(s)$ exists and has a finite sum, provided that s is not too large. Two sequences that are identical have the same power series generating function. Less obviously, if two generating functions are identical, then so are the sequences that they generate; this property of uniqueness is important.

A second form of generating function somewhat resembles an exponential series and is called an exponential generating function. The connection between an infinite sequence and its exponential generating function is

$$h(t) = \sum_{r=0}^{\infty} \frac{a_r t^r}{r!};$$

for a finite sequence the exponential generating function is $h(t) = \sum_{r=0}^{n} a_r t^r / r!$. Even if the a_r increase with r without bound, $h(t)$ will exist if $a_r / r!$ is finite for all r. An exponential generating function also has the property of uniqueness.

Generating functions are useful not only as generators of formulas for the individual items in a sequence. They give recurrence relations when the mathematical expressions for the a_r are complicated and they can provide good approximations. In discrete distribution theory they are particularly valuable for combining sequences of probabilities in various ways.

Moment Generating Functions

The rth uncorrected moment, μ'_r, of the **random variable** X is the expected value of X^r, where r is a positive integer, i.e. $\mu'_r = \mathrm{E}(X^r)$. If $\mu'_r / r!$ is finite for all r, then

$$\begin{aligned} U_X(t) &= \mathrm{E}\left(1 + Xt + \frac{X^2 t^2}{2!} + \frac{X^3 t^3}{3!} + \cdots\right) \\ &= \sum_{r \geq 0} \frac{\mu'_r t^r}{r!} = \mathrm{E}[\exp(Xt)] \end{aligned} \tag{1}$$

is the uncorrected-moment generating function (umgf). It is an exponential generating function. By successive differentiation

$$\begin{aligned} \mu'_1 &= \left[\frac{\mathrm{d}U_X(t)}{\mathrm{d}t}\right]_{t=0}, \\ \mu'_2 &= \left[\frac{\mathrm{d}^2 U_X(t)}{\mathrm{d}t^2}\right]_{t=0}, \ldots, \\ \mu'_r &= \left[\frac{\mathrm{d}^r U_X(t)}{\mathrm{d}t^r}\right]_{t=0}, \ldots. \end{aligned} \tag{2}$$

Example 1: Uncorrected Moments of the Binomial Distribution

Suppose that bacteria in a particular culture have fixed and independent probabilities p of being mutant and that n of them are examined under a microscope. Assume that the number of mutants has a **binomial distribution** with probability mass function (pmf) $\binom{n}{x}p^x(1-p)^{n-x}$, where n is a positive integer, $0 < p < 1$, $x = 1, 2, \ldots, n$. Then

$$\begin{aligned} U_X(t) &= \mathrm{E}[\exp(Xt)] \\ &= \sum_{x=0}^{n} \frac{n!p^x(1-p)^{n-x}}{x!(n-x)!} \cdot \mathrm{e}^{xt} \\ &= (1-p+p\mathrm{e}^t)^n \\ &= 1 + \frac{np(\mathrm{e}^t-1)}{1!} \\ &\quad + \frac{n(n-1)p^2(\mathrm{e}^t-1)^2}{2!} + \cdots \\ &= 1 + np\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right) \\ &\quad + \frac{n(n-1)p^2}{2!}\left(t^2 + \frac{2t^3}{2!} + \cdots\right) + \cdots . \end{aligned}$$

The uncorrected moments are therefore $\mu_1' = np$, $\mu_2' = np + n(n-1)p^2$, etc. A quicker way to find them is to set $t = 0$ in

$$\begin{aligned} \frac{\mathrm{d}U_X(t)}{\mathrm{d}t} &= n(1-p+p\mathrm{e}^t)^{n-1}p\mathrm{e}^t, \\ \frac{\mathrm{d}^2U_X(t)}{\mathrm{d}t^2} &= n(n-1)(1-p+p\mathrm{e}^t)^{n-2}p^2\mathrm{e}^{2t} \\ &\quad + n(1-p+p\mathrm{e}^t)^{n-1}p\mathrm{e}^t, \text{ etc.} \end{aligned}$$

The variable $X + c$ has the distribution of X shifted c units to the right of the origin. Because $\mathrm{E}\{\exp[t(X+c)]\} = \exp(ct)\mathrm{E}[\exp(tX)]$, we have

$$U_{X+c}(t) = \mathrm{e}^{ct}U_X(t); \tag{3}$$

this relates the moments of a shifted distribution to those of the original distribution.

Example 2: Uncorrected Moments of the Two- and Three-parameter Gamma Distribution

Let the survival time after a particular surgical procedure have an (unshifted) two-parameter **gamma distribution** with parameters a and b. Then the probability density function (pdf) for survival time is $f(x) = a^b\exp(-ax)x^{b-1}/\Gamma(b)$, $0 < a$, $0 < b$, $0 \le x < \infty$. The umgf is therefore

$$\begin{aligned} U_X(t) &= \int_0^\infty \frac{a^b\mathrm{e}^{-ax}x^{b-1}\mathrm{e}^{xt}}{\Gamma(b)}\,\mathrm{d}x \\ &= \left(\frac{a}{a-t}\right)^b \int_0^\infty \frac{\mathrm{e}^{-y}y^{b-1}\,\mathrm{d}y}{\Gamma(b)}\,\mathrm{d}x, \\ &\quad \text{where } y = x(a-t), \\ &= \left(\frac{a}{a-t}\right)^b \\ &= 1 + \frac{bt}{a} + \frac{b(b+1)t^2}{a^2 2!} \\ &\quad + \frac{b(b+1)(b+2)t^3}{a^3 3!} + \cdots, \end{aligned}$$

and the uncorrected moments are $\mu_1' = b/a$, $\mu_2' = b(b+1)/a^2$, $\mu_3' = b(b+1)(b+2)/a^3$, etc. Note that as a becomes large, the uncorrected moments tend to zero and the distribution tends to a degenerate distribution at the origin.

Suppose now that there is an initial constant length of time, c, during which the patient is kept alive in intensive care. Assume, then, that the distribution is shifted by an amount c to the right of the origin. This gives a three-parameter gamma distribution with umgf

$$\begin{aligned} U_{X+c}(t) &= \exp(ct)\left(\frac{a}{a-t}\right)^b \\ &= \left[1 + ct + \frac{(ct)^2}{2!} + \cdots\right] \\ &\quad \times \left[1 + \frac{bt}{a} + \frac{b(b+1)t^2}{a^2 2!} + \cdots\right], \end{aligned}$$

and its uncorrected moments are $\mu_1' = c + b/a$, $\mu_2' = c^2 + 2cb/a + b(b+1)/a^2$, etc. As a gets large, the new umgf tends to $\exp(ct)$, showing that the distribution tends to a degenerate distribution at $x = c$.

Some distributions have moments for which $\mu_r'/r!$ is unbounded and so the umgf does not exist. The **characteristic function** (cf) exists for all distributions, however. For a continuous distribution this is

$$\varphi_X(t) = \mathrm{E}[\exp(itX)] = \int_{-\infty}^{\infty} \exp(itx)\,\mathrm{d}F(x), \tag{4}$$

where $i = \sqrt{(-1)}$ and t is real; for a discrete distribution on 0, 1, 2, ... it is

$$\varphi_X(t) = \mathrm{E}[\exp(itX)] = \sum_{x\geq 0} \exp(ixt)\Pr(X = x). \tag{5}$$

When the umgf exists, $U_X(t) = \varphi_X(-it)$.

The rth moment about the mean of a distribution is $\mu_r = \mathrm{E}((X-\mu)^r)$, where μ is the mean; it is called the rth corrected moment (or the rth central moment). The first central moment, μ_1, is always zero. The second central moment, μ_2, is the variance, $\mathrm{var}(X)$. The central-moment generating function (cmgf) is

$$\begin{aligned} M_X(t) &= \mathrm{E}\left(1 + (X-\mu)t + \frac{(X-\mu)^2t^2}{2!} + \cdots\right) \\ &= 1 + \sum_{r\geq 2} \frac{\mu_r t^r}{r!} = \mathrm{E}\{\exp[(X-\mu)t]\}. \end{aligned} \tag{6}$$

Shifting a distribution leaves its central moments unaltered, since

$$M_{X+a}(t) = \mathrm{E}\{\exp[(X+a-\mu-a)t]\} = M_X(t). \tag{7}$$

Example 3: Central Moments of the Normal Distribution

The concentration of the antibiotic in tubes of chloramphenicol gel is assumed to have a **normal distribution** with parameters μ, σ^2 and pdf $f(x) = \exp[-(x-\mu)^2/2\sigma^2]/[\sigma(2\pi)^{1/2}]$, $-\infty < x < \infty$. The cmgf is

$$\begin{aligned} M_X(t) &= \mathrm{E}\{\exp[t(X-\mu)]\} \\ &= \int_{-\infty}^{\infty} \frac{\exp\left[\dfrac{-(x-\mu)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}} \\ &\quad \times \exp[(x-\mu)t]\,\mathrm{d}x \\ &= \frac{\exp\left(\dfrac{\sigma^2t^2}{2}\right)}{\sigma\sqrt{2\pi}} \\ &\quad \times \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu-\sigma^2t)^2}{2\sigma^2}\right] \mathrm{d}x \\ &= \exp\left(\frac{\sigma^2t^2}{2}\right) \\ &= 1 + \frac{\sigma^2t^2}{1!2} + \frac{\sigma^4t^4}{2!2^2} + \frac{\sigma^6t^3}{3!2^3} + \cdots. \end{aligned}$$

The odd central moments, μ_{2r+1}, are zero and the even ones are $\mu_r = (2r)!(\sigma^2/2)^r/r!$.

The relationship between the cmgf and the umgf is

$$M_X(t) = \mathrm{E}\{\exp[(X-\mu)t]\} = \exp(-\mu t)U_X(t); \tag{8}$$

therefore

$$\begin{aligned} \sum_{r\geq 0} \frac{\mu_r t^r}{r!} &= \left[1 - \mu t + \frac{(\mu t)^2}{2!} - \frac{(\mu t)^3}{3!} + \cdots\right] \\ &\quad \times \sum_{r\geq 0} \frac{\mu'_r t^r}{r!} \end{aligned} \tag{9}$$

and

$$\begin{aligned} \sum_{r\geq 0} \frac{\mu'_r t^r}{r!} &= \left[1 + \mu t + \frac{(\mu t)^2}{2!} + \frac{(\mu t)^3}{3!} + \cdots\right] \\ &\quad \times \sum_{r\geq 0} \frac{\mu_r t^r}{r!}. \end{aligned} \tag{10}$$

Because two sequences that are identical have the same generating function, we can equate the coefficients of $t^r/r!$ on the two sides of (9) to give

$$\mu_r = \sum_{j=0}^{r} (-1)^j \binom{r}{j} \mu'_{r-j}\mu^j, \tag{11}$$

i.e. $\mu_1 = -\mu'_1 + \mu = 0,$

$\mu_2 = \mu'_2 - \mu^2,$

$\mu_3 = \mu'_3 - 3\mu'_2\mu + 2\mu^3,$ etc.

Also, from (10), we have

$$\mu'_r = \sum_{j=0}^{r} \binom{r}{j} \mu_{r-j}\mu^j, \tag{12}$$

i.e. $\mu'_1 = \mu,$

$\mu'_2 = \mu_2 + \mu^2,$

$\mu'_3 = \mu_3 + 3\mu_2\mu + \mu^3,$ etc.

Example 1 continued: Central Moments of the Binomial Distribution

From Example 1 and the relationship between the cmgf and the umgf, (8), the cmgf of the **binomial distribution** is

$$M_X(t) = \left[1 - \mu t + \frac{(\mu t)^2}{2!} - \cdots\right]\left\{1 + npt + \frac{[np + n(n-1)p^2]t^2}{2!} + \cdots\right\},$$

where $\mu = np$. Hence $\mu_1 = 0$, $\mu_2 = np(1-p)$, etc.

The well-known binomial and exponential expansions were used in Examples 1 and 2. A standard way to create an unknown expansion is via a Maclaurin series:

$$h(t) = [h(t)]_{t=0} + \left[\frac{\mathrm{d}h(t)}{\mathrm{d}t}\right]_{t=0}\frac{t}{1!} + \left[\frac{\mathrm{d}^2h(t)}{\mathrm{d}t^2}\right]_{t=0}\frac{t^2}{2!} + \left[\frac{\mathrm{d}^3h(t)}{\mathrm{d}t^3}\right]_{t=0}\frac{t^3}{3!} + \cdots.$$

This gives $\mu_1 = [\,\mathrm{d}M_X(t)/\,\mathrm{d}t]_{t=0}$ (= 0 always), $\mu_2 = [\,\mathrm{d}^2U_X(t)/\,\mathrm{d}t^2]_{t=0}$, and so on.

Example 1 continued again: Central Moments of the Binomial Distribution

Here $h(t) = M_X(t) = \exp(-\mu t)(1 - p + pe^t)^n$, so

$$\frac{\mathrm{d}M_X(t)}{\mathrm{d}t} = -\mu M_X(t) + \frac{nM_X(t)pe^t}{1-p+pe^t},$$

$$\frac{\mathrm{d}^2M(t)}{\mathrm{d}t^2} = -\frac{\mu\,\mathrm{d}M_X(t)}{\mathrm{d}t} + \frac{n\,\mathrm{d}M_X(t)/\,\mathrm{d}t}{1-p+pe^t} + \frac{nM_X(t)pe^t}{1-p+pe^t} - \frac{nM_X(t)p^2e^{2t}}{(1-p+pe^t)^2}.$$

Therefore $\mu_1 = [\,\mathrm{d}M_X(t)/\,\mathrm{d}t]_{t=0} = -\mu + np = 0$ and $\mu_2 = [\,\mathrm{d}^2M_X(t)/\,\mathrm{d}t^2]_{t=0} = np - np^2$, as before.

An important property of the umgf is that if X, Y, Z, etc. are independent random variables, then

$$U_{X+Y}(t) = U_X(t)U_Y(t),$$

$$U_{X+Y+Z}(t) = U_X(t)U_Y(t)U_Z(t), \text{ etc.},$$

and

$$U_{X-Y}(t) = U_X(t)U_Y(-t).$$

Similarly, for the cmgf,

$$M_{X+Y}(t) = M_X(t)M_Y(t),$$

$$M_{X+Y+Z}(t) = M_X(t)M_Y(t)M_Z(t), \text{ etc.},$$

and

$$M_{X-Y}(t) = M_X(t)M_Y(-t).$$

These formulas give the moments of sums and differences of independent variables in terms of the moments of the individual variables.

Cumulant Generating Function

The cumulant generating function (cgf) of a distribution, $K(t)$, is sometimes simpler to handle than the umgf or cmgf. It is the logarithm to base e of the umgf. Expanding it as an exponential generating function gives the cumulants κ_r of the distribution, i.e.

$$K_X(t) = \ln U_X(t) = \ln\left(\frac{1 + \mu_1' t + \mu_2' t^2}{2!} + \cdots\right) = \sum_{r\geq 1}\frac{\kappa_r t^r}{r!}; \tag{13}$$

Expansion of the logarithm shows that there is no term in t^0 in (13) and that $\kappa_1 = \mu_1' = \mu$.

Example 4: Cumulants of the Poisson Distribution

The response in quanta of acetylcholine released per stimulus of a nerve cell has a **Poisson distribution** with parameter λ and pmf $e^{-\lambda}\lambda^x/x!$, $0 < \lambda$, $0 \leq x < \infty$. The umgf is

$$U_X(t) = \sum_{x=0}^{\infty}\left[\frac{e^{-\lambda}\lambda^x}{x!}\right]e^{xt} = \exp[\lambda(e^t - 1)],$$

so

$$K_X(t) = \lambda(e^t - 1) = \lambda + \lambda t + \frac{\lambda t^2}{2!} + \frac{\lambda t^3}{3!} + \cdots.$$

The cumulants of the Poisson distribution are all equal to the mean λ.

The cgf is an exponential generating function and so if need be the cumulants can be obtained as $\mu_1 = [\,\mathrm{d}K_X(t)/\,\mathrm{d}t]_{t=0}$, $\mu_2 = [\,\mathrm{d}^2K_X(t)/\,\mathrm{d}t^2]_{t=0}$, etc.

Because $K_{X+a}(t) = \ln[\exp(at)U_X(t)] = at + K_X(t)$, the coefficients of $t^r/r!$ in $K_{X+a}(t)$ and in $K_X(t)$ are the same for $r \geq 2$. This implies that when a constant is added to X, only κ_1 is changed; this is why the cumulants are sometimes called semi-invariants.

Consider

$$K_{X-\mu}(t) = \ln M_X(t) = \left(\frac{\mu_2 t^2}{2!} + \frac{\mu_3 t^3}{3!} + \frac{\mu_4 t^4}{4!} + \cdots\right) - \left(\frac{\mu_2^2 t^4}{(2!2!2)} + \cdots\right) + \cdots.$$

Equating the coefficients of t^r in $K_{X-\mu}(t)$ and $\ln M_X(t)$ shows that for $r \geq 2$ the cumulants are functions of the central moments

$$\kappa_2 = \mu_2, \quad \kappa_3 = \mu_3, \quad \kappa_4 = \mu_4 - 3\mu_2^2, \text{etc.} \tag{14}$$

Example 4 continued: Central Moments of the Poisson Distribution

Because $\kappa_r = \lambda$ for all r for the Poisson distribution, the mean is $\mu_1' = \lambda$ and, from (14), $\mu_2 = \lambda$, $\mu_3 = \lambda$, $\mu_4 = \lambda - 3\lambda^2$.

Let $X = X_1 + X_2 + \cdots + X_n$, where the X_js are independent random variables. The umgf of X is the product of the individual umgfs, and therefore the cgf of X is the sum of their cgfs

$$K_X(t) = \sum_{j=1}^{n} K_{X_j}(t). \tag{15}$$

Thus the kth cumulant of a sum is the sum of the individual kth cumulants.

Joint Moments and Cumulants

When X_j, $j = 1, 2, \ldots, m$, are not independent we need to look at their joint distribution. The uncorrected moments of a joint distribution are quantities like $\mathrm{E}(\prod_{j=1}^{m} X_j^{r_j})$; they are denoted by $\mu'_{r_1 r_2 \ldots r_m}$ and are called product moments about zero. The central product moments are

$$\mu_{r_1 r_2 \ldots r_m} = \mathrm{E}\left(\prod_{j=1}^{m} [X_j - \mathrm{E}(X_j)]^{r_j}\right). \tag{16}$$

For a bivariate distribution the central product moment, μ_{11}, is the covariance, cov (X_1, X_2).

The joint umgf of $X_1, X_2, \ldots, X_m$ is a function of m generating variables $t_1, t_2, \ldots, t_m$:

$$U_{X_1,\ldots,X_m}(t_1, t_2, \ldots, t_m) = \mathrm{E}\left[\exp\left(\sum_{j=1}^{m} t_j X_j\right)\right], \tag{17}$$

the joint cmgf is

$$\mathrm{E}\left(\exp\left\{\sum_{j=1}^{m} t_j [X_j - \mathrm{E}(X_j)]\right\}\right) = \exp\left\{-\sum_{j=1}^{m} t_j \mathrm{E}(X_j)\right\} \times U(t_1, t_2, \ldots, X_m), \tag{18}$$

and the joint cgf is $\ln U_{X_1 \ldots X_m}(t_1, t_2, \ldots, t_m)$. Their use is similar to that for univariate distributions.

Example 5: Moments of a Bivariate Gamma Distribution

Pairs of tumors were initiated at the same time, one on each side of the back of a mouse. Suppose that the sizes, X and Y, of the two tumors after one month's growth are not independent but have the joint pdf

$$f(x, y) = \begin{cases} a^2\{\mathrm{e}^{-ay} - \mathrm{e}^{-a(x+y)}\}, & 0 \leq x < y, \\ a^2\{\mathrm{e}^{-ax} - \mathrm{e}^{-a(x+y)}\}, & 0 \leq y < x, \end{cases}$$

$0 < a$. Then the marginal distribution of X has the pdf

$$f(x) = \int_0^y a^2\{\mathrm{e}^{-ay} - \mathrm{e}^{-a(x+y)}\}\,\mathrm{d}x + \int_y^\infty a^2\{\mathrm{e}^{-ax} - \mathrm{e}^{-a(x+y)}\}\,\mathrm{d}x = a^2 y \mathrm{e}^{-ay}$$

(a particular sort of gamma distribution); the marginal distribution of Y is similar. The joint umgf is

$$U_{X,Y}(t_1, t_2) = \int_0^\infty \int_0^y a^2\{\mathrm{e}^{-ay} - \mathrm{e}^{-a(x+y)}\}$$

$$
\begin{aligned}
&\times \exp[t_1 x + t_2 y]\,\mathrm{d}x\,\mathrm{d}y \\
&+ \int_0^\infty \int_y^\infty a^2\{\mathrm{e}^{-ax} - \mathrm{e}^{-a(x+y)}\} \\
&\times \exp[t_1 x + t_2 y]\,\mathrm{d}x\,\mathrm{d}y \\
&= a^2 \int_0^\infty \left[\frac{\exp[-(a - t_1 - t_2)y]}{t_1} \right. \\
&+ \frac{\exp[-(a - t_1 - t_2)y]}{(a - t_1)} \\
&\left. - \frac{\exp[-(a - t_1)y]}{t_1} - \frac{\exp[-(a - t_2)y]}{a - t_1} \right] \mathrm{d}y \\
&= a^3 (a - t_1)^{-1} (a - t_2)^{-1} (a - t_1 - t_2)^{-1}.
\end{aligned}
$$

The joint cgf is

$$
\begin{aligned}
K_{X,Y}(t_1, t_2) &= \ln[U_{X,Y}(t_1, t_2)] \\
&= -\ln\left(1 - \frac{t_1}{a}\right) - \ln\left(1 - \frac{t_2}{a}\right) \\
&\quad - \ln\left(1 - \frac{t_1}{a} - \frac{t_2}{a}\right) \\
&= \frac{2t_1}{a} + \frac{2t_2}{a} + \frac{2t_1^2}{a^2 2!} + \frac{t_1 t_2}{a^2 1!1!} \\
&\quad + \frac{2t_2^2}{a^2 2!} + \cdots,
\end{aligned}
$$

showing that $\mathrm{E}(X) = \kappa_{1,0} = 2/a$, $\mathrm{E}(Y) = \kappa_{0,1} = 2/a$, $\mathrm{var}(X) = \kappa_{2,0} = 2/a^2$, and $\mathrm{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \kappa_{1,1} = 1/a^2$.

Probability Generating Function

The probability generating function (pgf) is a very useful tool for studying discrete distributions; it is not applicable to continuous distributions. Let X be a discrete random variable X taking the values $0, 1, 2, \ldots$ with nonzero probability mass function (pmf) $p_x = \Pr(X = x)$. Then the pgf of the distribution is

$$
G_X(z) = \sum_{x=0}^{\infty} p_x z^x = \mathrm{E}(z^X). \tag{19}
$$

The conditions for $G_X(z)$ to be a pgf are

$$
G_X(0) \geq 0,
$$

$$
G_X(1) = 1 \text{ and } \left[\frac{\mathrm{d}^r G_X(z)}{\mathrm{d}z^r} \right]_{z=0} \geq 0, r > 0. \tag{20}
$$

Note that whereas the umgf, cmgf, and cgf are exponential generating functions where the a_r are the coefficients of $t^r/r!$, the pgf is a power-series generating function and the p_x are the coefficients of z^x. If probabilities are obtained via a Maclaurin series, then it is important to remember that $p_r = (r!)^{-1}[\,\mathrm{d}^r G_X(z)/\,\mathrm{d}z^r]_{z=0}$.

The pgf, cf, umgf, cmgf, and cgf of a discrete distribution are closely related. We have

$$
\begin{aligned}
G_X(z) &= \mathrm{E}(z^X), \\
\varphi_X(t) &= \mathrm{E}[\exp(itX)] = G(\mathrm{e}^{it}) \\
U_X(t) &= \mathrm{E}[\exp(tX)] = G(\mathrm{e}^t), \\
M_X(t) &= \mathrm{E}\{\exp[t(X - \mu)]\} \\
&= \mathrm{e}^{-\mu t} G(\mathrm{e}^t), \\
K_X(t) &= \ln \mathrm{E}[\exp(tX)] = \ln G(\mathrm{e}^t).
\end{aligned}
$$

Example 6: Probability Generating Function of an Exponentially Mixed Poisson Distribution (a Geometric Distribution)

Suppose that the number of schistosome ova per specimen has a Poisson distribution with parameter λ, where λ varies from patient to patient according to an exponential distribution with pdf $c\mathrm{e}^{-\lambda c}\,\mathrm{d}\lambda$. Then the resultant distribution of the number of ova per specimen, X, has the pmf

$$
\begin{aligned}
\Pr(X = x) &= \int_0^\infty \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} \cdot c\mathrm{e}^{-\lambda c}\,\mathrm{d}\lambda \\
&= \frac{c}{(1 + c)^{x+1}}, \quad x = 0, 1, \ldots.
\end{aligned}
$$

This is the pmf for a **geometric distribution**. Consider now the pgf. In Example 4 we saw that the umgf of the Poisson distribution is $\exp[\lambda(\mathrm{e}^t - 1)]$, so its pgf is $\exp[\lambda(z - 1)]$. Integrating over λ using an exponential distribution gives the pgf of X as

$$
\begin{aligned}
G_X(z) &= \int_0^\infty \mathrm{e}^{\lambda(z-1)} \times c\mathrm{e}^{-\lambda c}\,\mathrm{d}\lambda \\
&= \frac{c}{c + 1 - s}.
\end{aligned}
$$

The uniqueness property of pgfs ensures that this is the pgf of the same (geometric) distribution.

Let $X_1, X_2, \ldots, X_k$ be k independent random variables with pgfs $G_1(z), G_2(z), \ldots, G_k(z)$ and set

$X = \sum_{j=1}^{k} X_j$. Then

$$G_X(z) = \mathrm{E}(z^{X_1+X_2+\ldots X_k}) = \prod_{j=1}^{k} \mathrm{E}(z^{X_j}) = \prod_{j=1}^{k} G_j(z); \tag{21}$$

this method of combining distributions is called *convolution*. The pgf for the difference of two random variables, X_1 and X_2, is

$$G_{X_1-X_2}(z) = G_{X_1}(z)G_{X_2}(z^{-1}). \tag{22}$$

Consider now a "damage" process. If we have x items initially and each item has an independent probability $(1-\alpha)$ of removal, then the probability that k of them remain is $\binom{x}{k}\alpha^k(1-\alpha)^{n-k}$. When the initial number, X, is a random variable with pgf $G_X(z) = p_0 + p_1 z + p_2 z^2 + \cdots$ and R is the number remaining, then

$$\Pr(R = r) = p_r^* = \sum_{x \geq r} p_x \binom{x}{r} \alpha^r (1-\alpha)^{x-r} \tag{23}$$

and the pgf for R is

$$\begin{aligned} G_R(z) &= \sum_{r\geq 0} p_r^* z^r = \sum_{r\geq 0}\sum_{x\geq r} p_x \binom{x}{r} \alpha^r (1-\alpha)^{x-r} z^r \\ &= \sum_{x\geq 0}\sum_{r\leq x} p_x \binom{x}{r} \alpha^r (1-\alpha)^{x-r} z^r \\ &= \sum_{x\geq 0} p_x (1-\alpha+\alpha z)^x \\ &= G_X(1-\alpha+\alpha z). \end{aligned} \tag{24}$$

A similar but rather more complicated way of combining two distributions is illustrated in the next example.

Example 7: A Compound Process for the Negative Binomial Distribution

Suppose that each year brain fluid is sampled from anencephalic infants from a varying number of hospitals. If the number of anencephalics per hospital, X, has a logarithmic distribution with pgf $\ln(1-\theta z)/\ln(1-\theta)$ and the number of cooperating hospitals per year, N, has a Poisson distribution with pgf $\exp\lambda(z-1)$, then the total number of specimens collected per year, Y, has the distribution with pgf

$$\begin{aligned} G_Y(z) &= \sum_{x=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} \left[\frac{\ln(1-\theta z)}{\ln(1-\theta)}\right]^x \\ &= \exp\left\{\lambda\left[\frac{\ln(1-\theta z)}{\ln(1-\theta)} - 1\right]\right\} \\ &= \left(\frac{1-\theta}{1-\theta z}\right)^{-\lambda/\ln(1-\theta)} \end{aligned}$$

(the sum of x independent variables with the same pgf, $h(z)$, has the pgf $[h(z)]^x$). The outcome is a **negative binomial distribution**. This is a random-stopped sum distribution with a pgf of the form $G_1[G_2(z)]$, sometimes called a compound or a generalized distribution (these terms have other meanings as well).

There are also bivariate and multivariate pgfs. Given k dependent discrete variables $X_1, X_2, \ldots, X_k$, their joint pgf is

$$G_{X_1,X_2,\ldots,X_k}(z_1, z_2, \ldots, z_k) = \mathrm{E}\left(\prod_{j=1}^{k} t_j^{X_j}\right). \tag{25}$$

Factorial Moment Generating Functions

For a discrete distribution the factorial moment generating function (fmgf) is often easier to handle than other types of moment generating functions. The term "factorial moment" nearly always refers to a descending factorial moment; the rth descending factorial moment of X is

$$\mu'_{[r]} = \mathrm{E}\left[\frac{X!}{(X-r)!}\right] = \sum_{x\geq r} \frac{p_x x!}{(x-r)!}. \tag{26}$$

(The rth ascending factorial moment of X is $\mathrm{E}[(X+r-1)!/(X-1)!]$.) In the past other notations have been adopted for the $\mu'_{[r]}$; these are still used occasionally.

The fmgf is an exponential generating function. We have

$$\begin{aligned} \sum_{r\geq 0}\sum_{x\geq r} \frac{\mu'_{[r]} t^r}{r!} &= \sum_{x\geq 0}\sum_{r\leq x} \frac{p_x x!}{(x-r)!}\frac{t^r}{r!} \\ &= \sum_{x\geq 0} p_x (1+t)^x = G_X(1+t), \end{aligned} \tag{27}$$

where $G(z)$ is the pgf. So

$$\mu'_{[r]} = \left[\frac{\mathrm{d}^r G_X(1+t)}{\mathrm{d}t^r}\right]_{t=0} = \left[\frac{\mathrm{d}^r G_X(z)}{\mathrm{d}t^r}\right]_{z=1}. \tag{28}$$

A joint fmgf for a multivariate distribution with pgf $G_{X_1,X_2,\ldots,X_k}(z_1, z_2, \ldots, z_k)$ is defined similarly as $G_{X_1,X_2,\ldots,X_k}(1+t_1, 1+t_2, \ldots, 1+t_k)$.

The relationships between the univariate factorial moments and the uncorrected moments are

$$\begin{aligned}
\mu'_1 &= \mu'_{[1]} = \mu,\\
\mu'_2 &= \mu'_{[2]} + \mu'_{[1]},\\
\mu'_3 &= \mu'_{[3]} + 3\mu'_{[2]} + \mu'_{[1]}, \text{ etc.}
\end{aligned}$$

and

$$\begin{aligned}
\mu'_{[2]} &= \mu'_2 - \mu'_1,\\
\mu'_{[3]} &= \mu'_3 - 3\mu'_2 + 2\mu'_1, \text{ etc.}
\end{aligned} \tag{29}$$

Stuart & Ord [7] give further details and formulas. A particularly useful equation is

$$\mu_2 = \mu'_{[2]} + \mu - \mu^2. \tag{30}$$

The factorial cumulant generating function (fcgf) is the logarithm of the (descending) fmgf. It is an exponential generating function and the rth factorial cumulant, $\kappa_{[r]}$, is the coefficient of $t^r/r!$ in its expansion, i.e.

$$\ln G(1+t) = \sum_{r\geq 1} \frac{\kappa_{[r]} t^r}{r!}. \tag{31}$$

The formulas connecting $\{\kappa_r\}$ and $\{\kappa_{[r]}\}$ are analogous to those connecting $\{\mu'_r\}$ and $\{\mu'_{[r]}\}$, i.e.

$$\begin{aligned}
\kappa_1 &= \kappa_{[1]} = \mu,\\
\kappa_2 &= \kappa_{[2]} + \mu,\\
\kappa_3 &= \kappa_{[3]} + 3\kappa_{[2]} + \mu, \text{ etc.}
\end{aligned} \tag{32}$$

Examples 1 and 7 continued: The Factorial Moment Generating Functions and the Factorial Cumulant Generating Functions for the Binomial and Negative Binomial Distributions

In Example 1 the number of mutants had a binomial distribution. The pgf is $\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} z^x = (1-p+pz)^n$; the fmgf is therefore

$$\begin{aligned}
G_X(1+t) &= (1+pt)^n\\
&= 1 + npt + \frac{n(n-1)p^2t^2}{2!}\\
&\quad + \frac{n(n-1)(n-2)p^3t^3}{3!} + \cdots
\end{aligned}$$

and the factorial moments are $\mu'_{[1]} = np$, $\mu'_{[2]} = n(n-1)p^2$, $\mu'_{[3]} = n(n-1)(n-2)p^3$, etc. The fcgf is $\ln G_X(1+t) = n\ln(1+pt)$.

In Example 7 the number of specimens collected per year, Y, had a negative binomial distribution. Setting $\pi = \theta/(1-\theta)$ and $l = -\lambda/\ln(1-\theta)$ enables the pgf to be restated as $G_Y(z) = (1+\pi-\pi z)^{-l}$. Thus the fmgf is $(1-\pi t)^{-l}$, giving $\mu'_{[1]} = l\pi$, $\mu'_{[2]} = l(l+1)\pi^2$, $\mu'_{[3]} = l(l+1)(l+2)\pi^3$, etc. The fcgf is $-l\ln(1-\pi t)$.

Bibliography

Historical aspects of the development and use of generating functions in probability theory are discussed in [6] and the book on the history of probability theory and statistics by Hald [3]. A number of authors on probability theory and its applications give lucid accounts of generating functions in probability theory and statistics at a level that asks for only a moderate background in mathematics. The books by Feldman & Fox [2] and Port [5] can be recommended. Stuart & Ord [7] give a thorough treatment of all kinds of moments and their generating functions. Wilf's [8] lively and lucid text on the mathematics of generating functions is aimed at advanced undergraduate and postgraduate students. Douglas [1] gives a full discussion of the relationships between the various types of moments and cumulants in the context of discrete distribution theory. The book by Johnson et al. [4] on discrete distribution theory gives the probability generating function a central role and demonstrates its many uses.

References

[1] Douglas, J.B. (1980). *Analysis with Standard Contagious Distributions*. International Co-operative Publishing House, Burtonsville.

[2] Feldman, D. & Fox, M. (1991). *Probability: The Mathematics of Uncertainty*. Marcel Dekker, New York.

[3] Hald, A. (1990). *A History of Probability and Statistics and their Applications Before 1750*. Wiley, New York.

[4] Johnson, N.L., Kotz, S. & Kemp, A.W. (1992). *Univariate Discrete Distributions*, 2nd Ed. Wiley, New York.

[5] Port, S.C. (1994). *Theoretical Probability for Applications*. Wiley, New York.

[6] Seal, H.L. (1949). The historical development of the use of generating functions in probability theory, *Bulletin de l'Association des Actuaires Suisses* **49**, 209–228.

[7] Stuart, A. & Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 1: *Distribution Theory*, 6th Ed. Edward Arnold, London; and Halsted Press, New York.

[8] Wilf, H.S. (1990). *Generatingfunctionology*. Academic Press, San Diego.

(*See also* **Contagious Distributions**)

ADRIENNE W. KEMP

Genetic Correlations and Covariances

Fisher [3] devised a theoretical basis for making inferences about genetic and environmental causes of variation in continuous traits. This provided a synthesis of Mendelian inheritance (*see* **Mendel's Laws**) with the Darwinian theory of evolution through selection, and overcame serious problems associated with the theory of blending inheritance, a concept that Darwin had relied on despite recognizing that it presented a major obstacle for his theory of evolution.

Definitions

Genetic Variances

Consider a genetic locus defined by two alleles, A_1 and A_2. Assume that a population is in **Hardy–Weinberg equilibrium** and that mating within this population occurs at random. Specifically, if the frequency of A_1 is p, the frequencies of A_1A_1, A_1A_2, and A_2A_2 are $p^2, 2p(1-p)$, and $(1-p)^2$, respectively.

Suppose this genetic locus is the only factor that causes variation in the trait Y, and let $Y = G = \mu_{ij}$ for individuals with **genotype** A_iA_j. The mean of Y, μ, is therefore given by $p^2\mu_{11} + 2p(1-p)\mu_{12} + (1-p)^2\mu_{22}$, and the variance, σ^2, by $p^2(\mu_{11} - \mu)^2 + 2p(1-p)\ (\mu_{12} - \mu)^2 + (1-p)^2(\mu_{22} - \mu)^2$. Now $\sigma^2 = \sigma_g^2$ can be decomposed as $\sigma_a^2 + \sigma_d^2$, where

$$\begin{aligned}\sigma_a^2 &= 2p(1-p)[p\mu_{11} + (1-2p)\mu_{12} - (1-p)\mu_{22}]^2 \\ &= 2p(1-p)[p(\mu_{11} - \mu_{12}) + (1-p) \\ &\quad \times (\mu_{12} - \mu_{22})]^2 \end{aligned} \tag{1}$$

is called the *additive* component of variance, and

$$\sigma_d^2 = \{p(1-p)[\mu_{11} - 2\mu_{12} + \mu_{22}]\}^2 \tag{2}$$

is called the *dominance* component of variance. Note that if $\mu_{12} = (1/2)(\mu_{11} + \mu_{22})$ – i.e. the heterozygote is the average of the two homozygotes – then $\sigma_d^2 = 0$.

These genetic variance components, σ_a^2 and σ_d^2, have an interpretation. If the values μ_{11}, μ_{12}, and μ_{22} are plotted against the number of A_2 alleles, and a straight line fitted by weighted **least squares** with weight proportional to frequency, the mean squared deviation about the mean μ is σ^2, and σ_a^2 is the variance "explained" by the straight line, in the usual **linear regression** sense. This line represents the effects of alleles under the additive assumption that the effect of two A_2 alleles is twice that of one A_2 allele; hence, σ_a^2 is called the additive genetic variance. Note that, by considering the ratio of σ_a^2 to σ_d^2, if either p or 1-p is small, then most of the genetic variance is additive.

The variance not explained by the linear association, and therefore attributed to nonlinear effects of alleles, is the residual sum of squares about the straight line, σ_d^2; R.A. Fisher called this the dominance variance [3].

Note that dominance variance and dominant inheritance do not refer to the same concept. If a trait takes just two values, and is solely determined by a single gene expressed in a dominant fashion, then $\mu_{12} = \mu_{22} \neq \mu_{11}$, say, and then $\sigma_a^2 = 2p^3(1-p)(\mu_{11} - \mu_{22})^2 \neq 0$, so that σ_d^2 does not account for the total variance.

Genetic Covariance between Relatives

Assume as above that the trait is determined solely by a single genetic locus, and consider now a pair of genetically related individuals, such as a mother and her son. By summing over the three possible genotypes of the father, the genetic covariance between mother and son is $p^3{\mu_{11}}^2 + 2p^2(1-p)\mu_{11}\mu_{12} + p(1-p){\mu_{12}}^2 + 2p(1-p)^2\mu_{12}\mu_{22} + (1-p)^3{\mu_{22}}^2 - \mu^2$, which can be shown to be equal to $1/2\sigma_a^2$. This same expression, $1/2\sigma_a^2$, holds for *all* parent–offspring pairs.

For sibling pairs, including dizygotic (DZ) twins (*see* **Zygosity Determination**), if the summation is made over all nine possible genotypes of mother and father under the assumption that maternal and paternal genotypes are independent, the genetic covariance between siblings is found to be $(1/2\sigma_a^2 + 1/4\sigma_d^2)$.

The above procedure can be extended to determine the genetic covariance between any pair of relatives. For second-degree relatives, such as grandparent and grandchild or uncle and nephew, the genetic covariance is $1/4\sigma_a^2$, and for third degree it is $1/8\sigma_a^2$, and so on, the covariance being multiplied by a factor of 1/2 for each extra generation. For genetically

identical (monozygotic, MZ) twin pairs, obviously the covariance is $\sigma^2 = \sigma_g^2 = \sigma_a^2 + \sigma_d^2$. Apart from appearing in expressions for the covariance between sibling (including twin) pairs, the term σ_d^2 occurs only when there is **inbreeding**. Nonindependence between parents, or **assortative mating**, induces extra covariation.

The **genetic correlation** is the genetic variance divided by the total variance.

Genetic Covariance in Terms of Identity Coefficients

Let G_i and G_j represent the genotype of two individuals i and j. In general, under Hardy–Weinberg equilibrium and no inbreeding (unless $\sigma_d^2 = 0$), the genetic covariance can be expressed as

$$\text{cov}(G_i, G_j) = 2\phi_{ij}\sigma_a^2 + K_{2ij}\sigma_d^2, \qquad (3)$$

where ϕ_{ij}, the kinship coefficient or coefficient of coancestry, is the probability that a **gene** drawn at random from a given locus of i is identical by descent with a gene drawn at random from the same locus of j, and K_{2ij} is the probability that both genes at a given locus are identical by descent in i and j. Under random mating, for MZ twin pairs, $\phi_{ij} = 1/2$. In the absence of inbreeding, $\phi_{ij} = 1/4$ for first-degree relatives (parent–offspring and sibling pairs), 1/8 for second-degree relatives (such as grandparent–grandchild), and so on, the coefficient being halved for each successive generation separating the pair. For MZ pairs, $K_{2ij} = 1$, and for sibling pairs, $K_{2ij} = 1/4$. Otherwise, $K_{2ij} = 0$, except for pairs such as double-first cousins, in which case, $K_{2ij} = 1/16$, and for offspring of related persons; for example, of sib-matings, in which case $K_{2ij} = 7/32$ (*see* **Identity Coefficients**).

Genetic Covariance under Polygenic Inheritance

Assume G now represents a **polygenic** factor, i.e. the combined effect of a number of genetic factors $G_i, i = 1, 2, \ldots, n$, at different loci. Let

$$G = G_1 + G_2 + G_3 + \cdots + G_1 \odot G_2 + G_1 \odot G_2 \odot G_3 + \cdots + G_1 \odot G_2 \odot \cdots \odot G_n, \qquad (4)$$

where each G_k has variance $\sigma_k^2 = \sigma_{ak}^2 + \sigma_{dk}^2$ as above, and the $\odot$ notation represents interaction between the loci (epistasis). There may be interactions between the additive components of loci, between the dominance components of loci, and between additive and dominance components of loci. It can be shown that in this case (3) must be extended by including a term for each interaction of a different type; see [1]. An interaction between pairs of additive components involves a variance component, σ_{aa}^2 say, multiplied by $(2\phi_{ij})^2$, and between triples of additive components a term $(2\phi_{ij})^3\sigma_{aaa}^2$, and so on. Similar comments apply to dominance components, with the coefficient K_{2ij} replacing $2\phi_{ij}$. These interaction components all contribute fully to the covariance between MZ twin pairs, because the coefficients $2\phi_{ij}$ and K_{2ij} are both unity, but contribute far less to the covariance for other relationships. Therefore, while dominance contributes to sibling covariance but not to parent–offspring covariance, epistasis can contribute to both parent–offspring and sibling covariances.

Provided the G_i are independent, which is a reasonable assumption for unlinked loci, the additive genetic variance is the sum of additive variances at each loci. The same holds for the total variance. In the absence of epistasis, the genetic correlations between relatives in terms of additive and dominance variance components are the same for a polygenic factor as for a single locus.

If G represents a large number of genetic loci, each having a small additive impact on the trait, multivariate **central limit theorems** for the distribution of G across groups of relatives have been derived, for example, [8]. The theorems have been proved under several fairly stringent sufficient genetic restrictions: Hardy–Weinberg and linkage equilibrium (*see* **Linkage Disequilibrium**) for all loci; absence of assortative mating and epistasis; a small variance for each locus compared with the total variance over many loci; an upper bound on the number of loci per chromosome; and no inbreeding if there is dominance variance at any locus. Verification of these, and other technical conditions, in practice would be difficult, if not impossible, but it is usually assumed that these theoretical restrictions are not of practical importance.

Environmental Variances

Factors that are independent between individuals irrespective of their relationship to one another, such as environmental effects specific to an individual and

measurement error, induce no covariance between relatives. If variation in Y is caused only by such nongenetic factors E_i and E_j for distinct individuals i and j, then

$$\text{cov}(E_i, E_j) = 0. \tag{5}$$

Let $\sigma_e^2 = \text{var}(E)$.

If variation in Y is caused only by a factor, C, that is shared by or common to (classes of) relatives, these relatives will be correlated. For example, if C_i and C_j represent the effects of environmental factors common to the household for distinct individuals i and j, then

$$\text{cov}(C_i, C_j) = c_{ij}\sigma_c^2, \tag{6}$$

where $c_{ij} = 1$ if i and j live in the same house, else 0, and $\sigma_c^2 = \text{var}(C)$. This cohabitation effect may be more sophisticated, perhaps depending on the type of relationship between cohabiting individuals, increasing the longer these individuals live together, and attenuating the longer they live apart. Some theoretical models for this have been proposed and applied [2, 5, 9].

Covariance Between Relatives in Terms of Genetic and Environmental Variances

The combined effects of genetic, common environmental, and individual specific factors, G, C, and E, respectively, on the variance of a trait Y can be modeled in numerous ways. The simplest models suppose that the effects of each component of variation are independent, i.e.

$$Y = \mu + G + C + E, \tag{7}$$

in which case

$$\text{var}(Y) = \sigma_g^2 + \sigma_c^2 + \sigma_e^2 = \sigma^2, \tag{8}$$

and

$$\text{cov}(Y_i, Y_j) = \text{cov}(G_i, G_j) + \text{cov}(C_i, C_j). \tag{9}$$

Each of the covariance terms in (9) can take various forms, depending on whether dominance or epistasis is included and the sophistication with which the common environment is modeled.

This general model can be extended to include **gene–environment interactions**, in which the effects of genetic factors are not independent of those of environmental factors; i.e. the effect of a particular genotype can depend on the environment. Extra term(s) for interaction effects need to be included in both the variance and covariance. In defining genetic correlations, when there are gene–environment interactions it is not clear what the denominator should be. Should the effect of genetic factors depend on environmental factors that change as an individual ages, the genetic variance will change with age.

The general model can also be extended to allow for gene–environment covariation, in which the distribution of the genetic and environmental factors is not independent; i.e. certain genotypes are more common in particular environments.

The Role of the Mean and Measured Factors in Genetic and Environmental Modeling

The discussion above has focused on causes of variation without reference to the presumed population mean about which this variation occurs. The variance, covariance and hence variance components (i.e. **random effects**) may differ considerably depending on what factors are taken into account in modeling the trait mean (**fixed effects**). For example, if the trait mean depends on age, then this will induce a covariance between siblings if they are generally of a similar age. If the mean is adjusted for age, the siblings will be less correlated than if the mean is unadjusted for age [6].

The absolute values of genetic and environmental components of variance may change as the effects of different factors are modeled in the mean. For example, for traits whose mean values depend on height, adjustment for height usually leads to a reduction in the genetic variance, provided age has also been taken into account. This is presumably because variation in height for age appears to be predominantly attributable to genetic causes [3].

The effects of genetic variation at a particular locus for which the individual genotypes have been measured can be modeled as either a fixed effect or a random effect [4]. In the fixed effects modeling, the mean may be assumed to change linearly with the number of copies of a certain allele the individual possesses; i.e. an additive genetic variance at this locus is being modeled (cf. (1)). The mean can also be

assumed to take a different value for each genotype, in which case dominance genetic variation at this locus is also being considered. In the random effects modeling, similar to the way common environmental effects are modeled using (6), individuals may be assumed to be more correlated if they both have the same genotype, or the same **haplotype** if alleles at several closely linked loci have been measured.

For a review of issues surrounding the variance components models and the methods involved in their estimation, see [7].

References

[1] Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford, pp. 126–131.

[2] Eaves, L.J., Long, J. & Heath, A.C. (1986). A theory of developmental change to quantitative phenotypes applied to cognitive development, *Behavior Genetics* **16**, 143–162.

[3] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[4] Hopper, J.L. & Mathews, J.D. (1982). Extensions to multivariate normal models for pedigree analysis, *Annals of Human Genetics* **46**, 373–383.

[5] Hopper, J.L. & Mathews, J.D. (1983). Extensions to multivariate normal models for pedigree analysis. II. Modelling the effect of shared environment in the analysis of variation in blood lead levels, *American Journal of Epidemiology* **117**, 344–355.

[6] Hopper, J.L. (1992). The epidemiology of genetic epidemiology, *Acta Geneticae Medicae et Gemellologiae* **41**, 261–273.

[7] Hopper, J.L. (1993). Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health, *Statistical Methods in Medical Research* **2**, 199–223.

[8] Lange, K. (1978). Central limit theorem for pedigrees, *Journal of Mathematical Biology* **6**, 59–66.

[9] Lange, K. (1986). Cohabitation, convergence, and environmental covariances, *American Journal of Medical Genetics* **24**, 483–491.

(*See also* **Familial Correlations**; **Heritability**; **Population Genetics**)

JOHN L. HOPPER

Genetic Counseling

Genetic counseling is the communication process by which individuals and their family members are given information about the diagnosis of the disease in question, quantitation of risk, and the implications of this genetic information. The ability to predict accurately an individual's disease risk is important at personal and population levels. At a personal level, knowledge of risk promotes informed health care decisions. At a population level, risk prediction enables targeted public health interventions in high-risk groups. The main contribution of statistics to genetic counseling is to provide a probabilistic framework for the estimation of the genetic risk.

Genetic counseling is generally concerned with the probability that an individual will develop a specific genetic disease. The risk for individual i is then defined as

$$\Pr(\text{individual } i \text{ will have disease } D) = \Pr(D_i).$$

By definition, a genetic disease is associated with a particular **genotype** (or genotypes). This association can be expressed as

$$\Pr(D_i) = \Pr(D|X, G_i)\Pr(G_i),$$

where X represents environmental factors, and G_i is the event that individual i has a disease genotype. Environmental factors are taken to include endogenous characteristics of the individual such as age or physiological characteristics, as well as exogenous exposures to compounds that may cause or modify risk to develop disease. The conditional probability, $\Pr(D|X, G_i)$, is referred to as the **penetrance** function of the disease.

In most cases, the genetic counselor's main concern is $\Pr(G_i)$, the probability that an individual has a disease genotype. Alternately, the concern might center on determining the probability that an individual carries a genotype consistent with the inheritance of a disease genotype by his or her progeny. This is referred to as *carrier risk*. When the determination of the individual's risk is not based on direct determination of his or her genotype (e.g. by genetic testing) but on the genotypes and disease occurrence in relatives, the disease risk $\Pr(D_i)$ is referred to as *recurrence risk*.

Genetic risks can be classified into three categories, namely population-, pedigree-, and individual-based, according to the type of data available and the analytical tools used in **risk assessment**.

Population-based Risk

In population-based risks (also referred to as empirical risks) the data are the frequencies of affected and unaffected individuals sampled from populations. These data include disease **incidence rates** in defined populations, often obtained from **case–control** or **cohort studies**. The distinguishing characteristic of population-based risk is that there is no attempt to quantify either the penetrance function or $\Pr(G_i)$.

An example of this category of risk estimation is a woman's lifetime breast cancer risk. This risk is often quoted to be 0.125 since, from US cancer registries (*see* **Disease Registers**), it has been observed that one in eight women develop breast cancer. This number is based on population estimates of lifetime cancer incidence, and does not reflect the personal risk factors of the individual (e.g. family or reproductive history). It is applicable only to women who belong to the sampled population. Epidemiologic studies can be conducted to identify endogenous and exogenous factors that may affect disease risk. These risk factors can be used to construct models for risk prediction. For example, Gail et al. [1] generated breast cancer risk estimates using information about age at menarche, number of previous biopsies, age at first live birth, and limited family history.

Pedigree-based Risk

Most frequently, risk estimates in genetic counseling are based on pedigree information. Traditionally, one assumes an underlying genetic model for the disease. The models may include parameters such as gene frequencies, transmission probabilities, probability of being a sporadic case, penetrance of genotypes, mutation rates and, in the case of a linked marker, recombination fractions, proportion of families for which the disease gene is linked to the marker, and gene order. In general, the unknown parameters of a given genetic model are replaced by point estimates obtained by **segregation analysis** or **linkage analysis, model-based**.

As an example, let us assume that unaffected parents have had one child affected with a rare single locus dominant, incompletely penetrant, autosomal disease. What is the probability that the couple's next child will be affected? In other words, what is the recurrence risk for the next child? The evaluation of this risk is based on two conditional probabilities, the probability that the next child is affected given that the affected sib (1) inherited the disease gene present in the parents, or (2) has it as a result of a *de novo* mutation. The recurrence risk is obtained from these conditional probabilities through an application of **Bayes' theorem**.

Individual-based Risk

Once a disease gene is identified and cloned, its presence or absence in an individual can be determined by biochemical assays. In this case, there is very little uncertainty about the individual's genotype. Consequently, the probability of disease is equal to the penetrance if the individual has the disease genotype, and is equal to zero otherwise.

As an example, the cloning of the hereditary breast cancer susceptibility gene BRCA1 allowed individual-based risk estimates to be produced using direct genetic mutation analysis. Breast cancer risk in an individual who carriers a mutant form of the BRCA1 gene can be estimated using the result of this test. These estimates suggest that women who carry germline BRCA1 mutations have a substantial risk of developing breast cancer during their lifetimes.

Interval Estimates of Risk

Traditionally, a point estimate of risk is communicated to the counselee. Methods currently exist to provide interval estimates (*see* **Estimation, Interval**) [2]. This is particularly important when risks are estimated from small samples.

As an illustration, suppose that a linked **genetic marker** is available, and no recombinants have yet been observed between the marker and the putative disease gene. Thus, the estimated recombination probability will be zero, and the risk estimate will also be zero. Although the risk for a person who does not inherit the marker allele of risk may be small, the genetic counselor will be reluctant to provide a risk estimate of zero, and an upper confidence limit of risk may be more desirable.

References

[1] Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schaiere, C. & Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, *Journal of the National Cancer Institute* **81**, 1879–1886.

[2] Rogatko, A. (1995). Risk prediction with linked markers: theory, *American Journal of Medical Genetics* **59**, 13–24.

A. ROGATKO, J. BABB & T. REBBECK

Genetic Distance

The concept of genetic distance relates to the measurement of genetic difference between populations. It is designed to answer how dissimilar the populations are with respect to their genetic compositions. Since the genetic composition of a population is well approximated by the allele frequencies at most loci (*see* **Hardy–Weinberg Equilibrium**), the genetic distance between two populations has been traditionally defined by quantifying the differences in the allele frequencies between them by a single number. In devising such a measure it is necessary to determine the central position of each population, so that the distance between populations can also take into account any within-population genetic variation that may exist. This leads to a geometric interpretation of the distance concept, where a distance function is expected to satisfy the three mathematical properties of a *metric*; i.e. the distance D_{ij} between populations i and j should be: (i) nonnegative ($D_{ij} \geq 0$, with $D_{ij} = 0$ if and only if $i = j$), (ii) symmetric ($D_{ij} = D_{ji}$), and (iii) $D_{ik} + D_{jk} \geq D_{ij}$ for any three populations i, j, and k (the triangle inequality). The first condition immediately implies that the definition of D_{ij} must take into account the within-population variation in each population.

Although the methods of detecting genetic variation have changed considerably over the history of biology, and even more changes are expected to occur through the advent of recombinant DNA technology (*see* **DNA Sequences**), genetic distance studies predate the discovery of **genetic markers**. Since the introduction of the first distance measure [9], a variety of genetic distance measures have been proposed and used for a variety of purposes, and a detailed account of these is beyond the scope of this article. There are many excellent reviews on this subject, e.g. [5, 10, 14, 21], and [31], several of which are still up to date. For the purpose of applications, these distance measures may be classified into two broad classes: (i) those intended for population classification, and (ii) those intended for the study of evolution. Nei [22] classified into the first category Czekanowski's mean difference [9] and its variation (the "Manhattan metric" [32]), Pearson's coefficient of racial likeness [26], Roger's distance [28], **Mahanalobis' distance** [18], Sanghvi's distance [29] and its variant [1], Kurczynski's D^2 [13], Bhattacharyya's [2], and Cavalli-Sforza & Edwards' [4] distances. Into the second category Nei grouped the distance indices of based on Wright's F_{ST} index (e.g. [3] and [15]), Morton's kinship indices [19], and his own distances [20]. Cockerham & Weir's [8] coancestry measure of distance (*see* **Inbreeding**) also falls into this second category. Although as a measure of genetic dissimilarity between populations this categorization has little importance, there is a clear difference when the evolutionary dynamics of **gene frequencies** are formulated in terms of defined evolutionary mechanisms (such as mutation, drift (*see* **Population Genetics**), and migration). The first category of distance indices does not generally show a well-defined pattern or trend, such as increasing with the evolutionary time of separation. In contrast, genetic distance indices of the second category have been studied in the context of specific models of evolution, so that their expected trend with the time since divergence between populations is fairly well described [5, 22, 33].

Even though the definitions of these distances are based on different premises, several of them are analytically related [5, 27, 33], and even those that are mathematically dissimilar yield fairly similar inferences regarding interpopulation relationships, at least for genetically close populations [6]. Dissimilarities of distance measures, however, emerge from their being estimated by the sampling of allele frequencies from populations. All distance functions involve quadratic expressions (and often their ratios) of population allele frequencies, and hence their estimates are biased. Analytical correction for bias is not an easy task, and can only be achieved either by approximations that hold under some strict conditions, or by numerical resampling methods (*see* **Bootstrap Method**). Furthermore, the total variance of an estimated distance consists of both contemporary sampling effects as well as cumulative effects of the past evolution of populations [5, 17, 24]. Recent work on the subject suggests that, with the molecular data currently being employed to study genetic variation between populations, it is necessary to include the concept of the coalescence time of alleles (i.e. how long in the past two alleles had a common ancestor) in a distance measure for it to be appropriate for evolutionary studies, for example, [12, 30]. The statistical properties of such distance measures are still to be explored rigorously [11].

Finally, in the human context, although genetic distance provides a fairly good indicator of evolution and dispersal of the species, any graphic display of interpopulation genetic distances should be interpreted with caution. Genetic variation by geographic or racial classification is a mere reflection of the discontinuity of sampled populations. In reality, genetic variation is gradual in time and is continuous in space; most of the variation is interindividual, and the interpopulation component of variation constitutes only a small portion of the total genetic diversity of the species [7, 16, 23, 25].

References

[1] Balakrishnan, V. & Sanghvi, L.D. (1968). Distance between populations on the basis of attribute data, *Biometrics* **24**, 859–865.

[2] Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations, *Sankhyā* **7**, 401–406.

[3] Cavalli-Sforza, L.L. (1969). Human diversity, in *Proceedings of the Twelfth International Congress on Genetics*, Vol. 3. Tokyo, pp. 405–416.

[4] Cavalli-Sforza, L.L. & Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures, *American Journal of Human Genetics* **19**, 233–237.

[5] Chakraborty, R. & Rao, C.R. (1991). Measurement of genetic variation for evolutionary studies, in *Handbook of Statistics*, Vol. 8. Elsevier, Amsterdam, pp. 271–316.

[6] Chakraborty, R. & Tateno, Y. (1976). Correlations between some measures of genetic distance, *Evolution* **30**, 851–853.

[7] Chakraborty, R., Jin, L. & Deka, R. (1994). Intra- and inter-population variation at short tandem repeat, polymarker, and VNTR loci and their implications in forensic and paternity analysis, in *Proceedings of the Fifth International Symposium on Human Identification*. Madison, pp. 29–41.

[8] Cockerham, C.C. & Weir, B.S. (1987). Correlations, descent measures: drift with migration and mutation, *Proceedings of the National Academy of Sciences*, **84**, 8512–8514.

[9] Czekanowski, J. (1909). Zur Differential diagnose der Neandertalgruppe, *Korrespondenz-Blatt Deutschen Gesellschaft für Anthopologie, Ethnologie and Urgeschichte* **40**, 44–47.

[10] Jorde, L. (1985). Human genetic distance studies: present status and future prospects, *Annual Review of Anthropology* **14**, 343–373.

[11] Kimmel, M. & Chakraborty, R. (1997). Measures of variation at DNA repeat loci under a general stepwise mutation model, *Theoretical Population Biology* **50**, to appear.

[12] Kimmel, M., Chakraborty, R., Stivers, D.N. & Deka, R. (1996). Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci, *Genetics* **143**, 549–555.

[13] Kurczynski, T.W. (1970). Generalized distance and discrete variables, *Biometrics* **26**, 525–534.

[14] Lalouel, J-M. (1980). Distance analysis and multidimensional scaling, in *Current Developments in Anthropological Genetics: Theory and Methods*, Vol. 1, J.H. Mielke & M.H. Crawford, eds. Plenum, New York, pp. 209–250.

[15] Latter, B.D.H. (1973). Measures of genetic distance, in *Genetic Structure of Populations*, N.E. Morton, ed. University Press of Hawaii, Honolulu, pp. 27–37.

[16] Lewontin, R.C. (1972). The apportionment of human diversity, *Evolutionary Biology* **6**, 381–398.

[17] Li, W.-H. & Nei, M. (1975). Drift variances of heterozygosity and genetic distance in transient states, *Genetic Research* **25**, 229–248.

[18] Mahanalobis, P.C. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Sciences of India* **2**, 49–55.

[19] Morton, N.E. (1975). Kinship, information and biological distance, *Theoretical Population Biology* **7**, 246–255.

[20] Nei, M. (1972). Genetic distance between populations, *American Naturalist* **106**, 283–292.

[21] Nei, M. (1978). The theory of genetic distance and evolution of human races, *Japanese Journal of Human Genetics* **23**, 341–369.

[22] Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

[23] Nei, M. & Roychoudhury, A.K. (1974). Genetic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids, *American Journal of Human Genetics* **26**, 421–443.

[24] Nei, M. & Roychoudhury, A.K. (1974). Sampling variance of heterozygosity and genetic distance, *Genetics* **76**, 379–390.

[25] Nei, M. & Roychoudhury, A.K. (1982). Genetic relationship and evolution of human races, *Evolutionary Biology* **14**, 1–59.

[26] Pearson, K. (1926). On the coefficient of racial likeness, *Biometrika* **18**, 337–343.

[27] Rao, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach, *Theoretical Population Biology* **21**, 24–43.

[28] Rogers, J.S. (1972). Measures of genetic similarity and genetic distance, in *Studies in Genetics*, Vol. VII. University of Texas, Austin, pp. 145–153.

[29] Sanghvi, L.D. (1953). Comparison of genetical and morphological methods for a study of biological differences, *American Journal of Physical Anthropology* **11**, 385–404.

[30] Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies, *Genetics* **139**, 457–462.

[31] Smith, C.A.B. (1977). A note on genetic distance, *Annals of Human Genetics* **40**, 463–479.
[32] Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.
[33] Weir, B.S. (1996). *Genetic Data Analysis*, Vol. II. Sinauer, Sunderland.

RANAJIT CHAKRABORTY

Genetic Epidemiology

By 1967 discussions among researchers had led to the realization that the merger of methods to analyze family data from mathematical genetics and statistical tools from epidemiology was both inevitable and desirable [9]. This merger was designated *genetic epidemiology*, for which Morton [9] proposed the following definition:

> genetic epidemiology: A science that deals with etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations.

The formal definition of the new field followed two decades of discussion dating back to Neel & Schull [10] defining "epidemiological genetics" in 1954. This dynamic time period saw the emergence of genetic epidemiology from the broader field of **population genetics** and the synthesis of several parallel developments in mathematics and statistics as they applied to human disease. But to understand the role of this relatively new field, its impact on **human genetics**, and its evolving definition in the post Human Genome Project era, it is necessary to review its history.

In the beginning, genetics was "genetic epidemiology". By stating the observation that an offspring receives one of two factors from each parent and has a 50% chance of passing each factor to its offspring, Gregor Mendel defined the probabilities that set the mathematical and statistical tone for the broad scientific discipline called genetics (*see* **Mendel's Laws**).

Mendel's original work also set the stage for the three characteristics that provide the cultural milieu for genetic epidemiology. First, scientists ignored Mendel's seminal discovery for nearly 50 years. The mathematical proofs, computational details, and statistical arguments required by genetic epidemiology leave most geneticists bored and/or frightened, preferring to ignore genetic epidemiology. Secondly, the question of whether Mendel's results were "too good" [5], foretold a field that relishes controversy over methods, interpretations, and applications. Such "family fights" prove stimulating to the investigators involved but have had the tendency to convince other geneticists that the field lacks focus and rigor. And finally, until very recently genetic epidemiology was a small field easily dominated by a few creative and powerful figures whose scientific differences and personal animosity made for exciting and stimulating arguments and intense scientific fads.

As genetics developed at the turn of the century, the research was concentrated in two areas:

1. defining phenotypes, which for humans is clinical genetics, at that time including rudimentary biochemical genetics; and
2. studying the mathematical properties of genes in populations, population genetics.

By the 1930s cytogenetics (the study of chromosomes) was flourishing, primarily in *Drosophila*, and has continued to develop as a major area of research. And most recently the field of molecular genetics has exploded in a wealth of research and knowledge leading to the initiation of the Human Genome Project in 1989.

Genetic epidemiology shares the use of mathematics and statistics as its primary tools with the field of population genetics, but its definition of a unique niche within genetics results from its interaction with all of the other areas. The history will be divided into three broad and, to some extent, overlapping categories:

1. population genetics;
2. Mendel's first law: segregation of alleles at one locus; and
3. Mendel's second law: independent assortment of two loci.

The Beginning (1): Population Genetics

The seminal beginning to population genetics was the definition of the Hardy-Weinberg Law in 1908 (*see* **Hardy–Weinberg Equilibrium**). In 1932 Snyder [13] applied this law to demonstrate the mode of inheritance for tasting phenylthiocarbamide (PTC). Snyder not only provided a clear example of the appropriateness of the Hardy–Weinberg equilibrium for the human population but also demonstrated a mathematical method for testing that a phenotype was inherited. This was both useful and restricted by the fact that is was limited to traits sufficiently common to allow random sampling of a large

number of families. This restriction of population genetic principles to common traits has limited their usefulness until recently.

Another significant issue addressed by early population genetic principles was the role of Mendelian factors in the inheritance of quantitative traits. Early investigators argued that segregating alleles could not possibly account for quantitative traits with Gaussian distributions (*see* **Normal Distribution**). In 1918 Fisher [4] demonstrated conclusively that "many small, equal, and additive loci" would result in exactly the Gaussian distribution for a phenotype. From that finding grew the entire field of quantitative genetics, including **heritability**, breeding factors, and other crucial insights for plant and animal breeders. Eventually these quantitative genetic principles also contributed to our understanding of human inheritance (*see* **Genetic Correlations and Covariances**).

Population genetics research continued to develop in several areas relevant to studies in genetic epidemiology, including describing the structure of populations (*see* **Genetic Distance**; **Admixture in Human Populations**; **Assortative Mating**; **Heterozygosity**). The principal distinction for these areas is that they apply to all populations and traits, thus providing insight for the study of human disease but are not restricted to the study of specific disease etiology.

The Beginning (2): Mendel's First Law

A second early and significant area of research developed to address statistical issues arising from the use of family data. Mendel began with pure breeding parents and tested the F1, F2, and backcross data. For human diseases that are *rare* in the population, the approach is to ascertain families where the disease is *known* to exist thereby avoiding a huge random sample which might capture little or no information. However, the use of **ascertainment** introduces immediate bias into the sample, as recognized as early as 1912 by Weinberg [14]. Although numerous investigators have proposed solutions to this problem, it continues as a serious concern for genetic epidemiologists to this day.

Data were analyzed for the presence of genes using the Weinberg and a priori forms of ascertainment correction until 1958 when Morton [8] proposed a **likelihood** approach to segregation analysis. The use of likelihood scoring to estimate parameters permitted incorporation of an ascertainment probability, proportion of sporadic cases, and other concepts of interest. It also provided a direct estimate of the **penetrance** and a **likelihood ratio test** for whether it differed from 1.0. The likelihood model much more closely approximated reality than the simpler approaches and was fairly widely applied (*see* **Segregation Analysis, Classical**).

By the 1960s, sufficient numbers of genetic loci had been identified in humans to establish a need for other forms of statistical analysis. For example, the procedure for paternity exclusion was refined and applied both scientifically and in legal situations (*see* **Paternity Testing**). Interest developed in models specifying the relationship among family members. Methods originally developed by the brilliant but not oft published Charles Cotterman [11] brought binary numbers and matrix algebra to the forefront (*see* **Identity Coefficients**; **Inbreeding**). An elegant modeling system, path analysis, incorporated the relationships among relatives and the possibility of environmental factors in order to determine the etiology of more complex traits. Models were developed to utilize twin data as special cases in an attempt to estimate more rigorously the importance of environmental factors (*see* **Path Analysis in Genetics**; **Twin Analysis**; **Twin Concordance**; **Adoption Studies**).

Also during the 1960s quantitative genetics made its presence felt in human genetics. Falconer [3] introduced the idea of a normally distributed, quantitative trait as the "liability" or "susceptibility" for a genetic disorder. This underlying trait was polygenic in nature, conforming to the work of Fisher cited above. When environmental factors were known to influence the trait, such as birth order effects, the model was called multifactorial to include both polygenic genetic effects and the environment. The disorder, however, was dichotomous in phenotype, such as the presence or absence of a birth defect, and the underlying quantitative trait led to the phenotype through theoretical "thresholds", leading to the multifactorial threshold models (MF/T). This breakthrough in thinking was followed by a burst of research and what amounted to a fad in genetic epidemiology, as every trait of unknown etiology was shown to be inherited in a multifactorial fashion. The bubble burst in 1972 when Reich et al. [12] pointed out the simple statistical fact that the model was indeterminate unless there were at least two

thresholds. Fortunately for many human phenotypes, e.g. birth defects, the presence of at least two thresholds was a simple matter since the frequency of the disorder differed markedly in the two sexes (*see* **Polygenic Inheritance**).

The popularity of the multifactorial models helped genetic epidemiologists realize that many of the diseases and disorders of interest would have much more complex etiologies than single locus inheritance. But few if any traits survived a rigorous test of MF/T. The popularity of the MF/T model declined and was replaced by the concept of genetic heterogeneity.

The realization that etiologies were complex dictated the development of more mathematically sophisticated models and statistical tests. For example, until the development of the "mixed model", the presence of single locus inheritance and the presence of multifactorial inheritance were tested as separate hypotheses that did not contain the same parameter space, that of an overarching general model, and so were not *true* alternative hypotheses. The "mixed model" was so named because it defined that necessary general model and led directly to the more rigorous testing of the possible modes of inheritance (*see* **Segregation Analysis, Mixed Models**).

A split had developed in the field, however, and a serious controversy raged over the question of whether to use nuclear family data (parents and their offspring) or larger extended pedigrees for testing genetic hypotheses. As with the other previous disputes, the field experienced a surge forward as models were developed both for nuclear families and for extended pedigrees. There were two primary developments in the area of pedigree analysis. First the publication of the **Elston–Stewart algorithm** [2] provided an efficient, recursive mathematical approach to evaluating the likelihood over an extended family. Its importance to all of genetics was recognized with the presentation of the William Allan Award by the American Society of Human Genetics to Elston in 1996. (The first William Allan Award was presented in 1962 to Newton Morton for segregation analysis and the introduction of lod scores to linkage analysis.) The second development was the introduction of *transmission probabilities*. The use of these parameters expanded the ability to test genetic models to determine whether they fit the data best while remaining within the constraints of Mendel's first law. Which form of data to use, nuclear families or pedigrees, was resolved methodologically by a reformulation of both approaches to incorporate the parameters of the other and the definition of a single "complex segregation" analysis approach (*see* **Segregation Analysis, Complex**).

Two additional major developments were to occur in the arena of Mendel's first law. The first of these was the introduction, through a series of papers, of **regressive models** by Bonney and his colleagues. These models, the most complex to date, incorporate numerous **confounding** factors, such a cohort effects (*see* **Age–Period–Cohort Analysis**). However, by virtue of their complexity and the number of parameters involved, they require huge data sets and are often applied in more restricted forms.

The second remaining development was the merger of segregation analysis with linkage analysis, but first it is necessary to discuss linkage analysis itself.

The Beginning (3): Mendel's Second Law

The phenomenon of linkage, i.e. the violation of the independent assortment of Mendel's second law, was first studied extensively in experimental organisms, primarily *Drosophila* and mouse. The concepts of linkage groups, recombination frequency, genetic distance, and mapping functions (*see* **Genetic Map Functions**) were all observed, defined, and/or analyzed throughout the first half of this century. Linkage has proven to be a powerful genetic tool. In spite of this power, very little was done in humans before 1950 because so few marker loci existed. The first linkage in humans was demonstrated by Mohr [6] in 1954. About this time, however, the use of new laboratory techniques to identify genes (for example, electrophoresis) began and is still going on with the molecular biology revolution.

The breakthrough in human linkage studies required not only the means to develop marker loci but also the statistical tools to analyze the data. Studies with experimental organisms are done by using fixed mating schemes and counting the recombinants in the offspring. For humans the two restricting factors were unknown phase (which of the two marker alleles was on the chromosome with the disease allele) of the parents and the small family size. To overcome these obstacles the data had to

be pooled, but pooling matings of opposite phase would result in apparent independent assortment. Therefore each small family had to be analyzed and *then* pooled. By extending the principles of sequential sampling (*see* **Pedigrees, Sequential Sampling**), Morton [7] defined lod scores for linkage analysis in humans (*see* **Linkage Analysis, Model-based**). Over the course of the past 40 years untold numbers of lod scores have been calculated – at the beginning by hand [7] and later by a plethora of computer algorithms (*see* **Software for Genetic Epidemiology**).

The observation that insulin dependent diabetes mellitus type 1, IDDM, showed a strong association with alleles of the **HLA system** on chromosome 6 (*see* **Disease-marker Association**) defined a new situation for linkage analysis. The question debated extensively was whether this association represented **linkage disequilibrium** between HLA markers and an IDDM locus or a susceptibility role for the HLA alleles. The problem was further complicated by the nature of the HLA multi-locus region which contained numerous haplotypes in disequilibrium (*see* **Haplotype Analysis**). If IDDM were solely genetic, then reduced penetrance would have to be invoked when analyzing unaffected individuals. This problem stimulated the development of methods to detect linkage that were directed towards using only affected individuals (*see* **Linkage Analysis, Model-free**). Recent publications have independently demonstrated that the "model-free" methods can be considered as subsets of the model-based methods and by doing so have raised a set of statistical questions for genetic epidemiologists to resolve regarding application and interpretation of the different methods.

The rapid explosion in the development of the human genome linkage map led to the necessity for methods that could analyze multiple loci simultaneously. These methods (*see* **Linkage Analysis, Multipoint**) provide additional precision in determining the actual location of purported disease loci.

Likelihood models were derived for simultaneously estimating the segregation parameters and the linkage relationships for a disease locus (loci). More recently, alternative approaches of using linkage information alone to determine the mode of inheritance have been proposed.

In the End...

In summary, it is important to emphasize three points. First it is impossible in any overview to include all of the important factors that contributed to the development of a scientific discipline. Therefore, this discussion slighted some developments and emphasized others. The second point is to restate the long-term goal of genetic epidemiology: to understand the causes of genetic disease in humans. To this end, the research efforts must be translated into terms that can be discussed with patients in families at risk (*see* **Genetic Counseling**) and that leads to the formulation of the next set of research questions. Neither of these end points is simple to derive from the often convoluted results of analyzing complex disorders in less-than-ideal data.

And finally, the field of genetic epidemiology has been responsive to the need for interaction and the exchange of ideas. The journal *Genetic Epidemiology* was established in 1984 and the International Genetic Epidemiology Society was founded in 1991, adopting the journal as its official publication. The depth of methodological disagreements and their effect on the genetic community led directly to the establishment of the Genetic Analysis Workshop (GAW) series, the most recent being GAW 10 in 1996 [1]. These workshops provide a unique atmosphere where methods are compared and evaluated in a controlled workshop format.

The next challenge for the field of genetic epidemiology is to develop new approaches to utilize the vast amounts of information that will become available as the human genome is sequenced and to apply these techniques to the most common and complex of disorders rigorously and imaginatively.

References

[1] Bailey-Wilson, J.E., Borecki, I.B., Falk, C.T., Goldstein, A.M., Suarez, B.K. & MacCluer, J.W. (1997). Proceedings of Genetic Analysis Workshop 10, *Genetic Epidemiology* **14**, to appear.

[2] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[3] Falconer, D.S. (1965). The inheritance of liability of certain diseases, estimates from the incidence among relatives, *Annals of Human Genetics* **29**, 51–76.

[4] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[5] Fisher, R.A. (1936). Has Mendel's work been rediscovered?, *Annals of Science* **1**, 115–137.
[6] Mohr, J. (1954). *A study of Linkage in Man*. Munkegaard, Copenhagen.
[7] Morton, N.E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.
[8] Morton, N.E. (1958). Segregation analysis in human genetics, *Science* **127**, 79–80.
[9] Morton, N.E. (1982). *Outline of Genetic Epidemiology*. Karger, New York.
[10] Neel, J.V. & Schull, W.J. (1954). *Human Heredity*. University of Chicago Press, Chicago.
[11] Optiz, J.S. (1983). Cotterman and combinatorial genetics, *American Journal of Medical Genetics* **16**, 389–392.
[12] Reich, T., James, S.W. & Morris, C.A. (1972). The use of multiple thresholds in determining the mode of transmission of semi-continuous traits, *Annals of Human Genetics* **36**, 163–184.
[13] Snyder, L.H. (1932). Studies in human inheritance. IX. The inheritance of taste deficiency in man, *The Ohio Journal of Science* **32**, 436–440.
[14] Weinberg, J. (1912). Weitere Beitrage zur Theorie der Vererbung. 4. Uber Methode und Feglerquellen der Untersuchung auf Mendelsche Zahlen beim Menschen, *Arch. Rass. U. Geg. Biol.* **9**, 165–174.

M.A. SPENCE

Genetic Heterogeneity

When a trait has a genetically different etiology in different individuals, that trait is said to be genetically heterogeneous. Two major types of genetic heterogeneity are identified: allelic heterogeneity and locus heterogeneity. With allelic heterogeneity, different alleles at the same locus confer the same phenotype in different individuals. An example of allelic heterogeneity is the cystic fibrosis **gene** [17]. About 75% of the chromosomes in cystic fibrosis patients carry the ΔF_{508} mutant allele; the remaining chromosomes each carry one of a large number of different mutant alleles of the same gene. With locus heterogeneity, the phenotype in different individuals is due to different loci. A classic example is autosomal recessive albinism, which can be caused by a mutant allele in one of at least two different loci. A report by Trevor-Roper [34] of four normal offspring from the mating of two affected parents illustrates this concept. An earlier description of a disease exhibiting locus heterogeneity is given by Morton [20] for elliptocytosis.

When the phenotype (*see* **Genotype**) is relatively rare and the rate of new **mutations** is low, affected individuals in the same family (almost always) carry the same mutant allele. As a result, a set of families can, in theory, be grouped according to which mutant allele they carry. If the genes have not yet been identified, then the usual first step is to determine, using linkage analysis and additional positional cloning methods, the loci involved. Statistical analyses that allow for locus heterogeneity can be used to improve the power of **linkage analysis** and to group families according to the locus causing the phenotype. Once the loci have been identified, different mutant alleles within each locus can be identified using molecular methods and studied using statistical methods that model allelic heterogeneity.

If the phenotype is common, then affected individuals within a family may not carry the same mutant allele and families may not be as easily grouped. Nonetheless, the same overall plan may be followed (although researchers may instead opt to investigate specific **candidate genes** or use **linkage disequilibrium** methods to search for candidate loci). We consider primarily the problem of locus heterogeneity in the context of linkage analysis, since a substantial body of statistical work is available in this area, although locus heterogeneity is an equally important issue for other gene mapping methods.

Locus Heterogeneity

A disease exhibits locus heterogeneity if alleles at more than one locus confer susceptibility to the disease. Locus heterogeneity adversely affects the power of linkage analysis because the sample of families to be analyzed is in fact a mixture of families with different underlying genetic architectures. This is a particular problem for mapping complex diseases, so named because of their likely considerable degree of locus heterogeneity. Mapping **complex diseases** remains one of the biggest challenges in gene mapping. Power can be improved by employing a statistical model that allows for locus heterogeneity, either by representing the **likelihood** as a mixture likelihood (in the absence of additional phenotypic information) or by incorporating additional phenotypic information into the likelihood.

Let θ denote the recombination fraction between disease and **marker** loci. Consider a set of families in which a subset of the families is linked ($\theta < \frac{1}{2}$) to a locus denoted A and the remaining families are unlinked ($\theta = \frac{1}{2}$) to locus A. The disease in families unlinked to locus A may be caused by one or more unlinked loci or by environmental factors. Smith [31, 32] proposed the mixture likelihood

$$L_i(\alpha, \theta) = \alpha L_i(\theta) + (1 - \alpha) L_i\left(\tfrac{1}{2}\right), \qquad (1)$$

where $L_i(\theta)$ is the likelihood for family i and α is a mixture parameter. (*See* **Linkage Analysis, Model-based** for a description of likelihood models for linkage analysis.) A **likelihood ratio test** of homogeneity, given linkage, is given by

$$\chi^2 = 2[\ln L(\hat{\alpha}, \hat{\theta}) - \ln L(1, \hat{\theta})], \qquad (2)$$

where $L(\alpha, \theta) = \prod_i L_i(\alpha, \theta)$. The test statistic has asymptotically a distribution that is a 50:50 mixture of a point mass at zero and a chi-square with one degree of freedom (df). This test of homogeneity has been termed the A-test, with A representing "admixture".

Likelihood ratio tests of linkage in the presence of heterogeneity can also be constructed using the admixture likelihood:

$$\chi^2 = 2\left[\ln L(\hat{\alpha}, \hat{\theta}) - \ln L\left(1, \tfrac{1}{2}\right)\right]. \qquad (3)$$

The asymptotic distribution of this statistic is more difficult to determine than the test of homogeneity, because under the **null hypothesis** of no linkage, α is undetermined. A conservative approximation compares the likelihood ratio statistic to a χ^2 with 2 df. Faraway [7] showed that the asymptotic distribution is more accurately approximated as the maximum of two χ_1^2 deviates. More recently, Chiano & Yates [4] used a reparameterization to determine an approximate asymptotic distribution and recommended a critical lod score of 3.44 to conclude **genome-wide significance**.

Given estimates of α and θ, the conditional probability that family i is of the linked type can be estimated using

$$w_i(\hat{\alpha}, \hat{\theta}) = \frac{\hat{\alpha} L_i(\hat{\theta})}{\hat{\alpha} L_i(\hat{\theta}) + (1 - \hat{\alpha}) L\left(\frac{1}{2}\right)}. \qquad (4)$$

When α and θ are known, $w_i > \frac{1}{2}$ is an optimal classification rule to group families of the linked type [12]; when α and θ are estimated, the classification rule $w_i(\hat{\alpha}, \hat{\theta}) > \hat{\alpha}$ is generally more reliable in small samples [22]. Power to detect heterogeneity has been considered by other authors [2, 23].

Several generalizations of the admixture likelihood have been proposed. Ott [24, 25] describes a mixture model for two linked loci in which two recombination fractions and the heterogeneity parameter are estimated. A more complex mixture model, also described in [25] (see also [33]) specifies a mixture of three distributions, two of which are unlinked disease loci for which marker data are available and one of which represents a hypothetical third locus (or other factor) unlinked to the first two. These extensions are available in the HOMOG programs (see [25]). Vieland et al. [35] extended the admixture model to the combined analysis of multiple data sets by allowing the admixture parameter to vary across data sets and confirmed that this "compound" lod score, denoted HLOD-C, has better power in the presence of locus heterogeneity than HLOD alone or model-free linkage methods [16].

MacLean et al. [18] proposed the C-test, given by

$$C = \sum_{i=1}^{n} \max_{\theta} Z_i(\theta), \qquad (5)$$

where the sum is over all pedigrees in the sample and $Z_i(\theta)$ is the lod score for family i. MacLean et al. claimed that the C-test is much more powerful than the A-test, but this claim has been disputed [8, 19]. One can also fit two-locus (or more generally, **multilocus**) models to investigate locus heterogeneity. Goldin [11] compared the power of some two-locus linkage models with the admixture model. She found that the two-locus model was more powerful when all the families in the sample are segregating for both loci, but that this increase in power requires correct specification of the parameters of the genetic model, and as a result preferred the admixture lod score in real-data situations.

It is generally held that allowing for locus heterogeneity in the analysis improves power for detecting linkage for rare Mendelian disorders. For small numbers of families, the admixture parameter may not be well estimated. Chiano & Yates [3] proposed a **bootstrap** approach to estimating linkage parameters in the presence of locus heterogeneity. For common complex diseases, the problem of locus heterogeneity is more complex. Durner et al. [6] explored whether a phenocopy frequency can approximate or model locus heterogeneity and found that, in general, assuming a positive phenocopy frequency does not compensate for the presence of an unlinked form of genetic disease. Whittemore & Halpern [36] discuss the limitations of estimating the admixture proportion when the disease is not rare and when the underlying genetic model is complex. They argue that, for common diseases, the admixture likelihood relies on unverifiable assumptions (including correct specification of a phenocopy rate) that, if violated, may decrease power to detect linkage.

Additional phenotypic information can be used to help classify families into linked and unlinked types. A simple and common approach is the separate analysis of subgroups of families, particularly when families can be grouped according to some categorical variable. An alternative test, the M-test [20], can be applied when families can be grouped into c classes based on additional phenotypic information. The likelihood is maximized for each class separately ($i = 1, \ldots, c$) and compared with the likelihood maximized over the entire sample. The test statistic is written as

$$M = 2\ln(10)\left[\sum_i Z_i(\hat{\theta}_i) - Z(\hat{\theta})\right]. \qquad (6)$$

Asymptotically (i.e. as the number of opportunities for recombination in each class goes to infinity)

and under the assumption of homogeneity, the M statistic follows a χ^2 distribution with $c - 1$ df. The M-test requires estimation of a separate recombination fraction in each class. A more powerful approach based on a **hierarchical model** and termed the B-test (with B representing **Bayesian**) was proposed by Risch [28]. For this test, the recombination fractions are assumed to follow a beta distribution, the parameters of which are estimated from the posterior distribution of θ. The resulting likelihood ratio test is compared against a χ^2 with 1 df and is more powerful than the M-test.

Regression methods have been proposed recently that allow discrete and continuous phenotypes to be incorporated as **covariates**. A generalization of the admixture likelihood allows the mixture parameter α to be a function of covariates [30]. Let

$$\alpha = \frac{\exp(\gamma + \beta^{\mathrm{T}} x)}{1 + \exp(\gamma + \beta^{\mathrm{T}} x)}, \tag{7}$$

where x is a vector of covariates, γ is an intercept parameter, and β is a vector of regression parameters. Similar problems in determining the distribution of overall tests of linkage exist for this model as for other mixture likelihoods.

Some covariates, such as age-at-onset, can also be incorporated as part of the disease **penetrance** functions. When the associated parameters are estimated previously and fixed in the linkage analysis, such penetrance functions do not affect the null distribution of the linkage statistic. Alternatively, parameters in the penetrance function may be estimated as part of the linkage analysis. One method that incorporates a dichotomous covariate into an admixture heterogeneity model was proposed by Houwing-Duistermaat et al. [15], who proposed a mixture of four likelihoods that represent all combinations of the two-locus locations and two binomial covariate parameters.

For model-free linkage analysis (*see* **Linkage Analysis, Model-free**), the usual approach has been to subset the data and analyze each subset separately. More recently, several methods that better utilize or incorporate additional phenotypic information have been proposed. One method utilizes linkage information to optimally classify families [27]. Greenwood & Bull [13] incorporated covariates into the Risch [29] affected-sib-pair likelihood using a multinomial parameterization. A reparameterization in terms of genetic **risk ratios** was proposed by Olson [21]; this model also allows for covariates. A further modification with fewer parameters was later proposed by Goddard et al. [10]. Another method that incorporates a binary environmental exposure variable was introduced by Gauderman & Siegmund [9]. Most recently, Devlin et al. [5] propose a mixture model analogous to the A-test. In general, these new methods allow for covariate-related locus heterogeneity and, with an appropriate covariate, can greatly enhance the ability of model-free linkage methods to detect linkage to complex diseases.

Allelic Heterogeneity

Once a disease locus has been identified, molecular work is needed to identify the particular mutant alleles that contribute to disease susceptibility. Some methods are useful when a few mutant alleles (**polymorphisms**) of interest are common in the population of interest. In general, investigation of relationships between common mutations and disease characteristics can be carried out using standard biostatistical methods in random or selected samples of individuals or families. Another direction of study with its own methodology aims at understanding the underlying **population genetics**: the processes that govern mutation, selection, and the structures of modern populations. Only a few specialized statistical methods have been proposed for investigating extensive allelic heterogeneity (when, say, 100 mutational forms are known). One such class of methods is the analysis of mutational spectra to identify "hot spots" within a gene more likely to be mutated in affected individuals [1, 26]. Such studies aim to understand the structure and function of the gene (and resulting protein) at a molecular level (e.g. [14]).

References

[1] Adams, W.T. & Skopek, T.R. (1987). Statistical test for the comparison of samples from mutational spectra, *Journal of Molecular Biology* **194**, 391–396.

[2] Cavalli-Sforza, L.L. & King, M.-C. (1986). Detecting linkage for genetically heterogeneous disease and detecting heterogeneity with linkage data, *American Journal of Human Genetics* **38**, 599–616.

[3] Chiano, M.N. & Yates, J.R.W. (1994). Bootstrapping in human genetic linkage, *Annals of Human Genetics* **58**, 129–143.

[4] Chiano, M.N. & Yates, J.R.W. (1995). Linkage detection under heterogeneity and the mixture problem, *Annals of Human Genetics* **59**, 83–95.

[5] Devlin, B., Jones, B.L., Bacanu, S.-A. & Roeder, K. (2002). Mixture models for linkage analysis of affected sibling pairs and covariates, *Genetic Epidemiology* **22**, 52–65.

[6] Durner, M., Greenberg, D.A. & Hodge, S.E. (1996). Phenocopies versus genetic heterogeneity: can we use phenocopy frequencies in linkage analysis to compensate for heterogeneity?, *Human Heredity* **46**, 265–273.

[7] Faraway, J.J. (1993). Distribution of the admixture test for the detection of linkage under heterogeneity, *Genetic Epidemiology* **10**, 75–83.

[8] Faraway, J.J. (1994). Testing for linkage under heterogeneity: A test versus C test, *American Journal of Human Genetics* **54**, 563–564.

[9] Gauderman, W.J. & Siegmund, K.D. (2001). Gene-environment interaction and affected sib pair linkage analysis, *Human Heredity* **52**, 34–46.

[10] Goddard, K.A.B., Witte, J.S., Suarez, B.K., Catalona, W.J. & Olson, J.M. (2001). Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4, *American Journal of Human Genetics* **68**, 1197–1206.

[11] Goldin, L.R. (1992). Detection of linkage under heterogeneity: comparison of the two-locus vs. admixture models, *Genetic Epidemiology* **9**, 61–66.

[12] Goldstein, M. & Dillon, W.R. (1978). *Discrete Discriminant Analysis*. Wiley, New York.

[13] Greenwood, C.M.T. & Bull, S.B. (1999). Analysis of affected sib pairs, with covariates–with and without constraints, *American Journal of Human Genetics* **64**, 871–885.

[14] Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C.C. (1991). P53 mutations in human cancer, *Science* **253**, 49–53.

[15] Houwing-Duistermaat, J.J., Sandkuijl, L.A., Bergen, A.A.B. & van Houwelingen, H.C. (1995). Maximum-likelihood estimated in linkage heterogeneity models including additional information via the EM algorithm, *Genetic Epidemiology* **12**, 515–527.

[16] Huang, J. & Vieland, V.J. (2000). Comparison of "model-free" and "model-based" linkage statistics in the presence of locus heterogeneity: single data set and multiple data set applications, *Human Heredity* **51**, 217–225.

[17] Kerem, B.J., Rommens, M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. & Tsui, L.-C. (1989). Identification of the cystic fibrosis gene: genetic analysis, *Science* **245**, 1073–1080.

[18] MacLean, C.J., Ploughman, L.M., Diehl, S.R. & Kendler, K.S. (1992). A new test for linkage in the presence of locus heterogeneity, *American Journal of Human Genetics* **50**, 1259–1266.

[19] MacLean, C.J., Sham, P.C., Ploughman, L.M., Diehl, S.R. & Kendler, K.S. (1994). Reply to Faraway, *American Journal of Human Genetics* **54**, 564–567.

[20] Morton, N.E. (1956). The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type, *American Journal of Human Genetics* **8**, 80–96.

[21] Olson, J.M. (1999). A general conditional-logistic model for affected-relative-pair linkage studies, *American Journal of Human Genetics* **65**, 1760–1769.

[22] Ott, J. (1983). Linkage analysis and family classification under heterogeneity, *Annals of Human Genetics* **47**, 311–320.

[23] Ott, J. (1986). The number of families required to detect or exclude linkage heterogeneity, *American Journal of Human Genetics* **39**, 159–165.

[24] Ott, J. (1990). Genetic interpretation of disease clustering, in *Convergent Issues in Genetics and Demography*, J. Adams, D.A. Lam, A.I. Hermalin & P.E. Smouse, eds. Oxford University Press, New York, pp. 245–255.

[25] Ott, J. (1991). *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

[26] Piegorsch, W.W. & Bailer, A.J. (1994). Statistical approaches for analyzing mutational spectra: some recommendations for categorical data, *Genetics* **136**, 403–416.

[27] Province, M.A., Shannon, W.D. & Rao, D.C. (2001). Classification methods for confronting heterogeneity, *Advances in Genetics* **42**, 273–286.

[28] Risch, N. (1988). A new statistical test for linkage heterogeneity, *American Journal of Human Genetics* **42**, 353–364.

[29] Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs, *American Journal of Human Genetics* **67**, 1014–1019.

[30] Schaid, D.J., McDonnell, S.K. & Thibodeau, S.N. (2001). Regression models for linkage heterogeneity applied to familial prostate cancer, *American Journal of Human Genetics* **68**, 1189–1196.

[31] Smith, C.A.B. (1961). Homogeneity test for linkage data, *Proceedings of the Second International Congress of Human Genetics* **1**, 212–213.

[32] Smith, C.A.B. (1963). Testing for homogeneity of recombination fraction values in human genetics, *Annals of Human Genetics* **27**, 175–182.

[33] Sykes, B., Ogilvie, D., Wordsworth, P., Wallis, G., Mathew, C., Beighton, P., Nicholls, A., Pope, F.M., Thompson, E., Tsipouras, P., Schwartz, R., Jensson, O., Arnason, A., Borresen, A.-L., Heiberg, A., Frey, D. & Steinmann, B. (1990). Consistent linkage of dominantly inherited osteogenesis imperfecta to the type I collagen loci: COL1A1 and COL1A2, *American Journal of Human Genetics* **46**, 293–307.

[34] Trevor-Roper, P.D. (1952). Marriage of two complete albinos with normally pigmented offspring, *British Journal of Ophthalmology* **36**, 107–110.

[35] Vieland, V.J., Wang, K. & Huang, J. (2001). Power to detect linkage based on multiple sets of data in the present of locus heterogeneity: comparative evaluation of model-based linkage methods for affected sib pair data, *Human Heredity* **51**, 199–208.
[36] Whittemore, A.S. & Halpern, J. (2001). Problems in the definition, interpretation, and evaluation of genetic heterogeneity, *American Journal of Human Genetics* **68**, 457–465.

JANE M. OLSON

Genetic Liability Model

Many common diseases (e.g. hypertension, coronary artery disease and noninsulin-dependent diabetes mellitus) exhibit strong familial tendencies but classical **segregation analysis** fails to detect a simple Mendelian pattern of inheritance (*see* **Mendel's Laws**). The thorough analysis of these traits requires the formulation of models that incorporate both genetic and environmental sources of **familial correlations**. Two widely used models, the *multifactorial threshold model* [1–6, 9, 12–14, 22, 23, 28] and the *mixed model* [11, 17, 20] posit that an individual's liability to develop a disease results from the additive effects of many genetic and environmental factors. The probability of expressing the disorder is then modeled as a function of this latent liability. Both models assume that liability is continuous. In addition, the mixed model allows for the existence of a major locus that measurably alters an individual's liability.

Let l denote liability and D denote the affection status of an individual and assume, for the moment, that the trait exists in only two forms, *normal* ($D = 0$) and *affected* ($D = 1$). Under both the multifactorial and the mixed models, the probability that an individual expresses the disease is given by

$$P(D = 1) = \int_{-\infty}^{+\infty} f(l)S(l)\,\mathrm{d}l,$$

where $f(\cdot)$ is the probability distribution function of the liability and $S(\cdot)$ is a risk function, i.e. the conditional probability of expressing the disease given l. The joint probability of the affection statuses of n related individuals, $D_1, \ldots, D_n$, is

$$P(D_1, \ldots, D_n) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_n(l_1, \ldots, l_n) \times \prod_{i=1}^{n} S(l_i)^{D_i}[1 - S(l_i)]^{1-D_i}\,\mathrm{d}l_1, \ldots, \mathrm{d}l_n.$$

Under the multifactorial model, $f_n(\cdot)$ is the joint distribution of n correlated continuous variables. Under the mixed model, $f_n(\cdot)$ is a mixture of distributions whose weights (mixing proportions) are the probabilities of the **genotypes** at the major locus. There is little biological rationale underlying the choice of $S(\cdot)$. Any convenient risk function can be applied as long as it provides a good fit to the data when used in conjunction with $f_n(\cdot)$. For example, Eaves [7] uses the well-known logit transformation, $S_{\mathrm{L}}(l) = [1 + \exp(-l)]^{-1}$, in a study of **gene–environment interactions**. From a statistical perspective, there is nothing to distinguish this particular form of genetic analysis from any other statistical method designed to model a polychotomy (*see* **Polytomous Data**) as a function of continuous variables. The technical details of **estimation** will be addressed only briefly here.

Risk Functions

Since liability was assumed to be **normally distributed**, most early work concentrated on the appropriate form of the risk function; that is, on the nature of a "threshold". Falconer [12, 13] assumed that the disorder is expressed only when the liability exceeds a physical threshold, T, so that

$$S_{\mathrm{F}}(l) = \begin{cases} 1, & l > T, \\ 0, & l \leq T. \end{cases}$$

Edwards [8, 9], however, argued that the concept of a physical threshold is biologically implausible: an occasional individual with a very high liability may escape the disease; conversely, one with a very low liability may not. Accordingly, Edwards proposed to let the risk be a monotonically increasing function of l and suggested the use of

$$S_{\mathrm{E}}(l) = a\exp(bl)$$

for $a > 0$ and $b > 0$. $S_{\mathrm{E}}(\cdot)$ is particularly advantageous when the liability is normally distributed, since the conditional distribution of liability in affected individuals is still normal, rather than truncated normal as would be the case under $S_{\mathrm{F}}(\cdot)$. However, $S_{\mathrm{E}}(\cdot)$ may exceed unity for very large values of l and thus cannot be used to represent conditional probabilities. This observation led Curnow & Smith [5, 6, 23] to propose an alternative risk function,

$$S_{\mathrm{CS}}(l) = \mathbf{\Phi}\left[\frac{(l - \lambda)}{\xi}\right],$$

in which $\mathbf{\Phi}(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The sensitivity of the phenotype to the underlying liability is measured by $(1/\xi)$. For any given liability, higher values of $(1/\xi)$ yield higher probabilities of affection.

The parameter λ is equivalent to the *median lethal dose* (*see* **Median Effective Dose**) in toxicology, i.e. the value of the liability at which the probability of being affected is 50%. By appropriate definition of liability, $S_F(\cdot)$ and $S_{CS}(\cdot)$ can lead to mathematically identical expressions for $P(\mathbf{D})$ when the joint distribution of the liabilities is **multivariate normal** [5]. Moreover, since liability is not directly observed, the parameters ξ and λ will be **confounded**. For these two reasons $S_F(\cdot)$ is the most commonly used risk function.

Multifactorial Models

Under the multifactorial model (Figure 1), a continuous liability represents the sum of a large number of independent genetic and environmental factors. It is often assumed that all correlations between relatives stem from shared **genes** and not from a shared environment. This assumption may be relaxed but it then becomes impossible to estimate the heritability of the liability and to extrapolate recurrence risks from one type of relative pair to another. The basic model also assumes that liability is normally distributed. This is not a critical assumption since a normalizing transformation could always be used as long as the liability is continuously distributed. The **prevalence** of the disorder in the population, K, is given by

$$P(D = 1) = P(l > T) = 1 - \mathbf{\Phi}\left(\frac{T - \mu}{\sigma}\right),$$

in which $\mu = \mathrm{E}(l)$ and $\sigma^2 = \mathrm{var}(l)$. As the three parameters T, μ, and σ^2 are confounded, there is no loss of generality in taking $\mu = 0$ and $\sigma^2 = 1$. Thus, individual differences in susceptibility are modeled exclusively through the threshold T; for example, by making T a function of age and gender [5, 22, 23].

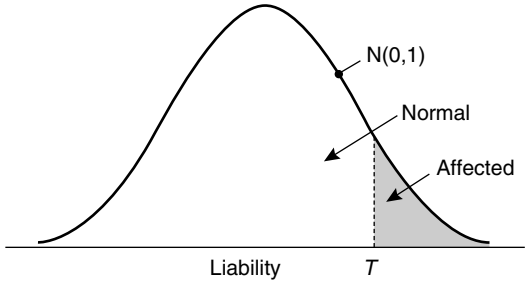

Figure 1 Multifactorial model with threshold. The liability is assumed normally distributed with mean zero and unit variance. An individual is affected when his/her liability exceeds the threshold T. The shaded area is the population prevalence

On the further assumption that the joint distribution of the liabilities of n related individuals is multivariate normal with correlation matrix $\mathbf{R}$, the probability of a pattern $\mathbf{D} = \{D_1, \ldots, D_n\}$ of affection statuses becomes

$$P(\mathbf{D}) = \int_{I_1}\int_{I_2}\cdots\int_{I_n} \phi_n(l_1, \ldots, l_n; \mathbf{R})\, \mathrm{d}l_1 \ldots \mathrm{d}l_n \tag{1}$$

where $\phi(\cdot; \mathbf{R})$ denotes the probability distribution function of the standardized multivariate normal distribution with correlation matrix $\mathbf{R}$. The open interval $I_j (j = 1, \ldots, n)$ is either $(-\infty, T_j)$ if $D_j = 0$, or $(T_j, +\infty)$ if $D_j = 1$, where T_j is the threshold for the jth individual. Eq. (1) can be expressed in a more compact form [26]:

$$P(\mathbf{D}) = (-1)^{|\mathbf{D}|} \sum_{\mathbf{\Delta} \le \mathbf{D}} (-1)^{|\mathbf{D}\cdot\mathbf{\Delta}|} B(\mathbf{\Delta}),$$

where $\mathbf{\Delta} = [\delta_1, \ldots, \delta_n]^t$ is a vector of ones and zeros, $|\mathbf{D} \cdot \mathbf{\Delta}| = \sum D_j \delta_j$, and $B(\cdot)$ is the integral (1) taken over $(-\infty, T_j]$ if $\delta_j = 0$ or over $(-\infty, +\infty)$ if $\delta_j = 1$ for $j = 1, \ldots, n$. That is to say, $B(\cdot)$ gives the probability that none of the members of a subset of $\{1, \ldots, n\}$ is affected. The summation is taken over all vectors $\mathbf{\Delta} = [\delta_1, \ldots, \delta_n]^t$ such that $0 \le \delta_j \le D_j (j = 1, \ldots, n)$.

The model parameters (thresholds and correlations) may be estimated by **maximum likelihood** [26]. A simpler form of analysis uses information on the prevalence and recurrence of the disease in several classes of individuals. The threshold for individuals in the ith class is estimated directly from K_i, the prevalence of the disease in that particular class: $\hat{T}_i = -\mathbf{\Phi}^{-1}(K_i)$. Let τ_{ij} denote the probability that two relatives, one from class i and the other from class j, are both affected. The estimate of the correlation in liability for this particular type of relative pair is obtained by finding ρ_{ij} such that $\tau_{ij} = \mathbf{\Phi}(-\hat{T}_i, -\hat{T}_j, \rho_{ij})$.

It is impossible to assign a genetic interpretation to the correlation coefficient unless it is assumed that the environmental components are uncorrelated and that all genetic factors are additive. In that case, $\rho_{ij} = \Delta_{ij} h^2$ where Δ_{ij} equals twice the kinship coefficient (*see* **Inbreeding**) and h^2 is the

heritability of the liability [22]. In the absence of inbreeding, $\Delta = 1/2$ for first-degree relatives (parent–offspring, sib–sib), 1/4 for second-degree relatives (uncle–niece, grandparent–grandchild) and so on. Thus, the correlation in liability obtained from different types of relative pairs is a linear function of the kinship coefficient. It is then possible to extrapolate the findings on one pair of relatives to another type of relative pair.

Unfortunately, there are multiple reasons why the estimates of the correlation coefficients should not be a linear function of the kinship coefficients. Possible explanations for a poor fit of the model include invalidity of the multivariate normality assumption, presence of dominance within loci, epistatic interactions between loci (*see* **Genotype**), or correlations between nongenetic familial effects [6, 24, 27]. In addition, the model may fail due to the presence of a major gene.

Mixed Models

One of the fundamental questions of genetic epidemiology is whether a disease is purely multifactorial or whether the available data suggest the presence of at least one locus of substantial effect. Morton & MacLean [17] introduced the mixed model to test this hypothesis. Under this model (Figure 2),

$$l = g_{\mathrm{m}} + \varepsilon,$$

where g_{m} is the effect of the major locus and ε is the residual liability, representing the cumulative effects

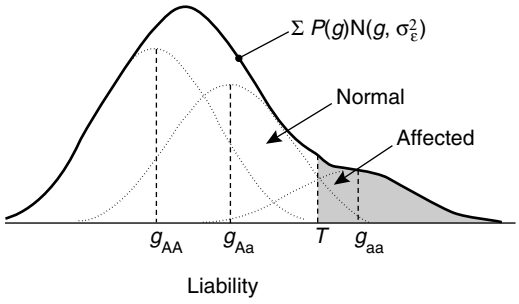

Figure 2 Mixed model with a single threshold. The liability distribution (solid line) is a mixture of three normal distributions with means g_{AA}, g_{Aa}, and g_{aa}, and common variance σ_ε^2. The area under each component curve (dashed curves, shown scaled) is the frequency of the associated genotype

of multiple genetic and/or environmental factors. For the sake of convenience it is often assumed that $\varepsilon = \varepsilon_{\mathrm{c}} + \varepsilon_{\mathrm{p}}$, in which $\varepsilon_{\mathrm{c}} \sim \mathrm{N}(0, \sigma_{\mathrm{c}}^2)$ is a random effect common to all offspring of a mating and $\varepsilon_{\mathrm{p}} \sim \mathrm{N}(0, \sigma_{\mathrm{p}}^2)$ represents uncorrelated effects specific to each individual (including measurement error). The major gene component of the mixed model usually admits only two alleles, A and a, thus introducing four new parameters: three means, g_{AA}, g_{Aa} and g_{aa} and the **gene frequency**, p_{A}. The probability of a pattern of affection, **D**, is computed by summing over all possible genotypes at the major locus,

$$P(\mathbf{D}) = \sum P(g_1, \ldots, g_n) \times \int_{I_1}\int_{I_2}\cdots\int_{I_n} \phi_n(l_1, \ldots, l_n; \mathbf{g}, \mathbf{R})\, \mathrm{d}l_1, \ldots, \mathrm{d}l_n,$$

in which $P(g_1, \ldots, g_n)$ is the joint probability of the genotypes and $\phi_n(\cdot)$, the probability distribution function of the residual terms given the **genotypes**, is multivariate normal with mean vector **g** and correlation matrix **R**.

Note that if the residual terms are uncorrelated, the mixed model reduces to a single gene model with incomplete **penetrance**:

$$\begin{aligned} P(\mathbf{D}) &= \sum P(g_1, \ldots, g_n) \times \int_{I_1}\int_{I_2}\cdots\int_{I_n} \phi_n(l_1, \ldots, l_n; \mathbf{g}, \mathbf{R})\, \mathrm{d}l_1, \ldots, \mathrm{d}l_n \\ &= \sum P(g_1, \ldots, g_n) \prod_{i=1}^{n} \int_{I_i} \phi(l_i; g_i)\, \mathrm{d}l_i \\ &= \sum P(g_1, \ldots, g_n) \prod_{i=1}^{n} S(g_i), \end{aligned}$$

where $S(g)$ is the penetrance of the genotype g. Thus, the mixed model differs from the single gene model only when there exists a source of familial correlation (**polygenic** or environmental) other than the major locus.

Multiple Thresholds

Many disorders are inadequately described by a simple dichotomy, such as *normal* vs. *abnormal*, because they exhibit gradations in severity. The question then arises whether these gradations reflect

the presence of multiple thresholds or the existence of biologically distinct forms of the disorder [18, 19]. For example, suppose that the trait exists in three forms, *normal* ($D = 0$), *mild* ($D = 1$) and *severe* ($D = 2$). This could result from a single liability distribution with two thresholds (Figure 3). To compute $P(\mathbf{D})$ under this assumption, one simply lets the intervals I_i in (1) range from $(-\infty, T_1)$ if $D_i = 0$, (T_1, T_2) if $D_i = 1$, and $(T_2, +\infty)$ if $D_i = 2$ [25]. Hackett & Weller [15] discuss this type of model in the context of **linkage analysis**. Alternately, the various forms of the disease may be determined by two different liabilities that may or may not be correlated. Figure 4 illustrates a situation in which there are two correlated liabilities (l_m and l_s). The severe form of the disease occurs when l_s exceeds the threshold T_s regardless of the value of l_m. The mild form is expressed when $l_m > T_m$ and $l_s < T_s$. The applicability of this model is limited by the rapid increase in the number of parameters. Reich et al. [18] detail several situations in which all parameters may be estimated from prevalence data.

Computational Considerations

Given the widespread availability of programs to compute the **bivariate normal** integral, the approximations to $P(\mathbf{D})$ provided by Falconer and others for the simplest types of familial data (pairs of relatives) are now primarily of historical interest. The calculation of $P(\mathbf{D})$ may still often involve the calculation of high dimensional integrals – a significant problem when these computations occur during an iterative estimation procedure. Some simplifications are possible when the structure of the data is simple and strong assumptions are made regarding the transmission of the trait. In particular, Curnow [5] has shown that if all relatives within a pedigree are of the same order (e.g. parents with offspring) and if it can be assumed that the correlation in liability is due entirely to additive genetic factors, then $P(D_1 = 1, \ldots, D_n = 1)$ may be calculated through the evaluation of a single integral:

$$P(D_1 = \cdots = D_n = 1) = \int_{-\infty}^{+\infty} \phi(x) \left[\mathbf{\Phi} \left(\frac{-(T + x\rho^{1/2})}{(1-\rho)^{1/2}} \right) \right]^n \mathrm{d}x,$$

where $\rho = h^2/2$ in the case of first-degree relatives.

Two approximations have remained popular for the general case [16, 21]. These assume that the conditional distribution of l_i, the liability for the ith individual given the affection statuses of the preceding individuals, is well represented by a normal distribution with mean $\mu_i^* = \mathrm{E}(l_i | D_1, \ldots, D_{i-1})$ and variance $\sigma_i^{2*} = \mathrm{var}(l_i | D_1, \ldots, D_{i-1})$. These expectations and variances are obtained from standard results regarding the moments of variables in truncated multivariate normal distributions. $P(\mathbf{D})$ is then approximated as

$$P(D_1, \ldots, D_n) \approx \prod_{i=1}^{n} \int_{I_i} \phi(l; \mu_i^*, \sigma_i^{2*}) \, \mathrm{d}l_i,$$

where the interval I_i depends on D_i as in integral (1). This approximation is very accurate when the correlations are moderate. The results may vary slightly depending upon the order in which the integrals are evaluated.

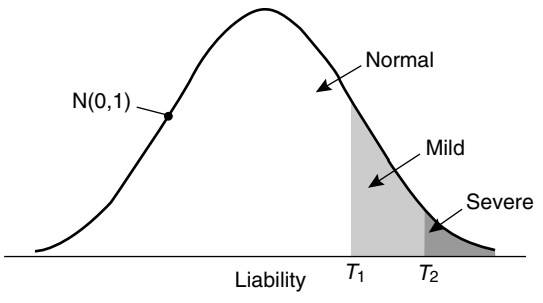

Figure 3 Single liability model for a multifactorial disease with three categories. An individual expresses the mild form of the disease if $T_1 < l < T_2$ and the severe form if $T_2 < l$

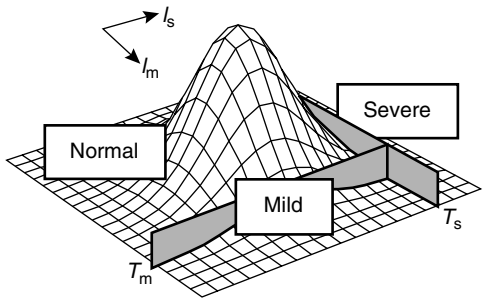

Figure 4 Bivariate liability model for a multifactorial disease with three categories. In this example, an individual expresses the mild form of the disease when $T_m < l_m$ and $l_s < T_s$ and the severe form when $T_s < l_s$, regardless of l_m

Discussion

Edwards [10] once questioned whether the multifactorial model "bore fruit or flower". Indeed, the finding that a multifactorial model provides the best explanation for the transmission of a disease is an anathema to many geneticists. Nonetheless, such traits certainly exist and the chances of encountering them increase as the focus of genetic epidemiology turns from the investigation of rare Mendelian diseases to that of common and complex disorders.

The prevailing modern strategy is to forsake the initial step of model fitting and to jump directly to the search for the genes. Accordingly, a simplified form of the mixed model is used in a series of linkage analyses. The accuracy of the model is subsidiary to whether it will allow the detection of the genes. Thus, estimates of the parameters are themselves of little interest. This is a radical departure from earlier work on the multifactorial and the mixed model, where the determination of these parameters was the goal of the analyses. Two factors contributed to this departure. First, the recognition that the best fitting model is not necessarily the correct one. Secondly, **genetic markers** now span the entire human genome and it has been shown that the inclusion of marker data increases the power of genetic analyses. Thus, the trend is toward the use of a hybrid of linkage and segregation analysis in which the mixed model will retain an important role.

References

[1] Carter, C.O. (1961). The inheritance of congenital pyloric stenosis, *British Medical Bulletin* **17**, 251.

[2] Carter, C.O. (1964). The genetics of common malformations, in *Second International Conference on Congenital Malformations*, M. Fishbein, ed. International Med. Cong. Ltd, New York, pp. 306–313.

[3] Carter, C.O. (1969). Genetics of common disorders, *British Medical Bulletin* **25**, 52–57.

[4] Crittenden, L.B. (1961). An interpretation of familial aggregation based on multiple genetic and environmental factors, *Annals of the New York Academy of Sciences* **91**, 769–780.

[5] Curnow, R.N. (1972). The multifactorial model for the inheritance of liability to disease and its implications for relatives at risk, *Biometrics* **28**, 931–946.

[6] Curnow, R.N. & Smith, C. (1975). Multifactorial models for familial diseases in man, *Journal of the Royal Statistical Society, Series A* **138**, 139–169.

[7] Eaves, L.J. (1984). The resolution of genotype × environment interaction in segregation analysis of nuclear families, *Genetic Epidemiology* **1**, 215–228.

[8] Edwards, J.H. (1967). Linkage studies of whole populations, in *Proceedings of the Third International Congress of Human Genetics*, J.F. Crow & J.V. Neel, eds. Johns Hopkins University Press, Baltimore, pp. 479–482.

[9] Edwards, J.H. (1969). Familial predisposition in man, *British Medical Bulletin* **25**, 58–64.

[10] Edwards, J.H. (1975). Discussion of the paper by Curnow and Smith, *Journal of the Royal Statistical Society, Series A* **138**, 157–161.

[11] Elston, R.C. & Rao, D.C. (1978). Statistical modeling and analysis in human genetics, *Annual Review of Biophysics and Bioengineering* **7**, 253–286.

[12] Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives, *Annals of Human Genetics* **29**, 51–71.

[13] Falconer, D.S. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus, *Annals of Human Genetics* **31**, 1–20.

[14] Gruenberg, H. (1951). The genetics of a tooth defect in the mouse, *Proceedings of the Royal Society, Series B* **138**, 437–451.

[15] Hackett, C.A. & Weller, J.I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions, *Biometrics* **51**, 1252–1263.

[16] Mendell, N.R. & Elston, R.C. (1974). Multifactorial qualitative traits: Genetic analysis and prediction of recurrence risks, *Biometrics* **30**, 41–57.

[17] Morton, N.E. & MacLean, C.J. (1974). Analysis of family resemblance. III. Complex segregation of quantitative traits, *American Journal of Human Genetics* **26**, 489–503.

[18] Reich, T., James, J.W. & Morris, C.A. (1972). The use of multiple thresholds in determining the mode of transmission of semi-continuous traits, *Annals of Human Genetics* **36**, 163–184.

[19] Reich, T., Rice, J., Cloninger, C.R., Wette, R. & James, J.W. (1979). The use of multiple thresholds and segregation analysis in analyzing the phenotypic heterogeneity of multifactorial traits, *Annals of Human Genetics* **42**, 371–389.

[20] Rice, J.P., Neuman, R. & Moldin, S.O. (1991). Methods for the inheritance of qualitative traits, in *Handbook of Statistics*, Vol. 8, C.R. Rao & R. Chakraborty, eds. Elsevier, Amsterdam, pp. 1–27.

[21] Rice, J.P., Reich, T. & Cloninger, C.R. (1979). An approximation to the multivariate normal integral: its application to multifactorial qualitative traits, *Biometrics* **35**, 451–459.

[22] Smith, C. (1970). Heritability of liability and concordance in monozygous twins, *Annals of Human Genetics* **34**, 85–91.

[23] Smith, C. (1971). Recurrence risks with multifactorial inheritance, *American Journal of Human Genetics* **23**, 578–588.
[24] Smith, C. (1974). Concordance in twins: Methods and interpretations, *American Journal of Human Genetics* **26**, 454–466.
[25] Smith, C., Falconer, D.S. & Duncan, L.J.P. (1972). A statistical and genetical study of diabetes. II. Heritability of liability, *Annals of Human Genetics* **35**, 281–299.
[26] Thompson, R. (1972). The maximum likelihood approach to the estimate of liability, *Annals of Human Genetics* **36**, 221–229.
[27] Wilson, S.R. (1979). On the use of multiple thresholds for the determination of the mode of inheritance of semi-continuous traits, *Annals of Human Genetics* **42**, 513–522.
[28] Wright, S. (1934). An analysis of the variability in number of digits in an inbred strain of guinea pigs, *Genetics* **19**, 506–536.

(*See also* **Genetic Correlations and Covariances**; **Segregation Analysis, Mixed Models**; **Path Analysis in Genetics**; **Segregation Analysis, Complex**)

A.A. TODOROV & B.K. SUAREZ

Genetic Map Functions

A genetic map function M gives a relation $r = M(d)$ connecting recombination fractions r and genetic map distances d between pairs of loci along a chromosome arm. Recombination fractions and map distances are summary statistics concerning potentially observable characteristics of the single chromosomes (also known as chromatids) that are the products of meiosis, and that go into gametes. The recombination fraction between two loci is the proportion of such chromosomes that are recombinant, that is, that have genetic material of differing parental origins, at the two loci (*see* **Linkage Analysis, Model-based**). The genetic map distance between two loci is the average number of exchange points that occur along such a chromosome between the loci, where an exchange point, also known as a crossover point, is a point where the parental origin of the genetic material changes. In these definitions, proportions and averages are calculated in the hypothetical infinite population of single chromosomes resulting from meiosis in a given organism, occurring under standard conditions. Variations between organisms within the same species, or of the conditions of meiosis, may lead to small, but observable, differences in these quantities. It should be noted that some authors (e.g. [1] and [9]) use the term map function for the function M^{-1} in the inverse relation $d = M^{-1}(r)$ expressing d in terms of r. We follow Karlin [7] and others in calling M a map function, mainly because the theoretical development is slightly simpler for M than for M^{-1}.

Map functions have been widely used in genetics because of two facts. The first is that genetic map distances are additive by definition, whereas recombination fractions are not. Thus, map distances are preferred for mapping chromosomes. The second is that recombination fractions are much easier to estimate from data, although with human data indirect techniques may need to be used, see [9]. This is because recombination refers only to features of chromosomes at the endpoints of intervals. By contrast, to estimate a map distance information concerning exchanges in the entire interval between two loci is required and, until recently, such information was rarely, if ever, available. Modern molecular genetic methods now exist permitting the identification of points of exchange along chromosomes, and in the near future it may become much easier to estimate map distances directly (see [8]).

The traditional use of map functions has been to take an estimated recombination fraction $\hat{r}$ between two loci and a map function M deemed appropriate for the organism in question, and estimate the map distance between the loci by the quantity $\hat{d} = M^{-1}(\hat{r})$. Perhaps the simplest case is the map function $r = d$, with inverse $d = r$. This is quite satisfactory for small r and d, say, in the interval (0, 0.05), but the relative error increases as the magnitudes of d and r increase. If two loci can be linked by a chain of intermediate loci, each having a recombination fraction of no more than 0.05 (say) with its successor, then a quite satisfactory estimate of the map distance between the initial and final locus can be obtained by adding the successive interlocus recombination fractions. The notion of map function is helpful in situations where such intermediate loci are not available.

The recombination fraction and map length of an interval will differ when there is a nonzero chance of multiple exchange points occurring in the interval. The chance of this occurring increases as the size of the interval increases. If we denote the distribution of exchange points in a particular interval by $(p_0, p_1, p_2, p_3, \ldots)$, so that p_k is the expected proportion of single chromosomes that have k exchange points in the interval, then the recombination fraction is

$$r = p_1 + p_3 + \cdots \tag{1}$$

(i.e. the probability of an odd number of exchange points), while the map length is

$$d = p_1 + 2p_2 + 3p_3 + \cdots. \tag{2}$$

For example, if $p_k = \mathrm{e}^{-d} d^k / k!$, then the map length is easily seen to be d, while the recombination fraction is

$$r = \mathrm{e}^{-d} + \mathrm{e}^{-d}\frac{d^3}{3!} + \cdots = \frac{1}{2}(1 - \mathrm{e}^{-2d}). \tag{3}$$

This relation is known as Haldane's map function, and it is widely used today, nearly 80 years after Haldane [6] first described it. Although simple and easy to use, especially for multilocus calculations, the **Poisson Process** underlying this map function has only rarely been found to fit recombination data. As a result, a sizeable body of work in the late

1940s and 1950s from R.A. Fisher and colleagues and students, excellently summarized in [1], supposed that the points of exchange along a chromosome follow a **renewal process** with independent interpoint distances distributed as $\frac{1}{4}\chi_4^2$ or $\frac{1}{6}\chi_6^2$, rather than the $\frac{1}{2}\chi_2^2$ that gives rise to Haldane's map function. Their model seemed to fit existing human, mouse, and other data quite satisfactorily, but possesses no map function.

The very notion of a map function embodies certain implicit biological assumptions about the process of recombination. For example, all map functions in the literature are bounded above by 1/2, thereby constraining recombination fractions to be $\leq 1/2$. This is widely believed to hold, but there have been instances where it was felt to be untrue, see [4]. Less obviously, the use of a map function presumes that distinct chromosomal intervals having the same map length necessarily have the same recombination fraction, and conversely. This form of stationarity or homogeneity is not observed in the one case in which there is enough data to test it [3]. A number of writers have discussed probability models for recombination that do not constrain recombination fractions to be $\leq 1/2$, and do not satisfy the stationarity properties leading to a map function, see [1] and [5]. Map functions are best viewed as an aspect of certain probability models for recombination. As such, they reflect modeling assumptions, and cannot be expected to be consistent with all the relevant biological knowledge. What matters is whether they are effective for the purposes to which they are put.

Map functions are also useful in contexts where all the products of meiosis remain together, as is the case with ordered or unordered tetrads or octads. In such situations, the model needs to be modified slightly, for although the concept of map distance remains appropriate, the classification of chromosomes as recombinant or not between loci is replaced by a classification of tetrads or octads depending on the parental origins of genetic material at the loci (see, for example, [2]). We will not give any details, here, but simply observe that this development leads us to consider probability models for recombination that refer to the four-strand bundle of chromatids, rather than to the single chromosome products of meiosis. In this approach, chiasmata (the chromosomal structures at points of exchange) are postulated to occur along the four-strand bundle according to some point process, and a mechanism for determining the strands involved in the chiasmata is also specified. The distribution of change-points along the resulting chromosomes is then a consequence of the interplay between the chiasma location process and the strand choice mechanism, and, in specifying the recombination process in this manner, we are also able to calculate the probabilities of interest concerning tetrad and octad types. The simplest assumption concerning strand choice is that the strands involved in any given chiasma are chosen at random from the four possible, independently of those chosen for other chiasmata. This is known as the assumption of no chromatid interference, interference being a term used in genetics to denote some form of dependence. In what follows we make this assumption, although (see [11]) map functions can be defined without it.

Under the assumption of no chromatid interference, a simple relationship widely attributed to K. Mather follows. It states that among meioses in which one or more chiasmata occur in a given interval along the four-strand bundle, on average half of the resulting chromosomes will be recombinant across that interval. More formally, if r is the recombination fraction between two loci, and c_0 is the chance of having no chiasma located in the interval in any meiosis, then

$$r = \tfrac{1}{2}(1 - c_0). \tag{4}$$

When $c_0 = c_0(d)$ depends only on the map length d of the interval, this relation is a map function. Now every chiasma involves just two of the four chromatids, and so the average number of chiasmata between two loci on the four-strand bundle is twice the average number of points of exchange between the same two loci on a single chromosome resulting from meiosis. Suppose that the number of chiasmata occurring in an interval along the four-strand bundle is Poisson distributed with mean $2d$. Then the map length of that interval is just d, and the chance of no chiasmata is e^{-2d}. Substituting into the above formula, we recover the Haldane map function (3) once more. It should be pointed out, however, that we can also recover this map function using a different distribution for the number of chiasmata and a different assumption concerning strand choice [13]. Keeping to the no chromatid interference assumption, we can derive many probabilistic models for recombination by postulating that chiasmata occur along the four-strand bundle according to a stationary renewal process (SRP). If the interchiasma density is f with

respect to twice the map length density, then simple arguments from renewal theory show that for such models,

$$c_0(d) = 2\int_d^\infty \int_y^\infty f(t)\,\mathrm{d}t\,\mathrm{d}y. \tag{5}$$

It is shown in [14] that most of the map functions in the literature can be realized by substituting this expression with a suitable f into Mather's formula (4). This includes certain empirical map functions, such as the following suggested by Haldane in 1919,

$$M^{-1}(r) = 0.7r - 0.15\log(1-2r). \tag{6}$$

Map functions must satisfy certain constraints as a result of their definition, see [11] for details. Some functions suggested in the literature as suitable map functions do not satisfy these constraints [12], and should probably not be used. More importantly, most map functions are associated with stationary renewal processes whose multilocus recombination probabilities are extremely difficult to calculate, and for this reason are not so useful. The class of SRPs with **chi-square distributed** interchiasma distances in the map distance metric has proved both tractable and fairly general [14]. Another family of recombination models in the literature are termed the count-location processes [7]. These require the specification of a distribution $(g_k : k \geq 0)$ for the number (count) of chiasmata along the four-strand bundle, and a sequence F_k of functions giving the distribution of the locations of k chiasmata, given that k occur, $k \geq 1$. In the special case that F_k is equivalent to specifying k locations independently and identically according to a fixed distribution F, and no chromatid interference, we easily find that

$$c_0(d) = g\left(1 - \frac{2d}{m}\right), \tag{7}$$

where $g(s) = \sum_k g_k s^k$, $(0 < s < 1)$ and $m = g'(0)$. Risch & Lange [10] found that this class of recombination models did not give a very good fit to certain large data sets involving *Drosophila melanogaster*.

For many people, map functions are related to the notion of interference. Crossover interference is said to exist when the chance of one or more exchange points in an interval depends on the occurrence of exchange points in other, disjoint intervals. When the points of exchange form a Poisson process, there is no crossover interference. In general, such interference is observed, which is another reason why Poisson processes do not form suitable general models for recombination. (Note that a similar definition of interference can be formulated that refers to chiasmata occurring along the four-strand bundle. The corresponding notion is termed chiasma interference.) The traditional measure of interference is the coincidence coefficient, this being, for adjacent intervals A and B,

$$C_{A,B} = \frac{r(A\&B)}{r(A)r(B)}, \tag{8}$$

where $r(A)$ and $r(B)$ are the recombination fractions of A and B, respectively, and $r(A\&B)$ denotes the chance of simultaneous recombination across A and B. It is easy to check that

$$r(A\&B) = \frac{r(A) + r(B) - r(A \cup B)}{2},$$

where $A \cup B$ is the union of the adjacent intervals A and B. Suppose now that A has map length d, while B has small map length h, and that we take a limit (assumed to exist) in the expression for $C_{A,B}$ as $h \to 0$. Assuming that $M'(0) = 1$, which is one of the conditions that a map function must satisfy, we obtain the differential equation

$$M'(d) = 1 - 2C(d)M(d), \tag{9}$$

where $C(d)$ is the limiting coincidence coefficient, assumed to depend only on the map length of A.

This argument is due to Haldane [6], and many familiar map functions are solutions of this equation when $C(d)$ has the form $(M(d))^{n-1}$. For example, when $n = 1$, we get the Kosambi map function widely used in human genetics:

$$M(d) = \tfrac{1}{2}\tanh(2d). \tag{10}$$

The foregoing discussion shows that there is a connection between map functions and one aspect of crossover interference. In fact, this connection is quite superficial. A more useful (and outside of genetics more widely used) measure of interference is the expression $C_4(d) = C_{A,B}$ where A and B are infinitesimal intervals separated by a map distance d. This measure cannot, in general, be expressed in terms of the map function. In fact, there exist distinct

probability models for recombination having the same map function, with one model having $C_4(d) =$ constant, while the other has $C_4(d)$, a function increasing almost monotonically from 0 at $d = 0$ to 1 for large d. In short, the two recombination models have the same map function, but very different interference properties, using the term interference in a general sense. Map functions do not adequately account for interference; this must be done using a probability model for recombination.

We close with some summary remarks. Map functions can be used to convert recombination fractions to map distances, correcting for multiple exchanges. They also correct for the effect of interference, but do not describe interference completely. They are essentially organism-dependent, and at best provide only rough approximations. It is not uncommon to see multilocus analyses carried out using the Poisson (no chiasma or crossover interference) model underlying Haldane's map function, at the end of the analysis correcting the estimated recombination fractions using Kosambi's or some other map function. This is necessary because map functions (such as Kosambi's) do not, in general, determine joint recombination probabilities for more than three loci. It is reassuring that this somewhat illogical approach gives estimated map distances that are not too different from those that would be obtained using (for example) a comparable stationary renewal process model with chi-square distributed interpoint distances. Ideally, multilocus mapping and linkage analyses should be carried out using a properly specified probability model for recombination suitable for the organism in question. When this is done, map functions are not needed.

References

[1] Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.

[2] Barratt, R.W., Newmeyer, D., Perkins, D.D. & Garnjobst, L. (1954). Map construction in Neurospora crassa, *Advances in Genetics* **6**, 1–93.

[3] Bridges, C.B. & Morgan, T.J. (1923). The second chromosome group of mutant characters, *Carnegie Institute of Washington Publication* **278**, Part II, 126–304.

[4] Fisher, R.A., Lyon, M.F. & Owen, A.R.G. (1947). The sex chromosome of the house mouse, *Heredity* **1**, 335–365.

[5] Goldgar, D.E. & Fain, P.R. (1988). Models of multilocus recombination: non-randomness in chiasma number and crossover location, *American Journal of Human Genetics* **43**, 38–45.

[6] Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors, *Journal of Genetics* **8**, 299–309.

[7] Karlin, S. (1984). Theoretical aspects of genetic map functions in recombination processes, in *Human Population Genetics: The Pittsburgh Symposium*, A. Chakravarti, ed. Van Nostrand Reinhold, New York, pp. 209–228.

[8] Nelson, S.F., McCusker, J.H., Sander, M.A., Kee, Y., Modrich, P. & Brown, P.O. (1993). Genomic mismatch scanning: A new approach to genetic linkage mapping, *Nature Genetics* **4**, 11–18.

[9] Ott, J. (1991). *Analysis of Human Genetic Linkage Data*. Johns Hopkins University Press, Baltimore.

[10] Risch, N. & Lange, K. (1979). An alternative model of recombination and interference, *Annals of Human Genetics* **43**, 61–70.

[11] Speed, T.P. (1996). What is a genetic map function?, in *Genetic Mapping and DNA sequencing*, T. Speed & M.S. Waterman, eds. Springer-Verlag, New York.

[12] Weeks, D.E. (1994). Invalidity of the Rao map function for three loci, *Human Heredity* **44**, 178–180.

[13] Zhao, H. (1995). Statistical analysis of genetical interference. PhD thesis, University of California at Berkeley.

[14] Zhao, H. & Speed, T.P. (1996). On genetic map functions, *Genetics* **142**, 1369–1377.

TERRY P. SPEED

Genetic Markers

Genetic markers are genetic **polymorphisms** used for the study of population structure (**inbreeding, admixture in human populations**), **linkage analysis**, **disease–marker association**, **haplotype analysis**, **paternity testing**, and forensics (*see* **Statistical Forensics**). Early markers included the **blood groups** and various easy-to-observe polymorphisms such as the ability to taste phenylthiocarbamide and anthroposcopic traits.

Beginning about 1960, protein polymorphisms became detectable by various electrophoretic and staining techniques, and the **HLA system** was developed; protein polymorphisms, together with the red cell blood groups, made it possible to type a whole battery of markers from a sample of blood. These markers segregate in a simple fashion according to **Mendel's Laws**, each **genotype** giving rise to a well-differentiated phenotype, but in many cases two genotypes would correspond to the same phenotype because of dominance.

Since about 1980, the genetic markers of choice have been deoxyribonucleic acid (**DNA**) polymorphisms. The first of these were the restriction fragment length polymorphisms (RFLPs), determined by digesting the DNA with restriction enzymes that cut the DNA at specific short sequences of base pairs; **mutations** in these sequences prevent the cutting, and differences in the cut and uncut fragments are detected by variation in the fragment lengths produced [1]. However, because so many fragments of similar sizes are produced when the whole DNA is digested, the fragments for a particular chromosomal locus have to be distinguished from the rest by a specific probe – a cloned sequence of DNA, marked in some way for detection, that will anneal with only the fragments from that locus. Such markers typically yield only two or three alleles per locus.

More recently, sequences of base pairs have been discovered that are repeated a different number of times from allele to allele (*see* **Gene**), and the variation in these repeats is the basis of a polymorphism. The first of these polymorphisms to be commonly used as markers were the minisatellites, also called "variable number of tandem repeat" (VNTR) polymorphisms [3]. In these, the sequence that is repeated varies from nine to 60 base pairs in length. These polymorphisms originally required a specific probe for their detection but can now also be detected via the polymerase chain reaction [5]. Since 1989, an abundance of multiallelic short tandem repeat polymorphisms (STRPs), also called microsatellites or simple sequence length polymorphisms, have been available, the sequence repeated varying from two to nine base pairs [6, 7]. (Some authors consider STRPs to be a type of VNTR polymorphism because the only difference between them is the size of the repeating unit and the complexity of the repeat.) Both minisatellites and microsatellites can have very high **polymorphism information content** (PIC) and **linkage information content** (LIC) values and, apart from typing errors, show a one-to-one correspondence between genotype and phenotype. They are particularly advantageous because, with the advent of the polymerase chain reaction, which allows specific segments of DNA to be amplified, they can be typed without any need for probes. They only require very small quantities of DNA, such as can be obtained from a cheek swab, and have largely supplanted all the earlier markers. During the 1990s, about 10 000 STRPs were identified and mapped, and many hundreds of genes have been mapped using them.

The most recent genetic markers are single nucleotide polymorphisms (SNPs, pronounced "snips") [2, 4], which are extremely abundant in the human genome – occurring approximately on average once every one or two thousand base pairs. SNPs are essentially diallelic RFLPs that involve a transition or transversion, or an insertion/deletion of a single nucleotide. SNPs can be detected by many different methods. They may supplant STRPs when the cost of typing them becomes cheaper (at most a third of the cost of typing multiallelic RFLPs) and they can be reliably typed because of their abundance and proximity to coding sequences. SNPs are often observed within the coding sequence of genes and may play a direct role in altering the function of a protein.

References

[1] Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980). Construction of a genetic-linkage map in man using restriction fragment length polymorphisms, *American Journal of Human Genetics* **32**, 314–331.

[2] International SNP Map Working Group, The (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature* **409**, 928–933.

[3] Jeffreys, A.J., Wilson, V. & Thein, S.L. (1985). Hypervariable "minisatellite" regions in human DNA, *Nature* **314**, 67–73.

[4] Kruglyak, L. (1997). The use of a genetic map of bialleleic markers in linkage studies, *Nature Genetics* **17**, 21–24.

[5] Mullis, K.B. & Faloona, F.A. (1987). Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction, *Methods of Enzymology* **155**, 335–350.

[6] Weber, J.L. & Broman, K.W. (2001). Genotyping for whole-genome scans: past, present and future, *Advances in Genetics* **42**, 77–96.

[7] Weber, J.L. & May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction, *American Journal of Human Genetics* **44**, 388–396.

ROBERT C. ELSTON

Genetic Risk Ratios

In human genetics and genetic epidemiology, risk ratios can assume a number of different forms, including risk ratios for relatives, for **candidate genes**, and for genetic **markers**. The goal of many genetic studies is to quantify the risk of disease occurrence associated with particular genetic factors. The strength of this association can depend on **interactions** between environmental and genetic factors, gene–gene interactions, and the distance along a causal pathway from a genetic variant to a disease outcome. Models that can incorporate these complexities have an important role. Here, however, we focus on risk ratios corresponding to **associations** between a single disease and a single genetic factor.

Let D denote the disease under study, and let G or $\overline{\text{G}}$ denote the presence or absence of a particular genetic characteristic in an individual. In these risk ratios, exposure is defined in terms of an individual's genetic information G. Further assume that every individual is correctly classified as having (D) or not having ($\overline{\text{D}}$) the disease. Then, a general genetic **relative risk** (RR) is defined as

$$\text{RR} = \frac{\Pr(\text{D}|\text{G})}{\Pr(\text{D}|\overline{\text{G}})}. \tag{1}$$

This measure of disease-by-genetic-factor association is just the ratio of the conditional probabilities of having the disease given the presence or absence of the genetic characteristic. These conditional probabilities are usually referred to as **penetrances**, and require cross-sectional or cohort designs for their direct estimation. Individuals or families are often sampled according to a disease-related phenotype, so the **odds ratio** is also useful because of its invariance to the direction of sampling. The genetic odds ratio (OR), that approximates the genetic RR for a rare disease, is

$$\text{OR} = \frac{\Pr(\text{D}|\text{G})/\Pr(\overline{\text{D}}|\text{G})}{\Pr(\text{D}|\overline{\text{G}})/\Pr(\overline{\text{D}}|\overline{\text{G}})}, \tag{2}$$

which can be written equivalently as

$$\text{OR} = \frac{\Pr(\text{G}|\text{D})/\Pr(\overline{\text{G}}|\text{D})}{\Pr(\text{G}|\overline{\text{D}})/\Pr(\overline{\text{G}}|\overline{\text{D}})}. \tag{3}$$

Risk Ratios for Relatives

When a genetic factor cannot be measured directly, but information on disease status is available on family members of an affected individual, genetic risk ratios can be derived indirectly.

Familial Aggregation

Evidence for familial aggregation, which is the tendency of disease to cluster in families, provides a rationale for subsequent genetic studies intended to assess particular genetic factors or to search for disease susceptibility **genes**. A measure of familial aggregation that uses information on family history (FH) of disease, say, in first-degree relatives, is

$$\text{OR} = \frac{\Pr(\text{D}|\text{FH})/\Pr(\overline{\text{D}}|\text{FH})}{\Pr(\text{D}|\overline{\text{FH}})/\Pr(\overline{\text{D}}|\overline{\text{FH}})}. \tag{4}$$

In a **case–control** design, D and $\overline{\text{D}}$ correspond to affected cases and unaffected controls, while FH is a surrogate for genetic loading in the family, and is thus subject to misclassification error [14]. Even in the absence of any genetic etiology, however, the probability of a positive FH increases with the number of relatives considered. Alternatively, FH scores can be defined to take into account the number of affected relatives and the family structure.

In a **family-based case–control** design, familial aggregation can also be assessed by comparing the risk of disease among relatives of cases with that among relatives of controls,

$$\frac{\Pr(\text{D in relative type R of case}|\text{affected case})}{\Pr(\text{D in relative type R of control}|\text{unaffected control})}. \tag{5}$$

When this ratio is greater than 1, family aggregation can be present, but, without additional environmental exposure information, aggregation due to shared genes can be indistinguishable from that due to shared environment [5]. For controls that are representative of the general population, the denominator of this ratio approximates the population risk.

Recurrence Risk

A related measure is known as the *recurrence risk* for a type R relative of an affected individual. It occurs in the numerator of the ratio, (5), for familial aggregation in family case–control designs. In Risch's [7]

development of **multilocus** models of inheritance (*see* **Segregation Analysis, Classical**) for complex traits that are useful in **linkage analysis**, he defines a risk ratio, λ_R, which compares the recurrence risk in relatives of type R of an affected individual with the population prevalence K. For example, when the relative type is a sibling, λ_s is defined as

$$\frac{\Pr(\text{D in Relative who is a sibling}|\text{affected case})}{K}. \tag{6}$$

Risch [8] also establishes how the power of affected relative pair studies to detect linkage critically depends on the value of λ_R. The relationship between d_s and the genotype RR, defined using (1) with G and G denoting individuals with or without the susceptibility genotype respectively, depends on allele frequency and the genetic inheritance models [9].

Risk Ratios for Candidate Genes

Investigations of candidate genes are usually based on a priori biologic hypotheses about a particular candidate gene. If genetic information is available, say in the form of a measured candidate gene, then the association of particular genetic variants with a disease can be evaluated using several versions of the genetic risk ratio. These include allelic, **genotype**, and **haplotype** RRs, although the latter is usually considered in the context of a genetic marker (see further below).

A common study design involves sampling individuals by disease status, assembling an appropriate control group, genotyping cases and controls at a **candidate gene** locus, and then comparing the distribution of the candidate gene between the case and control groups. Population (case–population control) and family-based (case–parental control) designs are two approaches used to assess risk ratios for the association of a candidate gene with a disease phenotype.

In case–control studies of candidate gene loci, a fundamental issue is the choice of a reference or control group. Controls can be randomly selected from the population of unaffected individuals or can be matched to the affected cases on relevant characteristics. Because genetic factors vary greatly by ethnic background and population history and geography, there is a serious potential for confounding by population stratification when unrelated individuals are used as controls or when matching on available measures of ethnicity is inadequate. To avoid this problem, the control group can be drawn from members (or potential members) of the family of a case instead of unrelated individuals. However, population controls may be less costly to recruit, and, in the absence of population stratification, may be more efficient statistically (*see* **Family-based Case–Control Studies**).

Population Risk Ratios

At a candidate gene locus, each individual inherits one genetic variant, known as an allele, from each parent and the two alleles together constitute a genotype. When only two variants occur in a population of individuals, there are three possible **genotypes**, while for a locus with multiple variants (n alleles), there are $n(n+1)/2$ possible genotypes. For a single multiallelic locus, the genetic factor G can be expressed as

$$G = \begin{cases} a_i & i = 1, \ldots, n, \quad \text{to denote a single allele, or} \\ a_i a_j & i, j = 1, \ldots, n, \quad \text{to denote a genotype.} \end{cases} \tag{7}$$

For a candidate gene with two alleles, a_1 and a_2, we can define *genotype RRs* in which G depends on the presence or absence of allele a_2, i.e.

$$\text{RR}_1 = \frac{\Pr(\text{D}|a_1a_2)}{\Pr(\text{D}|a_1a_1)}, \qquad \text{RR}_2 = \frac{\Pr(\text{D}|a_2a_2)}{\Pr(\text{D}|a_1a_1)}, \tag{8}$$

and the corresponding *genotype ORs* as

$$\begin{aligned} \text{OR}_1 &= \frac{\Pr(\text{D}|a_1a_2)/\Pr(\overline{\text{D}}|a_1a_2)}{\Pr(\text{D}|a_1a_1)/\Pr(\overline{\text{D}}|a_1a_1)}, \\ \text{OR}_2 &= \frac{\Pr(\text{D}|a_2a_2)/\Pr(\overline{\text{D}}|a_2a_2)}{\Pr(\text{D}|a_1a_1)/\Pr(\overline{\text{D}}|a_1a_1)}. \end{aligned} \tag{9}$$

When an increased risk of disease is associated with having only one copy of allele 2, i.e. a_2 is dominant to allele a_1, $\text{OR}_1 = \text{OR}_2 > 1$, whereas when an increased risk is associated with two copies of a_2, i.e. a_2 is recessive to allele a_1, $\text{OR}_2 > 1$ and $\text{OR}_1 = 1$. This formulation permits tests of overall association of the candidate gene with a disease by considering the two ORs jointly, as well as specific tests for mode of inheritance.

For a multiallele locus, the number of genotype categories can become large. Because comparisons

based on the presence or absence of an allele are more parsimonious, an alternative approach is to use the allele rather than the genotype as the unit of analysis, with two observations contributed from each individual. For example, one approach in a case–control study is to estimate each allelic RR_i by counting and comparing the frequency of the a_i alleles between cases and controls.

Family-based Risk Ratios with Parental Controls

One particular family-based design involves the ascertainment of an affected individual followed by genotyping of this case and their parents. Falk & Rubinstein [2] proposed that the maternal and paternal alleles transmitted to the affected child form a case genotype (D) while the nontransmitted alleles form a control genotype ($\overline{\text{D}}$).

For a two-allele candidate gene locus with alleles a_1 and a_2, a simple *genotype RR* is

$$\text{RR} = \frac{\Pr(\text{D}|\text{presence of allele } a_2)}{\Pr(\text{D}|\text{absence of allele } a_2)}, \qquad (10)$$

which does not distinguish between genotypes with one or two copies of allele a_2. When the measured candidate gene is the disease gene, this RR is insensitive to population stratification, and can be estimated without bias by the OR [4, 6]:

$$\text{OR} = \frac{\Pr(a_2 \text{ is present}|\text{D})/\Pr(a_2 \text{ is absent}|\text{D})}{\Pr(a_2 \text{ is present}|\overline{\text{D}})/\Pr(a_2 \text{ is absent}|\overline{\text{D}})}. \qquad (11)$$

This OR is sometimes referred to as a *haplotype RR*, particularly when a genetic marker rather than a candidate gene is used [2, 6]. Schaid & Sommer [11] suggest the use of two genotype RRs to distinguish between genotypes with one or two copies of allele a_2. Valid inference generally requires that the case–control **matching** be taken into account.

Examination at the level of the individual allele, rather than the genotype, leads to an *allelic RR*. The alleles transmitted and not transmitted by each of the parents are the basis of this risk measure, which is closely related to the transmission/disequilibrium tests (TDT) noted in the following section. However, inference can be complicated by a lack of independence between the parental transmissions, and models that condition on the parental genotypes are more suitable in this case [10].

For a multiallele locus with parents having alleles a_1a_2 and a_3a_4, there are four possible genotypes that could be observed in their offspring. If the affected child (D) received alleles a_1 and a_3, then the genotypes a_2a_4, a_1a_4, and a_2a_3 can be taken as control ($\overline{\text{D}}$) genotypes. Self et al. [12] formulated a general likelihood for a series of independent affected children based on indicator variables for the presence or absence of a given allele in the case and control genotypes. Their approach assumes uniform segregation of gametes apart from the genotype effect, but does not require the parental genotypes to be independent. The likelihood they develop has the same form as that for a logistic regression analysis of a matched case–control study with a single case and three controls.

Features of other family-based case–control designs that include siblings and cousins as controls have been of recent interest [3, 15].

Risk Ratios for Genetic Markers

Genotype, allelic, and haplotype RRs analogous to those described for candidate genes can be estimated using genetic markers. When specific genetic loci for a disease are unknown, finely spaced genetic markers with known locations may be used to screen the whole genome or selected genomic regions to help localize a disease-susceptibility gene.

Allelic information for a single locus can be extended to the multilocus setting by considering **haplotypes**. When two or more neighboring loci are considered together, a haplotype can be defined as a multilocus analog of an allele. A pattern of alleles from each of several loci that are transmitted together from one parent constitute one haplotype. For two loci, with n_1 and n_2 alleles, respectively, occurring in a population, there are $n_1 \times n_2$ possible haplotypes that can be expressed as $\text{G} = a_i b_j$, where a_i and b_j represent allele types $i, i = 1, \ldots, n_1$, and $j, j = 1, \ldots, n_2$, from the two loci comprising the haplotype. A pair of haplotypes, one inherited from each parent, constitutes a multilocus genotype. A multilocus haplotype of two genetic markers defined in this way can be used to construct a *haplotype RR* analogously to an allelic RR for a single genetic marker.

A multilocus haplotype can also be constructed from a genetic marker and disease gene, but the latter

is usually unobserved. When two loci forming a haplotype on a parental chromosome are close together, they are less likely to be separated by recombination when a gamete is formed during meiosis. This phenomenon can be exploited in disease–gene localization studies through the modeling of a haplotype consisting of an allele at a known marker location and an unobserved disease allele. Association between particular alleles of a genetic marker and specific alleles of the unobserved susceptibility gene that occurs across families in a population is known as *allelic association* or **linkage disequilibrium**, whereas the term linkage refers to association that occurs within a family. The absence of tight linkage disequilibrium between alleles at an unobserved causal disease gene locus and measured alleles at marker loci can induce attenuation bias in the RR estimates based on marker alleles [4, 6].

The class of methods known as TDT used in family-based designs involve hypothesis tests that detect linkage only in the presence of allelic association [6, 13]. These methods can also test for association in the presence of linkage, provided correlation between parental transmissions or among related individuals induced by the presence of linkage are taken into account [1, 13].

References

[1] Curnow, R.N., Morris, A.P. & Whittaker, J.C. (1998). Locating genes involved in human diseases, *Applied Statistics* **47**, 63–76.

[2] Falk, C.T. & Rubinstein, P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations, *Annals of Human Genetics* **51**, 227–233.

[3] Gauderman, W.J., Witte, J.S. & Thomas, D.C. (1999). Family-based association studies, in Innovative Study Designs and Analytic Approaches to the Genetic Epidemiology of Cancer. *Journal of the National Cancer Institute Monographs*, No. 26, pp. 31–37.

[4] Knapp, M., Seuchter, S.A. & Baur, M.P. (1993). The haplotype-relative-risk (HRR) method for analysis of association in nuclear families, *American Journal of Human Genetics* **52**, 1085–1093.

[5] Majumder, P.P., Chakraborty, R. & Weiss, K.M. (1983). Relative risks of diseases in the presence of incomplete penetrance and sporadics, *Statistics in Medicine* **2**, 13–24.

[6] Ott, J. (1989). Statistical properties of the haplotype relative risk, *Genetic Epidemiology* **6**, 127–130.

[7] Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models, *American Journal of Human Genetics* **46**, 222–228.

[8] Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs, *American Journal of Human Genetics* **46**, 229–241.

[9] Rybicki, B.A. & Elston, R.C. (2000). The relationship between the sibling recurrence-risk ratio and genotype relative risk. *American Journal of Human Genetic*, **66**, 593–604. (See also *American Journal of Human Genetics* **67**, 541).

[10] Schaid, D.J. (1996). Genetic score tests for associations of genetic markers with disease using cases and their parents, *Genetic Epidemiology* **13**, 423–449.

[11] Schaid, D.J. & Sommer, S.S. (1993). Genotype relative risks: methods for design and analysis of candidate-gene association studies, *American Journal of Human Genetics* **53**, 1114–1126.

[12] Self, S.G., Longton, G., Kopecky, K.J. & Liang, K.-Y. (1991). On estimating HLA/disease association with application to a study of aplastic anemia, *Biometrics* **47**, 53–61.

[13] Spielman, R.S. & Ewens, W.J. (1996). The TDT and other family-based tests for linkage disequilibrium and association, *American Journal of Human Genetics* **59**, 983–989.

[14] Weiss, K.M., Chakraborty, R., Majumder, P.P. & Smouse, P.E. (1982). Problems in the assessment of relative risk of chronic disease among biological relatives of affected individuals, *Journal of Chronic Disease* **35**, 539–551.

[15] Witte, J.S., Gauderman, W.J. & Thomas, D.C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs, *American Journal of Epidemiology* **149**, 693–705.

K.A. KOPCIUK & SHELLEY B. BULL

Genetic Transition Probabilities

Genetic transition probabilities represent one of four major components of **likelihoods** employed in **segregation analysis**, the statistical methodology used to elucidate, from family data, the mode of inheritance of a particular trait. The other components are concerned with the joint genotypic frequencies of mating pairs, the method by which the sample is **ascertained**, and the relationship between **genotype** and phenotype. The set of genetic transition probabilities describes how genetic variability is passed from one generation to the next. This approach to specification of the mode of inheritance was originally developed in the context of the "generalized major gene transmission model" [4], but subsequent synthetic efforts led to the formulation of a unified model incorporating elements of the classical mixed model in segregation analysis [5] while retaining the use of genetic transmission probabilities (see below) to describe genetic transition from one generation to the next.

A genetic transition probability specifies the probability that an offspring has a particular genotype, conditional on the genotypes of the parents, i.e. the probability p_{stu} that an offspring has genotype u, given that one parent has genotype s and the other has genotype t. For single-locus models, the **conditional probabilities** $\{p_{stu}\}$ form the elements of a three-dimensional stochastic matrix called the genetic transition matrix. These matrices represent mathematical summarizations of the genotypic distribution of offspring, conditional on the two parental genotypes, and possibly upon the gender of the offspring. They are commonly displayed as a two-dimensional matrix in which each element is a vector giving the probability distribution of the set of all possible offspring genotypes, conditional upon the parental mating type (pair of parental genotypes). The matrix of genetic transition probabilities under a simple Mendelian model (*see* **Mendel's Laws**) positing two allelic **gene** alternatives, A and a, at a single autosomal locus is given in Table 1, where each entry is a genotypic distribution $[p_{st1}\ \ p_{st2}\ \ p_{st3}]$, i.e. the vector of probabilities that the offspring is AA, Aa, and aa, respectively, conditional upon parental mating type $s \times t$. Thus, the mating type AA × AA can produce only AA offspring, and the AA × aa mating only Aa offspring, while the Aa × Aa mating type is associated with 1/4, 1/2, and 1/4 probabilities of producing AA, Aa, and aa offspring, respectively. Two such matrices are required if the locus of interest is X-linked, one for male offspring, and another for female offspring; dimensioning is also gender-dependent. The genetic transition matrix can be specified for any number of unlinked loci by utilizing Kronecker products to operate upon the individual genetic transition matrices for those loci, and this approach can also be generalized to accommodate linked loci [4].

Table 1

	t		
s	1 = AA	2 = Aa	3 = aa
1 = AA	$[1\ \ 0\ \ 0]$	$[\frac{1}{2}\ \ \frac{1}{2}\ \ 0]$	$[0\ \ 1\ \ 0]$
2 = Aa	$[\frac{1}{2}\ \ \frac{1}{2}\ \ 0]$	$[\frac{1}{4}\ \ \frac{1}{2}\ \ \frac{1}{4}]$	$[0\ \ \frac{1}{2}\ \ \frac{1}{2}]$
3 = aa	$[0\ \ 1\ \ 0]$	$[0\ \ \frac{1}{2}\ \ \frac{1}{2}]$	$[0\ \ 0\ \ 1]$

The genetic transition probabilities $\{p_{stu}\}$ can be expressed as functions of transmission probabilities that describe the probability that an individual with a given genotype transmits a particular allele (from the set of possible allelic variants) to his or her offspring. To illustrate the use of transmission probabilities, consider the case of an autosomal locus with two allelic alternatives, A and a, and define τ_t as the probability that an individual with genotype t transmits an A allele to the offspring; by symmetry, $(1 - \tau_t)$ is the probability that this individual instead transmits an a allele, and the entries of the matrix are generated by

$$[\,p_{st1}\quad p_{st2}\quad p_{st3}\,] = [\tau_s\tau_t\quad \tau_s(1-\tau_t)+\tau_t(1-\tau_s)\quad (1-\tau_t)(1-\tau_t)]\,.$$

If we specify the values $\tau_{AA} = 1$, $\tau_{Aa} = 1/2$, and $\tau_{aa} = 0$, which are appropriate under a Mendelian model that assumes that there is no mutation and that either parental allele is equally likely to be transmitted (no meiotic drive), then we obtain the values given in the above genetic transition matrix for a diallelic autosomal locus [2, 3]. Typically, the transmission probabilities are estimated jointly with other model parameters via **likelihood** and compared with results obtained under models specifying particular values or constraints on the transmission probabilities [1–3, 5] (*see* **Segregation Analysis, Complex**).

This general approach is also used in connection with **regressive models** for analyzing family data. Other extensions provide for the inclusion of phenomena such as mutation in the context of segregation analysis of pedigrees [7].

Care should be taken to distinguish the term *transition probabilities* (p_{stu}) from the component *transmission probabilities* (τ_t). It may be noted that the matrix of genetic transition probabilities is analogous to the matrix of transition probabilities from classical developments using **stochastic processes**, which have also been applied fruitfully to genetic problems: while the genetic transition matrix mediates change from one generation to the next, it also differs in that it describes a transition involving genetic transmission from two individuals (the parents) in one generation to a single individual (the offspring) in the next. The term "transition matrix" has also been applied to arrays of genotypic probabilities of individuals conditional on the genotypes of a relative of a particular type, such as the I and T matrices described by Li & Sacks [6] for the one-locus, two-allele case under panmixia.

In the case of **polygenic inheritance**, the genotype of interest is a polygenotype, which is typically modeled as having a Gaussian population distribution (*see* **Normal Distribution**) with variance σ_G^2. The genetic transition probability under such circumstances is modeled as a probability density; for example, under the classical additive polygenic model assuming panmixia, the expression for the offspring polygenotype H of parental polygenotypes F and G is a normal density with mean $(F + G)/2$ and variance $\sigma_G^2/2$. Genetic transition probabilities for a single locus of major effect and polygenic inheritance are both utilized in mixed models that contain both polygenic and monogenic components [2, 3].

References

[1] Demenais, F.M. & Elston, R.C. (1981). A general transmission probability model for pedigree data, *Human Heredity* **31**, 93–99.

[2] Elston, R.C. (1980). Segregation analysis, *Current Developments in Anthropological Genetics* **1**, 327–354.

[3] Elston, R.C. & Rao, D.C. (1978). Statistical modeling and analysis in human genetics, *Annual Review of Biophysics and Bioengineering* **7**, 253–286.

[4] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[5] Lalouel, J.M., Rao, D.C., Morton, N.E. & Elston, R.C. (1983). A unified model for complex segregation analysis, *American Journal of Human Genetics* **35**, 816–826.

[6] Li, C.C. & Sacks, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices, *Biometrics* **10**, 347–360.

[7] Morton, N.E. & Yasuda, N. (1980). Transition matrices with mutation, *American Journal of Human Genetics* **32**, 202–211.

DEBORAH V. DAWSON

Genitourinary Medicine

The application of biostatistics to genito–urinary medicine is largely focused on the epidemiology of sexually transmitted diseases (STDs). The evaluation of treatment efficacy for particular conditions is, in general, carried out using standard statistical methods, and these are not considered further. Some STDs in developed countries are ascertained with reasonably complete case notification or, alternatively, **prevalence** estimates are derived from data from sentinel clinics. However, for other important STDs, e.g. chlamydia and human papilloma virus (HPV), there is poor case ascertainment and little knowledge of prevalence or incidence in general populations.

The epidemiology of STDs differs in important respects from the epidemiology of many other infectious diseases. As with any infectious disease, **risk** is dependent on the prevalence of the disease in the population and, with bacterial STDs, multiple episodes are possible. However, the analysis of STD transmission has elements in common with problems often associated with environmental health exposures. Risk of disease is largely determined by individual risk behavior (i.e. numbers of partners, types of sexual acts, use of condoms, etc.). These different modes of behavior are heterogeneous [1, 5, 7] and difficult to measure accurately.

Historical Development

Hethcote & Yorke [4] carried out fundamental theoretical work in deterministic modeling of STD dynamics. This work has been extended to apply to human immunodeficiency virus (HIV) and **AIDS**, notably by May & Anderson [8]. The work essentially involves setting up differential equations to describe disease incidence in different strata of population, as a function of sexual behavior and disease prevalence throughout the population. In its simplest form, the basic reproductive rate (R_0) is a product of the probability of transmission (b), the rate of partner change (C), and the duration of infection (D). Another important area of theoretical research has been to investigate the possible effects of concurrent partnerships on the dynamic evolution of STDs and HIV in a population [2]. Researchers have also used stochastic **Monte Carlo** microsimulations, but these have been more focused on HIV and AIDS than other STDs.

For theoretical models to be useful, they must use reliable empirical data. However, possible behavioral determinants of STD risk are hard to define, let alone measure. Models are often posed in terms of rates of acquisition of new partners, a concept which is difficult to reconcile with the real experience of individuals. As a result, people may not necessarily report their sexual behavior reliably, nor interpret concepts such as partnerships or sexual contacts in the way a researcher would wish. Individuals may tend to report, and indeed remember, what is socially acceptable rather than what actually happened.

Development of Survey Methods

In the 1990s, a number of important studies were carried out in developed countries to attempt to generate accurate information on the distribution of sexual behavior in the general population of the UK [5], France [1], and the US [7]. Although standard **survey** methods were used, the sensitivity of the subject reinforced the need for rigorous attention to detail. Great care was taken to generate representative samples. The questionnaires provided appropriate and clear definitions of concepts (e.g. sexual intercourse, sexual partner) using unambiguous and nonjudgmental terms. Respondents were assured of **confidentiality** and the questions were designed to facilitate accurate recall of past events. Nevertheless, all studies to date have suffered to some extent from inconsistent responses from men and women, where, for example, men tend to report more opposite sex partners than women [10].

Sexual Networks and Mixing

There has been much interest in understanding the structure of sexual networks and tracing possible chains of infection. Theoretical and empirical studies have been carried out [6]. Routine STD clinic contact tracing (e.g. sexual contacts of gonorrhoea cases) can provide one possible avenue to identifying these contact networks. As well as studying partner choice at the micro (individual) level there has also been considerable research into understanding

mixing patterns at the macro level. In other words, is there a general propensity for people disproportionately to choose partners similar to themselves (**assortative mating**), dissimilar to themselves (disassortative mating), or is any partner equally likely to be selected irrespective of sexual behavior (random mating)? In matrix terminology, if men and women are allocated to strata i, j dependent on numbers of partners, then the proportion of men in stratum i who have female partners in stratum j will then be the i, jth element of the mixing matrix. In this notation, assortative mating is represented by disproportionately large values along the main diagonal, and conversely for disassortative mating. Renton et al. [9] have shown how the mating matrix describing a population may be inferred from information about STD transmission contact pairs attending a genito-urinary medicine (GUM) clinic.

Study Designs

Studies designed to elucidate risks for STDs are usually **observational**. Because of the difficulty of measuring behavioral variables, **residual confounding** will always be a potential problem in any attempt to measure a particular risk factor while controlling for others. There is limited scope for using randomized controlled trials (RCTs) (*see* **Clinical Trials, Overview**) to examine different risk factors for the acquisition of STDs, because risk factors such as numbers of partners or types of sex are only amenable to educational intervention. An important exception was the RCT of STD treatment carried out in Tanzania [3], where randomization was at the community level. **Case–control studies** of STD risk have the additional problem of identifying suitable **control** groups, and several published studies have reported misleading results. If cases are selected from an STD clinic setting, then others attending the same clinic will almost certainly be poor controls because of the nature of the risk that caused them to attend the clinic.

STD and HIV

It is possible that HIV risk is enhanced in the presence of concurrent STD, and this could explain some of the difference in HIV prevalence between the developed and the developing world. STD intervention in Tanzania reduced HIV incidence and Groskurth et al. [3] provide evidence for this. The trial was the first of its kind to use an RCT design to avoid any possible bias by confounding with behavioral factors, which had been a problem with all previous observational studies showing a link between HIV and STD. It is becoming increasingly apparent that silent epidemics of chlamydia and HPV may be much more important than was previously realized. The methodologic problems relating to ascertainment of reliable estimates of population prevalence for these silent diseases are considerable.

Future development

It is to be hoped that a better understanding of the sexual mixing matrix, both empirically and theoretically, will increase the utility of transmission models. Further examination of the interaction between HIV and STD may lead to better understanding of the epidemiology of HIV and in particular its geographic variability. The use of organism typing to determine chains of transmission in network and contact tracing studies, in particular with gonorrhoea, has the potential to improve vastly our knowledge of sexual networks and the spread of STD within social networks. Finally, improved **survey** methodology is required to facilitate the collection of better behavioral data.

References

[1] ACSF (1993). *Les Comportements Sexuels en France*. La Documentation Francaise. Paris.

[2] Dietz, K. & Hadeler, K.P. (1988). Epidemiological models for sexually transmitted diseases, *Journal of Mathematical Biology* **26**, 1–25.

[3] Groskurth, H, Mosha, F., Todd, J., Mwijarubi, E., Klokke, A., Sedkoro, K., Mayaud, P., Changalucha, J., Nicoll, A., Ka-Gina, G., Newell, J., Mugeye, K., Mabey, D. & Hayes, R. (1995). Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial, *Lancet* **346**, 530–536.

[4] Hethcote, H.W. & Yorke, J.A. (1984). Gonorrhea transmission dynamics and control, *Lecture Notes in Biomathematics*, Vol. 56. Springer-Verlag, New York.

[5] Johnson, A.M. Wadsworth, J., Wellings, K. & Field, J. (1994). *Sexual Attitudes and Lifestyles*. Blackwell, Oxford.

[6] Klovdahl, A.S., Potterat, J.J., Woodhouse, D.E., Muth, J.B., Muth, S.Q. & Darrow, W.W. (1994). Social

networks and infectious disease: the Colorado Springs, *Social Science and Medicine* **38**, 79–88.

[7] Laumann, E.O., Gagnon, J.H., Michael, J.T. & Maichaels, S. (1994). *The Social Organization of Sexuality*. University of Chicago Press, Chicago.

[8] May, R.M. & Anderson, R.M. (1988). The transmission dynamics of human immunodeficiency virus (HIV), *Philosophical Transactions of the Royal Society London, Series B* **321**, 565–607.

[9] Renton, A., Whitaker, L., Ison, C., Wadsworth, J. & Harris, J.R. (1995). Estimating the sexual mixing patterns in the general population from those in people acquiring gonorrhoea infection: theoretical foundation and empirical findings, *Journal of Epidemiology and Community Health* **49**, 205–213.

[10] Wadsworth, J., Johnson, A.M., Wellings, K. & Field, J. (1996). What's in a mean – an examination of the inconsistency between men and women in reporting sexual partnerships, *Journal of the Royal Statistical Society, Series A* **159**, 111–123.

JANE WADSWORTH* & LUKE WHITAKER

* Deceased July 1997

Genome-wide Significance

When **markers** distributed throughout a genomic region (perhaps 300–500 in present-day whole genome scans in humans) are used to map a **gene** or genes contributing to a phenotype (*see* **Genotype**) of interest, control of the overall false positive error rate involves a statistical issue of **multiple comparisons**.

For a single genomic locus t, possibly but not necessarily a marker locus, let $Z(t)$ be a statistic such that large values of $Z(t)$ are indicative of **linkage** of a trait to the locus t. The *nominal significance level* of the threshold b is $P_0\{Z(t) \geq b\}$. Here, P_0 denotes probability under the **null hypothesis** that the trait is unlinked to t. Suppose k genomic loci, $\{t_i : i = 1, \ldots, k\}$, each defined by its genomic map position given (for example) in centimorgans (cM) from a specified chromosomal location, are all tested for genetic linkage by means of $Z(t_i)$, $i = 1, \ldots, k$; and linkage to location t_i is declared if $Z(t_i)$ exceeds a suitable threshold, b. The *genome-wide significance level* for testing the hypothesis that *none* of the genomic loci is linked equals

$$P_0\left\{\max_i Z(t_i) \geq b\right\}. \tag{1}$$

Now the symbol P_0 denotes probability under the idealized hypothesis that none of the t_i is linked. Although with markers distributed throughout the genome it is unlikely that this idealized hypothesis is exactly true, it can, nevertheless, be regarded as a limiting approximation for the case that the phenotypic contributions of all linked genes are so small relative to the amount of data available that it is effectively impossible to detect those linkages. A slightly different interpretation is that (1) provides a bound on the probability that a false positive error occurs anywhere in the genome.

The simple **Bonferroni inequality** states that

$$P_0\left\{\max_i Z(t_i) \geq b\right\} \leq \sum_{i=1}^{k} P_0\{Z(t_i) \geq b\}. \tag{2}$$

In cases where the nominal significance levels are easily computed and the $Z(t_i)$ are not highly dependent (i.e. the t_i are sufficiently sparse that there is a reasonable amount of recombination between them), the Bonferroni inequality provides a simple, conservative approximation for the genome-wide significance level by the nominal significance levels. Its virtue is that it requires no assumptions about dependence among the different $Z(t_i)$. This is also its weakness, since by taking the dependence into account one can sometimes obtain more precise results.

For mapping quantitative traits in a backcross, Lander & Botstein [6] gave a simple approximation to (1) when the genomic locations t_i were an infinitely dense set of markers; they also provided simulated values when the t_i were markers equally spaced throughout an idealized tomato genome of 12 chromosomes of 100 markers each. Feingold et al. [5] showed that a similar analysis could be applied to allele-sharing statistics in **human genetics**, provided the sample size is large enough that a normal approximation to the distribution of the $Z(t_i)$ is reasonable; and they derived an approximation [(4) below] for the case of fully informative markers equally spaced at an intermarker distance Δ, which reduces to the earlier one in the limiting case that $\Delta = 0$.

In a debated, yet influential opinion piece, Lander & Kruglyak [7] argued that one should generally use the dense marker ($\Delta = 0$) approximation because (a) promising indications of linkage based on a sparse set of genomic locations was often followed up by saturating the promising area with what amounts to a dense set of markers; and (b) even in cases where this is not a feature of one particular study it is effectively the case when one considers the scientific community as a whole.

In particular studies, investigators have used **Monte Carlo methods** to determine the genome-wide significance level under the special conditions of their study. If direct simulation of the genome-wide significance level is forbiddingly complex, one may alternatively simulate the nominal significance levels and use the Bonferroni inequality to give a conservative approximation for the genome-wide significance level. This can be particularly effective when the nominal significance levels are all approximately the same.

Although simulations that are tailor made to the situation at hand are in many respects the best solution to the problem, simple analytic approximations can be useful to provide a rough check of a **simulation** program for gross errors and for theoretical comparisons of different experimental designs, when calculations must be repeated many times for different parameter values.

Remark

The problem of multiple comparisons in linkage analysis was recognized already in the classic paper of Morton [8], but it was dealt with differently. One assigned a prior probability of linkage to a randomly distributed marker and then determined a threshold for the detection of linkage that would make the posterior probability of linkage sufficiently large. See Ott [10] or Morton [9] for recent expositions of this idea, which with certain assumptions leads to the traditional threshold of 3 on the lod scale as the criterion to detect linkage. This analysis requires specific assumptions about the number, strength, and location of trait loci relative to linked markers. It seems adequate for Mendelian traits and a relatively small number of markers, where the linkage signal at a marker lying close by a trait locus is very strong and the principal impediment to the detection of linkage is the recombination fraction (distance) between a trait locus and the nearest markers. It seems less well suited for present-day genome scans designed to map genes for complex and/or quantitative traits. These genome scans involve large numbers of markers closely spaced throughout the genome, in order to map what may be multiple genes of variable penetrance, possibly interacting with each other and/or with the environment; and the major impediment to the detection of linkage is the absence of strong signals from the individual genes, even at markers lying next to or within the genes.

Approximation and Examples

Assume that fully informative markers are equally spaced at intermarker distance Δ throughout the genome. Let $Z(t_i)$ denote the statistic to be used for testing linkage to the marker t_i. We assume $Z(t)$ has been standardized so that at unlinked marker loci it has mean 0 and variance 1, and that it is reasonable to regard $Z(t)$ as approximately normally distributed. An additional condition is that as $s \to t$

$$\text{cov}[Z(t), Z(s)] = 1 - \beta|t - s| + o(|t - s|), \quad (3)$$

where $o(|t - s|)$ approaches 0 faster than $|t - s|$, and β is a parameter determined by the genetic relationships of the individuals contributing data to Z and on the form of the statistic Z.

A simple and important example is a sample of N independent sib pairs with $X(t)$ equal to the total number of alleles shared identical by descent (ibd) at the marker locus t (*see* **Identity Coefficients**). Let $Z(t) = [X(t) - N]/(N/2)^{1/2}$ be the ibd count standardized so that at an unlinked locus it has mean 0 and variance 1, while its mean is positive at a linked locus. Approximate normality of $Z(t)$ is a consequence of the central limit theorem if the number N of sib pairs is reasonably large. In this case $\beta = 0.04$/cM. A second example is the standardized regression statistic for testing that t is a quantitative trait locus in a backcross [6]. In this case, $\beta = 0.02$/cM. Other allele-sharing statistics and under certain conditions the signed square root of the (natural) log **likelihood ratio** statistic will also satisfy the required conditions. In human genetics the value of β will typically be slightly larger than 0.04, but its exact value will depend on the relationships of the individuals involved.

Under the assumed conditions

$$P_0\left\{\max_i Z(i\Delta) \geq b\right\} \approx 1 - \exp\{-C[1 - \Phi(b)] - L\beta b\varphi(b)\nu[b(2\beta\Delta)^{1/2}]\}. \quad (4)$$

In this formula, C is the number of chromosomes searched, L is their total genetic length, φ is the standard normal probability density function, Φ is the standard normal distribution function, and $\nu(x)$ is a special function, which is easily computed numerically and in the range $0 \leq x \leq 2$ is well approximated by $\exp[-0.583x]$.

For a numerical example, we consider an idealized human genome of 23 chromosomes of average genetic length of 140 cM. For intermarker spacings of $\Delta = 0, 1, 5$ and 10 cM, (4) yields a genome-wide significance level of 0.05 at the thresholds $b = 4.08$, 3.91, 3.73 and 3.6, respectively. On the lod scale, these thresholds are 3.62, 3.32, 3.02 and 2.82. The Bonferroni inequality (2) provides a very good approximation when $\Delta \geq 10$ cM, but it becomes overly conservative as Δ approaches 0. For a backcross, the smaller value of β indicates greater dependence between the values $Z(t_i)$ with the result that the conservatism of the Bonferroni inequality becomes apparent for somewhat larger Δ.

A number of related cases have been studied in detail and are reasonably well understood.

1. For qualitative traits studied using sib pairs or for quantitative traits in either an intercross in experimental genetics or in sibships, one may use a two degrees of freedom statistic in order to have more power to detect a gene having an additive effect and dominance deviation. A similar issue arises if we fix one locus, thought to be a trait locus, and search conditionally for a second locus using a model that provides for interaction between the two loci. Appropriate modifications of (4) are given by Dupuis & Siegmund [2]. For an intercross and an idealized mouse genome of 20 chromosomes of average length 80 cM, the 0.05 genome-wide thresholds for the same intermarker spacings used above are 4.43, 4.28, 4.12 and 4.01, respectively.
2. In experimental genetics, fairly large values of Δ (say $\Delta \approx 20$ cM) are common. If $Z(t)$ is maximized only over marker locations, either (2) or (4) (perhaps as modified in point 1 above for an intercross) would provide an adequate approximation. However, following the suggestion of Lander & Botstein [6], one often computes *estimated* values of $Z(t)$ at a dense set of positions between markers, then uses both the interpolated and the actual data. This leads to a larger value for $\max_t Z(t)$ and requires a slightly larger threshold. The Lander–Kruglyak suggestion to use (4) with $\Delta = 0$ is one possibility, but it is typically overly conservative. Appropriate approximations for genome-wide significance levels have been given by Rebai et al. [11, 12] and by Dupuis & Siegmund [1]. For the idealized mouse genome of the preceding example and $\Delta = 20$ cM, this would lead to an increase in the 0.05 genome-wide threshold from $b = 3.88$ to $b = 3.99$.
3. In human genetics when (a) sample sizes are small, (b) pedigrees contain more distant relatives than sibs or half sibs, or (c) there are more than two affecteds in a pedigree, the distribution of $Z(t)$ at a fixed genomic location t may be skewed. A simple modification of (4) has been suggested by Tang & Siegmund [13]. Since this is fundamentally a correction for non-normality of the marginal distribution of $Z(t)$, it can also be adapted for use with the Bonferroni inequality. For related approximations, which may be more accurate but require stronger assumptions and more complicated calculations, see [4] and [15].
4. For partially informative markers, there is evidence [14] that (4) is conservative but still provides a reasonable approximation if founders are available for genotyping and multipoint analysis is used. It appears that (4) can be overly conservative if founders are unavailable and anticonservative if multipoint analysis is not used. In the latter case the Bonferroni inequality may still be useful.
5. Simultaneous search for multiple linked genes involves a substantially larger number of tests and a larger threshold than the simple genome scans discussed in this article [3].

References

[1] Dupuis, J. & Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers, *Genetics* **151**, 373–386.

[2] Dupuis, J. & Siegmund, D. (2000). Boundary crossing probabilities in linkage analysis, in *Game Theory, Optimal Stopping, Probability and Statistics*, F. Thomas Bruss & L. Le Cam, eds. Institute of Mathematical Statistics, Hayward, pp. 141–152.

[3] Dupuis, J., Brown, P. & Siegmund, D. (1995). Statistical methods for linkage analysis of complex traits from high resolution maps of identity by descent, *Genetics* **140**, 843–856.

[4] Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method, *Journal of Applied Probability* **30**, 766–779.

[5] Feingold, E., Brown, P.O. & Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent, *American Journal of Human Genetics* **53**, 234–251.

[6] Lander, E.S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* **121**, 185–199.

[7] Lander, E.S. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**, 241–247.

[8] Morton, N.E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.

[9] Morton, N.E. (1998). Significance levels in complex inheritance, *American Journal of Human Genetics* **62**, 690–697.

[10] Ott, J. (1991). *Analysis of Human Genetic Linkage*, Revised Ed. Johns Hopkins University Press, Baltimore.

[11] Rebai, A., Goffinet, B. & Mangin, B. (1994). Approximate thresholds of interval mapping test for QTL detection, *Genetics* **138**, 235–240.

[12] Rebai, A., Goffinet, B. & Mangin, B. (1995). Comparing power of different methods for QTL detection, *Biometrics* **51**, 87–99.

[13] Tang, H.-K. & Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models, *Biostatistics* **2**, 147–162.

[14] Teng, J. & Siegmund, D. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers, *Biometrics* **54**, 1247–1265.

[15] Tu, I.-P. & Siegmund, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis, *Advances in Applied Probability* **31**, 510–531.

(*See also* **Linkage Analysis, Model-free**)

DAVID SIEGMUND

Genotype

The genotype of an organism is its **genes**, or genetic make-up, as opposed to its phenotype, or outward appearance. The physical basis of the human genotype lies in 23 pairs of chromosomes–microscopic bodies present in every cell nucleus. One pair are the sex chromosomes, and the other 22 pairs are known as autosomes, or autosomal chromosomes. The two alleles at each locus on the autosomes comprise the genotype for that locus. If the two alleles are the same, the genotype is called homozygous; if they are different, the genotype is called heterozygous. Persons with homozygous and heterozygous genotypes are called homozygotes and heterozygotes, respectively (*see* **Heterozygosity**).

If the phenotype, or phenotypic distribution, associated with a particular heterozygote is the same as that associated with one of the two corresponding homozygotes, then the allele in that homozygote is dominant, and the allele in the other corresponding homozygote is recessive, with respect to the phenotype; the locus is said to exhibit dominance. If the heterozygote expresses a phenotype that has features of both corresponding homozygotes–for example, persons with AB blood type have both A and B antigens on their red cells, determined by A and B alleles at the ABO locus–then there is said to be codominance (*see* **Blood Groups**). At the DNA level, that is, if the phenotype associated with a genotype is the DNA constitution itself, then all loci exhibit codominance.

The genotype being considered may involve the alleles at more than one locus. However, a distinction should be made between the genotype at multiple loci and the multilocus genotype. Whereas the former is specified by all the alleles at the loci involved, the latter is specified by the two haplotypes a person inherited, that is, the separate sets of maternal alleles and paternal alleles at the various loci (*see* **Haplotype Analysis**).

In the case of a quantitative trait, there is a dominance component to the variance if the heterozygote phenotype is not half-way between the two corresponding homozygote phenotypes, that is, if the phenotypic effects of the alleles at a locus are not additive. Similarly, if the phenotypic effect of a multilocus genotype is not the sum of the constituent one-locus genotypes, then there is epistasis. Dominance can be thought of as intralocus **interaction** and epistasis as interlocus interaction. Thus, in the case of a quantitative phenotype, the presence or absence of dominance and/or epistasis depends on the scale of measurement of the phenotype.

ROBERT C. ELSTON

Genotyping and Error-checking

Classifying Genetic Errors

Errors in genetic data can be classified into two categories: pedigree errors and genotyping errors. Pedigree errors arise when family relationships of the individuals under study are misspecified. Possible sources of misspecification include unrelated individuals such as adopted or switched sibs, nonpaternity cases such as half sibs who are recorded as full sibs, faulty family records and incorrect data entry (*see* **Paternity Testing** and **Relationship Testing** for more discussion of errors in family relationships). Genotyping errors arise when the true underlying **genotype** of an individual under study is misspecified. Possible sources include laboratory errors such as switched or contaminated samples, incorrect allele calls, or mistakes in data entry.

Genotyping error rates commonly reported in mapping and **linkage** studies range from 1% to 3% [1, 2, 16]. Genotyping errors in general either negate true recombinants or introduce false recombinants, as illustrated in the following equation for the expected recombination rate [21]:

$$\mathrm{E}(\hat{\theta}) = \theta_t + (1 - 2\theta_t)s,$$

where θ_t is the true recombination fraction and s is the misclassification frequency. For $\theta_t = 0.5$, unlinked **markers**, there is no **bias**, but for θ_t less than 0.5, the expected recombination rate is inflated by $(1 - 2\theta_t)s$. When θ_t is close to 0, $\mathrm{E}(\hat{\theta}) \approx s$, indicating most recombination events are false. False recombinants inflate estimated map distances between markers, obscure correct marker orders, and deflate lod scores, thus reducing the power to detect linkage. Negating true recombination events may inflate lod scores and lead to false linkage results.

Pedigree error rates due to nonpaternity vary widely among different populations that have been studied [3, 23]. Unlike genotype errors, however, pedigree errors that identify a previously unknown nonpaternity or adoption can have a profound effect on participants in the study. Common practice is not to reveal such errors to the participants.

Genotype Error Detection

Detecting errors in genetic data comprising a small number of markers and individuals may only require simple visual inspection, and resolving them may only require a quick inspection of a few genotypes in a database or a few entries from a family history when data entry error is the culprit. However, to handle the large number of genotypes generated for genome scans and the varied family structures used in linkage analysis, a variety of algorithmic and statistical methods have been developed that offer more powerful detection methods and can be implemented into software to automate the task of "debugging" genetic data. References to the methods contain details on the availability of software.

Violations of Mendelian Transmission

Since most genetic analysis software programs require the marker data to be **Mendelian** consistent, errors are often first detected after the program fails to execute. A marker is Mendelian consistent if the underlying genotype of the individuals in the pedigree are consistent with the laws of Mendelian transmission, otherwise the marker is Mendelian inconsistent. Although a genotype that contains a mutated allele (*see* **Gene**) may cause a Mendelian inconsistency, an inconsistency more often indicates the presence of an error, since mutation rates are much lower than genotyping error rates. Determining whether genotype data are Mendelian consistent can be done using genotype elimination algorithms [14, 20]. These algorithms use available genotyping information and family relationships to infer the sets of consistent genotypes of individuals with missing information. The pedigree has a Mendelian inconsistency if and only if, after the genotype elimination is performed, there is an individual with no consistent genotypes. Although genotype elimination is guaranteed to establish the existence of a Mendelian inconsistency, the algorithm may not be useful in identifying the possible source(s) of the inconsistency, since the effect of an inconsistency may propagate to different parts of the pedigree. Factors that improve the accuracy of identifying the source(s) are fewer individuals with missing data, highly polymorphic markers and larger sibships. The effect of the latter two factors was demonstrated by Gordon et al. [7], who studied the probability of detecting

genotyping errors in diallelic single-nucleotide **polymorphism** data for parent–offspring triples. They showed that in the case of this important sampling scheme for **linkage disequilibrium** studies the true error rate is over three times that of the error rate that would be reported using Mendelian consistency checks. Another very widely used paradigm in disease studies where Mendelian consistency checks are inadequate is sibling pairs. When parents are missing genotype information, two putative sibs will always be Mendelian consistent at codominant markers regardless of genotyping and/or pedigree error. Thus, alternative methods of error detection have been proposed.

Haplotyping

One method commonly used to check Mendelian consistent data is haplotyping [24]. Haplotyping is the processes of constructing the most likely gene flow in the pedigree assuming a known marker order, allele frequencies and intermarker recombination fractions. Finding double recombinants between closely spaced markers and/or many more recombinants than expected (based on the number of meiosis and recombination fractions in the most likely haplotype configuration) points to possible genotyping error. The location of suspect recombinants may not indicate the location of an error, since there may be other equally likely **haplotype** configurations with different locations of the recombinants. Thus, several **likelihood** methods that focus on identifying likely individuals and markers have been proposed.

Incomplete Penetrance Error Model

The genotyping error for an individual can be modeled within the likelihood at a codominant autosomal marker by defining an incomplete **penetrance** function [15]. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a vector of phenotypes, which are the observed genotypes, and $\mathbf{g} = (g_1, g_2, \ldots, g_n)$ be the vector of actual genotypes for n individuals in a pedigree. Let $G = m(m+1)/2$ be the number of combinatorially possible unordered genotypes given m alleles at the marker. If we assume the errors are independent and equally likely, then the conditional probability of the observed genotype x_j given the true genotype is g_j is

$$\Pr(x_j|g_j) = \begin{cases} 1-\varepsilon, & \text{if } x_j = g_j, \\ \dfrac{\varepsilon}{G-1}, & \text{if } x_j \neq g_j, \end{cases} \qquad (1)$$

where ε is the error rate. For an error rate of 3% this penetrance function assigns a high probability to the observed genotyping being correct and a low but equal probability to all other $G-1$ possible genotypes. More complex penetrance models assuming individual and locus-specific error rates can be used, but studies have shown that the penetrance model is relatively insensitive to the distribution of the error probability [5, 17, 25].

Although implementing an incomplete penetrance model into the likelihood of the pedigree is mathematically straightforward, the increased computational requirements of the model may be prohibitive, since each genotyped individual now has G nonzero penetrance values instead of 1. The complexity is less of an issue for small- to medium-sized pedigrees since the likelihood can be computed using the Lander–Green algorithm [13]. But for larger pedigrees that require the **Elston–Stewart algorithm** [6] to compute the likelihood, even analyzing a single polymorphic marker may be difficult when G is large. To reduce the complexity, several authors use a reduced incomplete penetrance model [5, 25], whereby only a single individual at a single marker has incomplete penetrance and all other genotypes are assumed correct.

Likelihood Methods

Stringham & Boehnke [25] proposed computing the posterior probability of genotyping error, $\Pr(E_k|\mathbf{x};\varepsilon)$, for each individual at each locus allowing for incomplete penetrance. They compute $\Pr(E_k|\mathbf{x};\varepsilon)$ using standard likelihood methods as $\Pr(E_k, \mathbf{x};\varepsilon)/\Pr(\mathbf{x};\varepsilon)$, where $\Pr(\mathbf{x};\varepsilon)$ is the likelihood for the data and $\Pr(E_k, \mathbf{x};\varepsilon)$ is the likelihood for the data assuming individual k has an error, which requires adjusting his/her penetrance function so the first term is zero. Individuals with high values are flagged as possible errors. They also proposed a less computationally intensive version using reduced incomplete penetrance, which can identify individuals who could be the sole source of a Mendelian inconsistency. Ott [22] suggested calculating for each individual k the sum of the squared differences between the conditional

probability for each genotype g_k given the phenotype x_k and each genotype g_k given the phenotypes of all the individuals in the pedigree **x**:

$$\sum[\Pr(g_k|x_k) - \Pr(g_k|\mathbf{x})]^2.$$

Individuals with large values are identified for further inspection. O'Connell & Weeks [19] proposed a method to identify the most likely source of Mendelian inconsistencies by comparing the most likely alternative Mendelian-consistent genotype configurations. The method first identifies minimal sets of individuals who could eliminate the inconsistencies by setting their phenotype to unknown, and then computes the likelihood of each possible consistent alternative phenotype for these sets of individuals. Sets of individuals with the highest likelihood are targeted as the most likely source of error, since they have the highest probability of having other genotypes than those observed.

Besides single marker tests, several authors have proposed multipoint tests, which have the advantage of being able to detect increased recombination and inflated map distances. Ehm et al. [5] proposed a hypothesis test of $\varepsilon > 0$ for each individual at each locus, assuming a reduced incomplete penetrance model. They proposed an exact multipoint **likelihood ratio test** statistic that assumed correct marker order, but jointly estimated the recombination fractions and the error rate. Owing to the complexity of the joint estimation, they also proposed an approximate test that used an estimate of the recombination fractions assuming no errors. ***P* values** for the test are determined by Monte Carlo **simulation**. Their simulation studies showed that the power of their method is directly related to the amount of typing in the pedigree and the number of loci analyzed jointly, and is inversely related to the recombination fraction. Douglas et al. [4] investigated the power of the multipoint posterior probability test assuming correct marker order for the special case of sibling pair pedigrees with no parents available. These pedigrees are small enough to allow incomplete penetrance for each individual and each marker. They showed that posterior error rate cutoffs vary substantially as a function of marker density, prior error rate and false positive rate. Although their method found less than 50% of the errors, they showed that the errors identified using their cutoffs were the most informative for restoring the true linkage information.

Although the above tests for genotyping errors assume correct pedigree structure, putative genotype errors may actually result from pedigree errors. For example, an adopted sib may be flagged as having genotyping errors, although the genotypes are correct. The likelihood depends not only on the penetrance, but also on allele frequencies and parent–child transmission probabilities. The effect of pedigree errors through the transmission probabilities may offset various penetrance values, just as a genotype with a very rare allele may have a high posterior probability of error due the small allele frequency. Lathrop et al. [16] proposed a method to distinguish between the two types of error by including an error model for parental misassignment with the incomplete penetrance model.

Resolving Errors

Resolving errors in genetic data by determining the true state of a pedigree relationship or genotype may not always be possible. Correcting data entry errors is rather easy, but resolving putative genotyping errors requires regenotyping the individuals involved. However, since regenotyping may not always be an option, due to cost or insufficient resources, the general practice is to prune the data. The methods presented can be used to identify the most appropriate genotypes to prune. Care must be taken, however, not to artificially increase evidence of linkage by removing false positives. Errors identified as pedigree errors due to misspecified relationships can often be corrected by removing individuals or by adding the appropriate relatives to the pedigree. For example, newly identified half sibs can be retained by adding in another parent, if necessary.

Other approaches besides pruning the data have been proposed. Morton & Collins suggest adjusting the recombination fractions for errors to take into account possible inflated map distance [18]. Göring & Terwilliger [9–12] propose a methodology to allow for errors in linkage and/or **linkage disequilibrium** analysis. Gordon et al. [8] propose a transmission/disequilibrium test for single-nucleotide polymorphism data that allows for errors. Finally, as stated above, the incomplete penetrance model can be incorporated into any exact likelihood-based analysis for small- to medium-sized pedigrees; for larger pedigrees, an approximate likelihood can be computed using **Markov chain Monte Carlo** methods.

Since genotyping error will probably be always with us, finding improved methods to identify errors or to allow for them in our analyses to extract maximum information from our data will continue to be an important area of research.

References

[1] Brzustowicz, L.M., Merette, C., Xie, X., Townsend, L., Gilliam, T.C. & Ott, J. (1993). Molecular and statistical approaches to the detection and correction of errors in genotype databases, *American Journal of Human Genetics* **53**, 1137–1145.

[2] Buetow, K.H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes, *American Journal of Human Genetics* **49**, 985–994.

[3] Cerda-Flores, R.M., Barton, S.A., Marty-Gonzalez, L.F., Rivas, F. & Chakraborty, R. (1999). Estimation of nonpaternity in the Mexican population of Nuevo Leon: a validation study with blood group markers, *American Journal of Physical Anthropology* **109**, 281–293.

[4] Douglas, J.A., Boehnke, M. & Lange, K. (2000). A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data, *American Journal of Human Genetics* **66**, 1287–1297.

[5] Ehm, M.G., Kimmel, M. & Cottingham, R.W. Jr (1996). Error detection for genetic data, using likelihood methods, *American Journal of Human Genetics* **58**, 225–234.

[6] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[7] Gordon, D., Heath, S.C. & Ott, J. (1999). True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms, *Human Heredity* **49**, 65–70.

[8] Gordon, D., Heath, S.C., Liu, X. & Ott, J. (2001). A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data, *American Journal of Human Genetics* **69**, 371–380.

[9] Goring, H.H. & Terwilliger, J.D. (2000). Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes, *American Journal of Human Genetics* **66**, 1095–1106.

[10] Goring, H.H. & Terwilliger, J.D. (2000). Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions, *American Journal of Human Genetics* **66**, 1107–1118.

[11] Goring, H.H. & Terwilliger, J.D. (2000). Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters, *American Journal of Human Genetics* **66**, 1298–1309.

[12] Goring, H.H. & Terwilliger, J.D. (2000). Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified, *American Journal of Human Genetics* **66**, 1310–1327.

[13] Lander, E.S. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans, *Proceedings of the National Academy of Sciences* **84**, 2363–2367.

[14] Lange, K. & Goradia, T.M. (1987). An algorithm for automatic genotype elimination, *American Journal of Human Genetics* **40**, 250–256.

[15] Lathrop, G.M., Hooper, A.B., Huntsman, J.W. & Ward, R.H. (1983). Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker mistyping, *American Journal of Human Genetics* **35**, 241–262.

[16] Lathrop, G.M., Huntsman, J.W., Hooper, A.B. & Ward, R.H. (1983). Evaluating pedigree data. II. Identifying the cause of error in families with inconsistencies, *Human Heredity* **33**, 377–389.

[17] Lincoln, S.E. & Lander, E.S. (1992). Systematic detection of errors in genetic linkage data, *Genomics* **14**, 604–610.

[18] Morton, N.E. & Collins, A. (1990). Standard maps of chromosome 10, *Annals of Human Genetics* **54**, 235–251.

[19] O'Connell, J.R. & Weeks, D.E. (1998). PedCheck: a program for identification of genotype incompatibilities in linkage analysis, *American Journal of Human Genetics* **63**, 259–266.

[20] O'Connell, J.R. & Weeks, D.E. (1999). An optimal algorithm for automatic genotype elimination, *American Journal of Human Genetics* **65**, 1733–1740.

[21] Ott, J. (1991). *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, pp. xxii, 302.

[22] Ott, J. (1993). Detecting marker inconsistencies in human gene mapping, *Human Heredity* **43**, 25–30.

[23] Sasse, G., Muller, H., Chakraborty, R. & Ott, J. (1994). Estimating the frequency of nonpaternity in Switzerland, *Human Heredity* **44**, 337–443.

[24] Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics, *American Journal of Human Genetics* **58**, 1323–1337.

[25] Stringham, H.M. & Boehnke, M. (1996). Identifying marker typing incompatibilities in linkage analysis, *American Journal of Human Genetics* **59**, 946–950.

(*See also* **Linkage Analysis, Model-based**; **Linkage Analysis, Model-free**)

JEFFREY R. O'CONNELL

Geographic Epidemiology

The purpose of this article is to give an overview of methods of analyzing epidemiologic data in which geographic location is of primary importance. Methods may be distinguished by the purposes of the analyses, which include (i) modeling **risk** as a function of geographically referenced variables; (ii) **hypothesis testing** about specific sources of risk; (iii) **mapping disease patterns** to provide a visual representation of risk variation; (iv) identifying areas of apparently elevated risk deserving further epidemiologic investigation (*see* **Environmental Epidemiology**); and (v) detecting specific or generalized clusters, which may be indicative of unsuspected sources of risk or of a contagious disease mechanism (*see* **Clustering**). These methods are reflected in the central sections of the article, which are preceded by an overview of the underlying models and followed by a discussion of the issues involved in the choice of method.

Interest in geographic epidemiology has increased greatly in recent years and the number of methods proposed in the literature is very large; two recent books are given as references [44, 64]. It would be quite impossible to review these methods comprehensively in this article; rather the object is to form a general classification according to their objectives and the underlying assumptions. Many proposed analyses will inevitably be omitted and an adverse judgment on them should not be inferred. On the face of it, geographic epidemiology is a topic in spatial statistics [27, 31, 36, 85], but there are special aspects which distinguish it from many of the other areas of application in the latter field.

It should be remarked at the outset that the likelihood of a geographic analysis revealing relationships of real scientific or clinical significance in a given case may be fairly low, for a number of reasons. First, in spite of considerable improvements in data availability over recent years, it is difficult to acquire accurate population sizes. A relatively high rate of individual migration and the development of new residential districts mean that population estimates can be seriously in error at the end of the intercensal period – 10 years in the UK, for example. Migration also implies that people are typically exposed to risks associated with different locations as they move around, which must inevitably dilute the sensitivity of any geographic analysis. Local mobility further confuses the picture: adults work and children go to school in areas which are often quite different from their place of residence, so that geographically mediated risks may be only weakly related to home address. It should also be remembered that most geographic observations – including those that have subsequently been found to be of real significance and value – have been anecdotal and the analyses have been executed *post hoc*. This inevitably adds to the difficulty of interpreting them.

Nevertheless, there have been notable triumphs of geographic epidemiology: the well-known story of John Snow and cholera [91]; the observations of Denis Burkitt leading to the recognition of a vector-related etiology of a human tumor [20, 21]; and more recently, the detection of the cause of an outbreak of epidemic asthma [3]; all are striking demonstrations of the epidemiologic potential of geographic observations. There is also a less positive but equally important reason for carrying out geographic analyses effectively. People are in practice very concerned about the impact of their environment on their health and it is important that anxieties are explored in a manner that inspires confidence, if only to provide reassurance in particular cases.

Underlying Models

Case-independence and Other General Issues

It is important to distinguish at the outset mechanisms of disease in which cases are or are not intrinsically related. Examples of case-dependence include contagious and familial diseases (*see* **Communicable Diseases**); any tendency of such cases to be close geographically is unlikely to identify any component of risk which is essentially geographic. Rather, interest centers on establishing the case-dependence as an intrinsic phenomenon independent of geographic location.

For case-independent disease processes, however, we suppose that cases occur independently of one another, *conditional on the underlying risk factors*, which may include some aspect of geographic location. On this assumption, any spatial **autocorrelation**, or general tendency of cases to be closer to one another than expected, will be inherited from the spatial structure of the underlying pattern of risk factors.

Within the frequentist framework, at least, it is *unnecessary to model it by means of a spatially correlated error process*, except to the extent that this may be a convenient way of allowing for spatially varying risk factors that we cannot observe. We return to this point below.

Models for case-dependent processes are not well developed. Spatial modeling of epidemics [4, 72] is concerned with the dynamics of spatial spread and has rather little interaction with epidemiologic analysis in the sense of this article. Some authors [1, 5, 16] have proposed clustering models in which "parent" cases give rise to "offspring" cases according to a defined stochastic mechanism; it is much easier to postulate models of this kind than to handle them analytically. "Second-order" **point processes**, which model the tendency of points to be clustered, have been studied using Ripley's K functions [84] in the epidemiologic context by Diggle & Chetwynd [37] and applied also by Diggle & Morris [39]; they are not easily related to specific models of person–person interaction, however.

In practice, much geographic epidemiology is applied to diseases – notably malignant diseases – for which there is very little evidence of case-dependency. Although it may be the objective of some analyses to detect such dependency, the assumption of independence is a reasonable basis at least for a **null hypothesis** H_0 of spatial uniformity. It is therefore quite sensible to discuss most of the methods of geographic epidemiology against a modeling background which assumes case-independence and this will be the standpoint of this article. Once this position is accepted, the basic models for the spatial distribution of disease become relatively simple.

Two other general aspects of the models we discuss should be mentioned. In the first place, it may be argued that much of epidemiologic analysis reduces to considering the **associations** between observations on a disease $\mathcal{D}$, a variable of primary interest $\mathcal{E}$, such as an exposure variable, and a set of other variables $\mathcal{C}$, which may be thought of as **covariates** and which may include possible **confounding** variables. In many analyses the variables in $\mathcal{C}$ are regarded as being fixed, even though they may be subject to error or sampling variation. For many purposes this will suffice; although the analysis may not be strictly correct [41], it can be justified in terms of a conditional argument – i.e. it is valid in the subspace of all outcomes in which $\mathcal{C}$ is as observed; alternatively, we may argue that it is reliably assessing the importance of a *modified* variable incorporating the unobservable error. In **observational studies** $\mathcal{D}$ and $\mathcal{E}$ are intrinsically random, but even here it is quite usual to condition on one or the other rather than to model the full joint distribution. This gives rise to a duality of analysis corresponding to the **case–control** and **cohort** approaches [19]. In this respect, geographic analyses are no different from any other epidemiologic analyses: geographic location may be treated as a covariate $\mathcal{C}$ or as a primary interest variable $\mathcal{E}$; in the latter case, $\mathcal{E}$ may be regarded as fixed with $\mathcal{D}$ random, or vice versa. This "duality principle", that either form of conditioning leads to valid and useful analyses of epidemiologic data, permits us to employ analyses of either kind interchangeably, which we will do below without further comment.

Secondly, we emphasize that, however we choose to model location, it is imperative to build in a reference distribution (normally in the form of a population density) at a fundamental level; not least this is because of the implication for the underlying variation in local **variance**. Analyses which start by assuming a homogeneous spatial distribution of cases and "correct for" heterogeneity of population distribution are to be regarded with suspicion and may give misleading results.

Areal Data

Most geographic data are in areal form, i.e. counts $Y_i, i = 1, 2, \ldots, k$ of numbers of cases in areas A_i, nearly always administratively defined, within a study region $\mathcal{R}$. These will be accompanied by population information, initially in the form of the sizes of the populations at risk in different relevant groups; such groupings will usually include age and sex, and often other factors such as socioeconomic status and ethnic group. If the risk within a given area A_i is constant, the Y_i are clearly **binomial**. Variation of this risk upsets this assumption, but only slightly in practice, particularly if the risk is small (which it will be for a rare disease or when modeling annual rates). In the latter case we can use the **Poisson** approximation even if there is identifiable heterogeneity of risk within the A_i.

We can therefore simplify the account by adopting the Poisson model and supposing that the population sizes have already been used in conjunction with reference rates to construct "expected" numbers e_i of

disease cases, either by a process of **standardization** or using a suitable **regression** model [13]. The case-independence assumption then implies that the Y_i are independently distributed with Poisson distributions having means $\theta_i e_i$, say, where we take θ_i as the **relative risk** (RR) in A_i – i.e. the risk relative to the assumptions under which the $\{e_i\}$ were computed. The factors θ_i may now be modeled in terms of possible explanatory variables using methods which are well understood, for example a **Poisson regression**; this employs a **loglinear model**, which is a particular case of a **Generalized Linear Model (GLM)** [66]. Such a model can be tested for **goodness of fit**; usually this may be satisfactorily accomplished using the residual deviance, though there are problems in interpreting this if the e_i are very small – say with appreciably many observations significantly less than around five. If there is evidence that the model does not fit well, this indicates that there is a component of risk that has not been incorporated into the model; it is said that there is "extra-Poisson" variation [17]. Modifications of the analysis in this case include postulating a distribution for θ over the A_i, for example from a **gamma distribution**, which leads to a **negative binomial distribution** for the Y_i [53]. The implications for model validity and **inference** do not depend on whether there is a geographic element to this variation; if there is, a richer model involving spatial autocorrelation may be fitted [28].

Continuous Data

The discrete structure of areal data, with the imposition of administrative boundaries, is not ideal either practically or mathematically. In principle, data may be available on a continuous basis, i.e. by the provision of the exact geographic coordinates of a sample of cases. The case-independence assumption and the duality principle then imply that these cases may be regarded as a **random sample** from a **bivariate** density function $\psi(x, y)$ of geographic location with coordinates (x, y). This density function determines the distribution of the place of residence of a randomly selected individual with the disease. Equivalently the case locations may be regarded as a realization of a **Poisson process** with intensity $f(x, y)$ proportional to $\psi(x, y)$. As before, we need a reference group and this may be taken to be the population density $\pi(x, y)$, where we use this term to refer not so much to a general demographic concept, but to a second mathematical density function describing the probability that a randomly selected member of the population will reside at a given point (x, y). We are now in a position to define a *risk function* giving the risk of being affected by the disease incurred by a randomly selected individual at location (x, y), namely.

$$\mu(x, y) = \frac{f(x, y)}{N\pi(x, y,)},$$

where N is the aggregate of the population in $\mathcal{R}$, in the form of **person-years at risk** if appropriate.

In practice, of course, these mathematically defined risk functions must be estimated from the data, the problem being one of estimating the ratio of two densities. Methods of estimating a single density have recently been much studied and developed [88, 89] and, although ratios present certain rather special problems, they are not insuperable. For the distribution of cases, the general methods apply, with the proviso that population-related densities are extremely multimodal and in this respect atypical of most of the examples to which density estimation has been applied. The estimation of the population density raises rather special problems, in that published data will still be in areal form, but the normal methods of **density estimation** can be adapted; alternative methods are available [69, 94].

It should also be noted that the population density may be satisfactorily estimated from a sample of **controls** [8], which has the advantage that the geographic resolution of their locations will be equal to that of the cases. Use of controls also removes the need to ascertain the entire population, though in the absence of information about whole-population risk the risk function will be determined only up to an unknown factor. In this case it may be regarded as a "relative risk function" (RRF), defined by $\theta(x, y) = \psi(x, y)/\pi(x, y)$, giving the risk at a particular location relative to the average in the whole of $\mathcal{R}$ [8].

The continuous model of geographic risk is attractive mathematically and opens up a number of new possibilities for analysis. In practice, however, there are considerable difficulties with obtaining suitable **sampling frames** for the controls. Case–control analyses have not been much used in practice, though they may reasonably be expected to assume greater significance in the future, as the geographic accuracy of address data improves.

Spatial Autocorrelation

The rationale for taking account of spatial autocorrelation is attributed by Cook & Pocock [28] to Lazar [65] as being that "failure to allow for spatially correlated errors may result in serious overestimates of the significance of relationships". Lazar demonstrates that this is true when the relationships in question are assessed using **correlation** coefficients with precision estimated from the data.

As argued above, however, case-independence implies that any spatial autocorrelation in observed disease rates must be due to a similar autocorrelation in one or more associated factors that have not been taken into account in the model, i.e. it is not really the *error* mechanism that should be modified by taking account of autocorrelations. Of course, the word "error" in statistical parlance has come to be synonymous with anything not accounted for in an explanatory model. The point is not merely semantic, however; use of the word in this context has the unfortunate effect of distracting attention from the importance of case-independence and to the construction of methods which do not make use of the known variances of binomial and Poisson data.

Cook & Pocock propose, in a study of heart disease, a model for the Standardized Mortality Ratio (SMR) (*see* **Standardization Methods**) of the form:

$$\ln\left(\frac{Y_i}{e_i}\right) = \mathbf{X}\beta + \varepsilon,$$

where the linear predictor $\mathbf{X}\beta$ is in the usual form for a **linear regression** model and

$$\varepsilon \sim \mathrm{N}[0, \sigma^2\mathbf{A}],$$

with $\mathbf{A}$ an autocovariance matrix. Exploratory analysis led them to propose an autocorrelation function decaying **exponentially** with distance. Estimation in this model did indeed lead to a reduction in the significance of the effect of water hardness, the regression coefficient being reduced by around 40% and its standardized value from -5.0 to -3.0 (*see* **Standardized Coefficients**).

Cook & Pocock conclude that failure to adjust for spatial autocorrelation leads us to overstate the significance of fitted regression coefficients. Although this is probably true in their example, it may not always be so, as may be seen more easily by considering a GLM in which terms may be assessed for significance in their own right, without recourse to an estimate of residual variance. It would be quite possible to have an unobserved covariate $\mathcal{C}$ which is spatially autocorrelated and which induces autocorrelations in the residuals, but which is independent of a fitted variable $\mathcal{E}$. Although taking account of $\mathcal{C}$ would improve the fit of the model, it would not necessarily reduce the contribution of $\mathcal{E}$ to the total deviance and need not, therefore, affect its significance. Nor is the reduction of the estimate of the regression coefficient conclusive, since $\mathcal{C}$ might be incidentally associated with $\mathcal{D}$ and $\mathcal{E}$; the latter could still be an important causative factor. Incorporation of autocorrelation is, nevertheless, a feature of many contributions to the field and it is generally supposed to be of considerable importance, both theoretically and practically. To some extent this view is encouraged by **Bayesian** modeling, where the emphasis is on the inclusion of terms to represent any unknown component of variation without consideration of model **parsimony**.

Bayesian Modeling

The models outlined above are essentially frequentist and assume the existence of unique but unknown parameters. The Bayesian alternative is becoming increasingly popular in statistics generally and particularly with the epidemiologic community [25, 32, 50, 55].

The seminal contribution by Clayton & Kaldor [26] distinguishes heterogeneity deriving from spatial and nonspatial sources. These give rise to corresponding methods of **smoothing rates** for disease mapping (see below) and were motivated primarily by this application. The methods have had considerable influence on geographic epidemiology and have been used to elaborate inferential modeling by numerous authors [63, 71] in the spirit of the frequentist approach outlined above. They are exemplified by the analysis by Richardson et al. [83] of childhood leukemia in relation to natural (background) **radiation**. These authors extend the standard Poisson regression to a Generalized Linear Mixed Model (GLMM) [18, 24], with **random effect** terms u_i, v_i specific to area A_i:

$$\ln\theta_i = \mathbf{X}_i\beta + u_i + v_i, \quad i = 1, 2, \ldots, k.$$

Here u_i models nonspatial heterogeneity through the assumption that

$$[u_i|u_j,\ j \neq i] \sim \mathrm{N}[\overline{u}_i, \lambda_u^{-1}], \quad \text{where}$$
$$\overline{u}_i = \sum_{j \neq i} \frac{u_j}{(k-1)}.$$

Spatial structure is modeled by v_i through the assumption that

$$[v_i|v_j,\ j \neq i] \sim \mathrm{N}[\overline{v}_i, (k_i\lambda_v)^{-1}], \quad \text{where}$$
$$\overline{v}_i = \sum_{j \neq i} \frac{W_{ij}v_j}{k_i},$$

and where the adjacency matrix element $W_{ij} = 1$ if areas i, j are adjacent, 0 otherwise, and k_i is the number of areas adjacent to A_i. We require $\sum u_i = \sum v_i = 0$ for **identifiability**.

The parameters λ_u, λ_v control the degree of variation in the dispersion terms. In an **empirical Bayes** treatment, these would be estimated; instead Richardson et al. pursue a "full Bayes" solution in which, together with β, they are given **prior distributions**. Choice of the "hyperparameters" in such prior distributions leads to the notion of a "**hierarchical** Bayes" model.

In the childhood leukemia analysis [83], the effect of radiation in a frequentist Poisson regression was very weak and was limited to acute nonlymphocytic leukemia and to one of three 5-year periods. It was reduced to the point of nonsignificance in the GLMM, and the variation in both the $\{u_i\}$ and the $\{v_i\}$ estimates appeared to be significant. Similar doubts attach to the interpretation of these results, however: to the extent that the random effects are independent of natural radiation, they should not affect **inference** for the latter. To the extent that they are associated with it, they may represent **confounders**, but the possibility that radiation is a primary and important effect cannot be excluded.

The models outlined above preserve the discreteness of areal data. A model for continuous data is more difficult mathematically. Typically it is assumed that the logarithm of the RRF is a realization of a spatial Gaussian process [62]. The mean of this process might be taken to be constant if the primary purpose is to model spatial relationship, which would be determined by an autocorrelation function; otherwise it could involve parameters designed to detect locational effects. The methods are comparatively new and untested.

The considerable computational difficulties with Bayesian methods have recently been facilitated by **Markov chain Monte Carlo** (MCMC) methods. These are computationally expensive and care needs to be taken to ensure that the chain is fully sampling the stationary distribution it is intended to estimate. It is also unclear how much the precision is affected by the need to estimate large numbers of parameters, though it should be remembered that the precision of classical frequentist methods incorporates information from the specification of the model and is therefore crucially dependent on its correctness. The book by Gilks [48] gives much technical detail about MCMC and available software, and includes a discussion of geographic applications, particularly issues affecting the choice of priors [70].

Analysis by Location

Areal Modeling

The simplest kind of geographic analysis merely attempts to model a disease rate at a particular location in relation to geographically defined variables associated with that location. Examples of such variables could include geographic variables such as altitude [20], possible measurable risk factors such as background radiation levels, and demographic characteristics of the population, such as socioeconomic status. The latter type of variable imputes to individuals at risk the average of some risk for the whole population in their immediate vicinity, giving rise to an "ecologic" analysis which is not without its dangers [38, 45] (*see* **Ecologic Fallacy**; **Ecologic Study**). Functions of location, such as distance from a specified point, also come into this category of analysis and are dealt with below.

Under the case-independence assumption discussed above, statistical analysis may proceed in an entirely classical way, for example using GLMs. As discussed above, a typical such model would be a Poisson regression with $\ln(e_i)$ as an offset. The residual deviance in this model may then be used to determine whether there is any extra-Poisson variation, which may be evidence of clustering (see below).

Continuous Analogs

Continuous data require an analogous method to analyze the risk function $\mu(x, y)$ or RRF $\theta(x, y)$. For case–control data with equal numbers of cases and controls, this can be achieved by a **conditional logistic regression**. We define

$$\rho(x, y) = \frac{\theta(x, y)}{[1 + \theta(x, y)]},$$

which gives the **conditional probability** that an individual sampled at (x, y) is a case rather than a control [8]. This probability can in principle be modeled logistically using any variable defined by location (x, y) as well as other attributes of the individual concerned [40, 57]. As remarked, above, however, case–control analyses of this sort have not to date been used to anything like the same extent as areal data methods.

Focused Tests of Point Source Hypotheses

A particular kind of locational analysis involves the study of the relationship between disease incidence and the location of some putative source of risk $\mathcal{S}$. Such an analysis imputes risk to geographic location or, equivalently, uses some function of location as a surrogate for risk. This inevitably requires the construction of a one-dimensional function of location. Distance from $\mathcal{S}$ is an obvious choice, though analyses can apply equally well to other measures, incorporating, for example, geographic characteristics such as altitude, bearing to prevailing wind, etc.

Most analyses of data in relation to point sources carried out to date are statistically elementary and consist in examining a single standardized incidence ratio (SIR) for a predefined region $\mathcal{R}$ around $\mathcal{S}$, comparing it with one or more control rates which would typically be national (*see* **Standardization Methods**). Such analyses have the great merit that they are easily understood, but they suffer from the severe disadvantage that they are not at all **powerful** against any sensible **alternative hypothesis**. They are also particularly dependent on the size of the region $\mathcal{R}$, though this is an intrinsic difficulty with other analyses too.

It is almost certainly better to use a method that makes explicit use of distance or other risk surrogate. Suppose that the true RR at a distance d is $\theta(d)$. Then, whether the data are in areal or continuous form, it follows easily from the **Neyman–Pearson lemma** [29] that the **most powerful test** of the null hypothesis of constant risk (H_0) is based on a Linear Risk Score (LRS) computed as the sum over n cases

$$T = \sum_j \ln(\theta(d_j)),$$

where d_j is the distance of the jth case from $\mathcal{S}$ [9]. It follows that the SIR test is only powerful against a hypothesis that supposes a dichotomization of risk inside and outside $\mathcal{R}$. Of course, the usefulness of the general result is limited by the fact that we do not know the true risk function $\theta(\cdot)$, but it provides a benchmark against which other methods may be calibrated. Moreover, it turns out that using a (canonical) risk score of $1/d$ or $1/\text{rank}(d)$ is quite powerful against a wide range of alternatives [12]. **Locally most powerful tests** have also been proposed by Waller [96] and others [61, 93, 97]; these score tests are clearly not "uniformly" most powerful against *all* alternatives and it is unclear how local their power properties might be.

The problem of our ignorance of the true risk function led Stone [14, 92] to propose a test designed to detect a general monotonic decreasing RRF. This test has recently become very popular in the UK. Known as Stone's MLR test, it is a (maximum) **likelihood ratio test** in which the alternative hypothesis is:

$$H_1: \theta_1 \geq \theta_2 \geq \cdots \geq \theta_k (\geq 1),$$

i.e. the areas A_i, ordered by distance from $\mathcal{S}$, have RRs θ_i estimated by maximizing the **likelihood** subject to the restriction that they are monotonic nonincreasing. The final constraint is optional though important, the issue being similar to the choice of a conditional or unconditional test discussed below [9]. This estimation problem is related to that of **isotonic regression** [86]; the computation required is nontrivial but feasible in practice. Stone also proposed the so-called "$P_{\max}$" test based on the first order-restricted RR estimate, $\hat{\theta}_1$. This turns out to be

$$\max_j \sum_{i=1}^{j} Y_i \Big/ \sum_{i=1}^{j} e_i,$$

i.e. the maximal cumulative RR as distance from $\mathcal{S}$ increases. These tests are undoubtedly important and in some situations very effective. The application

of the Neyman–Pearson lemma referred to above, however, implies that Stone's test is never most powerful, and it is not hard to find alternatives for which it has low power compared with canonical LRS tests.

Whichever test statistic is used, it is crucially important to distinguish between the *conditional* and *unconditional* form of the test [9]. The former conditions on the total number of cases observed in $\mathcal{R}$ and is affected only by the spatial information in the data. It would be appropriate whenever the rates used to compute the $\{e_i\}$ may not be reliable in $\mathcal{R}$; if they are trustworthy, however, the conditional analysis ignores potentially valuable information and can even lead to the rejection of H_0 resulting from a *deficit* of cases in the outer parts of the region $\mathcal{R}$. The unconditional form uses the overall disease incidence information as well as the spatial information, but is appropriate only when the $\{e_i\}$ are reliable. In practice, conditional and unconditional tests may produce very different results, especially from small data sets, and it is very important to decide *a priori* which form of test will be used.

Disease Mapping

History and Atlases

The mapping of the incidence of disease and mortality has a long history, well summarized by Howe [51]. Early endeavors were concerned with depicting the epidemic spread of infectious disease, often represented by contours of first occurrence date. More recently, with the diminishing importance of infectious disease in the developed world, the emphasis has changed to representing mortality or incidence in an attempt to infer geographically related explanations of variation in rates. For example, the *Atlas of United States Mortality* [76] contains color maps for each of 18 major causes of death. The colors indicate the 10, 20, 40, 60, 80, and 90 percentiles of the age-standardized death rates in Health Service Areas (HSAs); for example, the top band includes all HSAs whose rates are in the top 10% for the US as a whole. Rates based on small numbers are distinguished by hatching. Cancer atlases, in particular, have been produced for countries all over the world, including the US, continental Europe, China, and the UK [15, 46].

Some maps depict cartograms, in which regions are drawn with areas proportional to their population sizes (*see* **Mapping Disease Patterns**). This may help to overcome the problem of unequal population distribution, but it has the disadvantage that the resulting maps are geographically unfamiliar. We will confine our attention in this section to representational maps in which the geometry is preserved.

Areal Mapping by RR and Other Measures

Most atlases attempt to depict data in discrete form, using administratively defined areas. One of the major concerns in such mapping is the question of what to map. Plotting rates in small areas tends to produce a misleading picture, in that apparently high rates may appear in low-density regions by chance. This problem is exacerbated by the negative correlation usually found between population density and the sizes of administrative areas. Because areas of similar population density are often adjacent, this can induce an apparent spatial pattern where none exists. Some authors [30, 46] have employed a measure of statistical significance instead of or as well as SIRs; with these, however, small values of P which are statistically but not scientifically significant may arise in areas with large populations. It is now generally regarded as preferable to plot **rates** rather than ***P* values**, controlling the influence of sampling variation by using a degree of smoothing [22]. The latter may be *empirical* or *model-based*.

Nonspatial Smoothing

The rationale for **smoothing** is that the **maximum likelihood** estimate of the disease rate in a particular small area is, because of its statistical variability, a poor indicator of the true rate in that area. If, for example, there was very little evidence of geographic variation, we would probably abandon an area-specific estimate and use the global rate for the whole region. Smoothing may be seen as an attempt to compromise between the two positions, using both local and global information.

An attractive method of combining this information is the Bayesian formulation referred to above [26, 55, 68]. In this, it is assumed that the underlying RRs $\{\theta_i\}$ in areas $\{A_i\}$ have a **gamma**

prior distribution, with mean μ and variance σ^2. It follows that the mean of the posterior distribution of θ_i is

$$\tilde{\theta}_i = \frac{y_i + \mu^2/\sigma^2}{e_i + \mu/\sigma^2},$$

a formula that demonstrates how $\tilde{\theta}_i$ varies, according to the value of σ^2, between the area-specific ML estimate Y_i/e_i and the global mean μ, whose ML estimate is $\sum_i Y_i / \sum_i e_i$. Unfortunately the ML estimation of σ^2 requires an iterative method, though a simpler **moment**-estimator is available [26].

Estimating the posterior mean in this way is the basis of the **Empirical Bayes (EB)** method of smoothing. It shrinks the local estimates towards the global mean, but does not take any account of the spatial relationship of one area to another.

Spatial Smoothing Using Bayesian Models

The Bayesian modeling of spatial structure described above can be used to produce estimates of the RR which are smoothed by reference to adjoining areas as well as the overall mean. The original paper by Clayton & Kaldor [26] describes a method that did not take account of the varying number of areas adjacent to a given A_i and consequently lacked internal consistency; developments by Besag et al. [6] have led to a version that meets this objection. The goal of making the rates in adjacent areas more similar to one another than identical rates in well-separated areas may seem very reasonable. It must be remembered, however, that it will inevitably give an appearance of spatial relationship even where none exists; it is therefore essential that maps produced using these methods are clearly indicated as such. For a critique of the effect of smoothing on the interpretation of spatial relationship, see [47].

Empirical Smoothing

Although the Bayesian methodology is very attractive theoretically, it employs fairly sophisticated **algorithms** and is not easily related intuitively to the original data. Empirical methods of smoothing may be better in this respect and numerous methods have been proposed [15]. A particularly attractive method is described by Pukkala and applied to cancer in Finland [81]. At each point of a fine lattice, an estimated risk is computed as a weighted average of the rates in all the A_i whose centers are within a defined distance of the point. The weights take account of the distances of the A_i and also of their population sizes. The result is a map free of the original small area boundaries. A more elaborate method in which numerators and denominators of rates are smoothed separately is described by Kafadar [54].

Smoothing Based on the RRF

The formulation of a model for geographic data in terms of density functions discussed above suggests a simple, probability-based method of depicting risk continuously. All we need to do is to plot an estimate of the risk function or the RRF as the ratio of the densities for cases and controls [8, 56]. This method can also be adapted to employ areal data [10].

Degree of Smoothing and Other Issues

Whatever method of smoothing is employed, it is important to realize that determining the degree or scale of smoothing is intrinsic to any spatial method. For areal data, the sizes of the areas will be involved in this determination. Particular methods may offer choice at other levels of the analysis: for example, the bandwidth in the case of risk function estimation based on density estimates. The question of whether one may reasonably expect the data to determine how much smoothing should be applied is unclear. From a Bayesian standpoint, one should expect to build into the analysis a prior idea of the degree of risk variation. Estimating this degree from the data is akin to using empirical Bayes ideas and departs from the spirit of the "full Bayes" approach. Density estimation methods are certainly associated with data-driven methods of bandwidth determination, such as **cross-validation** [89]. However, these typically use criteria of doubtful relevance to the presentation of meaningful maps, and in practice may not produce satisfactory analyses.

The methods described above all incorporate information about the variation in the population density and this should be regarded as essential because of the variation in precision implied. Methods designed for homoscedastic continuous data (*see* **Scedasticity**), such as geophysical data, can give quite misleading results and should be avoided.

Clustering

Types of Clustering

Clustering may be defined as the tendency of observations to be situated closer to one another than would be expected; the role of chance in this expectation is crucial and much statistical effort is directed towards determining whether an observed cluster could easily be accounted for by chance. In the context of geographic analyses, the issue is that of whether people affected by a disease reside, work or otherwise congregate at places which are closer together than would be expected. From a mathematical point of view, the aggregation could be in any continuum and this gives rise to the notion of clustering in time, space or in the space–time product space. Mathematical considerations also suggest that the nature of the continuum should make rather little difference to the nature of the tests available, specifically that a test working in time, for example, should have an analog in space.

We remark also that it may be useful to distinguish various different types or modes of clustering. Cases may be close together because of a violation of the case-independence assumption; such clustering might provide evidence of a localized genetic effect or of a contagious process. Alternatively, aggregations might be due to variations in underlying risk. Either effect may be highly localized (as with a single familial cluster or a single environmental hazard) or may be widely disseminated. The statistical analysis is not capable of distinguishing these essentially different mechanisms, though different tests will be more sensitive to one rather than another.

We aim in this section to give a brief and relatively abstract overview of the methods available; further details of several of the methods may be found in **Clustering**.

Direct Methods

A set of rates or risk estimates, whether in discrete or continuous form, may be regarded as a risk function θ over geographic space and clustering should appear as some kind of nonuniformity of this function. Many functionals of θ suggest themselves as possible statistics to test for nonuniformity and to some extent the kind of alternative for which they should be powerful is intuitively obvious.

Thus, to detect a single isolated cluster, extremum statistics would be appropriate. These might include, for discrete areal data, occupancy statistics based on a large count in one or more areas [43], and for continuous case–control data, an analog in the form of the maximum height of the RRF estimated by density estimation [87] and tested under a permutation hypothesis. For continuous time, the **scan statistic** [66, 73] counts the maximum number of cases within a fixed length window as it moves through the study period; the distribution theory is analytically difficult [49, 95].

One of the problems with tests based on discrete areal data is that a cluster straddling two or more adjacent areas may be completely missed. A geographic analog of the scan statistic would solve this problem, but is presumably even less tractable analytically than in time, partly because of the dimensionality difference and partly because it is essential to allow for variation in population density. This is of minor importance, however, given that **Monte Carlo** testing provides a way to execute even the most complicated of tests.

The Geographical Analysis Machine of Openshaw [75] is effectively a scan test, though it was derived more by empirical than theoretically well-founded considerations. By varying the size and location of the scanning window, it entails a considerable amount of computation; the criterion of clustering is based on statistical significance. More recently, Anderson & Titterington [2] have used a method based on observed frequencies and applied it to case–control cancer data in South Lancashire. It adjusts the size of the scanning window at each point of $\mathcal{R}$ to ensure that it contains a constant expected number of cases under H_0; this involves extensive numerical integration.

If interest is more in general heterogeneity of risk, it would be more appropriate to use a dispersion statistic of some sort. For areal data this entails determining whether there is "extra-Poisson variation", i.e. whether the Y_i are appreciably different from the e_i. The Potthoff–Whittinghill test has recently become very popular with epidemiologists [80]. As a likelihood score test it is asymptotically locally most powerful, but it is hard to see why it should be better than the deviance statistic $\sum\{e_i - y_i + y_i \ln(y_i/e_i)\}$ when the null hypothesis is not nearly true and the e_i are moderately large.

For continuous data we can compute a suitable measure of overall dispersion, such as the weighted variance of $\hat{\theta}(x, y)$:

$$T_{\text{var}} = \int \int_{\mathcal{R}} \pi(x, y)\{\hat{\theta}(x, y) - 1\}^2 \, \mathrm{d}x \, \mathrm{d}y.$$

This is similar to the "integrated squared difference statistic" used by Anderson & Titterington [2], with rather different weighting which reflects the extent to which population density relates to the local information.

Distance-based Methods

For many problems in statistics, inverse sampling provides an alternative mode of analysis. In the clustering context, we may ask "What distance d from a given point P includes the nearest x cases?" rather than "What is the number x of cases within a distance d of P?". This gives rise to a class of tests based on nearest neighbor distances (NND) [33]. The simplest example, designed for case–control data, counts the number of individuals among the k nearest neighbors of each case that are cases (as opposed to controls). Cuzick & Edwards give analytical results for the null distributions of the different tests and evaluate their power. Further details are given in **Clustering**. The tests have been widely used in epidemiologic investigations. A related method for areal data, due to Besag & Newell [7], considers each case in turn and aggregates the areas around it that are necessary to include the xth nearest case. Tests of this kind have the feature that they adapt the scale on which they seek to detect clustering to varying population density; this may or may not be an advantage according to the clustering mechanism envisaged.

An historically earlier class of tests forms a kind of dual to the NND test in that they count the number of pairs of cases that are close in some sense. The original idea is due to Knox [59] and relates to space–time clustering. Because this is effectively detecting an **interaction** between the time and space variables, it can condition on the marginal distributions in time and space and so circumvent our ignorance about these distributions. To put it another way, the test uses the information on the marginal distributions already present in the data by asking the question "Given the number N_{T} of pairs of cases close in time and the number of pairs N_{S} close in space, what should be the number N_{TS} of pairs close in both time and space?" In fact, the distribution of N_{TS} depends on complex aspects of the configuration of the points in space and time, but Knox conjectured that N_{TS} should have approximately a Poisson distribution with mean

$$\mathrm{E}[N_{\text{TS}}] = N_{\text{T}} N_{\text{S}} \Big/ \binom{n}{2},$$

where the denominator is simply the number of pairs in the set of all n cases considered.

David & Barton [34] demonstrated that, in many situations, this conjecture is well-founded and derived expressions for the variance of N_{TS}. In a number of combinatorially impressive papers, the analysis has been extended (i) to allow for **latent periods** [79]; (ii) to permit a more general measure of closeness than the indicator function originally proposed by Knox [67]; (iii) to the analysis of cross-clustering of events of different types [58]; (iv) to a range of different distance categories [82]; and (v) to a permutation test for space-only clustering using a sample of controls [77, 78]. More recently, Jacquez [52] proposed a version in which closeness was defined in terms of belonging to the set of k nearest neighbors. The power advantages claimed for this may be practically important, though ultimately dependent on the form of the alternative.

Knox's original space–time test remains very popular, but most of these extensions have been used rather little. The space-only test (v) is particularly worthy of more attention, providing as it does an alternative to the Cuzick–Edwards test.

The power of the Knox-type tests is controversial. Barton et al. [5] demonstrated that the original space–time test is remarkably sensitive to the introduction of extra, "offspring" cases if $\mathbf{E}[N_{\text{ST}}]$ is small under the null hypothesis. Chen et al. [23], however, concluded that the test was not very powerful, though this seems to have been due to latent periods that were not allowed for in the analysis. Bradshaw [16] performed extensive simulations to estimate the power under a similar alternative and concluded that using a general continuous, distance-based closeness measure offered rather little improvement over a well-chosen step-function.

Spatial Relationship

The modeling of spatial structure suggests a number of tests based on estimates of spatial similarity. Theoretical development is possible as long as rates are modeled using the **multivariate normal distribution** as an asymptotic approximation to frequency data; the normal family is the only one that has an analytically tractable multivariate form, so that attempts to extend analytical methods to other families inevitably make limited progress.

Under the normal model the distribution of the data is completely specified by the **covariance matrix** which, in the spatial context, will be constructed by reference to postulated autocorrelations. These may be expressed as functions of distance, adjacencies of neighboring areas, lengths of common area boundaries, or other measures. To some extent these will be chosen for mathematical convenience, but the scientific relevance should not be overlooked: Euclidean distance may be important in some contexts and to ignore it could be misleading; in other contexts a measure of degree of adjacency may better reflect the variation of risk with population density.

Different structures in a model for spatial data are reflected in the numerous different tests available. Cliff & Ord [27] give a comprehensive account, though epidemiologic application requires attention to the need to take account of differing population sizes in different areas. This can be done by suitable modifications using weights, or by applying the methods to rates which have been standardized in respect of their sampling variability. Walter [99] reports an empirical investigation of the power of three popular tests of this kind. Munasinghe & Morris [74] use (local) estimates of "regional spatial autocorrelation" to identify particular locations with suspected clustering.

Issues of Interpretation

Much has been written on the interpretation of clusters (*see* **Clustering**; **Geographic Patterns of Disease**). An essential problem is that clusters are mostly reported *post hoc* and it is therefore impossible to assess their statistical significance formally. To say that a cluster is unusual begs the question of the reference set: an extreme that would be unusually high in a single administrative district might well occur quite frequently in a national investigation. From a statistical point of view, it would seem to be desirable to investigate the tendency of a disease to cluster by systematic analysis of a case register (*see* **Disease Registers**). Opinions are divided as to whether this is a good idea [90]. Certainly the investigator should have some idea of what to do if a new cluster is detected and ideally should work to an appropriate protocol.

Choice of Statistical Procedure

It is clear, even from the brief overview in this article, that there is a plethora of different methods for addressing questions raised by geographical data in epidemiology. Even allowing for the multiplicity of these questions and the different types of practical situation arising, it must be the case that some of the methods are worse than others and should be discarded. Very few studies have attempted to assess the comparative characteristics of competing methods and new ones are often introduced without any justification, either theoretical or empirical. We may consider optimality of the relevant procedures at three levels.

Theoretical Considerations

Because the underlying models are complicated it is difficult to obtain theoretical results on power, for example. Nevertheless, there are some guiding principles that should indicate whether a method is likely to work well. As discussed above, the principal of these is that any sensible method should recognize the importance of population density variation. This is important not only in relation to a satisfactory control or comparison group, but also because of the great variation in local information. Thus the homoscedasticity assumptions of geophysical methods such as kriging (*see* **Statistical Map**), for example, make them unsuitable for epidemiologic data unless suitably modified. Epidemiology is concerned essentially with counting people rather than measuring continuous quantities.

Another general principle applies to distance-based methods. In essence it is closeness of individuals, not their distance, that is important, yet it is surprising how many analyses compute mean distances, which are inevitably heavily influenced by the least interesting, large values. It is essential that distance-based

methods employ some inverse function to give most weight to the nearest individuals.

Many analyses are bad simply because they violate these principles and lose power or efficiency as a consequence. A smaller number are actually wrong as a result of serious statistical errors, perhaps concerned with sampling theory. This may occur, for example, when small areas are sampled randomly and assumed to be typical of areas in which index cases reside. This is almost certainly never the case, since the latter are **sampled with probability proportional to size** and administrative areas vary in size very considerably. Moreover, they do so in a way that is highly related to other geographic variables, such as population density. This can lead to artifactual associations which are highly misleading [11].

Statistical Performance

Frequently, it will be unclear which of a number of theoretically acceptable methods is best in the sense of having the best power or efficiency. It will very often be necessary to resort to **simulation**; this is a relatively straightforward way to address the issue, though it is not always easy to summarize the output from simulation experiments.

Some authors compare procedures by looking at the significance levels achieved in application to particular data sets, implicitly supposing that a smaller ***P* value** is evidence of a better test. Unfortunately, this is not so and it really is necessary to examine test performance on a large number of data sets simulated under a known alternative. Power is the usual performance characteristic considered, but it is worth remarking that the expected significance level (ESL) of Dempster & Schatzoff [35] has a number of advantages, including simplicity of simulation and the removal of type I error as a parameter of the experiment. Bithell & Dutton [12] follow Stone [92] in using the ESL in extensive simulations of methods for point source analyses.

There is a particular difficulty with Bayesian methods because of the essentially different philosophical standpoint involved. The Bayesian formulates a model on certain assumptions that are subjective, as with the choice of priors. Subject to these assumptions, the analysis will not only be optimal, but uniquely correct in some sense. Appropriate questions about method performance are therefore concerned more with issues of sensitivity than efficiency: how different would the answers be if the underlying assumptions were different in specified respects? On the whole, rather little of the literature on Bayesian methods in geographic epidemiology seems to address such issues.

Parameter Choice

Frequentist analyses also incur a problem of parameter choice, in that most analyses, even within a class known to work well in theory, will have one or more "tuning parameters". Probably the most important of these are the class of distance scale parameters: for example, how close is "close" in a clustering test? The distance scale parameter is intrinsic to every geographic analysis and appears in the guise of smoothing parameters, covariance functions, and so on. Other quantities which may have to be chosen in advance of the analysis include an analogous time scale parameter, study region size, age, and diagnostic groups. The practical choice of such parameters should ideally be informed by knowledge of the disease process under consideration. In practice this may be difficult and it will be tempting to do several analyses, with the obvious dangers of multiple testing. Allowance for multiple testing by parameter variation is always possible using Monte Carlo methods, but it does, of course, sacrifice power.

Some Studies to Date

There have been rather few systematic comparative studies of different methods of analysis to date. Chen et al. [23] and Walter [98] have carried out studies of the statistical properties of a limited number of tests. More general issues – involving the problems of parameter choice discussed above – are more difficult to study, involving as they do the decisions of the investigators.

In the spirit of comparative analysis, the Childhood Cancer Research Group in Oxford released a major set of childhood leukemia data in standard format to interested investigators, who reported the results of their differing analyses in a single volume [42]. There was no element of competition in this exercise, however, and, since the data were real, it was not possible to say which investigators had the "right" answers: simulated data sets are required to answer such questions.

The **International Agency for Research Against Cancer** have published the results of a "blind trial" in which investigators were presented with simulated data sets incorporating clustering known only to the organizers [1]. The results make it clear that the investigators' strategies and choice of test parameters are at least as important as the statistical properties of the procedures. This reflects real life, but makes it difficult to extrapolate conclusions to the way methods would perform in the hands of other investigators. Interestingly, no investigators in this experiment chose to use Bayesian methods, even though there was quite a lot of prior information about the distributions of the parameters governing the construction of the data sets.

Conclusion

We conclude from this discussion, as have others [100], that there is an urgent and widespread need, not for more elaborate statistical methods, but for clear principles by which existing methods should be judged, together with carefully designed simulation experiments where appropriate.

References

[1] Alexander, F.E. & Boyle, P. eds. (1996). *Methods for Investigating Localized Clustering of Disease*. IARC, Lyon.

[2] Anderson, N.H. & Titterington, D.M. (1995). Some methods for investigating spatial clustering, with epidemiological applications, *Journal of the Royal Statistical Society, Series A* **160**, 87–105.

[3] Antó, J.M. & Sunyer, J. (1992). Soya bean as a risk factor for epidemic asthma, in *Geographical Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. OUP for World Health Organization, Oxford, pp. 323–341.

[4] Bailey, N.T.J. (1978). Spatial models in the epidemiology of infectious diseases, in *Biological Growth and Spread. Lecture Notes in Biomathematics, No. 38*, W. Jäger, H. Rost & P. Tautu, eds. Springer-Verlag, Heidelberg.

[5] Barton, D.E., David, F.N., Fix, E., Merrington, M. & Mustacchi, P. (1966). Tests for space–time interaction and a power function, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. University of California Press, Berkeley, pp. 217–227.

[6] Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration with applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

[7] Besag, J. & Newell, J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A* **154**, 143–155.

[8] Bithell, J.F. (1990). An application of density estimation to geographical epidemiology, *Statistics in Medicine* **9**, 691–701.

[9] Bithell, J.F. (1995). The choice of test for detecting raised disease risk near a point source, *Statistics in Medicine* **14**, 2309–2322.

[10] Bithell, J.F. (1999). Disease mapping using the relative risk function estimated from areal data, in *Disease mapping and risk assessment for public health decision making*, A.B. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, R. Bertollini, eds. Wiley, Chichester, pp. 247–255.

[11] Bithell, J.F. & Draper, G.J. (1995). Apparent association between benzene and childhood leukaemia: methodological doubts concerning a report by Knox, *Journal of Epidemiology and Community Health* **49**, 437–439.

[12] Bithell, J.F. & Dutton, S.J. (1996). Optimal frequentist procedures for detecting raised disease risk near point sources, *American Statistical Association Proceedings of the 1995 Joint Meetings of the Section on Epidemiology*. American Statistical Association, Alexandria, pp. 1–10.

[13] Bithell, J.F., Dutton, S.J., Neary, N.M. & Vincent, T.J. (1995). Use of regression methods for control of socioeconomic confounding, *Journal of Epidemiology and Community Health* **49**,(Suppl. 2), S15–S19.

[14] Bithell, J.F. & Stone, R.A. (1989). On statistical methods for analyzing the geographical distribution of cancer cases near nuclear installations, *Journal of Epidemiology and Community Health* **43**, 79–85.

[15] Boyle, P., Muir, C.S. & Grundmann, E. eds. (1989). *Cancer Mapping*. Springer-Verlag, Berlin.

[16] Bradshaw D. (1982). A comparison of tests for space–time clustering of disease cases. D.Phil. Thesis, University of Oxford.

[17] Breslow, N.E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics* **33**(1), 38–44.

[18] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421), 9–25.

[19] Breslow, N.E. & Powers, W. (1978). Are there two logistic regressions for retrospective studies?, *Biometrics* **34**(1), 100–105.

[20] Burkitt, D. & Wright, D. (1966). Geographical and tribal distribution of the African lymphoma in Uganda, *British Medical Journal* **i**, 569–573.

[21] Burkitt, D.P. & Wright, D.H. (1970). *Burkitt's Lymphoma*. Churchill Livingstone, Edinburgh.

[22] Cartwright, R.A., Alexander, F.E., McKinney, P.A. & Ricketts, T.J. (1990). *Leukaemia and Lymphoma: An Atlas of Distribution within Areas of England and Wales 1984–1988*. Leukaemia Research Fund, Leeds.

[23] Chen, R., Mantel, N. & Klingberg, M.A. (1984). A study of three techniques for time–space clustering in Hodgkin's disease, *Statistics in Medicine* **3**, 173–184.

[24] Clayton, D.G. (1996). Generalized linear mixed models, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 275–301.

[25] Clayton, D. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, in *Geographical Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. OUP for World Health Organization, Oxford, pp. 205–220.

[26] Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671–682.

[27] Cliff, A.D. & Ord, J.K. (1981). *Spatial Processes: Models and Applications*. Pion, London.

[28] Cook, D.G. & Pocock, S.J. (1983). Multiple regression in geographical mortality studies, with allowance for spatially correlated errors, *Biometrics* **39**, 361–372.

[29] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[30] Craft, A.W., Openshaw, S. & Birch, J.M. (1985). Childhood cancer in the Northern Region, 1968–82: incidence in small geographical areas, *Journal of Epidemiology and Community Health* **39**, 53–57.

[31] Cressie, N.A.C. (1993). *Statistics for Spatial Data*, Revised Ed. Wiley, New York.

[32] Cressie, N.A.C. (1996). Bayesian and constrained inference for extremes in epidemiology, *American Statistical Association Proceedings of the 1995 Joint Meetings of the Section on Epidemiology*. American Statistical Association, Alexandria, pp. 11–17.

[33] Cuzick, J. & Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with discussion), *Journal of the Royal Statistical Society, Series B* **52**, 73–104.

[34] David, F.N. & Barton, D.E. (1966). Two space–time interaction tests for epidemicity, *British Journal of Preventive and Social Medicine* **20**, 44–48.

[35] Dempster, A.P. & Schatzoff, M. (1965). Expected significance levels as a sensitivity index for test statistics, *Journal of the American Statistical Association* **60**, 420–436.

[36] Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

[37] Diggle, P.J. & Chetwynd A.G. (1991). Second order analysis of spatial clustering for inhomogeneous populations, *Biometrics* **47**, 1155–1163.

[38] Diggle P. & Elliott P. (1995). Disease risk near point sources: statistical issues for analyses using individual or spatially aggregated data, *Journal of Epidemiology and Community Health* **49**, S20–S27.

[39] Diggle, P.J. & Morris, S. (1996). Second-order analysis of spatial clustering, in *Methods for Investigating Localized Clustering of Disease*, F.E. Alexander & P. Boyle, eds. IARC, Lyon, pp. 207–214.

[40] Diggle, P.J. & Rowlingson, B.S. (1994). A conditional approach to point process modelling of elevated risk, *Journal of the Royal Statistical Society, Series A* **157**, 433–440.

[41] Donnelly, C.A. (1995). The spatial analysis of covariates in a study of environmental epidemiology, *Statistics in Medicine* **14**, 2393–2409.

[42] Draper, G., ed. (1991). *The Geographical Epidemiology of Childhood Leukaemia and non-Hodgkin Lymphomas in Great Britain, 1966–83. Studies on Medical and Population Subjects, No. 53*. HMSO, London.

[43] Ederer, F., Myers, M.H. & Mantel, N. (1964). A statistical problem in space and time: do leukemia cases come in clusters?, *Biometrics* **20**, 626–638.

[44] Elliott P., Wakefield, J.C., Best, N.G. & Briggs, D.J. eds. (2000). *Spatial epidemiology: methods and applications*, Oxford University Press, Oxford, pp. 87–103.

[45] English, D. (1992). Geographical epidemiology and ecological studies, in *Geographical Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. OUP for World Health Organization, Oxford, pp. 3–13.

[46] Gardner, M.J., Winter, P.G., Taylor, C.P. & Acheson, E.D. (1983). *Atlas of Cancer Mortality in England and Wales, 1968–1978*. Wiley, Chichester.

[47] Gelman, A. & Price, P.N. (1999). All maps of parameter estimates are misleading, *Statistics in Medicine* **18**, 3221–3234.

[48] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[49] Glaz, J. (1993). Approximations for the tail probabilities and moments of the scan statistic, *Statistics in Medicine* **12**, 1845–1852.

[50] Heisterkamp, S.H., Doornbos, G. & Gankema, M. (1993). Disease mapping using empirical Bayes and Bayes methods on mortality statistics in the Netherlands, *Statistics in Medicine* **12**, 1895–1914.

[51] Howe, G.M. (1989). Historical evolution of disease mapping in general and specifically of cancer mapping, in *Cancer Mapping*, P. Boyle, C.S. Muir & E. Grundmann, eds. Springer-Verlag, Berlin, pp. 1–21.

[52] Jacquez, G.M. (1996). A k nearest neighbour test for space–time interaction, *Statistics in Medicine* **15**, 1935–1949.

[53] Johnson, N.L., Kotz S. & Kemp, A.W. (1992). *Univariate Discrete Distributions*, 2nd Ed. Wiley, New York.

[54] Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease, *Statistics in Medicine* **15**, 2539–2560.

[55] Kaldor, J. & Clayton, D. (1989). Role of advanced statistical techniques in cancer mapping, in *Cancer Mapping*, P. Boyle, C.S. Muir & E. Grundmann, eds. Springer-Verlag, Berlin, pp. 87–98.

[56] Kelsall, J.E. & Diggle, P.J. (1995). Non-parametric estimation of spatial variation in relative risk, *Statistics in Medicine* **14**, 2335–2342.

[57] Kelsall, J.E. & Diggle, P.J. (1998). Spatial variation in risk: a Nonparametric binary regression approach, *Applied Statistics* **47**, 559–573.

[58] Klauber, M.R. (1971). Two-sample randomization tests for space–time clustering, *Biometrics* **27**, 129–142.

[59] Knox, E.G. (1964). The detection of space–time interactions, *Applied Statistics* **13**, 25–29.

[60] Knox, E.G. & Lancashire, P.J. (1982). Detection of minimal epidemics, *Statistics in Medicine* **1**, 183–189.

[61] Lawson, A.B. (1993). On the analysis of mortality events associated with a prespecified fixed-point, *Journal of the Royal Statistical Society, Series A* **156**, 363–377.

[62] Lawson, A.B. (1994). Using spatial Gaussian priors to model heterogeneity in environmental epidemiology, *Statistician* **43**, 69–76.

[63] Lawson, A.B. (1995). MCMC methods for putative pollution source problems in environmental epidemiology, *Statistics in Medicine* **14**, 2473–2485.

[64] Lawson, A.B., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F. & Bertollini, R. (eds). (1999). *Disease Mapping and Risk Assessment for Public Health Decision Making*. Wiley, Chichester.

[65] Lazar, P. (1981). Geographical correlations between disease and environmental exposure, in *Perspectives in Medical Statistics*, J.F. Bithell & R. Coppi, eds. Academic Press, London.

[66] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[67] Mantel N. (1967). The detection of disease clustering and a generalized regression approach, *Cancer Research* **27**, 209–220.

[68] Manton, K.G., Woodbury, M.A., Stallard, E., Riggan, W.B., Creason, J.P. & Pellom, A.C. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates, *Journal of the American Statistical Association* **84**, 637–650.

[69] Martin D. & Bracken I. (1991). Techniques for modelling population-related raster databases, *Environment & Planning A* **23**, 1069–1075.

[70] Mollié, A. (1996). Bayesian mapping of disease, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 359–379.

[71] Mollié, A. & Richardson, S. (1991). Empirical Bayes estimates of cancer mortality rates using spatial models, *Statistics in Medicine* **10**, 95–112.

[72] Mollison D. (1977). Spatial contact models for ecological and epidemic spread, *Journal of the Royal Statistical Society, Series B* **39**, 283–326.

[73] Naus J.I. (1965). The distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association* **60**, 532–538.

[74] Munasinghe, R.L. & Morris, R.D. (1996). Localization of disease clusters using regional measures of spatial autocorrelation, *Statistics in Medicine* **15**, 893–905.

[75] Openshaw, S. & Craft, A. (1991). Using Geographical Analysis Machines to search for evidence of clusters and clustering in childhood leukaemia and non-Hodgkin lymphomas in Britain, in *The Geographical Epidemiology of Childhood Leukaemia and non-Hodgkin Lymphomas in Great Britain, 1966–83. Studies on Medical and Population Subjects, No. 53*, G. Draper, ed. HMSO, London, pp. 109–122.

[76] Pickle, L.W., Mungiole, M., Jones, G.K. & White, A.A. (1996). *Atlas of United States Mortality*. National Center for Health Statistics, Hyattsville.

[77] Pike, M.C. & Smith, P.G. (1974). A note on a "close pairs" test for space clustering, *British Journal of Preventive and Social Medicine* **28**, 63–64.

[78] Pike, M.C. & Smith, P.G. (1974). A case–control approach to examine disease for evidence of contagion, including disease with long latent periods, *Biometrics* **30**, 263–279.

[79] Pike, M.C. & Smith, P.G. (1968). Disease clustering: a generalization of Knox's approach to the detection of space–time interactions, *Biometrics* **24**, 541–556.

[80] Potthoff, R.F. & Whittinghill, M. (1966). Testing for homogeneity. II. The Poisson distribution, *Biometrika* **53**, 183–190.

[81] Pukkala, E. (1989). Cancer maps of Finland: an example of small-area based mapping, in *Cancer Mapping*, P. Boyle, C.S. Muir & E. Grundmann, eds. Springer-Verlag, Berlin, pp. 208–215.

[82] Raubertas R.F. (1988). Spatial and temporal analysis of disease occurrence for detection of clustering, *Biometrics* **44**, 1121–1129.

[83] Richardson, S., Montfort, C., Green, M., Draper, G. & Muirhead, C. (1995). Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain, *Statistics in Medicine* **14**, 2487–2501.

[84] Ripley, B.D. (1977). Modelling spatial patterns (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 172–212.

[85] Ripley, B.D. (1981). *Spatial Statistics*. Wiley, New York.

[86] Robertson, T., Wright F.T. & Dykstra R.L. (1988). *Order Restricted Statistical Inference*. Wiley, London.

[87] Rossiter, J.E. (1991). *Epidemiological applications of density estimation. D.Phil. Thesis*, University of Oxford.

[88] Scott, D.W. (1992). *Multivariate Density Estimation*. Wiley, London.

[89] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[90] Smith, D. & Neutra, R. (1993). Approaches to disease clustering investigations in a State Health Department, *Statistics in Medicine* **12**, 1757–1762.

[91] Snow, J. (1855). *On the Mode of Communication of Cholera*, 2nd Ed. Churchill Livingstone, London.

[92] Stone, R.A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test, *Statistics in Medicine* **7**, 649–660.

[93] Tango, T. (1995). A class of tests for detecting "General" and "Focused" clustering of rare diseases, *Statistics in Medicine* **14**, 2323–2334.

[94] Tobler, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions, *Journal of the American Statistical Association* **74**, 519–535.

[95] Wallenstein, S., Naus, J. & Glaz, J. (1993). Power of the scan statistic for detection of clustering, *Statistics in Medicine* **12**, 1829–1844.

[96] Waller, L.A. (1996). Statistical power and design of focused clustering studies, *Statistics in Medicine* **15**, 765–782.

[97] Waller, L.A. & Lawson, A.B. (1995). The power of focused tests to detect disease clustering, *Statistics in Medicine* **14**, 2290–2308.

[98] Walter, S.D. (1992). The analysis of regional patterns in health data. II. The power to detect environmental effects, *American Journal of Epidemiology* **136**, 742–758.

[99] Walter, S.D. (1993). Assessing spatial patterns in disease rates, *Statistics in Medicine* **12**, 1885–1894.

[100] Wartenburg, D. & Greenberg, M. (1993). Solving the cluster puzzle: clues to follow and pitfalls to avoid, *Statistics in Medicine* **12**, 1763–1772.

JOHN F. BITHELL

Geographic Patterns of Disease

The study of geographic patterns of disease is part of the classic triad in **descriptive epidemiology** of "time, person, place". Here, place is used as a surrogate for the mix of lifestyle, environmental, and possibly genetic factors that may underlie variations in rates of disease across populations. The purpose is both to *describe* such variations and to identify possible *causes* that could explain them.

Of course, apparent geographic variations in disease rates may be artifactual rather than real. Problems may occur either with the enumeration of cases (numerator) or with the population at risk (**denominator**), or both. Thus spurious geographic variations in disease could reflect differences between populations in case definition, completeness of ascertainment, diagnostic accuracy, and coding or (for mortality) survival rates. Enumeration of the population (e.g. at **census**) may be incomplete, or recent migration may distort population estimates. Great care is therefore required in interpretation. Despite these difficulties, publications such as *Cancer Incidence in Five Continents* [27], and international mortality statistics (*see* **Mortality, International Comparisons**) compiled by the **World Health Organization**, have provided an invaluable starting point for epidemiologic enquiry.

The analysis of geographic patterns of disease depends crucially on *scale*. Whereas broad-scale patterns may be apparent at an international level, for example, differences between developed and developing countries in the incidence of infectious diseases such as malaria, and in cardiovascular diseases [44], other patterns may only be apparent at a *local* level. These will include, for example, clusters of disease (*see* **Clustering**) and possible variations in disease risk near putative point sources of environmental pollution.

In this article we briefly discuss disease variations both at the broader and at the local (small-area) scale. We review issues involved in disease mapping (the usual means of presenting descriptive geographic data on disease occurrence) and discuss some of the problems associated with geographic **correlational studies (ecologic studies)**. Here the aim is to explore geographic variation in disease in terms of underlying spatially varying "risk factors". The emphasis is on **small-area** applications, where a number of recent advances in methodology have been made.

International Variations in Disease

International differences in disease occurrence may give important clues as to etiology, which may then be further studied in individual-level studies (e.g. cohort–control or case–control). Thus, in the Seven Countries Study, Keys [24] described large differences in population saturated fat intakes, which were predictive of population differences in the occurrence of coronary heart disease. The INTERSALT Study found cross-population differences in average blood pressure levels, and difference in blood pressure with age, that were associated positively with average levels of salt intake (measured by urinary sodium excretion); a similar positive relationship was also found at individual level [17]. Other examples include the incidence of malignant melanoma and multiple sclerosis, both of which are strongly related to latitude. While this relationship is inverse for melanoma (i.e. a tendency for higher rates near the equator, reflecting greater exposure to sunlight [19],), it is positive for multiple sclerosis (i.e. low incidence in countries near the equator [26]).

Migrant Studies

Migrant studies represent a special case of geographic study. Here, the disease experience of individuals or groups of people is examined as they move from one location or country to another. This affords a unique opportunity to examine the extent to which environmental or genetic influences might determine geographic variations in disease risk. Whereas genetic factors are important in determining which *individuals* become sick, at the population level, overwhelmingly, environmental and lifestyle factors predominate [33]. Thus, in the case of multiple sclerosis, migrants moving from a high risk to a low risk area retain their higher risk if migrating after the age of around 15 years, but attain the risk of the host country if migrating at younger ages [26]. These findings are compatible with an infectious etiology of multiple sclerosis, with infection acquired in childhood. Another example is the low levels of blood pressure, with little or no rise with age, found among

remote and isolated population groups around the world [17, 31]. Blood pressures are found to increase rapidly with migration to an urban environment [31], again indicating the overwhelming importance of environmental factors in determining the unfavorable blood pressure pattern among populations.

Local Variations in Disease

Variations in disease incidence or mortality at national [23] or subnational [20] level have been described, usually in the form of a disease atlas (see the section on *Disease Mapping*, below). Here we briefly address the occurrence of disease at the local (**small-area**) scale. Although in this context no satisfactory definition of the term "small area" exists, Cuzick & Elliott [10] suggest a working definition as follows:

> As a rough guide, any region containing fewer than about 20 cases of disease can be considered a small area ... Many cancers have annual incidence rates of around 5 per 100 000, so for a collective period of 5 years a small area constitutes a population of around 100 000 or fewer. In some instances, such as a cluster of disease in a remote area or small village, it could be much less, but usually populations of at least 10 000 are needed to form an aggregation of minimal size.

Of course, populations could be much smaller if the disease experience over many such areas is of primary interest – for example, in small-area disease mapping (see next section).

Disease Clusters and Clustering

A problem commonly facing public health authorities is how to deal with reports of apparent disease excess in their locality (i.e. disease "clusters"; *see* **Clustering**). These reports may subsequently be linked to a putative pollution source. This complicates interpretation since, for post hoc enquires of this type, formal statistical testing is no longer valid. Although there is little potential for isolated cluster investigations to yield new information on the cause of disease, nonetheless the public health authorities often feel compelled to respond. A careful review of cases, and selection of an appropriate **denominator** and time frame, may result in risk estimates (observed/expected ratios) that are close to 1. This is despite the potential for **bias** towards elevated risk ratios (*see* **Relative Risk**) – areas at apparently "low" risk do not come to the attention of the authorities! In some instances, replication of the study in other similarly polluted areas (if such can be found), or in a different time period, may be the only feasible way forward. It can also be helpful to place an alleged "cluster" in a wider context by carrying out small-area disease mapping across a larger region (see [43] for a recent example).

An alternative approach to the study of a single disease cluster is to examine more generally for evidence of clustering. Such evidence for Hodgkin's disease has been cited in support of ideas of an infectious etiology [1], although other explanations, including artifacts related to diagnostic coding, population mobility, or variations in birth rates, are also possible [10, 18].

Small-area Studies Near Sources of Environmental Pollution

Recently, high-resolution geographically referenced routine health data (*see* **Vital Statistics, Overview**) have become available in certain countries. Together with advances in computing and in statistical methodology, this has led to the development of largely automated systems to examine the distribution of disease near point sources of environmental pollution. In the UK, the Small Area Health Statistics Unit (SAHSU) has been established specifically to: respond rapidly to reports of disease excess ("clusters") near sources of environmental pollution; carry out studies of health statistics more generally around sources of pollution; carry out descriptive geographic studies at small-area level; and develop the methodology. Recent studies include an investigation of cancer incidence and mortality near a pesticide factory following media reports of excess cancers in the vicinity [43], and a national study of cancer incidence near radio and television transmitters [10] following reports of a leukemia excess near one of the transmitters [13] (*see* **Leukemia Clusters**).

A major problem in the interpretation of such studies is the issue of socioeconomic **confounding**. Measures of social deprivation (calculated from the census statistics) have been shown to be powerful predictors of the occurrence of disease [5], including stomach and lung cancer (though not leukemia). Deprived areas do not occur randomly throughout

a region, but tend to coincide with industrial sites and correlate with higher smoking rates. Failure to account for social deprivation could thus seriously bias investigation of other lifestyle or environmental risk factors and ill-health. This is illustrated by results of a national study of cancer risk near municipal solid waste incinerators in Great Britain [16]. **Excess risk** was found for a number of cancer sites, including stomach and lung, that persisted after adjustment for deprivation at the small-area scale. However, in the areas with available data, a similar excess was found also for the period before the incinerators were operational. This indicated the presence of **residual confounding** that had not been fully accounted for in the statistical analysis [16].

Disease Mapping

Maps have long been used to describe geographic patterns of disease (*see* **Mapping Disease Patterns**). For example, Stocks, in a series of atlases published in the 1930s, described the geographic variation in cancer mortality across counties in England and Wales (reproduced in [38]). A survey in 1991 [42] identified 49 international, national, and regional disease atlases; more recent examples include those by Swerdlow & dos Santos Silva [38] and Bernardinelli et al. [4]. Such maps typically show standardized mortality or incidence ratios (*see* **Standardization Methods**) for geographic areas such as countries, counties, or districts. The rate in area i is estimated by O_i/E_i, where O_i is the observed number of deaths or incident cases of disease in the area (assumed to follow an independent **Poisson distribution**) and E_i is the expected number of cases (calculated by applying age- and sex-specific death or disease rates to the census population counts for the area).

Maps convey instant visual information on the spatial distribution of disease and can identify subtle patterns which may be missed in tabular presentations. Their purpose is usually to display variations in ill-health (for example, related to the underlying sociodemography), formulate etiologic hypotheses, aid **surveillance** to detect areas of high disease incidence, and help place specific disease clusters and point source studies in proper context.

While disease maps have both visual and intuitive appeal, considerable caution is required to avoid overinterpretation. Apparent geographic variation in rates may simply reflect between-area differences in the quality of reporting, diagnosis, and classification of disease, or confounding due to **ethnic** and socioeconomic factors. Furthermore, disease maps implicitly assume that risk is homogeneous within areas. This is unlikely for the large areas used in many national and international atlases, and may result in misleading inference about individual-level risk.

There is currently considerable scientific interest in exploring more *local* geographic variations in disease. For example, in the UK, small-area mapping is often carried out at the level of electoral ward (average 5000 people) and census enumeration district (400 people).

Disease mapping at the small-area level raises a number of statistical issues. For relatively rare events such as death and cancer incidence, the observed numbers of cases tend to be small in areas with low population, and typically exhibit extra-Poisson sampling variation. This may be assessed formally using the Pothoff–Whittinghill test [30] (*see* **Clustering**). The sparseness of population data results in unreliable estimates of the area-specific standardized rate ratios, which may create the impression of spurious geographic variation when displayed on a map. These considerations have led to the use of statistical smoothing techniques, which pool information across areas. **Empirical Bayes** [8, 15] and hierarchical Bayes [7, 39] estimates of area-specific relative risk (*see* **Hierarchical Models in Health Service Research**) represent a compromise between the area-specific standardized rate ratios (*see* **Standardization Methods**) and the overall **mean** for the whole map.

Small-area disease data often exhibit spatial correlation due to the influence of unmeasured or unknown risk factors which themselves vary smoothly in space. Various hypothesis tests are available to assess such spatial **autocorrelation** – for example, the rank-adjacency D-statistic [23] and Smans' test [35].

Figure 1 shows a map of "unsmoothed" (standardized incidence ratio, adjusted for age, sex, and deprivation) and smoothed (empirical Bayes) estimates of brain cancer incidence for 1974–1986 across electoral wards in the West Midlands region of England [14]. As can be seen, much of the random variability is removed by smoothing, especially the apparent high rates found in the large, sparsely populated rural areas. Overall, there is only weak evidence of heterogeneity across the map

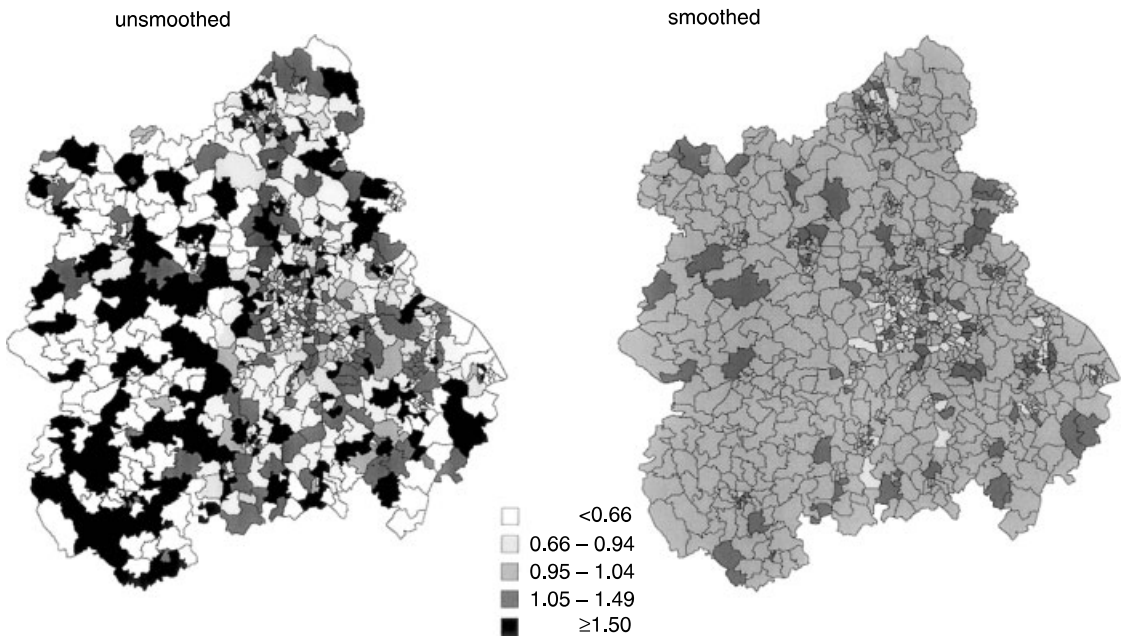

Figure 1 Age-, sex-, and deprivation-adjusted relative risks of brain and central nervous system tumors for electoral wards in West Midlands region, England, age 15–64 years, 1974–1986. Unsmoothed risks (left) and after map smoothing (right) using empirical Bayes method. Reproduced from Eaton et al. [14] by permission of *British Journal of Cancer*

(Potthoff–Whittinghill test; $p = 0.04$), and no evidence of spatial autocorrelation [14].

Bayesian prior distributions for the area-specific relative risks which allow smoothing towards a local mean, rather than the overall map mean, are also used to model spatial interdependence in small-area studies [7, 8, 11, 21, 25, 39]. Implementation of Bayesian hierachical–spatial models has been made feasible by recent computational [36] and *software* developments, namely BUGS [37] – involving **Markov chain Monte Carlo** simulation algorithms: this approach represents the current state of the art in small-area mapping of disease.

Technical Issues Concerning Presentation of Geographic Disease Data

Maps provide a succinct summary of geographic patterns in disease. However, visual perception may be influenced by various features of the map, such as the plotting symbols used (e.g. solid shading vs. hatching, color vs. gray scale) and the grouping of data into categories (e.g. percentiles of the distribution of risk, and numerically equidistant cutpoints) [35]. An empirical study [41] found that the manner of data display may have at least as much effect on observer perception of spatial variation as actual differences in the data. Recently, **nonparametric** mixture distributions (*see* **Contagious Distributions**) have been used to model the underlying relative risk of disease in small geographic areas [34]. This approach facilitates more objective mapping of disease patterns, since areas are categorized according to statistically driven estimation of the mixture components.

The summary statistic used for presentation may also influence visual interpretation of disease maps. Common choices include standardized rate ratios, smoothed relative risks, or ***P* values**. The former tend to yield erratic maps which are visually dominated by extreme estimates of low precision in sparsely populated areas; the latter are criticized for confusing statistical significance with biological importance (*see* **Clinical Significance Versus Statistical Significance**) and tend to overemphasize areas of high population in which even small deviations from the expected disease rate may achieve statistical significance. Significance testing of standardized rate ratios also suffers from the multiple decision problem (*see*

Multiple Comparisons), as each ratio is considered independently of the others on the map.

In our view, maps showing Bayesian **shrinkage estimates** of relative risk represent the best compromise, although it is important to realize that these estimates are not judgment-free. For example, they depend on the functions used to describe the distribution of relative risks across the map, and to define the local neighborhood over which spatial interdependence between the small areas is assumed. However, smoothing ensures that precision of the area-specific estimates is approximately comparable across the map, and Bayesian credible intervals derived from hierarchical models are not subject to the constraints of multiple significance testing. Mapping of posterior functions of Bayesian risk estimates is also possible. For example, a map showing the posterior probability that the relative risk in each area ranks above the **median** [2, 22] conveys information about the size *and* uncertainty associated with each area-specific estimate. Further advances in the application of Bayesian methods to disease mapping, and appropriate display methods, including measures of uncertainty, are to be expected.

Geographic Correlation Studies

Geographic correlation studies are a valuable means of formulating and testing etiologic hypotheses: disease patterns are compared with the geographic distribution of environmental and lifestyle exposures. They are particularly useful when individual-level measurements of exposure are either difficult or impossible to obtain for use in epidemiologic study (for example, air pollution) or are measured imprecisely (for example, diet, and sunlight exposure). (See [19] for further discussion and [32] for a review of the statistical methods.)

Examples of broad-scale **ecologic studies** are given in the section on *International Variations in Disease*. In some cases – for example, sunlight and melanoma, salt and blood pressure – the ecologic relationships have also been demonstrated at individual level. However, the potential for *bias* in such ecologic studies [19, 32] should be recognized. Exposure *within* areas is often heterogeneous; thus the ecologic (average group-level) association between exposure and disease may not equate to the relationship in individuals. To assume otherwise is to commit the **ecologic fallacy** [28]. Small-area studies may be less prone than broad-scale geographic studies to ecologic bias since the group data are closer to the level of the individual. Nonetheless, positive findings arising from ecologic analyses usually require replication in other data sets and, where possible, at individual level.

As already noted, a major problem in small-area disease studies is the potential for confounding by socioeconomic variables. Adjustment may be made by including, say, a deprivation score such as the Carstairs [6] index (based on small-area census statistics) as a **covariate** in the ecologic regression analysis. Alternatively, indirect standardization of the expected small-area disease counts can be done by stratifying on the socioeconomic status of the areas as well as on age and sex (*see* **Stratification**). Modeling of spatial autocorrelation between small areas in an ecologic regression study also provides some control for the effect of confounding due to location [9], but further development of these methods is required, and in particular their application to "real" data sets.

Interest has focused on ecologic designs which combine data on the general population with individual-level survey data to improve estimation of group exposure [29, 40]. Methods to adjust for random measurement error in exposure are also receiving attention [3]. Such techniques should enhance the ability of ecologic analyses to estimate the *size* of exposure–disease relationships, not merely to identify the possible presence of such associations.

Summary and Conclusions

The study of geographic patterns of disease plays a central role in descriptive epidemiology, and has led to some notable etiologic insights. However, geographic studies are associated with major problems of data quality, bias, confounding, and presentation which can seriously complicate their interpretation. The methodologic challenge is clear: to produce objective, statistically valid analyses of geographic variations in ill-health and its determinants, with particular emphasis on developments to combine the best features of individual-level and ecologic studies. Recent advances, particularly in methods for small-area studies, have begun to address these issues. As such techniques become routinely available, they should enhance our ability to quantify the effects

of environmental pollution (*see* **Environmental Epidemiology**) and lifestyle characteristics on human health.

References

[1] Alexander, F.E. (1990). Clustering and Hodgkin's disease, *British Journal of Cancer* **62**, 708–711.

[2] Bernardinelli, L., Clayton, D.G. & Montomoli, C. (1995). Mapping disease risk: how important are priors?, *Statistics in Medicine* **14**, 2411–2431.

[3] Bernardinelli, L., Pascutto, C., Best, N.G. & Gilks W.R. (1997). Disease mapping with errors in covariates, *Statistics in Medicine* **16**, 741–752.

[4] Bernardinelli, L., Maida, A., Marinoni, A., Clayton, D.G., Romano, G., Montomoli, C., Fadda, D., Solinas, M.G., Castiglia, P., Cocco, P.L., Ghislandi, M., Berzuini, C., Pascutto, C., Nerini, M., Styles, B., Capocaccia, R., Lispi L. & Mallardo, E. (1994). *Atlas of Cancer Mortality in Sardinia*, 1983–87. FATMA-CNR.

[5] Carstairs, V. (2000). Socio-economic factors at area level and their relationship with health, Ch. 4 in *Spatial Epidemiology: Methods and Application*, P. Elliott, J.C. Wakefield, N.G. Best,& D.J. Briggs, eds. Oxford University Press, Oxford, pp. 51–67.

[6] Carstairs, V. & Morris, R. (1991). *Deprivation and Health in Scotland*. Aberdeen University Press, Aberdeen.

[7] Clayton, D.G. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. Oxford University Press, Oxford, pp. 205–220.

[8] Clayton, D.G. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671–682.

[9] Clayton, D.G., Bernardinelli, L. & Montomoli, C. (1993). Spatial correlation in ecological analysis, *International Journal of Epidemiology* **22**, 1193–1202.

[10] Cuzick J. & Elliott, P. (1992). Small area studies: purpose and methods, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott,J. Cuzick,D. English & R. Stern, eds. Oxford University Press, Oxford, pp. 14–21.

[11] Denison, D.G.T. & Holmes, C.C. (2001). Bayesian partitioning for estimating disease risk, *Biometrics* **57**, 143–149.

[12] Dolk, H., Elliott, P., Shaddick, G., Walls, P. & Thakrar, B. (1997). Cancer incidence near radio and television transmitters in Great Britain. II. All high power transmitters, *American Journal of Epidemiology* **145**, 10–17.

[13] Dolk, H., Shaddick, G., Walls, P., Grundy, C., Thakrar, B., Kleinschmidt, I. & Elliott, P. (1997). Cancer incidence near radio and television transmitters in Great Britain. I. Sutton Coldfield transmitter, *American Journal of Epidemiology* **145**, 1–9.

[14] Eaton, N., Shaddick, G., Dolk, H. & Elliott, P. (1997). Small-area study of the incidence of neoplasms of the brain and central nervous system among adults in the West Midlands Region, 1974–86, *British Journal of Cancer* **75**, 1080–1083.

[15] Efron, B. & Morris, C. (1975). Data analysis using Stein's estimation and its generalisation, *Journal of the American Statistical Association* **70**, 311–319.

[16] Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J. & Grundy, C. (1996). Cancer incidence near municipal solid waste incinerators in Great Britain, *British Journal of Cancer* **73**, 702–710.

[17] Elliott, P., Stamler, J., Nichols, R., Dyer, A.R., Stamler, R., Kesteloot, H. & Marmot, M. (1996). INTERSALT revisited: further analyses of 24 hour sodium excretion and blood pressure within and across populations, *British Medical Journal* **312**, 1249–1253.

[18] Elliott, P. & Wakefield, J.C. (2000). Bias and confounding in spatial epidemiology, Ch. 5 in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakwfield, N.G. Best & D.J. Briggs, eds. Oxford University Press, Oxford, pp. 68–84.

[19] English, D. (1992). Geographical epidemiology and ecological studies, in *Geographical and Environmental Epidemiology: Methods for Small-Area studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. Oxford University Press, Oxford, pp. 3–13.

[20] Gardner, M.J., Winter, P.D. & Barker, D.J.P. (1984). *Atlas of Mortality from Selected Diseases in England and Wales 1968–1978*. Wiley, Chichester.

[21] Green, P.J. & Richardson, S. (2002). Hidden Markov models and disease mapping, *JASA*, to appear.

[22] Jarup, L., Best, N.G., Toledano, M.B., Wakefield, J.C. & Elliott, P. (2002). Geographical epidemiology of prostate cancer in Great Britain, *International Journal of Cancer* **97**, 695–699.

[23] Kemp, I., Boyle, P., Smans, M. & Muir C. (1985). *Atlas of Cancer in Scotland, 1975–1980, Incidence and Epidemiologic Perspective*. IARC Scientific Publication No. 72, International Agency for Research on Cancer, Lyon.

[24] Keys, A., ed. (1970). *Coronary Heart Disease in Seven Countries*. American Heart Association Monograph no. 29. American Heart Association, New York.

[25] Knorr-Held, L. & Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps, *Biometrics* **56**, 13–21.

[26] Kurtzke, J.F. (1985). Neurological system, in *Oxford Textbook of Public Health*, Vol. 4. Oxford University Press, Oxford, Chapter 12, pp. 203–249.

[27] Muir, C., Waterhouse, J., Mack, T., Powell, J. & Whelan, S., eds. (1987). *Cancer Incidence in Five Continents*, Vol. V. IARC Scientific Publication No. 88, International Agency for Research on Cancer, Lyon.

[28] Piantadosi, S., Byar, D.P. & Green, S.B. (1988). The ecological fallacy, *American Journal of Epidemiology* **127**, 893–904.

[29] Plummer, M. & Clayton, D. (1996). Estimation of population exposure in ecological studies, *Journal of the Royal Statistical Society, Series B* **58**, 113–126.

[30] Pothoff R.F. & Whittinghill, M. (1966). Testing for homogeneity. II. The Poisson distribution, *Biometrika* **53**, 183–190.

[31] Poulter, N.R., Khaw, K.T., Hopwood, B.E.C., Mugambi, M., Peart, W.S., Rose, G. & Sever, P.S. (1990). The Kenyan Luo migration study: observations on the initiation of a rise in blood pressure, *British Medical Journal* **300**, 967–972.

[32] Richardson, S. & Monfort, C. Ecological correlation studies, Ch. 11 in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best & D.J. Briggs, eds. Oxford University Press, Oxford, pp. 205–219.

[33] Rose, G. (1985). Sick individuals, sick populations, *International Journal of Epidemiology* **14**, 32–38.

[34] Schlattmann, P., Dietz, E. & Bohning, D. (1996). Covariate adjusted mixture models and disease mapping with the program DismapWin, *Statistics in Medicine* **15**, 919–929.

[35] Smans, M. & Esteve, J. (1992). Practical approaches to disease mapping, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. Oxford University Press, Oxford, pp. 141–157.

[36] Smith, A.F.M. & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, Series B* **55**, 3–23.

[37] Spiegelhalter, D.J., Thomas, A., Best, N.G. & Lunn, D. (2002). WinBUGS Bayesian inference Using Gibbs Sampling Manual Version 1.4, Imperial College, London and MRC Biostatistics Unit, Cambridge, `available from http://www.mrc-bsu.cam.ac.uk/bugs`.

[38] Swerdlow, A. & dos Santos Silva, I. (1993). *Atlas of Cancer Incidence in England and Wales 1968–85*. Oxford University Press, Oxford.

[39] Wakefield, J.C., Best, N.G. & Waller, L.A. (2000). Bayesian approaches to disease mapping, Ch. 7 in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best & D.J. Briggs, eds. Oxford University Press, Oxford, pp. 104–127.

[40] Wakefield, J.C. & Salway, R. (2001). A statistical framework for ecological and aggregate studies, *Journal of the Royal Statistical Society, Series A* **164**, 119–138.

[41] Walter, S.D. (1993). Visual and statistical assessment of spatial clustering in mapping data, *Statistics in Medicine* **12**, 1275–1291.

[42] Walter S.D. & Birnie, S.E. (1991). Mapping mortality and morbidity patterns: an international comparison, *International Journal of Epidemiology* **20**, 678–689.

[43] Wilkinson, P., Thakrar, B., Shaddick, G., Stevenson, S., Pattenden, S, Landon, M., Grundy, C. & Elliott, P. (1997). Cancer incidence and mortality around the Pan Britannica Industries pesticide factory, Waltham Abbey, *Occupational and Environmental Medicine* **54**, 101–107.

[44] World Bank (1993). *World Development Report 1993, Investing in Health*. Oxford University Press, Oxford.

PAUL ELLIOTT & NICOLA BEST

Geometric Distribution

In a coin-tossing problem, let X be the number of tosses required to obtain a head. The probability that $X = k$, $\Pr(X = k)$, equals $(1-p)^{k-1}p$, $k = 1, 2, \ldots$. Here p, $0 \le p \le 1$, is the probability of obtaining a head in a single toss. The **random variable** X with this probability mass function (pmf) is known to have the geometric distribution. The geometric distribution can also be identified with the experiment of drawing balls from an urn with replacement. Suppose an urn contains b black balls and w white balls. Let X be the number of drawings needed to draw a white ball from the urn; then X is a geometric random variable with $p = w/(b+w)$. For the geometric random variable, the probability $\Pr(X = k)$ is a monotone decreasing function of k. It is evident that $\Pr(X = k+t) = \Pr(X = k)(1-p)^t < \Pr(X = k)$. The random variable X has the memoryless property, i.e.

$$\begin{aligned}\Pr(X > r+s) &= \sum_{j=r+s}^{\infty} (1-p)^j p = (1-p)^{r+s} \\ &= \Pr(X > r)\Pr(X > s),\end{aligned}$$

where r and s are positive integers. The **hazard** function $h(x) = \Pr(X = x | X \ge x)$ is independent of x, $x = 1, 2, \ldots$. The probability **generating function** $G(s)$ of X is $G(s) = \mathrm{E}(s^X) = ps(1-qs)^{-1}$, where $q = 1-p$. We have the probability

$$\Pr(X = j) = \left[\frac{1}{j!}\frac{\mathrm{d}^j G(s)}{\mathrm{d}s^j}\right]_{s=0}.$$

The noncentral **moment generating function** is $G(\mathrm{e}^s)$. The rth noncentral **moment**, μ'_r, of X, namely $\mu'_r = \mathrm{E}(X^r)$, is

$$\sum_{j=1}^{\infty} j^r (1-p)^{j-1} p = \left[\frac{\mathrm{d}^r G(\mathrm{e}^s)}{\mathrm{d}s^r}\right]_{s=0}.$$

The central moment, $\mu_r = \mathrm{E}(X - \mu'_1)^r$, $r = 1, 2, \ldots$, has generating function $p\exp(-qs/p)(1 - q\mathrm{e}^s)^{-1}$, $t < -\ln(1-p)$. Using this central moment generating function, the following recurrence relations between the central moments can easily be established:

$$\mu_{r+1} = q\frac{\partial \mu_r}{\partial q} + \frac{rq}{p^2}\mu_{r-1},$$

with $r \ge 1$, $\mu_0 = 1$ and $\mu_1 = 0$. Thus the mean of X is

$$\mu'_1 = \mathrm{E}(X) = \sum_{j=1}^{\infty} j(1-p)^{j-1}p = \frac{1}{p},$$

and the variance of X is

$$\mathrm{var}(X) = \sum_{j=1}^{\infty} j^2(1-p)^{j-1}p - \mu_1^2 = \frac{1-p}{p^2}.$$

For instance, for an unbiased coin ($p = 1/2$) the number of tosses until a head appears has an expectation 2 and a variance of 2.

Suppose $X_{1,m} = \min(X_1, \ldots, X_m)$, where $X_1, \ldots, X_m$ are m independent copies of X, then $\Pr(X_{1,m} > x) = q^{mx}$, $x = 1, 2, \ldots$; $\Pr(X_{1,m} = x) = \Pr(X_{1,m} > x-1) - \Pr(X_{1,m} > x) = q^{m(x-1)}(1-q^m)$. Thus $X_{1,m}$ is a geometric random variable. Let $X_{2,2} = \max(X_1, X_2)$ and the **range**, $R_2 = X_{2,2} - X_{1,2}$. We have $\Pr(R_2 = 0) = \Pr(X_{2,2} = X_{1,2}) = \sum_{x=1}^{\infty}(pq^{x-1})^2 = p^2/(1-q^2)$. Since $\Pr(X_{1,2} = x, R_2 = 0) = \Pr(X_{1,2} = x, X_{2,2} = x) = (pq^{x-1})^2 = \Pr(X_{1,2} = x)\Pr(R_2 = 0)$, the events $\{R_2 = 0\}$ and $X_{1,2}$ are independent. $\Pr(R_2 = r) = \sum_{j=1}^{\infty}\Pr(X_{1,2} = j, X_{2,2} = j+r) = \sum_{j=1}^{\infty} 2p^2q^{2(j-1)+r} = 2pq^r/(1+q)$, $r = 1, 2, \ldots$. Using the equations

$$\sum_{r=1}^{\infty} 2r\frac{pq^r}{1+q} = \frac{2q}{p(1+q)}$$

and

$$\sum_{r=1}^{\infty} 2r(r-1)\frac{pq^{r-1}}{1+q} = \frac{4q^2}{p^2(1+q)},$$

we obtain $\mathrm{E}(R_2) = 2q/[p(1+q)]$ and $\mathrm{var}(R_2) = 2q(1+q^2)/p^2(1+q)^2$.

Further, $\Pr(X_{1,2} = j, R_2 = r) = \Pr(X_{1,2} = j, X_{2,2} = j+r) = 2p^2q^{2(j-1)+r} = \Pr(X_{1,2} = j)\Pr(R_2 = r)$ for all j and r, $1 \le j$, $r < \infty$; hence $X_{1,2}$ and R_2 are independent random variables. This independence property of the range and the minimum observation is also true for $m > 2$.

M. AHSANULLAH

Gerontology and Geriatric Medicine

We can distinguish some important features of geriatric medicine – a branch of medicine – and gerontology – the study of older people in general, healthy or ill. These include:

1. **Co-morbidity**; older patients typically suffer from complaints other than the presenting complaint, a fact which complicates both treatment and research.
2. An increasing **prevalence** of degenerative disease and sensory impairment, particularly in the later years of life; this has important consequences for treatment, and also for research.
3. The occurrence of almost any disease during this period of life, almost any of which can occur before. Nonetheless, the dementias are a high prevalence disease after age 75 but are almost unknown before age 55. Even when dementia is not at issue, cognitive performance may not be at a level to permit easy treatment or cooperation in research.

The exact age cutoff is arbitrary, and is based on historical sociopolitical considerations rather than on pathophysiological ones. Most research studies use 65 years, although 75 is becoming more popular. The term geriatric medicine itself has been superseded by some variant on health care of the elderly. Within this umbrella, **psychiatry** of old age is included, but there seems to be no specialty of, for example, geriatric surgery.

Statistical Issues

The field of gerontology, here encompassing geriatric medicine, has not spawned any specific statistical methods. Its journals publish papers using standard techniques, but do not stand out from other specialties statistically.

Randomized Control Trial (RCT) and the Elderly

The RCT (*see* **Clinical Trials, Overview**) has become established as the preferred method of evaluating therapeutic alternatives when it is possible to carry it out. The dominant ideology in selecting patients for an RCT has been that patients should be included as clear cases of the disease at issue, and no other (*see* **Eligibility and Exclusion Criteria**). This has brought three problems:

1. Because of the problem of comorbidity, trials recruit disproportionately from younger age groups, even when they do not employ an explicitly or implicitly ageist recruitment policy.
2. Trials exclusively in old age, for instance in dementia, have very high ratios of screened to enrolled patients.
3. High rates of mortality, sometimes from competing causes (*see* **Competing Risks**), have made **intention-to-treat analysis** vital. Dropout in an RCT in older people will seldom be at random.

The inevitable paradox is that older patient groups where most illness occurs have been least studied, and treatments must be used on the basis of optimistic generalization from younger age groups. For instance, in an overview of European studies of thrombolysis after acute myocardial infarction, it was found that patients aged 75 and over represented 33% of those requiring treatment in the population but only 10% of those enrolled in the major trials [1].

The issue of informed consent does not seem to be responsible for this deficit in recruitment. In general, workers in old age research have only seen this as a problem in the field of dementia, and have usually relied on consent from relatives in such trials.

Measurement and Scaling

As mentioned above, this field has to cope with varying degrees of sensory and motor impairment. This has caused workers here to be more interested in problems of measurement (*see* **Measurement Scale**), especially of functional ability, than in almost any other branch of medicine except psychiatry. There is awareness of the problems of constructing measures which span very large ranges of ability in the face of concurrent sensory and cognitive difficulties, but, as yet, no agreement on how to solve them.

Healthy Active Life Expectancy

Although this technique can be applied to all age groups it has been used extensively in gerontology [4]. Sullivan's index [2] takes standard **life table**

methods and attempts to combine information on mortality and morbidity.

Consider $T + 1$ age groups where the age at the start of each group is x_i, with $i = 0, 1, \ldots, T$. Based on the usual life table notation, l_i is the number of survivors to age x_i, L_i the number of **person-years** in age group i, and π_i the prevalence of the health status in question in age group i.

We partition the expectation at age k,

$$e_k = \frac{1}{l_k} \sum_{i=k}^{T} L_i, \tag{1}$$

into a diseased part D and a healthy part H:

$$e_k^{[D]} = \frac{1}{l_k} \sum_{i=k}^{T} \pi_i L_i, \tag{2}$$

$$e_k^{[H]} = \frac{1}{l_k} \sum_{i=k}^{T} (1 - \pi_i) L_i. \tag{3}$$

This device has been seen as a way of measuring and answering the question of whether increases in **life expectancy** have merely led to an increase in the period spent disabled before death, or whether the onset of disability has been postponed in parallel with death itself [4]. Clearly for diseases from which recovery is possible a more complex model is needed [2].

There has also been interest in the question of whether it is possible to estimate the limits on longevity [3]. This has usually taken the form of estimating what changes in q_i (the **conditional probability** of mortality in group i) would be needed to increase life expectancy to 85, 100 years, or beyond.

References

[1] European Secondary Prevention Study Group (1996). Translation of clinical trials into practice: a European population-based study of the use of thrombolysis for acute myocardial infarction, *Lancet* **347**, 1203–1207.

[2] Newman, S.C. (1988). A Markov process interpretation of Sullivan's index of morbidity and mortality, *Statistics in Medicine* **7**, 787–794.

[3] Olshansky, S.J., Carnes, B.A. & Cassel, C. (1990). In search of Methuselah: estimating the upper limits to human longevity, *Science* **250**, 634–640.

[4] Robine, J.M. & Ritchie, K. (1991). Healthy life expectancy: evaluation of global indicator of change in population health, *British Medical Journal* **302**, 457–460.

MICHAEL E. DEWEY

Gestational Age

Gestational age is the duration of a pregnancy. According to the **World Health Organization** (WHO), the duration of gestation is measured from the first day of the last normal menstrual period [3]. The rationale for this is that women are more likely to be able to recall the dates of their last period than they are to know the date when conception took place. Of course, there are circumstances when the opposite may be true.

If women conceive soon after stopping oral contraception, this may happen before they have had a spontaneous period. For women with irregular periods, the relationship between the beginning of their last period and their date of conception may be problematic. Because of this, other means are commonly used to estimate gestational age and hence the likely date of delivery. In settings where ultrasound examinations are standard practice, gestational age is assessed by measuring the biparietal diameter. In less developed countries, the fundal height is commonly used to estimate gestational age, but is generally considered to be unreliable. Finally, the gestational age of babies assessed by examination after birth is known as the "pediatric assessment". Given that none of the methods is perfect, the common practice is to choose the one considered most reliable as the "best estimate" of gestational age for a given woman.

Gestational age is expressed in completed days or weeks, so events occurring 280 to 286 days after the onset of the last normal menstrual period are considered to have occurred at 40 completed weeks of gestation. In calculating the gestational age, WHO stipulates that the first day of the last period should be regarded as day zero and not day one, so days zero to six therefore correspond to week zero. Thus, for example, events in the 40th week of gestation should be described as taking place after 39 completed weeks of gestation.

Within a given gestation, WHO distinguishes between preterm, term and postterm:

1. *Preterm*: Less than 37 completed weeks (less than 259 days) of gestation.
2. *Term*: From 37 completed weeks to less than 42 completed weeks (259 to 293 days) of gestation.
3. *Postterm*: From 42 completed weeks or more (294 days or more) of gestation.

With increasing interest in the preterm period, it is becoming common practice to identify two further categories: extremely preterm, which is less than 28 completed weeks or 196 completed days of gestation; and very preterm, which is less than 32 completed weeks or 224 completed days of gestation [1].

It is recommended that tabulations by week of gestational age group them as 22–23, 24–25, 26–27, 28–31, 32–36, 37–41 and 42 or more completed weeks of gestation [1].

In the past, the word "premature" was used to denote a birth that was either preterm or of low **birthweight** or both, but it is now considered insufficiently precise and therefore incorrect. The definitions above were introduced in the ninth revision of the **International Classification of Diseases** (ICD) [2] and should enable a distinction to be made between babies who are small because they are born too soon and those who are small for their gestational age.

References

[1] European Association of Perinatal Medicine (1996). *Perinatal Audit*. Parthenon, London.

[2] World Health Organization (1977). *International Classification of Diseases. Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*. 9th Rev., Vol. 1. WHO, Geneva.

[3] World Health Organization (1992). *International Classification of Diseases and Related Health Problems*, 10th Rev., Vol. 1. WHO, Geneva.

(*See also* **Reproduction**; **Vital Statistics, Overview**)

ALISON MACFARLANE

Ghosts

In the analysis of survival data X_i, left truncated at v_i, $i = 1, \ldots, n$ (*see* **Truncated Survival Times**) each observed life time $X_i = x_i > v_i$ "can be considered the remnant of a group, the size of which is unknown and all (except the one observed) with x-values less than or equal to v. (They can be thought of as X_i's ghosts.)". The quotation is (slightly adapted) from Turnbull's [5] use of the concept of *ghosts* to denote the number by which each truncated observation needs to be upweighted to provide the nonparametric maximum likelihood estimator of the survival function (*see* **Turnbull Estimator** for details).

This use of "ghosts" is similar to the handling of **censoring** by Inverse Probability of Censoring Weighted (IPCW) estimating equations, which is analogous to the **Horvitz–Thompson** device [1] in sampling theory introduced into survival analysis by Koul et al. [3] and used extensively by Robins (see [4], but also [2] and [6]). The rationale is again to compensate for the attrition due to censoring or truncation by letting each *observed* event represent not only him/herself, but also the "ghosts" who were not observed.

References

[1] Cochran, W.G. (1983). Horvitz-Thompson estimator, in *Encyclopedia of Statistical Sciences*, Vol. 3, N.L. Johnson & S. Kotz, eds. Wiley, New York, pp. 665–668.

[2] Keiding, N., Holst, C. & Green, A. (1989). Retrospective estimation of diabetes incidence from information in a prevalent population and historical mortality, *American Journal of Epidemiology* **130**, 588–600.

[3] Koul, H.L., Susarla, V. & Van Ryzin, J. (1981). Regression analysis with random right-censored data, *Annals of Statistics* **9**, 1276–1288.

[4] Robins, J.M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, in *AIDS Epidemiology, Methodological Issues*, N.P. Jewell, K. Dietz & V.T. Farewell, eds. Birkhäuser, Boston, pp. 297–331.

[5] Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 291–295.

[6] Whittemore, A.S. & Halpern, J. (1997). Multi-stage sampling in genetic epidemiology, *Statistics in Medicine* **16**, 153–167.

(*See also* **Survival Analysis, Overview**)

NIELS KEIDING

Gini, Corrado

Born: May 23, 1884.
Died: March 13, 1965.

Gini was the leading Italian statistician during the first half of the twentieth century. He became a university professor at the age of 26, and held chairs at the universities of Cagliari, Padua, and Rome. He founded *Metron* in 1920, owned the journal until 1962, and directed it until his death. He contributed widely to the development of statistical methods, with applications in **demography**, biometry, sociology, and economics. He was particularly influential in the development of mathematical approaches to descriptive statistics. In 1912, Gini introduced the *mean difference* as a measure of variation within a set of quantities, defined as the mean of the absolute differences between pairs of observations. In 1914, he showed that this is related to the *area of concentration* derived from a *Lorenz curve* representing inequalities in income distribution. (The Lorenz curve relates the proportion of a population, ordered by size of income, to the proportion of total income received by that group.)

PETER ARMITAGE

Gold Standard Test

In attempting to classify individuals as cases or noncases of disease, clinicians usually apply one or more **diagnostic tests**. Typically, however, these tests are subject to **measurement error**, and may thus display less than perfect **sensitivity** or **specificity**.

The term "gold standard" is applied to a test that, theoretically at least, is regarded as being error-free, that is, having 100% sensitivity and specificity. If such a test exists, then it can be used as a basis for comparison for any other candidate test. The gold standard may also be known as a "*reference test*" or method.

Unfortunately, definitive diagnosis through a gold standard will usually require an invasive or hazardous clinical intervention. For instance, while a chest X-ray or CT scan (relatively noninvasive tests) may provide suggestive evidence that a pulmonary tumor is present, a histologic evaluation of a biopsy specimen obtained by surgical intervention is conventionally required to confirm the diagnosis of a malignancy.

In practice, even tests that are consensually regarded as gold standards may be subject to error. In the lung cancer example, there is variation in where and how the surgeon elects to sample potentially malignant tissue; furthermore, there is some subjectivity in the pathologist's assessment of the biopsy material, with the result that inter- and intraobserver variation exists in the final diagnostic classification (*see* **Agreement, Measurement of**; **Observer Reliability and Agreement**). Such variation reveals the fact that the gold standard cannot be perfect.

In clinical practice, therefore, the term "gold standard" is often applied to tests that are regarded as the best available with current technology, even though the best method may sometimes be erroneous. A consequence of assuming a test to be gold standard, even though it is error-prone, is that any other competitor test that is compared with the gold standard will have apparently inflated error rates; in other words, estimates of sensitivity or specificity of a comparison test relative to the gold standard will be **biased** downwards. There is a growing literature on the appropriate analysis of data involving such "alloyed" gold standard tests (*see* **Diagnostic Test Evaluation Without a Gold Standard**)

STEPHEN D. WALTER

Goodman–Kruskal Measures of Association

The popularity of these measures stems from the articles by Goodman & Kruskal that span two decades [1–4]. All are used to examine the association between two categorical variables (A and B, say). The λ and τ measures are suitable for general use, with the measure γ being used when both A and B are ordinal. All are commonly included in computer packages.

The Lambda Measures

The simple idea underlying the λ and τ measures is that it may be possible to predict the category of one variable from knowledge of the category of the other variable.

As an example, suppose that two boys play a game in which boy A thinks of some living thing. Boy B has to guess whether it is a creature or a plant. If B knows that A thinks of plants on 60% of occasions, then one strategy for B would be to guess "Plant" every time (with a 60% success rate). However, if B is allowed to ask a question such as "Does it have legs?", then B's subsequent guess will be much more accurate!

Suppose that we know that, for variable A, category m is most frequent. In the absence of other information we would guess that a new individual belonged to m. However, if told that the individual belongs to category j of variable B, then we could examine all past records for this category. If these records showed i to be the commonest category of A for these individuals, then we would guess i and not m. Unless A and B are independent, this method reduces the probability of a guess being in error. The λ and τ measures are said to measure the *proportional reduction in error* (PRE), with 0 corresponding to independence and 1 corresponding to the complete elimination of error (e.g. as a result of the question "Can it move?"!).

Consider the data of Table 1. Knowing nothing of B we guess A_1 with Pr(error) $= 144/300 = 0.48$. If we know that an individual belongs to the second category of B, then we guess A_2 and otherwise guess A_1. The probabilities of the three categories

Table 1 Example data

	A_1	A_2	Total	Most common category	Guess	Probability of error
B_1	16	14	30	1	1	$\frac{14}{30}$
B_2	38	62	100	2	2	$\frac{38}{100}$
B_3	102	68	170	1	1	$\frac{68}{170}$
Overall	156	144	300	1	1	$\frac{144}{300}$

are estimated by 30/300, 100/300 and 170/300, so that the overall probability of an error is now:

$$\left(\frac{14}{30}\times\frac{30}{300}\right)+\left(\frac{38}{100}\times\frac{100}{300}\right)+\left(\frac{68}{170}\times\frac{170}{300}\right)=\frac{120}{300}=0.40.$$

Hence the proportionate reduction in error is given by:

$$\lambda_A=\frac{\frac{144}{300}-\frac{120}{300}}{\frac{144}{300}}=\frac{144-120}{144}=\frac{24}{144}=\frac{1}{6}.$$

The suffix A indicates that we are predicting the category of A from that of B. Suppose, instead, that we are asked to guess the category of B, knowing that of A. There is an immediate problem! For *both* categories of A we would guess B_3, and so with this procedure there is no reduction in error. Formally,

$$\lambda_B=\frac{\frac{130}{300}-\left\{\left(\frac{54}{156}\times\frac{156}{300}\right)+\left(\frac{76}{144}\times\frac{144}{300}\right)\right\}}{\frac{130}{300}}=\frac{130-130}{130}=\frac{0}{130}=0.$$

The problem is caused by the dominating size of a single category. The value of λ is therefore most informative when the categories are roughly equally likely.

In cases where it is equally reasonable for the predictions to proceed in either direction, the hybrid statistic λ is used. This combines the two numerators and denominators:

$$\lambda=\frac{24+0}{144+130}=\frac{24}{274}=0.09.$$

The Tau Measures

The τ measures differ from the λ measures by using a different guessing strategy. Instead of always choosing the most common category, categories are guessed in proportion to their known probabilities of occurrence. Although the resulting predictions are less successful, and the resulting formulae are more complicated, the problems with unbalanced category frequencies are avoided. The value 0 can only occur for a table displaying perfect independence. For the example data $\tau_A = 0.31$, $\tau_B = 0.18$, and $\tau = 0.24$.

Note that $\tau > \lambda$ tells us nothing about their relative merits.

The Measure Gamma

This too can be viewed as a PRE measure. It is suitable only for cases where both variables have ordered categories, as in the data of Table 2, which are obtained from wave 1 of the British Household Panel Study.

Suppose we predict that those in better health will also be more satisfied with their job. To test this idea we need to look at *pairs* of individuals. The usefully informative pairs belong both to different health categories and to different job categories. There are 962 individuals who felt satisfied and in excellent health. There are $(781 + 329 + 100 + 67) = 1277$ individuals who felt both less healthy and less satisfied. If we choose one individual out of each of these two groups (which we can do in 962×1277 different ways), then we will have a pair of individuals who match the prediction. There are other pairs also. The total number of concordant pairs is C:

$$C = (962 \times 1277) + \{1213 \times (329 + 67)\} + \{447 \times (100 + 67)\} + (781 \times 67) = 1\,835\,798.$$

However, there are also discordant pairs. The total number, D, is given by:

$$D = \{420 \times (447 + 781 + 62 + 100)\} + \{1213 \times (447 + 62)\} + \{329 \times (100 + 62)\} + (781 \times 62) = 1\,302\,937.$$

Table 2 Job satisfaction and self-reported health

	Self-reported health		
	Excellent	Good	Not good
Satisfied with job	962	1213	420
Neutral about job	447	781	329
Not satisfied with job	62	100	67

The statistic γ is given by

$$\gamma = \frac{C - D}{C + D}.$$

The range of possible values is from -1 (complete discordance) to $+1$ (complete concordance). For the given data $\gamma = 0.17$. The data suggest a very weak positive association between health and job satisfaction.

The significance of any of these measures can be assessed by comparison with its standard error. Software packages (such as SPSS and SAS) routinely report standard errors and tail probabilities in addition to the measure values.

References

[1] Goodman, L.A. & Kruskal, W.H. (1954). Measures of association for cross-classifications, I, *Journal of the American Statistical Association* **49**, 942–946.

[2] Goodman, L.A. & Kruskal, W.H. (1959). Measures of association for cross-classifications, II, *Journal of the American Statistical Association*, **54**, 123–163.

[3] Goodman, L.A. & Kruskal, W.H. (1963). Measures of association for cross-classifications, III, *Journal of the American Statistical Association* **58**, 310–364.

[4] Goodman, L.A. & Kruskal, W.H. (1972). Measures of association for cross-classifications, IV, *Journal of the American Statistical Association* **67**, 415–421.

(*See also* **Association, Measures of**; **Categorical Data Analysis**; **Contingency Table**)

G. UPTON

Goodness of Fit in Survival Analysis

As in contexts other than **survival analysis**, the idea of **goodness of fit** is to provide a more or less formal indication that some or all of the modeling assumptions made can be considered reasonable. Goodness of fit is related to but distinct from; regression **diagnostics**, **robust regression**, influential data analysis, **cross-validation** and, in particular, the **prediction** problem as quantified by R^2 measures. None of these are considered in this review, although it would seem important to point out that R^2-type prediction measures [26, 41, 46] are not goodness of fit measures as is very widely believed. Korn & Simon [27] provide some discussion on this. We will limit ourselves here to the observation that whereas a very poor fit will in general lead to a low value of R^2, the converse is not so, a simple example coming from linear regression with near zero slope and large residual variance, in which the fit may be perfect but we would anticipate the value of R^2 to be small.

Goodness of fit techniques help us address the very limited question of whether or not the observed data appear to come into conflict with one or more of the basic assumptions underscoring the adopted analytic approach. In this review the major part of our attention is on the **proportional hazards** model. This is for two reasons. First, the literature concerning parametric approaches is more classical and well reviewed elsewhere (*see* **Parametric Models in Survival Analysis**). Secondly, the proportional hazards model has become quite overwhelmingly the model of choice in applied studies so that there is much less interest in the goodness of fit question for parametric models. This having been said, it is quite likely that there will be renewed interest in parametric modeling in the future, and we include some discussion on the goodness of fit problem below.

Many tests have been proposed for testing the proportional hazards assumption, focusing on some feature of the assumed time independence of the regression effect, on the form of the link function, on the functional form of the covariates, and sometimes all of these. It needs to be stressed that real departures from proportional hazards can manifest themselves in different ways; an incorrect functional form appearing as a time-dependent effect, for example. Trying to separate out the different types of departure is most often not possible. So, although it is worthwhile considering tests having good power against a particular kind of departure, we ought be cautious about interpreting significant results; the departure can easily be of a quite different nature.

In the following sections we organize the different procedures under the headings omnibus tests, directional tests, and graphical techniques. A test designed to detect general nonspecific departures from the null hypothesis is called an omnibus test. In some cases a test might be especially designed to have high power against departures in certain directions; we refer to these as "tests for specific alternatives", although some authors call them "directional tests" (see, for example, Lawless [29]). Graphical techniques often provide an easy check on model assumptions. Although they are not formal tests, they are an important aspect of goodness of fit tests, and can sometimes be combined with formal ones to obtain significance levels. Finally, tests based on the empirical distribution function such as **Kolmogorov–Smirnov tests**, Cramér–von Mises type tests and the Anderson–Darling test are well known (*see* **Kolmogorov–Smirnov and Cramer–Von Mises Tests in Survival Analysis**). However, their generalization to arbitrary **censoring** is difficult, unless the censoring is type I or type II (*see* **Level of a Test**). The interested reader can find those in Lawless [29] and Andersen et al. [4]. Alternatively, resampling techniques could be used (*see* **Bootstrap Method**).

Notation and Data Sets

In the following, let T denote the failure time random variable. In a survival study with n subjects, let $T_1, T_2, \ldots, T_n$ be the true failure times, and $C_1, C_2, \ldots, C_n$ the potential **censoring** times for the individuals $i = 1, 2, \ldots, n$. We observe $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$ for each i. Denote $Y_i(t) = 1$ if $X_i \geq t$ and 0 otherwise. Often, we will suppose that the failure time of each subject is related to a vector of **covariates**, or **explanatory variables**, $\mathbf{Z}_i = (Z_{il}, \ldots, Z_{ip})', i = 1, 2, \ldots, n$.

We illustrate some of the techniques on two data sets. The first concerns the well-known Freireich et al. [17] study comparing two treatments in leukemia. The data consists of remission times in

weeks of 42 patients, 21 having received 6-MP and 21 having being treated with placebo. These data, used to illustrate the proportional hazards model in Cox [12], are widely believed to fit the model well. In contrast, the second data set, analyzed by Stablein et al. [49], indicates clear departures from model assumptions. These data concern a clinical trial in gastric carcinoma in which 90 patients were treated; 45 treated by chemotherapy alone and 45 treated by chemotherapy plus radiotherapy.

Parametric Models

There are various ways to assess the fit of parametric models without covariates. When covariates are included, it is often possible to adapt the methods to deal with this. For more general situations, very little has been done when compared with the proportional hazards model, and it is fair to say that such models are not often used in practice. We consider goodness of fit tests for several commonly seen parametric models without covariates used in survival analysis. These include the **exponential** model, the **Weibull** and **extreme value** models, and the **lognormal** models. Notice that the problem of testing a lognormal model is equivalent to that of testing a normal model.

Graphical Tests

The density or hazard functions are too noisy to estimate nonparametrically without resorting to some kind of smoothing technique (*see* **Smoothing Hazard Rates**). As a general principle, such smoothing requires some skill on the part of the user, and the whole subject is rather too vast to be given any coverage here. For our purposes we will focus attention on a nonparametric estimate of the survivorship function, either the **Kaplan–Meier** estimate or the **Nelson–Aalen** estimate.

Once we have an estimate of the survivorship function $S(t)$ in hand, then we can usually derive simple graphical techniques which cannot only provide some indication as to the goodness of fit but also some very simple estimates of the unknown parameters (*see* **Graphical Displays**; **Hazard Plotting**). Even if such estimates are only to be used as starting values to an iterative cycle in **maximum likelihood** estimation, this alone can be valuable computationally.

We can illustrate the above via some well known models. For the exponential distribution with parameter λ, we can plot $\log \hat{S}(X_i)$ against X_i. Under the model this will be a straight line with zero intercept and slope $-\lambda$. For the Weibull distribution with location parameter λ and scale parameter k, we can plot $-\log\{-\log \hat{S}(X_i)\}$ against $\log X_i$. Under the model this will be a straight line with intercept $\log \lambda$ and slope k. For the **Pareto** model with parameter c a plot of $-\log \hat{S}(X_i)$ against $\log X_i$ will have zero intercept and slope c. For the lognormal distribution any of the usual techniques for assessing normality will work, assuming that we can consistently estimate the mean and the variance of $\log T$. For the log **logistic** model with location parameter λ and shape parameter k, a plot of $\log\{1 - \hat{S}(X_i)\} - \log \hat{S}(X_i)$ vs. $\log X_i$ will be linear with intercept equal to $k \log \lambda$ and slope equal to k. For the extreme value model with mode a and variance equal to $\pi^2 b^2/6$, a plot of $\log\{-\log \hat{S}(X_i)\}$ vs. X_i will be linear with intercept equal to $-a/b$ and slope equal to $1/b$.

Apparent departures from the above as indicated by the plot will be indicative of one of three possibilities: the chosen parametric model is not of the correct form, the observed times are not independent or, thirdly and not to be overlooked, the assumption of an independent censoring mechanism is not sufficiently plausible.

One of the commonly used tools in graphical tests is the **residual** plot. The basic idea here is the same as that of the usual $Q-Q$ plots, which can be found in textbooks such as Rice [45] (*see* **Normal Scores**). In survival analysis, it is common to define residuals based on the unit exponentiality property: if T has cumulative hazard function $\Lambda(\cdot)$, then $\Lambda(T)$ has a unit exponential distribution. So define

$$\varepsilon_i = \Lambda(X_i) \quad \text{or} \quad \varepsilon_i = \Lambda(X_i|\mathbf{Z}_i) \tag{1}$$

in the presence of covariates, and their estimates

$$\hat{\varepsilon}_i = \hat{\Lambda}(X_i) \quad \text{or} \quad \hat{\varepsilon}_i = \hat{\Lambda}(X_i|\mathbf{Z}_i), \tag{2}$$

which involve the maximum likelihood estimates of the unknown parameters. One approach is to treat $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ as a possibly censored sample from the unit exponential distribution. One then calculates the Kaplan–Meier estimate of the survival function, $\hat{S}(\cdot)$, and plots $-\log\{\hat{S}(t)\}$ vs. t. When the model is adequate the plot should give roughly a straight line with slope one. A second approach uses the fact that if

ε is distributed as unit exponential, then $\mathrm{E}(\varepsilon|\varepsilon \geq t) = t + 1$. So in the event of a censored observation X_i, one replaces the censored residual $\hat{\varepsilon}_i$ by the adjusted residual $\tilde{\varepsilon}_i = \hat{\varepsilon}_i + 1$, and treats all the residuals as if they are uncensored. This procedure is convenient if there are only a few censored observations, or if one wants to plot the residuals against other factors such as the covariates.

For exponential distribution with hazard rate λ, Nelson [39] suggested the following technique for checking model adequacy. Under the model, the cumulative hazard function $\Lambda(t) = \lambda t$. Let the ordered failure times be $T_{(1)} \leq \cdots \leq T_{(n)}$. Then

$$\mathrm{E}\{\Lambda(T_{(i)})\} = \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-i+1}.$$

Thus he suggests plotting $T_{(i)}$ against the right-hand side of the above, and the points should cluster around a straight line for a reasonably good fit. He further points out that the procedure is satisfactory even in the presence of censoring.

Formal Tests

To check the fit of data to a certain parametric model, one type of test is obtained by embedding the null model into a larger class of parametric models (*see* **Hierarchical Models**). Standard procedures such as a score test (*see* **Likelihood**) or **likelihood ratio test** can then be used to test for particular parameter values, and arbitrarily censored data are accommodated. Of course, tests of this kind may not be effective at detecting departures that are not closely approximated by a member of the larger family of models. Specifically, an exponential model can be seen as a Weibull model

$$\lambda(t) = \kappa\theta(\theta t)^{\kappa-1} \tag{3}$$

with $\kappa = 1$, where $\lambda(\cdot)$ is the hazard function. A test of the hypothesis $\kappa = 1$ is against the alternative of monotone hazard functions. Meanwhile, an extreme value distribution (which is the distribution of the logarithm of a Weibull random variable) can be embedded in a three-parameter log **gamma** model with pdf $\int(x)$ equal to

$$\frac{|\lambda|(\lambda^{-2})^{\lambda^{-2}}}{\sigma\Gamma(\lambda^{-2})} \exp\left\{\frac{x-\mu}{\lambda\sigma} - \lambda^{-2}\exp\left[\frac{\lambda(x-\mu)}{\sigma}\right]\right\}, \quad \lambda \neq 0,$$

$$\frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad \lambda = 0. \tag{4}$$

A test of the extreme value model is obtained by testing $\lambda = 1$. Eq. (4) also includes the normal distribution as the special case $\lambda = 0$, and for $\lambda \neq 0$ it provides asymmetric alternatives to normality. There are many other families of distributions that include the normal distribution as a special case. For example, to check symmetric long- or short-tailed departures from normality, one can use the exponential power distribution with pdf

$$f(x) = \frac{k(\delta)}{\sigma}\exp\left(-\frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{\delta}\right), \tag{5}$$

where $k(\delta)$ is a normalizing factor. The normal distribution is recovered under $\delta = 2$. The generalized ***F* distribution** [24] incorporates all the above parametric models as special cases, although since some of these occur when parameters lie on the boundary of the parameter space its practical utility is limited.

More General Tests

Our suggestion is to consider any of the well known distance measures generalized to accommodate an independent right censoring mechanism. Even large sample results are complicated and difficult to obtain, although some specific cases have been worked out for particular censoring mechanisms [28, 29].

A general approach could be based on resampling techniques, the only difficulty here being to insure that resampling is carried out under the null hypothesis that the data are generated from F. We describe this via the Cramér–von Mises test, although the arguments apply equally well to a Kolmogorov–Smirnov or Anderson–Darling test.

Consider, then, the test statistic D^2, where

$$D^2 = \int \{F_u(t) - F_\beta(t)\}^2 \, \mathrm{d}F_\beta(t),$$

which calculates a distance between the observed empirical Kaplan–Meier estimate, F_{u}, and that obtained from fitting the model with parameters β, F_β. Our problem is to obtain the null distribution of D^2. In order to accomplish this we need to **simulate** under F_β. We also need to incorporate the censoring

mechanism C, for which we have a consistent estimate from its observed (censored by the failure times) distribution. Our simulated observations will be $X_i = \min(T_i, C_i)$, $i = 1, \ldots, n$, for which we refit the model and calculate a new value of D^2.

Proportional Hazards Regression Models

The Cox [12] proportional hazards **regression model** is given by

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) \exp\{\boldsymbol{\beta}'\mathbf{Z}(t)\}, \tag{6}$$

or, equivalently,

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{\beta}'\mathbf{Z}_i(t)\}, \quad i = 1, \ldots, n, \tag{7}$$

where $\lambda_0(t)$ is a fixed "baseline" hazard function, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of log relative risk parameters. Many tests have been proposed to check the model assumption of (6), some of them making use of **counting processes** and martingale notation. So let

$$N_i(t) = I\{T_i \le t, T_i \le C_i\}. \tag{8}$$

$N_i(\cdot)$ has the intensity process

$$\alpha_i(t)\,\mathrm{d}t = Y_i(t)\lambda_i(t)\,\mathrm{d}t, \tag{9}$$

and

$$M_i(t) = N_i(t) - \int_0^t \alpha_i(s)\,\mathrm{d}s \tag{10}$$

is a martingale.

Define

$$\pi_i(t;\boldsymbol{\beta}) = \frac{Y_i(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i(t)\}}{\sum_{j=1}^{n} Y_j(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_j(t)\}}, \tag{11}$$

and

$$\mathrm{E}(\mathbf{Z}|t;\boldsymbol{\beta}) = \sum_{j=1}^{n} \mathbf{Z}_j(t)\pi_j(t;\boldsymbol{\beta}); \tag{12}$$

that is, $\mathrm{E}(\cdot|t;\boldsymbol{\beta})$ denotes an expectation taken with respect to the discrete probability distribution $\{\pi_i(t;\boldsymbol{\beta})\}_i$. Also, the following notation is sometimes used:

$$S^{(\tau)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i(t)\}\mathbf{Z}_i(t)^{\otimes r}, \tag{13}$$

for $r = 0, 1, 2$, where for a column vector $\mathbf{a}$, $\mathbf{a}^{\otimes 2}$ refers to the matrix $\mathbf{a}\mathbf{a}'$, $\mathbf{a}^{\otimes 1}$ refers to the vector $\mathbf{a}$, and $\mathbf{a}^{\otimes 0}$ refers to the scalar 1. Then

$$\mathrm{E}(\mathbf{Z}|t;\boldsymbol{\beta}) = \frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}. \tag{14}$$

Finally, define

$$\mathbf{V}(\boldsymbol{\beta}, t) = \frac{S^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - \left\{\frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}\right\}^{\otimes 2}, \tag{15}$$

which is the **covariance matrix** of $\mathbf{Z}$ taken with respect to the discrete probability distribution $\{\pi_i(t;\beta)\}_i$.

Graphical Tests

Kay [25] suggested a test using a graphical procedure to check the assumption that a time-invariant covariate to be included in the Cox model affects the hazard in a multiplicative way. Suppose that the kth covariate Z_k is under consideration, $k = 1, \ldots, p$. Denote $\tilde{\mathbf{Z}}_i$ and $\tilde{\boldsymbol{\beta}}$ as $\mathbf{Z}_i$ and $\boldsymbol{\beta}$ with Z_{ik} and β_k omitted. For a binary Z_k, (7) is equivalent to

$$\lambda_i(t) = \begin{cases} \lambda_0(t)\exp(\beta_k)\exp\{\tilde{\boldsymbol{\beta}}'\tilde{\mathbf{Z}}_i(t)\}, & Z_{ik} = 1, \\ \lambda_0(t)\exp\{\tilde{\boldsymbol{\beta}}'\tilde{\mathbf{Z}}_i(t)\}, & Z_{ik} = 0. \end{cases}$$

The assumption that z_k acts on the hazard function in a multiplicative way can be tested by fitting the Cox model with two strata according to the values of z_k; that is,

$$\lambda_i(t) = \begin{cases} \lambda_{01}(t)\exp\{\tilde{\boldsymbol{\beta}}'\tilde{\mathbf{Z}}_i(t)\}, & Z_{ik} = 1, \\ \lambda_{02}(t)\exp\{\tilde{\boldsymbol{\beta}}'\tilde{\mathbf{Z}}_i(t)\}, & Z_{ik} = 0, \end{cases}$$

and estimating $\lambda_{01}(\cdot)$ and $\lambda_{02}(\cdot)$. Under the proportional hazards assumption, the log cumulative hazard functions, $\log \Lambda_{0j}(u)\,\mathrm{d}u$, $j = 1, 2$, should be parallel. Therefore it was suggested to plot $\log \hat{\Lambda}_{0j}(u)\,\mathrm{d}u$, $j = 1, 2$, against t, and constant differences should result. The procedure extends in an obvious way to discrete variables taking more than two values, and appropriate groupings allow similar techniques for continuous variables [4].

Andersen [1] took Kay's approach further in developing a formal test which could then be associated with this graphical output. This test was based on a piecewise model and some asymptotic approximations. In Figure 1 we illustrate such a plot for the

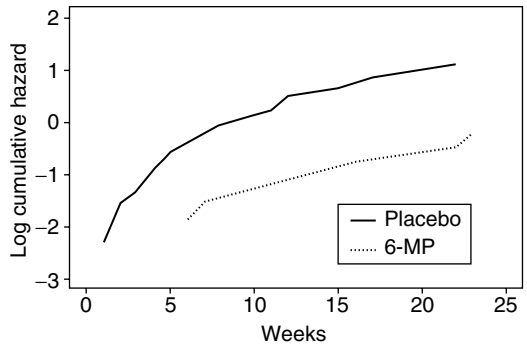

Figure 1 Log cumulative hazard plot for Freireich data

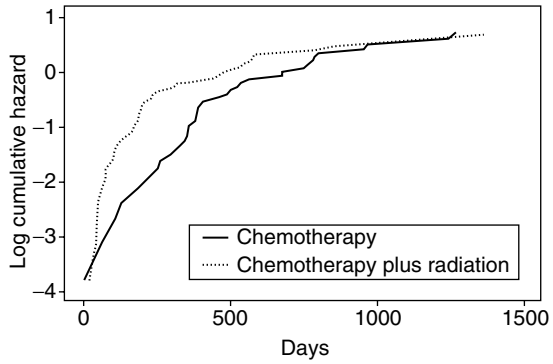

Figure 2 Log cumulative hazard plot for Stablein data

Freireich data. It is clear that the plot provides little evidence against the proportional hazards assumption. In contrast, these plots applied to the Stablein data look very different (Figure 2) and, even without any formal test, the proportional hazards assumption for these data appears doubtful. A number of tests have been proposed that use the unit exponentiality property. Kay [25] applied the residuals of Cox & Snell in order to obtain hazard-based residuals (assuming time-invariant covariates).

Define

$$\varepsilon_i = \exp(\boldsymbol{\beta}'\mathbf{Z}_i)\int_0^{T_i} \lambda_0(u)\,\mathrm{d}u, \quad i = 1, \ldots, n, \quad (16)$$

and their sample-based estimates

$$\hat{\varepsilon}_i = \exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i)\hat{\Lambda}_0(T_i), \quad i = 1, \ldots, n. \quad (17)$$

where $\hat{\boldsymbol{\beta}}$ is usually the maximum **partial likelihood** estimate, and

$$\hat{\Lambda}_0(t) = \sum_{T_k \le t} \frac{\delta_k}{\sum_1^n Y_j(T_k)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_j(T_k)\}}. \quad (18)$$

Under model (7) these quantities should exhibit approximately the properties of a random sample with independent right-censoring from a unit exponential distribution. This has survival function $S(\varepsilon) = \mathrm{e}^{-\varepsilon}$. A plot of an estimated log survival function of $\varepsilon_1, \ldots, \varepsilon_n$ provides a simple check of the model assumptions. However, Baltazar-Aban & Peña [7] pointed out that the critical assumption of approximate unit exponentiality of the residual vector will often not be viable. Their analytical and Monte Carlo results show that the model diagnostic procedures thus considered can have serious defects when the failure-time distribution is not exponential or when the residuals are obtained nonparametrically in the no-covariate model or semiparametrically in the Cox proportional hazards model. The difficulties stem from the complicated correlation structure arising through the estimation process of both the regression coefficients and the underlying cumulative hazard. It has also been argued that, even under quite large departures from the model, this approach may lack sensitivity.

Schoenfeld [47] defined the residuals $\{\mathbf{r}_i(\boldsymbol{\beta})\}$ as the discrepancy between the observed covariate value at time point X_i and its expectation over the risk set under the model:

$$\mathbf{r}_i(\boldsymbol{\beta}) = \mathbf{Z}_i(X_i) - \mathrm{E}(\mathbf{Z}|X_i; \boldsymbol{\beta}) \quad (19)$$

(*see* **Residuals for Survival Analysis**). For $\boldsymbol{\beta}$ fixed and known, these residuals are uncorrelated [13]. Schoenfeld [47] showed that $\{\mathbf{r}_i(\hat{\boldsymbol{\beta}})\}$ are asymptotically uncorrelated and $\mathrm{E}(\mathbf{r}_i(\hat{\boldsymbol{\beta}})) \approx 0$ under the proportional hazards model. Thus a plot of $r_{ik}(\hat{\boldsymbol{\beta}})$ vs. X_i should be centered about zero. However, if

$$\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i(t) + g(t)Z_{ik}\},$$

with $g(t)$ varying about 0, it can be shown that

$$\mathrm{E}(r_{ik}(\hat{\boldsymbol{\beta}})) \approx g(X_i)\{\mathrm{E}(Z_k^2|X_i; \boldsymbol{\beta}) - \mathrm{E}(Z_k|X_i; \boldsymbol{\beta})^2\}.$$

Since the term in the brackets is positive, the changes in $g(\cdot)$ will be reflected in a plot of $r_{ik}(\hat{\boldsymbol{\beta}})$ vs. X_i.

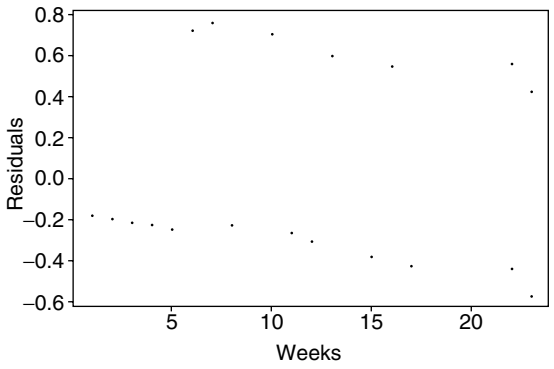

Figure 3 Schoenfeld residuals for Freireich data

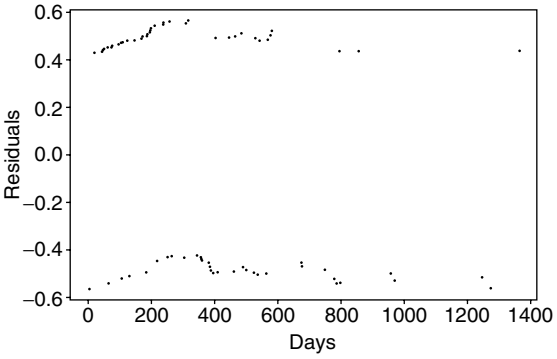

Figure 4 Schoenfeld residuals for Stablein data

Figures 3 and 4, illustrating a simple plot of the Schoenfeld residuals against the failure times, are not easily interpreted. Unlike Figures 1 and 2, as well as Figures 5 and 6 below, these direct residual plots are relatively insensitive to model departures. In practice, it is more instructive to examine the cumulative residuals.

Arjas [6] suggested a graphical method for testing the goodness of fit in Cox's regression model based on the martingale property of

$$M_i'(t;\beta) = N_i(t) - \int_0^t \pi_i(s;\boldsymbol{\beta})\,\mathrm{d}\overline{N}(s), \qquad (20)$$

where $\overline{N}(t) = \sum_1^n N_i(t)$. Suppose that $T_{(1)} < T_{(2)} < \cdots < T_{(K)}$ are the failure times. Let

$$\mathcal{M}_\rangle(\|;\boldsymbol{\beta}) = M_i'(T_{(k)};\boldsymbol{\beta}) = N_i(T_{(k)}) - \sum_{j\le k} \pi_i(T_{(j)};\boldsymbol{\beta}) \qquad (21)$$

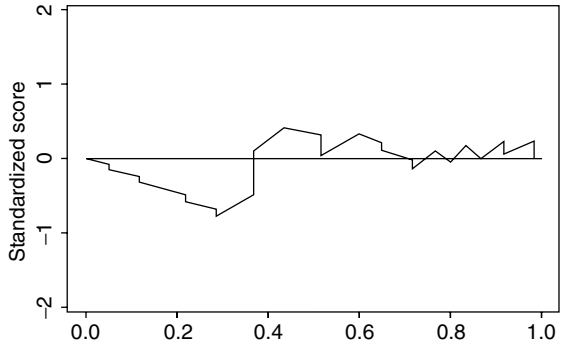

Figure 5 Standardized score process for Freireich data

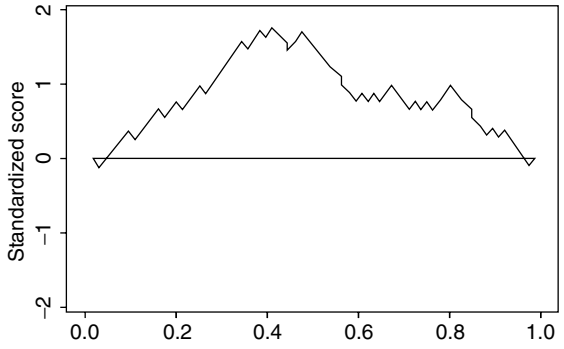

Figure 6 Standardized score process for Stablein data

be an "imbedded" discrete time martingale. Write

$$H(k;\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j\le k} \pi_i(T_{(j)};\boldsymbol{\beta}). \qquad (22)$$

Then the martingale property of

$$\overline{\mathcal{M}}(\|;\boldsymbol{\beta}) = \sum_{\rangle=\infty}^{\setminus} \mathcal{M}_\rangle(\|;\boldsymbol{\beta}) = \| - \mathcal{H}(\|;\boldsymbol{\beta}) \qquad (23)$$

reflects the collective balance between the actual failures and a corresponding cumulative hazard. After substituting the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, a plot of $H(k;\hat{\boldsymbol{\beta}})$ vs. k can be compared with the diagonal line $y = x$ to check the model assumption. A similar procedure can be applied to stratified data, where the sum in (22) and (23) is within a stratum and there will be one graph for each stratum. This approach relates closely to total time on test and other martingale-based residuals described by Andersen et al. [4].

Some of these techniques are summarized by Barlow & Prentice [8] and Therneau et al. [50], where they define a martingale-based residual for the ith subject as

$$e_i(\mathbf{f}_i) = \int_0^{t_0} \mathbf{f}_i(t)\,\mathrm{d}N_i(t) - \int_0^{t_0} \mathbf{f}_i(t)\alpha_i(t)\,\mathrm{d}t = \int_0^{t_0} \mathbf{f}_i(t)\,\mathrm{d}M_i(t). \quad (24)$$

Here t_0 is the maximum follow-up time for the sample, and $\mathbf{f}_i(\cdot)$ is a predictable process with $\mathbf{f}_i(t)$ defined in terms of data on the ith, and possibly other, subjects prior to time t. The estimated residual corresponding to (24) can then be written

$$\hat{e}_i(\hat{\mathbf{f}}_i) = \int_0^{t_0} \hat{\mathbf{f}}_i(t)\,\mathrm{d}N_i(t) - \int_0^{t_0} \hat{\mathbf{f}}_i(t)\,\mathrm{d}\hat{\Lambda}_i(t) = \int_0^{t_0} \hat{\mathbf{f}}_i(t)\{\,\mathrm{d}N_i(t) - \pi_i(t;\hat{\boldsymbol{\beta}})\,\mathrm{d}\overline{N}(t)\}. \quad (25)$$

It can be shown that asymptotically $\hat{e}_i(\hat{\mathbf{f}}_i)$ has mean zero, is uncorrelated with $\hat{e}_j(\hat{\mathbf{f}}_j)$ for $j \neq i$, and has variance estimated by

$$\frac{1}{n}\int_0^{t_0} \hat{\mathbf{f}}_i(t)^{\otimes 2}\pi_i(t;\hat{\boldsymbol{\beta}})\,\mathrm{d}\overline{N}(t). \quad (26)$$

Some special choices of $\mathbf{f}$ have been considered. If $\mathbf{f}_i(\cdot) = 1$, (25) is seen to be equivalent to the hazard-based residuals of Kay [25] when the covariates are time-invariant. $\mathbf{f}_i(t) = \mathbf{Z}_i(t)$ is also of interest, while $\mathbf{f}_i(t) = \mathbf{Z}_i(t) - \mathrm{E}(\mathbf{Z}|t;\boldsymbol{\beta})$ gives Schoenfeld [48] residuals. Henderson & Milner [22] point out that, for some particular choices of $\mathbf{f}$, plots of these residuals against time can exhibit systematic patterns even when the model is appropriate. Their suggestion is to superimpose estimates of expected mean or to standardize the residuals when plotting. For the Schoenfeld residuals, the ones we will most likely use in practice, the difficulty is not present unless the sample size is particularly small.

Lin et al. [32] propose procedures derived from cumulative sums of martingale-based residuals. The distribution of these stochastic processes under the proportional hazards model can be approximated by zero-mean Gaussian processes. They then compare the observed process with a number of simulated realizations from the approximate null distribution. Specifically, define the martingale residuals (see also (10) and (20))

$$\hat{M}_i(t) = M_i'(t;\hat{\boldsymbol{\beta}}) = N_i(t) - \int_0^t Y_i(s) \times \exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(s)\}\,\mathrm{d}\hat{\Lambda}_0(s), \quad (27)$$

where $\hat{\Lambda}_0(\cdot)$ is defined in (18). Assuming time-invariant covariates, define the following two classes of **stochastic processes**:

$$W_{\mathbf{z}}(t,\mathbf{z}) = \sum_{i=1}^n f(\mathbf{Z}_i)I(\mathbf{Z}_i \leq \mathbf{z})\hat{M}_i(t), \quad (28)$$

$$W_r(t,r) = \sum_{i=1}^n f(\mathbf{Z}_i)I(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i \leq r)\hat{M}_i(t), \quad (29)$$

where $f(\cdot)$ is a known smooth function, $\mathbf{z} = (z_1,\ldots,z_p)'$, and the event $\{\mathbf{Z}_i \leq \mathbf{z}\}$ means that all the p components of $\mathbf{Z}_i$ are no larger than the respective components of $\mathbf{z}$. It can be shown that under model (7) $n^{-1/2}W_{\mathbf{z}}(t,\mathbf{z})$ and $n^{-1/2}W_r(t,r)$ converge to zero-mean Gaussian processes. It is also shown that in large samples their null distributions can be approximated through simulations, where one repeatedly generates normal random samples while holding the observed data $\{X_i,\delta_i,\mathbf{Z}_i\}$ fixed [32]. Different specifications of (28) or (29) can be used to check different aspects of the model assumption. For example, $W_{\mathbf{z}}(\infty,\mathbf{z})$ with $f(\cdot) = 1$ and $z_k = \infty (k \neq j)$ provides a check of the functional form of the jth covariate, $j = 1,\ldots,p$; $W_r(\infty,r)$ with $f(\cdot) = 1$ checks the link function of the model; $W_{\mathbf{z}}(t,\mathbf{z})$ with $f(\mathbf{x}) = \mathbf{x}$ and $\mathbf{z} = \infty$ is the well-known score process, and checks the proportional hazards assumption; finally, $W_{\mathbf{z}}(t,\mathbf{z})$ with $f(\cdot) = 1$ can be viewed as an omnibus test when allowing t and $\mathbf{z}$ to vary. In all the cases, one can examine the fit visually by plotting the observed process along with a number of simulated ones. Furthermore, the graphical display may be supplemented with an estimated ***P* value** based on the distribution of the supremum of the process.

Omnibus Tests

Based on the Schoenfeld residuals, Harrell's [19] z-test computes the **correlation** ρ of the Schoenfeld

residuals and the **ranks** of survival times. Then the approximate z-statistic for nonzero correlation,

$$\frac{(n-3)^{1/2}}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \tag{30}$$

is treated as **standard normal**. The test is included in certain **software** packages (*see* **Survival Analysis, Software**). It has a disadvantage that the type I error rates tend to be larger than the nominal levels [40] and that power will not be strong for alternatives other than trend in regression effect (*see* **Level of a Test**). It has the advantage, outweighing the disadvantages in our view, of being simple to construct. Also, the most common alternatives of interest are likely to be trend alternatives (*see* **Ordered Alternatives**). For the Freireich data, we find $\rho = 0.049$ having an associated $z = 0.31$, which is not significant. For the Stablein data, we find $\rho = -0.35$ having an associated $z = -3.39$, which is significant at the 1% level.

Schoenfeld [47] proposed a class of omnibus **chi-square tests** based a partition of the time–covariate space. Let $J_1, \ldots, J_L$ be a partition of the $(p+1)$-dimensional $T \times \mathbf{Z}$ space. Let $I_t\{Z(t), t\}$ be the indicator function of J_1; then

$$f_l = \sum_{i \in D} I_l\{Z_i(T_i), T_i\} \tag{31}$$

is the observed number of failures that fall into J_l, where D is the set of individuals observed to fail, $l = 1, \ldots, L$. Denote $\mathbf{f} = (f_1, \ldots, f_L)'$. The "conditional" mean and variance–covariance matrix of $\mathbf{f}$ with respect to $\{\pi_i(\cdot; \boldsymbol{\beta})\}_i$ can be written down explicitly, and these are denoted by $\mathbf{e}$ and $\mathbf{V}$, respectively. On the basis of these observed and expected quantities, Schoenfeld constructed a chi-square type test, demonstrating the validity of the asymptotic null distribution as well as that for the noncentral chi-square under particular departures. Moreau et al. [35] showed that the Schoenfeld test could be derived as a score test for a broader proportional hazards model incorporating heterogeneity with respect to the regression effect together with a null hypothesis that the heterogeneity is zero. For the Freireich data and a simple division of the time axis into two intervals (less than or greater than 11 weeks) the test statistic, as well as a conservative approximation, were very far from being significant. For the Stablein data a division of the time axis into four intervals, containing approximately equal numbers of failures, resulted in a test statistic significant at the 2% level.

For the construction of the partition $\{J_l\}$, the following has been suggested. Divide the time axis into L_1 intervals that contain approximately the same number of observations. If $\mathbf{Z}$ is discrete, one may use each value as a partition. If $\mathbf{Z}$ has many values or is continuous, one can partition the range of $\mathbf{Z}$ by choosing a partition of the range of $\hat{\boldsymbol{\beta}}'\mathbf{Z}$. Assuming that the range of $\mathbf{Z}$ has been divided into L_2 sets $\{S\}$, one can define $\{J_l\}$ to be the Cartesian product of the above two partitions. Now suppose that the proportional hazards assumption does not hold, and that for a certain interval (τ_i, τ_{i+1}) on the time axis the effect of a covariate is greater than on other intervals. Suppose, furthermore, that when this component of $\mathbf{Z}$ has a high value in (τ_i, τ_{i+1}) the hazard is greater than if it is high on other intervals, and that $\mathbf{Z} \in S_j$ whenever this component is high. Then the partition formed by $(\tau_i, \tau_{i+1}) \times S_j$ will have more than the expected number of failures. On the other hand, if the hazard does not depend on $\mathbf{Z}$ in a loglinear manner, then, on each S_j, the expected and observed number of failures will not agree very well, and this pattern will repeat itself for each time interval. See the original paper for simplifications in computing the test statistic, verification of a technical requirement in order for the asymptotic result to hold, and calculation of the noncentrality parameter under the alternative hypothesis.

Mckeague & Utikal [34] use the doubly cumulative hazard function

$$A(t, z) = \int_0^z \int_0^l \lambda(s|x)\,\mathrm{d}s\;\mathrm{d}x$$

to test the goodness of fit of the Cox model when there is only one covariate to be considered. Their method compares two different estimates of $A(\cdot, \cdot)$. After stratifying over the covariate, they obtain a fully nonparametric estimator

$$\tilde{A}(t, z) = \int_0^z \tilde{\Lambda}(t|x)\,\mathrm{d}x, \tag{32}$$

where $\overline{\Lambda}(t|z)$ is the Nelson–Aalen estimator of the cumulative hazard function in one of the strata. Meanwhile, under the model, $A(t, z)$ can be estimated by

$$\hat{A}(t, z) = \hat{\Lambda}_0(t) \int_0^z \exp(\hat{\beta}x)\,\mathrm{d}x, \tag{33}$$

where $\hat{\Lambda}_0(t)$ is given in (18). Then under the model $\sqrt{n}(\tilde{A} - \hat{A})$ converges weakly to a Gaussian random field. So Kolmogorov–Smirnov type or Cramér–von Mises type test statistics, or resampling methods, may be applied. Alternatively, Mckeague & Utikal [34] developed a method that follows the above Schoenfeld [47] approach. The same idea is adapted to test Cox's model within general proportional hazards models.

Horowitz & Neumann [23] described a generalized moments test which does not require assigning data to predetermined cells. Their test is based on the unit exponentiality of (16), and the asymptotic results were studied. Let $g(\varepsilon, \delta, \mathbf{Z})$ be a vector-valued function with the property that $\mathrm{E}\{\mathbf{g}(\varepsilon, \delta, \mathbf{Z})\} = 0$ if ε has (possibly right-censored) unit exponential distribution, and does not equal to zero if not. Here δ is the censorship indicator of ε. Examples of $\mathbf{g}$ are $(1+\delta)\exp(-\varepsilon) - 1$, $\mathbf{Z}\{(1+\delta)\exp(-\varepsilon) - 1\}$, and $\varepsilon^2 - \varepsilon\delta$, although the first one was found to be the best in terms of finite sample size and power properties. The test is based on

$$\Omega_n = n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\hat{\varepsilon}_i, \delta_i, \mathbf{Z}_i). \tag{34}$$

Horowitz & Neumann [23] showed that under model (6) and regularity conditions, Ω_n is asymptotically **multivariate normal** as $n \to \infty$ with mean zero and covariance matrix that can be estimated consistently (see their appendix for details). As with all such tests, this generalized moments test does not indicate the sources of error in a rejected model, and may have little power against certain alternatives. However, it was shown to perform well against **accelerated failure time** models. A small-sample correction has also been derived.

Another omnibus test, not requiring arbitrary division of the time axis, was presented by Lin & Wei [31] Their idea, developed from White [53], was to contrast the observed information matrix and the squared score matrix. These matrices are both consistent for the inverse of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ under the model but are various ways of combining the elements of this matrix. Lin & Wej [31] made two suggestions: the first to take the largest element in absolute value and the second to use a Wald-type statistic (*see* **Likelihood**). The null distribution of the largest element was approximated via simulation, whereas an expression was given for the covariance matrix needed in the Wald statistic. The test has the advantage of good power for general alternatives and does not hinge upon arbitrary choices such as the time division. Its disadvantage is that the calculations necessary to obtain a significance level are relatively involved.

Marzec & Marzec [33] developed the goodness of fit inference based on Arjas's [6] graphical method. Let $I \subset \{1, \ldots, n\}$ be a given stratum. Similar to (13), define

$$S_I^{(r)}(\boldsymbol{\beta}, t) = |I|^{-1} \sum_{i \in I} Y_i(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_i(t)\} \mathbf{Z}_i(t)^{\otimes r}, \tag{35}$$

for $r = 0, 1, 2$, where $|I|$ means the size of I. Let $N_I(t) = \sum_I N_i(t)$. Define

$$M_I(t; \boldsymbol{\beta}) = N_I(t) - \frac{|I|}{n} \int_0^t \frac{S_I^{(0)}(\boldsymbol{\beta}, s)}{S^{(0)}(\boldsymbol{\beta}, s)} \, \mathrm{d}\overline{N}(s). \tag{36}$$

Under the model, $M_I(t; \boldsymbol{\beta})$ describes a collective balance in the stratum between the actual failures and a corresponding cumulative hazard. Assume that $|I|/n \to q$ as $n \to \infty, q \in (0, 1)$. Marzec & Marzec [33] showed that, depending on two different sets of conditions on $S_I^{(r)}(\boldsymbol{\beta}, t)$, $n^{-1/2} M_I(t; \hat{\boldsymbol{\beta}})$ converges weakly to a time transformed **Brownian motion**, or a more complex Gaussian process. In the first case,

$$\frac{\sup_{0<t<1} |M_I(t; \hat{\boldsymbol{\beta}})|}{\left\{ \int_0^1 \tau_n(\hat{\boldsymbol{\beta}}, s) \, \mathrm{d}\overline{N}(s) \right\}^{1/2}}, \tag{37}$$

where

$$\tau_n(\boldsymbol{\beta}, t) = \frac{|I| S_I^{(0)}(\boldsymbol{\beta}, t)}{n S^{(0)}(\boldsymbol{\beta}, t)} \left\{ 1 - \frac{|I|}{n} \frac{S_I^{(0)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} \right\},$$

converges weakly to $\sup_{0 \le t \le 1} |W(t)|$, where W is the Brownian motion. In the second case, the same test statistic (37) converges weakly to $\sup_{0 \le t \le 1} |W^0(t)|$, where W^0 is the Brownian bridge.

For a two-sample problem, the proportional hazards model can be simply written as $S_1(t) = S_2(t)^\theta$, where $S_i(t)$ is the survival function of group $i, i = 1, 2$. This is equivalent to (6) with $\theta = \mathrm{e}^\beta$. For this model, Wei [52] suggested an omnibus test based on the score process, which has also been used by others, including Barlow & Prentice [8] and Lin

et al. [32]. Because of the simplicity of this particular case, one can easily write down the explicit form of the statistic. Instead of the earlier notation that we have been using, denote (X_{ij}, δ_{ij}), $j = 1, \ldots, n_i$, the observations from sample i, and $Y_i(t) = \sum_j J(X_{ij} \geq t)$, $i = 1, 2$. Then the score (i.e. the derivative of the log partial likelihood) process can be seen to be $\theta^{-1}U_n(t; 0)$, where

$$U_n(t;\theta) = \sum_{j=1}^{n_i} \delta_{1j} I(X_{1j} \leq t) - \sum_{i=1}^{2}\sum_{j=1}^{n_i} \frac{Y_1(X_{ij})\theta}{Y_1(X_{ij})\theta + Y_2(X_{ij})} I(X_{ij} \leq t). \tag{38}$$

The test statistic is based on the supremum of the absolute value of the process (38), which has the interpretation as the observed number of failures from sample 1 minus the corresponding expected number of failures before or at time t. Replace θ by $\hat{\theta}$, the maximum partial likelihood estimate, and let $W_n(t) = n^{-1/2}U_n(t; \hat{\theta})$. Assuming that $n_i/n \to \rho_i, 0 < \rho_i < 1$, it can be shown that under the model $\{W_n(t) : 0 \leq t \leq \infty\}$ converges in law to

$$\{[\theta_\eta(\infty)]^{1/2} W^0\left[\frac{\eta(t)}{\eta(\infty)}\right] : 0 \leq t \leq \infty\},$$

where W^0 is the Brownian bridge, and $\eta(\infty)$ can be consistently estimated by

$$\hat{\eta}(\infty) = n^{-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i} \frac{\delta_{ij} Y_1(X_{ij}) Y_2(X_{ij})}{\{Y_1(X_{ij})\hat{\theta} + Y_2(X_{ij})\}^2}.$$

Note that $\eta(\infty)$ in the above was written $\eta(t)$ in the original paper, and that this seems to be a typographical error. Therefore the goodness of fit test statistic is

$$\{\hat{\theta}_{\hat{\eta}}(\infty)\}^{-1/2} \sup_{0<t<\infty} |W_n(t)|. \tag{39}$$

Wei [52] showed that (39) is consistent against the alternative of nonproportional hazards. A table of percentage points of $\sup_{0\leq s\leq 1} |W^0(s)|$ can be found in Koziol & Byar [28]. Alternatively, we can use the fact that

$$\Pr(\sup_t |W^0(t)| > \alpha) = 2\sum_{k=1}^{\infty}(-1)^{k+1} \times \exp(-2k^2\alpha^2), \alpha \geq 0, \tag{40}$$

a well known result in probability. It can be seen that absolute values of the standardized score process greater than around 1.4 can be considered significant at the 0.05 level. For the Freireich data the standardized score process (Figure 5) lies well within this limit. In contrast, this process for the Stablein data (Figure 6) has a maximum around 1.8, which is significant at 1%.

Tests for Specific Alternatives

Many of the tests against the alternative of time-varying regression effects can be summarized under the following model (for simplicity of notation, we assume parameters of dimension one):

$$\lambda_i(t) = \lambda_0(t)\exp\{[\beta + \alpha Q(t)]Z_i(t)\}, \tag{41}$$

where $Q(t)$ is a function of time which does not depend on the parameters β and α. Under $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ we recover the proportional hazards model (7). In the context of sequential group comparisons of survival data, this model has been considered by Tsiatis [51] and Harrington et al. [20]. Like (11) and (12), we denote by $E(\cdot|t; \beta, \alpha)$ the expectation taken with respect to the probability distribution $\{\pi_i(t; \beta, \alpha)\}_i$, where

$$\pi_i(t;\beta,\alpha) = \frac{Y_i(t)\exp\{[\beta + \alpha Q(t)]Z_i(t)\}}{\sum_{j=1}^{n} Y_j(t)\exp\{[\beta + \alpha Q(t)]Z_j(t)\}}. \tag{42}$$

Assume that $Q(t)$ is known: then the score vector $\mathbf{U}(\beta, \alpha)$ for model (41) has two components,

$$U_\beta(\beta,\alpha) = \sum_{i=1}^{n} \delta_i\{Z_i(X_i) - E(Z|X_i;\beta,\alpha)\} \tag{43}$$

and

$$U_\alpha(\beta,\alpha) = \sum_{i=1}^{n} \delta_i Q(X_i)\{Z_i(X_i) - E(Z|X_i;\beta,\alpha)\}; \tag{44}$$

while the **information matrix** is

$$\mathbf{I}(\beta,\alpha) = -\begin{pmatrix} U_{\beta\beta} & U_{\beta\alpha} \\ U_{\alpha\beta} & U_{\alpha\alpha} \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \tag{45}$$

where

$$I_{kl}(\beta,\alpha)=\sum_{i=1}^{n}\delta_i Q(X_i)^{k+1-2}\{\mathrm{E}(Z^2|X_i;\beta,\alpha) - \mathrm{E}(Z|X_i;\beta,\alpha)^2\},\quad k,l=1,2. \quad (46)$$

Let $\hat{\beta}$ be the maximum partial likelihood estimate of β under H_0: then $U_\beta(\hat{\beta},0)=0$. So the score test statistic arising under H_0 is

$$B=U_\alpha(\hat{\beta},0)G^{-1}U_\alpha(\hat{\beta},0), \quad (47)$$

where $G=I_{22}-I_{21}I_{11}^{-1}I_{12}$ and G^{-1} is the lower right corner element of $\mathbf{I}^{-1}$. Under H_0, S has an asymptotically χ^2 distribution with one degree of freedom.

As special cases of (41), Cox [12] considered $Q(t)=t$, Stablein et al. [49] considered $Q(t)=(t,t^2)'$, and Brown [11], Anderson & Senthilselvan [5], O'Quigley & Moreau [42], Moreau et al. [36], and O'Quigley & Pessione [43] assumed $Q(t)$ to be constant on predetermined intervals of the time axis; in other words $Q(t)$ is a step function. Although in the latter cases there is more than one parameter associated with $Q(t)$, the computation of the test statistic is similar to the above. Murphy [37] studied the size and the power of Moreau et al. [36] and found that, although it is consistent against a wide class of alternatives to proportional hazards, it is nonetheless an omnibus test which should be used when there is no specific alternative in mind.

Sometimes $Q(t)$ is chosen to be an unknown function and needs to be estimated; for example, $Q(t)=\Lambda(t)$ [10] and is estimated by the Nelson estimator. Because the estimates at time t depend only on the σ-field of events up to that time, the development of Cox [13] and Andersen & Gill [3] and thus, the above asymptotic theory, still applies. Breslow et al. [10] showed that their choice $Q(t)=\Lambda(t)$ has good power against the alternative of crossing hazards. Tsiatis [51], Harrington et al. [20], and Harrington & Fleming [21] used score processes based on (44) for **sequential** tests. They showed that after $Q(t)$ has been replaced by its estimate, the score process at different time points converge in distribution to multivariate normal. Their particular interest lies in the G^ρ family, where $Q(t)=S(t)^\rho$.

Another special case of (41) is found in O'Quigley & Pessione [44] and O'Quigley [40], where $Q(t)=I(t\le\gamma)-I(t>\gamma)$, with γ an unknown change point. Were γ known, the test statistic in (47) could be readily calculated and denote $B(\gamma)^{1/2}=U_\alpha(\hat{\beta},0;\gamma)/G^{1/2}$. For γ unknown, Davies [14, 15] demonstrates that an appropriate test should be based on the supremum $B(\cdot)$. In order to evaluate the significance level of the test statistic, the **autocorrelation function** is required. O'Quigley & Pessione [44] suggested approximating this via **bootstrap** resampling. Alternatively, a simpler approximation provided by Davies appeared to be satisfactory for most applications. O'Quigley & Pessione [44] showed these tests to be powerful for testing the equality of two survival distributions against the specific alternative of crossing hazards. These tests suffer only moderate losses in power, when compared with their optimal counterparts, if the alternative is one of proportional hazards.

Other authors [18, 30] have taken a slightly different starting point and introduced the function $Q(t)$ directly into a weighted score. This can be written

$$U_Q(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i Q(X_i)\{\mathbf{Z}_i(X_i)-\mathrm{E}(\mathbf{Z}|X_i;\boldsymbol{\beta})\}, \quad (48)$$

where $Q(\cdot)$ is a predictable process that converges in probability to a nonnegative bounded function uniformly in t. Let $\hat{\boldsymbol{\beta}}_Q$ be the zero of (48) and let $\hat{\boldsymbol{\beta}}$ be the partial likelihood estimate. Under the assumption that model (6) holds and that $(X_i,\delta_i,\mathbf{Z}_i)(i=1,\ldots,n)$ are iid replicates of $(X,\delta,\mathbf{Z})$, $n^{1/2}(\hat{\boldsymbol{\beta}}_Q-\hat{\boldsymbol{\beta}})$ is asymptotically normal with zero mean and covariance matrix that can be consistently estimated via derivations very close to those of (45). It then follows that a simple test can be based on the standardized difference between the two estimates. Lin [30] showed such a test to be consistent against any model misspecification under which $\boldsymbol{\beta}_Q\neq\boldsymbol{\beta}$, where $\boldsymbol{\beta}_Q$ is the probability limit of $\hat{\boldsymbol{\beta}}_Q$. In particular, it can be shown that choosing a monotone weight function for $Q(t)$ such as $\hat{F}(t)$, where $\hat{F}(\cdot)$ is the Kaplan–Meier estimate, is consistent against monotone departures (e.g. a decreasing regression effect) from the proportional hazards assumption.

A simple test of interaction between a linear combination of covariates $\mathbf{a}'\mathbf{Z}$ and time was developed by Nagelkerke et al. [38]. It uses the Schoenfeld residuals defined in (19). Let $U_{\mathbf{a}_i}(\boldsymbol{\beta})=\mathbf{a}'\mathbf{r}_i(\boldsymbol{\beta})$. Then successive values of $U_{\mathbf{a}_i}(\boldsymbol{\beta})$ are uncorrelated under model (6). Now suppose that Cox's model does

not hold because the impact of $\mathbf{a}'\mathbf{Z}$ on the hazard increases or decreases gradually with time. It can be seen that successive values of $U_{\mathbf{a}_i}(\boldsymbol{\beta})$ are positively correlated. So we can define a test statistic

$$V_{\mathbf{a}}(\hat{\boldsymbol{\beta}}) = \sum_i U_{\mathbf{a}_i}(\hat{\boldsymbol{\beta}}) U_{\mathbf{a}_{i-1}}(\hat{\boldsymbol{\beta}}) \tag{49}$$

and reject the Cox model for large values of $V_{\mathbf{a}}(\hat{\boldsymbol{\beta}})$. Notice that even under model (6), $V_{\mathbf{a}}(\hat{\boldsymbol{\beta}})$ does not have exactly zero expectation. A permutational approach is suggested to estimate the first two moments of the distribution of the test statistic $V_{\mathbf{a}}(\hat{\boldsymbol{\beta}})$. Nagelkerke et al. [38] suggest first testing with $\mathbf{a} = \hat{\boldsymbol{\beta}}$. If the model is rejected, one then goes on to test individual covariate by allowing $\mathbf{a}$ to have only one nonzero element. The test comes under the heading of omnibus tests and has rather weak power. This, together with the fact that computation is somewhat involved, has resulted in the test having seen little practical use. It nonetheless provides important insights into how the score process is affected by departures from the proportional hazards assumption and would seem to be worthy of further investigation.

Let $\lambda_1(t)$ and $\lambda_2(t)$ be the hazard functions for two groups. Gill & Schumacher [18] and Deshpande & Sengupta [16] have developed tests to check the assumption of proportional hazards vs. the alternative that the hazard ratio changes monotonically with time. Under the proportional hazards assumption, the hazard ratio, or the relative risk, $\theta = \lambda_2(t)/\lambda_1(l)$ can be estimated by the generalized rank estimator

$$\hat{\theta}_K = \frac{\int K(t)\, \mathrm{d}\hat{\Lambda}_2(t)}{\int K(t)\, \mathrm{d}\hat{\Lambda}_1(t)} \tag{50}$$

[2, 9], where $K(t)$ is a predictable random weight function, and $\Lambda_j(t)$ is the Nelson–Aalen estimator of the cumulative hazard function in group j. The integrals are over the range $(0, \tau)$, where τ is the upper limit of observable survival times. Gill & Schumacher [18] base their test statistic on the difference between two generalized rank estimators with two different weight functions say, $K_1(t)$ and $K_2(t)$. Specifically, let

$$K_{ij} = \int K_i(t)\, \mathrm{d}\hat{\Lambda}_j(t).$$

Under the proportional hazards assumption

$$D = K_{11}K_{22} - K_{21}K_{12} \tag{51}$$

is asymptotically normal with estimated variance

$$K_{21}K_{22}V_{11} - K_{21}K_{12}V_{12} - K_{11}K_{22}V_{21} + K_{11}K_{12}V_{22}, \tag{52}$$

where

$$V_{ij} = \int K_i(l)K_j(t)\{Y_1(l)Y_2(t)\}^{-1}\, \mathrm{d}\{N_1(t)+N_2(t)\},$$

$N_j(t)$ is the counting process, and $Y_j(t)$ is the number at risk at time l for group j. The test is shown to be consistent against alternatives with a monotone hazard ratio if $K_2(t)/K_1(t)$ is monotone as well, and this is the case for any two of the weight functions common in, for example, generalized linear rank tests. A related graphical method and discussion of the choice of appropriate weight functions in terms of **asymptotic relative efficiency** are also given in Gill & Schumacher [18].

Deshpande & Sengupta [16] use a ***U*-statistic** test to check the assumption that $\lambda_1(l)/\lambda_2(t)$ is equal to a constant versus the alternative that the hazard ratio increases with time. Under the alternative hypothesis, one can verify that

$$S_1(a)f_2(a)S_2(b)f_1(b) \geq S_2(a)f_1(a)S_1(b)f_2(b) \tag{53}$$

for $0 < a < b$, where $S_i(\cdot)$ and $f_i(\cdot)$ are the survival function and the density from group i, $i = 1, 2$. Their suggestion is to integrate the difference in (53) over the range $0 < a < b$ in order to obtain a particular functional, $\Delta(S_1, S_2)$, having the property of being zero under the null hypothesis and positive under the alternative. A U-statistic estimate of $\Delta(S_1, S_2)$ can be obtained by using the frequencies of observations falling into certain regions. The asymptotic distribution of the test statistic is normal and the asymptotic null variance can be estimated by the use of resampling techniques.

References

[1] Andersen, P.K. (1982). Testing goodness of fit of Cox's regression and life model, *Biometrics* **38**, 67–77.

[2] Andersen, P.K. (1983). Comparing survival distributions via hazard ratio estimates, *Scandinavian Journal of Statistics* **10**, 77–85.

[3] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[4] Andersen, P.K., Borgan, Ø., Gill, R. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, Berlin.

[5] Anderson, J. & Senthilselvan, A. (1982). A two-step regression model for hazard functions, *Applied Statistician* **31**, 44–51.

[6] Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model, *Journal of the American Statistical Association* **83**, 204–212.

[7] Baltazar-Aban, I. & Peña, E.A. (1995). Properties of hazard-based residuals and implications in model diagnostics, *Journal of the American Statistical Association* **90**, 185–197.

[8] Barlow, W.E. & Prentice, R.L. (1988). Residuals for relative risk regression, *Biometrika* **75**, 65–74.

[9] Begun, J.M. & Reid, N. (1983). Estimating the relative risk with censored data, *Journal of the American Statistical Association* **78**, 337–341.

[10] Breslow, N.E., Edler, L. & Berger, J. (1984). A two-sample censored-data rank test for acceleration, *Biometrics* **40**, 1049–1062.

[11] Brown, C.C. (1975). On the use of indicator variables for studying the time-dependence of parameters in a response-time model, *Biometrics* **31**, 863–872.

[12] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[13] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[14] Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* **64**, 247–254.

[15] Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* **74**, 33–43.

[16] Deshpande, J.V. & Sengupta, D. (1995). Testing the hypothesis of proportional hazards in two populations, *Biometrika* **28**, 251–261.

[17] Freireich, E.J. (1963). *See* Vol. 1, p. 697, 25.

[18] Gill, R. & Schumacher, M. (1987). A simple test of the proportional hazards assumption, *Biometrika* **74**, 289–300.

[19] Harrell, F.E. (1986). *The PHGLM Procedure, SAS Supplement Library User's Guide*, Version 5. SAS Institute Inc., Cary.

[20] Harrington, D.P., Fleming, T.R. & Green, S.J. (1982). Procedures for serial testing in censored survival data, in *Survival Analysis*, J. Crowley & R. Johnson, eds. Institute of Mathematical Statistics Lecture Notes, Monograph Series, pp. 269–286.

[21] Harrington, D.P. & Fleming, T.R. (1982). A class of rank test procedures for censored survival data, *Biometrika* **69**, 553–566.

[22] Henderson, R. & Milner, A. (1991). On residual plots for relative risk regression, *Biometrika* **78**, 631–636.

[23] Horowitz, J.L. & Neumann, G.R. (1992). A generalized moments specification test of the proportional hazards model, *Journal of the American Statistical Association* **87**, 234–240.

[24] Kalbfleisch, J. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[25] Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data, *Applied Statistics* **26**, 227–237.

[26] Korn, E.L. & Simon, R. (1990). Measures of explained variation for survival data, *Statistics in Medicine* **9**, 487–503.

[27] Korn, E.L. & Simon, R. (1991). Explained residual variation, explained risk, and goodness of fit, *American Statistician* **45**, 201–206.

[28] Koziol, J.A. & Byar, D.A. (1975). Percentage points of the asymptotic distributions of one and two sample K–S statistics for truncated or censored data, *Technometrics* **17**, 507–510.

[29] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley New York.

[30] Lin, D.Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, *Journal of the American Statistical Association* **86**, 725–728.

[31] Lin, D.Y. & Wei, L.J. (1991). Goodness-of-fit tests for the general Cox model, *Journal of the American Statistical Association* **86**, 725–728.

[32] Lin, D.Y., Wei, L.J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals, *Biometrika* **80**, 557–572.

[33] Marzec, L. & Marzec, P. (1993). On goodness of fit inference based on stratification in Cox's regression model, *Scandinavian Journal of Statistics* **20**, 227–238.

[34] Mckeague, I.W. & Utikal, K.J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards models, *Scandinavian Journal of Statistics* **18**, 177–195.

[35] Moreau, T., O'Quigley, J. & Lellouch, J. (1986). Concerning Schoenfeld's chi-squared statistic for testing the proportional hazards assumption, *Biometrika* **73**, 513–515.

[36] Moreau, T., O'Quigley, J. & Mesbah, M. (1985). A global goodness-of-fit statistic for the proportional hazards model, *Applied Statistics* **34**, 212–218.

[37] Murphy, S.A. (1993). Testing for a time dependent coefficient in Cox's regression model, *Scandinavian Journal of Statistics* **20**, 35–50.

[38] Nagelkerke, N.J.D., Oosting, J. & Hart, A.A.M. (1984). A simple test of goodness of fit of Cox's proportional hazards model, *Biometrics* **40**, 483–486.

[39] Nelson, W. (1972). Theory and application of hazard plotting for censored failure data, *Technometrics* **14**, 945–965.

[40] O'Quigley, J. (1994). On a two-sided test for crossing hazards, *Statistician* **43**, 563–569.
[41] O'Quigley, J. & Flandre, P. (1994). Predictive capability of proportional hazard regression, *Proceedings of the National Academy of Sciences* **91**, 2310–2314.
[42] O'Quigley, J. & Moreau, T. (1984). Testing the proportional hazards regression model against some general alternatives, *Revue d'Epidemiologie et de Santé Publique* **4**, 199–205.
[43] O'Quigley, J. & Pessione, F. (1989). Score tests for homogeneity of regression effect in the proportional hazards model, *Biometrics* **45**, 135–144.
[44] O'Quigley, J. & Pessione, F. (1991). The problem of a covariate-time qualitative interaction in a survival study, *Biometrics* **47**, 101–115.
[45] Rice, J. (1995). *Mathematical Statistics and Data Analysis*, 2nd Ed. Duxbury Press, North Scituate.
[46] Schemper, M. (1990). The explained variation in proportional hazards regression, *Biometrika* **77**, 216–218.
[47] Schoenfeld, D. (1980). Chi-squared goodness of fit tests for the proportional hazards regression model, *Biometrika* **67**, 145–153.
[48] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**, 239–241.
[49] Stablein, D.M., Carter, Jr, W.H. & Novak, J.W. (1981). Analysis of survival data with nonproportional hazard functions, *Controlled Clinical Trials* **2**, 149–159.
[50] Therneau, T. Grambsch, P. & Fleming, T. (1990). Martingale-based residuals for survival models, *Biometrika* **77**, 147–160.
[51] Tsiatis, A.A. (1982). Group sequential methods for survival analysis with staggered entry, in *Survival Analysis*, J. Crowley & R. Johnson, eds. Institute of Mathematical Statistics Lecture Notes, Monograph Series, pp. 257–268.
[52] Wei, L.J. (1984). Testing goodness of fit for proportional hazards model with censored observations, *Journal of the American Statistical Association* **79**, 649–652.
[53] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1–25.

(*See also* **Survival Analysis, Overview**; **Survival Distributions and Their Characteristics**)

JOHN O'QUIGLEY & RONGHUI XU

Goodness of Fit

Introduction

Probably the most famous test in statistics is Pearson's **chi-squared** goodness-of-fit test that assesses the agreement between an observed set of frequencies O_k in K classes and the expected numbers E_k in those classes. As an example, Kendall and Stuart [14, Section 30.7], take data from Mendel's classic experiments on pea breeding. Table 1 gives the frequencies of different kinds of seeds in crosses from plants with round yellow seeds and plants with wrinkled green ones. Also given are the theoretical probabilities from the Mendelian theory of inheritance (*see* **Mendel's Laws**).

The statistic is given by

$$X^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} = \sum_{k=1}^{K} \frac{(O_k - n\theta_k)^2}{n\theta_k} = 0.470. \tag{1}$$

With $K = 4$, the statistic is to be compared with the **chi-squared** (χ^2) distribution on $K - 1 = 3$ **degrees of freedom**. The value of X^2 is not significantly large ($\chi^2_{3,0.95} = 7.81$), so the data agree with the theory. However, the value is arguably not so small ($\chi^2_{3,0.05} = 0.352$) to indicate too close agreement between observations and theory, such as might have been caused by the knowing intervention of Mendel's gardener.

This well-known procedure shows many important characteristics of traditional goodness-of-fit tests. Most importantly, it employs many degrees of freedom to test for an unspecified general departure. If a particular departure is of interest, more powerful procedures can be found based on one or a few degrees of freedom. But, of course, such procedures may fail to detect departures other than those for which they are specific.

A second characteristic is that the test is based on aggregate statistics, that is, on quantities calculated over all the data. Attention is not paid to the contribution of individual observations or cases. The third characteristic is that the χ^2 distribution of X^2 only holds asymptotically. As with most asymptotic procedures, attention needs to be paid to the small sample distribution of the statistic. For X^2, this includes concerns about the effect of cells with few observations.

This entry can, with advantage, be read in conjunction with several others. Here we concentrate on tests based on aggregate statistics, simple plots of residuals and tests of distributional shape, either for samples or, more importantly, for residuals. Methods for determining the contribution of individual observations are described in the articles on **Residuals** and **Diagnostics**.

Regression

The Analysis of Variance

In the linear **multiple regression** model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters, and $\boldsymbol{X}$ is the $n \times p$ **matrix** of carriers, that is, of **explanatory variables** and perhaps functions of them, such as quadratics (*see* **Polynomial Regression**) and **interaction** terms. It is assumed that the additive errors of observation $\boldsymbol{\varepsilon}$ are independently distributed with constant variance σ^2. Many tests of goodness of fit of regression models either use the assumed distribution of the errors to provide t or F tests of models (*see* **Student's *t* Statistics**; ***F* Distributions**), or assess the model by the **normality** of the residuals. We give examples of both procedures.

The **least-squares** estimates of the parameters in (2) are

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{3}$$

yielding fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{4}$$

and least-squares residuals

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}. \tag{5}$$

In (5), $\boldsymbol{I}$ is the $n \times n$ identity matrix and, in (4), $\boldsymbol{H}$ is the "hat" matrix, so called because $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$.

Provided the model (2) fits, the residual sum of squares

$$S(\hat{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \tag{6}$$

is distributed as $\sigma^2\chi^2_{n-p}$. If σ^2 is known, the fit of the model can be tested by comparing $S(\hat{\beta})/\sigma^2$ with this

Table 1 Mendel's data on four kinds of pea seeds

Seeds	Observed frequency O_k	Theoretical probability θ_k	Expected frequency $E_k = n\theta_k$
Round and yellow	315	9/16	312.75
Wrinkled and yellow	101	3/16	104.25
Round and green	108	3/16	104.25
Wrinkled and green	32	1/16	34.75
Total	$n = 556$	1	556

chi-squared distribution, large values indicating lack of fit. Usually, σ^2 is not known, but is estimated by

$$s^2 = \frac{S(\hat{\beta})}{n - p}, \tag{7}$$

so that the goodness-of-fit test is not available. If however, an independent estimate of σ^2 is available, s_ν^2 on ν **degrees of freedom**, the ratio $S(\hat{\beta})/\{(n - p)s_\nu^2\}$ can be compared with the F distribution on $n - p$ and ν degrees of freedom.

Often, the value of s_ν^2 is found from replicate observations within the data. The subtraction of the replicate sum of squares from $S(\hat{\beta})$ leaves a lack of fit sum of squares, the mean square being compared with s_ν^2. The resulting test is typical of many goodness-of-fit tests in not being directed against any specific departure. To test specific departures, the model can be embedded in a more general one, yielding t or F tests for the extra terms. In summary, the two procedures are as follows:

- **Embedding**. Add extra terms to the model, typically higher-order polynomials and interactions in the explanatory variables. These yield t tests for individual coefficients or, more conservatively, F tests for groups of terms.
- **Lack of Fit Sum of Squares**. An estimate of σ^2 either from within or outside the experiment is used to test the residual sum of squares. Estimates from outside the data being analyzed should be treated with caution as they are typically too small through overoptimism about the accuracy of measurement.

If exact replicate observations are not available, observations "close" in X space can be grouped to give approximate replicates and so an estimate of σ^2. More frequently, it is assumed that the addition of higher-order terms removes any possibility of further departures from the model, so that $S(\hat{\beta})$ provides an **unbiased** estimate of error. The articles on **Residuals** and **Diagnostics** discuss procedures for detecting individual errant observations, such as **outliers**. The extension to groups of observations is covered in the article on **Forward Search**.

As an example in which there are replicate observations so that both embedding and a lack of fit sum of squares can be used to assess goodness of fit, consider the data on fetal growth given by Francis et al. [9, p. 328]. The data, plotted in Figure 1, show ultrasound measurements of the head to coccyx (htc) length, in millimeters, of the fetuses of 83 Brazilian women between 7 and 16 weeks pregnant. There is appreciable replication in the data as presented as the **gestational age** is rounded to the nearest week. This is then strictly an example in which observations "close" in X space have been grouped. Since there is at least one measurement at each age, there will be a

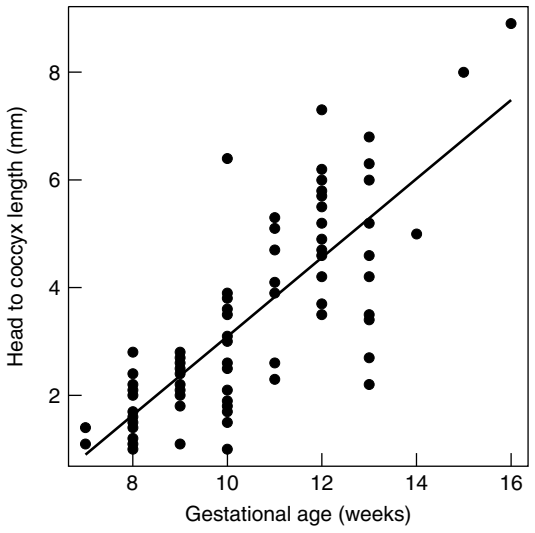

Figure 1 Scatter plot of head to coccyx length data. The replicate observations provide an estimate of pure error and so lead to a test of goodness of fit

maximum of 10 degrees of freedom for the model and 73 degrees of freedom for "pure" error (which will include woman to woman variation). The plot suggests that there is a linear relationship between the response htc and age and this fitted model is shown on the plot. To check the linear model, a quadratic in age is also included, giving seven degrees of freedom for lack of fit in the **analysis of variance** presented in Table 2.

The impression from the plot of a strong linear relationship between htc and age is overwhelmingly confirmed by the value of 160.182 for the F statistic on 1 and 73 degrees of freedom. There is no evidence of curvature, that is, of regression on $(\text{age})^2$. The lack of fit sum of squares has a P value of around $3\frac{1}{2}\%$, a slight, but uninformative, indication that all may not be well.

One possibility is that the linear model is inadequate. Fitting higher-order polynomial models up to $(\text{age})^9$ yields significant t values for age to the powers 4, 6, and 7, which powers are hard to credit or interpret. The further analysis of Francis et al. [9] suggests that variance increases with mean and that a **transformation** of htc is appropriate, probably the logarithmic transformation. With a nonnegative variable such as length, with a range from 1.0 to 8.9, it is extremely likely that variance will increase with observation size. Obtaining a satisfactory transformation is more systematically dealt with by the methods described in **Power Transformations**. The further analysis of Francis it et al. [9] was based on plots, particularly of residuals. We now consider some aspects of the use of residuals in determining goodness of fit.

Residuals

We assumed in (2) that the errors $\varepsilon_i, i, = 1, \ldots, n$, were identically and independently normally distributed. Even if this is true, the residuals e_i, defined in (5), although normally distributed, have **covariance matrix** $(\boldsymbol{I} - \boldsymbol{H})\sigma^2$ and will so not quite be a random sample from a normal distribution. However, it is customary in goodness-of-fit procedures to proceed as if the residuals from a satisfactory model will be such a sample. In **Residuals**, we describe the use of **simulation** envelopes in plotting to correct for this assumption.

Once a model has been fitted, the normality of the residuals, and so the goodness of fit of the model, can be checked in a variety of ways. These include histograms of the residuals (*see* **Frequency Distribution**), perhaps accompanied by a Pearson chi-squared goodness-of-fit test, a normal plot of the residuals and the calculation of a test of normality (*see* **Normality, Tests of**) such as that of Bowman and Shenton [4]. We first consider an example to illustrate the graphical techniques.

Royston and Altman [19] discuss the analysis of measurements of mandible length as a function of gestational age in 167 fetuses with ages from 8 weeks. The data are plotted in Figure 2. Although data with age >28 weeks are felt to be atypical, we include them in the present analysis.

To obtain a regression model we ignore the replicate observations over weeks. For illustration, we fit a simple regression on age (*see* **Linear Regression, Simple**). Other models for these data are discussed in

Table 2 Analysis of variance for regression models fitted to data on the head to coccyx length of 83 Brazilian fetuses as a function of gestational age in weeks

Source	df	Sum of squares	Mean square	F	Pr(F)
Age	1	165.992	165.992	160.182	0.0000
$(\text{Age})^2$	1	0.966	0.966	0.932	0.3375
Lack of fit	7	16.836	2.405	2.321	0.0340
Residual	73	75.648	1.036		
Total	82	259.442			

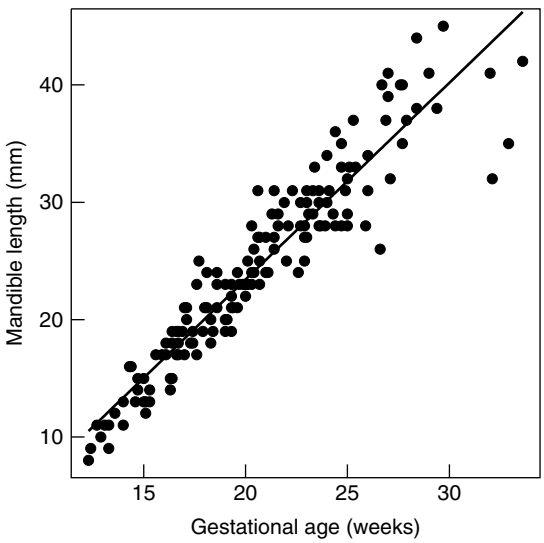

Figure 2 Scatter plot of mandible length data, showing a linear relationship and some increase of variance with age

the article on **Diagnostics**. The fitted values from this simple model are shown in Figure 2. It is noticeable that variance of the observations, and hence the size of the residuals, increases with age. This is evident in Figure 1 of **Residuals**, where the residuals are plotted against fitted values. Here we consider ways of testing goodness of fit on the basis of these residuals.

Figure 3 shows a normal plot of the least-squares residuals (5), which range from −11.71 to 6.56. It is clear from the figure that the residuals are not normally distributed. The histogram of the residuals in Figure 4 confirms the **skewness** apparent in Figure 3. One way of formally testing for the apparent nonnormality would be to calculate Pearson's X^2 for the 167 residuals, grouped as in the histogram, with the expected values given by a fitted normal distribution. However, the modification to the grouping necessary to avoid empty, or nearly empty, cells would remove much of the evidence of nonnormality, which comes from the long left-hand tail of the distribution. We use instead a **Monte Carlo** version of the Bowman–Shenton test which shows the statistical **power** that can be achieved from an appropriate test on one degree of freedom.

The Bowman–Shenton Test. The test is based on comparing a combination of the third and fourth **moments** of the residuals with those of the normal

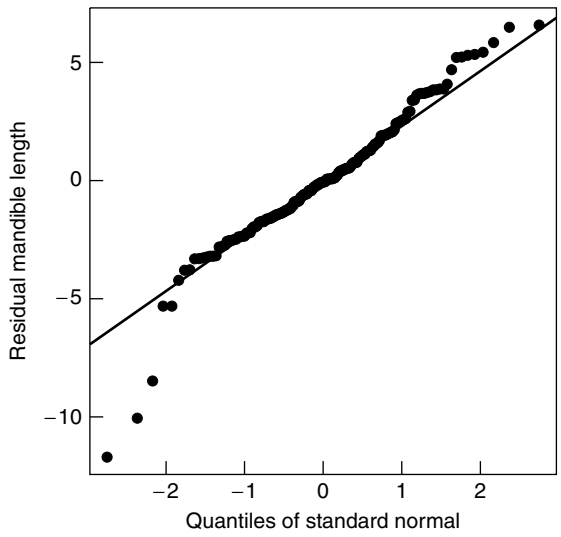

Figure 3 Mandible length data. Normal $Q-Q$ plot showing large negative residuals

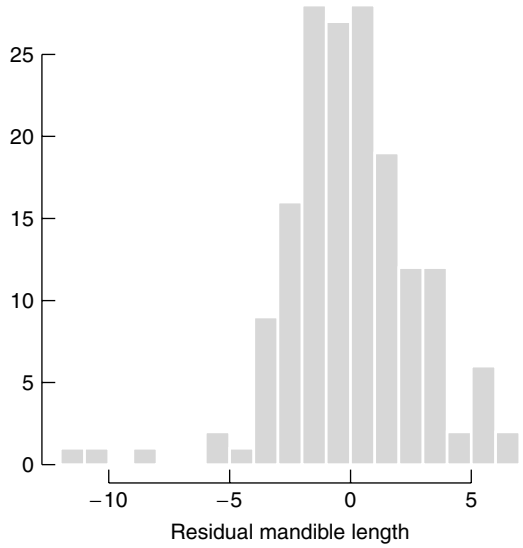

Figure 4 Mandible length data. Histogram of least-squares residuals showing a skewed distribution

distribution. Let

$$m_r = \sum_{i=1}^{n} \frac{(e_i - \overline{e})^r}{n}, \tag{8}$$

where, for regression models containing a constant, $\overline{e} = \sum e_i/n = 0$. Then the skewness measure is

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}}, \tag{9}$$

with the **kurtosis** measure given by

$$b_2 = \frac{m_4}{m_2^2}. \tag{10}$$

Asymptotically,

$$\sqrt{b_1} \sim N\left(0, \frac{6}{n}\right) \quad \text{and} \quad b_2 \sim N\left(3, \frac{24}{n}\right).$$

Since $\sqrt{(b_1)}$ and b_2 are approximately independent, the statistic

$$BS = \frac{nb_1}{6} + \frac{n(b_2 - 3)^2}{24} \tag{11}$$

should have approximately, a chi-squared distribution on two degrees of freedom. Bowman and Shenton [4] take transformations of $\sqrt{(b_1)}$ and b_2 to improve

normality. They also provide charts and tables for the distribution of the resulting statistic. In econometrics, it is customary to use (11) directly (for example, Harvey [10, p. 260]). When, as here, the statistic is based on residuals from regression, the effect of the **correlation** structure induced by the carrier matrix $\boldsymbol{X}$ needs to be accommodated. We accordingly use a form of Monte Carlo testing.

For the data of Figure 2, the value of BS is 49.36, clearly very large for a statistic with a nominal χ^2_2 distribution. 999 simulated values of the statistic based on the residuals from regressing random **standard normal deviates** on age had a maximum value of 28.68. The observed result is thus significant far past the 0.1% point. In contrast, application of the same procedure to the data of Figure 1 yields a value of 3.87. Ordering this within the 999 simulated values for this particular X matrix puts the observed value 90th, so that the significance level is 9%, implying no evidence of a departure from normality.

This Monte Carlo procedure is very general and could be applied to any version of the test. For example, the Bowman and Shenton [4] transformation to improve normality might be used, or the least-squares residuals could be standardized to have the same variance, the test statistic being calculated from values of $e_i/\sqrt{(1-h_i)}$, where h_i is the ith diagonal element of the hat matrix (4). However, by focusing narrowly on the normality of the errors, such procedures lose the information contained in the relationship between the residuals and other variables, such as fitted values. The increase of variance with fitted value seen in Figure 2 is one pointer that a power transformation of the data might be appropriate. The analysis of the mandible length data is taken further in **Residuals**, **Diagnostics**, and **Power Transformations** as well as in the articles on the **Fan Plot** and the **Forward Search**. In conclusion, we note that any transformation of the response, such as replacing $\boldsymbol{y}$ by $\log \boldsymbol{y}$, will affect the apparent linear relationship with age.

Many Regression Models

In the polynomial regressions of the previous section, there was an obvious **hierarchy of models** that could be successively tested for goodness of fit via the addition of extra terms. However, with regression on k different variables, there will be 2^k possible models, when all combinations of inclusion of each variable are considered, and there is no obvious order in which to test the models. If an estimate of σ^2 is available, perhaps from replicate observations, the fit of each model can be determined from the lack of fit sum of squares and potential models divided into two classes, those that fit the data and those that do not. Selection amongst those that do fit can then be based on grounds of parsimony, since if a model fits, models formed by adding terms to it will also fit. We now briefly describe some methods of testing goodness of fit in the absence of an estimate of σ^2.

The Coefficient of Determination, R^2

The coefficient of determination, or multiple R^2, measures the proportion of the variation in the data "explained" by the fitted model. Let the total corrected sum of squares be S_o, that is,

$$S_o = \sum_{i=1}^{n} (y_i - \overline{y})^2. \tag{12}$$

Then, the regression sum of squares equals $S_o - S(\hat{\beta})$ and

$$R^2 = \frac{S_o - S(\hat{\beta})}{S_o} = 1 - \frac{S(\hat{\beta})}{S_o}. \tag{13}$$

The theory is that a "good" model will have a value of R^2 near one.

There are two problems with the use of this measure of goodness of fit. One is the interpretation of the values, which do not have a standard distribution and may be misleading. The results in Table 2 for the data plotted in Figure 1 give an F value for linear regression of 160.182, whereas the value of R^2 is a rather uninspiring 0.6398. Likewise, for the data of Figure 2, $F = 1289$ with $R^2 = 0.8865$. Despite the overwhelming evidence for the regressions, the values of R^2 are far from one.

The second problem with R^2 is that the value increases as extra terms are added to the model, since $S(\hat{\beta})$ certainly cannot increase and usually decreases. Some allowance can be made by the use of **adjusted $\boldsymbol{R^2}$**

$$\overline{R^2} = 1 - \frac{S(\hat{\beta})/(n-p)}{S_o/(n-1)}. \tag{14}$$

Mallows' C_p

Mallows $\boldsymbol{C_p}$ measures goodness of fit by considering **prediction** by fitted models at the n observational

points. For satisfactory models

$$C_p = \frac{S(\hat{\beta})}{s^2} + 2p - n \qquad (15)$$

is approximately equal to p, the dimension of $\boldsymbol{\beta}$. Often s^2, the estimate of σ^2, is the residual mean square estimate from fitting the model including all terms being considered. If there are p^* terms, for this model $C_p = p^*$. The hope is that at least one reduced model will also have an acceptably small value of C_p. Interpretation of the values is often aided by a plot of C_p against p. Examples of the use of C_p to select models, and comparison with other procedures, are given in many statistics textbooks, for example, Hines and Montgomery [11]. The article on **Variable Selection** extends these comments and includes a discussion of **Bayesian methods** of selection.

Mallows criterion uses the mean squared error of prediction at the observational points to strike a balance between the improved predictions from fitting a larger model and the resulting increased variance of the predictions. The scaled residual sum of squares is penalized by twice the number of parameters in the model. The generalization of this idea to **likelihood**, known as **Akaike's information criterion**, or *AIC*, is mentioned in the next section.

Once a method such as the analysis of variance, $\overline{R}^2$ or C_p has been used to select one, or a few, satisfactory models, further checking should be undertaken. See **Residuals** and **Diagnostics** for methods based on the deletion of individual observations.

Generalized Linear Models

The Analysis of Deviance

Generalized linear models extend the regression model of the previous section to errors having a one-parameter **exponential family**, most importantly the **binomial** and **Poisson**, as well as the normal and **gamma**. The mean of each observation $\mu_i = E(Y_i)$ is related to the linear predictor $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ by the link function $g(\mu_i) = \eta_i$. The standard reference is McCullagh and Nelder [15], with an introduction by Dobson [7].

Inference procedures parallel those for regression, the generalized linear model with normal errors and identity link, $g(\mu_i) = \mu_i$. The analysis of deviance is the generalization of the analysis of variance of the previous section. The calculation and interpretation of residuals and other diagnostic quantities are aided by the reduction of **maximum likelihood** estimation to iteratively reweighted least squares.

Let the loglikelihood be $l(\boldsymbol{\beta}, \phi; \boldsymbol{y})$, where ϕ is a scale parameter, equal to σ^2 for the normal distribution. To test goodness of fit of models the loglikelihood is compared with the model which fits best, the one for which $\hat{\mu}_i = y_i$. Call these parameter estimates $\hat{\boldsymbol{\beta}}_{\max}$. The scaled deviance

$$D^*(\hat{\beta}) = 2\{l(\hat{\boldsymbol{\beta}}, \phi; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}_{\max}, \phi; \boldsymbol{y})\} \qquad (16)$$

equals $R(\hat{\beta})/\sigma^2$ for the regression model and can be used in the same way to test goodness of fit of models. If the hypothesis to be tested is that $\boldsymbol{\beta} = \boldsymbol{\beta}_o$, of dimension $p - s$, then $D^*(\hat{\beta}_o) - D^*(\hat{\beta})$ is compared with χ_s^2, the asymptotic distribution of the likelihood ratio. If ϕ is not known, it can be estimated from the residual unscaled deviance $D(\hat{\beta}) = \phi D^*(\hat{\beta})$, by setting $\hat{\phi} = D(\hat{\beta})/(n - p)$ for a model known to fit well.

We shall be particularly interested in generalized linear models with Poisson and binomial errors, for both of which $\phi = 1$. The value of either deviance then provides a test of goodness of fit of the model, which may again have an asymptotic χ^2 distribution provided the number of observations is sufficiently great relative to the number of parameters, the binomial case being particularly sensitive. A discussion and further references are given by Firth [8, Section 3.5.2]. The generalization of Mallows C_p to Akaike's information criterion,

$$\mathrm{AIC}_p = D^*(\hat{\beta}) + 2p = \frac{D(\hat{\beta})}{\phi} + 2p, \qquad (17)$$

allows the selection of models with small values of AIC_p for further checking. (Omission of the term in n which appears in the definition of C_p does not affect the ordering of the models). If it is necessary to estimate ϕ, the same estimate should be used for all comparisons between models. Both in this case and in regression, if the number of observations is large, there is the danger of selecting models with too many parameters. **Parsimonious** model choice can be achieved using penalties that increase with n. For example, Schwartz [20] suggests a criterion of the form

$$\frac{D(\hat{\beta})}{\phi} + p \log n,$$

which is more parsimonious than AIC for $n \geq 8$. Bias-corrected methods for the selection of regression and **time-series** models are described by Hurvich and Tsai [13].

Contingency Tables

Generalized linear models for **contingency tables**, such as the data of Table 1, can often be modeled with Poisson errors and the loglink $\log \mu_i = \eta_i$. The deviance for a model giving fitted values μ_i is

$$D(\mu) = 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right\}, \quad (18)$$

a statistic often called G^2 in the literature on the analysis of **categorical data** (for example, Agresti [1]). For the data of Table 1, the deviance is 0.475 for the expected frequencies given in the table, which do not involve any estimated parameters. This value yields the same inference as the value of 0.47 for Pearson's X^2. Pearson's statistic, written in this notation is

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\mu_i}. \quad (19)$$

The asymptotic equivalence of the deviance and X^2 for testing goodness of fit can be shown by expanding (18) in a Taylor series in $(y_i - \mu_i)/\mu_i$. (Kendall and Stuart [14, Section 30.8], McCullagh and Nelder [15, p. 197]). However, the asymptotic equivalence of the two tests says little about which has higher power for moderate samples, although there seems to be little difference. Some references are given by Agresti [1].

Binomial Data

Table 3 gives data from Bliss [3] on the mortality of adult beetles after five hours exposure to gaseous carbon disulfide. At each of the eight dose levels x_i, there are n_i binomial trials resulting in R_i successes (from the point of view of the experimenter). The question of interest is the relationship between x_i and the probability of success $\theta_i = E(R_i/n_i)$.

Table 3 Bliss's data on mortality of beetles

Log dosage, x_i	Number exposed, n_i	Number killed, R_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

A standard model for the analysis of such data is the linear logistic model (*see* **Logistic Regression**) in which

$$\log\left\{\frac{\theta_i}{1-\theta_i}\right\} = g(\theta_i) = \beta_o + \beta_1 x_i. \quad (20)$$

When this model is fitted, the residual deviance is 11.23 on 6 degrees of freedom. Here the number of insects at each dose level is sufficiently large that a central limit theorem might be expected to hold for each R_i and so it is sensible to compare 11.23 with the chi-square distribution on 6 degrees of freedom, for which the 5% point is 12.59. Although this is not a significant value, plots of observed and fitted values (**Residuals**, Figure 6) suggest a systematic lack of fit, which can be addressed either by adding a term in $\boldsymbol{x}^2$ to the linear predictor or by considering another link function. There are two ways in which a test of the link function may be obtained on a few degrees of freedom. Such tests are often called "goodness of link" tests.

An alternative to the logistic link is the complementary log–log link in which $g(\theta_i) = \log\{-\log(1 - \theta_i)\}$. These two can be combined in the parametric family

$$g(\theta_i, \lambda) = \log\left[\frac{\{1/(1-\theta_i)^{\lambda} - 1\}}{\lambda}\right], \quad (21)$$

which is the logistic link, as in (20), for $\lambda = 1$ and tends to the complementary log–log as $\lambda \to 0$. A test on one degree of freedom can be obtained either by finding the value of λ that maximizes the likelihood and by performing a **likelihood ratio test**, or by using a score test for $\lambda = 0$ or 1 (*see* **Likelihood**). A discussion and examples of other embeddings of links are given by Pregibon [16].

A more general test of the goodness of the link $g(\mu_i)$ is obtained by assuming that the true link is indeed $g^*(\mu_i) = \eta_i$, so that

$$g(\mu_i) = g\{g^{*-1}(\eta_i)\} = h(\eta_i). \quad (22)$$

Expansion of $h(\eta_i)$ in a Taylor series leads to a linear predictor that includes an added term in $\hat{\eta}_i^2$. Refitting the model with this extra variable without

recalculating the weights used in fitting leads to a reduction in deviance of 7.61 to be assessed on one degree of freedom – strong evidence, when compared with χ^2_1, of the need for another link with the linear predictor in (20).

For these data, two models giving satisfactory fits, as assessed by the residual deviance, are the complementary log–log model with a linear term in $\boldsymbol{x}$ and the logistic link including a quadratic in $\boldsymbol{x}$. Further details and related plots are in **Residuals** and the **Forward Search**. The interpretation of the residual deviances as chi-squared values need care. With the logistic link and all $n_i = 1$, the residual deviance is a function only of $\hat{\beta}$ and so contains no information about goodness of fit. McCullagh and Nelder [15, p. 121] discuss the distribution of the deviance and Pearson's X^2 for small n_i. Atkinson and Riani [2, Section 6.18], explore the relationship between the analysis of deviance and t tests for coefficients in the linear predictor when all $n_i = 1$.

The goodness of link test using $\hat{\boldsymbol{\eta}}^2$ is an example of an aggregate test, related to Tukey's one degree of freedom for nonadditivity in regression, which involves regression on $\hat{\boldsymbol{y}}^2$. A number of related aggregate tests are described by Hinkley [12]. In **Residuals**, we describe the use of added variable and related plots to determine the effect of individual observations on such procedures. Methods for the analysis of binary data are described, amongst others, by Cox and Snell [5, Section 2.7], who stress the importance of not relying on overall tests of goodness of fit, but of trying to detect scientifically important departures from the model.

Tests of Distributional Shape

A large part of the literature on goodness of fit is concerned with testing distributional shape. A full coverage is given by D'Agostino and Stephens [6]. One example is the Bowman–Shenton test, which uses moment properties of the normal distribution. A more general class of tests, applicable to a general continuous distribution, compares the empirical distribution function $F_n(y)$ of the sample with the cumulative distribution function $F_o(y)$ of the proposed distribution. As an example, we take the **Kolmogorov–Smirnov** test, based on the maximum difference between the two distributions

$$D = \sup_y |F_n(y) - F_o(y)|, \tag{23}$$

and apply it to the residuals of the data on mandible length.

The empirical distribution function (edf) of a sample of n observations y_i is the number of observations $\leq y$. If the ordered observations are denoted $y_{(i)}$, the e.d.f. is given more formally by

$$F_n(y) = \frac{i}{n} \quad y_{(i)} \leq y < y_{(i+1)}, \tag{24}$$

with

$$F_n(y) = 0, \quad y < y_{(1)} \quad \text{and} \quad F_n(y) = 1 \quad y_{(n)} \leq y. \tag{25}$$

The test is most easily performed by applying the probability integral transformation to the observations $\boldsymbol{y}$ to give a sample $\boldsymbol{z}$ that is **uniformly** distributed under the hypothesized distribution. For example, for the normal distribution with cdf $\Phi(w)$, let

$$w_i = \frac{y_{(i)} - \mu}{\sigma} \quad \text{and} \quad z_{(i)} = \Phi(w_i). \tag{26}$$

Figure 5 shows a plot of the edf of $F_n(z)$ against z, that is, of i/n against z. The two-sided Kolmogorov–Smirnov test for the test of normality is the maximum of the two one-sided tests calculated as

$$D^+ = \sup_i \left\{ \frac{i}{n} - z_{(i)} \right\}; \quad D^- = \sup_i \left\{ z_{(i)} - \frac{i-1}{n} \right\} \tag{27}$$

giving the test statistic

$$D = \sup\{D^+, D^-\}. \tag{28}$$

The value of the statistic for the mandible data is 0.0667. If the null distribution were completely specified, the statistic would not depend on the assumed distribution and the result would not be significant at the 5% level. In calculating the statistic, an estimate has had to be used for the standard deviation in the probability integral transformation, although the mean of the residuals is zero. Both the effect of estimation and of the correlation pattern of the residuals should be allowed for in calculating the significance of the statistic, although here there seems no doubt about the lack of evidence of departure from normality.

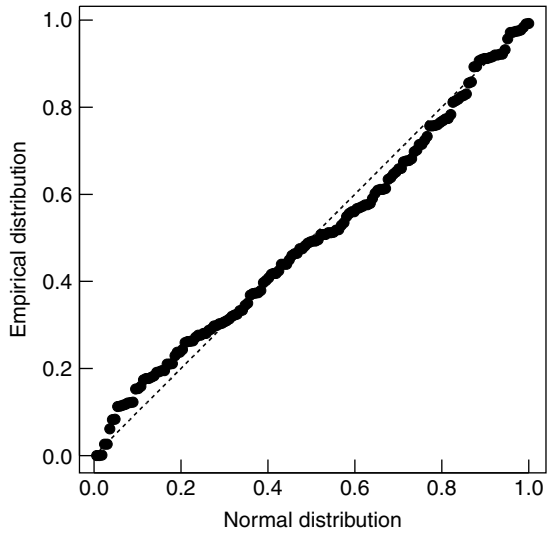

Figure 5 Mandible length data. The Kolmogorov–Smirnov test for normality is based on the greatest vertical distance between the empirical distribution and the dotted line

The comparison of Figure 5 with Figure 3 shows that it is not straightforward to interpret the edf of the z in terms of individual outliers. In particular, Figure 5 indicates why the Kolmogorov–Smirnov test failed to detect the outliers from an otherwise seemingly normal distribution. The general point, made in the introduction to this article, is the frequent lack of power of overall tests of fit against specific alternatives. Normal probability plots are more sensitive to **outliers** than the plot of Figure 5, although guidance, for example, from a simulation envelope, may be needed as to whether the line is sufficiently straight. If the interesting departures are likely to be in the tails of the distribution, statistics giving greater weight to the extremes will be more powerful. D'Agostino and Stephens [6] describe several such statistics. For discrete distributions methods such as Pearson's X^2 are often used.

Tests of **multivariate distributions** are in general much less well developed, an exception being the **multivariate normal distribution**. One possibility is to apply the multivariate Box–Cox transformation to normality (for example, Velilla [21], Riani and Atkinson [17]), testing to see whether a transformation is needed. A comparison of tests of **multivariate normality** is given by Romeu and Ozturk [18].

References

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. Wiley, New York.

[2] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

[3] Bliss, C.I. (1935). The calculation of the dosage-mortality curve, *Annals of Applied Biology* **22**, 134–167.

[4] Bowman, K.O. & Shenton, L.R. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2, *Biometrika* **62**, 243–250.

[5] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London.

[6] D'Agostino, R.B. & Stephens, M.A. eds. *Goodness-of-fit Techniques*. Marcel Dekker, New York, 1986.

[7] Dobson, A. (2001). *An Introduction to Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[8] Firth, D. (1991). Generalized linear models, in *Statistical Theory and Modelling*, D.V. Hinkley, N. Reid & E.J. Snell eds. Chapman & Hall, London, pp. 55–82.

[9] Francis, B., Green, M. & Payne, C. eds. (1993). *The GLIM System: Release 4 Manual*. Clarendon Press, Oxford.

[10] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, MA.

[11] Hines, W.M. & Montgomery, D.C. (2002). *Probability and Statistics in Engineering and Management Science*, 4th Ed. Wiley, New York.

[12] Hinkley, D.V. (1985). Transformation diagnostics for linear models, *Biometrika* **72**, 487–496.

[13] Hurvich, C.M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika* **76**, 297–307.

[14] Kendall, M.G. & Stuart, A. (1973). *The Advanced Theory of Statistics*, Vol. 2. Griffin, London.

[15] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[16] Pregibon, D. (1980). Goodness of link tests for generalized linear models, *Applied Statistics* **29**, 15–23.

[17] Riani, M. & Atkinson, A.C. (2001). A unified approach to outliers, influence and transformations in discriminant analysis, *Journal of Computational and Graphical Statistics* **10**, 513–544.

[18] Romeu, J.L. & Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality, *Journal of Multivariate Analysis* **46**, 309–334.

[19] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[20] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

[21] Velilla, S. (1995). Diagnostics and robust estimation in multivariate data transformations, *Journal of the American Statistical Association* **90**, 945–951.

(*See also* **Bayesian Methods for Model Comparison**; **Bayesian Measures of Goodness of Fit**; **Model Checking**; **Model, Choice of**)

A.C. ATKINSON

Gosset, William Sealy

Born: June 13, 1876, in Canterbury, UK.
Died: October 16, 1937, in London, UK.

William Sealy Gosset was born to Frederic Gosset, a Colonel in the Royal Engineers, and Agnes Sealy Vidal. He was a Scholar of Winchester College from 1889 to 1895, when he took up a scholarship at New College, Oxford, which he left in 1899 with a first class degree in Chemistry. In October 1899, he became a brewer with Arthur Guinness, Son & Co. Ltd and remained with the firm for the whole of his working life, mostly in Dublin, but moving to London in 1935. Gosset was married on January 16, 1906, to Marjorie Surtees Philpotts, the sister of a fellow brewer, and there were three children. He died aged 61.

The business of Guinness was then entirely concerned with the brewing of stout. Gosset was one of several men from Oxford and Cambridge who were appointed to make greater use of scientific methods. His work called for the analysis of experimental data from short runs, during which there were changes in both materials and the environment. He had no statistical training, and turned to the standard textbooks on the combination of observations in astronomy and geodesy. Their treatments assumed long series of observations made under stable conditions, whereas Gosset required methods that could be applied to small samples. He made contact with **Karl Pearson**, head of the Biometric Laboratory at University College, London, and attended lectures and tutorials there during the period 1906–1907. They included the idea of **correlation**, the Pearson system of nonnormal distributions (*see* **Pearson Distributions**), and the **chi-square test** for **goodness of fit**.

Gosset's early work is a fusion of classical methods for the combination of observations with the new ideas promoted at the Biometric Laboratory. His paper on the probable error of a mean, published in 1908, is the best example. The distribution of his test criterion, then called z, required him to find the distribution of the **variance** in normal samples, and prove its independence from the long-known distribution of the sample **mean**. He solved the first by fitting a Pearson curve with the same **moments**, and gave support to the second by establishing no correlation between mean and variance. The proof is not rigorous, but his distribution of z, now modified to t, is correct, and has been a discovery of the greatest importance throughout statistical methodology. Those who meet his test criterion for the first time usually do so as **Student's *t***, a pseudonym initially adopted to permit publication, and retained long after his work became well known.

After Gosset's return to Ireland, his published work was mainly concerned with agricultural field trials and the design of experiments, derived from his experience in the Irish barley fields. He always took an interest in the activities of his former teacher, and his private correspondence was gradually enlarged to deal with questions from other statisticians, including **R.A. Fisher**, from 1912. The effect on Fisher of Gosset's paper on the probable error of a mean was profound, leading on the one hand to the detailed exploration of **normal distribution** theory, and on the other to a central part of *Statistical Methods for Research Workers* [1], which eventually made Gosset's work known throughout the statistical world. Their correspondence turned to matters of experimental design after Fisher moved to Rothamsted in 1919, and here differences of opinion eventually emerged regarding the relative merits of **randomization** and balance. Towards the end of his career, Gosset became interested in evolutionary genetics, and Fisher helped with the promotion of his ideas.

Gosset gave advice, criticism, and enlightenment at a crucial phase in the career of **Egon S. Pearson** between 1926 and 1931, and also friendship, which the younger man never forgot. The foregoing account

is based on the book by E.S. Pearson [2] and appears by permission of Oxford University Press.

References

[1] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[2] Pearson, E.S. (1990). *"Student": A Statistical Biography of William Sealy Gosset*, G.A. Barnard & R.L. Plackett, eds. Oxford University Press, Oxford.

ROBIN L. PLACKETT

Graeco–Latin Square Designs

A Graeco–Latin square is a pair of **orthogonal Latin squares**, so named because Euler in 1782 used Roman and Greek letters to distinguish the symbols of the two Latin squares. Sometimes it is also referred to as an *Eulerian square*. If $S = (s_{ij})$ and $T = (t_{ij})$ are two Latin squares, each of *order* n, and defined on the set of symbols $1, 2, \ldots, n$, they are said to be *orthogonal* if each of the n^2 possible ordered pairs of the symbols occurs just once among the pairs (s_{ij}, t_{ij}). Any two of the Latin squares of order 4 in Figure 1 are orthogonal, and hence each pair forms a Graeco–Latin square.

A Graeco–Latin square is usually shown with the two symbols together in one square, often using Roman letters $A, B, C, \ldots$ for S and Greek letters $\alpha, \beta, \gamma, \ldots$ for T. Alternative representations use the T symbols as a suffix of the S symbols, or use ordered pairs (s_{ij}, t_{ij}). The first two squares in Figure 1 can be represented as in Figure 2.

Graeco–Latin squares exist for all orders n except 1, 2 and 6. In his original paper, Euler proved the existence of Graeco–Latin squares when n is an odd integer or a multiple of 4 but conjectured nonexistence of the squares for $n = 4k + 2$ for all $k = 1, \ldots$. In 1900, G. Tarry proved the conjecture for $k = 1$ but later, in 1960, Bose et al. [2] proved that it cannot hold for all $k > 1$. This constituted the first complete proof on the existence of Graeco–Latin squares (see [4] for more details, including other proofs).

A Graeco–Latin square can, in theory, be used for various experimental situations in which there are four effects that can be assumed to be additive. The four effects are orthogonal, leading to simple estimates and a simply calculated **analysis of variance**. With the usual constraints that the estimates sum to zero for each effect, the estimate for an effect symbol is the **mean** value for that symbol minus the overall mean. Each effect sums of squares is the sum of squares of the n estimates multiplied by n, and each is on $n - 1$ **degrees of freedom**. The residual degrees of freedom is therefore $(n - 1)(n - 3)$. As written, the design naturally appears to be a row–column design (a design with two orthogonal blocking structures) with n rows and n columns, and treatments formed by two orthogonal factors each of n levels. Only the main effects of these factors can be estimated. Although this use with two factors is rare, the design essentially arises when the units of an experiment on n treatments with row–column **blocking** using a Latin square design are to be used in a later experiment on another n treatments, and it is wished to have the new treatments orthogonal to the previous ones. The new treatments are said to be *superimposed* on the old ones.

Other possible situations when a Graeco–Latin square might be used are as a three-dimensional row–column–layer design (three orthogonal blocking effects each with n categories) for an experiment on n treatments (using the Roman letters for the layer levels); a $1/n$ **fractional factorial design** for three orthogonal factors with n categories (n^{3-1} design) having n blocks of size n (using the rows for blocks); and a $1/n^2$ single block fractional factorial design

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

1	2	3	4
3	4	1	2
4	3	2	1
2	1	4	3

1	2	3	4
4	3	2	1
2	1	4	3
3	4	1	2

Figure 1

$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
$B\gamma$	$A\delta$	$D\alpha$	$C\beta$
$C\delta$	$D\gamma$	$A\beta$	$B\alpha$
$D\beta$	$C\alpha$	$B\delta$	$A\gamma$

1_1	2_2	3_3	4_4
2_3	1_4	4_1	3_2
3_4	4_3	1_2	2_1
4_2	3_1	2_4	1_3

(1,1)	(2,2)	(3,3)	(4,4)
(2,3)	(1,4)	(4,1)	(3,2)
(3,4)	(4,3)	(1,2)	(2,1)
(4,2)	(3,1)	(2,4)	(1,3)

Figure 2

for four orthogonal factors with n categories (n^{4-2} design). In the factorial cases, only main effects are estimable. Using the example of Figure 2 and factors A, B, C, D, the latter two cases give the following designs in the usual notation with levels $0, 1, \ldots, n-1$ (brackets denote blocks, and factors may be superimposed treatments):
a 4^{3-1} design in four blocks:

$$(a_0b_0c_0, \quad a_1b_1c_1, \quad a_2b_2c_2, \quad a_3b_3c_3),$$
$$(a_0b_1c_2, \quad a_1b_0c_3, \quad a_2b_3c_0, \quad a_3b_2c_1),$$
$$(a_0b_2c_3, \quad a_1b_3c_2, \quad a_2b_0c_1, \quad a_3b_1c_0),$$
$$(a_0b_3c_1, \quad a_1b_2c_0, \quad a_2b_1c_3, \quad a_3b_0c_2);$$

a 4^{4-2} design in one block:

$$(a_0b_0c_0d_0, \quad a_0b_1c_1d_1, \quad a_0b_2c_2d_2, \quad a_0b_3c_3d_3,$$
$$a_1b_0c_1d_2, \quad a_1b_1c_0d_3, \quad a_1b_2c_3d_0, \quad a_1b_3c_2d_1,$$
$$a_2b_0c_2d_3, \quad a_2b_1c_3d_2, \quad a_2b_2c_0d_1, \quad a_2b_3c_1d_0,$$
$$a_3b_0c_3d_1, \quad a_3b_1c_2d_0, \quad a_3b_2c_1d_3, \quad a_3b_3c_0d_2).$$

A Graeco–Latin square can also be used to obtain a four-replicate resolvable (0, 1) (all within-block pairwise concurrences are 0 or 1) **Lattice square design** for n^2 treatments in n blocks of size n. The replicates are formed by writing the numbers 1 to n^2 in an n by n array, and then using the rows for the blocks for first replicate, the columns for the second, the Roman letters for the third, and the Greek letters for the fourth.

Two drawbacks to the widespread use of the Graeco–Latin square in practice are the need for all four effects to have n levels, and that it is rare for a factor to have more than about five levels. When used, a valid **randomization** of treatment labels to units is necessary – see [10]. Preece et al. [12] show that care is needed to ensure a valid randomization when treatments are superimposed.

An extension of the Graeco–Latin square is to have more than two orthogonal Latin squares. Situations in which $m > 2$ *mutually orthogonal Latin squares* (MOLS) can be used are natural extensions of those discussed for the Graeco–Latin square. The maximum possible value for m is $n-1$, which is attainable when n is a prime power. Then the $n-1$ orthogonal Latin squares are known as a *complete set*. Figure 1 gives an example for $n = 4 = 2^2$. Clearly, for $n = 6$ (and 1, 2) the maximum value of m is 1. The maximum value of m that is possible is denoted $N(n)$. Although $N(n)$ can be bounded below, its value is not known even for $n = 10$. It is known that $N(10) \geq 2$, $N(12) \geq 5$, and $N(n) \geq 3$ for all $n > 10$. A table of recent lower bounds for $N(n)$ for n up to 199 is given by Brouwer [3] (but note $N(27) = 26$, $N(97) = 96$). Complete sets of MOLS of orders 3 to 5, 7 to 9, are given in [5].

The existence of m MOLS of order n is equivalent to the existence of what is known as an orthogonal array OA $(n, m+2, 2, 1)$ of $m+2$ constraints, n levels, strength 2, and index 1. Some design constructions are given in terms of the OA (see [8]). If $m > 2$, then the above construction using a Graeco–Latin square for $m = 2$ can be extended to get an $(m+2)$- replicate resolvable (0, 1) *square lattice design* for n^2 treatments in n blocks of size n. When $m = n-1$, the complete set gives the $(n+1)$-replicate *balanced square lattice design* or the symmetric **balanced incomplete block design** (BIBD) $(n^2, \quad n(n+1), \quad n)$, equivalent to a *finite projective plane* with $n+1$ points on every line. A *balanced lattice square*, a nested row–column design with n^2 treatments for which each $n \times n$ block is a complete replicate, can be constructed from a balanced square lattice design, or equivalently from the $n-1$ MOLS. In general, m MOLS can be used to construct an $(m+2)$- replicate lattice square. They can also be used to construct an $(n \times n)/m$ Trojan square design for nm treatments replicated n times in an $n \times n$ row-column array of plots each with m subplots [11].

Sets of MOLS can be sought for particular types of Latin squares, such as *self-orthogonal, Knut Vik, row-complete* – see [6]. The construction of an r-replicate *rectangular lattice design* for $n(n-1)$ treatments uses $r-2$ MOLS of order n, which have each symbol on the main diagonal, a *semi-diagonal* Latin square. Bailey [1] has shown that complete sets of mutually orthogonal *quasi-complete* Latin squares exist when n is an odd prime power. These have every symbol next to every other symbol twice in the rows and twice in the columns. A complete set of MOLS can also be used as a valid Latin square randomization set. Sets of MOLS may then be used for a valid Latin square *restricted randomization* scheme. Martin [7] showed that a complete set of mutually orthogonal quasi- complete Latin squares can be chosen to have low variation under spatial dependence.

When m MOLS do not exist, it may be possible to find m Latin squares that are "nearly" orthogonal –

see [6]. Further discussion of Graeco–Latin squares and MOLS is in Keedwell [6] and Street & Street [13]; see also Preece [9, 10].

References

[1] Bailey, R.A. (1984). Quasi- complete Latin squares: construction and randomization, *Journal of the Royal Statistical Society, Series B* **46**, 323–334.

[2] Bose, R.C., Shrikhande, S.S. & Parker, E.T. (1960). Further results on the construction of mutually orthogonal Latin squares and the falsity of Euler's conjecture, *Canadian Journal of Mathematics* **12**, 189–203.

[3] Brouwer, A.E. (1991). Recursive constructions of mutually orthogonal Latin squares, in *Latin Squares: New Developments in the Theory and Applications*, J. Dénes & A.D. Keedwell, eds. North-Holland, Amsterdam, pp. 149–168.

[4] Dénes, J. & Keedwell, A.D. (1974). *Latin Squares and Their Applications*. English Universities Press, London, Chapter 11.

[5] Fisher, R.A. & Yates, F. (1963). *Statistical Tables for Biological, Agricultural and Medical Research*, 6th Ed. Oliver & Boyd, Edinburgh, Tables XV, XVI.

[6] Keedwell, A.D. (1983). Graeco- Latin squares, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 469–474.

[7] Martin, R.J. (1986). On the design of experiments under spatial correlation, *Biometrika* **73**, 247–277.

[8] Mukhopadhyay, A.C. (1985). Orthogonal arrays and applications, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 523–527.

[9] Preece, D.A. (1983). Latin squares, Latin cubes, Latin rectangles, etc., in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 504–510.

[10] Preece, D.A. (1991). Latin squares as experimental designs, in *Latin squares: New Developments in the Theory and Applications*, J. Dénes & A.D. Keedwell, eds. North-Holland, Amsterdam, pp. 317–342.

[11] Preece, D.A. & Freeman, G.H. (1983). Semi-Latin squares and related designs, *Journal of the Royal Statistical Society, Series B* **45**, 267–277.

[12] Preece, D.A., Bailey, R.A. & Patterson, H.D. (1978). A randomization problem in forming designs with superimposed treatments, *Australian Journal of Statistics* **20**, 111–125.

[13] Street, A.P. & Street, D.J. (1987). *Combinatorics of Experimental Design*. Clarendon Press, Oxford, Chapter 7.

(*See also* **Design Effects**; **Factorial Designs in Clinical Trials**; **Magic Square Designs**)

RICHARD J. MARTIN &
SARALEESAN NADARAJAH

Gramian Matrix

A symmetric matrix of real numbers $\mathbf{A}$ is said to be a Gramian matrix if there exists a matrix $\mathbf{B}$ of real numbers such that $\mathbf{BB}' = \mathbf{A}$ or $\mathbf{B}'\mathbf{B} = \mathbf{A}$ (see, for example, [1]).

Consider an application involving a sample of n subjects measured on k variables. Suppose each observation y_{ij}, $i = 1, \ldots, k$; $j = 1, \ldots, n$, is centered about its variable mean, i.e. $x_{ij} = (y_{ij} - \overline{Y}_i)$, and normalized by dividing by $\left(\sum_j x_{ij}^2\right)^{1/2}$. Call this $k \times n$ matrix of centered and normalized observations $\mathbf{W}$. Note that the elements of $\mathbf{W}$ are of the form $w_{ij} = x_{ij} \Big/ \left(\sum_j x_{ij}^2\right)^{1/2}$, $i = 1, \ldots, k$; $j = 1, \ldots, n$. The $k \times k$ **correlation** matrix, $\mathbf{R}$ is $\mathbf{WW}'$. Since every element of $\mathbf{W}$ and $\mathbf{R}$ is real, $\mathbf{R}$ is a Gramian matrix. The elements of $\mathbf{R}$ are of the form

$$r_{ii'} = \frac{\sum_j x_{ij} x_{i'j}}{\left(\sum_j x_{ij}^2\right)^{1/2} \left(\sum_j x_{i'j}^2\right)^{1/2}}.$$

Suppose a **principal components analysis** is performed on $\mathbf{R}$ (the $k \times k$ correlation matrix). $\mathbf{R}$ can be written as $\mathbf{R} = \mathbf{EDE}'$, where $\mathbf{D}$ is a $k \times k$ diagonal matrix whose diagonal elements are the **eigenvalues** of $\mathbf{R}$, and $\mathbf{E}$ is a $k \times k$ matrix which is orthonormal by columns [1]. Since $\mathbf{D}$ is a diagonal matrix it can be written as $\mathbf{D} = \mathbf{D}^{1/2}(\mathbf{D}^{1/2})'$, and therefore $\mathbf{R} = (\mathbf{ED}^{1/2})(\mathbf{ED}^{1/2})'$ is Gramian as long as the elements of $\mathbf{D}$ are real. Suppose we let $\mathbf{F} = (\mathbf{ED}^{1/2})$; $\mathbf{F}$ is called the principal components loading matrix of $\mathbf{R}$. The matrix $\mathbf{F}$ can be viewed as a sequence of row vectors $\mathbf{f}_i' = (f_{i1}, f_{i2}, \ldots, f_{ik})$, $i = 1, \ldots, k$. The determinant of the matrix the elements of which are the inner products of the vectors $\mathbf{f}_i$ and $\mathbf{f}_{i'}$, denoted $(\mathbf{f}_i'\mathbf{f}_{i'})$,

$$\begin{vmatrix} (\mathbf{f}_1'\mathbf{f}_1) & (\mathbf{f}_1'\mathbf{f}_2) & \ldots & (\mathbf{f}_1'\mathbf{f}_k) \\ (\mathbf{f}_2'\mathbf{f}_1) & (\mathbf{f}_2'\mathbf{f}_2) & \ldots & (\mathbf{f}_2'\mathbf{f}_k) \\ \vdots & & & \vdots \\ (\mathbf{f}_k'\mathbf{f}_1) & (\mathbf{f}_k'\mathbf{f}_2) & \ldots & (\mathbf{f}_k'\mathbf{f}_k) \end{vmatrix} = \begin{vmatrix} f_{11} & f_{12} & \ldots & f_{1k} \\ f_{21} & f_{22} & \ldots & f_{2k} \\ \vdots & & & \vdots \\ f_{k1} & f_{k2} & \ldots & f_{kk} \end{vmatrix}^2,$$

is called the Gramian determinant.

In **factor analysis** the correlation matrix $\mathbf{R}$ is replaced by the reduced correlation matrix $\mathbf{R}^*$, where the ones on the diagonal of $\mathbf{R}$ are replaced with the **communalities**. Factor analysis then searches for a $k \times m (m < k)$ matrix $\mathbf{F}^*$, where $\mathbf{F}^*(\mathbf{F}^*)' = \mathbf{R}^*$. Here m is the dimension of the common factor space (see [1, Chapter 5]).

In practice, Gramian matrices and determinants are used, for example, in principal components analysis and factor analysis (see, for example, [1]), and to estimate parameters in **multiple regression** applications (see, for example, [2]).

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale, pp. 97–98, 144–146.

[2] Seber, G.A.F. (1977). *Linear Regression Analysis*. Wiley, New York, pp. 302–348.

(*See also* **Matrix Algebra**)

LISA M. SULLIVAN

Gram–Schmidt Process

The Gram–Schmidt process, also called the Gram–Schmidt **orthogonalization** process, is a technique for constructing an orthogonal basis from a basis spanning the same subspace. Consider a square oblique matrix $\mathbf{X}$ with row vectors $\mathbf{x}_i'(i = 1, 2, \ldots, m)$. The orthogonal matrix $\mathbf{Y}$ the first row of which is identical to the first row of $\mathbf{X}$ can be determined through the Gram–Schmidt process using the following equations, which are developed sequentially:

$$
\begin{aligned}
\mathbf{y}_1' &= \mathbf{x}_1', \\
\mathbf{y}_2' &= \mathbf{x}_2' - \left(\frac{\mathbf{x}_2'\mathbf{y}_1}{\mathbf{y}_1'\mathbf{y}_1}\right)\mathbf{y}_1', \\
\mathbf{y}_3' &= \mathbf{x}_3' - \left(\frac{\mathbf{x}_3'\mathbf{y}_1}{\mathbf{y}_1'\mathbf{y}_1}\right)\mathbf{y}_1' - \left(\frac{\mathbf{x}_3'\mathbf{y}_2}{\mathbf{y}_2'\mathbf{y}_2}\right)\mathbf{y}_2', \\
&\vdots \\
\mathbf{y}_m' &= \mathbf{x}_m' - \left(\frac{\mathbf{x}_m'\mathbf{y}_1}{\mathbf{y}_1'\mathbf{y}_1}\right)\mathbf{y}_1' - \left(\frac{\mathbf{x}_m'\mathbf{y}_2}{\mathbf{y}_2'\mathbf{y}_2}\right)\mathbf{y}_2' - \cdots \\
&\quad - \left(\frac{\mathbf{x}_m'\mathbf{y}_{(m-1)}}{\mathbf{y}_{m-1}'\mathbf{y}_{(m-1)}}\right)\mathbf{y}'_{(m-1)}.
\end{aligned}
$$

Since the $(\mathbf{x}_i'\mathbf{y}_j/\mathbf{y}_j'\mathbf{y}_j)$ terms produce scalars, each $\mathbf{y}_j'(j = 1, 2, \ldots, m)$ is a row vector. The rows of $\mathbf{Y}$ are then normalized producing an orthogonal matrix (see, for example, [1, pp. 237–242] for a complete discussion of the process with examples).

The technique is used in a variety of settings, including **battery reduction** analysis and in estimating parameters in **multiple regression** applications. In battery reduction analysis, for example, it is of interest to reduce the number of variables. One means of achieving this is through Gram–Schmidt orthogonal rotations of the initial **factor loading matrix** from a **principal components analysis**. The process takes the initial factor matrix and finds the variable with the largest variance shared by other variables (called the **communality**).

A Gram–Schmidt rotation is performed so that the first component is identical to this variable. The process continues as follows. From the remaining variables, the one with the largest shared variance is found, and a Gram–Schmidt rotation is performed so that the second component is identical to this second variable. The process continues until m variables are identified, where $m < p$, and p represents the total number of original variables. The goal in battery reduction analysis is to reduce the number of variables while at the same time explaining as much variance in the original variables as possible (see, for example, [2]). In multiple regression applications, the goal is to reduce the n-dimensional space (where n represents the number of observations) to a p-dimensional space ($p < n$, where p represents the number of independent variables).

References

[1] Cureton, E.E., & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale, pp. 97–98, 144–146.

[2] D'Agostino, R.B., Dukes, K.A., Massaro, J.M. & Zhang, Z. (1992). Data/variable reduction by principal components, battery reduction and variable clustering, in *Proceedings of the Fifth Annual Conference of the Northeast SAS Users Group*. SAS Institute, Cary, pp. 464–474.

(*See also* **Matrix Algebra**)

KIMBERLY A. DUKES

Graphical Displays

Statistical graphics are abstract pictures of numbers that represent either the data themselves or quantities derived from the data. Their value stems from the fact that it is often easier to extract information from well-chosen pictures than from sets of numbers. Although a table of numbers can be considered a graphic, generally, graphics are distinguished from tables as alternative representations used for different purposes. For small sets of numbers or for situations in which the exact values of the numbers are required, tables are better than graphics; for examining less local properties (like patterns) in large data sets, graphics are better. More generally, graphics are invaluable tools for recording and storing large data sets, analyzing (describing, exploring, and summarizing) data, enhancing the use of other statistical tools, communicating numerical information, decorating articles, reports, and so on. Graphics are not an end in themselves, but are constructed for specific purposes and should be judged in this context. The best graphics are those that show clearly and forcefully, but without distortion, the presence or absence of features of the data that are important to the substantive questions underlying the analysis.

Statistical graphics have been in use at least since Playfair [73] and are in a continuing state of evolution; see [14, 35, 37, 44, 91] for historical background. Stimulating discussions of the principles and aesthetics of graphic construction can be found in a variety of references including [22, 24, 37, 82, 83, 86].

Many of the important ideas concerning graphic construction can be summarized by the statement that good graphics are those that convey complex numerical information with clarity, precision, and efficiency. The key elements of good graphics can be identified as substance, statistics, and design. Graphics should be based on good statistics, both in the sense of being integrated into statistical analysis and in the choice of what to plot. For instance, we may variously consider plotting differences or ratios rather than the raw data, plotting transformed variables rather than the original variables, removing gross structures such as linear trends so that we can concentrate on finer details, partitioning data that contains clusters and exploring the clusters separately, augmenting plots with fitted curves and other statistical summaries or even by plotting summaries of the data (e.g. boxplots) rather than the data itself. When possible, statistical methods should be used to help assess whether a feature in a graphic is an artifact or not. The design aspects of a graphic (size, aspect ratio, choice of axes, and plotting symbols) should be appropriately chosen to reflect balance, proportion, and a sense of scale. Length and direction are the preferred basis for plotting symbols; shape is generally better than size for coding information though size is widely used. Color is a potentially effective encoding method but can be difficult to use and restrained gray scales may be a better alternative. Finally, it seems natural to require the visual impact of a pattern to be matched to importance in the context of the analysis.

One of the more elusive ideas in graphic construction is that graphics should be simple. At a basic level, complex, cluttered graphics are difficult to interpret and can be misleading because clutter can interfere with our perception. Clutter can be reduced by eliminating gridlines, choosing appropriate scales (including plotting characters, transformation, etc.), and reducing the data (such as by the use of plotting symbols as in a sunflower plot). Careful, accurate labeling should be used to reduce complexity but needs to be balanced against a possible increase in clutter. Simplicity of perception can be achieved by noting that straight lines are perceived more clearly than curves; therefore, reference curves are preferably straight lines as in quantile–quantile (Q–Q) plots or residual plots. The difficulty is that simplicity of perception and simplicity of interpretation can conflict. For example, simplicity of perception can be achieved by ensuring that deviations of equal magnitude have equal importance in a graphic. In a histogram, longer bars are more variable than smaller ones but the square root transformation can be used to produce a rootogram [90] in which the bars have equal variability, at the expense of increasing the difficulty of interpretation. Similarly, dependence can have unexpected effects and complicate the interpretation of a graphic. For example, the **order statistics** in a Q–Q plot are dependent, so apparent structure can be due to this dependence. An alternative is to base the plot on the spacings between order statistics but again the requirements of simplicity of perception and simplicity of interpretation compete. To complicate matters, interpretability depends on the training and experience of the observer.

Successful graphic construction (and data analysis generally) requires flexible strategies and a computational environment for implementing these strategies. Graphics capabilities are hardware dependent (i.e. they depend on the particular terminal or device being used) but a minimal requirement includes high-resolution (even color) screens and printers, a "real-time" graphic capability to support dynamic graphics, and the ability to interact directly with the displays. The simplest interactions we need to make involve a single window. These include rotating, interpolating, scaling, subsetting, marking, and identifying points as well as controlling dynamic graphics. It is also desirable to be able to implement multiwindow interactions such as linking windows and brushing (highlighting, downlighting, deleting, and labeling). We need high-level abstract computing languages that express manipulations in the way we think about them. It is convenient for a language to include inbuilt default settings so that we do not always have to set explicitly every possible graphical parameter, but it is vital to be able to adjust all of these parameters quickly and easily. The graphics in this article are all constructed in **S-PLUS** running in a workstation environment.

The article is organized into six broad sections: relationships between two variables; relationships between several variables; the display of one-dimensional data; data developing through time; multivariate data; and survival data. We elected to treat relationships between two variables before the familiar one-dimensional displays because we felt that practitioners in biostatistics and statistics in medicine and the health sciences are more naturally concerned with relationships *between* variables than with simple univariate data. (e.g. in many applications in medical statistics, there is a well-defined **response variable** that is to be modeled in terms of several **explanatory variables**.) One-dimensional displays then arise naturally through looking at topics such as **residual** analysis from modeling relationships between variables.

Relationships Between Two Variables

Scatterplots

The simplest, and one of the most powerful, graphical tools for describing the relationship between two variables is the *scatterplot*, which represents each pair of data values using (x, y) coordinates in a Cartesian plane. The "shape" of the scatterplot is used to describe the relationship between the two variables. Two elements of the "shape" of a scatterplot that are most useful in describing relationships between variables are measures of "location" and "spread" for the (x, y) data. For example, location might be measured as a line or a curve that runs through the bulk of the data while spread might be measured in terms of deviations of (x, y) points from the estimated location. The use of notions like location and spread to describe the relationship observed from a scatterplot is essentially a form of smoothing applied to the scatterplot, and the results of this smoothing can be added to the scatterplot through the addition of fitted curves. The resultant plot contains information about individual data points as well as summarizing the relationship between the variables through the fitted curve. A seminal paper describing the use of scatterplots (and other graphics) in data analysis is [4]. Other relevant references for the general use of scatterplots include [20, 24].

Scatterplot Smoothing

Numerous strategies have been developed for smoothing scatterplots, primarily for estimating the location of the relationship. One suggestion, due to Cleveland [22, 23], is *lowess* (Locally Weighted Scatterplot Smoothing), which smooths by averaging local straight line fits to the data; see also [20]. Alternatives to lowess include running (weighted) mean smoothing, running (weighted) median smoothing, and **spline smoothing**; see [78, 82, 88]. All of these methods depend on the selection of bandwidths that control how local the fitting methods are. Smoothing approaches to estimate spread are usually based on *residuals*, deviations of the data from the estimated location, and are often based on location-smoothing of scatterplots of absolute residuals versus fitted values; see below for a discussion of methods based on residuals. Figure 1 shows a simple scatterplot of brain weight (in grams) versus body weight (in kilograms) for 62 mammals. This graphic demonstrates that as body weight increases, brain weight tends to also increase, but the precise relationship between brain weight and body weight is difficult to describe. The "location" of the scatterplot exhibits significant curvature, and the points seem to be very compressed in the bottom left corner of the plot. The real question

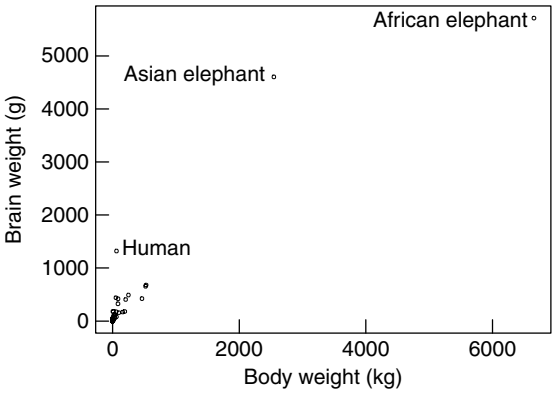

Figure 1 Scatterplot of brain weight versus bodyweight for 62 mammals. (Source: Weisberg [93]; excerpted from a larger study presented by Allison and Ciccheti [1]). Three "unusual" points are marked in the plot

arising from this basic graphic is: can it be improved to more precisely indicate the relationship between brain weight and body weight?

Use of Transformations to Straighten Scatterplots

One strategy often followed in presenting scatterplots is that of trying to find **transformations** of the original variables, so that a scatterplot involving the transformed variables displays a roughly linear relationship with a roughly constant spread of the response variable (y) for different values of the explanatory variable (x). Such strategies are desirable from a classical parametric modeling standpoint because they make the usual assumptions underlying parametric regression modeling more credible. They are also desirable graphically from the point of view of simplicity, visual appeal, and interpretability. It is important to keep in mind, of course, that the relationship between variables on the original scale is usually most important, and variables should not be transformed without thought for how a linear relationship on the transformed scale can be interpreted in terms of the original variables. A general discussion of the use of transformations in analyzing data is given by Mosteller & Tukey [71]. Broadly speaking, there are two approaches to transforming variables in order to understand their relationship: data-driven approaches and theoretical approaches. Data-driven approaches attempt to empirically straighten scatterplots. One data-driven method, suggested by Box and Cox [17], selects transformations from the class of **power transformations** to obtain the best linear fit to the data on the transformed scale. Once a transformation has been thus selected, scatterplots of the transformed variables can be employed to assess whether the transformation has been successful in linearizing the scatterplot. Theoretical approaches to transformations may be possible when prior information about the data can be used to help in selecting an appropriate transformation. For example, data arising from an exponential growth model typically yields a scatterplot that can be linearized by taking logs of the response variable. (Of course, it is critical to recognize that taking logs of the response variable affects not only its location but also its spread). Figure 2 shows how transformations can be used to effectively clarify the relationship between two variables. Shown is a plot of log brain weight versus log bodyweight for the data set described in the previous paragraph. There is a clear linear relationship between the logged variables. Points appearing unusual on the original scatterplot look less so on the transformed scale and the logged data also has much more uniform spread (along both sets of axes) than the original data. By positing a linear relationship between log brain weight and log body weight, it is possible to make clear the precise nature of the relationship between brain weight and body weight, a task rendered difficult using only the original scatterplot.

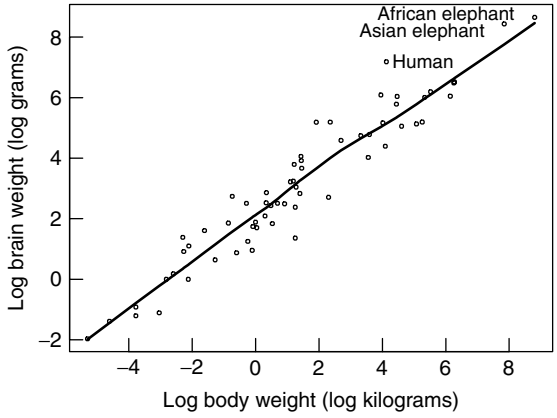

Figure 2 Scatterplot of log brain weight versus log body weight for 62 mammals. A lowess line is superimposed. Unusual points marked in Figure 1 are marked in this plot also for comparison

Relationships Between Several Variables

The Scatterplot Matrix

While the basic scatterplot is a satisfactory tool for describing the relationship between two variables, more sophisticated strategies become necessary when the goal is to explain a response variable in terms of several explanatory variables. A basic approach to this problem involves the use of a *scatterplot matrix*, an array of scatterplots relating every pair of variables for which we have data; see [20]. Individual panels of the scatterplot matrix can be smoothed, perhaps using a method like lowess, or with a fitted curve arising from a parametric model fit. The scatterplot matrix can be used to describe the marginal relationship between any pair of variables, but it is unable to provide information on joint relationships involving three or more variables. Figure 3 shows a scatterplot matrix for a data set from a study relating body fat to three explanatory variables, triceps skinfold thickness, thigh circumference, and midarm circumference. The scatterplot matrix shows that both triceps and thigh measurements have a positive relationship with body fat, but that midarm appears to have little relationship to body fat. Notably, triceps and thigh measurements appear highly correlated with one another.

Scatterplots Using Different Plotting Symbols

Basic scatterplots can be embellished to provide higher-dimensional information in several ways. One way is through the use of different plotting symbols on the scatterplot to yield information on other variables in the problem. For example, if one wished to plot information about three variables, one of which was **categorical** and the other two of which were measured continuously, one could plot a scatterplot involving the two continuous variables but the plotting symbol could reflect the level of the third variable. If all three variables were measured continuously, the magnitude of the third variable could be encoded as the length of a line segment that could be then used as a plotting symbol in a scatterplot of the other two variables. Higher-dimensional data could be accommodated by using higher-dimensional symbols: for example, the size of two variables could be conveniently encoded and plotted as the width and height of a rectangle. Other plotting symbols include: *metroglyphs*, circular symbols with rays extending from them [2]; stars where

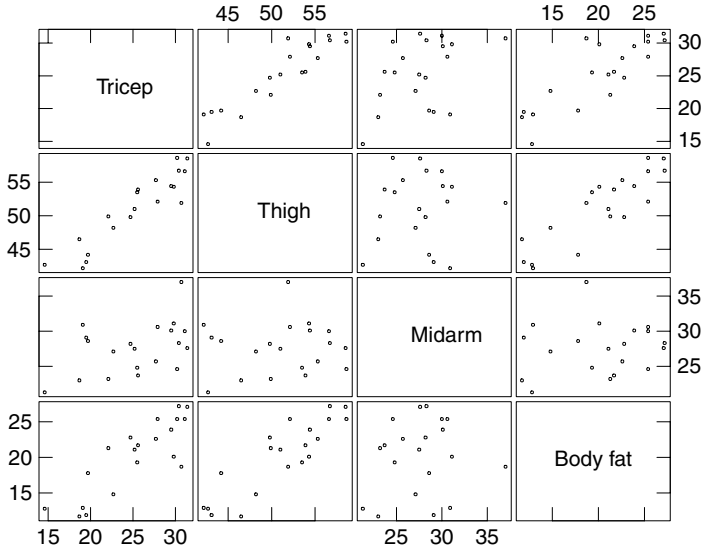

Figure 3 Scatterplot matrix for body fat data. (Source: Neter et al. [72, Chapter 8]). Each cell of the array of plots shows a simple scatterplot relating the pair of variables indicated in the corresponding row and column of the array. Response variable is body fat; explanatory variables are triceps skinfold thickness, thigh circumference, and midarm circumference. The data is based on a sample of 20 healthy females aged between 25 and 34 years

the number of rays reflects the number of variables encoded [45]; weather vane symbols [19]; and sunflowers [27]. Other ideas have been furnished by Hartigan [53], Bachi [8], Bertin [16], Everitt [42], and Tukey and Tukey [88]. An overview of the use of symbols in graphing data is given by Cleveland [24]. When choosing the type of symbol with which to encode information on several variables, a useful rule of thumb is to use a symbol having the same dimension as the number of variables it is encoding. Returning to the body fat data introduced in the previous paragraph, Figure 4 shows a scatterplot of body fat versus thigh circumference with midarm and triceps encoded as the width and height, respectively, of rectangles. Note that as thigh circumference increases, body fat tends to increase and the rectangles change from being shorter and flatter to taller and thinner, reflecting changes in the other variables.

Coplots

Another approach to presenting higher-dimensional data that uses simple scatterplots as its basis is the use of *coplots*. In a case where the data has two explanatory variables x_1 and x_2 and one response variable y, a set of coplots may consist of a sequence of scatterplots involving the response variable and one of the explanatory variables (say x_1), but points plotted within each scatterplot are those whose values of the second explanatory variable x_2 falls within some specified range. By using a sequence of (potentially overlapping) segments of x_2 that span its range, the sequence of scatterplots indicates how the relationship between y and x_1 changes as x_2 changes. If one of the variables is categorical, its levels serve as a natural set of values of the third variable for which to set up coplots. Geometrically, coplots can be thought of in terms of examining a sequence of slices parallel to the (x_1, y) plane rather than simply projecting all the data onto that plane, as occurs in a simple scatterplot. When there are more than two **covariates**, coplots can be constructed by conditioning on all but one of the covariates, although in this case care must be taken in organizing the order in which the covariates being conditioned on change so that the resultant sequence of coplots is readily interpretable. The idea of using conditioning as a tool for creating informative graphics was discussed by Tukey and Tukey [88], and the notion of coplots was introduced by Cleveland [25]. An extension of the coplot idea is the *trellis display* proposed by Becker et al. [11]. Trellis displays retain the simple idea from coplots of producing a sequence of graphics involving two variables after conditioning on the other variables in the problem. However, trellis displays make clever use of the way in which the plots are visually presented, and are not limited

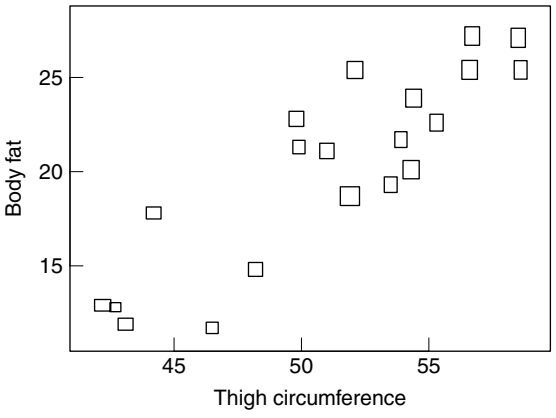

Figure 4 Scatterplot of body fat versus thigh using rectangular symbols to encode for midarm and tricep. The width of the rectangle encodes midarm; the height encodes tricep

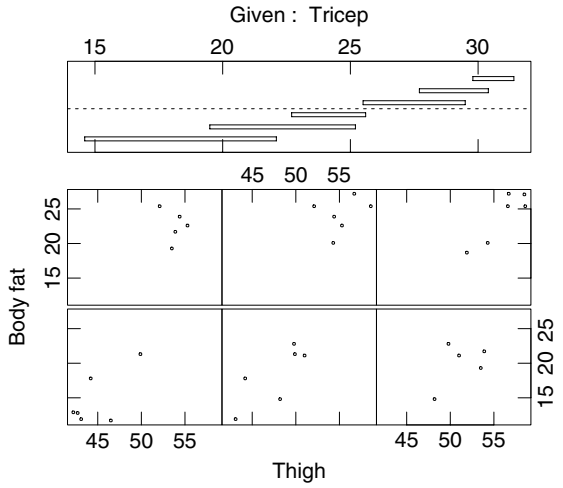

Figure 5 Coplot of body fat versus thigh given tricep. The top panel presents the range of tricep values for each of the coplot panels. The three lower plots correspond to segments of tricep below the dotted line in the top panel; the three higher plots correspond to segments above the dotted line. Note that segments of tricep are permitted to overlap, sometimes significantly, to allow for roughly equal numbers of points per coplot

to the use of two-dimensional scatterplots within the sequence of graphics displayed. An example of a coplot to investigate the relationship between body fat, thigh, and tricep for the body fat data is given in Figure 5.

Graphical Tools for Three-dimensional Data

Data in three dimensions lends itself particularly well to graphical analyses, and myriad tools are available for this case. For example, *perspective plots* and *contour plots* can be used to visualize relationships in three dimensions, and coplots are easiest to interpret in three dimensions. Extending these methods for higher-dimensional data is possible, for example, through the use of perspective coplots, or trellis displays whose panels comprise perspective plots or coplots. Figure 6 shows a perspective plot and a contour plot relating body fat to thigh and midarm circumferences for the body fat data. The plots clearly show that the response surface is fairly flat, suggesting that a linear model relating body fat to midarm and thigh may be appropriate.

One particularly promising approach to visualizing three- (and higher-) dimensional data is through the use of dynamic displays. Becker and Cleveland [10] and Cleveland and McGill [28] proposed two strategies, *brushing* and *spinning* displays, that can be used to graphically investigate three-dimensional data. Spinning displays amounts to producing a series of images of data point clouds, simulating motion as the data rotates about the origin. Structure within the data is then observed by interpreting the shape of the point cloud as it appears to move. Critical to the success of spinning displays is that the user can control the speed and direction of movement so that the data can be oriented into interesting and meaningful directions. Brushing displays can be used to embellish information obtained by spinning displays by highlighting (or "brushing") certain

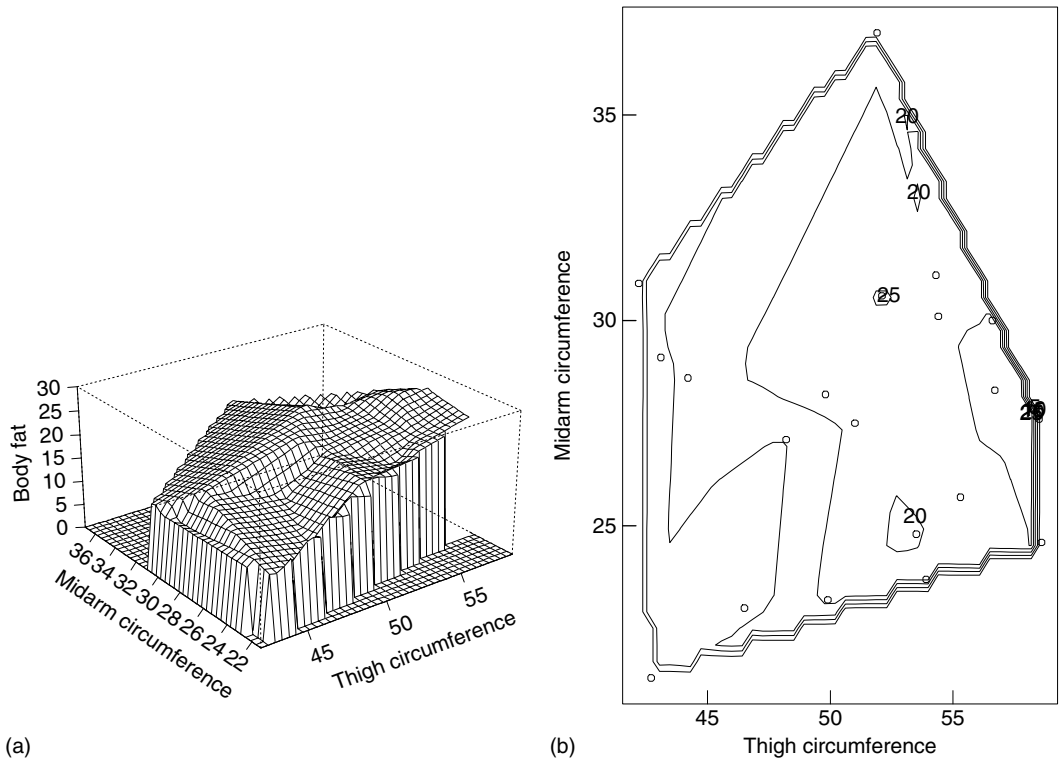

Figure 6 Plots specific to three-dimensional data. Plot (a) is a perspective plot relating body fat to the explanatory variables midarm circumference and thigh circumference. Plot (b) is a contour plot showing contours of the body fat surface projected onto the midarm–thigh plane

points or groups of points and watching how they individually affect the appearance of the point cloud. Brushing allows for the identification of outliers, influential points, and clustering within the data, as well as for other structures that may not be apparent from the simple scatterplot matrix. Dynamic graphics is a relatively new area of development in graphical data analysis, and is heavily reliant on powerful computing environments that allow the fluid movement of points in real time.

Exploring the Covariate Space

Apart from betraying relationships between the response variable and the various explanatory variables, the graphical tools described above also play an important role in finding potential problems for routine analysis in, and features of, the covariate space. Important features include the identification of outlying observations and influential points, identifying multicollinearity (*see* **Collinearity**) in the covariate space, and noting joint relationships between more than two variables that are not apparent from the two-dimensional plots in the scatterplot matrix. For example, coplots are useful tools for determining the marginal effect on the response of a certain explanatory variable (say x_1) after the effects of the other explanatory variables have been accounted for. If x_1 has a collinear relation with other explanatory variables in the data, the coplot of the response versus x_1 given the other explanatory variables exhibits no particular relationship between y and x_1 after the effects of the other variables has been accounted for, even though the original y versus x_1 scatterplot may have suggested a strong relationship between y and x_1. Note from Figure 5 that the relationship between body fat and thigh after the effect of tricep has been taken into account seems weaker than the marginal relationship observed from the scatterplot matrix. This discovery makes sense given that tricep and thigh appear highly correlated, and so contribute much the same information about body fat. Unfortunately, coplots can be difficult to interpret if the individual panels contain too few points. Coplots can also show the existence of joint relationships between more than two variables when the scatterplot matrix fails to show a strong marginal relationship between any pair of variables. The issues of multicollinearity and marginal versus joint relationships between variables are important because they directly impact our ability to describe the data in a compact way: multicollinearity because relationships between the explanatory variables prevent us saying how individual variables contribute to explaining the response; and complex joint relationships because they may not be obvious from usual, two- and three-dimensional graphical approaches. Where possible, a parsimonious model (that is, one that fits well with as few variables as possible) is the primary goal, and the reduction of the dimension of the covariate space made possible by visually examining its properties is an important part of any many-variable data analysis.

Outliers and Influential Points

Another important feature of a graphical exploration of data is the identification of unusual or aberrant data points. Here, it is important to distinguish between **outliers** (points that are far from the model fit), and *influential points* (points that do not seem consistent with the rest of the data, and whose presence dramatically affects the model fit; *see* **Diagnostics**). Outliers are usually easy to see in simple two-dimensional data sets, but can become very difficult to see in higher-dimensional data sets. In particular, the scatterplot matrix is not a reliable tool for finding outliers in three or more dimensions because the limited number of projections on which it gives information can miss unusual points in other directions in the data. Brushing and spinning tools can be highly effective for identifying outliers in three dimensions, but more general tools are needed for higher dimensions. *Residual plots*, discussed below, are often helpful for finding outliers with respect to a particular model fit, regardless of the dimension of the problem. In such cases, it is usually important that the models chosen be fit *robustly* so that the effect of outliers in the model fit is downweighted (*see* **Robustness**). Information about outliers can also be obtained by plotting the weight function used by the robust fitting technique and observing which points were downweighted in the analysis; see Figure 7(a). While robust fitting techniques can adequately deal with outliers, they may not be resistant to influential points. Influential points can be assessed through the use of *leverage plots* that display, for each point, a measure of how far the data point is from the "center" of the covariate space. Points far from the center of the data are then assessed as potentially influential. There is a considerable literature devoted

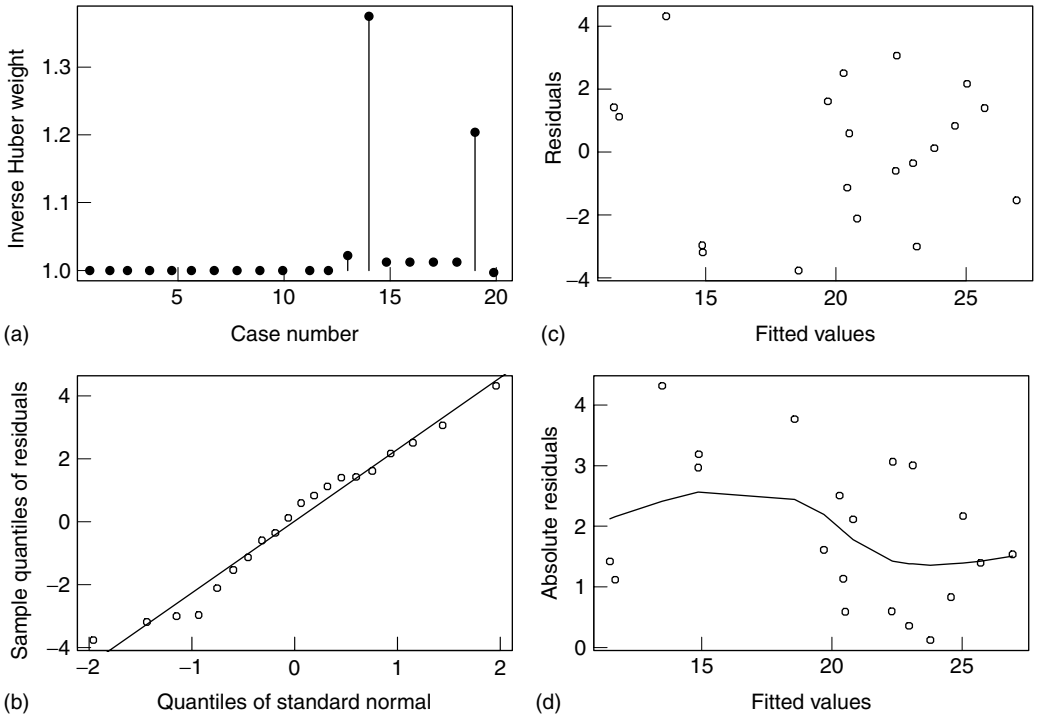

Figure 7 Residual diagnostic plots for the body fat data from a model relating body fat to all explanatory variables on the raw scale. The plots include: (a) a plot of robust regression weights for the model fit revealing data points 13, 14, and 19 as downweighted in the analysis (these points are potential outliers); (b) a normal quantile–quantile plot of the residual from the model fit showing the residuals as plausibly normal; (c) a plot of residuals versus fitted values; and (d) a plot of absolute residuals versus fitted values with a lowess line superimposed

to the subject of handling outliers in data analysis; see, for example, the monographs by Barnett and Lewis [9] and Hawkins [58], and the survey paper by Beckman and Cook [12]. Predominantly nongraphical approaches to measuring influence were discussed by Cook [29], Belsley et al. [13], Atkinson [7], and Cook and Weisberg [32].

Exploratory Methods

After an initial graphical exploration of the data has been carried out, it is usually desirable to model the data explicitly. The resultant model may be used to validate a theory about the process from which the data derived, or it may be used for prediction, or for some other purpose. Recently, several **exploratory** methods have been developed to suggest interesting directions in the data, and to suggest whether it is desirable to transform some or all of the variables in order to obtain a reasonable fit. Examples of these techniques include Additive VAriance Stabilized approach (AVAS), **Generalized Additive Models** (GAM), and **Projection Pursuit** Regression. The results of these methods are naturally interpreted graphically. The AVAS and GAM methods attempt to isolate appropriate transformations of both the response and explanatory variables to obtain the best additive fit of the transformed variables to the data. The AVAS method also attempts to transform the response variable to stabilize its spread. Scatterplots of the transformed AVAS variables against the original variables can be used to suggest appropriate transformations of the data for use in fitting formal parametric models. The AVAS method was introduced by Tibshirani [84], and the GAM method was proposed by Hastie and Tibshirani [55]; see also [54, 56, 83] and [57]. Projection pursuit regression, introduced by Friedman and Tukey [47] and discussed further by Friedman and Stuetzle [46], and Cook et al. [30, 31], attempts to find informative

projections in the covariate space, and then appropriate transformations of the projected data to obtain the best additive fit of the transformed projections to the response. Graphical exploration of the result of a projection pursuit fit can yield useful information about the effective dimension of the covariate space, although care must be taken not to overinterpret the results of projection pursuit regression.

Parametric Model Fitting and Plots Based on Residuals

Parametric approaches to modeling data assume a particular parametric form for the relationship between the response and explanatory variables, and then estimate the parameters of the relationship based on minimizing the distance between the fitted relationship and the original observations. Additional assumptions usually made in parametric modeling based on Gaussian models are that the errors associated with the model are additive, and have zero "location" and constant spread (*homoscedasticity; see* **Scedasticity**). For many applications, it is convenient to assume that the errors have the explicit structure of independent and identically distributed random, zero-mean, constant-variance normal variates. Many graphical tools for assessing the success or otherwise of the model fitting process are based on *residuals*, estimated deviations of the original observations from the model fit. The residuals estimate the unobservable errors described above. The model fit is represented by the *fitted values*, values of the estimated response for each of the observed covariate values. The *plot of residuals (or absolute residuals) versus fitted values* is a simple graphic often used to assess the validity of some of the error assumptions. If the assumptions are valid, this plot should exhibit a random pattern of residuals in a roughly uniform band about zero (homoscedasticity). Common deviations from this uniform pattern include patterns where the residuals seem to increase (or decrease) as fitted values increase (suggesting heteroscedasticity), or a strong relationship between the residuals and the fitted values (suggesting a deficiency in the model). Another useful graphic is a plot of residuals versus order, which can discover **serial correlation** among the residuals; see below for a discussion of dependent data. To assess the assumption of normality of residuals, a normal Q–Q plot is often used; a discussion of Q–Q plots is given later. Anscombe and Tukey [5] discuss the graphical use of residuals in regression diagnostics, and descriptions of standard residual plots are given in almost any text on basic regression methods; see, in particular, Cook and Weisberg [33]. Figure 7 shows a typical set of diagnostic plots based on residuals, there applied to a model for the body fat data. The four plots pictured are a plot of Huber weights, used to detect potential outliers in the data set; a plot of residuals versus fitted values and a plot of absolute residuals versus fitted values, useful for detecting outliers, assessing the error-variance assumption, as well as for assessing whether further modeling is desirable; and a Q–Q plot of residuals, used to assess the assumption that the error distribution is **normal**. The plots suggest no major concerns about the quality of the linear model fit.

Partial Residual Plots and Added Variable Plots

Plots based on residuals can also be used to assess which variables should enter the model, and in what form. Particular examples include *partial residual plots* and *added variable plots*. Partial residual plots were proposed by Ezekiel [43], and are described by Larsen and McLeary [68], and Wood [94] (who termed them component-plus-residual plots). Added variable plots were suggested by Gnanadesikan [49], and are described by Larsen and McCleary [67]. To describe partial residual plots, let $r_i, i = 1, \ldots, n$ denote residuals from the model fit, and let $\hat{\beta}_k$ be the estimated coefficient of the explanatory variable x_k in the model. A partial residual plot is a plot of $r_i + \hat{\beta}_k x_{ik}$ versus x_{ik}. If the model linear in x_k is appropriate, the partial residual plot for x_k should exhibit a linear trend; curvature in the partial residual plot suggests that x_k should be transformed before entering the model, or that a quadratic (or higher-order) term in x_k should be added to the model (*see* **Polynomial Regression**).

The added variable plot seeks to answer the question of whether a variable (say x_k) should be added to the model if other variables are already in the model. The graph plots the residuals of a model fitting the response to a set of explanatory variables not including x_k (representing that part of the response not adequately explained by those variables) versus the residuals of a model fitting x_k to the same reduced set of explanatory variables (representing that part of x_k not accounted for by the other explanatory variables). The resultant plot depicts the effective explanatory

power of x_k over y *after the other variables have been accounted for*. Provided a linear model involving x_k is appropriate and x_k is not **collinear** with the other variables, the plot should exhibit a linear trend with slope equal to the estimated regression coefficient of x_k in a linear model fit. Curvature in the additive variable plot suggests either that x_k needs to be transformed before entering the model, or that higher-order terms in x_k are warranted in the model. If the plot exhibits no particular structure and the original scatterplot matrix suggested a relationship between y and x_k, collinearity between x_k and the other explanatory variables is indicated. An example of an added variable plot applied to the body fat data is given in Figure 8. The question motivating this plot is that of whether thigh should enter a model for body fat if midarm is already present in the model (triceps have been removed from consideration because it is highly correlated with thigh). The strong linear trend of the resultant plot suggests that thigh should enter the model.

Display of One-dimensional Data

Diagnostic methods based on residuals include checks on the distributional assumptions made about the errors associated with the model fit. Hence, it is important to be able to describe one-dimensional data, such as the residuals, graphically. The simplest, and perhaps oldest [73], graphical display for one-dimensional data is the *histogram*, which divides the range of the data into bins and plots bars corresponding to each bin, the height of each bar reflecting the number of data points in the corresponding bin. Unfortunately, the way in which histograms depict the distribution of the data is somewhat arbitrary, depending heavily on the choice of bins and binwidths, with large binwidths tending to smooth over important features of the distribution, and small binwidths resulting in histograms that look too rough. Kernel-based smoothing (see, for example, Silverman [78]) can be used to provide a smooth **density estimate**, but many of the same issues of choosing an appropriate bandwidth arise. Examples of histograms appear in Figure 9 for a data set examining the effect of ozone (a component of smog) on body weight. *Stem-and-leaf plots* (see [89] for a discussion) are a variant on histograms that combine the features of a graphic and a table in that the original data values are explicitly shown in the display as a "stem" and a "leaf" for each value. The stems determine a set of bins into which the leaves are sorted, and the resulting list of leaves for each stem resembles a bar in a histogram. Tukey [87] developed the *boxplot* display, based on the five-number summary (minimum, first quartile, median, third quartile, maximum) of the data. The boxplot represents the middle half of the data (first to third quartiles) by a rectangle (box) with the median marked within, with *whiskers* extending from the ends of the box to the extremes of the data

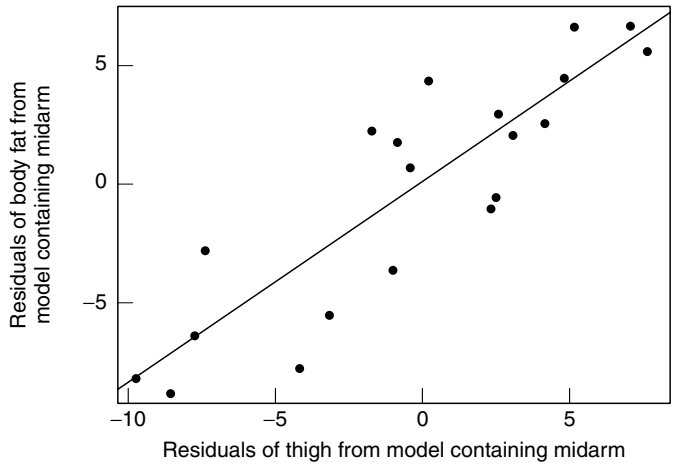

Figure 8 Added variable plot for thigh given the presence of midarm in the model for body fat. Plotted are the residuals from a model relating body fat to midarm versus the residuals from a model relating thigh to midarm. A robust regression line relating the two sets of residuals is superimposed, suggesting that a linear model incorporating thigh is appropriate

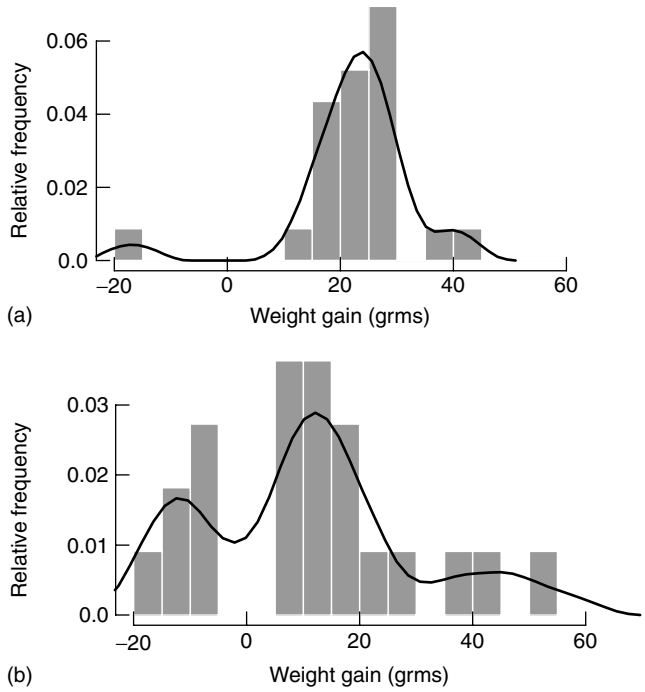

Figure 9 Histograms of weight loss in grams for control and ozone groups of rats from an experiment to measure the effects of ozone, a component of smog. (Source: Doksum and Sievers [41]). The control group contained 23 rats, the ozone group 22. All rats were around 70 days old, and weight gain was recorded after 7 days of exposure. Both histograms are drawn on the density scale, and smooth kernel density estimates are superimposed. The histograms are arranged vertically to facilitate comparisons between the two groups

or to one and a half times the interquartile range of the data, whichever is closer.

Each of the aforementioned displays tries to answer the question of how the data is distributed by showing what the data distribution "looks like", but they don't deal with the issue of how the data distribution compares with some theoretical distribution. This issue is important, particularly for examining residuals from a model fit, because it is often assumed that these residuals arise from a normal distribution with zero mean and constant variance. The most commonly used graphic to address this issue is the *normal Q–Q plot*; see Wilk & Gnanadesikan [92] for a general description. A similar graphical display is the P–P (probability–probability) plot. The normal Q–Q plot plots theoretical quantiles from a standard normal distribution against the empirical quantiles from the data (*see* **Normal Scores**). If the resultant plot appears linear, the data is consistent with the assumption of normality; departures from linearity, such as a characteristic "S" shape or concavity, suggest nonnormality. If nonnormality is indicated for a set of residuals, it may be useful to transform the data to achieve plausibly normal residuals, bearing in mind that the transformation affects other features of the data. An example of a normal Q–Q plot is given in Figure 7(c), applied there to the residuals of a model for the body fat data. Quantile–quantile plots can also be constructed against reference distributions other than the normal.

Comparison

Displays for one-dimensional data can often be used effectively for comparing data distributions. Key to the success of such ventures are notions of scale and placement: first, it is critical that the data sets to be compared are on roughly the same scale (e.g. they should be measured in the same units); and second, in combining graphical elements for purposes of comparison, it is useful to utilize a common set of basic

features (axes, orientation, etc.) for all elements, and, if possible, to place the individual components within the graphic to facilitate easiest comparison (e.g. side-by-side boxplots sharing a common set of axes form a more potent graphic for comparison than do side-by-side boxplots drawn on different axes, or boxplots aligned vertically). Other examples of comparative graphics include histograms using the same axes, arranged vertically, back-to-back stem-and-leaf plots (sharing a common stem), and h-plots; see Figure 9. Quantile–quantile plots relating the empirical quantiles from each of two data sets can also be used to compare the distributions of the two data sets. The histograms in Figure 9 facilitate easy comparison of the control and ozone groups of rats by virtue of their vertical arrangement. Figure 10 displays two comparative displays of one-dimensional data using side-by-side boxplots; the use of a log transformation in plot (b) to symmetrize the data distributions allows for an easier comparison between the two groups. Of course, while the use of a log transformation allows easier comparison between the two groups, it also makes for a more difficult interpretation. An important feature of graphic construction is simplicity, but, unfortunately, it is often the case in the construction of good statistical graphics that simplicity of perception and simplicity of interpretation conflict.

Graphical displays for one-dimensional data are often encountered in the popular media, but "presentation" graphics encountered in that setting, such as bar charts, line charts, and pie charts are not terribly effective tools for communicating numerical information. Pie charts, for example, usually summarize only a handful of numbers that would be more effectively presented in a small table. Tufte [86] puts this opinion rather bluntly: "A table is nearly always better than a dumb pie chart ... Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used". Bar charts differ from histograms in that the bars need not be ordered along the horizontal axis. However, bar charts published in many popular newspapers and magazines employ highly stylized bars (e.g. bars shaped like people) that distort the visual perception of the bars. Needless to say, many common embellishments of common presentation graphics such as three-dimensional bar charts, and three-dimensional or exploded pie charts, should be avoided altogether as they introduce redundant dimensions into the graphical display. These embellishments can also introduce unusual perspectives into the graphic, unintentionally (or, occasionally, intentionally!) causing the viewer to misinterpret the numerical information being portrayed. Perhaps the only graphic routinely used in the popular media that is also a good statistical graphic is the time series plot, although this display can also be ruined by careless use of shading, decoration, or other presentation errors. To faithfully record the various misuses and abuses of graphical displays in the popular media (and, indeed, in

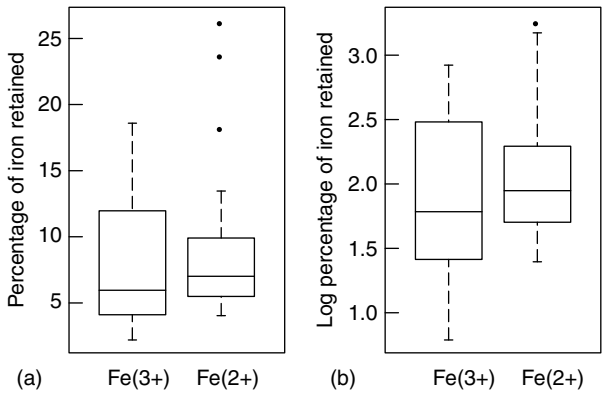

Figure 10 Side-by-side boxplots aid in comparison. Data obtained from an experiment to determine whether two forms of iron (Fe2+ and Fe3+) are retained differently (this feature impacts their usefulness as dietary supplements). The data relates to two groups of 18 mice each given a form of iron orally in the concentration 1.2 millimolar and the result is measured as percentage of iron retained. (Source: Rice [75, Chapter 11]) Plot (a) compares raw percentage retention for the two forms of iron. The boxplots indicate both distributions have similar features but are heavily skewed. Plot (b) shows boxplots for the logged scores, whose distributions are closer to symmetric

many scientific papers!) could fill many volumes. Tufte [86] has done an extraordinary job of describing the strengths as well as the pitfalls of graphical displays for visualizing quantitative information; his delightful book is filled with examples and discussions of the good, the bad, and (literally) the ugly of statistical graphics.

Data Developing in Time

Data in which development through time is an important aspect can be classified conveniently into data which consists of a few relatively long series of observations which may be dependent (as in traditional **time series** analysis) or into data which consists of a large number of relatively short series which are typically independent (as in repeated measures problems). In either case, analysis is simplest when the data in each series are collected at the same number of equally spaced time points, becoming more complicated as the lengths of the series vary and as the observation time points become irregularly spaced (*see* **Graphical Presentation of Longitudinal Data**).

Whether we have a few relatively long series or a large number of relatively short series, the most common representation of the data is by means of a *sequence plot* (see the top panel of Figure 11), which is a simple scatterplot of the data against time. Adjacent points in the same series are often connected by linear segments to create a piecewise linear curve which highlights the basic time series structure and distinguishes the different series. However, if we have a very large number of series, the resulting plot can be very difficult to interpret. For this kind of situation, Jones and Rice [61] proposed that we connect

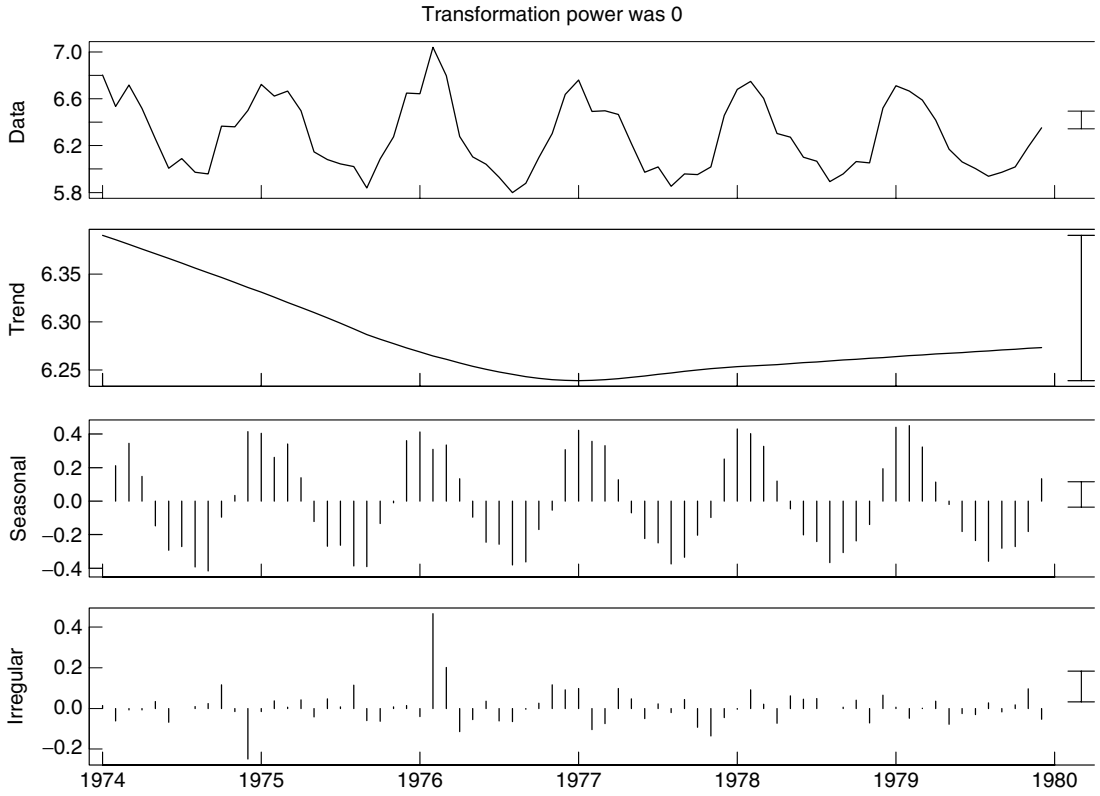

Figure 11 SABL Decomposition for United Kingdom lung disease data. Data are monthly UK deaths of females from bronchitis, emphysema and asthma from 1974–1979. (Source: Diggle [39, p. 238], data from Appleton) The plot produced by SABL decomposes the log number of deaths, given in the top panel, into trend, seasonal, and irregular components. The vertical bar at the right of each plotted component indicates the relative scales of the components

the points for only selected series in the data. While it is possible to choose the subset at random, it seems preferable to select quantiles from curves ordered with respect to some characteristic of the problem of interest. Diggle et al. [40, p. 38] have suggested ordering curves by means of robust measures of the level of the series, or by variability within series, or by trend or correlation between successive observations within the series, and then connecting the points of the selected series. Series that are grouped together (by virtue of receiving a common treatment, for instance) can be distinguished by being plotted separately, by being plotted with different lines or symbols or by being represented by common summary series. Such a plot is depicted in Figure 12 for data tracking pituitary growth in children of various ages, each series representing a single child.

Much of the analysis of traditional time series (which focuses on a few relatively long series of data) is based on the idea that on some scale determined by a function g, a series of data Y_t can be decomposed into three components as

$$g(Y_t) = T_t + S_t + E_t, \tag{1}$$

where T_t is the trend, S_t is the **seasonal** and E_t is the irregular component. The trend is supposed to describe the long run behavior of the series, so can be thought of as a smooth curve; the seasonal component is supposed to capture periodic variation about the trend, so can be thought of as a periodic function; and the irregular component is supposed to represent whatever is in the series that is not described by the trend and seasonal components. While the components of this decomposition are not unambiguously defined and the resulting decomposition is not therefore unique, it can be quite useful. Once the trend and seasonal components have been modeled and estimated, we can plot them and the irregular component against time. Further possibilities are to plot the detrended series (that is, the transformed data minus the estimated trend), the deseasonalized series (that is, the transformed data minus the estimated seasonal component) or the detrended and deseasonalized series that is simply the irregular component against time. These plots can be interpreted as plots of different kinds of residuals against time. The different components can be modeled parametrically or nonparametrically, and there is a substantial literature on the different approaches. As an illustration of the kind of graphic produced, the decomposition of a series obtained from SABL (see [26]) is shown in Figure 11 for data on mortality from lung disease. The decomposition shows a strong seasonal effect, with most deaths occurring in winter and fewest in summer. This effect can be observed clearly through the use of a *monthplot*, a plot of the monthly subseries for the data. Typically, a monthly average is also plotted; see Figure 13.

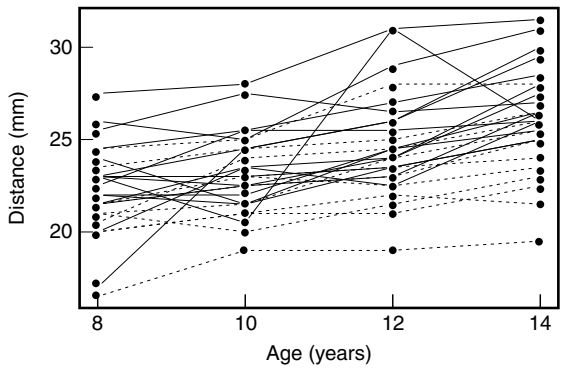

Figure 12 Sequence plots for several short series of pituitary data. Data measured are the distances (mm) between the center of the pituitary to the pteryomaxillary fissure in girls and boys aged 8, 10, 12, and 14 years. (Source: Potthoff and Roy [74]) The curves for boys are represented by solid lines and those for girls by dotted lines. The graphic shows that the distance generally increases with age, is generally larger in boys than girls, and that there is some tracking (children with large or small distances initially continue to have large or small distances throughout the study)

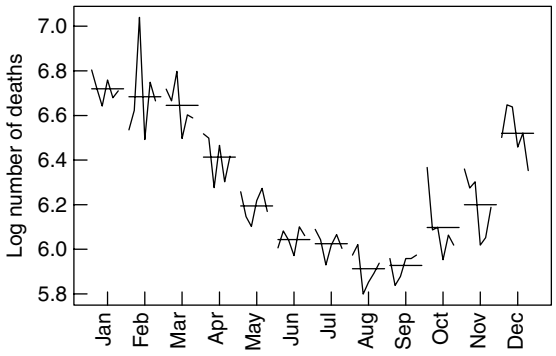

Figure 13 Monthplot of the log number of UK deaths of females from bronchitis, emphysema, and asthma from 1974–1979. The plot shows the six-year subseries for each month and the subseries mean. This plot confirms the strong seasonal effect

It can be important in dealing with time series data to keep in mind the dictum that time is not a real explanatory variable, and that we are often interested in modeling the relationship between the series and more interesting explanatory variables. Naturally, many of the methods developed for exploring relationships between variables are of use in this kind of problem and may be applied either to the raw data (in which case time effects may be explored in the residuals) or to residuals from which time effects have been removed.

An important feature of data evolving in time is that observations that are temporally close within a series are typically dependent. Thus, methods that explore the dependence between observations are important in analyzing time series. Once the effects of trend, seasonal, and other explanatory variables have been accounted for, we can explore the dependence structure remaining in the irregular or residual series, which we here denote Z_{it}. If the observations were made at equally spaced time points, we can consider plotting Z_{it} against $Z_{i,t-1}, t = 1, \ldots, n-1$, to detect first-order (lag 1) dependence. When the data consist of a large number of short, independent series of roughly equal length, we can go further and construct a scatterplot matrix in which Z_{it} is plotted against $Z_{i,t-1}$ (lag 1), $Z_{i,t-2}$ (lag 2), and so on, for $i = 1, \ldots, k$. If the scatter plots are in the form of bivariate Gaussian ellipses (*see* **Bivariate Normal Distribution**), we often observe a reduction in the strength of the relationship (i.e. dependence) as the temporal separation (represented by the lag) increases.

If the data in the series are Gaussian-distributed, we can summarize the dependence structure by various **correlation** coefficients. Furthermore, if the Z_{it} have constant mean and variance, and if the correlation between Z_{it} and $Z_{i,t+h}$ depends only on their temporal separation h, the series is said to be weakly stationary and we can readily estimate the **autocorrelation function** (ACF) at h by the sample correlation between observations lag h apart. Explicitly, the sample autocorrelation function is $\hat{\rho}_i(h) = \hat{\gamma}_i(h)/\hat{\gamma}_i(0)$, where

$$\hat{\gamma}_i(h) = \frac{1}{n} \sum_{t=h+1}^{n} (Z_{it} - \overline{Z}_i)(Z_{i,t-h} - \overline{Z}_i) \qquad (2)$$

is the autocovariance function. An *autocorrelation plot* or correlogram (Figure 14) is a plot of the sample autocorrelation function $\hat{\rho}_i$ against h. The *partial autocorrelation plot* is a closely related plot in which the partial autocorrelation function (PACF) is plotted against lag. These plots can be used to identify autoregressive-moving average (**ARMA**) models for the residual series; see, for example, [18]. Dependence between pairs of series can be explored similarly by the *cross-correlation plot*, which is a plot of the correlation between the observation at t in the one series and the residual at $t + h$ in the other against h; again, see [18] for details. If we have a large number of independent series which are assumed to have the same within-series dependence structure, we can average the within series estimates $\hat{\gamma}_i$ across series to obtain a better estimate of the common autocovariance function.

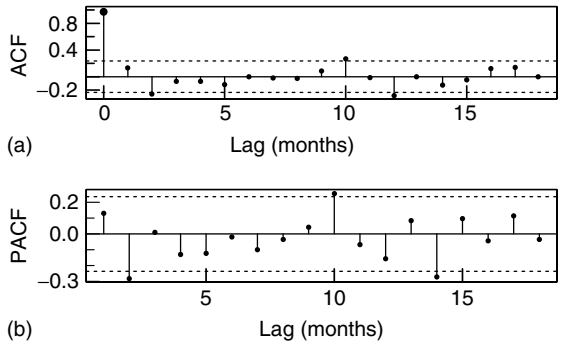

Figure 14 ACF and PACF of the detrended and deseasonalized (irregular) series of log UK deaths of females from bronchitis, emphysema, and asthma from 1974–1979. Dotted reference lines are plotted at height $\pm 2/\sqrt{n}$. These plots suggest that the observations in the irregular series are uncorrelated

For irregularly spaced data, the autocorrelation is more difficult to estimate unless we are prepared to round the observation times. An alternative function which describes the dependence structure and is easy to estimate even when we have irregular observation times is the **variogram** [62], originally referred to as the mean semi-squared difference curve or the serial variation curve. The *sample variogram* (Figure 15) is a scatterplot of the half-squared distances $0.5(Z_{it} - Z_{is})^2$ against the corresponding time differences $t - s, t \geq s$. The scatterplot is often usefully enhanced by scatterplot smoothing, which provides an estimate $\hat{\eta}_i(h)$ of the variogram at lag h and by the representation of the process variance estimated by $\hat{\sigma}_i^2$, the mean of the half-squared distances $0.5(Z_{it} - Z_{is})^2$ with $t \neq s$, by a horizontal

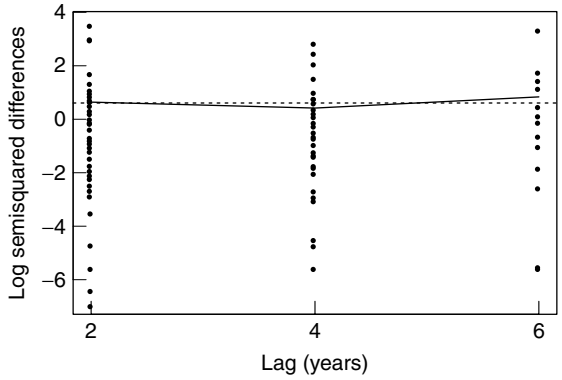

Figure 15 Log variogram of the pituitary data after removing the gender means at each time point. The dotted line represents the log process variance and the solid line, a piecewise linear estimate of the log variogram. The plot shows no evidence of correlation in the residuals

line. When the series are weakly stationary, the autocorrelation function at lag h can be estimated from $\hat{\eta}_i(h)$ by

$$\tilde{\rho}_i(h) = \frac{1 - \hat{\eta}_i(h)}{\hat{\sigma}_i^2}. \tag{3}$$

This relationship shows that $\hat{\eta}_i(h)$ increasing as h increases represents decreasing autocorrelation as h increases. Again, if we have a large number of independent series which are assumed to have the same within series dependence structure, we can plot the half-squared distances over $t \geq s$ and $i = 1, \ldots, k$ to obtain better estimates of the common variogram.

The issues that arise in the exploration of temporal dependence also arise in the exploration of spatial dependence, although this latter situation is rather more complicated. The basic graphical display is the *map* (the analog of the sequence plot) with superimposed data (*see* **Statistical Map**). Since at the very least we have three dimensions to display, surfaces, contour plots, spinning plots, and other methods for displaying high-dimensional data can be useful. The variogram is widely used to explore spatial dependence structure. In the simplest cases, lagged distances irrespective of direction can be plotted on the x-axis but in general, we may need to consider displaying separately variograms corresponding to different directions in space.

Multivariate Data

The use of graphics to explore high-dimensional data in the context of our having identified a single response variable that we are interested in relating to the remaining variables has been discussed earlier. The specification of a single response variable (and the inherent asymmetry it introduces into the analysis) determines many aspects of the subsequent graphical and analytical analyses. However, we can also be interested in looking for structure or patterns in the data (including looking for relationships between variables as well as outliers and clusters in the data) without having specified a response variable and a different set of graphics can be implemented for this purpose (*see* **Multivariate Graphics**).

We have already discussed a number of graphical techniques that can be useful in the preliminary examination of multivariate data. As alternative approaches to these, we can consider representing each independent vector of observations by a symbol, the features of which represent different variables. For example, Chernoff [21] suggested representing each observation by a face, the characteristics of which represent up to 18 variables. However, the fact that interpretation depends on how the variables are assigned to the facial characteristics and that it is difficult to examine simultaneously a large number of faces means that this graphic is probably more useful as an amusing way to present results than as a practical tool for data analysis. Other suggestions include using stars, weathervanes, parallel coordinate plots (each vector observation is plotted against $(1, 2, \ldots, p)$ and the points connected by lines), and Andrews' plots [3].

Unfortunately, experience suggests that considerable luck is needed to find structure using the above techniques. It is apparent that to get the most out of graphical methods, we need to supplement them by computations that simplify the task of searching for structure, primarily by *reducing the effective dimension of the data*. We will consider several techniques that may be useful for this task.

Projections

Pairwise scatterplots examine two-dimensional projections of the data onto the planes defined by pairs of axes in the coordinate system. Low-dimensional projections provide one of the most effective methods for examining points in high-dimensional spaces,

but there is no good reason to expect projections of the data onto the planes defined by pairs of axes in the coordinate system to reveal important structure and/or relationships in the data. In some cases, such as when we use a spinning display to examine three-dimensional data, it is easy to examine a large number of two-dimensional projections fairly quickly, but in general it is important to have available analytic methods for choosing revealing projections automatically.

The classical approach rotates the data to a new coordinate system called the **principal components** [59] which is determined by the directions of greatest variability in the data and then enables us to examine two-dimensional projections in this coordinate system by constructing a scatterplot matrix of the principal components. In order to achieve a reduction in dimension, it is usual to study only those principal components which make a substantial contribution to the variance by including principal components until the proportion of variance explained is above 80 or 90%, say. Hopefully, this strategy will result in the need to study only two or three variables rather than the original number. Outliers can also sometimes be detected in plots of the last two principal components; so as a diagnostic tool, we often also plot the last two principal components.

It is obvious from the fact that principal components analysis finds the projections of maximum variance that the scale of the data determines the result of the analysis. Even when the variables are on the same scale, there can be large differences in the variability of the variables, so some authors recommend standardizing the data by the standard deviations of the variables. This is equivalent to applying the analysis to the correlation matrix rather than the variance matrix. Working with the correlation matrix does not remove the scaling problem because it is simply another arbitrary scale. This strategy is not generally recommended because it is inappropriate when the variables are not of equal importance. Principal components analysis is also sensitive to the choice of the variables included in the data. It is obvious that omitting variables has an impact on an analysis, but it is less obvious that including redundant variables has an impact on both the last and the first principal components. In particular, linear relationships in the data can increase the weight given to a variable in the principal components. Finally, when a set of principal components has nearly equal standard deviations, they are arbitrary and cannot be interpreted in any meaningful way.

Projection Pursuit [47, 66] and *Grand Tours* [6] are modern alternatives to principal components analysis. Projection Pursuit defines revealing projections not in terms of variability but in terms of departure from normality. Thus projection pursuit tries to find the least ellipsoidal two-dimensional projections of the data. Grand Tours work by generating two points at random on the p-dimensional unit sphere and using these to generate a plane. A sequence of rotations is then applied to move the first point into the second while keeping the orthogonal complement of the plane fixed. The process is then repeated with another pair of points. The idea is that the resulting sequence of projections rapidly becomes dense among all possible projections, so selected revealing projections from the sequence can be examined for structure.

Biplots

Biplots [48] provide a graphical description of relationships among the observations (or rows), and the relationships among the variables (or columns) in a data set. It is based on the singular value decomposition (*see* **Matrix Computations**) of the mean-centered data matrix Z (so the columns of Z have mean zero), which can be written

$$Z = L\Delta M^T, \tag{4}$$

where L is an $n \times r$ matrix of rank r such that $L^T L = I_r$, M is a $p \times r$ matrix of rank r such that $M^T M = I_r$ and $\Delta = \text{diag}(\delta_1, \ldots, \delta_r)$ with $\delta_1 \geq \cdots \geq \delta_r > 0$ as the positive square roots of the nonzero **eigenvalues** of $Z^T Z$. A rank 2 approximation to Z is obtained by

$$Z_{(2)} = L_{(2)} \Delta_{(2)} M_{(2)}^T, \tag{5}$$

where $L_{(2)}$ represents the first two columns of L, $M_{(2)}$ represents the first two columns of M, and $\Delta_{(2)} = \text{diag}(\delta_1, \delta_2)$. The quality of this rank 2 approximation can be measured by $(\delta_1^2 + \delta_2^2)/\sum_{k=1}^r \delta_k^2$. See [51] for a detailed discussion of the statistical uses of the singular value decomposition. The biplot is the plot of the n rows of $G = (n-1)^{1/2} L_{(2)}$ and the p rows of $H = (n-1)^{-1/2} M_{(2)} \Delta_{(2)}$ represented as vectors. The plot of the p rows of H alone is called the h-plot [34].

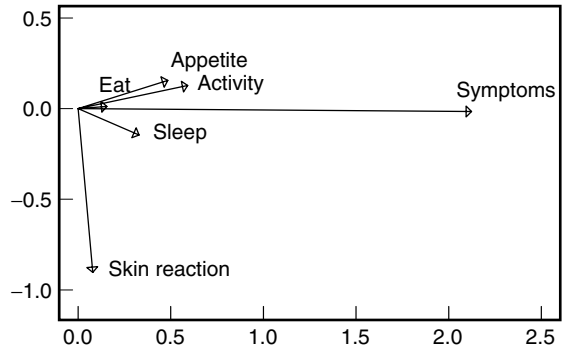

Figure 16 *h*-plot of the radiotherapy data. The data consist of average ratings over the course of radiotherapy of the number of symptoms (such as sore throat or nausea), amount of activity (1–5 scale), amount of sleep (1–5 scale), amount of food (1–3 scale), appetite (1–5 scale), and skin reaction (0–3 scale). (Source: Johnson and Wichern [60] from Tealey). The plot shows that the number of symptoms is far more variable than the other measures. The directions of arrows in the plot shows that the variables are highly correlated with the exception of skin reaction, which appears uncorrelated with the other variables

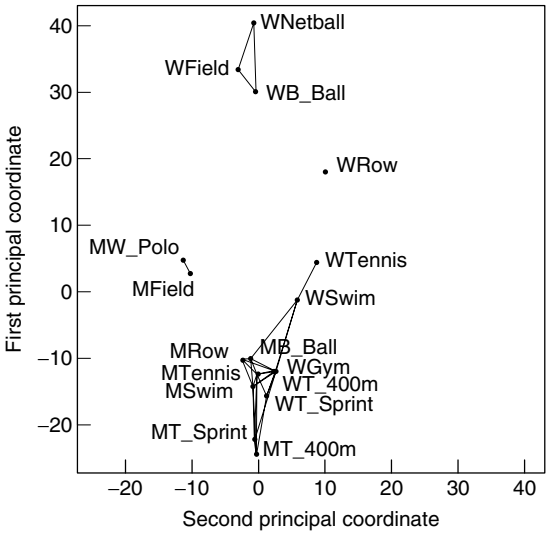

Figure 17 An ordination based on multidimensional scaling for the Australian Institute of Sport data (Source: Cook and Weisberg [33, p. 98] on data supplied by Telford and Cunningham). The data were physiological observations made on 202 different athletes of both sexes participating in 10 different sports (men's and women's basketball, swimming, rowing, track (400 meters and sprint), field, tennis, women's gymnastics and netball, and men's water polo) Classic multidimensional scaling based on Euclidean distances was used. The ordination suggests some clear clusters of sports, as indicated by the linked points

An example of an *h*-plot is shown in Figure 16 for data on cancer patients undergoing radiotherapy.

The distance between the points represented by two rows of G represents the **Mahalanobis distance** between the corresponding observations, the length of the vector represented by a row of H represents the standard deviation of the corresponding variable, the distance between the points represented by two rows of H is the sample variance of the difference between the corresponding variables, and the cosine of the angle between the vectors represented by two rows of H is the correlation between the corresponding variables. Thus the *h*-plot provides a visual representation of the covariance structure between variables and, additionally, the biplot contains information about the distance between observations. The matrix H from which the *h*-plot is constructed can be obtained directly from the principal components analysis of the covariance matrix of the data so *h*-plots can be viewed as an alternative representation of the results of a principal components analysis.

Ordination

Ordination methods attempt to arrange n data points in a high-dimensional space with proximity relationships represented by an $n \times n$ dissimilarity matrix D in a low (usually 2) dimensional space (*see* **Similarity, Dissimilarity, and Distance Measure**) by constructing an $n \times 2$ matrix Y such that the matrix of Euclidean distances between the points in Y is approximately the same as D. The columns of Y are referred to as **principal coordinates** and the plot of the first two principal coordinates is referred to as an *ordination*. An example of an ordination is given in Figure 17 for data from the Australian Institute of Sport. If the dissimilarity matrix D is determined directly by subjective assignment of dissimilarities, the ordination is a geometric representation of the proximity relationships between the objects. Alternatively, if the dissimilarity matrix D is calculated from a data matrix Z, ordination methods yield a lower dimensional representation of Z with an associated Euclidean distance matrix that is close to D. It is obvious that we can change the location and rotate the orientation of the ordination without changing the Euclidean distances between the points so that an ordination is not unique.

Multidimensional Scaling can be treated either as synonymous with ordination or as a particular ordination method. The subtle differences in perspective between these views are not particularly important in a graphical context, so we will treat them as synonymous. There are two types of multidimensional scaling depending on the nature of the proximity data:

i) Classical Metric Multidimensional Scaling [76, 85] or Principal Coordinates Analysis [50] which uses the magnitude of the proximities; and
ii) Nonmetric Multidimensional Scaling [65, 77], which uses the **ranks** of the proximities.

Nonmetric scaling is intended to facilitate subjective proximity assignments because it is usually easier to order proximities than to assign exact numerical values. However, if the ordering of the proximities is more important than the actual values, we may decide to use nonmetric scaling, even though the proximity values are available.

The two-dimensional representation of the dissimilarity structure may give a misleading impression of the actual distances between points because points that appear close in the first two principal coordinates may be far apart in higher dimensions. This is particularly likely when the first two principal coordinates fail to explain D. In such cases, the plot of the first two principal coordinates can be improved by linking units whose actual distance apart is less than some user-specified value; see Figure 17. It is of interest to vary the distance apart we specify to see how the picture changes. This method is useful for detecting one-dimensional structure in the data. If we obtain a two-dimensional representation and join points with dissimilarities below a user-specified value, we may obtain a horseshoe shape. This result suggests that the two-dimensional configuration is almost a one-dimensional configuration bent into a horseshoe shape and that ordering along the horseshoe curve may give a reasonable one-dimensional ordering of the data. This discovery is particularly relevant if we are seeking a one-dimensional ordering of the data (known as *seriation*).

It may be that in other problems we achieve a better ordination by using more than two principal coordinates. In general, we can choose the number of principal coordinates to consider by examining the proportion of the dissimilarities in D explained by the principal coordinates in the same way as we examine variance in principal components analysis (although we need to be careful of the fact that the eigenvalues can be negative).

One of the attractions of ordination methods is that they contain as special cases not only principal components analysis, but also a number of other classical multivariate techniques. These analyses are obtained by making particular choices of D. The most important relationship is that metric scaling based on the Euclidean distances between (mean-centered) points yields as principal coordinates the principal components. That is, metric scaling with Euclidean distances is principal components analysis. This suggests that the discussion of the issues underlying the application of principal components applies also to metric scaling. Principal coordinates analysis also includes **correspondence analysis** [15], linear **discriminant analysis**, and canonical variates analysis (*see* **Canonical Correlation**) as special cases. The important point is that ordination methods allow very general choices of D and so can be used when we have large sparse data matrices or subjectively assigned dissimilarities.

Minimum Spanning Trees

A spanning tree is a set of straight-line segments joining pairs of points such that there are no closed loops, each point is visited by at least one line segment, and every point is connected to every other point either directly or through a chain of other intermediary points. The length of a spanning tree is the sum of the lengths of the line segments so the minimum spanning tree is the spanning tree with shortest length [52].

One use of the minimum spanning tree is to assess the amount of distortion incurred in reducing the dimension of the data to, say, two dimensions. The idea is to incorporate the links in the minimum spanning tree into the two-dimensional scatterplot and check that the patterns of relative closeness described by the minimum spanning tree are preserved. When substantial distortions in the patterns of relative closeness occur, the structure in the two-dimensional representation of the data may not represent the same structure in the original data set.

Cluster Analysis

Although clusters can be detected by ordination methods, there are a number of techniques available for the specific purpose of detecting clusters in

data (*see* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**). The major difficulty with these techniques is that it is extremely difficult to define a cluster. It follows that it is difficult to specify the number of clusters in a data set and that the objectives of the analysis may be difficult to achieve.

Although there is a huge literature on clustering, there are basically three types of procedures: partitioning; overlapping clusters; and hierarchical clustering. From a graphical standpoint, hierarchical clustering is the most interesting method. Here, clusters are grouped within larger clusters to form a tree. The simplest techniques, called *agglomerative methods*, start by regarding each object as a cluster and proceed by merging near objects. Divisive methods work in the opposite direction by partitioning the objects into two groups and then by partitioning each group into subgroups, and so on. In either case, the results are best displayed in the form of a two-dimensional tree diagram called a **dendrogram** (Figure 18).

The simple class of agglomerative procedures known as *linkage methods* can be used to cluster either objects or variables. The basic algorithm is straightforward:

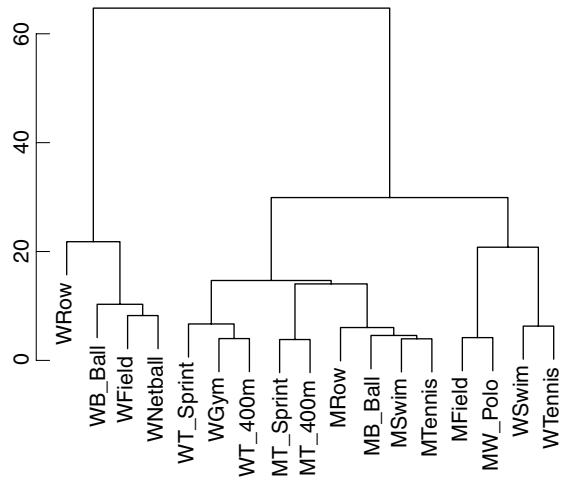

Figure 18 Dendrogram for the Australian Institute of Sport data. (Source: Cook and Weisberg [33, p. 98] on data supplied by Telford and Cunningham). The complete linkage method based on distances obtained from classic multidimensional scaling was used to construct the dendrogram. The clusters represented are not surprising given the clustering suggested by the ordination in Figure 17

1. Calculate an $n \times n$ symmetric matrix D of dissimilarities;
2. Regard each object or variable as being in a cluster so that there are n clusters;
3. Search D to find the most similar pair of clusters say U and V;
4. Merge U and V into a single cluster UV. Update D by deleting the rows and columns relating to U and V separately and introducing a new row and column for UV;
5. Repeat Steps 3 and 4 $n - 1$ times until there is a single cluster;
6. Construct a dendrogram from the record of mergers and the levels of dissimilarity at which they occur.

There are a large number of different ways of measuring the dissimilarity between two clusters in Step 2, and these essentially define different hierarchical clustering methods. For example, three commonly used approaches are:

(a) Single linkage or Nearest Neighbor [79]. The dissimilarity between clusters U and V is the smallest dissimilarity between a member of U and a member of V, namely,

$$d(U, V) = \min\{d(r, s) : r \in U, s \in V\}; \quad (6)$$

(b) Complete linkage or Farthest Neighbor [81]. The dissimilarity between clusters U and V is the greatest dissimilarity between a member of U and a member of V, namely,

$$d(U, V) = \max\{d(r, s) : r \in U, s \in V\}; \text{ and} \quad (7)$$

(c) Average linkage [80]. The dissimilarity between clusters U and V is the average of the $n_1 n_2$ dissimilarities between all members of U and V, namely,

$$d(U, V) = \frac{1}{n_1 n_2} \sum_{r \in U} \sum_{s \in V} d(r, s). \quad (8)$$

The results need not be the same in each case. For single and complete linkage, dissimilarities with the same initial ordering lead to the same configuration. However, for average linkage, dissimilarities with the same initial ordering need not result in the same configuration. All the methods are sensitive to outliers.

The sensitivity of the results to so many factors makes interpretation of the final configuration difficult. Also, a hierarchical method always yields clusters whether they are real or not. It is sensible to try different distances and different methods to check the stability of the configuration but even so, it is advisable to be cautious in interpreting the results.

Finally, the similarities between ordination and clustering mean that as an alternative to a dendrogram, we may be able to combine indicators of cluster membership within an ordination.

Analysis of Survival Data

The analysis of **survival** times for patients in medical studies is an important and common biostatistical application. The broad topic of survival analysis is covered in considerable detail elsewhere in these volumes; so, here we will focus primarily on graphical displays that are useful in modeling data of this type. Although the analysis of patient survival times is a typical setting for survival analysis, the term applies more generally to include the analysis of many other types of waiting times such as times to failure in an industrial process or, more positively, the time it takes for a patient to recover from an injury. A particular feature of survival data that makes its analysis distinctive is the problem of **censoring**, for example, due to individuals leaving the study before it is completed or to individuals surviving the entire period of study (so that their survival time cannot be properly observed). There are many excellent, comprehensive accounts of survival analysis available, including [38, 69] and [63].

Several functions are useful in the study of survival times (*see* **Survival Distributions and Their Characteristics**). In particular, the survival function $S(t) = P(T > t)$, the probability that an individual will survive beyond some particular time t is a central concept that is closely related to the distribution function of the survival times ($F(t) = 1 - S(t)$). Related functions include the hazard function (also called the force of mortality, or the **hazard rate**), which represents the instantaneous death (or failure) rate at time t, and the cumulative hazard $H(t) = -\log\{S(t)\} = \int_0^t h(z)\mathrm{d}z$. A plot of the survival function $S(t)$ versus time is referred to as a survival curve and represents how the probability of survival changes through time.

Of course, in almost all cases the survival function is unknown and must be estimated. The traditional **actuarial** approach, the **life table** method, groups survival times into a fixed set of intervals, and tabulates for each interval the number of subjects alive at the start of the interval, the proportion of subjects who die during the interval, and the estimated survival function (proportion surviving) at the start of the interval. While this approach is reasonable for large sets of survival times, for small data sets the potential for lengthy periods during which no deaths occur makes **grouping survival times** into intervals less attractive. The most commonly used estimator of the survival function is the **Kaplan–Meier** Product Limit estimator [64], which models the survival probability at time t as the product of the proportions surviving among those alive and present for the study at the beginning of each preceding time period. The Kaplan–Meier estimator takes account of censoring by calculating the proportion surviving during an interval by reducing both the numerator and the denominator of the proportion by the number of censored observations in that period. The term survival curve is usually used in the context of a plot of the Kaplan–Meier estimator $\hat{S}(t)$ of the survival function versus time. The Kaplan–Meier estimator is a step-function, constant over periods where there are no deaths, and falling at the time of each death. Times at which censoring events occur are represented on the curve by a distinguishing mark such as a tick mark or a +. A particularly effective graphical tool in understanding survival data where patients are divided into different groups (for example, treatment/placebo, male/female) is a plot of several survival curves on the same axes. This plot allows an effective comparison between the survival probabilities of several groups over time, for example, allowing the comparison of the effectiveness of different treatment regimes. Figure 19 shows estimated survival curves for male and female lung cancer patients in a Mayo Clinic study [70]. The graph shows that the survival experience of females is a little better than for males in this study. Note, however, that other factors that may influence survival (such as the age of patients) were not taken into account in this analysis.

A more sophisticated approach to modeling survival data allows survival time to be modeled not only through time but also in terms of other **covariates** that may influence survival (e.g. patient age or drug dosage). A commonly used model in this context is the **proportional hazards** model introduced by Cox [36] (*see* **Cox Regression Model**). This model

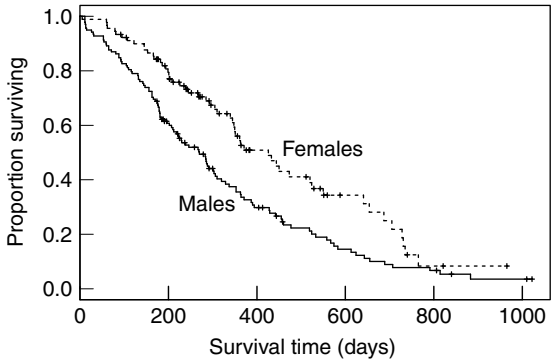

Figure 19 Survival curves comparing male and female groups of lung cancer patients. The data is described by Loprinzi et al. [70] from work carried out at the Mayo clinic. The data also contained information on other covariates such as patient age, weight loss, and some diagnostic measurements, although this information was not used in estimating survival probabilities. The estimated survival curves suggest a significant difference between male and female survival experience up to about 2 years survival, though the curves are similar beyond that time. Censored observations are marked on the plot by a plus (+) symbol

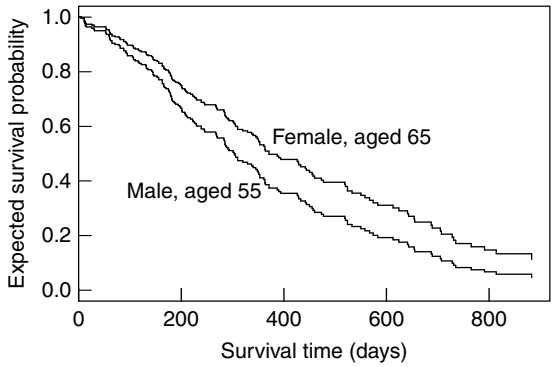

Figure 20 Expected survival probabilities arising from a proportional hazards model for the Mayo clinic lung cancer data (Loprinzi et al. [70]). The model was fit to incorporate both gender and age effects. Estimated expected survival curves are depicted for a male patient aged 55 and a female patient aged 65. Note that despite the 10-year age difference the female's expected survival experience remains better than the male's. In the model fit, the age covariate was only marginally significant

assumes that the hazard rate for a patient having covariate values z is proportional to some baseline hazard rate (usually corresponding to a patient having "average" covariates), with the constant of proportionality depending only on the values of the covariates z and not on time. A plot of a patient's expected survival curve which takes into account their particular covariate values can then be presented; see Figure 20 for an example.

Another graphic commonly used for analyzing survival data is a plot of the estimated survival function $\hat{S}(t)$ (or the estimated cumulative hazard $\hat{H}(t) = -\log\{\hat{S}(t)\}$) against a theoretical survival function $S(t)$ (or cumulative hazard $H(t)$, respectively). This graphic allows the comparison of an observed survival experience with a theoretical model for survival, with deviations from a straight-line pattern reflecting departures from the model. The visual basis for comparison in this case is similar to that used in Q–Q or P–P plots.

References

[1] Allison, T. & Ciccheti, D.V. (1976). Sleep in mammals: ecological and constitutional correlates, *Science* **194**, 732–734.

[2] Anderson, E. (1957). A semigraphical method for the analysis of complex problems, *Proceedings of the National Academy of Sciences* **13**, 923–927. (Reprinted in *Technometrics* (1960) **2**, 387–391).

[3] Andrews, D.F. (1972). Plots of high-dimensional data, *Biometrics* **28**, 125–136.

[4] Anscombe, F.J. (1973). Graphs in statistical analysis, *The American Statistician* **27**, 17–21.

[5] Anscombe, F.J. & Tukey, J.W. (1963). The examination and analysis of residuals, *Technometrics* **5**, 141–160.

[6] Asimov, D. (1985). The grand tour: a tool for viewing multivariate data, *SIAM Journal of Scientific and Statistical Computing* **6**, 128–143.

[7] Atkinson, A.C. (1982). Regression diagnostics, transformations and constructed variables (with discussion), *Journal of the Royal Statistical Society Series B* **44**, 1–35.

[8] Bachi, R. (1968). *Graphical Rational Patterns: A New Approach to Graphical Presentation of Statistics*. Israel University Press, Jerusalem.

[9] Barnett, V. & Lewis, T. (1978). *Outliers in Statistical Data*. John Wiley & Sons, New York.

[10] Becker, R.A. & Cleveland, W.S. (1987). Brushing scatterplots, *Technometrics* **29**, 127–142.

[11] Becker, R.A., Cleveland, W.S. & Shyu, M.-J. (1996). The visual design and control of trellis displays, *Journal of Computational and Graphical Statistics* **5**, 123–155.

[12] Beckman, R. & Cook, R.D. (1983). Outlier. . .s, *Technometrics* **25**, 119–149.

[13] Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics*. John Wiley & Sons, New York.

[14] Beniger, J.R. & Robyn, D.L. (1978). Quantitative graphics in statistics: a brief history, *The American Statistician* **32**, 1–11.

[15] Benzecri, P.J. (1973). L'analyse des correspondences, *L'analyse des Données*, Vol. 2. Dunod, Paris.

[16] Bertin, J. (1973). *Semiologie Graphique*, 2nd Ed. Mouton, Paris and The Hague (in French).

[17] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society Series B* **26**, 211–252.

[18] Box, G.E.P. & Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

[19] Bruntz, S.M., Cleveland, W.S., Kleiner, B. & Warner, J.L. (1974). The dependence of ambient ozone on solar radiation, wind, temperature, and mixing height, in *Symposium on Atmospheric Diffusion and Air Pollution*. American Meteorological Society, Boston, pp. 125–128.

[20] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Duxbury Press, Boston.

[21] Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association* **68**, 361–368.

[22] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

[23] Cleveland, W.S. (1981). LOWESS: a program for smoothing scatterplots by robust locally weighted regression, *The American Statistician* **35**, 54.

[24] Cleveland, W.S. (1985). *The Elements of Graphing Data*. Wadsworth, Monterey.

[25] Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, Summit.

[26] Cleveland, W.S. & Devlin, S.J. (1982). The SABL seasonal and calendar adjustment procedures, in *Time Series Analysis: Theory and Practice 1*, O.D. Anderson, ed. North Holland, Amsterdam, pp. 539–564.

[27] Cleveland, W.S. & McGill, M.E. (1984). The many faces of a scatterplot, *Journal of the American Statistical Association* **79**, 807–822.

[28] Cleveland, W.S. & McGill, M.E., eds. (1988). *Dynamic Graphics for Statistics*. Wadsworth and Brooks-Cole, Belmont.

[29] Cook, R.D. (1979). Influential observations in linear regression, *Journal of the American Statistical Association* **74**, 169–174.

[30] Cook, D., Buja, A. & Cabrera, J. (1993). Projection pursuit indices based on orthonormal function expansions, *Journal of Computational and Graphical Statistics* **2**, 225–250.

[31] Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995). Grand tour and projection pursuit, *Journal of Computational and Graphical Statistics* **4**, 155–172.

[32] Cook, R.D. & Weisberg, S. (1980). *Residuals and Influence in Regression*. Chapman & Hall, London.

[33] Cook, R.D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley & Sons, New York.

[34] Corsten, L.C.A. & Gabriel, K.R. (1976). Graphical exploration in comparing variance matrices, *Biometrics* **32**, 851–863.

[35] Costigan-Eaves, P. & Macdonald-Ross, M. (1990). William playfair (1759–1823), *Statistical Science* **5**, 318–326.

[36] Cox, D.R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society Series B* **34**, 187–220.

[37] Cox, D.R. (1978). Some remarks on the role in statistics of graphical methods, *Applied Statistics* **27**, 4–9.

[38] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[39] Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Oxford University Press, Oxford.

[40] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK.

[41] Doksum, K. & Sievers, G. (1976). Plotting with confidence. graphical comparisons of two populations, *Biometrika* **63**, 421–434.

[42] Everitt, B. (1978). *Graphical Techniques for Multivariate Data*. North Holland, New York.

[43] Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables, *Journal of the American Statistical Association* **19**, 431–453.

[44] Fienberg, S.E. (1979). Graphical methods in statistics, *The American Statistician* **33**, 165–178.

[45] Friedman, H.P., Farrell, E.S., Goldwyn, R.M., Miller, M. & Sigel, J.H. (1972). A graphic way of describing changing multivariate patterns, in *Proceedings of the Sixth Interface Symposium on Computer Science and Statistics*, University of California, Berkeley, pp. 56–59.

[46] Friedman, J.H. & Stuetzle, W. (1982). Projection pursuit methods for data analysis, in *Modern Data Analysis*, R.L. Launer & A.F. Siegel, eds. Academic Press, New York.

[47] Friedman, J.H. & Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers* **C-23**, 881–890.

[48] Gabriel, K.R. (1971). The biplot graphical display of covariance matrices with application to principal component analysis, *Biometrika* **58**, 453–467.

[49] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons, New York.

[50] Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**, 325–338.

[51] Gower, J.C. & Digby, P.G.N. (1981). Expressing complex relationships in two dimensions, in *Interpreting Multivariate Data*, V. Barnett, ed. John Wiley & Sons, New York, pp. 83–118.

[52] Gower, J.C. & Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis, *Applied Statistics* **18**, 54–64.

[53] Hartigan, J.A. (1975). Printer graphics for clustering, *Journal of Statistical Computation and Simulation* **4**, 187–213.

[54] Hastie, T.J. (1986). Generalized additive models: a GAIM analyst's toolbox, *American Statistical Association Proceedings in Statistical Computation*, 41–47.

[55] Hastie, T.J. & Tibshirani, R. (1986). Generalized additive models (with discussion), *Statistical Science* **1**, 297–310.

[56] Hastie, T.J. & Tibshirani, R. (1987). Generalized additive models: some applications, *Journal of the American Statistical Association* **82**, 371–386.

[57] Hastie, T.J. & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[58] Hawkins, D.M. (1980). *Identification of Outliers*. Chapman & Hall, London.

[59] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **26**, 139–142.

[60] Johnson, R.A. & Wichern, D.W. (1982). *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs.

[61] Jones, M.C. & Rice, J.A. (1992). Displaying the important features of large collections of similar curves, *The American Statistician* **46**, 140–145.

[62] Jowett, G.H. (1952). The accuracy of systematic sampling from conveyor belts, *Applied Statistics* **1**, 50–59.

[63] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

[64] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[65] Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**, 115–129.

[66] Kruskal, J.B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation', in *Statistical Computation*, R.C. Milton & J.A. Nelder, eds. Academic Press, New York, pp. 83–118.

[67] Larsen, W.A. & McCleary, S.J. (1969). The use of partial residuals in regression analysis, *Bell Laboratories Memorandum*. Murray Hill, NJ. Unedited version of Larsen and McCleary (1972).

[68] Larsen, W.A. & McCleary, S.J. (1972). The use of partial residual plots in regression analysis, *Technometrics* **14**, 781–790.

[69] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York.

[70] Loprinzi, C.L., Laurie, J.A., Wieand, H.S., Krook, J.E., Novotny, P.L., Kugler, J.W., Bartel, J., Law, M., Bateman, M. & Klatt, N.E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North central cancer treatment group, *Journal of Clinical Oncology* **12**, 601–607.

[71] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading.

[72] Neter, J., Wasserman, W. & Kutner, M.H. (1990). *Applied Linear Statistical Models*. Irwin, Boston.

[73] Playfair, W. (1786). *The Commercial and Political Atlas*. Corry, London.

[74] Potthoff, R.F. & Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika* **51**, 313–326.

[75] Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*, 2nd Ed. Duxbury Press, Belmont.

[76] Richardson, M.W. (1938). Multidimensional psychophysics, *Psychological Bulletin* **35**, 659–660.

[77] Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function, *Psychometrika* **27**, 125–140. 219–246.

[78] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[79] Sneath, P.H.A. (1957). The application of computers to taxonomy, *Journal Genetic Microbiology* **17**, 201–226.

[80] Sokal, R.R. & Michener, C.D. (1958). A statistical method for evaluating systematic relationships, *University of Kansas Scientific Bulletin* **38**, 1409–1438.

[81] Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. Freeman, San Francisco.

[82] Stone, C.J. (1977). Consistent nonparametric regression, *The Annals of Statistics* **5**, 595–620.

[83] Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models, *The Annals of Statistics* **14**, 590–606.

[84] Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization, *Journal of the American Statistical Association* **83**, 394–405.

[85] Torgerson, W.S. (1952). Multidimensional scaling: I – Theory and method, *Psychometrika* **17**, 401–419.

[86] Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire.

[87] Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.

[88] Tukey, P.A. & Tukey, J.W. (1981). Graphical display of data sets in 3 or more dimensions, in *Interpreting Multivariate Data*, V. Barnett, ed. John Wiley & Sons, New York, 189–257.

[89] Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*. Duxbury Press, Belmont.

[90] Wainer, H. (1974). The suspended rootogram and other visual displays: an empirical validation, *The American Statistician* **28**, 143–145.

[91] Wainer, H. (1990). Graphical visions from William Playfair to John Tukey, *Statistical Science* **5**, 340–346.

[92] Weisberg, S. (1980). *Applied Linear Regression*. John Wiley & Sons, New York.

[93] Wilk, M.B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data, *Biometrika* **55**, 1–17.

[94] Wood, F.S. (1973). The use of individual effects and residuals in fitting equations to data, *Technometrics* **15**, 677–695.

Further Reading

Tukey, J.W. (1990). Data-based graphics: visual display in the decades to come, *Statistical Science* **5**, 327–339.

(*See also* **Computer-intensive Methods**; **Software, Biostatistical**)

MICHAEL A. MARTIN & A.H. WELSH

Graphical Presentation of Longitudinal Data

Longitudinal data comprise repeated measurements over time on many individuals [1]. Typically, a longitudinal data analysis has the regression objective of describing the dependence of a **response variable** Y on time t and other **explanatory variables** $\mathbf{x}$. With a single response on each subject, the standard scatter-plot, perhaps enhanced with a **nonparametric regression** estimate of $E(Y|\mathbf{x})$, is an effective **graphical display** to explore this dependence. With repeated measurements, there is the opportunity to display other features of the data such as the relative variation in Y across time within a person and differences among persons in the dependence of Y on t or $\mathbf{x}$.

While there are not commonly accepted rules for the effective display of longitudinal data, the following guidelines [1] may be helpful:

1. Display original data relevant to the question at hand, not just data summaries. This communicates the degree of variation against which summaries should be judged.
2. Make apparent unusual observations and unusual persons (*see* **Outliers**).
3. Contrast variation within and among persons in the dependence of Y on t or $\mathbf{x}$.

A model for longitudinal data has two components: a regression model for the dependence of Y on $(\mathbf{x},t)$; and a model for the **autocorrelation** among the repeated observations for a person. Graphical displays are necessary for each component.

To illustrate two simple, but effective, displays, longitudinal data on the growth of Nepali preschool children, collected in the Nepal Nutrition Intervention Project [3], are used (*see* **Growth and Development**). This was a prospective investigation of the effects of vitamin A supplementation on children's morbidity and mortality. The data set here comprises 11 290 observations on 2237 children receiving placebos. Each child was measured in up to five visits at 4-month intervals.

Regression Displays

Figure 1 displays a standard scatter-plot of the 11 290 weights against age. Note that weight is recorded to the nearest tenth of a kilogram and age to the nearest month, which would create an unacceptable degree of granularity in the display. To overcome this problem, the x and y values have been jittered by adding a uniform error. The scale of the error was chosen to make nearly all points visible but also to communicate the limited precision of the raw data.

A smoothing **spline** with 22 equivalent degrees of freedom [2] was added to highlight the typical growth pattern. The number of degrees of freedom was chosen by eye so that the resulting smooth curve was sufficiently flexible to capture the strong nonlinearity in growth in the first year. Automatic selection criteria for longitudinal data have been developed (see [4] and references therein). But trial and error can be equally effective at choosing a degree of smoothness to enhance the data display.

The repeated measurements for a subset of 100 children have been connected to communicate the degree of consistency across time in a child's weight as well as the variation in weight among children. In some data sets, the repeated observations can usefully be connected for all subjects. In this case, doing so produces a useless ink blot. One simple, alternative strategy is to connect the repeated values for a judiciously selected subset of, say, 100 children. This selection can be done at random. However, this strategy will fail to identify unusual weights or children. As an alternative, we have chosen 50 children at random and a second 50 that were at the extremes of the data in the following sense. For each child, the mean time and weight was calculated. Mean weight was regressed on mean time and a residual obtained. The second 50 children comprise those with the largest residuals in absolute value. Hence, these children have average weights which are extreme for their age. Other criteria for identifying a subset of subjects are discussed by Diggle et al. [1].

In Figure 1, several important characteristics of the data set are now apparent. First, the average weight of Nepali children increases by about 1 kg per month for the first 6 months and then the growth rate dramatically slows to less than a quarter of the original rate. Secondly, we can see the degree to which a description of Nepali growth using these data will depend upon cross-sectional, in addition to longitudinal, information. Note that no child was observed for more than 18 months, as seen in the Figure by the lengths of the children's lines. Hence comparing average size at times separated by a

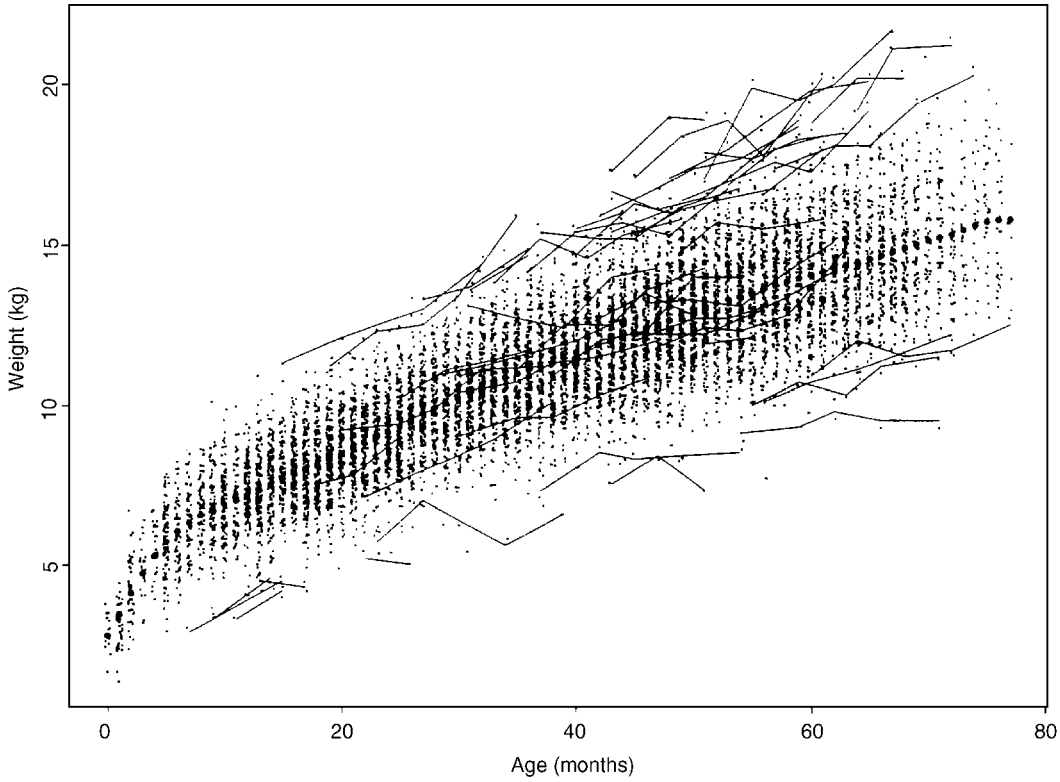

Figure 1 Scatter-plot of 11 290 weights on 2337 Nepali children. Bold dots indicate a smoothing spline with 22 equivalent degrees of freedom as an estimate of the mean weight at each age. The repeated observations for 100 children are connected – 50 chosen at random and 50 with extreme mean weights for their age

greater time interval uses entirely cross-sectional information. Thirdly, there is much greater variability in weight across children than across time for a given child. That is, there is strong *tracking* of children's weight so that repeated observations on the same child will be highly correlated. Finally, there is some indication that the rate of growth is greater for larger children, as evidenced by more positive slopes above the average curve than below. If substantiated by a more careful analysis, then this fact would be important to public health workers who could target interventions more appropriately.

Association Displays

A scatter-plot matrix is one effective approach to more directly view the nature and degree of association among repeated observations. First, the mean structure is removed from the data by regressing the response Y on predictors $\mathbf{x}$ yielding **residuals** r. If the data are observed at equally spaced times, the scatter-plot matrix simply displays r_{ij} vs. r_{ik} for all times $j \neq k$, where i denotes the subject. Unequally spaced times can be rounded to produce an equally spaced observation grid.

For example, in the Nepali data set, observation times were categorized into 6-month intervals. This often produced more than one observation per bin for a subject. All pairs of observations on each child – one from bin j and the other from bin k – could then be plotted in the (j, k) entry in the scatter-plot matrix. In fact, because there is such a large number of data pairs ($>30\,000$), only a single pair per child is used in each scatter-plot in the matrix.

Several characteristics of the data are apparent in Figure 2. First, we can again see that there is limited follow-up on each child by the absence of information away from the diagonal of the matrix.

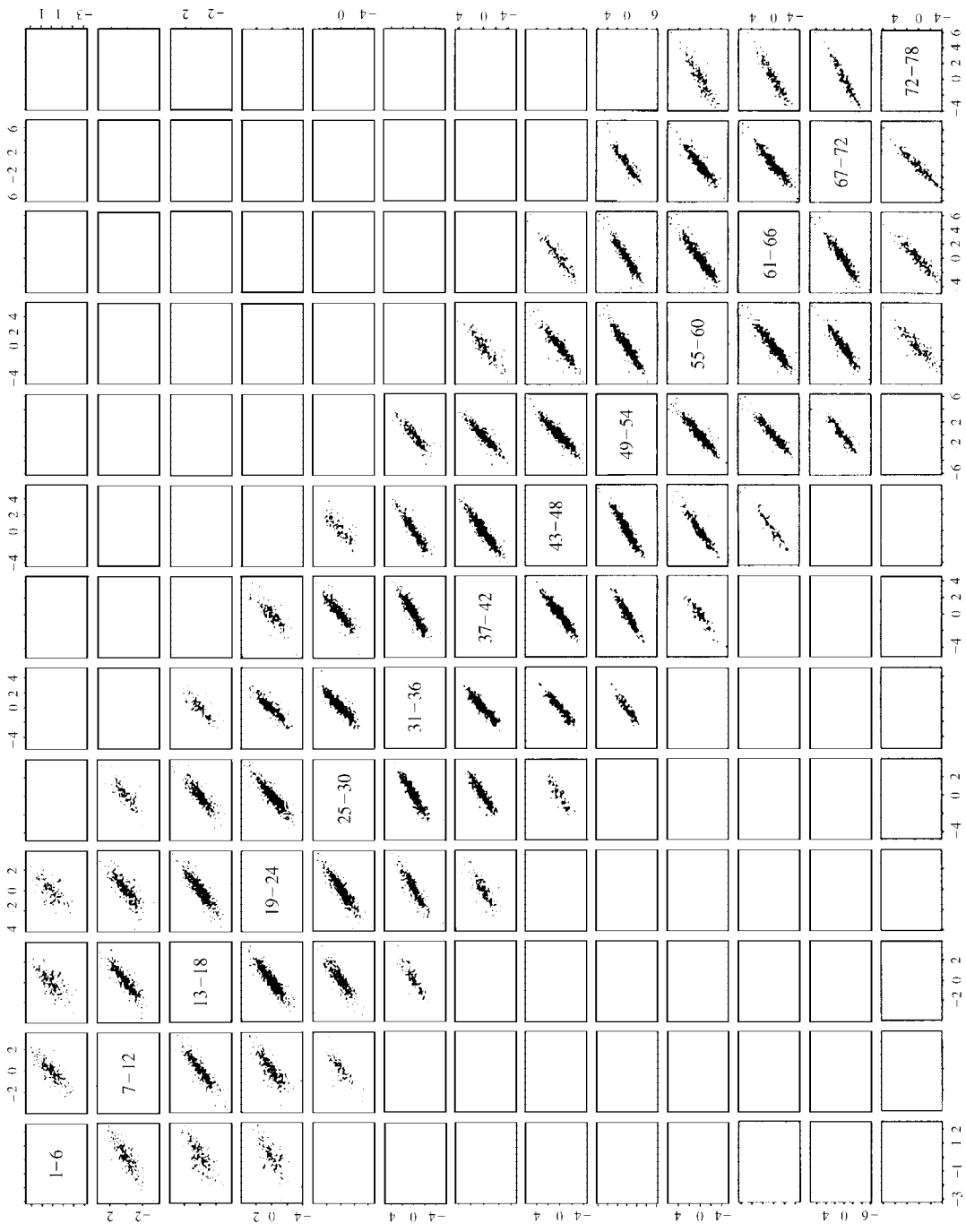

Figure 2 Scatter-plot matrix display of the residuals r_{ij} vs. r_{ik} for all $j > k$. The unequally spaced times are rounded to the nearest 6-month interval

Secondly, the extremely strong autocorrelation at all lags and ages is evident. Thirdly, the display suggests that the degree of association for two observations a fixed time apart is stronger at older ages than at the beginning of life. This is an example of a non-**stationary** autocorrelation pattern [1]. It might result if there is less stability across time in weight when the growth rate is faster.

If the autocorrelations appear to depend only on the lag and not on the absolute times of observation, the correlation pattern is said to be *stationary* and the scatter-plot matrix can be usefully summarized by a **variogram** [1]. Here, the normalized squared difference $0.5\ (r_{ij} - r_{ik})^2$ is plotted against the absolute time between the repeated observations $|t_{ij} - t_{ik}|$. A smooth regression curve fit to these data estimates the variogram, $0.5\text{E}\ [Y_i(t) - Y_i(t-u)]^2 = \sigma^2[1 - \rho(u)]$, $u > 0$, where t is the first observation time, u the lag, and $\rho(u)$ the autocorrelation function, which, because of stationarity, depends only upon u and not t. Diggle et al. [1] describe the use of the variogram in choosing models for longitudinal data.

References

[1] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, pp. 267–298.

[2] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, New York.

[3] West, K.P., Pokhrel, R.P., Katz, J., LeClerq, S.C., Khatry, S.B., Shrestha, S.R., Pradhan, E.K., Tielsch, J.M., Pandey, M.R. & Sommer, A. (1991). Efficacy of vitamin A in reducing preschool child mortality: a randomized double-masked community trial in Nepal, *Lancet* **338**, 67–71.

[4] Zeger, S.L. & Diggle, P.J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters, *Biometrics* **50**, 689–699.

(*See also* **Semiparametric Regression**)

SCOTT L. ZEGER & JOANNE KATZ

Graunt, John

Born: April 24, 1620, in London, UK.
Died: April 18, 1674, in London, UK.

John Graunt was a London draper who, in February 1662, published a small book *Natural and Political Observations Mentioned in a following Index and Made Upon the Bills of Mortality*. For this pioneer study of medical statistics and **demography** Graunt is rightly recognized as the founder of statistics as a scientific discipline.

The book attracted immediate attention. Within a month Graunt was elected to the Royal Society; a second edition appeared later in the year, and a third and fourth in the early weeks of the plague in 1665. A fifth edition appeared two years after his death, and there are modern reprints [1, 3, 7].

Graunt was a respected citizen, a Freeman, and eventually Renter Warden of the Drapers' Company. He held various civil and military offices, and his influence was sufficient, before he was 30, to procure the professorship of music at Gresham College for his friend **William Petty**. Graunt's house was destroyed in the fire of 1666, but despite assistance from Petty his business never recovered. A few years later he became a Catholic, resigned his offices, and died in poverty. No portrait is known.

The *London Bills of Mortality* were weekly accounts of the numbers of burials, distinguishing deaths from plague, and christenings, compiled from parish registers from the mid-sixteenth century. **Causes of death** were included from the early sixteen hundreds. Annual summaries were published, but initially only during plague years.

Graunt's study was based mainly on the annual *Bills* from 1604 to 1660. He had no information on population sizes. With this limited material his approach was thoroughly logical and scientific. He described in detail how the data were collected, and their nature; he was critical of their accuracy and completeness; he tabulated the material extensively and informatively, checked his first impressions against more extensive facts, and drew a wide variety of sensible and valid conclusions. Among much else, Graunt directed attention to the very high rates of mortality in infancy (*see* **Infant and Perinatal Mortality**), and showed that mortality was higher in London than in the country. He made the first realistic estimates of the numbers of men and women in London and the population of the whole country and showed that both were increasing, with a steady migration into London. He demonstrated that plague was under-recorded by about a quarter, examined the relative mortality in different plague years, discovered the extent to which London depopulated itself in plague years, and showed that it repopulated itself within a year. He distinguished between epidemic and endemic diseases, and noted the stability of accident and suicide rates from year to year, the under-recording of syphilis, and the increase of rickets. Graunt's methods and findings are reviewed in [2, 4], and [5].

Graunt had no information on the ages of the dead or the living. This led him to conceive the first **life table**, describing the dying-out of a population cohort in an attempt to estimate the number of men of military age (16–56) in London (see Table 1). However, not appreciating the need to use age-specific mortality rates, he mistakenly estimated the proportion of *deaths* between these ages from his survivors' column, and not the proportion living [2] (*see* **Standardization Methods**). His pioneer effort was nevertheless highly influential in stimulating the later **actuarial** development of the life table.

Suggestions that Graunt was not the author of the "Observations" and that these were the work of Petty only arose after Graunt's death, and were revived

Table 1 Graunt's life table

Exact age	Deaths	Survivors
0		100
	36	
6		64
	24	
16		40
	15	
26		25
	9	
36		16
	6	
46		10
	4	
56		6
	3	
66		3
	2	
76		1
	1	
80		0

during the twentieth century (see [3, 6], and [7]). A comprehensive re-examination by Glass [2] concluded that "there seems little reason to doubt that the volume published under Graunt's name was in all essential respects Graunt's work".

References

[1] Benjamin, B., ed. (1964). John Graunt's "Observations" (reprint of the first edition), *Journal of the Institute of Actuaries* **90**, 1–61.

[2] Glass, D.V. (1963). John Graunt and his natural and political observations, *Proceedings of the Royal Society, Series B* **159**, 1–37.

[3] Hull, C.H., ed. (1899, reprinted 1963). *The Economic Writings of Sir William Petty Together With the Observations on the Bills of Mortality More Probably by Captain John Graunt* (includes a reprint of the 5th Ed.). Cambridge University Press, Cambridge.

[4] Sutherland, I. (1963). John Graunt: a tercentenary tribute, *Journal of the Royal Statistical Society, Series A* **126**, 537–556.

[5] Sutherland, I. (1972). When was the Great Plague? Mortality in London, 1563–1665, in *Population and Social Change*, D.V. Glass & R. Revelle, eds. Edward Arnold, London, pp. 287–320.

[6] Willcox, W.F. (1937). The founder of statistics, *Review of the International Statistical Institute* **5**, 321–328. (Also reprinted in ref. [7].)

[7] Willcox, W.F., ed. (1939). *Natural and Political Observations Made Upon the Bills of Mortality by John Graunt* (reprint of the 1st Ed.). Johns Hopkins Press, Baltimore.

I. SUTHERLAND

Greenberg, Bernard George

Born: October 4, 1919, in New York City.
Died: November 24, 1985, in Chapel Hill.

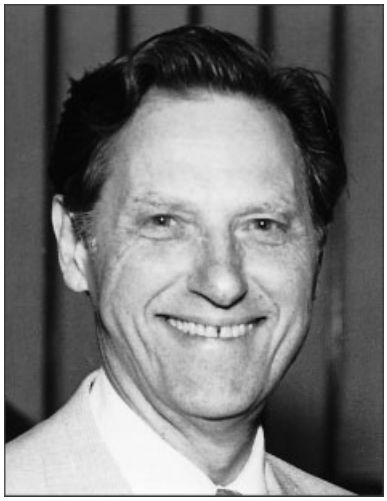

Bernard G. Greenberg, the founder of the Department of Biostatistics at the University of North Carolina and a pioneer in the field of public health, played a key role in the evolution of the discipline of biostatistics encompassing a large domain of public health, **demography** (population studies), and medical/clinical research.

Greenberg earned a B.S. degree in mathematics in 1939 from the City College of New York, and before being inducted into the US Army in 1941, served briefly as an assistant statistician in the US Bureau of Census, as well as in the New York State Department of Health. Following his discharge from the active military service (where he was raised to the rank of Captain), in 1946 he attended a special summer session of the Institute of Statistics at the North Carolina State College in Raleigh. His mentor, **Gertrude Cox**, organized instruction by eminent statisticians such as **R.A. Fisher**; **Harold Hotelling**; **George Snedecor**; **William Cochran**; **Chester Bliss**; Jacob Wolfowitz, and others. Greenberg was strongly influenced by this program and entered the North Carolina State College as a graduate student in experimental statistics, earning his Ph.D. degree in 1949. During the tenure of his graduate studies, Greenberg studied under Harold Hotelling, who headed the theoretical division of the Institute of Statistics of the University of North Carolina at Chapel Hill, though he was also influenced by William Cochran.

In July 1949, Edward G. McGavran, the Dean of the School of Public Health, University of North Carolina at Chapel Hill, invited Greenberg to organize a new, single faculty, Department of Biostatistics. Greenberg nurtured the growth of this department for more than 23 years as the Chairman (1949–1972), and also later (1972–1982), as the Dean of the School of Public Health. In the fall of 1982 he returned to the biostatistics faculty and began development of a curriculum for statistics of **communicable diseases**. Soon after, however, he was afflicted with cancer and, following a long period of illness, died in the fall of 1985, only a few months after his retirement. Thus the Greenberg era of outstanding leadership and phenomenal developments in biostatistics and public health started in 1949 and came to an unexpected halt after nearly 36 years. To outline Greenberg's major contributions, we need to focus on his organizational leadership and developmental vision before outlining his academic contributions.

In the early 1950s the primary mission of the new Department of Biostatistics at Chapel Hill was to offer a few service courses to meet the growing statistical needs of faculty and students in public health. This objective was quickly expanded to teaching courses in the entire field of Health Affairs (including the Medical, Nursing, Dental, Public Health, and Pharmacology Schools), and also providing biostatistical consultation support to faculty members. With generous support from the **National Institutes of Health (NIH)** and other Agencies, he introduced masters and doctorate level programs in biostatistics at Chapel Hill that ultimately have been recognized as among the finest programs in the nation.

Two of Greenberg's major contributions to the field of biostatistics prior to the 1960s deserve special mention. First, he introduced a field training program in which a master's degree student was assigned for several weeks to a counselor in a state health agency and this was designed to familiarize students with real problems in public health and to help them in choosing an appropriate professional career in the health sciences. Secondly, faculty members became engaged in cooperative, **multicenter clinical trials**, first in cancer clinical trials and later in the 1960s and

1970s on several large projects including the LIPIDS (1972–1984) trials sponsored by the National Heart, Lung and Blood Institute. Greenberg had a special skill in attracting distinguished statisticians, including D.R. Cox, David Duncan, and Herbert A. David, as visiting professors in the Department, with support from the training grants. In the 1950s and 1960s this aided the development of a biostatistics curriculum, and enabled the implementation of interactive research programs and research incentives for regular faculty and students in and around North Carolina. The creation of the population laboratory (PopLab) in the School of Public Health and with the association of the Carolina Population Center in the mid 1960s is an example of Greenberg's bringing together the key government and academic people with the broad objective of developing a research and training program.

Greenberg not only provided leadership to the Department of Biostatistics, but also contributed to research, often combining methodological research with applications. His interest in **order statistics** in the 1950s enabled him, in collaboration with A.E. Sarhan, to edit a fine volume entitled *Contributions to Order Statistics* [1] which gave an up-to-date account of developments in that field, including many of their joint works. In the field of **randomized response** models, Greenberg was instrumental in incorporating this basic statistical concept to reduce the **bias** of responses to sensitive questions in public health investigations. The methodological work began in the early 1960s and continued to have an important influence in the late 1970s. Perhaps Greenberg's major statistical contributions related to applications in the field of public health, with special emphasis on epidemiology, maternal and child health, nutrition and health administration. Over a long period he was associated with the North Carolina State Board of Health and introduced statistical tools and concepts in their areas of application. Though primarily an administrator and a leader in public health, he was a genuine promoter of the utilization of statistical tools and concepts in scientific inquiries.

His efforts and contributions were recognized with numerous honors and awards: the Bronfman Award (1966) from the **American Public Health Association**; a Kenan Professorship (1969–1985) in biostatistics, University of North Carolina; the Watson S. Rankin Award (1980) from the North Carolina Public Health Association; and the Order of the Golden Fleece, as well as the O. Max Gardner Award, from the University of North Carolina (1983). He was elected a Fellow of the **American Statistical Association**, the Institute of Mathematical Statistics, the American Public Health Association, the American College of Epidemiology, and he was an elected member of the **International Statistical Institute**, the American Epidemiological Society, and the Institute of Medicine. He served on the editorial boards of the *Journal of Statistical Planning and Inference, International Statistical Reviews, Journal of Chronic Diseases*, and the *American Journal of Obstetrics and Gynecology*.

Bernard Greenberg was a humane, caring person who, in spite of heavy professional, scholarly, and administrative responsibilities, took a genuine, personal interest in all his colleagues, former students, friends and family members. A volume of collected papers prepared by his colleagues, friends, and former students was written in 1985 to honor Greenberg on his 65th birthday [2].

References

[1] Greenberg, B.G. & Sarhan, A.E. (1962). *Contributions to Order Statistics*. Wiley, New York.

[2] Sen, P.K., ed. (1985). *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences, the Bernard G. Greenberg Volume*. North-Holland, Amsterdam.

PRANAB K. SEN

Greenhouse, Samuel W.

Born: January 13, 1918, in Bronx, New York.
Died: September 29, 2000, in Rockville, Maryland.

Samuel W. Greenhouse was one of the founding statisticians at the **National Institutes of Health**, who helped pioneer the use of statistical methods in epidemiological research (*see* **Epidemiology, Overview**), and was influential in the early development of the theory and practice of clinical trials (*see* **Clinical Trials, Early Cancer and Heart Disease**). He was also a distinguished professor of statistics at the George Washington University.

Sam, as he was known to all, received his BS in mathematics from the City College of New York in 1938 and thereafter moved to Washington, DC to begin his career in the bureau of census with Edward Deming (1940–1942). He served in the army during World War II and afterwards worked with the United Nations Relief and Rehabilitation Agency (1945–1948). In 1948, he was recruited by Harold **Dorn**, along with Jerome **Cornfield**, Jacob Lieberman, Nathan **Mantel**, and Marvin **Schneiderman**, to create the first biometry group at the National Institutes of Health (NIH) in the National Cancer Institute (NCI). In 1954, Sam left the NCI to become head of the theoretical statistics and applied mathematics section in the National Institute of Mental Health (NIMH).

In 1966, he was appointed chief of the epidemiology and biometry branch of the National Institute of Child Health and Human Development (NICHD), where he rose to the position of associate director for epidemiology and biometry (1970–1974) and acting associate director of the office of program planning and evaluation (1969–1974). He was the first statistician to hold such a high administrative position at the NIH.

While working full time at the NIH, Sam taught part-time and pursued his own graduate degrees under the direction of Solomon Kullback in the department of statistics at George Washington University (GWU).

When Sam retired from government service in 1974, he began a full-time academic career at GWU, where he served as chair of the department of statistics from 1976–1979 and again in 1985–1986. In 1988, he retired from the university faculty and was named professor emeritus. From 1988 until his death, he served as the associate director for research development of the GWU biostatistics center.

Sam articulated many times that the primary mission of the statisticians at the NIH was to collaborate and provide statistical support for the NIH scientists. Yet, it was always understood that these collaborations would lead to opportunities for statistical research in methodology and theory. It was not unusual to find Sam and the other early NIH statisticians coauthoring papers in subject matter journals and publishing corresponding theory and methods papers in statistics journals. This pattern was evident in his early papers on the evaluation of diagnostic tests. Although this work with Mantel [11] and Dunn [3] was rooted in the need to implement noninvasive methods for cancer **screening**, it also addressed methodological issues, such as deriving the estimated variance of **sensitivity** and **specificity** for the case when, the diagnostic cut-point for a quantitative test was also estimated from the data. While at the NIMH, he helped design and analyze the first multidisciplinary study of normal aging, and coedited the resulting book [2]. Recognizing the need for methods to analyze highly correlated psychological data from studies such as this one, led directly to new methodological work with Geisser [8, 10].

They derived an estimate of the degree of departure from the assumption of compound symmetry in the test of within-subjects effects in analysis of variance ANOVA, and an adjustment to the degrees of freedom of the F-ratio when that assumption is violated. The Greenhouse–Geisser correction is now provided in virtually all computer packages for repeated measures analysis. Their work [10] was recognized in 1982 as a science and social science citation classic (*see* **Analysis of Variance for Longitudinal Data**).

Sam was also influential in the early development of the theory and practice of clinical trials and shared an interest with Cornfield in methods for the **sequential analysis** of emerging data in clinical trials. While at the NICHD, his collaborations focused more on **observational** data, for example, assessing the safety of oral contraceptive use, and his interests returned to the development of methods for epidemiologic studies. His papers with Seigel [13, 14], for example, showed that **logistic regression** could be applied to matched and unmatched **case–control studies** to obtain an adjusted estimate of the prospective **odds**

ratio associated with a **risk factor**. At GWU in the late 1980s, Sam and Joe Gastwirth recognized similarities between a class of problems arising in legal settings and in epidemiologic studies (*see* **Epidemiology as Legal Evidence**). A collaboration began that was deeply grounded in the practical experiences of their respective fields of application (see, e.g. [5, 6]).

It is a tribute to his energy and enthusiasm for statistics that Sam received many honors for his intellectual and professional contributions. The **American Statistical Association** (ASA) recognized him with their prestigious Founders Award in 1993, and in 1997, videotaped a discussion with Sam as part of the ASA series of conversations with distinguished statisticians [1]. In addition to being elected as a Fellow in the major statistical professional societies, Sam was also an elected Fellow of the American College of Epidemiology and of the council of epidemiology of the American Heart Association. In 1969, he received the superior service honor award from the NIH and was named a Johns Hopkins University centennial scholar in 1976. In 1999, Sam was recognized by the Harvard Institute of Psychiatric Epidemiology and Genetics for his lifetime contributions to psychiatric epidemiology and biostatistics.

Sam was a much-loved presence in the profession. He attended the annual meetings of the Eastern North American Region (ENAR) (*see* **International Biometric Society (IBS)**), the **Society of Clinical Trials**, the joint statistical meetings (JSM), and the American Association for the Advancement of Science (AAAS) without fail, and the ISI (*see* **International Statistical Institute (ISI)**) as often as he could. He was amazingly current and had strong opinions on all matters. He was not shy about asking questions of speakers, especially when he did not understand a point (or felt that they did not), and it would not be unusual for the discussion to continue in the hall or even later via e-mail until he felt the issues were resolved. This was true whether the topic was statistics, literature, music, politics, religion, or sports. In his 1997 *Statistical Science* article on his reminiscences of the NIH, Sam wrote about how the group (Cornfield, Halperin, Mantel, he, and others) would often argue quite publicly over lunch about matters statistical and otherwise [9]. Although one of Sam's most endearing features was his personal warmth and smile, he could also be quite the provocateur. We fondly remember times in the late 1970s and early 1980s, when Sam would visit Max **Halperin** or Nathan Mantel at the biostatistics center. Sam loved nothing more than a friendly spirited argument and Max and Nathan were always eager to comply. Sam was capable of arguing either side of an issue and often would, especially if it would get a rise out of Max or Nathan.

Sam was passionate about statistics. He relished the opportunity to teach and engage both colleagues and young statisticians in statistical discourse. Whether giving a seminar, making a site visit, or on sabbatical, Sam was always a popular and stimulating visitor and speaker. However, if one asked Sam about the truly important work he was doing, he would inevitably talk about his scientific collaborations. For it was through the practice of statistics, he believed, that statisticians made their biggest impact on science, and it was through scientific collaborations that the important statistical problems were identified.

Sam died of cancer at the age of 82. For additional biographical information see [1, 4, 7, 9, 12].

References

[1] ASA Distinguished Statistician Video Series (1997). "A Conversation with Sam Greenhouse". DS044.

[2] Birren, J.E., Butler, R.N., Greenhouse, S.W., Sokoloff, L. & Yarrow, M. (1963). *Human Aging*. U.S. Government Printing Office, Washington, DC.

[3] Dunn, J.E., Jr. & Greenhouse, S.W. (1950). *Cancer diagnostic tests: Principles and criteria for development and evaluation*. Federal Security Agency, Public Health Service, \#9, Government Printing Office, Washington, DC.

[4] Ellenberg, J.H. (1995). Some perspectives on the career of Samuel W. Greenhouse: the first 75 years, *Statistics in Medicine* **14**, 1615–1619.

[5] Gastwirth, J.L. & Greenhouse, S.W. (1987). Estimating a common relative risk: application in equal employment, *Journal of American Statistical Association* **82**, 38–45.

[6] Gastwirth, J.L. & Greenhouse, S.W. (1995). Biostatistical concepts and methods in the legal setting, *Statistics in Medicine* **14**, 1641–1653.

[7] Gehan, E.A. (ed.) (2003). Perspectives on the biostatistical sciences: a symposium in memory of Samuel W. Greenhouse, *Statistics in Medicine* **22**(21) 3263–3430.

[8] Geisser, S. & Greenhouse, S.W. (1958). An extension of Box's results on the F distribution in multivariate analysis, *Annals of Mathematical Statistics* **29**, 885–891.

[9] Greenhouse, S.W. (1997). Some reflections on the beginnings and development of statistics in "your father's" NIH, *Statistical Science* **12**, 82–87.

[10] Greenhouse, S.W. & Geisser, S. (1959). On methods in the analysis of profile data, *Psychometrika* **24**, 95–112.

[11] Greenhouse, S.W. & Mantel, N. (1950). The evaluation of diagnostic tests, *Biometrics* **6**, 399–412.

[12] Lachin, J.M. (2003). A tribute to Samuel W. Greenhouse, *Statistics in Medicine* **22**, 3267–3276.

[13] Seigel, D.G. & Greenhouse, S.W. (1973). Multiple relative risk functions in case-control studies, *American Journal of Epidemiology* **97**, 324–331.

[14] Seigel, D.C. & Greenhouse, S.W. (1973). Validity in estimating relative risk in case-control studies, *Journal of Chronic Disease* **26**, 219–225.

JOEL B. GREENHOUSE & JOHN M. LACHIN

Greenwood, Major

Born: August 9, 1880, in the parish of Shoreditch, East London, UK.
Died: October 5, 1949, London, UK.

Reproduced by permission of the Royal Statistical Society

Major Greenwood played an important role in the development of biostatistics during the first half of the twentieth century, both as an epidemiologist and as a pioneer medical statistician. He was the only surviving son in the third generation of a family in general practice in Hackney; a family practice run in his youth by his father and grandfather, both also called Major Greenwood. With a view to keeping the practice in the family, his father insisted that young Greenwood should have a medical education, in spite of a taste for history and Latin (so often linked to mathematical ability) which he had developed at Merchant Taylor's School. He was awarded an entrance scholarship to the London Hospital, where he qualified in 1904. Having so far bowed to parental pressure, Greenwood turned to his own interests in medical research rather than practice, helped and influenced by two outstanding, very different, scientists. He became a demonstrator in the physiology department of the London Hospital Medical College under (Sir) Leonard Hill; and at the same time he began following the courses in the rapidly developing science of biological statistics given by **Karl Pearson** at University College London.

In Hill's department the young Greenwood was introduced to scientific methodology as applied to medical research, while Pearson's course made him one of the earliest medical converts to the use of biometric measurements in evaluating approaches to the treatment and prevention of disease. Throughout his subsequent career, Greenwood was to combine and develop what he regarded as his double intellectual legacy from his two mentors, to both of whom he remained a loyal pupil and friend. In the early decades of the century he succeeded in developing medical statistics in a way to make its methodology acceptable to an initially reluctant medical profession, primarily by adding good clinical judgment to the rigorous mathematics that characterized Pearson's work.

In 1909–10 Greenwood, by now a true disciple of Pearson, became involved in a debate between bacteriologists and clinicians. It concerned Almroth Wright's research on the so-called "opsonic index" as a technique for measuring a patient's capacity to deal with infection by phagocytosis. Greenwood based his criticism of Wright's statistical data on a technical distinction between "functional" errors of technique and "mathematical" inferential errors [2]. His cogent arguments came to the attention of C.J. Martin, Director of the Lister Institute, who was impressed by Pearson's promising disciple. As a result, Martin created a new post for Greenwood, the first of its kind in Britain, as resident statistician at the Lister. Here Greenwood's statistical investigations included studies of tuberculosis (which had killed his two younger siblings in infancy, and his mother in the year he himself qualified in medicine), **infant mortality**, and hospital fatality rates. He was also, from the beginning of his time at the Lister, involved in interpreting data from the Institute's ongoing major epidemiologic study of bubonic plague in India. Such studies were to provide a solid basis for Greenwood's subsequent career, with its felicitous blend of epidemiology and medical statistics.

During the Great War Greenwood served in the sanitary service of the Royal Army Medical Corps; in 1916 he was seconded to the Health and Welfare section of the Ministry of Munitions, in charge of statistical work. In 1919 the new Ministry of Health was created, and Greenwood was appointed its first Senior Statistical Officer, working closely with its Chief Medical Officer, George Newman. Never

based in Whitehall, Greenwood worked at the **Medical Research Council's** (MRC) National Institute for Medical Research in Hampstead; he was to be connected with the MRC, in one way or another, for the rest of his career. After nearly two decades, his statistical creed had developed from early dependence on Pearson's rigorous mathematics and "measurement as an end in itself" to an increasingly humane approach reflecting clinical judgment. His pioneering work on large-scale **clinical trials** designed to evaluate prophylactic and therapeutic measures, begun at the Lister, had gradually softened the medical profession's previously hostile attitude to statistical analysis.

Greenwood's mind now turned increasingly to the application of statistics in studies on experimental epidemiology when, in the early 1920s, he became associated with W.W.C. Topley. Topley needed Greenwood's statistical expertise in the studies he was initiating in "experimental epidemiology", using controlled populations of laboratory mice. The investigations were carried out, with various coworkers, until the mid-1930s [3]. By then, Greenwood and Topley had been colleagues, since 1927, as professors of Epidemiology and Vital Statistics, and of Bacteriology, respectively, at the London School of Hygiene and Tropical Medicine. The School was created, with generous Rockefeller support, as the successor to Patrick Manson's old School of Tropical Medicine between 1923 and 1929, when it could finally move into newly built premises in Keppel Street. Here teaching was, as ever, an important part of the staff's responsibilities. In 1935 Greenwood's lectures were published in book form as *Epidemics and Crowd Diseases* [1].

In his lifetime, Greenwood published well over a hundred books and papers on statistical, biometric, epidemiologic, and also historical subjects [4]. In print and in professional discussion his complex personality could on occasion make him appear cynical and censorious; but his intellectual integrity was absolute, as was his loyalty to those admitted to his limited circle of friends. He became a Member of the Royal College of Physicians in 1919, and a Fellow in 1924; in 1928 he was elected to Fellowship of the Royal Society. He retired in 1945 and died, suddenly, in London, during a meeting on cancer research, still active in his seventieth year. He was a founder member, in 1930, of the Socialist Medical Association. A few weeks before his death he argued, as its elected representative, the case for the Voluntary Euthanasia Society, in a broadcast debate arranged by the BBC.

References

[1] Greenwood, M. (1935). *Epidemics and Crowd Diseases*. Williams & Norgate, London.

[2] Greenwood, M. & White, J.D.C. (1909, 1910). A biometric study of phagocytosis with special reference to the opsonic index, 1st and 2nd memoir, *Biometrika* **6**, 376–401, and **7**, 505–530.

[3] Greenwood, M., Hill, A.B., Topley, W.W.C. & Wilson, J. (1936). *Experimental Epidemiology*, MRC Special Report Series No. 209. HMSO London.

[4] Hogben, L. (1950). Major Greenwood 1880–1949, *Obituary Notices of Fellows of the Royal Society* **7**, 139–154.

LISE WILKINSON

Grenander Estimators

Grenander [7], in a remarkable paper, derived the **nonparametric maximum likelihood** estimators (NPMLEs) of

1. a decreasing density
2. an increasing hazard
3. an increasing hazard for repeated events.

Of these, the first has been studied in particular detail by later authors.

NPMLE of a Decreasing Density

Grenander showed that the NPMLE is given by the derivative of the least convex majorant of the empirical distribution function (*see* **Goodness of Fit**). Barlow et al. [3], and its updated version by Robertson et al. [15], put this in the context of general order restricted inference (*see* **Isotonic Inference**). The asymptotic properties of the estimator $\hat{f}_n$ of the density f based on n independent, identically distributed (iid) replications are nonstandard: if $f'(t) < 0$, then

$$n^{1/3} \left|\tfrac{1}{2} f(t) f'(t)\right|^{-1/3} [\hat{f}_n(t) - f(t)]$$

converges in distribution to $2Z$, where Z is distributed as the location of the maximum of the process $\{W(u) - u^2 : -\infty < u < \infty\}$ with W the standard Wiener process on $(-\infty, \infty)$ with $W(0) = 0$ (*see* **Brownian Motion and Diffusion Processes**). Asymptotic distributions of more global measures of discrepancy between $\hat{f}_n$ and f were studied by Groeneboom & Pyke [10] and Groeneboom [8, 9]; Birgé [4] surveyed "nonasymptotic" properties.

For **censored** (survival) data with decreasing density the NPMLE was derived and discussed by Denby & Vardi [5].

Current Status Data

The order restriction for the above NPMLE problem reappears in the nonparametric *current status data* model, first studied by Ayer et al. [1], where n iid replications of pairs of independent random variables (X_i, Y_i) are considered (*see* **Interval Censoring**). Based on observation of $(I\{X_i \leq Y_i\}, Y_i)$ an estimate is sought of the survival function $S(x) = \Pr\{X_1 > x\}$. (Conceptually, all X_i are either left-censored or right-censored by Y_i.) Again, S is constrained to be a decreasing function, and a very similar algorithm is available, particularly studied by Groeneboom (see [11]), who also showed that similar asymptotic results ($n^{1/3}$-convergence) apply. A good general introduction to current status data is [6] (see also [12]). An example as well as details of the distribution of Z can be found in [13]. An important recent paper [2] derived a nonparametric likelihood ratio test for these contexts.

NPMLE of an Increasing Hazard Rate

Grenander's derivation [7] was shown by Barlow et al. [3] to be interpretable as the inverse of the slope (from the right) of the concave majorant of the **total time on test** plot. A similar $n^{1/3}$ asymptotic convergence result as above was given by Prakasa Rao [14].

References

[1] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information, *Annals of Mathematical Statistics* **26**, 641–647.

[2] Banerjee, M. & Wellner, J.A. (2001). Likelihood ratio tests for monotone functions, *Annals of statistics* **29**, 1699–1731.

[3] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. Wiley, London.

[4] Birgé, L. (1989). The Grenander estimator: a nonasymptotic approach, *Annals of Statistics* **17**, 1532–1549.

[5] Denby, L. & Vardi, Y. (1986). The survival curve with decreasing density, *Technometrics* **28**, 359–367.

[6] Diamond, I.D. & McDonald, J.W. (1992). Analysis of current-status data, in *Demographic Applications of Event History Analysis*, J. Trussell, R. Hankinson & J. Tilton, eds. Clarendon Press, Oxford, pp. 231–252.

[7] Grenander, U. (1956). On the theory of mortality measurement. Part II, *Skandinavisk Aktuarietidskrift* **39**, 125–153.

[8] Groeneboom, P. (1985). Estimating a monotone density, in *Proceedings of the Berkeley Conference in Honor of Neyman and Kiefer*, Vol. II. Wadsworth, Belmont, pp. 529–555.

[9] Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions, *Probability Theory and Related Fields* **81**, 79–109.

[10] Groeneboom, P. & Pyke, R. (1983). Asymptotic normality of statistics based on the convex minorants of empirical distribution functions, *Annals of Probability* **11**, 328–345.

[11] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Boston.

[12] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[13] Keiding, N., Begtrup, K., Scheike, T.H. & Hasibeder, G. (1996). Estimation from current-status data in continuous time, *Lifetime Data Analysis* **2**, 119–129.

[14] Prakasa Rao, B.L.S. (1970). Estimation for distributions with monotone failure rate, *Annals of Mathematical Statistics* **41**, 507–519.

[15] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester.

NIELS KEIDING

Group Randomized Trials

Introduction

Group-randomized trials (GRTs) are comparative studies in which investigators randomize identifiable groups to conditions and observe members of those groups to assess the effects of an intervention [8]. In this context, an identifiable group refers quite broadly to any group that is not constituted at random, so that there is some physical, geographic, social, or other connection among its members. Just as the randomized clinical trial (RCT) (*see* **Clinical Trials, Overview**) is the gold standard in public health and medicine when **randomization** of individuals to study conditions is possible, the GRT is the gold standard in public health and medicine when randomization of identifiable groups is required. This situation exists whenever the investigator wants to evaluate an intervention that operates at a group level, manipulates the social or physical environment, or cannot be delivered to individuals (*see* **Cluster Randomization**).

An Example

TEENS was a group-randomized trial occurring in 16 middle schools in the Twin Cities metropolitan area from 1997 to 2000 [7]. Schools agreeing to be in the study committed to the measurement protocol, randomization to condition, and if randomized to the intervention condition, to the following intervention protocol: (a) offer all 10 sessions of the TEENS curriculum in 7th and 8th grade; (b) allow the designated teacher to attend a full day training each year; (c) allow for provision of a family education component; and (d) allow school food service staff to be trained on modifying the school food environment. **Sample size** calculations, based on fruit, vegetables, and fat intake data from prior school-based studies, indicated that with 16 schools and at least 30 students measured per grade, we had 80% **power** to detect differences of 1.1 servings of fruits and vegetables and a 1.9% difference in energy from total fat intake between treatment groups. All students who were in 7th grade during the baseline data collection period were considered eligible to participate.

The primary **outcome measures** for evaluating the effectiveness of TEENS were student-level intake of fruits, vegetables, and energy from fat based on 24-hour dietary recalls. Data were collected at baseline at the beginning of the 7th grade in Fall 1998 and again at the end of the 7th and 8th grades in Spring 1999 and 2000.

Schools were randomly assigned from within matched pairs (*see* **Matching**) to either control or intervention condition. Schools were matched on both the proportion of 7th graders expected to receive the TEENS curriculum and on the proportion receiving free or reduced school lunch; randomization was constrained so that the four smallest schools were distributed with two in each of the two conditions. The eight intervention schools received the TEENS intervention and related training for two consecutive years beginning when the grade cohort was in the 7th grade (1998–1999) and continuing through the 8th grade year (1999–2000). The end of 7th grade data suggested a significant increase in consumption of fruits and vegetables in the intervention condition [1]; the end of 8th grade data suggested that those effects were not sustained through two years of follow-up [7].

Distinguishing Characteristics

There are four characteristics that distinguish the GRT from the more familiar RCT [8]. First, the unit of assignment is an identifiable group; such groups are not formed at random, but rather through some physical, social, geographic, or other connection among their members. Second, different groups are assigned to each condition, creating a nested or hierarchical structure for the design and the data (*see* **Hierarchical Models**). Third, the units of observation are members of those groups so that they are nested within both their condition and their group. Fourth, there usually are only a limited number of groups assigned to each condition.

These characteristics create several problems for the design and analysis of GRTs. The major design problem is that a limited number of often heterogeneous groups makes it difficult for randomization to distribute potential sources of **confounding** evenly in any single realization of the experiment. This increases the need to employ design strategies that will limit confounding and analytic strategies to deal with confounding where it is detected. The major analytic problem is that there is an expectation for positive intraclass **correlation** (ICC) among observations

on members of the same group [6]. That ICC reflects an extra component of variance (*see* **Variance Components**) attributable to the group above and beyond the variance attributable to its members. This extra variation will increase the variance of any group-level statistic beyond what would be expected with random assignment of members to conditions. Moreover, with a limited number of groups, the **degrees of freedom** (df) available to estimate group-level statistics are limited. Any test that ignores either the extra variation or the limited df will have a Type I error rate (*see* **Hypothesis Testing**) that is inflated [3].

The Development of Group-randomized Trials in Public Health

GRTs gained attention in public health in the late 1970s with the publication of a symposium on coronary heart disease prevention trials in the *American Journal of Epidemiology*. **Cornfield's** paper in particular has become quite well known among methodologists working in this area, as it identified the two issues that have vexed investigators who employ GRTs from the outset: extra variation and limited degrees of freedom [3].

The last 25 years have witnessed dramatic growth in the number of GRTs in public health and dramatic improvements in the quality of the design and analysis of those trials. Responding directly to Cornfield's warning, Donner and colleagues at the University of Western Ontario published a steady stream of papers on the issues of analysis facing group-randomized trials through the 1980s and 1990s. Murray and colleagues from the University of Minnesota began their examination of the issues of design and analysis in group-randomized trials in the mid-1980s. Other investigators from the **National Institutes of Health**, the University of Washington, the New England Research Institute, and elsewhere added to this growing literature in public health, especially in the 1990s.

In the 1998, the first textbook on the design and analysis of GRTs appeared [8]. It detailed the design considerations for the development of GRTs, described the major approaches to their analysis both for Gaussian (*see* **Normal Distribution**) and **binary data**, and presented methods for power analysis applicable to most GRTs. The second textbook on the design and analysis of GRTs appeared in 2000 [4]. It provided a good history on GRTs, examined the role of informed consent and other ethical issues (*see* **Ethics of Randomized Trials**), focused on extensions of classical methods, and included material on regression models for Gaussian, binary, count, and time-to-event data. Murray et al. recently reviewed a large number of articles on new methods relevant to the design and analysis of GRTs published between 1998 and 2003 [10].

Potential Design Problems and Methods to Avoid Them

For GRTs, there are four sources of **bias** that should be considered during the planning phase: **selection**, differential history, differential maturation, and contamination. These biases are well known and may also occur in RCTs. The first three are particularly problematic in GRTs where the number of units available for randomization is often small. GRTs planned with fewer than 20 groups per condition would be well served to include careful matching or **stratification** prior to randomization to help avoid these biases. Analytic strategies, such as regression adjustment for confounders, can be very helpful in dealing with any observed bias.

Potential Analytic Problems and Methods to Avoid Them

There are two major threats to the validity of the analysis of a GRT, which should be considered during the planning phase: misspecification of the analytic model and low power. **Misspecification** of the analytic model most commonly occurs when the investigator fails to reflect the expected ICC in the analytic model. Low power most commonly occurs when the design is based on an insufficient number of groups randomized to each condition.

There are several analytic approaches that can provide a valid analysis for GRTs [4, 8]. In most, the intervention effect is defined as a function of a condition-level statistic (e.g. difference in means, rates, or slopes) and assessed against the variation in the corresponding group-level statistic. These approaches included mixed-model ANOVA/ANCOVA (*see* **Analysis of Variance**; **Analysis of Covariance**) for designs having only one or two time intervals, random coefficient models for designs having three or more time intervals, and **randomization**

tests as an alternative to the model-based methods. Other approaches are generally regarded as invalid for GRTs because they ignore or misrepresent a source of random variation. These include analyses that assess condition variation against individual variation and ignore the group, analyses that assess condition variation against individual variation and include the group as a **fixed effect**, analyses that assess the condition variation against subgroup variation, and analyses that assess condition variation against the wrong type of group variation. Still other strategies may have limited application for GRTs. Application of fixed-effect models with post hoc correction for extra variation and limited df assumes that the correction is based on an appropriate ICC estimate. Application of survey-based methods or generalized estimating equations (GEE) and the sandwich method for **standard errors** requires a total of 40 or more groups in the study, or a correction for the downward bias in the sandwich estimator for standard errors when there are fewer than 40 groups in the study [10].

Low power will occur if the investigator employs a weak intervention, has insufficient replication, has high variance or intraclass correlation in the endpoints, or has poor reliability of intervention implementation. To avoid low power, investigators should plan a large enough study to ensure sufficient replication, employ more and smaller groups instead of a few large groups, employ strong interventions with good reach, and maintain the reliability of intervention implementation. In the analysis, investigators should consider regression adjustment for **covariates**, model time if possible, and consider post hoc stratification.

A detailed exposition on power for GRTs is beyond the scope of this article. Excellent treatments exist, and the interested reader is referred to those sources for additional information. Chapter 9 in the Murray text provides perhaps the most comprehensive treatment of detectable difference, sample size, and power for GRTs [8]. Even so, a few points bear repeating here. First, the increase in between-group variance due to the ICC in the simplest analysis is calculated as $1 + (m - 1)ICC$, where m is the number of members per group; as such, ignoring even a small ICC can underestimate standard errors if m is large. Second, while the magnitude of the ICC is inversely related to the level of aggregation, it is independent of the number of group members who provide data. For both these reasons, more power is available given more groups per condition with fewer members measured per group than given just a few groups per condition with many members measured per group, no matter the size of the ICC. Third, the two factors that largely determine power in any GRT are the ICC and the number of groups per condition. For these reasons, there is no substitute for a good estimate of the ICC for the primary endpoint, the **target population**, and the primary analysis planned for the trial, and it is unusual for a GRT to have adequate power with fewer than 8 to 10 groups per condition. Finally, the formula for the standard error for the intervention effect depends on the primary analysis planned for the trial, and investigators should take care to calculate that standard error, and power, based on that analysis.

The Future of Group-randomized Trials

Whenever the investigator wants to evaluate an intervention that operates at a group level, manipulates the social or physical environment, or cannot be delivered to individuals, the GRT is the best comparative design available. Even so, there remain many challenges facing GRTs. For example, there can be no question that it is harder to change the health behavior and risk profile of a whole community than it is to make similar changes in smaller identifiable groups such as those at worksites, physician practices, schools, and churches. And while no quantitative analysis has been published, it seems that the magnitude of the intervention effects reported for GRTs has been greater for trials that involved smaller groups than for trials involving such large aggregates as whole communities. With smaller groups, it is possible to include more groups in the design, thereby improving the validity of the design and the power of the trial. With smaller groups, it is easier to focus intervention activities on the target population. With smaller groups, the cost and difficulty of the implementation of the study generally are reduced. For these and similar reasons, future group-randomized trials may do well to focus on more and smaller identifiable groups rather than on whole cities or larger aggregates.

Another challenge is simply the difficulty in developing interventions strong enough to change the health behaviors of the target populations. The methods for the design and analysis of GRTs have evolved considerably from the 1970s and 1980s, but we continue to employ interventions that often prove ineffective. One of the problems for some time has been that

interventions are proposed, which lack even preliminary evidence of efficacy [9]. Efficacy trials in health promotion and disease prevention (*see* **Prevention Trials**) often are begun without the benefit of prototype studies, and often even without the benefit of adequate pilot studies. This has happened in large part because the funding agencies have been reluctant to support pilot and prototype studies, preferring instead to fund efficacy and effectiveness trials. Unfortunately, the interventions that lead to GRTs tend to be more complicated than those in other areas or those that lead to clinical trials. As such, it is even more important to subject them to adequate testing in pilot and prototype studies. These earlier phases of research can uncover important weaknesses in the intervention content or implementation methods. Moving too quickly to efficacy trials risks wasting substantial time and resources on interventions that could have been substantially improved through the experience gained in those pilot and prototype studies. One would hope that the funding agencies will recognize this point and begin to provide better support for pilot and prototype studies.

There are remaining methodological challenges as well. For example, there have been a number of recent studies that documented the downward bias in the sandwich estimator used in GEE when there are fewer than 40 groups in the study [5]. Corrections have been proposed, but none appear in the standard software packages (*see* **Software, Biostatistical**), so those corrections are relatively unavailable to investigators who analyze GRTs. Absent an effective correction, the sandwich estimator will have an inflated Type I error rate in GRTs having less than 40 groups, and investigators who employ this approach continue to risk overstating the significance of their findings.

As another example, consider studies that employ only one or a few groups per condition. With only one group per condition, it is not possible to separately estimate variation due to groups and condition, and so there is no valid analysis or absent strong and untestable assumptions. With only a few groups per condition, power is likely to be extremely limited, and so such studies are to be discouraged, except perhaps as pilot studies.

As another example, there have been a number of recent studies that proposed methods for **survival analysis** that could be applied to data from GRTs. Some of these methods involved use of the sandwich estimator, and so would be subject to the same concern as noted above for GEE.

As a third example, permutation tests (*see* **Randomization Tests**) have been advocated over model-based methods because they require fewer assumptions. At the same time, they tend to have lower power. To overcome this problem, Feng et al. developed an optimal randomization test that had nominal size and better power than alternative randomization tests or GEE, though it was still not as powerful as the model-based analysis when the model was specified correctly [2]. Additional research is needed to compare Braun & Feng's optimal randomization test and model-based methods under model misspecification.

There is every reason to expect that continuing methodological improvements will lead to better trials. There is also evidence that better trials tend to have more satisfactory results. For example, Rooney and Murray presented the results of a **meta-analysis** of group-randomized trials in the smoking-prevention field [11]. One of the findings was that stronger intervention effects were associated with greater methodological rigor. Stronger intervention effects were reported for studies that planned from the beginning to employ the unit of assignment as the **unit of analysis**, that randomized a sufficient number of assignment units to each condition, that adjusted for baseline differences in important confounding variables, that had extended follow-up, and that had limited attrition. One hopes that such findings will encourage the use of good design and analytic methods.

Acknowledgment

The material presented here draws heavily on work published previously by David M. Murray [8–10]. Readers are referred to those sources for additional information.

References

[1] Birnbaum, A.S., Lytle, L.A., Story, M., Perry, C.L. & Murray, D.M. (2002). Are differences in exposure to a multicomponent school-based intervention associated with varying dietary outcomes in adolescents? *Health Education and Behavior* **29**(4), 427–443.

[2] Braun, T. & Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials, *Journal of the American Statistical Association* **96**, 1424–1432.

[3] Cornfield, J. (1978). Randomization by group: a formal analysis, *American Journal of Epidemiology* **108**(2), 100–102.

[4] Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London.
[5] Feng, Z., McLerran, D. & Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with Gaussian error, *Statistics in Medicine* **15**, 1793–1806.
[6] Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, New York.
[7] Lytle, L., Perry, C., Murray, D.M., Story, M., Birnbaum, A., Kubik, M., & Varnell, S. School-based approaches to affecting adolescents' diets: Results from the TEENS study, *Health Education and Behavior*, **31**(2), 270–287.
[8] Murray, D.M. (1998). *Design and Analysis of Group-Randomized Trials*. Oxford University Press, New York.
[9] Murray, D.M. (2000). Efficacy and effectiveness trials in health promotion and disease prevention: design and analysis of group-randomized trials, in *Integrating Behavioral and Social Sciences With Public Health*, N. Schneiderman, M. Speers, J. Silva, H. Tomes, & J. Gentry, eds. American Psychological Association, Washington, pp. 305–320.
[10] Murray, D.M., Varnell, S.P. & Blitstein, J.L. Design and analysis of group-randomized trials: a review of recent methodological developments, *American Journal of Public Health*, **94**(3), 423–432.
[11] Rooney, B.L. & Murray, D.M. (1996). A meta-analysis of smoking prevention programs after adjustments for errors in the unit of analysis, *Health Education Quarterly* **23**(1), 48–64.

DAVID M. MURRAY

Grouped Data

We call data *grouped* when we observe only some set Y containing the variable of interest X rather than the value of X itself. For example, suppose X represents the length and width of an iris petal, which we can only measure to the nearest millimeter. If the observed measurements are 1.4 cm and 0.2 cm, then $X \in y = (1.35, 1.45) \times (0.15, 0.25)$. Because we can neither measure nor store data with infinite precision, all so-called "continuous" data are actually grouped.

There are two general approaches to analyzing grouped data. The first is to substitute the center $\tilde{Y}$ of Y for X, proceeding as if the data were known exactly. One may be able to correct such estimates to remove grouping effects, as with **Sheppard's corrections** for rounded data. The second approach is to base inferences directly on the distribution of Y or the **likelihood** arising from it. For example, if X follows a bivariate density $f(x;\theta)$, then the likelihood from nominal measurement $\tilde{y} = (1.4, 0.2)$ is

$$L(y;\theta) = \int_{(1.35,1.45)\times(0.15,0.25)} f(x;\theta)\,\mathrm{d}x.$$

A simple approximation is

$$\tilde{L}(y;\theta) = f(\tilde{y};\theta) = f(1.4, 0.2;\theta).$$

Typically, exact methods are difficult to implement but give better answers for large n and coarse groups.

An interesting special case of grouping is *heaping*, where a single data set contains items rounded with various levels of coarseness. Heaping can occur in self-reported variables such as age, income, and cigarette consumption. Another special case is **interval censoring**, where a continuous failure time is only known to lie between two (potentially random) limits.

Grouping has been the subject of research throughout the modern era of statistics. See [1], [2], and [3] for reviews.

References

[1] Gjeddebaek, N.F. (1968). Statistical analysis: grouped observations, in *International Encyclopedia of the Social Sciences*, Vol. 15. Macmillan and the Free Press, New York.

[2] Haitovsky, Y. (1982). Grouped data, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 527–536.

[3] Heitjan, D.F. (1989). Inference from grouped continuous data: a review, *Statistical Science* **4**, 164–183.

(*See also* **Categorizing Continuous Variables**)

DANIEL F. HEITJAN

Grouped Survival Times

In many investigations involving survival times, data are grouped prior to their statistical analysis (*see* **Grouped Data**). The grouped survival data consist of occurrence and exposure data over given time intervals and possible covariate strata. For grouped survival (or *failure-times*) data there is an assumed continuous underlying *hazard* function, in contrast to **discrete survival-time** data, with an intrinsically discrete time variable, discrete hazards, survival functions, etc.

One of the primary reasons for grouping can be found in studies involving large sample sizes such as epidemiologic studies [6]. Such studies typically involve the follow-up of large population groups over certain time periods to assess the cause and rate of death and/or to compare death rates among different population groups. Grouping data from such large sample sizes into tabular presentations (**life tables**) often provides a convenient format for presenting and summarizing life information. Also, grouping could be done intentionally, for example, to economize on data transmission and storage, to reduce computation, to protect the privacy of individual records, or to account for the limitations of a measurement instrument. Moreover, some large data sets are publicly released only in grouped form, as discussed by Haitovsky [16, 17]. Some examples that illustrate such grouped survival data are: the American Cancer Society study of 1 000 000 men and women [18] to determine the dose–time–response relationships between **smoking** and lung cancer or heart disease and the life span study of over 100 000 Japanese atom bomb survivors in Hiroshima and Nagasaki [4].

Another important reason for grouping data is that it is often difficult or even impossible to obtain exact lifetimes, because ethical, physical, or economic restrictions in research design allow the subjects in the follow-up study to be monitored only periodically. Thus, this type of study only provides grouped information, that is, the exact failure time is unknown and the only available information is whether the event of interest occurred between two inspection times. The following study illustrates situations where periodic inspection is used: the National Labor Survey of Youth (NLSY) study of time to weaning of breast-fed newborns in which 927 first-born children of mothers who chose to breast-feed their children were interviewed yearly.

Similar to continuous data in survival analysis, grouped survival data can involve **censored data** (right censoring, left censoring or double censoring) and/or **truncated** data. Moreover, the exact censoring or truncation times may be unknown for grouped data. For example, in the study of the time to weaning of breast-fed newborns, some infants were lost to follow-up and some infants were withdrawn from the study without being weaned. Also, grouped survival data can involve **covariates (explanatory variables)**. Some **parametric models** and the well-known **Cox regression model** are often fitted to grouped survival data [34].

The vast literature on grouped survival data involves: deriving the estimators of the hazard function and survival function under nonparametric or parametric models, test statistics for comparing the survival probabilities among different population groups, and large sample properties for these estimators and test statistics. Most estimates are derived based on **maximum likelihood** methods. Some references to such studies will be given later. The **Bayesian** approach to analyzing grouped survival data has also been studied in the literature [8, 24].

Notation of Grouped Survival Data

Let time be partitioned into a fixed sequence of intervals $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m$ with $\mathcal{T}_j = (t_{j-1}, t_j]$ and $0 = t_0 < t_1 < \cdots < t_m \leq \infty$. For grouped failure time data the only available information is:

n_j = number of subjects entering $\mathcal{T}_j$ not having experienced the event,

d_j = number of individuals experiencing the event in $\mathcal{T}_j$,

w_j = number of individuals lost to follow-up or withdrawn during $\mathcal{T}_j$ and,

Y_j = total time of individuals at risk during $\mathcal{T}_j$.

When the subjects are monitored periodically, the total time at risk Y_j is unknown. It is often approximated by $Y_j \approx [n_j - (d_j + w_j)/2](t_j - t_{j-1})$ for right-censored data.

Life Table

The **life table** is one of the oldest and most commonly used methods of presenting lifetime data. It is a table

for presenting and summarizing data, and estimating the survival function, the probability density function and the hazard function along with the variance of these estimators. For more details on the life table, see [15] and [7].

Interval Censored Grouped Data

For interval (doubly) censored grouped data, Turnbull [36, 37] proposed a "self-consistency" procedure, developed by Efron [11], to estimate the survival function $S(t)$. The Turnbull estimator is a **nonparametric maximum likelihood** estimator (NPMLE). Its derivation and asymptotic properties are discussed in the article on the **Turnbull estimator**. Some alternative approaches to maximizing the NPMLE are discussed in the article on **Interval Censoring**.

Logrank Test

Comparison of the survival probabilities with treatment groups or covariate strata in the grouped data can be done through rank tests. In the continuous data case Fleming & Harrington [13] studied a class of weighted **logrank tests**. These weighted logrank tests can be extended to the grouped failure time data. The usual logrank test (or evenly weighted logrank test) is most commonly and widely used in practice. Here we discuss the grouped data version of the logrank test. First, consider the two-sample case. Let n_{ij} and d_{ij}, $j = 1, \ldots, m$, $i = 1, 2$, be the number at risk at the beginning of the jth interval and observed failures in the jth interval, respectively, in sample i. Take n_j and d_j to be the corresponding values in the combined sample. The data, corresponding to the jth time interval, can be summarized as shown in Table 1. The grouped data based two-sample logrank test can be computed as

$$Q = \left[\sum_{j=1}^{m}(d_{1j} - E_{1j})\right]^2 \Big/ \sum_{j=1}^{m} V_{1j},$$

where E_{1j} and V_{1j} are the expected value and variance of d_{1j}, given by

$$E_{1j} = \frac{d_j n_{1j}}{n_j} \quad \text{and} \quad V_{1j} = \frac{d_j n_{1j} n_{2j}(n_j - d_j)}{n_j^2(n_j - 1)}.$$

Under the hypothesis of $S_1(t) = S_2(t)$, the two-sample logrank test statistic Q has approximately the **chi-square distribution** with 1 **degree of freedom** when the sample sizes are moderately large for each sample.

We can extend the two-sample logrank test to the k-sample comparison. The k-sample logrank test has a quadratic form with $(d_{1j} - E_{1j})$ replaced by the corresponding values from $k - 1$ samples and with V_{1j} replaced by the corresponding **covariance matrix**, where the hlth element is

$$\hat{\sigma}_{hl} = \frac{d_j n_{hj}}{n_j}\left(\delta_{hl} - \frac{n_{hj}}{n_j}\right)\frac{(n_j - d_j)}{(n_j - 1)}$$

and δ_{hl} is a Kronecker delta, that is, $\delta_{hl} = 1$ if $h = l$, and 0 otherwise.

Table 1

	Sample		
Failure	1	2	Total
Yes	d_{1j}	d_{2j}	d_j
No	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Total	n_{1j}	n_{2j}	n_j

Parametric Models and Regression Analysis

In survival analysis some parametric models have been studied extensively. The common parametric distributions considered are **exponential**, **gamma**, **Weibull**, **lognormal**, and Gompertz distributions (*see* **Parametric Models in Survival Analysis**). These parametric models are often fitted to grouped data as well. The parameters are usually estimated by maximizing the full (unconditional) **likelihood** or the conditional likelihood, that is, the likelihood function for the interval $(t_{j-1}, t_j]$ conditional on surviving till t_{j-1}. Many authors have given the maximum likelihood estimator (MLE) for grouped data [12, 27, 10]. Turnbull & Weiss [38] studied a **likelihood ratio test** statistic for testing **goodness of fit** for grouped failure data with possible double censoring.

It is important to assess the effects of covariates that may be associated with lifetimes in many applications of survival analysis. The regression model

for the conditional hazard function $\lambda(t|\mathbf{z})$ of the failure time given covariate $\mathbf{z}$ could be used to examine the covariate effects. Continuous covariates are often grouped into a fixed number of strata, and the value for each stratum is approximated by the midpoint of the covariate in the stratum (*see* **Stratification**). For simplicity we consider a one-dimensional covariate case. The methods and results discussed here can be extended to multidimensional cases. Let the cells into which the data are grouped be denoted $\mathcal{C}_{rj} = \mathcal{T}_r \times \mathcal{I}_j$, where $\mathcal{T}_1, \ldots, \mathcal{T}_{L_n}$ and $\mathcal{I}_1, \ldots, \mathcal{I}_{J_n}$ are the respective calendar periods (time intervals) and covariate strata. Grouped failure time data consist of the total number of failures (occurrence) and the total time at risk (exposure) in each cell $\mathcal{C}_{rj}$, given by d_{rj} and Y_{rj}. In the literature, most early work has been done under the piecewise exponential model, that is, the hazard function is assumed to be piecewise constant within each grouping cell. The natural estimate of the unknown hazard rate λ_{rj} is $\hat{\lambda}_{rj} = d_{rj}/Y_{rj}$ (occurrence/exposure rate). Deddens & Koch [10] showed that the maximum likelihood solution is approximately equivalent to maximizing the piecewise exponential likelihood function

$$L = \prod_{r,j} \lambda_{rj}^{d_{rj}} [\exp(-\lambda_{rj} Y_{rj})].$$

The occurrence/exposure rate estimator can also be obtained by solving the equations of $\partial \log L / \partial \lambda_{rj} = 0$.

The **counting process** approach and martingale techniques are applicable in grouped failure time data analysis. We assume that the counting process N_i, where $N_i(t)$ is the number of failures of the ith individual during time period $[0, t]$, has intensity

$$\lambda_i(t) = Y_i(t)\lambda[t, Z_i(t)],$$

where $Y_i(t)$ is a predictable $(0, 1)$-valued process indicating that the ith individual is at risk with $Y_i(t) = 1$, and $Z_i(t)$ is a predictable covariance process. The occurrence and exposure in each cell $\mathcal{C}_{rj}$ can be written as

$$d_{rj} = \sum_i \int_{\mathcal{T}_r} I[Z_i(t) \in \mathcal{I}_j] \, \mathrm{d}N_i(t)$$

and

$$Y_{rj} = \sum_i \int_{\mathcal{T}_r} I[Z_i(t) \in \mathcal{I}_j] Y_i(t) \, \mathrm{d}t.$$

When the censoring processes are independent of the survival time, we can show that $M_i(t) = N_i(t) - \int_0^t \lambda_i(u) \, \mathrm{d}u$ are local martingales. Under the piecewise constant model $[\lambda(t, z) = \lambda_{rj}$, for $(t, z) \in \mathcal{C}_{rj}]$,

$$\hat{\lambda}_{rj} = \frac{d_{rj}}{Y_{rj}} = \frac{M_{rj}}{Y_{rj}} + \lambda_{rj} \frac{Y_{rj}}{Y_{rj}},$$

where

$$M_{rj} = \sum_i \int_{\mathcal{T}_r} I[Z_i(t) \in \mathcal{I}_j] \, \mathrm{d}M_i(t)$$

is the martingale part of d_{rj}. Since for each $t \in \mathcal{T}_r$, Y_{rj} is not predictable, the martingale techniques are not applicable directly. However, in the independent, identically distributed (iid) case and some mild conditions, we can show that there exists a piecewise constant function f_{rj} bounded away from zero such that $n^{-1} Y_{rj}$ **converges** to f_{rj} in probability. Then we can replace M_{rj}/Y_{rj} by M_{rj}/nf_{rj} with the difference of $o_P(1)$. It follows that

$$\hat{\lambda}_{rj} = \frac{M_{rj}}{nf_{rj}} + \lambda_{rj} + o_P(1),$$

and the predictable variation process of M_{rj}/f_{rj} is

$$\left\langle \frac{M_{rj}}{f_{rj}} \right\rangle = \frac{\lambda_{rj} Y_{rj}}{f_{rj}^2}.$$

Therefore, $\hat{\lambda}_{rj}$ is an asymptotic **unbiased** estimator and the variance can be consistently estimated by

$$\hat{\sigma}_{rj} = \widehat{\mathrm{var}}(\hat{\lambda}_{rj}) = \frac{d_{rj}}{(Y_{rj})^2}.$$

For the general nonparametric model where the hazard function is unspecified, Holford [20] noted that this estimator is inconsistent unless the grouping becomes finer as the sample size increases (*see* **Consistent Estimator**).

The useful models for many applications are the **multiplicative** and **additive hazard models**. The model equations are given by

$$\lambda_{rj} = \lambda_{r0} \exp(\boldsymbol{\beta} \mathbf{z}_j) \quad \text{and} \quad \lambda_{rj} = \lambda_{r0} + \boldsymbol{\beta} \mathbf{z}_j,$$

where λ_{r0} is the baseline hazard rate for the rth time period. The parameters λ_{r0} and $\boldsymbol{\beta}$ are readily estimated by the MLE. Berry [5] and Frome [14] provide explicit MLE for this approach. For the multiplicative risk model the hazard function can be written

as $\lambda_{rj} = \exp(\alpha_r + \boldsymbol{\beta}\mathbf{z}_j)$, which has a loglinear form. It is often called the loglinear piecewise constant model. Holford [21] derived the log likelihood for this model:

$$L = \sum_r \alpha_r d_{r\cdot} + \sum_{r,j} d_{rj}\beta z_j - \sum_{r,j} Y_{rj}\exp(\alpha_r + \boldsymbol{\beta}\mathbf{z}_j),$$

where $d_{r\cdot} = \sum_{j=1}^{J_n} d_{rj}$ is the number of failures in the rth calendar period. Taking derivatives of L with respect to α_r and $\boldsymbol{\beta}$ and setting them equal to zero, the MLE estimator of $\boldsymbol{\beta}$ is given by solving the following equation:

$$\sum_{r,j} z_j d_{rj} - \sum_r \frac{\sum_j Y_{rj} z_j \exp(\boldsymbol{\beta}\mathbf{z}_j)}{\sum_j Y_{rj}\exp(\boldsymbol{\beta}\mathbf{z}_j)} d_{r\cdot} = 0.$$

As we discuss later, this MLE estimator of $\boldsymbol{\beta}$ can also be obtained by maximizing the grouped data version of Cox's **partial likelihood**.

The more general models are: Cox's proportional hazards model [9], where $\lambda(t, z) = \lambda_0(t)\exp(\boldsymbol{\beta}\mathbf{z})$; and **Aalen's additive regression model** [1], where $\lambda(t, z) = \lambda_0(t) + \boldsymbol{\beta}(t)\mathbf{z}$.

Cox's proportional hazards model has so far been the most popular model in survival analysis. The parameter estimator $\hat{\boldsymbol{\beta}}$ is obtained by maximizing Cox's partial likelihood function. Andersen & Gill [3] provide an excellent proof that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{P} \mathrm{N}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V}^{-1}$ is consistently estimated by $-n^{-1}\partial\mathbf{U}(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}$ and $\mathbf{U}$ is the partial likelihood score function $U(\boldsymbol{\beta}) = \partial \log L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ (*see* **Likelihood**). The grouped data based estimator $\hat{\boldsymbol{\beta}}_g$ can be obtained by maximizing the following approximation to the partial likelihood:

$$L_g(\boldsymbol{\beta}) = \prod_{r,j}\left(\frac{\exp(\boldsymbol{\beta}\mathbf{z}_j)}{\sum_k Y_{rk}\exp(\boldsymbol{\beta}\mathbf{z}_k)}\right)^{d_{rj}},$$

where the product is over the grouping cells, the sum is over the covariate strata, and z_j is the midpoint of the jth covariate stratum. This estimator has been studied by Kalbfleisch & Prentice [25], Holford [20], Prentice & Gloeckler [34], Breslow [6], Hoem [19], Selmer [35], and Huet & Kaddour [22]. It can be interpreted as the maximum likelihood estimator in a **Poisson regression** model, as shown by Laird & Olivier [26]. Under slightly stronger regularity conditions proposed in Andersen & Gill [3], it can be shown that $n^{1/2}(\hat{\beta}_g - \beta_0) \xrightarrow{P} \mathrm{N}(0, V)$ when the time intervals and covariate strata shrink at some suitable rate as the sample size increases. It is important to be able to assess estimation bias caused by grouping and to correct it if necessary. In the general grouped data analysis, a **Sheppard correction** can be used to reduce the bias to a higher order of the interval width [28]. McKeague & Zhang [33] obtained a Sheppard correction for Cox's proportional hazards model, provided a consistent estimator for the Sheppard correction, and derived the optimal rate of convergence for $\hat{\beta}_g$. The grouped data based estimator of the baseline hazard function, λ_0, is

$$\hat{\lambda}_0(t) = \frac{\sum_j d_{rj}}{\sum_j Y_{rj}\exp(\hat{\boldsymbol{\beta}}_g\mathbf{z}_j)}, \qquad \text{for } t \in \mathcal{T}_r.$$

Aalen's additive risk model provides a useful and sometimes biologically more plausible alternative to the Cox proportional hazards model. For continuous data, Aalen proposed a **least squares** estimator for the cumulative hazard functions, which has been studied by Aalen [1, 2], Mau [29, 30], and McKeague [31]. McKeague [32] and Huffer & McKeague [23] fit Aalen's additive risk model to the grouped data (when the covariates are observed for each individual and are non-time-dependent), and studied asymptotic results for the grouped data version of the least squares estimator and weighted least squares estimator. The estimators can be generalized to the more general grouped data setting where the only available information is d_{rj} and Y_{rj} for each cell $\mathcal{C}_{rj}$. More work is needed.

Finally, fitting parametric and regression models to grouped failure time data is based on d_{rj} and Y_{rj}. As we discussed in the univariate case, Y_{rj} may not be observable in some applications. It is usually approximated by $Y_{rj} \approx [n_{rj} - (d_{rj} + w_{rj})/2]l_r$, where n_{rj} is the number of individuals at risk at the beginning of the time period $\mathcal{T}_r$ for the jth covariate stratum, w_{rj} is the number of individuals who withdrew in cell $\mathcal{C}_{rj}$, and l_r is the width of the time interval $\mathcal{T}_r$. This approximation is based on the assumption that, on the average, the individuals failed or withdrew at the middle of the each time period. However, in

most applications, this assumption does not hold true. The bias introduced by this approximation could be severe. Caution must be taken when grouping the data to ensure that the number of grouping cells is sufficiently large (the widths of time periods and covariate strata are relative small) and each grouping cell contains enough individuals at risk.

References

[1] Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes, in *Springer Lecture Notes on Mathematical Statistics and Probability*, Vol. 2, W. Klonecki, A. Kozek & J. Rosinski, eds. Springer-Verlag, New York.

[2] Aalen, O.O. (1989). A linear regression model for the analysis of life time, *Statistics in Medicine* **8**, 907–925.

[3] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[4] Beebe, G.W. (1981). The atomic bomb survivors and problem of low dose radiation effects, *American Journal of Epidemiology*, **114**, 761–783.

[5] Berry, G. (1983). The analysis of mortality by the subject-years method, *Biometrics* **39**, 173–184.

[6] Breslow, N.E. (1986). Cohort analysis in epidemiology, in *A Celebration of Statistics: The ISI Centenary Volume*, A.C. Atkinson & S.E. Fienberg, eds. Springer-Verlag, New York, pp. 109–143.

[7] Breslow, N. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics* **2**, 437–453.

[8] Cornfield, J. & Detre, K. (1977). Bayesian life table analysis, *Journal of the Royal Statistical Society, Series B* **39**, 86–94.

[9] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[10] Deddens, J.A. & Koch, G.G. (1988). Survival analysis, grouped data, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 129–134.

[11] Efron, B. (1967). The two-sample problem with censored data, in *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. University of California Press, Berkeley, pp. 831–853.

[12] Elandt-Johnson, R.C. & Johnson, N.L. (1980). *Survival Models and Data Analysis*. Wiley, New York.

[13] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[14] Frome, E.L. (1983). The analysis of rates using Poisson regression models, *Biometrics* **39**, 665–674.

[15] Gehan, E.A. (1969). Estimating survival functions from the life table, *Journal of Chronic Diseases* **21**, 629–644.

[16] Haitovsky, Y. (1973). *Regression Estimation from Grouped Observations*. Griffin, London/Hafner Press, New York.

[17] Haitovsky, Y. (1983). Grouped data, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 527–536.

[18] Hammond, E.C. (1966). Smoking in relation to the death rates of one million men and women, *National Cancer Institute Monograph* **19**, 127–204.

[19] Hoem, J.M. (1987). Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review, *International Statistical Review* **55**, 119–152.

[20] Holford, T.R. (1976). Life tables with concomitant information, *Biometrics* **32**, 587–597.

[21] Holford, T.R. (1980). The analysis of rates and of survivalship using log-linear models, *Biometrics* **36**, 299–305.

[22] Huet, S. & Kaddour, A. (1994). Maximum likelihood estimation in survival analysis with grouped data on censored individuals and continuous data on failures, *Applied Statistics* **43**, 325–333.

[23] Huffer, F.W. & McKeague, I.W. (1991). Weighted least squares estimation for Aalen's additive risk model, *Journal of the American Statistical Association* **86**, 114–129.

[24] Johnson, W. & Christensen, R. (1986). Bayesian nonparametric survival analysis for grouped data, *Canadian Journal of Statistics* **14**, 307–314.

[25] Kalbfleisch, J.D. & Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model, *Biometrika* **60**, 267–278.

[26] Laird, N. & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques, *Journal of the American Statistical Association* **76**, 231–240.

[27] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

[28] Lindley, D.V. (1950). Grouping corrections and maximum likelihood equations, *Proceedings of the Cambridge Philosophical Society* **46**, 106–110.

[29] Mau, J. (1986). On a graphical method for the detection of time-dependent effects of covariates in survival data, *Applied Statistics* **35**, 245–255.

[30] Mau, J. (1988). A comparison of counting process models for complicated life histories, *Applied Stochastic Models and Data Analysis* **4**, 283–298.

[31] McKeague, I.W. (1988). Asymptotic theory for weighted least squares estimators in Aalen's additive risk model, *Contemporary Mathematics* **80**, 139–152.

[32] McKeague, I.W. (1988). A counting process approach to the regression analysis of grouped survival data, *Stochastic Processes and Their Applications* **28**, 221–239.

[33] McKeague, I.W. and Zhang, M.J. (1996). Fitting Cox's proportional hazards model using grouped survival data,

in N.P. Jewell, A.C. Kimber, M.L.T. Lee & G.A. Whitmore, eds. *Lifetime Data: Models in Reliability and Survival Analysis* Kluwer, Boston, pp. 227–232.

[34] Prentice, R.L. & Gloeckler, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics* **34**, 57–67.

[35] Selmer, R. (1990). A comparison of Poisson regression models fitted to multiway summary tables and Cox's survival model using data from a blood pressure screening in the city of Bergen, Norway, *Statistics in Medicine* **9**, 1157–1165.

[36] Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with double censored data, *Journal of the American Statistical Association* **69**, 169–173.

[37] Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 290–295.

[38] Turnbull, B.W. & Weiss L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data, *Biometrics* **34**, 367–375.

(*See also* **Survival Analysis, Overview**; **Survival Distributions and Their Characteristics**)

MEI-JIE ZHANG

Group-randomization Designs

Randomized **clinical trials** compare two or more intervention or treatment strategies, using random allocation (assignment) (*see* **Randomization**) to intervention condition. Within this framework, group randomization designs (sometimes also called cluster randomization designs) randomly allocate intact groups (clusters), rather than individuals, to intervention condition. Units of group randomization include communities, small towns or villages, factories (workplaces), schools or classrooms, religious institutions, chapters of social organizations, families, and clinical practices.

Randomized trials are the recommended approach for obtaining valid comparisons of competing intervention strategies. The advantages of randomization are no less important for group-based trials than for individual-based trials [14, 15]. **Randomized treatment assignment** avoids **bias**, achieves balance (on average) of both known and unknown predictive factors between intervention and comparison groups, and provides the basis for statistical tests (see [2]).

A variety of examples can be cited to illustrate the range of group-randomized trials. A trial implemented in Indonesia randomly assigned 450 villages to participate or not to participate in vitamin A supplementation, in the form of capsules distributed to preschool children at baseline and again six months later [26]. A trial of the impact of a community-level intervention for improved treatment of sexually transmitted diseases on HIV infection in the Mwanza region of rural Tanzania randomized 12 communities (six pairs) between intervention and control [16, 17]. The Community Intervention Trial for Smoking Cessation (COMMIT) randomized 22 communities (11 pairs) in North America to test a community-based, multichannel, four-year intervention [3, 12, 15]. The Child and Adolescent Trial for Cardiovascular Health (CATCH) randomized schools to test a behaviorally oriented cardiovascular health education program [22, 27]. As a final example here, the Eating Patterns Study randomized 28 physician practices (within six primary care clinics) to evaluate the effectiveness of a self-help booklet, with physician endorsement thereof, in lowering dietary fat intake and raising dietary fiber intake in a primary-care practice setting [1]. For some other examples, see [8] and [25].

Randomization by group is less efficient statistically than randomization by individual, because individuals in a group-randomized trial will contribute less information than if individually randomized, as discussed by Cornfield [4]. There are, however, reasons why randomization by group may be chosen, some of which are illustrated by the trials referenced above. As outlined by Green et al. [15], these reasons include feasibility of delivery of the intervention, political and administrative considerations, the need to avoid contamination between individuals allocated to competing interventions, the very nature of the intervention (i.e. group-level), the existence of ready-made endpoints measured at the group level (*see* **Outcome Measures in Clinical Trials**), the desire to use site-specific resources to decrease cost, and greater generalizability.

When a group entity such as a community is investigated, two distinct types of outcome measures may be used. In one approach, we identify a cohort of individuals in each community at the beginning of the trial (at the "baseline", just before randomization) and then follow the cohort prospectively to measure changes in behavior or other outcome. Alternatively, we determine changes in the **prevalence** of some condition or behavior in each community, using independent **cross-sectional** samples at the baseline and at the end of the trial (perhaps with intermediate assessments also). Reasons for selecting one or the other approach have been given [9, 13, 15, 20]; some trials include both of these approaches.

As outlined by Green et al. [15], possible advantages of the cohort approach are: that it can target specific segments of the population (e.g. smokers in a smoking cessation trial) and measure the effect of the intervention on such individuals directly; it tends to have better statistical **power**; and it is not influenced by certain changes in the nature of the community (e.g. migration). Possible disadvantages are the problem of losses to follow-up, and the theoretical risk that repeated contact may affect either the actual success of the intervention or at least the self-reports of outcome (such as behavior change). Also, during a long trial, it can be argued that a cohort may become less representative of the community from which it was selected; at a minimum, the surviving members of the cohort will be aging while the age distribution in the community may remain

unchanged. However, this latter issue may not be of concern for trials focusing on the outcome for individuals, even though group randomization had been employed. Conversely, possible advantages of the cross-sectional approach are that it avoids the problem of losses to follow-up (although, of course, **nonresponse** to surveys remains an issue); it reduces concern about repeated assessment influencing the outcome; and it can measure overall changes in a community. Possible disadvantages are that it is less efficient statistically, and it can be influenced by in/out migration. For additional discussions of this topic, see [9], [13], and [20].

Sample size determination for group-randomized trials needs to consider the extra source of variation resulting from the inherent heterogeneity across groups; see, for example, [7], [12], [18], and [27]. Expressed another way, the outcomes of individuals within a group are generally **correlated**, as measured by the intraclass (intracluster) correlation. It is important to have an adequate number of randomized groups to account for the between-group variation. Often, in practice, it is easier to obtain a reasonably large number of individuals per group than it is to obtain a large number of groups, so the latter becomes the factor controlling (and perhaps limiting) the power of the trial. Because of this, to increase power (efficiency), we may consider **stratification** or pair-**matching** of the groups prior to randomization, and then analyzing the data accordingly. While matching could actually produce a loss in efficiency when the number of pairs is particularly small, due to loss of **degrees of freedom** in the analysis [23], gain in power can be achieved depending on how effectively the matching reduces community heterogeneity within pairs (see [11]).

As with the sample size calculations, the method of analysis must account for the correlation of individuals within a group (see [5–7]). One approach that has been used to account for cluster randomization is to adjust a standard (individual-level) analysis for the estimated **"design effect"** (i.e. the variance inflation due to clustering), which depends on the estimated intraclass correlation $\hat{\rho}$ (see, for example, [6]); for clusters of equal size n, the design effect is $[1 + (n - 1)\hat{\rho}]$. Alternatively, a variety of models can be applied to group-randomized data, to account for individual-level **covariates** as well as group variation; for example, by introducing a **random group effect** in an **analysis of variance** or **analysis of covariance**; see, for example, [24], [21], and [9]. Recently, Feng et al. [10] have compared various estimation procedures under a linear model (including **maximum likelihood** under a normal mixed model (*see* **Multilevel Models**), **generalized estimating equations**, and a **bootstrap approach**), and Klar [19] has compared model-dependent and **robust** tests using a generalized-estimating-equations extension of **logistic regression** for **binary** outcome data.

Rather than using a model-based method to account for group randomization, we may prefer a randomization-based approach. With randomization-based inference, the outcome data are analyzed many times (once for each acceptable assignment that could have been employed, according to the randomization process) and then compared with the observed result, without dependence on additional distributional or model-based assumptions. Thus, this approach is robust; **hypothesis testing** (**randomization tests** or permutation tests) and corresponding test-based **confidence intervals** can be designed for group-randomized data, based specifically on the randomization distribution [14, 15]. Such randomization tests involve permuting the assignment of groups (clusters) to intervention condition. Therefore, by definition, they account for the fact that groups (rather than individuals) were randomized. As discussed by Gail et al. [13], one can construct randomization tests that are specific for the design (e.g. unmatched or matched).

For randomization-based **inference**, although permutation is at the group level, individual-level covariates can be incorporated in the analysis (see COMMIT [3] as an example). We can use individual-level data for purposes such as imputation for missing values (*see* **Multiple Imputation Methods**), adjustment (of the intervention effect) for baseline characteristics of the individual participants (*see* **Covariate Imbalance, Adjustment for**), or performing separate analyses in subsets defined by individual-level covariates such as demographic variables [13–15]. In the latter situation, in addition to testing the main effect of intervention, we can construct randomization tests for intervention–covariate **interactions**, to test formally whether the intervention effect differs according to the value of a covariate (*see* **Treatment-covariate Interaction**).

References

[1] Beresford, S.A.A., Curry, S.J., Kristal, A.R., Lazovich, D., Feng, Z. & Wagner, E.H. (1997). A dietary intervention in primary care practice: the Eating Patterns Study, *American Journal of Public Health* **87**, 610–616.

[2] Byar, D.P., Simon, R.M., Friedewald, W.T., Schlesselman, J.J., DeMets, D.L., Ellenberg, J.H., Gail, M.H. & Ware, J.H. (1976). Randomized clinical trials: perspectives on some recent ideas, *New England Journal of Medicine* **295**, 74–80.

[3] COMMIT Research Group (1995). Community Intervention Trial for Smoking Cessation (COMMIT): I. Cohort results from a four-year community intervention, *American Journal of Public Health* **85**, 183–192.

[4] Cornfield, J. (1978). Randomization by group: a formal analysis, *American Journal of Epidemiology* **108**, 100–102.

[5] Donner, A. & Klar, N. (1993). Confidence interval construction for effect measures arising from cluster randomization trials, *Journal of Clinical Epidemiology* **46**, 123–131.

[6] Donner, A. & Klar, N. (1994). Methods for comparing event rates in intervention studies when the unit of allocation is a cluster, *American Journal of Epidemiology* **140**, 279–289.

[7] Donner, A., Birkett, N. & Buck, C. (1981). Randomization by cluster: sample size requirements and analysis, *American Journal of Epidemiology* **114**, 906–914.

[8] Donner, A., Brown, K.S. & Brasher, P. (1990). A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989, *International Journal of Epidemiology* **19**, 795–800.

[9] Feldman, H.A. & McKinlay, S.M. (1994). Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model, *Statistics in Medicine* **13**, 61–78.

[10] Feng, Z., McLerran, D. & Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with Gaussian error, *Statistics in Medicine* **15**, 1793–1806.

[11] Freedman, L.S., Gail, M.H., Green, S.B. & Corle, D.K., for the COMMIT Research Group (1997). The efficiency of the matched-pairs design of the Community Intervention Trial for Smoking Cessation (COMMIT), *Controlled Clinical Trials* **18**, 131–139.

[12] Gail, M.H., Byar, D.P., Pechacek, T.F. & Corle, D.K., for the COMMIT Study Group (1992). Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT), *Controlled Clinical Trials* **13**, 6–21 [Erratum: *Controlled Clinical Trials* **14** (1993) 253–254].

[13] Gail, M.H., Mark, S.D., Carroll, R.J., Green, S.B. & Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials, *Statistics in Medicine* **15**, 1069–1092.

[14] Green, S.B. (1997). The advantages of community-randomized trials for evaluating lifestyle modification, *Controlled Clinical Trials* **18**, 506–513.

[15] Green, S.B., Corle, D.K., Gail, M.H., Mark, S.D., Pee, D., Freedman, L.S., Graubard, B.I. & Lynn, W.R. (1995). Interplay between design and analysis for behavioral intervention trials with community as the unit of randomization, *American Journal of Epidemiology* **142**, 587–593.

[16] Grosskurth, H., Mosha, F., Todd, J., Mwijarubi, E., Klokke, A., Senkoro, K., Mayaud, P., Changalucha, J., Nicoll, A., ka-Gina, G., Newell, J., Mugeye, K., Mabey, D. & Hayes, R. (1995). Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomized controlled trial, *Lancet* **346**, 530–536.

[17] Hayes, R., Mosha, F., Nicoll, A., Grosskurth, H., Newell, J., Todd, J., Killewo, J., Rugemalila, J. & Mabey, D. (1995). A community trial of the impact of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania: 1. Design, *Journal of Acquired Immune Deficiency Syndromes* **9**, 919–926.

[18] Hsieh, F.Y. (1988). Sample size formulae for intervention studies with the cluster as unit of randomization, *Statistics in Medicine* **7**, 1195–1201.

[19] Klar, N. (1996). Stratified analysis of correlated binary outcome data: A comparison of model dependent and robust tests of significance, *Communications in Statistics – Theory and Methods* **25**, 2431–2458.

[20] Koepsell, T.D., Diehr, P.H., Cheadle, A. & Kristal, A. (1995). Invited commentary: Symposium on Community Intervention Trials, *American Journal of Epidemiology* **142**, 594–599.

[21] Koepsell, T.D., Martin, D.C., Diehr, P.H., Psaty, B.M., Wagner, E.H., Perrin, E.B. & Cheadle, A. (1991). Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach, *Journal of Clinical Epidemiology* **44**, 701–713.

[22] Luepker, R.V., Perry, C.L., McKinlay, S.M., Nader, P.R., Parcel, G.S., Stone, E.J., Webber, L.S., Elder, J.P., Feldman, H.A., Johnson, C.C., Kelder, S.H. & Wu, M., for the CATCH Collaborative Group (1996). Outcomes of a field trial to improve children's dietary patterns and physical activity, *Journal of the American Medical Association* **275**, 768–776.

[23] Martin, D.C., Diehr, P., Perrin, E.B. & Koepsell, T.D. (1993). The effect of matching on the power of randomized community intervention studies, *Statistics in Medicine* **12**, 329–338.

[24] Murray, D.M., Hannan, P.J., Jacobs, D.R., McGovern, P.J., Schmid, L., Baker, W.L. & Gray, C. (1994). Assessing intervention effects in the Minnesota Heart Health Program, *American Journal of Epidemiology* **139**, 91–103.

[25] Simpson, J.M., Klar, N. & Donner, A. (1995). Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993, *American Journal of Public Health* **85**, 1378–1383.

[26] Sommer, A., Tarwotjo, I., Djunaedi, E., West, K.P., Jr, Loeden, A.A., Tilden, R. & Mele, L. (1986). Impact of vitamin A supplementation on childhood mortality: a randomized controlled community trial, *Lancet* **1**, 1169–1173.

[27] Zucker, D.M., Lakatos, E., Webber, L.S., Murray, D.M., McKinlay, S.M., Feldman, H.A., Kelder, S.H. & Nader, P.R., for the CATCH Study Group (1995). Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH): implications of cluster randomization, *Controlled Clinical Trials* **16**, 96–118.

SYLVAN B. GREEN

Growth and Development

Growth and development is the process through which the fetus changes in size, shape, composition, and function to become a reproductively mature adult. Growth is generally defined as the change (or rate of change) in one or more continuous measures of size, shape or composition, usually assessed by **anthropometry**. Indeed, the term "growth" is often used as a shorthand for child anthropometry. Development reflects maturation as measured by ordinal markers of function; some markers are inherently discontinuous, while others have an underlying continuum. Menarche, the onset of menstruation, is a marker of development, while height gain is a measure of growth.

General Introduction

Growth is a process that all children undergo, and normal growth is a proxy for good health. Equally, abnormal growth is a nonspecific marker for poor health, as many childhood diseases affect growth to a greater or lesser extent. It is primarily growth that distinguishes pediatrics from adult clinical medicine.

The most obvious manifestation of growth is an increase in physical size, which, unlike many other markers of clinical status, is easily quantified. Statisticians have always been attracted by the qualities of anthropometry data: they are accessible, cheap, and reproducible; in addition, measurements are often highly correlated from one age to another.

Broadly speaking, the rate of change of anthropometry with age is fairly constant. In detail, this does not hold, of course – growth in individuals progresses in fits and starts, and some markers of development, e.g. menarche, are, by their nature, sudden in onset. However, the population assessment of growth is assumed to smooth out these discontinuities, and a requirement of smoothness is imposed commonly on summary curves relating anthropometry to age.

There are broadly two aims to the study of growth and development – **screening** individuals and screening populations. Clinically, the biostatistical challenge is to characterize normal growth and development in a way that optimizes its value as a screening tool. In public health terms, the aim is to assess the health status (*see* **Quality of Life and Health Status**) of *groups* of children, defined on the basis of, say, geography, **ethnic** make-up, or culture.

The way that individual growth is characterized depends on the available information. This may be from one or many measurement occasions, and may involve anthropometry by itself or other **covariates** as well. Ideally, as much relevant information as possible should be used in the clinical assessment.

The reference centile (*see* **Quantiles**) chart, otherwise known as a growth reference or growth standard, is the fundamental tool for assessing growth, and it is used widely throughout the world for this purpose. It summarizes the distribution of anthropometry in a known reference population at different ages through childhood by plotting selected centiles of the distribution against age.

Weight and height are the measurements most commonly expressed as growth references, although other anthropometry has also been summarized in chart form. Figure 1 shows a height centile chart for British boys in 1990, made up of nine centile curves ranging from the 0.4th to the 99.6th centile. The curves are spaced two-thirds of a **standard deviation** score (SDS) apart over the range ± 2.67 SDS, the other centiles being (approximately) the 2nd, 9th, 25th, 50th, 75th, 91st, and 98th.

Also plotted on the chart are heights for a boy measured regularly from birth to 21 years. Until the age of 12 his height stays close to the **median**, but during the adolescent growth spurt it rises to the 75th centile and then falls back to the 25th. This is a common pattern of growth, with a fairly constant centile until puberty, then a rise or fall depending on whether the growth spurt occurs relatively early or late. Thus, the chart can identify unusual growth in two ways – single measurements on a relatively large or small centile, and serial measurements that cross centiles too rapidly (up or down) over time.

The paradox of the growth chart is that even though it contains no information about velocity, being based on single measurements, it is used almost universally to assess growth. The assumption is that subjects should grow parallel to the centile curves, yet this is wrong (see below). In addition, the chart cannot flag poor growth – the degree of centile crossing corresponding to an abnormal pattern of growth is not specified.

Addressing this problem is just one of several statistical issues that arise in the assessment of growth,

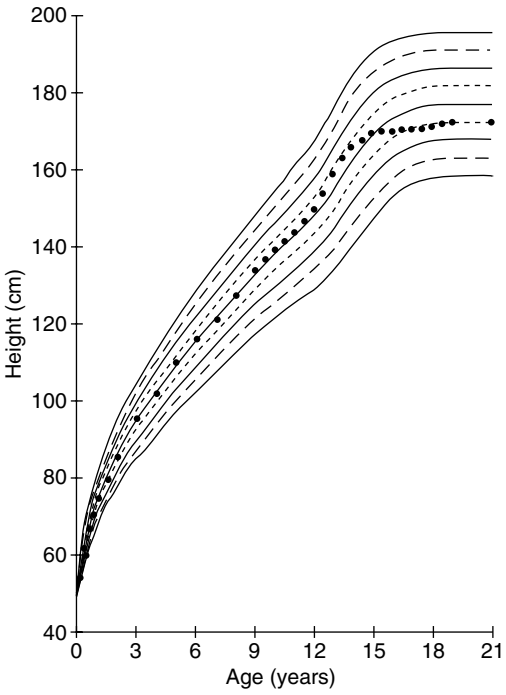

Figure 1 Height reference for British boys in 1990. The nine centiles are two-thirds of an SDS apart, at 0.4, 2.3, 9, 25, 50, 75, 91, 97.7 and 99.6. Superimposed are heights for a French boy measured from birth to 21 years

many of them directly related to the construction and use of growth charts.

Historical Development

In 1885, **Francis Galton** described centiles for summarizing the distribution of body measurements, and Bowditch later applied them to the heights of US schoolchildren. He divided the children into age groups, calculated a set of height centiles in each group, joined up the corresponding centiles across groups, and the centile chart was born.

Since then, centile charts have become very popular as a tool for monitoring growth. The statistical basis for the charts has, until recently, been simplistic – the data grouped by age, centiles for each group obtained by sorting and counting, and the centiles joined across groups with hand-smoothed or polynomially smoothed curves (*see* **Polynomial Regression**). For **normally distributed** measurements, the centiles can be estimated more efficiently using the mean and SD, each smoothed across age. For non-normal data such as weight, the centiles are often obtained assuming a **lognormal distribution**.

Two other types of chart have been developed to help assess growth over time: the velocity chart plots the rate of change in anthropometry against age, and the conditional chart plots the measurement against another covariate besides age. Where the conditional chart is based on two successive measurements, one adjusted for the other, this is a generalized form of velocity chart that has the advantage of adjusting for **regression to the mean**. The simple cross-sectional chart, by analogy to the velocity chart, is commonly called a *distance* chart.

During the 1960s and 1970s Tanner et al. [15] elevated growth chart production to something of an art form, producing hand-smoothed charts for a wide range of body measurements in British children. Since then, growth charts have continued to be developed, locally, nationally and internationally, and this interest has led over the past 10 years to a sharp increase in the statistical literature on the construction of reference centile curves and age-specific reference ranges. There has also been work on new forms of chart that are better suited to the assessment of growth over time.

Description of the Different Types of Study

Anthropometry data are collected either cross-sectionally or longitudinally. In a **cross-sectional** survey, the subjects each provide a single measurement, and the sample is chosen to be representative of some larger population. **Longitudinal** studies involve measuring the same subjects repeatedly, often at regular time intervals. The advantage of longitudinal studies is that they provide information on the distribution of growth velocity, which is not available in cross-sectional surveys.

A compromise design is the semi-longitudinal study, which combines the best features of single- and multiple-measurement designs. For the initial survey, a cross-sectional sample is drawn, then, for later surveys, a fraction of the original sample is redrawn and the balance is sampled afresh. As a simple example, efficient estimates of both height distance and annual height velocity can be obtained throughout childhood from just two surveys, one year apart.

The choice of reference sample is important. Ideally, it should be representative of the target population, obtained by random and, if necessary, by **stratified sampling**. However, there are two issues that complicate sampling for growth studies, widely debated under the titles "representative vs. healthy" and "national vs. local".

Since the chart's purpose is to identify abnormal growth, some people argue that the reference sample should be restricted to normally growing children (i.e. healthy rather than representative). The problem with this is that abnormal growth is hard to define, and it can lead to arbitrary exclusion rules. Using a representative sample avoids the arbitrariness and permits **random sampling**. The term "growth standard" implies a benchmark of healthy growth, whereas the "growth reference" is representative of the population and is neutral about its health status.

There is also concern about the appropriateness of using a nationally or internationally representative growth reference to assess individuals in areas of the world where local growth patterns are different. For instance, is it sensible to use an international reference based on North American children to assess growth in the developing world, as the **World Health Organization** (WHO) has, until recently, recommended?

The answer is that it depends. The growth chart is used in two ways – as a clinical tool and for public health purposes. Clinically, the chart should represent growth in the local **target population**, but in public health terms, international growth statistics need to be comparable – this requires the use of a common growth chart. So the question of "national vs. local" charts comes down to a question of how the chart is used.

Landmark Studies

During this century, much has been written on the statistics of growth, and it is unrealistic to mention all the relevant studies. On the construction of growth charts, Tanner et al. [15] introduced the concept of a velocity chart, and discussed the fundamentals of centile fitting in some detail. Healy [9] described the conditional **regression** chart, subsequently used by Cameron [3] and developed by Cole [5]. The fitting of centile curves became a recognized research topic in 1988, when Healy [10] and Cole [4] described separately fundamentally different methods for dealing with **skew** data. Goldstein [8] illustrated the value of **multilevel models** for analyzing data from semi-longitudinal growth studies, while Wade et al. [16] proposed a method for deriving centiles of ordinal developmental data, for example pubertal staging.

There have also been efforts over the past 60 years to model the shape of the individual human growth curve in infancy and childhood. Jenss & Bayley [11] first proposed the linear plus **exponential** model for infant growth, and other proposed models have been of the fractional polynomial family [14]. Preece & Baines [13] and Jolicoeur et al. [12] described ingenious five- and seven-parameter exponential models describing height growth during childhood.

Statistical Concepts and Techniques

Growth References

Metric for Calculation. The growth reference statistically adjusts anthropometry for age and sex, expressing it as either a centile, a fraction of the median, or a standard deviation score (SDS or Z score) (*see* **Normal Scores**). For statistical analysis, the centile is inappropriate, due to its nonlinear scale. The fraction of the median is simple to calculate and removes the age trend, but it does not adjust for age-related changes in the coefficient of variation (CV). The SDS adjusts for age trends in the mean and CV, and, in addition, it can, if the measurement is **normally distributed**, be converted to a centile. The relationship between measurement, fraction of the median and SDS is as follows:

$$\text{SDS} = \frac{(\text{measurement/median}) - 1}{\text{CV}},$$

which shows that if the CV is constant, then the fraction of the median and the SDS convey the same information.

In practice, measurements with a small CV, like height or head circumference, tend to be close to normally distributed, while other more variable measurements, like weight or skinfold thickness, are appreciably skew to the right. This nonnormality of much anthropometry has posed a problem for the construction of growth references.

Choice of Centiles. The centiles that appear on the growth chart need (i) to be symmetric about the

median, (ii) to be roughly evenly spaced across the distribution, and (iii) to provide reasonable cut-offs for screening purposes in the tails of the distribution. A common choice is the set: 3rd, 10th, 25th, 50th, 75th, 90th and 97th, approximately two-thirds of an SDS apart. Some charts prefer the 5th and 95th to the 3rd and 97th, and the WHO international reference uses whole SDSs from -3 to $+3$, reflecting the wide variation in the populations being assessed.

The scheme used in Figure 1 formalizes the two-thirds SDS spacing, and adds an extra centile in each tail to give cut-offs that screen in 4 per 1000 rather than 3% of the population. This lowers the **false positive rate**, which is desirable at a time when the costs associated with secondary referrals for growth disorder are rising steeply.

Distance Reference. The task of constructing a growth reference involves summarizing the distribution of the measurement as it changes with age. With normally distributed data, this is achieved by modeling the mean and SD (or CV) with suitable parametric or **nonparametric** functions, then calculating the required centiles. For example, if $M(t)$ and $S(t)$ are the mean and CV as functions of age t, and $C_{100\alpha}(t)$ is the 100_αth measurement centile, then

$$C_{100\alpha}(t) = M(t)[1 + S(t) \times z_\alpha],$$

where z_α is the normal equivalent deviate (NED) for tail area α. Parametric functions include fractional polynomials and nonlinear exponential functions, while nonparametric functions are the family of smoothers that includes lowess (*see* **Graphical Displays**), cubic **splines** and kernel smoothers (*see* **Density Estimation**).

Where the data are not normally distributed, they can often be **transformed**; e.g. to logarithms, to restore normality. However, this assumes that the same transformation is suitable at all ages. Where this is not the case, and the degree of nonnormality is age-related, two quite different approaches have been proposed. One is an extension of the "smoothed mean and SD" approach, where an extra parameter summarizing the third moment of the distribution is smoothed across age to give a third curve. This allows for the presence of skewness in the distribution, either as a constant or changing with age. This third parameter may be the power λ in the Box–Cox **power transformation** $g(t) = f(t)^\lambda$, or the shift c in the shifted **lognormal** $g(t) = \log[f(t) - c]$, or the shape parameter in the **gamma distribution**. The first of these alternatives, entitled the LMS method [4], has been used to construct several national growth references. The curves defining the Box–Cox power λ, the median μ, and the CV σ as functions of age t are denoted by $L(t)$, $M(t)$, and $S(t)$, respectively, and the curve for measurement centile 100α is given by

$$C_{100\alpha}(t) = M(t)[1 + L(t)S(t)z_\alpha]^{1/L(t)}.$$

The converse of this equation converts a measurement to a standard deviation score (SDS) Z:

$$Z = \frac{[\text{measurement}/M(t)]^{L(t)} - 1}{L(t)S(t)}.$$

The LMS curves are obtained either by grouping the data, estimating λ, μ and σ by group and then smoothing across groups, or, alternatively, by estimating the curves directly from ungrouped data. Cole et al. [6] have fitted reference curves for height and weight in British children using **penalized maximum likelihood**, leading to LMS curves that are cubic smoothing splines.

Estimating the third **moment** of the distribution as a function of age requires a very large sample ($n > 10^4$, ideally). There have been attempts to estimate the fourth moment in the same way, but it requires much larger samples than are currently available. The benefit is also likely to be small.

An entirely different approach, known as the Healy–Rasbash–Yang (HRY) method [10], estimates each of a set of centile curves using a form of scatterplot smoother, and models them with low-order polynomials in age. The coefficients of the polynomials are each constrained to follow a polynomial in z, the NED of the corresponding centile curves. This leads to a set of centile curves that are of similar shape, and where the spacings between centiles at each age are linked in a way analogous to a cumulative frequency distribution. The order of the polynomial in z determines the form of distribution, according to the shape of the Q–Q normality plot (*see* **Graphical Displays**). A straight line corresponds to a normal distribution, while a quadratic is similar in shape to a skew distribution and a cubic curve allows for **kurtosis**.

Velocity Reference. Distance references identify individuals whose measurement centile is extreme,

e.g. below the 4 per 1000 centile, but they give no information about *changes* in centile from one age to another. The velocity reference is designed to monitor centile change, and it provides centiles of velocity (i.e. growth per unit time) by age in the reference population. Thus, it is the direct analogy of the distance chart, and the same methods can be used to estimate distance and velocity centiles.

Velocity is calculated as $V = \Delta H/\Delta t$, where ΔH is the change in measurement H over time interval Δt. Thus, the population variance of V depends, *inter alia*, on twice the measurement error variance δ^2 of H, and Δt. For this reason, velocity charts have to be constructed for a prespecified time interval, commonly one year for height, and the centile spacings are particularly sensitive to δ, the size of the measurement error.

A relatively low velocity appears as centile crossing on the distance chart. If the distance chart is plotted on the SDS scale, so that the centile curves are horizontal straight lines, the traces of subjects growing abnormally depart from the horizontal. The advantage of working on the SDS scale is that the variability of centile change in the reference population is particularly simple to calculate: the change in SDS between two ages has a standard deviation given by

$$\mathrm{SD}(Z_2 - Z_1) = [2(1 - r)]^{1/2},$$

where Z_1 and Z_2 are the SDSs at each age and r is the correlation between them, based on the reference population.

Regression to the mean, which affects the interpretation of change over time, is not taken into account using the velocity reference. Individuals or groups below a given measurement centile can expect to become less extreme, i.e. to regress towards the mean, when measured again. This affects the velocity reference directly – subjects initially on a low centile have a greater expected velocity than those of the same age starting on a higher centile. For this reason, velocity references are intrinsically flawed, and to quantify regression to the mean, a regression-based conditional reference is needed instead.

Conditional Reference. A conditional reference adjusts (conditions) for another variable as well as age and sex (e.g. parental height or sibling weight). Healy [9] first suggested the use of a regression-based conditional reference to adjust velocity for regression to the mean by regressing the current measurement on the previous measurement:

$$H_2 = \alpha + \beta H_1 + \varepsilon.$$

Centiles for H_2 conditional on H_1 are then given by:

$$C_{100\alpha}(H_2|H_1) = \alpha + \beta H_1 + \sigma \times z_\alpha,$$

where σ is the residual standard deviation (RSD) from the regression. This regression-based reference of H_2 on H_1 then correctly predicts H_2 given H_1. Berkey et al. [2] extended the regression-based approach to two or more conditioning variables.

As with the velocity reference, the use of an SDS scale greatly simplifies the conditional reference, and the equation above becomes:

$$C_{100\alpha}(Z_2|Z_1) = rZ_1 + (1 - r^2)^{1/2} Z_\alpha.$$

This allows Z_2 to be calculated for a range of Z_1 values, which, when plotted on the SDS distance chart, define a line whose slope is the required 100αth conditional velocity centile. Several such lines can be superimposed on the SDS distance chart to provide for the simultaneous assessment of distance and velocity. The values of Z_1 and Z_2 can also be converted back and plotted on the original measurement scale. Figure 2 illustrates a combined distance and 5th centile conditional velocity chart for weight in British boys during the first year of life.

Importance of Measurement Error. The principle of growth monitoring is that individuals who cross centiles excessively are more likely to have an underlying growth disorder. The **sensitivity** and **specificity** of a **screening** test based on centile crossing, be it velocity $(Z_2 - Z_1)$ or conditional velocity $(Z_2 - rZ_1)$, depend critically on the **standard error** of the centile change, which in turn depends on the size of the **correlation** r between the measurements, which is mediated by the underlying measurement error. Unlike the distance chart, the velocity chart is highly sensitive to even a modest increase in the **measurement error**, which can increase the false positive rate dramatically. Effective growth monitoring demands high-quality measurements.

Developmental Markers. The age of occurrence of binary markers such as menarche is usually estimated

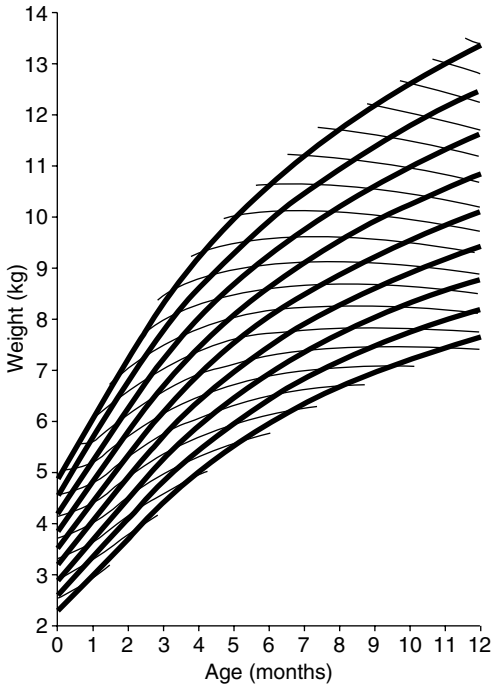

Figure 2 Conditional weight reference for British boys in 1990. The heavy lines are nine distance centiles two-thirds of an SDS apart, as in Figure 1. The slopes of the fainter lines represent the 5th centile of conditional velocity, as measured over a four-week period. The slope of an individual infant's weight curve needs to be compared with the slope of the nearest velocity line, to detect weight gain below the 5th centile

using probit or logit analysis (*see* **Quantal Response Models**), based on information from individual subjects about whether or not the event has occurred. This is a less biased estimate than the mean of the recalled age of the event, and adjusts for **censoring** in younger subjects.

Wade et al. [16] have described a method for estimating age-related centiles for ordinal data, which generalizes this approach.

Growth Curves

Growth curve analysis is a general term for the study of correlated measurements over time in individuals and groups, a topic that has been much studied over the last 30 years. It has tended to be of statistical rather than pediatric interest, and has been largely superseded in recent years by more powerful methods such as **generalized estimating equations** and **multilevel models**.

Growth Models and Adult Prediction

The shape of the human stature growth curve has always fascinated auxologists. Early attempts to model it parametrically focused on the early part of life, where the velocity decreases monotonically. Second-order fractional polynomials and the linear-exponential model both fit well over this age range [1]. However, the greater challenge has been to model height throughout childhood, including the pubertal growth spurt where the velocity rises to a peak then falls to zero. Preece & Baines [13] proposed a five-parameter exponential model, which fitted puberty well but failed to model infant growth. Subsequently, Jolicoeur et al. [12] came up with a seven-parameter model that has been found to fit well from birth to adulthood.

Gasser et al. [7] have developed nonparametric kernel-based methods to model the distance, velocity and acceleration curves of stature throughout childhood. By combining individual curves on suitably transformed age scales, they have demonstrated the existence of subtle features of the growth curve, most notably a transient peak in height velocity around age 7 (the mid-growth spurt).

Clinicians like to be able to predict adult height for children being treated for a growth disorder, using information on the height of parents or the child's stage of maturation. Statistically, the best way to do this is by **empirical Bayes**, combining the child's own measurements with information on the population child–adult growth curve. This is useful for monitoring the effect of clinical interventions (e.g. treatment with growth hormone).

Body-size Scaling

Weight and height are highly correlated with each other, and both are proxies for body size. Weight adjusted for height removes the size component, and leaves an index of shape. Shape in this context is viewed as a measure of body fatness, and the adjustment is carried out on the double logarithmic scale, leading to a power index of the form weight/height^n. When $n = 2$, the index is known as the Body Mass Index (BMI), and it is widely used as a measure of adiposity in adults and children of both sexes. The

choice of $n = 2$ gives an index that is only weakly correlated with height (ideally it should be uncorrelated), and that is also well correlated with direct measures of body fat. Like other variables measured during childhood, its distribution changes with age and needs to be adjusted accordingly.

Ratio indices like the BMI can be abused, usually when the form of the index is assumed (wrongly) to guarantee a lack of correlation between it and height. If this is of primary concern, then (log)weight should be adjusted for (log)height directly, using covariance analysis (*see* **Analysis of Covariance**). However, for clinical and epidemiologic use, the BMI is a useful tool for classifying overweight and obesity.

References

[1] Berkey, C.S. (1982). Comparison of two longitudinal growth models for preschool children, *Biometrics* **38**, 221–234.

[2] Berkey, C.S., Reed, R.B. & Valadian, I. (1983). Longitudinal growth standards for preschool children, *Annals of Human Biology* **10**, 57–67.

[3] Cameron, N. (1980). Conditional standards for growth in height of British children from 5.0 to 15.99 years of age, *Annals of Human Biology* **7**, 331–337.

[4] Cole, T.J. (1988). Fitting smoothed centile curves to reference data, *Journal of the Royal Statistical Society, Series A* **151**, 385–418.

[5] Cole, T.J. (1994). Growth charts for both cross-sectional and longitudinal data, *Statistics in Medicine* **13**, 2477–2492.

[6] Cole, T.J., Freeman, J.V. & Preece, M.A. (1998). British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood, *Statistics in Medicine* **17**, 407–429.

[7] Gasser, T., Kneip, A., Binding, A., Prader, A. & Molinari, L. (1991). The dynamics of linear growth in distance, velocity and acceleration, *Annals of Human Biology* **18**, 187–205.

[8] Goldstein, H. (1986). Efficient statistical modelling of longitudinal data, *Annals of Human Biology* **13**, 129–141.

[9] Healy, M.J.R. (1974). Notes on the statistics of growth standards, *Annals of Human Biology* **1**, 41–46.

[10] Healy, M.J.R., Rasbash, J. & Yang, M. (1988). Distribution-free estimation of age-related centiles, *Annals of Human Biology* **15**, 17–22.

[11] Jenss, R.M. & Bayley, N. (1937). A mathematical method for studying growth in children, *Human Biology* **9**, 556–563.

[12] Jolicoeur, P., Pontier, J., Pernin, M.O. & Sempé, M. (1988). A lifetime asymptotic growth curve for human height, *Biometrics* **44**, 995–1003.

[13] Preece, M.A. & Baines, M.J. (1978). A new family of mathematical models describing the human growth curve, *Annals of Human Biology* **5**, 1–24.

[14] Royston, P. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[15] Tanner, J.M., Whitehouse, R.H. & Takaishi, M. (1966). Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965 Parts I and II, *Archives of Disease in Childhood* **41**, 454–471, 613–635.

[16] Wade, A.M., Ades, A.E., Salt, A.T., Jayatunga, R. & Sonksen, P.M. (1995). Age-related standards for ordinal data: modelling the changes in visual acuity from 2 to 9 years of age, *Statistics in Medicine* **14**, 257–266.

T.J. COLE

Guidelines On Statistical Methods in Clinical Trials

In the early 1960s, the thalidomide disaster led to a tightening of the national regulations that dictate the evidence required for authorization of a medicinal product. Forty years on, these regulations still provide the basis upon which all regulatory authorities assess medicinal products. Although the statutory basis for regulations is different in all countries (the EU is governed by EC directives and the US is governed by Federal Government Statutes), the underlying principles for assessment of medicinal products are common, i.e. the evaluation of the quality, safety and efficacy of the medicinal product given the proposed product labeling (*see* **Drug Approval and Regulation**).

As with all statutes, some interpretation of the law is required. The EC has issued its Notice to Applicants [7]. The Food and Drug Administration (FDA) regularly publishes the Code of Federal Regulations and the Japan Pharmaceutical Reference is regularly issued in Japan [12]. These elucidate some of the high level issues arising in drug authorization but do not address specific clinical trial issues. These can only be addressed in guidelines or guidance documents. Guidelines are issued by a variety of organizations, including regulatory authorities, health organizations and expert groups, and as a result of differences in health care, populations and regional traditions in clinical research around the world, the detailed technical requirements stated in these guidelines often differ between countries and expert groups.

Many guidelines focus on the appropriate design and analysis of clinical trials in different therapeutic fields and mention statistical issues, but only in a relatively limited way. In the last 15 years there has been a clearer understanding that statistical excellence in the design and analysis of clinical trials is a necessary requirement to obtain evidence that is sufficient for a marketing authorization. This realization led to the publication of national regulatory guidelines which addressed statistical methodology in clinical trials and provided guidance on reporting [1, 8, 20].

In the 1990s there was increasing pressure on pharmaceutical companies to obtain more rapid global product approvals. The most efficient way to achieve this is to perform one drug development program that is suitable for all regulatory authorities. This is difficult to achieve when guidelines for drug development differ around the world. The need for the standardization of clinical trial methodology worldwide has led to the establishment of the International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use. In 1998, the ICH recommended for adoption the guideline on Statistical Principles for Clinical Trials [11], which is a consensus document covering the views of the US, Japan and the EU. This important document sets worldwide statistical standards for clinical trials.

In this article the background to the ICH process is explained. An overview of ICH E9 is presented and other important sources of guidance to statisticians working in clinical trials are discussed. The summary of ICH E9 presented here is inevitably limited in scope and perspective and cannot do justice to the carefully crafted wording of the complete document. The reader for whom the content of ICH E9 is important is strongly advised to refer to the complete guideline.

The ICH

The ICH was initiated in April 1990 to discuss scientific and technical aspects of product registration [3]. It is a joint initiative involving the regulatory authorities and pharmaceutical industry representatives from each of the three regions. The EU is represented by the European Agency for the Evaluation of Medicinal Products (EMEA) and the European Federation of Pharmaceutical Industries Associations (EFPIA). Japan is represented by the Ministry of Health and Welfare (MHW) and the Japan Pharmaceutical Manufacturers Association (JPMA) and the US is represented by the FDA and the Pharmaceutical Research and Manufacturers of America (PhRMA). There are also observers from other bodies and regions.

The purpose of the ICH is to make recommendations on the ways to achieve greater harmonization in the interpretation and application of technical guidelines and requirements for product registration in the development of new medicines. The objective of such harmonization is a more economical use of human, animal and material resources, and the elimination of unnecessary delay in the global development and availability of new medicines

whilst maintaining safeguards on quality, safety and efficacy, and regulatory obligations to protect public health. The ICH considers topics related to (preclinical) safety, quality and efficacy (including clinical safety) (*see* **Preclinical Treatment Evaluation**).

The basic principles of the ICH are to

- develop scientific consensus through discussions between regulatory and industry experts,
- provide wide consultation of the draft consensus documents through normal regulatory channels,
- produce a harmonized text, and
- gain commitment from regulatory authorities to implement the ICH harmonized texts.

Each new topic is tackled by an Expert Working Group (EWG) that includes members from the six cosponsors. The EWG then follows an extensive (five-step) process of consensus building and consultation during which comments are widely sought from all interested parties. When a final text is agreed by the EWG and accepted by the ICH Steering Committee, it is recommended for adoption by the authorities in each of the three regions.

Table 1 ICH E9 Expert Working Group

Europe – regulatory authority
Joachim Röhmel, BfARM, Berlin, Germany
John Lewis, MCA, London, UK

Europe – industry
Bernhard Huitfeldt, Astra Arcus AB, Södertälje, Sweden
Trevor Lewis, Pfizer Central Research, Kent, UK

Japan – regulatory authority
Isao Yoshimura, Science University of Tokyo, Tokyo, Japan
Tosiya Sato, Institute of Statistical Mathematics, Tokyo, Japan

Japan – industry
Tohru Uwoi, Yamanouchi Pharmaceutical Co. Ltd, Tokyo, Japan
Hiroyuki Uesaka, Eli Lilly, Kobe, Japan

US – regulatory authority
Robert O'Neill, FDA, Maryland, USA
Susan Ellenberg, FDA, Maryland, USA

US – industry
Bill Louv, GlaxoWellcome, North Carolina, USA
Stephen Ruberg, Hoechst Marion Roussel Inc, Missouri, USA

ICH E9: Statistical Principles for Clinical Trials

In November 1995, the ICH Steering Committee decided that an ICH guideline on statistical methodology should be developed (ICH E9). The CPMP *Note for Guidance on Statistical Methodology in Clinical Trials* [1] had been completed in the previous year. It had been created as a result of considerable collaborative effort among many European statisticians from regulatory agencies, industry and academia [14, 17]. It was the most up-to-date regional guideline on statistics and so formed the basis for the new ICH guideline. However, the ICH guideline was also heavily influenced by the earlier US and Japanese statistical guidelines [8, 20].

The EWG for ICH E9 consisted of 12 statisticians: two regulatory and two industry representatives from each of the three ICH regions. Table 1 gives a list of members of the Working Group. They decided that the guideline should concentrate on principles rather than detailed procedures and that it should attempt to address a broader audience than statisticians alone. The draft document (ICH Step 2) was discussed in the worldwide statistical community so that a large number of statisticians and other scientific experts from all three regions had an influence on the final content. In February 1998 the ICH Steering Committee recommended the guideline for adoption to the regulatory bodies of the EU, Japan and the US. CPMP adopted the guideline at the end of March 1998 (and it came into operation in Europe in September 1998). In September 1998 the guideline came into operation in the US and was published by the FDA in the Federal Register. The MHW adopted the guideline in November 1998. Many therapeutic guidelines now directly reference ICH E9 and this helps to ensure that common statistical standards apply in clinical trials. It also increases the awareness of all physicians leading drug development programs that appropriate statistical input is essential for regulatory approval. Therefore, the ICH E9 guideline is essential reading for all statisticians working in the environment of clinical trials.

The full guideline is published in the statistical literature [11] with an introduction by Lewis [15] or can be obtained from the World Wide Web in English and Japanese (see Table 2).

The scope and content of the guideline are now addressed by presenting key points from each section of the guideline, using the ICH E9 section headings.

Table 2 Useful websites

ICH website: www.ich.org
ICH E9 guideline: www.ich.org/MediaServer.jser?@_ID=485&@_MODE=GLB
CPMP (EU) efficacy guidelines: www.emea.eu.int/index/indexh1.htm
FDA (US) guidelines: www.fda.gov/cder/regulatory/default.htm
MHW (Japanese) guidelines: www.nihs.go.jp/drug/

Introduction

The guideline stresses the critical role of statistical expertise in clinical research within the whole drug development program. The focus is on statistical principles in confirmatory trials. It does not prescribe specific methods or procedures.

The underlying principles of the guideline deal with minimizing bias and maximizing precision. Potential sources of bias need to be identified as completely as possible so that attempts to limit such bias may be made, otherwise the ability to draw valid conclusions from the clinical trial may be seriously compromised.

Sources of bias may arise in the design, conduct or analysis of a clinical trial. Therefore, for each clinical trial it is assumed that important details of the design, conduct and proposed statistical analysis will be specified in a trial protocol. The extent to which the procedures in the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial. Since bias can occur in subtle or unknown ways and its effect is not measurable directly, it is important to evaluate the robustness of the results.

Considerations for Overall Clinical Development

Trial Context. The broad aim of the process of clinical development of a new drug is to find out whether there is a dose range and schedule at which the drug can be shown to be simultaneously safe and effective, to the extent that the risk–benefit relationship is acceptable (*see* **Benefit/Risk Assessment in Prevention Trials**). The particular subjects who may benefit from the drug, and the specific indications for its use, also need to be defined.

Confirmatory trials are required to provide robust evidence in support of all key claims related to efficacy or safety. Confirmatory trials should only address a limited number of questions with a predefined primary objective, which leads to the primary hypothesis and is the basis upon which the trial is designed and analyzed. Exploratory trials may have less precise objectives, a more flexible design and involve data exploration with data-dependent choice of hypotheses. Exploratory trials cannot form the basis of formal proof of efficacy but may contribute to the total body of evidence.

Scope of Trials. Clinical trial populations range from a narrow subgroup of patients in early trials, through to a wider representation of the target population in confirmatory trials.

Clinical trials generally contain one primary variable, which should be capable of providing the most clinically relevant and convincing evidence related to the primary objective of the trial (*see* **Outcome Measures in Clinical Trials**). The use of a reliable and validated variable with which experience has been gained in earlier studies or published literature is recommended. Secondary variables, their relative importance and roles in interpretation of results, should be defined in the protocol. Issues are discussed relating to specific forms of primary and secondary variables, namely composite variables, global assessment variables, multiple primary variables, surrogate variables and categorical variables.

Design Techniques to Avoid Bias. The optimal design of studies in a marketing application is the double-blind (*see* **Blinding or Masking**), randomized controlled trial. However, it is recognized that when such trials are not feasible, single-blind or open studies may be necessary.

Blinding should be maintained throughout the conduct of the trial and only when the data are cleaned to an acceptable level should unblinding occur. Some methods to overcome difficulties in blinding are discussed (double dummy treatment, separate assessors for patient care and outcome assessment, centralized randomization).

Randomization introduces a deliberate element of chance into the treatment assignment and provides a sound basis for quantitative evaluation of the

evidence relating to the treatment effects. It tends to produce treatment groups in which the distributions of prognostic factors, known and unknown, are similar. In combination with blinding, randomization helps to avoid possible bias in the selection and allocation of subjects arising from the predictability of treatment assignments. Detailed considerations about the production of a randomization schedule are discussed, with specific guidance for crossover trials, multicenter trials, stratified randomization and dynamic allocation.

Bias can also be reduced at the design stage by specifying procedures in the **protocol** aimed at minimizing any anticipated irregularities in trial conduct that might impair a satisfactory analysis, including various types of protocol violations, withdrawals and missing values. The protocol should consider ways to reduce the frequency of such problems and to handle the problems that do occur in the analysis of data.

Trial Design Considerations

Design Configuration. The most common clinical trial design for confirmatory trials is the parallel group design in which subjects are randomized to one of two or more treatment groups. **Crossover designs** in which each subject is randomized to a sequence of two or more treatments are discussed and the importance of avoiding carryover is stressed. Issues related to **factorial designs** are also presented.

Multicenter Trials. Multicenter trials are used to facilitate the accrual of subjects within a reasonable time frame and to provide a better basis for the subsequent generalization of findings. A common protocol is needed for all centers, the manner in which the protocol is implemented should be clear, and procedures should be standardized as completely as possible. It may be advantageous to avoid excessive variation in the numbers of subjects per center if it is later found necessary to take into account the heterogeneity of the treatment effect from center to center. Rules for combining centers in the analysis should be justified and specified in the protocol or at least at the time of the blind review.

The statistical model used in the primary analysis would not normally be expected to include a term for treatment-by-center interaction (*see* **Treatment-covariate Interaction**). In some trials with very few subjects per center there may be no reason to expect centers to have an effect of clinical importance. In other situations the limited numbers of subjects per center make it impracticable to include the center effects. Consequently, it is not appropriate to include a term for center in these models and it is not necessary to stratify randomization by center.

Type of Comparison. Efficacy is most convincingly established by demonstrating superiority to placebo, or an active control, or by demonstrating a dose–response relationship. Most of the guidance given in this document relates to superiority trials.

Trials that use an active control can also be used to test the objective of equivalence (*see* **Equivalence Trials**) or noninferiority (that the efficacy of an investigational product is no worse than that of the active control). Such trials, which do not also include a placebo control or multiple doses of the new drug, have no measure of internal validity and thus make external validation necessary. These trials are not conservative in nature, so that many flaws in the design or conduct of the trial will tend to bias the results towards a conclusion of equivalence. Consequently, the design and conduct of such trials should receive special attention. This discussion has been augmented in ICH E10 (on Choice of Control Groups), which provides greater insight into the problems of demonstrating the internal and external validity of noninferiority trials.

Group Sequential Designs. Group sequential designs are most commonly applied to facilitate the conduct of interim analyses (*see* **Data and Safety Monitoring**). The statistical methods employed should be specified in advance of the availability of information on subject treatment assignments.

Sample Size. The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed (*see* **Sample Size Determination**). This number is usually determined by the primary objective of the trial. If this is not the case, then it should be made clear and justified. The method by which the sample size is calculated and the estimates of quantities used in the calculations should be stated in the protocol. The basis of the estimates should be described and the sensitivity of the sample size estimate to deviations from these assumptions should be investigated.

Data Capture and Processing. The form and content of the information collected should focus on the data necessary to implement the planned analysis and be in full accordance with the protocol and with Good Clinical Practice (ICH E6).

Trial Conduct Considerations

Trial Monitoring and Interim Analysis. Careful conduct of a clinical trial according to the protocol has a major impact on the credibility of the results (*see* **Clinical Trials Audit and Quality Control**). Careful monitoring that oversees the quality of the trial can ensure that difficulties are noticed early and their occurrence or recurrence minimized. This type of monitoring does not require access to information on comparative treatment effects, nor unblinding of the data and therefore has no impact on the Type I error.

Interim analysis is another form of "monitoring". It involves breaking the blind to make treatment comparisons and should therefore be planned in the protocol with considerations of the potential biases that may be incurred.

Changes in Inclusion and Exclusion Criteria. If changes to the inclusion and exclusion criteria (*see* **Eligibility and Exclusion Criteria**) of a trial are necessary, then they should be made without breaking the blind and should be described in a protocol amendment, which should cover any statistical consequences.

Accrual Rates. If the rate of accrual falls appreciably below the projected level, then the reasons should be identified and remedial actions taken.

Sample Size Adjustment. If the sample size is revised during the course of a trial, then the steps taken to preserve blindness and consequences, if any, for the Type I error and confidence intervals should be explained.

Interim Analysis and Early Stopping. An interim analysis is any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to the formal completion of a trial (*see* **Data and Safety Monitoring**). Since the number, methods and consequences of these comparisons affect the interpretation of the trial, all interim analyses should be carefully planned in advance and described in the protocol. Special circumstances may dictate the need for an interim analysis that was not defined at the start of a trial, in which case a protocol amendment describing the interim analysis should be completed prior to the unblinding of the data for the interim analysis.

The protocol (or an amendment before the first interim analysis) should describe the schedule of interim analyses, or at least the considerations that will govern its generation. The stopping guidelines and their properties should be clearly described in the protocol.

The execution of an interim analysis should be a completely confidential process and all investigator and sponsor staff involved in the conduct of the trial should remain blind to the results of such analyses (except for those directly involved in the execution of the interim analysis). When a sponsor performs an interim analysis, particular care should be taken to protect the integrity of the trial and limit the dissemination of results.

Any interim analysis that is not planned appropriately (with or without the consequences of stopping the trial early) may flaw the results and weaken confidence in the results. Therefore, such analyses should be avoided. The reason for an unplanned interim analysis should be explained in the study report and an assessment of the potential magnitude of bias and impact on interpretation of results discussed.

Role of an Independent Data Monitoring Committee. An Independent Data Monitoring Committee (*see* **Data Monitoring Committees**) may be established to assess the progress, safety data and critical efficacy variables of a trial, and to make recommendations about whether to continue, modify or terminate the trial. It should have written operating procedures and maintain records of all its meetings, including interim results (which should be available for regulatory review).

Data Analysis Considerations

Prespecification of the Analysis. The statistical section of the **protocol** should include all the principal features of the proposed confirmatory analysis of the primary variable(s) and the way in which anticipated analysis problems will be handled.

A separate statistical analysis plan may be written after finalizing the protocol. It gives a more technical and detailed elaboration of the analyses of primary and secondary variables, but only results from analyses envisaged in the protocol can be regarded as confirmatory. The statistical analysis plan should be reviewed and possibly updated as a result of the blind review of the data (see Evaluation and Reporting section below). It should be finalized before breaking the blind. If the blind review suggests changes to the principal features in the protocol, then these should be documented in a protocol amendment. The timing of the finalization of the statistical analysis plan and the breaking of the blind should be formally recorded.

Analysis Sets. The sets of subjects whose data are to be included in the main analyses should be defined in the statistical section of the protocol.

The intention-to-treat principle [8, 16] (*see* **Intention to Treat Analysis**) implies that all randomized subjects should be included in the primary analysis (for superiority trials). Compliance with this principle would necessitate complete follow-up of all randomized subjects for study outcomes. Since this may be difficult to achieve in a clinical trial, the term "full analysis set" is introduced to describe the analysis set which is as complete as possible and as close as possible to the intention-to-treat ideal of including all randomized subjects. A few circumstances might lead to the exclusion of randomized subjects from the full analysis set (e.g. major eligibility violation, failure to take any medication or no data post-randomization). Concerns related to all these exclusion criteria are discussed in the guideline and it is noted that no analysis is complete unless the potential biases (arising from these exclusions or any other reasons) are addressed.

The "per protocol" set of subjects is sometimes described as the "valid cases" or "efficacy" or "evaluable subjects" sample. It refers to a subset of subjects in the full analysis set who are compliant with the protocol. The precise reasons for excluding subjects from the per protocol set should be fully documented before unblinding.

It is advantageous to demonstrate a lack of sensitivity of the principal trial results to alternative choices of analysis set. In superiority trials, the full analysis set is generally used in the primary analysis, but in an equivalence or noninferiority trial use of the full analysis set is not generally conservative and its role should be considered carefully.

Missing Values and Outliers. Missing values represent a potential source of bias in a clinical trial. A trial may be regarded as valid, none the less, provided the methods for dealing with missing values are sensible. If the number of missing values is substantial, then an investigation should be made into the sensitivity of the results to the method of handling missing values. The influence of outliers can be explored in a similar manner.

Data Transformation. The decision to transform key variables prior to analysis is best made during the design of the trial on the basis of similar data from earlier clinical trials and should be specified in the protocol. The decision on whether and how to transform a variable should be influenced by the preference for a scale that facilitates clinical interpretation.

Estimation, Confidence Intervals and Hypothesis Testing. The statistical section of the protocol should specify the hypotheses to be tested and the treatment effects to be estimated. Estimates should be accompanied by confidence intervals and their method of calculation specified. The underlying statistical model (including all factors and covariates to be fitted) should be fully specified. The primary analysis of the primary variable should be clearly distinguished from supporting analyses of the primary or secondary variables.

It is important to clarify whether one- or two-sided tests will be used and to justify prospectively the use of one-sided tests. The approach of setting Type I errors for one-sided tests at half the conventional Type I error used in two-sided tests is preferable in regulatory settings.

Adjustment of Significance and Confidence Levels. In confirmatory trials, any important aspects of **multiplicity** should be identified in the protocol and adjustments to the Type I error should be implemented and explained. Alternatively, an explanation of why adjustment is not thought necessary should be provided.

Subgroups, Interactions and Covariates. Pretrial deliberations should identify the covariates and factors expected to have an important influence on the primary variable and should consider how to account for these in the analysis in order to improve precision

and to compensate for any lack of balance between treatment groups (*see* **Covariate Imbalance, Adjustment for**). Special attention should be paid to the role of baseline measurements of the primary variable. It is not advisable to adjust the main analyses for covariates measured after randomization, because they may be affected by the treatments.

The treatment effect may vary with a subgroup or covariate (*see* **Treatment-covariate Interaction**). In some cases, such interactions are of particular interest and a subgroup analysis or model including interactions is part of the planned confirmatory analysis. In most cases, however, analyses of subgroups or interactions are exploratory and should be interpreted cautiously.

Integrity of Data and Computer Software Validity. The credibility of the results depends on the quality and validity of the methods and software used for data management and statistical analysis (*see* **Data Management and Coordination; Clinical Trials Audit and Quality Control**).

Evaluation of Safety and Tolerability

Scope of Evaluation. In early phase clinical trials, the evaluation of safety is mostly exploratory. In later phases, the safety and tolerability profile of the drug can be characterized more fully. Specific comparative safety claims should be supported by relevant evidence from confirmatory trials designed to evaluate safety.

Choice of Variables and Data Collection. The safety data collected will depend on various characteristics of the drug, the type of subjects to be studied and the duration of the trial. A consistent methodology for data collection and evaluation is recommended throughout the clinical trial program.

Set of Subjects to be Evaluated and Presentation of Data. The set of subjects to be summarized for safety is usually those who received at least one dose of the investigational drug. All adverse events should be reported, whether or not they are considered to be related to treatment, but the summarization of "treatment emergent" events is also helpful to reduce the noise caused by background signs and symptoms of the disease. If treatment is long-term and a substantial proportion of treatment withdrawals or deaths are expected, then time to event analyses should be considered.

Statistical Evaluation. The previous section noted that there are a number of methods for calculating the incidence of an adverse event and that the method used should be defined in the protocol. For laboratory data, it is recommended that both the treatment means and the numbers outside certain thresholds should be evaluated.

The calculation of P values is sometimes useful as an aid to evaluating a specific difference of interest, but the general lack of sensitivity of such safety comparisons means that small but clinically important differences (*see* **Clinical Significance Versus Statistical Significance**) may be overlooked (Type II error). In addition, when P values are used as a flagging device, the multiplicity of tests makes the P values difficult to interpret in a conventional manner.

Integrated Summary. Safety information is commonly summarized across trials, but the usefulness of this summary is dependent upon the trials being adequate and well-controlled with high quality data (See ICH M4: Common Technical Document).

Reporting

Evaluation and Reporting. The reporting of statistical work is covered in the ICH E3 guideline and is therefore covered relatively briefly in ICH E9. It is however noted that statistical judgment should bear on the analysis, interpretation and presentation of results, so the statistician should be a member of the team responsible for the clinical study report.

The use of the blind review is described and it is stated that decisions made at the time of the blind review should be described in the report and should be distinguished from those decisions made after the statistician was unblinded. Many of the detailed aspects of presentation and tabulation should be finalized at the time of the blind review. It is noted that statisticians or other staff involved in the unblinded interim analysis should not participate in the blind review or in making modifications to the statistical analysis plan. Attention should be paid to any differences between the planned analysis as described in the protocol, amendments and statistical

analysis plan based on the blind review and the actual analysis.

All subjects who entered the trial should be accounted for in the report. The effect of losses of data on the main analyses should be considered carefully. Descriptive statistics should illustrate the important features of the primary and secondary variables and of key prognostic and demographic variables. The results of significance tests should be reported with precise p-values.

Summarizing the Clinical Database. An overall summary and synthesis of the evidence on safety and efficacy from all the clinical trials is required for a marketing application and may be accompanied, when appropriate, by a statistical combination of results (*see* **Benefit/Risk Assessment in Prevention Trials**). It is always valuable to present the main results of a series of similar trials in an identical form to permit comparison, usually in tables or graphs that focus on estimates and confidence intervals. To facilitate these analyses, common methods for the evaluation of primary and secondary variables and methods for handling protocol deviators are worthwhile, and essential for **meta-analysis**. Any statistical procedures used to combine data across trials should be described in detail. Attention should be paid to the possibility of bias arising from selection of trials, homogeneity of results and the proper modeling of the various sources of variation. The sensitivity of conclusions to the assumptions and selections made should be explored.

In summarizing safety data, it is important to examine the safety database thoroughly for any indications of potential toxicity, and to follow up any indications by looking for an associated supportive pattern of observations. The risks associated with identified adverse effects should be appropriately quantified to allow a proper assessment of the risk–benefit relationship.

Other ICH guidelines

By December 2003, the text of 48 ICH guidelines had been finalized and most of them had been adopted in all three regions. Fifteen of the finalized guidelines are denoted as "Efficacy"topics and include diverse topics from clinical safety to good clinical practice that would be of interest to many statisticians. In December 2003, further three ICH guidelines were issued in draft from for consultation. This demonstrates that the ICH process has slowed considerably and its major work is probably complete. However, it remains possible to propose and develop new topics and the need to revise topics must always be remembered.

Table 3

ICH topic number	Guideline title
E3	Structure and Content of Clinical Study Reports
E4	Dose Response Information to Support Drug Registration
E5	Ethnic Factors in the Acceptability of Foreign Clinical Data
E6	Good Clinical Practice
E7	Studies in Support of Special Populations: Geriatrics
E8	General Considerations for Clinical Trials
E9	Statistical Principles for Clinical Trials
E10	Choice of Control Group in Clinical Trials
E11	Clinical Investigation of Medicinal Products in the Pediatric Population
E12A	Principles for Clinical Evaluation of New Antihypertensive Drugs
M4	Common Technical Document *(format for summary documents)*

Table 3 lists those ICH guidelines that have most impact on the statistician working on clinical trials. Day and Talbot [4] provide a brief overview of ICH E3, E9 and E10. The full text of all ICH guidelines can be obtained from the ICH website (Table 2).

Other guidelines

The CPMP, FDA and MHW produce many new guidelines each year covering issues in all areas of drug development (from pharmaceutical work-up to postmarketing surveillance), and across a broad spectrum of therapeutic indications (from the use of hormone replacement therapy to treatments for cardiac failure). Current details of the US, European and Japanese guidelines can be easily accessed on the World Wide Web (Table 2).

The CPMP develops two types of guideline. A Note for Guidance is produced when there is substantial experience in a particular field (hypertension, epilepsy). A Points to Consider

document is prepared when there is limited experience, for example, acute respiratory distress syndrome. All CPMP (and FDA) guidelines are released for public consultation before they are finalized.

In 1999, the CPMP announced that it was to prepare several Points to Consider guidance documents about biostatistical/methodological issues arising from discussions on licensing applications [19]. By December 2003, five of these had been issued, on the following topics:

- Switching between superiority and noninferiority;
- Applications with: 1. meta-analyses; 2. one pivotal study;
- Missing data;
- Multiplicity issues in clinical trials;
- Adjustment for baseline covariates.

A draft Points to Consider document on "Choice of non-inferiority margin"was issued for consultation in early 2004, and "Use of statistical methods for flexible design and analysis of confirmatory clinical trials"is expected to be released for consultation later in the year. A concept paper outlining the development of guidance on "Data monitoring committees"is also expected in 2004.

The most difficult debates arise when the guidance given by the FDA and CPMP is markedly different. The most fervent statistical debate in recent years (*Journal of Biopharmaceutical Statistics*, Part 1, Volume 7, 1997) has focused on the FDA guidance for statistical approaches to establishing bioequivalence, which includes methodology for establishing population and individual bioequivalence [10]. This compares with the publication of the revised CPMP Note for Guidance on bioavailability/bioequivalence [2], which promotes "average" bioequivalence. Another FDA guideline which has a major statistical content is that on population pharmacokinetics [10] and for this topic no CPMP guidance is available at present.

All therapeutic guidelines contain important information about trial design relevant to statisticians working in that field. In many cases these therapeutic guidelines also introduce interesting statistical topics such as twofold testing strategies to achieve a primary objective (e.g. CPMP Note for Guidance on Bipolar Disorder) and designs to prove lack of disease progression (e.g. CPMP Note for Guidance on Parkinson's Disease).

This article has focused on guidelines that are issued by the regulatory authorities in the US and EU. There are many other helpful guidelines that are issued from a number of sources such as other regulatory agencies (Canada, Australia, Nordic Regions), global and local health organizations (World Health Organization, Medical Research Council) and statistical organizations (**Statisticians in the Pharmaceutical Industry**).

Discussion

The ICH E9 guideline provides a consensus view on statistical principles in clinical trials. The guideline is a useful reference to promote the importance of statistical input to all aspects of a clinical trial (design, conduct, analysis and reporting). It also provides the basis for further discussion of unresolved or contentious statistical issues and specific topics of debate are now being addressed in the literature. For example, Edwards [5] explores the circumstances under which the Type I error is influenced by the failure to prespecify the statistical model and Lin [19] discusses the weighting of centers in a multicenter trial.

Consensus on Principles Must be Accompanied by Statistical Understanding

Statisticians working within the pharmaceutical environment are used to working within a highly regulated framework. Consequently, the creation of guidelines related to statistical matters is generally welcomed and considered as a tool to facilitate consensus amongst all parties involved in the worldwide development and authorization of medicinal products. However, statisticians working in other fields of application often consider the guidelines and need for careful prespecification of analysis plans in medical statistics as a restrictive block on their statistical prowess and expertise.

Lewis [13] clearly understood this concern,

> ... Finally, let me quieten the fears that I might be advocating *standardisation* of methodology. I will fight as determinedly as the next man to prevent our statistical work from becoming in any way an automatic unthinking process. There are, of course, strong pressures in this direction in the interest of efficiency and economy. I do ask for greater efforts to reach consensus agreement on general approaches.

> I am convinced we can move a long way down this road without stifling individual freedom.

It is important to remember that regulations are statutory and enforceable by law. Guidelines, however, are based on experience to date and seek to provide a consensus across different disciplines or regions. The use of a guideline as a "cookbook" with no scientific thought will rarely result in a recipe for success. Guidelines can never address all the issues that arise in the complicated world of clinical trial experimentation, but they can provide helpful advice and a starting point for scientific input. To this must be added the expertise of those involved in the experiment, in order to obtain results that are robust and meaningful.

Professionalism of Clinical Trials Statisticians

In section 5.4.1 of ICH E6 (on Good Clinical Practice) it is stated that

> The sponsor should utilise qualified individuals, (e.g. biostatisticians, clinical pharmacologists, and physicians) as appropriate, throughout all stages of the trial process, from designing the protocols and CRFs and planning the analyses to analysing and preparing interim and final clinical trial reports.

In section 1.2 of ICH E9 it is stated that:

> ... it is assumed that the actual responsibility for all statistical work associated with clinical trials will lie with an appropriately qualified and experienced statistician, as indicated in ICH E6. The role and responsibility of the trial statistician, in collaboration with other clinical trial professionals, are to ensure that statistical principles are applied appropriately in clinical trials supporting drug development. Thus, the statistician should have a combination of education/training and experience sufficient to implement the principles articulated in this guidance.

The final sentence of this quotation has led to great debate in the statistical community questioning the level of education/training and experience that are sufficient to qualify a person as a professional clinical trials statistician [6]. This is a particularly difficult matter in Japan where very few statisticians are academically qualified [18].

Implementation and Updating of Guidelines

The ICH initiative is committed to "ensure that there is a process for updating and supplementing the current ICH guidelines, when necessary and monitoring their use, so that the benefits of harmonisation achieved so far are not lost" (ICH website).

A guideline can only be properly judged when it has been put into practice for some time and reviewed with experience from its practical application. In January 2004, it is unclear whether ICH has a specific plan to review E9, but this will be necessary if the document is to continue to have maximum value.

References

[1] CPMP Working Party on Efficacy of Medicinal Products (1995). Biostatistical methodology in clinical trials in applications for marketing authorisations for medicinal products, *Statistics in Medicine* **14**, 1659–1682.

[2] CPMP (2001). *Note for Guidance on Clinical Investigation of Bioavailability and Bioequivalence*. CPMP/EWP/QWP/1401/98.

[3] D'Arcy, P.F. & Harron, D.W.G., eds (1992). *Proceedings of the First International Conference on Harmonisation*. Queen's University of Belfast.

[4] Day, S. & Talbot, D.J. (2000). Editorial. Statistical guidelines for clinical trials, *Journal of the Royal Statistical Society, Series A* **163**, 1–3.

[5] Edwards, D. (1999). On model prespecification in confirmatory randomised studies, *Statistics in Medicine* **18**, 771–785.

[6] EFSPI Working Group (1999). Qualified statisticians in the European pharmaceutical industry: report of a European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) Working Group, *Drug Information Journal* **33**, 407–415.

[7] European Commission (1996). *The Rules Governing Medicinal Products in the European Union*: Volume I: *The Rules Governing Medicinal Products for Human Use in the European Union*. Volume II: *Notice to Applicants for Marketing Authorization for Medicinal Products for Human Use in the European Union*. Volume III: *Guidelines on the Quality, Safety and Efficacy of Medicinal Products for Human Use*. European Commission, Brussels.

[8] FDA (1988). *FDA Guidelines for the Format and Content of the Clinical and Statistical Sections of a New Drug Application*. FDA, Rockville.

[9] FDA (1999). *Draft Guidance for Industry: Average, Population, and Individual Approaches to Establishing Bioequivalence*. FDA, Rockville.

[10] FDA (2001). *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence.* FDA, Rockville, MD, USA.

[11] ICH E9 Expert Working Group (1999). Statistical Principles for Clinical Trials: ICH Harmonised Tripartite Guideline, *Statistics in Medicine* **18**, 1905–1942.

[12] Japan Medical Products International Trade Association (1999). *Japan Pharmaceutical Reference – Pharmaceutical Administration and Regulations in Japan*, Version 5. Japan Medical Products International Trade Association, Tokyo.

[13] Lewis, J.A. (1983). Clinical trials: statistical development of practical benefit to the pharmaceutical industry, *Journal of the Royal Statistical Society, Series A* **146**, 362–393.

[14] Lewis, J.A. (1996). Editorial. Statistics and statisticians in the regulation of medicines, *Journal of the Royal Statistical Society, Series A* **159**, 359–362.

[15] Lewis, J.A. (1999). Statistical principles for clinical trials: an introductory note on an international guideline, *Statistics in Medicine* **18**, 1903–1904.

[16] Lewis, J.A. & Machin, D. (1993). Editorial. Intention to treat – who should use ITT?, *British Journal of Cancer* **68**, 647–650.

[17] Lewis, J.A., Jones, D.R. & Roehmel, J. (1995). Biostatistical methodology in clinical trials – a European guideline, *Statistics in Medicine* **14**, 1655–1682.

[18] Lewis, J.A., Louv, W., Rockhold, F. & Sato, T. (2000). The impact of the guideline entitled Statistical Principles for Clinical Trials (ICH E9), *Statistics in Medicine* **20**.

[19] Lewis, J., Louv, W., Rockhold, F. & Sato, T. (2001). The impact of the guideline entitled Statistical Principles for Clinical Trials (ICH E9), *Statistics in Medicine* **20**, 2549–2560.

[20] MHW (1992). *Guideline for the Statistical Analysis of Clinical Trials*. Ministry of Health and Welfare, Pharmaceutical Affairs Bureau, Tokyo.

KAREN M. FACEY & JOHN A. LEWIS

Guttman Scale

A Guttman scale is based on a set of items which address a one-dimensional latent variable or attribute. The response options for each of the items is of the form agree/disagree or Yes/No (*see* **Binary Data**). For example, consider an investigation in which we would like to measure a subject's physical ability. Suppose the following four items are developed: "Can you walk one city block without assistance?"; "Can you walk six city blocks without assistance?"; "Can you walk 1 mile without assistance?"; "Can you walk 2 miles without assistance?" In a Guttman scale the items are ordered hierarchically such that an affirmative response to one item implies affirmative responses to each of the items preceding it. The scale score is simply the sum of affirmative responses over the set of items. The score represents the subject's level of the attribute under investigation (in this example, the subject's physical ability).

In general terms, suppose we have k distinct items addressing the same attribute which are ordered hierarchically, i.e. ordered such that an affirmative response to one item theoretically implies affirmative responses to each of the items preceding it, and measured on each of n subjects. The data can be organized into a matrix $\mathbf{A}$ with n rows containing each subject's responses to the set of k items, which are called the subject's response profiles. For example, subject i has a response profile $\mathbf{a}_i' = (a_{i1}, a_{i2}, \ldots, a_{ik})$ with $a_{i1} \geq a_{i2} \geq \ldots \geq a_{ik}$, and $a_{ij} = 0$ or 1 representing No and Yes responses, respectively (see Figure 1). If subjects (rows of $\mathbf{A}$) are compared and sorted according to their response profiles, i.e. their scores on the set of items, a Guttman scale is said to exist if every pair of response profiles is comparable. For example, consider subjects i and j with response profiles $\mathbf{a}_i' = (a_{i1}, a_{i2}, \ldots, a_{ik})$ and $\mathbf{a}_j' = (a_{j1}, a_{j2}, \ldots, a_{jk})$, respectively. Subjects i and j are comparable if $a_{i1} \geq a_{j1}, a_{i2} \geq a_{j2}, \ldots$, and $a_{ik} \geq a_{jk}$ [1]. In this example, subject i has a higher level of the attribute under investigation than subject j since subject i scores as high or higher than subject j on every item. When every pair of response profiles is comparable, a Guttman scale is formed. The relationships among response profiles allow for subjects to be rank ordered according to their level of the attribute under investigation. Figure 1 displays the triangular relationship among response profiles in a Guttman scale (see, for example, [2]), where each item is scored as Y = Yes or N = No.

	Item					
Subject	1	2	3	.	.	k
1	Y	Y	Y	Y	Y	Y
2	Y	Y	Y	Y	Y	N
3	Y	Y	Y	Y	N	N
.	Y	Y	Y	N	N	N
.	Y	Y	N	N	N	N
n	N	N	N	N	N	N

Figure 1 Yes and No responses

Although the Guttman scale is theoretically appealing, it is highly structured and not often observed in practice [3]. In particular, Guttman scales do not work well for psychological attributes, since such attributes are generally less concrete. Guttman scaling is also known as scalogram analysis and **correspondence analysis**.

References

[1] DeVellis, R.F. (1991). *Scale Development: Theory and Applications*. Sage, Beverly Hills, pp. 62–63.

[2] Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Ed. McGraw-Hill, New York, pp. 72–75.

[3] Shye, S., ed. (1978). On the search for laws in the behavioral sciences, in *Theory Construction and Data Analysis in the Behavioral Sciences*, S. Shye, ed. Jossey-Bass, San Francisco, Chapter 1.

(*See also* **Psychometrics, Overview**; **Simplex Models**)

LISA M. SULLIVAN

Guy, William Augustus

Born: 1810, in Chichester, UK.
Died: September 10, 1885, in London, UK.

Guy studied medicine with **P. C. A. Louis**, qualified in Cambridge in 1837, and was appointed Professor of Forensic Medicine at King's College, London in 1838. He wrote an important appreciation of Louis's work in 1839, tinged, however, with doubt as to the applicability of statistical findings to the treatment of individual patients, and (as with Louis) a degree of scepticism about mathematical theory. He became a prolific writer on public health matters, and was a prominent member of the Statistical Society of London (later the **Royal Statistical Society**, which commemorates him with the Guy Medal).

PETER ARMITAGE

Haenszel, William M.

Born: Rochester, New York, June 19, 1910.
Died: Wheaton, Illinois, March 13, 1998.

Although William M. Haenszel is probably best known as the name to the right of the hyphen on the **Mantel–Haenszel** test and Mantel–Haenszel **odds ratio**, he has made other equally important contributions throughout a very long and productive career as a biostatistician and epidemiologist. Of particular importance is his leadership in establishing and maintaining the population-based **cancer registry** known as *SEER* (for *S*urveillance *E*pidemiology and *E*nd *R*esults); his comparative studies of cancer occurrence in foreign-born and native-born Japanese Americans; and his use and advocacy of biomarkers and pathological findings in studies of cancer etiology, especially those involving gastric and large bowel cancer. His work from the 1950s to 1970s at the National Cancer Institute (*see* **National Institutes of Health (NIH)**) was at the cutting edge of chronic disease **epidemiology** and has greatly enriched that field both substantively and methodologically.

His early life was spent in Buffalo, New York, where he went through the public elementary and high school systems, and through the University of Buffalo (now, State University of New York at Buffalo), receiving a B.A. *summa cum laude* in sociology and mathematics in 1931. A year later, he received his M.A. in statistics, also from the University of Buffalo. Shortly thereafter, he worked as a statistician at the New York State Department of Health and later, as Director of the Bureau of Vital Statistics at the Connecticut State Department of Health. His experience of nearly 20 years in health statistics and record keeping at the State level, gave him a unique expertise and perspective into registration and data collection at the "grass roots" level. This expertise became invaluable in his later work at the National Cancer Institute and at the University of Illinois at Chicago (UIC) where he collaborated extensively with State and Local Officials in his studies of cancer epidemiology. It was especially useful in the contributions that he made to establish a population-based cancer registry system that could be used for investigating **incidence** and etiology of cancers as well as survival in persons having these diseases.

In 1952, he accepted a position as Head of the Biometry Section of the newly created National Cancer Institute (NCI) and, in 1961, became Chief of the larger Biometry Branch of NCI. He stayed in this position until his retirement from NCI in 1976. During his tenure at the NCI, there was a cadre of statisticians who themselves made major contributions in biostatistics and epidemiology. Among them were **Nathan Mantel**, **Marvin Schneiderman**, **Sid Cutler**, Marvin Zelen, Ed Gehan, **David Byar**, John Gart, and John Bailar. During this period, and largely owing to Bill Haenszel's mentoring and management skills, the Biometry Branch of NCI became an internationally recognized center of excellence in the development of statistical methodology applied to **observational** and experimental health studies. He fostered a culture of productivity and achievement among statisticians at NCI and the Biometry Branch (presently under the direction of Mitchell Gail) has maintained this excellence throughout the quarter century since his retirement from NCI in 1976.

It was during his tenure at NCI, that he coauthored with Nathan Mantel his paper on what became most widely known as the Mantel–Haenszel test [4]. For many years, this was one of the most widely quoted articles in the entire scientific literature.

Although this test is now sometimes called the Cochran–Mantel–Haenszel test (or CMH test) in recognition of the test developed five years earlier by **William Cochran** [1], the two tests differ statistically, primarily in that the Mantel–Haenszel test is predicated on a **hypergeometric distribution**, whereas the Cochran test is based on a **binomial distribution**. The Mantel–Haenszel test and its associated odds ratio became, soon after its publication, a backbone of stratified analysis (*see* **Stratification**) and proved to have applications and extensions to scenarios far beyond the subject matter of the original article (e.g. tables larger than 2×2 (*see* **Contingency Table**), **survival analysis**, comparisons among adjusted rates, etc.).

There has been much discussion and speculation concerning the comparative contributions that each of the two authors made in the development of the test. Both Mantel and Haenszel, when queried individually, have always praised the contribution of the other. In a 1984 symposium at the University of Illinois at Chicago marking the 25th anniversary of the original publication, Haenszel indicated that his own major contribution was in the formulation of "how the fourfold table (*see* **Two-by-Two Table**) was entered and constructed". By this, he probably meant that, in the **case–control** scenario on which it was originally based, the appropriate inference issue is predicated upon the null distribution of exposed cases, given that the marginals for total cases, total controls, and total exposed persons are fixed. This leads to the hypergeometric formulation that is different (although generally indistinguishable numerically) from Cochran's earlier hypothesis test.

The Mantel–Haenszel test and its subsequent extensions represented major advances in statistical analysis. Its accuracy, however, as with all statistical tools, is entirely dependent on the quality of the underlying data, and it could lack generalizability if the cancer cases being studied were not representative of the universe of cancer cases with respect to the exposure being studied (*see* **Validity and Generalizability in Epidemiologic Studies**). Recognizing this, Haenszel and others in leadership roles at the NCI began efforts for establishing a registry of cancer cases that would be representative of all cancer cases in the United States. On the basis of their efforts, the population-based registry, SEER, was launched in 1973. It now contains over 30 years of data on cancer incidence and survival and remains as one of the flagship programs at the NCI.

Also, while at NCI, Haenszel launched his landmark studies comparing the high gastric cancer rates in Japan to the much lower rates in those Japanese who had migrated to Hawaii [3], and identified diet as a possible risk factor for gastric cancer (*see* **Migrant Studies**). He repeated in other ethnic groups, this model of comparing cancer rates that prevail in the parent country to those among persons of the same ethnic group who migrated to the United States as well as to those among persons of the same ethnicity who were born in the United States. This model has served as a tool for generating etiological hypotheses that can be verified in subsequent case–control or **cohort studies**.

He also began, while at NCI, his studies focusing on gastric cancer in Cali, Colombia in collaboration with Pelayo Correa, a Colombian pathologist and epidemiologist at Louisiana State University [2]. This program was innovative in that it focused on population surveys that included gastroscopic measurements, and that its major objectives were to identify **risk factors** for the known precursors of gastric cancer. By focusing on the precursors rather than the much rarer cancers themselves, they could conduct **cross-sectional**, case–control, and cohort studies that had adequate statistical **power**, whereas similar studies that focused on gastric cancer rather than the precursor would have much less statistical power. This collaboration continued for many years after Haenszel's retirement from NCI and provided a wealth of information about the epidemiology of gastric cancer, especially dietetic risk factors.

After his retirement from NCI in 1976, Haenszel relocated to the Chicago area where he took a joint position at the Illinois Cancer Council and the University of Illinois School of Public Health. Both of these Institutions were in their very early years, and Haenszel's initial plan was to help these Institutions for two years and then totally retire at age 68. The intended two years stretched out to nearly 20 and he had a truly rich career as a senior mentor and advisor to both Chicago Institutions, and as a senior consultant to many agencies engaged in cancer research and cancer control. Among these were Louisiana State University (where his collaboration with Correa and his group continued), NCI, the American Cancer Society, the Illinois Department of Public Health, and the **International Agency for Research**

in Cancer (IARC). During this period, he was a major force in the establishment in the 1980s, by the Illinois Department of Public Health of a population-based cancer registry, which has since matured into a high quality registry and a resource for investigations of cancer etiology. At the University of Illinois at Chicago School of Public Health, he was instrumental in enriching the curriculum in chronic disease epidemiology, developing and teaching a series of courses in quantitative epidemiology and in cancer epidemiology. In addition, he devoted much time in mentoring junior faculty and students, many of whom are now senior epidemiologists and biostatisticians. The Haenszel Memorial Award was established in the early 1990s and is given in his honor annually to a student in the Division of Epidemiology and Biostatistics at UIC who has performed outstanding research during his/her career as a student.

Haenszel's health began declining in the mid-1990s, and he retired for a final time in 1996, two years before his death. His legacy lives on in the continuing work of his collaborators and disciples and in the vitality of the Institutions on which he left his mark.

References

[1] Cochran, W.G. (1954). Some methods for strengthening the χ^2 test, *Biometrics* **10**, 417–457.

[2] Correa, P., Haenszel, W., Cuello, C., Zavala, D., Fontham, E., Zarama, G., Tannenbaum, S., Collazos, T. & Ruiz, B. (1990). Gastric precancerous process in a high risk population: cohort follow-up, *Cancer Research* **50**, 4737–4740.

[3] Haenszel, W., Kurihara, M., Segi, M. & Lee, R.K.C. (1972). Stomach cancer among Japanese in Hawaii, *Journal of the National Cancer Institute* **49**, 969–988.

[4] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

PAUL S. LEVY

Half-normal Distribution

The half-normal distribution will arise in sampling from a standard normal population when the signs of the negative observations are lost or not relevant. The half-normal distribution was introduced by Daniel [2] in connection with the **analysis of variance** of **factorial experiments**. An example of the use of the half-normal in biostatistics is given by Berlin et al. [1], where the outcome of interest was the difference in treatment effects between two treatments in a number of **clinical trials**. The treatment effects (and hence their differences) were assumed normally distributed, but there was no reason to assign either treatment as the first of the pair, so the sign of the difference was made to be positive.

Formally, if z is normally distributed with mean equal to zero and variance equal to one, then the half-normal distribution is the distribution of $\sigma|z|$. The probability density function (pdf) is given by

$$f(x) = \frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\left\{\exp\left[\frac{-(x/\sigma)^2}{2}\right]\right\}, \quad x \geq 0. \quad (1)$$

The first four central **moments** of the half-normal are given by Elandt [3]:

$$m_1 = \sigma\sqrt{\frac{2}{\pi}}, \quad (2)$$

$$m_2 = \left(1 - \frac{2}{\pi}\right)\sigma^2, \quad (3)$$

$$m_3 = \sqrt{\frac{2}{\pi}}\left(\frac{4-\pi}{\pi}\right)\sigma^3, \quad (4)$$

$$m_4 = \left(\frac{3\pi^2 - 4\pi - 12}{\pi^2}\right)\sigma^4. \quad (5)$$

The standardized third and fourth moments (**skewness** and **kurtosis**, respectively) are

$$\sqrt{\beta}_1 = \frac{(4-\pi)\sqrt{2}}{(\pi-2)^{3/2}} = 0.995272 \quad (6)$$

and

$$\beta_2 = \frac{3\pi^2 - 4\pi - 12}{(\pi-2)^2} = 3.86918. \quad (7)$$

The parameter σ can be estimated by equating the noncentral theoretical and sample moments:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{x_i^2}{n}, \quad (8)$$

where n is the sample size. According to Johnson [4], this is also the **maximum likelihood** estimator.

The half-normal distribution is a special case of the folded normal distribution, where the point of folding is at zero. The folded normal was investigated by Leone et al. [5], Elandt [3], and Johnson [4].

References

[1] Berlin, J.A., Begg, C.B. & Louis, T.A. (1989). An assessment of publication bias using a sample of published clinical trials, *Journal of the American Statistical Association* **84**, 381–392.

[2] Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments, *Technometrics* **1**, 311–341.

[3] Elandt, R. (1961). The folded normal distribution: two methods of estimating parameters from moments, *Technometrics* **3**, 551–562.

[4] Johnson, N.L. (1962). The folded normal distribution: accuracy of estimation by maximum likelihood, *Technometrics* **4**, 249–256.

[5] Leone, F.C., Nelson, L.S. & Nottingham, R.B. (1961). The folded normal distribution, *Technometrics* **4**, 543–550.

R.H. BYERS

Halley, Edmond

Born: November 8, 1656, in Haggerton, UK.
Died: January 14, 1742, in Greenwich, UK.

Edmond Halley was a major English astronomer, mathematician, and physicist, who was also interested in **demography**, insurance mathematics (*see* **Actuarial Methods**), geology, oceanography, geography, and navigation. Moreover, he was considered an engineer and a social statistician whose life was filled with the thrill of discovery. In 1705, he reasoned that the periodic comet – now known as Halley's comet – that appeared in 1456, 1531, 1607, and 1682, was the same comet that appears every 76 years, and accurately predicted that it would appear again in December 1758. His most notable achievements were his discoveries of the motion of stars, which were then considered fixed, and a scheme for computing the motion of comets and establishing their periodicity in elliptical orbits.

Edmond Halley, whose name was also spelled Edmund, was the eldest son of a prosperous landowner, soapmaker, and salter in London. He was tutored at home before attending St Paul's School, where he learned Latin, Greek, and mathematics, including geometry, algebra, the art of navigation, and the science of astronomy. In 1673, at the age of 17, he entered Queen's College, Oxford, and was introduced to John Flamsteed, who was appointed Astronomer Royal in 1676.

In November 1676, Edmond Halley sailed to the island of St Helena, where he cataloged the stars of the southern hemisphere, and incidentally discovered a star cluster in Centaurus, a constellation in the Southern Hemisphere. In 1677, he timed a transit of Mercury and of Venus across the sun and made rough calculations of the mean distance between Earth and the sun. In 1678, he published his results in Catalogus Stellarum Australium, was elected a fellow of The Royal Society, and received the M.A. degree from Oxford University. He married Mary Tooke in 1682, and they had three children – two daughters and one son. He established a home and small observatory center at Islington, and saw the comet of 1682.

Halley encouraged Newton to expand his studies on celestial mechanisms and contributed important editorial aid and financial support to the publication of Newton's major work, *Philosophiae Naturalis Principia Mathematica*, in 1686. From 1685 to 1696 he was assistant of the secretaries of the Royal Society, and from 1685 to 1693 he edited the *Philosophical Transactions of the Royal Society*. In 1698 he was the frequent guest of Peter the Great, who was studying British shipbuilding in England. He was the technical adviser to Queen Anne in the War of Spanish Succession, and in 1702 and 1703 she sent him on diplomatic missions to Europe to advise on the fortification of seaports.

Between 1687 and 1720 Halley published papers on mathematics, ranging from geometry to the computation of logarithms and trigonometric functions. He also published papers on the computation of the focal length of thick lenses and on the calculation of trajectories in gunnery. In 1684 he studied tidal phenomena, and in 1686 he wrote an important paper in geophysics about the trade winds and monsoons. From 1683 to 1692 he published two important papers in geophysics about terrestrial magnetism and made a chart of the variation of the compass. In 1716 he suggested that the aurora was governed by the terrestrial magnetic field.

Halley was a man of great curiosity who combined his astronomical knowledge to help in the dating of historical events. In 1691 he published a paper on the date and place of Julius Caesar's first landing in Britain, and in 1695 he published a paper on the ancient Syrian city of Palmyra. In 1695 he began an intensive study of the movement of the comets, using the hypothesis that cometary paths are nearly parabolic. In 1696 he became deputy controller of the mint at Chester. Between 1698 and 1700, Halley was appointed as a naval captain. He charted magnetic variations while crossing the Atlantic, and was the first to adopt isogonic lines to connect points of equal magnetic variation. In 1704 he was appointed Savilian Professor of Geometry at Oxford and was granted the degree of Doctor of Civil Law. In 1705 he published his cometary views in *Philosophical Transactions*, and *A Synopsis of the Astronomy of Comets*. In 1706 and 1710 he translated and published *Conics*, and *Sectio Rationis of Apollonius*. In 1712 Halley and Newton published *Historia Coelestis*, an edition of Flamsteed's observations, using material deposited at the Royal Society, and infuriated Flamsteed.

Although the major scientific interest of his life was astronomy, Halley wrote a seminal paper on **life**

tables. Since the end of the sixteenth century, registers of births and deaths by sex and age had been well kept in Breslau, Silesia. Caspar Neumann, a prominent evangelical pastor and scientist, used the data to combat some popular superstitions about the influence on health of the phases of the moon and certain ages (those divisible by seven and nine). Neumann sent his results to Leibniz, who in 1689 brought them to the attention of the Royal Society. Since the work of **Graunt** and **Petty**, members of the Royal Society were waiting to receive observations suitable for construction of a life table and sent the data to Halley for analysis. In 1693 Halley wrote the paper "An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the City of Breslaw, with an attempt to ascertain the price of annuities upon lives." Halley assumed a constant number of births per year, mortality by age constant in time, and no migration. He did not present the data in detail, but he calculated a life table based on the number of survivors by year, including the first empirical distribution of deaths according to age. He used the life table to calculate the number of men able to bear arms from age 18 to 56, the **median** remaining lifetime for an individual of age x (*see* **Life Expectancy**), the total population size, and certain calculations relating to annuities. He found that the value of an annuity is the sum of the expectation of the payments made to the living, a concept later pursued by **Abraham de Moivre**. His expectation became the fundamental quantity in life insurance, today called the pure endowment. Having written an important paper on life tables, Halley never returned to the topic, which was far from his main interests.

In 1715 Halley published a paper on novae, and nebulae, and recorded ideas and experiences of living underwater. In 1720 Halley succeeded John Flamsteed in his appointment as Astronomer Royal. In 1729 he was elected a Foreign Member of the Academie des Sciences at Paris. By 1731 he had published a method of using lunar observations for determining longitude at sea. He also studied the question of the size of the universe and the number of stars it contained.

At his death, Edmond Halley was 86 years old and widely mourned. He was a famous and a friendly man of rare intelligence who was always ready to support young astronomers. As Joseph Laland said about Halley, he was "the greatest of English astronomers ... ranking next to Newton among the scientific Englishmen of his time".

For more complete information about Halley's life, see the following references.

References

[1] Abbott, D. (1984). *The Biographical Dictionary of Scientists: Astronomers*. Peter Bedrick Books, New York.

[2] Armitage, A. (1966). *Edmond Halley*. Nelson, London.

[3] Gillipsie, C.C. (1972). *Dictionary of Scientific Biography*, Vol. VI. Charles Scribner's Sons, New York.

[4] Hald, A. (1987). On the early history of life insurance mathematics, *Scandinavian Actuarial Journal* **4**, 18.

[5] Hald, A. (1990). *A History of Probability and Statistics and Their Applications Before 1750*. Wiley, New York.

[6] Muirden, J. (1968). *The Amateur Astronomer's Handbook: A Guide to Exploring the Heavens*. Thomas Y. Crowell, New York.

[7] Ronan, C. (1969). *Astronomers Royal*. Doubleday, New York.

[8] Safra, J.E., chairman. (1997). *The New Encyclopedia Britanica*, Vol. 5. Encyclopedia Britanica, Chicago.

[9] Stephen, L. & Lee, S., eds (1968). *Dictionary of National Biography*, Vol. 8. Oxford University Press, Oxford.

LINA ASMAR

Halperin, Max

Born: November 5, 1917, in Omaha, Nebraska.
Died: February 1, 1988, in Fairfax, Virginia.

Max Halperin was a leading statistician in biostatistics for over 40 years both at the **National Institutes of Health** and at the Biostatistics Center at the George Washington University. At the time of his death he was Research Professor of Statistics and Director of the Biostatistics Center of the George Washington University.

Halperin graduated from the University of Omaha in 1940 with a B.S. degree and from the University of Iowa in 1941 with an M.S. degree, both in mathematics. He earned his Ph.D. in mathematical statistics from the University of North Carolina in 1950. From 1941 to 1946, Halperin served in the Armed Forces primarily with the US Air Force in the China–Burma–India theater of operations.

A brief review of his career begins with the year 1948–1949 when he was a research mathematician at the RAND Corporation where he worked with Alex Mood. He then spent the years 1950–1955 in the Biometrics Department of the US Air Force School of Aviation Medicine at Randolph Field, Texas. He first came to the National Institutes of Health (NIH) in 1951, joining Felix Moore in the Biometrics Research Branch of the National Heart Institute (NHLBI). From 1955 to 1958 he was Chief of the Biometrics Office of the Division of Biologic Standards, NIH. For the next eight years he held positions as statistician in private industry with the General Electric Company and with the Sperry-Rand Corporation. He returned to the NIH in 1966 as Assistant Chief and Chief of the Biometrics Research Branch, NHLBI. After retirement from the NIH in 1977, he spent the remaining years of his career as Research Professor of Statistics and Director of the Biostatistics Center of the Department of Statistics at the George Washington University.

Max Halperin entered the Statistics Department at the University of North Carolina shortly after **Harold Hotelling** became chairman of the Department. His first attempt at a dissertation had to be scrapped since a paper was published on the same topic. In 1948 he met Alex Mood who suggested that he write on the estimation of parameters in truncated samples. He successfully completed his dissertation on this theme, publishing one of the first papers on this subject in the *Annals of Mathematical Statistics* [3].

Max Halperin was widely respected and recognized for his contributions to theoretical and applied statistics and biostatistics. He took great joy in working on theoretical problems, particularly those initiated by his consultations with investigators engaged in scientific research. His theoretical work reflected his strength in **multivariate analysis** and his adeptness in deriving **large sample**, asymptotic distributions. His interests in both theoretical and applied research ranged over a broad spectrum of subjects. He contributed significantly to (i) various topics in **regression** such as inverse estimation [16], **errors in variables** [7, 13], interval estimation in **nonlinear regression** [9, 12]; (ii) interval estimation (*see* **Estimation, Interval**) of parametric nonlinear functions [11, 14, 21]; and (iii) distribution-free tests (*see* **Nonparametric Methods**) [4, 23, 25, 30]. In addition, he wrote on applied probability [5, 19], on reliability [10, 18], and on other problems in general statistical methodology [6, 8, 15, 20]. Halperin collaborated with **Cornfield** and others to write on an alternative solution for the **multiple comparison** problem [24] which turned out to be a powerful method to detect **outliers**, and to write on an adaptive procedure for sequential clinical trials (*see* **Sequential Analysis**) [1].

Halperin and his colleagues in the Biometrics Research Branch of the NHLBI were to a large extent responsible for developing the statistical foundations of the **multicenter clinical trial**. During this period,

as Chief of the Biometrics Branch in the NHLBI and later as Director of the Biostatistics Center, his interests were primarily directed to the **clinical trial**. The more he became involved in the conduct of clinical trials the more he realized that the clinical trial was much more complicated than a simple extension of a laboratory experiment into the community. He was led to consider special aspects of design [17, 31] and problems in **data and safety monitoring** [27]. His greatest effort at this time was devoted to two major topics: stochastic curtailment [28, 32] and early stopping of a clinical trial [2, 22]. His ideas and writings in these areas had a great impact on the planning and direction of clinical trials (*see* **Clinical Trials Protocols**). In addition to his personal research related to clinical trials, Max Halperin greatly influenced the design and conduct of clinical trials through his service on steering, policy advisory, or **data and safety monitoring boards** of many major clinical trials sponsored by the NHLBI.

Towards the end of his career, he was responsible for another novel idea related to multiple comparisons. Conventionally, statisticians looked for protection against making no errors – in an experiment, or family of experiments, etc. Halperin, however, relaxed the requirement by seeking protection against making at most one error, or at most two errors, etc. [29].

Halperin was a member of the Board of Directors of the **American Statistical Association** (1975–77) and served as an Associate Editor of the ***Journal of the American Statistical Association*** (1971–74) and of the *American Statistician* (1976–80). He was a member of a committee on standards for statistical symbols and notation together with H.O. Hartley and P.G. Hoel and was the senior author of the Committee's report [26]. He was Chairman of the Biometrics Section of the American Statistical Association in 1974.

Max Halperin received many honors. He was a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, the American Association for the Advancement of Science, and an elected member of the **International Statistical Institute**. He received a Superior Service Award from the Department of Health, Education and Welfare (1973) and the Statistics Section Award from the **American Public Health Association** in 1985.

Max not only worked on statistical problems – he also loved to talk about statistics, especially to point out the difficulties he was running into on a specific problem. Many of these discussions would occur at lunch, which for his associates became special occasions. The problems on which he worked were primarily those motivated by his work. The sole criterion: Was it real and interesting?

Max married Mary Ann Thomas whom he met while both were working at the National Heart Institute. They have a daughter, Martha.

References

[1] Cornfield, J., Halperin, M. & Greenhouse, S.W. (1969). An adaptive procedure for sequential clinical trials, *Journal of the American Statistical Association* **64**, 759–770.

[2] DeMets, D. & Halperin, M. (1982). Early stopping in the two-sample problem for bounded random variables, *Controlled Clinical Trials* **3**, 1–12.

[3] Halperin, M. (1952). Maximum likelihood estimation in truncated samples, *Annals of Mathematical Statistics* **23**, 226–238.

[4] Halperin, M. (1960). Extension of the Wilcoxon-Mann-Whitney test to samples censored at the same fixed point, *Journal of the American Statistical Association* **55**, 125–138.

[5] Halperin, M. (1960). Some asymptotic results for a coverage problem, *Annals of Mathematical Statistics* **31**, 1063–1076.

[6] Halperin, M. (1961). Almost linearly-optimum combination of unbiased estimates, *Journal of the American Statistical Association* **56**, 36–43.

[7] Halperin, M. (1961). Fitting of straight lines and prediction when both variables are subject to error, *Journal of the American Statistical Association* **56**, 657–669.

[8] Halperin, M. (1963). Approximations to the non-central "*t*", with applications, *Technometrics* **5**, 295–305.

[9] Halperin, M. (1963). Confidence interval estimation of non-linear regression, *Journal of the Royal Statistical Society, Series B* **25**, 330–333.

[10] Halperin, M. (1964). Some waiting time distributions for redundant systems with repair, *Technometrics* **6**, 27–40.

[11] Halperin, M. (1964). Interval estimation of non-linear parametric functions II. *Journal of the American Statistical Association* **59**, 168–181.

[12] Halperin, M. (1964). Note on interval estimation in non-linear regression when responses are correlated, *Journal of the Royal Statistical Society, Series B* **26**, 267–269.

[13] Halperin, M. (1964). Interval estimation in linear regression when both variables are subject to error, *Journal of the American Statistical Association* **59**, 1112–1120.

[14] Halperin, M. (1965). Interval estimation of non-linear parametric functions III. *Journal of the American Statistical Association* **60**, 1191–1199.

[15] Halperin, M. (1967). An inequality on a bivariate Student's "*t*" distribution, *Journal of the American Statistical Association* **62**, 603–606.

[16] Halperin, M. (1970). On inverse estimation in linear regression, *Technometrics* **12**, 727–734.

[17] Halperin, M. (in the MRFIT Group Report) (1977). Statistical design considerations in the NHLBI multiple risk factor trial, *Journal of Chronic Diseases* **30**, 261–275.

[18] Halperin, M. & Burrows, G.L. (1960). The effect of sequential batching for acceptance-rejection sampling upon sample assurance of total product quality, *Technometrics* **2**, 19–26.

[19] Halperin, M. & Burrows, G.L. (1961). An asymptotic distribution for an occupancy problem with statistical applications, *Technometrics* **3**, 79–89.

[20] Halperin, M. & Lan, K.K.G. (1987). A two sample ordered alternative test for means and variances, *Communications in Statistics – Theory and Methods* **16**, 1297–1313.

[21] Halperin, M. & Mantel, N. (1963). Interval estimation of non-linear parametric functions, *Journal of the American Statistical Association* **58**, 611–627.

[22] Halperin, M. & Ware, J.H. (1974). Early decision in a censored Wilcoxon two-sample test for accumulating survival data, *Journal of the American Statistical Association* **69**, 414–422.

[23] Halperin, M., Gilbert, P.R. & Lachin, J.M. (1987). Distribution free confidence intervals for $\Pr\{X(1) < X(2)\}$, *Biometrics* **43**, 71–80.

[24] Halperin, M., Greenhouse, S.W., Cornfield, J. & Zalokar, J. (1955). Tables of percentage points for the Studentized maximum absolute deviate in normal samples, *Journal of the American Statistical Association* **50**, 185–195.

[25] Halperin, M., Hamdy, M. & Thall, P.F. (1989). Distribution-free confidence intervals for a parameter of Wilcoxon-Mann-Whitney type for ordered categories and progressive censoring, *Biometrics* **45**, 509–521.

[26] Halperin, M., Hartley, H.O. & Hoel, P.G. (1965). Recommended standards for statistical symbols and notation, *American Statistician* **19**, 12–14.

[27] Halperin, M., Lan, K.K.G., Ware, J.H., Johnson, N.J. & DeMets, D.L. (1982). An aid to data monitoring of long term clinical trials, *Controlled Clinical Trials* **3**, 311–323.

[28] Halperin, M., Lan, K.K.G., Wright, E.C. & Foulkes, M.A. (1987). Stochastic curtailing for comparison of slopes in longitudinal studies, *Controlled Clinical Trials* **8**, 315–326.

[29] Halperin, M., Lan, K.K.G. & Hamdy, M. (1988). Some implications of an alternative definition of the multiple comparison problem, *Biometrika* **75**, 773–778.

[30] Halperin, M., Ware, J.H. & Wu, M. (1980). Conditional distribution-free tests for the two-sample problem in the presence of right censoring, *Journal of the American Statistical Association* **75**, 638–645.

[31] Lan, K.K.G., DeMets, D. & Halperin, M. (1984). More flexible sequential and non-sequential designs in long-term clinical trials, *Communications in Statistics – Theory and Methods* **13**, 2339–2353.

[32] Lan, K.K.G., Simon, R. & Halperin, M. (1982). Stochastically curtailed testing in long-term clinical trials, *Communications in Statistics – Theory and Methods* **1**, 207–219.

SAMUEL W. GREENHOUSE

Haplotype Analysis

Haplotype analysis examines and attempts to specify the genetic information descending through a pedigree, thus providing a useful visualization of the gene flow. Specifically, a haplotype for a given individual and set of loci is defined as the set of alleles inherited, one per locus, from the same parent (*see* **Gene**). Thus, for each person there are two haplotypes, one of maternal origin and the other paternal. Usually, the loci under consideration are syntenic, that is, the haplotype consists of alleles all on a single chromosome. Traditional haplotype analysis, also known as haplotype reconstruction or simply haplotyping, is the process of obtaining a "best" estimate for each of the two haplotypes for each person in a pedigree. This set of haplotypes is the inferred haplotype vector for that pedigree. For example, Figure 1 shows for a small fully typed pedigree the most likely of the 262 144 haplotype vectors consistent with the data [11, 15].

In addition to traditional haplotyping that simply specifies from which parent each child's allele is descended, there is a more complete form of haplotyping that specifies from which parental *allele* each child's allele is descended, that is, specifies grandparental source information. This more complete haplotype analysis includes sufficient information to describe completely gene flow through a pedigree. For example, consider Figure 2(a), which demonstrates a traditional haplotyping solution at a single locus, that is, everyone has been assigned an ordered **genotype**. (Here, an ordered genotype is listed with the maternal allele on the left, paternal allele on the right.) However, notice that the gene flow is not completely specified, in that the grandparental source of the "1" allele in the grandchildren cannot be determined. Grandparental source information can be displayed by using a gene flow representation. Figure 2(b) shows one of the four gene flow representations consistent with the data in Figure 2(a). Here, each individual is represented by two nodes at each locus: one for the allele of maternal origin and one for the paternal allele. The founders' nodes have specific alleles assigned to them and then arcs are drawn connecting each child node with the parental node from which it descended. This more complete form of haplotyping can be defined as the task of reproducing the complete gene flow information for a pedigree at the loci under consideration.

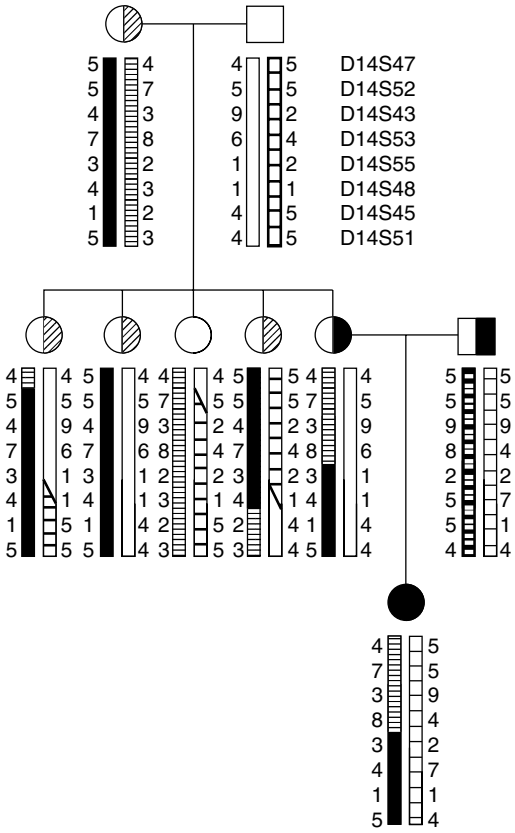

Figure 1 Example of a haplotyped pedigree in which each founder haplotype has been uniquely hatched. This is the most likely of the 262 144 haplotypes consistent with the fully typed pedigree [15]. This pedigree comes from a study of Krabbe disease: the full-black symbol indicates an affected person; half-black indicates obligate carriers; half-hatched indicates carriers identified by an enzyme assay [11] (*see* **Genetic Counseling**). Reproduced from Sobell et al. [15] by permission of Springer-Verlag

Applications of Haplotype Analysis

Haplotype analysis has several common applications. An early goal of haplotyping was to make the genetic data used in **linkage analysis** more informative. A locus is defined to be informative at a mating, that is, for two parents and their child, if the observed typing information at that locus allows one to infer from which parental allele each allele in the child is inherited (*see* **Polymorphism Information Content**). To illustrate this, consider a locus at which the parents are typed as **heterozygotes** with no alleles in common, for example, a/b and c/d, then one is

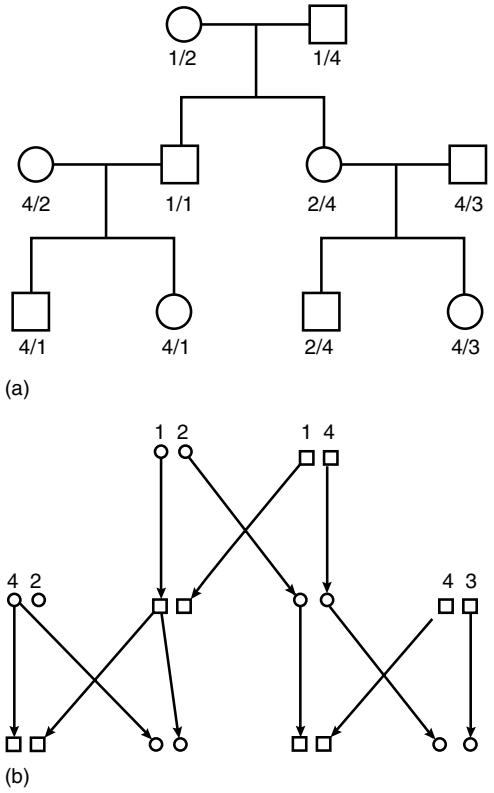

Figure 2 (a) Pedigree data set with one locus in which ordered genotypes have been inferred for each person. (Ordered genotypes are shown as maternal-allele/paternal-allele.) The grandparental origin of the "1" allele cannot be determined; (b) one of the four gene flow representations consistent with the data in (a). Reproduced from Sobel & Lange [14] by permission of the University of Chicago Press

assured an informative mating. Conversely, consider the mating in Figure 3(a) in which all three loci are individually uninformative. To increase informativeness, one may construct a single, highly polymorphic "mega-locus" from a number of less polymorphic, but closely linked, loci. (Indeed, some researchers still use the term haplotyping to refer only to this application or to the related problem of finding the population frequencies of the newly defined "mega-alleles".) The creation of a mega-locus is advantageous because each locus alone may be uninformative for many matings in the pedigree, while the combined mega-locus will often be informative at nearly all matings. For example, the mating in Figure 3(a) is uninformative at all three loci. However, if one can

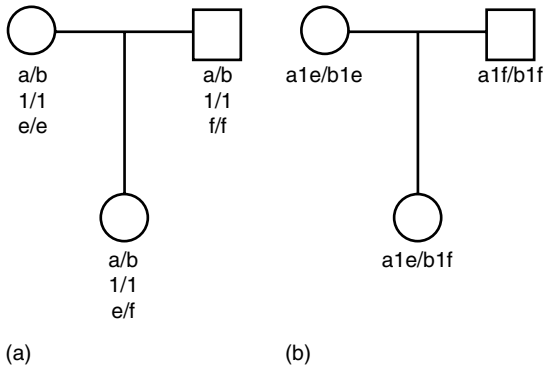

Figure 3 (a) Unordered genotypes at three loci. Each locus is uninformative in this mating; (b) shows the three loci haplotyped and combined into one "mega-locus" that is informative

create haplotypes for these three loci using the rest of the pedigree (not shown), then the newly defined mega-alleles may make the mega-locus informative. Such an informative mega-locus is seen in Figure 3(b). By treating the combined loci as a single point in the genome, the results of standard linkage analysis will often be improved. Clearly, this approach is best suited for closely linked loci, usually with no recombination between the loci.

Haplotyping is also used to identify genotyping or data-entry errors. Even relatively few mistyping errors can have a significant effect on the determination of genetic maps and gene localization [1, 8]. Haplotyping, by exhibiting the gene flow within a pedigree, permits a visual check of the data to find likely mistypings. Mistyping that results in non-Mendelian inheritance is easy to detect, for example, a 1/3 child from two 1/2 parents. (Of course, if these data stand up to retyping, then nonpaternity or non-maternity must be considered.) However, mistyping a true 2/2 child as a 1/2, when both parents are 1/2, is difficult to detect. Haplotyping across this locus may highlight the possibility that the child's typing was in error. For example, in Figure 4 haplotyping reveals a double recombination, one on either side of the questionable allele. If the distance between these flanking markers is small, then the "1" allele in the child would be a definite candidate for retyping. Several papers have discussed statistical tests (usually **likelihood ratio tests** or their approximation) that indicate which typings are most likely to be in error [1, 8].

Finally, haplotype analysis may provide a more precise localization of a putative trait locus than

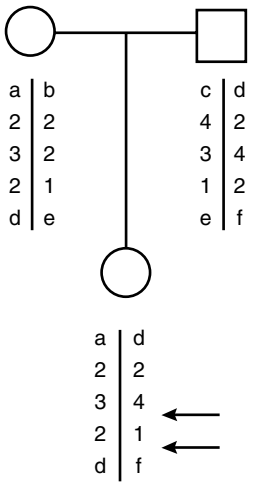

Figure 4 Haplotyping results that suggest that the child's typing may be in error at the "1" allele. The arrows indicate flanking recombination events

standard linkage analysis. The introduction of a rare trait in an isolated population, by new mutation or immigration, probably occurred in only a few ancient individuals. Many of the living affected persons will have inherited the trait from a common founder, even though it is not apparent that they are related. Haplotyping of these affected persons, over loci linked to the trait, may reveal a conserved haplotype inherited through many generations from a common founder. The conserved haplotypes will be flanked by alleles not part of the inferred ancient haplotype. These nonconserved alleles are evidence of recombination events that may have occurred in any generation since the introduction of the trait into the population. The interval contained within all such flanking recombination points is most likely to contain the trait locus. This localization technique, using conserved affected haplotypes in isolated populations, employs recombinations from the entire history of the trait's segregation within the population. Classical linkage analysis can only work with the recombination events manifested within the pedigrees observable today. Thus, it is not surprising that this use of haplotype analysis in an isolated population can often localize a trait to a smaller interval than can classical linkage analysis alone.

As an example of this use of haplotype analysis, consider the localization of the autosomal recessive disorder ataxia-telangiectasia (A-T). In early 1995, classical linkage analysis on an international consortium of 176 pedigrees generated an approximately 500 kb (kilobase) support interval for an A-T gene located on the short arm of chromosome 11 [7]. This support interval roughly spanned the region from S1819 to S1294, which contains the markers S384 and S535 (*see* **Genetic Markers**), as shown in Figure 5. No recombination events were seen in the 176 pedigrees between the A-T locus and either S384 or S535. Figure 6 shows an ancestral haplotype analysis of Costa Rican A-T affected persons and demonstrates that 20 out of the 27 seemingly unrelated affected individuals (and 34 out of the 54 haplotypes) contain a region from an identical ancestral haplotype [19]. Moreover, the boundaries of the conserved haplotypes (see individuals 26–3, 35–3, and 13–3) indicate that the trait locus must be distal to S384. (Here, distal is the direction away from the centromere and proximal is the reverse.) A similar haplotype analysis of a subpopulation of British A-T affected persons concluded that the locus must be proximal to S535 [16]. Haplotype analysis thus localized the A-T gene to the approximately 200 kb interval between S384 and S535; the approximately 100 kb A-T gene, now called ATM, was subsequently found within this interval [12].

Origins of Computational Complexity

Haplotyping is not conceptually difficult – it is simply determining the parental (and grandparental) origin of the children's alleles. What makes haplotyping

Figure 5 Map of four loci on chromosome 11 that are closely linked to the A-T gene. The 500 kb support interval is indicated for the A-T locus found using linkage analysis

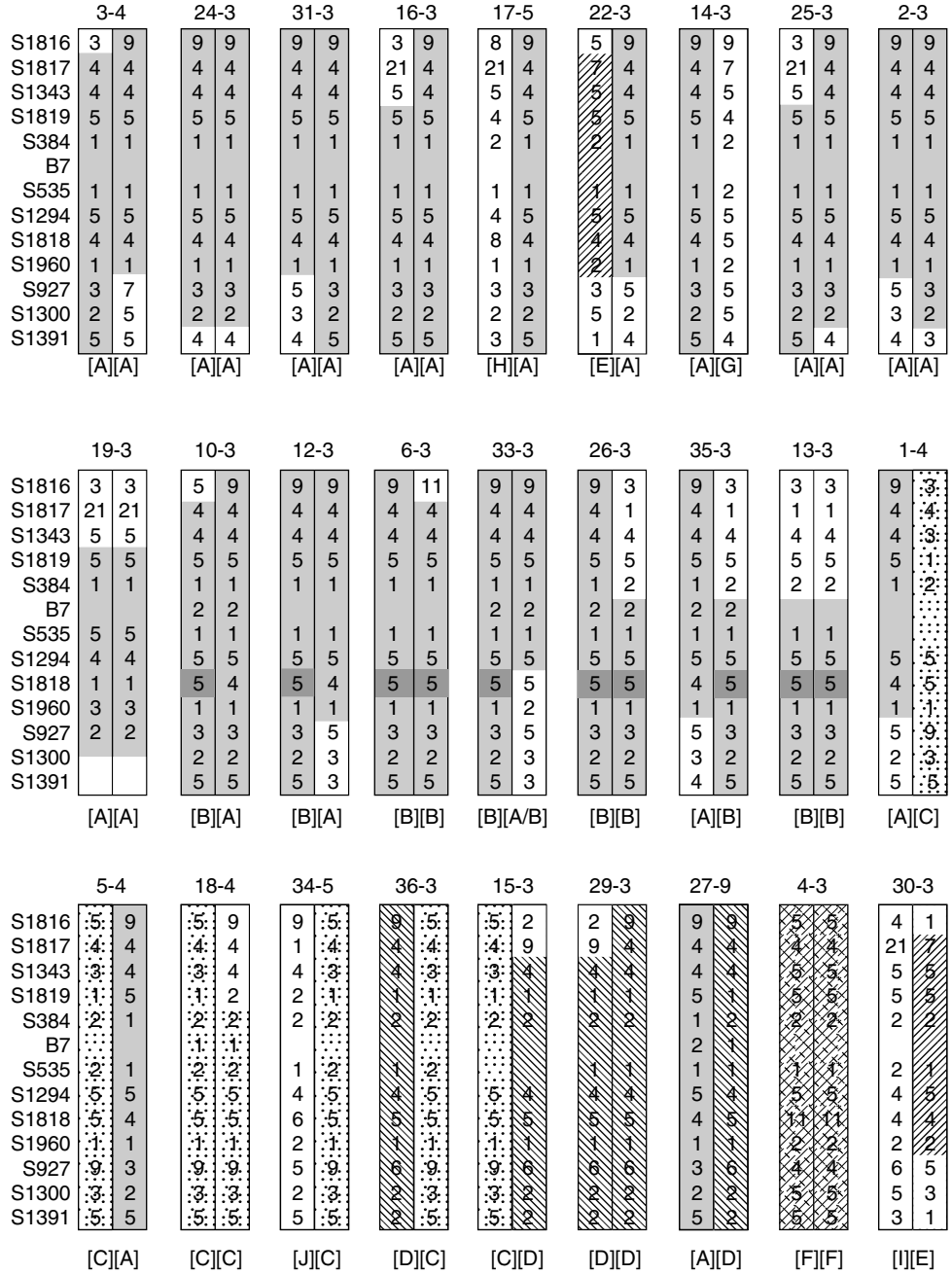

Figure 6 Haplotyping results of 27 living Costa Rican A-T affected persons. The individuals are labeled above the haplotypes; the haplotypes are labeled by bracketed capital letters. The haplotypes of common ancestral origin are similarly hatched. Unique haplotypes (G, H, I, and J) are not hatched. Reproduced from Uhrhammer et al. [19] by permission of the University of Chicago Press

exceedingly difficult in practice is the amount of missing information usually encountered in a pedigree data set. With even moderate amounts of missing data, the number of haplotype vectors that are consistent with the observed data can grow to astronomical levels. Thus, the search for the "best" haplotype vector is often a nontrivial combinatorial optimization problem.

For typical haplotyping problems, the missing data come in the following three forms: (i) Unknown typing is seen in real pedigrees because there are often some people who are simply unavailable for reliable typing at all the loci under consideration. These people may be too remote for sampling; they may decline to participate; or they may simply be deceased. (ii) Phase information for a locus at a typed individual specifies which allele is maternally inherited and which paternally. Modern genetic data, usually marker genotypes and trait phenotypes, do not specify phase. Marker loci are missing phase information because almost all are codominant loci that yield unordered genotypes. In the case of trait phenotypes, even the underlying alleles may be obscured, for example, a dominant allele will hide the value of the other allele. (iii) Grandparental source information is also required to be assured of complete knowledge of the gene flow through a pedigree (see Figure 2). However, no source information is specifically included in conventional pedigree data, although occasionally some can be directly inferred.

Much missing data of any type will result in a large number of possible haplotype vectors, each consistent with the data. For example, without phase information even a fully-typed pedigree can have an abundance of consistent haplotype vectors. Specifically, with p people fully typed over l loci there can be as many as 2^{pl} haplotype vectors consistent with the data [15]. For real fully typed pedigrees this value is usually significantly smaller, lowered by homozygosity in the founders and informativeness in the matings. The inheritance patterns in the vicinity of the people with missing data can help one infer the missing values, but the trend towards highly **polymorphic** marker loci increases the number of possible haplotypes the missing data imply. Considering the extent to which each of the three types of data is usually absent, it is not surprising that searching for the best haplotype vector is often computationally complex.

To demonstrate the size of the space that one must search to choose a best haplotype vector for a pedigree, consider the 36 person pedigree structure shown in Figure 7. This pedigree is from a study of dopa-responsive dystonia [10]. We simulated randomly complete gene flow data on the 14 linked polymorphic markers used in the linkage study. We considered then only the resulting unordered typing information for each individual, except for the six people unavailable for typing in the actual study, for whom no typing information was included. Table 1 lists for each person the number of haplotype pairs consistent with the simulated data. The number of haplotype vectors for the pedigree would be somewhat less than the product, over all people, of these numbers of haplotype pairs. Thus, an exhaustive search of the possible haplotype vectors is well beyond practical computability. (This analysis was previously reported by Sobel et al. [15].)

Algorithms for Haplotype Analysis

The algorithms that have been devised to overcome the computational complexities of haplotyping have evolved considerably with changes in technology. The trend has been away from heuristic and rule-based approaches toward more **likelihood**-based methods, while using sophisticated techniques to avoid as much as possible the computational bottleneck of calculating likelihoods for pedigrees with significant missing data. Manual haplotyping, the first

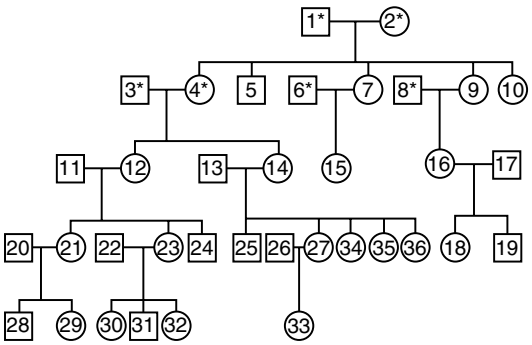

Figure 7 This 36-person pedigree structure is part of the data set used to generate Table 1. The individuals marked with an asterisk have no typing information assigned to them. Everyone else is typed at 14 polymorphic linked loci. Reproduced from Sobel et al. [15] by permission of Springer-Verlag

Table 1 The number of possible haplotype pairs for each individual in the pedigree in Figure 7. Reproduced from Sobel et al. [15] by permission of Springer-Verlag

Person	Haplotype pairs	Person	Haplotype pairs
1	29,421,583,551,360	19	1
2	29,421,583,551,360	20	2048
3	85,944,603,802,337,280	21	2
4	46,558,955,520,000	22	2048
5	75,776	23	2
6	1,024,479,830,005,632	24	2
7	2048	25	8
8	987,891,264,648,288	26	256
9	256	27	2
10	7168	28	2
11	2048	29	1
12	2048	30	2
13	256	31	8
14	2048	32	1
15	4	33	1
16	4	34	16
17	2048	35	8
18	2	36	1

method used, and still used by some, is likely not to consider all the possibilities, even for moderate-sized pedigrees. Indeed, the majority of published haplotype vectors we have examined can be improved; for examples, see Sobel et al. [15]. Another problem with manual haplotyping is that it is difficult to prove that one has performed sufficient analysis to enable others to have confidence in the results. Clearly, the tedious and error-prone nature of haplotyping lends itself to computer-based approaches.

The first class of widely-used computer **algorithms** specifically designed for haplotype analysis were rule-based: PATCH by Wijsman [24] and CHROMLOOK by Haines [4]. By using logical rules to transform the available typing information into inferred underlying haplotypes these programs avoided all likelihood calculations. This approach was developed because at that time the major likelihood calculation programs were fairly limited, by computer memory and time constraints, in the size and complexity of the pedigrees they could handle. Thus, these nonnumeric rule-based algorithms are faster than any other approach. However, rule-based algorithms are by their nature somewhat *ad hoc*. Particularly in the presence of nontyping or uninformative matings, these approaches may leave portions of the haplotypes undetermined. Also, by not considering the recombination fractions between loci, which can be quite varied, these methods may miss a more likely solution.

In contrast to the qualitative rule-based methods, several quantitative algorithms have been devised. All are based on searching for the haplotype vector that maximizes a likelihood calculation. The most straightforward is the exhaustive enumeration technique. For small pedigrees with few untyped people, one can consider systematically every possible haplotype vector and rank them by exact likelihood. This brute-force approach has become feasible with the speed of modern computers and likelihood calculation programs [21]. This algorithm is guaranteed to find the **maximum likelihood** solution, but is only practical for small pedigrees.

To handle larger pedigrees it is necessary to invoke some scheme to reduce the number of possible haplotype vectors that are considered or to reduce the number of calculations required for each vector. One such strategy is not to compute exact likelihoods over the untyped and uninformative loci, which is where the calculations become complex because the number of possible configurations becomes large. This is the strategy used by the CHROMPIC option of the CRI-MAP program developed by Green et al. [2]. Again, however, nontyping may lead to significant uncertainty or even leave portions of the haplotypes undetermined.

Another implemented strategy reduces the number of configurations that need be considered by using a sequential conditional probability algorithm. Here, the haplotype pairs are assigned to people in the pedigree in a sequential fashion. Once the first $i-1$ haplotype pairs have been assigned, individual i is assigned the most probable haplotype pair given the observed data and the previously assigned haplotypes. This is the method used in the program HAPLO developed by Weeks et al. [15, 21]. This method can accommodate pedigrees with a modicum of missing data, in which case the order in which the haplotypes are assigned can affect significantly the amount of computation required.

The next strategy for surveying the space of possible haplotype vectors in reasonable time employs the combinatorial optimization technique known as simulated annealing. This technique has been shown to work on many previously intractable optimization problems in many fields [20]. However, the stochastic nature of simulated annealing implies that one is assured only of reaching a near-optimal solution and repeated applications are normally suggested for confidence. SIMCROSS, developed by Weeks et al. [15, 21], includes a particularly fast implementation of simulated annealing for haplotyping. The speed is attained by using an easily calculated **pseudo-likelihood** for the haplotypes. Specifically, for each locus interval i with recombination fraction θ_i, let $\rho_i = \theta_i$ if interval i contains a crossover, and $1-\theta_i$, otherwise; SIMCROSS uses the pseudo-likelihood $\prod_i \rho_i$. Missing data and large pedigrees can be accommodated by simulated annealing, because it visits only a small fraction of the possible configurations and can escape local maxima of the search space.

Yet another strategy for efficiently searching the space of haplotype vectors uses the gene flow representation of pedigrees as modeled in Figure 2(b). If one ignores the actual alleles assigned to the founder nodes in such a complete gene flow representation, then one is left with only the graph of inheritance paths. Two facts are notable about the space of all such graphs that are consistent with the observed typing. First, the number of these graphs is smaller than the number of possible haplotype vectors – much smaller if there is nontyping in the pedigree. Secondly, given a graph it is straightforward to find the set of founder alleles such that the complete gene flow representation (that is, the combination of the graph and the founder alleles) has maximum likelihood [14]. Thus, one can search the relatively small space of graphs and rank each graph by the maximum likelihood of all haplotype vectors consistent with the graph. Moreover, by conducting this search of the space of graphs, one is, in effect, searching the space of all possible haplotype vectors.

Two haplotyping programs use inheritance graphs, although with different techniques for searching the graph space. For pedigrees in which $2n - f \leq 16$, where n is the number of nonfounders and f the number of founders, GENEHUNTER, by Kruglyak et al. [5], uses the Viterbi algorithm over hidden **Markov chains** to search the graph space comprehensively in reasonable time. Thus, as with the exhaustive enumeration technique, GENEHUNTER is guaranteed to find the maximum likelihood solution and yet is practical for pedigrees of small to modest size, including those with missing data.

However, for large complex pedigrees even the space of inheritance graphs can become too large for deterministic analysis. Similar to SIMCROSS, SIMWALK2, by Sobel & Lange [14], also uses simulated annealing, but to survey the graph space and thus obtain an estimate of the most likely graph. The simulated annealing is performed on a **Markov process** that moves between graphs using the Metropolis criterion [9], that is, in proportion to the ratio of their exact likelihoods. Although SIMWALK2 may be somewhat slower than the above programs for simple pedigrees (except for the exhaustive enumeration approach, which is, of course, the slowest), it can provide in reasonable time a good estimate of the best haplotype vector for even the most complex pedigree data set.

Conclusions

Haplotype analysis has evolved considerably, driven by the rapid change in genetic and computer technology. This evolution has moved from an *ad hoc* qualitative methodology to quantitative estimations based on maximum likelihood considerations. Despite this progress, haplotype reconstruction can still fail to recover the true haplotype vector for a pedigree, particularly in the presence of large intervals between loci and significant amounts of nontyping. It may simply be that the true state is not the most likely state consistent with the observed data [15]. However, the

process of haplotyping may well reveal those specific regions in which more information is needed to pinpoint the true underlying haplotype vector [5].

It is interesting to note that the most recent advances in haplotyping involve techniques that are also proving useful in other branches of pedigree analysis. For example, many areas of pedigree analysis have profited from the use of the gene flow representation of genetic data (see Sobel & Lange [13] and Thompson [18] for reviews). Clearly, haplotyping to the level of complete gene flow information is equivalent to specifying all the identity by descent (IBD) characteristics in the pedigree (*see* **Identity Coefficients**). Thus, it is not surprising that the directed inheritance graphs and the Markov process techniques mentioned above have also been proposed for use in robust nonparametric IBD-based linkage statistics [3, 5, 14, 22, 23] and for **Markov chain Monte Carlo** (MCMC) **multipoint linkage analysis** [5, 6, 13, 14, 17]. These multiple applications demonstrate the importance and flexibility of the gene flow representation and the Markov process methods employed. Moreover, it shows that haplotype analysis can play a central role in the confluence of statistics and genetics that exists in the field of pedigree analysis.

References

[1] Ehm, M.G., Kimmel, M. & Cottingham, R.W., Jr (1996). Error detection for genetic data, using likelihood methods, *American Journal of Human Genetics* **58**, 225–234.

[2] Green, P., Falls, K. & Crooks, S. (1990). Documentation for CRI-MAP 2.4. (Unpublished software documentation.).

[3] Guo, S.-W. (1995). Proportion of genome shared identical by descent by relatives: concept, computation, and applications, *American Journal of Human Genetics* **56**, 1468–1476.

[4] Haines, J.L. (1992). CHROMLOOK: An interactive program for error detection and mapping in reference linkage data, *Genomics* **14**, 517–519.

[5] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified approach, *American Journal of Human Genetics* **58**, 1347–1363.

[6] Lander, E. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans, *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2363–2367.

[7] Lange, E., Borresen, A.-L., Chen, X., Chessa, L., Chiplunkar, S., Concannon, P., Dandekar, S., Gerken, S., Lange, K., Liang, T., McConville, C., Polakow, J., Porras, O., Rotman, G., Sanal, O., Sheikhavandi, S., Shiloh, Y., Sobel, E., Taylor, M., Telatar, M., Teraoka, S., Tolun, A., Udar, N., Uhrhammer, N., Vanagaite, L., Wang, Z., Wapelhorst, B., Yang, H.-M. Yang, L., Ziv, Y. & Gatti, R.A. (1995). Localization of an ataxia-telangiectasia gene to an ∼500-kb interval on chromosome 11q23.1: linkage analysis of 176 families by an international consortium, *American Journal of Human Genetics* **57**, 112–119.

[8] Lincoln, S.E. & Lander, E.S. (1992). Systematic detection of errors in genetic linkage data, *Genomics* **14**, 604–610.

[9] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.

[10] Nygaard, T.G., Wilhelmsen, K.C., Risch, N.J., Brown, D.L., Trugman, J.M., Gilliam, T.C., Fahn, S. & Weeks, D.E. (1993). Linkage mapping of dopa-responsive dystonia (DRD) to chromosome 14q, *Nature Genetics* **5**, 386–391.

[11] Oehlmann, R., Zlotogora, J., Wenger, D.A. & Knowlton, R.G. (1993). Localization of the Krabbe disease gene (GALC) on chromosome 14 by multipoint linkage analysis, *American Journal of Human Genetics* **53**, 1250–1255.

[12] Savitsky, K., Bar-Shira, A., Gilad, S., Rotman, G., Ziv, Y., Vanagaite, L., Tagle, D.A., Smith, S., Uziel, T., Sfez, S., Ashkenazi, M., Pecker, I., Frydman, M., Harnik, R., Patanjali, S.R., Simmons, A., Clines, G.A., Sartiel, A., Gatti, R.A., Chessa, L., Sanal, O., Lavin, M.F., Jaspers, N.G.J., Taylor, A.M.R., Arlett, C.F., Miki, T., Weissman, S.M., Lovett, M., Collins, F.S. & Shiloh, Y. (1995). A single ataxia-telangiectasia gene with a product similar to PI-3 kinase, *Science* **268**, 1749–1753.

[13] Sobel, E. & Lange, K. (1993). Metropolis sampling in pedigree analysis, *Statistical Methods in Medical Research* **2**, 263–282.

[14] Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics, *American Journal of Human Genetics* **58**, 1323–1337.

[15] Sobel, E., Lange, K., O'Connell, J.R. & Weeks, D.E. (1996). Haplotyping algorithms, in *Genetic Mapping and DNA Sequencing, IMA Volume 81 in Mathematics and its Applications*, T.P. Speed & M.S. Waterman, eds. Springer-Verlag, New York, pp. 89–110.

[16] Taylor, A.M.R., McConville, C.M., Rotman, G., Shiloh, Y. & Byrd, P.J. (1994). A haplotype common to intermediate radiosensitivity variants of ataxia-telangiectasia in the UK, *International Journal of Radiation Biology* **66**, S35–S41.

[17] Thompson, E.A. (1994). Monte Carlo likelihood in genetic mapping, *Statistical Science* **9**, 355–366.

[18] Thompson, E.A. (1996). Likelihood and linkage: from Fisher to the future, *Annals of Statistics* **24**, 449–465.

[19] Uhrhammer, N., Lange, E., Parras, O., Naiem, A., Chen, X., Sheikhavandi, S., Chiplunkar, S., Yang, L.,

Dandekar, S., Liang, T., Patel, N., Teraoka, S., Udar, N., Calvo, N., Concannon, P., Lange, K. & Gatti, R.A. (1995). Sublocalization of an ataxia-telangiectasia gene distal to D11S384 by ancestral haplotyping in Costa Rican families, *American Journal of Human Genetics* **57**, 103–111.

[20] van Laarhoven, P.J.M. & Aarts, E.H.L. (1987). *Simulated Annealing: Theory and Applications*. Reidel, Dordrecht.

[21] Weeks, D.E., Sobel, E., O'Connell, J.R. & Lange, K. (1995). Computer programs for multilocus haplotyping of general pedigrees, *American Journal of Human Genetics* **56**, 1506–1507.

[22] Whittemore, A.S. & Halpern, J. (1994). Probability of gene identity by descent: computation and applications, *Biometrics* **50**, 109–117.

[23] Whittemore, A.S. & Halpern, J. (1994). A class of tests for linkage using affected pedigree members, *Biometrics* **50**, 118–127.

[24] Wijsman, E.M. (1987). A deductive method of haplotype analysis in pedigrees, *American Journal of Human Genetics* **41**, 356–373.

E. SOBEL & D.E. WEEKS

Hardy–Weinberg Equilibrium

The Hardy–Weinberg Equilibrium (HWE) principle, the most fundamental rule of **population genetics**, prescribes the **genotype** frequencies at a locus in terms of its allele frequencies in a population. In the most general form, it states that in the absence of mutation, selection, migration, and random genetic drift (*see* **Population Genetics**), and with random mating in a population, the genotype frequencies at an autosomal locus in a large population will reach equilibrium in a single generation and will continue to be in proportions given by the expansion of $(p_1A_1 + p_2A_2 + \cdots + p_kA_k)^2$, where p_i, $i = 1, 2, \ldots, k$, are the frequencies of k alleles $A_1, A_2, \ldots, A_k$ at the locus in the population. In other words, the frequency of a homozygote A_iA_i becomes p_i^2 and that of a **heterozygote** A_iA_j becomes $2p_ip_j$, the rule that was independently discovered by the British mathematician G.H. Hardy [8] and the German physician Weinberg [20]. Of course, before them, Yule [22], Pearson [15], and Castle [1] noted that this rule works for the special cases of allele frequencies at a biallelic locus (see [12] and [17] for historical notes on the discovery of HWE).

Several authors attempted to pay tribute to Castle's [1] work by renaming this rule as Castle–Hardy–Weinberg's (CHW) law (see, for example, [13]). HWE, as a predictive equation for genotype frequencies in a large population in terms of the allele frequencies at a locus, has played a pivotal role for many other population genetic principles. For example, since the equilibrium is reached in a single generation, it implies that if mating is at random, then to understand the genotypic composition of a population it is not necessary to investigate the past history of the population. Also, the rule implies that random mating (with regard to the locus under study) is equivalent to random union of gametes. Furthermore, under this rule, the frequency of a rare recessive gene is about one-half of its heterozygote carrier frequency, and, for rare dominant diseases, the frequency of affected individuals in a large population is approximately twice the allele frequency.

Since the conditions (i.e. no preferential mating, no viability and/or fertility differential of alleles, no immigration or emigration, no mutation, and infinite population size) under which HWE is strictly valid are quite severe, and perhaps no real population satisfies most of these conditions, the applicability of HWE in predicting genotype frequencies is still being questioned in current work (see, for example, [11]). The early optimism of the robustness of HWE, however, has turned out to be justified, since for most loci for which the allelic effects are not physiologically meaningful (e.g. **blood groups**, enzyme-proteins, DNA markers), the rule provides a good approximation to reality. This is so, because, in nonexperimental populations, the extent of deviation from HWE is generally so small that the statistical **power** of its detection is "notoriously" small [5, 19].

While HWE can be extended to X-linked loci, to polyploid genetic systems, and even to genotype frequencies at linked loci, the critical difference is that the approach to equilibrium under these systems is gradual, instead of being reached in a single generation (see [7] for discussions on these systems). Deviations from HWE, in the presence of "nondetectable" alleles, and/or mixture of subpopulations that do not completely interbreed, are also well studied (see, for example, [2, 3, 6, 16, 18], and [21]), indicating that unless subpopulations are genetically well differentiated, or the nondetectable alleles are at high frequency in the population, the approximation of HWE is accurate relative to the usual sampling error of genotype frequency evaluation. Both of these factors cause the expected frequencies of homozygotes to be increased, with corresponding deficiencies of heterozygotes in relation to the predictions of HWE, although the deviations are small and are not generally detectable [4, 14]. In contrast, the finite size of a population is expected to cause a reduction of homozygote frequency, with heterozygote frequencies correspondingly increased by a factor of the order of the inverse of twice the breeding size of a population [7, 9, 10].

References

[1] Castle, W.E. (1903). The laws of heredity of Galton and Mendel, and some laws governing race improvement by selection, *Proceedings of the American Academy of Arts and Sciences* **39**, 223–242.

[2] Chakraborty, R. & Danker-Hopfe, H. (1991). Analysis of population structure: a comparative analysis of different estimators of Wright's fixation index, in *Handbook of*

Statistics, Vol. 8, C.R. Rao & R. Chakraborty, eds. Elsevier, Amsterdam, pp. 203–254.

[3] Chakraborty, R. & Jin, L. (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting, *Human Genetics* **88**, 267–272.

[4] Chakraborty, R. & Kidd, K.K. (1991). The utility of DNA typing in forensic work, *Science* **254**, 1735–1739.

[5] Chakraborty, R. & Rao, D.C. (1972). Detection of the inbreeding coefficient from ABO blood-group data, *American Journal of Human Genetics* **24**, 352–354.

[6] Chakraborty, R., Zhong, Y., Jin, L. & Budowle, B. (1994). Non-detectability of restriction fragments and independence of DNA-fragment sizes within and between loci in RFLP typing of DNA, *American Journal of Human Genetics* **55**, 391–401.

[7] Crow, J.F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.

[8] Hardy, G.H. (1908). Mendelian proportions in a mixed population, *Science* **28**, 49–50.

[9] Hogben, L. (1946). *An Introduction to Mathematical Genetics*. Norton, New York.

[10] Levene, H. (1949). On a matching problem arising in genetics, *Annals of Mathematical Statistics* **20**, 91–94.

[11] Lewontin, R.C. & Hartl, D.L. (1991). Population genetics in forensic DNA typing, *Science* **254**, 1745–1750.

[12] Li, C.C. (1967). Castle's early work on selection and equilibrium, *American Journal of Human Genetics* **19**, 70–74.

[13] Li, C.C. (1976). *First Course in Population Genetics*. Pacific Grove, California.

[14] NRC (1996). *Evaluation of Forensic DNA Evidence*. National Research Council, Washington.

[15] Pearson, K. (1904). On a generalized theory of alternative inheritance, with special reference to Mendel's laws, *Philosophical Transactions of the Royal Society of London, Series A* **203**, 53–86.

[16] Smith, C.A.B. (1970). A note on testing the Hardy-Weinberg law, *Annals of Human Genetics* **33**, 377–383.

[17] Stern, C. (1943). The Hardy-Weinberg law, *Science* **97**, 137–138.

[18] Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationsercheinungen von Standpunkt der Vererbungslehre aus betrachtet, *Hereditas* **11**, 65–106.

[19] Ward, R.H. & Sing, C.F. (1970). A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations, *American Naturalist* **104**, 355–365.

[20] Weinberg, W. (1908). Uber den Nachweis der Vererbung beim Menschen, *Jahreshefte des Vereins für Vaterändische Naturkunde in Württemberg* **64**, 368–382.

[21] Weir, B.S. & Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure, *Evolution* **38**, 1358–1370.

[22] Yule, G.U. (1902). Mendel's laws and their probable relations to intra-racial heredity, *New Phytologist* **1**, 193–207 and 222–237.

RANAJIT CHAKRABORTY

Hawkins, Francis Bisset

Born: 1796
Died: 1894

Bisset Hawkins is most widely remembered as the author of the first book on medical statistics in the English language [2]. He was the first Professor of Materia Medica at King's College, London, and was a prolific author in the fields of industrial medicine and public health. He was a founder member in 1834 of the Statistical Society of London (later the **Royal Statistical Society**), although he does not appear to have played a very active part in its later proceedings, and his death at an advanced age went unrecorded in the Society's *Journal* (*see* ***Journal of The Royal Statistical Society***).

According to **Greenwood** [1], Hawkins "was instrumental in obtaining the insertion in the first Registration Act of a column containing the names of the diseases or causes by which death was occasioned", initially on a voluntary basis. That may prove to be a more lasting claim to fame than the celebrated book.

Unfortunately, *Elements of Medical Statistics* now has only curiosity value. Hawkins adopts a purely descriptive approach, relying heavily on crude death rates, but with some appreciation of the effects of the age structure of a population on demographic measures such as the **average age at death**. His detailed comments often show a remarkable lack of critical awareness. The book contains many complimentary remarks about Manchester, which he apparently thought to have a remarkably low death rate (1 in 74, as compared with 1 in 43 for Birmingham and 1 in 40 for London). Unfortunately, he had made an arithmetic slip, and in a copy of the book owned by the Royal Statistical Society all the complimentary references to Manchester are scored out, apparently in his own hand. Again, he seems to have accepted anecdotes from classical antiquity with a degree of naivety. He refers uncritically to reports of individuals living to ages greater than 150 years; and he uses a fatality rate from acute fevers quoted by Hippocrates as a control for more favorable figures recorded in 1825.

As Greenwood remarks, Hawkins "had been diligent and brought together numerical data from all parts of the world and was certainly one of the first physicians to advocate a serious study of hospital records". The work of pioneers often shows traces of fallibility, but they deserve to be remembered for their achievements rather than their weaknesses. Hawkins lacked the mathematical abilities of the younger physicians **W.A. Guy** in England and **Jules Gavarret** in France, but he helped to ensure that medical applications played a significant part in the enormous growth of statistical activity during the first half of the nineteenth century.

References

[1] Greenwood, M. (1948). *Medical Statistics from Graunt to Farr*. Cambridge University Press, Cambridge.

[2] Hawkins, F.B. (1829). *Elements of Medical Statistics*. Longman, Rees, Orme, Brown & Green, London.

PETER ARMITAGE

Hawthorne Effect

The Hawthorne effect is an effect on study participants that results from their knowing that they are being studied. For example, in a study of methods to promote smoking cessation, it might be necessary to contact study participants each year to determine smoking status. The Hawthorne effect could distort study results if this repeated annual contact affected smoking behavior or the reporting of smoking behavior.

MITCHELL H. GAIL

Hazard Models Based on First-passage Time Distributions

Introduction

Survival and **event history analysis** study the occurrence of events. It seems highly reasonable to imagine that these events depend on underlying processes that determine the occurrence. The dissolution of a marriage does not happen out of the blue; most often there is a process of deterioration prior to divorce. The occurrence of myocardial infarction, as another example, is the manifestation of a process of occlusion of coronary arteries. The reason why these underlying processes do not figure in the statistical analysis of survival data is that they are usually unobserved. Occasionally, partial information may be available through **covariates** or **marker processes**.

Although observation of the process is a problem, it might still be of interest to model survival data with the idea of an underlying process in mind. Such models will be highly idealized, but may still yield some insight. Typically, what one would do is to imagine that the process can be described roughly by some standard **stochastic process** with simple mathematical properties. It could be a **Markov** model on a finite state space, with a number of transient states and one absorbing state, such that the occurrence of the event in question corresponds to entrance into the absorbing state. Such distributions are called ***phase-type distributions***. Or the process could be modeled as a Wiener process (**Brownian motion**) with drift, such that the event occurs when the process passes a certain limit. The first-passage time in this case follows an ***inverse Gaussian distribution***. This inverse Gaussian family has interesting and useful properties and should definitely be used much more in survival analysis and other fields.

Assuming the stochastic process point of view allows for some general considerations of interest [2], we shall consider a population of individuals on the transient state space of the stochastic process (i.e. prior to absorption). In a sense, there are two forces operating on these individuals, namely, the general diffusion in the transient state space, and the attraction of the absorbing state. The rate at which absorption takes place (i.e. the **hazard rate** of the first-passage distribution) is created in a balance between these two forces. Understanding the shape of the hazard rate is in general difficult, and the present point of view might help in doing so.

A very useful concept in this context is that of an *attractor*. For many processes, the distribution of survivors (i.e. the distribution on the transient state space) moves towards an attractor, that is, a fixed distribution on the state space. The attractor is often termed a *quasi-stationary limit*. It is the limiting distribution given nonabsorption, and it is a somewhat surprising fact that such a limit often exists in spite of the continuous leaking of probability mass into the absorbing state.

A practical reason for considering the underlying processes is that many, if not most, covariates used in survival analysis are not really **risk factors**, but rather measures of how far advanced a disease is. An example is the staging measures used in cancer (*see* **Oncology**). A statistical application of this idea is given in [2].

Markov Chain with Absorption

The simplest stochastic process of use in this context, is a time-continuous **Markov chain** on a finite state space. One state is absorbing, and the idea is that the process starts out according to some distribution on the remaining (transient) state space. The first-passage time is the time to absorption, and its distribution is termed a *phase-type distribution*; see [1]. Such distributions play an important role in probability theory and there exists a lot of theory for their properties.

An example is given in Figure 1, which represents a Markov chain where an individual can move up and down in a state space before eventually being absorbed. This could model a disease where "repair" and improvement occur in various stages before the disease in the end moves into a nonreversible state. The interesting thing about this model is that it gives a nice illustration of how the shape of the hazard rate is influenced by the distance of the starting point from the absorbing state. Starting in states 1, 3, and 5 respectively and computing the hazard rates of time to absorption, one gets the result illustrated in Figure 2 (parameters equal to $\alpha = 1.5$ and $\beta = 1.0$). Hence, the three major shapes of hazard rates occurring in practice – the increasing, the decreasing, and the

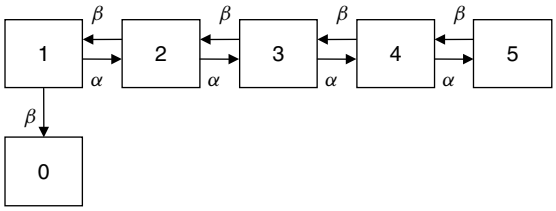

Figure 1 State space of a phase-type model. (Reprinted from Aalen and Gjessing, 2001. Reproduced by permission of the authors)

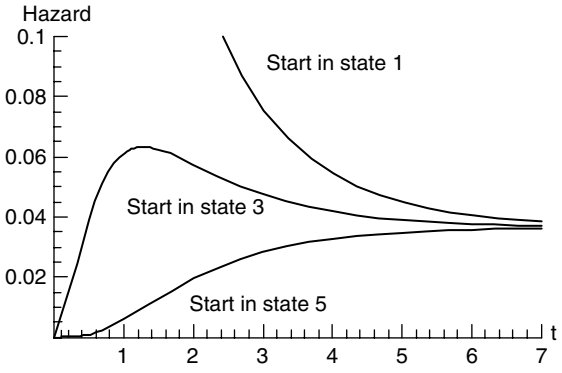

Figure 2 Hazard rates for time to absorption dependent on starting state in phase-type model. Parameters: $\alpha = 1.5$, $\beta = 1.0$. (Reprinted from Aalen and Gjessing 2001. Reproduced by permission of the authors)

unimodal one – arise naturally. This is no accident, but has to do with how the starting points relate to the quasi-stationary distribution. For the parameter values chosen here, the quasi-stationary distribution on states 1 to 5 is given as 0.037, 0.090, 0.167, 0.276, 0.430. With this starting distribution, the hazard rate of absorption is constant and equal to 0.0365. The three states 1, 3, and 5 can be considered to have, respectively, a short, medium, and long distance from the absorbing state, compared with a quasi-stationary distribution.

Sometimes, a unimodal hazard rate (increasing first and then decreasing) has been given the naive interpretation that its maximum point represents a point of "crisis". An example is the "seventh year itch" interpretation of divorce rates peaking after about seven years, the idea being that a typical marriage goes through a crisis after seven years, and then the divorce rate declines when it passes through the crisis. A more sophisticated interpretation of a declining hazard rate as a result of selection effects has been promoted within the context of ***frailty*** theory. The present interpretation of varying hazard rate shapes is yet another one. It tells us in fact that a unimodal hazard would be expected to arise quite often, as a natural phenomenon.

An extension of this theory to **semi-Markov processes** has been developed; see [6].

Wiener Process with Absorption

The Wiener process is the prototype of a continuous and completely random stochastic process. When combined with a drift term, it is an interesting model of a process that approaches some aim with a lot of random detours along the way; see [2, 4, 5] for statistical applications. We shall consider a Wiener process with drift $-\mu$ ($\mu > 0$, that is, drift towards zero) and variance coefficient σ^2, this meaning that the increments of the process over an interval of length Δt is normally distributed with mean $-\mu\Delta t$ and variance $\sigma^2\Delta t$. The process starts out in some positive state, c, and is absorbed when it hits zero at some time T. The distribution of T is the classical example of a first-passage time distribution, and is usually denoted the inverse Gaussian distribution. The density of T is given by:

$$f(t, c, \mu, \sigma) = \frac{c}{\sigma\sqrt{2\pi}}\, t^{-3/2} \exp\left[-\frac{(c-\mu t)^2}{2\sigma^2 t}\right] \tag{1}$$

In this case, there exist several quasi-stationary distributions [3], one of which is "canonical" in the following sense: starting out in a single given state with probability 1, the distribution of "survivors" will converge to the canonical one. The canonical one is given by: $(\mu^2/\sigma^4)\, x \exp(-\mu\, x/\sigma^2)$, which is a **gamma distribution**, and is illustrated in Figure 3. If the process is initiated with this distribution on the positive real line, then the hazard rate of absorption in zero is constant and given as $(\mu/\sigma)^2/2$. Note that this hazard rate *depends on the square of* μ. This means, for instance, that if the drift towards the event (i.e. absorption) doubles, then the rate at which the event occurs is multiplied by four. In practical applications of the model, such considerations may be of interest in the interpretation.

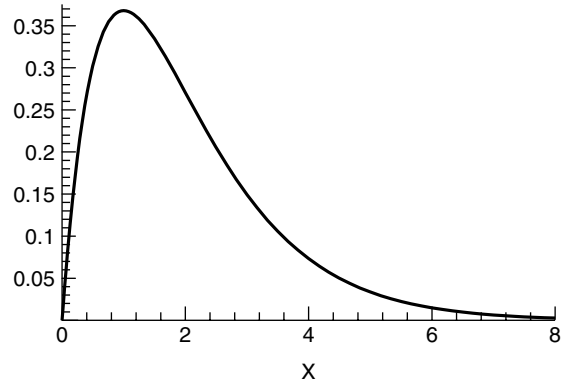

Figure 3 Quasi-stationary distribution for a Wiener process with absorption (parameters: $\mu/\sigma^2 = 1$). (Reprinted from Aalen and Gjessing 2001. Reproduced by permission of the authors)

The hazard rate of an inverse Gaussian distribution is unimodal, that is, increasing first and then decreasing. As pointed out by Aalen and Gjessing [2], when the starting point is close to the point of absorption compared to the quasi-stationary distribution, then the hazard rate is largely decreasing. When the starting point is far away, it is largely increasing. Finally, with a starting point in between, both the increasing and decreasing parts of the hazard rate will be of interest.

References

[1] Aalen, O.O. (1995). Phase type distributions in survival analysis, *Scandinavian Journal of Statistics* **22**, 447–463.

[2] Aalen, O.O. & Gjessing, H. (2001). Understanding the shape of the hazard rate, *Statistical Science* **16**, 1–22.

[3] Martinez, S. & San Martin, J. (1994). Quasi-stationary distributions for a Brownian motion with drift and associated limit laws, *Journal of Applied Probability* **31**, 911–920.

[4] Whitmore, G.A. (1986). First-passage-time models for duration data: regression structures and competing risks, *The Statistician* **35**, 207–219.

[5] Whitmore, G.A., Crowder, M.J. & Lawless, J.F. (1998). Failure inference from a marker process based on a bivariate Wiener model, *Lifetime Data Analysis* **4**, 229–251.

[6] Yau, C.L. & Huzurbazar, A.V. (2002). Analysis of censored and incomplete survival data using flowgraph models, *Statistics in Medicine* **21**, 3727–3743.

ODD O. AALEN

Hazard Plotting

Hazard plotting and **nonparametric** statistical inference for **hazard rate** (intensity) models have been vigorously studied in the mathematical framework of counting processes, as illustrated by articles, such as **Counting Process Methods in Survival Analysis**, **Nelson–Aalen Estimator**, **Kaplan–Meier Estimator**, **Aalen–Johansen Estimator**, **Repeated Events**, **Duration Dependence**, and **Goodness of Fit in Survival Analysis**. Common to these versions of the simple techniques for analysis of hazard rate models are some specific model assumptions that were not made by W. Nelson when he originally proposed what is here termed the *Nelson–Aalen estimator* of an integrated hazard.

As explained in detail in the article **Repeated Events**, the basic estimating equations (*see* **Estimating Functions**) for rate and mean functions are **unbiased** more generally than under the above model assumptions, provided that the observation intervals are independent of the process generating the events. (In particular, this means a more restrictive assumption on the **censoring** pattern.) Nelson described his approach in his monographs [1, 2].

References

[1] Nelson, W. (1982). *Applied Life Data Analysis*. John Wiley, New York.

[2] Nelson, W.B. (2002). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*. *ASA-SIAM Series on Statistics and Applied Probability*, SIAM, Philadelphia; ASA, Alexandria.

NIELS KEIDING

Hazard Rate

The hazard rate at time t of an event is the limit $\lambda(t) = \text{limit}_{\Delta \downarrow 0} \Delta^{-1} \Pr(t \preceq T \prec t + \Delta | t \preceq T)$, where T is the exact time to the event. Special cases and synonyms of hazard rate, depending on the event in question, include force of mortality (where the event is death), instantaneous incidence rate, **incidence rate**, and **incidence density** (where the event is disease occurrence).

For events that can only occur once, such as death or first occurrence of an illness, the probability that the event occurs in the interval $[0, t)$ is given by $1 - \exp(-\int_0^t \lambda(u)\,\mathrm{d}u)$ (*see* **Survival Analysis, Overview**; **Survival Distributions and Their Characteristics**). The quantity $\int_0^t \lambda(u)\,\mathrm{d}u$ is known as the **cumulative hazard**.

Often, the theoretical hazard rate $\lambda(u)$ is estimated by dividing the number of events that arise in a population in a short time interval by the corresponding **person-years at risk**. The various terms, hazard rate, force of mortality, incidence density, person–years incidence rate, and incidence rate are often used to denote estimates of the corresponding theoretical hazard rate.

MITCHELL H. GAIL

Hazard Ratio Estimator

In **survival analysis**, statistical models are frequently specified via the **hazard** function $\alpha(t)$. A simple model for the relation between the hazard functions in two groups (e.g. a treatment group 1 and a control group 0) is the **proportional hazards model** where

$$\alpha_1(t) = \theta\alpha_0(t), \tag{1}$$

with θ being the treatment effect. For a parametrically specified baseline hazard, $\alpha_0(t)$, both the treatment effect and the parameters in the baseline hazard are usually estimated using **maximum likelihood**. In a semiparametric model where the baseline hazard is left unspecified, several estimators for θ are available: the maximum **partial likelihood** ("Cox") estimator, cf. [5] (*see* **Cox Regression Model**), a class of **rank** estimators, and some *ad hoc* estimators.

Assume that the available data are

$$(X_{ij}, D_{ij}; \quad i = 1, \ldots, n_j, \quad j = 0, 1),$$

where the X_{ij} are the times of observation: a failure time if the corresponding indicator D_{ij} is 1, a right-censoring time if D_{ij} is 0. The Cox estimator, $\hat{\theta}$, is then the solution to the following equation:

$$O_1 = E_1(\theta), \tag{2}$$

where, for $j = 0, 1$, $O_j = \sum_i D_{ij}$ and

$$E_1(\theta) = \sum_{ij} \frac{Y_1(X_{ij})\theta}{Y_0(X_{ij}) + Y_1(X_{ij})\theta} D_{ij}.$$

Here, $Y_j(t) = \sum_i I(X_{ij} \geq t)$ is the number at risk at time $t-$ in group j, $j = 0, 1$. Notice that (2) expresses that for $\theta = \hat{\theta}$, the observed number, O_1, of failures in group 1 should be equal to a corresponding "expected" number, $E_1(\theta)$ under the proportional hazards assumption. The Cox estimator $\hat{\theta}$ is **consistent** and asymptotically normal under mild regularity conditions when n_0 and n_1 tend to infinity (*see* **Large-sample Theory**).

A class of explicit "rank" estimators originally introduced by Crowley et al. [6] and further discussed by Andersen [1] is for a given *weight process L(t)* given by

$$\widehat{\theta_L} = \frac{\sum_{i=1}^{n_1} L(X_{i1})(D_{i1}/Y_1(X_{i1}))}{\sum_{i=1}^{n_0} L(X_{i0})(D_{i0}/Y_0(X_{i0}))}. \tag{3}$$

For $L(t) = I(t \leq t^*)$, $\widehat{\theta_L}$ is simply the ratio between the **Nelson–Aalen estimators** for the cumulative hazards in groups 1 and 0 evaluated at t^*. Under the same kind of regularity conditions as for $\hat{\theta}$, the rank estimators given by (3) are also consistent and asymptotically normal if the weight process is well behaved in large samples. It was shown in the above-mentioned papers that the Cox estimator $\hat{\theta}$ given by (2) is always less dispersed than any $\widehat{\theta_L}$ given by (3). However, for the particular choice

$$L(t) = \frac{Y_0(t)Y_1(t)}{Y_0(t) + Y_1(t)}, \tag{4}$$

$\widehat{\theta_L}$ is nearly fully **efficient** when θ is close to 1. Furthermore, a fully efficient estimator is the two-step estimator of Begun & Reid [3] that is obtained with

$$L(t) = \frac{Y_0(t)Y_1(t)/\theta^*}{Y_0(t) + Y_1(t)\theta^*},$$

where θ^* is some preliminary, consistent estimator, e.g. $\widehat{\theta_{L=1}}$.

Using an estimator $\widehat{\theta_L}$ and its estimated variance (see, for example, Andersen et al. [2, Section V.3.1]) the hypothesis $\theta = 1$ of no treatment effect may be tested. This gives all the standard linear nonparametric two-sample tests for survival data (*see* **Linear Rank Tests in Survival Analysis**) and, in particular, the weight process given by (4) gives the **logrank test**.

Another explicit *ad hoc* estimator, discussed by Breslow [4], is given by

$$\tilde{\theta} = \frac{O_1/E_1(1)}{O_0/E_0(1)},$$

with

$$E_0(\theta) = \sum_{ij} \frac{Y_0(X_{ij})}{Y_0(X_{ij}) + Y_1(X_{ij})\theta} D_{ij}.$$

The estimator $\tilde{\theta}$ is generally inconsistent when $\theta \neq 1$ but it has gained some popularity due to its simplicity

and close connection to the logrank test, which is also based on the observed, O_0 and O_1, and expected, $E_0(1)$ and $E_1(1)$, numbers of failures.

Tests for the proportional hazards assumption (1) based on $\widehat{\theta_L}$ were studied by Gill & Schumacher [7] and further developed by Lin [9] and Sengupta et al. [10]; see, for example, Andersen et al. [2, Example VII.3.5].

The estimators discussed above only make sense under the proportional hazards model (1). However, Kalbfleisch & Prentice [8] defined, for a given survival function, S, the *average hazard ratio* by $\theta_1(S)/\theta_0(S)$ where, for $j = 0, 1$,

$$\theta_j(S) = -\int_0^\infty \frac{\alpha_j(t)}{\alpha_0(t) + \alpha_1(t)}\, \mathrm{d}S(t). \qquad (5)$$

Under the model (1), the average hazard ratio reduces to θ. Particular emphasis was paid to survival functions of the form $S(t) = (S_0(t)S_1(t))^\gamma$ where, for $j = 0, 1$, $S_j(t)$ is the survival function corresponding to the hazard function $\alpha_j(t)$. In this case (5) reduces to the following quantity:

$$-\int_0^\infty S_0(t)^\gamma\, \mathrm{d}S_1(t)^\gamma,$$

which, for a given value of γ, is easily estimated by replacing the survival function $S_j(t)$ by its **Kaplan–Meier estimator**.

References

[1] Andersen, P.K. (1983). Comparing survival distributions via hazard ratio estimates, *Scandinavian Journal of Statistics* **10**, 77–85.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[3] Begun, J.M. & Reid, N. (1983). Estimating the relative risk with censored data, *Journal of the American Statistical Association* **78**, 337–341.

[4] Breslow, N.E. (1975). Analysis of survival data under the proportional hazards model, *International Statistical Review* **43**, 45–58.

[5] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[6] Crowley, J.J., Liu, P.Y. & Voelkel, J.G. (1982). Estimation of the ratio of hazard functions, in *Lecture Notes – Monograph Series 2, Survival Analysis*, J.J. Crowley & R.A. Johnson, eds. Institute of Mathematical Statistics, Hayward, pp. 56–73.

[7] Gill, R.D. & Schumacher, M. (1987). A simple test of the proportional hazards assumption, *Biometrika* **74**, 289–300.

[8] Kalbfleisch, J.D. & Prentice, R.L. (1981). Estimation of the average hazard ratio, *Biometrika* **68**, 105–112.

[9] Lin, D.Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, *Journal of the American Statistical Association* **86**, 725–728.

[10] Sengupta, D., Bhattacharjee, A. & Rajeev, B. (1998). Testing the proportionality of hazards in two groups against the increasing cumulative hazard ratio alternative, *Scandinavian Journal of Statistics* **25**, 637–647.

(*See also* **Survival Distributions and Their Characteristics**)

PER KRAGH ANDERSEN

Health Care Financing

The methods used to finance personal health care service play a major role in shaping a country's health care system. Personal health care services are those services such as hospital care, physician care, dental services, and drugs that are provided directly to individuals. The financing methods influence the terms under which people access the health care delivery system, the types of health care provided, and the mechanisms used to allocate health care services. They also influence how the costs of health care services are distributed over the population by income and by health status.

Two aspects of health care financing are the focus of this article: the sources of funds for health care services and the mechanisms used to pay individuals and institutions who provide health care services. Figure 1 presents a diagrammatic representation of these two aspects of health care financing and provides a framework for the discussion that follows.

Sources of Payments for Health Care

Overview

Individuals may simply use their own incomes to purchase health care services from health care providers (physicians, hospitals, clinics, laboratories, and other firms/individuals). In most other markets, this is the way goods and services are purchased. However, the market for health care services has evolved quite differently. In this market, other mechanisms such as private insurance plans, sickness funds, and national health insurance systems have been developed to pay for a significant proportion of personal health care services.

Nevertheless, as indicated in Figure 1, all health care is eventually paid for by individuals. It is individuals who ultimately pay the premiums to insurance companies and the taxes to governments, which in turn pay for health care services. Furthermore, even though business may be a source for collecting premiums and/or taxes, the amounts paid by businesses are passed back to individuals and households through lower wages, higher prices, or lower returns on invested capital. The total payments made for health care are sometimes referred to as the *cost* of personal health care.

The way that the funds are raised to pay for health care services affects the distribution of the cost of health care within a country. Two types of distributions of the cost of health care are frequently considered: the distribution by health status (across the healthy and the sick), and the distribution by income. Analysts classify funding sources with respect to income as progressive, regressive, or proportional. In a progressive funding system, the fraction of a person's income paid in taxes (or premiums) rises as their income rises; in a regressive funding system, the fraction of a person's income paid in taxes (or premiums) declines as their income rises; while under a proportion funding system, this fraction is constant regardless of a person's income. This section will discuss the sources of funds in some detail and comment on the distribution of the cost of health care.

Payments by Third Parties

A basic characteristic of health care systems in all developed countries is that the majority of payments for medical services flows through third parties. A third party is an entity, usually an insurance company or government agency, that pays for medical services but does not receive or provide health care services. This payer is the third party, while the patient and the health care provider are the first two parties. This distinction between the third party and the providers is becoming blurred, particularly in the United States.

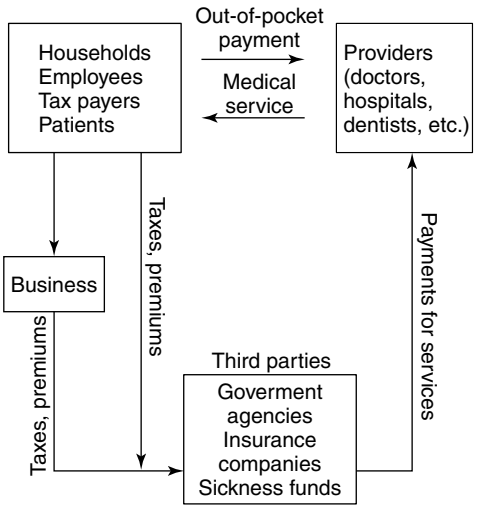

Figure 1 Health care financing. Source: adapted from [8]

In the US, groups of health care providers may assume some financial risk (thus acting as insurers) by contracting with governments, businesses, and/or individuals to provide medical care at a fixed rate per person covered. In general, third-party financing arose for two reasons: individuals wanted to insure against the large and uncertain financial costs associated with illness, and governments wanted to ensure that the population at large or certain vulnerable portions of the population had access to needed health care.

As indicated in Figure 1, there are three basic sources of third-party funds: general taxes, payroll taxes, and insurance premiums. General taxes include income taxes, value-added or consumption taxes, and other specific taxes. Payroll taxes are employment-related taxes, which are normally set as a percentage of payroll or wages. (Payroll taxes may apply to all wages, or to wages up to a certain amount.) (Additionally, governments may use other sources of revenue such as that from lotteries). Insurance premiums are the amount paid for an insurance policy – premiums vary according to the type and amount of insurance purchased as well as the characteristics of the individuals covered under the policy. Premiums are the price paid for an insurance policy. (An individual's policy may cover the individual, or the individual and other persons dependent on that individual.) Payments made by the insurance company for covered medical services used by individuals enrolled in the insurance plan are funded by the premiums. If premiums are the sole funding source for the plan, then premiums must reflect the expected cost of health care services used by the individuals covered under that plan as well as the administrative cost of running the plan.

There are two different approaches to setting the health insurance premiums: community rating and experience rating. Community rating systems use the average, expected cost of medical care for all individuals in a community and assign this premium to each individual in the community. An experience rating system groups individuals by some common characteristic (i.e. place of employment, age, whether or not they smoke) and assigns individuals of like characteristics the same premium, which is based on the average, expected cost of medical care for individuals with those characteristics (*see* **Actuarial Methods**).

Major Sources of Third-party Funds in Selected Countries

Payroll taxes are the major source of third-party funds in France and Germany. In both of these countries, employees and employers pay a certain percentage of their wages into sickness funds. In Germany, there are several sickness funds, and employees often have a wide choice of sickness funds in a region. Once an employee selects a fund, the fund has the right to collect its premiums (which vary from fund to fund) as a percent of the employee's gross wages – half the premium is paid by the employer and half by the employee. There is a limit on the amount of wages subject to premiums. In France, there is one large sickness fund (which covers about 80% of the population) and several small funds. The contributions paid by employers and employees are set by the Central Government: the employer pays about 12.6% of total wage bill and the employee pays about 6.8% of wages. In both of these countries, systems have been developed to cover individuals who are unemployed and those who are retired.

General taxes are the major source of third-party funds in the United Kingdom and Canada. The British National Health Service is a national program. The Canadian Medicare program is a decentralized program and the cost of the program is shared between the federal and provincial governments. Each Canadian province administers its own program and exercises some discretion regarding which medical services are covered. Third-party financing in the United States is fundamentally different from that in all other developed countries because of the large mix of funding methods used. Multiple types of third parties exist, including government programs such as Medicare and Medicaid, nonprofit insurance companies such as Blue Cross/Blue Shield, and numerous private insurance plans vended to employers, unions, and individuals. This variety of third-party payers results in a mix of sources of third-party funds, including premiums by individuals and businesses, general taxes, and payroll taxes.

In the United States, the majority of employed people and their dependants obtain health insurance through their employment. However, the provision of health insurance by employers is strictly voluntary (other than the State of Hawaii, which mandates that employers provide health insurance). The fact

that many employers voluntarily offer health insurance reflects the strong incentives in the current tax system for employers to provide health insurance. Under current tax codes, employer-paid premiums are considered a cost of doing business and are treated as a business expense. (Large companies frequently self-insure and hire administrators to manage their health benefits. Thus, these companies do not strictly pay premiums. Nevertheless, conceptually, one can think of expected, average health care costs as the premium.) Furthermore, these employer-paid premiums are not considered income for employees and are exempt from individuals' income taxes. Wide differences exist across firms with respect to the design of the benefit package and the proportion of the premium paid by the employer. (employers generally pay a higher proportion of the total premium for their employees than for their employees' dependants who might also be included in the health plan.). Employers sometimes offer employees a choice among health insurance plans. Active competition exists among insurance companies in the employment-offered health insurance market, and premiums are set competitively. Premiums tend to be experience rated rather than community rated, and businesses with healthy employees have lower premiums than businesses with generally sicker employees. In the United States, health insurance is also available to individuals and groups, independent of employment. Premiums are experience rated and reflect the average, expected cost of illness for individuals within the defined group covered by a particular policy. The sicker the pool of individuals covered by an insurance policy, the higher the premiums. However, a number of states have passed legislation that set limits on the range of premia.

Although the United States lacks universal health insurance, it provides public health insurance for the poor, the disabled, and the elderly through two publicly funded programs, Medicaid and Medicare. Medicaid, which is jointly funded by the Federal and State governments, provides medical services to low-income individuals who meet specific eligibility criteria. Eligibility criteria may vary slightly across states. In general, the Medicaid enrollees are medically disabled and/or poor. Medicaid contains specific benefits for low-income women and their children, and the majority of those covered by Medicaid are women and children. (The majority of payments, however, go for the disabled and the elderly to pay for long-term care services not covered under the Medicare program.) The federally funded Medicare program insures most people aged 65 and over, as well as a small subset of the general population who are medically disabled. The Medicare program is funded through a combination of payroll taxes, general tax revenues, beneficiary premiums, and direct beneficiary payments (*see* **Health Services Organization in the US**).

It should be pointed out that health insurance markets have developed in many countries to supplement government health insurance programs. For example, people in the United Kingdom may purchase private health insurance as an additional source of third-party funding that facilitates access to specialists and allows individuals to jump queues and to use private facilities not covered by government programs. Canadians may purchase supplemental health insurance policies to pay for medical services not covered under the provincial Medicare programs and to pay for nonessential amenities such as private rooms. Likewise, American Medicare beneficiaries may purchase supplemental insurance (called Medigap insurance) to cover expenses not included in the Medicare program, such as outpatient prescription drugs and extended nursing home care, as well as Medicare cost-sharing liabilities.

Direct Payments by Individuals

Individuals may also pay health care providers directly for services rendered. Direct payments to providers are often referred to as *out-of-pocket payments*. People may pay directly for health care services for several reasons: they are uninsured, or a particular service is not covered by their health insurance plan, or the health insurance coverage is not complete. For example, many insured individuals routinely make some out-of-pocket payments because their health insurance policies (including the national health insurance programs discussed above) contain explicit *cost-sharing* provisions mandating some amount of direct payment by the individual. Cost-sharing provisions take several different forms including: deductibles (a fixed amount that must be paid by the insured before any third party payment will be made); copayments such as a specific payment that the insured must pay for each service (e.g. $6.00 for each prescription filled or $5.00 for each visit to a physician); or a specific percentage of the bill (such

as the 20% copayment Medicare beneficiaries must pay for physician services). Wide variations exist across countries with respect to the level of out-of-pocket payments for health care. For example, in both Canada and in the United Kingdom there are no out-of-pocket payments for basic services covered under the respective national health insurance systems. In Germany, out-of-pocket payments are very limited, whereas France has substantial cost sharing (i.e. 20%–30%) for many health care services. In general, out-of-pocket payments in the United States are high relative to those in other developed countries for a number of reasons: a significant number of people are uninsured (no third-party payer), and private health insurance plans are not standardized with respect to the types of services covered or the level of cost-sharing provisions on covered services. Note that although the Medicare program does contain significant cost-sharing requirements, few Medicare beneficiaries actually pay much out-of-pocket for Medicare covered services. Indigent Medicare beneficiaries have their cost sharing covered under the Medicaid program, while the majority of other Medicare beneficiaries have their cost-sharing liabilities covered through employer sponsored retiree health benefits or privately purchased supplemental (Medigap) policies.

Uninsured Individuals

Among the developed countries, only the United States has a large number of people without insurance because it is the only country without a universal health insurance program. Approximately 17% of the United States nonelderly population was uninsured in 2002 [2]. The uninsured pay for services out of their own pocket. However, a large portion of the cost of medical care used by the uninsured is covered by publicly funded clinics, specifically designated charity funds, and "cost-shifting". Cost-shifting occurs when a provider charges some groups of patients higher than normal fees in order to cover the cost of services provided to other groups of patients for whom the provider receives no or inadequate payments.

Implication of Funding Sources on the Burden of Illness

In general, health insurance programs funded predominately by income taxes are the most progressive with respect to income. Individuals with the most income pay proportionately more than individuals with the lowest incomes, regardless of their use of medical services. Health insurance programs funded by premiums are the most regressive with respect to income. Payroll taxes tend to be moderately regressive because the percentage of a person's income from wages tends to decline as total income rises. Value-added or consumption taxes are also regressive with respect to income because low-income individuals tend to spend a higher proportion of their income on consumption goods than do upper-income individuals. However, the regressive nature of health insurance programs financed by consumption taxes can be reduced by excluding specific items (such as food) from taxation.

The broader and more inclusive the funding base for health insurance programs, the more the financial cost of illness is shifted from the sick to the healthy. The more out-of-pocket payments serve as the source of funds for health care, the more the burden of the financial cost of illness is borne by the sick. The more experience rating is used to set premiums, the more the relatively sick have to pay for health care. Thus, under national health insurance systems, such as those in the United Kingdom, Germany, and Canada, the *financial* cost of illness is borne socially and distributed broadly across the population. Actual expenditures (in out-of-pocket payments and premiums) made by the sick are much less than that of the total cost of their medical care. This is in contrast to the United States where a higher proportion of costs are paid out of pocket and a large number of insurance plans are experience rated.

Third-party Payments and Consumer Demand

Health insurance affects the price that a person pays for medical service, but the effect can be complicated. For example, a person who has complete insurance coverage pays a price of zero for each medical service used. A person with an insurance plan specifying a $500 deductible, 20% cost sharing, and a $3000 limit on out-of-pocket payments, pays out of pocket for the first $500 of medical services used and then pays 20% of charges for additional medical services until a total of $13 000 in medical services has been used. At that point, the person will have spent $3000 in total out-of-pocket expenses and the insurance policy will cover all additional costs. Thus, individuals covered

under the former plan may choose to use medical services differently than those covered under the second plan since they face very different prices for the same unit of medical care.

Not surprisingly, there is considerable interest in how the structure of cost sharing influences the number of health care services used, that is, in determining how responsive consumers are to out-of-pocket payments for medical services. Since the United States is the only country where direct patient payments are a significant source of funds, most research on this issue has been done in the United States using that country's data. Several different analytical structures have been employed to examine the use of services (*see* **Health Care Utilization Data**) by individuals who pay different prices for the same service including the examination of natural experiments (e.g. some exogenous event changes the price that an individual pays for health care), analyses of self-reported utilization data from large-scale surveys such as the National Medical Care Expenditure Survey (*see* **Medical Expenditure Panel Survey (MEPS)**), and analyses of claims data from different health insurance plans [7]. The need to control for nonprice factors that influence the use of health care services, as well as **selection bias** complicates this research. (Selection bias arises when people who need more services choose plans with more extensive coverage, and vice versa.) The best study examining these issues uses data from the Health Insurance Experiment, a randomized clinical trial (*see* **Clinical Trials, Overview**) conducted by the RAND corporation [6]. This experiment, conducted in the 1970s, **randomized** people from six different locations (both urban and rural) in the United States into one of 14 insurance plans, which differed by the amount of the deductible, consumer cost sharing, and maximum out-of-pocket expenses.

The research revealed that consumers generally respond to the price of health care services, and the extent to which they respond varies by the type of medical service. For example, the price of emergency services produces very little response, while people are more highly responsive to the price for elective services. The demand for mental health care services is more price responsive than that for physical health care services.

As argued above, the presence of health insurance affects an individual's behavior, relative to what it would be if he or she were not insured. This effect is sometimes referred to as *moral hazard*. Moral hazard is said to occur when insured individuals change their behavior because they have insurance. The term was first applied in the life and fire insurance markets. For instance, moral hazard is said to exist if a person burns down the house in order to collect fire insurance, or fakes death with the intent of collecting the life insurance payments. The use of the term in the life and fire insurance markets implies some immoral behavior on the part of the insured.

In the health insurance market, *moral hazard* refers to insured individuals engaging in riskier (in terms of their health) behavior because of the presence of health insurance or otherwise changing their health-related behavior, such as using more health care services or failing to seek out low-cost providers. However, most commonly the term *moral hazard* simply reflects the basic law of demand: as the out-of-pocket price of medical services decreases, people use more medical services. In this sense, moral hazard and price responsiveness (known as price elasticity of demand) are intimately related. Thus, there is nothing immoral about this type of moral hazard; it is simply a manifestation of rational human choice.

Funding Sources in the United States

As indicated above, the financing of health care services in the United States is more complex than in other countries. Table 1 presents information on the sources of funds for personal health care services in the United States; both total service and by the type of service. In 2002, expenditures on personal health care services were $1340.2 billion, of which 15.9 was paid by direct patient payments (out-of-pocket payments) and 84.1% by third parties. However, the proportion of expenditures covered by third parties ranges from 97% for hospital care to 44% for drugs and other services.

Data Sources

Information on health care expenditures is reported by the government agencies for each country. For example, the Center for Medicare and Medicare Services website has detailed information on program expenditures as well as national health care expenditures. Furthermore, both the *Health Care Financing Review and Health Affairs* annually publish detailed information on health care expenditures in the United

Table 1 Source of payment for selected personal health care 2002. Total: 1340.2 billion (US: 2002)

Type of service	Total (%)	Patient direct	Private health insurance	Other private	Federal[a]	State and local[b]
Hospital care	100	3.0	33.9	4.2	47.3	11.6
Physician care	100	10.1	49.1	6.9	27.9	5.9
Dentist services	100	44.0	49.5	0.1	3.8	2.6
Prescription drugs	100	29.9	47.8	–	12.9	9.5
Nursing homes	100	25.1	7.5	3.4	44.1	19.9
All personal health care services	100	15.9	35.8	4.2	33.6	10.6

[a]Includes medicaid SCHIP expansion and SCHIP.
[b]Subset of federal funds.

States. The Department of National Health and Welfare in Canada publishes data on Canadian health expenditures. The Organization for Economic Cooperation and Development (OECD) maintains an ongoing data collection and analyzes effort aimed at producing timely, consistent data for 24 nations in Asia, Europe, and North America. The journal *Health Affairs* periodically publishes data on the performance of health systems in OECD countries.

Paying Health Care Providers

A number of methods exist for paying health care providers (physicians, hospitals, clinics, labs, and other individuals/firms supplying health care services) for medical care services rendered to individuals. This section presents an overview of the most important of these payments methods.

Paying for Physician Services

Physicians are generally paid using one of three general methods of payment: fee-for-service, capitation, or salary. In some cases, physicians receive payments under more than one of these payment methods. The use of multiple payment methods occurs when either a given payer uses a combination of methods, or, as occurs in the United States, a physician receives payments from more than one third-party payer, each of which uses a different payment method.

Fee-for-service. Under the fee-for-service method of payment, physicians receive a fee for each service provided. The medical service rendered is the *unit of payment*, and there is a certain degree of discretion regarding how a service is defined. A service unit can be very distinct (e.g. a urinalysis) or relatively comprehensive (e.g. an appendectomy where the physician payment covers all care associated with the procedure, including the preoperative visit, the surgical procedure itself, and some follow-up care). Thus, the service on which the unit of payment is based can actually be some bundle of separate, discrete services.

Payments to physicians for medical services may be based on the fees that physicians set for their services or on a specific fee schedule. A fee schedule defines the amount or relative amount of fees for each physician service. In general, only third-party payers use fee schedules. Individuals without insurance for physicians' services are usually billed according to charges set by the physician.

In the United States, third-party payers using the fee-for-service method may pay physicians an amount based on the physician's charges, prenegotiated rates, or a fee schedule. Because different third-party payers may use different rates or schedules, physicians can receive different payment amounts for the same type of service depending upon the third party payer involved. By contrast, in most other countries using the fee-for-service method, physicians receive payment based on a single negotiated fee schedule or on regional negotiated fee schedules. The best-known fee schedule in the United States is the *Medicare fee schedule*, the fee schedule used by the Medicare program to pay physicians for services rendered to Medicare beneficiaries. The Medicare fee schedule assigns each defined unit of service a relative value quantifying the resources (such as physician time, skill, and use of support services) needed to produce the service. The Medicare fee schedule employs a conversion factor to translate the value of the resources used into a specific payment amount.

In addition to the Medicare program, some of the other third-party payers in the United States have adopted the Medicare fee schedule for use with their own resource conversion factors to set payments for physician services.

Capitation. The capitation method of payment provides physicians with a defined, periodic, per patient payment (usually monthly), regardless of the number or type of covered services the physician provides to a patient. Most commonly used to pay primary care physicians, the periodic payment reflects the expected cost of providing the covered services. The covered services and terms of the care provided under capitation vary with the actual capitation agreement. When used to pay primary care physicians, some subspecialty services provided to patients by other physicians may be charged to the primary care physician for payment out of the primary care physician's capitated fee. Likewise, some specified services provided by the primary care physician may not be included in the capitated fee and instead may be paid for on a fee-for-service basis. Again, these arrangements vary by the actual capitation agreement.

The capitation fee may be adjusted to reflect patient characteristics such as age in order to compensate physicians for variations in the expected use of services by groups of patients with similar characteristics. In the United States, managed care plans use the capitation method widely to pay primary care physicians, as does the British National Health Service.

Salary. The salary method of payment provides physicians with a fixed monthly or annual salary that does not vary with the number of patients treated or services provided. However, not all physicians are paid the same salary, which may be based on such factors as specialty, hours worked, special duties (such as administrative tasks), and years of experience. In many European countries, hospital-based physicians are paid using the salary method, while physicians working in the outpatient setting receive payment under other methods. In the United States, physicians working for government agencies, some Health Maintenance Organizations, or large group practices, often receive payment by the salary method.

It should be noted that a physician can receive payment under a single payment method, while third-party payers make payments for the physician's services using several different payment methods. For example, a physician belonging to a large group practice may receive a salary even though insurance plans pay for services rendered by physicians in the group via a capitation method.

Paying Other Professionals. For other health care professionals (physical therapists, dieticians, social workers, home care nurses, etc.), the fee-for-service and salary methods are widespread, while capitation is rarely used.

Paying Hospitals

Numerous methods are used to pay for hospital services, such as payment based on established charges, retrospective costs, per diem rates, per case rates, capitated payments, or budgets. Because there are many different third-party payers in the United States, hospitals located there frequently receive payments under a host of different methods. In contrast, hospitals in other countries tend to be paid according to a single payment method.

Charge-based Payments Method. Prevalent only in the United States, the charge-based method requires hospitals to define a price or "charge" for each service the hospital provides. This hospital-established charge (or a negotiated percent of that charge) for each service is then paid either directly by the patient or by the patient's health insurance company. If the insurance policy requires copayments, then the hospital's charge is split between the patient and the health insurance company according to the conditions of the insurance contract. The charge-based method allows the hospital to determine the price of hospital services. This method is not used by government payers.

Retrospective-cost-based Payment Method. The retrospective cost-based payment method pays hospitals on the basis of the actual costs of providing hospital services as opposed to a hospital set charge (which may not be linked to the cost of providing services). Under this method, a set of accounting rules allocates hospital costs to a group of patients. Although relatively common in the United States

from 1966 to 1983 because it was used by the Medicare program, most state Medicaid programs, and several Blue Cross plans, this method has lost importance since the mid-1980s when Medicare introduced the Prospective Payment System. When this method is used, hospital payments are typically subject to limitations – either limits on the extent to which reimbursable costs can rise from year to year and/or limits on the maximum allowable costs. Limitations on the maximum allowable costs are normally set relative to costs reported by other similar hospitals.

Per Diem Payment Method. The per diem payment method pays hospitals a set amount for each day that a patient spends in the hospital. In general, the per diem rate is independent of patient characteristics, (e.g. the same per diem rate is paid for patients undergoing heart surgery as for maternity cases). However, the per diem rate may vary by hospital. The rate is generally set via negotiations between the third-party payers and the hospital. The per diem method is relatively common in Europe. In Canada, provincial governments use the per diem method to pay hospitals located outside the province for hospital care rendered to residents of the province. These transfer payments represent only a small proportion of hospitals' budgets.

Per Case Payment Method. The per case method pays a hospital a set amount for each patient discharged from the hospital. In the most extreme form of the per case method, hospitals receive a defined amount per discharge irrespective of the patient's condition. More commonly, patients are classified into groups on the basis of the expected costs for necessary care (known as **case mix** formulations). Using a cost weight established for each group, the hospital receives a payment related to the patient's group classification. A number of patient classification systems exist, but the most frequently used systems are based on the **diagnosis related groups** (DRGs) developed at Yale University [3]; see [4, 5] for an overview of case-mix classification issues.

The *Medicare Prospective Payment System* is the best known of the case-mix-based per case payment systems. This system classifies patients into one of approximately 540 DRGs. A cost weight assigned to each group reflects the expected relative cost of treating patients within that group. For each patient discharged, the hospital receives a set payment that varies by the DRG assigned to the patient. This particular per case payment system also contains provisions for additional payments for patients whose treatment cost are exceptionally high (referred to as outlier payments) and adjustments for the higher costs of teaching hospitals.

Capitation Payment Method. The capitation payment method pays hospitals a fixed, periodic fee per patient for a defined group of patients, often referred to as a panel of patients. The capitation payment does not vary with the actual use of services. Thus, even if no hospital services are used by any patient in the hospital's panel of patients in a given period, the hospital still receives payment. Unexpectedly, high use of hospital services by the patients in the panel can result in net hospital losses for the period. This payment method shifts financial risk from the third-party payer to the hospital itself and its use is relatively rare.

Budget Payment Method. The budget payment method provides hospitals with a global budget designed to cover all services provided by the hospital over the course of the year. The global budget may be unilaterally set by some government agency; it may be established according to some generally accepted formulas, which account for inflation and changes in the size of the inpatient population; or it may be negotiated between the payer and the hospital. In some countries, global budgets involve the use of case-mix information. For example, in the Canadian provinces of Ontario and Alberta, the provincial governments use case-mix information to identify hospitals with global budgets, which may be over- or underfunded relative to other hospitals serving similar patients. Like a capitation system, a global budget system shifts financial risk from the third party to the hospital system. However, it differs from the capitation system because it is not so closely related to the number of covered lives.

Paying Other Institutional Providers. The same methods that have been developed to pay for hospitals are used to pay other institutional providers.

Data Sources

Government agencies publish reports that include the detailed specifications of their health care payments.

For example, the rules for both the Medicare Hospital Prospective Payment System and the Medicare Physician Payment system are published annually in the *Federal Register*. The former includes a listing of all DRGs, the associated costs, the relative cost weights, and other information needed to transform DRG relative costs into payment rates. The latter includes a list of physician services, associated codes, associated relative values, and the conversion factor. The Ontario Ministry of Health publishes the *Schedule of Benefits* for physician services under the Health Insurance Act. This includes the listing of services, the associated codes, and the payment amounts.

Incentives Embedded in the Payment Methods

Theoretical Effects of Provider Payment Mechanisms

The effects of these different methods of paying providers have been the focus of a large body of research (*see* **Health Services Research, Overview**). The basic approach to assessing providers' response to methods of payment is to identify those actions that either increase the providers' profits or decrease their losses. This can be done through formal theoretical modeling [1] as well as through a thoughtful consideration of the issues.

In general, analysts believe that the fee-for-service payment method creates incentives for providers to increase the number of services provided. Furthermore, the fee-for-service payment method lacks incentives for physicians to combine services in a way that minimizes the total cost of treating a patient to obtain a specific outcome. This effect, combined with a physician's desire to deliver thorough and comprehensive care, can result in too many services being provided. The capitation payment method eliminates the incentive in the fee-for-service method to increase the number of services rendered. Instead, the capitation method creates strong incentives for physicians to manage a patient's care efficiently – at least with respect to the services covered under the capitation fee. However, the capitated payment method may result in under-treatment of patients when physicians are not involved in the long-term planning of patient care. Furthermore, there may be some incentives for physicians actively to seek out or recruit relatively healthy (i.e. less costly) patients and to discourage relatively sick (i.e. more costly) patients from joining or remaining in the physician's panel of patients. This is possible because physicians can influence the nature of the interaction that they have with patients.

The salary payment method significantly reduces physician incentives to provide either too many or too few services. However, this method lacks any incentives for physicians to manage patient care efficiently. Furthermore, while the salary payment method removes the incentive for excessive use of medical services, physicians may respond by decreasing their work output. Thus, this method may necessitate productivity enhancement and monitoring measures to ensure an adequate level of work effort on the part of physicians.

Analysts examine the same factors when they consider the incentive effects embedded in the different methods used to pay hospitals. Under a charge-based payment system or a retrospective-cost-based payment system (when most of the payments are made by third parties), there are few financial incentives for hospital administrators to decrease costs or to develop systems that encourage physicians to manage care efficiently. Under a per diem payment method, financial incentives exist to manage the daily costs of hospital care but not the number of days. Therefore, as long as the per diem payment rate is higher than the marginal daily costs, incentives exist to increase the length of the hospital stay. The per case payment method creates strong incentives to manage the use of inpatient services efficiently but also creates incentives to shorten hospital stays. Hospitals may achieve shorter stays by transferring patients to other facilities. The per case payment method also contains financial incentives for hospital decision-makers to undertreat patients, to discriminate against relatively sick patients, and to encourage actively the admission of relatively healthy patients.

Empirical Research on Supply Response

Research on the supply response to payment methods has been concentrated in the United States because of the variety of payment schemes in effect there, the extensive changes in the level and structure of payments that have been made by third parties, and the accessibility of electronic databases (*see* **Administrative Databases**) suitable for testing hypotheses about supply response.

Different analytical approaches analyzing provider response to employer changes have been used. Two approaches are commonly used are before–after studies comparing the outcomes for a common set of providers before and after some specific change in payment method and **fixed effect** models analyzing the effect of payment method using categorical variables to characterize the payment method. Analysts also use a difference in the differences approach in which they analyze the relative differences across two panels of medical providers (or patients) over time where one panel has experienced a change in payment methods and the other panel has not. In general, empirical results are consistent with the incentive effects as discussed above (see [9] for a review of physician supply response).

Databases

As noted, research on supplier response has been facilitated by the existence of large, electronically available databases. For example, the Medicare program's administrative records include detailed information on all payments made under the traditional Medicare program. In the Medicare system, an electronic claim record is created for each service provided to a Medicare beneficiary by a physician or other individual medical care provider. A comprehensive electronic claim record is also created for each hospital admission for a Medicare beneficiary. Each provider of medical services (including hospitals and clinics) and each Medicare beneficiary has a unique identifier. Therefore, it is possible to develop records of episodes of care for beneficiaries and to assess the effect of payment changes for hospital care on length of stay, hospital transfers, the characteristics of hospital patients, and the use of non-hospital services. Detailed data linking providers and recipients (*see* **Record Linkage**) are also available for some state Medicaid programs and some private insurance companies and health plans. Additionally, some states require hospitals to report detailed diagnostic and charge information on all their discharges or mandate the collection of limited clinical data on all patients. These data are available from the relevant state agencies. Finally, the American Hospital Association (AHA) surveys all hospitals in the United States about the number of beds, the number of admissions, and costs. The results are reported in the *AHA Annual Guide to the Health Care Industry* as well as electronically.

Macroeconomic Concerns: A Comment

In all countries, the health care system is shaped by the general regulatory environment within which consumers make decisions about accessing the health care system, and providers make decisions about the types of treatments to provide or recommend. There are significant differences across countries with respect to the extent of centralized controls over the number and location of hospital beds, the number and specialties of physicians in training, physician licensing, practice location and mobility, and the ability of hospitals or groups of providers to establish clinics or purchase advanced technology. In addition, all health plans (government and private) define the types of services they will cover, the relative frequency with which some services (e.g. preventive services) will be paid for, and the conditions under which patients can seek specialty care. Under some plans, patients are allowed to self-refer to specialists; in others, they must obtain permission from a primary care physician to visit a specialist.

In addition to the regulatory controls, the level of control that government authorities have over aggregate health care budgets varies. In general, the more a health care system is directly budgeted, the more governmental control there is over the size of the health care system (subject of course to the give and play of the political environment). In those cases where governments or regulatory authorities control the prices of care (per service, per day, or per case), providers can influence the outcomes by altering the mix of services or the volume of care. However, it is possible to impose budgetary control in a system where prices are directly controlled. For example, the United States has imposed physician expenditure targets called Volume Performance Standards for physicians under the Medicare program. The conversion factor applied to determine the Medicare fee schedule is a function of how well physicians in the aggregate meet the volume performance standard. In the province of Ontario, the government sets income limits for individual physicians. As payments to individual physicians reach the limit, the proportion of the fee paid decreases. In general, there is much less aggregate control over the health care delivery system in the United States than there is in other countries.

The financing methods and the type of aggregate controls imposed on a country do not appear to have a major independent effect on the size of the health care sector relative to the aggregate economy. Health care expenditures per capita are highly correlated with gross domestic product per capita. In fact, a study of health care costs in countries that were members of the Organization for Economic Cooperation and Development (OECD) found that $R2$ (*see* **Correlation**) was 0.93 for a simple model in which the log of a country's per capita expenditures on health care was regressed against the log of the country's per capita domestic product [7].

References

[1] Ellis, R.P. & McGuire, T.G. (1986). Provider behavior under prospective reimbursement: cost sharing and supply, *Journal of Health Economics* **5**, 107–193.

[2] Fronstein P. (2003). *Sources of Health Insurance and Characteristics of the Uninsured, Analysis of the March 2003 Current Population Survey*. EBRI Issue Brief 264. Employee Benefit Research Institute, Washington, December 2003.

[3] Health Systems Management Group. (1982). *The New ICD-9-CM Diagnosis-Related Group Classification Scheme*, Final Report. Yale School of Organization and Management. New Haven, Connecticut.

[4] Hornbrook, M.C. (1982a). Hospital case mix: its definition, measurement, and use: Part I. The conceptual framework, *Medical Care Review* **39**, 1–43.

[5] Hornbrook, M.C. (1982b). Hospital case mix: its definition, measurement, and use: Part II. Review of alternative measures, *Medical Care Review* **39**, 75–123.

[6] Newhouse, J.P. (1993). *Free for All? Lessons from the RAND Health Insurance Experiment*. Harvard University Press, Cambridge.

[7] Phelps, C.E. (1992). *Health Economics*. Harper Collins, New York.

[8] Reinhardt, U.E. (1993). An "all-American" health reform proposal, *Journal of American Health Policy* **13**, 11–17.

[9] Rice, T. (1997). Physician payment policies: impacts and implications, in *Annual Review of Public Health*, J.E. Fielding, L.B. Lave & B. Starfield, eds. Annual Reviews Inc., Palo Alto, Vol. 18, 549–565.

JUDITH R. LAVE & PAMELA B. PEELE

Health Care Technology Assessment

What is HTA?

Health technology assessment (HTA) is the evaluation of the properties, effectiveness, and the direct and indirect impacts of health technologies. A health technology refers mainly to interventions or methods used to affect the health of an individual or populations. Thus, it includes health promotion, health care interventions to treat and rehabilitate (including drugs, devices, and procedures), and the systems for the support and delivery of care such as telemedicine and patient records. Health technology assessment (sometimes referred to as *Health Care* or *Health Service Technology Assessment*) can be considered a major component of **Health Services Research** (HSR), which is also concerned with broader issues like the financing, organization, and delivery of care. HTA is carried out in order to find out whether a technology works, for whom, at what cost and with what other intended or unintended consequences for the individuals, their families, the health service, and society in general. Therefore, it is best conducted by teams incorporating disciplines, such as statistics, epidemiology, economics, psychology, and sociology. The results are used to inform health policy and practice at a national or a local level.

Reasons for HTA

Over the last 25 or so years, health care spending in developed countries has been rising in real terms and as a percentage of gross domestic product. Much of this increased spending has been on new interventions for the prevention or management of disease. Therefore, health care funders are particular keen to ensure that this investment is worthwhile and delivers a sufficient return in the form of improved health outcomes. In addition to the cost implications of health technologies, several of them such as *in vitro* fertilization, genetic testing, organ transplantation, and life sustaining technologies raise important ethical and social concerns.

Research has also shown that there are significant geographical variations in the patterns of practice even within the same health care system [19] not explainable by variations in the underlying frequency of disease. Health care professionals and experts are not always reliable sources of information about the effects of a health technology. Clinical practice, therefore, rather than being determined purely by professional views, should be more firmly based on the research evidence. This has led to the call for "**Evidence-based Medicine**" and the standardization of health care informed by the research evidence base. HTA is the engine that provides much of the data for evidence-based health care and the evidence base for the development of clinical practice guidelines.

HTA grew out of general technology assessment that emerged in the United States in the mid-1960s in response to concern about the consequences of the rise of technology in modern life. In 1972, the US Congress established an Office for Technology Assessment (OTA) followed by the National Center for Health Care Technology and subsequent to that, the Office of Health Technology Assessment (OHTA). OHTA was a component of the National Center for Health Services Research and Health Care Technology Assessment (NCHSR), which later became the Agency for Health Care Policy and Research (AHCPR), the predecessor of the Agency for Healthcare Research and Quality (AHRQ).

OHTA's role was to advise the Healthcare Financing Administration now the Centers for Medicare & Medicaid Services, (CMS) on coverage decisions for new medical technologies under the Medicare program. The AHRQ continues its role as science advisor to CMS by providing health technology assessments to the Coverage and Analysis Group at CMS, whose coverage decisions are often followed by private health insurers. (*see* **Health Services Organization in the US**.)

Topics Addressed by HTA

Since there is no standard set of activities that can be said to form a part of all HTAs, much will depend on the aims of the HTA and the nature of the technology. However, one or more of the following elements are usually present [12].

- Assessment of the current state of development and use of the technology.
- Assessment of the technical characteristics of the technology if it is a device (*see* **Medical Devices**).
- Assessment of the effectiveness technology.

- Assessment of the safety of the technology.
- Economic evaluation of the technology to examine the resource use relative to the benefits/harms.
- Effect of the technology on the organization and delivery of services.
- Wider impact of the adoption of the technology.
- Ethical issues associated with the use of the technology.

Though HTA can encompass a wide range of topics spanning the technical to the social and ethical, in general most of them are limited to consideration of effectiveness and **cost-effectiveness** (*see also* **Pharmacoepidemiology, Adverse and Beneficial Effects**.)

Titles of some recent HTAs produced around the world include:

- Implantable Defibrillators (ICD)
- Positron Emission Tomography (PET) Imaging in Cancer Management
- Monitoring blood glucose control in diabetes: a systematic review
- Phakic Intraocular Lenses
- Systematic Review and Economic Evaluation of the Effectiveness of infliximab for the treatment of Crohn's Disease
- Evaluation of Molecular Tests for prenatal diagnosis of Chromosome Abnormalities
- Oseltamivir for the Treatment of Suspected Influenza: a clinical and economic assessment
- Sentinel Node Biopsy in Breast Cancer
- A Systematic Review of Atypical Antipsychotic Drugs in Schizophrenia
- Acupuncture: Evidence from Systematic Reviews
- Clinical and Cost-effectiveness of Routine Dental Checks: A Systematic Review and Economics Evaluation.

There is a tendency for HTA to focus on new technologies as these may represent the most pressing demands on health care budgets and so are of direct concern to health care funders. However, this can result in a bias, in which new technologies recommended as cost-effective displace older technologies that may be equally or more beneficial but which have not been assessed. Techniques have been developed to support a more rational and scientific approach to prioritizing research on technologies. Methods for estimating the payback from investments in such assessments have been developed but are quite crude, More recently, Claxton et al. [1] have developed a Bayesian decision theoretic approach (*see* **Bayesian Decision Models in Health Care**) to valuing additional information that helps funders or regulators decide whether the costs of reducing uncertainty by carrying out more research is justified by the expected value of improved information. The model helps regulators to decide if a health technology should be adopted on the basis of existing evidence or whether it is worth undertaking research to reduce uncertainty further.

Methods Used in HTA

Many HTAs consist of reviews of the results of existing research. Traditional reviews have now largely been discredited as being unreliable and often biased. It is important that reviews are systematic in order to ensure that there is an explicit, comprehensive and methodologically sound, and unbiased approach to identifying, appraising, and summarizing the research [10]. Quantitative approaches to combining the results of studies found in the review – **meta-analysis** – are now widely used [18]. This has been an important area for methodological development with different methods of handling variations between study heterogeneity and for combining evidence from different types of study designs, such as Bayesian **hierarchical models** [13]. Recently, there has been interest in methods for combining quantitative and qualitative data.

The most reliable evidence for estimating the effectiveness of a health technology (i.e. its effect on health outcomes) is a randomized controlled trial (RCT) (*see* **Clinical Trials, Overview**). If well conducted, this method reduces the susceptibility to bias and provides more reliable estimates of the effect of the health technology than other designs such as **quasi-experimental** and uncontrolled studies that often overestimate the effects. However, experimental evaluation of health technologies may be difficult if the technology is emerging, presents ethical challenges, and is very expensive to run. Questions are also raised about the degree to which the results of trials can be generalized to routine care situations. Thus, attempts are made to make trials as realistic or pragmatic as possible. Methods are also being developed for applying the results of studies to particular

patient subgroups or individuals, rather than just use average estimates.

If the technology is widespread and not applied to individuals, such as telemedicine or media health promotion campaigns, **randomization** is carried out at a higher level than the individual – cluster randomized trials (*see* **Cluster Randomization**) – which raises new methodological and statistical problems.

Assessing the benefits to patients of a technology involves deciding what endpoints or **outcome measures** should be used. Increasingly, it has been accepted that conventional clinical outcome measures such as lung function tests for people with asthma be replaced or supplemented by more patient-centered measures (which may be disease specific, dimension specific or cover the whole of health-related **quality of life**) [5]. All too often, technology assessments have based their conclusions on evaluations using **surrogate endpoints** or intermediate outcome measures, often physiological or biochemical markers that are taken to be predictive of important outcomes such as death. However, this often wrongly assumes that the predictive models are accurate and that in real life the pathway between the intermediate and eventual outcome is clear and unchanging. In addition, surrogates rarely predict harm and so a technology that looks beneficial on the basis of the performance of surrogate endpoints (e.g. reduction in cholesterol or reduction in cardiac dysrhythmias) may actually result in worse outcomes when measured by the effect on, say, total mortality [8].

Economic evaluation is often an integral part of a health technology assessment (*see* **Health Economics**). Using resources for one technology has an opportunity cost, in that the benefits derived from using those resources in some other effective intervention are forfeited. Thus, it is important to consider the resource implications associated with generating the benefits – cost-effectiveness. The standards for conducting economic evaluations have become more established over time [7] and now often employ complex statistical techniques for modeling the effectiveness and cost data [15].

Though most HTAs are predominantly quantitative in methods resulting in estimates of the costs and benefits of the technology, increasingly, qualitative approaches are being used to gain a richer understanding of the likely impact of technologies on patients and the health care system [11].

Timing

Health technologies have a life cycle with an early developmental phase, early adoption, then (if successful) rapid diffusion, and then a reduction in use as it becomes obsolete and is either displaced by newer technologies or abandoned as information about lack of effect or adverse side effects become available [3]. There is no best time to assess a health technology, the less established it is the more likely its uptake will be influenced by the assessment, for it is difficult to change practice once the use of a technology is widespread. However, it can be more difficult to assess a technology in its early phase because there are less data on effects and costs available on which the assessment can be based. Early assessments of some technologies, especially those involving technical skills, may underestimate the benefits if there is a learning effect. There is no right solution to this trade-off, which has been nicely captured as Buxton's paradox "its always too early until it is too late"! Several countries have established "horizon scanning" programs for the early identification and assessment of new and emerging technologies [17].

A good introduction to the wide range of methods used in conducting HTAs is provided by Stevens et al. [16].

Organization of HTA

Health Technology Assessment tends to be sponsored by national or regional health care funders such as the government, or by social or private insurance funds, or health care providers. In addition, there is substantial industrial HTA activity. HTA may be carried out in-house or by commissioning universities or specialist consultancy companies.

One of the earliest and sustained programs of HTA in the English-speaking world was established in 1993 as part of a national Research and Development program for the United Kingdom National Health Service (`www.ncchta.org`). It involves a national system of consultation to determine priorities for assessments reflecting areas where clinicians, health service managers or health service users felt there was sufficient uncertainty about the appropriate use of health technologies. More recently, the program also has a stream of work commissioned specifically by the National Institute for Clinical

Excellence, which was established to make recommendations to the NHS and government as to what health technologies should be publicly funded and to produce clinical practice guidelines. In most other European countries, there are well-established agencies for HTA. In Canada also, there have been various provincially funded HTA programs with a national Canadian Coordinating Office for HTA (CCOHTA: www.ccohta.ca).

HTA has had a long but more turbulent history in the United States [4] partly because of the fragmentation of the health care system (resulting in HTA being sponsored by several public and private organizations) and also because of the powerful professional and commercial interests. The congressional OTA and the federally funded OHTA were both closed down in the 1990s despite their good international reputation. Federally funded technology assessments are now conducted internally at the AHRQ or by contracting with one of their 14 Evidence-based Practice Centers (EPCs: www.ahrq.gov/clinic/epc/). One of these is the Blue Cross and Blue Shield Association Technology Evaluation Center (TEC), which provides technology assessment services to all independent Blue Cross and Blue Shield Member Plans, issuing around 20 to 25 TEC assessments per year.

Most health care systems are facing similar pressures and have to consider the appropriate use of the same technologies. Thus, an international organization was established in 1993 to promote the sharing of information from different assessments carried out across the world and to cooperate in the conduct of these assessments. Initially established by publicly funded HTA organizations, the International Network for Agencies for Health Technology Assessment (INAHTA: www.inahta.org) now has about 40 member agencies from 20 countries, stretching from North and Latin America to Europe, Australia, and New Zealand.

The conduct of HTA is also promoted by Health Technology Assessment International.(HTAi: www.htai.org), a new professional and scientific society. This society is the only professional society focusing specifically on HTAs around the world.

Dissemination of HTAs

The results of HTA are disseminated in a variety of ways including reports to the commissioner and articles in major medical journals. The specialist journal, the *International Journal of Technology Assessment in Health Care*, also serves as a forum for the wide range of professionals interested in the assessment of medical technology, its consequences for patients, and its impact on society. The Center for Reviews and Dissemination at the University of York, UK, in collaboration with the INAHTA, maintains a database of HTAs produced by a variety of agencies internationally (http://nhscrd.york.ac.uk/welcome.htm). It contains details of over 1600 HTA publications and hundreds of ongoing INAHTA projects. The National Library of Medicine also maintains a searchable collection of large full text technology assessments mainly carried out in the USA in *Health Services/Technology Assessment Text* (HSTAT: http://hstat.nlm.nih.gov). This includes the evidence reports from the AHRQ and the US Preventive Services Task Force's *Guide to Community Preventive Services*.

Impact of HTAs

HTAs are rarely undertaken as a purely academic exercise; they are carried out in order to inform policy and practice. The policy impact might be, for example, the decision of a health regulator to license a new drug or device of a health care funder to pay for a health technology and under what conditions (coverage decisions). At a clinical practice level, an HTA can feed into the production of clinical practice guidelines produced by a regulator, a health care organization or professional association. In ideal circumstances, HTA can result in the increased uptake of cost-effective technologies or lead to a reduction in the use of ineffective, unsafe technologies or those whose costs are too high, relative to any benefits. However, rarely does the evidence from an assessment lead to unambiguous policy or practice implications – the facts rarely "speak for themselves". Not only is there usually some uncertainty about the evidence, but also the way that evidence interplays with the practice, social, economic, and political context, can result in the same evidence being used to justify different recommendations. Thus, small benefits relative to cost may lead to recommendations to pay for a technology in a rich health economy such as the United States whilst a more cash limited system

such as the NHS in the United Kingdom might decide using the same evidence to restrict use or require more evidence [6].

At the level of clinical practice, many other factors will mediate the impact of an HTA, such as the mode of dissemination, and the degree to which it requires major professional change and the costs (e.g. financial, time, and skills acquisition) of adopting the recommended technology [14]. There has been little formal evaluation of the impact of HTAs on practice.

HTAs can influence decisions affecting the licensing or, more usually, the coverage of new drugs, devices and procedures, and so they have the potential to affect the sales and profits, the costs to health systems, and the clinicians' income. Health care professionals who use these technologies, and the industries that sell these are often suspicious that HTA is a device to cut costs and reduce professional control over health care.

This can make HTA reports the object of intense scrutiny by the health care, pharmaceutical and devices industries, government, and professional groups [6]. For example, the Canadian Coordinating Office of Health Technology Assessment (CCOHTA) had to use 13% of its annual budget to successfully beat a lawsuit from a pharmaceutical company seeking to prevent the release of a report on cholesterol lowering drugs (statins) [9]. In the United States, orthopedic surgeons orchestrated a campaign against the AHCPR in response to its guideline on the management of back pain, which stressed the importance of nonsurgical approaches. The backlash contributed to a cut in AHCPR's budget and the curtailment of guideline production [2]. This highly politicized environment makes it all the more important to have strong and independent procedures for the conduct of HTA.

References

[1] Claxton, K., Sculpher, M. & Drummond, M. (2002). A rational framework for decision making by the national institute for clinical excellence (NICE), *Lancet* **360**, 711–715.

[2] Deyo, R.A., Psaty, B.M., Simon, G., Wagner, E.H. & Omenn, G.S. (1997). The messenger under attack – Intimidation of researchers by special-interest groups, *New England Journal of Medicine* **336**, 1176–1180.

[3] Eisenberg, J. & Zarin, D. (2002). Health technology assessment in the united states: past, present, and future, *International Journal of Technology Assessment in Health Care* **18**(2), 192–198.

[4] Entwistle, V.A., Watt, I.S. & Johnson, F. (2000). The case of Norplant as an example of media coverage over the life of a new health technology, *Lancet* **355**, 1633–1636.

[5] Fitzpatrick, R., Davey, C., Buxton, M.J. & Jones, D.R. (2001). Criteria for assessing patient based outcome measures for use in clinical trials, in *The Advanced Handbook of Methods in Evidence Based Healthcare*, A. Stevens, K. Abrams, J. Brazier, R. Fitzpatrick & R. Lilford, eds. SAGE Publications, London, pp. 181–194, Chapter 11.

[6] Fox D.M. & Oxman A.D. eds. (2001). *Informing Judgment: Case Studies of Health Policy and Research in Six Countries*. Milbank Memorial Fund, New York.

[7] Gold, M.R., Siegel, J.E., Russell, L.B. & Weinstein, M.C. eds. (1996). *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York.

[8] Gotzsche, P.C., Liberati, A., Torri, V. & Rossetti, L. (1996). Beware surrogate endpoints, *International Journal of Technology Assessment in Health Care* **12**, 238–246.

[9] Hemminki, E., Hailey, D. & Koivusalo, M. (1999). The courts – a challenge to health technology assessment, *Science* **285**, 203–204.

[10] Khan, K.S., Riet, G., Glanville, J., Sowden, A.J. & Kleijnen, J. eds. (2001). *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews*, 2nd Ed., CRD Report 4 University of York, UK, (`www.york.ac.uk/inst/crd/report4.htm`)

[11] Leys, M. (2003). Health technology assessment: the contribution of qualitative research, *International Journal of Technology Assessment in Health Care* **19**, 317–329.

[12] Liberati, A., Sheldon, T.A. & Banta, D.H. (1997). EUR-ASSESS project subgroup report on methodology, *International Journal of Technology Assessment in Health Care* **13**, 186–219.

[13] Prevost, T.C., Abrams, K.R. & Jones, D.R. (2000). Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening, *Statistics in Medicine* **19**, 3359–3376.

[14] Sowden, A., Watt, I.S. & Wilson, P. eds. (1999). Getting evidence into practice, *Effective Health Care* **5**(1), York NHS Centre for Reviews and Dissemination. (`www.york.ac.uk/inst/crd/ehc51.pdf`)

[15] Spiegelhalter, D.J. & Best, N.G. (2003). Bayesian methods for evidence synthesis and complex cost-effectiveness models: an example in hip prostheses, *Statistics in Medicine* **22**, 3687–3709.

[16] Stevens, A., Abrams, K., Brazier, J., Fitzpatrick, R. & Lilford, R. eds. (2001). *The Advanced Handbook of Methods in Evidence Based Healthcare*. SAGE Publications, London.

[17] Stevens, A., Milne, R., Lilford, R. & Gabbay, J. (1999). Keeping pace with new technologies: systems needed to identify and evaluate them, *British Medical Journal* **319**, 1291–3.

[18] Sutton, A.J., Jones, D.R., Abrams, K.R., Sheldon, T.A. & Song, F. (2000). *Methods for Meta-analysis in Medical Research*. Wiley and Sons, Chichester.

[19] Wennberg, J.E. & Cooper M.M. eds. (1999). *The Quality of Medical Care in the United States: A Report on the Medicare Program, The Dartmouth Atlas of Health Care 1999*. American Health Association Press, Chicago.

TREVOR A. SHELDON

Health Care Utilization and Behavior, Models of

Models of health care utilization behavior provide guidance for defining variables, specifying the relationships between them, and evaluating programs and policies concerned with access to and utilization of health care services (*see* **Health Care Utilization Data**; **Health Care Utilization Data, Analysis**). Diagrams (as in Figure 1) are used to categorize the relevant variables and their interrelationships.

Models may be used to guide the conduct of descriptive, analytic, or evaluative studies of the operation and performance of the health services delivery system. Descriptive studies focus on profiling the variables in the model (represented by boxes in Figure 1) for a population or subgroup. Analytical designs speculate on the hypothesized relationships between the implied predictors (independent variables) and outcomes of interest (dependent variables) (displayed by arrows). **Experimental designs** or **quasi-experimental designs** test the impact of a specific program or intervention on desired outcomes (the end-points in Figure 1).

Four major types of conceptual models have been developed and applied in specifying the interrelationships of the array of possible predictors of health care utilization behavior, and in guiding the conduct of analytic and evaluative research in this area [2]. These include (i) models of patient decision making, grounded in sociological theory and research (particularly those developed by Suchman, Kosa and Robertson, and Mechanic); (ii) the health belief model, based in social psychological theory (developed by Becker); (iii) economic models of the demand for medical care (as amplified by Grossman); and (iv) the behavioral model of health services utilization (developed by Andersen and his colleagues, displayed in Figure 1) that has guided the conduct of much **health services research** on access to and utilization of health care services [1, 5]. The first three types of models will be reviewed next and the behavioral model discussed in more detail in the section that follows.

Models of Patient Decision Making

Suchman

Suchman's framework for stages of decision making about seeking medical care is focused on episodes of illness. In Suchman's paradigm, the sequence of seeking medical care for illness is divided into five stages: (i) experience of the symptom; (ii) assumption of the sick role; (iii) medical care contact; (iv) dependent patient role; and (v) recovery or rehabilitation. A group with more parochial or traditional, in contrast to more cosmopolitan, affiliations and a popular,

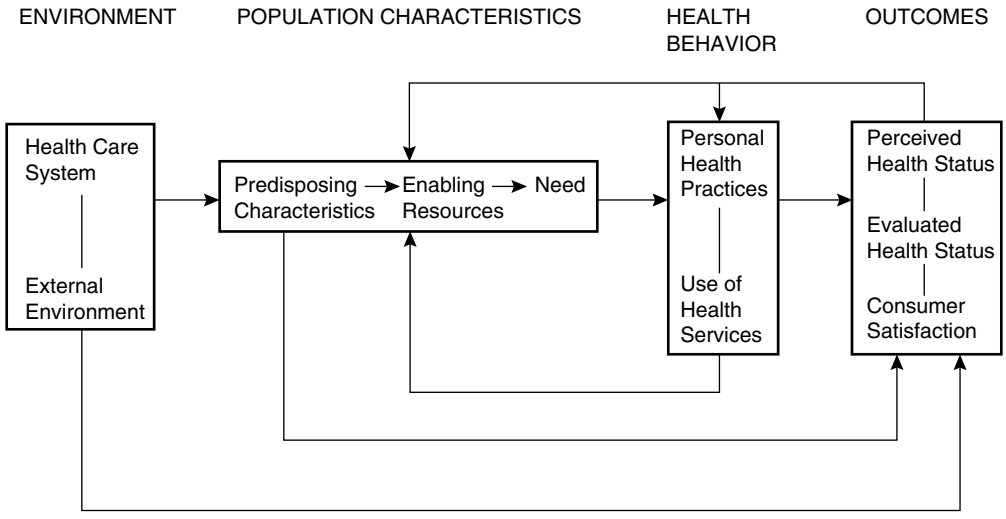

Figure 1 An emerging model – phase 4. Reprinted from [3] by permission of the publisher

rather than more scientific, orientation toward medical care would, he suggests, be more likely to delay in recognizing symptoms, linger longer in the stage of using home remedies, be suspicious of medical providers and perhaps shop around more, fail to adhere to prescribed therapies, and relinquish the sick role as soon as possible.

Kosa and Robertson

Whereas Suchman's model tended to offer more sociological or structural explanations for why individuals might respond differently at different stages of an illness episode, a model developed by Kosa and Robertson focused more on psychological explanations. Behavior is motivated by the individual's psychological need to reduce the anxiety aroused by the threat of illness. The Kosa and Robertson model also assumes stages of individual decision making in response to illness: (i) an assessment of a disturbance in usual functioning; (ii) anxiety arousal based on the perception of the symptoms; (iii) the application of one's medical knowledge to address the problem; and (iv) the performance of activities to alleviate the anxiety. Activities may be of two kinds: therapeutic interventions directed at the removal of the specific health problem, or interventions aimed at relieving the anxiety of satisfying other needs (e.g. fear) without addressing the health problem directly.

Each stage of decision making is influenced by these psychological dynamics as well as the culture and social groups (e.g. family) of which they are a part or with whom they come in contact (e.g. professional medical providers).

Mechanic

Mechanic's model catalogues an array of social and psychological factors that might influence the likely impact of symptoms on individuals care-seeking. These include: (i) perception of symptoms (e.g. salience, seriousness, disruptiveness, frequency); (ii) characteristics of individuals (e.g. tolerance of discomfort, knowledge of illness, competing needs); and (iii) accessibility of care causing disruption in the treatment process (e.g. inconvenient location or hours of service, out-of-pocket costs). However, the need for care from the point of view of the patient (self-defined illness) may not always agree with the need for care as defined by the provider (other-defined illness), which may have significant consequences for patient compliance and continuity [1].

Health Belief Model

The health belief model was originally conceived to understand preventive health care (health behavior) but has subsequently been applied to explaining care-seeking in response to illness (illness behavior) and those activities required for recovery from illness (sick role behaviors). The major components of this social–psychologically oriented model are as follows: (i) an individual's subjective state of readiness to take action, based on the individual's perceived likelihood of susceptibility to the illness, as well as its seriousness; (ii) an individual's assessment of engaging in a given health care-seeking behavior, based on weighing the benefits (reducing susceptibility or seriousness) relative to the likely costs (physical, financial, etc.); (iii) the presence of cues to action to trigger the appropriate action, coming from either internal (e.g. symptoms) or external (e.g. interpersonal interactions, mass media) sources; and (iv) the role of other modifying factors, such as demographic, sociopsychological, and structural. These factors all influence the perceived threat of the disease and the subsequent likelihood of taking action [1].

Models of Consumers' Demand for Medical Care

Economic models of consumer choice stress means (e.g. health insurance or income) through which people can attain services or translate their perceived need into economic demand for medical care. An important contribution to the demand models was made by Grossman, who argued that what consumers really demand when they purchase medical care is health [5]. A number of hypotheses might be generated by this model of joint demand for health and care including: (i) as people age and their stock of health declines they will increase their consumption of medical care to offset the decline; (ii) as people's income increases, their consumption of medical care will increase because they will place increased value on healthy days; and (iii) as people's education increases, their demand for medical care will

decline because they will be more efficient in producing health [4]. The best known application of the demand for medical care model is the RAND Health Insurance Study which employed randomized trials (*see* **Clinical Trials, Overview**; **Health Care Financing**) to estimate the effects of changes in health insurance benefits on people's use of medical care and their health status [6].

Behavioral Model of Health Services Utilization

The behavioral model of health services utilization is arguably the most comprehensive and widely applied model in health services research focusing on access to and use of health care services [2, 3]. The most current adaptation of the model is displayed in Figure 1.

The original version of the model developed in the 1960s suggested that people's use of services is a function of their predisposition to use services (predisposing variables), factors which enable or impede use (enabling variables), and their need for care (need variables). Predisposing variables include demographic and social structure factors (e.g. employment, social class, occupation, race) and health beliefs. The enabling component encompasses both resources specific to individuals (e.g. income, insurance coverage, regular source of care) and attributes of the community in which they live (e.g. physician and hospital bed supply). The need for care may be based on perceptions of the individuals themselves or diagnostic assessments by providers.

The model provides an empirical approach to assessing the equity of health services utilization. Andersen and Aday (originators of the model) assume that in an equitable system, need (rather than predisposing and enabling) components will be the primary basis for accounting for subgroup variations in use. They also distinguish those components which are more mutable (alterable by the health care system) – enabling factors – vs. those that are not – demographic or social structural characteristics.

In later versions of the model (1970s) the health care system was explicitly included in the model in recognition of the important impact of organizational and financial factors on the distribution and delivery of services. The dimensions of health services use measures (type, site, purpose, and time interval for care) were elaborated, and satisfaction added as another important (subjective) indicator of individuals experience of care-seeking.

More recently (during the 1980s and 1990s) the model allowed for the growing recognition of the importance of considering the impact of health care utilization in the context of other likely predictors of health outcomes. Revisions acknowledged that the external environment (physical, political, and economic) and personal health practices (such as diet, exercise, and self-care) influence formal health services utilization and (ultimately) health outcomes. The revised model added people's perception of their health status and clinical (evaluated) measures of health status as well as patient satisfaction with service as outcomes. Finally, it incorporated feedback loops showing that health outcomes, in turn, affect subsequent predisposing factors, perceived need for services, and health behavior.

Limitations of Dominant Models and New Directions

A number of criticisms have been offered of the original and expanded behavioral model of health services utilization, which may be seen as establishing the grounding for new substantive and methodological research based on the model [2, 7]. These criticisms relate principally to the specification of the independent and dependent variables in the model, the causal pathways between and among them, and the generalizability and policy relevance of research based on the model.

Independent Variables

The criticisms of the major predictors of utilization relate primarily to the validity or accuracy of the operational definitions used to measure major study concepts; the fact that most studies in which the model is used do not fully encompass all components of the model; the need to add other dimensions to capture adequately the relevant predictors for selected types of utilization or populations; and the likelihood of significant **interactions** between subcomponents of the model.

Dependent Variables

Important extensions of the model in terms of the utilization variables themselves would be to explore

more systematically the interrelationships (or trade-offs) between different types of service use (e.g. ambulatory vs. inpatient), as well as the relationship of utilization to both patient satisfaction and health outcomes.

Relationships between Variables

Full tests of the interrelationships between and among variables entail the use of stronger analytic and evaluative research designs and the application of more sophisticated modeling techniques in empirically examining these interrelationships.

Generalizability

Major criticisms of the model have focused on the fact that most studies in which it has been utilized explain only a small amount of the variation in health services utilization. Also, the program and policy relevance of the model would be enhanced by the design of studies and analyses to relate more directly the impact of utilization on patient satisfaction, health outcomes, and costs.

Summary

In summary, various factors may account for those who ultimately seek health care. Substantial progress has been made in specifying and measuring the relationships among these factors. The conceptual models reviewed here provide integrative frameworks for considering many of these factors and their interrelationships. However, health services research can provide continued guidance in refining these models, designing and implementing empirical studies to test and evaluate them, and shaping the formulation and interpretation of the policy relevance of research guided by such models.

References

[1] Aday, L.A. (1993). Indicators and predictors of health services utilization, in *Introduction to Health Services*, S.J. Williams & P.R. Torrens, eds. 4th Ed. Delmar, Albany, pp. 47–70.

[2] Aday, L.A. & Awe, W.C. (1997). Health services utilization models, in *Handbook of Health Behavior Research*, Vol. I *Determinants of Health Behavior: Personal and Social*, D.S. Gochman, ed. Plenum, New York, pp. 153–172.

[3] Andersen, R.M. (1995). Revisiting the behavioral model and access to medical care: does it matter?, *Journal of Health and Social Behavior* **36**, 1–10.

[4] Feldstein, P.J. (1979). *Health Care Economics*. Wiley, New York, pp. 78–79.

[5] Grossman, M. (1972). *The Demand for Health: A Theoretical and Empirical Investigation*, Occasional Paper 117. National Bureau of Economic Research, New York.

[6] Newhouse, J. & the Insurance Experiment Group (1993). *True for All? Lessons Learned from the RAND Health Insurance Experiment*. Harvard University Press, Cambridge, Mass.

[7] Pescosolido, B.A. & Kronenfeld, J.J. (1995). Health, illness, and healing in an uncertain era: challenges from and for medical sociology, *Journal of Health and Social Behavior* **Extra issue**, 5–33.

LU ANN ADAY & RONALD M. ANDERSEN

Health Care Utilization Data, Analysis

Data on health care utilization are used to address important questions about what is done for whom, and can shed light on why and with what outcome, in the large segment of economic life that "delivers" health care services to populations. We discuss the kinds of utilization studies and analytic issues that commonly arise. **Multivariable modeling** techniques are used to identify differences in use and possible reasons for these differences, although the available data often limit the questions that can be successfully addressed (*see* **Health Care Utilization Data**).

Important concerns when studying utilization are (a) identifying all instances of use, (b) identifying the at-risk population, (c) accounting for differential risk of individuals, and (d) distinguishing real differences from random noise. These issues are explored in more detail in this article, and in the articles on **health care utilization data** and **risk adjustment**.

An influential class of **small area variation** studies has established that people in different parts of the country and in different communities within the same region receive very different care, for example, **population-based rates** of tonsillectomy, hysterectomy, or hospitalization vary tremendously. Additional research has tried to understand these differences, for example, asking if areas with high hospitalization rates also have high rates of inappropriate hospital admissions (*see* **Health Care Utilization and Behavior, Models of**). Geographic variation in patterns of use may reflect lack of agreement within the medical community about treatment options, and help identify opportunities for improving care through standardization to better treatment protocols.

Disparities studies that explore variations in treatment by patient race or sex may reflect societal prejudice in allocating expensive resources. Variations by payer class, such as Medicaid versus private insurance, or payment method, such as **fee-for-service** versus **capitation**, may point to inequities or inefficiencies related to financing.

Other studies seek to estimate the effect of various factors on utilization during an instance of caregiving, such as a hospital admission. Potential predictors of utilization include patient characteristics (sociodemographic, medical), characteristics of doctors or other medical providers (sociodemographic, training, and experience), and characteristics of the conditions of practice (the particular site or its features, or the organizational/financial structures under which the care is given). Predictor variables can be heavily confounded; for example, most patients seen at high-volume hospitals may be city dwellers, making it difficult to sort out the separate effects. **Hierarchical** or **nested** models are needed to explore the influence of facility-specific factors, and may be of particular importance when the units of analysis differ from the units of inference (e.g. individual hospital admissions are studied for the purpose of comparing hospitals).

Another goal is to describe differences in **case mix**, as a guide to why providers may differ in the care they give and the outcomes achieved. Case-mix differences may serve as the basis for redistributing money among providers, so that those who treat the sickest patients receive the highest per-patient reimbursements.

Yet another goal is to provide "quality" reports, which guide patients and health care purchasers, assessing what health care providers do and what they achieve with the people they serve. An important initiative in the United States in the 1990s has been developing a protocol for comparing health plans (HEDIS) along measures such as what fraction of a plan's women of an appropriate age receive annual mammograms. However, HEDIS measures are generally not **risk-adjusted** and even when we agree that providers should encourage mammography, it is easier to achieve, say, 95% compliance with middle-class women than with a poor, transient population. Also, few HEDIS measures assess the quality of care given to the very sick, largely because this is so difficult to do.

Cost-effectiveness studies seek to relate the cost of the health care inputs used to the value of an achieved outcome. They require data on utilization and a plausible methodology for "pricing" this, as well as a numerical measure of the outcome, such as "quality-adjusted life years" (**QALYs**). Health care strategies with the highest QALY yield per dollar might be the first to be implemented (or the last to be eliminated) in a health care system with constrained dollars.

Provider profiling is used to identify individual doctors, hospitals, or health care systems with exemplary or problematic practices. Especially when providers are compared in public releases of analyzed

data, much harm can be done if good providers are "flagged" as problematic either because of random variation (small numbers and a "bad bounce") or because they care for an unusually difficult mix of patients. Appreciation of the "small numbers" problem and good **risk adjustment** is critical to useful provider profiling.

Special Features of Utilization Data that Require Analytic Attention

Skewness and Heteroscedasticity

Health care utilization variables often have a predominance of low values and long, heavy right tails. This is true when looking at numbers of office visits or hospitalizations per year, days of stay in a hospital for hip fractures, and costs of care for individual hospitalizations or during fixed periods of time. For example, in a working insured population during a one-year period, 20 to 30% of people eligible to receive health care may incur no costs, many more have low-level, nonhospital expenses, and the 5% with the most intense use may account for around 50% of all expenses; the standard deviation for expenditures is generally several times larger than the mean, and in a system with average costs of about $2000 per year, the most expensive cases exceed a million dollars. A few large outliers can substantially distort analyses, even in data sets that contain hundreds of thousands of cases, and decisions about whether to truncate or remove these extreme cases can affect study findings. For many purposes, top-coding (e.g. making the dependent variable equal to the smaller of $50 000 and actual cost) can effectively control the undue influence of extremely large observations.

A technique for addressing the concentration of zero values is to use a two-part model [2], in which one equation predicts the probability of any use (in the whole population) and a second equation predicts the level of use among users. The expected level of use for an individual is then calculated by multiplying these two estimates together. This framework has been extended to a four-part model in which the probability of hospital use is estimated for users, and then the costs for users without hospitalization and for those hospitalized are separately estimated [2].

To address skewness, many authors transform the utilization variable, for example, by modeling log (1 + dollars). (We add 1 because log(0) equals negative infinity). This may help in identifying factors that affect use because it makes the p-values on the significance of individual predictors more credible, but usually does not help predict actual levels of use (dollars, that is, not log dollars). Log transforms are especially problematic when the data contain both small observations and large ones, because the same multiplicative change has so different a meaning when, say, applying a 20% increase to a $10 versus a $1000 expense. When retransforming a log-transformed variable into its original scale, a smearing estimate [2] can be used to address bias, but the smearing only works under the assumption of constant variance, which is unlikely when the observations vary widely in magnitude. A simple way to put retransformed estimates on the "right scale" is to multiply all estimates by the number needed to make the average prediction equal to the average actual outcome (i.e. multiplying by $k =$ mean actual/mean predicted). Typically, R^2 values are higher for models in a log scale, although retransformed estimates fit the actual outcomes less well than models constructed on untransformed data. **General linear modeling** provides an attractive framework for simultaneously addressing the problems of skewness and nonconstant variance of the outcome variable, while predicting it in its original, untransformed scale. For example, the function that "links" costs to predictors can be specified as the log function and variances can be specified as proportional to means.

Lack of Independence

When studying hospital admissions, multiple hospitalizations for the same patient cannot necessarily be identified. Thus, the data may not be able to answer questions such as "now that patients are being discharged from hospitals earlier, have readmission rates increased?" Furthermore, random variation is greater when rehospitalizations are common than when single hospitalizations are the norm. When clustered data are analyzed as if independent, chance variability can be misinterpreted as evidence of systematic differences [1].

When examining the role of a site-specific characteristic such as "do hospitals which see lots of AIDS cases do better with them?" or "do major teaching hospitals see sicker patients?", analysis should account for the fact that patients are nested in hospitals, which are in turn nested within hospital type.

Otherwise, effects attributed to hospital "characteristics" can easily be determined by the experience of a few large facilities.

Death

Patients who die shortly after entering the hospital often use the least resources, while those who remain alive a few days but eventually die are among the most expensive. Some authors propose treating utilization (a nonnegative variable) as the outcome in a **survival analysis**, with death as the censoring variable. The resources that a patient would have used had he or she not died are then estimated, which may or may not be a useful concept. In addition, death is an "informative" reason for censoring, which may muddle the interpretation of findings. In general, it is unclear how and whether to use information about death in predicting utilization.

Study Scope

The Study Population

For many purposes, the "full" or "fully eligible" population, such as people who reside in a particular area, persons theoretically eligible to receive care from the Veteran's Administration, or enrollees in an HMO, is the right study frame. However, many people entitled to receive care in a system do not use it, perhaps because they require no care or because they receive it elsewhere. Sometimes the relevant study population is "all users", such as people who have received care at a particular VA center, or who use a particular physician as their primary care doctor.

To learn how medical problems are treated and what outcomes are achieved requires problem-defined study cohorts, such as people with diabetes, with hypertension, or with low back pain. Eligibility criteria affect what we see. For example, if the data from system A enable us to find all people who are even mildly diabetic, while system B data only identify hospitalized diabetics, we will probably see lower rates of diabetes, but more intensive treatment and worse outcomes per diabetic patient in system B.

Breadth of Use Within Relevant Populations

Gross measures of use (such as number of hospital admissions per thousand person-years of experience) matter to payers, but administrators need to understand where inefficiencies occur. Thus, for example, numbers of hospitalizations for respiratory problems for patients with asthma, and the prevalence of blood sugar tests and eye exams for diabetics provide more focused views of a health care delivery system. One difficulty in conducting such studies is that only some systems maintain disease registries that specifically allow the utilization of, say, diabetic patients, to be tracked.

Individual payers (such as state Medicaid programs) are principally interested in monitoring the utilization for which they pay, but the larger community has an interest in tracking the outcomes associated with all care that individuals receive.

Utilization in a Population Versus During a Period of Treatment

Some services are delivered at most once to any one person, such as hysterectomy, which is removal of the uterus. However, many services can be delivered more or less often and with variable intensity. Thus, each of the following questions may be of interest about inpatient hospital care for a population: What fraction was ever hospitalized during a given year? How many hospitalizations occurred per person-year of exposure? How many days of hospital care were incurred per person-year? How many days of intensive care unit (ICU) stay were used per person-year?

Examples of relevant measures when the hospital admission is the unit of analysis are: total length of stay, presence of any special care unit stays, number of days in special care units, number of x-rays ordered, and total cost of diagnostic testing. Comparisons are only meaningful among relatively similar cases; little insight is gained by pooling information for patients admitted for heart attacks with those admitted for hernias.

Which measure of utilization is examined depends upon the policy purpose. For example, when utilization is examined for quality-monitoring purposes the most relevant measure may be whether an appropriate medication or service was delivered, such as beta blockers for heart attack patients, rather than how often; with hospice programs that provide supportive care for people thought to be near death, per-person utilization, rather than services per month of enrollment, may be most relevant.

What is the Unit of Analysis?

When summing total hospital costs over a group of admissions, it does not matter whether an expense relates to a single admission, to several admissions for the same individual, or to one admission for each of several people. Such distinctions are important, however, to explore whether low costs per admission are due to frequent readmission for people discharged "early". "Unbundling" and "cost shifting" may also produce apparent shifts in utilization that are not real. (Are hospital costs lower simply because of separate billing for procedures that might have been subsumed in a global hospital bill, or because some services have been shifted to the outpatient setting?)

Episodes of Care and Calendar-based Time Frames

Utilization per "episode of care", ranging from first problem identification, through active treatment and follow-up can be used to compare providers on the efficiency with which they handle a defined medical problem, such as "stomach pain, due to an ulcer". Episodes can only be studied when care offered to the same person in different settings can be linked. Other potential difficulties arise in defining when one episode ends and a second one begins; the concept may not even make sense for chronic conditions. Also, when the same person has more than one medical problem, it is not obvious which services should be assigned to which episode.

When a medical event has a readily identifiable starting point, but no clear endpoint, it often makes sense to examine utilization within a fixed window of observation long enough for most follow-up care to occur. For example, we may study all stress tests that occur within 30 days following a hospital admission for heart attack, whether or not they are done in the hospital; or, all respiratory-related tests and services offered within the first six months after a breathing problem is identified.

Even costs per "episode" may not capture efficiencies associated with preventive care, since the number and/or severity of episodes may be affected by the presence and quality of preventive services. A yet more global way to examine utilization is through the lens of total use per person-year of coverage.

References

[1] Diehr, P., Cain, K., Connell, F. & Volinn, E. (1990). What is too much variation? The null hypothesis in small area analysis, *Health Services Reports* **24**, 741–771.

[2] Duan, N., Manning, W.G., Morris, C.N. & Newhouse, J.P. (1990). A comparison of alternative models for the demand for medical care, *Journal of Business and Economic Statistics* **1**, 115–126.

ARLENE S. ASH

Health Care Utilization Data

Many important questions about the fairness, efficacy or efficiency of health care delivery systems require utilization data. Ideally, we would like to have information on the nature, timing, cost, and setting of each person's utilization, and the ability to link that information with personal demographics, health status characteristics, and health outcomes, including declines in **functional status** and death.

Health services research works best when there is a complete list of persons whose care is being tracked and longitudinal records of the medical problems addressed, all care given, and health outcomes. Such records, maintained by the Centers for Medicare and Medicaid Services (CMS) for its Medicare program (which covers over 40 million people, including almost all US citizens over the age of 65) are a uniquely powerful resource for health services research (*see* **Medicare Data**). Unfortunately, loss of data, due to the failure to acquire "encounter records" (dummy bills) from managed care programs, is of growing concern in the Medicare program and elsewhere.

Another important source of health utilization data is the Veteran's Health Administration (VHA), with over 4 million users (mostly military veterans) annually. The main weakness of these data lies in the difficulty of identifying the set of people who, were they sick, would seek their care in this system. On the other hand, the VHA, as the nation's largest integrated health care delivery system, supports well-integrated, research quality, administrative and clinical data systems through the Veterans Affairs Information Resource Center (VIREC) [2].

In contrast to the completeness and continuity of Medicare data, and to the sophistication of the VHA's nationally integrated data system, is the fragmented information available for low-income persons with Medicaid health coverage. Each state administers its own Medicaid program, eligibility requirements constantly change, people move on and off the system as their incomes and other circumstances that affect eligibility change, and the same person can appear under different identification numbers.

Data from privately insured populations have intermediate-level quality, with most people remaining enrolled through the same employer year after year, although they may switch health plans during annual, open-enrollment periods. Other problems arise because coverage is usually offered to "families" of employees; thus, marriage, divorce, alternative coverage that becomes available (or is lost) to a spouse or child, job loss, and geographical movement can all disrupt the continuity of the data. Typically, employer-based coverage systems do not record why someone disenrolls; they may not even have an explicit record ("positive enrollment") of each person entitled to receive care in their "family" contracts.

Several government surveys address these gaps in US health care utilization data. One is CMS's Medicare Current Beneficiary Survey that seeks to capture all health care utilization (not just Medicare-covered utilization) from randomly selected program beneficiaries (*see* **Medicare Data**). Another is the National Health Interview Survey (NHIS) conducted, since 1957, by the National Center for Health Statistics (NCHS). Although NHIS data focuses on the health (rather than the health care utilization) of the US civilian, noninstitutionalized, household population, it contains some important utilization information, notably on child immunizations [3]. The Medical Expenditure Panel Survey (**MEPS**) contains extensive additional data on a subset of the NHIS sample, focusing on the nature, frequency, cost, and financing of health care utilization [4].

Only some health records are computer accessible and fewer can be reliably tracked at the patient level. This limits the questions that population-based studies can address. For example, many statewide databases capture in a uniform format and make available for research, a file with one record for each inpatient admission. Available variables include the age, sex, zip-code of residence of the patient, the principal problem that caused the admission and other medical problems present (using the International Classification of Diseases diagnostic coding system), major procedures (such as surgeries) received, dates of admission and discharge, discharge disposition (e.g. in-hospital death, transfer to another facility, discharge home), days of stay in special care units (e.g. ICUs), and hospital billing information, including "payer". These data capture very much the same information that Medicare requires for hospitalizations, in very much the same format, but for all persons, not just those in the Medicare program; see [1, Chapter 5] on administrative data. However,

despite the richness of these data, a great deal of important medical information is missing, some of which can (with time and effort) be captured by retrospectively abstracting information that is usually recorded in patients' medical charts; see [1, Chapter 6] on medical record data.

Another way to improve the completeness of available information is through prospective data collection, which assures that the desired elements will be present by collecting them while care is being administered. In certain health care systems, various additional computerized information, such as laboratory findings as well as tests ordered, drugs prescribed, and amounts dispensed, can round out the picture of what medical problems were being seen and what resources were used.

Patient surveys can capture "outside" utilization, such as purchases of nonprescription drugs and use of alternative or uncovered services like chiropractic or acupuncture, as well as provide insight into patients' views of their own health or of the care they receive. Although self-report is not an ideal way to capture health care utilization, it may be important when more reliable information is not available. Surveys can be very helpful in targeting people who might benefit from case management; see [1, Chapter 7] on survey data.

Measuring Utilization

Utilization studies may focus on a particular kind of use or on a summary measure of total utilization or "cost". Kinds of uses include hospital admissions, specific surgeries such as hysterectomies and tonsillectomies, ambulatory care (such as doctor's office visits and outpatient surgeries), readmissions to hospital (within, say, two weeks of an earlier discharge), diagnostic tests, referrals for specialist care, and intensive care unit stays. "Cost" measures can be constructed as weighted averages of different factors, where the weights reflect the intensity of resource consumption and not necessarily the dollars exchanged.

Average annual costs per person is a natural summary measure of health care expenses. In the case of a purchaser of medical care, such as CMS's Medicare program, "cost" is most commonly defined as the sum of all dollars that the purchaser pays for covered services; "cost" can also include overhead (the administrative costs associated with running the program). "Total" health care costs are larger than this, however, including at least the "out-of-pocket" expenses of health care consumers for covered services (copayments and deductibles) and consumers' expenditures for noncovered goods and services, such as over-the-counter medicines and devices, dental, psychiatric or long-term care, and visits to "nonorthodox" practitioners such as acupuncturists or chiropractors. Some calculations also attempt to capture costs ancillary to the receipt of care, such as the price of transportation to providers or the value of time lost in care seeking, as well as costs ancillary to being sick (lost productivity).

Most health care providers do not know the cost of particular instances of caregiving. If average (rather than marginal) costs are sought, the accounting system used to allocate fixed overhead expenses to particular cases matters, as does the universe of cases over which the average is computed (e.g. all admissions at the same hospital, admissions to the same administrative unit, pneumonia cases, admissions paid for from a single source). On the other hand, marginal costs, such as the cost of drawing and testing one additional blood sample in a fully equipped and staffed laboratory, may seriously understate the resources needed to provide services.

The "charges" that appear on many billing records often bear little relationship to either what payers actually pay or what expenses were actually incurred. Charge comparisons are suspect, especially when pooling or comparing cases from institutions with different accounting systems. Health services researchers sometimes use the method of "cost-to-charge ratios" to convert charges to a more credible estimate of costs.

Another way to summarize utilization is by "pricing" and counting each unit of care; a "synthetic" total cost is calculated by summing the imputed costs associated with the care given. The price of a service may be generated internally (e.g. as the average charge associated with it in the data) or externally, as a "book rate". The technique is credible so long as most relevant services are likely to be captured in the data and the relative prices, at least, form a believable weighting system. Any summary cost figure is essentially a weighted sum of inputs.

When comparing utilization across different delivery systems, it is important to recognize that some data-capture systems are substantially more complete than others. As a rule, records are most complete

and accurate when payments are linked to individual services through bills; "what was done" often cannot be tracked when capitated, and "lump sum" payment systems are used.

References

[1] Iezzoni, L.I. ed. (2003). Extensive discussion about billing records and coding conventions, *Risk Adjustment for Measuring Health Care Outcomes*, 3rd Ed., Health Administration Press, Ann Arbor.
[2] http://www.virec.research.med.va.gov/.
[3] http://www.cdc.gov/nchs/nhis.htm.
[4] http://www.meps.ahrq.gov/default.htm.

ARLENE S. ASH

Health Economics

The principal foundations of health economics are based in microeconomic theory and welfare economics [17]. Essentially, this field of specialization within the discipline of economics addresses the allocation of resources directed to health improvement and the organization, delivery, and financing of health services. Under this broad purview, practitioners in the field of health economics have tackled such questions as:

1. How does the uncertainty of health outcomes influence the optimal forms of organizing and paying for medical care [1]?
2. What mix of cost-sharing between patients, health plans (insurers), and health care providers (e.g. hospitals and physicians) will produce optimal outcomes in terms of the most improved health for the least incremental cost [9, 10]?
3. Under what conditions would increased competition among providers of health services be likely to produce improvements in health status and efficiency relative to existing market arrangements for health care [29]?
4. What are the economic costs to society of prevailing patterns of illness [32, 33]?
5. How does one measure the cost and benefit of programs directed toward the improvement of health [21, 27]?
6. What are the aggregate results of health care, when measured as increased **life expectancy** for given levels of health care expenditure [3]?
7. What are the differences in performance (e.g. in the quality and efficiency of services) between not-for-profit and for-profit organizations in health care [14]?
8. What contractual arrangements provide the greatest protection for health plans, providers, and patients from the "agency problems" inherent in the sometimes-conflicting interests of those parties [25]?

These questions illustrate, but of course do not fully characterize, the range of issues subsumed in health economics. A common theme cutting across theory and empirical work in health economics is to discover which forms of market structure, industrial organization, and individual behavior lead to efficient and equitable outcomes. Potential market structures range from the one extreme of "perfect competition", in which large numbers of consumers and providers interact in an environment of perfect information, to monopoly, with one large firm controlling the market. Virtually no health care services are provided in a market structure that is close to the perfectly competitive or monopoly model.

Indeed, health care is distinguished from other economic markets by specialized information, principal–agent relationships, a relatively small number of providers in any given local area (an "oligopoly"), an inherently intimate and highly personalized service, and the dominance of third-party insurance. Arrow [1] highlights the special nature of health care from the perspective of the economist as uncertain medical consequences resulting in demand for treatments determined by physicians with payment emanating from third-party insurance carriers. Thus, the consumer is unable to predict the illness, not responsible for selecting the services he will receive, and will not – for the most part – pay the bill. This peculiarity of medical care markets challenges the traditional theoretic paradigms.

Accordingly, health economists have approached market structure from a different tack: What kinds of economic incentives and countervailing power would induce consumers and concentrated provider oligopolies, characterized by few, large firms producing highly differentiated services for a wide array of consumers, to behave efficiently and to achieve equitable outcomes?

In attempting to answer such questions, the economist's attention inevitably turns to the related issue of how health care providers are organized; in terms of "vertical" integration among the suppliers of inputs (e.g. pharmaceutical manufacturers) and the "output" providers (e.g. hospitals and medical groups), and the "horizontal" integration in local markets of providers of similar services (e.g. hospital mergers and consolidations of medical groups) (*see* **Health Services Organization in the US**). Moreover, while market structure and industrial organization exert a strong influence on health services outcomes, the factors governing the individual behavior of consumers and providers are equally pivotal in the study of health economics (*see* **Health Care Utilization and Behavior, Models of**). Seminal studies of the role of coinsurance and deductibles (patient "cost-sharing") in encouraging the efficient use of services [30] have

contributed greatly to our understanding of health care, as has the theoretic and empirical research on different modes of provider payment [16, 17] (*see* **Health Care Financing**).

Ever since Kenneth Arrow's pioneering work on the role of uncertainty in shaping health economics [1], economists have investigated the importance of payment and organizational arrangements in determining health outcomes. In the past decade, economics has made great strides in addressing these issues through the lens of "agency theory" [7, 25]. This theory illuminates how the form of ownership, organizational rules, methods of paying health care providers, and the level of payment to providers interact to induce persons and organizations (the "agents") to act in conformance with the objectives of parties (the "principals") who delegate discretion and authority to those agents to act on their behalf. Health economic research has focused on health plan benefit design, experience rating of health plans, and risk adjustment–among other contract features–as mechanisms for ameliorating agency problems.

Defining "Efficiency"

In defining efficiency, the economist has in mind two distinct and quite specific types:

1. Technical efficiency refers to the production of a given amount of services ("output") for the least amount of resources ("cost"). This can be imagined as minimizing average cost per unit of output, or – equivalently – minimizing the total cost of producing a predetermined level of output.
2. Economic, or "allocative", efficiency examines "trade-offs" in the allocation of resources. An arrangement is allocatively efficient when the incremental benefits of services provided are equal to the incremental costs of those services. Thus, allocative efficiency is measured "at the margin": Is the change in total cost (marginal cost) of services matched by an equal change in patient health benefit (marginal benefit) from those services?

Following these definitions then, the search for efficient arrangements follows the so-called Pareto criterion: social welfare is optimized when no arrangement can be devised under which some individual(s) could be made better off without others being made worse off. This criterion effectively requires both technical and allocative efficiency. Either failure to produce at least cost or failure to deliver the correct output (aligning marginal benefit with marginal cost) would violate the Pareto principle.

Kaldor [20] and Hicks [15] refined the Pareto rule into a potential compensation criterion. Assuming that the gains to the "winner(s)" under some new arrangement could be measured, a situation was optimal if and only if no change could be effected that would leave winners with sufficient gains to compensate fully the losers. The use of a standard of "potential", rather than actual compensation, reflects the existence of real world costs of information and exchange (so-called "transaction costs") that impede actual exchange of compensation.

In recognition of the centrality of efficiency within economics, a set of methodologies for economic evaluation has been developed. Those methods can be broadly categorized as follows:

1. cost–benefit analysis;
2. cost–effectiveness analysis;
3. cost–utility analysis.

Over the past 20 years or so, a substantial literature has developed in the applied area of health economic evaluation [8]. Each of these techniques is grounded in economic theory, and their application to health services problems is illustrated in what follows.

Cost–Benefit Analysis

Cost–benefit analysis translates all costs and benefits into monetary units. The opportunity cost concept underlies the logic and implementation of all three economic evaluation methodologies. Opportunity cost is defined as the value of the resources used up in a given activity, measured as the value that those resources would have produced in their next best alternate use. Hence the value of opportunities foregone constitutes the opportunity cost of a given employment of resources. Even if no money changes hands – for example, in the case of time and assets donated for a particular activity – the resources do

have this opportunity cost, which includes both direct, observable monetary costs and implicit opportunities forgone.

The time value of resources, captured in normal financial dealings through the rate of interest, is another crucial element in cost–benefit analysis. Looking forward from today, a cost incurred a year from now is somewhat less onerous than the same monetary cost incurred today. Similarly, a benefit realized one year from now is perceived as less valuable than a benefit of equivalent magnitude delivered today.

This rate of time preference is reflected in the practice of "discounting" costs and benefits through the use of a discount rate – analogous to the interest rate or rate of return required in financial transactions. The use of a discount rate in valuing costs and benefits implies that those consequences are long-lived, or spread over a period of time. Thus the discounting approach is appropriate for health programs that take the form of investments, which involve commitment of resources over time in return for future benefits.

Whereas persons considering *financial* investments generally accept the logic of requiring *interest* (some additional amount above what they invested originally) as compensation for having to wait to receive returns in the future, a thought experiment may be useful for the reader seeking to convince himself or herself that this approach can be applied appropriately to investments the direct payoffs of which are in terms of *health*, not dollars. Suppose that investing in a new positron emission tomography (PET) scanner costs $1.2 million today, and is expected to produce health benefits starting three months from now and lasting for the useful life of the scanner (estimated to be 10 years, for example). Furthermore, assume that those health benefits are expressed as earlier detection of, and more rapid recovery from, a variety of acute health problems. For purposes of the hypothetical, let the incremental (specific) costs of caring for those health problems *if not* detected earlier by PET scanning, be valued cumulatively at $200 000 per year (say, 20 cases per year at $10 000 costs saved per case).

The use of costs of caring avoided as the value of the scanner's *health benefit* simplifies the illustration that benefits can be converted into monetary equivalents. Then, to complete the reasoning, if one assumes that health investments "compete" with financial investments for scarce resources, it makes sense that the expected rate of return on the next best alternate financial investment of comparable risk should become the rate of return required for a given health investment. Thus, a discount rate equal to the foregone financial return should be used to convert the future stream of benefits into its (time zero) present value equivalents.

Now let the required return on investments of comparable risk be 20%. In this case, if the initial costs of the scanner were subtracted from the discounted present value of the future health benefits (in a technique described in the next section), using 20% as the discount rate, one would calculate the scanner's net benefit valued as of now to be a negative $361 506. The reason is that, given that the benefits accrue over the future 10 years, they are not sufficiently large to offset the time costs of waiting (the 20% rate of return foregone) to recover the initial investment cost of $1.2 million.

Inflation would not affect these comparisons, because the health benefit values and the discount rate both would simply be increased by the same proportionate amount to reflect the rising cost of living. Thus, one can think of these examples as valuing costs and benefits in "real terms" (i.e. having abstracted from inflation).

Not only does the discount rate capture the time value of resources, but also the risk involved in the costs and payoffs from different programs. In reality, since health programs often do not adopt the language of owners, investors, and return on investment, the notions of business risk, financial risk, and systematic risk – so central to mainstream economic analysis of investments – have only infrequently been applied in health program applications of cost–benefit analysis. Nonetheless, to the extent that the benefits and costs of health programs are stochastic, not deterministic, it is appropriate to increase the discount rate from the level appropriate for a "riskless" investment, to compensate for the riskiness of the project.

To convince oneself of the appropriateness of building a positive "risk premium" into the discount rate for health investments, let us revisit the earlier thought experiment of the PET scanner. Suppose in that case that the benefits of the scanner were risky, in the sense that the estimated savings in cost of caring were dependent on alternate treatments available and environmental conditions affecting personal and public health for the kinds of health problems detected by the scanner. Then

it seems reasonable that to a risk-averse decision maker the health benefits per year would be worth something less than the stated amount of $200 000. Put differently, the decision maker would accept a "certainty equivalent" amount of something less than $200 000 per year *for sure* in exchange for the current risky "claim" to an *expected* value of $200 000 per year.

Modern finance theory allows one to calculate the size of the premium to be built into the discount rate (or to value the certainty equivalent amount per year) for a given level of riskiness. The risk premium to be added to the discount rate might, for example, be calculated from the capital asset pricing model (CAPM) developed by William Sharpe, Jan Mossin, and John Lintner [5]. The CAPM model assumes that: (i) the capital market is perfectly competitive; (ii) transaction costs are zero; (iii) investors have homogeneous beliefs about the risk and return on assets in the economy; and (iv) investors hold well-diversified portfolios (assumptions that, while not strictly true, seem to generate patterns of asset returns generally consistent with empirical experience in the security markets, although the CAPM's specific validity has come under recent challenge [11]. If these assumptions hold true, the required return on a particular investment – "health" or "financial" – is given by the following equation: $r_i = r_F + \beta_i(r_M - r_F)$, where r_i is the expected (required) return on a risky investment i, r_F is the return on a riskless investment (say, in 10 year Treasury bonds), r_M is the expected return on the "market portfolio" of economy-wide assets, and β_i is the systematic, or "market", risk of investment i. This systematic risk, or "beta", is measured by the **regression** coefficient (**covariance** of returns on i with returns on M, divided by the **variance** of the market portfolio's returns), and represents the notion that only such systematic risk will be "priced" in required returns. Nonsystematic risk, the unique variability associated with each asset's returns, will be averaged out ("diversified away", in the finance lexicon) by holding a large number of assets not perfectly correlated with each other in one's portfolio. Or one might find, equivalently under the CAPM, that all the risk in the payoffs from the scanner investment had been diversified away by the decision maker's holding of a well-diversified portfolio of health *and* financial investments. That is, suppose the risk was all "diversifiable"; in other words, the returns on the scanner investment had zero covariance with broader activity in the market. In this case, no "risk premium" would be built into the discount rate. For the purposes of the preceding example, assume that this "risk premium" accounts for, say, 10%, or half of the 20% required return. That implies that, if the health benefits were known with certainty, the appropriate discount rate would be a smaller amount, 10%. Then the net benefit in present value would change to +$28 913, and the scanner investment would be worth undertaking, on balance. The trick in factoring risk into health investments is to determine this risk premium, which in theory should reflect the extent to which the health payoffs (and costs) covary with returns on assets reflecting the larger economy (the "market portfolio"). In practice, this is extremely difficult to do, and analysts instead generally perform **sensitivity analyses** of the impact of different discount rate assumptions on estimated net benefits.

The analyst's perspective is crucial in implementing a specific cost–benefit analysis. Alternative points of view include the following:

1. society's – for example, in the case of public programs, in which the costs and benefits are broadly diffused among a large population (a publicly funded mobile coronary care unit for emergencies represents such an investment);
2. third-party payers' – for example, if a private health plan were structuring a new covered benefit (say, bone marrow transplantation), and wished to evaluate whether the long-term benefits in terms of market share (additional premium revenues) would offset the expected costs of the additional coverage;
3. health care providers' – for example, if a hospital or medical group were implementing a new information system and wished to compare the capital and operating costs of the investment with the future benefits of improved patient care and enhanced clinical efficiency over the long run; and
4. patients' – for example, in the case of a consumer cooperative organized for the health care of its members, the decision to develop a specialized home care unit (say, for persons with chronic obstructive pulmonary disease), with the costs to be fully funded through a surcharge to member premiums and with caregiver support to be provided by member volunteers.

Each one of these points of view suggests a potentially different perspective on costs and benefits. For example, the nature of the publicly funded mobile coronary care unit implies that a social opportunity cost viewpoint be used to assess that program. A full accounting of the direct monetary costs and implicit opportunity costs of the investment would be appropriate, as would a broad conception of population benefit. In contrast, the third-party payer might not incorporate the value of volunteer resources (e.g. donated time) in its calculation of program cost and would likely value benefit more narrowly in terms of gains in premium earned net of health care costs incurred for subscribers only. Similarly, providers and patients will view costs and benefits in narrower terms, based on the inflows and outflows of resources internalized by them.

Another important consideration in cost–benefit analysis is the methodology used to measure program benefits. Two basic approaches exist: (i) the human capital approach [32, 41], which measures program benefits as the sum of direct treatment costs (for illness) saved plus the value of increased production (in the work-for-pay labor force) attributable to the program; and (ii) "**willingness to pay** (WTP)", which values program benefits according to what prospective program "beneficiaries" would be willing to pay in return for receiving those benefits. The WTP technique has certain conceptual advantages relative to the human capital methodology, in that it includes the perceived value of leisure and nonmarket production as well as the **quality of life** and other indirect benefits of health program investments [18, 34].

These advantages come at a practical price, however. Measures of willingness to pay generally require either that careful population-based **surveys** be performed to collect sample estimates of benefit, or that intended beneficiaries' preferences be inferred by examining their choices in real-life situations in which health and money are "traded off". The "revealed preference" approach is exemplified in surveys of airline passengers, regarding their willingness to pay for improved airline travel safety [19] and for improved air quality [39]. Examples of the "revealed preference" approach include the classic work by Viscusi [38], which inferred the implicit value of human life by comparing the wage premium demanded for jobs at different levels of occupational health risk. Similar safety choices that have been analyzed include the use of automobile safety belts [4] and the decision to purchase new cars with improved safety features [2].

The Cost–Benefit Analysis Algorithm: A Geometrical and Numerical Example

The logic of cost–benefit analysis is displayed graphically in Figure 1. Consider the case of a small local health plan deciding whether to contract with an independent information systems company for its information technology support of its patient care arrangements with hospitals and physicians. The contract is for one year, and the present (year 0) value of the plan's total assets is \$2 million. The plan expects additional net revenues (revenues minus costs) next year of \$1 080 000 from the additional transaction processing efficiencies estimated from this one-year contract.

The curve BDE, labeled "project opportunity set", depicts the set of all projects available to the plan for investment. The "capital market line", drawn as CDF, reflects the tradeoffs for borrowing and lending (rates of return and interest rates) available for investing in comparably risky projects. The revenues and costs are quite risky for this project, in light of the control given up by "outsourcing" this traditional insurance function, so the plan assigns a 20% discount rate to the contract.

The project will require \$400 000 in initial investment (represented as a movement from point B to point A, drawing down plan assets from \$2 million to \$1.6 million), to cover the costs of canceling existing contracts for this information systems support function and the incremental costs of hiring staff to monitor the new arrangement. As shown in Figure 1, this proposed project would add an

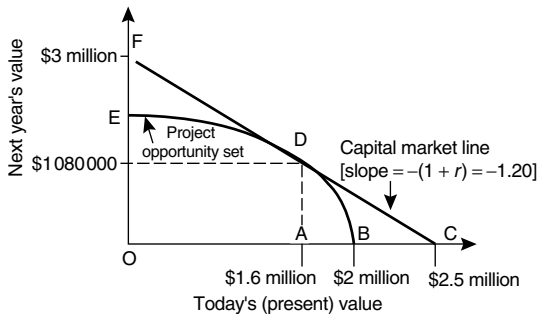

Figure 1 The logic of cost–benefit analysis

estimated \$500 000 to the net value of the health plan's assets. This is represented by the horizontal distance between the plan's original position (point B, pre-investment) of \$2 million and the ending position (point C) of \$2.5 million. This amount equals the net present value (NPV) of benefits less costs. Put another way, based on the plan's estimates, the company's next year value would be \$3 million, or \$600 000 more in year one terms than would have been the case if the plan's assets were simply invested in the capital markets for a 20% rate of return [\$3 million − 1.2(\$2 million) = \$0.6 million]. This \$600 000 in increased "future (year 1) value" is equivalent to receiving \$500 000 (\$600 000/1.2) in increased value today (year 0).

Next, consider the example of a local community not-for-profit hospital deciding whether or not to acquire 49% ownership interest in a large multispecialty medical group that is preeminent in its service area (*see* **Hospital Market Area**). The hospital's total assets are presently worth \$400 million. Acquisition of the group practice would not compromise the hospital's not-for-profit status, but would require an investment of \$60 million cash by the hospital and lease arrangements with the medical group for use of imaging and laboratory technology. The hospital estimates that the proposed lease contract represents a subsidy to the medical group (hospital costs greater than revenues recouped by the hospital) of \$20 million, for a total net investment of \$80 million in present value. The expected rate of return to the hospital on comparably risky investments is 15%, which the hospital's chief financial officer chooses as the discount rate for the costs and benefits of the hospital's investment in the medical group.

The hospital takes the "provider's" perspective in this cost–benefit analysis. The future revenues (benefits) are estimated at \$20 million per year for 25 years (the expected "economic life" of this investment), and annual costs to the hospital of supporting the medical group (e.g. with information systems, health plan contracting support, billing services) are estimated as \$5 million. Thus, the "net benefit" of this project, measured as total benefits minus total costs discounted to their present value, is represented as

$$\text{net present value benefit} = \sum_{t=0,1,\ldots,n} \frac{\text{benefits in year } t \text{ minus costs in year } t}{1/(1+\text{annual discount rate})^t},$$

summed over years $t = 0$ (now), $1, 2, \ldots, n$,

where n is the terminal year of the project. (1)

In this case the estimated net present value (NPV) is

$$\text{NPV (in \$millions)} = \left\{ \frac{\sum t = 1, \ldots, 25(20-5)}{(1.15)^t} \right\} - 80(\text{the time "0" net investment})$$

$$= 15(6.4641) - 80 = \$16\,962\,236. \quad (2)$$

Thus, based on this cost–benefit analysis, the investment should be undertaken because the present value of the net benefits of the project is positive. The figure of 6.4641 is termed the "annuity factor", and it represents the value today (time zero) of \$1 paid in each of n years (= 25 years in this case) invested at a rate of return r (= 0.15 in this case). The formula for calculating this factor is $\{(1/r)[1 - (1/(1+r)^n)]\}$.

In the case of publicly funded programs with widely dispersed beneficiaries and many implicit (rather than direct monetary) costs, it is likely to be more difficult to isolate the annual stream of costs, benefits, risk, correct discount rate, and the "economic life" of the program, but the principles remain the same.

Cost–Effectiveness Analysis

Cost–effectiveness analysis compares the incremental medical costs and health outcomes of alternate health care programs. In contrast to cost–benefit analysis, the denominator of the cost–effectiveness ratio represents health effects expressed in natural units (e.g. life-years gained (*see* **Person-years of Life Lost**), days free of symptoms, cases avoided) rather than monetary units. Valuation of outcome using monetary units favors those with greater income, to the extent that health outcomes are a "normal good", the value of which increases with income [28]. The cost–effectiveness approach to economic evaluation avoids the somewhat controversial monetary valuation of improved health outcomes such as lives saved. However, it should be noted that – in effect – even cost–effectiveness analysis requires an *implicit*

monetary value of health outcomes; otherwise, the decision maker would not know at what level to set the cost–effectiveness threshold for minimally acceptable projects. Under most conditions, results from cost–benefit and cost–effectiveness methods lead to similar conclusions [31]. The interest in the cost–effectiveness model currently stems from its broader acceptance within the health care field and, perhaps, from a working assumption that health decision makers generally operate under externally imposed budget constraints that effectively *fix* the threshold for minimally acceptable projects [23].

A maximization hypothesis underlies the method of cost–effectiveness analysis. Health outcomes are maximized for a given level of medical resource input [12]. The analytic perspective for cost–effectiveness analysis should be that of the health care decision maker, as it is the decision maker who seeks to maximize aggregate health benefits given a budget constraint. This perspective, however, has been criticized as inconsistent with the theoretic axioms of welfare economics, and not in the interest of society [13, 18]. Whereas cost–benefit analysis is strictly based upon a compensation test such as Kaldor–Hicks, cost–effectiveness analyses do not always result in a desirable reallocation from gainers to losers, unless the threshold for minimally acceptable investments is set according to the net benefit maximization rule [12].

The types of questions appropriate for cost–effectiveness analysis include the following:

1. Which of two drug products is most cost-effective for the treatment of major depression [35]?
2. Is heart transplantation a cost-effective strategy [37]?
3. Are work-site intervention programs for hypertension cost-effective [24]?

At the heart of the cost–effectiveness method is the determination of the average and marginal ratios of costs and effectiveness [40]. The average cost–effectiveness ratio is the net cost of each program divided by its measure of effectiveness, resulting in an estimate of the cost of the intervention per unit of outcome gained (e.g. cost per case avoided, or cost per life-year gained). The marginal cost–effectiveness ratio shows the costs and effectiveness of one program in relation to the alternate program.

Specifically, the marginal cost–effectiveness ratio is defined as the difference in medical care costs (net costs) over the difference in program effectiveness (net effectiveness) when comparing at least two alternatives. The marginal cost–effectiveness ratio can be expressed as

$$\frac{TC_a - TC_b}{O_a - O_b},$$

where TC is the total direct medical care costs associated with the intervention, O is the health outcome associated with the intervention, a is program a (new), and b is program b (existing level of care).

Cost–effectiveness analysis is most useful when there are multiple health programs with a common measure of effectiveness, thus allowing direct comparison between alternative programs. This is not always possible, for most medical interventions lack a common outcome measure. For example, medical interventions for hypertension involve an outcome measured in blood pressure units, while antibiotic treatments are assessed in terms of cases resolved.

Selection of the comparator is important, as the marginal cost–effectiveness ratio reflects a direct comparison of the new intervention compared to a base case. The cost–effectiveness ratio can vary dramatically depending upon the characteristics (cost and effectiveness) of the base case comparator.

The analytic horizon of a cost–effectiveness study should correspond to the expected period over which program costs and outcomes will be realized. For acute conditions such as treatment for infection, a less than one-year horizon is appropriate. However, programs for treatment of chronic disease (e.g. hypertension, diabetes, or asthma) and strategies for primary prevention (e.g. vaccination or disease screening) require a longer time frame. In those instances in which costs and outcomes extend beyond one year, discounting to adjust for time preference of the cost stream as a consequence of the program is recommended. Discounting nonmonetary clinical benefits is less widely accepted. Current recommendations for the selection of discount rates for economic evaluations of health care interventions range from 3% to 5% in the US and Canada [22].

Discounting health care costs and benefits has intended and unintended consequences. Time preference adjustment, also known as net present value calculation, of cost streams from comparator programs that accrue at different rates and at different time

periods allows for a standardized valuation of the numerator of the cost–effectiveness ratio. One potential down side to discounting is that public health prevention programs (e.g. immunization or health promotion interventions) often have relatively high up-front costs. Monetary savings that offset these costs may not be realized until much later. Thus, some programs may be judged as not cost-effective simply because of the discount factor.

For resource allocation purposes, the cost–effectiveness ratio can suggest that a new program: (i) improves allocative efficiency (same or better outcome at lower costs); (ii) reduces allocative efficiency (same or worse outcome at higher costs); or (iii) is potentially cost-effective (better outcome at higher cost) relative to the comparator. The latter result requires some interpretation by the decision maker as to the amount of additional resources that they are willing to allocate from other sources in order to realize the incremental gain in health outcome [6].

Cost–Utility Analysis

Cost–utility analysis is a special form of the cost–effectiveness model in which the lack of a common outcome measure is overcome by estimation of some composite metric such as the quality-adjusted life year (QALY) (*see* **Quality of Life and Health Status**). This single outcome measure incorporates the effect of the program or treatment on the quality and quantity of life and allows for comparison of a wide array of interventions [36]. The quality adjustment is derived from preference weights or health utility scores. Several direct and indirect approaches have been developed to measure health utilities or preferences for various outcomes (*see* **Health Status Instruments, Measurement Properties of**; **Outcomes Research**).

The QALY is not the only outcome measure used for cost–utility analysis. The healthy-year equivalent (HYE) has been proposed as an alternative to the QALY, in part because of the restrictive assumptions about preference measurement [26]. The acceptance of the HYE for cost–utility analysis remains controversial.

The primary application of cost–utility analysis is in cases in which programs or treatments generally impact the health status of individuals rather than improve survival or some other clinical outcome measure. Most importantly, cost–utility ratios can be used to compare programs and treatments across different disease states.

A Cost–Utility Analysis Example

Suppose that three different medical options are available for the treatment of inoperable stage 3 to 4 nonsmall cell lung cancer. Option 1 is supportive care without the use of chemotherapy agents. Option 2 is a chemotherapy regimen that consists of two concurrently administered agents. Option 3 is a chemotherapy regimen that consists of three concurrently administered agents. Patient survival, preference weights (utility scores), and cost data are depicted in Table 1. These data show that option 2 is more effective than either option 1 or option 3 from the standpoint of **median** survival. However, best supportive care without chemotherapy (option 1) provides better quality of life for patients. Options 2 and 3 are more expensive, owing in part to the additional cost of the chemotherapy agents, than option 1.

When comparing option 2 to a baseline of option 1, the incremental cost–utility ratio is $[(9985 - 4639/214 - 112) \times 365]$, or \$19 130 per QALY gained. Option 3 compared to option 1 yields an incremental cost–utility ratio of $[(6606 - 4639/165 - 112) \times 365]$, \$13 546 per QALY gained.

Table 1 A cost–utility example

	Treatment options		
Parameter	Option 1, best supportive care	Option 2, two-drug regimen	Option 3, three-drug regimen
Median survival (days)	112	214	165
Preference weight	0.61(±0.22)	0.34(±0.30)	0.34(±0.30)
QALY	0.187	0.199	0.154
Cost (\$)	4639	9985	6606

The Use of Cost–Effectiveness and Cost–Utility Studies by Decision Makers

How are cost–effectiveness and cost–utility analyses used to augment resource allocation decisions? The possible outcomes of a cost–effectiveness or cost–utility study comparing program or treatment A to program or treatment B are illustrated in Table 2. When the overall cost of A is less than B and the health outcomes associated with A are greater than for B, then A is considered to be "dominant", and should be adopted by providers and purchasers as they improve efficiency in the delivery of care. On the other hand, if A is more costly and provides reduced health benefits when compared to B, the new technology should be rejected. Most new programs or treatments are not consistent with the previous two examples. The third row of Table 2 shows the cost–outcome relationship of most new medical technology or programs. Health benefits are improved at some incremental cost for program A compared to program B. Clinicians, patients, and payors must decide whether the improvement in health outcome is "worth" the additional costs. Tradeoffs or substitutions must be made in order to finance and make available treatment A. The final possible result of a cost–effectiveness study is one in which A is less costly when compared to B and also is less effective. Again, if a health system or insurance plan were to adopt such a treatment (e.g. some population **screening** strategies), then tradeoffs would have to be made in terms of foregone benefit.

How attractive does a new treatment have to be to warrant adoption and reimbursement? At what level of cost per health outcome gained would decision makers choose to accept and use new medical innovations? Currently, a cutoff point for cost–effectiveness determination remains uncertain. The value of \$100 000 per QALY has been discussed by policy makers as the level below which a new program would be described as cost-effective and, therefore, worth the investment.

Table 2 Results of cost–effectiveness and cost–utility analyses for resource allocation decisions

Cost difference	Outcome difference	Implication
Cost (A) < Cost(B)	$O(A) > O(B)$	Accept A
Cost (A) > Cost(B)	$O(A) < O(B)$	Reject A
Cost (A) > Cost(B)	$O(A) > O(B)$	Tradeoff
Cost (A) < Cost(B)	$O(A) < O(B)$	Tradeoff

Summary

Health economics is a field of specialization within the discipline of economics that addresses the allocation of resources directed to health improvement and the organization (*see* **Health Services Organization in the US**), delivery, and financing of health services (*see* **Health Care Financing**). In recognition of the centrality of efficiency within economics, a set of methodologies for economic evaluation of health care programs and interventions has been developed. These methods include cost–benefit, cost–effectiveness, and cost–utility analysis. The application of these tools to health economic problems, and particularly to resource allocation decisions, is an area of intense interest. For analysts considering the use of these methods, a complete understanding of the role and limitations of each is necessary.

References

[1] Arrow, K. (1963). Uncertainty and the welfare economics of medical care, *American Economic Review* **53**, 941–973.

[2] Atkinson, S.E. & Halvorsen, R. (1990). The valuation of risks to life: evidence from the market for automobiles, *Review of Economics and Statistics* **72**, 133–136.

[3] Auster, R.D., Leveson, I. & Sarachek, D. (1969). The production of health: an exploratory study, *Journal of Human Resources* **4**, 411–436.

[4] Blomquist, G. (1979). Value of life saving: implications of consumption activity, *Journal of Political Economy* **87**, 540–558.

[5] Brealey, R.A. & Myers, S.C. (1986). *Principles of Corporate Finance*, 3rd Ed. McGraw-Hill, New York.

[6] Detsky, A.S. & Nagalie, G. (1990). A clinician's guide to cost-effectiveness analysis, *Annals of Internal Medicine* **113**, 147–154.

[7] Dranove, D. & White, W. Agency and the organization of health care delivery, *Inquiry* **24**, 405–415.

[8] Elixhauser, A., ed. (1993). Health care cost-benefit and cost-effectiveness analysis (CBA/CEA) from 1979 to 1990: a bibliography, *Medical Care* **31**, 1–141.

[9] Ellis, R.P. & McGuire, T.G. (1986). Provider behavior under prospective reimbursement: cost-sharing and supply, *Journal of Health Economics* **5**, 129–152.

[10] Ellis, R.P. & McGuire, T.G. (1990). Optimal payment systems for health services, *Journal of Health Economics* **9**, 375–396.

[11] Fama, E.F. & French, K.R. (1993). Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33**, 3–56.

[12] Garber, A.M. & Phelps, C.E. (1997). Economic foundations of cost-effectiveness analysis, *Journal of Health Economics* **16**, 1–31.

[13] Gold, M.R., Siegel, J.E., Russell, L.B. & Weinstein, M.C., eds (1996). *Cost-Effectiveness Analysis in Health and Medicine*. Oxford University Press, Oxford.

[14] Gray, B.H., ed. (1986). *For-Profit Enterprise in Health Care. An Institute of Medicine Report*. National Academy Press, Washington.

[15] Hicks, J.R. (1939). *Value and Capital, Part 1*. Oxford University Press, Oxford.

[16] Hillman, A.L., Pauly, M.V. & Kerstein, J.J. (1989). How do financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations?, *New England Journal of Medicine* **321**, 86–92.

[17] Hornbrook, M. & Rafferty, J. (1982). The economics of hospital reimbursement, in *Advances in Health Economics and Health Services Research*, Vol. 3, R.M. Scheffler & L.F. Rossiter, eds. JAI Press, Greenwich, pp. 79–115.

[18] Johannesson, M. (1996). *Theory and Methods of Economic Evaluation of Health Care*. Kluwer, Dordrecht.

[19] Jones-Lee, M.W. (1976). *The Value of Life: an Economic Analysis*. Martin Robertson, London.

[20] Kaldor, N. (1939). Welfare propositions of economists and interpersonal comparisons of utility, *Economic Journal* **September**, 549–552.

[21] Klarman, H.E., Francis, O'S. & Rosenthal, G.D. (1968). Cost-effectiveness analysis applied to the treatment of chronic renal disease, *Medical Care* **6**, 48.

[22] Krahn, M. & Gafni, A. (1993). Discounting in the economic evaluation of health care interventions, *Medical Care* **5**, 403–418.

[23] Laupacis, A., Feeny, D., Detsky, A.S. & Tugwell, P.X. (1992). How attractive does a new technology have to be to warrant adoption and utilization? Tenative guidelines for using clinical and economic evaluations, *Journal of the Canadian Medical Association* **146**, 473–481.

[24] Logan, A.G., Milne, B.J., Achber, C., Campbell, W.P. & Haynes, R.B. (1981). Cost-effectiveness of work-site hypertension treatment program, *Hypertension* **3**, 211–218.

[25] McGuire, T.G. Physician agency, in *Handbook of Health Economics*, Vol. 1A, Elsevier Publications, New York, pp. 461–536.

[26] Mehrez, A. & Gafni, A. (1989). Quality-adjusted life years, utility theory, and healthy-year equivalents, *Medical Decision Making* **9**, 142–149.

[27] Mishan, E.J. (1971). Evaluation of life and limb: a theoretical approach, *Journal of Political Economy* **75**, 139–146.

[28] Mishan, E.J. (1988). *Cost-Benefit Analysis*, 4th Ed. Unwin Hyman, London.

[29] Newhouse, J.P. (1978). Plan and market alternatives to the status quo: techniques for managing resource allocation in medical care, in *The Economics of Medical Care*, J.P. Newhouse, ed. Addison-Wesley, Menlo Park. Chapter 6.

[30] Newhouse, J.P. & Phelps, C.E. (1976). New estimates of price and income elasticities, in *The Role of Health Insurance in the Health Services Sector*, R.N. Rossett, ed. *Universities–NBER Series* 27. National Bureau of Economic Research, New York.

[31] Phelps, C.E. & Mushlin, A.I. (1991). On the (near) equivalence of cost-effectiveness and cost-benefit analysis, *International Journal of Technology Assessment in Health Care* **7**, 12–21.

[32] Rice, D.P. (1966). *Estimating the Cost of Illness*. US Department of Heath, Education, and Welfare, Washington, pp. 16–19.

[33] Salkever, D.S., Skinner, E.A., Steinwachs, D.M. & Katz, H. (1982). Episode-based efficiency comparisons for physicians and nurse practitioners, *Medical Care* **20**, 143–153.

[34] Schelling, T.C. (1968). The life you save may be your own, in *Problems in Public Expenditure Analysis*, S.B. Chase, Jr, ed. Brookings Institution, Washington.

[35] Simon, G.E., VonKorff, M., Heiligenstein, J.H., Revicki, D.A., Grothaus, L., Katon, W. & Wagner, E.H. (1996). Initial antidepressant choice in primary care: effectiveness and cost of fluoxetine vs tricyclic antidepressants, *Journal of the American Medical Association* **275**, 1897–1902.

[36] Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal: a review, *Journal of Health Economics* **5**, 1–30.

[37] Van Hout, B., Bansel, G., Habbema, D., van der Maas, P. & deCharro, F. (1993). Heart transplantation in the Netherlands; costs, effects and scenarios, *Journal of Health Economics* **12**, 73–93.

[38] Viscusi, W.K. (1978). Labor market valuations of life and limb: empirical estimates and policy implications, *Public Policy* **26**, 359–386.

[39] Viscusi, W.K., Magat, W.A. & Huber, J. (1991). Pricing environmental health risks: survey assessments of risk-risk and risk-dollar tradeoffs for chronic bronchitis, *Journal of Environmental Economics and Management* **21**, 32–51.

[40] Weinstein, M.C. & Stason, W.B. (1977). Foundations of cost-effectiveness analysis for health and medical practices, *New England Journal of Medicine* **296**, 716–720.

[41] Weisbrod, B. (1961). *The Economics of Public Health: Measuring the Economic Impact of Disease*. University of Pennsylvania Press, Philadelphia.

Further Reading

Newhouse, J.P., Manning, W.G., Duan, N., Morris, C.N., Keeler, E.B., Leibowitz, A., Marquis, S.M., Rogers, W.H., Davies, A.R., Lohr, K.N., Ware, J.E. & Brook, R.E. (1987). The findings of the Rand Health Insurance Experiment: a reply to Welch et al., *Medical Care* **25**, 157–179.

DOUGLAS A. CONRAD & SEAN D. SULLIVAN

Health Services Organization in the US

This article describes several of the key features and components of the health care system in the US.

Integrated Health Systems

The medical care system in the US has many separate components including physicians' offices, nursing homes, hospitals, drug stores, laboratories, and insurance companies. Historically, the US has operated primarily under the fee-for-service system, whereby a physician or other practitioner bills the patient for each encounter or service rendered (*see* **Health Care Financing**). Under this system, the components of the medical care system have usually operated independently.

When various elements of the delivery system necessary for the provision of care are formally interrelated, they are referred to as an *integrated health system*. An integrated health system may own all the components of the system, or it may own some components and contract for the others to achieve a complete system. The degree of integration can vary greatly. Some integrated systems include a direct insurance function, offering packaged insurance benefits to an enrolled population, with all services delivered through the integrated system. Alternatively, an employer or insurer may contract with the integrated system to use the delivery mechanism only.

One early US example of a long-standing, highly integrated health care system is the Kaiser Health Plan of California, which has its own salaried doctors who are usually required to treat patients in Kaiser's outpatient facilities. In Europe, an example is the British National Health Service, which pioneered integration and coordination with the control of resources.

The rapid increase in the cost of medical care has accelerated interest in, and prompted increased development of, integrated health care systems. It is believed that integrated systems promote more efficient and effective health care, in part because comprehensive management information systems permit administrators to monitor the use of services, the referrals to specialists, and access to specific, and especially expensive, services. A large system can obtain discounts on supplies and drugs.

Numerous organizational arrangements exist for structuring the components of an organized system. The essential features of an integrated delivery system are the degree of coordination in its network and its potential for controlling physician and patient behavior.

The best-known structure for an integrated system in the US is the *Health Maintenance Organization (HMO)*. It is so called because each patient pays a premium amount prospectively for all covered services, independent of which services were actually provided, which gives a physician or system an incentive to maintain the patient's health. There are several variants of the HMO model. In a *Closed-Panel HMO*, the HMO owns the outpatient and inpatient facilities and owns or contracts for most other services, is paid a fixed amount for each patient covered (*capitation*), and pays the doctors (the panel) a salary or other predetermined compensation. This model allows a great degree of control over both physicians and patients.

Another common HMO structure is the *Independent Practice Association (IPA)*, which contracts with some or all community-based physicians, hospitals, and other providers for services provided to enrolled clients. More structure is offered by the organizational form termed "group practice without walls" in which community-based physicians form a single legal entity while practicing independently with common office management services and shared contracting.

Another category of integrated system is based on legal and organizational relations between hospitals and physicians. These include *Physician Hospital Organizations (PHOs)* and *Management Service Organizations (MSOs)*. PHOs are usually initiated by hospitals for the development of partially integrated systems of care that can contract with insurance companies and employers. PHOs may work with a restricted universe of physicians or may be more broadly based. The MSO tends to be more physician-oriented than the PHO, is sometimes initiated by physicians, and provides greater practice support services for the physicians who are participating.

Health services research is easier to perform within an integrated health care system than in independent facilities, because there are enrolled populations and often a unified information system. Considerable research has been done to determine whether one organizational system yields higher **quality of care**, higher satisfaction, or lower cost

than other systems. The impact of such systems and differences between for-profit and not-for-profit systems will need a further evaluation as they evolve, although the evidence thus far is generally positive. For a further description of integrated health systems, see Brown [1].

Hospital and Health Systems

Hospitals were first termed "almshouses" or "poor houses" and provided care primarily to the homeless poor and chronically disabled. Wealthier people received care in their own residences. As medicine advanced, particularly after the turn of the century, and most notably after World War II, the hospital's role as a source of biomedical expertise and knowledge grew dramatically. In the US, the development of professional **nursing** and specialized technology and the increasing ability of physicians to intervene in disease and illness spurred on the growth of the nation's hospitals in the first half of the 1900s. The growth of private health insurance and government entitlement programs, and further advances in medical technology and the professional development of physicians in the years prior to and after World War II, gave a further thrust to hospital growth and development. In the US, health services are increasingly provided in an ambulatory setting, causing hospitals to become more a source of highly specialized services.

Hospitals and health systems may be organized and owned as government entities or as private for-profit or nonprofit entities. Hospital ownership and hospital management may be differentiated in that in some instances a hospital may be publicly owned, or owned by a nonprofit entity, but managed under contract by a for-profit corporation. Government hospitals may be owned by federal, state, or local entities. A for-profit hospital is typically part of a larger corporation, as may be the case also for a nonprofit entity. Publicly held, for-profit companies that own and operate or manage hospitals under contract are typically large corporations whose stock is traded on national stock exchanges.

In recent years, the for-profit hospital sector has experienced a high degree of turmoil with, most recently, increasing consolidation. Some controversy exists as to the extent to which for-profit hospitals are run more efficiently and have lower personnel-to-patient ratios. Economic pressures, however, are forcing all hospitals to improve their economic efficiency and management expertise and to focus on parameters of performance.

The traditional organizational structure of hospitals in the US includes three sources of power and authority. The governing board is ultimately responsible for all of the operations of the hospital. Hospital administration is delegated responsibility for the day-to-day management of the facility. The hospital medical staff is typically separately organized with delegated responsibility from the board for clinical matters, including the credentialing of physicians and assessing and assuring the quality of health services provided. Hospitals that are part of larger health systems, however, typically lose managerial and governance autonomy.

Hospitals and health systems in the US are facing increasing competition and cost pressures. Managed care requires an assumption of risk and participation in various new forms of reimbursement that have a variety of controls associated with them. Increasing vertical integration is occurring throughout the US in the hospital industry. Concerns over quality of care, malpractice litigation, excess bed capacity, and provision of care to the medically underserved are also common in most communities throughout the US.

Research needs to focus on issues of efficiency, outcomes, and costs of care in the hospital and health systems. Also relevant are issues associated with the integration of the hospital with other services and with systems of care.

Further information on the organization of hospitals in the US can be found in [13].

Ambulatory Care Services

Ambulatory (outpatient) care encompasses those services provided to a noninstitutional patient, as opposed to inpatient services, which are provided to a patient who has been admitted, at least overnight, to a hospital or other health care facility. Ambulatory services include a wide range of settings, professionals, and specific health care clinical services. Technological advances and financial pressures are increasingly leading to a shift of services from inpatient to outpatient care.

The typical US citizen has approximately six physician contacts per year. The most common setting for ambulatory care services is the physician's

office, incorporating solo practitioners, group practices, and hospital outpatient departments. Ambulatory services are also provided in a variety of other settings, including, most notably, ambulatory surgery centers, which have grown tremendously in importance and are now the setting for 70% of all surgery performed in the US, emergency rooms, and hospital clinics. Governmentally sponsored ambulatory care services include those in institutional settings such as Department of Veterans Affairs facilities and prisons, as well as military services and the Indian Health Service.

Ambulatory care plays an important role, particularly in managed care plans, in the coordination, organization, and control of all health care. Ambulatory care is usually less expensive than inpatient care.

Research has begun to address key issues such as the role of the gatekeeper in ambulatory care, appropriate use of services, coordination and access, and clinical practices and outcomes (*see* **Outcomes Research**).

Further details on ambulatory care services may be found in [12].

Group Practice

A group practice is a formal organizational arrangement for the affiliation of three or more health care professionals characterized by the sharing of income, expenses, medical records, staff, facilities, and other resources. The first physician group practice in the US was the Mayo Clinic in Rochester, Minnesota. Historically, most physicians were solo practitioners. With the advent of increasing specialization and administrative complexity, and, more recently, of insurance and prepayment, group practice has grown explosively, with approximately 40% of all physicians in the US practicing in a group.

Group practice provides professional management and shared financial and patient care responsibility. Group practice also limits a practitioner's clinical and financial freedom, requiring that practitioners conform to group norms and standards. Personal autonomy is exchanged for greater financial flexibility and contracting advantages.

Group practices may be organized as professional corporations, foundations, partnerships, and other legal forms. Other complex legal entities and contracting arrangements are used in groups that are involved in managed care. Some larger groups own their own hospitals, and most groups own ambulatory surgery, laboratory, and other specialized facilities. Under managed care, group practice assumes a particularly important role in managing physician resources and in controlling patient access to services.

Havlicek [5] provides further description of group practice.

Primary Care

Primary care is the provision of ongoing, day-to-day health care services, encompassing preventive services (*see* **Preventive Medicine**) as well as relatively routine and patient- and provider-initiated services. Primary care typically requires less intensive resources than more specialized care and can often be provided during a brief office visit. Primary care also includes follow-up and continuing care for chronic diseases.

Primary care is typically provided by a physician in the physician's office, but is also provided by other health professionals, such as nurse practitioners, especially in specialties such as pediatrics. Primary care is also available in hospital facilities and the patient's home.

Primary care provides an important entry into the health care system. It is the best setting for ongoing monitoring and coordination of care, and a reliable source of advice and guidance. It is the coordinating and controlling aspect of primary care, combined with increased reliance on primary care providers (i.e. general internists, family practitioners (*see* **General Practice**), pediatricians, and sometimes obstetrician/gynecologists), that is a key defining principle of many forms of managed care.

Wenzel [16] elaborates on the characteristics of primary care in the US.

Long-Term Care

Long-term care includes a broad array of physical health, mental health, and social services provided to individuals with significant, often permanent, illness and disability. In some instances, the need for long-term care may be only temporary, with eventual recovery. Long-term care services, in contrast to acute or short-term care, typically involve a broader array of social, and residential, services, as well as health services. The involvement of social and other

services may present financial difficulties for many individuals due to lack of external subsidies such as health insurance.

Long-term care services include skilled nursing facilities, such as nursing homes; inpatient hospital services, including medical, surgical, psychiatric, and rehabilitation facilities; ambulatory care services, and mental health facilities; alcohol and drug abuse programs; adult day care; home health services; hospice care; and social services, including meals on wheels, homemaker and personal care services; transportation, communication, health promotion activity programs, and recreational activities; and, finally, housing programs, including congregate care, retirement communities, assisted living facilities, and other living arrangements.

Long-term care services are primarily devoted to individuals with chronic physical or mental disability. An important component of long-term care is the rehabilitation service, particularly for individuals suffering from chronic disease, trauma, and accidents. The older population of long-term care users typically have multiple physical and/or mental health problems, as well as various social and financial constraints. A growing population of individuals in both the long-term and mental health systems is characterized by mental and/or physical disability attributable to various forms of dementia, such as Alzheimer's disease.

Nursing homes are an important component of long-term care. Nursing homes that are Medicare-certified are eligible to accept patients covered under the Medicare program. Medicare coverage of long-term care services is extremely limited, and most patients are required to spend-down most of their personal financial resources before becoming eligible for Medicaid program coverage. The nursing home resident is typically aged 85 and above with multiple health, and often mental health, problems and with a variety of dependency requirements.

Hospice is a form of organizing services for individuals with terminal illness. Hospice may be provided in specifically designated facilities or in the patient's home and involves a coordinated, multidisciplinary approach to addressing the patient's needs as well as those of the family.

Home health services is another growth area in long-term care. Technological advances allow a wider range of services, such as infusion therapy, to be provided in patients' homes, thereby decreasing the need for inpatient care. Increasing coverage under Medicare and insurance plans has spurred the growth of home health care in the US.

Long-term care services are often fragmented and need integration to match services to patient needs. Coordination requires integrated information systems (*see* **Administrative Databases**), care coordination, particularly by case managers, and integrated financing mechanisms. Current long-term care financing arrangements in the US limit this type of integration. Social services, in particular, are often inadequately coordinated with physical health and mental health needs. For further reading, see Evashwick [4].

Public Health and Preventive Services

Public health and preventive services are the front line of protection against injury, disease, and illness. Primary prevention and many public health services are population-based, such as the protection of food, water, and milk supplies, and the monitoring of disease (*see* **Surveillance of Diseases**) and disposal of wastes. Preventive services delivered to individuals with the purpose of avoiding illness include vaccinations and immunizations, physician examinations, and screening (*see* **Screening, Overview**). Of increasing importance in recent years is work site accident avoidance (*see* **Occupational Health and Medicine**).

Public health services are provided through state and local public health agencies. The core functions of public health agencies at all levels of government (see [7]), are: assessment, development, and assurance of public health services. State agencies have responsibility for the entire population in a state. Local agencies provide direct services, such as restaurant inspections and monitoring of food and water supplies. Provision of personal services such as immunization, venereal disease screening, and family planning clinics is a local function. State agencies intervene when local agencies do not perform adequately legally required public health services. In the US, federal agencies with responsibility for public health include the federal **Centers for Disease Control** and Prevention, which provides laboratory, epidemiologic, and advisory expertise. Federal grant support for selected priorities is provided to state and local agencies. Responsibility for protecting the nation's health is ultimately shared by

governmental agencies with front-line providers and by every citizen as well. Research needs in this area include cost/benefit analysis of screening (*see* **Screening Benefit, Evaluation of**) and routine personal preventive services. There is controversy about the appropriateness and frequency of most preventive measures. Further discussion on this topic appears in [7] and [14].

Mental Health Services

Mental health services involve services provided for psychiatric and neurological disease and illness pertaining to the brain and its function, as well as to emotional and behavioral deviance (*see* **Psychiatry**; **Neurology**).

Until the second half of the twentieth century, mental illness and dysfunctional behavior were treated by institutionalization and isolation as well as persecution. Developmental and experiential origins for behavior, codified by Freud and others, led to the establishment of psychiatry and psychoanalysis to diagnose and treat mental illness. Biomedical research is now vastly improving the identification and treatment of such illnesses as depression, schizophrenia, obsessive–compulsive behavior, and addictions.

Increasingly, physiologic etiologies are being identified for many forms of mental illness and aberrant behavior, leading to enhanced pharmacologic intervention. The introduction of psychotropic drugs in the 1950s, along with community-based outpatient services, led to the deinstitutionalization of mental health patients in the US. However, inadequate resources and lack of an integrated and comprehensive delivery system have also caused increases in the homeless population in many cities, multiple hospitalization episodes, and increased criminal activity by and against those with mental illness.

The US has both public and private mental health systems. The public system is the provider of last resort and is characterized by governmental facilities, while the private system cares for individuals with insurance, the ability to self-pay, or coverage under entitlement programs. The private system is characterized by private psychiatric hospitals and a greater role for psychiatrists as opposed to psychologists, who are more prevalent in the public system. Mental health services tend to be limited under private health insurance. Where physiologic origins for mental disorders are identified, prospects for enhanced coverage are brighter, as are the social advantages. However, problems such as developmental disabilities and severe organic brain disorders, dementia, including Alzheimer's disease, substance abuse, and criminal activity remain complex challenges. Mental health problems also raise numerous complex legal, ethical, and moral issues regarding individuals' rights to privacy, to treatment, and to involvement in society, as well as issues of access to care, appropriateness of various professional providers, and avenues for financing.

Research needs include continued epidemiologic investigations of the nature of illness and determination of cost-effective interventions (*see* **Health Economics**), as well as determination of the most appropriate sites for care, best practitioners to utilize, and financing arrangements.

The reader is referred to [6] and [11] for more detailed description of mental health services in the US.

Health Care Personnel

Approximately 8% of US, civilian employment is involved in the health care system as providers of care or in organizing or managing the system. Physicians are the key clinical decision-makers in the health care system. In the US, federal and state government initiatives in the mid-1960s led to substantial increases in the number of medical schools, as well as medical school graduates from existing schools. In addition, during the late 1960s and 1970s, federal policy allowed for an influx of substantial numbers of foreign medical school graduates. The result of these policy actions is a substantial increase in the number of physicians in training and in practice. Geographic dispersion of physician supply has also improved greatly over the past 30 years. Significant attention has been directed toward specialty distribution, to focus recently on increasing the supply of primary care practitioners at the expense of many surgical subspecialities. Managed care has endorsed this shift with an emphasis on providing care through primary care practitioners wherever possible.

There are over 2 000 000 registered nurses in the US. In recent years, there has been a dramatic shift in the education of individuals eligible to become registered nurses from hospital-based diploma programs

to baccalaureate and associate degree programs based in colleges and universities. The nursing field also includes many other roles, including various forms of nursing assistants and licensed practical nurses, as well as more heavily credentialed nurses, such as nurse practitioners.

Numerous other specialty professionals contribute to health services. For example, improvements in oral health combined with greater efficiency in dental practice have impacted demand for dental services and education; the number of dental schools and graduating dentists has begun to decline. However, dental services are still inequitably distributed due to financial constraints.

Research issues for the future are complex and include the increasing role of specialists in many areas, competition between different practitioners for employment and clinical roles, credentialing and licensure of professionals, and matching availability of personnel to the need for such individuals in an increasingly fiscally constrained environment. Issues of quality, cost, and appropriateness of utilization, will also need to be addressed in the future (*see* **Health Workforce Modeling**).

For further information on health care personnel in the US, see [2, 3], and [8].

Managed Care

Managed care is a general term representing the realignment of health care services and reimbursement in such a manner as to shift the risk, both financially and in other forms, from insurers to providers and consumers. Managed care is the more current form of what was previously termed prepaid health care. Although managed care is in a state of flux, some specific organizational structures are beginning to evolve.

Managed care plans are designed to reduce utilization, particularly of inpatient services, and at the same time often to provide a broader benefit structure with some degree of emphasis on preventive services in view of their potential long-term cost benefits. One important feature of managed care plans is the establishment of contractual arrangements with providers to allow the plan and its management to impose various forms of oversight and control over providers. Provider risk-sharing through contractual arrangements that provide incentive compensation is also common. Reduced consumer administrative burdens are typical of many forms of managed care, although various barriers are also introduced to reduce consumer incentives for utilization as well.

Managed care provides services in a more financially constrained framework, whereby both providers and consumers have greater incentive to control use of services and hence costs. Substitution of lower-cost clinical alternatives, rationing of services, coordination of care, reduction of duplication of services, and managerial efficiencies are among the approaches utilized in implementing managed care systems. Use of quantitative databases to monitor and evaluate clinical patterns of care, financial experience, and quality and utilization is an important component of managed care, which drives the need for a structured information system (*see* **Administrative Databases**) and for statistical evaluation methods for assessing use and patterns of care.

Managed care often incorporates various forms of HMOs. The percentage of the population enrolled in managed care plans has increased dramatically in the past decade. In the US, entitlement programs such as Medicare and Medicaid increasingly incorporate managed care principles and contractors to instill efficiencies and cost savings.

Managed care plans, through contractual arrangements with providers, establish networks of individual and institutional care sources. Less restrictive plans, such as Preferred Provider Organizations and Point-of-Service HMO plans use networks but allow out-of-plan use at higher cost. More restrictive forms of managed care, such as Closed-Panel HMOs often do not allow out-of-plan use of services. Other incentives and controls affecting consumers include copayments and deductibles, case management, benefit limitations and exclusions, and access barriers in various forms.

An increasingly popular mechanism for controlling utilization by consumers in managed care plans, particularly HMOs is the gatekeeper. The gatekeeper is a primary care physician who must either provide or approve referrals for any services within the plan. The gatekeeper concept has assigned much greater responsibility to the primary care physician and has reduced direct access to specialists by consumers.

In managed care plans, financial incentives for providers are often designed to provide rewards for the careful management of dollars. Various forms of incentives, such as bonuses and profit-sharing pools,

are used to encourage physicians to control carefully the use of services, particularly expensive inpatient and specialty care. Reimbursement of physicians has also shifted in many plans from traditional fee-for-service payment to salary and capitation, whereby incentives are much more clearly focused on rationing and control of utilization.

Further research is needed to test many of the principles of managed care and their long-term effects on cost, quality, and patient and provider satisfaction. Possible underutilization, lack of access, and adverse effects of financial incentives for physicians are other research issues.

Kongstvedt [9, 10] provides a good source for further reading regarding managed care.

Regulation and Controls

Regulatory mechanisms to control health care services may be imposed by governmental entities or by payers under contractual arrangements. Government intervention is usually designed to protect health and safety or to influence the health care market when the market fails to achieve those social goals desired by political forces.

Examples of US governmental regulations pertaining to health and safety include fire, health, and safety codes imposed by state and local governments. State regulation of health care personnel includes licensure of physicians, nurses, and other categories of professionals. Payers may also implement limited regulation of this nature, such as evaluation of participating physicians' qualifications and credentials.

Marketplace regulations in the US, particularly by government, date back to the Hill–Burton legislation, which allocated federal funds for hospital construction and renovation after World War II on the basis of simple health planning computations. The more recent era of government intervention dates to the Great Society in the mid-1960s. Numerous interventions attempted to influence costs and allocation of resources. Examples include subsidies aimed at individuals and institutions, such as grant and loan program tax exemptions. Entitlement programs such as Medicare and Medicaid represent large-scale, subsidy-type interventions. Restrictions on entry into professional fields through licensure requirements and facility licensure and capital expenditure controls are additional examples of interventions. Regulation through payment mechanisms includes requirements under the Medicare and Medicaid programs, rate-setting commissions, wage and price controls, payment restrictions under insurance programs, including contractual arrangements under managed care, and the determination of fee schedules.

Controls to assure the **quality of care** are also common in health services. These are usually associated with entitlement programs or with contractual obligations under insurance plans, particularly under managed care. These mechanisms have included professional review organizations, utilization review, preadmission authorization, second opinions for surgery, and other programs to assess quantitatively various aspects of the quality and utilization of services, and the control of use of services by providers and patients.

Numerous other regulatory mechanisms have been utilized in the past or are currently in place. These include: financial controls on consumers, including benefit limitations, deductibles, coinsurance exclusions, and other provisions of insurance plans; limitations on the supply of services through rationing, queues, and other restrictions on access; reviews of provider services, including claims reviews, medical audits, institutional reviews; and legal, regulatory, and practice-influencing effects from medical malpractice litigation.

Regulation and control of health services has historically had a focus on either affecting utilization and costs or influencing perceived inadequacies and misallocations of resources within the health care system. Although many regulatory efforts in the past have failed to provide adequate results or have simply not been cost-effective, marketplace mechanisms continue to be the primary focus at the present time with an emphasis on reduction in cost increases, on influencing patient expectations of behavior, and on affecting physician practice patterns and use of resources. Regulatory control mechanisms are increasingly focusing on economic considerations with some added focus on monitoring various aspects of the quality of care provided. Evaluation of consumers and providers by managed care organizations and self-evaluation of such managed care organizations themselves and by external organizations have grown substantially in recent years. Under the pro-competitive market approach in the US of recent years, the federal government's

direct role in the regulation and control of health services has focused primarily on costs, access, and quality issues in government entitlement programs with a less substantive contribution to the assessment of various aspects of the larger system.

The reader is referred to [17] for more details regarding regulating and controlling health services in the US.

Technology Assessment

Advances in technology can improve the diagnosis, treatment, and cure of disease and illness. However, important issues related to the diffusion and evaluation of technology impinge on policy-making regarding the role of technology in health services.

The evaluation of technology involves complex considerations including costs and benefits, regulation, efficacy, and clinical effectiveness. Diffusion of new technology is driven by financial considerations. Under fee-for-service reimbursement, technological advances have a tendency to be utilized as quickly as regulatory approval is achieved, while in a managed care environment the potential for some hesitation exists.

The principal federal agency responsible for technology assessment in the US is the **Food and Drug Administration (FDA)**. Drugs and medical devices must be demonstrated to be safe and efficacious to achieve FDA approval for clinical application. This complex and controversial process requires considerable time and financial resources.

Clinical research studies are usually necessary to determine the appropriate clinical situations when each technology should be utilized. Technologies that lead to overall cost reductions in health services due to their substitution for more expensive therapies are of particular research interest. Managed care organizations are interested in the appropriateness of various technologies and their associated costs. Controversy is building over the potential restriction of some technologies under insurance and entitlement programs due to costs and limited benefits. Ultimately, rationing of resources necessitates making judgments as to whether particular technologies are warranted in individual cases.

Further details may be found in Skorup [15].

References

[1] Brown, M. (1996). *Integrated Health Care Delivery*. Aspen, Gaithersburg.

[2] Bureau of Health Professions, Health Resources and Services Administration (1993). *Factbook: Health Personnel U.S.* (DHHS Pub. No. HRSA-P-AM-93-1). US Government Printing Office, Washington.

[3] Council on Graduate Medical Education (1995). *Seventh Report to Congress and the Department of Health & Human Services Secretary: Recommendations for Department of Health and Human Services' Programs*. US Government Printing Office, Washington.

[4] Evashwick, C.J. (1996). *The Continuum of Long-Term Care: An Integrated Systems Approach*. Delmar, Albany.

[5] Havlicek, P.L. (1996). *Medical Groups in the U.S.: A Survey of Practice Characteristics*. American Medical Association, Chicago.

[6] Howard, K.I., Cornille, T.A., Lyons, J.S., Vessey, J.T., Lueger, R.J. & Saunders, S.M. (1996). Patterns of mental health services utilization, *Archives of General Psychiatry* **53**, 696–703.

[7] Institute of Medicine (1988). *The Future of Public Health*. National Academy Press, Washington.

[8] Institute of Medicine (1996). *The Nation's Physician Workforce: Options for Balancing Supply and Requirements*. National Academy Press, Washington.

[9] Kongstvedt, P.R. (1997). *Essentials of Managed Care*, 2nd Ed. Aspen, Gaithersburg.

[10] Kongstvedt, P.R. (1996). *The Managed Health Care Handbook*, 3rd Ed. Aspen, Gaithersburg.

[11] Robins, L.N., Locke, B.Z. & Regier, D.A. (1991). An overview of psychiatric disorders in America, in *Psychiatric Disorders in America*. Free Press, New York.

[12] Ross, A., Williams, S.J. & Pavlock, E.J. (1997). *Ambulatory Care Management*, 3rd Ed. Aspen, Gaithersburg.

[13] Rowland, H.S. & Rowland, B.L. (1992). *Manual of Hospital Administration*. Aspen, Gaithersburg.

[14] Scutchfield, F. & Keck, C.W. (1996). *Principles of Public Health Practice*. Delmar, Albany.

[15] Skorup, T.E. (1994). Technology assessment and management, in *The AUPHA Manual of Health Services Management*, R.J. Taylor & S.B. Taylor, eds. Aspen, Gaithersburg.

[16] Wenzel, F.J. (1994). Primary care services, in *The AUPHA Manual of Health Services Management*, R.J. Taylor & S.B. Taylor, eds. Aspen, Gaithersburg.

[17] Williams, S.J. & Torrens, P.R. (1993). Influencing, regulating, and monitoring the health care system, in *Introduction to Health Services*, 4th Ed. Delmar, Albany.

STEPHEN J. WILLIAMS

Health Services Data Sources in Canada

The system is referred to as "national" because plans are linked by the federal government's Canada Health Act principles.

Similarly, the provinces and territories have separate health information systems to serve their particular purposes. These are a complex assortment of components operating in different places and at different levels, often quite independently. However, through agreements, the provinces and territories, federal government departments, and other organizations have instituted national health information databases and registries, maintained according to national standards. These databases do not yet form a coherent whole, but they are accessible to decision makers, planners, epidemiologists, researchers, and others to improve health services, and ultimately the health of Canadians (*see* **Administrative Databases**).

In 1989, the Conference of Deputy Ministers of Health, concerned about the limited and fragmented nature of health information, approved the establishment of the National Health Information Council (NHIC). In 1990, a National Task Force on Health Information was created to assist the NHIC with identifying health information needs. In addition, the Task Force was asked to develop priorities and organizational structures to bring about improvements and changes. Following consultations, the Task Force recommended establishing a Canadian coordinating council for health information.

These are available from the CIHI, two federal government departments – Statistics Canada (STC) and Health Canada (HC) – and a smaller but important special purpose agency, the Canadian Centre for Occupational Health and Safety (CCOHS). Each of these agencies is described briefly in the section "National Health Information Organizations in Canada". There are other data sources in the provinces/territories but national series and comparisons are mostly available through the national agencies.

Framework for Health Information

The evolution of health information has mirrored paradigm shifts in the view of health care – largely in-patient acute care and physician administrative data at first, and over time moving towards information across the continuum of care and on a broader definition of health. Canadian health care information consists mainly of information from government and hospital sources. Canada has standardized and comprehensive national databases, and these have the potential for much further development. Currently, the availability of data is most complete in the areas of hospitals, identifiable diseases or conditions, and utilization and costing. Surveys such as NPHS and CCHS are being used to explore nonmedical determinants of health. For convenience in identifying current data sources and development activities for the future, Canadian health information is categorized into three main areas.

1. *Health determinants*. The factors that influence or determine health. They include the environment, human biology, lifestyles, behaviors and risk factors, demographics, occupation, and socioeconomic factors. Examples are satisfaction with job, cigarettes smoked, proportion of aged, elderly below low income cutoffs, and labor force participation, or lack of it.
2. *Health services*. Services, interventions, and systems (whether public or private sector) allocated for restoring, maintaining, or improving health. These are subdivided into morbidity, health human resources, environmental and **occupational health**, and financial and operational data areas. Examples include mortality rates, readmission rates as well as physician to population ratios, numbers of health care facility beds and financial indicators. (*see* **Health Services Research, Overview**).
3. *Health status/population health*. Objective and subjective measures – including morbidity, disability, **life expectancy**, and **vital statistics** – of the health and well-being of populations as diagnosed by health care professionals or reported through self-assessment. Examples are life expectancy and **infant mortality**.

The remainder of the article discusses these broad areas of health information, listing major data holdings and indicating specific examples of information that may be derived from the sources. In addition, a sample of Canadian initiatives underway to widen the scope and fill the gaps in current information are described.

Health Determinants

Data for health determinants are mainly derived from general or special purpose population surveys, both periodic and occasional (*see* **Surveys, Health and Morbidity**). After the one-time Canada Health Survey in 1978, there was a gap until the mid-1980s when a number of federal and provincial surveys were initiated. These include surveys to provide information on current issues such as **AIDS**, tobacco control, and fitness. Other surveys have health-related components for population subgroups, including aboriginal Canadians, children and youth, disabled persons, seniors, and women. There have also been large provincial population sample surveys in Ontario and Quebec. Many of these activities are continuing.

A major ongoing survey is the biennial National Population Health Survey (NPHS) (1994–95, 1996–97, etc.). It includes a set of core population health status measures, longitudinal data on health determinants, and in its first three cycles, special, periodic **cross-sectional** information. This survey of 17 000 households is conducted by Statistics Canada. The self-reported information helps to monitor the health objectives of the provinces and territories, focusing on conditions responsive to prevention, treatment or intervention and examining the states of good health – not just illness. Specific survey categories are **health status**, use of health services (*see* **Health Care Utilization Data**), determinants of health, and demographic and economic information. The survey also allows the possibility of linking to the national health databases (*see* **Record Linkage**).

In 2000, Statistics Canada began conducting Canadian Community Health Survey (CCHS) as part of the Health Information Roadmap Initiative (see Further Initiatives). The CCHS, provides the basis for producing cross-sectional estimates to address priority health data gaps at national, provincial, and health region levels. The CCHS is composed of two cross-sectional components conducted over a two-year cycle. The first component is a health region-level survey has a sample of more than 130 000 with content adapted to health region needs while the second is a province-level survey in the second year with a specific, in-depth theme. With the introduction of the CCHS, the NPHS has now become a strictly longitudinal survey.

Other surveys which provide information either directly or indirectly on health determinants include: consumer income and expenditure; general social surveys; labor force surveys; the international literacy survey; environmental surveys, etc. These may either originate from or be conducted by Statistics Canada for another client (i.e. other government departments such as Industry Canada).

Examples of health determinants available from population surveys include:

1. Alcohol consumption (Statistics Canada – STC; Health Canada – HC).
2. Exercise frequency (STC, HC).
3. Measures taken to improve health (STC, HC).
4. Nutrition (HC).
5. Risk factors (lifestyle) (STC, HC).
6. Smoking/tobacco control and use (STC, HC).

Health Services

Health services data sources contain data elements on health-related curative, preventive, or promotional services provided to patients or to the general public. Many of these data have personal and/or institutional identifiers, so privacy, **confidentiality**, and security standards must be maintained. The major subcategories of data are morbidity, health human resources, environmental and occupational health, and financial and operational areas.

Morbidity

Morbidity databases include data from services provided in hospitals and those provided by physicians and other health practitioners. The Hospital Discharge Abstract Database (DAD) and its companion Hospital Morbidity Database at CIHI are major service event databases. DAD contains over 85% of all Canadian hospital patient discharges (about 4.3 million records annually). It provides data collection and processing services, reports to facilities, and carries out comparative reporting.

Major health care services data sources include:

1. Canadian Organ Replacement Register (Canadian Institute for Health Information – CIHI).
2. Hospital Discharge Abstract Database and Hospital Morbidity Database (CIHI, STC) – (i) most responsible diagnosis leading to hospital

length of stay; (ii) interventions; and (iii) relative resource use by grouping of cases.

CIHI is responsible for many databases and registries that capture information across the continuum of health care services in Canada. This information supports research and analysis for planning and policy making purposes. A detailed statement of purpose is included in the full description of each data holding. All CIHI's data holdings are subject to strict privacy and confidentially principles set out in *CIHI Principles and Policies for the Protection of Health Information* (PDF) 437 KB. Click on any of the links below for further information on the database, related publications, availability of data, and contact information.

Canadian Joint Replacement Register (CJRR) – captures information on hip and joint replacements performed in Canada and follows joint replacement patient time.

Continuing Care Reporting System (CCRS) – contains demographic, administrative and clinical data for residents in facility-based continuing care in Canada.

Hospital Mental Health Database (HMHDB) – contains demographic and medical diagnosis information for inpatient hospital stays for mental health disorders in Canada.

National Ambulatory Care Reporting System (NACRS) – includes data for all home-based and community-based ambulatory care: day surgery, outpatient clinics emergency departments. Currently contains Ontario emergency data only.

National Rehabilitation Reporting System (NRS) – A national health information system for adult inpatient rehabilitation services.

National Trauma Registry (NTR) – contains demographic, diagnostic and procedure information on all admissions to acute care hospitals in Canada due to injury.

Ontario Chronic Care Patient System (OCCPS) – contains demographic, administrative and clinical data for patients in designated chronic care beds in Ontario Hospital. Beginning in April 2003, these facilities will submit data to the national database (CCRS) and historical data will be converted.

Ontario Trauma Registry (OTR) – contains demographic, diagnostic and procedural data on all admissions to acute care hospitals in Ontario due to injury, plus detailed data on major trauma, and data on all deaths in Ontario due to injury.

Therapeutic Abortions Database (TADB) – contains basic demographic and medical information related to Canadian patients obtaining therapeutic abortions in Canada.

3. Health Promotion Surveys (HC).

Health Human Resources

Health human resources databases (*see* **Health Workforce Modeling**) track data elements related to medical and health practitioners, including numbers graduating and practicing, geographical distribution, services provided, and remuneration. The National Physician Database is a major source of information on the quantity of physician services, their costs, and limited patient information. It receives about 16 million records annually from the provincial/territorial health insurance (Medicare) system.

Major health human resources data sources include:

1. National Physician Database (CIHI).
2. Southam Medical Database (physician demographics) (CIHI).
3. Registered Nurses Database (CIHI).

Environment and Occupational Health

Environmental health databases (*see* **Environmental Epidemiology**) from Health Canada and Statistics Canada track chemical, biological, and physical hazards (*see* **Risk Assessment for Environmental Chemicals**), product safety, medical devices, **radiation** protection, and tobacco control (*see* **Smoking and Health**). Occupational health and safety databases, mainly from CCOHS also provincial departments of labour and worker compensation boards, include practical health and safety information on chemical and other contaminants and hazards, regulations, standards, and guidelines.

Major environmental health data sources include:

1. Environmental monitoring and analysis databases (STC).
2. Environmental health databases (HC).
3. Occupational health and safety databases and information (CCOHS).

Financial and Operational Areas

Financial and Operational databases contain details on both governmental and nongovernmental areas, in particular sectors (including hospitals and residential care facilities) or geographic areas (provinces and territories) or at the national level. The main database is National Health Expenditures which contains data from 1960 to the present by spending category and source of funding. The data originate from diverse public documents, including public accounts and annual reports, and private sector sources.

Major health expenditures data sources include:

1. National Health Expenditures Database (CIHI) – actual, real, and per capita expenditures by sector and category.
2. Annual Hospital Survey (CIHI) – beds per 1000 population by type of care.

Health Status/Population Health

Health status includes demographic information, general health status, health status of population subgroups, and illnesses or conditions by geographic or population groups.

The major sources of demographic information are the **census** (held every five years) and the regular reporting of vital statistics (historically – births, deaths, marriages, and divorces). These programs are managed by Statistics Canada. Vital statistics measures are traditional, if indirect, health status measures and may include such items as the age-specific fertility rate, births by **birthweight**, and age of mother (such as teenage births).

Health status data are derived from mandatory disease reporting systems (*see* **Disease Registers**), from general or special purpose population surveys (i.e. the National Population Health Survey), or from hospital or self-reported morbidity and general mortality databases (see details under the section "Health Determinants" above). Under the leadership of Health Canada, national **surveillance** networks are in place to create a picture of health risks, patterns, and trends across Canada. New early warning systems have been set up to detect **communicable diseases** of public health importance. These new surveillance networks represent combined laboratory and epidemiologic efforts. New surveillance systems are also in place to detect trends and risk factors in noncommunicable diseases. These include: acute coronary syndrome; myocardial infarction; childhood asthma; diabetes; congenital anomalies; breast, cervical, prostate, and brain cancer; perinatal health; and childhood injuries.

Selected data available include:

1. Demographic – birth rates (STC); fertility rates (STC); population size and distribution (STC).
2. General health status – health expectancy (STC); life expectancy (STC); health indicators (STC).
3. Health status of population subgroups – aboriginal health (HC, STC); immigrant health (STC) children's health (HC, STC); seniors' health (HC, STC); women's health (HC, STC).
4. Illnesses and conditions – accidents and injuries/trauma (CIHI, HC); AIDS (HC, also in STC health indicators); cancer (STC, HC); cardiovascular disease (STC, HC); chronic diseases (HC); communicable diseases (HC, also in STC health indicators); congenital abnormalities (HC); disability (STC); hospital mental health (CIHI, STC); hospital morbidity (CIHI, STC); mortality/causes of death (STC); notifiable diseases (HC, also in STC health indicators); therapeutic abortions (CIHI/STC); tropical diseases (HC).

Further Initiatives

In 1999, the Government of Canada launched the Health Information Roadmap initiative, a significant investment to enhance the gathering and sharing of information on the health of Canadians and the health of their health care system. The objective was to ensure the regular dissemination of timely and relevant information needed to enhance the public understanding and debate about issues of health and health care and to provide support to those responsible for developing policies, designing and managing programs and evaluating the health care system.

Health Indicator Conceptual Framework

In order to guide the identification of specific indicators that are primarily intended to support regional health authorities in monitoring progress in improving and maintaining the health of the population

Table 1 Health indicators conceptual framework

	Health status		
Well-being	Health conditions	Human function	Deaths
	Nonmedical determinants of health		
Health behaviors	Living and working conditions	Personal resources	Environmental factors
	Health system performance		
Acceptability	Accessibility	Appropriateness	Competence
Continuity	Effectiveness	Efficiency	safety
	Community and health system characteristics		
	Resources	Population	Health system

and the functioning of the health care system for which they are responsible, CIHI developed a health indicator conceptual framework (Table 1). In addition, the indicators should assist with reporting to governing bodies, the public, and health professional groups. The initial set of indicators was selected through a consultative process involving over 500 individuals including health administrators, researchers, caregivers, government officials, health advocacy groups, and consumers. Ongoing consultation with representatives from across the country ensures that existing indicators meet the needs of stakeholders, and that new indicators are added as new needs emerge and new data become available.

This framework is based on a population health, or determinants of health model. This framework reflects the principle, based on the supporting scientific evidence, that health is determined by a complex interaction of factors, including the social and physical environments, well-being, prosperity, health care, as well as genetic endowment and individual behavioral and biological response.

About these indicators. In September 2000, First Ministers issued a CommuniquØ on Health in which they agreed to provide clear accountability reporting to Canadians, beginning in September 2002. Over the past two years, health ministries from all provinces, territories, and the federal government have been working to select and to report on a set of comparable health indicators to the public.

As part (which map back to the conceptual framework) of their agreement in September 2000, First Ministers identified 14 areas for comparable health status and health system performance indicators reporting:

Health status

1. Life expectancy
2. Infant mortality
3. Low birth rate
4. Self-reported health

Health outcomes

5. Change in life expectancy
6. Improved quality of life
7. Reduced burden of disease, illness, and injury

Quality of service

8. Waiting times for key diagnostic and treatment services
9. Patient satisfaction
10. Hospital re-admission for selected conditions
11. Access to 24/7 first contact health services
12. Home and community care services
13. Public health surveillance and protection
14. Health promotion and disease prevention

National Health Information Organizations in Canada

Canadian Institute for Health Information (CIHI)

CIHI's mandate is twofold. It is mandated to be a national coordinator for the development and maintenance of an integrated health information system in Canada. Also, it provides accurate and timely information needed to: establish sound health policies, effectively manage the Canadian health care system, and generate awareness of factors affecting good health. The Institute was established in 1993 as a nongovernmental, nonprofit agency. CIHI is responsible

for data collection, processing, and analysis in wide areas of health, human resources, health care, and health expenditures. It also develops, promotes, and applies national standards to improve the accuracy and comparability of health statistics. Products and services are described in the Canadian Institute for Health Information *Catalogue of Products and Services* (2002) and on the CIHI Web site (www.cihi.ca)
Canadian Institute for Health Information. 377 Dalhousie Street, Suite 200, Ottawa, Ontario, Canada K1N 9N8. Tel. (613) 241-7860; fax. (613) 241-8120; internet: `www.cihi.ca.`

Statistics Canada (STC)

STC is recognized internationally for its expertise in statistics for all aspects of Canadian life. In health, it is responsible, mainly through its Health Statistics Division, for data in the areas of determinants of health, vital statistics, and health surveys to provide accurate and timely statistical information and analyses about the health of Canadians. Statistics Canada provides information on the health status of the population and other specialized information to diverse clients, including life insurance companies, health care associations, pharmaceutical companies, local health units, federal and provincial policy and program areas, and the general public. Products and services are described in the Statistics Canada web site (www.statcan.ea).
Statistics Canada. Director, Health Statistics Division, Statistics Canada, Main Bldg., Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Tel. (613) 951-1746; fax. (613) 951-0792; email: hd.ds@statcan.ca

Health Canada

Health Canada is the federal department responsible for helping the people of Canada maintain and improve their health. In partnership with provincial and territorial governments, Health Canada provides national leadership to develop health policy, enforce health regulations, promote disease prevention, and enhance healthy living for all Canadians. Health Canada ensures that health services are available and accessible to First Nations and Inuit communities. It also works closely with other federal departments, agencies, and health stakeholders to reduce health and safety risks to Canadians.

The Population and Public Health Branch (PPHB) collects the information on health determinants, diseases, and health services related to national programs on health promotion. Their role is to identify, investigate, prevent, and control disease on a national basis. The Healthy Environments and Consumer Safety Branch promotes safe living, working and recreational environments, and collects information on the environment in relation to human health. The Health Products and Food Safety Branch manages the risks and benefits to health products and food and collects information in this area. The first Nations and Inuit Health Branch collects data and information on Aboriginal Peoples.

The information collected through Health Canada and its multijurisdictional network is used by the department and other agencies and government jurisdictions for prevention, control, and policy formulation.

Health Canada. Internet: www.hc-sc.gc.ca

The Canadian Integrated Public Health Surveillance (CIPHS) program is one of several projects that are part of the Network for Health Surveillance in Canada. CIPHS' mandate is to bring standards-based management of public health data and to develop software applications that will allow for a convergence of the information in existing and new systems so that crucial public health information will be available to health professionals and decision-makers at the local, provincial, territorial, and national levels. CIPHS brings together a strategic alliance of public health and information technology professionals working collaboratively to build an integrated suite of computer and databases tools specifically for use by front-line Canadian public health professionals. These front-line public health workers will be able to more effectively undertake public health action through improved management of information and access to key data elements.
Health Canada. Assistant Deputy Minister, Health Protection Branch, Health Canada, Health Protection Bldg., Tunney's Pasture, Ottawa, Ontario, Canada K1A 0L2. Tel. (613) 952-7454; fax. (613) 957-4180; internet: `www.hc-sc.gc.ca`

Canadian Centre for Occupational Health and Safety (CCOHS)

CCOHS is a federal government agency under the Department of Human Resources Development. It

is an internationally recognized resource in occupational health and safety and in electronic information delivery systems. It provides information and advice about occupational health and safety in order to promote safe and healthy working environments.
Canadian Centre for Occupational Health and Safety. Customer Service, 250 Main Street East, Hamilton, Ontario, Canada L8N 1H6. Tel. (905) 572-4400; fax. (905) 572-4500; internet: `www.ccohs.ca`.

Further Reading

Adams, O., Ramsey, T. & Millar, W. (1992). Overview of selected health surveys in Canada, 1985–1991. *Health Reports* **4**, 25–52 [Cat. 82-003].

Canadian Institute for Health Information (2002). *Catalogue of Products and Services*. CIHI, Ottawa.

Last, J.M. (1995). *A Dictionary of Epidemiology*, 3rd Ed. Oxford University Press, Oxford.

McCullough, R. & Stephens, T. (1995). Status report on completed and planned national and provincial health-related surveys in Canada, 1990–94, in *Health Data Sharing in Canada*. Canadian Institute for Health Information, Ottawa, Appendix 1.

Shah, C.P. (1994). *Public Health and Preventive Medicine in Canada*, 3rd Ed. University of Toronto Press, Toronto.

Statistics Canada (1994). *Statistics Canada Catalogue, 1994; Supplement, 1995*. Statistics Canada, Ottawa.

Statistics Canada (1995). *National Population Health Survey Overview*. Statistics Canada, Ottawa.

Sutherland, R.W. & Fulton, M.J. (1992). *Health Care in Canada: A Description and Analysis of Canadian Health Services*. The Health Group, Ottawa.

Wilk, M.B. (1991). *Health Information for Canada: Report of the National Task Force on Health Information*. National Health Information Council, Health Canada, Ottawa.

Wolfson, M.C. (1995). Social Proprioception: Measurement, Data and Information from a Population Health Perspective, in *Why Are Some People Healthy and Others Not?* Evans , Barer & Marmor , eds. Aldine De Gruyter, New York.

S. TAILLON

Health Services Data Sources in Europe

Describing half a century of development for the more than 20 health systems of Europe, each of which has followed a somewhat different path, is difficult. However, one could summarize by noting that Europe's health systems, in general: (i) experienced a virtually unbridled expansion during the third quarter of the twentieth century; (ii) underwent a process of learning to operate under relative fiscal constraints between the mid-1970s and the late 1980s; and (iii) in the 1990s are experiencing a host of reforms that aim simultaneously at consolidating high but insufficient Equity achievements, and at doing more (Effectiveness) with fewer resources (Efficiency) while involving a broader array of participants in the decision processes (Empowerment) [3].

Europe's health systems are typically described as *Bismarckian* (i.e. predominantly nationwide public schemes to protect individuals against the financial risks associated with illness, through subsidized medical care benefits largely delivered through autonomous agents) or as *Beveridgian* (i.e. medical suppliers accessible to all residents at public expense, often publicly supplied, with out-of-pocket expenses limited – generously so in the early decades). In the Central and Eastern countries – not reviewed here – the appropriate label to describe collective production of, and universal entitlement to, a basic medical package was that of a *Semashko* model. That model nominally supported more preventive interventions than in the West but a considerably small array of high-technology procedures aimed at increasing survival rates in the older age strata. A fourth mixed model with heavier reliance on private insurance for sizable population segments and safety nets for frailer segments applies in a third of the Dutch population and, in a different way, in Swiss cantons. Although there is vocal advocacy for higher private insurance participation, this approach has not been a dominant model in the pattern of Europe's **health care financing** during the second half of the twentieth century.

There are, of course, many variations in European health systems within this broad classification scheme. For example, in Bismarckian France, insurees are required to pay their bills directly and are subsequently reimbursed, whereas in Bismarckian Germany vouchers relieve patients from the necessity of most cash transactions. These differences are more than cosmetic. Faced with broadly similar structural problems, some European public authorities opt for modest doses of reform in the apparent belief that structural adjustments can be painless, whereas others such as Britain choose to restructure on a large scale. Ambitious blueprints for a different system with competition among providers and among payers with a larger role for the insurees (the Dutch Dekker Plan in the late 1980s) appear to have yielded to more modest and politically easier ways using a small measure of competition. Stepwise priority setting, which started in the Netherlands and in the Nordic countries, is gaining ground in larger countries. After several decades of a pull towards public responsibility for medical care financing, a redefinition of the private–public mix is everywhere on the agenda, with population *well-being* still a dominant driver but with a strong concern for cost-effectiveness (*see* **Health Economics**) and greater codetermination.

European *averages* – used by necessity to describe expenditure trends and health states of the population to highlight the spread of real-world dispersions among the 22 countries – are published under the auspices of OECD (Organization for Economic Cooperation and Development)-Europe. Broadly similar measurement is available only for the 22 European country members of the OECD. OECD-Europe in 1997 comprised the 15 Member States of the European Union (formerly referred to as the European Community), which are Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, and the UK, together with the Czech Republic, Hungary, Iceland, Norway, Poland, Switzerland, and Turkey. OECD's membership also includes Australia, Canada, Japan, Korea, Mexico, New Zealand, and the US (these non-European countries are not reviewed in this article).

In the 1990s the institutions governing the European health systems and the mix of incentives and regulations governing them have become more diverse. However, there appears to be an underlying trend towards a separation of finance and delivery. Health services in Europe involve monetary transactions, but by and large they are not an activity like all others and obey a distinct set of principles. Money increasingly follows the patient. Monolithic structures are yielding, as in

Britain, in favor of fund-holding general practitioners or, as in Italy, in favor of hospitals run as still publicly owned but autonomous enterprises. The concern for quality has become pervasive, thereby generating closer monitoring and a growing demand for evaluation. External constraints (mainly of a financing nature, but also strained labor relations or ideologic debates as well as the time required to generate micromanagement models) slow down the transformation process of the multifaceted European health systems. An underlying converging current is, however, greatly facilitated by common external constraints (the need to abide by the discipline of a common currency unit, for instance, prompted Belgium, France, Italy, Spain, and others to accelerate reform proceedings) and by shared ethical principles, notably with respect to children, to ethnic minorities, and to underprivileged segments of the population. The OECD *Health Policy Studies* series (notably volumes 2, 5, 6, and 7) and the CD-ROM statistical compendium *OECD HEALTH DATA 97* embrace in a reasonably comprehensive way the health systems trends in the 22 countries referred to here.

In 1996 the European nations' *measured* effort to finance the range of medical goods and services varied by a 2 to 1 ratio (Germany, at 10.5% of GDP, leading Switzerland and France at the top; Poland and Turkey, at under 5% of GDP, at the bottom, with a sizable concentration at around 8% of GDP). Higher spending in some countries (about $2400 at purchasing power parity – a level of living exchange rate – in Switzerland, $2000 in Germany, and $2000 in France) primarily reflects higher per capita incomes but also higher costs of in-patient care episodes, physician contacts, and medicines. Poland and Turkey, whose *measured* expenditure on medical goods and services is closer to $300 per capita, are OECD-Europe's lowest real income countries. The elasticity of health expenditure to total domestic demand – which translates to a relative citizen and consumer preference for medical goods and services over all goods and services – has been well above unity during the second half of the twentieth century, although financing pressures have slowed this proclivity in most countries. Sizable public deficits and public debts have constrained the management of Europe's health systems in the 1990s while at the same time providing them with an opportunity to restructure.

The main driver of Europe's health systems is not simply a finance engine, but one which also has concerns for quantity and **quality of life**. Gains in **life expectancy** at birth have been greater than three months per year since the 1950s in Europe, with somewhat faster gains for disability-free life expectancy. Potential life years lost – a measure of avoidable mortality below 70 years of age – has shrunk by two-thirds, and morbidity from a number of causes has declined, even though the accelerating demand for services and the fluctuating indicators of patient satisfaction may convey the opposite perception. **Quality of care** and **outcomes** remain important concerns in the European health systems.

Europe's Health Systems Information

The intrinsic complexity of political and social organizations implies that – short of mammoth and virtually unmanageable integrated information systems – the data describing their features are compartmentalized and typically developed in isolated corners of the system. A catalog of data sources is required in each European country. No country releases information on inputs, on throughputs, on financing and on performance indicators, on health status and on outcome measurement, and on **population-based studies** in less than a dozen statistical collections. Multiplied by 22 (not to mention the **World Health Organization's** 50 European members), a combined catalog would reach booklet size. As **evidence-based medicine** and evidence-based health systems are only slowly maturing in most countries, the booklet would soon become a directory. A printout of the *OECD HEALTH DATA* sources and methods facility in hypertext – which ties to the 700–800 **time series** on 29 health systems contained in the software – exceeds 200 normal pages, but even that represents a considerable shortcut for a policy analyst in need of simultaneous access to information on several industrialized countries. Forerunners of the hypertext file have been published in conventional paper format [1, 2].

In particular, *OECD HEALTH DATA* covers macrohealth data pertaining to:

1. **Health status** (life expectancy, potential years of life lost, premature mortality, morbidity, and perceived health status)

2. Inputs and throughputs (health employment, medical education and training, high-technology medical facilities, health R&D, the pharmaceutical industry activity, and trade in medical goods)
3. Medical consumption and practice (average length of stay and admission/discharges by **case-mix** groupers or by **International Classification of Diseases (ICD)** categories, pharmaceutical deliveries by therapeutic classes, ambulatory surgical procedures, and other indicators of medical activity and their prices)
4. Lifestyle and environment (nutrition, nuisances and pollutions, behavioral parameters affecting health, and social protection arrangements)
5. Expenditure on health services and finance (*see* **Health Care Financing**) (total and public, outlays on medical functions and benefits: in-patient care, outpatient services, pharmaceuticals, therapeutic appliances; expenditure by age groups; and sources of funding)
6. Demographic and macroeconomic references related to the composition of the population and the labor force, general education, the national product, public finance, and monetary conversion rates.

A particular feature of the datafile is the inclusion of a hypertext facility that lists for every group of variables the intended content of each series, known deviations from the *standard* definition, and the sources of the data.

Other multicountry datafiles are available to the analyst. The European Office of the World Health Organization (WHO) periodically releases a *HEALTH FOR ALL* software which focuses on public health objectives set in the aftermath of the 1978 Alma Ata Conference. These include 38 targets designed to reduce substantially the toll of premature mortality and largely avoidable morbidity, resulting notably from lifestyle and environmental factors. These targets reinforce the continuous monitoring exercise of WHO in the areas of **communicable diseases** and of cancer and other chronic diseases. WHO supports a network of collaborating centers which collate information in their area of specialty, e.g. the **incidence** of **AIDS**. The total sum of these datafiles makes up a sizable amount of reasonably homogeneous data.

The Nordic Council, an informal institutional machinery set up by Denmark, Finland, Iceland, Norway, and Sweden, pioneered a harmonized development of epidemiologic as well as activity nomenclatures, notably in surgery. *Health Statistics in the Nordic Countries*, specialized publications on medicines and on social protection, and the relevant tables in the *Yearbook of Nordic Statistics* provide a limited but harmonized supply of data with detailed indications on the sources in the five countries.

The Commission of the European Union, chiefly through its Statistical Office (Eurostat), has over time developed data files dealing with policy areas pursued by the European Union or related to these pursuits. One example is a large database on congenital anomalies, Eurocat. Another is a database related to the use of case-mix (**diagnosis related groups** or related approaches) management. Many of these specific datasets are, like those developed under the auspices of the WHO collaborative work, the product of learned societies united by prospective economies of scale in pooling basic data. The European Union has an ambitious data development program on its agenda for the 5-year period 1997–2002 that expands its long-term concern for data on work accidents and occupational diseases, on the mobility of medical professionals, on trade in pharmaceuticals and medical equipment, and on selected areas of prevention, all domains in which the Treaties governing the Union provide the authority.

A few learned societies, like the European Dialysis and Transplant Association or the International Birth Defect Monitoring Association, have, over time, built sizable registries (*see* **Disease Registers**). These reasonably homogeneous datasets are similar to those generated by North American and Australasian bodies. Elsewhere, world associations, such as the International Dental Federation, which created a Working Party with participants from several continents, have instilled a world standard gradually disseminated through the national associations willing to take part in a survey of professional practice. Networks of social insurance administrators and payers and of public hospital managers etc. have instilled a culture in which their domestic reporting procedures are slowly converging.

The bulk of the data developmental effort lies, however, where the power to intervene lies: in public and in private national machineries. Health is virtually not defined anywhere; only its absence is recognized. In much the same way, the characteristics of health professionals and of delivery functions

such as hospital care and the measurement of quality of life attributes are not shared. Referring to one of the most established nomenclatures, the ICD, half a century of practice has not sufficed to erase variations in coding practices, let alone in aggregation. Medical culture is not uniform across countries, nor sometimes within countries. Since cross-national compendia are built from national time series or surveys, important attributes or characteristics may vary among individual entries. Lengthy commentaries are thus required for virtually every single component.

International agencies which cooperate in limiting the costly multiplication of nomenclatures are not immune to the risk of using similar headings for different datasets since the criteria underlying the construction of these datasets may vary. The reporting of drinking habits supplies an illustration: the *Health for All* compendium seeks comprehensiveness and identity of concepts; it borrows the basic data from a distillers' compilation which rests on identical conversion of beer, wine, and spirits into pure alcohol. *OECD Health Data*, on the other hand, seeks consistency between the hundreds of data elements it collates from each country more than cross-national comparability.

The preponderance of national datasets over international ones relates to the obvious policy relevance of the former. Companies develop datasets or purchase them from specialized service outfits because of their contribution to corporate strategy; governments develop sizable datasets to plan, implement, and evaluate incentives and regulatory interventions in their health systems. The statistical instruments developed in European countries much resemble those found in the US, comprising one-off surveys, recurrent surveys, **administrative databases**, and elaborate statistical constructs. Because the governance culture of most countries comprising OECD-Europe has been on the whole less quantitative than that prevailing in North America, and notwithstanding sociopolitical choices conducive to considerably greater involvement of public authorities in the management of health care delivery and in health systems financing, European countries have historically been providing less quantitative information on their respective health systems than what may be readily accessible in the US. The quantity and the quality of information released in Europe has been progressing at a rapid pace, however, since the oil crisis of the mid-1970s forced countries to reappraise the cost-effectiveness of public spending and since corporations operate less and less on captive markets but must live in a competitive environment. The list of these new statistical products is a long one – concerned analysts may turn to the *OECD Health Data* hypertext facility, to other international sources of quantitative information and, increasingly, to *Annual Statistical Yearbooks* or *Statistical Abstracts* of the countries in which they have a greater interest for preliminary indications of what is currently accessible. Furthermore, the stream of information is not drying up and, in several areas, new data collection activities have recently been started in Europe in support of evidence-based management systems.

References

[1] OECD (1985). *Measuring Health Care*. OECD, Paris.
[2] OECD (1993). *OECD Health Systems, Facts and Trends*. OECD, Paris.
[3] Poullier, J.-P. (1997). Public health policies in Europe–doing better and feeling worse, in *Oxford Textbook of Public Health*, Vol. 1, Part III, 3rd Ed. Oxford University Press, Oxford, pp. 275–295.

JEAN-PIERRE POULLIER

Health Services Data Sources in the US

The ability of the health care system to respond to the dynamic needs of a nation, and the adequacy of that response, is governed by the availability of relevant data. The necessary information is used for purposes of planning, and to establish baselines and assess change. Sociodemographic, economic, medical utilization, health care expenditure, insurance coverage, health status, and diagnostic measures are critical health care indices that assist in the determination of the demand for health care by the health service user population. For effective administration of health services, information is also needed on manpower and facility requirements and supply, access to care, satisfaction with care, manpower shortage areas, and financial constraints [9].

In the US, federally sponsored health care **surveys** and other related data systems such as inventories of health care providers have served as the primary data sources to assess the nation's overall level of health and health care needs, and to help identify deficiencies in the health care delivery system. Resultant analytic databases (*see* **Administrative Databases**) have been used to formulate analyses with policy implications, to model the impact of proposed changes in programs, and to evaluate the impact of policies over time. Numerous health surveys have also been conducted by state and local governments to address comparable health system evaluations at the subnational level. Health services information systems have also been developed with funding from private foundations and industries, but their focus is usually quite specific in nature. A brief description of these health services data systems in the US is presented below.

Population-based Surveys: Measures of Health Care Use, Expenditures, Access, and Need

National data on the incidence of acute illness, the **prevalence** of chronic conditions and impairments, the extent of disability, and the utilization of health care services are obtained in the US through the National Health Interview Survey (NHIS). The survey is an annual **cross-sectional** survey of approximately 40 000 households selected to represent the civilian, noninstitutionalized population of the US. Sponsored by the **National Center for Health Statistics (NCHS)**, the sample data measure demographic and socioeconomic characteristics, health status, and the use of health care services. Periodic supplements in the area of utilization, behavior, and health status are used on a rotating basis to collect more detailed information.

The NHIS national core sample also serves as the sampling frame for the Medical Expenditure Panel Survey (MEPS), which replaces the periodic National Medical Expenditure Survey (NMES). The survey is cosponsored by the Agency for Health Care Policy and Research (AHCPR) and the National Center for Health Statistics. This **panel** survey collects data to provide national annual estimates of health care utilization, expenditures, insurance coverage, and sources of payment for the civilian noninstitutionalized population, and for an oversample of policy-relevant subgroups that include the poor and near poor, the elderly, individuals with functional limitations, and individuals predicted to incur high levels of medical expenditures. Data collection for the redesigned MEPS was initiated in 1996, based on a sample of households from the 1995 NHIS. The MEPS survey is conducted annually, and obtains both cross-sectional and **longitudinal data** needed to monitor health care utilization, expenditures, and health insurance coverage continuously and to examine changes over time. The survey has sample size peaks, consisting of 13 000 households at five-year intervals starting in 1997, that satisfy national precision requirements for policy-relevant population subgroups. In the off-years of the survey (e.g. 1998–2001 and 2003–2006), the sample will be reduced in scale to approximately 9000 households, but with sufficient sample for national estimation and for large policy-relevant population subgroups. The survey obtains data necessary to make annual estimates and to model individual (and family-level) health status, access to care and use, expenditures, and insurance behavior over the two-year period [5]. Furthermore, the household data on utilization and expenditures is supplemented by linkage to data from medical providers [8] (*see* **Record Linkage**).

Person-specific data comparable to those collected in the MEPS are also obtained for a sample of the

Medicare eligible population in the Medicare Current Beneficiary Survey (MCBS), sponsored by the Health Care Financing Administration (HCFA). The MCBS is an ongoing longitudinal panel survey of 12 000 individuals selected from Medicare administrative files, that collects health care data covering a three-year period. Household respondents provide data on health care utilization, expenditures, insurance coverage, and health status, which is supplemented by linkage to Medicare Claims Information.

Prevalence data for specific diseases and health conditions and measurements of the nutritional status of the US population are collected in the National Health and Nutrition Examination Survey (NHANES), which is also sponsored by NCHS [20]. Data for NHANES are collected through direct physical examinations, laboratory analyses, and interviews. In the most recent NHANES, completed in 1994, approximately 30 000 persons aged two months and older, were examined in standardized mobile examination centers to obtain a wide range of medical measurements. The measurements include dietary intake, hematologic and biochemical tests, a physical examination and a nutritional assessment. The resultant database allows for the monitoring of national trends with respect to heart disease, diabetes, lead exposure, iron deficiency, and children's growth and development in addition to the nutritional health of the nation.

National estimates of the incidence, prevalence, consequences, and patterns of substance use and abuse are obtained from the National Household Survey on Drug Abuse (NHSDA). This annual survey, sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA), consists of about 18 000 household interviews of the population aged 12 and older, using special procedures to assure privacy and anonymity.

Analytic data on the use of medical services for family planning, infertility, and prenatal care are obtained in the National Survey of Family Growth (NSFG), conducted by the National Center for Health Statistics. The survey collects information from a nationally representative sample of over 8000 women in the child-bearing ages (15–44) on fertility, factors affecting childbearing (such as contraception, sterilization, and infertility), and related aspects of maternal and infant health [15]. The survey is usually conducted approximately every five years, and the NSFG survey design is consolidated with the NHIS, which serves as the sampling frame for the study.

Another survey with a focus on children, the NCHS-sponsored annual National Immunization Survey (NIS, formerly referred to as the State and Local Immunization Coverage and Health Survey), has been designed to produce estimates of early childhood immunization rates [1]. The annual survey consists of a telephone screening interview with 800 000 households each year to identify approximately 32 000 households with children between the ages of 19–35 months of age, in order to obtain more detailed immunization data on this target population.

The **Centers for Disease Control** and Prevention (CDC) sponsors the Behavioral Risk Factor Surveillance System (BRFSS), which is designed to collect state-specific general population data on forms of behavior that are related to the leading causes of premature death. The survey is a general population telephone **surveillance** system, which obtains data of particular interest to state health departments in targeted risk reduction and disease prevention activities [29].

Another source of both national and community-specific population based data on health services utilization, access to care, insurance coverage, and consumer satisfaction will be forthcoming from the household survey component of the Community Tracking Survey (CTS). The survey is being conducted by the Center for Studying Health System Change and is funded by the Robert Wood Johnson Foundation [16]. The study is designed to track changes in the health care system and their effects on care delivery and individuals. The household survey sample consists of 36 000 households to be interviewed in 1996–1997, primarily selected in 60 communities, and includes a longitudinal component with data collection in 1998–1999.

Another set of more targeted person-specific surveys have been designed with a special emphasis on obtaining statistical information on the older population. The Longitudinal Study of Aging, sponsored by NCHS and the National Institute on Aging (NIA), was designed to measure changes in functioning and in living arrangements in a cohort of older Americans. The survey was based on a supplement on aging to the 1984 National Health Interview Survey and consisted of 7500 participants aged 70 or older, who were interviewed in 1984, 1986, 1988, and 1990 [17]. The Health and Retirement Survey,

also sponsored by the National Institute on Aging, is a national panel survey consisting of a sample of individuals who were 51–61 in 1992 and their spouses (7600 households, over 12 600 persons) that are subsequently interviewed every two years over a 12-year period. The survey obtains information on health and cognitive conditions and status, retirement plans and perspectives, health insurance and pension plans, and income and net worth, to facilitate analyses of decisions affecting retirement [14]. In addition, the Asset and Health Dynamics Among the Oldest-Old (AHEAD) Survey, sponsored by the NIA, is a panel study of 10 000 persons born in 1923 or earlier that were primarily identified in the screening of 69 000 households for the Health and Retirement Survey. The survey obtains data on physical and functional health, cognitive functioning, economic status (assets and income), out-of-pocket costs for service use (community and nursing home), and other economic resources, in order to support analyses on the interplay of resources and late life health transitions [29]. Data collected in the AHEAD survey will be linked with information from the National Death Index.

The National Center for Health Statistics collects and publishes data on births, deaths, marriages, and divorces in the US through the National Vital Statistics System. In addition to demographic information, the death certificate data include items on educational attainment, Hispanic origin, and recent improvements in the medical certification information on **cause of death** [21] (*see* **Vital Statistics, Overview**).

A number of national household surveys that have been designed with a primary emphasis on socioeconomic issues also serve as important sources of health care estimates in the US. The Current Population Survey is an annual household survey consisting of approximately 60 000 housing units, sponsored by the Bureau of Labor Statistics and the Bureau of the Census, to obtain national estimates of employment, unemployment, and other socioeconomic characteristics of the general laborforce and the overall population [28]. The survey permits national and regional estimates of health insurance coverage for the US civilian noninstitutionalized population. The Survey of Income and Program Participation is a panel survey consisting of 36 700 households, sponsored by the Census Bureau, to produce national estimates on the economic situation of households, families and persons by detailed demographic characteristics covering a four-year period. The survey has included questions on work disability, functional limitations, and health insurance coverage which allow for the derivation of national population estimates for these health-related measures [28]. National household level estimates of out-of-pocket expenditures for health care can be obtained from the Consumer Expenditure Survey, sponsored by the Bureau of Labor Statistics [30]. The survey has been designed to provide data on the buying habits of American consumers, and consists of interviews with approximately 5000 consumer units each quarter.

Surveys of Health Care Institutions and Providers, and Hospital and Medical Information Systems

Surveys of medical providers and health care institutions both complement and enhance the information on health services utilization that are obtained from household surveys and serve as the primary source of clinical information on diagnostic and therapeutic services provided to patients (*see* **Health Care Utilization Data**). Physician-specific surveys provide information on practice characteristics, perceptions regarding clinical autonomy, scope of care provided, financial incentives derived through association with managed care organizations, and the impact of managed care arrangements on the practice of medicine. Health services utilization data obtained from institutions, such as from hospital discharge records, provides essential information on surgical procedure rates to help inform whether unexplained geographic variation exists for specific conditions. Furthermore, surveys of institutions such as nursing homes provide data for national estimates of institutional health services utilization, related expenses, and sources of payment, further distinguished by characteristics of the facility, including structure, size, certification, staffing, revenues, and expenses.

The MEPS Medical Provider Survey (MPS), sponsored by the Agency for Health Care Policy and Research, reflects a design strategy to enhance data reported by households on health services utilization and related expenditures through a contact with the associated medical providers. Data from the survey will be used to reduce the **bias** in national health care expenditure estimates that would occur if solely derived from household reported data. Individuals enrolled in the Medicaid program, in which financial

transactions occur only between the provider and the state Medicaid agency, and enrollees of managed care plans are often unaware of the total amount billed or how much the provider is paid for the services they received [6]. Furthermore, detailed information on the specific types and intensity of the services provided, such as physician procedure codes (CPT-4s), diagnosis codes (ICD-9s and DSM-IVs), and classification codes for inpatient stays (DRGs), need to be obtained directly from the medical providers (*see* **International Classification of Diseases (ICD)**). To satisfy these design objectives, the annual MEPS Medical Provider Survey targets a nationally representative sample of the physicians, facilities and home health providers that were reported to provide medical care to MEPS household respondents [7].

The National Ambulatory Medical Care Survey (NAMCS), sponsored by NCHS, is a perennial source of statistical data on the ambulatory medical care provided by office-based physicians to the US population. The target population consists of office based visits to physicians engaged in the provision of direct care to ambulatory patients. The survey data collected can be used for research on the use, organization, and delivery of medical care. For the physician practices selected into the sample, information is collected on patient visits, date and duration of visit, patient characteristics, diagnostic and therapeutic services provided, and the disposition and duration of the visit [25]. For the 1992 survey a sample of 3000 physicians was selected, with data obtained from approximately 34 000 patient records.

The National Hospital Ambulatory Medical Care Survey (NHAMCS), sponsored by NCHS, is an annual survey of visits by patients to emergency departments and outpatient departments of non-Federal short-stay or general hospitals. In 1993, utilization data were collected for approximately 36 000 patient visits to emergency departments and 35 000 patient visits to outpatient departments [21]. Non-Federal short-stay general hospitals that have a 24-hour emergency room are also eligible for the Drug Abuse Warning Network (DAWN) sample. The DAWN is an ongoing drug abuse data collection system sponsored by the Substance Abuse and Mental Health Services Administration, which obtains data on drug abuse occurrences that have resulted in a medical crisis or death. The primary objective of the data system is to facilitate the monitoring of drug abuse patterns and trends [21].

National estimates of the utilization of non-Federal short-stay hospitals can be obtained from data collected through the NCHS sponsored National Hospital Discharge Survey (NHDS), from a national sample of the hospital records of discharged patients. Estimates are provided by the demographic categories of the patients discharged, geographic regions of hospitals, conditions diagnosed, and surgical and nonsurgical procedures performed. Measurements of hospital use include frequency, rate and percent of discharges, and days of care and average length of stay [11]. For the 1991 survey, 466 hospitals participated and data were abstracted from about 235 000 medical records.

The Agency for Health Care Policy and Research's Healthcare Cost and Utilization Project (HCUP-3) uses encounter-level administrative data collected by state governments and state hospital associations to create research databases. There are two HCUP-3 inpatient databases. The State Inpatient Database (SID) contains discharge abstract records for all discharges from community hospitals in 17 states, comprising half the discharges in the US. The Nationwide Inpatient Sample (NIS) is a sample of the SID and approximates a 20% sample of US community hospitals. The NIS contains all discharges from 900 hospitals. The HCUP-3 hospital inpatient databases include patient demographics, diagnoses, and procedures, length of stay, hospital charges, expected pay source, and hospital and physician identifiers. The databases are designed to support research in the following areas: variations in medical practice, diffusion in medical technology, effectiveness of medical treatments, quality of health services, hospital economic behavior, impacts of market structure, changes in delivery systems, and impact of state and federal health care reform initiatives. HCUP-3 also includes an Alternative Services Database that contains records for all hospital-based ambulatory surgeries in five states. The Alternative Services Database enables studies examining the shift of health services from inpatient to outpatient settings [10].

The US Department of Health and Human Services sponsors three distinct surveys of nursing homes: the institutional portion of the Medicare Current Beneficiary Survey (MCBS); the National Nursing Home Survey (NNHS), conducted by NCHS; and the National Nursing Home Expenditure Survey (NNHES), conducted by AHCPR. The MCBS includes an annual institutional component; the

NNHS was last conducted in 1995; and the NNHES was fielded in 1996 as part of the MEPS. To complement the 1996 MEPS Household Survey, the National Nursing Home Expenditure Survey collected data from a sample of 800 nursing homes and more than 5000 residents nationwide on the characteristics of the facilities and services offered, expenditures, and sources of payment on an individual resident level, and resident characteristics, including functional limitation, cognitive impairment, age, income, and insurance coverage for calendar year 1996. The survey also collected information on the availability and use of community-based care prior to admission to nursing homes and data on the capacity, staffing, and services provided by the institutions [5, 24]. NCHS also sponsors the annual National Home and Hospice Care Survey, which obtains facility characteristics and patient specific health service utilization information from home health agencies and hospices.

Other related nongovernment data sources of health care providers and institutions in the US include the American Medical Association's (AMA) Annual Physician Survey [4], which obtains information on practice characteristics, patient profiles, hours and weeks worked, professional income, professional expenses, and fees. The Community Tracking Survey (CTS), conducted by the Center for Studying Health System Change and funded by the Robert Wood Johnson Foundation [16], includes a physician survey to obtain data necessary to track changes in service delivery, access, and perceived ability to provide quality care. The sample design complements the CTS household survey, consisting of a sample of 12 600 physicians in 60 communities.

Health Insurance Data Systems

Coverage and Costs

In the US, the population's access to health services is influenced by the presence and generosity of their health insurance coverage (*see* **Health Care Financing**). Population-based surveys such as the MEPS and the NHIS provide critical data on the sources of insurance coverage that characterize the population. The 1997 Integrated MEPS-Insurance Component (IC), sponsored by AHCPR, will consist of interviews with approximately 9200 employers, 300 union officials, and 400 insurers, to obtain supplemental information on the health insurance held by respondents to the 1996 MEPS Household Survey. This linked survey will provide data to support analyses of individual behavior and choices made with respect to health care use and expenditures and insurance coverage (*see* **Health Care Utilization and Behavior, Models of**).

In a complementary fashion, the 1994 National Employer Health Insurance Survey (NEHIS), co-sponsored by the Agency for Health Care Policy and Research, the National Center for Health Statistics, and the Health Care Financing Administration, was designed to obtain national and state-level estimates of the number of employers offering health insurance, their costs, and the coverage and characteristics of their respective health plans. The 1997 MEPS-IC will include an establishment component that conducts interviews at more than 30 000 establishments to obtain national and regional estimates of the availability of health insurance at the workplace. The analytic objective is to derive estimates of the amount, types, and costs of health insurance provided to Americans by their employers [5, 27].

The Community Tracking Survey (CTS) also includes an employer survey to measure changes in employers' offering of insurance, the types of insurance offered, premiums, and employees' share of premiums [16]. The sample design complements the CTS household survey, consisting of a sample of approximately 10 000 employers in 60 communities.

Utilization

In addition to survey data, **administrative databases** such as data on insurance claims provide another mechanism to measure health services utilization (*see* **Health Care Utilization Data**). Claims data are generally gathered and maintained at the patient level in order to report charges and monitor the use of medical services and resources [2]. In the US there are three major sources of claims data: Medicare, Medicaid, and private insurers [19]. As of 1991, Medicare claims are available for all Medicare enrollees in the US. The Medicare claims database includes information on cost, diagnoses, and procedures. Alternately, state-specific Medicaid claims data are available as part of the Medicaid Statistical Information System

(MSIS) for 21 states, although the level of detail provided on diagnosis and procedures varies widely, given the nonmandatory reporting requirements for these data elements. Complete Medicaid diagnosis and procedure coding is available in the Tape-to-Tape Medicaid database, but is limited to four states (California, Georgia, Michigan, and Tennessee). Claims data from private insurers are generally employer-based, and vary in the level of detail provided regarding information on cost, diagnoses and procedures, and enrollment data. Insurers such as Blue Cross/Blue Shield, United Health Care, and Kaiser Permanente maintain comprehensive claims databases, as do commercial vendors such as MEDSTAT/SysteMetrics, Inc. and Shared Medical Systems, Inc. [23]. These claims data also have been used in the conduct of cost analyses of clinical practice guidelines.

Health System Inventories and Related Federal Program Data

The US Department of Health and Human Services now obtains data on the level, characteristics, and distribution of the **health workforce** and the physical capital in the health system through a number of separate inventories and surveys, with several more in the early planning stages.

The Health Resources and Services Administration (HRSA) has developed and maintained the Area Resource File System (ARFS), which is designed to be used by health professionals seeking consistent, current, and compatible information for conducting research on the nation's health care delivery system [13] (*see* **Health Services Organization in the US**). The Area Resource File System consists of four major components: (i) the Area Resource File (ARF) which is a county-specific database that consolidates many disparate data elements useful in the analysis of health professions issues and developments on a geographic basis (*see* **Small Area Estimation**); (ii) a State/National Timeseries database; (iii) a microcomputer data series containing demographic, health facilities, and health professions data extracts for use on microcomputers; and (iv) detailed hospital data files. This data system provides the necessary information to allow for research and analysis of the geographic distribution and maldistribution of health manpower, the analysis of health manpower supply, utilization, requirements and cost, and the development of long-range **forecasts** of the health profession's supply and requirements [13].

The National Health Provider Inventory was developed and is maintained by the National Center for Health Statistics to provide counts of the number of health care facilities such as nursing homes and board and care homes in the United States. It also includes an inventory of all home health agencies and hospices in the US, and has served as a **sampling frame** for more detailed surveys of these facilities and agencies [26]. The inventory was last conducted in 1991. The American Medical Association maintains a master file containing data on physician specialty and current employment status for nearly every physician in the US [4]. Hospital-level inventory information is obtained in the American Hospital Association's annual survey of all non-Federal hospitals in the US [3]. A biennial inventory of mental health organizations and general hospital mental health services (IMHO/GHMHS) is maintained by the Substance Abuse and Mental Health Services Administration.

The Health Care Financing Administration maintains the Medicare Statistical System (MSS), which provides data for examining the program's effectiveness and for tracking the eligibility of enrollees and the benefits that they use, the certification status of institutional providers, and the payments made for covered services. The MSS consists of four distinct databases: the health insurance master file, containing demographic and benefit utilization data for Medicare enrollees; the service provider file, which contains information on hospitals, home health care agencies, skilled nursing facilities, clinical laboratories, and suppliers of outpatient physical therapy services that participate in the Medicare program; the Hospital Insurance (HI) claims file, which includes information on the beneficiaries' entitlement and use of benefits for hospital, skilled nursing facility, and home health agency services; and the Supplementary Medical Insurance (SMI) payment records file, which provides information on whether the enrollee has met the deductible and on amounts paid for physician services and other SMI covered services and supplies [12, 21]. The Health Care Financing Administration also compiles estimates of health expenditures on an annual basis by type of expenditure and source of funds in their National Health Accounts [18].

Other administrative data systems with a health services focus are maintained by Federal government departments outside Health and Human Services, in order to satisfy program-specific objectives. The Department of Defense maintains several health-related data systems within the Office of the Deputy Assistant Secretary of Defense. One such system, the Defense Enrollment Eligibility Reporting System (DEERS), has information on eligibility for medical, dental, and other related benefits on approximately 13.8 million uniformed services beneficiaries [22]. The Defense Department also maintains the Defense Medical Information System (DMIS), which contains patient data with data elements comparable to those found on the Uniform Hospital Discharge Data Set or the UB-82 [2]. The clinical and administrative data in DMIS on all inpatient episodes at Defense Department facilities are obtained for the Automated Quality of Care Evaluation Support System (AQCESS).

The Department of Veterans Affairs (DVA) maintains four main health related data files: the Patient Treatment File (PTF), which includes patient-specific claims type data (admission date, diagnosis, and procedures) for care received at VA facilities; the Out Patient Care file (OPC), which includes patient specific outpatient utilization data; the Long-Term Care Patient Assessment Instrument (PAI) file, which contains patient-specific demographic, treatment, and diagnostic data for residents of DVA hospital intermediate medicine wards or nursing home units; and the Annual Patient Census (APC) file, which contains utilization data on patients in DVA hospitals at the end of the fiscal year [2].

Summary

In totality, the set of health service information systems that are available in the US are quite comprehensive in their capacity to measure the demand for health services by the US population, and to assess the ability of the health system to satisfy that demand. Future efforts at health service information system expansions at the federal level will be directed to a broader systems view that allows for characterization of the health system as a whole, the analysis of interactions between supply and demand, and the analysis of the relationship between capacity, functioning of the system, and cost. Such information will allow modeling of the impact of change in one aspect of the system on others (e.g. the interaction of the private and public health systems under various health reform scenarios). Similarly, a stronger focus on systems-wide or community perspectives will allow for analysis of the overall structure of the system in terms of regionalization, organization, and redundancy [5].

References

[1] Abt Associates, Inc. (1994). *State and Local Area Immunization Coverage Health Survey Final Sampling Plan*. Abt Associates, Inc., Chicago.

[2] Agency for Health Care Policy and Research (1991). *Report to Congress: the Feasibility of Linking Research-related Data Bases to Federal and Non-Federal Medical Administrative Data Bases*. AHCPR Publ. No. 91-0003.

[3] American Hospital Association (1993). *Hospital Statistics, 1993–94 Edition. Data from the American Hospital Association 1992 Annual Survey*. American Hospital Association, Chicago.

[4] American Medical Association (1994). *Physician Characteristics and Distribution in the U.S.*, 1994 Ed. American Medical Association, Chicago.

[5] Arnett, R.A., Hunter, E., Cohen, S., Madans, J. & Feldman, J. (1996). e Department of Health and Human Services' Survey Integration Plan, *American Statistical Association 1996 Proceedings of the Section on Government Statistics*. American Statistical Association, Alexandria, pp. 142–147.

[6] Cohen, J.W., Monheit, A.C., Beauregard, K.M., Cohen, S.B., Lefkowitz, D.C. Potter, D.E.B., Sommers, J., Taylor, A. & Arnett, R.A. (1997). The National Medical Panel Survey: a national health information resource, *Inquiry* **33**, 373–389.

[7] Cohen, S.B. (1996). mple design of the MEPS medical provider survey, in *American Statistical Association 1996 Proceedings of the Section on Government Statistics*. American Statistical Association, Alexandria, pp. 152–157.

[8] Cohen, S.B. (1997). The redesign of the Medical Expenditure Panel Survey – a component of the DHHS Survey Integration Plan, Seminar on Statistical Methodology in the Public Service, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, 211–249.

[9] Cox, B.G. & Cohen, S.B. (1985). *Methodological Issues for Health Care Surveys*. Marcel Dekker, New York.

[10] Elixhauser, A. (1996). *Clinical Classifications for Health Policy Research, Version 2: Software and User's Guide*. AHCPR Publ. No. 96-0046.

[11] Graves, E.J. (1994). National Hospital Discharge Survey: annual summary, 1992, in *Vital and Health Statistics Series 13: Data from the National Health Survey No. 117*. DHHS Publ. No. (PHS) 94–1779.

[12] Health Care Financing Administration (1988). *Medicare Statistical File Manual*. HCFA Publ. No. 03272.

[13] Health Resources and Services Administration (1994). *The Area Resource File (ARF) System: Information for Health Resources Planning and Research*. Office of Health Professions Analysis and Research, OHPAR Report No. 4–94.

[14] Heeringa, S. & Conner, J. (1995). *Technical Description of the Health and Retirement Survey Sample Design*. Institute for Social Research, Ann Arbor.

[15] Judkins, D.R., Moser, W.D. & Botman, S. (1991). National Survey of Family Growth: design, estimation and inference, in *Vital and Health Statistics Series 2, Data Evaluation and Methods Research, No. 109*. DHHS Publ. (91–1386).

[16] Kemper, P., Blumenfeld, D., Corrigan, J., Felt, S., Grossman, J., Kohn, L., Metcalf, C., St. Peter, R., Strouse, R. & Ginsburg, P. (1996). The design of the Community Tracking Study: a longitudinal study of health system change and its effects on people, *Inquiry* **33**, 195–206.

[17] Kovar, M., Fitti, J. & Chyba, M. (1992). The Longitudinal study of aging, *Vital and Health Statistics, Series 1, No. 28* DHHS Publ.

[18] Lazenby, H., Levit, K.R. & Waldo, D.R. (1992). National health accounts: lessons from the U.S. experience, *Health Care Financing Review* **14**(4),. Health Care Financing Administration, Washington.

[19] Mitchell, J.B. (1995). Cost analysis of clinical guidelines: which data to use and how to find them, in *Conference Proceedings: Cost Analysis for Clinical Practice Guidelines*. AHCPR Publ. No. 95-0001.

[20] National Center for Health Statistics (1994). Plan and operation of the Third National Health and Nutrition Examination Survey 1988–94, in *Vital and Health Statistics Series 1: Programs and Data Collection Procedures No. 32*. DHHS Publ. No. (PHS) 94–1308.

[21] National Center for Health Statistics (1995). *Health, United States, 1994*. DHHS Publ. No. (PHS) 95–1232.

[22] Office of the Deputy Assistant Secretary of Defence and Defence Medical Support Systems Center (1990). *Program Fact Book*.

[23] Paul, J.E., Weis, K.A. & Epstein, R.A. (1993). Data bases for variations research, *Medical Care* **31**, 96–102.

[24] Potter, D.E.B. (1996). e MEPS National Nursing Home Survey: design and preliminary round 1 field progress, in *American Statistical Association 1996 Proceedings of the Section on Government Statistics*. American Statistical Association, Alexandria, pp. 158–163.

[25] Schappert, S.M. (1994). National Ambulatory Medical Care Survey: 1991 summary, in *Vital and Health Statistics Series 13: Data from the National Health Survey No. 116*. DHHS Publ. No. (PHS) 94–1777.

[26] Sirocco, A. (1994). Nursing homes and board and care homes: data from the 1991 National Health Provider Inventory, in *Advance Data from Vital and Health Statistics, No. 244*. National Center for Health Statistics.

[27] Sommers, J. & Chapman, D. (1996). mpling issues for the 1997 MEPS-IC, in *American Statistical Association 1996 Proceedings of the Section on Government Statistics*. American Statistical Association, Alexandria.

[28] US Bureau of the Census (1996). *Current Population Survey, March 1996 Technical Documentation*. Administrative and Customer Services Division, Microdata Access Branch. Washington.

[29] US Bureau of the Census (1996). *Data Base News in Aging*. Federal Interagency Forum on Aging-Related Statistics.

[30] Walden, D., Miller, R. & Cohen, S. (1994). Comparison of out of pocket health expenditure estimates from the 1987 National Medical Expenditure Survey with the Consumer Expenditure Survey, *Journal of Economic and Social Measurement* **20**, 139–158.

ROSS ARNETT

Health Services Research, Overview

Health services research (HSR) is the "multidisciplinary field of scientific investigation that studies how social factors, financing systems, organizational structures and processes, health technologies, and personal behaviors affect access to health care, the quality and cost of health care, and ultimately our health and well-being" [12]. In both basic and applied research, HSR aims to increase knowledge and understanding of the structure, processes, and effects of health services for individuals, families, organizations, institutions, communities, and populations [9, 12]. The origins of health services research can be traced to the 1920s in the US, and several experts since the 1970s have developed descriptions or definitions of the field [3, 4, 9, 13, 14, 17]. HSR has grown most prominent in the 1990s in both the US and abroad, where it is sometimes referred to as health systems research, or simply health research.

Definitional Concepts

Several important concepts set the field of health services research apart from other academic or clinical disciplines. First, HSR is multidisciplinary in that it involves a wide range of disciplines, clinical specialties, and distinct academic fields; those who work in health services research tend to be identified not by their academic training but rather by the nature of the research they conduct. Core areas generally include clinical specialties (e.g. medicine, nursing, dentistry, pharmacy, social work, and public health), economics (or **health economics**), epidemiology, statistics, and biostatistics. Depending on the research or policy question at hand, however, a considerable range of fields can play major roles in HSR, including anthropology, bioengineering, business administration, computer sciences, decision analysis, ethics and bioethics, history, law, management sciences and administration, psychology, **operations research**, and sociology.

Secondly, HSR involves investigations into basic questions of the behavior of individuals, organizations, and systems within health care; more commonly, HSR comprises applied studies concerned with practical questions of health policy, health care delivery and management, evaluation of health care interventions, and the use of information for public and private health care decision-making. Thirdly, HSR directly generates new or better knowledge about this range of topics, and it also contributes to conceptual, theoretic, and methodological structures by which empirical work can be framed, conducted, and interpreted. Fourthly, HSR is concerned with issues of health services that are broadly defined and involve populations (i.e. members of groups defined by sociodemographic characteristics, health conditions or diagnoses, cultural or ethnic factors, geography or geopolitical jurisdictions, or public or private health insurance plans); it is not focused solely on personal health care for individuals.

Finally, HSR is an expansive field that can include clinical studies (*see* **Clinical Trials, Overview**), **outcomes research**, and health technology assessment; it is sometimes characterized as boundary-crossing [2] when multiple fields, disciplines, and methods are brought to bear on a single question. HSR is distinguished, however, from basic biomedical research and clinical investigation in that it is concerned more with the effectiveness of health care interventions (what works in health care in average or day-to-day practice and health care delivery) than with their efficacy (what works and how safely in ideal settings or controlled trial circumstances). Thus, for many decisions about allocation of resources in the health care sector and day-to-day clinical practice, HSR often provides the critical information that biomedical research cannot [10].

Topics Addressed by Health Services Research

The breadth of health services research is explained by the fact that the field endeavors to understand and improve all aspects of the processes and outcomes of health care delivery and to overcome significant problems of making high-quality health care available to all members of a given society at an affordable cost to that society. The costs and the quality of health care have been the subject of study for the longest period (several decades); more recently, HSR has also been concerned with access to care, health care reform and restructuring of public- and private-sector health care systems, computer-based and electronic communications and information systems, and the size and changing roles of the health care workforce.

Costs of Health Services

The costs of health care and public (e.g. national) and private (e.g. individual) levels of expenditures on care have long been an important area of investigation in health services research (*see* **Health Care Financing**; **Health Economics**). Most basic are studies of the total expenditures on health care, often described in terms of the percentage of gross domestic product (GDP) of a country that is devoted to health care. The effects of various elements of private or public health insurance, including the impact of so-called cost-sharing (coinsurance and deductibles), have been a major area of research; the most prominent HSR investigation of these issues was the Health Insurance Experiment, conducted by the RAND Corporation in the 1970s and 1980s [15]. Related issues concern what services or benefits are included in health insurance packages, how insurance is priced, how health insurance plans reinsure themselves against catastrophic loss, and how insurance plans should be regulated. Because consumer choice of health insurance plans can significantly affect how well and how extensively health care costs are shared across healthy and sick individuals, and even undermine the basic idea of insurance, HSR has directed considerable attention to biased (i.e. adverse or favorable) selection of risk and **risk** (or **case-mix**) adjustment techniques. HSR also involves studies of who pays for what portions of the cost of different types of services (such as health care for physical ailments, as contrasted with mental or emotional disorders, or sociomedical problems such as substance abuse). Most recently, approaches to consumer-designed, individualized health insurance have been the object of HSR investigation in the US.

Organization of Health Care

Closely tied to questions of the costs of health care are issues of how health care delivery is organized and financed (*see* **Health Services Organization in the US**). Health services researchers investigate a wide range of ways in which to structure health care systems: for example, national health systems (or universal national or provincial health insurance), systems in which some portions of a population are enrolled in publicly supported health plans or insurance schemes, private-sector approaches based largely on fee-for-service reimbursement, and private-sector entities of various sorts that are characterized as health maintenance or managed care organizations. HSR illuminates how the structure of health care delivery systems affects the practices and performance of clinicians and of persons seeking or obtaining care, and it documents how different organizational structures and ways of reimbursing health care facilities or clinicians pose incentives for inducing or constraining the provision of services. It is also concerned with the effects of different attempts to control national health care spending through various regulatory controls and use of competition and free-market principles. More recently, HSR has examined aspects of "health care reform", such as the shift in the US from a fee-for-service to a prepaid, capitated, or managed care orientation, and the movement in countries with national health systems to introduce various aspects of private-sector health care delivery or insurance.

Quality of Care and Satisfaction with Care

An important component of HSR is the study of how populations and individuals can obtain efficacious, effective, appropriate, competent, and compassionate health care services – in short, high-quality health care. **Quality of care** has been defined as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge" [6]. HSR aims to identify problems with quality of care, such as overuse of unnecessary or inappropriate services, underuse of needed and appropriate care, and good or poor technical and interpersonal care. It measures the structural aspects of care (e.g. professional credentials or characteristics of facilities), processes of care (e.g. what is done to and for patients and consumers), and outcomes of care (e.g. death, disease, disability, or discomfort). Investigators in this field also study patient safety (or medical errors) and patient or consumer satisfaction with health care amenities, delivery system procedures, and/or outcomes. Of particular note in the US has been the development of the "CAHPS" (Consumer Assessment of Health Plans Survey) family of surveys. HSR studies that combine issues of costs and quality are said to be concerned with the "value" of health care.

HSR also contributes to the measurement and improvement of the quality of health care by providing data collection and analysis tools for programs

of quality assurance and continuous quality improvement or total quality management. Because some of these programs rely heavily on gathering and disseminating information to patients and consumers, HSR has contributed to the design of reliable, valid, and practical means by which information can be obtained, synthesized, and made available in forms such as so-called report cards to purchasers and consumers. Such efforts typically imply comparisons among health care providers and plans, so HSR has been expected to develop techniques by which differences in patient severity of illness, presence of other health problems ("**co-morbidity**"), or other factors can be taken into account. These "**risk adjustment**" questions are considered to pose among the most difficult research questions facing the field in the late 1990s.

To provide adequate guidance on these quality-of-care issues, HSR is also deeply involved in evaluating the clinical effectiveness (e.g. expected benefit of a health care intervention under average conditions of use) of health services. Such studies typically focus on the expected benefits and harms of alternate approaches to prevent, diagnose, treat, or palliate illnesses in different patient populations; they may specifically address the cost–effectiveness of alternate health care interventions [5]. These activities may involve assessing and comparing specific health care technologies (i.e. technology assessment) or developing clinical practice guidelines ("systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [7]). HSR directed at these areas also targets questions of how patients (and their families) and clinicians make treatment decisions (*see* **Health Care Utilization and Behavior, Models of**), the role of shared (or informed) decision-making, and the contributions of medical informatics and decision support systems.

Access to Health Care

Health services research has for decades been concerned with the extent to which individuals can seek and successfully obtain health care when it is needed – in short, access to care, defined as the timely receipt of appropriate care [8]. Among the topics studied are the numerous financial and nonfinancial barriers that confront individuals or groups in gaining access to care. These can include costs (especially for those who have no public or private health insurance), geographic difficulties (travel distances or times to obtain care, especially for persons in rural or frontier areas), **ethnic** and racial factors, cultural and attitudinal barriers, and language or literacy impediments. Investigators study the demographic, cultural, financial, and other factors that influence people to choose among health insurance plans, to seek preventive services and health care, to follow healthy lifestyle or treatment recommendations and regimens, and to acquire information about illnesses and problems. Also of concern to HSR investigators are mechanisms for expanding access to care and the effects on access (and hence health) of the lack or loss of public or private health insurance coverage.

Information Systems, Informatics, and Clinical Decision Making

HSR depends heavily on computer-based health services information systems (*see* **Administrative Databases**). These supply health care providers and researchers with faster and easier access to better and more complete health care information on both individuals and groups than was ever possible in the past. Many different information systems are now available to clinicians and to patients and consumers, and these provide information on clinical problems, practice guidelines, and other data needed to make informed decisions about clinical care. In addition, computer-based systems often include tools to assist clinicians in real-time decision making, such as automatic alerts or reminders at the time of patient visits or when the results of laboratory tests or diagnostic procedures are obtained, and computerized physician order entry for tests or medications. How such clinical decision making tools should be developed, deployed, and evaluated in terms of costs and quality of care are questions of considerable interest to HSR (*see* **Decision Analysis in Diagnosis and Treatment Choice**).

Computer-based information and telecommunications systems also permit individuals to communicate with others about health problems of concern to them and to learn about different treatment options. "Telemedicine," "Telemonitoring," and the rapidly expanding availability of health information on the Internet are additional communications phenomena that expand interactions within the health care sector. The impact of all these resources and communications technologies on the attitudes and behavior of

clinicians and patients or consumers, and ultimately on health care systems, is another important area of HSR.

Health Care Professions and Workforce

Investigators in the HSR field track the supply of and demand for different types of health care professionals and workers, including the development of various types of models that permit educators and policy makers to predict the need for and plan for education and training of health personnel (*see* **Health Workforce Modeling**). In addition, researchers examine how individual and team education and training, professional socialization, and cultural and ethnic background affect practitioner attitudes, behavior, and performance. HSR also concerns itself with ethical and bioethical questions involving the health professions, such as how health care professionals, particularly physicians, reconcile their professional duties to act in their patients' best interests with their responsibility to society as a whole, especially when resources are scarce and economic incentives pose difficult or conflicting obligations.

Methods Used in Health Services Research

Health services research employs virtually all quantitative and qualitative methods found in statistics and biostatistics, economics (*see* **Health Economics**), sociology (*see* **Social Sciences**) and anthropology (*see* **Anthropometry**), psychology, epidemiology (*see* **Analytic Epidemiology**; **Descriptive Epidemiology**), **operations research**, **actuarial** sciences; finance, management, political science, policy analysis, and law. The types of studies done in HSR can include randomized controlled trials (*see* **Clinical Trials, Overview**), a wide array of **quasi-experimental** investigations involving simple or complex **case–control studies, observational studies** and descriptive studies, and community-based demonstrations and evaluations; the **units of analysis** can be nations; regions, states, or provinces; municipalities of all sizes, and communities or neighborhoods (*see* **Small Area Variation Analysis**); groups of individuals defined according to many different sociodemographic (*see* **Social Classifications**), cultural, or health characteristics; health care providers, specified according to type of clinician or facility, health care plan, or setting of care; and families or individuals. HSR places significant emphasis on understanding the end results of health care programs and health care delivery and on obtaining self-reported information on processes and outcomes of care from patients and consumers (*see* **Quality of Life and Health Status**; **Outcomes Research**; **Quality of Care**). The field has generated many reliable and valid instruments for obtaining such information (*see* **Health Status Instruments, Measurement Properties of**) and pursues sophisticated methods research in outcomes measurement [11]. HSR studies employ many sources of information (*see* **Health Services Data Sources in Canada**; **Health Services Data Sources in Europe**; **Health Services Data Sources in the US**), including various types of interviews and questionnaires, focus groups, **surveys** and polls, so-called **administrative data** from various types of computer-based information systems (e.g. insurance billing claims, or hospital discharge abstracts), administrative records of health care programs and plans in the private or public sector, community health information networks, and patient medical records – both paper- and computer-based. Generally, the biostatistical methods required in HSR are similar to those used in biomedical research, except that the sets of variables of interest in HSR tend to be more broadly defined, more concerned with functional and quality-of-life outcomes of interest to patients, families, consumers, and policy makers, and sometimes more difficult or costly to measure than variables of interest in biomedical or clinical investigations.

Major Funders of Health Services Research

Globally, the US funds and produces the great majority of health services research work: of this, the largest portions are supported by agencies of the US federal government. Since the late 1960s, the leading agency was the National Center for Health Services Research (variously titled over the years), a unit of the Department of Health and Human Services (DHHS, formerly the Department of Health, Education and Welfare). In 1989, the Agency for Health Care Policy and Research (AHCPR) was created from this Center; reauthorised and renamed the Agency for Healthcare Research and Quality in 1999, AHRQ continues to

this day to be the central public-sector funding source for HSR. As of 2002, AHRQ had centers focused on evidence-based practice, outcomes and effectiveness research, primary care, organization and delivery of health care, cost and financing of health care, quality of health care and patient safety.

Other DHHS agencies, notably several in the **National Institutes of Health** (especially the National Institute for Mental Health, National Institute for Drug Abuse, and the National Institute for Alcoholism and Alcohol Abuse) and the Centers for Medicare and Medicaid Services (formerly the Health Care Financing Administration) also support projects that fall within the HSR rubric. The US Department of Veterans Affairs has a formal program to support HSR, and elements of the US Department of Defense conduct activities focused on HSR issues such as prevention, quality, or efficient delivery of services. Numerous private philanthropic organizations (foundations) also support research (or demonstrations and evaluations) on HSR topics, especially in areas related to access to care, quality of care, and organization and financing of care; often their focus is on state or local, rather than national or international, issues. Internationally, some governments have programs of health systems research within their national health services (e.g. the UK) or support related efforts in health technology assessment (e.g. Sweden and Canada).

Compared with levels of spending on health care or biomedical research, the support for HSR is small. As of 2000 in the US, approximately $1.35 billion (US dollars) was spent on HSR, a figure that approached only 0.10% of the $1.3 trillion spent that year on health care in that country. Few nations, however, support HSR at these or higher levels.

Personnel Engaged In or Trained In Health Services Research

The number of professionals in the HSR workforce has always been difficult to estimate, for it consists of researchers trained to design, supervise or carry out, and report on HSR work, individuals who assist in such investigations, and users who analyze HSR information or apply HSR for management and policy purposes. A mid-1990s estimate put the number of current health services researchers at 5000, largely in the US; of these, about one-half are trained at the doctoral level and just over one-quarter (mostly physicians) have clinical degrees [9]. This workforce has been trained through many different organizations and programs supported by both public and private funds; only a small minority of these programs are formally established to train individuals at the doctoral level in health services research *per se*.

Professional Organizations and Publications

The most prominent professional organization for health services researchers is Academy Health, a private, nonprofit organization, established in 1981 (as the Association of Health Services Research) and based in Washington, DC. Related organizations include international societies focused on specific areas that HSR studies, including quality of care (International Society for Quality in Health Care) and technology assessment (International Society for Technology Assessment in Health Care). Another global effort is the **Cochrane Collaboration** with centers in Australia, Canada, Denmark, France, Italy, the Netherlands, Norway, the UK, and the US; these centers, like the evidence-based Practice Centers supported by AHRQ in the US, prepare, maintain, and disseminate systematic reviews of the effectiveness of health care, generally using information from randomized controlled trials or other reliable evidence (*see* **Meta-analysis of Clinical Trials**).

Journals available internationally that exclusively or frequently publish on HSR-related topics include *Health Care Financing Review, Health Economics, Health Services Research, Inquiry, Journal of Health Economics, Medical Care*, and *Medical Care Review*; some have been publishing since the 1960s. Newer journals include *Health Services Management Research, Journal of Evaluation in Clinical Practice, Journal of Health Services Research & Policy*, and *Quality of Life Research*. Health policy publications, which also typically feature HSR-related work, include *Health Affairs, Health Policy, International Journal of Health Services, Journal of Health Politics, Policy and Law*, and the *Milbank Quarterly*. Journals with a public health, epidemiologic, or clinical orientation that also publish HSR-related work include the *American Journal of Public Health, Annals of Internal Medicine, BMJ (British Medical Journal), Journal of the American Medical Association, Journal*

of Clinical Epidemiology, Journal of General Internal Medicine, Lancet, Medical Decisionmaking, New England Journal of Medicine, and publications of other professional and clinical societies in the United States and other nations. HSR is often at the core of material published in the journals of international societies, such as the *International Journal for Quality in Health Care* and the *International Journal of Technology Assessment in Health Care*. Several monographs published since 1990 provide substantial overviews of the primary issues that HSR has covered since that time [1, 4, 9, 16].

References

[1] Altman, S.H. & Reinhardt, U.E., eds. (1996). *Strategic Choices for a Changing Health Care System*. AHSR and Health Administration Press, Chicago.

[2] Brook, R.H. & Lohr, K.N. (1985). Efficacy, effectiveness, variations, and quality: boundary-crossing research, *Medical Care* **23**, 710–722.

[3] Flook, E.E. & Sanazaro, P.J. (1973). Health services research: origins and milestones, in *Health Services Research and R&D in Perspective*, E.E. Flook & P.J. Sanazaro, eds. Health Administration Press, Ann Arbor, pp. 1–81.

[4] Ginzberg, E. (1991). The challenges ahead, in *Health Services Research. Key to Health Policy*, E. Ginzberg, ed. Harvard University Press, Cambridge, Mass., pp. 315–331.

[5] Gold, M.R., Siegel, J.E., Russell, L.B. & Weinstein, M.C., eds. (1996). *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York.

[6] Institute of Medicine (1990). *Medicare: a Strategy for Quality Assurance*, Vol. I, K.N. Lohr, ed. National Academy Press, Washington.

[7] Institute of Medicine (1992). *Guidelines for Clinical Practice: From Development to Use*, M.J. Field & K.N. Lohr, eds. National Academy Press, Washington.

[8] Institute of Medicine (1993). *Access to Health Care in America*, M. Millman, ed. National Academy Press, Washington.

[9] Institute of Medicine (1995). *Health Services Research. Work Force and Educational Issues*, M.J. Field, R.E. Tranquada & J.C. Feasley, eds. National Academy Press, Washington.

[10] Lohr, K.N. (1996). The role of research in setting priorities for health care, *Journal of Evaluation in Clinical Practice* **2**, 79–82.

[11] Lohr, K.N. (2000). Health outcomes methodology symposium. Summary and recommendations. *Medical Care*, **38**, II–194 II–208.

[12] Lohr, K.N. & Steinwachs, D.M. (2002). Health services research: an evolving definition of the field, *Health Services Research* **37**, 7–9.

[13] Marshall, J.E. (1985). Introduction, *Medical Care* **23**, 381–382.

[14] Neuhauser, D. (1985). Health services research, 1984, *Medical Care* **23**, 739–742.

[15] Newhouse, J.P. and the Health Insurance Group (1993). *Free for All? Lessons from the RAND Health Insurance Experiment*. Harvard University Press, Cambridge, Mass.

[16] Shortell, S.M. & Reinhardt, U.E., eds. (1992). *Improving Health Policy and Management: Nine Critical Research Issues for the 1990s*. Health Administration Press, Ann Arbor.

[17] Steinwachs, D.M. (1991). Health services research: its scope and significance, in *Promoting Health Services Research in Academic Health Centers*, P. Forman, ed. Association of Academic Health Centers, Washington, pp. 9–19.

KATHLEEN N. LOHR

Health Services Resources, Scheduling

Operations Research had its origins in World War II, where speedy solutions were required for new operational problems, usually of a military kind. The underlying philosophy of innovation and simplicity was adapted after the war for application in industry, where it proved to be of considerable commercial value [8]. It was a natural extension of scope to explore the possibilities of applying it to public services and there was a spate of applications to the health services in the 1960s and 1970s [7] (*see* **Health Services Research, Overview**). Applications to specific organizational problems included the planning of appointment systems, blood bank inventories, ambulance routing, and the control of hospital bed occupancy. In practice the implementation of optimal policies was somewhat limited by political and economic constraints and the emphasis of research in the area changed to more qualitative questions and resource planning at the population level, for example in the assessment of demographic trends [12]. Important though such issues may be, the efficient solution of small-scale operational problems is a significant ingredient of efficient resource utilization. In particular, the unpredictability of demand and patient progress means that the stochastic nature of the processes involved is important at all but the macroscopic level.

It is the purpose of this article to outline one particular area, namely that of the utilization of discrete and equal units of resource, such as beds. When such units are in use we say they are occupied and describe the class of corresponding processes as "occupancy processes", though the latter expression may be applied to the utilization of resources that we would not normally describe as being occupied.

Occupancy Models

Occupancy problems and the models which may be used to study them have certain common ingredients. Generally speaking they will fall naturally into the continuous or discrete time frameworks, the former being more appropriate for activities that fit into the span of a working day, the latter for modeling longer-term activities in which resource utilization at a fixed point of the day is of interest. The mathematics of discrete time models tends to be less elegant but more tractable: rather few continuous time models of any sophistication can be handled analytically. In particular, nonhomogeneity in time is much more easily handled with discrete time models. The ability to permit variation of parameters with time is an important practical consideration; so also is the question of whether we are interested essentially in the short- or long-run properties of a process. The latter can be obtained rather more easily than the former by the derivation of a stationary distribution (*see* **Stationarity**).

The most tractable models have the first-order **Markov** property, which means, in practical terms, that the state of the process must incorporate all the information necessary for predicting its future course. This imposes a severe restriction on the situations that can be modeled simply, though a number of devices for mitigating its effect are referred to below. In the face of analytical difficulties, numerical or **simulation** methods will usually provide solutions quite feasibly, though they bring their own problems of interpretation, particularly with complex models involving many parameters. A simple analytical model, however, even if not particularly realistic in points of detail, may give insight into the structural aspects of a problem that are concealed by more complex formulations.

The models and situations we consider all have available at any given time a fixed number N of units of resource, e.g. beds, operating suites, casualty nurses. These resource units are assumed to be **exchangeable**, i.e. to be equally capable of servicing the requirements of one unit of demand, i.e. one patient in each of the latter examples. In **queuing** theory applications, these resource units would be identified with servers. The occupancy of the system is defined as the number of the units, $X(t)$, that are actually in use at time t; this would often be expressed as a percentage of N.

Arrivals to and departures from the system will typically increase or decrease the occupancy, except that, if $X(t) = N$, then arrivals may form a queue. Queuing processes put emphasis on characteristics such as the length of the queue and the **mean** queuing time. For many health service applications, however, these characteristics are less important than those of the occupancy process, $X(t)$, and its determinants, such as the nature of the arrival and departure

processes. Arrivals and departures will typically be determined by times in a continuous time model, but by counts associated with the fixed time units of the process in a discrete time model. Emergency inputs to the system will result in arrivals determined by points in a **Poisson process** (not necessarily time homogeneous) in the continuous case or by **Poisson distributed** counts in the discrete case. The Poisson process and distribution form important ingredients of all such models, both in theory and in practice.

Some Example Applications

We now consider some areas of application of the general class of models outlined. Each of these areas has been studied by a number of investigators, but not necessarily from a modeling point of view that is consonant with our discussion above.

Appointment Systems

The idea of reducing patient waiting time by providing a realistic schedule of appointments in an out-patient facility or in **general practice** has a long history [1, 2], though most contributions to the subject have been concerned with practical feasibility rather than statistical issues. Stochastic elements of the problem include the possibility of the nonarrival or lateness of scheduled patients, and also the unpredictability of the lengths of times of individual appointments. There is a conflict between the time a patient may expect to wait and the time for which a doctor might be idle; tuning the parameters of the system should achieve a balance consistent with preassigned **utilities**.

Models for the situation are typically formulated in continuous time and use results from queuing theory, although nonstandard assumptions, such as deterministic but unequally spaced arrival times, quickly lead to analytic difficulties; moreover, the stationary theory is not generally useful. Certainly, a queue is an inevitable feature of an appointment system, but interest may focus as much on the final finishing time of a clinic as on the waiting and idle times. Brahimi & Worthington [5] describe a recent application of queueing theory to appointment systems, while O'Keefe discusses practical issues and limitations of theoretic systems [10].

Operating Theater Utilization

This problem has quite a lot in common with that of appointment systems, though the inputs are typically under the complete control of the hospital and patient waiting time would normally be given little weight. Length of service time, i.e. of individual operations, would be the main stochastic element and the optimization criteria might well include the probability that a scheduled operation would need to be postponed. In practice, the exchangeability property of the general model would probably fail and it would be necessary to put particular operations in particular theaters, thus introducing a combinatorial element into the problem.

Casualty Units

Here the units of resource might be regarded as multidimensional, in the sense that waiting may be occasioned by a shortage of doctors, nurses or X-ray facilities, for example. It could well make sense, however, to model a casualty department by treating bed-spaces as the fundamental units of resource; in this case N could be regarded as effectively infinite, the phenomenon of patients waiting on trolleys being a not unfamiliar feature of busy units. A patient in the system could reasonably be assured of emergency attention by a rearrangement of priorities among those present, though the average length of stay within the unit would then be dependent on the occupancy; this would not preclude using a first-order Markov model. Continuous time would probably be a first choice for a casualty unit, though a discrete time approximation would make it easier to incorporate the time heterogeneity that could be expected. All arrivals would effectively be from Poisson processes and the total occupancy would therefore also follow a Poisson distribution under weak assumptions (see below).

In-Patient Wards

The hospital ward provides a classic example of the models we have discussed. Typically, we should expect a mixture of emergency and scheduled admissions. At one extreme an orthopedic surgical facility would have very few emergency admissions; at the other, certain medical facilities may have nearly all their admissions as emergencies – even a general medical ward would probably admit over half its cases as emergencies.

Early models of bed occupancy [14] borrowed the ideas of queuing theory, regarding beds as servers. They were consequently formulated in continuous time; we would argue, however, that discrete-time models are much more appropriate. Thus, it is the number of patients in a hospital ward overnight that is of crucial importance; during the day patients are arriving and departing in a way that gives a certain flexibility of bed use.

The models described in the section below also mostly assume that N is infinite, largely because this permits a number of analytic results of great importance that would otherwise not apply. It may at first sight seem to be an unrealistic feature of the models. However, we can argue that results from such models may provide useful approximations to the finite capacity case if a ward is not too near to full capacity. Moreover, it is often possible in practice to extend the capacity of a ward by using emergency beds or borrowing from other facilities. Such devices would entail organizational costs which can then be quantified by the predictions of the model.

Queuing theory has also been applied, rather more naturally, to the study of waiting lists [13].

Maternity Units

Before the practice of inducing births, admissions to a maternity unit would have been very close to a homogeneous Poisson process. Nowadays admissions are likely to be both of an emergency and scheduled character. In this respect, maternity units have become more like other hospital in-patient facilities, though continuous time might have advantages that do not apply in the general case.

Some Theoretic Results for Occupancy Processes

Generally speaking, occupancy models are analytically much more tractable if we make assumptions of the independent behavior of the individuals in the system. Assuming Poisson inputs, this implies a potentially infinite capacity, which clearly conflicts with reality. However, the models would give a good guide to the behavior of finite systems that are not operating at full capacity and in practice there often are, as argued above, overflow possibilities for emergencies entailing extra costs which can be quantified by the predictions of the model.

Relaxing assumptions such as the infinite capacity or nonindependent behavior of individuals in the system quickly leads to analytic intractability. More sophisticated models can still be studied straightforwardly using simulation methods. However, it is not always easy to extract useful generalizations from large-scale simulations. On the philosophy that modeling is as much concerned with providing an understanding of underlying phenomena as making precise numerical predictions, analytic methods are to be preferred. It is in this spirit that in the following section we summarize simple analytic results concerning the behavior of stochastic occupancy models.

In the following exposition we consider patients occupying beds and moving through a hospital with a number of different wards according to well-defined stochastic rules, though we emphasize that the results would apply to any system with a number of compartments providing units of resource to service the needs of individuals representing units of demand. Similarly, we will, for discrete time models, consider daily time units, though these can also be arbitrarily defined.

Random Inputs

Suppose that patients are admitted as emergencies to a single ward with infinite capacity modeled in discrete time, i.e. the numbers admitted each day are independent **random variables** with a Poisson distribution. If their lengths of stay (LOS) in the system are determined independently from an arbitrary distribution, then the distribution of the occupancy process, or number of beds occupied, also has a Poisson distribution [11]; so too does the daily number of discharges. In the steady state, the mean occupancy is equal to the mean length of stay multiplied by the mean admission rate, a result which we refer to as the mean occupancy theorem and which applies very generally. With suitable modification of the theorem, this important result remains true even when there is day-to-day variation in the mean number of patients admitted per day and in their LOS distributions [4, 9].

If a system has many wards through which patients move independently, then an analogous result applies, i.e. independent random inputs to one or more wards result in Poisson-distributed occupancies of them and, furthermore, the numbers in the different compartments are independent. The same result is obviously

true for hypothetically defined "compartments" which may be defined to represent abstract stages of a patient's progress, for example.

There is a continuous time analog of these results. Specifically, if we have a multiward system with independent inputs to one or more wards resulting from Poisson processes, and if patients move independently through the different wards, then the resulting occupancies are independently Poisson distributed. In particular, we may model a set of hypothetical compartments in each of which a patient spends a length of time, which necessarily has an **exponential distribution**, in such a way as to reproduce any total LOS distribution with a rational Laplace transform [6]. Such models may be useful for modeling flow through a maternity facility, for example, but do not permit the study of any kind of control mechanism.

Deterministic Inputs and Independent Progress

If inputs are deterministic and the lengths of stay, though having an arbitrary distribution, are determined independently of one another and of the occupancy of the process, then the **variance** of the latter is approximately proportional to the **standard deviation** of the LOS distribution [4], with a mean determined by the mean occupancy theorem. This result determines the component of variability in a system that is due to the unpredictability of the length of stay and permits the study of a system in which admissions are planned according to a predetermined pattern of bookings.

General Inputs and Independent Progress

The analysis underlying the above result provides the occupancy distribution resulting from general input distributions and general LOS distributions, which need not be time homogeneous, i.e. may differ on different weekdays, for example. The variance of the occupancy process partitions approximately into components due to the input variance and the variability in the LOS. Analytic expressions are available that completely determine the behavior of an infinite-capacity system in which patients are admitted and progress through the system independently of one another and of the state of the system [4].

Scheduled Inputs

The variance of an occupancy process can, in theory, be reduced by controlling either the admissions or the discharges, though in practice it would be unsatisfactory to determine discharge from a unit merely to control occupancy unless some transfer to an intermediate or predischarge ward were arranged. Scheduling the admissions to bring the occupancy closer to an ideal value, however, is administratively feasible, for example by admitting from a short-notice list a number of patients determined by the occupancy at a given time. The consequential non-independence in the system can be modeled using first-order **Markov chains**. The Markov requirement implies that the lengths of stay have a **geometric distribution**, but there are arguments permitting the relaxation of this rather severe assumption [3]. More seriously, the Markov assumption implies, in the first instance, that patients can be given a maximum of one time-unit of notice, typically 1 day. This restriction can also be relaxed with a non-Markovian process which is a functional of a first-order Markov process and for which the marginal distribution is easily computed [3].

Prediction of Discharge

Using only the current occupancy of a ward to schedule the number of admissions obviously wastes information about the probable short-term future of the process. Predicting the probability of the discharge of individual patients has obvious potential for improvement and can generally be expected to reduce the variance of $X(t)$. The degree of reduction obviously depends on the degree of predictability, but empirical and theoretic considerations suggest that a factor of the order of one-third may be achievable [4].

Discussion

This article has been based on the concept of occupancy, with particular emphasis on bed occupancy. It may reasonably be argued that concentration on beds ignores the demand for other resources, most of which are more expensive than the mere provision of a bed. It is indeed true that this emphasis has led to progressively shorter lengths of stay in hospital and a consequent rise in the cost per bed day. At the

same time, the availability of a bed in an emergency is certainly a key question and deployment of other resources exhibits an elasticity in practice, so that the effects on the system as a whole of an extra patient are not as hard to accommodate as the requirement for an extra bed when a facility is full.

More generally, it may be argued that concentrating on any one unit of resource is clearly to risk oversimplifying the problem, given that in practice we typically require multiple resources, any one of which may be in short supply. Moreover, resources are often interrelated in ways that make independent modeling unsatisfactory. Nevertheless, we need to start somewhere, and it is the object of modeling as much to throw light on general principles and underlying structural relationships as to provide accurate representations or predictions.

References

[1] Bailey, N.T.J. (1952). A study of queues and appointment systems in hospital outpatient departments, *Journal of the Royal Statistical Society, Series B* **14**, 185–199.

[2] Bevan, J.M. & Draper, G.J. (1967). *Appointment Systems in General Practice*. Nuffield Provincial Hospitals Trust, Oxford University Press, Oxford.

[3] Bithell, J.F. (1969). A class of discrete-time models for the study of hospital admission systems, *Operations Research* **17**, 48–69.

[4] Bithell, J.F. (1971). Some generalized Markov chain occupancy processes and their application to hospital admission systems, *Review of the International Statistical Institute*, **39**, 170–184.

[5] Brahimi, M. & Worthington, D.J. (1991). Queueing models for out-patient appointment systems – a case study, *Journal of the Operational Research Society* **42**, 733–746.

[6] Cox, D.R. (1955). A use of complex probabilities in the theory of stochastic processes, *Proceedings of the Cambridge Philosophical Society* **51**, 313–319.

[7] Duncan, I.B. & Curnow, R.N. (1978). Operational research in the health and social services, *Journal of the Royal Statistical Society, Series A* **141**, 153–194.

[8] Hillier, F.S. & Lieberman, G.J. (1990). *Introduction to Operations Research*, 5th Ed. McGraw-Hill, New York.

[9] Huang, X.-M. (1995). A planning model for requirement of emergency beds, *Journal of Mathematics Applied in Medicine and Biology* **12**, 345–353.

[10] O'Keefe, R.M. (1985). Investigating outpatient departments: implementing policies and qualitative approaches, *Journal of the Operational Research Society* **36**, 705–712.

[11] Pike, M.C., Proctor, D.M. & Wyllie, J.M. (1963). Analysis of admissions to a casualty ward, *British Journal of Preventive and Social Medicine* **17**, 172–176.

[12] Weinberg, J. (1995). The impact of ageing upon the need for medical beds: a Monte Carlo simulation, *Journal of Public Health Medicine* **17**, 290–296.

[13] Worthington, D.J. (1987). Queueing models for hospital waiting lists, *Journal of the Operational Research Society* **38**, 413–422.

[14] Young, J.P. (1966). Administrative control of multiple-channel queuing systems with parallel input streams, *Operations Research* **14**, 145–156.

JOHN F. BITHELL

Health Statistics, History of

The field of statistics in the twentieth century (*see* **Statistics, Overview**) encompasses four major areas; (i) the **theory of probability** and mathematical statistics; (ii) the analysis of uncertainty and errors of measurement (*see* **Measurement Error in Epidemiologic Studies**); (iii) design of experiments (*see* **Experimental Design**) and sample surveys; and (iv) the collection, summarization, display, and interpretation of observational data (*see* **Observational Study**). These four areas are clearly interrelated and have evolved interactively over the centuries. The first two areas are well covered in many histories of mathematical statistics while the third area, being essentially a twentieth-century development, has not yet been adequately summarized. Although the fourth area has been going on since man first learned to think inductively, it relies on the state of the art in the first three areas. In this brief survey of health statistics during the past five centuries, emphasis will be given to the development of official health statistics systems in Europe and the US.

Early Interest in Statistics

At the end of the fifteenth century, mathematics was at a rather primitive stage and the threshold of the "scientific revolution" was still two generations away. The mathematics of the Greeks had only re-entered European thinking in the twelfth century, and although some progress had been made in practical applications in navigation and commercial arithmetic, the burgeoning of numeracy was only beginning. Mathematicians still did not recognize the number zero or know how to deal with negative numbers. Except for a few examples of probabilistic thinking such as that in the talmudic literature [10], there was scant evidence of the use of a mathematical approach to probabilities to estimate **risks** or assess the reliability of measurements until the mid-seventeenth century.

Most historians of statistics trace the origins of modern probability theory to the efforts to solve certain gambling problems [e.g. Pacioli (1494), Cardano (1539), and Forestani (1603)] which were first solved definitively by Pierre de Fermat (1601–1665) and Blaise Pascal (1623–1662). These efforts gave rise to the mathematical basis of probability theory, statistical distribution functions (*see* **Sampling Distributions**), and statistical **inference**.

The analysis of uncertainty and errors of measurement had its foundations in the field of astronomy which, from antiquity until the eighteenth century, was the dominant area for use of numerical information based on the most precise measurements that the technologies of the times permitted. The fallibility of their observations was evident to early astronomers, who took the "best" observation when several were taken, the "best" being assessed by such criteria as the quality of observational conditions, the fame of the observer, etc. But, gradually an appreciation for averaging observations developed and various techniques for fitting the observational data to **parametric models** evolved. Many of the founders of modern statistics contributed to the early development of the theory of measurement errors including **Jacob Bernoulli** (1654–1705), **Abraham De Moivre** (1667–1754), **Pierre Simon Laplace** (1749–1827), and **Carl Friedrich Gauss** (1777–1855).

A systematic approach to the collection of data and tabulating observations in a rational manner began with the teachings of Francis Bacon (1561–1626). In his influential treatise *Novum Organum* (1620), he attacked the scholastic philosophy which had developed in the Middle Ages on the basis of the methods of Aristotle. One of the first areas influenced by Bacon's approach was **demography** and **vital statistics** and the social utility of systematic observations is clearly reflected in these early efforts.

The utilitarian nature of statistics is evident in the origins of the word from the Italian *stato* (state), and the original meaning of statistics was a collection of facts of interest to a statesman. Initially such facts were not primarily numerical, but included information on geography, politics, and customs of a region. The compilers of such facts were called statists, a term which survived into the nineteenth century, when the word statistics came to be used for numerical data only, replacing the term "political arithmetic", and the word "statistician" came into vogue.

The Origins of Demography and Vital Statistics

Since ancient times, sporadic surveys of people and property were done to set tax assessments and levies for military service. But after the fall of the Roman empire, regular **censuses** covering an entire state did not occur until the eighteenth century. However, there were intermittent attempts to keep track of the births and deaths in some areas through church records of weddings, christenings, and burials. The City of London was one of the first to regularize the maintenance of such records in 1538, but only within the Church of England. Also at about this time a **surveillance** or early warning system of plague deaths was started in London. To detect the onset of a plague epidemic, parish clerks submitted weekly reports on the numbers and causes of deaths. These weekly *Bills of Mortality* were noted by the authorities who were to take actions if they detected the onset of an epidemic, and by the wealthier citizens for "an indication of when to leave the city for the fresh air of the country" [7]. The weekly bills were published regularly from 1604 until 1842 when they were superseded by reports from the Registrar General.

In 1662, **John Graunt** (1620–1674), a London tradesman who had been active in local politics and intellectual society, published his *Natural and Political Observations Made Upon the Bills of Mortality*, which historians of statistics have referred to as "a remarkable book [12]", "one of the great classics of science [6]", and "a paragon for descriptive statistical analysis of demographic data [7]". Hald summarizes Graunt's contributions to the origins of statistics thus:

> Graunt's critical appraisal of the rather unreliable data, his study of mortality by cause of death, his estimation of the same quantity by several different methods, his demonstration of the stability of statistical ratios, and his life table set up new standards for statistical reasoning. Graunt's work led to three different types of investigations: political arithmetic; testing the stability of statistical ratios; and calculation of expectations of life and survivorship probabilities [7].

At a time when denominator data on the size of the population by age were not available, Graunt used several ingenious lines of reasoning to generate the first **life table** ever published, perhaps his most famous contribution.

Owing to the widespread influence of Graunt's work, bills of mortality similar to the London bills were introduced in Paris in 1667, and soon after in other cities in Europe.

Graunt's life table was brought to the attention of Christiaan Huygens (1629–1695) and his brother Ludwig (1631–1699) who proceeded to develop a probabilistic interpretation of the life table, which was rediscovered independently by **Nicholas Bernoulli** (1687–1759). These investigations, together with the more applied techniques of **Edmond Halley** (1656–1742) based on the births and funerals in the City of Breslau (1693), and the work of Deparcieux (1703–1768) in France who used data from tontines to construct the first correct life tables, formed the foundation of the **actuarial** sciences for life insurance and annuities. These were developed further by Abraham DeMoivre (1667–1754), Thomas Simpson (1710–1761), Benjamin Gompertz (1779–1865), and William Makeham (1826–1891).

It was not until 1766 in Sweden that Per Wargentin (1717–1783) published the first mortality tables for a whole country based on enumerations of the living population as well as on deaths. These mortality tables demonstrated for the first time in a general population that the mortality rate of females was less than that of males.

Graunt's methods of statistical analysis were widely adopted by seventeenth-century statists. **William Petty** (1623–1687), who was a protégé of Graunt, and after Graunt's financial bankruptcy in 1666, his patron, coined the term "political arithmetick" and was one of the founders of the field of political economy. Gregory King (1648–1712) and Charles Davenant (1656–1714) contributed to improvements in the estimates of the population of England. Sebastien de Vauban (1633–1707) described the extent of poverty in France, for which he suffered public disgrace because of its embarrassment of the royal government. Nicholas Struyck (1678–1769) instituted town censuses in the Netherlands and improved the recording of births and deaths. The revelations of statistical data were also used to support religious positions such as the claim of John Arbuthnott (1667–1735), who was a vigorous proponent of political arithmetic, that the stability of the sex ratio "is not the effect of chance but divine providence". Somewhat later, Johann Peter

Suessmilch (1707–1767) in Germany gathered vital statistics from virtually every source then available as evidence of certain tenets of orthodox Lutheran theology. He maintained that the life span (*see* **Life Expectancy**) was constant and that little could be done to improve mortality rates. His work directly influenced the thinking of **Thomas Robert Malthus** (1766–1834). These diverse endeavors eventually led to the establishment of governmental statistical offices in the nineteenth century.

Among the developments in mathematical statistics that occurred during the eighteenth century, two had special relevance for health statistics. **Daniel Bernoulli** (1700–1782), who first developed the **normal** approximation to the **binomial distribution** and used it in studies of the stability of the sex ratio at birth, applied the methods of calculus to mortality rates by treating them as continuous functions. This enabled him to obtain a solution in 1760 to an important public health question of his day: to estimate the impact on life expectancy of eliminating smallpox through a proposed program of mandatory vaccination. His invention of the method of **competing risks**, with some improvement by d'Alembert (1761) and by Makeham (1874), still forms the basic tool for such analyses.

A second development expanded the techniques used by Vauban. Laplace proposed a nonrandom sampling method to estimate the size of the population in 1786. It was based on a notion similar to that of current **ratio estimates**, i.e. that the size of the population of a region was proportional to the annual number of births in that region and that the constant of proportionality could be determined from a purposive sample of subregions. Graunt had used a similar assumption implicitly a century earlier.

Laplace's method was severely criticized, most notably by Baron de Keverberg (1827) [11, p. 164]. These criticisms clearly reflected an appreciation that there were a multitude of factors that could influence any chosen characteristic of a population, that subgroups of the population were not homogeneous with regard to the array of factors influencing the characteristic, and, therefore, purposive samples of the population could not reflect the total population. Only complete censuses of the population would do, and these would have to amass immense amounts of information. At this time there was not yet an appreciation for the power of random sampling methods (*see* **Probability Sampling**).

Applying Statistics to Medical and Social Issues

Just as **demographic** and economic statistics began with the name of "political arithmetic" in the seventeenth century, medical statistics began with the name of "the numerical method" early in the nineteenth century. Although some of his methods were evident in the works of **Phillipe Pinel** (1745–1826) and other French physicians, **Pierre-Charles-Alexandre Louis** (1787–1872) has been described "as the first modern clinician, the man who made bedside medicine a science as well as an art, and who established the principle of learning medicine from thoughtful observation of patients [1]." His studies on the inefficacy of blood letting were the beginning of quantitative medicine and earned him the title of "father of medical statistics" [12]. Louis's hopes for his numerical method were echoed by Giacomo Tommasini (1768–1846) in Italy, and **F. Bisset Hawkins** (1796–1894) in England, who published in 1829 the first English textbook on medical statistics with the rather grand title of *Elements of Medical Statistics; Containing the Substance of the Gulstonian Lectures Delivered at the Royal College of Physicians with Numerous Additions Illustrative of the Comparative Salubrity, Longevity, Mortality, and Prevalence of Diseases in the Principal Countries and Cities of the Civilized World.* Although by later standards Louis's statistical attempts were often inadequate, suffering particularly from sparse numbers, he had a crucial influence on **William Farr** who attended his lectures during his two years in Paris, as did several American physicians who were influential in the early development of public health and epidemiology.

Louis's methods were not immediately accepted for many of the same reasons that Laplace's methods were not: the variability between cases was thought to be highly individualistic and not subject to statistical summarization. For example, **William A. Guy** (1810–1885), who contributed much to public health and occupational statistics, felt "the formulae of the mathematician have a very limited application to the results of observation" [12, p. 151].

The Belgian, **Adolphe Quetelet** (1796–1874), who dominated the field of social statistics for half a century, may have gone too far in the other direction. Impressed by the **central limit theorem** and believing that averages based on large numbers of

observations from a population had remarkable stability, he introduced the concept of the "average man" (*l'homme moyen*) which had considerable popular appeal. He was also enamored of the normal distribution and fitted it to many characteristics, marvelling at the statistical homogeneity of large bodies of data which detracted from further exploration of valid heterogeneities. However, he influenced a large number of statisticians including **Louis Adolphe Bertillon** (1821–1883), Wilhelm Lexis (1837–1914), **Francis Galton** (1822–1911), **Karl Pearson** (1857–1936), and **Ronald A. Fisher** (1890–1962) [11].

Development of Health Statistics in England

During the eighteenth century many physicians and registrars in England recognized the inadequacies of the bills of mortality. There were frequent calls for reforms but because of concerns about personal liberties, religious arguments, and beliefs that population figures were crucial state secrets, it was not until 1800 that Parliament passed a population act that set up the census of 1801. By the 1830s, as in the mid-seventeenth century (with Graunt and Petty), London "witnessed a flash of enthusiasm for vital statistics and political arithmetic" [5, p. 13]. The Statistical Society of London was founded in 1834 by the same group that had founded the statistics section (Section F) of the British Association for the Advancement of Science in 1833, and started publication of its Journal in 1838. These and other early statistical societies in England were greatly concerned with social problems, conducting several surveys to document conditions in England and continuing to push for social reforms long after the surveys proved too expensive to continue. Although they claimed scientific objectivity, these statists were superficial in their use of mathematical methods, paid little attention to the validity or accuracy of their data, but were aware that using numeric data gave credibility to political arguments [5].

A more balanced contribution was made by William Farr (1807–1883) in the area of vital statistics. Starting his career as an unsuccessful London clinician, he quickly became an acknowledged authority on vital and health statistics with a strong interest in medical and social reform. He founded his own weekly journal, *British Annals of Medicine, Pharmacy, Vital Statistics, and General Science*, which lasted only eight months, January to August 1837, but allowed him to write major articles on medical reform and vital statistics. The Births and Deaths Registration Act of 1836 had inaugurated the modern system of civil registration and led to the establishment of the General Register Office in 1837. Farr joined the staff of the General Register Office in 1839, serving forty years, first as compiler of abstracts and then as superintendent of the Statistical Department.

Farr "insisted that the statistician adopt a critical approach, investigating the accuracy of his data, questioning the appropriateness of the units used, and attempting with the help of ratios, logarithms, and the calculus of probabilities to discover relationships and regularity in order to make predictions" [5, p. 29]. Farr's philosophy had an almost immediate impact on improving British statistics. The first four censuses were fraught with many problems. The 1841 census was the first conducted under the supervision of the General Register Office and Farr was one of the key advisors. It was a great improvement over its predecessors and, together with the annual vital statistics data, enabled Farr to put together tables and analyses which placed England at the forefront of this discipline. Between 1836 and the Registration Act of 1874, Farr was largely responsible for establishing the procedures for collecting and analyzing the official mortality statistics. He introduced the standard **death certificate** in 1845 which saw almost no change until 1902. Through Farr's influence the census of 1851 introduced questions on physical disabilities and other medical items which were continued through 1911.

Farr was greatly interested in statistical nosology, introducing his first classification of diseases in 1839. The first International Statistical Congress in 1853 took up the issue, but Farr's nosology did not win the support of other European countries. It was not until 1893 that Jacques Bertillon (1851–1922) proposed a system that became the International List of Causes of Death (*see* **International Classification of Diseases (ICD)**).

Problems noted in the vital registration system in the mid-nineteenth century are still of concern at the end of the twentieth, namely accuracy of diagnoses was not reliable, selection of a single underlying cause of death (*see* **Cause of Death, Underlying and Multiple**) from among several listed

conditions, "the temptation of practitioners to obscure or falsify the cause of death to save respectable families embarrassment in certain sorts of death" [5, p. 62]. Henry Wyldbore Rumsey (1809–1876), one of the chief proponents of sound vital statistics, was vigorous in pointing out statistical fallacies and shortcomings of the existing systems that bear rereading today.

Many of Farr's statistical methods have had a lasting impact: defining mortality rates precisely and basing them on **person-years at risk**, establishing the standard expression of mortality as "deaths per thousand", using the life table and life expectancy as key instruments to assess mortality, using the method of indirect standardization (*see* **Standardization Methods**) to compare mortality rates of localities (although he seems to have made little use of the direct method first demonstrated by F.G.P. Neison in his refutation of the proposal of Edwin Chadwick (1800–1890) to use **average age at death** as a criterion for the health of communities), recommending the establishment of longitudinal **cohort studies** [9], and proposing a paradigm for the estimation of the economic value of human life at each age and social class. Farr's association with **Florence Nightingale** (1820–1910) also resulted in contributions to the use of statistical information for health policy purposes, particularly in respect to the graphic presentation of data (*see* **Graphical Displays**).

Development of Vital Statistics in the United States

As interest in statistical information burgeoned in Europe in the first third of the nineteenth century, a similar phenomenon was occurring on the other side of the Atlantic [4]. Although medicine, statistics, and science generally, in the US lagged behind that in Europe, America had actually preceded other countries in two important respects. Whereas other areas relied on church-maintained records of christenings and burials as the basis for vital statistics, the Massachusetts Bay Colony enacted a law in 1639 requiring the reporting of every birth and death within its jurisdiction, thus establishing the collection of vital statistics as a governmental function covering the entire population. The other colonies gradually adopted similar regulations but for at least the next two hundred years the quality and completeness of the reports were decidedly deficient. The second precedent was when the US became the first nation to establish by constitutional mandate a periodic census requiring complete enumeration of the entire population, conducting its first census in 1790.

At about this time death reports were being used on occasion in port cities to institute quarantine measures in efforts to control epidemics of cholera, yellow fever, and typhus. As the Benthamite social reform interests reached America and evidence for the harmful effects of poverty, industrialization, and unsanitary conditions was sought from vital statistics, the inadequacies of the city and local registration systems became evident. In 1826, Walter Channing (1786–1876) in Boston outlined some of the requirements for valid data on causes of death, including the requisite for medical certification. In 1827 Nathaniel Niles and John D. Russ published the first report on public health statistics in a comparison of mortality data from New York, Philadelphia, Baltimore, and Boston. Other analyses soon followed which became models for the quantitative health reports produced by subsequent generations of health officials which led to increasing pressures for improving the quality of the information. In 1842 Massachusetts again achieved a first by establishing a statewide vital registration system. The effort to establish similar systems in other states marked the beginning of an organized public health movement and contributed to the professionalization of statisticians in this country [2, 3].

Following on the foundation of the Statistical Society of London, statistical societies were started in New York and other American cities. Most did not last very long but the **American Statistical Association**, founded in Boston in 1839, proved to be enduring. It is significant that 14 of the original 54 local members were physicians. But it was a publisher and bookseller, Lemuel Shattuck (1793–1859), who was the Society's key "statist" for health-related issues. He consulted with, among others, Quetelet and was a prime mover for the Massachusetts Registration Act of 1842. He also played a role in the origins of national vital statistics by having mortality queries included in the 1850 census.

In 1846, the first national medical convention (which led to the founding of the American Medical Association) formed two committees relevant to health statistics: (i) a committee on registration

whose report "provided for the convention to formally petition every state government to enact effective registration legislation and to request state and local medical societies to take the lead in lobbying for such laws" [3, p. 201], and (ii) a committee on disease nomenclature which adopted a modification of Farr's classification. Neither of these recommendations was widely adopted for at least 50 years. Although there were many attempts, these efforts were often failures since "the registration movement had moved too far ahead of its base of community support" [3, p. 204]. At the end of the century, no state had a system as good as those in several European countries.

During the last two decades of the nineteenth century, the initiative for improving vital statistics shifted to the Federal government [8]. Under Dr John Shaw Billings (1838–1913), who directed vital statistics in the 1880 and 1890 US censuses, improvements were made in gathering mortality data. The **American Public Health Association** joined with the Census Bureau, which was established in 1902, in drafting a model vital statistics law and standard birth and death certificates that each state could adopt. Because of the early efforts of Cressy L. Wilbur (1865–1928), Chief Statistician for Vital Statistics from 1906 to 1914, the birth- and death-registration areas grew, reaching completeness in 1933, nearly a century after several European countries. The Division of Vital Statistics of the Bureau of the Census was transferred to the Public Health Service in 1946, becoming the National Office of Vital Statistics, with Dr Halbert L. Dunn (1896–1975) as Director. In 1960, NOVS was combined with the National Health Survey to become the **National Center for Health Statistics** with **Forrest E. Linder** (1906–1988) as its first Director.

Development of Health Surveys in the United States

The establishment of the National Health Survey in 1957 marked a milestone in health statistics. With only a few exceptions, previous data relating to health came from vital statistics or from diagnosed diseases seen in hospitals or included in various notifiable **diseases registers**. As public health concerns in the US shifted from the surveillance and control of acute **communicable diseases** to the prevention of chronic diseases, it was necessary to develop data systems that would better describe the current health status of the population (*see* **Quality of Life and Health Status**) and shed some light on health-associated behaviors and use of health care services (*see* **Health Services Organization in the US**). The National Health Survey was the first continuous nationwide survey to gather information from randomly drawn representative samples (*see* **Probability Sampling**) of the noninstitutionalized population of the country to accomplish these aims (*see* **Surveys, Health and Morbidity**). It consists of two distinct surveys: the National Health Interview Survey (NHIS) and the National Health Examination Survey, the latter subsequently expanded to the National Health and Nutrition Examination Survey (NHANES). The NHIS conducts interviews in about 1000 households each week to obtain information on acute illnesses, chronic conditions, health-related knowledge and behaviors, and use of health services. The NHANES involves detailed standardized medical examinations, including laboratory studies and special tests such as ECGs and X-rays, and extensive questionnaires on nutrition and previous health conditions. The NHANES is a periodic survey and NHANES III (actually the sixth cycle of these surveys), being carried out from 1988 to 1994, examined a sample of about 30 000 persons aged 6 months and over. Health interview surveys have now been conducted in many countries and examination surveys have been used effectively in several developing countries to assess the population's health.

These surveys would not have been feasible without the development of survey methodologies which occurred in the twentieth century. Anders N. Kiaer (1838–1919), the first director of the Norwegian Central Bureau of Statistics, reintroduced the idea of a survey sample in what he called the "representative method", in which the sample was to be selected purposively as Laplace had suggested a century earlier, rather than randomly. Arthur Lyon Bowley (1869–1957) is credited with being the first statistician to use random sampling (1906). The seminal breakthrough for sampling methodology came in 1934 when **Jerzy Neyman** (1894–1981) established the theoretical basis for **stratified sampling** with unequal inclusion probabilities. He made another major contribution when he introduced the use of cost functions into survey sampling theory (1938). In the early 1940s, Morris Hansen (1910–1990) and William Hurwitz (1908–1969) at the Bureau of

the Census perfected the methodologies for complex **multistage sampling** designs that are the basis for most modern large-scale surveys.

Conclusion

At the end of the twentieth century, most industrialized countries have effective vital statistics systems in place and many have established periodic interview surveys to assess the health status and needs of their citizens. Much remains to be done in developing countries to institute health services information systems (*see* **Administrative Databases**) that can guide public policies and programs. As the public health burden continues to shift from infectious diseases to problems of an aging population, to concerns about health promotion and disease prevention, and to assuring adequate health care for all citizens, the needs for reliable, relevant, and timely health statistics become ever greater. Fortunately, the methodologies developed over several centuries and the data systems that have been established can, if appropriate resources are provided, meet these needs.

References

[1] Bollet, A.J. (1973). Pierre Louis: the numerical method and the foundation of quantitative medicine, *American Journal of Medical Science* **266**, 92–101.

[2] Cassedy, J.H. (1969). *Demography in Early America. Beginning of the Statistical Mind. 1600–1800*. Harvard University Press, Cambridge, Mass.

[3] Cassedy, J.H. (1984). *American Medicine and Statistical Thinking 1800–1860*. Harvard University Press, Cambridge, Mass.

[4] Cohen, P.C. (1982). *A Calculating People. The Spread of Numeracy in Early America*. University of Chicago Press, Chicago.

[5] Eyler, J.M. (1979). *Victorian Social Medicine: The Ideas and Methods of William Farr*. Johns Hopkins University Press, Baltimore.

[6] Greenwood, M. (1941–1943). Medical statistics from Graunt to Farr. *Biometrika* **32**, (1941), 101–127; **32** (1942), 203–225; **33** (1943), 1–24. Published by Cambridge University Press, Cambridge, 1948, as the Fitzpatrick Lectures for 1941 and 1943.

[7] Hald, A. (1990). *A History of Probability and Statistics and Their Applications Before 1750*. Wiley, New York.

[8] Lawrence, P.S. (1976). The health record of the American People, in *Health in America: 1776–1976*. US Department of Health, Education, and Welfare, DHEW Pub. No. (HRA)76–616.

[9] Nissel, M. (1987). *People Count. A History of the General Register Office*. Office of Population Censuses and Surveys, HMSO, London.

[10] Rabinovitch, N.L. (1973). *Probability and Statistical Inference in Ancient and Medieval Jewish Literature*. University of Toronto Press, Toronto.

[11] Stigler, S.M. (1968). *The History of Statistics. The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, Mass.

[12] Westergaard, H. (1932). *Contributions to the History of Statistics*. King, London.

MANNING FEINLEIB

Health Status Instruments, Measurement Properties of

Introduction

In this article, we present the key measurement properties necessary for a useful health status instrument. This article also includes a comment on the issue of respondent and administrative burden. The discussion is drawn largely from a previous publication [5]. The concepts are most relevant for measurement of health status, but apply to measurements of any human attribute or characteristic.

The Structure of Health Status Measures

Since semantic issues in health status measurement are both controversial and important, we will clarify how we shall use words in our discussion. Some measures consist of a single question which essentially asks "How would you rate the quality of your life?" [25].

This question may be asked in a simple or a very sophisticated fashion, but either way yields limited information. More commonly, health status instruments are questionnaires made up of a number of *items*, or questions. These items are added up in a number of *domains* (also sometimes called *dimensions*). A domain or dimension refers to the area of behavior or experience that we are trying to measure. Domains might include mobility and self-care, which could further be aggregated into physical function, or depression, anxiety, and well-being, which could be aggregated to form an emotional function domain. For some instruments, investigators have undertaken rigorous valuation exercises in which the importance of each item is rated in relation to the others. More often, items are equally weighted, implying an assumption that their value is equal.

What Makes a Good Health Status Instrument?

Current strategies for evaluating health status measures build on close to 100 years' work in the measurement of attributes such as intelligence and attitudes [1]. These strategies have evolved, incorporating insights from studies directly relating to health status and **quality of life** [23].

Measuring at a Moment in Time versus Measuring Change

The goals of health status measures include differentiating between people who have a better health status and those who have a worse health status (a *discriminative instrument*), and measuring how much health status has changed (an *evaluative instrument*) [16]. The construction of instruments for these two purposes can be quite different. For instance, let us take the example of thyroid disease. If we are trying to discriminate between those with and without thyroid disease, we would be unlikely to include fatigue as an item, because fatigue is too common among people who do not have thyroid disease. On the other hand, in measuring improvement in health status with treatment, fatigue, because of its importance in the day-to-day lives of people with thyroid disease, would be a key item. In the next sections, concerned with what makes a good health status instrument, we list key measurement properties separately for discriminative and evaluative instruments. The properties that make useful discriminative and evaluative instruments are presented in Table 1.

Signal and Noise

Investigators examining physiologic endpoints have long been aware that reproducibility and validity are the necessary attributes of a good test. For health status instruments, reproducibility translates into having a high ratio of signal to noise, and validity translates into whether they are really measuring what they are intended to measure. For discriminative instruments, the way of quantifying the signal-to-noise ratio is called *reliability*. If the variability in scores between subjects (the signal) is much greater than the variability within subjects (the noise), an instrument will be deemed reliable. Reliable instruments will generally demonstrate that stable subjects show more or less the same results on repeated administration. For evaluative instruments, those designed to measure changes within individuals over time, the way of determining the signal-to-noise ratio is called *responsiveness*. Responsiveness refers to an instrument's ability to

Table 1 What makes a good health status measure

Instrument property	Evaluative instruments (measuring differences within subjects over time)	Discriminative instruments (measuring differences between subjects at a moment in time)
High signal-to-noise ratio	Responsiveness	Reliability
Validity	Correlations of changes in measures over time consistent with theoretically derived predictions	Correlations between measures at a moment in time consistent with theoretically derived predictions
Interpretability	Differences within subjects over time can be interpreted as trivial, small, moderate, or large	Differences between subjects at a moment in time can be interpreted as trivial, small, moderate, or large

detect change. If a treatment results in an important difference in health status, investigators wish to be confident that they will detect that difference, even if it is small. Responsiveness will be directly related to the magnitude of the difference in score in patients who have improved or deteriorated (the signal) and the extent to which patients who have not changed obtain more or less the same scores (the noise).

Validity When There is a Gold Standard

If we have a **gold standard** or criterion standard for some aspect of health, it implies that we have endorsed a particular measurement tool as providing the underlying truth about that aspect. The concept of a reference, gold, or criterion standard is most easily applied for physiologic measures. For instance, experts may agree that the cardiac angiogram is a gold standard for measurement of various aspects of cardiac anatomy and function, and noninvasive tests should be judged in relation to this criterion.

Although there is no gold standard for health status, there are instances in which there is a specific target for a health status measure that can be treated as a criterion or gold standard. Under these circumstances, one determines whether an instrument is measuring what is intended using *criterion validity*, according to which an instrument is valid insofar as its results correspond to those of the criterion standard. For instance, criterion validity is applicable when a shorter version of an instrument (the test) is used to predict the results of the full-length index (the gold standard). Another example is using a health status instrument to predict mortality. In this instance, to the extent that variability in survival between patients (the gold standard) is explained by the questionnaire results (the test), the instrument will be valid. Self-ratings of health such as more comprehensive and lengthy measures of general health perceptions include an individual's evaluation of her or his physiologic, physical, psychologic, and social well-being. Perceived health, measured through self-ratings, is an important predictor of mortality [17].

Validity When There is No Gold Standard

Validity has to do with whether the instrument is measuring what it is intended to measure. When there is no gold or criterion standard, health status investigators have borrowed validation strategies from clinical and experimental psychologists, who have for many years been dealing with the problem of deciding whether questionnaires examining intelligence, attitudes, and emotional function are really measuring what they are supposed to measure.

The types of validity that psychologists have introduced include face, content, and construct validity.

Face validity refers to whether an instrument appears to be measuring what it is intended to measure, while *content validity* refers to the extent to which the domain of interest is comprehensively sampled by the items, or questions, in the instrument. Quantitative testing of face and content validity are rarely attempted. Feinstein [4] has reformulated these aspects of validity by suggesting criteria for what he calls the *sensibility*, including the applicability of the questionnaire, its clarity and simplicity, likelihood of **bias**, comprehensiveness, and whether redundant items have been included. Some of these criteria compete with one another: redundant items may help to ensure comprehensiveness, and reduce the likelihood of bias, while increasing the burden on respondents. Because of their specificity, Feinstein's criteria facilitate quantitative rating of an instrument's face and content validity [18].

The most rigorous approach to establishing validity is called *construct validity*. A construct is a theoretically derived notion of the domain(s) that we wish to measure. An understanding of the construct will lead to expectations about how an instrument should behave if it is valid. Construct validity therefore involves comparisons between measures, and examination of the logical relationships that should exist between a measure and characteristics of patients and patient groups. The first step in construct validation is to establish a "model" or theoretic framework that represents an understanding of what investigators are trying to measure. That theoretic framework provides a basis for understanding how the system being studied behaves, and allows hypotheses or predictions about how the instrument being tested should relate to other measures. Investigators then administer a number of instruments to a population of interest, and examine the data. Validity is strengthened or weakened according to the extent to which the results confirm or refute the hypotheses. For example, a discriminative health status instrument may be validated by comparing two groups of patients; those who have undergone a very toxic chemotherapeutic regimen and those who have undergone a much less toxic chemotherapeutic regimen. A health status instrument should distinguish between these two groups, and, if it does not, it is very likely that something has gone wrong. Alternately, **correlations** between symptoms and functional status can be examined, the expectation being that those with a greater number and severity of symptoms will have lower functional status scores on a health status instrument.

Another example is the validation of an instrument discriminating between people according to some aspect of emotional function. Results from such an instrument should show substantial correlations with existing measures of emotional function. We call construct validity that deals with measurements at one point in time *cross-sectional construct validity*. The principles of validation are identical for evaluative instruments, but their validity is demonstrated by showing that *changes* in the instrument being investigated correlate with *changes* in other related measures in the theoretically derived predicted direction and magnitude (*longitudinal construct validity*). For instance, the validity of an evaluative measure of health status for patients with chronic lung disease was supported by the finding of moderate correlations with changes in walk test scores [7].

Validation is not an all-or-nothing process. We may have varying degrees of confidence that an instrument is really measuring what it is supposed to measure. *A priori* predictions of the strength of relationship with other measures that one would expect if a new instrument is really measuring what is intended strengthen the validation process. Without such predictions, it is generally easy to rationalize whatever correlations between measures are observed.

Validation does not end when the first study with data concerning validity is published, but continues with repeated use of an instrument. The more frequently an instrument is used, and the wider the situations in which it performs as we would expect if it were really doing its job, the greater is our confidence in its validity. Perhaps, we should never conclude that a questionnaire has "been validated"; the best we can do is to suggest that strong evidence for validity has been obtained in a number of different settings and studies.

Interpretability

A final key property of a health status measure is *interpretability*. For a discriminative instrument, we could ask whether a particular score signifies that a patient is functioning normally, or has mild, moderate, or severe impairment of health status. For an evaluative instrument, we might ask whether a particular change in score represents a trivial, small but important, moderate, or large improvement or deterioration. Considerable research has focused on establishing what constitutes the minimal important difference (MID) in health status. One can define the MID as "The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and which would lead the patient or clinician to consider a change in the management" [22]. However, any change in management will depend on the downsides, including cost, associated with that outcome and the values and preferences patients place on these outcomes.

A number of strategies are available for trying to make health status scores interpretable and describe the MID [6, 9, 11]. The first is called an anchor-based approach. Investigators have often used global ratings of change (patients classifying themselves as unchanged, or experiencing small, medium, and large improvements or deteriorations) as the independent

standard when correlations between the ratings and the instrument for which information on interpretability is sought are strong, generally greater than 0.5. Several disease-specific instruments use seven-point scales with an associated verbal descriptor for each level on the scale. For each questionnaire domain, one could divide the total score by the number of items so that domain scores can range from 1 to 7. Using this approach to framing response options, the smallest difference that patients consider important is often approximately 0.5 per question [10, 14]. A moderate difference corresponds to a change of approximately 1.0 per question, and changes of greater than 1.5 can be considered large. So, for example, in a domain with four items, patients will consider a one point change in two or more items as important. This finding seems to apply across different areas of disease, including patients with chronic airflow limitation [10]; patients with both adult [14] and child [13] asthma patients, and parents of child asthma patients [12]; and adults with rhinoconjunctivitis [15].

The approach described above relies on within-patient comparisons as the independent standard. One alternative is between-patient comparisons. In one example of this approach, groups of patients with chronic airflow limitation participating in a respiratory rehabilitation program completed the Chronic Respiratory Questionnaire [20]. The patients conversed with one another long enough to make judgments about their relative experience of fatigue in daily life. While there was a bias in their assessment (patients generally considered themselves better off than one another), their relative ratings allows estimates of what differences in the Chronic Respiratory Questionnaire score constitute small, medium, and large differences [20].

Another anchor-based approach uses HRQL instruments for which investigators have established the MID. Investigators can apply **regression** or other statistical methods to compute the changes on a new instrument that correspond to those of the instrument with the established MID [21]. Similar to the anchor-based approach using transition ratings, investigators should ensure that the strength of the correlation between the change scores of these instruments exceeds a minimum (e.g. a correlation coefficient of 0.5). Yet another approach to estimate the MID involves enrolling panels of experts or patients and qualitative research methods, such as Delphi techniques [26]. The experts establish a consensus what constitutes the MID of the CRQ. Investigators have also proposed distribution-based methods to determine interpretability of HRQL instruments. Distribution-based methods differ from anchor-based methods in that they interpret results in terms of the relation between the magnitude of effect and some measure or measures of variability in results [9]. The magnitude of effect can be the difference in an individual patient's score before and after treatment, a single group's score before or after treatment, or the difference in score between treatment and control groups. As a measure of variability, investigators may choose between-patient variability (the **standard deviation** of patients at baseline, for instance) or within-patient variability (the standard deviation of change that patients experienced during a study).

If an investigator used the distribution-based approach, the clinician would see a treatment effect reported as, for instance, 0.3 standard deviation unit. The great advantage of distribution-based methods is that the values are easy to generate for almost any HRQL instrument because there will always be one or more measures of variability available. This contrasts with the work needed to generate an anchor-based interpretation. The problem related to this methodology is that the units do not have intuitive meaning to clinicians. It is possible, however, that clinicians could gain experience with standard deviation units in the same way they learn to understand other HRQL scores. Cohen addressed this problem in a seminal work by suggesting that changes in the range of 0.2 standard deviation unit represent small changes, those in the range of 0.5 standard deviation unit represent moderate changes, and those in the range of 0.8 standard deviation unit represent large changes [3]. To further respond to this problem, investigators have attempted to provide empirical evidence about the relationship between distribution-based and anchor-based results. These studies address the question, "What is the appropriate interpretation of a particular magnitude of effect, in distribution-based units, as judged by the results of anchor-based studies?".

The **standard error** of measurement (SEM) presents another distribution based method and is defined as the variability between an individual's observed score and the true score and is computed as the baseline standard deviation multiplied by the square root of 1 minus the reliability of the

Table 2 Modes of administration of health status measures

Mode of administration	Strengths	Weaknesses
Self-administered	Minimal resources required	Greater likelihood of low response rate, missing items, or misunderstandings
	Willingness to respond to personal questions	Need modest literacy and numeracy skills
	Reduced risk of erroneous interpretation by interviewer	
	Can be computer- administered	
Interviewer-administered	Maximizes response rate	Requires considerable resources,
	Few, if any, missing items	Training of interviewers
	Minimizes errors of misunderstanding	May reduce willingness to acknowledge problems
Telephone-administered	Few, if any, missing items	Limits format of instrument
	Minimizes errors of misunderstanding	Access to phone necessary
	Less resource-intensive than interviewer-administered	
Postal-administered	Modest resources required	High likelihood of low response rate, missing items, or misunderstandings
		Interference by family members
	Willingness to respond to personal questions	Need modest literacy and numeracy skills
Surrogate responders	Reduces stress for target group (very elderly or sick)	Perceptions of surrogate may differ from those of target group
	Improves response rate (young children and those incapable of responding)	

QOL measure. In theory, a QOL measure's standard error of measurement is sample independent, whereas its component statistics, the standard deviation and the reliability estimate, are sample dependent and vary around the standard error of measurement [27]. When the between-person variability in the population increases, the standard deviation will increase (tending to raise the standard error of measurement), but the reliability will also increase (tending to lower the standard error of measurement). Thus, the standard error of measurement largely reflects within-person variability over time.

Knowing the change or difference in score that is meaningful enables the clinically useful estimation of the number of patients who need to be treated for one individual to have an additional clinically meaningful improvement [8]. (*see* **Number Needed to Treat (NNT)**)

Respondent and Administrative Burden

Alternate approaches to obtaining information from patients have different resource implications. The strengths and weaknesses of the different modes of administration are summarized in Table 2. Health status questionnaires are either administered by trained interviewers or self-administered. The former method is resource intensive, but ensures compliance and minimizes errors and missing items. The latter approach is much less expensive, but increases the number of missing patients and missing responses. A compromise between the two approaches is to have the instrument completed under supervision.

Another compromise is the telephone interview, which minimizes errors and missing data but may necessitate a relatively simple questionnaire structure unless the response options are provided to the respondent in advance. With clear instructions postal completion can provide valid data but may have a low response rate [19]. Computer administration has become a common method of questionnaire administration and yields similar responses to paper versions [2] (*see* **Computer-assisted Interviewing**).

Another issue in administrative and respondent burden is the length of the questionnaires. This may be less of an issue in research settings in which,

once one has invested the resources in setting up the interview, the incremental resource expenditure of a longer interview is relatively minor. On the other hand, it may be necessary to find a short questionnaire for clinical settings in which one needs to obtain information at regular intervals (*see* **Questionnaire Design**).

Under these circumstances, distilling the measurement of health status into a few key questions would be a dream come true. One approach to achieving this goal is to develop a long instrument, test it, and use its performance to choose key questions to include in a shorter index. This approach has been used, for example, to create shorter questionnaires based on the lengthy instruments from the Medical Outcomes Studies [24].

How would one determine if the shortened questionnaire is an adequate substitute for the full version? The issue for discriminative purposes is the extent to which people are classified similarly by the short and long forms of the questionnaire. Statistically, one would examine the extent to which **variance** or variability in scores, in the full instrument is predicted or explained by scores of the abbreviated version: the greater the extent to which the rating of people's quality of life by the shorter instrument corresponds to ratings by the longer version, the more comfortable we should be with the substitution.

For evaluative purposes, the responsiveness and validity of the shorter version should be tested against the full instrument. If both correlations of change with independent measures and instrument responsiveness were comparable, substitution of the shorter instrument would be desirable. If measurement properties deteriorated, the investigator would face a decision about trading off respondent burden with increases in sample size necessitated by a less responsive instrument. Before comparing results generated by original and shortened versions of a questionnaire, one should check that a single patient sample yields the same between-subject differences (discriminative) and within-subject changes (evaluative) with both versions.

References

[1] American Psychological Association, Washington, D.C. (1985). *Standards for Educational and Psychological Testing*.

[2] Caro, J.J., Sr., Caro, I., Caro, J., Wouters, F. & Juniper, E.F. (2001). Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Quality of Life Research* **10**, 683–691.

[3] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed., Lawrence Erlbaum Associates, Hillsdale.

[4] Feinstein, A. (1987). *Clinimetrics*. Yale University Press, New Haven, 141–166.

[5] Guyatt, G.H., Feeny, D. & Patrick, D. (1993). Measuring health-related quality of life: basic sciences review, *Annals of Internal Medicine* **70**, 225–230.

[6] Guyatt, G.H. (2000). Making sense of quality-of-life data, *Medical Care* **38**, II175–II179.

[7] Guyatt, G.H., Berman, L.B., Townsend, M., Pugsley, S.O. & Chambers, L.W. (1987). A measure of quality of life for clinical trials in chronic lung disease, *Thorax* **42**, 773–778.

[8] Guyatt, G.H., Juniper, E.F., Walter, S.D., Griffith, L.E. & Goldstein, R.S. (1998). Interpreting treatment effects in randomised trials, *BMJ* **316**, 690–693.

[9] Guyatt, G.H., Osoba, D., Wu, A.W., Wyrwich, K.W., Norman, G.R. & Clinical Significance Consensus Meeting, Group. (2002). Methods to explain the clinical significance of health status measures, *Mayo Clinic Proceedings* **77**, 371–383.

[10] Jaeschke, R., Guyatt, G., Keller, J. & Singer, J. (1989). Measurement of Health Status: Ascertaining the meaning of a change in quality-of-life questionnaire score, *Control Clinical Trials* **10**, 407–415.

[11] Jaeschke, R., Singer, J. & Guyatt, G.H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference, *Control Clinical Trials* **10**, 407–415.

[12] Juniper, E., Guyatt, G., Feeny, D., Ferrie, P., Griffith, L. & Townsend, M. (1996). Measuring quality of life in the parents of children with asthma, *Quality Life Research* **5**, 27–34.

[13] Juniper, E., Guyatt, G., Feeny, D., Ferrie, P., Griffth, L.E. & Townsend, M. (1996). Measuring quality of life in children with asthma, *Quality Life Research* **5**, 35–46.

[14] Juniper, E., Guyatt, G., Willan, A. & Griffth, L. (1994). Determining a minimal important change in a disease-specific quality of life questionnaire, *Journal of Clinical Epidemiology* **47**, 81–87.

[15] Juniper, E.F., Guyatt, G.H., Griffith, L.E. & Ferrie, P.J. (1996). Interpretation of rhinoconjunctivitis quality of life questionnaire data, *Journal of Allergy & Clinical Immunology* **98**, 843–845.

[16] Kirshner, B. & Guyatt, G. (1985). A methodological framework for assessing health indices, *Journal of Chronic Diseases* **38**, 27–36.

[17] Mossey, J.M. & Shapiro, E. (1982). Self-rated health: a predictor of mortality among the elderly, *American Journal of Public Health* **72**, 800–808.

[18] Oxman, A.D. & Guyatt, G.H. (1991). Validation of an index of the quality of review articles, *Journal of Clinical Epidemiology* **44**, 1271–1278.

[19] Pinnock, H., Sheikh, A. & Juniper, E. (2004). Evaluation of an intervention to improve successful completion of the Mini-AQLQ: comparison of postal and supervised completion, *Primary Care Respiration*; In press.

[20] Redelmeier, D., Guyatt, G. & Goldstein, R. (1996). Assessing the minimal important difference in symptoms: a comparison of two techniques, *Journal of Clinical Epidemiology* **49**, 1215–1219.

[21] Schünemann, H., Griffith, L., Jaeschke, R., Stubbing, D., Goldstein, R. & Guyatt, G.H. (2003). Evaluation of the minimal important difference for the feeling thermometer and St. Georges Respiratory questionnaire in patients with chronic airflow limitation, *Journal of Clinical Epidemiology* **56**, 1170–1176.

[22] Schünemann, H., Puhan, M., Goldstein, R., Jaeschke, R. & Guyatt, G. (2004). Measurement properties and interpretability of the Chronic Respiratory Disease Questionnaire (CRQ), *Journal of COPD*; In Press.

[23] Scientific-Advisory-Committee (1995). Instrument review criteria, *Medical Outcomes Trust Bulletin* **3**.

[24] Stewart, A.L., Hays, R.D. & Ware, J.E., Jr. (1988). The MOS short-form general health survey. Reliability and validity in a patient population. *Medical Care*, **26**, 724–735.

[25] Torrance, G. (1986). Measurement of health state utilities for economic appraisal, *Journal of Health Economics* **5**, 1–30.

[26] Wyrwich, K.W., Fihn, S.D., Tierney, W.M., Kroenke, K., Babu, A.N. & Wolinsky, F.D. (2003). Clinically important changes in health-related quality of life for patients with chronic obstructive pulmonary disease. An Expert Consensus Panel Report, *Journal of General Internal Medicine* **18**, 196–202.

[27] Wyrwich, K.W., Nienaber, N.A., Tierney, W.M. & Wolinsky, F.D. (1999). Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life, *Medical Care* **37**, 469–478.

HOLGER SCHÜNEMANN, ELIZABETH JUNIPER & GORDON GUYATT

Health Workforce Modeling

Health workforce modeling is generally concerned with projecting the future supply of and requirements for a particular type of health professional. The objective of such an effort is to assess the relative balance between supply and requirements under various assumptions and alternative future workforce policies. *Health workforce modeling* is a term that has come into usage over the last two decades as a more gender-neutral formulation of what had traditionally been called *health manpower* planning [6]. In addition, modelers have in more recent years used the more neutral term *requirements* as a generic term which may reflect, depending on the disciplinary background and/or political orientation of the modeler, the "needs", "wants", "demand", or "expected utilization" for health services of a relevant population (*see* **Health Care Utilization Data**). Health workforce modeling, when employed at a regional or national level, is directed toward alerting policy makers to current or potential future imbalances between supply and requirements or to identifying maldistributions of professionals by geographic region, specialty, or practice setting which may adversely affect access to care, **quality of care**, or health care costs. The deceptively simple goal of these analyses is to develop policies, typically affecting the supply side, to ensure that the proper number and type of health professionals will be available to deliver required services to a specified future population. In practice, the achievement of this goal is complicated by the incompleteness of data necessary to implement the models, lack of agreement on essential definitions, competing perspectives of diverse stakeholders, and lack of agreement on what constitutes the "correct" balance between supply and requirements.

Workforce modeling has become of greater interest as governments wrestle with fundamental reforms of the structure and financing of their health care systems (*see* **Health Care Financing**). The undersupply of health professionals can adversely affect the health status and economic viability of populations. Alternatively, because the education of health professionals is supported in large part by public funds, the oversupply of highly trained professionals wastes scarce societal resources that could be better employed elsewhere. The unemployment or underutilization of health professionals carries enormous personal costs as well. However, it has been persuasively argued by Reinhardt [11] that the role of governments in attempting to bring supply in line with requirements ought to be limited to making information on health professions markets freely available to all affected parties so that the market can adjust supply and demand as it does in most other professions.

Invariably, a health professions model will develop a forecast of the future supply of one or more types of personnel and a forecast of the requirements for the personnel in a future time period. Occasionally, modelers will verify their models by "backcasting" to determine if the model, under known conditions and parameter settings, would have predicted correctly a previously recorded level of supply. Some intrepid researchers have assessed the historical performance of alternate forecasts made in prior periods to actual data after they became available [1].

At the national or regional level, three categories of models have been employed:

1. Supply models which forecast the number of a particular kind of health professional expected to be practicing at some future time period (usually expressed either in full-time equivalent persons or in head count).
2. Requirements models which translate the expected utilization or need for specified health services into requirements for a particular kind of professional.
3. Integrated models which explicitly represent the interaction of supply and requirements and other exogenous factors such as disposable personal income, health insurance coverage, and managed care penetration simultaneously to develop estimates of supply and requirements.

Health Workforce Supply Models

Health professions supply models have taken several forms. Conceptually, the most simple is a model that forecasts the future stock of particular kinds of health professionals by obtaining from professional associations or licensure data a count of those practicing in one year, adding to it the expected entrants and subtracting those who leave the profession owing to retirement or death, to produce an estimate of the

active workforce in a future period. Rates of addition, separation, and labor force participation in a cohort will depend on age, gender, and geographical location, among other factors. These labor stock models are appealing for policy analysis because training program enrollments, graduation rates, class composition, and licensure or certification rates are at least partially controllable through policy interventions. As described below, the US Bureau of Health Professions has developed a set of labor stock models to estimate supply [13].

The Bureau of Labor Statistics (BLS) of the US government uses a complex econometric model to estimate the occupational employment of 507 occupations in 258 industrial groupings. The model depends in part on projections of the gross domestic product (GDP) contributions of various industry sectors, the interrelationships between sectors, demand for goods and services, personal income, and other factors. The GDP, demand, and income projections alone require the solution of 400 equations with 213 exogenous variables. Despite the complexity of the BLS models at the macro level, these techniques cannot capture the micro-level details of training program structure and career choice that are found in workforce stock models which frequently drive the production of health professionals.

Bureau of Health Professions Physician Supply Models

Because the investment by society into the training of physicians is greater than for any other health profession and because the length of the supply "pipeline" is the longest of any health profession, considerable effort has been devoted to modeling the physician supply process. At the US federal level, the Bureau of Health Professions [3] utilizes a physician supply forecasting model that consists of five submodels: three at the national level and two at the level of states and census regions. An aggregate supply model forecasts the total national supply of physicians by age, gender, and country of medical education. A specialty model allocates the total supply among 36 specialties in eight practice settings (inpatient, outpatient, long-term care, etc.) and to nonpatient care activities (administration, teaching, and research). A model of the graduate medical education process projects the distribution of residents by specialty and by year of training for future years. The results of the graduate medical education model may be influenced by changing the size, fill rates, and proportions of US and international medical graduates in residency programs. These, and the dynamics of specialty choice and specialty switching, are policy variables that can be influenced, in part, by government initiatives.

Health Workforce Requirements Models

Unlike supply models, which are relatively transparent in their assumptions, requirements modeling is influenced at least as much by the philosophical perspective taken by the model as by the analytic approach. Supply models are largely descriptive. Requirements models are either explicitly or implicitly normative in that they describe what the number and type of health care professionals should be to provide health care to a given population. The simplest (and least useful) kind of requirements model is to form a ratio of providers-to-population, e.g. dentists per 100 000 persons. These provider-to-population ratios give a gross measure of supply which can be used to compare one nation with another or one region with another but tell us nothing about what care is delivered, how it is delivered, to whom it is delivered, and in what facilities it is delivered.

A utilization-based model will forecast health services utilization for a particular population, usually in the form of office visits, inpatient episodes of care, nursing home days, etc. Each encounter type can then be described in terms of who is involved and how long a particular person or team is typically involved. Person-hours are then aggregated over all delivery venues applicable to a given health profession to determine the number of full-time equivalent providers required to deliver an assumed volume of services. In their pure form, utilization-based models forecast only what *will* be rather than what *should* be the number and types of providers required under certain utilization and task allocation assumptions. Utilization models will account for variation in utilization of services by age, gender, and geographic region and they may account for differential access owing to insurance status, provider availability, travel distance, and social and economic factors. Utilization models do not attempt to provide estimates of the number and kind of services needed by a population to maintain that population in optimum health.

Need-based models, on the other hand, start from the perspective of a population's need for a certain mix of health services as recommended by knowledgeable health professionals usually convened in consensus panels. Gaining consensus on what should be done, to whom it should be done, and by whom it should be done is not an easy process, especially when competing specialties and professions are involved. The term *adjusted need-based model* refers to an approach that tempers the requirements estimate with information about how populations actually use services based on assumptions relating price and accessibility.

In the US in the early 1980s, the Graduate Medical Education National Advisory Committee (GMENAC) [14] developed an adjusted need-based model from the work of a number of disease area expert panels representing medical or surgical specialists, primary care providers, and nonphysician providers (such as physician assistants and nurse practitioners). A modeling panel integrated the findings of the disease area panels to resolve problems with overlap and variations in assumptions.

The opportunity for enormous variation in requirements estimates exists at two points in the process. First, professionals (and clinical evidence) may not agree on what services should be provided. Secondly, the rate of service provision and the mix of professionals providing the service can be affected greatly by health system organization (*see* **Health Services Organization in the US**) and financing structures. Analysts have, for example, found variations of 25% or more in the number of physicians required, depending on whether traditional fee-for-service or aggressive managed care utilization rates and staffing ratios are assumed [12, 16]. Unless one is willing to accept wildly unrealistic estimates of the number of health professionals required, estimates must be based on supportable assumptions regarding the utilization and delivery patterns that will actually occur at the specified future time [10]. Presentation of supply and requirements estimates under alternative future scenarios is one way to illustrate the sensitivity of estimates to changes in the settings of model parameters or policy options.

Another recent development is the Bureau of Health Professions' Integrated Requirements Model for Primary Care for Physicians' Assistants (PAs), Nurse Practitioners (NPs), and Certified Nurse-Midwives (CNMs) [9]. Known as the IRM, this system has been used to forecast US requirements for physicians and other nonphysician primary care providers for the delivery of primary health care services, using a variety of assumptions (or scenarios). These assumptions can be adjusted by the users and are designed so that users can also forecast requirements under an unlimited number of scenarios by varying model inputs and parameters. At the heart of the model is the assumption that requirements will differ depending on how certain primary care tasks are allocated to the various health professions and where the boundaries of "primary care" tasks lie within the health services domain.

A Recent Application of Requirements Forecasting in an Integrated Delivery System

To determine the number and kinds of physicians required to staff the US Department of Veterans' Affairs (VA) health care delivery system, the Institute of Medicine [7] utilized three distinct but complementary approaches: (i) empirical models based on current practice in the VA; (ii) expert judgment models; and (iii) comparisons to other large integrated systems operating in the US. In practice, these three approaches interacted to a great extent, and the final recommendation was for an informed blending of alternative requirements forecasts.

The empirical models developed were of two forms: (i) production functions (PF), in which physicians were one of several factors leading to the production of patient care workload, and (ii) inverse production functions (IPF), in which the required number of physicians in a given specialty was estimated directly from workload and other staffing inputs. In the PF variant, the patient workload (measured, for example, in weighted workload units) for one of 14 patient care areas (e.g. inpatient psychiatric service) was hypothesized to be related to the number of physician FTEs by specialty allocated to the area, the number of residents by year of training, nurse FTE per physician, other support FTE per physician, and other institutional factors possibly affecting productivity.

In the IPF variant the required number of physicians in 11 specialties for a given facility was assumed to be a function of the estimated required workload in all settings (inpatient, outpatient, and long-term care), the number of residents assigned in

the specialty by year of training, support staff allocated to the specialty, and other productivity-related factors such as hospital type.

The second major approach to forecasting physician requirements was to use 11 expert panels organized around specialty (e.g. neurology, rehabilitation medicine, and radiation oncology) or multidisciplinary care area (e.g. long-term care). Rather than simply critiquing the empirically derived estimates, the panels developed independent quantitative estimates of physician requirements under a variety of alternative scenarios of care provision. The work of the panels was informed by the results of the empirical models and external norm data from other organizations.

Ultimately, estimates of requirements from the empirical models were formally reconciled with estimates from the expert judgment methods through a weighting and smoothing process. In the end, no "cook book" approach was developed. Rather, it was recommended that the empirical, expert judgment, and external norm approaches be continually enhanced and coordinated to produce demand-driven staffing requirement estimates that would guide management decisions and resource allocation. On the basis of the Institute of Medicine (IOM) experience, Lipscomb et al. [8] have developed a **Bayesian** statistical approach to combine expert panel judgments through **hierarchical models**.

Integrated Supply and Requirements Models

One currently used example of an integrative model is the Econometric Model of the Dental Sector (EMODS) developed at the US Bureau of Health Professions [1]. EMODS employs an interactive system of equations explicitly to represent the impact of population changes, disease etiology, dental insurance, cost of services, and personal income on demand for dental care. Also included on the production side is the technology of care delivery, use of auxiliaries, hours of work, and the labor content of procedures. Supply equations (exogenous to the model) include not only the stock of dentists, but also the stock of various auxiliaries as well. Prices for care affect consumption of services, which in turn affects employment of dentists. The full model contains a set of 195 equations that represent the interactions in the dental care sector. The model has been tested and **calibrated** by comparing model estimates to actual data over a period of several years (*see* **Model Checking**). Among the complex econometric models that have been formulated for various health professions, the most highly developed is the model of the dental sector, which is self-contained and relatively easy to describe. Even in dentistry, however, it is difficult to obtain adequate data to permit full utilization of econometric models. In fact, to reduce the data burdens, researchers have found that a single-equation **regression** model, while not providing the richness of insight available in the full model, does permit adequate forecasts of dental prices and expenditures [1].

Data Requirements for Workforce Modeling

Figure 1, adapted from *Data Systems to Support State Health Personnel Planning and Policy making: A Resource Guide for State Agencies* [15], outlines three different levels of sophistication in both the supply and requirements domains and identifies the kinds of data required at each level. An important feature of this approach is that it allows one to move from the simplest approaches to the more complex approaches. This is essential because one can get lost quite easily in sophisticated details of modeling and equations before one has had a chance to answer the more basic questions about supply and requirements. From a practical perspective, modeling efforts should proceed sequentially, collecting data specified in the innermost portions of Figure 1 and then proceeding outward in both the supply and requirements directions consistent with not only the required precision and planning time horizon, but also the resources available to the task.

On the supply side, these basic analyses and data include the counts of licensees in state and employment counts. Such basic data can be augmented by national and state trends in the number of practitioners, as indicated by the growth in newly licensed persons in the state or nation and trends in employment. Professional association data can be useful in understanding and projecting supply, but, because of duplication in licenses, national estimates of supply remain problematic. Such an approach, however, can be used at the state or regional level, where universal unduplicated licensure ensures that supply can be

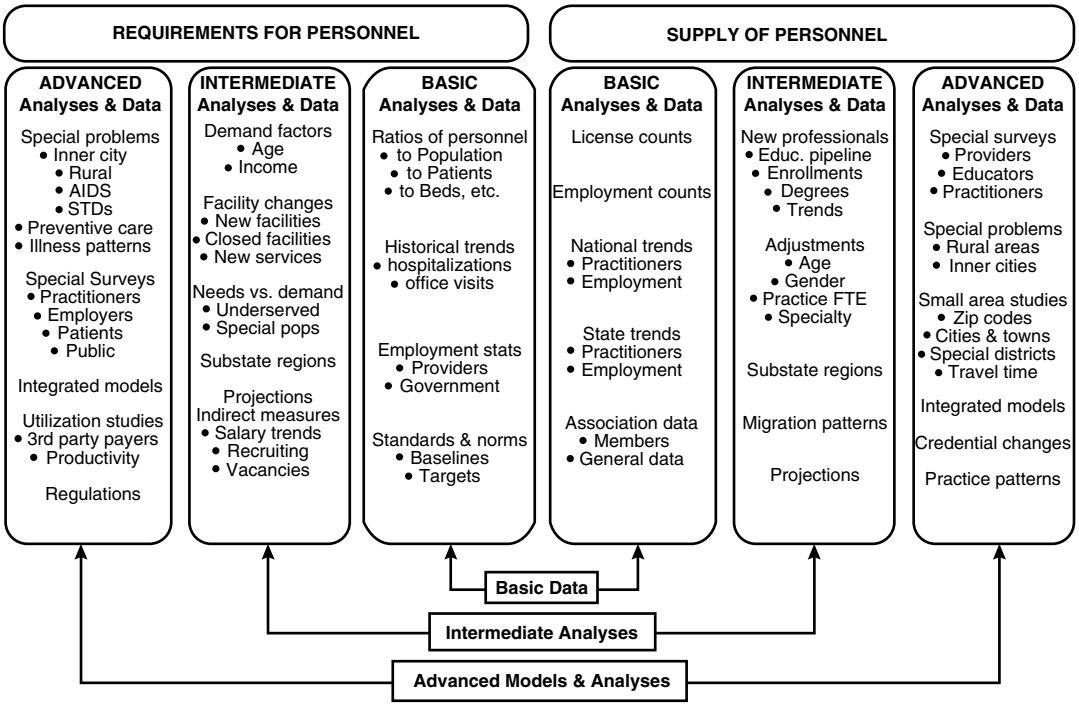

Figure 1 The health workforce data analysis hierarchy. Adapted from [4], based on [15, Figure 1]. Reproduced from *Physical Therapy* by permission of the American Physical Therapy Association

adequately measured and trends can give us a hint as to what the underlying demand might be.

A comprehensive licensure data system – such as maintained in the state of North Carolina – takes time to implement, but, with periodic resurveys, the quality of data improves. Additional items can be added to increase the usefulness of the database. Data generated in North Carolina using this model are now available on location, employment setting, and type of employment. These data provide helpful information by projecting demand as well as supply, as mobility in and out of employment sectors can be quite sensitive to economic trends.

One of the ways in which demand can be estimated from supply data is by looking at the different sectors in which health professionals are employed, and by comparing the kinds of employment for the entire workforce with those of the newly licensed individuals in a given year. For example, in North Carolina, 36.8% of the currently licensed individuals in physical therapy are working primarily in hospitals. This figure has fluctuated between 35% and 40% over the past decade, even as the total number of therapists has increased. Among newly licensed personnel in 1994, including both new graduates and immigrants to the state, the proportion employed in hospitals is 64%. This proportion suggests that hospital employment is especially attractive to new graduates and immigrants. By examining employment trends, it may be possible to determine whether this percentage increase will continue (representing an increase in employment in that particular sector) or whether it represents the employment patterns of new entrants who subsequently move to other sectors.

The methods used here to assess requirements emphasize trends rather than needs or demands. Obviously, such an approach is subject to misinterpretation, but it uses available data from supply to assess changes in different sectors to provide "reality checks" on the more idealistic need-based models and the more abstract and data-intensive demand-based models. Such an approach also provides information at the state and local levels, where decisions about expansion in the number and size of training institutions are likely to occur.

Emerging Issues

As the data requirements for both supply and demand models become more complex and sophisticated, and as various elements in national health systems are becoming decentralized and privatized, new issues are coming to the fore. Chief among these are how to resolve inherent tensions between the need for regulation and accountability expressed by central health authorities, credentialling bodies, national payment systems, and health professional educational systems, and the need for privacy and **confidentiality** expressed by individual health professionals, their associations, and health service delivery systems, which increasingly employ health professionals. As data requirements for workforce planning become more complex, they also become more onerous to individuals. Health workforce planners of the next generation will be faced with the challenge to use creative and innovative strategies to acquire data at increasingly detailed levels while preserving the confidentiality of the sources.

New information technologies also provide unprecedented opportunities for health workforce modeling. For example, the availability of various kinds of **simulation** software for microcomputers allows the development and display of simulations and **sensitivity analyses** in real time. In addition, the use of geographic information systems (GIS) allows cartographic presentation of databases containing disease patterns (*see* **Mapping Disease Patterns**) and demographic data. Overlaying the location of health professions or activity space data on such maps can provide dramatic opportunities to identify issues of distribution which might well remain opaque in the absence of such visual displays. Creatively applied, such processes can be conducted in group settings with panels of health professions, education administrators, and health policy analysts in attendance in such a way that not only engages their attention, but also serves to close the loop between planning and policy.

Conclusion

Health workforce modeling has been employed at a variety of organizational levels and has either concentrated on a single health profession or considered multiple health professions interacting to provide a spectrum of health care services. At the micro level, models have been developed to analyze one or more specific types of personnel in a specific delivery setting (e.g. physician assistants employed in the office of a generalist physician). Models have been developed to analyze various kinds of personnel in an organized delivery system (e.g. all physician specialties in an integrated health care delivery system which includes inpatient, outpatient, long-term, and home health care). Some models have covered specific kinds of personnel in a regional, state, or national framework with the purpose of forecasting future workforce structure and affecting health workforce policy (e.g. physical therapists in the US or all licensed health professions practicing within the boundaries of a given state, perhaps at a county level of disaggregation).

Although not formally workforce modeling, population-based "benchmarking" also has been applied recently in health workforce studies as a way to get at the notion of "requirements", while avoiding the question of whether "needs" or "demands" are being met. This approach compares a priori standard ratios of health professionals to populations (either normative or those extant in particular health systems) to the range of these ratios across **hospital market areas**, broad regions, or national health systems. A recent application of this approach used managed care ratios in the US to examine how variations in supply and composition of the physician workforce relate to the organization of the health care delivery system in different areas [2].

This article concentrates on applications at the health system or national levels because most of the recently published material is directed at macro-level analyses. This concentration of published material at higher levels of aggregation is a result of the recent emphasis on workforce reform as a part of health system reform proposals, and is also a consequence of single-site studies being described most frequently in less accessible internal documents of the firms in which the analyses were performed. An excellent earlier summary of **operations research** applications that spanned this entire spectrum appears in [6]. In addition, Hall & Mejia [5] provided a comprehensive monographic summary of the various approaches up to the mid-1970s, with a special focus on techniques applicable to developing countries and feasible for health workforce planning as a component of more general health and development strategies.

References

[1] Capilouto, E. (1995). A review of methods used to project the future supply of dental personnel and the future demand and need for dental services, *Journal of Dental Education* **59**, 237–257.

[2] Goodman, D.C. (1996). Benchmarking the US physician workforce: An alternative to needs-based or demand-based planning, *Journal of the American Medical Association* **276**, 1811–1817.

[3] Greenberg, L. (1992). *Forecasting the Future Supply of Physicians: Logic and Operation of the BHPr Physician Supply Model*. OHPAR Report No. 3-93, BHPr. Rockville.

[4] Hack, L.M. & Konrad, T.R. (1995). Determination of supply and requirements in physical therapy: Some considerations and examples, *Physical Therapy* **75**, 52.

[5] Hall, T.L. & Mejia, A. (1978). *Health Manpower Planning: Principles, Methods, Issues*. World Health Organization, Geneva.

[6] Levin, E. & Kahn, H.D. (1975). Health manpower models, in *Operations Research in Health Care: A Critical Analysis*, L.J. Shulman, R.D. Speas, Jr & J.P. Young, eds. The Johns Hopkins University Press, Baltimore, pp. 337–364.

[7] Lipscomb, J. & Alexander, B.J. eds (1992). *Institute of Medicine, Physician Staffing for the VA*, Vol. II. National Academy Press, Washington.

[8] Lipscomb, J., Parmigiani, G. & Hasselbad, V. (1997). Combining expert judgment by hierarchical modeling: an application to physician staffing, *Management Science* (to appear).

[9] Moses, E. & Sekscenski, T. (1997). Bureau of Health Professions' integrated requirements model, in *Combined Proceedings of the Seventh and Eighth Federal Forecasters' Conferences*, D.E. Gerald, ed. US Department of Education, National Center for Educational Statistics, Washington, pp. 69–77.

[10] Reinhardt, U.E. (1981). The GMENAC forecast: An alternative view, *American Journal of Public Health* **71**, 1149–1157.

[11] Reinhardt, U.E. (1996). The economic and moral case for letting the market determine the health workforce, in *The U.S. Health Workforce: Power, Politics, and Policy*. M. Osterweis et al. eds. Association of Academic Health Centers, Washington, pp. 3–13.

[12] Schwartz, W.B. (1988). Why there will be little or no physician surplus between now and the year 2000, *New England Journal of Medicine* **318**, 892–897.

[13] Traxler, H. (1994). Physician supply modeling in the United States of America and its uses in assisting policy making, *World Health Statistics Quarterly* **47**, 118–125.

[14] US Department of Health and Human Services (1981). *Summary Report of the Graduate Medical Education National Advisory Committee to the Secretary, Department of Health and Human Services*, Vol. I (GMENAC Summary Report), DHHS Publ. No. (HRA) 81-651, Government Printing Office, Washington.

[15] US Department of Health and Human Services, Public Health Service, Health Resources and Services Administration, Bureau of Health Professions, Office of Health Professions (1992). *Data Systems to Support State Health Personnel Planning and Policymaking: A Resource Guide for State Agencies, Washington*. Analysis and Research Report No. 2–93. Washington.

[16] Weiner, J.P. (1994). Forecasting the effects of health reform on US physician workforce requirements. Evidence from HMO staffing patterns, *Journal of the American Medical Association* **272**, 222–230.

KERRY E. KILPATRICK & THOMAS R. KONRAD

Healthy Worker Effect

Epidemiological studies of occupationally or environmentally exposed cohorts are useful in identifying and quantifying environmental risks, because workers are generally exposed to higher levels of toxic materials than the general population. Although higher exposures make it easier to detect and characterize modest elevations in risk, such studies are plagued by a particularly pervasive form of selection bias, referred to as the healthy worker effect (HWE). The HWE is reflected in better health status of workers relative to the general population. This bias arises because workers were healthy enough to be hired initially and those who remain at work stayed healthy enough to maintain employment, whereas general populations include persons unfit for work because of impaired health.

The HWE can be regarded as a form of *selection bias* because workers are selected preferentially on the basis of health status, either by themselves or as a result of job requirements. There are two sources of the HWE: the initial selection of healthier individuals at the time of hire and the survival of healthier individuals among the exposed workers. The latter, healthy worker survivor bias, results from less healthy workers leaving the workforce as well as less healthy workers transferring to jobs with lower exposure(s). Both forms of self-selection can distort the shape of the exposure–response relationship and lead to bias towards the null.

Identifying the Bias

The HWE was originally identified as a feature of cohort mortality studies [6]. In such time-to-event studies, the standardized mortality ratio (SMR) for all causes of death combined provides a convenient measure of the extent to which the bias is operating. By this measure, the HWE has been found to wear off with length of follow-up [8]. This same pattern of rising SMRs with increasing time since hire or length of follow-up has also been consistently observed for specific causes of death, including cardiovascular disease and nonmalignant respiratory and digestive diseases [8]. The increase in the SMR during the time period just after leaving work provides evidence that the HWE is stronger for cardiovascular disease mortality than for cancers [2]. Despite evidence that the HWE also affects cancer mortality rates, particularly in the period close to hire, the lack of consistency in this finding suggests that a weaker bias is operating.

Selection bias due to the healthy worker survivor effect also arises in cross-sectional studies of morbid outcomes and longitudinal studies of recurrent events, although it cannot be as easily identified [3]. The bias is particularly prominent in cross-sectional studies of active workers. In such studies, the undersampling of short-term workers results in a disproportionate number of survivors, i.e. workers who are more resistant to the effects of exposure. This is often found in cross-sectional studies of acute toxicity where the most responsive or sensitive workers are ‘selected out’ of exposure.

By contrast, longitudinal studies with repeated measures of exposure and response over time are frequently used to study the development of chronic conditions by observing changes in a continuous physiological parameter. Like cohort studies, longitudinal studies often involve measures of exposure that vary over time. Time-varying exposures and confounders complicate exposure–response analyses because the inter-relationships between time-dependent variables makes it more difficult to control the HWE. On the other hand, information about changes in exposure over time allows for analytical strategies to address the HWE in ways that are impossible in more limited study designs.

The hypothesis in most studies of time-varying occupational exposures is that past or current exposure affects current health status. However, causality can also move in the reverse direction, with past health status impacting subsequent exposure either via job transfer or leaving work. This reversal in the direction of a causal relationship has been referred to as *feedback*. Feedback is a key feature of the HWE, whereby current exposure and health status may both be affected by past exposure and health status as well as affect subsequent measures.

Over the past 25 years much research has focused on methods to reduce this form of selection bias. This effort has produced a large literature on the topic of the HWE, but the problem has so far proven to be easier to detect than to fix. Several strategies have been proposed to address the HWE in longitudinal studies of time-to-event or repeated measures. The empirical basis of each method and its effectiveness in reducing bias due to selective hire, leaving work,

or job transfer in studies of occupational hazards is described below.

Reducing the Bias

Using an Internal Reference Group

In its simplest form, the HWE is recognizable as a lower mortality risk among an occupational cohort compared to a national or regional reference population. Whether this is viewed as a problem of unmeasured confounding or selection bias, some of the distortion can be eliminated by the selection of a more appropriate comparison group. An unexposed group of workers is needed to avoid unmeasured confounding by health status. An internal analysis that relies on an unexposed, or less exposed, subset of the study population as the reference group will reduce the HWE. This approach will eliminate the component of the bias due to the hiring of healthier workers, since all subjects were similarly hired.

The component of the HWE due to the survival of the healthier workers in the active workforce, however, will not be corrected by the use of an internal reference group. If sicker workers systematically leave work or transfer out of the more highly exposed jobs, the HWE will persist. For example, as described in more detail below, because workers who develop asthma in jobs with high exposure are more likely to leave work or be transferred than workers in jobs with low exposure, comparisons between workers with more and less exposure may result in a negative exposure–response curve even in the presence of a hazard.

Lagging Exposure

Lagging exposure is a method used in cancer epidemiology to account for disease latency by assigning zero weight to exposures that occur in a time period just prior to the observed health event. It has also been proposed as a strategy to reduce the bias caused by selective leaving [7]. The rationale is that recent exposures should be ignored because only the healthiest survivors remain exposed. For example, lagging exposure to arsenic by 10 to 20 years increased the rate ratios for respiratory cancers in a reanalysis of Enterline's cohort study described by Arrighi and Hertz-Picciotto [1].

A similar approach involving exposure weighting was applied to address HWE due to selective job transfer in a reanalysis of a cross-sectional study of asthma with information on date of diagnosis and retrospective exposure information over time [4]. In the original unadjusted analysis, the odds ratio for asthma prevalence was estimated in a **logistic regression** model based on cumulative exposure to metalworking fluids up to the time of the cross-sectional survey. To account for the possibility that workers transfer to lower exposure jobs after developing asthma, zero weight was assigned to exposures that occurred after the reported age of onset. Using a proportional hazards model, exposure was cumulated only up to the age of onset for each case and all subjects in the risk set for that case. In contrast with the unlagged analysis, a significantly elevated incidence rate ratio was observed for synthetic fluid exposure.

Although exposure lagging has effectively reduced the HWE in several instances, this approach has limitations. Lagged exposures cannot be used in studies of acute health effects because the most recent exposures are probably the most biologically relevant. Even in studies of chronic conditions, such as cardiovascular disease, current exposures may continue to exert an effect and exposure lagging may not be a viable option.

Adjusting for Time Since Hire

The HWE, as measured by the SMR for all causes of death combined, declines with increasing time since hire (or length of follow-up) [8]. It follows, as suggested by Fox and Collier [6], that analysis restricted to the subgroup with longer follow-up will be less biased by the HWE. Alternatively, time since hire can be viewed as a simple confounder, i.e. a risk factor for the health outcome in the absence of exposure, and associated with exposure. It then follows that HWE can be addressed using either of the standard methods to control confounding: stratification or inclusion of a marker for the HWE as a covariate in a multiple regression model. The diminished HWE that occurs with increasing time since hire may even occur independent of exposure. Flanders has shown that even when there is no effect, adjusting for time since hire reduces bias [5].

By contrast, the HWE does *not* decline with increasing duration of employment. In fact, it

increases because the healthiest workers manage to remain employed longest. Steenland and Staynor suggested that an explanation for the difference is that duration of employment includes mostly active person years whereas time since hire includes an increasing proportion of inactive person time [12].

Adjusting for Employment Status

Observed mortality rates for workers after leaving employment are double the rates among active employees [12, 13]. Steenland and Staynor have suggested that employment status acts as a negative confounder under the null hypotheses [12]. In this case, the HWE might be reduced by treating employment status (on or off work) as a time-varying confounder and identifying each person year as either one in which the subject was actively working or off work. Based on combined data from 10 cohorts selected by the US National Institute for Occupational Safety and Health, adjustment for employment status reversed the negative trends between duration of exposure and all cause mortality [13].

Adjustment for employment status as a confounder will, however, itself lead to bias if subjects in more highly exposed jobs leave the workforce at a different rate than workers in jobs with less exposure. The bias will occur whether the exposure causes individuals to leave work because it seriously impairs their health (in which case leaving work is an intermediate variable on the causal pathway from exposure to disease) or simply because high-exposure jobs have a higher or lower turnover rate due either to the irritant effects of exposure (say on the eyes) or because more highly exposed jobs are often preferentially lost when economic downturns occur. Steenland et al. simulated data based on two alternative hypotheses: (a) cumulative exposure affects leaving but does not necessarily affect disease, and (b) cumulative exposure also affects disease. Results demonstrated that under either hypothesis, adjusting for confounding by controlling for employment status in the analysis results in bias [13]. Epidemiologists recognize that if a covariate is a confounder it must be controlled for in the analysis to reduce bias. However, we have argued here that when a covariate is also affected by earlier exposure (e.g. the covariate is an intermediate variable on the causal pathway from exposure to disease), controlling for that variable will, in fact, introduce bias.

None of the straightforward approaches described above can adequately address this problem: employment status is likely to act as a confounder and also be influenced by exposure. It is this aspect of HWE, which we refer to as the healthy worker survivor effect, that limits the effectiveness of all the methods reviewed to this point, and provides the motivation for the more analytically complex method of estimation known as G-estimation.

G-estimation

In studies with information on exposure and work status over time, change in work status can both depend on past exposure and affect future exposure. Since being off work is often related to health status as well as exposure, work status (active vs. inactive) can act as a confounder. If high exposure causes workers to preferentially leave work then, as noted above, bias can result whether or not one controls for work status in a standard analysis. G-estimation provides an appropriate adjustment for the effect of a time-varying exposure in the presence of a time-varying confounder, such as employment status, which is influenced by past exposure as well [9–11].

G-estimation is an example of structural equations modeling in which causal structures are integrated into observational data analysis to model causal relationships [9, 10]. The causal models are structural nested failure time models for the effect of a time-dependent exposure on a survival time outcome. The causal parameter is estimated using a semiparametric method by treating the cohort data as a sequential randomized trial in which exposure at time t is randomly assigned, conditional on past exposure and employment history and on baseline variables such as race, gender, age at hire and calendar period of hire.

G-estimation of structural nested failure time models also allows one to estimate the magnitude of the exposure effect. The exposure–response parameter has a direct interpretation as the fraction of years of life lost due to continuous exposure at a unit dose. Robins et al. show how to convert this estimate into an estimate of the causal mortality ratio comparing an always exposed to a never exposed worker population [11]. Thus the method can theoretically be used to provide unbiased estimates of the effect of the cumulative exposure variable typically used in

studies of chronic disease, even in the presence of selection bias due to the healthy worker survivor effect.

In a structural nested failure time model the causal effect of exposure is quantified by contrasting observed outcomes and exposure histories with counterfactual outcomes, e.g. the outcomes exposed subjects *would have had* in the absence of exposure. Although such counterfactual outcomes are not directly observed for exposed workers, under the sequential randomization assumption mentioned above this contrast can be unbiasedly estimated from the available data. Simply compare, at each time t, the subsequent survival experience of each subject at work at time t with other subjects also at work at time t with the same exposure and employment history prior to t but with a different exposure at t. Subjects who are off work at time t are not used in this comparison because less healthy workers preferentially leave work and thus are not comparable to subjects remaining at work. The need for exact matching on past exposure and work history can be overcome by modeling the mean exposure at time t as a function of past exposure and employment history using logistic, polytomous logistic, or least squares regression models depending on whether exposure is dichotomous, polychotomous, or continuous.

In order to apply the G-estimation algorithm, data on whether a worker was actually on or off work at each time t must be available. If, instead, one incorrectly assumes all workers were continuously employed from their date of hire until their date of last employment, when in fact long lay-off periods are common, serious bias may result, regardless of the method of analysis.

G-estimation has been used to estimate the effect of azidothymidine (AZT) on Kaposi's sarcoma, the effect of aerosolized pentamidine on survival in acquired immune deficiency syndrome (AIDS) patients, the effect of smoking cessation on heart disease mortality in the Mr Fit study, the effect of graph vs. host disease on time to recurrence in leukemic patients treated with bone marrow transplantation, and the effect of systolic hypertension on coronary artery disease in the Framingham cohort. To date, however, there has been no successfully completed application of this methodology to the analysis of occupational or environmental cohort data, although several studies are currently underway.

References

[1] Arrighi, H.M. & Hertz-Picciotto, I. (1996). Controlling healthy worker survivor effect: an example of arsenic exposure and respiratory cancer, *Occupational and Environmental Medicine* **53**, 455–462.

[2] Blair, A., Stewart, P., O'Berg, M., Gaffey, W., Walrath, J., Ward, J., Bales, R., Kaplan, S. & Cubit, D. (1986). Mortality among workers exposed to formaldehyde, *Journal of the National Cancer Institute* **76**, 1071–1084.

[3] Eisen, E.A. (1995). Healthy worker effect in morbidity studies, *Medicina del Lavoro* **86**(2), 127–140.

[4] Eisen, E.A., Holcroft, C.R., Greaves, I.A., Wegman, D.H., Woskie, S.R. & Monson, R.R. (1997). A strategy to reduce healthy worker effect in a cross-sectional study of asthma and metal working fluids, *American Journal of Industrial Medicine* **31**, 671–677.

[5] Flanders, W.D., Cardenas, V.M. & Austin, H. (1993). Confounding by time since hire in internal comparisons of cumulative exposure in occupational cohort studies, *Epidemiology* **4**, 336–341.

[6] Fox, A.J. & Collier, P.F. (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry, *British Journal of Social Medicine* **30**, 225–230.

[7] Gilbert, E.S. (1982). Some confounding factors in the study of mortality and occupational exposures, *American Journal of Epidemiology* **116**, 177–188.

[8] Monson, R.R. (1986). Observations on the healthy worker effect, *Journal of Occupational Medicine* **28**, 425.

[9] Robins, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-Applications to the control of the healthy worker survivor effect, *Mathematical Modelling* **7**, 1393–1512.

[10] Robins, J.M. (1992). Estimation of the time-dependent accelerated failure-time model in the presence of confounding factors, *Biometrika* **79**, 321–334.

[11] Robins, J.M., Blevins, D., Ritter, G. & Wufsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for *Pneumocystis carini* pneumonia on the survival of AIDS patients, *Epidemiology* **3**, 19–336.

[12] Steenland, K. & Staynor, L. (1991). The importance of employment status in occupational cohort studies, *Epidemiology* **2**, 418–423.

[13] Steenland, K., Deddens, J., Salvan, A. & Staynor, L. (1996). Negative bias in exposure-response trends in occupational studies: modeling the healthy worker survivor effect, *American Journal of Epidemiology* **143**, 202–210.

(*See also* **Attributable Risk**; **Occupational Mortality**)

ELLEN A. EISEN & JAMES M. ROBINS

Hepatology

Hepatology is the study of diseases of the liver. These can be mainly classified as hepatitis, hepatocellular carcinoma (or liver cancer), and liver cirrhosis.

Hepatitis

Several distinct infections are included under the generic title of hepatitis. There are many similarities between these different forms of hepatitis, but their epidemiologies and methods of prevention and control vary. These infections are labeled as hepatitis A, B, C, D, and E. Hepatitis D is sometimes called delta hepatitis [3]. Most statistical work has been done on hepatitis A and B, with little on other forms of hepatitis.

Hepatitis A and B occur worldwide. Outbreaks of hepatitis A are patchy and tend to occur in regular cycles. For developed countries disease spreads in day-care centres for children in diapers, to household and sexual contacts of acute cases, intravenous drug users, and travelers to endemic countries. Hepatitis A is spread by the fecal–oral route. Contaminated water supplies, handling and preparation of food by infected foodhandlers, and shellfish have all been responsible for outbreaks. Hepatitis B is endemic with little seasonal variation in incidence. In developed countries such as the US, infection is most common in young adults, whereas in developing countries widespread infection occurs in infancy. Hepatitis B infection is common in certain high risk groups: intravenous drug injectors, promiscuous heterosexuals, male homosexuals, and workers in some health care and public safety occupations. It is spread by infectious blood, saliva, semen, and vaginal fluids.

Hepatitis C is transmitted by infected blood and blood products, and occurs virtually everywhere in the world. It accounts for 15%–40% of community-acquired hepatitis cases. High-risk groups include transfusion recipients, intravenous drugabusers, and dialysis patients. Hepatitis D closely resembles and is often associated with hepatitis B infection. Its mode of transmission is also very similar. Hepatitis E closely resembles hepatitis A in both its clinical symptoms and its epidemiology. The attack rate is highest amongst young adults, especially males.

Hepatitis A

Frösner et al. [8] discussed the decrease in incidence of hepatitis A infections in Germany using serological data. They used a catalytic model (*see* **Communicable Diseases**) with a sigmoidal decrease in the force of infection (*see* **Hazard Rate**). The force of infection fell from 0.04 per year in 1945 to 0.005 per year in 1965. Frösner et al. [9] and Schenzle et al. [23] discussed antibodies against hepatitis A in seven European countries. Prevalence was highest in Greece and France and lowest in Scandinavia. The force of infection had declined almost everywhere in the period leading up to 1979. Keiding [14] considered nonparametric estimation (*see* **Nonparametric Methods**) of the age-specific force of infection applied to serological hepatitis A data for Bulgaria. These data are ideal for statistical **estimation** as they were collected before the advent of mass vaccination. Keiding estimated the proportion of people of different ages who must be vaccinated to eliminate hepatitis A in Bulgaria. Greenhalgh & Dietz [10] extended this work to an age-structured model and vaccination at several different ages. They examined the effect of different mixing patterns on vaccination campaigns.

Hadeler et al. [11, 12] performed a statistical analysis of the outbreaks of hepatitis A in Maricopa County, Arizona. These studies strongly link the spread of hepatitis A in the US to very young children in day-care centres and provide a framework for designing disease control strategies. Sattenspiel [20] developed a matrix-migration model for the spread of hepatitis A in US day-care centers using these results. The theoretical results of Sattenspiel's model were applied to data on the incidence of hepatitis A in Alberquerque, New Mexico, in 1979. Analysis of the data suggested that local clusters were at higher risk for epidemics. Close social ties linked up these centers in small local clusters which helped explain the disproportionate number of cases associated with these centers. Sattenspiel [21] described two stochastic **simulation** models which supported these results. Sattenspiel & Simon [22] pushed the theoretical development of the model further.

Liu [15] considered a differential equation **epidemic model** for hepatitis A where the duration of the **latent period** depends on the number of infectious individuals. He showed that nonlinearity due to a dose-dependent latent period can cause periodicity. This model was compared with US hepatitis data.

Hepatitis B

Hepatitis B infects people worldwide. The highest rates of infection are in sub-Saharan Africa and East Asia. Early mathematical models for hepatitis B were due to Cvetanovic et al. [5] and Pasquini & Cvetanovic [19], who used a compartment model (*see* **Pharmacokinetics and Pharmacodynamics**) in which the host population was stratified clinically and epidemiologically to investigate a variety of control strategies in Mediterranean countries. Anderson & May [1] developed a differential equation model for the spread of hepatitis B. A key feature was that around 1% of infected individuals became carriers and continued to transmit the disease for the rest of their lives. These carriers are an important reservoir of infection and their presence represents a complication for immunization programs. Anderson et al. [2] described a model for the sexual transmission of hepatitis B in developed countries which included heterogeneous mixing with respect to age and sexual activity class. They used this model to assess the effects of vaccination campaigns. The first dynamic model of hepatitis B transmission in developing countries has been developed by McLean & Blumberg [18].

Edmunds et al. [7] discussed the influence of age on the development of the carrier state. A model was fitted to the data using **maximum likelihood**. Infants infected perinatally were found to have a high probability, 0.885, of becoming carriers. Over early childhood there is a sharp decrease in the proportion of infections which lead to the carrier state. By adulthood the probability of becoming a carrier was about 0.1. Implications for vaccination programs were also discussed. Edmunds et al. [6] outlined a deterministic compartmental model to describe the transmission dynamics and control of hepatitis B in the Gambia. The model included a class of carriers. They examined the impact of mass vaccination on the incidence of liver cancer (as carriers have a higher than average chance of developing liver cancer). They used age-structured serological data to estimate parameters. Two models were outlined which assumed that infection in adults was due to horizontal and sexual transmission, respectively.

Cirrhosis

Liver cirrhosis is a chronic disease of the liver, normally suffered by alcoholics, but it can also be caused by chronic hepatitis C infection. Carriers of hepatitis B also have a higher risk of developing cirrhosis [3]. Hepatocellular carcinoma occurs in 10%–25% of cirrhotic patients [17]. The prevalence of cirrhosis in the population is not known exactly. This is partly due to the fact that many cases are clinically silent. Up to 30% or even 40% of cases may be discovered at autopsy, and an unknown proportion remains clinically silent. There may be marked geographical differences in incidence from one country to another, or even between different regions in the same country [16]. Moreover, the proportion of alcoholic and nonalcoholic cirrhosis differs from one country to another, the prevalence of alcoholic cirrhosis being highest generally in wine-producing countries [17].

Hepatocellular Carcinoma

Primary hepatocellular cancer (PHC) or hepatocellular carcinoma (HCC) is recognized worldwide. It is among the most common malignant neoplasms in China, many parts of Asia, and Africa. It is relatively uncommon in the US and Europe. Chronic infection with hepatitis B virus is an important risk factor in most cases; hepatitis C may also be involved. Most patients go through a stage of liver cirrhosis before development of the tumor [3].

Berman [4] and later Higginson [13] called world attention to the extremely high incidence rate of HCC amongst the black male population in Mozambique. From the statistical data from various geographical locations it seems that the greater the incidence rate, the younger the peak age. Among Mozambican males the peak age is between 25 and 34 years, the average age in Japan is 56.8 years in males and 59.9 years in females, and it is higher in Northern Europe [17]. HCC occurs in more advanced ages in alcoholic cirrhosis.

References

[1] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[2] Anderson, R.M., Medley, G.F. & Nokes, D.J. (1992). Preliminary analyses of the predicted impacts of various vaccination strategies on the transmission of hepatitis B virus, in *The Control of Hepatitis B: the Role of Prevention in Adolescence*, D.L. Bennett, ed. Gower Medical Publishing, London, pp. 95–130.

[3] Beneson, A.S. (1990). *Control of Communicable Diseases in Man*, 16th Ed. American Public Health Association, Washington.

[4] Berman, C. (1951). *Primary Carcinoma of the Liver*. Lewis, London.

[5] Cvetanovic, B., Delimar, B., Kosicek, M., Likar, M. & Spoljaric, B. (1984). Epidemiological model of hepatitis B, *Annals of the Academy of Medicine* **13**, 175–184.

[6] Edmunds, W.J., Medley, G.F. & Nokes, D.J. (1997). The transmission dynamics and control of hepatitis B in The Gambia, *Statistics in Medicine* **15**, 2215–2234.

[7] Edmunds, W.J., Medley, G.F., Nokes, D.J., Hall, A.J. & Whittle, H.C. (1993). The influence of age on the development of the hepatitis B carrier state, *Proceedings of the Royal Society of London, Series B* **253**, 197–201.

[8] Frösner, G., Willers, H., Müller, R., Schenzle, D., Deinhardt, F. & Höpken, W. (1978). Decrease in incidence of hepatitis A infections in Germany, *Infection* **6**, 259–260.

[9] Frösner, G., Papavangelou, G., Butler, R., Iwarson, S., Lindholm, A., Courouce-Pauty, A., Hass, H. & Deinhardt, F. (1979). Antibodies against hepatitis A in seven European countries, *American Journal of Epidemiology* **110**, 63–69.

[10] Greenhalgh, D. & Dietz, K. (1994). Some bounds on estimates for reproductive ratios derived from the age-specific force of infection, *Mathematical Biosciences* **124**, 9–57.

[11] Hadeler, S.C., Erben, J.J., Francis, D.P., Webster, H.M. & Maynard, J.E. (1982). Risk factors for hepatitis A in day-care centers, *Journal of Infectious Diseases* **145**, 255–261.

[12] Hadeler, S.C., Webster, H.M., Erben, J.J., Swanson, J.E. & Maynard, J.E. (1980). Hepatitis A in day-care centres, *New England Journal of Medicine* **302**, 1222–1227.

[13] Higginson, J. (1963). The geographical pathology of primary liver cancers, *Cancer Research* **23**, 1624–1633.

[14] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective, *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[15] Liu, W. (1993). Dose-dependent latent period and periodicity of infectious diseases, *Journal of Mathematical Biology* **31**, 487–494.

[16] Marubini, E. (1987). Epidemiology of cirrhosis, in *Cirrhosis of the Liver*, N. Tygstrup & F. Orlandi, eds. Elsevier, Amsterdam, pp. 275–294.

[17] McIntyre, N., Benhamou, J.-P., Bircher, J., Rizetto, M. & Rhodes, J. (1991). *Oxford Textbook of Clinical Hepatology*, Vols 1 and 2. Oxford University Press, Oxford.

[18] McLean, A.R. & Blumberg, B.S. (1994). Modelling the impact of mass vaccination against hepatitis B. I. Model formulation and parameter estimation, *Proceedings of the Royal Society of London, Series B* **256**, 7–15.

[19] Pasquini, P. & Cvetanovic, B. (1988). Mathematical models of hepatitis infection, *Annali Istituto Superiore di Sanita* **24**, 245–250.

[20] Sattenspiel, L. (1987). Population structure and the spread of disease, *Human Biology* **59**, 411–438.

[21] Sattenspiel, L. (1987). Epidemics in nonrandomly mixing populations: a simulation, *American Journal of Physical Anthropology* **73**, 251–261.

[22] Sattenspiel, L. & Simon, C.P. (1988). The spread and persistence of infectious diseases in structured populations, *Mathematical Biosciences* **90**, 341–366.

[23] Schenzle, D., Dietz, K. & Frösner, G. (1979). Hepatitis A antibodies in European countries. II. Mathematical analysis of cross-sectional surveys, *American Journal of Epidemiology* **110**, 70–76.

DAVID GREENHALGH

Heritability

Before discussing what genetic heritability is, it is important to be clear about what it is not. For a binary trait, such as whether or not an individual has a disease, heritability is not the proportion of disease in the population attributable to, or caused by, genetic factors. For a continuous trait, genetic heritability is not a measure of the proportion of an individual's score attributable to genetic factors. Heritability is not about cause *per se*, but about the causes of variation in a trait across a particular population.

Definitions

Genetic heritability is defined for a quantitative trait. In general terms it is the proportion of variation attributable to genetic factors. Following a genetic and environmental variance components approach, let Y have a mean μ and variance σ^2, which can be partitioned into genetic and environmental components of variance, such as additive genetic variance σ_a^2, dominance genetic variance σ_d^2, common environmental variance σ_c^2, individual specific environmental variance σ_e^2, and so on.

Genetic heritability in the narrow sense is defined as

$$\frac{\sigma_a^2}{\sigma^2}, \tag{1}$$

while genetic heritability in the broad sense is defined as

$$\frac{\sigma_g^2}{\sigma^2}, \tag{2}$$

where σ_g^2 includes all genetic components of variance, including perhaps components due to epistasis (gene–gene interactions; *see* **Genotype**) [3]. In addition to these random genetic effects, the total genetic variation could also include that variation explained when the effects of measured **genetic markers** are modeled as a **fixed effect** on the trait mean.

The concept of genetic heritability, which is really only defined in terms of variation in a quantitative trait, has been extended to cover categorical traits by reference to a **genetic liability model**. It is assumed that there is an underlying, unmeasured continuous "liability" scale divided into categories by "thresholds". Under the additional assumption that the liability follows a **normal distribution**, genetic and environmental components of variance are estimated from the pattern of associations in categorical traits measured in relatives. The genetic heritability of the categorical trait is then often defined as the genetic heritability of the presumed liability (latent variable), according to (1) and (2).

Comments

There is no unique value of the genetic heritability of a characteristic. Heritability varies according to which factors are taken into account in specifying both the mean and the total variance of the population under consideration. That is to say, it is dependent upon modeling of the mean, and of the genetic and environmental variances and covariances (*see* **Genetic Correlations and Covariances**). Moreover, the total variance and the variance components themselves may not be constants, even in a given population. For example, even if the genetic variance actually increased with age, the genetic heritability would decrease with age if the variation in nongenetic factors increased with age more rapidly. That is to say, genetic heritability and genetic variance can give conflicting impressions of the "strength of genetic factors".

Genetic heritability will also vary from population to population. For example, even if the heritability of a characteristic in one population is high, it may be quite different in another population in which there is a different distribution of environmental influences.

Measurement error in a trait poses an upper limit on its genetic heritability. Therefore traits measured with large measurement error cannot have substantial genetic heritabilities, even if variation about the mean is completely independent of environmental factors. By the definitions above, one can increase the genetic heritability of a trait by measuring it more precisely, for example by taking repeat measurements and averaging, although strictly speaking the definition of the trait has been changed also. A trait that is measured poorly (in the sense of having low **reliability**) will inevitably have a low heritability because much of the total variance will be due to measurement error (σ_e^2). However, a trait with relatively little measurement error will have a high heritability if all the nongenetic factors are known and taken into account in the modeling of the mean.

Fisher [1] recognized these problems and noted that

> whereas ... the numerator has a simple genetic meaning, the denominator is the total variance due to errors of measurement [including] those due to uncontrolled, but potentially controllable environmental variation. It also, of course contains the genetic variance ... Obviously, the information contained in [the genetic variance] is largely jettisoned when its actual value is forgotten, and it is only reported as a ratio to this hotch-potch of a denominator.

Historically, other quantities have also been termed heritabilities, but it is not clear what parameter is being estimated, e.g. Holzinger's $H = (r_{MZ} - r_{DZ})$ (the correlation between monozygotic twins minus the correlation between dizygotic twins) (*see* **Twin Analysis**) [2], Nichol's $HR = 2(r_{MZ} - r_{DZ})/r_{MZ}$ [5], the E of Neel & Schull [4] based on twin data alone, and Vandenburg's $F = 1/[1 - \sigma_a^2/\sigma^2)]$ [6]. Furthermore, the statistical properties of these estimators do not appear to have been studied.

References

[1] Fisher, R.A. (1951). Limits to intensive production in animals, *British Agricultural Bulletin* **4**, 217–218.

[2] Holzinger, K.J. (1929). The relative effect of nature and nurture influences on twin differences, *Journal of Educational Psychology* **20**, 245–248.

[3] Lush, J.L. (1948). Heritability of quantitative characters in farm animals, *Suppl. Hereditas* **1948**, 256–375.

[4] Neel, J.V. & Schull, W.J. (1954). *Human Heredity*. University of Chicago Press, Chicago.

[5] Nichols, R.C. (1965). The National Merit twin study, in *Methods and Goals in Human Behaviour Genetics*, S.G. Vandenburg, ed. Academic Press, New York.

[6] Vandenberg, S.G. (1966). Contributions of twin research to psychology, *Psychological Bulletin* **66**, 327–352.

JOHN L. HOPPER

Heterozygosity, Loss of

Loss of tumor suppressor **gene** function has long been implicated as an important event in the onset and progression of cancer. One assay commonly used to define chromosomal regions likely to contain novel tumor suppressor genes is loss of heterozygosity (LOH). LOH compares the chromosomal organization and stability of cells from normal tissues with those from tissues derived from various stages of tumor development, thus highlighting chromosomal regions that may harbor tumor suppressor genes. To date, multiple regions of LOH have been defined for essentially all tumor types, in some situations facilitating our understanding of the function of the underlying genes. For other cancers, however, the delineation of LOH boundaries for tumor subtypes is ongoing as investigators seek to determine the molecular chain of events leading to tumor progression. As characterization of individual tumor suppressor genes progresses, LOH will become a powerful diagnostic tool for clinical assessment of tumor stage and grade.

LOH detects chromosomal deletions within tumor cell populations by comparing allele patterns from a single individual's normal and tumor cells at a set of ordered genetic **markers** referred to as a "**haplotype**." The point at which a pattern changes from two haplotypes, representing the heterozygous state of the normal cell, to a single haplotype, representing the loss of all or part of one chromosomal arm containing a putative cancer gene, is used to define the boundaries of LOH for a single tumor (Figure 1). Historically, one problem associated with precisely defining such boundaries has been a lack of reproducibility across studies. This has been due, in part, to the small numbers of tumors analyzed by many studies, as well as the heterogeneous collections of tumors used. This problem can theoretically be solved by the development of tumor banks featuring large numbers of well-characterized tissues. A bigger concern, however, has been contamination of samples from adjacent, noncancerous tissues, which reduces the quantitative potential of the technique, as well as its ability to detect subtle molecular events. The introduction of laser capture microdissection to isolate virtually pure populations of tumor cells has almost eliminated this problem [11].

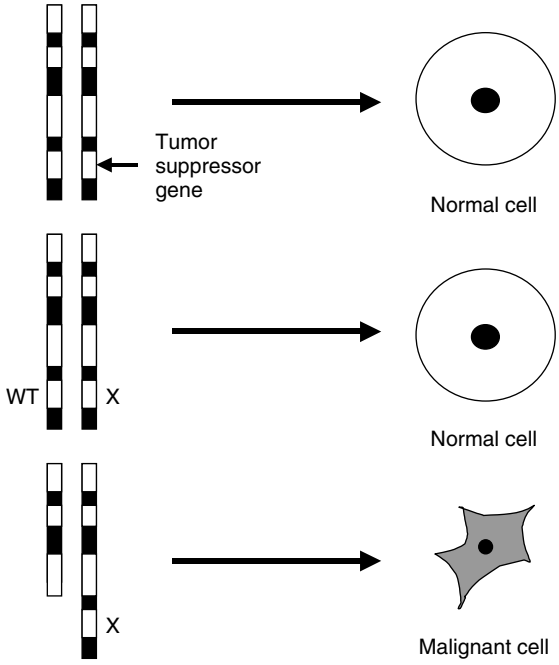

Figure 1 Normal cells require at least one functional copy of a critical tumor suppressor gene to remain noncancerous. When one inherited copy is nonfunctional because of a germ-line inactivation event and the remaining copy is lost as evidenced by LOH, malignancy results. WT: wild-type functional tumor suppressor gene. X: germ-line mutant nonfunctional tumor suppressor gene

Within any study, the utility of LOH for detecting the hierarchy of molecular events leading to disease is limited largely by the number of tumors representing the full range of pathologic stages and grades. Although stratification of data by clinical features of disease may limit overall power to detect rare events, the resulting increase in genetic homogeneity may facilitate detection of important chromosomal regions with statistical significance.

Because of the unique challenges faced in understanding the development of a tumor that is both common in the population and genetically heterogeneous (*see* **Genetic Heterogeneity**), we will consider the example of prostate cancer as we highlight the ways in which LOH can provide novel and useful information towards the identification of genes that drive the initiation, progression, and, ultimately, metastasis of cancer.

An Example: LOH and Prostate Cancer

It is estimated that nearly 198 100 men in the US will be newly diagnosed with prostate cancer, and approximately 31 500 men will die of the disease in 2001 [14]. Hence, identifying the genes responsible for the initiation and progression of the disease is of paramount interest [18]. The epidemiology of prostate cancer is consistent with a multistep process, with the premalignant lesion for most prostatic carcinomas being prostatic intraepithelial neoplasia (PIN) [2]. Autopsy results often reveal microscopic foci of well-differentiated prostate carcinoma, so-called "latent tumors". These tumors appear to be present in over 30% of men in their 60s and upwards of 70% of men in their 80s [3]. These are, in general, microscopic low-grade tumors that may never manifest clinically. The key to understanding prostate cancer, therefore, is to exploit technologies that allow us to understand the genetic differences between latent tumors and those that develop to clinically significant disease.

In the case of prostate cancer, there are several specific chromosomal regions that are frequently highlighted by LOH. One of the most comprehensive studies to date has been that of Saric et al. [20] who compared LOH profiles of 49 high-grade PIN, 22 primary prostate tumors, and 34 metastatic tumor foci using 37 microsatellite markers from 15 chromosomes. Tumor samples were microdissected from paraffin embedded tissues to greater than 75% purity. PIN was identified based on enlarged luminal secretory epithelium, frequently associated with enlarged nucleoli. Primary tumors were included if they were Gleason scores 3–8. Metastatic foci were obtained from 26 autopsies, collecting samples from lymph node, peritoneum, liver, and other tissues. Microsatellite markers were selected to span LOH regions defined by previously published studies, many of which were limited to the examination of only single chromosomes and/or included a heterogeneous mix of tumors. In the resulting analysis, significant levels of LOH were observed at only two regions (5q13–14 and 16q24.2) when considering high grade PIN, 15 regions when analyzing primary tumors, and 20 when metastatic tumors were analyzed. This concurs with results from others describing a relationship between increased genomic instability from PIN to primary cancer [26] and primary tumors to metastases [5]. Interestingly, both of these regions observed in high grade PIN were also lost in either primary or metastatic tumors, and 12 of the 15 LOH events defined in primary tumors (8p12–p21, 8p22, 8p23.1, 10q22–23, 11pter-q13, 16q22.1, 16q24.2, 16q24.3, 16q24, 17p13, 18q23, 21q22.3) were also observed in metastatic tumors.

Using LOH to Select Candidate Genes

Among the most frequently reported regions lost in studies of prostate tumor LOH are chromosomes 11 [6, 23] and 8 [8, 17, 19, 24]. Indeed, using a set of 38 microdissected samples of normal prostatic epithelium and invasive carcinoma, Dahiya and colleagues have described four distinct regions of LOH on chromosome 11, two each on the p and q arms [6]. Among the most interesting candidate genes on 11p is KAI1, which maps to 11p11.2, and has been shown to suppress metastasis when introduced into rat prostate cancer cells. In addition, expression of the gene is reduced in human cell lines derived from metastatic prostate tumors [9, 10]. Not surprisingly, levels of protein are inversely correlated with tumor grade [1]. The percentage of LOH or allelic imbalance at the KAI1 locus in metastatic tissues from autopsy cases is estimated at 70%, compared with 33% for clinically localized cases [16]. All of these studies suggest a key role for KAI1 in metastatic progression of prostate tumors and nicely illustrate the ways in which LOH can be used to map and subsequently enhance our understanding of tumor suppressor biology. Particular **candidate genes** have been investigated as a follow-up to many other regions of LOH. A comprehensive listing of such genes is included in a recent review by De Marzo et al. [7].

Most LOH studies to date have focused on a limited number of chromosomes. To best utilize the technique, high density, genome-wide scans of 300–400 markers are needed. At least one such study should include tumors derived from high-risk families of the type currently being used in genome-wide scans aimed at mapping prostate cancer susceptibility loci [13, 21, 22]. Given the apparent heterogeneity of prostate cancer, a comparison of potential germline and somatic events may prove useful for finding genes important in disease etiology. Indeed, Xu et al. have recently reported **linkage** to 8p22–23 in a set of 159 high-risk prostate cancer families [25]. Chromosome 8 is arguably the most frequently reported region of LOH for prostate

tumors [8, 17, 19, 24], with some studies reporting as many as 50% of tumors showing loss of all or part of the chromosome [17]. Joint analysis of representative data sets is clearly needed. In addition, comparing regions of frequent LOH from prostate tumors with other hormonally influenced tumors, such as breast and ovarian, may lend further insight to the existence of generalized tumor suppressor genes important in multiple hormonally regulated tissues. Several such regions have already been reported [4, 12, 15], suggesting some commonality between different tumor types.

Summary

LOH has been and continues to be one of the most important tools available for assessing genes important in cancer. The utility of the technique is limited only by characterization of the samples to which it is applied. Given the current abilities to isolate, purify, and characterize tumors, there is great potential for LOH to further inform us about tumor etiology.

References

[1] Bouras, T. & Frauman, A.G. (1999). Expression of the prostate cancer metastasis suppressor gene kai1 in primary prostate cancers: a biphasic relationship with tumour grade, *Journal of Pathology* **188**, 382–388.

[2] Brawer, M.K. (1992). Prostatic intraepithelial neoplasia: a premalignant lesion, *Journal of Cellular Biochemistry*, Supplement **16G**, 171–174.

[3] Carter, B.S., Ewing, C.M., Ward, W.S., Treiger, B.F., Aalders, T.W., Schalken, J.A., Epstein, J.I. & Isaacs, W.B. (1990). Allelic loss of chromosomes 16q and 10q in human prostate cancer, *Proceedings of the National Academy of Sciences* **87**, 8751–8755.

[4] Chen, C., Brabham, W.W., Stultz, B.G., Frierson, H.F., Jr, Barrett, J.C., Sawyers, C.L., Isaacs, J.T. & Dong, J.T. (2001). Defining a common region of deletion at 13q21 in human cancers, *Genes, Chromosomes & Cancer* **31**, 333–344.

[5] Cher, M.L., Bova, G.S., Moore, D.H., Small, E.J., Carroll, P.R., Pin, S.S., Epstein, J.I., Isaacs, W.B. & Jensen, R.H. (1996). Genetic alterations in untreated metastases and androgen-independent prostate cancer detected by comparative genomic hybridization and allelotyping, *Cancer Research* **56**, 3091–3102.

[6] Dahiya, R., McCarville, J., Lee, C., Hu, W., Kaur, G., Carroll, P. & Deng, G. (1997). Deletion of chromosome 11p15, p12, q22, q23–24 loci in human prostate cancer, *International Journal of Cancer* **72**, 283–288.

[7] De Marzo, A.M., Putzi, M.J. & Nelson, W.G. (2001). New concepts in the pathology of prostatic epithelial carcinogenesis, *Urology* **57**, 103–114.

[8] Deubler, D.A., Williams, B.J., Zhu, X.L., Steele, M.R., Rohr, L.R., Jensen, J.C., Stephenson, R.A., Changus, J.E., Miller, G.J., Becich, M.J. & Brothman, A.R. (1997). Allelic loss detected on chromosomes 8, 10, and 17 by fluorescence in situ hybridization using single-copy p1 probes on isolated nuclei from paraffin-embedded prostate tumors, *American Journal of Pathology* **150**, 841–850.

[9] Dong, J.T., Lamb, P.W., Rinker-Schaeffer, C.W., Vukanovic, J., Ichikawa, T., Isaacs, J.T. & Barrett, J.C. (1995). Kai1, a metastasis suppressor gene for prostate cancer on human chromosome 11p11.2, *Science* **268**, 884–886.

[10] Dong, J.T., Suzuki, H., Pin, S.S., Bova, G.S., Schalken, J.A., Isaacs, W.B., Barrett, J.C. & Isaacs, J.T. (1996). Down-regulation of the kai1 metastasis suppressor gene during the progression of human prostatic cancer infrequently involves gene mutation or allelic loss, *Cancer Research* **56**, 4387–4390.

[11] Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss, R.A. & Liotta, L.A. (1996). Laser capture microdissection, *Science* **274**, 998–1001.

[12] Filippova, G.N., Lindblom, A., Meincke, L.J., Klenova, E.M., Neiman, P.E., Collins, S.J., Doggett, N.A. & Lobanenkov, V.V. (1998). A widely expressed transcription factor with multiple DNA sequence specificity, ctcf, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers, *Genes, Chromosomes and Cancer* **22**, 26–36.

[13] Gibbs, M., Stanford, J.L., Jarvik, G.P., Janer, M., Badzioch, M., Peters, M.A., Goode, E.L., Kolb, S., Chakrabarti, L., Shook, M., Basom, R., Ostrander, E.A. & Hood, L. (2000). A genomic scan of families with prostate cancer identifies multiple regions of interest, *American Journal of Human Genetics* **67**, 100–109.

[14] Greenlee, R.T., Hill-Harmon, M.B. & Murray, T. (2001). Cancer statistics, 2001, *Ca: a Cancer Journal for Clinicians* **51**, 15–36.

[15] Huang, H., Qian, C., Jenkins, R.B. & Smith, D.I. (1998). Fish mapping of yac clones at human chromosomal band 7q31.2: identification of yacs spanning fra7g within the common region of loh in breast and prostate cancer, *Genes, Chromosomes and Cancer* **21**, 152–159.

[16] Kawana, Y., Komiya, A., Ueda, T., Nihei, N., Kuramochi, H., Suzuki, H., Yatani, R., Imai, T., Dong, J.T., Yoshie, O., Barrett, J.C., Isaacs, J.T., Shimazaki, J., Ito, H. & Ichikawa, T. (1997). Location of kai1 on the short arm of human chromosome 11 and frequency of allelic loss in advanced human prostate cancer, *Prostate* **32**, 205–213.

[17] Macoska, J.A., Trybus, T.M., Benson, P.D., Sakr, W.A., Grignon, D.J., Wojno, K.D., Pietruk, T. & Powell, I.J. (1995). Evidence for three tumor suppressor gene loci

on chromosome 8p in human prostate cancer, *Cancer Research* **55**, 5390–5395.

[18] Ostrander, E.A. & Stanford, J.L. (2000). Genetics of prostate cancer: too many loci, too few genes, *American Journal of Human Genetics* **67**, 1367–1375.

[19] Prasad, M.A., Trybus, T.M., Wojno, K.J. & Macoska, J.A. (1998). Homozygous and frequent deletion of proximal 8p sequences in human prostate cancers: identification of a potential tumor suppressor gene site, *Genes, Chromosomes and Cancer* **23**, 255–262.

[20] Saric, T., Brkanac, Z., Troyer, D.A., Padalecki, S.S., Sarosdy, M., Williams, K., Abadesco, L., Leach, R.J. & O'Connell, P. (1999). Genetic pattern of prostate cancer progression, *International Journal of Cancer* **81**, 219–224.

[21] Smith, J.R., Freije, D., Carpten, J.D., Grönberg, H., Xu, J., Isaacs, S. et al. (1996). Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search, *Science* **274**, 1371–1374.

[22] Suarez, B.K., Lin, J., Burmester, J.K., Broman, K.W., Weber, J.L., Banerjee, T.K., Goddard, K.A.B., Witte, J.S., Elston, R.C. & Catalona, W.J. (2000). A genome screen of multiplex prostate cancer sibships, *American Journal of Human Genetics* **66**, 933–944.

[23] Virgin, J.B., Hurley, P.M., Nahhas, F.A., Bebchuk, K.G., Mohamed, A.N., Sakr, W.A., Bright, R.K. & Cher, M.L. (1999). Isochromosome 8q formation is associated with 8p loss of heterozygosity in a prostate cancer cell line, *Prostate* **41**, 49–57.

[24] Vocke, C.D., Pozzatti, R.O., Bostwick, D.G., Florence, C.D., Jennings, S.B., Strup, S.E., Duray, P.H., Liotta, L.A., Emmert-Buck, M.R. & Linehan, W.M. (1996). Analysis of 99 microdissected prostate carcinomas reveals a high frequency of allelic loss on chromosome 8p12-21, *Cancer Research* **56**, 2411–2416.

[25] Xu, J., Zheng, S.L., Hawkins, G.A., Faith, D.A., Kelly, B., Isaacs, S.D., Wiley, K.E., Chang, B., Ewing, C.M., Bujnovszky, P., Carpten, J.D., Bleecker, E.R., Walsh, P.C., Trent, J.M., Meyers, D.A. & Isaacs, W.B. (2001). Linkage and association studies of prostate cancer susceptibility: evidence for linkage at 8p22-23, *American Journal of Human Genetics* **69**, 341–350.

[26] Zitzelsberger, H., Engert, D., Walch, A., Kulka, U., Aubele, M., Hofler, H., Bauchinger, M. & Werner, M. (2001). Chromosomal changes during development and progression of prostate adenocarcinomas, *British Journal of Cancer* **84**, 202–208.

HAWKINS B. DE FRANCE &
ELAINE A. OSTRANDER

Heterozygosity

Genes can exist in different allelic forms and there are several ways to quantify the degree of allelic variation in a population. One way is simply to report the frequencies of the different alleles. Other parameters are used to address specific genetic questions. Individuals with two different alleles for some gene are said to be heterozygous for that gene, whereas those with two alleles that are the same are homozygous. The continued existence of heterozygotes implies continued genetic variation, and there have been several reports of correlation between growth rate and heterozygosity (see [2]).

If a gene has alleles a_i, then the frequencies of **genotypes** $a_i a_i$ and $a_i a_j$, $j \neq i$, are written as P_{ii} and P_{ij}, and the frequency of allele a_i is written as p_i. For large random mating populations, the **Hardy–Weinberg** law states that

$$P_{ii} = p_i^2,$$
$$P_{ij} = 2p_i p_j, \quad i \neq j.$$

When it is of interest to be able to quantify H, the total frequency of heterozygotes, under the Hardy–Weinberg situation this requires only the following allele frequencies:

$$\begin{aligned} H &= \sum_i \sum_{j \neq i} P_{ij} \\ &= \sum_i \sum_{j \neq i} p_i p_j \\ &= 1 - \sum_i p_i^2. \end{aligned}$$

This last expression is often referred to as *heterozygosity*, but this is a misnomer since it provides the frequency of heterozygotes only under Hardy–Weinberg equilibrium. It is more appropriate to define "gene diversity" D by

$$D = 1 - \sum_i p_i^2.$$

For populations with an **inbreeding** coefficient of f, heterozygote frequencies are modified to

$$P_{ij} = 2p_i p_j (1 - f),$$

so that

$$H = (1 - f)D.$$

The most likely cause of a difference between H and D in human populations is population **admixture**. If a proportion α_k of the population belongs to subpopulation k, in which frequencies for alleles a_i are p_{ki}, then the frequency of $a_i a_{i'}$ heterozygotes in the whole population is

$$P_{ii'} = 2p_i p_{i'} + \sum_k \alpha_k (p_{ki} - p_i)(p_{ki'} - p_{i'}),$$

where the total allele frequencies are given by

$$p_i = \sum_k \alpha_k p_{ki}.$$

This result assumes Hardy–Weinberg frequencies within each subpopulation. There may be more or less of a particular heterozygote than expected from the Hardy–Weinberg law in the whole population, although the overall heterozygosity is diminished:

$$H = D - \sum_i \sum_k \alpha_k (p_{ki} - p_i)^2.$$

In linkage studies it is necessary to determine whether or not recombination has occurred between two loci, and this in turn puts constraints on the genotypes of individuals in successive generations. The **polymorphism information content** (PIC) characterizes the extent to which a marker gene (*see* **Genetic Markers**) is useful for linkage studies, with higher values being better. It cannot be greater than H.

Variance of Heterozygosity

If sample allele and genotype frequencies are written as $\tilde{p}_i$ and $\tilde{P}_{ij}$, the sample heterozygosity is

$$\tilde{H} = \sum_i \sum_{j \neq i} \tilde{P}_{ij}.$$

Taking expectation $\mathcal{E}$ over repeated samples of n individuals from the same population, assuming genotype counts are **multinomially distributed**, provides

$$\mathcal{E}(\tilde{H}) = H,$$

whereas the expected value of sample diversity,

$$\tilde{D} = 1 - \sum_i \tilde{p}_i^2,$$

is

$$\mathcal{E}(\tilde{D}) = \left(1 - \frac{1+f}{2n}\right) D$$

The variance over repeated samples from the same population is just the **binomial** variance for $\tilde{H}$,

$$\text{var}(\tilde{H}) = \frac{1}{n} H(1-H),$$

whereas for diversity [3],

$$\text{var}(\tilde{D}) = \frac{2(1+f)}{n}\left[\sum_i p_i^3 - \left(\sum_i p_i^2\right)^2\right].$$

If heterozygosity is averaged over loci, then the variance of the average depends on two-locus heterozygosities. If $H_{ll'}$ is the probability of an individual being heterozygous at loci l and l', then the sample single-locus heterozygosities are correlated:

$$\text{cov}(\tilde{H}_l, \tilde{H}_{l'}) = \frac{1}{n}(H_{ll'} - H_l H_{l'}),$$

so that the variance within populations of heterozygosity averaged over m loci is

$$\text{var}(\tilde{H}) = \frac{1}{nm^2}\sum_l H_l(1-H_l) + \frac{1}{nm^2}\sum_l \sum_{l' \neq l} (H_{ll'} - H_l H_{l'}).$$

Brown et al. [1] pointed out that the two-locus heterozygosity depends on **linkage disequilibrium** between the loci, and the variance of average single-locus heterozygosity therefore serves as a summary statistic for linkage disequilibrium. The same holds for average gene diversity.

References

[1] Brown, A.H.D., Feldman, M.W. & Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum spontaneum*, *Genetics* **96**, 523–530.

[2] Hartl, D.L. & Clark, A.G. (1989). *Principles of Population Genetics*, 2nd Ed. Sinauer, Sunderland.

[3] Weir, B.S., Reynolds, J. & Dodds, K.G. (1990). The variance of sample heterozygosity, *Theoretical Population Biology* **37**, 235–253.

B.S. WEIR

Hidden Markov Models

Hidden Markov models (HMMs) are a class of stochastic models that have proven to be useful in a wide range of applications for modeling highly structured sequences of data. Some applications of HMMs include machine speech recognition [15], ion channel kinetics [9, 10], and biomolecular sequence analysis [1, 4–6, 8] (*see* **DNA Sequences**).

A hidden Markov model can be viewed as a black box that generates sequences of observations. The unobservable internal state of the box is stochastic and is determined by a finite state **Markov chain**. The observable outputs of the black box are stochastic, with distribution determined by the current state of the hidden Markov chain. In more detail, let $(s_t, t = 0, 1, 2, \ldots)$ be an unobserved Markov chain on the state space $(1, 2, \ldots, L)$ and let $(y_t, t = 0, 1, 2, \ldots)$ be an observed process that takes values in the set $(1, 2, \ldots, K)$. The restriction to discrete observations is not essential but it is adequate for the applications considered here.

There are three inference problems that arise in the development or application of hidden Markov models: **estimation** of model parameters, restoration of the hidden states, and model selection (*see* **Model, Choice of**). In this article we define an HMM as a stochastic model that generates sequences of observations, provide examples of HMMs that are used in applications, and discuss approaches to the first two inference problems.

Model Specification

An HMM with L hidden states and K observable outputs is specified by three sets of distributions. First is the *initial distribution* of the hidden Markov chain

$$\Pr(s_0 = i), \quad i \in \{1, \ldots, L\}. \tag{1}$$

Second is the *transition distribution* of the hidden Markov chain as represented by the $L \times L$ matrix $\mathbf{\Lambda} = [\lambda_{ij}]$ with elements

$$\lambda_{ij} = \Pr(s_{t+1} = j | s_t = i), \quad i \in \{1, \ldots, L\}, \quad j \in \{1, \ldots, L\}. \tag{2}$$

Third is the set of *output distributions* of the hidden states as represented by the $L \times K$ matrix $\mathbf{\Pi} = [\pi_{ij}]$ with elements

$$\pi_{ij} = \Pr(y_t = j | s_t = i), \quad i \in \{1, \ldots, L\}, j \in \{1, \ldots, K\}. \tag{3}$$

Both of the matrices $\mathbf{\Lambda}$ and $\mathbf{\Pi}$ are stochastic, i.e. they are formed by nonnegative numbers and their row sums are equal to one. Thus the parameter $\boldsymbol{\theta} \equiv (\mathbf{\Lambda}, \mathbf{\Pi})$ takes values in a compact set Θ which is a direct product of L L-dimensional and L K-dimensional simplexes.

Models with continuous output distributions can be developed by replacing the probability mass function in (3) with an appropriate density function, e.g. **normal**. With some minor modifications, the results below can be applied to continuous data.

The number of hidden states and their connectivity, i.e. the set of nonzero λ_{ij}, define the *architecture* of an HMM. The choice of an architecture is typically driven by an application for which the HMM is intended. In some cases the architecture is an attempt to model a physical system (e.g. ion channels) and in other cases the HMM is merely a convenient fiction that is useful for **classification** or **prediction** (e.g. speech recognition). The states of the hidden Markov chain may be recurrent or transient. It is worthwhile to consider two classes of architectures. First is the *recurrent* architecture in which any hidden state may be reached from any other hidden state. Second is the *left-to-right* architecture, in which the hidden states are transient. Of course, arbitrarily complex HMMs can be constructed with both recurrent and nonrecurrent components.

It is often convenient to consider transient chains and to introduce two states *begin* (B) and *end* (E) that do not produce any output. Without loss of generality we assume that the initial distribution is concentrated in the state B. Thus $\Pr(s_0 = \mathrm{B}) = 1$. The state transition matrix $\mathbf{\Lambda}$, whose dimension becomes $(L + 2) \times (L + 2)$, is modified as follows:

1. The state B is unattainable from any state including itself; $\lambda_{i\mathrm{B}} = 0$, for all i.
2. State E is absorbing, so that $\lambda_{\mathrm{EE}} = 1$ and is recurrent, so there is a stopping time $n^* = \min(k : s_k = \mathrm{E}, k \geq 0)$ such that $\Pr(n^* \leq \infty) = 1$.
3. The direct transition from state B to state E is not allowed; $\lambda_{\mathrm{BE}} = 0$.

Introduction of the absorbing state E allows us to deal with finite realizations of the HMM up to the

stopping time n^*. We put $n = n^* - 1$ and use the following notation for the sequence of hidden states and the corresponding sequence of outputs

$$\mathbf{s} \equiv s_1 s_2 \ldots s_n,$$

$$\mathbf{y} \equiv y_1 y_2 \ldots y_n.$$

The states $s_0 = \mathrm{B}$ and $s_{n+1} = \mathrm{E}$ will be suppressed in the notation, except where they are explicitly needed below.

Suppose we observe N independent realizations of an HMM and denote the set of observed outputs by

$$\mathbf{Y} \equiv \left\{ \begin{array}{l} \mathbf{y}_1 = y_{1,1} y_{1,2} \ldots, y_{1,n_1} \\ \vdots \quad \vdots \\ \mathbf{y}_N = y_{N,1} y_{N,2} \ldots, y_{N,n_N} \end{array} \right\}.$$

The sequences of paths through the hidden Markov chain that produced $\mathbf{Y}$ will be denoted by

$$\mathbf{S} \equiv \left\{ \begin{array}{l} \mathbf{s}_1 = s_{1,1} s_{1,2} \ldots, s_{1,n_1} \\ \vdots \quad \vdots \\ \mathbf{s}_N = s_{N,1} s_{N,2} \ldots, s_{N,n_N} \end{array} \right\}.$$

In this formulation there is a one-to-one correspondence between the states of the hidden Markov chain and the elements of the observed sequence. The model can be generalized to include null states (other than B and E). Null states may be visited by the hidden Markov chain but do not produce any observable output.

Hidden Markov models can have large parameter spaces because there may be many possible state transitions and because each state can have its own unique output distribution. Depending on the application, it may be desirable to allow all nonzero parameter values to vary freely. At the other extreme, we may require that some subsets of parameters take identical values. Constraints of this type are referred to as "tied" parameterizations. A less extreme form of combining information can be achieved by imposing a hierarchical model on the parameters in which sets of parameter values are assumed to be drawn from a common distribution [7, 19].

There are known **identifiability** problems with the recurrent HMM model due to the labeling of the states, and some convention for state labeling is needed. There can also be identifiability problems if the output distributions in different states are not distinct. These issues are discussed by Leroux [15]. We note that, for the two-state model, identifiability problems also arise when $\lambda + \mu = 1$. This result suggests that further investigation into the identifiability of HMMs may be worthwhile.

Examples of HMMs

Finite-State Recurrent Architecture

Consider a hidden Markov chain with two main states denoted by 0 and 1. The two-state recurrent architecture is illustrated in Figure 1. Its transition probability matrix, defined on the extended state space (B, 0, 1, E), is

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & \lambda_{\mathrm{B}0} & \lambda_{\mathrm{B}1} & 0 \\ 0 & \lambda_{00} & \lambda_{01} & \lambda_{0\mathrm{E}} \\ 0 & \lambda_{10} & \lambda_{11} & \lambda_{1\mathrm{E}} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

For the case of binary (0, 1) data sequences, the output distribution is given by

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}.$$

This HMM generates nonhomogeneous binary sequences that consist of homogeneous regions of two types, with distinct frequencies of zeros and ones. This model and the more general L-state, K-output recurrent model were applied by Churchill [4, 5] to identify regions with distinct functions in DNA sequences based on differences in local base frequencies.

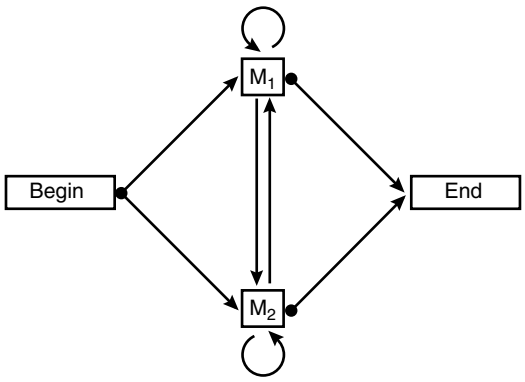

Figure 1 Two-state recurrent HMM architecture

Left-to-Right Architectures

An example of a left-to-right architecture is shown in Figure 2. Left-to-right models are used in applications to speech recognition as "word" models. Each state has an output distribution that characterizes part of the acoustic signal that defines a word. The state transitions follow the evolution of the word from left to right and allow for compression or expansion of the duration of the signal over time. Collections of word models can be nested inside a larger HMM for purposes of word classification and recognition. Further details and references on HMM applications to speech recognition can be found in [13].

Biomolecular Models

Another example of a left-to-right architecture is the mutation–deletion–insertion (MDI) model shown in Figure 3. The MDI model has become a very popular tool for the problem of aligning multiple DNA or protein sequences [1, 14]. In this model, there are three different types of states. The backbone of the model consists of *mutation* states ($M_1, M_2, \ldots, M_L$). Each mutation state M_i has a corresponding *deletion* state D_i. Following the state B there is an *insertion* state I_0, and following each of the mutation states M_i there is an insertion state I_i. There are two sets of output distributions in the MDI model. Outputs from M-states are generated according to $\Pr(j|M_i)$. These distributions will typically vary from state to state and reflect the position-specific frequencies of nucleotide or amino acid subunits as they occur along the length of a molecule. Outputs from I-states are generated according to $\Pr(j|I_i)$. These states allow for site-specific insertion of letters into the sequence. The D-states are silent and do not produce any output. These states allow specific positions (modeled by M-states) to be skipped in the generation of an output sequence. The length of an output sequence will typically be close to the number of M-states in the model, but any realization may be shorter or longer due to insertion and deletion events.

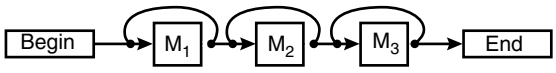

Figure 2 A simple left-to-right HMM architecture

The presence of silent states in the MDI model introduces a minor complication into our description of these HMMs. It was implicit in our earlier definition of an HMM that there is a one-to-one correspondence between outputs and hidden states. However, in the MDI model, as it is typically implemented, there may be hidden states (D-states) that are visited but have no corresponding output. Thus the length of **y** may be less than the length of the corresponding hidden state sequence **s**. We note that the output of an MDI model can be viewed as the

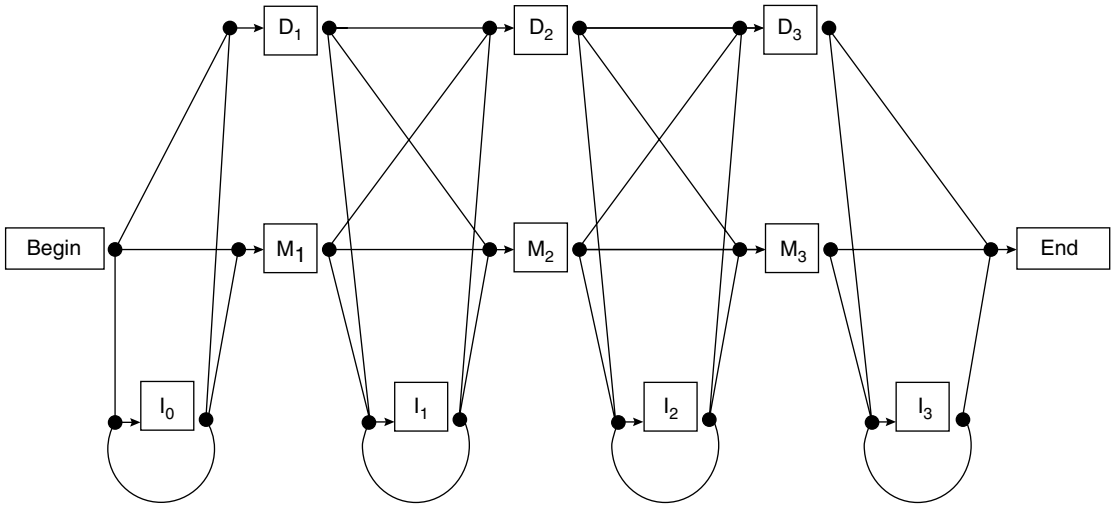

Figure 3 Mutation–deletion–insertion (MDI) architecture with three M-states

output of a standard HMM consisting of only M-states and I-states. This MI chain is embedded within the MDI chain and can be constructed by simply removing the D-states. The architecture of the MI chain includes additional transitions to replace the removed D-states. Unfortunately the additional transition parameters must be constrained in a rather complicated fashion to recover exactly the original MDI model. The output distributions of the MI model are identical to those of the MDI model. It follows that results derived for standard HMMs apply equally to MDI models.

A hidden Markov model has been developed for the problem of DNA sequence assembly [6, 7]. We consider a collection of DNA sequences (see Table 1) that are independently copied from a common *prototype* sequence, $\mathbf{r} = r_1, \ldots, r_L; r_i \in$ (A,C,G,T), by a process that introduces errors in the form of *substitutions, deletions*, and *insertions*. Each realization, $i = 1, \ldots, N$, of the MDI chain will generate a sequence $\mathbf{y}_i$ with elements $y_{ij} \in$ (A,C,G,T,N). The output character N is sometimes generated by DNA sequencing devices to represent ambiguous determination of a base. Each M-state in the MDI chain is associated with an element of the prototype sequence, i.e. M_i is associated with r_i. This association will determine the output distribution of the M-state. For example, if the state M_i is associated with $r_i = \mathrm{A}$, the most likely output of state M_i is the letter A. A substitution error occurs when the output is a letter other than A. A deletion error occurs when the state D_i is visited, thus bypassing M_i, and no letter is generated as output. An insertion error occurs when the state I_i is visited, thus generating extraneous letters in the output sequence. To summarize, a visit of the D_i-state results in a deletion of r_i in the copying process; k successive visits of the I_i-state result in an insertion of k letters after the ith position in the prototype; a visit of the M_i-state results in copying r_i, with possible substitution error.

The output from N realizations of an MDI chain will be a set of sequences of letters. The sequences will generally be similar to one another but may vary in length as well in the identity of specific letters. Often the goal of applying an MDI model to a sequence is to restore the hidden state sequence. Restoration of $\mathbf{s}_i$ establishes a correspondence between the elements of $\mathbf{y}_i$ and the states of the MDI model. Furthermore, the multiple path restoration of $\mathbf{S}$ establishes a correspondence among all elements of all the DNA sequences via their correspondence with the M-states. This correspondence is a *multiple sequence alignment* [21]. An example of an HMM-generated sequence alignment of the DNA sequences from Table 1 is shown in Figure 4.

Inference for HMMs

Likelihood

In this section, we describe an **algorithm** to compute the **likelihood** of an observed sequence $\mathbf{y}$, i.e. $\Pr(\mathbf{y}|\boldsymbol{\theta})$. In the case of multiple independent observations, the likelihood is simply the product $\Pr(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \Pr(\mathbf{y}_i|\boldsymbol{\theta})$.

We can express the likelihood as a summation over all possible hidden state sequences

$$\Pr(\mathbf{y}|\boldsymbol{\theta}) = \sum_{s} \Pr(\mathbf{y}|\mathbf{s},\boldsymbol{\Pi}) \Pr(\mathbf{s}|\boldsymbol{\Lambda}), \tag{4}$$

where

$$\Pr(\mathbf{y}|\mathbf{s},\boldsymbol{\Pi}) = \pi_{s_1,y_1} \cdot \pi_{s_2,y_2} \ldots \pi_{s_n,y_n} \tag{5}$$

and

$$\Pr(\mathbf{s}|\boldsymbol{\Lambda}) = \lambda_{\mathrm{B}s_1} \cdot \lambda_{s_1 s_2} \ldots \lambda_{s_n \mathrm{E}}. \tag{6}$$

However, this summation is generally intractable, and an alternative approach is needed to compute the likelihood.

The likelihood can also be written in the form

$$\begin{aligned} \Pr(\mathbf{y}|\boldsymbol{\theta}) &= \prod_{t=1}^{n} \Pr(y_t|\mathbf{y}^{t-1}) \\ &= \prod_{t=1}^{n} \sum_{s_t=1}^{L} \Pr(y_t|s_t) \Pr(s_t|\mathbf{y}^{t-1}), \end{aligned} \tag{7}$$

where $\mathbf{y}^{t-1} = y_1, \ldots, y_{t-1}$. We assume that the distribution of y_t depends only on s_t. The first term in (7) is the output distribution and the second term is the predictive density. The predictive density can be computed using the *forward pass algorithm*. This algorithm is the basis for a number of other computations and is presented here.

The Forward Pass Algorithm. To begin, suppose that $\Pr(s_{t-1}|\mathbf{y}^{t-1})$ is known. A prediction of the state

Table 1 An unaligned set of DNA sequences

TAGACAGGNGCCCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
TAGACAGGGNCCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
TAGANAGGGCCTCCACTGGGGAAATGAAGGTACCNACCAACCTTCAAAACTT
TAGACCAGGNGCTCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
TAGACAGGGCCTCCACTGGAGATNTGAGGTCACCAACCAACCTTCAAAAACTT
TAGACAGGGGCTCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT

```
                TAGACAGGGCC-CCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
                      N
                TAGACAGGGNC-CCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
                TAGANAGGGCCTCCACTGG-GGAATGAGGT-ACCNACCAACCTTC-AAAACTT
                                  A A
                TAGACAGGNGCTCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
                    C
                TAGACAGGGCCTCCACTGGAG-ATTGAGGTCACCAACCAACCTTCAAAAACTT
                                   N
                TAGACAGGGGCTCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
Consensus       TAGACAGGGNCTCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
```

Figure 4 A multiple sequence alignment of the DNA sequences in Table 1. The alignment was generated using an MDI architecture with 52 M-states. Letters aligned in each column correspond to the same M-states. Insertions are shown above the main sequence and deletions are shown as dashes. An estimated consensus or prototype sequence is shown below the multiple alignment

at time t can be computed, using the law of total probability, as

$$\Pr(s_t|\mathbf{y}^{t-1}) = \sum_{s_{t-1}=1}^{L} \Pr(s_t|s_{t-1}, \mathbf{y}^{t-1}) \times \Pr(s_{t-1}|\mathbf{y}^{t-1}). \quad (8)$$

This conditional distribution is called the *predictive density*. Next, the information in the current observation is incorporated by updating the predictive density. The so-called *filtered density* is

$$\Pr(s_t|\mathbf{y}^t) = \frac{\Pr(y_t|s_t, \mathbf{y}^{t-1}) \Pr(s_t|\mathbf{y}^{t-1})}{\Pr(y_t|\mathbf{y}^{t-1})}, \quad (9)$$

by **Bayes' theorem**, where

$$\Pr(y_t|\mathbf{y}^{t-1}) = \sum_{s_t=1}^{L} \Pr(y_t|s_t, \mathbf{y}^{t-1}) \Pr(s_t|\mathbf{y}^{t-1}).$$

Restoration of the Hidden Markov Chain

The problem of restoring the hidden state sequence $\mathbf{s}$ from a given observation sequence $\mathbf{y}$ is addressed here. We assume that the model parameters $\boldsymbol{\theta}$ are given and suppress $\boldsymbol{\theta}$ in the notation of this section.

There are two general approaches to the state restoration problem. A local restoration uses the marginal conditional densities to find the most probable state at each point t. This marginal restoration is given by

$$s_{i,t}^* = \text{argmax}_{s_t} \Pr(s_t|\mathbf{y}_i).$$

A global restoration maximizes the full conditional density to find a most probable restoration. Thus,

$$\mathbf{s}_t^* = \text{argmax}_s \Pr(\mathbf{s}|\mathbf{y}_i).$$

These two approaches can result in quite different solutions. The first problem is solved using the *backward algorithm* and the second is solved using the *Viterbi algorithm* [20].

The Backward Algorithm. The backward algorithm uses the results of the forward algorithm to compute the conditional distribution of the hidden state sequence given the complete sequence of

observed data. It is sufficient to specify the distribution of pairs of adjacent states because we assume the Markov property.

The joint distribution of two adjacent states is computed recursively, starting with the last step of the forward algorithm

$$\Pr(s_t, s_{t+1}|\mathbf{y}) = \frac{\Pr(s_{t+1}|\mathbf{y}) \Pr(s_{t+1}|s_t) \Pr(s_t|\mathbf{y}^t)}{\Pr(s_{t+1}|\mathbf{y}^t)}$$

The marginal distribution can be obtained by summing the expression over s_{t+1}. Thus

$$\Pr(s_t|\mathbf{y}) = \Pr(s_t|\mathbf{y}^t) \sum_{s_{t+1}=1}^{L} \frac{\Pr(s_{t+1}|\mathbf{y}) \Pr(s_{t+1}|s_t)}{\Pr(s_{t+1}|\mathbf{y}^t)}.$$

See Rabiner [17] or Churchill [4] for more detail.

The Viterbi Algorithm. Our goal is to find the sequence of states $\mathbf{s}$ that maximizes over the space of all state sequences the conditional probability of $\mathbf{s}$ given the observation sequence $\mathbf{y}$ and known model parameters. Notice that

$$\Pr(\mathbf{s}|\mathbf{y}) = \frac{\Pr(\mathbf{s}, \mathbf{y})}{\Pr(\mathbf{y})}, \tag{10}$$

and thus it is sufficient to find the sequence of states which maximizes $\Pr(\mathbf{s}, \mathbf{y})$.

The joint probability can be factored as

$$\begin{aligned}\Pr(\mathbf{s}, \mathbf{y}) &= \prod_{t=1}^{n} \Pr(s_t, y_t|\mathbf{s}^{t-1}, \mathbf{y}^{t-1}) \\ &= \prod_{t=1}^{n} \Pr(y_t|\mathbf{s}^t, \mathbf{y}^{t-1}) \Pr(s_t|\mathbf{s}^{t-1}, \mathbf{y}^{t-1}) \\ &= \prod_{t=1}^{n} \Pr(y_t|s_t) \Pr(s_t|s_{t-1}). \end{aligned} \tag{11}$$

Let $\delta_t(i)$ denote the maximum probability up to time t over all state sequences which end at the state $s_t = i$,

$$\delta_t(i) = \max_{s^{t-1}} \Pr(\mathbf{s}^{t-1}, s_t = i, \mathbf{y}^t). \tag{12}$$

The joint probability, (11), can be maximized by the following procedure:

1. initialization,

$$\delta_1(i) = \Pr(s_1 = i) \Pr(y_1|s_1 = i), 1 \le i \le r; \tag{13}$$

2. recursion,

$$\delta_t(j) = \max_{1 \le i \le r} [\delta_{t-1}(i)\lambda_{ij}] \Pr(y_t|s_t = j); \tag{14}$$

3. termination,

$$\max_{\mathbf{s}} \Pr(\mathbf{s}, \mathbf{y}) = \max_{1 \le i \le r} \delta_n(i). \tag{15}$$

We are asking at each time point t, if the present state is $s_t = j$, which state at time $t - 1$ maximizes the joint probability over all past state sequences. Roughly, if we are in state j at time t, where did we come from at time $t - 1$?

For each state $1 \le j \le r$ at time t we wish to keep track of the state at time $t - 1$ which gives us the maximum. To do this, we define the quantity

$$\psi_t(j) = \arg \max_{1 \le i \le r} [\delta_{t-1}(i)\lambda_{ij}]. \tag{16}$$

When the process is terminated, we have computed $\delta_n(i)$ and $\psi_n(i)$ for $1 \le i \le r$. A state sequence which attains the maximal probability can be constructed by a traceback. Let

$$s_n^* = \arg \max_{1 \le i \le r} [\delta_n(i)] \tag{17}$$

be the best final state. The traceback is completed by the recursion

$$s_t^* = \psi_{t+1}(s_{t+1}^*), \ t = n - 1, \ n - 2, \ldots, 1. \tag{18}$$

Note that $\delta_t(j) \to 0$ fast. Thus computations are more easily executed on a log scale. The recursion, (13), will look like

$$\begin{aligned}\log \delta_t(j) = &\max_{1 \le i \le r} [\log \delta_{t-1}(i) + \log \lambda_{ij}] \\ &+ \log \Pr(y_t|s_t = j). \end{aligned} \tag{19}$$

Parameter Estimation

In the **maximum likelihood** approach to HMM restoration, no prior information on the parameter θ is assumed and the inference problems of parameter estimation and state restoration are addressed by first finding an estimator for $\boldsymbol{\theta}$ and then restoring $\mathbf{S}$ conditionally given the estimated value.

In general, the likelihood is intractable for direct maximization. However, given a state sequence $\mathbf{S}$, the augmented data likelihood,

$$\Pr(\mathbf{Y}, \mathbf{S}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \Pr(\mathbf{y}_i, \mathbf{s}_i|\boldsymbol{\theta}) = \prod_{i=1}^{N} \Pr(\mathbf{y}_i|\mathbf{s}_i, \boldsymbol{\Pi}) \Pr(\mathbf{s}_i|\boldsymbol{\Lambda}), \quad (20)$$

is quite well behaved. The problem of maximizing the augmented data likelihood is trivial. The augmented data **sufficient statistics** for this problem are matrices $\mathbf{C}^{\boldsymbol{\Lambda}} \equiv [c_{ij}^{\boldsymbol{\Lambda}}]$ and $\mathbf{C}^{\boldsymbol{\Pi}} \equiv [c_{ij}^{\boldsymbol{\Pi}}]$, where $c_{ij}^{\boldsymbol{\Lambda}}$ is the number of transitions to the j-state from the i-state and $c_{ij}^{\boldsymbol{\Pi}}$ is the number of outputs j from state i. When some parameter values are tied, the dimensions of the sufficient statistics can be reduced.

The problem of maximizing the observed data likelihood is solved by the *Baum–Welch* algorithm [2, 3, 17]. An initial estimate $\boldsymbol{\theta}^0$ is chosen. The algorithm iterates the following steps:

1. Use the forward and backward algorithms to compute the conditional expectation of $\mathbf{C}^{\boldsymbol{\Lambda}}$ and $\mathbf{C}^{\boldsymbol{\Pi}}$ with respect to $\Pr(\mathbf{S}|\mathbf{Y}, \boldsymbol{\theta}^m)$.
2. Use the expected sufficient statistics from step 1 to obtain a new estimate $\boldsymbol{\theta}^{m+1}$. The MLEs for $\boldsymbol{\Pi}$ and $\boldsymbol{\Lambda}$ are simply the row-normalized expectations of $\mathbf{C}^{\boldsymbol{\Lambda}}$ and $\mathbf{C}^{\boldsymbol{\Pi}}$, respectively.

As $m \to \infty$ the sequence of parameter estimates $\boldsymbol{\theta}^m$ will converge to a point of maximum, not necessarily global, of the likelihood function [3, Eq. (4)]. HMM likelihoods can be multimodal, and thus it is recommended to try several starting values for the estimation algorithm.

From a computational point of view the first step can be carried out in time proportional to $N \times L^2 \times$ (average sequence length) and the estimator in step 2 can be obtained in closed form. However, the implementation can be rather tedious and in many applications a *modified* version, called the segmental k-means algorithm [12], is used. Step 1 is replaced by:

1. Obtain a most probable path $\mathbf{s}_i^{m+1} = \operatorname{argmax}_{\mathbf{s}} \Pr(\mathbf{s}|\mathbf{y}_i, \boldsymbol{\theta}^m)$ for each $i = 1, \ldots, N$.
2. Obtain a new estimate $\boldsymbol{\theta}^{m+1}$ that maximizes the augmented data likelihood $\Pr(\mathbf{Y}, \mathbf{S}^{m+1}|\boldsymbol{\theta})$.

The first step can be accomplished by dynamic programming [20], and a closed-form estimator is available for step 2. This algorithm produces an estimator $\tilde{\boldsymbol{\theta}}$ that maximizes the objective function $\max_{\mathbf{s}} \Pr(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta})$. Although this is not a maximum likelihood estimator, the two estimators will generally be very similar [16].

Having obtained some parameter estimate $\boldsymbol{\theta}^*$, we can restore $\mathbf{S}$ using either of the methods in the section, "Restoration of the Hidden Markov Chain".

A Bayesian Approach

A weakness of the likelihood approach to the state restoration problem is that the final solution is based on the point estimator of $\boldsymbol{\theta}$ and fails to take into account other "reasonable" values of $\boldsymbol{\theta}$. Furthermore, it may be of interest to find not only an *optimal* multiple path but also to have access to reasonable alternative restorations. These concerns motivate a **Bayesian** approach to the state restoration problem.

We assume a **prior distribution** $P_0(\boldsymbol{\theta})$ for the parameter $\boldsymbol{\theta} \equiv (\boldsymbol{\Lambda}, \boldsymbol{\Pi})$ so that the posterior distribution of the pair $(\mathbf{S}, \boldsymbol{\theta})$ is

$$\Pr(\mathbf{S}, \boldsymbol{\theta}|\mathbf{Y}) \propto P_0(\boldsymbol{\theta}) \prod_{i=1}^{N} \Pr(\mathbf{y}_i|\mathbf{s}_i, \boldsymbol{\Pi}) \Pr(\mathbf{s}_i|\boldsymbol{\Lambda}), \quad (21)$$

where the last two terms are defined in (5) and (6), respectively. Integrating out the parameter $\boldsymbol{\theta}$ in (21) we obtain the marginal posterior $\Pr(\mathbf{S}|\mathbf{Y})$, which will be our primary interest. Similarly, summing over all multiple paths, we obtain the marginal posterior of $\Pr(\boldsymbol{\theta}|\mathbf{Y})$. These marginal posterior distributions are not practically computable, in part because of unassessable normalizing constants.

Two slightly different solutions have been proposed for this problem. Both approaches use a Gibbs sampler (e.g. [11]) to approximate the desired posterior distributions (*see* **Markov Chain Monte Carlo**). The Gibbs sampler alternately generates random samples from the full conditional distributions $\Pr(\mathbf{s}|\mathbf{y}, \boldsymbol{\theta})$ and $\Pr(\boldsymbol{\theta}|\mathbf{y}, \mathbf{s})$. In Robert et al. [18] the hidden state sequence is sampled elementwise. In Churchill & Lazareva [7], the full sequence is sampled in one step at each iteration of the Gibbs sampler. Sampling of $\boldsymbol{\theta}$ is trivial when conjugate Dirichlet or Dirichlet mixture priors are used (*see* **Loglinear Model**).

Summary

Hidden Markov models provide a powerful class of models that can be applied to analyze data that consist of sequences of dependent observations with an underlying heterogeneous structure. The inference problems of parameter estimation and state restoration can be addressed using algorithms described in this article, but many subtle and difficult issues may arise in any particular application. The widespread use and success of hidden Markov models in speech recognition and molecular sequence analysis suggest that there may be many fruitful areas of application to which these methods can be applied.

References

[1] Baldi, P., Chauvin, Y., Hunkapillar, T. & McClure, M.A. (1994). Hidden Markov models of biological primary sequence information, *Proceedings of the National Academy of Sciences* **91**, 1059–1063.

[2] Baum, L.E. & Petrie, T. (1966). Statistical inference for probabilistic or finite state Markov chains, *Annals of Mathematical Statistics* **37**, 1554–1563.

[3] Baum, L.E., Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics* **41**, 164–171.

[4] Churchill, G.A. (1989). A stochastic model for heterogeneous DNA sequences, *Bulletin of Mathematical Biology* **51**, 79–94.

[5] Churchill, G.A. (1992). Hidden Markov chains and the analysis of genome structure, *Computers and Chemistry* **16**, 107–115.

[6] Churchill G.A. (1995). Accurate restoration of DNA sequences, in *Case Studies in Bayesian Statistics*, Vol. II, C. Gatsaris, J.S. Hodges, R.E. Kass & N.D. Singpurwalla, eds. Springer-Verlag, New York, pp. 90–148.

[7] Churchill, G.A. & Lazareva, B. (1998). Bayesian restoration of a hidden Markov chain with applications to DNA sequencing, unpublished manuscript.

[8] Eddy, S.R. (1996). Hidden Markov models, *Current Opinion in Structural Biology* **6**, 361–365.

[9] Fredkin, D.R. & Rice, J.A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings, *Proceedings of the Royal Society of London, Series B* **249**, 125–132.

[10] Fredkin, D.R. & Rice, J. (1992). Bayesian restoration of single channel patch clamp recordings, *Biometrics* **48**, 427–448.

[11] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[12] Juang, B.H. & Rabiner, L.R. (1990). The segmental k-means algorithm for estimating parameters of hidden Markov models, *IEEE Transactions on Acoustics, Speech and Signal Processing* **38**, 1639–1641.

[13] Juang, B.H. & Rabiner, L.R. (1991). Hidden Markov models for speech recognition, *Technometrics* **33**, 251–272.

[14] Krogh, A., Brown, M., Mian, I.S., Sjölander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling, *Journal of Molecular Biology* **235**, 1501–1531.

[15] Leroux, B.G. (1992). Maximum-likelihood estimation for hidden Markov models, *Stochastic Processes and their Applications* **40**, 127–143.

[16] Merkav, N. & Ephraim, Y. (1991). Maximum likelihood hidden Markov modeling using a dominant sequence of states, *IEEE Transactions on Signal Processing* **39**, 2111–2115.

[17] Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**, 257–286.

[18] Robert, C.P., Celeux, G. & Diebolt J. (1994). Bayesian estimation of hidden Markov chains: a stochastic implementation, *Statistics and Probability Letters* **16**, 77–83.

[19] Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Main, I.S. & Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology, *CABIOS* **12**, 327–345.

[20] Viterbi, J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Transactions on Information Theory* **13**, 260–269.

[21] Waterman, M.S. (1995). *Introduction to Computational Biology*. Chapman & Hall, London.

GARY A. CHURCHILL

Hierarchical Models in Genetics

A problem with the evaluation of large amounts of genomic information is the issue of **multiple comparisons**. This problem arises from performing numerous analyses on the same data without taking into consideration the increased likelihood of falsely detecting an association. Hierarchical modeling – incorporating higher-level "prior" models into a conventional analysis – offers a solution to problems of multiple inferences. Furthermore, this approach can give estimates that are more plausible and stable than conventional estimates by "borrowing information" from the similarities in one's data.

Previous work has shown that parameter estimates from hierarchical models can be more plausible and stable than estimates from conventional models [5, 9, 10, 13]. This potential improvement results from modeling similarities among parameters of interest by using a second-stage model. In addition to providing parameter estimates of effect that are more accurate and more plausible than those from conventional models, it allows one to incorporate multiple levels of information on genetic and environmental factors into a single analysis and provides a solution to problems of multiple comparisons [12]. For example, Thomas et al. [9] presented a hierarchical modeling approach for evaluating **candidate genes** and environmental factors, and applied it to simultaneously investigate the **associations** between multiple human leukocyte antigen (HLA) alleles and insulin-dependent diabetes mellitus (IDDM). Furthermore, hierarchical modeling can simultaneously address issues of population stratification [6] (*see* **Bias in Case–Control Studies**).

Hierarchical Modeling

First-stage Model

Assume that one collects data on multiple correlated exposures of interest $\mathbf{x}$ (i.e. **genotype**), and phenotype y. Further assume that one wants to use these data to estimate coefficients β for the effects of genotype on phenotype. One can estimate β from the following **generalized linear model** for the expectation of y conditional on $\mathbf{x}$:

$$g_1[\mathrm{E}(y|\mathbf{x})] = \alpha + \mathbf{x}\beta, \qquad (1)$$

where g_1 is a monotonic differentiable strictly increasing link function between the random and systematic components, and y has mean $\mathrm{E}(y|\mathbf{x})$ and **variance** σ^2. Conventional analytic approaches to estimating β using (1) include: (a) fitting a (full) model that contains all the exposures; (b) reducing a full model with a preliminary testing algorithm (e.g. stepwise); (c) constructing numerous one-at-a-time models (i.e. evaluating the multiple parameter inference problem as multiple one-parameter inference problems); and (d) estimating haplotypes from the genotype information and the effects of haplotypes on phenotype (*see* **Haplotype Analysis**). Unfortunately, none of these approaches provides entirely satisfactory estimates of β. Approach (a) is impracticable if a full model will not fit one's data; moreover, even when the model does fit, this approach can give biased and inefficient estimates. Approach (b) excludes statistically "nonsignificant" exposures from the full model regardless of their biologic importance, and produces biased point and variance estimates [5, 7, 8]. Approach (c) takes no account of correlations among the exposures. Approach (d) requires estimation of haplotypes, and assumes that effects are homogeneous across haplotypes. Finally, none of the approaches properly addresses issues of multiple comparisons [7, 10].

Second-stage Model

Instead of undertaking a conventional single-stage analysis, one can use a hierarchical model to estimate β. This approach provides a coherent framework for multiple inference problems, using shrinkage estimation to improve estimation accuracy [5, 13]. In particular, this approach uses higher level "priors" to model the parameters of interest (here β) as random variables whose joint distribution is a function of hyperparameters. Assume that in addition to the above data (i.e. $\mathbf{x}$ and y), one has information about similarities between the components of β (e.g. physical distance between genetic **markers**). One can use such additional information in a second-stage generalized linear model for the expectation of β conditional on this information:

$$g_2[\mathrm{E}(\beta|\mathbf{Z})] = \mathbf{Z}\pi, \qquad (2)$$

where g_2 is a strictly increasing link function, β has mean $\mathrm{E}(\beta|\mathbf{Z})$ and variance τ^2, and $\mathbf{Z}$ is a second-stage

design matrix expressing the similarities between the β. For example, one could assume a linear link function and define a second-stage design matrix $\mathbf{Z}$, where z_{ij} indicates a function of the physical distance between each **polymorphism** (i.e. element z_{ij} of $\mathbf{Z}$ is equal to $e^{-d_{ij}t}$, where t is a scale parameter estimated from the data, and d_{ij} is the distance between polymorphism i and polymorphism j). Thus, the coefficients for each polymorphic marker β_i would be related through a second-stage covariate (e.g. physical distance) that is thought to be relevant to the strength of the marker-specific effects (which are measured by β_i).

Hierarchical Estimates

Hierarchical (i.e. posterior) estimates are then obtained by combining results – essentially taking weighted averages – from the different level models. Weights used in combining these results reflect how well each stage was able to estimate that level's parameters. Specifically, more unstable estimates will be given smaller weight, and vice versa. Hence, if sufficient data exist to estimate adequately first-stage parameters, then adding a second-stage will have limited effect on these estimates. To fit the levels in a hierarchical model separately one can use iterative weighted least squares. Assume that one has conventional **maximum likelihood** coefficient estimates $\hat{\beta}$ from fitting a first-stage model. One can then compute hierarchical estimates $\tilde{\beta}_{\mathbf{W}}$ of the coefficients β by averaging $\hat{\beta}$ with the fitted $E(\beta)$ from the second-stage regression of $\hat{\beta}$ on $\mathbf{Z}$. In particular, one can estimate the second-stage regression coefficients π using weighted-least-squares:

$$\tilde{\pi} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}\mathbf{W}\hat{\beta}, \tag{3}$$

where $\mathbf{W} = [\hat{\mathbf{V}} + \mathbf{diag}(\tau^2)]^{-1}$, and $\hat{\mathbf{V}}$ is inverse information for β evaluated at $\hat{\beta}$. The fitted value for β from the second-stage regression is therefore $\mathbf{Z}\tilde{\pi}$. Averaging $\mathbf{Z}\tilde{\pi}$ with the maximum likelihood estimates $\hat{\beta}$ gives the hierarchical estimate

$$\tilde{\beta}_{\mathbf{W}} = \mathbf{B}\mathbf{Z}\tilde{\pi} + (\mathbf{I} - \mathbf{B})\hat{\beta}, \tag{4}$$

with estimated covariance matrix

$$\tilde{\mathbf{C}} = \hat{\mathbf{V}}[\mathbf{I} - (\mathbf{I} - \mathbf{H})'\mathbf{B}], \tag{5}$$

where $\mathbf{B} = \mathbf{W}\hat{\mathbf{V}}$, $\mathbf{I}$ = the identity matrix, and $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}$. Equation (4) shows how two-stage hierarchical modeling compromises between first- and second-stage estimates: the distance of the hierarchical estimate from either stage estimates is indirectly proportional to its stability. More specifically, the larger the elements in $\hat{\mathbf{V}}$ are, the farther $\tilde{\beta}_{\mathbf{W}}$ will be from $\hat{\beta}$. Conversely, the larger the τ^2 are, the farther $\tilde{\beta}_{\mathbf{W}}$ will be from $\mathbf{Z}\tilde{\pi}$.

Using the weighted-least-squares approach requires that enough data exist to fit a first-stage model. When a full model will not fit one's data, a penalized **likelihood** hierarchical approach can be used instead. This approach entails thinking of the second-stage design matrix $\mathbf{Z}$ as a rational basis for choosing a penalty function for penalized likelihood. Specifically, the penalty function based on (2),

$$\mathbf{P} = \tilde{\beta}'_{\mathbf{P}}(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\tilde{\beta}_{\mathbf{P}}, \tag{6}$$

could be used to compute hierarchical estimates of the coefficients $\tilde{\beta}_{\mathbf{P}}$ by penalized likelihood. This corresponds to a weighted sum-of-squared-residual penalty for departures of $\tilde{\beta}_{\mathbf{P}}$ from the linear model $\mathbf{Z}\pi$. The maximum penalized likelihood estimates are obtained by maximizing $\mathbf{L} - \mathbf{P}/2\tau^2$, where $\mathbf{L}$ is the conventional log likelihood derived from (1). Hence, τ^2 is the inverse of the usual smoothing parameter in penalized likelihood [11].

When distributions are not conjugate, one can use Gibbs sampling to fit hierarchical models (*see* **Markov Chain Monte Carlo**). Gibbs sampling is a Monte Carlo method for estimating the joint and marginal posterior distributions of a set of random variables when direct calculation of these distributions is infeasible. It requires only specification of the set of conditional distributions of each random variable, given all other variables, and possibly data [3]. If the full conditional distribution does not have a simple form, one can still use Gibbs sampling by applying a derivative-free adaptive rejection sampling procedure to generate samples from the exact posterior distribution [4].

Application of Hierarchical Models

As an example of hierarchical modeling, Thomas et al. [9] presented an empirical-Bayes approach (*see* **Markov Chain Monte Carlo**) for testing associations with large numbers of candidate genes in the

presence of environmental risk factors. They investigated this approach by application to **HLA** associations in IDDM, and by a **simulation** study designed to reflect situations they have observed in family studies of IDDM. Their hierarchical approach assumed that the log **relative risks** for all alleles at a given locus are exchangeable (presuming that there is no preferential zygotic assortment and negligible recombination; *see* **Linkage Analysis, Model-based**). Furthermore, they considered modeling the covariance between two haplotypes as a function of the number of alleles they share, and the marginal strength of the effects of these alleles. Simulation results indicated that hierarchical empirical-Bayes was superior to maximum likelihood. In particular, when there were no haplotype effects, empirical-Bayes estimates were closer to the true value than maximum likelihood estimates for 75% of the alleles, and the empirical-Bayes estimates were more stable as well. When there were haplotype effects, empirical-Bayes was also superior because maximum likelihood models were often unable to fit without first using the data to select a subset of variables.

Aragaki et al. [1] used a hierarchical model to estimate NAT2 genotype-specific dietary effects on adenomatous polyps. The first stage used **logistic regression** to model the joint effects of **genotype** and diet, as well as their main effects and other **covariates**. However, using this conventional approach to estimate dietary effects within nine genotypes – with the 910 case–control subjects in the study – gave unstable results. Therefore, to improve precision they modeled the joint effect of genotype and diet as a function of initial rate of carcinogenic conversion of dietary heterocyclic amines to aryl nitrenium ions [1]. In comparison with the conventional results, the hierarchical model gave more precise and reasonable estimates.

Another potentially valuable application of hierarchical modeling arises in combining linkage results across different studies (*see* **Meta-analysis in Human Genetics**). In this situation, the fact that different studies give results for the same marker locus is exploited in an attempt to improve estimates of recombination or lod scores. Specifically, one can use hierarchical modeling to combine likelihoods [2], or corresponding lod scores and recombination fractions. For example, when estimating **confidence** (or support) **intervals** for a particular recombination fraction, one can apply a hierarchical model that "borrows strength" from all recombination fractions estimated at the marker of interest. The general concept is similar to that presented above, except that here we are interested in estimation of likelihoods instead of regression coefficients.

The potential improvement available with hierarchical approaches does require reasonable higher-level models; but, as long as this requirement is met, hierarchical modeling will generally give better estimates than (one-stage) conventional maximum likelihood. In contrast, if a higher-level model cannot provide an even remotely reasonable approximation to reality, this approach may produce invalid confidence intervals, and parameter estimates may be less accurate than those obtained with typical analyses.

References

[1] Aragaki, C.C., Greenland, S., Probst-Hensch, N. & Haile, R.W. (1997). Hierarchical modeling of gene-environment interactions: estimating NAT2* genotype-specific dietary effects on adenomatous polyps, *Cancer Epidemiology, Biomarkers and Prevention* **6**, 307–314.

[2] Efron, B. (1996). Empirical Bayes methods for combining likelihoods, *Journal of the American Statistical Association* **91**, 538–565.

[3] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[4] Gilks, W.R., Roberts, G.O. & George, E.I. (1994). Adaptive direction sampling, *Statistician* **431**, 179–189.

[5] Greenland, S. (1993). Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression, *Statistics in Medicine* **12**, 717–736.

[6] Kim, L.-L., Fijal, B.A. & Witte, J.S. (2001). Hierarchical modeling of the relation between sequence variants and a quantitative trait: addressing multiple comparison and population stratification issues, *Genetic Epidemiology*, to appear.

[7] Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion), *Journal of the American Statistical Association* **78**, 47–65.

[8] Sclove, S.L., Morris, C. & Radhakrishna, R. (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution, *Annals of Mathematical Statistics* **43**, 1481–1490.

[9] Thomas, D.C., Langholz, B., Clayton, D., Pitkaniemi, J., Tuomilehto-Wolf, E. & Tuomilehto, J. (1992). Empirical Bayes methods for testing associations with large numbers of candidate genes in the presence of environmental risk factors, with applications to HLA association in IDDM, *Annals of Medicine* **24**, 387–392.

[10] Thomas, D.C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M. & Armstrong, B.G. (1985). The problem of multiple inference in studies designed to generate hypotheses, *American Journal of Epidemiology* **122**, 1080–1095.

[11] Titterington, D.M. (1985). Common structure of smoothing techniques, *International Statistics Review* **53**, 141–170.

[12] Witte, J.S. (1997). Genetic analysis with hierarchical modeling, *Genetic Epidemiology* **14**, 1137–1142.

[13] Witte, J.S. & Greenland, S. (1996). A simulation study of hierarchical regression, *Statistics in Medicine* **15**, 1161–1170.

(*see also* **Maximum Likelihood**)

JOHN S. WITTE

Hierarchical Models in Health Service Research

Hierarchical models provide a natural framework for conceptualizing and quantifying systematic and random components of variation in **multilevel** data. For example, hierarchical nested structures are present in data describing **health care utilization**, cost (*see* **Health Economics**), and **outcomes** for patients treated in specific hospitals, which may in turn belong to particular health care systems or may be clustered by geographic or **hospital market areas** (*see* **Small Area Variation Analysis**).

The analysis of variations in health care processes and outcomes seeks to quantify and characterize variability across clusters, such as physicians, hospitals, and geographic or market areas, at each level of the hierarchical data structure. In particular, the analysis seeks to determine whether comparable patients receive similar treatment and experience similar outcomes across clusters. If differences exist, the analysis turns to the examination of how patient, hospital, or regional characteristics may be related to these differences. In addition, the analysis examines the link between measures of outcome, such as patient mortality, morbidity, and functioning and indicators of process, such as descriptors of regional or provider practice patterns. When the focus of the analysis is on comparative measures of performance of health care providers, the term *profiling* analysis is often used (*see* **Profiling Providers of Medical Care**).

A number of methodologic issues confront the investigator in the analysis of variations in health care. First, sample size can vary across clusters, resulting in substantially different precision of cluster-specific estimates. For example, the number of patients with a particular condition in each hospital may vary from a handful to several hundred in a typical analysis of hospital variations. Secondly, the analytic strategy needs to take into account the **correlation** of the responses within each cluster (*see* **Clustering**). Failure to do so may result in understating the error associated with the estimates of effects of case-specific **covariates**, such as the effect of patient characteristics on medical procedure utilization across different geographic areas. Thirdly, the analyst needs to derive reliable cluster-specific estimates, such as mortality rates for each hospital, and also to estimate the effects of cluster-level covariates, such as hospital characteristics. The usual approach of fitting a single **regression** model to the entire data set does not account for correlations and cannot accommodate both cluster-specific indicator variables and cluster-level covariates.

Hierarchical regression modeling goes a long way toward meeting these methodologic challenges. The approach enables the analyst to separate sampling variability from variability across clusters. It also allows the latter to be further partitioned into a *systematic* component (*see* **Fixed Effects**), which is linked to cluster characteristics, and a *random* component (*see* **Random Effects**). The hierarchical model accommodates within-cluster correlations and makes it possible to estimate case- and cluster-level covariate effects and **variance components** simultaneously. The model also makes it possible to pool information across clusters in order to derive more precise estimates of cluster-specific parameters and cluster-level effects.

Examples

Although hierarchical models have been extensively discussed in the statistical literature (see [4], [7], [8], and [14], and references in **Multilevel Models**), their use in **health services research** is relatively recent [1–3, 5, 6, 9, 10, 12, 13]. In a particular study, the complexity of the hierarchical model will be commensurate with the research question and the level of detail in the data. The following examples illustrate two typical scenaria.

Aggregate Responses: Hierarchical Poisson Model

Consider studies in which a **Poisson** count, Y_i, of events is observed in the ith of K clusters. For example, Y_i can be the number of patients who experience complications after undergoing a specific operation in the ith hospital during a particular year. The number of patients receiving the operation in the ith hospital is denoted by n_i. If there is no reason to suspect systematic differences across hospitals, the following hierarchical model with an **exchangeable** second-level structure can be considered:
Level I (within-hospital). $Y_i|\theta_i \sim$ Poisson $(\theta_i n_i)$.
Level II (between-hospitals). $\log(\theta_i) \sim \mathrm{N}(\mu, \sigma)$.

Fully **Bayesian** formulations of the hierarchical model include a third level, in which **priors** on the population parameters μ and σ are specified. Vague (but proper) priors are often used at this stage.

A model of this type was employed in a recently published analysis of teenage conception rates for different health boards in Scotland [5]. The model was used to derive estimates and posterior intervals of the individual health board rates (θ_i) and of the relative ranking for each health board. Similar models have been used with **binomial** outcomes in Level I and with cluster-level covariates in Level II of the hierarchy.

Case-level Responses: Hierarchical Logistic Model

The models for binomial and Poisson responses are applicable to studies in which only aggregate data are available on each cluster. If case-specific data are available, more intricate hierarchical regression models can be employed. In many studies a **binary** response, Y_{ij}, is observed on the jth case in cluster i, where $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. The available data include a K-dimensional vector of covariate values X_{ij} on the ijth case and an L-dimensional vector of covariate values Z_i on cluster i. For example, in a study of geographic patterns of utilization of coronary angiography in elderly patients who had a heart attack, the binary response of interest was the indicator of whether angiography was performed on a patient within a specific time interval after the infarction. Data on patient sociodemographics (such as age, gender, and race) and **co-morbid** conditions were represented by the vector X_{ij}, and characteristics of the geographic area (such as location, and availability of angiography in local hospitals) were represented by Z_i [2]. As a second example, in a profiling analysis of the performance of hospitals which treat heart-attack patients, the binary response of interest was the indicator of whether a patient survived past the initial 30-day period after the heart attack. Data on patient sociodemographic characteristics and severity at hospital entry were represented by the vector X_{ij} and selected hospital characteristics by the vector Z_i [12]. As a third example, in an analysis of data from the National Health Interview Survey, the binary response was an indicator of whether an individual had a physician visit in the past year. The analysis included data on characteristics of the individual and the county of the individual's residence [9].

The following hierarchical **logistic regression** model was used in the analysis of the angiography data:

Level I (within-area variability). A logistic model was assumed within each area. Specifically, if $p_{ij} = \Pr(Y_{ij} = 1)$, then

$$\text{logit}(p_{ij}) = \beta_{0i} + \beta_{1i} X_{1ij} + \beta_{2i} X_{2ij} + \cdots + \beta_{Ki} X_{Kij}.$$

Level II (between-areas variability). The variation across areas was partitioned into a systematic and a random component. The systematic component was expressed by a **multiple linear regression** model linking the within-area logistic coefficients to area-level covariates. Specifically,

$$\beta_{ki} = \gamma_{k1} Z_{1i} + \cdots + \gamma_{kL} Z_{Li} + \varepsilon_{ki}.$$

The error terms ε_{ki} were assumed to have a **multivariate normal distribution** with mean zero and covariance structure such that (i) the error terms for different units are independent, and (ii) the within-unit $(K + 1) \times (K + 1)$ **covariance matrix** D is the same for all units. Heavier-tailed distributions, such as **multivariate *t*** may also be used for modeling the variability of the β_i [1]. In a third and final level, vague proper priors can be assumed on the components of γ and on the covariance matrix D.

The above hierarchical logistic model makes it possible to combine data across geographical areas in order to derive smoothed estimates of the effects of patient characteristics, such as sociodemographics and co-morbidity, both over the entire country and within each area. Therefore, we can determine whether a specific patient characteristic (such as race) has a differential impact on practice patterns in different areas of the country. The process of combining information across areas takes into account differences in sample size and results in improved precision of estimates for areas with small sample sizes. The hierarchical model estimates of the logistic coefficients can be conceptually described (and numerically approximated) as a weighted combination of (i) the coefficients resulting from fitting the logistic model solely to the data of the particular area and (ii) average values of these coefficients across areas, as determined by the area-level characteristics. In the angiography example, the hierarchical model estimates are effectively obtained by shrinking the

coefficients from the fully stratified analysis towards a regression line determined by the area characteristics included in the vector **Z**. The degree of **shrinkage** is different for each covariate, being influenced by (i) the accuracy with which the particular covariate can be estimated via the stratified analysis (*see* **Stratification**) and (ii) by the degree to which the estimate for a particular area differs from the estimates for the other areas. Shrinkage is generally going to be higher for coefficients of areas with smaller overall sample sizes and/or small cell counts for a particular covariate.

The hierarchical model also makes it possible to derive area-specific estimates of the probability of the outcome (in this case, performance of angiography) for each stratum of patients that can be defined using the covariate vector **X**. These probabilities can be presented graphically using maps (*see* **Mapping Disease Patterns**). In addition, the model makes it possible to examine the relation between area-level covariates (**Z**) and area-specific event rates or area-specific effects of patient characteristics. In the angiography analysis, for example, it was determined that an area's rate of angiography for an average patient was positively related to an index measuring the availability of the procedure to patients in that area. It was also found that the effect of race differed across census regions in the country.

There are two simpler *fixed-effects* alternatives to the hierarchical logistic model of the analysis of variations across areas: (i) regression analysis stratified by area, and (ii) regression analysis using a single logistic model for the entire country, with indicator variables for each area and their interactions. The fully stratified analysis is close to the spirit of Level I of the hierarchical model. However, such an analysis may not be an efficient approach and may lead to highly imprecise estimates of effects, especially in areas with small sample sizes overall or in some categories of patients. The analysis via a single logistic regression model is more common in practice. However, the **standard errors** of the coefficients from this analysis do not account for the effects of clustering of patients within areas and will, therefore, need to be adjusted. This correlation is accounted for by the hierarchical analysis. In addition, the analysis via a single regression model for the entire country cannot incorporate both patient-level and area-level covariates without leading to model indeterminacy. For example, if **dummy variables** for areas are included then it is no longer possible to include variables indicating the location of the area and other area characteristics. Such area characteristics can be accommodated via the hierarchical model or via a two-stage approach in which a fully stratified analysis is first carried out and the resulting coefficients are used as the dependent variable in regression models similar to those in Level II of the hierarchical model. The two-stage analysis will generally lead to consistent estimates of the second-stage coefficients but is likely to understate the standard error of the estimates without careful adjustment.

Further Applications

Hierarchical regression modeling techniques are by now available for most response data of interest in health services and outcomes research. In particular, the response may be binary or a count as above; **polytomous**, e.g. utilization of one of several alternative treatments [1]; **ordered categorical**, e.g. appropriateness of care; or continuous. The latter may be observed completely, e.g. cost of care; or above a threshold, e.g. vulnerability to malpractice claim [3]. For each type of response the models can include cluster-level covariates, such as hospital size and teaching status. Aggregate data on patient mix can also be included in Level II of the model, but that would provide only a rudimentary method for **case-mix** adjustment. More substantial case-mix adjustment can be implemented with models such as the hierarchical logistic model above. The approach requires the use of patient-level information and can be accomplished with hierarchical models in which Level I describes the relation of the response on an individual patient to patient characteristics. However, it should be noted that the use of hierarchical modeling does not necessarily address the effects of **selection bias**, especially if such bias is related to covariates that are not represented in the database.

Further levels can be added to the above hierarchical models in order to accommodate additional structure in the data. In the Poisson example, longitudinally observed counts may be available on each cluster over several years (*see* **Longitudinal Data Analysis, Overview**). In the logistic example, primary clusters such as hospitals may be further grouped by geographic region or market area. In each case, the incorporation of further levels and corresponding covariates is straightforward. In

addition to cluster-level covariates, the hierarchical structure may also be used to model spatial dependence. Such models have already been developed and used in epidemiologic studies and can be readily adapted for health services research data.

Model Fitting and Checking

Simulating observations from the posterior distribution of the parameters is generally recognized to be the most flexible and broadly applicable approach to fitting hierarchical regression models. This fully Bayesian framework provides a more realistic account of uncertainty in the estimates without the need for rather complex adjustments. A key practical advantage of the approach is that it makes it possible to **simulate** values and derive estimates of any function of the parameters, with little additional computational burden. For example, in profiling analyses, it is generally straightforward to simulate values from the posterior distribution of any measure of hospital performance, to derive estimates and to account for the uncertainty in these estimates. The most common algorithms for generating simulated values involve **Markov chain Monte Carlo** (MCMC) methodology [4]. Although special programs may have to be developed for some of the more complex, multilevel models, a large class of problems can be analyzed using the publicly available software BUGS [11]. A number of recent authors have proposed **diagnostics** for checking the convergence of MCMC runs [4]. Some of these diagnostics are now available in BUGS and other MCMC software. A recent account of approaches to checking model fit (*see* **Model Checking**) and comparing alternative models can be found in [4].

An alternative computational approach to posterior simulation has been developed using weighted **least squares** methods and can be implemented via the software package MLn (*see* **Multilevel Models**). Some classes of hierarchical regression models can also be fitted using special SAS subroutines (*see* **Software, Biostatistical**) for mixed models as well as a plethora of more specialized software.

References

[1] Daniels, M. & Gatsonis, C. (1997). Hierarchical polytomous regression models with applications to health services research, *Statistics in Medicine* **16**, 2311–2325.

[2] Gatsonis, C.A., Epstein A.M., Newhouse, J.P., Normand, S.L. & McNeil, B.J. (1995). Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction: An analysis using hierarchical logistic regression, *Medical Care* **33**, 625–642.

[3] Gibbons, R.D., Hedeker, D., Charles, S.C. & Frisch P. (1994). A random-effects probit model for predicting medical malpractice claims, *Journal of the American Statistical Association* **89**, 760–767.

[4] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[5] Goldstein, H. & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical Society, Series A* **159**, 385–444.

[6] Jenks, S.F., Daley, J., Draper, D., Thomas, N., Lehnart, G. & Walker, J. (1988). Interpreting hospital mortality data: the role of clinical risk adjustment, *Journal of the American Medical Association* **260**, 3611–3616.

[7] Kass, R.E. & Steffey, D. (1989). Approximate Bayesian inference for conditionally independent hierarchical models, *Journal of the American Statistical Association* **84**, 717–726.

[8] Lindley, D. & Smith, A. (1972). Bayes estimates for the linear model, *Journal of the Royal Statistical Society, Series B* **34**, 1–41.

[9] Malec, D., Sedransk, J. & Tompkins, L. (1993). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey, in *Case Studies in Bayesian Statistics*, C. Gatsonis, J. Hodges, R. Kass & N. Singpurwall, eds. Springer-Verlag, New York, pp. 377–389.

[10] McNeil, B.J., Pederson, S. & Gatsonis, C. (1992). Current issues in profiling quality of care, *Inquiry* **29**, 298–307.

[11] MRC Biostatistics Unit (1996). *BUGS Manual*. Institute of Public Health, Cambridge.

[12] Normand, S.L., Glickman, M. & Gatsonis, C. (1997). Statistical methods for profiling providers: issues and applications. *Journal of the American Statistical Association* **92**, 803–814.

[13] Shwartz, M., Ash, A., Anderson, J., Iezzon, L., Payne, S. & Restuccia, J. (1994). Small area variation in hospitalization rates: how much you see depends on how you look, *Medical Care* **32**, 189–201.

[14] Wong, G. & Mason, W.M. (1991). Contextually specific effects and other generalizations of the hierarchical linear model for comparative analysis, *Journal of the American Statistical Association* **86**, 487–503.

CONSTANTINE GATSONIS

Hierarchical Models

This term is currently used in a variety of contexts. The most traditional one is in the sense that two statistical models are said to be hierarchical if one is a submodel of the other. A set of models, $H_1, H_2, \ldots, H_k$, is similarly called a hierarchy if

$$H_1 \subset H_2 \subset \cdots \subset H_k.$$

Hierarchical models specified by a finite set of parameters are of particular importance, because the comparison of models can be based on standard likelihood ratio tests. When models are not hierarchical, or nested, then special procedures are required (*see* **Separate Families of Hypotheses**).

The term "hierarchical" can also be used to refer to a single model, usually in the context of **regression** or **analysis of variance**. In this usage, a model is said to be hierarchical if the presence of an **interaction** term implies the inclusion of all lower-order interactions and main effects for the explanatory variables involved in the interaction. The model is hierarchical in the sense that it includes all submodels in a hierarchy as special cases. It has been argued that only such models should be considered [1], but this is not universally accepted. Significance tests for the presence of interactions are, however, best considered in the context of hierarchical models.

Another usage of the term "hierarchical models" is as a synonym for **multilevel models**. This usage derives from the hierarchical nature of data in which observations are nested within higher level classifications. For example, individuals may be nested within families, or patients may be nested within clinics. **Bayesian** hierarchical models provide another use of the term. **Markov chain Monte Carlo** methods are very important in this context.

Reference

[1] Nelder, J.A. (1977). A reformulation of linear models (with discussion), *Journal of the Royal Statistical Society, Series A* **140**, 48–77.

VERN T. FAREWELL

Hill, Austin Bradford

Born: July 8, 1897, in Hampstead, London, UK.
Died: April 18, 1991, in Cumbria, UK.

Reproduced by permission of the Royal Statistical Society

Austin Bradford Hill was Professor of Medical Statistics and Director of the Medical Research Council's Statistical Research Unit at the London School of Hygiene and Tropical Medicine, 1946–61; he introduced the principle of **randomization** into the conduct of controlled trials (*see* **Clinical Trials, Overview**) in clinical medicine and established particularly clearly the role of smoking in the production of lung cancer and, subsequently, many other diseases (*see* **Smoking and Health**). As a result of the experience gained in interpreting the observed association between smoking and lung cancer, Hill drew up "guidelines" to help reach a positive conclusion about causality that have come to be used widely in epidemiology and, on occasions, in the law (*see* **Hill's Criteria for Causality**).

Career

Hill, who was always known as Bradford Hill in scientific circles and as Tony to his friends, had wanted to study medicine, but he was diverted from doing so by the outbreak of World War I. He enlisted, at the first opportunity, in 1916 and opted for a commission in the Royal Naval Air Service. After being posted to the Greek islands in support of the attack on the Dardanelles, he developed pulmonary tuberculosis and, in November 1917, was invalided out of the service and sent home. Instead of causing his death, which was the anticipated outcome, the development of pulmonary tuberculosis probably saved his life, for the expectation of life of fighter pilots in World War I was measured in weeks. The downhill progress of the disease was, however, arrested after he was given an artificial pneumothorax to rest his lung, and by 1919 he was sufficiently recovered to think again about his future. Medicine was out of the question and he decided to study economics as an external student of London University. With the aid of a correspondence course and by reading in bed, he succeeded in obtaining a second class honours degree in 1922, having attended the university itself only to take examinations.

Hill had no desire to make a career of economics and he managed to enter medicine with the help of **Major Greenwood**, a friend of his father and one of the few medical statisticians of the day. He obtained a grant from the **Medical Research Council** to investigate the reasons for the high mortality of young adults in rural areas and, whilst holding it, attended part of **Karl Pearson's** course on statistics for the London B.Sc. at University College. From then on he worked consistently with Greenwood in a variety of capacities in the conduct of epidemiologic research and, later, in the teaching of medical statistics at the London School of Hygiene and Tropical Medicine, where Greenwood had been appointed to the professorship of Medical Statistics. On the outbreak of World War II he was seconded to the Research and Experimental Department of the Ministry of Home Security and subsequently to the Medical Directorate of the Royal Air Force. In 1946 Greenwood retired and Hill was appointed to succeed him, both as professor at the School and as director of the MRC's unit.

Teaching Medical Statistics

Hill described himself as an arithmetician rather than a statistician, and it was the clarity of his exposition of simple arithmetic and statistical procedures and of the logic that justified conclusions from epidemiologic studies, combined with his sensitivity to the ethical concerns of practicing clinicians, that enabled him to

influence British academic medicine as greatly as he did. From his first appointment at the London School of Hygiene in 1933 he found himself responsible for **teaching** the elements of statistics to medical postgraduates who, as a group, had little liking or aptitude for mathematics in any form. At that time, the need for some sort of statistical analysis had been recognized in the field of public health and had begun to be appreciated in laboratory medicine, but it was hardly understood in clinical medicine at all. Hill responded, not by pressing the need for deferring to a statistical consultant, but by urging research workers in all branches of medicine to learn enough about statistical techniques to appreciate their value in both the planning of experiments (*see* **Experimental Design**) and in the interpretation of figures and so to accept the statistician as a partner in their research, while the statistician, for his part, had to steep himself in the realities of medical practice. His lectures on medical statistics proved to be so effective that he was asked to publish them in a series of articles in the *Lancet* and to republish them in book form. The book, entitled *Principles of Medical Statistics*, was published in 1937 [4] and republished and expanded in a further 10 editions, some of which were translated into Spanish, Korean, Indonesian, Polish, and Russian, before a twelfth enlarged edition appeared shortly after his death, with his son, I. D. Hill, as joint author [8]. The fact that statistical analysis is now an integral part of almost every medical publication is a result of the work of many gifted statisticians throughout the world (*see* **Statistical Review for Medical Journals, Journal's Perspective**). The fact that the medical profession awoke to its need in the middle of the century was largely due to its exposition by Hill.

The Introduction of Randomization

In the first edition of his book, Hill made no reference to randomization in the planning of controlled trials. He urged only the need for concurrent controls, obtained, for example, by giving different treatment to alternate patients, a technique that had been recommended since the end of the nineteenth century, but was still the exception rather than the rule. This method was, however, far from ideal, as practice proved that a doctor's decision to enter a patient into a trial could be **biased** if he knew what treatment he or she would receive. Hill appreciated this, but he explained, shortly before his death, that he had deliberately omitted any reference to randomization in his 1937 articles

> because I was trying to persuade doctors to come into controlled trials in the very simplest form and I might have scared them off. I think the concepts of "randomization" and "random sampling numbers" are slightly odd to the layman or, for that matter, to the lay doctor, when it comes to statistics. I thought it would be better to get doctors to walk first, before I tried to get them to run [6].

By the end of World War II, the situation had changed and Hill felt able to introduce physicians to the idea, and in 1946 he persuaded two committees of the Medical Research Council to adopt the method: first, to test the value of a pertussis vaccine to prevent whooping cough [11] and second, a few months later, to test the efficacy of streptomycin in the treatment of pulmonary tuberculosis (*see* **Medical Research Council Streptomycin Trial**) [10]. The results of the latter study were, however, published first and it is usually, but undeservedly, described as the first randomized clinical trial.

The idea of randomization in biological experiments was not new. It had been introduced by **R. A. Fisher** 20 years before as a basic principle of experimental design in agriculture; but it was unheard of in clinical medicine and was anathema on first presentation to many clinicians who thought it conflicted with their responsibility for doing the best they could for individual patients and resulted in beneficial effects being diluted by giving the new treatment to patients who were unsuitable for it. Neither objection was, of course, valid, as entry to the trial was in the clinician's own hands and required him or her not to know which was the better treatment and to exclude patients if they were thought to be unsuitable for either of the therapies under trial (*see* **Ethics of Randomized Trials**; **Medical Ethics and Statistics**). Gradually clinical opposition was overcome, largely, in the UK, as a result of Hill's emphasis on ethical considerations, which won the respect of practising clinicians, and within 10 years randomization had become the standard technique for the conduct of controlled clinical trials. Recent claims that randomization had been introduced earlier by others as, for example, in the trial of patients for the treatment of the common cold, do not bear close investigation [9].

Smoking and Lung Cancer

Of Hill's many epidemiologic studies the most outstanding are those that demonstrated the importance of cigarette smoking as a cause of lung cancer. In one, comparisons were made between the smoking habits of patients with lung cancer admitted to 20 London hospitals and the habits of other patients of the same sex and age admitted to the same hospitals with other diseases (*see* **Case–Control Study, Hospital-based**). The results showed sharp differences between the two groups and led to the conclusion that cigarette smoking was an important cause of the disease [1]. This was not the first study to have shown that patients with lung cancer tended to have smoked more than other patients, but it was the first in which a firm conclusion about causality had been reached on logical grounds and it set out clearly the basis for it. The conclusion was not, however, widely accepted, and Hill set out to test it by means of a prospective study, in which information was obtained about the smoking habits of 40 000 British doctors, who were then followed to determine the mortality rates in different groups of men and women who smoked different amounts. Within a few years, results were obtained that were almost identical to those predicted from the case–control study [2, 3] and the validity of the earlier conclusion quickly came to be accepted. Neither the case–control study nor the prospective, or **cohort study** as it has come to be called, was the first of their type to have been carried out; but they set standards of design and analysis by which subsequent similar studies have come to be assessed.

Guidelines for Determining Causality

In reaching the conclusion that the association observed between smoking habits and the development of lung cancer reflected cause and effect, Hill had first to exclude chance, bias, and **confounding** as alternative explanations. The first two were not difficult to exclude, but the third was, and positive evidence had to be sought that would justify the choice of causality. Koch's postulates that had been valuable in determining the microbiological causes of infectious disease were not appropriate for other types of disease that could have multiple causes, and Hill suggested a set of guidelines to replace them, based on his experience in interpreting the results of his studies of lung cancer [5]. Only one feature had to be present (the temporal relationship of the suspected cause and its effect), none alone was conclusive, and Hill emphasized that the guidelines were no more than a help to constructive thought and that each case had to be considered on its merits. For this purpose, they have proved to have lasting value to both scientists and lawyers.

Hill, the Man

Hill was not a prolific writer of scientific papers. Apart from his lecture series on medical statistics and the many editions of his textbook, his bibliography lists only 140 publications, including 28 letters to journals and 13 reviews or historical notes [7]. His influence on British medicine was, however, disproportionately great; not only because of the importance of some of the papers, but because of his teaching, the advice he gave personally to the many individuals who sought it, and his contribution to the work of the Medical Research Council through membership of many committees and, in 1954, membership of the Council itself. In committee, he expressed his opinion cogently and firmly, but he never imposed it and he was, in consequence, always listened to with respect and his advice was almost always taken. In public he avoided controversy and, though distressed by Sir Ronald Fisher's attacks on his interpretation of the association between smoking and the development of lung cancer, he preferred to let the facts speak for themselves rather than embark on a public dispute. He took immense trouble over his lectures, which he gave without the use of visual aids and rehearsed so often and read so well that his audience often thought that he spoke without a text. Even those whose interest flagged were kept attentive by the occasional witty aside. As a department head, he kept his door open to any junior who sought his advice and he saw his job as providing the conditions under which his university and research staff could be most productive. No one who worked in his department ever wanted to leave, and it was only with the greatest difficulty that they could be persuaded to take up more senior positions elsewhere.

References

[1] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung. Preliminary report, *British Medical Journal* **2**, 739–748.

[2] Doll, R. & Hill, A.B. (1954). The mortality of doctors in relation to their smoking habits. A preliminary report, *British Medical Journal* **1**, 1451–1455.

[3] Doll, R. & Hill, A.B. (1956). Lung cancer and other causes of death in relation to smoking. A second report on the mortality of British doctors, *British Medical Journal* **2**, 1071–1076.

[4] Hill, A.B. (1937). *Principles of Medical Statistics.* The Lancet, London.

[5] Hill, A.B. (1965). The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[6] Hill, A.B. (1990). Memories of the British Streptomycin Trial in Tuberculosis, *Controlled Clinical Trials* **11**, 77–79.

[7] Hill, A.B. (1993). Bibliography: publications (in English) of Sir Austin Bradford Hill, *Statistics in Medicine* **12**, 797–806.

[8] Hill, A.B. & Hill, I.D. (1991). *Bradford Hill's Principles of Medical Statistics.* Edward Arnold, London.

[9] Medical Research Council Patulin Trials Committee (1944). Clinical trial of patulin in the common cold, *Lancet* **ii**, 373–374.

[10] Medical Research Council Streptomycin in Tuberculosis Trials Committee (1948). Streptomycin treatment for pulmonary tuberculosis, *British Medical Journal* **2**, 769–782.

[11] Medical Research Council Whooping-Cough Immunization Committee (1951). The prevention of whooping-cough by vaccination, *British Medical Journal* **1**, 1463–1471.

(*See also* **Bradford Hill Lectures**)

R. DOLL

Hill's Criteria for Causality

Despite philosophic criticisms of inductive **inference**, inductively oriented causal criteria have commonly been used to make such inferences. If a set of necessary and sufficient causal criteria could be used to distinguish causal from noncausal **associations** in **observational studies**, the job of the scientist would be eased considerably. With such criteria, all the concerns about the logic or lack thereof in causal inference could be forgotten: it would only be necessary to consult the checklist of criteria to see if a relation were causal. We know from philosophy that a set of sufficient criteria does not exist [3, 6]. Nevertheless, lists of causal criteria have become popular, possibly because they seem to provide a road map through complicated territory.

A commonly used set of criteria was proposed by **Sir Austin Bradford Hill** [1]; it was an expansion of a set of criteria offered previously in the landmark Surgeon General's report on Smoking and Health [11], which in turn were anticipated by the inductive canons of John Stuart Mill [5] and the rules of causal inference given by Hume [3]. Hill suggested that the following aspects of an association be considered in attempting to distinguish causal from noncausal associations: strength, consistency, specificity, temporality, biologic gradient, plausibility, coherence, experimental evidence, and analogy. The popular view that these criteria should be used for causal inference makes it necessary to examine them in detail:

Strength

Hill's argument is essentially that strong associations are more likely to be causal than weak associations because, if they could be explained by some other factor, the effect of that factor would have to be even stronger than the observed association and therefore would have become evident (*see* **Cornfield's Inequality**). Weak associations, on the other hand, are more easily explained by undetected **biases**. To some extent this is a reasonable argument, but, as Hill himself acknowledged, the fact that an association is weak does not rule out a causal connection. A commonly cited counterexample is the relation between cigarette smoking and cardiovascular disease.

Counterexamples of strong but noncausal associations are also not hard to find; any study with strong **confounding** illustrates the phenomenon. For example, consider the strong but noncausal relation between Down syndrome and birth rank, which is confounded by the relation between Down syndrome and maternal age. Of course, once the confounding factor is identified, the association is diminished by adjustment for the factor. These examples remind us that a strong association is neither necessary nor sufficient for causality, nor is weakness necessary nor sufficient for absence of causality. In addition to these counterexamples, we have to remember that neither **relative risk** nor any other measure of association is a biologically consistent feature of an association; as described by many authors [4, 7], it is a characteristic of a study population that depends on the relative **prevalence** of other causes. A strong association serves only to rule out hypotheses that the association is entirely due to one weak unmeasured **confounder** or other source of modest bias.

Consistency

Consistency refers to the repeated observation of an association in different populations under different circumstances. Lack of consistency, however, does not rule out a causal association, because some effects are produced by their causes only under unusual circumstances. More precisely, the effect of a causal agent cannot occur unless the complementary component causes act, or have already acted, to complete a sufficient cause. These conditions will not always be met. Thus, transfusions can cause HIV infection but they do not always do so: the virus must also be present. Tampon use can cause toxic shock syndrome, but only when other conditions are met, such as presence of certain bacteria. Consistency is apparent only after all the relevant details of a causal mechanism are understood, which is to say very seldom. Even studies of exactly the same phenomena can be expected to yield different results simply because they differ in their methods and **random errors**. Consistency serves only to rule out hypotheses that the association is attributable to some factor that varies across studies.

Specificity

The criterion of specificity requires that a cause leads to a single effect, not multiple effects. This argument has often been advanced to refute causal interpretations of exposures that appear to relate to myriad effects, especially by those seeking to exonerate smoking as a cause of lung cancer. The criterion is wholly invalid, however. Causes of a given effect cannot be expected to lack other effects on any logical grounds. In fact, everyday experience teaches us repeatedly that single events or conditions may have many effects. Smoking is an excellent example: it leads to many effects in the smoker. The existence of one effect does not detract from the possibility that another effect exists. Thus, specificity does not confer greater validity to any causal inference regarding the exposure effect. Hill's discussion of this criterion for inference is replete with reservations, and many authors regard this criterion as useless and misleading [8, 9].

Temporality

Temporality refers to the necessity that the cause precede the effect in time. This criterion is unarguable, insofar as any claimed observation of causation must involve the putative cause C preceding the putative effect D. It does *not*, however, follow that a reverse time order is evidence against the hypothesis that C can cause D. Rather, observations in which C followed D merely shows that C could not have caused D in these instances; they provide no evidence for or against the hypothesis that C can cause D in those instances in which it precedes D.

Biologic Gradient

Biologic gradient refers to the presence of a monotone (unidirectional) **dose–response** curve. We often expect such a monotonic relation to exist. For example, more smoking means more carcinogen exposure and more tissue damage, hence more carcinogenesis. Such an expectation is not always present, however. The somewhat controversial topic of alcohol consumption and mortality is an example. Death rates are higher among nondrinkers than among moderate drinkers, but ascend to the highest levels for heavy drinkers. Because modest alcohol consumption can have beneficial effects on serum lipid profiles, such a J-shaped dose–response curve is at least biologically plausible.

Conversely, associations that do show a monotonic trend in disease frequency with increasing levels of exposure are not necessarily causal; confounding can result in a monotonic relation between a noncausal risk factor and disease if the confounding factor itself demonstrates a biologic gradient in its relation with disease. The noncausal relation between birth rank and Down syndrome mentioned above shows a biologic gradient that merely reflects the progressive relation between maternal age and the occurrence of Down syndrome.

Thus the existence of a monotonic association is neither necessary nor sufficient for a causal relation. A nonmonotonic relation only conflicts with those causal hypotheses specific enough to predict a monotonic dose–response curve.

Plausibility

Plausibility refers to the biologic plausibility of the hypothesis, an important concern but one that is far from objective or absolute. Sartwell [9], emphasizing this point, cited the remarks of Cheever, in 1861, who was commenting on the etiology of typhus before its mode of transmission (via body lice) was known:

> It could be no more ridiculous for the stranger who passed the night in the steerage of an emigrant ship to ascribe the typhus, which he there contracted, to the vermin with which bodies of the sick might be infested. An adequate cause, one reasonable in itself, must correct the coincidences of simple experience.

What was to Cheever an implausible explanation turned out to be the correct explanation, since it was indeed the vermin that caused the typhus infection. Such is the problem with plausibility: it is too often not based on logic or data, but only on prior beliefs. This is not to say that biological knowledge should be discounted when evaluating a new hypothesis, but only to point out the difficulty in applying that knowledge.

The **Bayesian** approach to inference attempts to deal with this problem by requiring that one quantify, on a probability (0 to 1) scale, the certainty that one has in prior beliefs, as well as in new hypotheses. This quantification displays the dogmatism or open-mindedness of the analyst in a public fashion, with certainty values near 1 or 0 betraying a strong commitment of the analyst for or against a hypothesis. It

can also provide a means of testing those quantified beliefs against new evidence [2]. Nevertheless, the Bayesian approach cannot transform plausibility into an objective causal criterion.

Coherence

Taken from the Surgeon General's report on Smoking and Health [11], the term *coherence* implies that a cause and effect interpretation for an association does not conflict with what is known of the natural history and biology of the disease. The examples Hill gave for coherence, such as the histopathologic effect of smoking on bronchial epithelium (in reference to the association between smoking and lung cancer) or the difference in lung cancer incidence by sex, could reasonably be considered examples of plausibility as well as coherence; the distinction appears to be a fine one. Hill emphasized that the absence of coherent information, as distinguished, apparently, from the presence of conflicting information, should not be taken as evidence against an association being considered causal. On the other hand, presence of conflicting information may indeed undermine a hypothesis, but one must always remember that the conflicting information may be mistaken or misinterpreted [12].

Experimental Evidence

It is not clear what Hill meant by experimental evidence. It might have referred to evidence from laboratory experiments on animals, or to evidence from human experiments. Evidence from human experiments, however, is seldom available for most epidemiologic research questions, and animal evidence relates to different species and usually to levels of exposure very different from those that humans experience. From Hill's examples, it seems that what he had in mind for experimental evidence was the result of removal of some harmful exposure in an intervention or prevention program, rather than the results of laboratory experiments [10]. The lack of availability of such evidence would at least be a pragmatic difficulty in making this a criterion for inference. Logically, however, experimental evidence is not a criterion but a test of the causal hypothesis, a test that is simply unavailable in most epidemiologic circumstances.

Although experimental tests can be much stronger than other tests, they are not as decisive as often thought, because of difficulties in interpretation. For example, one can attempt to test the hypothesis that malaria is caused by swamp gas by draining swamps in some areas and not in others to see if the malaria rates among residents are affected by the draining. As predicted by the hypothesis, the rates will drop in the areas where the swamps are drained. As Popper emphasized, however, there are always many alternative explanations for the outcome of every experiment. In this example, one alternative, which happens to be correct, is that mosquitoes are responsible for malaria transmission.

Analogy

Whatever insight might be derived from analogy is handicapped by the inventive imagination of scientists who can find analogies everywhere. At best, analogy provides a source of more elaborate hypotheses about the associations under study; absence of such analogies only reflects lack of imagination or experience, not falsity of the hypothesis.

Conclusion

As is evident, the standards of epidemiologic evidence offered by Hill are saddled with reservations and exceptions. Hill himself was ambivalent about the utility of these "standards" (he did not use the word *criteria* in the paper). On the one hand he asked "in what circumstances can we pass from this observed *association* to a verdict of *causation?*" (original emphasis). Yet, despite speaking of verdicts on causation, he disagreed that any "hard-and-fast rules of evidence" existed by which to judge causation:

> None of my nine viewpoints [criteria] can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*.

Actually, the fourth criterion, temporality, is a *sine qua non* for causality: If the putative cause did not precede the effect, that indeed is indisputable evidence that the observed association is not causal (although this evidence does not rule out causality in other situations, for in other situations the putative cause may precede the effect). Other than this one condition, however, which may be viewed as part

of the definition of causation, there is no necessary or sufficient criterion for determining whether an observed association is causal.

Acknowledgment

This article is adapted from Chapter 2 of *Modern Epidemiology* 2nd Ed. [8], with permission from the publisher.

References

[1] Hill, A.B. (1965). The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[2] Howson, C. & Urbach, P. (1993). *Scientific Reasoning. The Bayesian Approach*, 2nd Ed. Open Court, LaSalle.

[3] Hume, D. (1978). *A Treatise of Human Nature* (originally published in 1739). Oxford University Press edition, with an Analytical Index by L. A. Selby-Bigge, published 1888. 2nd Ed. with text revised and notes by P.H. Nidditch, published 1978.

[4] MacMahon, B. & Pugh, T.F. (1967). Causes and entities of disease, in *Preventive Medicine*, D.W. Clark & B. MacMahon, eds. Little, Brown & Company, Boston.

[5] Mill, J.S. (1862). *A System of Logic, Ratiocinative and Inductive*, 5th Ed. Parker, Son and Bowin, London.

[6] Popper, K.R. (1968). *The Logic of Scientific Discovery*. Harper & Row, New York.

[7] Rothman, K.J. (1976). Causes, *American Journal of Epidemiology* **104**, 587–592.

[8] Rothman, K.J. & Greenland, S. (1997). *Modern Epidemiology*, 2nd Ed. Lippincott, Philadelphia, Chapter 8.

[9] Sartwell, P. (1960). On the methodology of investigations of etiologic factors in chronic diseases – further comments, *Journal of Chronic Diseases* **11**, 61–63.

[10] Susser, M. (1988). Falsification, verification and causal inference in epidemiology: reconsiderations in the light of Sir Karl Popper's philosophy, in *Causal Inference*, K.J. Rothman, ed. Epidemiology Resources, Inc., Boston.

[11] US Department of Health, Education and Welfare (1964). Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service, *Public Health Service Publication No. 1103*. Government Printing Office, Washington.

[12] Wald, N.A. (1985). Smoking, in *Cancer Risks and Prevention*, M.P. Vessey & M. Gray, eds. Oxford University Press, New York, Chapter 3.

(*See also* **Causation**)

KENNETH J. ROTHMAN &
SANDER GREENLAND

Historical Controls in Survival Analysis

The use of historical controls in treatment evaluation is a large and controversial topic, and a general discussion is given elsewhere (*see* **Bias from Historical Controls**). The purpose of this article is the more technical one of surveying current contributions to the centuries-old statistical tradition [4] of comparing the observed mortality of a study group with that expected under "standard" (historical) rates.

If the historical rates are derived from a specific statistical analysis, then the straightforward modern approach would usually be to formulate a general statistical model containing the current data as well as the historical information, and then simply test the hypothesis of equality of the relevant mortality rates, perhaps taking into account **covariates**. Partly because the historical information is not always available as concrete statistical estimators, but also to some extent motivated by tradition, there is considerable interest in rephrasing the question as "how would these individuals have survived had they been subject to standard (historical) conditions?" Note that the so-called Peters–Belson approach in regression analysis similarly predicts study group responses from a statistical model fitted only to a control group, and then compares observed with expected [1–3].

A separate article recalls the classical calculation of **expected number of deaths** and contrasts it with a sometimes more easily interpretable calculation called the "prospective method", which however requires knowledge of the potential **censoring** time for each individual, including those who died during the study. Another separate article surveys several recent approaches to defining an **expected survival curve**, all of which have been illustrated through asymptotic statistical results of Nielsen [5], as well as some further topics and pitfalls in using **Cox regression models** in this area.

References

[1] Cochran, W.G. (1969). The use of covariance in observational studies, *Applied Statistics* **18**, 270–275.

[2] Cochran, W.G. & Rubin, D.B. (1973). Controlling bias in observational studies: a review, *Sankhyā, Series A* **35**, 417–446.

[3] Gastwirth, J.L. & Greenhouse, S.W. (1994). Biostatistical concepts and methods in the legal setting, *Statistics in Medicine* **14**, 1641–1653.

[4] Keiding, N. (1987). The method of expected number of deaths, 1786-1886-1986, *International Statistical Review* **55**, 1–20.

[5] Nielsen, B. (1997). Expected survival in the Cox model, *Scandinavian Journal of Statistics* **24**, 275–287. Addendum Vol. 26, 159.

(*See also* **Survival Analysis, Overview**)

NIELS KEIDING

HLA System

The HLA (human leukocyte antigen) **gene** complex on the short arm of chromosome 6 has been of widespread interest to scientists and physicians for more than 25 years. The most well-characterized genes are HLA-A, B, C, DR, DQ, and DP [4, 9, 14]. These code for transmembrane glycoproteins which function as receptors. They bind degraded pieces of proteins (peptides, 8–15 amino acids long) and present them to T lymphocytes to initiate immune responses.

The HLA-A, B, and C genes were identified first and are often referred to as class I genes. They are expressed on nearly all nucleated cells of the body where they allow cytotoxic T cells to recognize and eliminate tumor cells or cells infected with viruses or other intracellular pathogens [8, 10]. HLA-DR, DQ, and DP molecules (class II molecules) are expressed only on B cells, macrophages, and antigen-presenting cells. These present peptides to T helper cells to induce inflammatory immune responses [5, 8, 10].

The HLA molecules are extremely **polymorphic**. The number of alleles at each locus currently ranges from 38 to over 100 [1, 3]. Widespread amino acid substitutions occur around the molecule's peptide binding groove, and it is thought that the maintenance of polymorphism is due to selection and the evolutionary advantage of **heterozygotes** in combatting infection [7, 11, 12].

The polymorphism and immunologic function of the HLA molecules has made them of considerable interest in transplantation and disease pathogenesis. HLA molecules can induce rejection of HLA-mismatched cells and organs; matching is currently performed for kidneys and bone marrow transplants [6]. Some HLA-DR and HLA-DQ alleles show strong associations with susceptibility to autoimmune and inflammatory diseases such as rheumatoid arthritis, type I insulin-dependent diabetes mellitus, and multiple sclerosis [2, 13]. The reasons for the disease **associations** are not clear, but presumably relate to peptide-binding.

HLA allele frequencies can vary dramatically between ethnic groups [3]. Differences in allele frequencies have been used to monitor population movements and trace ancestral derivations. Some combinations of alleles at adjacent loci are inherited together on a haplotype (*see* **Haplotype Analysis**) more frequently than expected. Such **linkage disequilibrium** is often observed between HLA-B and HLA-C and between HLA-DR and HLA-DQ, and sometimes for longer distances or across the entire gene complex.

For population studies, HLA gene frequencies and two locus linkage disequilibrium coefficients can be determined by standard methods. HLA and disease associations are evident as statistically significant differences in allele frequency between patients and controls. The controls must be matched for ethnic group and adjustment made for multiple comparisons.

Although most interest has centered on HLA-A, B, C, DR, DQ, and DP, the HLA region actually covers more than 4 million base pairs of DNA and includes numerous other genes, including those for cytokines, complement components, olfactory receptors, chaperones, and many with still undefined functions [14]. Analogous gene complexes are found in other mammals [9]. The generic term *major histocompatibility complex* or MHC is frequently used to designate these gene complexes irrespective of species.

References

[1] Bodmer, J.G., Marsh, S.G.E., Albert, E.D., Bodmer, W.F., Dupont, B., Erlich, H.A., Mach, B., Mayr, W.R., Parham, P., Sasazuki, T., Schreuder, G.M.Th., Strominger, J.L., Svejgaard, A. & Terasaki, P.I. (1994). Nomenclature for factors of the HLA system, *Tissue Antigens* **44**, 1–18.

[2] Campbell, R.D. & Milner, C.M. (1993). MHC genes in autoimmunity, *Current Opinion in Immunology* **5**, 887–893.

[3] Charon, D. ed. (1997). Genetic diversity of HLA: functional and medical implications, in *Proceedings of the Twelfth International Histocompatibility Workshop and Conference*, to appear.

[4] Corzo, D., Salazar, M., Granja, C.B., & Yunis, E.J. (1995). Advances in HLA genetics, *Experimental and Clinical Immunogenetics* **12**, 156–170.

[5] Cresswell, P. (1995). Assembly, transport and function of MHC class II molecules, *Annual Review of Immunology* **12**, 259–293.

[6] Field, H. & Garavoy, M.R. (1994). Positive impact of DNA typing on solid organ transplantation, *Transplantation Review* **8**, 151–173.

[7] Hill, A.V. (1996). Genetic susceptibility to malaria and other infectious diseases: from the MHC to the whole genome, *Parasitology* **112**, Supplement, S75–S84.

[8] Germain, R.N. (1994). MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation, *Cell* **76**, 287–299.

[9] Klein, J. (1986). *Natural History of the Major Histocompatibility Complex*. Wiley, New York.

[10] Morris, A., Hewitt, C. & Young, S. (1994). The major histocompatibility complex: its genes and their roles in antigen presentation, *Molecular Aspects of Medicine* **15**, 377–503.

[11] Nei, M. & Hughes, A.L. (1991). lymorphism and evolution of the major histocompatibility complex in mammals, in *Evolution at the Molecular Level*, R.K. Selander, A.G. Clark & T.S. Whittam, eds. Sinauer, Sunderland, pp. 222–247.

[12] Parham, P., Adams, E.J. & Arnett, K.L. (1995). The origins of HLA-A,B,C polymorphism, *Immunology Reviews* **143**, 141–180.

[13] Thomson, G. (1995). HLA disease associations: models for the study of complex human genetic disorders, *Critical Reviews in Clinical Laboratory Science* **32**, 183–219.

[14] Trowsdale, J. (1993). Genomic structure and function in the MHC, *Trends in Genetics* **9**, 117–122.

DONNA D. KOSTYU

Hogben, Lancelot Thomas

Born: December 9, 1895.
Died: August 22, 1975.

Hogben had a brilliant academic career in biology, with chairs at the London School of Economics, Aberdeen, and Birmingham, and election to Fellowship of the Royal Society in 1936. He wrote popular books on mathematics, science, and linguistics. During the 1939–1945 war he became Acting Director of Medical Statistics at the War Office, and from 1947 to 1961 he was Professor of Medical Statistics at the University of Birmingham. His main interests were in procedures for recording and tabulating medical data, and in the philosophical basis of statistics. In the latter context, he was critical of the claims made for randomized **clinical trials,** and, in a 1957 book on *Statistical Theory,* he expressed dissatisfaction with probabilistic **inference** as a basis for the interpretation of statistics. A very full biography and a complete bibliography are given in [1].

Reference

[1] Wells, G.P. (1978). Lancelot Thomas Hogben, *Biographical Memoirs of Fellows of the Royal Society* **24**, 183–221.

Horvitz–Thompson Estimator

The estimator known as the *Horvitz–Thompson estimator* (HTE) was developed by Horvitz & Thompson in their classic 1952 paper [4]. In that article they propose the following estimator of a population total, X, that is valid for any **sampling design with or without replacement**:

$$x'_{\text{hte}} = \sum_{i=1}^{v} \frac{x_i}{\pi_i},$$

where x_i is the value of the variable for the ith enumeration unit in the sample, π_i is the probability of the ith enumeration unit being selected into the sample and v is the number of distinct enumeration units sampled (as distinguished from n, which is the total sample size). Clearly, $n = v$ when sampling is without replacement.

In that same paper, the authors showed that the HTE is **unbiased** with **standard error** given by the following expression:

$$\text{se}(x'_{\text{hte}}) = \left[\sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i} \right) x_i^2 + \sum_{i=1}^{N} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) x_i x_j \right]^{1/2},$$

where N is the number of enumeration units in the population and π_{ij} is the probability that both enumeration units i and j are included in the sample.

In addition, they showed that the estimator, $\overline{v}(x'_{\text{hte}})$, given by the expression

$$\overline{v}(x'_{\text{hte}}) = \sum_{i=1}^{v} \frac{1-\pi_i}{\pi_i^2} x_i^2 + \sum_{i=1}^{v} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{x_i x_j}{\pi_{ij}},$$

is an unbiased estimator of the **variance** of x'_{hte}.

The Horvitz–Thompson estimator differs from an earlier unbiased estimator generally referred to as the *Hansen–Hurwitz estimator*:

$$x'_{\text{hh}} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i}{\pi'_i}$$

proposed by Hansen & Hurwitz [2] which is valid when sampling is with replacement, and where π'_i is the probability of selecting the ith enumeration unit at any drawing of the sample.

These estimators are illustrated in Table 1, in which $N = 3, n = 2, X_1 = 1, X_2 = 3, X_3 = 4, \pi'_1 = 1/6, \pi'_2 = 2/6, \pi'_3 = 3/6$, and the sampling is with replacement. As can be seen, the two estimators do not necessarily produce the same numerical estimate for the same sample, and both are unbiased estimators of the population total. Both estimators are used in unequal probability sampling, including widely used applications such as **sampling with probability proportionate to size** and **network sampling**. Because the Horvitz–Thompson estimator is appropriate for situations in which sampling is without replacement, however, it has been especially important in the development of design-based **estimation** theory and methodology for sample surveys.

Table 1 Comparison of estimators

Enumeration units in sample (ordered)	Probability of sample occurring	Horvitz–Thompson estimator	Hansen–Hurwitz estimator
X_1, X_1	0.0278	3.27	6.0
X_1, X_2	0.0556	8.67	7.5
X_1, X_3	0.0833	8.61	7.0
X_2, X_1	0.0556	8.67	7.5
X_2, X_2	0.1111	5.40	9.0
X_2, X_3	0.1667	10.73	8.5
X_3, X_1	0.0833	8.61	7.0
X_3, X_2	0.1667	10.73	8.5
X_3, X_3	0.2500	5.33	8.0

For more detailed discussions of the Horvitz–Thompson estimator, we refer the reader to the texts by Hedayat & Sinha [3] and Thompson [5]. Also, **Cochran** discusses this estimator in the *Encyclopedia of Statistical Sciences* [1] (this was one of Professor Cochran's last articles before his death in 1980).

References

[1] Cochran, W.G. (1983). Horvitz–Thompson estimator, in *Encyclopedia of Statistical Sciences*, Vol. 3. S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 665–668.

[2] Hansen, M.M. & Hurwitz, W.N. (1943). On the theory of sampling from finite populations, *Annals of Mathematical Statistics* **14**, 333–362.

[3] Hedayat, A.S. & Sinha, B.K. (1991). *Design and Inference in Finite Population Sampling*. Wiley, New York.

[4] Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**, 663–685.

[5] Thompson, S.K. (1992). *Sampling*. Wiley, New York.

PAUL S. LEVY

Hospital Market Area

A hospital market area (HMA) is the geographic area served by a hospital or a group of hospitals. Market areas are usually defined on the basis of patient origin studies, which examine the zip (postal) codes in which the patients of a hospital reside. HMAs are used in **health services research** to define the populations which provide the denominator for hospital admission rates. For example, one might use the number of back surgeries performed on people who live in a particular HMA, divided by the HMA population, as the admission rate for that area's hospital. In **small-area variation analysis**, one would examine the admission rates for different HMAs to find areas with particularly high or low admission rates, suggesting inappropriate use of services, perhaps attributable to the hospital in that HMA. Unfortunately, however, several hospitals may serve the same area, and a particular hospital may draw patients from many areas, especially in urban and suburban areas. These considerations make a hospital's admission rate both conceptually unclear and technically difficult to estimate.

Two methods have been proposed for defining HMAs and their corresponding hospital-based admission rates from population-based data. The plurality rule of Wennberg & Gittelsohn [3] assigns the population and the hospital admissions from each zip code to the hospital which is the recipient of the plurality of the admissions from the area. While simple to apply, this method is flawed by considerable **misclassification error** since many (potentially even a majority) of the persons and admissions from any given small area will be assigned to one hospital when, in truth, they "belong" to another. Furthermore, this method underemphasizes the utilization of small hospitals, since small hospitals infrequently constitute a plurality in any small area.

Griffith et al. [1] propose a "proportional allocation" method, which allocates a proportion of each small area's population to each hospital based on the proportion of that area's admissions to each hospital. Thus, if Hospital X received 24% of area A's admissions, it would be allocated 24% of Area A's population as well. By summing the populations allocated to Hospital X across all small areas, one can estimate a theoretical catchment population for Hospital X. Dividing this theoretic denominator into Hospital X's admissions yields an "admission rate" for Hospital X. The principal flaw in this method is that, because the population at risk is allocated in proportion to the numerator, it tends to diminish any true differences between hospitals in their propensity to admit.

There are other problems in defining hospital market areas. HMAs based on one patient origin study may not be appropriate for all conditions that might be studied. For example, HMAs defined by a patient origin study of all hospital admissions would not be appropriate for a study of trauma admission rates if one of the hospitals had a renowned trauma center that attracted patients from a large area. Origin studies are often based on Medicare data [2], which, since it is available primarily for people over age 65, may not be appropriate for services for younger people.

References

[1] Griffith J.R., Restuccia, J.D., Tedeschi, P.J., Wilson, P.A. & Zuckerman, H.S. (1981). Measuring community hospital services in Michigan, *Health Services Research* **16**, 135–160.

[2] Makuc, D.M., Haglund, B., Ingram, D.D., Kleinman, J.C. & Feldman, J.J. (1991). Health service areas for the United States, *Vital and Health Statistics, Series 2: Data Evaluation and Methods Research* **112**, 1–102.

[3] Wennberg, J. & Gittelsohn, A. (1973). Small area variations in health care delivery, *Science* **182**, 1102.

PAULA DIEHR

Hotelling, Harold

Born: September 29, 1895, in Fulda, Minnesota.
Died: December 26, 1973, in Chapel Hill, North Carolina.

Reproduced by permission of the Royal Statistical Society

Harold Hotelling was responsible for much original theoretical work in both statistics and mathematical economics and did much to advance the teaching of statistics at US universities, including Columbia and the University of North Carolina. Hotelling's undergraduate degree was a B.A. in journalism from the University of Washington in 1919, but his mathematical talent was recognized and encouraged by Eric T. Bell. Hotelling received an M.S. degree in Mathematics at the University of Washington in 1921 and a Doctorate of Philosophy at Princeton University in 1924, with a dissertation in the field of topology.

Following his doctorate, he spent seven years at Stanford University, first as Research Associate in the Food Research Institute and later as a Associate Professor of Mathematics. During his time at Stanford, he applied mathematical ideas to problems in journalism and political science, population and food supply, and theoretic economics. In 1929, he spent six months with **R.A. Fisher** at the Rothamstead Experimental Station at Harpenden in England which helped to develop his strong interest in mathematical statistics. In 1931 he published perhaps his most important contribution to statistics when he generalized to the multivariate case **Student's** t test for the **mean** of a univariate **normal distribution** [4] (*see* **Multivariate** t **Distribution**). This test has become known as **Hotelling's generalized** T^2 test and was later recognized as having wide applicability in statistics.

In 1931 Hotelling was appointed Professor of Economics at Columbia University where he stayed for 15 years. During World War II he organized the Statistical Research Group, which was engaged in statistical work relating to military problems. The group included **Abraham Wald**, W. Allen Wallis, and Jacob Wolfowitz. During this time Wald developed his theory of **sequential analysis**. In 1946 Hotelling was invited by **Gertrude Cox** to organize a Department of Mathematical Statistics at the University of North Carolina at Chapel Hill (UNC-Chapel Hill), which became an important center for statistical research and teaching. He recruited many outstanding statisticians, including R.C. Bose, S.N. Roy, W. Hoeffding, W.G. Madow, H.E. Robbins, W.L. Smith, and N.L. Johnson. Hotelling remained at UNC–Chapel Hill until his death.

Hotelling proposed a method of **principal components** [6] which is applicable to problems of **factor analysis** arising in educational testing. Using ideas of n-dimensional geometry, the principal components are linear functions of multivariate observations, the first of which has the greatest variability and each subsequent one less variability. A similar mathematical idea underlies Hotelling's theory of **canonical correlations** [7]. Among Hotelling's other contributions to statistics are his paper on differential equations subject to error [3], one of the first dealing with statistical problems related to **stochastic processes**; a paper (jointly with H. Working) on the interpretation of trends [18] which had one of the first examples of a **confidence region** and the idea of **multiple comparisons**; the derivation of the distribution of Spearman's **rank correlation** coefficient [11]; and the experimental determination of the maximum of a function [10].

In economic theory, he dealt with problems in depreciation and the importance of maximizing principles [2]; the interrelated demand and supply functions of profit maximizers [5]; and welfare economics [8], possibly his most important contribution to mathematical economics.

Hotelling had a talent for attracting excellent faculty members, both at Columbia and the University of

North Carolina, and played an important role in raising standards in statistical research and developing mathematical statistics as a respected academic discipline. He was a strong advocate of the importance of **teaching statistics** [9], which had an impact on the academic community and aided in the establishment of departments of statistics at American universities. Levene paid tribute to his excellence as a teacher and lecturer [12].

In 1955 Hotelling received an honorary LL.D. from the University of Chicago. In 1963 he received an honorary D.Sc. from the University of Rochester and was an Honorary Fellow of the **Royal Statistical Society** and a Distinguished Fellow of the American Economic Association. In 1936–37 he was the President of the Econometric Society and in 1941 of the Institute of Mathematical Statistics. In 1970 he was elected to the National Academy of Sciences and in 1972 received the North Carolina Award for Science. His final award, in 1973, was his election to membership of the Accademia Nazionale dei Lincei in Rome, which occurred shortly before his death.

Hotelling's contributions to statistics have been memorialized by Anderson [1], Madow [13, 14], **Neyman** [15], and Olkin [16], and to mathematical economics by Samuelson [17].

References

[1] Anderson, T.W. (1960). Harold Hotelling's research in statistics, *American Statistician* **14**, 17–21.

[2] Hotelling, H. (1925). A general mathematical theory of depreciation, *Journal of the American Statistical Association* **20**, 340–353.

[3] Hotelling, H. (1927). Differential equations subject to error, and population estimates, *Journal of the American Statistical Association* **20**, 340–353.

[4] Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.

[5] Hotelling, H. (1932). Edgeworth's taxation paradox and the nature of demand and supply functions, *Journal of Political Economy* **40**, 577–616.

[6] Hotelling, H. (1933). Analysis of complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441, 498–520.

[7] Hotelling, H. (1936). Relations between two sets of variates, *Biometrika* **28**, 321–377.

[8] Hotelling, H. (1938). The general welfare in relation to problems of taxation and of railway and utility rates, *Econometrica* **6**, 242–269. (Presidential address to the Econometric Society at the meeting in Atlantic City, N.J., December 28, 1937).

[9] Hotelling, H. (1940). The teaching of statistics, *Annals of Mathematical Statistics* **11**, 457–470.

[10] Hotelling, H. (1941). Experimental determination of the maximum of a function, *Annals of Mathematical Statistics* **12**, 20–45.

[11] Hotelling, H. & Pabst, M.R. (1936). Rank correlation and tests of significance involving no assumption of normality, *Annals of Mathematical Statistics* **7**, 29–43.

[12] Levene, H. (1974). In memoriam: Harold Hotelling, 1895–1973, *American Statistician* **28**, 71–73.

[13] Madow, W.G. (1960). Harold Hotelling, in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow & H.B. Mann, eds. Stanford University Press, Stanford, pp. 3–5.

[14] Madow, W.G. (1960). Harold Hotelling as a teacher, *American Statistician* **14**, 15–17.

[15] Neyman, J. (1960). Harold Hotelling: a leader in mathematical statistics, in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow & H.B. Mann, eds. Stanford University Press, Stanford, pp. 6–10.

[16] Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G. & Mann, H.B., eds (1960). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford.

[17] Samuelson, P.A. (1960). Harold Hotelling as mathematical economist, *American Statistician* **14**, 21–25.

[18] Working, H. & Hotelling, H. (1929). Applications of the theory of error to the interpretation of trends, *Journal of the American Statistical Association* **24**, Supplement, 73–85.

EDMUND A. GEHAN

Hotelling's T^2

The Hotelling T^2 statistic is a generalization of the squared univariate t (*see* **Student's t Distribution**) for testing hypotheses on the normal distribution mean, when the population variance is unknown and must be estimated from the sample observations. For a single random sample of N p-dimensional observations from the **multivariate normal distribution**, the Hotelling statistic for testing the hypothesis $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ on the mean vector $\boldsymbol{\mu}$ is

$$T^2 = N(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0).$$

In the squared univariate statistic the means have been replaced by mean vectors and the reciprocal of the sample variance has become the inverse of the sample **covariance matrix S**. T^2 is thus a measure of the distance of the sample mean vector from the hypothesized population vector, but in the metric of **S**. This case of Hotelling's T^2 test, its general derivation, and its application to two independent samples and repeated measures designs are covered in this article.

The Hotelling T^2 Test

Derivation

The T^2 statistic was originally proposed by Hotelling [3]. Hotelling's account of its derivation by the invariance properties of the roots of a certain determinantal equation is contained in [4]. T^2 is also the sample analog of **Mahalanobis distance** [5] of $\overline{\mathbf{x}}$ and $\boldsymbol{\mu}_0$. Construction of the hypothesis test by the generalized **likelihood ratio** principle gives a statistic that is a monotonic function of T^2 [1]. Roy [10, 11] derived the T^2 statistic by his **union–intersection principle**; the explicit single-sample case has been given by Morrison [7].

Distribution of T^2

Hotelling [3] first found the distribution of T^2 by a geometrical argument. More recently, Rao [9] has given an ingenious and simple derivation of the distribution for both the null and alternative hypotheses. We give a very general definition of T^2, state its distribution, and then apply it to cases of T^2 computed from sample observations. Let $\mathbf{Y}$ be a $p \times 1$ random vector with the multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The sums of squares and products matrix $n\mathbf{S}$ has the **Wishart distribution** [[1], Chapter 7] with parameters degrees of freedom n and covariance matrix $\boldsymbol{\Sigma}$, and is distributed independently of $\mathbf{Y}$. Then the general Hotelling statistic is

$$T^2 = \mathbf{Y}'\mathbf{S}^{-1}\mathbf{Y},$$

and its linear transformation,

$$F = \left[\frac{(n - p + 1)}{np}\right] T^2,$$

has the noncentral F **distribution** with degrees of freedom p, $n - p + 1$, and noncentrality parameter $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. If $\mathbf{E}(\mathrm{Y}) = \boldsymbol{\mu} = \mathbf{0}$, F has the usual central F distribution with p and $n - p + 1$ degrees of freedom. In the context of a single random sample and a test of the null hypothesis $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$, $\mathbf{Y} = \overline{\mathbf{y}} - \boldsymbol{\mu}$ and $n = N - 1$. Under $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$, $\overline{\mathbf{y}}$ is $\mathrm{N}[\boldsymbol{\mu}_0, (1/N)\boldsymbol{\Sigma}]$, $T^2 = \mathrm{N}(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)$, and $F = [(N - p)/(N - 1)p]T^2$ has the central F distribution with p and $N - p$ degrees of freedom. For the general alternative hypothesis $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_1$, $[(N - p)/(N - 1)/p]T^2$ has the noncentral F distribution with p and $N - p$ degrees of freedom and noncentrality parameter $\delta^2 = N(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. **Power** probabilities of the T^2 test can be found from the Pearson–Hartley charts of the noncentral F distribution [7, 8], or by statistical *software* (e.g. [2] and [6]). **Sample size determination** for a given α-level test and power probability must, of course, be made iteratively, since the sample size appears both in the second degrees of freedom and in the noncentrality parameter.

Affine Invariance Property

The T^2 statistic has an important invariance property: it is unaffected by affine **transformations**

$$\mathbf{W} = \mathbf{AY} + \mathbf{h},$$

in which $\mathbf{A}$ is a $p \times p$ matrix of real constants with a nonzero determinant, and $\mathbf{h}$ is a $p \times 1$ vector of constants. The transformation must be applied to the sample mean vector $\overline{\mathbf{y}}$ as well as the population mean

vector $\boldsymbol{\mu}_0$. Use of the transformation in the single-sample Hotelling statistic gives

$$\begin{aligned} T_{\mathrm{W}}^2 &= N(\mathbf{A}\bar{\mathbf{y}} + \mathbf{h} - \mathbf{A}\boldsymbol{\mu}_0 - \mathbf{h})'(\mathbf{ASA}')^{-1} \\ &\quad \times (\mathbf{A}\bar{\mathbf{y}} + \mathbf{h} - \mathbf{A}\boldsymbol{\mu}_0 - \mathbf{h}) \\ &= N(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{A}'(\mathbf{ASA}')^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \\ &= T_{\mathrm{Y}}^2, \end{aligned}$$

and, of course, the affine invariance property can be verified for other more general forms of T^2. The statistic is not only unaffected by scale and location changes, but is also unchanged by oblique linear transformations of the coordinate system as well.

Tests of Hypotheses

Single Sample

We have already introduced the single-sample T^2 test of $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $\mathrm{H}_1\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$: reject H_0 at the α level if $F = [(N-p)/(N-1)p]T^2 > F_{\alpha;p,N-p}$. The hypothesized mean vector $\boldsymbol{\mu}_0$ is given by the analyst from a substantive context: psychological test score means, dimension, or other specification means in quality assurance, or normative values of the random vector components.

Two Samples

The model for the two-sample T^2 test for equality of multivariate normal mean vectors assumes that independent random samples have been drawn from each population, and that the populations have a common covariance matrix $\boldsymbol{\Sigma}$. The observation vectors in the respective samples will be denoted by $\mathbf{x}_{11}, \ldots, \mathbf{x}_{N1}$, $\mathbf{x}_{12}, \ldots, \mathbf{x}_{M2}$. The sample mean vectors are

$$\bar{\mathbf{x}}_1 = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathbf{x}_{i1}, \qquad \bar{\mathbf{x}}_2 = \left(\frac{1}{M}\right)\sum_{i=1}^{M}\bar{\mathbf{x}}_{i2}$$

and the pooled, or within-sample, covariance matrix estimating $\boldsymbol{\Sigma}$ is

$$\mathbf{S} = \left(\frac{1}{N+M-2}\right)\left[\sum_{i=1}^{N}(\mathbf{x}_{i1} - \bar{\mathbf{x}}_1)(\mathbf{x}_{i1} - \bar{\mathbf{x}}_1)' + \sum_{i=1}^{M}(\mathbf{x}_{i2} - \bar{\mathbf{x}}_2)(\mathbf{x}_{i2} - \bar{\mathbf{x}}_2)'\right].$$

The two-sample T^2 statistic is

$$T^2 = \left(\frac{NM}{N+M}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

When $\mathrm{H}_0\colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ is true, $F = [(N+M-p-1)/(N+M-2)p]T^2$ has the F distribution with p and $N+M-p-1$ degrees of freedom. The null hypothesis is rejected if $F > F_{\alpha;p,N+M-p-1}$. When the alternative $\mathrm{H}_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ holds, F has the noncentral F distribution with degrees of freedom $p, N+M-p-1$, and noncentrality parameter $\delta^2 = [NM/(N+M)](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Repeated Measurements

Frequently, p observations are taken successively on each of N independent sampling units for a test of the hypothesis that the p means are equal (*see* **Longitudinal Data Analysis, Overview**). For example, plasma-free fatty acid levels might be measured in blood samples taken at $p = 6$ 15-min intervals from normal subjects after they had ingested a particular food or drug. The hypothesis of a common-mean-free fatty acid level at six times might be of interest, and could be tested by Hotelling's T^2 statistic.

The repeated-measures test is equivalent to testing the hypothesis that the $p-1$ successive differences of the variables have zero means. We begin by transforming the p response variables $X_1, \ldots, X_p$ to the successive differences $Y_1, \ldots, Y_{p-1}$ by the linear transformation

$$\begin{aligned} \mathbf{Y} &= \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_{p-1} \end{bmatrix} \\ &= \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_p \end{bmatrix} \\ &= \mathbf{CX}. \end{aligned}$$

We test $\mathrm{H}_0\colon \mathrm{E}(Y_1) = \cdots = \mathrm{E}(Y_{p-1}) = 0$ by

$$T^2 = N\bar{\mathbf{y}}'\mathbf{S}^{-1}\bar{\mathbf{y}},$$

or equivalently in terms of the observations on the original variables,

$$T^2 = N\bar{\mathbf{x}}'\mathbf{C}'(\mathbf{CSC}')^{-1}\mathbf{C}\bar{\mathbf{x}},$$

where $\mathbf{C}$ is the $(p-1)\times p$ matrix of the successive difference transformation. As in the single-sample case, $F = [(N-p+1)/(N-1)(p-1)]T^2$ has the F distribution with $p-1$ and $N-p+1$ degrees of freedom, and we reject the null hypothesis of equal response variable means if $F > F_{\alpha;p-1,N-p+1}$.

Paired Response Variables

Some repeated-measurements experiments consist of the same p response variables observed at two different times or conditions on the same subjects or other sampling units. If we represent the $p\times 1$ response vectors at the two times by $\mathbf{X}_1$ and $\mathbf{X}_2$ we can test the hypothesis $\mathrm{H}_0 : \mathrm{E}(\mathbf{X}_1) = \mathrm{E}(\mathbf{X}_2)$ by the T^2 statistic. A random sample of N independent observation vectors partitioned according to the two times as $[\mathbf{x}_{i1}', \mathbf{x}_{i2}']$ yields the respective partitioned sample mean vector and covariance matrix

$$\begin{bmatrix}\bar{\mathbf{x}}_1'\\ \bar{\mathbf{x}}_2'\end{bmatrix}, \qquad \begin{bmatrix}\mathbf{S}_{11} & \mathbf{S}_{12}\\ \mathbf{S}_{12}' & \mathbf{S}_{22}\end{bmatrix},$$

where

$$\mathbf{S}_{ij} = \sum_{h=1}^{N}(\mathbf{x}_{hi}-\bar{\mathbf{x}}_i)(\mathbf{x}_{hj}-\bar{\mathbf{x}}_j)', \quad i,j = 1,2.$$

The Hotelling statistic is

$$T^2 = N(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'(\mathbf{S}_{11}+\mathbf{S}_{22}-\mathbf{S}_{12}-\mathbf{S}_{12}')^{-1} \times (\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2),$$

and reflects the correlations between the two times through the elements of the submatrix $\mathbf{S}_{12}$. T^2 is merely the extension of the **paired *t* test** to p pairs of response variables. If the data were transformed to an $N\times p$ matrix of paired differences, then the T^2 statistic would reduce to the single-sample T^2 described previously. When H_0 is true, $F = [(N-p)/(N-1)p]T^2$ has the F distribution with p and $N-p$ degrees of freedom. H_0 would be rejected when F exceeds the right-tail α-level critical value for that distribution.

Confidence Statements Obtained from T^2

Confidence Region for a Single Mean Vector

The distribution of T^2 for a single random sample from the multinormal distribution can be used to obtain this ellipsoidal $100(1-\alpha)\%$ confidence region for the population mean vector $\boldsymbol{\mu}$:

$$N(\boldsymbol{\mu}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{x}}) \le [(N-1)p/(N-p)]F_{\alpha;p,N-p}$$

(*see* **Confidence Intervals and Sets**).

Simultaneous Confidence Intervals

Rejection of the null hypothesis by the T^2 test still does not indicate *which* of the p responses may have contributed to that decision. Roy's union–intersection derivation of T^2 [10] leads directly to simultaneous tests and confidence intervals for linear compounds of the population means (*see* **Simultaneous Inference**). "Simultaneous" means that one may construct an unlimited number of confidence intervals and still have an overall coverage probability of $1-\alpha$, or test infinitely many hypotheses and still enjoy an overall type I error rate no greater than α. For the single-sample case the $100(1-\alpha)\%$ Roy–Bose [12] simultaneous confidence interval for the linear compound $\mathbf{a}'\boldsymbol{\mu}$ is

$$\mathbf{a}'\bar{\mathbf{x}} - \left\{(1/N)\mathbf{a}'\mathbf{Sa} \times[(N-1)p/(N-p)]F_{\alpha;p,N-p}\right\}^{1/2} \le \mathbf{a}'\boldsymbol{\mu} \le \mathbf{a}'\bar{\mathbf{x}} + \left\{(1/N)\mathbf{a}'\mathbf{Sa} \times[(N-1)p/(N-p)]F_{\alpha;p,N-p}\right\}^{1/2},$$

where $\mathbf{a}$ is any $p\times 1$ vector of constants chosen by the investigator. For the two-sample situation the $100(1-\alpha)\%$ simultaneous confidence interval for the linear compound $\mathbf{a}'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ of the differences of the mean vector elements is

$$\mathbf{a}'(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) - \left\{[(N+M)/NM]\mathbf{a}'\mathbf{Sa} \times[(N+M-2)p/(N+M-p-1)] \times F_{\alpha;p,N+M-p-1}\right\}^{1/2} \le \mathbf{a}'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \le \mathbf{a}'(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) + \left\{[(N+M)/NM]\mathbf{a}'\mathbf{Sa}\right.$$

$$\times [(N + M - 2)p/(N + M - p - 1)]$$
$$\times F_{\alpha;p,N+M-p-1}\}^{1/2}.$$

If the interval contains zero, then the hypothesis $H_0: \mathbf{a}'\boldsymbol{\mu}_1 = \mathbf{a}'\boldsymbol{\mu}_2$ is tenable at the α level in the simultaneous testing sense. Alternatively, in both cases families of hypotheses can also be tested with an overall type I error rate no greater than α.

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Galen Research, Inc. (1990). *Electronic Tables*. Galen Research, Inc., Salt Lake City.

[3] Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.

[4] Hotelling, H. (1954). Multivariate analysis, in *Statistics and Mathematics in Biology*, O. Kempthorne, T. Bancroft, J. Gowen & J.L. Lush, eds. Hafner, New York, pp. 67–80.

[5] Mahalanobis, P.C. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Sciences of India* **2**, 49–55.

[6] Mehta, C.R. & Patel, N.R. (1994). *StaTable^{TM} Electronic Tables for Statisticians and Engineers*. Cytel Software Corporation, Cambridge, Mass.

[7] Morrison, D.F. (1990). *Multivariate Statistical Methods*, 3rd Ed. McGraw-Hill, New York.

[8] Pearson, E.S. & Hartley, H.O. (1951). Charts of the power function of the analysis of variance tests, derived from the non-central F distribution, *Biometrika* **38**, 112–130.

[9] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. Wiley, New York.

[10] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[11] Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

[12] Roy, S.N. & Bose, R.C. (1953). Simultaneous confidence interval estimation, *Annals of Mathematical Statistics* **24**, 513–536.

(*See also* **Multivariate Analysis of Variance**; **Multivariate Analysis, Overview**)

DONALD F. MORRISON

Human Genetics, Overview

Mendel's laws underlie the distribution of genetic traits observed in individuals. At each genetic locus, an individual receives one **gene** which is a copy of a randomly chosen one of the two genes of the father, and one which is a copy of a randomly chosen one of the two genes of the mother. Each individual passes on to each offspring a randomly chosen one of his two genes, independently to each offspring and independently of the gene contributed by his spouse. The different allelic forms of the genes at a locus, acting in combination with alleles at other loci and with environmental effects, give rise to different phenotypes, the observable characteristics of individuals. Alleles at loci on different chromosomes are inherited independently, but alleles at loci on the same chromosome are *linked*, or correlated, in their inheritance, owing to the process of meiosis which gives rise to the gamete cells.

At the population level, new alleles arise by mutation, and frequencies of alleles are influenced by the genetic forces of selection and the demographic forces of migration and population structure. Since populations are finite, allele frequencies will change over time under random genetic drift, even in the absence of directional genetic or demographic forces (*see* **Population Genetics**). Whereas genetic analysis of other species has been directed towards an understanding of evolution and population biology, or to the increase of crop yields and animal produce, human genetics has been primarily focused on an understanding of the genetic determinants of human disease.

The year 1900 saw the rediscovery of Mendel's work, 1901 the first discovery of a human **blood group** system, and 1902 the first application of Mendelian principles in medical genetics, setting the stage for the development of human medical and population genetics. With the analysis of data from human (as opposed to experimental) populations, came the need to address questions of **ascertainment** [11, 37]. With the discovery of blood group systems, came the first array of **genetic markers** that could be used both to assess human diversity and as markers in **linkage analysis**.

In the 1930s there was a rapid expansion in the development of approaches to the statistical analysis of human genetic data, with the work of Haldane, Hogben, and Fisher. Although Mendelian principles had been applied earlier in assessing the proportion of affected offspring in families ascertained for segregation of rare recessive diseases [1, 37], this period also saw the earliest formal **segregation analyses**, comparing alternative models for the underlying basis of a genetic trait [12, 16], and consequently further development in a statistical framework and model for ascertainment [8]. Also at this time came the recognition that the methods of linkage analysis already used in experimental populations could be applied also to data collected from human families ascertained for a genetic disease [9, 13].

There is no strict separation between inferences of the genetic basis of traits from family data and from population data. One of the earliest applications of population genetic principles to human disease was Haldane's consideration of the expected frequencies of Mendelian genetic diseases in terms of mutation-selection balance [14]. One of the first statistical analyses of population data was that of Bernstein [2] leading to a resolution of the basis of the ABO blood types, while Fisher [10] regarded his analysis of the rhesus blood group system as a fine example of scientific inference. The resolution of human genetic blood groups and enzyme systems not only provided genetic markers for linkage analysis but also a source of extensive information on human diversity.

In fact, the discovery of many blood group systems in the first half of the century prompted many studies of the extent of human diversity, and a search for explanations of observed data on the basis of models of selection, and of the migration patterns of human history. Many of the population genetic ideas and models underlying such inferences were described by Cavalli-Sforza & Bodmer [5]. While **demography** may be the major factor influencing global patterns of human diversity, a gene need not itself be subject to differential selection in order for its selection to have an impact. The discovery of the many variants present in the human white blood cells **(HLA system)**, and their multiple and complex associations with disease prompted renewed effort in understanding patterns of human variation. The phenomenon of "hitchhiking" [20], where the selective effects of genes at closely linked loci affect patterns of observed variation, has been used to explain unusually high frequencies of

some human disease alleles in some human populations [35]. Thomson et al. [34] used human HLA data to examine the evolutionary interactions of selection, migration, and linkage. A more recent review of the statistical approaches to an understanding of associations of HLA and disease is given in [33]. Although current research in human genetics is often more focused towards individual data than to information at the population level, the data painstakingly compiled by Mourant et al. [23] remain a rich source of information, while the major work of Cavalli-Sforza et al. [6] shows how population allele frequencies reflect the imprint of human history.

While demography and genetic selection affect population allele frequencies, mutation is the source of new genetic variation. The estimation of mutation rates is therefore an important aspect of statistical genetics. It is hard to obtain precise estimates of human mutation rates by direct methods, since mutation rates are small. Indirect methods of estimation use current levels of genetic variation, and require assumptions about population size and structure. Over the 45 years since 1951 [25], the leader in study of human mutation rates by both direct [27] and indirect [26] methods has been J.V. Neel.

From 1935 to 1975 there were many developments in the statistical analysis of human genetic data observed on relatives, but the basic framework of segregation and linkage analysis, as developed by J.B.S. Haldane and R.A. Fisher, remained largely unchanged. For computational reasons, early analyses had been restricted to nuclear families or small pedigrees. With the widespread availability of digital computers, increasing interest in analysis of data on more extended pedigrees led to the development of new computational approaches [7]. With computational power permitting the analysis of larger data sets, and hence perhaps resolving more complex traits, came the necessity for more complex trait models such as the **mixed model in segregation analysis** [22]. In linkage analysis particularly, there were further developments, leading to a better understanding of how inferences could be drawn [15, 21] and to a better understanding of their properties of linkage likelihoods [31]. Ott [29] covers many of these developments.

Since 1980, with the development of molecular biology, there has been an explosion in the number of **polymorphisms** available for use as genetic markers in linkage analysis. Human genome maps at centimorgan density are now a reality [24], and the limitation in linkage analysis is no longer the availability of segregating markers, but the trait information. Simple Mendelian traits are rapidly being mapped, and the relevant genes identified. However, if a trait is exceedingly rare, or shows genetic heterogeneity, or if its genetic basis is uncertain or complex involving alleles at several loci, then problems in linkage analysis remain.

The computation of a linkage likelihood over a pedigree requires a specific segregation analysis model for the trait to be assumed. For traits whose basis is uncertain, particularly of incomplete **penetrance** or delayed onset, linkage detection methods using only affected individuals have been developed. These are more robust to trait model assumptions; indeed, under the null hypothesis of no linkage, the distribution of the test statistic is often independent of trait model assumptions. Such methods date back to the 1930s, when Penrose [30] introduced sib pair methods, but more recently have been extended to other types of relationship [3, 19, 36]. In many cases, the use of only affected pedigree members can greatly increase robustness with little loss of power.

Once linkage has been detected, **multipoint linkage analysis** can help to localize the position of the gene more precisely. However, multipoint methods are computationally exceedingly intensive, particularly where there are many unobserved members of the pedigree. Moreover, there are limits to the resolution of linkage mapping (*see* **Genetic Map Functions**) [4]. The scale of resolution depends on the number of segregations that can be (explicitly or implicitly) observed. Where genetic homogeneity can be assumed, disequilibrium mapping [17] or **haplotype analysis** provides an alternative. Here the exact ancestry of current carriers of a disease allele is unknown, but their shared ancestry results in **linkage disequilibrium** with marker loci at small **genetic distance**. The large number of ancestral segregations provides for a finer mapping scale. Ultimately it may be possible to map at still finer scales by considering the matching and nonmatching segments of individual genomes [28].

Genetic heterogeneity is one of the major difficulties in resolving the genetic basis of any trait. Studies within a given population or of data on a single extended pedigree reduce the chance of heterogeneity

within the data set, but such data sets are often limited in size, and results may not be relevant outside the particular population studied. Gradually, however, more complex traits are being resolved through advances both in the available genetic data and in methods of analysis and computation.

The classic text on human genetics is that of Stern [32]. The more recent text by Khoury et al. [18] provides a thorough overview of approaches in modern genetic epidemiology, while Ott [29] is the best reference text on linkage analysis in human genetics.

References

[1] Alpert, E. (1914). The laws of Naudin-Mendel, *Journal of Heredity* **5**, 492–497.

[2] Bernstein, F. (1925). Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen, *Zeitschrift für induktive Abstammungs- und VererbungsLehre* **37**, 237–270.

[3] Bishop, D.T. & Williamson, J. (1990). The power of identity-by-state methods for linkage analysis, *American Journal of Human Genetics* **46**, 254–265.

[4] Boehnke, M. (1994). Limits of resolution of genetic linkage studies: implications for positional cloning of human disease genes, *American Journal of Human Genetics* **55**, 379–390.

[5] Cavalli-Sforza, L.L. & Bodmer, W.F. (1971). *The Genetics of Human Populations*. Freeman, San Francisco.

[6] Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton.

[7] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[8] Fisher, R.A. (1934). The effects of methods of ascertainment on the estimation of frequencies, *Annals of Human Genetics* **6**, 13–25.

[9] Fisher, R.A. (1934). The amount of information supplied by records of families as a function of the linkage in the population sampled, *Annals of Eugenics* **6**, 66–70.

[10] Fisher, R.A. (1947). The *Rhesus* factor: a study in scientific method, *American Scientist* **35**, 95–102, 113.

[11] Galton, F. (1904). Average number of kinfolk in each degree, *Nature* **70**, 529, 626.

[12] Haldane, J.B.S. (1932). A method for investigating recessive characters in man, *Journal of Genetics* **25**, 251–255.

[13] Haldane, J.B.S. (1934). Methods for the detection of autosomal linkage in man, *Annals of Eugenics* **6**, 26–65.

[14] Haldane, J.B.S. (1935). The rate of spontaneous mutation of a human gene, *Journal of Genetics* **31**, 317–326.

[15] Haldane, J.B.S. & Smith, C.A.B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man, *Annals of Eugenics* **14**, 10–31.

[16] Hogben, L.T. (1931). The genetic analysis of familial traits. I. Single gene substitutions, *Journal of Genetics* **25**, 97–112.

[17] Kaplan, N.L., Hill, W.G. & Weir, B.S. (1995). Likelihood methods for locating disease genes in nonequilibrium populations, *American Journal of Human Genetics* **56**, 18–32.

[18] Khoury, M.J., Beaty, T.H. & Cohen, B.H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press, Oxford.

[19] Lander, E.S. & Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children, *Science* **236**, 1567–1570.

[20] Maynard Smith, J. & Haigh, J. (1974). The hitch-hiking effect of a favourable gene, *Genetical Research* **23**, 23–35.

[21] Morton, N.E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.

[22] Morton, N.E. & MacLean, C.J. (1974). Analysis of family resemblance. III. Complex segregation analysis of quantitative traits, *American Journal of Human Genetics* **26**, 489–503.

[23] Mourant, A.E., Kopec, A.C. & Domaniewska-Sobczak, K. (1976). *The Distribution of Human Blood Groups and Other Polymorphisms*. Oxford University Press, Oxford.

[24] Murray, J.C., Buetow, K.H., Weber, J.L. (and 24 others) (1994). A comprehensive human linkage map with centimorgan density, *Science* **265**, 2049–2064.

[25] Neel, J.V. & Falls, H.F. (1951). The rate of mutation of the gene responsible for retinoblastoma in man, *Science* **114**, 419–422.

[26] Neel, J.V. & Rothman, E.D. (1978). Indirect estimates of mutation rate in tribal Amerindians, *Proceedings of the National Academy of Sciences* **75**, 5585–5588.

[27] Neel, J.V., Satoh, C., Goriki, K., Fujita, M., Takahashi, N., Asakawa, J. & Hazama, R. (1986). The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides, *Proceedings of the National Academy of Sciences* **83**, 389–393.

[28] Nelson, S.F., McCusker, J.H., Sander, M.A., Kee, Y., Modrish, P. & Brown, P.O. (1993). Genomic mismatch scanning: a new approach to genetic linkage mapping, *Nature Genetics* **4**, 11–18.

[29] Ott, J. (1991). *Analysis of Human Genetic Linkage*, 2nd Ed. Johns Hopkins University Press, Baltimore.

[30] Penrose, L.S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage, *Annals of Eugenics* **6**, 133–138.

[31] Smith, C.A.B. (1953). Detection of linkage in human genetics, *Journal of the Royal Statistical Society, Series B* **15**, 153–192.

[32] Stern, C. (1960). *Principles of Human Genetics*, 2nd Ed. Freeman, San Francisco.

[33] Thomson, G. (1981). A review of theoretical aspects of HLA and disease associations, *Theoretical Population Biology* **20**, 168–208.

[34] Thomson, G., Bodmer, W.F. & Bodmer, J. (1976). The HL-A system as a model for studying the interaction between selection migration and linkage, in *Population Genetics and Ecology*, S. Karlin & E. Nevo, eds. Academic Press, New York, pp. 465–498.

[35] Wagener, D.K. & Cavalli-Sforza, L.L. (1975). Ethnic variation in genetic diseases: possible roles of hitch-hiking and epistasis, *American Journal of Human Genetics* **27**, 348–364.

[36] Weeks, D.E. & Lange, K. (1988). The affected-pedigree-member methods of linkage analysis, *American Journal of Human Genetics* **42**, 315–326.

[37] Weinberg, W. (1912). Zur Verebung der Anlage der Bluterkrankheit mit methodol. Erganzungen meiner Geschwistermethode, *Archiv für Rassen- und Gesellschaftsbiologie* **9**, 694–709.

E. THOMPSON

Human Genome Project

The Human Genome Project (see http://www.ornl.gov/hgmis/) was formally initiated on 1 October 1990 as an international scientific effort to (a) map all of the approximately 30 000 functional human **genes** and (b) sequence the approximately 3 billion deoxyribonucleic acid (DNA) nucleotides that make up these genes [30] (*see* **DNA Sequences**; **Human Genetics, Overview**; **Sequence Analysis**). Originally envisioned as a 15-year project, the Human Genome Project has proceeded far more rapidly than anticipated, due in large part to rapidly accelerating improvements in molecular technology. A major milestone of the Project was reached in 2001 with the publication of a first draft sequence of the human genome [10]; a parallel industry effort undertaken by Celera Genomics was published contemporaneously [32].

The Project's primary goals were to enable a fuller understanding of the genetic basis of **complex diseases** and to allow new insights into human evolution through the comparison of the human genome with the genomes of other organisms. Secondary goals of the Human Genome Project included: developing informatics tools to store, disseminate and analyze the very large amounts of data produced (*see* **Bioinformatics**); technology transfer to the private sector; and developing appropriate responses to any ethical, legal or social concerns arising from the Project.

The availability of a draft outline of the human genome will greatly assist the investigation of the complex interrelationships of genetically programmed phenotypes with the constituent genes comprising each individual's unique version of the human genome. Several kinds of biologic information are being produced by the Human Genome Project to enable such investigations, including **DNA sequences**, physical maps, genetic maps (*see* **Genetic Map Functions**) and genetic **polymorphisms**. These data have increased exponentially over the last decade and are being collated in large, web-based databases.

There are many potential benefits anticipated from the Human Genome Project. These include advances in **molecular epidemiology**, allowing benefits such as improved disease diagnosis; improved risk models allowing early detection of increased disease risk; enhanced drug design; gene therapy; and drugs tailored to individual patients based on their **pharmacogenetic** profiles. Other potential benefits expected to arise from the Human Genome Project include enhancements in the fields of **statistical forensics**, bioarchaeology, evolutionary biology and **population genetics**.

Genetic Polymorphism Discovery

An important component of the Human Genome Project that is currently the focus of intense research effort internationally is the discovery and utilization of genetic **polymorphisms** for genetic **association** studies (*see* **Disease-marker Association**). The simplest and most abundant class of polymorphism derives from a single-base substitution of one nucleotide for another – a single nucleotide polymorphism (SNP; pronounced "snip") [17]. SNPs are recognized through a variety of techniques that exploit the known DNA sequence variant [17]. SNPs may be found in coding or regulatory regions of a gene and thus can directly affect gene function or expression.

The generation of SNP maps from high-throughput sequencing projects [17, 24, 31, 33] has continued to accelerate over the last decade [1, 5, 6, 18, 22, 25], with the hope that these data will facilitate the process of gene discovery in complex human disease. In addition to large government-sponsored projects in England [http://www.sanger.ac.uk/], the US [4] and Japan [23], there are now several major industrial group efforts [14, 15], a large academic–industry consortium effort [19], and a number of smaller academic programs [e.g. http://pga.bwh.harvard.edu/] devoted to large-scale SNP discovery. A current focus in human genetics is thus on SNP discovery, leading to the creation of SNP catalogues, and on improving technologies for SNP genotyping.

As part of the intense research effort to improve our ability to discover the genetic determinants of complex human disease over the last decade, technologic advances in the laboratory related to sequencing and SNP genotyping have proceeded at a very rapid rate (*see* **Genetic Markers**). Although the pace of technologic development in SNP analysis is rapid [7, 8], using microarray and other technologies [16], there are many technical problems with these systems that limit their utility at present, such as cost and the inherent lack of flexibility in hardwiring markers on a chip.

There are six primary areas of potential application for SNP technologies in improving our understanding of the etiology of complex human disease: gene mapping; candidate polymorphism association testing; pharmacogenetics; diagnostics and risk profiling; prediction of response to nonpharmacologic environmental stimuli; and homogeneity testing and epidemiologic study design [25]. While only a few of these areas are currently areas of active research in human genetics, it is likely that some or all of these areas will become relevant to investigations of the genetic susceptibility to human disease.

Implications of the Human Genome Project for Biostatistics

The Human Genome Project has had important implications for the fields of biostatistic genetics and **genetic epidemiology**. Catalyzed in part by the vast amounts of data generated by the Human Genome Project and the SNP genotyping efforts in complex human disease, it has become clear that concomitant statistical advances in the mapping of complex traits will also be required [12, 28, 35] (*see* **Linkage Analysis, Model-based**; **Linkage Disequilibrium**). The Human Genome Project and SNP genotyping efforts have caused a substantial rethinking of mapping methodologies and study designs in complex human disease [9, 21, 27, 34]. The testing of large numbers of genotypes or other genetic parameters, such as gene expression profiles for association with one or more traits, raises important statistical issues regarding the appropriate false positive rate of the tests and the level of statistical significance to be adopted given the multiple testing involved [21, 27]. Other important and unresolved issues include the appropriate use of haplotypes (*see* **Haplotype Analysis**) and the modeling of linkage disequilibrium. The required methodologic development in genetic statistics and bioinformatics is nontrivial given the complexity of many common diseases and the genetic databases being collated. Some current areas of methodologic development include haplotype analysis [11, 29, 36], distance-based mapping measures [3, 26], combined linkage and association analyses [13], techniques for modeling linkage disequilibrium and population history [36], and **Markov chain Monte Carlo**-based approaches [20] (*see* **Bioinformatics** for other areas of methodologic development).

Conclusion

The sequencing of the human genome, pursued both by government and industry, is rapidly informing us as to genetic structure and diversity [2]. The availability of a complete reference sequence for the human genome, together with new advances in high-throughput genotyping, functional genomics, chemistry, proteomics and in bioinformatics and biostatistical genetics, will likely accelerate the gene discovery process in complex human disease.

References

[1] Bentley, D.R. (2000). The Human Genome Project – an overview, *Medical Research Review* **20**, 189–196.

[2] Broder, S. & Venter, J.C. (2000). Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium, *Annual Review of Pharmacology and Toxicology* **40**, 97–132.

[3] Collins, A. & Morton, N.E. (1998). Mapping a disease locus by allelic association, *Proceedings of the National Academy of Sciences* **95**, 1741–1745.

[4] Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. & Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998–2003, *Science* **282**, 682–689.

[5] Eberle, M.A. & Kruglyak, L. (2000). An analysis of strategies for discovery of single-nucleotide polymorphisms, *Genetic Epidemiology* **19**, S29–S35.

[6] Gray, I.C., Campbell, D.A. & Spurr, N.K. (2000). Single nucleotide polymorphisms as tools in human genetics, *Human Molecular Genetics* **9**, 2403–2408.

[7] Kurian, K.M., Watson, C.J. & Wyllie, A.H. (1999). DNA chip technology (editorial), *Journal of Pathology* **187**, 267–271.

[8] Landegren, U., Nilsson, M. & Kwok, P.Y. (1998). Reading bits of genetic information: methods for single-nucleotide polymorphism analysis, *Genome Research* **8**, 769–776.

[9] Lander, E. & Schork, N. (1994). Genetic dissection of complex traits, *Science* **265**, 2037–2048.

[10] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome, *Nature* **409**, 860–921.

[11] Li, T., Ball, D., Zhao, J., Murray, R.M., Liu, X., Sham, P.C. & Collier, D.A. (2000). Family-based linkage disequilibrium mapping using SNP marker haplotypes: application to a potential locus for schizophrenia at chromosome 22q11, *Molecular Psychiatry* **5**, 452.

[12] Long, A.D. & Langley, C.H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits, *Genome Research* **9**, 720–731.

[13] MacLean, C.J., Morton, N.E. & Yee, S. (1984). Combined analysis of genetic segregation and linkage under an oligogenic model, *Computers and Biomedical Research* **17**, 471–480.

[14] Marshall, E. (1997). Snipping away at genome patenting (news), *Science* **277**, 1752–1753.

[15] Marshall, E. (1998). A second private genome project (news), *Science* **281**, 1121.

[16] Marshall, A. & Hodgson, J. (1998). DNA chips: an array of possibilities, *Nature and Biotechnology* **16**, 27–31.

[17] Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., et al. (1999). A general approach to single-nucleotide polymorphism discovery, *Nature Genetics* **23**, 452–456.

[18] Martin, E.R., Lai, E.H., Gilbert, J.R., Rogala, A.R., Afshari, A.J., Riley, J., et al. (2000). SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease, *American Journal of Human Genetics* **67**, 383–394.

[19] Masood, E. (1999). As consortium plans free SNP map of human genome (news), *Nature* **398**, 545–546.

[20] Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms, *Genetics* **154**, 931–942.

[21] Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science* **273**, 1516–1517.

[22] Roberts, L. (2000). Human genome research. SNP mappers confront reality and find it daunting (news), *Science* **287**, 1898–1899.

[23] Saegusa, A. (1999). Japan bids to catch up on gene sequencing (news), *Nature* **399**, 96.

[24] Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270**, 467–470.

[25] Schork, N.J., Fallin, D. & Lanchbury, J.S. (2000). Single nucleotide polymorphisms and the future of genetic epidemiology, *Clinical Genetics* **58**, 250–264.

[26] Terwilliger, J.D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci, *American Journal of Human Genetics* **56**, 777–787.

[27] Terwilliger, J.D. & Goring, H.H. (2000). Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design, *Human Biology* **72**, 63–132.

[28] Terwilliger, J.D. & Weiss, K.M. (1998). Linkage disequilibrium mapping of complex disease: fantasy or reality?, *Current Opinion in Biotechnology* **9**, 578–594.

[29] Toivonen, H.T., Onkamo, P., Vasko, K., Ollikainen, V., Sevon, P., Mannila, H., et al. (2000). Data mining applied to linkage disequilibrium mapping, *American Journal of Human Genetics* **67**, 133–145.

[30] van Ommen, G.J., Bakker, E. & den Dunnen, J.T. (1999). The human genome project and the future of diagnostics, treatment, and prevention, *Lancet* **354**, Supplement 1, SI5–SI10.

[31] Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. (1995). Serial analysis of gene expression, *Science* **270**, 484–487.

[32] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., et al. (2001). The sequence of the human genome, *Science* **291**, 1304–1351.

[33] Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, *Science* **280**, 1077–1082.

[34] Weeks, D. & Lathrop, G. (1995). Polygenic disease: methods for mapping complex disease traits, *Trends in Genetics* **11**, 513–519.

[35] Zhao, L.P., Aragaki, C., Hsu, L. & Quiaoit, F. (1998). Mapping of complex traits by single-nucleotide polymorphisms, *American Journal of Human Genetics* **63**, 225–240.

[36] Zollner, S. & von Haeseler, A. (2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms, *American Journal of Human Genetics* **66**, 615–628.

LYLE J. PALMER

Hypergeometric Distribution

Consider a clinical study of five patients A, B, C, D, and E, two of whom are randomly assigned to a new therapy (surgery plus drug) and the remaining three to surgery alone. As healthy skeptics, we wish to test the statement that the patients' fates are unaffected by the new drug. Suppose that patients A and C respond. What is the probability distribution of the number of responders in the group assigned to the new therapy?

We can easily enumerate the possible outcomes (Table 1). There are 10 equally likely assignments of patients to the treatment groups, and under our assumption of predestined fate, the probability of two, one, or no responders in the new treatment group are 10%, 60%, and 30%, respectively.

Enumeration works well when the number of possibilities is small, such as the example in Table 1 of 10 possible assignments. However, as shown later, there is a rapid increase in the number of possible assignments with a modest increase in the scope of the number of patients.

The Hypergeometric Distribution

Consider a population of N patients, among whom A are "responders" and $B = N - A$ are "failures". Suppose we select a random sample of n patients. (That is, any subset of n patients has an equal chance of being the actual sample.) What is the probability distribution of the number of responders in the sample? The answer defines the *hypergeometric distribution*.

Combinatorial Considerations

Question: How many distinct n letter words can we make from an N letter alphabet, when no letter is repeated?

Table 1 Assignments to new therapy (others to standard)

AC (two responders in new therapy):	1 way
AB, AD, AE, BC, CD, CE (one responder):	6 ways
BD, BE, DE (no responders):	3 ways

Answer:

$$\prod_{i=0}^{n-1}(N-i) = N(N-1)(N-2)\dots(N-n+1). \tag{1}$$

The first letter can be selected N ways. For each of these, the second can be chosen in $(N-1)$ ways. Hence there are $N(N-1)$ two-letter words. For each of these two-letter words there are $(N-2)$ ways to select the third letter, or $N(N-1)(N-2)$ three-letter words. The general formula follows inductively.

Question: How many ways can we select n distinct letters from an N letter alphabet, if order is unimportant?

Answer:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \left[\prod_{i=0}^{n-1}(N-i)\right] \Big/ n!, \tag{2}$$

where, by definition, $r! = r(r-1)(r-2)\dots 1$ and $0! = 1$. [We define $\binom{N}{n}$ as zero if $n < 0$ or $n > N$.]

By applying (1) with $N = n$, there are $n!$ ways to arrange each collection of n distinct letters into words. Hence, the number of distinct selections of n letters from an N letter alphabet is the number of n letter words per (1) divided by $n!$ This gives us the right-most result in (2). By multiplying the numerator and denominator of the right-most part of (2) by $(N-n)!$, we obtain the middle expression of (2).

Hypergeometric Probability Function

The solution of the original question posed is defined as the probability of observing x responses in a random sample of n patients from a population containing A responders and $N - A$ nonresponders:

$$h(x; n, A, N) = \binom{A}{x}\binom{N-A}{n-x} \Big/ \binom{N}{n} \tag{3}$$

$$= \binom{n}{x}\binom{N-n}{A-x} \Big/ \binom{N}{A}. \tag{4}$$

Note that (4) tells us that the roles of n and A are interchangeable.

To derive (3), note that from (2), there are $\binom{N}{n}$ possible samples. Also from (2), we can select x responders from the population of A responders in $\binom{A}{x}$ ways and for each of those, we can complete the sample by selecting the nonresponders in $\binom{N-A}{n-x}$ ways. Hence, the numerator of (3) represents the number of possible samples with exactly x responders.

To obtain (4) from (3), one can simply replace the combinatorial terms by factorials [middle part of (2)].

In the example posed in the Introduction, $N = 5$ patients, $A = 2$ responders, and $n = 2$ sampled in the experimental treatment. From (3):

$$h(2; 2, 2, 5) = \frac{1}{10},$$
$$h(1; 2, 2, 5) = \frac{6}{10},$$
$$h(0; 2, 2, 5) = \frac{3}{10}.$$

If $N = 20$ and $n = 10$, then there are 184 756 possible samples. Enumeration of the possible samples, as in the case where $N = 5$ and $n = 2$, would quickly become too time-consuming.

Properties of the Hypergeometric Distribution

Property 1. By straightforward algebra applied to (3):

$$h(x+1; n, A, N) = h(x; n, A, N)\left[\frac{(n-x)(A-x)}{(x+1)(N-A-n+x+1)}\right]. \quad (5)$$

Property 2. Since from (5), the term in square brackets decreases with increasing x, the distribution is "unimodal" and has its mode at one (or both) of the integer values of x surrounding the x value where the term in square brackets is equal to unity. That is, the mode is adjacent to or equal to (if an integer):

$$x = \frac{(A+1)(n+1)}{N+2} - 1. \quad (6)$$

A "unimodal" discrete distribution over a finite set of integers $0, 1, \ldots, K$ has probabilities that behave in one of the following ways: (i) increase to a peak and then decrease; (ii) have a peak at zero and decrease; or (iii) increase from zero to its peak at the highest possible x, K. Looking at increasing values of the random variable, it cannot show an increase in probability to the right of a decrease.

Property 3. The mean of the hypergeometric distribution, μ, is

$$\mu = \frac{nA}{N}. \quad (7)$$

Property 4. The variance of the hypergeometric distribution, σ^2, is

$$\sigma^2 = \frac{n(N-n)A(N-A)}{N^2(N-1)}. \quad (8)$$

Property 5. Although the mean and mode do not seem to be related, it can be shown that the mean is always larger than the value defined by x in (6), and the mean is always within one unit of the value of x defined in (6). Thus, the mode must occur close to the mean.

Approximations

The approximations below can be "proven" by limit theory. The astute question, however, is: How large must the various quantities be before the approximation works to our satisfaction? Hence, rather than limit theory, we use exhaustive computer searches to explore the accuracy. The demonstrations are convincing in terms of closeness over a broad range of applications.

We approximate the cumulative distribution, which has exact value:

$$\Pr[X \le x] = H(x; n, A, N) = \sum_{j=0}^{x} h(j; n, A, N), \quad (9)$$

where h is defined in (3).

Binomial Approximation

If the population, N, is "large", and the sample size, n, is a "small" fraction of the smaller of A (responders) and $(N - A)$ (failures), then the cumulative distribution satisfies the following approximation:

$$H(x; n, A, N) \cong \sum_{j=0}^{x} \binom{N}{j} p^j (1-p)^{n-j}, \quad (10)$$

where $p = A/N$.

The right-hand side of (10) is the cumulative **binomial distribution**.

Because of the "drop in the bucket" effect, successive trials are close to independent. Sampling without replacement (hypergeometric) is similar to sampling with replacement (binomial). (*see* **Sampling With and Without Replacement**).

Reality Check. We studied each of the 432.9 million binomial approximations where $100 \leq A \leq 1000$, $100 \leq N - A \leq 1000$, $n \leq 0.1A$ and $n \leq 0.1(N - A)$, and $x = 0, 1, \ldots, n$. The largest deviation between the cumulative distributions occurred where $A = 100$, $N - A = 100$, $n = 10$, and $x = 3$. Eq. (9) gave an exact value of 0.1656, while (10) gave an approximate value of 0.1719, for a difference of 0.0063.

Normal Approximation

Let us define the "smallest expected value" as

$$EC = \min\left\{\frac{An}{N}, \frac{(N-A)n}{N}, \frac{A(N-n)}{N}, \frac{(N-A)(N-n)}{N}\right\}, \tag{11}$$

the smallest of the four expectations obtained per (7), interchanging the symmetric roles of n vs. $N - n$ and A vs. $N - A$.

If EC is "large", then

$$H(x; n, A, N) \cong \Phi\left(\frac{(x-\mu)}{\sigma}\right), \tag{12}$$

where Φ is the **standard normal** cumulative distribution. The values μ and σ are defined in (7) and (8).

Since the hypergeometric distribution is discrete, Yates [8] suggested that a better approximation might result by using the following, noting that all the probabilities occur at integer values:

$$\Pr[X \leq x + 0.5] = \Pr[X \leq x] \text{ for } x = \text{integer in the hypergeometric distribution.}$$

The basic idea is to approximate the discrete probability that the hypergeometric variable is equal to an integer, by the normal probability of falling within ± 0.5 of the integer (*see* **Yates's Continuity Correction**).

The "corrected approximation" is

$$H(x; n, A, N) \cong \Phi\left(\frac{(x-\mu+0.5)}{\sigma}\right). \tag{13}$$

A rule of thumb, supposedly attributed to R.A. Fisher, claims that values of EC as low as 5 give satisfactory results.

Reality check of (13). We ran a computer check of all situations where $N \leq 250$ and $EC \geq 5$. The largest deviation between (13) and (9) occurred where $N = 250$, $A = 36$, $n = 37$, and $x = 5$. The exact probability from (9) is 0.5517, while the corrected normal approximation, (13), yielded 0.5347, for a deviation of 0.0170. The largest deviation in the "tail" (where the cumulative probability was small), occurred at $N = 245$, $A = 35$, $n = 35$, and $x = 1$. The exact probability is 0.0230, while the corrected normal approximation yields 0.0342 for a deviation of 0.0112. The computer routine compared over 19 million contingencies.

We also ran a computer check of all situations where $N \leq 250$, $EC \geq 7$, and the cumulative hypergeometric probability is below 20%. The largest deviation here was 0.0080, which occurred at $N = 250$, $A = 42$, $n = 42$, and $x = 3$. The exact value is 0.0462, while the corrected normal approximation is 0.0542.

The uncorrected normal approximation may be much more unreliable in the tails, where it is most important. For example, if $N = 50$, $A = 16$, $N = 17$, and $x = 3 (EC > 5)$, then the exact hypergeometric cumulative probability (9), is 0.1056, while the uncorrected normal approximation, (12), is only 0.0611, a deviation of 0.0445.

For more information on other approximations, see [5] and [2].

Final Commentary

The reader may wonder why we introduced approximations in an era when exact calculations are routinely available on computers. Ironically, the very hardware and software that allowed us to investigate the adequacy of the approximations, in the most extensive study yet conducted, are the very same tools

that allow us to use exact methods for every application. However, these approximations have been used in countless past research projects, and will continue to be employed by others. The investigation in this article indicates that the ***P* values** reported in these articles using binomial or corrected normal approximations are reasonably accurate, and that the inferences are qualitatively correct, provided that one is not an all-or-none type inference maker, based on a P value of 5% or 1%. Haber [2] shows that if the goodness criterion is a ratio of probabilities, then in the tails and with low expected numbers, the Yates correction was perceived to perform relatively poorly. This should not dissuade users.

In 1990, *Statistics in Medicine* devoted considerable coverage to the Yates correction (see [3] and the ensuing discussion). Shuster [6] used the binomial approximation for the analysis of **clinical trials** where the sample size is large but the events are rare. In effect, he interchanged the roles of A and n, as noted below (4).

Suissa & Shuster [7] relied heavily upon the hypergeometric distribution when they derived sample size requirements for clinical trials involving two independent samples. Their exact unconditional methods require fewer patients than a corresponding **Fisher's exact test**, when type I error and power are prespecified.

The term hypergeometric distribution [1] is based on the connection with the "hypergeometric series" defined by Euler in 1769. His series produced as special cases the geometric series which he was generalizing, and a polynomial whose coefficients are constant multipliers of the hypergeometric probabilities. For further details, see [4].

Generalization

The idea of the hypergeometric distribution can be extended to a multivariate setting as follows. Suppose a population contains N_j subjects of type j, $j = 1, 2, \ldots, J$. Suppose we wish to partition this population randomly into subgroups of size M_i, $i = 1, 2, \ldots, I$. Let X_{ij} be the number of subjects of type j in subgroup i. Then

$$\Pr\left[X_{ij} = x_{ij} : 1 \le i \le I, 1 \le j \le J\right] = \prod_{i=1}^{I} M_i! \prod_{j=1}^{J} N_j! \left\{ N! \prod_{j=1}^{J} \prod_{i=1}^{I} x_{ij}! \right\}^{-1}, \quad (14)$$

where

$$N = \sum_{i=1}^{I} M_i = \sum_{j=1}^{J} N_j.$$

If $I = J = 2$, $N_1 = A$, $N_2 = N - A$, $M_1 = n$, and $M_2 = N - n$, then (14) reduces to (3), the hypergeometric. The marginal distribution of each X_{ij} is hypergeometric with $A = N_j$ and $n = M_i$.

Finally, note that for the fixed constants N_j and M_i, the probability given in (14) is inversely proportional to

$$\prod_{j=1}^{J} \prod_{i=1}^{I} x_{ij}!.$$

This fact drives computer programs dedicated to the exact analysis of two-dimensional **contingency tables** (*see* **Exact Inference for Categorical Data**).

The same concept can be extended to multidimensional situations with probabilities inversely proportional to the product of the factorials of the individual cell counts.

Acknowledgments

The author wishes to thank Professors James Kepner, P.V. Rao, Andrew Rosalsky, Andre Khuri, and Instructor Maria Ripol for helpful material.

References

[1] Guenther, W.C. (1983). Hypergeometric distributions, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 707–712.

[2] Haber, M. (1980). A comparison of some continuity corrections for the chi-squared test on 2×2 tables, *Journal of the American Statistical Association* **75**, 510–515.

[3] Haviland, M.G. (1990). Yates correction for continuity and the analysis of 2×2 contingency tables (with discussion), *Statistics in Medicine* **9**, 363–367.

[4] Larsen, R.J. & Marx, M.L. (1985). *An Introduction to Probability and its Applications*. Prentice-Hall, New York, Chapter 3.

[5] Patel, J.K. & Read, C.B. (1982). *Handbook of the Normal Distribution*. Marcel Dekker, New York, Chapter 7.

[6] Shuster, J.J. (1993). Fixing the number of events in large comparative trials with low event rates: a binomial approach, *Controlled Clinical Trials* **14**, 198–208.

[7] Suissa, S.S. & Shuster, J.J. (1985). Exact unconditional sample sizes, for the 2×2 binomial trial, *Journal of the Royal Statistical Society, Series A* **148**, 317–327.

[8] Yates, F. (1984). Tests of significance for 2×2 contingency tables, *Journal of the Royal Statistical Society, Series A* **147**, 426–463.

(*See also* **Logistic Regression, Conditional**)

JONATHAN J. SHUSTER

Hypothesis Testing

The global responses of patients with diabetic neuropathy who had been randomly assigned to one of two different treatments (*see* **Randomized Treatment Assignment**) are displayed in Table 1. If the treatments were equally efficacious, we would expect to see the same percentage of patients deteriorating or improving in both. Since the responses of individual patients will differ, even if given the same treatment, the resulting random variation means that we would not expect to see exactly the same percentage in each group. However, note that there are 14 patients with moderate or excellent response to treatment A and only three such patients on treatment B. Could this great a difference have happened "at random"? The basic idea behind hypothesis testing is to compute the **probability** of the pattern of data that we have observed, under the assumption that any differences are "purely random". If that probability is very low, then we would be tempted to reject the hypothesis that the differences between treatments is due to "random noise" alone. If the pattern that is seen also suggests a consistent difference in response between the treatments, we would be even more inclined to reject the hypothesis of equal effect.

In Table 1, around 30% of the patients on treatment A have a moderate or excellent response versus only 6% on treatment B. If we treat these two numbers as coming from independent **binomial** random variables with the same underlying probability of response, the probability of seeing as great or greater a difference is less than 0.001. Yet, the only excellent response was under treatment B. How can we claim that treatment A is better? What would have happened if we had compared only the percentage of patients with an excellent response? Or the percentage of patients with any improvement? Or the percentage of patients who deteriorated? Of all parts of Table 1, the break between moderate improvement or better and all other responses is most favorable to treatment A. Is it acceptable to choose the most favorable part of the data before calculating the probability? What is the "right" way of applying hypothesis tests? Is there an "optimal" method of testing? Questions like this have generated a vast literature of books and articles, ranging from abstract mathematical dissertations, to philosophical discussions of the meaning of probability, to the interpretation of hypothesis tests run on medical, epidemiologic, and other biological data. There are at least two major schools of thought, and how one uses hypothesis tests or interprets them may differ, depending upon which school of thought is being invoked.

Historical Development

The basic idea behind hypothesis testing has been used in many branches of science for at least 200 years. One author [1] claims to have found the germ of the idea in a medical discussion from 1662. Other early references have included astronomical and sociological investigations (see [16]). However, the earliest clearly thought-out use of hypothesis testing probably belongs to **Karl Pearson**. Pearson was collecting biological data from all over the world and attempting to fit these data to specific probability distributions (see Galton et al. [12], for a formal statement of this program). The plan was to show the effects of natural selection and evolution on shifts in these distributions under the pressure of changes in the environment. To determine whether a given distribution fit the data, Pearson ordered the numbers and divided them into bins containing 5–20 adjacent numbers in a bin. He then computed the expected number of observations that he should have seen in each bin and compared the expected number to the observed number. If $O_i =$ the observed number in bin i, and $E_i =$ the expected number in bin i, then the sum

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_i - E_i)^2}{E_i} + \cdots$$

Table 1 Responses of patients with diabetic neuropathy, to two randomly assigned treatments

	Deterioration			Improvement		
	Severe	Slight	No change	Slight	Moderate	Excellent
Treatment A	1	2	20	9	14	0
Treatment B	3	2	26	15	2	1

was used to determine if the fit was good. If this sum was too large, the proposed distribution was rejected. Pearson proved that, regardless of the underlying probability distribution being tested, if the sample size was large enough, this sum had a specific distribution, which he called a **chi-square(d) distribution**. Thus, he was able to test the hypothesis that the data followed a specific random pattern with an omnibus test.

Pearson's proof was not completely rigorous, and his exact calculations were in need of some minor adjustments (derived in [6]). However, his work contains the basic components of any modern hypothesis test:

1. a well defined probability distribution that describes the hypothesis that the differences in pattern are "purely random".
2. A test statistic that can be calculated from the data, which:
   (i) has a distribution that is the same regardless of the definition of "purely random"; and
   (ii) can be used to compute a probability that measures how well the observed data fit the distribution that defines "purely random".

R.A. Fisher, a younger contemporary of Pearson, derived most of the test statistics that we now use, in a series of papers and books during the 1920s and 1930s. Fisher also published a "cook book" of methods to popularize these tests [7], which went through 10 editions. **G.W. Snedecor**, who founded the first statistics department in the US at Iowa State University, published a textbook [20] that spread Fisher's methods and test statistics into even further use. In the 1970s, a review of the *Science Citation Index* showed that Snedecor's textbook was the single most frequently cited paper or book in the scientific literature of the time.

However, there were many questions about how to use these test statistics and which test statistics to use under which circumstances. In the late 1920s, Karl Pearson's son, **Egon Pearson**, approached the young Polish mathematician **Jerzy Neyman** with a question that was bothering him. If you test whether data fit a particular probability distribution, and the test statistic is not large enough to reject that distribution, how do you know that this is the "best" that could be done? How do you know that some other test statistic might not have rejected that probability distribution? The resulting collaboration between Egon Pearson and Neyman over the next few years produced a series of papers that revolutionized the nature of hypothesis testing and introduced some of the basic ideas that now govern this field.

Following on from this work by Neyman and Pearson, Eric Lehmann published a definitive textbook [15] that elaborated on the original Neyman–Pearson formulation. This version of the Neyman–Pearson formulation is the interpretation of hypothesis testing that is usually taught in elementary statistics courses, and it dominates much of the medical and epidemiologic literature, where hypothesis testing is used.

Fisher, the creator of most of our modern methods, was critical of the Neyman–Pearson formulation (see [9]; developed more fully in [11]). He felt that the formulation may have been very nice mathematics, but that it had nothing to do with the way in which hypothesis tests are actually used in scientific investigations. In addition, the statistical literature is filled with other objections to the validity of the Neyman–Pearson formulation in terms of its use of probability and its ability to interpret experimental results (for a survey of this work, see [2]). In general, these objections come from two schools of statistical reasoning. One school follows Fisher's approach and views hypothesis tests as rough tools of **inference** that should be used only in conjunction with other tools (for a full discussion, see [5]). The other school criticizes what they consider to be irrational components of hypothesis testing and proposes that inference should be based on the **likelihood** function. These critics, in turn, fall into two general categories, the Bayesians and those who would rest all inference on likelihood alone (*see* **Bayesian Methods**; **Likelihood**). Bayesian techniques do not make use of hypothesis tests but base their inference on credibility intervals that describe a highly probable range of values for a given parameter, based on prior knowledge and the data (*see* **Prior Distribution**). The likelihood approach to inference also rejects formal hypothesis testing and bases inference on ranges of the parameters that produce relatively high likelihoods for the observed data.

Since Lehmann's definitive text, hypothesis testing has been a fruitful area for statistical research. More recent developments include locally most powerful tests, restricted tests, investigations into the **robustness** of tests, and tests of nested (or **hierarchical**) and nonnested models (*see* **Separate Families of**

Hypotheses). Some of these will be discussed briefly in what follows, but the reader should be aware that this continuing research means that the nature of hypothesis testing and the applications of these techniques will continue to change.

The Neyman–Pearson Formulation

We shall start with a simple model and build on that. Consider two hypotheses about the nature of reality. In Table 1, the two hypotheses might be

H_0: the probability of moderate or better improvement is the same for patients on treatment A as it is on treatment B.
H_1: The probability of moderate or better improvement is twice as great on treatment A as it is on treatment B.

H_0 is called the **"null hypothesis"**. H_1 is called the **"alternative hypothesis"**. We are presented with data from a study and we are asked to make one of two decisions:

D_0: H_0 is true.
D_1: H_1 is true.

This situation can be displayed as a two-by-two table, as shown in Table 2. If the decision matches the true state of nature, there is no error. Otherwise, two types of error are possible. The probability of a type I error (deciding for the alternative hypothesis when the null hypothesis is true) is labeled α. The probability of a type II error (deciding for the null hypothesis when the alternative hypothesis is true) is labeled β.

One "solution" for this simple setup is to consider all possible patterns that the data might have and order them in terms of increasing evidence in favor of the alternative hypothesis. For instance, in comparing two treatments with respect to the frequency of improvements in Table 1, we have a total of 41 $(9 + 14 + 0 + 15 + 2 + 1)$ patients improving.

Table 2 A decision table for choice between two hypotheses

	True state of nature	
Decision	H_0	H_1
D_0	No error	Type II error
D_1	Type I error	No error

A result that would be most favorable to the alternative hypothesis (that treatment A is better than treatment B) would be to assign all 41 to treatment A. The next most favorable would be to assign 40 to treatment A and one to treatment B, etc. Once the possible outcomes are ordered, the analyst can pick a specific outcome (say, a break of 30A and 11B) and calculate the probability, based on the assumption of H_0, for each outcome that was as favorable or more favorable than that specific break point. Similarly, the analyst could compute the probability, based on the assumption of H_1, for each outcome that is less favorable than that specific break point. Let the decision be as follows:

D_1: choose H_1 if the observed outcome is at that break or at one more favorable.
D_0: choose H_0 if the observed outcome is one of the events less favorable than that break.

Then, the sum of the favorable probabilities at that break and beyond, under the null hypothesis, is the probability of a type I error, α. Similarly, the sum of the unfavorable probabilities less than that break, under the alternative hypothesis, is the probability of a type II error, β.

Another approach is to decide in advance what level of α (perhaps 0.05) error the analyst is willing to have (*see* **Level of a Test**). Then, each time the analyst is faced with a decision involving a simple null and a simple alternative hypothesis, the analyst can choose a break point that corresponds to that level of α. Then, in the long run, regardless of the exact problem at hand, the proportion of times the analyst will make a type I error will be α. However, for this to hold, the complete decision process (the choice of α, of the test statistic, and of the cut-point) must be set up in advance of seeing any data and independent of the outcome of a particular trial.

This situation of a simple null and simple alternative hypothesis seldom holds in real life. For instance, in Table 1 we can propose a simple null (that the probability of response is the same for both treatments), but a simple alternative would require that we pick a specific difference in probabilities for each type of response. If we want the alternative hypothesis to be more general, such as that the probability of improvement is greater for A and that the probability of deterioration is less for A, then we have to consider an infinitude of possible differences in probabilities of response. In such a case, the alternative hypothesis

is called a composite hypothesis. However, although the comparison of a simple null and a simple alternative hypothesis seldom occurs in real-life problems, it is a useful first step in visualizing how one might proceed.

To be more realistic, let us consider a simple null hypothesis and a composite hypothesis of alternatives that are farther and farther away from the null. For instance, we might consider the null hypothesis that the probability of improvement is the same for both treatments and the composite alternative that the probability of improvement for A has the relationship

$$p_A = kp_B, \quad k > 1.$$

This includes the simple alternatives that $k = 2$ (probability of improvement for A is twice that of B), $k = 1.001$, $k = 10$, etc. As k increases, the "distance" between the simple null and the alternative increases. (There is a technical problem, here. If k gets large enough, then p_A will be greater than 1, but this can be taken care of by considering **odds** (*see* **Odds Ratio**) rather than probabilities.)

At this point, the Neyman–Pearson formulation has three parameters that govern the decision process: α, β, and δ the latter being the "distance" from the null to a specific simple alternative that is part of the composite alternative. Once they had reached this point in their development, Neyman and Pearson sought an optimum solution to the problem. Some of the critics of this formulation have attached on this attempt to find an optimum solution as one of the inherent problems. This is because one has to define what is meant by optimum, and the act of defining it often limits the nature of the solution. In this case, Neyman realized that there was no single definition of optimum when dealing with three freely ranging parameters. However, it was possible to define an optimum if the problem was constrained. Neyman's solution was the following:

1. fix α;
2. find a decision process that minimizes β for a range of δ-values.

They called $1 - \beta$, the probability of correctly deciding in favor of the alternative hypothesis, the **power** of the test procedure. Thus, in words, the optimum solution is one that fixes the probability of a type I error in advance (at, say, 0.05) and then has the greatest power for a specific range of alternative hypotheses. In this way, the analyst will make type I errors $100\alpha\%$ of the time across the entire spectrum of decisions that use this same α-level. At the same time, the analyst will be testing hypotheses is a way that is most favorable to the set of alternatives that have been chosen as important for each decision.

In such a model, the ideal decision process is one that is more powerful than any other for all possible alternatives. This is the uniformly **most powerful**, or UMP, test. If a UMP test exists, to follow the Neyman–Pearson formulation, the analyst should always use it. Unfortunately, as Neyman noted in the final paper that he wrote in this series [17], UMP tests seldom exist. In particular, they do not exist for the types of hypothesis tests most often used in medical and epidemiologic research.

There are two ways of overcoming this problem. The analyst can narrow the class of alternatives, or the analyst can seek the best decision process from a collection of decision processes that are constrained in some way. The first method (narrowing the class of alternatives) occurs with the use of restricted hypothesis tests and locally most powerful hypothesis tests (for a more complete discussion, see [18]). The second approach is done by requiring that the hypothesis tests have certain properties. Some of these properties are:

1. *unbiasedness* (the probability of a type II error is never less than α);
2. *symmetry* (the test statistic should produce the same value if the data are permuted – the nature of the permutation defining the type of symmetry);
3. *invariance* (the test statistic should produce the same value if all the data are subjected to a specific monotone transformation, such as multiplication by a constant, and an appropriate transformation is applied to the parameter value).

This has led to a plethora of terms to describe different types of optimum tests, such as UMP in the class of unbiased tests. Whether such tests are "optimum" for a given situation should depend upon whether the restriction of the alternative hypotheses or the properties defining the class of tests are

appropriate to that situation. Just because a test has a nice name (such as "exact") does not mean that it is the "best" to use.

Criticisms of the Neyman–Pearson Formulation

The major criticisms of this formulation for hypothesis testing are three-fold: (i) the validity of the α-level depends upon the definition of probability as the long-run frequency of errors that might occur; (ii) the computation of the significance level uses the probabilities of events that have not been observed; and (iii) the definition of optimum is purely arbitrary.

The Use of the Long-Run Frequency of Errors to Define α

Fisher [9] was one of the first to object on these grounds. Fisher pointed out that long-run frequency of error is a concept appropriate to quality control, where an inspector wants to be sure than no more than $100\alpha\%$ of defective products will pass. However, scientific investigation, said Fisher, is a process which involves a sequence of experiments, in which the conditions of each experiment are dependent upon the outcome of previous experiments. G.E.P. Box [3] added to this description by noting that the data from previous experiments are often reexamined in the light of later results. To Fisher, the fact that the analyst uses a cutoff significance level of 0.05 does not mean that the analyst will be wrong 5% of the time. For, according to Fisher, the analyst has no right to declare something is so until he can design a study that will invariably produce a significant result in favor of it.

There is a further problem with the Neyman–Pearson formulation whenever the observed ***P* value** is less than or equal to α. It does not allow the analyst to make any other decision. Thus, if α is set at 0.05, a P value of 0.04999 has the same interpretation as a P value of 0.00001. In Neyman–Pearson hypothesis testing, there is no such thing as "more significant", and the use of symbols such as * (for $P \leq 0.05$), ** (for $P \leq 0.01$), and *** (for $P \leq 0.001$) has no meaning. Attempts have been made to develop a theory of "evidence" that will allow for multiple decisions within this frequentist definition of probability. However, all have failed. (For a complete description of this problem, see [14].)

The Use of Events not Observed to Compute P Values

The power of an hypothesis test depends upon its ability to reject the null for events that are more favorable to the alternative. However logical the Neyman–Pearson development may seem, critics point out that it ends up with a counterintuitive procedure. Why should outcomes more extreme than the one observed play any role in the decision process? These critics point out that the only reasonable computations involve the likelihood of the observations under the null hypothesis and the likelihood under the alternative hypothesis. The ratio of these likelihoods should be used to compare the hypotheses with respect to the data. This ratio is called the Bayes factor in the development of Bayesian statistical procedures (*see* **Bayesian Methods**).

The Definition of Optimum is Arbitrary

Neyman was faced with a three-parameter problem: α (the probability of a type I error), β (the probability of a type II error), and δ (the "distance" between the null and alternative hypotheses). His definition of optimum was to fix α and minimize β over a range of δ. Other definitions are possible. One could minimize the sum $\alpha + \beta$ or the odds of α over β. The major justification for Neyman's definition of optimum is that the resulting decision process mimics closely the way in which prior workers such as Fisher had used hypothesis tests. However, there is no reason why fixing α is appropriate for medical or epidemiologic problems. In fact, some critics have asked sarcastically why the analyst's long-run probability of error should have anything to do with whether treatment A is a life-saving procedure that should be used in medicine. One alternative, called sequential Bayes, proposes that there is a finite number of patients who will be treated (*see* **Adaptive and Dynamic Methods of Treatment Assignment**). Some of those will be treated in a controlled trial that compares treatment A to treatment B. Once the trial is over, all of the remaining patients will be given the treatment that has been declared better. The criterion

proposed is to minimize the number of patients on the poorer treatment.

Cox's Formulation of Significance Testing

Many authors have agreed with Fisher's objections and proposed alternative approaches to hypothesis tests. Like Fisher, these authors would treat P values as rough tools for inspecting data. A full development of this idea is due to D.R. Cox [5]. To distinguish between his formulation and that of Neyman and Pearson, Cox called the informal use of P values "significance testing" (as opposed to hypothesis testing). Cox would have the analyst compute the P value of a test statistic but treat it as one of many descriptors of the data. If the experiment was a difficult one to duplicate, or if the data from an epidemiologic study were difficult to accumulate, then the analyst should consider higher P values as "significant". If alternative experiments were easily and cheaply done, then the analyst should require lower P values before taking any decision in favor of an alternative hypothesis. In Cox's view, the cutoff P value is not set in advance, but is dependent upon the importance of the question, the ease of replication, and, to some extent, the data themselves. At all times, Cox warned, the evidence presented by a small P value should be part of a more general analysis of the data that pays attention to the estimated mean effects and to the plausibility of the results.

Cox claimed that there are two general ways in which significance tests and P values are used in scientific research. He called these "hypothesis dividing" and "hypothesis refining". In the hypothesis dividing mode, the scientist proposes two distinctly different hypotheses as explanations of reality. The scientist constructs an experiment or an observation that will lean one way for one hypothesis or the other way for the other hypothesis. The significance test is used to determine if there is enough information in the data to allow for a decision between the two hypotheses. The significance test is not necessarily used to decide in favor of one hypothesis or the other. That decision depends upon the design of the study and the nature of the data. The significance test is used only to discard certain studies as not providing enough information. (This echoes Fisher's view of significance tests.) In the hypothesis refining mode, the scientist has a complicated model of reality, involving many parameters, and he or she wishes to eliminate some of those parameters as having minor or negligible effects. Suppose, for instance, that we have been following a cohort of individuals for many years and wish to determine which baseline characteristics were predictive of some future event (such as a myocardial infarction). We might run a **logistic regression** using all the baseline variables that were collected and use significance tests to eliminate those that do not have a "significant" slope (*see* **Variable Selection**).

The Meaning of *P* Values

The P value of a test statistic is computed as the probability of a **critical region** of possible observations under the null hypothesis. However, it is difficult to define what that means in real life. In fact, the whole problem of linking the mathematical theory of probability to real life is a controversial one. Neyman finessed the problem by fixing the α-level and defining its meaning as the long-run proportion of times that an analyst will make a type I error. However, one cannot use this formulation to justify the statement that we are "$100(1-\alpha)\%$ sure" that the null is false. Nor does this definition make sense if a hypothesis test is going to be applied to a nonreplicated event such as a definitive placebo controled clinical study of a potentially life-saving treatment (since, once the null hypothesis has been rejected, it is unethical to do another study).

In the mathematical theory of **probability**, we propose that there exists a space of "events". Probability is a measure (similar to length or area) on that space of events. The link between mathematical probability and real life is how we define that space of events. **W.S. Gossett** (who wrote under the pseudonym "Student") applied probability theory to the outcome of experiments and said that the space of events was the set of all possible outcomes of such an experiment. But, since only one outcome is seen, this is not a well-defined idea. In sample survey theory, the population is fixed and a sample of that population is chosen at random. The space of events is, then, the set of all possible **random samples** that might have been chosen. Since the **randomization** mechanism is known, this space is well defined. The uncertainty described by probability theory for sample surveys is not about the characteristics of the population (which are fixed)

but about the estimates of those characteristics that are derived from the sample. In a **case–control study** the concept of a sample and population remains, but the calculated probabilities are the **conditional probabilities** that a person has a prior condition (e.g. a heavy coffee drinker), given that the person has the disease (for an excellent discussion of this, see [4]).

To justify the use of probability theory in controlled experiments, Fisher noted that the act of randomly assigning experimental units to treatments in a randomized controlled experiment generated a space that consisted of all possible random assignments. He was able to show that the classical distributions of test statistics that he had developed are approximations of the permutation probabilities that would result (*see* **Randomization Tests**). However, this meant that hypothesis tests were valid only in the framework of a randomized controlled study (for a discussion of some of the consequences of this view, *see* **Intention to Treat Analysis**). Fisher objected to the observational studies connecting smoking with cardiovascular disease and cancer [10], because they used hypothesis tests to "prove" the case. Following the same logic, Fisher would have objected to the use of all hypothesis tests in epidemiology, in case–control studies, or in any type of clinical study that did not involve randomized assignment to treatment.

It is tempting to use probability statements to describe how "sure" one is about the results of the investigation. This would result in phrases such as "The probability that coffee is an important factor in the development of pancreatic tumors is less than 10%". However, the only way this can be done is to describe a space of events that is either related to the state of mind of the observer (personal or **subjective probability**; for a complete discussion, see [19]), or to a general opinion that one might expect in a community of knowledgeable scholars (an idea developed in [13]). All of these fall under the heading of Bayesian statistical methods (*see* **Bayesian Methods**).

Power and the Acceptance of the Null Hypothesis

An important element of the Neyman–Pearson formulation of hypothesis testing is the concept of power. The quality of a statistical test is determined by its power. One could construct a great many test statistics from the data in Table 1. A chi-square test for the independence of the rows and columns, for instance, can be computed based on the null hypothesis that the row (treatment) has nothing to do with the columns (responses) (*see* **Chi-square Tests**). The P value for this test is greater than 0.50. On the other hand, one can use a Cochran–Armitage test (which is a restricted test) that concentrates on the class of alternatives where the probability of a patient's being in treatment A increases with the response (*see* **Trend Test for Counts and Proportions**). The P value for the Cochran–Armitage test is 0.04. The theory of restricted tests says that, if we believe that the only viable alternatives are that the better treatment is consistently better across all the possible responses, then the Cochran–Armitage test is more powerful and should be used rather than the chi-square test.

If we do not pay attention to power, then any test would be equally as "good". A *reductio ad absurdum* of this is to ignore the data from a study and pick a number from a table of **uniformly distributed** random numbers between zero and one. If the number is less than α, declare significance. Such a test is "exact". It also protects the α-level. But, the power is also exactly equal to α (since, regardless of the alternative hypothesis, it will reject the null $100\alpha\%$ of the time.)

In spite of the fact that the Neyman–Pearson formulation involves a decision to accept the null hypothesis, there is a general consensus among statisticians that hypothesis tests are not really designed for that. To quote Fisher [8],

> ... tests of significance (are) ... cogent for the rejection of hypotheses, but...by no means cogent for their acceptance...the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning.

To deal with this problem, some have advocated that articles which describe the results of studies include information about the power of the study to detect a meaningful degree of effect. Alternatively, it has been urged that studies which result in a finding of no significant difference should include **confidence intervals** on the differences in effect that would be reasonable from the data. If the power of the study is inadequate to detect a meaningful effect or if the

confidence interval contains meaningful differences in effect, then the study is inadequate to accept the null hypothesis (*see* **Clinical Significance Versus Statistical Significance**).

References

[1] Armitage, P. (1983). Trials and errors: the emergence of clinical statistics, *Journal of the Royal Statistical Society, Series A* **146**, 321–334.

[2] Berger, J. (1983). The frequentist viewpoint and conditioning, in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. 1, L.M. LeCam & R.A. Olshen, eds. Wadsworth, Monterey.

[3] Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modeling and robustness, *Journal of the Royal Statistical Society, Series A* **143**, 383–340.

[4] Cornfield, J. (1954). Statistical relationships and proof in medicine, *American Statistician* **8**, 19–23.

[5] Cox, D.R. (1977). The role of significance tests, *Scandinavian Journal of Statistics* **4**, 49–70.

[6] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, Chapter 30.

[7] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[8] Fisher, R.A. (1935). Statistical tests, *Nature* **136**, 474–475.

[9] Fisher, R.A. (1955). Statistical methods and scientific inference, *Journal of the Royal Statistical Society, Series B* **17**, 69–78.

[10] Fisher, R.A. (1958). Cigarettes, cancer, and statistics, *Centennial Review* **2**, 151–166.

[11] Fisher, R.A. (1959). *Statistical Methods and Scientific Inference*. Oliver & Boyd, Edinburgh.

[12] Galton, F., Pearson, K. & Weldon, R. (1898). Charge of this journal, *Biometrika* **1**, 1.

[13] Keynes, J.M. (1921). *A Treatise on Probability*. Macmillan, London.

[14] Kiefer, J. (1976). Admissibility of conditional confidence procedures, *Annals of Statistics* **4**, 836–865.

[15] Lehmann, E. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

[16] Moroney, M.J. (1951). *Facts from Figures*. Penguin, Harmondsworth, Middlesex, Chapter 15.

[17] Neyman, J. (1935). Sur la verification des hypotheses statistiques composées, *Bulletin de la Société Mathematique de France* **63**, 246–266; *Statistics* **4**, 49–70.

[18] Salsburg, D.S. (1992). *The Use of Restricted Significance Tests in Clinical Trials*. Springer-Verlag, New York.

[19] Savage, L.J. (1954). *The Foundations of Statistics*. Wiley, New York.

[20] Snedecor, G.W. (1940). *Statistical Methods*. Iowa State University Press, Ames.

DAVID SALSBURG

Identifiability

Consider a vector $\mathbf{Y}$ of random variables having a distribution $F(\mathbf{y};\boldsymbol{\theta})$ that depends on an unknown parameter vector $\boldsymbol{\theta}$. $\boldsymbol{\theta}$ is *identifiable* by observation of $\mathbf{Y}$ if distinct values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ for $\boldsymbol{\theta}$ yield distinct distributions for $\mathbf{Y}$, that is, if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ implies $F(\mathbf{y};\boldsymbol{\theta}_1) \neq F(\mathbf{y};\boldsymbol{\theta}_2)$ for some $\mathbf{y}$ [1]. A function $g(\boldsymbol{\theta})$ is *identifiable* by observation of $\mathbf{Y}$ if $g(\boldsymbol{\theta}_1) \neq g(\boldsymbol{\theta}_2)$ implies $F(y;\boldsymbol{\theta}_1) \neq F(y;\boldsymbol{\theta}_2)$ for some $\mathbf{y}$. Note that $\boldsymbol{\theta}$ is identifiable if and only if all functions of $\boldsymbol{\theta}$ are identifiable.

There is some variation in the definition of identifiability, the preceding being the most general. Variants typically employ the density $f(\mathbf{y};\boldsymbol{\theta})$ or the expectation $\mathrm{E}(\mathbf{Y};\boldsymbol{\theta})$ in place of the distribution; the latter variants may explicitly involve a design matrix X of regressors; for example, $\mathrm{E}(\mathbf{Y}; X, \boldsymbol{\theta})$. The basic concept, however, is that $\boldsymbol{\theta}$ [or $g(\boldsymbol{\theta})$] is a function of the $\mathbf{Y}$ distribution, and hence observations of realizations of $\mathbf{Y}$ can be used to discriminate among distinct values of $\boldsymbol{\theta}$ [or $g(\boldsymbol{\theta})$].

The term *estimable* is sometimes used as a synonym for identifiable, but is also used in more specific ways, especially in the context of linear models. For example, Scheffé [4] defines a linear function $\mathbf{c}'\boldsymbol{\theta}$ of $\boldsymbol{\theta}$ to be estimable if there exists an unbiased estimator of $\mathbf{c}'\boldsymbol{\theta}$ that is a linear function of the observed realizations of $\mathbf{Y}$. This property has also been referred to as linear estimability. In epidemiology, estimability of $g(\boldsymbol{\theta})$ is sometimes used to mean that $g(\boldsymbol{\theta})$ can be consistently estimated from observed realizations of $\mathbf{Y}$. Several other definitions have been given; see, for example, [2, 3] and [5].

References

[1] Bickel, P.J. & Doksum, K.A. (1977). *Mathematical Statistics*. Holden–Day, Oakland.

[2] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[3] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

[4] Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

[5] Seber, G.A.F. & Wild, C.J. (1989). *Nonlinear Regression*. Wiley, New York.

(*See also* **General Linear Model**)

SANDER GREENLAND

Identity Coefficients

Identity (k-, or kinship) coefficients were introduced by Cotterman [3], Malécot [10, 11], and Gillois [7] to answer questions of the following sort: if an individual X, is of **genotype** Aa, what is the probability that X's relative, Y, is aa? To answer a question of this kind efficiently, it is necessary to partition the problem into two parts: (i) a measure of the relationship connecting X and Y, and (ii) genotype probabilities conditioned on the relationship. The first part depends upon the concept of "identity by descent" (ibd); the second on assumptions about the mating system in the population.

The Concept of Identity by Descent

In Figure 1, Z is the offspring of X and Y, and X is the offspring of V and W. The transmitted gametes are labeled a, b, c, and d. It is clear from a consideration of Mendelism (*see* **Mendel's Laws**) that the **gene** a is an immediate replicate of either c or d, but not both. Let R denote the relation "is an immediate replicate of" and Pr denote probability, then

$$\Pr(\mathrm{a}R\mathrm{c}) + \Pr(\mathrm{a}R\,\mathrm{d}) = 1 \quad \text{and} \quad \Pr(\mathrm{a}R\mathrm{c}, \mathrm{a}R\,\mathrm{d}) = 0.$$

The following relations are defined in terms of the fundamental relation, R, where x, y, z, and z' are arbitrary genes at a single locus:

$R^0 =$ the identity relation
(a gene is identical to itself),

$$R^2 = [(x, y) : \exists\ z\ (xRz, zRy)],$$

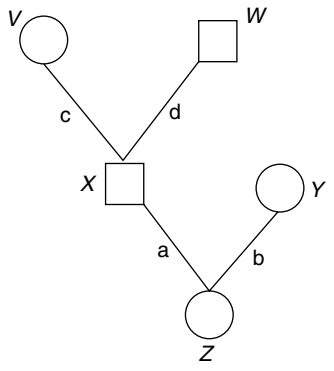

Figure 1 The concept of identity by descent

$$R^3 = [(x, y) : \exists\ z\ \exists\ z'(xRz, zRz', z'Ry)],$$

$$R^n = [(x, y) : \exists\ z\ \exists\ z'(xRz, zR^{n-2}z', z'Ry)].$$

These are the powers of the relation, R. For example, zR^2y means that x is the immediate replicate of a gene which is itself the immediate replicate of y. If we let $R^{\mathrm{U}} = R^0 \cup R \cup R^2 \cup R^3 \cup \ldots \cup R^n$, the relation of identity by descent, I, is defined as

$$I = \left[(x, y) : \exists\ z\ \left((xR^{\mathrm{U}}y) \cup (yR^{\mathrm{U}}x) \cup (xR^{\mathrm{U}}z, yR^{\mathrm{U}}z)\right)\right].$$

In practice, instead of writing xIy, it is customary to write $x \equiv y$ to mean x is identical by descent (ibd) to y. This definition is simply a restatement in set-theoretic notation of Cotterman's original definition for which he used the term "derivative". Cotterman [3] states: "...derivative genes are genes which are relatively recently descended one from the other or both from some common gene". The qualification introduced by the phrase "relatively recently" is included in the definition presented above by restricting the size of n in the definition of R^{U}.

Arbitrary Relationships

Consider two related diploid individuals, X and Y. Label the genes of X a_1 and a_2, and the genes of Y a_3 and a_4. There are 15 "identity by descent" relations among the four genes, as shown in Figure 2. A line connecting two genes denotes that those genes are ibd, the lack of a line denotes that they are not ibd. Some of the events are combined; for example, ${}_{00}k_1$ is the probability of the union of two events. The notation is mnemonic: the first prescript is 0 if X is allozygous (i.e. has nonibd genes at the locus) and 1 if X is autozygous (i.e. has ibd genes at the locus), the second prescript refers to Y in the same way, and the postscript indicates the number of genes X and Y share in common, that is, the number that are ibd. An alternative notation [9] is: ${}_{11}k_2 = \Delta_1$, ${}_{11}k_0 = \Delta_2$, $2_{10}k_1 = \Delta_3$, ${}_{10}k_0 = \Delta_4$, $2_{01}k_1 = \Delta_5$, ${}_{01}k_0 = \Delta_6$, ${}_{00}k_2 = \Delta_7$, $2_{00}k_1 = \Delta_8$, ${}_{00}k_0 = \Delta_9$.

As an example, the k-coefficients for X and Y in Figure 3 are as shown. All coefficients with a leading prescript equal to one are zero because X is assumed to be allozygous. The probability that Y is autozygous and X and Y share no gene in common is ${}_{01}k_0 = 2/32$, because this can happen only if both gametes

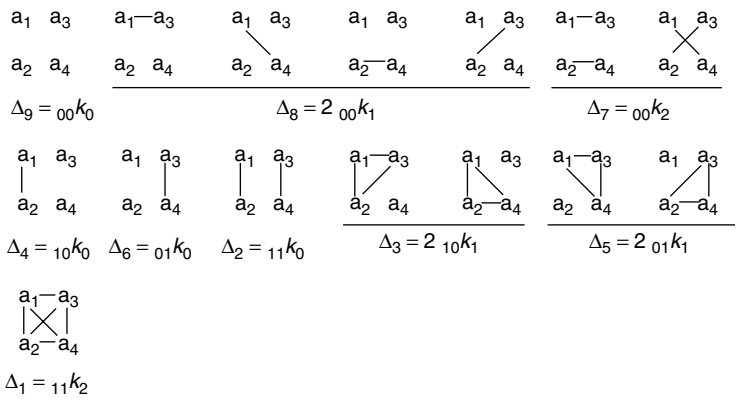

Figure 2 The 15 ibd relationships between two individuals, X and Y, and their probabilities (k-coefficients). (The genes of X are labeled a_1 and a_2, those of Y, a_3 and a_4)

that form Y are derived from the gene in X's daughter that did not come from X; the probability of this is $(1/2)^4$. The probability that both X and Y are allozygous and they share no gene in common is ${}_{00}k_0 = (1/2)^3 + (1/2)^3 + (1/2)^4 = 10/32$, and so on. In this pedigree, the k-coefficients can be calculated from a simple application of basic Mendelism. In general, the nine k-coefficients are more difficult to calculate except in very simple pedigrees [5, 12, 14].

Genotype Pair Probabilities for Pairs of Relatives

The joint genotype distribution for a two-allelic locus in terms of the k-coefficients and the gene frequencies, p and q, of A and a, respectively, is shown in Table 1. For example, the probability that X and Y are both of genotype Aa is, according to the table, $\Pr(X = \text{Aa}, Y = \text{Aa}) = 2_{00}k_2 pq + 2_{00}k_1 pq + 4_{00}k_0 p^2 q^2$.

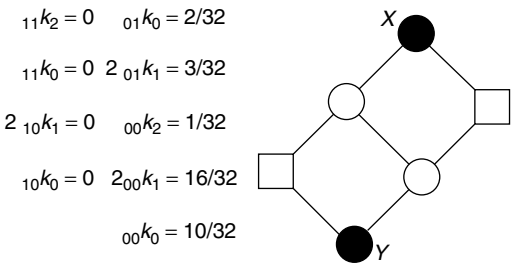

Figure 3 A sample pedigree: calculating the k-coefficients

Noninbred Relatives

Often, one does not need the full set of k-coefficients because many relationships do not involve inbreeding. When neither X nor Y is inbred (when each has an **inbreeding** coefficient, $F = 0$), then all the k-coefficients are zero except for ${}_{00}k_2$, $2_{00}k_1$, and ${}_{00}k_0$.

The k-coefficients for a number of simple noninbred relationships are shown in Table 2. Bilineal relations are those for which ${}_{00}k_2 > 0$; unilineal relationships have ${}_{00}k_2 = 0$.

To calculate the k-coefficients for any noninbred relationship is straightforward. Let a and b be the gametes that form X and c and d those that form Y. Let f_{ac} be the probability that a and c are ibd, f_{ad} be the probability that a and d are ibd, and so on. Let F_{XY} be the probability that a random gamete from X is identical to a random gamete from Y, i.e. the inbreeding coefficient of a, perhaps hypothetical, offspring of X and Y. Then

$$k_2 = f_{ac} f_{bd} + f_{ad} f_{bc}, \qquad 2k_1 = 4F_{XY} - 2k_2,$$
$$k_0 = 1 - 2k_1 - k_2$$

and the calculation is reduced to the methods used to calculate inbreeding coefficients. All of the k-coefficients in Table 2 can be calculated from these formulas.

Extensions

The k-coefficients described above apply to a single autosomal gene in a diploid organism. Several kinds

Table 1 The joint genotype distribution of X and Y

X	Y	${}_{11}k_2$	$2_{01}k_1$	$2_{10}k_1$	${}_{11}k_0$	${}_{10}k_0$	${}_{01}k_0$	${}_{00}k_2$	$2_{00}k_1$	${}_{00}k_0$
AA	AA	p	p^2	p^2	p^2	p^3	p^3	p^2	p^3	p^4
AA	Aa	0	0	pq	0	$2p^2q$	0	0	p^2q	$2p^3q$
AA	Aa	0	0	0	pq	pq^2	p^2q	0	0	p^2q^2
Aa	AA	0	pq	0	0	0	$2p^2q$	0	p^2q	$2p^3q$
Aa	Aa	0	0	0	0	0	0	$2pq$	pq	$4p^2q^2$
Aa	aa	0	pq	0	0	0	$2pq^2$	0	pq^2	$2pq^3$
aa	AA	0	0	0	pq	p^2q	pq^2	0	0	p^2q^2
aa	Aa	0	0	pq	0	$2pq^2$	0	0	pq^2	$2pq^3$
aa	aa	q	q^2	q^2	q^2	q^3	q^3	q^2	q^3	q^4

Table 2 The k-coefficients for some simple relationships

Relationship	${}_{00}k_2$	$2_{00}k_1$	${}_{00}k_0$
Unilineal			
Parent–offspring	0	1	0
Grandparent–grandchild	0	1/2	1/2
Half sibs	0	1/2	1/2
Avuncles	0	1/2	1/2
First cousins	0	1/4	3/4
Bilineal			
MZ twins	1	0	0
Full sibs	1/4	1/2	1/4
Double first cousins	1/16	6/16	9/16

of extensions are immediately apparent: to more than one autosomal locus [1, 2, 6], to more than two individuals [13], and to X-linked loci [4]. For a recent review and more extensive bibliography, see [8].

References

[1] Campbell, M.A. & Elston, R.C. (1971). Relatives of probands: models for preliminary genetic analysis, *Annals of Human Genetics* **35**, 225–236.

[2] Cockerham, C.C. & Weir, B.S. (1973). Descent measures for two loci with some applications, *Theoretical Population Biology* **4**, 300–330.

[3] Cotterman, C.W. (1940). A calculus for statistico-genetics, *Unpublished PhD thesis*. Ohio State University, Columbus, Ohio.

[4] Denniston, C. (1967). Probability and genetic relationship. *Unpublished thesis*, University of Wisconsin, Madison.

[5] Denniston, C. (1974). An extension of the probability approach to genetic relationships: one locus, *Theoretical Population Biology* **6**, 58–75.

[6] Denniston, C. (1975). Probability and genetic relationship: two loci, *Annals of Human Genetics* **39**, 89–104.

[7] Gillois, M. (1964). La relation d'identité en génétique. Thèse Faculté des Sciences de Paris.

[8] Gillois, M. (1988). Consanguinity, in *Proceedings of the Second International Conference on Quantitative Genetics*, B.S. Weir, E.J. Eisen, M.M. Goodman & G. Namkoong, eds. Sinauer, Sunderland, pp. 353–359.

[9] Jacquard, A. (1974). *The Genetic Structure of Populations*. Springer-Verlag, Berlin.

[10] Malécot, G. (1941). Etude mathématique des populations "mendéliennes", *Annales de l'Université de Lyon, Sciences, Section A* **2**, 25–37.

[11] Malécot, G. (1948). *Les mathématiques de l'hérédité*. Masson, Paris.

[12] Nadot, R. & Vaysseix, G. (1973). Apparentement et identité. Algorithme du calcul des coefficients d'identité, *Biometrics* **29**, 347–359.

[13] Thompson, E.A. (1974). Gene identities and multiple relationships, *Biometrics* **30**, 667–680.

[14] Vu Tien Khang, J., De Rochambeau, H., Chevalet, C. & Gillois, M. (1979). Analyse des pedigrees et calcul des coefficients d'identité par les arbres géniques, *Biometrical Journal* **21**, 367–387.

C. DENNISTON

Image Analysis and Tomography

Seeing is believing: sight is fundamental to our understanding of the world. This is as true in science as in everyday life. The collection of much statistical data is dependent upon human vision. For example, the examination of samples under a microscope, observing animal behavior, and the identification and counting of plant species in a field are all forms of image analysis. We are superb at analyzing the images projected onto our retinas, using one-third of our brains for vision. However, computers are being used more and more to automate and extend the potential of image analysis. Computers are better at extracting quantitative information from images than human observers: they can be more accurate and more consistent from day to day. Furthermore, computers may spare us from much tedious image interpretation.

We see effortlessly, most of the time. Progress was expected to be rapid when research commenced in the 1960s on computer-based image analysis. The task, however, has proved to be far more difficult. At least in part, this is because we are not conscious of the processes we go through in seeing. Biological objects present an even greater challenge to computer interpretation than man-made ones, because they tend to be more irregular and variable in shape.

Application Areas

Images to be analyzed in biostatistics may come from microscopy, medical scanning systems, electrophoresis, or simply from photographing illuminated objects. Figure 1 shows several such examples.

Figure 1(a) is a back-illuminated optical *microscope image* of cashmere goat fibers whose diameters were to be measured [38]. Measurement is made more difficult because the microscope has a shallow depth of focus and some fibers are out of focus, producing either dark or light edges to the fibers, so-called "Becke lines". There is a danger of misinterpretation if the optics that produced a particular image are not correctly understood. For example, the bas-relief type of images typical of differential interference contrast microscopy may be mistaken for three-dimensional features. However, tailoring image processing **algorithms** to particular forms of microscopy poses a considerable challenge. There are many optical microscope systems, including brightfield, darkfield, phase contrast, interference contrast, fluorescence, and confocal systems (see, for example [91]). There are also many other types of microscope systems such as scanning electron microscopes and confocal microscopes. Also, the theory of microscopy is complicated and agreement with data is less than perfect.

Figure 1(b) shows an example of an image produced by a *medical imaging system*, in this case a reconstructed slice (tomogram) from positron emission tomography (PET). It shows a transverse cross-section through a woman's thorax, with a tumor circled. There are many other medical imaging systems, such as conventional radiology, angiography, X-ray transmission computed tomography (CT), ultrasound imaging, magnetic resonance imaging (MRI), and single photon emission (computed) tomography (SPET or SPECT), each with its own characteristics requiring attention in analysis (for example, [16] and [72]). Tomographic methods, including CT, SPET, and PET, seek to reconstruct slices *within* the body from observations *outside*.

Figure 1(c) shows a type of *electrophoresis gel*, a DNA sequencing gel autoradiograph, produced as one stage in the **DNA sequencing** of gene fragments. About 50 mixtures of radioactively labeled fragments are positioned as distinct spots along one side of the gel. Each mixture then migrates down the gel, and DNA fragments produce separate, approximately horizontal bands. Finally, a photographic plate is placed over the gel. This blackens in response to radioactive emissions, thus producing an autoradiograph. Electrophoresis has many variants, including two-dimensional (2D) electrophoresis, electrofocusing, isotachophoresis, and several forms of immunoelectrophoresis [45]. Various forms of chromatography and chemical assays also produce pictorial information which can be interpreted by image analysis.

Finally, Figure 1(d) is an image of *illuminated objects*, in this case of 50 wheat grains, obtained using a video camera. This was part of an experiment to see if it was possible to estimate flour yield by digital image analysis [8]. Opportunities are almost limitless for digitally analyzing images of objects illuminated in many different ways. See, for example, the review by Price & Osborne [77] of imaging applications in agriculture and plant science, and by

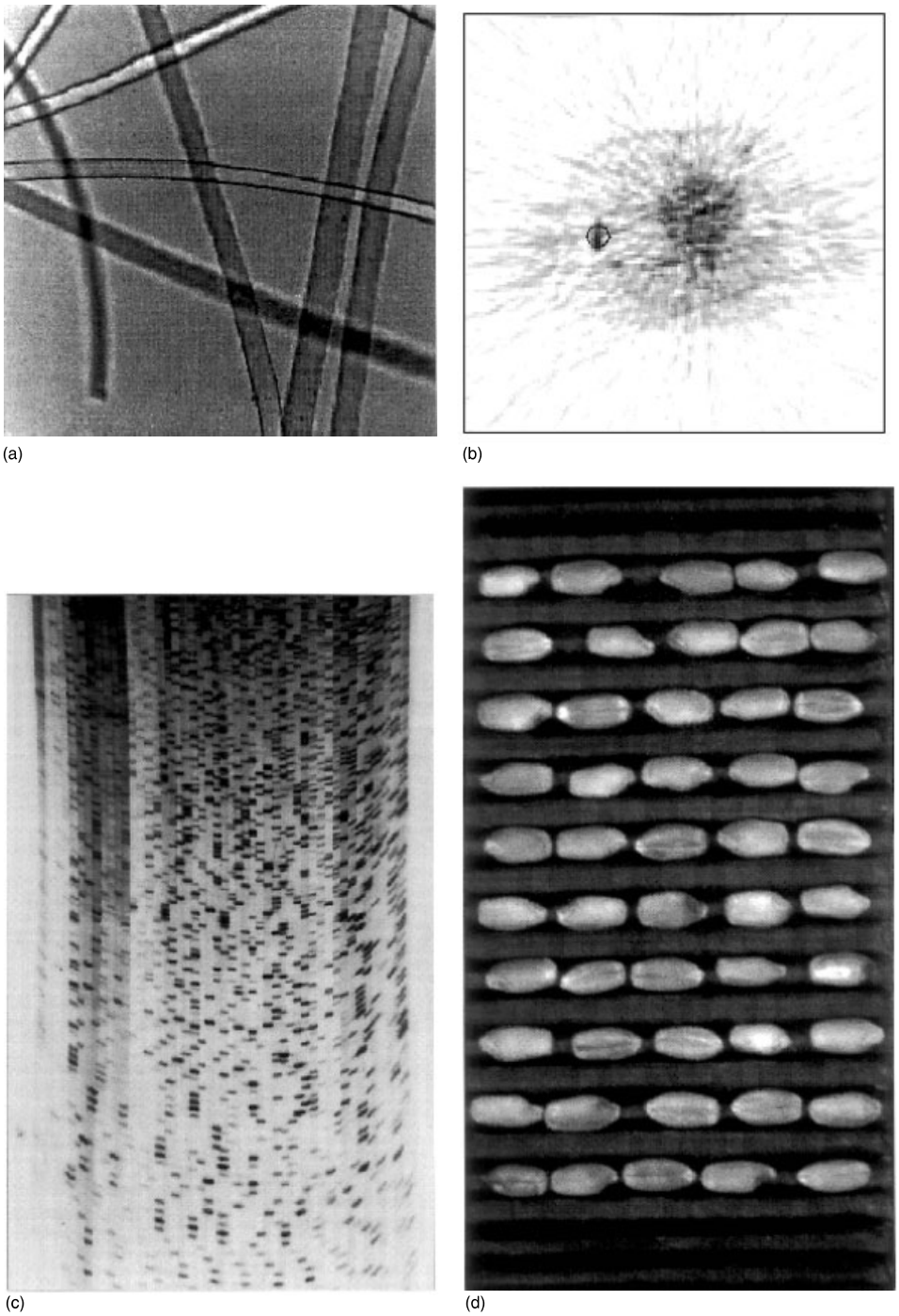

Figure 1 Examples of images: (a) microscope image of cashmere fibers; (b) positron emission tomogram (PET) of a transverse cross-section through a woman's thorax, with a tumor circled, reconstructed using the filtered-backprojection (FBP) algorithm (by courtesy of Max Lonneux and C. Michel, Positron Tomography Laboratory, UCL Belgium); (c) DNA sequencing gel autoradiograph; (d) wheat grains

Sapirstein [85] of cereal variety identification from grains.

Types of Image

Digital images are obtained via an appropriate image capture device, such as a video camera or scanner. A 2D digital image usually consists of a rectangular array of tiny squares called "picture elements", or *pixels* for short. Associated with each pixel is a number, representing the average *brightness* of that part of the original picture covered by the pixel. Usually, the brightness will be discretized to 8-bit resolution, i.e. there are $256 = 2^8$ shades of grey, with 0 representing black and 255 representing white.

The pixel brightness may represent *any* variate which has been measured on a 2D grid. Typically it is a measure of the intensity of reflected light, as in the wheat grains example [Figure 1(d)], or of transmitted light, as with the cashmere fibers [Figure 1(a)]. However, it could alternatively depend upon reflected or transmitted radiation in another part of the electromagnetic spectrum [such as gamma rays, Figures 1(b) and (c)].

The object being imaged may be essentially 2D, as with the DNA sequencing gel [Figure 1(c)], or three dimensional (3D). In the latter case, the *sampling procedure* may involve taking a cross-section, either physically or by a computer reconstruction [e.g. Figure 1(b) is a cross-section of a 3D tomography reconstruction: *tomography* techniques are, literally, those creating pictures of a slice, free of the effects of layers outside the focused plane]. Alternatively, a 3D object with either an opaque [wheat grains, Figure 1(d)] or a semi-transparent surface [cashmere fibers, Figure 1(a)], could be imaged simply by viewing it from a particular direction. Some sensors, such as confocal microscopes and magnetic resonance images, can collect *3D arrays* of data. These can be analyzed using similar methods to those for 2D images.

Although we only consider univariate, so-called *grayscale*, images in this article, it is worth pointing out the increasing use of *color* and *multispectral* image analysis. A color image actually consists of three grayscale images, representing light at red, green, and blue wavelengths, respectively. The wheat grains image [Figure 1(d)] is in fact the green component of a color image.

Methodologies

The ultimate aim of image analysis is usually to extract quantitative information, which may be in the form of binary presence/absence categories, or of measures of object location, length or area, shape statistics, etc. In some applications it may only be possible or desirable to automate some stages in an analysis, leaving the rest to human interpretation. For example, in medical diagnosis the radiologist will want to look at the SPET image. Image analysis methods constitute an eclectic collection of techniques derived from many different theoretical standpoints:

1. The first, and probably most widely used, approach arose in the 1960s from the engineering discipline of *signal processing*, as typified by the books of Rosenfeld & Kak [83] and Jain [55]. Methods include histogram transformations, linear and nonlinear filters (*see* **Spectral Analysis**) and thresholding – techniques that we illustrate later.
2. An elegant approach, termed *mathematical morphology*, emerged from the Ecole des Mines in Fontainebleau, France, in the 1970s (*see* **Stereology**). It is based on the assumption that an image consists of structures which may be handled by set theory, leading to such highly effective methods as openings, closings, skeletonization, and watershed segmentation. The seminal works are Serra [86, 87]. Soille & Rivest [93] provide a useful introduction to the subject from an applications perspective.
3. From **artificial intelligence** have arisen approaches such as *syntactic* **pattern recognition** [28] and *computer vision* [6], but these methods have not often been applied in biostatistical contexts.
4. The 1980s saw the development of **Bayesian** *image analysis* [9, 31]. **Prior** *information* on an appropriate model for an image is combined with *data*, imperfect information about the image (such as pixel values affected by noise), in order to derive the *posterior distribution* for the image.
5. Yet another aspect of image analysis, namely that of extracting *measurements* such as lengths, areas, histograms, etc. from images, is identified as a distinct approach by Serra [87, p. 10]. These descriptors are subsequently interpreted

using **stereology**, shape statistics, or classification methods.

Serra [87, p. 11] acknowledges that, although the different approaches to image analysis are somewhat contradictory, they each have their place. He suggests that an analysis might first require linear methods, then morphological ones, and finally either measurements or syntactic methods.

In the rest of this article we consider first the tomographic reconstruction of images, then the three major components of image analysis: enhancement, segmentation, and taking measurements, drawing on techniques from each of the above approaches and illustrating them using some of the images in Figure 1. We are concerned with the application of image analysis. Therefore, in this article we emphasize methods that we have found useful in practice. We are also conscious of only having space to present a subset from a very large field.

Tomography

In tomography the measured data are *projections*, from which the spatial distribution within a body is reconstructed. For example, Figure 2(a) shows PET recordings similar to those used to reconstruct the cross-section shown in Figure 1(b): these are data from only one of many planes collected in a single study. (The cross indicates the position of the cursor used as an aid to navigate through the 3D data set.) In PET, a radionuclide is introduced into the patient's bloodstream and then distributes throughout the body. Radionuclide decays are recorded on a PET scanner. The *distribution* and *intensity* of activity is recorded by the PET scanner, but the accumulated distribution of radionuclide in the body can only be inferred indirectly (by mathematical analysis) from the scanner projection data.

Projection measurements may be modeled as known linear functionals of an unknown spatial distribution (or image). The goal of reconstruction is to infer from the data the distribution within the specimen. Rosenfeld & Kak [83, Chapter 8] and Jain [55, Chapter 10] present the filtered-backprojection (FBP) algorithm [as used in Figure 1(b)] for reconstruction from data given by the Radon-transform (ray sums) of an image. See also Girard [34] and Bickel & Ritov [11], who study estimation of linear functionals and asymptotic convergence of FBP in PET.

Statistical interest has centered on PET and SPET, which exhibit significant statistical variability in camera recordings. A recent introduction to PET and SPET is given by Kay [58]. McColl et al. [68] describe statistical methods in neuroimaging, with special reference to these imaging modalities. However, many of the methods described apply more widely. We consider further: reconstruction methodologies, use of prior information in fusing reconstructions from different modalities, and parametric mapping based on **pharmacokinetic modeling**. Reconstructions are used not only for clinical studies, providing a basis for individual patient diagnosis, but they also provide data for research studies.

Statistical Reconstruction

In pioneering work, Shepp & Vardi [88] and Lange & Carson [63] applied the **EM algorithm** for **maximum likelihood** (ML) estimation of **Poisson** count data in tomography. Both transmission and emission tomography are well modeled by a description [92] based on a spatially inhomogeneous **Poisson point process**. If $f(s)$ is the (unknown) distribution, then the distribution of recorded activity [in projections such as Figure 2(a)] is expressible as

$$g(t) = \int_{s \in \mathcal{X}} a(t|s) f(s) \, \mathrm{d}s,$$

or, in suitably discretized form, $g = Af$, with $A = (a_{ts})$. Conditional probabilities $a(t|s)$ are determined by the resolution and geometry of the acquisition camera and by the physics of photon transport through the body, and may be regarded as known. There is great generality in this specification, whether in radiology, CT, SPET, or PET, for precise modeling of physical effects (attenuation, scattering) including the ability to incorporate information from other modes. See, for example, Aykroyd & Green [3], Fulton et al. [29], Hutton et al. [54], Vardi & Lee [98], and Weir & Green [100]. This model flexibility and resulting gains in restoration quality have led to a growing interest in statistical reconstruction in clinical use.

Reconstruction (i.e. estimating f from observations on g) constitutes an *inverse problem*. Many reconstruction methods apply corrections using *backprojection* (*see* **Back-calculation**), redistributing residual projection errors ($z = g - \hat{g}$) to provide the

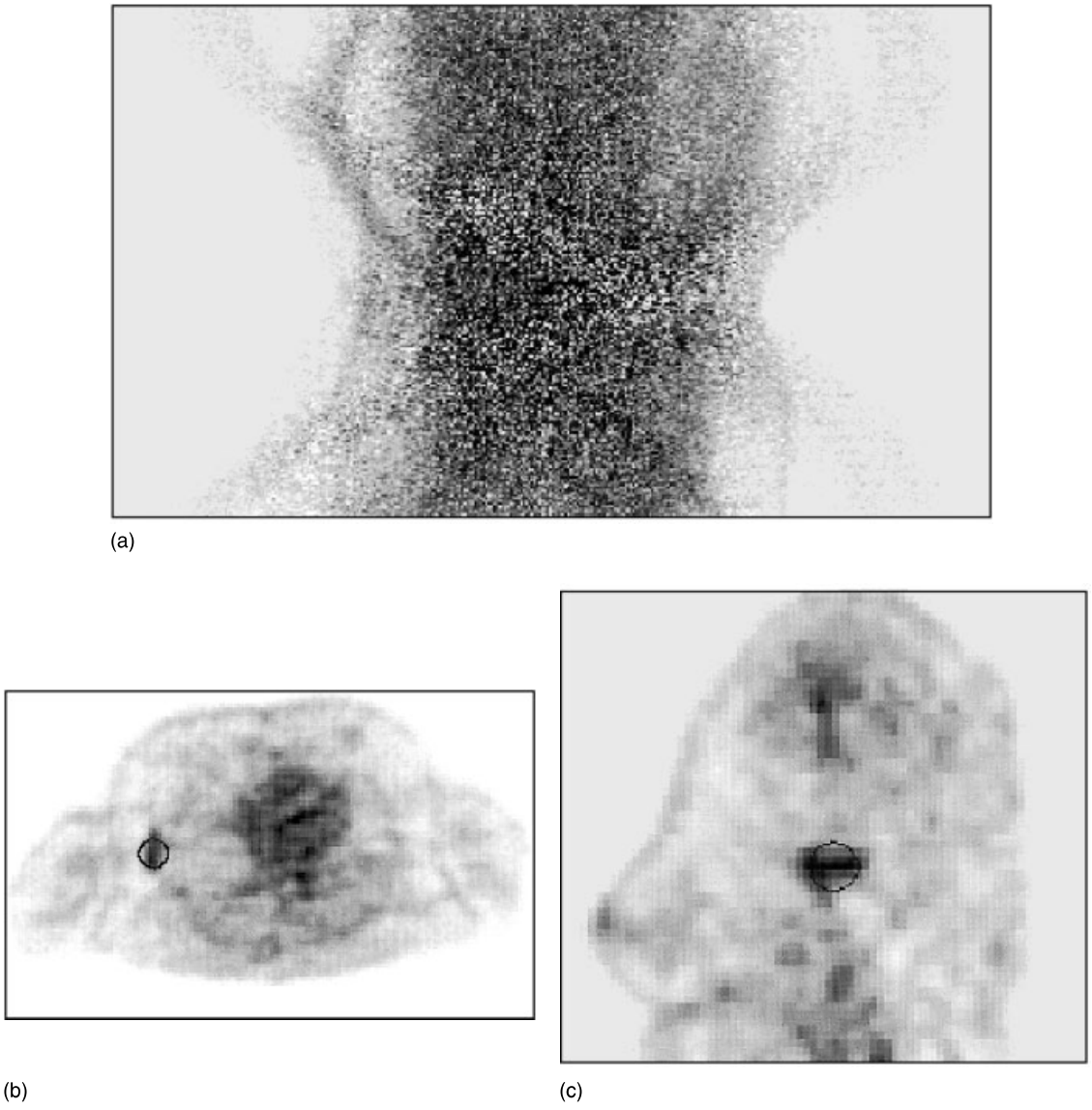

Figure 2 PET clinical study of a woman's thorax: (a) one plane of the projection data (with the cross indicating the position of the cursor used as an aid to navigate through the 3D data set); (b) a transverse cross-section, as in Figure 1(b), but reconstructed using four iterations of the OSEM algorithm; (c) a Sagittal cross-section, with a tumor circled (by courtesy of Max Lonneux and C. Michel, Positron Tomography Laboratory, UCL Belgium)

correction $\hat{f} + \hat{\delta f}$ to an initial estimate $\hat{f}$ by

$$\hat{\delta f}(s) \propto \int_{t \in y} a(t|s) z(t)\, \mathrm{d}t.$$

For example, ML–EM computations provide iterative improvements on a starting image, which is typically taken to be a **uniform distribution** within the body. ML–EM requires repetition of two steps: (i) project the current source estimate to produce fitted projection data, and (ii) backproject the ratio between observed and fitted projections to determine multiplicative corrections to be applied to the current source distribution. The usual convergence theory for EM algorithms shows that each iteration increases the

likelihood and ML–EM converges from any starting image to an ML solution. There is a heavy computational burden, with arrays typically of size 128^3, but this can be eased by exploiting the sparse structure of the matrix A.

Figure 2(b) shows a 2D slice of a reconstruction, based on the EM algorithm, of the same projection data used with the FBP algorithm in Figure 1(b). In addition, Figure 2(c) shows a Sagittal slice through the 3D reconstruction. The gain in restoration quality of the statistical reconstruction is evident, with Figure 1(b) exhibiting streaking artifacts typical of FBP. The reconstruction was produced using the ordered subsets EM (OSEM) algorithm described in Hudson & Larkin [52], which is an adaption of Shepp & Vardi's iterative ML–EM algorithm. OSEM accelerates EM in its ML and Bayesian forms. Here four OSEM iterations were employed, each requiring similar computational effort to that required for the full FBP reconstruction, but far fewer than would be required in ML–EM for the same result.

Related Issues and Approaches

While the resolution of ML–EM images continues to improve with further iterations, they also exhibit an undesirable increase in noise. The effect is similar to bias–variance tradeoffs in nonparametric **density estimation**, and is attributable to the ill-posed inverse problem formulation. A choice of a regularized solution is therefore required. Approaches here include:

1. early stopping of iterations [as in Figure 2(b)];
2. a Bayesian specification of prior information or penalized likelihood criterion (*see* **Penalized Maximum Likelihood**) (see Green & Silverman [42]);
3. post-reconstruction smoothing (see Beekman & Viergever [7]).

Silverman et al. [89] propose an approach to reduce the buildup in noise within iterative reconstruction by local smoothing. The Shepp–Vardi ML-EM algorithm is readily modified to accomplish reconstruction by adopting prior information in a Bayesian formulation (e.g. [32], [41], and [48]) as required for regularization. Multiscale reconstruction may also be advantageous, and there are obvious applications of **wavelet** methodology with body organs creating discontinuities within the imaged region. Efficient convergence is also a critical factor.

Dynamics

Parametric mapping involves modeling functional parameters (e.g. metabolism or blood flow) on the basis of the time-varying distribution of activity of a tracer introduced into the bloodstream in a controlled manner. Time sequences of images result, with the aim of reconstruction being to provide maps of parameters of the model specifying dynamics, not the activity distribution itself.

Cunningham & Jones [22] propose a semiparametric spectral decomposition useful in compartmental models of **pharmacokinetic** studies. In this approach the total activity within prespecified regions of interest (or pixels) are collected over consecutive time intervals. The methodology is nonparametric. No specific compartmental model is assumed, but the time activity curves are expressed in terms of a dense set of basis functions. Cunningham & Jones provide a number of illustrations of the interpretability of such models; the review of O'Sullivan [73] extends this methodology. The method can be applied to time activity curves of indirect observations (projection data) equally well to determine significant modes. With indirect data a staged approach separating the spatial and temporal stages of the reconstruction may be adopted, as provided in the EMPIRA algorithm of Carson & Lange [17].

Enhancement

All images are subject to some degradation from their ideal forms, whether this is the presence of noise, blurring, or a warping/distortion of the image frame. Image enhancement is a set of methods for modifying images to reduce these effects, both to aid human interpretation and as a precursor to segmentation or other digital methods of analysis. In some images the degradation is relatively minor, and image enhancement is unnecessary for the particular application. However, in many cases this will not be so. We look at methods for correcting for warping, at filters, and at deconvolution, using the DNA sequencing gel in Figure 1(c) for illustration.

Registration and Unwarping

Unwarping of images is an important stage in many applications of image analysis. It may be needed to

remove optical distortions introduced by a camera or viewing perspective [96], or to register an image with a reference grid such as a map, or to align two or more images. For example, matching is important in reconstructing a 3D shape from either a series of 2D sections or stereoscopic pairs of images. There is considerable interest in registering images produced by medical sensing systems with body atlas information [18, Section 3; 30] and in image fusion [2]. In tomography studies, MRI or CT provide accurate maps of *anatomy* while PET or SPET provide much lower resolution maps of *function*. Linking function to the anatomy is of interest. Image registration and segmentation techniques are required here.

There have been many approaches to finding an appropriate warp, but a common theme is the compromise between insisting that the distortion be *smooth* and achieving a *good match*. In some recently published cases the warp seems unnecessarily rough [19, Figure 8b; 44, Figure 7f]. Smoothness can be ensured by assuming a parametric form for the warp, such as the affine transformation, or by insisting that the warp satisfies partial differential equations such as Navier's equilibrium equations for elastic bodies [5]. Depending on the application, matching might be specified by points which must be brought into alignment [12], by local measures of **correlation** between images, or by the coincidence of edges [15].

In the DNA sequencing gel, shown in Figure 1(c), it is clear that bands are not aligned, because of a relative lengthening of the tracks near the center of the gel, known colloquially as a "smile" on the gel. Interpretation of electrophoretic gels often involves making comparisons between tracks, or between spot positions on different gels. Distortions are common. Figure 3(a) shows the result of an unwarping operation proposed by Glasbey & Wright [37]. Horgan et al. [51] show how affine and thin-plate spline transformations can be used to align two or more 2D electrophoretograms.

Filters

Filters have two roles in image analysis, either to *reduce noise* by smoothing or to *emphasize edges*, i.e. boundaries between objects or parts of objects. Filters are *linear* if the output values are linear combinations of the pixels in the original image, otherwise they are *nonlinear*.

Linear filters are well understood and fast to compute. They can be studied and implemented in either spatial or frequency domains. Linear filters can be categorized as *low-pass* or *high-pass*, according to whether they smooth by removing high-frequency components in images, or emphasize edges by removing low-frequency components. A third category, *band-pass* filters, remove both the lowest and highest frequencies from images. Use of the **Fast Fourier Transform** leads to efficient computation for filters larger than 5×5. Further details can be found in Glasbey & Horgan [36, Chapter 3]. Note that smoothing filters are a form of kernel regression (*see* **Nonparametric Regression**). See, for example, Hastie & Tibshirani [47, Chapter 2] for a review of this and alternative statistical approaches to smoothing.

In filtering to reduce noise levels, linear smoothing filters inevitably blur edges, because both edges and noise are high-frequency components of images. Nonlinear filters are able to simultaneously reduce noise and preserve edges, but they have less secure theoretical foundations and can be slow to compute. The simplest, most studied, and most widely used nonlinear filter is the moving median. However, many other *robust estimators* of location have also been used [27]. Multiresolution methods based on wavelets are a new approach to smoothing images [25], which also offer great potential in other areas of image analysis.

Morphological filters are a subclass of nonlinear filters, the simplest of which are based on "max" and "min" operations. Substantial improvements in images can often be achieved using sequences of such filters. For example, another problem with Figure 1(c) is that the brightness in the background varies. This is a common problem in image analysis, and makes comparison of similar features in different parts of the image difficult. A morphological *closing* of the image can be used to estimate the background trend. The simplest closings are obtained by first replacing each pixel by the maximum local intensity in a region (e.g. using a *structuring element* which is a disc of radius R centered on each pixel), and then performing a similar operation on the resulting image, using the local minimum. Mathematically, the pixels, z_{ij}, in the closed image will be given by

$$z_{ij} = \min_{k,l} x_{i+k,j+l} \quad \text{and} \quad x_{ij} = \max_{k,l} y_{i+k,j+l},$$

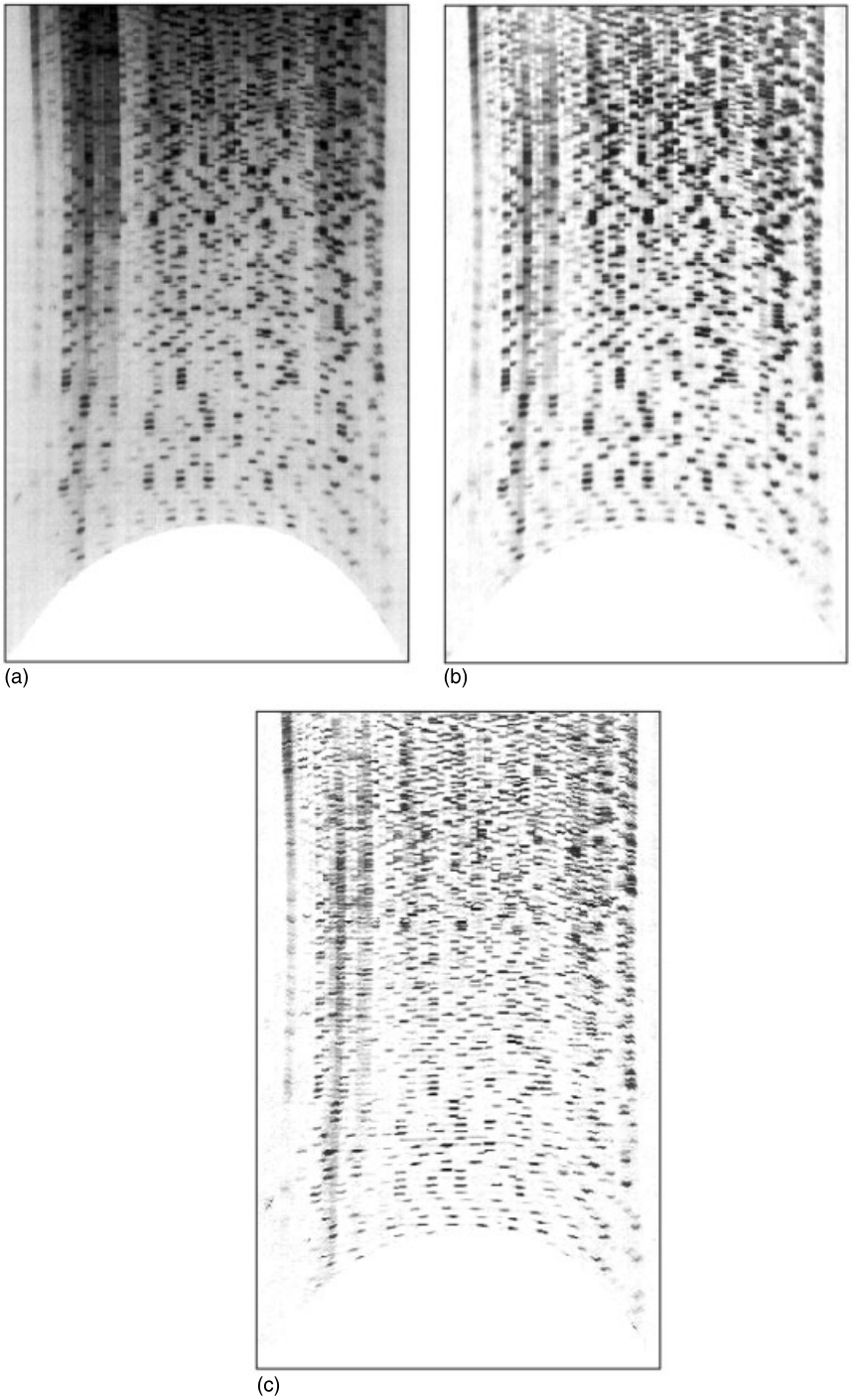

Figure 3 Enhancement of image of DNA sequencing gel autoradiograph: (a) after unwarping of Figure 1(c); (b) after application of top-hat transform to Figure 3(a) to remove background trend; (c) after constrained least squares deconvolution of Figure 3(b)

where $(k^2 + l^2)^{1/2} \leq R$ and $y_{i,j}$ denotes the original pixel value in row i, column j. If this filter is applied to Figure 3(a), then only the small groups of pixels which are darker than their surroundings will be substantially changed from y_{ij} to z_{ij}. These are the bands. By subtracting z from y, these bands will be made more distinct. Figure 3(b) shows the result using a disc of radius 10 pixels. This is known as a *top-hat filter*. Further morphological filters are discussed in [86], and [93].

Deconvolution

If an image has been contaminated by noise and blurring of forms which are either known or can be estimated, then filters can be constructed which optimally restore the original image. There are both linear and nonlinear deconvolution methods (see, for example [83, Chapter 7]). The fundamental linear method is the *Wiener filter*. Nonlinear restoration algorithms can do better than linear ones, but require substantially more computation. For example, *maximum entropy restoration* [90] is one method which exploits the constraint that the restored image is nonnegative. However, as Donoho et al. [24] point out, there are many alternate methods which are equally good. Nonparametric methods for deconvolution generally require the selection of hyperparameters that control smoothing. Rice [80] evaluates generalized **cross validation** (GCV) in this context. See also Thompson et al. [97]. O'Sullivan & Pawitan [74] describe methods for indirect estimation problems and apply them in tomography.

Examination of the pixel values in Figure 3(b) shows the blurring to be well approximated by a Gaussian distribution with variance $\sigma^2 = 2$. This suggests the following model, in which we only consider blur down columns:

$$y_{ij} = \sum_{k=-m}^{m} w_k \, x_{i+k,j} + e_{ij},$$
$$\text{for } i = 1, \ldots, M, \; j = 1, \ldots, N,$$

where

$$w_k = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-k^2}{2\sigma^2}\right),$$
$$\text{for } k = -m, \ldots, m,$$

M and N are the image dimensions, m is the integer part of 3σ, x_{ij} is an ideal unblurred version of the image, which is constrained to be nonnegative, and e_{ij} is uncorrelated noise. For a more general approach to blur estimation, see, for example, [79].

We can use information about the nature of the degradations to design a filter that will smooth y and enhance the edges, so as to get as close as possible to restoring x. Deconvolution can be posed as a constrained **optimization** problem:

$$\text{minimize } S = \sum_{i=1}^{M} \sum_{j=1}^{N} \left(y_{ij} - \sum_{k=-m}^{m} w_k x_{i+k,j} \right)^2$$
$$\text{with respect to } x_{ij},$$
$$\text{for } i = 1, \ldots, M, \;\; j = 1, \ldots, N,$$
$$\text{subject to } x_{ij} \geq 0.$$

In the absence of the inequality constraint, and provided that we can consider x to be the realization of a random process, the optimal solution is the Wiener filter:

$$\hat{x}^*_{kl} = \frac{y^*_{kl}}{w^*_{kl}} \frac{|w^*_{kl}|^2}{|w^*_{kl}|^2 + S^e_{kl}/S^x_{kl}},$$

where y^* denotes the Fourier transform of y, and S^x_{kl} denotes the spectrum of x at frequency (k, l). For a derivation, see, for example, [83, Section 7.3]. The constrained problem can be solved iteratively by gradient descent. Further details are given in Horgan & Glasbey [50]. Figure 3(c) shows the result of deconvolving Figure 3(b). It can be seen that bands which are very close together have been separated in Figure 3(c), although they are indistinguishable in Figure 3(b).

Segmentation

Image segmentation is the division of an image into *regions* or *objects*. This is often a necessary step before the desired quantitative analysis can be carried out. As an example, we wish to segment the wheat grains image [Figure 1(d)], which was one of 38 such images. Each image consisted of 50 grains of the same type, and different images represented grains of different varieties or sites. The aim of the experiment was to see how well flour yield could be predicted from summary size and shape statistics obtained from each image. It is therefore natural to segment the images into individual grains before measuring and accumulating relevant summary statistics.

In some instances, of course, it is possible to estimate the parameters of interest without first resorting to segmentation. However, typically this requires strong model assumptions, upon which indirect inference can be based. Unfortunately, images are large data sets, e.g. a 512×512 image consists of more than 250 000 pixels. Therefore, there is considerable scope (as in other large data sets) for model assumptions to be violated. Often this has the consequence that the optimal solution for the theoretical model is a poor solution to the real problem! Consequently, in most problems it is necessary to segment an image first before measuring and analyzing the result. In this section we will briefly examine four classes of segmentation: thresholding, edge-based segmentation, region-based segmentation, and Bayesian approaches. The wheat grains image will illustrate some of the techniques discussed.

Thresholding

The simplest method of segmentation is thresholding, i.e. whenever a pixel's value is less than or equal to a certain number, t say, its value is replaced by 1, and otherwise given the value 2.

An obvious question is: How does one choose the threshold(s)? The simplest way is by applying some classification technique to the *histogram* of the pixel values. Glasbey [35] reviewed 11 histogram-based methods for choosing the thresholds automatically, most of which are fairly naive. Perhaps the most sophisticated is the minimum error thresholding technique of Kittler & Illingworth [61], which models the histogram as a mixture of Gaussian distributions. The parameters are estimated iteratively in such a way that the observed and estimated means and variances are equated.

Figure 4(a) shows the histogram of the wheat grains image. Clearly, there are two identifiable groups of pixels: light ones largely belonging to wheat grains, with a mean a little above 100, and dark ones, predominantly associated with the background. Note the non-Gaussian shape of the part of the histogram representing dark pixels, and especially the spike at zero due to the camera setting. Despite the fact that the histogram is not a mixture of Gaussians, we nevertheless applied the minimum error thresholding technique. It gave a value of $t = 66$. (Many other algorithms give a similar value.) Figure 4(b) shows the original wheat grains image, but with pixels whose values exceed 66 overlaid in black. This figure demonstrates a number of relevant issues. First, around each region overlaid in black is a grey halo. For the most part, these halos are 1 or 2 pixels wide. Mostly these represent "mixed" pixels, which are not definitively grain or background, but a mixture of both, caused by camera blur and shadows. In any event, where these halos are narrow, the boundaries of the overlaid regions are close enough to the true grain boundaries for most practical purposes. However, note that the darker parts of some of the grains have not been properly classified. This is perhaps not surprising, because histogram-based thresholding takes no account of spatial context. The remaining classes of segmentation discussed in this section attempt to account for spatial context in various ways. It is also possible to define an *adaptive threshold* which varies across an image (see, for example, [14]).

Edge-Based Segmentation

As the name implies, in edge-based segmentation an attempt is made to find edges in images, often by estimating a "derivative"; see [36, Chapter 3] for a description of some of the more popular edge detectors. One of the simplest edge detectors is *Prewitt's gradient filter*, which implicitly assumes a planar surface in a 3×3 window centered on each pixel, estimates the surface by least squares, and computes its maximal gradient. Figure 4(c) shows this gradient for the wheat grains image. Most of the grain boundaries are apparent, although there are some obvious gaps. Figure 4(d) is the result of thresholding Figure 4(c) at $t = 10$. Less obvious gaps are now apparent, as are some spurious features. This highlights the fundamental problem of edge-based segmentation, namely the absence of parts of boundaries and the presence of spurious edges. Edge tracking methods have been proposed by Hueckel [53], Martelli [67] and Breen & Peden [14], among others, but success is often only partial, especially in images that are more complex than the one analyzed here.

Region Growing and Merging

The basic idea behind region growing is the following. Suppose that one can find distinct points, or clusters of points, such that each distinct cluster belongs to a distinct object in the image, and the number of clusters equals the number of objects.

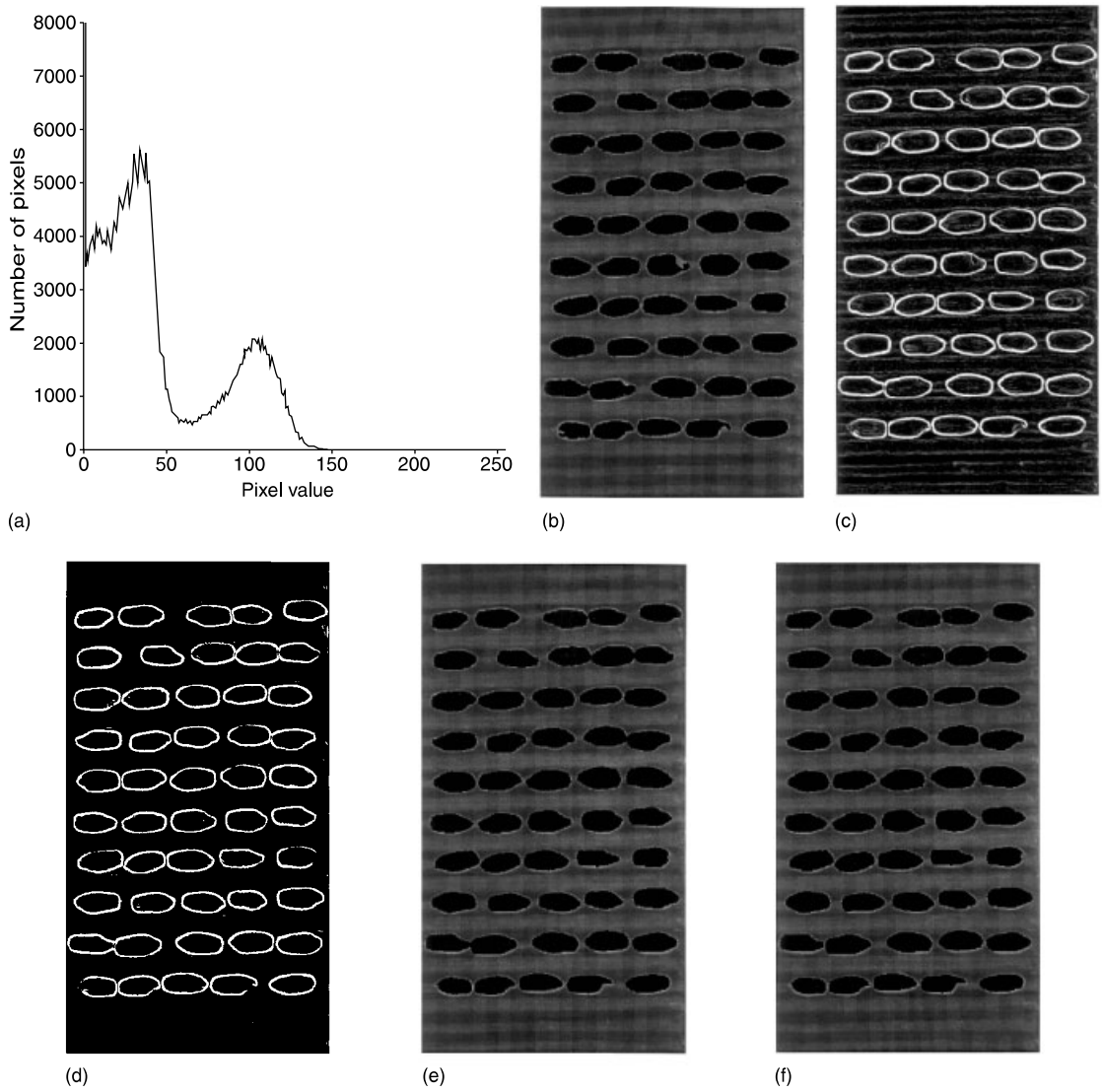

Figure 4 Approaches to segmenting the wheat grains image: (a) histogram of Figure 1(d); (b) after thresholding Figure 1(d) at $t = 66$ (pixels greater than the threshold are overlaid in black); (c) after applying Prewitt's gradient filter to Figure 1(d); (d) after thresholding Figure 4(c) at $t = 10$; (e) after applying seeded region-growing to Figure 1(d); (f) after applying a modified watershed transform to Figure 4(e)

Such points are typically called *seeds* or *markers*. Now grow out spatially from each cluster of seeds according to some mechanism, allocating pixels to objects as they grow in a way that preserves the connectedness of the objects. This process will produce objects with complete boundaries, thereby overcoming a problem with edge-based segmentation mentioned above. Fast algorithms for a number of important region-growing algorithms have been developed in recent years by using data structures that come under the collective name of *priority queues* [13]. In this subsection we apply two important region-growing algorithms to the wheat grains image.

Seeded region growing [1] first computes the mean grayscale of each cluster. Next, all neighboring pixels

of clusters are examined, and the one whose grayscale value is closest to the mean of its neighboring cluster is assigned to that cluster, and the mean value of the cluster is updated. This process continues, one pixel at a time, until all pixels are assigned to a cluster (which by the end of the process is a complete object or region). For the wheat grains image we have chosen our seeds for the grains to be all pixels with a grayscale greater than 80, and our seeds for the background to be all pixels with a grayscale less than 40 [see Figure 4(a)]. Some of the pixels greater than 80 form small (spurious) islands near the bottom of the image [see Figure 4(b)]. Any connected region of pixels less than 100 pixels in area is therefore removed as a seed for the grains. The result of applying seeded region growing using these seeds is shown in Figure 4(e). Apart from the halos mentioned above, the segmentation appears to have found the grains very well. Seeded region growing appears to be quite robust to the choice of parameters; the important thing is to obtain a reasonable number of "representative" seeds for each distinct connected region in an image.

A point to notice in Figure 4(e) is that some of the grains are touching. It is important to separate these grains for subsequent measurements relevant to size, and particularly shape. To do this we employ a variant of a widely used region-growing technique called *watershedding* [69, 99]. The result, shown in Figure 4(f), is a reliable segmentation of the wheat grains. The remaining 37 images were mostly segmented as well and required very little manual intervention.

There are many other split-and-merge algorithms in the literature, most of them more complex than the one presented above. Haralick & Shapiro [46, Chapter 10] discuss a variety of such algorithms and Gordon [40] surveyed methods for constrained classification. The *Hough transform* (see, for example, [64]) can also be used for segmentation by identifying the linear or curved features in images.

Bayesian Approaches

The Bayesian approach to image segmentation received its initial impetus from the pioneering papers of Geman & Geman [31] and Besag [9]. Since then there has been a large number of papers on the subject. However, in the authors' opinion, these techniques are still only applicable for a specialized class of images, in which the models used are good representations of the data. As pointed out earlier, there is plenty of scope for the relatively simple model assumptions used in the Bayesian literature to be violated, because images are such large data sets. However, because of its importance in the statistical literature, we give a brief survey of the area.

Many of the Bayesian approaches to image segmentation rest on variants of the following model as described in Besag [9]. Let S denote the set of all pixels in an image, and let $n = MN$ be the number of pixels in S. Assume that all pixels in the image belong to one of c classes, labeled $1, 2, \ldots, c$, respectively; we do not allow for mixed pixels. Let X_i denote the class to which pixel i belongs (double indexing of subscripts is unnecessary for the present discussion), and let $\mathbf{X} = (X_1, \ldots, X_n)$. Let y_i denote the value recorded at pixel i, and let $\mathbf{Y} = (y_1, \ldots, y_n)$.

Let $f(\mathbf{Y}|\mathbf{X}, \theta)$ denote the conditional density function of $\mathbf{Y}$ given $\mathbf{X}$, with parameter θ. Often (but not always, e.g. [59]) it is assumed that the observations are conditionally independent, i.e. $f(\mathbf{Y}|\mathbf{X}, \theta) = \prod f(y_i|X_i, \theta)$. Let $g(\mathbf{X}, \beta)$ denote the prior distribution of $\mathbf{X}$, with parameter β. In what follows we drop reference to θ and β. It is common to model g as a *locally dependent Markov Random Field* (MRF) [60]. Often, but not always, the local dependence is on the immediate eight neighbors of each pixel. MRFs usually produce a relatively simple structure for g (apart from a normalizing factor); they are also appealing because they can be modeled as limits to (possibly inhomogeneous) **Markov chains**. This means that they can be approximately simulated via **Markov chain Monte Carlo** (MCMC) techniques [10], and are therefore amenable to (computationally intensive) inference (*see* **Computer-intensive Methods**).

The maximum a posteriori (MAP) estimator chooses $\mathbf{X}$ to maximize the posterior likelihood, which is proportional to $f(\mathbf{Y}|\mathbf{X})g(\mathbf{X})$. Unfortunately, this maximization is usually difficult because of the normalization factor mentioned above. In special cases, exact maximization (e.g. [43]) or approximate maximization [26] is possible. However, to circumvent this, Geman & Geman [31] used *simulated annealing* (an inhomogeneous MCMC technique) to find the global maximum of the posterior likelihood. Apart from being computationally intensive, this method sometimes produces gross mislabeling in certain classification problems and "oversmoothing" in related surface reconstruction and image restoration problems [9, 23,

66]. This phenomenon is most probably due to the method's strong dependence on the particular model chosen.

Partly as a consequence of these apparent limitations, Besag [9] introduced the *iterated conditional modes* (ICM) algorithm. Let $h(X_i|X_{S\setminus i})$ denote the distribution of X_i conditional on the other X_js; this will usually have a simple structure for an MRF. Let $\hat{\mathbf{X}}$ denote a provisional estimate of $\mathbf{X}$. ICM *iteratively* chooses $\hat{X}_i$ to maximize

$$p(X_i|\mathbf{Y}, \hat{X}_{S\setminus i}) \propto f(\mathbf{Y}|X_i, \hat{X}_{S\setminus i})h(X_i|\hat{X}_{S\setminus i}).$$

This simplifies in an obvious way when the y_is are independent conditional on $\mathbf{X}$. Besag shows that ICM never decreases the posterior likelihood and so will usually converge to a *local* maximum.

Variants of the above model include those of Geman et al. [33], who imposed constraints on the shapes of class boundaries, and Helterbrand et al. [49], who used boundary closure constraints. A somewhat different and interesting approach is adopted by Baddeley & van Lieshout [4]. They used prior distributions on $\mathbf{X}$ more appropriate for objects of a given shape and size; for instance, in the wheat grains example, these might be ellipses with given radii. The centers of these objects were modeled as nearest-neighbor Markov **point processes**. An algorithm similar to ICM was used to find a local maximum of the posterior distribution. One of Baddeley & van Lieshout's two examples involved fitting circles to an image of (roughly) circular pellets. Their segmentation fitted reasonably well in most places, but not everywhere, in part because the circularity assumptions were not quite right. Similar discrepancies might occur if the wheat grains were modeled as ellipses. Rather than assuming a fixed size and shape, Grenander and coworkers (see Grenander & Miller [44] and references therein) used *deformable templates* to define the boundaries of objects. This requires knowledge of the mean shape of objects, and variability about the mean. They also used jump-diffusion processes to model and simulate the process of interaction between objects. The associated segmentation process appears to be extremely computationally intensive. A related method is where segment boundaries are constrained to be smooth by including roughness penalties such as bending energies in an optimization criterion [71]. This is referred to as the fitting of "snakes" [57]. For further work in this area and a range of applications, see [3], [20], [76], [78], and [81].

We applied a form of ICM [9, Eq. (7)] to the wheat grains image, but the results were only slightly better than those produced by thresholding [Figure 4(b)]. It would seem that stronger prior constraints need to be incorporated. An appropriate Bayesian model and associated estimation procedure would almost certainly segment the wheat grains image as well as the region-based methods. However, it would require a lot of research (and probably data) to find the appropriate model and the estimation procedure is likely to be computationally intensive.

Measurement

The extraction of quantitative information is the endpoint of most image analysis in biostatistics. The aim may simply be to count the number of objects in a scene, or measure their areas, or it may be more complex, such as describing the shapes of objects to discriminate between them.

It is straightforward to count the number of objects in an image provided that the segmentation has successfully associated one, and only one, component with each object. If this is not the case, then manual intervention may be necessary to complete the segmentation. However, short-cuts can sometimes be taken. For example, if the mean size of objects is known, then the number of objects in an image can be estimated, even when they are touching, through dividing the total area covered by all the objects by this average size. It is even possible to make allowance for objects overlapping each other provided that this process can be modeled, for instance by assuming that objects are positioned at random over the image and making use of the properties of Boolean models [21, pp. 753–759]. For example, Jeulin [56] has estimated the size distribution of a powder in such a way. Rudemo et al. [84] used a marked point process model to obtain estimates of plant densities in images of field crops.

Moments offer one method for summarizing segmented objects. If the object we are interested in is represented by all pixels $(i, j) \in A$, then the *(k, l)th moment* is

$$\mu_{kl} = \sum\sum_{(i,j)\in A} i^k j^l, \quad \text{for } k, l = 0, 1, 2, \ldots .$$

In particular, the zeroth-order moment, μ_{00}, specifies the area of the object. First-order moments specify the location of an object. Higher-order moments are also mainly determined by an object's location. *Central moments*, defined by

$$\mu'_{kl} = \sum_{(i,j)\in A}\sum \left(i - \frac{\mu_{10}}{\mu_{00}}\right)^k \left(j - \frac{\mu_{01}}{\mu_{00}}\right)^l,$$
$$\text{for } k + l > 1,$$

are *locationally* – but not *rotationally* – invariant. If orientation is an important feature of an object, as it will be in some applications, then it is probably desirable for the moments to be sensitive to it. However, in other cases orientation is irrelevant and moment statistics are more useful if they are invariant to rotation as well as to location. One such method is based on first specifying the direction in which the object has the maximum value for its second-order moment. This direction is

$$\phi = \frac{1}{2}\tan^{-1}\left(\frac{2\mu'_{11}}{\mu'_{02} - \mu'_{20}}\right), \quad \text{if } \mu'_{02} > \mu'_{20},$$

and is otherwise this expression plus $\pi/2$. Direction ϕ, the *major axis* of the object, has second-order moment:

$$\lambda_1 = \mu'_{20}\sin^2\phi + \mu'_{02}\cos^2\phi + 2\mu'_{11}\sin\phi\cos\phi.$$

The direction perpendicular to ϕ, i.e. the *minor axis*, has the smallest second-order moment of

$$\lambda_2 = \mu'_{20}\cos^2\phi + \mu'_{02}\sin^2\phi - 2\mu'_{11}\sin\phi\cos\phi.$$

For a derivation, see Rosenfeld and Kak [83, Volume 2, pp. 288–290].

Perimeters of objects are also useful summary statistics. Let P denote the number of pixels on the boundary of object A, specified as follows. Pixel (i, j) is on the boundary if $(i, j) \in A$, but one of its four horizontal or vertical neighbors is outside the object, i.e.

$$(i+1, j) \notin A \quad \text{or} \quad (i-1, j) \notin A \quad \text{or}$$
$$(i, j+1) \notin A \quad \text{or} \quad (i, j-1) \notin A.$$

This gives an *8-connected* boundary, with pixels linked either horizontally, vertically, or diagonally. An unbiased estimator of the perimeter is given by

$$\frac{4}{\pi}\frac{P}{\sqrt{2}}$$

provided that either all orientations in the boundary occur equally often or the sampling grid is positioned randomly on the object. This, and more complicated methods for estimating perimeters, are considered by Koplowitz & Bruckstein [62]. The use of scaling factors is part of stereology, a field which has traditionally been concerned with inference about objects using information from lower-dimensional samples, such as estimating volumes of objects from the areas of intersection with randomly positioned cutting planes (see, for example, [95, Chapter 11]). In particular, the scaling factor of $\pi/4$ arises in two of the so-called "six fundamental formulae" of classical stereology. However, the last 10 years have seen a revolution in stereology, with the discovery of the *disector* [*sic*] and other 3D sampling strategies [94]. Note, furthermore, that mathematical morphology can be used to study size distributions of objects in images. By performing openings, using structuring elements at a range of different sizes, a *granulometry* can be obtained [86, Chapter 10].

Shape information is what remains once location, orientation, and size features of an object have been dealt with. One commonly used shape statistic is a measure of *compactness*, which is defined to be the ratio of the area of an object to the area of a circle with the same perimeter. Another statistic often used to describe shape is a measure of *elongation*. This can be defined in many ways, one of which is as the ratio of the second-order moments of the object along its major and minor axes.

Summary statistics of area, perimeter, and major- and minor-axis lengths were obtained for the 50 wheat grains given by segmented regions in Figure 4(f). To illustrate these results, a **principal components analysis** was performed on the log transformed data. Table 1 gives the principal component coefficients. Figure 5 is a scatterplot of the first two

Table 1 Principal component coefficients and percentages of correlation matrix explained for log-transformed summary statistics from 50 wheat grains given by segmented regions in Figure 4(f)

Component:	1	2	3	4
Percent variability:	80.1	18.9	0.9	0.003
Area	0.55	−0.21	0.34	0.74
Perimeter	0.53	0.32	−0.78	0.06
λ_1	0.49	0.54	0.52	−0.45
λ_2	0.42	−0.75	−0.06	−0.50

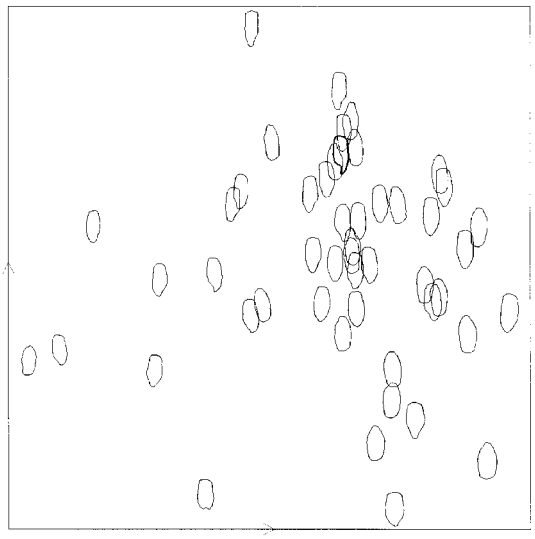

Figure 5 A scatterplot of the first two principal components of summary statistics from the segmented wheat grains [Figure 4(f)], with the first principal component along the horizontal axis. Each point is represented by that grain's outline, and the outline displayed in bold is an outlier in the third principal component

principal components, which account for 99.1% of the variation in the correlation matrix. Each point is represented by that grain's outline. Examination of Table 1 and Figure 5 reveals that the first principal component is an indicator of grain size, while the second is a composite measure of compactness and elongation. The third principal component discriminates between one unusual grain outline, that shown in bold in Figure 5, and the rest. Comparison with Figure 1(d) shows that this grain is not particularly unusual, but rather that the segmentation has failed to recognize a particularly dark part of the grain. Berman et al. [8] used these summary statistics, together with those from a further 37 images, to predict flour yield. They found that the average area of grains in each image, together with averages of λ_1, λ_2, an estimate of volume of a prolate ellipsoid proportional to $\lambda_1\lambda_2^2$, and grain weight explained 65% of the variation in flour yield.

The description of shape is an open-ended task, because there are potentially so many aspects to an object even after location, orientation, and size effects have been removed. Other approaches include the use of *landmarks* [39] and warpings such as *thin-plate splines* and other *morphometric methods* [12], which consider image plane distortions needed to move landmarks to designated locations. Rohlf & Archie [82] and Mou & Stoermer [70] compared alternate forms of *Fourier descriptors* to approximate object boundaries, and applied Zahn & Roskies' [101] method to describe the outlines of mosquito wings and diatoms, respectively. Further methods are discussed in the reviews of shape analysis by Pavlidis [75] and Mardia et al. [65].

References

[1] Adams, R. & Bischof, L. (1994). Seeded region growing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**, 641–647.

[2] Ardekani, B.A., Braun, M., Hutton, B. & Kanno, I. (1996). Minimum cross-entropy reconstruction of PET images using prior anatomical information obtained from MR, in *Quantification of Brain Function Using PET*, R. Myers, V. Cunningham, D. Bailey & T. Jones, eds. Academic Press, London, pp. 113–117.

[3] Aykroyd, R.G. & Green, P.J. (1991). Global and local priors, and the location of lesions using gamma-camera imagery, *Philosophical Transactions of the Royal Society, London, Series A* **337**, 323–342.

[4] Baddeley, A.J. & van Lieshout, M.J.M. (1993). Stochastic geometry models in high-level vision, *Advances in Applied Statistics* **20**, Supplement, 233–258.

[5] Bajcsy, R. & Kovacic, S. (1989). Multiresolution elastic matching, *Computer Vision, Graphics and Image Processing* **46**, 1–21.

[6] Ballard, D.H. & Brown, C.M. (1982). *Computer Vision*. Prentice-Hall, Englewood Cliffs.

[7] Beekman, F.J. & Viergever, M.A. (1996). Evaluation of fully 3D iterative scatter correction and post-reconstruction filtering in SPECT, in *Three Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, P. Greangeat & J.-L. Amans, eds. Kluwer, Dordrecht, pp. 163–175.

[8] Berman, M., Bason, M.L., Ellison, F., Peden, G. & Wrigley, C.W. (1996). Image analysis of whole grains to screen for flour-milling yield in wheat breeding, *Cereal Chemistry* **73**, 323–327.

[9] Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion), *Journal of the Royal Statistical Society, Series B* **48**, 259–302.

[10] Besag, J. & Green, P.J. (1993). Spatial statistics and Bayesian computation, *Journal of the Royal Statistical Society, Series B* **55**, 25–38.

[11] Bickel, P.J. & Ritov, Y. (1995). Estimating linear functionals of a PET image, *IEEE Transactions on Medical Imaging* **14**, 81–87.

[12] Bookstein, F.L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge.

[13] Breen, E.J. & Munro, D.H. (1994). An evaluation of priority queues for mathematical morphology, in *Mathematical Morphology and its Applications to Image Processing*, J. Serra & P. Soille, eds. Kluwer, Dordrecht, pp. 249–256.

[14] Breen, E.J. & Peden, G.M. (1994). Automatic thresholding and edge linking of ferritic steel weld images, *Journal of Computer-Assisted Microscopy* **6**, 167–179.

[15] Burr, D.J. (1981). A dynamic model for image registration, *Computer Graphics and Image Processing* **15**, 102–112.

[16] Bushberg, J.T., Seibert, J.A., Leidholdt, E.M. & Boone, J.M. (1994). *The Essential Physics of Medical Imaging*. Williams & Wilkin, Baltimore.

[17] Carson, R.E. & Lange, K. (1985). The EM parametric image reconstruction algorithm, *Journal of the American Statistical Association* **80**, 20–22.

[18] Colchester, A.C.F. & Hawkes, D.J., eds (1991). Information processing in medical imaging, in *Proceedings of the Twelfth International Conference on Information Processing in Medical Imaging*. Springer-Verlag, Berlin.

[19] Conradsen, K. & Pedersen, J. (1992). Analysis of 2-dimensional electrophoretic gels, *Biometrics* **48**, 1273–1287.

[20] Cootes, T.F., Taylor, C.J., Cooper, D.H. & Graham, J. (1995). Active shape models – their training and application, *Computer Vision and Image Understanding* **61**, 38–59.

[21] Cressie, N.A.C. (1991). *Statistics for Spatial Data*. Wiley, New York.

[22] Cunningham, V.J. & Jones, T. (1993). Spectral analysis of dynamic PET studies, *Journal of Cerebral Blood Flow and Metabolism* **13**, 15–23.

[23] Devijver, P.A. & Dekesel, M.M. (1987). Learning the parameters of a hidden Markov random field model: a simple example, in *Pattern Recognition Theory and Applications*, P.A. Devijver & J. Kittler, eds. Springer-Verlag, Heidelberg, pp. 141–163.

[24] Donoho, D.L., Johnstone, I.M., Hoch, J.C. & Stern, A.S. (1992). Maximum entropy and the nearly black object (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 41–81.

[25] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion), *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

[26] Ferrari, P.A., Frigessi, A. & Gonzaga de Sa, P. (1995). Fast approximate maximum a posteriori restoration of multicolour images, *Journal of the Royal Statistical Society, Series B* **57**, 485–500.

[27] Fong, Y., Pomalaza-Raez, C.A. & Wang, X. (1989). Comparison study of nonlinear filters in image processing applications, *Optical Engineering* **28**, 749–760.

[28] Fu, K.S. (1982). *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs.

[29] Fulton, R., Hutton, B., Braun, M., Ardekani, B. & Larkin, R. (1994). Use of 3D reconstruction to correct for patient motion in SPECT, *Physics in Medicine & Biology* **39**, 563–574.

[30] Gee J.C., Reivich, M. & Bajcsy, R. (1993). Elastically deforming 3D atlas to match anatomical brain images, *Journal of Computer Assisted Tomography* **17**, 225–236.

[31] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–735.

[32] Geman, S. & McClure, D. (1987). Statistical methods for tomographic image reconstruction, *Bulletin of the International Statistical Institute* **52**, 5–21.

[33] Geman, D., Geman, S., Graffigne, C. & Dong, P. (1990). Boundary detection by constrained optimization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 609–628.

[34] Girard, D.A. (1987). Optimal regularized reconstruction in computerized tomography, *SIAM Journal on Scientific and Statistical Computing* **8**, 934–950.

[35] Glasbey, C.A. (1993). An analysis of histogram-based thresholding algorithms, *CVGIP: Graphical Models and Image Processing* **55**, 532–537.

[36] Glasbey, C.A. & Horgan, G.W. (1995). *Image Analysis for the Biological Sciences*. Wiley, Chichester.

[37] Glasbey, C.A. & Wright, F.G. (1994). An algorithm for unwarping multitrack electrophoretic gels, *Electrophoresis* **15**, 143–148.

[38] Glasbey, C.A., Hitchcock, D., Russel, A.J.F. & Redden, H. (1994). Towards the automatic measurement of cashmere-fibre diameter by image analysis, *Journal of the Textile Institute* **85**, 301–307.

[39] Goodall, C. (1991). Procrustes methods in the statistical analysis of shape (with discussion), *Journal of the Royal Statistical Society, Series B* **53**, 285–339.

[40] Gordon, A.D. (1996). A survey of constrained classification, *Computational Statistics and Data Analysis* **21**, 17–29.

[41] Green, P. (1990). Bayesian reconstruction from emission tomography data using a modified EM algorithm, *IEEE Transactions on Medical Imaging* **9**, 84–93.

[42] Green, P. & Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

[43] Greig, D.M., Porteous, B.T. & Seheult, A.H. (1989). Exact maximum a posteriori estimation for binary images, *Journal of the Royal Statistical Society, Series B* **51**, 271–279.

[44] Grenander, U. & Miller, M.I. (1994). Representations of knowledge in complex systems (with discussion), *Journal of the Royal Statistical Society, Series B* **56**, 549–603.

[45] Hames, B.D. & Rickwood, D. eds. (1981). *Gel Electrophoresis of Proteins: A Practical Approach*. IRL Press, London.

[46] Haralick, R.M. & Shapiro, L.G. (1992). *Computer and Robot Vision*, Vol. 1. Addision-Wesley, Reading.

[47] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[48] Hebert, T. & Leahy, R. (1992). Statistic-based MAP image reconstruction from Poisson data using Gibbs priors, *IEEE Transactions on Signal Processing* **40**, 2290–2302.

[49] Helterbrand, J.D., Cressie, N. & Davidson, J.L. (1994). A statistical approach to identifying closed object boundaries in images, *Advances in Applied Probability* **26**, 831–854.

[50] Horgan, G.W. & Glasbey, C.A. (1995). Uses of digital image analysis in electrophoresis, *Electrophoresis* **16**, 298–305.

[51] Horgan, G.W., Creasey, A.M. & Fenton, B. (1992). Superimposing two-dimensional gels to study genetic variation in malaria parasites, *Electrophoresis* **13**, 871–875.

[52] Hudson, H.M. & Larkin, R.S. (1994). Accelerated image reconstruction using ordered subsets of projection data, *IEEE Transactions on Medical Imaging* **13**, 601–609.

[53] Hueckel, M.H. (1971). An operator which locates edges in digitized pictures, *Journal of the Association for Computing Machinery* **18**, 113–125.

[54] Hutton, B.F., Hudson, H.M. & Beekman, F.J. (1997). A clinical perspective of accelerated statistical reconstruction, *European Journal of Nuclear Medicine* **24**, to appear.

[55] Jain, A.K. (1989). *Fundamentals of Digital Image Processing*. Prentice-Hall, Englewood Cliffs.

[56] Jeulin, D. (1993). Random models for morphological analysis of powders, *Journal of Microscopy* **172**, 13–21.

[57] Kass, M., Witkin, A. & Terzopoulos, D. (1988). Snakes: active contour models, *International Journal of Computer Vision* **1**, 321–331.

[58] Kay, J.W. (1994). Statistical models for PET and SPECT data, *Statistical Methods in Medical Research* **3**, 5–21.

[59] Kiiveri, H.T. & Campbell, N.A. (1992). Allocation of remotely sensed data using Markov models for image data and pixel labels, *Australian Journal of Statistics* **34**, 361–374.

[60] Kinderman, R. & Snell, J.L. (1980). *Markov Random Fields and Their Applications*, Contemporary Mathematics, Vol. 1. American Mathematical Society, Providence.

[61] Kittler, J. & Illingworth, J. (1986). Minimum error thresholding, *Pattern Recognition* **19**, 41–47.

[62] Koplowitz, J. & Bruckstein, A.M. (1989). Design of perimeter estimators for digitized planar shapes, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 611–622.

[63] Lange, K. & Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography, *Journal of Computer Assisted Tomography* **8**, 306–316.

[64] Leavers, V.F. (1992). *Shape Detection in Computer Vision Using the Hough Transform*. Springer-Verlag, London.

[65] Mardia, K.V., Kent, J.T. & Walder, A.N. (1991). Statistical shape models in image analysis, in *Proceedings of the 23rd Symposium on the Interface: Computer Science and Statistics*, E.M. Keramidas, ed. Interface Foundation of North America, Fairfax Station, pp. 550–557.

[66] Marroquin, J., Mitter, S. & Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision, *Journal of the American Statistical Association* **82**, 76–89.

[67] Martelli, A. (1976). An application of heuristic search methods to edge and contour detection, *Communications of the Association for Computing Machinery* **19**, 73–83.

[68] McColl, J.H., Holmes, A.P. & Ford, I. (1994). Statistical methods in neuroimaging with particular application to emission tomography, *Statistical Methods in Medical Research* **3**, 63–86.

[69] Meyer, F. & Beucher, S. (1990). Morphological segmentation, *Journal of Visual Communication and Image Representation* **1**, 21–46.

[70] Mou, D. & Stoermer, E.F. (1992). Separating Tabellaria (Bacillariophyceae) shape groups based on Fourier descriptors, *Journal of Phycology* **28**, 386–395.

[71] Mumford, D. & Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems, *Communications on Pure and Applied Mathematics* **42**, 577–685.

[72] National Academy of Science, USA, Committee on the Mathematics and Physics of Emerging Dynamic Biomedical Imaging (1996). *Mathematics and Physics of Emerging Biomedical Imaging*. National Academy Press, Washington.

[73] O'Sullivan, F. (1994). Metabolic images from dynamics positron emission tomography studies, *Statistical Methods in Medical Research* **3**, 87–101.

[74] O'Sullivan, F. & Pawitan, Y. (1996). Bandwidth selection for indirect density estimation based on corrupted histogram data, *Journal of the American Statistical Association* **91**, 610–626.

[75] Pavlidis, T. (1978). A review of algorithms for shape analysis, *Computer Graphics and Image Processing* **7**, 243–258.

[76] Phillips, D.B. & Smith, A.F.M. (1994). Bayesian faces via hierarchical template modelling, *Journal of the American Statistical Association* **89**, 1151–1163.

[77] Price, T.V. & Osborne, C.F. (1990). Computer imaging and its application to some problems in agriculture and plant science, *Critical Reviews in Plant Science* **9**, 235–266.

[78] Qian, W. & Titterington, D.M. (1991). Pixel labelling for 3-dimensional scenes based on Markov mesh models, *Signal Processing* **22**, 313–328.

[79] Reeves, S.J. & Mersereau, R.M. (1992). Blur identification by the method of generalized cross-validation, *IEEE Transactions on Image Processing* **1**, 301–311.

[80] Rice, J. (1986). Choice of smoothing parameter in deconvolution problems, *Contemporary Mathematics* **59**, 137–151.

[81] Ripley, B.D. & Sutherland, A.I. (1990). Finding spiral structures in images of galaxies, *Philosophical Transactions of the Royal Society of London, Series A* **332**, 477–485.

[82] Rohlf, F.J. & Archie, J.W. (1984). A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae), *Systematic Zoology* **33**, 302–317.

[83] Rosenfeld, A. & Kak, A.C. (1982). *Digital Picture Processing*, 2nd Ed. Academic Press, San Diego.

[84] Rudemo, M., Sevestre, S. & Andreasen, C. (1995). Marked point process models for cropweed images, *Scandinavian Image Analysis Conference – 95SCIA*. Swedish Society for Automated Image Analysis, Uppsala, pp. 23–31.

[85] Sapirstein, H.D. (1995). Variety identification by digital image analysis, in *Identification of Food-Grain Varieties*, C.W. Wrigley, ed. American Association of Cereal Chemists, St Paul, pp. 91–130.

[86] Serra, J. (1982). *Image Analysis and Mathematical Morphology*. Academic Press, London.

[87] Serra, J. ed. (1988). *Image Analysis and Mathematical Morphology*, Vol. 2: Theoretical Advances. Academic Press, London.

[88] Shepp, L. & Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography, *IEEE Transactions on Medical Imaging* **2**, 113–122.

[89] Silverman, B.W., Jones, M.C., Wilson, J.D. & Nychka, D.W. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion), *Journal of the Royal Statistical Society, Series B* **52**, 271–324.

[90] Skilling, J. & Bryan, R.K. (1984). Maximum entropy image reconstruction: general algorithm, *Monthly Notices of the Royal Astronomical Society* **211**, 111–124.

[91] Slayter, E.M. & Slayter, H.S. (1992). *Light and Electron Microscopy*. Cambridge University Press, Cambridge.

[92] Snyder, D. & Miller, M. (1985). The use of sieves to stabilize images produced with the EM algorithm for emission tomography, *IEEE Transactions on Nuclear Science* **32**, 3864–3872.

[93] Soille, P. & Rivest, J.-F. (1992). Principles and applications of morphological image analysis. Workshop Lecture Notes from 11th IAPR International Conference on Pattern Recognition.

[94] Stoyan, D. (1990). Stereology and stochastic geometry, *International Statistical Review* **58**, 227–242.

[95] Stoyan, D., Kendall, W.S. & Mecke, J. (1987). *Stochastic Geometry and Its Applications*. Wiley, Chichester.

[96] Tang, Y.T. & Suen, C.Y. (1993). Image transformation approach to nonlinear shape restoration, *IEEE Transactions on Systems, Man and Cybernetics* **23**, 155–171.

[97] Thompson, A.M., Brown, J.C., Kay, J.W. & Titterington, D.M. (1991). A study of methods of choosing the smoothing parameter in image restoration by regularization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 326–339.

[98] Vardi, Y. & Lee, D. (1993). From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems, *Journal of the Royal Statistical Society, Series B* **55**, 569–612.

[99] Vincent, L. & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 583–598.

[100] Weir, I.S. & Green, P.J. (1994). Modelling data from single photon emission computed tomography, in Statistics and Images, Vol. 2, K.V. Mardia, ed. Carfax, Abingdon, pp. 313–338.

[101] Zahn, C.T. & Roskies, R.Z. (1972). Fourier descriptors for plane closed curves, *IEEE Transactions on Computers* **21**, 269–281.

C.A. GLASBEY, M. BERMAN & H.M. HUDSON

Immunotoxicology

Immunotoxicology is a speciality of toxicology aimed at the detection, quantification, and interpretation of xenobiotic-induced direct and indirect alterations (stimulatory and/or suppressive) in the immune system and the resulting effects on morbidity (incidence of infection, duration of infection, incidence of tumors, etc.) and mortality. The immune system is a highly complicated network of lymphoid cells, nonlymphoid cells, soluble factors and regulatory molecules which protect humans from foreign substances and disease. The assays used in immunotoxicology (*see* **Bioassay**) can be divided into *in vivo* assays (in living animals) and *in vitro* assays (in cultured cells), and these can be further divided into immune function assays (measuring the responsiveness of the immune system to stimulation) and host resistance assays (measuring the ability of the immune system to protect the host from infectious agents and neoplasia) [4]. All host resistance and some immune function assays are done *in vivo*, with the remaining assays being done *in vitro* or through a combination of *in vitro* and *in vivo* methods [6]. In some cases, the same or similar immune function assays can be done in both *in vivo* and *in vitro* settings.

Host Resistance Assays

The most obvious endpoint for a host resistance assay is survival. These assays tend to be short-term and are performed in a two-step process [1, 8, 9]. In the first step, animals are exposed to a xenobiotic, generally using three dose groups and a control group. After a brief waiting period, the animals are then exposed to a carefully titrated concentration of some infectious agent (e.g. influenza virus), resulting in some known degree of mortality in the population (usually targeted for 20% in 4 days). The resulting data (the number dead out of the number exposed) are generally analyzed using pairwise comparisons (**Fisher's exact test** or its equivalent), sometimes accounting for multiple comparisons, and, when the study supports it, trend analysis (the Cochran–Armitage linear trend test). Seldom does the analysis include an analysis of the impact of titration variability in administering the infectious agent, a common problem.

More recent work utilizing infectious agents has moved away from mortality as an endpoint and has focused on body burden or tissue load of the infectious agent. Analysis of these assays is generally through the use of normality based statistical methods (e.g. **analysis of variance** (ANOVA)).

Other host resistance assays are similar to the infectious agent assays described above in that the endpoint can be viewed as a survival endpoint. An example would be administration of live tumor cells into the animal (PYB6 or B16F10 assays) with counts of the number of animals with and without tumors after a prescribed period. This type of assay is conducted in the same manner as the infectious agent survival assay described above, with administration of the xenobiotic followed by administration of the live tumor cells.

Another way in which this assay is examined is to count the number of tumors appearing in the animals. Here, the usual analysis is a comparison of the mean numbers of tumors in each dose group via a t test or a similar method. Seldom are these data analyzed by more complicated methods useful for count data such as **Poisson regression**, although there is some use of the Freeman–Tukey transformation method (*see* **Multinomial Distribution**). This is an area for further statistical development.

Immune Function Assays

Immune function assays are generally used to measure the functional competency of the immune system for dealing with antigenic response. In animals, these assays can be conducted *in vivo* by exposing animals to a xenobiotic followed by an antigen and then, following sacrifice, studying key components of the immune system (e.g. the numbers of antibodies producing B-cells in the spleen [2]). In addition, for some species it is possible to perform immune function tests in peripheral blood lymphocytes, either through exposing the host to the xenobiotic and then removing blood and performing the assay or by removing blood and doing the entire assay *in vitro*. These types of studies can also be carried out in exposed and control human populations. In most cases, the data derived from these assays are count data with very high numbers (e.g. the number of plague-forming cells in a Petri dish following stimulation of lymphocytes with sheep red blood cell antigen) and they are analyzed

under the basic assumption of normality (t tests and ANOVA) or through the use of similar **nonparametric** analyses. In some cases, the counts are converted to ratios (B-cells per million spleen cells) to control for fluctuations in physiology. Finally, some immune system markers are simple organ sizes (e.g. thymus weight, spleen weight, and cell counts) which are analyzed via assumptions of normality or **lognormality**.

In most of the analyses of immune function assays, care is taken to control for **multiple comparisons** (usually using Dunnett's' method) and, as with the analysis of host resistance assays, when data on dose–response are available, analyses for trend are common (e.g. **linear regression** and/or Jonkheere's test; *see* **Nonparametric Methods**).

Immunotoxicity and Risk Assessment

There have been considerable efforts in the past few years to develop methods to apply findings from immunotoxicity to the assessment of risks from exposure to xenobiotics (*see* **Risk Assessment**). The major challenge here is to synthesize a large array of assays into a single standard. There has been some work in this area, focusing on the relationship between immune function assays and host resistance assays [3, 5, 7], in which formal regression methods have been used. However, as general area of research, the utility of immunotoxicology for setting exposure standards is still emerging. One area of keen interest is the use of mechanistic models of immune function and response as a tool for understanding alterations due to xenobiotics.

References

[1] Luster, M.I., Germolec, D.R., Bruccoleri, A. & Simeonova, P.P. (1997). Immunotoxicological methods and applications: animal models, in *Comprehensive Toxicology*, I.A. Sipes, C.A. McQueen & A.J. Gandolfi, eds. Elsevier Science, Oxford.

[2] Luster, M.I., Germolec, D.R., Kayama, F., Rosenthal, G.J., Comment, C.E. & Wilmer, J.L. (1995). Approaches and concepts in immunotoxicology, in *Experimental Immunotoxicology*, R. Smialowicz & M. Holsapple, eds. CRC Press, Boca Raton, pp. 103–123.

[3] Luster, M.I., Portier, C., Pait, D.G., White, K.L., Gennings, C., Munson, A.E. & Rosenthal, G.J. (1992). Risk assessment in immunotoxicology. I. Sensitivity and predictability of immune tests, *Fundamental and Applied Toxicology* **18**, 200–210.

[4] Luster, M.I., Munson, A.E., Thomas, P., Holsapple, M.P., Fenters, J., White, K., Lauer, L.D., Germolec, D.R., Rosenthal, G.J. & Dean, J.H. (1988). Development of a testing battery to assess chemical-induced immunotoxicity: National Toxicology Program's guidelines for immunotoxicity evaluation in mice, *Fundamental and Applied Toxicology* **10**, 2–19.

[5] Luster, M.I., Portier, C., Pait, D.G., Rosenthal, G.J., Germolec, D.R., Corsini, E., Blaylock, B.L., Pollock, P., Kouchi, Y., Craig, W., White, K.L., Munson, A.E. & Comment, C.C. (1993). Risk assessment in immunotoxicology. II. Relationships between immune and host resistance tests, *Fundamental and Applied Toxicology* **21**, 71–82.

[6] Munson, A.E. & LeVier, D. (1995). Experimental design in immunotoxicology, in *Methods in Immunotoxicology*, G. Burleson, J.H. Dean & A.E. Munson, eds. Wiley-Liss, New York, pp. 11–24.

[7] Selgrade, M.K., Daniels, M.J. & Dean, J.H. (1992). Correlation between chemical suppression of natural killer cell activity in mice and susceptibility to cytomegalovirus: rationale for applying murine cytomegalovirus as a host resistance model and for interpreting immunotoxicity testing in terms of risk of disease, *Journal of Toxicology and Environmental Health* **37**, 123–137.

[8] Thomas, P.T. & Sherwood, R. (1995). Host resistance models in immunotoxicology, in *Experimental Immunotoxicology*, R. Smialowicz & M. Holsapple, eds. CRC Press, Boca Raton, pp. 29–46.

[9] Vos, J.G., Smialowicz, R.J. & van Loveren, H. (1994). Animal models for assessment, in *Immunotoxicology and Immunopharmacology*, 2nd Ed. J. Dean, M. Luster, I. Kimber & A. Munson, eds. Raven Press, New York, pp. 19–30.

(*See also* **Animal Screening Systems**; **Dose–Response Models in Risk Analysis**)

CHRISTOPHER J. PORTIER & DORI GERMOLEC

Importance Sampling

Importance sampling is an extremely useful statistical technique with a long history. In the last ten to fifteen years, driven by the popularity of **Monte Carlo** and computationally intensive methods, it has been enriched and extended in many exciting directions. Importance sampling is based on a very simple idea. When one wants to estimate population means or expectations of random variables with respect to a distribution of interest, referred to as the target distribution, but samples are drawn from another distribution, referred to as the trial distribution, importance sampling assigns weight to the samples to make the necessary adjustments. There are three possible reasons for using samples from a trial distribution instead of the target distribution.

1. ***Variance** reduction*: the trial distribution is deliberately chosen so that the estimate obtained is actually superior to an estimate based on samples drawn from the target distribution with the same sample size.
2. *Feasibility and convenience*: drawing samples from the target distribution can be very difficult or outright impossible.
3. *Reusing and mixing samples*: sometimes means and expectations with respect to multiple distributions are of interest, and importance sampling allows us to use samples drawn from one distribution to obtain estimates of expectations under various distributions. When we have multiple sets of samples drawn from different trial distributions, an obvious approach would be to first obtain an appropriate estimate from each set of samples and then compute a weighted average of them as the overall estimate. This, however, is usually not the optimal way of using the samples. The most efficient estimates are usually obtained by treating the multiple sets of samples as a single set drawn from a mixture distribution [7]. Some of the most interesting developments in theory and applications occur in this area.

Importance sampling is used extensively in sample surveys, where the samples are concrete units such as people or households [8]. However, many of the recent developments in applications and methodology are associated with Monte Carlo estimation, where samples are computer generated. Interestingly, while many of the techniques developed end up to be applicable to a wide range of applications, human genetics, particularly pedigree analysis, has been the driving force behind many of them. Also, it was demonstrated recently that a novel recursive estimation technique developed for analyzing coalescence data in **population genetics** can also be considered as an elaborate form of importance sampling.

The basic theory behind importance sampling is as follows. Suppose x is random with outcome space Ω, $g(x)$ is some function of x, and of interest is

$$\mu = \mathrm{E}_{p_1}[g(x)] = \int_{x\in\Omega} g(x)p_1(x)\,\mathrm{d}x, \qquad (1)$$

the expectation of $g(x)$ with respect to the probability density $p_1(x)$, i.e. $p_1(x)$ is the target density. If x is discrete, then $p_1(x)$ is a probability mass function, and the integral in (1) is replaced by a summation. If it is not feasible to compute μ analytically, but Monte Carlo samples, $x_1, \ldots, x_n$, can be generated from the distribution $p_1(x)$, then

$$\frac{\sum_i g(x_i)}{n} \qquad (2)$$

is an unbiased estimate of μ. Now, suppose the samples $x_1, \ldots, x_n$ are not drawn from $p_1(x)$, but instead are generated from some trial $p_0(x)$, whose support contains the support of $p_1(x)$. Since μ can be rewritten as

$$\mathrm{E}_{p_0}\left\{\left[\frac{p_1(x)}{p_0(x)}\right]g(x)\right\} = \int_{x\in\Omega}\left[\frac{p_1(x)}{p_0(x)}\right] \times g(x)p_0(x)\,\mathrm{d}x = \int_{x\in\Omega} g(x)p_1(x)\,\mathrm{d}x, \qquad (3)$$

μ can be estimated by either

$$\tilde{\mu} = \frac{\sum_i w(x_i)g(x_i)}{n}, \qquad (4)$$

where $w_i = p_1(x_i)/p_0(x_i)$ is the importance sampling weight of x_i, or

$$\hat{\mu} = \frac{\sum_i w_i g(x_i)}{\sum_i w_i} = \sum_i w_i^* g(x_i), \qquad (5)$$

where $w_i^* = w_i/(\sum_j w_j)$ are normalized weights that sum to 1. While $\tilde{\mu}$ is the natural **unbiased** estimate, $\hat{\mu}$ is in the form of a ratio estimate that usually is biased in the technical sense. However, the **bias** is of order $1/n$, whereas the standard deviation is of order $1/\sqrt{n}$. Hence, the contribution of the bias to the mean-squared-error is often negligible for large n. There are two possible reasons for using $\hat{\mu}$ instead of $\tilde{\mu}$. If w_i is strongly positively correlated with $g(x_i)$, then $\hat{\mu}$ can have a substantially smaller variance than $\tilde{\mu}$, and hence is preferred. Often that is not the case, but we still use $\hat{\mu}$ because $\tilde{\mu}$ is not computable. Note that while evaluating $\tilde{\mu}$ requires the exact values of the w_i, $\hat{\mu}$ can be computed if the w_i are known only up to a multiplicative constant (because the constant will be canceled out in w_i^*), something that occurs in many new applications of importance sampling. For reference, Hammersley & Handscomb [5] contains many useful results with regard to Monte Carlo **simulations**, importance sampling and the choice of $p_0(x)$. On a purely theoretical level, the ideal $p_0(x)$ to use in conjunction with $\tilde{\mu}$ is $cp_1(x)/g(x)$ for some constant c. Notice that with such a $p_0(x)$, $[p_1(x)/p_0(x)]g(x) = 1/c$, which means $\tilde{\mu}$ has zero variance and c must be equal to $1/\mu$. Hence, while this choice of $p_0(x)$ may serve as a guide, implementing it exactly implies that μ is computed directly and it is no longer a Monte Carlo problem. In general, the choice of $p_0(x)$ always involves a compromise between the variance of the estimate, the ease of generating the samples and the work needed to compute the estimate given the samples. In practice, the problem is further complicated when the expectations of multiple gs, instead of a single one, are of interest. For $\hat{\mu}$ and some extensions of it, expressions of large sample variances and some useful results for comparisons of different estimates can be found in Kong et al. [12]. A rule of thumb is that the efficiency of $\hat{\mu}$ tends to be inversely proportional to one plus the variance of the importance sampling weight w_i.

An example of importance sampling being used in pedigree analysis because of reason 1 is described in Kong et al. [13]. There a trial distribution is deliberately chosen to oversample the tail of the target distribution, which is the distribution of a test statistic under the **null hypothesis** so as to obtain a much more efficient estimate of the P-value. There $\tilde{\mu}$ can be computed and is superior to $\hat{\mu}$. However, most recent developments in methodology are focused on situations falling under scenario 2, where it is difficult to draw samples from the target distribution, and often $\hat{\mu}$ is used instead of $\tilde{\mu}$. For resolving the difficulty of directly obtaining samples from the target distribution, it is important to acknowledge the existence of an alternative strategy. **Markov Chain Monte Carlo** (MCMC) techniques, which were first invented by physicists, have enjoyed unequalled attention in the last two decades in the statistics community and much progress has been made. There a **Markov chain** is constructed with the target distribution as its stationary distribution. In pedigree analysis, examples of how this is done can be found in Lange & Sobel [15], Heath [6], Jensen & Kong [10], and Thompson [20]. Under ideal conditions, this technique allows us to obtain samples that have essentially the right distribution and are assigned equal weights, but the samples are correlated instead of independent, in contrast with importance sampling, which generates independent but weighted samples. As a solution to reason 2, importance sampling and MCMC are competitors and which one performs better depends on individual applications. However, even when samples are generated by MCMC, importance sampling can often be used to expand the utility of the samples because of reason 3, and the two techniques become complementary. One good example of this is Monte Carlo expectation–maximization algorithm **EM** [21], where Monte Carlo estimates are used to perform the expectation step of the algorithm. Since EM is an iterative procedure and the expectation of the score function has to be calculated for a series of parameter values, importance sampling allows us to reuse samples generated with respect to one parameter value for subsequent steps when expectations have to be calculated for other parameter values, regardless of whether the original samples are direct samples, importance sampling samples, or MCMC samples. Irwin et al. [9] demonstrate how this can be done for estimating the recombination fractions (*see* **Linkage Analysis, Model-based**) among a set of **polymorphic markers** using pedigree data. This application also serves as an example of another way that MCMC and importance sampling can potentially be used together. Because of interference, the crossover/recombination process is not a Poisson process, and the crossover events of two neighboring but nonoverlapping regions are not independent. Even with MCMC, it is not easy to generate samples with

the desired distribution. One possible solution is to use MCMC techniques to generate samples based on a no-interference model that is as close to the desired model as possible, and then use importance sampling to do the final adjustment.

Because many of the interesting new applications occur in the area of missing data problems and **likelihood** calculations, we will go into these in more detail. Before doing that, it is worth noting that while μ is the expectation of $g(x)$ under $p_1(x)$, it is also the expectation of $g'(x) = g(x)p_1(x)/p_0(x)$ under $p_0(x)$. Because of that, while the target distribution $p_1(x)$ may have a natural meaning in a specific application, on a certain level, it does not have any special relevance independent of $g(x)$. Indeed, in many of the modern applications of importance sampling, the role of the target distribution is blurred and mathematically the problem is simply one of Monte Carlo integration.

Given any two distributions $p_0(x)$ and $p_1(x)$, let $q_0(x) = c_0 p_0(x)$ and $q_1(x) = c_1 p_1(x)$ for some constants c_0 and c_1. Obviously

$$\mathrm{E}_{p_0}\left[\frac{q_1(x)}{q_0(x)}\right] = \int \frac{c_1 p_1(x)}{c_0 p_0(x)} p_0(x)\, \mathrm{d}x = \left(\frac{c_1}{c_0}\right)\int p_1(x)\, \mathrm{d}x = \frac{c_1}{c_0}. \tag{6}$$

With samples $x_i, i = 1, \ldots, n$, from $p_0(x)$,

$$\left(\frac{1}{n}\right)\sum_i \left[\frac{q_1(x_i)}{q_0(x_i)}\right] \tag{7}$$

is an unbiased estimate of c_1/c_0. This rather simple setting actually incorporates one of the most interesting applications of Monte Carlo estimation. In a missing data problem, let y denote observed data and of interest is the likelihood function

$$L(\theta) = p(y|\theta) \tag{8}$$

as a function of θ. Suppose we cannot compute $p(y|\theta)$ directly, but if y is augmented by x, which is not actually observed, then $p(x, y|\theta)$ can be easily computed for any value of x and θ. Suppose we have samples $x_i, i = 1, \ldots, n$, drawn from some distribution $p_0(x)$ and we can compute $p_0(x_i)$. Let $p_1(x) = p(x|y, \theta)$ and $q_1(x) = p(x, y|\theta)$ so that $c_1 = q_1(x)/p_1(x) = p(y|\theta) = L(\theta)$, and let $q_0(x) = p_0(x)$ so that $c_0 = 1$. It follows from (7) that

$$\left(\frac{1}{n}\right)\sum_i \left[\frac{q_1(x_i)}{q_0(x_i)}\right] = \left(\frac{1}{n}\right) \times \sum_i \left[\frac{p(x_i, y|\theta)}{p_0(x_i)}\right] \tag{9}$$

is an unbiased estimate of $c_1 = L(\theta)$. Now consider an alternative scenario where we cannot compute $p_0(x_i)$, but we can compute $q_0(x_i) = c_0 p_0(x_i)$ for some unknown constant c_0. While the likelihoods cannot be estimated directly here, they can be estimated up to an unknown constant that allows us to estimate likelihood ratios. Specifically, if θ_1 and θ_2 are two possible values of θ, then $L(\theta_2)/L(\theta_1)$ can be estimated by

$$\frac{\sum_i [p(x_i, y|\theta_2)/q_0(x_i)]}{\sum_i [p(x_i, y|\theta_1)/q_0(x_i)]}. \tag{10}$$

This important observation originated from Ott [18] and Geyer & Thompson [3]. One typical choice of $p_0(x)$ will be $p(x|y, \theta_0)$ for some value θ_0 of θ, so that $q_0(x) = p(x, y|\theta_0)$ and $c_0 = p(y|\theta_0) = L(\theta_0)$. Note that (9) can be considered as a form of $\tilde{\mu}$ or simply a form of (2), while (10) is a ratio estimate. By rearrangement of the terms, one can show that (9) can be considered as $\hat{\mu}$ with $p_1(x) = p(x|y, \theta_1)$, $p_0(x) = p(x|y, \theta_0)$, and $g(x) = p(x, y|\theta_2)/p(x, y|\theta_1)$. In particular,

$$w_i = \frac{p(x_i|y, \theta_1)}{p(x_i|y, \theta_0)} = \frac{p(x_i, y|\theta_1)}{p(x_i, y|\theta_0)}\frac{p(y|\theta_0)}{p(y|\theta_1)}, \tag{11}$$

where $\{p(y|\theta_0)/p(y|\theta_1)\}$ is the unknown constant that cancels in

$$w_i^* = w_i \Big/ \left(\sum_j w_j\right) = \frac{p(x_i, y|\theta_1)/p(x_i, y|\theta_0)}{\sum_j \{p(x_i, y|\theta_1)/p(x_i, y|\theta_0)\}}. \tag{12}$$

From (10) and its relationship with (6), one can see that the problem of estimating **likelihood ratios** mathematically falls within the framework of estimating ratios of normalizing constants of functions [e.g. c_0^{-1} and c_1^{-1} are the normalizing constants for $q_0(x)$ and $q_1(x)$, respectively], which happens to have very

broad applications [17]. In the case here, of interest is the whole likelihood function and so, in a sense, of interest are an infinite number of likelihood ratios. Estimates of the ratios can often be greatly improved, meaning that the variance can be greatly reduced, if samples are simulated from multiple trial distributions [1], and the combined samples are treated as draws from a mixture distribution. In the case here, the multiple trial distributions will naturally be $p(x|y, \theta)$ for different values of θ. For importance sampling in general, the step of combining samples is straightforward if the $p_0(x_i)$ can be calculated easily for each of the trial distributions, but in this setting, by design, $p_0(x_i)$ is only known up to a constant and this constant is different for each of the trial distributions. In a way, the problem looks amusingly circular – we want to combine the samples to better estimate the ratios of the normalizing constants, but to properly combine the samples we need to know the ratios of the normalizing constants. Fortunately, this problem actually has a solution and the search for this solution leads to development of new methods [2, 17] and new theory [14].

It is usually the case that when samples from $p_0(x)$ are generated using MCMC, $p_0(x_i)$ can at best be computed up to a constant. By contrast, if independent samples are drawn directly from $p_0(x)$, then most likely $p_0(x_i)$ can be computed. However, the distributions one can draw from directly are much more limited than those that can be drawn from using MCMC. One way to enrich the class of distributions one can draw from directly is a technique called sequential imputations. The idea is as follows. Suppose the observed data y and the missing data x can each be partitioned as $y = [y(1), y(2), \ldots, y(T)]$ and $x = [x(1), x(2), \ldots, x(T)]$. While it is difficult to draw directly from $p(x|y)$, suppose we can draw from $p[x(1)|y(1)]$ and, for $t = 2, \ldots, T$, it is possible to draw $x(t)$ from

$$p[x(t)|y(1), \ldots, y(t), x(1), \ldots, x(t-1)] \quad (13)$$

for any fixed values of $x(1), \ldots, x(t-1)$. Instead of drawing a sample $x_i[x_i(1), x_i(2), \ldots, x_i(T)]$ in one step, sequential imputation simulates the components $x_i(t), t = 1, \ldots, T$, one at a time with each component simulated conditional on values already drawn for previous components. With $p(x|y)$ considered as the target distribution $p_1(x)$, the trial distribution is

$$p_0(x) = p[x(1)|y(1)] \times \prod_{t=2}^{T} p[x(t)|y(1), \ldots, y(t), x(1), \ldots, x(t-1)]. \quad (14)$$

As is true in many other cases, this is a technique that had been invented many times over by researchers in different areas before it was formalized and given a name. Often there is a parameter θ involved in the conditional distribution of x given y, and $p(x|y)$ can either be, in a Bayesian setup, a distribution with θ analytically integrated out [11], or it can be the distribution corresponding to some chosen value θ^* of θ. In genetics, Irwin et al. [9] used it to perform multipoint linkage analysis, while Stephens & Donnelly [19] used it to assist inference in molecular population genetics and also pointed out that an earlier recursive simulation method invented by Griffiths & Tavaré [4] could also be viewed as a form of sequential imputation. Some recent developments can be found in Liu et al. [16].

References

[1] Gelman, A. & Meng, X.L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling, *Statistical Science* **13**, 163–185.

[2] Geyer, C.J. (1994). Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo. Technical Report 568. School of Statistics, University of Minnesota.

[3] Geyer, C.J. & Thompson, E. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 657–699.

[4] Griffiths, R.C. & Tavaré, S. (1994). Simulating probability distributions in the coalescent, *Theoretical Population Biology* **46**, 131–159.

[5] Hammersley, J.M. & Handscomb, D.C. (1979). *Monte Carlo Methods*, Chapman & Hall, New York.

[6] Heath, S.C. (1997). Markov Chain Monte Carlo segregation and linkage analysis for oligogenic models, *American Journal of Human Genetics* **61**, 748–760.

[7] Hesterberg, T.C. (1995). Weighted average importance sampling and defensive mixture distributions, *Technometrics* **37**, 185–194.

[8] Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**, 663–685.

[9] Irwin, M., Cox, N. & Kong, A. (1994). Sequential imputation for multilocus linkage analysis, *Proceedings of the National Academy of Sciences* **91**, 11 684–11 688.

[10] Jensen, C.S. & Kong, A. (1999). Blocking Gibbs and linkage analysis, *American Journal of Human Genetics* **65**, 885–901.

[11] Kong, A., Liu, J.S. & Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems, *Journal of the American Statistical Association* **89**, 278–288.

[12] Kong, A., Liu, J.S. & Wong, W.H. (1997). The cross-match estimate and importance sampling, *Annals of Statistics* **25**, 2410–2432.

[13] Kong, A., Frigge, M., Irwin, M. & Cox, N. (1992). Importance sampling, I. Computing multimodel *p*-values in linkage analysis, *American Journal of Human Genetics* **51**, 1413–1429.

[14] Kong, A., McCullagh, P., Meng, X.L. & Nicolae, D. (2001). Statistical Methods for Monte Carlo Integration. Technical report. Department of Statistics, University of Chicago.

[15] Lange, K. & Sobel, E. (1991). A random walk method for computing genetic location scores, *American Journal of Human Genetics* **49**, 1320–1334.

[16] Liu, J.S., Chen, R. & Wong, W.H. (1999). Rejection control and sequential importance sampling, *Journal of the American Statistical Association* **92**, 1022–1031.

[17] Meng, X.L. & Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical explanation, *Statistica Sinica, Series B* **6**, 831–860.

[18] Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees, *American Journal of Human Genetics* **31**, 161–175.

[19] Stephens, M. & Donnelly, P. (2000). Inference in molecular population genetics (with discussion), *Journal of the Royal Statistical Society, Series B* **62**, 605–655.

[20] Thompson, E. (2000). MCMC estimation of multilocus genome sharing and multipoint gene location scores, *International Statistics Review* **68**, 53–73.

[21] Wei, G.C.G. & Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm, *Journal of the American Statistical Association* **85**, 699–704.

A. KONG

Inbreeding

Individuals with common ancestors are said to be related, and their offspring are inbred. If no further qualifications are made, then all humans are both inbred and related to everyone else simply because the population is finite. Each of us has two parents, and if we had four grandparents, eight grandparents, 16 great-grandparents, and so on, it would take only a few hundred years back in time before we would have more ancestors than there were people living on the planet at that time. Obviously our parents have some ancestors in common, but conventional definitions of inbreeding refer only to children whose parents are related through people in the past few generations.

Inbreeding in Pedigrees

The genetic consequences of inbreeding follow directly from basic Mendelian principles (*see* **Mendel's Laws**). For each **gene**, an individual receives two alleles, one from each parent, and is generally equally likely to transmit either of these two alleles to a child. The random element in such transmission means that statements about inbreeding are usually expressed as probabilities. Because related people share ancestors, there is a chance that they receive copies of the same allele from those ancestors. Half-sibs, for example, may each receive copies of the same allele from their one common parent. Because this common parent has two alleles, there is a probability of one-half that the half-sibs receive alleles that are identical by descent (ibd). There is a further one-half probability that they would each transmit these ibd alleles to an offspring. A child of half-sibs, therefore, would have an inbreeding coefficient, F, of one-eighth.

A general approach is to specify some initial or reference population, in which all members are assumed to be unrelated, and inbreeding is then measured relative to that generation. It is generally accepted, for example, that Finland was settled by a relatively small group of people about 4000 years ago. It would be convenient to quantify inbreeding, for a random member of the presently living descendants of those founders, as the probability that the person receives two alleles that trace back to a single allele among the founders. Alleles that trace to distinct founding alleles will be considered not ibd.

If the common parent in the half-sib example itself had related parents, and had an inbreeding coefficient of F, then one-half the time it would transmit copies of the same allele to two offspring and the other half of the time it would transmit alleles that had probability F of being ibd. Such arguments lead to "path-counting" equations for inbreeding coefficients. If the parents of individual I have common ancestors A, with inbreeding coefficients F_A, and if there are n_A people in the loop from one parent through A and back to the other parent, then the inbreeding coefficient of I is

$$F_I = \sum_A \left(\frac{1}{2}\right)^{n_A} (1 + F_A).$$

In the half-sib case, the common parent A is the only common ancestor and $n_A = 3$ so that $F_I = 1/8$ as before. Full sibs have two parents in common, so the inbreeding coefficient of their children would be $1/8 + 1/8 = 1/4$. First cousins have four distinct parents, two of whom are full sibs, so they have two grandparents in common ($n = 5$ for each). The inbreeding coefficient of the children of first cousins is therefore 1/16, and this is the maximum amount of inbreeding tolerated by most marriage laws.

Just as the concepts of inbreeding and relatedness are closely connected, so are the probabilities of these events. The usual measure of relatedness for individuals X and Y, the coancestry coefficient (also called the coefficient of kinship) θ_{XY}, is defined as the probability that two alleles, one taken at random from the same locus of each of X and Y, are ibd. This definition provides a value of 1/4 for full sibs, 1/8 for half sibs and 1/16 for first cousins. If individuals X and Y have a child I, then

$$F_I = \theta_{XY}.$$

Although there is not complete independence among different genes, an inbreeding coefficient of F can be interpreted as meaning that a fraction F of the genes in such an individual has two alleles that are ibd. For alleles that are both harmful and recessive, such as the ΔF508 allele responsible for most cases of cystic fibrosis, inbreeding increases the proportion of people with the harmful trait by virtue of having two copies of the deleterious allele, not masked by a normal allele. The ΔF508 allele in Caucasian populations has a (relative) frequency of about $p = 0.05$. Among individuals whose parents

are unrelated, the probability of having two copies of the allele, and therefore having cystic fibrosis, is about $p^2 = 0.0025$. Among people whose parents are cousins, however, $(1 - F) = 15/16$ of the time the probability is p^2, but $F = 1/16$ of the time it is the higher value of p. The total probability is more than doubled, to 0.0063. In general, the probability that an individual with an inbreeding coefficient of F is homozygous aa for allele a that has a frequency of p_{a} is

$$P_{\mathrm{aa}} = p_{\mathrm{a}}^2 + F p_{\mathrm{a}}(1 - p_{\mathrm{a}}). \tag{1}$$

The increased homozygosity brought about by inbreeding must be accompanied by an equivalent decrease in **heterozygosity**. If $\overline{\mathrm{a}}$ indicates an allele different from a, then for inbred individuals

$$P_{\mathrm{a}\overline{\mathrm{a}}} = 2 p_{\mathrm{a}} p_{\overline{\mathrm{a}}}(1 - F). \tag{2}$$

Homozygotes have two alleles that have the same chemical composition, and so are identical in state. Such alleles may or may not be ibd. Heterozygotes have alleles that are not identical in state, and these alleles cannot be ibd.

There is often interest in the joint probability $P_{\mathrm{a,a}}$ with which two individuals carry a specific allele a. It may be that one person, X, is alleged to be the father of a child but some other person Y is actually the father, and a is the allele known to have been received by the child from its father. The probability with which an allele chosen at random from one individual is ibd to one from the other is just the coancestry, so

$$P_{\mathrm{a,a}} = p_{\mathrm{a}}^2 + \theta p_{\mathrm{a}}(1 - p_{\mathrm{a}}) \tag{3}$$

when the two individuals have coancestry θ. There is a more complicated expression if the individuals are also inbred.

Inbreeding in Populations

Equations (1) and (2) have been derived for inbred individuals, where F is necessarily positive. Alternatively, they could be used to relate **genotypic** frequencies P_{aa} and $P_{\mathrm{a}\overline{\mathrm{a}}}$ in some population to allele frequencies p_{a} and $p_{\overline{\mathrm{a}}}$ in the same population, although it is then conventional to use the symbol f in place of F. A general treatment allows for variation among fs for different loci, and for different genotypes at a locus. Writing the frequency of allele a_i as p_i, the frequency of $\mathrm{a}_i\mathrm{a}_i$ homozygotes as P_{ii} and the frequency of $\mathrm{a}_i\mathrm{a}_j$ heterozygotes as P_{ij}:

$$P_{ii_k} = p_{i_k}^2 + \sum_{j \neq i} f_{ij} p_{i_k} p_{j_k},$$

$$P_{ij_k} = 2 p_{i_k} p_{j_k}(1 - f_{ij}), \quad i \neq j,$$

where the subscript k emphasizes that the equations hold for some particular population k.

Inferences about the f parameters are based on a model of repeated sampling from the population, and if this statistical sampling is random, the **multinomial distribution** is appropriate for large populations. Population data provide direct estimates of the genotypic frequencies, and **maximum likelihood** estimates for the p_{a}s and fs, based on sample genotype frequencies $\tilde{P}_{ijk}$ are

$$\hat{p}_{i_k} = \tilde{P}_{ii_k} + \tfrac{1}{2} \sum_{j \neq i} \tilde{P}_{ij_k},$$

$$\hat{f}_{ij_k} = 1 - \frac{\tilde{P}_{ij_k}}{2 \hat{p}_{i_k} \hat{p}_{j_k}}.$$

If a common value f, the "within-population inbreeding coefficient" is assigned to all the f_{ij}s, then iterative procedures are needed for maximum likelihood estimation. These procedures typically produce estimates of the order of 0.001 for human populations. The quantity f was written as F_{IS} by Wright [2], referring to the relation of alleles within individuals (I) relative to a subpopulation (S).

For a specific population, the within-population inbreeding coefficient f quantifies the excess homozygosity over that expected for random mating populations. **Population-genetic** analyses are likely to be concerned with the evolutionary processes that lead to extant populations, and therefore recognize that the present population itself results from genetic sampling. Because of the random processes involved in the choice of alleles transmitted between generations, as well as in other evolutionary forces such as selection and mutation, the genetic composition of a population cannot be specified with certainty over time. Instead, probabilistic models are needed.

Taking expectations $\mathcal{E}$ over populations (or over the evolutionary process) the frequency of allele a_i is

written as p_i, and of genotype $a_i a_j$ is written as P_{ij}:

$$\mathcal{E}p_{i_k} = P_i,$$
$$\mathcal{E}P_{ij_k} = P_{ij}.$$

At this total-expectation level,

$$P_{ii} = p_i^2 + Fp_i(1 - p_i),$$
$$P_{ij} = 2p_i p_j(1 - F), \quad i \neq j, \tag{4}$$

where F is the "total inbreeding coefficient." Evidently, then (1) and (2) refer in expectation to all individuals with the pedigree leading to a specific F value – even though any particular individual is either inbred or not – and invoke the expected allele frequency rather than the frequency for a specific population.

The usual interpretation of (4) is that they apply as an average over populations. One application concerns a large population considered to consist of a number of subpopulations indexed by k. Equations (1) and (2) hold for the subpopulations, but (4) holds for the total. Any variation in p_{i_k} over subpopulations causes $\mathcal{E}p_{i_k}^2$ to exceed p_i^2, so that $P_{ii} > p_i^2$ even if $P_{ii_k} = p_{i_k}^2$. This result is known as the "Wahlund principle." Wright [2] wrote F as F_{IT}, referring to the relation of alleles within individuals (I) relative to the total population (T).

The quantity $p_{i_k}^2$ can be regarded as the probability of two alleles in population k both being of type a_i:

$$P_{i,i_k} = p_{i_k}^2.$$

Taking expectations over populations:

$$P_{i,i} = p_i^2 + \theta p_i(1 - p_i),$$

illustrating why Wright wrote θ as F_{ST}, for the relationship between alleles within subpopulations (S) relative to the total population (T). The three measures of inbreeding are related by

$$f = \frac{F - \theta}{1 - \theta}.$$

It needs to be stressed that $P_{i,i}$ is the joint probability of two alleles in the same subpopulation being ibd, averaged over all subpopulations. Estimation of the inbreeding and coancestry coefficients F and θ requires data from more than one population. Otherwise there is no knowledge of the variation in allele frequencies among populations. If there is random mating within populations, then two alleles have the same relationship whether they are in the same or different individuals, $F = \theta$, and $f = 0$.

One method of estimation, under the random mating assumption, is to compare allelic variation within and among populations. The two means squares are MSW for within and MSA for among. For allele a_i and samples of size n alleles from each of r populations:

$$\text{MSW} = \frac{n}{r(n-1)} \sum_k p_{i_k}(1 - p_{i_k}),$$
$$\text{MSA} = \frac{n}{r-1} \sum_k (p_{i_k} - \overline{p}_i)^2,$$

where

$$\overline{p}_i = \frac{1}{r} \sum_k p_{i_k}.$$

The **variance components** for allele frequencies within and between populations are

$$\sigma_w^2 = p_i(1 - p_i)(1 - \theta),$$
$$\sigma_b^2 = p_i(1 - p_i)\theta,$$

and θ can be estimated [1] as

$$\hat{\theta} = \frac{\text{MSA} - \text{MSW}}{\text{MSA} + (n-1)\text{MSW}}.$$

Conditional Probabilities

It is now possible to return to the disputed **paternity** example. If the alleged father has been typed and found to carry the obligate paternal allele a, then the quantity of interest is the conditional probability $P_{a|a}$ with which some other man also carries the allele. If this unknown and untested other man belongs to the same population as the alleged father,

$$P_{a|a} = \frac{P_{a,a}}{p_a}$$
$$= p_a + \theta(1 - p_a),$$

which applies as an average over all populations. The allele frequency could be estimated from a sample from the total population, as opposed to the particular population to which the two men belong. Data from several populations would be needed if θ

is to be estimated. If the two men belong to different populations, then

$$P_{a|a} = p_a.$$

Other Measures of Inbreeding

Identity coefficients describe other measures of inbreeding. These coefficients are the probabilities that sets of more than two alleles are ibd, and they are needed to express joint and conditional probabilities of genotypes, as opposed to alleles. They find use in questions of disputed identity where one person is found to have a particular genotype and is alleged to be the donor of some biological sample. This use is described in **statistical forensics**.

References

[1] Weir, B.S. & Cockerham, C.C. (1984). Estimating *F*-statistics for the analysis of population structure, *Evolution* **38**, 1358–1370.

[2] Wright, S. (1951). The genetical structure of populations, *Annals of Eugenics* **15**, 323–354.

B.S. WEIR

Incidence Density Ratio

The incidence density ratio is the ratio of the **incidence density** in one group to that in another group. The incidence density ratio approximates the **hazard ratio** if time intervals are small and can be estimated both from **cohort studies** and from **case–control studies** in which controls are selected by **density sampling**.

MITCHELL H. GAIL

Incidence Density

An incidence density is an **incidence rate** and can be used to estimate a **hazard rate**.

MITCHELL H. GAIL

Incidence Rate

The incidence rate is the number of persons who develop a disease of interest over a defined interval of time or age divided by the corresponding **person-years at risk** among members of the source population. Subjects are only "at risk" before they develop the disease of interest if, as is common, the incidence rate describes the rate of first occurrence of a disease. Usually, relatively short time intervals are used, compared with the timescale for development of disease, such as five-year intervals for a cancer incidence study. When individual follow-up data are not available to compute person-years at risk, the person-years are often estimated as the interval width times the population size at the midpoint of the interval. Synonyms for incidence rate include **incidence density** and person-years incidence rate. Incidence rate sometimes denotes a population **hazard rate**, rather than the estimate defined above. Sometimes the term incidence rate is used instead of **cumulative incidence rate**, but the concepts are distinct.

MITCHELL H. GAIL

Incidence–Prevalence Relationships

This article attempts a statistical view on the classical epidemiologic concepts of (age-specific) incidence and **prevalence**. Each individual's dynamics in the **Lexis diagram** is modeled by a simple three-state illness–death **stochastic process** in the age direction and individuals are recruited from a **Poisson process** in the time direction. Observable quantities are regarded as *estimators* of the *parameters* (incidence, prevalence, mortality, **mean** duration, etc.) of the statistical model.

The next section discusses increasingly complex versions of the classical epidemiologic relation

$$\text{prevalence} = \text{incidence} \times \text{duration},$$

and its generalization to age- and duration-specific incidence and mortality. Then some comments are provided on statistical techniques for estimating **incidence rates** from prevalence surveys, while the following section considers, conversely, the feasibility of estimating prevalence from information on incidence and mortality. The material is also relevant in the theory of **screening**, as briefly pointed out later.

A related topic *not* touched in this article is **inference** on mortality (or further morbidity) from follow-up of a **cross-sectional** sample, the so-called *prevalent cohort study*. This topic is treated in the articles **Delayed Entry** and **Biased Sampling of Cohorts in Epidemiology**.

Prevalence, Incidence, and Duration

Most – even rather elementary – textbooks in epidemiology contain versions of the statement

$$\text{prevalence} = \text{incidence} \times \text{duration}, \qquad (1)$$

see, for example, [19, pp. 65–66] or [10, pp. 64–66]. In broad generality, (1) is a conservation equation called Little's equation in queuing theory:

time-average number of units in the system
= arrival rule × average delay time per unit.

See Little [17] for the first general proof in the context of strictly stationary processes in steady state conditions and Ramalhoto et al. [25] for a comprehensive discussion.

In epidemiology, the archetypical situation concerns irreversible transitions between a healthy state H, a diseased state I, and the dead state D, simplest in the time- and age-homogeneous **Markov illness–death process** specified by intensities as follows:

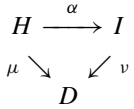

and fed by a stationary homogeneous **Poisson process** with (birth) intensity β. Here α is disease intensity for a healthy individual (the connection to the epidemiologic concept of disease incidence to be discussed below) and μ and ν are death intensities for healthy and diseased, respectively. Sometimes ν is called the **case fatality** rate or just lethality.

In this stochastic process our approach to prevalence is to imagine a **cross-sectional** sample taken at a particular time t, say $t = 0$. We may then calculate the expected number of healthy at $t = 0$ as

$$\int_0^\infty \beta \exp[-(\alpha+\mu)a]\,\mathrm{d}a = \frac{\beta}{\alpha+\mu}$$

since a person born at time $-a$ has probability $\exp[-(\alpha+\mu)a]$ of remaining alive and healthy until time 0; similarly the expected number of diseased at $t = 0$ is

$$\int_0^\infty \int_0^a \beta \exp[-(\alpha+\mu)y]\alpha \exp[-\alpha(a-y)]\,\mathrm{d}y\,\mathrm{d}a = \frac{\alpha\beta}{(\alpha+\mu)\nu}.$$

Under the present assumptions, disease duration is **exponentially** distributed with mean ν^{-1}. Definition of disease incidence requires more care. The intensity α refers to the healthy only, while *disease incidence in the population* may be defined as the rate of occurrence of new disease in the whole population. This is

$$\beta \int_0^\infty \exp[-(\alpha+\mu)\alpha]\,\mathrm{d}a = \frac{\beta\alpha}{\alpha+\mu}$$

and we see that

$$\begin{aligned}\text{E(diseased)} &= \frac{\beta\alpha}{\alpha+\mu}\nu^{-1} \\ &= \text{disease incidence} \times \text{mean duration},\end{aligned}$$

yielding (1) in the present interpretation of units of individuals (rather than the often used prevalence proportion in relative units).

Note, furthermore, that what we shall often term prevalence odds satisfies

$$\frac{\text{E(diseased)}}{\text{E(healthy)}} = \frac{\alpha\beta/[(\alpha+\mu)\nu]}{\beta/(\alpha+\mu)} = \frac{\alpha}{\nu},$$

that is,

$$\text{prevalence odds} = \text{incidence} \times \text{mean duration},$$

where incidence is now understood as *intensity of getting diseased for a healthy individual*.

Alho [5] viewed the above relations between prevalence, incidence, and duration in the macrodemographic context of stable population theory.

The above discussion may be generalized to *time-*, *age-* and disease *duration*-dependent intensities $\beta(t), \alpha(t,a), \mu(t,a)$, and $\nu(t,a,d)$, as documented by Keiding [11]. We may then also discuss such concepts as *age-specific prevalence*, expressing the probability of having the disease for a person at age a alive at time t. The general formulas become complicated and are not reproduced here, although some applications will be indicated below.

In the particular case of *time homogeneity*, which, though not very realistic nevertheless underlies most epidemiologic folklore, similar relations between prevalence, incidence, and duration result as above. In particular, the rate of occurrence of new cases in the population becomes

$$\beta\int_0^\infty \exp\{-[\alpha(a)+\mu(a)]\}\alpha(a)\,\mathrm{d}a,$$

and the expected number of diseased at $t=0$ (prevalence on the population scale, "absolute" prevalence) becomes

$$\beta\int_0^\infty\int_0^a \exp\{-[\alpha(y)+\mu(y)]\}\alpha(y) \times \exp[-\nu(a,a-y)]\,\mathrm{d}y\,\mathrm{d}a.$$

In the simple case where the case fatality rate $\nu(a,d)$ depends only on duration d but not age a, a change of order of integration yields

$$\int_0^\infty \beta\exp\{-[\alpha(y)+\mu(y)]\}\alpha(y)\,\mathrm{d}y \times \int_0^\infty \exp[-\nu(v)]\,\mathrm{d}v,$$

where the first factor is incidence as just specified, while since $\exp[-\nu(v)]$ is the survival function of a diseased, the second factor is mean survival. This provides an interpretation of

$$\text{prevalence} = \text{incidence} \times \text{duration}$$

in the age-dependent case, and Keiding [11] specified how to obtain a similar interpretation when $\nu(a,d)$ depends also on a.

The relation *prevalence odds = disease intensity × mean duration* discussed in the time/age/duration homogeneous special case above, also generalizes to the age/duration inhomogeneous case, see again Keiding [11] and O'Neill et al. [23].

Inference on Incidence from Prevalence Data

As has been known in population statistics (**demography**) for hundreds of years, it is true under very restrictive stationarity assumptions (no dependence of birth and death rates on calendar time, no migration) that the age distribution of the living has density proportional to the survival function (= 1 − distribution function) of the mortality. Inference on mortality rates is therefore in principle available from the age distribution of the living.

The simplest generalization of this to morbidity (disease incidence) is analysis of *current status data* where age-specific incidence rates are estimated from the age distributions of diseased and healthy in a cross-sectional sample. Diamond & McDonald [8] gave a survey based on **parametric models** in discrete and continuous time while Keiding [11] and Keiding et al. [14] focused on variants of current **nonparametric survival analysis** techniques. Ades & Nokes [2] gave a useful practical discussion of the range and limitations of these ideas in modeling infectivity rates from seroprevalence studies; and Marschner [20] gave **sample size** calculations.

As emphasized by Preston [24], the crucial **stationarity** assumption may only be verified from at least two successive cross-sectional samples, which however might then be directly used for inference without the stationarity assumption. Recent work in this direction is due particularly to Marschner ([21, 22]) and Ades [1] as well as a series of papers by Brunet & Struchiner (for example [6]), in the pseudo-stochastic mathematical biology tradition.

Inference on Prevalence from Incidence and Mortality Data

It is not uncommon that disease incidence and mortality are more directly estimable (e.g. from a historically prospective incidence study with follow-up) than prevalence. In that case the relations between prevalence, incidence and duration may be used to estimate prevalence, possibly calendar time-and/or age-specifically, see Keiding [11]. Such calculations will often be variations of the nonparametric **Aalen–Johansen estimator** of a transition probability in a nonhomogeneous Markov illness–death process, and this link provides a methodology for derivation of **standard errors**. See [12, 13, 15, 16] for applications to bone marrow transplantation.

Application of such ideas has been primarily in the context of cancer [7, 9, 27], although there are also examples from neuroepidemiology [26, 28], reference [26] containing counterfactual and predictive "what if" calculations under specified past or future structures in incidence and mortality.

Screening

There are strong relations between the above material and the mathematical theory of **screening** for chronic disease [29, 30], in the simplest but also most important case by having the three states *Healthy, Preclinical* (where the patient feels healthy but screening can identify the disease), and *Clinical* (ly manifest) diseased. The same relations are valid, properly interpreted, and Zelen & Feinleib [30] actually also obtained a *prevalence* = *incidence* × *mean duration* result. O'Neill et al. [23] formalized a concept of *initiation*, equivalent to subclinical disease onset. The comprehensive exposition of the theory of screening by Albert et al. [3, 4] and Louis et al. [18] is based on probability densities rather than intensities as in this article and most of the other references.

References

[1] Ades, A.E. (1995). Serial HIV seroprevalence surveys: interpretation, design and role in HIV/AIDS prediction, *Journal of Acquired Immunodeficiency Syndrome* **9**, 490–499.

[2] Ades, A.E. & Nokes, D.J. (1993). Modeling age- and time specific incidence from seroprevalence: toxoplasmosis, *American Journal of Epidemiology* **137**, 1022–1034.

[3] Albert, A., Gertman, P.M. & Louis, T.A. (1978). Screening for the early detection of cancer – the temporal natural history of a progressive disease state, *Mathematical Biosciences* **40**, 1–59.

[4] Albert, A., Gertman, P.M., Louis, T.A. & Liu, S.-I. (1978). Screening for the early detection of cancer – II. The impact of screening on the natural history of the disease, *Mathematical Biosciences* **40**, 61–109.

[5] Alho, J.M. (1992). On prevalence, incidence and duration in general stable populations, *Biometrics* **48**, 587–592.

[6] Brunet, R.C. & Struchiner, C.J. (1996). Rate estimation from prevalence information on a simple epidemiologic model for health interventions, *Theoretical Population Biology* **50**, 209–226.

[7] Capocaccia, R. & de Angelis, R. (1997). Estimating the completeness of prevalence based on cancer registry data, *Statistics in Medicine* **16**, 425–440.

[8] Diamond, I.D. & McDonald, J.W. (1992). Analysis of current status data, in *Demographic Applications of Event History Analysis*, J. Trussel, R. Hankinson & J. Tilton, eds. Clarendon Press, Oxford, pp. 231–252.

[9] Feldman, A.R., Kessler, L., Myers, M.H. & Naughton, M.D. (1986). The prevalence of cancer. Estimates based on the Connecticut Tumor Registry, *New England Journal of Medicine* **315**, 1394–1397.

[10] Hennekens, C.H. & Buring, J.E. (1987). *Epidemiology in Medicine*. Little, Brown & Company, Boston.

[11] Keiding, N. (1991). Age specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[12] Keiding, N. (1999). Event history analysis and inference from observational epidemiology. *Statistics in Medicine* **18**, 2353–2363.

[13] Keiding, N., Klein, J.P. & Horowitz, M.M. (2001). *Multistate models and outcome prediction in bone marrow transplantation.*

[14] Keiding, N., Begtrup, K., Scheike, T.H. & Hasibeder, G. (1996). Estimation from current-status data in continuous time, *Lifetime Data Analysis* **2**, 119–129.

[15] Klein, J.P., Keiding, N. & Copelan, E.A. (1993). Plotting summary predictions in multistate survival models:

Probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine* **12**, 2315–2332.

[16] Klein, J.P., Keiding, N., Shu, Y., Szydlo, R.M. & Goldman, J.M. (2000). Summary curves for patients transplanted for chronic myeloid leukemia salvaged by a donor lymphocyte infusion: The current leukemia free survival curve. *British Journal of Haematology* **109**, 148–152.

[17] Little, J.D.C. (1961). A proof for the queuing formula: $L = \lambda W$, *Operations Research* **9**, 383–387.

[18] Louis, T.A., Albert, A. & Heghinian, S. (1978). Screening for the early detection of cancer. III. Estimation of disease natural history, *Mathematical Biosciences* **40**, 111–144.

[19] MacMahon, B. & Pugh, T.F. (1970). *Epidemiology: Principles and Methods*. Little, Brown & Company, Boston.

[20] Marschner, I.C. (1994). Determining the size of a cross-sectional sample to estimate the age-specific incidence of an irreversible disease, *Statistics in Medicine* **13**, 2369–2381.

[21] Marschner, I.C. (1996). Fitting a multiplicative incidence model to age- and time-specific prevalence data, *Biometrics* **52**, 492–499.

[22] Marschner, I.C. (1997). A method for assessing age-time disease incidence using serial prevalence data, *Biometrics* **53**, 1384–1398.

[23] O'Neill, T.J., Tallis, C.M. & Leppard, P. (1985). The epidemiology of a disease using hazard functions, *Australian Journal of Statistics* **27**, 283–297.

[24] Preston, S.H. (1987). Relations among standard epidemiologic measures in a population, *American Journal of Epidemiology* **126**, 336–345.

[25] Ramalhoto, M.F., Amaral, J.A. & Cochito, M.T. (1983). A survey of J. Little's formula, *International Statistical Review* **51**, 255–278.

[26] Somnier, F.E., Keiding, N. & Paulson, O.B. (1991). Epidemiology of Myasthenia Gravis in Denmark: a longitudinal and comprehensive population survey, *Archives of Neurology* **48**, 733–739.

[27] Verdecchia, A., Capocaccia, R., Egidi, V. & Colini, A. (1989). A method for the estimation of chronic disease morbidity and trends from mortality data, *Statistics in Medicine* **8**, 201–216.

[28] Werdelin, L. & Keiding, N. (1990). Hereditary ataxias and associated disorders. Epidemiological aspects, *Neuroepidemiology* **9**, 321–331.

[29] Zelen, M. (1986). A review of the theory of screening for chronic diseases: single exam and the scheduling of examinations, in *Statistical Design: Theory and Practice*. Cornell University Press, Ithaca, pp. 27–41.

[30] Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic diseases, *Biometrika* **56**, 601–614.

NIELS KEIDING

Incident Case

An incident case is a subject who has just developed the disease or condition of interest for the first time. Incident cases of chronic diseases are particularly valuable for etiologic investigations because disease incidence, unlike disease **prevalence**, is determined by etiologic factors only and not by factors that influence survival following disease onset. To contrast incident with prevalent cases, (*see* **Biased Sampling of Cohorts**; **Case–Control Study, Prevalent**; **Cross-sectional Study**; **Incidence–Prevalence Relationships**; **Prevalent Case**).

MITCHELL H. GAIL

Incomplete Block Designs

Experimental designs with fixed block size k in which the number of treatments (or levels of a single factor) v to be compared exceeds the available block size are called *incomplete block designs*. Such designs first arose in agricultural experiments and were studied by, among others, **Fisher**, **Yates**, and Bose. Incomplete block designs are currently used in a wide variety of subject areas, including agricultural field and animal experiments, food-tasting experiments, industrial processes, toxicology, educational psychology, and, occasionally, in **clinical trials**. For example, in animal experiments it may be desirable to compare the test treatments within a litter, but the number of treatments may exceed the available litter sizes. In food and beverage tasting experiments, the number of items to be tasted is often greater than the number of items a judge can taste within a reasonable time period. In an experiment to compare the tread wear on different kinds of automobile tires, each car can have, at most, four distinct tires and so, if the number of treatments to be tested is greater than four, the blocks (cars) are necessarily incomplete.

Most graduate and advanced undergraduate text books on experimental design contain some material on incomplete block designs, particularly **balanced incomplete block designs** (BIBDs) and **Youden squares**. Two such books, with an applied flavor, are by Lentner & Bishop [9], and John & Quenouille [6]. Das & Giri [3] contains a full account of all the major types of incomplete block designs, their analyses, and some selected construction results. John [5] is a more theoretical discussion of the mathematical structure and analysis of incomplete block designs, and of the construction of such designs by the cyclic development of one or more initial blocks. This article describes briefly the different kinds of incomplete designs, their relationships to each other and to other well-known kinds of complete block designs (*see* **Randomized Complete Block Designs**), and general methods of analysis of such designs.

The incomplete block design problem is one of arranging the test treatments. Two technical concepts needed in a discussion of incomplete block designs are *binary* designs and *connected* designs.

An incomplete block design is said to be *binary* if no test treatment occurs more than once in any block. It can be shown that a design that is not binary may always be improved (in the sense of average or total **variance** of the treatment contrasts) by replacing all duplicates of test treatments in a given block with treatments not already in that block. Hence, we may restrict our attention to binary designs.

A design is said to be *disconnected* if the blocks of the design may be split into two groups in such a way that none of the test treatments that occurs in one group of blocks occurs in the other group of blocks. Treatments that occur in the different groups of blocks may not be compared due to **confounding** with the block effects. As a consequence, we also restrict our attention to designs that are not disconnected; that is, to designs that are *connected*.

There are many different types of incomplete block designs that are both binary and connected. *Balanced incomplete block designs* (BIBDs) have the property that all treatments occur in the same number of blocks, say r, and all pairs of treatments occur in the same number of blocks together, say λ. In BIBDs, all **paired comparisons** of treatments are estimated with equal precision. Kiefer [7, 8] has shown that BIBDs are optimal in the sense of having smallest average or total variance for the paired treatment comparisons.

A standard assumption in the analysis of block designs is that measurements on different experimental units are statistically independent. There are situations in which this is not a reasonable assumption. For example, in agricultural field experiments, fertilizer or irrigation may spill over from an experimental plot to its neighboring experimental plots. A family of experimental designs that is useful in such a situation is the *equineighbored* BIBDs in which each pair of test treatments occurs adjacent to each other in the same number of blocks.

One shortcoming of BIBDs is that for a given number of treatments and a given block size, the number of blocks required to construct a balanced incomplete block design may be too large to be of practical use. Yates [10] addressed this problem with a series of designs that he called **lattice designs**. Lattice designs exist only when $v = s^2$ and $k = s$, for some positive integer s, and are constructed using sets of mutually orthogonal **Latin squares**. Of course, the requirements that $v = s^2$ and $k = s$ are quite restrictive, so the application of lattice designs is somewhat limited.

Bose & Nair [1] discovered a more general alternative to BIBDs. In **partially balanced incomplete**

block designs, all treatments occur in the same number of blocks and pairs of treatments occur together in λ_1 or λ_2 or λ_3 or ... or λ_m blocks together. (Two treatments that occur in the same block λ_l times are said to be in the *lth associate class*). BIBDs are special cases of partially balanced incomplete block designs for which $\lambda_1 = \lambda_2 = \cdots = \lambda_m = \lambda$. Lattice designs are also special kinds of partially balanced incomplete block designs.

Partially balanced incomplete block designs exist for more combinations of parameters than do BIBDs. Das [2] and Giri [4] showed how to construct incomplete block designs for still more combinations of parameters. Their **algorithm** starts with a partially balanced incomplete block design for v treatments in b blocks of size k. Augment each block in this design with α treatments that are not among the original v treatments. The result is a design for $v + \alpha$ treatments in b blocks of size $k + \alpha$. Such designs are called *reinforced designs*.

Youden squares [11] are incomplete block designs in which two sources of variation (**blocking**) may be eliminated. First used by Youden in greenhouse studies, these designs are related to Latin square designs. Indeed, removing any row and any column from a Latin square always yields a Youden square, but Youden squares may also be constructed from certain kinds of BIBDs.

Analysis of Incomplete Block Designs

Youden squares are distinct from the other types of incomplete block designs in that they involve two blocking factors rather than one.

In the analysis of incomplete block designs with a single treatment factor and a single blocking factor, the following linear model (*see* **General Linear Model**) is usually assumed:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, \quad \text{for } i = 1, 2, \ldots v \text{ and } j = 1, 2, \ldots, b,$$

where the τ_is denote the treatment effects and the β_js denote the block effects. No **interaction** among block and treatment effects is assumed. Indeed, because not all treatments occur in every block, not all block-treatment interactions are even estimable. The validity of the assumption of no interactions may be evaluated graphically by plotting the **residuals** from the fitted model against, for example, the predicted values. Degrees of freedom for the different factors are summarized in Table 1.

Table 1 Degrees of freedom for the different factors

Source of variation	Degrees of freedom
Blocks	$b - 1$
Treatments	$v - 1$
Error	$bk - b - v + 1$
Corrected total	$bk - 1$

Computation of appropriate sums of squares is complicated in incomplete block designs due to the lack of **orthogonality** of block and treatment effects. Simple formulae for sums of squares do exist for BIBDs but for other incomplete block designs, even for a partially balanced incomplete block design with just two associate classes, the formulae become extremely complicated. For this reason, it is recommended that analyses be carried out using a computer package. If algebraic formulae are required, the reader is referred to the books by Das & Giri [3] and Lentner & Bishop [9].

The most common analysis of incomplete block designs is often called the *intrablock* analysis, because block differences are eliminated and all treatment **contrasts** may be expressed as differences among observations in the same blocks. This is essentially a **least squares** analysis, assuming that the block effects are fixed. Usually, the block effects are not of intrinsic interest and so the block sum of squares is computed *without* adjusting for treatment effects. Then, the treatment sum of squares is computed after adjusting for block effects. In linear models jargon, this is a *type I* **analysis of variance**, and is standard in most statistical packages, including SAS (*see* **Software, Biostatistical**). If the block effects *are* of intrinsic interest, a block sum of squares adjusted for treatment effects may be computed. This is commonly called a *type II* analysis of variance.

Yates [10] proposed an alternative analysis, which he called the *interblock* analysis, in which additional information about the treatment effects might be obtained by comparing experimental units in different blocks. In modern linear models terminology, the interblock analysis is a mixed effects analysis of variance in which the treatment effects are regarded as **fixed effects**, while the block effects are viewed as independent **random variables** with **mean** zero

and variance σ_β^2. Modeling block effects by random variables is particularly appropriate when the blocks may be viewed as a sample from some population of blocks. For example, in **multicenter clinical trials** the centers at which the study takes place may be viewed as a representative sample of all possible centers (*see* **Random Effects**).

The interblock (mixed effects) analysis may be carried out using, for example, PROC GLM and PROC MIXED in the SAS computer package. Although the intrablock analysis has been the standard analysis for many years, the interblock analysis seems to be gaining in popularity as researchers are more inclined to view their block effects as random quantities.

References

[1] Bose, R.C. & Nair, K.R. (1939). Partially balanced incomplete block designs, *Sankhyā* **4**, 337–372.

[2] Das, M.N. (1958). Reinforced incomplete block designs, *Journal of the Indian Society of Agricultural Statistics* **10**, 73–77.

[3] Das, M.N. & Giri, N.C. (1979). *Design and Analysis of Experiments*. Halstead Press, New York.

[4] Giri, N.C. (1958). On reinforced P.B.I.B. designs, *Journal of the Indian Society of Agricultural Statistics* **12**, 45–56.

[5] John, P.W.M. (1980). *Incomplete Block Designs. Lecture Notes in Statistics*, Vol. 1. Marcel Dekker, New York.

[6] John, J.A. & Quenouille, M.H. (1977). *Experiments: Design & Analysis*, 2nd Ed. Macmillan, London.

[7] Kiefer, J. (1958). On the nonrandomized optimality and randomized non-optimality of symmetrical designs, *Annals of Mathematics and Statistics* **29**, 675–699.

[8] Kiefer, J. (1959). Optimum experimental designs, *Journal of the Royal Statistical Society, Series B* **21**, 272–319.

[9] Lentner, M. & Bishop, T. (1986). *Experimental Design and Analysis*. Valley Book Co., Blacksburg.

[10] Yates, F. (1940). The recovery of interblock information in balanced incomplete block designs, *Annals of Eugenics* **10**, 317–325.

[11] Youden, W.J. (1940). Experimental designs to increase accuracy of greenhouse studies, *Contributions from Boyce Thompson Institute* **11**, 219–228.

D.R. CUTLER

Incomplete Follow-up

Longitudinal (or follow-up) study data analysis is complicated by the diversity of possible outcomes and the different lengths of observation time. Some subjects die or relapse ("failures"), some remain alive or in remission ("survivors"), and some are *lost to follow-up* (e.g. drop out or withdraw from treatment) (Colton [2, pp. 299–302]). Even if one had the time to wait until all subjects met the outcome failure criterion, the problem of accounting for those who were lost to follow-up would remain. Thus, because of their *incompleteness*, longitudinal studies often are subject to selective influences (Hill & Hill [4, p. 27]).

Substantial **bias** in longitudinal studies can result from not considering the duration of the study and from inappropriate handling of incomplete data, even when only a small proportion of the observations is missing (Andersen [1, p. 80], Colton [2, pp. 237–250], Hill & Hill [4, pp. 188–203], Murray & Findlay [6].) Average duration of survival may be a convenient way to summarize "mean observation time" for the "failures" (Colton [2, pp. 299–302]). However, it is a meaningful term *vis-à-vis* mortality or relapse only when all study subjects have had the outcome; it has no meaningful interpretation in terms of survival or prognosis (Colton [2, pp. 237–250, 299–302]). Averaging survival time only among the failures, while ignoring those who have not lived long enough to experience the outcome within the period of observation, selects for early failure and overemphasizes negative outcomes (Colton [2, pp. 299–302]; Hill & Hill [4, pp. 188–203]).

Furthermore, conclusions based solely upon individuals with complete follow-up data presume that results recorded for the failure subgroup would not be affected by including those with incomplete data, i.e. both the survivor and lost-to-follow-up subgroups. In other words, this assumption presumes that the characteristic of "being followed up" does not correlate with the characteristic being measured, for example, survival (Hill & Hill [4, pp. 23–33]). However, the characteristic of being "lost to follow-up" may correlate with being either more or less likely to be alive or dead, so that the ratio of alive/dead may differ in traced versus untraced (i.e. lost) cases (Hill & Hill [4, pp. 188–203]). For example, in a study of treatment for alcoholism, treatment drop-outs may be more likely to relapse to heavy drinking than subjects who remain in treatment.

The magnitude of the incomplete follow-up problem increases as larger numbers of individuals drop out or are withdrawn. Consequently, conclusions drawn from the analysis of outcomes from follow-up data can be considerably biased by incomplete data. Between group differences in mean survival times can be attributed to the incomplete follow-up fallacy.

In all cases, the obvious data management solution is to conduct a comprehensive follow-up or, in the case of **clinical trials**, to emphasize study retention so that concerns regarding **missing data** due to incomplete follow-up do not arise. However, this can be a lengthy, not to mention costly, undertaking. Various statistical techniques have been developed to try to deal with this problem (Gibbons et al. [3], Little & Rubin [5], Rubin [7]).

References

[1] Andersen, B. (1990). Bias I, in *Methodological Errors in Medical Research: An Incomplete Catalogue*. Blackwell Scientific, Oxford, pp. 72–83.

[2] Colton, T. (1974). Fallacies in numerical reasoning, in *Statistics in Medicine*. Little, Brown & Company, Boston, pp. 299–302.

[3] Gibbons, R.D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H.C., Greenhouse, J.B., Shea, T., Imber, S.D., Sotsky, S.M. & Watkins, J.T. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data: application to the NIMH Treatment of Depression Collaborative Research Program dataset, *Archives of General Psychiatry* **50**, 739–750.

[4] Hill, A.B. & Hill, I.D. (1991). Collection of statistics: bias, in *Bradford Hill's Principles of Medical Statistics*, 12th Ed. Edward Arnold, London, pp. 23–33.

[5] Little, R. & Rubin, D. (1987). *Statistical Analysis With Missing Data*. Wiley, New York.

[6] Murray, G.D. & Findlay, J.G. (1988). Correcting for the bias caused by drop-outs in hypertension trials, *Statistics in Medicine* **7**, 941–946.

[7] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

(*See also* **Nonignorable Dropout in Longitudinal Studies**)

HOWARD M. KRAVITZ

Incubation Period of Infectious Diseases

The incubation period is the time interval between exposure to a disease-causing agent and the onset of symptomatic disease. For example, the incubation period of an infectious disease refers to the time interval between infection or exposure to a viral or bacterial agent and the onset of symptomatic (clinical) disease. The incubation period is also called the clinical latency period (*see* **Latent Period**). The focus of this article is on modeling and estimating the incubation period of infectious diseases. However, some of the ideas may also be applicable to the incubation period of noninfectious disease, for example the incubation period of radiation-induced cancer that refers to the time interval from **radiation** exposure to cancer diagnosis.

The length of the incubation period depends on the disease and the infectious agent. It can be very short, perhaps only several days in the case of a streptococcal sore throat, or perhaps several weeks in the case of smallpox, or perhaps a decade in the case of the acquired immune deficiency syndrome **(AIDS)**. After an individual is exposed to an infectious agent, the agent multiplies, and the host defenses are weakened. Eventually, the individual may experience the onset of clinical disease. Individuals may or may not be infectious (that is, capable of transmitting the infection to others) during the incubation period or subsequently.

The incubation period of a disease can be very variable among individuals [2, 21]. A single number, such as the **mean** or **median** incubation period, does not reveal the significant heterogeneity in incubation periods in a population for a given infectious diseases. The incubation period distribution, $F(t)$, is the probability that the incubation period is less than or equal to t time units. The probability density function of incubation periods usually is asymmetric and is **skewed** to the right. Sartwell [23, 24] suggested that the **lognormal distribution** adequately describes the incubation period distribution of a number of diseases. However, other **parametric models for survival** data may also adequately describe incubation period distributions, including the **Weibull**, **gamma**, log-logistic (*see* **Logistic Distribution**), and piecewise **exponential** models [9]. There is no requirement that all infected individuals eventually develop clinical disease. Thus, the distribution function, F, may not be proper. For example, one may postulate that a proportion, p, of infected individuals eventually develop clinical disease with incubation distribution, F_1, and the remaining proportion of infected individuals, $1 - p$, never develop disease; then we have the mixture model $F(t) = pF_1(t)$.

Studies of the incubation period distribution are important for several reasons. First, the incubation period distribution is important for forecasting the course of epidemics, and is used with either transmission models [1] or **back-calculation** approaches [9]. If the incubation period is long, then infected individuals may be silently and unknowingly spreading the infection to others. Secondly, identification of **covariates** or cofactors that may lengthen the incubation period may lead to the development of effective therapeutic interventions. Thirdly, knowledge of the incubation period is useful in counseling infected patients about their prognosis. Finally, the incubation period is a critical parameter in designing **clinical trials** of early interventions and vaccines (*see* **Communicable Diseases**; **Infectious Disease Models**; **Vaccine Studies**).

The ideal study for estimating the incubation period is to monitor a **cohort** of uninfected individuals, determine the dates of infection, and then to follow the infected patients to determine the dates of the onset of clinical disease. The data for estimating the incubation period distribution would consist of the time interval between infection and disease for those patients who became infected. If an infected individual did not develop clinical disease at the time of last follow-up, then the data would be right **censored** at that time. Classical **survival analysis** techniques could be used to estimate the incubation period distribution from right-censored *data* [12]. **Kaplan–Meier** survival curves could be used to estimate $F(t)$ nonparametrically, the cumulative distribution function of incubation periods. Parametric models could also be fit to the right-censored incubation period data (*see* **Parametric Models in Survival Analysis**). A simple example is the case of a single point source epidemic, as might occur with salmonellosis where infection is transmitted from contaminated food or water [21]. In this example a cohort may be defined as all individuals who were exposed (e.g. individuals

who are in a restaurant on the given day that contaminated food was served), in which case the date of exposure is known precisely. Another example of a point source epidemic for a noninfectious disease is the onset of leukemia associated with radiation exposure following the 1945 atomic bomb explosion in Hiroshima [11]. The incidence of leukemia appeared to peak about six years after exposure. Survival analyses could be performed on the time intervals from exposure to clinical disease, and of course some of these intervals may be right censored at the times of last follow-up (*see* **Epidemic Curve**).

Unfortunately, the ideal study of incubation periods can seldom be performed because of a number of important complications. First, it may not be possible to identify a cohort of initially uninfected individuals, and to follow them over time. Instead, we may only have available a sample of cases who already have clinical disease (see the section "Retrospective ascertainment" below). Even if a cohort is assembled and followed over time, it may not be possible to ascertain either the exact dates of infection (exposure) or the onset of clinical disease (see the section "Cohort studies" below). For example, an individual may already be infected at the time of enrollment in a cohort study, but the time that incident infection occurred is unknown. Many of these problems have surfaced in studies of the incubation period of AIDS, and have been the subject of active methodologic research among statisticians in recent years. In the next sections we discuss more fully these complexities and the methodologic approaches to address them. The issues are illustrated with studies of AIDS, although the methods are applicable more generally to other infectious diseases.

Retrospective Ascertainment

The first data on the incubation period of AIDS (time from HIV infection to AIDS diagnosis) were based on transfusion-associated AIDS cases [22]. In that study, AIDS cases were identified who had become infected by receiving a transfusion of infected blood. The date of infection was estimated retrospectively as the date of blood transfusion. There was an important selection criterion to get into the study, namely that subjects had to have AIDS. Early in an epidemic of a new disease the only data about incubation periods that may be gathered rapidly may come from symptomatic cases of disease who have already been identified. These cases of disease are then retrospectively studied to determine dates of exposure to the infectious agent. Such studies have been referred to as having "retrospective ascertainment" because only individuals with symptomatic disease are included and then they are retrospectively studied. A naive analysis of this type of data, which did not account for the selection criteria, could lead to serious underestimation of the incubation period. This is because the data are right **truncated**. Individuals with long incubation periods may not yet have symptomatic disease, and thus could not possibly be included in the data set. To analyze such data properly, the analysis must condition properly on the selection criteria [18, 19].

There are other **biases** with studies based on retrospective ascertainment. For instance in the transfusion example, patients who receive blood transfusion are often elderly and sick with chronic diseases, and thus they may die from other causes of death before developing AIDS. This leads to length-biased sampling: we are more likely to observe patients with shorter incubation periods, because patients with long incubation periods may die first from another disease and thus are never included in the data set. In a series of papers, statisticians have developed methods to correct for these and other biases (see, for example, [18], [19], and [25]). However, none of these methods can correct for the fundamental limitation of this sort of data: they are retrospective and involve only cases of disease and so without strong parametric assumptions they provide essentially no information about the prospective probability of getting a disease once one is infected.

Cohort Studies

A second type of study involves identifying a cohort of uninfected individuals, ascertaining as best one can the subsequent dates of infection, and following the infected individuals to ascertain the date of onset of clinical disease. The first issue concerns the difficulty in identifying the date of infection. The usual method is to test individuals serially with a laboratory assay such as the test for antibodies to the infectious agent. In the case of AIDS, individuals may be serially tested with ELISA or Western Blot assays to identify the dates of seroconversion to HIV antibodies [16].

A complication is that the date of seroconversion does not correspond to the date of infection. Infected individuals will be seronegative for antibodies to the virus until they develop detectable antibodies, usually within several months. Although we define the incubation period as the time from infection to the clinical diagnosis of disease, many studies cannot identify the actual dates of infection but only the time of antibody seroconversion. However, in the case of AIDS, the time from infection to antibody seroconversion is relatively short (approximate median is two months) compared with the much longer period from seroconversion to the onset of disease. Accordingly, many studies define the incubation period to be the time interval from antibody seroconversion (becoming antibody positive) to the onset of clinical disease. Nevertheless, this points out that the results of studies of incubation periods may depend on the choices of the assays that are used to ascertain infection or exposure to the infectious agent. PCR (polymerase chain reaction) testing may identify evidence of infection considerably earlier than antibody testing [17].

If individuals are periodically screened by laboratory tests for evidence of infection, then the date of infection can at best be determined up to an interval (i.e. interval censored). This interval is defined by the time of the latest screening test that was negative for infection, L, and the earliest screening test that was positive for infection, R. The term *doubly censored data* refers to time to event data for which both the time origin and failure time are censored. In cohort studies of the incubation period the data are frequently doubly censored because the date of infection is interval censored and the date of onset of clinical disease is right censored for those individuals who have not developed clinical disease by the time of the last follow-up.

A popular *ad hoc* approach for analyzing doubly censored data on incubation periods is to estimate (impute) the calendar date of infection by the midpoint of the interval. The imputed midpoint calendar date of infection is $S = (L + R)/2$. Then, standard survival analysis techniques for right-censored data are used on the incubation periods with imputed dates of infection. However, such approaches will typically be biased and give incorrect **variance** estimates. The bias of the estimated incubation resulting from midpoint imputation depends critically on the width of the intervals, $R - L$, the incubation distribution, and the density of infection times. For example, in the exponential growth phase of simple epidemics, midpoint imputation will tend to underestimate the time of infection and thus overestimates the incubation period. Law & Brookmeyer [20] studied the impact of midpoint imputation, and concluded that with a median incubation period of 10 years in the case of AIDS, the bias resulting from midpoint imputation associated with intervals even as large as two years is relatively small.

A more formal parametric approach for analyzing the doubly censored data in studies of the incubation period involves specifying parametric models and joint estimation of both the probability densities of infection times and of incubation times. The **likelihood** function is maximized to obtain the **maximum likelihood** estimators. This approach was used by Brookmeyer & Goedert [10] to estimate the incubation period of HIV infection among hemophiliacs. Bacchetti & Jewell [4] used a weakly **semiparametric** approach. A discrete time scale was used with a separate parameter to represent the discrete **hazard** for each month. To avoid irregularities that result from trying to estimate a large number of parameters (e.g. wildly varying hazards from one month to the next with large variances), a **penalized likelihood** function was used that penalized for "roughness" in the estimated hazard function. A completely **nonparametric** approach to the problem has been given by De Gruttola & Lagakos [14]. However, the completely nonparametric estimate of the incubation period distribution, $F(t)$, is often numerically unstable, and it is not defined for all values of t.

Deconvolution Methods

Occasionally, population data may be available both about the incidence of clinical disease and infection rates in the population. The expected **cumulative incidence** of clinical disease up to calendar time t, $D(t)$, is related to infection rates $g(s)$ at calendar time s (numbers of new infections per unit time) and the incubation period distribution, by the convolution equation

$$D(t) = \int_0^t g(s)F(t-s)\,\mathrm{d}s.$$

The basic idea is to use data on $D(t)$ and an estimate of $g(s)$ to glean information about F. This

method was pioneered by Bacchetti & Moss [5] and Bacchetti [3] in connection with estimating the incubation period of HIV infection. The usefulness of the method depends on the availability of accurate information on the infection rates in the population, $g(s)$, and accurate disease **surveillance** data over time. For example, detailed information about historical infection rates was available in San Francisco on the basis of several epidemiologic surveys and cohort studies [5, 26]. The statistical framework is as follows. Let y_j represent the number of cases of disease in calendar interval I_j. Suppose that N, the cumulative number of infections that have occurred, is known. Then the vector of counts of cases of disease, **y**, has a **multinomial distribution** with sample size N and cell probabilities that involve the incubation distribution and the known infection rates. Maximum likelihood estimation methods are used to estimate the parameters of the incubation period distribution. The method is closely related to the back-calculation methodology which uses data on $D(t)$ and an estimate of F to estimate historical infection rates $g(s)$. Back-calculation is a method for estimating past infection rates from disease surveillance data. The method requires reliable counts of numbers of cases of disease diagnosed over time and a reliable estimate of the incubation period distribution. The method has been used to obtain short-term projects of disease incidence and to estimate **prevalence** of infection [6, 9]. Early references on back-calculation are [7] and [8]

Synthesis of Studies of the Incubation Period

The main complications in the analysis and interpretation of studies of the incubation period include uncertainty in the dates of infection and the sampling criteria by which individuals are included in the data set. Accordingly, it is important to synthesize and compare estimates across studies because the estimates may be used on different methodologies with different underlying assumptions.

In the case of AIDS, many different methodologies outlined in this article have been used to study the incubation period distribution. The results from several different methodologies have been compared [15] and a general picture emerges [9]. The probability of developing AIDS within the first two years of HIV antibody seroconversion is very small, less than 0.03. Then the hazard of progression to AIDS begins to rise rapidly so that the cumulative probability of developing AIDS within seven years of seroconversion is approximately 0.25 and the median incubation period is nearly 10 years. When comparing incubation period estimates from different studies, an important consideration is whether treatments were available to delay progression and thus alter the incubation period distribution. Treatments such as AZT became available beginning in 1987 which may lengthen the incubation period. In the case of AIDS, the one covariate that has been shown to influence the length of the incubation period in multiple studies is the age at infection [13].

References

[1] Anderson, R.M. & May, R.M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[2] Armenian, H.K. & Lilienfeld, A.M. (1974). The distribution of incubations periods of neoplastic diseases, *American Journal of Epidemiology* **99**, 92–100.

[3] Bacchetti, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnostic patterns, *Journal of the American Statistical Association* **85**, 1002–1008.

[4] Bacchetti, P. & Jewell, N.P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times, *Biometrics* **47**, 947–960.

[5] Bacchetti, P. & Moss, A.R. (1989). Incubation period of AIDS in San Francisco, *Nature* **338**, 251–253.

[6] Bacchetti, P., Segal, M. & Jewell, N.P. (1993). Back-calculation of HIV infection rates (with discussion), *Statistical Science* **8**, 82–119.

[7] Brookmeyer, R. & Gail, M.H. (1986). Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States, *Lancet* **2**, 1320–1322.

[8] Brookmeyer, R. & Gail, M.H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic, *Journal of the American Statistical Association* **83**, 301–308.

[9] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, New York.

[10] Brookmeyer, R. & Goedert, J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS, *Biometrics* **45**, 325–335.

[11] Cobb, S., Miller, M. & Wald, N. (1959). On the estimation of the incubation period in malignant disease, *Journal of Chronic Disease* **9**, 385–393.

[12] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[13] Darby, S.C., Doll, R. & Thakrar, R., Rizza, C. & Cox, D.R. (1990). Time from infection with HIV to

onset of AIDS in patients with hemophilia in the United Kingdom, *Statistics in Medicine* **9**, 681–689.

[14] De Gruttola, V. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data with applications to AIDS, *Biometrics* **45**, 1–11.

[15] Gail, M.H. & Rosenberg, P.S. (1992). in *AIDS Epidemiology: Methodologic Issues*, N. Jewell, K. Keietz, & V. Farewell, eds. Birkhauser, Boston, pp. 1–38.

[16] Haseltine, W.A. (1989). Silent HIV infections, *New England Journal of Medicine* **320**, 1487–1489.

[17] Horsborgh, C.R., Qu, C.Y., Jason, I.M., Holmberg, S., Longini, I., Schable, C., Mayer, K., Lifson, A., Schochetman, G., Ward, J., Rutherford, G., Evatt, B., Seage, G. & Jaffe, H. (1989). Duration of human immunodeficiency virus infection before detection of antibody, *Lancet* **2**, 637–640.

[18] Kalfbleisch, J.D. & Lawless, J.F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion related AIDS, *Journal of the American Statistical Association* **84**, 360–372.

[19] Lagakos, S., Barraj, L. & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data with application to AIDS, *Biometrika* **75**, 515–523.

[20] Law, C.G. & Brookmeyer, R. (1992). Effects of midpoint imputation on the analysis of doubly censored data, *Statistics in Medicine* **11**, 1569–1578.

[21] Lilienfeld, A.M. & Lilienfeld, D.E. (1980). *Foundations of Epidemiology*, 2nd Ed. Oxford University Press, Oxford.

[22] Lui, K.J., Lawrence, D.N., Morgan, W.M., Peterman, T., Haverkos, H. & Bregman, D. (1986). A model based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome, *Proceedings of the National Academy of Sciences* **83**, 3051–3055.

[23] Sartwell, P.E. (1950). The distribution of incubation periods of infectious disease, *American Journal of Hygiene* **51**, 310–318.

[24] Sartwell, P.E. (1966). The incubation period and the dynamics of infectious of disease, *American Journal of Epidemiology* **83**, 204–216.

[25] Wang, M.-C. (1992). The analysis of retrospectively ascertained data in the presence of reporting delays, *Journal of the American Statistical Association* **87**, 397–406.

[26] Winkelstein, W., Samuel, M., Padian, N.S., Wiley, J., Lang, W., Anderson, R. & Levy, J. (1987). The San Francisco Men's Health Study III. Reduction in human immunodeficiency virus transmission among homosexual/bisexual men, 1982–1986, *American Journal of Public Health* **77**, 685–689.

(*See also* **Screening, Sojourn Time**)

RON BROOKMEYER

Independence of a Set of Variables, Tests of

Suppose that p variables have been measured on each of n sample individuals. Denote the value of the jth variable for the ith individual by x_{ij}, and collect together the p values observed on the ith individual into the vector $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$. Then

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = (\overline{x}_1, \ldots, \overline{x}_p)'$$

is the sample mean vector,

$$\mathbf{A} = (a_{ij}) = \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$$

is the corrected sum of squares and products matrix,

$$\mathbf{S} = (s_{ij}) = \frac{1}{n-1}\mathbf{A}$$

is the sample **covariance matrix**, and

$$\mathbf{R} = \mathbf{DSD}$$

is the sample **correlation** matrix, where

$$\mathbf{D} = \mathrm{diag}(s_{11}^{-1/2}, \ldots, s_{pp}^{-1/2}).$$

The two sample matrices $\mathbf{S}$ and $\mathbf{R}$ can be viewed as estimates of the corresponding population quantities $\boldsymbol{\Sigma}$ and $\boldsymbol{\Upsilon}$, respectively.

Situations often arise in which the p variables can be divided a priori into k distinct sets with p_i variables in the ith set ($i = 1, 2, \ldots, k$). For example, each child in a school class may have to sit an examination that comprises p separate tests, on each of which a mark of between 0 and 100 is awarded. However, these tests may be identifiably of four types: p_1 of them examine verbal ability, p_2 of them examine arithmetic ability, p_3 of them examine general knowledge, and p_4 of them examine logical reasoning. Clearly, *within* each set the individual tests are likely to be (highly) correlated, but a question of interest would then be whether the variables in different sets can be treated as independent. In this section, we consider testing independence of such sets of variables.

Without loss of generality, we can assume the variables to be arranged so that the first p_1 of them fall in the first set, the next p_2 in the second set, and so on. Denote by $\mathbf{S}_{ii}$ the sample covariance matrix of the variables in the ith set and by $\mathbf{S}_{ij}$ the matrix of sample covariances between those pairs of variables in which one variable comes from the ith set and the other from the jth set. Apply the same notation to the matrices $\mathbf{A}, \mathbf{R}, \boldsymbol{\Sigma}$, and $\boldsymbol{\Upsilon}$. Then the overall sample covariance matrix $\mathbf{S}$ can be expressed in partitioned form as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \ldots & \mathbf{S}_{1k} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \ldots & \mathbf{S}_{2k} \\ \vdots & \vdots & & \vdots \\ \mathbf{S}_{k1} & \mathbf{S}_{k2} & \ldots & \mathbf{S}_{kk} \end{pmatrix},$$

and each of the matrices $\mathbf{A}, \mathbf{R}, \boldsymbol{\Sigma}$, and $\boldsymbol{\Upsilon}$ can be partitioned similarly. The null hypothesis that we are concerned with is

$$\mathrm{H}_0\colon\ \boldsymbol{\Sigma}_{ij} = \mathbf{0}$$

for all $i \neq j$, the only requirement on the remaining unspecified matrices $\boldsymbol{\Sigma}_{ii}$ being that they are positive definite for all i. The alternative hypothesis is the general one, i.e. that $\boldsymbol{\Sigma}_{ij} \neq \mathbf{0}$ for at least one $i \neq j$.

Assuming **multivariate normality** of the data, the **likelihood ratio test** for this situation is obtained by maximizing the likelihood of the sample under the null hypothesis and dividing the result by the unconditional maximum of the likelihood. After some algebraic simplification the test statistic can be written

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A}_{11}| \cdots |\mathbf{A}_{kk}|},$$

and elementary properties of determinants establish that equivalent expressions for this statistic are

$$\Lambda = \frac{|\mathbf{S}|}{|\mathbf{S}_{11}| \cdots |\mathbf{S}_{kk}|}$$

or

$$\Lambda = \frac{|\mathbf{R}|}{|\mathbf{R}_{11}| \cdots |\mathbf{R}_{kk}|}.$$

Unfortunately, the exact sampling distribution of Λ is complicated and difficult to handle, so large-sample approximations are generally employed in practice. Standard likelihood-ratio theory provides the basic result that $-n \log \Lambda$ asymptotically follows the **chi-square distribution** with $\nu = \frac{1}{2}(p^2 - \sum_i p_i^2)$ **degrees of freedom** when H_0 is true, so this distribution can be used to find an approximate significance **level** for the test. However, a more accurate large-sample approximation was obtained by Box [2], who showed that when H_0 is true, then

$$\Pr(-a \log \Lambda \le z) = \Pr(\chi_\nu^2 \le z) + ba^{-2}[\Pr(\chi_{\nu+4}^2 \le z) - \Pr(\chi_\nu^2 \le z)] + O(a^{-3}),$$

where

$$a = n - \frac{3}{2} - \frac{1}{3}\left(p^3 - \sum_i p_i^3\right)\left(p^2 - \sum_i p_i^2\right)^{-1}$$

and

$$b = \frac{1}{48}\left(p^4 - \sum_i p_i^4\right) - \frac{5}{96}\left(p^2 - \sum_i p_i^2\right) - \frac{1}{72}\left(p^3 - \sum_i p_i^3\right)^2\left(p^2 - \sum_i p_i^2\right)^{-1}$$

(see also [1, p. 385], [6, p. 534], or [8, p. 90]). Alternatively, Muirhead [6, p. 537] reproduces tables from Davis & Field [3] that contain correction factors to make the percentage points of $-a \log \Lambda$ exactly those of χ_ν^2.

Two special cases of the above test are commonly of interest. The first is when $k = p$, in which case the null hypothesis becomes the hypothesis that all the variables are mutually uncorrelated (independent if normality of data is assumed); in other words, that $\mathbf{\Sigma}$ is a diagonal matrix. In this case the likelihood ratio statistic becomes

$$\Lambda = \frac{|\mathbf{S}|}{s_{11}s_{22}\cdots s_{pp}} = \frac{|\mathbf{A}|}{a_{11}a_{22}\cdots a_{pp}} = |\mathbf{R}|,$$

where $a_{jj} = \sum_{i=1}^n (x_{ij} - \overline{x}_j)^2$ and $s_{jj} = a_{jj}/(n-1)$ for $j = 1, \ldots, p$. Exact percentage points of $-(n - [2p+11]/6)\log \Lambda$ are given by Mathai & Katiyar [5] and reproduced by Seber [8, p. 612]. Alternatively, any of the above approximations can be used with $p_i = 1$ for all i.

The second special case is when there are just two a priori groups of variables; that is, $k = 2$ with p_1 and p_2 variables in the two groups respectively. In this case the sample covariance matrix $\mathbf{S}$ has the partitioning

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix},$$

with a corresponding form for each of the matrices $\mathbf{A}, \mathbf{R}$, $\mathbf{\Sigma}$, and $\mathbf{\Upsilon}$. The null hypothesis is now simply

$$H_0\colon\ \mathbf{\Sigma}_{12} = \mathbf{0},$$

and the likelihood ratio test statistic becomes

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A}_{11}||\mathbf{A}_{22}|} = \frac{|\mathbf{S}|}{|\mathbf{S}_{11}||\mathbf{S}_{22}|} = \frac{|\mathbf{R}|}{|\mathbf{R}_{11}||\mathbf{R}_{22}|}.$$

When the null hypothesis is true this statistic has Wilks's lambda distribution $\Lambda_{p_2, p_1, n-p_1-1}$, which has been tabulated extensively (see, for example, Seber [8, p. 565]), so that exact significance levels are easily found.

Note also that in this case we have $|\mathbf{S}| = |\mathbf{S}_{11}||\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}|$ (using results for patterned matrices given, for example, by Seber [8, p. 519]), and equivalent expressions exist for $\mathbf{A}$ and $\mathbf{R}$. The likelihood ratio statistic can thus be reduced to the form

$$\Lambda = \prod_{i=1}^{q}(1 - r_i^2),$$

where $q = \min(p_1, p_2)$ and the r_i^2s are the nonzero **eigenvalues** of $\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ (or, equivalently, of $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$, or of either of these expressions with $\mathbf{A}_{ij}$ or $\mathbf{R}_{ij}$ replacing $\mathbf{S}_{ij}$ for $i, j = 1, 2$). These are the squared **canonical correlations** between the two sets of variables, which are important multivariate descriptors of the inter-set associations.

In this particular special case, it is possible also to derive a **union–intersection test** of the null hypothesis (which has not been found possible to date for the general case of k sets). To derive this test, we consider the univariate hypothesis

$$\rho_{ab}^2 = \frac{(\mathbf{a}'\mathbf{\Sigma}_{12}\mathbf{b})^2}{[(\mathbf{a}'\mathbf{\Sigma}_{11}\mathbf{a})(\mathbf{b}'\mathbf{\Sigma}_{22}\mathbf{b})]} = 0,$$

where ρ_{ab} is the correlation between two arbitrary linear combinations, one from each of the two sets of variables. A suitable test statistic for this univariate hypothesis is $(\mathbf{a}'\mathbf{S}_{12}\mathbf{b})/[(\mathbf{a}'\mathbf{S}_{11}\mathbf{a})(\mathbf{b}'\mathbf{S}_{22}\mathbf{b})]^{1/2}$

and, on maximization with respect to both **a** and **b**, the union–intersection test statistic is found to be $\max_i r_i^2$. Critical values of this statistic have also been tabulated extensively; see, for example Pearson & Hartley [7, Tables 48 and 49] or Seber [8, p. 593].

Various other test statistics have been proposed for this last situation. Invariance arguments lead to statistics which are functions of the **eigenvalues** r_i^2, and the most popular variants are $\sum_{i=1}^{s} r_i^2$ or $\sum_{i=1}^{s} [r_i^2/(1 - r_i^2)]$. Muirhead [6, p. 548] discusses some power comparisons among the various statistics.

A final point concerns the behavior of all these test statistics when the data are not normal. Relatively few systematic studies have been conducted, although both Muirhead [6, p. 546] and Fang & Zhang [4, p. 170] give some results relevant to samples from elliptic distributions. Fang & Zhang derive forms of the likelihood ratio statistic appropriate for such samples, while Muirhead considers asymptotic null distributions of normal-based likelihood ratio test statistics when the data are actually from elliptic distributions. He quotes some **Monte Carlo** studies which indicate that the normal likelihood ratio test statistics should only be used with care if the data come from elliptic distributions.

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika* **36**, 317–346.

[3] Davis, A.W. & Field, J.B.F. (1971). Tables of some multivariate test criteria, *Division of Mathematical Statistics Technical Paper No. 32*. CSIRO, Melbourne, Australia.

[4] Fang, K.-T. & Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*. Science Press, Beijing/Springer-Verlag, Berlin.

[5] Mathai, A.M. & Katiyar, R.S. (1979). Exact percentage points for testing independence, *Biometrika* **66**, 353–356.

[6] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

[7] Pearson, E.S. & Hartley, H.O. (1972). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge University Press, Cambridge.

[8] Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

(*See also* **Multivariate Analysis, Overview**; **Multivariate Bartlett Test**; **Sphericity Test**)

W.J. KRZANOWSKI

Indian Statistical Institute

Research in the theory and applications of statistics as a new scientific discipline began in India in the early 1920s through the pioneering initiative and efforts of **Prasanta Chandra Mahalanobis**. Soon after his return from England, Mahalanobis began to carry out statistical studies with the help of some part-time assistants. A chance meeting with Nelson Annandale (the then Director of the Zoological and Anthropological Survey of India) and subsequent interactions with him led to the first scientific paper by Mahalanobis on the statistical analysis of stature of Anglo-Indian males of Calcutta. This was followed by further research in **anthropometry**, in meteorology and in problems of flood control in North Bengal and Orissa. Gradually, a small group of young scientists was picked up by him in the Department of Physics, Presidency College, Calcutta, where he was a professor. This group formed the nucleus of a laboratory which later came to be known as the Statistical Laboratory.

In the early 1930s, realizing the necessity for a concerted effort for the advancement of theoretical and applied statistics in India, Mahalanobis, together with P.N. Banerjee and N.R. Sen, both professors of Calcutta University, convened a meeting on December 17, 1931, to consider various steps to be undertaken for the establishment of an association for the advancement of statistics in the country. As a result of this meeting, the Indian Statistical Institute (ISI) was registered as a non-Government and nonprofit-distributing learned society on April 28, 1932, with Sir R.N. Mookerjee as President and Professor P.C. Mahalanobis as (Honorary) Secretary. The total staff strength then was only two or three. From such a modest beginning, the Institute grew, under the remarkable leadership of Mahalanobis into an all-India organization which now has a staff strength of about 1600, including about 500 scientific staff. The Institute has its headquarters in Calcutta and centers at Bangalore and Delhi and a branch at Giridih. In addition, it has a network of service units of the Statistical Quality Control and Operations Research Division at Bangalore, Baroda, Calcutta, Chennai (formerly Madras), Coimbatore, Delhi, Hyderabad, Mumbai (formerly Bombay), Pune, and Tiruvananthapuram.

From the very beginning, Mahalanobis and his associates, who included S.S. Bose, R.C. Bose, S.N. Roy, K.R. Nair, K. Kishen, and H.C. Sinha, worked with zeal and enthusiasm for the development of statistical theory and methods, and in promoting research and practical applications in different areas of the natural and social sciences. *Sankhyā*, the Indian Journal of Statistics, was started in 1933 with Mahalanobis as its Editor, and received instant international recognition, which continues till today. Pioneering research activities were carried out in many areas of statistical theory, especially in the core areas of **multivariate analysis**, sample surveys and **experimental design**. Such activities were strengthened and new directions were opened up by Professor C.R. Rao and many others who joined the Institute in the 1940s and the tradition continues. The Institute pioneered the development of statistical methods in agricultural research and in the conduct of large-scale agricultural enquiries. This led to a large number of research publications and to the introduction of training activities offering short-term courses in statistics for officers in government departments and scientific institutions. The scientists of ISI, led by Mahalanobis, helped in introducing the first post-graduate degree course in Statistics in India at the Calcutta University in 1941, and in securing a separate section for Statistics in the Indian Science Congress.

Activities of the Institute gained further momentum from 1938. Mahalanobis started sample surveys to estimate the area under the jute crop in Bengal in 1937 as an exploratory project, which later grew to a full-scale survey of the entire province in 1941. Gradually, sample surveys of agricultural crops, and other socioeconomic surveys, became some of the most important activities of the Institute, and earned the Institute and Mahalanobis international reputation. Mahalanobis was appointed Honorary Statistical Advisor to the Cabinet, Government of India, and in 1950, through his initiative, the National Sample Survey (NSS) was started for conducting socioeconomic surveys of all-India coverage on a continuing basis. This was the first ever attempt in India to have a database for various developmental programs and the five year plans.

The ISI played a pioneering role in starting the Statistical Quality Control (SQC) movement in India by organizing a visit of W.A. Shewhart, the father of SQC, to India in 1948 and later by inviting

other experts like W.E. Deming. SQC promotional work was gradually spread all over the industrial centers in India under a comprehensive program covering education and training, applied research, and consultancy services.

Research in economics was greatly stimulated when in 1954 Prime Minister Jawaharlal Nehru entrusted the preparation of the draft Second Five-Year Plan of the country to Mahalanobis and the Institute. The "draft" submitted by Mahalanobis and the plan models formulated by him in that connection have since been regarded as major contributions to economic planning in India. Since then many economists of the Institute have continued to work on various aspects of national planning and, until 1970, were directly helping the Planning Commission of the Government of India in the preparation of the long-term prospective plans for the country. Research in other disciplines of social sciences was also started in the Institute in the late 1950s. Mahalanobis's participation in 1946 in the annual scientific conferences of the Milbank Foundation led to the initiation of systematic studies in India on population growth. Earlier, the well-known Y-sample estimates for the 1941 census population were also derived by the ISI. Theoretical and empirical research in sociology using statistical techniques was started in the Institute for the first time in South-East Asia. Similarly, the development and introduction of psychometric tests for selection processes in different organizations was first made by the ISI in India besides carrying out basic research in psychometry (*see* **Psychometrics, Overview**). Studies of the phonetic structure of some major Indian languages have been made on a continuing basis in the Institute under the guidance and collaboration of the famous linguist Djordje Kostic.

The Institute, since its inception, recognized the need for development and use of accurate and fast computing equipment for the processing and analysis of data. Mahalanobis strongly believed that to be a good theoretical statistician one must also compute and must therefore have the best computing aids. The Institute has lived up to this tradition from the very beginning. In 1953 a small analog computer was designed and built in the Institute. In 1956 the Institute acquired a HEC-2M machine from the UK which was the first digital computer in India. In 1958 a digital computer URAL was received as a gift from the then USSR. From 1956 to the mid-1960s the Institute had been serving as a *de facto* national computer center for the country. In the early 1960s the Institute, in collaboration with Jadavpur University, undertook the design, development, and fabrication of a fully transistorized digital computer, called ISIJU-1, which was commissioned in 1966 by Mr M.C. Chagla, the then Minister of Education, Government of India.

Quantitative analysis in the physical and earth sciences was one of the novel ideas that Mahalanobis pursued in the true spirit of the Institute. In addition to evolving some interesting techniques and obtaining some very interesting results from the analysis of directional geological data, the Institute also made a significant contribution by discovering the bones of a 16 m (+) long sauropod dinosaur named *Barapasaurus tagoreii*, from the lower Jurassic Kota rocks near Sironcha, Gadchiroli district, Maharashtra, in the 1960s. The discovery has helped in understanding the interesting problem about the origin and evolution of sauropod dinosaurs. It represents the only intermediate form between the prosauropods and the sauropods, and is called a "missing link" in the evolution of the sauropod dinosaur.

The Institute expanded its research, teaching, training, and project activities and earned national and international recognition over time. The substantial contributions of the Institute to theoretical and applied statistical work have culminated in the recognition of the Institute by the Government of India enacting *The Indian Statistical Institute Act, 1959 (No. 57)* which declared the Institute as an "Institution of National Importance" and empowered it to award degrees and diplomas. None other than Pandit Jawaharlal Nehru, the then Prime Minister of India, piloted the bill in Parliament. With this recognition, the already existing teaching and training programs were consolidated and expanded and courses for the degrees of Bachelor of Statistics (B.Stat. (Honors)) and Master of Statistics (M.Stat.) were started in June 1960. The Institute was also empowered to award Ph.D./D.Sc. degrees from the same time. Later on, courses leading to Master of Technology degrees were started in Computer Science and in Quality, Reliability and Operations Research. Recently, the Institute has also been empowered to grant degrees and diplomas in mathematics, quantitative economics, computer science and subjects related to statistics as well as statistics itself. A master's

degree programme in quantitative economics has just been initiated.

The role and importance of ISI in conducting and promoting teaching of statistics has been appreciated by international bodies as well. In 1950 the **International Statistical Institute** initiated the International Statistical Education Centre (ISEC), Calcutta, jointly with ISI, to impart training in theoretical and applied statistics to participants selected from developing countries. The center is run by ISI under the auspices of UNESCO, the International Statistical Institute and the Government of India.

Recognition of the Institute by the Act of Parliament provided greater encouragement to research activities not only in statistics and mathematics but also in various branches of the natural and **social sciences**, without whose live contact, it was believed, the methodology of statistics could not grow. It is also due to this fact that "Unity in Diversity" has been adopted as the motto of the Institute.

The objectives of the Institute are:

1. to promote the study and dissemination of knowledge of statistics, to develop statistical theory and methods, and their use in research and practical applications generally, with special reference to problems of planning for national development and social welfare;
2. to undertake research in various fields of natural and social sciences with a view to the mutual development of statistics and these sciences; and
3. to provide for, and undertake, the collection of information, investigations, projects, and operational research for purposes of planning and the improvement of efficiency of management and production.

From the early days, the Institute has been in touch with many internationally famous scientists in different disciplines from all over the world. Some of these scientists have worked in the Institute for several months or even longer. **R.A. Fisher**, a pioneer of modern statistics, was a regular visitor to the Institute and lent it considerable support. J.B.S. Haldane, a geneticist of international repute, was a member of the faculty for several years beginning 1957. At the inspiration of these stalwarts and other renowned scientists, the Institute began to expand and/or undertake research activities in several areas of the natural and social sciences with the hope that collaboration under the same roof would foster the mutual development of statistics and other disciplines. In fact, the Institute stood up to R.A. Fisher who called statistics a "key technology" of the century, in view of its intimate relevance to all scientific endeavors which involve experimentation, measurement and **inference** from sample to aggregate.

Coming to more recent times, the Institute has continued to pursue its goal of attainment of excellence in various fields of science. Fundamental research in statistics with its roots in applications has been the bottom line ever since the inception of the Institute. The contributions from the Institute in multivariate analysis, design and analysis of experiments, sample surveys, statistical methods of data analysis and statistical inference have found their places in textbooks and monographs, and the tradition continues. In addition, **probability theory** and **stochastic processes** have also been major areas of research in the Institute. The mathematicians of the Institute, in addition to collaborating with the statisticians, are also making fundamental contributions in several fields – topology, functional analysis, harmonic analysis, algebra, combinatorics, quantum mechanics, game theory, to name a few. The current trend of research in statistics not only carries forward the traditions set up in the Institute, but is also setting new directions, both in theory and applications, in different disciplines.

The Institute has been maintaining its tradition of high-quality research and development in the field of computer science. In 1979, a microprogrammed signal processing system using the **Fast Fourier Transform (FFT)** was designed and developed. Keeping pace with the global advances in computer technology, the activities of the Institute in the field of computer science gathered a tremendous momentum in the late 1970s, resulting in diversification of research in different areas including **algorithms** and complexity, parallel and distributed processing, fault-tolerant computing, VLSI, computational geometry, fuzzy sets and systems, wave propagation, atmospheric remote sensing, speech signal processing, cybernetics, **pattern recognition**, **neural networks**, **artificial intelligence**, image processing (*see* **Image Analysis and Tomography**), computer vision, document analysis, natural language processing, particle physics, fluid dynamics, plasma physics, etc. In recognition of its contributions in the field of computer science,

the Government of India established, in collaboration with the United Nations Development Program (UNDP), one of the five national nodal centers for knowledge-based computing systems (NCKBCS) in ISI in the year 1988.

The different disciplines under the social sciences also continued to develop and flourish over time by carrying out basic research as well as inter- and multidisciplinary programs. In economics, the Institute has come to be known as a specialized center for its significant contributions in different branches of theory and also for studies on such areas as demand analysis, poverty and levels of living, measurement of inequalities, production and prices, national income and allied topics, development and planning, etc. In **demography**, sociology, psychometry and linguistics also, the Institute maintained its distinctive feature for the focus and emphasis on quantitative aspects. Mention may be made, in this context, of the pioneering theory for teaching and training for hearing-impaired children, developed by D. Kostic. Based on this theory the Electronics Unit of the Institute, in collaboration with the Linguistic Research Unit and the Government of Tripura, designed, developed and fabricated a set of instruments for hard-of-hearing children of the Institute of Speech Rehabilitation, Government of Tripura, Agartala. This has come to be regarded as having significant impact on social welfare. Recently, the Institute has established a Policy Planning Research Unit at its Delhi Center and a Survey Research and Data Analysis Center in Calcutta.

Plant and human biology have been major areas of research in biological sciences in the Institute. Both basic and applied research are conducted, with emphasis on quantification, statistical design and analysis, and modeling. In the area of plant biology, research has included quantification of natural variability and modeling animal behavior, effect of interaction of rice varieties on yield, use of protein extracted from leaves to supplement human food, mathematical modeling of ecological and embryological phenomena, etc. In the area of human biology, researches have included anthropometric, genetic and biochemical studies on population affinities, microevolution, studies on utilizing data on anthropometric variability in designing car seats, human adaptation to differing environments, human ecology and growth, (*see* **Growth and Development**), and **genetic epidemiology**.

Over the years, the SQC & OR Division has grown to the size of having ten operating units all over the country and has uniquely served for promotion, education and training and technical guidance in total quality management methodology and quality assurance systems for the benefit of the manufacturing and service industry. It has thus, as was intended, played a leading role in the dissemination of new concepts, methods and techniques in the areas of quality and productivity.

The central library of the Institute is located at Calcutta with a network extending to other locations of the Institute. Over the years, the library of the Institute has attained the distinction of being one of the richest libraries in the country, particularly in the fields of statistics and related disciplines. The library has developed a well-equipped reprography and photography unit. The library's gift collections include the personal libraries of Mahalanobis and Shewhart. The library has been recognized as the depository library for World Bank Publications. A separate collection of books and journals in mathematics, statistics, etc. known as the Eastern Regional Center of the National Board of Higher Mathematics (NBHM), has been developed out of the grants from the NBHM.

The Documentation Research and Training Centre (DRTC) established at Bangalore in 1962 by the late S.R. Ranganathan, a doyen in the field of library and information science, is engaged in research, teaching and training in documentation and information science. The Institute awards post-graduate diplomas in documentation sciences.

The continual publication of many books and monographs and a large number of scientific papers in national and international journals by the scientific staff of the Institute give a good idea of the nature and extent of the contributions of the Institute to statistics and related fields. Scientists of the Institute have also received recognition from many national and international organizations by way of awards, titles, and fellowships. With a dynamic group pursuing and guiding research work in some of the most modern topics and frontier areas of statistics, mathematics, and in various fields of the natural and social sciences, there is close interaction with scientists from all over the world.

S.B. RAO

Infant and Perinatal Mortality

An infant death is defined as the death of a live-born baby before a completed year after birth [24]. The concept of infant mortality did not emerge until the latter half of the nineteenth century, although data for much earlier periods have subsequently been used to construct infant mortality rates [3, 11]. Similarly, the idea that stillbirths and deaths in the first week of life could be grouped together and described as perinatal deaths was not put forward until 1948 [16], but perinatal mortality rates have been constructed retrospectively for earlier years.

The Emergence of the Concept of "Infantile Mortality"

In 1858, Sir John Simon, Medical Officer to the General Board of Health used the term "infantine death rate" for mortality among children under the age of five. In his introduction to *Papers Relating to the Sanitary State of the People of England* [18], he expressed the view that this rate was a proxy measure of the health of the population. Drawing attention to the wide differences between districts, he commented that these infantine death rates

> ... furnish a very sensitive test of sanitary circumstances; so that differences of infantine death-rate are, under certain circumstances, the best proof of differences of household condition in any number of compared districts. And, secondly, those places where infants are most apt to die are necessarily the places where survivors are most apt to be sickly ... [18].

He went on to suggest that, "Deaths which occur in excess within five years of birth are mainly due to two sets of causes; first to the common infectious diseases of childhood prevailing with unusual fatality; and secondly to the endemic prevalence of convulsive disorders, diarrhea and pulmonary inflammation". A factor that he did not mention was differences in the completeness of registration of births. It was likely that some babies who died shortly after birth were not registered; in particular, babies born outside marriage in big cities.

William Farr first used the current definition of infant mortality indirectly when reporting deaths in 1875, although he did not explicitly use the term "infant", nor the word "infantile", which was more commonly used in the succeeding decades. He wrote, "I show that in 1000 infants born in 1875 no less than 158 died in the first year of life ..." [5].

Infantile Mortality and Stillbirth Registration

In the same report, William Farr commented on the implications of changes in the law that had made the registration of live births compulsory in 1875. He pointed out that, "In the case of children born alive – or who breathe – both the birth and death are registered, but still-born children are not registered in England" but "Under the provisions of the new Registration Act, no still-born children, however, should be buried without a certificate stating that they were still-born" [5]. There is good evidence that these certificates were also used to bury victims of infanticide [11].

An international survey undertaken for the Select Committee on Stillbirth Registration and published in 1893 showed that Britain and Ireland lagged behind many other countries in not having stillbirth registration [10]. Nearly 20 years later, a second and fuller survey was done by the "Special Committee on Infantile Mortality" set up by the **Royal Statistical Society** [17].

These surveys covered European countries, New Zealand, states of Australia and the US and provinces of Canada. The Royal Statistical Society's survey also covered other British colonies and some Latin American countries. It found that stillbirth registration was compulsory in most countries, but that, "The large majority of the countries where registration is not required are under the British Crown, and it may be concluded that the Registration Laws in force in such countries have been based on the English model." In contrast, Sweden had introduced registration of both live and still births and deaths as early as 1749, followed by Denmark and Norway in 1801.

The surveys found wide differences between the countries in their criteria for birth registration and for distinguishing between infant deaths and stillbirths. As William Farr had already pointed out, "In France, under the provisions of the Code Napoleon, children who die (either before or after birth) before registration, are recorded as still-born. Dr **Bertillon** estimates

that twenty-two in 100 of the children registered in France as still-born breathed, and such children in England would be registered among the births and deaths" [5].

It was this problem that prompted the Royal Statistical Society's enquiry. When presenting the Committee's report to the Society, Reginald Dudfield focused his attention on the need for a definition of stillbirth, as none of the countries with stillbirth registration appeared to have one in their legislation [4]. He considered two sets of issues. The first was the question of "viability". This was linked to the **gestational age** after which the fetus should be considered a child capable of independent life. The second was how to establish whether the fetus or child was, or had been, alive at birth.

After asking the Obstetrical Section of the Royal Society of Medicine for a definition of stillbirth, he recommended the following slightly amended version:

> A "still-born child" means a child whose body at birth measures not less than 13 inches (32 centimeters) in length from the crown of the head to the sole of the heel and who, when completely born (the head, body and limbs of the child, but not necessarily the afterbirth being extruded from the body of the mother), exhibits no sign of life – that is to say whose heart has ceased to function, as demonstrated by the absence of pulsation in the cord at its attachment to the body of the child and the absence of any heart-sounds or impulses.
>
> NOTE: Crying and/or breathing – being secondary signs of life, manifested only when the heart is acting – can be relied upon as signs of life, but in the absence of either or both is not to be held to be proof of absence of life in the child [4].

When stillbirth registration was eventually introduced in England and Wales in 1927, a shorter definition based on gestational age was used:

> "Stillborn" and "stillbirth" shall apply to any child who has issued forth from its mother after the twenty-eighth week of pregnancy and which did not at any time after being completely expelled from its mother breathe or show any other signs of life [6].

Public Concern About Infantile Mortality and Developments in Analysis

The Royal Statistical Society's enquiry came at a time when there had been a growing concern about infant mortality in a number of countries. In Britain, this had been prompted by the discovery that many potential recruits for the Boer War were unfit and by the campaign by the Women's Co-operative Guild for maternity services.

The Royal Statistical Society Committee also discussed the way in which the infantile mortality rate was calculated. It had defined this as the ratio of the deaths during the first year of life to births. Its enquiries had revealed, however, that some countries had used the estimated numbers alive under the age of one year instead. Given the relative inaccuracy of population estimates, the Committee recommended using births instead.

Having pointed out that some countries compiled their birth statistics by year of registration and others by year of occurrence, it recommended using occurrences. It also recommended that stillbirths should be tabulated separately and that in countries where live-born babies who died before registration were registered as stillbirths, they should actually be counted as infant deaths [17].

As a result of public concern about infant mortality, analyses of infant mortality by age at death in the Annual Reports of the Registrar General from 1904 were more detailed than in earlier years. In addition, a series of four reports on infant mortality was published by the Local Government Board, the government department responsible for public health. In the first of these, the Board's Chief Medical Officer, Arthur Newsholme reiterated John Simon's view in stating that, "Infant mortality is the most sensitive index we possess of social welfare and of sanitary administration, especially under urban conditions" [14]. These reports compared the infant mortality rates for different parts of England and Wales and discussed the comparisons and local data in relation to factors such as sex, legitimacy, family size, the quality of help available in childbirth, the ages of mothers, poverty, overcrowding and defective sanitation.

A similar concern about infant mortality in the US at the same period has been attributed to its emergence as a world power.

> The problem of infant mortality is one of the great social and economic problems of our day ... A nation may waste its forests, its water power, its mines, and to some degree, even its lands, but if it is to hold its own in the struggle for supremacy, its children must be conserved at any cost. On

> the physical, intellectual and moral strength of the children of today the future depends [9].

One response to this was the setting up of the Children's Bureau and its enquiry in 1913 into infant mortality in eight cities. This enquiry took a cohort approach, following up children born in a given year, and was analyzed by a statistician, Robert Morse Woodbury (*see* **Birth Cohort Studies**). Having considered the same broad range of factors as Arthur Newsholme, he concluded that the level of the father's earnings was the strongest "causal" factor associated with infant mortality [21].

These conclusions underpinned calls for political action to improve the conditions for young children and their parents, but these were not the only views held at the time. Followers of the **eugenics** movement took the view that heredity was the prime factor in infant mortality and that attempts to reduce it hindered natural selection by delaying or preventing the death of children who would survive as "weaklings" [15].

The introduction of new technology in the form of punched card equipment increased the extent to which infant mortality could be analyzed by cause, age at death, and other factors [3]. Peter McKinlay's analysis of the decline in infant mortality in England and Wales in the first quarter of the twentieth century showed that, "... all ages have not shared in this amelioration to the same extent ... as a general rule, the nearer to birth the less has the mortality been affected" [12].

In his analysis, he subdivided infant mortality into two categories, "(a) the death rate from 'congenital debility, malformation and premature birth' (number 28 of causes of death given for each separate district in the Annual Reports of the Registrar General), and (b) the remainder of the infant deaths under one year". He labeled these as "neo-natal" and "post-natal", respectively, and called stillbirths "ante-natal" deaths.

He concluded from his analysis of differences between areas of England and Wales that

> only the provision of skilled assistance to mothers in childbed is of importance in connection with ante-natal mortality. ... The neo-natal death rate is related both to variations in external environment and in the obstetrical assistance available to mothers in childbed. ... The postnatal death rate seems to offer the greatest scope for administrative measures. In this case the health of the mother would appear to come first in order of appearance, environment also is of some importance, whereas the effects of variations in obstetrical services have now ceased to be reflected on the mortality of infancy [12].

The term "neonatal" was also used a few years later in an international analysis for the League of Nations [19]. This had a demographic focus and started by looking at trends in countries' infant mortality in relation to their birth rates, population changes and overall death rates. It brought together the two streams of opinion on infant mortality in stating that, "It is evident that the causes of infant mortality may be divided into two distinct categories: (a) those depending on the fitness of the infant to live at all, and (b) those arising from the unfitness of the surroundings to support infant life" [19].

In comparing the death rates for different countries, the author grouped together deaths of live-born babies under the age of one month with stillbirths, partly to get over the differences in stillbirth registration referred to earlier. The term "birth mortality" was suggested for this combined rate. This rate varied far less between countries than that for older babies. The author commented that, "Infant mortality has repeatedly been stated to be the best measure of the sanitary state of a country ... if the infant mortality rate is employed for this purpose, it should clearly be only the part relating to infants over 1 month" [19].

The Establishment of Current Definitions

In the latter half of the twentieth century, the current definitions of fetal death, stillbirth, and the components of the infant mortality rate have become established. They are shown in Figure 1. Introducing these definitions, the Registrar General's Statistical Review for England and Wales for 1951 commented that the use of the term "neonatal period" was "now traditional among obstetricians and compilers of vital statistics" and its first use by writers of Annual Reviews had been in 1936 [7]. It also pointed out the term "perinatal mortality" had first been used in 1950. The term had been coined by a demographer Sigismund Peller, who took the view that time trends in early neonatal deaths had more in common with those in stillbirths than with those in the rest of the first year of life. [16]

In most developed countries, infant mortality rates have fallen persistently and dramatically in the latter

$$\text{Stillbirth rate} = \frac{\text{still births} \times 1000}{\text{live births} + \text{stillbirths}}$$

$$\text{Perinatal mortality rate} = \frac{(\text{stillbirths} + \text{deaths at 0–6 days after live birth}) \times 1000}{\text{live births} + \text{stillbirths}}$$

$$\text{Early neonatal mortality rate} = \frac{\text{deaths at 0–6 days after live birth} \times 1000}{\text{live births}}$$

$$\text{Late neonatal mortality rate} = \frac{\text{deaths at 7–27 days after live birth} \times 1000}{\text{live births}}$$

$$\text{Neonatal mortality rate} = \frac{\text{deaths at 0–27 days after live birth} \times 1000}{\text{live births}}$$

$$\text{Postneonatal mortality rate} = \frac{\text{deaths at 1–11 months after live birth} \times 1000}{\text{live births}}$$

$$\text{Infant mortality rate} = \frac{\text{deaths under the age of 1 year after live birth} \times 1000}{\text{live births}}$$

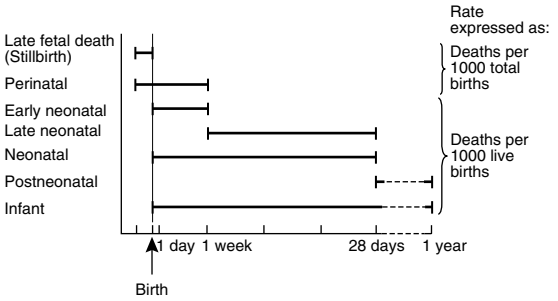

Figure 1 Definitions of stillbirth and infant morality rates. Reproduced from Macfarlane & Mugford [11] by permission of the office for National Statistics. © Crown copyright 1984

half of the twentieth century to well below 10 infant deaths per 1000 live births. As the survival rates of preterm and immature babies have risen, the definitions used have been extended to include ever smaller babies and fetuses and countries still differ considerably in their criteria for registering live and still births. [8, 13].

The **World Health Organization**'s Expert Committee on Vital Statistics recommended in 1950 that, as a minimum, all countries register and tabulate all fetal deaths after the 28th completed week of gestation [22]. This was endorsed in the seventh revision of the **International Classification of Diseases (ICD)**. This was the first to incorporate a definition of stillbirth that separates the definition of a dead-born fetus from the criteria for registration.

A quarter of a century later, a different approach was used in the ninth revision of the ICD. This recommended that *national* perinatal statistics should include all fetuses and babies delivered "weighing at least 500 g or, where birthweight is unavailable, the corresponding gestational age (22 weeks) or body length (25 cm crown–heel), whether alive or dead" [23]. It went on to acknowledge that countries' legal requirements might have different criteria for registration purposes and that international comparisons should be restricted to fetuses and babies "weighing 1000 g or more (or, where birthweight is unavailable, the corresponding gestational age (28 weeks) or body length (35 cm crown–heel)" [23].

The tenth revision of the ICD took yet another approach and defined the perinatal period "which commences at 22 completed weeks (154 days) of gestation (the time when birthweight is normally 500 g) and ends seven completed days after live birth" [24]. Although the ICD no longer uses the term stillbirth, the term still appears in the legislation of individual countries, such as the countries of the UK.

The relevance of the upper cutoff point for the perinatal period has often been questioned in recent years. Increasingly, the use of intensive care is enabling very immature babies to survive, but there is also a tendency for those that die to do so later after birth. One response to this is to redefine perinatal deaths as the sum of all stillbirths and neonatal deaths, as is done in Australia. Another, which takes into account the view that there are increasing

differences between stillbirths and neonatal deaths, is to tabulate stillbirths, neonatal and postneonatal deaths separately and drop the use of the perinatal mortality rate.

The ninth revision of the ICD recommended using a special form of certificate for perinatal deaths, with the cause section subdivided into "main and other diseases or conditions in the fetus or infant," "main and other maternal conditions affecting the fetus or infant" and "other relevant circumstances" [23]. It did not indicate how these data should be analyzed. In response to this problem, the Office of Population Censuses and Surveys, now known as the **Office for National Statistics**, has devised a hierarchical classification to group causes of stillbirth and neonatal death from the forms of certificate it introduced in 1986 [1, 2]. This classification uses categories first proposed by Jonathan Wigglesworth for use with information derived from case notes [20] and also builds on the extensive research done over many years in Aberdeen, Scotland.

References

[1] Alberman, A., Botting, B., Blatchley, N. & Twidell, A. (1994). A new hierarchical classification of causes of infant deaths in England and Wales, *Archives of Disease in Childhood* **70**, 403–409.

[2] Alberman, A., Blatchley, N., Botting, B., Schuman, J. & Dunn A. (1997). Medical causes on stillbirth certificates in England and Wales; distribution and results of hierarchical classifications tested for the Office for National Statistics, *British Journal of Obstetrics and Gynaecology* **104**, 1043–1049.

[3] Armstrong, D. (1986). The invention of infant mortality, *Sociology of Health and Illness* **8**, 211–232.

[4] Dudfield, R. (1912). Still-births in relation to infant mortality, *Journal of the Royal Statistical Society* **76**, 1–26.

[5] Farr, W. (1877). Letter to the Registrar General, in *Thirty-eighth Annual Report of the Registrar General of Births, Deaths and Marriages in England*. Abstracts of 1875. Cd 1786. HMSO, London.

[6] General Register Office (1929). *The Registrar General's Statistical Review of England and Wales for the Year 1927*. HMSO, London.

[7] General Register Office (1954). *The Registrar General's Statistical Review of England and Wales for the Year 1951*. HMSO, London.

[8] Gourbin, C. & Masuy-Stroobant, G. (1995). Registration of vital data: are live and stillbirths comparable all over Europe?, *Bulletin of the World Health Organization* **73**, 449–460.

[9] Holt, L.E. (1913). Infant mortality, ancient and modern. An historical sketch, *Archives of Pediatrics* **30**, 885–915.

[10] House of Commons (1893). *Still-births in England and Other Countries*. Return to House of Commons. No 279. HMSO, London.

[11] Macfarlane, A.J. & Mugford, M. (1984). *Birth Counts: Statistics of Pregnancy and Childbirth*. HMSO, London.

[12] McKinlay, P.L. (1929). Some statistical aspects of infant mortality, *Journal of Hygiene* **28**, 394–417.

[13] Mugford, M. (1983). A comparison of reported differences of vital events and statistics, *WHO Statistics Quarterly* **26**, 201–212.

[14] Newsholme, A. (1910). *Report by the Medical Officer on Infant and Child Mortality*, Supplement to the Thirty-Ninth Annual Report of the Local Government Board for 1909–10. Cd 5263. HMSO, London.

[15] Pearson, K. (1912). The intensity of natural selection in man, *Proceedings of the Royal Society of London, Series B* **85**, 469–476.

[16] Peller, S. (1948). Mortality past and future, *Population Studies* **1**, 405–456.

[17] Royal Statistical Society (1912). Report of Special Committee on Infantile Mortality, *Journal of the Royal Statistical Society* **76**, 27–87.

[18] Simon, J. (1858). Introductory report, in *Papers Relating to the Sanitary State of the People of England*. HMSO, London.

[19] Stouman, K. (1934). The perilous threshold of life. League of Nations, *Quarterly Bulletin of the Health Organisation* **3**, 531–612.

[20] Wigglesworth, J.S. (1980). Monitoring perinatal mortality – a patho-physiological approach, *Lancet* **ii**, 684–686.

[21] Woodbury, R.M. (1925). *Causal Factors in Infant Mortality. A Statistical Study Based on Investigation in Eight Cities*, Children's Bureau Publication No 25. Government Printing Office, Washington.

[22] World Health Organization (1957). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*, 7th Rev., Vol. 1. WHO, Geneva.

[23] World Health Organization (1977). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*, 9th Rev., Vol. 1. WHO, Geneva.

[24] World Health Organization (1992). *International Classification of Diseases and Related Health Problems*, 10th Rev., Vol. 1. WHO, Geneva.

(*See also* **Birthweight**; **Cause of Death, Underlying and Multiple**; **Death Certification**; **Midwifery, Obstetrics, and Neonatology**; **Vital Statistics, Overview**)

ALISON MACFARLANE

Infectious Disease Models

There are two major roles for stochastic infectious disease models. Their study provides insights into the spread of disease in a community, and they are an essential component in the analysis of data from empirical studies of infectious disease (*see* **Epidemic Models, Stochastic**).

The Epidemic Threshold Theorem

A major insight provided by epidemic models is that major epidemics can be prevented in a large community by immunizing only a fraction of the individuals. This property is sometimes referred to as herd immunity, and is quantified by the **epidemic threshold** theorem. Deterministic models for infectious diseases (*see* **Epidemic Models, Deterministic**) indicate this result, but these models assume that both the group of susceptible individuals and the group of infective individuals are large throughout the epidemic. The stochastic version of the threshold theorem also requires a large susceptible group, but the infection process may start with only one infective individual. The stochastic threshold theorem is also richer in that it quantifies the probability of a major epidemic when a small number of infective individuals enter a large community that is currently free from the disease.

In the overly simple setting of a large community of homogeneous individuals, who mix uniformly (*see* **Random Mixing**), the threshold theorem indicates that the probability of a major epidemic is zero when the proportion of individuals who are susceptible to infection is less than $1/\theta$. The parameter θ, known as the basic **reproduction number**, is the mean number of individuals infected by the direct contacts of an infective entering the community when all other individuals are susceptible.

The epidemic threshold theorem holds under quite general conditions, but the bound $1/\theta$ then depends on the community structure and the heterogeneity among individuals (see [7] and [8]).

Data on Outbreaks in Households

Infectious disease data have three features that distinguish them from other data. There is usually some knowledge about the mechanism that generates the data, the data are dependent, and the infection process is only partially observable. A consequence of these features is that the analysis of data is usually most effective when it is based on a model that describes aspects of the infection process. The level of detail that should be incorporated into the model depends on the objective of the study.

Disease transmission and the natural history of diseases evolve in continuous time, but discrete time models are often appropriate for data analysis. It may be that events are only recorded to the nearest day, say, or only the eventual outcomes of outbreaks are observed. Data on the eventual number of cases in households are often collected, because households are a manageable unit size and data on eventual infection can be verified by laboratory tests, which makes them relatively reliable.

Chain Binomial Models

In a household having initially s susceptible individuals, there will be $1, 2, \ldots,$ or s eventual cases. The probability of a specified number of eventual cases in an infected household is computed in terms of disease transmission probabilities by considering the likelihood of the various chains of infection. To illustrate, suppose that one of a total of five susceptible individuals of a household is infected and starts an outbreak in the household. Assume that the outbreak evolves without further infection from outside. Four eventual cases in the household could result via a number of different chains of infection. One such chain is $1 \to 2 \to 1 \to 0$, which means that the single initial infective infected exactly two household members, who in turn infected exactly one member, and the last remaining susceptible member escaped infection throughout.

A simple **chain binomial model** would compute the probability for this chain, given one introductory case, as

$$\binom{4}{2} p_1^2 q_1^2 \binom{2}{1} p_2 q_2 \binom{1}{0} p_1^0 q_1 = 12 p_1^2 q_1^3 p_2 q_2,$$

where q_i is the probability that a susceptible escapes infection when exposed to i infectives for the duration of their infectious periods and $p_i = 1 - q_i$. The probability that the number of eventual cases in a household is x is the sum of chain probabilities over all chains with x eventual cases.

The **EM algorithm** is a convenient tool for finding maximum likelihood estimates when fitting chain binomial models to size of household outbreak data. This is pointed out with reference to **partner studies** for HIV infection in [11] and is discussed more fully in the review paper [6] (*see* **AIDS and HIV**).

Models that capture the infection mechanism of the data generally contain parameters with clear interpretations and are well suited for testing epidemiologically important hypotheses. For example, with a chain binomial model for the size of household outbreaks, we can test the Reed–Frost hypothesis $q_2 = q_1^2$, or the Greenwood hypothesis $q_2 = q_1$. The Reed–Frost assumption is appropriate for diseases that spread primarily by direct person-to-person contact.

Many methods of analysis of household data assume that each household outbreak evolves essentially independently after the initial infection of the household. This assumption is often of concern. Longini & Koopman [10] propose an analysis based on a pragmatic chain binomial model that also allows infection from outside the household.

Epidemic Chain Models with Random Effects

It is instructive to think about disease transmission in terms of a continuous infectivity function $\lambda(t)$ that indicates how infectious an infective is t time units after being infected. The infectivity function reflects both the level of infectious agent emitted by the infective and his or her rate of making contacts with others. Often, the infectivity function is zero for a period immediately after infection, because the infectious organism is developing within the body and no infectious agent is emitted. When disease transmission is person-to-person, the probability that a given susceptible individual escapes infection when exposed to a given infective is $q_1 = \exp[-\int_0^T \lambda(t)\,\mathrm{d}t]$, where T is the duration of time from infection until the end of the infectious period.

Epidemic chain binomial models assume that infectives are homogeneous, in the sense that they all have the same infectivity function. When infectives have different infectivity functions, we still use chain binomial models if the infectives can be partitioned into homogeneous groups. Otherwise, we proceed by considering the q_1 for each infective to be a realization from a probability distribution. In these **random effects** models, see [4, Chapter 3], the probabilities of the epidemic chains are expressed in terms of the **moments** of q_1. This allows for heterogeneity in the infectivity of infected individuals. Heterogeneity in susceptibility or among households can be allowed for in a similar way. An application of random effects models to data on *Shigella sonnei* in households is given by Baker & Stevens [3].

A comprehensive analysis of infectious disease data on household outbreaks, allowing infection from outside the household, variation in the duration of the infectious period, and **covariates**, is described by Addy et al. [1].

Continuous Time Data for Households

Sometimes, when daily data are available on symptoms shown by infected individuals, the analysis is based on a continuous time model. The standard model used is a compartmental model for the irreversible compartments Susceptible $\rightarrow$ Exposed $\rightarrow$ Infective $\rightarrow$ Removed, referred to as the SEIR model. An individual in the exposed category is infected, but not yet infectious, and said to be in the **latent period**. The final category is called removed, because these individuals play no further part in the infection process. These individuals may simply have recovered and have acquired immunity from further infection for the duration of the epidemic. It is of interest to estimate characteristics, such as the mean and variance, of the latent and infectious periods. This can be done by assuming a parametric model for the distribution of the latent and infectious periods, as described in [2, Chapter 15] and [4, Chapter 4]. It is also of interest to make inferences about the functional form of the infectivity function, which is considered in the context of transmission of the human immunodeficiency virus (HIV) by Shiboski & Jewell [12] on the basis of data on partners of individuals infected with HIV.

Data on an Epidemic in a Community

Regression Analysis

When data are available on the days on which individuals show symptoms of disease, and these can be used to deduce the date of infection, with reasonable accuracy, then a comprehensive regression analysis is possible. The response variable is the

indicator of infection for each susceptible individual on each day. The **Mantel–Haenszel** test statistic has been suggested as a way of reducing the number of covariates, see [4, Chapter 5]; however, a **logistic regression** model is also convenient for determining which covariates are needed in the model. When a final set of covariates is arrived at it is useful to fit a **loglinear** regression model in these covariates to the binary data. The preference for the loglinear model stems from the more direct epidemiologic interpretation of its parameters in the infectious disease context. More specifically, if Y is the indicator of escaping infection for a given susceptible on a given day, then fitting the model $Y \simeq \text{binomial}[1, \exp(-\boldsymbol{\beta}'\mathbf{x})]$ is useful, because with this model $\boldsymbol{\beta}'\mathbf{x}$ can be interpreted as the force of infection acting on the susceptible on that day. The covariate $\mathbf{x}$ might include the number of infectives in the community and the number of infectives in the susceptible's household, for example. An illustration of such a regression analysis is given in [4, Chapter 6].

Martingale Methods

The fact that the infection process is observed only partially causes the likelihood function based on continuous time data to be very complicated. This has encouraged the development of pragmatic methods based on simplifying assumptions and approximations. In contrast, methods of analysis derived from martingales for **counting processes** have proved successful for developing simple methods of statistical inference for some crucial parameters, such as the basic reproduction number, for quite general models. Tutorial accounts of these methods are given in [5] and [4, Chapter 7].

Vaccine Efficacy

A major motivation for the study of infectious diseases is to gain insight into ways in which they can be controlled and to determine requirements for their control. The most successful method of intervention continues to be vaccination (*see* **Vaccine Studies**). The epidemic threshold theorem plays a key role here, but it can only be applied if parameter estimates are available. A crucial parameter is the vaccine efficacy. Traditionally, vaccine efficacy has been estimated by $1 - (AR_V/AR_U)$, where AR_V is the attack rate among vaccinated individuals and AR_U is the attack rate among unvaccinated individuals. The attack rate is the proportion of individuals infected in the specified risk group over a nominated period of time. As a measure of the protective effect that the vaccine provides, this concept of vaccine efficacy suffers from depending on both the community from which the data come and on the time period over which the data are collected. Recently, there has been a more careful study of the interpretation and estimation of vaccine efficacy, see [9]. Typically, as a concept of protection against infection, vaccine efficacy might be interpreted as α, where the force of infection acting on vaccinated individuals is $\alpha g(t)$ at chronological time t when the force of infection exerted on an unvaccinated susceptible is $g(t)$. Depending on the vaccine, α may be a constant in [0, 1] or a separate realization on a random variable for each vaccinated individual.

The HIV/AIDS Epidemic

The appearance of AIDS stimulated new interest in the problems of modeling and data analysis for infectious disease studies. A distinguishing feature of infection with HIV is the very long time between infection and diagnosis with AIDS. This has made it feasible, and of interest, to assess the size of the epidemic, forecast its progress, and study characteristics of disease progression during the course of the epidemic (*see* **AIDS and HIV**).

References

[1] Addy, C.L., Longini, I.M. & Haber, M. (1991). A generalized stochastic model for the analysis of infectious disease final size data, *Biometrics* **47**, 961–974.

[2] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

[3] Baker, R.D. & Stevens, R.H. (1995). A random effects model for analysis of infectious disease final-state data, *Biometrics* **51**, 956–968.

[4] Becker, N.G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

[5] Becker, N.G. (1993). Martingale methods for the analysis of epidemic data, *Statistical Methods in Medical Research* **2**, 93–112.

[6] Becker, N.G. (1997). Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases, *Statistical Methods in Medical Research* **6**, to appear.

[7] Becker, N.G. & Dietz, K. (1995). The effect of household distribution on transmission and control of

highly infectious diseases, *Mathematical Biosciences* **127**, 207–219.

[8] Becker, N.G. & Hall, R. (1996). Immunization levels for preventing epidemics in a community of households made up of individuals of different types, *Mathematical Biosciences* **132**, 205–216.

[9] Halloran, M.E., Haber, M. & Longini, I.M. (1992). Interpretation and estimation of vaccine efficacy under heterogeneity, *American Journal of Epidemiology* **136**, 328–343.

[10] Longini, I.M. & Koopman, J.S. (1982). Household and community transmission parameters from final distributions of infections in households, *Biometrics* **38**, 115–126.

[11] Madger, L. & Brookmeyer, R. (1993). Analysis of infectious disease data from partner studies with unknown source of infection, *Biometrics* **49**, 1110–1116.

[12] Shiboski, S.C. & Jewell, N.P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data, *Journal of the American Statistical Association* **87**, 360–372.

(*See also* **Communicable Diseases**; **Incubation Period of Infectious Diseases**)

NIELS G. BECKER

Infectivity Titration

In an experiment to assay the virulence of a suspension of living, self-reproducing organisms (which we refer to here as "particles"), doses derived by successive dilution of the original suspension are administered to groups of host organisms, and the proportion of hosts infected at each dilution is recorded. The "independent action" or "one-hit" theory [13–15] assumes that infection can be initiated by one particle, which, for some reason or other, is "effective". Particles act independently, any one particle having a probability, p, of being effective on a particular occasion. The biological interpretation of p depends on the host–pathogen system. It may be the probability of a particle being retained in the host, or reaching a totally susceptible site; or, on a stochastic model, it may depend on the relative rates at which particles divide and die within a host [4] (*see* **Stochastic Processes**).

Situations for which this model has been proposed include the infection of plants by viruses [6], the titration of viruses in egg membranes [16] or portions of membrane [10], the infection of animals by bacteria [20, 21], and the initiation of tumors in animals by viruses [7].

Suppose that at the ith dilution, the mean number of particles per inoculum is λ_i, and that n_i hosts are inoculated, of which r_i are infected. If the probability of infection, p, is the same for all hosts, the probability that a host receiving this inoculum will not be infected is the first term in the **Poisson distribution**,

$$P_i = \exp(-\lambda_i p). \tag{1}$$

If this dose has a concentration equal to a fraction x_i of the original preparation, we can define $\gamma_i = \lambda_i p = \gamma x_i$, say, where γ is the mean number of effective particles per inoculum in the undiluted preparation. From a set of results at a series of different dilutions one could estimate γ as in the **dilution method for bacterial density estimation**, or the "most probable number" method (*see* **Serial Dilution Assay**).

Note, first, that the parameter to be estimated here depends both on the density of the particles in the original preparation and on the probability of infection, p. In the dilution method for counting viable bacteria it is assumed that a particle present in the inoculum will be detected without fail, so that $p = 1$ for all hosts. The absolute density of particles in a preparation can then be estimated. In the more general situation considered here, the absolute density of particles cannot be estimated without some further assumption about the probability of infection. Nevertheless, an infectivity titration can be used to compare two or more microbial populations, inoculated into randomly assigned hosts (*see* **Biological Assay, Overview**).

Note, secondly, that if p is not equal to unity universally, it may vary between hosts, and this feature leads to a number of important modifications of the model described above. However, variation in infectivity between individual particles does not invalidate the simple model, provided the hosts are identical in their susceptibility.

Host Variability

Suppose that p varies from host to host with a distribution function $F(p)$. (We avoid the use of a capital letter for the **random variable** p, as P is customarily used in the different sense of (1) above.) Then, the probability of noninfection is

$$P_i = \int_0^1 \exp(-\lambda_i p)\, \mathrm{d}F(p). \tag{2}$$

Expression (2) is a **moment generating function** for the distribution $F(p)$, and may be expanded in terms of the moments. Since p is restricted to the range (0,1), the terms involving the higher moments are, in practice, negligible for all except very large values of λ_i, and a good approximation is

$$P_i \cong \exp(-\lambda_i \mu)\frac{1 + (\lambda_i^2 \mu_2)}{2}, \tag{3}$$

where μ and μ_2 are, respectively, the mean and variance of p. Expression (3) shows that the effect of host variability is to flatten the dose–response curve relating P_i to λ_i or to the known concentration factor x_i, the proportionate effect being greater at the higher concentrations.

The general effect of host variability was noted in [21]. The precise effect has been studied for a number of specific distributional forms, including the **gamma distribution** [1, 16], truncated **exponential** [1, 5], **beta** [5] and two-point [2, 7] distributions.

For the gamma (or type III) distribution with density function

$$f(p) = \exp \frac{(-p/\mu k)(p/\mu k)^{(1/k)-1}}{(\mu k)\Gamma(1/k)}, \qquad (4)$$

with mean μ and variance $\mu_2 = k\mu^2$, (2) gives

$$P_i = (1 + \lambda_i \mu k)^{-(1/k)}, \qquad (5)$$

which is equivalent to (3) to $O(k)$ and tends to (1) as $k \to 0$. The gamma distribution, having infinite range, is strictly inappropriate as a distribution for p, but may be regarded as an adequate model for small values of μ, when truncation at $p = 1$ would have little effect.

The two-point distribution places probability masses α_i at values $p = \pi_i, i = 1, 2$, with $\pi_1 < \pi_2$. If $\pi_2 \gg \pi_1$, the effect on the dose–response curve relating the probability of infection, $Q_i = 1 - P_i$, to x_i or $\log x_i$, is to suggest a "shelf" at approximately $Q = \alpha_2$, since only the hosts with the higher level of susceptibility will be infected at the lower concentrations.

Further insight into the flattening effect [5] follows by regarding the response curve as the distribution function of a tolerance distribution (*see* **Quantal Response Models**). Let V denote the variance of the tolerance distribution of $\log x$ from (2), V_0 the corresponding value from (1) for homogeneous hosts, and $V_{\log p}$ the variance of $\log p$. Then, as noted in [5],

$$V = V_0 + V_{\log p}. \qquad (6)$$

This increase in variance caused by host variability corresponds to the flattening of the dose–response curve. Many authors [8, 10, 20] have used probit analysis to analyze dose–response curves in infectivity titrations (*see* **Quantal Response Models**). It is known [11] that the exponential model (1) leads to a tolerance distribution for the log dose closely similar to a normal distribution, and that, with logs taken to base 10, the expected slope of a probit line is about 1.8–2.0. The expression (6) shows that, in general, probit slopes against log dose for infectivity experiments will usually be less than 1.8. Systematic values below 1.8 would indicate host variability, whilst values above 2.0 would suggest departure from independent action.

Detection and Estimation of Host Variability

Probit analysis provides a rough way of checking the evidence for host variability and of estimating its magnitude through the parameter $V_{\log p}$. A better approach is based on the more correct models (1) and (2).

For a test of the null hypothesis of zero variability in a titration experiment with n hosts at each dilution, Moran [16, 17] proposed the statistic

$$T = \sum r_i(n - r_i), \qquad (7)$$

and evaluated its null distribution for series with various dilution ratios. Moran's statistic is symmetric as between infected and noninfected hosts, and thus does not use the fact, shown by (3), that departures from the null model (1) will tend to be associated with excessive numbers of noninfections at high doses.

Armitage [1] proposes a score test (*see* **Least Squares**), based on the likelihood function derived from (3). The test statistic is

$$\phi = \sum \left[\frac{n_i \hat{Q}_i - r_i}{\hat{Q}_i} \right] \frac{(\hat{\gamma}_0 x_i)^2}{2}. \qquad (8)$$

Here, $\hat{\gamma}_0$ is the maximum likelihood estimate of γ under the null model (1), and $\hat{Q}_i$ is the corresponding estimate of Q_i. Note that in (8) the discrepancies between observed and expected frequencies are more heavily weighted at the higher doses. A modified statistic ϕ' is available when γ is estimated by a consistent but inefficient estimator, such as that suggested by Fisher [9] based on $\sum r_i$.

Moran's T is identically zero for series with $n = 1$, and is therefore inappropriate in that situation. Experiments are likely to require larger values of n, but may sometimes be designed in blocks, each with $n = 1$.

Stevens [22] proposes the statistic R, equivalent to Moran's [18] $D + 1$, defined as the number of dilutions between (and including) the first at which not all hosts are infected, and the last at which at least one is infected. As an example, in the following series of increasing dilutions with single observations (+ representing infection),

$$\cdots + + 0 + 0 + + 0\,0 \ldots,$$

the value of R is 5. As with T, there is no differential weighting of the two extremes of the response curve. An alternative simple statistic [3] is J, defined as the number of infected hosts at dilutions beyond that at which the first noninfection occurs. In the example, $J = 3$. The statistic J has been shown [3] to be highly correlated with the efficient statistic ϕ and to be more powerful than R in the detection of small departures from the null model.

All tests of host variability must be one-sided, since the null hypothesis lies at one end of the parameter range. It might be argued that such tests are pointless, since some degree of host variability must exist except in the extreme case where $p = 1$ for all hosts. However, one situation leading to a mean μ less than unity might arise if a proportion, μ, of hosts were invariably susceptible, the remaining proportion, $1 - \mu$, being totally resistant. In that case, there would be no variability. Contradiction of the null hypothesis in a test for variability, then, at least rules out that possible scenario. In general, though, it will be more useful to estimate the degree of variability than to test for evidence of its existence.

Maximum likelihood estimation of k follows from the likelihood equations based on (3) [1], and less efficient estimates may be based on the simple test statistics such as T.

Dependent Action Models

Most log dose–response curves encountered in infectivity experiments are sufficiently flat to support the independent action or "one-hit" theory, for which other evidence exists [13, 15]. If infection depended on cooperation between more than one particle, the resulting "multi-hit" log dose–response curve would be steeper than that given by (1). Iwaszkiewicz & Neyman [12] discuss the estimation of the critical number of effective particles required for infection, and introduce the concept of host variability by allowing the critical number to vary between hosts, the effect again being to flatten the log dose–response curve. Independent action could conceivably also give rise to a steeper curve. In the production of tumors by viruses, for example, tumors might be too faint to be detected unless several occurred close together [19]. In the absence of host variability, the response curve would then be steeper than the exponential form (1).

References

[1] Armitage, P. (1959). Host variability in dilution experiments, *Biometrics* **15**, 1–9.

[2] Armitage, P. (1959). An examination of some experimental cancer data in the light of the one-hit theory of infectivity titrations, *Journal of the National Cancer Institute* **23**, 1313–1330.

[3] Armitage, P. & Bartsch, G.E. (1960). The detection of host variability in a dilution series with single observations, *Biometrics* **16**, 582–592.

[4] Armitage, P., Meynell G.G. & Williams, T. (1965). Birth–death and other models for microbial infection, *Nature* **207**, 570–572.

[5] Armitage, P. & Spicer, C.C. (1956). The detection of variation in host susceptibility in dilution counting experiments, *Journal of Hygiene* **54**, 401–414.

[6] Bald, J.G. (1937). The use of numbers of infections for comparing the concentration of plant virus suspensions. I. Dilution experiments with purified suspensions, *Annals of Applied Biology* **24**, 33–35.

[7] Bryan, W.R. (1956). Biological studies on the Rous sarcoma virus. IV. Interpretation of tumor–response data involving one inoculation site per chicken, *Journal of the National Cancer Institute* **16**, 843–863.

[8] Finter, N.B. & Armitage, P. (1957). The membrane piece technique for *in vitro* infectivity titrations of influenza virus, *Journal of Hygiene* **55**, 434–456.

[9] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

[10] Fulton, F. & Armitage, P. (1951). Surviving tissue suspensions for influenza virus titration, *Journal of Hygiene* **49**, 247–262.

[11] Irwin, J.O. (1942). The distribution of the logarithm of survival times when the true law is exponential, *Journal of Hygiene* **42**, 328–333.

[12] Iwaszkiewicz, K. & Neyman, J. (1931). Counting virulent bacteria and particles of virus, *Acta Biologiae Experimentalis, Varsovie* **6**, 101–142; Reprinted in Neyman, J. (1967). *A Selection of Early Statistical Papers of J. Neyman*. Cambridge University Press, Cambridge.

[13] Meynell, G.G. (1957). The applicability of the hypothesis of independent action to fatal infections in mice given *Salmonella typhimurium* by mouth, *Journal of General Microbiology* **16**, 396–404.

[14] Meynell, G.G. (1957). Inherently low precision of infectivity titrations using a quantal response, *Biometrics* **13**, 149–163.

[15] Meynell, G.G. & Meynell, E.W. (1958). The growth of micro-organisms *invivo* with particular reference to the relation between dose and latent period, *Journal of Hygiene* **56**, 323–346.

[16] Moran, P.A.P. (1954). The dilution assay of viruses. I, *Journal of Hygiene* **52**, 189–193.

[17] Moran, P.A.P. (1954). The dilution assay of viruses. II, *Journal of Hygiene* **52**, 444–446.

[18] Moran, P.A.P. (1958). Another test for heterogeneity of host resistance in dilution assays, *Journal of Hygiene* **56**, 319–322.

[19] Parker, R.F., Bronson, L.H. & Green, R.H. (1941). Further studies of the infectious unit of vaccinia, *Journal of Experimental Medicine* **74**, 263–281.

[20] Peto, S. (1953). A dose–response equation for the invasion of micro-organisms, *Biometrics* **9**, 320–335.

[21] Reid, D.B.W. & MacLeod, D.R.E. (1954). The relation between dose and mortality for *Salmonella dublin*, *Journal of Hygiene* **52**, 18–23.

[22] Stevens, W.L. (1958). Dilution series: a statistical test of technique, *Journal of the Royal Statistical Society, Series B* **20**, 205–214.

PETER ARMITAGE

Inference, Foundations of

Humans learn considerable information (and misinformation) from birth and throughout their lives. This ability arises through some combination of genetic influences, direct experience in the surrounding environment, and input from others, both through interpersonal interaction (i.e. parents, friends, teachers, etc.) and stored and supplied societal information (i.e. media, books, paintings, digital information, analog information, etc.). Some learning is relatively direct and easy to assimilate (e.g. falling off a bicycle can be painful and lead to injury), but much other learning is more difficult, because of both variability in observed relationships and the complex and sophisticated underlying models and concepts that "explain" observations (e.g. the currently accepted models of particle physics or molecular biology). Multiple distinct intellectual pathways have contributed to the current methods of biostatistical inference. These pathways include randomness, **probability**, regular variability, statistical modeling, and observational vs. experimental data collection.

The ideas of randomness go back at least to biblical times where the casting of lots was used. However, the modern ideas of probability were initially developed with respect to games of chance [8]. **Jacob Bernoulli's** proof of the strong **law of large numbers** [12] set the scene for interpreting probabilities as the limit of the proportion of times that an event would occur in a long sequence of repeated identical trials. In this context it was natural that probability was thought of in a frequentist sense: "fair" games could be repeated with equal probabilities of differing outcomes in cards or dice.

A second intellectual thread was the observation that in repeatable situations with variable outcomes there was a regularity or **pattern** to the variability observed. Repeated measurements taken in the (presumed) same or similar situation clustered around some (presumed) true value. It was natural to consider ways of dealing with this variation in outcome, and natural that at some point the mathematical theory of probability theory would be used as one possible way of assessing the variability.

Probability as a concept becomes more difficult as one thinks about it. Einstein, for example, did not believe that probability was an inherent property of the universe: his view is often quoted as "God does not play dice with the universe" [3]. However, most physicists believe that probability is basic to the quantum mechanical structure of the universe. Regular variability that could "mimic" probability theory could also result from the mathematics of **chaos theory** that make it clear that very small changes in initial conditions, for even relatively simple nonlinear systems, can lead to dramatic changes in outcome over time. If for no other reason than an inability to delineate precisely the initial conditions, variability in biological systems can be expected to be the commonly observed situation. As an example of possible chaos theory, multiple card shuffling could be expected (with a number of shuffles) to approximate the usual mathematical model that all permutations are equally likely; yet there are skilled individuals who can shuffle cards with perfect knowledge of how the cards will interleave.

At the same time there are (seemingly) unique situations where only one event will be observed. For example, one presidential election, one football game, or the treatment of an individual patient may be at issue. Yet individuals evaluate such situations and at least implicitly attach **odds** or probabilities to these situations; some consistently do an excellent job, while others do not do so well. This suggests that the evaluation of probabilities or the **likelihood** of outcomes also can be related to the personal or individual beliefs of humans. This **subjective** version of probability, or **Bayesian** probability, has become another contributor to current thinking about statistical inference. Of course, thinking Bayesians also believe in an external frequentist probability (otherwise one could not talk about the data swamping the prior probability, or a state of nature). Savage [11] argues that unless one follows a Bayesian behavior system one will be in a position to lose, *no matter what the true state of nature*, if forced to bet. Note, however, that in this formulation the state of nature has an external (frequentist?) probability associated with it. Modern philosophers of science have discussed extensively the concepts of probability and **causality** when the relationships are statistical rather than deterministic. Such considerations are expected, given the current physical models of the universe. Probabilities in deterministic situations, where maximal information is not available, that have some frequency distribution of the outcome, lead to frequency-based probabilities. Because one may wish to use probabilities in a setting where an

experimental set-up with inherent indeterminism may not be repeated, there is also a need for probability that is not personal or subjective but that also is not justified by appeal to the strong law of large numbers. However, here the probability is inherent in the laws governing certain situations (e.g. quantum mechanics in physics); such probabilities are sometimes called propensity-based probabilities (e.g. [6]). In most statistical settings these latter two concepts of probability are combined and called *frequentist*.

Inference in the statistical sense has been used to describe procedures that analyze data that come from (at least conceptually) some underlying set of probability distributions. When this is the case statements can often be made that only make sense in the context of the underlying distributions. For example, one constructs an interval in such a way that 95% of the time the mean of the underlying probability distribution for repeated samples from the distribution will lie in the interval; that is, one constructs a 95% **confidence interval**. Or one compares the change in sitting diastolic blood pressure from a baseline measurement to 12 weeks in two groups: one that is **randomized** to a placebo treatment and another group randomized to a new presumed antihypertensive drug. The ***P* value** for a treatment difference is used to summarize the strength of evidence for a treatment difference. Or, in the same experiment, a Bayesian **prior distribution** about the treatment differences is updated given the data from the experiment, and the probability that the change in blood pressure is more in the new drug group is used as a summary to show the new drug is effective. In each case the inference (in the every-day use of the word) is summarized by *statements that depend upon an underlying probability model combined with the observed data*. Such inference is *statistical inference*. Statisticians often perform other activities that are not statistical inference *per se*, but when combined with further processing of the data are associated with statistical inference in many situations. Descriptive statistics, including summary statistics (e.g. sample **mean, median, standard deviation**, minimum, and maximum), **graphic** plots (e.g. scatter diagrams, histograms, line graphs of mean or median values), computer visualization of multiple variables at a time, or two-dimensional projections of higher dimensional space, etc., are not statistical inference *per se*. However, plots with confidence sets or intervals would directly involve statistical inference.

The concepts of probability theory were integrated into one formal theory of statistical inference by **Jerzy Neyman** and **Egon Pearson** (e.g. [10]). They developed their formal framework for **hypothesis testing** introducing the familiar concepts of the **null hypothesis**, the **alternative hypothesis**, type I and II errors (*see* **Level of a Test**), **power**, etc. Some philosophers of science conclude that scientific paradigms can never be essentially (i.e. up to statistical variability) proven to be true. They can only be shown to be consistent with the facts at hand; however, further data or theory combined with data may show them to be inconsistent. While this is true for complex theories, particularly in physics, for other situations (e.g. a drug lowers blood pressure on the average in some population), "theories" or "facts" could be more clearly established conceptually. Hypothesis testing, combined with Occam's razor (the simplest possible explanation is to be preferred; *see* **Parsimony**), gives a paradigm for scientific endeavor. Hypothesis testing fits nicely into this paradigm as "null hypotheses" (usually straw men shown to be false) are to be rejected. The new theory had many other benefits: by selecting conventional levels for the significance level it led to an acceptable level of scientific proof that is largely used today both by scientific journals and regulatory authorities; it allowed experiments to be designed and sample sizes to be computed using the concept of statistical power, or equivalently type I and type II errors. Hypothesis testing about the parameters of a state of nature leads naturally to the concept of confidence regions and intervals because of the duality between hypothesis testing and confidence intervals (e.g. [2]).

Hypothesis testing was not without its problems. Cornfield [4, 5] lists problems with the formal hypothesis testing paradigm. For example, if the significance level is formally set at 0.05 and one performs an experiment and fails to reject a null hypothesis, then no amount of additional evidence should ever be allowed to lead to rejection of the null hypothesis! Furthermore, in a complex situation hypothesis testing can be used to plan experiments, but if one wants formally to take into account already known, but very complex, facts and/or beliefs, then this is difficult to incorporate rationally into the formal scientific inference. Bayesian statistics would appear to be the solution to the incorporation of complex facts already known. The Bayesian prior estimate of the state of nature (or the distribution of a parameter(s)

of interest) seems ideal for such situations. However, there are also problems with using Bayesian statistical methods [7]: (i) different individuals will have (often drastically) different prior beliefs – whose prior distribution should be used; (ii) humans are at best very imperfect Bayesians and cannot process data as probability models suggest we should [9]; and (iii) like the frequentist models, in practice difficulties and new data can arrive that would not have been adequately addressed in the elicitation of prior beliefs. (For example, new animal data suggest a toxicity problem with the long-term use of a drug or biologic; another drug with the same molecular mechanism of action reports findings.) There is no known method of statistical inference that can withstand all rational criticism. Thus, the actual application of statistics to important biostatistical problems necessarily is far from **algorithmic**; scientific judgment and reasoning, as well as intuition, often enter into important decision making in addition to formal statistical inference – as embodied, say, by hypothesis testing, confidence regions, model building, or Bayesian analysis. That is to say, while statistical inference is based upon the mathematical theory of probability, the decisions and understandings that result from the statistical inference have many arbitrary aspects, both technical (e.g. the significance level to be used in a test) and judgmental (e.g. an experiment has so much **missing data** that one decides to disregard it altogether). This results in important decisions using both statistical inference and other human decision making capacities in many instances. There is good reason to believe that humans are at best very imperfect Bayesian or frequentist statisticians [9].

The rapid continuing increase in computing power has led to innovations in statistical inference methodology. For example, the ability to resample from samples of distributions (e.g. **bootstrap** techniques), to implement **Markov chain Monte Carlo** methods, and to simulate from permutation distributions (especially the randomization distribution; *see* **Randomization Tests**) allow approaches to inference that use the same underlying ideas of the last 60 years but are new techniques that were not feasible a generation ago.

Another important path to understanding modern concepts of inference – especially in biostatistics – is to understand the important difference between observationally collected data and experimental data where the observer can intervene in the system to establish stronger scientific inference. For example, the history of medicine is replete with harmful treatments given for hundreds of years [1]. The **scientific method** and appropriate experimentation has led to rapid progress in science in general, and biology and medicine in particular. Until the advent of the scientific method (or, more realistically, a growing appreciation of the scientific method) plausible, but incorrect, systems of understanding human and animal biology were seemingly accepted if they were internally self-consistent and advocated by authority. Selected subject-matter application areas of biostatistics, besides medical biostatistics, may have other difficulties. For example, epidemiologic studies, ecologic, and wild animal studies are often necessarily restricted to observational data collection and/or mathematical modeling. The inability to experiment gives less cogent scientific inference and potentially a larger probability of mistaken "knowledge" due to potential underlying **biases**. No matter how firm the basis of statistical inference, the possibility of bias from unknown sources cannot be discounted. Other areas, such as plants for food and animal breeding, are more amenable to more classic experimentation; yet even here the heterogeneity is a large issue (say compared with electrons which are all assumed to be the same in particle physics). The idea of randomization, as introduced by **R. A. Fisher**, allows much more cogent experimentation, especially in human populations, than observational data or even less controlled experimental data that may be subject to unknown important biases.

In the previous paragraph the term *the scientific method* was used. There appears to be no entirely satisfactory definition that encompasses all the situations where one might use this term. Often books on the philosophy of science introduce it by implication. One definition is: "a method of research in which a problem is identified, relevant data are gathered, a hypothesis is formulated from these data, and the hypothesis is empirically tested" [13]. Such a definition is in accord with the statistical theory of hypothesis testing.

Statistical inference in the medical biological sciences has difficulties that do not arise in the physical sciences, at least to the same degree. One of the cornerstones of modern science is the ability to replicate results. If one group reports a simple method of cold fusion, then other experimenters around the

world may try to replicate the results. In experiments with animals, and especially humans, the sanctity of life introduces ethical concerns. In certain situations there may be a strong **ethical** and practical prohibition against replicating a result. If a therapy has been "shown" to prolong life as compared with a placebo, then further placebo-controlled experimentation may be considered unethical and/or impracticable. The lack of the ability to replicate can lead to the unchecked propagation of a false "fact". Furthermore, because of the need to monitor for patient benefit and/or safety, a minimal amount of information adequate for showing benefit or harm is the rule not the exception. That is to say, if a society decides that proof consists of rejecting a null hypothesis at the 0.05 significance level, then a randomized **clinical trial** with mortality as an endpoint might be argued to be unethical if the trial does not stop when reaching this level of significance, taking into account the **multiple comparison** issues of sequential monitoring. All experimentation seems beset with difficulties, and without a doubt Murphy was an optimist (*cf.* Murphy's Law: Anything that can go wrong will). However, human experimentation may have even deeper difficulties. Subjects may deliberately unblind therapy in a randomized trial (*see* **Blinding or Masking**). Subjects may exercise the right to withdraw, go on vacation, and miss a crucial follow-up visit, etc. The combination of these and other factors leads to more dispute *that is unresolved by convincing data* than in the more "hard" sciences. Statistical inference may investigate the **sensitivity** to such experimental deviations, but the cogency of the results is usually lessened in ways difficult to quantify.

Statistical inference and associated experimental design for drugs, biologics, and devices for human use is further complicated by very important practical matters. The rewards and development costs of new therapies and diagnostic tests for human use are both extremely large in many situations. Also, the competitive nature of the market-place plays a large role in the development of new modalities of treatment or diagnosis. The first sponsor of an approved modality is in a very favorable position. This, plus possible humanitarian reasons, puts a premium on the speed of development. The large stakes place intense pressures on both industry and regulatory agencies. This can result in a very strict adherence to statistical inference guidelines for such research. However, treatments – for example a drug – are rarely all good or bad. An appropriate dose needs to be found; furthermore, the appropriate amount and/or method of delivery may differ for important human subgroups (e.g. race, older individuals and children, individuals with impaired organ function, and genetically distinct subsets). The best designs and development programs from a statistical inference point of view may not be used because of other considerations.

In summary, statistical inference in biostatistics is formally the same as in other applied areas. However, practical and ethical issues can introduce limitations not seen in many other areas of applied statistics.

References

[1] Ackerknecht, E.H. (1973). *Therapeutics from the Primatives to the 20th Century*. Hafner, New York (translated from the German).

[2] Bickel, P.J. & Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco, p. 178.

[3] Clark, R.W. (1973). *The Life and Times of Einstein. An Illustrated Biography*. Wings Books, New York, p. 216.

[4] Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle, *American Statistician* **20**, 18–22.

[5] Cornfield, J. (1969). The Bayesian outlook and its application (with discussion), *Biometrics* **25**, 617–657.

[6] Fetzer, J.H. (1993). *Philosophy of Science*. Paragon House, New York, p. 97.

[7] Fisher, L.D. (1996). Commentary: comments on Bayesian and frequentist analysis and interpretation of clinical trials, *Controlled Clinical Trials* **17**, 423–434.

[8] Hald, A. (1990). *A History of Probability and Statistics and Their Applications before 1750*. Wiley, New York, Chapters 3–5.

[9] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York.

[10] Neyman, J. & Pearson, E.S. (1967). *Joint Statistical Papers of J. Neyman and E.S. Pearson*. University of California Press, Berkeley.

[11] Savage, L.J. (1972). *The Foundations of Statistics*. Dover Publications, New York.

[12] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. The Belknap Press of Harvard University, Cambridge, Mass., Chapter 2.

[13] *The Random House Dictionary of the English Language*. The Unabridged Edition (1967). Random House, New York, p. 1279.

(*See also* **Foundations of Probability**; **Inference**)

LLOYD D. FISHER

Inference

Inference is usually defined as the process of drawing conclusions from facts, available evidence, and premises. *Statistical inference* is the term associated with the process of making conclusions on the basis of data that are governed by probability laws. More generally conclusions are made that are uncertain. The objective measurement of the uncertainty is one of the principal goals of statistical inference. In practical applications of statistical inference data are available and the aim of the inference is to draw conclusions about models which potentially may have generated the data. *Data analysis* is the colloquial expression which is often used to describe the statistical inferential process.

Examples of such situations are: a randomized **clinical trial** is being carried out to determine which of two therapy programs is superior for the treatment of **AIDS**; a sample survey is taken in a community having a contaminated public water supply to determine if families having higher concentrations of contaminated water also have higher rates of congenital abnormalities; data are available from individuals diagnosed as having an acute myocardial infarct – how is the infarct incidence related to age and gender?

Implicit in the inferential process is a defined population and an experimental plan which describes how data is generated from the population. The experimental plan for data collection may be very well laid out, as in a clinical trial, or may be quite informal, such as data collected from a hospital to carry out an **observational study**. The less formal the experimental plan for data collection, the greater the opportunities for blunders and systematic biases. Furthermore, any conclusions drawn from the data, strictly speaking, apply only to the population from which the data have been generated. Populations that are not well defined also create opportunities for the injection of systematic errors in the inferential process. In what follows it will always be assumed that there is both a well-defined population and a data collection plan which does not create opportunities for systematic error.

There is a general lack of agreement on the best ways to carry out statistical inferences. These differences have led to different "schools of statistical inference". These "schools" are often referred to as the frequentist, likelihood, Bayesian, and fiducial schools of inference. Major articles appear in this Encyclopedia reflecting the different schools of inference. Important criticisms have been made against some of the ideas in each of these schools of inference. Even within a school there may be sharp disagreements. In this article the major views of different schools of inference will be compared, with special emphasis on the frequency school of inference.

The problem may be formulated by considering that data represent realizations of a **random variable** X having a family of **probability** distributions $\{P_\theta(x)\}$ which is indexed by θ. The random variable X and parameter θ may be vector valued. Realizations of X are denoted by x. To concentrate on ideas, we will assume that all operations described below are defined and any required regularity conditions are satisfied. We define $f_\theta(x)$ to be a probability density function (pdf) or frequency function of X. The main goal of the inference is to draw conclusions about θ. More generally, if θ is vector valued, then the inference may be concerned with drawing inferences on a subset of values of θ. The remaining parameters are referred to as **nuisance parameters**.

The most important class of inference problems is when the vector X is composed of independent identically distributed (iid) random variables having the joint distribution $\prod_i f_\theta(x_i)$. The **likelihood** is central to nearly all schools of inference and is defined by being proportional to the joint distribution, i.e. $L(\theta|x) \propto \prod_i f(x_i|\theta)$. Usually the likelihood is defined as equal to the joint distribution except for the omission of a multiplicative constant which does not depend on θ.

In some cases the likelihood function can be written as

$$L(\theta|x) = L_1(\theta|t(x))L_2(x),$$

where $t(x)$ is a function of the observations and may be a vector. In this case $t(x)$ is called a **sufficient statistic**. Hence the probability distribution of X, conditional on $t(x)$, is not a function of θ. As a result, $t(x)$ contains all the relevant data for making inferences about θ. A minimal sufficient statistic corresponds to the smallest dimension of $t(x)$ for which the distribution of X, conditional on $t(x)$, is not a function of θ. Note that the likelihood function is a sufficient statistic. Using only the sufficient statistic to make inferences on θ leads to a data reduction method

in which the sample x is replaced by $t(x)$. This data reduction does not lose any information on θ.

Frequentist School of Inference

The frequentist school of inference is the most widely used method of inference in practice. Much of the foundations were laid by Fisher [12–18], Neyman [23], and Neyman & Pearson [24–28]. However, Fisher and Neyman & Pearson have serious disagreements about basic issues. Articles describing aspects of the frequency school of inference are scattered throughout this Encyclopedia. Major articles are: **Hypothesis Testing**, **Estimation**, and **Maximum Likelihood**. Other articles discussing aspects of estimation are: **Confidence Intervals and Sets**, **Consistent Estimator**, **Cramér–Rao Inequality**, **Efficiency and Efficient Estimators**, **Generalized Maximum Likelihood**, **Minimum Variance Unbiased (MVU) Estimator**, **Sufficient Statistic**, and **Unbiasedness**. Articles discussing the frequentist theory of hypothesis testing and related topics are: **Alternative Hypothesis**, **Critical Region**, **Likelihood Ratio Tests**, **Most Powerful Test**, **Neyman–Pearson Lemma**, **Null Hypothesis** and **Level of a Test**. Two basic texts on frequentist inference are Lehmann [21, 22].

The basic idea underlying the frequentist school of inference is to evaluate the inferential process by assuming that an "experiment" is repeated an infinite number of times. Procedures having "better properties", as judged by long-term behavior, are deemed superior. Throughout the frequentist formulation of inference there is an attempt to derive statistical methods that have "optimal" properties in the context of an infinite number of repetitions of the experiment.

In general, all probability statements generated by the frequency theory of inference are based on the frequentist interpretation of probability. Yet the conclusions are targeted at specific data sets or specific experiments. Critics dismiss the concept of using methods based on properties associated with an infinite repetition of experiments. Outcomes that did not happen should not be used to evaluate observed outcomes. They point out that the goal of a data analysis is to make an inference from the particular experiment which has generated the data, not from a hypothetical infinite repetition of experiments. Widely used methods such as tests of significance and confidence procedures are subject to these criticisms. The critics agree that frequentist ideas may be relevant prior to carrying out an experiment, but are irrelevant after the experiment is carried out. Nevertheless, these frequentist-based methods have proven to be very useful in practice. Their applicability is continuing to expand despite the presence of sharp criticisms.

For example, suppose $X_1, X_2, \ldots, X_n$ represent iid random variables following a $\mathrm{N}(\theta, \sigma^2)$ distribution with θ unknown and σ^2 known. The $100(1-2\alpha)\%$ **confidence interval** is $\overline{x} \pm z_\alpha \sigma \sqrt{n}$, where $\overline{x}$ is the sample mean and z_α is the normal deviate which cuts off probability α in the tail of the normal distribution. The formal probability statement is $\Pr\{\overline{X} - z_\alpha \sigma/\sqrt{n} < \theta < \overline{X} + z_\alpha \sigma/\sqrt{n}\} = 1 - 2\alpha$. For any fixed $\overline{x}$, the population mean is either within the interval or is outside the interval. Hence, this statement only assigns a probability 0 or 1 that the population mean θ is included within the interval. The probability $(1-2\alpha)$ refers to the process of calculating such intervals over infinite repetitions of the experiment. Operationally, confidence coefficients are usually chosen to be high (95% or 99%) and individuals "act" as if the statement is correct that the population mean is included within the interval. An additional criticism of confidence intervals is that no distinction is made as to whether the population parameter is likely to have a higher probability of being in the neighborhood of $\overline{x}$ compared with being at the ends of the interval. Intuitively, most individuals would agree that the value of the parameter is more likely to be in a neighborhood of $\overline{x}$ compared with being located in a neighborhood around the boundary of the confidence interval.

The operational use of confidence regions is essentially associating a "degree of belief" with the statement that the parameter θ is located within the calculated confidence region on the basis of specific data. The high values chosen for confidence coefficients are so close to unity, that practitioners behave as if the statement is "certain". However, this same idea of using a degree of belief to measure the uncertainty of an inference can also be used to ascribe different degrees of belief for comparing values within a confidence region. To illustrate ideas, consider the calculation of confidence intervals for a mean as described earlier. Suppose a spectrum of confidence intervals is calculated by choosing different

confidence coefficients. The point $\overline{x}$ is a degenerate confidence interval with confidence zero. Then for any potential value of the population parameter θ', there corresponds a confidence interval for which θ' is the end-point, i.e., the normal deviate corresponding to $z_{\alpha'} = \sqrt{n}(\theta' - \overline{x})/\sigma$ (if $\theta' > \overline{x}$) or $z_{\alpha'} = \sqrt{n}(\overline{x} - \theta')/\sigma$ (if $\theta' < \overline{x}$). Then the ratio of degrees of belief comparing the value θ' relative to $\overline{x}$ is α' for the population mean. Hence each end-point of a 95% two-sided confidence interval has a degree of belief of 0.025 of being the population mean compared with the sample average, whereas each end-point of a 10% two-sided confidence interval ($z_{0.45} = 0.12$) has a degree of belief of 0.45 (collectively 0.90) of being the population mean relative to $\overline{x}$.

Among the most widely used techniques in the frequentist theory of inference is the test of significance (*see* **Hypothesis Testing**). Central to a test of significance are the **null** and **alternative hypotheses**, i.e. $H_0 : \theta = \theta_0$ and $H_1: \theta \neq \theta_0$. The alternative hypothesis can also be one-sided, $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$. The test of significance consists of calculating evidence which is "unfavorable" to the null hypothesis. The test of significance consists of using a statistic $T(x)$ so that large values of $T(x)$ indicate departures from the null hypothesis $H_0: \theta = \theta_0$. If $T_0(x)$ represents the value of the statistic from an experiment, then the test of significance calculates $P = \Pr\{T(X) \geq T_0(x) | \theta = \theta_0\}$. This so-called ***P* value** is interpreted as discrediting the null hypothesis if P is small (usually $P \leq 0.05$) and in favor of the null hypothesis if P is large. The role of the alternative hypothesis is to specify the statistic $T(x)$.

The P value is widely interpreted as summarizing the statistical evidence of an experiment. If a small value is calculated ($P \leq 0.05$), then either one has observed a rare event if H_0 is true, or if H_0 is not true, then the model for carrying out the calculation (assuming $\theta = \theta_0$), is wrong. Ordinarily the conclusion is made that the model is incorrect and H_0 is rejected.

The logic of the significance test is that if the observed $T_0(x)$ is evidence against the null hypothesis, then larger values of $T(x)$ would constitute even stronger evidence against H_0. The logic of significance tests is questioned, in that a hypothesis may be rejected on the basis of experimental outcomes which were not observed.

The test of significance does not recognize that there may be different interpretations on the P value which are dependent on both the sample size and the magnitude of the deviation from H_0. An experiment in which there is a negligible deviation from the null hypothesis, but having a very large sample size, will result in a small P value, whereas an experiment in which there is a large deviation from H_0, but with a small sample size, may not result in small P value. The uncritical use of tests of significance may result in misleading conclusions. It is often recommended that: (i) if P is small, then information should be presented on the magnitude of the deviation from the null hypothesis, and (ii) if P is large, then evidence should be presented on the power of the test.

Randomization and Permutation Tests

One of the important applications of significance tests is when the sample space is generated by the investigator. This has no analogy with any of the other methods of inference. It occurs whenever an experimental design makes use of randomization. We shall refer to these tests as **randomization tests**.

To illustrate ideas, consider an experiment where the location parameters of two treatment groups are to be compared by a significance test. The most widely used example is a randomized clinical trial in which patients are assigned to each of two treatment groups such that each patient has the same probability of being assigned to each group. If there are $2n$ patients available for the experiment and each group is assigned n patients, then there will be

$$N = \binom{2n}{n}$$

possible assignments. Hence there will be N points in the sample space. Suppose the null hypothesis is that there is no difference in outcome among the two groups. If $\overline{x}_1$ and $\overline{x}_2$ represent the sample average for each group, then $\overline{x} = (\overline{x}_1 + \overline{x}_2)/2$ will always be constant for the experiment. Hence, to show a difference in the location parameters the differences between the two sample averages $\overline{x}_1 - \overline{x}_2 = (\overline{x}_1 - \overline{x})$ will be considered. The sample space will consist of N possible values of the differences between the two sample averages, each having probability equal to $1/N$ of arising due to the randomization. Hence, if $D_0 = |\overline{x}_1 - \overline{x}_2|$ represents the absolute value of

the observed difference between the group sample averages, then the test of significance calculates $P =$ (number of absolute differences $\geq D_0)/N$. No further assumptions need to be made about the probability distributions of the outcomes. Essentially the randomization tests are distribution-free. The validity of the procedure is justified by the randomization which injected probability into the experiment. Of course, the inference only applies to patients who are in the clinical trial, i.e. the inference procedure concludes which is the best treatment for the population of patients that have been entered in the trial. To have a broader inference of making the conclusions apply to the population of patients having disease, it would be necessary to have a random sample of patients entered on the clinical trial.

A closely related set of procedures are *permutation tests*. We distinguish between randomization and permutation tests. Randomization tests are characterized by the investigator purposely introducing probability into the experimental design, whereas in permutation tests it is *assumed* that the sample space consists of equally likely outcomes. Consider the following example of a permutation test. Suppose the water supply in a community was a blend coming from several sources. Depending on the location of the residence, there would be different amounts of water from each source in the blend of water available to each residence. One of the water sources was found to be contaminated. From the time the contaminated source was put into service until the discovery of contamination there were 20 live births in the community. Two babies were born with congenital abnormalities. The amount of water going to each mother's residence during her pregnancy was known during this period of time. Is there an association between the contaminated water and birth defects? A permutation test would assume that each baby has the same risk of having a congenital abnormality. Hence there will be

$$\binom{20}{2} = 190$$

different ways in which two birth defect infants can be distributed among the 20 infants. If the residences with the two highest amounts of contaminated water also had the two birth defects, then this could happen with probability $P = 1/190 = 0.005$ if there is no relationship between birth defects and the contaminated water supply. This probability is so low that the frequentist would conclude a relationship between contaminated water and birth defects. Most people would intuitively agree there may be a relationship. A P value of 0.05 would arise if the two birth defect babies came from residences in which the amount of contaminated water delivered to the households during pregnancy ranked fourth and fifth highest. There are nine other more extreme outcomes than the observed fourth and fifth. If pairs of numbers represent the rankings then these outcomes are: (1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4) and (3, 5). Since under the permutation test assumption any one of the possible 190 outcomes has the same chance of occurring, the number of outcomes equal to or more extreme than the one observed is $10/190 = 0.053$.

The randomization and permutation tests have generated the field of distribution-free or **nonparametric** methods. These methods do not require knowledge of the probability distribution of the observed outcome, as our two examples illustrate. To ease computations the observations are often replaced by **ranks** or **scores**. Very little statistical efficiency is lost by these substitutions.

Conditioning

An important modification of frequentist inference is the possibility of using information (data) to consider only a subset of the sample space. This may be done by conditioning on some aspect of the observed data which will result in a reduced sample space (*see* **Conditionality Principle**). The conditioning can only be done after the experiment has been completed and the data are available. Fisher [17] has advocated conditioning on the relevant subset of the sample space. The conditional sample space is also referred to as *recognizable subsets* or *reference sets*.

Cox [10] presented a very interesting example which leads one to make a conditional inference on a recognizable subset. His example is as follows. Suppose there are two normal populations, $\mathrm{N}(\theta, \sigma_1^2)$ and $\mathrm{N}(\theta, \sigma_2^2)$, having the same mean, but different variances. The mean is unknown, but the variances are known with $\sigma_1^2 \gg \sigma_2^2$. The experiment consists of choosing a population with probability 1/2 and drawing one observation, x. The population is known which is sampled. Consider a test of H_0: $\theta = 0$ vs. H_1: $\theta = \theta' \simeq \sigma_1$. Consider two tests – a conditional and an unconditional test to be made at an $\alpha = 0.05$ level of significance. The conditional test is

made on the population from which the sample was drawn. This leads to rejection regions $x > 1.64\sigma_1$ or $x > 1.64\sigma_2$ depending on which population has been sampled. However, this is not the most powerful test over the entire sample space. Application of the Neyman–Pearson theory results in a test which approximately has rejection regions $x > 1.28\sigma_1$ or $> 5\sigma_2$ depending on which population has been sampled. If the sample is from the second population, then almost complete discrimination is made between $\theta = 0$ against a much larger value of θ. As a result we can have a significance level of 10% if one is sampling from the first population. The power of the first test is 0.26 if population one is sampled and nearly unity if population 2 is sampled. Alternatively, the unconditional test has a power of 0.80. Thus, considering the overall sample space the first test has an average power of 0.63. Cox states

> if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test just given is in order .... If, however, our object is to say "what can we learn from the data we have", the unconditional test is surely no good. The unconditional test says that we can assign a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some other distributions. But this fact seems irrelevant to the interpretation of an observation which we know may come from a distribution with variance σ_1^2. That is, our calculations of power, etc. should be made conditionally within the distribution known to have been sampled.

Cox's example shows that the inference procedure should not be determined solely by considerations of power when one is considering repetitions of the experiment. In this example the indicator of the population sampled is called an **ancillary statistic** because it contains no information about θ. In general an ancillary statistic is defined as a function of the observations whose distribution is not a function of the parameter. In Cox's example, if δ is an indicator variable indicating which experiment is chosen, then the data are (δ, x) and δ is an ancillary statistic. The recognizable subset conditions on δ, i.e. $f(x|\delta)$.

Another illuminating example is provided by Berger & Wolpert [5]. Assume $X_1, X_2 \ldots, X_n$ are iid having a **uniform distribution** over the interval $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. The sufficient statistics are $U = \min(X_i), V = \max(X_i)$. Their joint distribution is $f(\mu, v) = n(n-1)(v-u)^{n-2}, \theta - \frac{1}{2} < \mu \leq v < \theta + \frac{1}{2}$. However, $R = V - U$ is an ancillary statistic because its distribution is independent of θ. The distribution of (U, V) conditional on $R = r$ is uniform over the interval $\theta - \frac{1}{2} \leq \mu < \theta + \frac{1}{2} - r$ and should be the starting point for the statistical inference. In particular, a $100(1-\alpha)\%$ confidence interval is $(u+v)/2 \pm (1-r)(1-\alpha)/2$. In this example it is clear that the range is an ancillary statistic. However, in other situations the proper ancillary statistic may not be obvious and there may be competitive ancillary statistics.

The problem becomes more complex when $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and the inference is to be made on $\boldsymbol{\theta}_1$ with $\boldsymbol{\theta}_2$ being regarded as a vector of nuisance parameters. If the likelihood factors into

$$L(x|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = L_1(\boldsymbol{\theta}_1|x, a(x))L_2(\boldsymbol{\theta}_2|a(x)),$$

then the distribution of $a(x)$ is independent of $\boldsymbol{\theta}_1$ and $a(x)$ is ancillary for $\boldsymbol{\theta}_1$. An example of this likelihood decomposition is when the data consists of pairs $(Y_i, X_i), i = 1, 2, \ldots, n$, which are iid following a bivariate normal distribution with $\mathrm{E}(Y_i) = \mu_y$, $\mathrm{var}(Y_i) = \sigma_y^2$, $\mathrm{E}(X_i) = \mu_x$, $\mathrm{var}(X_i) = \sigma_x^2$, and $\mathrm{cov}\,(X_i, Y_i) = \rho\sigma_x\sigma_y$. The object of the inference is on the parameters $\alpha = \mu_y - \beta\mu_x$ and $\beta = \rho\sigma_y/\sigma_x$ as $\mathrm{E}(Y_i|X_i) = \alpha + \beta X_i$. The likelihood can be written

$$\begin{aligned} &L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho|x, y) \\ &\quad = L_1(\mu_y, \tau^2, \alpha, \beta|x, y)L_2(\mu_x, \sigma_x^2|x), \end{aligned}$$

where $\tau^2 = \mathrm{var}(Y_i|X_i) = \sigma_y^2(1-\rho^2)$ and

$$\begin{aligned} &L_1(\mu_y, \tau^2, \alpha, \beta|x, y) \\ &\quad = \tau^{-n} \exp\left\{-\sum_{i=1}^{n} \frac{[y_i - (\alpha + \beta x_i)]^2}{2}\tau^2\right\} \\ &L_2(\mu_x, \sigma_x^2|x) \\ &\quad = \sigma_x^{-n} \exp\left\{-\sum_{i=1}^{n} \frac{(x_i - \mu_x)^2}{2\sigma_x^2}\right\}. \end{aligned}$$

Thus the regression analysis is conditional on the observed values of x_i which are treated as fixed constants.

When the likelihood factors can be written in terms of minimal sufficient statistics which factor into $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|t_1, t_2) = L_1(\psi|t_1, t_2), L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|t_2)$, then it is possible to consider the distribution of t_1 conditional

on t_2 in order to make inferences on $\psi = \psi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. This is the case for the **two by two contingency table** for comparing two **binomial distributions**. Conditioning on the total number of successes allows an inference to be made on the ratio of two **odds** in which one of the success probabilities is regarded as a nuisance parameter. Similarly, for comparing two **Poisson distributions**, with rate parameters (θ_1, θ_2), the likelihood can be written

$$L(\theta_1, \theta_2 | s_1, s_2) = \left(\frac{\theta_1}{\theta_2}\right)^{s_1} \theta_2^t \exp[-(\theta_1 + \theta_2)]$$
$$= L_1\left(\psi = \frac{\theta_1}{\theta_2} | s_1\right) \times L_2(\theta_1, \theta_2 | t = s_1 + s_2),$$

which admit of an inference on ψ by conditioning on the sum of the observed events. In both of these examples t contains "no information" about the parameter of interest, ψ. Cox [10] proposed a criterion for determining if t gives no information about ψ when nuisance parameters are present.

An important use of invoking a conditional inference arises in some censoring situations. To illustrate ideas suppose an investigator is testing a drug on patients in which the outcome is success or failure. The experiment is carried out until one observes a single failure. Hence the number of observations is a random variable following a **geometric distribution**. However, the investigator has only enough drug to treat 10 patients. The experiment could then have a maximum of 10 patients and the truncated distribution for the sample size would be

$$\Pr\{N = n | N \leq 10\} = \frac{\theta^{n-1}(1-\theta)}{(1-\theta^{10})}, \quad \text{for } n = 1, 2, \ldots, 10,$$
$$\Pr\{N = 10\} = \theta^9.$$

The experiment is carried out and the fifth patient had a failure. Should the statistician use the likelihood $\theta^4(1-\theta)$ in making the inference, or the truncated likelihood? To continue the story, the investigator tells the statistician afterwards that just before the experiment started the drug manufacturer had agreed to make available as much drug as needed. Hence there would be no need for the truncated distribution. In a final development, the drug manufacturer changed its offer to only make available the amount of drug for a maximum of 20 patients. This would change the truncated distribution. What should the statistician do? The actual experiment did not need the extra drugs, but an unconditional inference would have required taking account of the limited supply. The change of mind of the drug manufacturer would change the truncated probability distribution. Yet the manufacturer's decision had nothing to do with the actual experiment that was carried out with the available drug supply. Common sense dictates that the inference should be made conditional on what happened, not what could have happened. There was enough drug to carry out the experiment as planned. The likelihood should be conditional on the actual drug supply expended.

Estimation

The most widely used methods of **estimation** among frequentists are **minimum variance unbiased (MVU) estimation** and the method of **maximum likelihood**. Other methods in use are the **method of moments** and **generalized estimating equations**. The principle of having **unbiased** estimates sometimes leads to problems. For example, suppose $\overline{X}$ is the sample average of a sequence of n iid random variables having a $N(\theta, \sigma^2)$ distribution with σ^2 known. Then the minimum variance unbiased estimate of θ is the sample average. However, if it is desired to estimate θ^2, then we note that $E(\overline{X}^2) = \theta^2 + \sigma^2/n$. Therefore, if $T = \overline{X}^2 - \sigma^2/n$, $E(T) = \theta^2$ and T is an unbiased estimate of θ^2. However, θ^2 is always nonnegative, but there is a positive probability that $T = \overline{X}^2 - \sigma^2/n$ will be negative giving a nonsensical estimate. In general, if $g(\theta)$ is some function of θ, and T is an unbiased estimate of θ, then $g(T)$ is ordinarily not an unbiased estimate of $g(\theta)$. Despite some anomalies with the concept of unbiased estimation, the applications of the minimum variance unbiased criteria to estimation problems is very useful in applications – especially for models linear in the parameters.

Another widely used estimation procedure is the method of maximum likelihood. Estimates of θ, denoted by $\hat{\theta}$, are formed by maximizing the likelihood function, i.e. $L(\theta|x) = \max_\theta L(\theta|x)$. The properties of the maximum likelihood estimates are that they are: consistent, asymptotic minimum variance unbiased, and asymptotically normal. In addition, the

maximum likelihood estimate has the property that the estimate of $g(\theta)$ is $g(\hat{\theta})$.

Both minimum variance unbiased estimation and maximum likelihood estimation supply both point and confidence region estimates. Their justification is based on properties associated with infinite repetitions of sampling. However, Bayesian ideas have been used to justify maximum likelihood estimation.

Likelihood School of Inference

The use of the likelihood function is basic to many of the methods associated with frequentist inference. It was introduced by Fisher [12–16] as an information summary. It is essentially a minimal sufficient statistic for θ. Edwards [11] discusses the history of the likelihood function, and Berger & Wolpert [5] contains a thorough development of the statistical implications when the likelihood function is used as a basis for inference. Many of the ideas building on Fisher's early work are attributed to Barnard [1–4] and Birnbaum [6, 7].

Fisher initially introduced the likelihood function to obtain maximum likelihood estimates of parameters. The justification for maximum likelihood estimation has been made in terms of its large-sample properties relying on the frequency concepts of probability. Furthermore, ratios of likelihoods form the basis of likelihood ratio tests and the general Neyman–Pearson theory of hypothesis tests. However, all properties of these methods are judged by their behavior over the entire sample space of possible observations. This is at odds with the likelihood school of inference who regard the sample space as irrelevant after the experiment has been done. The only relevant quantities are the sample data, x, and its incorporation into the likelihood function, $L(\theta|x)$.

The basis of all **Bayesian inference** is the likelihood function. If $p(\theta)$ is the **prior distribution** on θ and $\pi(\theta|x)$ is the posterior distribution, then we have the well-known relationship $\pi(\theta|x) \propto L(\theta|x)p(\theta)$. Hence, whatever implications for inference arise from the likelihood function also apply to Bayesian inference methods.

The basis for the use of the likelihood in statistical inference is contained in the likelihood principle (*see* **Foundations of Probability**). It states that all information about θ from an experiment is contained in the likelihood function. Furthermore, two likelihood functions (from the same or different experiments) contain the same information about θ if they are proportional to one another.

The importance of the likelihood function in inference arises from its justification based on the widely accepted ideas of sufficiency and conditionality. Conversely, the likelihood function has been shown to lead to sufficiency and conditionality. These proofs were originally made by Birnbaum [6] and are correct for discrete observations. Others have modified his arguments for the continuous case. We summarize the main ideas borrowing from the development of Berger & Wolpert [5]. The ideas of conditionality and sufficiency have been discussed earlier in this article. There are two versions of each which are modified by the adjectives "weak" and "strong". We informally state the weak versions, which are all that is necessary in the proofs found in the literature.

Weak Conditionality Principle (WCP). Suppose there are two or more possible experiments, each having possibly different probability distributions, and each giving rise to different experimental outcomes. However, they all have in common the same parameter, θ. Consider the mixed experiment in which experiment i is chosen to be carried out with probability p_i. Then the WCP states that the evidence about θ from the mixed experiment is the experiment actually performed.

Weak Sufficiency Principle (WSP). Consider an experiment in which $t(X)$ is a sufficient statistic for θ. Then if x_1 and x_2 represent two different outcomes, but $t(x_1) = t(x_2)$, then the evidence about θ is the same for each outcome.

It has been proved that the WCP and the WSP imply the likelihood principle and conversely the likelihood principle implies both the WCP and WSP. The proof for the continuous case can be found in Berger & Wolpert [5].

The proponents of inference based on the likelihood principle view the rejection of the likelihood principle as also logically rejecting the WSP or WCP. However, the WSP is one of the basic ideas in frequency inference. The WCP is regarded as simply "common sense".

The likelihood principle is incompatible with many of the methods used in the frequency theory of inference. For example, **randomization**, significance tests, hypothesis testing, confidence intervals, and randomization tests are all contraindicated by the

likelihood principle. It is of interest that although randomization is rejected by the WCP, there has been no effort to negate randomized clinical trials. The idea of randomized clinical trials and the general idea of randomization appear to have been accepted by the likelihood and Bayesian schools of inference, notwithstanding the variance with the likelihood principle. One reason for accepting randomization is that it is regarded as a way to obtain "balance" among different groups being compared with respect to unknown factors affecting the outcomes.

Censoring which did not occur is considered irrelevant. For example, suppose in a clinical trial patients are only followed for a maximum period of time (say 10 years). If one had data from such a trial where the end-point was death and all patients died within the 10-year period, then the censoring at 10 years is of no consequence. Yet in calculating the behavior of a frequentist statistical procedure, it is necessary to consider infinite repetitions of the trial in which some patients may have survived 10 years and would be censored.

Stopping rules are deemed irrelevant. Hence, **sequential** methods are treated as if data arose from fixed-size experiments. For example, if X is $\mathrm{N}(\theta, \sigma^2)$ then one may sample from this population until the sample average exceeds a fixed constant, i.e. $\overline{x} > k\sigma/\sqrt{n}$. By the law of the iterated logarithm (*see* **Limit Theorems**), there is a finite probability that the event will happen. However, for (say) $\theta = 0$ and large k, the necessary number of observations may be very large. If a frequentist desires to exclude $\theta = 0$ from a 95% confidence limit, then it is only necessary to choose $k = 1.96$. Of course, this entire procedure would be misleading from a frequentist point of view. The likelihood argument is that the inference should not interpret the usual confidence interval in the frequency sense. A frequentist would also take account of the ultimate sample size in placing confidence intervals on θ. Alternatively, the Bayesian approach to this problem is to incorporate the possibility of $\theta = 0$ in a prior distribution.

Suppose one observed four successes in 10 trials from sampling a binomial distribution. This gives rise to the likelihood function $L(\theta|x) = \theta^4(1-\theta)^6$, where θ is the probability of success in a single trial. Alternatively, suppose one samples from this population until four successes are observed. If this experiment took 10 observations to observe four successes, then it will give rise to the same likelihood as observing four successes in 10 trials. The likelihood principal would treat both experiments as generating the same information even though the sampling distributions associated with each experiment are different.

One of the principal criticisms of the use of the likelihood function for inference is the need to specify the model generating the data. Furthermore, it does not encompass any nonparametric methods.

Methods for solely using the likelihood function for inference are not well developed. We cite a few examples of the use of a likelihood function for making inferences.

The ratio of likelihoods can measure the relative support of two values of θ, e.g. $L(\theta_1|x)/L(\theta_2|x)$. Since the maximum likelihood estimate $\hat{\theta}$ maximizes the likelihood, it can serve as a normalization factor and one may consider $L(\theta|x)/L(\hat{\theta}|x)$ to measure the relative support of any θ to $\hat{\theta}$. Likelihood contours may be calculated by setting $L(\theta|x)/L(\hat{\theta}|x) = k$ for a range of values of k. If we consider values of the parameter satisfying

$$\frac{L(\theta|x)}{L(\hat{\theta}|x)} \leq k,$$

then, we can obtain an expression analogous to a confidence region for θ.

Using the likelihood when nuisance parameters are present raises complications. Suppose $\theta = (\theta_1, \theta_2)$ and θ_2 is a nuisance parameter. One approach is to substitute the maximum likelihood estimate of $\hat{\theta}_2(\theta_1)$ in the ratio $L(\theta_1, \hat{\theta}_2(\theta_1)|x)L(\hat{\theta}_1, \hat{\theta}_2|x)$. For example, consider the likelihood function of a sample of n observations from a $\mathrm{N}(m, \sigma^2)$ distribution, i.e.

$$L(m, \sigma^2|x) = \sigma^{-n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m)^2}{\sigma^2}\right\}.$$

The maximum likelihood estimate of σ^2, as a function of m, is $\hat{\sigma}^2(m) = \sum_{i=1}^{n}(x_i - m)^2/n$. Then one has

$$\frac{L(m, \hat{\sigma}^2(m)|x)}{L(\hat{m}, \hat{\sigma}^2|x)} = \left\{1 + \frac{(\hat{x} - m)^2}{s^2}\right\}^{-n/2} \cong \left\{1 + \frac{t^2}{n}\right\}^{-n/2},$$

where $ns^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$ and $t^2 = n(\hat{x} - m)^2/s^2$. This is the **Student *t* distribution** (except for the normalizing constant), but with n in place of $(n-1)$. If we desired bounds on m, then we could set

$L(m, \hat{\sigma}^2(m)|x)/L(\hat{m}, \hat{\sigma}^2|x) \leq k$. Then for large n we have the interval $\hat{x} \pm s(-2\log k)^{1/2}/\sqrt{n}$.

The general application of likelihood methods may be extended by noting that for large samples

$$\frac{L(\theta|x)}{L(\hat{\theta}|x)} \simeq \exp\left\{\frac{-(\theta-\hat{\theta})^2}{2\sigma^2(\hat{\theta})}\right\},$$

where

$$\sigma^2(\hat{\theta}) = -\left(\frac{\partial^2 \log\theta}{\partial\theta^2}\right)^{-1}_{\theta=\hat{\theta}}$$

Hence, the maximum likelihood estimate and the estimate of its asymptotic variance can be used to make likelihood-type inferences for large samples. The result extends to the multivariate situation.

Bayesian School of Inference

In this section we illustrate and contrast Bayesian methods of inference with frequency methods. The application of Bayesian methods to problems of medicine and biology is growing. In large measure the expansion in applications is due to the development of new computing **algorithms** which allow the calculations of posterior distributions having large numbers of parameters. This class of computer algorithms are called **Markov chain Monte Carlo** methods. Two widely used algorithms for this purpose are the Gibbs sampler and the Metropolis-Hastings algorithms: cf. Tanner & Wong [34], Smith [31], Tierney [35], Smith & Gelfand [32], and Smith & Roberts [33]. Breslow [8], in a review paper, cites many applications of Bayesian methods to biostatistics, i.e. **longitudinal data** models, **small area estimation**, **risk assessment** based on species to species **extrapolation**, **bioequivalence** and sequential clinical trials (*see* **Data and Safety Monitoring**).

The Bayes paradigm is that all inferences are based on calculating the posterior distribution of θ. The difficulty in utilizing Bayesian methods is due to both the dependence on model specification, and the meaning of Bayesian probability statements. To utilize Bayesian methods it is necessary to specify both the likelihood function and a prior distribution. Prior distributions may be chosen by subjective opinion or may reflect previous data or knowledge. Issues arise when informationless prior distributions are chosen which contain no information or parameters. The interpretation of a Bayesian probability is that it measures a degree of belief. It allows attaching a posterior probability (degree of belief) associated with hypotheses. This is in contrast to frequentist inference which attaches a value of 0 or 1 to the truth of a hypothesis. Bayesian methods are often judged in practice by their behavior over infinite repetitions. The book by Savage [30] still remains as a cogent treatise advocating the use of Bayesian methods.

Elements

The basis of all Bayesian inference is that, "Any inferential process that does not follow from some likelihood function and some set of priors has objectively verifiable deficiencies", cf. [9]. Bayesian inference places probability distributions on parameters. The elements of Bayesian inference are: that a prior distribution $p(\theta)$ summarizes information about θ prior to experimentation; the likelihood $L(\theta|x)$ incorporates information utilizing data; and the posterior distribution $\pi(\theta|x)$ depicts the probability distribution of θ after incorporating the data. More formally, the relationship between these quantities is

$$\pi(\theta|x) \propto L(\theta|x)p(\theta).$$

This expression is a direct consequence of Bayes' theorem and shows how prior beliefs are changed with the availability of data.

The interpretation of $\pi(\theta|x)$ is that it is a degree of belief, taking on values within the unit interval. In what follows it will be assumed for simplicity that θ has a prior distribution having a probability density function. This assumption is not necessary, but it eases the formalism.

Ratios of posterior distributions are often used to indicate "support" for comparing two different values of θ, i.e. $\pi(\theta_1|x)/\pi(\theta_2|x)$. If one of the θs is the mode of $\pi(\theta|x)$ (denoted by θ_{m}), then $\pi(\theta|x)/\pi(\theta_{\mathrm{m}}|x)$ compares the degree of belief of an arbitrary θ with the modal value. If θ is one dimensional, then a graph of the ratio vs. θ is particularly useful.

Although there is a great deal of debate on the choice of the prior distribution, the importance of the prior distribution diminishes when the sample size is large. This is easily seen by noting that if we write $L(\theta|x) = \prod_{i=1}^{n} L(\theta|x_i)$, then

$$\log[\pi(\theta|x)] = \log[L(\theta|x)p(\theta)]$$
$$= \sum_{i=1}^{n}\left[\log L(\theta|x_i) + \frac{1}{n}p(\theta)\right],$$

and noting that the second term in brackets goes to zero as $n \to \infty$.

The normalizing constant $P(x)$ is $P(x) = \int_\Omega L(\theta|x)p(\theta)\,\mathrm{d}\theta$ and is the expected value of the likelihood averaged over the prior distribution. The integral is over the parameter space of θ. Hence, the posterior distribution can be written

$$\pi(\theta|x) = \frac{L(\theta|x)p(\theta)}{P(x)}.$$

Note that $P(x)$ is proportional to the posterior probability of observing the data x.

An important aspect of Bayesian inference is predicting a future observation (*see* **Prediction**). If y represents a future observation and x represents data already observed, then the predictive likelihood of y is

$$L(y|x) = \int_\Omega L_1(\theta|y)\pi(\theta|x)\,\mathrm{d}\theta,$$

where $L_1(\theta|y)$ is the likelihood of a single new observation. The quantity $L(y|x)$ is the likelihood of observing a future observation given the data represented by x. If $L(y|x)$ is normalized, then

$$f(y|x) = \frac{L(y|x)}{\int_{-\infty}^{\infty} L(y|x)\,\mathrm{d}y}$$

is the predictive distribution of y. It is clear that the predictive distribution is not necessarily restricted to predicting a single observation, but can predict an arbitrary number of observations. The book by Geisser [19] is a principal reference for predictive distributions.

A fundamental problem in statistical inference is to carry out an inference when nuisance parameters are present. The methods of Bayesian inference can deal with the problem in a relatively straightforward way. Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and inference is to be made on $\boldsymbol{\theta}_1$ with $\boldsymbol{\theta}_2$ regarded as a vector of nuisance parameters. The methods of Bayesian inference deal with this problem by considering the marginal posterior distribution of $\boldsymbol{\theta}_1$, i.e.

$$\pi(\boldsymbol{\theta}_1|x) = \int_{\Omega_2} \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|x)\,\mathrm{d}\boldsymbol{\theta}_2.$$

The comparable problem in the frequency theory of inference can be carried out only in special cases when minimal sufficient statistics exist.

Example

To illustrate ideas we shall consider the Bayesian analysis for comparing two binomial distributions. If (p_i, s_i, n_i) represent the success probabilities, number of successes, and sample sizes, respectively, for $i = 1, 2$, then the likelihood is

$$L(p_1, p_2|s_1, s_2) = p_1^{s_1}(1-p_1)^{n_1-s_1} p_2^{s_2}(1-p_2)^{n_2-s_2}.$$

Define the new parameters (α, β) by the logit transformations (*see* **Logistic Regression**), i.e.

$$\log\left[\frac{p_1}{1-p_1}\right] = \alpha, \quad \log\left[\frac{p_2}{1-p_2}\right] = \alpha + \beta.$$

Note that $\mathrm{e}^\beta = p_2(1-p_1)/p_1(1-p_2)$. The reparameterized likelihood can be written

$$L(\alpha, \beta|s, t) = \frac{\mathrm{e}^{\alpha t + \beta s}}{(1+\mathrm{e}^\alpha)^{n_1}(1+\mathrm{e}^{\alpha+\beta})^{n_2}},$$

where $t = s_1 + s_2$ and $s = s_2$. The reparameterization allows one to test H_0: $p_1 = p_2$ by considering β only. The parameter α is a nuisance parameter.

The sampling theory of inference considers the distribution of s conditional on t, which results in α being dropped. This is the basis of the analysis of 2×2 tables with all marginal totals fixed. The explicit conditional distribution is

$$f(s|t, \beta) = \frac{C(s,t)\mathrm{e}^{\beta_s}}{\sum_{z=0}^{r} C(z,t)\mathrm{e}^{\beta z}}, \quad s = 0, \ldots, r,$$

when $r = \min(t, n_2)$ and

$$C(s,t) = \binom{n_1}{t-s}\binom{n_2}{s}.$$

The Bayesian analysis finds the posterior distribution of (α, β) from which the marginal distribution of β can be calculated. Let the prior distributions of (α, β) be taken as

$$\frac{p(\alpha, \beta) \propto \mathrm{e}^{\alpha t' + \beta s'}}{(1+\mathrm{e}^\alpha)^{n_1'}(1+\mathrm{e}^{\alpha+\beta})^{n_2'}},$$

where the prime quantities are parameters of the prior distribution, but subject to the condition $0 \le s' \le t' \le$

$n_1' + n_2'$. Then the posterior distribution is

$$\frac{\pi(\alpha, \beta|s, t) \propto e^{\alpha t'' + \beta s''}}{(1 + e^{\alpha})^{n_1''}(1 + e^{\alpha+\beta})^{n_2''}},$$

where $s'' = s + s'$, $t'' = t + t'$, $n_1'' = n_1 + n_1'$, and $n_2'' = n_2 + n_2'$. Since the form of the posterior is the same as the likelihood, the prior distribution is called a *natural conjugate* or *conjugate distribution*. Finally, the marginal posterior distribution is

$$\pi(\beta|s'', t'') \propto e^{\beta s''} \int_0^1 \frac{v^{t''-1}(1-v)^{n''-t''-1}\,dv}{[1 - v + ve^{\beta}]^{n_2''}},$$

where $n'' = n_1'' + n_2''$. The integral can be found using numerical methods (*see* **Numerical Integration**).

Prior Distributions

The choice of a prior distribution is an important first step in implementing a Bayesian analysis. Prior distributions may be chosen on the basis of other similar experiments, subjective opinion, or the acknowledgment that nothing is known about the parameters, and the prior is informationless.

There is a great deal of debate when the prior distribution is informationless. Jeffreys [20] proposed that if a parameter takes on values over the real line, that the prior distribution be uniform, whereas if θ takes on values over the positive real line, that $\log\theta$ have a uniform distribution. Therefore the informationless prior for location and scale parameters, as suggested by Jeffreys, are $p(m) = 1\ (-\infty < m < \infty)$ for a location parameter and $p(\sigma) = 1/\sigma\ (0 < \sigma < \infty)$ for a scale parameter. Both are improper distributions in that their integrals over the parameter space do not exist. Another view of the improper prior distribution for the scale parameter is that it is flat over the range of the parameters of the likelihood functions. Therefore the posterior distribution is essentially the likelihood function.

Nevertheless the posterior distributions do satisfy all conditions for distribution functions. For example, suppose $X_1, \ldots, X_n$ are iid $N(m, \sigma^2)$ with m and σ^2 both unknown. The likelihood function is

$$L(m, \sigma^2|\mathbf{x}) = \sigma^{-n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m)^2}{2\sigma^2}\right\},$$

resulting in the posterior distribution

$$\pi(m, \sigma^2|\mathbf{x}) \propto \frac{L(m, \sigma)}{\sigma}.$$

If the marginal distribution of m is obtained by integrating over σ, then we obtain

$$\int_0^{\infty} \pi(m, \sigma^2|\mathbf{x})\,d\sigma \propto \left[1 + \frac{t^2}{(n-1)}\right]^{-(n-1)/2},$$

with $t^2 = n(\overline{x} - m)^2/s^2$, $s^2 = \sum_{1=1}^{n}(x_i - \overline{x})^2/(n - 1)$, which is Student's t with $n - 1$ degrees of freedom. Finally, integrating our Student's t results in unity, i.e.

$$\int_{-\infty}^{\infty}\int_0^{\infty} \pi(m, \sigma^2|\mathbf{x})\,d\sigma\,dm = 1.$$

Jeffreys has also suggested an algorithm for constructing informationless priors on the basis of the Fisher information. If $X_1, X_2, \ldots, X_n$ are iid with likelihood function $L(\sigma|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$, then the Fisher **information** in the sample is

$$I_n(\theta) = E\left(\frac{\partial \log L(\theta|x)}{\partial\theta}\right)^2 = nE\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right)^2 = nI(\theta).$$

Jeffreys' algorithm for an informationless prior is $p(\theta) \propto I(\theta)^{1/2}$. Since the Fisher information is invariant under transformations, the Jeffreys algorithm is also invariant with respect to transformations. The Jeffreys algorithm for many parameters is to take $p(\theta) \propto |I(\theta)|^{1/2}$, where $|I(\theta)|$ is the determinant of the matrix of partial cross derivatives.

An important class of prior distributions is when the posterior distribution belongs to the same family of distributions as the prior distributions. These are called *conjugate* or *normal conjugate* prior distributions. For example, consider the distribution of the sample mean arising from n iid observations arising from a $N(m, \sigma^2)$ with σ^2 known. The likelihood is $L(m|x) = \exp[-n(\overline{x} - m)^2/2\sigma^2]$. The conjugate prior distribution is $p(m) \propto \exp[-n'(m - m')^2/2\sigma^2]$, where (n', m') are parameters of the prior distribution. Then the posterior distribution of m is

$$\pi(m|x) \propto L(m|x)p(m) \propto \exp\left[\frac{-n''(m - m'')}{2\sigma^2}\right],$$

where $n'' = n + n'$ and $m'' = (n\overline{x} + n'm')/n''$. Although the quantity n in the likelihood is an integer, the parameter n' is not restricted to be an integer but only to be nonnegative. Note that as $n' \to 0$,

$p(m) \propto 1$, which results in an improper prior in the limit. The form of the posterior distribution shows that the prior distribution contributed n' observations having a mean of m' to the likelihood. The informationless prior corresponding to $n' \to 0$ contributes no information to the likelihood.

There exist conjugate prior distributions for all of the distributions having minimal sufficient statistics. The book by Raiffa & Schlaifer [29] contains a compendium and an extensive discussion of conjugate prior distributions. In all of these conjugate prior distributions it is possible to interpret the parameters of the prior as adding additional information to the data.

To cite another example, in addition to the normal distribution, consider the likelihood arising from observing s successes out of n trials from a binomial distribution. The likelihood is $L(\theta|s) = \theta^s(1-\theta)^{n-s}$ and the conjugate prior is $p(\theta) \propto \theta^{s'}(1-\theta)^{n'-s'} (0 \leq s' \leq n')$ resulting in the beta posterior distribution. $\pi(\theta|s) \propto \theta^{s''}(1-\theta)^{n''-s''}$, with $s'' = s + s'$ and $n'' = n + n'$. The prior distribution has contributed s' successes from n' trials. Jeffreys [10] proposed that the informationless prior for this situation take $s' = 1/2$ and $n' = 1$, i.e. $p(\theta) \propto [\theta(1-\theta)]^{-1/2}$. The prior contributes a "half" a success from a single trial to the likelihood.

In any event, there is a determination of the information for every conjugate prior distribution. As a result, if the prior is based on past information, than it can lead to the fitting of the parameters of the conjugate prior distributions. The same remark holds for choosing priors by a subjective assessment. Any subjective assessment that is incorporated into a conjugate prior can be interpreted with regard to its information content.

Over the past several decades much of the debate concerning Bayesian inference has been concentrated on the use of prior distributions. It is unlikely that there will be closure on this topic. However, when the prior distribution can be interpreted with respect to information content, priors having information equivalent or less than one unit of information and which are smooth over the parameter space are unlikely to have a major effect on the likelihood function.

References

[1] Barnard, G.A. (1947). The meaning of significance level, *Biometrika* **34**, 179–182.

[2] Barnard, G.A. (1949). Statistical inference (with discussion), *Journal of the Royal Statistical Society, Series B* **11**, 115–139.

[3] Barnard, G.A. (1967). The use of the likelihood function in statistical inference, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley.

[4] Barnard, G.A. (1974). On likelihood, in *Proceedings of the Conference on Foundational Questions in Statistical Inference*, O. Barndorff-Nielsen, P. Blaesild, & G. Schou, eds. Department of Theoretical Statistics, University of Aarhus.

[5] Berger, J.O. & Wolpert, R.L. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward.

[6] Birnbaum, A. (1962). On the foundations of statistical inference, *Journal of the American Statistical Association* **57**, 269–306.

[7] Birnbaum, A. (1972). More on concepts of statistical evidence, *Journal of the American Statistical Association* **67**, 858–861.

[8] Breslow, N. (1990). Biostatistics and Bayes, *Statistical Science* **5**, 269–298.

[9] Cornfield, J. (1969). The Bayesian outlook and its application (with discussion), *Biometrics* **25**, 617–657.

[10] Cox, D.R. (1958). Some problems connected with statistical inference, *Annals of Mathematical Statistics* **29**, 357–371.

[11] Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, Cambridge.

[12] Fisher, R.A. (1921). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

[13] Fisher, R.A. (1970). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[14] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

[15] Fisher, R.A. (1934). Two new properties of mathematical likelihood, *Proceedings of the Royal Society of London, Series A* **144**, 285–307.

[16] Fisher, R.A. (1935). The fiducial argument in statistical inference, *Annals of Eugenics* **6**, 391–398.

[17] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Oliver & Boyd, Edinburgh.

[18] Fisher, R.A. (1966). *The Design of Experiments*, 8th Ed. Oliver & Boyd, Edinburgh.

[19] Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall, New York.

[20] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Oxford University Press, Oxford.

[21] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[22] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd Ed. Wiley, New York.

[23] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London, Series A* **236**, 333–380.

[24] Neyman, J. & Pearson, E.S. (1933). On the testing of statistical hypotheses in relation to probabilities a priori, *Proceedings of the Cambridge Philosophical Society* **29**, 492–510.

[25] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London, Series A* **231**, 289–337.

[26] Neyman, J. & Pearson, E.S. (1936). Sufficient statistics and uniformly most powerful tests of statistical hypotheses, *Statistical Research Memoirs* **1**, 113–137.

[27] Neyman, J. & Pearson, E.S. (1936). Unbiased critical regions of Type A and Type A_1, *Statistical Research Memoirs* **1**, 1–37.

[28] Neyman, J. & Pearson, E.S. (1938). Contributions to the theory of testing statistical hypotheses, *Statistical Research Memoirs* **2**, 25–57.

[29] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Graduate School of Business Administration, Harvard University, Cambridge, Mass.

[30] Savage, L.J. (1954). *The Foundations of Statistics*. Methuen, London.

[31] Smith, A.F.M. (1991). Bayesian computational methods, *Philosophical Transactions of the Royal Society of London, Series A* **337**, 369–386.

[32] Smith, A.F.M. & Gelfand, A.F. (1992). Bayesian statistics without tears: a sampling-resampling perspective, *American Statistician* **46**, 84–88.

[33] Smith, A.F.M. & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, Series B* **55**, 3–23.

[34] Tanner, M. & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association* **82**, 528–550.

[35] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics* **22**, 1701–1762.

(*See also* **Inference, Foundations of**)

M. ZELEN

Influence Function in Survival Analysis

The influence function of an estimator was introduced by Hampel [5] in the context of robust estimation (*see* **Robustness**). Broadly speaking, the influence function evaluated at a possible data point x indicates how the estimator is changed by the addition of a data point with value x. As an example, the influence function for the sample mean is identically equal to x, showing that a single data point has an influence on the mean directly proportional to its value. This reflects the fact that the sample mean is very sensitive to **outliers**. However, the influence function for the sample **median** is a step function. The median is the simplest example of an estimator with bounded influence function. Several more **efficient** estimators with bounded influence functions have been proposed as more robust estimators [5]. An introduction to the influence function is given in [9].

The influence function of an estimator is computed by first writing the estimator as a functional of a distribution function. For an estimator that is a function of independent, identically distributed observations, this distribution function will be the empirical distribution function, $F_n(x)$ (*see* **Goodness of Fit**). For example, the sample mean $\overline{X} = n^{-1} \sum X_i$ can be expressed as $\int x \, \mathrm{d}F_n(x)$. This estimates the same functional of the true distribution function $\int x \, \mathrm{d}F(x)$: estimators with this property are called Fisher **consistent**. We can write the sample median as $F_n^{-1}\left(\frac{1}{2}\right)$, (with a suitable definition of inverse for a noncontinuous function), and this estimates $F^{-1}\left(\frac{1}{2}\right)$.

We use the general notation $T(F)$ for a functional of a distribution function. Then we define the influence function for $T(F)$ by

$$IC(x; T, F) = \lim_{t \to 0} t^{-1}\{T[(1-t)F + t\delta_x] - T(F)\}, \tag{1}$$

if this limit exists. In (1) δ_x is the distribution function that puts mass 1 at the point x. For $T(F) = \int x \, \mathrm{d}F(x)$, we have $IC(x; T, F) = x$, and for $T(F) = F^{-1}\left(\frac{1}{2}\right)$, we have

$$IC(x; T, F) = \begin{cases} -\dfrac{1}{2f[F^{-1}(1/2)]}, & \text{if } x < F^{-1}\left(\frac{1}{2}\right), \\ +\dfrac{1}{2f[F^{-1}(1/2)]}, & \text{if } x > F^{-1}\left(\frac{1}{2}\right). \end{cases}$$

The influence function defined in (1) is constructed from the so-called Gâteaux derivative of the functional T. The definition can be extended by computing the Gâteaux derivative of more complex functionals, such as functionals $T(F, u)$ that depend on an additional real parameter, or bivariate functionals $T(F, G)$, say.

In survival data analysis, the estimators of interest are typically more complex than simple functions of independent and identically distributed observations, and the definition of the influence function needs these more complex functionals. Consider the case of a single sample of independent, possibly **censored**, observations $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, where X_i is the observed failure or censoring time, and δ_i is 1 if X_i is uncensored, and 0, otherwise. Assuming the random censorship model, we write $X = \min(X^0, Y)$, where X^0 has distribution $F(\cdot)$, the failure time distribution of interest, and Y has distribution G.

The influence function of the **Kaplan–Meier** estimate of the survival distribution F was introduced in [8, Eqs. (2.1), (2.2)]. This used a representation, due to Peterson [6], of the cumulative hazard function as a functional of two subsurvival functions: $S_u(\cdot)$, and $S_c(\cdot)$, where $S_u(t) = \Pr(X > t, \delta = 1)$, and $S_c(t) = \Pr(X > t, \delta = 0)$. This gives a pair of influence functions for the Kaplan–Meier estimator of $S(t) = 1 - F(t)$:

$$IC_1(s; T, S_u, S_c)(t) = S(t) \left\{ \int_0^{\min(s,t)} \frac{\mathrm{d}S_u(x)}{(S_u + S_c)^2(x)} + \frac{1(s \le t)}{(S_u + S_c)(s)} \right\},$$

$$IC_2(s; T, S_u, S_c)(t) = S(t) \left\{ \int_0^{\min(s,t)} \frac{\mathrm{d}S_u(x)}{(S_u + S_c)^2(x)} \right\}.$$

The first term in IC_1 is the effect a new observation at time s has on the estimate of $S(t)$ by increasing the size of the risk set, if $s \le t$. This is the only effect of a new censored observation, as is seen from the expression for IC_2. The second contribution to IC_1 corresponds to the additional jump point in the Kaplan–Meier estimate of $S(t)$ when a new, uncensored, observation at time s is added.

In addition to providing a descriptive summary of how sensitive an estimator can be to outliers, the influence function can be used to compute the asymptotic variance of an estimator. We assume for notational convenience that we have a simple functional of one distribution function $T(F)$. If this functional

is differentiable, then we can write

$$T(G) = T(F) + \mathrm{d}T_F(G - F) + R, \qquad (2)$$

where G is some distribution function, $\mathrm{d}T_F$ is the differential of $T(F)$ and is a linear functional, and R is a remainder term. For many statistical functionals, $\mathrm{d}T_F(G - F)$ will take the form

$$\mathrm{d}T_F(G - F) = \int IC(x; T, F)\,\mathrm{d}G(x),$$

where IC is defined in (1), but has been standardized if necessary so that $\int IC(x)\,\mathrm{d}F(x) = 0$. If we now let $G = F_n$, then we have an expression for the estimator $T(F_n)$ as the true value $T(F)$, plus a linear combination $n^{-1}\sum IC(X_i; T, F)$ and a random remainder term. Under some conditions that, in particular, ensure that the remainder goes to 0 in a suitable sense as $n \to \infty$, we may conclude that $\sqrt{n}[T(F_n) - T(F)]$ is asymptotically normally distributed with mean 0 and variance $\int IC^2(x; T, F)\,\mathrm{d}F(x)$.

The argument sketched above is an example of the *functional delta method*, described in [1, II.8], and [4]. The ordinary **delta method** uses an approximate linearization of a nonlinear function to find the limiting distribution of an estimator. The functional delta method uses the same argument with the functional derivative $\mathrm{d}T$ defined by (2). In fact, the functional derivative is not well defined by (2): we need to specify in what sense R converges to 0, as G becomes arbitrarily close to F. There are three main notions of convergence, leading to Gâteaux, compact, and Fréchet differentiability. While the definition of the influence function in (1) uses the weak notion of Gâteaux differentiability, the asymptotic argument requires that T be either compact or Fréchet differentiable. For a fuller discussion of this, see [1, II.8]. Since functional derivatives also obey a chain rule, the functional delta method can be used to find the asymptotic distribution for functions of estimators as well. The asymptotic variance of the cumulative hazard estimator was computed using this method in [8]. The functional delta method is applied to the Kaplan–Meier estimator and several functions of it in [1, IV.3] and to more complex product limit estimators in [1, IV.4].

Another important use of the influence function in survival analysis is to suggest influence **diagnostics**, or case-deletion diagnostics, for use in the **proportional hazards** regression model, analogously to the way regression diagnostics are computed routinely for linear regression models. It is shown in [3] and [10] that a sample estimate of the influence function can be used to approximate $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i}$, where $\hat{\boldsymbol{\beta}}$ is the usual estimate of the regression parameter in Cox's proportional hazards regression model and $\hat{\boldsymbol{\beta}}_{-i}$ is the estimate obtained when the ith observation is deleted. Storer & Crowley [11] consider various other ways to estimate $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_i$. Barlow & Prentice [2], in a discussion of various definitions of **residuals** for proportional hazards regression, relate the estimated influence function to a particular type of residual, and thereby also extend the definitions of [3] and [10] to **time-dependent covariates**. Further development, with emphasis on applications, is given in [7]. There is also a helpful summary in [1, VII.3].

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Barlow, W.E. & Prentice, R.L. (1988). Residuals for relative risk regression, *Biometrika* **75**, 65–74.

[3] Cain, K.C. & Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data, *Biometrics* **40**, 493–500.

[4] Gill, R.D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1), *Scandinavian Journal of Statistics* **16**, 97–128.

[5] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**, 383–394.

[6] Peterson, A.V. (1977). Expressing the Kaplan–Meier estimator as a function of empirical sub-survival functions, *Journal of the American Statistical Association* **72**, 854–858.

[7] Pettitt, A.N. & Bin Daud, I. (1989). Case-weighted measures of influence for proportional hazards regression, *Applied Statistics* **38**, 51–68.

[8] Reid, N. (1981). Influence functions for censored data, *Annals of Statistics* **9**, 78–92.

[9] Reid, N. (1983). Influence functions, in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 117–120.

[10] Reid, N. & Crépeau, H. (1985). Influence functions for proportional hazards regression, *Biometrika* **72**, 1–9.

[11] Storer, B.E. & Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods, *Journal of the American Statistical Association* **80**, 139–147.

(*See also* **Residuals for Survival Analysis**; **Survival Distributions and Their Characteristics**)

N. REID

Informatics in the Health Sciences

Information technology, in its widest sense, increasingly serves multiple roles in professional activities and practice. Besides being the servant that enables a number of activities, it also provides the environment to model and study those activities and domains of interest. Informatics is the use of information technology, broadly conceived, to advance a domain of work or inquiry. *Medical* informatics is the application of informatics principles and practice to clinical care; *nursing* informatics, to nursing care; *public health* informatics, to issues of public health; **bioinformatics** to biology and to the work of biologists; *statistics* informatics, to epistemology and the practice of statistics.

The word was first used in A.I. Mikhailov's (Scientific Department of the Moscow State University) book *Oznovi Informatiki* (Foundations of Informatics); it was then turned into the French word *informatique* (de medecine), or medical computing. The term was then brought into English by MF Collen in 1977 [2].

Friedman [4] has proposed a scaffold of four levels of activity within informatics: model formulation, system development, system installation, and study of effects. A complementary model is shown here, where *System* refers to a model of the target domain, *Role* speaks of the role an individual plays within that domain (in health sciences/care, the roles are clinical care, learning, teaching, research, and administration), *Functions* refers to those activities that support a particular role, and *Workflow* represents the model of the activities that support each activity. The next layer, of information resources, divides into two components: *Information Repositories* refers to the storage facilities needed to accomplish a step in the workflow and *Information Tools* refers to the software tools (*see* **Software, Biostatistical**) or **algorithms** by which an information repository is used to support that step. *Standards* refer to common formats and protocols that support the information resources to promote interoperability, and *Technology* refers to hardware and software products and protocols that support the whole enterprise (*see* **Computer Architecture and Organization**).

<table>
<tr><td colspan="2">System</td></tr>
<tr><td colspan="2">Role</td></tr>
<tr><td colspan="2">Functions</td></tr>
<tr><td colspan="2">Workflow</td></tr>
<tr><td>Information repositories</td><td>Information tools</td></tr>
<tr><td colspan="2">Standards</td></tr>
<tr><td colspan="2">Technology</td></tr>
</table>

In medical informatics, *system* refers to the clinical environment. Physicians have specific *roles* (clinician, educator, learner, manager, researcher), and tasks within those roles, in that environment, and different workflows apply to each function. For instance, in clinical care, two functions are *diagnose* and *treat*. A diagnostic decision support system helps the clinician to diagnose. In doing so, it embodies the workflow of diagnosis (the hypothetico-deductive loop of hypothesizing a number of conditions, suggesting data to collect, collecting the data, and modifying the differential diagnosis) (*see* **Computer-aided Diagnosis**). To accomplish that support, it needs an information repository of a knowledge base and an algorithm of an inference engine. (Inference engines have generally been based on logical inference, **neural networks**, or Bayesian probability networks.) For decision support systems to be interoperable, standards, such as Arden syntax, have been designed and promulgated. The decision support system can be implemented in a range of technologies, ranging from mainframe to handheld devices.

For the core clinical functions of *retrieve clinical data*, *manage*, *transact*, and *document*, the computer-based patient record and computer-based provider order entry are key repositories. *Decision support* is often sited in the latter. For the clinical function of *communicate*, the technologies of telemedicine have received attention, especially in settings with difficult access to clinicians (prisons, rural areas, shut-ins).

A range of communication, vocabulary, and technology standards are increasingly viewed as vital to integrating American healthcare into a National Health Information Infrastructure (NHII). An advisory office in the Department for Health and Human Services has been established to foster the NHII. The Consolidated Health Initiative of the eGov effort, working with the National Committee on Vital and Health Statistics, has advised government adoption of the following standards: Health Level 7® (HL7®, communication), National Council on Prescription

Drug Programs (NCDCP, standards for ordering drugs from retail pharmacies) , Institute of Electrical and Electronics Engineers 1073 (IEEE1073, "Medical Informatics Bus", for physiological systems communication), Laboratory Logical Observation Identifier name Codes® (LOINC®, laboratory values), and Digital Imaging Communications in Medicine® (DICOM®, image transmission and storage).

In all cases, efforts have been taken to understand the optimal methods of implementing and deploying these systems, as well as evaluating the systems' impact, either at the level of return on investment or at the level of impact on patient outcomes and on the processes of care. These issues have been a particular focus of nursing informatics.

In their review of the history of definitions of nursing informatics, Staggers and Thompson [6] provide the following synthesis:

> Nursing informatics is a specialty that integrates nursing science, computer science, and information science to manage and communicate data, information, and knowledge in nursing practice. Nursing informatics facilitates the integration of data, information, and knowledge to support patients, nurses, and other providers in their decision making in all roles and settings. This support is accomplished through the use of information structures, information processes, and information technology.

In supporting the role of clinical care, nursing informatics focuses on training nurse users and in specifying and designing systems to support the work of nurses. Nursing informatics tends to focus more on supporting the administrative role than does medical informatics, with attention to the functions of quality improvement, performance improvement, outcome measurement, process redesign, and disease management.

Public health informatics is a more recently articulated discipline [3, 7]. With populations as its focus, its primary roles are prevention (*see* **Preventive Medicine**), **surveillance**, **risk assessment**, and research. Recent activity has focused on recognizing the range of functions that can be supported by existing or novel information technology; enabling communication between information repositories or integration among them; the creation of novel algorithms to perform, for instance, syndromic surveillance, with these new sources of information, and enabling the dissemination of tools for using those repositories in an inexpensive manner; the modification of existing, or creation of new standards; and the creation of novel technologies, such as environmental sensors, also to support these roles.

There are disciplines of informatics to support particular subdomains of medicine, like radiology, pathology, **dermatology**, **psychiatry**, and primary care.

On the research side, **bioinformatics** tends to divide into two areas. One is the view of biology as a set of digital processes; this view leads to computational biology and the like. Second is the view of research as an activity performed by people, which leads to research informatics and the creation of **databases** and other repositories and tools to support the work of research. The importance of information technology to support clinical research has led to solutions at each level of the hierarchy: repositories of research instruments and of participant data; statistical packages for clinical research; the use of clinical standards in coding research results; and data collection tools, like those based on personal digital assistants.

Statistics also had a dual relationship with informatics. On the one hand, statistical algorithms are used in all domains, when the view of data goes across more than one **unit of analysis**, whether in the description of a population of patients, or the analysis of a set of natural-language text. On the other hand, relatively little support is available for the *work* of statistics, outside the statistical packages that ease the calculations. The statistical *workflow* involves the creation of data sources and communication among them, with an eye toward the analytic and decision goals of the data; involves the assembly of the data, with an understanding of the errors and limitations conferred by the different data sources; involves "cleaning" of the assembled data, taking the data sources and the analytic goals into account; involves the choices and sequencing of the statistical models and processes for analysis [1, 5]; and involves the reporting of the data, with the caveats and limitations of the inferences made explicit. There are few standards to support this workflow that take into account the needs of the statistician, outside the growing library of programming modules. The knowledge and experience of professionals that have been encoded in guidelines and protocols in other domains is lacking in statistics, and there are no statistical decision support tools to support this workflow.

Attention to the workflow in the spirit of informatics more generally should fill this void.

References

[1] St Amant, R. & Cohen, P. (1998). Intelligent support for exploratory data analysis, *Journal of Computational and Graphical Statistics* **7**(4), 545–558.

[2] Collen, M.F. (1995). *A History of Medical Informatics in the United States*. American Medical Informatics Association, Bethesda.

[3] Friede, A., McDonal, M. & Blum, H. (1995). Public health informatics: how information-age technology can strengthen public health, *Annual Review of Public Health* **16**, 239–252.

[4] Friedman, C.P. (1994). The research we should be doing, *Academic Medicine* **69**(6), 455–457.

[5] Oldford, R.W. & Peters, S.C. (1986). Implementation and study of statistical strategy, in *Artificial Intelligence and Statistics*, W.A. Gale, ed. Addison-Wesley, Reading, pp. 335–353.

[6] Staggers, N. & Thompson, C.B. (2002). The evolution of definitions for nursing informatics: a critical analysis and revised definition, *Journal of the American Medical Informatics Association* **9**, 255–261.

[7] Yasnoff, W., Overhage, J., Humphrey, B. & LaVenture, M. (2001). A national agenda for public health informatics, *Journal of the American Medical Informatics Association* **8**(6), 535–545.

HAROLD P. LEHMANN

Information Matrix

Let $\mathbf{X}$ be an n-vector of observations relating to a vector parameter $\boldsymbol{\theta}$. In addition, let $f(\mathbf{X}|\boldsymbol{\theta})$ denote the joint density (or mass function) of $\mathbf{X}$ under $\boldsymbol{\theta}$, which, viewed as a function of $\boldsymbol{\theta}$, is the **likelihood** function. A key role in statistics is played by the matrices $\mathbf{I}^{(\mathrm{o})}$ and $\mathbf{I}$ defined as follows, where the expectation below is taken under $\boldsymbol{\theta}$:

$$\mathbf{I}^{(\mathrm{o})}_{pq}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_p \partial\theta_q} \log f(\mathbf{X}|\boldsymbol{\theta}),$$

$$\mathbf{I}_{pq}(\boldsymbol{\theta}) = \mathrm{E}\left[\mathbf{I}^{(\mathrm{o})}_{pq}(\boldsymbol{\theta})\right].$$

The matrix $\mathbf{I}(\boldsymbol{\theta})$ is called the Fisher information matrix, or sometimes, for emphasis, the expected information matrix. For a scalar parameter θ, the term used is simply Fisher information (*see* **Information**). The matrix $\mathbf{I}^{(\mathrm{o})}(\boldsymbol{\theta})$ is called the observed information matrix. Generally, the elements of the matrices $\mathbf{I}(\boldsymbol{\theta})$ and $\mathbf{I}^{(\mathrm{o})}(\boldsymbol{\theta})$ will have magnitude of order n.

The basic classic results concerning the information matrix are stated below. The most familiar setting for these results is that of independent, identically distributed (iid) observations, but they extend to more general settings. For the results to hold, certain technical conditions are required (see, for example, [4, Chapter 5]); such conditions generally hold in typical applied statistics settings. Below we denote the inverse of the information matrix $\mathbf{I}(\boldsymbol{\theta})$ by $\mathbf{V}^*$.

Cramér–Rao Inequality

Let $\tilde{\boldsymbol{\theta}}$ be any unbiased estimate of $\boldsymbol{\theta}$, and denote by $\mathbf{V}$ its **covariance matrix**. Then the matrix $\mathbf{V} - \mathbf{V}^*$ is nonnegative definite. In particular, in the case of a scalar parameter θ, this result says that the variance of any unbiased estimator of θ cannot be less than V^*. There also exists an asymptotic version of this result in which unbiasedness is replaced by asymptotic unbiasedness and covariance is replaced by asymptotic covariance.

Behavior of MLEs

The **maximum likelihood** estimate (MLE) $\hat{\boldsymbol{\theta}}$ of the true $\boldsymbol{\theta}$ is defined as the maximizer for given data $\mathbf{X}$ of the function $f(\mathbf{X}|\boldsymbol{\theta})$ over all $\boldsymbol{\theta}$. The estimator $\hat{\boldsymbol{\theta}}$ is approximately distributed according to the **multivariate normal distribution** with mean vector $\boldsymbol{\theta}$ and covariance matrix $\mathbf{V}^*$. In light of what was said above, this indicates an asymptotic optimality property of the MLEs.

Estimation of $\mathbf{V}^*$

The matrix $\mathbf{I}^{(\mathrm{o})}(\hat{\boldsymbol{\theta}})$ is a **consistent** estimator of $\mathbf{I}(\boldsymbol{\theta})$ in the sense that $\mathbf{I}(\boldsymbol{\theta})^{-1}\mathbf{I}^{(\mathrm{o})}(\hat{\boldsymbol{\theta}})$ **converges** to the identity matrix in large samples. Correspondingly, $\mathbf{I}^{(\mathrm{o})}(\boldsymbol{\theta})^{-1}$ consistently estimates $\mathbf{V}^*$.

The above results extend to more general forms of likelihood often encountered in biostatistics, including **quasi-likelihood** [3, Chapter 9] and partial likelihood [2]. Asymptotic optimality properties of MLEs in these more general settings also have been developed; see, respectively, [3, Section 9.5] and [1].

References

[1] Bickel, P.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Inference in Semiparametric Models* John Hopkins University Press, Baltimore.

[2] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[3] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[4] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. Wiley, New York.

D.M. ZUCKER

Information

It is natural to try to quantify the amount of information provided by a set of data concerning an unknown quantity of interest. R.A. Fisher, in a classic 1925 work [1], proposed that the statistical information provided by a set of data on a parameter θ be defined as the inverse of the variance of an **efficient** estimator of θ. This quantity is equal to the Fisher information $I(\theta)$ defined in the article, **Information Matrix**. More broadly, one may define the information provided by a given estimator $\tilde{\theta}$, not necessarily an efficient one, to be the inverse of its variance. Intuitively, the lower the variance of an estimator, the more precisely it estimates the underlying parameter. Thus, the definition just given says simply that information equals precision.

The above concept of information arises in various contexts; here, we give two examples of interest in biostatistics.

The first example is the situation of combining several independent (asymptotically) **unbiased** estimates of the same parameter θ. This situation arises in a number of settings in biostatistics; for example, **stratified** analysis and **meta-analysis**. Typically, estimates are combined by weighted averaging. It is a classical result that the optimal method of weighting, in terms of minimizing the variance of the combined estimator, is to weight each individual estimate, $\tilde{\theta}_i$, according to the inverse of its variance, v_i. This result, which is an easy consequence of the Cauchy–Schwarz inequality, is the most basic version of the weighted **least squares** schemes that abound in applied statistics.

Intuitively, the weight assigned to a given estimate is in proportion to the amount of information contributed by that estimate. The variance of the combined estimator is easily derived, and by reciprocation the information content of the combined estimate is found to be the sum of v_i^{-1}. Thus, it is seen that the information content of an optimally weighted average of several independent estimates is equal to the sum of the information contents of the individual estimates.

The second example is the situation of sequential monitoring in clinical trials (*see* **Data and Safety Monitoring**). With certain popular monitoring schemes, particularly that of Lan & DeMets [3], it is necessary at each interim analysis to have some measure of how far the trial has progressed. A simple measure of trial progress is elapsed calendar time from the date the trial began. Some workers, however, argue that a more appropriate measure of trial progress is the proportion of statistical information accumulated by the time of the interim analysis, relative to the total amount of information that is expected to be accumulated by the planned end of the trial.

This measure of trial progress is often referred to as *information time.* The proportion of information accumulated is reflected fully by the sample size in certain special cases, but not in general. For instance, in a study monitored sequentially using the **logrank test**, the information is reflected by the number of events, while in a longitudinal study analyzed using a mixed linear model with a random slope and intercept for each subject, the information content is a function of the ratio of within-subject to between-subject variance and the observation pattern of the various individuals. See [4] for a more detailed discussion.

The foregoing conceptualizations of information are related, as will be indicated below, to S. Kullback's notion of discrimination information. Kullback's information measure, also known as relative entropy and by various other names, is in turn related to the entropy-based definition of information used in the Shannon–Weaver theory of information and coding. Consider a data vector $\mathbf{X}$ relating to the parameter θ. Let $f(\mathbf{X}|\theta)$ denote the joint density (or mass function) of $\mathbf{X}$ under θ, which, viewed as a function of θ, is the **likelihood** function. In addition, define

$$Z(\theta_1 : \theta_0) = \log \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)},$$

which is the **likelihood ratio test** statistic for testing $\mathrm{H}_0 : \theta = \theta_0$ vs. $\mathrm{H}_1 : \theta = \theta_1$. Kullback defines the mean information for discriminating between these two hypotheses when the true θ value is θ_1 to be

$$\begin{aligned}\mathrm{Inf}(\theta_1 : \theta_0) &= \mathrm{E}_{\theta_1}[Z(\theta_1 : \theta_0)] \\ &= \int f(\mathbf{X}|\theta_1) \log \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \, \mathrm{d}\mathbf{X},\end{aligned}$$

where the integral is replaced by a sum when $\mathbf{X}$ is discrete. Kullback demonstrates that this information measure possesses a number of basic properties that one would intuitively expect from an information measure. For example, the information provided by two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is equal to the sum

of the information provided by **X** and the information associated with the conditional distribution of **Y** given **X**. For more detailed discussion, see [2].

Kullback shows, furthermore, by a Taylor expansion of the likelihood ratio, that if θ_0 and θ_1 are close and suitable regularity conditions hold, then

$$\text{Inf}(\theta_1 : \theta_0) \doteq \tfrac{1}{2} I(\theta_1)(\theta_1 - \theta_0)^2;$$

in the case of a vector parameter, the right-hand side becomes $\frac{1}{2}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{I}(\boldsymbol{\theta})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)$, where $\mathbf{I}(\boldsymbol{\theta})$ here is the Fisher information matrix. Thus, there is a direct connection between Kullback's notion of information and Fisher's.

If the data are reduced down to some (not necessarily efficient) estimator $\tilde{\theta}$ of θ that is approximately normally distributed with mean θ, then the Fisher information of the reduced data is approximately equal to the inverse of $\text{var}_\theta(\tilde{\theta})$. This observation provides a further way of viewing inverse variance as information.

In particular, considering the **maximum likelihood** estimator (MLE) $\hat{\theta}$ of θ, assuming suitable conditions, it is known that when the sample size is large, the estimator $\hat{\theta}$ is approximately normal. Thus, if the data were reduced down to the MLE $\hat{\theta}$, then the Fisher information for the reduced data would be approximately equal to the inverse of $\text{var}_\theta(\hat{\theta})$. Now this quantity is equal precisely to the Fisher information for the unreduced data. In other words, the information contained in the MLE is approximately equal to the information in the entire data, reflecting the efficiency of the MLE.

This finding may be arrived at also by a route starting from the Kullback definition of information. Suppose that θ_0 and θ_1 are close and the sample size is large. Then, assuming suitable conditions, it is known that the likelihood ratio statistic $Z(\theta_1 : \theta_0)$ is approximately equivalent to the Wald statistic $Z_{\text{Wald}} = (\hat{\theta} - \theta_0)/\text{var}_\theta(\hat{\theta})^{1/2}$ [in practice $\text{var}_\theta(\hat{\theta})$ has to be estimated on the basis of $\hat{\theta}$] (*see* **Likelihood**). This statistic depends on the data only through $\hat{\theta}$, so that we see again that the information in $\hat{\theta}$ alone is approximately equal to the information in the entire data.

References

[1] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

[2] Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York (Dover, New York, 1968; Peter Smith Publisher, Magnolia, Mass., 1978).

[3] Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.

[4] Lan, K.K.G. & Zucker, D.M. (1993). Sequential monitoring of clinical trials: the role of information and Brownian motion, *Statistics in Medicine* **12**, 753–765.

(*See also* **Large-sample Theory**)

D.M. ZUCKER

Instrumental Variables in Health Services Research

The technique of instrumental variables (IV) was developed by econometricians in the 1930s and 1940s to address situations in which an **explanatory variable** or variables are correlated with the error term in an Ordinary **Least Squares regression**. In such a case, Ordinary Least Squares (OLS) methods produce inconsistent estimates of the regression coefficients. Because economists can rarely conduct controlled experiments, this situation arises frequently in analyses of economic data. As a result, IV methods have been widely used – many might say overused – in applied econometrics and are described in virtually every econometrics text (e.g. [5, Chapter 13], [6, Chapters 9, 13], [2, Chapter 7], [4, Chapters 9, 13, 16], [10, Chapter 10]). In the 1990s, IV methods began to be used in **health services research** to analyze observational data, where they have achieved a modicum of popularity [3, 7, 8].

I first give an intuitive explanation of the method and then sketch it more formally. I next describe an actual example and conclude by emphasizing the limitations of the IV estimator.

The essence of the IV method is to purge the explanatory variable(s) in a regression equation of the portion of its (their) **variance** that is not independent of the error term and then estimate the relationship between the dependent variable and the remaining variance. (In econometrics, the variation that is independent of the error term is referred to as exogenous, whereas the variation that is not independent is referred to as endogenous.) (*See* **Structural Equation Models**.) The purging is done by finding another variable or variables that are termed instrumental variables and that satisfy two assumptions. First, the IVs are independent of the error term in the regression of interest. Second, they are correlated with the explanatory variable(s) in question. Another way to say this is that the IVs have a direct effect only on the explanatory variable; they have no direct effect on the dependent variable. This is illustrated in Figure 1. The arrows indicate causal relationships; for a variable to be an IV, there must be no arrow from the IV directly to the dependent variable. There must be at least one IV for each explanatory variable that is not independent of the error term.

The effects on the dependent variable of the variation that the IV induces in an explanatory variable of interest can be estimated; that is, the covariation between the induced variation and the dependent variable can be used to estimate a regression coefficient. The results, however, should not be extrapolated outside the region of induced variation. Although such a caution is always appropriate in regression analysis, it applies with even more force here because the range of variation may be sharply reduced relative to the total variation in the observed explanatory variable (i.e. the sum of the exogenous and endogenous variations).

Although IV methods were developed for nonexperimental data, the simplest example of an IV comes from the **randomization** process in a **clinical trial**. Suppose subjects are assigned to the treatment or the control group by flipping a fair coin. As is well known, assuming no refusal and no attrition, the difference in the mean outcomes between the experimental and control groups is an **unbiased** estimate of the treatment effect.

In this case, the IV is the outcome of the coin flip. By definition, the outcome of the flip has no effect on the observed clinical outcome in any patient other than through its effect on assignment to the treatment group. Hence, the variation that the coin flip induces in assigning subjects to treatment and control groups is independent of the error term in a regression explaining outcomes. Moreover, given no refusal or attrition, it perfectly explains the assignment of subjects to treatment or control groups. Thus, it satisfies both assumptions of an IV. By contrast, in observational data, those assigned a particular medical treatment may be sicker or healthier in unobserved ways than those not given the treatment, which means the difference in outcomes between the two groups is a biased estimate of the effect of the treatment. Indeed, this is the reason randomized controlled trials are preferred to **observational studies** for estimating the effects of medical treatments.

More formally, suppose one has a sample of N observations that come from the following structure:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon. \tag{1}$$

Let $\mathbf{Y}$ be an $N \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ an $N \times k$ matrix of observations on k explanatory variables (the first column of X may be a vector of ones to allow for an intercept term), β a $k \times 1$ vector of constants to be

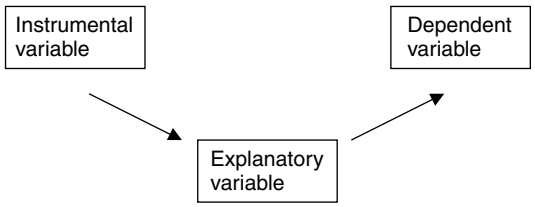

Figure 1 There is no arrow from the IV directly to the dependent variable because all of the effect of the IV on the dependent variable is through its effect on the explanatory variable. Moreover, only the variation in the explanatory variable induced by the IV is used to estimate the relationship between the dependent variable and the explanatory variable

estimated, and ε an $N \times 1$ vector of errors, which has distribution $F(0, \sigma^2)$. Let

$$b = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \text{ be the OLS estimator of } \beta. \quad (2)$$

Substituting (1) into (2), $\mathrm{E}(b) = \beta + \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\varepsilon)$. If $\mathbf{X}$ and ε are not independent, the second term is nonzero and b is a biased estimator of β.

Let X_i represent a subset k' of the k variables of $\mathbf{X}$ that are not independent of ε ($k' \leq k$), and let $\mathbf{Z}$ be an $N \times m$ matrix of observations on m variables that are correlated with X_i but are independent of ε ($k' \leq m$). Then

$$b^{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y} \text{ is the IV estimator of } \beta. \quad (3)$$

b^{IV} is a consistent estimator of β, as can be seen by substituting (1) into (3).

I turn now to an example of IV in health services research that is taken from [7]. These authors worked with a sample of Medicare patients over 65 who had suffered an Acute Myocardial Infarction (AMI or heart attack) (*see* **Medicare Data**). Motivated by the high geographic variation in procedure rates, they sought to answer the question whether cardiac catheterization followed by possible revascularization (either a coronary artery bypass graft or angioplasty, both of which open coronary arteries and provide for increased blood flow to the heart) achieved better outcomes than no cardiac catheterization. (Catheterization is always done prior to either revascularization procedure.) There were, of course, clinical trials of all these procedures, but the clinical trial results posed two difficulties in this context: (1) The trials often excluded those over 65, but around half the AMIs annually in the United States occurred among that group. Yet the benefits of the procedures had been well enough established among younger persons that it almost certainly would have been unethical to conduct a trial among the elderly; (2) The trials had been conducted in major medical centers, and it was not clear that the results applied to patients in community hospitals.

The principal IV the authors used in this example was the incremental distance from the patient's residence to a hospital with a catheterization facility relative to a hospital with no such facility. If the nearest hospital had a catheterization facility, the incremental distance was zero. This IV exploited the notion that patients who suffer an AMI tend to go to the nearest hospital and that therefore the distribution of the severity of the AMI, an unobserved determinant of both mortality and treatment, would be approximately the same at various distances. Thus, distance acted something like the flip of a coin in the clinical trial example; those living near a hospital with a catheterization facility were more likely to be treated in that fashion than patients living further away, but distance seemed independent of the severity of the heart attack, the principal factor other than treatment determining survival.

The assumption that distance is independent of mortality except through its effect on treatment must ultimately be taken on faith, although there are various tests of the plausibility of this assumption. One straightforward test is that observable variables that affect mortality should also be approximately independent of distance. Of course, the observable variables that affect mortality can be controlled for in a regression. As a result, if they are not independent of distance, that dependence will not cause the estimates to be inconsistent; nonetheless, such dependence undermines one's confidence that unobserved variables are independent of distance.

The first two columns of data in Table 1 show differences in both observable variables and outcomes for individuals who received a catheterization within 90 days of their AMI and those who did not; the last two show differences between those individuals who lived within 2.5 incremental miles of a hospital with a catheterization facility and those who did not. Both the observable covariates as well as the outcomes differ markedly between the groups that did and did not receive a catheterization (the first two columns). Mortality at four years in these

Table 1 Relationship between outcome and catheterization in groups sorted by procedure and by differential distance (%, except for age and number of observations)

	No catheterization within 90 days	Catheterization within 90 days	Incremental distance ≤ 2.5 miles	Incremental distance > 2.5 miles
Female	53.5	39.7	51.3	49.5
White	90.4	91.8	89.0	92.3
Age in years	77.4	71.6	76.1	76.1
Cancer	2.2	0.85	10.4	11.0
Pulmonary disease, uncomplicated	11.1	9.3	18.1	18.0
Diabetes	18.3	17.1	4.8	4.8
Cerebrovascular disease	5.4	2.8	45.4	5.0
Admit to cath hospital[a]	40.9	62.9	45.4	5.0
Admit to revasc hospital[a]	21.6	41.6	41.7	10.7
Admit to high-volume hosp[a]	50.0	58.0	67.1	36.5
90-day catheterization rate	0.0	100.0	26.2	19.5
One-day mortality	10.3	0.9	7.50	8.88
Seven-day mortality	22.0	3.3	16.80	18.59
30-day mortality	26.6	7.4	24.86	26.35
One-year mortality	47.1	16.6	39.79	40.54
Two-year mortality	55.3	21.3	47.20	47.89
Four-year mortality	66.7	29.9	58.06	58.52
Number of observations	158 261	46 760	102 516	102 505

[a]Catheterization hospital is one that carried out five or more catheterization procedures on patients in the sample and that is not a revascularization hospital; revascularization hospital is one that carried out ten or more revascularizations on patients in the sample; high-volume hospital is one that treated 75 or more of the AMI admissions in the sample.

two groups differs by more than a factor of two (29.9 versus 66.7%), and controlling for the observed covariates such as age does little to diminish this difference (results not shown). If those catheterized had the same distribution of unobserved factors affecting mortality as those not catheterized, one would conclude that catheterization followed by possible revascularization had a huge effect in reducing mortality.

But cardiologists and cardiac surgeons are less aggressive with those who are less healthy; for example, those who might not survive an invasive procedure are unlikely to be given one, but they are at greater risk for mortality independent of whether they receive a procedure. Hence, receipt of the procedure is correlated with the error term. The difference between the groups that did and did not receive a catheterization is strikingly illustrated by the difference in mean age; the group receiving catheterization was on average 71.6 years old, much younger than the 77.4 years average in the group that did not receive the procedure (recall that everyone in the sample was over 65 years of age). Furthermore, the group receiving the procedure had fewer **co-morbidities**, consistent with their being a generally healthier group.

The two rightmost columns divide the sample into approximately equal halves according to differential distance. Although one can exploit all the variation in an IV (in this example that means treating distance as a continuous variable), simply dividing the sample into two groups according to the value of the IV, as is done here, will show whether the observed covariates tend to be independent of the IV. In this example, the mean age in the groups sorted by incremental distance is the same to three significant digits, and the **prevalence** of co-morbidities is much more similar than in the first two columns, suggesting that distance is a reasonable choice as an IV.

From the data in these two columns, one can also derive a simple IV estimate of the effect of catheterization and possible subsequent revascularization on four-year mortality. The estimator is the difference in the mortality rates divided by the difference in the

catheterization status or:

$$\text{Estimated catheterization effect} = \frac{\Delta(\text{4-year mortality})}{\Delta(\text{catheterization})} = \frac{(58.06 - 58.52)}{(26.2 - 19.5)} = \left(\frac{-0.46}{6.7}\right) = -0.069. \quad (4)$$

In other words, an increment of one percentage point in the catheterization rate reduces the four-year mortality by about 0.07 percentage points, or about 7 in 10 000, a vastly smaller effect than would have been estimated from the first two columns, where a 100 percentage point difference in catheterization rates was associated with a 36.8 percentage point difference in mortality.

The exercise shown in the table of splitting the sample on the value of the IV is also helpful in showing the range of variation that the IV induces. In this case, just under 20% of the more distant group received a catheterization, whereas just over 26% of the group in the sample closer to the hospital received a catheterization. A finer grouping of individuals by distance somewhat expands this range from 18 to 28%.

This range illustrates an important difference between IV and a clinical trial. Ignoring refusal and attrition, the randomization in a clinical trial results in one group in which 100% of the subjects receive a treatment, the treatment or experimental group, and another group in which none of the subjects receives the treatment, the control group. The denominator in the analog to (4) thus is 100-0 or 100. The numerator is simply the difference between the outcomes in the two groups, so in this case (4) gives the usual trial result that the difference between the treatment and control group is the estimated treatment effect (dividing the percentage difference by 100 means the difference is expressed as a proportion).

By contrast, the IV result only gives the effect of increasing the catheterization rate from approximately 20% to approximately 26% (with the finer gradation of distance, this range becomes 18 to 28%). The difference between these two rates represents patients who were catheterized if they lived near a hospital with a facility but were not if they lived farther away. Another implication is that around 18 to 20% of patients were catheterized no matter where they lived (they were presumably transferred if they were initially admitted to a hospital with no catheterization facility), and around 72 to 74% of patients were not catheterized irrespective of where they lived. The data are uninformative about each of these two latter groups because there is no variation in treatment according to place of residence. It is therefore imprudent to use the estimated effect size of −0.069 to estimate what might happen to mortality if the catheterization rate fell much below 18% or rose much above 28%.

Another way to say this is that the IV estimate is informative about the person who might be called the marginal patient, namely the patient who receives a catheterization if it is convenient but not otherwise. A clinical trial, on the other hand, yields the effect of the procedure on the average patient. If the range of the IV is very large, from near zero to near 100%, the IV estimate tends toward the clinical trial estimate. As the example of the clinical trial as IV illustrates, the limiting case of the marginal effect is the average effect.

Although the IV estimator can be a highly useful tool, the two necessary assumptions required for a valid IV may be limiting. In practice, it can be difficult to find variables that are plausibly independent of the outcome of interest, except through their effect on the explanatory variable, but that induce enough variation in the explanatory variable to yield an estimate of interest. If there is little induced variation there is little power for estimating an effect, and large samples tend to be required. Less obviously, if the IV induces only a small amount of variation in the explanatory variable, the IV result will be biased toward the OLS result [1] and [9]. How much induced variation is needed to render the bias unimportant? A rough rule of thumb comes from running a regression of the explanatory variable of interest on all the covariates as well as the IV(s). In the simple case of one explanatory variable, the bias is proportional to the reciprocal of the incremental F-statistic on the IV(s) in this regression; with more than one variable correlated with the error term, the reciprocal of the F-statistic is an approximation (*see* ***F*** **Distributions**). For a derivation of this result see [9].

References

[1] Bound, J., Jaeger, D.A. & Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association* **90**(430), 443–450.

[2] Davidson, R. & MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

[3] Earle, C.C., Tsai, J.S., Gelber, R.D., Weinstein, M.C., Neumann, P.J., & Weeks, J.C. (2001). Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variables and propensity analysis, *Journal of Clinical Oncology* **19**(4), 1064–1070.

[4] Greene, W.H. (2000). *Econometric Analysis*, 4th Ed. Prentice Hall, Upper Saddle River.

[5] Judge, G.G., Griffiths, W.E., Hill, R.C. Lee, T.-C. (1980). *The Theory and Practice of Econometrics*. John Wiley & Sons, New York.

[6] Kmenta, J. (1986). *Elements of Econometrics*, 2nd Ed. Macmillan, New York.

[7] McClellan, M., McNeil, B.J. & Newhouse, J.P. (1994). Does more intensive treatment of acute myocardial infarction reduce mortality? *Journal of the American Medical Association* **272**(11), 859–866.

[8] McClellan, M.B. & Newhouse, J.P. eds. (2000). Instrumental variables analysis: applications in health services research, *Health Services Research* **35**(5, Part II), 1061–1202.

[9] Staiger, D. & Stock, J.H. (1997). Instrumental variables regression with weak instruments, *Econometrica* **65**(3), 557–586.

[10] Stock, J.H. & Watson, M.W. (2003). *Introduction to Econometrics*. Addison-Wesley, Boston.

JOSEPH P. NEWHOUSE

Instrumental Variables

Permutt & Hebel [13] describe an analysis of the results of conducting a randomized **clinical trial** [15] using what is often called an encouragement design. Pregnant women who were cigarette smokers were randomized into two groups: those who received encouragement to reduce or stop their smoking, and those who did not (the controls). Two outcome measures were recorded: (a) the number of cigarettes smoked per day in the eighth month of pregnancy (S), and (b) the birth weight of the baby (B). Permutt & Hebel were concerned with the estimation of the **causal** effect of S on B, whilst acknowledging that the observed **association** between S and B might be subject to **confounding** (and, although they do not mention this characteristic, the measure of smoking frequency is also subject to **measurement error**). In order to solve their estimation problem, the authors made use of the randomization indicator (R), assuming that R is likely to be highly correlated with S but only has any effect on B through its effect on S (that is, conditional on S, randomization has no effect on B). In this context, the variable R is called an *instrumental variable*. Although instrumental variable methods had previously been widely used in econometrics, the Permutt & Hebel paper was one of the first examples of its use in a medical application. Another starting point for much of the work on adjustment for **selection** effects (particularly to adjust for the effects of noncompliance in randomized controlled trials (RCTs)) is the work of Bloom [1], Sommer and Zeger [16] and Robins [14] – (*see* **Noncompliance, Adjustment for**).

Consider two observed **random variables**, X and Y. X and Y could be quantitative or **categorical** (**binary**, for example) and they might be subject to measurement errors (**misclassification errors** in the case of binary variables). X and Y are assumed to be associated but their association might also be subject to hidden confounding (selection effects). The variables T_x and T_y are the true underlying values of X and Y (i.e. without measurement or misclassification errors), respectively. E_x and E_y are the corresponding measurement errors (assumed to be independent). Our aim is to estimate the strength of association between T_x and T_y after adjustment for any possible confounding. This might be via the estimation of a **regression** coefficient, for example, a **correlation** or an **odds ratio**. Is it possible to find a **consistent estimator**? Not without some unrealistic assumptions or further information. One might record the values of several potential confounders, for example, but there still might be the possibility of the existence of confounders that we have never thought of. We might also be able to obtain information on the characteristics of the measurement processes for X and Y (the variance of the respective measurement errors if they are quantitative, or of their **sensitivities** and **specificities** if they are binary), but there is no guarantee that the measures are performing in exactly the same way in the different circumstances. An alternative approach, and potentially a very useful one, is to obtain further information in the form of measurements of an *instrumental variable*. Suppose that we have access to a third variable, Z, which is strongly associated with T_x, but conditional on T_x has no effect on Y. Z is called an instrumental variable. Its potential will be illustrated through several familiar applications.

But, let us start by considering three hypothetical quantitative measures, X, Y, and Z. We describe the influence of Z on X through a simple **linear regression** model and, similarly the influence of X on Y through another simple linear model. But remember that X and Y are both subject to measurement error and we are really interested in the influence of T_x on T_y. A graphical representation to illustrate this general situation is provided by the **path** diagram in Figure 1. T_x and T_y are placed within circles to stress the fact that they are latent variables (unobserved, and possibly unobservable), whereas the observed or manifest variables, X, Y, and Z are themselves placed in square boxes. The key parameter we wish to estimate is the regression coefficient marked as β in this path diagram. The other important regression coefficient is marked as γ. Confounding is represented by the correlation (ρ) between the two random disturbance terms, D_x and D_y. A "1" next to an arrow on the path diagram implies that the corresponding regression coefficient is set to be 1 (i.e. $X = T_x + E_x$, for example). Although the full model described by the path diagram is hopelessly under **identified**, the parameter β is identified and it can be estimated in a number of equivalent ways. The approach we start with is to write down the expected

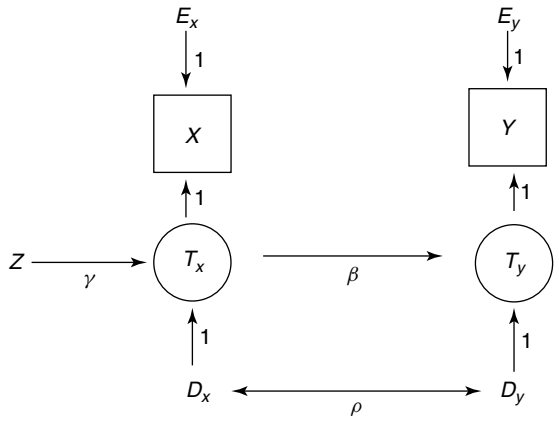

Figure 1 Illustration of a model in which we wish to estimate the effect of T_x on T_y in the presence of measurement errors (we only have observations on the error-prone variables X and Y) and selection effects (correlation between D_x and D_y). Unbiased estimation is the only problem because we also have a measure on an instrumental variable, Z

values of the covariance of Z and X, and that of Z and Y:

$$\text{Cov}(Z, X) = \gamma \text{Var}(Z) \tag{1}$$

$$\text{Cov}(Z, Y) = \gamma\beta \text{Var}(Z). \tag{2}$$

It follows immediately that:

$$\beta = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}. \tag{3}$$

One possible estimator for β is then obtained by simply estimating these two covariances from the data and substituting these estimates into (3). This is equivalent to

$$\widehat{\beta_{IV}} = \frac{\sum (Z - \overline{Z})(Y - \overline{Y})}{\sum (Z - \overline{Z})(X - \overline{X})}, \tag{4}$$

where the summation is over all subjects/observations in the dataset. It is also equivalent to

$$\widehat{\beta_{IV}} = \frac{\widehat{\beta_{yz}}}{\widehat{\beta_{xz}}}, \tag{5}$$

where $\widehat{\beta_{yz}}$ is the slope estimate obtained from the ordinary **least squares** (OLS) regression of Y on Z, and $\widehat{\beta_{yz}} = \widehat{\gamma}$ is the corresponding estimate from the OLS regression of X on Z.

Another starting point is to write down the OLS estimate for the intercept, α, and slope, β, as the solution to

$$\frac{1}{N} \sum X(Y - \widehat{\alpha_{OLS}} - \widehat{\beta_{OLS}} X) = 0, \tag{6}$$

where N is the number of observations. In our situation, of course, these estimates will be biased. It can be shown that the corresponding instrumental variable (IV) estimates are obtained from the solution to

$$\frac{1}{N} \sum Z(Y - \widehat{\alpha_{IV}} - \widehat{\beta_{IV}} X) = 0. \tag{7}$$

These are both special cases of generalized **method of moments** (GMM) estimators [7].

By far the most common IV estimation procedure, however, is the use of OLS regression in two stages – ***two-stage least squares*** (2SLS or TSLS). First one regresses X on Z to obtain the predicted value of X, that is $\widehat{X}$. The second stage involves the regression of Y on $\widehat{X}$. The estimate of the slope from the second stage is the equivalent to $\widehat{\beta_{IV}}$ above. If we were to actually do this in two stages, we would have to be aware that the second regression would give an incorrect **standard error** for $\widehat{\beta_{IV}}$. Most general-purpose **software** packages have a 2SLS routine that provides the correct standard errors and corresponding ***P* values**. The asymptotic variance of $\widehat{\beta_{IV}}$ is, in fact,

$$\text{Var}(\widehat{\beta_{IV}}) = \frac{\sigma_u^2}{N\sigma_X^2 \rho_{XZ}^2}, \tag{8}$$

where σ_X^2 is the variance of X, σ_u^2 is the variance of the deviations $u = Y - \widehat{\alpha_{IV}} - \widehat{\beta_{IV}} X$ and ρ_{XZ}^2 is the square of the correlation between X and Z (see [6] or [21] for further details). Note that, keeping everything else constant, this variance decreases with increasing values of the correlation between X and the instrumental variable Z. Note also that neither the IV estimate itself nor its asymptotic variance is dependent on any distributional assumptions. In particular, both Z and X could be binary (see section on adjustment for noncompliance in an RCT, below). If we have covariates other than the instrumental variable (or instrumental variables if we can find more than one) then the first stage of 2SLS involves the regression of X and the covariates. The second stage involves regressing Y on $\widehat{X}$ and the same covariates [21].

One final approach to IV estimation for the linear model is again a two-stage procedure. The first stage, as before, involves the regression of X on Z. But in this case we keep the **residuals** from the regression for the next stage, rather than the predicted values of X. The second stage involves the regression of Y on *both* the observed X *and* the residual from the first stage regression (see [11], for example). Again, the estimate is equivalent to $\widehat{\beta_{IV}}$ as defined above. Once again, we need care in the estimation of the correct standard errors.

IV estimation, in general, and 2SLS estimation, in particular, appear to be rarely used in medical applications (and when IV methods are used they frequently do not get an explicit description as such), with the obvious exception of the application of **econometric applications in health services** research. The methods are rarely mentioned in medical statistics texts. The general econometrics literature on IV estimation via 2SLS (and other more complex estimation methods), however, is vast. It is one of the most widely used statistical methodologies in econometrics (after OLS methods). Any textbook on econometrics will contain at least a chapter on IV/2SLS methodology. Good introductions can be found in [19, 20, 21]. Advanced topics can be found in [7]. A range of methods of IV estimation for the linear model has been provided in this section, partly to point out the equivalence of the several different approaches in the context of linear models, but mainly because they each can provide an approximate solution for nonlinear problems (**logistic regression** models in epidemiology, for example).

Estimation of Treatment or Exposure Effects in Biostatistics

Elementary reviews of the use of instrumental methods in outcomes of treatment research and in epidemiology are provided by [8, 12]. Research on the outcomes of treatment using instrumental variable methodologies has tended to concentrate on the problems of selection effects in the choice of treatments (for the special case of noncompliance in randomized studies, *see* **Noncompliance, Adjustment for**). On the whole, measurement errors in the classification of treatment received, or in the amount of treatment received, have not received much attention. In epidemiology on the other hand, it is the problem of **exposure** measurement that has lead to a surge of interest in instrumental variable methods (*see* **Measurement Error in Epidemiologic Studies**). Stefanski and Buzas [18], for example, have developed instrumental variable estimation methods in binary regression models using exposure measurements subject to measurement errors. These authors have used approximations to those in the linear model, such as the use of (3), for example, in which the two regression coefficients are obtained from the appropriate logistic regressions instead of OLS. The use of 2SLS itself does not appear to be very satisfactory for nonlinear instrumental variable regression problems (see Foster [5], for example, who advocates GMM-based methods). Nagelkerke et al. [11], however, successfully apply the second of the two-stage methodologies described above to a range of nonlinear models involving noncompliance with randomized treatment allocation. An interesting example of the development of the IV method to test some of the assumptions of a measurement error model is provided by Spiegelman et al. [17]. For a general and detailed discussion of exposure **measurement error problems in epidemiology**, readers are referred to the article in the present encyclopedia and to [2]. Greenland [8] describes how an instrumental variable can be used to adjust for hidden confounders in an epidemiological study.

Method Comparison Studies

Consider a study in which we wish to compare either two quantitative methods of measurement or two binary **diagnostic tests**. In neither case do we have access to a **gold standard** (*see* **Diagnostic Test Evaluation Without a Gold Standard**). An appropriate measurement model for two quantitative measuring instruments, X and Y might have the following form:

$$\begin{aligned} \mathrm{X} &= \alpha_{\mathrm{x}} + \beta_{\mathrm{x}}\tau + \delta, \\ \mathrm{Y} &= \alpha_{\mathrm{y}} + \beta_{\mathrm{y}}\tau + \varepsilon, \end{aligned} \tag{9}$$

where τ is "true" value of the material being measured, the α's and β's are the parameters of the two regression equations, and δ and ε are random measurement errors (assumed to be independent of each other and of the true value, τ). Our aim is to estimate the regression coefficients, together with the

variances of the measurement errors. The data from such a study can be summarized by the means and variances of X and Y, together with their covariance. Clearly, the model described by (9) is underidentified (there are far too many parameters to be estimated from the small number of summary statistics). We first need to specify a scale of measurement. One convenient way of doing this is to consider one of the methods as a standard (X, for example, by setting $\alpha_x = 0$ and $\beta_x = 1$). But the model is still underidentified. We cannot estimate the free parameters of the model without either making further assumptions or by obtaining further data. One useful approach for the latter is to obtain a third measurement, Z, that is correlated with the true value (τ) but conditional on τ is uncorrelated with both X and Y. Z is an instrumental variable. One obvious choice for Z is a third measurement of τ, taking care to ensure that the measurement errors for Z are not correlated with those of either X or Y. It is now straightforward to estimate all of the free parameters of the model (β_y, for example, is obtained from (3)). Details can be found in [3, 4, 6].

In the case of the comparison of two binary diagnostic tests (X and Y), we start by postulating the existence of a binary **latent class** (the "true" diagnosis) and related the observed tests results to the true status of the patient through a latent class or finite mixture model. With only two diagnostic tests, however, the model is underidentified [9]. We need more information. Again, we resort to an instrumental variable, Z. Z could be either categorical (but not necessarily binary) or quantitative. The only requirement is that it is strongly associated with the true status of the patient, but conditional on that true status is independent of both X and Y. Z, of course, could be a third diagnostic test. Details can be found in [4, 9, 10]. In all the applications of instrumental variable methods, the existence of measurements on a single instrumental variable enable us to get a handle on the estimation of otherwise unidentified model parameters. If we have data on more than a single instrumental variable, however, not only can we get even better estimates, but we might also be able to use the extra information to check the validity of our modeling assumptions. In the context of method comparison studies, for example, models involving three variables are just identified. That is, they fit the data perfectly and there is no way of testing whether any of the necessary assumptions might be invalid. If we have four or more measurements, however, we have the possibility of an overidentified model together with degrees of freedom to test model fit. Nagelkerke et al. [10] provide a nice example involving diagnostic tests. Dunn & Roberts [3] illustrate the point with quantitative measures.

References

[1] Bloom, H.S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **8**, 225–246.

[2] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

[3] Dunn G. (2004). *The Statistical Evaluation of Measurement Errors*, 2nd Ed. Arnold, London.

[4] Dunn, G. & Roberts, C. (1999). Modelling method comparison data. *Statistical Methods in Medical Research* **8**, 161–179.

[5] Foster, E.M. (1997). Instrumental variables for logistic regression: an illustration. *Social Science Research* **26**, 487–504.

[6] Fuller, W.A. (1987). *Measurement Error Models*. Wiley, New York.

[7] Greene, W.H. (2000). *Econometric Analysis*, 4th Ed. Prentice Hall, Upper Saddle River.

[8] Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29**, 722–729 (Erratum p. 1102).

[9] Hui, S.L. & Walter, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36**, 167–171.

[10] Nagelkerke, N., Fidler, V. & Buwalda, M. (1988). Instrumental variables in the evaluations of diagnostic test procedures when the true disease status is unknown. *Statistics in Medicine* **7**, 739–744.

[11] Nagelkerke, N., Fidler, V. Bersen, R. & Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* **19**, 1849–1864.

[12] Newhouse, J.P. & McClellan. M. (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual Reviews of Public Health* **19**, 17–34.

[13] Permutt, T. & Hebel, J.R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking and birth weight. *Biometrics* **45**, 619–622.

[14] Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested means models. *Communications in Statistics – Theory and Methods* **23**, 2379–2412.

[15] Sexton, M. & Hebel, J.R. (1984). A clinical trial of change in maternal smoking and its effect on birth weight. *Journal of the American Medical Association* **251**, 911–915.

[16] Sommer, A. & Zeger, S.L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10**, 45–52.

[17] Spiegelman, D., Schneeweiss, S. & McDermott, A. (1997). Measurement error correction for logistic regression models with an "Alloyed Gold Standard". *American Journal of Epidemiology* **145**, 184–196.

[18] Stefanski, L. & Buzas, J.S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association* **90**, 541–549.

[19] Verbeek, M. (2000). *A Guide to Modern Econometrics*. Wiley, New York.

[20] Wooldridge, J.M. (2001). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

[21] Wooldridge, J.M. (2003). *Introductory Econometrics: A Modern Approach*, 2nd Ed. United.

GRAHAM DUNN

Intention to Treat Analysis

Intention-to-treat (ITT) analysis (also referred to as "as-randomized" or "method effectiveness" analysis [25]) is defined in the context of a randomized **clinical trial** (RCT). In a RCT design for the comparison of treatments, subjects are randomly assigned to different treatments. Once this **randomized assignment to treatment** is made, the ITT principle requires that any comparison of the treatments is based upon comparison of the outcome results of all patients in the treatment groups to which they were randomly assigned. This approach is recommended to maintain the benefits of randomization.

Randomization provides two important features. First, the treatment assignment is based on chance alone. The characteristics of the group of patients receiving the different treatments should be roughly equivalent at the onset of the trial, with the only difference being their treatment assignment. If, during the implementation of the trial, the groups continue to be distinguished only by their treatment assignment, then any differences in the outcomes of the groups at the end of the study (*see* **Outcome Measures in Clinical Trials**) can be attributed solely to difference in treatment. Secondly, randomization provides the theoretical foundation for the statistical tests of significance that are used to test for observed differences (*see* **Randomization Tests**) [2, 7]. These two benefits of randomization are the foundations of the science of comparative clinical trials.

Unfortunately, once a clinical trial begins, several predictable, as well as unforeseen, conditions can (and usually do) influence the actual vs. the intended protocol under which individual subjects in each group are studied (*see* **Clinical Trials Protocols**). Thus, the groups, as treated, may no longer be comparable at the end of the trial. These circumstances may include: subjects who do not adhere to the assigned treatment regimen (*see* **Compliance Assessment in Clinical Trials**); subjects whose eligibility for trial participation has changed or was incorrect at the start of the trial (*see* **Eligibility and Exclusion Criteria**); subjects whose treatment assignment was incorrect; or subjects who terminate participation in the trial prior to the measurement of the main clinical outcome.

Partial or complete noncompliance with the regimen of the assigned treatment is fairly common in medical treatment trials. Patients may not be able to tolerate the side effects of their randomly assigned treatment and may request the termination of all medication or switch to another of the study treatments. Less common is a clerical or computer error that results in a patient being given medication other than that randomly assigned. Although such patients may take their medication faithfully, they will be noncompliant with their assigned treatment. In the extreme, widespread failure to comply with the assigned regimen can destroy a study. For example, in a trial comparing an active treatment with a control, if nearly all of the subjects randomly assigned to the active treatment do not take the drug, then a comparison of efficacy between the two groups will be meaningless. Less extreme, but very common, are instances where patients do not take the prescribed dosage of a treatment, or take the drug intermittently or for a limited duration [9, 25].

It is expected that, in general, there will be reasonable adherence to protocol in a substantial proportion of patients within each of the treatment groups. For those subjects not adhering to the protocol, the question may be raised as to whether they should be excluded from the analysis, the concern being that they do not provide information relevant to the efficacy of treatment taken as prescribed. By definition, ITT analysis does not allow treatment comparisons using only those subjects compliant with the therapies under test. Use of ITT analysis requires, with very few exceptions, that all subjects with valid outcomes be (i) included in the analysis and (ii) analyzed according to their randomly assigned treatment. Subjects who comply with their randomly assigned treatment for only a short time after its initiation, or switch to a competing treatment, are still considered, for analysis purposes, in their randomly assigned treatment group. This ensures that the randomization is protected; that is, the treatment groups can be assumed to include patients equivalent prior to the onset of therapy, and the possibility that the analysis could have been inadvertently biased by "selective" exclusion or inclusion of subjects into treatment groups is eliminated (*see* **Selection Bias**).

Contrary to the strict interpretation of the ITT principle, investigators may attempt to compare groups defined not solely by the original randomization but by factors that might be influenced by the treatments

under test. In the face of substantial noncompliance, the comparison of only those patients in the trial that actually complied with the prescribed treatment is intuitively appealing. However, in addition to the difficulty of defining compliance in an objective manner, it has been seen that subjects who comply tend to fare differently and in a sometimes unpredictable way from those who do not comply [1, 3, 14, 27]. Thus, any observed differences among treatment groups constructed in this manner may be due not to treatment, but to factors associated with compliance.

The ITT analysis approach often provides a conservative estimate of the effect of a treatment administered as prescribed. The inclusion of noncompliant patients in the assessment of efficacy, barring some peculiar **dose–response** relationship, dilutes the difference between outcomes in the treatment and control groups. For the same reasons, ITT analysis may underestimate the risk of adverse side effects.

The ITT approach may be viewed as evaluating a *treatment strategy,* as contrasted to evaluating the efficacy of a treatment taken as prescribed [14]. The effectiveness of a treatment strategy is a reasonable approximation to the effectiveness of a prescribed regimen in the community [18]. Patients prescribed a treatment outside of a clinical trial often exhibit the same or an increased level of noncompliance, without the extra encouragement to comply that is provided in most clinical trials.

In contrast to the issue of noncompliance with the study treatment regimen, the determination of a subject's eligibility for trial participation after initiation of randomized treatment is a circumstance where many clinical trialists feel that a strict interpretation of the ITT principle can produce study results that are simply not credible [9, 15]. For example, in a comparative study of treatments for sepsis in newborns, treatment must usually be started as soon as there is a presumptive diagnosis of sepsis. Sepsis is too dangerous to be left untreated, and definitive laboratory tests are not immediately available, so treatment is usually started at the first sign of infection. If subjects with a presumptive diagnosis are assigned treatment by randomization, then a portion in each of the treatment groups will be later proven not to have had sepsis. Strict application of the ITT principle requires that the subjects proved not to have had sepsis nevertheless be included in the analysis in the treatment group to which they were assigned. A more relaxed application acknowledges that sepsis was or was not present *prior to randomization* and that its presence *was not known* due to the absence of confirmatory laboratory information. Thus, the exclusion from analysis of those patients without sepsis, and thus ineligible for the trial, would be appropriate since there is no conceivable way in which the treatment assignment could have influenced which subjects previously had sepsis. The choice of analysis in this example affects not only the credibility of the study report, which might have included in the analysis patients without sepsis, but also the measure of the effect of treatment. Including the subjects without sepsis in the treated groups would provide biased estimates of efficacy, since those without the disease would be counted as cured. In contrast, for the assessment of safety, inclusion of all subjects with and without a definitive diagnosis of sepsis is reasonable. The larger sample size will allow for the detection of differences in rarer side effects.

There is a concern that if exceptions to ITT analysis, as in the sepsis example above, are routinely allowed, then analysts will inevitably be tempted to adopt exceptions or exclusions that do **bias** the results of a study. An example, not as clear as the sepsis trial, is the administration of a study drug other than that actually assigned. This can happen on rare occasions because of a clerical error at the start of treatment, or, more commonly, because a subject decides to change medication early in the course of their assigned treatment. The first example seems simple enough; an unintentional clerical mistake resulted in an incorrect drug being dispensed. It might, therefore, be reasonable to include the subject in the analysis as having taken the received treatment. The second example, however, raises severe problems, since the drug itself may have caused the switch, possibly due to a perceived lack of efficacy or unpalatable side effects [19]. Most analysts would consider the latter example to constitute crossing the ITT "line in the sand".

Patients dropping out of a clinical trial before their endpoint can be measured can bias a study severely regardless of the analysis approach utilized. If, for example, one of the tested treatments has unpleasant side effects, patients with mild disease might view the side effects as a hardship when contrasted to their mild affliction, and discontinue participation. This would create an imbalance among the treatment groups with regard to severity of disease. Information on some clinical outcomes such as mortality might be obtainable given enough time, even if a subject drops

out from a study. Other outcomes, such as laboratory evaluations at a specific time point after baseline, will not be available from alternative sources. Clinical trials should be designed and organized with the resources available and directed so that dropouts will be minimal and hopefully at random [8]. Several authors have proposed methods for imputing data for patient endpoints on the basis of previous data obtained in the study [4, 11, 13, 16], although Lachin [12] points out that of "... the majority of methods in common use ... none allow or adjust for the bias introduced by nonrandomly censored or missing observations" (*see* **Missing Data in Clinical Trials**).

The above discussion has centered around comparative trials of efficacy, with the primary example the comparison of an active treatment with a placebo, and the inherently conservative nature of an ITT analysis has been mentioned. A large number of clinical trials, known as **equivalence trials**, seek to show that a new treatment is equal to an established treatment with regard to efficacy, while being less toxic or less expensive. Suppose that the new drug, however, is less efficacious than the established drug. If subjects who fail to comply with their assignment to the new treatment, or who switch to the standard treatment, are, for analysis purposes, considered in the group to which they were originally assigned, the difference between outcomes in the two groups is brought closer together. The established treatment group will include subjects taking the less effective treatment or taking no treatment, and the new treatment group will include subjects taking the more effective treatment and/or no treatment. Thus, the ITT approach for equivalence trials may tend to mask true differences, making it easier to conclude that treatments are equivalent when they are, in fact, not [15].

Statisticians and clinical trialists generally agree that some form of the ITT principle is appropriate for most efficacy trials. The strictness of application of the principle still raises considerable discussion.

Alternative Analysis Strategies

Alternatives to ITT analysis generally attempt either to restrict analysis to those subjects who have adhered to a treatment protocol, or to incorporate measures of compliance to treatment in comparative analyses. Many titles have been given to the first of these approaches including "as-treated analysis" [6], "treatment received" [20], "explanatory approach" [24], "method effectiveness" [25], "per-protocol" [25], "efficacy analyzable patients" [9], and "biologic efficacy" [26]. The phrase "as-treated" (AT) analysis will be used in this article.

In AT analysis (i) only patients considered compliant with one of the study treatments are included in the analysis, and (ii) outcomes of subjects are attributed to the treatment groups on the basis of the treatment actually taken, regardless of their randomly assigned treatment. It is argued that the primary interest of a comparative clinical trial is in testing whether a treatment, taken as prescribed, is effective. In contrast, the ITT approach dictates that all subjects be included in the analysis, even those who have not taken the prescribed treatment; and that subjects who have complied with a study treatment other than that assigned by randomization nevertheless be counted as having taken the assigned treatment. The ITT approach is counterintuitive to many clinicians and other scientists. It is considered by AT proponents as incorrect, and it is understood by those arguing for either ITT or AT analysis that ITT analysis will, in the face of noncompliance, provide a diluted estimate of efficacy.

The primary argument against the AT approach is that it can lead to biased comparisons of treatment groups, in contrast to the ITT approach as outlined above. That is, it can lead to claims of efficacy even when a treatment is nonefficacious. There are also some difficult practical problems with AT analysis, including difficulty in defining compliance [14, 21], difficulty in determining which subjects are compliant, and loss of sample size when analysis is limited to compliant subjects.

Another alternative to ITT analysis has been developed in recent work [5, 10, 17, 22, 23, 25], which attempts to incorporate measures of compliance into the statistical analysis. This approach focuses on the information that compliance has to offer, rather than on considering compliance as a defining characteristic for the inclusion of patients in, or the exclusion of patients from, analysis. The goal of the new work is to use compliance data to provide better estimates of, or understanding of, the clinical response to treatment. These model-based approaches currently require assumptions about compliance, either its relation to treatment or to outcome, that may be difficult to accept or verify. Nonetheless,

this work may help to breach the gap between the current ITT and AT positions, and may be particularly useful for analyzing equivalence trials (*see* **Noncompliance, Adjustment for**).

Current Status

ITT analysis is a widely used strategy for the analysis of comparative clinical trials in the definitive comparison of treatments for both regulatory and nonregulatory assessments. "AT analysis" is used frequently as a secondary and confirmatory analysis to the ITT analysis, or for explanatory or exploratory assessment of efficacy in subgroups defined by characteristics or factors that could have differential response rates, such as compliance, gender or ethnic groups.

References

[1] Azurin, J.C. & Alvero, M. (1971). Cholera incidence in a population offered cholera vaccination: comparison of cooperative and uncooperative groups, *Bulletin of the World Health Organization* **44**, 815–819.

[2] Byar, D.P., Simon, R.M., Friedewald, W.T., Schlesselman, J.J., DeMets, D.L., Ellenberg, J.H., Gail, M.H. & Ware, J.H. (1976). Randomized clinical trials: perspectives on some recent ideas, *New England Journal of Medicine* **295**, 74–80.

[3] Canner, P.L., Forman, S.A., Prud'homme, G.J., Berge, K.G. & Stamler, J. (1980). Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project, *New England Journal of Medicine* **303**, 1038–1041.

[4] Efron, B. (1994). Missing data and the bootstrap (Abstract), *Journal of the American Statistical Association* **89**, 463–474.

[5] Efron, B. & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials, *Journal of the American Statistical Association* **86**, 9–26.

[6] Ellenberg, J.H. (1996). Intent-to-treat analysis versus as-treated analysis, *Drug Information Journal* **30**, 535–544.

[7] Fisher, L., Dixon, D.O., Herson, J., Frankowski, R., Hearron, M.S. & Peace, K.E. (1990). Intention to treat in clinical trials, in *Statistical Issues in Drug Research and Development*, K.E. Peace, ed. Marcel Dekker, New York.

[8] Freidman, L.M., Furberg, C.D. & DeMets, D.L. (1996). *Fundamentals of Clinical Trials*, 3rd Ed. Mosby-Year Book, St. Louis.

[9] Gillings, D. & Koch, G. (1991). The application of the principle of intention-to-treat to the analysis of clinical trials, *Drug Information Journal* **25**, 411–424.

[10] Goetghebeur, E., Molenberghs, G. & Katz, J. (1998). Estimating the causal effect of compliance on binary outcome in randomized controlled trials, *Statistics in Medicine* **17**, 341–355.

[11] Greenless, J.S., Reece, W.S. & Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed (Abstract), *Journal of the American Statistical Association* **77**, 251–261.

[12] Lachin, J.M., (2000). Statistical considerations in the intent-to-treat principle, *Controlled Clinical Trials* **21**, 167–189.

[13] Laird, N.M. (1988). Missing data in longitudinal studies (Abstract), *Statistics in Medicine* **7**, 305–315.

[14] Lee, Y.J., Ellenberg, J.H., Hirtz, D.G. & Nelson, K.B. (1991). Analysis of clinical trials by treatment actually received: is it really an option? *Statistics in Medicine* **10**, 1595–1605.

[15] Lewis, J.A. & Machin, D. (1993). Intention-to-treat–who should use ITT? *British Journal of Cancer*, **68**, 647–650.

[16] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[17] Nagelkerke, N., Fidler, V., Bernsen, R. & Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance, *Statistics in Medicine*, **19**, 1849–1864.

[18] Newell, D.J. (1992). Intention-to-treat analysis: implications for quantitative and qualitative research, *International Journal of Epidemiology* **21**, 837–841.

[19] Peduzzi, P., Detre, K., Wittes, J. & Holford, T. (1991). Intent-to-treat analysis and the problem of crossovers, *Journal of Thoracic and Cardiovascular Surgery* **101**, 481–487.

[20] Peduzzi, P., Wittes, J. & Detre, K. (1993). Analysis as-randomized and the problem of non-adherence: an example from the veterans affairs randomized trial of coronary artery bypass surgery, *Statistics in Medicine* **12**, 1185–1195.

[21] Redmond, C., Fisher, B. & Wieand, H.D. (1983). The methodologic dilemma in retrospectively correlating the amount of chemotherapy received in adjuvant therapy protocols with disease-free survival, *Cancer Treatment Reports* **67**, 519–526.

[22] Rochon, J. (1995). Supplementing the intent-to-treat analysis: accounting for covariates observed postrandomization in clinical trials, *Journal of the American Statistical Association* **90**(429), 292–300.

[23] Rubin, D.B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance, *Statistics in Medicine* **17**, 371–385.

[24] Schwartz, D. & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials, *Journal of Chronic Diseases* **20**, 637–648.

[25] Sheiner, L.B. & Rubin, D.B. (1995). Intention-to-treat analysis and the goals of clinical trials, *Clinical Pharmacology and Therapeutics* **57**, 6–15.

[26] Sommer, A. & Zeger, S.L. (1991). On estimating efficacy from clinical trials, *Statistics in Medicine* **10**, 45–52.

[27] Tarwotjo, I., Sommer, A., West, K.P., Djunaedi, E., Loedin, A.A., Mele, L. & Hawkins, B. (1987). Influence of participation on mortality in a randomized trial of vitamin A prophylaxis, *American Journal of Clinical Nutrition* **45**, 1466–1471.

JONAS H. ELLENBERG

Interaction in Factorial Experiments

An important benefit obtained by using **factorial experiments** is the ability to determine whether **interactions** are present. In this context, interaction may be defined as the modification of the effect of a factor on a response, due to the influence of another factor. Put another way, the presence of an interaction means that the relationship between one factor and a response is different for different levels of another factor.

A typical factorial model which includes the interaction between factors A and B can be written as

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + \gamma_{ij}, \quad \text{for all } i, j,$$

where μ_{ij} represents the mean response across all observations when factor A is at level i and factor B is at level j, $\mu_{..}$ represents the overall mean, α_i represents the main effect for the ith level of factor A, β_j represents the main effect for the jth level of factor B, and γ_{ij} represents the interaction effect of the ith level of factor A and the jth level of factor B. This model clearly indicates that the effects of the factors are not simply additive, as would be the case if the interaction term were deleted (*see* **Additive Model**). Note that since the present discussion is concerned with understanding the meaning and utility of the interaction term, no constraints are imposed on the model parameters, and so models given will generally be overparameterized models.

By manipulation of the model given above, and defining the main effects as $\alpha_i = \mu_{i.} - \mu_{..}$ and $\beta_j = \mu_{.j} - \mu_{..}$, the interaction term can be expressed as a difference of differences; namely

$$\begin{aligned}\gamma_{ij} &= (\mu_{ij} - \mu_{i.}) - (\mu_{.j} - \mu_{..}) \\ &= (\mu_{ij} - \mu_{.j}) - (\mu_{i.} - \mu_{..}),\end{aligned}$$

where the "." in the subscript indicates summation over all levels of that subscript. Some authors refer to functions like $\mu_{ij} - \mu_{aj} - \mu_{ib} + \mu_{ab}$, i.e. differences of differences of cell means, as interaction effects. While such quantities are readily evaluated and interpreted from tables or graphs of treatment combination means, they are not equal to the interaction effects defined above. Such functions are actually the corresponding functions of the interaction effects, namely, $\gamma_{ij} - \gamma_{aj} - \gamma_{ib} + \gamma_{ab}$. Hence, it is true that if any of the functions $\mu_{ij} - \mu_{aj} - \mu_{ib} + \mu_{ab}$ are nonzero, then interactions are present.

Traditionally, the presence of interactions in a model has been viewed as something to be avoided, if possible. Didactic presentations of factorial models often emphasize testing for the significance of the interaction terms (*see* **Hypothesis Testing**), with the hope that the test statistics would be nonsignificant. If this no-interaction model is tenable, the relationship between the two factors and the response is easy to explain. However, in many real-life situations, interactions are appropriately included in the statistical model, since the relationship between a set of factors and the response goes beyond the simple additive (i.e. no-interaction) model.

Indeed, in certain situations the presence of interactions is viewed as desirable. For example, consider a randomized **clinical trial** where repeated measurements are taken through the course of a study comparing the effects of two different treatments on a disease of interest. The researcher anticipates that at the time of **randomization**, the average response will be the same in both treatment groups, but will eventually become different as the treatments have their desired effect. It is this divergence of response that is appropriately reflected by the treatment–time interaction effect in the analysis model.

The number of two-factor interaction terms in the model, corresponding to the **degrees of freedom** associated with that interaction effect, is the product of the numbers of levels of the two main effects in the models. By extension, the number of degrees of freedom associated with higher-order interaction terms increases multiplicatively with the number of effects included in the interaction term. This is especially noticeable when the number of levels of one or more of the factors increases beyond a simple dichotomy. In this situation, estimates of interaction effects may become unstable if the number of such terms increases but the sample size remains fixed. This instability is partly due to the loss of degrees of freedom associated with the estimated residual **variance**, since the number of observations at the combinations of the factor levels may become too small. Furthermore, computational problems can occur because the model specification may have included some interaction terms in the model which involve unobserved treatment combinations. Strong computational **algorithms** will build in methods for

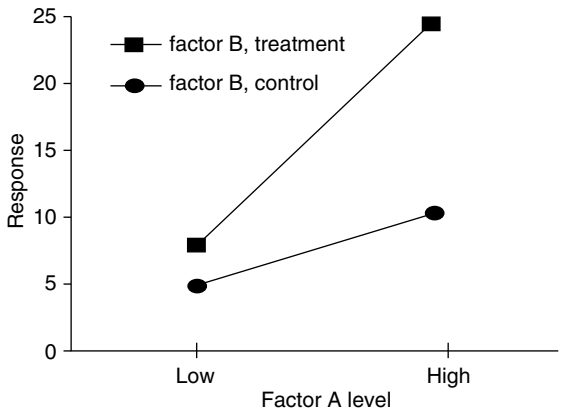

Figure 1 Graph of A–B interaction effect

recognizing and dealing appropriately with such possibilities.

As indicated in the graph (Figure 1), the interaction of two classification factors, or of a classification and a continuous factor, is straightforward to graph, and hence to interpret. However, comparable interpretation of interaction effects containing two or more continuous factors is considerably more complex. Practical experience suggests that it is next to impossible to explain and/or interpret interaction effects containing more than three terms in a straightforward verbal manner. While graphs may be useful in this regard, if the graphs involve more than one, or possibly two, continuous factors, the dimensionality of the required graph may be unreasonable.

In some situations, clarity of interpretation of interaction effects can be obtained by modeling these effects with nested effects (*see* **Multilevel Models**) rather than as true interaction effects. That is, instead of using the terms α_i, β_j, and γ_{ij} in the model that deals with factors A and B and the A $*$ B interaction, we use α_i and terms for effect B that are different for each level of A. This may be more useful as one attempts to understand the underlying relationships among the variables.

When constructing a model with the possible inclusion of interactions, the investigator should be attentive to the convention of using the hierarchy principle for determining which interactions should be included in a model. For a model to be hierarchical, the inclusion of any interaction term mandates that all lower-order effects which include the effects in the interaction must also be included in the model (*see* **Hierarchical Models**). That is, if the AB interaction effect is in the model, then the A and B main effects must also be included. By extension, if the ABC interaction effect is used in a model, then the interaction effects AB, AC, and BC, as well as the A, B, and C main effects, must be included.

Interaction is something different from what epidemiologists refer to as **confounding**. In that literature, confounding refers to the change found in the relationship between a factor (often called an effect in epidemiology) and a response when another effect (the confounding effect) is added to the model. Interaction effects need not be included in a model for confounding to be present. Rather, confounding refers to the presence of significant **correlation** among the effects.

The use of interaction effects can be especially useful, and perhaps difficult to implement, when **time-dependent covariates** are used in the model. For example, in the analysis of growth of infants in the first year of life (*see* **Growth and Development**), it would be useful to utilize information about whether the child is being breast-fed at each measurement time. Inclusion of an indicator variable about the breast-feeding condition will simply yield a linear shift of the growth curve. However, if the interaction term involving feeding condition and age is included in the model, the rate of growth can be modeled differently under the two feeding conditions.

References

[1] Fisher, L.D. & van Belle, G. (1993). *Biostatistics: A Methodology for the Health Sciences*. Wiley, New York, pp. 435–443.

[2] Forthofer, R.N. & Lee, E.S. (1995). *Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery*. Academic Press, San Diego, pp. 395–404.

[3] Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman, W. (1996). *Applied Linear Statistical Models*. Richard D. Irwin, Chicago, pp. 805–812.

ROBERT ANDERSON

Interaction Model

Interaction models for **categorical data** are **loglinear models** describing association among categorical variables. They are called interaction models because of the analytic equivalence of loglinear **Poisson regression** models describing the dependence of a count variable on a set of categorical explanatory variables and loglinear models for **contingency tables** based on **multinomial** or product multinomial sampling. The term is, however, somewhat misleading, because the interpretation of parameters from the two types of models are very different. *Association models* would probably be a better name.

Instead of simply referring the discussion of interaction and association models to the section on loglinear models, we will consider these models from the types of problems that one could address in connection with analysis of **association**. The first problem is a straightforward question of whether or not variables are associated. To answer this question, one must first define association and dissociation in multivariate frameworks and, secondly, define multivariate models in which these definitions are embedded. This eventually leads to a family of so-called graphical models that can be regarded as the basic type of interaction or association. The second problem concerns the properties of the identified associations. Are associations homogeneous or heterogeneous across levels of other variables? Can the strength of association be measured and in which way? To solve these problems, one must first decide upon a natural measure of association among categorical variables and, secondly, define a parametric structure for the interaction models that encapsulates this measure. Considerations along these lines eventually lead to the family of hierarchical loglinear models for nominal data and models simplifying the basic loglinear terms for **ordered categorical data**.

Graphical Interaction Models

What is meant by association between two variables? The most general response to this question is indirect. Two variables are dissociated if they are *conditionally* independent given the rest of the variables in the multivariate framework in which the two variables are embedded. Association then simply means that the two variables are not dissociated.

Association in this sense is, of course, not a very precise statement. It simply means that conditions exist under which the two variables are not independent. Analysis of association will typically have to go beyond the crude question of whether or not association is present, to find out what characterizes the conditional relationship – for instance, whether it exists only under certain conditions, whether it is homogeneous, or whether it is modified by outcomes on some or all the conditioning variables. Despite the inherent vagueness of statements in terms of unqualified association and dissociation, these statements nevertheless define elegant and useful models that may serve as the natural first step for analyses of association in multivariate frames of inference. These so-called *graphical* models are defined and described in the subsections that follow.

Definition

A graphical model is defined by a set of assumptions concerning pairwise conditional independence given the rest of the variables of the model.

Consider, for instance, a model containing six variables, A to F. The following set of assumptions concerning pairwise conditional independence defines four constraints for the joint distribution $\Pr(A, B, C, D, E, F)$. The family of probability distributions satisfying these constraints is a graphical model:

$$\begin{aligned}
A \perp C|BDEF &\Leftrightarrow \Pr(A, C|BDEF) \\
&= \Pr(A|BDEF)\ \Pr(C|BDEF), \\
A \perp D|BCEF &\Leftrightarrow \Pr(A, D|BCEF) \\
&= \Pr(A|BCEF)\ \Pr(D|BCEF), \\
B \perp E|ACDF &\Leftrightarrow \Pr(B, E|ACDF) \\
&= \Pr(B|ACDF)\ \Pr(E|ACDF), \\
C \perp E|ABDF &\Leftrightarrow \Pr(C, E|ABDF) \\
&= \Pr(C|ABDF)\ \Pr(E|ABDF).
\end{aligned}$$

Interaction models defined by conditional independence constraints are called "graphical interaction models", because the structure of these models can be characterized by so-called interaction graphs, where

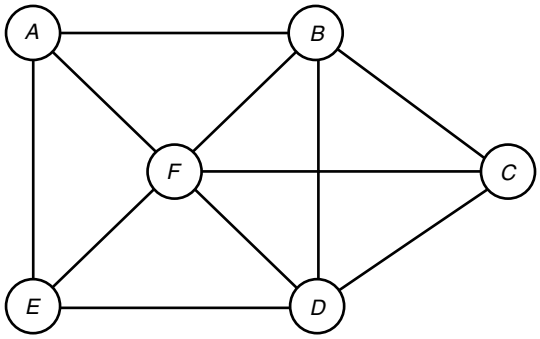

Figure 1 An interaction graph

variables are represented by nodes connected by undirected edges if and only if association is permitted between the variables. The graph shown in Figure 1 corresponds to the set of conditional independence constraints above, because there are no edges connecting A to C, A to D, B to E, and C to E.

Interaction graphs are visual representations of complex probabilistic structures. They are, however, also mathematical models of these structures, in the sense that one can describe and analyze the interaction graphs by concepts and **algorithms** from mathematical graph theory and thereby infer properties of the probabilistic model. This connection between probability theory and mathematical graph theory is special to the graphical models.

The key notion here is conditional independence, as discussed by Dawid [5]. While the above definition requires that the set of conditioning variables always includes all the other variables of the model, the results described below imply that conditional independence may sometimes be obtained if one conditions with certain subsets of variables.

Graphical models for multidimensional tables were first discussed by Darroch et al. [5]. Since then, the models have been extended both to continuous and mixed categorical and continuous data and to regression and block recursive models. Whittaker [9], Edwards [7], Cox & Wermuth [4], and Lauritzen [8] present different accounts of the theory of graphical models. The sections below summarize some of the main results from this theory.

The Separation Theorem

The first result connects the concept of graph separation to conditional independence.

First, we present a definition: a subset of nodes in an undirected graph separate two specific nodes, A and B, if all paths connecting A and B intersect the subset. In Figure 1, (B, D, F) separate A and B, as does (B, E, F). E and C are separated by both (A, D, F) and (B, D, F).

The connection between graph separation and conditional independence is given by the following result, sometimes referred to as the separation theorem.

Separation Theorem. If variables A and B are conditionally independent given the rest of the variables of a multivariate model, A and B will be conditionally independent given any subset of variables separating A and B in the interaction graph of the model.

The four assumptions on pairwise conditional independence defining the model shown in Figure 1 generate six minimal separation hypotheses:

$$A \perp C|BDF, \quad A \perp C|BEF, \quad A \perp D|BEF,$$
$$B \perp E|ADF, \quad C \perp E|ADF, \quad C \perp E|BDF.$$

Closure and Marginal Models

It follows from the separation theorem that graphical models are closed under marginalization, in the sense that some of the independence assumptions defining the model transfer to marginal models.

Collapsing, for instance, over variable C of the model shown in Figure 1 leads to a graphical model defined by conditional independence of A and D and B and E, respectively, because the marginal model contains separators for both AD and BE (Figure 2).

Loglinear Representation of Graphical Models for Categorical Data

No assumptions have been made so far requiring variables to be categorical. If all variables are categorical, however, the results may be considerably strengthened both with respect to the type of model defined by the independence assumptions of graphical models and in terms of the available information on the marginal models.

The first published results on graphical models [5] linked graphical models for categorical data to loglinear models:

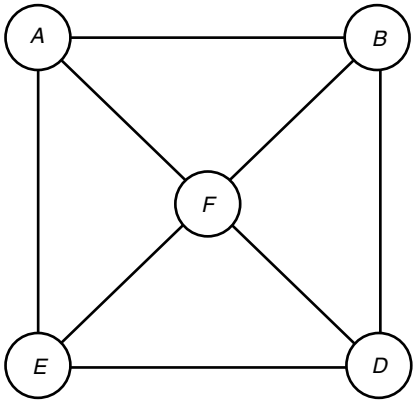

Figure 2 An interaction graph obtained by collapsing the model defined by Figure 1 over variable C

> A graphical model for a multidimensional contingency table without **structural zeros** is loglinear with generators defined by the cliques of the interaction graph.

The result is an immediate result of the fact that any model for a multidimensional contingency table has a loglinear expansion. Starting with the saturated model, one removes all loglinear terms containing two variables assumed to be conditional independent. The loglinear terms remaining after all the terms relating to one or more of the independence assumptions of the model have been deleted define a hierarchial loglinear model with parameters corresponding to each of the completely connected subsets of nodes in the graph.

The interaction graph for the model shown in Figure 1 has four cliques, $BCDF$, ABF, AEF, and DEF, corresponding to a loglinear model defined by one four-factor interaction and three three-factor interactions.

Separation and Parametric Collapsibility

While conceptually very simple, graphical models are usually complex in terms of loglinear structure. The problems arising from the complicated parametric structure are, however, to some degree to be compensated for by the properties relating to collapsibility of the models.

Parametric collapsibility refers to the situation in which model terms of a complete model are unchanged when the model is collapsed over one or more variables. Necessary conditions implying parametric collapsibility of loglinear models are described by Agresti [1, p. 151] in terms which translate into the language of graphical models:

> Suppose variables of a graphical model of a multidimensional contingency table are divided into three groups. If there are no edges connecting variables the first group with connected components of the subgraph of variables from the third group, then model terms among variables of the first group are unchanged when the model is collapsed over the third group of variables.

Parametric **collapsibility** is connected to separation in two different ways. First, parametric collapsibility gives a simple proof of the separation theorem, because a vanishing two-factor term in the complete model also vanishes in the collapsed model if the second group discussed above contains the separators for the two variables. Secondly, separation properties of the interaction graph may be used to identify marginal models permitting analysis of the relationship between two variables. If one first removes the edge between the two variables, A and B, and secondly identifies separators for A and B in the graph, then the model is seen to be parametric collapsible on to the model containing A and B and the separators with respect to all model terms relating to A and B.

The results are illustrated in Figure 3, where the model shown in Figure 3(a) is collapsed on to marginal models for $ABCD$ and $CDEF$. The separation theorem is illustrated in Figure 3(b). All terms relating to A and B vanish in the complete model. The model satisfies the condition for parametric collapsibility, implying that these parameters also vanish in the collapsed model. The second property for the association between E and F is illustrated in Figure 3(c). C and D separate E and F in the graph from which the EF edge has been removed. It follows, therefore, that E and F cannot be linked to one and the same connected component of the subgraph for the variables over which the table has been collapsed. The model is therefore parametric collapsible on to CDEF with respect to all terms pertaining to E and F.

Decomposition and Reducibility

Parametric collapsibility defines situations in which inference on certain loglinear terms may be performed in marginal tables because these parameters

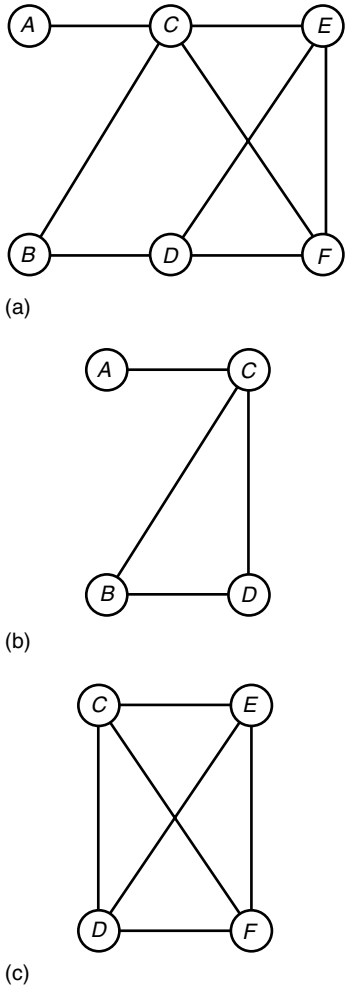

Figure 3 Collapsing the model given in (a) illustrates the separation theorem for A and B (b), and parametric collapsibility with respect to E and F (c)

are unchanged in the marginal tables. Estimates of, and test statistics for, these parameters calculated in the marginal tables will, however, in many cases differ from those obtained from the complete table. Conditions under which calculations give the same results may, however, also be stated in terms of the interaction graphs.

An undirected graph is said to be *reducible* if it partitions into three sets of nodes – X, Y, and Z – if Y separates the nodes of X from those of Z and if the nodes of Y are completely connected. If the interaction graph meets the condition of reducibility, it is said to decompose into two components, $X + Y$ and $Y + Z$. The situation is illustrated in Figure 4, which decomposes into two components, $ABCD$ and $CDEF$.

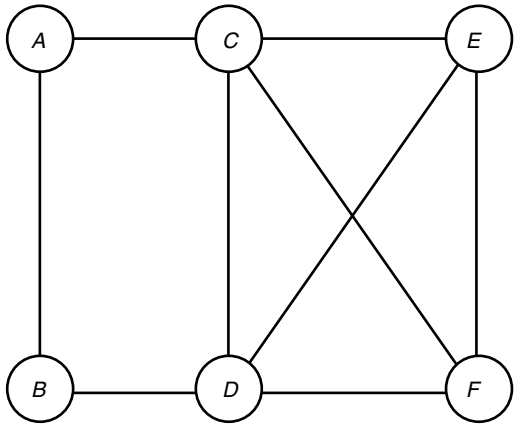

Figure 4 An interaction graph of a reducible model

It is easily seen that reducibility above implies parametric collapsibility with respect to the parameters of X and Z, respectively. It can also be shown, however, that likelihood-based estimates and test statistics obtained by analysis of the collapsed tables are exactly the same as those obtained from the complete table.

Regression Models and Recursive Models

So far, the discussion has focused on models for the joint distribution of variables. The models can, however, without any problems, be extended first to multidimensional regression models describing the conditional distribution of a vector of dependent variables given another vector of **explanatory variables** and, secondly, to block recursive systems of variables. In the first case, the model will be based on independence assumptions relating to either two dependent variables or one dependent and one independent variable. In the second case, recursive models have to be formulated as a product of separate regression models for each recursive block conditionally given variables in all prior blocks. To distinguish between symmetric and asymmetric relationships edges between variables in different recursive blocks, interaction graphs are replaced by arrows.

Parametric Structure: Homogeneous or Heterogeneous Association

The limitations of graphical models for contingency tables lie in the way in which they deal with higher-order interactions. The definition of the graphical models implies that higher-order interactions *may* exist if more than two variables are completely connected.

It is therefore obvious that an analysis of association by graphical models can never be anything but the first step of an analysis of association. The graphical model will be useful in identifying associated variables and marginal models where associations may be studied, but sooner or later one will have to address the question of whether or not these associations are homogeneous across levels defined by other variables and, if not, which variables modify the association. The answer to the question of homogeneity of associations depends on the type of measure that one uses to describe or measure associations. For categorical data, the natural measures of association are measures based on the so-called cross product ratios [2] (*see* **Odds Ratio**). The question therefore reduces to a question of whether or not cross product ratios are constant across different levels of other variables, thus identifying loglinear models as the natural framework within which these problems should be studied.

Ordinal Categorical Variables

In the not unusual case of association between ordinal categorical variables, the same types of argument apply against the hierarchical loglinear models as against the graphical models. Loglinear models are basically interaction models for nominal data; and, as such, they will give results that are too crude and too imprecise for ordinal categorical data. The question of whether or not the association between two variables is homogeneous across levels of conditioning variables can, for ordinal variables, be extended to a question of whether or not the association is homogeneous across the different levels of the associated variables. While not abandoning the basic loglinear association structure, the answer to this question depends on the further parameterization of the loglinear terms of the models. We refer to a recent discussion of these problems by Clogg & Shihadeh [3].

Discussion

The viewpoint taken here on the formulation of interaction models for categorical data first defines the family of graphical models as the basic type of models for association and interaction structure. Loglinear models are, from this viewpoint, regarded as parametric graphical models, meeting certain assumptions on the nature of associations not directly captured by the basic graphical models. Finally, different types of models for ordinal categorical data represent yet further attempts to meet assumptions relating specifically to the ordinal nature of the variables.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Altham, P.M.E. (1970). The measurement of association of rows and columns for an $r \times s$ contingency table, *Journal of the Royal Statistical Society, Series B* **32**, 63–73.

[3] Clogg, C. & Shihadeh, E.S. (1994). *Statistical Models for Ordinal Variables*. Sage, Thousand Oaks.

[4] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies. Models, Analysis and Interpretation*. Chapman & Hall, London.

[5] Darroch, J.N., Lauritzen, S.L. & Speed, T.P. (1980). Markov fields and log-linear models for contingency tables, *Annals of Statistics* **8**, 522–539.

[6] Dawid, A.P. (1979). Conditional independence in statistical theory, *Journal of the Royal Statistical Society, Series B* **41**, 1–15.

[7] Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer-Verlag, New York.

[8] Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press, Oxford.

[9] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

SVEND KREINER

Interaction

Interaction is most often considered in the context of regression models, including the special case of models underlying the **analysis of variance** (ANOVA). In these models, the response variable is linked in some manner to a linear predictor of the form

$$\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k,$$

where the X_is represent **explanatory variables**, and α and the β_is represent parameters to be estimated. Here, for exposition purposes, the X_is will be regarded as representing separate factors of interest, or perhaps functions of a measurement or coding of a single factor. In this case, the linear predictor reflects an additive relationship such that a change in X_i induces the same change in the linear predictor whatever the values of the other explanatory variables.

In this framework, an interaction term is defined by the product of two or more X_is. Consider the special case of two explanatory variables. Then the linear predictor can be expanded and be represented by

$$\nu(X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2.$$

The coefficient β_{12} then represents a departure from an **additive model** for the simultaneous effect of X_1 and X_2 on the response. A test of the hypothesis $\beta_{12} = 0$ is used to examine whether there is evidence for such a departure. Technically, such a test is undertaken as a standard test for a nonzero regression coefficient in the **regression** model being considered.

If X_1 is continuous, then plots of ν against X_1, with X_2 fixed, provide an illustration of interaction effects. In the absence of interaction, the curves are parallel for different values of X_2. If X_2 is also continuous, then parallel curves also arise when ν is plotted against X_2 with X_1 fixed. The nature of the variables may determine the most natural means of presentation. For example, if X_1 represents an experimental treatment level and X_2 a covariate that specifies some intrinsic characteristic of a subject, then it is natural to plot ν against X_1 with X_2 fixed. If X_1 and X_2 are *categorical*, then interaction effects are often displayed by the presentation of values of ν for different values of X_1 and X_2 in a two-way table.

The absence of interaction, when there is particular interest in the effect of both X_1 and X_2 on the response variable, indicates that the separate effects of the two variables are additive. If interest primarily focuses only on X_1, and X_2 is regarded as a covariate, then the lack of interaction indicates that the effect of X_1 is independent of X_2. Particularly in analysis of variance procedures, the interaction of a treatment variable X_1, with a covariate X_2 which varies in a haphazard or largely uncharacterizable way, may be regarded as random variation that may be used in the estimation of the error of treatment contrasts (*see* **Random Effects**).

When the term $\beta_{12} X_1 X_2$ is referred to as an interaction term, terms of the form $\beta_i X_i$ are often referred to as main-effect terms. This derives predominantly from the ANOVA literature, and is particularly relevant to the orthogonal effects that derive from the coding of explanatory variables commonly used there (*see* **Analysis of Variance**). More generally, the interpretation of main-effect terms may depend very critically on the particular representation of the explanatory variables used to define X_i, particularly in the presence of interaction terms.

A distinction is sometimes made between qualitative and quantitative interactions. A qualitative interaction is one in which the direction of the effect of X_1, say, differs depending on the value of X_2. A quantitative interaction would reflect changes in the magnitude of the X_1 effect with X_2, which do not induce a change in the direction of the effect.

Another distinction is between synergistic (*see* **Synergy of Exposure Effects**) and antagonistic interactions. Assume that a change in X_1 induces a change δ_1 in the linear predictor through the term $\beta_1 X_1$, and a change in X_2 similarly induces a change δ_2. If δ_1 and δ_2 have the same sign, say "+", then a synergistic interaction is one which causes the change in the linear predictor due to changes in both X_1 and X_2 to be greater than $\delta_1 + \delta_2$. In contrast, an antagonistic interaction will result in a change less than the sum of the individual effects.

Interactions are always defined in terms of a specific model. Another model which is defined with a transformation of the response variable or a different relationship between the response variable and the linear predictor will not necessarily manifest the same interactions. Some formal attention has been paid to defining "removable interactions", but it is

probably best to consider alternative models for this purpose on a case by case basis.

For models with more than two factors, products of all pairs of variables can be considered, and would be termed second-order or two-way interactions. In the obvious way, interactions of order m can be defined by introducing a product of m variables. When factors are defined with a set of binary **dummy variables**, interactions between factors involve products of these dummy variables, and the set of cross products corresponding to a pair of factors can be regarded as a single interaction term with **degrees of freedom** corresponding to the number of nonlinearly dependent cross products that can be defined. For factors with I and J levels, the degrees of freedom would be $(I - 1)(J - 1)$.

It has been argued that any model with an interaction term must have all main effects corresponding to terms in the interaction in the model. Such an approach produces what are called **hierarchical models**. While there are examples in which this requirement is viewed as too strong, it is in almost all situations sensible formally to test for nonzero interaction effects in the presence of the main effects.

A comprehensive review of interaction has been given by Cox [1]. In epidemiology, interaction is closely linked with the term **effect modification**.

Reference

[1] Cox, D.R. (1984). Interaction, *International Statistical Review* **52**, 1–31.

(*See also* **Experimental Design**; **Multiple Linear Regression**)

VERN T. FAREWELL

Interim Analysis of Censored Data

In many chronic disease **clinical trials**, the major endpoint of interest is time to an event, such as time to disease progression or time to death. Often, the focus of the clinical trial is the comparison of time to event among different treatment groups. In such trials, patients enter the study during some **staggered entry** accrual period, and the final analysis is planned after a predetermined follow-up period. Usually, at the final analysis, not all events are observed, giving rise to **censored** survival data.

For ethical as well as practical reasons, these trials are monitored periodically and interim analyses are performed (*see* **Data and Safety Monitoring**). It is now common practice for all large-scale clinical trials to be monitored formally. Independent data-monitoring boards have been established for most large-scale government-sponsored clinical trials, and, increasingly, such monitoring boards are being established for pivotal clinical trials conducted by private industry such as pharmaceutical companies. The role of the data-monitoring board is to serve as an external oversight committee that reviews periodically the data from the trial as they accrue and to advise on the early termination of the trial or modification of the protocol on the basis of the emerging results (*see* **Clinical Trials Protocols**). Reasons for the termination of a trial are complex, and include serious toxicity, unexpected adverse events, design and/or logistical issues too serious to address, such as very low accrual or event rates, external information, established benefit, or no trend of interest. The board also considers carefully issues of data quality and the consistency of results across various endpoints and over time before making their recommendation.

There has been a great deal of statistical research devoted to early termination of a trial, if, during an interim analysis, a sufficiently large or small treatment difference is observed in the primary endpoint. The major question is: How large or small must the treatment difference be during the interim analysis to warrant terminating the trial? To this end, a test statistic is computed at each interim analysis and compared with a stopping boundary. If the test statistic crosses the boundary at an interim analysis, then the trial is terminated; otherwise, the trial continues until the next interim analysis. This process is continued, if necessary, until the time of a planned final analysis, and is referred to as **sequential** testing. We focus our discussion on upper boundaries only; that is, we allow the possibility of stopping the trial only if a sufficiently large treatment difference is observed during an interim analysis. In some settings, both upper and lower boundaries, allowing termination if either a sufficiently large or small treatment difference is observed, may be implemented (*see* **Data and Safety Monitoring**).

Statistical methods for sequential testing have been available for a long time, but only since the early 1980s have they been used routinely in monitoring clinical trials. One reason is that standard sequential methods require that the trial be monitored continually. Although there are many experimental conditions where this is feasible, it is generally not flexible enough to accommodate the needs of most large-scale clinical trials, where, administratively, it is too difficult for the data to be maintained for continual monitoring. Moreover, continual review is not feasible in a system where the data are monitored by an independent board that, of necessity, can meet only, at most, several times a year. The flexible method proposed by Lan & DeMets [10] has proven to be a useful way of monitoring trials that allow the number and timing of interim analyses to be left unspecified. Their method depends on specifying an alpha-spending function that may be translated into stopping boundaries. We discuss this strategy for sequential monitoring in more detail later in the article. Other strategies for monitoring clinical trials that include the use of the triangular test, the truncated sequential probability ratio test, and the restricted procedure are discussed in [25] (*see* **Sequential Analysis**).

To derive sequential tests, we must be able to characterize the joint distribution of the sequentially computed test statistics used with censored survival data. The difficulty is that there are two time axes that must be considered in evaluating the distribution of sequentially computed test statistics. Time to event for individuals is measured from the time they enter the trial, and it is the distribution of these patient times that are compared among treatments. However, sequential monitoring occurs over calendar time, which is measured as time from the start of the study. These issues are considered in more detail

below. Later, we describe how the results for the joint distribution of sequentially computed test statistics for right censored data may be used in conjunction with the flexible methods of Lan & DeMets to construct stopping rules.

Formalization of the Problem

We assume that n individuals enter the trial at calendar times $E_1, \ldots, E_n$. Each individual i has a potential survival time T_i, possibly unobserved, measured from the time of entry (*see* **Survival Analysis, Overview**). The distribution of T_i may depend on a vector of **covariates**, which includes a treatment indicator, denoted by Z_{0i}, and possible additional covariates $\mathbf{Z}_{1i}$. The relationship between survival time and the covariates is often modeled through the hazard function given by

$$\lambda(u|Z_0, \mathbf{Z}_1, \boldsymbol{\beta}) = \lim_{h \to 0} h^{-1} \times \Pr(u \le T < u + h | T \ge u, Z_0, \mathbf{Z}_1),$$

where $\boldsymbol{\beta}$ denotes a vector of parameters that may be finite dimensional (parametric model) or infinite dimensional (**semiparametric** or **nonparametric** models). The main objective is testing the null hypothesis of no treatment effect on the survival distribution (*see* **Hypothesis Testing**). For example, if we consider only treatment indicator Z_0, then the nonparametric null hypothesis may be posed as

$$\mathrm{H}_0\colon \lambda(u|Z_0 = 1) = \lambda(u|Z_0 = 0), \quad u \ge 0. \qquad (1)$$

Parametric or semiparametric models may also be used for this purpose; for example, we may assume the hazard function follows a **proportional hazards** model:

$$\lambda(u|Z_0, \mathbf{Z}_1) = \lambda_0(u) \exp(\beta_0 Z_0 + \boldsymbol{\beta}_1^{\mathrm{T}} \mathbf{Z}_1), \qquad (2)$$

where $\lambda_0(u)$ is some unknown baseline hazard function, $\mathbf{Z}_1$ is a vector of additional covariates, and the null hypothesis is $\mathrm{H}_0 : \beta_0 = 0$.

If an interim analysis is conducted at calendar time t (measured from the start of the study), then individual i will have censored survival data if $T_i > t - E_i$. Censoring may also occur from other random loss-to-follow-up causes. We define V_i to be the potential censoring time due to causes unrelated to the time of an interim analysis. Thus, assume that for individual i there exists a vector of random variables $(E_i, T_i, V_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, some of which are possibly unobserved. At analysis time t, the observable random variables are $\{X_i(t), \delta_i(t), \mathbf{Z}_i\}$, for all $i = 1, \ldots, n$, such that $E_i \le t$. Here, $X_i(t) = \min(T_i, V_i, t - E_i)$ is the observed time-on-study at analysis time t, and $\Delta_i(t) = 1$ if $T_i \le \min(t - E_i, V_i)$, 0 otherwise, denotes the failure indicator at time t. It is important to note that the data available for an individual at different interim analysis times may vary. For example, an individual with censored time-to-event data at time $t[\Delta_i(t) = 0]$ may at some later time t' be uncensored $[\Delta_i(t') = 1]$.

Typically, a test statistic is computed using all the available data at time t. This statistic, which we denote by $W(t)$, is used to test the null hypothesis H_0 of no treatment difference; that is, the null hypothesis is rejected when $W(t)$ or $|W(t)|$ is sufficiently large, depending on whether we are considering one-sided or two-sided alternatives. The most widely used methods for testing the nonparametric null hypothesis given by (1) are the class of *weighted logrank tests*. Special cases of this general class include the **logrank test** [12, 14]. Prentice's [16] generalization of the Wilcoxon test (*see* **Wilcoxon–Mann–Whitney Test**), and the G^ρ tests of Harrington & Fleming [8]. If, instead, the null hypothesis is stated using a parametric or semiparametric model such as (2), then $W(t)$ may be a standard Wald or score test statistic derived from the **likelihood** for a parametric model or from Cox's [4] **partial likelihood** for the semiparametric model (2).

In a sequential time-to-event trial, the study is monitored at interim times $t_1, \ldots, t_K$. At each analysis time, t_j, we compute the test statistic $W(t_j)$, which incorporates all of the information up to the analysis time. If this statistic exceeds the stopping boundary value b_j, i.e.

$$|W(t_j)| \ge b_j,$$

then we may terminate the study and reject the null hypothesis. The boundary values b_j must be chosen in such a way as to preserve the **level of the test**. For example, if we wish to test at level of significance α, then the b_j must satisfy

$$\Pr_{\mathrm{H}_0}\left\{\bigcup_{j=1}^{K} [|W(t_j)| \ge b_j]\right\} = \alpha, \qquad (3)$$

where $\Pr_{H_0}$ denotes probability computed under the null hypothesis. To evaluate probabilities such as those in (3), we require the joint distribution of $[W(t_1), \ldots, W(t_K)]$. The particular challenge in deriving this joint distribution arises from the fact that the data for any individual contributing to the test statistic at different interim times may vary. A **Lexis diagram** is very helpful in explaining the interrelationship of these two time-scales. For an excellent example that illustrates the use of a Lexis diagram, we refer the reader to [9]. Careful consideration of patient time vs. calendar time has allowed derivation of the joint sequential distribution for most test statistics commonly used with right censored data. Some of the main results are as follows.

A random vector has normal independent increments if its joint distribution is the same as a vector of partial sums or independent normal random variables. Tsiatis [20, 21], Slud [18], and Gu & Lai [6] show that a general class of time sequential nonparametric statistics (i.e. weighted logrank statistics) are asymptotically distributed with an independent increments normal structure. The independent increments structure is the basis for most sequential designs and analyses, enabling the immediate application of standard group sequential methods and software to compute probabilities such as those given by (3). Assuming the proportional hazards model of Cox [3] (*see* **Cox Regression Model**), Gu & Ying [7], generalizing the work of Tsiatis [21], Sellke & Siegmund [17], and Tsiatis, et al. [23], showed that the test statistic based on maximizing the partial likelihood [4] also has this independent increment structure.

Recently, Tsiatis et al. [22], under the assumption of a parametric model with a single test parameter of interest (usually corresponding to a treatment difference) and a finite number of **nuisance parameters**, proved that the joint distribution of sequentially computed maximum likelihood estimators, and the joint distribution of sequentially computed score tests, have this independent increments structure.

In summary, most test statistics for right censored data, properly normalized, have a joint asymptotic distribution corresponding to an independent increments multivariate normal random vector with variance proportional to statistical **information**. Here, information refers to the usual notion of Fisher information for parametric models. An extended definition of information for semiparametric and nonparametric models is given by Bickel et al. [2]; this is beyond the scope of this article. One important example worth mentioning is the logrank test. In a randomized trial, the information for the logrank test is proportional to the number of events.

Flexible Sequential Boundaries

We now describe how sequential boundaries $b_1, \ldots, b_K$ may be constructed satisfying (3), using the flexible method proposed by Lan & DeMets [10]. The key to this method is to note that the rejection region given in (3) may be partitioned as follows:

$$
\begin{aligned}
&|W(t_1)| \geq b_1, \text{ or}\\
&|W(t_1)| < b_1, |W(t_2)| \geq b_2, \text{ or}\\
&\cdots\\
&|W(t_1)| < b_1, \ldots, |W(t_{K-1})|\\
&\qquad < b_{K-1}, |W(t_K)| \geq b_K.
\end{aligned}
$$

Denote these mutually exclusive rejection regions as $R_1, \ldots, R_K$. If we define the rejection probabilities γ_j such that $\Pr_{H_0}(R_j) = \gamma_j$, $j = 1, \ldots, K$, then (3) will be satisfied when

$$\sum_{j=1}^{K} \gamma_j = \alpha. \tag{4}$$

If we know the joint distribution of $[W(t_1), \ldots, W(t_K)]$, then for any set of γ_j, $j = 1, \ldots, K$, satisfying (4), we may recursively derive the boundary values b_j, $j = 1, \ldots, K$, so that $\Pr_{H_0}(R_j) = \gamma_j$. This is the method proposed by Slud & Wei [19] to be used with the sequentially computed Gehan–Wilcoxon test [5], (*see* **Nonparametric Methods**).

Lan & DeMets suggest that the rejection probabilities be linked directly to the information available at the different interim analyses through the use of an "α-spending function". The use of information-based methods is discussed by Lan & Zucker [11]. Specifically, define a monotone increasing function $\alpha(\pi)$ for $0 \leq \pi \leq 1$ such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$, where $\pi = \pi(t)$ denotes the proportion of statistical information at an interim analysis time t, with 100% information at the time of a final analysis. If we define MI as the maximum information and $I(t)$ as information at interim analysis time t, then

the proportion of information at t would be $\pi(t) = I(t)/MI$. The rejection probabilities, γ_j, are set equal to $\{\alpha[\pi(t_j)] - \alpha[\pi(t_{j-1})]\}$, where $t_0 = 0$ and $\pi(0) = 0$. By definition, these satisfy (4) and may be used to define stopping boundaries b_j, $j = 1, \ldots, K$. The term α-spending function refers to the fact that the probability of rejecting the null hypothesis using this strategy, by time t_j, if the null hypothesis is true, is $\alpha[\pi(t_j)]$, or, "we have spent $\alpha[\pi(t_j)]$ of the significance level by time t_j". The procedure guarantees that the level of significance will be equal to some prespecified α regardless of the number of interim analyses or the timing of these analyses.

As an example of the application of this method, consider the use of the logrank test in a randomized trial to test for the equality of the survival distribution between two treatments. In this case, the information is proportional to the number of deaths. Let $D(t)$ denote the number of deaths observed until time t and D^* denote the maximum number of deaths that determines the end of the trial. At the first analysis time, t_1, we compute the proportion of information $\pi(t_1) = D(t_1)/D^*$. The first boundary value, b_1, is the solution to

$$\Pr_{H_0}[|W(t_1)| \geq b_1] = \alpha[\pi(t_1)].$$

After this is determined, we compute the observed value of the test statistic. If it exceeds b_1, then we stop and reject H_0; otherwise, we continue to the next monitoring time.

Consider the jth ($j = 2, \ldots, K-1$) analysis time t_j, and suppose that the boundary values $b_1, \ldots, b_{j-1}$ have been computed. At t_j, the proportion of information $\pi(t_j)$ is equal to $D(t_j)/D^*$, and the boundary value b_j solves the following equation:

$$\Pr_{H_0}[|W(t_1)| \leq b_1, \ldots, |W(t_{j-1})| \leq b_{j-1}, \\ |W(t_j)| \geq b_j] = \alpha[\pi(t_j)] - \alpha[\pi(t_{j-1})].$$

The solution is easily computed using the independent increments property of the logrank statistic and the recursive numerical integration **algorithm** of Armitage, et al. [1]. The computations may be carried out using available statistical software such as EAST (Early Stopping, Cytel Corporation). Again, the observed value of the test statistic is determined and compared with the cutoff. If it exceeds b_j, then we stop and reject H_0; otherwise, we continue to the next monitoring time.

If we continue until the final analysis time, then we "use up" the remaining significance level; that is, we compute b_K, where

$$\Pr_{H_0}[|W(t_1)| \leq b_1, \ldots, |W(t_{K-1})| \leq b_{K-1}, \\ |W(t_K)| \geq b_K = \alpha - \alpha[\pi(t_{K-1})].$$

To implement these methods for sequential stopping, we must specify the maximum information and the α-spending function prior to the initiation of the trial. The choice of group sequential stopping rules has received a great deal of attention by many authors, including Pocock [15], O'Brien & Fleming [13], and Wang & Tsiatis [24]. The expected stopping times at various alternatives is the criterion that is most often used for comparing competing group sequential tests with the same significance level and **power**. Because α-spending functions and stopping boundaries have a one-to-one relationship, results on the choice of stopping boundaries may be used to determine the choice of the α-spending functions. Space limitations preclude detailed discussion of these issues; we note that two common α-spending functions have received considerable attention in the literature. These functions correspond to what are referred to as the O'Brien–Fleming boundary and the Pocock boundary, respectively [10]. The former function tends to be very conservative at the early stages of the study, while the latter is more liberal. For an O'Brien–Fleming boundary, we take $\alpha(\pi) = \alpha_1(\pi)$, where

$$\alpha_1(\pi) = 4 - 4\Phi\left(\frac{z_{\alpha/4}}{\sqrt{\pi}}\right),$$

and for a Pocock boundary, we take $\alpha(\pi) = \alpha_2(\pi)$, where

$$\alpha_2(\pi) = \alpha \log[1 + (e-1)\pi].$$

In these formulas, e is a constant whose natural logarithm is equal to one, and $\Phi(\cdot)$ and z_x are the cumulative density function and $1 - x$ **quantile** of a **standard normal** random variable, respectively.

The choice of maximum information (MI) is closely related to the power necessary to detect a clinically important alternative. When designing a clinical trial, the information necessary to detect a clinically important difference, with some predetermined power, when using a test at a specified level of significance, is computed. For example, if we test the null hypothesis (1) using the logrank test with

significance level α for a clinical trial where patients are randomized with probability 0.5 to each of two treatments, then the number of events necessary to detect a treatment difference corresponding to a log hazard ratio of β_0, with power $1 - \eta$, is given by

$$4\left(\frac{z_{\alpha/2} + z_\eta}{\beta_0}\right)^2.$$

However, when the data are monitored at several interim analysis times with the possibility of early stopping, there is a loss of power. Hence, the maximum information must be inflated by a factor that depends on the spending function, the significance level, power, and the number of interim analyses. For example, if we use the O'Brien–Fleming-type spending function with five interim analyses at the 0.05 level and 90% power, then the information must be increased by 3%. In contrast, an increase of 21% would be necessary with a Pocock-type spending function. The results of Wang & Tsiatis [24] may be used to determine the inflation factor as a function of K, α, η, and the type of spending function.

In summary, we have described a class of flexible and comprehensive methods for developing stopping rules for clinical trials with censored data, which may be used with parametric, semiparametric, and nonparametric models. These methods are used commonly by **data and safety monitoring boards**, as they guarantee the preservation of the significance level and power of the test while still allowing for early termination at interim times that do not have to be specified in advance.

References

[1] Armitage, P., McPherson, C.K. & Rowe, B.C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

[2] Bickel, P.J., Klaassen, C.A., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[3] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[5] Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples, *Biometrika* **52**, 203–223.

[6] Gu, M. & Lai, T. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials, *Annals of Statistics* **19**, 1403–1433.

[7] Gu, M. & Ying, Z. (1995). Group sequential methods for survival data using partial score processes with covariate adjustment, *Statistica Sinica* **5**, 793–804.

[8] Harrington, D.P. & Fleming, T.R. (1982). A class of rank test procedures for censored survival data, *Biometrika* **69**, 553–566.

[9] Keiding, N., Bayer, T. & Watt-Boolsen, S. (1987). Confirmatory analysis of survival data using left truncation of the life times of primary survivors, *Statistics in Medicine* **6**, 939–944.

[10] Lan, G.K.K. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.

[11] Lan, G.K.K. & Zucker, D.M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion, *Statistics in Medicine* **12**, 753–765.

[12] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports* **50**, 163–170.

[13] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[14] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society, Series A* **135**, 185–206.

[15] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.

[16] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika* **65**, 167–179.

[17] Sellke, T. & Siegmund, D. (1983). Sequential analysis of the proportional hazards model, *Biometrika* **70**, 315–326.

[18] Slud, E. (1984). Sequential linear rank statistics for two sample censored survival data, *Annals of Statistics* **12**, 551–571.

[19] Slud, E. & Wei, L.J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic, *Journal of the American Statistical Association* **77**, 862–868.

[20] Tsiatis, A.A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time, *Biometrika* **68**, 311–315.

[21] Tsiatis, A.A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis, *Journal of the American Statistical Association* **77**, 855–861.

[22] Tsiatis, A.A., Boucher, H. & Kim, K. (1995). Sequential methods for parametric survival models, *Biometrika* **82**, 165–173.

[23] Tsiatis, A.A., Rosner, G.L. & Tritchler, D.L. (1985). Group sequential tests with censored survival data adjusting for covariates, *Biometrika* **72**, 365–373.

[24] Wang, S.K. & Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials, *Biometrics* **43**, 193–199.

[25] Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*, 2nd Ed. Ellis Horwood, New York.

(*See also* **Sample Size Determination in Survival Analysis**)

A.A. TSIATIS

International Agency for Research Against Cancer (IARC)

The International Agency for Research on Cancer (IARC) was established in May 1965, through a resolution of the XVIIth World Health Assembly as an extension of the **World Health Organization** after a French initiative. IARCs founding members were the Federal Republic of Germany, France, Italy, the UK, and the US. The Agency's headquarters' building was provided by its host, and is located in Lyon, France. Today, IARC's membership has grown to 16 countries (founding states plus Australia, Belgium, Canada, Denmark, the Russian Federation, Finland, Japan, Norway, the Netherlands, Sweden, and Switzerland). IARC activities are mainly funded by the regular budgetary contributions paid by its participating states. Each contribution is according to a formula which shares the first 70% equally amongst all participating states and apportions the remaining 30% depending upon the individual country's GNP.

A major goal of the IARC is the identification of causes of cancer, so that preventive measures may be adopted against them. The Governing Council has repeatedly stated that research dealing with treatment and other aspects of cancer patient care should not be a part of IARC's mission, nor should the Agency be directly involved in the implementation of control measures, except in cases where it is necessary in order to assess the effectiveness of the mechanisms of carcinogenicity, or when the experimental intervention is needed to permit identification of causes. Nor does IARC deal in the formulation of policies or legislation aimed at controlling carcinogens.

The main emphasis of research is on epidemiology, environmental carcinogenesis, and research training. This emphasis reflects: (i) the generally accepted notion that 80% of all cancers are, directly or indirectly, linked to environmental factors, and thus are preventable; (ii) the recent recognition of the fact that epidemiology may play an important part in cancer prevention and in the evaluation of prevention measures; and (iii) the fact that geographic variations in cancer incidence almost certainly reflect differences in the environment and are therefore particularly well suited for international research efforts.

Epidemiologic research is in two main areas: **descriptive epidemiologic** studies show the trends of cancer incidence and mortality in different populations and geographic areas, and **analytic epidemiologic** studies focus on the **associations** between incidence and mortality and specific risk factors (diet, some professional exposures, etc.).

Recent years have seen renewed interest for the study of genetic factors (*see* **Genetic Epidemiology**) and other host factors contributing to cancer. This trend came about after an increasing body of evidence showing that genetic mutations play a critical part in carcinogenesis, because of the potential importance of host factors in the modification of the carcinogenic effect of environmental agents (*see* **Environmental Epidemiology**), and because of the potential usefulness of genetic methods in the identification of people at high cancer risk who could benefit from a specific intervention.

Throughout its existence, IARC has had an active programme in biostatistics, involving several professional biostatisticians. Emphasis has been given not only to the optimal utilization of methods, but also to the development of new methodology in response to the needs of cancer research. Contribution to **case–control** and **cohort study** methodology, evaluation of **screening** programs, long-term animal experiments (*see* **Tumor Incidence Experiments**) and descriptive epidemiology are among the fields in which IARC has made notable methodologic contributions.

IARC publishes several series: the *Scientific Publications* series (154 volumes), the *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* (82 titles and eight supplements), the *Technical Reports* series, the *Directory of Agents being Tested for Carcinogenicity* and a few other nonserial publications. Of particular importance to statisticians are the volumes on statistical methods in cancer research [1–4].

IARC has around 150 staff members at the Agency's Headquarters in Lyon, and welcomes every year an average of over 600 visiting scientists and trainees from over 30 countries.

References

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1: *The Analysis of Case-Control Studies*. IARC, Lyon.

[2] Breslow, N.E. & Day, N.E. (1986). *Statistical Methods in Cancer Research*, Vol. 2: *The Design and Analysis of Cohort Studies*. IARC, Lyon.

[3] Estève, J., Benhamou, E. & Raymond, L. (1994). *Statistical Methods in Cancer Research*, Vol. 4: *Descriptive Epidemiology*. IARC, Lyon.

[4] Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research*, Vol. 3: *The Design and Analysis of Long-Term Animal Experiments*. IARC, Lyon.

JAQUES ESTEVE & DOUGLAS G. ALTMAN

International Biometric Society (IBS)

The International Biometric Society is an

> international society for the advancement of biological science through the development of quantitative theories and the application, development, and dissemination of effective mathematical and statistical techniques. To this end the Society welcomes to membership biologists, mathematicians, and others interested in applying similar techniques.

The Society was founded on September 6, 1947, at the First International Biometric Conference at Woods Hole, Massachusetts, in the US. The first President of the Society was **R.A. Fisher** from Britain and the first Secretary was **Chester I. Bliss**, from the US. The founders of the Society were motivated by the need for an organization that would foster international cooperation in the methodology and applications of statistics to biology. Biological research was defined broadly and included medicine, agronomy, public health, epidemiology, psychometrics, crop forecasting, paleontology, plant and animal husbandry, design of experiments, etc.

Structure of the Society

The Society is comprised of geographically delimited Regions or Groups that operate both independently and in consort with the international parent organization. Each Region or Group has its own set of officers and operates scientific and educational programs within its own geographic areas as well as maintaining an active role in the activities of the parent organization. The Governing Body of the Society is its Council, with members of Council elected by the membership at large. The election procedures ensure that all geographic areas have appropriate representation on the Council. The Society had seven Regions in 1948 and in 1995 had 18 Regions and 17 Groups covering virtually the entire world. The total membership in 1995 was approximately 6300.

Publications of the Society

The Society publishes ***Biometrics*** (Founding Editor **Gertrude M. Cox**, US) a peer-reviewed journal with the general purpose "to promote and extend the use of mathematical and statistical methods in pure and applied biological sciences". Potential authors do not need to be members of the Society to submit articles for publication. *Biometrics* was first published in 1945 as the *Biometrics Bulletin* and is currently published quarterly with special issues from time to time. The *Biometric Bulletin*, first published in 1983 (Founding Editor, Robert O. Kuehl, US) also published quarterly by the Society provides information on Society activities, Regional and Group activities, scientific abstracts from Biometric Society Regional meetings, as well as expository papers on biometric applications in various areas of the world. The Society, in collaboration with the **American Statistical Association** publishes quarterly the *Journal of Agricultural, Biological and Environmental Statistics* (Founding Editor, Dallas E. Johnson, US), which focuses on the methodology and applications of statistics to the named fields. The first issue appeared in 1996.

The journal *Biometrics* is currently published by Blackwell Publishing. The journal is available online through the JSTOR initiative since 2002, with back issues also available in electronic form. The newsletter, *Biometric Bulletin*, has been converted to electronic-only publication, with the exception of the last issue of each year, which also contains the Society's Business Plan and its Strategic Plan.

Society Networks

The Society supports a series of Regional biostatistical networks. The networks provide a linkage between countries with established biostatistical centers and those with little or no expertise. Joint conferences, short courses, individual training and provision of journals and computer **software** are components of the activities that make available biometric design and analysis to the scientists in developing countries. The Society had 19 Regions and 17 Groups in 2004, as well one Network, the Sub-Saharan Network.

International Biometric Society Conferences

Since the first International Biometric Conference (IBC) in 1947 there have been 17 IBCs in countries around the world. The conferences provide

a forum for the presentation of scientific papers, discussion of these papers and interactions with biostatistical colleagues with different types of problems and perspectives. The Regions apply to host the biannual meetings and the conferences attract a large number of attendees from many different countries. The Society celebrated the fiftieth anniversary of its founding at the eighteenth IBC in Amsterdam in 1996. Over 700 people attended from more than 60 countries. Recent International Biometric Conferences were held in Cape Town, South Africa (1998), Berkeley, USA (2000), Freiburg, Germany (2002), Cairns, Australia (2004), and a meeting scheduled for Montreal, Canada in 2006.

For additional information about the International Biometric Society, contact the IBS Business Office, 808 17th Street, NW, Suite 200, Washington, DC, 20006-3910, USA. Tel: 1-202-223-9669; Fax: 1-202-223-9569. The Society now has a website `www.tibs.org`.

JONAS H. ELLENBERG &
GEERT MOLENBERGHS

International Classification of Diseases (ICD)

The International Classification of Diseases (commonly known as the ICD) is a classification system designed to group together similar diseases, injuries, and related health problems to facilitate statistical analysis of these conditions. The classification is designed to have a finite number of categories encompassing the entire range of morbid conditions. A specific disease or condition is given its own separate category title in the classification only when separate identification is warranted because of its frequency of occurrence or importance as a medical or public health concern. However, many category titles in the classification contain groups of separate but usually related morbid conditions. There is a unique place for inclusion into one of the categories for every disease or morbid condition; therefore, a number of residual categories are reserved throughout the classification for those conditions which do not belong under one of the more specific titles. The International Classification of Diseases is a statistical classification, not a nomenclature or extensive list of approved names for morbid conditions; however, the concepts of classification and nomenclature are closely related. Some classifications are so detailed (e.g. in zoology and botany) that they in fact become nomenclatures, but these very detailed classifications often lose their value for statistical purposes.

History and Development of the International Classification of Diseases

Interest in classifying diseases and studying disease patterns is usually traced back to the work of **John Graunt** and his tabulations of causes of death based on the London Bills of Mortality in the seventeenth century. During the eighteenth and early nineteenth centuries, several classifications of diseases were prepared. The first to approach classification of diseases systematically was François Bossier de Lacroix (1706–1777), writing under the name Sauvages, in his treatise, *Nosologia Methodica*. During the same period, the naturalist and physician Carolus Linnaeus (1707–1778) prepared, in addition to his seminal classification of botany, a treatise entitled *Genera Morborum*. By the beginning of the nineteenth century, the disease classification in general use was *Synopsis Nosologiae Methodicae*, prepared by William Cullen (1710–1790) and published in 1785 [1].

When the General Register Office of England and Wales was established in 1837, **William Farr** (1807–1883) was named as its first medical statistician. Farr found the Cullen classification, still in use, to be outdated and not sufficiently useful for statistical summarization. In his annual "Letters", published in the Annual Reports of the Registrar General, Farr urged the adoption of a new, uniform, statistical classification of diseases. He noted that many diseases were denoted by more than one term, some terms were used to describe more than one disease, vague terms were used, and complications were recorded instead of primary diseases [2].

The importance of a uniform statistical classification was recognized at the first International Statistical Congress meeting in Brussels in 1853. The Congress asked Farr and Marc d'Espine of Geneva to prepare an internationally acceptable uniform classification of causes of death. At the next meeting of the Congress in 1855 in Paris, Farr and d'Espine each submitted his own classification and the Congress adopted a compromise list of 139 rubrics; the compromise list reflected Farr's arrangement into five groups: epidemic diseases, constitutional (general) diseases, local diseases arranged according to anatomical site, developmental diseases, and diseases directly resulting from violence. Over the next 30 years, this classification was revised four times but it maintained the general structure proposed by Farr.

In 1891, the **International Statistical Institute**, successor to the International Statistical Congress, charged a committee to prepare a new classification of causes of death. The committee, chaired by Jacques **Bertillon** (1851–1922), submitted its classification to the Institute in 1893, and it was adopted. This Bertillon Classification, as it was called, consisted of 161 rubrics as well as an abridged classification of 44 titles and another of 99 titles. These were based on Farr's principle of distinguishing between general diseases and those localized to a particular organ or anatomical site. The Bertillon Classification received general approval and was put into use by several countries and a number of cities. The 1899 meeting of the Institute passed a resolution acknowledging

the use of this "system of cause of death nomenclature" in all the statistical offices in North America, and some in South America and Europe. The resolution further "insists vigorously that this system of nomenclature be adopted in principle and without revision, by all the statistical institutions of Europe" and "approves. . .the system of decennial revision proposed by the American Public Health Association . . .".

The French Government, as a response to the International Statistical Institute's 1899 resolution, convened in Paris, in 1900, the first International Conference for the Revision of the Bertillon or International List of Causes of Death. This conference adopted a classification consisting of 179 groups and an abridged list of 35 groups, and it reaffirmed the desirability of decennial revisions. Accordingly, the International List of Causes of Death, and its successor classifications, has been revised approximately every 10 years thereafter.

Bertillon continued his leadership in classification matters, and the revisions of 1900, 1910, and 1920 were carried out under his guidance. During the decade following his death in 1922, there was an increasing interest in expanding the classification to accommodate morbidity and other **vital statistics** interests. At the same time, there was recognition of the need to involve other international agencies, particularly the Health Organization of the League of Nations, in future revision activity. To coordinate efforts, an international commission, known as the Mixed Commission, was created with equal representation from the International Statistical Institute and the Health Organization of the League of Nations. This Commission drafted the proposals for the fourth (1929) and fifth (1938) revisions of the International List of Causes of Death.

In 1946, the newly established **World Health Organization** was given the responsibility for the next (sixth) revision of the International List of Causes of Death and to develop an International List of Causes of Morbidity. In 1948, the International Conference for the Sixth Revision of the International Lists of Diseases and Causes of Death met in Paris. The Conference secretariat was the joint responsibility of competent French authorities and the World Health Organization. The Sixth Decennial Revision Conference introduced a new era in international vital and health statistics. In addition to recommending a comprehensive list of conditions for both morbidity and mortality, the *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death*, the Conference agreed on rules for selecting the underlying **cause of death**, a Medical Certificate of Cause of Death form (*see* **Death Certification**), and special lists and guidelines for tabulation. These recommendations were endorsed by the first World Health Assembly in 1948, resulting in World Health Organization Nomenclature Regulations which member countries have agreed to follow.

The International Conference for the Seventh Revision of the International Classification of Diseases was held under WHO auspices in 1955; the Eighth Revision Conference took place in 1965. The seventh revision was limited to a few essential changes and amendments or corrections. The eighth revision, while more extensive than the seventh, still maintained the basic structure of the classification and the general concept of classifying diseases according to etiology rather than manifestation.

The International Conference for the Ninth Revision of the International Classification of Diseases, again convened by WHO, took place in 1975. During the period when the seventh and eighth revisions were in force, there was a growing use of the International Classification of Diseases for indexing hospital records and for other morbidity applications. These expanding uses were recognized in the ninth revision, which added considerable detail and specificity to the classification. Also introduced was an optional method of classifying selected conditions according to their manifestation in a particular organ or site as well as by the underlying general disease. In addition, based on recommendations of the Ninth Revision Conference, the World Health Assembly approved the publication by WHO of two supplementary classifications on a trial basis: one for Impairments, Disabilities, and Handicaps [4] and one for Procedures in Medicine [3] (*see* **Classifications of Medical and Surgical Procedures**). These were to be adjuncts to the International Classification of Diseases, not integral parts of the basic classification.

Planning for the preparation of the tenth revision began even before the publication of the ninth revision. Early on, it was apparent that the expanded uses of the classification and the resultant complexities and additional detail required more than the usual 10-year cycle for this revision. The longer time period would not only allow broad solicitation of input from users and producers of the data but would also permit trials

of some of the major changes being proposed. Therefore, WHO, with the concurrence of member states, postponed the Tenth Revision Conference from 1985 to 1989, with the planned implementation of the tenth revision consequently also delayed.

Characteristics of the Tenth Revision

The formal title of the tenth revision of the International Classification of Diseases (usually referred to as ICD-10) is *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* [6]. It comprises three volumes: Vol. 1 contains the main classifications; Vol. 2 contains guidance and rules for use of the ICD; and Vol. 3 is the alphabetic index.

ICD-10 is a variable-axis classification evolved from the original principles of organization proposed by Farr. It is designed as a three-character code with fourth-character subdivisions where appropriate. A letter is used in the first position and a numerical digit in the second, third, and fourth positions. The fourth character is preceded by a decimal point. Therefore, individual alphanumeric codes range from A*nn.n* to Z*nn.n*, where *n* represents any of the ten digits from 0 to 9. The letter U is not used. The alphanumeric characteristic of ICD-10 codes is an innovation designed to permit more flexibility in maintaining a hierarchical sequence of diseases while adding more detail to the classification; previous revision code numbers were completely numeric. Vol. 1 contains the list of three-character categories and the tabular list of inclusions and four-character subdivisions. The "core" classification is the list of three-character categories representing the level of reporting required for the WHO mortality database and for routine international comparisons. Many countries use the ICD only at this level of detail; further subdivision of disease categories may not be possible given the quality of the original diagnostic data. Both the core classification and the fully detailed tabular list with its fourth-character detail are arranged into 21 main chapters, and chapters into blocks of related conditions headed by an appropriate block title. In the tabular list, but not in the list of three-character categories, inclusion terms are provided under each code number as examples or guides to the intended content of the category. However, the inclusion terms so listed are not intended to be exhaustive for any given category, and the Alphabetic Index (Vol. 3) serves as a much more detailed guide to the correct placement of conditions into ICD categories.

Vol. 1 also contains a separate classification of morphology of neoplasms which may be used in addition to the main ICD codes which usually classify neoplasms only by behavior and site. These morphology codes are the same as those appearing in the adaptation of the International Classification of Diseases called the *International Classification of Diseases for Oncology* (ICD-O) [5]. In addition, Vol. 1 contains key definitions adopted by the World Health Assembly to facilitate international comparisons of data, and special tabulation lists recommended for the uniform statistical summarization and presentation of both morbidity and mortality data based on the International Classification of Diseases.

ICD-10 came into force on January 1, 1993; however, the actual implementation of this revision of the classification in countries around the world did not begin in earnest until 1995 and the next several years thereafter.

References

[1] Knibbs, G.H. (1929). The International Classification of Disease and Causes of Death and its Revision, *Medical Journal of Australia* **1**, 2–12.

[2] Registrar General of England and Wales (1839). *First Annual Report*. Registrar General of England and Wales, London.

[3] World Health Organization (1978). *International Classification of Procedures in Medicine*, Vols 1 and 2. World Health Organization, Geneva.

[4] World Health Organization (1980). *International Classification of Impairments, Disabilities and Handicaps. A Manual of Classification Relating to the Consequences of Disease*. World Health Organization, Geneva.

[5] World Health Organization (1990). *International Classification of Diseases for Oncology*, 2nd Ed. World Health Organization, Geneva.

[6] World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems*, 10th Rev., 3 Vols. World Health Organization, Geneva.

(*See also* **Cause of Death, Automatic Coding**; **Mortality, International Comparisons**)

ROBERT A. ISRAEL

International Society for Clinical Biostatistics (ISCB)

The International Society for Clinical Biostatistics (ISCB) was founded in May 1979 with the aim of stimulating research into the principles and methodology used in the design and analysis of clinical research, to increase the relevance of statistical theory to clinical practice, and to further the communication between statisticians and clinicians. The Society also has the policy to work with other societies and organizations in the advancement of biostatistics and to provide a common forum for clinicians and statisticians through meetings, seminars and publications.

The ISCB is constituted by an executive committee and led by a President. The Executive Committee consists, in addition to the President, of a Vice-President, a Treasurer, a Secretary and up to eight members, the past President, the News Editor and the Webmaster. The Society's Permanent Office is located in Denmark at the following address: ISCB Permanent Office, PO Box 130, DK-3460 Birkerod, Denmark (tel.: +45 4567 2279; fax: +45 7022 1571; email: `office@iscb.info`. The Society has a website: `http://www.iscb.info`.

Scientific Meetings, Courses, and Publications

The ISCB organizes an annual scientific meeting open to anybody with an interest in statistical methodology and applications in the broad field of medical research, including statisticians, clinicians, epidemiologists, and pharmacologists.

Between its foundation in 1979 and 2003, 24 annual meetings have been held (there was no annual meeting in 1981), all but one of them in Europe – the 1997 meeting was in Boston, US. Four of these meetings have been joint meetings with other societies; in 1985 and 1999 with the German Gesellschaft für Medizinische Dokumentation, Informatik und Statistik (GMDS), and in 1991, 1997 and 2003 with the Society for Clinical Trials (SCT).

A special feature of the meeting is a mini-symposium devoted to a particular medical or statistical field. In recent years these have included environmental epidemiology, statistical challenges in pediatric research, cancer genetics, human fertility, and fecundity, emerging issues in clinical trial data monitoring.

The Society does not publish its own journal, but by an arrangement with the publishers, John Wiley & Sons Ltd, reviewed papers from the annual meetings are published in issues of ***Statistics in Medicine***. In addition, a twice/thrice yearly newsletter, *ISCB News*, is published.

The Society also performs an educational role in the sense of organizing courses on particular statistical topics relevant to the application of statistics in medicine (*see* **Teaching Medical Statistics to Statisticians**). These have generally been run in conjunction with annual meetings, either as pre- or postconference activities, with faculties of foremost researchers in their field.

The Society recognizes the political and economic difficulties of some countries and actively promotes and supports the establishment of national groups. The Society organizes courses in these countries, to enhance the development of biostatistics, and has a Conference Awards for Scientists Program for biostatisticians to attend and present papers at the annual meetings.

In order to promote participation of young researchers, the Society has a Student Conference Awards Program for postgraduate students from all over the world to attend and present papers at the annual meetings.

Special Working/Interest Groups

In recent years the Society has developed a number of working groups, called subcommittees, which deal with scientific, regulatory, or organizational issues. Among the first established subcommittees, there are the one on "Statistics in Regulatory Affairs", whose remit is to consider and influence the development of regulatory requirements, guidelines, and other documents concerning the scientific aspects of data collection, management, analysis, and reporting (*see* **Drug Approval and Regulation**); the subcommittee on **fraud,** which ended its activity after the publication of the paper " *The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials"*, by Buyse M., George S.L., Evans S., Geller N., Ranstam J., Scherrer B., Lesaffre E, Murray G., Edler

L., Hutton J., Colton T., Lachenbruch P., Verma B, on behalf of ISCB, on *Statistics in Medicine*, 18(1999), p. 3435-3451. Other subcommittees currently active are those on Education, Dentistry, National Groups, Student Conference Awards, Conference Organising, and Communication. Their terms of reference and members are published on the Society's News and Web site.

Membership

The membership during the mid-1990s has been fairly stable and numbering around 800 members from some 40 countries from around the world, with the majority coming from Europe. Nonmembers attending an annual meeting automatically become members for that year. The annual membership fee in 2004 is 40 Euro.

MARIA GRAZIA VALSECCHI

International Statistical Institute (ISI)

The International Statistical Institute (ISI) was established in 1885 in London, closely following an exploratory contact of statisticians in Paris. Its predecessor organization, the International Statistical Congress (ISC), was started in 1853 in Brussels, under the leadership of the famous Belgian statistician **Adolphe Quetelet**. The ISC remained in existence until 1876 when the German–French rivalries of the time, apparently through an intervention of Chancellor Bismarck, led to the dissolution of this intergovernmental statistical cooperation. The ISI, under the circumstances, was started as a nongovernmental instrument of international statistical cooperation, with heavy reliance on its elected membership, which consists of outstanding statisticians of the world in their personal professional capacity. ISI is considered today the world academy of statisticians.

At the beginning of 2004, ISI had 1884 elected members, distinguished statisticians active in academia, government, and the private sector of about 130 countries. At the same time, 166 persons who are the heads of national and international statistical offices participate as *ex officio* members in ISI. In addition to its elected and *ex officio* members, ISI involves in its Associations (Sections) a substantial number of statisticians who are active in specialized areas of statistics.

These Associations are: the Bernoulli Society for Mathematical Statistics and Probability (1241 members); the International Association for Official Statistics (474 members); the International Association for Statistical Computing (495 members); the International Association of Survey Statisticians (1266 members); and the International Association for Statistical Education (463 members). The Sections of ISI maintain open membership for all interested statisticians; many elected or ex officio ISI members also hold membership in one or more of these specialized Associations. The total of about 5800 association members include those ISI members who hold membership both in these Sections and in the ISI. ISI is incorporated as a not-for-profit institution in the Netherlands (with its Permanent Office, established in 1913, located in Voorburg, a town adjacent to The Hague).

The goal of ISI is the development and improvement of statistical methods, and their application throughout the world, all in the widest sense of the word. The role played by ISI has changed since its inception. In the nineteenth century, for example, the promotion of standardization of statistical methodology in the official statistics of countries was a key task: the acceptance of the first **International Classification of Diseases** at the 1893 Chicago Session of the ISI was a historic step in this regard. Today, the intergovernmental organizations such as the United Nations (UN), its specialized agencies, the European Union, and the Organization for Economic Cooperation and Development (OECD) are the main forums for these types of endeavor. It is convenient to group present-day ISI activities into five areas: (i) conference services, (ii) publications, (iii) research activities, (iv) membership services, and (v) other functions.

In respect of *conferences*, the biennial ISI Sessions are the most outstanding.

The 2003 Session in Berlin was the fifty-fourth such undertaking (during World Wars I and II no Sessions were held). Recent Sessions were held in Seoul (2001), Helsinki (1999), Istanbul (1997), Beijing (1995), and Florence (1993). The number of participants in recent Sessions has reached about 2300 with nearly 285 invited, 728 contributed, and 84 poster papers presented on a wide array of statistical, theoretical, methodological, and application questions. Smaller, and more specialized conferences were held in Szczecin (2003), Cape Town (2002) amongst others, in cooperation with other national and international statistical organizations.

The *publications* of ISI include scientific journals, such as the *International Statistical Review* [3], *Bernoulli* [1], abstracting resources such as *Statistical Theory and Method Abstracts* [4], books such as *The Oxford Dictionary of Statistical Terms* [2], which has been published in several editions (between 1957 and 2003), as well as the *Newsletter of the ISI*, and so on. In addition, there are numerous journals and publications by the five ISI Sections dealing with areas of their specialization.

ISI has been a promoter of *research activities* since its inception.

The *Bulletins of the ISI* go back to the nineteenth century: these volumes have been issued after each of the ISI Sessions and printed in the host country where the Sessions were held. These “Bulletins”

are a repository of, and a testimonial to, the manifold research efforts undertaken by statisticians in numerous countries over the last 150 years. Later, in connection with the "World Fertility Survey" in the 1970s and 1980s, ISI set up an internal research facility regarding population issues, albeit budgetary restrictions by the end of the 1980s made this venture financially unsustainable. The research function of ISI, however, has been maintained. Today, it involves holding conferences on acute methodological, theoretical, or topical matters such as the statistical issues of derivatives trading, the index numbers of stock markets, or the demographic crisis of the transition countries. It also involves projects at the ISI Permanent Office such as the multilingual glossary of statistical terms and historical statistical investigations and commemorations. Moreover, the five ISI Sections are involved in a wide range of similar activities.

The *membership services* of ISI are primarily administrative in nature and result in the publication of the *Directory of ISI*, which lists all ISI and ISI Section members as well as containing a listing of national and international statistical organizations and societies.

Among the *other functions* of ISI mention should be made of the site (home page) maintained on the **internet** (`http://www.cbs.nl/isi`). Also, every second year (at the time of the world-wide Session) ISI awards the "Jan Tinbergen Prize" to the three most deserving statistical studies submitted by young statisticians from developing countries. Each winner receives 2269 Euros, transportation to, and free stay at the Session, and an opportunity to present their winning papers.

ISI also attempts to promote cooperation with statisticians active in other, primarily nonstatistical, organizations dealing with biometrics, econometrics, psychometrics, astronomy, classification science, linguistics, and so on. It is believed that in addition to the studies emanating from specialized and subspecialized areas within statistics identified as such, there are significant intellectual and practical gains to be made by fostering more integration with the rather dispersed statistical professionals active in all other fields. The Sessions of ISI, therefore, are being opened up for meetings with the "sister organizations" active in statistics.

References

[1] *Bernoulli*, McCullagh, P. ed. Published every two months. ISI, Voorburg, The Netherlands.

[2] Dodge, Y. ed. (2003). *The Oxford Dictionary of Statistical Terms*, 6th Ed., Oxford University Press, Oxford.

[3] *International Statistical Review*, Seneta, E. & Manninen, A. eds. Published three times per year in April, August, and December. ISI, Voorburg, The Netherlands.

[4] *Statistical Theory and Method Abstracts*, van Eeden, C., van Harn, K. & van Es, B. eds. Published twice a year (CD ROM) and online. ISI, Voorburg, The Netherlands.

Further Reading

Newsletter of the International Statistical Institute, Berze, D. & Mehta, S. eds. Published every four months. ISI, Voorburg, The Netherlands.

Z. KENESSEY

International Studies of Infarct Survival (ISIS)

The ISIS began in 1981 as a collaborative worldwide effort to evaluate the effects of several widely available and practical treatments for acute myocardial infarction (MI). The ISIS Collaborative Group randomized more than 134 000 patients into four large simple trials assessing the independent and synergistic effects of beta-blockers, thrombolytics, aspirin, heparin, converting enzyme inhibitors, oral nitrates, and magnesium in the treatment of evolving myocardial acute infarction (Table 1). More than 20 countries participated in these trials, which were coordinated worldwide by investigators in Oxford, England.

ISIS-1: Atenolol in Acute MI [1]

Beta-blocking agents reduce the heart rate and blood pressure, as well as their product, inhibit the effects of catecholamines, and increase thresholds for ventricular fibrillation. Thus, it is not surprising that beta-blockers were among the first agents to be evaluated in randomized trials of evolving acute MI. Even by 1981, the available trials of beta-blocking agents for acute infarction were too small to demonstrate a significant benefit. However, based on an overview of the available evidence (*see* **Meta-analysis of Clinical Trials**), it was judged that the prevention of even one death per 200 patients treated with beta-blockers (*see* **Number Needed to Treat (NNT)**) would represent a worthwhile addition to usual care. Unfortunately, detecting such an effect would require the **randomization** of over 15 000 patients. It was toward this end that the First International Study of Infarct Survival (ISIS-1) trial was formed.

In a collaborative effort involving 245 coronary care units in 11 countries, the ISIS-1 trial randomized 16 027 patients with suspected acute MI to a regimen of intravenous atenolol versus no beta-blocker therapy. Patients assigned to active treatment received an immediate intravenous injection of 5–10 mg atenolol, followed by 100 mg/day orally for seven days. Similar agents were avoided in those assigned at random to no beta-blocker therapy unless it was believed to be medically indicated. As in the subsequent ISIS collaborations, all other treatment decisions were at the discretion of the responsible physician.

During the seven-day treatment period in which atenolol was given, vascular mortality was significantly lower in the treated group (3.89% vs. 4.57%, $P < 0.04$), representing a 15% mortality reduction. Almost all of the apparent benefit was observed in days 0 to 1 during which time there were 121 deaths in the atenolol group as compared with 171 deaths in the control group. The early mortality benefit attributable to atenolol was maintained at 14 days and at the end of one year follow-up (10.7% atenolol vs. 12.0% control). Treatment did not appear to decrease infarct size substantially, although the ability of a large and simple trial such as ISIS-1 to assess such a reduction was limited. Despite its large size, the 95% confidence limits of the risk reductions associated with atenolol in ISIS-1 were wide and included relative risk reductions between 1% and 25%. However, an overview that included ISIS-1 and 27 smaller completed trials of beta-blockade suggested a similar sized mortality reduction (14%). When a combined endpoint of mortality, nonfatal cardiac arrest and nonfatal reinfarction was considered from all available trials, the 10%–15% reduction persisted with far narrower confidence limits. Taken together, these data suggest that early treatment of 200 acute MI patients with beta-blocker therapy

Table 1 The International Studies of Infarct Survival (ISIS)

Trial	Year completed	Agents studied	Patients randomized
ISIS-1	1985	Atenolol vs. control	16 027
ISIS-2	1988	Streptokinase vs. placebo Aspirin vs. placebo	17 187
ISIS-3	1991	Streptokinase vs. tPA vs. APSAC Aspirin + SC heparin vs. aspirin	41 299
ISIS-4	1993	Captopril vs. placebo oral mononitrate vs. placebo Magnesium vs. control	58 050

would lead to avoidance of one reinfarction, one cardiac arrest, and one death during the initial seven-day period. Unfortunately, beta-blocker use in the setting of acute MI remains suboptimal with utilization rates ranging between 30% in the US to less than 5% in the UK. This underutilization appears related in part to poor physician education. In the GUSTO-1 trial, beta-blockers were encouraged by the study protocol and almost 50% of all patients received the drugs without any apparent increase in adverse effects.

ISIS-2: Streptokinase and Aspirin in Acute MI [2]

As with beta-blockers, data from randomized trials of thrombolytic therapy completed prior to 1985 did not yield truly reliable results. Indeed, the largest of the early studies enrolled 750 patients, a totally inadequate sample size to detect the most plausible 20%–25% reduction in mortality.

Given this situation, the Second International Study of lnfarct Survival (ISIS-2) was designed to test directly in a single randomized, double-blind, placebo-controlled trial (*see* **Blinding or Masking**) the risks and benefits of streptokinase and aspirin in acute MI. To accomplish this goal, the ISIS-2 collaborative group randomized 17 187 patients presenting within 24 hours of symptom onset using a 2×2 **factorial design** to one of four treatment groups: 1.5 million units of intravenous streptokinase over 60 minutes; 162.5 mg/day of oral aspirin for 30 days; both active treatments; or neither.

In brief, the primary endpoint (*see* **Outcome Measures in Clinical Trials**) of the trial, total vascular mortality, was reduced 25% by streptokinase alone (95% Cl, −32 to −18, $P < 0.0001$) and 23% by aspirin alone (95% CI, −30% to −15%, $P < 0.00001$). Patients allocated to both agents had a 42% reduction in vascular mortality (95% Cl, −50 to −34, $P < 0.00001$), indicating that the effects of streptokinase and aspirin are largely additive. When treatment was initiated within six hours of the onset of symptoms, the reduction in total vascular mortality was 30% for streptokinase, 23% for aspirin, and 53% for both active agents.

For aspirin, the mortality benefit was similar when the drug was started 0–4 hours (25%), 5–12 hours (21%), or 13–24 hours (21%) after the onset of clinical symptoms. Aspirin use also resulted in highly significant reductions for nonfatal reinfarction (49%) and nonfatal stroke (46%). As regards side-effects, for bleeds requiring transfusion, there was no significant difference between the aspirin and placebo groups (0.4% vs. 0.4%), although there was a small absolute increase of minor bleeds among those allocated to aspirin (0.6%, $P < 0.01$). For cerebral hemorrhage, there was no difference between the aspirin and placebo groups.

For streptokinase, those randomized within four hours of pain onset experienced the greatest mortality reduction, although statistically significant benefits were present for patients randomized throughout the 24 hour period. As expected, there was an excess of confirmed cerebral hemorrhage with streptokinase (7 events vs. 0; $2P < 0.02$), all of which occurred within one day of randomization. Reinfarction was slightly more common among those assigned streptokinase alone, but this difference was not statistically significant. Furthermore, aspirin abolished the excess reinfarction attributable to streptokinase.

In addition to demonstrating the independent as well as synergistic effects of streptokinase and aspirin, ISIS-2 also supplied important information concerning which patients to treat. Because the ISIS-2 entry criteria were broad, the trial included the elderly, patients with left bundle branch block, and those with inferior as well as anterior infarctions. In each of these subgroups, clear mortality reductions were demonstrated.

Thus, in addition to changing radically the premise that thrombolysis should be avoided in patients already on aspirin, the ISIS-2 trial was largely responsible for widening the **eligibility criteria** for patients who would benefit from thrombolytic therapy.

ISIS-3: Streptokinase vs. APSAC vs. tPA and Subcutaneous Heparin vs. No Heparin in Acute MI [3]

While ISIS-2 (streptokinase), the first Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico (GISSI-1, streptokinase), the APSAC Intervention Mortality Study (AIMS, anisoylated plasminogen streptokinase activator complex [APSAC]), Anglo–Scandinavian Study of Early Thrombolysis (ASSET, tissue plasminogen activator [tPA]) and ISIS-2 all documented clear mortality benefits for

thrombolysis, they did not provide information that allowed for directly comparing these agents. It was also unclear whether patients given aspirin would further benefit from the addition of heparin. These questions were the focus of the Third International Study of Infarct Survival (ISIS-3).

In brief, the ISIS-3 collaborative group randomized 41 299 patients to streptokinase, APSAC, and tPA. Patients presenting within 24 hours of the onset of evolving acute MI and with no clear contraindication to thrombolysis were assigned randomly to IV streptokinase (1.5 MU over one hour), IV tPA (duteplase, 0.50 million U/kg over four hours), or IV APSAC (30 U over three minutes). All patients received daily aspirin (162.5 mg), with the first dose crushed or chewed in order to achieve a rapid clinical antithrombotic effect. In addition, half were randomly assigned to receive subcutaneous heparin (12 500 IU twice daily for seven days), beginning four hours after randomization.

ISIS-3 demonstrated no differences in mortality between the three thrombolytic agents. Specifically, among the 13 780 patients randomized to streptokinase, there were 1455 deaths (10.5%) within the initial 35-day follow-up period as compared with 1448 deaths (10.6%) among the 13 773 patients randomized to APSAC and 1418 deaths (10.3%) among the 13 746 randomized to tPA.

Long-term survival was also virtually identical for the three agents at both three and six months. With regard to in-hospital clinical events, cardiac rupture, cardiogenic shock, heart failure requiring treatment, and ventricular fibrillation were similar for the three agents. For nonfatal reinfarction, there was a reduction with tPA, while streptokinase and APSAC allocated patients had higher rates of allergy and hypotension requiring treatment. Streptokinase produced fewer noncerebral bleeds than either APSAC or tPA.

While there were no major differences between thrombolytic agents in terms of lives saved or serious in-hospital clinical events, significant differences were found in ISIS-3 for rates of total stroke and cerebral hemorrhage. Specifically, there were 141 total strokes in the streptokinase group as compared with 172 and 188 in the APSAC and tPA groups, respectively. For cerebral hemorrhage there were 32 events (two per 1000) in the streptokinase group as compared with 75 (five per 1000) in the APSAC group and 89 (seven per 1000) in the tPA group. While the absolute rates for cerebral hemorrhage for all three agents was low, this apparent advantage for streptokinase was highly statistically significant ($P < 0.0001$ for streptokinase vs. APSAC, $P < 0.00001$ for streptokinase vs. tPA).

With regard to the addition of delayed subcutaneous heparin to thrombolytics there was no reduction in the prespecified endpoint of 35 day mortality. During the scheduled seven day period of heparin use, there were slightly fewer deaths in the aspirin plus heparin group compared with the aspirin group alone, a difference of borderline significance. There was, however, a small but significant excess of strokes deemed definite or probable cerebral hemorrhages among those allocated aspirin plus heparin (0.56% vs. 0.40%, $P < 0.05$). In contrast, reinfarction was more common among those randomized to aspirin alone as compared with those receiving aspirin plus subcutaneous heparin.

ISIS-4: Angiotensin Converting Enzyme Inhibition, Nitrate Therapy, and Magnesium in Acute MI [4]

In 1991 the ISIS collaboration chose to investigate several other promising but unproven approaches to the treatment of acute MI. Specifically, the Fourth International Study of Infarct Survival (ISIS-4) sought to examine treatment strategies that would benefit both high- and low-risk patients presenting with acute MI, not simply those who are eligible for thrombolysis.

To attain this goal, the ISIS collaborative group chose to study three promising agents: a twice daily dose of the angiotensin converting enzyme (ACE) inhibitor captopril for 30 days, a once daily dose of controlled release mononitrate for 30 days, and a 24-hour infusion of intravenous magnesium. As was true in each of the preceding ISIS trials, the available data were far too limited to allow reliable clinical recommendations concerning these therapies. For example, while ACE inhibiting agents had been shown to be successful in reducing mortality in patients with congestive heart failure and in patients a week or two past acute infarction, it was unclear whether these agents provided a net benefit for all patients in the setting of evolving acute MI. Similarly, while nitrates were often used in evolving MI because of their ability to reduce myocardial

afterload and potentially limit infarct size, barely 3000 patients had received intravenous nitroglycerin in randomized trials and even fewer patients had been studied on oral nitrate preparations. Finally, because of its effects on calcium regulation, arrhythmia thresholds, and tissue preservation, magnesium therapy had often been considered as an adjunctive therapy for acute infarction even though no data from a randomized trial of even modest size had been available.

Based on statistical overviews, the ISIS investigators estimated that each of these therapies had the potential to reduce mortality in acute infarction by as much as 15%–20%. However, because many patients presenting with acute infarction were treated with thrombolytic therapy and aspirin, the estimated mortality rates at one month were estimated to be as low as 7%–8%. Thus, to assess reliably whether these potentially important clinical effects were real required the randomization of a very large number of patients, perhaps as many as 60 000. To achieve this goal, a $2 \times 2 \times 2$ factorial design was employed in which patients were randomized first to captopril or captopril placebo, then to mononitrate or mononitrate placebo, and then to magnesium or magnesium control. Thus, it was possible in the trial for any given patient to receive all three active agents, no active agents, or any combination.

Captopril

Use of the ACE inhibitor captopril was associated with a significant 7% decrease in five-week mortality (2088 [7.19%] deaths among patients assigned to captopril vs. 2231 [7.69%] deaths among those assigned to placebo), which corresponds to an absolute difference of 4.9 ± 2.2 fewer deaths per 1000 patients treated for one month. The absolute benefits appeared to be larger (possibly as high as 10 fewer deaths per 1000) in some higher-risk groups, such as those presenting with heart failure or a history of MI. The survival advantage appeared to be maintained at 12 months. In terms of side-effects, captopril produced no excess of deaths on days 0–1, even among patients with low blood pressure at entry. It was associated with an increase of 52 patients per 1000 in hypotension considered severe enough to require termination of study treatment, of five per 1000 in reported cardiogenic shock, and of five per 1000 in some degree of renal dysfunction.

Mononitrate

Use of mononitrate was not associated with any significant improvements in outcomes. There was no significant reduction in overall five-week mortality, nor were there reductions in any subgroup examined (including those not receiving short-term nonstudy intravenous or oral nitrates at entry). Continued follow-up did not indicate any later survival advantage. Somewhat fewer deaths on days 0–1 were reported among individuals allocated to active treatment, which is reassuring about the safety of using nitrates early in evolving acute MI. The only significant side-effect of the mononitrate regimen was an increase in hypotension of 15 per 1000 patients.

Magnesium

As with mononitrate, use of magnesium was not associated with any significant improvements in outcomes, either in the entire group or any subgroups examined (including those treated early or late after symptom onset or in the presence or absence of fibrinolytic or antiplatelet therapies, or those at high risk of death). Further follow-up did not indicate any later survival advantage. In contrast to some previous small trials, there was a significant excess of heart failure with magnesium of 12 patients per 1000, as well as an increase of cardiogenic shock of five patients per 1000 during or just after the infusion period. Magnesium did not appear to have a net adverse effect on mortality on days 0–1. In terms of side-effects, magnesium was associated with an increase of 11 patients per 1000 in hypotension considered severe enough to require termination of the study treatment, of three patients per 1000 in bradycardia, and of three patients per 1000 in a cutaneous flushing or burning sensation.

Because of its size, ISIS-4 provided reliable evidence about the effects of adding each of these three treatments to established treatments for acute MI. Collectively, GISSI-3, several smaller studies, and ISIS-4 have demonstrated that, for a wide range of patients without clear contraindications, ACE inhibitor therapy begun early in evolving acute MI prevents about five deaths per 1000 in the first month, with somewhat greater benefits in higher-risk patients. The benefit from one month of ACE inhibitor therapy persists for at least the first year. Oral nitrate

therapy, while safe, does not appear to produce a clear reduction in one-month mortality. Finally, intravenous magnesium was ineffective at reducing one-month mortality.

Conclusion

Because of their simplicity, large size, and strict use of mortality as the primary endpoint, the ISIS trials have played a critical substantive role in establishing rational treatment plans for patients with acute MI. Methodologically, they have clearly demonstrated the utility of large simple randomized trials.

Three principles guided the design and conduct of the ISIS trials. The first was the belief that a substantial public health benefit would result from the identification of effective, widely practical treatment regimens that could be employed in almost all medical settings, as opposed to those that can be administered only at specialized tertiary care facilities. For this reason, the ISIS investigations focused on strategies to decrease mortality which, in and of themselves, did not require cardiac catheterization or other invasive procedures for either diagnostic or therapeutic purposes.

The second principle was that the benefits of truly effective therapies would be applicable to a wide spectrum of patients with diverse clinical presentations. Thus, the entry criteria for the ISIS trials were intentionally broad and designed to mimic the reality all health care providers encounter when deciding whether or not to initiate a given treatment plan. This is one reason that the ISIS trials focused on evolving acute MI in the view of the responsible physician.

The third and perhaps most important principle was that most new therapies confer small to moderate benefits, on the order of 10%–30%. While such benefits on mortality are clinically very meaningful, these effects can be detected reliably only by randomized trials involving some tens of thousands of patients. Thus, the ISIS protocols were streamlined to maximize randomization and minimize interference with the responsible physician's choice of nonprotocol therapies and interventions. Nonetheless, by selectively collecting the most important entry and follow-up variables that relate directly to the efficacy or adverse effects of the treatment in question, the ISIS trials yielded reliable data for providing a rational basis for patient care. By limiting paperwork and not mandating protocol-driven interventions, the ISIS approach proved to be remarkably **cost-effective**. Indeed, the large ISIS trials were conducted at a small fraction of the usual cost of other smaller trials which, because of their inadequate sample sizes, failed to demonstrate either statistically significant effects or informative null results.

References

[1] ISIS-1 (First International Study of lnfarct Survival) Collaborative Group (1986). Randomised trial of intravenous atenolol among 16027 cases of suspected acute myocardial infarction: ISIS-1, *Lancet* **2**, 57–65.

[2] ISIS-2 (Second International Study of lnfarct Survival) Collaborative Group (1998). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2, *Lancet* **2**, 349–360.

[3] ISIS-3 (Third International Study of lnfarct Survival) Collaborative Group (1992). ISIS-3: Randomised comparison of streptokinase vs. tissue plasminogen activator vs. anistreplase and of aspirin plus heparin vs. aspirin alone among 41299 cases of suspected acute myocardial infarction: ISIS-3. *Lancet* **339**, 75–70.

[4] ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group (1995). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction: ISIS-4. *Lancet* **345**, 669–685.

CHARLES H. HENNEKENS & P.J. SKERRETT

Internet

Introduction

The internet (lower case i) is the world's biggest computer network connecting millions of machines worldwide. It offers exciting new ways for people to communicate with each other and new ways to disseminate and access information. The Internet (upper case I), which is a term used to describe what can be done over the internet, has been described as potentially the most exciting and revolutionary development in information since Caxton's printing press. It opens up whole new vistas for academics, the business community, and the general public, providing easier access to information, faster means of communication and exchange of ideas, and new ways of using leisure time.

The internet began with the military in the late 1960s. The Advanced Research Projects Agency (ARPA), a branch of the American defence department, sought a way of exchanging military research information between sites. One essential criterion was that the network had to be able to survive a nuclear war. If one computer in the network was destroyed, then the information would simply take another route. This led to a computer network known as Advanced Research Projects Agency Network (ARPANET). From its military origins, it grew considerably as more US government departments and agencies gained access to the network. In 1983, the military network moved to a separate, more secure system and in 1984, America's National Science Foundation (NSF) created the NSFNET, which linked supercomputers together to allow access by any US educational establishment, irrespective of location.

The ARPANET and NSFNET networks laid the foundations for the wide area network of computers that span the globe and which is now known as the internet. Essentially, it is a network of networks. Each country has many networks for educational, government, and commercial purposes. Each of these separate networks, such as AARNET in Australia, NSFNET in the United States and JANET in the United Kingdom, is an entity in itself. There are also networks operated by commercial Internet Service Providers, who provide access to homes and businesses throughout the world. Each of these networks is made up of smaller, local networks. This whole collection of networks interconnects across the world, allowing each site access to every other site, irrespective of the starting point. This vast interconnecting network forms what is referred to as the internet.

The interconnection of computers to form the internet is based on TCP/IP protocols that allow computers of different types, running different operating systems to communicate with each other. Each computer on the internet is identified by its IP number, which is a hierarchical number similar in nature to a telephone number with country, area, and district codes. However, there are also names for the machines, which are easier to remember, and when a name is known, it can be looked up in a Domain Name Service (DNS).

In the same way that ARPANET passed its information through any available route, depending on which machines were available, information can be routed across the internet by any available path. This makes the whole system extremely robust to failures of individual machines or subnetworks. The machines and networks that make up the internet change from day to day and year to year, but that is not important, as it was designed from its earliest beginnings to be robust to such evolution.

Types of Use

Running over the physical structure of the internet are many different types of services, in the same way that telephone companies supply many types of services over telephone lines. This array of services is generally termed the Internet (upper case I). The Internet can be used in a variety of ways, which see exciting new developments every few months. The principal types of activity are communication between individual people or within a group, gaining access to information, and operating **software**, which is held on a remote computer. These main areas are outlined below.

Communication

Electronic mail (often referred to as "e-mail") is a form of communication where text entered by an individual into one computer system is then sent in electronic form to another, where it can be read by the intended recipient. This type of communication can take place through any system that is connected

to the internet and which has software installed for sending and receiving e-mail. The messages are not transferred around the world instantaneously, but are delivered in seconds rather than days. This has made e-mail a popular form of communication.

In addition to messages between individuals, it is also possible for one person to send a single message to a group of others. Messages sent to a particular Internet address are then automatically broadcast to a list of participants of a "discussion group". Discussion groups of this kind exist for an enormous number of different interests.

Transfer of messages to the individual mailboxes of people interested in a particular subject can be avoided by conducting the discussion through UseNet News groups, where all the messages are brought together on electronic bulletin boards and accessed by interested parties at their convenience.

Again, using text exchanges, technologies such as Internet Relay Chat (IRC) and Talker services enable people to conduct discussions where the text typed by each individual is seen by all the others involved, in real time. Each day there are many people holding discussions, limited only by the speed at which they can type.

As the bandwidth of the internet increases, so the possibility of conducting real-time audio conversations, with full motion video, becomes a reality. This is currently not possible everywhere, but it is spreading rapidly. Again, this need not be confined to one-to-one exchanges. Linking many people together allows computer conferencing to take place, with the consequent reduction in the need for people to travel the world to meet face-to-face for discussion.

File Transfer

When information is transferred in message form, it is usually also possible to attach other files to the main message. However, computer files of any type, such as programs, word processor files, data, or text, can also be transferred from one system to another over the internet. This method of transfer uses the File Transfer Protocol (FTP). This facility is very convenient for a variety of purposes and it is particularly effective for large files. It also forms the backbone of many of the systems for gaining access to information.

Information Searching and Browsing

The Internet has provided a means by which a vast amount of information can be mounted in electronic form and accessed by an enormous number of people. The variety and extent of information now available is staggering. Methods of access to the information are independent of the type of computer a user has, which takes us closer to universal access to information. One of the difficulties faced by users is locating the information of particular interest to them. This has been aided by the development of "hypertext" links, where a keyword in one document is linked to another related document. This is a very flexible system of organization, which has in turn led to a crucial need for searching mechanisms to locate material of interest.

A great deal of attention has been focussed on the richer form of presentation of information within a framework called the World Wide Web. The Web, as it is often called, started at CERN (the European Laboratory for Particle Physics in Geneva) in 1989 as a system to allow researchers in high energy physics to exchange papers and information about their experiments. Its use grew exponentially and it is now a major part of the Internet. The Web is accessed by a "browser" that can be used to navigate through this information in a convenient manner, simply by clicking on words or images (called hyperlinks) which transfer the reader to another related document or resource. Resources are identified by their Uniform Resource Locator (URL), which identifies the machine containing the resource and the location of the resource in the machine's filestore. Hyperlinks may point to resources on the same machine or on another system anywhere else in the world. Navigation between systems is handled by the browser and is invisible to the user, who is free to appreciate the information available without concern for the technicalities of its retrieval. The most commonly used browsers are Internet Explorer™, Netscape™, Opera™, Mozilla™ and Firefox™. The Web allows close integration of text, graphics, sound, animation, and "virtual reality" 3D worlds. Recent developments may be found at the World Wide Web Consortium site [10]. It is good practice to construct web pages that are accessible to those with disabilities. In some countries this is a legal requirement.

There is a large number of "search engines" for locating material on the web. One of the most commonly used is Google™ [2], which has an advanced

ranking system for ordering the results of a search. One of the factors in this is the number of websites that link to the site in question. Altavista™ is another popular tool for searching the web. Most search engines feature a directory section that organizes links in a logical tree structure.

Remote Software

One feature of the Internet is that a user in one geographical location can access a computer system in another place. This allows software or processing power available on one machine to be operated from a remote location. However, recent developments have allowed the reverse to happen, so that software can be downloaded from a remote site onto a local machine and then run automatically without further expertise required on the part of the user. This makes computing much more accessible to the general public, as very little knowledge about the technology is required before using it. Currently, the most popular computer language for such developments is Java™. Code written in Java™ will run on many types of machines without modification. Information on recent developments may be found at [4].

Programs may be embedded within web pages and these are often used to create further web pages in a dynamic manner or to produce particular images. The programs may run either on the machine that is hosting the web page or on the machine that is accessing it. Examples of systems that can be used to create web pages of this type are Perl, JavaScript, and VBScript. Web browser facilities for JavaScript, Java, or ActiveX are often included with the browser software.

Intranets

Many organizations have seen the potential that the Internet has for the free distribution of information and have applied the Internet principles to their own organizations. They have set up what are called Intranets, which are freely accessible by everyone within the organization but inaccessible from outside. It is easy to see the benefits that may be gained by running such a private Internet world.

Security

Connection to the internet exposes a computer to attempts to modify or destroy files (*see* **Confidentiality and Computers**), take control, install viruses, or deny service. Viruses can be attached to e-mail and it is therefore important to install a current virus-checking program and to keep the virus definition files up-to-date. Attacks can also be made without using e-mail to both workstations and servers by exploiting loopholes in operating systems. Most institutions will have a protective "firewall" for all machines connected to the internet but an individual user at home may not. On the web, information sent by form is not encrypted as standard. To ensure the privacy of information, the HTTPS rather than the HTTP protocol should be used, as this encrypts communication in both directions.

E-science

Increasingly, scientific work needs to make use of either very large sets of data or very large amounts of computational resource. One response to these demands is to share the geographically distributed computing resources, storage capacity, and networks of many organizations and individuals, using an architecture designed for this purpose called a GRID. Once users have authenticated themselves to a particular GRID, they are free to make use of data and computational resources, within agreed limits, irrespective of location as though it were a single unified resource. This kind of activity is often referred to as e-Science.

Internet Uses in Biostatistics

In the world of biostatistics, the potential benefits of using the Internet, which are outlined above, are all available. In particular, the use of e-mail has had a large effect on the working lives of many people and the Web is having an increasing effect on the way information is made available in biostatistics (and indeed on the construction of this encyclopedia [1]). One of the great benefits of this revolution is that it allows biostatisticians who are geographically isolated from colleagues to maintain contact in a convenient way, and hence feel part of

a wider community. In addition to individual contacts, e-mail discussion lists provide a particularly helpful forum for this. There are many lists and information sources that are relevant to the interests of biostatisticians. Some Internet sites provide helpful listings of other sites that are of interest in a particular subject area. An example for statistics is at [5].

For the provision of information, many organizations have now created websites. This includes professional societies, and websites now exist for most of the major societies whose interests include biostatistics. These sites provide valuable, up-to-date information on professional activities. In particular, it is becoming increasingly common for conference information and registration facilities to be provided on the Web. Research organizations of all types are also making use of the Internet. In addition to general information on their activities, research papers are often made available. This provides a very fast means of disseminating research ideas and results. A useful list of organizations in the general statistics area which have websites is provided at [3]. In addition, information about statistical software and sometimes the software itself is readily accessible. A useful starting point for exploring this area is [8].

A significant issue in biostatistics is the availability of data and information from application areas in medicine and health sciences. From this perspective, the Internet can be thought of as providing access to a vast library of information held at a large number of sites throughout the world. Where the data itself are not made available, for copyright or other reasons, the Internet can still be extremely useful in identifying and contacting the site where information may be held. In addition to the lists of professional and research organizations mentioned above, a useful starting point for information is the **World Health Organization** (WHO) site at [9], which also provides links to a large number of organizations with medical and health interests.

The Internet has allowed the development of virtual communities of individuals with mutual interest(s). This has led to the emergence of virtual development teams, with members working on a project from worldwide locations, meeting rarely if ever. The pace and scale of development of open-source and public domain software has increased enormously over the last few years as a result of this style of development. There are a number of such projects in statistics such as the development of the statistical programming software **R** [7] and the umbrella project for statistical computing collaboration, Omegahat [6], which facilitates the sharing of ideas and interworking of statistical computing tools.

The web is also beginning to have significant effects on the practice of biostatistics in areas such as **clinical trials**. Facilities for patient entry, **randomization**, patient tracking and other aspects of the conduct of a trial are now available (*see* **Clinical Trials Audit and Quality Control**). These tools, which are particularly useful in **multicenter** studies, are rapidly evolving and are likely to be used increasingly in future studies.

The development and use of the Internet has taken place at such a phenomenal rate that it is difficult to predict what further changes will take place in the future. One area that is already developing very fast is commercial electronic publishing, with many journals available over the Internet. Another area is conferences, where a "virtual conference" over the Internet can provide a cheaper and more accessible alternative to the traditional physical event. While further developments are difficult to predict, it is certain that the Internet will continue to have an increasingly large effect on the way biostatistics is conducted.

References

[1] The Encyclopedia of Biostatistics website: `http://www.wiley.co.uk/eob/`

[2] The Google search facility `http://www.google.com/`

[3] The International Association for Statistical Computing, where lists of organisations relevant to statistics and some of its application areas are held. `http://www.stat.unipg.it/iasc/`

[4] The JAVA language, information on: `http://www.sun.com/java/`

[5] The LTSN Centre for Mathematics, Statistics & OR, a useful starting point in finding information or discussion lists of interest in the general area of statistics: `http://mathstore.gla.ac.uk/`

[6] The Omegahat project: `http://www.omegahat.org/`

[7] The R project: `http://www.r-project.org/`

[8] The Statlib archive, which is particularly useful for locating information about software in statistics `http://lib.stat.cmu.edu/`

[9] The World Health organisation: `http://www.who.ch/`
[10] The World Wide Web consortium site: `http://www.w3.org/`

ADRIAN W. BOWMAN, EWAN CRAWFORD,
JAMES CURRALL & STUART G. YOUNG

Interpenetrating Samples

Interpenetrating sampling (IPS), also known as interpenetrating subsampling and replicated sampling, was introduced in the pioneering contribution of **P. C. Mahalanobis** [17, 18]. Mahalanobis used IPS for the jute and rice acreage surveys in Bengal and Bihar, eastern states of India, as early as 1937. Since then, many countries of different continents have started using IPS in their large-scale **sample surveys**. The United Nations Subcommission on Statistical Sampling strongly recommended the use of IPS in 1949 [26]. IPS was originally proposed in assessing the **nonsampling errors** as the so-called "interviewer errors". Interviewer effects in the measurements from IPS are compared in the **fixed-effects** setup, and the **variance component** due to interviewers is measured in the **random-effects** setup. IPS has also turned out to be an effective method of estimating the **variance** of the **estimator** of a parameter of interest in complex surveys (see Deming [3, 4], Lahiri [14], and Yates [27]). In fact, IPS is the foundation of modern **resampling** methods like **jackknife** [24] and **bootstrap** [5], and also replication methods [20, 21].

IPS consists of selecting a sample in the form of $k (k \geq 2)$ samples using the identical sampling design from the same population. Sample sizes in k samples may or may not be equal. If k interviewers are assigned to collect information from k samples, then the interviewer effects can be studied and compared. Samples may or may not be drawn independently. The sampling design can be a complex design, that is, **multistage**, **stratified**, with equal or unequal probabilities. Let θ be the parameter of interest and $t_1, \ldots, t_k$ be the k estimators of θ based on k IPS. First, assume that

$$\mathrm{E}(t_j) = \theta, \qquad \mathrm{var}(t_j) = \sigma^2,$$
$$\mathrm{cov}(t_j, t_{j'}) = \rho\sigma^2, \quad j \neq j'.$$

Consider the following estimator $\hat{\theta}$ of θ

$$\hat{\theta} = \sum_{j=1}^{k} w_j t_j,$$

where the w_js are fixed constant (known or unknown), with $\sum_{j=1}^{k} w_j = 1$. It can be seen that

$$\mathrm{var}(\hat{\theta}) = \sigma^2(1-\rho)\sum_{j=1}^{k} w_j^2 + \rho\sigma^2.$$

An estimator of $\mathrm{var}(\hat{\theta})$ is

$$\widehat{\mathrm{var}}(\hat{\theta}) = \frac{\sum_{j=1}^{k} w_j^2}{1 - \sum_{j=1}^{k} w_j^2} \sum_{j=1}^{k} w_j (t_j - \hat{\theta})^2.$$

It can be checked that

$$\mathrm{E}[\widehat{\mathrm{var}}(\hat{\theta})] = \mathrm{var}(\hat{\theta}) - \rho\sigma^2.$$

As a result, $\widehat{\mathrm{var}}(\hat{\theta})$ is an **unbiased** estimator of $\mathrm{var}(\hat{\theta})$ when $\rho = 0$, $\widehat{\mathrm{var}}(\hat{\theta})$ overestimates $\mathrm{var}(\hat{\theta})$ when $\rho < 0$, and $\widehat{\mathrm{var}}(\hat{\theta})$ underestimates $\mathrm{var}(\hat{\theta})$ when $\rho > 0$. If $w_1 = \cdots = w_k = 1/k$, then $\hat{\theta} = [(t_1 + \cdots + t_k)/k] = \bar{t}$ and

$$\widehat{\mathrm{var}}(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{j=1}^{k} (t_j - \bar{t})^2.$$

In the case where k samples are drawn independently, $\widehat{\mathrm{var}}(\hat{\theta})$ is an unbiased estimator of $\mathrm{var}(\hat{\theta})$. Thus, IPS provides a quick, simple, and effective way of estimating the variance of the estimator even in a complex survey. The case $\mathrm{var}(t_j) = \sigma_j^2$, $j = 1, \ldots, k$, is considered in Murthy [22], Koop [12] and others. Suppose that $t(j)$ is the value of $\bar{t}$ when the jth estimator of θ from the jth investigator is omitted. Then

$$t(j) = \frac{t_1 + \cdots + t_{j-1} + t_{j+1} + \cdots + t_k}{k-1}, \quad j = 1, \ldots, k.$$

Let

$$t(\cdot) = \frac{t(1) + t(2) + \cdots + t(k)}{k}.$$

The jackknife version of $\bar{t}$ is given by

$$\hat{\theta}^J = k\bar{t} - (k-1)t(\cdot).$$

It can now be seen that $\hat{\theta}^J = \bar{t}$, and, consequently, the expression of $\mathrm{var}(\hat{\theta}^J)$ in Efron & Stein [6]

$$\widehat{\mathrm{var}}(\hat{\theta}^J) = \frac{k-1}{k} \sum_{j=1}^{k} [t(j) - t(\cdot)]^2,$$

is exactly the same as $\widehat{\mathrm{var}}(\hat{\theta}) = \widehat{\mathrm{var}}(\bar{t})$ given above. For comparison of several **ratio and regression estimators** based on IPS with or without jackknife, see Ghosh & Gomez [9, 10].

In IPS, k samples are drawn from the same population using the identical sampling design. If k samples are selected with replacement so that they are independent, then one can see its similarity in principle with the modern Bootstrap Sampling (BSS) (see Efron & Tibshirani [7]). In BSS, the observed data are a **random sample** of size n from an unknown probability distribution F. Bootstrap samples are random samples of size n drawn with replacement from the observed data or the empirical distribution $\hat{F}$. If we treat the observed data as a finite population of size n, then BSS are, in fact, IPS with $k = n$. Of course, there is no interviewer effect for bootstrap samples.

In IPS, three basic principles of **experimental designs**, namely, **randomization**, replication, and local control, are used. The main purpose of IPS is to identify, reduce, and control errors due to interviewers. IPS is used extensively not only in agriculture but also in **social sciences**, **demography**, epidemiology, public health, and many other fields (see Hansen et al. [11], Lahiri [15], Som [25], Fellegi [8], Bailey et al. [1], Levy & Lemeshow [16]). Fractile graphical analysis developed in [19], based on IPS, is used in comparing, analyzing, and testing the separation between two populations when a concomitant variable (*see* **Covariate**) is measured in addition to the response variable. Details on the theory of IPS are available in [2], [13], and [23].

References

[1] Bailey, L., Moore, T.F. & Bailer, B.A. (1978). An interviewer variance study for the eight impact cities of the national crime survey cities sample, *Journal of the American Statistical Association* **73**, 16–23.

[2] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[3] Deming, W.E. (1960). *Sample Design in Business Research*. Wiley, New York.

[4] Deming, W.E. (1963). On some of the contributions of interpenetrating network of samples, in *Contribution to Statistics*, C.R. Rao, ed. Statistical Publishing Society, Calcutta, pp. 57–66.

[5] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**, 1–26.

[6] Efron, B. & Stein, C. (1981). The jackknife estimate of variance, *Annals of Statistics* **9**, 586–596.

[7] Efron, B. & Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[8] Fellegi, I. (1964). Response variance and its estimation, *Journal of the American Statistical Association* **59**, 1016–1041.

[9] Ghosh, S. & Gomez, R. (1986). Comparison of ratio estimators based on interpenetrating sub-samples with or without jackknifing, *Journal of the Indian Society of Agricultural Statistics* **38**, 200–210.

[10] Ghosh, S. & Gomez, R. (1987). Interpenetrating subsampling regression estimation with or without jackknifing, *Communications in Statistics – Simulation and Computation* **16**, 1105–1116.

[11] Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample Survey Methods and Theory*. Wiley, New York.

[12] Koop, J.C. (1967). Replicated (or interpenetrating) samples of unequal sizes, *Annals of Mathematical Statistics* **38**, 1142–1147.

[13] Koop, J.C. (1988). The technique of replicated or interpenetrating samples, in *Handbook of Statistics*, Vol. 6, P.R. Krishnaiah & C.R. Rao, eds. Elsevier, Amsterdam, pp. 336–368.

[14] Lahiri, D.B. (1954). Technical paper on some aspects of the development of the sample design: National Sample Survey Report No. 5, *Sankhyā* **14**, 264–316.

[15] Lahiri, D.B. (1957). Observations on the use of interpenetrating samples in India, *Bulletin of the International Statistical Institute* **36**, Part 3, 144–152.

[16] Levy, P.S. & Lemeshow (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

[17] Mahalanobis, P.C. (1944). On large-scale sample surveys, *Philosophical Transactions of the Royal Society of London, Series B* **231**, 329–451.

[18] Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute, *Journal of the Royal Statistical Society* **109**, 325–378.

[19] Mahalanobis, P.C. (1960). A method of fractile graphical analysis, *Econometrica* **28**, 325–351.

[20] McCarthy, P.C. (1966). Replication: an approach to the analysis of data from complex surveys, *PHS Publication No. 1000, Series E, No. 14*. US Government Printing Office, Washington.

[21] McCarthy, P.C. (1969). Pseudo-replication: half-samples, *International Statistical Review* **37**, 239–264.

[22] Murthy, M.N. (1964). On Mahalanobis' contributions to the development of sample survey theory and methods, in *Contribution to Statistics*, C.R. Rao, ed. Statistical Publishing Society, Calcutta, pp. 282–316.

[23] Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

[24] Quenouille, M. (1949). Approximate tests of correlation in time series, *Journal of the Royal Statistical Society, Series B* **11**, 18–44.

[25] Som, R.K. (1965). Use of interpenetrating samples in demographic studies, *Sankhyā, Series B* **27**, 329–342.

[26] United Nations (1949). *Recommendations for the Preparation of Sample Survey Reports, Series C, No. 1*. United Nations, New York.

[27] Yates, F. (1981). *Sampling Methods for Censuses and Surveys*, 4th Ed. Oxford University Press, London.

SUBIR GHOSH

Interval Censoring

Interval **censoring** is commonly used to denote a type of sampling scheme or to describe a type of incomplete data. By interval-censored data, we mean that a **random variable** of interest is known only to lie in an interval, instead of being observed exactly. In most applications of **survival analysis**, the random variable is the time to some event such as death or a disease. A common example of interval-censored survival data occurs in medical or health studies that entail periodic follow-up. Many **clinical trials** and **longitudinal** studies fall into this category [18]. In this situation, an individual due for the scheduled observations for a clinically observed change in disease status may miss some observations and may return with a changed status, thus contributing an interval-censored time of the occurrence of the change. Another example arises in the acquired immune deficiency syndrome (**AIDS**) studies [13] that concern the human immunodeficiency virus (HIV) infection and the AIDS **incubation period** (the time from HIV infection to AIDS diagnosis). In this case, if a subject is HIV positive at the beginning of the study, his or her HIV infection time is usually determined by a **retrospective study** of the subject's history. Thus, only an interval given by the last HIV negative test and the first HIV positive test is known for the HIV infection time.

An important special case of interval-censored data is current status data [29] and [61]. In this situation, each subject is observed only once for the status of the occurrence of the event of interest at the observation time. In other words, the observation of the time to the event is either left- or right-censored (the survival time is less or greater than the observation time). **Cross-sectional** data provide one example of current status data [31] and another example is given by nonlethal tumor data when the time to tumor onset is of interest, but not directly observable [14] (*see* **Tumor Incidence Experiments**). Note that for the first example, current status data occur because of study designs, while for the second case, they are observed because of the inability of measuring the variable directly and exactly. Current status data are also sometimes referred to as case I interval-censored data and in correspondence, the general case is referred to as case II interval-censored data [23].

Another special case of interval-censored data that may occur is left-censored data, in which the time to the event of interest is either left censored or exactly known. One reason behind the occurrence of the left-censored data is the inability of measuring the variable of interest when it is below a certain level. Such an example is given by the severity of an adverse event related to a drug, which sometimes can be determined only when it is over a certain degree.

To this point, survival time has been defined in a way that starts from time zero or a known time point. A more general framework is to define survival time as the time between two related events. This illustrates a more complicated type of interval-censored data: doubly interval-censored data [60], in which the times of the occurrences of both events defining the survival time are interval censored. An example of such data is provided by the AIDS studies discussed above when the variable of interest is AIDS incubation time [13].

For the analysis of interval-censored failure-time data, in the following, we will first discuss nonparametric estimation of the distribution function as well as the **hazard** function for survival time (*see* **Survival Distributions and Their Characteristics**). Secondly, regression analysis of interval-censored data will be investigated under various regression models. The comparison of several survival functions will be considered thirdly and followed by discussion on some other topics. These will include doubly interval-censored failure-time data, interval-censored data with truncation, multivariate interval-censored data, and interval-censored data with informative censoring. Finally, some concluding remarks will be given. Unless specified otherwise, we will assume that the censoring mechanism resulting in censoring intervals is independent of the survival time.

Nonparametric Estimation

In medical and health studies, estimation of the cumulative distribution function (cdf) of survival time or the survival function is perhaps the most important and common task. Let T denote the survival time of interest in a survival study and $F = \Pr(T \leq t)$ its cdf. Suppose that observed data can be represented by $\{I_i\}_{i=1}^n$, where $I_i = (L_i, R_i]$ is the interval known to contain the unobserved survival time associated with the ith subject. If $L_i = 0$, we have a left-censored observation and if $R_i = \infty$, we have a

right-censored observation. Let $\{s_j\}_{j=0}^{m+1}$ denote the unique ordered elements of $\{0, \{L_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \infty\}$, α_{ij} be the indicator (*see* **Dummy Variables**) of the event $(s_{j-1}, s_j] \subseteq I_i$, and $p_j = F(s_j) - F(s_{j-1})$. Then the **likelihood** function of $p = (p_1, \ldots, p_{m+1})'$ is proportional to

$$L(p) = \prod_{i=1}^{n} \{F(R_i) - F(L_i)\} = \prod_{i=1}^{n} \left(\sum_{j=1}^{m+1} \alpha_{ij} p_j \right) \tag{1}$$

and the problem of finding the **nonparametric maximum likelihood estimator** (NPMLE) of F becomes that of maximizing $L(p)$ with respect to p subject to $\sum_{j=1}^{m+1} p_j = 1$ and $p_j \geq 0$ $(j = 1, \ldots, m+1)$ [20, 34, 39]. Note that a more general way to express an interval-censored observation is to use the finite union of disjoint intervals [64] and in this case, the discussion here equally applies.

To maximize $L(p)$ with respect to p, a simple and common way is to use the self-consistency **algorithm** proposed by **Turnbull** [64]. It can be seen as an application of the **EM algorithm** and iterates the equation $p_j^{\text{new}} = n^{-1} \sum_{i=1}^{n} [\alpha_{ij} p_j^{\text{old}} / (\sum_{l=1}^{m+1} \alpha_{il} p_l^{\text{gold}})]$ until convergence. This approach is easy to implement, but is known to have a slow convergence rate. An alternative is to apply the convex minorant algorithm introduced by Groeneboom and Wellner [23], which promises to converge faster than the self-consistency algorithm. Böhning et al. [10] proposed to use the vertex-exchange or other algorithms proposed for the mixture model problem because of the similarity of the two problems. All the above algorithms are iterative and in fact, there is no closed form for the NPMLE of F.

For current status data, however, a closed form of the NPMLE can be found. In this case, the NPMLE of F can be shown to be equal to the **isotonic regression** of $\{d_1/n_1, \ldots, d_m/n_m\}$ with weights $\{n_1, \ldots, n_m\}$, where $d_j = \sum_{i \varepsilon S_j} I(T_i \leq s_j)$, $n_j = |S_j|$ and S_j denotes the set of subjects who are observed at s_j, $j = 1, \ldots, m$. Thus, by using the max–min formula for an isotonic regression [3], the NPMLE of F can be written as

$$\hat{F}_n(s_j) = \max_{u \leq j} \min_{v \geq j} \left(\frac{\sum_{l=u}^{v} d_j}{\sum_{l=u}^{v} n_j} \right). \tag{2}$$

A self-consistent estimate of F may not be an NPMLE. To verify this, one approach is to use the so-called Kuhn–Tucker conditions given in [20]. Note that at the NPMLE, p_j can be nonzero only if s_{j-1} is a left endpoint L_i for some subject i and s_j is a right endpoint R_k for some possibly different subject k. Some of the p_j's that satisfy this criterion may still be zero. The Kuhn–Tucker conditions can also be applied to identify these zero p_j's, thus speeding up the self-consistency algorithm. Another approach is to use the fact that an estimate $\hat{p}$ is an NPMLE if and only if $\sup_{1 \leq j \leq m+1} \sum_{i=1}^{n} (\alpha_{ij} / \sum_{l=1}^{m+1} \alpha_{il} \hat{p}_l) = n$ [10]. Gentleman and Geyer [20] also discussed the conditions required for the uniqueness of NPMLEs.

It can be shown that the above NPMLE $\hat{F}_n$ is **consistent** [23, 66]. Furthermore, as $n \to \infty$ and at fixed time point t_0, $\hat{F}_n(t_0)$ has a limiting, nonnormal distribution at $n^{1/3}$ or $(n \log n)^{1/3}$ **convergence** rate depending on if the probability of observing $T = t_0$ is zero or away from zero [23, 65]. Note that this is different than the usual $n^{1/2}$-convergence rate. However, the integral of $\hat{F}_n$ and its linear functionals can be shown to have asymptotic normal distributions with $n^{1/2}$-convergence [21, 27]. For variance estimation of $\hat{F}_n$, Sun [59] presented two methods, a generalization of Greenwood's formula (*see* **Survival Analysis, Overview**) and a **bootstrap** approach.

Sometimes estimation of the hazard function of survival time may be of interest. In this case, one way is to use the empirical estimator, which is usually rough and difficult to interpret. Corresponding to this, several smooth estimators, which are more descriptive than the empirical estimator, of the hazard function have been proposed (*see* **Smoothing Hazard Rates**). Among others, Kooperberg and Stone [33] and Rosenberg [48] gave **spline**-based estimators (*see* **Spline Smoothing**). Bebchuk and Betensky [4] and Betensky et al. [7] considered the **multiple imputation** and local likelihood approaches, respectively.

Regression Analysis

Regression analysis of survival data is commonly performed to study the effect of various **covariate** factors such as treatment, sex, and age on survival time. Let X_i denote the covariate vector associated with the ith subject and assume that the censoring mechanism is independent of the covariates. By using the notation given above, the likelihood then has the

form:

$$L = \prod_{i=1}^{n} \{F(R_i|X_i) - F(L_i|X_i)\}$$
$$= \prod_{i=1}^{n} \left[\sum_{j=1}^{m+1} \alpha_{ij} \{F(s_j|X_i) - F(s_{j-1}|X_i)\} \right], \quad (3)$$

where $F(T|X)$ denotes the cdf of T given covariates X.

In survival analysis, the most commonly used regression model is perhaps the **proportional hazards** (PH) model, which has the form

$$\lambda(t|X) = \lambda_0(t) \exp(\beta' X) \quad (4)$$

[12], where $\lambda_0(t)$ denotes an unknown baseline hazard function and β the regression coefficients (*see* **Cox Regression Model**). For inference about β and the cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$, a natural method is the full likelihood approach, which maximizes L over β and $\Lambda_0(t)$ simultaneously and was first discussed by Finkelstein [18]. Huang [24] also studied this approach for the case of current status data and showed that the **maximum likelihood estimator** (MLE) of the regression coefficients is consistent and **efficient** and has an asymptotic **normal distribution** with $n^{1/2}$-convergence rate.

An alternative to the above full likelihood approach is the **marginal likelihood** approach. This approach defines a marginal likelihood as the summation of the probabilities of the ranking of the T_i's over the set of all possible rankings of the T_i's that are consistent with observed interval-censored data [50]. It is a generalization of the corresponding approach for right-censored data [30] and has the advantage of not involving $\lambda_0(t)$. The disadvantage is that it does not have a simple and easily manageable form, resulting in the need for great computational effort. In addition, little is known about its properties. Another choice for inference about the PH model is to use some types of imputation approaches, which involve generating right-censored data based on observed interval-censored data [41, 51].

Although the PH model yields sound results in many cases, there are situations in which other models may provide a better fit to interval-censored data. For example, the **proportional odds regression** model given by

$$\log \left\{ \frac{[F(t|X)]}{[1 - F(t|X)]} \right\} = \alpha(t) + \beta' X \quad (5)$$

is often used for **environmental** health data, where $\alpha(t)$ is a monotone-increasing function. Among others, Rossini and Tsiatis [49] discussed the fitting of this model to current status data. Huang and Rossini [26] and Rabinowitz et al. [46] considered the sieve estimation and the approximated score function methods, respectively.

Another alternative to the PH model that has been discussed for interval-censored data is the **accelerated** regression model defined by $\log(T) = \beta' X + \varepsilon$, where the distribution of ε is unknown [9, 47]. The **additive hazards** regression model and the **logistic** model have also been investigated for regression analysis of interval-censored data. In these cases, the discussion has been confined to current status data for the former [37, 38] and to discrete survival data for the latter [54]. Some of the other models proposed recently for the regression analysis of interval-censored data can be found in references [8, 11, 25, 32].

Comparison of Survival Functions

The comparison of survival functions is often the primary goal of clinical and follow-up studies. In the case of interval-censored data, as in other situations, two commonly used approaches are to base the comparison on regression techniques and to develop distribution-free test procedures. The application of the regression techniques involves defining covariates X as group or treatment indicators and then using the Wald or score test (*see* **Likelihood**) for testing regression coefficients equal to zero [18, 47]. In this, one can apply the regression methods described above. An alternative, which was not discussed above, is to consider some **rank**-based regression models and to derive rank test procedures for the comparison [17]. For example, Self and Grossman [52] considered the linear regression model $T = \alpha + \beta' X + \varepsilon$, where α is a constant and ε denotes the error term. Under the model, they derived the **linear rank tests** defined as the score tests for $\beta = 0$ from the marginal likelihood of the ranking of the survival times T's.

Several distribution-free test procedures have been proposed for the comparison of survival functions

based on interval-censored data. Among these, Andersen and Ronn [2] and Sun and Kalbfleisch [62] proposed some **nonparametric** tests for current status data. Sun [53] presented a nonparametric procedure of **logrank** type and Pan [40] developed an approach based on multiple imputation. Most of the above mentioned procedures are rank-oriented or sensitive to ordered hazard differences. Sometimes test procedures that are sensitive to ordered survival differences may be preferred. For this purpose, Petroni and Wolfe [45] and Fang et al. [16] considered approaches based on differences between estimated survival functions. Furthermore, Lim and Sun [35] proposed three classes of nonparametric test procedures that include most of existing tests as special cases. Note that most of the existing comparison procedures require the same censoring distribution.

Other Topics

This section will discuss a few subjects not investigated above about interval censoring. One is the analysis of doubly interval-censored data, which was first studied by De Gruttola and Lagakos [13]. In this case, the analysis is more complicated since the likelihood function involves not only the distribution of the survival time, but also the distribution of the originating event that defines the survival time. Following De Gruttola and Lagakos [13], who proposed a self-consistency algorithm for estimation of the distribution function of survival time, many authors have considered the inference about doubly interval-censored data. Some recent contributions include Fang and Sun [15], who discussed the consistency of the NPMLE of the distribution function, and Sun [58], who developed a nonparametric test for treatment comparison. Also Goggins [22], Pan [42] and Sun et al. [63] studied the regression problem under the PH model. More discussion and references about doubly interval-censored data can be found in [60].

Truncation is another feature of survival data and may sometimes occur together with interval censoring. One of the early papers discussing this is given by Turnbull [64], who proposed a self-consistency algorithm for the NPMLE of a distribution for interval-censored and truncated data. More recently, among others, Pan and Chappell [43] and Sun [55] considered the one sample estimation problem when left-truncation and general truncation are involved, respectively. Lim et al. [36] also discussed the one sample estimation problem with general truncation and the existence of a **change-point** and Alioum and Commenges [1] and Pan and Chappell [44] considered the regression problem under the PH model.

The discussion so far has been focusing on univariate failure-time data and the research on multivariate failure-time data in the literature is relatively limited. This is especially the case for multivariate interval-censored data (*see* **Multivariate Survival Analysis**). One problem investigated in this case is the one sample estimation problem for bivariate interval-censored failure time data [5]. Also for bivariate failure-time data, Betensky and Finkelstein [6] generalized the Kendall's coefficient (*see* **Rank Correlation**) to the interval-censoring situation. Another subject, that often occurs in practice and has not been discussed much in the literature, is the analysis of interval-censored data when the censoring mechanism resulting in interval censoring is informative or depends on the survival of interest. For this, Sun [57] developed a nonparametric test for treatment comparison for a special case in which observed data are current status data and observation times depend on treatments. Finkelstein et al. [19] discussed both one sample and regression problems.

Concluding Remarks

For the analysis of interval-censored failure-time data, the discussion here has been focusing on more recently developed and **semiparametric** and nonparametric methods. One can find early references about interval-censored data in [28, 56]. An alternative to the semiparametric and nonparametric methods is to use **parametric methods** [56], which are usually relatively straightforward. It is worth noting that interval-censored data discussed here are different from **grouped survival** data, which are sometimes also referred to as interval-censored data in the literature. The interval-censored data reduce to grouped data if all observed intervals either completely overlap on each other or have no overlaps.

There are still a lot of open questions in the analysis of interval-censored data. One is that the properties of many proposed methods remain unknown and this is especially the case for doubly interval-censored data. Although a great deal of research for regression analysis of interval-censored data has been done,

there is no approach available as simple as the **partial likelihood** method for right-censored data. Also, there are no methods available for **model checking** for all regression models discussed above and more research is needed for multivariate and informatively interval-censored failure-time data.

References

[1] Alioum, A. & Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data, *Biometrics* **52**, 512–524.

[2] Andersen, P.K. & Ronn, B.B. (1995). A nonparametric test for comparing two samples where all observations are either left- or right-censored, *Biometrics* **51**, 323–329.

[3] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

[4] Bebchuk, J.D. & Betensky, R.A. (2000). Multiple imputation for simple estimation of the hazard function based on interval censored data, *Statistics in Medicine* **19**, 405–419.

[5] Betensky, R.A. & Finkelstein, D.M. (1999). A nonparametric maximum likelihood estimator for bivariate interval censored data, *Statistics in Medicine* **18**, 3089–3100.

[6] Betensky, R.A. & Finkelstein, D.M. (1999). An extension of Kendall's coefficient of concordance to bivariate interval censored data, *Statistics in Medicine* **18**, 3101–3109.

[7] Betensky, R.A., Lindsey, J.C., Ryan, L.M. & Wand, M.P. (1999). Local EM estimation of the hazard function for interval-censored data, *Biometrics* **55**, 238–245.

[8] Betensky, R.A., Lindsey, J.C., Ryan, L.M. & Wand, M.P. (2002). A local likelihood proportional hazards model for interval censored data, *Statistics in Medicine* **21**, 263–275.

[9] Betensky, R.A., Rabinowitz, D. & Tsiatis, A.A. (2001). Computationally simple accelerated failure time regression for interval censored data, *Biometrika* **88**, 703–711.

[10] Böhning, D., Schlattmann, P. & Dietz, E. (1996). Interval censored data: a note on the nonparametric maximum likelihood estimator of the distribution function, *Biometrika* **83**, 462–466.

[11] Chen, Y.Q. & Jewell, N.P. (2001). On a general class of semiparametric hazards regression models, *Biometrika* **88**, 687–702.

[12] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[13] De Gruttola, V.G. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics* **45**, 1–11.

[14] Dinse, G.E. & Lagakos, S.W. (1983). Regression analysis of tumour prevalence data, *Applied Statistics* **32**, 236–248.

[15] Fang, H. & Sun, J. (2001). Consistency of nonparametric maximum likelihood estimation of a distribution function based on doubly interval-censored failure time data, *Statistics and Probability Letters* **55**, 311–318.

[16] Fang, H., Sun, J. & Lee, M.-L.T. (2002). Nonparametric survival comparison for interval-censored continuous data, *Statistica Sinica* **12**, 1073–1083.

[17] Fay, M.P. (1996). Rank invariant tests for interval censored data under the grouped continuous model, *Biometrics* **52**, 811–822.

[18] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data, *Biometrics* **42**, 845–854.

[19] Finkelstein, D.M., Goggins, W.B. & Schoenfeld, D.A. (2002). Analysis of failure time data with dependent interval-censoring, *Biometrics* **58**, 298–304.

[20] Gentleman, R. & Geyer, C.J. (1994). Maximum likelihood for interval censored data: Consistency and computation, *Biometrika* **81**, 618–623.

[21] Geskus, R. & Groeneboom, P. (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2, *Annals of Statistics* **27**, 627–674.

[22] Goggins, W.B., Finkelstein, D.M. & Zaslavsky, A.M. (1999). Applying the Cox proportional hazards model for analysis of latency data with interval censoring, *Statistics in Medicine* **18**, 2737–2747.

[23] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. DMV Seminar, Band 19, Birkhauser, New York.

[24] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring, *Annuals of Statistics* **24**, 540–568.

[25] Huang, J. (1999). Efficient estimation of the partly linear additive Cox model, *Annals of Statistics* **27**, 1536–1563.

[26] Huang, J. & Rossini, A.J. (1997). Sieve estimation for the proportional odds failure-time regression model with interval censoring, *Journal of the American Statistical Association* **92**, 960–967.

[27] Huang, J. & Wellner, J.A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I, *Statistica Neerlandica* **49**, 153–163.

[28] Huang, J. & Wellner, J.A. (1997). Interval censored survival data: a review of recent progress, in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D. Lin & T. Fleming, eds. Springer-Verlag, New York, 105–123.

[29] Jewell, N.P. & Laan, Mv. (1995). Generalizations of current status data with applications, *Lifetime Data Analysis* **1**, 101–110.

[30] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[31] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[32] Kooperberg, C. & Clarkson, D.B. (1997). Hazard regression with interval-censored data, *Biometrics* **53**, 1485–1494.

[33] Kooperberg, C. & Stone, C.J. (1992). Logspline density estimation for censored data, *Journal of Computational and Graphical Statistics* **1**, 301–328.

[34] Li, L., Watkins, T. & Yu, Q. (1997). An EM algorithm for estimating survival functions with interval-censored data, *Scandinavian Journal of Statistics* **24**, 531–542.

[35] Lim, H.J. & Sun, J. (2003). Nonparametric tests for interval-censored failure time data, *Biometrical Journal* **45**, 263–276.

[36] Lim, H.J., Sun, J. & Matthews, D.E. (2002). Maximum likelihood estimation of a survival function with a change-point for truncated and interval-censored data, *Statistics in Medicine* **21**, 743–752.

[37] Lin, D.Y., Oakes, D. & Ying, Z. (1998). Additive hazards regression with current status data, *Biometrika* **85**, 289–298.

[38] Martinussen, T. & Scheike, T.H. (2002). Efficient estimation in additive hazards regression with current status data, *Biometrika* **89**, 649–658.

[39] Ng, M.P. (2002). A modification of Peto's nonparametric estimation of survival curves for interval-censored data, *Biometrics* **58**, 439–442.

[40] Pan, W. (2000). A two-sample Test with Interval Censored Data via Multiple Imputation, *Statistics in Medicine* **19**, 1–11.

[41] Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data, *Biometrics* **56**, 199–203.

[42] Pan, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies, *Biometrics* **57**, 1245–1250.

[43] Pan, W. & Chappell, R. (1998). Estimating survival curves with left-truncated and interval-censored data under monotone hazards, *Biometrics* **54**, 1053–1060.

[44] Pan, W. & Chappell, R. (2002). Estimation in the Cox proportional hazards model with left truncated and interval censored data, *Biometrics* **58**, 64–70.

[45] Petroni, G.R. & Wolfe, R.A. (1994). A two-sample test for stochastic ordering with interval-censored data, *Biometrics* **50**, 77–87.

[46] Rabinowitz, D., Betensky, R.A. & Tsiatis, A. (2000). Using conditional logistic regression to fit proportional odds models to interval censored data, *Biometrics* **56**, 511–518.

[47] Rabinowitz, D., Tsiatis, A. & Aragon, J. (1995). Regression with interval-censored data, *Biometrika* **82**, 501–513.

[48] Rosenberg, P.S. (1995). Hazard function estimation using B-splines, *Biometrics* **51**, 874–887.

[49] Rossini, A. & Tsiatis, A.A. (1996). A semiparametric proportional odds regression model for the analysis of current status data, *Journal of the American Statistical Association* **91**, 713–721.

[50] Satten, G.A. (1996). Rank-based inference in the proportional hazards model for interval censored data, *Biometrika* **83**, 355–370.

[51] Satten, G.A., Datta, S. & Williamson, J.M. (1998). Inference based on imputated failure times for the proportional hazards model with interval-censored data, *Journal of the American Statistical Association* **93**, 318–327.

[52] Self, S.G. & Grossman, E.A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers, *Biometrics* **42**, 521–530.

[53] Sun, J. (1996). A nonparametric test for interval-censored failure time data with application to AIDS studies, *Statistics in Medicine* **15**, 1387–1395.

[54] Sun, J. (1997). Regression analysis of interval-censored failure time data, *Statistics in Medicine* **16**, 497–504.

[55] Sun, J. (1997). Self-consistency estimation of distributions based on truncated and doubly censored data with applications to AIDS cohort studies, *Lifetime Data Analysis* **3**, 305–313.

[56] Sun, J. (1998). Interval censoring, *Encyclopedia of Biostatistics*, 1st Ed. John Wiley & Sons, New York, pp. 2090–2095.

[57] Sun, J. (1999). A nonparametric test for current status data with unequal censoring, *Journal of the Royal Statistical Society, Series B* **61**, 243–250.

[58] Sun, J. (2001). Nonparametric test for doubly interval-censored failure time data, *Lifetime Data Analysis* **7**, 363–375.

[59] Sun, J. (2001). Variance estimation of a survival function for interval-censored survival data, *Statistics in Medicine* **20**, 1249–1257.

[60] Sun, J. (2004). Statistical analysis of doubly interval-censored failure time data, in *Handbook of Statistics: Survival Analysis*, N. Balakrishnan & C.R. Rao, eds; 105–122.

[61] Sun, J. & Kalbfleisch, J.D. (1993). The analysis of current status data on point processes, *Journal of the American Statistical Association* **88**, 1449–1454.

[62] Sun, J. & Kalbfleisch, J.D. (1996). Nonparametric tests of tumor prevalence data, *Biometrics* **52**, 726–731.

[63] Sun, J., Liao, Q. & Pagano, M. (1999). Regression analysis of doubly censored failure time data with applications to AIDS studies, *Biometrics* **55**, 909–914.

[64] Turnbull, B.W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 290–295.

[65] Wellner, J. (1995). Interval censoring case 2: alternative hypothesis, in *Analysis of Censored Data*, H.L. Koul & J.V. Deshpande eds. IMS, Hayward, pp. 271–291.

[66] Yu, Q., Li, L. & Wong, G. (2000). On consistency of self-consistent estimator of survival functions with interval-censored data, *Scandinavian Journal of Statistics* **27**, 35–44.

JIANGUO SUN

Intervention Analysis in Time Series

When data are collected in the form of time series there are important questions concerning "changes" in the series. Changes may be "man-made" or they may arise "naturally". How efficient was a preventive program to decrease the monthly number of accidents? How did the frequency of traditional neurological diagnostic methods change after the introduction of computer tomography? How did the pattern of morbidity in a population change after an environmental accident? Notifications of diseases, entries in a hospital, injuries due to accidents, etc. are usually collected in fixed equally spaced intervals. Such time series observations are likely to be dependent. ARIMA models [autoregressive integrated moving average models (*see* **ARMA and ARIMA Models**) and Box–Jenkins models [1]] allow the stochastic dependence of consecutive data to be modeled. Intervention analysis proposed by Box & Tiao [2] is an extension of ARIMA modeling allowing study of the magnitude and structure of changes of ARIMA processes.

The well-known two-sample t test for a change in level after an intervention may not be appropriate in this situation due to the possible dependency of the observations. In addition, this test allows only an assessment of the magnitude of a change and not of its structure. Since the series may be nonstationary, large changes of the series could occur even when no intervention takes place (*see* **Stationarity**). Intervention analysis may allow an investigator to distinguish between what can be expected due to nonstationarity alone and what cannot.

Analogous questions of "change" may arise when studying time series data recorded in an individual patient; changes of time series, for example, may occur after the intervention "treatment". Dependence of consecutive observations may be important when data such as blood glucose are recorded within a single patient over time. Such studies on individual subjects may be interesting and relevant in basic medical research and in clinical applications. In clinical research they may allow physicians the assessment of individual treatment effects. Decisions on treatment strategy may be based on knowledge of the stochastic processes representing the observed time series, thus allowing full use to be made of the recorded data [4].

Intervention Models

Let $y_{t-1}, y_t, y_{t+1}, \ldots$ denote the observations (number of entries in a hospital, etc.) at equally spaced times, $t-1, t, t+1, \ldots$ (e.g. yesterday, today, tomorrow, etc.). The intervention model states that y_t (or a suitably transformed version of the series) may be decomposed into two parts, an "explained" part u_t and an "unexplained" or "noise" part n_t,

$$y_t = u_t + n_t. \tag{1}$$

The Noise Series n_t

The noise series (*see* **Noise and White Noise**) or unexplained part n_t is an autoregressive integrated moving average ARIMA(p, d, q) process given by

$$w_t = \nabla^d n_t, \tag{2}$$

$$w_t = \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \text{ or}$$

$$\phi(B) w_t = \theta(B) a_t. \tag{3}$$

∇ is the differencing operator such that $\nabla n_t = n_t - n_{t-1}$ and the integer d is the number of times n_t has to be differenced to obtain a stationary series w_t. B is the backward shift operator (*see* **Backward and Forward Shift Operators**) such that $Bw_t = w_{t-1}$, $B^k w_t = w_{t-k}$, $\phi(B)$ and $\theta(B)$ are polynomials in B:

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \text{ and}$$

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q. \tag{4}$$

$\phi(B)$ is called the autoregressive operator of order p and $\theta(B)$ the moving average operator of order q. The parameters of the noise process $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are constrained such that the roots of $\phi(z)$ and $\theta(z)$ in the complex z-plane lie outside the unit circle. a_t is a *white noise* series consisting of independent identically distributed normal random variables with mean zero and variance σ_a^2. The ARIMA model for the noise n_t may be extended to the seasonal ARIMA model by including seasonal autoregressive and moving average operators. In the

absence of u_t, the observed series y_t is just the ARIMA process n_t.

The Explained Part u_t

The explained part u_t is the "response" of a system to a dummy input variable I_t:

$$u_t = f(I_t). \tag{5}$$

The input I_t is usually taken as the unit pulse function p_t,

$$p_t = \begin{cases} 1, & \text{for } t = T, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

or the unit step function s_t,

$$s_t = \begin{cases} 1, & \text{for } t \geq T, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The pulse function p_t may represent, for example, an unusual event which acts only at time T. The step function s_t represents, for example, a preventive measure starting at time T. Since the "noise process" n_t may be nonstationary, large changes of the series could occur even when no intervention takes place.

Basic Patterns of Response

Figure 1 shows basic intervention models. In the first line the two dummy input variables s_t and p_t are depicted. The lines (a), (b), and (c) below show "responses" corresponding to the following three models:

1. Figure 1(a): *the simplest, case*

$$f(I_t) = \omega_0 I_t. \tag{8}$$

   The response is just the pulse- or step-input I_t multiplied by ω_0. The parameter ω_0 measures the "strength" of the effect. In this model, and with the step function as input, the new level is reached immediately.
2. Figure 1(b): *a refined model*:

$$\begin{aligned} f(I_t) &= \omega_0 I_t - \omega_1 I_{t-1}, \text{ or} \\ f(I_t) &= (\omega_0 - \omega_1 B) I_t, \text{ or} \\ f(I_t) &= \omega(B) I_t, \end{aligned} \tag{9}$$

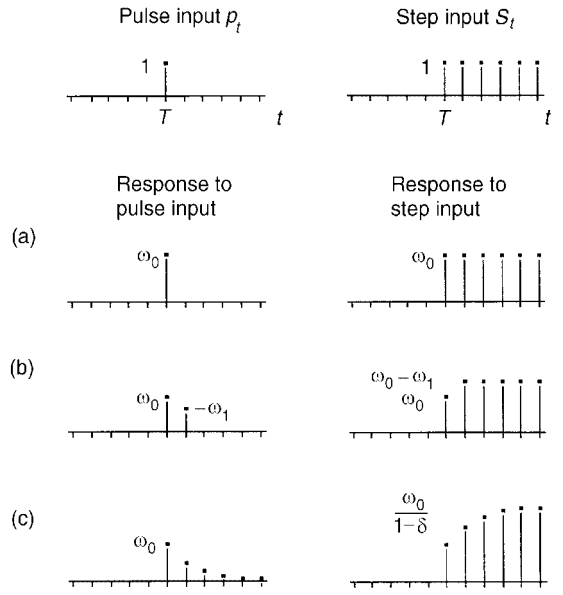

Figure 1 Responses to a unit pulse and step input: (a), (b), and (c) basic patterns of response

   where $\omega(B) = \omega_0 - \omega_1 B$ is a polynomial of first order in B.

   In this refined model, and with the step function as input, the final level is reached in two steps. If $\omega_0 = 0$ the response is as in (a) but 1 time unit delayed.
3. Figure 1(c): *gradual approach to equilibrium*:

$$\begin{aligned} f(I_t) &= \omega_0(I_t + \delta I_{t-1} + \delta^2 I_{t-2} + \cdots) \\ &= \omega_0(I_t + \delta B I_t + \delta^2 B^2 I_t + \cdots) \\ &= \omega_0(1 + \delta B + \delta^2 B^2 + \cdots) I_t \\ &= \left[\frac{\omega_0}{(1 - \delta B)}\right] I_t. \end{aligned} \tag{10}$$

   In this basic type of response the final level is reached only gradually.

The above three types of response are special cases of the *general response*

$$\begin{aligned} f(I_t) &= v(B) I_t \\ &= \left[\frac{\omega(B)}{\delta(B)}\right] I_t, \end{aligned} \tag{11}$$

where $\omega(B) = \omega_0 - \omega_1 B - \cdots - \omega_s B^s$ and $\delta(B) = 1 - \delta_1 B - \cdots - \delta_r B^r$ are polynomials in B. The

corresponding model,

$$y_t = f(I_t) + n_t, \tag{12}$$

is called an intervention model of order r. $v(B)$ is the transfer function containing infinitely many parameters in general. $\omega(B)/\delta(B)$ is a parsimonious "rational lag representation" of the transfer function $v(B)$ containing only $s + r + 1$ parameters [2].

The noise part of the model is usually obtained from the preintervention period in the same way as the ordinary ARIMA model. Standard software such as SAS or BMDP (*see* **Software, Biostatistical**) allows maximum likelihood estimation of ARIMA models and intervention models. Inspection of the data may suggest a pattern by which the known event has changed the series. Additional help may be obtained by inspection of the residuals from the corresponding model. A different way to obtain a model for the response consists in postulating one or several expected "types of change" and studying if the data provide evidence for a particular type of change.

Example

During an investigation concerned with the relationship between air pollution and respiratory diseases the environmental accident of "Schweizerhalle" occurred. In that investigation a series of medical data had been collected during about one year: the daily number of respiratory symptoms per child in a randomly selected group of preschool children (called "SYMPTOMS"). On November 1, 1986, a Sandoz storehouse containing chemical substances burned down in Schweizerhalle, located near Basle. After many people experienced symptoms and, additionally, when dead fish appeared in the Rhine, public pressure demanded investigation of possible health effects. In addition to studies specially set up for this purpose, it seemed recommendable to analyze the ongoing study with regard to the question of whether health effects could be discovered on the date of the accident.

For the preaccident period of the series SYMPTOMS an *AR(1) model* was identified. Figure 2 illustrates the process of intervention model building. Three intervention models of increasing complexity are fitted to the series SYMPTOMS in a way that each

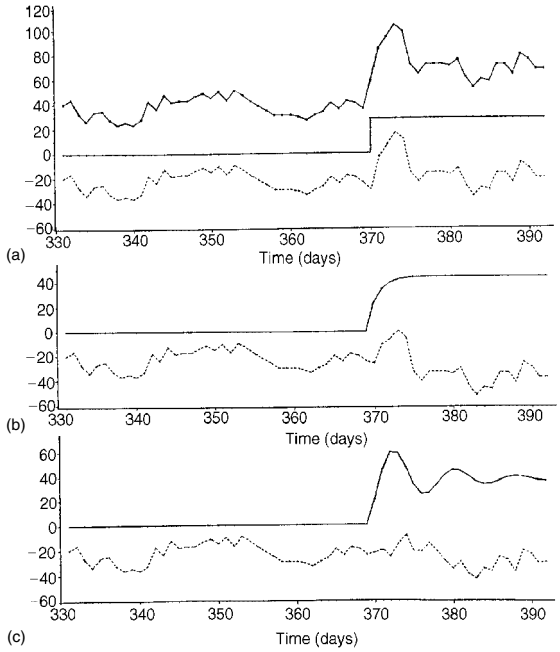

Figure 2 Three intervention models of increasing complexity. (a) Model of order zero. Upper curve: series y_t (SYMPTOMS). Second curve: "explained" part u_t. Third curve: noise series $n_t = y_t - u_t$ (shifted downwards). (b) Model of order one. Upper curve: "explained" part u_t. Lower curve: noise series $n_t = y_t - u_t$ (shifted downwards). (c) Model of order two. Same arrangement as in (b). Curves are shown multiplied by 100. Explanation in the text

additional parameter allows for a refined explanation of the data. The simplest model is as follows:

1. Intervention model of *order zero*,

$$y_t = \omega_0 s_t + n_t, \tag{13}$$

where s_t is the unit step function. The estimated parameters of this model are shown in Table 1 below the univariate model. Figure 2(a) shows, in the second row, the estimated function $\omega_0 s_t$ (the point labeled 370 corresponds to the date of the accident). This model gives a better fit to the data ($\sigma_a^2 = 0.00454$) than the univariate model ($\sigma_a^2 = 0.00476$). However, this simplified model does not fully represent all characteristic properties of the series: it predicts, for example, that the final level is reached immediately. It is

Table 1 Summary of intervention models

Model type	Estimated parameters ± se	Residual variance
Univariate	$\phi_1 = 0.91 \pm 0.02$ $\mu = 0.368 \pm 0.039$	0.00476
Intervention (order 0)	$\phi_1 = 0.87 \pm 0.03$ $\mu = 0.343 \pm 0.026$ $\omega_0 = 0.274 \pm 0.058$	0.00454
Intervention (order 1)	$\phi_1 = 0.87 \pm 0.03$ $\mu = 0.331 \pm 0.027$ $\omega_0 = 0.239 \pm 0.065$ $\delta_1 = 0.46 \pm 0.16$	0.00441
Intervention (order 2)	$\phi_1 = 0.87 \pm 0.03$ $\mu = 0.334 \pm 0.026$ $\omega_0 = 0.203 \pm 0.033$ $\delta_1 = 1.21 \pm 0.07$ $\delta_2 = 0.75 \pm 0.07$	0.00420

therefore natural to consider the following more elaborate model.

2. Intervention model of *order one*,

$$y_t = \omega_0(1 - \delta_1 B)^{-1} s_t + n_t. \qquad (14)$$

The parameters are given in Table 1 and the corresponding curves are presented in Figure 2(b). The residual variance decreases to $\sigma_a^2 = 0.00441$. This model allows for gradually reaching the final level [upper curve of Figure 2(b)]. However, the additional parameter δ_1 has a relatively large standard error. In addition, one recognizes from the lower curve of Figure 2(b) that the noise series n_t still has an unexplained "bump". This suggests introducing an additional refinement of the model.

3. Intervention model of *order two*,

$$y_t = \omega_0(1 - \delta_1 B - \delta_2 B^2)^{-1} s_t + n_t. \qquad (15)$$

The "explained" part of the model $u_t = \omega_0(1 - \delta_1 B - \delta_2 B^2)^{-1} s_t$ may be rewritten

$$(1 - \delta_1 B - \delta_2 B^2) u_t = \omega_0 s_t, \text{ or}$$

$$u_t - \delta_1 u_{t-1} - \delta_2 u_{t-2} = \omega_0 s_t \text{ or}$$

$$u_t = \delta_1 u_{t-1} + \delta_2 u_{t-2} + \omega_0 s_t. \qquad (16)$$

This second-order difference equation may represent vibrations of discrete systems (in analogy to the differential equations of order two in continuous physical systems). The lowest part of Table 1 shows the estimated parameters of this model. The parameters ω_0, δ_1 and δ_2 have small standard errors and the residual variance drops to $\sigma_a^2 = 0.00420$. The "bump" in the noise series n_t [lower curve of Figure 2(c)] has disappeared.

The upper curve of Figure 2(c), u_t, shows the characteristic behavior of a "damped vibration". The final increase of the series over the level of the preaccident period is estimated by the gain $g = \omega_0(1 - \delta_1 - \delta_2)^{-1} = 0.376$, i.e. an increase of approximately 0.38 respiratory symptoms per child per day. The introduction of additional parameters into the model did not reduce the residual variance any further. Thus, identification of a sequence of models of increasing complexity showed that the response of the series SYMPTOMS after the accident of Schweizerhalle may be parsimoniously represented by an intervention model of second order. The model corresponds to what is known in continuous physical systems as a "damped vibration". After an initial overshoot, the series settles down to a new equilibrium at a higher level.

The results obtained support the hypothesis that the number of symptoms per child increased after the accident. The identified intervention model states that after an initial overshoot following the accident the series settles down to a new level. Unfortunately, data were not available for a longer period after the accident; thus, there is a possibility that "return to normal" could have been missed. The question of

whether under the impression of the accident more symptoms were recorded (than were actually present) cannot be answered entirely satisfactorily. However, other studies conducted in this context point toward an increase in respiratory symptoms in the general population. A more detailed presentation of this study may be found in [6].

Remarks

Extensions of models as described and illustrated above are possible: intervention analysis with input consisting of multiple pulses, for example, may be used to analyze questions such as "Are there more deaths due to infarction in years with influenza A than in years without?" The sequence of pulses then represents years with influenza A.

Responses to an intervention need not to occur instantaneously; a preventive program may show an effect eventually after a delay. In addition, effects of preventive programs need not show a "permanent" effect. Decomposing the dummy input into a short-term and a long-term component may help to decide if an effect is only transient. Outliers in ARIMA time series may be detected and removed by introducing corresponding pulse inputs.

A complementary method to intervention analysis is **forecasting**: a forecast obtained from data before the intervention may be compared with actual data obtained after the intervention [3].

Literature

A nontechnical introduction to intervention analysis is given by McCleary & Hay [8]. The classical reference to ARIMA models by Box & Jenkins [1] does not include intervention analysis. The authoritative presentation of intervention analysis is that of Box & Tiao [2]. Jenkins [7] provides instructive case studies in the fields of business, industry, and economics. A review, examples, and references of studies concerned with intervention analysis in medicine may be found in [5]. Applications and references of studies concerned with intervention modeling of *single patient* time series are presented in [4].

References

[1] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, Rev. Ed. Holden-Day, San Francisco.

[2] Box, G.E.P. & Tiao, G.C. (1975). Intervention analysis with applications to economic and environmental problems, *Journal of the American Statistical Association* **70**, 70–79.

[3] Box, G.E.P. & Tiao, G.C. (1976). Comparison of forecast and actuality, *Applied Statistics* **25**, 195–200.

[4] Crabtree, B.F., Ray, S.C., Schmidt, P.M., O'Connor, P.J. & Schmidt, D.D. (1990). The individual over time: Time series applications in health care research, *Journal of Clinical Epidemiology* **43**, 241–260.

[5] Helfenstein, U. (1996). Box-Jenkins modelling in medical research, *Statistical Methods in Medical Research* **5**, 3–22.

[6] Helfenstein, U., Ackermann-Liebrich U., Braun-Fahrländer, Ch. & Wanner, H.U. (1991). The environmental accident of "Schweizerhalle" and respiratory diseases: a time series analysis, *Statistics in Medicine* **10**, 1481–1492.

[7] Jenkins, G.M. (1979). *Practical Experiences with Modelling and Forecasting Time Series*. Gwilym Jenkins & Partners (Overseas) Ltd, St. Helier.

[8] McCleary, R. & Hay, R.A. (1980). *Applied Time Series Analysis for the Social Sciences*. Sage, Beverly Hills.

ULRICH HELFENSTEIN

Interviewer Bias

Interviewer bias is a type of information **bias** (*see* **Bias in Observational Studies**; **Bias, Overview**) that arises when an interviewer consciously or unconsciously elicits inaccurate information from study subjects. Interviewer bias can result in **differential error**, which can seriously distort disease–exposure **associations**, if the interviewer is aware of the disease status and exposure hypothesis in a **case–control study**, or if the interviewer is aware of the exposure status and outcome hypothesis in a **cohort study**. In the former case, the interviewer may probe more deeply for evidence of exposure among cases than among **controls**. In the latter case, the interviewer may try to elicit evidence of health effects more assiduously in exposed than in unexposed cohort members. Methods used to minimize interviewer bias include providing structured questionnaires (*see* **Questionnaire Design**), training interviewers to follow a fixed pattern of questioning, and, where possible, keeping interviewers unaware of the disease status and exposure hypotheses of greatest interest in case–control studies, and unaware of exposure status and health outcome hypotheses of greatest interest in cohort studies (*see* **Blinding or Masking**).

(*See also* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Interviewing Techniques**).

MITCHELL H. GAIL

Interviewing Techniques

Interviewers have a variety of roles and responsibilities in conducting a **survey**. Primarily, they are responsible for the collection of data by administering a data collection instrument (*see* **Questionnaire Design**). Other roles include conducting screening interviews to ensure the respondent selected meets the requirements of the survey design plan, gaining respondent cooperation, accurately coding and editing data, and representing the survey sponsor to the public. Twenty years ago, interviewer techniques were almost exclusively designed for household, face-to-face interviews. In recent years, the advent of **computer-assisted interviewing** (CAI), which includes computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI), have had a marked effect on the techniques used by survey interviewers. The following standard techniques, as well as techniques used specifically for CAI, are based on contemporary interviewer training manuals [1, 2].

Introducing the Survey and Gaining Cooperation

An interviewer's first contact with the respondent is crucial for several reasons. First, the purpose and importance of the survey is communicated. Secondly, rapport between the interviewer and respondent is established. Thirdly, and perhaps most important, cooperation of the respondent to participate is usually received.

The interviewer must convey the purpose and importance of the study in a simple and direct manner. This is sometimes awkward if the interviewer must read verbatim a long and complex script. After reading the introduction, the interviewer must be able to paraphrase effectively what was read so that the respondent begins to feel and understand that it is important to participate. Discussing participation in a confident, friendly, empathetic tone can help to eliminate hesitancy on the respondent's part. Interviewers can allay concerns by suggesting that they start the interview and reiterating that the respondent does not have to answer any question that may be too personal. This is an effective way to coach a reluctant respondent to participate and allows the respondent to feel that he/she has some control over the interview situation.

Typically, respondents are given assurances of **confidentiality** and, if under the auspices of an official agency, authorization for the survey during the first contact. While this may seem like a technical requirement, respondents often feel more at ease and are more willing to respond once they know the purpose of the survey, and they understand that the information they are providing will be held strictly confidential. Reluctant respondents often are concerned about how they were selected. Virtually every interviewer has had a respondent ask: "How did you get my name?" If the survey is a **random sample**, then interviewers need to explain the process in a clear and concise manner. Interviewers should listen carefully to all questions and comments from the respondent and should answer only what is asked.

Handling respondents who refuse to participate (*see* **Nonresponse**) is probably the most challenging component of an interviewer's job. In any survey, there are respondents who simply do not want to be interviewed. Some respondents refuse outright and others indicate refusal indirectly by avoiding the interview or constantly rescheduling. While a relatively high number of initial interview contacts result in rescheduling, few refusals actually do occur in a well-planned survey. The interviewer is the major influence on the motivation of the respondent and on the quality of the responses received. Interviewers can subtly communicate their interest in the study, their enthusiasm about their work, and even their positive feelings about the respondent, all of which increase survey cooperation.

Interviewers are trained to convert a potential refusal into cooperation. A good interviewer will use techniques that reduce covert negative issues likely to cause refusals. Respondents may be mistrustful of the interviewer, may see participating in a survey as an invasion of their privacy, or may not understand or believe assurances of confidentiality and anonymity. Respondents may also feel threatened about the survey's topic, especially if it is sensitive. Interviewers can discuss the respondent's concerns in an open, relaxed manner and usually can convince them to start the interview.

Interviewing a Survey Respondent

Ask Each Question Exactly as Worded. This is a standard long held in survey design. If questions are not read exactly as worded, they may not yield comparable results. Research has shown that even minor changes in wording can change response distributions. It is true, however, that interviewers are allowed to depart from the standard wording, but that is only after they have first asked the standard form of the question and attempted to get an answer.

Ask Every Question. Although the answer to a given question may seem obvious, interviewers must ask the question and obtain a response. Occasionally, the respondent provides an answer which applies to a question asked later in the interview. In this case, the interviewer should verify the answer to the question.

Maintain Positive Rapport. Offering the respondents some assurance that they are doing well and that their responses are valuable can increase their willingness to participate. Comments such as "Yes, I see" show the respondents that their answers are important and interesting to the interviewer. This can stimulate a respondent to talk further and to engage more actively in the survey process.

Use Effective Probing Techniques. When the respondent's answer does not meet the question's objective, or when a respondent seems to have misunderstood the question, interviewers need to probe for clarification or correction. One of the most common probes is to ask the respondent to repeat the answer – this often prompts the respondents to expand on the answer, offer more information, or correct an answer. Another common probe technique is to reread the question. This often results in the respondent paying closer attention to the question and revising the answer.

In general, interviewers should not probe by paraphrasing a question, offering additional explanations (unless this is provided for in the interviewer's manual), or assuming responses that may seem obvious from prior answers. Rather, probing should take the form of more general questions ("What do you mean?") or should be an attempt to improve the specificity of a response ("Could you put your answer in terms of days?").

The danger of probing is that the interviewer can influence the respondent's answer and approach to other questions. Also, it can unnecessarily prolong the interview and convey a more conversational, informal tone. It is important for the interviewer to probe only when necessary and then return to asking survey questions. Probes should be neutral and not convey a right or wrong answer. For example, if the respondent says she had between five and eight visits to the doctor, a **biased** probe would be "So, would you say five is about right?" and a neutral probe would be "Would you say the number of visits was closer to five or closer to eight?" Interviewers can also probe using statements and questions that indicate their uncertainty, such as "I don't know quite what you mean," or "Which figure would you say comes closer?" It is important when using these sorts of probes to keep the tone positive, and not intimate that the respondent gave a wrong answer. In fact, some interviewers find it useful to probe with a suggestion that it is the interviewer who is misunderstanding (e.g. "I'm not sure what you mean by that; could you tell me a little more?").

One exception to the probing guidelines above is when an interviewer is conducting a cognitive interview. Cognitive interviews are typically used in pretests to identify flawed questions prior to fielding the survey. Cognitive interviewers specialize in applying cognitive psychology principles to understanding the survey response process. These interviewers depart freely from the questionnaire to probe intensively. Probes are not used to help record or code information given by the respondent; rather, they are qualitative in nature and used to determine whether the respondent is having difficulty comprehending the question, recalling the information asked, or using an inappropriate response strategy (like estimating or guessing). Examples of typical probes used in cognitive interviews include "Can you paraphrase for me what you think this question is asking?", "Tell me how you arrived at your answer?", and "How sure are you that your answer is correct?"

Interviewer's Manner and Nonverbal Communication

Whether in person or on the telephone, interviewers must uniquely combine a friendly approach with an official, businesslike manner. Respondents often begin to talk about the subject matter of a questionnaire, and if the interviewer becomes distracted or

becomes too engaged in social conversation with the respondent, it is often difficult to return to asking survey questions.

It is important that interviewers maintain an objective attitude. They should never indicate a personal opinion about a reply or a survey topic. Furthermore, they need to be acutely aware of facial expressions, mannerisms, tone of voice, and other spontaneous reactions to respondent answers. Expressions of surprise, amusement, disapproval, or even sympathy may cause respondents to give untrue answers or to withhold information. Objectivity is the most effective method for putting respondents at ease and making them feel free to answer questions honestly.

For telephone interviews, it is essential that the interviewer's tone be pleasant and friendly, that they speak clearly, and that they are familiar enough with the instrument to avoid long pauses or delays. Pauses on the telephone are awkward, and may give the impression that the interviewer is waiting for an explanation from the respondent. Hesitation and expressed uncertainty about what to ask may create a negative impression in a telephone contact that would not have necessarily occurred in a face-to-face interview.

Telephone surveys, in general, are administered more quickly than a face-to-face interview because interviewers tend to read the questions faster and respondents tend to answer quicker and are less inclined to engage in social discourse. However, rushing can also give the appearance of lacking confidence and may cause the listener to misunderstand. Telephone interviewing should be confined to shorter data-collection instruments (*see* **Telephone Sampling**).

Computer-Assisted Interviewing (CAI)

Field data collection using computers is a new approach in many surveys. Clear benefits of CAI are that it eliminates editing responsibilities of the interviewer and keying of questionnaire data, resulting in quicker availability of results. Using a computer to collect interview data offers some important advantages to actual interviewer techniques as well.

First, the computer presents the correct sequence of questions based on the information and the responses already entered. This relieves the interviewer of a burden as they do not need to follow skip instructions, check items, and so forth. The computer also checks responses to ensure that all applicable items are answered appropriately. For example, where possible answers to a question are 1 (Yes) or 2 (No), the computer will reject other answers, and prompt the interviewer either to reask the question or to check the entry they made. Clearly, the use of computers is expected to help interviewers do their job more efficiently by eliminating tedious paperwork and freeing them to concentrate on actual data collection and building rapport with respondents.

One advantage of CAI interviews is that the program can easily provide on-screen instructions or other helpful information. For example, the screen may display previously provided names of family members so that the interviewer can refer to them by name in follow-up questions. This helps the interviewer administer the questions in a more friendly, casual manner. During the interview, the disadvantage to a lot of on-screen instructions and information is that the interviewer must not devote much time to reading the screen and trying to familiarize her/himself with instructions and other screen entries. This can hurt the interviewer's credibility as a well-trained professional, and can serve to distract and/or disengage the respondent. The interviewer's training must thoroughly familiarize the interviewer with the screens and instrument flow.

Another advantage of CAI is that each displayed screen can have an accompanying HELP screen, readily providing interviewers with screen-specific information, definitions, and explanations. This preserves the interviewer's sense of confidence, and improves data quality.

In CAPI surveys, interviewers have to be sensitive to the respondents' perception of the use of a computer to enter data. Some respondents perceive the computer as a means of entering their data into some sort of an open "information highway", and need to be assured that computers actually go further to protect confidential data. The computer may also serve as a barrier to good eye contact between the interviewer and respondent. In addition, if the interviewer is not comfortable with the computer program, they may spend too much time attending to the computer to the exclusion of the respondent. One point that interviewers should remember is that the respondent usually cannot see the screen that the interviewer is looking at, in contrast to being able to see a paper and pencil questionnaire. Also, because the interviewing

program will perform internal consistency edits based on information previously and subsequently provided, the programs often identify inconsistent answers that the interviewer has to probe about. If this is not done in a sensitive manner (e.g. "I must have entered something wrong – let me ask that question again."), respondents may begin to feel defensive about their answers and disengage from the survey process.

Last, interviewers may have more difficulty showing cards, life history calendars, or other tools that they need to give to the respondent during the interview. Depending on where the computer is set up, if it is sitting on the interviewer's lap and the respondent is not close by, too much distance may be created which will make these survey aides awkward to use. Also, if interviewers are forced to stand while conducting CAPI interviews, they lose the mobility to use ancillary interviewing tools smoothly.

Conclusion

Good interviewing techniques apply to all modes of interviewing: face-to-face, telephone, CATI, and CAPI. Each interviewing mode also makes unique demands on the interviewer. The current trend towards increasing use of CAI requires interviewers to fortify themselves with all the standards of good interviewing techniques, and develop automation skills as an additional requirement. Interviewers are also being required to administer a survey for an increasingly resistant respondent pool, and to manipulate the dynamics of the interviewing situation to maintain respondent cooperation and interest. Newer and more effective training strategies need to be developed to equip interviewers with all the skills demanded by a fast-paced, changing survey environment.

References

[1] US Bureau of the Census (1996). HIS-100 C CAPI Manual for HIS Field Representatives, *National Health Interview Study*. US Bureau of the Census, Washington..

[2] Westat, Inc. (1996). *Interviewer Training Manual, Race/Ethnicity Study*. Westat Inc., Rockville.

(*See also* **Interviewer Bias**)

SUSAN SCHECHTER &
ADRIENNE ONETO QUASNEY

Inverse Gaussian Distribution

In 1828 the English botanist Robert Brown described observations made on the motion of plant pollen immersed in water. He found a swimming, dancing motion from pollen from many different plants and he extended his research to include particles of fossilized plants and mineral specimens. Apparently, he believed that he had discovered a new type of particle. His work led to the realization that the motion was a physical phenomenon rather than biologic.

Bachelier (1900) and Einstein (1905) derived the normal distribution as the model for **Brownian motion**. Wiener (1923) gave the theory of a measure on the path space. Schrodinger (1915) considered Brownian motion with a positive drift, and obtained the distribution of the first passage time to describe the position of a particle performing Brownian motion. Tweedie [5] noticed the inverse relationship between cumulant **generating functions**, and proposed the name inverse Gaussian distribution. Wald [6] obtained the distribution as an approximation of the sample size distribution in a **sequential** probability ratio test. The distribution is sometimes known as Wald's distribution.

The probability density function of the inverse Gaussian distribution, denoted by $\mathrm{IG}(\mu, \lambda)$, is

$$f_x(x : \mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left[\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right],$$

with $\mu > 0$, $\lambda > 0$, and $x > 0$. The mean is μ and the variance is μ^3/λ. The unimodal density function, a member of the **exponential family**, is skewed to the right (*see* **Skewness**). Its shape resembles other skewed density functions such as the **lognormal, Weibull**, and **gamma**. It can be obtained from the Wiener process $X(t)$ with positive drift ν and variance parameter σ^2 (*see* **Brownian Motion and Diffusion Processes**). Starting at zero, the time, T, for $X(t)$ to reach the barrier $a(a > 0)$ for the first time is called the *first passage time* and has an inverse Gaussian distribution with $\mu = a/\nu$ and $\lambda = a^2/\sigma^2$.

Unlike the Weibull or gamma distributions, the inverse Gaussian distribution allows exact sampling distributions. The **sufficient statistics** $\overline{X}$ and $T = \sum(1/X - 1/\overline{X})$, or a one-to-one function of $\overline{X}$ and T, are the basic quantities which are used in all of the **hypothesis tests, confidence intervals**, etc. $\overline{X}$ is also inverse Gaussian and λT is independently distributed as a **chi-square distributed** variable with $n - 1$ **degrees of freedom**.

Statistical methods based on the distribution have been developed to include the point and interval **estimation** of parameters, prediction intervals, estimation of the cumulative distribution function (cdf), analysis of **residuals** (one-way and two-way), **regression** analysis, and reliability analysis (*see* **Survival Analysis, Overview**).

The distribution has been used to describe phenomena in many of the sciences. In the biosciences, Sheppard & Savage [4] made use of the distribution to describe the length of time a particle remains in the blood. Since then it has been used in modeling maternity data, crop field size, shelf life, and many other types of data. Eaton & Whitmore [2] modeled the length of stay in a hospital as an inverse Gaussian variable.

For additional information, see [1] and [3].

References

[1] Chhikara, R.S. & Folks, J.L. (1989). *The Inverse Gaussian Distribution*. Marcel Dekker, New York.

[2] Eaton, W.W. & Whitmore, G.A. (1977). Length of stay as a stochastic process: a general approach and application to hospitalization for schizophrenia, *Journal of Mathematical Sociology* **5**, 273–292.

[3] Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 1. Wiley, New York.

[4] Sheppard, C.W. & Savage, L.J. (1951). The random walk problem in relation to the physiology of circulatory mixing, *Physical Review* **83**, 489–490.

[5] Tweedie, M.C.K. (1957). Statistical properties of inverse Gaussian distributions I, *Annals of Mathematical Statistics* **28**, 362–377.

[6] Wald, A. (1944). On cumulative sums of random variables, *Annals of Mathematical Statistics* **15**, 283–296.

J. LEROY FOLKS

Inverse Probability Weighting in Survival Analysis

Introduction

Modern epidemiologic and clinical studies aimed at analyzing a time to an event endpoint T routinely collect, in addition to (possibly **censored**) information on T, high-dimensional data often in the form of baseline (i.e. time-independent **covariates** $V(0)$) and **time-dependent covariates** $V(t), t > 0$, measured at frequent intervals. Scientific interest, however, often focuses on a low-dimensional functional $\beta = \beta(F_X)$ of the distribution F_X of the (intended) full data $X = \left(T, \overline{V}(T)\right)$ where $\overline{V}(t) \equiv \{V(u) : 0 \leq u \leq t\}$. Inverse probability weighted augmented (AIPW) estimators of β meet the analytic challenge posed by these high-dimensional data because they are **consistent** and asymptotically normal (CAN) under models that do not make assumptions about the parts of F_X that are of little scientific interest. As such, they are not subject to biases induced by **misspecification** of models for these secondary parts of F_X.

AIPW estimators were originally introduced by Robins and Rotnitzky [21] as part of a general **estimating function** methodology in coarsened, that is, incompletely observed, data models under **nonparametric** or **semiparametric** models for arbitrary full-data configurations X when the data are **coarsened at random** (CAR) [5, 6, 8] and the coarsening, that is, censoring or missingness, mechanism is either known or correctly modeled (*see* **Missing Data**). The AIPW estimators generalized and made efficient the nonaugmented inverse probability weighted estimators proposed by Koul, Susarla, and van Ryzin [12], and Keiding, Holst, and Green [10]. Robins and Rotnitzky [21] derived their methodology drawing from the modern theory of semiparametric **efficiency** due to Bickel, Klaassen, Ritov, and Wellner [3], Newey [14], van der Vaart [39, 40] among others. In this article, we restrict the discussion to coarsened data in the form of right-censored failure-time data. In this setting, the full data are $X = \left(T, \overline{V}(T)\right)$, the observed data are $Y = \left(\widetilde{T} = \min(T, C), \Delta = I(T \leq C), \overline{V}\left(\widetilde{T}\right)\right)$, where C is a censoring variable, and C and T are continuous positive random variables. Furthermore, the CAR assumption is equivalent to

$$\lambda_C(u|X) = \lambda_C\left(u|\overline{V}(u)\right) \text{ for all } u \geq 0, \qquad (1)$$

where $\lambda_C(u|\cdot) = \lim_{h \to 0^+} \Pr(u \leq C < u + h|\cdot, C \geq u, T \geq u)$ is the cause-specific hazard for censoring. Therefore, the coarsening mechanism is determined by the stochastic process $G \equiv G(\cdot)$, where $G(u) \equiv \exp\left\{-\int_0^u \lambda_C\left(t|\overline{V}(t)\right) \mathrm{d}t\right\}$.

AIPW Estimators Under CAR

Motivation: the Curse of Dimensionality

When the data are CAR, the **likelihood** $\mathcal{L}_n(F_X, G)$ based on n i.i.d. copies of Y factorizes as $\mathcal{L}_n(F_X, G) = \mathcal{L}_{1,n}(F_X)\mathcal{L}_{2,n}(G)$, where G denotes the coarsening mechanism, that is, the conditional distribution of the observed data Y given the full data X, and F_X is the cumulative distribution function of X. Thus, for models in which G and F_X are variation independent, any method that obeys the **likelihood principle** must result in the same inference about β regardless of whether G is known, completely unknown or known to follow a model. However, under non- or big semiparametric models for F_X, Robins and Ritov [20] have shown that due to the curse of dimensionality, with high- dimensional coarsened at random data, any method of inference that obeys the likelihood principle and thus ignores G must perform poorly in realistic sample sizes as the following example illustrates.

Example Suppose that no covariates $V(t)$ are measured for any $t > 0$ and, to simplify the notation, let W denote $V(0)$. Under CAR, $\mathcal{L}_n(F_X, G) = \mathcal{L}_{1,n}(F_X)\mathcal{L}_{2,n}(G)$, where $\mathcal{L}_{1,n}(F_X) = \prod_{i=1}^n f_{T|W}\left(\widetilde{T}_i|W_i\right)^{\Delta_i}\left(1 - F_{T|W}\left(\widetilde{T}_i|W_i\right)\right)^{1-\Delta_i} f_W(W_i)$. Suppose that we are interested in estimating $\beta = \Pr(T \leq t)$ for a fixed t. Because $\beta = \beta(F_X) = E_{F_W}\left\{F_{T|W}(t|W)\right\}$, its MLE $\widehat{\beta}$ is equal to $\beta\left(\widehat{F}_X\right)$, where $\widehat{F}_X$ is the **Maximum likelihood** estimator (MLE) of F_X. Since the MLE of F_W is the empirical cdf of W, then $\widehat{\beta} = n^{-1}\sum_i \widehat{F}_{T|W}(t|W_i)$. The MLE $\widehat{F}_{T|W}$ of $F_{T|W=W_i}$ is given by the Kaplan–Meier estimator based on the subsample $\{Y_i : W = W_i\}$. If W is continuous, each subsample consists of one observation, so the **nonparametric MLE** of $F_{T|W=W_i}$ assigns

probability 1 to $\widetilde{T}_i$; however, when $\Delta_i = 0$, the NPMLE of $F_{T|W=W_i}$ assigns probability 0 to the interval $\left[0, \widetilde{T}_i\right]$ and it is undefined on $\left(\widetilde{T}_i, \infty\right)$. Thus, when W is continuous the NPMLE of $F_{T|W=W_i}$ is undefined for some observed values of W_i and hence the MLE of β is also undefined. One could assume that $F_{T|W}$ was smooth in W and use multivariate smoothing techniques (*see* **Smoothing Hazard Rates**). However, when W is high dimensional, $F_{T|W}$ would not be well estimated with the moderate sample sizes found in practice, because no two units would have values of W close enough to allow the borrowing of information needed for smoothing. Thus, unrealistically large sample sizes would be required for any estimator of β to have an approximately centered normal sampling distribution with variance small enough to be of substantive use.

AIPW estimators depend on a model for G and thus violate the likelihood principle, yet they yield estimators that are well behaved with moderate sample sizes. Locally efficient AIPW estimators of β (defined in the next section) simultaneously correct for bias due to dependent censoring attributable to the covariate process $V(t)$ and recover information from the censored observations by nonparametrically exploiting the correlation between the process $V(t)$ observed up to the censoring time and the unobserved failure time T.

Even under CAR, because of the curse of dimensionality, well-behaved estimators of β in finite samples do not exist unless one imposes additional restrictions on either the coarsening mechanism G or on the non- or semiparametric model for F_X. Hence, the best that can be hoped for is an estimator that is CAN under the CAR assumption (1) when either (but not necessarily both), a lower-dimensional model for G or a lower-dimensional model for F_X is correct. Such an estimator, when it exists, is called *doubly robust* (DR). Scharfstein, Rotnitzky, and Robins [29], Robins [18], Robins, Rotnitzky, and van der Laan [25] and van der Laan, and Robins [38] provide a broad theory of double **robustness** in CAR models. Using this theory, these authors show that the locally efficient AIPW estimators (described in the next section) are doubly robust. Robins and Rotnitzky [22] provide a summary of known results on double-robust estimation, including estimation in nonignorable models, that is, when CAR is not assumed.

Locally Efficient AIPW Estimation

Suppose that $\beta = \beta(F_X)$ is a smooth $k \times 1$ parameter under a non- or semiparametric model $\mathcal{M}_{\text{Full}}$ for F_X. That is, $\beta(F_X)$ is estimable at rate $\sqrt{n}$ under all laws F_X in model $\mathcal{M}_{\text{Full}}$ when X is fully observed for all n sample units. The general **algorithm** for the construction of locally efficient AIPW estimators of a $k \times 1$ vector $\beta(F_X)$ starts with the specification of a full-data *orthogonal* estimating function $D(\beta, \rho) = d(X; \beta, \rho)$ for β. This is a $k \times 1$ vector function of the full data X, of β and, possibly, of a **nuisance parameter** ρ, such that each component of $D(\beta(F_X), \rho(F_X))$ has mean zero and covariance zero with any nuisance score under F_X, for all F_X in $\mathcal{M}_{\text{Full}}$. We need to allow the orthogonal estimating equation to possibly depend on a nuisance parameter ρ so as to make the general methodology applicable to a broad class of estimation problems. For instance, in the Cox proportional hazards model with time independent covariates, $\rho \equiv \rho(\cdot)$, where $\rho(u)$ is the mean of the covariate among subjects who fail at time u.

The full data-estimating function $D(\beta, \rho)$ gives rise to the observed data-estimating function

$$U\{D(\beta, \rho), G\} - A\left(h_{F_X}, G\right) \tag{2}$$

The term $U\{D(\beta, \rho), G\}$ is an inverse probability weighted estimating function defined as

$$U\{D(\beta, \rho); G\} = \frac{\tau D(\beta, \rho)}{G(T^*)}, \tag{3}$$

where T^* is the minimum time such that $D(\beta, \rho)$ is observed and $\tau = I(T^* < C)$ is the indicator that $D(\beta, \rho)$ is observed. The critical point of inverse probability weighting is that when $\Pr(\tau = 1|X) > 0$, the estimating function is an **unbiased** *inverse probability* estimating function because, under CAR, $\Pr(\tau = 1|X) = G(T^*)$ and, thus, $E_G[U\{D(\beta, \rho); G\}|X] = D(\beta, \rho)$.

The second term in (2) is a mean zero augmentation term which, for any function $h\left(u, \overline{V}(u)\right)$, is defined as

$$A(h, G) = \int_0^{\widetilde{T}} \left[h\left(u, \overline{V}(u)\right)/G(u)\right] \mathrm{d}M_C(u), \tag{4}$$

where $\mathrm{d}M_C(u) = I\left(\widetilde{T} = u, \Delta = 0\right) - I\left(\widetilde{T} \geq u\right) \lambda_C\left(u|\overline{V}(u)\right) \mathrm{d}u$. The function h_{F_X} in (2) depends on

F_X and G and is defined as

$$h_{F_X}\left(u, \overline{V}(u)\right) = -E\left[U D(\beta, \rho) | \overline{V}(u), T \geq u\right]. \tag{5}$$

A doubly robust, locally efficient AIPW estimator $\widehat{\beta}(D)$ that uses $D(\beta, \rho)$ is the solution to

$$\sum_{i=1}^{n}\left[U\left\{D_i(\beta, \widehat{\rho}(\beta)); \widehat{G}\right\} - A_i\left(h_{\widehat{F}_X}, \widehat{G}\right)\right] = 0, \tag{6}$$

where $\widehat{F}_X$ and $\widehat{G}$ are the maximum likelihood estimators of F_X and G under parametric or semiparametric working models $\mathcal{M}_{\text{work}} \subset \mathcal{M}_{\text{Full}}$ for F_X and $\mathcal{G}_{\text{work}}$ for G and $\widehat{\rho}(\beta)$ is an estimator of ρ such that $\widehat{\rho}(\beta)$ evaluated at the true β converges at an appropriate rate (usually $n^{1/4}$) to $\rho(F_X)$ if either, but not necessarily both, working models are true; see van der Laan and Robins, 2003, for construction of $\widehat{\rho}(\beta)$. The estimator $\widehat{\beta}(D)$ has the following properties.

(a) If $\Pr(\tau = 1|X) > \sigma > 0$, $\widehat{\beta}(D)$ is doubly robust. That is, provided $\widehat{F}_X$ and $\widehat{G}$ converge at a sufficiently fast rate to F_X and G under $\mathcal{M}_{\text{work}}$ and $\mathcal{G}_{\text{work}}$ respectively, $\widehat{\beta}(D)$ is CAN in the union model that assumes that $F_X \in \mathcal{M}_{\text{Full}}$, CAR and either $F_X \in \mathcal{M}_{\text{work}}$ or $G \in \mathcal{G}_{\text{work}}$.

(b) There exists $D_{\text{opt}}(\beta, \rho)$ such that $\widehat{\beta}\left(D_{\text{opt}}\right)$ is locally semiparametric efficient in the union model of (a) at the intersection submodel where both $\mathcal{M}_{\text{work}}$ and $\mathcal{G}_{\text{work}}$ are correct. Robins and Rotnitzky [21], and van der Laan and Robins (2002) show how to derive D_{opt}.

The previous results can be derived from the results in Robins and Rotnitzky [21], which provide a general representation for the influence functions and the efficient score of estimators of smooth parameters $\beta = \beta(F_X)$ of non- or semiparametric models $\mathcal{M}_{\text{Full}}$ for distributions F_X of arbitrary full-data configurations X under parametric, semiparametric or nonparametric CAR models for the censoring or missingness mechanism.

Example (continued) When $\beta = P(T \leq t)$ for a fixed t, then $D(\beta, \rho) = I(T \leq t) - \beta$ does not depend on a nuisance parameter and, because the full-data model is nonparametric $D(\beta, \rho)$ is, up to a multiplicative constant, the unique (orthogonal) unbiased estimating function. In this setting, we have $T^* = \min(T, t)$ and

$$\begin{aligned} h_{F_X}(u, W) &= E\{I(T \leq t) - \beta | W, T \geq u\} \\ &= F_{T|W, T \geq u}(t|W, T \geq u) - \beta. \end{aligned} \tag{7}$$

To obtain a locally efficient DR estimator of β, we specify a low dimensional, for example, parametric, model $F_{T|W}(u|W; \eta)$ for $F_{T|W}(u|W)$ and compute the maximum likelihood estimator $\widehat{\eta}$ of η under the model. We leave the marginal distribution of W unrestricted so its MLE is the empirical distribution of W. In addition, we specify a low-dimensional model for $G(u)$, for example, we may postulate a Cox proportional hazards model $\lambda_C(u|W) = \lambda_0(u)\exp\left(\gamma' m(W)\right)$ and estimate γ with the Cox partial likelihood estimator $\widehat{\gamma}$ and $\lambda_0(u)$ with the Cox baseline hazard estimator $\widehat{\lambda}_0(u)$ regarding the censoring times as the outcomes and the failure times T as the censoring times for C. We then compute $\widehat{G}(u) = \prod_{\left\{i: \widetilde{T}_i \leq u, \Delta_i = 0\right\}}\left[1 - \widehat{\lambda}_0\left(\widetilde{T}_i\right)\exp\left(\widehat{\gamma}' m(W)\right)\right]$. We compute $h_{\widehat{F}_X, \widehat{G}}(u, W)$ using $\widehat{F}_{T|W}(u|W; \eta) = F_{T|W}(u|W; \widehat{\eta})$ and $\widehat{\beta}_{\text{MLE}} = n^{-1}\sum_{i=1}^{n}\widehat{F}_{T|W}(t|W_i)$. Then we solve (6) to obtain $\widehat{\beta}(D)$. Note that $\widehat{\beta}_{\text{MLE}}$ is the MLE of β under model $F_{T|W}(u|W; \eta)$. Thus, $\widehat{\beta}(D)$ is CAN, well behaved in finite samples and generally more efficient than $\widehat{\beta}(D)$ if the working model $F_{T|W}(u|W; \eta)$ holds. However, $\widehat{\beta}_{\text{MLE}}$, in contrast to the doubly robust estimator $\widehat{\beta}(D)$, is inconsistent if the model $F_{T|W}(u|W; \eta)$ is mispecified.

A Survey of Applications of the AIPW Methodology under CAR in Survival Analysis Problems

The book of van der Laan and Robins [38] contains a comprehensive treatment of the AIPW methodology for estimation in censored and missing data models and in counterfactual models for causal inference under the CAR assumption. Here, we review the literature on AIPW estimation restricting attention to censored failure-time data in noncounterfactual models.

Robins and Rotnitzky [21], Robins [16], and Robins and Finkelstein [19] constructed locally efficient AIPW estimators of the survival distribution of

T and of regression parameters of **Cox** proportional hazards models and **accelerated failure-time models** for the conditional distribution of T given baseline covariates. Robins [17] constructed AIPW estimators of median regression models for right-censored failure-time data; and Robins, Rotnitzky, and Zhao [26], and Nan, Emond, and Wellner [13] described AIPW estimation of Cox proportional hazards regression parameters with missing covariates. Robins, Rotnitzky, and Bonetti [23] described AIPW estimation of a failure-time distribution under **double sampling** with follow-up of dropouts. Hu and Tsiatis [7] used the AIPW methodology to construct estimators of a survival function from right-censored data subject to reporting delays. Zhao and Tsiatis [41–43], and Van der Laan, and Hubbard [35] constructed AIPW estimators of the quality of life adjusted survival-time distribution from right-censored data. Bang and Tsiatis [1, 2], and Strawderman [33] derived respectively AIPW estimators of a median regression model for medical costs from right-censored data and of the mean of an increasing **stochastic process**. Van der Laan, Hubbard, and Robins [36], and Quale, van der Laan and Robins [15] constructed locally efficient AIPW estimators of a **multivariate survival** function when failure times are subject to a common censoring time and to a failure-time-specific censoring respectively. In the same setting, Keles, van der Laan, and Robins [9] derived AIPW estimators that are easier to compute and almost as efficient than the Quale et al. estimators.

Robins and Rotnitzky [21] restricted their investigation to data configurations for which the full data X has a positive probability of being completely observed. Their work was later extended to censored data structures under the CAR assumption in which X is never completely observed. These extensions include the estimation of the marginal survival function of T and of regression parameters of an accelerated failure-time model for the law of T, given baseline covariates $V(0)$ from current status and/or interval censored data when the intensity function for monitoring whether T has occurred by t depends on the observed covariate history $\overline{V}(t)$ [34, 37].

AIPW Estimation without the CAR Assumption

The CAR assumption (1) implies that the data $V(t), t \geq 0$ include all the time-dependent and time-independent **prognostic factors** for failure that also predict censoring. In most studies, however, data are typically available on some but not all joint prognostic factors for censoring and survival and hence, CAR fails. Scharfstein, Robins, Eddings, and Rotnitzky [31], and Scharfstein and Robins [30] have extended the AIPW methodology to allow estimation of the marginal survivor function at a fixed t, $\beta = \Pr(T > t)$, of a discrete and continuous failure time T respectively under non-CAR models. Their work was an extension to the analysis of failure-time data of the AIPW methodology in non-CAR models for non–failure-time endpoints derived in a series of papers by Rotnitzky and Robins [27], Rotnitzky, Shcarfstein, and Robins [28], Robins, Rotnitzky, and Scharfstein [24], and Scharfstein, Rotnitzky and Robins [29]. This methodology allows the analyst to appropriately adjust for informative censoring due to measured prognostic factors while simultaneously quantifying the sensitivity of inference to nonidentifying assumptions concerning residual dependence between the failure time and censoring due to unmeasured factors. For continuous failure-time data, their approach relies on the assumption that the censoring mechanism follows the model

$$\lambda_C\left(u|\overline{V}(u), T\right) = \lambda_{0,C}\left(u, \overline{V}(u)\right) \times \exp\left\{q\left(u, \overline{V}(u), T\right)\right\} \text{ for all } u \geq 0, \quad (8)$$

where $\lambda_{0,C}\left(u, \overline{V}(u)\right)$ is an unknown function and $q\left(u, \overline{V}(u), T\right)$ is a user-specified (i.e. known) function. The function $q\left(u, \overline{V}(u), T\right)$ quantifies, for those who remain at risk at time u, the dependence measured on the hazard ratio scale between T and censoring just after u after having adjusted for prognostic factors $\overline{V}(u)$. The choice $q\left(u, \overline{V}(u), T\right) = 0$ corresponds to an assumption slightly less stringent than the CAR assumption (1). Scharfstein and Robins, arguing like in Scharfstein, Rotnitzky, and Robins [29], showed that their model, like models proposed in the **competing risks** literature without auxiliary data $V(t)$ [4, 11, 32, 44, 45] is a nonparametric, just identified model for the law of the observed data Y; that is, the function $q\left(u, \overline{V}(u), T\right)$ is not identified, but, once specified, the survivor parameter β is identified but the law of Y is not restricted. Following the lead in the competing risks

literature, these authors recommended drawing inference about β by varying $q\left(u, \overline{V}(u), T\right)$ over a plausible range, and described a useful parameterization of this function for conducting such analysis.

Model (8) is stringent enough to allow identification for β. However, as in the CAR model, the model is not stringent enough to allow well-behaved estimation of β in finite samples when the covariate process is high dimensional because the function $\lambda_{0,C}\left(u, \overline{V}(u)\right)$ cannot be estimated well due to the curse of dimensionality. In order to reduce the dimension of the unknown function $\lambda_{0,C}\left(u, \overline{V}(u)\right)$, Scharfstein and Robins [30] assumed a lower-dimensional model for $\lambda_{0,C}\left(u, \overline{V}(u)\right)$ of the form

$$\lambda_{0,C}\left(u, \overline{V}(u)\right) = \lambda^*_{0,C}(u) \exp\left\{\gamma' w\left(u, \overline{V}(u)\right)\right\}, \tag{9}$$

where $\lambda^*_{0,C}(u)$ and γ are unknown and $w\left(u, \overline{V}(u)\right)$ is a user-specified function, and described AIPW estimators of β under this model. Unlike the CAR model, this model does not admit doubly robust estimators [22].

References

[1] Bang, H. & Tsiatis, A.A. (2000). Estimating medical cost with censored data, *Biometrika* **87**, 329–343.

[2] Bang, H. & Tsiatis, A.A. (2002). Median regression with censored medical cost data, *Biometrics* **58**, 643–650.

[3] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[4] Fisher, L. & Kanarek, P. (1974). Presenting censored survival data when censoring and survival times may not be independent, *Reliability and Biometry: Statistical Analysis of Lifelength*, SIAM, Philadelphia, pp. 303–326.

[5] Gill, R.D., van der Laan, M.J. & Robins, J.M. (1997). Coarsening at random: characterizations, conjectures and counterexamples, in Proceedings of the First Seattle Symposium on Survival Analysis, pp. 255–294.

[6] Heitjan, D.F. & Rubin, D.B. (1991). Ignorability and coarse data, *Annals of Statistics* **19**, 2244–2253.

[7] Hu, P.H. & Tsiatis, A.A. (1996). Estimating the survival function when ascertainment of vital status is subject to delay, *Biometrika* **83**, 371–380.

[8] Jacobsen, M. & Keiding, N. (1995). Coarsening at random in general samoke spaces and random censoring in continuous time, *Annals of Statistics* **23**, 774–786.

[9] Keles, S., van der Laan, M.J. & Robins, J. (2004). Estimation of the bivariate survival function in the presence of time dependent covariates, in *Survival Analysis Volume of the Handbook of Statistics*, Vol.23 C.R. Rao & N. Balakrishnan, eds Elsevier, North-Holland, pp. 143–175.

[10] Keiding, N., Holst, C. & Green, A. (1989). Retrospective estimation of diabetes incidence from information in a current prevalent population and historical mortality, *American Journal of Epidemiology* **130**, 588–600.

[11] Klein, J.P. & Moeschberger, M.L. (1998). Bounds on net survival probabilities for dependent competing risks, *Biometrics* **44**, 529–538.

[12] Koul, H., Susarla, V. & van Ryzin, J. (1981). Regression analysis with randomly right censored data, *Annals of Statistics* **9**, 1276–1288.

[13] Nan, B., Emond, M. & Wellner, J. (2004). Information bounds for Cox regression models with missing data, *Annals of Statistics* **32**, 2.

[14] Newey, W.K. (1990). Semiparametric efficiency bounds, *Journal of Applied Econometrics* **5**, 99–135.

[15] Quale, C.M., van der Laan, M.J. & Robins, J.M. (2003). Locally Efficient Estimation with Bivariate Right Censored Data, Technical Report, University of California, Department of Statistics, Berkeley.

[16] Robins, J.M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, *Proceedings of the Biopharmaceutical Section*, American Statistical Association, Alexandria, pp. 24–33.

[17] Robins, J.M. (1996). Locally efficient median regression with random censoring and surrogate markers, Proceedings of the 1994 Conference on Lifetime Data Models in Reliability and Survival Analysis, Boston. in *Lifetime Data: Models in Reliability and Survival Analysis*, N.P. Jewell, A.C. Kimber, M.-L. Ting Lee & G.A. Whitmore eds. Kluwer, New York, 263–274.

[18] Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models, in Proceedings 1999 Joint Statistical Meetings, Washington, DC.

[19] Robins, J.M. & Finkelstein, D. (2000). Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests, *Biometrics* **5**(3), 779–788.

[20] Robins, J.M. & Ritov, Y. (1997). A curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models, *Statistics in Medicine* **16**, 285–319.

[21] Robins, J.M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, in *AIDS Epidemiology – Methodological Issues*, N. Jewell, K. Dietz & V. Farewell eds. Birkhäuser, Boston, pp. 297–331.

[22] Robins, J. & Rotnitzky, A. (2001). Discussion of the paper by Bickel and Kwon, inference for semiparametric models: some questions and an answer, *Statistica Sinica* **11**(4), 863–960.

[23] Robins, J., Rotnitzky, A. & Bonetti, M. (2001). Discussion of the paper by Frangakis C and Rubin D. "A

note on addressing an idiosyncrasy in estimating survival curves using double-sampling in the presence of self-selected right censoring", *Biometrics* **57**, 343–347.

[24] Robins J.M., Rotnitzky A. & Scharfstein D.O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models, in: *Statistical Models for Epidemiology, the Environment, and Clinical Trials*, E. Halloran & D. Berry, eds. Springer-Verlag, New York, pp. 1–95.

[25] Robins, J.M., Rotnitzky, A. & van der Laan, M.J. (2000). Discussion of "On profile likelihood" by Murphy and van der Vaart, *Journal of the American Statistical Association* **95**, 477–482.

[26] Robins, J., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some of the regressors are nor always observed, *Journal of the American Statistical Association* **89**, 846–866.

[27] Rotnitzky, A. & Robins, J. (1997). Analysis of semiparametric regression models with non-ignorable nonresponse, *Statistics in Medicine* **16**, 81–102.

[28] Rotnitzky, A., Robins, J.M. & Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse, *Journal of the American Statistical Association* **93**(444), 1321–1339.

[29] Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion), *Journal of the American Statistical Association* **94**(448), 1096–1120.

[30] Scharfstein, D.O. & Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative censoring, *Biometrika* **89**, 617–634.

[31] Scharfstein, D.O., Robins, J.M., Eddings, W. & Rotnitzky, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints, *Biometrics* **57**(2), 404–413.

[32] Slud, E.V. & Rubinstein, L.V. (1983). Dependent competing risks and summary survival curves, *Biometrika* **70**, 643–649.

[33] Strawderman, R.L. (2000). Estimating the mean of an increasing stochastic process at a censored stopping time, *Journal of the American Statistical Association* **95**, 1192–1208.

[34] van der Laan, M.J. & Hubbard, H. (1997). Estimation of interval censored data and covariates, *Lifetime Data Analysis* **3**, 77–91.

[35] van der Laan, M.J. & Hubbard, H. (1999). Locally efficient estimation of the quality adjusted lifetime distribution with right-censored data and covariates, *Biometrics* **55**, 530–536.

[36] van der Laan, M.J., Hubbard, H. & Robins, J. (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies, *Journal of the American Statistical Association* **97**, 494–507.

[37] van der Laan, M.J. & Robins, J. (1998). Locally efficient estimation with current status data and time dependent covariates, *Journal of the American Statistical Association* **93**, 693–701.

[38] van der Laan, M.J. & Robins, J. (2003). *Unified Methods for Censored and Longitudinal Data and Causality*. Springer Verlag, New York.

[39] van der Vaart, A.W. (1988). *Statistical Estimation in Large Parameter Spaces*. CWI Tract, Centre for Mathematics and Computer Science, Amsterdam.

[40] van der Vaart, A.W. (1991). On differentiable functionals, *Annals of Statistics* **19**, 178–204.

[41] Zhao H. & Tsiatis A.A. (1997). A consistent estimator for the distribution of quality adjusted survival time, *Biometrika*, **84**, 339–348.

[42] Zhao, H. & Tsiatis, A.A. (1999). Efficient estimation of the distribution of quality adjusted survival time, *Biometrics* **55**, 1101–1107.

[43] Zhao, H. & Tsiatis, A.A. (2000). Estimating mean quality adjusted lifetime with censored data, *Sankhya* **62**, Series B(1), 175–188.

[44] Zheng M. & Klein J.P. (1994). A self-consistent estimator of marginal survival functions based on dependent competing risk data and an assumed copula, *Communications in Statistics, Part A - Theory and Methods* **23**, 2299–2311.

[45] Zheng, M. & Klein, J.P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika* **82**, 127–138.

ANDREA ROTNITZKY & JAMES M. ROBINS

Ion Channel Modeling

Biological cells are enclosed by a phospholipid bilayer that is almost impermeable to water and water soluble molecules. Ion channels are proteins that span the membrane with a central pore that can open under certain conditions, allowing electrically charged ions to pass through it forming a minute flow of electrical current. Different kinds of channels allow the passage of different ions, such as Na^+ or K^+.

The opening and closing of ion channels is called *gating*. The major types of gating mechanism are *voltage gated* (channels respond to changes in membrane potential) and *ligand activated* (channels are activated by binding with molecules of certain chemicals). All electrical activity in the nervous system is regulated by ion channel gating, thereby controlling many diverse activities. Studying their behavior increases our understanding of normal physiology and the effect of drugs and toxins.

Measurements of the superposition of currents through many channels are called *macroscopic measurements.* For example, in the decay of a miniature endplate current at the neuromuscular junction, several thousand channels are involved, enough to produce a smooth curve in which the contribution of individual channels is impossible to see. Experimental macroscopic measurements are made following a jump change in conditions and the time course of the subsequent current can be fitted by the sum of several exponential curves with different time constants. If, on the other hand, we record from a moderate number of channels, the fluctuations about the average behavior become large enough to measure and they can be studied by **time series** methods.

Since the pioneering patch–clamp experiments of Neher and Sakmann [39] (see also [42]), it has become routinely possible to observe currents of a few picoamperes flowing through a single channel. Apart from noise and inertia in the recording system, we are essentially observing the opening and closing of the channel. When the channel is open there is a current of approximately constant amplitude (shown as a downward deflection in Figure 1); when the channel closes the current stops.

Single-channel Models

An ion channel consists of a single macromolecule that can exist in a number, m, say, of different chemical states, either by itself or in association with molecules of a specific ligand. Under stable conditions, we model transitions between these states by a homogeneous continuous-time **Markov chain** with transition rate matrix $\mathbf{Q}$ with elements

$$q_{ij} = \frac{\lim_{\Delta t \to 0} P(X(t + \Delta t) = j | X(t) = i)}{\Delta t}, \quad i \neq j \tag{1}$$

where $X(t)$ denotes the state occupied at time t. Define $q_{ii} = -\sum_{k \neq i} q_{ik}$ so the elements of each row sum to zero.

For example, a five-state model has been used to describe the nicotinic acetylcholine receptor. In this mechanism, there may be one agonist molecule (A) or two molecules (A_2) bound to the shut channel (R) or the open channel (R*). In Figure 2, three shut states (3, 4, 5) are shown on the bottom row and two open states (1, 2) on the top; on the right two agonist molecules are bound, one in the middle and none on the left.

The possible transitions are marked with appropriate rate constants. The rate constant for binding one molecule when the channel is free is written as $2k_{+1}$ because there are two free receptor sites; similarly, the dissociation rate for unbinding one of two occupied receptors is written as $2k_{-2}$ (for the shut channel) and $2k^*_{-2}$ (for the open channel). For a reaction involving a single molecule (e.g. a conformation change), the *transition rate* is simply the *reaction rate constant*. The same is true of dissociation (unbinding) of a single molecule of ligand that is bound to a receptor. For a reaction in which a free ligand

Figure 1 An example record of a current flowing through a single channel

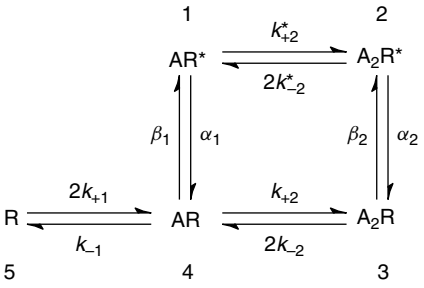

Figure 2 A simple model of an acetylcholine channel showing 5 possible states

molecule binds to a receptor, the transition rate is the product of the rate constant and the ligand concentration. The assumption that transition rates are constant over time, implies that the ligand concentration and membrane potential are constant; this is usually not true in daily life but may be in controlled experiments.

The transition rate matrix is

$$\mathbf{Q} = \begin{pmatrix} -(\alpha_1 + k^*_{+2}x) & k^*_{+2}x & 0 & \alpha_1 & 0 \\ 2k^*_{-2} & -(\alpha_2 + 2k^*_{-2}) & \alpha_2 & 0 & 0 \\ 0 & \beta_2 & -(\beta_2 + 2k_{-2}) & 2k_{-2} & 0 \\ \beta_1 & 0 & k_{+2}x & -(\beta_1 + k_{+2}x + k_{-1}) & k_{-1} \\ 0 & 0 & 0 & 2k_{+1}x & -2k_{+1}x \end{pmatrix}. \tag{2}$$

Note the multiplying factor, x, the free ligand concentration for the transition rates involving binding. In choosing numerical values for the parameters, we assume the principle of *microscopic reversibility*, so that, in the absence of an energy source, each individual reaction will proceed, on average, at the same rate in each direction. In particular, the product of transition rates around the cycle (1, 2, 3, 4) is the same in both directions. A physically plausible example, with $x = 100\,\text{nM}$, is

$$\mathbf{Q} = \left(\begin{array}{cc:ccc} -3050 & 50 & 0 & 3000 & 0 \\ 0.667 & -500.667 & 500 & 0 & 0 \\ \hdashline 0 & 15000 & -19000 & 4000 & 0 \\ 15 & 0 & 50 & -2065 & 2000 \\ 0 & 0 & 0 & 10 & -10 \end{array}\right), \tag{3}$$

where transition rates are in units of s^{-1}. The matrix has been partitioned with open states in the top left corner. The doubly occupied state opens much quicker and closes slower than the singly occupied state, which is slow to open and quick to shut.

Many models can be constructed in this way; some models can get quite large. Ball [1] studied a model for the nicotinic acetylcholine receptor based on molecular structure with 128 states of which 4 are open; by exploiting symmetry this reduces to 3 open and 69 closed states – still quite big!

Although one can eliminate some models for a particular channel on the basis of observable characteristics, some indeterminacy arises because we cannot see which state the channel is in; only if it is open or closed. Two or more distinct models may give rise to the same observable features under fixed conditions, [24, 34]. Further discrimination between models is possible by observing the same channel under different conditions of voltage or agonist concentration.

Standard Markov chain theory yields the transition probability matrix with elements $p_{ij}(t) = P(X(t) = j | X(0) = i)$ is given by $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$, $t > 0$. Spectral expansion of the matrix $-\mathbf{Q}$ with **eigenvalues** λ_i leads to the expression

$$\mathrm{e}^{\mathbf{Q}t} = \sum_{i=1}^{m} \mathrm{e}^{-\lambda_i t} \mathbf{A}_i. \tag{4}$$

Using this expansion it can be shown, [19], that macroscopic current relaxes as a mixture of $m - 1$ exponential components, omitting the zero eigenvalue, as does the **autocorrelation function** in noise measurements. Thus, for example, we expect to see four components in the five-state model. However, some components of the mixture may correspond to short-lived components (large λ) with small weight, so they may be difficult to detect in practice. The apparent number of components observed in such experiments, plus one, can thus only be taken as a lower bound for the number of states in the model.

Behavior of Single Channels Under Equilibrium Conditions

Open Times and Shut Times

If there is more than one open state, then an open time starts when the channel leaves a shut state for an open state, takes a tour round various open states and ends by entering a shut state. For example, in the five-state model an open time might start with a transition from state 4 to state 1, make several transitions back and forth between states 1 and 2, then end with a transition from state 1 to state 4 or from state 2 to state 3.

It is convenient to arrange that all the open states be labelled as states $1, 2 \ldots m_o$ where m_o is the number of open states; the shut (or closed) states having the highest numbered labels $m_o + 1$ to $m_o + m_c = m$. Then the $\mathbf{Q}$-matrix can be partitioned as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{oo} & \mathbf{Q}_{oc} \\ \mathbf{Q}_{co} & \mathbf{Q}_{cc} \end{pmatrix}, \tag{5}$$

where $\mathbf{Q}_{oo}$, a square matrix of dimension $m_o \times m_o$, contains all the transition rates between open states; $\mathbf{Q}_{cc}$ contains all the transition rates between closed states; $\mathbf{Q}_{oc}, \mathbf{Q}_{co}$ contain, respectively, the transition rates from open to closed states and from closed to open states, see (3).

The process may be seen as an alternating process of open and closed intervals. If we note the durations of these intervals and the states of the underlying system in which such intervals begin, we have a **semi-Markov process** with kernel matrix

$$\mathbf{G}(t) = \begin{pmatrix} \mathbf{0} & \mathbf{G}_{oc}(t) \\ \mathbf{G}_{co}(t) & \mathbf{0} \end{pmatrix}. \tag{6}$$

The ijth element of $\mathbf{G}(t)$ is a joint probability density for the duration of an interval and the probability that it ends with a transition into state j, conditional on starting an interval in state i. To obtain this kernel, let $\mathbf{R}_o(t)$ be a matrix function whose ijth element, where i and j are open states, is

$$\Pr\text{ob}(X(t) = j \text{ and channel open throughout time 0 to } t | X(0) = i). \tag{7}$$

Then $\mathbf{R}_o(t) = \mathrm{e}^{\mathbf{Q}_{oo}t}$, $t > 0$, and $\mathbf{G}_{oc}(t) = \mathbf{R}_o(t)\mathbf{Q}_{oc} = \mathrm{e}^{\mathbf{Q}_{oo}t}\mathbf{Q}_{oc}$. Similarly, $\mathbf{G}_{co}(t) = \mathrm{e}^{\mathbf{Q}_{cc}t}\mathbf{Q}_{co}$.

A Markov chain embedded at the points where intervals begin, records the states occupied at those times. The transition probability matrix of this chain is obtained by integrating with respect to t, yielding $\mathbf{G} = \int_0^\infty \mathbf{G}(t)\,\mathrm{d}t = \begin{pmatrix} \mathbf{0} & \mathbf{G}_{oc} \\ \mathbf{G}_{co} & \mathbf{0} \end{pmatrix}$, where $\mathbf{G}_{oc} = -\mathbf{Q}_{oo}^{-1}\mathbf{Q}_{oc}$ and $\mathbf{G}_{co} = -\mathbf{Q}_{cc}^{-1}\mathbf{Q}_{co}$. If we consider only the start of open intervals, the chain embedded at those points has transition matrix $\mathbf{G}_{oc}\mathbf{G}_{co}$. Let the equilibrium distribution of this chain be denoted by the row vector $\mathbf{\Phi}_o$. The equilibrium distribution of the entry states for closed intervals is then given by $\mathbf{\Phi}_c = \mathbf{\Phi}_o\mathbf{G}_{oc}$.

To get the pdf of open times, we have to sum over the possible closed states the channel might move to at the end and, suitably weighted, sum over the states that an open time might start in: so we get

$$\begin{aligned} f_o(t) &= \mathbf{\Phi}_o\mathbf{G}_{oc}(t)\mathbf{u}_c = \mathbf{\Phi}_o\mathrm{e}^{\mathbf{Q}_{oo}t}\mathbf{Q}_{oc}\mathbf{u}_c \\ &= -\mathbf{\Phi}_o\mathrm{e}^{\mathbf{Q}_{oo}t}\mathbf{Q}_{oo}\mathbf{u}_o, \end{aligned} \tag{8}$$

with mean open time $\mu_o = -\mathbf{\Phi}_o\mathbf{Q}_{oo}^{-1}\mathbf{u}_o$. In this equation $\mathbf{u}_o$, $\mathbf{u}_c$ are column vectors of 1's of appropriate length. Note that, because the rows of the matrix $\mathbf{Q}$ sum to zero, $\mathbf{Q}_{oo}\mathbf{u}_o + \mathbf{Q}_{oc}\mathbf{u}_c = \mathbf{0}$. If we use the spectral resolution of the matrix $-\mathbf{Q}_{oo}$, we see that the probability density function $f_o(t)$can be expressed as a mixture of exponential components with time constants given by the m_o eigenvalues of $-\mathbf{Q}_{oo}$. Unlike the matrix $-\mathbf{Q}$, it will not have a zero eigenvalue.

Similarly, interchanging o and c, we get the pdf of shut times as

$$f_c(t) = -\mathbf{\Phi}_c\mathrm{e}^{\mathbf{Q}_{cc}t}\mathbf{Q}_{cc}\mathbf{u}_c, \text{ with mean shut time}$$
$$\mu_c = -\mathbf{\Phi}_c\mathbf{Q}_{cc}^{-1}\mathbf{u}_c. \tag{9}$$

The pdf can be expressed as a mixture of exponential components with time constants given by the m_c eigenvalues of $-\mathbf{Q}_{cc}$.

So the distributions of open times and shut times tell us something about the numbers of open states and shut states, again with the caveat that we might not be able to distinguish all components from an experimental record.

Standard Markov chain theory tells us that the duration of sojourns in a single state, state i say, follow a simple exponential distribution with pdf $f(t) = -q_{ii}\mathrm{e}^{q_{ii}t}$, for $t > 0$; and mean $-1/q_{ii}$. This is a special case of the above distributions. In particular, for the five-state model, the reciprocals of (minus) the diagonal elements of $\mathbf{Q}$ give the mean lifetimes

of sojourns in individual states 1–5 as 0.328 ms, 1.997 ms, 52.6 μs, 0.484 ms, and 100 ms respectively.

Joint Distributions

The joint probability density of an open time T_o and the immediately following shut time T_c is given by

$$\mathbf{\Phi}_o \mathbf{G}_{oc}(t_o)\mathbf{G}_{co}(t_c)\mathbf{u}_o.$$

Similar results may be obtained for any pair of intervals. Joint distributions are useful in distinguishing between mechanisms; see [23, 35, 36]. We can also build up a **likelihood** for a complete sequence of open times and shut times. If there are M pairs of open and following shut times (t_j, s_j), this takes the form

$$\mathbf{\Phi}_o \prod_{j=1}^{M} (\mathrm{e}^{\mathbf{Q}_{oo}t_j}\mathbf{Q}_{oc}\mathrm{e}^{\mathbf{Q}_{cc}s_j}\mathbf{Q}_{co})\mathbf{u}_o.$$

This can be maximized to estimate the parameters of a given model and to test the fit of a model to data – see [13, 24, 32].

Correlations between open times or shut times can occur if there are at least two open states and two shut states. The correlation between the duration of an open time and the nth subsequent open time has the form

$$\rho_n = \sum w_i \lambda_i^n, \tag{10}$$

where the number of terms in the summation is $V - 1$. Here λ_i are those eigenvalues of $\mathbf{G}_{oc}\mathbf{G}_{co}$ that are neither zero nor one. V, the (vertex) connectivity of the mechanism is the smallest number of states that need to be removed (together with any links they have) in order to separate the set of open states from the set of shut states. Correlations between intervals therefore tell us something about the connectivity between open and shut states. In the five-state model, we need to remove at least two states to separate the open and shut states; so $V = 2$ and correlations die away with lag n as a single geometric term.

Results on correlations are given in [4, 10, 12, 25]. Colquhoun and Hawkes [21] also studied the distribution of openings and shuttings after a jump in agonist concentration or voltage. The behavior of various subsequent open and shut times (first, second etc.) also depends on V.

Bursting Behavior

Openings usually seem to occur in bursts of activity. A sequence of openings will be separated by brief shuttings and then there will be a long shut period before the activity starts again. This behavior can be explained by dividing the shut states into two categories: short-lived shut states and long-lived shut states.

A detailed treatment of bursting behavior is given in [20]. For the five-state model, the mean duration of a stay in state 5 is very much longer than that in any other state. Then shut times (gaps) within bursts of openings are almost certainly sojourns within the pair of shut states (3, 4); gaps between bursts will consist of sojourns within the shut states that include at least one visit to state 5. The distribution of gaps within bursts is a mixture of two exponentials with a mean of 57.6 μs, whereas gaps between bursts have a mean of 3790 ms.

Information about gaps between bursts is unreliable because there may be more than one channel in the recording environment so that, while the activity within a burst of openings almost certainly arises from one channel, different bursts may arise from different channels. This is one reason for studying the behavior of within-burst activity, as the information arising from it should be fairly reliable. Distributions derived include those for the number of openings per burst, duration of a burst, total open time per burst, individual openings within a burst (the first, second etc.).

Time Interval Omission

A big problem in observing single-channel records is that of time interval omission (TIO). Because of noise and inertia in the recording system, very short openings or shuttings, are likely to be missed. Results get distorted because, for example, what appears to be one long open time may actually be two or three open times separated by shut times too short to be distinguished. One way to cope with this is to study the total burst length or the total open time per burst, as these should be insensitive to missing short shut times.

In order to allow for missed events when dealing with the individual openings and shuttings it is usual to assume a critical constant dead-time, ξ, such that all open or shut intervals greater than this are observed accurately but shorter intervals are missed (a safe ξ value can be imposed retrospectively on recorded data). We then work with *apparent open*

times defined as periods that start with an open time of duration greater than ξ that may then be extended by a number of openings separated by shut periods each of duration less than ξ; they are terminated at the start of a shut period of duration greater than ξ. Apparent shut times are similarly defined.

Approximate solutions for the distributions of apparent open times and apparent shut times were used before Ball and Sansom, [11, 12], obtained exact results in the form of Laplace Transforms and also considered the effect of TIO on correlations between intervals. Exact expressions for the pdfs of apparent open times and shut times were found by Hawkes et al. [29]. These are fine for small to moderate values of time t, but can be numerically unstable for large t. These results were also studied in a general semi-Markov framework, [6, 7]. Asymptotic approximations can be found, [30, 33], that are extremely accurate for values of t from very large right down to fairly small; for small t the exact results are readily obtainable, so that the distributions are obtained over the whole range. If the true distribution is a mixture of k exponentials, then the approximation to the distribution of apparent times allowing for TIO is also a mixture of k exponentials; the time constants are, however, different.

These methods were used to study the effect of TIO on joint distributions of apparent open and shut times, [23], and to calculate the likelihood of a complete series of intervals, and thus estimate the model parameters; see also [18] for a study of the performance of likelihood estimation. TIO can induce some indeterminacy in the estimation of parameters. For data recorded under fixed conditions there can be two sets of parameters that seem to fit the data equally well: typically a *fast solution* and a *slow solution.* These can, however, be discriminated by observing the same channel under different conditions of voltage or ligand concentration, see [1, 3].

The TIO problem has been studied in the context of recording apparent open and shut intervals stimulated by a pulse of agonist concentration or voltage change; see [22, 38].

Multiple Levels

So far we have discussed only channels that are open or closed. Some channels, however, show several different levels of current, corresponding to different sets of states. It is possible to ignore this and just analyze the system as open or closed, but this loses information. Ball et al. [9] give a general treatment of a multilevel system, deriving burst properties including distributions of total charge transfer, total sojourn time, and number of visits to each conductance level during a burst. Merlushkin [37] studied various apparent sojourn distributions allowing for TIO in the multilevel case.

Multiple levels sometimes arise from the presence of more than one channel: if so, they are usually treated as acting independently. However, various models for systems of interacting channels are studied in [5, 8, 14].

Hidden Markov Methods of Analysis

Several authors (e.g. [15–17, 26–28, 40, 41, 43]) have applied **Bayesian** or **Hidden Markov** methods to the original noisy signals obtained from patch clamp experiments. These can be used to extract the ideal step-function signals (representing opening and shutting) from the noise; they are also used to estimate parameters in the models directly without identifying individual open and shut times. These techniques can cope with multilevel records as well as the simple open/shut case. **Markov chain Monte Carlo** methods of Bayesian analysis were applied in [2, 31].

References

[1] Ball, S.S. (2000). *Stochastic Models of Ion Channels*. PhD Thesis, University of Nottingham.

[2] Ball, F.G., Cai, Y., Kadane, J.B. & O'hagan, A. (1999). Bayesian inference for ion-channel gating mechanisms directly from single-channel recordings, using Markov chain Monte Carlo, *Proceedings of the Royal Society of London A* **455**, 2879–2932.

[3] Ball, F.G. & Davies, S.S. (1995). Statistical inference for a two-state Markov model of a single ion channel, incorporating time interval omission, *Journal of the Royal Statistical Society B* **57**, 269–287.

[4] Ball, F.G., Kerry, C.J., Ramsey, R.L., Sansom, M.S.P. & Usherwood, P.N.R. (1988). The use of dwell time cross-correlation functions to study single ion channel gating kinetics, *Biophysical Journal* **54**, 309–320.

[5] Ball, F.G., Milne, R.K., Tame, I.D. & Yeo, G.F. (1997). Superposition of interacting aggregated continuous-time Markov chains, *Advances in Applied Probability* **29**, 56–91.

[6] Ball, F., Milne, R.K. & Yeo, G.F. (1991). Aggregated semi-Markov processes incorporating time interval omission, *Advances in Applied Probability* **23**, 772–797.

[7] Ball, F.G., Milne, R.K. & Yeo, G.F. (1993). On the exact distribution of observed open times in single ion channel models, *Journal of Applied Probability* **30**, 529–537.

[8] Ball, F.G., Milne, R.K. & Yeo, G.F. (2000). Stochastic models for systems of interacting ion channels, *IMA Journal of Mathematics Applied in Medicine and Biology* **17**, 263–293.

[9] Ball, F.G., Milne, R.K. & Yeo, G.F. (2002). Multivariate semi-Markov analysis of burst properties of multiconductance single ion channels, *Journal of Applied Probability* **39**, 179–196.

[10] Ball, F.G. & Rice, J.A. (1989). A note on single-channel autocorrelation functions, *Mathematical Biosciences* **97**, 17–26.

[11] Ball, F. & Sansom, M.S.P. (1988a). Aggregated Markov processes incorporating time interval omission, *Advances in Applied Probability* **20**, 546–572.

[12] Ball, F.G. & Sansom, M.S.P. (1988b). Single-channel autocorrelation functions: the effects of time interval omission, *Biophysical Journal* **53**, 819–832.

[13] Ball, F.G. & Sansom, M.S.P. (1989). Ion-channel gating mechanisms: model identification and parameter estimation from single channel recordings, *Proceedings of the Royal Society of London B* **236**, 385–416.

[14] Ball, F.G. & Yeo, G.F. (1999). Superposition of spatially interacting aggregated continuous time Markov chains, *Methodology and Computing in Applied Probability* **2**, 93–115.

[15] Chung, S.H. & Cage, P.W. (1998). Signal processing techniques for channel current analysis based on hidden Markov models, *Methods in Enzymology* **293**, 420–437.

[16] Chung, S.H., Krishnamurthy, V. & Moore, J.B. (1991). Adaptive processing techniques based on hidden Markov models for characterising very small channel currents buried in noise and deterministic interference, *Philosophical Transactions of the Royal Society of London B* **334**, 357–384.

[17] Chung, S.H., Moore, J.B., Xia, L., Premkumar, L.S. & Gage, P.W. (1990). Characterization of single channel currents using digital signal processing techniques based on hidden Markov models, *Philosophical Transactions of the Royal Society of London B* **329**, 265–285.

[18] Colquhoun, D., Hatton, C.J. & Hawkes, A.G. (2003). The quality of maximum likelihood estimation of ion channel rate constants, *Journal of Physiology London* **547**, 699–728.

[19] Colquhoun, D. & Hawkes, A.G. (1977). Relaxation and fluctuations of membrane currents that flow through drug-operated channels, *Proceedings of the Royal Society of London B* **199**, 231–262.

[20] Colquhoun, D. & Hawkes, A.G. (1982). On the stochastic properties of bursts of single ion channel openings and of clusters of bursts, *Philosophical Transactions of the Royal Society of London B* **300**, 1–59.

[21] Colquhoun, D. & Hawkes, A.G. (1987). A note on correlation in single ion channel records, *Proceedings of the Royal Society of London B* **230**, 15–52.

[22] Colquhoun, D., Hawkes, A.G., Merlushkin, A. & Edmonds, B. (1997). Properties of single ion channel currents elicited by a pulse of agonist concentration or voltage, *Philosophical Transactions of the Royal Society of London A* **355**, 1743–1786.

[23] Colquhoun, D., Hawkes, A.G. & Srodsinski, K. (1996). Joint distributions of apparent open times and shut times of single ion channels and the maximum likelihood fitting of mechanisms, *Philosophical Transactions of the Royal Society of London A* **354**, 2555–2590.

[24] Fredkin, D.R., Montal, M. & Rice, J.A. (1985). Identification of aggregated Markovian models: application to the nicotinic acetylcholine receptor, in *Proceedings of the Berkeley Conference in Honour of Jerzy Neyman and Jack Kiefer*, L.M. Le Cam & R.A. Ohlsen, eds. Wadsworth, Belmont, pp. 269–289.

[25] Fredkin, D.R. & Rice, J.A. (1987). Correlation functions of a function of a finite-state Markov process with application to channel kinetics, *Mathematical Biosciences* **87**, 161–172.

[26] Fredkin, D.R. & Rice, J.A. (1992a). Maximum likelihood estimation and identification directly from single-channel recordings, *Proceedings of the Royal Society of London B* **249**, 125–132.

[27] Fredkin, B.R. & Rice, J.A. (1992b). Bayesian restoration of single-channel patch clamp recordings, *Biometrics* **48**, 427–448.

[28] Fredkin, B.R. & Rice, J.A. (2001). Fast evaluation of the likelihood of an HMM: ion channel currents with filtering and coloured noise, *IEEE Transactions on Signal Processing* **49**, 625–633.

[29] Hawkes, A.G., Jalali, A. & Colquhoun, D. (1990). The distributions of the apparent open times and shut times in a single channel record when brief events cannot be detected, *Philosophical Transactions of the Royal Society of London A* **332**, 511–538.

[30] Hawkes, A.G., Jalali, A. & Colquhoun, D. (1992). Asymptotic distributions of apparent open times and shut times in a single channel record allowing for the omission of brief events, *Philosophical Transactions of the Royal Society of London B* **337**, 383–404.

[31] Hodgson, M.E.A. (1999). A Bayesian restoration of an ion channel signal, *Journal of the Royal Statistical Society B* **61**, 95–114.

[32] Horn, R. & Lange, K. (1983). Estimating kinetic constants from single channel data, *Biophysical Journal* **43**, 207–233.

[33] Jalali, A. & Hawkes, A.G. (1992). Generalised eigenproblems arising in aggregated Markov processes allowing for time interval omission, *Advances in Applied Probability* **24**, 302–321.

[34] Kienker, P. (1989). Equivalence of aggregated Markov models of ion-channel gating, *Proceedings of the Royal Society of London B* **236**, 269–309.

[35] Magleby, K.L. & Weiss, D.S. (1990). Identifying kinetic gating mechanisms for ion channels by using two-dimensional distributions of simulated dwell times, *Proceedings of the Royal Society of London B* **241**, 220–228.

[36] Mcmanus, O.B., Blatz, A.L. & Magleby, K.L. (1985). Inverse relationship of the duration of adjacent open and shut intervals for Cl and K channels, *Nature* **317**, 625–628.

[37] Merlushkin, A.I. (1996). *Some Problems Arising in Stochastic Modelling of Ion Channels due to Time Interval Omission*. PhD Thesis, University of Wales.

[38] Merlushkin, A.I. & Hawkes, A.G. (1997). Stochastic behaviour of ion channels in varying conditions, *IMA Journal of Mathematics Applied in Medicine and Biology* **14**, 125–149.

[39] Neher, E. & Sakmann, B. (1976). Single-channel currents recorded from membrane of denervated frog muscle fibres, *Nature* **260**, 799–802.

[40] Qin, F., Auerbach, A. & Sachs, F. (2000a). A direct optimisation approach to hidden Markov modeling for single channel kinetics, *Biophysical Journal* **79**, 1915–1927.

[41] Qin, F., Auerbach, A. & Sachs, F. (2000b). Hidden Markov modeling for single channel kinetics with filtering and correlated noise, *Biophysical Journal* **79**, 1928–1944.

[42] Sakmann, B. & Neher, E. eds. (1995). *Single Channel Recording*, 2nd Ed. Plenum Press, New York.

[43] Venkataramanan, L. & Sigworth, F.J. (2002). Applying hidden Markov models to the analysis of single ion channel activity, *Biophysical Journal* **82**, 1930–1942.

(*See also* **Compartment Models**; **Mathematical Biology, Overview**)

ALAN G. HAWKES

Irwin, Joseph Oscar

Born: December 17, 1898, in London, UK.
Died: July 27, 1982, in Schaffhausen, Switzerland.

Reproduced by permission of the Royal Statisical Society

As the leading theoretician amongst British medical statisticians in the 1930s and in subsequent decades, Oscar Irwin played an important role in linking developments in statistical theory to applications in medical research.

At school, Irwin had specialized in classics before he took up mathematics. In 1917 his entry to Cambridge on a scholarship was delayed first by illness, and then by a crucial period working under **Karl Pearson** on anti-aircraft trajectories. On achieving his degree in 1921, he joined Pearson's staff at University College. Renewed illness led to a period of recuperation in Switzerland which initiated a life-long love of that country and, in later life, to his marriage to a Swiss wife.

In 1928, Irwin joined **R. A. Fisher's** department at Rothamsted, and thus became one of the few statisticians to work with both Karl Pearson and Fisher. A decade later, when Fisher and **Egon Pearson** occupied adjacent floors at University College, Irwin was said to be one of the few people to be *persona grata* in both departments. During his period with Fisher, ending in 1931, Irwin came to grips with the mathematical theory published during the 1920s by Fisher, who always retained a high opinion of Irwin's mathematical ability.

In 1931, Irwin joined the staff of the **Medical Research Council (MRC)**, housed at the London School of Hygiene and Tropical Medicine, where he was to stay for most of the next 30 years. As an MRC worker, Irwin had only a part-time university appointment. However, for about 25 years he taught a course in statistical methods, introducing many relatively recent developments in the subject. During the war years (1940–1945) he worked in Cambridge, teaching statistics to mathematicians, many of whom followed a subsequent career in statistics.

Irwin retired in 1965, after which he worked for a short time at the Galton Laboratory, University College London, before moving to Switzerland. He was a Visiting Professor at the University of North Carolina, Chapel Hill during three sabbatical periods.

Irwin's early papers reveal great mathematical fluency, which he retained throughout his life. In a paper of 1927 [1] he derived the distribution of the **mean** (*see* **Sampling Distributions**) from various distributions using the **characteristic function**. At Rothamsted he wrote on the influence of climatic factors on crop yield, but was perhaps more intrigued by theoretic work on topics such as the **analysis of variance**. In 1931, he started a series of expository papers on 'Recent advances in mathematical statistics', with bibliographies, which were particularly influential at a time at which few books on statistical theory existed.

His move to the MRC enlarged his research interests. He wrote several papers on **factor analysis**, but during the 1930s, while his colleague and close contemporary **Austin Bradford Hill** devoted himself largely to epidemiologic and (later) clinical research, Irwin's interests focused on laboratory experimentation. There were papers with H. Barkwith on the **dilution method** of estimating bacterial densities, and a developing interest in the methodology of **biological assay**, stimulated by his membership of a committee of the British Pharmacopoeia Commission. There was a major paper in the 1937 ***Journal of the Royal Statistical Society***, *Supplement* [3], and papers with E. A. Cheeseman clarifying the **maximum likelihood** solution in probit analysis (*see* **Quantal Response Models**).

His 1935 paper in *Metron* [2] described the "exact" test for **two-by-two tables**, derived and published independently from **Yates'** 1934 paper and Fisher's insertion in the 1934 edition of *Statistical*

Methods for Research Workers (*see* **Fisher's Exact Test**). He also wrote extensively on theories of **accident proneness**.

After the war he embarked on many long-term collaborative research programs, often for official committees. These included collaborative assays, especially for the standardization of vitamins, nutritional studies, work on physiologic responses to hot climates, laboratory tests for pertussis vaccines, and tests for the carcinogenicity of tars and mineral oils (*see* **Tumor Incidence Experiments**). The latter work led to papers on the analysis of animal carcinogenicity tests by **actuarial methods**. His earlier work on accident proneness stimulated a revived interest in long-tailed discrete distributions, with some pioneering studies of the Waring distributions.

Irwin played a very active role in the affairs of the **Royal Statistical Society**, as President in 1962–1964, Editor of the *Journal, Series B* from 1949 to 1959, Chairman of the Study Section in 1934–1935 and of the Research Section in 1947–1949, and recipient of the Guy Medal in Silver. He was President of the British Region of the **International Biometric Society** during 1958 and 1959.

Oscar Irwin was a man of wide cultural interests and fine sensitivity. In some ways he was ill-adapted to the more robust features of professional life, and preferred quiet and intimate conversation to public forum and debate. He exhibited great kindness to visiting scientists and students; several young medical visitors to the London School of Hygiene were given tutorials on Fisher's *Statistical Methods for Research Workers*, a book for which Irwin retained undying respect throughout his life.

References

[1] Irwin, J.O. (1927). On the frequency distribution of the means of samples from a population having any law of frequency with finite moments with special reference to Pearson's Type II, *Biometrika* **19**, 225–239.

[2] Irwin, J.O. (1935). Tests of significance for differences between percentages based on small numbers, *Metron* **12**, 83–94.

[3] Irwin, J.O. (1937). Statistical method applied to biological assays, *Journal of the Royal Statistical Society, Supplement* **4**, 1–60.

PETER ARMITAGE

Isolated Populations

From the perspective of **population genetics**, a population is described as a group of individuals who can intermix freely such that there is no restriction of **gene** flow among the members within the group [11]. However, intermixing or gene flow does not always occur freely among the members of natural populations. To account for this phenomenon, population genetic theory developed the concept of "substructured populations", in which partitions within a natural population are allowed with incomplete mixing between the subpopulations [32]. In certain instances, a very small number of individuals are isolated from their parental group and become the founders of a new population. This population is often small in size and geographically so isolated that intermixing or sexual mating becomes almost exclusively restricted to the members within the group. Such groups of individuals established by a few founders having limited contact with other groups, have come to be known as "isolated populations". The degree of isolation can be varied, determined by the number of founders, time of isolation, and the extent of isolation (i.e. lack of gene flow with others).

Examples

In the context of humans, there are numerous examples of isolated populations; the relatively better known in the genetic literature are the Finns, Sardinians, Icelanders, Bedouins, Lapps, Basques, Amish, Hutterites, and some of the Polynesians islanders, among others. Each of them has a unique evolutionary and demographic history in terms of the number of founders, age of the population since founding and other population-related factors, such as the growth and expansion during the life of an isolate. Some of the isolated populations were established far back in time compared with others; some have maintained a relatively constant population size over time, e.g. the Saami of Scandinavia [14], while others like the Finns experienced a large population expansion after their founding [23, 26]. These demographic factors have profound effects on the genetic make-up of a population, which consequently reflect on the phenotype.

Genetic Characteristics of Isolated Populations

The distinctive characteristics of isolated populations, namely, a limited number of founders and lack of gene flow (or contact) with other populations, have some genetic consequences, which may be used as signatures of isolation of the population. First, being formed by a limited number of founders, irrespective of the source of the founders, an isolated population starts its evolution from a somewhat restricted amount of genetic variation. This restriction of genetic variability should be reflected in heterozygosity (i.e. proportion of heterozygous individuals averaged over loci), or gene diversity (i.e. complement of sum of squares of allele frequencies, averaged over loci), as well as in the number of segregating alleles [5, 17]. Secondly, the lack of contact with other populations (i.e. the isolation) also impacts genetic variation within isolated populations in a number of ways. New **mutations**, arising in an isolated population, do not have a chance to traverse easily to other populations, together with which, genetic drift (being particularly strong due to the small size of the isolated population) tends to reduce its genetic variation. As a consequence, an isolated population accumulates genetic divergence at a detectably fast rate from its sister populations from which it diverged after separation from their common ancestry [4, 18].

While the above genetic signatures of population isolation are generally seen at individual locus levels, there are consequences of isolation at a multilocus level of genetic variation as well. For example, the limited number of founders in an isolated population necessarily brings in a limited supply of **haplotypes** (i.e. multilocus combination of alleles on chromosomes), signifying strong **linkage disequilibrium** (LD, nonrandom association of alleles) between loci. The LD between loci in an isolated population is expected to remain strong, since the limited population size does not allow recombination to occur as frequently as in a large population, because of the smaller number of meiosis events per generation. Thus, isolated populations that are of recent origin should demonstrate stronger LD between loci in comparison with large cosmopolitan populations.

Of course, the demographic history of an isolated population, following its foundation, also plays

a role in molding its genetic variation in subsequent generations. The small founding population, combined with effects of genetic drift in subsequent generations, results in likely elimination of certain genetic attributes and enrichment of others. This makes the genome of isolates more homogeneous than that of large cosmopolitan populations. Geographic and reproductive isolation also leads to limited mate choice and consequently results in a higher coefficient of **inbreeding**. The drift effect and inbreeding level lead to increased **prevalence** of certain diseases (particularly the ones of recessive mode of inheritance). Another important characteristic feature is that the members of an isolate generally share a common environment, e.g. climate, nutrition, cultural habits, occupation and education levels, substantially reducing the **confounding** effects of environmental heterogeneity – an important factor in understanding the etiologies of common diseases, which have both genetic and environmental components.

Isolated Populations in Gene Mapping

Population isolates, thus, offer two very important advantages for gene mapping: a homogeneous genome, and a shared environment (see, for example, [7] and [27]. In fact, the confounding effects of genetic and environmental heterogeneity have been a source of discouragement for mapping genes in large cosmopolitan populations. Some isolated populations also have kept demographic records through parish or other registries enabling reconstruction of genealogies for several generations. It is, therefore, not surprising that a large number of **Mendelian** disorders have been mapped in isolated populations. The aforementioned features, typical of many isolated populations, are exemplified convincingly by the demographic history of the Finnish population [23]. A catalogue listing about 35 diseases, mostly recessive, has come to characterize the "Finnish Disease Heritage" [6, 20, 22]; and utilizing the uniqueness of the Finnish population, mutations in 19 of these diseases have thus far been identified and chromosomal regions for 13 others have been localized [31]. Other isolated populations have also been studied for mapping disease genes, such as the Amish, Ashkenazi Jews, Sardinians, French Canadians, Bedouins [2, 12, 19, 25].

This success in identifying the genetic basis of single-gene disorders had raised expectations that isolated populations would also be greatly useful in mapping **complex** traits. An important consideration of using isolated populations in this endeavor is the notion, mentioned earlier, that LD would be higher in recently founded isolated populations compared with large cosmopolitan populations. Indeed, on a global scale, the extent of LD is significantly lower in African populations compared with other world populations [24, 30]. Further isolation prevents the influx of foreign genes, leading to the retention of older haplotypes in a relatively stable fashion. In addition, reduced genetic variability in isolated populations would likely enhance the possibility of capturing alleles with minor individual effects (oligogenic) underlying complex traits because reduction in variation would lead to the enrichment of a few predisposed alleles in the population. As a result of this optimism, several major studies have been undertaken in several isolated populations for finding genes of common diseases, such as asthma, among the Hutterites [21] and Tristan da Cunha [34], type 2 diabetes among the Finns [9], Pima Indians [10], and schizophrenia in Palau [16]. The impact of this resurgence is particularly noticed in privately funded biotechnology companies; for example, the launching of the deCODE Genetics project for studying the entire Icelandic population, initiation of a similar project in Tonga by AutoGen Ltd, as well as joint collaboration between a nonprofit and a for-profit venture for launching a similar project in Estonia. Notwithstanding such excitement and several ongoing studies referred to above, however, apart from rare Mendelian forms, such as MODY [1], no genetic variant associated with a common disease has thus far been identified even though several potential genomic regions have been localized. Current literature presents a series of conflicting views, based on both empirical data and theoretical modeling, on the advantage of using isolates in mapping complex traits [8, 13, 15, 28, 29].

Based on simulation studies, Kruglyak [13] demonstrated that LD in general populations would not extend beyond a distance of 3 kb, and more importantly, isolated populations are unlikely to harbor a higher level of LD compared with general populations, minimizing the importance of population isolates in complex disease studies. Two empirical studies supported these observations: Eaves et al.

[8] in Sardinia, and Taillon Miller et al. [29] in Finland). Over-generalization of these results may, however, be erroneous. For example, Shifman & Darvasi [28] have shown that the extent of LD between SNPs at a distance of up to 200 kb does not differ between isolated (Finnish, Sardinian and Ashkenazi Jew) and outbred populations. However, when the distance between SNPs is increased beyond 200 kb, LD in the three isolated populations noted above increases by an average of 5.6 times. Consequently, the importance of isolated populations in efficiently mapping complex trait genes should not be readily dismissed.

Comments

This resurgence of interest in isolated populations points to a number of existing gaps in our knowledge of the consequences of the genetic characteristics of such populations. For example, in the context of the comparison of LD in isolated vs. outbred cosmopolitan populations, current data do not always specify what truly constitutes an isolation of a population. Since all recombination events that have occurred in the population are key determinants of the decay of LD, a critical demographic factor in designing a mapping study should be the number of generations to the most recent common ancestor (MRCA) in the population [33]. Thus, since the age of the MRCA in a recently expanded population is younger than that in a stable population, isolation alone may not be sufficient to guarantee extended LD blocks in the genome. Likewise, the founder population size at the stage of expansion is important because this initial genetic structure of the population dictates the extent of LD that would go through subsequent decay because of accumulation of subsequently occurring recombination events. Thus, comparisons of isolated vs. cosmopolitan populations with regard to the extent of contemporary LD values should be adjusted for differences of their initial population structure, which has been done rarely in the literature. Finally, since drift effects are more pronounced for loci with smaller mutation rates [3], it is important to demonstrate empirically the genome homogeneity of isolated populations at regions of the genome where mutation rates may not be so small.

References

[1] Bell, G.I., Xiang, K.J., Newman, M.V., Wu, S.H., Wright, L.G., Fajans, S.S. et al. (1991). Gene for non-insulin-dependent diabetes mellitus (maturity-onset diabetes of the young subtype) is linked to DNA polymorphism on human chromosome 20q, *Proceedings of the National Academy of Sciences* **88**, 1484–1488.

[2] Bonné-Tamir, B., Nystuen, A., Ser rousi, E., Kalinsky, H., Kwitek-Black, A.E., Korostishevsky, M. et al. (1997). Usher syndrome in the Samaritans: strengths and limitations of using inbred isolated populations to identify genes causing recessive disorders, *American Journal of Physical Anthropology* **104**, 193–200.

[3] Chakraborty, R. & Jin, L. (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting, *Human Genetics* **88**, 267–272.

[4] Chakraborty, R. & Nei, M. (1977). Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model, *Evolution* **31**, 347–356.

[5] Crow, J.F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.

[6] de la Chapelle, A. (1993). Disease gene mapping in isolated human populations: the example of Finland, *Journal of Medical Genetics* **30**, 857–865.

[7] de la Chapelle, A. & Wright, F. (1998). Linkage disequilibrium mapping in isolate populations: the example of Finland revisited, *Proceedings of the National Academy of Sciences* **95**, 12 416–12 423.

[8] Eaves, I.A., Merriman, T.R., Barber, R.A., Nutland, S., Wolf-Tuomilehto, E., Tuomilehto, J. et al. (2000). The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes, *Nature Genetics* **25**, 320–323.

[9] Ghosh, S., Watanabe, R.M., Valle, T.T., Hauser, E.R., Magnuson, V.L., Langefeld, C.D., Ally, D.S. et al. (2000). The Finland-United States investigation of non-insulin dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to Type 2 diabetes, *American Journal of Human Genetics* **67**, 1174–1185.

[10] Hanson, R.L., Ehm, M.G., Pettitt, D.J., Prochazka, M., Thompson, D.B., Timberlake, D., Foroud, T. et al. (1998). An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians, *American Journal of Human Genetics* **63**, 1130–1138.

[11] Hedrick, P.W. (2000). *Genetics of Populations*. Jones & Bartlett, Boston.

[12] Howen, R.H., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L.A. & Freimer, N.B. (1994). Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis, *Nature Genetics* **8**, 380–386.

[13] Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes, *Nature Genetics* **22**, 139–144.

[14] Laan, M. & Pääbo, S. (1997). Demographic history and linkage disequilibrium in human populations, *Nature Genetics* **17**, 435–438.

[15] Lonjou, C., Collins, A. & Morton, N.E. (1999). Allelic association between marker loci, *Proceedings of the National Academy of Sciences* **96**, 1621–1626.

[16] Myles-Worsley, M., Coon, H., Tiobech, J., Collier, J., Dale, P., Wender, P. et al. (1999). Genetic epidemiological study of schizophrenia in Palau, Micronesia: prevalence and familiality, *American Journal of Medical Genetics* **88**, 4–10.

[17] Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

[18] Nei, M., Maruyama, T. & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations, *Evolution* **29**, 1–10.

[19] Nikali, K., Suomalainen, A., Terwilliger, J., Koskinen, T., Weissenbach, J. & Peltonen, L. (1995). Random search for shared chromosomal regions in four affected individuals: the assignment of a new hereditary ataxia locus, *American Journal of Human Genetics* **56**, 1088–1095.

[20] Norio, R., Nevanlinna, H.R. & Perheentupa, J. (1973). Hereditary diseases in Finland, rare flora in rare soul, *Annals of Clinical Research* **5**, 109–141.

[21] Ober, C., Tsalenko, A., Parry, R. & Cox, N.J. (2000). A second-generation genomewide screen for asthma susceptibility loci in a founder population, *American Journal of Human Genetics* **67**, 1154–1162.

[22] Peltonen, L. (2000). Positional cloning of disease genes: advantages of genetic isolates, *Human Heredity* **50**, 66–75.

[23] Peltonen, L., Palotie, A. & Lange, K. (2000). Use of population isolates for mapping complex traits, *Nature Reviews* **1**, 182–190.

[24] Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T. et al. (2001). Linkage disequilibrium in the human genome, *Nature* **411**, 199–204.

[25] Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S. et al. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population, *Nature Genetics* **9**, 152–159.

[26] Sajantila, A., Abdel-Halim, S., Savolainen, P., Bauer, K. & Pääbo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population, *Proceedings of the National Academy of Sciences* **93**, 12 035–12 039.

[27] Sheffield, V.C., Stone, E.M. & Carmi, R. (1998). Use of isolated inbred human populations for identification of disease genes, *Trends in Genetics* **14**, 391–396.

[28] Shifman, S. & Darvasi, A. (2001). The value of isolated populations, *Nature Genetics* **28**, 309–310.

[29] Taillon-Miller, P., Sardina-Bauer, I., Saccone, N.L., Putzel, J., Laitinen, T., Cao, A., Kere, J. et al. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28, *Nature Genetics* **25**, 324–328.

[30] Tishkoff, S.A., Dietzch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonné-Tamir, B. et al. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins, *Science* **271**, 1380–1387.

[31] Varilo, T., Laan, M., Hovatta, I., Wiebe, V., Terwilliger, J.D. & Peltonen, L. (2000). Linkage disequilibrium in isolated populations: Finland and a young subpopulation of Kuusamo, *European Journal of Human Genetics* **8**, 604–612.

[32] Wright, S. (1951). The genetical structure of populations, *Annals of Eugenics* **15**, 323–354.

[33] Wright, A.F., Carothers, A.D. & Pirastu, M. (1999). Population choice in mapping genes for complex diseases, *Nature Genetics* **23**, 397–404.

[34] Zamel, N., McClean, P.A., Sandell, P.R., Siminovitch, K.A. & Slutsky, A.S. (1996). Asthma on Tristan da Cunha: looking for genetic link. The University of Toronto Genetics of Asthma Research Group, *American Journal of Respiratory and Critical Care Medicine* **153**, 1902–1906.

RANAJIT CHAKRABORTY & RANJAN DEKA

Isotonic Inference

Isotonic inference concerns situations in which a set of parameters is assumed, a priori, to satisfy certain order restrictions. In the most common case, where data are arranged in ordered groups, the **mean** value of a **random variable** is assumed to change monotonically with the ordering of the groups. It is then reasonable to take account of the order restrictions in making inferences about the group means, such as point or interval estimations or significance tests. Isotonic inference extends more generally to situations where there are various shape constraints on response curves, such as convexity, concavity, or sigmoidicity.

One approach to such problems is to assume a parametric model that incorporates those order or shape constraints such as a **linear regression** equation or a particular **dose–response** function. The inference based on the parametric model can, however, be considerably **biased** and variable when the specified model is incorrect. It has been pointed out in environmental toxicology applications, for example, that no parametric dose–response model can be assumed to hold generally at very low doses of interest, and yet a monotone and convex relationship might reasonably, and more reliably, be assumed. We are therefore concerned in this article mainly with methods of inference that avoid the need to specify a rigid parametric model, but nevertheless allow for those order restrictions.

There is a large literature on **estimation** and testing (*see* **Hypothesis Testing**) in the areas of isotonic and order-restricted inferences, and comprehensive surveys of these areas include [3] and [43].

One general approach to the isotonic inference is **maximum likelihood** estimation. The problem of finding order-restricted maximum likelihood estimates is often solved by using **isotonic regression**. In its simplest case an explicit solution is obtained by the pool-adjacent-violators method, but in more general cases it is solved only by some nonlinear programming, see [46] and [10], for example, or by the aid of a formal **Bayesian** approach, as in [42]. For a restricted **likelihood ratio test** the usual asymptotic **chi-square distribution** theory does not apply. In some cases, the resulting distributions are known to be a mixture of χ^2 distributions, but in other cases some **computer-intensive methods** such as parametric **bootstrap** tests [10], or an asymptotic conservative approximation method [46] may be used. The maximum likelihood approach is outlined in another article (*see* **Isotonic Regression**). Here we are concerned mainly with other approaches to isotonic inference. As a natural method of incorporating prior knowledge in particular applications, a Bayesian approach is also briefly mentioned.

The Case for Isotonic Inference

The data in Table 1 are measurements of the half-life of an antibiotic drug in relation to the dose administered. The usual **analysis of variance** (ANOVA) is obviously inappropriate, because of the ordering of the doses, and one possible approach is to assume a parametric model. The simplest model for the monotone relationship is linear regression. However, it is generally difficult to assume that a linear relationship holds over a wide range of an **explanatory variable**. For the dose–response relationship there are of course more natural response curves, such as a sigmoid function, but it is still often difficult to assume a particular model for the given set of data. Furthermore, it also sometimes suffices to show an overall upward trend or to detect a steep **change-point** in the responses. It is then unnecessary to assume a rigid parametric model, and a nonparametric trend test or some **multiple comparisons** procedure is more appropriate (*see* **Simultaneous Inference**). We need assume only a monotone relationship in the mean half-life,

$$\mathrm{H}_1\colon \mu_1 \leq \cdots \leq \mu_a, \tag{1}$$

where at least one inequality is strong, so that the null model $\mathrm{H}_0\colon \mu_1 = \cdots = \mu_a$ is excluded.

The data in Table 2 show ordinal (i.e. **ordered) categorical data** typical of a Phase III comparative **clinical trial**. Assuming a **multinomial** model with cell probabilities p_{ij}, the **null hypothesis** that the two

Table 1 Half life of an antibiotic in rats

Dose (mg/kg)	Data (h)					Average
5	1.17	1.12	1.07	0.98	1.04	1.076
10	1.00	1.21	1.24	1.14	1.34	1.186
25	1.55	1.63	1.49	1.53		1.550
50	1.21	1.63	1.37	1.50	1.81	1.504
200	1.78	1.93	1.80	2.07	1.70	1.856

Table 2 Efficacy in a phase III trial of antibiotics

Drug	Not effective	Slightly effective	Effective	Excellent
AMPC	3	8	30	22
S6472	8	9	29	11

treatments are equal can be expressed as $p_{1j} = p_{2j}$, $j = 1, \ldots, 4$, or equivalently as

$$p_{ij} = p_{i\cdot} p_{\cdot j}, \tag{2}$$

where a dot denotes the summation with respect to the suffix replaced by the dot. Eq. (2) is the familiar independence hypothesis for a two-way **contingency table**. However, the usual **goodness-of-fit chi-square test** is inappropriate, since we are interested in a more restricted alternative

$$\begin{aligned} &p_{11}/p_{21} \le \cdots \le p_{14}/p_{24} \\ \mathrm{H}_2 : &\text{ or} \\ &p_{11}/p_{21} \ge \cdots \ge p_{14}/p_{24}, \end{aligned}$$

where at least one inequality is strong, implying that treatment 1 is superior to treatment 2 in efficacy or vice versa. **Ordered categorical data** are a special case of rank data with many ties, and any method for rank data can be applied to ordered categorical data, and vice versa.

If ordered categorical data are obtained at several doses, as in Table 3, then we are interested in testing the two-way **ordered alternative**:

$$\begin{aligned} \mathrm{H}_3 : p_{i+1,j}/p_{i,j} &\le p_{i+1,j+1}/p_{i,j+1}, \\ i &= 1, \ldots, a-1; \\ j &= 1, \ldots, b-1, \end{aligned}$$

which implies that higher doses are superior to lower doses in efficacy.

A similar hypothesis

$$\begin{aligned} \mathrm{H}_4 : \mu_{i+1,j+1} - \mu_{i+1,j} - \mu_{i,j+1} + \mu_{i,j} \ge 0, \\ i = 1, \ldots, a-1; j = 1, \ldots, b-1, \end{aligned}$$

has been considered for normal means from a two-way layout experiment, which implies that the differences, $\mu_{ij} - \mu_{i'j}$, tend upwards as the level j increases for any $i > i'$; see [14].

Various Extensions of the Monotone Relationship

A monotone dose–response relationship may be disturbed by toxicity at higher doses, and a **nonparametric** testing procedure for the downturn (or "umbrella") hypothesis,

$$\begin{aligned} \mathrm{H}_5 : \mu_1 \le \cdots \le \mu_{\tau+1} \ge \mu_{\tau+2} \ge \cdots \ge \mu_a, \\ \tau = 1, \ldots, a-1, \end{aligned}$$

has been proposed in [50]; here τ is an unknown turning point.

Some other extensions arise when responses show monotone relationships with the passage of time. Frequently encountered examples include the monotonic change of occurrence probabilities of some events, increasing treatment effects, and increasing **hazard rates** with time. For instance, the hypothesis

$$\mathrm{H}_6 : \mu_2 - \mu_1 \le \mu_3 - \mu_2 \le \cdots \le \mu_a - \mu_{a-1}$$

arises from the analysis of the age–period–cohort effects model (*see* **Age–Period–Cohort Analysis**) where only the second-order differences are estimable in each effect along with the time axis. Hypothesis H_6 is equivalent to $\mathrm{H}_6' : \mu_i - 2\mu_{i+1} + \mu_{i+2} \ge 0$, and may be called the "convexity hypothesis". Convexity, concavity, and sigmoidicity constraints are commonly employed also in the field of bioassay as reasonable shape constraints on a

Table 3 Usefulness in a dose-finding clinical trial

Drug	Undesirable	Slightly undesirable	Not useful	Slightly useful	Useful	Excellent
Placebo	3	6	37	9	15	1
AF 3 (mg/kg)	7	5	33	21	10	1
AF 6 (mg/kg)	5	6	21	16	23	6

dose–response relationship (*see* **Biological Assay, Overview**; **Quantal Response Models**).

As seen from the above examples, isotonic inference is closely related to **change-point** analysis. Actually, a one-sided change-point model may be formulated as a set of particular monotone relationships,

$$H_7 : \mu_1 = \cdots = \mu_\tau < \mu_{\tau+1} = \cdots = \mu_a, \quad \tau = 1, \ldots, a-1, \tag{3}$$

with τ an unknown change-point parameter, so that a useful statistic for change-point analysis is useful also for isotonic inference. Interestingly, (3) defines $a-1$ edges of the convex cone defined by the simple ordered alternative (1); see [24].

For other extensions, including **tree-structured**, star-shaped, unimodality, and symmetry models, the reader is referred to [3] and [43].

Testing a Simple Ordered Alternative in Normal Means

We wish to test a simple ordered alternative H_1 in the one-way layout model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, a; j = 1, \ldots, n_i,$$

where the ε_{ij} are assumed to be independently distributed as **normal** $N(0, \sigma^2)$ with known **variance** σ^2. Then there are two major streams of overall trend tests and multiple contrast type tests. Most cases of unknown variance can be dealt with similarly, if an **unbiased** variance estimator distributed as a multiple of χ^2 is available.

Overall Trend Tests

One possible approach is the restricted likelihood ratio test developed extensively in [3] (*see* **Isotonic Regression**). The approach does not, however, possess any obvious optimal property for such restricted alternatives, and is rather difficult to extend to higher-way problems.

Abelson & Tukey [1] proposed a linear score statistic which maximizes the minimum **power** in the region defined by H_1 within the class of linear tests. This has been extended to the most stringent and somewhere most powerful (MSSP) test for a more general restricted alternative by Schaafsma [44, 45]. In the balanced case, Abelson & Tukey's score is determined by equalizing powers at all the $a-1$ edges of H_1 and is given by

$$c_i \propto -\left[i\left(1-\frac{i}{a}\right)\right]^{1/2} + \left[(i-1)\left(1-\frac{i-1}{a}\right)\right]^{1/2}, \quad i = 1, \ldots, a.$$

Extending Taguchi's idea [52], the cumulative χ^2 test was introduced in [15], and its power has been compared with that of the previous two approaches. The test statistic χ^{*2} is the sum of squares of the standardized accumulated statistics

$$y_i^* = \frac{1}{\sigma}\left(\frac{1}{N_i} + \frac{1}{N_i^*}\right)^{-1/2}(\overline{Y}_i^* - \overline{Y}_i), \quad i = 1, \ldots, a-1, \tag{4}$$

where $N_i = n_1 + \cdots + n_i$, $N_i^* = n_{i+1} + \cdots + n_a$, and $\overline{Y}_i = (y_{1\cdot} + \cdots + y_{i\cdot})/N_i$, $\overline{Y}_i^* = (y_{i+1\cdot} + \cdots + y_{a\cdot})/N_i^*$ with $y_{i\cdot} = (y_{i1} + \cdots + y_{in_i})$, $i = 1, \ldots, a$. The χ^{*2} statistic is characterized by the strong positive **correlations** between the serial components y_i^*, and in particular by the expansion for the balanced case in a series of independent χ^2 variables,

$$\chi^{*2} = \frac{1}{1\cdot 2}\chi_{(1)}^2 + \frac{a}{2\cdot 3}\chi_{(2)}^2 + \cdots + \frac{a}{(a-1)\cdot a}\chi_{(a-1)}^2,$$

where $\chi_{(l)}^2$ is the 1 df χ^2 statistic for detecting the departure from the null model in the direction of Chebyshev's lth order orthogonal polynomial. Hence χ^{*2} tests mainly, but not exclusively, a linear trend; see [19] and [36] for details.

Multiple Contrast Type Tests

Several multiple comparison procedures have been proposed for ordered parameters. Williams [55] proposed a closed testing procedure based on the maximum likelihood estimator for defining the maximal noneffective dose level. Marcus [30] modified the method by changing the estimator at the control level from $\overline{y}_1$. to $\hat{\mu}_1$, the maximum likelihood estimator of μ_1, so that his statistic is the maximal component of Bartholomew's $\overline{\chi}^2$. The limiting distribution of the latter statistic is obtained in [56] and more recently, an exact recursive integration procedure for calculating its distribution

function is obtained in [29]. The maximal component of χ^{*2} has been proposed also for this purpose, and is called the "max t" method, where t stands for the y_i^* of (4). The statistic is characterized as the likelihood ratio test for the change-point hypothesis H_7, and an exact and very efficient algorithm for calculating the ***P* value** has been obtained by Hawkins [12] (*see* **Change-point Problem**). The power functions of these closed multiple testing procedures have been compared in [30], [49], and [26]. More general multiple tests for ordered parameters are obtained in [32].

Confidence Interval

A **confidence interval** taking advantage of order restrictions can be obtained by inverting an appropriate test for order restricted alternatives. For example, Marcus & Peritz [31], Schoenfeld [47] and Hirotsu & Srivastava [25] obtain confidence intervals for normal means by inverting a multiple contrast type test, the restricted likelihood ratio test, and the max t test, respectively. Wynn [59] gives a general methodology for obtaining one-sided confidence intervals, and Hayter [13] obtains confidence intervals based on the one-sided **studentized range** test. Miwa & Hayter [35] obtain confidence intervals for the differences of the ordered normal means in the one-way layout setting taking the advantages of the one- and two-sided procedures. In particular, in the bioassay problem, Schmoyer [46] obtains improved upper confidence bounds for the responses at very low doses by assuming sigmoidicity in the dose–response curve.

Hwang & Peddada [28] develop a methodology under a general order restriction, which has been extended recently to a test procedure by Peddada, Prescott and Conaway [37].

There is no extensive work on the design of experiments on the ordered parameters, although an optimal allocation has been discussed in [23] (*see* **Optimal Design**).

Applications

A test of Abelson & Tukey, the cumulative χ^2 test, and some of the multiple comparison procedures, are now applied to the data in Table 1. Since the variance σ^2 is unknown, it is replaced by the usual unbiased estimate of variance, $\hat{\sigma}^2 = \sum\sum(y_{ij} - \overline{y}_{i\cdot})^2/(24 - 5) = 0.020741$.

The linear score statistic of Abelson & Tukey is calculated as

$$(-c_1\overline{y}_{1\cdot} - c_2\overline{y}_{2\cdot} + c_3\overline{y}_{3\cdot} + c_4\overline{y}_{4\cdot} + c_5\overline{y}_{5\cdot})/\hat{\sigma} = 9.1206,$$

with scores $c_1 = c_5 = (\sqrt{6} + 1)/\sqrt{5} = 1.543$, $c_2 = c_4 = (4 - \sqrt{6})/\sqrt{20} = 0.3467$, and $c_3 = 0$, giving a P value of 2.2×10^{-8} as evaluated by the t distribution with 19 df.

The null distribution of cumulative χ^2 statistic $\sum y_i^{*2}$ is well approximated by $d\chi_f^2$, a multiple of the χ^2 variable with df f, where the constants d and f are given by

$$d = 1 + \frac{2}{a-1} \times \left(\frac{\lambda_1}{\lambda_2} + \frac{\lambda_1 + \lambda_2}{\lambda_3} + \cdots + \frac{\lambda_1 + \cdots + \lambda_{a-2}}{\lambda_{a-1}}\right),$$
$$f = \frac{a-1}{d}, \tag{5}$$

with $\lambda_i = N_i/N_i^*$. An even better approximation based on the expansions by Laguerres' orthogonal polynomials, and also the approximation under the alternative hypothesis, are given in [16]. Then the P value of the statistic

$$F^* = (a-1)^{-1}\chi^{*2}|_{\sigma^2=\hat{\sigma}^2} = 54.739$$

can be evaluated as 1.1×10^{-8} by the ***F* distribution** with df $(f, \sum n_i - a)$, where $f = 2.067$ from (5).

The maximal component of the $\chi^{*2}|_{\sigma^2=\hat{\sigma}^2}$ is obtained at the partition between levels 2 and 3:

$$\max t = \left[\left(\frac{1}{10} + \frac{1}{14}\right)(0.02741)\right]^{-1/2} \times \left(\frac{23.00}{14} - \frac{11.31}{10}\right) = 8.584,$$

the one-sided P value of which is evaluated as 1.1×10^{-7} by the recurrence formula based on the Markov property of y_i^*s. According to the closed testing procedure of [32], the process proceeds to the final step where the t statistic between levels 1 and 2 shows a nonsignificant result at the one-sided significance level 0.10, thus suggesting finally the difference between the dose levels (1,2) and (3,4,5).

In applying the Williams [55] procedure and the modified Williams procedure of [30], we need the maximum likelihood estimaters of the μ_i, which are

$$\hat{\mu}_1 = 1.076, \qquad \hat{\mu}_2 = 1.186,$$
$$\hat{\mu}_3 = \hat{\mu}_4 = 1.524, \quad \hat{\mu}_5 = 1.856,$$

by the pool-adjacent-violators method. Since $\hat{\mu}_1 = \overline{y}_{1\cdot}$, both statistics coincide and equal

$$w = \max \frac{\sqrt{m}(\hat{\mu}_i - \overline{y}_{1\cdot})}{\hat{\sigma}}$$
$$= \frac{\sqrt{m}(\overline{y}_{5\cdot} - \overline{y}_{1\cdot})}{\hat{\sigma}} = 8.357,$$

where we take the repetition number m as the harmonic mean of the n_is for referring approximately to the tables for upper percentiles in the balanced case by [55] and [56], respectively. In any case, the statistic w is highly significant and the closed testing procedure stops with the nonsignificant result between levels 1 and 2, thus again suggesting a difference between the dose levels (1, 2) and (3, 4, 5).

For a more general likelihood $L(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\nu})$ with the ordered parameter $\boldsymbol{\theta}$, and possibly with the **nuisance parameter** $\boldsymbol{\nu}$, arguments similar to those used above apply if the asymptotic normality of the likelihood estimators is assured. In particular, the cumulative χ^2 and the max t statistics can be based on the cumulative efficient scores evaluated at the null hypothesis and extended easily to two-way problems; see [7, 17, 18] for details.

Testing Ordered Alternatives in Binomial Probabilities

The data in Table 4 are from a dose–response **clinical trial**. Assuming that the y_i are independently distributed as **binomial** $B_{\text{in}}(n_i, p_i)$ we are interested in testing the simple **ordered alternative**

$$\mathrm{H}_1\colon\ p_1 \le \cdots \le p_a.$$

Table 4 Dose finding trial for a heart disease drug

Dose (mg/day)	Improved	Not improved
100	20	16
150	23	18
200	27	9
225	26	9
300	9	5

If the quantitative measures $d_1 < \cdots < d_a$ are attached to the y_i, then the locally most powerful test against a wide range of monotone relationships of p_i to d_i is obtained by Cochran [6] and Armitage [2] (*see* **Trend Test for Counts and Proportions**). For the case where there is no information on d_i, the likelihood ratio test has been developed by Chacko [5]. The tests based on the cumulative χ^2 and its maximal component have also been extended as follows:

$$\text{the cumulative } \chi^2 : \chi^{*2} = \sum y_i^{*2},$$
$$\text{the maximal component of } \chi^{*2} : \max t = \max y_i^*, \tag{6}$$

where y_i^* is given by (4) with σ replaced by $[\overline{Y}(1-\overline{Y})]^{1/2}$, $\overline{Y} = \sum y_i / \sum n_i$, and $y_{i\cdot}$ replaced by y_i in defining Y_i. Formula (5) is also valid for the χ^{*2} to give a two-sided P value of 0.113 when applied to Table 4. For max t, another exact algorithm is available based on the Markov property of the y_i^* to give a one-sided P value of 0.044 for Table 2 at the partition between levels (1, 2) and (3, 4, 5); see [26, 57, 58] for the algorithm. The Cochran–Armitage test gives a slightly larger one-sided P value of 0.049 since there is a slight downturn tendency in this example.

Analyzing the Two-Way Contingency Table with Ordered Column Categories

Two-Sample Problem

First consider the two-sample problem presented by Table 2. A popular approach to the analysis is to use a nonparametric test based on a linear score statistic such as Wilcoxon's (*see* **Wilcoxon–Mann–Whitney Test**). Now, for the two-sided alternative H_2 the two statistics,

$$\text{the cumulative } \chi^2 : \chi^{*2} = \chi_1^2 + \cdots + \chi_{b-1}^{*2}, \tag{7}$$
$$\text{the maximal component of } \chi^{*2} :$$
$$\max \chi^2 = \max \chi_j^2, \tag{8}$$

can be defined in terms of the accumulated efficient scores, where χ_j^2 is the goodness-of-fit χ^2 statistic for the **2 × 2 table** formed by accumulating the first j and the remaining $b - j$ columns. The χ_j^2 is, however, identical to the y_i^{*2} of (6) if the binomial data

are arranged in a $2 \times b$ table in an obvious way and exactly the same distribution theory applies also to this case. The two-sided P values are 0.039 for χ^{*2} and 0.154 for max χ^2, whereas it is 0.025 for the Wilcoxon test. In this case the Wilcoxon test shows the smallest P value, since approximately a linear trend is observed in p_{1j}/p_{2j}, $j = 1, \ldots, 4$. If these tests are applied to the last two rows of Table 3 for comparing AF 3 mg and AF 6 mg, then the two-sided P values are 0.0128, 0.0096, and 0.0033 for the Wilcoxon, the χ^{*2}, and max χ^2 methods, respectively. It has been verified by **simulation** that, when evaluated as two-sample nonparametric tests, the Wilcoxon method is useful for the location shift of the underlying symmetrical and light-tailed distributions such as the **logistic** or normal, the max χ^2 method is useful for skewed or heavy-tailed distributions, and the cumulative χ^2 method is characterized by its **robustness**, having relatively high **power** over a wide range of underlying distributions – normal, heavy-tailed, or skewed.

Another important approach to the problem is to assume an underlying continuous distribution for each treatment and to compare the parameters describing those distributions. The **proportional-odds** and **proportional-hazards** models are important examples; see [34] for details.

General a-*Sample Problem*

For a general a-sample problem the Wilcoxon test is extended to the Kruskal–Wallis test. The same type extensions are available for the χ^{*2} and its maximal component by defining the χ_j^2 in (7) and (8) as the goodness-of-fit χ^2 statistic for the accumulated $a \times 2$ table for the partition between columns j and $j+1$. The constants for the χ^2 approximation of χ^{*2} are obtained by

$$d = 1 + \frac{2}{b-1} \times \left(\frac{\gamma_1}{\gamma_2} + \frac{\gamma_1 + \gamma_2}{\gamma_3} + \cdots + \frac{\gamma_1 + \cdots + \gamma_{b-2}}{\gamma_{b-1}} \right),$$

$$f = \frac{(a-1)(b-1)}{d},$$

with $\gamma_j = C_j / C_j^*$, $C_j = y_{\cdot 1} + \cdots + y_{\cdot j}$, and $C_j^* = y_{\cdot j+1} + \cdots + y_{\cdot b}$. The max χ^2 can be evaluated by the calculation **algorithm** based on the Markov property of the subsequent χ_j^2s [26].

For the row-wise multiple comparisons based on the cumulative χ^2, the statistic

$$S = \max ||(\mathbf{a}' \otimes \mathbf{C}^{*'})z||^2$$

is defined where $\otimes$ is a Kronecker product, $\mathbf{z}$ a vector of $\sqrt{y_{\cdot\cdot}}\, y_{ij} / (y_{i\cdot} y_{\cdot j})^{1/2}$ arranged in dictionary order, $\mathbf{C}^{*'}$ a $b-1 \times b$ matrix defined so that the (j, j') th element of $\mathbf{C}^{*'}\mathbf{C}^*$ is $(\gamma_j / \gamma_{j'})^{1/2}$ for $j \le j'$ and the maximum is taken over all $\mathbf{a}$ that satisfy $\mathbf{a}'\mathbf{a} = 1$, and $(\sqrt{y_{1\cdot}}, \ldots, \sqrt{y_{a\cdot}})\, \mathbf{a} = 0$. When $a \ge b$ and under the null model, the statistic S is asymptotically distributed as the largest root of the Wishart matrix $W(\mathbf{C}^{*'}\mathbf{C}^*, a-1)$, which is well approximated by $\gamma_{(1)} \chi^2 (a-1)$ with $\gamma_{(1)}$ the largest root of $\mathbf{C}^{*'}\mathbf{C}^*$. The statistic S gives the Scheffé-type multiple comparison test, and has been applied to taste-testing data of five foods in five ordered categorical responses of [4] to obtain the significant classification of rows (foods) (1, 2), (3, 4) and (5). The max χ^2 is also applied to the data for multiple comparisons of the columns to obtain a highly significant classification (1, 2, 3) and (4, 5). The resulting block interaction model is expressed as

$$p_{ij} = p_{i\cdot} p_{\cdot j} q_{\mu\nu}, \quad \mu = 1, 2, 3, \nu = 1, 2,$$

if i belongs to the μth subgroup of rows and j to the νth subgroup of columns. The goodness-of-fit χ^2 has been compared with the fitting of the proportional-odds model [51] and its extension [33]; see [20] for details. The Scheffé-type multiple comparison method is applied to the normal distribution model in [21], for classifying subjects based on the upward, flat, and downward tendencies defined by repeated measurements.

Two-Way Contingency Table with Natural Orderings in Both Rows and Columns

Assuming a multinomial model $M(y_{\cdot\cdot}, p_{ij})$ for the data y_{ij} in Table 3 the cumulative χ^2 statistic and its maximal component are defined for testing H_3 using the cumulative efficient scores evaluated at the null hypothesis. These are

the doubly cumulative $\chi^2 : \chi^{**2} = \sum \sum \chi_{ij}^2$,

the maximal component of $\chi^{**2} : \max \max \chi_{ij}^2$,

with the χ_{ij}^2 being the goodness-of-fit χ^2 for the 2×2 tables obtained from partitioning and accumulating rows and columns at $i = 1, \ldots, a-1$, and $j = 1, \ldots, b-1$, respectively. The χ^{**2} is for the two-sided version of H_3, and max max χ^2 is applicable to both one- and two-sided problems. When applied to Table 3 the two-sided P values are approximately 0.0065 for the χ^{**2} and exactly 0.0142 for max max χ^2. The details of the P value calculations are given in [22]. As a semiparametric model for the ordered two-way table the constant-**odds ratio** model has been proposed by Wahrendolf [54] based on Plackett's [38] coefficient of association for **bivariate distributions** (*see* **Association, Measures of**).

As an example of the higher-way layouts, a $2 \times J \times K$ table comparing two treatments based on bivariate allele frequencies, is analyzed in [27]. An example of highly fractional factorial experiments with ordered categorical responses is given in [11]; see also the discussion following that article (*see* **Factorial Experiments**).

Bayesian Approach to Isotonic Inference

Since the purpose of an isotonic inference is to make use of the prior knowledge to enhance the efficiency of test and estimation, it is natural to consider a Bayesian approach. For example, an essentially complete class of tests for orderly constrained hypothesis is obtained as the whole set of Bayes tests with a **prior distribution** defined on those constrained supports. The cumulative χ^2 and max t methods are derived from this idea; see [7, 17, 53]. More specifically, in bioassay problems, the Dirichlet prior has been introduced for the successive differences of the responses for doses d_i, $p(d_i) - p(d_{i-1}), i = 1, \ldots, a+1$; $p(d_0) = 0$, $p(d_{a+1}) = 1$, reflecting the nondecreasing nature of the dose–response relationship, see [40, 41], for example. Shaked & Singpurwalla [48] discuss the defect of the Dirichlet prior, and introduce concavity constraints on the shape of a dose–response curve, reflecting a situation encountered in practice. Because of computational difficulties, however, they are unable to compute posteriors beyond modal estimates. The computational problem was overcome later by Gelfand & Kuo [8], who showed how a sampling-based approach could be used to develop the desired marginal posterior distributions and their features, for Dirichlet and product–beta priors; see also [9]. Ramgopal et al. [39] consider convex, concave, and ogive constraints to specify the shape of dose–response curves, and extend the sampling-based approach to calculating any posterior feature of interest in these generalized constrained problems.

References

[1] Abelson, R.P. & Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: general theory and the case of the simple order, *Annals of Mathematical Statistics* **34**, 1347–1369.

[2] Armitage, P. (1955). Test for linear trends in proportions and frequencies, *Biometrics* **11**, 375–386.

[3] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, Chichester.

[4] Bradley, R.A., Katti, S.K. & Coons, I.J. (1962). Optimal scaling for ordered categories, *Psychometrika* **27**, 355–374.

[5] Chacko, Y.C. (1966). Modified chi-square test for ordered alternatives, *Sankhyā, Series B* **28**, 185–190.

[6] Cochran, W.G. (1954). Some methods for strengthening the common χ^2 test, *Biometrics* **10**, 417–451.

[7] Cohen, A. & Sackrowitz, H.B. (1991). Tests for independence in contingency tables with ordered categories, *Journal of Multivariate Analysis* **36**, 56–67.

[8] Gelfand A.E. & Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response, *Biometrika* **78**, 657–666.

[9] Gelfand, A.E., Smith, A.F. & Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling, *Journal of the American Statistical Association* **87**, 523–532.

[10] Geyer, C.J. (1991). Constrained maximum likelihood exemplified by isotonic convex logistic regression, *Journal of the American Statistical Association* **86**, 717–724.

[11] Hamada, M. & Wu, C.F.J. (1990). A critical look at accumulation analysis and related methods (with discussion), *Technometrics* **32**, 119–130.

[12] Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location, *Journal of the American Statistical Association* **72**, 180–186.

[13] Hayter, A.J. (1990). A one-sided Studentized range test for testing against a simple ordered alternative, *Journal of the American Statistical Association* **85**, 778–785.

[14] Hirotsu, C. (1978). Ordered alternatives for interaction effects, *Biometrika* **65**, 561–570.

[15] Hirotsu, C. (1979). The cumulative chi-squares method and Studentized maximal contrast method for testing an ordered alternative in a one-way analysis of variance model, *Reports of Statistical Application Research, Union of Japanese Scientists and Engineers* **26**, 12–21.

[16] Hirotsu, C. (1979). An F-approximation and its application, *Biometrika* **66**, 577–584.

[17] Hirotsu, C. (1982). Use of cumulative efficient scores for testing ordered alternatives in discrete models, *Biometrika* **69**, 567–577.

[18] Hirotsu, C. (1983). Defining the pattern of association in two-way contingency tables, *Biometrika* **70**, 579–589.

[19] Hirotsu, C. (1986). Cumulative chi-squared statistic as a tool for testing goodness of fit, *Biometrika* **73**, 165–173.

[20] Hirotsu, C. (1990). Discussion on Hamada and Wu's paper, *Technometrics* **32**, 133–136.

[21] Hirotsu, C. (1991). An approach to comparing treatments based on repeated measures, *Biometrika* **78**, 583–594.

[22] Hirotsu, C. (1997). Two-way change-point model and its application, *Australian Journal of Statistics* **39**, 205–218.

[23] Hirotsu, C. & Herzberg, A.M. (1987). Optimal allocation of observations for inference on k ordered normal population means, *Australian Journal of Statistics* **29**, 151–165.

[24] Hirotsu, C. & Marumo, K. (2002). Changepoint analysis as a method for isotonic inference, *Scandinavian Journal of Statistics* **29**, 125–138.

[25] Hirotsu, C. & Srivastava, M.S. (2000). Simultaneous confidence intervals based on one-sided max t test, *Statistics & Probability Letters* **49**, 25–37.

[26] Hirotsu, C., Kuriki, S. & Hayter, A.J. (1992). Multiple comparison procedure based on the maximal component of the cumulative chi-squared statistic, *Biometrika* **79**, 381–392.

[27] Hirotsu, C., Aoki, S., Inada, T. & Kitao, Y. (2001). An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis, *Biometrics* **57**, 769–778.

[28] Hwang, J.T.G. & Peddada, S. (1994). Confidence interval estimation subject to order restrictions. *Annals of Statistics* **22**, 67–93.

[29] Kuriki, S., Shimodaira, H. & Hayter, T. (2002). On the isotonic range statistic for testing against an ordered alternative, *Journal of Statistical Planning and Inference* **105**, 347–362.

[30] Marcus, R. (1976). The powers of some tests of the equality of normal means against an ordered alternative, *Biometrika* **63**, 177–183.

[31] Marcus, R. & Peritz, E. (1976). Some simultaneous confidence bounds in normal models with restricted alternatives, *Journal of the Royal Statistical Society, Series B* **38**, 157–165.

[32] Marcus, R., Peritz, E. & Gabriel, K.R. (1976). On closed testing procedure with special reference to ordered analysis of variance, *Biometrika* **63**, 655–660.

[33] McCullagh, P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[34] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[35] Miwa, T. & Hayter, T. (1999). Combining the advantages of one-sided and two-sided test procedures for comparing several treatment effects, *Journal of the American Statistical Association* **94**, 302–307.

[36] Nair, V.N. (1986). On testing against ordered alternatives in analysis of variance models, *Biometrika* **73**, 493–499.

[37] Peddada, S.D., Prescott, K.E. & Conaway, M. (2001). *Biometric* **57**, 1219–1227.

[38] Plackett, R.L. (1965). A class of bivariate distributions, *Journal of the American Statistical Association* **60**, 516–522.

[39] Ramgopal, P., Laud, P.W. & Smith, A.F.M. (1993). Nonparametric Bayesian bioassay with prior constraints on the shape of the potency curve, *Biometrika* **80**, 489–498.

[40] Ramsey, F.L. (1972). A Bayesian approach to bioassay, *Biometrics* **28**, 841–858.

[41] Ramsey, F.L. (1973). Correction, *Biometrics* **29**, 830.

[42] Robert, C.P. & Hwang, J.T.G. (1996). Maximum likelihood estimation under order restrictions by the prior feedback method, *Journal of the American Statistical Association* **91**, 167–172.

[43] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[44] Schaafsma, W. (1966). Hypothesis testing problems with the alternative restricted by a number of inequalities, *Doctoral dissertation.* University of Groningen, Noordhoff, Groningen.

[45] Schaafsma, W. (1968). A comparison of the most stringent and the most stringent somewhere most powerful tests for certain problems with restricted alternatives, *Annals of Mathematical Statistics* **39**, 531–546.

[46] Schmoyer, R.L. (1984). Sigmoidally constrained maximum likelihood estimation in quantal bioassay, *Journal of the American Statistical Association* **79**, 448–453.

[47] Schoenfeld, D.A. (1986). Confidence bounds for normal means under order restrictions, with application to dose-response curves, toxicology experiments, and low-dose extrapolation, *Journal of the American Statistical Association* **81**, 186–195.

[48] Shaked M. & Singpurwalla, N.D. (1990). A Bayesian approach for quantile and response probability estimation with applications to reliability, *Annals of the Institute of Statistical Mathematics* **42**, 1–19.

[49] Shirley, E. (1979). The comparison of treatment with control group means in toxicological studies, *Applied Statistics* **28**, 144–151.

[50] Simpson, D.G. & Margolin, B.H. (1986). Recursive nonparametric testing for dose response relationships subject to downturn at high doses, *Biometrika* **73**, 589–596.

[51] Snell, E.J. (1964). A scaling procedure for ordered categorical data, *Biometrics* **20**, 592–607.

[52] Taguchi, G. (1966). *Statistical Analysis* (in Japanese). Maruzen, Tokyo.

[53] Takeuchi, K. (1979). Test and estimation problems under restricted null and alternative hypotheses (in Japanese), *Journal of Economics* **45**, 2–10.

[54] Wahrendolf, J. (1980). Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions, *Biometrika* **76**, 15–21.

[55] Willams, D.A. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control, *Biometrics* **27**, 103–117.

[56] Williams, D.A. (1977). Some inference procedures for monotonically ordered normal means, *Biometrika* **64**, 9–14.

[57] Worsley, K.J. (1983). The power of likelihood ratio and cumulative sum tests of a change in a bionomial probability, *Biometrika* **70**, 455–464.

[58] Worsley, K.J. (1986). Confidence regions and tests for a change point in a sequence of exponential family of random variables, *Biometrika* **73**, 91–104.

[59] Wynn, H.P. (1975). Integrals for one-sided confidence bounds: A general result, *Biometrika* **62**, 393–396.

C. HIROTSU

Isotonic Regression

Many regression problems involve minimizing a weighted sum of squares subject to the restriction that the solution must satisfy certain side conditions. A function $f(x)$ defined on a finite index set of numbers $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$ is isotonic or order preserving if x, y are in $\mathcal{X}$ and $x < y$ implies $f(x) \leq f(y)$. Isotonic regression minimizes a weighted sum of squares subject to the condition that the regression function is isotonic. A simple example is a one-way **analysis of variance** for ordered normal means, in which the variable $x_i = i$ indexes the groups. With each point x_i in $\mathcal{X}$ we associate a positive weight w_i, usually the number of observations on which it is based. Suppose that $g(x)$ is a given function defined on $\mathcal{X}$, then the isotonic regression of $g(x)$ with weights $w_1, w_2, \ldots, w_k$ is denoted by $g^*(x)$ and minimizes the weighted sum of squares,

$$\sum_{i=1}^{k} w_i[g(x_i) - f(x_i)]^2,$$

in the class of all isotonic functions $f(x)$ defined on $\mathcal{X}$. Note that $g^*(x)$ is the isotonic function closest to $g(x)$ as measured in weighted **least squares** distance. For a one-way analysis of variance for nondecreasing ordered normal means, $g(x_i)$ is the observed mean for group i, w_i is the number of observations on which the mean is based and $f(x_i) = \mu_i$ is the mean for group i and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$. For the simple case of ordinary **linear regression** with a single independent variable, $g(x_i)$ is the mean value of the independent **explanatory** variable at x_i, w_i is the number of observations on which it is based, and $f(x) = a + bx$ is a linear regression function. Isotonic regression allows $f(x)$ to be any isotonic function rather than restricting the regression function $f(x)$ to be linear. While isotonic regression can be viewed as a smoothing procedure, one disadvantage is that the isotonic estimates are essentially step functions and, hence, are not smooth everywhere. The degree of smoothness depends on the type of assumptions made; for example, that the function is nondecreasing or convex. Another disadvantage is that the isotonic estimators are biased. Isotonic regression is important because it provides **maximum likelihood** estimators for a large class of problems involving ordered parameters, as well as solving many more constrained statistical problems than the weighted least squares problem stated above. One simple example is unimodal simple regression, which consists of an up-phase in which $\mathrm{E}(Y|X = x)$ is increasing with x and a down-phase in which $\mathrm{E}(Y|X = x)$ is decreasing with x. If the turning point is known, then it is possible to use isotonic regression for each phase separately. If the turning point is unknown, a simple modification of this idea yields the solution (see [3] for details and examples).

A number of efficient **algorithms** for isotonic regression are available, especially for the case of a single independent variable (see [2] for details). The pooled-adjacent-violators algorithm is widely used, but is only applicable for the case of a simple order. A simple order is when $x_1 < x_2 < \cdots < x_k$, and this implies $f(x_1) \leq f(x_2) \leq \cdots \leq f(x_k)$. This algorithm basically involves, possibly repeated, weighted averages of the unconstrained estimates. For recent extensions of this algorithm, including the case of concave regression and additive isotonic models, see [5] and [1], respectively. Other types of ordering such as quasi- and partial ordering exist; for example, when we have more than one independent variable, partial orderings, which deal with situations such as noncomparable elements in $\mathcal{X}$, arise (see [4] for details on types of ordering).

References

[1] Bacchetti, P. (1989). Additive isotonic models, *Journal of the American Statistical Association* **84**, 289–294.

[2] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

[3] Frisen, M. (1986). Unimodal regression, *Statistician* **35**, 479–485.

[4] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[5] Tang, D. & Lin, S.P. (1991). Extension of the pool-adjacent-violators algorithm, *Communications in Statistics – Theory and Methods* **20**, 2633–2643.

(*See also* **Isotonic Inference**)

JOHN W. MCDONALD

Iterative Proportional Fitting

Iterative proportional fitting (IPF), also known as iterative proportional scaling, is an **algorithm** for constructing tables of numbers satisfying certain constraints. In its simplest form, the algorithm enables one to construct two-way **contingency tables** with specified marginal totals and a prescribed degree of association; from a more general perspective, it may be viewed as a cyclic ascent algorithm which maximizes a specific objective function. The algorithm can also be used to construct **maximum likelihood** estimators for table entries based upon hierarchical **log-linear models** for **Poisson**, **multinomial**, or product multinomial models. We will illustrate these aspects of the algorithm and its applications by describing some simple cases.

Suppose that we are given two pairs $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ of positive numbers satisfying $u_1 + u_2 = v_1 + v_2$, and a further positive number ψ. The IPF algorithm will enable us to construct the *unique* **two-by-two table** $\mathbf{b} = (b_{ij})$ such that, for all i and j,

$$b_{i+} = u_i, \qquad b_{+j} = v_j, \qquad \frac{b_{11}b_{22}}{b_{12}b_{21}} = \psi,$$

where the subscript $+$ denotes the result of summing over the subscript it replaces. The algorithm goes like this. Begin with the 2×2 table $\mathbf{a} = (a_{ij})$ defined by $a_{11} = \psi$, $a_{12} = a_{21} = a_{22} = 1$, noting that the cross-ratio $a_{11}a_{22}/a_{12}a_{21} = \psi$ (*see* **Odds Ratio**). Next, scale the rows of $\mathbf{a}$ to form the table $\mathbf{a}' = (a'_{ij})$:

$$a'_{ij} = a_{ij} \times \frac{u_i}{a_{i+}}, \tag{1}$$

for $i = 1, 2$ and $j = 1, 2$. It is easy to check that $\mathbf{a}'$ has the desired row sums, as well as having cross-ratio ψ. We now scale the columns of $\mathbf{a}'$ to form the table $\mathbf{a}'' = (a''_{ij})$:

$$a''_{ij} = a'_{ij} \times \frac{v_j}{a'_{+j}}. \tag{2}$$

One can check that $\mathbf{a}''$ has the desired column sums and cross-ratio, although the row sums are no longer (u_i). This completes one cycle of the IPF algorithm, beginning with the table $\mathbf{a}$.

The algorithm continues by repeatedly scaling the rows, as in (1), and then the columns, as in (2), to have the desired totals. After a number of cycles, the row totals are closer to (u_i) than they were initially, the column totals are exactly (v_j), and the cross-ratio is exactly ψ. The sequence of tables so defined converges pointwise to a 2×2 table $\mathbf{b}$ with all the desired properties; uniqueness also follows.

It is instructive to examine why these assertions are true, for in doing so we obtain further insights into the IPF algorithm. To do this, we introduce the notion of **information** (or I-) divergence between two tables $\mathbf{c} = (c_{ij})$ and $\mathbf{d} = (d_{ij})$, satisfying $c_{++} = d_{++}$, defined as follows:

$$I(\mathbf{c}|\mathbf{d}) = \sum_{ij} c_{ij} \log\left(\frac{c_{ij}}{d_{ij}}\right).$$

(A similar definition applies to singly indexed arrays.) It can be proved that $I(\mathbf{c}|\mathbf{d}) \geq 0$, and that $I(\mathbf{c}|\mathbf{d}) = 0$ if and only if $\mathbf{c} = \mathbf{d}$. Although not a symmetric function of its arguments, I behaves in many ways like a metric on tables, and it provides the basis of a proof of convergence of the IPF algorithm. We return to our construction of a table $\mathbf{b}$ having row totals $\mathbf{u}$, column totals $\mathbf{v}$, and cross-ratio ψ. First define the table $\mathbf{c} = (c_{ij})$ as follows:

$$c_{ij} = \frac{u_i v_j}{w},$$

where $w = u_+ = v_+$. The tables $\mathbf{a}$, $\mathbf{a}'$, $\mathbf{a}''$, ... become closer to $\mathbf{c}$ as the iterations continue, closeness here being in the sense of I-divergence. More precisely, we can check that

$$I(\mathbf{c}|\mathbf{a}) = I(\mathbf{c}|\mathbf{a}'') + I(\mathbf{v}|\mathbf{a}'_2) + I(\mathbf{u}|\mathbf{a}_1), \tag{3}$$

where $\mathbf{a}_1 = (a_{i+})$ and $\mathbf{a}'_2 = (a'_{+j})$. The convergence and uniqueness assertions above all follow from repeated use of this expansion and the stated properties of I. As long as there exists at least one table $\mathbf{c}$ with the desired marginal totals, we can begin the IPF algorithm with any table having the desired cross-ratio, and expect to converge to the stated limit. The repeated scaling gives tables closer and closer in the sense of I-divergence to the table $\mathbf{c}$, all the while retaining the original cross-ratio, and the row and column totals converge to their desired values.

All of the discussion so far applies with minimal changes to $r \times s$ tables; in the more general case, there are further cross-ratios to take into account.

Whereas in a 2×2 table there is only one cross-ratio whose value can be fixed, in an $r \times s$ table, there are $(r-1)(s-1)$ multiplicatively independent cross-ratios. A convenient set (cf. [12]) is the following:

$$\psi_{ij} = \frac{b_{ij} b_{rs}}{b_{is} b_{rj}}, \quad i = 1, \ldots, r-1; j = 1, \ldots, s-1.$$

Here we constructed our cross-ratios in relation to the index values r and s. Other choices give equivalent results; indeed there are quite different ways of defining the quantities which are preserved. This issue is addressed in the theory of **loglinear models**; see [2, 11] and [12]. Given an arbitrary set of $(r-1)(s-1)$ positive numbers (ψ_{ij}), and positive numbers $\mathbf{u} = (u_i)$ and $\mathbf{v} = (v_j)$ satisfying $u_+ = v_+$, the IPF algorithm may be initiated with the table $\mathbf{a} = (a_{ij})$ given by $a_{ij} = \psi_{ij}, i = 1, \ldots, r-1; j = 1, \ldots, s-1$, and $a_{rj} = 1 = a_{is}, i = 1, \ldots, r; j = 1, \ldots, s$. With this initial table, the steps are just as before, and the resulting sequence of tables converges to the unique table having row totals (u_i), column totals (v_j), and cross-ratios (ψ_{ij}).

We turn now to reasons for constructing such tables. One is simply to demonstrate the fact that the row totals, column totals, and cross-ratios of two-way tables may be specified independently, and to show how to obtain tables with arbitrarily specified (but consistent) values of these quantities. Historically, the algorithm was first used to adjust sample frequencies to expected marginal totals. In the examples in Deming [5], we have a table $\mathbf{n} = (n_{ij})$ based upon a **sample survey**, and marginal totals (N_{i+}) and (N_{-j}), but *not* the individual cell frequencies $\mathbf{N} = (N_{ij})$, from a census of the population. The result of applying the IPF algorithm with initial table $\mathbf{n}$, and desired marginal totals (N_{i+}) and (N_{+j}), can then be regarded as an estimate of what would have been obtained by cross-tabulating the entire population, instead of only a sample thereof. A modern treatment of these ideas can be found in [2], where the procedure is known as *raking* the table $\mathbf{n}$. The third application of the algorithm we note is to the construction of maximum likelihood estimates of table entries under loglinear models. We simply describe the results here; the reader may consult standard references such as [2, 11], or [1] for fuller details. Suppose that $\mathbf{n} = (n_{ij})$ is a two-way table of independent Poisson counts with parameters $\boldsymbol{\lambda} = (\lambda_{ij})$. Then the maximum likelihood estimate $\hat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ under the *multiplicative model* for the (λ_{ij}), has the same row and column totals as $\mathbf{n}$, and all cross-ratios equal to 1. In this case, the IPF algorithm begins with a table all of whose entries are 1, and scales the row and column totals to match those of the data $\mathbf{n}$. The algorithm converges after a single cycle to the unique maximum likelihood estimator $\hat{\boldsymbol{\lambda}}$.

Three- and Higher-Way Tables

There are a number of ways in which the IPF algorithm may be used with three-way tables. We illustrate two of these. Suppose that we have an $r \times s$ table $\mathbf{u} = (u_{ij})$ and an $s \times t$ table $\mathbf{v} = (v_{jk})$ of positive numbers satisfying $u_{+j} = v_{j+}$ for $j = 1, \ldots, s$. By analogy with our earlier construction, we might be interested in obtaining an $r \times s \times t$ table $\mathbf{b} = (b_{ijk})$ having

$$b_{ij+} = u_{ij}, \qquad b_{+jk} = v_{jk}.$$

This can be solved rather straightforwardly. For example, the table $\mathbf{c} = (c_{ijk})$ given by

$$c_{ijk} = \frac{u_{ij} v_{jk}}{w_j},$$

where $w_j = u_{+j} = v_{j+}, j = 1, \ldots, s$, is readily checked to have ij-margin $\mathbf{u}$ and jk-margin $\mathbf{v}$.

Of course, this is not the end of the story. We may also be interested in any further structure concerning the table $\mathbf{b}$ which may be specified, in addition to these marginal totals. It turns out that we may also ask that the table has predetermined values of certain cross-ratios. In this example, and more generally, we need rules to tell us which marginal totals and which cross-ratios can be specified independently. The issue is best discussed in the language of **hierarchical** loglinear models for multiway tables, where these are commonly described in terms of the marginal subtables which constitute the **sufficient statistics** for the models (under either independent Poisson, multinomial, or independent multinomial sampling). We refer to [1, 2], and [11] for details concerning these models. In this language, the cross-ratios that we have been specifying are the antilogarithms of elements of subspaces **orthogonal** to those that define the hierarchical loglinear model corresponding to the specified marginal totals. For example, by specifying margins corresponding to the indices

ij and jk, as we did in our example, we are also able to specify independently cross-ratios corresponding to the pair ik and the triple ijk – that is, all interactions other than those involved in the loglinear model defined by the prescribed marginal totals.

Now let us suppose that, in addition to **u** and **v** as above, we are given a $t \times r$ table $\mathbf{w} = (w_{ki})$ of positive numbers satisfying $w_{k+} = v_{+k}$ and $w_{+i} = u_{i+}$ for all k and i. Can we use IPF to construct a table $\mathbf{b} = (b_{ijk})$ satisfying

$$b_{ij+} = u_{ij}, \quad b_{+jk} = v_{jk}, \qquad b_{i+k} = w_{ki},$$

and having prescribed values for the ijk cross-ratios? One might think that this would be quite straightforward. Begin with a suitable initial table **a**. Then scale to achieve the ij, jk, and ki marginal totals **u**, **v**, and **w**, respectively. One cycle of the algorithm would be three such scalings, and after a few cycles, we might expect to have a table with the specified cross-ratios, and essentially the desired marginal totals.

How can this version of IPF go wrong? A clue is provided by our indication of the method used to prove that IPF converges. We made use of the existence of a table **c** satisfying the marginal constraints, and then everything followed. However, in the case of three-way tables, it is not hard to specify three consistent, positive two-way tables, for which *no* three-way table exists having positive entries, and the three specified tables as two-way marginal totals. A simple example is given by three 2×2 tables each having 1 in the diagonal cells and 2 in the off-diagonal cells. Although they are clearly consistent, it is easy to check that no $2 \times 2 \times 2$ table can exist with positive entries and these margins. Use of the IPF algorithm with an initial table whose entries are all 1, and these three marginal tables, results in a cycle through the same three tables. The tables constructed do not converge. Summarizing this discussion, we can say that only if there exists a three-way table with the given two-way tables as marginal totals is the IPF algorithm guaranteed to converge to a limiting table with the desired marginal totals and three-way cross-ratios. When it does, this table is uniquely specified by these properties.

We note that in the application of this result to maximum likelihood estimation with loglinear models, the assumption of the existence of *some* table with the given marginal totals is trivially satisfied as long as the observed table $\mathbf{n} = (n_{ijk})$ has positive entries, for in this case **n** itself suffices. If the observed table has some zero entries, but positive two-way marginal totals, the IPF algorithm still converges, but to a table with some zero entries. In a sense, this is an extended maximum likelihood estimator: one on the boundary of the natural parameter space.

The foregoing discussion applies without change to higher-way tables. For example, suppose that we have an initial four-way table $\mathbf{a} = (a_{ijkl})$, and we wish to scale it to have prescribed ij, jk, kl, and li marginal totals. What cross-ratios (equivalently, what loglinear structure) of this initial table will be preserved throughout the iterations, and could therefore be specified independently of the marginal totals? The answer is: all interactions other than those involved in the loglinear model defined by the prescribed marginal totals, that is, the ik, jl, ijk, ijl, ikl, jkl, and $ijkl$ interactions. Note that we still need to know that there exists a table with the specified marginal totals before the algorithm is guaranteed to converge to a limiting table with all the desired properties.

Finite Termination: Decomposable Models

Decomposable models are a class of loglinear models for complete multiway tables which possess closed-form expressions for their MLEs under the standard sampling models; see [11] and [2]. It turns out that the IPF algorithm behaves rather well for this class of models. Suppose that a set of marginal totals to be fitted via IPF defines a decomposable loglinear model. If the initial table is constant, and the margins to be fitted are taken in a suitable order, the algorithm converges after just one cycle. Furthermore, there *always* exists a table with the given set of tables as marginal subtables, when the corresponding model is decomposable. Finally, as long as the specified tables are all positive, the table whose existence has just been described has positive entries.

History

Fienberg [7] presents a discussion of the history of the IPF algorithm. Some additional references can be found in [8]. The most important early papers are [6] and [14].

Numerical Aspects

Haberman [11] proves that tables constructed by the IPF algorithm converge to their limit at a geometric (also called first-order) rate. This means that, asymptotically, the difference between the nth iterate and the limit is bounded above by ρ^n for some ρ between zero and unity. (This compares unfavorably with the behavior of Newton or modified Newton algorithms, which typically exhibit what is known as quadratic convergence.) In many cases, ρ may be quite close to unity, and so convergence may be rather slow, giving rise to a literature concerning speeding up of the algorithm. However, at that point, the algorithm ceases to be the one we are discussing.

The great advantage of the IPF algorithm is its simplicity, stability, and economy of space. When a table is large, and the number of iterations is not a limiting factor, it is the method of choice for the problems we have discussed. For other problems, such as the calculation of MLEs under loglinear models, Newton-type methods are preferred, because of their speed of convergence and the fact that variance–**covariance matrices** are an automatic byproduct. FORTRAN IV versions of the IPF algorithm can be found in [9] and [10].

Variants and Generalizations

It is implicit in the foregoing discussion that the tables being considered are all *complete*, that is, are fully rectangular, or rectangular parallelepipeds, etc., and have no so-called **structural zeros**. This was because the algorithm is mostly used, and its properties are most easily discussed, in that context. However, variant forms of the algorithm are used successfully with tables having a variety of other structures, and preserving features corresponding to models other than hierarchical loglinear models; see [11].

For generalizations of a different kind, see [3, 4, 13]. In these papers, applications of the algorithm beyond contingency tables are given, and its connections to the information measure I and entropy are more fully explored.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

[3] Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems, *Annals of Probability* **3**, 146–158.

[4] Darroch, J.N. & Ratcliff, D. (1972). Generalized iterative scaling for loglinear models, *Annals of Mathematical Statistics* **43**, 1470–1480.

[5] Deming, W.E. (1964). *Statistical Adjustment of Data*. Dover, New York.

[6] Deming, W.E. & Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics* **11**, 427–444.

[7] Fienberg, S.E. (1970). An iterative procedure for estimation in contingency tables, *Annals of Mathematical Statistics* **41**, 907–917.

[8] Fienberg, S.E. & Meyer, M.M. (1983). *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, p. 2275.

[9] Haberman, S.J. (1972). Loglinear fit for contingency tables, *Applied Statistics* **21**, 218–225.

[10] Haberman, S.J. (1973). Printing multidimensional tables, *Applied Statistics* **22**, 118–126.

[11] Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

[12] Plackett, R.L. (1981). *The Analysis of Categorical Data*, 2nd Ed. Griffin, London.

[13] Ruschendorf, L. (1995). Convergence of the iterative proportional fitting procedure, *Annals of Statistics* **23**, 1160–1174.

[14] Stephan, F.F. (1942). An iterative method of adjusting sample frequency tables when the expected marginal totals are known, *Annals of Mathematical Statistics* **13**, 166–178.

(*See also* **Categorical Data Analysis**)

TERRY P. SPEED

Jackknife Method

The primary purpose of this technique is the estimation of the **standard errors** and the **bias** of estimators, $T(\mathbf{x})$. These Ts may be either too complicated to admit analytical derivation of the **sampling distributions**, or based on $\mathbf{x}$s from a probability model that is too difficult or impossible to specify. The essence of the computations for random samples of size n is the re-evaluation of the estimator on subsamples, which are typically produced by leaving out one observation at a time. For instance, the result of leaving out the ith datum may be denoted by $T_i = T[\mathbf{x}(i)]$, where $\mathbf{x}(i)$ is the particular subsample of size $n-1$ without the observation x_i.

The idea of appropriately differencing to reduce biases of order n^{-1} is credited to Quenouille. The original (1949) article [6] involves a **serial correlation** context, where the two subsets are the first and second half of the series. The second [7] has a more general context. Tukey [12] named the tool in 1958 for its parallel with the rough-and-ready boy-scout implement. He also coined the term *pseudovalues* for the individual differences, $nT - (n-1)T_i$. These are the simple ingredients for the standard error estimator. Tukey argued that in many instances these may be treated as approximately independent and their ordinary sample variance would be a reasonable estimator of var $[T(\mathbf{x})]$. This variance estimator was shown to be appropriate in large samples for the bias-corrected point estimator as well. Approximate **confidence intervals** and **hypothesis tests** are based on treating a standardized estimator as a normal or a **Student** t.

The methodology was extended to more general bias structures by Gray & Schucany [3]. Important early contributions to the theoretical foundations, by Rupert Miller, his students, and others were reviewed in 1974 [5]. There were early results on consistency of variance estimators for a broad classes of problems, including functions of **maximum likelihood** estimators (MLEs) and functions of U**-statistics**. A rigorous demonstration of the asymptotics for MLEs is given in [8]. The estimators that do not jackknife well have discontinuous *influence functions* (see [2, Section 11.6] or [9, Section 2.2.1]), of which the most notable example is the **median**. The approximate confidence intervals work better after symmetrizing **transformations**, for example, $\log s^2$ and $\tanh^{-1} r$.

The encyclopedia entry by Hinkley [4] contains the logical foundation, elementary notation, and some illustrative calculations. Efron & Tibshirani [2] give an excellent overview and the relationship of the jackknife to the **bootstrap**. These are distinct sample reuse approaches to getting information about the sampling distribution of $T(\mathbf{x})$. The bootstrap does this by simulating from the empirical distribution function, $F_n(x)$, the best estimate of $F(x)$ in a certain sense. The jackknife may be viewed as studying T in the neighborhood of F_n by quadrature (*see* **Numerical Integration**) rather than by **Monte Carlo**. Davison & Hinkley [1] present the pseudovalues as approximations of the empirical influence values for a nonparametric delta method. Distinct from this theoretical connection, they also illustrate the computation of jackknife-after-bootstrap diagnostics. For a more theoretical treatment of the jackknife and bootstrap, see [9].

Censored data are an important feature of some biostatistical problems. A recent examination of the suitability of jackknifing Kaplan–Meier integrals (*see* **Kaplan–Meier Estimator**) may be found in [11]. Stefanski & Cook [10] establish a relationship between the jackknife and SIMEX, which is a **simulation**-based method of inference for measurement error models.

References

[1] Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, New York.

[2] Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[3] Gray, H.L. & Schucany, W.R. (1972). *The Generalized Jackknife Statistic*. Marcel Dekker, New York.

[4] Hinkley, D.V. (1983). Jackknife methods, in *Encyclopedia of Statistical Sciences*, Vol. 4, N.L. Johnson, S. Kotz, & C.B. Read, eds. Wiley, New York, pp. 280–287.

[5] Miller, R.G. (1974). The jackknife – a review, *Biometrika* **61**, 1–17.

[6] Quenouille, M. (1949). Approximate tests of correlation in time series, *Journal of the Royal Statistical Society, Series B* **11**, 18–44.

[7] Quenouille, M. (1956). Notes on bias in estimation, *Biometrika* **43**, 353–360.

[8] Reeds, J.A. (1978). Jackknifing maximum likelihood estimates, *Annals of Statistics* **6**, 727–739.

[9] Shao, J. & Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer-Verlag, New York.

[10] Stefanski, L.A. & Cook, J.R. (1995). Simulation–extrapolation: the measurement error jackknife, *Journal of the American Statistical Association* **90**, 1247–1256.
[11] Stute, W. & Wang, J.-L. (1994). The jackknife estimate of a Kaplan–Meier integral, *Biometrika* **81**, 602–606.
[12] Tukey, J.W. (1958). Bias and confidence in not quite large samples (abstract), *Annals of Mathematical Statistics* **29**, 614.

(*See also* **Cross-validation**)

WILLIAM R. SCHUCANY

James–Stein Estimator

The discovery of Stein [6] that the sample mean of a normal population is inadmissible in three or more dimensions was based on an argument using the estimator

$$\mathbf{d}^1(\mathbf{x}) = \left(1 - \frac{b}{a + |\mathbf{x}|^2}\right)\mathbf{x},$$

where we observe $\mathbf{X} = \mathbf{x}$, with $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\theta}, \mathbf{I})$, a p-dimensional normal random variable (*see* **Multivariate Normal Distribution**). If $p \geq 3$, Stein showed that, for sufficiently small b and sufficiently large a,

$$\mathrm{E}_\theta|\mathbf{d}^1(\mathbf{X}) - \boldsymbol{\theta}|^2 < \mathrm{E}_\theta|\mathbf{X} - \boldsymbol{\theta}|^2, \quad \text{for all } \boldsymbol{\theta}, \qquad (1)$$

demonstrating the inadmissibility of $\mathbf{X}$ under squared error loss. This result only demonstrated the existence of a better estimator, as Stein did not give specific values of a and b that would satisfy (1). This was remedied in James & Stein [4], where it was shown that the estimator

$$\mathbf{d}^{\mathrm{JS}}(\mathbf{x}) = \left(1 - \frac{c}{|\mathbf{x}|^2}\right)\mathbf{x} \qquad (2)$$

dominates $\mathbf{X}$ as long as $0 \leq c \leq p - 2$. In fact, James & Stein [4] show that the optimal value of c is $c = p - 2$, and using this value (2) is usually referred to as the *James–Stein estimator*. Starting from (2), entire families of improved estimators of $\boldsymbol{\theta}$ have been derived. Note, in particular, that since $\mathbf{X}$ is a **minimax** estimator of $\boldsymbol{\theta}$, any estimator that dominates it is also a minimax estimator. Thus, research began into finding better families of minimax estimators of a multivariate normal mean.

One of the most important developments was due to Baranchik [1], who proved that estimators of the form

$$\mathbf{d}^{\mathrm{B}}(\mathbf{x}) = \left(1 - \frac{r(|\mathbf{x}|)}{|\mathbf{x}|^2}\right)\mathbf{x}$$

are minimax provided that (i) $0 \leq r(\cdot) \leq 2(p-2)$; and (ii) the function r is nondecreasing.

An immediate consequence of Baranchik's result was the minimaxity of (and the dominance of $\mathbf{X}$ by) the *positive-part Stein estimator*

$$\mathbf{d}^{+}(\mathbf{x}) = \left(1 - \frac{p-2}{|\mathbf{x}|^2}\right)^{+}\mathbf{x}, \qquad (3)$$

where $(\cdot)^+$ indicates that the quantity in parentheses is replaced by 0 whenever it is negative. This represents a great improvement over (2), as it does not suffer from aberrant behavior when $\mathbf{x}$ is near 0. (There, the James–Stein estimator can actually get infinitely large). In fact, the positive-part estimator (3) is so good that even though it is known to be inadmissible, it took over 25 years to exhibit an estimator that dominates it. (The inadmissibility of (3) follows from Brown [2], who showed that the admissible estimators must be generalized Bayes estimators. Because of the "point" at $|\mathbf{x}|^2 = p - 2$, (3) is not smooth enough to be generalized Bayes. The work of Efron & Morris [3, Section 5] showed that (3) was close to being a Bayes rule, and hence close to admissible, so it was suspected that it would be difficult to dominate. Finally, Shao & Strawderman [5] exhibited a dominating estimator.)

References

[1] Baranchik, A.J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution, *Annals of Mathematical Statistics* **41**, 642–645.

[2] Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems, *Annals of Mathematical Statistics* **42**, 855–903. (Corrigenda: *Annals of Statistics* **1**, 594–596.)

[3] Efron, B. & Morris, C.N. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach, *Journal of the American Statistical Association* **68**, 117–130.

[4] James, W. & Stein, C. (1961). Estimation with quadratic loss, in *Proceedings of the Fourth Berkeley Symposium on Mathematics Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 311–319.

[5] Shao, P.Y.-S. & Strawderman, W.E. (1994). Improving on the James–Stein positive-part estimator. *Technical Report*, Department of Statistics, Rutgers University.

[6] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 197–206.

(*See also* **Decision Theory**; **Shrinkage**; **Shrinkage Estimation**)

GEORGE CASELLA

Jeffreys, Harold

Born: April 22, 1891, in Fatfield, Co. Durham, UK.
Died: March 18, 1989, in Cambridge, UK.

The career of Harold Jeffreys is easily described. From his local school he went up to Cambridge, where he stayed for the rest of his life. His continuous 75 years as a fellow of St John's College is a record for any Oxbridge college. He was Plumian Professor of Astronomy and Experimental Philosophy, received numerous scientific awards, and was knighted.

During most of his life, and certainly up until his retirement from the Chair in 1958, he was best known for his important work in geophysics and related fields. The data he studied therein, and the general interest in the philosophy of science present in Cambridge in the 1920s, combined and culminated in the publication in 1939 of his book, called simply *Theory of Probability*. The substantially revised, third edition appeared in 1961 [1]. It is still in print and considered by many statisticians to be essential reading, not just for historical reasons, but because of its modern manner of thought. He was a poor oral communicator but his writing is superb. He stands with literature's greatest in the effective use of the English language.

There are two major novelties in the Theory, as he liked to call his book. The first lies in the concept of probability: the second in the development, from this concept, of operational procedures for handling data. He addressed the problem of how one's uncertainty about quantities of scientific interest, like hypotheses or values of constants, should be described. In the first chapter he demonstrated, on the basis of some simple ideas, that this could only be done through probability; so that one could speak of the probability of a hypothesis being true. Furthermore, statements of these uncertainties had to combine according to the rules of probability. One of these rules is **Bayes' theorem** and because of its ubiquity, the subject, when treated from this viewpoint, has become known as Bayesian statistics (*see* **Bayesian Methods**). The Theory was the first modern book on Bayesian statistics. This attitude towards probability was quite different from that of his near-contemporary, **R.A. Fisher**, who was, in the 1930s, revolutionizing statistics. Fisher used only the probability of data, given the hypothesis, whereas Jeffreys was advocating and justifying the concept of the probability of the hypothesis, given the data. Fisher's ideas found general acceptance and Jeffreys was initially treated as a maverick.

Although, at the time, their results seemed in good numerical agreement, it is now appreciated that they typically differ. If data x on hypothesis H has density $p(x|H)$, then Fisher used the tail-area probability $\int_x^\infty p(t|H)\,\mathrm{d}t$, or ***P*** **value**, to describe the status of H. Jeffreys used a direct probability $p(H|x) \propto p(x|H)p(H)$. In the use of the integral in the former but not in the latter, which satisfies the **likelihood** principle, the ideas contrast and the numerical values differ.

Jeffreys differed from **de Finetti** in regarding the numerical value of a probability as being shared by all rational persons, whereas de Finetti thought of it as subjective. If Jeffreys was right, then he had to have some way of producing the rational probability. The way he explored, and which later workers have followed, is first to describe a rational view of ignorance. This forms a reference point from which other states can be described, using Bayes' theorem. The invariance ideas he used have been extended into a modern development of reference priors.

Jeffreys' views have influenced the philosophy of science, and are in marked contrast to those of **Karl Popper**, who advocated the view that a hypothesis could only be disproved, whereas probability admitted values near one, effectively amounting to proof. Jeffreys was a great geophysicist who also created an original way of conducting the scientific method.

Reference

[1] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Clarendon Press, Oxford.

DENNIS V. LINDLEY

Job-exposure Matrices

Epidemiologic investigation of occupational hazards requires information on illness among workers and on their occupations or occupational exposures. Two families of epidemiologic investigations can be distinguished: industry-based studies and community-based studies. Each has unique advantages and disadvantages. Historically, community-based studies were based on analyses of job titles. With growing realization that there can be substantial variation in exposure profiles among workers who share the same job title, and that workers in different occupations can have common exposures, increasing attention has been paid to ascertaining subjects' occupational exposures (*see* **Occupational Epidemiology**; **Occupational Health and Medicine**; **Occupational Mortality**).

Since taking measurements in subjects' current workplaces is usually neither feasible nor useful for diseases of long latency, other approaches have been developed to ascertain subjects' past occupational exposures. If subjects can be interviewed, they can be asked about their exposure to various chemicals, but information thereby obtained is not sufficiently valid. Another approach is to obtain information about the jobs that subjects did and then have experts in industrial hygiene estimate the chemicals that may have been present in such workplaces. If the information collected about subjects' jobs is reasonably detailed, and the experts knowledgeable, then this can lead to quite valid exposure estimates. However, it is an expensive labor-intensive enterprise.

The job exposure matrix (JEM) approach was developed to provide a relatively inexpensive way of inferring exposures when the investigator has information on subjects' job histories. A JEM is simply a correspondence system for translating any occupation code into a list of exposures. The JEM provides the means for bringing together, for the purpose of statistical analysis, groups of subjects who share common exposures, irrespective of their occupations. A JEM consists of two primary axes, an exhaustive and mutually exclusive classification of occupations, and a list of substances. The occupation axis can be further subdivided by industries, by time periods, and conceivably by geographic areas. In the simplest form, the entry in the matrix could be a **binary** indicator of whether a worker in occupation i should be considered exposed to substance j. Applying each column in turn to a set of occupation histories allows the investigator to infer the exposure status of each study subject to each substance in the JEM. A more refined JEM could contain quantitative indicators of the probability of exposure to the substance in the job and estimates of the degree of exposure.

If the number of JEM substances is lengthy and the matrix entries are valid, this could generate useful data. While a handful of community-based JEMs have been developed in a few countries [1], they have not found wide applicability. The main limiting factor is the lack of valid and generalizable JEMs which are sufficiently broad in scope as to satisfy a wide range of research needs [3] (*see* **Validity and Generalizability in Epidemiologic Studies**). By contrast, a JEM can also be developed in the context of a **cohort study** and can be very useful if based on company records or expertise [2]. Such a JEM would not normally be applicable outside the cohort for which it was developed.

References

[1] Coughlin, S.S. & Chiazze, L. (1990). Job-exposure matrices in epidemiologic research and medical surveillance, *State of the Art Reviews in Occupational Medicine* **5**, 633–646.

[2] Goldberg, M., Kromhout, H., Guenel, P., Fletcher, A.C., Gerin, M., Glass, D.C., Heedrok, D., Kauppinen, T. & Ponti, A. (1993). Job-exposure matrices in industry, *International Journal of Epidemiology* **22**, Supplement 2, S10–S15.

[3] Siemiatycki, J. (1996). Exposure assessment in community-based studies of occupational cancer, *Occupational Hygiene* **3**, 41–58.

JACK SIEMIATYCKI

Joint Modeling of Longitudinal and Event Time Data

Many scientific investigations generate both *longitudinal measurement data*, with repeated measurements of a response variable at a number of time points (*see* **Longitudinal Data Analysis, Overview**), and **event history** *data*, in which times to transient or terminating events are recorded. Methods for the separate analyses of the two data components are well developed but procedures for their simultaneous treatment are still under study.

To fix ideas, consider **clinical trials** or observational studies in renal **transplantation**. After transplant, the performance of the graft can fluctuate over time. This can be measured through a proxy, perhaps serum creatinine, which will be high under abnormal function. The study design may call for creatinine to be measured regularly, maybe weekly during the first few months following transplantation, and this will generate a sequence $\mathbf{Y} = (Y(t_1), Y(t_2), Y(t_3), \ldots)$ of longitudinal data for each patient. Sometimes the data will be balanced, with measurement at the same time points on all patients, but often, especially in observational studies, the exact timing of measurements will be patient-specific, though nonetheless decided by the clinician or experimenter, perhaps informed by the patient's condition. In parallel to the collection of the longitudinal data, a **stochastic process** of times to or between certain events may also be observed. These may be transient events, such as intermittent reversible short-term rejection episodes (*see* **Repeated Events**), or single survival time outcomes such as death of the patient or failure of the graft (*see* **Survival Analysis, Overview**). This leads to event time data T, and joint modeling methods are appropriate when it is considered that the $\mathbf{Y}$ and T processes may not be independent. In principle, transient event data can be handled using similar techniques to single event data, but in practice, most development to date has assumed the latter. Note that T may be **censored** and may also be cause-specific: grafts can fail for a variety of reasons and the relationship between $\mathbf{Y}$ and T may well differ between causes. For instance, failures due to technical surgical reasons and failures due to rejection are likely to be associated with differing serum creatinine profiles.

Much work so far in joint longitudinal and event time modeling, henceforth just "joint modeling", has been based on specific case studies. Clearly, the statistical and scientific objectives of investigations of this kind will depend upon the application of interest. In particular, the primary focus for inference may be on:

(a) adjustment of inferences about longitudinal measurements to allow for possibly outcome-dependent dropout; (*see* **Nonignorable Dropout in Longitudinal Studies**);
(b) the distribution of time to a terminating or transient event conditional on intermediate longitudinal measurements;
(c) the possibility of using relatively quickly measured responses $\mathbf{Y}$ as a **surrogate** for survival time; or
(d) the joint evolution of the measurement and event-time processes.

The difference between (a) and (b) is reflected in the history of joint modeling, which derives principally from two originally distinct subject areas. Longitudinal researchers came to joint modeling through the 1990s as a result of the need to account for dropout from trials, which terminates observation at a random time T and is considered to be a nuisance. If the reason for dropout is related to the unobserved response, then severe bias can occur unless the analysis takes this association into account. In parallel with development in longitudinal data methodology, survival analysts arrived at effectively the same point at almost the same time through efforts to incorporate into **proportional hazards** analyses occasionally observed **covariates** subject to measurement error (*see* **Errors in the Measurement of Covariates**; **Measurement Error in Survival Analysis**). **Partial likelihood** methods for fitting **Cox regression models** can be used with **time-dependent covariates** over time, but require these to be observed at all failure times at which the patient remains at risk. In practice, this is rare for many clinical covariates, which are observed only occasionally, usually at clinic visits, and can vary between visits. In addition, many biomarkers or disease progression indicators can be measured only by proxy and/or with substantial measurement error. Denoting the covariate process by $\mathbf{Y}$, joint modeling approaches for $\mathbf{Y}$ and T were developed by survival specialists to account for the uncertainty in $\mathbf{Y}$. By the late 1990s, the two schools

merged and now use essentially the same methods. A slight difference is that longitudinal trials usually have balanced data, with a relatively small number of measurement points, and dropout is defined to be the time of first missed or last observed measurement, leading to discrete T. From the survival side, T is usually treated as continuous and the $\mathbf{Y}$ are more likely to have unbalanced measurement times.

Significant papers in the development of joint modeling methods include [1, 7, 8, 15, 17, 21, 22, 28]. To a lesser extent the methods have been motivated also by research in degradation, with $\mathbf{Y}$ a measure of wear before failure time T, and in surrogacy, with attempts to treat $\mathbf{Y}$ as a proxy for T [4, 30].

Modeling Strategies

Writing X for the fully observed covariates, interest is in the distribution $f(Y, T|X)$.

A *pattern mixture* factorization is often the simplest. Here we write

$$f(Y, T|X) = f(Y|T, X)f(T|X) \tag{1}$$

and in principle, fit a different Y-model for each cohort of patients defined by the values of T, using standard longitudinal approaches. Separately, a **marginal model** is fitted to T, using standard survival techniques. This method works well when there are only a few potential T values so that cohort sizes are fairly large. There can be problems, however, with the treatment of censored T, or when there are multiple cause events, such as dropout for a variety of reasons, not all equally associated with $\mathbf{Y}$. The approach is nonetheless easy to employ using standard software and attractive intuitively when it is believed that the population does indeed partition naturally by event times. Survival analysts, used to conditioning on the past through hazard modeling (*see* **Hazard Rate**), can be uncomfortable with the conditioning on the future implicit in $f(Y|T, X)$.

The second broad strategy reverses the conditioning through a *selection* factorization

$$f(Y, T|X) = f(T|Y, X)f(Y|X) \tag{2}$$

This approach makes censoring and multiple event causes easier to handle, but is again best employed when there are only a few possible event times T and values of $\mathbf{Y}$ are available where required in the $f(T|Y, X)$ model. An interpretive danger arises when $\mathbf{Y}$ is known to be a proxy for some unobserved true status of the patient, but $\mathbf{Y}$ is not in itself causal for event times. In that case, $f(T|Y, X)$ will describe an empirical relationship only and this needs to be recognized.

Random effects approaches usually postulate unobserved subject-specific effects U, which affect both $\mathbf{Y}$ and T. Assuming conditional independence between the observable components given the random effect, these models assume

$$f(T, Y|X) = \int f(T|X, U)f(Y|X, U)f(U)\,\mathrm{d}U \tag{3}$$

and estimation methods use standard **missing-data** approaches, almost invariably **EM** or **Markov chain Monte Carlo** (MCMC). Such models are conceptually attractive in that realistically, $\mathbf{Y}$ and T are rarely directly associated but intuitively depend jointly on some underlying true health status of the patient, captured by U. Disadvantages are that fitting is computationally intensive and the models rely on strong assumptions about the components $f(T|X, U)$, $f(Y|X, U)$, and $f(U)$, which are not directly testable from the observed data. An example of a random effects model is given later.

Variations of random effects models without conditional independence include **random-coefficient** pattern-mixture models

$$f(Y, T|X) = \int f(Y|T, X, U)f(U|X, T)f(T|X)\,\mathrm{d}U \tag{4}$$

and random-coefficient selection models

$$f(Y, T|X) = \int f(T|Y, X, U)f(Y|X, U)f(U)\,\mathrm{d}U \tag{5}$$

as described in [21]. Illustrations of the use of the above joint modeling strategies in a variety of application areas include [1–3, 8, 9, 13–16, 19, 27–29]. Other approaches to joint modeling include: parametric **multivariate analyses** with a **transformation** of T considered as a response, for example, a **multivariate normal** model for $(Y, \log T)$ [4, 6]); a **latent class** approach under which subjects are grouped into relatively homogeneous but unobserved subgroups with

differing longitudinal and event time properties [20]; and multistate modeling [12]. If interest is in estimation of treatment effects only rather than modeling per se, a variety of authors have developed methods intended to be **robust** to modeling assumptions (e.g. [10, 24]).

Ignoring Association

Given that methods for the separate analysis of **Y** and T are well developed and easily applied, the question arises as to what is gained through considerably more involved joint analysis. The answer to this is **efficiency** and, importantly, bias reduction (*see* **Unbiasedness**). Failure to account properly for the association between **Y** and T can lead to severe parameter bias in both sub-models.

A small **simulation** experiment illustrates a random effects model specification and the effect of ignoring association. Assume three **Y** measurements are scheduled, at times 0, 1, and 2 months. Responses are taken from a Gaussian linear model with a treatment indicator x (0 = placebo, 1 = active), a linear time trend, and subject-specific random intercept and slopes. Specifically,

$$\mathrm{E}[Y(t)|U_1, U_2] = \beta_x x + \beta_t t + U_1 + U_2 t$$
$$\mathrm{Var}(Y(t)|U_1, U_2) = \sigma_\varepsilon^2 \quad (6)$$

with $U_1 \sim \mathrm{N}(0, \sigma_1^2)$ and $U_2 \sim \mathrm{N}(0, \sigma_2^2)$ independently. Observation is terminated by an event at time T, taken from a hazard model

$$\lambda(t|U_1, U_2) = \lambda_0(t) \exp\{\alpha x + \gamma(U_1 + U_2 t)\} \quad (7)$$

with fixed-point censoring at three months. The data were simulated with constant baseline $\lambda_0(t)$, and although distributions are improper if $\gamma(U_1 + U_2 t)$ is negative, that is, the survival curves need not fall to zero, the censoring prevents this being a problem.

The parameter γ induces the association between the conditional distributions $f(Y|X, U)$ and $f(T|X, U)$. The following table illustrates the effect of γ on parameter estimates when the observed **Y** values are analyzed alone, with no allowance for association. The table gives mean values from 100 simulations each with sample size 250.

	$\beta_x = 1$	$\beta_t = 1$	$\sigma_1^2 = 1$	$\sigma_2^2 = 0.25$	$\sigma_\varepsilon^2 = 0.25$
$\gamma = 0$	1.00	1.00	1.00	0.26	0.25
$\gamma = 0.5$	0.99	0.93	1.02	0.25	0.25
$\gamma = 1$	0.97	0.84	1.01	0.23	0.26
$\gamma = 1.5$	0.98	0.78	0.96	0.23	0.26

Some parameters, including the mean treatment effect, are hardly affected in this scenario. Others, and in particular, the slope estimate β_t, are severely biased by ignoring the association between **Y** and T. This bias is not due to the *amount* of information lost, as for all values of γ, about 45% of patients had all three **Y** measurements. Rather, it is due to the trajectories of those missing: when $\gamma > 0$, subjects with steep slopes are more likely to have early T and hence the average and variance of observed slopes are attenuated.

Turning to event times, a Cox proportional hazards model was fitted in two ways. First, only treatment was included (OT), and second, the longitudinal **residuals** $R(t) = Y(t) - \hat{\beta}_x x - \hat{\beta}_t t$ were additionally included as time-dependent covariates, carrying forward the last value (LV) until another became available. In this second case, there is no true parameter for comparison and so results are given for the treatment effect α only.

	$\alpha = 1$	
	OT	*LV*
$\gamma = 0$	1.04	1.05
$\gamma = 0.5$	0.86	0.96
$\gamma = 1$	0.70	0.93
$\gamma = 1.5$	0.54	0.80

Ignoring the longitudinal data completely is equivalent to ignoring **frailty** effects in survival, which is well known to severely bias coefficient estimates toward zero. The simple model using longitudinal residuals as time-dependent covariates helps but does not remove the bias, which arises in part because of model **misspecification**, and in part because of the carrying forward of the last recorded value, which in general, is known to cause attenuation of regression effects and poor coverage of **confidence intervals** (e.g. [25]).

Sensitivity and Diagnostics

Fitting a correctly specified joint model overcomes bias but this gain is not made without cost. As well as

the need for more **computer-intensive** fitting methods, the joint model relies heavily on the assumptions made about the mechanism, which generates the association between $\mathbf{Y}$ and T. If the longitudinal data are divided into observed and missing components, Y_{obs} and Y_{miss}, there can sometimes be an infinite number of possible distributions for the missing data Y_{miss}, each of which is consistent with the observed data Y_{obs} and T. For example, certain illnesses involve sudden crises in condition, corresponding to rapid degradation (e.g. [18]). As a simple model, suppose that a crisis terminates observation and at the same time causes a step change in the response $\mathbf{Y}$ by an amount θ. This step change affects only Y_{miss} and therefore θ is not **identifiable** from the observed data Y_{obs} and T. Moreover, any false presumption as to the value of θ will lead to a biased estimate of the relationship between $\mathbf{Y}$ and the event-time hazard. In this example and more generally, inference about the missing data relies crucially upon the assumptions made, and these assumptions cannot be checked from the observed data alone. In applications, careful consideration of the modeling assumptions by statisticians and collaborators together is required.

If a fairly simple summary is of interest, such as the effect of treatment on the change in $\mathbf{Y}$ from start to end of the study, then **sensitivity analyses** are practicable and highly recommended. In essence, important parameters should be allowed to vary over a realistic grid of values and the effect on the summary measure investigated. In the simulation example above, a joint model could be fitted for a range of γ and changes in the treatment effects could be monitored. Stability of the estimate brings some credence to the results, although within the particular model class only. Sensitivity procedures are discussed by, for example, [5, 23, 26].

Although it is not possible to declare any particular model as correct, it *is* possible to declare models that do not fit the observed parts of the data as incorrect. This aspect is sometimes overlooked in applications but is important. Residual analysis on the longitudinal data should be carried out but with proper allowance for the event time as properties of residuals depend upon T. Conditional means, variances, and covariances of longitudinal residuals can be calculated, at least in the Gaussian case, and can be used to produce standardized residuals for assessment.

Comments

A great deal of research over the last decade has been directed toward the development of sophisticated methods for the joint analysis of longitudinal and event time data. These new methods are useful in their own right but also reveal two uncomfortable truths: ignoring the association between the two data components, or using overly simple methods to deal with it, can lead to seriously misleading conclusions; but sophisticated methods rely on assumptions that are difficult, or even impossible, to validate from the data alone. This makes it easy, if unhelpful, to conclude that the best way to deal with dropout in longitudinal studies, or measurement error in intermittently observed covariates or markers, is not to have these problems in the first place. A more constructive suggestion is that all reasonable steps should be taken during the design and data collection stages to minimize these problems.

If the primary interest is in the event times, then joint modeling procedures are helpful and conclusions are likely to be fairly robust to modeling assumptions, especially for balanced designs with a regular measurement schedule. There can be problems, however, if the decision to schedule a measurement at all is in itself informative, for instance, if measurements are taken during acute episodes of a chronic condition (e.g. [11, 18]). Methods that include the decision to take a measurement as a third process are being developed but are as yet unproven. Another issue from the survival side is that follow-up times can be long, so that simple slope-and-intercept longitudinal models are often inadequate and more sophisticated stochastic models are needed. The models need to be sufficiently flexible to describe a variety of possible trajectories of the covariate or marker, including rapid deterioration just before the event.

If the main interest is in the longitudinal data, then the question of what would happen if there was no dropout sometimes cannot be avoided. Here, sensitivity analysis is certainly required but brings new difficulty in how it should be undertaken. Sensitivity is easy to assess when there is a simple summary, such as a treatment effect on mean response, and a single parameter that links the longitudinal and event time processes. If there are multiple parameters driving the association, then sensitivity assessment is more awkward, and this problem is exacerbated further when sensitivity to model choice rather than

parameter values is to be considered. This is an area of ongoing research.

One of the cited purposes of joint modeling is to enable quick and easy prognostic information to be gleaned from longitudinal measurements, which can then be used as surrogates for an event-time outcome that may not be observed until a considerable time into the future. This suggestion needs more application-specific testing and experience before advantages of joint modeling can be conclusively argued.

Finally, almost all joint modeling techniques are computationally intensive and as yet there are no general and user-friendly routines available in statistical packages. Until these are written, there is likely to be little widespread use of the methods. This is perhaps no bad thing as it will allow specialists to gain more experience and advise influential software developers as to which models and methods can and cannot be recommended.

References

[1] Berzuini, C. & Larizza, C. (1996). A unified approach for modelling longitudinal and failure time data, with application in medical monitoring, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**, 109–123.

[2] Billingham, L.J. & Abrams, K.R. (2002). Simultaneous analysis of quality of life and survival data, *Statistical Methods in Medical Research* **11**, 25–48.

[3] Birmingham, J., Rotnitzky, A. & Fitzmaurice, G.M. (2002). Pattern-mixture and selection models for analyzing longitudinal data with monotone missing patterns, *Journal of the Royal Statistical Society, Series B* **65**, 275–298.

[4] Cox, D.R. (1999). Some remarks on failure times, surrogate markers, degradation, wear, and the quality of life, *Lifetime Data Analysis* **5**, 307–314.

[5] Daniels, M.J. & Hogan, J.W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout, *Biometrics* **56**, 1241–1248.

[6] De Gruttola, V. & Tu, X.M. (1994). Modelling progression of CD-4 lymphocyte count and its relationship to survival time, *Biometrics* **50**, 1003–1014.

[7] Diggle, P.J. & Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with Discussion), *Applied Statistics* **43**, 49–93.

[8] Faucett, C.L. & Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach, *Statistics in Medicine* **15**, 1663–1686.

[9] Faucett, C.L., Schenker, N. & Elashoff, R.M. (1998). Analysis of censored survival data with intermittently observed time-dependent binary covariates, *Journal of the American Statistical Association* **93**, 427–437.

[10] Glidden, D.V. & Wei, L.J. (1999). Rank estimation of treatment differences based on repeated measurements subject to dependent censoring, *Journal of the American Statistical Association* **94**, 888–895.

[11] Grüger, J., Kay, R. & Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models, *Biometrics* **47**, 595–608.

[12] Guihenneuc-Jouyaux, C., Richardson, S. & Longini, I.M. (2000). Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4 cell decline, *Biometrics* **56**, 733–741.

[13] Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of measurements and event time data, *Biostatistics* **1**, 465–480.

[14] Hogan, J.W. & Daniels, M.J. (2002). A hierarchical modelling approach to analysing longitudinal data with drop-out and non-compliance, with application to an equivalence trial in paediatric acquired immune deficiency syndrome, *Applied Statistics*, **51**, 1–22.

[15] Hogan, J.W. & Laird, N.M. (1997a). Increasing efficiency from censored survival data using random effects to model longitudinal covariates, *Statistical Methods in Medical Research* **7**, 28–48.

[16] Hogan, J.W. & Laird, N.M. (1997b). Mixture models for the joint distribution of repeated measures and event times, *Statistics in Medicine* **16**, 239–257.

[17] Hogan, J.W. & Laird, N.M. (1997c). Model-based approaches to analysing incomplete longitudinal and failure time data, *Statistics in Medicine* **16**, 259–272.

[18] Liestøl, K. & Andersen, P.K. (2002). Updating for covariates and choice of time origin in survival analysis: problems with vaguely defined disease states, *Statistics in Medicine* **21**, 3701–3714.

[19] Lin, H., McCulloch, C.E. & Mayne, S.T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables, *Statistics in Medicine* **21**, 2369–2382.

[20] Lin, H., Turnbull, B.W., McCulloch, C.E. & Slate, E.H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data, *Journal of the American Statistical Association* **97**, 53–65.

[21] Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**, 1112–1121.

[22] Pawitan, Y. & Self, S. (1993). Modelling disease marker processes in AIDS, *Journal of the American Statistical Association* **88**, 719–726.

[23] Rotnitzky, A., Scharfstein, D., Su, T.-L. & Robins, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring, *Biometrics* **57**, 103–113.

[24] Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association* **94**, 1096–1146.

[25] Tsiatis, A.A. & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error, *Biometrika* **88**, 447–458.

[26] Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. & Kenward, M.G. (2001). Sensitivity analysis for nonrandom dropout: a local influence approach, *Biometrics* **57**, 7–14.

[27] Wang, Y. & Taylor, J.M.G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome, *Journal of the American Statistical Association* **96**, 895–905.

[28] Wulfsohn, M.S. & Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error, *Biometrics* **53**, 330–339.

[29] Xu, J. & Zeger, S.L. (2001a). Joint analysis of longitudinal data comprising repeated measures and time to events, *Applied Statistics* **50**, 375–388.

[30] Xu, J. & Zeger, S.L. (2001b). The evaluation of multiple surrogate endpoints, *Biometrics* **57**, 81–87.

A fuller list of selected references is available at `www.maths.lancs.ac.uk/~henderr1/jntmod`

(*See also* **Model, Choice of**; **Nonlinear Mixed Effects Models for Longitudinal Data**)

ROBIN HENDERSON

Journal of Biopharmaceutical Statistics

The *Journal of Biopharmaceutical Statistics* (*JBS*) is an international, applied, biopharmaceutical statistical journal, published four times per year by Marcel Dekker, Inc., a division of Taylor & Francis, UK. It was established in 1988 by Karl E. Peace, Ph.D., the Georgia Cancer Coalition's distinguished cancer scholar at Georgia Southern. The past editors include Karl E. Peace (1990–1999) and A. Lawrence Gould (2000–2002). Current editor-in-chief is Shein-Chung Chow, Ph.D., Vice President of Biostatistics and Medical Writing of Millennium Pharmaceuticals, Inc., Cambridge, MA 02139. Beginning 2004, JBS also include book review section. The first issue of the *JBS* appeared in 1991.

The *JBS* provides an information resource for applied statisticians working in biopharmaceutical areas through publication of (i) high quality applications of statistics in biopharmaceutical research and development (*see* **Pharmacoepidemiology, Overview**) and (ii) expositions of statistical methodology with clear and immediate applicability to such work. Although not exhaustive, biopharmaceutical areas include particularly those attendant to the drug, device, or biologic research development processes; drug screening; assessment of pharmacological activity; pharmaceutical formulation and scale-up; preclinical safety assessment; **bioavailability**, **bioequivalence**, pharmacokinetics, pharmacodynamics, and genomics; **phase I, Phase II** and Phase III clinical development (*see* **Clinical Trials, Overview**); pre-market approval assessment of clinical safety (*see* **Drug Approval and Regulation**); **post-marketing surveillance**; manufacturing and quality control; technical operations; and regulatory issues (*see* **Drug Approval and Regulation**). Papers submitted to the *JBS* for publication consideration should emphasize the application of statistical methods, rather than the methods *per se* – whether new or established. Substantive aspects of the application should be presented. The process of problem formulation appropriate to the statistical method should be specifically addressed. Of particular importance is attention to statistical design (*see* **Experimental Design**) and protocol development (*see* **Clinical Trials Protocols**). In reflecting applied statistics as a scientific discipline, the *JBS* aims to provide models in the biopharmaceutical areas of proper design, analysis, and interpretation of both experimental and **observational studies**.

Digital submission of all manuscripts submitted for publication consideration in JBS is required. JBS currently accepts two options for digital submission: digital storage media, including $3-\frac{1}{2}$ diskettes, zip disks, or CD-ROMs, and e-mail attachments. Manuscripts may be submitted by e-mail to `journaledit@dekker.com`. The manuscript documents, along with all files containing references, figures, tables, outline, and abstract should be attached to the e-mail. Manuscript that is embedded in the body of the e-mail cannot be accepted. Manuscripts submitted to the *JBS* for publication consideration are reviewed by an editorial board member and two referees or appropriate experts. Review below regional editors is blinded to author identity. Authors are blinded to reviewer identity. General criteria for acceptance include originality, quality, and significance of the application or methods as well as the quality of the presentation.

Publications in the *JBS* reflect the international nature of biopharmaceutical research, with contributions by authors from every continent and many countries, e.g. Australia, Canada, China, England, Finland, Germany, Japan, South Africa, Switzerland, Taiwan, and the US. As of May 2004, the second issue of the fourteenth volume is in press, and all issues of the fourteenth volume are compiled.

KARL E. PEACE & SHEIN-CHUNG CHOW

Journal of Clinical Epidemiology

The *Journal of Clinical Epidemiology* represents a continuation, under a new name, of the *Journal of Chronic Diseases* (*JCD*), which was inaugurated in 1955. During that era, before the proliferation of specialty journals in such fields as **gastroenterology**, geriatrics (*see* **Gerontology and Geriatric Medicine**), and **rheumatology**, journals concerned specifically with medical research were oriented almost exclusively to studies of pathophysiology and biologic mechanisms. Clinical studies of chronic disease were seldom encouraged or accepted, because the care of chronic disease was seldom regarded as a scientific activity in the explicatory type of laboratory research usually conducted as "clinical investigation". The investigative methods needed to study patient care and to do **clinical trials**, however, were different from those of laboratory experiments; and reports using those methods would be either unappreciated by laboratory scientists, or regarded as too pragmatic for the often theoretical orientation of biostatistical and other journals concerned with methodology. Thoughtful reviews of clinical topics and appraisals of research methods would also usually take more space than most journals were willing to allocate. The *JCD* thus offered an orientation, appreciation, editorial policy, and space that had previously been unavailable.

The first volume of the journal immediately showed its new orientation, and its lively interest in methodology. The first paper in the first issue was **Merrell** & Shulman's "Determination of prognosis in chronic disease, illustrated by systemic lupus erythematosus". The paper described the medical use of **life-table** analyses for the course of clinical ailments, and was repeatedly cited thereafter for many years as the classic publication in that field. Another classic methodologic publication in the first volume was Louis Lasagna's discussion of an investigative method that was then in its infancy: "The controlled clinical trial: theory and practice". The discussion had a substantial influence on the planning and analyzing of cancer trials at the National Institutes of Health (NIH) during the late 1950s. A third methodologic classic in Volume 1 was **Harold Dorn's** essay, "Some applications of biometry in the collection and evaluation of medical data". The latter paper contains Dorn's often quoted remark that

> Reproducibility does not establish validity, since the same mistake can be made repeatedly; but without reproducibility an observed relationship becomes merely an isolated historical event and adds nothing to accumulated scientific knowledge.

During the next few early years, the journal's continuing focus on clinical issues in chronic disease was reflected by publications concerned with topics that today might be classified as **neurology**, metabolism, rheumatology, gastroenterology, **hepatology, psychiatry**, neonatology, congenital anomalies (*see* **Teratology**), atherosclerosis (*see* **Cardiology and Cardiovascular Disease**), hematology, **oncology**, and such chronic infections as tuberculosis and syphilis. Victor McKusick's pioneering work in clinical genetics first appeared in the *JCD* in a series of instalments under the general title of "Heritable disorders of connective tissue".

In addition to these clinical topics, the early volumes of the journal continued to offer a forum for methodology. The papers referred to classification of arthritis, evaluation of **screening** tests, discussions of epidemiologic principles, uses of interview data to assess **prevalence** of disease, the measurement of pain and pain relief, uses of nonmedical interviewers to obtain data about specific symptoms, and variability of daily blood pressure measurements. The methodologic studies, then as now, revealed the frequent, but often unrecognized, problems of **bias** in research with human groups. **Donald Mainland**, after analyzing results of a questionnaire given to a class of 129 first-year medical students, concluded that "... more than one-half of them held opinions which, if allowed to influence the selection of subjects in a forward-going etiologic survey, would bias the results".

Sidney Cobb et al., reporting on "differences between respondents and nonrespondents in a morbidity survey involving clinical examination" demonstrated the type of bias that might arise from low response rates in studies for the estimation of prevalence. The authors recommended a procedure that (like many other recommendations about how to deal with bias) has often been subsequently neglected: "A study of the nonresponse problem (should) be built into each new field investigation as it is planned."

The journal also become involved in topics that were overtly controversial or that would later generate controversy. In a controlled trial reported elsewhere in 1952, anticoagulant therapy had been found unequivocally effective in reducing short-term deaths in patients with acute myocardial infarction. Many clinicians claimed, however, that the results of the trial were inconsistent with their own clinical experience, particularly for the predominance of "low risk" patients who had excellent prognoses without treatment. Only much later would it be realized that the anticoagulant trial, despite an untreated control group, was not randomized (*see* **Randomization**) or double-blind, and that the proanticoagulant results could easily be attributed to a **biased** assignment of patients. Nevertheless, anticoagulant therapy was being so vigorously advocated that physicians who failed to use it might be sued for malpractice if a patient with myocardial infarction died. In a pair of editorials in the journal in 1956, the virtues of anticoagulant therapy received a spirited denunciation by David Rytand and a vigorous defense by William Foley.

The *JCD*, in its early years, published several instalments of research, conducted by the US Public Health Service, as the Tuskegee study of "Untreated syphilis in the male Negro". The research was originally regarded as a splendid investigation of the **natural history** of a disease, but the work later became controversial and received many ethical rebukes for continuing to assess "natural history" at a time when presumably effective therapy (with penicillin) had become available. During subsequent debates about the moral and methodological issues, T.G. Benedek (in the 1978 *JCD*) pointed out that the rebukers had often overlooked the ethical context of the era in which the research was done.

The modern fervor (and dispute) about lowering cholesterol was just beginning. Several papers on how to reduce cholesterol with diet or medication had appeared in the *JCD* (and elsewhere), but an international group of experts, after a meeting in Geneva, stated in 1957 that

> There was no clear cut scientific evidence to show that any particular factor causes... coronary artery disease. Numerous public statements by scientific and other writers... give the impression that the atherosclerosis problem has been largely solved. The chief culprit is purported to be fat and diet. According to these opinions, all one has to do to mend the situation is to change one's eating habits so as to include a special kind of low-fat diet. Unfortunately, scientific proof of a causal relationship between fats in the diet (and) coronary artery disease is still lacking.

Forty years later, many experts would claim that the proof has now become convincing; but others would still argue that it is not.

In 1957, after the death of J. Earle Moore (the founding editor), the co-editors became Louis Lasagna and David Seegal. In 1966, David P. Earle became editor with Martin Branfonbrener added as co-editor a year later. In 1978, Earle again became sole editor, with Brandfonbrener and Walter O. Spitzer working as associate editors. When Earle retired, the journal resumed the geographically separated, dual-editor pattern that had originally been set by J.E. Moore in Baltimore and D. Seegal in New York. The co-editors after 1982 were Alvan R. Feinstein in New Haven and Walter O. Spitzer in Montreal.

Throughout its 33 years, the journal has published some outstanding, memorable papers. Some of them have already been mentioned, but several others can be noted as "golden oldies". They are: Seegal's 1962 editorial on the virtues of saying "I don't know"; E. Schimmel's 1963 editorial on "The physician as pathogen"; another 1963 editorial in which the author concluded that without better clinical "science at the bedside, modern medical research may yield an intricately designed, expensively produced, doubly-blind controlled, statistically significant chaos"; a 1965 reprinting of J. Evelyn's treatise, originally published in 1961, on the hazards of air pollution in London; and a randomized double-blind trial, by C.R.B. Joyce and R.M.C. Welldon, in 1965, on "The objective efficacy of prayer".

With further passage of time, the contents of the journal gradually changed as other journals became available in subspecialty medical domains, and in new specialties such as rehabilitation and geriatrics. Many of the clinical reviews and symposia that formerly might have appeared in the journal became submitted and published elsewhere. The journal's symposia increasingly began to reflect its additional methodologic orientation, with topics that included the role of computers in medicine, quantitative principles in the design and analysis

of long-term studies (*see* **Longitudinal Data Analysis, Overview**), and the development of indexes (or rating scales) to measure **health status** or to describe functional status and **quality of life**. The statistical philosophies that guided the US National Institutes of Health (NIH) in its approach to clinical-trial research were first described by several statisticians in a 1966 *JCD* "biometrics seminar" called "The role of hypothesis testing in clinical trials". The discussion included the often-quoted remark by **Jerome Cornfield** that, "I do not believe that anything that is good science can be bad statistics." The many unresolved controversies about retrospective **case–control studies** were discussed in a frequently cited 1979 symposium edited by Michel Ibrahim and W.O. Spitzer. Many other methodologic issues in epidemiology were considered in a memorial "festschrift", in 1986, for Abraham Lilienfeld.

The journal's clinical scope was still broad and the clinical topics still emphasized diagnosis, prognosis, course, and therapy, but most of the clinical publications began to include the kinds of group data and statistical analyses that today would make the work be classified as **clinical epidemiology**. The latter domain expanded the scope of "epidemiology" to include many topics in which the people under study were in clinical, rather than community, settings. The topics under study were also expanded to include behavioral, social, and familial factors – personality traits, emotional adjustment, urbanization, **social class**, social isolation, and familial structures and relationships – that could affect the development or management of human ailments such as heart disease, cancers, renal disease, schizophrenia, and disability.

Although all of these topics could today be included in the broad scope of contents for "clinical epidemiology", the journal's most striking expansion was in methodology. The *JCD* became the prime publication for creative scholarship in the analysis and development of methods for research in quantitative clinical epidemiology. The orientation required a special blend of thought: a sophisticated knowledge of clinical distinctions in human ailments; an intense awareness of epidemiologic subtleties in the way that groups of people are formed and collected; and mathematical attention to the statistical strategies with which results can be quantitatively summarized and interpreted.

Scholars who work in the multidisciplinary intersection produced by the manifold methodologic challenges of quantitative clinical epidemiology are also relatively homeless. Their interests are often too quantitative or epidemiologic to be appreciated by clinical journals, too clinical to be approved by journals of epidemiology or public health, and too clinically or epidemiologically "applied" to be welcomed by journals of mathematical or biologic statistics. By opening this multidisciplinary forum, the *JCD* became a leading outlet for methodologic advances in medical research concerned with groups of people.

The methodologic analyses and advances were sometimes presented within the text of publications on specific ailment-oriented topics, but often the methods themselves were the main focus of discussion. The methods included such clinical issues as the role of **co-morbidity** (a term and concept introduced in the *JCD* in 1970); the acquisition of cogent, high-quality data in interviews and questionnaires (*see* **Questionnaire Design**); the evaluation of diagnostic and technologic tests (*see* **Diagnostic Tests, Evaluation of**); decisions about what kind of data to assemble in describing personality, behavior, **quality of care**, or quality of life; defining the "range of normal" (*see* **Normal Clinical Values, Reference Intervals for**); and the role of necropsy research in revealing the fallacies of using "cause-specific" death certificate data for individual decisions or collective concepts about the distribution of disease.

The epidemiologic methods referred to problems that produce biased or inaccurate results in data for groups. The problems included diverse issues in life-table analysis, the first empirical demonstration of the distortion known as **Berkson's fallacy**, and attention to every aspect of the assembly, maintenance, and collection of data for the people investigated in randomized trials, **cohort studies**, case–control studies, community **cross-sections, ecologic studies**, and other architectural structures for research. The statistical concepts included problems in planning sample sizes for diverse investigative situations (*see* **Sample Size Determination**), estimating relative risks, assessing the role of repeated measures and **regression to the mean**, determining **prevalence** in longitudinal or cross-sectional studies, and understanding or evaluating the virtues and hazards of old multivariable analytic methods (such as **linear regression**) and newer **multivariate analysis** (such

as binary regression, logistic **transformations**, **discriminant functions**, and the **proportional hazards** model).

To acknowledge its focus on the intimate interchange between qualitative challenges in clinical science and quantitative issues in statistics, the *Journal of Chronic Diseases* in 1988 changed its name to the *Journal of Clinical Epidemiology*. The fertile interchange has led to many useful collaborations among clinicians, epidemiologists, psychosocial scientists, and statisticians, while producing valuable interdisciplinary cross-fertilization and communication. It has made medical people aware of the need for satisfactory "numeracy" and "reliability" in communicating with their statistical and psychometric colleagues, while making statisticians and psychometricians aware of the need for satisfactory "literacy" and "sensibility" in communicating with medical people.

The satisfactory adjustment of interdisciplinary communication is not easy. Sometimes a statistical author may submit a paper that is aimed exclusively at statisticians, and that is incomprehensible to a medical reader. By 1979, the problem was happening often enough in the *JCD* to make David Earle publish an editorial urging "potential authors to write their manuscripts so that the clinical relevance is clearly apparent, and put as much of the derivation as possible in an appendix or to publish the mathematics elsewhere". The editors of the *JCE* have often made analogous requests.

Sometimes a reviewer who is a rigorous quantitative methodologist may make demands that are impossible for an investigator to attain. A psychometric reviewer may ask, almost as a matter of routine, that **Cronbach's alpha** be calculated for a five-category ordinal scale for which the calculation (which requires an inventory of individual items) cannot be done. A biostatistical reviewer may ask for analyses of data that cannot be obtained because the research project is completed, or may want the authors to use complex multivariable procedures (beyond those already employed) that may bring more smoke and heat to the results but little light.

From the medical-content side of the spectrum, an epidemiologist may insist that the research is worthless because it was a cohort study rather than a randomized trial, or vice versa; or the reviewer may dismiss the research as useless and beyond repair because the investigators should have chosen an entirely different control group. A clinical reviewer, unwilling to learn some basic principles of numeracy, may complain about the "obscurity" of statistical writing that contains nothing more esoteric than simple regression and **correlation** coefficients. These problems have not been common, however. In general, the journal seems to have had outstanding success in attracting suitable authors, getting capable, open-minded reviewers, and producing relatively clear interdisciplinary communications. The term "clinical epidemiology" covers many methods and orientations. They include causal elucidations by a classical epidemiologist, patient care decisions by a classical clinician, analytic improvements by a classical statistician, and diverse mixtures of some or all of these activities. (At some levels of definition – such as "disease", "chronic disease", "clinical", or "epidemiology" – an encompassing vagueness may be more satisfactory than an excluding precision.)

Under the new title, the *JCE* continued its basic policies, but added some new sections. Controversial topics were regularly published as trios containing a presentation, dissent, and response. Some of the more prominent controversies discussed in this variance-and-dissent format have been classification and directionality in epidemiologic research, the use of the **kappa** statistic, and the role of Popperian causality (*see* **Causation**) in scientific reasoning. This format, which allows direct airing for controversial topics, has been popular with readers, and has subsequently been emulated at other journals. A "Second Thoughts" section was made available for "lighter" essays that would be "fun" to read. James F. Jekel was invited to prepare periodic summaries, called "Rainbow Reviews", of the reports (with multicolored covers) regularly issued by the US **National Center for Health Statistics**. With the continued application of clinical epidemiology to studies of pharmaceutical agents, a new section on **pharmacoepidemiology** was added in 1991.

In 1996, the *JCE* began publishing supplements containing the abstracts submitted to the annual meeting of the International Clinical Epidemiology Network (INCLEN). The *JCE* has also published supplements on the SUPPORT study of outcomes and risks of treatment, policy for management of asymptomatic

hypercholesterolemia, pharmacoepidemiology in developing countries, ethics in epidemiology, long-term health effects of silicone breast implants, and postvasectomy sequelae.

Since 1995, the journal has been published by Elsevier. As at 2004, the Editors are A. Knottnerus (Netherlands School of Primary Care Research) and P. Tugwell (University of Ottawa).

ALVAN R. FEINSTEIN

Journal of The American Statistical Association

The **American Statistical Association** (ASA) was founded in 1839 by men concerned about issues surrounding the nation's decennial **census** taking. Almost 50 years later, General Walker (ASA President 1883–1896), a towering figure who was impassioned in his desire and beliefs that all workers (most especially government workers and researchers throughout the land) should embrace the statistical method in their daily work, led the association beyond the then-local horizons of Boston with the adoption of a number of measures, the most fundamental one being the establishment in 1888 of a *New Series* of publications of the ASA. This title reflected the fact that previously there had been a *Collection* of the ASA (with the first and possibly only volume in 1847) plus other occasional papers, many of which had been destroyed in the Great Fire of Boston in 1872. An account of these earlier collections can be found in [4]. This *New Series* subsequently assumed the title *Quarterly Publications of the American Statistical Association* and in 1922 was renamed *Journal of the American Statistical Association (JASA)*. The header *New Series* however continued for 44 years, eventually being removed with the 1932 volume. Today, *JASA* still appears quarterly.

For its first 40 years with only two exceptions, ASA held four quarterly (or three quarterly and one annual) meetings per year; after 1894 the quarterly meetings were dropped but the annual meetings continued. Papers read at these meetings constituted a large proportion of the articles in *JASA* in its first 20 years or so. In 1928, read papers were assembled together as a Proceedings section of *JASA*, a practice continued until 1937 after which there was a return to the earlier custom whereby such read papers intermingled with general papers. This was finally discontinued with the publication of separate Proceedings of the Business and Economic Section in 1954 and of the Social Statistics Section in 1958.

Articles reflected the interests of ASA members who were primarily economists, accountants, social scientists, political scientists, historians, health professionals, and the like. That is, members were users of statistical science in their substantive field, and so articles were focused on advancing knowledge and new theories in those fields rather than in statistical methodology *per se*. Indeed, the first article of Volume 1, on water power [5] and the second one on parks and open spaces [3], illustrate amply the concerns of members with societal issues (in these cases people's basic well-being). In addition, there were numerous Reports, Miscellany, News and Notices, and Reviews entries. These articles typically included reports on **vital statistics** (of every imaginable stripe – deaths, births, divorces, diseases, suicides, etc. recorded for national, state and local municipalities, as well as international regions); they covered reviews of important papers from abroad (most often mathematical developments from British publications); and they included book reviews, among other topics. Starting in 1897, regular reports of the ASA Secretary as read at the annual meetings were included; see [1]. The Proceedings and Scientific Program of these meetings began to be published in 1910 [2]. Reports of various ASA committees would at times also be published. In short, *JASA* was the vehicle to convey information – both scientific results and operational news – to the membership.

Today, the nature and content of *JASA* have changed, at least on the surface if not in its aims. In a formal sense, the general articles now appear within the Applications and Case Studies Section or the Theory and Methods Sections. The split into separate divisions was implemented in 1968 and visible in 1970 when papers in each section were assembled together. The Applications Section was expanded to the Applications and Case Studies Section in 1989. Book reviews had been collected together from the beginning. In 1989, this section was expanded to the (General) Review Section, though book reviews still constitute the bulk of the material in this section. The extensive Reports and much of the Miscellany articles have disappeared from *JASA*; the News and Notices articles also no longer appear in *JASA*. These were essentially moved to *The American Statistician* when it began in 1947 and later to the monthly *Amstat News* begun in 1974, with the exceptions that the Reports of the Annual Meetings continued to appear in *JASA* until the 1971 meeting and the Board of Directors and related Reports until the 1969 Report, when these reports shifted to *The American Statistician*. *The American Statistician* from its inception also included articles demonstrating the uses of statistics, articles that previously occupied many pages of *JASA*. As its founding editors said, *The American*

Statistician would serve as an adjunct to the technical papers published in *JASA*.

From the content viewpoint, a reading of Information for Authors, which currently appears in each issue of *JASA*, reveals that the Applications and Case Studies Section seeks articles that contribute to a substantive field through the use of sound or innovative uses of statistical methods, present data useful to such fields, or discuss and evaluate such data and findings; methodological innovations are not requirements. This descriptor reflects very accurately and keeps intact the tenor and the goals of ASA as exemplified in the articles of the very early issues of *JASA*. Then and now the important new theoretical results were directed at the field of application with new statistical theory that may have emerged being incidental to the major thrust.

By the time of ASA's Seventy-fifth Anniversary in 1914, changes were looming on the horizon. The so-called mathematical method had appeared through **correlation**, and has remained as an integral part of statistical science. Nevertheless, mathematical statistical articles still assumed a relatively small proportion of *JASA* articles until about the 1950s. By the 1960s, mathematically based articles had become more dominant. These articles now appear in the Theory and Methods Section. As defined in the Information for Authors, this Section "publishes articles that make original contributions to the foundations, theoretical development, and methodology of statistics and probability". This mandate is "interpreted broadly. . . and may include computational and graphical methods as well as more traditional mathematical methods". However, such articles should be, and are, motivated by a practical problem arising from a substantive application.

The General Section includes the traditional Book Review Section plus Review Papers covering an area of applied statistics or a review of some specific statistical theory. Special topic papers may also appear in this general section.

In addition, at its April 1907 meeting, it was decided to ask the ASA President to address the association at its annual meeting and that this Address be published as the lead article in the following March issues of *JASA*. Accordingly, the first Presidential Address was delivered by Wright on January 17, 1908 (see [6]). Interestingly, though the ASA had but five presidents for its first 70 years, each serving till death (Fletcher 1839–45, Shattuck 1846–51, Jarvis 1852–82, Walker 1883–96, and Wright 1897–1909), subsequent presidents were limited to one year terms.

Today, in substance *JASA* is dominated by theoretical mathematically based statistical methodology, although its authors continue to be motivated by real world problems. The advances in the substantive field that totally dominated *JASA* prior to about 1950–60 are now found in other ASA journals, namely *Journal of Business and Economic Statistics* (begun in 1983), *Journal of Educational and Behavorial Statistics* (1976), *Journal of Computational and Graphical Statistics*, (1992), *Journal of Agricultural Biological and Environmental Statistics* (1996 and a joint venture with the International Biometric Society), and *Technometrics* (1959 and a joint venture with the American Society of Quality). In addition, the magazines *STATS* (1989) and *Chance* (1988) are targeted to the student and/or man-in-the-street nonstatistician audience. Earlier, ***Biometrics*** (called *Biometrics Bulletin* 1945–46) was launched under **Gertrude Cox's** editorship by the Biometrics Section of ASA to publish articles on the use of mathematical and statistical methodology in biology (including agriculture); this journal was fully assumed by the **International Biometric Society** in 1951. This applications orientation of the membership is still vibrant today and has been a strong and persistent thread throughout. The strength of this view was reflected by the unsuccessful efforts of the ASA's mathematical members to have the *Annals of Mathematical Statistics* (begun in 1930, and first edited by Harry Carver) as an ASA publication.

The first *JASA* Editor was Davis Dewey, who served from 1888 to 1907. The division in 1969 brought with it a separate editor for each section: the Applications (predecessor of the Applications and Case Studies) Section, Theory and Methods Section, and Book Review (now the General Review) Section.

With the exception of the occasional invited paper and also Invited Discussants to selected articles typically addressing a major applied topic, articles are unsolicited. Potential authors submit manuscripts to the Editor(s). Papers deemed not to fit the overall aims of the journal would be returned to the authors without going through the formal reviewing process. That said, in a typical scenerio, the Editor disseminates the submissions to a cadre of Associate Editors, who take the responsibility for selecting and monitoring the refereeing process. Double-blind refereeing

for the Applications and Case Studies and the Theory and Methods Sections was instituted in 1996.

The 2001 Editors reported that the Applications and Case Studies Section received 123 new manuscripts (125 in 2000) and that the acceptance rate (of new and resubmissions) continues to be just under 25%. The Theory and Methods Section received 374 new submissions in 2000 (370 in 2001) with an acceptance rate of 20% (19% in 2001). In the General Review Section, in 2000, 62% (57% in 2001) of new books received were sent out for reviews; three of the six review manuscripts received in 2000 were rejected and three were accepted in 2001, while four of the nine received in 2001 were rejected and three were accepted in 2001 (and one in 2002). Subsequently, there were 48 Applications and Case Studies papers, 66 Theory and Methods papers, 3 Review papers, 57 Book Reviews, and 27 Telegraphic Review articles published in 2001, plus the Presidential Address and relevant editorially related information, occupying 1543 pages. In contrast, Volume 1 published 11 Leading Articles and 49 Reviews and Miscellany articles in 492 pages. Volume 1 spanned two years; it was not until Volume 19 in 1924 that one volume per year began.

Throughout its long history, JASA has continued to be a premier journal serving the entire international statistical community. It remains the flagship publication of the ASA.

References

[1] Dewey, D.R. (1897). Report of the Secretary of the American Statistical Association, *Journal of the American Statistical Association* **5**, 234–235.

[2] Doten, C.W. (1910). Proceedings of the Seventy-first Annual Meeting of the American Statistical Association, New York, December 27–30, 1909, *Journal of the American Statistical Association* **12**, 36–39.

[3] Gould, E.R.L. (1888). Park areas and open spaces in American and European Cities, *Journal of the American Statistical Association* **1**, 49–61.

[4] Green, S.A. (1889). An account of the collections of the American Statistical Association, *Journal of the American Statistical Association* **1**, 328–330.

[5] Swain, G.F. (1888). Statistics of water power employed in manufacturing in the United States, *Journal of the American Statistical Association* **1**, 1–44.

[6] Wright, C.D. (1908). Address of Carroll D. Wright, President of the American Statistical Association, at its Annual Meeting in Boston, January 17, 1908, *Journal of the American Statistical Association* **11**, 1–16.

LYNNE BILLARD

Journal of The Royal Statistical Society

In May 1838, following the establishment of the Society in 1834, the Council launched the first volume of the *Journal of the Statistical Society of London*:

> The Council of the Statistical Society of London is of opinion that the time has arrived when the Fellows of the Society, and the public, will hail with satisfaction the appearance of a Journal devoted to the collection and comparison of Facts which illustrate the condition of mankind, and tend to develop the principles by which the progress of society is determined.

Since the "Science of Statistics" was in its infancy in the 1830s, the Council felt it necessary to add some explanation about the objects of the Society and its journal in their introduction to the new journal. Within the extensive scope of the subject, the importance of medical statistics was already recognized:

> Mechanics discover the means of abridging human labour; Chemistry enters largely into the economy of Arts; Medicine practises on the bodies of men; all these sciences operate upon human interests and their powers and effects are susceptible of statistical exposition.

Reviewing the content of the journal in 1865, the Council divided the subject matter into seven classes: "commercial", "industrial", "financial", "moral and social", "vital", and "miscellaneous", the last category comprising papers not presented for reading to the Society. W.A. **Guy** was an early presenter of papers, such as "Influence of the seasons and weather on sickness and mortality" and "Influence of employment and health". William **Farr**, of still greater authority, made his debut with the paper "Mortality of lunatics" and "The influence of elevation on the fatality of cholera". Other contributions included the reports of the Committee on Medical Statistics (1837) and of the Committee on Sickness among the Metropolitan Police Force (1839–1840). Although, as long ago as 1863, the Society had been urged to publish discussions at its meetings, the proposal had been defeated. It was not until the June issue in 1873 that reports of the oral discussion and the authors' replies began to appear, and the practice of having both a formal proposing and seconding of a vote of thanks, as now, did not begin until 1909.

In January 1887, the Society was granted its Royal Charter, and the journal accordingly changed its name to the ***Journal of the Royal Statistical Society***. Another important change at that time was the introduction of reviews of books on statistical and economic subjects in 1886, under the heading "Notes on some recent Additions to the Library".

Up to that point in the journal's history, the papers published had been almost entirely of a descriptive nature, with large numbers of tables presenting data, but without detailed analysis. It was only around the turn of the century that the mathematical foundations began to be developed in the pages of the journal, with papers such as "On the theory of correlation" by G. Udny **Yule** (1897) and "On the representation of statistics by mathematical formulae" by F.Y. **Edgeworth** in four parts, and concluded in the 1899 volume.

In June 1928, the Society formed its first "Study Group" for holding informal meetings. This was followed in 1933 by the formation of the first of its Sections, the Industrial and Agricultural Research Section. The papers presented at the first two meetings of the Section were published in a supplementary issue with part II of the main journal in the Society's centenary year. A second supplement was issued with part IV of the journal. The *Supplements*, two parts per volume, were initially designed to cater for the "considerable developments in the application of modern statistical methods to technical problems met with in industry and agriculture" during the previous two decades.

World War II caused the research activities of the Industrial and Agricultural Research Section to be abandoned. Only about four meetings could be held in each session during the early years. The 1940–1941 volume of the *Supplement* was slim and the next volume did not appear until 1946. However, papers continued to be accepted and published as "read" papers even though they could not be presented. One effect was therefore that written, rather than oral, contributions to the discussion appeared, and this practice persisted and grew after the war.

Some work, however, was carried out around the country under the Industrial Applications Group formed for the purpose. This prompted the Society to split the Section into two: the Industrial Applications Section and the Research Section.

The Industrial Applications Section was constituted of several Local Groups around the country whose purpose was primarily to organize meetings, whereas the *Supplement* was intended to be primarily the vehicle for publication of the proceedings of Research Section meetings and to fulfill the need "for a medium of publication for research work (not necessarily theoretical or mathematical), which is of general interest to statisticians". An editorial panel was set up for the journal under the editorship of M.G. **Kendall**, B.L. Welch, and F. **Yates**. Research Section meetings did not have the status of the Ordinary Meetings of the Society until 1958 when the meetings were called Research Methods Meetings (later changed to the current "Ordinary Meetings organized by the Research Section" in 1969).

By 1947, the *Supplement* had grown into a scientific journal of high repute. The Council therefore decided that, from 1948, the Society's two publications would both be issued under the main title of the ***Journal of the Royal Statistical Society***, the original journal being distinguished by the subtitle "Series A (General)" and the *Supplement* by the subtitle "Series B (Methodological)". Though the names had changed, the volume numbers ran on sequentially. Series A remained the organ of the Society as a whole, with publication of papers from Ordinary Meetings (other than the research type), the annual report of Council, book reviews, obituaries, and other features from time to time.

Since the war, the Council had been aware of the absence of a publication devoted to the practical statistical problems that arise in the many fields of human activity. To fill this gap, *Applied Statistics* was launched in 1952. It officially became the ***Journal of the Royal Statistical Society***, Series C, in 1964. The President, Austin Bradford **Hill**, defined its aims in the first issue as follows.

> *Applied Statistics* has been founded, therefore, to meet the needs of all workers concerned with statistics – not of professional statisticians only but also of those innumerable workers in industry, commerce, science, and other branches of daily work, who must handle and understand statistics as part of their tasks. Its aim, in short, is to present, in one way or another but always simply and clearly, the statistical approach and its value, and to illustrate in original articles modern statistical methods in their everyday applications.

The journal published three issues per volume and initially was designed more as a magazine than as a learned journal. It contained reports of the meetings of the various Groups, articles expounding statistical methods and illustrating their application, and features entitled "Questions and answers", "Notes and comments", "Letters to the Editors", and book reviews. Gradually, though, the journal became more technical in character.

In December 1966, J.A. Nelder and B.E. Cooper organized a meeting on "Statistical programming – the present situation and future prospects" at the Atlas Laboratory, Chilton, UK. The five papers presented and an account of the discussion were published in *Applied Statistics* in 1967. As a result of the meeting, the "Statistical algorithms" section of the journal was started in 1968. The main aims of the section were

> to ensure that published algorithms are clearly organized, well documented, and standardized in notation and terminology, so that they will be readily understandable to a large number of readers; ... also, that the algorithms are programmed as far as possible in languages which are widely available and clearly defined independently of individual implementations.

These aims only began to outlive their usefulness by the mid-1990s when the publication of statistical **algorithms** ceased. By this time, over 300 had been published.

In 1968, Series A made a break with tradition when it ceased to publish the Sauerbeck index of wholesale prices, which it had published annually since 1886, latterly from material supplied by the editor of the *Statist*. Other changes arising from concern over overlap with Series C refocused its editorial policy and included a change in its subtitle to "Statistics in society" and a decrease from four issues per volume to three in 1988. In contrast, Series B and Series C increased from three issues to four in 1993.

The year 1993 also saw the addition of *The Statistician* to Series A, B, and C following the merger with the Institute of Statisticians. This resulted in an integration and reorganization of the content of the four journals, with the intention that *The Statistician* would be aimed particularly at the professionally qualified members of the Society as well as at a wide international audience of practising statisticians.

Around the turn of the century, rapid advances in information technology took place, and the Society's journals evolved to take advantage of them.

Firstly, a complete run of every issue since the first in 1838 was digitized and housed in the on-line archive JSTOR. Not only is each page available to view in a facsimile form of the original but also the database has sophisticated search facilities for each item in the journals, providing powerful research tools. These tools have been extended even further for volumes since 1997 by inclusion in the CrossRef scheme, which links participating publishers' individual databases of journals to enable researchers to follow up references smoothly on line from one journal to another and from one discipline to another. Access is through Blackwell Synergy, which provides searchable hypertext on-screen versions of each paper in the Society's journals as well as downloadable versions in portable document format for subscribers.

At the prepublication end of the process, manuscripts are now almost exclusively submitted and handled in the refereeing process in electronic form.

Responding to these changes and the changing requirements of the statistical community, in 2002 the Society conducted a review of its publications. The outcome was a consolidation of its journals back into three series: a widely accessible subject-matter journal (Series A), a methodological journal (Series B) and a journal for innovative applications (Series C). In addition, in 2004 the Society launched a magazine titled *Significance*, which, as well as being of interest to all its members and other statisticians, performs an outreach role in promoting statistics to nonstatisticians.

The aims and scope of the three series and the magazine are as follows.

Series A (*Statistics in Society*) publishes papers that demonstrate how statistical thinking, design, and analyses play a vital role in all walks of life and benefit society in general. For example, important applications of statistical methods in medicine, business and commerce, industry, economics and finance, education and teaching, physical and biomedical sciences, the environment, the law, government and politics, demography, psychology, sociology, and sport all fall within the journal's remit. It is aimed at a wide statistical audience and at professional statisticians in particular. Its emphasis is on quantitative approaches to problems in the real world rather than the exposition of technical detail. Of particular interest are papers on topical or contentious statistical issues, papers that give reviews or *exposés* of current statistical concerns and papers demonstrating how statistics has contributed to our understanding of important substantive questions. Historical, professional, and biographical contributions are also welcome, as are discussions of methods of data collection and of ethical issues, provided that all such papers have substantial statistical relevance.

Series B (*Statistical Methodology*) publishes papers that contribute to the understanding of statistical methodology and/or develop and improve statistical methods. The kinds of contribution considered include descriptions of new methods of collecting or analyzing data, with the underlying theory, an indication of the scope of application and preferably a real example. Also considered are comparisons, critical evaluations, and new applications of existing methods, contributions to **probability theory**, which have a clear practical bearing (including the formulation and analysis of stochastic models), statistical computation or **simulation** where original methodology is involved and original contributions to the foundations of statistical science. Reviews of methodological techniques are also considered.

Series C (*Applied Statistics*) promotes papers that both are driven by real life problems and make a novel contribution to the subject, for example by developing methodology or by demonstrating the proper application of new or existing statistical methods to them. Applications are central, and case studies may therefore be particularly appropriate. Papers describing interdisciplinary work are especially welcome, as are those that give novel applications of existing methodology or new insights into the practical application of techniques. Methodological papers that are not motivated by a genuine application are not within the scope; nor are papers that include only brief numerical illustrations or describe simulations of properties of statistical techniques. However, papers describing developments in statistical computing are within the scope, provided that they are driven by practical examples. Other types of papers considered are those on design issues (e.g. in relation to experiments (*see* **Experimental Design**), surveys (*see* **Sample Surveys in the Health Sciences**) or **observational studies**) that arise from specific practical problems and feature an adequate description of a substantial application and a justification for any new theory.

Significance is a quarterly magazine for anyone interested in statistics and the analysis and

interpretation of data. Its aim is to communicate and demonstrate in an entertaining and thought-provoking way the practical use of statistics in all walks of life and to show how statistics benefit society. Articles are largely nontechnical and hence accessible and appealing not only to members of the profession but also to all users of statistics. Students and teachers of statistics will find articles of interest in *Significance*, as will people working in central and local government, medicine and health care, administration, economics, business and commerce, industry, social studies, survey research, science, and the environment. As well as promoting the discipline and covering topics of professional relevance internationally, *Significance* contains a mixture of statistics in the news, case studies, reviews of existing and newly developing areas of statistics, the application of techniques in practice and problem solving.

(*See also* **Royal Statistical Society**)

M.C. OWEN

J-shaped Distribution

As a sequel to Khinchin's definition of **unimodality**, a J-shaped distribution function is defined. A characterization for a related distribution is given using a well-known result of Khinchin on unimodality and a characterization theorem for a **U-shaped** probability density function by Ghosh and Shanbhag.

We define the following:

Definition 1. A distribution function $F(x)$ is said to be negative-tailed (positive-tailed) J-shaped if there exists a value $x = a$ such that $F(x)$ is convex (concave) for $x < a$ ($x > a$) and $F(a) = 1$ (0). The point $x = a$ is called a negative (positive) pivot of $F(x)$.

Definition 2. If a J-shaped distribution $F(x)$ is differentiable except at a countable subset of the set of reals, then the derivative $F'(x)$ is called a J-shaped probability density function.

We observe that for a negative-tailed J-shaped distribution function $F(x)$ with a negative pivot at $x = a$ we have $F(x) = 1$ at $x = a$ and for $x > a$. Since a constant function is both concave and convex, we conclude that a negative-tailed J-shaped distribution is unimodal with vertex at $x = a$. Hence, by Khinchin's theorem [3, p. 92] its **characteristic function** $p(t)$ has the following representation:

$$p(t) = \left[\frac{\exp(ita)}{t}\right] \int_0^t q(u)\,\mathrm{d}u, \quad \text{for a real } t.$$

Similarly, the characteristic function $r(t)$ of a positive-tailed J-shaped distribution has the following representation:

$$r(t) = \left[\frac{\exp(itc)}{t}\right] \int_0^t s(u)\,\mathrm{d}u, \quad \text{for a real } t,$$

where c is the positive pivot.

Examples

The following are two examples of J-shaped density functions:

1. Let a random variable X measure the level of nicotine intake by human beings. Then the frequency distribution of X amongst patients with lung cancer is likely to be negative-tailed J-shaped.
2. If a cohort of children is followed from birth to age 5 years, the distribution of age at death, amongst those who die, is likely to be positive-tailed J-shaped. Conversely, among those who die in a cohort of individuals followed from, say, age 65 to 70 years, the distribution of age at death is likely to be negative-tailed J-shaped.

Related Distributions

The concepts of a U-shaped probability density function and a bimodal distribution are related to J-shaped distribution. We define the following:

Definition 3. Let X be an absolutely continuous random variable. Then the probability density function of X is said to be U-shaped if there exist real numbers a, b, and c such that $a < b < c$, $\Pr\{X < a\} = 0$, $\Pr\{X > c\} = 0$, $\Pr\{X < x\}$ is concave in (a, b) and convex in (b, c).

The following theorem [2] gives a characterization for a U-shaped density function.

Theorem. Let X be an absolutely continuous random variable with probability density function $h(x)$. Then $h(x)$ is bounded and U-shaped if and only if the characteristic function of X is given by

$$f(t) = \frac{p[\exp(itb) - \exp(ita)]}{it} - q\frac{\exp(itc)}{t}\int_0^t r(u)\,\mathrm{d}u, \quad \text{for a real } t,$$

where $r(u)$ is a characteristic function, a, b, c, and p are real numbers, and q is real and positive.

Definition 4. A distribution function $F(x)$ is said to be bimodal if there exist real numbers a, b, and c with $a < b < c$ such that (i) $F(x)$ is convex for $x < a$; (ii) $F(x)$ is concave for $a < x < b$; (iii) $F(x)$ is convex for $b < x < c$; and (iv) $F(x)$ is concave for $x > c$. The points $x = a$ and $x = c$ are called two vertices of $F(x)$. The point $x = b$ is called an antimode of $F(x)$.

If $a = c$, then $F(x)$ is unimodal in Khinchin's sense.

The following theorem [1] characterizes a bimodal distribution.

Theorem. The function $f(t)$ is the characteristic function of a bimodal distribution function $F(x)$ with

vertices a and c, with $a < c$, if and only if

$$f(t) = F(a)h(t) + \{F(c) - F(a)\}g(t) + \{1 - F(c)\}k(t),$$

where $h(t)$ is the characteristic function of a negative-tailed J-shaped distribution with a negative pivot at $x = a$, $g(t)$ is the characteristic function of a U-shaped distribution over (a, c), and $k(t)$ is the characteristic function of a positive-tailed J-shaped distribution with a positive pivot at $x = c$.

References

[1] Ghosh, P. (1978). A characterization of a bimodal distribution, *Communications in Statistics – Theory and Methods* **7**, 475–477.

[2] Ghosh, P. & Shanbhag, D.N. (1972). A note on the characterization of a U-shaped probability density function, *Journal of Applied Probability* **9**, 684–685.

[3] Lukacs, E. (1970). *Characteristic Function*. Griffin, London.

PANKAJ GHOSH

Kalman Filter

The Kalman filter is the recursive **algorithm** devised by Kalman [7, 8] often used to provide estimates of parameters in state-space models of **time series**. The original ideas and those that flowed from them have had so much impact, see [4], that Kalman filtering is sometimes taken to denote the entire state-space approach.

State-space models were proposed by several authors in times series, for example, [9, 5] and [3], but have their roots in control engineering. They relate observations to "state variables" and have wide application. Suppose that an observed series $\{x_t\}$, which may be a vector series, can be written in terms of d unobserved state variables, say a $d \times 1$ vector $\{\boldsymbol{\alpha}_t\}$. The explicit form of our model relating state and observation is

$$x_t = \mathbf{H}\boldsymbol{\alpha}_t + \varepsilon_t \tag{1}$$

where $\mathbf{H}$ is a known $1 \times d$ **matrix** and ε_t is, as usual white noise, often called measurement noise.

To make life simpler, we also assume that the state variables satisfy

$$\boldsymbol{\alpha}_t = \boldsymbol{\phi}\boldsymbol{\alpha}_{t-1} + \mathbf{K}\boldsymbol{\eta}_t \tag{2}$$

where $\boldsymbol{\phi}$ is the $d \times d$ transition matrix, and $\boldsymbol{\eta}_\mathbf{t}$ is a n dimensional noise vector, uncorrelated with ε_t, with **covariance matrix** $\boldsymbol{\Sigma}$. The matrix $\mathbf{K}$ is a $d \times n$ parameter matrix.

This rather curious set of equations can be justified in terms of **conditional** expectations in a **multivariate normal distribution**, however, the main reason for our interest is in the set of updating equations, the so called *Kalman recursions*. Many authors, notably Harvey [4] have used the state-space equations explicitly giving rise to what are known as **structural equation models**. These are a subset of the ARIMA family (*see* **ARMA and ARIMA Models**) but are valuable in modeling terms as they give an alternative approach via state variables. It is worth noting that ARIMA models can be written in state-space form but without the ε_t term in (1); see [6].

Perhaps the most important point to appreciate is that the state-space form and the Kalman algorithm give us a relatively simple model and an effective method of computing the **likelihood**.

For any estimation method based on the state-space formulation of a model, we need estimates of the state variables. Finding the best approximation to the state variables (in a conditional mean sense) is known as the filtering problem, while the corresponding problem of finding the best approximation to the observations is the smoothing problem.

Suppose that our estimate of $\boldsymbol{\alpha}_t$ at time t is a_t while the estimate made at time t-1 is $\mathbf{a}_{t|t-1}$. These estimates will have variance matrices

$$\mathbf{C}_t = \mathrm{E}[(\boldsymbol{\alpha}_t - \mathbf{a}_t)(\boldsymbol{\alpha}_t - \mathbf{a}_t)^{'}] \tag{3}$$

$$\mathbf{C}_{t|t-1} = \mathrm{E}[(\boldsymbol{\alpha}_t - \mathbf{a}_{t|t-1})(\boldsymbol{\alpha}_t - \mathbf{a}_{t|t-1})^{'}] \tag{4}$$

The beauty of the state-space representation is that we have the Kalman filter updating equations; see [7] or [6]. A comprehensive discussion can also be found in [1]

The Kalman Filter Recursions

The prediction equations

$$\mathbf{a}_{t|t-1} = \boldsymbol{\phi}\mathbf{a}_{t-1}. \tag{5}$$

$$\mathbf{C}_{t|t-1} = \boldsymbol{\phi}\mathbf{C}_{t-1}\boldsymbol{\phi}^{'} = \mathbf{K}\boldsymbol{\Sigma}\mathbf{K}^{'}. \tag{6}$$

The updating equations

$$F_t = \mathbf{H}\mathbf{C}_{t|t-1}\mathbf{H}^{'} + \sigma_\varepsilon^2 \tag{7}$$

$$\mathbf{C}_t = \mathbf{C}_{t|t-1} - \frac{1}{F_t}\mathbf{C}_{t|t-1}\mathbf{H}^{'}\mathbf{H}\mathbf{C}_{t|t-1} \tag{8}$$

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} + \frac{1}{F_t}\mathbf{C}_{t|t-1}\mathbf{H}^{'}(x_t - \mathbf{H}\mathbf{a}_{t|t-1}). \tag{9}$$

For constructing the likelihood,

$$V_t = x_t - \mathbf{H}\mathbf{a}_{t|t-1}. \tag{10}$$

Given some starting values, we step through the recursions and at each stage we obtain the **prediction** errors V_t and the prediction error variances F_t for any given parameter set. This means we can compute the likelihood for any parameter set, and with a suitable maximization procedure, we can get **maximum likelihood** estimates. The Kalman approach provides compact computer code with the possibility of fast execution by comparison with alternative approaches.

The Start-up Problem

If we start our recursions at time $t = 1$, then we need the state estimate $\mathbf{a}_0$ together with a covariance estimate $\mathbf{C}_0$ at time zero. There are various possibilities, using the unconditional expected values to rather exotic ones; see [3].

A simple approach is to use the fact that the effect of the starting values is soon lost, especially with long series and we will set $\mathbf{a}_0$ to zero and the covariance $\mathbf{C}_0$ to M (a large number) times the unit matrix. This is a simple, an effective technique that is widely used. It is also common for the recursions to settle down to a steady state when we have stationary series. By this we mean that $\mathbf{C}_t$, $\mathbf{C}_{t|t-1}$, and F_t converge to fixed (time independent) values. The advantage is that when this happens, we can skip equations (6), (7), and (8). While there is no analytic result to tell us when this has happened, we can put a numerical check in the recursions.

The reader may have noticed that we could have used a vector value of x_t in most of the algebra above. Indeed, we can easily extend all our state-space models to vector series.

An Example. We take as an example a tree ring series from 1700 to 1987 [2] shown in Figure 1

Take a simple model of a drifting mean, $x_t = \mu_t + \varepsilon_t$ and

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \text{ with } \beta_t = \beta_{t-1} + \zeta_t \quad (11)$$

This simple model includes trend and mean effects but no seasonal. We start at t = 1 with

$$\boldsymbol{\phi} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{pmatrix},$$
$$\mathbf{C}_0 = \begin{pmatrix} 1000 & 0 \\ 0 & 1000 \end{pmatrix}, \quad \mathbf{a}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (12)$$

The $\mathbf{K}$ matrix is just a unit matrix.

Our life can be made a little simpler by noting that we have three variances $\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\zeta^2$ What we can do is *concentrate out a parameter*. We set $\sigma_\varepsilon^2 = 1$ and regard the other two unknown variances as being scaled by σ_ε^2; see [6]. This saves us some computation. We can now use the Kalman filter equations to produce for given $\sigma_\eta^2, \sigma_\zeta^2$ a set of predictions and predictions errors, which will give us values for the likelihood. The table below shows how the $\mathbf{C}_t$ matrix converges as the value of t increases. In fact, we would normally drop the first six iteration values from the likelihood computation because they are unreliable.

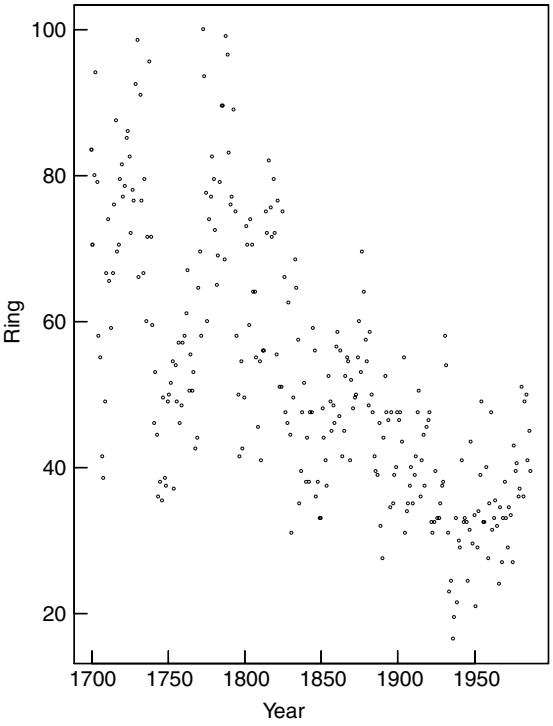

Figure 1 Model fit to tree ring widths

t	$\mathbf{a}_t$	$\mathbf{C}_t$		V_t
2	70.5	5002.25	5001.25	−54.74791
	−12.98906	5001.25	5002.25	
4	94	2.666644	1.666644	11.99685
	9.501143	1.666644	2.666644	
10	49	2.618034	1.618034	16.65551
	4.138162	1.618034	2.618034	
14	59	2.618034	1.618034	4.825807
	−4.6567	1.618034	2.618034	

As we can now construct a likelihood via the filter, we are able to seek the maximum of the likelihood. After 100 iterations, we find that the (scaled) values are

$$\sigma_\varepsilon^2 = 1, \quad \sigma_\eta^2 = 204.1846, \quad \sigma_\zeta^2 = 137.2181. \quad (13)$$

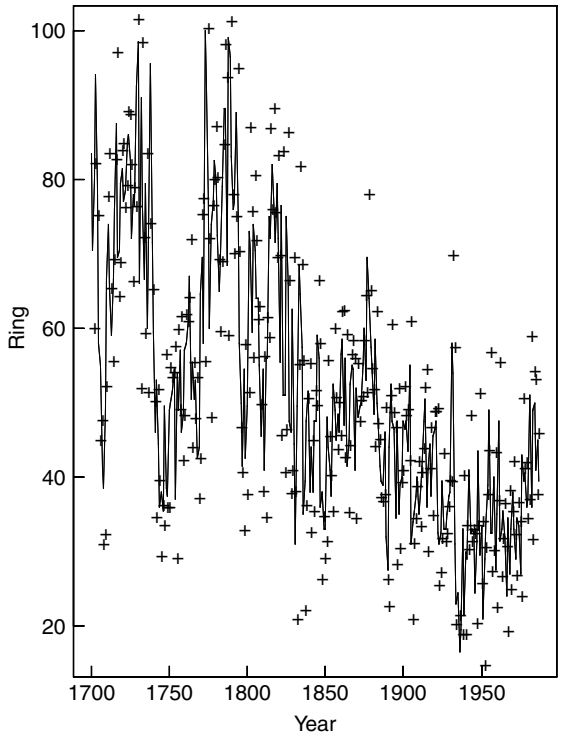

Figure 2 Tree ring widths by year

Standard errors are also available from the derivatives of the likelihood.

This is a really rather simple model and as we can see, the predicted values "+" are not very close (see Figure 2). This is, however, a fault of the model rather than the filter!

References

[1] Durbin, J. & Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. University Press, Oxford.

[2] Earle, C.J., Brubaker, L.B., Segura, G. *Douglas Fir Ring Widths, Silver Creek, Washington State*, International Tree Ring Data Base, NOAA/NGDC Paleoclimatology Program, Boulder, Colorado, USA. see http://www.ngdc.noaa.gov/paleo/treering.html

[3] Harvey, A.C. (1989). *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, UK.

[4] Harvey, A.C. (1993). *Time Series Models*. Prentice Hall, USA.

[5] Harrison, P.J. & Stevens, C.F. (1976). Bayesian forecasting, *Journal of the Royal Statistical Society Series B* **38**, 205–247.

[6] Janacek, G.J. & Swift, A.L. (1990). *Times series*, Ellis Horwood, Chichester, UK.

[7] Kalman, R.E. (1960). A new approach to linear filtering, *Journal of Basic Engineering, Transactions of the ASM, Series D* **82**, 35–45.

[8] Kalman R.E. & Bucy, R.S. (1961). New results in linear filtering and prediction theory, *Journal of Basic Engineering, Transactions of the ASM, Series D* **83**, 95–108.

[9] West, M. & Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Methods*. Springer, New York.

G.J. JANACEK

Kaplan–Meier Estimator

The Kaplan–Meier estimator is a **nonparametric** estimator which may be used to estimate the **survival distribution** function from **censored data**. The estimator may be obtained as the limiting case of the classical actuarial (**life table**) estimator, and it seems to have been first proposed by Böhmer [2]. It was, however, lost sight of by later researchers and not investigated further until the important paper by Kaplan & Meier [12] appeared. Today the estimator is usually named after these two authors, although sometimes it is denoted the product–limit estimator (*see* **Aalen–Johansen Estimator**). Below we describe the Kaplan–Meier estimator, illustrate its use in one particular case, and discuss estimation of the median and mean survival times. Furthermore, we show how the Kaplan–Meier estimator can be given as the product–integral of the **Nelson–Aalen estimator**, and indicate how this may be used to study its statistical properties. For almost four decades the Kaplan–Meier estimator has been one of the key statistical methods for analyzing censored survival data, and it is discussed in most textbooks on survival analysis. Rigorous derivations of the statistical properties of the estimator are provided in the books by Fleming & Harrington [7] and Andersen et al. [1]. In particular the latter presents formal proofs of almost all the results reviewed below as well as an extensive bibliography.

The Estimator and Confidence Intervals

Consider the survival data situation where we want to study the time to death (or some other event) for a homogeneous population with survival distribution function $S(t)$ representing the probability that an individual will be alive at time t. Assume that we have a sample of n individuals from this population. Our observation of the survival times for these individuals will typically be subject to right-censoring, meaning that for some individuals we only know that their true survival times exceed certain censoring times. The censoring is assumed to be independent in the sense that the additional knowledge of censorings before any time t does not alter the risk of failure at t. We denote by $t_1 < t_2 < \cdots$ the times when deaths are observed and let d_j be the number of individuals who die at t_j.

The Kaplan–Meier estimator for the survival distribution function then takes the form

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{r_j}\right), \tag{1}$$

where r_j is the number of individuals at risk (i.e. alive and not censored: in the **risk set**) just prior to time t_j. If there are no censored observations, then (1) reduces to one minus the empirical distribution function. The variance of the Kaplan–Meier estimator is estimated by Greenwood's formula:

$$\hat{\sigma}^2(t) = \hat{S}(t)^2 \sum_{t_j \le t} \frac{d_j}{r_j(r_j - d_j)}. \tag{2}$$

In the case of no censoring, (2) reduces to $\hat{S}(t)[1 - \hat{S}(t)]/n$, the standard **binomial** variance estimator.

In large samples the Kaplan–Meier estimator, evaluated at a given time t, is approximately normally distributed so that a standard $100(1-\alpha)\%$ confidence interval for $S(t)$ takes the form

$$\hat{S}(t) \pm z_{1-\alpha/2}\hat{\sigma}(t), \tag{3}$$

with $z_{1-\alpha/2}$ the $1-\alpha/2$ fractile of the **standard normal** distribution. The approximation to the normal distribution is improved by using the log-minus-log transformation (*see* **Quantal Response Models**) giving the **confidence interval**

$$\hat{S}(t)^{\exp\{\pm z_{1-\alpha/2}\hat{\sigma}(t)/[\hat{S}(t)\ln\hat{S}(t)]\}}. \tag{4}$$

This interval is satisfactory for quite small sample sizes [3]. Confidence intervals with small-sample properties which are comparable with (4), or even slightly better, may be obtained by using the arcsine-square-root transformation [3] or by basing the confidence interval on the **likelihood ratio test** [5, Section 4.3; 16]. Note that all these confidence intervals should be given a pointwise interpretation. Simultaneous confidence bands for the survival distribution function are considered below.

Right-censoring is not the only kind of data incompleteness in survival analysis. Often, e.g. in epidemiologic applications, individuals are not followed from time zero (in the relevant time scale, typically age), but only from a later entry time (conditional on survival until this entry time). Thus, in addition to right-censoring, the survival data are subject to left truncation. For such data we may, in

principle at least still use the Kaplan–Meier estimator (1) and estimate its variance by (2). The number at risk, r_j, is now the number of individuals who have entered the study before time t_j and are still in the study just prior to t_j. However, for left-truncated data the numbers at risk, r_j, will often be low for small values of t_j. This will result in estimates $\hat{S}(t)$ which have large sampling errors and which therefore may be of little practical use. What can be usefully estimated in such situations is the conditional survival distribution function, $S(t|t_0) = S(t)/S(t_0)$, representing the probability of survival to time t given that an individual is alive at time $t_0 < t$. It may be useful to estimate such conditional distribution functions for several values of t_0 (at which there are reasonable numbers at risk), there being nothing canonical about any particular value. The estimation is performed as described earlier, the only modification being that the product in (1) and the sum in (2) are restricted to those t_j for which $t_0 < t_j \leq t$.

An Illustration

As an illustration we use data from a randomized clinical trial for patients with histologically verified liver cirrhosis. Patients were recruited from several hospitals in Copenhagen between 1962 and 1969 and were followed until death, lost to follow-up, or until the closing date of the study, October 1, 1974. The time variable of interest is time since entry into the study. Patients are right censored if alive on October 1, 1974, or if lost to follow-up before that date.

We consider only the 138 placebo-treated male patients. Their median age at entry was 57 years, while the lower and upper quartiles were 51 and 66 years, respectively. Of the 138 patients, 88 died during the study. The Kaplan–Meier estimate of the survival distribution function for these patients is shown in Figure 1 with 95% confidence intervals computed according to (4). From the figure we see, for example, that the five years survival probability is estimated as 43.0% with a 95% confidence interval from 34.0% to 51.9%, while the estimated 10 years survival probability is 18.4% with a confidence interval from 9.7% to 29.3%. We return to the liver cirrhosis example below in connection with median and mean survival times and simultaneous confidence bands. A further discussion and analysis of the data are given by Schlichting et al. [15]. The data were also used for illustrative purposes by Andersen et al. [1].

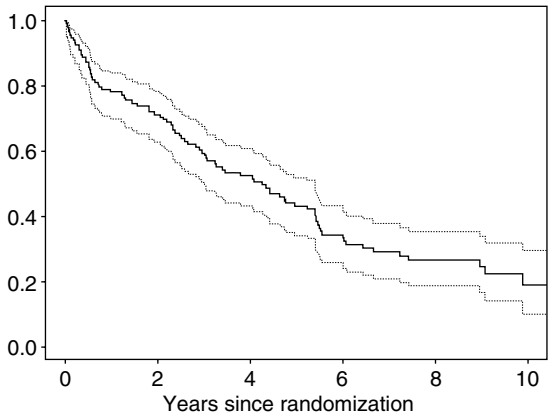

Figure 1 Kaplan–Meier estimate of the survival distribution function for 138 placebo-treated male patients with liver cirrhosis with 95% log-minus-log-transformed confidence intervals

Median Survival Time and Related Quantities

The use of the Kaplan–Meier estimator is not restricted to estimating survival probabilities for given times t. It may also be used to estimate fractiles such as the **median survival time** and related quantities like the interquartile range (*see* **Quantiles**).

Consider the pth fractile, ξ_p, of the cumulative distribution function $F(t) = 1 - S(t)$, and assume that $F(t)$ has a positive density function $f(t) = F'(t) = -S'(t)$ in a neighborhood of ξ_p. Then ξ_p is uniquely determined by the relation $F(\xi_p) = p$, or equivalently, $S(\xi_p) = 1 - p$. The Kaplan–Meier estimator is a step function and hence does not necessarily attain the value $1 - p$. Therefore a similar relation cannot be used to define the estimator $\hat{\xi}_p$ of the pth fractile. Rather, we define $\hat{\xi}_p$ to be the smallest value of t for which $\hat{S}(t) \leq 1 - p$, i.e. the time t where $\hat{S}(t)$ jumps from a value greater than $1 - p$ to a value less than or equal to $1 - p$. In large samples $\hat{\xi}_p$ is approximately normally distributed with a variance that may be estimated by

$$\widehat{\text{var}}(\hat{\xi}_p) = \frac{(1-p)^2 \hat{\sigma}^2(\hat{\xi}_p)}{[\hat{f}(\hat{\xi}_p)\hat{S}(\hat{\xi}_p)]^2}. \tag{5}$$

Here $\hat{f}(t)$ is an estimator for the density function $f(t) = -S'(t)$ (*see* **Density Estimation**). One may, for example, use

$$\hat{f}(t) = \frac{1}{2b}\left[\hat{S}(t-b) - \hat{S}(t+b)\right] \qquad (6)$$

for a suitable bandwidth b (corresponding to a kernel function estimator with uniform kernel). Furthermore, for $p < q$, $\hat{\xi}_p$ and $\hat{\xi}_q$ are approximately binormally distributed, and their correlation may be estimated by

$$\widehat{\text{corr}}(\hat{\xi}_p, \hat{\xi}_q) = \frac{\hat{\sigma}(\hat{\xi}_p)\hat{S}(\hat{\xi}_q)}{\hat{\sigma}(\hat{\xi}_q)\hat{S}(\hat{\xi}_p)}. \qquad (7)$$

Note that $\hat{S}(\hat{\xi}_p)$ in (5) and (7) is equal to or only slightly less than $1 - p$, and that (5) could have been simplified if we had used this approximate equality. We have chosen not to do so since then $\hat{S}(\hat{\xi}_p)$ in (5) and (7) cancels with the same factor in $\hat{\sigma}(\hat{\xi}_p)$; cf. (2).

The above results may be used in the usual way to determine approximate confidence intervals, e.g. for the median survival time $\xi_{0.50}$ and the interquartile range $\xi_{0.75} - \xi_{0.25}$, as illustrated below. For the purpose of determining a confidence interval for a quantile (fractile) like the median it is, however, better to apply the approach of Brookmeyer & Crowley [4]. For the pth fractile one then uses as a confidence interval all hypothesized values ξ_p^0 of ξ_p which are not rejected when testing the null hypothesis $\xi_p = \xi_p^0$ against the alternate hypothesis $\xi_p \neq \xi_p^0$ at the α level (*see* **Hypothesis Testing**). Such test-based confidence intervals can be read directly from the lower and upper confidence limits for the survival distribution function in exactly the same manner as $\hat{\xi}_p$ can be read from the Kaplan–Meier curve itself (*see* **Median Survival Time**).

For the liver cirrhosis data an estimate of the median survival time is 4.27 years (standard error 0.66 years), while the lower and upper quartiles are estimated as 1.46 years (0.35 years) and 8.97 years (1.13 years), respectively, with an estimated correlation of 0.28. In these computations the bandwidth $b = 1$ year was used in (6). An estimate of the interquartile range of the survival distribution function is $8.97 - 1.46 = 7.51$ years, with standard error $(0.35^2 + 1.13^2 - 2 \times 0.35 \times 1.13 \times 0.28)^{1/2} = 1.09$ years. From this an approximate 95% confidence interval for the median survival time is $4.27 \pm 1.96 \times 0.66$, i.e. from 2.98 to 5.56 years, while 95% confidence limits for the interquartile range are from 5.37 to 9.65 years. For the median survival time it is, as mentioned earlier, better to read the confidence limits directly from the pointwise confidence intervals for the survival distribution function given in Figure 1. This gives 95% confidence limits for the median survival time from 3.02 years to 5.41 years. Note that no estimate of the density function is needed here.

Mean Survival Time

Owing to right-censoring, in most survival studies it will not be possible to obtain reliable estimates for the mean survival time $\mu = \int_0^\infty t f(t)\,\mathrm{d}t = \int_0^\infty S(t)\,\mathrm{d}t$ (*see* **Life Expectancy**). This is one important reason why, in survival analysis, the median is a more useful measure of location than the mean. What may be usefully estimated from right-censored survival data is the expected time lived in a given interval $[0, t]$, i.e. $\mu_t = \int_0^t S(u)\,\mathrm{d}u$. This is estimated by

$$\hat{\mu}_t = \int_0^t \hat{S}(u)\,\mathrm{d}u,$$

the area below the Kaplan–Meier curve between 0 and t. Such an estimate may be of interest in its own right, or it may be compared with a similar population-based estimate to assess the expected number of years lost up to time t for a group of patients. In large samples, $\hat{\mu}_t$ is approximately normally distributed with a variance that may be estimated by

$$\widehat{\text{var}}(\hat{\mu}_t) = \sum_{t_j \le t} \frac{(\hat{\mu}_t - \hat{\mu}_{t_j})^2 d_j}{r_j(r_j - d_j)},$$

a result which may be used to give approximate confidence limits for μ_t. By letting t tend to infinity, the above results may be extended to the estimation of the mean μ itself [8]. However, the conditions (mainly on the censoring) needed for such an extension to be valid are usually not met in practice.

In the liver cirrhosis study no patient was followed for more than 13 years, making the estimation of the mean survival time impossible. We may, however, estimate the expected number of years lived up to a given time t. In particular, estimates for the expected number of years lived up to 5 years and 10 years after

the start of the study are 3.29 years (standard error 0.17 years) and 4.73 years (0.33 years), respectively.

Redistribute-to-the-right Algorithm and Self-consistency

We mentioned earlier the relationship between the Kaplan–Meier estimator and the empirical distribution function in the case of no censoring. The redistribute-to-the-right algorithm and the concept of self-consistency, both due to Efron [6], further illustrate this relation.

For notational convenience we assume that there are no ties, and we denote by $t_1^0 < t_2^0 < \cdots < t_n^0$ the ordered times of deaths and censorings combined. The redistribute-to-the-right algorithm is as follows. First, we construct the ordinary empirical (survival) distribution function which places probability mass $1/n$ at each of the observed times t_j^0. If $t_{j_1}^0$ is the smallest t_j^0 that corresponds to a censored observation, then we remove its mass and redistribute it equally among the $n - j_1$ time-points to the right of it. Then, if $t_{j_2}^0$ is the second smallest censored observation, we remove its mass, which will be $1/n + 1/[n(n - j_1)]$, and redistribute it equally among the $n - j_2$ time-points to its right, etc. This algorithm will converge in a finite number of steps to the Kaplan–Meier estimator (1) (with the modification that it is set equal to zero after t_n^0 also when this last time-point corresponds to a censored observation).

A self-consistent estimator $\tilde{S}(t)$ for the survival distribution function equals $1/n$ times an estimate for the number of individuals who survive time t. More precisely,

$$\tilde{S}(t) = \frac{1}{n}\left[\#(t_j^0 > t) + \sum_{t_j^0 \le t} a_j(t)\right], \qquad (8)$$

where $a_j(t) = \tilde{S}(t)/\tilde{S}(t_j^0)$ if the observation at t_j^0 corresponds to a censored observation, and $a_j(t) = 0$ if it corresponds to an observed death. It turns out that the Kaplan–Meier estimator (modified as just indicated) is the unique self-consistent estimator. Turnbull [17] (*see* **Turnbull Estimator**) used the idea of self-consistency to derive an iterative procedure (a version of the **EM algorithm**) for estimating the survival distribution function nonparametrically from arbitrarily grouped, censored, and truncated data, while Gill [9] showed that the self-consistency equation, (8), may be interpreted as a generalized score equation.

Product–Integral Representation and Relationship to the Nelson–Aalen Estimator

Usually one assumes that the survival distribution function $S(t)$ is absolute continuous with density function $f(t) = -S'(t)$, hazard rate function $\alpha(t) = f(t)/S(t)$, and cumulative hazard rate function $A(t) = \int_0^t \alpha(u)\,\mathrm{d}u$. However, the Kaplan–Meier estimator is discrete in nature, and the same applies to the Nelson–Aalen estimator for the cumulative hazard rate function. This makes it useful to be able to handle both discrete and continuous distributions within a unified framework. Let us therefore review how the survival distribution function $S(t)$ and the cumulative hazard rate function $A(t)$ are related for distributions which need neither to be continuous nor discrete. For such distributions

$$A(t) = -\int_0^t \frac{\mathrm{d}S(u)}{S(u-)}, \qquad (9)$$

where $S(t-)$ denotes the left-hand limit of the survival distribution function at t. For an absolute continuous distribution, (9) specializes to $A(t) = -\ln S(t) = \int_0^t \alpha(u)\,\mathrm{d}u$. For a discrete distribution it gives $A(t) = \sum_{u \le t} \alpha_u$, where the discrete hazard, α_t, is the conditional probability of death exactly at time t given that death has not occurred earlier. To express the survival distribution function by the cumulative hazard rate function it is convenient to use the product–integral $\prod\!\!\!\!\prod$, defined as the limit of approximating finite products in a similar manner as the ordinary integral $\int$ is defined as the limit of approximating finite sums (*see* **Product-integration**). With the use of the product–integral we may write

$$S(t) = \prod_{u \le t}\!\!\!\!\prod [1 - \mathrm{d}A(u)]. \qquad (10)$$

For a continuous distribution, (10) specializes to the well-known relation $S(t) = \exp[-A(t)]$, while for a discrete distribution it takes the form $S(t) = \prod_{u \le t}(1 - \alpha_u)$.

The Nelson–Aalen estimator for the cumulative hazard rate function is $\hat{A}(t) = \sum_{t_j \le t} d_j/r_j$. This corresponds to a distribution with all probability mass

concentrated at the observed failure times and with discrete hazard $\hat{\alpha}_j = d_j/r_j$ at t_j. Using (10), the corresponding survival distribution function takes the form

$$\hat{S}(t) = \prod_{u \le t} [1 - \mathrm{d}\hat{A}(u)] = \prod_{t_j \le t} (1 - \hat{\alpha}_j), \qquad (11)$$

i.e. it is the Kaplan–Meier estimator (1). Thus the Kaplan–Meier and Nelson–Aalen estimators are related in exactly the same way as are the survival distribution function and the cumulative hazard rate function themselves. This fact is lost sight of when one considers the relations $A(t) = -\ln S(t)$ and $S(t) = \exp[-A(t)]$ which are only valid for the continuous case. In fact, the latter relations have led researchers to suggest the estimators $-\ln \hat{S}(t)$ and $\exp[-\hat{A}(t)]$ for the cumulative hazard rate function and the survival distribution function, respectively. The numerical differences between these two estimators and the Nelson–Aalen and Kaplan–Meier estimators will be of little importance in most cases. But the fact that the Nelson–Aalen and Kaplan–Meier estimators are related through (9) and (10) indicates that they are the canonical nonparametric estimators for the cumulative hazard rate function and the survival distribution function. This statement is supported by the fact that they may both be given a **nonparametric maximum likelihood** interpretation [11].

Martingale Representation and Statistical Properties

The product–integral formulation (11) of the Kaplan–Meier estimator shows its close relationship to the Nelson–Aalen estimator, and it is the key to the study of its statistical properties. In fact, these are closely related to those of the Nelson–Aalen estimator. We here indicate a few main steps and refer to Andersen et al. [1, Section IV.3] for a detailed account.

Let $J(t) = 1$ if there is at least one individual at risk just before time t; $J(t) = 0$ otherwise. Furthermore, introduce $A^*(t) = \int_0^t J(u)\,\mathrm{d}A(u)$, and let

$$S^*(t) = \prod_{u \le t} [1 - \mathrm{d}A^*(u)]. \qquad (12)$$

We note that (12) is almost the same as $S(t)$ [cf. (10)] when there is only a small probability that there is no one at risk at times $u \le t$. By a general result for product–integrals (Duhamel's equation), we may write

$$\frac{\hat{S}(t)}{S^*(t)} - 1 = -\int_0^t \frac{\hat{S}(u-)}{S^*(u)}\,\mathrm{d}(\hat{A} - A^*)(u). \qquad (13)$$

Here $\hat{A} - A^*$ is a square integrable martingale (*see* **Nelson–Aalen Estimator**). It follows that the right-hand side of (13) is a stochastic integral and hence itself a mean zero square integrable martingale. As a consequence of this, $\mathrm{E}[\hat{S}(t)/S^*(t)] = 1$ for any given t, so the Kaplan–Meier estimator is almost **unbiased**. Furthermore, the predictable variation process of the martingale on the right-hand side of (13) may be used to arrive at an estimator for the variance of $\hat{S}(t)/S^*(t)$. From this, Greenwood's formula (2) follows provided one adopts a general model, not necessarily continuous. Greenwood's formula may also be derived through a standard **information** calculation starting with a binomial-type likelihood for such a general model.

A further consequence of (13) is that $\sqrt{(n)}(\hat{S} - S)/S$ is asymptotically equivalent to $-\sqrt{(n)}(\hat{A} - A)$ and therefore converges weakly to a mean zero Gaussian martingale. In particular, for a fixed t, the Kaplan–Meier estimator (1) is asymptotically normally distributed, a fact that was used in connection with the confidence intervals (3) and (4). Also, the asymptotic distributional results of the estimators for the median and mean survival times reviewed earlier are consequences of this weak convergence result.

Confidence Bands

The weak convergence of $\sqrt{(n)}(\hat{S} - S)/S$ to a mean zero Gaussian martingale also makes it possible to derive confidence bands for the survival distribution function, i.e. limits that contain $S(t)$ for all t in an interval $[\tau_1, \tau_2]$ with a prespecified probability. Two important types of such confidence bands are the equal precision bands [14] and the Hall–Wellner bands [10]. Borgan & Liestøl [3] derived transformed versions of these confidence bands and compared them with the nontransformed ones.

The standard and log-minus-log transformed equal precision bands are obtained by replacing $z_{1-\alpha/2}$ in (3) and (4) by $d_{1-\alpha}(\hat{c}_1, \hat{c}_2)$, the $1 - \alpha$ fractile in the distribution of the supremum of the absolute value

of a standardized Brownian bridge over the interval from $\hat{c}_1$ to $\hat{c}_2$ (*see* **Brownian Motion and Diffusion Processes**). Here

$$\hat{c}_i = \frac{n[\hat{\sigma}(\tau_i)/\hat{S}(\tau_i)]^2}{1 + n[\hat{\sigma}(\tau_i)/\hat{S}(\tau_i)]^2}, \quad i = 1, 2. \tag{14}$$

The fractile $d_{1-\alpha}(\hat{c}_1, \hat{c}_2)$ may be found (approximately) by solving (with respect to d) the following nonlinear equation:

$$\frac{4\phi(d)}{d} + \phi(d)\left(d - \frac{1}{d}\right)\ln\left[\frac{\hat{c}_2(1-\hat{c}_1)}{\hat{c}_1(1-\hat{c}_2)}\right] = \alpha,$$

with $\phi(d)$ the standard normal density. The equal precision bands require $0 < \hat{c}_1 < \hat{c}_2 < 1$, so they cannot be extended all the way down to $t = 0$. Typically, one will also omit the largest values of t.

The nontransformed Hall–Wellner band takes the form

$$\hat{S}(t) \pm n^{-1/2} e_{1-\alpha}(\hat{c}_1, \hat{c}_2)\left\{1 + n\left[\frac{\hat{\sigma}(t)}{\hat{S}(t)}\right]^2\right\}\hat{S}(t). \tag{15}$$

Here $e_{1-\alpha}(\hat{c}_1, \hat{c}_2)$ is the $1-\alpha$ fractile in the distribution of the supremum of the absolute value of a Brownian bridge over the interval from $\hat{c}_1$ to $\hat{c}_2$; cf. (14). For completely observed survival data the Hall–Wellner band reduces to the well-known Kolmogorov band $\hat{S}(t) \pm n^{-1/2} e_{1-\alpha}(\hat{c}_1, \hat{c}_2)$. For the band (15), one will often let $\tau_1 = 0$, in which case tables of $e_{1-\alpha}(\hat{c}_1, \hat{c}_2) = e_{1-\alpha}(0, \hat{c}_2)$ are given, for example by Koziol & Byar [13] and Hall & Wellner [10] for selected values of α and $\hat{c}_2$. We note that (15) is obtained from (3) by substituting $n^{-1/2} e_{1-\alpha}(\hat{c}_1, \hat{c}_2)\{1 + n[\hat{\sigma}(t)/\hat{S}(t)]^2\}\hat{S}(t)$ for $z_{1-\alpha/2}\hat{\sigma}(t)$. The same substitution in (4) gives the log-minus-log transformed Hall–Wellner band. This transformed band requires $\hat{c}_1 > 0$, so it cannot be extended all the way down to $t = 0$. Owing to the approximation $e_{1-\alpha}(\hat{c}_1, \hat{c}_2) \approx e_{1-\alpha}(0, \hat{c}_2)$, the above-mentioned tables may also be used for the transformed bands when $\hat{c}_1$ is close to zero.

The nontransformed equal precision band tends to achieve too high error rates when the number of observations is low, and the use of transformed bands is recommended, even for samples of a hundred or more. The achieved error rates of the nontransformed Hall–Wellner band are fairly close to the nominal ones even in small samples, and the improvement obtained by using transformed bands is of less importance.

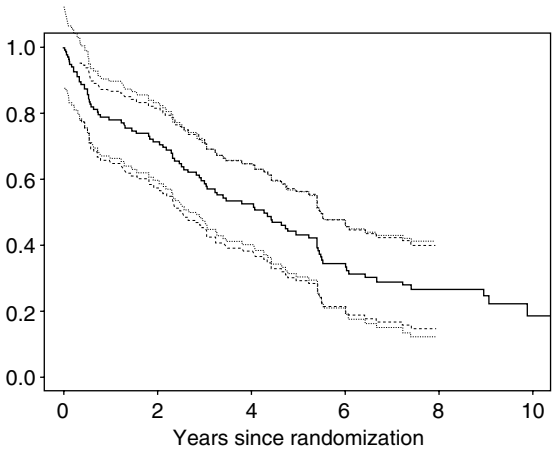

Figure 2 Kaplan–Meier estimate of the survival distribution function for 138 placebo-treated male patients with liver cirrhosis with 95% confidence bands: log-minus-log transformed equal precision band over the interval from 4 months to 8 years (- - - -); Hall–Wellner band over the interval [0, 8] years (· · ·)

Figure 2 shows the Kaplan–Meier estimate for the liver cirrhosis data with 95% confidence bands. The bands shown are the log-minus-log transformed equal precision band over the interval from 4 months to 8 years and the nontransformed Hall–Wellner band valid from time zero to 8 years. Since $\tau_1 = 1/3$ year and $\tau_2 = 8$ years correspond to $\hat{c}_1 = 0.090$ and $\hat{c}_2 = 0.789$, the fractiles $d_{0.95}(\hat{c}_1, \hat{c}_2) = 2.99$ and $e_{0.95}(0, \hat{c}_2) = 1.36$ were used. It is seen that the equal precision band is narrower than the Hall–Wellner band both for low and high values of t, while the Hall–Wellner band is slightly narrower than the equal precision band for intermediate values.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Böhmer, P.E. (1912). Theorie der unabhängigen Wahrscheinlichkeiten, *Reports, Memoirs and Proceedings, Seventh International Congress of Actuaries, Amsterdam*, Vol. 2, pp. 327–343.

[3] Borgan, Ø. & Liestøl, K. (1990). A note on confidence intervals and bands for the survival curve based on transformations, *Scandinavian Journal of Statistics* **17**, 35–41.

[4] Brookmeyer, R. & Crowley, J.J. (1982). A confidence interval for the median survival time, *Biometrics* **38**, 29–41.

[5] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[6] Efron, B. (1967). The two sample problem with censored data, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. Prentice Hall, New York, pp. 831–853.

[7] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[8] Gill, R.D. (1983). Large sample behavior of the product-limit estimator on the whole line, *Annals of Statistics* **11**, 49–58.

[9] Gill, R.D. (1989). Non- and semi-parametric maximum likelihood estimation and the von Mises method (Part 1), *Scandinavian Journal of Statistics* **16**, 97–128.

[10] Hall, W.J. & Wellner, J.A. (1980). Confidence bands for a survival curve from censored data, *Biometrika* **67**, 133–143.

[11] Johansen, S. (1978). The product limit estimator as maximum likelihood estimator, *Scandinavian Journal of Statistics* **5**, 195–199.

[12] Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481, 562–563.

[13] Koziol, J.A. & Byar, D.P. (1975). Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data, *Technometrics* **17**, 507–510.

[14] Nair, V.N. (1984). Confidence bands for survival functions with censored data: a comparative study, *Technometrics* **26**, 265–275.

[15] Schlichting, P., Christensen, E., Andersen, P.K., Fauerholdt, L., Juhl, E., Poulsen, H. & Tygstrup, N., for The Copenhagen Study Group for Liver Diseases (1983). Prognostic factors in cirrhosis identified by Cox's regression model, *Hepatology* **3**, 889–895.

[16] Thomas, D.R. & Grunkemeier, G.L. (1975). Confidence interval estimation of survival probabilities for censored data, *Journal of the American Statistical Association* **70**, 865–871.

[17] Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 290–295.

ØRNULF BORGAN

Kappa and its Dependence on Marginal Rates

The **kappa** statistic was proposed by Cohen [4] as a measure of reliability for **nominal** classification procedures and was constructed specifically to "correct" the proportion of raw agreement for agreement expected purely by random classifications given the marginal rates. Since its introduction there have been many generalizations and extensions developed and it has been applied widely in medical research. There has also been considerable debate about the utility of this index [9, 13, 14], arising in part as a result of a genuine lack of consensus on precisely how to model and measure the reliability of nominal classification procedures. The issues are most easily illustrated in the assessment of the reliability of a simple **binary** test. Let T_k denote the outcome of the kth application of a binary test for which $T_k = 1$ and $T_k = 2$ indicate the presence and absence of disease, respectively, $k = 1, 2$. Upon two applications of this test to a sample of n subjects, cross-classifying the results leads to a **2 x 2 table** (see Table 1), where x_{ij} denotes the frequency with which $T_1 = i$ and $T_2 = j$, $x_{i\cdot} = \sum_{j=1}^{2} x_{ij}$, and $x_{\cdot j} = \sum_{i=1}^{2} x_{ij}$. Conditioning on n leads to a **multinomial distribution** for $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})'$, where p_{ij} is the probability of $T_1 = i$ and $T_2 = j$, $p_{i\cdot} = \sum_{j=1}^{2} p_{ij}$, and $p_{\cdot j} = \sum_{i=1}^{2} p_{ij}$. The raw agreement is $p_0 = \sum_{i=1}^{2} p_{ii}$ and, given the marginal rates $p_{i\cdot}, i = 1, 2$ and $p_{\cdot j}, j = 1, 2$, and under the assumption of independent classifications, the expected level of agreement is $p_e = \sum_{i=1}^{2} p_{i\cdot} p_{\cdot i}$ The kappa index takes the form $\kappa = (p_0 - p_e)/(1 - p_e)$. The estimate of kappa, subsequently referred to as the kappa statistic and denoted $\hat{\kappa}$, is obtained by replacing p_0 and p_e by the corresponding estimates $\hat{p}_0 = \sum_{i=1}^{2} \hat{p}_{ii}$ and $\hat{p}_e = \sum_{i=1}^{2} \hat{p}_{i\cdot}\hat{p}_{\cdot i}$ respectively, where $\hat{p}_{ii} = x_{ii}/n$, $\hat{p}_{i\cdot} = x_{i\cdot}/n$, and $\hat{p}_{\cdot i} = x_{\cdot i}/n, i = 1, 2$.

While on the surface the kappa statistic is intuitively appealing as a measure of reliability, paradoxical results can arise from computing the kappa statistic for tables of various configurations. The most often cited paradox with kappa is termed the "base rate" or "prevalence" problem and refers to the fact that for a fixed $\hat{p}_0$, values of $\hat{p}_{1\cdot} \approx \hat{p}_{\cdot 1}$ away from 0.50 in either direction lead to smaller values of $\hat{\kappa}$. Thus, a diagnostic test with fixed **sensitivity** and **specificity** when applied twice to a sample of patients with $\hat{p}_{1\cdot} \approx \hat{p}_{\cdot 1} \approx 0.50$, will generate a kappa statistic larger than would be obtained from a similar application of the test in a very low-risk population (with $\hat{p}_{1\cdot} \approx \hat{p}_{\cdot 1} \approx 0.10$), or in an extremely high-risk population (with $\hat{p}_{1\cdot} \approx \hat{p}_{\cdot 1} \approx 0.90$). Given that the diagnostic instrument is the same in both studies, this result is argued to be counter-intuitive. This paradox also raises concerns about the utility of the ranges for the kappa statistic given by Landis & Koch [10] said to correspond to poor, fair, good, and excellent agreement. Owing to this dependence on the marginal frequencies, a comparison of reliability findings, as measured by $\hat{\kappa}$, is difficult across studies involving populations with different prevalences.

Table 1

	$T_2 = 1$	$T_2 = 2$	Total
$T_1 = 1$	x_{11}	x_{12}	$x_{1\cdot}$
$T_1 = 2$	x_{21}	x_{22}	$x_{2\cdot}$
Total	$x_{\cdot 1}$	$x_{\cdot 2}$	$x_{\cdot\cdot} = n$

At the population level, some insight can be gained into the reason for this behavior. Let θ denote the **prevalence** of the disease in the population from which the subjects under study were randomly sampled. Let α and β denote the **false positive** and **false negative** error rates for the diagnostic test, respectively. Then, if successive applications of the test may be assumed to be independent, $p_{11} = \theta(1-\beta)^2 + (1-\theta)(1-\alpha)^2$, $p_{12} = \theta(1-\beta)\beta + (1-\theta)(1-\alpha)\alpha$, $p_{21} = \theta(1-\beta)\beta + (1-\theta)(1-\alpha)\alpha$, $p_{22} = \theta\beta^2 + (1-\theta)\alpha^2$, $p_0 = p_{11} + p_{22} = \theta(1-2\beta) + (1-\theta)(1-2\alpha)$, and $p_e = 1 - 2a(1-a)$, where $a = [\theta(1-\beta) + (1-\theta)(1-\alpha)]^2$. Kraemer [8] derives the relation

$$\kappa = \frac{2\theta(1-\theta)(1-\alpha-\beta)^2}{2[(\theta(1-\beta)+(1-\theta)\alpha) \times (1-\theta(1-\beta)-(1-\theta)\alpha)]}. \qquad (1)$$

For fixed (α, β), plots of κ as a function of θ are concave down taking on the value zero at $\theta = 0$ and $\theta = 1$ [13].

Several solutions to this problem have been proposed ranging from supplementing $\hat{\kappa}$ with additional statistics to facilitate disentangling the nature of the

agreement, to entirely new approaches. Feinstein & Cicchetti [7] effectively illustrate the dependence of kappa on the marginal frequencies by considering several sample tables in which the observed raw agreement $\hat{p}_0$ is fixed, but the marginal frequencies vary. In a companion paper, Cicchetti & Feinstein [3] then propose that the kappa statistic should always be reported with two accompanying statistics called the index of average positive agreement, $\hat{q}_{\text{pos}} = 2x_{11}/(x_{1\cdot} + x_{\cdot 1})$, and the index of average negative agreement $\hat{q}_{\text{neg}} = 2x_{22}/(x_{2\cdot} + x_{\cdot 2})$. The motivation is that these statistics may be used to gain insight into the marginal agreement and imbalance in the marginal frequencies and hence allow interpretation of $\hat{\kappa}$ accordingly. Byrt et al. [1] propose using what they refer to as a bias-adjusted kappa, which reduces to an index previously proposed by Scott [12]. Lantz & Nebenzahl [11] suggest that kappa statistics be accompanied with statistics $\hat{\kappa}_{\min} = \hat{p}_0^2/[(1-\hat{p}_0)^2 + 1]$ and $\hat{\kappa}_{\max} = \hat{p}_0 - 1/\hat{p}_0 + 1$ for $\hat{p}_0 < 1$, which correspond to the minimum and maximum values of κ for a given level of observed agreement, and $\hat{\kappa}_{\text{nor}} = 2\hat{p}_0 - 1$, which is also the so-called prevalence-adjusted, bias-adjusted kappa statistic of Byrt et al. [1].

Much of the work on kappa has been carried out on intuitive, but largely ad hoc, grounds. For example, there is no underlying probability function, and hence **likelihood** function, for which κ is a sole parameter of interest. Rather, it has been proposed as an "index", a function of parameters which may be estimated and, when done so, is thought to have some attractive properties. Thus, it appears that the paradox arises since the kappa statistic is not model-based, and depends in a complicated way on the observed raw agreement and the marginal frequencies. The key factor in the paradoxes is the role of $\hat{p}_e$, which serves as a "correction factor" in the numerator of $\hat{\kappa}$, as well as a rescaling factor in the denominator $1 - \hat{p}_e$. The extent to which the lack of a likelihood function relates to the above prevalence problem is worthy of consideration.

Another difficulty is that it is not generally well understood precisely what is, or should be, meant by the "reliability" of a **binary** diagnostic test. With a view to exploring this, Cook & Farewell [5] describe a likelihood-based approach for the separate examination of the marginal agreement (relative magnitude of $p_{1\cdot}$ and $p_{\cdot 1}$) and subject-specific agreement (as measured by the **odds ratio**). Likelihood factorizations, conditioning arguments, and exact distributions facilitate detailed examination of well-defined and interpretable aspects of reliability.

In all of the recommended procedures cited above it must be borne in mind that the raw data for the 2×2 table under consideration consist only of four numbers and at most three **degrees of freedom**. Hence, presentation of three or four "summary" statistics does not serve the purpose of data reduction. Nevertheless, there appears to be general agreement that for the purpose of assessing reliability of a diagnostic test with a binary outcome, a single summary statistic is not adequate. The influence of the marginal frequencies on the kappa statistic is also present in the case of multiple nominal categories, but the precise nature of this influence is more difficult to characterize and is not well understood [10]. Chamberlin & Sprott [2] and Farewell and Sprott [6] derive a discrete conditional distribution which may be used as a basis for conditional inference (*see* **Conditionality Principle**) on subject-specific agreement in this context.

References

[1] Byrt, T., Bishop, J. & Carlin, J.B. (1993). Bias, prevalence and kappa, *Journal of Clinical Epidemiology* **46**, 423–424.

[2] Chamberlin, S.R. & Sprott, D.A. (1991). On a discrete distribution associated with the statistical assessment of nominal scale agreement, *Discrete Mathematics* **92**, 39–47.

[3] Cicchetti, D.V. & Feinstein, A.R. (1990). High agreement but low kappa. II. Resolving the paradoxes, *Journal of Clinical Epidemiology* **43**, 551–558.

[4] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[5] Cook, R.J. & Farewell, V.T. (1995). Conditional inference for subject-specific and marginal agreement: two families of agreement measures, *Canadian Journal of Statistics* **23**, 333–344.

[6] Farewell, V.T. & Sprott, D.A. (1999). Conditional inference for predictive agreement, *Statistics in Medicine* **18**, 1435–1449.

[7] Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low kappa. I. The problems of two paradoxes, *Journal of Clinical Epidemiology* **43**, 543–549.

[8] Kraemer, H.C. (1979). Ramifications of a population model for κ as a coefficient of reliability, *Psychometrika* **44**, 461–472.

[9] Kraemer, H.C. & Bloch, A.D. (1988). Kappa coefficients in epidemiology. An appraisal of a reappraisal, *Journal of Clinical Epidemiology* **41**, 959–968.

[10] Landis, R.J. & Koch, G.G. (1977). The measurement of observer agreement for categorical data, *Biometrics* **33**, 159–174.

[11] Lantz, C.A. & Nebenzahl, E. (1996). Behaviour and interpretation of the κ statistic: resolution of the two paradoxes, *Journal of Clinical Epidemiology* **49**, 431–434.

[12] Scott, W.A. (1955). Reliability and content analysis: the case of nominal scale coding, *Public Opinion Quarterly* **19**, 321–325.

[13] Thompson, W.D. & Walter, S.D. (1988). A reappraisal of the kappa coefficient, *Journal of Clinical Epidemiology* **41**, 949–958.

[14] Thompson, W.D. & Walter, S.D. (1988). Kappa and the concept of independent errors, *Journal of Clinical Epidemiology* **41**, 969–970.

(*See also* **Agreement, Measurement of**; **Observer Reliability and Agreement**)

RICHARD J. COOK

Kappa

In medical research it is frequently of interest to examine the extent to which results of a classification procedure concur in successive applications. For example, two psychiatrists may separately examine each member of a group of patients and categorize each one as psychotic, neurotic, suffering from a personality disorder, or healthy. Given the resulting data, questions may then be posed regarding the diagnoses of the two psychiatrists and their relationship to one another. The psychiatrists would typically be said to exhibit a high degree of agreement if a high percentage of their diagnoses concurred, and poor agreement if they often made different diagnoses. In general, this latter outcome could arise if the categories were ill-defined, the criteria for assessment were different for the two psychiatrists, or their ability to examine these criteria differed sufficiently, possibly as a result of different training or experience. Poor empirical agreement might therefore lead to a review of the category definitions and diagnostic criteria, or possibly retraining with a view to improving agreement and hence consistency of diagnoses and treatment. In another context, one might have data from successive applications of a test for dysplasia or cancer from cervical smears. If the test indicates normal, mild, moderate, or severe dysplasia, or cancer, and the test is applied at two time points in close proximity, ideally the results would be the same. Variation in the method and location of sampling as well as variation in laboratory procedures may, however, lead to different outcomes. In this context, one would say that there is empirical evidence that the test is reliable if the majority of the subjects are classified in the same way for both applications of the test. Unreliable tests would result from the sources of variation mentioned earlier. Again, empirical evidence of an unreliable test may lead to refinements of the testing procedure (*see* **Observer Reliability and Agreement**).

The Kappa Index of Reliability for a Binary Test

For convenience, consider a diagnostic testing procedure generating a binary response variable T indicating the presence ($T = 1$) or absence ($T = 2$) of a particular condition. Suppose this test is applied twice in succession to each subject in a sample of size n. Let T_k denote the outcome for the kth application with the resulting data summarized in the **two-by-two table** (Table 1). where x_{ij} denotes the frequency at which $T_1 = i$ and $T_2 = j$, $x_{i\cdot} = \sum_{j=1}^{2} x_{ij}$, and $x_{\cdot j} = \sum_{i=1}^{2} x_{ij}$, $i = 1, 2$, $j = 1, 2$. Assuming that

Table 1

	$T_2 = 1$	$T_2 = 2$	Total
$T_1 = 1$	x_{11}	x_{12}	$x_{1\cdot}$
$T_1 = 2$	x_{21}	x_{22}	$x_{2\cdot}$
Total	$x_{\cdot 1}$	$x_{\cdot 2}$	$x_{\cdot\cdot} = n$

test results on different subjects are independent, conditioning on n leads to a **multinomial distribution** for the outcome of a particular table with

$$f(\mathbf{x}; \mathbf{p}) = \binom{n}{x_{11}\ x_{12}\ x_{21}\ x_{22}} \prod_{i=1}^{2}\prod_{j=1}^{2} p_{ij}^{x_{ij}},$$

$\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})'$, $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})'$, and $p_{22} = 1 - p_{11} - p_{12} - p_{21}$. Let $p_{i\cdot} = \sum_{j=1}^{2} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{2} p_{ij}$. Knowledge of $\mathbf{p}$ would correspond to a complete understanding of the reliability of the test. Since knowledge of $\mathbf{p}$ is generally unattainable and estimation of $\mathbf{p}$ does not constitute a sufficient data reduction, indices of reliability/agreement typically focus on estimating one-dimensional functions of $\mathbf{p}$ (*see* **Agreement, Measurement of**).

A natural choice is $p_0 = \sum_{i=1}^{2} p_{ii}$, the probability of raw agreement, which is estimated as $\hat{p}_0 = \sum_{i=1}^{2} x_{ii}/n$. If $p_0 = 1$, then the test is completely reliable since the probability of observing discordant test results is zero. Similarly, if $\hat{p}_0$ is close to unity, then it suggests that the outcomes of the two applications concurred for the vast majority of the subjects. However, several authors have expressed reluctance to base **inferences** regarding reliability on the observed level of raw agreement (see [3] and references cited therein). The purported limitations of $\hat{p}_0$ as a measure of reliability stem from the fact that p_0 reflects both "chance" agreement and agreement over and above that which would be expected by chance. The agreement expected by chance, which we denote by p_e, is computed on the basis of the marginal distribution, defined by $p_{1\cdot}$ and $p_{\cdot 1}$, and under the assumption that the outcomes of the two

tests are independent conditional on the true status. Specifically, $p_e = \sum_{i=1}^{2} p_{i\cdot} p_{\cdot i}$ is estimated by $\hat{p}_e = \sum_{i=1}^{2} x_{1\cdot} x_{\cdot 1}/n^2$. To address concerns regarding the impact of nonnegligible chance agreement, Cohen [3] defined the index kappa which takes the form

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

and indicated that it can be interpreted as reflecting "the proportion of agreement *after* chance agreement is removed from consideration". This can be seen by noting that $p_0 - p_e$ is the difference in the proportion of raw agreement and the agreement expected by chance, this being the agreement arising due to factors not driven by chance. If $p_0 - p_e > 0$, then there is agreement arising from nonchance factors; if $p_0 - p_e = 0$, then there is no additional agreement over that which one would expect based on chance; and if $p_0 - p_e < 0$, then there is less agreement than one would expect by chance. Furthermore, $1 - p_e$ is interpreted by Cohen [3] as the proportion "of the units for which the hypothesis of no association would predict disagreement between the judges". Alternatively, this can be thought of as the maximum possible agreement beyond that expected by chance. An estimate of κ, denoted $\hat{\kappa}$, is referred to as the kappa statistic and may be obtained by replacing p_0 and p_e with their corresponding point estimates, giving

$$\hat{\kappa} = \frac{\hat{p}_0 - \hat{p}_e}{1 - \hat{p}_e}. \tag{1}$$

The Kappa Index of Reliability for Multiple Categories

When the classification procedure of interest has multiple nominal categories, assessment of agreement becomes somewhat more involved. Consider a diagnostic test with R possible outcomes and let T_k denote the outcome of the kth application of the test, $k = 1, 2$. Then T_k takes values on $\{1, 2, 3, \ldots, R\}$ and interest lies in assessing the extent to which these outcomes agree for $k = 1$ and $k = 2$. An $R \times R$ **contingency table** may then be constructed (see Table 2), where again x_{ij} denotes the frequency with which the first application of the test led to outcome i and the second led to outcome $j, i = 1, 2, \ldots, R, j = 1, 2, \ldots, R$. A category-specific measure of agreement may be of interest to examine the extent to which the two applications tend to lead to consistent conclusions with respect to outcome r, say. In this problem there is an implicit assumption that the particular nature of any disagreements are not of interest. One can then collapse the $R \times R$ table to a 2×2 table constructed by cross-classifying subjects with **binary** indicators such that $T_k = 1$ if outcome r was selected at the kth application, $T_k = 2$ otherwise, $k = 1, 2$. A category-specific kappa statistic can then

Table 2

	$T_2 = 1$	$T_2 = 2$	$T_2 = 3$	$\cdots$	$T_2 = R$	Total
$T_1 = 1$	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1R}	$x_{1\cdot}$
$T_1 = 2$	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2R}	$x_{2\cdot}$
$T_1 = 3$	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3R}	$x_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$T_1 = R$	x_{R1}	x_{R2}	x_{R3}	$\cdots$	x_{RR}	$x_{R\cdot}$
Total	$x_{\cdot 1}$	$x_{\cdot 2}$	$x_{\cdot 3}$	$\cdots$	$x_{\cdot R}$	$x_{\cdot\cdot} = n$

be constructed in the fashion indicated earlier. This can be repeated for each of the R categories giving R such statistics.

In addition to these category-specific measures, however, an overall summary index of agreement is often of interest. The kappa statistic in (1) is immediately generalized for the $R \times R$ ($R > 2$) table as follows. Let p_{ij} denote the probability of $T_1 = i$ and $T_2 = j$, one of the R^2 multinomial probabilities, $p_{i\cdot} = \sum_{j=1}^{R} p_{ij}$, and $p_{\cdot j} = \sum_{i=1}^{R} p_{ij}, i = 1, 2, \ldots, R, j = 1, 2, \ldots, R$. Then, as before, $\hat{p}_{ij} = x_{ij}/n$, $\hat{p}_{i\cdot} = x_{i\cdot}/n$, $\hat{p}_{\cdot j} = x_{\cdot j}/n$, $\hat{p}_0 = \sum_{i=1}^{R} \hat{p}_{ii}$, $\hat{p}_e = \sum_{i=1}^{R} \hat{p}_{i\cdot}\hat{p}_{\cdot i}$, and the overall kappa statistic takes the same form as in (1). This overall kappa statistic can equivalently be written as a weighted average of category-specific kappa statistics [6].

The kappa statistic has several properties that are widely considered to be attractive for measures of agreement. First, when the level of observed agreement, reflected by $\hat{p}_0$, is equal to the level of agreement expected by chance ($\hat{p}_e$), $\hat{\kappa} = 0$. Secondly, $\hat{\kappa}$ takes on its maximum value of 1 if and only if there is perfect agreement (i.e. $\hat{p}_0 = 1$ arising from a diagonal table). Thirdly, the kappa statistic is never less than -1. The latter two features require further elaboration, however, as the actual upper and lower limits on $\hat{\kappa}$ are functions of the marginal frequencies. In particular, $\hat{\kappa}$ takes on the value 1 only when the

marginal frequencies are exactly equal and all off-diagonal cells are zero. Values less than 1 occur when the marginal frequencies are the same but there are different category assignments in the table or, more generally, when the marginal frequencies differ (when the marginal frequencies differ there are necessarily nonzero diagonal cells and hence some disagreements). It is natural then to expect the kappa statistic for such a table to be less than unity. Cohen [3] shows that the maximum possible value of $\hat{\kappa}$ takes the form

$$\hat{\kappa}_{\mathrm{M}} = \frac{x_{\cdot\cdot}\sum_{i=1}^{R}\min(x_{i\cdot}, x_{\cdot i}) - \sum_{i=1}^{R}x_{i\cdot}x_{\cdot i}}{x_{\cdot\cdot}^2 - \sum_{i=1}^{R}x_{i\cdot}x_{\cdot i}}, \qquad (2)$$

and argues that this is intuitively reasonable since differences in the marginal frequencies necessarily lead to a reduction in the level of agreement and hence $\hat{\kappa}$. Cohen then suggests that if one is interested in assessing the proportion of the agreement permitted by the margins (correcting for chance), then one computes $\hat{\kappa}/\hat{\kappa}_{\mathrm{M}}$. We return to the topic of marginal frequencies and their influence on the properties of κ later in the article.

If the marginal frequencies for the two tests are uncorrelated (as measured by the product–moment **correlation** of the margins [3]), then the lower bound for $\hat{\kappa}$ is $\hat{\kappa}_{\mathrm{L}} = -(R-1)^{-1}$. When the marginal frequencies are negatively correlated, $\hat{\kappa}_{\mathrm{L}} > -(R-1)^{-1}$. However, when the marginal frequencies are positively correlated, $\hat{\kappa}_{\mathrm{L}} < -(R-1)^{-1}$. It is only as the number of categories reduces to two, the correlation of the marginal frequencies approaches 1, and the **variances** of the marginal frequencies increase, that $\hat{\kappa}_{\mathrm{L}}$ approaches -1 [3].

Having computed a kappa statistic for a given contingency table it is natural to want to characterize the level of agreement in descriptive terms. Landis & Koch [11] provide ranges that suggest, beyond what one would expect by chance, $0.75 < \hat{\kappa}$ typically represents excellent agreement, $0.40 < \hat{\kappa} < 0.75$ fair to good agreement, and $\hat{\kappa} < 0.40$ poor agreement. While there is some appeal to this convenient framework for the interpretation of $\hat{\kappa}$, caution is warranted (*see* **Kappa and its Dependence on Marginal Rates**).

Frequently, it will be of interest to construct **confidence intervals** for the index kappa. Fleiss et al. [8] derive an approximate large sample estimate for the variance of $\hat{\kappa}$, $\widehat{\mathrm{var}}(\hat{\kappa})$, as

$$\left(\sum_{i=1}^{R}\hat{p}_{ii}[1-(\hat{p}_{i\cdot}+\hat{p}_{\cdot i})(1-\hat{\kappa})]^2 + (1-\hat{\kappa})^2\sum_{i}\sum_{j\neq i}\hat{p}_{ij}(\hat{p}_{\cdot i}+\hat{p}_{j\cdot})^2 - [\hat{\kappa}-\hat{p}_{\mathrm{e}}(1-\hat{\kappa})]^2\right) \Big/ [x_{\cdot\cdot}(1-\hat{p}_{\mathrm{e}})^2], \qquad (3)$$

and Fleiss [6] recommends carrying out tests (*see* **Hypothesis Testing**) and constructing confidence intervals by assuming approximate **normality** of $(\hat{\kappa}-\kappa)/[\widehat{\mathrm{var}}(\hat{\kappa})]^{1/2}$ and proceeding in the standard fashion. For tests regarding the **null hypothesis** $\mathrm{H}_0 : \kappa = 0$, an alternate variance estimate may be derived from (3) by substituting 0 for $\hat{\kappa}$, and $\hat{p}_{i\cdot}\hat{p}_{\cdot j}$ for $\hat{p}_{ij}$, giving

$$\widehat{\mathrm{var}}_0(\hat{\kappa}) = \left(\sum_{k=1}^{R}\hat{p}_{i\cdot}\hat{p}_{\cdot i}[1-(\hat{p}_{i\cdot}+\hat{p}_{\cdot i})]^2 + \sum_{i\neq j}\hat{p}_{i\cdot}\hat{p}_{\cdot j} \times (\hat{p}_{\cdot i}+\hat{p}_{j\cdot})^2 - p_{\mathrm{e}}^2\right) \Big/ [x_{\cdot\cdot}(1-\hat{p}_{\mathrm{e}})^2], \qquad (4)$$

with tests carried out as described above.

The Weighted Kappa Index

The discussion thus far has focused on situations in which the test serves as a **nominal** classification procedure (e.g. as in the psychiatric diagnosis example at the beginning of the article). In such settings, since there is no natural ordering to the outcomes, any disagreements are often considered to be equally serious and the methods previously described are directly applicable. In some circumstances with nominal scales, however, certain types of disagreements are more serious then others and it is desirable to take this into account. Furthermore, when the outcome is ordinal (as in the cervical cancer screening example) (*see* **Ordered Categorical Data**), it is often of interest to adopt a measure of

agreement that treats disagreements in adjacent categories as less serious than disagreements in more disparate categories. For the test based on cervical smears designed to classify the condition of the cervix as healthy, mildly, moderately, or severely dysplastic, or cancerous, if on one occasion the test suggested mild dysplasia and on another moderate, this type of disagreement would be considered less serious than if a cervix previously diagnosed as cancerous was subsequently classified as mildly dysplastic. In general, the seriousness reflects clinical implications for treatment and the consequences of wrong decisions.

Weighted versions of the kappa statistic were derived by Cohen [4] to take into account the additional structure arising from ordinal measures or from nominal scales in which certain types of disagreement are of more importance than others. In particular, the objective of adopting a weighted kappa statistic is to allow "different kinds of disagreement" to be differentially weighted in the construction of the overall index. We begin by assigning a weight to each of the R^2 cells; let w_{ij} denote the weight for cell (i, j). These weights may be determined quite arbitrarily but it is natural to restrict $0 \leq w_{ij} \leq 1$, set w_{ii} to unity to give exact agreement maximum weight, and set $0 \leq w_{ij} < 1$ for $i \neq j$, so that all disagreements are given less weight than exact agreement. The selection of the weights plays a key role in the interpretation of the weighted kappa statistic and also impacts the corresponding variance estimates, prompting Cohen [4] to suggest these be specified prior to the collection of the data.

Perhaps the two most common sets of weights are the quadratic weights, with $w_{ij} = 1 - (i - j)^2/(R - 1)^2$, and the so-called Cicchetti weights, with $w_{ij} = 1 - |i - j|/(R - 1)$ [1, 2]. The quadratic weights tend to weight disagreements just off the main diagonal more highly than Cicchetti weights, and the relative weighting of disagreements farther from the main diagonal is also higher with the quadratic weights. Clearly, these two weighting schemes share the minimal requirements cited above. The weighted kappa statistic then takes the form

$$\hat{\kappa}^{(\mathrm{w})} = \frac{\hat{p}_0^{(\mathrm{w})} - \hat{p}_{\mathrm{e}}^{(\mathrm{w})}}{1 - \hat{p}_{\mathrm{e}}^{(\mathrm{w})}}, \tag{5}$$

where $\hat{p}_0^{(\mathrm{w})} = \sum_{i=1}^{R} \sum_{j=1}^{R} w_{ij} \hat{p}_{ij}$ and $\hat{p}_{\mathrm{e}}^{(\mathrm{w})} = \sum_{i=1}^{R} \sum_{j=1}^{R} w_{ij} \hat{p}_{i\cdot} \hat{p}_{\cdot j}$. If $\overline{w}_{i\cdot} = \sum_{j=1}^{R} \hat{p}_{\cdot j} w_{ij}$ and $\overline{w}_{\cdot j} = \sum_{i=1}^{R} \hat{p}_{i\cdot} w_{ij}$, then the large-sample variance of $\hat{\kappa}^{(\mathrm{w})}$ is estimated by

$$\begin{aligned} &\widehat{\mathrm{var}}(\hat{\kappa}^{(\mathrm{w})}) \\ &= \left(\sum_{i=1}^{R} \sum_{j=1}^{R} \hat{p}_{ij} [w_{ij} - (\overline{w}_{i\cdot} + \overline{w}_{\cdot j})(1 - \hat{\kappa}^{(\mathrm{w})})]^2 \right. \\ &\quad \left. - [\hat{\kappa}^{(\mathrm{w})} - \hat{p}_{\mathrm{e}}^{(\mathrm{w})}(1 - \hat{\kappa}^{(\mathrm{w})})]^2 \right) \Big/ [x_{\cdot\cdot}^2 (1 - \hat{p}_{\mathrm{e}}^{(\mathrm{w})})^2] \end{aligned} \tag{6}$$

and, as before, tests and confidence intervals may be carried out and derived in the standard fashion assuming asymptotic normality of the quantity $(\hat{\kappa}^{(\mathrm{w})} - \kappa^{(\mathrm{w})})/[\widehat{\mathrm{var}}(\hat{\kappa}^{(\mathrm{w})})]^{1/2}$. As in the unweighted case, a variance estimate appropriate for testing $\mathrm{H}_0 : \kappa^{(\mathrm{w})} = 0$ may be derived by substituting $\hat{p}_{i\cdot} \hat{p}_{\cdot j}$ for $\hat{p}_{ij}$, and 0 for $\hat{\kappa}^{(\mathrm{w})}$ in (6).

We note in passing that the weighted kappa with quadratic weights has been shown to bear connections to the intraclass correlation coefficient. Suppose that with an ordinal outcome the categories are assigned the integers 1 through R from the "lowest" to "highest" categories, respectively, and assignment to these categories is taken to correspond to a realization of the appropriate integer value. Fleiss & Cohen [7] show that the intraclass correlation coefficient computed by treating these integer responses as coming from a Gaussian **general linear model** for a two-way **analysis of variance**, is asymptotically equivalent to the weighted kappa statistic with quadratic weights.

The Kappa Index for Multiple Observers

Thus far we have restricted consideration to the case of two applications of the classification procedure (e.g. two successive applications of a diagnostic test, two physicians carrying out successive diagnoses, etc.). In many situations, however, there are multiple (>2) applications and interest lies in measuring agreement on the basis of several applications. Fleiss [5] considered the particular problem in which a group of subjects was examined and classified by a fixed number of observers, but where it was not necessarily the same set of observers carrying out the assessments for each patient. Moreover, Fleiss [5] assumed that it was not possible to identify which observers were involved in examining the patients.

For this problem, we require some new notation. Let M denote the number of subjects, N denote the number of observers per subject, and R denote the number of categories as before. Therefore, NM classifications are to be made. Let n_{ij} denote the number of times the ith subject was assigned to the jth category. A measure of overall raw agreement for the assignments on the ith subject is given by

$$\hat{q}_i = \frac{\sum_{j=1}^{R} n_{ij}(n_{ij}-1)}{N(N-1)},$$

which can be interpreted as follows. With N observers per subjects there are $\binom{N}{2}$ possible pairs of assignments. There are $\binom{n_{ij}}{2}$ which agree on category j and hence a total number of $\sum_{j=1}^{R}\binom{n_{ij}}{2}$ pairs of assignments which concur altogether for the ith subject. Thus, (7) simply represents the proportion of all paired assignments on the ith subject for which there was agreement on the category. The overall measure of raw observed agreement over all subjects is then given by $\hat{q}_0 = M^{-1}\sum_{i=1}^{M}\hat{q}_i$, which equals

$$\hat{q}_0 = \frac{\sum_{i=1}^{M}\sum_{j=1}^{R} n_{ij}^2}{MN(N-1)} - \frac{1}{N-1}. \tag{8}$$

As before, however, some agreement would be expected among the observers simply by chance and the kappa statistic in this setting corrects for this. The expected level of agreement is computed by noting that

$$\hat{p}_j = \frac{\sum_{i=1}^{M} n_{ij}}{MN}$$

is the sample proportion of all assignments made to category j, with $\sum_{j=1}^{R}\hat{p}_j = 1$. So if pairs of observers were simply assigning subjects to categories at random and independently one can estimate that they would be expected to agree according to

$$\hat{p}_e = \sum_{j=1}^{R}\hat{p}_j^2, \tag{9}$$

then the kappa statistic is computed by correcting for chance in the usual way as

$$\hat{\kappa} = \frac{\hat{q}_0 - \hat{p}_e}{1-\hat{p}_e}. \tag{10}$$

The sample variance for (10) is derived by Fleiss et al. [9] to be

$$\widehat{\mathrm{var}}(\hat{\kappa}) = 2\left[\left(\sum_{j=1}^{R} p_j(1-p_j)\right)^2 - \sum_{j=1}^{R} p_j(1-p_j)(1-2p_j)\right] \Big/ MN(N-1)\left(\sum_{j=1}^{R} p_j(1-p_j)\right)^2 \tag{11}$$

and is typically used for tests or interval estimation in the standard fashion.

When the same set of raters assesses all subjects and individual raters scores are known, it is not possible to use the results of Fleiss [5] without ignoring the rater-specific assignments. For this context, Schouten [13] proposed the use of indices based on weighted sums of pairwise measures of observed and expected levels of agreement. In particular, for a given pair of raters and a given pair of categories, observed and expected measures of agreement may be computed as earlier. Then, for each pair of raters, a measure of overall observed agreement may be obtained by taking a weighted average of such measures over all pairwise combinations of categories. Given a corresponding measure of expected agreement, an overall kappa statistic can be computed in the usual fashion. Schouten [13] then described how to obtain kappa statistics reflecting agreement over all observers, agreement between a particular observer and the remaining observers, and agreement within and between subgroups of observers.

General Remarks

MaClure & Willett [12] provide a comprehensive review and effectively highlight a number of limitations of the kappa statistics. In particular, they stress that for ordinal data derived from categorizing underlying continuous responses, the kappa statistic depends heavily on the often arbitrary category definitions, raising questions about interpretability. They also suggest that the use of weights, while attractive in allowing for varying degrees of disagreement, introduces another component of subjectivity into the computation of kappa statistics. Perhaps the issue of

greatest debate is the so-called prevalence, or base-rate, problem of kappa statistics (*see* **Kappa and its Dependence on Marginal Rates**). Several other authors have examined critically the properties and interpretation of kappa statistics [10, 14, 15], and the debate of the merits and demerits continues unabated. Despite the apparent limitations, the kappa statistic enjoys widespread use in the medical literature and has been the focus of considerable statistical research.

References

[1] Cicchetti, D.V. (1972). A new measure of agreement between rank ordered variables, *Proceedings of the American Psychological Association* **7**, 17–18.

[2] Cicchetti, D.V. & Allison T. (1973). Assessing the reliability of scoring EEG sleep records: an improved method, *Proceedings and Journal of the Electrophysiological Technologists' Association* **20**, 92–102.

[3] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[4] Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213–220.

[5] Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin* **76**, 378–382.

[6] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[7] Fleiss, J.L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurement* **33**, 613–619.

[8] Fleiss, J.L., Cohen, J. & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa, *Psychological Bulletin* **72**, 323–327.

[9] Fleiss, J.L., Nee, J.C.M. & Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters, *Psychological Bulletin* **86**, 974–977.

[10] Kraemer, H.C. & Bloch, D.A. (1988). Kappa coefficients in epidemiology: an appraisal of a reappraisal, *Journal of Clinical Epidemiology* **41**, 959–968.

[11] Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data, *Biometrics* **33**, 159–174.

[12] MaClure, M. & Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic, *American Journal of Epidemiology* **126**, 161–169.

[13] Schouten, H.J.A. (1982). Measuring pairwise interobserver agreement when all subjects are judged by the same observers, *Statistica Neerlandica* **36**, 45–61.

[14] Thompson, W.D. & Walter S.D. (1988). A reappraisal of the kappa coefficient, *Journal of Clinical Epidemiology* **41**, 949–958.

[15] Thompson, W.D. & Walter S.D. (1988). Kappa and the concept of independent errors, *Journal of Clinical Epidemiology* **41**, 969–970.

RICHARD J. COOK

Kempthorne, Oscar

Born: January 31, 1919, in St. Tudy, Cornwall, England.
Died: November 15, 2000, in Annapolis, Maryland.

Oscar Kempthorne made important contributions to both statistics and statistical genetics. His books *The Design and Analysis of Experiments* [10] and *An Introduction to Genetic Statistics* [15], both published in the 1950s, brought him early recognition and have become classics, still much cited. He also had a deep interest in statistical inference, especially **randomization** theory. Kempthorne was profoundly influenced and inspired by R.A. **Fisher's** writings. He was a great admirer of Fisher, but not an uncritical one.

Born on a farm in Cornwall, the young Kempthorne soon decided that farm work was not for him. He worked hard to win needed scholarships to Cambridge, teaching himself additional mathematics in a then remote and backward county.

During his three years at Cambridge he became a Wrangler (equivalent to a first class honors) in the Mathematical Tripos examinations, which emphasized pure and applied mathematics. His interest in statistics was aroused in lectures by John Wishart and J.O. **Irwin**. Upon graduation in May 1940, Kempthorne, who had been reserved for technical work in World War II, spent a term assisting Irwin on a drug assay project. After a "useless" six months in the Ministry of Supply he joined Rothamsted Experimental Station, then directed by Frank Yates. Kempthorne worked with the influential zoologist and military advisor Solly Zuckerman on operations research associated with the war effort, and began his research career with papers on sampling (*see* **Sample Surveys in the Health Sciences**) and **experimental design** [9].

In 1946, Kempthorne was appointed to an allied mission set up by the U. S. Department of State to observe the Greek parliamentary elections and a plebiscite on whether George II was to be retained as King of the Hellenes. Kempthorne was the British member of a group of statisticians that included W. Edwards Deming and Jerzy **Neyman**. The group was led by the sampling expert Ray Jessen of the Iowa State College Statistical Laboratory. Two reports were published (e.g. [8]) and Kempthorne's life was changed when he was offered an associate professorship at Iowa State. His appointment, made possible by W.G. **Cochran's** resignation, continued the Statistical Laboratory's connection with Rothamsted Experimental Station begun some 20 years earlier when George **Snedecor** was one of the first in the United States to recognize the importance of Fisher's work.

Oscar Kempthorne arrived in Ames in January 1947 and was a key faculty member for the next 42 1/2 years. He soon became involved in consulting with agricultural research workers and continued his research on experimental design begun at Rothamsted [18]. Also, while teaching this subject, he published in 1952, the 600-page masterly Wiley text [10] which, though unchanged for many years, has been highly influential. Apart from its comprehensive coverage, the book was the first statistics text to use **matrix algebra** intensively. In the year of publication, Kempthorne was elected Fellow of the **American Statistical Association**, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science (AAAS). A two-volume revision of the book is finally underway, in conjunction with Kempthorne's former student, Klaus Hinkelmann, who is preparing the second volume, the first [7] having appeared in 1994.

A special feature of [10] is the justification of the usual **analysis of variance F**-tests by their closeness to corresponding **randomization tests**. This point is elaborated in [13, 22], with due recognition of Fisher's famous illustration for the paired t-test [2] (*see* **Student's *t* Statistics**) and the pioneering work of Welch [28] and Pitman [24].

The design of experiments merged with Kempthorne's growing interest in genetic statistics in a series of papers on diallel crosses, especially when there were too many pure lines for all possible crosses to be tested [14, 17, 20]. Another way such merging occurred was via the transfer of ideas from **factorial experimentation** to genetics. Thus, if observations on a quantitative character are affected by many loci and there is random mating and independent assortment, the total effect of a **genotype** can be expressed as a sum of main effects and **interactions**. Main effects are those of individual alleles, two-way interactions between alleles are dominance deviations, if the alleles are at the same locus, and additive $\times$ additive effects if they are from different loci. The next term in the sum is an additive $\times$ dominance effect, which is a three-way interaction between an

allele at one locus and a pair of alleles at a second locus. Proceeding in this way, one is led to an expression for the total variance among genotypes as the sum of **variance components**, each of which is associated with main effects or particular interactions.

It is also the case that genetic covariances between relatives are expressible as linear combinations of the aforementioned variance components (*see* **Genetic Correlations and Covariances**). The coefficients are probabilities of identity by descent of genes or sets of genes, chosen appropriately to fit the degree of relationship between the relatives.

Fisher had, in 1918, expressed the sum of variance components, other than those associated with main effects of alleles and dominance deviations, as a single term, the epistatic variance. He had also derived expressions for covariances between relatives, involving the additive and dominance components of variance, in some special cases. Kempthorne [11, 12] generalized this work by partitioning the epistatic variance into genetically interpretable components associated with two-way and higher-order interactions, as described above. These terms also appeared in expressions for covariances between relatives. Another of his contributions in this regard was a clear and elegant notation, which easily allowed for multiple alleles at individual loci, rather than only two.

Kempthorne's research on covariances between relatives was summarized in his text [15] on genetic statistics. It was also described therein how an experimenter could estimate genetic variance components from mean squares in analysis of variance tables. Another feature of this book was the first presentation in a textbook of Sewall Wright's results on inbreeding theory [29] in terms of probabilities of **identity** by descent, rather than path coefficients (*see* **Path Analysis in Genetics**).

Kempthorne was also interested in "Fisher's Fundamental Theorem of Natural Selection". An attempt to interpret the cryptic description of this theorem in Fisher's *The Genetical Theory of Natural Selection* [3] is in his text. Further work on this subject is in three joint papers with E. Pollak [21, 25, 26], in which attempts were made to explicitly spell out mathematical consequences of definitions of fitness in populations that either have discrete generations or are age-structured. It was found, for example, in [21], that some of these definitions lead to different consequences than others in the writings of Fisher and other authors.

Another aspect of Kempthorne's thinking on genetics was his concern over misuses of quantitative genetic theory (*see* **Polygenic Inheritance**) by people who concluded that aid, such as "head start", to members of some socioeconomic or racial groups, is useless. Supporters of this view assert that the low average scores on IQ tests of people in these groups are largely due to genetic deficiencies. To support this assertion, they claim that the **heritability**, or the genetically transmissible fraction of the total variance, of the attribute IQ is high. The usual methods for estimating this fraction are indeed applicable to populations such as those of crops or livestock, which are under the control of experimenters. Kempthorne pointed out, however, that this approach is very questionable for populations of humans, who are not randomly assigned to environments [16, 23]. This is because the total variance among phenotypes contains, in this case, the covariance between genotypes and environments, as well as a variance component associated with variability among environments within genotypic groups.

These extra terms are eliminated by appropriately designed experiments with crops or livestock. Mathematical details are given by Emigh [1]. Kempthorne [16] also attacks the notion that data analysis can establish causation and "that one can establish effects of an intervention process when it does not occur".

Throughout his long career Kempthorne was struggling to understand the logic of theories of **inference**. The philosophers were ultimately a disappointment and Fisher, while much admired, remained obscure on many points. Perhaps the best summary of Kempthorne's views on Fisher's statistical writings is given in just two pages of [27]. He is deeply puzzled by the discrepancy between Fisher's strong advocacy of randomization, a procedure resting on satisfactory long-run behavior, and the complete lack of reference to long-run considerations in Fisher's theory of inference. Not that the Neyman–**Pearson** theory escapes questioning. In [19], a text for seniors and beginning graduate students, a discussion of tests of hypotheses (*see* **Hypothesis Testing**) concludes with the statement that apparently the choice of the size α of the test "has to be based on **decision theory**, with introduction of prior opinion and loss function".

A man of wide interests, Kempthorne was much sought after as a speaker. His provocative style was a further attraction. He was active in editorial work on several journals and was chief organizer and

proceedings editor of some major conferences. The offices held by Kempthorne included terms as president of the Biometric Society (*see* **International Biometric Society (IBS)**), Eastern North American Region, in 1961, as chairman, Section U (Statistics) of AAAS, in 1981, and as president of the Institute of Mathematical Statistics, 1984 to 1986.

For his researches, Kempthorne received the Sc.D. degree from Cambridge University in 1960. In 1965, he was elected to membership of the International Statistical Institute and in 1988, to Honorary Fellow of the **Royal Statistical Society**.

Kempthorne brought energy and flair to all his activities. He was a challenging teacher. Of the 42 PhD students he directed, 12 so far have become Fellows of the American Statistical Association.

For further information and insights, see the Festschrift [5], the interview [4], and the memorial article [6], the last containing a complete bibliography.

References

[1] Emigh, T.H. (1977). Partition of phenotypic variance under unknown dependent association of genotypes and environments, *Biometrics* **33**, 505–514.

[2] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[3] Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

[4] Folks, J.L. (1995). A conversation with Oscar Kempthorne, *Statistical Science* **10**, 321–336.

[5] Hinkelmann, K. ed. (1984). *Experimental Design, Statistical Models, and Genetic Statistics. Essays in Honor of Oscar Kempthorne*. Marcel Dekker, New York.

[6] Hinkelmann, K. (2001). Remembering Oscar Kempthorne (1919–2000), *Statistical Science* **16**, 169–183.

[7] Hinkelmann, K. & Kempthorne, O. (1994). *Design and Analysis of Experiments*. Vol. 1, *Introduction to Experimental Design*. John Wiley & Sons, New York.

[8] Jessen, R.J., Blythe, R.H., Kempthorne, O. & Deming, W.E. (1947). On a population sample for Greece, *Journal of the American Statistical Association* **42**, 357–384.

[9] Kempthorne, O. (1947). A simple approach to confounding and fractional replication in factorial experiments, *Biometrika* **34**, 255–272.

[10] Kempthorne, O. (1952). *The Design and Analysis of Experiments*. John Wiley & Sons, New York.

[11] Kempthorne, O. (1954). The correlation between relatives in a random mating population, *Proceedings of the Royal Society of London, Series B* **143**, 102–113.

[12] Kempthorne, O. (1955a). The theoretical values of correlations between relatives in random mating populations, *Genetics* **40**, 153–167.

[13] Kempthorne, O. (1955b). The randomization theory of experimental inference, *Journal of the American Statistical Association* **50**, 946–967.

[14] Kempthorne, O. (1956). The theory of the diallel cross, *Genetics* **41**, 451–459.

[15] Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. John Wiley & Sons, New York.

[16] Kempthorne, O. (1978). Logical, epistemological and statistical aspects of nature-nurture data interpretation, a Biometrics invited paper, *Biometrics* **34**, 1–23.

[17] Kempthorne, O. & Curnow, R.N. (1961). The partial diallel cross, *Biometrics* **17**, 229–250.

[18] Kempthorne, O. & Federer, W.T. (1948). The general theory of prime-power lattice designs: I. Introduction and designs for p^n varieties in blocks of p plots, *Biometrics* **4**, 54–79.

[19] Kempthorne, O. & Folks, L. (1971). *Probability, Statistics, and Data Analysis*. Iowa State University Press, Ames.

[20] Kempthorne, O. & Hinkelmann, K. (1963). Two classes of group divisible partial diallel crosses, *Biometrika* **50**, 281–291.

[21] Kempthorne, O. & Pollak, E. (1970). Concepts of fitness in Mendelian populations, *Genetics* **64**, 125–145.

[22] Kempthorne, O. & Wilk, M.B. (1955). Fixed, mixed, and random models, *Journal of the American Statistical Association* **50**, 1144–1167.

[23] Kempthorne, O. & Wolins, L. (1982). Testing reveals a big social problem, *The Behavioral and Brain Sciences* **5**, 327–336.

[24] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any populations: III. The analysis of variance test, *Biometrika* **29**, 322–335.

[25] Pollak, E. & Kempthorne, O. (1970). Malthusian parameters in genetic populations: Part I. Haploid and selfing models, *Theoretical Population Biology* **1**, 315–345.

[26] Pollak, E. & Kempthorne, O. (1971). Malthusian parameters in genetic populations: Part II. Random mating populations in infinite habitats, *Theoretical Population Biology* **2**, 357–390.

[27] Savage, L.J. (1976). On rereading R.A. Fisher (with discussion), *Annals of Statistics* **4**, 441–500.

[28] Welch, B.L. (1937). On the z-test in randomized blocks and Latin squares, *Biometrika* **29**, 21–52.

[29] Wright, S. (1921). Systems of mating, *Genetics* **61**, 111–178.

HERBERT A. DAVID & EDWARD POLLAK

Kendall, Maurice George

Born: September 6, 1907, in Kettering, UK.
Died: March 29, 1983, in Redhill, UK.

Reproduced by permission of the Royal Statistical Society

Despite showing only a belated interest in mathematics at school, Maurice Kendall obtained a scholarship to read mathematics at St John's College in Cambridge. He played cricket for his college and was a keen chess player, and gained a first class in both parts of the mathematical tripos.

After graduating, he entered the administrative class of the Civil Service. Here, at the Ministry of Agriculture and Fisheries, he was responsible for statistical work. A chance meeting with **G. Udny Yule** in 1935 led to Kendall becoming co-author of a revision of Yule's classic textbook [9]. In 1941 Kendall became statistician to the British Chamber of Shipping, and in the following years he published many papers on theoretical statistics. This work was wide-ranging, but major themes were the theory of **rank correlation** coefficients (one of which he discovered in 1938 and now bears his name), **paired comparison** experiments, k statistics, and **time series**. At the same time, Kendall was working on his *Advanced Theory of Statistics*, the first advanced textbook on the subject. This was published as two volumes in 1943 and 1946 [1].

He was appointed professor of statistics at the London School of Economics in 1949, where he founded a research techniques division which carried out large sample surveys. In addition to this work and further theoretical research, in this period Kendall published the first important dictionary of statistical terms [4] and worked on the first comprehensive bibliography of statistical literature [5].

In 1961, during his presidency of the **Royal Statistical Society**, he again changed career, becoming scientific director (and ultimately chairman) of a computer consultancy (later called SCICON). During this spell he completed the rewriting of his influential book [1] into three volumes [6].

On retiring in 1972, Kendall embarked on another, testing career as the first director of the World Fertility Survey. This was a huge multinational sample survey project, which fully tested his extraordinary organizational powers. Ill health forced his retirement from this position in 1980.

Kendall was a prolific author, producing 17 books and around 75 papers on theoretical statistics alone. Seventeen of his papers are reprinted in [7], which also contains a bibliography. His other books included [2] and [3].

His interest in language was demonstrated by his literary style – "lucid, balanced, often ironical" [8] – but also by the word play in the spoof story of Lamia Gurdleneck and Sara Nuttal by K.A.C. Manderville in Volume 2 of *The Advanced Theory of Statistics* [6], in which all the names are anagrams of either Maurice (G.) Kendall or Alan Stuart, and his Longfellow pastiche *Hiawatha Designs an Experiment* (reprinted in [1]).

Kendall was much honored. He received the Guy medal in gold from the Royal Statistical Society. In 1974 he was knighted for his services to statistics, and on retiring from the World Fertility Survey he was awarded the United Nations peace medal.

References

[1] Kendall, M.G. (1943 & 1946). *The Advanced Theory of Statistics*, Vols. 1 & 2. Griffin, London.

[2] Kendall, M.G. (1948). *Rank Correlation Methods*. Griffin, London.

[3] Kendall, M.G. (1975). *Multivariate Analysis*. Griffin, London.

[4] Kendall, M.G. & Buckland, W.R. (1957). *Dictionary of Statistical Terms*. Oliver & Boyd, Edinburgh.

[5] Kendall, M.G. & Doig, A.G. (1962, 1965, 1968). *Bibliography of Statistical Literature*, Vols. 1–3. Oliver & Boyd, Edinburgh.

[6] Kendall, M.G. & Stuart, A. (1958, 1961, 1966). *The Advanced Theory of Statistics*, Vols. 1–3. Griffin, London.

[7] Stuart, A. ed. (1984). *Statistics Theory and Practice. Selected Papers by Maurice Kendall* (1907–1983) Griffin, High Wycombe.

[8] Stuart, A. (1984). Obituary of Sir Maurice Kendall, *Journal of the Royal Statistical Society, Series A* **147**(,), 120–122.

[9] Yule, G.U. & Kendall, M.G. (1937). *An Introduction to the Theory of Statistics*, 11th Ed. Griffin, London.

DOUGLAS G. ALTMAN

Kin–Cohort Studies

The basic idea behind a kin–cohort design is that one can estimate **penetrance** by genotyping a set of unrelated individuals and obtaining phenotype information about their relatives [16, 18]. In an important, albeit narrow set of circumstances, a kin–cohort analysis has some distinct advantages over more conventional applications of **segregation analysis** and epidemiologic designs used to estimate the penetrance of a **genotype** for its phenotype.

A kin–cohort analysis is characterized by an unusual feature [18]. The phenotypes of the individuals who are genotyped (we will call them the volunteers) are not used directly for estimating the penetrance because of the difficulty in determining the pattern of **ascertainment**; instead, the penetrance is estimated from the phenotypes of the volunteers' relatives, whose genotypes are not determined directly, but are inferred from the genotypes of the volunteers.

These characteristics give the kin–cohort approach its notable strengths and its notable weaknesses. A kin–cohort study can be implemented quickly and economically. Because the kin–cohort procedure does not (necessarily) focus on members of high-risk families, the penetrance estimate may be applicable to those who carry the **mutation**, regardless of whether they have extensive family history. The ability to study multiple phenotypes simultaneously is a further advantage; the range of questions that can be addressed from a kin–cohort study is demonstrated below. A major disadvantage is that there must be a way to determine the phenotypes of the relatives of the genotyped individuals accurately; this is easier to accomplish where there is a complete disease registry against which relatives can be checked [17] or where there is little stigma attached to a disease phenotype so that each relative is likely to know about another's diagnosis. A second major disadvantage arises from reliance on volunteers who may be more likely to participate if they themselves or members of their families have been affected by the disease being studied.

The kin–cohort method can exploit the existence of an identifiable population with a higher than average frequency of the at-risk genotype. For example, the prevalence in the US of carriers of any of the BRCA1 and BRCA2 alleles (*see* **Gene**) that confer excess risk of breast and ovary cancer is probably less than 0.5%, while the frequency of three BRCA1 and BRCA2 founder mutations in Ashkenazi Jews, an endogamous ethnic group descended from a small number of founders, is close to 2% [14]. A founder population (*see* **Founder Effect**) has the additional advantage of not needing to sequence the entire gene in each individual to search for mutations [15].

If volunteering is independent of genotype, conditional on phenotype, then the disease history of the volunteers can be incorporated into the estimate of penetrance [6]. This assumption would be violated if known carriers were more or less likely to volunteer or if survival after diagnosis depended on genotype; the survival of the relatives after diagnosis can be related to mutation status without causing bias because their disease phenotype is reported by the volunteer. In extensions of the original approach, phenotypes of volunteers can be used indirectly for estimating penetrance under some assumptions [6].

Origin

The special characteristics of the kin–cohort design are best considered in the context of the Washington Ashkenazi Study (WAS), for which it was first conceived and implemented. Earlier, Struewing et al. [15] had discovered the 185delAG mutation in BRCA1 in Ashkenazi breast/ovary cancer families; they subsequently noted that the mutation had a high enough frequency [14] in Jews to make a study of its effects worthwhile.

Jeff Struewing, Patricia Hartge, Larry Brody, Margaret Tucker, and Sholom Wacholder of the National Institutes of Health, Bethesda, MD, planned a cross-sectional study of Ashkenazi Jewish volunteers in the Washington, DC, area. The investigators noted that a study of allele **prevalence** could be used to estimate penetrance in a population without the high levels of family history seen in the Breast Cancer Linkage Consortium (BCLC) families used to derive the earlier breast cancer penetrance estimate of 85% by age 70 [4]. They recognized that if BRCA1 was 100% penetrant, over 50% of the carriers' mothers would be affected eventually, but that if the mutation were unrelated to risk, the carriers' and noncarriers' mothers should show little difference in risk, so there must be information about penetrance in the data from the study. Therefore, they determined that the main goal

of the study would be to estimate the penetrance of the 185delAG BRCA1 carriers for breast cancer in a setting without large numbers of multiply affected families. This study could address a key question: Would the penetrance estimate in carrier women from less loaded family families be as high as the 85% estimated by the BCLC? After the study began, the investigators added genotyping of two other newly identified founder mutations to the focus of the study.

The investigators developed a simple method-of-moments argument to develop an estimator of penetrance that could be applied to a study of volunteers from whom information about their relatives' years of birth and death and diagnosis of disease, if applicable, was obtained [18]. The main ideas behind the estimator are:

1. The disease experience of the volunteers themselves should not be used directly to estimate penetrance if the disease is often fatal. There would be a substantial number of individuals in the study population who had developed disease and had died and thus would not contribute to the count of disease. Even if there is no mortality from disease, affected individuals may be more likely to volunteer, leading to upward bias in the estimation of penetrance. Estimation of **relative risk** may be **unbiased**, however, unless survival after diagnosis or participation depends on genotype. The distribution of genotypes of the volunteers is used indirectly in the estimation of penetrance [16, 18]. When the **likelihood** approach of Gail et al. [6] incorporates the phenotypes of the volunteers, the ascertainment based on disease status does not lead to bias if the genotype is unrelated to volunteering, conditional on phenotype, and to time of survival after diagnosis.
2. There is information about the genotype of relatives from the genotypes of the volunteers. Simple rules of Mendelian inheritance (*see* **Mendel's Laws**) can be used to infer the relative's genotype if the mode of inheritance and the allele frequency are known. Essentially, since the allele frequency is low (and could be estimated from the study; *see* **Gene Frequency Estimation**), slightly more than half of the first-degree relatives of mutation carriers are themselves carriers, while only a small fraction (close to half of the carrier frequency) of first-degree relatives of noncarriers are themselves carriers (*see* **Genetic Counseling**). Thus, the **cumulative incidence** function for disease (time from birth until disease) among the relatives of carriers and among the relatives of noncarriers is each weighted average of the survival functions in carriers and noncarriers, but with different weights. The weights themselves are proportional to the probabilities of the relative's genotype, conditional on the volunteer's (known) genotype, and can be derived by Mendelian principles if the mode of inheritance is known (*see* **Segregation Analysis, Classical**) [18]. Because the mutation is rare, solving for two equations in two unknowns at any specified age gives an estimate of the survival functions for carriers and noncarriers from the allele frequency estimate and the survival functions of the relatives of carriers and noncarriers. A bootstrap, with resampling based on family, can be used to estimate the **variance** or point-wise **confidence intervals** (*see* **Bootstrap Method**) [16, 18].

The name "kin–cohort" derives from the use of cohorts of relatives of the volunteers to estimate penetrance.

Several assumptions need to be made to operationalize the plan; they might not be perfectly satisfied in a study that was actually feasible.

1. The volunteers must have the same level of family history and the same allele frequency as a random sample. The investigators in the WAS were not convinced that they could choose a random sample of the Jewish population in an American setting economically. Instead, they relied on volunteers; they took advantage of the community interest in breast cancer. This had the unfortunate consequence, most likely, that the penetrance estimate was too high, since those with a family history of breast cancer were more likely to volunteer than those without; the extent of the **bias** may have been mitigated to some extent by the concern in the community about breast cancer, even in those without a personal or family history.
2. Reporting of year of birth, and year of death and diagnosis of disease, if applicable, must be accurate and complete for each relative. Since the population was well-educated and breast

cancer seems not to stigmatize those affected, the authors felt that most of the reports on first-degree relatives would be complete. Reporting of ovarian cancer, particularly in parents, seemed more difficult because the exact site of a mother's reproductive-tract-organ cancer may not have been known to children.

3. The penetrance from the mutations found in the study population must be the same as the average penetrance over all mutations in the population to which one wishes to extrapolate. If the Ashkenazi founder mutations studied in the WAS are more penetrant than other mutations, or if the risk of breast cancer for any mutation among the Jewish women were higher than among non-Jewish women, the estimate from the WAS would be biased upwards *vis-à-vis* women with other mutations or non-Jewish women.
4. The distribution of any common environmental or genetic factor that is an important modifier of risk of disease and can affect penetrance estimates must be the same in the study population and in the population to which the estimates are to be applied (*see* **Gene-environment Interaction**).
5. It is the allele frequency at conception that needs to be estimated for the Mendelian arithmetic to be accurate. The estimates of penetrance may be biased if the allele frequency changes with age, perhaps due to survival differences from the disease under study or other factors.
6. Like **linkage** and segregation (and conditional and unconditional **logistic regression**), confidence intervals and test statistics from kin–cohort data depend on the assumption of conditional (on genotype) independence of the disease within families. In reports from the WAS [16, 18] the authors used a bootstrap approach where the sampling unit was the family to address this problem. Chatterjee & Wacholder [2] proposed a **marginal likelihood** approach to remove, or at least reduce, the bias in penetrance estimation due to residual **familial correlation**.
7. To estimate the survival curves for the relatives of noncarriers and carriers, one needs to assume that censoring does not depend on the unknown genotype of the relatives. The penetrance estimates, therefore, also depend on the assumption of no competing risks.

The features of the WAS carried out at the National Cancer Institute (NCI) were chosen by the investigators to be able to apply this method to estimate penetrance of three founder mutations [16]. Jewish volunteers were solicited through Jewish communal organizations, media publicity, and advertisements targeted at Jews. Each volunteer gave consent for genotyping and was asked about the year of birth and death and sites and dates of any cancer diagnoses in all first-degree relatives.

The main results of the investigation are reported in Struewing et al. [16]. Kaplan–Meier estimates of cumulative risk of breast cancer revealed a clear difference between breast cancer incidence in first-degree relatives of carriers and noncarriers. The method-of-moments estimator for cumulative incidence in carriers (penetrance) and noncarriers was substantially lower than the BCLC, particularly at older ages [4]. The authors noted that their estimate of breast cancer penetrance was probably too high because relatives of women with breast cancer were more likely to participate. But the important message from their paper was that the penetrance in a population without extreme family history is lower than in families with multiply affected individuals, even assuming correction for ascertainment. Indeed, other estimates based on women with less extensive family history [7, 17, 19] have been consistently below those from the BCLC [4]. The discrepancy may be due, in part, to the difficulty of correcting completely for ascertainment in a study with data from many collaborators; regardless of the minimal requirements to include a family, it seems reasonable that investigators tried increasingly hard to get families with increasing numbers of affected. Furthermore, and perhaps more important, modifier genes that segregate within families or environmental factors that modify risk and aggregate in families may lead to variation in penetrance even among different families with the same mutation; if so, a study that chose families on the basis of a high number of affected members would selectively pick families with higher risk, leading to a higher estimate of penetrance than a more population-based study [18].

The methods-of-equations estimator used in the early reports [16, 18] has several defects, notably the possibility, realized in the original report, that the estimate of the survival function might not be monotone. Nevertheless, this seed concept opened a floodgate of scientific and methodologic investigation.

Methodologic Extensions

Gail et al. considered a **likelihood** framework for the estimation of penetrance and studied the sample size needed for the kin–cohort design, or as they called it, a "genotyped proband design" [6]. They also considered incorporating the volunteer's disease history data into the analysis after accounting for possible differential participation by their disease status. In the likelihood approach, the information on penetrance from the use of the phenotypes of genotyped individuals is the risk ratio for the effect of the genotype. It therefore requires the same assumptions as in a cross-sectional or, at least, a **case–control** study; in particular, because the study uses prevalent cases, survival after diagnosis of disease cannot be related to genotype.

The likelihood formulation [6], which avoids the possibility of nonincreasing penetrance estimates, forms the basis for subsequent theoretical work. Moore et al. [10] considered piecewise exponential models for carriers' and noncarriers' survival curves. They found that optimization of the likelihood becomes difficult due to the complex nature of the volunteers' contribution to the likelihood, particularly for a large number of hazard intervals. They developed a **pseudo-likelihood** approach that iterates between estimating the penetrance parameters from the relatives' contributions to the likelihood and estimating the allele frequency from the volunteers' contributions to the likelihood. They did not use the phenotypes of the volunteers directly in estimating penetrance. Chatterjee & Wacholder [2] developed a marginal likelihood that treats each relative of the volunteer individually rather than jointly, as considered by Gail et al. [5]. This method had several advantages: (a) it enjoys flexibility of the likelihood approach and can correct for the monotonicity problems Wacholder et al.'s original method [16, 18] had; (b) it is computationally simpler and faster than the full likelihood approach; and (c) under the assumptions of a kin–cohort analysis listed above it produces an unbiased estimate of penetrance even if there are sources of familial aggregation of the disease other than the genes being studied (*see* **Genetic Correlations and Covariances**). A joint likelihood approach under any specific assumption of the residual correlation, although it could be slightly more efficient than the marginal likelihood approach, will produce a biased estimate of the penetrance when the assumed degree of residual correlation is wrong. However, a disadvantage of the marginal likelihood approach is that it is unclear how to account for complex ascertainment in this approach as can be done in a joint likelihood approach. Kaufman [8] took a traditional segregation analysis approach to the kin–cohort data. He used the families of 114 carrier volunteers as his pedigrees. Using the likelihood framework of Gail et al. [5] Carroll et al. developed a score test for the existence of residual familial aggregation that incorporates the volunteer's disease history data into the analysis [1]. Using a multivariate survival modeling approach, Chatterjee et al. [3] showed how one can quantify and estimate the residual familial correlation from the kin–cohort data.

Thus, a kin–cohort design can be seen as a minimalist approach to segregation analysis. In its extreme idealized form, the data available are the genotypes of a random sample of unrelated individuals and the phenotypes of their relatives. The method-of-moments analysis, with its comparison of the survival times to disease in the relatives of carriers and noncarriers, provides some additional insight beyond the segregation analysis.

Scientific Applications

Thorlacius et al. [17] was able to apply a kin–cohort design without relying on volunteers or reporting of relatives' phenotypes. Genotyping was performed from stored pathology tissue from breast cancer cases in Iceland. The family registry and cancer registry were linked to identify relatives and when and whether they developed cancer. Carriers of the 999delT founder mutation in BRCA2 were found to have a 37% penetrance by age 70; the lower penetrance estimate could be due to a lower penetrance from BRCA2 mutations than from BRCA1 generally [13], a founder mutation less penetrant than others, or to the absence of bias from volunteering and self-report.

Several other questions – clinical and etiologic – can be addressed directly from a kin–cohort design or by extensions. Lee et al. [9] investigated survival after diagnosis using the WAS data. Woodage et al. [20] explored whether there was excess risk associated with a founder mutation in the APC gene among Ashkenazi Jews using the WAS data. Moslehi et al. [11] and Risch et al. [12] have estimated penetrance for ovarian cancer using kin–cohort as well.

Design Questions

Many questions about the design of studies using the kin–cohort method remain open. Gail et al. [5] examine the degree of bias from various violations of assumptions. The tables from Gail et al. [6] can be used to compare the numbers of genotypes and numbers of individuals needed to estimate penetrance with a given precision; they also consider a variant design where some of the volunteers' relatives are also genotyped. However, in using these tables, one must consider that a special population with a high allele frequency is an ideal setting for kin–cohort but may not be practicable for a cohort or **case–control** study; also, the tables ignore residual familial aggregation, as they note [6], and use phenotypes of the genotyped individuals to estimate penetrance, which requires the attendant assumptions noted above.

A kin–cohort analysis of subjects in a case–control or cohort study may be feasible if genotyping and the collection of family histories have already been completed. The estimate of penetrance based entirely or extensively on affected cases is likely to be slightly higher than one based on controls only or on a random sample if there is any residual familial aggregation of risk, because families with more affected will be selectively included [18].

References

[1] Carroll, R.J., Gail, M.H., Benichou, J. & Pee, D. (2000). Score tests for familial correlation in genotyped-proband designs, *Genetic Epidemiology* **18**, 293–306.

[2] Chatterjee, N. & Wacholder, S. (2001). A marginal likelihood approach for estimating penetrance from kin-cohort designs, *Biometrics* **57**, 245–252.

[3] Chatterjee, N., Shih, J., Hartge, P., Brody, L., Tucker, M. & Wacholder, S. (2001). Association and aggregation analysis using kin-cohort designs with applications to genotype and family history data from the Washington Ashkenazi Study, *Genetic Epidemiology* **21**, 123–138.

[4] Easton, D.F., Ford, D. & Bishop, D.T. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium, *American Journal of Human Genetics* **56**, 265–271.

[5] Gail, M.H., Pee, D. & Carroll, R. (1999). Kin-cohort designs for gene characterization, *Journal of the National Cancer Institute Monographs* 55–60.

[6] Gail, M.H., Pee, D., Benichou, J. & Carroll, R. (1999). Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped-proband designs, *Genetic Epidemiology* **16**, 15–39.

[7] Hopper, J.L., Southey, M.C., Dite, G.S., Jolley, D.J., Giles, G.G., McCredie, M.R. et al. (1999). Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. Australian Breast Cancer Family Study, *Cancer Epidemiology, Biomarkers and Prevention* **8**, 741–747.

[8] Kaufman, D. (2001). Penetrance of BRCA1 and BRCA2 mutations among Ashkenazi Jews: validation of the kin-cohort method of estimation using segregation analysis, Unpublished Doctoral Thesis. Johns Hopkins University School of Public Health, Baltimore.

[9] Lee, J.S., Wacholder, S., Struewing, J.P., McAdams, M., Pee, D., Brody, L.C. et al. (1999). Survival after breast cancer in Ashkenazi Jewish BRCA1 and BRCA2 mutation carriers, *Journal of the National Cancer Institute* **91**, 259–263.

[10] Moore, D.F., Chatterjee, N., Pee, D. & Gail, M.H. (2001). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study, *Genetic Epidemiology* **20**, 210–227.

[11] Moslehi, R., Chu, W., Karlan, B., Fishman, D., Risch, H., Fields, A. et al. (2000). BRCA1 and BRCA2 mutation analysis of 208 Ashkenazi Jewish women with ovarian cancer, *American Journal of Human Genetics* **66**, 1259–1272.

[12] Risch, H.A., McLaughlin, J.R., Cole, D.E., Rosen, B., Bradley, L., Kwan, E. et al. (2001). Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer, *American Journal of Human Genetics* **68**, 700–710.

[13] Satagopan, J.M., Offit, K., Foulkes, W., Robson, M.E., Wacholder, S., Eng, C.M. et al. (2001). The lifetime risks of breast cancer in Ashkenazi Jewish carriers of brca1 and brca2 mutations, *Cancer Epidemiology, Biomarkers and Prevention* **10**, 467–473.

[14] Struewing, J.P., Abeliovich, D., Peretz, T., Avishai, N., Kaback, M.M., Collins, F.S. et al. (1995). The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals, *Nature Genetics* **11**, 198–200.

[15] Struewing, J.P., Brody, L.C., Erdos, M.R., Kase, R.G., Giambarresi, T.R., Smith, S.A. et al. (1995). Detection of eight BRCA1 mutations in 10 breast/ovarian cancer families, including 1 family with male breast cancer, *American Journal of Human Genetics* **57**, 1–7.

[16] Struewing, J.P., Hartge, P., Wacholder, S., Baker, S.M., Berlin, M., McAdams, M. et al. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews, *New England Journal of Medicine* **336**, 1401–1408.

[17] Thorlacius, S., Struewing, J.P., Hartge, P., Olafsdottir, G.H., Sigvaldason, H., Tryggvadottir, L. et al.

(1998). Population-based study of risk of breast cancer in carriers of BRCA2 mutation, *Lancet* **352**, 1337–1339.

[18] Wacholder, S., Hartge, P., Struewing, J.P., Pee, D., McAdams, M., Brody, L. et al. (1998). The kin-cohort study for estimating penetrance, *American Journal of Epidemiology* **148**, 623–630.

[19] Whittemore, A.S., Gong, G. & Itnyre, J. (1997). Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer: results from three U.S. population-based case-control studies of ovarian cancer, *American Journal of Human Genetics* **60**, 496–504.

[20] Woodage, T., King, S.M., Wacholder, S., Hartge, P., Struewing, J.P., McAdams, M. et al. (1998). The APCI1307K allele and cancer risk in a community-based study of Ashkenazi Jews, *Nature Genetics* **20**, 62–65.

(*See also* **Population Genetics**)

SHOLOM WACHOLDER &
NILANJAN CHATTERJEE

Kolmogorov, Andrey Nikolayevich

Born: April 23, 1903, in Tambov, Russia.
Died: October 20, 1987, in Moscow, Russia.

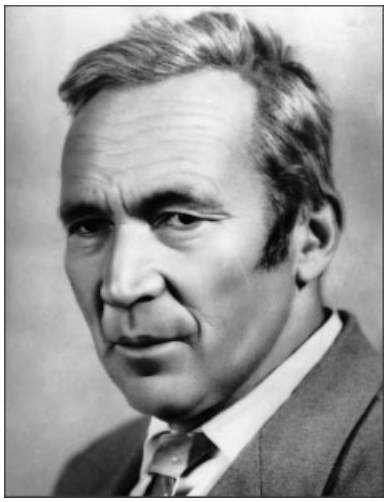

Reproduced by permission of the Royal Statistical Society

Kolmogorov is widely considered to be one of the greatest mathematicians of the twentieth century. He made important contributions to the theory of functions, topology, **probability theory**, statistics, logic, theory of dynamic systems, information theory, ergodic theory, theory of **algorithms**, mathematical education, and various applications of the above fields of mathematics. His works were mainly concerned with the intermediate areas between several "traditional" branches of mathematics and its applications, and he used fresh and striking ideas illuminating the relations between them. In the field of probability theory and statistics, Kolmogorov's main achievement is perhaps the introduction of the axioms and the clarification, through a rigorous approach, of various basic concepts. He was also famous as the leader of a school of numerous researchers, mainly his students and associates from the former USSR and Eastern block countries, whose work shaped probability theory (and to a lesser extent mathematical statistics) through the 1950s and 1960s.

During his career he held important administrative posts in the Moscow State (Lomonossov) University (MSU) and the USSR Academy of Sciences, including the headship of the Mechanics and Mathematics Department of the MSU, the Laboratory of Statistical Methods of MSU, and the chairmanship of the Mathematics Section of the Academy. Kolmogorov's personality had a great impact on everyone who came into contact with him, and in particular on hundreds of pupils at the specialist mathematical school for gifted children gathered from around the former Soviet Union, which he ran from the 1960s to the early 1980s.

In mathematical statistics, Kolmogorov is acclaimed worldwide for introducing the so-called **Kolmogorov–Smirnov statistic**. Based on this statistic (and its modifications), the Kolmogorov–Smirnov type tests of **goodness of fit** have been developed, which are among the most widely used in statistical practice. The original references are [5], [8], [12], and [22–24], a detailed account of work done before 1970 can be found in [3], and further developments are commented on in [7].

In Soviet statistics, he is also considered as a founder of the modern approach to the **least squares** method, and his papers [13, 18] are widely quoted. In the West these papers became known much later (see, for example, [20] and [21]). The third direction stemming from Kolmogorov's theoretical work is related to **unbiased** estimators and their relation to **sufficient statistics** [15]. The first application of his approach was connected with industrial quality inspection and discussed in the (almost unobtainable) brochure [14]; it was further developed in [1] (for recent references, see [2]).

Kolmogorov was deeply interested in applied statistics, specifically in regard to the analysis of genetic experiments, turbulence, weather forecasting, analysis of geologic deposits, analysis of artillery fire precision (research conducted before and during the early part of World War II), and analysis of Russian poetry. Many of his ideas were later used in practical recommendations in various fields, including the Soviet nuclear and space programs, although he was never directly involved in any of these projects (unlike most of his contemporaries of a similar stature in the USSR, a fact of which he was rather proud). He was also a keen and original popularizer of statistics, notably through his articles for the *Great Soviet*

Encyclopedia. In biostatistics, he was active in commenting on statistical confirmation of **Mendel's law** of genetics (see below) as well as in introducing and developing various mathematical models. For example, in [19] a nonlinear equation was analyzed rigorously in detail, describing the spread of an "advantageous gene". A similar equation was simultaneously proposed by **Fisher** [4], who predicted the long-term behavior of its solution, but did not provide a formal proof; this was done in [19] and subsequent papers. The equations studied in [19] and [4] are now often called Fisher or Kolmogorov–Petrovsky–Piskunov equations (another frequently used name is reaction–diffusion equations); their popularity in combustion theory far surpassed that in the analysis of biologic populations. Another notion connected with Kolmogorov's long-time interest in genetics was that of a **branching process**, introduced for the first time in [17], where the term "branching random processes" was first introduced (for related statistical considerations, see [9], [10], and [16]); again, the popularity of this concept in other applications exceeded that in the original field of theoretical biology.

Kolmogorov's participation in the discussion of the validity of Mendel's laws deserves a detailed account not only as being directly relevant to biostatistics, but also to illustrate the relation between statistics and "real life" at that time. The 1930s and the years following were a period of sharp struggle in Soviet biology. A group led by the infamous T.D. Lysenko started a ferocious campaign against Mendel's theory (and genetics in general) and its use in practice. Capitalizing on the support by Soviet officialdom, the followers of Lysenko denounced genetics as a "bourgeois pseudo-science", useless (or even harmful) for socialism and the future communist society. Their campaigning created an atmosphere of hysteria (matched by the general fear of repression of the period); as a result, many Soviet geneticists lost their jobs and some their lives. Kolmogorov had friendly ties with many of the leading USSR geneticists and was deeply interested in their experiments. In 1939 a collaborator of Lysenko published the results of a series of experiments with plants claiming that they disproved Mendel's 3:1 law. Kolmogorov [11] analyzed her data and, by performing a straightforward **chi-square test**, discovered that the experiments actually confirmed the 3:1 law. Given the circumstances of the time, this was an extraordinarily bold step. The paper [11] provoked an angry reply by Lysenko and his cronies, but luckily it did not cause serious harm to Kolmogorov. [At the same time similar experiments were conducted by another researcher, a follower of Vavilov, the leader of Soviet genetics (by that time dismissed from his positions and replaced by Lysenko; soon after, Vavilov himself was arrested and later died in prison in inhumane conditions). When the results of this series were shown to Kolmogorov, he immediately spotted that the author had "doctored" the data to make them fit exactly to the theoretical curve. I thank Prof. V.M. Tikhomirov, from Moscow State University, for providing me with this episode.] For a detailed account, see, for example, [6] and [25].

This and other episodes served to deter Kolmogorov from further experimentation in the subject. However, he kept a deep interest in biostatistics: in the 1960s he organized the department of medical statistics in the Laboratory of Statistical Methods and took an active part in its work. His ideas are widely used in medical statistical practice, mostly in Russia, but their detailed account still awaits publication.

References

[1] Belyaev, Yu.K. (1975). *Probability Methods of Sample Control*. Nauka, Moscow (in Russian).

[2] Belyaev, Yu.K. & Lumel'skii, Ya.P. (1992). Unbiased estimators, in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, pp. 585–587.

[3] Durbin, J. (1973). Distribution theory for tests based on the sample distribution function, in *Regional Conference Series in Applied Mathematics*, Vol. 9, SIAM, Philadelphia.

[4] Fisher, R.A. (1937). The wave of advance of advantageous genes, *Annals of Eugenics* **7**, 355–369.

[5] Glivenko, V. (1933). On the empirical determination of a probability law, *Giornale dell'Istituto Italiano degli Attuari* **4**, 92–99 (in Italian).

[6] Joravsky, D. (1986). *The Lysenko Affair*. University of Chicago Press, Chicago.

[7] Khmaladze, E.V. (1992). Empirical distribution, in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, pp. 574–582.

[8] Kolmogorov, A.N. (1933). On the empirical determination of a distribution law, *Giornale dell'Istituto Italiano degli Attuari* **4**, 83–91 (in Italian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 139–146.

[9] Kolmogorov, A.N. (1935). Deviations from Hardy's formulas under partial isolation, *Doklady Akademii Nauk SSSR* **3**, 129–132 (in Russian); English translation

in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 179–181.

[10] Kolmogorov, A.N. (1938). Solution of a biological problem, *Izvestiya NII Matematiki i Mekhaniki Tomskogo Universiteta* **2**(1), 7–12 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 216–221.

[11] Kolmogorov, A.N. (1940). On a new confirmation of Mendel's laws, *Doklady Akademii Nauk SSSR*. **27**, 38–42 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 222–227.

[12] Kolmogorov, A.N. (1941). Confidence limits for an unknown distribution function, *Annals of Mathematical Statistics* **12**, 461–463.

[13] Kolmogorov, A.N. (1946). Justification of the method of least squares, *Uspekhi Matematicheskih Nauk* **1**, 57–70 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 285–302.

[14] Kolmogorov, A.N. (1950). *Statistical Inspection Control When the Admissible Number of Defects is Zero*. Izdatel'stvo Doma Nauchno-Tehnicheskoi Propagandy, Leningrad (in Russian).

[15] Kolmogorov, A.N. (1950). Unbiased estimators, *Izvestiya Akademii Nauk SSSR* **14**, 303–326 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 369–394.

[16] Kolmogorov, A.N. (1959). Transition of branching processes to diffusion processes and related genetic problems, *Teoriya Veroyatnosteii i E e Primeneniya* **4**, 233–236 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 466–469.

[17] Kolmogorov, A.N. & Dmitriev, N.A. (1947). Branching random processes, *Doklady Akademii Nauk SSSR* **56**, 7–10 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, pp. 309–314.

[18] Kolmogorov, A.N., Petrov, A.A. & Smirnov, Yu.M. (1947). A formula of Gauss in the method of least squares, *Izvestiya Akademii Nauk SSSR Matematicheskaya* **11**, 561–566 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 2, A.N. Shiryayev, ed. Kluwer, Dordrecht, 1992, 303–308.

[19] Kolmogorov, A.N., Petrovsky, I.G. & Piskunov, N.S. (1937). A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem, *Vestnik Moskovskogo Gosudarstvennogo Universiteta* **1**, 1–26 (in Russian); English translation in *Selected Works of A.N. Kolmogorov*, Vol. 1, V.M. Tikhomirov, ed. Kluwer, Dordrecht, 1991, pp. 242–270.

[20] Linnik, Yu.V. (1961). *Method of Least Squares and Principles of the Theory of Observations*. Oxford.

[21] Scheffé, H. (1959). *Analysis of Variance*. Wiley, New York.

[22] Smirnov, N.V. (1939). An estimate of the discrepancy between empirical distribution curves in two independent samples, *Bulleten' Moskovskogo Gosudarstvennogo Universiteta, Ser. Matematika, Mekhanika* **2**, 3–14 (in Russian).

[23] Smirnov, N.V. (1939). On the deviation of the empirical distribution curve, *Matematicheskii Sbornik* **6**, 3–24 (in Russian).

[24] Smirnov, N.V. (1948). Tables for estimating the goodness of fit of empirical distribution, *Annals of Mathematical Statistics* **19**, 279–281.

[25] Soyfer, V. (1994). *Lysenko and the Tragedy of Soviet Science*. Rutgers University Press, New Brunswick.

YURII SUHOV

Kolmogorov–Smirnov and Cramer–Von Mises Tests in Survival Analysis

The classical one-sample Kolmogorov goodness-of-fit statistic and the two-sample Smirnov statistic are well-known general statistical procedures. They are collectively known as **Kolmogorov–Smirnov (K–S) tests**. Using these techniques in **survival analysis**, modifications have to be made to deal with **censored data**. There are various types of censoring. What is generally known as type I involves observations being known precisely if they are less than a fixed value and only known to exceed the value otherwise. For so-called type II censoring, the smallest r, say, observations out of a possible n are observed. These definitions of censoring are somewhat restricted in terms of survival analysis. Generally, it is assumed that censoring is random for each observation, and special modifications and assumptions are required. For further discussions of censoring see Michael & Schucany [7] for a succinct account, or Andersen et al. [1] for a full account of censoring in the context of survival data.

For type I or type II censoring, adaptations of the Kolmogorov statistic have been made by Barr & Davidson [2], where small-sample percentage points are tabulated; see also Dufour & Maag [3] for modified statistics which can be used with the asymptotic percentage points found by Koziol & Byar [6].

For survival analysis the random censoring case is more commonly met. Fleming et al. [4] proposed one sample and two sample K–S type procedures with randomly censored data. Efficient procedures exist for comparing two populations, such as the **logrank test** for **proportional hazards** and Gehan–Wilcoxon tests (*see* **Nonparametric Methods**), when survival distributions have **proportional odds**. For some alternatives, such as where the difference between two survival curves occurs primarily at a given time, the K–S two-sample statistic should have good power. Examples of this situation occur, for example, in a treatment regime where individuals might obtain short-term benefits but when compared with controls there is no benefit in the longer term. Other examples for which the logrank and Gehan–Wilcoxon tests would have little power include the class of models known as *crossing-hazards*.

Fleming et al. [4] modified the K–S procedures so as to deal with randomly right-censored data. Asymptotic results are obtained for the censoring mechanism being independent of the survival time. The K–S statistics are defined in terms of the difference of two distribution functions, $F_u(t)$ and $F_v(t)$, where $F_u(t)$ is the empirical (sample) distribution function (edf) and, for the one-sample goodness-of-fit statistic, $F_v(t)$ is a hypothesized distribution function and, for the two-sample case, $F_v(t)$ is the edf of the second sample. The classical two-sided K–S statistic is

$$D = \sup_t |F_u(t) - F_v(t)|$$

or, alternatively, with survivor functions $S_\cdot(\cdot)$ replacing distribution functions $F_\cdot(\cdot)$:

$$D = \sup_t |S_u(t) - S_v(t)|.$$

To obtain statistics which can deal with censored samples, Fleming et al. [4] modify the statistic D and they give computing formulas for the one-sample and two-sample statistics. They also give a simple formula to calculate ***P* values** based on the asymptotic distribution of the statistic, which, in **simulation** studies, were found to be conservative for the modified two-sample Smirnov statistic with heavily censored data in small samples. Later work of Guilbaud [5] gives exact small sample percentage points for the K–S type statistic.

In **Monte Carlo** simulations, Fleming et al. [4] investigate the power of the modified Smirnov two-sample statistic and compare it with Gehan–Wilcoxon and logrank statistics for various alternatives. For those alternatives purposely designed to have substantial differences between the two survivor functions at a given time and not necessarily at other times, the modified Smirnov statistic had good power. It is suggested that the modified statistics should be used in conjunction with the Gehan–Wilcoxon and logrank statistics and, of course, plots of the survivor functions.

Further results on the **power** of the modified Smirnov statistic are given by Stablein & Koutrouvelis [10] who consider crossing hazards alternatives. For such alternatives the modified Smirnov statistic has very good power, considerably in excess of the logrank statistic.

Schumacher [9] also investigates a Smirnov statistic (called by the author a K–S statistic), which has

as its asymptotic distribution the supremum of the Brownian bridge (*see* **Brownian Motion and Diffusion Processes**). In simulation studies the finite sample distributions are found to converge slowly to the asymptotic distribution, and consequently the author is not keen to recommend their use. However, with cheap computing it is a straightforward matter to simulate such statistics, and this is a harsh conclusion.

In the book by Andersen et al. [1] further references are given to more recent work and the authors take an approach which defines statistics having their asymptotic distribution given by distributions derived from the Brownian bridge.

Cramér–von Mises (C–VM) statistics can be defined for the one-sample problem for types I and II censoring as described above. In general form they are defined by

$$\omega^2 = \int [F_u(t) - F_v(t)]^2 \psi(t)\, \mathrm{d}t,$$

where the same notation is used as above for the K–S statistics and ψ is a weight function. Pettitt & Stephens [8] modify the C–VM statistic to deal with this type of censoring and give computing formulas for the one-sample statistics and tabulate asymptotic distributions when the null hypothesis completely determines the survival function.

Koziol & Green [6] introduce a Cramér–von Mises statistic to test the **goodness of fit** for randomly censored data survival times. They also assume independent censoring where the survival function of the censoring distribution is that of the survival times raised to a positive power, β. They find the asymptotic distribution of the C–VM statistic, which is quite sensitive to the value of β. They discuss how to estimate β from the sample. However, this restriction seems rather harsh for application to survival data found in practice.

Schumacher [9] considers C–VM statistics for the two-sample problem which have their asymptotic distributions given by the standard form, i.e. the integral of the square of the Brownian bridge, and in simulation studies finds that the asymptotic distribution is acceptable for small samples. In a power study, the two-sample C–VM statistic was found to have good power for a crossing survival curve alternative and an "early difference" case. Schumacher [9] also gives references to earlier works and tables of percentage points of various functionals of the Brownian bridge which arise as asymptotic distributions of K–S and C–VM type test statistics.

Values of the various goodness-of-fit and two-sample statistics should be enhanced by plots of data, or vice versa, and Michael & Schucany [7] give details of a number of plots for censored data.

In conclusion, the K–S and C–VM statistics provide useful procedures to detect differences, either in the one- or two-sample cases, which are less likely to be detected using tests based on proportional hazards or proportional odds assumptions.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Barr, D.R. & Davidson, T. (1973). A Kolmogorov-Smirnov test for censored samples, *Technometrics* **15**, 739–757.

[3] Dufour, R. & Maag, U.R. (1978). Distribution results for modified Kolmogorov–Smirnov statistics for truncated or censored data, *Technometrics* **20**, 29–32.

[4] Fleming, T.R., O'Fallon, J.R., O'Brien, P.C. & Harrington, D.P. (1980). Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right-censored data, *Biometrics* **36**, 607–625.

[5] Guilbaud, O. (1988). Exact Kolmogorov-type tests for left-truncated and/or right censored data, *Journal of the American Statistical Association* **83**, 213–221.

[6] Koziol, J.R. & Byar, D.P. (1975). Percentage points of the asymptotic distribution of one and two sample K-S statistics for truncated or censored data, *Technometrics* **17**, 507–510.

[7] Michael, J.R. & Schucany, W.R. (1986). Analysis of data from censored samples, in *Goodness-of-Fit Techniques*, R.B. D'Agostino & M.A. Stephens, eds. Marcel Dekker, New York, pp. 461–496.

[8] Pettitt, A.N. & Stephens, M.A. (1976). Modified Cramér–von Mises statistics for censored data, *Biometrika* **63**, 291–298.

[9] Schumacher, M. (1984). Two-sample tests of Cramér–von Mises and Kolmogorov–Smirnov-type for randomly censored data, *International Statistical Review* **52**, 263–281.

[10] Stablein, D.M. & Koutrouvelis, I.A. (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored data, *Biometrics* **41**, 643–652.

(*See also* **Hypothesis Testing**)

ANTHONY N. PETTITT

Kolmogorov–Smirnov Test

Consider two independent groups of subjects. To be concrete, suppose an experimental group consists of sons of alcoholic fathers, and each subject consumes a precise amount of alcohol. Suppose some outcome, X, is measured, such as hangover symptoms, and let Y be the outcome for a control group. Let $F(x)$ be the probability that a randomly sampled subject from the experimental group gets a score less than or equal to x. Similarly, let $G(x)$ be the probability that a randomly sampled subject from the control group gets a score less than or equal to x. The Kolmogorov distance between these two distributions is the maximum possible value of $|F(x) - G(x)|$, the maximum being taken over all possible values of x. If the distributions are identical, meaning that $F(x) = G(x)$ for all possible values of x, then the Kolmogorov distance is zero. From a graphical point of view the Kolmogorov distance is the largest vertical distance between the two cumulative distribution functions.

Kolmogorov-type tests are methods for comparing distributions that are based on the Kolmogorov distance function. In some cases one of the distributions might be specified. For example, it might be hypothesized that $F(x)$ is a normal distribution with specified mean and variance, and the goal might be to determine whether this hypothesis is reasonable based on observations that are available. This hypothesis can be tested with what is called a Kolmogorov test. A Kolmogorov–Smirnov test is a Kolmogorov-type test where the goal is to compare two unknown distributions. That is, $F(x)$ and $G(x)$ are not known for any x, but they can be estimated based on randomly sampled subjects from each group, and the goal is to test $H_0 : F(x) = G(x)$, for any x, the hypothesis that the two distributions are identical.

Let $X_1, \ldots, X_m$ be a random sample of observations from the first group, let $Y_1, \ldots, Y_n$ be a random sample from the second, and let $Z_1, \ldots, Z_N$ be the pooled observations, where $N = n + m$. That is, $Z_i = X_i, i = 1, \ldots, m$, and $Z_{n+i} = Y_i, i = 1, \ldots, n$. For any x, let $a_i = 1$ if $X_i \leq x$, otherwise $a_i = 0$. Similarly, let $b_i = 1$ if $Y_i \leq x$, otherwise $b_i = 0$. Let $\hat{F}(x) = \sum a_i/m$ and $\hat{G}(x) = \sum b_i/n$. The Kolmogorov distance between the distributions F and G is estimated by

$$D = \max |\hat{F}(Z_i) - \hat{G}(Z_i)|,$$

where the maximum is taken over all $i = 1, \ldots, N$. If D is sufficiently large, then reject H_0. When there are no ties, the exact probability of a type I error (*see* **Level of a Test**) can be determined using an algorithm derived by Kim & Jennrich [3]. When there are tied values, results in [4] can be used. Details about these algorithms, together with appropriate software, are summarized in [7].

A common criticism of the Kolmogorov–Smirnov test is that it has low **power** under normality. Table 1 compares its power with several other methods for comparing measures of location. The methods are **Student's *t*** (T), Welch's adjusted degrees of freedom procedure (W), Yuen's [8] method for trimmed means that reduces to Welch's test when there is no trimming (Y), and a method for comparing one-step M-estimators (*see* **Robustness**) using a **bootstrap method** (OSM). (For details about these tests, see [6].) In Table 1 the first three distributions are normal with variance one, and the difference between the means is δ. The notation CN1 refers to a symmetric heavy-tailed distribution that is a mixture of two normal distributions. It has distribution

$$H(x) = 0.9\Phi(x) + 0.1\Phi\left(\frac{x}{k}\right),$$

Table 1 Estimated power, $m = n = 25, \alpha = 0.05$

Distributions	δ	T	W	Y	OSM	KS (exact)	KS (α = 0.052)
Normal	0.6	0.529	0.536	0.464	0.531	0.384	0.464
Normal	0.8	0.778	0.780	0.721	0.751	0.608	0.700
Normal	1.0	0.925	0.931	0.890	0.921	0.814	0.872
CN1	1.0	0.326	0.278	0.784	0.788	0.688	0.780
CN2	1.0	0.191	0.162	0.602	0.760	0.698	0.772
Expo	0.6	0.539	0.697	0.623	0.592	0.866	0.867
Logn	0.6	0.232	0.243	0.409	0.363	0.666	0.678

with $k = 10$, and where $\Phi(x)$ is the standard normal distribution. That is, with probability 0.9, an observation is sampled from a standard normal distribution, otherwise sampling is from a normal distribution having standard deviation $k = 10$. The difference between the standard normal and CN1 is small as measured by the Kolmogorov distance – it is less than 0.04. Despite this, the variance is equal to 10.9 vs. 1 for the standard normal. The distribution CN2 is the same as CN1, only $k = 20$. Finally, Expo indicates an **exponential distribution**, and Logn is **lognormal**. The column headed KS (exact) means that the smallest critical value is used such that the probability of a type I error does not exceed $\alpha = 0.05$. The exact probability of a type I error is 0.036. The last column reports power when the critical value is chosen so that the probability of a type I error is as close as possible to 0.05, which in this case is 0.052.

As would be expected, methods for comparing measures of location have more power when sampling from normal distributions, but, with even slight departures from normality (CN1 and CN2), the Kolmogorov–Smirnov test has substantially more power than methods based on means, and it competes well with methods based on robust measures of location.

A criticism of the Kolmogorov–Smirnov test is that when the sample sizes are small, there are situations where the exact probability of a type I error might not be acceptably close to some desired level, because the test statistic, D, has a discrete distribution. Suppose, for example, $\alpha = 0.05$, and consider $n = m = 10$. The exact probability of a type I error, based on the critical value in [1], is 0.035. For $n = m = 11$, 12, and 13 the exact type I error probabilities are 0.036, 0.031, and 0.044, but for $n = m = 14$ it is 0.019, which might be considered too small. However, the next highest significance level is 0.12 which might be considered too high. This problem might be used to argue for comparing some measure of location, but most methods for comparing measures of location can also yield unsatisfactory control over the probability of a type I error, particularly methods based on means (see, for instance [5] and [6]). An exception appears to be a percentile t bootstrap combined with a 20% **trimmed** mean (see [7]).

An advantage of the Kolmogorov–Smirnov test is that it is sensitive to several features of the data which can be revealed using the method described in [2]. This uses the Kolmogorov–Smirnov test statistic to compute a **confidence interval** for the difference between any two **quantiles** such that the simultaneous probability coverage of all such intervals is determined exactly. Note that the 0.2 quantile, for example, of the first group might be larger than the 0.2 quantile of the second, but when the 0.8 quantiles are compared the reverse can be true. That is, the confidence band can indicate a difference between subpopulations of subjects that is completely missed when attention is restricted to a measure of location. Doksum & Sievers [2] suggest plotting the estimated quantiles of the first group vs. the difference between the quantiles. Letting $\hat{x}_q$ and $\hat{y}_q$ be the quantiles of the two groups, plot $\hat{x}_q$ vs. $\delta = \hat{x}_q - \hat{y}_q$. For illustrations and software, see [7].

References

[1] Conover, W.J. (1980). *Practical Nonparametric Statistics*. Wiley, New York.

[2] Doksum, K.A. & Sievers, G.L. (1976). Plotting with confidence: graphical comparisons of two populations, *Biometrika* **63**, 421–434.

[3] Kim, P.J. & Jennrich, R.I. (1973). Tables of the exact sampling distribution of the two-sample Kolmogorov–Smirnov criterion, $D_{mn}, m \leq n$, in *Selected Tables in Mathematical Statistics*, Vol. I, H.L. Harter & D.B. Owen, eds. American Mathematical Society, Providence.

[4] Schroër, G. & Trenkler, D. (1995). Exact and randomization distributions of Kolmogorov–Smirnov tests for two or three samples, *Computational Statistics and Data Analysis* **20**, 185–202.

[5] Westfall, P.H. & Young, S.S. (1993). *Resampling-Based Multiple Testing*. Wiley, New York.

[6] Wilcox, R.R. (1996). *Statistics for the Social Sciences*. Academic Press, San Diego.

[7] Wilcox, R.R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego.

[8] Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances, *Biometrika* **61**, 165–170.

(*See also* **Nonparametric Methods**)

R. WILCOX

Kullback–Leibler Information

The Kullback–Leibler [5] information number, $I(P\|Q)$, determined for two probability measures defined on the same measurable space $(\chi, \mathcal{F})$, is a nonnegative number (possibly $+\infty$) which represents "distance" (in a certain sense) between P and Q. This "distance" is not symmetric [in general, $I(P\|Q) \neq I(Q\|P)$], but does have the property that $I(P\|Q) = 0$ if and only if $P = Q$. This quantity is also referred to by a myriad of other names, including information for discrimination, discrimination information, Renyi's information gain, entropy distance, entropy of P relative to Q, cross-entropy, directed divergence, Kullback information.

If P is absolutely continuous with respect to $Q (P \ll Q)$, the **Radon–Nikodym** derivative $P_Q(x)$ is defined almost surely (Q) and serves as a basis for the general definition:

$$I(P\|Q) = \begin{cases} \displaystyle\int_\chi \ln P_Q \, \mathrm{d}P = \int_\chi P_Q \ln P_Q \, \mathrm{d}Q, \\ \qquad\qquad P \ll Q, \\ +\infty, \qquad P \not\ll Q. \end{cases}$$

Note that, if R is a sigma-finite measure such that $P \ll Q \ll R$, we can also write

$$I(P\|Q) = \int_\chi P_R \ln \left(\frac{P_R}{Q_R} \right) \mathrm{d}R.$$

Thus, for the common situation where P and Q are discrete (and R is the counting measure), the Kullback–Leibler information number reduces to

$$I(P\|Q) = \sum_i p_i \ln \left(\frac{p_i}{q_i} \right),$$

where p_i and q_i are the standard probability mass functions. Note also that, if X is a random vector with probability distribution P, then $I(P\|Q)$ can be interpreted as the expectation of the log of a **likelihood ratio** statistic.

If Q is a uniform measure over a finite set of points and P is a probability measure on the same points, then $I(P\|Q)$ is just the negative of the well-known Shannon entropy [8]. This quantity is very important in statistical information theory, which has its mathematical roots in the concept of entropy in thermodynamics and statistical mechanics.

Given a probability measure Q on a measurable space, the convex set of probability measures defined by

$$S(Q, \rho) = [P : I(P\|Q) < \rho]$$

is often called the I-sphere with center Q and radius ρ. If $\mathcal{E}$ is a convex set of probability measures intersecting $S(Q, \infty)$, a probability measure $R \in \mathcal{E}$ satisfying

$$I(R\|Q) = \inf_{P \in \mathcal{E}} I(P\|Q) \tag{1}$$

is called the I-projection of Q onto ε. Csiszár [1] has shown that the I-projection exists if the convex set $\mathcal{E}$ is closed in variation distance and has also developed an appealing "geometric" approach which characterizes the "tangent hyperplanes" of I-spheres $S(Q, \rho)$.

Problems of the type (1) play a basic role in the information theoretic approach to statistics [4], the theory of large deviations [6], and in statistical physics [3]. Dykstra & Lemke [2] have shown that problems of the form (1) are often equivalent to **multinomial** maximum likelihood problems under various types of constraint regions.

Kullback–Leibler information numbers between distributions in a common family are often quite tractable. Thus, if P and Q are **Poisson distributions** with respective means m_1 and m_2, it is easily shown that $I(P\|Q) = m_1 \ln(m_1/m_2) + m_2 - m_1$.

If P and Q are k-variate normal distributions (*see* **Multivariate Normal Distribution**) with respective mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and respective **covariance matrices** $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, then

$$\begin{aligned} I(P\|Q) = {} & \tfrac{1}{2} \ln \left(\det \boldsymbol{\Sigma}_2 \big/ \det \boldsymbol{\Sigma}_1 \right) \\ & + \tfrac{1}{2} \mathrm{tr} \boldsymbol{\Sigma}_1 \left(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1} \right) \\ & + \tfrac{1}{2} \mathrm{tr} \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'. \end{aligned}$$

Thus, when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, $I(P\|Q)$ reduces to the natural **Mahalanobis distance** $(1/2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ between the two mean vectors.

Though Kullback–Leibler information seems very different from the statistical concept of Fisher **information**, there is actually a rather remarkable connection. For example, if we have multinomial distributions whose probabilities $\pi_i(\theta), i = 0, 1, \ldots, k$, depend upon the parameter θ, then

$$\lim_{\Delta\theta \to 0} \frac{I[\Pi(\theta + \Delta\theta) \| \Pi(\theta)]}{(\Delta\theta)^2} = \frac{1}{2} I_{\mathrm{F}}(\theta),$$

where I_{F} denotes the Fisher information and the Kullback–Leibler information number is calculated from the appropriate multinomial distributions [7, Chapter 15].

If P_n denotes an empirical distribution from a random sample and $\mathcal{E}$ denotes a family of possible models, then a natural estimate of a model from the family is the one (in $\mathcal{E}$) closest to P_n. "Closest" here means in the sense of the Kullback–Leibler information number, i.e. the I-projection of P_n onto $\mathcal{E}$.

Moreover, $2n \inf_{P \in \mathcal{E}} I(P \| P_n)$ is often a desirable test statistic (with nice asymptotic properties) for testing whether the actual distribution is contained in $\mathcal{E}$ (*see* **Large-sample Theory**).

References

[1] Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems, *Annals of Probability* **3**, 146–158.

[2] Dykstra, R.L. & Lemke, J. (1988). Duality of I-projections and maximum likelihood estimation for log-linear models under cone constraints, *Journal of the American Statistical Association* **83**, 546–554.

[3] Jaynes, E.T. (1957). Information theory and statistical mechanics, *Physical Review* **106**, 620–630.

[4] Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York, 1968; Peter Smith Publisher, Magnolia, 1978.

[5] Kullback, S. & Leibler, R.A. (1951). On information and sufficiency, *Annals of Mathematical Statistics* **22**, 79–86.

[6] Sanov, I.N. (1957). On the probability of large deviations of random variables, *Matemateceskii Sbornick N.S.* **42**, 11–44.

[7] Savage, Leonard J. (1972). *The Foundations of Statistics*. Dover, New York.

[8] Shannon, C.E. (1948). A mathematical theory of communication, *Bell Systems Technical Journal* **27**, 379–423, 623–656.

RICHARD DYKSTRA

Kurtosis

Kurtosis is related to the standardized fourth **moment** of a distribution. It is expressed in a number of ways, the most common being

$$\beta_2 = \mu_4/\sigma^4 \quad \text{and} \quad \gamma_2 = \beta_2 - 3,$$

where μ_4 and σ^2 are, respectively, the fourth central moment and the variance of the distribution. Often it is used as a measure to judge the deviation of a distribution from normality. For the **normal distribution**, $\beta_2 = 3$ and $\gamma_2 = 0$. Unimodal distributions with values of β_2 greater than 3 (called *leptokurtic*) usually indicate that the distribution displays a higher "peak" around the mean, and also more probability in the tails of the distribution, than does the normal (see Figure 1). These distributions are also called thick-(or long-)tailed distributions [3, Chapter 9]. Those with β_2 less than 3 (called *platykurtic*) usually are more concentrated about the mean and flatter than the normal. β_2 cannot be less than 1. Those with $\beta_2 = 3$ are called *mesokurtic*.

Sample measures of kurtosis are used in tests of normality. A common test statistic is $b_2 = m_4/m_2^2$, where m_2 and m_4 are the second and fourth moments about the mean (see [3, Chapter 9]). Extensive tables of the **sampling distribution**, as well as approximations, exist for the null distribution of b_2 [3, 4, 7]. Also, it is often used jointly with a measure of **skewness** ($b_1 = m_3/m_2^{3/2}$) to evaluate deviations from normality [2, 3]. D'Agostino & Pearson [2] and D'Agostino et al. [5] developed a **chi-square distribution** approximation that combines b_2 with b_1 for an omnibus test. It is a useful adjunct to normal probability plots [5] (*see* **Normal Scores**).

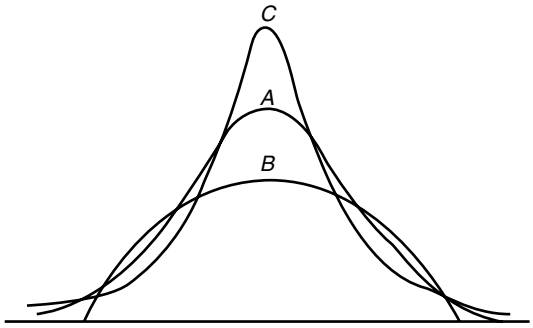

Figure 1 Unimodal distributions: $A = \beta_2 = 3$; $B = \beta_2 < 3$; $C = \beta_2 > 3$

There have been a number of attempts to generalize kurtosis to the multivariate setting. These have often been in the context of developing **tests of multivariate normality**.

In the case of nonnormality, knowledge of kurtosis is useful in evaluating the **robustness** of standard statistical procedures such as the **Student's *t* tests** [8] and in developing measures of location [1]. For this latter situation in particular, the tail thickness of a distribution is often very important and some have questioned the usefulness of b_2 for evaluating it. Alternative measures based on sample **quantiles** have been suggested [6, 9].

References

[1] D'Agostino, R.B. & Lee, A.F. (1977). Robustness of location estimators under changes of population kurtosis, *Journal of the American Statistical Association* **72**, 393–396.

[2] D'Agostino, R.B. & Pearson, E.S. (1973). Testing for departures from normality. I. Fuller empirical results for the distribution of b_2 and b_1, *Biometrika* **60**, 613–622.

[3] D'Agostino, R.B. & Stephens, M.A. (1986). *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

[4] D'Agostino, R.B. & Tietjen, G.L. (1971). Simulation probability points for b_2 for small samples, *Biometrika* **58**, 669–672.

[5] D'Agostino, R.B., Belanger, A.J. & D'Agostino, R.B., Jr (1990). A suggestion for using powerful and informative tests of normality, *American Statistician* **44**, 316–321.

[6] Hogg, R.V. (1972). More light on the kurtosis and related statistics, *Journal of the American Statistical Association* **67**, 422–424.

[7] Pearson, E.S. & Hartley, H.O. (1954). *Biometrika Tables for Statisticians*, Vol. 1. Cambridge University Press, Cambridge.

[8] Pearson, E.S. & Please, N.W. (1975). Relation between the shape of a population and the robustness of four simple test statistics, *Biometrika* **62**, 223–241.

[9] Royston, P. (1992). Which measures of skewness and kurtosis are best?, *Statistics in Medicine* **11**, 333–343.

RALPH B. D'AGOSTINO, SR

Lagged Dependent Variables

In **longitudinal** studies, several observations are taken from each individual at different time points. Often, an observation depends on previous observations; for example, in a **crossover** clinical trial, observations in one period may depend on the observations in the previous periods. A simple model for this scenario might include a lag-1 dependent variable as an explanatory variable [2]:

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}\boldsymbol{\beta} + u_i + e_{it}, \qquad (1)$$

where y_{it} is the observation from subject i in period t, $\mathbf{x}_{it}$ is a vector of **covariates**, u_i is a subject effect, and e_{it} is an error term. This model can be extended to include multiple lagged variables by replacing $\gamma y_{i,j-1}$ by $\sum_{l=1}^{p} \gamma_l y_{i,t-l}$ in (1). Model (1) is different from a **serially correlated** model with the same covariates. In the latter, $y_{i,t}$ depends on $\mathbf{x}_{it}$ only (not $y_{i,t-1}$), while in the former it depends on all $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}$ [4].

Statistical inference based on model (1) includes model fitting, **model checking** and **hypothesis tests**. In biostatistics, the number of subjects is often large, but the number of observations from each subject is small. In this situation we should be careful when using the asymptotic properties of the estimated parameters. For n subjects and times $1, 2, \ldots, T$, and conditional on u_i, the log **likelihood** function of this model can be written as

$$l(\boldsymbol{\beta}, \gamma, \mathbf{u}) = \sum_{i=1}^{n}\sum_{j=1}^{T} \log[p(y_{it}|y_{i,t-1}, \boldsymbol{\beta}, \gamma, u_i)]. \quad (2)$$

When there are no subject effects ($u_i = 0$), this model can be fitted easily using the lagged dependent variables as covariates [3]. When u_i is fixed and $u_i \neq 0$ the **maximum likelihood** estimates (mle) of γ and $\boldsymbol{\beta}$ are not **consistent** for fixed T when the total sample size $n \to \infty$ [2]. To obtain consistent estimates, the **instrumental variable** procedure can be used either for fixed or random u_i. To illustrate how this procedure works we write model (1) as

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + \boldsymbol{\beta}(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + e_{it} - e_{i,t-1}. \qquad (3)$$

Directly using $(y_{i,t-1} - y_{i,t-2})$ as a covariate may lead to inconsistency, since it and $e_{it} - e_{i,t-1}$ are correlated. However, $(y_{i,t-2} - y_{i,t-3})$ or $y_{i,t-2}$ is independent of $e_{it} - e_{i,t-1}$ and can be used as an instrumental variable. When assuming $u_i \sim \mathrm{N}(0, \sigma_u^2)$ the log likelihood function is more complicated than (2), but the mle can be obtained by the Newton–Raphson method (*see* **Optimization and Nonlinear Equations**). In this case the mle is consistent for fixed T and $n \to \infty$.

Model (1) can be extended to include discrete outcomes. One approach is to discretize y_{ij} by letting $y_{ij}^* = 1$ if $y_{ij} > 0$ and $y_{ij}^* = 0$ otherwise. This approach leads to the autoregressive probit model [1]. A more general approach is to use (1) as the linear predictor in a **generalized linear model**, and a wide range of data such as count data can then be modeled. Again, the model fitting without u_i is easy but the mle for random u_i is very difficult to obtain.

When there are missing data in the repeated measurements we may need to write (1) in another form. For example, when y_{i2} is missing we can write the equation for y_{i3} with y_{i1} as a covariate. This can be done by replacing y_{i2} by its regression model. However, the model becomes nonlinear, and nonlinear regression procedures should be used [4].

There are two special issues in the models with lagged dependent variables. One is the distinction between these models and other models for correlated outcomes. To distinguish these models from the models with random subject effects, we may test for given u_i if y_{ij} depends on the previous outcomes. To distinguish these models from serially correlated models, we may test if y_{ij} depends on previous covariates. Another issue concerns the initial observation y_{i0}. Assuming y_{i0} as fixed leads to a simple model, but it may not be reasonable for models with random u_i. The case when y_{i0} is random is more complicated; see [2] for details.

References

[1] Hamerle, A. & Ronning, G. (1995). Panel analysis for qualitative variables, in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg. & M.E. Sobel, eds. Plenum, New York.

[2] Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge University Press, Cambridge, Chapter 4.

[3] Jones, B. & Kenward, M.G. (1989). *Design and Analysis of Cross-over Trials*. Chapman & Hall, London.

[4] Rosner, B. & Munoz, A. (1992). Conditional linear models for longitudinal data, in *Statistical Models for Longitudinal Studies of Health*, J.H. Dwyer, M. Feinleib, P. Lippert & H. Hoffmeister, eds. Oxford University Press, New York.

B. JONES & J. WANG

Lambda Criterion, Wilks'

In 1932, Wilks [35] proposed the **likelihood ratio test** criterion, known usually as Wilks' Λ criterion, for testing the equality of the mean vectors of k p-variate normal distributions with common but unknown **covariance matrix** (*see* **Multivariate Normal Distribution**). Later, Wilks [36] and Bartlett [2] extended its use for testing regression coefficients; see Anderson [1] (*see* **Multiple Linear Regression**).

The problem in its canonical form can be expressed as follows. Let $(\mathbf{X}) : p \times r$, $\mathbf{Y} : p \times m$, and $\mathbf{Z} : p \times n$ be random matrices such that the columns of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are independently distributed as p-variate normal distributions with the same covariance matrix $\mathbf{\Sigma}$. The problem is to test $\mathrm{H}_0 : \mathbf{\Theta} \equiv \mathrm{E}(\mathbf{X}) = \mathbf{0}$ against $\mathrm{H}_1 : \mathbf{\Theta} \neq \mathbf{0}$, given that $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$, $\mathbf{\Sigma}$ being unknown. The likelihood-ratio test, evaluated by Wilks [34], rejects H_0 if and only if

$$V_{p,r,n} \equiv \frac{\det(\mathbf{ZZ}')}{\det(\mathbf{XX}' + \mathbf{ZZ}')}$$

is too small; here "det" denotes the determinant. The Λ criterion is the $\frac{1}{2}(r+m+n)$th power of $V_{p,r,n}$. In the context of the original problem or the **multivariate analysis of variance** (MANOVA) problem, the matrices $\mathbf{XX}'$ and $\mathbf{ZZ}'$ denote the sums of products and cross products matrices due to the hypothesis H_0 and due to error, respectively. It is tacitly assumed that $n \geq p$.

Wilks [35] derived the null distribution of $V_{p,r,n}$ explicitly for $p = 1, 2, 3$ with $r = 3$, and for $p = 4$ with $r = 4$; see also Consul [5] and Mathai [17].

The null distribution of $V_{p,r,n}$ can be expressed as the distribution of $U_1, U_2, \ldots, U_p$, where the U_is are independently distributed, with the distribution of U_i being the **beta distribution** $\mathrm{B}(\frac{1}{2}(n+1-i), r/2)$; moreover, the distribution of $V_{p,r,n}$ is the same as that of $V_{r,p,n+r-p}$; see [1]. For $p = 1, 2$ and $r = 1, 2$ the distributions of $V_{p,r,n}$ take simple ***F* distribution** forms as follows [1]:

$$\frac{1 - V_{1,r,n}}{V_{1,r,n}} \frac{n}{r} \sim F_{r,n},$$

$$\frac{1 - V_{p,1,n}}{V_{p,1,n}} \frac{n+1-p}{p} \sim F_{p,n+1-p},$$

$$\frac{1 - (V_{2,r,n})^{1/2}}{(V_{2,r,n})^{1/2}} \frac{n-1}{r} \sim F_{2r,2(n-1)},$$

$$\frac{1 - (V_{p,2,n})^{1/2}}{(V_{p,2,n})^{1/2}} \frac{n+1-p}{p} \sim F_{2p,2(n+1-p)}.$$

Wald & Brookner [37] presented a method for obtaining the null distribution of $V_{p,r,n}$ for even values of p and r; see also Schatzoff [29] and Anderson [1]. For other results on the null distribution, see Pillai & Gupta [24] and the books by Seber [32] and Muirhead [20].

Tables for significance points of $V_{p,r,n}$ are obtained by Schatzoff [29], Pillai & Gupta [24], Lee [16] and Davis [9, 10]; see also Anderson [1], Muirhead [20], and Pearson & Hartley [22].

Bartlett [3] has shown that the null distribution of $-[n - \frac{1}{2}(p - r + 1)] \log V_{p,r,n}$ tends to the **chi-square distribution** with pr **degrees of freedom** as $n \to \infty$; see [1]. Mudholkar & Trivedi [19] have suggested a normal approximation to the distribution of $-\log V_{p,r,n}$ for large p or r; see [1]. This approximation is better than the chi-square approximation when n is small. Rao [27] has suggested an F approximation as follows:

$$\frac{1 - V^{1/s}}{V^{1/s}} \frac{ks - q}{pr} \sim F_{pr,ks-q},$$

where $s = [(p^2r^2 - 4)/(p^2 + r^2 - 5)]^{1/2}$, $q = (pr/2) - 1$, and $k = r - (p - r + 1)/2$. For small r, this approximation is more accurate than the chi-square approximation.

An asymptotic expansion of the null distribution of $V_{p,r,n}$ (in powers of $1/n$) in terms of chi-square distributions has been given in Rao [26], Anderson [1], and Muirhead [20].

Constantine [4] has obtained the moments of the nonnull distribution of $V_{p,r,n}$. For asymptotic expansion of the nonnull distribution in terms of noncentral chi-square distributions, see Muirhead [20], Suguira [33], Suguira & Fujikoshi [34], Fujikoshi [12], and Pillai [23].

Schwarz [31] has shown that the likelihood ratio test is **Bayes** and admissible; see also Anderson [1]. The **power** function of this test depends on the parameters only through the characteristic roots **(eigenvalues)** $\nu_1, \ldots, \nu_p$ of $\mathbf{\Theta\Theta}'\mathbf{\Sigma}^{-1}$. DasGupta et al. [8] have shown that the power of the likelihood ratio test monotonically increases as each ν_i increases; see also a review paper by DasGupta [6]. The power of this test has been studied by DasGupta & Perlman [7].

The power functions of the likelihood ratio test, the **Lawley–Hotelling trace** test, and **Pillai's trace test** for the MANOVA problem have been compared by Rothenburg [28] on the basis of asymptotic expansions. It is shown that if the coefficient of variation of the ν_is is large enough, then the power of the Lawley–Hotelling trace test is greater than that of the likelihood ratio test, which in turn is greater than that of Pillai's test; in the opposite situation, the ordering of power is reversed. For comparisons of the power function of the likelihood ratio test with the power functions of other standard tests for the MANOVA, see Itô [13], Lee [15], Mikhail [18], Olson [21], Pillai & Jayachandran [25], and Schatzoff [30]. Olson's study [21] indicates that the likelihood ratio test is quite robust under departure from covariance homogeneity. For a review of results on **robustness**, see Itô [14] (*see* **Multivariate Techniques, Robustness**).

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Bartlett, M.S. (1934). The vector representation of a sample, *Proceedings of the Cambridge Philosophical Society* **30**, 327–340.

[3] Bartlett, M.S. (1938). Further aspects of the theory of multiple regression, *Proceedings of the Cambridge Philosophical Society* **34**, 33–40.

[4] Constantine, A.G. (1963). Some noncentral distributional problems in multivariate analysis, *Annals of Mathematical Statistics* **34**, 1270–1285.

[5] Consul, P.C. (1966). On the exact distributions of the likelihood ratio criterion for testing linear hypothesis about regression coefficients, *Annals of Mathematical Statistics* **37**, 1319–1330.

[6] DasGupta, S. (1980). Monotonicity and unbiasedness property of ANOVA and MANOVA tests, in *Handbook of Statistics*, Vol. 1, P.R. Krishnaiah, ed. North-Holland, New York, pp. 179–198.

[7] DasGupta, S. & Perlman, M.D. (1973). On the power of Wilks' U-test for MANOVA, *Journal of Multivariate Analysis* **3**, 220–225.

[8] DasGupta, S., Anderson, T.W. & Mudholkar, G.S. (1964). Monotonicity of the power functions of some tests of multivariate linear hypothesis, *Annals of Mathematical Statistics* **35**, 200–205.

[9] Davis, A.W. (1971). Percentile approximations for a class of likelihood ratio criteria, *Biometrika* **58**, 349–356.

[10] Davis, A.W. (1979). On the differential equation for Meijer's $G^{p,0}_{p,p}$ function and further tables of Wilks's likelihood ratio criterion, *Biometrika* **66**, 519–531.

[11] Davis, A.W. (1980). On the effects of moderate multivariate abnormality on Wilks's likelihood ratio criterion, *Biometrika* **67**, 419–427.

[12] Fujikoshi, Y. (1973). Asymptotic formulas for the distributions of three statistics for multivariate linear hypothesis, *Annals of the Institute of Statistical Mathematics* **25**, 423–437.

[13] Itô, K. (1962). A comparison of the powers of two multivariate analysis of variance tests, *Biometrika* **49**, 455–482.

[14] Itô, P.K. (1980). Robustness of ANOVA and MANOVA test producers, in *Handbook of Statistics*, Vol. 1, P.R. Krishnaiah, ed. North-Holland, New York, pp. 199–236.

[15] Lee, Y.S. (1971). Asymptotic formulae for the distribution of a multivariate test statistic: power comparisons of certain multivariate tests, *Biometrika* **58**, 647–651.

[16] Lee, Y.S. (1972). Some results on the distribution of Wilks' likelihood ratio criterion, *Biometrika* **59**, 649–664.

[17] Mathai, A.M. (1971). On the distribution of the likelihood ratio criterion for testing linear hypothesis on regression coefficients, *Annals of the Institute of Statistical Mathematics* **23**, 181–197.

[18] Mikhail, N.N. (1965). A comparison of tests of the Wilks-Lawley hypothesis in multivariate analysis, *Biometrika* **52**, 149–156.

[19] Mudholkar, G.S. & Trivedi, M.C. (1981). A normal approximation for the multivariate likelihood ratio statistics, in *Statistical Distributions in Scientific Work*, Vol. 5, C. Tallie et al., eds. Reidel, Dordrecht pp. 219–230.

[20] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

[21] Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 874–908.

[22] Pearson, E.S. & Hartley, H.O. (1972). *Biometrika Tables for Statisticians*, Vol. II. Cambridge University Press, Cambridge.

[23] Pillai, K.C.S. (1977). Distributions of characteristic roots in multivariate analysis, part II: non-null distributions, *Canadian Journal of Statistics* **5**, 1–62.

[24] Pillai, K.C.S. & Gupta, A.K. (1969). On the exact distribution of Wilks's criterion, *Biometrika* **56**, 109–118.

[25] Pillai, K.C.S. & Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria, *Biometrika* **54**, 195–210.

[26] Rao, C.R. (1948). Tests of significance in multivariate analysis, *Biometrika* **35**, 58–79.

[27] Rao, C.R. (1956). An asymptotic expansion of the distribution of Wilks' criterion, *Bulletin of the International Statistical Institute* **33**, 177–180.

[28] Rothenburg, T.J. (1977). Edgeworth expansions for multivariate test statistics, *IP-255*, Center for Research in Management Science, University of California, Berkeley.

[29] Schatzoff, M. (1966). Exact distribution of Wilks's likelihood ratio criterion, *Biometrika* **53**, 347–358.

[30] Schatzoff, M. (1966). Sensitivity comparisons among tests of the general linear hypotheses, *Journal of the American Statistical Association* **61**, 415–435.

[31] Schwarz, R. (1967). Admissible tests in multivariate analysis of variance, *Annals of Mathematical Statistics* **38**, 698–710.

[32] Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

[33] Sugiura, N. (1973). Further asymptotic formulas for the non-null distributions of three statistics for multivariate linear hypothesis, *Annals of the Institute of Statistical Mathematics* **25**, 153–163.

[34] Sugiura, N. & Fujikoshi, Y. (1969). Asymptotic expansions of the non-null distributions of the likelihood ratio criteria for multivariate linear hypothesis and independence, *Annals of Mathematical Statistics* **40**, 942–952.

[35] Wilks, S.S. (1932). Certain generalizations in the analysis of variance, *Biometrika* **24**, 471–494.

[36] Wilks, S.S. (1935). On the independence of k sets of normally distributed statistical variables, *Econometrica* **3**, 309–326.

[37] Wald, A. & Brookner, R.J. (1941). On the distribution of Wilks' statistic for testing the independence of several groups of variables, *Annals of Mathematical Statistics* **12**, 137–152.

(*See also* **Multivariate Analysis, Overview**)

SOMESH DASGUPTA

Lancaster, Henry Oliver

Born: February 1, 1913, in Sydney, New South Wales.
Died: December 2, 2001, in Sydney.

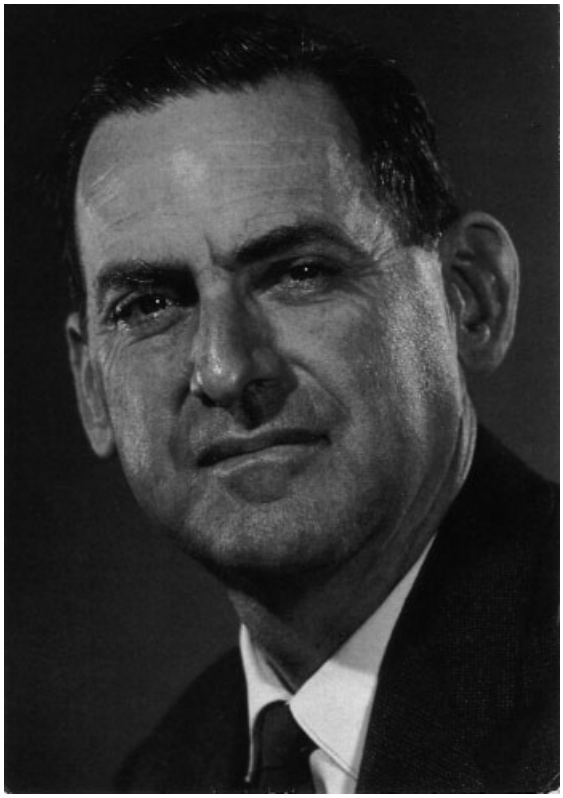

During the twentieth century, many qualified physicians contributed in different ways to the theory or practice of statistics. Oliver Lancaster was unique in having held university chairs in both medical and mathematical statistics, in making important contributions to the history of quantitative medicine, and in his pioneering work in the bibliography of statistics.

Lancaster was the son of a doctor practicing in the country town of Kempsey on the Macleay River in northeastern New South Wales. He was born in Sydney because his mother had accompanied his father who was playing in a chess championship there, a suitable advent for a boy who was to have a precocious childhood and develop strong aptitude in all forms of games. He showed all-round ability in school with particular strength in mathematics and chemistry, and then started on an actuarial career by taking evening classes in economics at the University of Sydney, whilst working for an insurance company. He quickly switched to Arts degree classes with a broader range of science subjects and in 1931 enrolled as a medical student, qualifying in 1937. During the next three years, he served as Resident Medical Officer and then pathologist, and sought to extend his scientific range by further studies in chemistry and a reading of Yule's *Introduction to the Theory of Statistics*. In 1940, after two other hospital appointments, he joined the Royal Australian Army Medical Corps as a pathologist, a period of service in the Middle East, Australia, and New Guinea which was to last until 1946.

His work as an army pathologist introduced him to statistical problems such as the analysis of $2 \times 2 \times 2$ tables in the study of parasitic infections and **Bayesian** concepts in diagnosis. His target at this stage was a possible career in demography and he began a serious study of mathematics, enrolling as an external student in mathematics within the Arts degree. After demobilization, he had a temporary appointment as a Lecturer in Statistics at the Sydney School of Public Health and Tropical Medicine and obtained an Arts degree in 1947.

In 1948, he was awarded a Rockefeller Scholarship in Medicine, to study at the London School of Hygiene and Tropical Medicine under A. Bradford **Hill**. He arrived there with a sheaf of draft articles on a variety of topics in medical statistics, including the analysis of amoebic surveys (involving **overdispersed** binomial distributions (*see* **Beta-binomial Distribution**), the control of routine blood counting, partition of the **chi-square** statistic in **contingency tables** to identify particular contrasts with disproportionate frequencies, and the use of what is now known as the mid-P test for discrete data (*see* **Fisher's Exact Test**). He benefited from the guidance of **J.O. Irwin**, who recognized Lancaster's remarkable insight but realized that it was not matched by experience in the writing-up of his research. He also completed an analysis of the large data set published by A. Geissler in 1889 on the **sex ratio** in families of different sizes, concluding that the evidence for genetic variability in the ratio was extremely scanty.

He returned to Sydney in 1949, as a member of the Commonwealth Health Department located at the

School of Public Health, with some teaching responsibilities. He now embarked on a study of trends in Australian mortality from different causes, extending eventually to some 50 papers. He approached **vital statistics** with an investigative mind and made important epidemiologic findings. N.M. Gregg had discovered that maternal rubella in the first trimester of pregnancy can cause developmental defects, such as cataracts, in the fetus, and it was thought that this effect might be due to a recent viral mutation. Lancaster discovered that excessive incidences of deaf-mutism had occurred as long ago as 1898 and 1899, following rubella epidemics. He also found that melanoma, which was known to be associated with exposure to sunlight, was more specifically related to latitude in Australia, mortality increasing toward the equator.

In 1959, he was appointed Associate Professor of Medical Statistics in the University of Sydney, but almost immediately he applied for, and was appointed to, the new chair in Mathematical Statistics. This marked a new phase in his career. He was never to abandon his interest in medical statistics (as his later book [3] shows), but his concern now was to develop the new department and plan its courses. His research included further work on the chi-square distribution, leading to his book [2]. In this, he made use of orthogonal functions (*see* **Orthogonality**). He had published several papers on this topic, particularly in relation to their use in measuring dependence in **bivariate distributions** with given marginal distributions. His work in this area is summarized in [4]; see also [8].

Throughout his career, Lancaster had been interested in historical scientific literature, especially in relation to statistics. In his book [1], he listed bibliographies contained in a wide range of papers and continued to produce 21 addenda at approximately annual intervals until 1989. For this purpose, he created a vast card index in Sydney, the maintenance of which became one of his principal interests in later life. After his retirement in 1978, he decided to continue and consolidate his work on medical statistics, and produced a massive book [6] on trends in mortality subdivided by diseases and regions, and a history of quantitative medicine [7]. These books are important sources of information although the author's rather laconic and occasionally cryptic style makes few concessions to his less erudite readers.

Lancaster obtained doctorates from Sydney in philosophy (1953), medicine (1967), and science (1971). He was a Fellow of the Australian Academy of Science (1961) holding the Academy's Lyle Medal (1961) and the Pitman Medal of the Statistical Society of Australia (1980). In 1992, he was appointed Officer of the Order of Australia. He held many honorary fellowships and honorary life memberships of learned societies throughout the world.

Lancaster played a major role in the establishment of statistical and mathematical organizations in Australia. In 1947, the Statistical Society of New South Wales was formed and Lancaster became Secretary in 1949 and President from 1952 to 1953. He was largely responsible for the Society's *Bulletin* and was joint Editor of its successor, the *Australian Journal of Statistics*, from 1959 to 1971. After the Society became the Statistical Society of Australia, he served as President from 1965 to 1966. He helped to found the Australian Mathematical Society from 1956 to 1957 and served as its General Secretary (1959–1963) and President (1966–1967).

Lancaster had a hesitant manner of conversation, probably because of a childhood stammer, which he had gradually overcome as an adult, and this sometimes conveyed the impression of a diffident personality. He was, in fact, a man of strong opinions, often forcibly expressed. He was held in high respect and affection by those who knew him, not least his former research students, many of whom came to occupy senior positions in academia and public service.

Lancaster published his own account of his career in [5].

References

[1] Lancaster, H.O. (1968). *Bibliography of Statistical Bibliographies*. Oliver & Boyd, Edinburgh.

[2] Lancaster, H.O. (1969). *The Chi-squared Distribution*. Wiley, New York.

[3] Lancaster, H.O. (1974). *An Introduction to Medical Statistics*. Wiley, New York.

[4] Lancaster, H.O. (1980). Orthogonal models for contingency tables, in *Developments in Statistics*, Vol. 3, P.R. Krishnaiah, ed. Academic Press, New York, pp. 99–157.

[5] Lancaster, H.O. (1982). From medicine, through medical to mathematical statistics: some autobiographical notes, in *The Making of Statisticians*, J. Gani, ed. Springer, New York, pp. 236–252.

[6] Lancaster, H.O. (1990). *Expectations of Life: A Study in the Demography, Statistics, and History of World Mortality*. Springer-Verlag, New York.

[7] Lancaster, H.O. (1994). *Quantitative Methods in Biological and Medical Sciences: A Historical Essay*. Springer-Verlag, New York.

[8] Seneta, E. (2002). In memoriam: Emeritus Professor Henry Olive Lancaster, AO FAA, 1 February 1913 – 2 December 2001, *Australia and New Zealand Journal of Statistics* **44**, 385–400.

PETER ARMITAGE

Laplace, Pierre–Simon

Born: March 23, 1749, in Beaumont-en-Auge, France.
Died: March 5, 1827, in Paris, France.

"Laplace was among the most influential scientists in all history" [1]. The son of a well-to-do tradesman, he entered the University of Caen in 1766 to study theology, but left prematurely to study mathematics in Paris under d'Alembert. He secured an appointment at the Ecole Militaire (where he examined Napoleon), and was elected to the Académie des Sciences in 1773. His career continued to flourish after the Revolution, and for a short time he was Napoleon's Minister of the Interior. He became Chancellor of the Senate in 1803 and a Marquis in 1806.

During the first 20 years of his academic life, he worked prolifically in several areas of mathematical science, notably celestial mechanics, differential equations, and probability and statistics. These remained the central themes throughout his career. As his research findings proliferated and matured, Laplace incorporated them into two major treatises, *Mécanique céleste* (1799–1825) and *Théorie analytique des probabilités* (1812). The second edition of the latter, in 1814, was accompanied by a new introduction, *Essai philosophique sur les probabilités*.

Laplace's early work on probability adopted a **Bayesian** approach. Laplace had discovered **Bayes' theorem**, possibly unaware in 1774 of Bayes' posthumous publication of 1763. He applied this to combinatorial and demographic problems, and used the **beta** prior distribution for **binary data**. Other techniques and results introduced by Laplace included **generating functions** for discrete distributions, **characteristic functions**, the Laplace transform (at least in embryo), a form of the **central limit theorem**, and various aspects of **regression** and **least squares**. Commentators on his career usually refer to his somewhat cavalier attitude towards results obtained by other workers. According to [1], "not a single [contemporary] testimonial bespeaking congeniality survives". Nevertheless, as Grattan-Guiness [2] remarks,

> Laplace's contributions to probability and statistics were fundamental . . .He . . .changed the emphasis of probability from its preoccupation with moral sciences and jurisprudence to include also applications in scientific contexts, wither it had hitherto infrequently strayed. His most important early successors were Quetelet and Poisson; after them, both probability and statistics moved to adulthood in the family of sciences, and the heritage from Laplace began to be recognized.

The major memoir [1], written in collaboration with others, contains extensive bibliographic information. The relation between Laplace's work and that of his near-contemporaries, especially **Gauss**, is described in [3].

References

[1] Fox, R., Gillispie, C.C. & Gratton-Guinness, J. (1981). Laplace, Pierre-Simon, Marquis de, in *Dictionary of Scientific Biography*, Vol. 15, C.C. Gillispie, ed. Scribner, New York, pp. 273–403.

[2] Grattan-Guiness, I. (1983). Laplace, Pierre Simon, in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 469–473.

[3] Stigler, S.M. (1986). *The History of Statistics: the Measurement of Uncertainty before 1900*. Belknap Press, Cambridge, Mass.

PETER ARMITAGE

Large-sample Theory

Large-sample theory (LST) plays a fundamental role in biostatistics in the prescription of fruitful methodology that can be well adapted in practical applications, often under conditions weaker than in standard (finite sample) parametrics. The basic clause of *large samples* is usually satisfied in real biostatistical applications, especially in investigations involving large-scale data collection. The advent of modern computers has strengthened the case for LST. The research literature on LST has gone through a phenomenal growth during the past three decades wherein delicate concepts from **probability** theory and **stochastic processes** have been blended towards a unified resolution, though at the cost of mathematical abstractions and sophistications often beyond the normal range of comprehension in biostatistics. Moreover, in biostatistics, various experimental or observational factors generally impose certain constraints on underlying statistical models, so that the classical parametric theory may not be universally adoptable, and increasingly **nonparametrics** and **semiparametrics** are being used; the LST has an even more dominant role in this setup. Yet there is a hierarchy in the methodological developments within the domain of LST with respect to their validity in moderate sample sizes, and many modern developments are geared toward a better resolution for moderate to large sample sizes. The interesting point in this context is the interplay between *validity robustness* and *efficiency robustness*; modern LST addresses this aspect quite well.

The basic concepts in probability theory underlying the evolution of LST in biostatistics are the following:

1. *stochastic, almost sure*, and other *modes of* **convergence**;
2. *probability inequalities*, and **laws of large numbers**; and
3. *weak convergence* or *convergence in distribution (law)*.

In the classical sense these concepts were mostly developed for sample *statistics* that are generally expressible as the sum or average of independent random elements. Yet, in applications one often encounters more general forms of statistics violating this postulation, and even sometimes sample functions that are *stochastic processes* in a general sense. The *empirical distribution* (*see* **Goodness of Fit**) and *survival* functions (*see* **Survival Distributions and Their Characteristics**) are classical examples of this type. In this context the emergence of *martingales, reverse martingales* and related *dependent sequences* has greatly reshaped the adaptability of LST in diverse setups, and our discussion remains somewhat incomplete without their introduction and role (*see* **Counting Process Methods in Survival Analysis**). The intricate role of LST in **transformations** on variables or statistics also deserves a closer look.

The main theme of LST relates to the *asymptotic distribution theory* for various statistics that arise in statistical analysis in biostatistics, where *point* and *confidence set* **estimation**, and **hypothesis testing** occupy a focal point. In this context, *linear*, **generalized linear**, **categorical data** models, and some semiparametric and nonparametric models deserve detailed discussion. LST in survival analysis is also a vital component of this development. For a comprehensive view, we also briefly present some *invariance principles* that play a fundamental role in these developments.

Stochastic Convergence

Let T_n be a statistic based on a sample of size N from a population with distribution function F, and let θ be a parameter which can generally be defined as a function of F. Then T_n is said to converge in probability (or stochastically) to θ if, for every positive η and ε, there exists a positive integer $n_0 = n_0(\eta, \varepsilon)$, such that

$$\Pr\{|T_n - \theta| > \eta\} < \varepsilon, \quad \text{for all } n \geq n_0. \tag{1}$$

If we view T_n as an estimator of θ, then the above definition coincides with the notion of (*weak*) **consistency** in estimation theory. Similarly, T_n converges almost surely (or strongly) to θ if

$$\Pr\{|T_n - \theta| > \eta, \text{ for some } N \geq n\} < \varepsilon, \quad \text{for all } n \geq n_0. \tag{2}$$

Again, in estimation theory this corresponds to the notion of *strong consistency*. In the same vein, T_n is said to converge in the rth mean, for some $r > 0$, if

$$\mathrm{E}\{|T_n - \theta|^r\} \to 0, \quad \text{as } n \to \infty. \tag{3}$$

Note that both almost sure convergence and convergence in the rth mean imply convergence in probability, but the converse may not be true generally. These definitions extend readily for the case of vectors T_n and θ where we need to use the Euclidean or other norms, and also to more general cases by using suitable norms. As an example consider the case of T_n being the sample distribution function defined as $F_n(\cdot)$, and consider the *sup-norm* $||F_n - F|| = \sup\{|F_n(x) - F(x)| : x \in \mathbf{R}\}$. With respect to this metric, the definitions all extend to this case of functional statistics and parameters.

Probability Inequalities

The *Chebyshev inequality*. For a nonnegative random variable U with $\mu = EU$, $\Pr\{U \geq \mu t\} \leq t^{-1}$, for all $t > 0$, provides the genesis of all probability inequalities. Letting $U = (T_n - \theta)^2$ and denoting by $\sigma_n^2 = E\{(T_n - \theta)^2\}$, we have the derived Chebyshev inequality:

$$\Pr\{|T_n - \theta| \geq \varepsilon\} \leq \varepsilon^{-2}\sigma_n^2, \quad \text{for every } \varepsilon > 0, \quad (4)$$

so that a sufficient condition for the stochastic convergence of T_n is that $\sigma_n \to 0$ as $n \to \infty$. Although this characterization does not require T_n to have independent summands, in fact it is the second (and generally rth) mean convergence property. For almost sure convergence and related results some sharper inequalities are useful. Among these, special mention may be made of (i) the *Bernstein inequality*, (ii) the *Kolmogorov–Hájek–Rényi inequality*, and (iii) the *Hoeffding inequality*, all of which were initially formulated for independent summands, but later were generalized to some dependent cases as well. Other useful inequalities in probability theory include the c_r *inequality* ($r > 0$), the *Holder inequality*, the *Cauchy–Schwarz inequality*, and the *Jensen inequality*. For details we refer the reader to Sen & Singer [21] and Ferguson [7], where other pertinent references are all cited.

Laws of Large Numbers (LLN)

For independent and identically distributed (iid) random variables, the *Khintchine Strong LLN* asserts the almost sure convergence of the sample mean to the population mean whenever the latter exists. The *Borel SLLN* refers to the particular case of Bernoulli variables (*see* **Binary Data**). However, without the iid clause, extra regularity conditions are needed for such LLNs to hold. The *Kolmogorov SLLN*, in the case of independent but not necessarily identically distributed summands, is based on the convergence of the series $\sum_{n\geq 1} n^{-2}\sigma_n^2$, where σ_n^2 stands for the variance of X_n, for $n \geq 1$. These LLNs have also been extended to some dependent sequences. The *Markov LLN* relates to stochastic convergence for the possibly nonidentically distributed case, and does not require the second moment condition, but a condition slightly more stringent than the first.

Martingales and Reversed Martingales

Let $\{T_n; n \geq 1\}$ be a sequence of random variables with finite expectations. If $E\{T_n|T_j, j \leq n-1\} = T_{n-1}$ almost everywhere for every $n \geq 1$ (where T_0 can be taken as a constant), then $\{T_n; n \geq 1\}$ is termed a martingale. If in the above (conditional) expectation, for all n, the $=$ is replaced by a $\geq$ (or $\leq$), then we have a submartingale (or supermartingale) sequence. A sequence $\{T_n\}$ forms a reversed martingale if for every n, $E\{T_n|T_{n+1}, T_{n+2}, \ldots\} = T_{n+1}$, almost everywhere, and a similar definition holds for reversed sub (or super) martingales. The sample mean, ***U*-statistics** and many other symmetric estimators can be characterized as reversed martingales. Similarly, sample sums, **likelihood ratio test statistics**, and other forms of **rank** statistics can be characterized as martingales. Score statistics, arising in parametric models [16] as well as in various nonparametric and semiparametric applications [10], are abundant in biostatistics (*see* **Likelihood**). In most of these cases, either a reversed martingale or a forward martingale characterization holds. The empirical distribution function F_n is also a reversed martingale (process). Such dependent sequences show up frequently in survival analysis and other areas in biostatistics. Most of the probability inequalities and LLNs have been extended to such dependent sequences, and hence they enjoy similar convergence properties. We refer to some of these later in the article.

Weak Convergence and CLT

Consider a sequence $\{T_n; n \geq n_0\}$ of random variables or statistics with distribution functions $\{G_n; n \geq$

$n_0\}$. Then T_n is said to converge weakly (or in distribution/law) to a possibly degenerate random variable T with distribution function G, if $||G_n - G|| \to 0$ as $n \to \infty$, i.e. G_n converges to G at all points of continuity of G. Of particular interest is the classical **central limit theorem** (CLT) which relates to the case of a normal G. In the case of iid random variables $\{X_i, i \geq 1\}$ with finite mean μ and variance σ^2, $T_n = n^{-1/2} \sum_{i=1}^n (X_i - \mu)/\sigma$ converges in law to T, where T has the standard normal distribution function. The *Liapounoff* theorem established this weak convergence result in the nonidentically distributed case under a moment condition of order higher than 2, while the classical *Lindeberg–Feller* CLT pertains to the same result under a less stringent *uniform integrability* condition: for all $\varepsilon > 0$.

$$s_n^{-2} \sum_{i=1}^n \mathrm{E}[(X_i - \mathrm{E}X_i)^2 I(|X_i - \mathrm{E}X_i| > \varepsilon s_n)] \to 0, \quad \text{as} \quad n \to \infty \tag{5}$$

where $s_n^2 = \sum_{i=1}^n \mathrm{var}(X_i)$. These CLTs have been extended to more general *triangular* schemes as well as to some multivariate situations. In this context, it may not be necessary to assume that the limiting distribution is of full rank, i.e. degenerate limit laws are also allowed. Moreover, the CLTs hold for various dependent summands, including the martingales and reverse martingales, under some extra mild regularity conditions [5]. In general, in biostatistics, often a statistic T_n does not have independent summands, and may not even be strictly a martingale or reversed martingale, so a CLT may not be applied on it. However, in this context the well-known *Slutsky* theorem, presented below, provides an easily verifiable approach.

Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables not necessarily independent, such that $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{\mathcal{P}} c$, a constant. Then the following results hold:

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + c,$$

$$X_n Y_n \xrightarrow{\mathcal{D}} cX,$$

and

$$\frac{X_n}{Y_n} \xrightarrow{\mathcal{D}} \frac{X}{c}, \quad \text{if } c \neq 0. \tag{6}$$

In a variety of situations, we have the following *projection* result:

$$T_n = T_n^0 + R_n; \quad T_n^0 = \sum_{i=1}^n \mathrm{E}[T_n | X_i] - (n-1)\mathrm{E}[T_n], \tag{7}$$

and the remainder term, R_n, having the nice property that $\mathrm{E}(R_n^2) = \mathrm{E}(T_n - \theta)^2 - \mathrm{E}(T_n^0 - \theta)^2$, stochastically converges to 0 at a rate faster than the standard error of T_n^0, whenever the projection technique yields an asymptotic quadratic mean equivalence. In such a case, the CLT holds for $\{T_n^0\}$ (which has independent summands), while the Slutsky theorem leads to the asymptotic normality of the standardized form of T_n. Hoeffding [8] used this projection result for U-statistics, where T_n^0 is a sample average of iid random variables, and he also indicated how the nonidentically distributed clause can be accommodated in the same vein. During the past 50 years a vast amount of research work has been accomplished in this direction. We refer to Jurečková & Sen [10] for deeper results on *asymptotic representations* of possibly nonlinear statistics in terms of T_n^0 and a remainder term, wherein the algebraic complications underlying the projection technique's asymptotic quadratic mean equivalence has further been replaced by a less stringent weaker expansion. As a simple illustration, consider the case of the sample variance $S_n^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$, where $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the sample mean. Here $T_n^0 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$ and $R_n = O_p(n^{-1})$. Thus, the CLT applies to $\sqrt{n}(S_n^2 - \sigma^2)$, though it is itself not an average of iid random variables.

Some weak convergence results allied to the above CLTs deserve mention. First, the asymptotic version of the *Cochran* theorem. Let $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Gamma})$, and let $\mathbf{A}_n$ be a possibly stochastic matrix, converging (in probability) to $\mathbf{A}$, a generalized inverse of $\boldsymbol{\Gamma}$, with $\mathrm{Tr}(\mathbf{A}) = q(\leq p)$. Then $n(\mathbf{T}_n - \boldsymbol{\theta})' \mathbf{A}_n (\mathbf{T}_n - \boldsymbol{\theta})$ has an asymptotically **chi-square distribution** with q **degrees of freedom**. This theorem has many uses in biostatistics, and we will discuss some of them later in the article. As an illustration, we consider the case of the **Hotelling T^2 statistic** (in the multivariate one-sample model):

$$T_n^2 = n(\overline{\mathbf{X}}_n - \boldsymbol{\mu})' \mathbf{S}_n^{-1} (\overline{\mathbf{X}}_n - \boldsymbol{\mu}), \tag{8}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the population mean vector and **covariance matrix**, and $\overline{\mathbf{X}}_n$ and $\mathbf{S}_n$ are their sample counterparts. Whenever the underlying distribution function has finite second-order moments, $\mathbf{S}_n \xrightarrow{\mathcal{P}} \boldsymbol{\Sigma}$ and the multivariate CLT applies for $\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu})$. Therefore, by the above version of the Cochran theorem, we claim that as n increases, T_n^2 has closely the central chi-square distribution function with p degrees of freedom. This result is useful in testing suitable null hypotheses on $\boldsymbol{\mu}$ as well as for obtaining *confidence sets* for $\boldsymbol{\mu}$, without specifically making the multinormality assumption. In the general context of linear models (without normality of the errors), the conventional least squares procedures lead to various estimates and test statistics where the Slutsky theorem and the above version of the Cochran theorem provide access to related asymptotic distribution theory. We refer to Sen & Singer [21] for most of these details.

Secondly, under the same setup, consider a real valued function $Z_n = g(\mathbf{T}_n)$ and define $\nu = g(\boldsymbol{\theta})$. Then under differentiability of $g(\cdot)$ at $\boldsymbol{\theta}$, we have

$$\sqrt{n}(Z_n - \nu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \gamma^{*2}), \tag{9}$$

where $\gamma^{*2} = (\dot{g})'\boldsymbol{\Gamma}(\dot{g})$, and $\dot{g}$ is the gradient (vector) of $g(\cdot)$ at $\boldsymbol{\theta}$, which is assumed to be nonnull (as otherwise we would have a degenerate normal law). This basic CLT result has numerous applications in biostatistics. As a simple illustration, we consider the case of the sample *coefficient of variation*, $V_n = S_n/\overline{X}_n$, where it is assumed that the population mean, μ, is nonnull (usually taken to be positive). Thus, V_n is a function of $(\overline{X}_n, S_n^2)$, and the above result directly yields the asymptotic normality of $\sqrt{n}(V_n - \xi)$, where $\xi = \sigma/\mu$. Sample correlation and regression coefficients are also notable examples of this type of statistics.

Thirdly, it should be clearly kept in mind that such *weak convergence results may not imply moment convergence*, i.e. the convergence of the mean, variance, and other moments of the statistics T_n or $g(T_n)$ to their asymptotic counterparts as specified by their asymptotic distributions. A simple example to this effect is the parameter θ, the reciprocal of the binomial (probability) parameter π, where T_n is the sample proportion, and $g(T_n) = T_n^{-1}$ is the natural plug-in estimator of θ. This model arises in the context of the well known **Capture–mark–release–recapture** (CMRR) procedure for estimating the size of a finite population; see, for example, Sen & Singer [21]. Since T_n can assume the value 0 with a positive probability $(1-\pi)^n$, no matter how large n is, $g(T_n)$ does not have any finite positive order moment. However, the asymptotic normality result pertains to $g(T_n)$ as long as $\pi > 0$. As a historical note it may be mentioned that during the 1940s and 1950s considerable attempts were made to obtain the exact **skewness** and **kurtosis** coefficients of sampling distributions of various statistics (or estimators) in showing that they are asymptotically null, so that their asymptotic distribution would be normal. While the *Fréchet–Shohat* theorem (based on moment convergence of all order) justifies such an approach, obviously, the convergence of the first four moments may not suffice. Moreover, this limits the scope to a more restricted class of statistics which have finite moments all of finite order. In (bio-)statistical applications, for example, for setting a (large-sample) confidence interval for a parameter or testing a null hypothesis on the same, all we need is the asymptotic distribution and estimates of the parameter(s) that appear in these laws. Thus, weak convergence results are generally enough, and moment convergence is usually not needed. Finally, in such applications we may like to have some deeper weak convergence results which can match with more complexities and also may accelerate the goodness of fit of the asymptotics for moderate to large samples. These are presented separately in the following three sections.

Weak Convergence: Conditional Distributions

In the context of *resampling plans*, such as **jackknifing** and **bootstrapping**, one encounters a somewhat different asymptotic situation which requires additional care. We illustrate this with the simple bootstrap methodology. Let $X_1, \ldots, X_n$ be n iid random variables drawn from a distribution function F, and let $T_n = T(X_1, \ldots, X_n)$ be a suitable statistic whose population counterpart is denoted by θ. In a variety of cases it may be possible to establish that under suitable regularity assumptions, the distribution function $G_n(\cdot)$ of $\sqrt{n}[T_n - \theta]$ converges to a limiting distribution function $G(\cdot)$ as n becomes large; however

this distribution function G may not be normal, or even if it is so, it may have a scale parameter γ which is an involved functional of F. Therefore, we may like to estimate G_n in a nonparametric manner. From $(X_1, \ldots, X_n)$, we draw with replacement a sample of n observations, and denote these by $X_1^*, \ldots, X_n^*$, respectively. Let $T_n^* = T(X_1^*, \ldots, X_n^*)$ be the bootstrap version of T_n, and let us denote by $Z_n^* = \sqrt{n}[T_n^* - T_n]$. Note that under the conditional law $\Pr\{X_i^* = X_k | X_1, \ldots, X_n\} = n^{-1}$, for all $i, k = 1, \ldots, n$, the exact (conditional) distribution of Z_n^* can be obtained by enumeration, though such a task becomes prohibitively laborious as n increases (check the growth of the number n^n). Moreover, this conditional (bootstrap law) is intended as an estimator of the unconditional law G_n. This naturally imposes some restraints on the type of G_n for which a passage from the conditional to the unconditional distribution is well lighted. For example, if G_n is not attracted by a normal limit, then this postulation may not be true. This objective often precludes small sample size cases, even when G_n has a normal limit. As such, when n is large, M (a large number of) repetitions of the bootstrapping yields conditionally independent and identically distributed copies of Z_n^*, and this set is used to estimate G_n as well as a measure of its scale parameter. A very similar case arises in multivariate nonparametrics where conventional rank statistics are not usually genuinely distribution-free even under suitable null hypotheses (of invariance), and hence, their permutation distribution (*see* **Randomization Tests**) (corresponding to the case of simple random **sampling without replacement** (SRSWOR)) is used to generate the (conditional) null distribution of such rank statistics. In such applications, too, the passage from the conditional to unconditional distributions is generally fortified for large samples when the asymptotic (multi-)normality can be incorporated in a suitable manner; we refer to Puri & Sen [15] for a detailed account of such permutational LST. In jackknifing, a similar SRSWOR scheme arises, and it rests on the permutational probability measure generated by the $n!$ equally likely permutations of the observations. In such a case, the classical weak convergence results may not directly hold, though under additional mild regularity assumptions, the passage from the conditional limit law to an unconditional one can be fortified. Usually the (multi-)normality of the conditional distribution and its asymptotic homoscedasticity suffice for the purpose. But that may exclude some important applications in practice. For example, if the limit distribution is (scale) mixed-normal, this convergence of conditional limit laws to their unconditional forms may not generally hold. A word of caution: contrary to the belief and heuristic practice of using such resampling schemes for moderate to small sample sizes as well, there is no sound methodological justification for such usages. In many cases, they may be misleading.

Weak Invariance Principles

Let $X_1, \ldots, X_n$ be n iid random variables with finite mean μ and variance σ^2. Set $S_k = \sum_{i \le k}(X_i - \mu), k \ge 1$, and let $S_0 = 0$. Then the CLT asserts that for large n, $S_n/\{\sqrt{n}\sigma\}$ has closely a standard normal distribution. Let us construct a stochastic process $W_n = \{W_n(t), t \in (0, 1)\}$, by letting $W_n(k/n) = \{S_k/\{\sigma\sqrt{n}\}, k = 0, 1, \ldots, n\}$ and completing the definition by linear interpolation on (0, 1). This way we map the *partial sum process* $\{S_k; k \le n\}$ into a stochastic process W_n with continuous sample paths on the unit interval (0, 1). Now let $W = \{W(t), t \in (0, 1)\}$ be a *Gaussian process* on the unit interval (0, 1), such that $\mathrm{E}W(t) = 0$ and $\mathrm{E}[W(s)W(t)] = \min(s, t), s, t \in (0, 1)$. Then W is termed a *standard* **Brownian motion process** on (0,1). As a generalization of the CLT, we have the following:

$$W_n \xrightarrow{\mathcal{D}} W, \quad \text{as } n \to \infty. \tag{10}$$

The implications of this weak convergence result are (i) the finite dimensional distributions of W_n converge to those of W, and (ii) like W, W_n is *tight* or relatively compact. The first result follows by using a multivariate version of the CLT, while (ii) can be established by using some maximal inequalities, and both accomplished under no extra regularity conditions. This result extends directly to martingales/reversed martingales and to the nonidentically distributed case as well.

A second weak invariance principle having a profound impact on LST in biostatistics is the following. Let $F_n(x) = n^{-1}\sum_{i=1}^n I(X_i \le x), x \in \mathbf{R}$, be the sample distribution function, and define a stochastic process $W_n^0 = \{W_n^0(t), t \in (0, 1)\}$ by letting $W_n^0(t) = \sqrt{n}[F_n(x) - F(x)]$, at $t = F(x)$, for $t \in (0, 1)$. Also, let $W^0 = \{W^0(t), t \in (0, 1)\}$ be a Gaussian function on (0, 1), such

that $E W^0(t) = 0$ and $E[W^0(s)W^0(t)] = \min(s,t) - st, s, t \in (0,1)$. W^0 is termed a *standard Brownian bridge* or a tied-down Brownian motion. Note that at $t = 0$ or 1, $W_n^0(t)$ is equal to 0 with probability 1, and hence the term *tied-down* has been affixed. Here also, we have for all continuous F,

$$W_n^0 \xrightarrow{\mathcal{D}} W^0, \quad \text{as } n \to \infty, \tag{11}$$

and the implications of this weak convergence result are the same as in (10). Extensions to higher dimensional distribution functions and more general functionals of the sample distribution functions have been considered at great depth. We refer to Jurečková & Sen [10], for some details. Some applications of these invariance principles will be discussed later in the article.

Variance Stabilizing Transformations

In a general setup, whenever $\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, the asymptotic variance σ^2 may depend on the unknown parameter θ; therefore, we write $\sigma^2 = h(\theta)$ and assume that the form of $h(\cdot)$ is known. To use the above result for drawing a confidence interval for θ or to test a suitable null hypothesis on θ, it may be more desirable to consider a transformation: $T_n \to g(T_n)$, such that $\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^2)$, where $g(\cdot)$ is a manageable function and c does not depend on θ. While in general such a transformation may not exist, but for the single parameter case there are some well-known cases where it has worked out well; these are therefore termed variance stabilizing transformations (*see* **Delta Method**). It follows from (9) that a sufficient condition for this to be achieved is that

$$g'(\theta) = c\{h(\theta)\}^{-1/2} \quad \text{or} \quad g(\theta) = c \int \{h(y)\}^{-1/2}\, dy. \tag{12}$$

For **binomial**, **Poisson**, normal variance and **correlation** coefficient parameters, (12) work out well, and furthermore, in all these cases, some small corrections have been incorporated, mostly on empirical grounds, to provide a faster rate of convergence of the asymptotic normality result: the statistical motivation, however, stems primarily from LST. We refer to Sen & Singer [21, Chapter 3] for details. There are, however, some impasses in the multiparameter case where the dependence pattern of the coordinate estimators may violate the applicability of the variance stabilizing transformation for their covariance terms. A classical example is the **multinomial** distribution where for each cell probability one may use the arcs in transformation to stabilize its asymptotic variance but then their covariances would still be dependent on the unknown cell probabilities.

Order Statistics and Empirical Distribution

Order statistics and empirical distribution functions are interrelated (one-to-one) in the classical univariate setup, and together they play a fundamental role in LST, particularly in *robust* as well as *nonparametric* inference problems. In biostatistics, in the active area of **survival analysis**, their role is overwhelming. The order statistics are neither independent nor identically distributed, even when the unordered collection relates to iid random variables. Hence LST pertaining to sums of independent random variables may not be directly applicable here. But with a reformulation in terms of indicator functions, most of these standard LST's can be adopted for order statistics, for sample quantiles, as well as extreme values. For example, for LLNs for sample quantiles, the Borel SLLN applies with little modification, while for the CLT, under the positivity and continuity of the density function at the population quantile, this approach via Bernoulli variables works out better, not only for a single quantile in a univariate setup but also for multiple quantiles in a general multivariate setup. More conveniently with adaptations from weak invariance principles for the empirical distributional processes, the related LST for order statistics and empirical distributions has emerged in a very elegant form. Interestingly enough, reversed martingales also play a very prominent role in this context. We refer to Sen & Singer [10, Chapter] for some details. In **robust** estimation, covering both parametric and nonparametric models, we often use *L-estimators*, which are linear combinations of functions of order statistics, and *M-estimators*, which are solutions of implicit equations involving suitable *score functions* and the empirical distribution function. Likewise, *R-estimators* of location and regression parameters are based on suitable

rank-order statistics which can be expressed as functionals of the empirical distributions. In this way we can conceive of a statistic $T_n = T(F_n)$ as a general functional of the empirical distribution function, F_n, and use LST pertaining to invariance principles for such processes, as outlined in (11). Naturally, the nature of $T(\cdot)$ will dictate the LST approach, and suitable differentiability properties of such functionals provide the necessary tools. A detailed treatment of this area of LST is beyond the scope of this article, but we refer the reader to Jurečková & Sen [10, Chapters 3 and 7], where an up-to-date and unified account has been provided. Hoeffding's [8] *U-statistics*, von Mises [23] statistical functionals and their (multisample) generalizations occupy a prominent place in nonparametrics, and they are abundant in biostatistics applications. Fortunately, they are statistical functionals and there are various martingale–reversed martingale representations for such statistics, discussed in detail in Sen [19], which pave the way for adoption of standard LST tools for the study of asymptotics for such statistics.

LST for MLE and BAN Estimators

In biostatistics, in actual applications, often, for easier interpretations and simpler statistical analysis, suitable parametric statistical models are postulated, and this approach naturally tilts the flavor to using optimal parametric statistical estimators (and tests) for the model parameters. *Maximum likelihood estimators* (MLE) are known to have various optimality properties, at least asymptotically, and in this depiction LST plays a basic role. In the case of the so-called **exponential family** *of densities*, granted *sufficiency* and *continuous differentiability* of the *likelihood function* (of the sample observations), the LST is generally based on the standard tools described earlier. However, for a density not belonging to such a class (namely the **Cauchy**, *Laplace*), the treatment of LST becomes more complex and involves additional regularity assumptions. Basically the approach is to explore a *quadratic approximation* for the likelihood ratio statistic in a suitable neighborhood of the true parameter point, and this in turn provides an asymptotic representation for the MLE in terms of the *likelihood score statistics* which yield the desired asymptotic normality, consistency, as well as asymptotic efficiency properties of the MLE $\hat{\theta}_n$ (in the regular case). Basically, if we denote the score statistics $(\partial/\partial\theta)\ \ln L_n(X_1, \ldots, X_n)$ by $U_n(\theta)$, and the *Fisher* **information** *per observation* by $I(\theta) = n^{-1}\mathrm{E}\{U_n^2(\theta)\}$, then under appropriate regularity conditions we have a first-order asymptotic representation:

$$\hat{\theta}_n - \theta = \{nI(\theta)\}^{-1}U_n(\theta) + o_p(n^{-1/2}), \qquad (13)$$

where $U_n(\theta)$ involves independent summands with zero mean and variance $I(\theta)$, and hence the CLT applies there. This leads to the following:

$$\sqrt{n}[\hat{\theta}_n - \theta] \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1}(\theta)), \qquad (14)$$

where by the classical Fréchet–**Cramér–Rao information** inequality, $[nI(\theta)]^{-1}$ is the lower bound to the mean square error of an unbiased estimator of θ; this yields the asymptotic efficiency (and asymptotic unbiasedness) of the MLE. A similar situation holds in the multiparameter case. The regularity conditions, classically known as the *Cramér conditions*, have gone through some evolution during the past 50 years. Although least stringent regularity conditions may be formulated as in LeCam [12], from a biostatistical applications point of view, a somewhat intermediate set of conditions hinging on the following compactness condition of the second derivative of the log density function provides a much simpler and more easily verifiable scenario. As $\delta(>0)$ approaches 0,

$$\mathrm{E}\left\{\sup\left(\left|\left(\frac{\partial^2}{\partial\theta^2}\right)\ln f(X, \theta + h) - \left(\frac{\partial^2}{\partial\theta^2}\right)\ln f(X, \theta)\right| : |h| < \delta\right\} \to 0; \quad (15)$$

Cramér's conditions involve the third derivative instead of this compactness of the second derivative. For the exponential family of densities, the above condition follows from the continuity of the parametric functions, and in many other cases it can be verified by standard manipulations; we refer to Sen & Singer [21, Chapter 5] for details.

It is quite pertinent here to make some comments about LST for the MLE. First, in a nonregular case, the MLE may not be asymptotically normal, and may even lose its asymptotic efficiency property. Secondly, the MLE may not be the only estimator that is asymptotically efficient in the above sense.

There are alternate estimators which may often share the asymptotic normality and efficiency properties along with the MLE; such estimators are termed best asymptotically normal **(BAN) estimators**. In the context of *categorical data models*, such BAN estimators based on the *minimum chi-square* and *modified minimum chi-square* criteria have been extensively studied in the literature (see for example, Agresti [1] and Sen & Singer [21] where detailed references are also cited); often, they are computationally less cumbersome than the MLE. Thirdly, the MLE are generally not robust to plausible model departures, and in that respect, alternative estimators, particularly *adaptive* estimators, may combine the BAN property with robustness to a greater extent. Finally, with the increase in the number of parameters, the performance characteristics of the MLE may deteriorate, and they may even become inconsistent or inefficient; the classical Neyman–Scott problem with a large number of nuisance parameters is a glaring example. Hence, modifications are often made to enhance the efficiency of the MLE. Among various such modifications, we refer to (*partial*) PMLE based on suitable partial likelihood functions [4], and quasi- and profile MLE based on **quasi-** and **profile likelihood** functions, which in a semiparametric context will be treated briefly later in the article.

LST and WLSE

Linear models (*see* **General Linear Model**) and *linear statistical inference* are household words in biostatistics. With the primary objective of interacting with researchers in biomedical and environmental sciences, in biostatistics it is customary to pose simple linear models that can be easily interpreted to collaborative scientists and can thereby be adapted to conventional linear statistical inference tools. Yet, in many cases the basic assumptions underlying such conventional procedures may not be tenable, and hence suitable modifications are often necessary to cope with the valid and efficient use of statistical inference tools. In biostatistics often we have nonnegative response variables where suitable transformations are used to induce linearity of the model to a greater extent, albeit at the cost of having nonnormal distributions (or vice versa). Therefore in linear statistical inference the basic assumption of normality of the errors may not be always tenable, and without this, the MLE based on the normality assumption may lose its appeal of validity and efficiency. The classical least squares estimators (LSE) and (large-sample) tests based on them occupy a focal point in this situation, and for such linear statistics, standard LST applies well. Weighted (WLSE) and generalized (GLSE) least squares estimators are the hybrids of the LSE that suit such nonstandard applications to a greater extent. The *heteroscedastic* linear model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $E(\mathbf{e}) = \mathbf{0}$, $V(\mathbf{e}) = \text{diag}\,(\sigma_1^2, \ldots, \sigma_n^2)$, provides a typical application of the WLSE when the σ_j^2 are not equal but known up to an unknown scalar constant (*see* **Scedasticity**). Thus, if we take $\sigma_j^2 = c_j\sigma^2$, $j \geq 1$, where the c_j are known constants, not all equal, while σ^2 is unknown, and if we denote the ith row of $\mathbf{X}$ by $\mathbf{x}_i'$, $i = 1, \ldots, n$, then we can consider the weighted sum of squares due to residuals:

$$\sum_{i=1}^{n} c_i^{-1}\{Y_i - \mathbf{x}_i'\boldsymbol{\beta}\}^2 \tag{16}$$

and minimize this with respect to $\boldsymbol{\beta}$. This leads to the WLSE of $\boldsymbol{\beta}$. This procedure extends readily to the multivariate case where the $\mathbf{Y}_i$ are p vectors for some $p \geq 1$, provided the covariance matrices of the associated error vectors satisfy a similar heteroscedastic condition. In that setup it is generally referred to as the (generalized) GLSE, and if such a matrix is diagonal, it is termed the WLSE. The GLSE or WLSE are linear estimators, and hence the LLNs, CLTs, and other standard asymptotics apply here under some extra mild conditions on the c_i. However, in biostatistical applications, such as in **loglinear models** for categorical data, the exact variance–covariances of the transformed response statistics are not known, and are estimated from the sample itself. That brings the relevance of LST into a broader perspective. Estimated variance–covariances are used in the above minimization problem, often requiring an iterative procedure to update these estimates along with the estimates of the main parameters of interest. The two-step (Aitken) estimator belongs to this class. For details of related LST we refer to Sen & Singer [21, Chapter 7]. Other related procedures for linear (as well as location) models include the so called **trimmed** *LSE* (TLSE) and *regression quantiles* (*see* **Quantile Regression**); these are discussed in detail in Jurečková & Sen [10], and the related LST runs parallel to the case of WLSE.

From a robustness prospect, however, such TSLE or regression quantiles are more appealing than the classical LSE.

LST of Statistical Tests

For testing a simple null hypothesis H_0 against a simple alternative H_1, the **Neyman–Pearson** *Fundamental Lemma* characterizes the *likelihood ratio test* (LRT) as most powerful, and this extends to *uniformly* **most powerful** (UMP) tests for *one-sided* alternatives. However, in a general multiparameter case with possibly **nuisance parameters**, one has typically composite null and composite alternative hypotheses. Here an exact (similar) test may not always exist, and even if one exists, it may be difficult to characterize one that will be uniformly best. For this reason, characterizations of optimality of statistical tests have often been made in an asymptotic framework, and there are competing tests sharing such properties in some way or other. Among such classes of tests, the following (parametric) deserve special mention: (i) likelihood ratio test, (ii) Rao's score test, and (iii) Wald's test; we refer to Rao [16] for a nice comparative account. The LRT is based on two sets of MLE, computed under H_0 and H_1, respectively, providing the ratio of the two maximized likelihood functions. LST pertaining to such LRTs, covering their null as well as alternative hypothesis distributions, is interlinked with the LST for the MLE, and hence they involve parallel regularity assumptions. Rao's score test, on the other hand, is based on the likelihood score statistics and their modifications, so that their asymptotics can be studied directly by using the standard LST tools. Computationally, Rao's score test is usually less cumbersome than the LRT. Wald's test is directly based on the MLE and the parametric constraints imposed by the hypotheses, and hence the general asymptotics for the MLE provide the access for parallel results for this type of tests. For local alternatives all the three types of tests share common asymptotic properties (*see* **Locally Most Powerful Tests**), although for nonlocal alternatives the LRT may have some advantages in a special way of interpretation [9]. Here also, on robustness considerations, such likelihood-based tests may not be very suitable. Moreover, for *restricted alternatives*, such as one-sided multiparameter hypotheses, such tests may have quite complicated forms and may even lose their asymptotic optimality properties to a greater extent (*see* **Isotonic Inference**). Roy's [17] **union–intersection principle** has added a lot of flexibility to this testing scenario, and their asymptotics have been studied under similar regularity assumptions. From robustness and nonparametric considerations, alternative tests based on L-, M- and R-estimators and suitable U-statistics have been extensively studied in the literature; we refer to Jurečková & Sen [10, Chapter 10] for a good account of these. In these developments, naturally, the asymptotics for such estimators play a basic role. In passing, we should also comment on **sequential** and *multistage* tests which have been considered in the literature. In this context, the classical Wald [25] *sequential probability ratio test* (SPRT) and its generalizations are all aimed at capturing some optimality properties in an interpretable manner. In the general multiparameter (composite hypothesis testing) case, again such optimality properties in an exact sense are hard to establish, and there is a good deal of asymptotics in the interpretation and derivation of such plausible optimality properties. The domain is by no means restricted to classical parametric setups, and nonparametric as well as robust procedures have been developed along the same vein. These procedures exploit the weak convergence and invariance principles introduced earlier and inherit the robustness aspects of the estimators or test statistics on which they are based. For details, we refer to Sen [19]. Group sequential procedures and related repeated significance testing (RST) procedures in **clinical trials** and biomedical studies have received considerable attention during the past two decades (*see* **Data and Safety Monitoring**). In this domain, too, the development of the methodology inherits a lot of asymptotics, and LST plays a vital role. In these developments, exact computations of boundary crossing probabilities, even for binomial or normal distributions, may become prohibitively laborious, if not impossible, and weak convergence of the encountered stochastic processes to suitable Gaussian functions (e.g. Brownian motion or Brownian bridge) provides the adaptability of standard results for Gaussian processes, and a general account of these developments is given in Sen [19, Chapters 9 and 10]. There are some other variations of such schemes, and we shall refer to some of them later in the article.

Semiparametric and Generalized Linear Models

Linear models are abundant in biostatistics, and yet in many applications the *normality* of the error components, their *homoscedasticity*, or even the basic *linearity of the model* may not be tenable. The classical MLE in the normal case agree with the *least squares estimators* (LSE), and for such linear estimators, standard LST can be adopted without many problems [21, Chapter 7]. Nevertheless, the optimality of the LSE may no longer be true when there are model departures. Therefore alternate models have been introduced to deemphasize the three basic assumptions underlying the normal theory MLE or the LSE, and in that way alternative classes of estimators have evolved.

Box–Cox Type Transformations

In biomedical applications the response variables are mostly nonnegative with (highly) positively skewed distributions. Although in such a case asymptotic normality of the LSE can be justified methodologically, in applications it may require an enormously large sample size. For this reason, logarithm, square-root, or cube-root transformations (*see* **Power Transformations**) are used to induce more symmetry in these response distributions so that moderate sample asymptotics can be justified to a greater extent. On the other hand, if the original model is closely linear, such nonlinear transformations can affect the regression relation considerably. Thus, one may require some **nonlinear regression** *models* to validate such transformations in practice. Either way, the LSE may not retain their normal-theory optimality, even asymptotically, although consistency and asymptotic normality would be retained under fairly general conditions [21, Chapter 7].

Generalized Linear Models

In **biological assays** and many survival analysis models, a response variable may be *quantal* (i.e. all or nothing) in nature (*see* **Quantal Response Models**). For such dichotomous (or even polychotomous) response variables, standard LSE may not work out well. *Logit* (**logistic regression**), *probit* and other models in bioassay are the precursors of **generalized linear models** (GLM). A more unified approach to such GLMs is outlined in McCullagh & Nelder [14]; their treatment addresses mostly the finite sample (or exact) methodology, and the findings are quite relevant to a general exponential family of densities. Nevertheless, in biostatistical applications such exact GLMs may not be tenable in all cases, and often (weighted) WLSE methodology is incorporated to facilitate suitable large-sample solutions (see [21, Chapter 7]). Such GLMs yield suitable *estimating equations* (EEs) (*see* **Estimating Functions**) which provide the estimators (mostly) as implicit solutions; this way the situation is similar to the case of the MLE. However, to cope with variations from most ideal situations, such EEs are replaced by suitable **generalized estimating equations** (GEE), and in their asymptotic treatment one needs additional regularity assumptions and manipulations too; see [21, Chapter 7]. Viewed from a practical perspective in a biostatistics context, such as in bioassays, dosimetric and mechanistic models in *toxicological* studies, the doses may be subject to *measurement errors* (*see* **Errors in Variables**) or latent effects, so that even if a simple GLM were pertinent to the basic dose–response pattern, such perturbations may cause great damage to their adoption without reservation. In this manner, one ignores the GLM methodology and has to take recourse in alternative LST where robustness and nonparametrics may dominate the scenario.

Nonparametric Linear Models

While assuming the linearity or additivity of the basic model, no specific distributional assumption is made on the response variables. An extensive literature relates to L-, M- and R- procedures in a variety of linear models, and an up-to-date treatment of the related asymptotics is contained in Jurečková & Sen [10]. Such procedures are generally more robust, consistent, and have asymptotic normality properties. Within this bigger class, one can also have suitable *adaptive* estimators which are asymptotically efficient and robust as well.

Semiparametric Models

The **Cox model** [3] or **proportional hazards model** (PHM) is a very simple illustration of a semiparametric model. In a simple two-sample model, if F and G are the respective distribution functions,

and we denote the corresponding survival functions by $\overline{F}(x) = 1 - F(x)$ and $\overline{G}(x) = 1 - G(x), x \in \mathbf{R}$, then in a Lehmann [13] model (*see* **Lehmann Alternatives**), we set $\overline{G}(x) = [\overline{F}(x)]^c$, for some $c > 0$. The null hypothesis of the homogeneity of F and G then reduces to $c = 1$. If the distribution functions are absolutely continuous with densities f and g, respectively, then we define equivalently the **hazard** *functions* $h_F(x)$ and $h_G(x)$ as $f(x)/\overline{F}(x)$ and $g(x)/\overline{G}(x)$, respectively. Then the Lehmann model can be put equivalently as $h_G(x) = ch_F(x)$, for all x, i.e. the two hazard functions are proportional. Motivated by this simple observation, Cox [3] considered a general situation where conditionally on a set of concomitant variates, say, $\mathbf{z}$, the hazard function for the primary variate, denoted by $h(y|\mathbf{z})$, is assumed to satisfy the following model:

$$h(y|z) = h_0(y) \exp\{\boldsymbol{\beta}'\mathbf{z}\}, \tag{17}$$

where the nonnegative $h_0(y)$, the baseline hazard rate, is independent of the concomitant variates and is of arbitrary form (i.e. nonparametric in nature), and the regression on the concomitant variates is of a specified parametric form. This also leads to the following:

$$\ln h(y|\mathbf{z}) = \ln h_0(y) + \boldsymbol{\beta}'\mathbf{z}; \tag{18}$$

in the literature this is known as the *hazard regression*. In either setup, note that $h_0(y)$ is a functional while β is a finite dimensional (regression) parameter. For this reason, this is referred to as a semiparametric model. Typically, in a general setup one may have a functional parameter space, and in that way the MLE or other conventional estimators may lose their efficacy, and often, consistency properties too. It may be possible in some cases to reparameterize in such a way that the parameters of interest constitute a finite-dimensional vector, while the nuisance parameter space may be very large. In this setup, often a conditional approach leads to a **partial likelihood** function whereby the finite-dimensional parameters of interest can be estimated consistently by the (partial) PMLE and with reasonable efficacy. Martingale methods play a basic role in the related asymptotics, and in this context, *counting processes* have evolved to be of prime interest in the study of general asymptotics; we refer to Andersen et al. [2] for a nice account of related asymptotics.

Nonparametric Regression and Smoothing Techniques

The past 20 years have witnessed a phenomenal growth of research literature in this domain, and these developments are of considerable use in biostatistics. Both the *kernel* and *nearest neighbor* methods are popular in this context (*see* **Density Estimation**). In terms of model flexibility, such models are the most desirable ones. However, in terms of precision of derived estimators, such a model may have the opposite flavor. Compared to the usual $\sqrt{n}$-consistency of the classical estimators in the parametric or semiparametric models, here one has n^{λ}-consistency for some positive $\lambda < 1/2$. Moreover, the asymptotic bias and asymptotic standard error may be of comparable order of magnitude, and hence *adaptive bandwidth* selection procedures are often prescribed to achieve asymptotic optimality within this class. Generally, much larger sample size is required for the adaptibility of asymptotics in the nonparametric regression case than in other models; we refer to Thompson & Tapia [22] for a treatise of **nonparametric regression** and function estimation problems.

LST for Time-sequential Schemes

This is one of the most important areas of current research activities in biostatistics, and LST has a fundamental role in this field. In clinical trials or medical investigations, generally one obtains data sets accumulating over time, so that statistical conclusions are drawn at the termination of the study. On the contrary, most of these studies relate to comparisons of different treatments or subgroups, and involve human beings. Thus, for ethical reasons, it is often advised that if there is any real difference in the response patterns for the various subgroups, then the trial should be able to detect it as early as possible, and the better treatment be made available to the entire set of subjects for their better prospects. On the other hand, lacking any real difference, the trial, if conducted up to the end of the planned duration, may contain valuable information for other scientific studies as well. This motivates the need for *interim analysis* in such clinical trials, and these may be made either on a periodic (namely fixed calendar-month/year interval) basis or on a monitoring basis resulting in the so called time-sequential procedures. The main points of difference between the classical

sequential and time-sequential procedures are the following:

1. The number of subjects to be included in the study is prefixed in a time-sequential scheme, but is itself a random variable in a sequential one. Thus the formulation of an average sample number (ASN) is quite different in the two schemes.
2. The observations in a sequential scheme are typically iid, whereas in a time-sequential one they typically represent the ordered failure points along with other concomitant variables, and hence the iid clause generally is not tenable.
3. The emphasis on type I and type II errors in a sequential test is somewhat different from that in a time-sequential one.
4. **Censoring** (of various types) is a typical phenomenon in a time-sequential scheme, and to a greater extent the statistical modeling and analysis depend on such deviations.
5. Typically, in view of point 4, nonparametrics and semiparametrics play a more dominant role in time-sequential schemes than in the classical sequential schemes, where the probability (or likelihood) ratio statistics have a more visible parametric flavor.

Nonparametric and semiparametric procedures for time-sequential schemes have been studied extensively in the literature during the past two decades. The basic foundation was laid down by the development of martingale methodology for various rank statistics [19], as well as for counting processes related to such stochastic events [2]. In this context the classical LST may not be directly applicable; nevertheless, they are quite pertinent and justifiable through the modifications based on adoption of martingale theory. We conclude this discussion with a brief introduction to LST pertinent to the **Kaplan–Meier** [11] *product-limit* (PL) estimator of the survival function under random censoring. In random censoring schemes the set of censoring variables $T_1, \ldots, T_n$ are iid according to a distribution function G, and T_i and X_i are stochastically independent for every i; note that the X_i are iid with a survival function $\overline{F}$. The observable random elements are $Z_i = \min(X_i, T_i)$ and $I_i = I(Z_i = X_i), i = 1, \ldots, n$. Define $N_n(t) = \sum_{i \le n} I(Z_i > t), t$ real, and set $\alpha_i(t) = I(Z_i \le t, I_i = 1), i \ge 1, t$ real, and let $\tau_n = \max\{Z_i : i \le n\}$. Then the PL-estimator of $\overline{F}$ is given by

$$\overline{P}_n(t) = \prod_{i=1}^{n} \left\{ \frac{N_n(Z_i)}{N_n(Z_i)+1} \right\}^{\alpha_i(t)} I(t \le \tau_n)$$

$$= \prod_{\{i: Z_i \le t\}} \left\{ \frac{n\overline{H}_n(Z_i)}{n\overline{H}_n(Z_i)+1} \right\}^{I_i}, \qquad (19)$$

where $\overline{H}_n(t) = n^{-1} N_n(t), t$ real. Thus, this estimator (a stochastic process) can be viewed as a functional of the counting process $\{N_n(t), t \text{ real}\}$, and hence LST relating to such counting processes can be imported here to study the asymptotic properties of the PL-estimator. Alternatively, suitable martingale characterizations of the PL-estimator can also be incorporated in the study of related LST. We refer to Andersen et al. [2] for details, albeit at a much higher level of mathematical sophistication. Generally test statistics or estimators in time-sequential schemes are functionals of the PL-estimator (in the censored case) or the original empirical survival function (in the uncensored case), having some sort of time-sequential flavor, and hence suitable stochastic processes relating to such functionals can be incorporated to formulate the *stopping* and *decision rules*. For example, in survival analysis, the *mean residual life* (MRL) is an important tool to measure the effectiveness of treatment protocols (*see* **Life Table**). Corresponding to the population measure

$$\mu(x) = \{\overline{F}(x)\}^{-1} \int_x^\infty \overline{F}(y)\, dy, \qquad (20)$$

the sample measure is defined by $\hat{\mu}_n = \{\overline{P}_n(x)\}^{-1} \int_x^\infty \overline{P}_n(y)\, dy$, and one may like to study the weak or strong consistency and asymptotic normality of $\hat{\mu}_n(x)$ for a given x, and more generally for a range of values of x. Weak convergence of such stochastic processes naturally provides the key to subsequent developments, and a systematic account of this type of LST in the context of clinical trials is given in Sen [20].

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[3] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[5] Dvoretzky, A. (1971). Asymptotic normality for sums of dependent random variables, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 2. University of California Press, Berkeley, pp. 513–535.

[6] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

[7] Ferguson, T. (1996). *Large Sample Theory*. Chapman & Hall, London.

[8] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *Annals of Mathematical Statistics* **19**, 293–325.

[9] Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions (with discussion), *Annals of Mathematical Statistics* **36**, 369–408.

[10] Jurečková, J. & Sen, P.K. (1996). *Robust Statistical Procedures*. Wiley, New York.

[11] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481, 562–563.

[12] LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

[13] Lehmann, E.L. (1953). The power of rank tests, *Annals of Mathematical Statistics* **24**, 23–43.

[14] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[15] Puri, M.L. & Sen, P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[16] Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

[17] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[18] Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau, *Journal of the American Statistical Association* **63**, 1379–1389.

[19] Sen, P.K. (1981). *Sequential Nonparametrics*. Wiley, New York.

[20] Sen, P.K. (1985). *Theory and Applications of Sequential Nonparametrics*, CBMS/NSF Ser. 49. SIAM, Philadelphia.

[21] Sen, P.K. & Singer, J.M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall, London.

[22] Thompson, J.R. & Tapia, R.A. (1990). *Nonparametric Function Estimation, Modeling and Simulation*. SIAM, Philadelphia.

[23] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions, *Annals of Mathematical Statistics* **18**, 309–348.

[24] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society* **54**, 426–482.

[25] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

(*See also* **Limit Theorems**)

PRANAB K. SEN

Latent Class Analysis

Latent class analysis is a discrete variable analog of **factor analysis**. Latent class analysis was originally developed [11, 16] to investigate the classification of subjects according to an underlying categorical trait, such as an attitude or psychological state, that is not directly observable. Membership in a particular class of the underlying variable is estimated from a subject's responses to a set of categorical items. Overviews of latent class modeling are given by Lazarsfeld & Henry [17], Andersen [3], Henry [13], McCutcheon [19], and Clogg [6].

Examples

The following example is a simplified version of the items considered by Rimer [20] in a study of methods for promoting smoking cessation. Subjects were asked to agree or disagree with a series of items, a subset of which are:

1. Smoking cigarettes relieves your tension.
2. Smoking helps you concentrate and do better work.
3. You are more relaxed and more pleasant when smoking.
4. You like the image of yourself as a smoker.

In principle, responses to these items reveal an underlying attitude towards smoking, perhaps reflecting a subject's "resistance" to quitting smoking. The trait of "resistance" could be dichotomized into the categories: not resistant or resistant. The observed items, 1–4 above, are called the manifest variables, while the underlying trait is the latent variable. In the example, both the manifest variables and the latent variable are dichotomous, but latent class models can be applied more generally, with manifest and latent variables that are ordinal or polytomous categorical variables (*see* **Ordered Categorical Data**; **Polytomous Data**).

A fundamental assumption in latent class modeling is "local independence", which states that given the latent class membership, the manifest variables are conditionally independent. A numerical example serves to illustrate this assumption. Consider a population that can be cross-classified according to the manifest variables, A and B, in the proportions displayed in Table 1. Suppose that the population can be divided into equal proportions by a binary latent variable, and that within levels of the latent variable the population can be cross-classified according to the manifest variables A and B as displayed in Table 2. The values in the cells of the table represent the conditional probabilities associated with A and B, given the level of Z. Despite the marginal association between variables A and B displayed in Table 1, the variables A and B are conditionally independent, given the level of the latent variable.

In practice, displays such as Table 2 cannot be constructed because the latent variable is unobservable directly. In many cases, however, the existence of a latent variable can be derived from theoretical models of attitudes, behavior, or psychology. Latent class analysis can be used to investigate the degree to which inferences about the unobservable latent trait can be derived from the manifest or observed

Table 1 Cross-classification of two manifest variables

		A		
		Agree	Disagree	Total
B	Agree	0.4	0.2	0.6
	Disagree	0.2	0.2	0.4
	Total	0.6	0.4	1

Table 2 Illustration of local independence of manifest variables (A and B) given the latent variable, Z

$Z = 1$

		B		
		Agree	Disagree	Total
A	Agree	0.64	0.16	0.80
	Disagree	0.16	0.04	0.20
	Total	0.80	0.20	1

$Z = 0$

		B		
		Agree	Disagree	Total
A	Agree	0.16	0.24	0.40
	Disagree	0.24	0.36	0.60
	Total	0.40	0.60	1

variables. The next section provides a mathematical formulation of latent class analysis.

Mathematical Model

Suppose that each of n subjects is observed on K categorical manifest variables, $\mathbf{Y} = (Y_1, \ldots, Y_K)$, with each variable taking on one of C categories. The cells of the K-way cross-classification table are indexed by $\mathbf{y} = (y_1, \ldots, y_k)$, with $n_{\mathbf{y}}$ denoting the observed number of subjects and $\pi_{\mathbf{y}}$ denoting the probability associated with response profile $\mathbf{y}$. The cell frequencies are assumed to have a **multinomial distribution** with $\mathrm{E}(n_{\mathbf{y}}) = n\pi_{\mathbf{y}}$ and $\Sigma_{\mathbf{y}}\pi_{\mathbf{y}} = 1$. The latent variable, Z, is assumed to take on one of T classes, with θ_z denoting the proportion of the population in class $z, z = 1, \ldots, T, \sum_{z=1}^{T} \theta_z = 1$. In the segment of the population in latent class $Z = z$, the proportion of the population classified into the cell indexed by $\mathbf{y}$ is denoted $\pi_{\mathbf{y}|t} = \Pr(\mathbf{Y} = \mathbf{y}|T = t)$. The assumption of local independence states that given the latent class membership, the manifest variables are conditionally independent:

$$\Pr[\mathbf{Y} = \mathbf{y}|Z = z] = \prod_{i=1}^{K} \Pr[Y_i = y_i|Z = z]. \quad (1)$$

Estimates of the conditional response probabilities, $\Pr[Y_i = y_i|Z = z], i = 1, \ldots, K$, and the latent class proportions, $\theta_z, z = 1, \ldots, T$, are derived from the observed counts, $n_{\mathbf{y}}$, through the equations $\mathrm{E}(n_{\mathbf{y}}) = n\pi_{\mathbf{y}}$ and

$$\pi_{\mathbf{y}} = \sum_{z=1}^{T} \prod_{i=1}^{K} \Pr[Y_i = y_i|Z = z]\theta_z. \quad (2)$$

The latent class model can also be viewed as a **loglinear model** [10, 12] for the expected counts in a $(K + 1)$-way cross-classification of the K manifest variables $(Y_1, Y_2, \ldots, Y_K)$ and the latent variable, Z. Denoting the expected counts by $m_{y,z}$, and with the usual constraints on the parameters of the loglinear model [1], the loglinear model

$$\ln m_{y,z} = \mu + \lambda_z^Z + \lambda_{y_1}^{Y_1} + \cdots + \lambda_{y_K}^{Y_K} + \lambda_{y_1,z}^{Y_1 Z} + \cdots + \lambda_{y_K,z}^{Y_K Z} \quad (3)$$

expresses conditional independence of the manifest variables, $\mathbf{Y}$, given the latent class. Although the cell frequencies in the $(K + 1)$-way classification are not observed, estimates in the loglinear model (3), are derived from the observed frequencies in the K-way cross-classification of the manifest variables, n_y.

Maximum likelihood is the most widely used method for estimating the parameters of the latent class model. Goodman [10] proposed an iterative algorithm for obtaining maximum likelihood estimates; the algorithm is an example of a general procedure now known as the **EM algorithm** [8]. Once the maximum likelihood estimates, $\hat{m}_{y,z}$, have been computed, the goodness-of-fit of the latent class model can be tested. The most commonly used statistics for testing goodness-of-fit are the generalized **likelihood ratio test** statistic,

$$G^2 = 2\sum_{\mathbf{y}} n_y \ln\left(\frac{n_y}{\hat{m}_y}\right),$$

and the Pearson X^2 statistic,

$$X^2 = \sum_{y} \frac{(n_y - \hat{m}_y)^2}{\hat{m}_y},$$

where $\hat{m}_y = \Sigma_z \hat{m}_{y,z}$ (*see* **Chi-square Tests**). Under the null hypothesis that the latent class model fits, the statistics are asymptotically distributed as a χ^2 random variable, with degrees of freedom equal to the number of cells in the table cross-classifying the manifest variables, minus the number of parameters being estimated in (3). With K manifest variables, each representing C categories, and with one latent trait having T classes, the degrees of freedom for testing the goodness-of-fit of the model equals $C^K - T[1 + K(C - 1)]$. A comparison of these statistics is given in [14].

Despite the similarity in form to standard loglinear model analysis, fitting a latent class model involves the additional issue of the **identifiability** of parameters. For example, with K binary manifest variables $(C = 2)$ and T latent classes, for $T > 2^K/(1 + K)$, the number of parameters in the model exceeds the number of "observations", the number of cells in the cross-classification of the manifest variables. Having more observations than parameters is a necessary condition for all parameters to be identifiable, but it is not sufficient. The identifiability of parameters is discussed in [10] and [18].

Extensions and Other Applications

Clogg & Goodman [7] extended the single population latent class analysis to simultaneous modeling of latent classes across several populations. Latent class methods specific to ordinal manifest variables were considered by Clogg [5]. In addition to its original applications in the study of attitudes, latent class modeling has been applied to the study of inter-rater reliability [2, 23] (*see* **Observer Reliability and Agreement**), survey response errors [4], incomplete data [9, 24], chronic disease epidemiology [15], medical diagnosis [21], and repeated measurements (*see* **Longitudinal Data Analysis, Overview**) [22].

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Agresti, A. & Lang, J. (1993). Quasi-symmetric latent class models, with application to rater agreement, *Biometrics* **49**, 131–139.

[3] Andersen, E.B. (1982). Latent structure analysis: a survey, *Scandinavian Journal of Statistics* **9**, 1–12.

[4] Bye, B. & Schechter, E. (1986). A latent Markov model approach to the estimation of response errors in multiwave panel data, *Journal of the American Statistical Association* **81**, 375–380.

[5] Clogg, C. (1979). Some latent structure models for the analysis of Likert-type data, *Social Science Research* **8**, 297–301.

[6] Clogg, C. (1992). The impact of sociological methodology on statistical methodology (with discussion), *Statistical Science* **7**, 183–196.

[7] Clogg, C. & Goodman, L. (1984). Latent structure analysis of a set of multidimensional contingency tables, *Journal of the American Statistical Association* **79**, 762–771.

[8] Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[9] Espeland, M. & Handelman, S. (1989). Using latent class models to characterize and assess relative error in discrete measurements, *Biometrics* **45**, 587–599.

[10] Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika* **61**, 215–231.

[11] Green, B.F. (1952). Latent structure analysis and its relation to factor analysis, *Journal of the American Statistical Association* **47**, 71–76.

[12] Hagenaars, J. (1993). *Log-Linear Models with Latent Variables*. Sage, Newbury Park.

[13] Henry, N. (1983). Latent structure analysis in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz and N.L. Johnson, eds. Wiley, New York, pp. 497–504.

[14] Holt, J. & Macready, G. (1988). Comparison of maximum likelihood and Pearson chi-square statistics for assessing latent class models, in *American Statistical Association 1988 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 167–171.

[15] Kaldor, J. & Clayton, D. (1985). Latent class analysis in chronic disease epidemiology, *Statistics in Medicine* **4**, 327–335.

[16] Lazarsfeld, P.F. (1950). The logical and mathematical foundations of latent structure analysis, in *Measurement and Prediction*, S.A. Stouffer et al., eds. Princeton University Press, Princeton.

[17] Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*. Houghton-Mifflin, Boston.

[18] Lindsay, B., Clogg, C. & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis, *Journal of the American Statistical Association* **86**, 96–107.

[19] McCutcheon, A. (1987). *Latent Class Analysis*. Sage, Newbury Park.

[20] Rimer, B. (1993). Enhancing Cancer Control in a Community Health Center, *R01 CA59734-03*. National Cancer Institute.

[21] Rindskopf, D. & Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis, *Statistics in Medicine* **5**, 21–27.

[22] Skene, A. & White, S. (1992). A latent class model for repeated measurements experiments, *Statistics in Medicine* **11**, 2111–2122.

[23] Uebersax, J. & Grove, W. (1993). A latent trait finite mixture model for the analysis of rating agreement, *Biometrics* **49**, 823–835.

[24] Winship, C. & Mare, R. (1989). Loglinear models with missing data: a latent class approach, *Sociological Methodology* **7**, 331–367.

(*See also* **Contingency Table**; **Rasch Models**)

MARK R. CONAWAY

Latent Period

Latency or *latent period* is defined as the time interval between the initiation time, say t_0, of a disease process and the time, say t_1, of the first occurrence of a specifically defined manifestation of the disease. For infectious diseases (*see* **Communicable Diseases**), t_0 is the time of infection by the infectious agent and the manifestation may either be a specific serologic marker, or a laboratory abnormality, or a symptom [31]. If the manifestation is the occurrence of a symptom, then the latent period is the same as the **incubation period**, which is the term usually used by statisticians for infectious diseases (e.g. Alcabes [1]). In the case of cancer epidemiology, t_0 is the time of initial exposure to a carcinogen (cancer initiation) and t_1 the time of the first clinical occurrence of the disease [3, 14]. For example, the initial exposure may be the time of exposure to **radiation** or the time of exposure to a chemical carcinogen, and the first clinical occurrence may be detected by a biological marker for cancer or by clinical evidence of a tumor [15]. For A-bomb survivors such as those from Hiroshima or Nagasaki, Japan, t_0 is thus the actual time of explosion of the bomb whereas t_1 is the time the disease first appears.

Other Definitions

For infectious diseases, Bailey [5] and Anderson & May [2] have used "the time to first become infectious" as the specified manifestation so that they define the latent period of the disease as the time interval from the point of infection to the beginning of the state of infectiousness of the infected host. This latter definition is not necessarily synchronous with the incubation period except in cases (e.g. yellow fever) in which both the average intervals from the point of infection to the infectiousness of the host and from the point of infection to the onset of a symptom are very short. For many infectious diseases caused by parasites, a distinction can usually be made between infection according to some laboratory criteria and symptoms of illness [2]. For infectious diseases caused by viruses and bacteria, however, such a distinction may be difficult; furthermore, for some viral diseases such as smallpox and yellow fever, an infected individual may be immune to the disease so that illness may never occur in some individuals [2, 30].

For exposure to a carcinogen, distinctions have been made between the biologic latent period and the epidemiologic latent period. For exposure to radiation, such as in A-bomb survivors, the biologic latent period is defined in [39] as the interval during which an elevation of the risk of the disease occurs between the exposed and nonexposed individuals (see Example 3 in the next section for illustration), whereas the epidemiologic latent period is defined in [39] as the interval between the first exposure and the time of death from the cause of interest. For exposure to a chemical carcinogen, the beginning of the biologic latent period is the time that a DNA adduct of the carcinogen first appears because carcinogenesis starts with the interaction between the DNA adduct of the carcinogen and the genome of the host [18, 37]. The endpoint of the biologic latent period is the time of first occurrence of a cancer tumor cell; see [36]. For the epidemiologic latent period, the initial time is the time of first exposure to the carcinogen, whereas the endpoint is the time of first appearance of a detectable cancer tumor. It is shown in [18] that it is not the exposed dose but the dose of the DNA adduct of the agents that gives a linear **dose–response** curve for small doses; furthermore, detectable cancer tumors arise by clonal expansion from cancer tumor cells [45]. Thus, in most cases there are significant differences between the biologic latent period and the epidemiologic latent period.

Some Examples

The latent period of a disease may be very short and fairly constant. In some chronic infectious diseases, and in cancer, the latent period may be very long and varies greatly among individuals, in which case one should treat it as a **random variable** and work with the probability distribution of this variable.

Example 1. Yellow Fever

Yellow fever is an infectious disease caused by a yellow fever virus which is the prototype of the flavivirus genus (family Flaviviridae). It is an acute, mosquito-borne viral infection that occurs in epidemic and endemic form in tropical America and

Africa. Clinical symptoms of this disease include fever, headache, malaise, and lassitude which persist for 2 to 4 days and occur in 10%–20% of the infected individuals. For this disease, the incubation period is very short (3 to 6 days) and can be considered as fairly constant [30].

Example 2. Malaria

Malaria is an infectious disease caused by parasites called Plasmodia. This disease occurs mainly in tropical areas and is transmitted to humans by the bite of malaria-infected female Anopheles mosquitoes. The four major *Plasmodium* species are *P. falciparum* (Africa, Asia, Oceania, Central America, and South America), *P. vivax* (Asia, Oceania, Central America, and South America), *P. ovale* (Africa and Oceania), and *P. malariae* (Africa and South America). The incubation periods for these four *Plasmodium* species are 8–27 days (average 12 days), 8–27 days (average 14 days), 9–17 days (average 15 days), and 16–28 days, respectively [33]. For this disease the human host becomes infectious with the accumulation of gametocytes in the blood. Hence the interval from infection to infectiousness is the time from initial infection to the first appearance of gametocytes in the blood. (This is the definition of latent period used by Anderson & May [2].) For the above four species, this period is given by 9–10 days, 9–10 days, 10–14 days, and 15–16 days, respectively [2].

Example 3. Leukemia in A-Bomb Survivors

Land & Norman [24] have studied the biologic latent periods of radiogenic cancers occurring among Japanese A-bomb survivors in Hiroshima and Nagasaki, Japan. The leukemias (acute leukemia and chronic granulocytic leukemia) are particularly interesting since the cumulative distributions of those who have been exposed to an A-bomb with kerma doses of 100 rads or more lie on the far left of those who have not been exposed to an A-bomb or those who have been exposed to an A-bomb but with the kerma doses of 0–9 rads. The magnitude of the elevation of the cumulative probability of leukemia over the biologic latent period depends on the age of the survivor at the time of exposure, with the age group 10–19 years at exposure having the largest elevation followed by the age groups 20–34 and 35–49 years at exposure. The biologic latent periods for leukemia are intervals from five years since exposure (time of explosion of the bomb) to an endpoint, say t_1, which is less than 29 years since exposure and which depends on the age of the survivor at the time of exposure. For the age groups 10–19 and 20–34 years at exposure, t_1 is 29 years since exposure, but for age groups 0–9 and 35–49 years at exposure, t_1 is approximately 25 years since exposure.

Example 4. Incubation Period of AIDS

The infectious chronic disease **AIDS** is caused by a retrovirus called HIV (human immunodeficiency virus). This is an endemic fatal infectious disease without cure at the present time. (For a summary of basic facts about AIDS, see [34].) Following infection by HIV, it usually takes several months to develop HIV antibodies in the blood. (For the time interval from infection to the development of antibodies, the estimate by Horsburgh et al. [20] is 3.5 months.) According to the 1993 surveillance definition of AIDS used by the US **Centers for Disease Control** and Prevention (CDC), the incubation period is the time interval between infection by HIV and the first time that the total CD4 T-cell counts falls below 200/mm^3 or the first time that the absolute percentage of CD4 T-cells falls below 14% or the first time that one of the 25 symptoms listed in [12] appears. This period is usually several years and depends on age [35], treatment with antiviral drugs [29], the presence of mutations of the gene CCR5 [16] (long or short AIDS survivors), and possibly other **covariates**. For untreated subjects aged 20–50 years at infection, the average incubation period is about 10 years. Note, however, that the AIDS definition used by CDC has been broadened three times, first in June 1985, next in July 1987, and then in December 1992. Hence, the incubation times measured before 1993 tend to be longer than the incubation times based on the 1993 AIDS definition.

The Latent Period of Infectious Diseases

For some infectious diseases such as yellow fever and malaria, the incubation period is relatively short and can be regarded as approximately constant. However, for some chronic infectious diseases such as AIDS,

the incubation period is long and variable. In this latter case it makes more sense to treat the incubation period as a random variable, rather than as a fixed constant "latency", and to describe the process in terms of the probability distribution of incubation times. For example, the probability distribution of the incubation period of AIDS has been studied extensively, as summarized in Brookmeyer & Gail [9], Becker & Motika [6], and Tan et al. [38]. This probability distribution has been estimated by both parametric and **nonparametric methods**. However, all the estimates in the literature are based on the 1987 definition of AIDS: estimates of the HIV incubation period based on data and the 1993 AIDS definition have yet to be published.

The Latent Period of Cancer

Some researchers have used the concept of latency and average latent period to describe the interval between exposure to the A-bomb and the subsequent cancer onset in Hiroshima and Nagasaki, Japan [24, 7]. For example, leukemias tend to arise about five years following exposure to nuclear radiation [7]. However, Brookmeyer [8] pointed out that such estimates might be misleading because of censoring (*see* **Censored Data**) and competing causes (*see* **Competing Risks**) of death.

Many investigators prefer to consider the distribution of time to cancer onset, especially investigators who study cancer onset in animals exposed to low doses of a carcinogen [15]. In such cases the latent period is usually very long, and the expected time-to-tumor may exceed vastly the normal life span of the animal. In such circumstances, information on the mean latent period is not sufficient to determine the probability of developing a tumor before dying of some other causes. Moreover, different distributions may have the same **mean** time to tumor in the high dose range but give vastly different risk estimates when extrapolated to low doses [17, 43]. Thus, some scientists have avoided the use of mean latency for risk assessment based on low dose extrapolation (*see* **Extrapolation, Low Dose**) [15]; rather, they describe carcinogens as altering the probability distribution of time to detectable cancer. This probability distribution depends on the mechanism of carcinogenesis and is influenced by many factors. In particular, the incidence of cancer is altered by changing the dose of the carcinogen to which the individual is exposed.

Armitage & Doll [4] developed the first stochastic model of carcinogenesis for the time-to-tumor distribution. This model is referred to as the multistage model (*see* **Multistage Carcinogenesis Models**) as described in reviews by Whittemore & Keller [42] and Kalbfleisch et al. [21]. The Armitage–Doll multistage model (*see* **Dose–Response Models in Risk Analysis**) assumes that a tumor develops from a normal stem cell by k ($k \geq 2$) consecutive and irreversible genetic changes. These assumptions and the assumptions of low transition rates imply a **Weibull** model for the cancer **incidence rate**, $\lambda(t)$, and the following dose–response relationship between cancer incidence rate and the dose, d, of carcinogen:

$$\lambda(t) \propto \eta(d) \times t^{k-1}, \tag{1}$$

where $\eta(d)$ is a function of the dose d and is independent of time t.

The Armitage–Doll multistage model has been widely used by statisticians to assess how exposure to carcinogens alters the cancer incidence rates and the distributions of time-to-tumor. Breslow & Day [7] and others [13, 10] applied this model to study the effects of cigarette smoking on lung cancer risk, of asbestos exposure on risk of lung cancer and mesothelioma, and of radiation exposure on risks of leukemia, breast cancer, and bone cancer. While it is widely accepted that cancer results from a multistage process, recent results from molecular biology and molecular genetics have raised questions about some details of the assumptions in the Armitage–Doll multistage model (see [36] and [19]).

For risk assessment of carcinogens by low dose extrapolation, it has been documented that the same observable data can be fitted equally well by different models that yield very different estimates of risk at low doses [41]. Such extrapolation should be based on biologically plausible models, preferably models suggested by data. Thorslund et al. [40] and Moolgavkar et al. [32] proposed the MVK two-stage model (see [36]) for risk assessment. In this model, the first stage is a **Poisson process** describing how normal stem cells are changed into initiated cells by mutation (initiation); in the second stage the model incorporates stochastic birth and death (*see* **Stochastic Processes**) for proliferation of initiated cells (promotion), that change into malignant tumor cells by another mutation. Dose–response curves based on the MVK two-stage model have been developed by Chen & Moini [11], and by Krewski &

Murdoch [23]. They have used these dose–response curves to assess how a carcinogen alters cancer incidence through its effects on initiating mutations or on the rate of proliferation of initiated cells. If the carcinogen is a pure initiator, then the dose–response curve for cancer incidence can be factorized as a product of a function of dose and a function of time and age; in these cases, the pattern of dose–response curves of the MVK model is quite similar to that of the Armitage–Doll multistage model. However, if the carcinogen is a promoter or a complete carcinogen, then the dose–response curves of the MVK model cannot be factorized, and they differ qualitatively from the Armitage–Doll model.

The MVK two-stage model, and extensions of it, together with many other biologically supported models have been analyzed in Tan [36] and in Yakovlev & Tsodikov [44]. Some extensions and modifications have recently been developed by Little and his colleagues [25–28]. (Little [25, 26] has called the multievent model in Tan [36] the generalized MVK model.) By merging initiation and promotion, alternate modeling approaches have been proposed by Klebanov et al. [22] for radiation carcinogenesis.

References

[1] Alcabes, P. (1993). The incubation period of human immunodeficiency virus, *Epidemiologic Reviews* **15**, 303–318.

[2] Anderson, R.M. & May, R.M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[3] Armitage, P. & Doll, R. (1961). Stochastic models for carcinogenesis, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Biology and Problems of Health*. University of California Press, Berkeley, pp. 19–38.

[4] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

[5] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*. Griffin, London.

[6] Becker, N.G. & Motika, M. (1993). Smoothed nonparametric back-projection of AIDS incidence data with adjustment for therapy, *Mathematical Biosciences* **118**, 1–23.

[7] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[8] Brookmeyer, R. (1988). Time and latency considerations in the quantitative assessment of risk, in *Epidemiology and Health Risk Assessment*, L. Gordis, ed. Oxford University Press, Oxford, pp. 178–188.

[9] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, Oxford.

[10] Brown, C.C. & Chu, K.C. (1983). Implications of multistage theory of carcinogenesis applied to occupational arsenic exposure, *Journal of the National Cancer Institute* **70**, 455–463.

[11] Chen, C.W. & Moini, A. (1990). Cancer dose–response models incorporating clonal expansion, in *Scientific Issues in Quantitative Cancer Risk Assessment*, S.H. Moolgavkar, ed. Birkhauser, Boston, pp. 153–175.

[12] CDC (1992). Revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults, *Morbidity and Mortality Weekly Report* **41**(RR-17), 1–19.

[13] Day, N.E. & Brown, C.C. (1980). Multistage models and primary prevention of cancer, *Journal of the National Cancer Institute* **64**, 977–989.

[14] Druckrey, H. (1967). Quantitative aspects of carcinogenesis, in *Potential Carcinogenic Hazards from Drugs*, R. Truhaut, ed. UICC Monograph Series, Vol. 7, Springer-Verlag, New York, pp. 60–78.

[15] Guess, H.A. & Hoel, D.G. (1977). The effect of dose on cancer latency period, *Journal of Environmental Pathology and Toxicology* **1**, 279–286.

[16] Hill, C.M. & Littman, D.R. (1996). Natural resistance to HIV, *Nature* **382**, 668–669.

[17] Hoel, D.G., Gaylor, D.W., Kirschstein, R.L. & Saffiotti, U. (1975). Estimation of risks of irreversible delayed toxicity, *Journal of Toxicology and Environmental Health* **1**, 133–151.

[18] Hoel, D.G., Kaplan, N.L. & Anderson, N.W. (1983). Implication of nonlinear kinetics on risk estimation in carcinogenesis, *Science* **210**, 1032–1037.

[19] Hopkin, K. (1996). Tumor evolution: survival of the fittest cells, *Journal of NIH Research* **8**, 37–41.

[20] Horsburgh, C.R. Jr, Qu, C.Y. & Jason, I.M. (1989). Duration of human immunodeficiency virus infection before detection of antibody, *Lancet* **2**, 637–640.

[21] Kalbfleisch, J.D., Krewski, D.R. & Van Ryzin, J. (1983). Dose–response models for time-to-response toxicity data, *Canadian Journal of Statistics* **11**, 25–50.

[22] Klebanov, L.B. Rachev, S.T. & Yakovlev, A.Y. (1993). A stochastic model of radiation carcinogenesis: latent time distributions and their properties, *Mathematical Biosciences* **113**, 51–75.

[23] Krewski, D.R. & Murdoch, D.J. (1990). Cancer modeling with intermittent exposure, in *Scientific Issues in Quantitative Cancer Risk Assessment*, S.H. Moolgavkar ed. Birkhauser, Boston, pp. 196–214.

[24] Land, C.E. & Norman, J.E. (1978). Latent periods of radiogenic cancers occurring among Japanese A-bomb survivors, in *Late Biological Effects of Ionizing*

Radiation, Vol. 1. International Atomic Energy Agency, Vienna.

[25] Little, M.P. (1995). Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon and Knudson, and of the multistage model of Armitage and Doll, *Biometrics* **51**, 1278–1291.

[26] Little, M.P. (1996). Generalizations of the two-mutation and classical multi-stage models of carcinogenesis fitted to the Japanese atomic bomb survivor data, *Journal of Radiology Protection* **16**, 7–24.

[27] Little, M.P., Muirhead, C.R., Boice, J.D. & Kleinerman, R.A. (1995). Using multistage models to describe radiation-induced leukaemia, *Journal of Radiology Protection* **15**, 315–334.

[28] Little, M.P., Muirhead, C.R. & Stiller, C.A. (1996). Modeling lymphocytic leukaemia incidence in England and Wales using generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon and Knudson, *Statistics in Medicine* **15**, 1003–1022.

[29] Longini, I.R. Jr, Clark, W.S. & Karon, J. (1993). The effect of routine use of therapy in showing the clinical course of human immunodeficiency virus (HIV) infection in population-based cohort, *American Journal of Epidemiology* **137**, 1229–1240.

[30] Monath, T.P. (1994). Yellow fever, in *Infectious Disease: A Treatise of Infectious Diseases*, 5th Ed. P.D. Hoeprich, M.C. Jordan & A.R. Ronald, eds. Lippincott, Philadelphia pp. 826–828.

[31] Mosley, J.W. (1994). Epidemiology, in *Infectious Disease: A Treatise of Infectious Diseases*, 5th Ed. P.D. Hoeprich, M.C. Jordan & A.R. Ronald, eds. Lippincott, Philadelphia pp. 20–31.

[32] Moolgavkar, S.H. Cross, F.T. & Luebeck, E.G. (1990). A two-mutation model for radon-induced lung tumors in rats, *Radiation Research* **121**, 28–37.

[33] Redd, S.C. & Campbell, C.C. (1994). Malaria, in *Infectious Disease: A Treatise of Infectious Diseases*, 5th Ed. P.D. Hoeprich, M.C. Jordan & A.R. Ronald, eds. Lippincott, Philadelphia, pp. 1335–1344.

[34] Rhame, F.S. (1994). Acquired immunodeficiency syndrome, in *Infectious Disease: A Treatise of Infectious Diseases*, 5th Ed. P.D. Hoeprich, M.C. Jordan & A.R. Ronald, eds. Lippincott, Philadelphia, pp. 628–652.

[35] Rosenberg, P.S. (1995). Scope of the AIDS epidemic in the United States, *Science* **270**, 1372–1375.

[36] Tan, W.Y. (1991). *Stochastic Models of Carcinogenesis*. Marcel Dekker, New York.

[37] Tan, W.Y. & Singh, K.P. (1987). Assessing the effects of metabolism of environmental agents on cancer tumor development by a two-stage model of carcinogenesis, *Environmental Health Perspective* **74**, 203–210.

[38] Tan, W.Y. Tang, S.C. & Lee, S.R. (1996). Characterization of the HIV incubations and some comparative studies, *Statistics in Medicine* **15**, 197–220.

[39] Thomas, D.C. & McNeill, K.G. (1982). Risk estimates for the health effects of alpha radiation, in *Appendix M, Research Report*. Atomic Energy Control Board, Ottawa.

[40] Thorslund, T.W., Brown, C.C. & Charnley, C. (1987). Biologically motivated cancer risk models, *Risk Analysis* **7**, 109–119.

[41] Van Ryzin, J. (1980). Quantitative risk assessment, *Occupational Medicine* **22**, 321–326.

[42] Whittemore, A.S. & Keller, J.B. (1978). Quantitative theories of carcinogenesis, *SIAM Review* **20**, 1–30.

[43] Whittemore, A. & Altshuler, B. (1976). Lung cancer incidence in cigarette smokers: further analysis of Doll and Hill's data for British physicians, *Biometrics* **32**, 805–816.

[44] Yakovlev, A.Y. & Tsodikov, A.D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

[45] Yang, G.L. & Chen, C.W. (1991). A stochastic two-stage carcinogenesis model: a new approach to computing the probability of observing tumor in animal bioassays, *Mathematical Biosciences* **104**, 247–258.

WAI-YUAN TAN & CHAO W. CHEN

Latin Square Designs

A Latin square design is a **balanced incomplete block design** for comparing t treatments in which heterogeneity is eliminated in two ways. It is an **incomplete block design** insofar as not every combination of row, column, and treatment is assigned to an experimental unit. It is a *balanced* design insofar as the number of treatments is equal in each row and in each column.

Example 1

Suppose a toxicologist wants to compare a series of t treatments. Suppose the experiments are to be carried out in r different laboratories $L_1, \ldots, L_r$ on c different animal species $S_1, \ldots, S_c$. The experimenter wants to take into account the heterogeneity coming from both these factors. The simplest experimental design would be a **randomized complete blocks design** in which every treatment is assigned to every combination of both heterogeneity factors. In this whole experiment we would have rct experimental units corresponding to rc blocks and t treatments. This number of experimental units can become enormous, even for moderate r, c, and t. If $r = c = t$, this design can be replaced by a Latin Square Design (LSD) in which each treatment, traditionally denoted with Latin letters, occurs exactly once in each row and once in each column, so that the number of experimental units is only t^2.

In other settings a complete block design is physically impossible, as in Examples 2 and 3, and Latin square designs are a natural alternative.

Example 2

In a field experiment where the experimental field exhibits a gradient in two orthogonal directions, each spot can only be assigned to a single treatment.

Example 3

In a **clinical trial** the **blocking** factors may be the individual subject and successive time periods. If the objective of the trial is to compare treatments, then only one treatment can be given to a given subject at a given time. This design is called a **crossover** trial (see [5]).

Construction of the Design

If we want to use an LSD, we have first to choose a Latin square. A Latin square of order t is an arrangement of t letters or numbers (representing the treatments) in a square of t columns and t rows (representing the two heterogeneity factors), such that each letter appears once and only once in each column and each row (whence the balance of the design). Latin squares of any order exist, as can be seen from Figure 1.

The enumeration of all the Latin squares of any order t becomes tedious as t increases. However, permuting rows, columns or treatments of a Latin square gives another Latin square. Particular Latin squares are those for which the first column and the first row are ordered (A, B, C, ...); these are called *standard* Latin squares. By permuting rows and columns of a standard Latin square we can obtain $t!(t-1)!$ different Latin squares.

Thus, sampling a Latin square consists in:

1. sampling a standard Latin square with equiprobability among all the standard Latin squares;
2. randomly permuting the t rows, the $t-1$ first columns and the t treatments.

For more details of this procedure and tables of standard Latin squares, see [2]; for tables of random permutations, see [1].

The **randomization** procedure of a Latin square is equivalent to the observation of the t^3 random variables δ_{ijk}, where $\delta_{ijk} = 1$ if the treatment k is affected in row i and column j, and 0 otherwise. These random variables have the following

1	2	3	...	$t-1$	t
2	3	4	...	t	1
3	4	5	...	1	2
...					...
t	1	2	...		$t-1$

Figure 1 Example of Latin square of order t

properties:

$$\sum_i \delta_{ijk} = \sum_j \delta_{ijk} = \sum_k \delta_{ijk} = 1. \tag{1}$$

$$\mathrm{E}(\delta_{ijk}) = \frac{1}{t}, \quad \text{for all } i, j, k. \tag{2}$$

$$\mathrm{E}(\delta_{ijk}\delta_{i'j'k'}) = \begin{cases} 1/t, & \text{if } i = i' \text{ and } j = j' \text{ and } k = k', \\ 0, & \text{if either } i \neq i' \text{ or } j \neq j' \text{ or } k \neq k', \\ 1/t(t-1), & \text{if either } i = i' \text{ or } j = j' \text{ or } k = k', \\ 1/t(t-1)^2(t-2), & \text{if } i \neq i' \text{ and } j \neq j' \text{ and } k \neq k'. \end{cases} \tag{3}$$

Estimation and Analysis of Variance

Let Y_{ijk} be the response that would be observed if the kth treatment were assigned on the ith row and the jth column. Under unit-treatment additivity (in the terminology of Hinkelmann & Kempthorne [3]), i.e. under the hypothesis that there are no **interactions** between the treatments on one side and rows or columns on the other side, we can write:

$$Y_{ijk} = \mu + \alpha_i^{\mathrm{C}} + \alpha_j^{\mathrm{R}} + \alpha_k^{\mathrm{T}} + \alpha_{ij}^{\mathrm{RC}} + \varepsilon_{ijk}, \tag{4}$$

where superscripts C, R, and T indicate columns, rows, and treatments respectively. The technical errors ε_{ijk} are assumed independent with zero **mean** and equal **variance** σ_ε^2, and are independent of the randomization procedure. We can also write the usual side conditions which imply no loss of generality:

$$\alpha_\cdot^{\mathrm{C}} = \alpha_\cdot^{\mathrm{R}} = \alpha_\cdot^{\mathrm{T}} = \alpha_{i\cdot}^{\mathrm{RC}} = \alpha_{\cdot j}^{\mathrm{RC}} = 0, \quad \text{for all } i, j. \tag{5}$$

We note that within an LSD not all Y_{ijk} are actually measured but only those for which the **random variable** δ_{ijk} is equal to 1.

The observed means can thus be written:

$$Y_{i\cdot\cdot} = \frac{1}{t}\sum_j \sum_k \delta_{ijk} Y_{ijk},$$

$$Y_{\cdot j\cdot} = \frac{1}{t}\sum_i \sum_k \delta_{ijk} Y_{ijk},$$

$$Y_{\cdot\cdot k} = \frac{1}{t}\sum_i \sum_j \delta_{ijk} Y_{ijk},$$

$$Y_{\cdots} = \frac{1}{t^2}\sum_{ijk} \delta_{ijk} Y_{ijk}. \tag{6}$$

If we substitute Y_{ijk} in (6) by its expression from (4), then, after applying the side conditions (5) and the relations (1), we obtain:

$$\begin{aligned} Y_{i\cdot\cdot} &= \mu + \alpha_i^{\mathrm{R}} + \frac{1}{t}\sum_j \sum_k \delta_{ijk}\varepsilon_{ijk}, \\ Y_{\cdot j\cdot} &= \mu + \alpha_j^{\mathrm{C}} + \frac{1}{t}\sum_i \sum_k \delta_{ijk}\varepsilon_{ijk}, \\ Y_{\cdot\cdot k} &= \mu + \alpha_k^{\mathrm{T}} + \frac{1}{t}\sum_i \sum_j \delta_{ijk}\varepsilon_{ijk}, \\ Y_{\cdots} &= \mu + \frac{1}{t^2}\sum_i \sum_j \sum_k \delta_{ijk}\varepsilon_{ijk}. \end{aligned} \tag{7}$$

The total sum of squares $SS_{\mathrm{tot}} = \sum_{ijk} \delta_{ijk}(\varepsilon_{ijk})^2$ can be decomposed as follows:

$$SS_{\mathrm{tot}} = t\sum_i (Y_{i\cdot\cdot} - Y_{\cdots})^2 + t\sum_j (Y_{\cdot j\cdot} - Y_{\cdots})^2 + t\sum_k (Y_{\cdot\cdot k} - Y_{\cdots})^2 + SS_{\mathrm{e}}. \tag{8}$$

The **expectations** of these sums of squares depend on the **moments** of the δ_{ijk} given in (2) and (3). Somewhat tedious computations yield the **analysis of variance** table given in Table 1 where:

$$\sigma_{\mathrm{R}}^2 = \frac{1}{t-1}\sum_i (\alpha_i^{\mathrm{R}})^2,$$

$$\sigma_{\mathrm{C}}^2 = \frac{1}{t-1}\sum_j (\alpha_j^{\mathrm{C}})^2,$$

$$\sigma_{\mathrm{T}}^2 = \frac{1}{t-1}\sum_k (\alpha_k^{\mathrm{T}})^2, \tag{9}$$

and

$$\sigma_{\mathrm{RC}}^2 = \frac{1}{(t-1)^2}\sum_i \sum_j (\alpha_{ij}^{\mathrm{RC}})^2.$$

To test the hypothesis (*see* **Hypothesis Testing**) of equality of the treatment effects (i.e. $\sigma_{\mathrm{T}}^2 = 0$), we are

Table 1 Analysis of variance of an LSD

Source	df	SS	E(MS)
Rows	$t-1$	$t\sum_i (Y_{i\cdot\cdot} - Y_{\cdots})^2$	$\mathrm{E}(MS_{\mathrm{T}}) = \sigma_\varepsilon^2 + t\sigma_{\mathrm{R}}^2$
Columns	$t-1$	$t\sum_j (Y_{\cdot j\cdot} - Y_{\cdots})^2$	$\mathrm{E}(MS_{\mathrm{C}}) = \sigma_\varepsilon^2 + t\sigma_{\mathrm{C}}^2$
Treatments	$t-1$	$t\sum_k (Y_{\cdot\cdot k} - Y_{\cdots})^2$	$\mathrm{E}(MS_{\mathrm{T}}) = \sigma_\varepsilon^2 + \sigma_{\mathrm{RC}}^2 + t\sigma_{\mathrm{T}}^2$
Error	t^2-3t+2	by subtraction $= SS_{\mathrm{e}}$	$\mathrm{E}(MS_{\mathrm{e}}) = \sigma_\varepsilon^2 + \sigma_{\mathrm{RC}}^2$
Total	t^2-1	$\sum_{ijk} \delta_{ijk}(Y_{ijk} - Y_{\cdots})^2$	

led by Table 1 to consider the ratio

$$F = \frac{MS_{\mathrm{T}}}{MS_{\mathrm{e}}}. \tag{10}$$

This statistic can be referred to $F(t-1, t^2-3t+2)$ under normal theory (*see* ***F*** **Distributions**). Nevertheless, Welch [6] investigates its distribution under the randomization process for LSD described above and Hinkelmann & Kempthorne [3], reviewing this work, "assume that normal theory gives satisfactory approximations to corresponding randomization tests".

Another immediate consequence of Table 1 is that there do not exist any legitimate tests for row and column effects unless we suppose absence of interaction between them ($\sigma_{\mathrm{RC}}^2 = 0$).

Other Topics

Departure from Additivity

The effects of row × treatment, column × treatment, row × column × treatment can severely affect the results obtained above, as discussed in detail by Scheffé [4]. Nevertheless, Wilk & Kempthorne [7] show that the usual F test is still appropriate, even in the presence of interactions, if the t rows have been sampled from a population of R rows, the t columns have been sampled from a population of C columns, the t treatments have been sampled from a population of T treatments with $R >> t, C >> t$, and $T >> t$ (*see* **Random Effects**).

Limitations of the LSD and Some Extensions

It can be argued that the LSD is limited from a practical point of view. Some extensions have been proposed for the following problems (see [3]):

1. the numbers of rows, columns and treatments have to be the same (**Youden squares** and Latin rectangles are some alternatives);
2. for small values of t the number of **degrees of freedom** for error is insufficient (this can be mended by replicating the LSDs);
3. the number of heterogeneity factors is restricted to two (**Graeco–Latin squares** deal with three or more heterogeneity factors).

References

[1] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York.

[2] Fisher, R.A. & Yates, F. (1963). *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver & Boyd, Edinburgh.

[3] Hinkelmann, K. & Kempthorne, O. (1994). *Design and Analysis of Experiments*, Vol. 1. Wiley, New York.

[4] Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

[5] Senn, S. (1994). The AB/BA crossover: past, present and future?, *Statistical Methods in Medical Research* **3**, 303–324.

[6] Welch, B.L. (1937). On the z-test in randomized blocks and Latin squares, *Biometrika* **29**, 21–52.

[7] Wilk, M.B. & Kempthorne, O. (1957). Non-additivities in a Latin square design, *Journal of the American Statistical Society* **52**, 218–236.

PASCAL WILD & MICHEL GRZEBYK

Lattice Designs

The term *lattice design* encompasses two different types of equireplicate designs. The first type, *square lattice designs* and their extensions, are resolvable (0, 1) incomplete block designs. The second type, *lattice squares* and their extensions, are nested *row-column designs* (*see* **Youden Squares and Row–Column Designs**) for which each block is a complete replicate.

An **incomplete block design** for t treatment each replicated r times with all blocks of size $k < t$ is *resolvable* if the b blocks can be grouped into $r = bk/t$ sets of $s = t/k$ blocks, such that each set is a complete replicate. A (0, 1)-*design* has each pairwise treatment concurrence equal to 0 or 1. A square lattice design, introduced by **F. Yates** in 1936, has $t = k^2$ and $s = k$. It can be constructed by writing the numbers 1 to t in a $k \times k$ array. The *simple square lattice design* has $r = 2$. The k blocks of the first replicate are the rows of the array. The second replicate uses the columns. The *triple square lattice design* has $r = 3$. The third replicate is found by superimposing a **Latin square** of order k on the array, and using the Latin square symbols to define the blocks.

If $k \neq 6$, then the Greek letters of a **Graeco–Latin square** can be used to get a fourth replicate (*quadruple square lattice design*). If $k \neq 3$, 6, 10, then at least five replicates can be obtained using *mutually* **orthogonal** *Latin squares* (MOLS). If k is a prime power, a complete set of $k - 1$ MOLS exist, and up to $k + 1$ replicates are possible. If, in this case, $k + 1$ replicates are used, then the variance-balanced design, called a *balanced square lattice*, is a symmetric **balanced incomplete block design** BIBD $[k^2, k(k + 1), k]$.

For example, using the three MOLS of order 4:

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

1	2	3	4
3	4	1	2
4	3	2	1
2	1	4	3

1	2	3	4
4	3	2	1
2	1	4	3
3	4	1	2

on the array

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

gives the balanced square lattice design (parentheses denote blocks, each replicate is a row):

(1 2 3 4),	(5 6 7 8),	(9 10 11 12),	(13 14 15 16);
(1 5 9 13),	(2 6 10 14),	(3 7 11 15),	(4 8 12 16);
(1 6 11 16),	(2 5 12 15),	(3 8 9 14),	(4 7 10 13);
(1 7 12 14),	(2 8 11 13),	(3 5 10 16),	(4 6 9 15);
(1 8 10 15),	(2 7 9 16),	(3 6 12 13),	(4 5 11 14).

For $r < 5$, the first r replicates (say) can be used.

The need for t to be a perfect square (and not 36 if more than three replicates are required) can be a severe restriction. In some experiments, such as variety trials which often use a large number of varieties and few replicates (two or three), it may be possible to add a few extra treatments, or even remove some, to get $t = k^2$. Various extensions have been proposed to allow some other values of t. *Cubic lattices* have $t = k^3$ and $s = k^2$ in a similar way using a cubic array and orthogonal Latin cubes; and m-dimensional lattices with $t = k^m$ for $m > 3$ are possible if k is prime. A *rectangular lattice* with $t = k(k + 1)$ and $s = k + 1$ can be constructed in a similar way to the square lattices using the numbers 1 to t in a $(k + 1) \times (k + 1)$ array with the leading diagonal omitted. For $r > 2$ the Latin squares used must have each symbol occurring on the main diagonal. The idea for constructing the rectangular lattice can be used for any $s > k + 1$ by, if possible, omitting further (wrap-around) diagonals. The α-*designs* are an extension to resolvable designs with any s – see John [3, Section 4.8] and John & Williams [4, Sections 4.4 and 4.5]. Optimal two-replicate resolvable designs are discussed by John & Williams [4, Section 4.7].

Although specific formulas can be given, as in John [3, Section 3.4], John & Williams [4, Sections 4.2 and 4.3], the usual *intrablock* analysis for the model with fixed treatment and block effects – see John [3, Sections 1.3–1.5] and John & Williams [4, Sections 1.4–1.6] – is easily carried out on a computer. If s is not small, then it may be worth using the *interblock* information also. Various methods are available for combining the information, but a good choice nowadays would be to use a computer package (*see* **Software, Biostatistical**) such as GENSTAT or SAS which performs REML (**restricted maximum likelihood**) estimation – see John & Williams [4, Sections 7.5 and 7.6].

Lattice squares, introduced by F. Yates in 1940, are nested row–column designs for $t = k^2$ in $r > 1$ complete blocks (or groups) of k rows and k columns.

If k is a prime power, then a variance-balanced design, the *balanced lattice square*, is possible. This needs $r = (k+1)/2$ for k odd, and $r = k+1$ for k even. The balanced lattice square can be constructed from the balanced square lattice design. If k is odd, then each replicate of a balanced square lattice forms either the rows or the columns of one block. If k is even, then each replicate of a balanced square lattice forms the rows of one block, and the columns of another. For example, the balanced lattice square with $t = 3$ is

1	2	3
4	5	6
7	8	9

1	6	8
9	2	4
5	7	3

and one with $t = 4$ is

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

1	5	9	13
6	2	14	10
11	15	3	7
16	12	8	4

1	6	11	16
12	15	2	5
14	9	8	3
7	4	13	10

1	7	12	14
8	2	13	11
10	16	3	5
15	9	6	4

1	8	10	15
2	7	9	16
3	6	12	13
4	5	11	14

If k is odd, then a balanced lattice square with $r = k+1$ repeats the original design with the rows and columns interchanged.

Lattice squares which are not variance-balanced are obtained if other numbers of replicates are used. Care is then needed in the choice of the r replicates to ensure that the combined row and column treatment concurrences are as equal as possible – see Cochran & Cox [1, Section 12.12] and John & Williams [4, Section 6.2]. Designs may be possible for some $r > 1$ if k is not a prime power, depending on the maximum number of MOLS of order k. Balanced lattice squares for $k = 3, 4, 5, 7, 8, 9, 11, 13$ are given by Cochran & Cox [1, Chapter 12].

As before, the restriction on the possible values of t may be severe. Extensions to *lattice rectangles* are possible. John & Williams [4] discuss other resolvable nested row–column designs: α-designs (Section 6.3), two-replicate designs (Section 6.6), and designs generated using **algorithms** (Section 6.4). If the design is not balanced, then the usual intrablock analysis can again be easily carried out on a computer. There is no interblock information when each block is a complete replicate, but there may sometimes be useful interrow or intercolumn information, which again is best combined using REML estimation.

The lattice square designs and square lattices can be used for replicated blocked **factorial experiments** if the product of the levels is k^2 – see Cochran & Cox [1, Sections 10.12 and 12.11]. For the balanced designs, main effects and **interactions** are equally partially **confounded**. If $t_1 = t_2 = k$, then the first replicate of the square lattice design confounds one main effect, and the second replicate confounds the other.

Some further details on lattice designs are given by Cornelius [2].

References

[1] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York.

[2] Cornelius, P.L. (1983). Lattice designs, in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 510–518.

[3] John, J.A. (1987). *Cyclic Designs*. Chapman & Hall, London.

[4] John, J.A. & Williams, E.R. (1995). *Cyclic and Computer Generated Designs*, 2nd Ed. of [3]. Chapman & Hall, London.

(*See also* **Experimental Design**)

RICHARD J. MARTIN

Law of Large Numbers

The first theorem recognizable as a precise form of a limiting-frequency statement (or "law of large numbers" in the terminology introduced by Poisson) is the famous "*weak law of large numbers for Bernoulli trials*" of James Bernoulli (1654–1705) [2] (*see* **Bernoulli Family**). In our current notation and terminology his theorem, published posthumously in *Ars Conjectandi* (1713), now reads as follows. Let $X_1, \ldots, X_n$ be the outcomes of n independent 0–1 trials (Bernoulli trials) with success probability p. With $S_n = \sum_{i=1}^{n} X_i$ the number of successes (the number of ones) and $\overline{X}_n = S_n/n$, we have for each $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - p| > \varepsilon\right) = 0.$$

In words: the ratio of the number of successes and the number of trials or the proportion of successes **converges in probability** to the success probability p. A universally applied notation is $\overline{X}_n \xrightarrow{\Pr} p$ (i.e. $\overline{X}_n$ converges to p in probability).

Remark 1. The result provides some empirical confirmation for the axioms of probability theory [7, Chapter 1] (*see* **Foundations of Probability**). See also [3, pp. 20–21] for an introductory discussion.

Remark 2. The proof is a simple application of the Chebyshev inequality:

$$\Pr\left(|\overline{X}_n - p| > \varepsilon\right) \le \frac{1}{\varepsilon^2}\mathrm{E}\left(\overline{X}_n - p\right)^2$$
$$= \frac{p(1-p)}{n\varepsilon^2} \longrightarrow 0, \quad n \longrightarrow \infty.$$

For $X_1, \ldots, X_n$ a sequence of independent random variables with common distribution function F, let $S_n = \sum_{i=1}^{n} X_i$, $\overline{X}_n = S_n/n$ and $\mu = \mathrm{E}(X_1)$. Then the following generalization of Bernoulli's theorem has been obtained.

Theorem 1. Khinchin's law of large numbers

$$\mathrm{E}|X_1| < \infty \text{ implies } \overline{X}_n \xrightarrow{\Pr} \mu.$$

In words: the sample mean $\overline{X}_n$ is a "weakly consistent" estimator of the population mean μ.

A refined version of this result is the following characterization due to Kolmogorov (see [2] for further details).

Theorem 2. In order that there exist constants μ_n such that for each $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu_n| > \varepsilon\right) = 0,$$

it is necessary and sufficient that

$$n \Pr\left(|X_1| > n\right) \longrightarrow 0, \quad n \longrightarrow \infty.$$

In this case, $\mu_n = \int_{-n}^{n} x \, \mathrm{d}F(x)$.

Strong Law of Large Numbers

Almost sure (a.s.) convergence is a mode of convergence that describes the behavior of a statistic (e.g. the sample mean, $\overline{X}_n$) outside an unspecified set of probability zero. Almost sure convergence is stronger than convergence in probability (see Remark 4 below). Synonyms for almost sure convergence are convergence with probability one and convergence almost everywhere.

The Kolmogorov strong law of large numbers, the a.s. version of Khinchin's law, reads as follows.

Theorem 3. For $X_1, \ldots, X_n$ a sequence of independent random variables with common distribution function F:

$$\mathrm{E}|X_1| < \infty \text{ if and only if } \Pr\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1$$

with $\mu = \mathrm{E}(X_1)$.

Remark 3. The standard way to write this result is $\mathrm{E}|X_1| < \infty$ if and only if $\overline{X}_n \to \mu$ a.s.
In words: the sample mean $\overline{X}_n$ is a "strongly **consistent**" estimator of the population mean μ.

Remark 4. This simple characterization of strong convergence makes it clear why probabilists like the a.s. convergence mode. For statisticians the difference between the weak and the strong law of large numbers is subtle and cannot be adequately explained without measure theory. See [3, p. 233] for an intuitive discussion and some amusing quotations. From an applied point of view, most statisticians seem to be satisfied with convergence in probability.

Applications

Application 1

For $X_1, \ldots, X_n$ a sequence of independent random variables with common distribution function F let $\mathbb{F}_n$ denote the empirical distribution function, i.e.

$$\mathbb{F}_n(x) = \frac{\#\{i : X_i \le x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{X_i \le x\}, \quad x \in \mathbb{R}$$

(*see* **Goodness of Fit**). For fixed x, $\mathbb{F}_n(x) \to F(x)$ a.s. by the strong law of large numbers. This property strengthens to almost sure convergence uniform in x.

Theorem 4. Glivenko–Cantelli theorem.

$$\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| \longrightarrow 0 \text{ a.s.}$$

In words: uniformly in x we can rediscover F from the data. Taking n large enough, this can be done to any desired degree of precision.

Application 2

We now discuss the statistical relevance of the consistency results given in the previous sections. Many problems in statistics are of the following type: for a given (unknown) parameter θ ($\mathbb{R}$-valued, $\mathbb{R}^d$-valued, or a function), find a consistent estimator T_n and obtain the limit distribution of $n^\gamma(T_n - \theta)$. Typical values for γ are $\gamma = 1/2$ for the classical **central limit theorem** (normality) and $\gamma = 1$ for a limit distribution that corresponds to a weighted sum of centered **chi-square distributed** random variables. Typically, $\mathrm{var}[n^\gamma(T_n - \theta)] \to \sigma^2$, with σ^2 a **nuisance parameter** that needs to be estimated in a consistent way in order to construct, for example, approximate **confidence intervals** for θ.

To make this point clear, consider the following simple example. Given a sequence of independent identically distributed random variables $X_1, \ldots, X_n$ let $\theta = \mu$ and $T_n \equiv \overline{X}_n$. If $0 < \sigma^2 = \mathrm{var}(X_1) < \infty$, then the limit distribution of $n^{1/2}(\overline{X}_n - \mu)$ is a zero mean **normal distribution** with variance σ^2. The sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right\}$$

is a (strongly) consistent estimator for σ^2. To see this note that

$$\overline{X}_n \longrightarrow \mu \text{ a.s. implies } \overline{X}_n^2 \longrightarrow \mu^2,$$

and that the strong law of large number gives

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \longrightarrow \mathrm{E}X_1^2 = \sigma^2 + \mu^2.$$

Hence we have $S_n^2 \to \sigma^2$ a.s.

We therefore obtain, using Slutsky's theorem (for which weak consistency is sufficient) that the limit distribution of $n^{1/2}(\overline{X}_n - \mu)/S_n$ is standard normal. Hence, an approximate $100(1-\alpha)\%$ confidence interval for μ is of the form $\overline{x}_n \pm z_{1-(\alpha/2)}(s_n/n^{1/2})$ with $\overline{x}_n$ and s_n^2 the actual values of $\overline{X}_n$ and S_n^2 and $z_{1-(\alpha/2)}$ the $[1-(\alpha/2)]$-percentile of the standard normal distribution.

Extensions

Extension 1

Let $X_1, X_2, \ldots$ be a sequence of random variables with a common distribution function F and replace the independence assumption by the **stationarity** assumption: for every n the joint distribution of $X_1, \ldots, X_n$ is the same as the joint distribution of $X_{1+k}, \ldots, X_{n+k}$ for all positive integers k.

On the basis of general results from ergodic theory one can show that, given the stationary process, $X_1, X_2, \ldots, \overline{X}_n$ still obeys laws of large numbers. Note that ergodic theory applies to a large number of problems in probability theory and analysis. See [8] for further reading.

Extension 2

Given a sequence of independent zero-mean random variables $\Delta_1, \Delta_2, \ldots$, define $X_n = \sum_{i=1}^{n} \Delta_i$. With $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, the σ-algebra generated by $X_1, \ldots, X_n$, we then have that

$$\mathrm{E}(X_{n+1}|\mathcal{F}_n) = X_n.$$

This simple property delineates a very useful class of **stochastic processes**: martingales. The study of laws of large numbers for martingales is an intrinsic part of modern probability theory and the results

are extremely important for a variety of applications in, for example, survival analysis [5] (*see* **Counting Process Methods in Survival Analysis**). Further discussion is beyond the scope of this article; we refer to [1, Section 35] for an excellent introductory discussion and a number of interesting examples. A specialized reference is [6].

Finally, note that thorough discussions on the law of large numbers can be found in [4, 9], and [10].

References

[1] Billingsley, P. (1979). *Probability and Measure*. Wiley, New York.

[2] Bingham, N.H. (1989). *Theory of Probability and Its Applications* **34**, 129–139.

[3] Chung, K.L. (1975). *Elementary Probability Theory with Stochastic Processes*. Springer-Verlag, New York.

[4] Doob, J.L. (1953). *Stochastic Processes*. Wiley, New York.

[5] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[6] Hall, P. & Heyde, C.C. (1980). *Martingale Limit Theory and its Applications*. Academic Press, New York.

[7] Kolmogorov, A.N. (1950). *Foundations of the Theory of Probability*. Chelsea, New York.

[8] Krengel, U. (1985). *Ergodic Theorems*. Walter de Gruyter, Berlin.

[9] Révész, P. (1968). *The Laws of Large Numbers*. Academic Press, New York.

[10] Stout, W.F. (1974). *Almost Sure Convergence*. Academic Press, New York.

(*See also* **Limit Theorems**)

PAUL JANSSEN

Lawley–Hotelling Trace

To test the equality of mean vectors of k p-variate normal distributions (*see* **Multivariate Normal Distribution**) with common but unknown **covariance matrix**, Hotelling [14] proposed a test, known as the Hotelling's generalized trace (or T_0^2) test, which could be considered as a generalization of **Hotelling's T^2 test** proposed for $k = 2$. This test statistic was also considered by Lawley [21], Bartlett [4], and Hsu [16]; it is often known as the Lawley–Hotelling trace. The test can be expressed as follows in its canonical form.

Consider random matrices $\mathbf{U} : p \times r$, $\mathbf{V} : p \times m$, and $\mathbf{W} : p \times n$, such that the columns of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ are independently distributed as p-variate normal distributions with a common covariance matrix $\mathbf{\Sigma}$. The problem is to test $\mathrm{H}_0 : \mathbf{\Theta} \equiv \mathrm{E}(\mathbf{U}) = \mathbf{0}$ against $\mathrm{H}_1 : \mathbf{\Theta} \neq \mathbf{0}$, given that $\mathrm{E}(\mathbf{W}) = \mathbf{0}$, $\mathbf{\Sigma}$ being unknown. The Lawley–Hotelling's trace test rejects H_0 if and only if

$$T_0^2 \equiv \mathrm{trace}[(\mathbf{U}\mathbf{U}')(\mathbf{W}\mathbf{W}')^{-1}]n$$

is too large; it is assumed that $n \geq p$. The **multivariate analysis of variance** (MANOVA) problem can be reduced to the above canonical form; in that case, $\mathbf{U}\mathbf{U}'$ and $\mathbf{W}\mathbf{W}'$ denote the sums of products and cross products matrices due to the hypothesis H_0 and due to error, respectively. This trace test can be deduced from Roy's **union–intersection principle**; see Mudholkar et al. [25].

The Lawley–Hotelling trace criterion can be considered for testing independence between two sets of variates jointly distributed as a normal distribution, as well as for testing equality of covariance matrices of two p-variate normal distributions; for this correspondence, see the article on **Pillai's trace test** and the review papers by Pillai [30, 31].

Hotelling [15] derived an explicit form of the null distribution of T_0^2 for $p = 2$; see Hsu [16] and Anderson [1]. Constantine [5] expressed the density of T_0^2 as an infinite series in generalized Laguerre polynomials, and also as an infinite series in zonal polynomials; for details, see Muirhead [27]. Pillai [29] suggested to approximate the null distribution of T_0^2 as follows:

$$F_{\nu_1,\nu_2} = \frac{\nu_2}{\nu_1} \times \frac{T_0^2}{ns},$$

where $s = \min(p, r)$, $\nu_1 = s(t + s + 1)$, $\nu_2 = s(n - p - 1)$, $t = |r - p| - 1$, and $F_{a,b}$ denotes the ***F* distribution** with a and b df; for details on this approximation, see Pillai [29].

Tables of the significance points of T_0^2 have been given by Grubbs [13] for $p = 2$, and by Davis [7–9] for $p = 3(1)10$; see Anderson [1]. Approximate significance points of T_0^2 have been suggested by Pillai [29]; see also Pillai & Samson [33] and Hughes & Saw [17].

The asymptotic (as $n \to \infty$) null distribution of T_0^2 is the **chi-square distribution** with rp **degrees of freedom**. For asymptotic expansion and approximation of the null (nonnull) distribution of T_0^2 in terms of chi-square (noncentral chi-square) distributions, see Itô [18, 19], Fujikoshi [10, 11], Siotani [37], Davis [8], Pillai [30, 31], Khatri & Pillai [20], Pillai & Young [35], and Muirhead [26, 27], in particular.

The power function of the Lawley–Hotelling trace test depends on the parameters through the characteristic roots $\nu_1, \ldots, \nu_p$ of $\mathbf{\Theta}\mathbf{\Theta}'\mathbf{\Sigma}^{-1}$ in the above set-up (*see* **Eigenvalue**). Ghosh [12] has shown that this test is admissible; for a different proof, see Anderson [1]. DasGupta et al. [6] have shown that the **power** of this test increases monotonically as each of the ν_i's increases. For monotonicity of the power function of the trace test for testing independence between two sets of variates, see Anderson & DasGupta [3], and for testing equality of covariance matrices, see Anderson & DasGupta [2]. Simultaneous confidence regions for $\mathbf{\Theta}$ based on the Lawley–Hotelling's trace criterion are given in [1].

The power of the Lawley–Hotelling test has been compared with the powers of the other standard tests for the MANOVA problem by a number of authors; see Mikhail [24], Schatzoff [36], Pillai & Jayachandran [32], Fujikoshi [11], and Lee [22]. If the characteristic roots ν_i are substantially unequal such that the coefficient of variation of the roots is large enough, then the Lawley–Hotelling test is more powerful than the **likelihood ratio test**, which, in turn is more powerful then Pillai's trace test; the reverse is true if the ν_i's are close. Lee [22] has noted that the power of the Lawley–Hotelling test is nearly constant on the region trace $[\mathbf{\Theta}\mathbf{\Theta}'\mathbf{\Sigma}^{-1}] =$ constant. For **robustness** of the Lawley–Hotelling test, see Mardia [23], Olson [28], Pillai & Sudjana [34], the review papers by Pillai [30, 31], and the article on **robustness of multivariate techniques**.

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Anderson, T.W. & DasGupta S. (1964). A monotonicity property of the power functions of some tests of the equality of two covariance matrices, *Annals of Mathematical Statistics* **35**, 1059–1063.

[3] Anderson, T.W. & DasGupta, S. (1964). Monotonicity of power functions of some tests of independence between two sets of variates, *Annals of Mathematical Statistics* **35**, 206–208.

[4] Bartlett, M.S. (1934). A note on tests of significance in multivariate analysis, *Proceedings of the Cambridge Philosophical Society* **34**, 33–40.

[5] Constantine, A.G. (1966). The distribution of Hotelling's generalized T_0^2. *Annals of Mathematical Statistics* **37**, 215–225.

[6] DasGupta, S., Anderson, T.W. & Mudholkar, G.S. (1964). Monotonicity of the power functions of some tests of multivariate linear hypothesis, *Annals of Mathematical Statistics* **36**, 1174–1184.

[7] Davis, A.W. (1970). Exact distribution of Hotelling's generalized T_0^2, *Biometrika* **57**, 187–191.

[8] Davis, A.W. (1970). Further applications of a differential equation for Hotelling's generalized T_0^2, *Annals of the Institute of Statistical Mathematics* **22**, 77–87.

[9] Davis, A.W. (1980). Further tabulation of Hotelling's generalized T_0^2, *Communications in Statistics – Simulation and Computation* **9**, 321–336.

[10] Fujikoshi, Y. (1970). Asymptotic expansions of the distributions of test statistics in multivariate analysis, *Journal of Science Hiroshima University, Series A-1* **32**, 293–299.

[11] Fujikoshi, Y. (1973). Asymptotic formulas for the distributions of three statistics for multivariate linear hypothesis, *Annals of the Institute of Statistical Mathematics* **25**, 423–437.

[12] Ghosh, M.N. (1964). On the admissibility of some tests of MANOVA, *Annals of Mathematical Statistics* **35**, 789–794.

[13] Grubbs, G.E. (1954). Tables of 1% and 5% probability levels of Hotelling's generalized T^2 statistic, *Technical note no. 926*, Ballistic Research Laboratory, Aberdeen, Proving Ground, Maryland.

[14] Hotelling, H. (1947). Multivariate quality control, illustrated by the air-testing of sample bombsights, in *Techniques of Statistical Analysis*, C. Eisenhart, M.W. Hastay & W.A. Wallis, eds. McGraw-Hill, New York, pp. 111–184.

[15] Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 23–41.

[16] Hsu, P.L. (1940). On generalized analysis of variance (I), *Biometrika* **31**, 221–237.

[17] Hughes, D.T. & Saw, J.G. (1972). Approximating the percentage points of Hotelling's generalized T_0^2 statistic, *Biometrika* **59**, 224–226.

[18] Itô, K. (1956). Asymptotic formulae for the distribution of Hotelling's generalized T_0^2 statistic, *Annals of Mathematical Statistics* **27**, 1091–1105.

[19] Itô, K. (1960). Asymptotic formulae for the distribution of Hotelling's generalized T_0^2 statistic, II, *Annals of Mathematical Statistics* **31**, 1148–1153.

[20] Khatri, C.G. & Pillai, K.C.S. (1966). On the moments of trace of a matrix and approximations to its noncentral distribution, *Annals of Mathematical Statistics* **37**, 1312–1318.

[21] Lawley, D.N. (1938). A generalization of Fisher's Z-test, *Biometrika* **30**, 180–187.

[22] Lee, Y.S. (1971). Asymptotic formulae for the distributions of a multivariate test statistic: power comparison of some multivariate tests, *Biometrika* **58**, 647–651.

[23] Mardia, K.V. (1971). The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model, *Biometrika* **58**, 105–127.

[24] Mikhail, N.N. (1965). A comparison of tests of the Wilks–Lawley hypothesis in multivariate analysis, *Biometrika* **52**, 149–156.

[25] Mudholkar, G.S., Davidson, M.L. & Subbiah, P. (1974). A note on the union–intersection character of some MANOVA procedures, *Journal of Multivariate Analysis* **4**, 486–493.

[26] Muirhead, R.J. (1970). Asymptotic distributions of some multivariate tests, *Annals of Mathematical Statistics* **41**, 1002–1010.

[27] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

[28] Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

[29] Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis, *Annals of Mathematical Statistics* **26**, 117–121.

[30] Pillai, K.C.S. (1976). Distributions of characteristic roots in multivariate analysis, part I: null distributions, *Canadian Journal of Statistics* **4**, 157–184.

[31] Pillai, K.C.S. (1977). Distributions of characteristic roots in multivariate analysis, part II: non-null distributions, *Canadian Journal of Statistics* **5**, 1–62.

[32] Pillai, K.C.S. & Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria, *Biometrika* **54**, 195–210.

[33] Pillai, K.C.S. & Samson, P. (1959). On Hotelling's generalization of T^2, *Biometrika* **46**, 160–168.

[34] Pillai, K.C.S. & Sudjana (1975). Exact robustness studies of tests of two multivariate hypotheses based on four criteria and their distribution problems under violations, *Annals of Statistics* **3**, 617–638.

[35] Pillai, K.C.S. & Young, D.L. (1966). On the exact distribution of Hotelling's generalized T_0^2, *Journal of Multivariate Analysis* **1**, 90–107.

[36] Schatzoff, M. (1966). Sensitivity comparisons among tests of the general linear hypotheses, *Journal of the American Statistical Association* **61**, 415–435.

[37] Siotani, M. (1971). An asymptotic expansion of the non-null distributions of Hotelling's generalized T_0^2 statistic, *Annals of Mathematical Statistics* **42**, 560–571.

(*See also* **Lambda Criterion, Wilks'**; **Multivariate Analysis, Overview**)

SOMESH DASGUPTA

League Tables

When making comparisons between, say, hospitals, it is very tempting to summarize the results into a set of scores, which are then sorted into a "league table", thus giving an explicit rank to each hospital (*see* **Profiling Providers of Medical Care**). This follows the sporting model in which attention is focused on which "team" is at the top and which at the bottom, and what movements have occurred since the last ranking exercise. Consider, for example, risk-adjusted 30-day mortality rates following cardiac artery bypass grafts in New York State between 1997 and 1999. Figure 1 shows the ranked rates for individual surgeons with 95% **confidence intervals** – in the original publication [6] the surgeons are named.

Such exercises can be easily criticized: someone always has to be top and bottom, even if there is really no difference between the surgeons and the league table is only the consequence of the play of chance. The widths of the intervals express the considerable uncertainty about the true underlying mortality rates, and yet none of this uncertainty is reflected in the rank given to a surgeon. The crucial insight is that the rank is itself a summary statistic that can be subject to standard **inferential** procedures [3–5], although the **sampling distributions** of observed ranks are not generally amenable to closed-form analysis. The simplest solution is to use **Monte Carlo** techniques: essentially, by thinking of the intervals in Figure 1 as expressing probability distributions for the true mortality rates, sampling from those distributions, ranking each of the generated samples, and so obtaining a set of plausible "true ranks" for the surgeons. This can be a fully **Bayesian** procedure in which the intervals summarize posterior distributions, or an approximate **maximum likelihood** analysis based on a "parametric **bootstrap**''.

Figure 2 shows the results of the ranking exercise (carried out using a Bayesian **Markov chain Monte Carlo** technique). It is clear there is substantial uncertainty concerning the true ranks, with the majority of surgeons having a very wide interval; only 2 out of

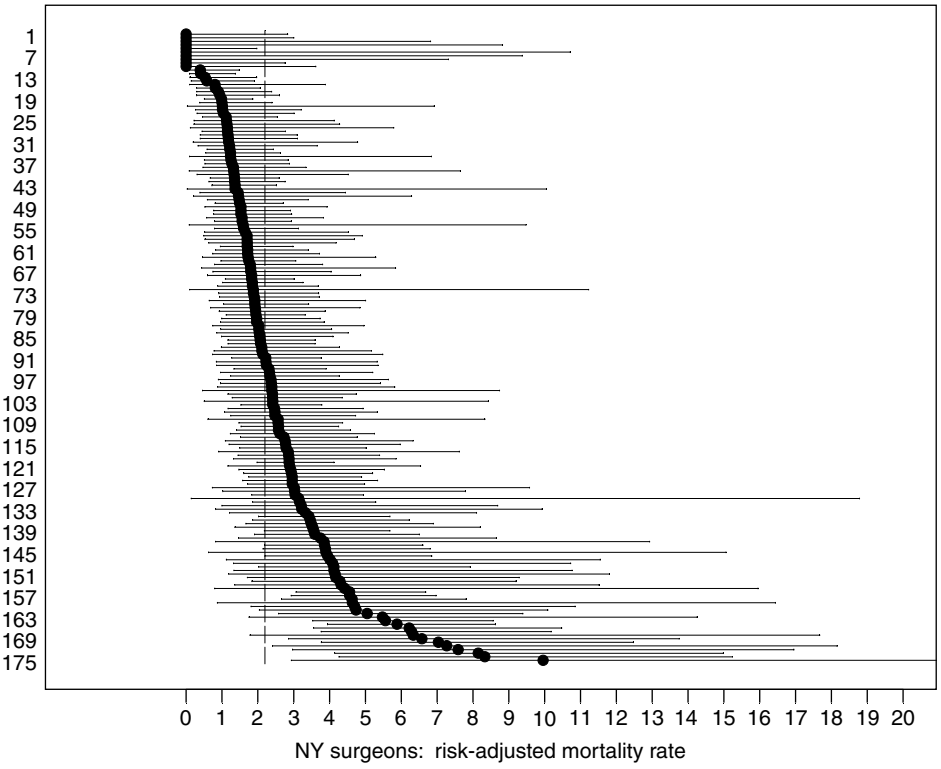

Figure 1 Observed mortality rates for 175 surgeons, with 95% confidence intervals

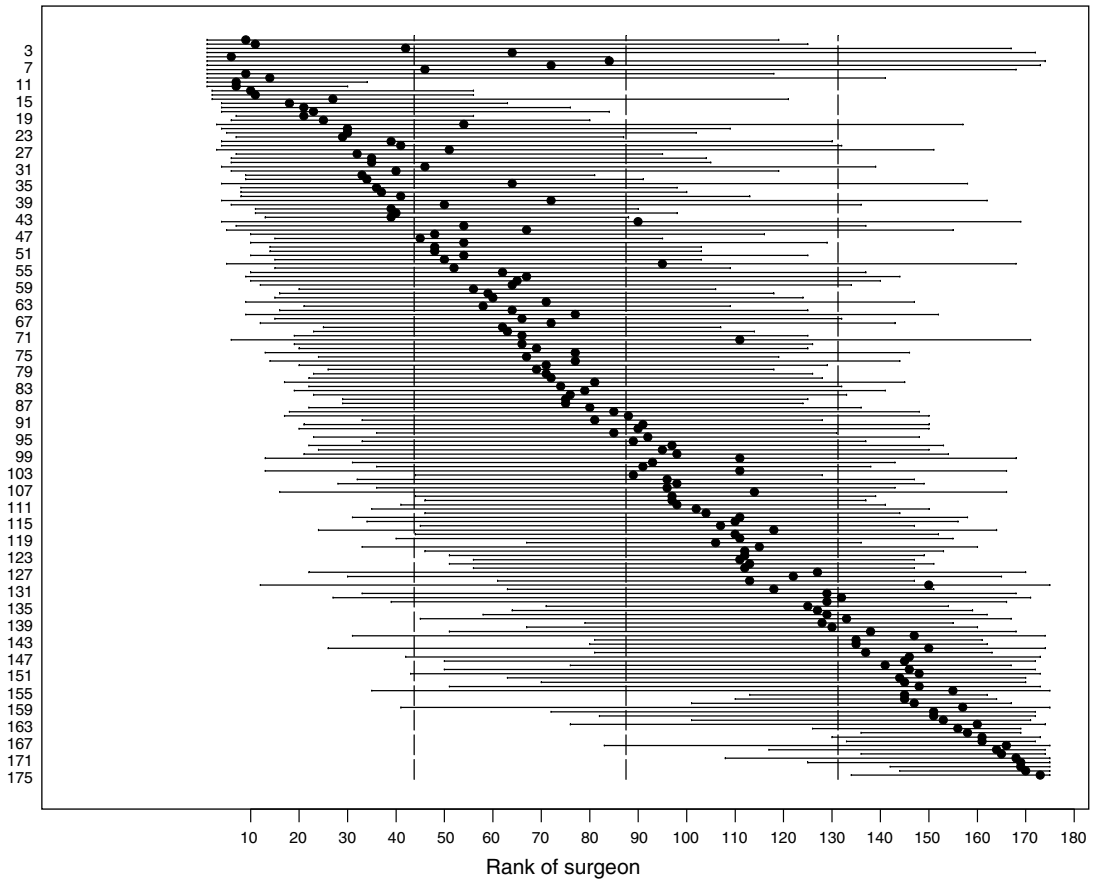

Figure 2 Median estimates and 95% intervals for true ranks of 175 New York surgeons

175 can be confidently placed in the "best" quarter, and only 6 in the "worst" quarter.

This procedure is straightforward to carry out if the league table has been based on a single indicator with a known sampling distribution. In practice, a league table may be based on a composite score, which may comprise both objective measures and subjective assessments of, for example, "reputation". Nevertheless, it should still be possible to assign a plausible measure of error to each of the items making up the score, and hence simulate a distribution of the "true" score, and obtain a distribution for the "true" rank. The ranking procedure can also apply after more complex models have been fitted. In particular, **hierarchical models** are often recommended in this context [1–3, 7] as a means of dealing with the inadequacies of risk-adjustment and "**regression-to-the-mean**", and the resulting "shrunk" estimates (*see* **Shrinkage Estimation**) will tend to lead to even more uncertainty about the true ranks.

Inferences on ranks may be useful in many other contexts, such as selecting promising entities for further research from among a large number of, say, drug compounds, crop varieties, or genes being screened.

The consequence of formal inference on ranks is generally to realize that little can be said and that any "league table" is largely spurious, apart from possibly identifying some extreme cases that can confidently be placed in, say, the top or bottom quarter. An alternative comparative tool that does not make an explicit ranking and yet draws attention to genuine extremes is the *funnel plot* (*see* **Meta-analysis of Clinical Trials**), which can be thought of as a Shewhart control chart (*see* **Quality Control in Laboratory Medicine**) around a target performance measure [8]. The outcome is plotted against volume and predictive

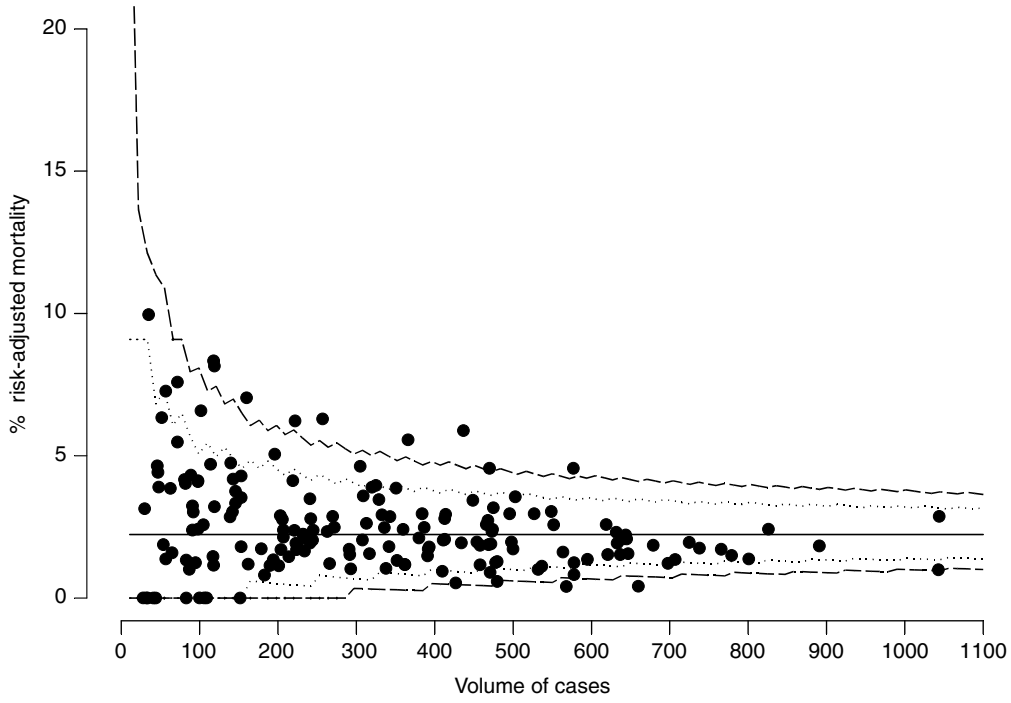

Figure 3 Funnel plot of risk-adjusted mortality rates for 175 New York surgeons: 95% and 99.9% prediction intervals are shown

limits superimposed, say 95% (≈2 standard deviations) and 99.9% (≈3 standard deviations). Figure 3 shows this applied to the New York surgeons data, and it is immediately apparent that the vast majority follow precisely the pattern expected and the only ones worthy of some attention are easily identified. The plot makes clear that there is no point in carrying out a ranking exercise on those in the "funnel".

References

[1] Burgess, J., Christiansen, C., Michalak, S. & Morris, C. (2000). Medical profiling: improving standards and risk adjustments using hierarchical models, *Journal of Health Economics* **19**, 291–309.

[2] Christiansen, C. & Morris, C. (1997). Improving the statistical approach to health care provider profiling, *Annals of Internal Medicine* **127**, 764–768.

[3] Goldstein, H. & Spiegelhalter, D.J. (1996). Statistical aspects of institutional performance: league tables and their limitations (with discussion), *Journal of the Royal Statistical Society, Series A* **159**, 385–444.

[4] Marshall, E.C. & Spiegelhalter, D.J. (1998). Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates, *British Medical Journal* **317**, 1701–1704.

[5] Morris, C.N. & Christiansen, C.L. (1996). Hierarchical models for ranking and for identifying extremes, with applications, in *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Clarendon Press, Oxford, pp. 277–296.

[6] New York State Department of Health (2002). *Coronary Artery Bypass Surgery in New York State, 1997-9*, New York State Department of Health: `http://www.health.state.ny.us/nysdoh/heart/heart_disease.htm`, Albany: New York.

[7] Shahian, D., Normand, S., Torchiana, D., Lewis, S., Pastroe, J., Kuntz, R. & Dreyer, P. (2001). Cardiac surgery report cards: comprehensive review and statistical critique, *Annals Thoracic Surgery* **72**, 1845–1848.

[8] Spiegelhalter, D.J. (2002). Funnel plots for institutional comparisons (letter), *Quality Safety in Health Care* **11**, 390–391.

DAVID J. SPIEGELHALTER

Least Squares

Because of the controversy between Legendre and **Gauss** about the priority in the discovery of least squares, it is useful to distinguish between the principle of least squares and the theory of least squares.

Let $\mathbf{y}, \boldsymbol{\xi}(\boldsymbol{\theta})$, and $\mathbf{e}$, be $n \times 1$ vectors of observations, of known parametric functions of $\boldsymbol{\theta}$, and of random observational or experimental errors, respectively, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of unknown parameters, $k < n$. The model is

$$\mathbf{y} = \boldsymbol{\xi}(\boldsymbol{\theta}) + \sigma \mathbf{e}, \tag{1}$$

where σ is a scalar parameter specifying the measurement scale. The problem is to estimate $\boldsymbol{\theta}$. If the observations were free from errors, $\mathbf{e} \equiv \mathbf{0}$, and if the model were exactly correct, the resulting n equations $\mathbf{y} = \boldsymbol{\xi}(\boldsymbol{\theta})$ in the k parameters $\boldsymbol{\theta}$ would have to be consistent in the mathematical sense that there exists a value of $\boldsymbol{\theta}$ satisfying all n equations. However, in general this is not the case, and the equations are inconsistent – there is no value of $\boldsymbol{\theta}$ satisfying (1). The observations are assumed to have been taken with equal care, so that all observations should contribute equally to the estimation of $\boldsymbol{\theta}$. The problem is therefore to combine the observations to extract all of their information about $\boldsymbol{\theta}$. This problem used to be referred to as the combination of observations. There is no value of $\boldsymbol{\theta}$ that in general minimizes the errors $\mathbf{e}$ uniformly. They therefore have to be minimized in some global sense. Laplace suggested minimizing the sum of absolute deviations $\sum |e_i|$, sometimes called L_1 regression (*see* **Robust Regression**). The principle of least squares minimizes the sum of squared deviations $Q = \sum e_i^2 = \sum [y_i - \boldsymbol{\xi}_i(\boldsymbol{\theta})]^2$. The resulting least squares estimate, $\hat{\boldsymbol{\theta}}$, is a solution of the least square equations $\partial Q / \partial \boldsymbol{\theta}_j = \sum (y_i - \boldsymbol{\xi}_i) \partial \boldsymbol{\xi}_i / \partial \boldsymbol{\theta}_j = 0$.

In the special but widespread case where $\boldsymbol{\xi}(\boldsymbol{\theta})$ is a linear function of $\boldsymbol{\theta}$, $\boldsymbol{\xi} = \mathbf{X}\boldsymbol{\theta}$, $\mathbf{X}$ being an $n \times k$ matrix of rank k of known constants x_{ij}, (1) becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \sigma \mathbf{e}, \tag{2}$$

the Gauss linear model, discussed in text books as linear regression. The corresponding least squares equation and least squares estimate are

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{y}, \qquad \hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3}$$

Legendre has priority in the principle of least squares, having published in 1805 [4], whereas Gauss's first publication on least squares was in 1809 [2]. Legendre derived the models $\mathbf{y} = \boldsymbol{\xi} + \sigma \mathbf{e}$ and the special case (2), and the corresponding least squares equations including (3). But Gauss [2, 3] went on to develop the theory of least squares, producing the treatment of linear regression as given in textbooks today. In fact, according to Fisher [1, p.88], Gauss's method only lacked for completeness the refinement of the use of **Student's *t* distribution**, appropriate for samples of rather small numbers of observations.

Standard Normal Errors

This is Gauss's first, or parametric inferential, approach [2]. Gauss assumed the errors $\mathbf{e}$ to be n independent random variables having density function $\prod f(e_i) = \prod f(y_i - \boldsymbol{\xi}_i)$. He then used a typically Bayesian argument assuming independent uniform prior distributions for $\boldsymbol{\theta}_i$ to obtain a posterior density function for $\boldsymbol{\theta}$. The estimate $\hat{\boldsymbol{\theta}}$ was chosen to be the mode of this posterior density function. Since the posterior distribution of $\boldsymbol{\theta}$ is proportional to $\prod f(e_i)$, which is proportional to the likelihood function of $\boldsymbol{\theta}$, the resulting equations of estimation are

$$\sum_{i=1}^{n} \left[\frac{\partial f(y_i - \boldsymbol{\xi}_i)}{\partial \boldsymbol{\xi}_i} \right] \left(\frac{\partial \boldsymbol{\xi}_i}{\partial \boldsymbol{\theta}_j} \right), \quad j = 1, \ldots, n, \tag{4}$$

which are equivalent to the equations of maximum likelihood, and $\hat{\boldsymbol{\theta}}$ is the **maximum likelihood** estimate. To proceed further requires specifying f. To do this Gauss assumed that for the single-parameter model in which $\boldsymbol{\xi}_i \equiv \boldsymbol{\theta}, i = 1, \ldots, n$, the appropriate estimate of the scalar $\boldsymbol{\theta}$ is $\overline{y}$. This was the procedure commonly used in the physical sciences, such as astronomy. Substituting $\boldsymbol{\xi}_i \equiv \boldsymbol{\theta}$ into (4), and requiring the solution $\hat{\boldsymbol{\theta}}$ to be $\overline{y}$, implies that $f(e_i)$ is the normal N(0, 1) density. Then $\prod f(e_i) \propto \exp(-Q^2/2)$, where $Q = \sum (y_i - \boldsymbol{\xi}_i)^2$. Maximizing f is equivalent to minimizing Q, which is the method of least squares. Gauss then specialized to the linear model (2), produced (3), and went on to develop the theory of normal linear regression as presented in textbooks today. He also generalized (1) to $\sigma = \text{diag}(\sigma_i)$, where the observations have different scale

parameters σ_i. This results in weighted least squares in which $Q = \sum[y_i - \xi_i(\boldsymbol{\theta})]^2/\sigma_i^2$ is to be minimized. This approach to least squares can be considered as a generalization of the use of the arithmetic **mean**. It is based upon the normal error distribution, in which case its use is equivalent to the use of the more general method of maximum likelihood.

Unspecified Distribution

This is Gauss's second, or nonparametric decision-theoretic, approach [3] based on the following assumptions.

1. The model. The model is (2) where the errors have zero means, unit variances, and are uncorrelated.
2. Linear estimates. The measuring instrument is sufficiently precise that the squares and higher powers of the errors **e** can be ignored, thus restricting attention to linear error-consistent estimates
$$\tilde{\boldsymbol{\theta}} = \mathbf{C}\mathbf{y}, \quad \text{where } \mathbf{C}\mathbf{X} = \mathbf{I}, \tag{5}$$
and **C** is a $k \times n$ matrix. The error-consistency requirement, $\mathbf{C}\mathbf{X} = \mathbf{I}$, is to ensure that when the observations are free from errors, $\mathbf{e} \equiv \mathbf{0}$, the estimate should be the true value, $\tilde{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}$.
3. Squared error loss. Estimation is a game with a potential loss (*see* **Loss Function**) and no hope for gain. The loss is taken to be proportional to the squared error $(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^2$.

The requirement is that the estimate (5) should minimize the expected loss, the well-known **mean square error** (EMS) criterion, $\mathrm{E}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^2$. Gauss's Theorem proves that among all estimates (5), the least squares estimate $\hat{\boldsymbol{\theta}}$ minimizes the EMS.

Gauss then went on to examine various decompositions of sums of squares, and essentially develop the **analysis of variance**. He also showed that if $Q_{\mathrm{m}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})$ is the residual or minimum sum of squares, then $\mathrm{E}(Q_{\mathrm{m}}) = (n-k)\sigma^2$. He thus recommended $s = [Q_{\mathrm{m}}/(n-k)]^{1/2}$ as the appropriate estimate of σ. Notice that while s^2 is an unbiased estimate of σ^2, s is a biased estimate of σ. Thus Gauss did not seem to be preoccupied with unbiased estimates.

Discussion

This last point is particularly relevant as textbooks almost always treat (5) as a requirement of **unbiasedness**, implying that the purpose of least squares is to seek **minimum variance unbiased estimates**. The resulting formalized theorem is usually referred to as the Gauss–Markov Theorem.

Textbooks also usually ignore the justification of the linearity requirement (assumption 2), and assume it is simply reasonable to restrict attention to linear estimates. This, together with unbiasedness and variance, lead to best linear unbiased estimates (BLUEs) and uniformly minimum variance (UMV) estimates. Justification for such estimates thus appears to be based more on their mathematical convenience than on their scientific relevance.

The domain of the application of least squares to parametric models is (1) where **e** is a vector of independent N(0,1) errors. In this case least squares is identical to maximum likelihood. When **e** is not normal, least squares is no longer relevant, but maximum likelihood usually is applicable to the estimation of components $\boldsymbol{\theta}_i$ if due attention is paid to the shape of the likelihood functions of $\boldsymbol{\theta}_i$.

The domain of application of the nonparametric approach, where the model is not specified, is to areas where it is appropriate to assume **e** is small, so that it makes sense to restrict attention to linear estimates and to a squared error loss function. For example, Gauss originally applied least squares to the calculation of a planetary orbit, and the prediction of where the planet will be seen. He also used it in map-making, or geodesy, where the above assumptions are fulfilled. These assumptions may not be generally appropriate in many applications in biology and the life sciences, and hence in biostatistics. There the principal source of error is usually the variability of the experimental material, which may be large and asymmetric.

These considerations make it seem unlikely that least squares, as a method of estimation per se, has a widespread application in biostatistics. However, *weighted* least squares has computational applications in iterative procedures required to find the maximum likelihood estimate (*see* **Optimization and Nonlinear Equations**). In particular it is useful in obtaining an initial value $\boldsymbol{\theta}_0$ to start the iterations using the Newton–Raphson or Fisher's scoring methods. For example, consider the binomial **logistic**

regression $\mathbf{y} \sim \text{bin}(\mathbf{s}, \mathbf{p})$, where $\zeta = \log[\mathbf{p}/(1 - \mathbf{p})] = \mathbf{X}\boldsymbol{\theta}$. If $\hat{\zeta} = \log[\hat{\mathbf{p}}/(1 - \hat{\mathbf{p}})] = \log[\mathbf{y}/(\mathbf{s} - \mathbf{y})]$, an initial value $\boldsymbol{\theta}^{(0)}$ is the weighted least squares estimate of $\boldsymbol{\theta}$ in the regression of $\hat{\zeta}$ on $\mathbf{X}$. This is obtained by minimizing with respect to $\boldsymbol{\theta}$ the weighted sum of squares

$$Q = \sum_{i=1}^{n} \left(\hat{\zeta} - \sum_{j=1}^{k} x_{ij}\theta_j \right)^2 n_i \hat{p}_i (1 - \hat{p}_i)$$
$$= (\hat{\zeta} - \mathbf{X}\boldsymbol{\theta})' D (\hat{\zeta} - \mathbf{X}\boldsymbol{\theta}),$$

where D is the diagonal matrix of elements $n_i \hat{p}_i (1 - \hat{p}_i)$, the reciprocals of the estimated variances of the $\hat{\zeta}_i$. The successive correction terms in the Newton–Raphson iterations can be obtained in a similar way (*see* **Optimization and Nonlinear Equations**).

For further details on Gauss and least squares, see Sprott [5].

References

[1] Fisher, R.A. (1973). *Statistical Methods and Scientific Inference*. Hafner, New York.

[2] Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium*. Werke 7. (English translation: C.H. Davis. Dover, New York, 1963).

[3] Gauss, C.F. (1821, 1823, 1826). *Theoria Combinationis Erroribus Minimis Obnoxiae*, Parts 1, 2, and Supplement. Werke 4, pp. 1–108.

[4] Legendre, A.M. (1805). Nouvelles méthodes pour la determination des orbits des cométes. (Appendix: Sur la méthode des moindres carrés.).

[5] Sprott. D.A. (1978). Gauss's contributions to statistics, *Historia Mathematica* **5**, 183–203.

(*See also* **Estimation**)

D.A. SPROTT

Lehmann Alternatives

To study the effect of a drug one often needs to make a **nonparametric** comparison of the cumulative distribution function, G, of **scores** from treated subjects, with the distribution, F, of scores from the untreated control group. In such comparisons, Lehmann [5] pointed out that under any alternative hypothesis the test is not distribution-free. To overcome this difficulty, Lehmann suggested a functional relationship $G = f(F)$, where f is a specified function between G and F. The importance of this formulation is that the distribution of the **rank** vector under the alternative hypothesis will depend only on f and hence every rank statistic will have a nonparametric distribution-free property. As an illustration, Lehmann derived the power of various rank tests for simple alternatives $G = F^2$ and $G = F^3$.

One of the major applications of the Lehmann alternatives has been in the formulation of the alternative hypothesis for testing the effect of a drug when some subjects in the treatment group are not affected by the treatment. This so-called "nonresponse" phenomenon occurs, for example, in the development of a new drug. In such studies, one may exhibit greater variability as well as mean response in the treatment group. This increased variability can be considered to be due to the presence of subjects in the treatment group who are unaffected by the treatment. Thus, if p is the proportion of subjects in the treatment group who respond to the treatment, then Salsburg [9] suggests testing the null hypothesis of no treatment effect,

$$\mathrm{H}_0\colon G(x) = F(x), \tag{1}$$

against a Lehmann-type alternative of the form

$$\mathrm{H}_1\colon G(x) = (1-p)F(x) + p[F(x)]^{\gamma}, \tag{2}$$

where $\gamma > 1$ is a known constant. Salsburg's argument is based on maintaining the same range of observations in the treatment and control groups. In this form of the Lehmann alternative the response of each subject in the treatment group who is affected by treatment is assumed to have the same distribution as the maximum of γ responses in the control group. Salsburg suggests using a rank test, where the two samples are ranked together and the score for a given subject is

$$s(i) = \left(\frac{i}{N+1}\right)^k, \quad k > 1, \tag{3}$$

where i is the combined rank of that subject and N is the total number of subjects in the combined set. Conover & Salsburg [1] show that when $\gamma = 5$ and $k = \gamma - 1 = 4$, then a rank test based on (3) provides a test with maximum **asymptotic relative efficiency**. The recommended value of $\gamma = 5$ is based on some empirical results. Razzaghi & Nanthakumar [8] formulated the problem of testing for treatment effect in terms of the parameter p as

$$\mathrm{H}_0\colon p = 0$$

against

$$\mathrm{H}_1\colon p > 0$$

and derived a **locally most powerful test**. Such a test is based on the statistic

$$S_{n,m} = \sum_{i=1}^{n} F_m^{\gamma-1}(y_i), \tag{4}$$

where m is the number of subjects in the control group and F_m is the empirical distribution of the control observations defined at any point x as the proportion of the control observations not exceeding x, and $y_1, \ldots, y_n$ are the observations from the treatment group. The test of H_0 against H_1 will reject the hypothesis of no treatment effect when

$$S_{n,m} > \frac{n}{\gamma} + \frac{\gamma-1}{\gamma}\left(\frac{n}{2\gamma-1}\right)^{1/2} Z_\alpha,$$

where Z_α is the $100(1-\alpha)$th fractile (or **quantile**) of the **standard normal distribution.** The power of the test based on (4) increases as the proportion of the responders, p, rises. A value of $\gamma = 5$ is again recommended on the basis of an analysis of the **power** of the test.

Conover & Salsburg [1] also present an argument for using the other form of the Lehmann alternative and expressed the distribution of the treatment in the presence of nonresponders as the hypothesis

$$\mathrm{H}_2\colon (1-p)F(x) + p\{1 - [1-F(x)]^{1/\gamma}\}, \quad \gamma > 1, \tag{5}$$

which implies that the distribution of each control score is assumed to have the same distribution as the minimum of γ responses from the treatment group. Razzaghi & Nanthakumar [7] proposed a locally optimal test for the alternative (5). The test is based on the score statistic (*see* **Likelihood**)

$$T_n = \sum_{i=1}^{n} [1 - F(y_i)]^{-(\gamma-1)/\gamma}. \tag{6}$$

It is shown that when $\gamma = 2$, the asymptotic distribution of T_n under the null hypothesis is normal, while for $\alpha > 2$ this distribution is in the domain of attraction of a stable distribution. More specifically, a locally most powerful test of H_0 against H_2 with $\gamma = 2$ rejects the null hypothesis, H_0, when

$$T_n > 2n\left[1 - (n \ln n)^{-1/2}\right] + (n \ln n)^{1/2} Z_\alpha, \tag{7}$$

and the test for $\gamma > 2$ rejects H_0 when

$$\frac{\{T_n - n\gamma(1 - a_n^{-(\gamma-1)^{-1}})\}}{a_n} \tag{8}$$

exceeds the $(1 - \alpha)$th fractile of a stable distribution with indices $\gamma/(\gamma - 1)$ and -1. In (8), a_n is given by

$$a_n = \frac{\left\{\left[\frac{8\gamma n}{\ln n}\right]^{(\gamma-1)/\gamma} - \left[\frac{16n}{\ln n}\right]^{1/2}\right\}}{(\gamma - 2)^{2(\gamma-1)/\gamma}}. \tag{9}$$

Example

To illustrate the methodologies described here, we use a data set from an experiment on **pain** scores. The data first appeared in Conover & Salsburg [1]. Values for patients from a study of acute painful diabetic neuropathy were recorded at baseline and after four weeks of treatment on an analog scale. The changes from baseline were described as the natural logarithm of the ratio of baseline to final scores. Table 1 is reproduced from Conover & Salsburg for completeness. For these data using $\gamma = 5$, the value of $s(i)$ is 0.261 for the treated patients and 0.133 for the control subjects, leading to a P value of 0.031. The value of $S_{n,m}$ is 2.10, leading to a P value of 0.0179. Computation of T_n for these data gives

$$T_n = \begin{cases} 100.13, & \gamma = 2, \\ 210.31, & \gamma = 3, \\ 421.28, & \gamma = 5, \end{cases}$$

and in all cases the test indicates a highly significant treatment effect. The fractiles of a stable distribution may be obtained from Cross [3].

There is a vast body of literature on the theoretical development and applications of Lehmann alternatives. Wijsman [10] provides a comprehensive and thorough list of early references. Halperin & Ware [4], Cox [2], and Peto [6] discuss the use of Lehmann alternatives in the analysis of data from **clinical trials**. The intent here has been to demonstrate more recent applications in biostatistical problems.

References

[1] Conover, W.J. & Salsburg, D.S. (1988). Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to "respond to treatment", *Biometrics* **44**, 189–196.

Table 1 Pain scores ln(baseline/final)

Treated patients, Y_i						
−1.535	−0.547	−0.201	−0.201	−0.154	−0.095	−0.049
0.000	0.000	0.000	0.105	0.111	0.201	0.251
0.310	0.406	0.511	0.531	0.575	0.575	0.773
0.981	1.299	1.299	1.322	1.386	1.792	2.398
Controls, X_i						
−1.490	−0.021	−0.128	−0.087	−0.054	0.000	0.000
0.000	0.000	0.000	0.028	0.039	0.049	0.061
0.080	0.105	0.134	0.193	0.216	0.223	0.273
0.288	0.330	0.357	0.487	0.541	0.793	1.042
1.099	1.609					

Reproduced from [1] by permission of the publisher.

[2] Cox, D.R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society, Series B* **39**, 79–85.

[3] Cross, M.J. (1973). Tables of finite-mean nonsymmetric stable distributions as computed from their convergent and asymptotic series, *Journal of Statistical Computation and Simulation* **3**, 1–27.

[4] Halperin, M. & Ware, J. (1974). Early decisions in a censored Wilcoxon two-sample test for accumulating survival data, *Journal of the American Statistical Association* **69**, 414–422.

[5] Lehmann, E.L. (1953). The power of rank tests, *Annals of Mathematical Statistics* **24**, 23–43.

[6] Peto, R. (1972). Rank tests of maximal power against Lehmann-type alternatives, *Biometrika* **59**, 467–471.

[7] Razzaghi, M. & Nanthakumar, A. (1992). On using Lehmann alternatives with nonresponders, *Mathematical Biosciences* **109**, 69–83.

[8] Razzaghi, M. & Nanthakumar, A. (1994). A locally most powerful test for detecting a treatment effect in the presence of nonresponders, *Biometrical Journal* **3**, 373–384.

[9] Salsburg, D.S. (1986). Alternative hypotheses for the effects of drugs in small-scale clinical studies, *Biometrics* **42**, 671–674.

[10] Wijsman, R.A. (1985). Lehmann alternatives, in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz & N.L. Johnson, eds. Wiley, New York.

(*See also* **Cure Models**; **Proportional Hazards, Overview**)

MEHDI RAZZAGHI

Length Bias

The length-biased distribution is a **probability** distribution resulting from a biased sampling scheme in which the probability of observing a positive-valued **random variable** is proportional to the value of the variable.

Length Bias in Renewal Theory

The presence of length bias is a natural phenomenon in **renewal** theory. Consider a sequence of random variables

$$X_1, \quad X_1 + X_2, \quad X_1 + X_2 + X_3, \ldots,$$

where the Xs are positive-valued, nondegenerate, independent and identically distributed (iid) random variables. Suppose the process starts from time 0 and is observed at the time τ_0, where τ_0 is a positive constant. Let α be the index so that

$$\sum_{i=1}^{\alpha-1} X_i < \tau_0 \leq \sum_{i=1}^{\alpha} X_i.$$

Let Y, T, and R respectively denote the length of the interval containing τ_0, the backward recurrence time, and the forward recurrence time, or equivalently,

$$Y = X_\alpha, \qquad T = \tau_0 - \left\{\sum_{i=1}^{\alpha-1} X_i\right\},$$

$$R = \left\{\sum_{i=1}^{\alpha} X_i\right\} - \tau_0.$$

Let f, S, and μ represent the density function, survivorship function, and mean of X_1, respectively. When τ_0 is sufficiently large so that an *equilibrium condition* is reached [3], the joint density of (T, R) can then be derived as

$$p_{T,R}(t, r) = \frac{f(t + r)I(t \geq 0, r \geq 0)}{\mu}. \tag{1}$$

The marginal density functions of Y, T, and R can be derived, based on (1), as

$$p_Y(y) = \frac{yf(y)I(y \geq 0)}{\mu}, \tag{2}$$

$$p_T(t) = \frac{S(t)I(t \geq 0)}{\mu}, \tag{3}$$

$$p_R(r) = \frac{S(r)I(r \geq 0)}{\mu}. \tag{4}$$

The distribution of (2) is generally referred to as the *length-biased distribution*.

Although the length-biased distribution in renewal theory is usually derived under the iid assumption on the Xs, as a general result the independence assumption can be removed and the density formulas (1)–(4) still remain valid [6].

Statistical Methods

Length-biased sampling is recognized in many research fields including epidemiology, ecology, and reliability. A number of methods for length-biased data have been developed in the statistical literature. Cox [4] proposed estimating the survivorship function by a weighted empirical distribution function (*see* **Goodness of Fit**), with weight inversely proportional to y_i:

$$\hat{S}_n(y) = n^{-1}\hat{\mu}\sum_{i=1}^{n}\left[y_i^{-1}I(y_i > y)\right],$$

where $\hat{\mu} = \{n^{-1}\sum_j y_j^{-1}\}^{-1}$ serves as an appropriate estimate of μ, since $n^{-1}\sum_j y_j^{-1}$ estimates μ^{-1}. The estimator $\hat{S}_n$ can be proven to be the **nonparametric maximum likelihood estimator** of S, a special case under Vardi's **selection bias** models [12, 13]. Following the same weighting procedure, a kernel estimator of the density function f (*see* **Density Estimation**) was proposed in [7] as

$$\hat{f}_n(y) = n^{-1}\hat{\mu}\sum_{i=1}^{n}\left[y_i^{-1}K_h(y - y_i)\right],$$

where $K_h(x) = h^{-1}K(h^{-1}x)$, $h > 0$, with K a kernel function. Alternatively, one could first estimate the length-biased density, (2), by an ordinary kernel estimator and then use the relationship of (2) and f to obtain an estimator of f [2]. Under the proportional hazards model [5], a risk set sampling technique was developed in [17] for estimating regression parameters. For $y_j \geq y_i$, let $\Delta_j(y_i)$ be a binary variable which equals 1 with probability y_i/y_j, and 0

with probability $1 - (y_i/y_j)$. The indicators $\Delta_j(y_i)$ are used to identify bias-adjusted **risk sets** and to construct **pseudo-likelihood** equations. Regression parameter estimates are then derived by solving the score equations (*see* **Likelihood**).

Length Bias in Prevalent Cohorts

Length-biased sampling could arise in many epidemiologic studies when survival data are collected from a disease population (*see* **Prevalent Case**). As an illustration, suppose a random sample of women with breast cancer (b.c.) are recruited for observation of survival. Assume (i) the rate of occurrence of b.c. remains constant over time, and (ii) the density function of the time from b.c. to death, f, is independent of the calendar time when b.c. occurred. Conditions (i) and (ii) together are referred to as the *equilibrium condition*. Denote by τ_i the calendar time when woman i with b.c. is recruited, t_i the time from the initial diagnosis of b.c. to τ_i, and y_i the time from the initial diagnosis of b.c. to death. Under the equilibrium condition, the joint density of (t_i, y_i) is an equivalent of (1), namely

$$p_{T,Y}(t, y) = \frac{f(y)I(y \geq t \geq 0)}{\mu}, \qquad (5)$$

and the distribution of y_i is length-biased with density (2). Suppose a sample of iid $(t_1, y_1), \ldots, (t_n, y_n)$ is observed. By the factorization theorem, the observed failure times $\{y_i\}$ serve as **sufficient statistics** for parameters of f. In this case, the variables $\{t_i\}$ do not contain additional information for f.

The preceding length-biased sample can be described more generally as disease prevalent data. Suppose there are two chronologically ordered and nonrecurrent events, termed the initiating and terminating events. Replacing the events of b.c. and death by the initiating and terminating events, the sample $\{y_i\}$ is length-biased when study individuals are recruited from those who have experienced the initiating event but have not experienced the terminating event [16]. Samples of this type could also be collected in a **screening** program for chronic diseases. It was indicated by Zelen & Feinleib [20] that the screen does not detect people at random, but detects people with longer preclinical **sojourn times**.

Although statistical methods can be formulated on the basis of length-biased observations as discussed earlier, the analysis could be further complicated by the presence of right **censoring**. We next make connection between length-biased sampling and left **truncation** in this context.

Length Bias and Left Truncation

Using formula (5), the density function of y_i given t_i can be derived as $f(y)I(y > t)/S(t)$, a truncated density function. The observed t_i in left truncation models [8, 10, 11, 15, 16, 19] is usually termed the *truncation time* and has density function $S(t)I(t > 0)/\mu$. Given the observations $(t_1, y_1), \ldots, (t_n, y_n)$, the full density can be expressed as the product of the marginal density of the t_i,

$$\prod_{i=1}^{n} \left[\frac{S(t_i)}{\mu} \right],$$

and the conditional density of y_i given the t_i,

$$\prod_{i=1}^{n} \left[\frac{f(y_i)}{S(t_i)} \right]. \qquad (6)$$

In length-biased models the truncation times in general do not serve as **ancillary statistics** for parameters of f, and thus the conditional likelihood (6) is used subject to loss of information.

Suppose now the observation of the terminating event is subject to right-censoring. Assume the following independent censoring condition: conditional on the observed t_i, the time from τ_i to the terminating event, r_i, is independent of the time from τ_i to censoring, d_i. This independent censoring condition does not, however, imply independence between the length-biased time, $y_i (= t_i + r_i)$, and the censoring time, $c_i (= t_i + d_i)$ [14, 16]. Let $w_i = \min\{y_i, c_i\}$ be the time from the initiating event to the end of observation, and $\delta_i = I(w_i = y_i)$ the censoring indicator. Conditional on t_i, under the independent censoring condition the density of (w_i, δ_i) is proportional to $f(w_i)^{\delta_i} S(w_i)^{1-\delta_i}/S(t_i)$. Given a sample of iid observations $(t_1, w_1, \delta_1), \ldots, (t_n, w_n, \delta_n)$, statistical approaches based on the conditional likelihood,

$$\prod_{i=1}^{n} \left[\frac{f(w_i)^{\delta_i} S(w_i)^{1-\delta_i}}{S(t_i)} \right],$$

are considered as methods for left-truncated and right-censored data. These approaches replace the usual risk sets $R(w) = \{w_i : w_i \geq w\}$ by $R^*(w) = \{w_i : t_i \leq w \leq w_i\}$ and result in an interesting contrast with the familiar techniques used in survival analysis [11, 15, 16, 19]. These methods can be alternately derived using **counting process** techniques with left-filtering [1, 8, 10]. The connection between renewal processes and left truncation can also be made by various approaches [9, 18]. While the methods provide "simple solutions" for analyzing censored length-biased data, these conditional approaches, similar to the left-truncation case, are used subject to loss of **efficiency** because marginal information from the truncation times is not used in the construction of the methods. Furthermore, the applicability of these methods requires that the truncation time, t_i, be observable, and such a requirement might not be met in some applications.

Length Bias and Cross-sectional Sampling

In the example of prevalent cohorts, the initiating and terminating events are required to be nonrecurrent. Nevertheless, the problem of length bias could also be encountered in studies that adopt **cross-sectional** sampling techniques to collect failure times from univariate or bivariate recurrent event processes (*see* **Repeated Events**). In these studies the outcome variable of interest is the length between two successive events. The crucial condition assumed, for the validity of length-biased distribution, is the equilibrium condition for the recurrent events. With cross-sectional samplings, the intervals which contain the sampling times are observed and form the length-biased sample. Examples include cross-sectional samples of (i) fibre length [4], where the recurrent events are of the same type and the location of an event is specified as the left end of a fibre, and (ii) length of stay in a hospital, in which the bivariate recurrent events are admission to and discharge from a hospital.

References

[1] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Bhattacharyya, B.B., Franklin, L.A. & Richardson, G.D. (1988). A comparison of nonparametric unweighted and length-biased density estimation of fibres, *Communications in Statistics – Theory and Methods*, **17**, 3629–3644.

[3] Cox, D.R. (1962). *Renewal Theory*. Methuen, London.

[4] Cox, D.R. (1969). Some sampling problems in technology, in *New Development in Survey Sampling*, N.L. Johnson & H. Smith, Jr, eds. Wiley-Interscience, New York, pp. 506–527.

[5] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[6] Cox, D.R. & Isham, V. (1980). *Point Processes*. Chapman & Hall, London.

[7] Jones, M.C. (1991). Kernel density estimation for length biased data, *Biometrika* **78**, 511–519.

[8] Keiding, N. (1992). Independent delayed entry, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Dordrecht, pp. 309–326.

[9] Keiding, N. & Gill, R.D. (1988). Random truncation models and Markov processes. *Technical Report*. Centre for Mathematics and Computer Science, Amsterdam, The Netherlands.

[10] Keiding, N. & Gill, R.D. (1990). Random truncation models and Markov processes, *Annals of Statistics* **18**, 582–602.

[11] Lai, T.-L. & Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data, *Annals of Statistics* **19**, 417–442.

[12] Vardi, Y. (1982). Nonparametric estimation in the presence of length bias, *Annals of Statistics* **10**, 616–620.

[13] Vardi, Y. (1985). Empirical distributions in selection bias models, *Annals of Statistics* **13**, 178–203.

[14] Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation, *Biometrika* **76**, 751–761.

[15] Wang, M.-C., Jewell, N.P. & Tsai, W.-Y. (1986). Asymptotic properties of the product-limit estimate under random truncation, *Annals of Statistics* **14**, 1597–1605.

[16] Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data, *Journal of the American Statistical Association* **86**, 130–143.

[17] Wang, M.-C. (1996). Hazards regression analysis for length-biased data, *Biometrika* **83**, 343–354.

[18] Winter, B.B. & Foldes, A. (1988). A product-limit estimator for use with length-biased data, *Canadian Journal of Statistics* **16**, 337–355.

[19] Woodroofe, M. (1985). Asymptotic properties of the product-limit estimate under random truncation, *Annals of Statistics* **13**, 163–177.

[20] Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic diseases, *Biometrika* **56**, 601–614.

(*See also* **Weighted Distributions**)

MEI-CHENG WANG

Leukemia Clusters

Interest in the possibility that cases of cancer and leukemia tend to occur in clusters has a long history [6]. Early reports were of cancer in particular families or houses; more recently interest has centred on spatial and space–time **clustering**. Most recent interest concerns leukemia in children and has been stimulated by suspicions of an environmental etiology.

The possible explanation of clusters most considered is that environmental **radiation** might be responsible for leukemia in children, and public concern is so great that it has had a major impact on the development of civil nuclear power programs. Natural though such apprehensions may be, they are out of all proportion to the strength of the epidemiologic evidence. Radiation is certainly a known leukemogen, but most environmental doses are too low to account for significant risk. Recently, other leukemogenic mechanisms have been receiving more attention, notably the possibility of an infectious etiology. Although there is some evidence of geographical clustering, this is not strong and, despite intensive investigation, no actual leukemia cluster has led to the identification of a specific cause.

The Nature of Leukemia

Acute leukemia is a malignant disease characterized by rapid proliferation of leucocytes from a single malignant clonal cell; chronic forms develop slowly over a long period of time but are capable of becoming acute. The tumor is relatively rare in adults, but commonest in children, accounting for around a third of all cases of malignant disease under the age of 15. The most distinctive of the numerous forms is acute lymphocytic leukemia (ALL). In practice, many epidemiologic investigations distinguish only between ALL and acute nonlymphocytic leukemia (ANLL).

Types of cancer generally, and leukemia in particular, show very different relative frequencies in children and adults. Thus ALL accounts for around 80% of all leukemia in children, with a peak at around 3 years of age [15]. Most of the remaining cases are of ANLL, chronic leukemia being rare in childhood. Among adults, however, the commonest form is chronic lymphocytic leukemia, ALL being less common than either the chronic or acute myelocytic forms [25].

Reported **incidence rates** of leukemia show some variation internationally; these may be partly due to genetic factors, but are probably also a consequence of differences in reporting and diagnostic procedures [25]. In addition, the disease tends to be masked by acute infections, which may explain some of the international differences and the more marked historical trends [24]. Intranational rates show less variation for children [14] than for adults [7].

The Etiology of Leukemia

The etiology of leukemia is only partly understood. The importance of genetic factors is clear from its association with certain conditions having a clear genetic etiology, notably Down's syndrome (trisomy 21), in which the risk of childhood leukemia is increased about 15-fold [35].

Of the various exogenous factors that have been proposed as having etiological significance for leukemia, ionizing radiation is by far the most important. That it causes leukemia in high doses (of the order of 100 mSv or more) has been established beyond doubt by various epidemiologic studies [29, 34], which, however, predominantly involved adults. There is also significant evidence that the much lower doses exposing the fetus in obstetric X-ray investigations (typically 2–20 mSv) are leukemogenic [4, 13]. The epidemiologic evidence therefore justifies the interest in environmental radiation, especially the possibility of a risk near nuclear installations [33] or following nuclear accidents. In most cases, however, the excess radiation levels in such environments are small compared with the natural background radiation and are unlikely to explain observed excesses of leukemia [11].

A viral component of the etiology has also been suspected for a long time. The first strong epidemiologic evidence relates to a cohort of children born in March 1958 in the UK [17], for which there was a ninefold increase in the risk of leukemia and lymphoma among children whose mothers contracted influenza during pregnancy. Chance may have at least partly exaggerated the finding, and it is noteworthy also that the relevant exposure was to a particularly virulent epidemic of the Asian strain; in any event, subsequent corroboration was only partial [30]. More

recently, Greaves [21] has postulated that children exposed to below average levels of infection postnatally may fail to develop a fully effective immune system, making them more vulnerable to tumor initiation. Such a mechanism might explain the known association between childhood leukemia and high socioeconomic status [14] and also the population-mixing phenomena demonstrated by Kinlen [23] and discussed below.

Other known or possible causes of leukemia include chemotherapeutic agents [12]; exposure to chemical carcinogens, which is normally occupational and may be parental [28]; and exposure to electromagnetic fields [10]. With the exception of the last of these, for which the evidence is least convincing, a genuine causal relationship would be unlikely to result in demonstrable **geographic patterns**.

The Nature of Clustering

Clustering may be defined as the tendency of observations to be situated closer to one another than would be expected. The reference space within which a cluster appears may be discrete or continuous, the former being exemplified by clustering within families. The problem of analyzing excesses within family or other groups raises few special problems, however. Greater interest, both theoretically and practically, attaches to clustering in a continuum, which is normally taken to be geographical space, time, or their product, the latter giving rise to space–time clustering.

Clustering in time only would presumably be indicative of a widely dispersed short-term hazard; few instances of such hazards have been proposed for leukemia. Seasonal variation (*see* **Seasonal Time Series**) could also induce this form of clustering, though the term would not normally be taken to include such an effect and specific, period-related methods of analysis would be more appropriate than general tests. Rather little evidence of seasonality in leukemia incidence has been advanced [16].

Clustering in space could be ascribed to a number of possible mechanisms, mostly involving geographical variation of risk. Local variation of genetic factors could in principle produce spatial clustering, but there is little or no evidence of this in the case of leukemia. Spatial variation is more likely to be due to local variation in risk due to some environmental factor.

Space–time clustering – i.e. an **interaction** between the space and time distributions – could be indicative of some infective mechanism in the etiology of the disease. The evidence for this is briefly summarized below. Space–time interaction tests have the apparent attraction that they can be executed without knowledge of the marginal distributions in time and space, the latter being particularly hard to estimate accurately. However, they are vulnerable to space–time interactions in the denominators, i.e. to changes in population distribution over the study period (*see* **Denominator Difficulties**). Little work has been done on how sensitive they are to such changes and how much this may affect published findings.

Assessment of clusters is inevitably bound up with the methods used to detect them (*see* **Geographic Epidemiology**). Here we emphasize only the importance of distinguishing between situations where there is or is not a hypothesis identified a priori; the distinction crucially affects the choice of method as well as the interpretation of the results.

Evidence of Leukemia Clustering

Most of the study of leukemia clustering has concentrated on childhood leukemia, and this is reflected in the following brief review.

Nuclear Installations

The specific environmental issue that has received greatest attention is that of possible risk in proximity to nuclear installations. Early concern following the accident at Three Mile Island in the US gained new impetus in the UK, particularly after a television program in 1983 identified an abnormally large number of cases in Seascale, a village near the nuclear reprocessing plant at Sellafield in Cumbria.

Although Sellafield is one of the largest nuclear reprocessing plants in the world and has released significant quantities of radiation into the environment, detailed radiologic analyses considering available estimates of risk coefficients and parallel exposures from nuclear fallout make it very unlikely that radiation alone could account for the observed **relative risks** [11, 33]. The hypothesis that paternal preconception irradiation might be the crucial pathway [19] is inconsistent with current dosimetric estimates of genetic risk and with other epidemiologic studies [26].

Nevertheless, public concern remains high and a number of clusters in the vicinity of other nuclear installations have been reported more or less anecdotally; significantly, perhaps, these include excesses at other reprocessing plants at Dounreay in Scotland [8] and La Hague in France [32]. Public concern has extended to nuclear power generating stations although they normally have very much lower emissions. This concern has prompted a number of studies of nuclear installations in the UK [5], the US [22], and Canada [27]. These studies have not generally uncovered further significant excesses.

The excess in Seascale is particularly difficult to interpret in view of the mode of its discovery and initial reporting. Its statistical significance is in some measure diluted by the observation that Seascale is one of almost 10 000 similar areal units in the UK. The fact remains, however, that the Sellafield plant is unique in terms of its history and activity and any prior hypothesis would presumably have put high odds on this being the most likely location of any excess. It is disturbing too that, since the initial finding, further cases have occurred: between 1984 and 1992 there were a further three cases of ALL and non-Hodgkin lymphoma, bringing the total since 1963 to eight, compared with around 0.65 expected [9]. A useful collection of abstracts and papers on childhood cancers near nuclear installations was published in 1993 and dedicated to the late **Martin J. Gardner** [3].

Viral Etiologies

The difficulty of explaining the Seascale cluster in terms of radiation has prompted a search for other possibilities. Foremost of these is the possibility that the risk of childhood leukemia is increased by exposure to an infective agent consequent on increased levels of "population mixing". In a remarkable series of papers, Kinlen and colleagues studied other populations in which a similar mixing effect could be expected [23]. These include new towns, the vicinities of major construction sites, and communities receiving wartime evacuees; they showed consistently raised risks of childhood leukemia. An explanation might be that children in the indigenous population are vulnerable to an infectious agent or agents not previously encountered or, in line with Greaves' argument [21], that they have a generally higher susceptibility to leukemogenesis resulting from a reduced exposure to infections in the postnatal period. The geographic data do not permit the identification of a specific organism and are consequently unlikely to throw more light on these possible mechanisms.

Generalized Clustering of Childhood Leukemia

The evidence discussed above is noteworthy, and perhaps more convincing, because it stands out from the relative uniformity of childhood leukemia incidence. Attempts to demonstrate widespread and generalized clustering have not, generally speaking, produced striking results. An atlas of leukemia incidence covering around a third of the population of England and Wales in the years 1984–1988 [7] demonstrates moderate variation between counties; this is probably largely due to the contribution for adults, which was not separated out. Tests of spatial clustering at a more local level, however, were broadly negative.

As far as childhood leukemia is concerned, the largest register of data in the world is the (UK) National Registry of Childhood Tumours maintained by the Childhood Cancer Research Group in Oxford (*see* **Disease Registers**). Geographically referenced cases occurring in Britain in the years 1966–1983 were made available to a group of researchers, who tried out their different methodologies; the results were reported in a monograph [14]. The evidence of generalized spatial clustering was rather slight, was related to ALL under the age of 5 and appeared to be strongest in rural areas [1]; the latter association may be a reflection of the socioeconomic effect already noted. An analysis of space–time clustering by Gilman et al. [20] reported some statistically significant results, but these were hard to interpret because of the problems of multiple testing (*see* **Multiple Comparisons**) and of sensitivity to population changes referred to above; the latter problem is especially severe in the analysis of large data sets, where statistically significant results may correspond to very small real effects. The overview by Gardner [18] concluded: "Overall, there are apparently no dramatic findings in the results of the analyses carried out for this volume."

This negative view of the importance of generalized clustering is consistent with many other papers and reviews [25, 31], though the review by Alexander [2] concludes that the data as a whole are "consistent with their interpretation as an imperfect reflection of

some underlying population infective process". It is only to be expected that some significant findings will be reported. They should certainly not be discounted, but need critical appraisal, particularly where there may be doubts about the methodology.

Discussion

We conclude from this review that the evidence for any significant general tendency of leukemia to exhibit clustering is equivocal at best. It would be as unsatisfactory to conclude that there is no such effect as to conclude that the evidence is strong enough to provide real pointers toward the etiology. Although there is some evidence of geographic and of space–time clustering, the effects are at most weak and the scope for methodologic and data error is considerable. Methodologic limitations work both ways: it is possible that some stronger effects are being masked by inefficient methods and inadequate data. In particular, it is likely that place of birth is at least as important as place of residence in the etiology of childhood leukemia; unfortunately, it is in practice harder to obtain extensive data on place of birth and most published results relate to residence at diagnosis or death.

There is no particular reason why geography should hold the clue to the etiology of leukemia. Even if environmentally varying factors were known to be very important, the mobility of the population will inevitably dilute the impact on individuals. In practice, so little is known about the etiology of the disease that we cannot assume prima facie that environment should be significant. None of this is likely to allay the anxieties of people who believe that their own form of the disease is directly related to their own circumstances. If only to put their anxieties into perspective and offer the best possible reassurance, it is necessary to maintain research effort on the possibility that there is a much stronger environmental component in the etiology of the disease than appears likely at present.

References

[1] Alexander, F.E. (1991). Investigations of localized spatial clustering, and extra-Poisson variation, in *The Geographical Epidemiology of Childhood Leukaemia and non-Hodgkin Lymphomas in Great Britain, 1966–83. Studies on Medical and Population Subjects, No. 53*, G. Draper, ed. HMSO, London, pp. 69–76.

[2] Alexander, F.E. (1993). Viruses, clusters and clustering of childhood leukaemia: a new perspective?, *European Journal of Cancer, Series A* **29**, 1424–1443.

[3] Beral, V., Roman, E. & Bobrow, M., eds (1993). *Childhood Cancer and Nuclear Installations*. British Medical Journal Publishing Group, London.

[4] Bithell, J.F. & Stewart, A.M. (1975). Pre-natal irradiation and childhood malignancy: a review of British data from the Oxford survey, *British Journal of Cancer* **31**, 271–287.

[5] Bithell, J.F., Dutton, S.J., Draper, G.J. & Neary, N.M. (1994). The distribution of childhood leukaemias and non-Hodgkin lymphomas near nuclear installations in England and Wales, *British Medical Journal* **309**, 501–505.

[6] Boyle, P., Walker, A.M. & Alexander, F.E. (1996). Historical aspects of leukaemia clusters, in *Methods for Investigating Localized Clustering of Disease*, F.E. Alexander & P. Boyle, eds. IARC, Lyon, pp. 1–20.

[7] Cartwright, R.A., Alexander, F.E., McKinney, P.A. & Ricketts, T.J. (1990). *Leukaemia and Lymphoma: An Atlas of Distribution within Areas of England and Wales 1984–1988*. Leukaemia Research Fund, London.

[8] Committee on Medical Aspects of Radiation in the Environment (COMARE) (1988). *Second Report. Investigation of the Possible Increased Incidence of Leukaemia in Young People near the Dounreay Nuclear Establishment, Caithness, Scotland*. HMSO, London.

[9] Committee on Medical Aspects of Radiation in the Environment (COMARE) (1996). *Fourth Report. The Incidence of Cancer and Leukaemia in Young People in the Vicinity of the Sellafield Site, West Cumbria*. HMSO, London.

[10] Committee on the Possible Effects of Electromagnetic Fields on Biologic Systems (1966). *Possible Health Effects of Exposure to Residential Electric and Magnetic Fields*. National Academy Press, Washington.

[11] Darby, S.C. & Doll, R. (1987). Fallout, radiation doses near Dounreay, and childhood leukaemia, *British Medical Journal* **294**, 603–607.

[12] Doll, R. (1989). The epidemiology of childhood leukaemia, *Journal of the Royal Statistical Society, Series A* **152**, 341–351.

[13] Doll, R. & Wakeford, R. (1997). Risk of childhood cancer from fetal irradiation, *British Journal of Radiology* **70**, 130–139.

[14] Draper, G., ed. (1991). *The Geographical Epidemiology of Childhood Leukaemia and non-Hodgkin Lymphomas in Great Britain, 1966–83. Studies on Medical and Population Subjects, No. 53*. HMSO, London.

[15] Draper, G.J., Birch, J.M., Bithell, J.F., Kinnier Wilson, L.M., Leck, I., Marsden, H.B., Morris Jones, P.H., Stiller, C.A. & Swindell, R. (1982). *Childhood Cancer in Britain: Incidence, Survival and Mortality. Studies on Medical and Population Subjects, No. 37*. HMSO, London.

[16] Ederer, F., Myers, M.H. & Mantel, N.A. (1964). A statistical problem in space and time: do leukemia cases come in clusters?, *Biometrics* **20**, 626–638.

[17] Fedrick, J. & Alberman, E.D. (1972). Reported influenza in pregnancy and subsequent cancer in the child, *British Medical Journal* **ii**, 485–488.

[18] Gardner, M.J. (1991). Overview of the geographical approach to investigating childhood leukaemia, in *The Geographical Epidemiology of Childhood Leukaemia and non-Hodgkin Lymphomas in Great Britain, 1966–83. Studies on Medical and Population Subjects, No. 53*, G. Draper, ed. HMSO, London, pp. 127–131.

[19] Gardner, M.J., Snee, M.P., Hall, A.J., Powell, C.A., Downes, S. & Terrell, J.D. (1990). Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria, *British Medical Journal* **300**, 423–429.

[20] Gilman, E.A. & Knox, E.G. (1991). Temporal-spatial distribution of childhood leukaemias and non-Hodgkin lymphomas in Great Britain, in *The Geographical Epidemiology of Childhood Leukaemia and non-Hodgkin Lymphomas in Great Britain, 1966–83. Studies on Medical and Population Subjects, No. 53*, G. Draper, ed. HMSO, London, pp. 77–99.

[21] Greaves, M.F. (1997). Aetiology of acute leukaemia, *Lancet* **349**, 344–349.

[22] Jablon, S., Hrubec, Z. & Boice, J.D. (1991). Cancer in populations living near nuclear facilities – a survey of mortality nationwide and incidence in 2 states, *Journal of the American Medical Association* **265**, 1403–1408.

[23] Kinlen, L.J. (1995). Epidemiological evidence for an infective basis in childhood leukaemia, *British Journal of Cancer* **71**, 1–5.

[24] Kneale, G.W. (1971). Excess sensitivity of pre-leukaemics to pneumonia, *British Journal of Preventive and Social Medicine* **25**, 152–159.

[25] Linet, M.S. (1985). *The Leukemias: Epidemiologic Aspects*. Oxford University Press, New York.

[26] Little, M.P., Charles, M.W. & Wakeford, R. (1995). A review of the risks of leukemia in relation to parental pre-conception exposure to radiation, *Health Physics* **68**, 299–310.

[27] McLaughlin, J.R., King, W.D., Anderson, T.W., Clarke, E.A. & Ashmore, J.P. (1993). Paternal radiation exposure and leukaemia in offspring: the Ontario case-control study, *British Medical Journal* **307**, 959–966.

[28] O'Leary, L.M., Hicks, A.M., Peters, J.M. & London, S. (1991). Parental occupational exposures and risk of childhood cancer: a review, *American Journal of Industrial Medicine* **20**, 17–35.

[29] Preston, D.L., Kusumi, S., Tomonaga, M., Izumi, S., Ron, E., Kuramoto, A., Kamada, N., Dohy, H., Matsui, T., Nonaka, H., Thompson, D.E., Soda, M. & Mabuchi, K. (1994). Cancer incidence in atomic bomb survivors. Part III: Leukemia, lymphoma and multiple myeloma, 1950–1987, *Radiation Research* **137**, S68–S97.

[30] Shore, R.E., Pasternack, B.S. & McCrea Curnen, M.G. (1976). Relating influenza epidemics to childhood leukemia in tumor registries without a defined population base: a critique with suggestions for improved methods, *American Journal of Epidemiology* **103**, 527–535.

[31] Smith, P.G. (1982). Spatial and temporal clustering, in *Cancer Epidemiology and Prevention*, D. Schottenfeld & J.F. Fraumeni, eds. W.B. Saunders, Philadelphia.

[32] Viel, J.-F., Pobel, D. & Carré, D. (1995). Incidence of leukemia in young people around the La Hague nuclear waste reprocessing plant: a sensitivity analysis, *Statistics in Medicine* **14**, 2459–2472.

[33] Wakeford, R. & Binks, K. (1989). Childhood leukaemia and nuclear installations, *Journal of the Royal Statistical Society, Series A* **152**, 61–86.

[34] Weiss, H.A., Darby, S.C., Fearn, T. & Doll, R. (1995). Leukemia mortality after X-ray treatment for ankylosing spondylitis, *Radiation Research* **142**, 1–11.

[35] Zipursky, A., Poon, A. & Doyle, J. (1992). Leukemia in Down syndrome: a review, *Pediatric Hematology and Oncology* **9**, 139–149.

JOHN F. BITHELL

Level of a Test

The level of a statistical test, often called the level of significance, is the probability of rejecting the **null hypothesis** of "no effect" when in fact it is true. In classical **hypothesis testing** a null hypothesis, denoted by H_0, is assumed to be true, and the observed data are evaluated by a statistical test procedure to decide if the data provide sufficient evidence to reject this hypothesis. The possible outcomes of this test procedure are summarized in Table 1.

The level of the test is the probability of making a type I error. The decision to accept or reject H_0 is based on a comparison of the prespecified level of the test (generally 0.05 or 0.01) with the test procedure's ***P* value**, i.e. the calculated probability of finding a difference at least as great as the one actually observed, assuming that H_0 is true. If the P value is less than the level of the test, then the experimental data are considered to be inconsistent with the null hypothesis, H_0 is rejected, and the result is declared to be "statistically significant" at that particular level. If the P value is greater than the level of the test, then H_0 is accepted. A statistically significant departure from the null hypothesis may or may not be of practical importance in a given study.

Table 1

		State of Nature	
		H_0 true	H_0 false
Decision	Accept H_0	Correct	Type II error
	Reject H_0	Type I error	Correct

Although the level of a test is traditionally taken as 0.05 or 0.01, this choice may vary, depending upon the probability of type I and type II errors that the experimenter is willing to accept. The level of a test is also one of several factors (together with sample size, the magnitude of the difference to be detected and the underlying variability) that determine the **power** of a test procedure for detecting departures from the null hypothesis (power is defined as one minus the probability of a type II error). Thus, for example, if 0.01 rather than 0.05 is selected as the level of the test, the probability of a type I error is reduced, but so is the power for detecting departures from H_0.

Other statistical terms that are often used to refer to the level of a test include the size of the test, the alpha error, and the **false positive rate**.

JOSEPH K. HASEMAN

Levinson–Durbin Algorithm

The Levinson–Durbin **algorithm** is a method for finding the solution $\boldsymbol{b} = [b_1, b_2, \ldots b_p]^{\mathrm{T}}$ to a system of p linear equations

$$\mathbf{A} \times \boldsymbol{b} = \mathbf{c}, \tag{1}$$

where $\mathbf{A}$ is a $p \times p$ symmetric Toeplitz **matrix** (a Toeplitz matrix is composed of elements that are constant along the diagonals)

$$\mathbf{A} = \begin{bmatrix} a_0 & a_1 & \ldots & a_{p-2} & a_{p-1} \\ a_1 & a_0 & \ldots & a_{p-3} & a_{p-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{p-2} & a_{p-3} & \ldots & a_0 & a_1 \\ a_{p-1} & a_{p-2} & \ldots & a_1 & a_0 \end{bmatrix} \tag{2}$$

and $\mathbf{c}$ is a known p-dimensional column vector.

Systems like (1) with $\mathbf{A}$ both symmetric and Toeplitz can be found in several applications, like **spectral** estimation, filter design (*see* **Kalman Filter**), or linear **prediction**. For instance, given two N-point sequences $\{x_k\}$ and $\{y_k\}$, the linear prediction problem deals with finding the coefficients of a moving-average filter of order p that predicts $\{y_k\}$ from $\{x_k\}$, minimizing the sum of the squared-errors S:

$$S = \sum_{k=0}^{N-1} [y_k - (b_1 x_k + \cdots + b_p x_{k-p})]^2. \tag{3}$$

The coefficients $\boldsymbol{b} = [b_1, b_2, \ldots b_p]^{\mathrm{T}}$ are the solution to (1), when $a_i = r_{xx}(i)$ are the **autocorrelation** coefficients of $\{x_k\}$, and $\boldsymbol{c} = [r_{xy}(0), r_{xy}(1), \ldots, r_{xy}(p-1)]^{\mathrm{T}}$ are the cross-correlation coefficients (Wiener–Hopf equations).

Another application is finding the coefficients b_k of an autoregressive model (*see* **ARMA and ARIMA Models**) of order p approximating a time series $\{y_k\}$

$$y_N = -\sum_{k=1}^{p} b_k y_{N-k} + e_N \tag{4}$$

when the autocorrelation coefficients $[r_{yy}(i)]$ are known, or estimated from $\{y_k\}$. The solution $\boldsymbol{b} = [b_1, b_2, \ldots, b_p]^{\mathrm{T}}$ satisfies (1), where $a_i = r_{yy}(i)$, and $\boldsymbol{c} = [-r_{yy}(1), -r_{yy}(2), \ldots, -r_{yy}(p)]^{\mathrm{T}}$ (**Yule–Walker equations**).

To solve the matrix equation (1), one should calculate $\boldsymbol{b} = \boldsymbol{A}^{-1} \times \mathbf{c}$. Common general methods require a number of operations proportional to p^3 to compute $\boldsymbol{A}^{-1}$. By exploiting the special structure of the Toeplitz matrix in connection with the linear prediction problem, N. Levinson developed a computationally efficient algorithm [2] that requires a number of operations proportional only to p^2. The method was rediscovered by J. Durbin, who applied it to fit an autoregressive model to a given correlation sequence [1]; therefore, the algorithm is commonly referred to as the "Levinson–Durbin" algorithm.

The method is an iterative procedure that solves a series of truncated problems. It is initialized by solving the equation $a_0 b_1^{(1)} = c_1$. Then, at each step m the size of the problem is incremented by considering a new row and column of $\boldsymbol{A}$, that is, the set of m-linear equations:

$$\sum_{i=1}^{m} a_{i-1} b_i^{(m)} = c_j (j = 1, \ldots, m). \tag{5}$$

When $m = p$, the vector $\boldsymbol{b} = [b_1^{(p)}, b_2^{(p)}, \ldots b_p^{(p)}]^{\mathrm{T}}$ is the solution to (1). Figure 1 shows a scheme of the algorithm when it is used to solve the Yule–Walker equations; further details can be found in [4]. A generalization to the case of the nonsymmetric Toeplitz matrix is given in [6].

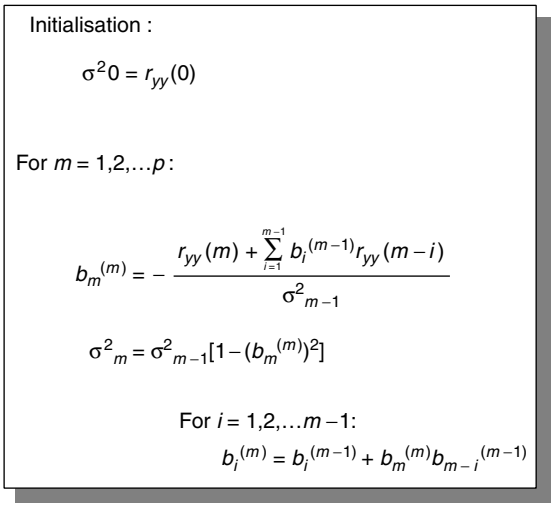

Figure 1 Scheme of the Levinson–Durbin algorithm for solving the Yule–Walker equations

The algorithm has been widely applied in several biomedical fields like speech analysis [3] and spectral analysis of EEG [5, 7] and cardiovascular signals [8] (*see* **Clinical Signals**). The popularity of the Levinson–Durbin algorithm in biomedicine is based on its effectiveness in solving the Yule–Walker equations. In fact, many nonstationary biomedical signals–like the EEG or the beat-by-beat series of cardiovascular data–can be considered "locally stationary" over short time windows. Sometimes these segments of local stationarity are too short to be analyzed by traditional **Fast Fourier transform** (FFT) spectra, which might not provide the required frequency resolution. Alternatively, the best autoregressive model of order p fitting the data is identified by solving the Yule–Walker equations (owing to its recursive structure, the Levinson–Durbin algorithm also provides all the models of order $m < p$), and a high-resolution spectrum is calculated from the theoretical expression for autoregressive processes [4].

In the following example, the algorithm is applied to model a respiratory signal $\{y_N\}$. Respiratory waves can be derived from the ECG by assessing beat-by-beat changes in the cardiac electrical axis. From a short ECG recording consisting of 15 consecutive heartbeats, the following 14 samples of a respiratory time series $\{y_N\}$, in mV, were derived:

$$\{88; -141; -154; 129; 7; -135; -26;$$
$$161; -72; -13; 262; -48; -134; 75\}.$$

The sampling frequency of $\{y_N\}$ is the mean heart rate that, in this case, was 70 bpm. The coefficients $b_i^{(3)}$ of the best autoregressive model of order $m = 3$ fitting the data

$$y_N = -b_1^{(3)} y_{N-1} - b_2^{(3)} y_{N-2} - b_3^{(3)} y_{N-3} + e_N \quad (6)$$

are found solving the Yule–Walker equations. First, the autocorrelation coefficients are estimated from $\{y_N\}$ obtaining $r_{yy}(0) = 15\,195$, $r_{yy}(1) = -3196$, $r_{yy}(2) = -10\,776$, and $r_{yy}(3) = 9640$. Then, the algorithm is applied as shown in Figure 1. From

$$\sigma_0^2 = r_{yy}(0) = 15\,195 \quad (7)$$

we iteratively obtain

$$m = 1 : b_1^{(1)} = \frac{-r_{yy}(1)}{\sigma_0^2} = \frac{3196}{15\,195} = 0.210$$
$$\sigma_1^2 = \sigma_0^2[1 - b_1^{(1)2}] = 15\,195$$
$$\times (1 - 0.210^2) = 14\,525, \quad (8)$$

$$m = 2 : b_2^{(2)} = \frac{-(r_{yy}(2) + b_1^{(1)} r_{yy}(1))}{\sigma_1^2}$$
$$= \frac{10\,776 + 0.210 \times 3196}{14\,525} = 0.788$$
$$b_1^{(2)} = b_1^{(1)} + b_2^{(2)} b_1^{(1)} = 0.210 + 0.788$$
$$\times 0.210 = 0.375$$
$$\sigma_2^2 = \sigma_1^2[1 - b_2^{(2)2}] = 14\,525$$
$$\times (1 - 0.788^2) = 5506, \quad (9)$$

and finally, for $m = 3$:

$$b_3^{(3)} = \frac{-(r_{yy}(3) + b_1^{(2)} r_{yy}(2) + b_2^{(2)} r_{yy}(1))}{\sigma_2^2}$$
$$= \frac{-(9640 - 0.375 \times 10\,776 - 0.788 \times 3196)}{5506}$$
$$= -0.559$$
$$b_1^{(3)} = b_1^{(2)} + b_3^{(3)} b_2^{(2)} = 0.375 - 0.559 \times 0.788$$
$$= -0.065$$

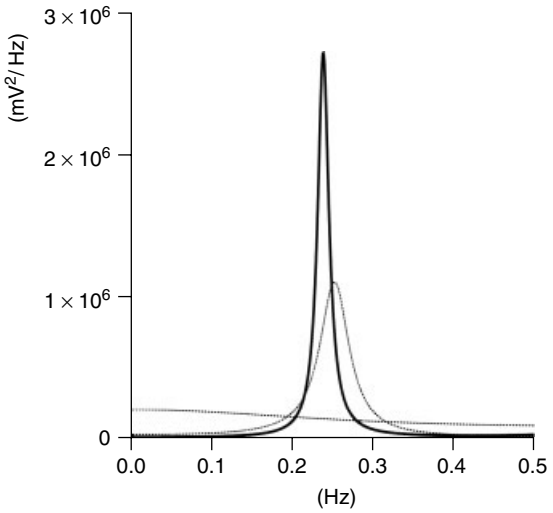

Figure 2 Power spectrum of the autoregressive model of order 3 fitting the respiratory signal of the example (see text); the model coefficients were identified by solving the Yule–Walker equations through the Levinson–Durbin algorithm. Spectra corresponding to models of order 2 and 1 are also shown (dotted lines)

$$b_2^{(3)} = b_2^{(2)} + b_3^{(3)} b_1^{(2)} = 0.788 - 0.559 \times 0.375 = 0.578$$
$$\sigma_3^2 = \sigma_2^2[1 - b_3^{(3)2}] = 5506 \times (1 - 0.559^2) = 3785. \quad (10)$$

In this way, we identify not only the coefficients $b_i^{(3)}$ but also the power of the model error, σ_3^2. Figure 2 shows the power spectrum of $\{y_N\}$ calculated from this model.

References

[1] Durbin, J. (1960). The fitting of time-series models, *Review of the International Statistical Institute* **28**, 233–243.

[2] Levinson, N. (1949). Appendix B of Wiener, N, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Wiley, New York.

[3] Markel, J.D. & Gray, A.H. (1976). *Linear Prediction of Speech.* Spinger-Verlag, New York.

[4] Marple, S.L. (1987). *Digital Spectral Analysis with Applications*. Prentice Hall, Englewood Cliffs, NJ.

[5] Pardey, J., Roberts, S. & Tarassenko L. (1996). A review of parametric modelling techniques for EEG analysis, *Medical Engineering and Physics* **18**, 2–11.

[6] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, New York.

[7] Zetterberg, L.H. (1978). Recent advances in EEG data processing, *Electroencephalography and Clinical Neurophysiology. Supplement* **34**, 19–36.

[8] Baselli, G., Cerutti, S. (1985). Identification techniques applied to processing of signals from cardiovascular systems, *Medical Informatics* **10**, 223–235.

PAOLO CASTIGLIONI

Lexis Diagram

A Lexis diagram is a (time, age) coordinate system, representing individual lives by line segments of unit slope, joining (time, age) of birth and death [14] (see Table 1 and Figure 1). The Lexis diagram is an important descriptive tool in epidemiology and **demography**. However, it also has several applications in **survival analysis** and analytical epidemiology as a tool for several classes of statistical models, as surveyed by Keiding [8]. These uses of the Lexis diagram are less common and it is the aim of this article to indicate some recent developments.

Lexis [14] in his Figure 1, reproduced here as Figure 2, originally considered a diagram of (calendar time at birth, age) in which life lines will be vertical rather than having unit slope. In his Figure 2, reproduced here as Figure 3, he also mentioned an equilateral diagram in which the time units in the calendar time, age, and cohort (i.e. time of birth) directions are of the same length. See [11] for more on the early history of the Lexis diagram. Lexis further discussed a three-dimensional extension allowing for an intermediate (irreversible) life event, in Lexis's case exemplified by marriage. This corresponds to the three-state model basic to the modern statistical description of **incidence** and **prevalence** (cf. [9, 15]).

Despite its long history, the Lexis diagram is still being rediscovered among statisticians, cf. Goldman [6] for the standard Lexis diagram and Weinkam & Sterling [20] for the equilateral Lexis diagram.

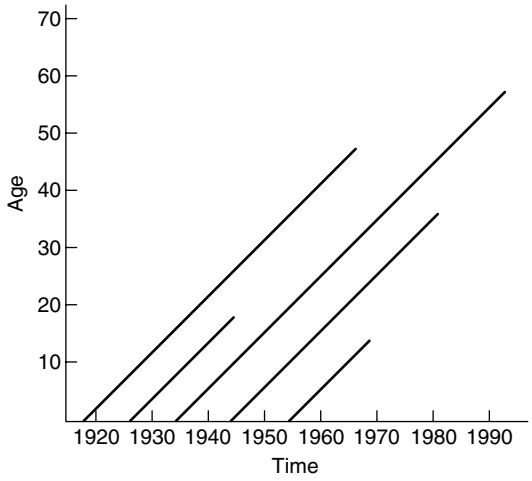

Figure 1 A Lexis diagram representing the five lives of Table 1

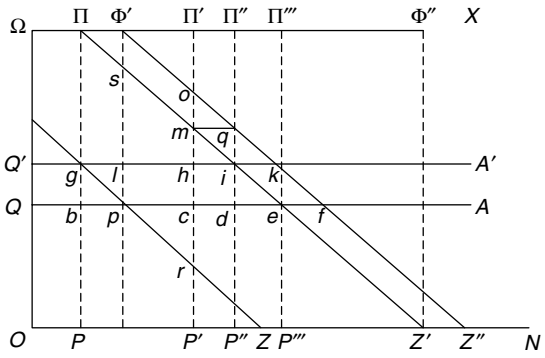

Figure 2 Lexis's diagram [14, Figure 1]

Applications of the Lexis Diagram in Survival Analysis and Analytical Epidemiology

Clinical Trials with Staggered Entry

In many **clinical trials** patients arrive sequentially in calendar time but the substantive interest is on survival time since entry. As explained in the articles **Interim Analysis of Censored Data** and **Staggered Entry**, the resulting interplay between the two time scales (calendar time and duration) has generated considerable complications in the development of a satisfactory statistical theory, particularly if comparisons between treatments are intended along the way at certain fixed time points (interim analysis) or sequentially (*see* **Data and Safety Monitoring**).

As mentioned by Keiding et al. [12] (*see* **Delayed Entry**), it is sometimes feasible to exploit the remaining life times of individuals (counted with delayed entry) from an interim analysis to supplement new individuals in a confirmatory analysis. This idea is explained in the Lexis diagram of Figure 4.

Table 1 Five lives illustrated in Figure 1

Born	Died	Age at death
1918	1966	48
1926	1944	18
1934	1992	58
1944	1978	34
1954	1968	14

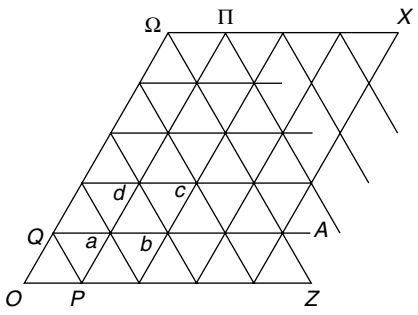

Figure 3 Lexis's equilateral diagram [14, Figure 2]

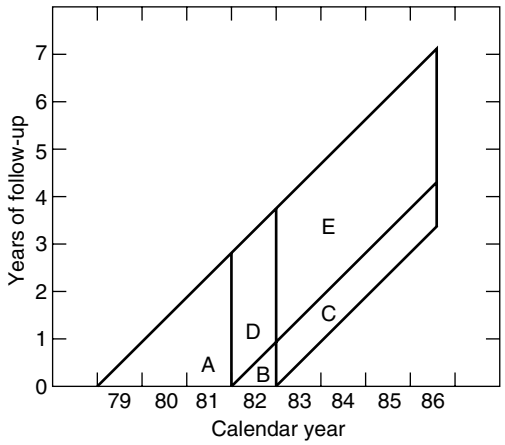

Figure 4 Lexis diagram of the DBCG-77 clinical trials on adjuvant treatment of breast cancer. The traditional independent data set for verifying an unexpected finding in A would be based on B and C. However, much more information is obtained by including also D and E, and in fact B and D already would have yielded the independent confirmation not achieved by B and C. Reproduced from Keiding et al. [12] by permission of John Wiley & Sons Ltd

Disease Incidence Studies

Lexis diagram representations of classical (often historically) prospective studies (*see* **Cohort Study**; **Cohort Study, Historical**) of (calendar time, age)-specific disease incidence are common, and we return to some of the statistical issues below. More intricate sampling plans may also take advantage of this representation, such as the **retrospective** incidence study of a **cross-sectional** sample of prevalent diabetics by Keiding et al. [13], where each incident and surviving case needed to be weighted (in a **Horvitz–Thompson** fashion) by its inverse survival probability from disease onset to the sampling date.

Prevalent Cohort Studies

A prevalent cohort study is based on a cross-sectional sample of diseased patients, with or without retrospective information on disease onset. Patients are followed until death or a fixed later calendar time, whichever comes first, (see Figure 5, and also [7] for additional examples and the link to the Arjas–Haara theory of innovative and noninnovative marks in the marked **point process** that accounts for the partial observation). The articles **Biased Sampling of Cohorts in Epidemiology** and **Delayed Entry** provide surveys of design and analysis problems for such studies.

Statistical Inference in the Lexis Diagram

Piecewise Constant Intensity Models

Many disease incidence and mortality studies (perhaps particularly in cancer) have taken piecewise constant intensity models (*see* **Grouped Survival Times**) as method of choice (see Clayton & Schifflers [4, 5] for a definitive survey). As is also well known in sociology, there is an inherent

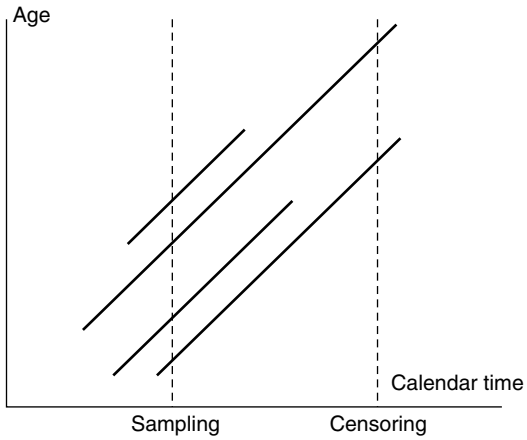

Figure 5 Lexis diagram of a prevalent cohort study. Four patients are sampled and their disease onset is known. During the follow-up period two of them die; the other two are still alive at the end of follow-up, where they are censored

unidentifiability of the linear component in a model allowing for dependence on both **age, period and cohort** (*see* **Identifiability**), although Nakamura [17] showed that a **Bayesian** framework allowed roughness penalties in the three directions to decide the matter.

Point Processes, Continuous Time

Brillinger [3] initiated an exact use of point processes as a basis for statistical models for incidence and mortality in the Lexis diagram, generalized to morbidity (incidence) and prevalence by Keiding [9, 10] (see also [19]). Without parametric assumptions, statistical analysis requires smoothing formally studied by McKeague & Utikal [16] and embedded in an **empirical Bayes** interpretation of **penalized likelihood** by Berzuini et al. [2], Berzuini & Clayton [1] and Ogata et al. [18], who reanalyzed the retrospective diabetes incidence study by Keiding et al. [13] quoted above.

References

[1] Berzuini, C. & Clayton, D. (1994). Bayesian analysis of survival on multiple time scales, *Statistics in Medicine* **13**, 823–838.

[2] Berzuini, C., Clayton, D. & Bernardinelli, L. (1993). Bayesian inference on the Lexis diagram, *Bulletin of the International Statistics Institute* **55**, 149–165; with discussion **55**, 42–43.

[3] Brillinger, D.R. (1986). The natural variability of vital rates and associated statistics (with discussion), *Biometrics* **42**, 693–734.

[4] Clayton, D. & Schifflers, E. (1987). Models for temporal variation in cancer rates. I: age-period and age-cohort models, *Statistics in Medicine* **6**, 449–467.

[5] Clayton, D. & Schifflers, E. (1987). Models for temporal variation in cancer rates. II: age–period–cohort models, *Statistics in Medicine* **6**, 469–481.

[6] Goldman, A.I. (1992). Eventcharts: Visualizing survival and other timed-events data, *American Statistician* **46**, 13–18.

[7] Keiding, N. (1989). Discussion of E. Arjas: Survival models and martingale dynamics, *Scandinavian Journal of Statistics* **16**, 209–213.

[8] Keiding, N. (1990). Statistical inference in the Lexis diagram, *Philosophical Transaction of the Royal Society of London, Series A* **332**, 487–509.

[9] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[10] Keiding, N. (1992). Independent delayed entry (with discussion), *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Dordrecht, pp. 309–326.

[11] Keiding, N. (2000). Mortality measurement in the 1870s: diagrams, stereograms, and the basic differential equation. `http://www.demogr.mpg.de/Papers/workshops/ws.000828.htm`.

[12] Keiding, N., Bayer, T. & Watt-Boolsen, S. (1987). Confirmatory analysis of survival data using left truncation of the life times of primary survivors, *Statistics in Medicine* **6**, 939–944.

[13] Keiding, N., Holst, C. & Green, A. (1989). Retrospective estimation of diabetes incidence from information in a prevalent population and historical mortality, *American Journal of Epidemiology* **130**, 588–600.

[14] Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Trübner, Strassburg.

[15] Lund, J. (2000). Sampling bias in population studies – how to use the Lexis diagram. *Scand. J. Statist.* **27**, 589–604.

[16] McKeague, I.W. & Utikal, K.J. (1990). Inference for a nonlinear counting process regression model, *Annals of Statistics* **18**, 1172–1187.

[17] Nakamura, T. (1986). Bayesian cohort models for general cohort table analyses, *Annals of the Institute of Statistical Mathematics* **38**, 353–370.

[18] Ogata, Y., Katsura, K., Keiding, N., Holst, C. & Green, A. (2000). Empirical Bayes age-period-cohort analysis of retrospective incidence data, *Scand. J. Statist.* **27**, 415–432.

[19] Wang, M.-C., Brookmeyer, R. & Jewell, N.P. (1993). Statistical models for prevalent cohort data, *Biometrics* **49**, 1–11.

[20] Weinkam, J.J. & Sterling, T.D. (1991). A graphical approach to the interpretation of age-period-cohort data, *Epidemiology* **2**, 133–137.

NIELS KEIDING

Life Expectancy

Life expectancy is both the most summary and the most significant measure derived from a **life table**. Life expectancy at age **x** is the average number of years a person aged **x** will live if subject to the mortality rates contained in the life table.

In life table notation, life expectancy at age **x**, $\mathring{e}_x$, is given by

$$\mathring{e}_x = \frac{T_x}{l_x},$$

where T_x is the total years lived in the life table population after exact age **x**, and l_x = the number of survivors in the life table population at exact age **x**. The method for calculating these quantities can be found in standard textbooks [3].

Like the life table itself, life expectancy is determined by the force of mortality or mortality hazard function, $\mu(x)$, over the entire age range (*see* **Hazard Rate**). In continuous notation

$$\mathring{e}(x) = \int_x^\infty \frac{l(x)\,\mathrm{d}x}{l(x)},$$

and since

$$l(x) = \exp\left[-\int_0^x \mu(u)\,\mathrm{d}(u)\right],$$

it can be seen that life expectancy at age **x** reflects both the **cumulative hazard** from birth to age **x** (through l_x), and the cumulative hazard from **x** to the oldest age (through T_x, itself an integral of l_x).

In **actuarial** analysis it is normal to distinguish between the complete (exact) expectation of life and the curtate or whole year expectation, but in **demography** and epidemiology the complete expectation is universally employed.

In most populations $\mathring{e}(x)$ tends to rise between birth and age 1, and to decline linearly thereafter, although in very low mortality countries the decline is virtually linear throughout the age range.

The most common life expectancy encountered is $\mathring{e}(0)$, the expectation of life at birth. Because $\mathring{e}(0)$ incorporates the entire mortality experience of the cohort or life table population, it may be considered as an age standardized (*see* **Standardization Methods**) measure of mortality, where the standard age distribution is derived from the age pattern of mortality itself.

Life expectancy at birth has increased substantially with modernization, rising from a preindustrial level of perhaps 40 years to current levels of over 80 in countries like Japan. Because of its cumulative impact throughout the age range, improved survival in infancy has made the greatest contribution to this increase.

In recent years there have been unexpected gains in expectation of life at older ages in some very low mortality countries, leading to predictions that expectation of life could rise to 100 years, with significant effects on social security and pension systems. However, 85 seems a more likely upper limit [2].

Traditionally, life table theory has not had a strong statistical component: the focus has primarily been on the average expectation of life, rather than on its distribution. However, Chiang [1] has addressed the **sampling** theory of the life table, and more recently the homogeneity/heterogeneity of life expectancy has received renewed attention, particularly in the context of expectation of life at advanced ages.

Life expectancy is also used as a powerful tool in association with multistate or multilevel life tables. Expectations of working life, of a healthy life, or of a life free of disability are examples of this. Unlike death, individuals may move in and out of these states, requiring modifications to the logic of the life table to incorporate nonabsorbing states.

References

[1] Chiang, C.L. (1984). *The Life Table and Its Applications*. Krieger, Malabar.

[2] Olshansky, S.J. & Carnes, B.A. (1994). Demographic perspectives on human senescence, *Population and Development Review* **20**, 57–80.

[3] Shryock, H.S. & Siegel, J.S. (1973). *The Methods and Materials of Demography*. US Department of Commerce, Bureau of the Census, Washington.

L. SMITH

Life Table

A life table is a tabular representation of central features of the distribution of a positive **random variable**, say T, with an absolutely continuous distribution. It may represent the lifetime of an individual, the failure time of a physical component, the remission time of an illness, or some other duration variable. In general, T is the time of occurrence of some event that ends individual survival in a given status. Let its cumulative distribution function (cdf) be $F(t) = \Pr(T \leq t)$ and let the corresponding survival function be $S(t) = 1 - F(t)$, where $F(0) = 0$. If $F(\cdot)$ has the probability density function (pdf) $f(\cdot)$, then the risk of event occurrence is measured by the **hazard** $\mu(t) = f(t)/S(t)$, for t where $S(t) > 0$. Because of its sensitivity to changes over time and to risk differentials between population subgroups, $\mu(t)$ is a centerpiece of interest in empirical investigations.

In applications to human mortality, which is where life tables originated, the time variable normally is a person's attained age and is denoted x. The function $\mu(x)$ is then called the *force of mortality* or *death intensity* (*see* **Hazard Rate**). The life-table function $l_x = 100\,000\ S(x)$ is called the *decrement function* and is tabulated for integer x in *complete life tables*; in *abridged life tables* it is tabulated for sparser values of x, most often for five-year intervals of age. The *radix* l_0 is selected to minimize the need for decimals in the l_x table; a value different from 100 000 is sometimes chosen. Other life-table functions are the expected number of deaths $d_x = l_x - l_{x+1}$ at age x (i.e. between age x and age $x + 1$), the single-year death probability $q_x = \Pr(T \leq x + 1 | T > x) = d_x/l_x$, and the corresponding survival probability $p_x = 1 - q_x$. Simple integration gives

$$q_x = 1 - \exp\left[-\int_x^{x+1} \mu(s)\,\mathrm{d}s\right]. \qquad (1)$$

Life-table construction consists in the estimation and tabulation of functions of this nature from empirical data. If ungrouped individual-level data are available, then the **Kaplan–Meier estimator** can be used to estimate l_x for all relevant x and estimators of the other life-table functions can then be computed subsequently. Alternatively, a segment of the **Nelson–Aalen estimator** can be used to estimate $\int_x^{x+1} \mu(s)\,\mathrm{d}s$; (1) can then be used to estimate q_x for each x, and the rest of the computations follow suit. From any given schedule of death probabilities $q_0, q_1, q_2, \ldots$, the l_x table is easily computed sequentially by the relation $l_{x+1} = l_x(1 - q_x)$ for $x = 0, 1, 2, \ldots$. Much of the effort in life-table construction therefore is concentrated on providing such a schedule $\{q_x\}$.

More conventional methods of life-table construction use **grouped survival times**. Suppose for simplicity that the range of the lifetime T is subdivided into intervals of unit length and that the number of failures observed during interval x is D_x. Let the corresponding total person-time recorded under risk of failure in the same interval be R_x. Then, if $\mu(t)$ is constant over interval x (the assumption of *piecewise constancy*), then the *death rate* $\hat{\mu}_x = D_x/R_x$ is the **maximum likelihood** estimator of this constant. Relation (1) can again be used to provide an estimator

$$\hat{q}_x = 1 - \exp(-\hat{\mu}_x), \qquad (2)$$

and the crucial first step in the life-table computation has been achieved. Instead of (2), $\hat{\mu}_x \big/ \left(1 + \frac{1}{2}\hat{\mu}_x\right)$ is often used to estimate q_x. This solution is of older vintage and may be regarded as an approximation to (2).

Two kinds of problems may arise: (i) the exact value of R_x may not be known, and (ii) the constancy assumption for the hazard may be violated.

When the exact risk time R_x is not known, some approximation is often used. An Anglo-Saxon tradition is to use the midyear population in the age interval. Alternatively, suppose that the number N_x of survivors to exact age x and the number W_x of withdrawals (losses to follow-up) in the age interval are known. What has become known as the **actuarial method** then consists in approximating R_x by $N_x - \frac{1}{2}(D_x + W_x)$. If there are no withdrawals and N_x is known, then D_x/N_x is the maximum likelihood estimator of q_x, and this provides a suitable starting point for the life-table computations.

For the case where only grouped data are available and the piecewise-constancy assumption for the intensity function is implausible, various methods have been developed to improve on (2). For an overview, see Keyfitz [12]. Even if single-year age groups are used, mortality drops too fast in the first year of life to merit an assumption of constancy over this interval. Demographers often use $\hat{\mu}_0/[1 + (1 - a_0)\hat{\mu}_0]$ to estimate q_0, where a_0 is some small figure, say between 0.1 and 0.15 [2]. If it is

possible to partition the first year of life into subintervals in each of which mortality *can* be taken as constant, then it is statistically more efficient essentially to build up a life table for this year. This leads to an estimate like $\hat{q}_0 = 1 - \exp(-\sum_i \hat{\mu}_i)$, where the sum is taken over the first-year intervals. See Dublin et al. [5, p. 24] for an example.

The force of mortality is sometimes represented by a function $h(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. Actuaries most often use the classical Gompertz–Makeham function $h(x;a,b,c) = a + bc^x$ for the force of mortality in their life tables (*see* **Parametric Models in Survival Analysis**). When individual-level data are available, it would be statistically most efficient to estimate the parameters by the maximum likelihood method, but most often they are estimated by fitting $h(\cdot;\boldsymbol{\theta})$ to a schedule of death rates $\{\hat{\mu}_x\}$, perhaps by **least squares**, minimum chi-square (*see* **Ban Estimates**), or some **method of moments**. This approach is called *analytic graduation*; for its statistical theory, see [11]. One of many alternatives to modeling the force of mortality is to let [10]

$$\frac{q_x}{p_x} = A^{(x+B)^C} + D\exp[-E(\ln x - \ln F)^2] + GH^x.$$

So far we have tacitly assumed that the data come from a group of independent individuals who have all been observed in parallel and whose lifetimes have the same cdf. **Staggered** (delayed) **entries** into the study population and voluntary exits (withdrawals) from it are permitted provided they contain no information about the risk in question, be it death, recurrence of a disease, or something else. Nevertheless, the basic idea is that of a connected cohort of individuals that is followed from some significant starting point (like birth or the onset of some disease) and which is diminished over time due to *decrements (attrition)* caused by the risk's operation. In demography, this corresponds to following a **birth cohort** through life or a marriage cohort while their marriages last, and the ensuing tables are called *cohort life tables*.

Because such tables can only be terminated at the end of a cohort's life, it is more common to compute age-specific attrition rates $\hat{\mu}_x$ from data collected for the members of a population during a limited period and to use the mechanics of life-table construction to produce a *period life table* for the population from such rates. If mortality patterns are tied to cohorts, then individuals who live at widely differing ages in the period of observation cannot be expected to have the same risk structure, and the period table is said to reflect the patterns of a *synthetic* (fictitious) cohort exposed to the risk of the period at the various ages.

Multiple-decrement Tables

When two or more mutually exclusive risks operate on the study population (*see* **Competing Risks**), one may correspondingly compute a *multiple-decrement table* to reflect this. For instance, a period of sickness can end in death or, alternatively, in recovery. Suppose that an integer random variable K represents the *cause of decrement* and define $F_k(t) = \Pr(T \le t, K = k)$, $f_k(t) = \mathrm{d}F_k(t)/\mathrm{d}t$, and $\mu_k(t) = f_k(t)/S(t)$, assuming that all $F_k(\cdot)$ are absolutely continuous. Then $\mu_k(\cdot)$ is the cause-specific hazard (intensity) for risk cause k and $\mu(t) = \sum_k \mu_k(t)$ is the total risk of decrement at time t. For the multiple-decrement table, we define the decrement probability

$$\begin{aligned} q_x^{(k)} &= \Pr(T \le x+1, K = k | T > x) \\ &= \int_0^1 \exp\left[-\int_0^t \mu(x+s)\,\mathrm{d}s\right] \mu_k(x+t)\,\mathrm{d}t. \end{aligned} \tag{3}$$

For given risk intensities, $q_x^{(k)}$ can be computed by numerical integration in (3). The expected number of decrements at age x as a result of cause k is $d_x^{(k)} = l_x q_x^{(k)}$. When estimates are available for the cause-specific risk intensities, one or two columns can therefore be added to the life table for each cause to include estimates of $d_x^{(k)}$ and possibly $q_x^{(k)}$.

Several further life-table functions can be defined by formal reduction or elimination of one or more of the intensity functions in formulas like those above. In this manner, a *single-decrement life table* can be computed for each cause k, depicting what the normal life table would look like *if* cause k were the only one that operated in the study population and *if* it did so with the risk function estimated from the data. The purpose is to see the effect of the risk cause in question without interference from other causes. Some demographers call this abstraction the risk's *pure* effect. No assumption is made that in practice the total attrition risk can actually be reduced to the level of the one which is in focus or that this cause operates independently of other causes. For instance, a single-decrement life table of recovery from an

illness reflects the pure timing effect of the duration structure of the intensity of recovery even though the elimination of mortality is unattainable.

A single-decrement life table is at an extreme end of a class of tables produced by deleting one (or more) of the cause intensities in formulas like those above. To obtain a *cause-deleted life table*, where only cause k has been eliminated, one may introduce $\mu_{-k}(t) = \mu(t) - \mu_k(t)$,

$$q_x^{(-k)} = \int_0^1 \exp\left[-\int_0^t \mu_{-k}(x+s)\,\mathrm{d}s\right]\mu_{-k}(x+t)\,\mathrm{d}t$$
$$= 1 - \exp\left[\int_x^{x+1} \mu_{-k}(s)\,\mathrm{d}s\right], \qquad (4)$$

and so on, and a "normal" life table may be computed with $\mu(t)$ replaced by $\mu_{-k}(t)$ everywhere. A corresponding cause-deleted multiple-decrement life table may be based on reduced cause-specific decrement probabilities like

$$\int_0^1 \exp\left[-\int_0^t \mu_{-k}(x+s)\,\mathrm{d}s\right]\mu_j(x+t)\,\mathrm{d}t,$$
$$\text{for } j \neq k.$$

Such a table would show what a normal table would look like *if* it were possible to eliminate cause k without changing the risk of any other cause. Again no assumption needs to be made about the feasibility of such elimination in real life nor about cause independence. The computations are based on a pure abstraction. The interpretation for real-life applications must be based on substantive considerations and is a different matter.

Life Expectancy

An individual's **life expectancy** (at birth) is the expected value

$$\mathring{e}_0 = \mathrm{E}(T) = \int_0^\infty [1 - F(x)]\,\mathrm{d}x = \int_0^\infty \frac{l_x}{l_0}\,\mathrm{d}x$$

of his or her lifetime T, computed for the probability distribution $F(\cdot)$ operating at the time of birth. When the individual has survived to (exact) age x, his or her remaining lifetime, $U = T - x$, is positive and has the survival function $S_x(u) = S(x+u)/S(x) = l_{x+u}/l_x$, and the *residual life expectancy* is

$$\mathring{e}_x = \mathrm{E}(T - x | T > x) = \int_0^\infty S_x(u)\,\mathrm{d}u = \int_0^\infty \frac{l_{x+u}}{l_x}\,\mathrm{d}u.$$

If $L_x = \int_0^1 l_{x+t}\,\mathrm{d}t$, we get $L_x \cong \frac{1}{2}(l_x + l_{x+1})$ by the trapezoidal rule of numerical integration, and

$$\mathring{e}_x = \sum_{t=0}^\infty L_{x+t} \cong \sum_{t=0}^\infty \frac{l_{x+t}}{l_x} - \frac{1}{2}, \qquad (5)$$

which is normally used to compute values for $\mathring{e}_x$.

Equivalent names for the life expectancies are *mean survival time* for $\mathring{e}_0$ and *mean residual survival time at age x* for $\mathring{e}_x$. The *median length of life* is the median in the distribution of T; it used to be called the *probable length of life* (*see* **Median Survival Time**). Correspondingly, the *median residual length of life* at age x used to be called the *probable residual length of life*. If we denote the latter by ξ_x, then it is defined by the relation $l_{x+\xi_x} = \frac{1}{2}l_x$.

The above functions can be computed for cohort life tables and for period life tables. Figure 1 shows plots of the function $\mathring{e}_x$ according to the mortality experience for Swedish women in 1891–1900 and 1990–1994. The life expectancy at birth has increased from 53.6 years in the older table to 80.8 some one hundred years later. Note that in the older table $\mathring{e}_x$ increases with x up to age 2 and remains above $\mathring{e}_0$ up through age 11. When mortality is high at very young ages, surviving the first part of life *increases* your expected remaining lifetime. As a consequence of mortality improvements for very young children, these features have

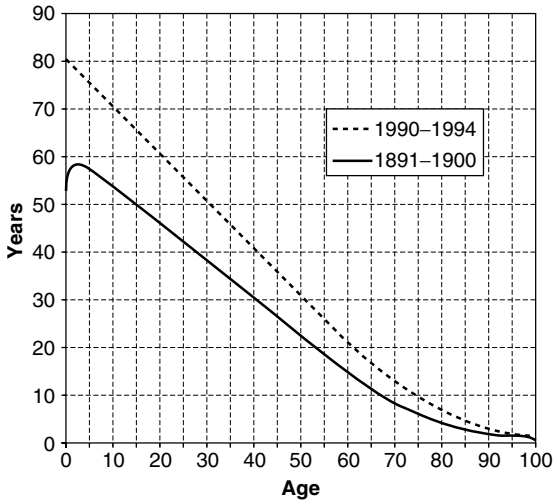

Figure 1 Residual life expectancy for Swedish women, 1891–1900 and 1990–1994

disappeared in the younger table. Note that the expected *total* lifetime, $x + \mathring{e}_x$, always increases with x throughout the human lifespan. (One can show that the derivative of this function is always positive.) The longer you have lived already, the longer you can expect the total length of your life to be.

In a multiple-decrement situation, formula (5) can be used to compute a residual life expectancy $\mathring{e}_x^{(-k)}$ from the decrement series of the cause-deleted life table for risk k. The difference $\mathring{e}_x^{(-k)} - \mathring{e}_x$ is the gain one would get in residual life expectancy at age x *if* it were possible to eliminate risk cause k without changing the risk intensity of any other cause of decrement. Dublin et al. [5, p. 96] note that according to the cause-specific mortality of the US in 1939–1941 the gain would be 9.01 years for white men and 8.80 years for white women at age 0 if one could eliminate the risk of death due to cardiovascular–renal diseases at all ages (and change no other cause-specific mortality risks). The gains from eliminating the risk of death in cancer alone were much less (1.39 years for men and 2.05 years for women).

History and Literature

The first step toward the development of the life table was taken when **Graunt** [9] published his famous *Bills of Mortality*. There were subsequent contributions by **Halley**, Huygens, Leibniz, Euler, and others. Deparcieux [4] clarified the definition of the life expectancy and identified the need for separate tables for men and women. Wargentin [17] was the first to publish real age-specific death rates, and the first to do so for a whole country. Price [14] included most of the columns now associated with the life table, and the tables by Duvillard [7] contained them all. The basic notions of cause-eliminated life tables go back to Bernoulli [1]. Cournot [3] developed the essentials of their mathematics. See Dupâquier [6] and Seal [15] for historical overviews. Smith & Keyfitz [16] have collected extracts from many original texts.

Life-table techniques are described in most introductory textbooks on the methods of actuarial statistics, biostatistics, demography, or epidemiology. See for example, Chiang [2], Elandt-Johnson & Johnson [8], or Manton & Stallard [13].

References

[1] Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Académie Royale des Sciences, Mémoires, Année 1760*, pp. 1–45.

[2] Chiang, C.L. (1984). *The Life Table and its Applications*. Krieger, Malabar.

[3] Cournot, A. (1843). *Exposition de la théorie des chances et des probabilités*. Hachette, Paris.

[4] Deparcieux, A. (1746). *Essai sur les probabilités de la durée de la vie humaine*. Guérin Frères, Paris.

[5] Dublin, L.I., Lotka, A.J. & Spiegelman, M. (1947). *Length of Life.* Ronald Press, New York.

[6] Dupâquier, J. (1996). *L'invention de la table de mortalité*. Presses Universitaires de France, Paris.

[7] Duvillard, E. (1806). Analyse des tableaux de l'influence de la petite vérole, et de celle qu'un préservatif tel que la vaccine peut avoir sur la population et la longévité. Paris.

[8] Elandt-Johnson, R.C. & Johnson, N.L. (1980). *Survival Models and Data Analysis*. Wiley, New York.

[9] Graunt, J. (1662). *Natural and Political Observations Made Upon the Bills of Mortality*. London.

[10] Heligman, L. & Pollard, J. (1980). The age pattern of mortality, *Journal of the Institute of Actuaries* **107**, 49–75.

[11] Hoem, J.M. (1972). Analytic graduation, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability 1970*, Vol. 1, L.M. Le Cam, J. Neyman & E.L. Scott, eds. University of California Press, Berkeley, pp. 569–600.

[12] Keyfitz, N. (1982). Keyfitz method of life-table construction, in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 371–372.

[13] Manton, K.G. & Stallard, E. (1984). *Recent Trends in Mortality Analysis*. Academic Press, New York.

[14] Price, R. (1783). *Observations of Reversionary Payments: On Schemes for Providing Annuities for Widows, and for Persons in Old Age; and on the National Debt*. Cadell & Davies, London.

[15] Seal, H. (1977). Studies in the history of probability and statistics, XXV: multiple decrements or competing risks, *Biometrika* **64**, 429–439.

[16] Smith, D. & Keyfitz, N. (1977). *Mathematical Demography: Selected Papers*. Springer-Verlag, Heidelberg.

[17] Wargentin, P. (1766). *Mortaliteten i Sverige, i anledning af Tabell-Verket*. Kongl. Vetenskaps-Academiens Handlingar, Stockholm.

(*See also* **Demography**; **Vital Statistics, Overview**)

JAN M. HOEM

Likelihood Principle

All statisticians will agree that many, if not most, modern procedures in all approaches to statistics involve some use of the **likelihood** function. However, they will be much less in agreement about the applicability of some general likelihood principle. Although such principles have generated animated debate in the past, in recent years, fewer statisticians have shown interest in the foundations of their discipline, so discussion has waned (*see* **Inference, Foundations of**).

Suppose that we are interested in obtaining empirical information about a fixed set of completely specified models $P(\boldsymbol{\psi}) \in \mathcal{P}$, indexed by the unknown parameters $\boldsymbol{\psi}$. We observe values $\mathbf{y}$ of the relevant **random variable** $Y \in \mathcal{Y}$, using an appropriate study design. Then, the probability of these specific observations can be calculated for any member of the set $\mathcal{P}$. This is called the *likelihood function* $L(\boldsymbol{\psi}; \mathbf{y})$ for $\boldsymbol{\psi}$ given the set of models $\mathcal{P}$ and the observed data $\mathbf{y}$ [7]. Observed likelihood functions are said to be proportional if the proportionality constant contains only functions of $\mathbf{y}$, but not of $\boldsymbol{\psi}$.

On the basis of the likelihood function, a number of *likelihood principles* have been formulated. These are listed below from the weakest to the strongest.

1. Any model indexed by $\boldsymbol{\psi}_1$ is more plausible or *likely* than another $\boldsymbol{\psi}_2$, in the light of only the given observed data, written $L(\boldsymbol{\psi}_1; \mathbf{y}) > L(\boldsymbol{\psi}_2; \mathbf{y})$ if it makes these data more probable [8].
2. Data sets coming from replications of the same study design, thus having *the same sample space*, and having proportional likelihood functions contain the same information about $\boldsymbol{\psi}$ given $P(\boldsymbol{\psi})$ [6].
3. *Any* data sets with proportional likelihood functions contain the same information about $\boldsymbol{\psi}$ given $P(\boldsymbol{\psi})$ [2, 4].
4. From Principle 3, all inferences must be based on the likelihood function in such a way that proportional likelihoods lead to the same conclusions about $\boldsymbol{\psi}$ given $P(\boldsymbol{\psi})$ [3].

These should all be distinguished from **maximum likelihood** (sometimes also called a principle) that provides only a point estimate, perhaps with an asymptotic interval of precision. **Inferences** using likelihood principles involve the complete likelihood function, not just one value of it.

Principle 1 refers only to the observed data, making no reference to repetitions, to alternative designs, or to prior information not contained in $\mathcal{P}$. Principle 2 has been called the weak likelihood principle and Principle 3, the strong likelihood principle. These principles are closely related to **sufficient statistics** and to *conditioning* on the design used as well as on the observed outcome. At least for discrete data spaces, likelihood principle 3 can be derived from such sufficiency and **conditionality principles** [4], whereas Principle 2 is equivalent to sufficiency.

In contrast to the first two principles, Principles 3 and 4 imply that the same conclusions would be made from samples drawn from different probability spaces, that is, with different sample designs, if the resulting likelihood functions are proportional. Then, inference should not depend on the sample space $\mathcal{Y}$, but only on $\mathcal{P}$ and on the *observed* values $\mathbf{y}$.

The most common simple example used to illustrate the import of the likelihood principle is repeated Bernoulli trials (*see* **Binary Data**). Thus, the family of models $\mathcal{P}$, indexed by the constant unknown probability π, will describe a series of (supposedly) independent Bernoulli trials to be performed. Notice that independence and constant probability are *assumptions* of the models.

1. In a fixed sample size design, a coin is tossed a fixed number of times n and the number of heads y recorded. If the tosses are independent with constant probability of heads, the likelihood function is **binomial**,

$$L(\pi; y, n) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \qquad (1)$$

   and the sample space is $\mathcal{Y} = \{y : 0 \leq y \leq n\}$.
2. In a sequential design (*see* **Sequential Analysis**), the coin will be tossed until a fixed number of tails c has been observed. Thus, y is the number of heads recorded in a random total sample size of $n = c + y$ tosses. With the same conditions of independence and constant probability as in Design 1, the likelihood function is **negative binomial**,

$$L(\pi; y, c) = \binom{c+y-1}{c-1} \pi^y (1-\pi)^c \qquad (2)$$

   and the sample space is $\mathcal{Y} = \{y : y \in \mathbf{N}\}$.

The second design differs in several important ways from the first. Not only is the sample size random, but we must also have available the result of each trial, as it occurs, in order to know whether to stop or not, and, hence, we must necessarily have the time-ordered sequence of results. This will often be the reason why such a design is chosen. However, this information is always lost in the negative binomial likelihood given above so that, in such cases, it would not be the appropriate likelihood function. On the other hand, Design 1 can be performed without this sequential information being available. If it is available, it also is discarded in constructing the above likelihood.

Once the total number of tosses, $n = c + y$, and the number of heads, y, are known, both likelihood functions, for a model of independent events with constant probability, are proportional to

$$\pi^y(1-\pi)^{n-y} = \pi^y(1-\pi)^c \tag{3}$$

and y is the sufficient statistic for π (given n or c fixed). However, in specifying the likelihood for Design 2, and perhaps that for Design 1, we discard relevant information about π. This is only necessarily available in the sequential order of results produced by Design 2. This lost information would allow us to check the assumptions of independence and constant probability π.

The difference between the two designs (fixed versus random sample size) is reflected in the frequentist approach to the problem. With the same null hypothesis and the same observed results, a ***P* value** will generally be different in the two designs because the sample spaces differ. Such inferences violate Principles 3 and 4.

In this sense, a frequentist procedure does not clearly separate testing the **null hypothesis** of a given fixed value for π from checking these more fundamental assumptions of the model. However, application of Principle 3 or 4 in this example clearly involves the assumption that the set of models under consideration, indexed by π, and defining the likelihood function, contains the true model. Dependence or nonconstant probability are excluded by hypothesis and information to check these assumptions is discarded.

Application of the stronger likelihood principles involves a number of strong assumptions that should be made explicit:

1. The set of models $P(\boldsymbol{\psi})$ must be fully specified and it must be assumed that one of them is true.
2. All uncertainty must be included in $\boldsymbol{\psi}$. It must index all unknowns in the problem, such as unknown variables and values to be predicted, and not just parameters in the classical sense.
3. If different designs are involved, $\boldsymbol{\psi}$ must be identical in all of them.
4. Choice among different designs must be noninformative.

In a decision problem, such as testing for treatment effect in a Phase III **clinical trial**, these may all be reasonable assumptions. However, in scientific research, Assumptions 1 and 4 are usually questionable.

- Scientific models are always approximations and can never be "true". The scientist must be able to question *all* assumptions and, as far as possible, check them with the data.
- The choice between two designs is rarely indifferent (made by random selection). One design will be used rather than another because it provides more appropriate information, given its costs.

Discussion of the likelihood principle necessarily involves the importance of the *stopping rule*. This specifies how observation is ended in a study. Stopping rules must only use information available up until the point in time when stopping is to occur. The role of the stopping rule can clearly be seen in the above Bernoulli example. The implication of likelihood principle 3 is that only the likelihood based on the final result of a properly conducted sequential trial should be used in making inferences. It is irrelevant that intermediate checks of the data were made to determine when to stop (perhaps by examining intermediate likelihoods). This contrasts with frequentist analyses involving "spending" the test probability over the course of a sequential trial (*see* **Interim Analysis of Censored Data**). Here, the final test depends on what happened throughout the trial and not just on the final outcome.

The way to handle **censoring** and **missing data** such as dropouts is closely related to the stopping rule. According to the stronger likelihood principles, if the censoring or missing data mechanism is noninformative about the process of interest, it need not explicitly appear in the likelihood function. Planned censoring will generally be of this form but missing

data, by definition, are generally unplanned and hence more difficult. Empirically, missingness, if not completely independent of the process under study (the broken test tube), will invariably be closely implicated with that process and must be modeled and included in the likelihood function no matter what approach is used.

Different approaches to inference necessarily imply differing appreciation of the various likelihood principles. The likelihood school maintains that Principle 1 is adequate to draw many scientific conclusions. However, the relative plausibilities of a set of models must be tempered by their complexity using a penalty, such as is done with information criteria (AIC [1], BIC [9], or their modifications; *see* **Akaike's Criteria**).

The frequentist school holds that Principle 1 is inadequate because indication of performance in repeated application of inferences is necessary. However, its techniques involve the use of the sample space, that is, what might have been observed, thus excluding Principles 3 and 4. If obeyed to the letter, this has drastic consequences. Although a study might have no censoring or missing data, if it could have had some, then the sample space for calculating test and confidence probabilities should include the possibility of censoring or missingness.

The **Bayesian** school upholds Principle 4 and provides strong arguments for the need to use **prior** probabilities in implementing it. Historically, up until about 1990, most debate was centered around this principle [3]. More recently, Principle 1 has received increasing attention, notably in the context of applying model selection criteria [5] such as the AIC or BIC (*see* **Bayesian Methods for Model Comparison**).

The type of design adopted for a given study will depend on the type of information required to be collected, as well as on other factors such as cost. Rarely will different designs provide proportional likelihood functions, as in the example of Bernoulli trials above. Thus, in most applied statistical practice, Principles 3 and 4 will not be of major concern in this context.

A more fundamental issue involves whether or not inferences should be conditioned solely on the observed data. Frequentist tests and confidence intervals involve the probabilities of events that might have been observed, given the design, and not just on what was observed. The Bayesian school argues that these are invalid. However, the latter argument can only be sustained if one is certain that the true model is contained in the set under consideration.

Thus, the strength of likelihood principle that one is prepared to apply, in the classification outlined above, must depend on one's confidence in the validity of the set of models under study. If one is certain about the set of models and only wishes to distinguish among them in the light of new data, a strong likelihood principle can be justified. On the other hand, if all aspects of the models are under scientific scrutiny, choice of inference based on a much weaker likelihood principle will be more prudent. However, restricting inference only to a likelihood principle narrows statistical analysis to routine **decision** problems, making scientific *discovery* impossible.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Inference Theory*, B.N. Petrov & F. Csàki eds. Akadémiai Kiadó, Budapest, pp. 267–281.

[2] Barnard, G.A. (1949). Statistical inference, *Journal of the Royal Statistical Society* **B11**, 115–149 (with discussion).

[3] Berger, J.O. & Wolpert, R.L. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward.

[4] Birnbaum, A. (1962). On the foundations of statistical inference, *Journal of the American Statistical Association* **57**, 269–306 (with discussion).

[5] Burnham, K.P. & Anderson, D.R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, Berlin.

[6] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

[7] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London* **A222**, 309–368.

[8] Fisher, R.A. (1959). *Statistical Methods and Scientific Inference*. Oliver & Boyd, Edinburgh.

[9] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

(*See also* **Foundations of Probability**)

J.K. LINDSEY

Likelihood Ratio Tests

Suppose $\mathbf{x}_{\text{obs}}$ is a vector of data that have been collected in order to test a hypothesis. The likelihood ratio test is a hypothesis testing procedure that can be performed in a wide variety of situations. To apply the procedure, we must be able to regard the observed data $\mathbf{x}_{\text{obs}}$ as having been drawn at random from a population whose distribution is described by a joint density function $f(\mathbf{x};\boldsymbol{\theta})$ depending on an unknown parameter vector $\boldsymbol{\theta}$, and we must formulate the hypothesis as a statement about $\boldsymbol{\theta}$. The distribution of the population may be discrete or continuous or may have both discrete and continuous components, such as occurs with some censored data.

The function $L(\boldsymbol{\theta}) = f(\mathbf{x}_{\text{obs}};\boldsymbol{\theta})$ is called the **likelihood** function. Let Θ denote the set of possible parameter vectors and let Θ_0 be a subset of Θ. For testing the null hypothesis H_0: $\boldsymbol{\theta} \in \Theta_0$ vs. the alternative hypothesis H_a: $\boldsymbol{\theta} \notin \Theta_0$, Neyman & Pearson [11] introduced the *likelihood ratio test* (abbreviated as *LR test*; also called the *generalized likelihood ratio test* or *maximum likelihood ratio test*). Let $L(\hat{\boldsymbol{\theta}})$ be the maximum value of the likelihood as $\boldsymbol{\theta}$ varies over Θ, let $L(\hat{\boldsymbol{\theta}}_0)$ be the maximum as $\boldsymbol{\theta}$ varies over Θ_0, and let

$$\lambda = \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}.$$

The maximizing value $\hat{\boldsymbol{\theta}}$ of the parameter vector is called the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ (*see* **Maximum Likelihood**), and $\hat{\boldsymbol{\theta}}_0$ is the MLE under the null hypothesis. We can expect the MLE $\hat{\boldsymbol{\theta}}$ to be close to the true parameter vector. If the null hypothesis were true, i.e. if the true parameter vector were in Θ_0, then we would expect both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_0$ to be close to the true parameter vector, and hence we would expect λ to be close to 1. The likelihood ratio test rejects the null hypothesis if λ is significantly smaller than 1, i.e. if the maximum likelihood under the null hypothesis is significantly smaller than the maximum likelihood under the alternative hypothesis.

In special situations the ***P* value** of a likelihood ratio test can be calculated exactly, but in general it must be approximated. If $f(\mathbf{x};\boldsymbol{\theta})$ satisfies certain regularity conditions (discussed below), then an approximate P value can be obtained as the proportion of a **chi-square distribution** that is larger than $-2\log\lambda$, where the number of **degrees of freedom** is the number of independent conditions that the null hypothesis imposes on the parameter vector $\boldsymbol{\theta}$. This proportion can be calculated by using the chi-square cumulative distribution function that is available in some computer packages, or bounds can be put on it by using a chi-square table.

Example 1

The diastolic blood pressures of 15 patients with moderate essential hypertension were measured immediately before and two hours after taking the drug captopril [4, p. 72]. A common way to analyze such data is to calculate the differences ("after" minus "before") and regard them as a random sample from a normally distributed population with unknown mean, δ, and unknown standard deviation, σ. The null hypothesis that the drug has no effect on blood pressure can be formulated as H_0: $\delta = 0$. The likelihood function is

$$L(\delta,\sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\delta)^2\right],$$

where n is the sample size and the x_is are the differences. For the observed blood pressure data this becomes

$$L(\delta,\sigma) = \frac{1}{(2\pi)^{15/2}\sigma^{15}} \times \exp\left[-\frac{1}{2\sigma^2}(15\delta^2+278\delta+2327)\right].$$

The likelihood attains its maximum value, $L(\hat{\delta},\hat{\sigma})$, at $\hat{\delta} = -9.27$ and $\hat{\sigma} = 8.32$. Under the null hypothesis, the likelihood attains a maximum value, $L(0,\hat{\sigma}_0)$, at $\hat{\sigma}_0 = 12.46$. Then $-2\log\lambda = 2[\log L(-9.27, 8.32) - \log L(0, 12.46)] = 12.10$. Since the null hypothesis imposes only one condition on the parameters, the number of degrees of freedom is 1. From a chi-square table we see that the approximate P value is less than 0.001, and we conclude that the drug has an effect.

Example 1 is simple enough that there are explicit formulas for the MLEs: $\hat{\delta} = \overline{x}$, where $\overline{x}$ is the sample mean; $\hat{\sigma} = [(n-1)/n]^{1/2}s$, where s is the sample standard deviation; and $\hat{\sigma}_0 = (\sum x_i^2/n)^{1/2}$. Therefore there are explicit formulas for $L(\hat{\delta},\hat{\sigma})$ and $L(0,\hat{\sigma}_0)$ and hence for λ, namely $\lambda = (\hat{\sigma}/\hat{\sigma}_0)^n$. In general, however, $L(\hat{\boldsymbol{\theta}})$ and $L(\hat{\boldsymbol{\theta}}_0)$ must be calculated by

numerical optimization methods (*see* **Optimization and Nonlinear Equations**).

In Example 1 the LR test is equivalent to the usual t test. In fact, if $x_1, \ldots, x_n$ is a random sample from a normally distributed population with unknown mean μ and unknown standard deviation σ, then the LR test statistic, $-2\log\lambda$, for testing H_0: $\mu = \mu_0$, is an increasing function of the **Student's *t* statistic** $|t| = |\overline{x} - \mu_0|/(s/\sqrt{n})$. Therefore, in this situation it is possible to obtain the exact P value of the LR test. (But of course it is exact only if the population is exactly normally distributed.) From a t table it is seen that the exact P value also is less than 0.001.

Example 2

For the data in Example 1, the assumption that the blood pressure differences come from a normally distributed population can be justified by arguing that the data contain no outliers and that the t test is robust against nonnormality. But if one felt uncomfortable about assuming normality, a different test of the null hypothesis of no drug effect could be performed by regarding the differences as a random sample from a continuous population with an unknown proportion π of positive values. Here we are assuming nothing about the population other than that it is continuous. The null hypothesis can be formulated as H_0: $\pi = 0.5$. Let x denote the number of patients whose blood pressure differences were positive. The likelihood function is

$$L(\pi) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}.$$

This is another simple example in which the MLE has an explicit formula: $\hat{\pi} = x/n$. For the blood pressure data, $-2\log\lambda = 2[\log L(2/15) - \log L(0.5)] = 9.01$. The number of degrees of freedom is 1. From a chi-square table we see that the approximate P value is between 0.01 and 0.001, and we again conclude that the drug has an effect. The exact P value is also available in this example. By using the fact that, under our assumptions, the exact distribution of x is **binomial**, one obtains $P = 0.007$.

Example 3

One half of a group of 42 leukemia patients were treated with the drug 6-mercaptopurine and the other half were given a placebo [3]. Their remission times, in weeks, were recorded during a period of one year. At the end of the year some patients still had had no remission, and so these observations were censored. Let us assume that the patients were selected and treated independently of one another. A reasonable model for these data is that they are two independent **censored** random samples from two **Weibull** distributions. The likelihood function is

$$L(\kappa_1, \rho_1, \kappa_2, \rho_2) = L_1(\kappa_1, \rho_1)L_2(\kappa_2, \rho_2),$$

where

$$L_i(\kappa_i, \rho_i) = \kappa_i^{d_i}\rho_i^{d_i\kappa_i} \times \exp\left[(\kappa_i - 1)\sum_{\text{unc}}\log x_{ij} - \rho_i^{\kappa_i}\sum_{\text{all}} x_{ij}^{\kappa_i}\right],$$

in which x_{ij} is the jth observation in the ith subgroup ($i = 1, 2$, $j = 1, \ldots, 21$), d_i is the number of uncensored observations in the ith subgroup, "unc" indicates summation over the uncensored observations, and "all" indicates summation over all the observations, including the censored ones. The null hypothesis of no treatment effect can be expressed as H_0: $\kappa_1 = \kappa_2$ and $\rho_1 = \rho_2$.

No explicit expression for λ is available in this example. However, there is an explicit expression for the MLE $\hat{\rho}_i$ as a function of κ_i, namely $\hat{\rho}_i = (d_i/\sum_{\text{all}} x_{ij}^{\kappa_i})^{1/\kappa_i}$. Substitute this into $L_i(\kappa_i, \hat{\rho}_i)$ to obtain the **profile likelihood** $L_{\text{P}i}(\kappa_i)$. A numerical procedure must be used to maximize the profile likelihood, yielding the MLE $\hat{\kappa}_i$, from which we obtain $L(\hat{\kappa}_1, \hat{\rho}_1, \hat{\kappa}_2, \hat{\rho}_2) = L_{\text{P}1}(\hat{\kappa}_1)L_{\text{P}2}(\hat{\kappa}_2)$. Similarly, combining the two subgroups into a single sample under the assumption that H_0 is true, we can obtain $L_{\text{P}0}(\hat{\kappa}_0)$ and then $-2\log\lambda = 2[\log L_{\text{P}1}(\hat{\kappa}_1) + \log L_{\text{P}2}(\hat{\kappa}_2) - \log L_{\text{P}0}(\hat{\kappa}_0)] = 66.17$. The number of degrees of freedom is 2. From a chi-square table we see that the approximate P value is less than 0.001, and we conclude that the treatment has an effect.

Likelihood ratio tests are commonly used in a number of different statistical areas. In **multiple linear regression** and **analysis of variance** for models with independent and identically normally distributed errors, the usual ***F* tests** are equivalent to LR tests. In **multivariate analysis**, tests using the Wilks lambda criterion (*see* **Discriminant Analysis, Linear**) are equivalent to LR tests. In the analysis of **generalized linear models**, the reduction in deviance between a

model and an extended model is equal to the LR test statistic $-2\log\lambda$ for testing whether the two models are significantly different. To test hypotheses about **contingency tables**, one typically uses either the Pearson **chi-square test** or the LR test.

Likelihood ratio tests have been proposed in many other areas too. For any parametric statistical model that satisfies certain, fairly general, regularity conditions, the null distribution of $-2\log\lambda$ is well approximated by a chi-square distribution, and so it is straightforward, at least in theory, to apply the LR test to test the parameters of the model. In practice, calculation of the likelihood ratio often requires numerical optimization, which can involve substantial computation; but computation is becoming less of a concern as computer capabilities increase. The regularity conditions mentioned above assume that the support $\{\mathbf{x} : f(\mathbf{x};\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$, that $f(\mathbf{x};\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$, and a few other requirements that are mathematically technical but often satisfied. Under such conditions the null distribution of the LR test statistic can be approximated by a chi-square distribution with d degrees of freedom, where d is the difference in the dimensions of Θ and Θ_0. By the dimension of Θ is meant the number of "freely varying" components in the vector $\boldsymbol{\theta}$. More precisely, the dimension of Θ is k if it contains a solid k-dimensional cube and does not contain a solid $(k+1)$-dimensional cube. The degrees of freedom d can also be described as the number of independent conditions that the null hypothesis imposes on the parameter vector $\boldsymbol{\theta}$.

The chi-square approximation for the null distribution of the LR test statistic is based on asymptotic theory (*see* **Large-sample Theory**) and so it may not work well if the sample size is small. For example, for small categorical data sets the chi-square approximation is usually poor [1] and produces inaccurate P values. The approximation may also be inadequate if the number of parameters is large, such as when testing the goodness of fit of a **generalized linear model** against a saturated model [5], or if the null parameters are on the boundary of the parameter set, such as when testing whether a variance component is zero [9, p. 501].

In situations where the chi-square distribution poorly approximates the null distribution of the likelihood ratio test statistic it is sometimes possible to obtain a better approximation, or even the exact distribution. For example, for the Wilks lambda criterion, its null distribution is better approximated by using an F distribution rather than a chi-square distribution [12], and there are numerical procedures for calculating its exact critical values to any desired precision. Higher-order asymptotic methods, such as Bartlett adjustment [10] (*see* **Bartlett's Test**) or modified profile likelihood [8, 13], can be used to adjust a likelihood ratio so that its null distribution can be well approximated.

In the simple case of testing H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. H_a: $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, the **Neyman–Pearson lemma** states that the likelihood ratio test is the **most powerful** test for any given level. In a one-parameter **exponential family** the LR test of a one-sided hypothesis is a uniformly most powerful test. Under certain more general conditions, LR tests have been shown to be optimal in several senses: asymptotically most stringent [15], asymptotically locally most powerful unbiased [2], and Bahudur efficient [14]. However, LR tests may not be robust to violations of the model assumptions [6, 7].

The likelihood ratio test has wide applicability and has given reasonable results in a large number of cases in which its performance has been studied.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[3] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, New York.

[4] Cox, D.R. & Snell, E.J. (1981). *Applied Statistics: Principles and Examples*. Chapman & Hall, New York.

[5] Firth, D. (1991). Generalized linear models, in *Statistical Theory and Modelling*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, New York, pp. 55–82.

[6] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

[7] Kent, J.T. (1982). Robust properties of likelihood ratio tests, *Biometrika* **69**, 19–27.

[8] Lindsey, J.K. (1996). *Parametric Statistical Inference*. Oxford University Press, Oxford.

[9] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute Inc., Cary.

[10] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, New York.

[11] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika* **20A**, 175–240, 263–295.

[12] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. Wiley, New York.

[13] Severini, T.A. (2000). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.

[14] van der Vaart, A.W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[15] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* **54**, 426–482.

(*See also* **Likelihood Ratio**)

D. BIRKES

Likelihood Ratio

Let $\mathbf{x}_{\text{obs}}$ be a vector of observed data. To analyze the data, one often assumes that $\mathbf{x}_{\text{obs}}$ has been randomly drawn from a population whose distribution is described by a joint density function $f(\mathbf{x};\boldsymbol{\theta})$ depending on an unknown parameter vector $\boldsymbol{\theta}$. The distribution may be discrete or continuous or may have both discrete and continuous components, such as occurs with some censored data. The **likelihood** of the parameter vector $\boldsymbol{\theta}$ based on the data vector $\mathbf{x}_{\text{obs}}$ is defined to be $L(\boldsymbol{\theta}) = f(\mathbf{x}_{\text{obs}};\boldsymbol{\theta})$. Given two possible values of the parameter vector, the one with the greater likelihood is regarded as being more likely to be the true parameter value. That is to say, the value $\boldsymbol{\theta}_1$ is more likely than the value $\boldsymbol{\theta}_2$ if the *likelihood ratio* $L(\boldsymbol{\theta}_1)/L(\boldsymbol{\theta}_2)$ is greater than 1. In fact, the likelihood ratio can be used as a quantitative measure of the strength of support that the data provide for $\boldsymbol{\theta}_1$ in comparison with $\boldsymbol{\theta}_2$ [1, 4, 5].

For example, consider an urn containing 10 balls, θ of which are red and $10-\theta$ are green. Suppose we draw two balls at random without replacement from the urn and both balls are red. The probability of such an outcome is $f(2;\theta) = (\theta/10)[(\theta-1)/9] = \theta(\theta-1)/90$. To compare the possibility that $\theta = 6$ with the possibility that $\theta = 5$, we can calculate the likelihood ratio $L(6)/L(5) = f(2;6)/f(2;5) = 1.5$ and state that $\theta = 6$ is 1.5 times as likely as $\theta = 5$.

The likelihood ratio $L(\boldsymbol{\theta}_0)/L(\boldsymbol{\theta}_1)$ is a sensible test statistic for testing the simple null hypothesis $\mathrm{H}_0\colon \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. the simple alternative hypothesis $\mathrm{H}_1\colon \boldsymbol{\theta} = \boldsymbol{\theta}_1$ (*see* **Hypothesis Testing**). If the ratio is small, then the likelihood of $\boldsymbol{\theta}_0$ is small in comparison with the likelihood of $\boldsymbol{\theta}_1$, and so it makes sense that we should reject H_0. Moreover, Neyman & Pearson [3] showed that this test is the most powerful one among all tests having the same level (*see* **Neyman–Pearson Lemma**). For testing a general null hypothesis $\mathrm{H}_0\colon \boldsymbol{\theta} \in \boldsymbol{\theta}_0$ vs. a general alternative hypothesis $\mathrm{H}_1\colon \boldsymbol{\theta} \in \boldsymbol{\theta}_1$, Neyman & Pearson [2] proposed the test statistic $\lambda = L(\hat{\boldsymbol{\theta}}_0)/L(\hat{\boldsymbol{\theta}})$, where $L(\hat{\boldsymbol{\theta}}_0)$ is the maximum value of the likelihood as $\boldsymbol{\theta}$ varies over $\boldsymbol{\theta}_0$ and $L(\hat{\boldsymbol{\theta}})$ is the maximum value of the likelihood as $\boldsymbol{\theta}$ varies over $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ (*see* **Likelihood Ratio Tests**). The null hypothesis is rejected if λ is small. Some authors call λ a *likelihood ratio* but it is also called a *likelihood ratio criterion* or *generalized likelihood ratio* or *maximum likelihood ratio* or *likelihood ratio test statistic*. The quantity $-2\log\lambda$ is also sometimes called a *likelihood ratio test statistic*.

Other Interpretations of the Likelihood Ratio

Although the concept of likelihood is distinct from the concept of probability, it is possible to give the likelihood ratio $L(\boldsymbol{\theta}_1)/L(\boldsymbol{\theta}_2)$ a direct probability interpretation if we take a **Bayesian** viewpoint. Suppose that, on the basis of past experience, it can be assumed that the true parameter is either $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$ and that both are equally probable. Then our **prior distribution** is given by $\Pr(\boldsymbol{\theta}_1) = \Pr(\boldsymbol{\theta}_2) = 0.5$. The likelihood ratio coincides with the ratio $\Pr(\boldsymbol{\theta}_1|\mathbf{x}_{\text{obs}})/\Pr(\boldsymbol{\theta}_2|\mathbf{x}_{\text{obs}})$ of the posterior probabilities of the two parameters.

If the likelihood ratio is viewed as a random function, in the manner indicated below, then it is a minimal sufficient statistic (*see* **Sufficient Statistic**). Choose a fixed parameter vector $\boldsymbol{\theta}_0$ and, for each fixed value of $\mathbf{x}$, regard the likelihood ratio $R(\mathbf{x})(\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{x})/L(\boldsymbol{\theta}_0;\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta})/f(\mathbf{x};\boldsymbol{\theta}_0)$ as a real-valued function of $\boldsymbol{\theta}$. Let $\mathbf{X}$ be a random vector with density $f(\mathbf{x};\boldsymbol{\theta})$. It is a consequence of the factorization theorem for sufficient statistics that $R(\mathbf{X})$ is a minimal sufficient statistic.

References

[1] Blume, J.D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine* **21**, 2563–2599.

[2] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika* **20A**, 175–240, 263–295.

[3] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society, Series A* **231**, 289–337.

[4] Reid, N. (2000). Likelihood. *Journal of the American Statistical Association* **95**, 1335–1340.

[5] Royall, R. (2000). On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association* **95**, 760–780.

D. BIRKES

Likelihood

In general use the word *likelihood* is a synonym for **probability** but in statistics it has a more specific meaning; it is the probability (or probability density) of the observed data given the probability model which gave rise to the data. Likelihood is used to compare different possible candidate values for the *parameters* of the model, and for this purpose it needs to be defined only up to a constant of proportionality: any constant multiple of the likelihood serves equally well. When comparing two candidate values for a parameter, the one with the greater likelihood is said to be *more likely*, and parameter values for which the probability of the observed data is greatest are known as *most likely* values, or **maximum likelihood** estimates. The concept of likelihood is central to both the *frequency* and the **Bayesian** theory of **inference**. In addition there have been many attempts to found a theory of inference on likelihood alone.

A Simple Example

Let 10 subjects be followed for five years, and a record made of whether they die (fail) or survive. A simple probability model is that the outcome for each subject is independently random with probability π for failure and $1-\pi$ for survival. The probability π is the *parameter* of the model. When four subjects fail, and six survive, the probability of the observed data is found from the **binomial distribution** to be

$$L(\pi) = 210\pi^4(1-\pi)^6.$$

Suppose we wish to compare $\pi = 0.1$ with $\pi = 0.5$ as possible values for the true value which gave rise to the data. The two likelihoods are $L(0.1) = 0.0112$ and $L(0.5) = 0.2051$, so $\pi = 0.5$ is more likely than $\pi = 0.1$. The most likely value is $\pi = 0.4$, which has likelihood 0.2508. Since the likelihood can be scaled by any constant without altering such comparisons it is often convenient to scale it to take the value 1 when π takes its most likely value. The scaled likelihood for π is then the **likelihood ratio** $L(\pi)/L(\hat{\pi})$, where $\hat{\pi}$ is the most likely value for π.

Part (a) of Figure 1 shows the likelihood ratio for a range of possible values for π. Values of π corresponding to a high likelihood ratio are said to be *supported* by the data; those with a low likelihood ratio are not supported. The distinction between supported and not supported depends on where the cut-point is placed on the likelihood ratio scale. A convenient summary of the information about π in the data is provided by the most likely value of π and

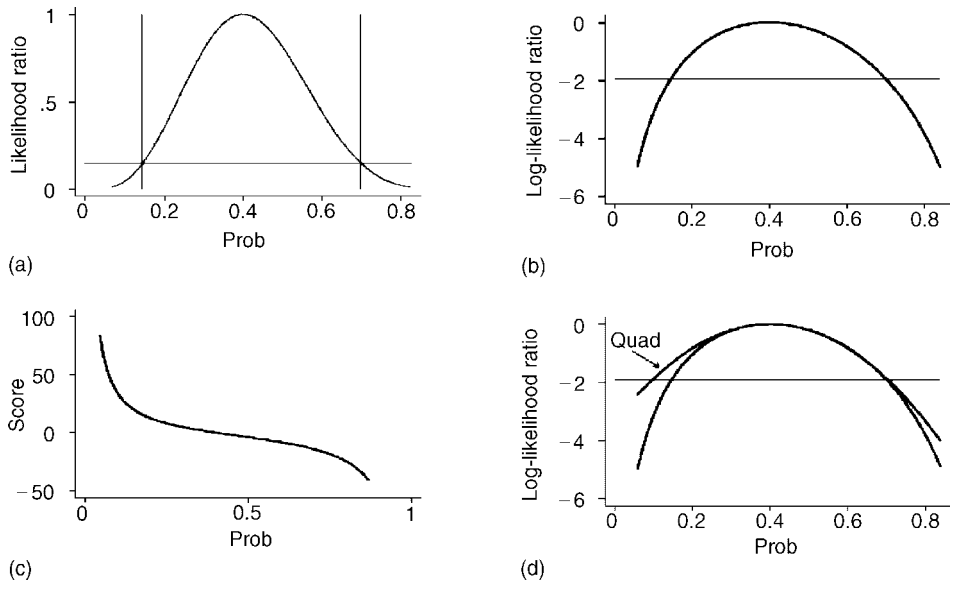

Figure 1 (a) Likelihood, (b) log likelihood, (c) score, and (d) quadratic approximation

a range of values which are supported at a given cut-point. The choice of cut-point can be regarded as a matter of convention; for example, we might all agree that parameter values with likelihood ratios above 0.15 are supported, while those with values below are not supported. Another approach is to choose the cut-point in terms of how well the supported range works when evaluated for repeated samples from the probability model assumed to have given rise to the data. This is called the *frequency* approach to statistics.

For any particular value of π the cut-point 0.1465 produces a supported range which includes the value of π in approximately 95% of repeated samples, provided the likelihood curve has roughly a *normal* bell shape. For this cut-point, then, the range of supported values corresponds to a 95% **confidence interval**. The supported range may also be thought of as a Bayesian *plausibility* interval based on a uniform *prior* belief about the true value of π. With a cut-point 0.1465 the area under the curve in part (a) of Figure 1, between the two verticals, is approximately 95% of the total area, so the *posterior* probability that π lies between the two limits is approximately 0.95. For all but very small studies the likelihood will have a normal bell shape (this is called the **central limit theorem**), and the three approaches (given likelihood ratio, given confidence level, and given posterior probability) will lead to almost the same range of values for the parameter.

Likelihood ratios are most easily studied as differences in log likelihoods. In this example the log likelihood is

$$l(\pi) = 4\log(\pi) + 6\log(1-\pi),$$

and the log of the likelihood ratio is $l(\pi) - l(\hat{\pi})$. This log-likelihood function is shown in part (b) of Figure 1; the cut point for the supported range is now $\log(0.1465) = -1.921$ and the shape of the log-likelihood ratio curve is quadratic rather than a bell. The shape can be further explored by examining the gradient at each value of the parameter, given by

$$l'(\pi) = \frac{4}{\pi} - \frac{6}{(1-\pi)}.$$

This is called the *score* function and it is usually written as $u(\pi)$. The graph of the score function is shown in part (c) of Figure 1; note that the score is zero when π takes its most likely value of 0.4.

Some General Definitions

Let the data consist of observations $x_1, x_2, \ldots, x_N$, with probability model $f(x;\theta)$, which depends on a parameter θ. When there are only a limited number of possible values for x the function $f(x;\theta)$ specifies the probability of each outcome, and when there are infinitely many outcomes $f(x;\theta)$ specifies the probability density. The log likelihood for θ is

$$l(\theta) = \sum_{i=1}^{N} \log f(x_i;\theta),$$

and the score function is

$$u(\theta) = l'(\theta) = \sum_{i=1}^{N} \frac{f'(x_i;\theta)}{f(x_i;\theta)}.$$

The most likely value of θ is $\hat{\theta}$, satisfying $u(\hat{\theta}) = 0$.

In the neighborhood of $\theta = \hat{\theta}$ the score function is approximately linear [see part (c) of Figure 1], and using Taylor's expansion

$$u(\theta) \approx u(\hat{\theta}) + (\theta - \hat{\theta})u'(\hat{\theta}).$$

The quantity $u'(\hat{\theta}) = l''(\hat{\theta})$ is negative, and its numerical value, namely $-l''(\hat{\theta})$, is called the observed **information** and is referred to as $j(\hat{\theta})$, or j for short. Since $u(\hat{\theta}) = 0$ the linear approximation can be written as

$$u(\theta) \approx -j(\theta - \hat{\theta}).$$

In part (c) of Figure 1 j is the numerical value of the gradient of the score function at $\pi = 0.4$. The steeper this gradient the more precise $\hat{\theta}$ is as an estimate of θ.

When considering the frequency properties of $\hat{\theta}$ as an estimate of θ it is best to write the function $l(\theta)$ as $l(\theta;x)$, stressing that it is a function of both the parameter values and the data. In a strictly likelihood approach the data are fixed and θ varies, which is why we write the function as $l(\theta)$. In the frequency approach it is x (the data) which varies and θ which is fixed, so that $l(\theta;x)$ is a **random variable**. The value of θ should be thought of as fixed at its true value, i.e. the value which gave rise to the data. The score $u(\theta;x)$ is now also a random variable, and is particularly important in frequency theory because its mean is zero and its variance can be calculated from the log likelihood. In fact

$$\mathrm{var}(u) = -\mathrm{E}[l''(\theta;x)].$$

The quantity $-\mathrm{E}\{l''(\theta; x)\}$ is called the *expected* or *Fisher information* and referred to as $i(\theta)$. When evaluated at $\theta = \hat{\theta}$ the expected information usually reduces to the same thing as the observed information, i.e. $i(\hat{\theta}) = j(\hat{\theta})$. Using the approximation $u(\theta) \approx -j(\theta - \hat{\theta})$, it follows that

$$\hat{\theta} - \theta \approx j^{-1}u(\theta),$$

and

$$\mathrm{var}(\hat{\theta}) \approx j^{-1}\mathrm{var}(u)j^{-1}.$$

Since $\mathrm{var}(u) \approx j$, the right-hand side of this equation becomes $j^{-1}jj^{-1} = j^{-1}$, so the variance of $\hat{\theta}$ in repeated samples is approximately j^{-1}, the inverse of the observed information.

The expression $\mathrm{var}(\hat{\theta}) \approx j^{-1}\mathrm{var}(u)j^{-1}$ is called the *information sandwich*. There are situations when it is unwise to assume that $\mathrm{var}(u) = j$, and better to replace $\mathrm{var}(u)$ by an empirically based estimate, using the individual values $u(\hat{\theta}; x_i)$. In this case the sandwich does not reduce to j^{-1} but provides instead a more robust estimate of the variance of $\hat{\theta}$.

When combining data from several sources, about the same parameter, the total log likelihood is obtained by adding the log likelihoods from the different sources. Since the score is the first derivative of the log likelihood, the total score is also found by adding the scores from the different sources, and since the information is the second derivative of the log likelihood it too is found by adding over sources. This additive property of the log likelihood and its derivatives makes the combining of data from different sources straightforward.

Approximate Log Likelihoods

The function $l(\theta)$ can be expanded around $\theta = \hat{\theta}$ using Taylor's expansion:

$$l(\theta) \approx l(\hat{\theta}) + (\theta - \hat{\theta})l'(\hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta}).$$

Since $l'(\hat{\theta}) = 0$ and $j = -l''(\hat{\theta})$, it follows that

$$l(\theta) - l(\hat{\theta}) \approx -\tfrac{1}{2}j(\theta - \hat{\theta})^2.$$

This may also be written as

$$l(\theta) - l(\hat{\theta}) \approx -\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{S}\right)^2,$$

where $S^2 = j^{-1}$ is the variance of $\hat{\theta}$. The fact that the log likelihood is approximately quadratic shows that the frequency distribution of $\hat{\theta}$ is approximately normal. A 95% confidence interval for θ is therefore given by $\hat{\theta} \pm 1.960S$. Alternatively, solving

$$-\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{S}\right)^2 = \log(0.1465) = -1.921$$

for θ leads to the same expression.

Two or More Parameters

Extending the results from one parameter to two or more parameters is largely a question of notation. For simplicity we concentrate on two parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The log likelihood $l(\theta)$ is now a function of both parameters, and there are two score functions $\mathbf{u} = (u_1, u_2)$, where u_1 is the derivative of $l(\theta_1, \theta_2)$ with respect to θ_1, and u_2 is the derivative of $l(\theta_1, \theta_2)$ with respect to θ_2. Similarly, there are two most likely values $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$ which together maximize the value of $l(\theta)$. The observed information becomes

$$j_{11}(\hat{\boldsymbol{\theta}}) = \frac{-\partial^2 l(\boldsymbol{\theta})}{\partial\theta_1^2},$$

$$j_{22}(\hat{\boldsymbol{\theta}}) = \frac{-\partial^2 l(\boldsymbol{\theta})}{\partial\theta_2^2},$$

$$j_{12}(\hat{\boldsymbol{\theta}}) = j_{21}(\hat{\boldsymbol{\theta}}) = \frac{-\partial^2 l(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2},$$

where all the derivatives are evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. These quantities are often written as a 2×2 information matrix $\mathbf{j}(\hat{\boldsymbol{\theta}})$ with elements $j_{rs}(\hat{\boldsymbol{\theta}})$, where $r = 1, 2$ and $s = 1, 2$. Similarly, the expected information becomes a 2×2 matrix $\mathbf{i}(\boldsymbol{\theta})$ with elements

$$i_{rs}(\theta) = -\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_r\partial\theta_s}\right].$$

All the results for one parameter extend in a fairly straightforward way to two or more parameters. In particular, the distribution of $\mathbf{u} = (u_1, u_2)$ has zero mean (0, 0), and covariance matrix equal to $\mathbf{i}(\boldsymbol{\theta})$, the expected information matrix. When evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ this becomes equal to $\mathbf{j}(\hat{\boldsymbol{\theta}})$, the observed

information matrix. The linear approximation to the score functions becomes

$$u_1 \approx -j_{11}(\theta_1 - \hat{\theta}_1) - j_{12}(\theta_2 - \hat{\theta}_2),$$
$$u_2 \approx -j_{12}(\theta_1 - \hat{\theta}_1) - j_{22}(\theta_2 - \hat{\theta}_2),$$

which may be written in matrix terms as

$$\mathbf{u} \approx -\mathbf{j}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{j}^{-1}\mathbf{u}.$$

The mean of the distribution of $\hat{\boldsymbol{\theta}}$ is approximately (0, 0) and the covariance matrix is approximately

$$\mathbf{j}^{-1}\mathrm{var}(\mathbf{u})\mathbf{j}^{-1},$$

which reduces to $\mathbf{j}^{-1}$ when var($\mathbf{u}$) is replaced by $\mathbf{j}$.

Finally, the quadratic approximation to the log likelihood in the neighborhood of $(\hat{\theta}_1, \hat{\theta}_2)$ is

$$l(\theta_1, \theta_2) - l(\hat{\theta}_1, \hat{\theta}_2) \approx -\tfrac{1}{2} j_{11}(\theta_1 - \hat{\theta}_1)^2 - \tfrac{1}{2} j_{22}(\theta_2 - \hat{\theta}_2)^2 - j_{12}(\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2),$$

which shows that the joint distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ is approximately bivariate normal with mean (0, 0) and covariance matrix $\mathbf{j}^{-1}$.

Nuisance Parameters

A supported *region* for (θ_1, θ_2) can be found by solving $l(\theta_1, \theta_2) = -1.921$, but in most practical applications one of the two parameters (say θ_1) is of interest and the other is a *nuisance*; so one wants a supported range for θ_1. It is straightforward to find a supported range for θ_1 for a given value θ_2^0 for θ_2, by solving $l(\theta_1, \theta_2^0) = -1.921$ for θ_1, but the answer will in general depend on θ_2^0. Only rarely will the supported range be independent of the value chosen for θ_2^0.

There are two possible ways of obtaining a supported range for θ_1 which is not dependent on choosing a particular value of θ_2. The first way is to find some aspect of the data which, when held fixed, leads to a *conditional* log likelihood which depends only on θ_1. Provided the aspects of the data which are held fixed are uninformative about θ_1, no information is lost by using the conditional log likelihood.

The second way is to replace the nuisance parameter θ_2 by its most likely value given θ_1, that is by $\hat{\theta}_2(\theta_1)$. The resulting log likelihood, $l_p(\theta_1) = l(\theta_1, \hat{\theta}_2(\theta_1))$, is called the *profile* log likelihood for θ_1, and can be used to find a confidence interval for θ_1 by solving $l_p(\theta_1) = -1.921$, as before. The idea of profile log likelihood extends to more than one nuisance parameter, but should not be used when there are many nuisance parameters to be eliminated, but not very much data. This is because each $\hat{\theta}_2(\theta_1)$ is too poorly estimated for the resulting profile log likelihood to be useful. A well-known example where this happens is a matched case–control study where there is a nuisance parameter for each new matched set. In this situation it is necessary to use a conditional log likelihood, and indeed this is generally the best thing to do provided one is available. Unfortunately there are many situations where it is not possible to find a conditional likelihood which depends only on the parameter of interest.

A quadratic approximation to the **profile likelihood** for θ_1 is found by starting from the quadratic approximation to $l(\theta_1, \theta_2)$, in the neighborhood of $(\hat{\theta}_1, \hat{\theta}_2)$, and then obtaining the profile likelihood for θ_1. This gives a quadratic approximation to $l_p(\theta_1)$, the profile likelihood for θ_1, of the form

$$l_p(\theta_1) \approx -\frac{1}{2}\left(\frac{\theta_1 - \hat{\theta}_1}{S}\right)^2,$$

where S is given by

$$S^2 = \frac{j_{22}}{j_{11}j_{22} - j_{12}^2}.$$

This is the first diagonal element in the inverse of $\boldsymbol{j}$, the observed information matrix, so the quadratic approximation to the profile likelihood for θ_1 coincides with the normal approximation to $\hat{\theta}_1$.

Hypothesis Testing

From a strictly likelihood point of view, support for a specific *null* value of a parameter, say θ^0, is measured by the likelihood ratio for this value in the same way as for any other value of θ. The likelihood ratio is a measure of how different θ^0 is from $\hat{\theta}$. Cut-points on the likelihood ratio scale are a matter of convention; some useful ones are shown in Table 1.

From a frequency point of view we measure how far θ^0 is from $\hat{\theta}$ in terms of how often the value of some statistic in repeated samples exceeds the value observed. To do this requires a statistic whose distribution when $\theta = \theta^0$ is known, and a natural one

Table 1

Likelihood ratio	Evidence against the null value
>0.25	None
0.15–0.25	Slight
0.05–0.15	Strong
<0.05	Very strong

to choose is the score $u(\theta^0; x)$. When the true value of θ is θ^0, the score has mean zero and variance $i(\theta^0)$, so the distribution of

$$z = \frac{u(\theta^0) - 0}{[i(\theta^0)]^{1/2}}$$

is approximately N(0, 1), a normal distribution with unit variance, and the probability of observing a value greater than $|z|$ is obtained by looking z^2 up in a χ^2 distribution with one degree of freedom (df). This probability is called the ***P* value**, and the test is called a *score test*.

Another candidate for the choice of statistic is $\hat{\theta}$, the most likely value of θ. When the true value of θ is θ^0, this statistic has an approximately normal distribution with mean θ^0 and variance j^{-1}. The probability of observing a value greater than $|z|$, where z is now

$$z = \frac{\hat{\theta} - \theta_0}{j^{-1/2}},$$

is obtained by looking z^2 up in a χ^2 distribution with one df. The test is now called a *Wald test*.

The last and generally the best statistic which is used is the log-likelihood ratio itself. Provided the log-likelihood curve is reasonably close to a quadratic shape, the distribution of

$$d = 2[l(\hat{\theta}) - l(\theta^0)]$$

is approximately χ^2 with one df. The probability of observing a value of d which is greater than the one actually observed is found by looking up d in tables of the **chi-square distribution** on 1 df. The test is now called the (log) **likelihood ratio test**.

All three tests extend to null hypotheses in which several parameters take their null values; for example the distribution of

$$2[l(\hat{\theta}_1, \hat{\theta}_2) - l(\theta_1^0, \theta_2^0)]$$

is approximately χ^2 on two df.

Further Reading

The concept of likelihood was introduced to statistics by Fisher in 1925, but the first book to discuss statistical inference from an exclusively likelihood point of view is Edwards [4]. More recent accounts which stress the central role of likelihood in statistical inference are given by Lindsey [5] at an elementary level, Azzalini [1] at an intermediate level, and Barndorff-Nielson & Cox [2] at an advanced level. An elementary account of the use of likelihood in the context of epidemiology is given by Clayton & Hills [3].

References

[1] Azzalini, A. (1996). *Statistical Inference Based on the Likelihood*. Chapman & Hall, London.

[2] Barndorff-Nielson, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[3] Clayton, D.G. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

[4] Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, London.

[5] Lindsey, J.K. (1995). *Introductory Statistics: The Modeling Approach*. Oxford University Press, Oxford.

(*See also* **Foundations of Probability**; **Hypothesis Testing**)

MICHAEL HILLS

Likert Scale

A Likert scale, or summated rating scale, is computed by summing responses over several items hypothesized to measure the same latent variable or construct [1]. Likert scales are often used to measure opinions, beliefs, or attitudes regarding a particular underlying construct. Items that comprise Likert scales are often measured on Likert response formats, which represent degrees of endorsement of those items [2]. For example, consider the following item measured on a five-point Likert response format:

Item:	Diet is an important part of a healthy lifestyle				
Response options:	Strongly agree	Agree	Neither	Dis-agree	Strongly disagree
	1	2	3	4	5

Items measured on Likert response formats are generally presented as declarative statements. For respondents to discriminate among response options, it is recommended that the items be presented as strong declarative statements.

Likert response formats may have three, four, five or more response options. An example of a six-point Likert response format which could have been used for the sample item above is: Strongly agree, Moderately Agree, Mildly agree, Mildly Disagree, Moderately disagree, and Strongly disagree. The five-point Likert response format is widely used. The number of response options for a given item depend upon the item being measured and subjects' abilities to discriminate between response options. Investigators should try to provide respondents with response options that are approximately equally spaced across the continuum of endorsement.

Many applications involve the measurement of a single underlying construct using multiple items, since in many cases single items are not adequate to measure the construct with sufficient precision. A Likert scale is constructed by summing or averaging responses over the set of all items to produce an overall score. In constructing the Likert scale, usually each item is equally weighted. If an investigation involves k such items, each measured on the same r-point Likert response format (responses coded as $1, 2, \ldots, r$, with higher scores reflecting more endorsement of each item), then the theoretic range of the Likert scale is k to kr.

Assumptions underlying the construction of multiple-item Likert scales are that each item is linearly related to the overall scale score, and that each item comprising the scale has approximately the same distribution (e.g. similar means and standard deviations). As an aside, when investigators present a set of items related to a single construct to respondents, some of the items should be reverse coded so as to reduce the likelihood that respondents consistently select the same response (e.g. strongly agree) for each item.

There are a number of techniques used to evaluate the **reliability** and validity (*see* **Validation Study**) of multiple-item Likert scales. These include, for example, the internal consistency reliability of the Likert scale, which is generally assessed using the **Cronbach's alpha** coefficient, and construct validity, which is assessed through **factor analysis**.

References

[1] DeVellis, R.F. (1991). *Scale Development: Theory and Applications*, Sage, Newbury Park.

[2] Stewart, A.L. & Ware, J.E., Jr (1992). *Measuring Functioning and Well-Being: the Medical Outcomes Study Approach*. Duke University Press, Durham.

(*See also* **Principal Components Analysis**; **Psychometrics, Overview**)

KIMBERLY A. DUKES

Limit Theorems

The earliest results in mathematical probability (dating from 1654) involved computations for finite sample spaces having equally likely outcomes, and thus could be regarded merely a branch of elementary combinatorics. The subject took a major step forward, however, both in the depth of its results and the sophistication of its methods after the discovery of its first limit theorems: James Bernoulli's **law of large numbers** (*c*. 1685, published posthumously in his *Ars conjectandi* of 1713) and Abraham De Moivre's **central limit theorem** [5] for sequences of dichotomous **binary** trials. These results extracted order from chaos by demonstrating that random phenomenon in the small (a limited number of observations) can exhibit regularities and deterministic behavior in the large.

Such results were often motivated by, and provided a theoretical basis for, the process of statistical **estimation**; in modern terminology, the law of large numbers amounts to nothing other than a statement that the sample mean is a **consistent** estimator of the population mean. If, for example, S_n denotes the number of successes in n dichotomous trials having probability of success p, then Bernoulli proved that $\lim_{n\to\infty} S_n/n = p$; using Stirling's approximation (including correction terms) to estimate the individual terms in the **binomial distribution** and then summing, De Moivre dramatically refined Bernoulli's result to discover the remarkable fact that

$$\lim_{n\to\infty} \Pr\left[a \le \frac{S_n - np}{[np(1-p)]^{1/2}} \le b\right] = \frac{1}{(2\pi)^{1/2}} \int_a^b \exp\left(-\frac{1}{2}x^2\right) \mathrm{d}x.$$

During the nineteenth and twentieth centuries, this result was extended far beyond the simple coin-tossing setup considered by De Moivre, important contributions being made by the French school of **Laplace** and **Poisson**, the Russian school of Chebyshev, Markov, Liapunov, Bemstein, Khinchin, and **Kolmogorov**, and the varied contributions of von Mises, Cantelli, Lindeberg, Lévy, and Feller in the period between the two world wars. Fueling these advances were the use of increasingly sophisticated methods such as the introduction of **characteristic functions** (by Laplace) and the **method of moments** (by Markov). The introduction of measure theory by Lebesgue and its use in the axiomatization of mathematical probability by Kolmogorov (*see* **Probability Theory**) led in turn to a sharp distinction between different forms of limit theorem, corresponding to different concepts of convergence: most commonly, **convergence in distribution**, probability, almost sure, and in L^p.

It is useful to regard a sequence of random variables as a single function $X(n, \omega)$, n being an integer and ω an element of a sample space. If one first fixes n, the result is a **random variable** $X_n(\cdot)$ (a function on the sample space), and one can then investigate the behavior of the distribution of X_n as $n \to \infty$; limiting behavior in this case corresponds to convergence in distribution. If, on the other hand, one first fixes ω, the result is a sequence $\omega(\cdot)$ (a function on the set of integers), and one can investigate the behavior of this *sample path* for typical values of ω; this corresponds to the case of almost sure convergence provided that the sequence converges except on a set of ω having probability 0.

The two behaviors can be very different. If, for example, the sequence represents an aperiodic finite **Markov chain**, then the distribution of X_n converges to the stationary distribution of the chain, but the sample paths $\omega(\cdot)$ cycle endlessly among the finite states of the chain; thus, in one sense, the chain exhibits ordered behavior over time, but in another it remains chaotic. (This phenomenon was first pointed out by Paul Ehrenfest, who introduced his celebrated urn model to illustrate that the Zermelo *recurrence paradox* does not contradict Boltzmann's demonstration that in statistical mechanics the entropy of a system, properly understood, is an increasing function of time. One can, in fact, prove some limit theorems in probability by first showing that an associated entropy function for a sequence of random variables increases with time.)

Thus, the law of large numbers has two versions, weak and strong, corresponding to convergence in probability and almost sure convergence. Just as the central limit theorem can be regarded as a refinement of the weak law of large numbers, the *law of the iterated logarithm* may be regarded as a refinement of the strong law of large numbers (almost sure convergence of the sequence of sample means to the population mean). This result (due in increasing generality to Khinchin, Kolmogorov, and Hartmann–Wintner),

one of the three pearls of the classical limit theorems, states that the sample path behavior for a sum of independent and identically distributed random variables, properly normalized, is at once both simple and unexpected: if $X_1, X_2, X_3, \ldots$ are independent and identically distributed random variables, such that $\mathrm{E}[X_k] = 0$, $\mathrm{var}[X_k] = 1$ and $S_n = X_1 + \cdots + X_n$, then

$$\Pr\left[\limsup_{n\to\infty} \frac{S_n}{(2n \log\log n)^{1/2}} = 1\right] = 1.$$

(It is simple to deduce from this that the set of limit points for the normalized sequence is almost surely the closed interval $[-1, 1]$.) Other such "zero–one" laws include the celebrated Borel–Cantelli lemmas used to prove the strong law.

This classical theory was subsequently extended to the study of sums and triangular arrays of sequences of random variables not having two **moments**, and using forms of normalization other than the mean and standard deviation. The stable and infinitely divisible distributions then arise as possible limiting distributions; the entire edifice is summarized in spare and elegant fashion in the classic and beautiful book of Gnedenko & Kolmogorov [8]. Lamperti [9] provides an attractive and accessible account of many of the key features of this classical theory.

Other generalizations of the classical theorems include the Birkhoff *ergodic theorem* and the *martingale convergence theorem*. The ergodic theorem (a direct descendant of Ehrenfest's attempts to explain and justify Boltzmann's theories) extended convergence of sample means from the domain of independence to that of stationary sequences (sequences invariant under shift); the martingale convergence theorem extracts a key property of centered sums (that one's expected future gain in a sequence of fair games is the same as one's present fortune) to derive other limiting forms of behavior. In the hands of Doob [6] and his successors, the martingale concept and its use became a fundamental and pervasive aspect of modern probability theory.

Two important modern advances in limit theorems after this classical period were the concepts of invariance principles and functional limit theorems. *Invariance principles* establish that if one sequence in a class of possible sequences converges to a limit, then all sequences in that class must converge to the same limit. Thus, one proof of the central limit theorem demonstrates that if the normalized sum of a sequence of independent and identically distributed random variables having a second moment (such as Bernoulli trials) converges to the standard normal distribution, then all sequences in this class must also converge to this limit (see [4]). *Functional central limit theorems* generalize the central limit theorem by considering functionals of sample paths (such as the maximum) and determining their limiting distribution by computing the distribution of that functional applied to the limiting distribution of sample paths: **Brownian motion**. The abstraction of these two theories led to the creation of the subject of *weak convergence* (see [2]).

There is a simple hierarchy that applies to the most common modes of convergence for a sequence of random variables: convergence almost surely $\Rightarrow$ convergence in probability $\Rightarrow$ convergence in distribution. These implications admit of limited reversal: if a sequence of random variables converges in probability, then every subsequence contains a further subsequence converging almost surely; if a sequence of random variables converges in distribution, then one can find a sequence of random variables having the same one-dimensional distributions that converges almost surely (*Skorokhod's theorem*).

This last result is related to a distinctively modern element in the proof of limit theorems: the use of *coupling methods* to construct versions of the random elements in question that live on the same probability space. Due originally to the gifted French probabilist Doeblin (who died tragically at the beginning of the Second World War), the method only gained currency decades later. Its use provides perhaps the most elegant derivation of the limiting behavior for countable Markov chains (see, for example, [3]).

The increasing use of the **bootstrap** [7], **Markov chain Monte Carlo**, simulated annealing, and other **computer-intensive methods** only recently possible, points to an emerging post-modern period of limit theorems in mathematical probability, the outlines of which are only just beginning to be clear. For a number of interesting applications of modern limit theorems to the biological sciences, see Waterman [10]. Such applications include methods as diverse as the Aldous [1] Poisson clumping heuristic, the Erdös–Renyi law of large numbers, and the theory of large deviations.

References

[1] Aldous, D. (1989). *Poisson Approximations via the Poisson Clumping Heuristic*. Springer-Verlag, New York.

[2] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[3] Billingsley, P. (1995). *Probability and Measure*. Wiley, New York.

[4] Breiman, L. (1968). *Probability*. Addison-Wesley, Reading.

[5] De Moivre, A. (1738). *The Doctrine of Chances*, 3rd Ed., 1756. Reprinted Chelsea, New York, 1967.

[6] Doob, J.L. (1953). *Stochastic Processes*. Wiley, New York.

[7] Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[8] Gnedenko, B.V. & Kolmogorov, A.N. (1949). *Limit Distributions for Sums of Independent Random Variables*, 2nd English Ed.; translated, annotated, and revised by K.L. Chung. Addison-Wesley, Reading, 1968.

[9] Lamperti, J. (1996). *Probability*, 2nd Ed. Wiley, New York.

[10] Waterman, M.S. (1995). *Introduction to Computational Biology: Maps, Sequences, and Genomes*. Chapman & Hall, New York.

(*See also* **Large-sample Theory**)

S.L. ZABELL

Linder, Forrest E.

Born: November 21, 1906, in Waltham, Massachusetts.
Died: August 18, 1988, in Washington, DC.

Forrest E. Linder devoted a lifetime of service to the worldwide development of vital and health statistics (*see* **Vital Statistics, Overview**). Although born in Massachusetts, he grew up in Iowa where, at Iowa State University, he received a doctorate in mathematics and statistics. Following a brief assignment with a foundation in Massachusetts, he moved to Washington, DC, in 1935, where he served until 1944 as a statistician with the US Bureau of the Census. In 1939 he spent a year in Montevideo assisting the Uruguayan government in establishing a vital statistics system. This was the beginning of Linder's interest in international vital statistics. He subsequently established an international vital statistics program at the Bureau of the Census, which provided training in civil registration and vital statistics to foreign national officials as well as an on-site consulting program for Latin American countries.

During World War II he served as Assistant Chief of the Medical Statistics Division of the US Navy, where he was responsible for establishing a morbidity reporting system, including the necessary data-processing support to provide current estimates of morbidity and mortality for the personnel of a greatly expanded wartime navy. After the war, Linder joined the United Nations where he became the first Chief of the Demographic and Social Statistics Branch of the UN Statistical Office. There, his contributions to the improvement of the **demographic** statistics of developing countries included projects he conceived and inspired, such as the world **censuses** of population and housing of 1950 and 1960 and a series of regional seminars on vital and health statistics. He was instrumental in designing and producing the first *United Nations Demographic Yearbook, Principles for a Vital Statistics System*, and the *Handbook of Vital Statistics Methods*.

In 1957, Linder returned to the federal civil service in Washington, DC, to become the director of the newly established National Health Survey, a program of the US Public Health Service. When the National Office of Vital Statistics was merged with the National Health Survey in 1960 to form the **National Center for Health Statistics (NCHS)**, Linder was named as its first director, a post he held until his retirement in 1967.

Upon leaving the Public Health Service, he joined the faculty at the University of North Carolina where he was both professor of biostatistics and the first director of the International Program of Laboratories for Population Statistics, popularly known as POPLAB [3]. While at POPLAB, he began to lay the groundwork for an international organization that would address the professional interests and needs of national officials responsible for civil registration. Recognizing that these officials were a diverse group, often working in isolation and without an international focus for information exchange and guidance, he founded the International Institute for Vital Registration in 1974, an organization to encourage and promote the improvement of civil registration throughout the world with special attention to lesser developed countries. When he retired from the university he served as President and Executive Director of the Institute until his death in 1988.

During a career that spanned half a century, Forrest Linder had an impressive list of important technical publications; these covered a wide range of topics in demography and health, such as fertility measurement, morbidity and mortality analysis, and survey methodology applied to public health (*see* **Surveys, Health and Morbidity**) [1, 2, 4–7]. He served on international and national advisory committees, such as the World Fertility Survey Steering Committee, US Agency for International Development Research Advisory Committee, **World Health Organization** Expert Committees on Health Statistics, and the US National Committee on Vital and Health Statistics; and he received the Distinguished Service Award from the US Department of Health, Education and Welfare and the Bronfman Prize from the **American Public Health Association**.

In each of his professional endeavors Forrest Linder exhibited a strong pioneering spirit at the highest technical level. Throughout his career he continued his abiding interest in the improvement of national and international vital and health statistics and he had an unfailing belief in the importance of his undertakings. His work has had a positive influence not only on civil registration and vital statistics programs

throughout the world, but also on the many statisticians, demographers and others with whom he came in contact.

References

[1] Linder, F.E. (1965). National health interview surveys, in *Trends in the Study of Morbidity and Mortality*, Public Health Papers, No. 27. World Health Organization, Geneva.

[2] Linder, F.E. (1967). Sources of data on health in the United States, in *Preventive Medicine*, D. Clark & B. MacMahon, eds. Little, Brown & Company, Boston, pp. 55–56.

[3] Linder, F.E. (1971). *The Concept and the Program of the Laboratories for Population Statistics*, International Program of Laboratories for Population Statistics, Scientific Series No. 1. University of North Carolina, Chapel Hill.

[4] Linder, F.E. (1971). Fertility and family planning in relation to public health, *Milbank Memorial Fund Quarterly*, Vol. XLIX, No. 4, Part 2, New York.

[5] Linder, F.E. (1974). The dual-record system of collecting demographic data, *United Nations Publication ST/ECLA/Conf. 47/L.3*. United Nations, New York.

[6] Linder, F.E. & Grove, R. (1965). Techniques of Vital Statistics, Reprint of Chapters I-IV, *Vital Statistics Rates in the United States, 1900–1940*. National Center for Health Statistics, Washington.

[7] Linder, F.E. & Lingner, J.W. (1975). *Systems of Demographic Measurement: General Evaluation – The Measurement Problem*, International Program of Laboratories for Population Statistics, Scientific Series No. 22. University of North Carolina, Chapel Hill.

ROBERT A. ISRAEL

Lindley's Paradox

A sharp **null hypothesis** may be strongly rejected by a standard sampling theory test of significance (*see* **Hypothesis Testing**) and yet be awarded high odds by a **Bayesian** analysis based on a small **prior probability** for the null hypothesis and a diffuse distribution of one's remaining probability over the **alternative hypothesis**. This disagreement between sampling theory and Bayesian methods was first studied by Jeffreys [2], and it was first called a paradox by Lindley [3].

The paradox can be exhibited in the simple case where we are testing $\theta = 0$ using a single observation Y from a **normal distribution** with variance one and mean θ. If we observe a large value y for Y ($y = 3$, for example), then standard sampling theory allows us to reject confidently the null hypothesis. But the Bayesian approach advocated by Jeffreys can give quite a different result. Jeffreys advised that we assign a nonzero prior probability π_0 to the null hypothesis and distribute the rest of our probability over the real line according to a fairly flat probability density, $\pi_1(\theta)$. If the range of possible values for θ is very wide, then the set of values within a few units of y will be very unlikely under $\pi_1(\theta)$, and consequently the overall **likelihood** of the alternative hypothesis,

$$L_1 = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left[-\frac{1}{2}(y-\theta)^2\right] \pi_1(\theta)\, \mathrm{d}\theta,$$

will be very small. It may even be so much smaller than the likelihood of the null hypothesis,

$$L_0 = \frac{1}{(2\pi)^{1/2}} \exp\left(\frac{-y^2}{2}\right),$$

that the odds in favor of the null hypothesis,

$$\frac{\Pr(\theta = 0 | Y = y)}{\Pr(\theta \neq 0 | Y = y)} = \frac{\pi_0}{1-\pi_0} \frac{L_0}{L_1}, \qquad (1)$$

are substantial.

We can think of (1) as a way of balancing arguments for and against the null hypothesis. *Against* the null hypothesis is its small initial probability (small π_0) and the unlikeliness of the observation under the null hypothesis (small L_0). *For* the null hypothesis is the unlikeliness of alternative values of θ near y [small $\pi_1(\theta)$, leading to small L_1]. There is no strong constraint between the arguments for and against. No matter how small π_0 and L_0 are, a sufficiently diffuse $\pi_1(\theta)$ can make L_1 small enough to counterbalance them.

If we are confident of the specified prior distribution – if, for example, we are working with a series of problems involving θs that are zero about π_0 of the time and distributed roughly according to $\pi_1(\theta)$ the rest of the time – then the Bayesian analysis is unassailable, and hence we must reject the standard sampling theory. An observation three standard deviations from the null hypothesis is not adequate to reject the null hypothesis if that observation is even more unlikely under the alternative hypothesis. This has led many authors to suggest that we make tests increasingly stringent as measurements become more precise relative to the range of possible values for what is being measured. We should, for example, lower the significance level (*see* **Level of a Test**) as the sample size grows. More sophisticated suggestions are made by Berger & Delampady [1].

However, if diffuseness of $\pi_1(\theta)$ reflects merely a wide uncertainty about θ rather than a positive prior confidence that values of θ near y are likely to occur, then the conflict seems to constitute a criticism of the Bayesian analysis. If we have no idea how θ arises, then our mere ignorance cannot justify a skepticism about values close to y so strong as to outweigh real evidence against the value of zero. This has motivated non-Bayesian approaches discussed by Shafer [4].

References

[1] Berger, J. & Delampady, M. (1987). Testing precise hypotheses (with discussion), *Statistical Science* **2**, 317–352.

[2] Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press, Oxford.

[3] Lindley, D.V. (1957). A statistical paradox, *Biometrika* **44**, 187–192.

[4] Shafer, G. (1982). Lindley's paradox (with discussion), *Journal of the American Statistical Association* **77**, 325–351.

GLENN SHAFER

Linear Mixed Effects Models for Longitudinal Data

Introduction

Linear mixed-effects (LME) models [9] have become a popular tool for analyzing **longitudinal data** that arise in areas as diverse as **clinical trials**, **epidemiology**, agriculture, economics, and geophysics. The increasing popularity of these models is explained by the flexibility they offer in modeling the within-subject **correlation** often present in longitudinal data, by the handling of both balanced and unbalanced data, and by the availability of reliable and efficient **software** for fitting them [16, 20].

As a motivating example to illustrate the key ideas behind LME models, we consider data from a longitudinal study on the heights of a sample of 26 boys from Oxford, England, described and analyzed in [8] and presented in Figure 1.

As typically occurs with longitudinal data, the **growth** curves show a similar pattern across subjects (a linear trend, in this case), but important individual differences, both in intercept and in slope, are also observed.

Different approaches can, in principle, be used to model longitudinal data: (a) ignore the between-subject differences, concentrating on the estimation of the overall trend; (b) use a separate model for each subject, thus accounting for between-subject differences; and, (c) use a mixed-effects model. In the case of the Oxford data, the first approach consists of assuming the *population average* model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \tag{1}$$

to represent the height measurement y_{ij} on subject i at time x_{ij}. The within-subject errors ε_{ij} are assumed independently distributed as $\mathcal{N}(0, \sigma^2)$. The parameters of interest, the population intercept β_0, the population slope β_1, and the error variance σ^2, are estimated via ordinary **least squares** (OLS) [7]. The between-subject differences in growth pattern, which are not accounted for in the model, lead to incorrect **standard errors** for the OLS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and this is the main drawback of this approach.

The model associated with the second approach uses *subject-specific* coefficients β_{0i} and β_{1i} to accommodate differences between subjects

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \varepsilon_{ij}. \tag{2}$$

However, it fails to take into account the common growth pattern observed across subjects, requires a large number of estimates (53 in this case), and does not scale-up with the number of subjects. Population inferences require a second-stage analysis, in which the individual estimates are treated as data. This is fairly inefficient and can lead to poor population estimates, especially when the data are unbalanced.

The linear mixed-effects approach strikes a balance between the population average model (1) and the subject-specific model (2), being expressed as

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) x_{ij} + \varepsilon_{ij}. \tag{3}$$

It accommodates between-subject variation in growth pattern via **random effects** $\mathbf{b}_i = [b_{0i}, b_{1i}]'$, while capturing the population average behavior through **fixed effects** $\boldsymbol{\beta} = [\beta_0, \beta_1]'$. The random effects are assumed to be independently distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ vectors (*see* **Multivariate Normal Distribution**) and this common distribution ties together the observations from different subjects. A total of six parameters (the two fixed effects, the within-subject variance, and the three unique parameters in $\boldsymbol{\Psi}$) need to be estimated, irrespective of the number of subjects observed, so that the model is scalable in the number of subjects. The LME fit provides, in a single step, information about the population behavior as well as the between-subject variation in growth pattern.

In the next section, we describe the general linear mixed-effects model for Gaussian longitudinal data, its assumptions and estimation methods. Software for fitting the LME model is mentioned in the section "Software" and extensions to other mixed-effects models are briefly discussed in the section "Extensions".

Model Definition and Assumptions

The linear mixed-effects model for a Gaussian response measured longitudinally on a set of M subjects, proposed by [13], is a generalization of (3) that is expressed as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{4}$$

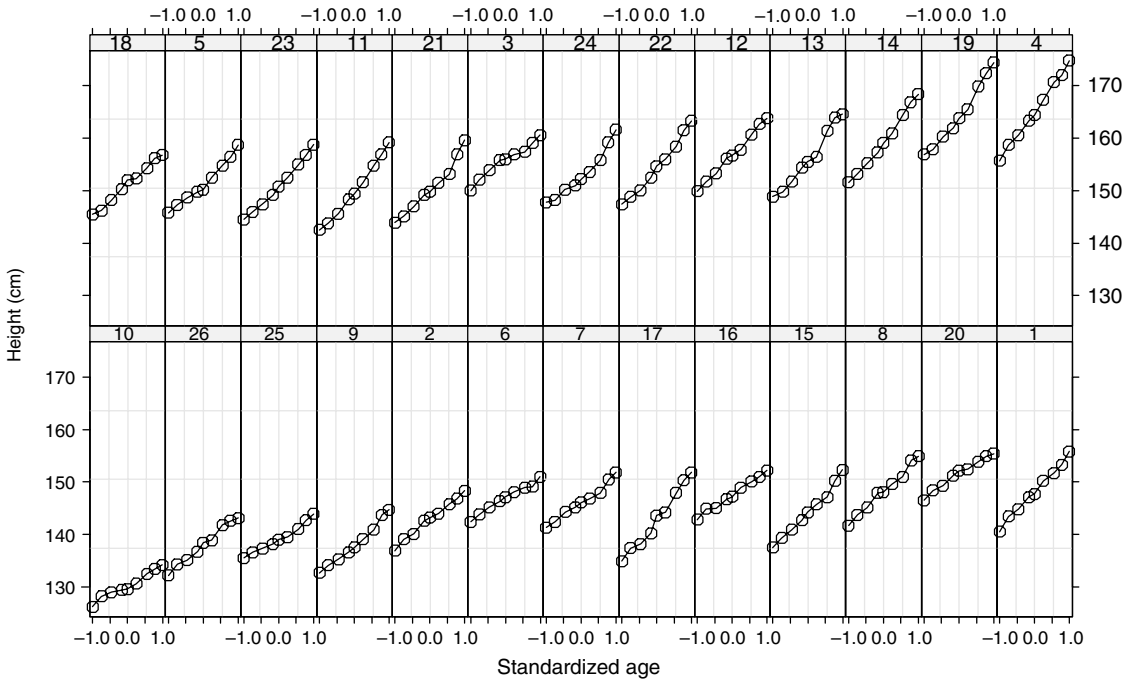

Figure 1 Heights of 26 boys from Oxford, England, each measured on nine occasions. The ages have been standardized to allow an easier comparison of the individual growth curves

where i is the subject index, $\mathbf{y}_i$ is an n_i-dimensional vector of observed responses, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known $n_i \times p$ and $n_i \times q$ regression **matrices** corresponding to the p-dimensional fixed-effects vector $\boldsymbol{\beta}$ and the q-dimensional random-effects vector respectively, and $\boldsymbol{\varepsilon}_i$ is an n_i-dimensional vector of within-subject errors. Note that the number of observations n_i, as well as the regression matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ are allowed to vary with subjects, so that unbalanced data are naturally handled. In the case of the Oxford data, the LME model formulation (4) would use $\mathbf{X}_i = \mathbf{Z}_i = [\mathbf{1} \ \mathbf{x}_i]$, with $\mathbf{1}$ representing a column vector of ones and $\mathbf{x}_i$ a column vector with the standardized times corresponding to subject i.

The $\mathbf{b}_i$ are assumed to be independent with distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ and the $\boldsymbol{\varepsilon}_i$ are assumed to be independent with distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_i)$, independent of the $\mathbf{b}_i$. The $\boldsymbol{\Psi}$ **covariance matrix** may be unstructured or structured – for example, diagonal [12], being parameterized by a set of parameters $\boldsymbol{\theta}$. The $\boldsymbol{\Lambda}_i$ matrices are typically assumed to depend on i only through their dimensions, being parameterized by a fixed, generally small, set of parameters $\boldsymbol{\lambda}$ – for example, an AR(1) structure [1] (*see* **ARMA and ARIMA Models**).

Even though the random effects are useful and intuitive quantities to represent between-subject differences in the coefficients, they are not observable in practice. Therefore, estimation and inference generally rely on the *marginal* distribution of the observed response vectors $\mathbf{y}_i$. Because of the linearity of the random effects in the LME model (4), the assumptions on the random effects and the within-group errors, and the properties of the multivariate normal distribution, it can be shown that the $\mathbf{y}_i$ are marginally distributed as independent $\mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$ random vectors, where the marginal covariance matrix is given by $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i' + \boldsymbol{\Lambda}_i$.

There are two ways in which the LME model (4) can account for within-subject correlation and heteroscedasticity (nonconstant variance) (*see* **Scedasticity**): through the random effects $\mathbf{b}_i$, and through the within-subject errors $\boldsymbol{\varepsilon}_i$. Because the random effects $\mathbf{b}_i$ are fixed by subject, and do not vary with time, the within-subject observations share the same random effects and are, therefore, correlated. This is represented by the $\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i'$ component of $\boldsymbol{\Sigma}_i$. Note, also, that the diagonal elements of $\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i'$ need not be constant, so that it can also accommodate

heteroscedasticity. The within-subject error contribution to the marginal covariance matrix is given directly by $\mathbf{\Lambda}_i$, which can be nondiagonal (correlation) and have different diagonal elements (heteroscedasticity). These two-model components may actually compete for explaining the marginal covariance structure of the response vectors and some care should be exercised when specifying their structure to avoid numerical problems in the **optimization** algorithm used to estimate the model parameters [20].

Estimation

Several methods of parameter **estimation** have been proposed for linear mixed-effects models. We concentrate here on two general methods: **maximum likelihood** (ML), and **restricted maximum likelihood** (REML). Descriptions and comparisons of the various estimation methods used for LME models can be found, for example, in [21] and [24]. For a **Bayesian** perspective, see [25].

The log-**likelihood** function corresponding to the LME model (4), based on the marginal distribution of the $\mathbf{y}_i$ (*see* **Marginal Likelihood**), is given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{y}) = -\frac{1}{2}\left\{N\log(2\pi) + \sum_{i=1}^{M}\left[\log|\mathbf{\Sigma}_i| + (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\,\mathbf{\Sigma}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right]\right\}, \tag{5}$$

where $\mathbf{y}$ denotes the entire response vector, and N the total number of observations. Conditional on $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, it is easy to show to that the maximum likelihood estimate (MLE) of the fixed effects is given by

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \left(\sum_{i=1}^{M}\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{M}\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{y}_i. \tag{6}$$

The MLEs of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ cannot be expressed in closed form, except in trivial cases, and numerical optimization of the loglikelihood function (5) must be employed. The MLE of $\boldsymbol{\beta}$ is then obtained by replacing $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ in (6) with their corresponding MLEs.

The most popular optimization methods for likelihood estimation in LME models are the **EM algorithm** [5] and Newton–Raphson or quasi-Newton methods [23] (*see* **Optimization and Nonlinear Equations**). A detailed discussion of numerical optimization in LME models, including efficient methods for calculating the loglikelihood (5) on the basis of orthogonal-triangular decompositions, is given in [15].

Maximum likelihood estimates of **variance components** tend to underestimate these parameters [10]. Restricted (or residual) maximum likelihood (REML) methods [10, 19] were developed to circumvent this problem. The REML loglikelihood is defined as the loglikelihood of a set of ordinary least-squares **residual** contrasts of the response vector $\mathbf{y}$ with respect to the fixed-effects regression matrix $\mathbf{X}$. By the definition of the LME model, this loglikelihood does not contain any information about the fixed effects and can be expressed as [10]

$$\ell_R(\boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{y}) = \ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{y}) - \tfrac{1}{2}\log\left|\sum_{i=1}^{M}\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{X}_i\right|. \tag{7}$$

REML estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are obtained via numerical optimization of (7) using similar algorithms as in the ML case. Although the REML loglikelihood does not have information on the fixed effects, REML estimates of $\boldsymbol{\beta}$ are obtained by plugging the REML estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ into (6). Table 1 presents the ML and REML estimates for the Oxford example, model (3). The random-effects parameters $\boldsymbol{\theta}$ are represented by the standard deviations σ_0 and σ_1 corresponding, respectively, to b_{0i} and b_{1i}, and their correlation coefficient ρ. The within-subject parameters $\boldsymbol{\lambda}$ include only the within-subject standard deviation σ.

In this example, because of the balanced structure of the data, the only differences between ML and REML estimation is in the standard deviations of the random effects, with REML giving larger estimates, as expected.

Table 1 ML and REML estimates for the Oxford data LME model

Method	β_0	β_1	σ_0	σ_1	ρ	σ
ML	149.37	6.53	7.92	1.65	0.64	0.66
REML	149.37	6.53	8.08	1.68	0.64	0.66

Even though the random effects are not regarded as parameters in the LME model, it is also of interest in practice to obtain predicted values for them. The best linear **unbiased predictors** (BLUPs) of the random effects, given $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, are the conditional means of the $\mathbf{b}_i$ given the responses $\mathbf{y}_i$:

$$\widehat{\mathbf{b}}_i = \widehat{\mathbf{b}}_i(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{\Psi}\mathbf{Z}_i'\boldsymbol{\Sigma}_i^{-1}\left[\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right]. \quad (8)$$

Estimated BLUPs are obtained, in practice, by replacing $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ in (8) with their corresponding (RE)ML estimates.

Inference

As a consequence of the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ not being expressible in closed-form, inference on the LME model parameters usually relies on approximate distributions for the (RE)ML estimates, derived from asymptotic results. Under certain regularity conditions, generally satisfied in practice, the MLEs in the LME model (4) are **consistent** and asymptotically normal [11, 18] (*see* **Large-sample Theory**). Furthermore, the MLEs $\widehat{\boldsymbol{\beta}}$ are asymptotically uncorrelated with the MLEs $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\lambda}}$. The approximate distributions for the MLEs in the LME model are

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\left(\boldsymbol{\beta},\ \left(\sum_{i=1}^{M}\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\right),$$

$$\begin{bmatrix}\widehat{\boldsymbol{\theta}}\\ \widehat{\boldsymbol{\lambda}}\end{bmatrix} \dot{\sim} \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{\lambda}\end{bmatrix},\ \mathbf{I}^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right), \quad (9)$$

where $\mathbf{I}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ denotes the **information matrix** [3] corresponding to $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, that is, minus the expected value of the second-order derivative of the loglikelihood function (5) with respect to $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ (which does not depend on $\boldsymbol{\beta}$).

The REML estimates in the LME model are also consistent and asymptotically normal [11], with the same approximate distributions as in (9), but with the information matrix $\mathbf{I}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ calculated using the REML loglikelihood ℓ_R defined in (7).

In practice, the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are replaced by their respective (RE)ML estimates in the approximate distributions (9). These approximate distributions are then used to produce **hypothesis tests** and **confidence intervals** in the model parameters.

Likelihood-ratio tests [14] are generally used to test hypothesis about $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ in *nested* models with the same fixed effects. Corrections on the number of **degrees of freedom** are needed in the case of tests involving boundary conditions [22]. Likelihood-ratio tests are less frequently used to compare nested models with different fixed-effects because (a) they cannot be used under REML estimation (the REML loglikelihoods of models with different $\mathbf{X}_i$ matrices are not comparable); and (b) they tend to produce tests that are too liberal, in the sense that the actual significance **levels** tend to be considerably higher than their nominal values. For these reasons, Wald tests (*see* **Chi-square Tests**) are preferred to test hypothesis on the fixed effects, with different approximations being used for the denominator degrees of freedom in the corresponding **Student's *t*-** and ***F***-tests [16].

Software

The availability of reliable and efficient software for fitting linear mixed-effects models in commercial packages is one of the main reasons for their increasingly widespread use.

The SAS system includes several procedures and macros for fitting mixed-effects models, with `PROC MIXED` [16] being solely devoted to the linear mixed-effects model. It implements both ML and REML estimation and allows separate specifications of the fixed-effects model ($\mathbf{X}_i$), the random-effects model ($\mathbf{Z}_i$ and $\boldsymbol{\Psi}$), and the within-subject error model ($\boldsymbol{\Lambda}_i$). Similar capabilities are available in the **S-PLUS** and **R** languages, implemented in the `lme` function, which is part of the more comprehensive `NLME` library for mixed-effects models [20]. Both `PROC MIXED` and `lme` have the advantage of being part of general purpose statistical packages, so they can be used in conjunction with other features in the system, such as graphics and programming language capabilities.

Stand-alone, commercial software for fitting linear mixed-effects models is also available, including MLwiN [26], HLM [27], Mplus [28], and AS Reml [29]. All of these include capabilities for fitting and analyzing linear mixed-effects models for longitudinal data, plus varying additional features and capabilities. See the corresponding URLs in the References section for more information on each software.

Extensions

The linear mixed-effects model described in the section "Model Definition and Assumptions" can be

extended in a variety of ways. The LME model (4) is intended for longitudinal data collected on subjects, characterizing a single level of grouping. **Multilevel** linear mixed-effects models [2, 8] handle the case of multiple nested levels of grouping, which often occurs in education and sociology, for example. The generalization of model (4) to the multilevel case is straightforward, with the same estimation methods and similar optimization algorithms being used for estimation [20].

Nonlinear mixed-effects (NLME) models [4, 20, 24] extend linear mixed-effects models by allowing the regression function to depend nonlinearly on fixed and random effects. Because of its greater flexibility, an NLME model is generally more interpretable and **parsimonious** than a competitor empirical LME model based, say, on a **polynomial** or **spline function**. The greater flexibility of NLME models does not come without cost, however. Because the random effects are allowed to enter the model nonlinearly, the marginal likelihood function, obtained by integrating the joint density of the response and the random effects with respect to the random effects, does not have a closed-form expression, as in the LME model. As a consequence, an approximate likelihood function needs to be used for the estimation of parameters, leading to more **computer-intensive** estimation algorithms and to less reliable inference results.

Generalized linear mixed-effects models (GLMMs) [6, 17] have been developed for grouped data with non-Gaussian response variables, like **binary** and count data. As with NLME models, the marginal loglikelihood of GLMMs generally does not have a closed-form expression because of model nonlinearity with respect to the random effects. Estimation methods use different approximations to the loglikelihood, resulting in more computationally intensive algorithms and less reliable estimation results than in LME models.

References

[1] Box, G.E.P., Jenkins, G.M. & Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd Ed. Holden-Day, San Francisco.

[2] Bryk, A. & Raudenbush, S. (1992). *Hierarchical Linear Models for Social and Behavioral Research*. Sage, Newbury Park.

[3] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[4] Davidian, M. & Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London.

[5] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37), *Journal of the Royal Statistical Society, Series. B* **39**, 1–22.

[6] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

[7] Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd Ed. Wiley, New York.

[8] Goldstein, H. (1987). *Multilevel Models in Education and Social Research*. Oxford University Press, Oxford.

[9] Hartley, H.O. & Rao, J.N.K. (1967). Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika* **54**, 93–108.

[10] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**, 320–340.

[11] Pinheiro, J.C. (1994). Topics in Mixed-Effects Models. PhD thesis, University of Wisconsin, Madison.

[12] Jennrich, R.I. & Schluchter, M.D. (1986). Unbalanced repeated measures models with structural covariance matrices, *Biometrics* **42**(4), 805–820.

[13] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[14] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. Wiley, New York.

[15] Lindstrom, M.J. & Bates, D.M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data (corr: 94v89 p1572), *Journal of the American Statistical Association* **83**, 1014–1022.

[16] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute Inc., Cary.

[17] McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.

[18] Miller, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance, *The Annals of Statistics* **5**, 746–762.

[19] Patterson, H.D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal, *Biometrika* **58**, 545–554.

[20] Pinheiro, J. & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

[21] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

[22] Stram, D.O. & Lee, J.W. (1994). Variance components testing in the longitudinal mixed-effects models, *Biometrics* **50**, 1171–1177.

[23] Thisted, R.A. (1988). *Elements of Statistical Computing*. Chapman & Hall, London.

[24] Vonesh, E.F. & Chinchilli, V.M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measures*. Marcel Dekker, New York.
[25] Wakefield, J.C., Smith, A.F.M., Racine-Poon, A. & Gelfand, A.E. (1994). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler, *Applied Statistics* **43**, 201–221.
[26] `http://multilevel.ioc.ac.uk/features`.
[27] `http://www.ssicentral.com/hlm/hlm.htm`.
[28] `http://www.statmodel.com/mplus`.
[29] `http://www.vsn-intl.com/ASReml`.

(*See also* **Segregation Analysis, Mixed Models**; **Random Coefficient Repeated Measures Model**)

JOSÉ C. PINHEIRO

Linear Programming

Linear programming (LP) is a decision model (*see* **Decision Theory**) that was developed early in the history of **operations research** and has wide applicability. It is a technique for finding optimal solutions, i.e. solutions to a decision problem that optimize some objective (maximize profit, minimize cost, etc.) subject to a set of constraints. In health care, LP has been applied both to management decision making and to decisions regarding clinical care. A few examples of LP are: (i) to improve breast cancer diagnosis on the basis of cell characteristics from a fine needle biopsy and to model the likelihood of recurrence in surgically treated patients [10]; (ii) to optimize scheduling nurses to meet coverage needs at the lowest cost [7]; (iii) to develop a model of costs and revenues based on patient diagnostic groups for strategic planning at a major university medical center [1]; (iv) to identify underutilized resources and inefficient production of services at the Department of Veterans Affairs medical centers [14]; (v) to develop a severity index for emergency medicine patients with cardiac problems [11]; (vi) to determine a treatment plan for radiation therapy that optimizes tumor exposure while reducing the exposure of healthy tissue [12]; and (vii) to compare alternative methods to develop a state rate-setting formula for nursing homes [2]. Greenberg [3–6] provides a tutorial with an overview of LP methodology and applications, while Hillier & Lieberman [8] is an excellent basic reference.

Although LP is the decision model most widely used by corporations, health care applications have been limited in the past by inaccessible software, unavailable data and low demand. Charge based fee-for-service clinical practice and cost-plus reimbursement for hospitals coupled with less competition in the past produced a low perceived need to optimize. The growth of managed care with its emphasis on global budgets and capitation for the care of populations is changing this situation; consequently, LP is likely to become more important in health care management and in **health services research** (*see* **Health Services Organization in the US**).

Model

An LP model has three main components: (i) a set of *decision variables* which represent quantities over which management has control; (ii) an *objective function*, defined on the decision variables, representing the quantity that the decision maker wishes to optimize; and (iii) a set of *constraints* representing the limitations imposed on the decision choices. The word *linear* refers to the form of the functions of the decision variables appearing in the objective function and in the constraints. The form of these functions is a summation of terms, each term being a single decision variable multiplied by a coefficient. Thus, linear programming, strictly defined, does not permit forms that have variables raised to powers or multiplied by other variables. There are techniques for nonlinear programming, but they involve different algorithms. The term *programming* refers to the iterative nature of solution techniques, not to computer programming.

Solving an LP problem involves finding a set of values to assign to the decision variables that will maximize (or minimize) the objective function without violating any of the constraints. Constraints may reflect limitations on resources, policy requirements, proportional relationships that must be maintained, or other requirements of the situation. There are three types of functional constraints: requirements – "greater than or equal to", limitations – "less than or equal to", and strict equality. Sign constraints are requirements that a variable be nonnegative.

There are three stages in LP analysis: formulation, solution, and interpretation. Formulation involves expressing the decision problem as an LP model in *standard form*. In standard form the constraints have the variables on the left-hand side of the operator and a constant on the right-side (known as RHS or right-hand-side quantities). Next, the "optimal solution" is found and its general properties are determined, including sensitivity to parameter variations (*see* **Sensitivity Analysis**) and shadow prices for all constraints. Then, interpretation involves translating the numbers produced by the solution technique into their meaning in the context of the decision to be made.

Solution Techniques

Once an LP model has been formulated, a solution is sought. Some LP models do not have solutions. Actually, with any LP model exactly one of the following will be the case: (i) it will have at least one optimal solution; (ii) it will be infeasible; or (iii) it will

be unbounded. An *infeasible model* is one in which there is no solution that satisfies all of the constraints. An *unbounded model* is one in which the objective function can move infinitely far in the desired direction. In the latter case, it is likely that the model was formulated incorrectly, or does not represent a real-life situation. There are three widely known *techniques for solving* LP problems: the Graphical Solution Technique, the **Simplex** Method, and the Interior Point Method (also known as "Karmarkar's Algorithm").

Graphical Solution Technique

This technique is generally used only for problems that contain two variables; its value is more as a teaching tool than as a practical problem-solving technique. Coordinate axes represent the decision variables and constraints are plotted, often resulting in an enclosed area; assuming that a feasible solution does exist, the set of all feasible solutions in this enclosed area is called the *feasible region* (Figure 1). Each corner of this enclosed area is known as a *corner-point feasible* (CPF) solution; the importance of CPF solutions is that the optimal value of the objective function will come from this set. The objective function is then set to an arbitrary constant, yielding an equation which is plotted as a line (called a contour line or isoquant). Another arbitrary constant is then used to generate a parallel line with the same slope and in the direction of improving the objective function. The last CPF solution to intersect an isoquant line as it leaves the feasible region in the direction of optimization is the optimal solution. The optimal solution(s) will become apparent from this analysis. If there is an optimal solution, then there will be at least one corner point optimal solution; if two corner points are optimal, then all of the points in between them are also optimal.

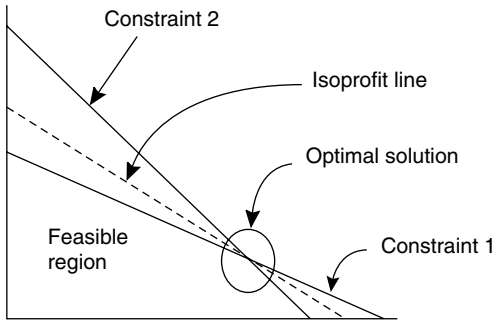

Figure 1 Graphical method for solving linear programming problems

If constraints do not stop the contour lines from moving infinitely far in the desired direction, then the problem is *unbounded*. An unbounded problem has failed to include or appropriately value at least one relevant constraint.

If there is no simultaneous optimal solution to all of the constraints in the problem as formulated, then the model may be infeasible, also known as *inconsistent*, or misspecified. Infeasible LP problems result from the constraint equations, not the objective function, and they should be reviewed if the model is infeasible. In practice, constructing large LP models may involve constraint inputs from many different individuals or teams, so initial infeasibility of a model is not uncommon.

As mentioned above, this technique is applied to two variable problems. LP models with only one decision variable are not of practical significance. With three decision variables, the feasible region would typically be a three-dimensional figure with flat surfaces, and the objective function would be represented by contour planes finding the best corners of this figure. With more than three variables, visual representation becomes impracticable.

Example

A simplified problem that could be solved by the graphical technique is the decision for assigning the mix of appointment slots for two types of patients in a fee-for-service Nurse Practitioner clinic to produce the optimal amount of revenue. Assume that the revenue from a hypertensive patient visit (H) is \$10 and for a patient with diabetes (D) \$18, that Nurse Practitioners (NP) see the former on average for six minutes and the latter for 15 minutes while Nursing Assistants (NA) are with both types of patients for 18 minutes, and that total available Nurse Practitioner hours equals 7000 while those for Nursing Assistants equals 14 000. The objective function is then $10H + 18D$, the first constraint is $0.1H + 0.25D \leq 7000$, and the second constraint is $0.3H + 0.3D \leq 14\,000$. The optimal solution is to schedule 31 111 hypertension patients and 15 555 diabetes patients.

The Simplex Method

George Dantzig developed the Simplex Method in the 1940s and it has proven to be both robust and versatile. This method produces successive outputs, known as tableaus, as part of its iterative search for an optimal solution. To establish an initial tableau, remove inequalities and satisfy the nonnegativity constraints of the model, additional variables are introduced: *slack* variables for constraint equations ("$\leq$") and *surplus* variables for requirements ("$\geq$"). In the optimal solution to the LP, slack or surplus variables will be zero for active constraints and positive for inactive constraints.

The standard equation constraint form of the LP in matrix notation is then

$$\begin{aligned} \text{optimize } Z &= c\mathbf{x} \\ \text{subject to } A\mathbf{x} + \mathbf{s} &= \mathbf{b} \\ \mathbf{x}, \mathbf{s} &\geq 0. \end{aligned}$$

Continuing with the graphical analogy, the Simplex Method first identifies a corner point; if there is none, then the problem is said to be *inconsistent*, or infeasible. Once a corner point is identified, then the **algorithm** moves from corner point to adjacent corner point of the feasible region, with each successive iteration produced by the Gaussian elimination equaling or improving the value of the objective function. The algorithm terminates when an optimal CPF solution is identified; the resulting basic feasible solution consists of the nonnegative variables in the set.

Example

In an article on hospital financial planning [1] linear programming is used at a leading academic medical center to evaluate the resource and revenue implications of changes in patient acuity level and primary insurer. The goal in this formulation is to maximize net revenue after variable expenses:

$$\max \sum_j (r_j - vc_j)^* x_j,$$

where r_j = total revenue from patient type j, vc_j = total variable cost from patient type j, and x_j = number of patients of type j. Here, the decision variables can be optimized for a given situation or varied to explore the financial impact under differing scenarios, for example, in contract negotiations, in considering major capital renovations, or in considering a shift in patient mix based on marketing emphases or regulations. Examples of constraints in this model include the number of beds in each clinical service, ancillary services, requirements that the institution meet at least minimum levels of demand for admission by populations it has traditionally served, and limits on patient demand by various groups – based on the output from a separate forecasting study.

For an LP (the *primal*) there also exists an alternate formulation of the problem, called the *dual*, in which the number of decision variables in the primal equals the number of constraints in the dual and the number of constraints in the primal equals the number of decision variables in the dual. Optimization of the primal also results in optimization of the dual; however, if the primal is a maximization model, then the dual will be a minimization model and vice versa. When the optimal solution contains fewer positive variables than constraints, it is said to be *degenerate*. One result of degeneracy is that there is restricted ability to do postoptimality analyses.

Most problems of practical significance are too large to be solved manually, so specialized software has been written to perform this function. The time to solve an LP is essentially determined by the number of constraints in the problem rather than the number of decision variables; therefore, using the relationship between the primal and dual specifications, solving the formulation with the smaller number of constraints will be faster. Although computing speed and power continue to increase, this is a useful observation since the sizes of the problems to be solved are also increasing.

Beyond computational efficiency there is economic information contained in the correspondence between the primal and dual. Much of the value of an LP solution comes from post-optimality, or sensitivity, analysis. The typical computer output for an optimized LP model will contain values for the objective function, for basic structural variables (those having nonzero values), shadow prices for the constraints and change vectors for Right-Hand-Side and objective ranging or sensitivity analyses. These *change vectors* report how much the basic variables change when the Right-Hand-Side constraints are increased or decreased; *shadow prices* show how much the

value of the objective function will change as the values of the RHS change. Taken together, these results answer the questions of how much of which resources should be purchased, if any, and at what price.

The range of values over which these sensitivity analyses are valid is restricted: information on the lower and upper bounds of these ranges for each variable will be included in the printout under sensitivity analysis. If the quantities being considered in the sensitivity analyses lie outside these ranges, then the model must be rerun with different inputs since the original optimal solution would no longer be valid. Finally, sensitivity results are only correct when the values reported lie within the upper and lower bounds, and only when one variable is changed at a time. If the analyst is interested in simultaneous changes to more than one variable, again the model must be rerun with these changes in the input.

Interior Point Method

An alternative to the Simplex Method was reported by Narendra Karmarkar in 1984. Karmarkar's approach and similar barrier algorithms are based on progressing along successive points interior to the feasible region toward an optimal solution. Since the algorithm does not move from corner to corner as the Simplex does, it is potentially much faster for very large problems and may be the only option for extremely large problems. Research continues on interior point algorithms for LP; however, they are currently inferior to the Simplex Method in supporting sensitivity analysis [8, 9].

Software

Few problems of practical significance can be solved manually; fortunately, developments in hardware and software have moved the ability to solve large linear programming problems to the desktop. Depending on the operating system, size of the problem, price, data input source(s), and other features desired, many options exist to support a knowledgeable user; however, online support for a novice is infrequently provided [15].

Perhaps the most immediately useful of the LP software to a broad audience are the **spreadsheet** packages that include optimization routines or have transparent "add-ins" for this function. These functions, coupled with stored models and report writer capability, allow users easy import of data into widely used applications packages to produce outputs clearly on the basis of an array of scenarios or assumptions. Some commercial packages, e.g. SAS OR [13] (*see* **Software, Biostatistical**) employ enhancements that allow the user greater flexibility of features than the Simplex Method alone would permit.

Quantitative management courses, such as those typically taught in masters programs in business or health administration, increasingly rely on spreadsheet applications to teach these methods – increasing the likelihood that the techniques may be applied more often than in the past.

Internet

Typical LP resources on the **internet** include web sites for professional associations, university courses on LP, computer routines that permit LP problems to be solved over the internet, and other areas of specialization. Although the internet is dynamic, a useful overview of resources and potential uses is provided by Sodhi [16].

Summary

Until recent advances in data availability, hardware, software, and trained users, the trade-off between the time and cost of modeling, obtaining the data, and the likely payoff for what might be one-time efforts, discouraged widespread use of LP in health care. As the expectations for efficient, quality health care increase and integrated financing and delivery systems look for ways to address simultaneously demands for lower cost and higher quality, the opportunities for the application of LP should increase.

References

[1] Brandeau, M.L. & Hopkins, D.S.P. (1984). A patient mix model for hospital financial planning, *Inquiry* **21**, 32–44.

[2] Diehr, G. & Tamura, H. (1989). Linear programming models for cost reimbursement, *Health Services Research* **24**, 329–347.

[3] Greenberg, H.J. (1993). How to analyze the results of linear programs – part 1: preliminaries, *Interfaces* **23**, 56–67.

[4] Greenberg, H.J. (1993). How to analyze the results of linear programs – part 2: price interpretation, *Interfaces* **23**, 97–114.

[5] Greenberg, H.J. (1993). How to analyze the results of linear programs – part 3: infeasibility diagnosis, *Interfaces* **23**, 120–139.

[6] Greenberg, H.J. (1993). How to analyze the results of linear programs – part 4: forcing substructures, *Interfaces* **24**, 121–130.

[7] Harmeier, P.E. (1991). Linear programming for optimization of nurse scheduling, *Computers in Nursing* **9**, 149–151.

[8] Hillier, F.S. & Lieberman, G.J. (1995). *Introduction to Operations Research*, 6th Ed. McGraw-Hill, New York.

[9] Hooker, J.N. (1986). Karmarkar's linear programming algorithm, *Interfaces* **13**, 75–90.

[10] Mangasarian, O.L., Street, W.N. & Wolberg, W.H. (1995). Breast cancer diagnosis via linear programming, *Operations Research* **43**, 570–577.

[11] Nagurney, F. (1992). A regression-like approach to developing a severity index for EMS patients, *Computers in Biology and Medicine* **22**, 123–133.

[12] Rosen, I.I., Lane, R.G., Morrill, S.M. & Belli, J.A. (1991). Treatment plan optimization using linear programming, *Medical Physics* **18**, 141–152.

[13] SAS Institute Inc. (1989). *SAS/OR® User's Guide, Version 6*, 1st Ed. SAS Institute Inc., Cary.

[14] Sexton, T.R., Leiken, A.M., Nolan, A.H., Liss, S., Hogan, A. & Silkman, R.H. (1989). Evaluating managerial efficiency of Veterans Administration medical centers using data envelopment analysis, *Medical Care* **27**, 1175–1188.

[15] Sharda, R. (1995). Linear programming solver software for personal computers: 1995 report, *OR/MS Today* **22**, 49–57.

[16] Sodhi, M.S. (1995). An OR/MS guide to the internet, *Interfaces* **25**, 14–29.

ALAN LYLES

Linear Rank Tests in Survival Analysis

Linear rank tests for survival data are generalized **nonparametric methods** for testing the null hypothesis of equal **survival distributions** among groups.

A number of approaches to generalizing rank tests to censored data have appeared in the literature – approaches which are often quite different from one another. The earliest statistic to reach widespread use was that of Gehan [13], who generalized the **Wilcoxon–Mann–Whitney** scores for the two-group problem. Mantel [24] used arguments based on the construction of the **Mantel–Haenszel** test for stratified $2 \times K$ contingency tables to propose a test that later became known as the **logrank test**. Efron [8] then proposed a statistic based on combining the values of the estimated survival distributions of two groups across time. In 1970 Breslow [5] provided a generalization of the Kruskal–Wallis statistic that reduced to the Gehan–Wilcoxon statistic in two samples. The work of Peto & Peto [25] made important progress in studying the properties of these and other tests. In 1972, the **proportional hazards** model of Cox [6] (*see* **Cox Regression Model**) provided a setting in which the logrank test could be derived as a **partial likelihood** score test from a regression model. Prentice [26] showed in 1978 that many of these tests were asymptotically equivalent to tests that were natural generalizations of the classical linear rank tests for uncensored data described in Hájek & Šidák [17]. In his seminal doctoral thesis and later published work, Aalen [1, 2] showed that the theory of **counting processes** and martingales could be used to recast two-sample tests with right-censored data in the multiplicative intensity model and to study their asymptotic theory. Gill [14] extended this work to a complete study of the operating characteristics of two-sample tests with censored data, and Andersen et al. [3] illustrated the use of this methodology for tests used to compare more than two groups. Remarkably, all these approaches point to essentially the same tests.

Censored data rank tests are most well-developed for right-censored failure time data. Data are right-censored if, for each subject in a study, the underlying data consist of the time, T, to some event and a censoring time, U, while the observable data are $X = \min(T, U)$ and $\delta = I(T \leq U)$, where $I(A)$ is the usual indicator random variable of the event A. The variable T is commonly called a failure time or a survival time. The underlying survival function, $\Pr(T > t)$, is usually denoted by $S(t)$, the cumulative hazard, $-\log S(t)$, by $\Lambda(t)$ for continuous T, and the hazard function for absolutely continuous T by $\lambda(t)$ (*see* **Survival Distributions and Their Characteristics**). If S has discontinuities or is otherwise not differentiable, the cumulative hazard function is given more generally by $\Lambda(t) = -\int_0^t [S(u-)]^{-1}\, \mathrm{d}S(u)$. We let $\pi(t) = \Pr(X \geq t)$, the probability that a subject is at risk at time t.

Linear rank tests are most commonly used when making comparisons among K groups, $K \geq 2$, or when comparing a single group with a known or hypothesized population. For the K-group problem, the observable data consist of the pairs (X_{ij}, δ_{ij}), $1 \leq j \leq n_i$, $i = 1, \ldots, K$; that is, there are n_i observations in the ith of K groups, with underlying survival distribution $S_i(t)$ and probability $\pi_i(t)$ of being at risk. The validity of all the tests discussed below depends centrally on the assumption that the observed failure rate among cases at risk of failure is the same rate that would be observed if censoring were not present. This is satisfied if T_{ij} and U_{ij} are independent random variables for all pairs i, j, and we assume this condition throughout.

We use the counting process setting here. That methodology is not only the most recent and the most successful in studying the asymptotic theory of these tests, but it also provides a surprisingly useful framework for a less formal exploration of their properties. The theory for these tests is best understood for the two-sample problem, and, since the two-group comparison problem is the most prevalent testing problem with censored data, we discuss that case in greater detail.

Gill [15] provides an accessible and intuitive introduction to the martingale approach to survival analysis in the context of the proportional hazards model.

The Counting Process Approach

Counting process methods are now widely used for survival data, and these methods have a particularly simple form when used to study rank tests. Generally, a **stochastic process** $N = [N(t) : t \geq 0]$ is a counting

process if $N(0) = 0$ and it has increasing, right-continuous step functions for paths, with jumps of size 1 at each discontinuity. The process

$$N_{ij} = [N_{ij}(t) = I(X_{ij} \le t, \delta_{ij} = 1); t \ge 0]$$

has simple right-continuous step functions for paths, beginning at 0 at $t = 0$ and taking a single jump to 1 at time t if and only if $T_{ij} = t$ and $T_{ij} \le U_{ij}$. The information in the pair (X_{ij}, δ_{ij}) is equivalent to that in the complete path of N_{ij} as well as in the path of the process $N_{ij}^{\mathrm{U}}(t) = I(X_{ij} \le t, \delta_{ij} = 0)$. Formally, the information up to time t in the pair N_{ij}, N_{ij}^{U} is represented as the σ-algebra $\mathcal{F}_t^{ij}$ generated by the set of variables $[N_{ij}(u), N_{ij}^{\mathrm{U}}(u); 0 \le u \le t]$. The **information** in all K groups up to time t is the product σ-algebra $\mathcal{F}_t = \otimes_{ij} \mathcal{F}_t^{ij}$. Since that information increases with time, the collection of σ-algebras $\mathcal{F} = (\mathcal{F}_t; t \ge 0)$ forms a filtration, i.e. an increasing sequence (in t) of σ-algebras.

The counting process approach uses the stochastic calculus of martingales in its representation of test statistics and the martingale **central limit theorems** [29] for the asymptotic theory (*see* **Large-sample Theory**). A process $M = [M(t); t \le 0]$ is a martingale with respect to a filtration $(\mathcal{G}_t; t \ge 0)$ if

1. $M(t)$ is adapted to $\mathcal{G}_t$ for each t
2. $\mathrm{E}|M(t)| < \infty$ for all $t < \infty$, and
3. $\mathrm{E}[M(t+s)|\mathcal{G}_t] = M(t)$ a.s. for all $s \ge 0, t \ge 0$.

Condition 3 implies that $\mathrm{E}[M(t) - M(u)|\mathcal{G}_u] = 0$ for all $u \le t$, and this is sometimes written informally as $\mathrm{E}[\,\mathrm{d}M(t)|\mathcal{G}_{t-}] = 0$. A process M is called a submartingale if the equation for the conditional expectation in condition 3 above is replaced by the inequality $\mathrm{E}[M(t+s)|\mathcal{G}_t] \ge M(t)$. Because the martingale property depends on the underlying filtration, we sometimes say that M is a $\mathcal{G}_t$-martingale. The martingale definition implies that a linear combination of processes which are martingales with respect to a common filtration $\mathcal{G}$ will itself be a $\mathcal{G}_t$-martingale.

It is possible to show that, when T_{ij} and U_{ij} are independent for all pairs i, j, the process

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(u)\,\mathrm{d}\Lambda_i(u) \qquad (1)$$

is a martingale with respect to the filtration $\mathcal{F}$ defined above, where $Y_{ij}(u) = I(X_{ij} \ge u)$ is the process denoting whether or not subject i, j has failed or been censored before time t (cf. Theorems 1.3.1 and 1.3.2 in Fleming & Harrington [10]). The integral on the right-hand side of (1) is called the compensator for the process N_{ij}. For simplicity of notation we usually write (1), and others like it, as

$$M_{ij} = N_{ij} - \int Y_{ij}\,\mathrm{d}\Lambda_i.$$

It is not surprising that M_{ij} is an $\mathcal{F}_t$-martingale. Conditional on the history of the failure and censoring processes before time t, the conditional probability of a jump in N_{ij} at t is approximately $Y_{ij}(t)\,\mathrm{d}\Lambda_i(t)$, so that $E(\mathrm{d}N_{ij} - Y_{ij}\,\mathrm{d}\Lambda_i|\mathcal{F}_{t-}) = 0$. This result is an example of the more general Doob–Meyer decomposition for submartingales (cf. [9]), which states that for any submartingale Z (subject to boundedness conditions) there exists a predictable process A, called the compensator for Z, such that $Z - A$ is a martingale.

Nearly all commonly used linear rank statistics for survival data can, under H_0, be represented, or at least approximately so, as $\sum_{ij} \int H_{ij}\,\mathrm{d}M_{ij}$, or as sums of stochastic integrals of "predictable" processes with respect to the fundamental martingale processes. This construction allows the use of the stochastic calculus for martingales outlined below. More detail may be found in Fleming & Harrington [10] and Andersen et al. [4]. To keep the technical material to a minimum, we have not given the most general versions of these results.

There are several definitions of a predictable process; the following is one of the more accessible.

Definition. A stochastic process $H = [H(t); t \ge 0]$ defined on a probability space $(\Omega, \mathcal{A}, P)$ is predictable with respect to a filtration $\mathcal{F} = (\mathcal{F}_t; t \ge 0)$ on that space if H is measurable with respect to the smallest σ-algebra on $[0, \infty) \times \Omega$ generated by the adapted left-continuous processes.

This definition implies that any left-continuous $\mathcal{F}_t$-adapted process will be $\mathcal{F}_t$-predictable; operationally, that is how predictability is checked.

Slightly more general versions of the following theorems appear in Fleming & Harrington [10, cf. Theorem 2.4.1].

Theorem 1. Suppose H is a bounded, $\mathcal{F}_t$-predictable process and M an $\mathcal{F}_t$-martingale with

$M(0+) - M(0) = 0$. Then the process

$$\int H \, dM = \left[\int_0^t H(u) \, dM(u); t \geq 0 \right]$$

is an $\mathcal{F}_t$-martingale.

Theorem 2. Suppose M_1 and M_2 are square integrable $\mathcal{F}_t$-martingales (i.e. $\sup_{t \geq 0} EM_i^2(t) < \infty, i = 1, 2$). Then there exists a unique $\mathcal{F}_t$-predictable process $\langle M_1, M_2 \rangle$ such that $M_1 M_2 - \langle M_1, M_2 \rangle$ is an $\mathcal{F}_t$-martingale.

The process $\langle M_1, M_2 \rangle$ is called the predictable covariation process for the martingales M_1 and M_2. When M_1 and M_2 are the same process M, $\langle M, M \rangle$ is called the predictable quadratic variation process, and is often denoted by $\langle M \rangle$. Since martingales have constant expected value, $EM_1(t)M_2(t) = E\langle M_1, M_2 \rangle(t)$ whenever $M_1(0)M_2(0) - \langle M_1, M_2 \rangle(0) = 0$. This formula is particularly valuable for computing second **moments** when the quadratic variation process takes a simple form, as it does for the counting process martingales arising in survival analysis.

It is possible to show that

$$\left\langle \int H_1 \, dM_1, \int H_2 \, dM_2 \right\rangle = \int H_1 H_2 \, d\langle M_1, M_2 \rangle$$

(cf. Fleming & Harrington [10, Theorem 2.4.2]).

The following summarizes results on quadratic variation processes for counting process martingales.

Definition 1. A k-dimensional counting process $(N_1, N_2, \ldots, N_k)$ is called a multivariate counting process if each component N_j is a counting process and no two component processes jump at the same time.

Theorem 3. Let $(N_1, N_2, \ldots, N_k)$ be a multivariate counting process, and let A_j be the compensator of N_j. Then $\langle M_j, M_j \rangle = \int (1 - \Delta A_j) \, dA_j$ and $\langle M_i, M_j \rangle = -\int \Delta A_i \, dA_j$ for $i \neq j$.

When a survival distribution is continuous, $\Delta A = 0$, and these formulas are particularly simple. In that case $\langle M_i, M_j \rangle = 0$ and the martingales M_i and M_j are called orthogonal. The more general formula is useful, however, in estimating second moments when there are **ties** in observed failure times, even when the underlying model is continuous, as will be seen below.

Common Linear Rank Tests

Despite its demanding technical foundation, the martingale approach to rank tests is a useful setting for formulating tests. This is most easily seen for two-sample tests. Suppose two groups have survival functions S_1 and S_2 and cumulative hazard functions Λ_1 and Λ_2. Let $\overline{N}_i = \sum_j N_{ij}$, $\overline{Y}_i = \sum_j Y_{ij}$, and $\overline{M}_i = \overline{N}_i - \int \overline{Y}_i \, d\Lambda_i$, $i = 1, 2$. The observed number of failures at time t in group 1 is $dN_1(t)$; with independent censoring and under H_0: $\Lambda_1 = \Lambda_2$, the conditionally expected number of failures in group 1 at t, given that a failure has been observed at t, is

$$\frac{\overline{Y}_1(t)\{ d[\overline{N}_1(t) + \overline{N}_2(t)]\}}{[\overline{Y}_1(t) + \overline{Y}_2(t)]}.$$

A simple test of H_0 can be constructed by comparing

$$\int_0^\infty d\overline{N}_1 - \overline{Y}_1 \frac{d(\overline{N}_1 + \overline{N}_2)}{\overline{Y}_1 + \overline{Y}_2} \tag{2}$$

with 0. This statistic is the numerator of the logrank statistic. Simple algebra shows that under H_0 the above expression is equal to

$$\begin{aligned}
\int_0^\infty \frac{\overline{Y}_1 \overline{Y}_2}{\overline{Y}_1 + \overline{Y}_2} \left(\frac{d\overline{N}_1}{\overline{Y}_1} - \frac{d\overline{N}_2}{\overline{Y}_2} \right) &= \int_0^\infty \frac{\overline{Y}_1 \overline{Y}_2}{\overline{Y}_1 + \overline{Y}_2} \\
&\quad \times \left[\left(\frac{d\overline{N}_1}{\overline{Y}_1} - d\Lambda_1 \right) - \left(\frac{d\overline{N}_2}{\overline{Y}_2} - d\Lambda_2 \right) \right] \\
&= \int_0^\infty \frac{\overline{Y}_2}{\overline{Y}_1 + \overline{Y}_2} \, d\overline{M}_1 - \int_0^\infty \frac{\overline{Y}_1}{\overline{Y}_1 + \overline{Y}_2} \, d\overline{M}_2 \\
&= \sum_{j=1}^{n_1} \int_0^\infty \frac{\overline{Y}_2}{\overline{Y}_1 + \overline{Y}_2} \, dM_{1j} \\
&\quad - \sum_{j=1}^{n_2} \int_0^\infty \frac{\overline{Y}_1}{\overline{Y}_1 + \overline{Y}_2} \, dM_{2j}.
\end{aligned} \tag{3}$$

Gill [14] used an expression similar to (3) as a foundation for a generalized class of statistics that includes many previously discussed in the literature.

Definition. Suppose one has two samples of right-censored observations (X_{ij}, δ_{ij}), $1 \leq j \leq n_i$, $i = 1, 2$, giving rise to the counting and at risk processes $\overline{N}_i, \overline{Y}_i$, $i = 1, 2$, and let $(\mathcal{F}_t; t \geq 0)$ be the filtration generated by $[N_{ij}(u), Y_{ij}(u); 0 \leq u \leq t, 1 \leq j \leq$

$n_i, i = 1, 2]$. Let K be a bounded nonnegative $\mathcal{F}_t$-predictable process satisfying $K(t) = 0$ whenever $\overline{Y}_1(t)\overline{Y}_2(t) = 0$. Then

$$G_K = \int_0^\infty K\left(\frac{\mathrm{d}\overline{N}_1}{\overline{Y}_1} - \frac{\mathrm{d}\overline{N}_2}{\overline{Y}_2}\right)$$

is called a statistic of the class $\mathcal{K}^+$. When

$$K = \left(\frac{n_1 + n_2}{n_1 n_2}\right)^{1/2} W\left(\frac{\overline{Y}_1\overline{Y}_2}{\overline{Y}_1 + \overline{Y}_2}\right)$$
$$= \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} W\frac{\overline{Y}_1}{n_1}\frac{\overline{Y}_2}{n_2}\frac{n_1 + n_2}{\overline{Y}_1 + \overline{Y}_2},$$

the statistic is called a weighted logrank statistic. The fraction $\overline{Y}_1\overline{Y}_2/(\overline{Y}_1 + \overline{Y}_2)$ appears in the usual logrank statistic; as will be seen later, the terms involving sample sizes ensure **convergence** under null and alternative hypotheses. The function W reweights the observed minus expected increments in (2). Although weighted logrank statistics are a subset of the statistics of class $\mathcal{K}^+$, they are the most common in applications, and we give those somewhat more attention here. Since the function W is the important part of the weight function in these statistics, we denote weighted logrank statistics by G_W.

The upper limit of integration in the integral representation for statistics of class $\mathcal{K}^+$ occasionally causes confusion. Because the weight function in the two-sample statistic takes value 0 as soon as at least one of the **risk sets** is empty, the integral as written denotes a statistic computed for the portion of the time axis over which there are cases at risk in both groups. Contributions to the statistic stop when all cases in either one of the groups have failed or have been censored; that is the most natural way for the practitioner to think of these nonparametric statistics. In the asymptotic theory for some of these statistics, the integral may be computed only over a prespecified time interval for which the probability of a subject being at risk at the right end point is bounded away from zero. That is not necessary for most of the statistics discussed here. Finally, the upper limit of integration may be thought of as a variable t when the statistic is considered a process with changing values as time increases. This last perspective is used in the martingale calculus.

If W is a predictable process, a weighted logrank statistic can, under H_0, be represented as sums of stochastic integrals of predictable processes with respect to martingales, so that G_W is itself a martingale. This representation and its quadratic variation process can be used to derive formulas for the first two moments of these test statistics. The following summarizes Theorems 3.3.1 and 3.3.2 in Fleming & Harrington [10]; except for some regularity conditions and tedious algebra, it follows directly from Theorems 1 and 2.

Theorem 4. Let G_K be a statistic of the class $\mathcal{K}$. When $\Lambda_1 = \Lambda_2 = \Lambda, \mathrm{E}G_K = 0$ and

$$\mathrm{E}G_K^2 = \mathrm{E}\sum_{i=1}^{2}\int_0^\infty \frac{K^2}{\overline{Y}_i}(1 - \Delta\Lambda)\,\mathrm{d}\Lambda.$$

The variance estimator

$$\hat{\sigma}^2 = \int_0^\infty \sum_{i=1}^{2}\frac{K^2}{\overline{Y}_i}\left(1 - \frac{\Delta\overline{N}_1 + \Delta\overline{N}_2 - 1}{\overline{Y}_1 + \overline{Y}_2 - 1}\right)$$
$$\times \frac{\mathrm{d}(\overline{N}_1 + \overline{N}_2)}{\overline{Y}_1 + \overline{Y}_2} \tag{4}$$

is an **unbiased** estimate of $\mathrm{E}G_K^2$.

When Λ is continuous, $\Delta\Lambda = 0$ in the expression for $\mathrm{E}G_K^2$, and the second term in the sum comprising the integrand in $\hat{\sigma}^2$ would seem unnecessary. That term is present only when two or more observed failure times are equal, however, and seems to improve the small-sample behavior of the estimator in tied data.

When $W = 1$, the statistic is the logrank statistic originally proposed by Mantel [24]. The same statistic arises as a partial likelihood score statistic in a proportional hazards regression model with a single binary covariate, although that approach leads to the variance estimate which assumes $\Delta\Lambda$ is identically 0. When $W(t)$ is a function of the proportion of cases at risk at time t, $[\overline{Y}_1(t) + \overline{Y}_2(t)]/(n_1 + n_2)$, the statistic is a member of the family proposed by Tarone & Ware [32]. If W is exactly the proportion of cases at risk, then G_W is the Gehan [13] generalization of the Wilcoxon statistic. If $W = \hat{S}^-$, the left-continuous version of the **Kaplan–Meier** [19] estimator computed with the two groups combined, then the statistic is asymptotically equivalent to the Wilcoxon generalization proposed by Prentice [26] and to a similar statistic proposed by Peto & Peto [25]. When $W = (\hat{S}^-)^\rho, \rho > 0$, the resulting family of statistics is that proposed by Harrington &

Fleming [18]. In this family, the logrank statistic corresponds to $\rho = 0$ and the Prentice–Wilcoxon to $\rho = 1$. Gray & Tsiatis [16] have shown that this family may be extended to allow $\rho < 0$. The statistic proposed by Efron [8] is equivalent to one with weight

$$W = \frac{(\overline{Y}_1 + \overline{Y}_2)\hat{S}_1^- \hat{S}_2^- I(\overline{Y}_1 \overline{Y}_2 > 0)}{(\overline{Y}_1 \overline{Y}_2)}.$$

Since both $\overline{Y}_1$ and $\overline{Y}_2$ are $\mathcal{F}_t$-adapted and left-continuous, the use of the left-continuous version of the Kaplan–Meier estimator in these statistics ensures predictability of the integrand K.

The counting process representation of these statistics provides insight into the term "linear rank tests". In uncensored data, where $X_{ij} = T_{ij}$ for all pairs i, j, a classical linear rank statistic as discussed in Hájek & Šidák [17] has the form $\sum_{i,j} a(R_{ij})$, where R_{ij} is the rank of T_{ij} in the combined sample and a is function assigning scores to the ranks. When the weight function W in a weighted logrank statistic depends only on the order of the possibly censored observations in the combined sample, as in all the statistics discussed above, then the Stieltjes integral representation of the statistic consists of a linear combination of scores assigned to observed failures according to the ordering of the observations X_{ij}.

The counting process representation also sheds light into the operating characteristics of these two-sample tests. The weighted logrank statistics can be written as

$$c \int_0^\infty W \left[\mathrm{d}\overline{N}_1 - \overline{Y}_1 \frac{\mathrm{d}(\overline{N}_1 + \overline{N}_2)}{\overline{Y}_1 + \overline{Y}_2} \right],$$

where c is a constant depending only on sample size. When W is constant, the observed minus expected failures are weighted equally, so that deviations from 0 in these terms in the right tail of the observations, where the risk sets are small, have as much influence on the value of the statistic as early deviations at times with large risk sets. If S_1 and S_2 are absolutely continuous and $\lambda_1 - \lambda_2$ changes sign at some time t, i.e. the underlying distributions have crossing hazard functions, then the logrank statistic may have a value that is not significantly different from zero, regardless of sample size (cf. Fleming et al. [11] for an example from a clinical study and Prentice & Marek [27] for an extended discussion). If W decreases when the observations increase, as in the Harrington–Fleming family when $\rho > 0$, then earlier differences in the observed minus expected failures will be emphasized. Gill [14] shows that statistics of the class $\mathcal{K}^+$ are consistent as long as $\Lambda_1(t) \geq \Lambda_2(t)$ for all t or vice versa, with strict inequality on at least one interval containing nonzero mass for the two distributions. Formal results about asymptotic operating characteristics under alternative hypotheses are summarized later.

The K-sample, $K > 2$, statistics are natural generalizations of the two-sample tests. Let $\overline{N}_i, \overline{Y}_i, i = 1, \ldots, K$, be defined as with two groups, and let $\overline{N} = \sum_i \overline{N}_i$ and $\overline{Y} = \sum_i \overline{Y}_i$. Under H_0: $\Lambda_1 = \Lambda_2 = \cdots = \Lambda_K$, the K-dimensional statistic with the ith component given by

$$G_{W,i} = \int_0^\infty W \left(\mathrm{d}\overline{N}_i - \overline{Y}_i \frac{\mathrm{d}\overline{N}}{\overline{Y}} \right)$$

is, for each group, a weighted sum of observed minus conditionally expected number of failures. It is not difficult to show that $\sum_i G_{W,i} = 0$, so that there are only $K - 1$ linearly independent components in the statistic. More detailed information about the covariance of the components of the statistic comes from the equivalent (under H_0) martingale representation

$$G_{W,i} = \sum_{l=1}^K \int_0^\infty W \left(r_{il} - \frac{\overline{Y}_i}{\overline{Y}} \right) \mathrm{d}\overline{M}_l,$$

where $r_{il} = 1$ when $i = l$ and 0 otherwise. If we assume that the underlying cumulative hazard functions are continuous and that there are no ties in the observed data, the components of the statistic can be written as a sum of integrals with respect to orthogonal martingales. The simpler formulas for quadratic variation and covariation can be used to show that, under the hypothesis that all groups have a common cumulative hazard Λ,

$$\langle G_{W,i} \rangle = \sum_{l=1}^K \int_0^\infty W^2 \left(r_{il} - \frac{\overline{Y}_i}{\overline{Y}} \right)^2 \overline{Y}_l \, \mathrm{d}\Lambda$$

and

$$\langle G_{W,i}, G_{W,k} \rangle = \sum_{l=1}^K \int_0^\infty W^2 \left(r_{il} - \frac{\overline{Y}_i}{\overline{Y}} \right)$$

$$\times \left(r_{kl} - \frac{\overline{Y}_k}{\overline{Y}} \right) \overline{Y}_l \, \mathrm{d}\Lambda$$

$$= \int_0^\infty W^2 \frac{\overline{Y}_i}{\overline{Y}} \left(r_{ik} - \frac{\overline{Y}_i}{\overline{Y}} \right) \overline{Y} \, \mathrm{d}\Lambda.$$

The last expression leads to a natural estimator $\hat{\Sigma}$ of the **covariance matrix** of the statistic, with elements $\hat{\sigma}_{ik}$ given by

$$\hat{\sigma}_{ik} = \int_0^\infty W^2 \frac{\overline{Y}_i}{\overline{Y}} \left(r_{ik} - \frac{\overline{Y}_k}{\overline{Y}} \right) \mathrm{d}\overline{N}. \qquad (5)$$

As with the two-sample statistic, it is possible to show that $\mathrm{E}(\hat{\sigma}_{ik}) = \mathrm{cov}(G_{W,i}, G_{W,k})$.

The basic martingale may also be used to construct a statistic for comparing a single sample with a known or hypothesized population distribution with failure rate $\mathrm{d}\Lambda_0$. The natural analog of the weighted logrank statistic is

$$c \int_0^\infty W(\, \mathrm{d}\overline{N}_1 - \overline{Y}_1 \, \mathrm{d}\Lambda_0 \, \mathrm{d}u).$$

These statistics are discussed in detail in Andersen et al. [4] and Woolson [33].

Asymptotic Distribution Theory

Large-sample normality for the K-sample statistics under both null and alternative distributions has been established by a number of authors. The original derivations of these tests by Gehan, Mantel, and others contained strong plausibility arguments for asymptotic distributions, and Schoenfeld [31] may have been one of the first to establish formally the asymptotic **efficiency** of the logrank test under proportional hazards alternatives. Using the original martingale formulation of Aalen and the martingale central limit theorem of Rebolledo [29], Gill provided a thorough study of the large-sample operating characteristics of the two-sample tests under both null and alternative hypotheses (*see* **Power**). Andersen et al. [3] used the same methodology to study the large-sample behavior of tests for more than two samples. The theorems below summarize the major results in this area. The first results provide asymptotic distributions under the null hypothesis, first for the two-sample case, and then for the general K-sample statistics.

The most general theorems about the convergence of these statistics require more regularity conditions than might at first be expected. Beyond the usual conditions needed for central limit theorems, the asymptotic normality in these statistics can be disturbed if the weight function W becomes too large or if the cumulative hazard approaches infinity too quickly. The following theorem for the two-sample case covers nearly all the statistics used in practice and, because of the form of the weight function, does not require many conditions. This theorem appears as Theorem 7.2.1 in Fleming & Harrington [10], and relies for its proof on the more general result of Corollary 4.3.1 in Gill [14].

Theorem 5. In the two-sample testing problem with right-censored data (as described above), let $\hat{S}(t)$ denote the Kaplan–Meier estimator computed from the combined samples. Let $\hat{\pi}(t)$ denote the pooled sample estimator of the probability that a subject is alive and uncensored at time t, i.e. $\hat{\pi}(t) = [\overline{Y}_1(t) + \overline{Y}_2(t)]/(n_1 + n_2)$. Let f be a nonnegative bounded continuous function of bounded variation on [0, 1]. Suppose the weighted logrank statistic G_W has weight function of the form

$$\left(\frac{n_1 n_2}{n_1 + n_2} \right)^{1/2} W(t) \frac{\overline{Y}_1(t)}{n_1} \frac{\overline{Y}_2(t)}{n_2} \frac{n_1 + n_2}{\overline{Y}_1 + \overline{Y}_2},$$

where $W(t) = f[\hat{S}(t-)]$ or $W(t) = f[\hat{\pi}(t)]$. Suppose that $\lim_n n_i/n = a_i$ exists and lies in (0, 1), and let $\hat{\sigma}$ be as in (4). Then under H_0: $\Lambda_1 = \Lambda_2$, $G_W/\hat{\sigma}$ converges in distribution, as $n \to \infty$, to a normally distributed random variable with mean 0 and variance 1.

Theorem 6 follows from Theorem V.2.1 and Example V.2.10 in Andersen et al. [4]. Because of the linear dependence of the terms in the K-dimensional statistic, the last component is usually dropped when computing the quadratic form for a Wald test (*see* **Likelihood**).

Theorem 6. In the K-sample testing problem with right-censored data (as described above), let $\hat{S}(t)$ denote the Kaplan–Meier estimator computed from the combined samples. Let $\hat{\pi}(t)$ denote the pooled sample estimator of the probability that a subject is at risk at time t, i.e.

$$\hat{\pi}(t) = [\overline{Y}_1(t) + \cdots + \overline{Y}_K(t)]/(n_1 + \cdots + n_K).$$

Let f be a nonnegative bounded continuous function of bounded variation on [0, 1]. Suppose the weight function W in the K-sample weighted logrank statistic is of the form $W(t) = f[\hat{S}(t-)]$ or $W(t) = f[\hat{\pi}(t)]$, and let the ith component of the standardized statistic be given by

$$G_{W,i} = \left[\frac{n}{n_i(n-n_i)}\right]^{1/2} \int_0^\infty W\left(\mathrm{d}\overline{N}_i - \overline{Y}_i \frac{\mathrm{d}\overline{N}}{\overline{Y}}\right),$$

where $n = \sum_i n_i$. Let the column vector $\mathbf{G}$ be given by $\mathbf{G}' = (G_{W,1}, \ldots, G_{W,K-1})$, and let the estimated covariance matrix for this $(K-1)$ dimensional vector be denoted by $\hat{\mathbf{\Sigma}}_{K-1}$ with elements given in (5). Suppose that $\lim_n n_i/n$ exists and lies in (0, 1) for each i. Then, under H_0: $\Lambda_1 = \cdots = \Lambda_k$, and as $n \to \infty$, the quadratic form $\mathbf{G}'\hat{\mathbf{\Sigma}}_{K-1}^{-1}\mathbf{G}$ converges in distribution to a χ^2 random variable with $K-1$ **degrees of freedom** (*see* **Chi-square Distribution**).

When all K components of the statistic are used in the quadratic form, a generalized inverse of the complete, singular covariance matrix may be used (*see* **Matrix Algebra**).

Asymptotic distributions under alternative hypotheses provide information about the power of the statistics. Gill [14] has shown that all two-sample tests of class $\mathcal{K}^+$ have asymptotic power 1, i.e. are consistent, under ordered hazards alternatives. Consequently, asymptotic power comparisons must be made under sequences of alternatives approaching the null hypothesis. Results with the counting process formulation are again simplest in the two-sample case. If a test statistic is asymptotically normal under a sequence of alternative hypotheses converging to the null hypothesis as the sample size increases, then the ratio of the square of the asymptotic mean to the asymptotic variance is called the (asymptotic) *efficacy*. The efficacy will be the noncentrality parameter in the χ^2 distribution for the square of the statistic, and the ratio of efficacies for two statistics, computed under the same sequence of alternatives, has the same value as the asymptotic ratio of the sample sizes needed for the two tests to have equal power. To avoid technical details, we will argue only heuristically here. Generally, much more care must be taken when establishing limiting distributions under sequences of alternative distributions. Gill [14] contains detailed results for the two-sample problem; the results for the K-sample problem may be found in Andersen et al. [3, 4]. The asymptotic theory of testing, especially for rank-based methods, may be found in Randles & Wolfe [28] and Hájek & Šidák [17].

For simplicity, we assume that the underlying survival distributions are absolutely continuous. We let $n = n_1 + n_2$ index the underlying survival and hazard functions in the sequence of alternative distributions. When the two hazard functions λ_1^n and λ_2^n are not equal, a two-sample statistic of the class $\mathcal{K}^+$ can be written

$$\begin{aligned} G_K &= \int_0^\infty K\left(\frac{\mathrm{d}\overline{N}_1}{\overline{Y}_1} - \frac{\mathrm{d}\overline{N}_2}{\overline{Y}_2}\right) = \int_0^\infty K\left(\frac{\mathrm{d}\overline{N}_1}{\overline{Y}_1} - \lambda_1^n\right) \\ &\quad - \int_0^\infty K\left(\frac{\mathrm{d}\overline{N}_2}{\overline{Y}_2} - \lambda_2^n\right) + \int_0^\infty K(\lambda_1^n - \lambda_2^n) \\ &= \int_0^\infty \frac{K}{\overline{Y}_1}\,\mathrm{d}\overline{M}_1 - \int_0^\infty \frac{K}{\overline{Y}_2}\,\mathrm{d}\overline{M}_2 \\ &\quad + \int_0^\infty K\left(\frac{\lambda_1^n}{\lambda_0} - \frac{\lambda_2^n}{\lambda_0}\right)\lambda_0, \end{aligned} \tag{6}$$

where λ_0 is the hypothetical common hazard function under the null hypothesis. For the weighted logrank statistics,

$$K = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} W \frac{\overline{Y}_1}{n_1}\frac{\overline{Y}_2}{n_2}\frac{n_1 + n_2}{\overline{Y}_1 + \overline{Y}_2},$$

where W is usually a function converging to some deterministic function w. The martingale central limit theorem implies that the first two terms in (6) converge to a mean zero Gaussian **(normal)** random variable, and the strong **law of large numbers** implies that the last two terms in the equation for K above converge collectively to $\pi_1\pi_2/(a_1\pi_1 + a_2\pi_2)$. Loosely speaking, the convergence of the statistic to a Gaussian variable under a sequence of alternatives will depend on the convergence of

$$\int_0^\infty \frac{\pi_1\pi_2}{a_1\pi_1 + a_2\pi_2} w \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \left(\frac{\lambda_1^n}{\lambda_0} - \frac{\lambda_2^n}{\lambda_0}\right)\lambda_0.$$

The last three terms in the integrand above may be written as

$$\left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \left[\left(\frac{\lambda_1^n}{\lambda_0} - 1\right) - \left(\frac{\lambda_2^n}{\lambda_0} - 1\right)\right]\lambda_0.$$

Convergence under the sequence of alternative distributions will thus depend on convergence of

$$\left[\frac{n_1 n_2}{(n_1+n_2)}\right]^{1/2}\left[\left(\frac{\lambda_i^n}{\lambda_0}\right)-1\right]$$

to a function $g_i, i=1,2$. This implies when $i=1$, for instance, that the ratio of functions λ_1^n/λ_0 must converge to 1 at rate $n_1^{1/2}$, and that λ_1^n/λ_2^n must also converge to 1 at the same rate. If $g=g_1-g_2$, then the asymptotic mean of the statistic under this sequence of alternatives will be

$$\mu=\int_0^\infty \frac{\pi_1\pi_2}{a_1\pi_1+a_2\pi_2} w g \lambda_0.$$

The asymptotic variance of the statistic will be determined by the first two integrals in (6), and turns out to equal the asymptotic variance under the null hypothesis,

$$\sigma^2=\int_0^\infty \frac{\pi_1\pi_2}{a_1\pi_1+a_2\pi_2} w^2 \lambda_0.$$

The asymptotically best weighted logrank test statistic is the member of that class that maximizes this efficacy with respect to the asymptotic weight function w. Gill [14] uses a Lagrange multiplier argument to show that, in fact, the asymptotic efficacy is maximized over all of $\mathcal{K}^+$ when a statistic is a weighted logrank statistic with weight function W converging to an asymptotic weight function w proportional to g. This result can be used to calculate the best test from $\mathcal{K}^+$ against particular types of alternatives.

Suppose that $\lambda_i^n=\lambda_{\theta_i^n}$. Then,

$$\left(\frac{n_1 n_2}{n_1+n_2}\right)^{1/2}\left(\frac{\lambda_1^n}{\lambda_0}-1\right) = \frac{\lambda_{\theta_i^n}-\lambda_{\theta_0}}{\theta_i^n-\theta_0}\times\left(\frac{n_1 n_2}{n_1+n_2}\right)^{1/2}\frac{\theta_i^n-\theta_0}{\lambda_{\theta_0}}. \qquad (7)$$

If $\theta_i^n\to\theta_0$ such that

$$\lim_{n\to\infty}\left(\frac{n_1 n_2}{n_1+n_2}\right)^{1/2}(\theta_i^n-\theta_0)=c_i^*,$$

then both sides in (7) will approach $c_i^*\partial/\partial\theta\log\lambda_\theta$, where the derivative with respect to θ is evaluated at θ_0. Consequently, the function g appearing in the asymptotic mean and efficacy under a sequence of alternatives will be proportional to

$$\left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0}\log\lambda_\theta.$$

This result confirms that, for instance, when $\lambda_{\theta_i^n}(t)=\lambda_0(t)\exp(\theta_i^n), i=1,2$ (i.e. proportional hazards alternatives), the most efficient statistic of class $\mathcal{K}^+$ has a constant weight function W (i.e. is the logrank statistic).

Computing asymptotic relative efficiencies against optimal tests for parametric models is more difficult, but uses the same approach. Gill [14] shows that, when the hazard function in the two-sample problem is known up to a single parameter $\theta_i^n, i=1,2$, satisfying

$$\theta_i^n-\theta_0=(-1)^{i+1}c\left[\frac{n_{i'}}{n_i(n_1+n_2)}\right]^{1/2},\quad i\neq i',$$

then the asymptotic efficacy of the **likelihood ratio test** of equality of the two hazard functions is given by

$$\int_0^\infty\left(\frac{\partial}{\partial\theta}\log\lambda_\theta\big|_{\theta=\theta_0}\right)^2(a_2\pi_1+a_1\pi_2)\lambda_{\theta_0}.$$

Gill also shows that the ratio of the asymptotic efficacies comparing an optimal test of class $\mathcal{K}^+$ to the likelihood ratio test is bounded above by 1, as expected, but that the ratio may equal 1 when $\pi_1=\pi_2$. Under random censoring, $\pi_i(t)=\Pr(T_{ij}\geq t)\Pr(U_{ij}\geq t)$. Since in the limit $\Pr(T_{1j}\geq t)=\Pr(T_{2j}\geq t)$, fully efficient tests of the class $\mathcal{K}^+$ can be found when asymptotic censoring distributions are equal.

The study of asymptotic operating characteristics of general K-sample tests uses similar tools; the interested reader can find a detailed treatment in Andersen et al. [4].

The counting process and martingale framework provides methods for a rigorous study of the operating characteristics of linear rank tests for censored data, but a variety of other approaches have been used in special cases. As mentioned in the introduction, Gehan [13] originally generalized the two-sample Wilcoxon test by extending the notion of the scores used in the Mann–Whitney version of the Wilcoxon statistic. Gehan then used a permutation argument

to compute a distribution under the null hypothesis, conditional on the observed pattern of censorship (*see* **Randomization Tests**), and argued that the permutation distribution would approach normality in large samples. Since the Gehan–Wilcoxon statistic corresponds to a weighted logrank statistic with weight function $W(t) = [\overline{Y}_1(t) + \overline{Y}_2(t)]/(n_1 + n_2)$, which asymptotically depends on both the underlying survival and censoring distributions in the groups, Gill's results show that this version of the Wilcoxon statistic has operating characteristics that depend on the censoring distribution. Leurgans [22, 23] also provides extensive applied and theoretical discussions of the asymptotic operating characteristics of rank statistics for censored data.

Mantel's [24] original derivation of the two-sample logrank statistic treated the observations as a series of **2 × 2 contingency tables**, with one table at each observed failure time. The marginal classifications of the tables denoted the number of subjects at risk in each group just prior to the observed failure time, and the numbers of subjects failing or not failing at the observed time. Mantel argued that, conditional on the risk sets in the two groups at the observed failure times, the set of observed minus conditionally expected number of failures in group 1 were independent, and that standard central limit theorems could be used to justify asymptotic normality. Mantel's arguments were heuristic, but the differences between observed and conditionally expected failures in the tables are exactly the increments in the integral representation, (2). The martingale representation shows that these increments are uncorrelated and consequently asymptotically independent. Mantel also argued that, since the Mantel–Haenszel statistic on which the logrank test is based is efficient at detecting a constant **odds ratio** different from 1 in stratified 2×2 tables, the logrank should have good power against proportional hazards alternatives.

Prentice [26] generalized the theory of linear rank tests, as described in Hájek & Šidák [17], to censored observations by suggesting a modification to the efficient score. This approach outlined a general context for linear rank tests for censored data, and showed that many of the statistics finding widespread use could be thought of as special cases of this general approach. Prentice was the first to show that Gehan's generalization of the Wilcoxon statistic was not the only natural way to create a Wilcoxon-type statistic for censored data. The Wilcoxon statistic in uncensored data arises from the optimal scoring function for shift alternatives in the **logistic distribution** [17], and Prentice's generalized scoring function for censored data led to a statistic that was approximately the weighted logrank statistic with weight function $\hat{S}(t-)$. Cuzick [7] later showed that many of the asymptotic results on linear rank tests for uncensored data could be extended to test statistics using the Prentice scoring function.

Because of space constraints, many important contributions to this field have necessarily been omitted. First, in the interest of simplicity, we have suppressed much of the generality that results from the use of martingale theory. The more complete treatments in [4], [14], and elsewhere discuss the use of local martingales, which relax some of the implicit boundedness conditions in the martingale definition. Local martingales allow more general weight functions in $\mathcal{K}^+$ statistics and also are used, sometimes in subtle ways, in the proofs of many of the results stated here. The asymptotic theory for these tests has been described only briefly (with mathematical details kept to a minimum) and small-sample properties not at all. It is possible to derive a permutation distribution for tests such as the logrank under the null hypothesis and the assumption of equal censoring in both groups, but most small-sample studies have relied on **simulation** (cf. Lee et al. [21] and Latta [20]). Alternative variance estimators to those given here are discussed in Andersen et al. [4]. Stratified tests are available for situations when differences among groups within strata are constant but baseline failure rates across strata differ (cf. [4]) (*see* **Stratification**). Gastwirth [12] and others have discussed the problem of how to combine linear rank tests for censored data when it is difficult to specify clearly the form of an alternative hypothesis. Robins & Rotnitzky [30] have proposed tests that relax the independent censoring assumption so prominently used here. Many authors have studied nonparametric estimates of survival distributions with left- or interval-censored data, and these estimates often lead to test statistics. Several nonlinear rank tests based on the sample path behavior of statistics from $\mathcal{K}^+$ have been proposed; some generalize the classical **Kolmogorov–Smirnov test**. There is an extensive literature on **sequential** methods that can be used with censored data linear rank tests in clinical trials and other prospective studies (*see* **Interim Analysis of Censored Data**). Finally,

many more linear rank tests have been proposed than just those discussed in this article.

References

[1] Aalen, O.O. (1975). Statistical Inference for a Family of Counting Processes, *Ph.D. Dissertation*. University of California, Berkeley.

[2] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[3] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1982). Linear nonparametric tests for comparison of counting processes with application to censored survival data (with discussion), *International Statistical Review* **50**, 219–258.

[4] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[5] Breslow, N.E. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship, *Biometrika* **57**, 579–594.

[6] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[7] Cuzick, J. (1985). Asymptotic properties of censored linear rank tests, *Annals of Statistics* **13**, 133–141.

[8] Efron, B. (1967). The two-sample problem with censored data, *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics & Probability*, Vol. 4, Prentice Hall, New York, pp. 831–853.

[9] Elliot, R.J. (1982). *Stochastic Calculus and Applications*. Springer-Verlag, New York.

[10] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[11] Fleming, T.R., O'Fallon, J.R., O'Brien, P.C. & Harrington, D.P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right censored data, *Biometrics* **36**, 607–626.

[12] Gastwirth, J.L. (1985). The use of maximum efficiency robust tests in combining contingency tables and survival analysis, *Journal of the American Statistical Association* **80**, 380–384.

[13] Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrary singly-censored samples, *Biometrika* **52**, 203–223.

[14] Gill, R.D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.

[15] Gill, R.D. (1984). Understanding Cox's regression model: A martingale approach, *Journal of the American Statistical Association* **79**, 441–447.

[16] Gray, R.J. & Tsiatis, A.A. (1989). A linear rank test for use when the main interest is in differences in cure rates, *Biometrics* **45**, 899–904.

[17] Hájek, J. & Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

[18] Harrington, D.P. & Fleming, T.R. (1982). A class of rank test procedures for censored survival data, *Biometrika* **69**, 133–143.

[19] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimator from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[20] Latta, R.B. (1981). A Monte Carlo study of some two-sample rank tests with censored data, *Journal of the American Statistical Association* **76**, 713–719.

[21] Lee, E.T., Desu, M.M. & Gehan, E.A. (1973). A Monte Carlo study of the power of some two-sample tests, *Biometrika* **62**, 425–432.

[22] Leurgans, S. (1983). Three classes of censored data rank tests: Strengths and weaknesses under censoring, *Biometrika* **70**, 651–658.

[23] Leurgans, S. (1984). Asymptotic behavior of two-sample rank tests in the presence of random censoring, *Annals of Statistics* **12**, 572–589.

[24] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports* **50**, 163–170.

[25] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series A* **135**, 185–206.

[26] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika* **65**, 167–179.

[27] Prentice, R.L. & Marek, P. (1979). A qualitative discrepancy between censored data rank tests, *Biometrics* **35**, 861–867.

[28] Randles, R.H. & Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

[29] Rebolledo, R. (1980). Central limit theorems for local martingales, *Zeitschrift für Wahrlichtkeitstheorie und Verwandte Gebeite* **51**, 269–286.

[30] Robins, J. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring, in *AIDS Epidemiology: Methodologic Issues*, N. Jewell, K. Dietz & V. Farewell, eds. Birkhauser, New York, pp. 297–331.

[31] Schoenfeld, D.A. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions, *Biometrika* **68**, 316–319.

[32] Tarone, R.E. & Ware, J. (1977). On distribution free tests for equality of survival distributions, *Biometrika* **64**, 156–160.

[33] Woolson, R.F. (1981). Rank tests and a one-sample logrank test for comparing observed survival data to a standard population, *Biometrics* **37**, 687–696.

(*See also* **Survival Analysis, Overview**)

DAVID HARRINGTON

Linear Regression, Simple

Historically, the term "regression" was introduced by **Galton** [3, p. 246] to describe the tendency for the offspring of seeds "to be always more mediocre [i.e. more average] than their parent seeds . . .. The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it". Pearson & Lee [4] subsequently collected data on the heights of 1078 father–son pairs in order to study Galton's "law of universal regression" which they summarized as "Each peculiarity in a man is shared by his kinsmen, but on the average in a less degree" (*see* **Regression to the Mean**).

Modern applications rarely involve the element of "regression" as Galton meant it; however, the word is now too established to change. Consequently, regression now describes any relationship between a **response** (dependent, outcome) variable, Y, and a **covariate** or **explanatory** (independent, predictor) variable, X. Strictly speaking, only the response, Y, is assumed to vary randomly; however, in many applications the observed values of X are not known or fixed. We assume that any inherent variation in the measurement of X can be ignored. If this is not the case, we strongly advise resorting to methods that are appropriate when there is measurement error in an explanatory variable (*see* **Errors in Variables**). Simple linear regression involves finding the best-fitting curve that relates $\mathrm{E}(Y|X)$, the mean value of Y given X, and X, using an equation with a suitable functional form, such as $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$. This regression equation is called linear because $\mathrm{E}(Y|X)$ is a linear (straight-line) function with respect to the unknown model parameters, β_0 and β_1. It is not essential that $\mathrm{E}(Y|X)$ also depend linearly on X, although this is frequently the case in applications. For example, the model $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X^2$ describes a linear regression model that is a straight-line function of β_0 and β_1, but is quadratic in X (*see* **Polynomial Regression**). Whatever the model form, the goals of regression modeling are

1. to determine whether Y and X are associated in some systematic way; and/or
2. to estimate or **predict** the value of Y, or its mean, corresponding to a known value of X.

The unknown parameters, β_0 and β_1, are estimated from data – ordered pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ – using the method of **least squares**, which was discovered independently by **Gauss** and Legendre; see Plackett [5].

Estimating β_0 and β_1

Before fitting a linear regression model to data, it is wise to examine a scatterplot of Y vs. X in order to ensure that the proposed relationship is a sensible one (*see* **Graphical Displays**). Such a scatterplot is shown in Figure 1 for measurements of systolic blood pressure and age obtained from 21 males between the ages of 25 and 80. For these data, the notion that average systolic blood pressure increases systematically, in a roughly linear manner, with age seems plausible.

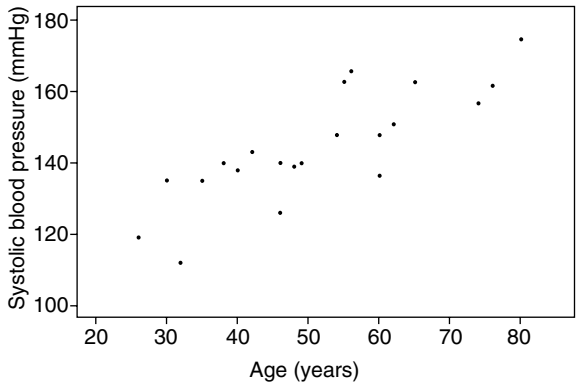

Figure 1 A scatterplot of systolic blood pressure (Y) vs. age (X) for a sample of 21 males between 25 and 80 years old

All linear regression models consist of a systematic component – the model equation, $E(Y|X) = \beta_0 + \beta_1 X$ – and a residual (random, error) component, ε; the sum, $\beta_0 + \beta_1 X + \varepsilon$, constitutes the regression model for Y. The **residual**, $\varepsilon = Y - \beta_0 - \beta_1 X$, represents the amount by which an observed value of Y deviates from the predicted mean, $\beta_0 + \beta_1 X$. Not all (X_j, Y_j) pairs for a given set of data will lie on the predicted line (curve). The method of least squares identifies the unique values of β_0 and β_1 that minimize the average of the squared residuals. Specifically, $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, where $\overline{x} = \sum_{i=1}^{n} x_i/n$, $\overline{y} = \sum_{i=1}^{n} y_i/n$, $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, and $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$. The equation of the estimated regression of Y on X is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = \overline{y} + \hat{\beta}_1 (X - \overline{x}).$$

Least squares estimates can be derived based only on the assumptions that the residuals, $\varepsilon_1, \ldots, \varepsilon_n$, are uncorrelated and have a mean value of zero and constant variance, σ^2. We use the estimated residuals $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i, i = 1, \ldots, n$, to estimate σ^2. The formula

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

which involves $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, the estimated residual sum of squares, emphasizes that two parameters, β_0 and β_1, are estimated; hence the divisor $n - 2$. Adopting the additional assumption that the residuals are **normally distributed** gives rise to various statistical procedures that we will discuss subsequently. First, however, we examine linear regression as an explanation for the observed variability in the response, Y.

Partitioning the Variability in Y

To account for the variability in Y, we can always resort to the simplest explanation, namely that Y varies about a fixed mean, μ. The corresponding linear regression model is $Y_i = \beta_0 + \varepsilon_i$, where $\beta_0 = \mu$. In this case, the residuals, i.e. $\varepsilon_i = Y_i - \beta_0 = Y_i - \mu$, are usually large, and result in a substantial estimate of σ^2. For the data concerning blood pressure and age, these estimated residuals are shown in Figure 2(a). In the absence of additional information, this is the only explanation we can devise for the observed variability in Y.

However, when Y appears to depend systematically on X, we can use the known values of X that were measured concurrently with Y to estimate $E(Y|X)$. Using $\hat{Y} = E(Y|X)$, we can partition the observed variability in Y into two components – the change in $E(Y|X)$ accounted for by the change in X, and the residual variability of Y values that have the same value of X, and hence the same value of $E(Y|X)$. These two components of variability correspond to the systematic component, $\beta_0 + \beta_1 X$, and the residual component, ε, respectively, in the linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$. By estimating $E(Y|X) = \beta_0 + \beta_1 X$, we can reduce the estimated residuals and hence the estimate of σ^2, the residual variation. Clearly, the estimated residual sum of squares in Figure 2(b) will be much less than the corresponding value based on the estimated residuals in Figure 2(a). Equivalently, knowing a subject's age provides important information about what his blood pressure is likely to be.

This partition of the variability in Y is usually summarized in an **analysis of variance** (ANOVA) table, such as the one corresponding to the example shown in Table 1. This partitioning is represented by the equation

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

which is alternatively described by the relationship

$$\text{total sum of squares} = \text{model sum of squares} + \text{residual sum of squares.}$$

The ratio of the model sum of squares to the total sum of squares is called R^2, and represents the proportion of the observed variability in Y that is accounted for by modeling the mean response for Y as the function, $\beta_0 + \beta_1 X$, of X.

Table 1 An ANOVA table summarizing the partition of observed variability in blood pressure measurements into the systematic (model) component represented by the estimated regression model, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, and the residual component. The value of R^2 for these data is 0.69

Source	SS	df	MS
Model	3453.5	1	3453.5
Residual	1545.8	19	81.36
Total	4999.3	20	

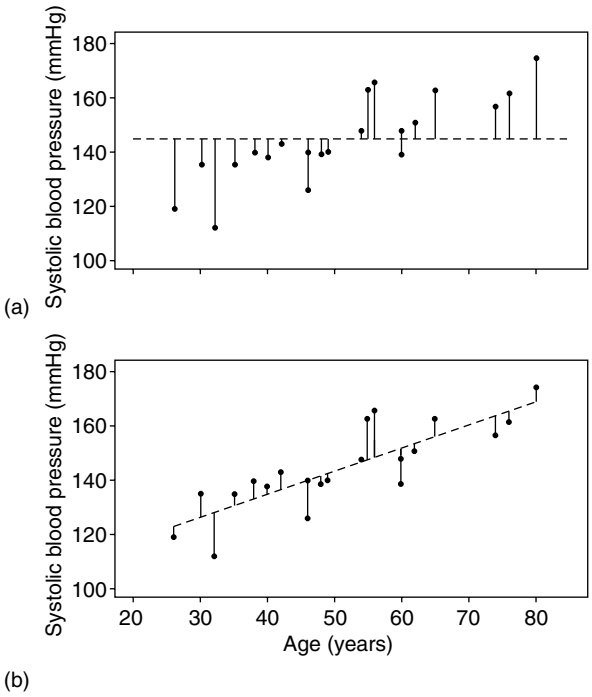

Figure 2 Estimated residuals (solid lines) for systolic blood pressure measurements (Y) in a sample of 21 males between 25 and 80 years old. (a) Based on the simple model, $Y = \mu + \varepsilon$; the dashed line indicates $\hat{\mu} = 144.7$. (b) Based on the linear regression model, $Y = \beta_0 + \beta_1 X + \varepsilon$, relating Y and age (X); the dashed line indicates the least squares estimated regression equation $\hat{Y} = 100.6 + 0.86X$

Interpreting the Regression Estimates

If two values of the explanatory variable differ by one unit, the corresponding values of the model equation differ by β_1. Therefore, $\hat{\beta}_1$ represents the estimated change in the mean response associated with a unit increase in the explanatory variable. Of course, this estimate and its interpretation are only valid within the range of X values used in fitting the linear regression model.

The value β_0 represents the mean response when $X = 0$. Frequently, this mean response may be of no interest, or may not belong to the range of X values used in fitting the model to data. A more useful parameter in many situations is $\gamma = \beta_0 + \beta_1 \overline{x}$, which represents the mean response when $X = \overline{x}$, the observed average of the explanatory variable. The estimated value is $\hat{\gamma} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y}$, the sample mean of Y.

The estimates of β_0 and β_1 for the example are 100.6 and 0.86, respectively. From these data we conclude that 0.86 mmHg is the estimated increase in mean systolic blood pressure associated with each additional year of age for men 25–80 years old. At an age of $\overline{x} = 51.1$ years, the estimated mean value is $\hat{\gamma} = \overline{y} = 145$ mmHg.

Statistical Inference in Linear Regression

Under the additional assumption that the residuals, $\varepsilon_1, \ldots, \varepsilon_n$, are normally distributed, the estimators of β_0 and β_1 have normal sampling distributions. Estimated **standard errors** (est. se) for $\hat{\beta}_0$ and $\hat{\beta}_1$ are routinely produced by most computing packages (*see* **Software, Biostatistical**). The ratio of each difference, $\hat{\beta}_0 - \beta_0 \hat{\beta}_1 - \beta_0$, to its corresponding estimated standard error follows a **Student's *t* distribution** with $n - 2$ **degrees of freedom** (df). From these results, **hypothesis tests** and/or **confidence intervals** for β_0 and β_1 can be evaluated. Likewise, the **sampling distribution** of $\hat{\gamma} = \overline{Y}$ is

normal, and the corresponding estimated standard error is $s/\sqrt{n}$, where $s^2 = \hat{\sigma}^2$.

A test of the null hypothesis, $H_0 : \beta_1 = 0$, is routinely used to assess the significance of the regression; that is, to determine whether the data constitute statistical evidence of an association between Y and X. This test can be based either on the ratio $\hat{\beta}_1$/est. se$(\hat{\beta}_1)$, which has a Student's t distribution with $n - 2$ df, or on $[\hat{\beta}_1/\text{est. se}(\hat{\beta}_1)]^2$, which has an ***F* distribution** with 1 and $n - 2$ df. The latter test statistic is equal to the ratio of the model mean square to the residual mean square in the corresponding ANOVA table for the regression analysis, and usually appears in an additional column labelled F ratio.

For the blood pressure vs. age example, the estimated standard errors for $\hat{\beta}_0, \hat{\beta}_1$ are 7.05 and 0.13, respectively. The 95% confidence interval for β_1 is (0.58, 1.14), and for γ the interval is (136, 153).

Model Diagnostics

A fitted regression model and associated statistical inferences are based on various assumptions concerning the functional form of the model for $\mathrm{E}(Y|X)$ and distributional properties of the residuals. Violations of these assumptions may invalidate conclusions based on the regression analysis. Therefore, it is essential to check these assumptions, using various types of **diagnostic** plots.

The estimated residuals, $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$, $i = 1, \ldots, n$, play an essential role in model diagnostics. Many computer packages offer the option of using these ordinary residuals or the corresponding standardized or studentized residuals, which have a common variance. Use of either of the latter two is preferable, since the $\hat{\varepsilon}_i$s do not all have the same variance.

The following diagnostic plots furnish graphical evidence that one or more of the model assumptions may be contradicted by the data:

1. Residuals vs. the fitted values, $\hat{Y}_i$. An unsuitable functional form is usually revealed by the systematic appearance of this plot, as is nonconstant variance.
2. Residuals vs. the explanatory variable, X_i. Systematic patterns in this plot can indicate violations of the mean 0, constant variance assumptions, or inappropriate model form.
3. Normal probability plot of the residuals. This plot checks the normal distribution assumption from which all the statistical inference procedures arise (*see* **Normal Scores**).
4. Residuals vs. the temporal/spatial order of data collection. Unexpected regularity in this plot suggests that the Y_is may be correlated. To prepare this diagnostic check, it is essential to record the temporal/spatial ordering when data are first collected.
5. Index plots (plot against the case number, i.e. the observation label) of the leverages and Cook's distance (*see* **Normal Scores**). The former are a measure of the amount of influence exerted on $\hat{Y}_i$ by the corresponding observed response, Y_i. Cook's distance is a summary measure of the influence that each case exerts on the estimated regression coefficients. These two diagnostic plots can reveal outliers (values of Y that are anomalous with respect to the rest of the data) or influential points (values of (X_j, Y_j) that strongly influence the estimated values of $\hat{\beta}_0, \hat{\beta}_1$ and s^2).

Deviations from the expected (null) pattern in any of these plots may indicate problems that require further investigation or remedial action. For additional details concerning model diagnostics, see Belsley et al. [1] or Cook & Weisberg [2]. Further details concerning examination of the adequacy of a fitted regression model are found in the article on **Goodness of Fit**.

References

[1] Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

[2] Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London.

[3] Galton, F. (1885). Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute* **15**, 246–263.

[4] Pearson, K. & Lee, A. (1903). On the laws of inheritance in man. I. Inheritance of physical characters, *Biometrika* **2**, 357–462.

[5] Plackett, R.L. (1972). Studies in the history of probability and statistics, XXIX: the discovery of the method of least squares, *Biometrika* **59**, 239–251.

(*See also* **Multiple Linear Regression**)

DAVID E. MATTHEWS

Linearization Methods of Variance Estimation

The **variance** of a linear function of variables is a linear function of variables. An approximation by a linear function of a nonlinear function enables one to derive an approximate variance of a complex nonlinear function. The most common approach consists of taking linear terms of Taylor series expansion of the nonlinear function of the observations around their expected values. The approach has been widely used for approximating large-sample variances (*see* **Large-sample Theory**), and is referred to as Taylor Series Linearization or the **Delta method**. Another linearization approach was suggested by Quenouille [6] and made well known by Tukey [10] as **jackknife**. A good review of the jackknife and other methods appears in [7].

Tepping [9] suggested the use of Taylor series linearization for estimating variances in complex sample surveys. Applications to **mean** and **linear regression** coefficients for complex surveys were presented by Kish & Frankel [4] and Folsom [3]. Some **simulation** results were presented by Shah et al. [8]. Woodruff [11] presented a general application of the linearization method to explicit functions of observed data. Binder [1] extended the results to parameters defined as implicit functions or estimating equations. Binder also proved the asymptotic **normality** of the estimates. Binder [2] presented an application of Taylor series linearization to the estimation of parameters for Cox's **proportional hazard** model for the survival data collected from a complex sample survey. We present here a brief summary of the results by Woodruff [11] and Binder [1].

The Taylor series linearization method is illustrated here for statistics that can be defined explicitly as functions of linear statistics estimated from a survey sample. Means, totals, proportions, general ratios of the form $\sum wx / \sum wy$, and linear regression coefficients all fall into this category of functions. A linearized variable, Z_i, is defined on the basis of the Taylor series expansion of the function, and then substituted into the variance formula appropriate under the specified design for any linear statistic estimated from the sample.

The technique will be illustrated for a statistic which is a function of two linear statistics, although it extends to any number of linear statistics and to statistics that are vectors. Let $\hat{\theta}$ be an estimate of the population parameter θ, with $\hat{\theta} = F(X, Y)$ where X and Y are two linear sample statistics. Let $\mu_x = \mathrm{E}(X)$ and $\mu_y = \mathrm{E}(Y)$, where the **expectation** operator E denotes averaging over repeated sampling from the **target population**. θ can be expanded in a Taylor series about μ_x and μ_y, so that

$$\begin{aligned}\hat{\theta} = {} & F(\mu_x, \mu_y) + \partial F_x(\mu_x, \mu_y)(X - \mu_x) \\ & + \partial F_y(\mu_x, \mu_y)(Y - \mu_y) \\ & + \text{higher order terms},\end{aligned}$$

where the $\partial F_x(\mu_x, \mu_y)$ and $\partial F_y(\mu_x, \mu_y)$ functions are first-order partial derivatives of F with respect to X and Y evaluated at their respective expectations, μ_x and μ_x. If the higher order terms are negligible, then

$$\begin{aligned}\mathrm{var}[\hat{\theta}] &\doteq \mathrm{E}[\hat{\theta} - F(\mu_x, \mu_y)]^2 \\ &= \{(\partial F_x)^2 \mathrm{E}(X - \mu_x)^2 + (\partial F_y)^2 \mathrm{E}(Y - \mu_y)^2 \\ &\quad + 2(\partial F_x)(\partial F_y) \\ &\quad \times \mathrm{E}[(X - \mu_x)(Y - \mu_y)](Y - \mu_y)\} \\ &\quad \times \{(\partial F_x)^2 \mathrm{var}(X) + (\partial F_y)^2 \mathrm{var}(Y) \\ &\quad + 2(\partial F_x)(\partial F_y)\mathrm{cov}(X, Y)\},\end{aligned}$$

where $\partial F_x = \partial F_x(\mu_x, \mu_y)$ and $\partial F_y = \partial F_y(\mu_x, \mu_y)$.

An equivalent computational procedure for producing the Taylor series variance estimate suggested by Woodruff [11] recognizes that the variable portion of the linearization in his Eq. (3.2) is

$$Z = (\partial F_x)X + (\partial F_y)Y,$$

and therefore

$$\begin{aligned}\mathrm{var}[\hat{\theta}] &\doteq \mathrm{var}[(\partial F_x)X + (\partial F_y)Y] \\ &= \mathrm{var}(Z).\end{aligned}$$

Noting that X and Y are linear statistics formed from the corresponding response variates x_i and y_i, measured on the ith sample unit, the variance approximation can be produced by substituting the linearized variable

$$Z_i = (\partial F_x)X_i + (\partial F_y)Y_i$$

for x_i or y_i in the variance formula appropriate for computing var(X) or var(Y) under the specified sample design. To obtain a sample estimate for the Taylor series variance approximation, one replaces the population-evaluated derivative functions in Z_i with the corresponding sample analog, i.e.

$$Z_i = [\partial F_x(X, Y)]x_i + [\partial F_y(X, Y)]y_i.$$

Binder [1, 2] proposed and justified using an implicit differentiation method for estimating the variance for a vector of survey statistics. Binder's results are particularly useful when the parameters are implicitly defined, but the results also cover the explicit case.

Logistic regression coefficients and survival models (*see* **Survival Analysis, Overview**) fall into this category of parameters that are implicitly defined.

Let $\theta = (\theta_1, \ldots, \theta_p)'$ be the finite population parameter vector which is defined by

$$W(\theta) = \sum_{k=1}^{N} U(Z_k; \theta) - v(\theta) = 0,$$

where $Z_k = (z_{1k}, \ldots, z_{qk})$ are the data values for the kth unit, and $W(\theta)$ is a vector with the ith element:

$$W_i(\theta) = \sum_{k=1}^{N} V_i(Z_k; \theta) - v_i(\theta) = 0.$$

Let $U(\theta) = \sum_{k=1}^{N} U(Z_k; \theta)$ be estimated from the sample by $\hat{U}(\theta)$. $\hat{U}(\theta)$ is the estimator of the total based on the functions of data values $U(Z_1; \theta), \ldots, U(Z_n; \theta)$, for example $\hat{U}(\theta) = \sum_{i \in S} w_i U(Z_i; \theta)$. Then, $\hat{W}(\theta) = \hat{U}(\theta) - v(\theta)$. Assuming that a unique solution exists, $\hat{\theta}$, the estimate of θ, is defined as the solution to

$$\hat{W}(\hat{\theta}) = 0.$$

To approximate the variance of $\hat{\theta}$, Binder expands $\hat{W}(\hat{\theta})$ in a Taylor series about the point $\hat{\theta} = \theta$, where θ is the true unknown parameter. Defining $\hat{J}(\theta) = \partial \hat{W}(\theta)/\partial\theta$ as the $p \times p$ matrix whose ij element is the partial derivative $\partial \hat{W}_i(\theta)/\partial\theta_j$, and expanding $\hat{W}(\hat{\theta})$ about $\hat{\theta} = \theta$, gives

$$0 = \hat{W}(\hat{\theta}) \approx \hat{W}(\theta) + \hat{J}(\theta)(\hat{\theta} - \theta),$$

or, if $\hat{J}^{-1}(\theta)$ exists,

$$\hat{\theta} - \theta \doteq \hat{J}^{-1}(\theta)\hat{W}(\theta).$$

This leads to the approximation of the variance matrix of $\hat{\theta}$:

$$V(\hat{\theta}) \doteq [\hat{J}^{-1}(\theta)][V(\hat{W}(\theta))][\hat{J}^{-1}(\theta)]',$$

where $V(\hat{W}(\theta))$ is the **covariance matrix** of $\hat{W}(\theta)$. Finally, θ is replaced by its estimator $\hat{\theta}$, in both $\hat{J}^{-1}(\theta)$ and $V(\hat{W}(\theta))$ to obtain the estimator of the covariance matrix of $\hat{\theta}$:

$$\hat{V}(\hat{\theta}) = [\hat{J}^{-1}(\hat{\theta})][\hat{V}(\hat{W}(\hat{\theta}))][\hat{J}^{-1}(\hat{\theta})]'.$$

Binder [1] gives regularity conditions that are needed to ensure the asymptotic normality of the parameters $W_i(\hat{\theta})$ and the consistency of $\hat{V}(\hat{\theta})$. These conditions include:

1. the existence of a parameter space that contains a neighborhood of the parameter θ;
2. the existence of a sequence of sample designs and populations which admits asymptotically normal estimators for certain population totals and consistent estimators for the variance of the estimate of the totals: in particular, $\hat{W}(\theta)$ is approximately normally distributed for fixed θ;
3. some continuity and limiting conditions on $W(\theta)$ and its partial derivatives, and a continuity condition on the variance of the estimated total.

Furthermore, Binder [1, Corollary 2] shows that the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is the same as the asymptotic distribution of a **random variable** that is **multivariate normal** with mean zero and variance–covariance matrix $n[\hat{J}^{-1}(\hat{\theta})][\hat{V}\hat{W}(\hat{\theta})][\hat{J}^{-1}(\hat{\theta})]'$.

The question is often raised as to how good the linearization method is compared with other alternatives. There are basically three major competing methods: balanced repeated replication (BRR), jackknife, and **bootstrap** methods.

The drawbacks of the linearization methods are:

1. Linearization methods require computation of derivatives, and hence are more difficult to program than BRR or jackknife. Linearization methods are also limited to smooth functions of observations.
2. The impact of weight adjustments, such as **post-stratification** or **nonresponse**, on variance estimation is difficult to account for, and is often ignored in most implementations.

The advantages for the linearization method are:

1. They require substantially less computational resources than jackknife and BRR and are most suitable for large datasets.
2. They are applicable to a large number of situations, and can be applied to **multistage** designs with or without replacements. BRR and jackknife are somewhat limited in this respect.

Krewski & Rao [5], and Rao & Wu [7] have compared the three methods. Linearization methods are less **biased** and more stable than BRR or jackknife methods. Asymptotically, all of the methods provide **consistent estimators** of the variances, and hence the differences between them get smaller as the sample size increases. Overall there are no compelling reasons to choose one method over the others, and the decision to select a method should be based on convenience and available software. Currently, three software packages have implemented variance estimation using the linearization method. These are: SUDAAN, by Research Triangle Institute; PCCARP, by Iowa State University; and STATA, by Stata Corporation (*see* **Software, Biostatistical**).

References

[1] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review* **51**, 279–292.

[2] Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data, *Biometrika* **79**, 139–147.

[3] Folsom, R.E. (1974). *National Assessment Approach to Sampling Error Estimation, Sampling Error Monograph*, Prepared for National Assessment of Educational Progress, Denver.

[4] Kish, L. & Frankel, M.R. (1974). Inference from complex surveys, *Journal of Royal Statistical Society, Series B* **36**, 1–37.

[5] Krewski, D. & Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods, *Annals of Statistics* **9**, 25–45.

[6] Quenouille, M.H. (1956). Note on bias in estimation, *Biometrika* **43**, 353–360.

[7] Rao, J.N.K. & Wu, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for nonlinear statistics, *Journal of the American Statistical Association* **80**, 620–630.

[8] Shah, B.V., Holt, M.M. & Folsom, R.E. (1977). Inference about regression models from survey data, *Bulletin of the International Statistical Institute* **47**, 43–57.

[9] Tepping, B.J. (1968). The estimation of variance in complex surveys, in *American Statistical Association 1968 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 11–18.

[10] Tukey, J.W. (1958). Bias and confidence in not quite large samples, *Annals of Mathematical Statistics* **29**, 614.

[11] Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association* **66**, 411–414.

BABUBHAI V. SHAH

Linkage Analysis, Model-based

In the field of **human genetics**, the main goal has been to identify the **genes** that are responsible for various phenotypes (*see* **Genotype**) – typically diseases. The primary method for doing so has been through looking at phenotypically silent marker loci that are randomly distributed in the genome (*see* **Polymorphism**), and trying to determine which such loci have alleles that tend to cosegregate in families with the trait of interest. In linkage analysis one tries to estimate the frequency with which the marker and disease gene segregate together in families. When we talk of model-based linkage analysis, it is assumed that one can fully describe the mode of action of the disease gene, i.e. its allele frequency (*see* **Gene Frequency Estimation**) and the **penetrances** for each disease locus genotype (*see* **Gene**). When one does not know these quantities accurately, one often applies model-free linkage analysis methods (*see* **Linkage Analysis, Model-free**), which effectively correspond to special cases of model-based linkage analysis. Similarly, it can be shown that other methods for detecting correlations between a marker locus and disease gene on a population level through **linkage disequilibrium** analysis can also be considered as special cases of model-based linkage analysis. For this reason, it is important to understand the principles of model-based linkage analysis, because their principles are behind all statistical gene mapping techniques (*see* **Genetic Map Functions**).

Biological Basis of Linkage

The human genetic material is composed of large linear units of deoxyribonucleic acid (DNA) called chromosomes, each of which contains a long linear sequence of genes. These genes are **DNA sequences** that tell the cell how to construct a specific protein, and thus these genes provide the blueprint from which a person is assembled. There are 22 pairs of chromosomes in each human cell, plus a pair of sex chromosomes (X and Y) which determine the sex of an individual (XX individuals are female, and XY individuals are male). Each person receives one copy of each chromosome from his mother and one copy from his father, and thus each person has 50% of his DNA inherited from each parent. Since 50% of one's genes come from each parent, there is a **correlation** between related individuals at the phenotypic level, and thus the common observation that certain diseases "run in families". The male sperm cell and the female egg cell each contain one copy of each chromosome (haploid state) from the father and mother, respectively, instead of two copies of each chromosome as in a normal somatic cell (diploid state). When the sperm and egg combine to form a new child, this infant again has two copies of each chromosome. The important step in determining the genetic makeup of the new child is to look at how the single copy of each chromosome is selected for each gamete. The process by which these haploid gametes are generated is called meiosis. If we label the two copies of chromosome N that a given parent has as N_a and N_b (where this individual received chromosome N_a from his father and chromosome N_b from his mother), then **Mendel's laws** dictate that which copy of each chromosome is transmitted to any given gamete is determined at random, and that each chromosome is inherited independently of every other. Thus, the probability that a gamete receives chromosome 1_b equals the probability that it receives 1_a, and the same holds for chromosomes 2,3, and so on. In this simplified model of inheritance, if an allele located somewhere on chromosome 1_a was received by a given gamete, then the probability of another allele on chromosome 1_a also being inherited would be 1, while the probability of an allele located on chromosome 2_a being inherited would be 0.5. If life were this simple, then we could easily test whether or not two genes were on the same chromosome (syntenic) by checking whether any gamete ever received the a allele at one gene and the b allele at the other. Under the **null hypothesis** that two genes are on different chromosomes, 50% of gametes would receive the a allele at one gene and the b allele at the other, and it would thus be very easy to test for synteny.

In human meiosis, however, it is not as simple as this. There is an additional source of variation in the genetic material. During meiosis the two copies of each chromosome line up next to each other and undergo a random process called recombination in which the two copies of each chromosome can exchange their genetic material with each other (as illustrated in Figure 1). In fact, there may be many

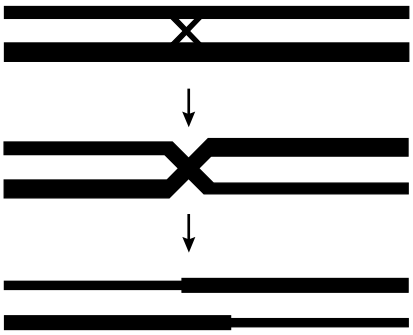

Figure 1 Pictorial representation of recombination. The (thick and thin) lines represent homologous chromosomes during meiosis; the black X in the top of the figure represents a recombination event, one outcome of which is indicated in the figure

such crossover events per chromosome in each meiotic event, and thus there is a great deal of variation between even the most tightly related people. When an odd number of crossover events occur between two genes, the alleles from different ancestral chromosome are received (i.e. allele a at one locus and allele b at the other) – in this situation we say that a recombination of the genetic material has occurred between these two loci – the combination of alleles a at one locus and b at the other is a new combination that was not present on either parental chromosome. If an even number of crossovers occurs between two genes, then the same combination of alleles as in the parents is present in the gamete – this is termed a nonrecombination. If two loci are on different chromosome, they recombine with probability 50% (since, as was indicated earlier, given allele a was inherited at a given point on chromosome 1, the probability of an allele on chromosome 2 being present in the gamete in its b form is 50%). Similarly, if two loci are very far apart on the same chromosome (and we assume an absence of chromatid interference) they also recombine with probability 50%, but when two loci are very close together on the same chromosome, the probability of a recombination between them tends toward 0 as the distance between them decreases. Thus it becomes possible to devise a means of testing for linkage between two loci by looking at whether they recombine with probability 50% or less.

There are a large number of loci spread randomly throughout the human genome called marker loci that have no known function, but that are very variable from person to person, and whose positions in the genome are well known. It is straightforward to determine the genotype of any individual at these polymorphic DNA sequence variations, and thus we can look for genes with unknown position in the genome by testing whether the gene of interest recombines with a marker locus with probability less than 50%. Since there are marker loci spread throughout the genome, at least one of them should be linked to any new gene we wish to isolate. The goal of linkage analysis is to identify marker loci that recombine with our trait locus with low probability, such that we may significantly narrow down the portion of the total genome where this gene can be found by subsequent labor-intensive molecular analysis. The closer we can get to the gene through the simple process of linkage analysis, the easier it will be to identify, though typically it is impossible to get within less than 2 cM – approximately 2 000 000 base pairs – through linkage analysis with a disease having a known mode of inheritance, while for complex traits with an unknown mode of inheritance, it may be difficult to get within less than 10 cM (10 000 000 base pairs), further complicating the subsequent molecular analyses required. The next Sections introduce the mathematical techniques employed in the testing and estimation of linkage in humans, starting from the simplest situations and continuing through to the general case.

Lod Score Analysis

The most commonly employed statistic in human genetic linkage analysis is based on the principle of **maximum likelihood**. In this, we compute the probability of the observed data under different assumptions about the unknown parameter – here the recombination fraction (typically denoted as θ). The null hypothesis is that $\theta = 0.5$, and the alternative is that $\theta < 0.5$, so the test statistic is based on the maximum of the **likelihood ratio** $\Lambda(\theta) = L(\theta)/L(\theta = \frac{1}{2})$. For purely historical reasons, the lod score function commonly used is $Z(\theta) = \log_{10}[\Lambda(\theta)]$; however, the quantity $2 \ln \Lambda$ is much more theoretically pleasing since, asymptotically, $\max_\theta[2 \ln \Lambda(\theta)](= \Lambda_{\max})$ is distributed as a mixture of a 50% point mass at 0 and 50% $\chi^2_{(1)}$ – the 50% point mass at zero because of the one-sided alternative, $\theta < 0.5$ (if $\hat{\theta} = 0.5$, $\Lambda_{\max} = 0$) and other arguments (cf. [15], [20], and [7]). The

conventional critical value used in linkage analysis for calling a test significant is $Z \geq 3$, which corresponds to a ***P* value** of 0.0001 if we assume the distribution given above. The reason for insisting on such a small P value is due to the multiple testing employed (*see* **Multiple Comparisons**) if one were to conduct a full genome scan with many markers, and the low prior probability of linkage to a randomly selected marker locus (see [6], [11], [16], and [21] for more details).

Counting Recombinants and Nonrecombinants – Phase-known Pedigrees

In some situations it is possible to directly observe whether a recombination occurred or not between two loci in a given meiosis. This is commonly the case in animal crosses, where you can breed animals to the point where you know which of their offspring are the results of recombinant meioses and which are nonrecombinant. Let us assume that our data are in the form of genotypes, so we can write the pedigree **likelihood** as

$$L(\theta) = \Pr(\text{genotypes}_{\text{parents}}) \times \Pr(\text{genotypes}_{\text{offspring}}|\text{genotypes}_{\text{parents}}),$$

and the lod score can then be written as

$$\log_{10}\frac{L(\theta)}{L\left(\theta=\frac{1}{2}\right)} \log_{10} = \frac{\Pr(\text{genotypes}_{\text{parents}}) \times \Pr(\text{genotypes}_{\text{offspring}}|\text{genotypes}_{\text{parents}};\theta)}{\Pr(\text{genotypes}_{\text{parents}}) \times \Pr\left(\text{genotypes}_{\text{offspring}}|\text{genotypes}_{\text{parents}};\theta=\frac{1}{2}\right)}.$$

Note that in the ratio we can factor out Pr(genotypes$_{\text{parents}}$), since this factor is independent of the recombination fraction, as the genotypes are known and identical in numerator and denominator. Therefore, assuming we can count the number, k, of recombinants out of n meioses (leaving $n-k$ nonrecombinants), the likelihood is simply $L(\theta) = \theta^k(1-\theta)^{n-k}$. The maximum likelihood estimate of the recombination fraction in this case is trivial to compute as $\hat{\theta} = k/n$ if $k < n/2$; $\hat{\theta} = 0.5$ if $k \geq n/2$ (since values of $\theta > 0.5$ are inadmissible). While this estimate of θ is biased (because of the inadmissibility of estimates of $\theta > 0.5$), it can be shown to be asymptotically unbiased (cf. [16]). It is possible to count recombinants and nonrecombinants in situations when the phase is known. When we say the phase is known we mean that it is possible to tell which alleles were inherited from each grandparent. Consider the pedigree shown in Figure 2; in that pedigree it is apparent that allele A at the first locus and allele B at the second locus in the father were inherited together from the grandfather; and similarly alleles a and b were inherited together from the grandmother. If the two loci are syntenic, it would mean they are on the same chromosome in the father. In this case, among the children all A B or a b haplotypes are nonrecombinants (parental types), while all children who received haplotypes A b or a B are recombinants (or nonparental types). In this example, there are four nonrecombinant children and two recombinants, so our lod score is computed as

$$Z(\theta) = \log_{10}\frac{\theta^2(1-\theta)^4}{(0.5)^2(0.5)^4} = \log_{10} 2^6\theta^2(1-\theta)^4.$$

In this pedigree, the maximum likelihood estimate of θ is $2/6 = 1/3$, so the maximum lod score is $Z(1/3) = 0.1475$. Note that the information coming

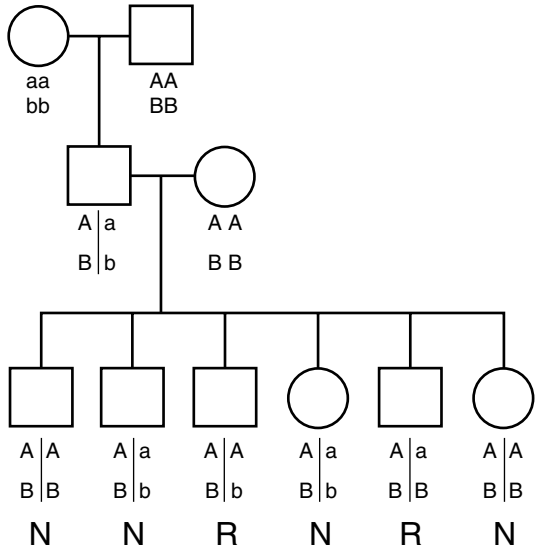

Figure 2 Sample phase-known pedigree with two codominant loci indicated–the first locus has two alleles (A and a) and the second locus has alleles B and b. Nonrecombinant meioses are indicated by "N" and recombinant meioses by "R"

from alleles inherited from the mother by the offspring was not included in this computation. This is because there is no linkage information coming from the mother as she is homozygous (*see* **Heterozygosity**) at both loci, and thus she transmits alleles A and B to all children with probability 1, independently of the recombination fraction. In fact, a parent has to be heterozygous at both loci to be informative for linkage. If the mother were Aa at the first locus and BB at the second, then every child would receive a B allele with probability 1, and at the other locus A is inherited with probability 0.5, independently of the recombination fraction (since the inheritance of the B allele is ubiquitous).

Lod Scores in Phase-unknown Pedigrees

Often there is some ambiguity about the parental phase – for example, if the grandparents were unavailable for genotyping. Consider the same pedigree without the grandparents having been typed (as shown in Figure 3). In this situation, there are two possible phases for the father – either he could have phase A B/a b or he could have phase A b/a B. A priori these two phases are equally likely (assuming an absence of linkage disequilibrium), so we either have four recombinants and two nonrecombinants or we have four nonrecombinants and two recombinants. Early human geneticists would throw these pedigrees away because it was thought that they contained no useful information about linkage. However, there is information in these families. The likelihood is computed as

$$\begin{aligned} L(\theta) &= \Pr(\text{parents})\,\Pr(\text{off}|\text{parents};\theta) \\ &= \sum_{\text{phases}} \Pr(\text{phase}_{\text{father}}) \\ &\quad \times \Pr(\text{genotypes}_{\text{off}}|\text{phase}_{\text{father}};\theta), \end{aligned}$$

since the only ambiguity in the parental genotypes is in the paternal phase. In this example, each phase has probability 0.5, so

$$\begin{aligned} L(\theta) &= \tfrac{1}{2}\theta^4(1-\theta)^2 + \tfrac{1}{2}\theta^2(1-\theta)^4 \\ &= \tfrac{1}{2}\theta^2(1-\theta)^2[\theta^2 + (1-\theta)^2]. \end{aligned}$$

Note that the maximum likelihood estimate (MLE) of θ is not trivial to compute by hand, although it can be estimated using numerical maximization techniques; using the ILINK program of the LINKAGE package [13] (*see* **Software for Genetic Epidemiology**), the MLE is found to be 0.5. Even though there are not 50% of meioses in this case showing evidence of recombination, as analyzed in detail by Nordheim et al. [15], there are an enormous number of potential phase-unknown pedigrees that all yield an estimate of $\theta = 0.5$. That is not to say that phase-unknown pedigrees do not provide information about linkage in general. Consider a phase-unknown pedigree with N children all of whom received the identical alleles at both loci – this would either represent N recombinants or N nonrecombinants. The likelihood would then be equal to $L(\theta) = \frac{1}{2}\theta^N + \frac{1}{2}(1-\theta)^N$, which can be shown to be maximized when $\theta = 0$, giving a lod score of

$$\begin{aligned} Z_{\text{PU}}(\theta = 0) &= \log_{10} \frac{\frac{1}{2}(0)^N + \frac{1}{2}(1)^N}{\frac{1}{2}\left(\frac{1}{2}\right)^N + \frac{1}{2}\left(\frac{1}{2}\right)^N} \\ &= \log_{10} \frac{\frac{1}{2}}{\left(\frac{1}{2}\right)^N} = (N-1)\log_{10}(2), \end{aligned}$$

where the subscript PU stands for phase-unknown. If the pedigree were phase-known with six nonrecombinants, then the likelihood would be $L(\theta) = (1 -$

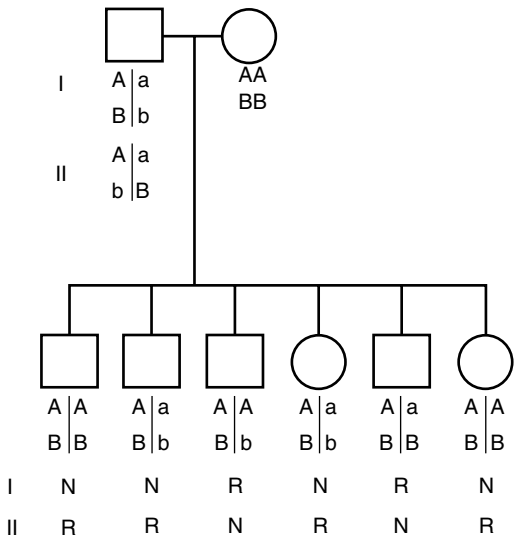

Figure 3 Sample phase-unknown pedigree with two possible phases for the father – indicated as I and II – under each of his offspring is an indication of whether the individual had a recombination or not between these two loci under each possible phase for the father

$\theta)^N$, which is also maximized when $\theta = 0$, but giving a lod score of $Z_{\mathrm{PK}}(\theta = 0) = \log_{10}(1)^N/\left(\frac{1}{2}\right)^N = N\log_{10}(2)$, where the subscript PK indicates phase-known. So, in each pedigree there is a cost of one meiosis when the phase is unknown if there is no recombination – when there is recombination in the pedigree the cost is even higher, as was illustrated by the previous example. It may seem like a small cost initially, but if you consider that the average sibship size is between two and three, the cost can be huge in a large set of pedigrees. The lod score in each phase-unknown pedigree is $Z_{\mathrm{PU}}(\theta = 0) = [(N-1)/N]Z_{\mathrm{PK}}(\theta = 0)$ when all sibs are nonrecombinant. Note that this is an upper bound on the phase-unknown lod score as a function of the phase-known lod score. In general, $Z_{\mathrm{PU}}(\hat{\theta}) \leq [(N-1)/N]Z_{\mathrm{PK}}(\tilde{\theta})$, where θ is estimated separately in the phase-known and phase-unknown cases, and there is linkage. In the US, most sibships are of size two or three, so that the lod scores in each pedigree are at least 1.5–2 times higher for phase-known pedigrees than for phase-unknown pedigrees, when there is linkage between the two loci. Since lod scores can be added across pedigrees at the same recombination fraction values (since all pedigrees are independent, and independent likelihoods can be multiplied), the sum over a large set of pedigrees will also be 1.5–2 times larger or more if there is linkage and the phase can be established unequivocally.

Genotypes Unknown

We are often confronted with the complication that not all individuals' genotypes are known. Remember that, when we know the genotypes of the parents, the pedigree likelihood for a phase-known nuclear pedigree is

$$L(\theta; \text{data}) = \Pr(g_{pa})\Pr(g_{ma}) \prod_{\text{offspring}} \times \Pr(g_{offs}|g_{ma}, g_{pa}; \theta),$$

where g_{ma} is the genotype of the mother, etc.; and in the case of unknown phase it is just

$$L(\text{data}) = \sum_{\text{phase}} \Pr(g_{ma}, \text{Phase}_{ma})\Pr(g_{pa}, \text{Phase}_{pa}) \times \prod_{\text{offspring}} P(g_{offs}|g_{pa}, g_{ma}, \text{Phase}_{pa}, \text{Phase}_{ma}; \theta).$$

If we were interested in the actual probability of the data, then we would have to use the allele frequencies at each marker locus in order to compute $\Pr(g_{pa})$ from the allele frequencies for the two loci. For example, if g_{ma} were AA at one locus and BB at the other, then $\Pr(g_{ma})$ would be Pr(A) Pr(A) Pr(B) Pr(B) if we assume **Hardy–Weinberg equilibrium** and absence of linkage disequilibrium. Note that, in the phase-known situation, the lod score is

$$Z(\theta) = \log_{10} \frac{\Pr(g_{ma})\Pr(g_{pa}) \times \prod_{\text{offspring}} \Pr(g_{offs}|g_{ma}, g_{pa}; \theta)}{\Pr(g_{ma})\Pr(g_{pa}) \times \prod_{\text{offspring}} \Pr\left(g_{offs}|g_{ma}, g_{pa}; \theta = \frac{1}{2}\right)},$$

and we do not need to worry about the exact values of the probability of the parental genotypes, since they can be factored out of this ratio and have no effect on the lod score – this makes sense because the parental genotypes tell us nothing in and of themselves about the recombination fraction. Similarly, in the phase-unknown case, we also know the genotypes of the parents, so we can factor the independent genotype and phase probabilities as $\Pr(g_{ma}, \text{Phase}_{ma}) = \Pr(g_{ma})\Pr(\text{Phase}_{ma})$, assuming absence of linkage disequilibrium, and so

$$\begin{aligned} Z(\theta) = \log_{10} \Bigg(& \Pr(g_{ma})\Pr(g_{pa}) \sum_{\text{phase}_{ma}} \Pr(\text{Phase}_{ma}) \\ & \times \sum_{\text{phase}_{pa}} \Pr(\text{Phase}_{pa}) \prod_{\text{offspring}} \Pr(g_{offs}|g_{ma}, g_{pa}, \\ & \frac{\text{Phase}_{ma}, \text{Phase}_{pa}; \theta)}{\Pr(g_{ma})\Pr(g_{pa})} \\ & \times \sum_{\text{phase}_{ma}} \Pr(\text{Phase}_{ma}) \sum_{\text{phase}_{pa}} \Pr(\text{Phase}_{pa}) \\ & \times \prod_{\text{offspring}} \Pr\left(g_{offs}|g_{ma}, g_{pa}, \text{Phase}_{ma}, \right. \\ & \left. \text{Phase}_{pa}; \theta = \tfrac{1}{2}\right) \Bigg). \end{aligned}$$

Again, the parental genotype probabilities factor out of this equation, but the phase probabilities do not.

In fact, it is possible to consider the phase as an integral part of the parental genotype, and in that

case we see that we are effectively taking the sum over all possible parental genotypes (with phase), and weighting them by their probabilities. This argument can be easily extended to cover situations in which we known nothing about the genotype of the parents. In that case, the likelihood can be written as

$$L(\theta; \text{data}) = \sum_{G_{\text{ma}}} \sum_{G_{\text{pa}}} \Pr(G_{\text{ma}}) \Pr(G_{\text{pa}}) \times \prod_{\text{offspring}} \Pr(\text{g}_{\text{offs}} | G_{\text{ma}}, G_{\text{pa}}; \theta),$$

where G_i is the genotype with the phase of individual i; note that for the offspring we do not know the phase. If we wished to express this formula in terms of each offspring's genotype with phase, then it would be

$$L(\theta; \text{data}) = \sum_{G_{\text{ma}}} \Pr(G_{\text{ma}}) \sum_{G_{\text{pa}}} \Pr(G_{\text{pa}}) \times \prod_{\text{offspring}} \sum_{G_{\text{offs}}} \Pr(G_{\text{offs}} | G_{\text{ma}}, G_{\text{pa}}; \theta).$$

In this way we can also include offspring whose genotype is unknown, or who have some ambiguity in their genotypes. Only the parental genotypes are functions of the allele frequencies, while the genotypes of their children are dependent solely on the parental genotypes and the recombination fraction in each case.

Genotype Unknown – Phenotype Known

It is important to note that thus far we have been dealing only with genotypes, and we have assumed that we can either identify the genotype or else we know nothing. In reality we are usually somewhere in the middle; the most important way linkage analysis is used is to identify genes which affect the expression of some trait, typically by increasing the probability of becoming affected with some disease. In this situation, there are additional parameters needed to perform the linkage analysis – we need to quantify the probability of the phenotype conditional on each of the possible genotypes at the locus in question. For example, if we have a dominant disease with full penetrance, then this means that we have a disease-predisposing locus with, typically, two alleles, D and +, where $\Pr(\text{disease}|\text{DD}) = \Pr(\text{disease}|\text{D}+) = 1$; $\Pr(\text{disease}|++) = 0$. Note that this also uniquely determines the penetrances for the phenotype "unaffected" as well, because $\Pr(\text{disease}|\text{DD}) + \Pr(\text{unaffected}|\text{DD}) = 1$. In this case $\Pr(\text{unaffected}|\text{DD}) = \Pr(\text{unaffected}|\text{D}+) = 0$, and $\Pr(\text{unaffected}|++) = 1$. For a fully penetrant recessive disease, $\Pr(\text{disease}|\text{DD}) = 1$, and $\Pr(\text{disease}|\text{D}+) = \Pr(\text{disease}|++) = 0$. These are the two most classical situations in which the genotypes are not uniquely determined by the phenotypes. These penetrances can be factored into the likelihood in a straightforward manner as

$$L(\theta; \text{data}) = \Pr(\text{Ph}_{\text{ma}}) \Pr(\text{Ph}_{\text{pa}}) \times \prod_{\text{offspring}} \Pr(\text{Ph}_{\text{offs}} | \text{Ph}_{\text{ma}}, \text{Ph}_{\text{pa}}; \theta),$$

where Ph_i is the observed phenotype for individual i at all loci. If we allow for the penetrances as described above, we know that $\Pr(\text{Ph}) = \sum_G \Pr(G) \Pr(\text{Ph}|G)$. For parents, $\Pr(G)$ can be computed from the allele frequencies at each locus, and $\Pr(\text{Ph}|G)$ is the penetrance that must be specified for each locus. The sum is taken over all possible genotypes at all loci. If the individual is an offspring in the pedigree, then $\Pr(G)$ is replaced by $\Pr(G|G_{\text{ma}}, G_{\text{pa}}; \theta)$, and the penetrance remains unchanged, since this is considered to depend only on the individual's genotype. In this way we can take into account any possible relationships between genotype and phenotype. Note that many genotypes will not be possible, as they are incompatible with Mendelian laws – for example, you cannot have parents who are AA and AA having a child who has genotype aa. For this reason, many terms will have zero probability. Complicated computer programs have been written to perform these calculations for any set of penetrances and general pedigree structures. For further details about the technical aspects of likelihood calculations in linkage analysis the reader is referred to [1] and [16]. The important thing here is to see why it is necessary to specify all the parameters, and how they affect the analysis.

Fully Penetrant Recessive Traits

The simplest mode of inheritance to consider for a disease is one with full penetrance and no phenocopies. Let us start by considering a simple recessive trait, which means that $\Pr(\text{affected}|\text{DD}) = 1$ and $\Pr(\text{affected}|\text{D}+ \text{ or } ++) = 0$. Look at the

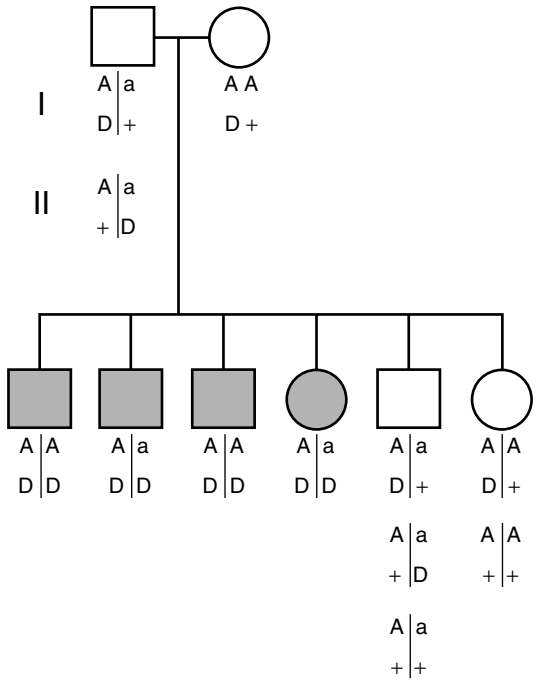

Figure 4 Sample pedigree with a fully penetrant autosomal recessive disease segregating. Solid shapes indicate affected individuals with this trait, and open figures are unaffected

pedigree in Figure 4. In this pedigree, the possible genotypes (with phase) for each individual are indicated – the probabilities of all other genotypes can be shown to be 0. Because the trait is fully penetrant recessive, all affected individuals must have genotype DD, and all unaffected individuals are either D+ or ++. Because the parents have affected children, they must each have at least one D allele, and because they are unaffected, they cannot have two – therefore, they are D|+. The only ambiguous cases are the two unaffected children, who could be either ++ or D+, since both of these are compatible with the parental genotypes and the phenotype unaffected. Because the mother is not heterozygous at both loci, we cannot tell whether recombination occurred between the trait and marker loci (since all children must receive the A allele from her with probability 1, irrespective of what disease allele they received). Because there is no ambiguity in the genotype of the affected individuals, it is clear that they provide most of the information about linkage. The ambiguity of the other sibs' disease locus genotypes adds noise to the analysis. Because disease alleles are typically rare, and recessive diseases are quite often lethal, it is rare for parents to be affected themselves, and thus the majority of pedigrees which one will ascertain (*see* **Ascertainment**) are either nuclear pedigrees, as in Figure 4, or inbred (*see* **Inbreeding**) pedigrees, where the disease alleles in the affected kids are identical by descent from some common ancestor.

In the case of inbred pedigrees, most affected children would be homozygous at a marker locus tightly linked to the recessive trait locus [19] (this is also the fundamental cause of linkage disequilibrium if one thinks of populations as large extended inbred pedigrees). Smith [19] proposed that an efficient strategy for detecting linkage with rare recessive traits in inbred pedigrees would be to look for marker loci at which affected individuals are more frequently homozygous than expected – a technique that has come to be known as homozygosity mapping. It is critical to point out that homozygosity mapping is not a different statistical technique to analyze data – in fact, one normally applies standard model-based linkage analysis to the pedigree in question. It is merely an efficient technique for minimizing the amount of genotyping one needs to detect linkage, by only typing one affected child from a consanguineous marriage segregating a recessive disease in the initial genome screen – later, of course, one should go back and genotype the parents and other family members to make sure that the individual has received the marker alleles identical by descent (IBD) from one common ancestor.

Fully Penetrant Dominant Disease

The second classical model for trait inheritance is fully penetrant dominant, in which $\Pr(\text{affected}|\text{DD or D+}) = 1$ and $\Pr(\text{affected}|{+}{+}) = 0$. The majority of affected individuals in such a disease are going to have genotype D+, because the disease allele is typically very rare, and all affected children have at least one affected parent. As a result of this, pedigrees segregating fully penetrant dominant diseases are typically large and extend over multiple generations, with smaller sibships than in recessive disease pedigrees. This is because, in recessive diseases, when we ascertain pedigrees to have more than one affected child, we are biased toward

large sibships, whereas for dominant diseases there are typically affected individuals in many generations. For this reason, dominant traits are often transmitted in phase-known meioses, whereas most meioses in recessive pedigrees are phase-unknown (typically only the bottom generation has affected individuals).

Complex Disease

A complex disease is one for which either the mode of inheritance is unknown, there are multiple genes involved, diagnosis is uncertain, or environmental factors are the entire cause of the disease, and no genes are involved at all [12]. Typically, the penetrances are not 0 and 1, but somewhere in the middle, even for single gene disorders. It may be that a disease has a late age of onset, which is variable from individual to individual, or it may be that certain environmental factors are necessary in combination with the genes to produce a phenotype, etc. (*see* **Penetrance**).

When such complexities are present, it becomes difficult to do a good model-based linkage analysis, because the linkage analysis is based on specified parameters, as indicated above. When these parameters are incorrectly specified, the recombination fraction is usually overestimated, and the lod scores may be smaller than if the model was correct. That is not to say the power is always highest when analysis is done under the correct model – power to detect linkage often tends to be higher when the genetic effect of a locus on the trait is overestimated, especially if the mode of inheritance is actually very weak – hence the high power of **"model-free" linkage methods**, which are in many cases mathematically equivalent to lod score analyses under models with very strong genetic effects, as shown later in this article. However, when the mode of inheritance is incorrectly specified, it is impossible to get accurate estimates of the location of a disease gene from the recombination fraction estimates, and the test statistic itself is the only means of determining the location of the disease gene. For this and other reasons it is very difficult to fine-map a disease gene for a complex trait in an equivalently sized dataset, and the accuracy will be orders of magnitude less with complex traits. As an extension of this argument, it will also probably be very difficult to find linkage disequilibrium with complex trait predisposing genes, as those genes are often very common, thus leading to extreme allelic and nonallelic heterogeneity in the population-as-pedigree.

Testing for Linkage – Positive and Negative

Originally, the lod score method was proposed as a sequential procedure in which one would continue to add more and more families until the lod score at some predetermined recombination fraction either exceeded 3, in which case linkage was accepted, or fell below -2, in which case linkage was said to be excluded [14]. However, the common practice changed such that people now maximize the lod score over the recombination fraction to prove linkage with greater **power**, while to exclude linkage they simply look at all values of θ for which the lod score remains below -2 (all points to the left of the upward pointing arrow in Figure 5). The example shown in Figure 5 would allow for the conclusion that there is linkage with MLE of a recombination fraction equal to $\hat{\theta}$, as indicated in the figure, and yet linkage is also excluded at small recombination fractions in the same pedigree. Since it is known, for linkage analysis under an incorrect mode of inheritance assumption, that the recombination fraction MLEs are biased in an upward direction [17], such decision rules for "exclusion" mapping are not very useful when the mode of inheritance is not well characterized and correctly modeled.

In modern human genetic linkage studies a large number of markers are tested in a fixed set of pedigrees. In this context, the spirit of lod score analysis has changed dramatically since the days of Morton [14]. We are no longer testing one specific marker, but an entire genome-wide set of markers, typically spaced at intervals from 5–10 cM, in which case there would be about 600 such markers in a full genomic scan. In analyzing this situation, there is a multiple testing problem to be taken into account. If there is a single gene disorder, it has been argued that the gene must be somewhere, and thus as more markers are shown to be unlinked, the prior probability that one of the remaining markers is linked is increased, and this would theoretically compensate for the large number of tests. However, this argument does not hold if we are using linkage analysis to prove there is a gene. For certain complex diseases we have no proof of a genetic component at

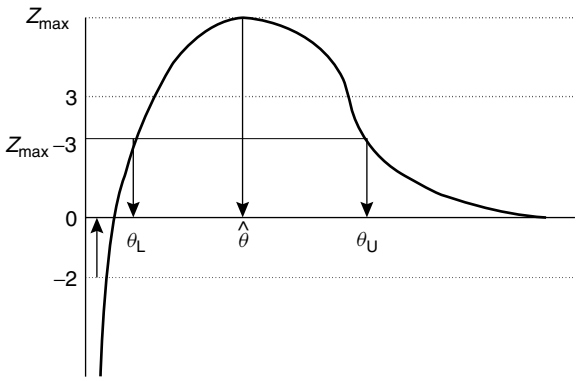

Figure 5 Sample graph of two-point lod scores as a function of the recombination fraction, with: indicators of the maximum lod score, Z_{max}; the maximum likelihood estimate of the recombination fraction, $\hat{\theta}$; the upper and lower limits of its 3-unit support interval, (θ_U and θ_L); and the exclusion region (to the right of the upwardly pointing arrow)

all, despite the best efforts of **segregation analysis**. If there is no gene, the prior probabilities of linkage are not increased after many markers are all found to be unlinked to the trait.

The current theoretical arguments about critical values are based on the null hypothesis distribution of the lod score maximized over all markers in the genome. If there is no linkage, different analyses have predicted the probability of a lod score of 3 arising by chance for some marker to be between 0.005 and 1, depending on the pedigree structure, marker informativeness, and other assumptions (e.g. [9] and [22]). For example, for a fixed critical value, with more informative markers, there is a lower rate of **false** (and true) **positives** over the whole genome. There is thus a small but nontrivial chance of getting at least one false positive in a genome screen with a large number of markers. In practice, one might consider how their best lod scores compare with the lod scores at other markers throughout the genome – obviously the highest lod scores are the most promising, and the lower ones are less so (cf. [22]). There are no hard and fast rules, because molecular technology has progressed to the point where one can feasibly obtain genotypes for as many markers as one wants. Ideally, one should consider the lod scores for all markers in a completed genome scan jointly before interpreting borderline results, in contrast to the old days (i.e. 3–4 years ago) of linkage analysis where typing small numbers of markers was a huge chore. For complex diseases, a lod score of 3 is not so convincing today, unless it is with a **candidate gene**, or it is interpreted in the context of a full genome scan.

Estimation of the Recombination Fraction

When the mode of inheritance is known, it is possible to compute the exact likelihood for the data, and the maximum likelihood estimates of the recombination fraction, while typically biased, are consistent. To demonstrate the **bias** in a very simple example, let us consider a phase-known pedigree with four informative meioses, and let us assume that the recombination fraction is 0.50. Then we compute the expectation of the MLE. The possible outcomes are given in Table 1. Because $E(\hat{\theta}) = \sum_{Data} \hat{\theta} Pr(data)$, where the sum is taken over all possible outcomes, in this case, while $\theta = 0.5, E(\hat{\theta}) = 0.40625$, which shows a considerable downward bias. However, asymptotically, as the number of informative meioses approaches infinity, the **expectation** of the estimate approaches its true value, and therefore the MLE of θ is **consistent**.

That the MLE of the recombination is consistent is valuable to researchers in human genetics, as it gives a means not only to detect linkage through use of the lod score as a test statistic, but also to estimate the distance between the marker locus and the disease-predisposing gene. Normally, because these estimates are not very accurate in small samples, we construct support intervals around the MLE, and say that the true recombination lies somewhere in that interval with reasonable certainty. The k-lod-unit support interval for the MLE of the recombination

Table 1 Sample demonstration of the bias in estimates of the recombination fraction in small samples

Pedigree	Outcome			
Recombinants	Nonrecombinants	Pr(data)	$\hat{\theta}$	Pr(data)*$\hat{\theta}$
0	4	(0.5)	0	0
1	3	4(0.5)	0.25	0.0625
2	2	6(0.5)	0.50	0.1875
3	1	4(0.5)	0.50	0.1250
4	0	(0.5)	0.50	0.03125
Total		1		0.40625

fraction consists of all values of θ for which $Z(\theta) \geq Z(\hat{\theta}) - k$. Historically, researchers used $k = 1$ to construct a support interval which asymptotically was approximately equivalent to a 95% confidence interval. However, an inconsistency arises when the maximum lod score is greater than 1 and less than 3, because in that case a 1-unit support interval would exclude the null hypothesis $\theta = 0.5$, even though the null hypothesis has not been rejected. Two possible solutions to this problem are in common practice. The original recommendation was to think of linkage analysis as a two-step procedure, i.e. first one tests the null hypothesis of no linkage, and if this null hypothesis is rejected, then and only then do you consider the estimates of the recombination fraction and its 1-unit support interval [3]. Another argument suggests that, since testing and estimation are based on the same statistic, it is impossible to separate the two logically, and for this reason, a 3-lod-unit support interval should be constructed (because the test is typically performed with the critical value of $Z > 3$) [21]. Fundamentally, either argument works, and it all depends on what the end-user wants to believe – the latter procedure will allow support intervals to be constructed even without significant test results, while the former does not; however, the latter procedure gives much wider support intervals, and thus does not narrow down the region in which the investigators would have to look for the gene as much as a 1-unit support interval would. The counter argument is that 1 in 20 times (approximately) the gene would be outside the 1-unit support interval, while only 1 in 10 000 times would the gene fall outside the 3-unit support interval. The choice of support criteria is dependent largely on the desires of the investigator and whether one thinks a 5% chance of missing the gene is a gamble worth taking.

The caveat of all this discussion about estimating recombination fractions is that it is dependent on the accuracy of the parametric model for the disease *and* the marker locus. It presumes that the allele frequencies are all accurately estimated and that the penetrances are correct as well. For complex diseases, this is never the case, especially at the trait locus. In many cases, it is impossible to do this correctly when we are restricted to the confines of a single-locus parametric model of disease with no environmental cofactors. There have been attempts made to extend linkage analysis methods to multiple gene traits, and to mixed environmental/genetic models, but they are computationally intensive, and do not tend to increase the power of the test statistic greatly. The only gain from these complications, for most cases studied thus far, is an increase in the accuracy of the recombination fraction estimates. However, in practice, if one is using a single gene model which is known with certainty to be incorrect, one will find that the recombination fraction is always overestimated, and loses all of its meaning; in those situations, the value of the test statistic itself is all we have to use in fine-mapping the trait locus.

Relationships Between "Model-free" and "Model-based" Methods

In linkage analysis, **nonparametric methods** have been employed to "increase robustness" [9] and to make the calculations fast and simple [2]. Initially it was impossible to do likelihood-based analysis on complex pedigrees, as for general pedigrees the likelihoods could not be computed in the absence of recent technological and theoretical innovations, while most sib-pair and relative pair analyses could be performed on the back of an envelope. However, as soon as there were additional affected relatives beyond the initial affected relative pair, the analyses become problematic, as there are higher-order correlations among the marker genotypes of multiple affected individuals within a single pedigree. In special situations the multiple relative pairs may be asymptotically pairwise-independent [2], but it is still an approximation to looking at the entire set of affected individuals jointly, as in likelihood-based pedigree analysis.

Recent theoretical studies have demonstrated that one can often use model-based likelihood methods

to compute statistics with equivalent properties to the pairs-based statistics, in which the entire set of affected individuals in a pedigree are analyzed jointly, e.g. [8]. Below, the simplest cases are examined in detail for sib-pair analysis and extended-pedigree identity-by-descent (IBD) analyses (*see* **Linkage Analysis, Model-free**).

Sib-pair Analysis

It has recently been demonstrated [7, 8] that there is an algebraic equivalence between the sib-pair mean test [2] and parametric linkage analysis under a recessive model. More accurately, it has the same statistical properties as a likelihood-based linkage analysis between the marker locus of interest and a "pseudo-marker", which has genotype 1–2 in the mother, 3–4 in the father, and 2–3 in each affected child. If one sets $\theta = 1/2$ between these two loci, then from each parent the children share one allele at the marker locus IBD with probability $p = \theta^2 + (1-\theta)^2 = 0.5$, which is the null hypothesis of the sib-pair mean test statistic. When θ is allowed to take on all values between 0 and 0.5, there is a 1:1 mapping of the interval $[\theta : 0, 0.5] \rightarrow [p : 0.5, 1]$, and the lod score between the marker locus and this "pseudo-marker" is a simple transformation of the sib-pair mean test statistic, $R = (x-y)^2/(x+y)$, where $Z_{\max} = R/2\ln(10)$, x is the number of alleles shared IBD over all sib-pairs, and y is the number of alleles not shared IBD across all sib-pairs). In light of this equivalence when one is analyzing only sibling pairs, the analogy can be extended to multiplex sibships [10, 18] by computing the lod scores between a marker locus and such a "pseudo-marker" where all affected siblings in a sibship have genotype 2/3. The statistic $R = 2\ln[L(\hat{\theta})/L(\theta = 0.5)]$ has a well-defined distribution which converges rapidly, in as few as 20 sibpairs, to a 50–50 mixture of $\chi^2_{(1)}$ and a point mass at $R = 0$. The traditional mean test, no matter what weighting function is assumed for multiplex sibships [2], has a skewed distribution when larger sibships are analyzed. Analysis of the power of this likelihood-based extension has been shown to be consistently robust and powerful over a wide variety of modes of inheritance [4].

Other Affected Relative Pairs

In other "model-free" methods based on extended pedigrees, it is customary to select a set of pairs of relatives and see if they share more alleles IBD than would be expected if there were no linkage. Traditionally, this has been most frequently done by breaking multiplex pedigrees into all possible pairs of affected relatives, and pretending they are independent of each other, when really there is a complicated set of interdependencies which only go away asymptotically (i.e. in unrealistically large datasets). Ultimately, one is interested in testing whether or not a given marker locus segregates independently of the trait in the entire pedigree. Following the aforementioned logic, in pedigrees without consanguinities, any pair of individuals who are not sibs can share at most one allele IBD. If we have a simple pedigree with only two affected individuals, an artificial "pseudo-marker" locus can be created in which they share the one marker allele IBD that is the most they could possibly share, i.e. they each are assigned a pseudo-marker genotype 1/2, where all founder individuals who are not ancestors of all affected individuals are given genotype 1/1. Performing a likelihood-based analysis of this locus (setting the allele frequency of the 2 allele to be very small) against a marker would represent a test equivalent to an IBD test on this relative pair because the recombination fraction again is a 1:1 transformation of the probability that the two relatives share an allele IBD. The model of the trait "pseudo-marker" genotypes is essentially a rare dominant mode of inheritance in the likelihood calculations. In this way it is possible to develop statistics with properties analogous to various nonparametric IBD methods within the unified context of likelihood-based lod score analysis, giving us a common currency and a feel for the underlying symmetries between conceptually different methods of linkage analysis.

Linkage Disequilibrium

Ultimately, linkage disequilibrium analysis is very similar to the extended pedigree analysis described above. In linkage disequilibrium analyses, the population under study is thought of as one giant pedigree, in which the disease predisposing allele is assumed to have entered a population once, or very few times (*see* **Linkage Disequilibrium**). Then, many generations later, at tightly linked marker loci, the allele which was on the founder chromosome would still be present with a higher frequency among affected individuals. In essence, the null hypothesis is that

the marker locus genotypes are independent of the disease, i.e. the disease and marker alleles have segregated independently in this population. Under the alternative hypothesis, however, the assumption is that the affected individuals would share more alleles IBD from this common ancestor than any two randomly selected individuals from the population. Again, an analogy can be made to likelihood-based linkage analysis in the population, following the paradigm from the previous section, assuming each individual had received 2 alleles IBD (i.e. we typically have to look at genotypes in **case–control** disequilibrium studies). However, in a population, we assume that the individuals under study are so distantly related that we can approximate the linkage analysis by simply comparing the genotype frequencies in affecteds and unaffecteds. If there are known to be closer relationships between certain sets of individuals in the population, it would behoove the analyst to take these correlations into account where relationships can be identified, to avoid erroneous assumptions that all genotypes within case and control samples are really iid.

Algebraic Equivalence ≠ Identical Assumptions

It is, of course, erroneous to say that IBD analysis in sib-pairs "assumes" the mode of inheritance to be recessive – rather, it is more appropriate to say that the sib-pair mean test statistic is algebraically equivalent to lod score analysis under a recessive model (with additional assumptions). The subtle difference between these two statements is critical to appreciate, for the null hypothesis properties of likelihood-based lod score analyses do not depend in any way on the true mode of inheritance, and are valid irrespective of the true state of nature. This same statement holds for nonparametric tests as well: under the null hypothesis, the marker is assumed to segregate randomly and independently of the trait, and thus the true mode of inheritance is irrelevant to the validity of any of these tests.

Conclusion

Likelihood-based parametric linkage analysis is the **gold standard** for detecting disease genes through reverse genetics in pedigree data. There are a number of other procedures (*see* **Linkage Analysis, Model-free**) which provide simple and rapid approximations to this type of analysis, but ultimately it remains the standard which the nonparametric methods attempt to emulate. Bearing in mind the **Neyman–Pearson lemma**, which states that if there is a best test of a given hypothesis it will be in the form of a likelihood ratio, and also the manner in which full likelihood analysis can use all of the data and not just a small subset thereof, it remains the method of choice. It has been very successful in mapping hundreds of disease-predisposing genes, and while there are no easy answers to the questions of the future – involving common complex diseases – it seems likely that it will be the best unified framework at our disposal to build upon in answering these more complicated but very important problems.

References

[1] Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Clarendon Press, Oxford.

[2] Blackwelder, W.C. & Elston, R.C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci, *Genetic Epidemiology* **2**, 85–97.

[3] Conneally, P.M., Edwards, J.H., Kidd, K.K., Lalouel, J.M., Morton, N.E., Ott, J. & White, R. (1985). Report on the committee on methods of linkage analysis and reporting, *Cytogenetics and Cell Genetics* **40**, 356–359.

[4] Davis, S. & Weeks, D.E. (1996). Comparison of nonparametric statistics for detecting linkage in affected-sir-pair data, *American Journal of Human Genetics* **59**, A216.

[5] Doerge, R.W. (1995). Testing for linkage: phase known/unknown, *Journal of Heredity* **86**, 61–62.

[6] Ginsburg, E.Kh., Axenovich, T.I. & Goodman, D.W. (1996). On estimation of linkage test power, *Genetic Epidemiology* **13**, 355–366.

[7] Hyer, R.N., Julier, C. Buckley, J.D. Trucco, M., Rotter, J., Spielman, R., Barnett, A., Bain, S., Boitard, C., Deschamps, I., Todd, J.A., Bell, J.I. & Lathrop, G.M. (1991). High-resolution linkage mapping for susceptibility genes in human polygenic disease: insulin-dependent diabetes mellitus and chromosome 11q, *American Journal of Human Genetics* **48**, 243–257.

[8] Knapp, M., Seuchter, S.A. Baur, M.P. (1994). Linkage analysis in nuclear families: relationship between affected sib-pair tests and lod score analysis, *Human Heredity* **44**, 44–51.

[9] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.

[10] Kuokkanen, S., Sundvall, M., Terwilliger, J.D., Tienari, P.J. Wikström, J., Holmdahl, R., Pettersson, U. & Peltonen, L. (1996). A putative vulnerability locus to

multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus EAE2, *Nature Genetics* **13**, 447–480.

[11] Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**, 241–247.

[12] Lathrop, G.M., Terwilliger, J.D. & Weeks, D.E. (1996). Multifactorial inheritance and genetic analysis of multifactorial disease, in *Emory & Rimoin's Principles and Practice of Medical Genetics*, 3rd Ed., D.L. Rimoin, J.M. Connor & R.E. Pyeritz, eds. Churchill Livingstone, New York, pp. 333–346.

[13] Lathrop, G.M., Lalouel, J.M. Julier, C. & Ott, J. (1984). Strategies for multilocus linkage analysis in humans, *Proceedings of the National Academy of Sciences* **81**, 3443–3446.

[14] Morton, N.E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.

[15] Nordheim, E.V. (1984). On the performance of a likelihood ratio test for genetic linkage, *Biometrics* **40**, 785–790.

[16] Ott, J. (1984). *Analysis of Human Genetic Linkage*, 1st Ed. Johns Hopkins University Press, Baltimore.

[17] Risch, N. & Giuffra, L. (1992). Model misspecification and multipoint linkage analysis, *Human Heredity* **42**, 77–92.

[18] Satsangi, J., Parkes, M., Louis, E., Hashimoto, L., Kato, N., Welsh, K., Terwilliger, J.D., Lathrop, G.M., Bell, J.I. & Jewell, D.P. (1996). Two-stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3,7, and 12, *Nature Genetics* **14**, 199–202.

[19] Smith, C.A.B. (1953). The detection of linkage in human genetics, *Journal of the Royal Statistical Society, Series B* **15**, 153–184.

[20] Tai, J.J. & Chen, C.L. (1989). Asymptotic distribution of the lod score for familial data, *Proceedings of the National Science Council of the Republic of China, Series B* **13**, 38–41.

[21] Terwilliger, J.D. & Ott, J. (1994). *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

[22] Terwilliger, J.D., Shannon, W.D. Lathrop, G.M. Nolan, J.P. Goldin, L.R. Chase, G.A. & Weeks, D.E. (1997). True and false positive peaks in genome-wide scans: applications of length-biased sampling to linkage mapping. *American Journal of Human Genetics* **61**, 430–438.

J.D. TERWILLIGER

Linkage Analysis, Multipoint

Multipoint linkage analysis is the analysis of linkage data involving three or more linked loci (*see* **Linkage Analysis, Model-based**). Such analyses are carried out to order or map a set of loci, to position a new locus in relation to a mapped set of loci, or perhaps to exclude a locus from a region containing two or more loci.

Until the mid-1980s, most linkage mapping was two-point; that is, it involved the estimation or testing of a single recombination fraction. Although inefficient from the statistical viewpoint, three or more loci can be mapped using only two-point data, since linear maps are determined by pairwise distances. When there are plenty of data, such as with *Drosophila*, multipoint analyses may be unnecessary. However, in most contexts, data are scarce. In such cases, multipoint linkage analysis can be viewed as an attempt to make more efficient use of recombination data to further the aims of **linkage analysis** [15, 17, 28]. By making fuller use of available data, greater precision or **power** is achievable; at times the differences can be large.

Multipoint linkage analyses are more complex than two-point analyses in several important ways. First, they require the specification of an order for the loci: if we have n linked loci, there are $\frac{1}{2}n!$ potentially distinguishable orders. Secondly, they require the specification of a joint distribution for all possible recombination patterns: for n loci, there are 2^{n-1} such patterns (including the parental one). Thirdly, from the perspective of parametric statistical inference, joint distributions over recombination patterns corresponding to distinct orderings of the loci define noncomparable statistical models. Most of the difficulties of multipoint linkage analysis stem from these facts, particularly the rate of increase of the number of orders or patterns with the number of loci. When linkage analysis is being done using pedigree data, the size (number of individuals) and complexity (presence of one or more loops) of the pedigrees are additional limiting factors.

As with two-point linkage analyses, a major complication in multipoint linkage analyses can be the incompleteness of data. For example, there may be missing data due to some individuals not being typed. All data may be available, but phenotype may not determine genotype, as with dominant traits and other types of incomplete penetrance. **Genotypes** may be known, but haplotypes may not. That is, phase – which allelic combinations across loci are together on the same chromosome – may be unknown (*see* **Haplotype Analysis**). With known genotypes at n loci, there are 2^{n-1} possible haplotypes. While these incompleteness problems can slow down two-point analyses, they can quickly make exact multipoint analyses impossible. However, multipoint analyses can make use of data that cannot be used in two-point analyses; for example, when only uninformative data are available at a locus intermediate between two fully informative loci [18, 24]. In multipoint linkage analysis using pedigree data, the feasibility of an exact analysis will depend on the number of loci, the size and complexity of the pedigrees involved, and the nature and extent of incompleteness in the data.

To explore the topics in a little more detail, it is necessary to introduce some notation. Suppose that we are discussing data from n loci, written $A-B-\cdots-C$ in an arbitrary, but fixed, order. Then the joint recombination probabilities may be denoted by $\mathbf{p} = (p_{i_1 i_2 i_3, \ldots, i_{n-1}})$, where the subscript $i_k = 1$ corresponds to recombination across the kth interval, and $i_k = 0$ corresponds to no recombination across the same interval. For example, if $n = 3$, and the loci are ordered $A-B-C$, then we have four probabilities $\mathbf{p} = (p_{00}, p_{01}, p_{10}, p_{11})$, corresponding to the four patterns of recombination or not across $A-B$ and $B-C$. The order with respect to which these probabilities are defined does not need to be the true one, and if we change it, the probabilities need only be relabeled. For example, if we go from the order O: $A-B-C$ with probabilities $\mathbf{p}$, to O': $A-C-B$ with probabilities $\mathbf{p'}$, then $\mathbf{p'}$ is related to $\mathbf{p}$ as follows:

$$p'_{00} = p_{00}, \qquad p'_{10} = p_{10},$$
$$p'_{01} = p_{11}, \qquad p'_{11} = p_{01}.$$

Our first remark is that three-point phase known crosses have been used for decades to order loci in experimental organisms without any explicit model assumptions. This works because, under very general conditions, the smallest of the four probabilities $(p_{i_1 i_2})$ corresponds to the event of double recombination across two consecutive intervals when the loci are correctly ordered. For example, if the correct order is O: $A-B-C$, then (assuming no chromatid

interference (*see* **Genetic Map Functions**):

$$p_{11} \leq p_{10}, \qquad p_{01} \leq p_{00}.$$

If, however, O': $A-C-B$ is the correct order, but we have written our probabilities relative to O, then $p'_{11} = p_{01}$ will be the smallest probability. It follows that with sufficiently large samples of data, any set of loci can be ordered by inspection, with only a small chance of error. Naturally, this is also possible using only the pairwise recombination fractions, but that would take more data to achieve the same level of confidence in the ordering. More generally, it is possible to show that under the assumption of no chromatid interference, a multipoint recombination probability decreases, or at least does not increase, when any nonrecombinant interval is changed to recombinant status [26]. Lathrop et al. [16] discuss three-point mapping from the point of view of hypothesis testing.

Historically, the first formal linkage analysis involving more than two loci was given by Fisher [4]. There, he showed how to combine data from a number of two-point analyses in order to get **efficient** estimates of a set of recombination fractions. Although the data were all two-point, Fisher needed to express the recombination fraction across the union $A-C$ of two adjacent intervals $A-B$ and $B-C$ in terms of their individual recombination fractions. He did so by making the assumption of *complete interference*; that is, by assuming that, at most, one recombination could occur across any pair of adjacent intervals. This is equivalent to the following joint distribution:

$$p_{00} = 1 - r_1 - r_2, \quad p_{01} = r_2,$$
$$p_{10} = r_1, \qquad\qquad p_{11} = 0,$$

where r_1 and r_2 are the recombination fractions across $A-B$ and $B-C$, respectively. This model would not be appropriate for the analysis of three-point data in which double recombinants are observed, but it has been used in modern times with very short intervals, see, for example [24], section 6.7. The first satisfactory class of recombination models were the χ^2 **renewal process** models discussed by Fisher and his students and colleagues [5]; Bailey [1] gives a good overview of this research. The simplest of these joint probabilities is too complex to be given here, and this is probably the reason that this class of models has not been used with human data until recently [20]. In human linkage analysis one finds almost exclusive use made of the extremely tractable **Poisson** or *no interference* model, whose joint probabilities for three loci take the form:

$$p_{i_1 i_2} = r_1^{i_1}(1-r_1)^{1-i_1} r_2^{i_2}(1-r_2)^{1-i_2},$$

where, for $i = 1, 2$, the recombination fractions r_i may be expressed in terms of map distances d_i by

$$r_i = \tfrac{1}{2}(1 - \exp(-2d_i)).$$

It seems that although this model and its extension to more than three loci fail to fit most data sets of any size, the recombination fractions and locus orderings obtained are generally satisfactory [26]. However, the map distances estimated under this model may be seriously in error, and so use is typically made of a suitable map function at the end of the analysis. We refer to **Genetic Map Functions** for more on this point, and for further details concerning probability models for recombination, within which multipoint recombination probabilities must be calculated. The major alternatives to the χ^2 renewal models introduced by Fisher et al. are due to Karlin & Liberman [10] and Risch & Lange [25] (independently), called count-location or generalized no interference models, and the model of Goldgar & Fain [6].

Ordering More Than Three Loci

Many strategies exist for ordering a set of loci on the basis of multipoint recombination data concerning those loci. Some of these are mentioned in [24], and for others we refer to [29]. There is no evidence to suggest that a method exists that is generally better than choosing that order that maximizes the **likelihood** of the data using a suitable recombination model, at least not when the calculation of the likelihoods corresponding to each of the $\frac{1}{2}n!$ distinct orders is possible. The Poisson or no interference model is the one typically used in this context. Although there does not appear to be a systematic study of this issue, the available evidence suggests that only small gains in the efficiency of ordering loci are to be found by using a more suitable model when interference exists (see [2], [6], [7], [18], and [26] for related results).

It is not possible to examine all of the different orders when the number of loci grows beyond 15–20.

At that point it becomes necessary to adopt some deterministic or stochastic search strategy, which concludes with an order that may be suboptimal. In this respect, the locus ordering problem resembles the traveling salesman problem widely discussed in the field of combinatorial optimization [9]. However, the programs that are currently widely used to order loci on the basis of human or other pedigree data make little or no use of recent research from that field.

Location Scores

The idea behind the use of location scores [17] is that we wish to compare two simple hypotheses concerning the location of an unmapped locus. One states that it is at a specific position, B, in the interior of the chromosomal interval subtended by two known loci A and C; the other asserts that it is elsewhere, unlinked to either A or C. This comparison will be carried out using the log-likelihood ratio based on data concerning the known loci A and C, and the unmapped locus. For the present illustration we will suppose that we have complete information, including phase, concerning the three loci, and the question of interest is whether these data provide more support for the hypothesis that the locus is at a specific position, B, in the interval $A-C$, than the alternative that it is elsewhere, unlinked to both. In practice, there may be further loci to the left of A and to the right of C, as well as incomplete data, but we will ignore these possibilities here.

The probabilities ($p_{i_1 i_2}$) under the first hypothesis must be specified using a suitable model, and we suppose that after this is done, the likelihood for the data is $L(ABC)$. However, if the locus is unlinked to both A and C, then we will have

$$p_{00} = p_{11} = \tfrac{1}{2}(1-r), \qquad p_{01} = p_{10} = \tfrac{1}{2}r,$$

where r is the recombination fraction between A and C, assumed known. This assignment will lead to a likelihood for the data that we denote $L(AC)$. A **likelihood ratio test** of the null hypothesis that the locus is unlinked to either A or C, against the alternative that it is at B, will then be based on the quantity

$$L = \log\left[\frac{L(ABC)}{L(AC)}\right],$$

and the null hypothesis will be rejected if L is sufficiently large. If we regard B as a variable point in the interval $A-C$, this quantity can be viewed as a function $L = L(B)$ of the position B along the interval, and this is called a *location score*. A significant peak in the function at some point may then be interpreted as suggesting that the unmapped locus is at that point. Of course, the threshold determining significance must be decided upon taking into account the model and the length of the interval.

In practice, this calculation may be repeated in each of a series of adjacent intervals, and the global maximum, or all maxima *above* some threshold, noted. As long as there are recombinations in a given data set in every such interval, the resulting plot will go to $-\infty$ as the endpoints of each interval are approached, and be roughly parabolic in between, although the shape can be somewhat different with certain patterns of incomplete data. Further details and graphs can be found in [27].

An alternative use of these location scores leads to what is known as *exclusion mapping*. The idea here is that if the location score stays *below* a suitable threshold throughout an interval, then this may be interpreted as suggesting that the unmapped locus lies nowhere in the interval. As with the direct use of location scores, care needs to be taken with thresholds. Practices differ, and we refer to [24] for more details on this matter.

Algorithms and Programs for Multipoint Linkage Mapping

I will refer mainly to human linkage analysis using qualitative data on one or more pedigrees. Multipoint linkage analysis with quantitative traits is a specialized subarea that I cannot discuss here, although some of the references given below cover that topic as well. Programs for carrying out multipoint linkage analysis for crosses of experimental organisms are much more straightforward to write, but no such general purpose programs seem to have gained widespread use. However, particular programs do circulate among communities of scientists studying the same or similar organisms.

Overviews of available computer programs for multipoint linkage mapping in humans can be found in Ott [24] and Terwilliger & Ott [27]. These also contain excellent bibliographies.

Algorithms for carrying out multipoint linkage analysis with human (and other) pedigree data are of two kinds: those based upon the **Elston–Stewart** [3] approach, using what is known as peeling, and those based upon the Lander & Green [14] hidden Markov model formulation. Each of these classes of algorithms has its strengths and weaknesses, and there are problems that cannot be solved exactly with either of them. The Elston–Stewart approach underlies most of the algorithms discussed in [24] and [27], and we refer to these for further details. For a recent improvement of the implementation of these algorithms, see [23]. One new package that has become available since the publication of [24] and [27] uses the basic Lander & Green algorithm in a number of different human linkage problems [13]. These include analyses with sib-pairs [11], the analysis of recessive traits with nuclear families [12], and multipoint linkage with many markers for general pedigrees of moderate size [13]. Another recent program assists with map construction [22].

When exact linkage analysis methods fail because of time or space constraints, **Monte Carlo methods** may be used. At present, these are more research tools than approaches suitable for routine use, but they are developing rapidly, and should become more widely used in the near future. Lin [19] is a recent review, discussing both the sequential imputation approach of Irwin et al. [8] and **Markov chain Monte Carlo** methods [21].

References

[1] Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.

[2] Bishop, D.T. & Thompson, E.A. (1988). Linkage information and bias in the presence of interference, *Genetics and Epidemiology* **5**, 107–119.

[3] Elston, R.C. & Stewart, J. (1971). A general model for the analysis of pedigree data, *Human Heredity* **21**, 523–542.

[4] Fisher, R.A. (1922). The systematic location of genes by means of crossover observations, *American Naturalist* **56**, 406–411.

[5] Fisher, R.A., Lyon, M.F. & Owen, A.R.G. (1947). The sex chromosome of the house mouse, *Heredity* **1**, 335–365.

[6] Goldgar, D.E. & Fain, P.R. (1988). Models of multilocus recombination: Non-randomness in chiasma number and crossover location, *American Journal of Human Genetics* **43**, 38–45.

[7] Goldstein, D.R., Zhao, H. & Speed, T.P. (1995). Relative efficiencies of chi-square models of recombination for exclusion mapping and gene ordering, *Genomics* **27**, 265–273.

[8] Irwin, M., Cox, N. & Kong, A. (1994). Sequential imputation for multilocus linkage analysis, *Proceedings of the National Academy of Sciences* **91**, 11684–11688.

[9] Johnson, D.S. (1990). Local optimization and the travelling salesman problem, in *Lecture Notes in Computer Science*, Vol. 443, M.S. Paterson, ed. Springer-Verlag, New York.

[10] Karlin, S. & Liberman, U. (1978). Classification and comparison of multilocus recombination distributions, *Proceedings of the National Academy of Sciences* **75**, 6332–6336.

[11] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.

[12] Kruglyak, L., Daly, M.J. & Lander, E.S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping, *American Journal of Human Genetics* **56**, 519–527.

[13] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.

[14] Lander, E.S. & Green, P. (1987). Construction of multilocus genetic maps in humans, *Proceedings of the National Academy of Sciences* **84**, 2363–2367.

[15] Lathrop, G.M., Lalouel, J.-M. & White, R. (1984). Construction of human linkage maps: Likelihood calculations for multilocus linkage analysis, *Genetic Epidemiology* **3**, 39–52.

[16] Lathrop, G.M., Chotai, J., Ott, J. & Lalouel, J.-M. (1987). Tests of gene order from three-locus linkage data, *Annals of Human Genetics* **51**, 235–249.

[17] Lathrop, G.M., Lalouel, J.-M., Julier, C. & Ott, J. (1984). Strategies for multilocus linkage analysis in humans, *Proceedings of the National Academy of Sciences* **81**, 3443–3446.

[18] Lathrop, G.M., Lalouel, J.-M., Julier, C. & Ott, J. (1985). Multilocus linkage analysis in humans: Detection of linkage and estimation of recombination, *American Journal of Human Genetics* **37**, 482–498.

[19] Lin, S. (1996). Monte Carlo methods in genetic analysis, in *Genetic Mapping and DNA Sequencing*, T. Speed & M.S. Waterman, eds. Springer-Verlag, New York.

[20] Lin, S. & Speed, T.P. (1996). Incorporating crossover interference into pedigree analysis using the chi-square model, *Human Heredity* **46**, 315–322.

[21] Lin, S. & Wijsman, E. (1994). Monte Carlo multipoint linkage analysis, *American Journal of Human Genetics* **55**, A40.

[22] Matise, T.C., Perlin, M. & Chakravarti, A. (1994). Automatic construction of genetic linkage maps using an expert system (MultiMap): A human genetic linkage map, *Nature Genetics* **6**, 384–390.

[23] O'Connell, J.R. & Weeks, D.E. (1995). The Vitesse algorithm for rapid exact multilocus linkage analysis via genotype set recoding and fuzzy inheritance, *Nature Genetics* **11**, 402–408.

[24] Ott, J. (1991). *Analysis of Human Genetic Linkage Data*. Johns Hopkins University Press, Baltimore.

[25] Risch, N. & Lange, K. (1979). An alternative model of recombination and interference, *Annals of Human Genetics* **43**, 61–70.

[26] Speed, T.P., McPeek, M.S. & Evans, S.N. (1992). Robustness of the no interference model for ordering genetic markers, *Proceedings of the National Academy of Sciences* **89**, 3103–3106.

[27] Terwilliger, J. & Ott, J. (1994). *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

[28] Thompson, E.A. (1984). Information gain in joint linkage analysis, *IMA Journal of Mathematical Applications in Medicine and Biology* **1**, 31–49.

[29] Weeks, D.E. (1991). Human linkage analysis: Strategies for locus ordering, in *Advanced Techniques in Chromosome Research*, K.W. Adolph, ed. Marcel Dekker, New York, pp. 297–330.

TERRY P. SPEED

Linkage Analysis, Multivariate

For correlated traits, such as those predicting cardiovascular disease risk, multivariate approaches for genetic **linkage** can increase the **power** and precision of estimators for genetic effects [13, 29]. For traits influenced by several genetic factors, the specific genetic loci may induce distinct correlation structures among the measures, so that one can separate the effects of each genetic locus by **multivariate analysis**, even though this might not be possible with simple univariate analyses. Finally, multivariate analysis provides a statistically efficient mechanism for controlling the analysis-wise significance level, when there are multiple trait observations for each subject (*see* **Multiple Comparisons**). In multivariate analysis of quantitative traits, it is not always apparent whether a variable should be treated as a **covariate** or as an outcome. For example, in analysis of blood pressure, body-mass index (BMI), which is a measure of obesity, is often treated as a covariate. However, if a genetic factor influences both BMI and blood pressure, then adjusting blood pressure for BMI would reduce the effects from the major-genetic locus. Therefore, using methods that can analyze several traits jointly is essential. Genetic model-free methods [5, 10, 25, 30] are more easily applied than full **likelihood** methods, which require modeling the **prevalences** of genetic factors along with the parameters to describe the **genotype** specific phenotype distributions.

De Andrade et al. [18] and Almasy et al. [4] developed and applied models for performing multivariate linkage analysis using **variance components** (VC) procedures. Eaves et al. [20] developed **structural equations models** for partitioning multivariate sources of variation among major-genetic, polygenic, and nongenetic sources of variation. Vogler et al. [35] used VC analysis to jointly perform a multivariate analysis of five traits that were simulated as a part of the Genetic Analysis Workshop. Multivariate VC analysis was primarily used as a descriptive tool, without detailed discussion of issues related to **hypothesis testing** or **estimation**. Todorov et al. [34] recently constructed a very general framework extending VC analysis as a general aspect of structural equations modeling. The framework that they provided can incorporate arbitrarily large families, but the optimization procedures are simplified for data from sibpairs. Iturria and Blangero [26] proposed an **EM algorithm** for obtaining **maximum likelihood** estimates in a multivariate VC linkage model parallel to the commonly used scoring algorithm. Programs to perform multivariate VC analysis are currently available as components of the **software** packages ACT [17], SEGPATH [28], SOLAR [3], and EMVC [26].

The Multivariate Model

The multivariate variance components (MVC) approach is an extension of the univariate approach described by

$$\mathbf{Y}_j|\mathbf{X}_j = \boldsymbol{\mu} + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{a}_j + \mathbf{g}_j + \mathbf{e}_j, \qquad (1)$$

where $\mathbf{Y}_j$ is a vector of dimension N_j of trait values for the family j, $\boldsymbol{\mu}$ is the overall mean vector of dimension N_j for family j, $\mathbf{X}_j$ is an $N_j \times p$ matrix of covariates, $\boldsymbol{\beta}$ is a p-vector of **regression** coefficients, $\mathbf{g}_j$ is a N_j-vector of genetic effects by which the major locus affects the trait values for family j, $\mathbf{a}_j$ is a $1 \times N_j$ vector expressing how the additive polygenic factor affects the trait values for family j, and $\mathbf{e}_j$ is residual variation (or environmental effects) from the model; for more details, see [5, 15]. The MVC approach is also a model-free approach, and it has advantages over model-dependent approaches. To simply describe these models, we use the vec transformation [4, 6, 18] to string out the observations as a single vector and then allow elements of this vector to be correlated, according to the model proposed by equation (1). Let $\mathbf{Y}_j = (Y_{11}, \ldots, Y_{1N_j}, \ldots, Y_{mN_j})'$ be a vector of m multivariate trait values for N_j members of the jth family. Let N be the total number of families, $\boldsymbol{\beta}$ a vector of dimension mp of the regression coefficients for the p covariates (including a vector of 1's corresponding to the overall mean), $\mathbf{X}_j = \mathbf{I}_m \otimes \mathbf{X}_{N_j,p}$ an $mN_j \times mp$ known matrix of covariate values for the jth family, where $\otimes$ is the Kronecker product. Then, the variance–covariance matrix of the m traits, $\mathbf{V}_j$, with dimension $mN_j \times mN_j$ is

$$\mathbf{V}_j = \mathbf{A} \otimes \mathbf{R}_j + \mathbf{B} \otimes \boldsymbol{\pi}_j + \mathbf{C} \otimes \mathbf{I}_j, \qquad (2)$$

where, $\mathbf{R}_j$ is the $N_j \times N_j$ matrix of the coefficients of relationship for the jth family; $\boldsymbol{\pi}_j$ an $N_j \times N_j$

matrix of estimated proportion of alleles identical by descent (IBD) for pairs of related individuals for the jth family; $\mathbf{I}_j$ is the $N_j \times N_j$ identity matrix; and **A**, **B**, and **C** are, respectively, polygenic, major-gene, and residual variance–covariance matrices each of dimension $m \times m$. A fourth term to measure dominance components can be added. Because the dominance component of variance (*see* **Population Genetics**) is usually much smaller than the additive component, it is ignored here, but can be modeled by including increased covariance among pairs sharing two alleles IBD. Similarly, additional terms to model-shared environment can be added. When longitudinal data are considered, the error variance structure can be modified to take account of **serial correlation** among the observations [16]. A special approach can be taken for discrete/quantitative traits. In this approach, a decomposition is effected in which the quantitative trait is first conditioned on the discrete trait [38].

Hypothesis Tests for Multivariate Analysis

The Multivariate Haseman–Elston (MH–E) Test

Amos et al. [7] developed a multivariate analog of the Haseman–Elston test [24]. To test for linkage, the procedure evaluates the regression expression

$$E\left[\sum_{k=1}^{m}\left(c_k\left(Y_{ik}-Y_{lk}\right)\right)^2 \mid \pi_{il}\right] = \alpha + \beta\pi_{il} \qquad (3)$$

subject to the constraints $\left(\sum_{k=1}^{m} c_k^4 + \sum_{l<k<3}^{m} 4\, c_l^2\, c_k^2\right) = 1$, to ensure that the variance of this linear function remains constant during the optimization. MAXFUN [33] was used to optimize the coefficients, c. The union–intersection approach can be used to develop a hypothesis test for **multivariate regression**, and results in the evaluation of a ratio of quadratic forms. This ratio would be expected to follow an **F-distribution** with m and $\eta - m - 1$ degrees of freedom for a test that includes a single regression coefficient (and hence effects from a single linked major gene) and η independent sibpairs except that the constraint on c that one imposes has a slightly different form from typical multivariate regression. Allison et al. [2] used the direct search option of MAXFUN to evaluate a grid of values, subject to the constraint that $\sum_{i=1}^{m} c_i^2 = 1$ to find the values c that maximize equation (3). However, they did not provide a limiting distribution for their test statistic and therefore one must depend upon empirical critical values to assess significance. Amos et al. [6] found that the critical values were similar to those provided by the F-distribution, suggesting that an F-distribution can be used to obtain significance levels. However, for smaller samples, the distribution of the test statistic was slightly wider than predicted by an F-distribution. Elston et al. [21] developed a regression approach in which the centered cross product of sibpairs are regressed upon IBD and showed that this method can be more powerful than the older H–E test. They also showed that a simple multivariate test could be constructed by first obtaining the **principal components** (PC) and then combining the ***P* values** from testing each of the separate analysis for each PC. Gorlova et al. [22] studied the statistical properties of these PC analysis and found the PC analysis to have slightly higher power than the MH–E test.

The Multivariate Variance Components Test

To test for genetic linkage, we construct a **likelihood ratio test**. Under the **null hypothesis**, the major-gene parameter(s) **B** of equation (2) are constrained to **0**. For simplicity, let us consider bivariate traits. For bivariate linkage analysis of an additive genetic effect, the parameters are σ^2_{a1}, σ^2_{a2}, and $\sigma_{a1,a2}$ where the first two components measure the major-genetic variance of the traits and the third component measures the major-gene covariance for the traits. We also usually constrain the major-gene variances to be positive so that they fall in the admissible part of the parameter space. As a result, the distribution of the bivariate test that the major-gene components and covariance are zero, is a mixture of 1/4 χ^2_0, 1/2 χ^2_1, and 1/4 χ^2_3 as suggested by Self and Liang [31]. This follows because for one-quarter of the parameter space, both genetic variance parameters are estimated to be positive and hence lead to a **chi-squared test** having three degrees of freedom; for one-half of the parameter space, one of the genetic variances is constrained to be 0 and hence the major-gene covariance is 0 so that the **chi-squared distribution** has one degree of freedom, while for the remaining one-quarter of the parameter space, both genetic variances are constrained to be zero, resulting in a degenerate distribution of a point mass at 0.

Because the same major-gene alleles are assumed to be determining the two traits, it is logical to consider imposing the constraint $\sigma_{a1,a2} = \pm\sigma_{a1}\sigma_{a2}$, which is always satisfied whenever there is a single genetic factor in a region and the dominance components of variance affecting each trait is 0. As discussed by Almasy et al. [4], the observed correlation attributable to a locus may not be one if there are multiple loci affecting both traits in a region. Therefore, they proposed testing the hypotheses of pleiotropy, which presumes that the trait(s) are influenced by the same gene versus coincident linkage, which presumes that there are two or more linked loci that separately influence the traits. If the covariance is constrained to be the product of the square root of the variances, then the hypothesis test of linkage for either of the traits becomes a mixture of 1/4 χ_0^2, 1/2 χ_1^2, and 1/4 χ_2^2. In this case, the covariance is no longer a parameter to be estimated. Amos et al. [6] compared the efficacy of fitting data either with or without this constraint on the covariance. They found rather similar power for either the unconstrained or constrained tests when the empirically derived critical values were used.

Power of Tests

Theoretical studies have evaluated the power of multivariate procedures for genetic studies of crosses between inbred animal lines and the general conclusions are relevant to studies of outbred animals and humans. Jiang and Zeng [27] provided analytical forms describing the distribution of bivariate likelihood ratio and regression-based tests. These forms can be used for comparing the asymptotic performance of tests for linkage using bivariate or univariate data. To allow for the possibility that observations may not be **normal** and tests may not converge rapidly to limiting distributional forms, Churchill and Doerge [14] advocated a permutation based approach for hypothesis testing. Under this approach data are resampled, allowing the investigator to compare test statistics constructed from data to the empirical null distribution of a regression-based test statistic, generated using the same set of data. Deterministic studies indicated that when there is a genetic correlation between traits being studied, multivariate approaches are more powerful than univariate approaches [29]. However, the relative gain in power was highly dependent upon the strength of the correlation as well as its direction relative to residual familial (or polygenic) components, as was also noted for animal models [27]. The greatest gain in power occurs when the polygene and major-gene correlations between traits have opposite signs. VC are usually implemented under an assumption of **multivariate normality**, as this ensures computational simplicity in obtaining estimates. However, recent studies [1, 36] have shown lack of convergence of the test for linkage to a limiting chi-squared distribution when the data are not normally distributed and there is a strong residual correlation among sibs, after allowing for the major-gene effect. Methods to allow the construction of accurate tests for nonnormal data include data **trimming** [37], application of **generalized estimating equations** or robust variance estimation [9], construction of hypothesis tests that allow for **kurtosis** in the data [11] or the construction of permutation tests (*see* **Randomization Tests**) [23]. Application of each of these approaches for multivariate data could be rather complex, and tests using either permutation tests or generalized estimating equations are computationally intensive. Sham et al. [32], following some earlier work by Dudoit and Speed [19], have proposed regressing the IBD sharing of relatives onto the trait values. This approach has the advantage of being relatively insensitive to distributional assumptions while maintaining power, compared with variance-components procedures here discussed. This approach requires some further development for application to multivariate traits.

Comparison of multivariate tests have been presented recently in the literature by Allison et al. [2] and Amos et al.[6]. Allison et al. presented results from a large **simulation** study to assess the effectiveness of a bivariate H–E test for linkage versus the univariate H–E test. Their results showed that bivariate analyses can improve the power to detect linkage, with a greater gain in power when the genetic covariance due to a major locus linked to the marker studied is negative and the residual covariance among the traits is positive. They applied slightly different test statistic that was earlier proposed by Amos et al. [7], although the general form is similar. Their alternative test did not follow any well-characterized distributional form. Therefore, they used a **Monte Carlo method** to develop hypothesis tests. Amos et al. [6] performed extensive simulation studies to

compare the power of VC and the multivariate H–E tests proposed by Amos et al. [7]. They also validated the use of software for use in multivariate analysis of quantitative traits, and provided empirical power results for the study of bivariate traits. Amos et al. [6] showed that bivariate VC procedures provided more power than the bivariate Haseman–Elston test at a higher computational cost. Although the MH–E test had lower power than bivariate VC procedures, this procedure still holds promise as a tool for rapidly studying combinations of three or more traits. Because by using simple correlation methods the investigator cannot easily decide which traits to study in multivariate linkage analysis, two approaches can be taken. First, biologically relevant attributes of the traits may allow the investigator to choose combinations of traits for study. Second, the MH–E procedure or newer modifications of it [21, 22] could be used to screen for combinations of variables to be studied. The PC approach suggested by Elston et al. [21] and further studied by Gorlova et al. [22] provides a rapid and efficient method for screening traits using either regression or ML variance components methods. However, full multivariate variance components linkage analysis is suggested where linkages are noted since it provides more meaningful and interpretable results. For sufficiently large sample sizes, the formulation of the MH–E test appeared to be approximately distributed as an F-distribution [6]. Thus, results from analysis using this method should be easily interpretable. A step-up or step-down procedure can be used to decide among traits to be included in analysis [8]. When a set of traits has been identified for further analysis by using the MH–E procedure, VC analysis can be implemented. Ultimately, as combinations of traits are studied, multivariate applications of fully parametric linkage models might be implemented to characterize the specific effects of each allele at a locus.

Model-dependent approaches for genetic linkage fit parameters to model the prevalence of each allele influencing a trait along with a model to describe the distribution of trait values, conditional upon the genotypes and covariates. Because a large number of parameters must be fitted, model-dependent multivariate methods have not been widely used in linkage analysis [12]. As a result, no model-dependent genetic linkage studies of more than two genetically influenced risk factors have been published.

Acknowledgment

This research was partially supported by NIH grants R01HL71917, R01HG02275, R01ES09912, and P30CA16672.

References

[1] Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I. & Blangero, J. (1999). Testing the robustness of the likelihood ratio test in a variance-component quantitative trait loci (QTL) mapping procedure, *American Journal of Human Genetics* **65**, 531–545.

[2] Allison, D.B., Thiel, B., St. Jean, P., Elston, R.C., Infante, M.C. & Schork, N.J. (1998). Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages, *American Journal of Human Genetics* **63**, 1190–1201.

[3] Almasy, L. & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees, *American Journal of Human Genetics* **62**, 1198–1121.

[4] Almasy, L., Dyer, T.D. & Blangero, J. (1997). Bivariate quantitative trait linkage analysis: pleiotropy versus coincident linkages, *Genetic Epidemiology* **14**, 953–958.

[5] Amos, C.I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees, *American Journal of Human Genetics* **54**, 535–543.

[6] Amos, C.I., de Andrade, M. & Zhu, D. (2001). Comparison of multivariate tests for genetic linkage, *Human Heredity* **51**, 133–144.

[7] Amos, C.I., Elston, R.C., Bonney, G.E., Keats, B.J.B. & Berenson, G.S. (1990). A multivariate method for detecting genetic linkage with application to the study of a pedigree with an adverse lipoprotein phenotype, *American Journal of Human Genetics* **47**, 247–254.

[8] Amos, C.I. & Murigande, C. (1992). Preliminary evaluation of linkage between chromosome 1p markers and nevus densities in the Utah data, *Cytogenetic and Cell Genetics* **59**, 173–175.

[9] Amos, C.I., Zhu, D. & Boerwinkle, E. (1996). Assessing genetic linkage and association with robust components of variance approaches, *Annals of Human Genetics* **60**, 143–160.

[10] Blangero, J. & Almasy, L. (1997). Multipoint oligogenic linkage analysis of quantitative traits, *Genetic Epidemiology* **14**, 959–964.

[11] Blangero, J., Williams, J.T. & Almasy, L. (2001). Variance components methods for detecting complex trait loci, *Advances in Genetics* **42**, 151–181.

[12] Bonney, G.E., Lathrop, G.M. & Lalouel, J.-M. (1988). Combined linkage and segregation analysis using regressive models, *American Journal of Human Genetics* **43**, 29–37.

[13] Boomsma, D.I. & Dolan, C.V. (1998). A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores, *Behavior Genetics* **28**, 329–340.

[14] Churchill, G.A. & Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping, *Genetics* **138**, 963–971.

[15] de Andrade, M., Amos, C.I. & Thiel, T.J. (1999). Methods to estimate genetic components of variance for quantitative traits in family studies, *Genetic Epidemiology* **17**, 64–76.

[16] de Andrade, M., Gueguen, R., Visvikis, S., Sass, C., Siest, G. & Amos, C.I. (2002). Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis, *Genetic Epidemiology* **22**, 221–232.

[17] de Andrade, M., Krushkal, J., Yu, L., Zhu, D. & Amos, C. (1998). ACT – A computer package for analysis of complex traits, *American Journal of Human Genetics* **63**, A287.

[18] de Andrade, M., Thiel, T.J., Yu, L. & Amos, C.I. (1997). Assessing linkage in chromosome 5 using components of variance approach: univariate versus multivariate, *Genetic Epidemiology* **14**, 773–778.

[19] Dudoit, S. & Speed, T.R. (2000). A score test for linkage analysis of qualitative and quantitative traits based on identity by descent in sibpairs, *Biostatistics* **1**, 1–26.

[20] Eaves, L.J., Neale, M.C. & Maes, H. (1996). Multivariate multipoint linkage analysis of quantitative trait loci, *Behavior Genetics* **26**, 519–525.

[21] Elston, R.C., Buxbaum, S., Jacobs, K.B. & Olson, J.M. (2000). Haseman and Elston revisited, *Genetic Epidemiology* **19**, 1–17.

[22] Gorlova, O.Y., Amos, C.I., Zhu, D.K., Wang, W., Turner, S. & Boerwinkle, E. (2002). Power of a simplified multivariate test for genetic linkage, *Annals of Human Genetics* **66**, 407–417.

[23] Guerra, R., Wan, Y., Jia, A., Amos, C.I. & Cohen, J.C. (1999). Testing for linkage under robust genetic models, *Human Heredity* **49**, 146–153.

[24] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[25] Hopper, J.L. & Mathews, J.D. (1982). Extensions to multivariate normal models for pedigree analysis, *Annals of Human Genetics* **46**, 373–383.

[26] Iturria, S.J. & Blangero, J. (2000). An EM algorithm for obtaining maximum likelihood estimates in the multiphenotype variance components linkage model, *Annals of Human Genetics* **64**, 349–369.

[27] Jiang, C. & Zeng, Z.B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci, *Genetics* **140**, 1111–1127.

[28] Province, M.A. & Rao, D.C. (1995). General purpose model and a computer program for combined segregation and path analysis (SEGPATH): automatically creating computer programs from symbolic language model specifications, *Genetic Epidemiology* **12**, 203–221.

[29] Schmitz, S., Cherny, S.S. & Fulker, D.W. (1998). Increase in power through multivariate analyses, *Behavior Genetics* **28**, 357–363.

[30] Schork, N.J. (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations, *American Journal of Human Genetics* **53**, 1306–1319.

[31] Self, S.G. & Liang, K.-L. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of American Statistical Association* **82**, 605–610.

[32] Sham, P.C., Purcell, S., Cherny, S.S. & Abecasis, G.R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees, *American Journal of Human Genetics* **71**, 238–253.

[33] Sorant, A.J.M. & Elston, R.C. (1991). *MAXFUN: A Subroutine Package for Function Maximization (A user's Guide to MAXFUN, Version 5.1)*. Department of Biometry and Genetics, LSU Medical Center, New Orleans, LA.

[34] Todorov, A.A., Vogler, G.P., Gu, C., Province, M.A., Li, Z., Heath, A.C. & Rao, D.C. (1998). Testing causal hypotheses in multivariate linkage analysis of quantitative traits: general formulation and application to sibpair data, *Genetic Epidemiology* **15**, 263–278.

[35] Vogler, G.P., Tang, W., Nelson, T.L., Hofer, S.M., Grant, J.D., Tarantino, L.M. & Fernandez, J.R. (1997). A multivariate model for the analysis of sibship covariance structure using marker information and multiple quantitative traits, *Genetic Epidemiology* **14**, 921–926.

[36] Wan, Y., de Andrade, M.A. & Amos, C.I. (1998). Genetic Linkage analysis using lognormal variance components, *Annals of Human Genetics* **62**, 521–530.

[37] Wang, J., Guerra, R. & Cohen, J. (1998). Statistically robust approaches for sib-pair linkage analysis, *Annals of Human Genetics* **62**, 349–359.

[38] Williams, J.T., Van Eerdewegh, P. & Blangero, Almasy.L. (1999). Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results, *American Journal of Human Genetics* **65**, 1134–1147.

MARIZA DE ANDRADE &
CHRISTOPHER I. AMOS

Linkage Disequilibrium

Linkage disequilibrium, more appropriately termed allelic association or allelic disequilibrium, refers to the nonrandom association between the alleles at two or more genetic loci in a natural breeding population (*see* **Gene**). The concept of linkage disequilibrium was postulated by population geneticists in theoretical studies of the consequences of random mating on the distribution of alleles and allelic combinations (**genotypes**) at multiple loci. Further theoretical studies have shown that in most instances this nonrandom association declines rapidly with evolutionary time, and as a function of the recombination frequency between the loci. However, with natural selection, allelic associations may persist in a population for long evolutionary time periods. With the availability of high-resolution genetic maps in the human and many other species, empirical studies of linkage disequilibrium in different genomic segments have been initiated. These studies have not only shed light on the distribution of alleles and allelic combinations in the genome, but also on the **population genetic** mechanisms that are likely to lead to the observed distribution of linkage disequilibrium across loci. In particular, these studies suggest that in certain circumstances, linkage disequilibrium can be used to infer the location of a disease-causing gene by studying **disease–marker associations**.

Hardy–Weinberg Law

The rediscovery of Mendelism (*see* **Mendel's Laws**) coincided with the identification of hundreds of phenotypes, in diverse species, whose inheritance could be explained by a dominant or recessive allele. In crosses, the familiar $3:1$ segregation ratio (*see* **Segregation Analysis, Classical**) was consistently observed for these phenotypes, leading to the suggestion that the population frequency of dominant to recessive genotypes should also be in the 3:1 ratio. This suggestion was, however, easily refuted by observations in natural populations. This dilemma was resolved by Hardy and Weinberg who showed by theoretical analysis that the frequency of genotypes in populations, under the simplifying assumptions of random mating in a population of infinite size and the absence of mutation, migration and selection, was solely determined by the frequencies of the constituent alleles [7] (*see* **Hardy–Weinberg Equilibrium**). Thus, if the allele frequencies of the dominant, D, and recessive, d, alleles at an autosomal locus were p and q, respectively, then the frequencies of the genotypes DD, Dd, and dd are $p^2, 2pq$, and q^2, respectively. The same results apply to females at an X-linked locus, with males having alleles in proportion to their population frequencies. In the presence of dominance, the dominant and recessive phenotypes have population frequencies of $p^2 + 2pq = 1 - q^2$ and q^2, respectively. Under these assumptions, Hardy and Weinberg also showed that, irrespective of the initial genotype frequency distributions, one generation of random mating generates the above genotype frequencies and that these allele and genotype frequencies do not change further over time, i.e. the frequencies are at an equilibrium state.

Shortly after, in the 1910s, with the discovery of genetic linkage, theoretical studies were initiated to investigate the consequences of random mating when two linked loci were considered, and under the same simplifying assumptions used by Hardy [7]. Jennings [10] and Robbins [18] showed that with two linked genes, the population, once again, approached an equilibrium state in which the alleles at the two loci associated at random; Geiringer [6] solved the problem with three linked factors. If two loci, one with alleles D and d with frequencies p and q, respectively, and a second with alleles E and e with frequencies r and s, respectively, are linked with recombination frequency θ ($0 \leq \theta \leq 1/2$), then, at equilibrium, the frequency of homozygotes such as DDEE is p^2r^2, the frequency of single **heterozygotes** such as DdEE is $2pqr^2$, and the frequency of the double heterozygote DdEe is $4pqrs$. In random mating populations, the equilibrium population frequency of any genotype, at one or more loci, is determined solely by the constituent allele frequencies at individual loci.

Linkage Disequilibrium

Jennings [10], Robbins [18], and Geiringer [6] also determined the rate of approach to equilibrium when more than one locus is involved and gave the genotypic distributions after any finite number of generations. These authors showed that, under random mating, the frequency of any multilocus genotype, g generations from an initial condition, is determined

by the products of frequencies of the four haplotypes (allele combinations) DE, De, dE, and de, and that the haplotype frequencies change over time. After g generations, the haplotype frequencies are:

$$\mathrm{DE}: h_1 = pr + \varepsilon,$$
$$\mathrm{De}: h_2 = ps - \varepsilon,$$
$$\mathrm{dE}: h_3 = qr - \varepsilon,$$
$$\mathrm{de}: h_4 = qs + \varepsilon.$$

In each generation there is a specific departure of the haplotype frequencies from the equilibrium values; this excess or deficiency is denoted as ε and termed the coefficient of allelic association or linkage disequilibrium. Note that the numerical value of ε is bounded, since each haplotype frequency is nonnegative, less than unity, and the four haplotype frequencies add to 1, as follows:

$$-\min(pr, qs) \leq \varepsilon \leq \min(ps, qr).$$

With genetic recombination between the D and E locus, the coefficient of allelic association declines every generation so that in successive generations they are related as:

$$\varepsilon' = (1 - \theta)\varepsilon.$$

Thus, in g generations, the total decline is,

$$\varepsilon_g = (1 - \theta)^g \varepsilon_0,$$

where ε_0 is the coefficient of allelic association at generation zero (initial condition). Note that in any generation, $\varepsilon = h_1 h_4 - h_2 h_3$.

The above equations suggest that linkage disequilibrium occurs only as a nonequilibrium phenomenon since it always declines to zero. However, the rate of decline is determined by the recombination value with a half-life of $-\ln 2/\ln(1 - \theta)$; the inverse relationship with the recombination value suggests that, for very close linkage, disequilibrium may persist for long evolutionary time periods.

Causes of Linkage Disequilibrium

An extensive body of population genetics literature shows how various forms of natural selection acting on individual genes can lead to permanent, equilibrium association of the alleles at two loci, even in the absence of linkage [15]. Additionally, when two loci are linked, natural selection at one locus can lead to the apparent selection at a linked locus ("hitchhiking") and permanent linkage disequilibrium. However, and as stated earlier, in the absence of natural selection no permanent associations are expected at two linked loci. Thus, the term linkage disequilibrium is a misnomer since linkage is neither necessary nor sufficient for permanent associations to occur; the descriptive term allelic (nonrandom) association is more appropriate. Other circumstances under which allelic associations can occur, although not permanently, are population admixture and population subdivision (*see* **Admixture in Human Populations**). In the latter case, a population with hidden subpopulations, each of which differs in allele frequencies and/or allelic associations, but treated as a single population, can also create allelic disequilibrium. Further details on these and other theoretical models are provided in [14–16].

In many large, random mating populations allelic associations are nevertheless observed. In humans, such observations, restricted to very closely linked **blood group** genes, such as those within the Rhesus (C, D, and E loci) and MNS (M and S loci), or within the **HLA system** of genes, have been known for a long time [1]. More recently, with the availability of multiple DNA **polymorphisms** in a small genomic region, these studies have gained in popularity. In fact, polymorphic alleles at multiple sites all within 20–30 kilobases of DNA demonstrate allelic associations [3]. These observations are best explained by the nonequilibrium state of the human population and the expectation of a slow decay of linkage disequilibrium when the maximal recombination rate within a region is less than 0.0005 per generation [3]. These observations have found multiple uses such as for associating specific mutations with a molecular haplotype in a genomic region [12] and for mapping the location of a mutation to a small genetic interval [13].

Parameterization and Estimation

Allelic associations, parameterized as the coefficient of linkage disequilibrium ε, are most easily and efficiently estimated from haplotype data. However, for most diploid organisms it is not possible to derive haplotypes from diploid two-locus genotypes unless family data (i.e. parents, offspring, and other relatives are sampled and studied) are also available (*see*

Haplotype Analysis). Then, haplotype frequencies are estimated from the relative frequencies of haplotypes in a sample of independent families and by counting only independent haplotypes within each family. Tests of allelic associations ($H_0: \varepsilon = 0$) are based on the significance of a chi-squared statistic with 1 df comparing the observed and expected (under equilibrium) haplotype frequencies. In a sample of n haplotypes studied, if allelic associations are present, the expected value of this **chi-square** statistic is $\chi^2 = n\rho^2$, where,

$$\rho = \frac{\varepsilon}{(pqrs)^{1/2}}.$$

Thus, the **correlation** in the chi-square **contingency table** of observations is another measure of linkage disequilibrium, one that is a natural measure based on the test of association. In many investigations, this latter measure has been used since ρ appears to be less dependent on the allele frequencies than ε. A measure that is not dependent on allele frequencies and finding more popular use in these studies is Yule's measure of association [13]:

$$A = \frac{h_1h_4 - h_2h_3}{h_1h_4 + h_2h_3}.$$

The above results are for loci with two alleles per locus, whereas many polymorphic markers have multiple alleles. Then, coefficients for allelic association have to be defined for each allele pair at two loci, allele triples for three loci, and so on. A discussion of these multiple disequilibria and tests of hypotheses on them is discussed in [19].

When family data are unavailable, linkage disequilibrium parameters can be efficiently estimated from population samples using iterative methods such as the **EM algorithm**. In the genetic context, this was first proposed by Hill [9] for two loci; additional methods and calculations of sample size for predefined statistical **power** was studied by Brown [2]. In addition, Brown [2] considers the appropriate parameterization of disequilibria for multiple loci taken together (i.e. the nonrandom association of alleles at multiple loci) once the pairwise disequilibria have been considered. Tests of significance (*see* **Hypothesis Testing**) are performed by appropriate modifications of the chi-square test alluded to above. Of greater importance, particularly with DNA polymorphism data, is the existence of multiple alleles at each locus studied. In this circumstance there are several possible tests of disequilibrium, such as an omnibus test or conditional tests on specific collection of alleles. Weir & Cockerham [19] have provided a theoretical and statistical account of this situation.

It is clear that accurate estimates of allelic associations require very large sample sizes when the allele frequencies, at one of the two loci studied, are close to 0 or 1. This occurs whenever **disease–marker associations** are evaluated. In these circumstances a conditional sampling strategy, in which marker genotypes are evaluated within classes of affected and unaffected individuals, is very efficient. Chakravarti et al. [4] have provided a **maximum likelihood** method for estimation of linkage disequilibrium statistics from conditional marker genotype data. These methods are useful for mapping disease genes with respect to a map of DNA markers (*see* **Genetic Map Functions**) [11, 13].

Current Applications

Studies of allelic associations across the genome and in various natural populations are now beginning, with the availability of high-resolution genetic maps in a number of species. So far, the primary purpose of these studies has been to probe the population structure of natural populations. Since in large randomly mating populations no allelic associations are expected, the finding of widespread associations will suggest that these populations are either not in equilibrium or that they have been recently established from a small number of founders. The fitting of specific population genetic models can then elucidate whether the nonequilibrium nature of these populations is due to genetic drift, natural selection, subdivision, or migration. In the human, these types of studies all suggest that modern humans have descended from a limited pool of founders ($\sim$10 000) approximately 200 000 years or so before the present.

In the human, the greatest application of studies of allelic association is to find the location of a disease gene with respect to a map of DNA markers, once the gene has been genetically localized to a DNA segment under 1000 kilobases. This was first demonstrated with the molecular cloning of the gene for cystic fibrosis [13]. A theoretical basis for disease–marker associations owing to common descent of a specific mutation from an ancestor, such as for cystic fibrosis, has been given by Hastbacka et al.

[8] and Puffenberger et al. [17]. The prospects for such disease mapping, as an aid to the molecular cloning of the mutant gene, has also been discussed by Jorde [11]. Recent studies of variation patterns of the human genome suggests that nonrandom associations between the most common type of sequence variant, the single nucleotide polymorphism (SNP), is highly clustered. Gabriel et al. [5] have shown that nonrandom associations in several human populations occur as blocks of high association with apparent random association between blocks. This highly punctuate pattern seems to be more prominent in non-African than African samples. An international project, the HapMap project, to decipher these patterns in multiple samples from across the world is currently underway and promises to uncover a wide variety of statistical problems that beg for a solution.

References

[1] Bodmer, W.F. & Cavalli-Sforza, L.L. (1976). *Genetics, Evolution, and Man*. W.H. Freeman, San Francisco.

[2] Brown, A.H.D. (1975). Sample sizes required to detect linkage disequilibrium between two or three loci, *Theoretical Population Biology* **8**, 184–201.

[3] Chakravarti, A., Buetow, K.H., Antonarakis, S.E., Waber, P.G., Boehm, C.D. & Kazazian, H.H. (1984). Nonuniform recombination within the human β-globin gene cluster, *American Journal of Human Genetics* **36**, 1239–1258.

[4] Chakravarti, A., Li, C.C. & Buetow, K.H. (1984). Estimation of the marker gene frequency and linkage disequilibrium from conditional marker data, *American Journal of Human Genetics* **36**, 177–186.

[5] Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. & Altshuler, D. (2002). The structure of haplotype blocks in the human genome, *Science* **296**, 2225–2229.

[6] Geiringer, H. (1945). Further remarks on linkage in Mendelian heredity, *Annals of Mathematical Statistics* **16**, 390–393.

[7] Hardy, G.H. (1908). Mendelian proportions in a mixed population, *Science* **28**, 49–50.

[8] Hastbacka, J., de la Chapelle, A., Kaitial, I., Sistonen, P., Weaver, A. & Lander, E.S. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland, *Nature Genetics* **2**, 204–211.

[9] Hill, W.G. (1974). Estimation of linkage disequilibrium in randomly mating populations, *Heredity* **33**, 229–239.

[10] Jennings, H.S. (1917). The numerical results of diverse systems of breeding with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage, *Genetics* **12**, 97–154.

[11] Jorde, L.B. (1995). Linkage disequilibrium as a gene-mapping tool, *American Journal of Human Genetics* **56**, 11–14.

[12] Kazazian, H.H., Stuart, O.H., Markham, A.F., Chapman, C.R., Youssoufian, H. & Waber, P.G. (1984). Quantification of the close association between DNA haplotypes and specific β-thalassaemia mutations in mediterraneans, *Nature* **310**, 152–154.

[13] Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. & Tsui, L.C. (1989). Identification of the cystic fibrosis gene: Genetic analysis, *Science* **245**, 1073–1808.

[14] Kimura, M. & Crow, J.F. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.

[15] Kimura, M. & Ohta, T. (1971). *Theoretical Aspects of Population Genetics*, Princeton University Press, Princeton.

[16] Li, C.C. (1976). *First Course in Population Genetics*. The Boxwood Press, Pacific Grove.

[17] Puffenberger, E.G., Kauffman, E.R., Bolk, S., Matise, T.C., Washington, S.S., Angrist, M., Weissenbach, J., Garver, K.L., Mascari, M., Ladda, R., Slaugenhaupt, S.A. & Chakravarti, A. (1994). Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22, *Human Molecular Genetics* **3**, 1217–1225.

[18] Robbins, R.B. (1918). Some applications of mathematics to breeding problems. III, *Genetics* **3**, 375–389.

[19] Weir, B.S. & Cockerham, C.C. (1978). Testing hypotheses about linkage disequilibrium with multiple alleles, *Genetics* **88**, 633–642.

ARAVINDA CHAKRAVARTI

Linkage Information Content

Several authors have studied the measurement of **marker polymorphism**. Chakraborty et al. [2] proposed a criterion to evaluate whether a mating pair randomly drawn from a population is potentially informative for **linkage** studies. They defined a mating pair to be potentially informative if it is capable of producing offspring of at least two different phenotypes. Botstein et al. [1] proposed a similar measure, called the **polymorphism information content** (PIC) value. The PIC value is defined as the probability that the marker **genotype** of a given offspring will allow one to deduce which of the two marker alleles (*see* **Gene**) it received from the affected parent (in the absence of crossing-over) (*see* **Linkage Analysis, Model-based**), assuming one of the parents is affected with a rare dominant disease. Guo & Elston [3] introduced the concept of transmission informativeness, the probability of being able to deduce which of the two marker alleles a parent has transmitted given the marker genotypes of the offspring and both parents, and showed that the PIC value is equal to the average transmission informativeness over all possible parent–offspring trios. They hence proved that the PIC value is in fact a general measure of how informative a marker is regardless of the mode of inheritance of the trait being studied.

The use of model-free linkage analysis (*see* **Linkage Analysis, Model-free**) on samples of relative pairs has become common practice in genetic studies of **complex diseases**. This is based on identity-by-descent (ibd) probabilities; hence, a marker's usefulness for detecting linkage by such a method depends on the probability of being able to determine the ibd probabilities for each particular type of relative pair. Because the informativeness measures discussed above are not adequate for such a purpose, the linkage information content (LIC) value was developed [3, 6] to measure the informativeness of a marker for determining the ibd-sharing status of particular types of relative pairs. In the following section, the concept and the calculation of LIC are introduced for five types of relative pairs: full sib, half-sib, grandparent–grandchild, first cousin, and avuncular pairs.

Linkage Information Content Value for Specific Types of Relative Pairs

For a particular marker and a particular pair of relatives A and B, $\mathrm{LIC_{AB}}$ is defined as the probability of knowing how many alleles A and B share ibd at that marker. For a particular type of relative pair R, $\mathrm{LIC_R}$ is the average value of $\mathrm{LIC_{AB}}$ over all AB pairs in the population, i.e.

$$\mathrm{LIC_R} = \sum_{\substack{\text{all possible} \\ \text{AB pairs}}} \Pr(\mathrm{AB})\Pr(\text{knowing ibd status} \mid \text{AB pair}).$$

In what follows the alleles at a marker locus are represented by A_i, A_j, A_k and A_l, and it is understood that different subscripts indicate different alleles.

Full Sib

To be able to determine whether a sib pair shares the same allele from a parent, that parent has to be heterozygous. If one parent is heterozygous and the other is homozygous (mating types: $\mathrm{A}_i\mathrm{A}_i \times \mathrm{A}_i\mathrm{A}_j$ or $\mathrm{A}_i\mathrm{A}_i \times \mathrm{A}_j\mathrm{A}_k$), then the ibd status for the sib pair can be determined for only one (the heterozygous) parent. In the case when both parents are heterozygous (either the same heterozygous genotype, i.e. mating type: $\mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_i\mathrm{A}_j$; or different heterozygous genotype, i.e. $\mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_i\mathrm{A}_k$ or $\mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_k\mathrm{A}_l$) and we know which allele each parent has transmitted to at least one of the sibs, then the ibd status for the sib pair from both parents is determined [6]. The LIC value of a marker for sib pairs is defined to be the sum of the probability of being able to determine the sharing status from both parents and half of the probability of being able to determine the sharing status from only one parent, i.e. $\mathrm{LIC_S} = \mathrm{LIC_{S2}} + \frac{1}{2}\mathrm{LIC_{S1}}$, where

$$\begin{aligned}\mathrm{LIC_{S2}} = {} & \Pr(\mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_i\mathrm{A}_j \text{ parents})\Pr(\text{knowing sib pair sharing status} \mid \mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_i\mathrm{A}_j \text{ parents}) \\ & + \Pr(\mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_i\mathrm{A}_k \text{ parents})\Pr(\text{knowing sib pair sharing status} \mid \mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_i\mathrm{A}_k \text{ parents}) \\ & + \Pr(\mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_k\mathrm{A}_l \text{ parents})\Pr(\text{knowing sib pair sharing status} \mid \mathrm{A}_i\mathrm{A}_j \times \mathrm{A}_k\mathrm{A}_l \text{ parents}),\end{aligned}$$

and

$$\begin{aligned}\mathrm{LIC_{S1}} = {} & \Pr(\mathrm{A}_i\mathrm{A}_i \times \mathrm{A}_i\mathrm{A}_j \text{ parents})\Pr(\text{knowing sib pair sharing status} \mid \mathrm{A}_i\mathrm{A}_i \times \mathrm{A}_i\mathrm{A}_j \text{ parents}) \\ & + \Pr(\mathrm{A}_i\mathrm{A}_i \times \mathrm{A}_j\mathrm{A}_k \text{ parents})\Pr(\text{knowing sib pair sharing status} \mid \mathrm{A}_i\mathrm{A}_i \times \mathrm{A}_j\mathrm{A}_k \text{ parents})\end{aligned}$$

The overall LIC for sib pairs, on simplification, is given for a marker with allele frequencies p_{i} by

$$\mathrm{LIC_S} = 1 - \sum_i p_i^2 + \frac{1}{2}\sum_i p_i^4 - \frac{1}{2}\left(\sum_i p_i^2\right)^2.$$

Half-sib

The LIC value for half-sib pairs is the probability of being able to deduce which marker allele the common parent has transmitted to each child in the half-sib pair. To be informative, the common parent of a half-sib pair has to be heterozygous, and neither the child nor the spouse can have the same heterozygous genotype in each of the two nuclear families [3]. Therefore, the LIC value for half-sibs is given by

$$\begin{aligned}\mathrm{LIC_H} &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 2p_i p_j (1 - p_i p_j)^2 \\ &= 1 - \sum_i p_i^2 - 2\left(\sum_i p_i^2\right)^2 + 2\sum_i p_i^4 \\ &\quad + \left(\sum_i p_i^3\right)^2 - \sum_i p_i^6.\end{aligned}$$

Grandparent–Grandchild, First Cousin, and Avuncular Pairs

The LIC value for grandparent–grandchild, first cousin, and avuncular pairs can be calculated by

$$\begin{aligned}\mathrm{LIC_R} = \sum {} & \Pr(\text{grandparental mating type}) \\ & \Pr(\text{middle generation genotypes} \mid \text{grandparental mating type}) \\ & \Pr(\text{knowing ibd status} \mid \text{grandparental mating type and middle generation genotypes})\end{aligned}$$

for type R relative pairs. The mating type probabilities are determined assuming random mating. The calculation of the probability of the middle generation genotype(s) given the grandparental mating type is straightforward. The conditional probability of knowing ibd status for a particular type of relative pair can be obtained for a given mating type and middle generation genotypes (see Table 1 of [6] for details). For example, if the mating type is $\mathrm{A}_i\mathrm{A}_i \times \mathrm{A}_j\mathrm{A}_j$, then all the children will have genotype $\mathrm{A}_i\mathrm{A}_j$ with probability 1. In the case of grandparent–grandchild pairs, we know whether the grandchild has inherited an allele from a grandparent if and only if we know which allele the parent ($\mathrm{A}_i\mathrm{A}_j$) transmitted to the grandchild, which has a probability of $1 - p_i p_j$. That is, the conditional probability of knowing the ibd status for the grandparent–grandchild pair is $1 - p_i p_j$. In the case of avuncular pairs, the conditional probability of knowing ibd status is 0 because we have no

Table 1 The maximum LIC value for full sib, half-sib, grandparent–grandchild, first cousin, and avuncular pairs, for a marker with n alleles

Type of relative pair	Maximum LIC value
Full sib	$\dfrac{(n-1)(2n^2-1)}{2n^3}$
Half-sib	$\dfrac{(n+1)^2(n-1)^3}{n^5}$
Grandparent–grandchild	$\mathrm{LIC_G} = \dfrac{(n+1)^2(n-1)^3}{n^5}$
First cousin	$\mathrm{LIC_C} = \dfrac{(n-1)(4n^6-n^5-7n^4+2n^3+5n^2-2n-2)}{4n^7}$
Avuncular	$\mathrm{LIC_A} = \dfrac{(n-1)(2n^4-n^3-3n^2+1)}{2n^5}$

way to identify whether the A_i and/or A_j alleles are identical by descent for the two siblings in the middle generation. The LIC value for each of the three types of relative pairs can thus be obtained by summing up the products of the probabilities of mating type, the middle generation genotype(s), and the conditional probabilities of knowing ibd status. Specifically, the LIC value for grandparent–grandchild pair is

$$\begin{aligned}\mathrm{LIC_G} &= \sum_i \sum_{j\neq i} p_i^2 p_j^2 (1-p_i p_j) + \sum_i \sum_{j\neq i} 4p_i^3 p_j \\ &\times \left(\frac{1}{2}\right)(1-p_i p_j) + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} 2p_i^2 p_j p_k \left(\frac{1}{2}\right) \\ &\times [(1-p_i p_j) + (1-p_i p_k)] \\ &+ \sum_i \sum_{j\neq i} \sum_{k\neq i,j} 4p_i^2 p_j p_k \left(\frac{1}{4}\right)[(1-p_i p_j) \\ &+ (1-p_i p_k) + (1-p_j p_k)].\end{aligned}$$

The LIC value for first cousin pairs is given by

$$\begin{aligned}\mathrm{LIC_C} &= \sum_i \sum_{j\neq i} p_i^2 p_j^2 \left(\frac{1}{2}\right)(1-p_i p_j)^2 + \sum_i \sum_{j\neq i} 4p_i^3 \\ &\times p_j \left(\frac{1}{4}\right)\left[\frac{1}{2}(1-p_i p_j) + \frac{1}{2}(1-p_i p_j)\right. \\ &\left. + \frac{3}{4}(1-p_i p_j)^2\right] + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} 2p_i^2 p_j p_k \left(\frac{1}{4}\right) \\ &\times \left[\frac{3}{4}(1-p_i p_j)^2 + \frac{3}{4}(1-p_i p_j)(1-p_i p_k)\right. \\ &\left. + \frac{3}{4}(1-p_i p_j)(1-p_i p_k) + \frac{3}{4}(1-p_i p_k)^2\right] \\ &+ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 4p_i^2 p_j^2 \left(\frac{1}{16}\right)[(1-p_i p_j) + 1 \\ &+ (1-p_i p_j) + 2(1-p_i p_j)^2 + (1-p_i p_j) \\ &+ 1 + (1-p_i p_j)] + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} 4p_i^2 p_j p_k \\ &\times \left(\frac{1}{16}\right)\left[\frac{1}{2}(1-p_i p_j) + \frac{1}{2}(1-p_i p_k) + 1\right. \\ &\left. + \frac{1}{2}(1-p_i p_j) + (1-p_i p_j)^2 + 1 + (1-p_i p_j)\right. \\ &\left. \times (1-p_j p_k) + \frac{1}{2}(1-p_i p_k) + 1 + (1-p_i p_k)^2\right. \\ &\left. + (1-p_i p_k)(1-p_j p_k) + 1 + (1-p_i p_j)\right. \\ &\left. \times (1-p_j p_k) + (1-p_i p_k)(1-p_j p_k)\right. \\ &\left. + (1-p_j p_k)^2\right] + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} \sum_{l\neq i,j,k} p_i p_j p_k p_l \\ &\times \left(\frac{1}{16}\right)[(1-p_i p_k)^2 + (1-p_i p_k)(1-p_i p_l) \\ &+ (1-p_i p_k)(1-p_j p_k) + 1 + (1-p_i p_l) \\ &\times (1-p_i p_k) + (1-p_i p_l)^2 + 1 + (1-p_i p_l) \\ &\times (1-p_j p_l) + (1-p_j p_k)(1-p_i p_k) + 1 \\ &+ (1-p_j p_k)^2 + (1-p_j p_k)(1-p_j p_l) + 1 \\ &+ (1-p_j p_l)(1-p_i p_l) + (1-p_j p_l) \\ &\times (1-p_j p_k) + (1-p_j p_l)^2].\end{aligned}$$

The LIC value for avuncular pairs is

$$\begin{aligned}\mathrm{LIC_A} &= \sum_i \sum_{j\neq i} 4p_i^3 p_j \left(\frac{1}{4}\right)\left[\frac{1}{2}(1-p_i p_j)\right. \\ &\left. + \frac{1}{2}(1-p_i p_j)\right] + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} 2p_i^2 p_j p_k \left(\frac{1}{4}\right) \\ &\times \left[\frac{1}{2}(1-p_i p_j) + \frac{1}{2}(1-p_i p_k) + \frac{1}{2}(1-p_i p_j)\right. \\ &\left. + \frac{1}{2}(1-p_i p_k)\right] + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 4p_i^2 p_j^2 \left(\frac{1}{16}\right) \\ &\times [1 + 2(1-p_i p_j) + 1 + 1 + 2(1-p_i p_j) + 1] \\ &+ \sum_i \sum_{j\neq i} \sum_{k\neq i,j} 4p_i^2 p_j p_k \left(\frac{1}{16}\right)[1 + (1-p_i p_j) \\ &+ (1-p_i p_k) + 3 + (1-p_j p_k) + 2 \\ &+ (1-p_j p_k) + 1 + (1-p_i p_j) + (1-p_i p_k) + 1] \\ &+ \sum_i \sum_{j\neq i} \sum_{k\neq i,j} \sum_{l\neq i,j,k} p_i p_j p_k p_l \left(\frac{1}{16}\right) \\ &\times [1 + (1-p_i p_l) + (1-p_j p_k) + 1 + (1-p_i p_k) \\ &+ 1 + 1 + (1-p_j p_l) + (1-p_i p_k) + 1 + 1 \\ &+ (1-p_j p_l) + 1 + (1-p_i p_l) + (1-p_j p_k) + 1].\end{aligned}$$

The maximum value of LIC is attained when a marker has n equally frequent alleles and is summarized in Table 1 for the five types of relative pairs.

Conclusion

The success of linkage analysis in studying complex diseases is significantly dependent on the informativeness of the markers used. We often assume that markers are fully informative, which is usually not true in practice. LIC values measure how informative a marker is for ibd sharing for each type of relative pair, and so provide an appropriate measure for a model-free linkage analysis, especially in the design of linkage studies that use relative pairs [4, 5]. Niu et al. [7] developed a software package called POLYMORPHISM that calculates LIC values for each of these five types of relative pairs. The LIC values calculated by POLYMORPHISM may be used to determine the average informativeness of markers for IBD sharing status and can be entered into DESPAIR, a program in the software package S.A.G.E. [8], to obtain the optimal two-stage global search design for linkage studies.

References

[1] Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms, *American Journal of Human Genetics* **32**, 314–331.

[2] Chakraborty, R., Fuerst, P.A. & Ferrell, R.E. (1979). Potential information in family studies of linkage, in *Genetics Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease*, C.F. Sing & M. Skolnick, eds. Alan R. Liss, New York, pp. 297–303.

[3] Guo, X. & Elston, R.C. (1999). Linkage information content of polymorphic markers, *Human Heredity* **49**, 112–118.

[4] Guo, X. & Elston, R.C. (2000). Two-stage global search designs for linkage analysis. I. Use of the mean statistic for affected sib pairs, *Genetic Epidemiology* **18**, 97–110.

[5] Guo, X. & Elston, R.C. (2000). Two-stage global search designs for linkage analysis. II. Including discordant relative pairs in the study, *Genetic Epidemiology* **18**, 111–127.

[6] Guo, X., Olson, J.M., Elston, R.C. & Niu, T. (2001). The linkage information content value of polymorphism genetic markers in model-free linkage analysis, *Human Heredity*, in press.

[7] Niu, T., Struk, B. & Lindpaintner, K. (2001). Statistical considerations for genome-wide scans: design and application of a novel software package POLYMORPHISM, *Human Heredity* **52**, 102–109.

[8] S.A.G.E.: Statistical Analysis for Genetic Epidemiology, Release 3.1 (1998). Computer program package available from the Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio.

XIUQING GUO

Linkage Analysis, Model-free

Model-free linkage methods, in contrast to **model-based linkage methods**, do not depend on prior specification of a model of inheritance for the disease or trait of interest. In other words, the frequencies and **penetrances** of disease **genotypes** need not be known in advance, and functions of these quantities may in fact be estimated in conjunction with linkage parameters. It is important to recognize, however, that many of the methods do rely on assumptions about the underlying genetic model and some methods are in fact parametric and semiparametric in nature. In this Section, two general types of model-free linkage methods will be distinguished – those designed for qualitative traits and those designed for quantitative traits – although both theory and applications of these two groups of methods overlap. Model-free linkage methods typically evaluate marker locus identity-by-descent (IBD) relationships among family members, often pairs of siblings, and thus are often referred to as relative-pair, or sib-pair, methods.

Identity by Descent

A pair of related individuals shares an allele IBD if that allele has a common ancestral source, i.e. the same chromosome of the same ancestor. In the context of linkage analysis, the common ancestor is taken to be a recent ancestor – one within the sampled pedigree. For example, if the pair are siblings, the common ancestors are their parents, and the sibs may have inherited the same paternal allele and/or maternal allele. Let f_i, $i = 0, 1, 2$, be the prior (unconditional) probability that a relative pair shares i alleles IBD at a single marker locus, and $\hat{f}_i$ the estimate of f_i conditional on available marker data (*see* **Genetic Markers**). Now, let π be the proportion of alleles a relative pair shares IBD at a single locus, and $\hat{\pi} = \frac{1}{2}\hat{f}_1 + \hat{f}_2$ the estimate of π conditional on available marker data. Table 1 gives the prior (unconditional) distribution of π for different types of relative pairs.

Computation of the $\hat{f}_i$ for a single marker locus, using available pedigree marker data, was first proposed by Haseman & Elston [24] for nuclear families, and later by Amos et al. [3] for extended pedigrees.

Table 1 Prior distribution of π for relative pairs

Type of relative pair	π			
	0	$\frac{1}{2}$	1	E(π)
Sibling	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
Second degree	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{4}$
Third degree	$\frac{3}{4}$	$\frac{1}{4}$	0	$\frac{1}{8}$

Let I_{m} represent the available family marker data. Then

$$\hat{f}_i = \frac{\Pr(\pi = i/2, I_{\mathrm{m}})}{\Pr(I_{\mathrm{m}})}$$

is a general form for estimating $\hat{f}_i$. The denominator is the probability, or likelihood, of the pedigree marker data, and may be computed using an **Elston–Stewart** ("peeling") **algorithm**; the numerator can be written as a sum of the terms of the denominator consistent with sharing i alleles IBD, each term representing a phase-known pedigree genotype. More detailed algorithms for computing $\hat{f}_i$ for different types of relative pairs were given by Amos et al. [3]. Whittemore & Halpern [52] also presented a peeling algorithm for computing the probabilities of IBD relationships among the genes of pedigree members, and then showed how to use these probabilities to calculate the probability of any combination of genotypes or phenotypes for the pedigree members.

Estimation of multipoint IBD probabilities has also been explored. Let d represent the **genetic distance** from an arbitrary origin on a marker map with known intermarker distances. To compute the $\hat{f}_{di}$, the estimated allele-sharing distribution at location d, Kruglyak & Lander [29] employed a hidden **Markov chain** model that assumes that the π at consecutive loci behave in a first-order Markov manner; i.e. $\Pr(\pi_k|\pi_1, \pi_2, \ldots, \pi_{k-1}) = \Pr(\pi_k|\pi_{k-1})$, for loci ordered 1 through k on a chromosome. Inheritance at each location d is represented by an inheritance vector $V(d)$, in which each component corresponds to a particular meiosis in the pedigree and the component takes a value of 0 or 1 according to whether the paternal or maternal allele is transmitted. Computation of the probability distribution for $V(d)$ conditional on the marker data can be accomplished by considering pairs of loci successively. Additional computational speed is achieved by taking advantage of the fact that phase differences in the founders

are equivalent and have equal probabilities [31]. An even faster algorithm uses a divide-and-conquer method and allows for meiosis-specific recombination fractions at virtually no additional cost [27]. These algorithms are all modifications of the Lander–Green algorithm [30, 32]. Sobel & Lange [46] developed a **Markov chain Monte Carlo** algorithm to approximate multipoint IBD-sharing estimates in larger pedigrees.

Given multipoint IBD-sharing estimates at marker locus locations, **regression** models can also be used to obtain IBD-sharing estimates at points between two markers. Let f_{ij} be the prior (unconditional) joint probability that a sib-pair shares i alleles IBD at one marker locus and j alleles IBD at a second, usually linked, marker locus, and let $\hat{f}_{ij}$ be estimates, conditional on the available marker data, that account for the recombination fraction, assumed to be known, between the two markers. Let $\hat{f}_{i.} = \sum_j \hat{f}_{ij}$, $\hat{f}_{.j} = \sum_i \hat{f}_{ij}$, $\hat{\pi}_1 = \hat{f}_{1.}/2 + \hat{f}_{2.}$, $\hat{\pi}_2 = \hat{f}_{.1}/2 + \hat{f}_{.2}$, and $\widehat{\pi_1\pi_2} = \hat{f}_{11}/4 + (\hat{f}_{12} + \hat{f}_{21})/2 + \hat{f}_{22}$. Given these multipoint estimates of IBD-sharing at two adjacent loci and assuming no crossover interference (*see* **Genetic Map Functions**), IBD-sharing at a point d between the two loci can be obtained using the regression equations

$$\hat{\pi}_d = \rho_0 + \rho_1\hat{\pi}_1 + \rho_2\hat{\pi}_2$$

and

$$\hat{f}_{1d} = \omega_0 + \omega_1(\hat{\pi}_1 + \hat{\pi}_2 - 2\widehat{\pi_1\pi_2}) + \omega_2\hat{f}_{1.} + \omega_3\hat{f}_{.1} + \omega_4\hat{f}_{11},$$

where expressions for the regression parameters in terms of the recombination fractions between the two loci (θ_{m}), between the first marker and $d(\theta_1)$, and between the second marker and $d(\theta_2)$ are given in Table 2 [34].

Table 2 Regression parameters in the expressions for $\hat{\pi}_{\mathrm{d}}$ and $\hat{f}^{\mathrm{a}}_{1\mathrm{d}}$

ρ_0	$(1-\psi_1)(1-\psi_2)/\psi_{\mathrm{m}}$
ρ_1	$-\psi_2(1-\psi_2)(1-2\psi_1)/[\psi_{\mathrm{m}}(1-\psi_{\mathrm{m}})]$
ρ_2	$-\psi_1(1-\psi_1)(1-2\psi_2)/[\psi_{\mathrm{m}}(1-\psi_{\mathrm{m}})]$
ω_0	$2\psi_1(1-\psi_1)\psi_2(1-\psi_2)/\psi^2_{\mathrm{m}}$
ω_1	$2(1-2\psi_1)(1-2\psi_2)\psi_1(1-\psi_1)\psi_2(1-\psi_2)/[\psi^2_{\mathrm{m}}(1-\psi^2_{\mathrm{m}})]$
ω_2	$(1-2\psi_1)^2\psi^2_2(1-\psi_2)^2/[\psi^2_{\mathrm{m}}(1-\psi^2_{\mathrm{m}})]$
ω_3	$(1-2\psi_2)^2\psi^2_1(1-\psi_1)^2/[\psi^2_{\mathrm{m}}(1-\psi^2_{\mathrm{m}})]$
ω_4	$(1-2\psi_1)^2(1-2\psi_2)^2\psi_1(1-\psi_1)\psi_2(1-\psi_2)/[\psi^2_{\mathrm{m}}(1-\psi^2_{\mathrm{m}})(1-2\psi_{\mathrm{m}})^2(1-2\psi_{\mathrm{m}}+2\psi^2_{\mathrm{m}})]$

[a]From Olson [34], reproduced by permission of the publisher; $\psi_j = \theta_j^2 + (1-\theta_j)^2$, $j = 1, 2, \ldots, m$.

Linkage Between Marker and Qualitative Trait

Model-free linkage methods designed for qualitative traits often consider samples of affected sib-pairs or sibships with at least two affected members. If a trait and marker are linked, affected sib-pairs should share more alleles IBD than expected by chance. Under the **null hypothesis** of no linkage, sib-pairs are expected to share exactly 0, 1, or 2 alleles IBD at a single marker locus with respective probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. Early test statistics generally assumed that the marker IBD state can be determined with certainty, so that a sample of n pairs can be partitioned into n_0, n_1, and n_2 pairs corresponding to sharing 0, 1, or 2 alleles IBD. Day & Simons [14] and Suarez et al. [48], assuming such a fully informative marker locus, proposed a one-sided (*see* **Alternative Hypothesis**) **nonparametric** test statistic (T_2) that compares the observed proportion of sib-pairs that share exactly two marker alleles IBD to its null value of $\frac{1}{4}$:

$$T_2 = \frac{n_2/n - 1/4}{[3/(16n)]^{1/2}}.$$

Green & Woodrow [22] proposed a one-sided nonparametric test statistic (T_{m}) that compares the observed mean proportion of marker alleles shared IBD to its null value of $\frac{1}{2}$:

$$T_{\mathrm{m}} = \frac{n_2 + n_1/2 - 1/2}{[1/(8n)]^{1/2}}.$$

Extensions for use with larger sibships were proposed by Green & Woodrow [22] and deVries et al. [15]. These more general test statistics can substitute $\bar{\hat{\pi}} = \sum_{j=1}^{n}\hat{\pi}_j/n$ for $n_2 + n_1/2$ and an empirical variance estimate for the denominator when IBD sharing cannot be determined with certainty.

A **goodness-of-fit** statistic that compares the observed IBD distribution to that expected by chance was proposed by Weitkamp et al. [50]. Blackwelder

& Elston [5] compared the **power** of this test statistic, T_2, and T_m and found that T_m has greater power for most one-locus genetic models; T_2 has more power for some recessive models. Schaid & Nick [44] studied the asymptotically most powerful linear combination of the $\hat{f}_i$ and determined that T_m has power close to optimal for a broad range of single-locus models. Knapp et al. [28] determined that, provided $\delta_1^2 = \delta_0\delta_2$, where $\delta_s = \Pr(\text{affected}|\text{trait genotype with } s \text{ susceptibility alleles})$, T_m is uniformly most powerful in θ, the recombination fraction between trait and marker loci.

Suarez et al. [48] characterized the distribution of sib-pair IBD sharing in terms of the population trait prevalence K, additive genetic variance σ_a^2, dominance genetic variance σ_d^2, and θ, the recombination fraction between trait and marker loci. Suarez et al. [44] determined the boundaries of this parameter space under a one-locus model. Risch [38–41] developed a parametric strategy for detecting linkage to complex diseases using affected sib-pairs and extended the methodology to multilocus trait models and to other types of relative pairs. Let z_{ri} be parameters defined as the probability that an affected relative pair of type r shares i marker alleles IBD. Then the lod score [the log base 10 of the ratio of the likelihoods under the alternative and the null hypothesis of no linkage (*see* **Likelihood Ratio**)] for the pedigree marker data I_m given **ascertainment** of an affected relative pair of type r (arp_r) can be written

$$Z(I_m|\text{arp}_r) = \log_{10} \sum_{i=0,1,2} \frac{\hat{f}_i z_{ri}}{f_{ri}}. \qquad (1)$$

Under the null hypothesis $\theta = \frac{1}{2}$, $z_{ri} = f_{ri}$, for $i = 0, 1, 2$, and $Z(I_m|\text{arp}_r) = 0$. The lod score (1) is maximized over the z_{ri} at regular (e.g. 1 cM) intervals in a chromosomal region containing the typed markers. Alternatively, Hauser et al. [25] maximized the lod score over both the z_{ri} and the recombination fraction between one of two flanking markers and a disease locus assumed to lie between them.

Now define K_r to be the probability that a relative of type r of an affected individual is also affected, and let $\lambda_r = K_r/K$ be the **relative risk** of disease to a relative of type r. Let the subscripts s, o, and m denote sibling, parent/offspring, and monozygotic twins (*see* **Heterozygosity**), respectively. Under the assumption that a single locus confers susceptibility to disease, the z_{ri} are related to the relative risks and the recombination fraction as shown in Table 3 [39]. For affected sib-pairs, when $\theta = 0$, $z_{s0} = 1/(4\lambda_s)$, $z_{s1} = \lambda_0/(2\lambda_s)$, and $z_{s2} = \lambda_m/(4\lambda_s)$. Constraints on the z_{si} consistent with a one-locus genetic model are: $z_{s0} \geq 0$, $z_{s2} + z_{s0} \geq z_{s1}$, and $z_{s1} \geq 2z_{s0}$. Holmans [26] determined the asymptotic distribution of the maximum lod score under these constraints to be a mixture of χ_1^2 and χ_2^2 random variables; a lod score of 2.3 corresponds to a significance level of 10^{-3}. For various types of relative pairs, Lander & Kruglyak [33] proposed that pointwise significance levels (*see* **Hypothesis Testing**) and lod scores corresponding to a genome-wide significance of 0.05 be considered "significant" evidence in favor of linkage, assuming a single disease locus and an infinitely dense marker map. Davis et al. [13] developed **simulation**-based nonparametric statistics that condition on the marker genotypes of the unaffected family members.

Table 3 Parameters of z_{ri} for relative pairs[a]

Type of relative pair	z_{r0}	z_{r1}	z_{r2}
Full sibling	$\frac{1}{4} - \frac{1}{4\lambda_s}(2\psi - 1)[(\lambda_s - 1) + 2(1 - \psi)(\lambda_s - \lambda_o)]$	$\frac{1}{2} - \frac{1}{2\lambda_s}(2\psi - 1)^2(\lambda_s - \lambda_o)$	$\frac{1}{4} + \frac{1}{4\lambda_s}(2\psi - 1)[(\lambda_s - 1) + 2\psi(\lambda_s - \lambda_o)]$
Grandparental	$\frac{1}{2} - \frac{1}{2\lambda_g}(1 - 2\theta)(\lambda_g - 1)$	$\frac{1}{2} + \frac{1}{2\lambda_g}(1 - 2\theta)(\lambda_g - 1)$	–
Avuncular	$\frac{1}{2} - \frac{1}{2\lambda_a}(1 - \theta)(1 - 2\theta)^2(\lambda_a - 1)$	$\frac{1}{2} + \frac{1}{2\lambda_a}(1 - \theta)(1 - 2\theta)^2(\lambda_a - 1)$	–
Half-sibling	$\frac{1}{2} - \frac{1}{2\lambda_h}(2\psi - 1)(\lambda_h - 1)$	$\frac{1}{2} + \frac{1}{2\lambda_h}(2\psi - 1)(\lambda_h - 1)$	–
First cousin	$\frac{3}{4} - \frac{1}{2\lambda_c}[(1 - \theta)^4 + \theta^2(1 - \theta)^2 + \frac{\theta^2}{2} - \frac{1}{4}](\lambda_c - 1)$	$\frac{1}{4} + \frac{1}{2\lambda_c}[(1 - \theta)^4 + \theta^2(1 - \theta)^2 + \frac{\theta^2}{2} - \frac{1}{4}](\lambda_c - 1)$	–

[a]From Risch [39], reproduced by permission of the publisher; $\psi = \theta^2 + (1 - \theta)^2$; s = full sibling, g = grandparental, a = avuncular, h = half-sibling, c = first cousin, o = parent/offspring.

For samples of affected sib-pairs, a test statistic with more power than the lod score for some recessive models is

$$\frac{T_{z_2} = \left(\hat{z}_{s2} - \frac{1}{4}\right)}{\text{var}(\hat{z}_{s2})},$$

where $\text{var}(\hat{z}_2)$ is obtained from the observed **information matrix** [35].

Affected sib-pair methods are particularly useful in detecting linkage to complex diseases, which are expected to be oligogenic. The single-locus lod score provides a valid test of linkage even if other loci contribute to a disease. Multilocus models are also of interest. A general two-locus lod score with eight free parameters may be written as

$$Z(I_m|\text{arp}_s) = \log_{10} \sum_{i=0,1,2} \sum_{j=0,1,2} \frac{z_{ij} \hat{f}_{i1} \hat{f}_{j2}}{f_{i1} f_{j2}},$$

where the subscripts 1 and 2 refer to the two loci, and z_{ij} are parameters representing the probability that an affected sib-pair shares i alleles IBD and the first locus and j at the second locus. The two disease loci are assumed to be unlinked. If the two loci interact in a multiplicative fashion, then $z_{ij} = z_i z_j$ by definition, and four free parameters are required. Let λ_{ki} be the relative risk to a relative that shares i alleles IBD with an affected individual at disease locus k, and let K_k and λ_{ks} be the contributions to the overall prevalence and sibling relative risk due to locus k. If the two loci interact in an additive fashion, then

$$\frac{z_{ij}}{f_{1i} f_{2j}} - 1 = \frac{1}{\lambda_s} \left(\frac{K_1}{K}\right)^2 (\lambda_{1i} - \lambda_{1s}) + \frac{1}{\lambda_s} \left(\frac{K_2}{K}\right)^2 (\lambda_{2j} - \lambda_{2s}),$$

an additive property [39]; this formulation also requires four free parameters.

The lod score (1) may also be written as

$$Z(I_m|\text{arp}_s) = \log_{10} \left[1 + \beta \left(\hat{\pi} - \tfrac{1}{2}\right) + \gamma \left(\hat{f}_1 - \tfrac{1}{2}\right)\right],$$

where $\beta = (\sigma_a^2 + \sigma_d^2)/(K K_s) = 4(z_2 - z_0)$ and $\gamma = -\sigma_d^2/(2K K_s) = 2(z_1 - z_0 - z_2)$. The general two-locus model may similarly be parameterized in terms of **variance components** $\sigma^2_{a_j}, \sigma^2_{d_j}, \sigma^2_{a_1 a_2}, \sigma^2_{d_1 d_2}$, and $\sigma^2_{a_j d_{3-j}}$ – the contribution to the total genetic variance due to the **interaction** between the additive (a) or dominance (d) component of the jth locus, $j = 1, 2$ [11, 16, 23, 35]. One such model [35] is

$$\begin{aligned} Z(I_m|\text{arp}_s) = \log_{10} \Big\{ 1 + (K K_s)^{-1} \Big[& B_1 \left(\hat{\pi}_1 - \tfrac{1}{2}\right) \\ & + C_1 \left(\hat{f}_{11} - \tfrac{1}{2}\right) + B_2 \left(\hat{\pi}_2 - \tfrac{1}{2}\right) \\ & + C_2 \left(\hat{f}_{12} - \tfrac{1}{2}\right) + D \left(\hat{\pi}_1 \hat{\pi}_2 - \tfrac{1}{4}\right) \\ & + F_1 \left(\hat{\pi}_1 \hat{f}_{12} - \tfrac{1}{4}\right) + F_2 \left(\hat{\pi}_2 \hat{f}_{11} - \tfrac{1}{4}\right) \\ & + G \left(\hat{f}_{11} \hat{f}_{12} - \tfrac{1}{4}\right) \Big] \Big\}, \end{aligned}$$

where

$$\begin{aligned} B_j &= \sigma^2_{a_j} + \sigma^2_{d_j}, \quad j = 1, 2, \\ C_j &= \frac{-\sigma^2_{d_j}}{2}, \quad j = 1, 2, \\ D &= \sigma^2_{a_1 a_2} + \sigma^2_{a_1 d_2} + \sigma^2_{a_2 d_1} + \sigma^2_{d_1 d_2}, \\ F_j &= \frac{-\sigma^2_{a_j d_{3-j}} + \sigma^2_{d_1 d_2}}{2}, \quad j = 1, 2, \end{aligned}$$

and

$$G = \frac{\sigma^2_{d_1 d_2}}{4}.$$

When K and K_s are known, the variance components may be estimated directly; otherwise, the model can be fitted by reparameterizing so that $B_1^* = B_1/(K K_s)$, and similarly for the remaining parameters. **Additive** and **multiplicative models** can also be fitted using variance components parameterizations [35].

For linkage analysis using small pedigrees, Kruglyak et al. [31] proposed calculating a scoring function $S(v, \Psi)$ (*see* **Scores**) that depends on an inheritance vector v and the observed disease phenotypes Ψ in the pedigree. When the inheritance vector is unknown, one computes its conditional expectation

$$\overline{S}(\Psi) = \sum_v S(v, \Psi) P(v),$$

where $P(v)$ is estimated using available marker data. The authors further discuss a model-free scoring function that considers IBD-sharing among sets of affected family members; this scoring function was first proposed by Whittemore & Halpern [51]. Let a denote the number of affected individuals in a

pedigree, let h be a collection of alleles obtained by choosing one allele from each of these individuals, and let $b_i(h)$ denote the number of times that the ith founder allele appears in h. The scoring function is defined as

$$S_{\text{all}}(v) = 2^{-a} \sum_h \left[\prod_i b_i(h)! \right],$$

where the sum is over the 2^a possible ways to choose h. Kruglyak et al. then standardize the score to obtain $Z(v) = [S(v) - \mu]/\sigma$, where μ and σ are the mean and standard deviation of S under the **uniform distribution** of inheritance vectors. A global score is obtaining by taking a weighted average of standardized scores; weights depend on pedigree size.

The affected-pedigree-member (APM) method of Weeks & Lange [49] can also be used to analyze extended pedigrees, and uses identity-by-state sharing to incorporate information from multiple markers; this method is less powerful than methods based on IBD sharing [21]. Curtis & Sham [12] proposed comparing observed and expected numbers of alleles shared IBD between all affected relative pairs in a pedigree. Guo [23] proposed plotting $\hat{\pi}$ over each chromosomal interval and examining further regions for which $\hat{\pi}$ is substantially larger than $\frac{1}{2}$.

Elston [17] and Elston et al. [18] developed and studied two-stage global search designs for linkage analysis to complex diseases using pairs of affected relatives. In the first of the two stages, a genome scan is performed on n affected pairs using m equally spaced marker loci. For each marker with a pointwise ***P* value** less than α^*, k additional markers in the region are typed and a more stringent significance level α applied. Given the relative risk λ_r for a particular trait locus, the desired power of the study, and the ratio of the cost of recruiting one person into the study to the cost of performing one marker assay, n, m, k, and α may be chosen to minimize the total cost of the study. Typically, an optimal two-stage procedure halves the cost of a study, compared to a procedure involving only the first stage and the criterion α.

Linkage analysis may also be done using discordant pairs, i.e. pairs in which one member is affected and the other unaffected. Such pairs provide good power for linkage if the disease is rare and dominant, or if the disease is common. The lod score takes the same form as (1), except that the z_{ri} are now the probabilities that a discordant pair of type r share i alleles IBD. For discordant sib-pairs, genetic constraints are obtained by reversing the roles of z_{s0} and z_{s2} in the inequalities given above for affected sib-pairs.

Linkage Between Marker and Quantitative Trait

The problem of detecting linkage between a marker locus and a locus underlying a quantitative trait using sib-pair data was first considered by Penrose [37], who proposed comparing the covariance of the sib-pair trait and marker differences with that expected when the trait and marker are not linked. Haseman & Elston [24] and Blackwelder [4] expanded this idea and included available marker information from the parents.

Assume that a single locus with alleles T and t underlies a quantitative trait. For an observation X from the trait distribution, the genetic model may be written

$$X = \mu + g + e,$$

where μ is an overall mean, g is a major gene effect such that $g = a, d$, or $-a$ for trait genotypes TT, Tt, or tt, respectively, and e is a residual effect with an unspecified distribution. Putting $p = \Pr(\text{T})$ and $q = 1 - p$,

$$\sigma_{\text{a}}^2 = 2pq[a - d(p - q)]^2,$$
$$\sigma_{\text{d}}^2 = 4p^2q^2d^2,$$

and $\sigma_{\text{g}}^2 = \sigma_{\text{a}}^2 + \sigma_{\text{d}}^2$. Also, let $\sigma_\varepsilon^2 = \text{E}(e_1 - e_2)^2$, for a pair of sibs indexed 1 and 2, and let $\psi = \theta^2 + (1 - \theta)^2$, where θ is the recombination fraction between trait and marker loci.

The squared difference $Y = (X_1 - X_2)^2$ between the measurements of a quantitative trait for a randomly sampled pair of siblings is a linear function of the **Bayesian** estimate of the proportion of marker alleles shared IBD between the members of the pair ($\hat{\pi}$) and the estimated probability that the pair share exactly one marker allele IBD ($\hat{f}_1$), i.e.

$$\text{E}(Y|I_{\text{m}}) = \alpha_{\text{s}} + \beta_{\text{s}}\hat{\pi} + \gamma_{\text{s}}\hat{f}_1,$$

where

$$\alpha_{\text{s}} = \sigma_\varepsilon^2 + 2\sigma_{\text{g}}^2\psi + 2\sigma_{\text{d}}^2\psi(1 - \psi),$$
$$\beta_{\text{s}} = 2\sigma_{\text{g}}^2(1 - 2\psi),$$

and

$$\gamma_s = \sigma_d^2(1-2\psi)^2.$$

If $\theta = \frac{1}{2}$ or $\sigma_g^2 = 0$, then $\beta = 0$; otherwise, $\beta < 0$. After fitting the regression model using **least squares**, an asymptotically normal one-sided test of linkage may be constructed.

Similar regression relationships have been developed for other types of relative pairs, specifically half-sib, grandparental, avuncular, and cousin pairs [2]; all take the form

$$E(Y|I_m) = \alpha_r + \beta_r\hat{\pi},$$

where α_r and β_r are functions of $\sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2$, and θ that are specific to relative pair type r (Table 4). Olson & Wijsman [36] used **generalized estimating equations** to combine information from different types of relative pairs in a set of pedigree data. Assume that p types of relative pairs are of interest and that the data set consists of N pedigrees, each with n_i relative pairs. Under a working independence model, α_r and β_r are estimated separately for each type of relative pair; the robust **covariance matrix** of these estimates is given by

$$\text{var}(\alpha, \beta) = \left(\sum_{i=1}^{N} \mathbf{D}_i'\mathbf{D}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{D}_i'\hat{\mathbf{S}}_i\hat{\mathbf{S}}_i'\mathbf{D}_i\right) \times \left(\sum_{i=1}^{N} \mathbf{D}_i'\mathbf{D}_i\right)^{-1},$$

where $\mathbf{D}_i$ is an $n_i \times 2p$ matrix containing the $\hat{\pi}_r$ and $\hat{f}_1$ (analogous to the design matrix in a linear regression model), and $\mathbf{S}_i$ is a vector of length n_i with elements $y - \alpha_r - \beta_r\hat{\pi}_r$. A one-sided, asymptotically normal, test of linkage takes the form

$$\frac{T = N^{1/2}\mathbf{c}^T\hat{\boldsymbol{\beta}}}{[\mathbf{c}^T\text{var}(\hat{\boldsymbol{\beta}})\mathbf{c}]^{1/2}},$$

where $\mathbf{c}$ is a p-dimensional vector of weights, chosen a priori, with elements $\text{var}(\pi_r)n_r$, n_r is the total number of pairs of type r, and $\hat{\boldsymbol{\beta}}$ is a vector of the regression estimates $\hat{\beta}_r$. In a multipoint setting, or when a candidate locus is being tested, the model with common slope

$$E(Y|I_m) = \alpha_r + \boldsymbol{\beta}\hat{\pi}$$

may be fitted; $\mathbf{D}_i$ becomes an $n_i \times p + 1$ matrix and no a priori weights are required to test linkage. Because $\beta_s = \beta_h$ for all values of the recombination fraction, sib-pairs and half-sib-pairs may be combined in a similar manner for a genome scan using single markers. Schaid et al. [45] combine these two types of pairs into a single test of linkage, using an empirically derived adjustment to the **degrees of freedom** of the t-statistic (*see* **Student's *t* Distribution**) to allow for correlated pairs.

Regression models have been developed for sampling schemes other than random sampling. Assume that probands are sampled from the upper tail of the trait distribution. Let $X_1 > c$ be the proband trait value, and X_2 the trait value for the proband's sibling. **Consistent estimates** of regression coefficients may be obtained by fitting

$$X_2 = A + B_1X_1 + B_2\left(\hat{\pi} - \tfrac{1}{2}\right) + B_3\left(\left|\hat{\pi} - \tfrac{1}{2}\right| - \tfrac{1}{4}\right),$$

[7]. $B_2 = 0$ if $\theta = \frac{1}{2}$; otherwise, $B_2 > 0$. ($B_2 < 0$ if probands are sampled from the lower tail of the

Table 4 Coefficients of the regression of squared pair differences on the proportion of alleles shared IBD for relative pairs[a]

Type of relative pair	α_r	β_r
Half-sibling	$\sigma_\varepsilon^2 + 2\sigma_g^2 - 2\theta(1-\theta)\sigma_a^2$	$-2(1-2\theta)^2\sigma_a^2$
Grandparental	$\sigma_\varepsilon^2 + 2\sigma_g^2 - \theta\sigma_a^2$	$-2(1-2\theta)\sigma_a^2$
Avuncular	$\sigma_\varepsilon^2 + 2\sigma_g^2 - (\frac{5}{2}\theta - 4\theta^2 + 2\theta^3)\sigma_a^2$	$-2(1-2\theta)^2(1-\theta)\sigma_a^2$
First cousin	$\sigma_\varepsilon^2 + 2\sigma_g^2 - (\frac{4}{3}\theta - \frac{5}{2}\theta^2 + 2\theta^3 - \frac{2}{3}\theta^4)\sigma_a^2$	$-2(1-2\theta)^2(1-\frac{4}{3}\theta + 2\theta^2)\sigma_a^2$

[a]From Amos & Elston [2], reproduced by permission of the publisher.

trait distribution.) Conditioning on the ascertainment process, rather than the proband's trait value, gives

$$\mathrm{E}(X_2|I_\mathrm{m}, X_1 > c) = A^* + B_2^*\left(\hat{\pi} - \tfrac{1}{2}\right) + B_3^*\left(\left|\hat{\pi} - \tfrac{1}{2}\right| - \tfrac{1}{4}\right).$$

Use of this sampling scheme greatly increases the power to detect linkage [6, 7], particularly for a rare allele with a large effect.

This "selected sampling" scheme provides excellent power provided that one samples probands from the tail of the distribution with the rarer allele. A second design, which is uniformly powerful in all genetic situations, is sampling of extreme discordant sib-pairs – pairs such that one sib has a trait value from the upper tail and the other from the lower tail. For example, given a large sample of probands with an extreme value in one direction (usually that indicating disease), one might genotype only those pairs for which the sibling has a trait value in the opposite tail. Such extreme discordant pairs provide good power for detecting linkage for additive, dominant, and recessive models [42, 43]. As there may be little trait variation within each tail, it is useful to ignore this variation and model

$$z_i \equiv \Pr(\pi = i/2|\mathrm{edsp}), \quad i = 0, 1, 2,$$

where edsp denotes "extreme discordant sib-pair". The lod score is the same as in the case of sampling sib-pairs discordant for a dichotomous trait, which may be considered a special case of edsp for which the two tails share the same cutpoint, with different constraints on the z_i.

Other approaches to quantitative trait linkage have been proposed. To assess evidence for genetic linkage from pedigrees, Amos [3] modeled the covariance matrix of pedigree trait values in terms of variance components, IBD sharing and the recombination fraction. Let a general model for trait values X be

$$X_i = \mu + g_i + G_i + \boldsymbol{\beta}^\mathrm{T}\mathbf{w}_i + e_i,$$

where μ, g_i, and e_i are the overall mean, major genotype effect, and environmental error, as before, G_i is a random effect of polygenes (*see* **Polygenic Inheritance**), $\mathbf{w}_i$ a vector of fixed covariates, and $\boldsymbol{\beta}$ a set of regression parameters. Without loss of generality, take $\mathrm{E}(g_i) = \mathrm{E}(e_i) = \mathrm{E}(G_i) = 0$. Then

$$\mathrm{E}(X_i) = \mu + \boldsymbol{\beta}^\mathrm{T}\mathbf{w}_i,$$
$$\mathrm{var}(X_i) = \sigma_\mathrm{a}^2 + \sigma_\mathrm{d}^2 + \sigma_\mathrm{G}^2 + \sigma_\mathrm{e}^2,$$

and

$$\mathrm{cov}(X_i, X_j|\pi_{ij}) = f(\theta, \pi_{ij})\sigma_\mathrm{a}^2 + g(\theta, f_{2ij})\sigma_\mathrm{d}^2 + \phi_{ij}\sigma_\mathrm{G}^2 \quad \text{for } i \neq j,$$

where ϕ_{ij} is the coefficient of relationship between family members i and j (*see* **Inbreeding**), $f(\theta, \pi_{ij})$ is given for various relative pairs in Table 5, and $g(\theta, f_{2ij})$ equals 0 for all but sib-pairs, in which case

$$\begin{aligned}\mathrm{cov}(X_i, X_j|\pi_{ij}) = {} & 2\theta(1-\theta)\sigma_\mathrm{g}^2 \\ & + 2(\theta-1)\theta(1-2\theta+2\theta^2)\sigma_\mathrm{d}^2 \\ & + [(1-2\theta)^2\sigma_\mathrm{g}^2 - (1-2\theta)^4\sigma_\mathrm{d}^2]\pi_{ij} \\ & + (1-2\theta)^4\sigma_\mathrm{d}^2 f_{2ij}.\end{aligned}$$

Parameters may be estimated using **maximum likelihood** methods, if **multivariate normality** of errors is assumed, or by estimating-equation approaches.

Another approach, the weighted pairwise correlation (WPC) statistic, was proposed by Commenges [8]. This score test may be applied to sets of large pedigrees with several types of relative pairs. Consider a set of F pedigrees, each with n_f members. For an individual pedigree, a general form for the score statistic is

$$S_\mathrm{L} = \sum_{i<j} W_{ij}U_iU_j,$$

where $W_{ij} = \hat{\pi_{ij}} - \overline{\hat{\pi}}_{rij}$ are centered IBD-sharing estimates for the pair of relatives i and j, and U_i is the residual $x_i - \mathrm{E}(X_i|\text{covariates})$, based on some parametric model for the mean of X. If more **robustness** is desired, the U_i may be replaced by their centered **ranks**, to give

$$S_\mathrm{R} = \sum_{i<j} W_{ij}(R_i - \overline{R})(R_j - \overline{R}),$$

Table 5 Expressions for $f(\theta, \pi_{ij})$[a]

Relative pair	$f(\theta, \pi_{ij})$
Sibling	$\frac{1}{2} + (1-2\theta)^2(\pi_{ij} - \frac{1}{2})$
Half-sibling	$\frac{1}{4} + (1-2\theta)^2(\pi_{ij} - \frac{1}{4})$
Avuncular	$\frac{1}{4} + (1-2\theta)^2(1-\theta)(\pi_{ij} - \frac{1}{4})$
Grandparental	$\frac{1}{4} + (1-2\theta)(\pi_{ij} - \frac{1}{4})$
First cousin	$\frac{1}{8} + (1-2\theta)^2(1 - \frac{4}{3}\theta + \frac{2}{3}\theta^2)(\pi_{ij} - \frac{1}{8})$

[a]From Amos [1], reproduced by permission of the publisher.

where R_i is the rank of the ith residual, and $\overline{R}$ is the mean of the ranks. For a set of F pedigrees, linkage may be tested using

$$S = \frac{\sum_{f=1}^{F}[S_{\mathrm{Rf}} - \mathrm{E}(S_{\mathrm{Rf}})]}{\left(\sum_{f=1}^{F}\mathrm{var}S_{\mathrm{Rf}}\right)^{1/2}},$$

where

$$\mathrm{E}(S_{\mathrm{Rf}}) = -\frac{n_f+1}{12}\sum_{i<j} W_{ijf},$$

$$\mathrm{var}(S_{\mathrm{Rf}}) = A\sum_{i<j} W_{ijf}^2 + B\sum_{i<j}\sum_{r<s;r,s\neq i,j} \times W_{ijf}W_{rsf} - 2C\sum_{i,j\neq i}\sum_{r\neq i,j} W_{irf}W_{jrf},$$

$A=(n_f+1)(n_f-2)(5n_f^2+n_f-8)/720$, $B=(n_f+1)(10n_f+16)/720$, and $C=(n_f+1)(5n_f^2-3n_f-16)/720$. This test is called the weighted pairwise rank correlation (WPRC) test. Simulations show that the WPRC can be more powerful than the Haseman–Elston test for single large pedigrees, or in the presence of genotype-by-environment interaction (*see* **Gene-environment Interaction**) or family-specific residual variance [10]. For larger samples of small pedigrees, the Haseman–Elston test is generally the more powerful, particularly for highly heritable dominant traits. The WPC or WPRC can be applied to dichotomous traits as well as to quantitative traits and can substitute identity-by-state sharing for IBD sharing estimates. Commenges & Abel [9] proposed a **transformation** that yields uncorrelated **residuals** and a more robust test statistic.

Goldgar [19] proposed a multipoint IBD method that assumes that the quantitative trait is due to additive genetic effects and a normally distributed random environmental component. The method is parameterized using the proportion of the total trait variance due to additive genetic effects (h^2) and the proportion (P) of the genetic variance due to loci in the chromosomal interval defined by the marker loci. For each sib-pair, the proportion of the chromosomal region shared IBD, conditional on the marker data, is estimated. A covariance matrix of the sibship trait values is then constructed as a function of the IBD-sharing estimates, h^2, and P. The likelihood for the trait values, conditional on IBD-sharing, is assumed to be multivariate normal; numerical maximum-likelihood techniques are used to estimate P and to test the null hypothesis $P = 0$. Limited simulation suggests that this multipoint method is more powerful than the single-marker Haseman–Elston method. The method also has power comparable to model-based linkage analysis when parental data are unknown, the effect of the major locus is small and there is additional genetic variation, or the parameters of the model-based analysis are misspecified [20].

Software

Estimates of multipoint IBD-sharing may be obtained using MAPMAKER/SIBS (nuclear families) and GENEHUNTER (small pedigrees). These programs also apply a variety of parametric and nonparametric tests of linkage for both quantitative and qualitative data. For single markers, the SAGE program SIBPAL performs nonparametric tests of linkage for affected sib-pairs, and applies the Haseman–Elston regression method for sib and half-sib pairs. The SAGE program RELPAL estimates single-marker allele-sharing probabilities for large pedigrees and applies the Olson–Wijsman test of linkage. The SAGE program DESPAIR provides optimal two-stage designs for genome searches using affected relative pairs. Other software, including APM, ASPEX, ERPA, ESPA, GAS, MFLINK, MIM, NOPAR, and SIMIBD, that perform various aspects of model-free linkage analysis, are available from their respective authors (*see* **Software for Genetic Epidemiology**).

References

[1] Amos, C.I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees, *American Journal of Human Genetics* **54**, 535–543.

[2] Amos, C.I. & Elston, R.C. (1989). Robust methods for the detection of genetic linkage for quantitative data from pedigrees, *Genetic Epidemiology* **6**, 349–360.

[3] Amos, C.I., Dawson, D.V. & Elston, R.C. (1990). The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees, *American Journal of Human Genetics* **47**, 842–853.

[4] Blackwelder, W.C. (1977). Statistical Methods for Detecting Genetic Linkage from Sibship Data, *Institute*

of Statistics Mimeo Series No. 1114. Department of Biostatistics, University of North Carolina, Chapel Hill.

[5] Blackwelder, W.C. & Elston, R.C. (1985). A comparison of sib-pair linkage tests from disease susceptibility loci, *Genetic Epidemiology* **2**, 85–97.

[6] Cardon, L.R. & Fulker, D.W. (1994). The power of interval mapping of quantitative trait loci using selected sib pairs, *American Journal of Human Genetics* **55**, 825–833.

[7] Carey, G. & Williamson, J.A. (1991). Linkage analysis of quantitative traits: increased power by using selected samples, *American Journal of Human Genetics* **49**, 786–796.

[8] Commenges, D. (1994). Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic, *Genetic Epidemiology* **11**, 189–200.

[9] Commenges, D. & Abel, L. (1996). Improving the robustness of the weighted pairwise correlation test for linkage analysis, *Genetic Epidemiology* **13**, 559–574.

[10] Commenges, D., Olson, J. & Wijsman, E. (1994). The weighted pairwise correlation statistic for linkage analysis: simulation study and application to Alzheimer's disease, *Genetic Epidemiology* **11**, 201–212.

[11] Cordell, H.J., Todd, J.A., Bennett, S.T., Kawaguchi, Y. & Farrall, M. (1995). Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 in type 1 diabetes, *American Journal of Human Genetics* **57**, 920–934.

[12] Curtis, D. & Sham, P.C. (1994). Using risk calculation to implement an extended relative pair analysis, *Annals of Human Genetics* **58**, 151–162.

[13] Davis, S., Schroeder, M., Goldin, L.R. & Weeks, D.E. (1996). Nonparametric simulation-based statistics for detecting linkage in general pedigrees, *American Journal of Human Genetics* **58**, 867–880.

[14] Day, N.E. & Simons, M.J. (1976). Disease susceptibility genes – their identification by multiple case family studies, *Tissue Antigens* **8**, 109–119.

[15] deVries, R.R.P., Fat, R.F.M., Lai, A., Nijenhuis, L.E. & van Rood, J.J. (1976). HLA-linked genetic control of host response to *Mycobacterium leprae*, *Lancet* **ii**, 1328–1330.

[16] Dupuis, J., Brown, P.O. & Siegmund, D. (1995). Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent, *Genetics* **140**, 843–856.

[17] Elston, R.C. (1992). Designs for the global search of the human genome by linkage analysis, in *Proceedings of the Sixteenth International Biometric Conference*, Hamilton, New Zealand, December 7–11, pp. 39–51.

[18] Elston, R.C., Guo, X. & Williams, L.V. (1996). Two-stage global search designs for linkage analysis using pairs of affected relatives, *Genetic Epidemiology* **13**, 535–558.

[19] Goldgar, D.E. (1990). Multipoint analysis of human quantitative genetic variation, *American Journal of Human Genetics* **47**, 957–967.

[20] Goldgar, D.E. & Oniki, R.S. (1992). Comparison of a multipoint identity-by-descent method with parametric multipoint linkage analysis for mapping quantitative traits, *American Journal of Human Genetics* **50**, 598–606.

[21] Goldin, L.R. & Weeks, D.E. (1993). Two-locus models of disease: comparison of likelihood and nonparametric linkage methods, *American Journal of Human Genetics* **53**, 908–915.

[22] Green, J.R. & Woodrow, J.C. (1977). Sibling method for detecting HLA-linked genes in a disease, *Tissue Antigens* **9**, 31–35.

[23] Guo, S.-W. (1995). Detection of genome similarity as an exploratory tool for mapping complex traits, *Genetic Epidemiology* **12**, 877–882.

[24] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[25] Hauser, E.R., Boehnke, M., Guo, S.-W. & Risch, N. (1996). Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations, *Genetic Epidemiology* **13**, 117–137.

[26] Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis, *American Journal of Human Genetics* **52**, 362–374.

[27] Idury, R.M. & Elston, R.C. (1997). A faster and more general hidden Markov model algorithm for multipoint likelihood calculations, *Human Heredity* **47**, 197–202.

[28] Knapp, M., Seuchter, S.A. & Baur, M.P. (1994). Linkage analysis in nuclear families. 1: Optimality criteria for affected sib-pair tests, *Human Heredity* **44**, 37–43.

[29] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative trait data, *American Journal of Human Genetics* **57**, 439–454.

[30] Kruglyak, L., Daly, M.J. & Lander, E.S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping, *American Journal of Human Genetics* **56**, 519–527.

[31] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.

[32] Lander, E.S. & Green, P. (1987). Construction of multilocus genetic maps in humans, *Proceedings of the National Academy of Sciences* **84**, 2363–2367.

[33] Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**, 241–247.

[34] Olson, J.M. (1995). Robust multipoint linkage analysis: an extension of the Haseman-Elston method, *Genetic Epidemiology* **12**, 177–193.

[35] Olson, J.M. (1997). Likelihood-based models for linkage analysis using affected sib pairs, *Human Heredity* **47**, 110–120.

[36] Olson, J.M. & Wijsman, E.M. (1993). Linkage between quantitative trait and marker loci: methods using all relative pairs, *Genetic Epidemiology* **10**, 87–102.

[37] Penrose, L.S. (1938). Genetic linkage in graded human characters, *Annals of Eugenics* **6**, 133–138.

[38] Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models, *American Journal of Human Genetics* **46**, 222–228.

[39] Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs, *American Journal of Human Genetics* **46**, 229–241.

[40] Risch, N. (1990). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs, *American Journal of Human Genetics* **46**, 242–253.

[41] Risch, N. (1992). Corrections to "Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs", *American Journal of Human Genetics* **51**, 673–675.

[42] Risch, N. & Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans, *Science* **268**, 1584–1589.

[43] Risch, N. & Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations, *American Journal of Human Genetics* **58**, 836–843.

[44] Schaid, D.J. & Nick, T.G. (1990). Sib-pair linkage tests for disease susceptibility loci: Common tests vs. the asymptotically most powerful test, *Genetic Epidemiology* **7**, 359–370.

[45] Schaid, D.J., Elston, R.C., Wilson, A.F. & Tran, L. (1994). In *SAGE Statistical Analysis for Genetic Epidemiology, Release 2.2*. Computer program available from the Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA.

[46] Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics, *American Journal of Human Genetics* **58**, 1323–1337.

[47] Suarez, B.K., Reich, T. & Trost, J. (1976). Limits of the general two-allele single locus model with incomplete penetrance, *Annals of Human Genetics* **40**, 231–243.

[48] Suarez, B.K., Rice, J. & Reich, T. (1978). The generalized sib paid IBD distribution: its use in the detection of linkage, *Annals of Human Genetics* **42**, 87–94.

[49] Weeks, D.E. & Lange, K. (1988). The affected-pedigree-member method of linkage analysis, *American Journal of Human Genetics* **42**, 315–326.

[50] Weitkamp, L.R., Stancer, H.C., Persad, E., Flood, C. & Guttormsen, S. (1981). Depressive disorders and HLA: a gene on chromosome 6 that can affect behavior, *New England Journal of Medicine* **305**, 1301–1306.

[51] Whittemore, A.S. & Halpern, J. (1994). A class of tests for linkage using affected pedigree members, *Biometrics* **50**, 118–127.

[52] Whittemore, A.S. & Halpern, J. (1994). Probability of gene identity by descent: computation and applications, *Biometrics* **50**, 109–117.

(*See also* **Identity Coefficients**)

JANE M. OLSON

LISREL

The acronym LISREL was coined by Jöreskog [5–7]: it is derived from LInear Structural RELations. Researchers use the term LISREL to refer either to **structural equations models** or to Jöreskog & Sörbom's [8] popular **software** program to estimate such statistical models. The LISREL *model* consists of two primary parts: a latent variable model and a measurement model. The former allows linear relationships between latent (unobserved) variables. This is much like a simultaneous equation model used in econometrics, except that it has latent rather than observed variables. It formulates the relation between the latent variables free of the confounding effects of measurement errors. The measurement model provides the linkages between the latent and observed variables. This model enables a researcher to use multiple indicators of the latent variables and to assess the "quality" of the measures. Many popular linear models (e.g. simultaneous equations, **confirmatory factor analysis**, **multiple regression, analysis of variance**, **analysis of covariance**, etc.) are special cases of Jöreskog's LISREL model.

In the original LISREL model, linear relations were assumed between continuous latent and continuous observed variables. Extensions of the LISREL model (see, for example, [8] and [11]) maintain the assumption of continuous latent variables but allow noncontinuous observed variables; for example, **censored**, ordinal (*see* **Ordered Categorical Data**), or dichotomous variables (*see* **Binary Data**). The relation between the latent variables and the noncontinuous observed variables is nonlinear. Other extensions allow equations that are nonlinear in the latent variables [3, 9, 10]. The article **Structural Equation Models** gives a more complete description of the LISREL model.

The second use of the term LISREL refers to a computer software program. One of the primary reasons for the rise in popularity of structural equation models was the availability of Jöreskog & Sörbom's [8] LISREL software package. For many years, LISREL was the only widely available program capable of estimating and testing these models. It is partly for this reason that both the structural equation model and the software were referred to by the same LISREL term. Since about the mid-1980s, other structural equation software programs have become more common (e.g. [1, 2, 4], and [12]). In addition, Jöreskog & Sörbom have continuously updated the LISREL program. The greater availability of software has contributed both to the further spread of these models as well as to the trend to refer to the statistical models as "structural equation models". The latter term helps to distinguish the model from the software needed to analyze the model.

References

[1] Arbuckle, J.L. (1997). *AMOS User's Guide, Version 3.6*. Small Waters Company, Chicago.

[2] Bentler, P.M. (1992). *EQS Structural Equations Program Manual*. BMDP Statistical Software, Los Angeles.

[3] Bollen, K.A. (1995). Structural equation models that are nonlinear in latent variables: a least-squares estimator, in *Sociological Methodology 1995*, P.M. Marsden, ed. American Sociological Association, Washington, pp. 223–251.

[4] Hartmann, W.M. (1990). *The CALIS Procedure: Extended User's Guide*. SAS Institute, Cary.

[5] Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system, in *Structural Equation Models in the Social Sciences*, A.S. Goldberger & O.D. Duncan, eds. Academic Press, New York, pp. 85–112.

[6] Jöreskog, K.G. (1977). Structural equation models in the social sciences: specification estimation, and testing, in *Applications of Statistics*, P.R. Krishnaiah, ed. North-Holland, Amsterdam, pp. 265–287.

[7] Jöreskog, K.G. & Sörbom, D. (1981). *LISREL V: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. National Educational Resources, Chicago.

[8] Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8.* Scientific Software, Mooresville.

[9] Jöreskog, K.G. & Yang, F. (1996). Nonlinear structural equation models: the Kenny–Judd model with interaction effects, in *Advanced Structural Equation Modeling*, G. Marcoulides & R. Schumacker, eds. Lawrence Erlbaum, Mahwah, pp. 57–88.

[10] Kenny, D.A. & Judd, C.M. (1984). Estimating the nonlinear and interactive effects of latent variables, *Psychological Bulletin* **96**, 201–210.

[11] Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika* **49**, 115–132.

[12] Muthén, B. (1988). *LISCOMP: Analysis of Linear Structural Equations with a Comprehensive Measurement Model*, 2nd Ed. Scientific Software, Mooresville.

KENNETH A. BOLLEN

Locally Most Powerful Tests

The classical **Neyman–Pearson Lemma** gives a **most powerful (MP) test** for the problem of testing a simple null hypothesis $\theta = \theta_0$ against a simple alternative hypothesis $\theta = \theta_1$. The Neyman–Pearson tests turn out to be uniformly most powerful (UMP) in some situations, but this is not true in general. For example, when $H_0: \theta \in \Theta_0 \subset \Theta$ and $H_1: \theta \in \Theta_1 \subset \Theta$ are one-sided, then the existence of a UMP test for every level α is essentially equivalent to the requirement that the joint density function has a monotone **likelihood ratio** property [11].

When a UMP test does not exist, one may restrict the class of tests to, say, the class of **unbiased** and/or invariant tests, and then look for a UMP test in this smaller class. Alternatively, one may look for tests that have maximum power against alternatives in a subset of Θ_1. The case when the subset of alternatives is "close" to the null parameter values has received a good deal of attention, presumably because tests that have good power for "local alternatives", which are the hardest to detect, may also retain good power for "nonlocal" alternatives.

Locally Most Powerful Tests

We focus attention to the case when θ is a real parameter, and use the Neyman & Pearson [8, 9] framework. Consider the problem of testing $H_0: \theta \leq \theta_0$ against $H_1: \theta > \theta_0$. Let ϕ_0 be a test function with **power** function $\beta_{\phi_0}(\theta) = E_\theta \phi_0(X)$. Then ϕ_0 is a locally most powerful (LMP) test of size α if there exists a $\Delta > 0$ such that for any other test ϕ with $\alpha = \sup_{\theta \leq \theta_0} \beta_{\phi_0}(\theta) \geq \sup_{\theta \leq \theta_0} \beta_\phi(\theta)$, $\beta_{\phi_0}(\theta_0) \geq \beta_\phi(\theta)$ for every $\theta \in (\theta_0, \theta_0 + \Delta]$. Thus, an LMP test maximizes

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\beta(\theta)\bigg|_{\theta=\theta_0} = \beta'(\theta)\bigg|_{\theta=\theta_0}$$

subject to the size constraint. Under some smoothness conditions one can show that any test of the form $\phi_0(\mathbf{x}) = 1$ if $\partial \log f(\mathbf{x};\theta)/\partial\theta|_{\theta=\theta_0} > k, = 0$ if $\partial \log f(\mathbf{x};\theta)/\partial\theta|_{\theta=\theta_0} < k$ will maximize $\beta'(\theta)|_{\theta=\theta_0}$. Here $f(\mathbf{x};\theta)$ is the joint probability density function (pdf) of a **random sample** $X_1, X_2, \ldots, X_n$ with common pdf $f(\mathbf{x};\theta)$.

Consider for example, the problem of testing $H_0: \theta \leq 0$ against $H_1: \theta > 0$, where θ is the **median** of a **Cauchy** density function $f(x;\theta) = \pi^{-1}[1 + (x-\theta)^2]^{-1}$, $-\infty < x < \infty$. Let $x_1, x_2, \ldots, x_n$ be n observations. It is easy to see that MP size α tests of $\theta = 0$ against $\theta = \theta_1, \theta_1 > 0$, depend on θ_1 and hence a UMP test for testing H_0 against H_1 does not exist. An LMP test of H_0 against H_1 is of form

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} 2x_i/(1+x_i^2) > k, \\ 0, & \text{elsewhere,} \end{cases}$$

where one chooses k so that the size of ϕ_0 is α (*see* **Critical Region**).

This LMP test, although good at detecting small departures from $H_0: \theta \leq 0$, is quite unsatisfactory in detecting values of θ much larger than 0. In fact, $\beta_{\phi_0}(\theta) \to 0$ as $\theta \to \infty$ if $\alpha < 1/2$.

Locally Most Powerful Unbiased Tests

The definition of an LMP test can be extended to the case of two-sided alternatives. In general, there do not exist LMP tests for two-sided alternatives. The LMP test ϕ_0 above is trivially unbiased in some interval $[\theta_0, \theta_0 + \Delta)$. It follows that $\beta'_{\phi_0}(\theta_0) \geq 0$, suggesting that for testing $\theta = \theta_0$ against $\theta \neq \theta_0$ we seek a test ϕ_0 with power function

$$\beta_{\phi_0}(\theta_0) = \alpha, \quad \beta'_{\phi_0}(\theta_0) = 0 \quad \text{and}$$
$$\beta''_{\phi_0}(\theta_0) \text{ maximum.}$$

Such a test is called LMP unbiased of size α for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$.

An LMP unbiased test is of the form

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & \text{if } \dfrac{\partial^2}{\partial\theta^2} f(\mathbf{x};\theta)\Big|_{\theta=\theta_0} \\ & \quad > k_1 f(\mathbf{x};\theta_0) + k_2 \dfrac{\partial}{\partial\theta} f(\mathbf{x};\theta)\Big|_{\theta=\theta_0}, \\ \boldsymbol{\gamma(\mathbf{x})}, & \text{if } \dfrac{\partial^2}{\partial\theta^2} f(\mathbf{x};\theta)\Big|_{\theta=\theta_0} \\ & \quad = k_1 f(\mathbf{x};\theta_0) + k_2 \dfrac{\partial}{\partial\theta} f(\mathbf{x};\theta)\Big|_{\theta=\theta_0}, \\ 0, & \text{if } \dfrac{\partial^2}{\partial\theta^2} f(\mathbf{x};\theta)\Big|_{\theta=\theta_0} \\ & \quad < k_1 f(\mathbf{x};\theta_0) + k_2 \dfrac{\partial}{\partial\theta} f(\mathbf{x};\theta)\Big|_{\theta=\theta_0}, \end{cases}$$

where k_1, k_2, and $\gamma(\cdot)$ are chosen to satisfy $\beta_{\phi_0}(\theta_0) = \alpha$ and $\beta'_{\phi_0}(\theta_0) = 0$.

For the Cauchy density function in the first section, the critical region of the LMP test is the set of points $\mathbf{x}$ such that

$$2\sum_{i=1}^{n} \frac{x_i^2 - 1}{(1 + x_i^2)^2} + \left[\sum_{i=1}^{n} \frac{2x_i}{1 + x_i^2}\right]^2 > k_1,$$

where k_1 is chosen to satisfy $\beta_{\phi_0}(\theta_0) = \alpha$. This test is not a two-sided version of the LMP test given in the first section.

Locally Most Powerful Invariant Tests

Similar considerations apply when attention is restricted to the class of tests that are invariant under a group of transformations on the sample space. Then it is sufficient to consider test statistics that are functions of the maximal invariant and local optimality criteria may be applied to its density.

Consider, for example, the **nonparametric** two-sample problem. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be random samples from respective (continuous) distribution functions F and G. Suppose we wish to test $\mathrm{H}_0\colon F(x) \geq G(x)$ for all x against $\mathrm{H}_1\colon F(x) \leq G(x)$ for all x [$F(x) \neq G(x)$ for some x]. Restricting attention to the **sufficient statistics** $X_{(1)} < X_{(2)} < \cdots < X_{(m)}$ and $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$, the problem is invariant under continuous monotone transformations and a maximal invariant is the set of ranks $(R_1, R_2, \ldots, R_m, S_1, S_2, \ldots, S_n)$, where $R_i =$ **rank** of X_i in the combined sample and $S_j =$ rank of Y_j in the combined sample. Invariance considerations lead us to focus attention on tests that depend only on $R_1, R_2, \ldots, R_m$. Again, a UMP rank test of H_0 does not exist, but one can obtain LMP rank tests.

Suppose, for example, that we fix g, the probability density function corresponding to G and consider the location problem of testing $\mathrm{H}_0\colon f(x) = g(x)$ against $\mathrm{H}_1\colon f(x) = g(x - \theta)$ for values of $\theta > 0$. Then the methods of the first section lead to the LMP test: reject $\mathrm{H}_0\colon \theta = 0$ against $\mathrm{H}_1\colon \theta > 0$ for large values of the linear rank statistic $\sum_{i=1}^{m} a(R_i)$, where

$$a(i) = E\left[-\frac{g'(G^{-1}(U_{(i)}))}{g(G^{-1}(U_{(i)}))}\right], \quad i = 1, 2, \ldots, m,$$

and $U_{(1)} < U_{(2)} < \cdots < U_{(N)}$ are the **order statistics** for a random sample of size $N = m + n$ from a **uniform** (0, 1) distribution. The special case when g is normal (0, 1) leads to the well-known Fisher–Yates test (*see* **Normal Scores**), while when g is logistic, the resulting test is the **Wilcoxon–Mann–Whitney test**.

LMP rank tests are especially useful when the data are **censored**. Rank tests with type II censored data have been discussed by Johnson [4] and Mehrotra et al. [6], and by Prentice [12] and Peto & Peto [10] for arbitrarily censored data.

Ferguson [1, Sections 5.5 and 5.7] is an easily accessible source for LMP tests. Both Lehmann [5] and Schmetterer [14, Section III.6], give a more measure-theoretic treatment. Hájek & Šidák [2, Section III.4] give a fairly general treatment of LMP rank tests for various hypotheses of invariance. At a somewhat lower level, one can refer to Randles & Wolfe [13, Section 9.1]. For the multiparameter case, see Isaacson [3], Schmetterer [14], Neyman [7], and Neyman & Pearson [9].

References

[1] Ferguson, T. (1967). *Mathematical Statistics*. Academic Press, New York.

[2] Hájek, J. & Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

[3] Isaacson, S.L. (1951). On the theory of unbiased tests of simple statistical hypotheses specifying the values of two or more parameters, *Annals of Mathematical Statistics* **22**, 217–234.

[4] Johnson, R. (1974). *Reliability and Biometry*. SIAM, Philadelphia.

[5] Lehmann, E. (1997). *Testing Statistical Hypotheses*. 2nd Ed. Springer-Verlag, New York.

[6] Mehrotra, K., Johnson, R. & Bhattacharya, G. (1977). Locally most powerful tests for multiple-censored data, *Communications in Statistics – Theory and Methods* **6**, 459–470.

[7] Neyman, J. (1935). Sur la vérification des hypothèses statistiques composées, *Bulletin de la Société Mathematique de France, Paris* **63**, 246–266.

[8] Neyman, J. & Pearson, E. (1936). Contributions to the theory of testing statistical hypotheses, *Statistical Research Memoirs* **1**, 1–37.

[9] Neyman, J. & Pearson, E. (1938). Contributions to the theory of testing statistical hypotheses, *Statistical Research Memoirs* **2**, 25–57.

[10] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society, Series A* **135**, 185–198.

[11] Pfanzagl, J. (1963). Überall trennscharfe tests und monotone Dichtequotienten, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **1**, 109–115.

[12] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika* **65**, 167–179.

[13] Randles, R. & Wolfe, D. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

[14] Schmetterer, L. (1974). *Introduction to Mathematical Statistics*. Springer-Verlag, New York.

(*See also* **Large-sample Theory**; **Linear Rank Tests in Survival Analysis**)

EDSEL A. PEÑA & VIJAY K. ROHATGI

Location–Scale Family

A set of random variables $X_1, \ldots, X_n$ is said to have a location–scale family distribution with parameter (μ, σ) if their joint cumulative distribution function (cdf) can be expressed as

$$F(x_1, \ldots, x_n|\mu, \sigma) = F\left(\frac{x_1 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}\right), \quad \mu \text{ real}, \sigma > 0,$$

for some cdf $F(\cdot)$. Equivalently $(X_1, \ldots, X_n)$ has a location–scale family with parameter (μ, σ) if the joint cdf of $(T_1, \ldots, T_n)$ is $F(t_1, \ldots, t_n)$, where $T_i = (X_i - \mu)/\sigma$, $F(t_1, \ldots, t_n)$ is any n-dimensional cdf, and different $F(\cdot)$s correspond to different location–scale families. The parameter μ is the location parameter and σ is the scale parameter. The parameter (μ, σ) is defined as the location-scale parameter of a random variable X if and only if the distribution of $(x - \mu)/\sigma$ under (μ, σ) is free from μ and σ.

From any location–scale family of distributions, two important subfamilies are obtained; namely, a location family with the parameter μ when σ is fixed (and without loss of generality $\sigma = 1$), and a scale family with the parameter σ when μ is fixed (and without loss of generality $\mu = 0$).

Corresponding to any location–scale family $F(x_1, \ldots, x_n|\mu, \sigma)$, the member of the family with $\mu = 0$ and $\sigma = 1$ has a cdf $F(x_1, \ldots, x_n)$ and is referred to as the "standard" or "generator" of the family, generated through a group of location and scale transformations. If $F(x_1, \ldots, x_n)$ has a probability density function (pdf) $f(x_1, \ldots, x_n)$ with respect to a Lebesgue measure, then the continuous location–scale family has a pdf

$$\frac{1}{\sigma^n} f\left(\frac{x_1 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}\right).$$

Some important examples of location–scale family distributions are **uniform** $(\mu - \sigma, \mu + \sigma)$, **normal** (μ, σ) (here σ is the standard deviation), and **Cauchy** (μ, σ).

Parameter Estimation

Least Squares Estimation

Order statistics play an important role in the estimation of μ and σ. We assume that $X_1, \ldots, X_n$ are independent, identically distributed (iid) with a location scale pdf $(1/\sigma)h[(x - \mu)/\sigma]$, where $h(\cdot)$ is known. From the property of the location–scale family, it follows that $X_i = \mu + \sigma Z_i$, $i = 1, \ldots, n$, where $Z_1, \ldots, Z_n$ are iid with pdf $h(z)$. If $Y_1, \ldots, Y_n$ are the order statistics based on $X_1, \ldots, X_n$, and $Z_{(1)}, \ldots, Z_{(n)}$ are the order statistics based on $Z_1, \ldots, Z_n$, then $Y_i = \mu + \sigma Z_{(i)}$, $i = 1, \ldots, n$. Since $h(\cdot)$ is a known pdf, $\mathrm{E}(Z_{(i)}) = \alpha_i$, and $\mathrm{cov}(Z_{(i)}, Z_{(j)}) = w_{ij}$, $i, j = 1, \ldots, n$, are known [assuming the first two moments of $h(\cdot)$ exist]. Then, we get

$$\mathrm{E}(\mathbf{Y}) = \mu\mathbf{1} + \sigma\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\theta},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$, and $\mathbf{1}$ is a vector with unit elements, $\mathbf{A} = (\mathbf{1}, \boldsymbol{\alpha})$, $\boldsymbol{\theta} = (\mu, \sigma)^{\mathrm{T}}$ and $\mathrm{var}(\mathbf{Y}) = \sigma^2\mathbf{w}$, where $\mathbf{w}$ is the matrix of the elements w_{ij}. Then weighted **least squares** estimates of μ and σ are given by

$$\hat{\mu} = -\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{Y}, \quad \hat{\sigma} = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{Y},$$

where $\boldsymbol{\Gamma} = \boldsymbol{\Omega}(\mathbf{1}\boldsymbol{\alpha}^{\mathrm{T}} - \boldsymbol{\alpha}\mathbf{1}^{\mathrm{T}})\boldsymbol{\Omega}/\Delta$, $\boldsymbol{\Omega} = \mathbf{w}^{-1}$ and $\Delta = |\mathbf{A}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{A}|$. The variance–**covariance matrix** of these estimates is given by

$$\frac{\sigma^2}{\Delta}\begin{pmatrix} \mathbf{alpha}^{\mathbf{T}}\boldsymbol{\Omega}\boldsymbol{\alpha} & -\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\alpha} \\ -\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\alpha} & \mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{1} \end{pmatrix}.$$

For details, see Lloyd [6].

Minimum Risk Equivariant Estimation

Since a location–scale family is a group family (see [4, pp. 19–21]), invariance consideration plays an important role in inference. For a location family of distributions $f(x_1 - \mu, \ldots, x_n - \mu)$ and a group of location transformations, a maximal invariant statistic is given by $(x_1 - x_n, \ldots, x_{n-1} - x_n)$. This is an **ancillary statistic** since its distribution does not depend on μ.

Under a squared error **loss**, the minimum risk equivariant (MRE) estimate (if it exists) of μ is given

by (see, for example, [4, p. 160])

$$\frac{\int \mu f(x_1 - \mu, \ldots, x_n - \mu)\, \mathrm{d}\mu}{\int f(x_1 - \mu, \ldots, x_n - \mu)\, \mathrm{d}\mu}, \tag{1}$$

and is known as the Pitman estimate of μ. If $X_1, \ldots, X_n$ are iid $\mathrm{N}(\mu, \sigma)$ with σ known, then the above estimate reduces to $\overline{x}$. In this case, it is also the uniformly **minimum variance unbiased estimate** (UMVUE) of μ.

Similarly, for a scale family of distributions $(1/\sigma^n) f(x_1/\sigma, \ldots, x_n/\sigma)$ and a group of scale transformations, a maximal invariant statistic is given by $(x_1/x_n, \ldots, x_{n-1}/x_n, x_n/|x_n|)$ (see [4, p. 174]) which is an ancillary statistic. Under the loss function $(a/\sigma^r - 1)^2$, the MRE estimate (if it exists) of σ^r is given by (see, [4, p. 177])

$$\frac{\int_0^\infty \sigma^{n+r-1} f(\sigma x_1, \ldots, \sigma x_n)\, \mathrm{d}\sigma}{\int_0^\infty \sigma^{n+2r-1} f(\sigma x_1, \ldots, \sigma x_n)\, \mathrm{d}\sigma}. \tag{2}$$

For $X_1, \ldots, X_n$ iid $\mathrm{N}(0, \sigma)$ and $r = 2$, the above estimate reduces to $\sum x_i^2/(n+2)$.

For a location–scale family of distributions $(1/\sigma^n) f[(x_1 - \mu)/\sigma, \ldots, (x_n - \mu)/\sigma]$ and a group of location–scale transformations, a maximal invariant statistic is given by (see [4, p. 179])

$$\left(\frac{x_1 - x_n}{x_{n-1} - x_n}, \ldots, \frac{x_{n-2} - x_n}{x_{n-1} - x_n}, \frac{x_{n-1} - x_n}{|x_{n-1} - x_n|} \right).$$

Estimation of $\beta\mu + \gamma\sigma$ for known β and γ is important. (The case $\beta = 1$, $\gamma = 0$, corresponds to the estimation of μ, whereas $\beta = 0$, $\gamma = 1$, corresponds to the estimation of σ, and $\beta = 1$ and given γ corresponds to the estimation of a certain percentile.)

Under an invariant loss function $L(\mu, \sigma, a) = w[(a - \beta\mu - \gamma\sigma)/\sigma]$, the MRE estimator (if it exists) of $\beta\mu + \gamma\sigma$ has been discussed in detail in Datta & Ghosh [2]. For the loss function $w(x) = x^2$, and for $X_1, \ldots, X_n$ iid $\mathrm{N}(\mu, \sigma)$, the MRE estimator of $\beta\mu + \gamma\sigma$ is given by $\beta\overline{X} + \gamma k S$, where

$$k = \frac{(n-1)^{1/2}\Gamma(n/2)}{\sqrt{2}\Gamma[(n+1)/2]},$$

$$S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2$$

(see [3, p. 182]). The UMVUE of μ and σ^2 for the $\mathrm{N}(\mu, \sigma)$ problem are $\overline{X}$ and S^2, respectively.

It follows from Berger [1, p. 410] that the MRE estimates for μ in (1), for σ^r in (2), and for $\beta\mu + \gamma\sigma$ are generalized Bayes estimates with respect to the right invariant Haar density for the respective group of location, scale, and location–scale transformations (*see* **Decision Theory**).

Hypothesis Tests

To test for location and scale parameters, the most widely used assumption is that $X_1, \ldots, X_n$ are iid $\mathrm{N}(\mu, \sigma)$. In this setup, to test $\mathrm{H}_0 : \mu = \mu_0$ vs. $\mathrm{H}_1 : \mu \neq \mu_0$, for example, the rejection region for known σ is

$$\left| \frac{n^{1/2}(\overline{X} - \mu_0)}{\sigma} \right| \geq z_{\frac{\alpha}{2}},$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$th percentile point of a **standard normal** distribution. For unknown σ, the corresponding rejection region is obtained by replacing σ and $z_{\alpha/2}$ in the preceding rejection region by S and $t_{\alpha/2}$, the $100(1 - \alpha/2)$th percentile point of **Student's *t* distribution**, with $n - 1$ degrees of freedom, respectively. The above tests can be shown to be uniformly most powerful unbiased (UMPU) tests and can be derived as **likelihood ratio tests**. To test for $\sigma^2, \mathrm{H}_0 : \sigma^2 = \sigma_0^2$ vs. $\mathrm{H}_1 : \sigma^2 \neq \sigma_0^2$, the widely used test which rejects if

$$\frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{\frac{1-\alpha}{2}} \quad \text{or} \quad \frac{(n-1)S^2}{\sigma_0^2} \geq \chi^2_{\frac{\alpha}{2}}$$

is an approximate (for large n) UMPU size α test, where $\chi^2_{\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the χ^2_{n-1} distribution (**chi-square distribution** with $n - 1$ **degrees of freedom**). For details on these tests and other distribution-free tests for the location and scale parameters, the reader is referred to Lehmann [5].

References

[1] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

[2] Datta, G.S. & Ghosh, M. (1988). Minimum risk equivariant estimators of percentiles in location–scale families of distributions, *Calcutta Statistical Association Bulletin* **37**, 201–207.

[3] Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

[4] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[5] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. Wiley, New York.

[6] Lloyd, E.H. (1952). Least squares estimation of location and scale parameters using order statistics, *Biometrika* **39**, 88–95.

GAURI SANKAR DATTA

Logistic Distribution

The logistic **random variable** X with mean μ and variance σ^2 has a cumulative distribution function

$$F(x, \mu, \sigma) = \left\{1 + \exp\left[\frac{-\pi(x-\mu)}{(\sigma\sqrt{3})}\right]\right\}^{-1}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0, \tag{1}$$

and density function f, which is simply related to its distribution function by

$$f(x, \mu, \sigma) = \frac{\pi}{\sigma\sqrt{3}} F(x, \mu, \sigma)[1 - F(x, \mu, \sigma)]. \tag{2}$$

We denote this distribution by $\mathcal{L}(\mu, \sigma^2)$. These functions may also be expressed as

$$F(x, \mu, \sigma) = \frac{1}{2}\left\{1 + \tan h\left[\frac{\pi}{2}\frac{(x-\mu)}{(\sigma\sqrt{3})}\right]\right\} \tag{3}$$

and

$$f(x, \mu, \sigma) = \frac{\pi}{4\sigma\sqrt{3}} \operatorname{sech}^2\left[\frac{\pi}{2}\frac{(x-\mu)}{(\sigma\sqrt{3})}\right], \tag{4}$$

with the latter expression providing the logistic with the sech-square(d) distribution label. The density f is bell-shaped and symmetrical, with heavier tails than a normal density with the same mean and variance.

To describe some of the basic properties of the logistic distribution, it is simpler to use the "canonical form", $\mathcal{L}(0, \pi^2/3)$, which corresponds to the random variable Z, with mean $\mu = 0$ and variance $\sigma^2 = \pi^2/3$, and has cumulative distribution and density functions

$$G(z) = \frac{1}{1 + \mathrm{e}^{-z}}, \tag{5}$$

$$g(z) = G(z)[1 - G(z)], \tag{6}$$

and a monotonic *hazard* function

$$\lambda(z) = \frac{g(z)}{1 - G(z)} = G(z). \tag{7}$$

Eq. (6), and therefore (2), characterizes the logistic distribution and is equivalent to the linearity of the **transformation**

$$\log\left[\frac{G(z)}{1 - G(z)}\right] = z. \tag{8}$$

This transformation, which is labeled *logit* by Berkson [7], is perhaps the single best known and most popular application of the logistic distribution, especially in the context of modeling **quantal response** data, and performing **logistic regression**.

The distribution function of the standardized random variable $Z/(\pi/\sqrt{3})$, is very close to the **standard normal** distribution, and even closer to the distribution function of a normal random variable with zero mean and standard deviation 15/16 [32]. However, this distribution function is even better approximated by that of a standardized **Student's *t* distribution** with nine **degrees of freedom** [38]. Moreover, unlike the normal distribution, the sum of independent logistic random variables is not a logistic random variable. Goel [23] and George & Mudholkar [19] give closed-form expressions for the distribution function, and the latter authors also propose a simple Student's t approximation.

Characteristic Function

The **characteristic function** of Z may be expressed in the forms

$$\phi_Z(t) = \Gamma(1 - it)\Gamma(1 + it) = \prod_{j=1}^{\infty}\left(1 - \frac{t^2}{j^2}\right)^{-1} \tag{9}$$

and

$$\phi_Z(t) = \sum_{k=1}^{\infty}(-1)^{k-1}\frac{2(2^{2k} - 1)}{(2k)!}B_{2k}(\pi i t)^{2k}, \tag{10}$$

where the B_{2k}s are Bernoulli numbers [54]. The characteristic function (9), or direct integration, may be used to obtain the absolute **moments**

$$\mathrm{E}|Z|^k = 2\Gamma(k+1)\left[1 - \frac{1}{2^{k+1}}\zeta(k)\right], \tag{11}$$

where $\zeta(k) = \sum_{j=1}^{\infty} j^{-k}$, is the zeta function.

From (9), we get the following equalities in distribution [18]:

$$Z \stackrel{\mathcal{D}}{=} \sum_{j=1}^{\infty} W_j$$
$$\stackrel{\mathcal{D}}{=} \sum_{j=1}^{\infty} (E_{1j} - E_{2j}) \stackrel{\mathcal{D}}{=} Y_1 - Y_2, \qquad (12)$$

where the W_js are independent Laplace or double exponential random variables, and the E_{ij}s are independent **exponential** random variables with respective densities $f_{W_j}(w) = (j/2)\exp(-j|w|), -\infty < \omega < \infty$, $f_{E_{ij}}(x) = j\exp(-jx), i = 1, 2; j = 1, 2, \ldots$ and Y_1, Y_2 are iid **extreme value** random variables with density $h(y) = e^{-y}\exp(-e^{-y}), -\infty < y < \infty$. The logistic distribution is also obtained from a mixture of the extreme value distribution and the exponential distribution [13]. From (12) we may conclude immediately that the logistic distribution is infinitely divisible.

Order Statistics

Let $Z_{1:n} \le Z_{2:n} \le \cdots \le Z_{n:n}$ be **order statistics** of a random sample from $\mathcal{L}(0, \pi^2/3)$. Then it can be shown that the characteristic function of $Z_{r:n}$ may be expressed as

$$\phi_{r:n}(t) = \prod_{j=1}^{r-1}\left(1 + \frac{it}{j}\right)\prod_{k=1}^{n-r}\left(1 - \frac{it}{k}\right)\phi_Z(t) \qquad (13)$$

(see [10, 28, 46], and [47]). Consequently,

$$Z_{r:n} + \sum_{k=1}^{n-r} E_{1k} - \sum_{j=1}^{r-1} E_{2j} \stackrel{\mathcal{D}}{=} Z_1, \qquad (14)$$

where the E_{ij}s are independent exponential random variables with densities $f_{E_{ij}}$ given above, $i = 1, 2$, $j = 1, \ldots, n-1$. Gupta & Shah [28] provide percentage points for the rth order statistics, $Z_{r:n}$, for $1 \le n \le 25$. Shah [52] and Gupta & Balakrishnan [27] provide an extensive list of recurrence relations for the moments of order statistics of the logistic distribution. Gupta & Shah [29] and Malik [37] give closed-form expressions for the **range**, $R_n = Z_{n:n} - Z_{1:n}$, and the rth quasi-range, $Z_{n-r:n} - Z_{r+1:n}$. By expressing the distribution of the range in terms of an associated Legendre function, George & Rousseau [22] obtain the recurrence relation

$$nP(R_{n+2} \le x) = (2n+1)\left(\frac{1 + e^{-x}}{1 - e^{-x}}\right) \times \Pr(R_{n+1} \le x) - (n+1) \times \Pr(R_n \le x). \qquad (15)$$

George & Rousseau [21] show that the characteristic function of the midrange $(Z_{n:n} + Z_{1:n})/2$ is a well-poised hypergeometric function, which may be expressed as

$$\varphi_n(t) = \begin{cases} \prod_{j=1}^{p-1}\left(1 + \frac{t^2}{4j^2}\right)[\phi_Z\left(\frac{t}{2}\right)]^2, \\ \qquad \text{if } n = 2p, \\ \prod_{j=1}^{p}\left[1 + \frac{t^2}{(2j-1)^2}\right]\phi_Z(t), \\ \qquad \text{if } n = 2p+1, \end{cases} \qquad (16)$$

and obtain a closed-form expression for its distribution. For a sample of size three, they establish the rather interesting relationship

$$\frac{Z_{1:3} + Z_{3:3}}{2} \stackrel{\mathcal{D}}{=} Z_{2:3}. \qquad (17)$$

Gumbel [24], relating the logistic to extreme value distributions, shows that, for a large family of symmetric distributions satisfying a general set of conditions that are formalized by de Haan [11], the limiting distribution of the midrange is logistic [17]. This result is extended by Gumbel to the "mth midrange", i.e. $(Z_{m:n} + Z_{n-m+1:n})/2$. In this case, the asymptotic distribution is generalized logistic. Gumbel & Keeney [26] show that the logistic is the asymptotic distribution of a family of extremal quotients.

Generalized Logistic

It is easy to see that if U is **uniformly distributed** on the unit interval (0,1), then the logit transform of U, $\log\left[U/(1-U)\right]$, has the logistic distribution function G. In fact, one of the many generalizations of the logistic distribution is obtained by simply replacing the uniform random variable U (which is equal in distribution to a **beta** (1,1) random variable), with a beta

(α, β) random variable, [20, 48]. When $\alpha = \beta$, the symmetric generalized logistic is obtained. Like the logistic distribution, the generalized logistic is used for modeling binary response data [48] and the log of survival times [35]. In the context of application to quantal assay data, Stukel [53] proposes another generalization of the logistic distribution by introducing different shape parameters at the tails of the distribution (*see* **Quantal Response Models**).

Parametric Estimation

The simplicity of the logistic distribution, as expressed by (1)–(6), belies the complexity of the process of estimating its parameters. No closed forms exist for the MLE (**maximum likelihood** estimator), BLUE (best linear **unbiased** estimators), or UMVUE (uniform **minimum variance unbiased estimators**) of the mean μ and variance σ^2. For example, given a random sample $X_1, \ldots, X_n$ from an $\mathcal{L}(\mu, \sigma^2)$ population, the estimating equations for the MLE of μ and σ^2, which must be solved iteratively, may be expressed by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\exp[\pi(X_i-\mu)/(\sigma\sqrt{3})]}=\frac{1}{2} \tag{18}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)\left(\frac{1-\exp\left[\dfrac{\pi(X_i-\mu)}{(\sigma\sqrt{3})}\right]}{1+\exp[\pi(X_i-\mu)/(\sigma\sqrt{3})]}\right)=\frac{\sqrt{3}}{\pi}. \tag{19}$$

From Gupta & Gnanadesikan [28] (in which explicit approximate expressions for the BLUE estimates of μ and σ are given based on selected order statistics $X_{n_1:n} \le X_{n_2:n} \le \cdots \le X_{n_k:n}$), Gupta et al. [30] and Harter & Moore [31] (in which linear estimates are calculated from censored data), a vast literature has evolved on the use of **censored** logistic random variables to estimate μ and σ. Using the large sample variance–**covariance matrix** of the MLEs, Antle et al. [1] construct **confidence intervals** for μ and σ. Bain [4], Eastman [14], Schafer & Sheffield [50], and Bain et al. [5] discuss applications in life-testing using complete and censored data. Other accounts involving the use of linear functions of order statistics for estimating the logistic parameters are given by several authors in Balakrishnan [6, Chapter 4].

Multivariate Distributions

A model for a **bivariate distribution** with logistic marginals first proposed by Gumbel [25] is extended by Malik & Abraham [38] to an m-dimensional multivariate distribution function

$$F_{\mathbf{Z}}(\mathbf{z}) = F_{Z_1,\ldots,Z_m}(z_1,\ldots,z_m) = \left(1+\sum_{i=1}^{m}\exp(-z_i)\right)^{-1}, \tag{20}$$

with density function

$$f_{\mathbf{Z}}(\mathbf{z}) = m!\frac{\exp\left(-\sum_{i=1}^{m}z_i\right)}{\left[1+\sum_{i=1}^{m}\exp(-z_i)\right]^{m+1}}, \tag{21}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_m)$ and $\mathbf{z} = (z_1, \ldots, z_m)$. This distribution, which is sometimes referred to as the Gumbel–Malik–Abraham model, suffers from the restriction that the **correlation** between any pair Z_i, Z_j, is 1/2.

The joint **moment generating function** of $\mathbf{Z}$ is given by

$$M_{\mathbf{Z}}(t_1,\ldots,t_m) = \Gamma\left(1+\sum_{j=1}^{m}t_j\right)\prod_{j=1}^{m}\Gamma(1-t_j). \tag{22}$$

From this generating function, Arnold [3] observes that, analogous to the univariate logistic distribution, the joint distribution of $(Z_1, \ldots, Z_m)$ is the same distribution as $(Y_1 - Y_0, \ldots, Y_m - Y_0)$, where $Y_0, Y_1, \ldots, Y_m$ are independent, identically distributed (iid) extreme value random variables with density given by $h(y) = \mathrm{e}^{-y}\exp(-\mathrm{e}^{-y})$, $-\infty < y < \infty$.

The Gumbel–Malik–Abraham model is one example of a multivariate logistic distribution that can be constructed by using a multivariate analog of a property of the univariate logistic distribution. Others are described by Arnold [3]. These include

a representation in terms of a multivariate survival function:

$$\Pr(\mathbf{Z} \geq \mathbf{z}) = \left[1 + \sum_{j=1}^{m} \exp(z_i) + \sum_{j_1 \neq j_2} c_{j_1 j_2} \exp(z_{j_1} + z_{j_2}) + \cdots + c_{1\ldots m} \exp(z_1 + z_2 + \cdots + z_m) \right]^{-1}, \quad (23)$$

where $\mathbf{Z} \geq \mathbf{z}$ denotes the event $Z_1 \geq z_1, Z_2 \geq z_2, \cdots, Z_m \geq z_m$ and the cs are chosen to satisfy conditions that make (23) a true survival function [2] (*see* **Survival Distributions and Their Characteristics**). This expression can be obtained from a multivariate analog of the following result: if $Z_1, Z_2, \ldots,$ are iid $\mathcal{L}(0, \pi^2/3)$ variables and N is a **geometric** random variable with $\Pr(N = n) = pq^{n-1}$, $q = 1 - p$, then

$$Z_{1;N} - \log p \stackrel{\mathcal{D}}{=} Z_{N;N} + \log p \stackrel{\mathcal{D}}{=} Z_1. \quad (24)$$

Eq. (23) clearly generalizes the Gumbel–Malik–Abraham representation. As an example, the bivariate logistic distribution function obtained from (23) is given by

$$F_{Z_1, Z_2}(z_1, z_2) = [1 + \exp(-z_1) + \exp(-z_2) + \theta \exp(-z_1 - z_2)]^{-1}, \quad (25)$$

where $0 \leq \theta \leq 2$.

Another representation given by Arnold [3] uses the concept of **frailty** from survival analysis to obtain

$$\Pr(\mathbf{Z} \geq \mathbf{z}) = \Lambda_F \left\{ \sum_{j=1}^{m} \Lambda_F^{-1}[1 + \exp(z_j)]^{-1} \right\}, \quad (26)$$

where Λ_F denotes the Laplace transform of a given distribution function F. Using distribution functions instead of survival functions leads to a different, but related, family of multivariate logistic distribution functions

$$\Pr(\mathbf{Z} \leq \mathbf{z}) = \Lambda_F \left\{ \sum_{j=1}^{m} \Lambda_F^{-1}[1 + \exp(-z_j)]^{-1} \right\}. \quad (27)$$

Examples of multivariate logistic distributions from these models are:

$$\Pr(\mathbf{Z} \geq \mathbf{z}) = \left\{ \sum_{j=1}^{m} [1 + \exp(z_j)]^{(1/\alpha)} - m + 1 \right\}^{-\alpha} \quad (28)$$

and

$$\Pr(\mathbf{Z} \leq \mathbf{z}) = \left\{ \sum_{j=1}^{m} [1 + \exp(-z_j)]^{(1/\alpha)} - m + 1 \right\}^{-\alpha}, \quad (29)$$

corresponding to a choice of **gamma** $(\alpha, 1)$ for F and

$$\Pr(\mathbf{Z} \leq \mathbf{z}) = \exp\left[- \left(\sum_{j=1}^{m} \{\log[1 + \exp(-z_j)]\}^{(1/\alpha)} \right)^{\alpha} \right], \quad (30)$$

corresponding to choosing $\Lambda_F(t) = \exp(-t^\alpha)$, $\alpha \leq 1$.

The Farlie–Gumbel–Morgenstern model of a multivariate logistic [3, 33, 34] is yet another representation. This model may be described by

$$\Pr(\mathbf{Z} \leq \mathbf{z}) = \prod_{j=1}^{m} G(z_j) \left\{ 1 + \alpha \prod_{j=1}^{m} [1 - G(z_j)] \right\}, \quad (31)$$

where $|\alpha| < 1$ and $G(z) = (1 + \mathrm{e}^{-z})^{-1}$. This model suffers from a restriction in correlation: $\rho(Z_i, Z_j) = 3\alpha/\pi^2$ for every pair Z_i, Z_j. The correlation structure limits the use of the model. The bivariate version of this model is due to Gumbel [25].

Historical Notes and Applications

The logistic function is one of the oldest models for analyzing **demographic** and organismic growth data. Verhulst [56], Pearl [43, 44], Pearl & Reed [45], Yule [57], and, more recently, Oliver [41, 42] and Leach [36] discuss applications to **population growth**. Other biological applications of the logistic function include the modeling of the growth of yeast cells [40, 45, 51] and the use of the logistic function in analysis of survival data [47].

Reed & Berkson [49], who are usually credited with the logit label, for the inverse transformation of the logistic function, and Berkson [7–9] have championed the use of the logistic distribution function for modeling **dose–response** curves in **bioassay** (see also Finney [15, 16]). Berkson's minimum logit chi-square estimates are easier to compute than maximum likelihood estimates. However, with the availability of sophisticated software, this is no longer a significant advantage over the efficiency of the maximum likelihood estimates. From the limited use of the logistic distribution for quantal bioassay has emerged logistic regression analysis, which is currently a very popular **generalized linear model** procedure for analyzing **binary data**. In the context of applications of logistic regression to health and social sciences, Tsokos & DiCroce [55] give an extensive bibliography. Prentice [48] Stukel [53], Devidas, et al. [12] and others discuss applications of generalizations of the logistic models in low-dose bioassays.

References

[1] Antle, C., Klimko, L. & Harkness, W. (1970). Confidence intervals for the parameters of the logistic distribution, *Biometrika* **57**, 397–402.

[2] Arnold, B.C. (1990). A flexible family of multivariate Pareto distributions, *Journal of Statistical Planning and Inference* **24**, 249–258.

[3] Arnold, B.C. (1992). Multivariate logistic distributions, in *Handbook of the Logistic Distribution*, N. Balakrishnan, ed. Marcel Dekker, New York, Chapter 11.

[4] Bain, L.J. (1978). *Statistical Analysis of Reliability and Life-Testing Models – Theory and Practice*. Marcel Dekker, New York.

[5] Bain, L.J., Balakrishnan, N., Eastman, J.A., Engelhart, M. & Antle, C.A. (1992). Reliability estimation based on MLEs for complete and censored samples, in *Handbook of the Logistic Distribution*, N. Balakrishnan, ed. Marcel Dekker, New York, Chapter 5.

[6] Balakrishnan, N. (1992). Maximum likelihood estimation based on complete and Type II censored samples, *Handbook of the Logistic Distribution*, N. Balakrishnan, ed. Marcel Dekker, New York, Chapter 3.

[7] Berkson, J. (1994). Application of the logistic function to bioassay, *Journal of the American Statistical Association* **37**, 357–365.

[8] Berkson, J. (1951). Why I prefer logits to probits, *Biometrics* **7**, 327–339.

[9] Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bioassay and quantal response, based on the logistic function, *Journal of the American Statistical Association* **48**, 565–599.

[10] Birnbaum, A. & Dudman, J. (1963). Logistic order statistics, *Annals of Mathematical Statistics* **34**, 658–663.

[11] de Haan, L. (1975). *On Regular Variation and Its Application to Weak Convergence of Sample Extremes*, 3rd Ed. Mathematical Center Tracts, Vol. 32, Amsterdam.

[12] Devidas, M., George, E.O. & Zelterman D. (1993). Generalized logistics models for low-dose response data, *Statistics in Medicine*, **12**, 881–892.

[13] Dubey, S.D. (1969). A new derivation of the logistic distribution, *Naval Research Logistics Quarterly* **16**, 37–40.

[14] Eastman, J.A. (1972). Statistical Issues of Various Time-to-Fail Distributions, *Doctoral Thesis*. University of Missouri-Rolla, Missouri.

[15] Finney, D.J. (1947). The principles of biological assay, *Journal of the Royal Statistical Society, Series B* **9**, 46–91.

[16] Finney, D.J. (1952),. *Statistical Methods in Biological Assay*. Hafner, New York.

[17] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd Ed. Krieger, Melbourne, Florida.

[18] George, E.O. & Devidas, M. (1992). Some related distributions, in *Handbook of the Logistic Distribution*, N. Balakrishnan, ed. Marcel Dekker, New York, Chapter 10.

[19] George, E.O. & Mudholkar, G.S. (1983). On the convolution of logistic random variables, *Metrika* **30**, 1–13.

[20] George, E.O. & Ojo, M.O. (1980). On a generalization of the logistic distribution, *Annals of the Institute of Statistical Mathematics* **32**, 161–169.

[21] George, E.O. & Rousseau, C.C. (1987). On the logistic midrange, *Annals of the Institute of Statistical Mathematics* **39**, 627–635.

[22] George, E.O. & Rousseau, C.C. (1992). Asymptotics of the logistic range, *Sankhyā, Series B* **54**, 165–169.

[23] Goel, P.K. (1975). On the distribution of standardized mean samples from the logistic population, *Sankhyā, Series B* **37**, 165–172.

[24] Gumbel, E.J. (1944). Ranges and midranges, *Annals of Mathematical Statistics* **15**, 414–422.

[25] Gumbel, E.J. (1961). Bivariate logistic distributions, *Journal of the American Statistical Association* **56**, 335–349.

[26] Gumbel, E.J. & Keeney, R.D. (1950). The extremal quotient, *Annals of Mathematical Statistics* **21**, 523–538.

[27] Gupta, S.S. & Balakrishnan, N. (1992). Logistic order statistics and their properties, in *Handbook of the Logistic Distribution*, N. Balakrishnan, ed. Marcel Dekker, New York, Chapter 2.

[28] Gupta, S.S. & Gnanadesikan, M. (1966). Estimation of the parameters of the logistic distribution, *Biometrika* **53**, 565–570.

[29] Gupta, S.S. & Shah, B.K. (1965). Exact moments and percentage points of the order statistics and the distribution of the range from the logistic distribution, *Annals of Mathematical Statistics* **36**, 907–920.

[30] Gupta, S.S., Qureishi, A.S. & Shah, B.K. (1967). Best linear unbiased estimators of the parameters of the

logistic distribution using order statistics, *Technometrics* **9**, 43–56.

[31] Harter, H.L. & Moore, A.H. (1967). Maximum likelihood estimation, from censored samples, of the parameters of a logistic distribution, *Journal of the American Statistical Association* **62**, 675–684.

[32] Johnson, N.L. & Kotz, S. (1970). *Distribution in Statistics, Continuous Univariate Distributions*, Vol. 2. Wiley, New York.

[33] Johnson, N.L. & Kotz, S. (1975). On some generalized Farlie–Gumbel–Morgenstern distributions, *Communications in Statistics – Theory and Methods* **4**, 415–427.

[34] Johnson, N.L. & Kotz, S. (1977). On some generalized Farlie–Gumbel–Morgenstern distributions, II: Regression, correlations and further generalizations, *Communications in Statistics – Theory and Methods* **6**, 485–496.

[35] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[36] Leach, D. (1981). Re-evaluation of the logistic curve for human populations, *Journal of the Royal Statistical Society, Series A* **144**, 94–103.

[37] Malik, H.J. (1980). Exact formula for the cumulative distribution function of the quasi-range from the logistic distribution, *Communications in Statistics – Theory and Methods* **9**, 1527–1534.

[38] Malik, H.J. & Abraham, B. (1973). Multivariate logistic distribution, *Annals of Statistics* **1**, 588–590.

[39] Oliver, F.R. (1964). Methods of estimating the logistic growth function, *Applied Statistics* **13**, 57–66.

[40] Oliver, F.R. (1966). Aspects of maximum likelihood estimation of the logistic growth function, *Journal of the American Statistical Association* **61**, 697–705.

[41] Oliver, F.R. (1982). Notes on the logistic curve for human populations, *Journal of the Royal Statistical Society, Series A* **145**, 359–363.

[42] Pearl, R. (1925). *The Biology of Population Growth*, Knopf, New York.

[43] Pearl, R. (1940). *Medical Biometry and Statistics*, Sanders, Philadelphia.

[44] Pearl, R. & Reed, L.J. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation, *Proceedings of the National Academy of Sciences* **6**, 275–288.

[45] Plackett, R.L. (1958). Linear estimation from censored data, *Annals of Mathematical Statistics* **29**, 131–142.

[46] Plackett, R.L. (1959). The analysis of life test data, *Technometrics* **1**, 9–19.

[47] Prentice, R.L. (1976). A generalization of the probit and logit methods for dose-response curves, *Biometrics* **32**, 761–768.

[48] Reed, L.J. & Berkson, J. (1929). The application of the logistic function to experimental data, *Journal of Physical Chemistry* **33**, 760–779.

[49] Schafer, R.E. & Sheffield, T.S. (1973). Inferences on the parameters of the logistic distribution, *Biometrika* **29**, 449–455.

[50] Schultz, H. (1930). The standard error of a forecast from a curve, *Journal of the American Statistical Association* **25**, 139–185.

[51] Shah, B.K. (1970). Note on the moments of a logistic order statistics, *Annals of Mathematical Statistics* **41**, 2151–2152.

[52] Stukel, T. (1988). Generalized logistic models, *Journal of the American Statistical Association* **83**, 426–431.

[53] Tarter, M.E. & Clark, V.A. (1965). Properties of the median and other order statistics of the logistic variates, *Annals of Mathematical Statistics* **36**, 1779–1786.

[54] Tsokos, C.P. & DiCroce, P.S. (1992). Applications in health and social sciences, in *Handbook of the Logistic Distribution*, N. Balakrishnan, ed. Marcel Dekker, New York, Chapter 17.

[55] Verhulst, P.J. (1845). Recherches mathematiques sur la loi d'accroissement de la population. *Académie de Bruxelles* **18**, 1–38.

[56] Yule, G.U. (1925). The growth of population and factor which controls it, *Journal of the Royal Statistical Society, Series A* **88**, 1–58.

Further Reading

Mudholkar, G.S. & George, E.O. (1978). A remark on the shape of the logistic distribution, *Biometrika* **65**, 667–668.

E. OLUSEGUN GEORGE

Logistic Regression, Conditional

An important extension of the **logistic regression** model is the analysis of data from stratified samples (*see* **Stratification**). Examples of this application include studies where data are collected from several different sites such as schools, hospitals, or clinics as well as analyses where **covariates** are controlled for by defining *post hoc* stratification variables. The most frequently encountered stratified study design employing the logistic regression model is the matched **case–control study** used in epidemiology (*see* **Matched Analysis**). A discussion of the rationale for these matched studies may be found in epidemiology texts such as Breslow & Day [1], Kleinbaum et al. [5], Schlesselman [8], Kelsey et al. [4], and Rothman [7].

The basic idea is to expand the logistic model by inclusion of stratification variables. Assume the sampled data may be represented as a triple $(y_{kj}, \mathbf{x}_{kj}, \mathbf{z}_k)$, where $j = 1, 2, \ldots, n_k$ represents the particular subject observed within stratum $k = 1, 2, \ldots, K$, $y_{kj} = 0$ or 1 is the observed value of the binary outcome variable for subject j in stratum k, $\mathbf{x}'_{kj} = (x_{kj1}, x_{kj2}, \ldots, x_{kjp})$ is a vector of p nonconstant covariates, and $\mathbf{z}'_k = (z_{k1}, z_{k2}, \ldots, z_{kq})$ is a vector of q covariates defining stratum characteristics. The quantity n_k denotes the number of observations in stratum k. The vector $\mathbf{z}$ may simply contain one variable to indicate the stratum, $z_k = k$, or a set of values of q covariates may be used to define strata. For example, if one defined strata by gender and race coded at three levels, then $\mathbf{z}'_k = (z_{k1}, z_{k2})$ with $z_{k1} = 0$ or 1, $z_{k2} = 1$, 2, or 3, and $k = 1, 2, \ldots, 6$.

A number of different stratified logistic regression models are possible. The simplest logistic regression model has a logit function with one design variable for the stratum specific effect and constant slope across strata for the covariates, namely

$$g(\mathbf{x}_{kj}, z_k) = \beta_0 + \alpha_k + \boldsymbol{\beta}'\mathbf{x}_{kj}. \tag{1}$$

The logit function is discussed in detail in the article on **Logistic Regression**. It is defined in terms of the model conditional probability as $g(\mathbf{x}_{kj}, z_k) = \ln\{\pi(\mathbf{x}_{kj}, z_k)/[1 - \pi(\mathbf{x}_{kj}, z_k)]\}$ and $\pi(\mathbf{x}_{kj}, z_k) = \Pr(Y_{kj} = 1|\mathbf{x}_{kj}, z_k)$. In the parameterization in (1) one may think of the values of α_k as the coefficients for design variables generated by the K levels of the stratum variable. These design variables may be created using any method but the most frequent choice is either referent cell or deviation from means coding. There are $K - 1$ parameters or degrees of freedom associated with the stratification variable. The model in (1) has a stratum-specific intercept and constant slopes. Thus the effect of the covariates is the same for all strata. The covariate vector, $\mathbf{x}$, may contain both main effects as well as higher-order terms such as interactions and squared terms, but may not contain terms that indicate the stratum.

An extension of the model in (1) is possible when the vector $\mathbf{z}$ contains covariates that measure stratum characteristics, e.g. gender and race as noted above. The vector may also contain continuous covariates. Age is often used as a stratification variable. In this setting one may add interactions to (1), which yield a model with stratum-specific slopes. Suppose strata are defined by gender and $z_k = 0$ or $1(1 = \text{male})$ records the gender of the subject. The logit for an extended model is

$$g(\mathbf{x}_{kj}, z_k) = \beta_0 + \alpha_k z_k + \boldsymbol{\beta}'\mathbf{x}_{kj} + z_k \times \boldsymbol{\gamma}'\mathbf{x}_{kj}. \tag{2}$$

The model for females is

$$g(\mathbf{x}_{kj}, z_k = 0) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_{kj},$$

and the model for males is

$$g(\mathbf{x}_{kj}, z_k = 1) = \beta_0 + \alpha_1 + (\boldsymbol{\beta} + \boldsymbol{\gamma})'\mathbf{x}_{kj}.$$

The model in (2) allows for stratum-specific intercepts as well as stratum-specific slopes. Maximum likelihood estimators of the parameters in (1) or (2) are obtained by extending the likelihood function (*see* **Logistic Regression**) to include a product over strata. The **likelihood** function for the model in (1) is

$$l(\alpha, \beta,) = \prod_{k=1}^{K} \prod_{j=1}^{n_k} \zeta(\mathbf{x}_{kj}, z_k), \tag{3}$$

where $\zeta(\mathbf{x}_{kj}, z_k) = \pi(\mathbf{x}_{kj}, z_k)^{y_{kj}}[1 - \pi(\mathbf{x}_{kj}, z_k)]^{1-y_{kj}}$. Application of the likelihood function in (3) to the model in (2) is accomplished by adding the requisite additional terms to the logit. Estimators of the parameters may be obtained from logistic regression software (*see* **Software, Biostatistical**) by inclusion of the variables recording stratum-specific data into the model.

Thus the model as shown in (1) or (2) does not represent anything particularly new or difficult for the investigator familiar with the logistic regression model. The model-building issues and details are identical to those of the ordinary logistic model, or for that matter any regression model.

Problems begin to arise which require a different approach when the number of strata becomes large and, at the same time, the number of observations within each stratum remains fixed. Application of the logistic regression model to this setting will be described in the remainder of this article.

Logistic Regression with Highly Stratified Data

A convenient setting to illustrate the use of logistic regression with highly stratified data is the matched case–control study design. In this study design subjects are stratified on the basis of covariates believed to be associated with the outcome. Age and gender are examples of commonly used stratification variables. Within each stratum a sample of subjects with the outcome present, called cases ($y = 1$), and a sample of subjects without the outcome, called controls ($y = 0$), is chosen. The number of cases and controls need not be constant across strata, but the most common matched design is one where each stratum includes one case and one control. Study variables are collected on all subjects. We develop the methods for analysis of highly stratified data for the general case. Greater detail is provided for the one-to-one matched design because it can be analyzed using standard logistic regression software.

The methods to be described may be used in settings other than matched case–control studies. For example, suppose that, in a study of student performance, data were collected from 1000 different schools and a fixed number of students was selected from each school. The outcome variable is whether the student "passed" a particular course or standardized test. In this example there are 1000 strata defined by school. The conditional likelihood approach described below is the same for both the case–control study and the general highly stratified design. More stringent sampling assumptions are required in the case–control study, see [3, Chapter 6].

We begin by providing some motivation for the need for special methods for the highly stratified study. We noted in (1) that we could handle the stratified sample by including variables created from the stratification variables in the model. This approach works well when the number of subjects in each stratum is large and strata are few. However, matched studies have few subjects per stratum. For example, in the one-to-one matched design with K case–control pairs we have only two subjects per stratum. A fully stratified analysis of the model in (1) with p covariates would require estimation of $(K + p)$ parameters, the $p + 1$ slope coefficients for the covariates, and the $K - 1$ coefficients for the stratum-specific design variables, using a sample of size $2K$. The optimality properties of the method of maximum likelihood, derived by letting the sample size, K, become large, hold only when the number of parameters remains fixed. In any matched study this is not the case, as the number of parameters increases at the same rate as the sample size. For example, when analyzing a matched one-to-one design via the fully stratified likelihood in (3) using a logistic regression model containing one dichotomous covariate and the $K - 1$ design variables for strata, it can be shown (see [1, p. 250]) that the bias in the estimate of the coefficient is 100%. If we regard the stratum-specific parameters as (nuisance) parameters whose values are neither of great interest to us nor are essential for the inferences required in the study, and we are willing to forgo their estimation, then we can create a conditional likelihood which will yield maximum likelihood estimators of the slope coefficients in the logistic regression model that are consistent and asymptotically normally distributed. The mathematical details of conditional likelihood analysis may be found in [2] (*see* **Likelihood**). We summarize its application to the matched design. Liang [6], in related work, considers a general approach to the analysis of highly stratified data.

The Conditional Logistic Regression Model

Suppose that there are K strata with n_{k1} cases (subjects with $y = 1$) and n_{k0} controls (subjects with $y = 0$) in stratum k, $k = 1, 2, \ldots, K$. The conditional likelihood for the kth stratum is obtained as the probability of the observed data conditional on the stratum total sample size (fixed by the sampling design) and the total number of cases, the **sufficient statistic** for the stratum-specific **nuisance parameter**. This probability is the ratio of the probability of the observed

outcome to the probability for all possible assignments of n_{k1} subjects with $y = 1$ and n_{k0} subjects with $y = 0$ to $n_k = n_{k0} + n_{k1}$ subjects. The number of possible assignments is the n_k choose n_{k1} combinations. Let the subscript j denote any one of these assignments. For any assignment we let subjects 1 to n_{k1} correspond to the subjects with $y = 1$ and subjects $n_{k1} + 1$ to n_k to the subjects with $y = 0$. This will be indexed by i for the observed data and by i_j for the jth possible assignment. The contribution to the conditional likelihood for the kth stratum is

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{k1}} \Pr(y_{ki} = 1|\mathbf{x}_{ki}) \prod_{i=n_{k1}+1}^{n_k} \Pr(y_{ki} = 0|\mathbf{x}_{ki})}{\sum_j \left[\begin{array}{l} \prod_{i_j=1}^{n_{k1}} \Pr(y_{ki_j} = 1|\mathbf{x}_{ki_j}) \\ \times \prod_{i_j=n_{k1}+1}^{n_k} \Pr(y_{ki_j} = 0|\mathbf{x}_{ki_j}) \end{array} \right]}, \quad (4)$$

where the summation over j in the denominator is over the n_k choose n_{k1} combinations. The full conditional likelihood is the product of the $l_k(\boldsymbol{\beta})$ over the K strata,

$$l(\boldsymbol{\beta}) = \prod_{k=1}^{K} l_k(\boldsymbol{\beta}). \quad (5)$$

If we substitute the logistic regression model with the logit defined in (1), $\pi(\mathbf{x}_{ki}) = \Pr(y_{ki} = 1|\mathbf{x}_{ki})$, into (4), then (5) simplifies to

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{k1}} \pi(\mathbf{x}_{ki}) \prod_{i=n_{k1}+1}^{n_k} [1 - \pi(\mathbf{x}_{ki})]}{\sum_j \left\{ \prod_{i_j=1}^{n_{k1}} \pi(\mathbf{x}_{ki_j}) \prod_{i_j=n_{k1}+1}^{n_k} [1 - \pi(\mathbf{x}_{ki_j})] \right\}}. \quad (6)$$

Since the terms of the form $\exp(\beta_0 + \alpha_k)/[1 + \exp(\beta_0 + \alpha_k + \mathbf{x}'_{ki}\boldsymbol{\beta})]$ appear equally in both the numerator and denominator of (6) they cancel out, and (6) simplifies to

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{k1}} \exp(\boldsymbol{\beta}'\mathbf{x}_{ki})}{\sum_j \left(\prod_{i_j=1}^{n_{k1}} \exp(\boldsymbol{\beta}'\mathbf{x}_{ki_j}) \right)}, \quad (7)$$

which depends only on the unknown parameter vector $\boldsymbol{\beta}$. The conditional maximum likelihood estimator for $\boldsymbol{\beta}$ is that value which maximizes (5) when the expression in (7) is used for $l_k(\boldsymbol{\beta})$. Most software packages performing logistic regression have the capability to fit this conditional logistic regression model (*see* **Software, Biostatistical**).

The argument leading to expression (7) is more complicated for a case–control study and requires assumptions about sampling of cases and controls and applications of **Bayes' theorem**. The details will not be presented here but may be found in [3, Chapters 6 and 7].

One must always keep in mind when using the conditional likelihood in (7) that it was obtained by beginning with the usual logistic regression model. Thus, one still interprets the coefficients as "log-odds ratios". The original logistic regression model (1) or (2) tends to become lost in the arithmetic process of re-expressing the likelihood in (7). This point can be especially confusing to those analyzing data from a one-to-one matched case–control study.

The one-to-one matched design is probably the most frequent example of the use of a conditional logistic regression model. We show how one may analyze this design using standard logistic regression software, since not all packages have the capability to perform conditional logistic regression. More general software must be used in other matched designs and in the general highly stratified setting.

Logistic Regression Analysis for the One-to-One Matched Study

In the one-to-one matched study there are two subjects within each stratum. To simplify the notation, let $\mathbf{x}_{k1}$ denote the covariate vector for the case and $\mathbf{x}_{k0}$ the covariate vector for the control in the kth stratum. Using this notation, the conditional likelihood, (7), for the kth stratum is

$$l_k(\beta) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{k1})}{\exp(\boldsymbol{\beta}'\mathbf{x}_{k1}) + \exp(\boldsymbol{\beta}'\mathbf{x}_{k0})}. \quad (8)$$

Further simplification is obtained by dividing the numerator and denominator of (8) by $\exp(\boldsymbol{\beta}'\mathbf{x}_{k0})$, yielding

$$l_k(\beta) = \frac{\exp[\boldsymbol{\beta}'(\mathbf{x}_{k1} - \mathbf{x}_{k0})]}{1 + \exp[\boldsymbol{\beta}'(\mathbf{x}_{k1} - \mathbf{x}_{k0})]}. \quad (9)$$

The expression on the right-hand side of (9) is identical to a logistic regression model with the constant term set equal to zero, $\beta_0 = 0$, and covariate vector equal to the value of the case minus the value of the control, $\mathbf{x}_k^* = \mathbf{x}_{k1} - \mathbf{x}_{k0}$. This algebraic simplification allows one to use standard logistic regression software to compute the conditional maximum likelihood estimators of the coefficients and their standard errors. To accomplish this, one performs the following data modifications: define the sample size as the number of case–control pairs, compute the difference vector $\mathbf{x}_k^*$, compute a *pseudo*-response variable equal to 1, $y_k^* = 1$, and exclude the constant term from the model, e.g. force its value to be equal to zero. Thus, from a computational point of view, the one-to-one matched design presents no new challenges.

We have found that in the process of creating the differences and setting the "outcome" equal to 1, one can lose sight of the model. It is important to distinguish between the logistic regression model being fit to the data and the computational manipulations required to fit this model with standard logistic regression software. The process is less confusing if one focuses on the logistic regression model first and then considers the computations needed to obtain the parameter estimates. A few examples should help to illustrate this point.

Suppose we have a dichotomous independent variable coded zero or one. This variable is correctly modeled via a single coefficient in the logit, irrespective of whether we enter the variable via a design variable or treat it as continuous. The difference variable which we obtain by subtracting the value of the case from that of the control may take on one of three possible values: (−1, 0 or 1). If we had mistakenly thought of the difference variable as being the actual data, then we would have incorrectly modeled the variable by including two design variables in the model. The correct method is to create a difference variable and treat it as if it were continuous.

As a second example, suppose we have a variable such as race, coded at three levels. To model this variable correctly in the one-to-one matched design, we create, for each case and control in a pair, the values of the two design variables representing race. We compute the difference between these two design variables for the case and control and model each of these differences as if it were continuous. The same process is followed for any categorical scaled covariate. Note that the computer software may not recognize the differences in design variables as being created from the same variable, so one has to be sure that all design variables are included in the model. Another point to keep in mind is that differences between variables used to form strata are equal to zero for all strata and thus will not be useful as main effects. However, one may include interaction terms between stratification variables and other covariates, because differences in these interaction variables will likely not be zero.

In summary, the conceptual process for modeling matched or highly stratified data is identical to that of the usual logistic regression model. If one develops the modeling strategy for highly stratified data as if one had unstratified data, and then uses the conditional likelihood, then one will always be proceeding correctly.

Examples of the Use of the Conditional Logistic Regression Model

For illustrative purposes we use a small one-to-one matched data set obtained from a study of factors associated with the birth of a low birthweight baby (less than 2500 g). These data are in [3, Appendix 3]. These data, as well as the other data sets used in [3], may be obtained in the logistic regression menu at **internet** address `http://www-unix.oit.umass.edu/~statdata`. A one-to-one matched data set was obtained from an unmatched study of 189 births of which 59 were low weight. The matched data were obtained by randomly selecting, for each woman who gave birth to a low birthweight baby, a mother of the same age who did not give birth to a low birthweight baby. For three of the young mothers (age less than 17) it was not possible to identify a match since there were no mothers of normal weight babies of that age. The data consist of 56 age-matched case–control pairs. Variables selected for use in this example are a prior pre-term delivery (ptd, 1 = yes, 0 = no), smoking status (during pregnancy) of the mother (smoke, 1 = yes, 0 = no), history of hypertension (ht, 1 = yes, 0 = no), presence of uterine irritability (ui, 1 = yes, 0 = no), and the weight of the mother at the last menstrual period (lwt, pounds).

In ordinary logistic regression the coefficient for a model containing only one dichotomous variable is equal to the log of the cross-product ratio (odds ratio) from the **two-by-two table** of outcome by

the dichotomous variable. The same result is true when the conditional logistic model is used with a one-to-one matched study and the model contains a single dichotomous variable. The estimator of the odds ratio in a one-to-one matched study is the ratio of the frequencies of the discordant pairs. These are the frequencies in the off main diagonal cells of a 2×2 table cross-classifying the dichotomous variable for the case by the control. For example, consider the smoking status of the mother. The 2×2 table is shown in Table 1 and the results from fitting the conditional logistic regression model containing this variable are shown in Table 2. The odds ratio computed from Table 1 is $\hat{\psi} = 22/8 = 2.75$ and its log is $\ln \hat{\psi} = 1.012$. The results presented in Table 2 show that the coefficient for smoke is identically equal to the log of the odds ratio from Table 1. A confidence interval for the odds ratio may be obtained by exponentiating the end points of the confidence interval for the coefficient shown in Table 2. The resulting interval is (1.22, 6.18) indicating that, in these data, smoking during pregnancy is a risk factor for giving birth to a low birthweight baby. The significance of the coefficient may be tested using the Wald statistic (*see* **Likelihood**), labeled as z in Table 2, and whose two-tailed ***P* value** is 0.014. The appropriateness of both the confidence interval and test depend on an assumption that the sample size, 56 in this case, is large enough to employ the large-sample distributional properties (normality) of maximum likelihood estimators.

If one did not have available software specifically to perform conditional logistic regression, then the previously described method of creating difference variables could be used before beginning full-scale modeling of the data. This technique is not as important as it once was as most of the commonly available packages either have specific conditional logistic regression routines, or methods for adapting other routines are explained in their manuals. Again we wish to reinforce the point that the method of creating difference variables will only work for the one-to-one matched study. Any other design must be modeled through specific conditional logistic regression software.

We present in Table 3 the results of fitting a more complex model. The purpose of this model is to illustrate the use and interpretation of results from a multivariable conditional logistic regression model. See [3, Chapter 7] for a discussion of the issues involved in developing a model within the context of the current example and conditional logistic regression.

We obtain estimates of the odds ratios and their confidence intervals by exponentiating the estimated coefficients and end points of their confidence intervals in Table 3. These are shown in Table 4. The odds ratio and confidence interval presented for the weight of the mother at the last menstrual period is for a 10 pound increase in weight. The results for last menstrual period (lwt) are obtained from Table 3 by multiplying the coefficient and end points of the confidence interval by 10 before exponentiating. This is done since lwt is measured in pounds, and an odds ratio for a one pound weight difference is likely not to be clinically meaningful.

The odds ratios in Table 4 suggest an important increase in risk of delivering of a low birthweight baby for prior pre-term deliveries, smoking during pregnancy, presence of hypertension, and presence of uterine irritability. The odds ratio for the weight

Table 1 Cross-classification of the smoking status of the case by the control

	Control		
Case	No	Yes	Total
No	18	8	26
Yes	22	8	30
Total	40	16	56

Table 2 Results from fitting a conditional logistic regression model containing the dichotomous variable, smoking status of the mother

Variable	Coeff.	Std. error	z	P	95% CIE
Smoke	1.012	0.413	2.45	0.014	(0.202, 1.821)

Table 3 Results from fitting a conditional logistic regression model containing prior pre-term delivery, smoking status of the mother, presence of hypertension, presence of uterine irritability, and the weight of the mother at the last menstrual period to 56 matched pairs

Variable	Coeff.	Std. error	z	P	95% CIE
ptd	1.671	0.747	2.24	0.025	(0.207, 3.135)
smoke	1.480	0.562	2.63	0.009	(0.378, 2.582)
ht	2.330	1.003	2.32	0.020	(0.364, 4.296)
ui	1.345	0.694	1.94	0.052	(−0.015, 2.705)
lwt	−0.015	0.008	−1.88	0.060	(−0.031, 0.001)

Table 4 Estimated odds ratios and 95% confidence intervals for prior pre-term deliveries, smoking status of the mother, presence of hypertension, presence of uterine irritability, and the weight of the mother at the last menstrual period (10 lb increase)

Variable	Odds ratio	95% CIE
ptd	5.32	(1.23, 22.99)
smoke	4.39	(1.46, 13.22)
ht	10.28	(1.44, 73.41)
ui	3.84	(0.99, 14.95)
lwt	0.86	(0.73, 1.01)

of the mother at the last menstrual period suggests an approximate 14% decrease in risk per 10 pound increase in weight. This interpretation assumes that the logit is linear in lwt. One should always check the scale of all continuous variables in any regression model. We did this using a method based on design variables for the quartiles of lwt (see [3, p. 194]), which supported the linearity assumption for lwt.

The confidence interval estimates in Table 4 are quite wide for the dichotomous variables. This instability is due to the fact that the variance estimator is inversely related to the number of discordant pairs. The analysis presented in Tables 2 and 3 is based on 56 pairs and the numbers of discordant pairs are 19, 30, 10, and 16, respectively, for the dichotomous variables. The widths of the confidence intervals in Table 4 are a result of the relatively few discordant pairs. This points out an important consideration that must be kept in mind at the design stage of a study. The gain in precision obtained from matching and using conditional logistic regression may be offset by a loss owing to few discordant pairs for dichotomous covariates. In general, the variance estimator of the slope coefficient is a function of how different the subjects with $y = 1$ are from those with $y = 0$ within each stratum.

Likelihood ratio tests may be used for model testing and refinement in a manner similar to that discussed in the article on logistic regression. In the case of conditional logistic regression the likelihood for model zero, "the no data model", is obtained by setting the coefficient vector equal to zero in (7). This model is essentially a coin toss with stratum specific probability $\Pr(Y_{kj} = 1) = n_{k1}/n_k$.

Application of the conditional logistic regression model to other, more complicated, matched or highly stratified designs is, for all intents and purposes, identical to the one-to-one matched study discussed. The essential point to keep in mind is that one uses and interprets the estimated coefficients in a manner identical to ordinary logistic regression. Although not illustrated in the example, because of relatively few matched pairs, one may use matching or stratification variables to form interactions with variables in the model but one may not include them as main effect terms. Much of the content of this article is based on [3].

References

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. 1. *The Analysis of Case–Control Studies*. Oxford University Press, New York.

[2] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, New York.

[3] Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd Ed. Wiley, New York.

[4] Kelsey, J.L., Thompson, W.D. & Evans, A.S. (1986). *Methods in Observational Epidemiology*. Oxford University Press, New York.

[5] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Van Nostrand Reinhold, New York.

[6] Liang, K.Y. (1987). Extended Mantel–Haenszel estimating procedure for multivariate logistic regressions, *Biometrics* **43**, 289–300.

[7] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown & Company, Boston.

[8] Schlesselman, J.J. (1982). *Case–Control Studies*. Oxford University Press, New York.

(*See also* **Binary Data**; **Correlated Binary Data**)

DAVID W. HOSMER JR &
STANLEY LEMESHOW

Logistic Regression

The goal of a logistic regression analysis is to find the best fitting and most parsimonious, yet biologically reasonable, model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. What distinguishes the logistic regression model from the **linear regression** model is that the outcome variable in logistic regression is categorical and most usually *binary* or *dichotomous* (*see* **Binary Data**).

In any regression problem the key quantity is the mean value of the outcome variable, given the value of the independent variable. This quantity is called the *conditional mean* and will be expressed as $\mathrm{E}(Y|x)$, where Y denotes the outcome variable and x denotes a value of the independent variable. In linear regression we assume that this mean may be expressed as an equation linear in x (or some transformation of x or Y), such as

$$\mathrm{E}(Y|x) = \beta_0 + \beta_1 x.$$

This expression implies that it is possible for $\mathrm{E}(Y|x)$ to take on any value as x ranges between $-\infty$ and $+\infty$.

Many distribution functions have been proposed for use in the analysis of a dichotomous outcome variable. Cox & Snell [2] discuss some of these. There are two primary reasons for choosing the logistic distribution. These are: (i) from a mathematical point of view it is an extremely flexible and easily used function, and (ii) it lends itself to a biologically meaningful interpretation.

To simplify notation, let $\pi(x) = \mathrm{E}(Y|x)$ represent the conditional mean of Y given x. The logistic regression model can be expressed as

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \qquad (1)$$

The *logit transformation*, defined in terms of $\pi(x)$, is as follows:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x. \qquad (2)$$

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. The logit, $g(x)$, is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$ depending on the range of x.

The second important difference between the linear and logistic regression models concerns the conditional distribution of the outcome variable. In the linear regression model we assume that an observation of the outcome variable may be expressed as $y = \mathrm{E}(Y|x) + \varepsilon$. The quantity ε is called the *error* and expresses an observation's deviation from the conditional mean. The most common assumption is that ε follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. It follows that the conditional distribution of the outcome variable given x is normal with mean $\mathrm{E}(Y|x)$, and a variance that is constant. This is not the case with a dichotomous outcome variable. In this situation we may express the value of the outcome variable given x as $y = \pi(x) + \varepsilon$. Here the quantity ε may assume one of two possible values. If $y = 1$, then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$, then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus, ε has a distribution with mean zero and variance equal to $\pi(x)[1 - \pi(x)]$. That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean, $\pi(x)$.

Fitting the Logistic Regression Model

Suppose we have a sample of n independent observations of the pair $(x_i, y_i), i = 1, 2, \ldots, n$, where y_i denotes the value of a dichotomous outcome variable and x_i is the value of the independent variable for the ith subject. Furthermore, assume that the outcome variable has been coded as 0 or 1 representing the absence or presence of the characteristic, respectively. To fit the logistic regression model (1) to a set of data requires that we estimate the values of β_0 and β_1, the unknown parameters.

In linear regression the method used most often to estimate unknown parameters is **least squares**. In that method we choose those values of β_0 and β_1 that minimize the sum of squared deviations of the observed values of Y from the predicted values based upon the model. Under the usual assumptions for linear regression the least squares method yields estimators with a number of desirable statistical

properties. Unfortunately, when the least squares method is applied to a model with a dichotomous outcome the estimators no longer have these same properties.

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is **maximum likelihood**. This is the method used to estimate the logistic regression parameters. In a very general sense the maximum likelihood method yields values for the unknown parameters that maximize the probability of obtaining the observed set of data. To apply this method we must first construct a function called the *likelihood function* (*see* **Likelihood**). This function expresses the probability of the observed data as a function of the unknown parameters. The *maximum likelihood estimators* of these parameters are chosen to be those values that maximize this function. Thus, the resulting estimators are those that agree most closely with the observed data.

If Y is coded as 0 or 1, then the expression for $\pi(x)$ given in (1) provides (for an arbitrary value of $\boldsymbol{\beta}' = (\beta_0, \beta_1)$, the vector of parameters) the conditional probability that Y is equal to 1 given x. This will be denoted $\Pr(Y = 1|x)$. It follows that the quantity $1 - \pi(x)$ gives the conditional probability that Y is equal to zero given x, $\Pr(Y = 0|x)$. Thus, for those pairs (x_i, y_i), where $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$, where the quantity $\pi(x_i)$ denotes the value of $\pi(x)$ computed at x_i. A convenient way to express the contribution to the likelihood function for the pair (x_i, y_i) is through the term

$$\xi(x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{3}$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in (3) as follows:

$$l(\beta) = \prod_{i=1}^{n} \xi(x_i). \tag{4}$$

The principle of maximum likelihood states that we use as our estimate of β the value that maximizes the expression in (4). However, it is easier mathematically to work with the log of (4). This expression, the *log likelihood*, is defined as

$$\begin{aligned} L(\beta) &= \ln[l(\beta)] \\ &= \sum \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \end{aligned} \tag{5}$$

To find the value of β that maximizes $L(\beta)$ we differentiate $L(\beta)$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. These equations are as follows:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{6}$$

and

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0, \tag{7}$$

and are called the *likelihood equations*.

In linear regression, the likelihood equations, obtained by differentiating the sum of squared deviations function with respect to β, are linear in the unknown parameters, and thus are easily solved. For logistic regression the expressions in (6) and (7) are nonlinear in β_0 and β_1, and thus require special methods for their solution. These methods are iterative in nature and have been programmed into available logistic regression software. McCullagh & Nelder [6] discuss the iterative methods used by most programs. In particular, they show that the solution to (6) and (7) may be obtained using a generalized weighted least squares procedure.

The value of β given by the solution to (6) and (7) is called the maximum likelihood estimate, denoted as $\hat{\boldsymbol{\beta}}$. Similarly, $\hat{\pi}(x_i)$ is the maximum likelihood estimate of $\pi(x_i)$. This quantity provides an estimate of the conditional probability that Y is equal to 1, given that x is equal to x_i. As such, it represents the fitted or predicted value for the logistic regression model. An interesting consequence of (6) is that

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i).$$

That is, the sum of the observed values of y is equal to the sum of the predicted (expected) values.

After estimating the coefficients, it is standard practice to assess the significance of the variables in the model. This usually involves testing a statistical hypothesis to determine whether the independent

variables in the model are "significantly" related to the outcome variable. One approach to testing for the significance of the coefficient of a variable in any model relates to the following question. *Does the model that includes the variable in question tell us more about the outcome (or response) variable than does a model that does not include that variable?* This question is answered by comparing the observed values of the response variable with those predicted by each of two models; the first with and the second without the variable in question. The mathematical function used to compare the observed and predicted values depends on the particular problem. If the predicted values with the variable in the model are better, or more accurate in some sense, than when the variable is not in the model, then we feel that the variable in question is "significant". It is important to note that we are not considering the question of whether the predicted values are an accurate representation of the observed values in an absolute sense (this would be called *goodness of fit*). Instead, our question is posed in a relative sense.

For the purposes of assessing the significance of an independent variable we compute the value of the following statistic:

$$G = -2\ln\left(\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right). \quad (8)$$

Under the hypothesis that β_1 is equal to zero, the statistic G will follow a chi-square distribution with one degree of freedom. The calculation of the log likelihood and this generalized **likelihood ratio test** are standard features of any good logistic regression package. This makes it possible to check for the significance of the addition of new terms to the model as a matter of routine. In the simple case of a single independent variable, we can first fit a model containing only the constant term. We can then fit a model containing the independent variable along with the constant. This gives rise to a new log likelihood. The likelihood ratio test is obtained by multiplying the difference between the log likelihoods of the two models by -2.

Another test that is often carried out is the Wald test, which is obtained by comparing the maximum likelihood estimate of the slope parameter, $\hat{\beta}_1$, with an estimate of its standard error (*see* **Likelihood**). The resulting ratio

$$W = \frac{\hat{\beta}_1}{\widehat{\text{se}}(\hat{\beta}_1)},$$

under the hypothesis that $\beta_1 = 0$, follows a standard normal distribution. Standard errors of the estimated parameters are routinely printed out by computer software. Hauck & Donner [3] examined the performance of the Wald test and found that it behaved in an aberrant manner, often failing to reject when the coefficient was significant. They recommended that the likelihood ratio test be used. Jennings [5] has also looked at the adequacy of inferences in logistic regression based on Wald statistics. His conclusions are similar to those of Hauck & Donner.

Both the likelihood ratio test, G, and the Wald test, W, require the computation of the maximum likelihood estimate for β_1. For a single variable this is not a difficult or costly computational task. However, for large data sets with many variables, the iterative computation needed to obtain the maximum likelihood estimates can be considerable.

The logistic regression model may be used with matched study designs. Fitting **conditional logistic regression** models requires modifications, which are not discussed here. The reader interested in the conditional logistic regression model may find details in [4, Chapter 7].

The Multiple Logistic Regression Model

Consider a collection of p independent variables which will be denoted by the vector $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. Assume for the moment that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present be denoted by $\Pr(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$. Then the logit of the multiple logistic regression model is given by

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad (9)$$

in which case

$$\pi(x) = \frac{\exp[g(\mathbf{x})]}{1 + \exp[g(\mathbf{x})]}. \quad (10)$$

If some of the independent variables are discrete, nominal scaled variables (*see* **Nominal Data**) such as race, sex, treatment group, and so forth, then it is inappropriate to include them in the model as if they

were interval scaled. In this situation a collection of *design variables* (or **dummy variables**) should be used. Most logistic regression software will generate the design variables, and some programs have a choice of several different methods.

In general, if a nominal scaled variable has k possible values, then $k-1$ design variables will be needed. Suppose, for example, that the jth independent variable, x_j has k_j levels. The $k_j - 1$ design variables will be denoted as D_{ju} and the coefficients for these design variables will be denoted as $\beta_{ju}, u = 1, 2, \ldots, k_j - 1$. Thus, the logit for a model with p variables and the jth variable being discrete is

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} + \beta_p x_p.$$

Fitting the Multiple Logistic Regression Model

Assume that we have a sample of n independent observations of the pair $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, n$. As in the univariate case, fitting the model requires that we obtain estimates of the vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$. The method of estimation used in the multivariate case is the same as in the univariate situation, i.e. maximum likelihood. The likelihood function is nearly identical to that given in (4), with the only change being that $\pi(\mathbf{x})$ is now defined as in (10). There are $p + 1$ likelihood equations which are obtained by differentiating the log likelihood function with respect to the $p + 1$ coefficients. The likelihood equations that result may be expressed as follows:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(\mathbf{x}_i)] = 0,$$

for $j = 1, 2, \ldots, p$.

As in the univariate model, the solution of the likelihood equations requires special purpose software which may be found in many packaged programs. Let $\hat{\beta}$ denote the solution to these equations. Thus, the fitted values for the multiple logistic regression model are $\hat{\pi}(\mathbf{x}_i)$, the value of the expression in (13) computed using $\hat{\beta}$ and $\mathbf{x}_i$.

Before proceeding further we present an example that illustrates the formulation of a multiple logistic regression model and the estimation of its coefficients.

Example

To provide an example of fitting a multiple logistic regression model, consider the data for the low birth weight study described in Appendix 1 of Hosmer & Lemeshow [4]. The code sheet for the data set is given in Table 1.

The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 g). In this study data were collected on 189 women; $n_1 = 59$ of them

Table 1 Code sheet for the variables in the low birth weight data set

Variable	Abbreviation
Identification code	ID
Low birth weight (0 = birth weight ≥2500 g, 1 = birth weight <2500 g)	LOW
Age of the mother in years	AGE
Weight in pounds at the last menstrual period	LWT
Race (1 = white, 2 = black, 3 = other)	RACE
Smoking status during pregnancy (1 = yes, 0 = no)	SMOKE
History of premature labor (0 = none, 1 = one, etc.)	PTL
History of hypertension (1 = yes, 0 = no)	HT
Presence of uterine irritability (1 = yes, 0 = no)	UI
Number of physician visits during the first trimester (0 = none, 1 = one, 2 = two, etc.)	FTV
Birth weight (g)	BWT

delivered low birth weight babies and $n_0 = 130$ delivered normal birth weight babies. In this example the variable race has been recoded using the two design variables shown in Table 2. FTV was recoded to 0 = some, 1 = none, and PTL was recoded to 0 = none, 1 = one or more. The two newly coded variables are called FTV01 and PTL01.

The results of fitting the logistic regression model to these data are given in Table 3.

In Table 3 the estimated coefficients for the two design variables for race are indicated in the lines denoted by "RACE 1" and "RACE 2". The estimated logit is given by

$$\begin{aligned}\hat{g}(\mathbf{x}) = {} & 0.545 - 0.035 \times \text{AGE} - 0.015 \times \text{LWT} \\ & + 0.815 \times \text{SMOKE} + 1.824 \times \text{HT} + 0.702 \\ & \times \text{UI} + 1.202 \times \text{RACE 1} + 0.773 \times \text{RACE 2} \\ & + 0.121 \times \text{FTV01} + 1.237 \times \text{PTL01}.\end{aligned}$$

The fitted values are obtained using the estimated logit, $\hat{g}(\mathbf{x})$, as in (10).

Table 2 Coding of design variables for RACE

	Design variable	
RACE	RACE 1	RACE 2
White	0	0
Black	1	0
Other	0	1

Testing for the Significance of the Model

Once we have fit a particular multiple (multivariate) logistic regression model, we begin the process of assessment of the model. The first step in this process is usually to assess the significance of the variables in the model. The likelihood ratio test for overall significance of the p coefficients for the independent variables in the model is performed based on the statistic G given in (8). The only difference is that the fitted values, $\hat{\pi}$, under the model are based on the vector containing $p + 1$ parameters, $\hat{\boldsymbol{\beta}}$. Under the null hypothesis that the p "slope" coefficients for the covariates in the model are equal to zero, the distribution of G is **chi-square** with p **degrees of freedom**.

As an example, consider the fitted model whose estimated coefficients are given in Table 3. For that model the value of the log likelihood is $L = -98.36$. A second model, fit with the constant term only, yields $L = -117.336$. Hence $G = -2[(-117.34) - (-98.36)] = 37.94$ and the ***P* value** for the test is $\Pr[\chi^2(9) > 37.94] < 0.0001$ (see Table 3). Rejection of the **null hypothesis** (that all of the coefficients are simultaneously equal to zero) has an interpretation analogous to that in multiple linear regression; we may conclude that at least one, and perhaps all p coefficients are different from zero.

Before concluding that any or all of the coefficients are nonzero, we may wish to look at the univariate Wald test statistics, $W_j = \hat{\beta}_j / \widehat{\text{se}}(\hat{\beta}_j)$. These are given in the fourth column (labeled z) in Table 3.

Table 3 Estimated coefficients for a multiple logistic regression model using all variables from the low birth weight data set

Logit estimates — Number of obs. = 189
$\chi^2(9) = 37.94$
Prob > $\chi^2 = 0.0000$
Log likelihood = −98.36

Variable	Coeff.	Std. error	z	Pr > |z|	[95% conf.	interval]
AGE	−0.035	0.039	−0.920	0.357	−0.111	0.040
LWT	−0.015	0.007	−2.114	0.035	−0.029	−0.001
SMOKE	0.815	0.420	1.939	0.053	−0.009	1.639
HT	1.824	0.705	2.586	0.010	0.441	3.206
UI	0.702	0.465	1.511	0.131	−0.208	1.613
RACE 1	1.202	0.534	2.253	0.024	0.156	2.248
RACE 2	0.773	0.460	1.681	0.093	−0.128	1.674
FTV01	0.121	0.376	0.323	0.746	−0.615	0.858
PTL01	1.237	0.466	2.654	0.008	0.323	2.148
cons	0.545	1.266	0.430	0.667	−1.937	3.027

Under the hypothesis that an individual coefficient is zero, these statistics will follow the **standard normal** distribution. Thus, the value of these statistics may give us an indication of which of the variables in the model may or may not be significant. If we use a critical value of 2, which leads to an approximate level of significance (two-tailed) of 0.05, then we would conclude that the variables LWT, SMOKE, HT, PTL01 and possibly RACE are significant, while AGE, UI, and FTV01 are not significant.

Considering that the overall goal is to obtain the best fitting model while minimizing the number of parameters, the next logical step is to fit a reduced model, containing only those variables thought to be significant, and compare it with the full model containing all the variables. The results of fitting the reduced model are given in Table 4.

The difference between the two models is the exclusion of the variables AGE, UI, and FTV01 from the full model. The likelihood ratio test comparing these two models is obtained using the definition of G given in (8). It has a distribution that is chi-square with three degrees of freedom under the hypothesis that the coefficients for the variables excluded are equal to zero. The value of the test statistic comparing the models in Tables 3 and 4 is $G = -2[(-100.24) - (-98.36)] = 3.76$ which, with three degrees of freedom, has a P value of $P[\chi^2(3) > 3.76] = 0.2886$. Since the P value is large, exceeding 0.05, we conclude that the reduced model is as good as the full model. Thus there is no advantage to including AGE, UI, and FTV01 in the model. However, we must not base our models entirely on tests of statistical significance. Numerous other considerations should influence our decision to include or exclude variables from a model.

Interpretation of the Coefficients of the Logistic Regression Model

After fitting a model the emphasis shifts from the computation and assessment of significance of estimated coefficients to interpretation of their values. The interpretation of any fitted model requires that we can draw practical inferences from the estimated coefficients in the model. The question addressed is: *What do the estimated coefficients in the model tell us about the research questions that motivated the study?* For most models this involves the estimated coefficients for the independent variables in the model. The estimated coefficients for the independent variables represent the slope or rate of change of a function of the dependent variable per unit of change in the independent variable. Thus, interpretation involves two issues: (i) determining the functional relationship between the dependent variable and the independent variable, and (ii) appropriately defining the unit of change for the independent variable.

For a linear regression model we recall that the slope coefficient, β_1, is equal to the difference between the value of the dependent variable at $x + 1$ and the value of the dependent variable at x, for any value of x. In the logistic regression model $\beta_1 = g(x + 1) - g(x)$. That is, the slope coefficient represents the change in the logit for a change of one unit in the independent variable x. Proper interpretation of the coefficient in a logistic regression model depends on being able to place meaning

Table 4 Estimated coefficients for a multiple logistic regression model using the variables LWT, SMOKE, HT, PTL01 and RACE from the low birth weight data set

Logit estimates — Number of obs. = 189; $\chi^2(6)$ = 34.19; Prob > χ^2 = 0.0000

Log likelihood = 100.24

Variable	Coeff.	Std. error	z	Pr > |z|	[95% conf.	interval]
LWT	−0.017	0.007	−2.407	0.016	−0.030	−0.003
SMOKE	0.876	0.401	2.186	0.029	0.091	1.661
HT	1.767	0.708	2.495	0.013	0.379	3.156
RACE 1	1.264	0.529	2.387	0.017	0.226	2.301
RACE 2	0.864	0.435	1.986	0.047	0.011	1.717
PTL01	1.231	0.446	2.759	0.006	0.357	2.106
cons	0.095	0.957	0.099	0.921	−1.781	1.970

on the difference between two logits. Consider the interpretation of the coefficients for a univariate logistic regression model for each of the possible measurement scales of the independent variable.

Dichotomous Independent Variable

Assume that x is coded as either 0 or 1. Under this model there are two values of $\pi(x)$ and equivalently two values of $1 - \pi(x)$. These values may be conveniently displayed in a **2 x 2 table**, as shown in Table 5.

The **odds** of the outcome being present among individuals with $x = 1$ is defined as $\pi(1)/[1 - \pi(1)]$. Similarly, the odds of the outcome being present among individuals with $x = 0$ is defined as $\pi(0)/[1 - \pi(0)]$. The **odds ratio**, denoted by ψ, is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$, and is given by

$$\psi = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}. \tag{11}$$

The log of the odds ratio, termed log odds ratio, or *log odds*, is

$$\ln(\psi) = \ln\left\{\frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}\right\} = g(1) - g(0),$$

which is the *logit difference*, where the log of the odds is called the logit and, in this example, these are

$$g(1) = \ln\left\{\frac{\pi(1)}{1 - \pi(1)}\right\}$$

and

$$g(0) = \ln\left\{\frac{\pi(0)}{1 - \pi(0)}\right\}.$$

Using the expressions for the logistic regression model shown in Table 5 the odds ratio is

$$\psi = \frac{\left(\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}\right)\left(\frac{1}{1 + \exp(\beta_0)}\right)}{\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right)\left(\frac{1}{1 + \exp(\beta_0 + \beta_1)}\right)}$$
$$= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

Hence, for logistic regression with a dichotomous independent variable

$$\psi = \exp(\beta_1), \tag{12}$$

and the logit difference, or log odds, is

$$\ln(\psi) = \ln[\exp(\beta_1)] = \beta_1.$$

This fact concerning the interpretability of the coefficients is the fundamental reason why logistic regression has proven such a powerful analytic tool for epidemiologic research. A confidence interval (CI) estimate for the odds ratio is obtained by first calculating the endpoints of a **confidence interval** for the coefficient β_1, and then exponentiating these values. In general, the endpoints are given by

$$\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2} \times \widehat{\text{se}}(\hat{\beta}_1)\right].$$

Because of the importance of the odds ratio as a measure of association, point and interval estimates are often found in additional columns in tables presenting the results of a logistic regression analysis.

In the previous discussion we noted that the estimate of the odds ratio was $\hat{\psi} = \exp(\hat{\beta}_1)$. This is correct when the independent variable has been

Table 5 Values of the logistic regression model when the independent variable is dichotomous

			Independent variable X	
			$x = 1$	$x = 0$
Outcome variable	Y	$y = 1$	$\pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$\pi(0) = \frac{\exp\beta_0}{1 + \exp\beta_0}$
		$y = 0$	$1 - \pi(1) = \frac{1}{1 + \exp(\beta_0 + \beta_1)}$	$1 - \pi(0) = \frac{1}{1 + \exp\beta_0}$
		Total	1.0	1.0

coded as 0 or 1. This type of coding is called "reference cell" coding. Other coding could be used. For example, the variable may be coded as -1 or $+1$. This type of coding is termed "deviation from means" coding. Evaluation of the logit difference shows that the odds ratio is calculated as $\hat{\psi} = \exp(2\hat{\beta}_1)$ and if an investigator were simply to exponentiate the coefficient from the computer output of a logistic regression analysis, the wrong estimate of the odds ratio would be obtained. Close attention should be paid to the method used to code design variables.

The method of coding also influences the calculation of the endpoints of the confidence interval. With deviation from means coding, the estimated standard error needed for confidence interval estimation is $\widehat{\text{se}}(2\hat{\beta}_1)$, which is $2 \times \widehat{\text{se}}(\hat{\beta}_1)$. Thus the endpoints of the confidence interval are

$$\exp\left[2\hat{\beta}_1 + z_{1-\alpha/2} \times 2 \times \widehat{\text{se}}(\hat{\beta}_1)\right].$$

In summary, for a dichotomous variable the parameter of interest is the odds ratio. An estimate of this parameter may be obtained from the estimated logistic regression coefficient, regardless of how the variable is coded or scaled. This relationship between the logistic regression coefficient and the odds ratio provides the foundation for our interpretation of all logistic regression results.

Polytomous Independent Variable

Suppose that instead of two categories the independent variable has $k > 2$ distinct values (*see* **Polytomous Data**). For example, we may have variables that denote the county of residence within a state, the clinic used for primary health care within a city, or race. Each of these variables has a fixed number of discrete outcomes and the scale of measurement is nominal.

Suppose that in a study of coronary heart disease (CHD) the variable RACE is coded at four levels, and that the cross-classification of RACE by CHD status yields the data presented in Table 6. These data are hypothetical and have been formulated for ease of computation. The extension to a situation where the variable has more than four levels is not conceptually different, so all the examples in this section use $k = 4$.

At the bottom of Table 6 the odds ratio is given for each race, using white as the reference group. For example, for hispanic the estimated odds ratio is $(15 \times 20)/(5 \times 10) = 6.0$. The log of the odds ratios are given in the last row of Table 6. This display is typical of what is found in the literature when there is a perceived referent group to which the other groups are to be compared. These same estimates of the odds ratio may be obtained from a logistic regression program with an appropriate choice of design variables. The method for specifying the design variables involves setting all of them equal to zero for the reference group, and then setting a single design variable equal to one for each of the other groups. This is illustrated in Table 7.

Table 7 Specification of the design variables for RACE using white as the reference group

	Design variables		
RACE (code)	D_1	D_2	D_3
White (1)	0	0	0
Black (2)	1	0	0
Hispanic (3)	0	1	0
Other (4)	0	0	1

Table 6 Cross-classification of hypothetical data on RACE and CHD status for 100 subjects

CHD status	White	Black	Hispanic	Other	Total
Present	5	20	15	10	50
Absent	20	10	10	10	50
Total	25	30	25	20	100
Odds ratio ($\hat{\psi}$)	1.0	8.0	6.0	4.0	
95% CI		(2.3, 27.6)	(1.7, 21.3)	(1.1, 14.9)	
$\ln(\hat{\psi})$	0.0	2.08	1.79	1.39	

Table 8 Results of fitting the logistic regression model to the data in Table 6 using the design variables in Table 7

Variable	Coeff.	Std. error	z	$P > \|z\|$	[95% conf. interval]	
RACE 1	2.079	0.632	3.288	0.001	0.840	3.319
RACE 2	1.792	0.645	2.776	0.006	0.527	3.057
RACE 3	1.386	0.671	2.067	0.039	0.072	2.701
cons	−1.386	0.500	−2.773	0.006	−2.367	−0.406

Variable	Odds ratio	[95% conf. interval]	
RACE 1	8	2.32	27.63
RACE 2	6	1.69	21.26
RACE 3	4	1.07	14.90

Use of any logistic regression program with design variables coded as shown in Table 7 yields the estimated logistic regression coefficients given in Table 8.

A comparison of the estimated coefficients in Table 8 with the log odds in Table 6 shows that $\ln[\hat{\psi}(\text{black, white})] = \hat{\beta}_{11} = 2.079$, $\ln[\hat{\psi}(\text{hispanic, white})] = \hat{\beta}_{12} = 1.792$, and $\ln[\hat{\psi}(\text{other, white})] = \hat{\beta}_{13} = 1.386$.

In the univariate case the estimates of the standard errors found in the logistic regression output are identical to the estimates obtained using the cell frequencies from the contingency table. For example, the estimated standard error of the estimated coefficient for design variable (1), $\hat{\beta}_{11}$, is $0.6325 = (1/5 + 1/20 + 1/20 + 1/10)^{1/2}$. A derivation of this result appears in Bishop et al. [1].

Confidence limits for odds ratios may be obtained as follows:

$$\hat{\beta}_{ij} \pm z_{1-\alpha/2} \times \widehat{\text{se}}(\hat{\beta}_{ij}).$$

The corresponding limits for the odds ratio are obtained by exponentiating these limits as follows:

$$\exp[\hat{\beta}_{ij} \pm z_{1-\alpha/2} \times \widehat{\text{se}}(\hat{\beta}_{ij})].$$

Continuous Independent Variable

When a logistic regression model contains a continuous independent variable, interpretation of the estimated coefficient depends on how it is entered into the model and the particular units of the variable. For purposes of developing the method to interpret the coefficient for a continuous variable, we assume that the logit is linear in the variable.

Under the assumption that the logit is linear in the continuous covariate, x, the equation for the logit is $g(x) = \beta_0 + \beta_1 x$. It follows that the slope coefficient, β_1, gives the change in the log odds for an increase of "1" unit in x, i.e. $\beta_1 = g(x+1) - g(x)$ for any value of x. Most often the value of "1" will not be biologically very interesting. For example, an increase of 1 year in age or of 1 mmHg in systolic blood pressure may be too small to be considered important. A change of 10 years or 10 mmHg might be considered more useful. However, if the range of x is from zero to one, as might be the case for some created index, then a change of 1 is too large and a change of 0.01 may be more realistic. Hence, to provide a useful interpretation for continuous scaled covariates we need to develop a method for point and interval estimation for an arbitrary change of c units in the covariate.

The log odds for a change of c units in x is obtained from the logit difference $g(x+c) - g(x) = c\beta_1$ and the associated odds ratio is obtained by exponentiating this logit difference, $\psi(c) = \psi(x + c, x) = \exp(c\beta_1)$. An estimate may be obtained by replacing β_1 with its maximum likelihood estimate, $\hat{\beta}_1$. An estimate of the standard error needed for confidence interval estimation is obtained by multiplying the estimated standard error of $\hat{\beta}_1$ by c. Hence the endpoints of the $100(1-\alpha)\%$ CI estimate of $\psi(c)$ are

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2} c\widehat{\text{se}}(\hat{\beta}_1)].$$

Since both the point estimate and endpoints of the confidence interval depend on the choice of c, the

particular value of c should be clearly specified in all tables and calculations.

Multivariate Case

Often logistic regression analysis is used to *adjust statistically* the estimated effects of each variable in the model for differences in the distributions of and associations among the other independent variables. Applying this concept to a multiple logistic regression model, we may surmise that each estimated coefficient provides an estimate of the log odds adjusting for all other variables included in the model. The term confounder is used by epidemiologists to describe a covariate that is associated with both the outcome variable of interest and a primary independent variable or risk factor. When both associations are present the relationship between the risk factor and the outcome variable is said to be *confounded* (*see* **Confounding**). The procedure for adjusting for confounding is appropriate when there is no interaction.

If the association between the covariate and an outcome variable is the same within each level of the risk factor, then there is no interaction between the covariate and the risk factor. When interaction is present, the association between the risk factor and the outcome variable differs, or depends in some way on the level of the covariate. That is, the covariate modifies the effect of the risk factor (*see* **Effect Modification**). Epidemiologists use the term effect modifier to describe a variable that interacts with a risk factor.

The simplest and most commonly used model for including interaction is one in which the logit is also linear in the confounder for the second group, but with a different slope. Alternative models can be formulated which would allow for other than a linear relationship between the logit and the variables in the model within each group. In any model, interaction is incorporated by the inclusion of appropriate higher order terms.

An important step in the process of modeling a set of data is to determine whether or not there is evidence of interaction in the data. Tables 9 and 10 present the results of fitting a series of logistic regression models to two different sets of hypothetical data. The variables in each of the data sets are the same: SEX, AGE, and CHD. In addition to the estimated coefficients, the log likelihood for each model and minus twice the change (deviance) is given. Recall that minus twice the change in the log likelihood may be used to test for the significance of coefficients for variables added to the model. An interaction is added to the model by creating a variable that is equal to the product of the value of the sex and the value of age.

Examining the results in Table 9 we see that the estimated coefficient for the variable SEX changed from 1.535 in model 1 to 0.979 when AGE was added in model 2. Hence, there is clear evidence of a confounding effect owing to age. When the interaction term "SEX $\times$ AGE" is added in model 3 we see that the change in the deviance is only 0.52 which, when compared with the chi-square distribution with one degree of freedom, yields a P value of 0.47, which clearly is not significant. Note that the coefficient for sex changed from 0.979 to 0.481. This is not surprising since the inclusion of an interaction term, especially when it involves a continuous variable, will usually produce fairly marked changes in the estimated coefficients of dichotomous variables involved in the interaction. Thus, when an interaction term is present in the model we cannot assess confounding via the change in a coefficient. For these data we would prefer to use model 2 which suggests that age is a confounder but not an effect modifier.

The results in Table 10 show evidence of both confounding and interaction due to age. Comparing model 1 with model 2 we see that the coefficient for sex changes from 2.505 to 1.734. When the age by sex interaction is added to the model we see that the deviance is 4.06, which yields a P value

Table 9 Estimated logistic regression coefficients, log likelihood, and the likelihood ratio test statistic (G) for an example showing evidence of confounding but no interaction

Model	Constant	SEX	AGE	SEX $\times$ AGE	Log likelihood	G
1	−1.046	1.535			−61.86	
2	−7.142	0.979	0.167		−49.59	24.54
3	−6.103	0.481	0.139	0.059	−49.33	0.52

Table 10 Estimated logistic regression coefficients, log likelihood, and the likelihood ratio test statistic (G) for an example showing evidence of confounding and interaction

Model	Constant	SEX	AGE	SEX × AGE	Log likelihood	G
1	−0.847	2.505			−52.52	
2	−6.194	1.734	0.147		−46.79	11.46
3	−3.105	0.047	0.629	0.206	−44.76	4.06

of 0.04. Since the deviance is significant, we prefer model 3 over model 2, and should regard age as both a confounder and an effect modifier. The net result is that any estimate of the odds ratio for sex should be made with respect to a specific age.

Hence, we see that determining if a covariate, X, is an effect modifier and/or a confounder involves several issues. Determining effect modification status involves the parametric structure of the logit, while determination of confounder status involves two things. First, the covariate must be associated with the outcome variable. This implies that the logit must have a nonzero slope in the covariate. Secondly, the covariate must be associated with the risk factor. In our example this might be characterized by having a difference in the mean age for males and females. However, the association may be more complex than a simple difference in means. The essence is that we have incomparability in our risk factor groups. This incomparability must be accounted for in the model if we are to obtain a correct, unconfounded estimate of effect for the risk factor.

In practice, the confounder status of a covariate is ascertained by comparing the estimated coefficient for the risk factor variable from models containing and not containing the covariate. Any "biologically important" change in the estimated coefficient for the risk factor would dictate that the covariate is a confounder and should be included in the model, regardless of the statistical significance of the estimated coefficient for the covariate. On the other hand, a covariate is an effect modifier only when the interaction term added to the model is both biologically meaningful and statistically significant. When a covariate is an effect modifier, its status as a confounder is of secondary importance since the estimate of the effect of the risk factor depends on the specific value of the covariate.

The concepts of adjustment, confounding, interaction, and effect modification may be extended to cover the situations involving any number of variables on any measurement scale(s). The principles for identification and inclusion of confounder and interaction variables into the model are the same regardless of the number of variables and their measurement scales.

Much of this article has been abstracted from [4]. Readers wanting more detail on any topic should consult this reference.

References

[1] Bishop, Y.M.M., Fienberg, S.E. & Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Boston.

[2] Cox, D.R. & Snell, E.J. (1989). *The Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London.

[3] Hauck, W.W. & Donner, A. (1977). Wald's Test as applied to hypotheses in logit analysis, *Journal of the American Statistical Association* **72**, 851–853.

[4] Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd Ed. Wiley, New York.

[5] Jennings, D.E. (1986). Judging inference adequacy in logistic regression, *Journal of the American Statistical Association* **81**, 471–476.

[6] McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models*. Chapman & Hall, London.

(*See also* **Categorical Data Analysis**; **Loglinear Model**; **Proportional-odds Model**; **Quantal Response Models**)

STANLEY LEMESHOW &
DAVID W. HOSMER, JR

Loglinear Model

Multivariate analysis has occupied a prominent place in the classical development of statistical theory and methodology. The analysis of cross classified **categorical data**, or **contingency table** analysis as it is often called, represents the *discrete* multivariate analog of **analysis of variance** for continuous response variables, and now plays an important role in biostatistical practice. This article provides an introduction to some of the more widely used techniques for the analysis of contingency table data using loglinear models and to the statistical theory that underlies them, revising and extending an earlier review article [27] (for additional material on this topic, and related methods, *see* **Categorical Data Analysis**; **Contingency Table**).

The term *contingency*, used in connection with tables of cross classified categorical data, seems to have originated with **Karl Pearson** [61], who for an $s \times t$ table defined contingency to be any measure of the total deviation from "independent probability". The term is now used to refer to the table of counts itself. Prior to this formal use of the term, statisticians going back at least to Quetelet [64] worked with cross classifications of counts to summarize the association between variables. Pearson [59] had laid the groundwork for his approach to contingency tables when he developed his **chi-square test** for comparing observed and expected (theoretic) frequencies. Yet Pearson preferred to view contingency tables involving the cross classification of two or more polytomies as arising from a partition of a set of **multivariate, normal** data, with an underlying continuum for each polytomy. This view led Pearson [50] to develop his tetrachoric correlation coefficient for **2 × 2 tables**, and this work in turn spawned an extensive literature well chronicled by Lancaster [54] (*see* **Association, Measures of**).

The most serious problems with Pearson's approach were (i) the complicated infinite series linking the tetrachoric correlation coefficient with the frequencies in a 2×2 table, and (ii) his insistence that it always made sense to assume an underlying continuum, even when the dichotomy of interest was dead–alive or employed–unemployed, and that it was reasonable to assume that the probability distribution over such a continuum was normal. In contradistinction, Yule [72] chose to view the categories of a cross classification as fixed, and he set out to consider the structural relationship between or among the discrete variables represented by the cross classification, via various functions of the cross product ratio. Especially impressive in this, Yule's first paper on the topic, is his notational structure for n attributes or 2^n tables, and his attention to the concept of partial and joint association of dichotomous variables.

The debate between Pearson and Yule over whose approach was more appropriate for contingency table analysis raged for many years (see, for example, Pearson & Heron [63]), and the acrimony it engendered was exceeded only by that associated with Pearson's dispute with R.A. Fisher over the adjustment in the **degrees of freedom** (df) for the **chi-square test** of independence in the $s \times t$ table. [In this latter case Pearson was simply incorrect; as Fisher [34] first noted, $\mathrm{df} = (s-1)(t-1)$.]

Although much work on two-dimensional contingency tables followed the pioneering efforts by Pearson and Yule, it was not until 1935 that Bartlett, as a result of a suggestion by Fisher, utilized Yule's cross product ratio to define the notion of second-order **interaction** in a $2 \times 2 \times 2$ table, and to develop an appropriate test for the absence of such an interaction [6]. The multivariate generalizations of Bartlett's work, beginning with the work of Roy & Kastenbaum [67], form the basis of the loglinear model approach to contingency tables, which is described in detail below.

The past 40 years have seen a burgeoning literature on the analysis of contingency tables. Some of this literature emphasizes the use of the minimum modified chi-square approach (e.g. Grizzle et al. [47]) or the use of the minimum discrimination information approach (e.g. Gokhale & Kullback [K]), but the bulk of it follows Fisher in the use of **maximum likelihood**. For most contingency table problems, the minimum discrimination information approach yields maximum likelihood estimates. More recently, attention has turned to the development of hierarchical **Bayesian** approaches, which lead to the computation of *posterior distributions* (rather than point estimates) for quantities of interest (see, for example, Leonard [55], Albert & Gupta [3], Epstein & Fienberg [23], and Gelman et al. [37]).

Except for a few attempts at the use of additive (linear) models (see, for example, Bhapkar & Koch [7]), almost all of the papers written on the topic

emphasize the use of loglinear or logit models. Key papers by Birch [8], Darroch [14], Good [38], and Goodman [41, 42], plus the availability of high-speed computers, served to spur renewed interest in the problems of categorical data analysis, and culminated in a series of books first published in the 1970s and which focused in large part on the use of loglinear models for both two-dimensional and multidimensional tables (see, for example, [E, H, K, L, M, N, Q]). The past decade has seen an even greater flourishing of this expository literature, only some of which we reference here (see the review in [29]).

The subsequent sections of this presentation deal primarily with the analysis of contingency table data using loglinear models. The next section describes three examples that will serve to illustrate some of the methods of analysis, and the third section discusses briefly sampling models and estimation methods used in conjunction with categorical data analysis. The fourth section outlines the basic statistical theory associated with maximum likelihood estimation and loglinear models, including brief descriptions of the family of graphical loglinear models emanating from the work of Darroch et al. [19], and related work on the collapsing of contingency tables, as well as brief discussions of **capture–recapture** methods and latent trait (*see* **Latent Class Analysis**) and **Rasch models**, and their linkage to loglinear models. The fifth section contains examples of analysis to illustrate the basic theoretic results. The sixth section presents a brief introduction to Bayesian hierarchical approaches to loglinear models. We end with a guide to some computer programs for loglinear model analysis.

Three Examples

In this article we use three examples to illustrate the models and methods described. Two of these are classic examples, from Bartlett [6] and Waite [71], which have been analyzed repeatedly in the literature and have been used in many texts. The third, due to Edwards & Havranek [22], appears in the more recent texts by Edwards [I] and by Whitaker [S].

The data reported by Bartlett [6] in his pioneering article, and included here in Table 1, are from an *experiment* giving the response (alive or dead) of 240 plants for each combination of two explanatory variables, time of planting (early or late), and length of cutting (high or low).

Table 1 $2 \times 2 \times 2$ table

Time of planting:		Early		Late	
Length of cutting:		High	Low	High	Low
Response	Alive	156	107	84	31
	Dead	84	133	156	209
Total		240	240	240	240

Source: Bartlett [6].

Table 2 Fingerprints of the right hand classified by the number of whorls and small loops

	Small loops						
Whorls	0	1	2	3	4	5	Total
0	78	144	204	211	179	45	861
1	106	153	126	80	32		497
2	130	92	55	15			292
3	125	38	7				170
4	104	26					130
5	50						50
Total	593	453	392	306	211	45	2000

Source: Waite [71].

The questions to be answered are as follows: (i) What are the effects of time of planting and length of cutting on survival? (ii) Do they interact in their effect on survival?

The data in Table 2, from Waite [71], give the cross classification of right-hand fingerprints according to the number of whorls and small loops. The total number of whorls and small loops is at most five, and the resulting table is triangular. There the question of interest is more complicated because, as a result of the constraint forcing the data into the triangular structure, the number of whorls is "related to" the number of small loops. Such an array of counts is referred to as an *incomplete contingency table*, and the incomplete structure, in the case of the Waite data, was the cause of yet another controversy involving Pearson [62], this time with J.A. Harris (see Harris & Treloar [48]). The fit of a relatively simple model to these data is explored below. (See also [68], for a reexamination of Pearson's introduction of the methods used by Waite.)

The data in Table 3 come from a prospective epidemiologic study of 1841 workers in a Czechoslovakian car factory, intended to investigate the potential risk factors for coronary thrombosis (see

Table 3 Prognostic factors in coronary heart disease

F	E	D	C	B / A	No	Yes	No	Yes
Negative	<3	<140	No		44	40	112	67
			Yes		129	145	12	23
		≥140	No		35	12	80	33
			Yes		109	67	7	9
	≥3	<140	No		23	32	70	66
			Yes		50	80	7	13
		≥140	No		24	25	73	57
			Yes		51	63	7	16
Positive	<3	<140	No		5	7	21	9
			Yes		9	17	1	4
		≥140	No		4	3	11	8
			Yes		14	17	5	2
	≥3	<140	No		7	3	14	14
			Yes		9	16	2	3
		≥140	No		4	0	13	11
			Yes		5	14	4	4

Source: Edwards & Havranek [22].

Edwards & Havranek [22]). There are six variables corresponding to prognostic factors in the table:

A (smoking: yes, no)
B (strenuous mental work: yes, no)
C (strenuous physical work: yes, no)
D (systolic blood pressure: <140, ≥140)
E (ratio of beta and alpha lipoproteins: <3, ≥3)
F (family anamnesis of coronary heart disease: yes, no).

Sampling Models and Estimation for Contingency Tables

Let $\mathbf{x}' = (x_1, x_2, \ldots, x_t)$ be a vector of observed counts for t cells, structured in the form of a cross classification such as in Tables 1 and 2, where $t = 2^3 = 8$ and $t = 21$, respectively. Now let $\mathbf{m}' = (m_1, m_2, \ldots, m_t)$ be the vector of expected values that are assumed to be functions of unknown parameters $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_s)$, where $s < t$. Thus one can write $\mathbf{m} = \mathbf{m}(\boldsymbol{\theta})$.

There are three standard sampling models for the observed counts in contingency tables.

1. *Poisson model*. The $\{x_i\}$ are observations from independent **Poisson** random variables with means $\{m_i\}$ and **likelihood** function

$$\prod_{i=1}^{t} \left[\frac{m_i^{x_i} \exp(-m_i)}{x_i!} \right]. \tag{1}$$

2. *Multinomial model*. The total count $N = \sum_{i=1}^{t} x_i$ is a random sample from an infinite population, where the underlying cell probabilities are $\{m_i/N\}$, and the likelihood is

$$N! \cdot N^{-N} \prod_{i=1}^{t} \left(\frac{m_i^{x_i}}{x_i!} \right) \tag{2}$$

(*see* **Multinomial Distribution**).
3. *Product-multinomial model*. The cells are partitioned into sets, and each set has an independent multinomial structure, as in the multinomial model.

For the Bartlett data in the preceding section, the sampling model is product-multinomial – there are actually four independent binomials, one for each of the four experimental conditions corresponding to the two factors, time of planting and length of cutting.

For each of these sampling models the estimation problem can typically be structured in terms of a "distance" function, $K(\mathbf{x}, \mathbf{m})$, where parameter estimates $\hat{\boldsymbol{\theta}}$ are chosen so that the distance between $\mathbf{x}$ and $\mathbf{m} = \mathbf{m}(\boldsymbol{\theta})$, as measured by $K(\mathbf{x}, \mathbf{m})$, is minimized. The *minimum chi-square method* uses the distance function

$$X^2(\mathbf{x}, \mathbf{m}) = \sum_{i=1} \frac{(x_i - m_i)^2}{m_i}, \tag{3}$$

and the *minimum discrimination information method* uses

$$G^2(\mathbf{x}, \mathbf{m}) = 2 \sum_{i=1}^{t} x_i \log \left(\frac{x_i}{m_i} \right). \tag{4}$$

For the three basic sampling models for contingency tables, choosing $\hat{\boldsymbol{\theta}}$ to minimize $G^2(\mathbf{x}, \mathbf{m})$ in (4) is equivalent to maximizing the likelihood function provided that

$$\sum_{i=1}^{t} m_i(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{t} x_i \tag{5}$$

(and that constraints similar to (5) hold for each of the sets of cells under product-multinomial sampling).

Moreover, the estimators that minimize each of (3) to (5) in such circumstances belong to the class of *best asymptotic normal* (**BAN**) *estimates* for **m** (see Bishop et al. [E] for further discussion of asymptotic equivalence). Because of various additional asymptotic properties, and because of the smoothness of maximum likelihood estimates in relatively sparse tables, many authors have preferred to work with maximum likelihood estimates (MLEs), which minimize (4). We restrict our attention to MLEs, except for the "related" material on Bayesian estimation in a later section.

Basic Theory for Loglinear Models

Set-up for Two- and Three-way Tables

For expected values $\{m_{ij}\}$ for a 2×2 table,

A \ B	1	2
1	m_{11}	m_{12}
2	m_{21}	m_{22}

a standard measure of association for the row and column variables, A and B, respectively, is the *cross product ratio* (also referred to as the **odds ratio**) proposed by Yule [72]:

$$\alpha = \frac{m_{11}m_{22}}{m_{12}m_{21}}. \tag{6}$$

Independence of A and B is equivalent to setting $\alpha = 1$, and can also be expressed in loglinear form:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}, \tag{7}$$

where

$$\sum_{i=1}^{2} u_{1(i)} = \sum_{j=1}^{2} u_{2(j)} = 0. \tag{8}$$

Note that the choice of notation here parallels that for analysis of variance models.

Bartlett's [6] no-second-order interaction model for the expected values in a $2 \times 2 \times 2$ table,

m_{111}	m_{121}
m_{211}	m_{221}

m_{112}	m_{122}
m_{212}	m_{222}

is based on equating the values of α in each layer of the table; that is

$$\frac{m_{111}m_{221}}{m_{121}m_{211}} = \frac{m_{112}m_{222}}{m_{122}m_{212}}. \tag{9}$$

The expression given in (9) can be represented in loglinear form as

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}, \tag{10}$$

where, as in (8), each subscripted u-term sums to zero over any subscript; for example,

$$\sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0. \tag{11}$$

All of the parameters in (10) can be written as functions of cross product ratios (see Bishop et al. [E]). Our u-term notation follows that in Bishop et al. [E] and Fienberg [J], and differs somewhat from the λ notation adopted for example by Goodman [43, 44, 46] and by Agresti [B, C]. Furthermore, we have used symmetric linear constraints in (11), whereas other authors often choose to set selected u-terms equal to zero. The following results hold independent of the choice of parameterization and constraints.

For the sampling schemes described in the preceding section, the minimal **sufficient statistics** (msss) are the two-dimensional marginal totals, $\{x_{ij+}\}, \{x_{i+k}\}$, and $\{x_{+jk}\}$ (except for linearly redundant statistics included for purposes of symmetry), where a "+" indicates summation over the corresponding subscript. The MLEs of the $\{m_{ijk}\}$ under the model given in (10) must satisfy the likelihood equations,

$$\begin{aligned} \hat{m}_{ij+} &= x_{ij+}, \quad i, j = 1, 2, \\ \hat{m}_{i+k} &= x_{i+k}, \quad i, k = 1, 2, \\ \hat{m}_{+jk} &= x_{+jk}, \quad j, k = 1, 2, \end{aligned} \tag{12}$$

usually solved by some form of iterative procedure (*see* **Iterative Proportional Fitting**). For the Bartlett data, the third set of equations in (12) corresponds to the binomial sampling constraints.

General Results

The results described in the preceding section generalize directly to ones that are applicable to any form

of cross classification, and to a variety of models that are linear in the logarithmic scale for the expected cell values. It is helpful to have these results available in this general form so that we can adapt them to specific models for specific circumstances. Four results are described here: (i) the form of the data summaries or sufficient statistics for a model, which take the form of linear combinations of counts, often sums as in the preceding section; (ii) the form of the equations that produce maximum likelihoods, setting these data summaries equal to their expected values; (iii) the equivalence of MLEs under different sampling models; and (iv) the large-sample chi-square distribution for the usual **goodness-of-fit** statistics. The technical details follow, and some readers may wish to skip the remainder of the section until they have seen additional special cases.

Suppose that we have a collection of counts organized in the form of a vector, $\mathbf{x}$, with a corresponding vector of expected values, $\mathbf{m}$. We are interested in models for $\mathbf{m}$ such that we can represent the log expectations $\boldsymbol{\lambda}' = (\log m_1, \ldots, \log m_t)$ as linear combinations of the parameters $\boldsymbol{\theta}$. Then the following results hold under the Poisson and multinomial sampling schemes:

1. Corresponding to each parameter in $\boldsymbol{\theta}$ is an MSS that is expressible as a linear combination of the $\{x_i\}$. (More formally, if $\mathcal{M}$ is used to denote the loglinear model specified by $\mathbf{m} = \mathbf{m}(\boldsymbol{\theta})$, then the MSSs are given by the projection of $\mathbf{x}$ on to $\mathcal{M}$, $P_{\mathcal{M}}\mathbf{x}$. For a more detailed discussion, see Haberman [N].)
2. The MLE, $\hat{\mathbf{m}}$, of $\mathbf{m}$, if it exists, is unique and satisfies the likelihood equations

$$P_{\mathcal{M}}\hat{\mathbf{m}} = P_{\mathcal{M}}\mathbf{x}. \tag{13}$$

   (Note that the equations in (12) are a special case of those given by (13).)
   Necessary and sufficient conditions for the existence of a solution to the likelihood equations, (13), are relatively complex (see Haberman [L]). A sufficient condition is that all cell counts be positive – that is, $x > \mathbf{0}$ – but MLEs for loglinear models exist in many sparse situations in which a large fraction of the cells have zero counts.
   For product-multinomial sampling situations, the basic multinomial constraints (i.e. that the counts must add up to the multinomial sample sizes) must be taken into account. Typically, some of the parameters in $\boldsymbol{\theta}$ that specify the loglinear model $\mathcal{M}$, such as $\mathbf{m} = \mathbf{m}(\boldsymbol{\theta})$, are fixed by these constraints.
   More formally, let $\mathcal{M}$ be a loglinear model for $\mathbf{m}$ under product-multinomial sampling which corresponds to a loglinear model $\mathcal{M}$ under Poisson sampling such that the multinomial constraints "fix" a subset of the parameters, $\boldsymbol{\theta}$, used to specify $\mathcal{M}$. Then:
3. The MLE of $\mathbf{m}$ under product-multinomial sampling for the model $\mathcal{M}$ is the same as the MLE of $\mathbf{m}$ under Poisson sampling for the model $\mathcal{M}$.
   As a consequence of result 3, the expressions given in (12) are the likelihood equations for the $2 \times 2 \times 2$ table under the no-second-order interaction model for Poisson or multinomial sampling, as well as for product-multinomial sampling when any set of one-way or two-way marginal totals is fixed (i.e. these correspond to the multinomial constraints).
   A final result, which is used to assess the fit of loglinear models, can be stated in the following informal manner:
4. If $\hat{\mathbf{m}}$ is the MLE of $\mathbf{m}$ under a loglinear model, and if the model is correct, then the statistics

$$X^2 = \sum_{i=1}^{t} \frac{(x_i - \hat{m}_i)^2}{\hat{m}_i} \tag{14}$$

   and

$$G^2 = 2\sum_{i=1}^{t} x_i \log\left(\frac{x_i}{\hat{m}_i}\right) \tag{15}$$

   have asymptotic χ^2 distributions with $t - s$ degrees of freedom, where s is the total number of independent constraints implied by the loglinear model and the multinomial sampling constraints (if any). If the model is not correct, then X^2 and G^2, in (14) and (15), are stochastically larger than χ^2_{t-s}.

Expression 15) is the minimizing value of the distance function, (5), but (14) is not the minimizing chi-square value for the function given in (3). Both χ^2 and G^2 are special cases of the family of *power-divergence statistics*, the distance function of which takes the form

$$K(\mathbf{x}, \hat{\mathbf{m}}) = \frac{2}{\phi(\phi+1)} \sum_{i=1}^{t} x_i \left[\left(\frac{x_i}{\hat{m}_i}\right)^{\phi} - 1\right], \tag{16}$$

where ϕ is a real-valued parameter in the interval $-\infty < \phi < \infty$. The statistic χ^2 corresponds to $\phi = 1$, and the statistic G^2 corresponds to the limit as $\phi \to 0$. For further details on the properties of the general family of power divergence statistics, see Read & Cressie [66].

In the next section, these basic results are applied in the context of three data sets presented in the previous section.

Loglinear Models for High-dimensional Tables

The same ideas and ANOVA-like models in the logarithmic scale are useful for multiway tables. For such models, the minimal sufficient statistics are sets of marginal totals of the full table. Furthermore, all independence or conditional independence relationships are representable as loglinear models, and these models have estimated expected values that can be computed directly. There is a somewhat larger class of loglinear models with this direct or *decomposable* representation described below. For all loglinear models that are not decomposable, we require an iterative solution of likelihood equations.

Not all applications of loglinear models involve such simple structures as 2^3 tables or incomplete 6×6 arrays. Indeed, much of the methodology was developed in the mid-1960s to deal with very large, highly multidimensional tables. For example, in the National Halothane Study [10], investigators considered data on the use of (i) five anesthetic agents in operations involving (ii) four levels of risk, and patients of (iii) two sexes, (iv) ten age groups, with (v) seven differing physical statuses (levels of anesthetic risk) and (vi) previous operations (yes, no) for (vii) three different years, from (viii) 34 different institutions. Two sets of data were collected, the first consisting of all deaths within six weeks of surgery, and the second consisting of a sample (of comparable size) of all those exposed to surgery. Thus the data consisted of two very sparse $5 \times 4 \times 2 \times 10 \times 7 \times 2 \times 3 \times 34$ tables, each containing in excess of 57 000 cells. One of the more successful approaches used in the analysis of the data in these tables was based on loglinear models and the generalizations of the methods illustrated in this section.

One of the key reasons why loglinear models have become so popular in such analyses is that they lead to a simplified description of the data in terms of marginal totals – the minimal sufficient statistics of result 1 of the section "Basic theory for loglinear models". This is especially important when the table of data is large and sparse. For more details on the halothane study analyses, see Bishop et al. [E].

The doubly subscripted u-term notation introduced in the previous sections generalizes immediately to multiway tables. As in the previous section, we restrict attention to hierarchical models, in which if any u-term is set equal to zero, all of its higher order relatives must be set equal to zero; for example, setting $u_{12(ij)} = 0$ for all i, j implies setting $u_{123(ijk)} = 0$ for all i, j, k.

We need to think of parameters in loglinear models as deviations from lower-order parameters; they can also be represented and interpreted as a function of generalized odds ratios or cross product ratios; for example, see Fienberg [J] or Agresti [A, B].

In a multiway contingency table, the model that results from setting exactly one two-factor term (and all its higher-order relatives) equal to zero is called a *partial association* model. For example, in four dimensions, if we set $u_{12(ij)} = 0$ for all i, j, then the minimal sufficient statistics are $\{x_{i+kl}\}$ and $\{x_{+jkl}\}$, and the resulting partial association model corresponds to the conditional independence of variables 1 and 2 given 3 and 4. The corresponding maximum likelihood estimates for the expected cell frequencies are

$$\hat{m}_{ijkl} = \frac{x_{i+kl}x_{+jkl}}{x_{++kl}} \quad \text{for all } i, j, k, l. \tag{17}$$

For more details on partial association models and their uses, see Bishop et al. [E].

Loglinear Models and Graphical Representations

One of the major innovations of the past 15 years has been the development of methods associated with a subfamily of loglinear models known as *graphical loglinear models*. The formulation of graphical models is due originally to Darroch et al. [19], and has now found its way into several introductory textbooks on loglinear models (e.g. [F, I, R]), and serves as the basis for several monographs [O, S].

We begin with some special notation and then define the class of graphical models. We denote the situation in which F and G are conditionally independent given H by $F \perp G|H$. Thus, in a three-way table, if variables 1 and 2 are conditionally

independent given 3, we denote this by

$$1 \perp 2|3.$$

Similarly, in a four-way table, if variables 1 and 2 are conditionally independent given 3 and 4, we denote this by $1 \perp 2|\{3, 4\}$.

In formal mathematics, a *graph* $G = (K, E)$ is based on a set of vertices, K, and a set of edges, E, which consists of pairs of elements from K. We depict such a graph using a picture with vertices linked by edges. For example, the graph $G = (K, E)$ with $K = \{1, 2, 3, 4, 5, 6\}$ and $E = \{(1, 2), (1, 4), (2, 3), (3, 4), (3, 5), (4, 5), (4, 6)\}$ corresponds to Figure 1.

The following definition links partial association models to the absence of edges in a graph. Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ be a vector of random variables, and $K = \{1, 2, \ldots, k\}$. Furthermore, let $K\backslash\{i, j\}$ be the set of vertices in K excluding i and j. An undirected graph is an independence graph if there is no edge between two vertices whenever the variables they represent are conditionally independent given all remaining variables, that is $i \perp j|K\backslash\{i, j\}$ for all $(i, j) \notin E$, which corresponds to the partial association model discussed in the preceding subsection. This simply means that we can represent all possible conditional independence relationships in terms of the absence of edges in an undirected graph.

In the above example with $K = \{1, 2, 3, 4, 5, 6\}$ and $E = \{(1, 2), (1, 4), (2, 3), (3, 4), (3, 5), (4, 5), (4, 6)\}$, there are $\binom{6}{2} = 15$ possible edges that could have connected the six vertices, nine of which are absent and each of which corresponds to a conditional independence statement of the form $1 \perp 3|\{2, 4, 5, 6\}$. Some of these conditional independence relationships can be combined and expressed in a more succinct form that is intuitive from the graph. For example, the single conditional relationship, $\{1, 2\} \perp \{5, 6\}|\{3, 4\}$, which can be seen from the "separation" of $\{1, 2\}$ from $\{5, 6\}$ by $\{3, 4\}$ in the graph, summarizes four different conditional independence relationships.

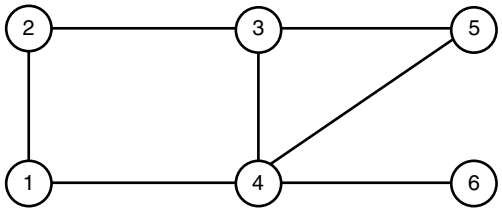

Figure 1 An illustrative graph with six vertices and seven edges, and a four-cycle

Independence graphs can be used in connection with all random variables with positive density (continuous or discrete), and many of the results for independence graphs discussed in Lauritzen [O] and Whittaker [S] are applicable to such random variables. For the purposes of this article, however, we can think in terms of categorical variables, and the models that have independence graph representations are said to be *graphical models*. For categorical variables, all graphical models are loglinear. (For further details, see Lauritzen [O] and Whittaker [S].)

For three-way tables, the models of complete independence (no edges), joint independence (two absent edges), and conditional independence (one absent edge) are graphical, but the no-second-order interaction model is not, because the graph with all three edges present corresponds to the saturated model. Thus, all graphical models for three-way tables are *decomposable*; that is, the expected values can be written *explicitly* as a product and/or ratio of the expected marginal totals corresponding to the sufficient statistics. In this sense, the expected values can be *directly decomposed* in terms of the corresponding margins. In such circumstances, the MLEs can be written out directly and have a simple interpretation; see [1, 18, 33]. There are also a number of especially interesting technical results that apply to decomposable loglinear models [O] and many of these results have proved useful for computing bounds on cell counts given the margins corresponding to a decomposable loglinear model (see [12]).

For four-way tables, the graph with four edges in a cycle; that is, corresponding to the joint occurrence of $2 \perp 3|\{1, 4\}$ and $1 \perp 4|\{2, 3\}$ and with $E = \{(1, 2), (2, 3), (3, 4), (1, 4)\}$, is graphical but also *nondecomposable*. This means that the model is not decomposable and thus the expected values under the model cannot be expressed as an explicit function of the marginal totals corresponding to the sufficient statistics $\{x_{ij++}\}$, $\{x_{+jk+}\}$, $\{x_{++kl}\}$, and $\{x_{i++l}\}$.

As we saw above, all graphical models for three-way tables are decomposable. For higher-way tables, a graphical model is nondecomposable whenever its independence graph includes a cycle involving four or more vertices. Thus in the above example involving six variables where $K = \{1, 2, 3, 4, 5, 6\}$ and $E = \{(1, 2), (1, 4), (2, 3), (3, 4), (3, 5), (4, 5), (4, 6)\}$,

there is a four-cycle involving the edges linking $\{1, 2, 3, 4\}$ and thus the corresponding graphical loglinear model is nondecomposable.

In general, we have the following relationship among classes of loglinear models:

hierarchical models

$\cup$

graphical models

$\cup$

decomposable models

$\cup$

conditional independence models.

Model Selection and Collapsing

Many authors have devised techniques for selecting among the class of loglinear models applicable for contingency table structures. These typically (although not always) resemble corresponding model selection procedures for analysis of variance and regression models (*see* **Variable Selection**). See, for example, the discussions in Agresti [B], Bishop et al. [E], and Fienberg [J]. Edwards [I] and Whitaker [S] have special sections on model selection that take special advantage of the form of graphical models and the link between edges and two-factor effects.

A special aspect of model selection relates to the issue of when is it possible to work with and report loglinear model effects from a reduced table, collapsing over one or more variables of initial interest. The problem of collapsing was first taken up in Bishop [9] and Bishop et al. [E], who defined the concept in terms of the parameters of the loglinear model itself, now referred to as *parametric collapsibility*. Their discussion led to an extensive literature, in which differing definitions of collapsibility were proposed, including *model collapsibility* in which MLEs for the probabilities associated with a subset of variables, say A, can be performed directly in the A-margin of the table (see Asmussen & Edwards [4] and Lauritzen [O] for further details and references).

When a contingency table is not collapsible with respect to a subset of variables, inferences about the relationships among the remaining variables drawn from the corresponding marginal table are inevitably misleading. The best known example of this problem is referred to as **Simpson's paradox** or Yule's paradox. It is usually described as a situation involving three binary variables A, B, and C, such that the cross product ratio between A and B is greater than 1 for each level of C (i.e. there is a positive conditional relationship) but the cross product ratio in the marginal table for A and B is less than 1 (i.e. there is a negative marginal relationship).

Capture Multiple Recapture Analyses

This type of analysis estimates the size of a nonchanging population (see, for example, Bishop et al. [E] and Fienberg [26]). If the members of nonchanging populations are sampled k successive times (possibly dependent), the resulting recapture history data can be displayed in the form of a 2^k table with one missing cell, corresponding to those never sampled. Such an array is amenable to loglinear analysis, the results of which can be used to project a value for the missing cell.

In recent years there have been a number of major applications of the capture–recapture methodology ($k = 2$), especially in the context of the US decennial **census**; for example, see Zaslavsky & Wolfgang [73] and Fienberg [28], and in a variety of epidemiologic settings. For a detailed description of the history of the methodology and its potential for use in the context of epidemiology and public health, see Hook & Regal [49], and the International Working Group for Disease Monitoring and Forecasting [50, 51].

A key assumption in the use of standard loglinear models for capture–recapture and multiple-recapture population estimation is that of constant capture probabilities, or homogeneity. A traditional approach to allow for heterogeneity has been stratification, with separate models used for individual homogeneous strata. The problem with stratification as a strategy is that it often leads to very sparse cross classifications, and a substantial increase in the variability associated with population estimates. Recent developments linked to a variation of the Rasch model have led to extensions of the standard models that allow for special multiplicative forms of heterogeneity; for example, see [1, 20].

(For further details on these and related models for population size estimation, *see* **Capture–Recapture**; **Rasch Models**.)

Latent Trait and Rasch Models

In psychologic tests or attitude studies, we are often interested in quantifying the value of an unobservable *latent trait*, such as mathematic ability or manual dexterity, on a sample of individuals. While latent traits are not directly measurable, we can either assume something about the way in which the latent trait relates to the observable or *manifest* variables or assume that we can assess indirectly a person's value for the latent trait from his/her responses to a set of well chosen items on a test (see, for example [5]).

In the 1970s, Goodman [44] and Haberman [N] developed a special representation for the analysis of contingency tables for manifest variables using loglinear models in the presence of latent variables, beginning with the traditional model in which the manifest variables are conditionally independent given the latent variables.

The second approach is prevalent in educational testing and has recently found its way into a wide variety of applications. The simplest model in this domain was introduced by Rasch [65], and is known as the Rasch model. Given responses from n individuals to k items in a test, the Rasch model permits the estimation of parameters associated with individuals and with items, as well as prediction of the person's behavior when confronted with a different set of items from the same domain. In the 1980s, an important relationship between the Rasch model and loglinear models was recognized by Tjur [70], Cressie & Holland [11], and Duncan [21]. The representation of these models in terms of symmetry and **quasi-symmetry** was presented in Fienberg & Meyer [32], Darroch [16] and Darroch & McCloud [17]. See also Anderson [D] and the article on **Rasch Models** for a presentation of this topic.

Association Models for Ordinal Variables

Loglinear models as described in this article ignore any structure linking the categories of variables, yet biostatistical problems often involve variables with ordered categories; for example, differing dosage levels for a drug or the severity of symptoms or side effects. Beginning in the late 1970s, methods for a special class of models, known as *association* models, moved to the forefront of methodological research. Goodman [45] provided a framework for association models that builds on extensions to standard loglinear models and utilizes multiplicative interaction terms. For a detailed description of these and other methods for ordinal variables, see Agresti [A] and Clogg & Shidadeh [G]. Etzioni et al. [25] provide a useful review of association models for ordinal variables in medical research using notation compatible with this article.

Loglinear Model Analyses

Bartlett's Data and No-second-order Interaction

For the 2^3 table of Bartlett, variables 2 (time of planting) and 3 (length of cutting) are fixed by design, so that $\hat{m}_{+jk} = 240$, and the estimated expected values under the no-second-order interaction model of the expressions given in (12) are shown in Table 4. These values were computed by Bishop et al. [E] using the method of *iterative proportional fitting* (IPF). Bartlett originally found the solution to (14) by noting that the constraints in his specification (11), reduced (14) to a single cubic equation for the discrepancy $\Delta = \hat{m}_{111} - x_{111}$. Note that the expected values satisfy (14), e.g., $\hat{m}_{12+} = 78.9 + 36.1 = 115 = 84 + 31 = x_{12+}$. The goodness-of-fit statistics for this model are $X^2 = 2.27$ and $G^2 = 2.29$. Using result 4 of the preceding section, one compares these values to tail values of the chi-square distribution with 1 df, for example $\chi_1^2(0.10) = 2.71$, and this suggests that the no-second-order interaction model provides an acceptable fit to the data.

Since the parameters u, $\{u_{2(j)}\}$, $\{u_{3(k)}\}$, and $\{u_{23(jk)}\}$ are fixed by the binomial sampling constraints for these data, the model given by (12)

Table 4 Observed and expected values for the Bartlett data, including the no-second-order interaction model

Cell	Observed x	Estimated expected $\hat{m}$
1,1,1	156	161.1
2,1,1	84	78.9
1,2,1	84	78.9
2,2,1	156	161.1
1,1,2	107	101.9
2,1,2	133	138.1
1,2,2	31	36.1
2,2,2	209	203.9

is often rewritten as

$$\log\left(\frac{m_{1jk}}{m_{2jk}}\right) = 2[u_{1(1)} + u_{12(1j)} + u_{13(2k)}]$$
$$= w + w_{2(j)} + w_{3(k)}, \qquad (18)$$

where

$$\sum_j w_{2(j)} = \sum_k w_{3(k)} = 0.$$

Expression (18) is referred to as a *logit* model for the log odds for alive versus dead (*see* **Logistic Regression**). The simple additive structure corresponds to Bartlett's notion of no-second-order interaction.

Waite's Fingerprint Data and Quasi-independence

For the Waite fingerprint data of Table 2, one model that has been considered is the simple additive loglinear model of (9), but only for those cells where positive counts are possible; that is, in the upper triangular section. For cells with $i > j, m_{ij} = 0$ a priori. This restricted version of the independence model is referred to as a **quasi-independence** model, and the results of the preceding section can be used in connection with it. The MSSs are still the row and column totals (result 1). The likelihood equations under multinomial sampling are (applying results 1 and 2):

$$\hat{m}_{i+} = x_{i+}, \quad i = 0, 1, 2, \ldots, 5,$$
$$\hat{m}_{+j} = x_{+j}, \quad j = 0, 1, 2, \ldots, 5, \qquad (19)$$

where $m_{ij} = 0$ for $i > j$. A solution of (19) satisfying the model can be found directly, or by using a standard iterative procedure. The estimated expected values for the fingerprint data under the model of quasi-independence are given in Table 5, and they satisfy the marginal constraints in (19).

The goodness-of-fit statistics for this model are $X^2 = 399.8$ and $G^2 = 450.4$, which correspond to values in the very extreme right-hand tail of the χ^2_{10} distribution. Thus the model of quasi-independence seems inappropriate. Darroch [15] describes the loglinear model of F-independence (with more parameters than the quasi-independence model), which takes into account the way in which the constraint – that the number of small loops plus the number of whorls cannot exceed 5 – makes the usual definition of independence inappropriate. This model in loglinear form is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{3(5-i-j)}, \qquad (20)$$

where the u_3 parameters correspond to diagonals along which the sum of the numbers of whorls and small loops is constant. Darroch & Ratcliff [18] illustrate the fit of the F-independence model to a related set of fingerprint data involving large rather than small loops.

Table 5 Estimated expected values for fingerprint data under quasi-independence

	Small loops						
Whorls	0	1	2	3	4	5	Total
0	200.6	167.4	166.6	150.3	131.1	45.0	861
1	122.2	101.9	101.4	91.6	79.9		497
2	85.5	71.4	71.0	64.1			292
3	63.8	53.2	53.0				170
4	70.9	59.1					130
5	50.0						50
Total	93	453	392	306	211	45	2000

Application of Graphical Loglinear Models: Prognostic Factors for Coronary Disease

To illustrate some of the features of graphical models, we now analyze the data in Table 3 on prognostic factors for coronary heart disease among 1841 men in a Czech car factory. Our analysis follows closely that in Whittaker [S]. We begin by examining all partial association models and computing G^2 for each of the 15 partial association models found by setting one two-factor term equal to zero (see Table 6).

There are 16 df associated with each G^2 value. If we drop edges in a graph using the 0.05 P value for a χ^2 variable with 16 df, that is 26.30, then we end up with the graph shown in Figure 2.

Table 6 Goodness of fit of partial association models for coronary heart disease data

A	*					
B	22.65	*				
C	42.80	689.99	*			
D	28.72	12.23	14.81	*		
E	40.02	17.24	18.63	31.06	*	
F	21.31	22.79	22.15	18.35	18.32	*
	A	B	C	D	E	F

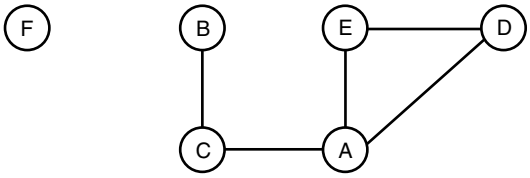

Figure 2 A preliminary graphical model for the prognostic factors example based on partial association models

The likelihood ratio statistic for the loglinear model corresponding to this graph is $G^2 = 83.75$ with 51 df. This corresponds to a P value of 0.0026, and suggests that we have deleted too many edges. A few additional steps of addition and deletion yield the model

$$[ABCE][ADE][BF],$$

for which the likelihood ratio statistic $G^2 = 44.59$ with 42 df suggests a well fitting model. The corresponding independence graph is shown in Figure 3.

The alternative graphical model shown in Figure 4, with two fewer edges, is somewhat simpler than the preferred model reported by Edwards [I] and fits the data well.

In Figure 4, F (family history) is conditionally independent of {D, E} (the physical symptoms), given {A,B,C} (the behavioral conditions). For further details on the analysis of this data set using graphical models, see Edwards [I] and Whitaker [S].

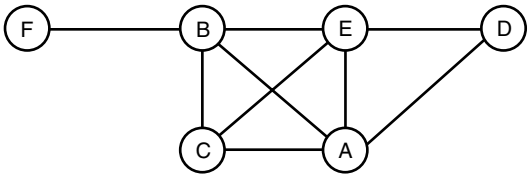

Figure 3 An intermediate graphical model for the prognostic factors example

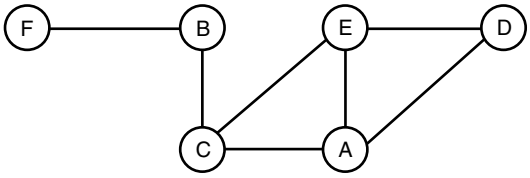

Figure 4 The final graphical model for the prognostic factors example

Bayesian Approaches

Background

In recent years, much attention has been directed to the development of Bayesian approaches for contingency tables and loglinear models. Early references include Good [39, 40] and Fienberg & Holland [31]. Initial attempts at formulating a Bayesian approach to estimation in contingency tables concentrated on the problem of incorporating *prior knowledge* about unobservable cell proportions or about expected cell counts. Those early efforts resulted in the derivation of the Dirichlet family of distributions as the *conjugate* family of **prior distributions** for cell proportions and expected counts. More recently, emphasis has been placed on formulation of more complex prior distributions that permit incorporating information about the underlying structure in a contingency table (e.g. Albert & Gupta [2, 3], Knuiman & Speed [52], and Epstein & Fienberg [23, 24]), and on computational issues (e.g. Gelman & Rubin [36], Epstein & Fienberg [23], and Gelman et al. [37]).

The general approach adopted by most of these authors has been as follows: using either the multinomial or the Poisson sampling models, incorporate uncertainty about the expected cell proportions (or expected cell counts) via the Dirichlet conjugate family of prior distributions, with variations to account for more or less structure in the expected cell means. **Markov chain Monte Carlo** methods (see, for example, Gelfand & Smith [35]) can then be used to approximate the marginal posterior distributions of the expected cell values or of any continuous functions of m.

A different, but parallel line of development is described in the work of, for example, Leonard [55], Laird [53], and Nazaret [58]. These authors address the problem of estimation in contingency tables from a loglinear model approach, and attempt to use Bayesian results similar to those developed by Lindley & Smith [56] for the linear model. Except where convenient for computations, we will describe the Bayesian methods that address contingency tables directly, without resorting to results from the linear models literature.

To introduce the Bayesian approach to estimation, we consider the counts $\{x_i\}, i = 1, \ldots, t$ (t equals the number of cells), to be observations from independent Poisson random variables, with means or expected

values $\{m_i\}$. In the Bayesian framework, the $\{m_i\}$ are random variables having some prior distribution, and the statistician's task is to update the prior to a *posterior distribution* by incorporating the information about the $\{m_i\}$ provided by the observed counts $\{x_i\}$. For the multinomial sampling model, we are interested in the marginal posterior distributions of parameters of a loglinear model, $\{\theta_i\}$, or the posterior distribution of the expected cell values, $\{m_j\}$. For the Poisson and the multinomial sampling models, likelihood functions are given in (1) and (2).

Estimation and Computation

Here we refer to the Poisson sampling model. Results, however, also apply to the multinomial model after appropriate normalization (see, for example, Gelman et al. [37]).

The simplest way to incorporate prior information about the value of the expected cell counts $\{m_i\}$ is via the Dirichlet conjugate family of prior distributions. If $\{m_i\}$ are jointly distributed a priori as independent Dirichlet random variables with parameters k and $\{\eta_i\}$, then the joint prior distribution has density function

$$p(\mathbf{m}|k, \boldsymbol{\eta}) \propto \prod_{i=1}^{t} m_i^{k\eta_i - 1}, \tag{21}$$

where $k > 0$ and $\eta_i > 0$. For $\boldsymbol{\eta} = \{\eta_i\}$, parameters η_i can be thought of as our prior "guess" about the value of $\{m_i\}$, while the *flattening constant* k [30, 39] represents our prior certainty about those guesses.

For the likelihood corresponding to the Poisson sampling model in (1), and for $p(\mathbf{m}|k, \boldsymbol{\eta})$ as in (21), the posterior distribution of the expected cell counts $\mathbf{m}$ is proportional to the product of the likelihood function and the prior distribution

$$\begin{aligned} p(\mathbf{m}|\mathbf{x},k, \boldsymbol{\eta}) &\propto \prod_i p(x_i|m_i) \times p(m_i|k, \eta_i) \\ &\propto \prod_i \exp\{-m_i\} m_i^{x_i + k\eta_i - 1}. \end{aligned} \tag{22}$$

Gelman & Rubin [36] and Gelman et al. [37] have developed a Bayesian version of IPF (BIPF) starting from (22) and using the multiplicative version of the loglinear model. Their algorithm produces estimates of marginal posterior distributions of the $\{m_i\}$ (or of continuous functions of the $\{m_i\}$) rather than point estimates (as does the standard IPF). In this sense, BIPF is a misnomer, since it incorrectly suggests that the Bayes estimates obtained are, as in the frequentist case, just point estimates of the $\{m_i\}$. For some details on the derivation of BIPF, the reader is referred to Gelman & Rubin [36].

Examples

In the previous section, we applied the BIPF to the Bartlett 2^3 table and to Edwards & Havranek's [22] six-way table on prognostic factors for heart disease. In both examples we used a multinomial sampling model with the constraints imposed by the sampling scheme.

The loglinear models that we fit to each data set were those described earlier. Thus, for the $2 \times 2 \times 2$ table of Bartlett, we used a loglinear model with main effects and all two-way interactions, while for the six-way table we used a model corresponding to

$$[ABC][ACE][ADE][BF].$$

We incorporated prior information via the Dirichlet conjugate family (see (21)), with $k\eta_i = 0.5$ for all i, that results is noninformative prior densities for the expected cell counts $\{m_i\}$. For a somewhat different Bayesian analysis of these data along with an excellent discussion of Bayesian model search focusing on graphical models, see [57]. Dobra, Karr, Sanil, and Fienberg [13] use similar tools in the context of disclosure limitation problems and apply them to these data as well.

Our Bayesian results for the Bartlett data are given in Table 7. To highlight the differences between results obtained from the frequentist and the Bayesian approaches (see Table 4 for the former), we provide not only a point estimate for the expected counts $\{m_i\}$ (we chose the means of the marginal posterior distributions as our point estimates) but also the posterior 5th, 25th, 50th, 75th, and 95th percentiles of the distributions of each $\{m_i\}$.

Note that, as required by the sampling scheme, $\hat{m}_{12+} = 78.75 + 36.01 = 114.8 = 84 + 31 = x_{12+}$ (within numerical error), and that the same holds true for $\hat{m}_{11+}$, $\hat{m}_{21+}$, and $\hat{m}_{22+}$.

We give results obtained for the prognostic factors for heart disease data in Table 8, which shows the means of the marginal posterior distributions of the expected cell counts $\{m_i\}$. Note that, as required, the

Table 7 Observed values, posterior means of expected values, and percentiles of posterior distribution of expected values for the Bartlett data, including the no-second-order interaction model

Cell	Observed x	Posterior mean	Posterior percentiles				
			5th	25th	50th	75th	90th
1,1,1	156	161.13	150.67	156.82	161.31	165.37	170.97
2,1,1	84	78.87	69.03	74.63	78.69	83.18	89.33
1,2,1	84	78.75	68.67	74.45	78.43	82.78	89.67
2,2,1	156	161.25	150.33	157.22	161.57	165.55	171.33
1,1,2	107	101.89	91.06	97.36	101.87	106.43	112.83
2,1,2	133	138.11	127.17	133.57	138.13	142.64	148.94
1,2,2	31	36.01	29.56	32.89	35.81	38.91	43.67
2,2,2	209	203.99	196.33	201.09	204.19	207.11	210.44

Table 8 Prognostic factors in coronary heart disease: posterior means of expected counts

F	E	D	C	B / A	No	Yes	No	Yes
Negative	<3	<140	No		41.2	33.6	104.7	68.2
			Yes		122.1	139.0	15.0	24.6
		≥140	No		32.9	16.5	83.5	33.4
			Yes		97.4	68.1	12.0	12.0
	≥3	<140	No		27.0	31.7	68.6	64.3
			Yes		52.7	83.4	6.5	14.8
		≥140	No		26.9	26.5	68.4	53.9
			Yes		52.6	69.8	6.5	12.4
Positive	<3	<140	No		6.3	5.1	21.2	13.8
			Yes		18.5	21.1	3.0	5.0
		≥140	No		5.0	2.5	16.9	6.8
			Yes		14.8	10.3	2.4	2.4
	≥3	<140	No		4.1	4.8	13.9	13.0
			Yes		8.0	12.6	1.3	3.0
		≥140	No		4.1	4.0	13.8	10.9
			Yes		8.0	10.6	1.3	2.5

sum of the estimated expected counts equals N, the total number of individuals in the study.

Brief Guide to Computer Programs for Loglinear Model Analysis

As with other forms of multivariate analysis, the analysis of multidimensional contingency tables relies heavily on computer programs. A large number of these have been written to compute estimated parameter values for loglinear models and associated test statistics, and most computer installations at major universities have one or more programs available for users (*see* **Software, Biostatistical**).

The most widely used numerical procedure for the calculation of maximum likelihood estimates for loglinear models a decade ago was IPF, which iteratively adjusts the entries of a contingency table to have marginal totals equal to those used in specifying the likelihood equations. The IPF approach has been implemented in the BMDP 4F Program and in SPSS. The major advantage of the IPF method is that it requires limited computer memory capabilities since it does not require matrix inversion or equivalent computations, and thus can be used in connection with the analysis of very high-dimensional tables. Its major disadvantage is that it does not provide, in an easily accessible form, estimates of the basic loglinear model parameters (and an estimate of their asymptotic covariance matrix); it provides only estimated expected values.

MIM is an excellent Windows program for graphical modeling that is useful for fitting graphical (and other loglinear) models designed in part to accompany Edwards [I]. A student version is available free from `www.hypergraph.dk`.

The other numerical approaches suggested for the computation of maximum likelihood estimates are typically based on classical procedures for solving nonlinear equations, such as modifications of Newton's method or the Newton–Raphson method (*see* **Optimization and Nonlinear Equations**). Since such approaches can be implemented as part of the methods for the broader class of **generalized linear models** of which loglinear and logit models are special cases (see, for example, McCullagh & Nelder [P]), a common approach in several computer packages is to embed loglinear and logit models approaches as part of GLM routines, see; for example **S-PLUS** GLIM, SAS (PROC GENMOD),

STATA, and SYSTAT. The virtue of these programs is that they produce both estimated expected values and estimated parameter values, and an estimate of the asymptotic covariance matrix. The user of a GLM package should be sure to check the specific parameterization used, as the constraints on the loglinear models typically vary from package to package. No matter what the choice of parameterization for the loglinear model parameters, the estimates of the expected values and the goodness of fit statistic values should agree with those computed using the IPF algorithm. Some packages such as BMDP and SPSS also have separate subroutines for logit and logistic regression models. SAS's JMP has only a logistic regression routine.

Agresti [C] includes an especially nice appendix with a guide to SAS and SPSS programming, and Agresti [B] includes examples from GLIM and BMDP. Stokes et al. [69] provide a detailed guide with examples for the SAS PROC CATMOD, which can be used for loglinear models as well as a number of other approaches to the analysis of categorical data.

Before running any of the loglinear model or generalized linear model routines referred to above on sparse multiway tables with one or more zero entries, users should read the package documentation with care, since some packages may treat zeros in an unexpected fashion.

Textbook References

[A] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York (An introduction to the analysis of categorical data with special emphasis on loglinear models and their variants for ordinal data. It emphasizes the use of cross product ratios to describe association in different ways.).

[B] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York (A second generation introduction to loglinear models for the analysis of categorical data. It includes a mix of theory and methods, as well as many examples.).

[C] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York (A second generation nonmathematical introduction to loglinear models for the analysis of categorical data. It includes many examples and a guide to computing in SAS and SPSS.).

[D] Anderson, E.B. (1990). *The Statistical Analysis of Categorical Data*. Springer-Verlag, Heidelberg (A second generation introduction to loglinear models, including such topics as association models for ordinal data, **correspondence analysis**, graphical models, and the Rasch model.).

[E] Bishop, Y.M.M., Fienberg, S.E. & Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass (A systematic exposition and development of the loglinear model for the analysis of contingency tables through the early 1970s, primarily using maximum likelihood estimation, and focusing on the use of iterative proportional fitting. It includes chapters on measures of association, and others on special related topics. It contains both theory and numerous examples from many disciplines with detailed analyses.).

[F] Christensen, R. (1990). *Loglinear Models*. Springer-Verlag, New York (An intermediate introduction to loglinear models, with special emphasis on interpretation including graphical models. It also contains chapters on logistic regression and logistic discrimination (*see* **Discriminant Analysis, Linear**.)).

[G] Clogg, C.C. & Shidadeh, E.S. (1994). *Statistical Models for Ordinal Variables*. Sage, Thousand Oaks (An introduction to association models logit-type regression models for ordinal variables.).

[H] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman & Hall, New York (The second edition of an early guide to the analysis of categorical data by D.R. Cox, including basic results for loglinear and logit models. It contains numerous examples and references.).

[I] Edwards, D. (2000). *Introduction to Graphical Modeling*. 2nd Ed. Springer-Verlag, New York.

[J] Fienberg, S.E. (1980). *The Analysis of Cross-classified Categorical Data*, 2nd Ed. MIT Press, Cambridge, Mass (A comprehensive introduction, for those with some training in statistical methodology, to the analysis of categorical data using loglinear models and maximum likelihood estimation. The emphasis is on methodology, with numerous examples and problems.).

[K] Gokhale, D.V. & Kullback, S. (1978). *The Information in Contingency Tables*. Marcel Dekker, New York (A development of minimum discrimination information procedures for linear and loglinear models. It contains a succinct theoretical presentation, followed by numerous examples.).

[L] Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago (A highly mathematical, advanced presentation of statistical theory associated with loglinear models and of related statistical and computational methods. It contains examples, but is suitable only for mathematical statisticians who are familiar with the topic.).

[M] Haberman, S.J. (1978). *Analysis of Qualitative Data*, Vol. 1: *Introductory Topics*. Academic Press, New York (See next entry.).

[N] Haberman, S.J. (1979). *Analysis of Qualitative Data*, Vol. 2: *New Developments*. Academic Press, New York (An intermediate-level, two-volume introduction to the analysis of categorical data via loglinear models, emphasizing maximum likelihood estimates computed via the Newton–Raphson algorithm. Volume 1 examines complete cross classifications, and Volume 2 considers multinomial response models, incomplete tables, and related

topics. The volumes contain many examples, problems, and solutions, and a computer program listing (for two-way tables) is included in Volume 2.).

[O] Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press, Oxford (An advanced mathematical treatment of graphical models and their analysis, for both continuous and categorical variables, with emphasis on topics such as decomposability and exact tests.).

[P] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London (The definitive introduction to generalized linear models and their properties, including loglinear and logit models as special cases. It includes many examples.).

[Q] Plackett, R.L. (1974). *The Analysis of Categorical Data*. Griffin, London (A concise introduction to statistical theory and methods for the analysis of categorical data. It assumes a thorough grasp of basic principles of statistical inference. There is considerable emphasis on two-way tables. It contains many examples and exercises.).

[R] Santner, T.J. & Duffy, D.E. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York (A graduate level introduction to loglinear models and other methods for the analysis of discrete data. It includes discussions of Bayesian methods, graphical models, and logistic regression.).

[S] Whitaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York (A somewhat mathematical introduction to graphical loglinear models with extensive examples and applications. It deals with both directed and undirected graphs.).

Other References

[1] Agresti, A. (1994). Simple capture–recapture models permitting unequal catchability and variable sampling effort, *Biometrics* **50**, 494–500.

[2] Albert, J.H. & Gupta, A.K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables, *Annals of Statistics* **10**, 61–68.

[3] Albert, J.H. & Gupta, A.K. (1983). Estimation in contingency tables using prior information, *Journal of the Royal Statistical Society, Series B* **45**, 60–69.

[4] Asmussen, S. & Edwards, D. (1983). Collapsibility and response variables in contingency tables, *Biometrika* **70**, 567–578.

[5] Baker, F.B. (1992). *Item Response Theory. Parameter Estimation Techniques*. Marcel Dekker, New York.

[6] Bartlett, M.S. (1935). Contingency table interactions, *Journal of the Royal Statistical Society, Supplement* **2**, 248–252.

[7] Bhapkar, V.P. & Koch, G. (1968). On the hypotheses of "no interaction" in contingency tables, *Biometrics* **24**, 567–594.

[8] Birch, M.W. (1963). Maximum likelihood in three-way contingency tables, *Journal of the Royal Statistical Society, Series B* **25**, 229–233.

[9] Bishop, Y.M.M. (1971). Effects of collapsing multidimensional contingency tables, *Biometrics* **27**, 545–562.

[10] Bunker, J.P., Forrest, W.H. Jr, Mosteller, F. & Vandam, L. (1969). *The National Halothane Study*, Report of the Subcommittee on the National Halothane Study of the Committee on Anesthesia, Division of Medical Sciences, National Academy of Sciences – National Research Council, National Institutes of Health, National Institute of General Medical Sciences, Bethesda, US Government Printing Office, Washington.

[11] Cressie, N.E. & Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models, *Psychometrika* **48**, 129–141.

[12] Dobra, A. and Fienberg, S.E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, *Proceedings of the National Academy of Sciences* **97**, 11885–11892.

[13] Dobra, A., Karr, A., Sanil, A. P., and Fienberg, S.E. (2002). Software systems for tabular data releases, *The International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, in press.

[14] Darroch, J.N. (1962). Interaction in multi-factor contingency tables, *Journal of the Royal Statistical Society, Series B* **24**, 251–263.

[15] Darroch, J.N. (1971). A definition of independence for bounded-sum, nonnegative, integer-valued variables, *Biometrika* **58**, 357–368.

[16] Darroch, J.N. (1986). Quasi-symmetry, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 469–473.

[17] Darroch, J.N. & McCloud, P.I. (1990). Separating two sources of dependence in repeated influenza outbreaks, *Biometrika* **77**, 237–243.

[18] Darroch, J.N. & Ratcliff, D. (1973). Tests of F-independence with reference to quasi-independence and Waite's fingerprint data, *Biometrika* **60**, 395–402.

[19] Darroch, J.N., Lauritzen, S.L. & Speed, T.P. (1980). Markov fields and log-linear interaction models for contingency tables, *Annals of Statistics* **8**, 522–539.

[20] Darroch, J.N., Fienberg, S.E., Glonek, G. & Junker, B. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability, *Journal of the American Statistical Association* **88**, 1137–1148.

[21] Duncan, O.D. (1983). Rasch measurement: further examples and discussion, in *Survey Measurement of Subjective Phenomena*, Vol. 2, C.F. Turner & E. Martin, eds. Russell Sage, New York, Chapter 12, pp. 367–403.

[22] Edwards, D.E. & Havranek, T. (1985). A fast procedure for model search in multidimensional contingency tables, *Biometrika* **72**, 339–351.

[23] Epstein, L.D. & Fienberg, S.E. (1991). Using Gibbs sampling for Bayesian inference in multidimensional contingency tables, in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, E.M. Keramidas & S.M. Kaufman, eds. pp. 215–223.

[24] Epstein, L.D. & Fienberg, S.E. (1992). Bayesian estimation in multidimensional contingency tables, in *Bayesian*

Analysis in Statistics and Econometrics, P.K. Goel, & N.S. Iyengar, eds. Lecture Notes in Statistics, Vol. 75, Springer-Verlag, New York, pp. 27–41.

[25] Etzioni, R.D., Fienberg, S.E., Gilula, Z. & Haberman, S.J. (1994). Statistical models for the analysis of ordered categorical data in public health and medical research, *Statistical Methods in Medical Research* **3**, 179–204.

[26] Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables, *Biometrika* **59**, 591–603.

[27] Fienberg, S.E. (1982). Contingency tables, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 161–170.

[28] Fienberg, S.E. (1992). Bibliography on capture–recapture modeling with application to census undercount adjustment, *Survey Methodology* **18**, 143–154.

[29] Fienberg, S.E. (2000). Contingency tables and loglinear models: Basic theory and new developments, *Journal of the American Statistical Association* **95**, 643–647.

[30] Fienberg, S.E. & Holland, P.W. (1972). On the choice of flattening constants for estimating multinomial probabilities, *Journal of Multivariate Analysis* **2**, 127–134.

[31] Fienberg, S.E. & Holland, P.W. (1973). Simultaneous estimation of multinomial cell probabilities, *Journal of the American Statistical Association* **68**, 683–691.

[32] Fienberg, S.E. & Meyer, M.M. (1983). Loglinear models and categorical data analysis with psychometric and econometric applications, *Journal of Econometrics* **22**, 191–214.

[33] Fienberg, S.E., Johnson, M., and Junker, B. (1999). Classical multi-level and Bayesian approaches to population size estimation using data from multiple lists, *Journal of the royal Statistical Society, Series A* **162**, 383–406.

[34] Fisher, R.A. (1922). On the interpretation of chi-square from contingency tables, and the calculation of P, *Journal of the Royal Statistical Society* **85**, 87–94.

[35] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[36] Gelman, A. & Rubin, D.B. (1991). Simulating the posterior distribution of loglinear contingency tables, *Unpublished Technical Report*, Department of Statistics, Harvard University.

[37] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

[38] Good, I.J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables, *Annals of Mathematical Statistics* **34**, 911–934.

[39] Good, I.J. (1965). *The Estimation of Probabilities: An Essay in Modern Bayesian Methods*. MIT Press, Cambridge, Mass.

[40] Good, I.J. (1967). A Bayesian significance test for multinomial distributions (with discussion), *Journal of the Royal Statistical Society, Series B* **29**, 399–431.

[41] Goodman, L.A. (1963). On methods for comparing contingency tables, *Journal of the Royal Statistical Society, Series A* **126**, 94–108.

[42] Goodman, L.A. (1964). Simultaneous confidence limits for cross-product ratios in contingency tables, *Journal of the Royal Statistical Society, Series B* **26**, 86–102.

[43] Goodman, L.A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications, *Technometrics* **13**, 33–61.

[44] Goodman, L.A. (1978). *Analyzing Quantitative/Categorical Data*. Abt Books, Cambridge, Mass. (a collection of papers).

[45] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classification having ordered categories, *Journal of the American Statistical Association* **74**, 537–552.

[46] Goodman, L.A. (1984). *Analysis of Cross-classified Data Having Ordered Categories*. Harvard University Press, Cambridge, Mass. (a collection of papers).

[47] Grizzle, J.E., Starmer, C.F. & Koch, G.G. (1969). Analysis of categorical data by linear models, *Biometrics* **25**, 489–504.

[48] Harris, J.A. & Treloar, A.E. (1927). On a limitation in the applicability of the contingency coefficient, *Journal of the American Statistical Association* **22**, 460–472.

[49] Hook, E.B. & Regal, R.R. (1995). Capture–recapture methods in epidemiology: methods and limitations, *Epidemiological Reviews* **17**, 243–264.

[50] International Working Group for Disease Monitoring and Forecasting (1995). Capture–recapture and multiple-record systems estimation I: history and theoretical development, *American Journal of Epidemiology* **142**, 1047–1058.

[51] International Working Group for Disease Monitoring and Forecasting (1995). Capture–recapture and multiple-record systems estimation II: applications in human diseases, *American Journal of Epidemiology* **142**, 1059–1068.

[52] Knuiman, M.W. & Speed, T.P. (1988). Incorporating prior information into the analysis of contingency tables, *Biometrics* **44**, 1061–1071.

[53] Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables, *Biometrika* **65**, 581–590.

[54] Lancaster, H.O. (1969). *The Chi-Squared Distribution*, Wiley, New York, Chapters 11 and 12.

[55] Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables, *Journal of the Royal Statistical Society, Series B* **37**, 23–37.

[56] Lindley, D.V. & Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 1–42.

[57] Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of the American Statistical Association* **89**, 1535–1546.

[58] Nazaret, A. (1987). Bayesian log linear estimates for three-way contingency tables, *Biometrika* **74**, 401–410.

[59] Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine, 5th Series* **50**, 157–175.

[60] Pearson, K. (1900). Mathematical contributions to the theory of evolution in the inheritance of characters not capable of exact quantitative measurement, VIII, *Philosophical Transactions of the Royal Society of London, Series A* **195**, 79–150.

[61] Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation, *Draper's Company Research Memoirs, Biometric Series I*, 1–35.

[62] Pearson, K. (1930). On the theory of contingency. Note on Professor J. Arthur Harris' papers on the limitation in the applicability of the contingency coefficient, *Journal of the American Statistical Association* **25**, 320–323.

[63] Pearson, K. & Heron, D. (1913). On theories of association, *Biometrika* **9**, 159–315.

[64] Quetelet, M.A. (1849). *Letters Addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha on the Theory of Probabilities as Applied to the Moral and Political Sciences* (translated from the French by Olinthus Gregory Downs). Charles and Edwin Layton, London.

[65] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The Danish Institute of Educational Research; expanded Ed. (1980), The University of Chicago Press, Chicago.

[66] Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.

[67] Roy, S.N. & Kastenbaum, M.A. (1956). On the hypothesis of no "interaction" in a multiway contingency table, *Annals of Mathematical Statistics* **27**, 749–757.

[68] Stigler, S. (1992). Studies in the history of probability and statistics XLIII. Karl Pearson and quasi-independence, *Biometrika* **79**, 563–575.

[69] Stokes, M.E., Davis, C.S. & Koch, G.G. (1995). *Categorical Data Analysis Using the SAS System*. SAS Institute, Cary.

[70] Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model, *Scandinavian Journal of Statistics* **9**, 23–30.

[71] Waite, H. (1915). Association of fingerprints, *Biometrika* **10**, 421–478.

[72] Yule, G.U. (1900). On the association of attributes in statistics: with illustration from the material of the childhood society, &c., *Philosophical Transactions of the Royal Society of London, Series A* **194**, 257–319.

[73] Zaslavsky, A.M. & Wolfgang, G.S. (1993). Triple-system modeling of census, post-enumeration survey, and administrative-list data, *Journal of Business Economics and Statistics* **11**, 279–288.

ALICIA L. CARRIQUIRY &
STEPHEN E. FIENBERG

Lognormal Distribution

The lognormal distribution is one of the most commonly used distributions for modeling the data arising in biostatistical studies.

The **random variable** X has a two-parameter lognormal distribution with parameters μ and σ^2 if $Y = \ln X$ has a **normal distribution** with mean μ and variance σ^2. The probability density function of the lognormal distribution is

$$f(x) = \frac{1}{x\sigma(2\pi)^{1/2}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right],$$
$$x > 0, -\infty < \mu < \infty, \quad \sigma > 0,$$

where σ is called the shape parameter.

The three-parameter lognormal distribution, a generalization of the two-parameter lognormal distribution, is obtained when X in the above definition is replaced by $(X - c)$ with $x > c$, where c is any real number (see [10] for further details). c is called the location parameter.

Properties

The properties of the two-parameter lognormal distribution are:

1. Mean $= \exp(\mu + \sigma^2/2)$.
2. Variance $= \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$.
3. Median $= \exp(\mu)$.
4. Mode $= \exp(\mu - \sigma^2)$.
5. The 100 pth percentile **(quantile)** $= \exp(\mu + z_p\sigma^2)$, where z_p is the pth percentile of the standard normal distribution.
6. The standardized lognormal distribution tends to the standard normal distribution as σ tends to zero.
7. The **moment generating function** of the lognormal distribution does not exist.

For more properties of the two- and three-parameter lognormal distributions, see [6] and [10].

Estimation of Parameters

The estimation of parameters μ and σ^2 of the two-parameter lognormal distribution, in general, follows from the above logarithmic transformation (of the data) and related estimation methods for the normal distribution. We refer the reader to [19] and [13] for an extensive discussion of statistical inference for the two-parameter lognormal distribution. The methods of estimation for the three-parameter lognormal and the truncated lognormal distributions are more complicated [4, 5]. The estimation in the presence of **censored data** is discussed in [4].

Applications

Aitchison & Brown [1] provide an extensive discussion of early history, the geneses and applications of lognormal distributions. Koch [11, 12] has discussed the geneses of lognormal distributions arising from biological and pharmacological mechanisms; for example, he considered the lognormal distribution for modeling the metabolic turnover. Many applications of the lognormal distributions in biochemistry are discussed in [15] and its references.

The lognormal distributions are useful for modeling data arising in many medical studies. The *hazard* function of the lognormal distribution first increases and then decreases. In many cancer studies the lognormal distribution is used as a survival distribution [2, 8, 9, 20] (*see* **Parametric Models in Survival Analysis**).

Lawrence [14] and the references cited in his paper provide an extensive review of the applications of the lognormal distribution in medical studies such as the incubation period of disease, the time to recovery, and duration of survival.

The delta–lognormal distribution, a variant of a lognormal distribution, appears in the analysis of ichthyoplankton data [17]. The **Poisson** mixture using the lognormal distribution arises in many ecologic studies such as the analyses of species frequency data [3, 8] (*see* **Contagious Distributions**). For more applications of the lognormal distribution arising in ecologic studies, see [7].

Mosimann & Campbell [16] discuss the lognormal distribution as a model for tissue growth. In the same paper, they also discuss the uses of the multivariate lognormal distribution for size and shape analyses arising in **allometry** studies.

References

[1] Aitchison, J. & Brown, J.A.C. (1957). *The Lognormal Distribution*. Cambridge University Press, Cambridge.

[2] Bennett, S. (1983). Log-logistic regression models for survival data, *Applied Statistics* **32**, 165–171.

[3] Bliss, C.I. (1966). An analysis of some insect trap records, *Sankhyā, Series A* **28**, 123–136.

[4] Cohen, A.C. (1988). Censored, truncated, and grouped estimation, in *Lognormal Distributions, Theory and Applications*, E.L. Crow & K. Shimizu, eds. Marcel Dekker, New York, pp. 139–172.

[5] Cohen, A.C. (1988). Three-parameter estimation, in *Lognormal Distributions, Theory and Applications*, E.L. Crow & K. Shimizu, eds. Marcel Dekker, New York, pp. 113–137.

[6] Crow, E.L. & Shimizu, K. eds (1988). *Lognormal Distributions, Theory and Applications*. Marcel Dekker, New York.

[7] Dennis, B. & Patil, G.P. (1988). Applications in ecology, in *Lognormal Distributions, Theory and Applications*, E.L. Crow & K. Shimizu, eds. Marcel Dekker, New York, pp. 303–330.

[8] Farewell, V.T. & Prentice, R.L. (1979). A study of distributional shape in life testing, *Technometrics* **19**, 69–75.

[9] Feinleib, M. (1960). A method for analysing log-normally distributed survival data with incomplete follow-up, *Journal of the American Statistical Association* **55**, 534–545.

[10] Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 1. Wiley, New York, Chapter 14.

[11] Koch, A.L. (1966). The logarithm in biology: I. Mechanisms generating the log-normal distribution, *Journal of Theoretical Biology* **12**, 276–290.

[12] Koch, A.L. (1969). The logarithm in biology: II. Distributions simulating the log-normal, *Journal of Theoretical Biology* **23**, 251–268.

[13] Land, C.E. (1988). Hypothesis tests and interval estimates, in *Lognormal Distributions, Theory and Applications*, E.L. Crow & K. Shimizu, eds. Marcel Dekker, New York, pp. 87–112.

[14] Lawrence, R.J. (1988). The lognormal as event-time distribution, in *Lognormal Distributions, Theory and Applications*, E.L. Crow & K. Shimizu, eds. Marcel Dekker, New York, pp. 211–266.

[15] Masuyama, M. (1984). A measure of biochemical individual variability, *Biomedical Journal* **26**, 337–346.

[16] Mosimann, J.E. & Campbell, G. (1988). Applications in biology: simple growth models, in *Lognormal Distributions, Theory and Applications*, E.L. Crow & K. Shimizu, eds. Marcel Dekker, New York, pp. 287–302.

[17] Pennington, M. (1983). Efficient estimators of abundance, for fish and plankton surveys, *Biometrics* **39**, 281–286.

[18] Preston, F.W. (1948). The commonness, and rarity, of species, *Ecology* **29**, 254–283.

[19] Shimizu, K. (1988). Point estimation, in *Lognormal Distributions, Theory and Applications*, E.L. Crow & K. Shimizu, eds. Marcel Dekker, New York, pp. 27–86.

[20] Whittemore, A. & Altschuler, B. (1976). Lung cancer incidence in cigarette smokers: further analysis of Doll and Hill's data for British physicians, *Biometrics* **32**, 805–816.

M. RATNAPARKHI

Logrank Test

The *logrank test* (so named by Peto & Peto [1]) is a rank test for comparing two samples of right-censored survival data. A careful description of its several origins, history, and connection to other tests for one-, two- and k-sample problems in censored survival data is given in the article **Linear Rank Tests in Survival Analysis**. Here we merely define the test in the simplest situation.

Let $\tilde{X}_{hi}$, for $i = 1, \ldots, n_h, h = 1, 2$, be independent nonnegative **random variables** with absolutely continuous distribution function F_h and **hazard rate** α_h. We do not observe the $\tilde{X}_{hi}$ but rather the right-censored samples (X_{hi}, D_{hi}), $X_{hi} = \tilde{X}_{hi} \wedge U_{hi}$, and $D_{hi} = I\{X_{hi} = \tilde{X}_{hi}\}$ for some censoring times $U_{hi}, i = 1, \ldots, n_h, h = 1, 2$. It is assumed that there is *independent censoring* (*see* **Censored Data**), which would be true if the U_{hi} were independent random variables, independent of the $\tilde{X}_{hi}$. In accordance with the assumption of absolutely continuous distributions, we assume that all $\tilde{X}_{hi}$ are distinct (no ties). *See* **Tied Survival Times** for further generalization.

The (two-sample) logrank test tests the hypothesis $\mathrm{H}_0 : F_1 = F_2$ (or equivalently $\alpha_1 = \alpha_2$) by comparing the observed number of events in group 1,

$$O_1 = \sum_{i=1}^{n_1} D_{hi},$$

with the so-called expected number of events in group 1 under H_0 (*see* **Expected Number of Deaths**). The latter is estimated as

$$E_1 = \sum_{h=1}^{2} \sum_{i=1}^{n_h} D_{hi} \frac{Y_1(X_{hi})}{Y_1(X_{hi}) + Y_2(X_{hi})},$$

where $Y_h(t) = \sum_{k=1}^{n_h} I\{X_{hk} \geq t\}$ is the *number at risk* in group h at time t (*see* **Risk Set**). Indeed, it may be shown (still assuming no ties) that in large samples, $(O_1 - E_1)/\sqrt{V}_1$ is asymptotically **standard normal**, with

$$V_1 = \sum_{h=1}^{2} \sum_{i=1}^{n_h} D_{hi} \frac{Y_1(X_{hi}) Y_2(X_{hi})}{(Y_1(X_{hi}) + Y_2(X_{hi}))^2}.$$

Reference

[1] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series A* **135**, 195–206.

(*See also* **Survival Analysis, Overview**)

NIELS KEIDING

Lomb Periodogram

The Lomb periodogram is a generalization of the periodogram (*see* **Spectral Analysis**) for unequally spaced series. Many biomedical time series are sampled irregularly because of missing data due to instrumental failures or because the nature of the measured variable makes the sampling interval intrinsically uneven. Since traditional spectral estimators need even sampling rates, unevenly spaced series should be interpolated and resampled before spectral analysis. By contrast, the method proposed by Lomb does not require interpolation and resampling [4]. Given N data $[y_i]$ sampled at times t_i, the normalized Lomb periodogram $P(\omega)$ (with ω the angular frequency $2\pi f$) is:

$$P(\omega) = \frac{1}{2\sigma^2}\left\{\frac{\left[\sum_{i=1}^{N}(y_i - \overline{y}) \times \cos\omega(t_n - \tau(\omega))\right]^2}{\sum_{i=1}^{N}\cos^2\omega(t_n - \tau(\omega))} + \frac{\left[\sum_{i=1}^{N}(y_i - \overline{y}) \times \sin\omega(t_n - \tau(\omega))\right]^2}{\sum_{i=1}^{N}\sin^2\omega(t_n - \tau(\omega))}\right\}, \quad (1)$$

where $\overline{y} = (1/N)\sum_1^N y_i$ and $\sigma^2 = (1/(N-1))\sum_1^N (y_i - \overline{y})^2$ are the mean and the variance of the data and

$$\tau(\omega) = \frac{1}{2\omega}\arctan\left\{\frac{\sum_{i=1}^{N}\sin 2\omega t_i}{\sum_{i=1}^{N}\cos 2\omega t_i}\right\}$$

is an offset that makes $P(\omega)$ invariant to time translation. An implementation in code by a fast algorithm can be found in [5]. Equation (1) defines a *normalized* periodogram because of the term σ^2 in the denominator. Scargle showed that with this normalization, $P(\omega)$ of a white Gaussian noise approximately follows an **exponential** probability distribution with unit mean [6]. A statistical test for detecting a periodicity in the data derives from this property. Given M independent spectral lines, if the highest periodogram ordinate exceeds the critical value $P_\alpha = -\ln(1-(1-\alpha)^{1/M})$, then it is significant at the false-alarm probability α, that is, the null hypothesis that the time series is a white Gaussian noise is rejected. A procedure for searching multiple periodicities is reported in [8].

The performances of the Lomb periodogram were extensively analyzed in studies assessing the heart-rate variability [3], the respiratory arrhythmia in neonates [2], and the **circadian** rhythms in oral temperature and urinary cortisol secretion [8]. This method shows some limitations when the data contain fractions of non-Gaussian noise or periodic signals with nonsinusoidal shapes [7] or when the sampling rate is not random, but depends on the value of the signal [1].

References

[1] Castiglioni, P. & Di Rienzo, M. (1996). On the evaluation of heart rate spectra: the Lomb Periodogram, *Computers in Cardiology 1996*. IEEE, Piscataway NJ, 505–508.

[2] Chang, K.L., Monahan, K.J., Griffin, M.P., Lake, D. & Moorman, J.R. (2001). Comparison and clinical application of frequency domain methods in analysis of neonatal heart rate time series, *Annals of Biomedical Engineering* **29**, 764–774.

[3] Laguna, P., Moody, G.B. & Mark, R.G. (1998). Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals, *IEEE Transactions on Biomedical Engineering* **45**, 698–715.

[4] Lomb, N.R. (1976). Least-squares frequency analysis of unequally spaced data, *Astrophysics and Space Science* **39**, 447–462.

[5] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, New York.

[6] Scargle, J.D. (1982). Studies in astronomical time series analysis. II Statistical aspects of spectral analysis of unevenly spaced data, *Astrophysical Journal* **263**, 835–853.

[7] Schimmel, M. (2001). Emphasizing difficulties in the detection of rhythms with Lomb-Scargle Periodograms, *Biological Rhythm Research* **32**, 341–345.

[8] Van Dongen, H.P.A., Olofsen, E., VanHartevel, J.H. & Kruyt, E.W. (1999). A procedure of multiple period searching in unequally spaced time-series with the Lomb-Scargle method, *Biological Rhythm Research* **30**, 149–177.

PAOLO CASTIGLIONI

Longitudinal Data Analysis, Overview

Longitudinal data arise when each member of one or more *cohorts* or *panels* of subjects provides a measurement on a number of occasions [21]. The cohort may be defined, for example, in terms of the date of birth of its members, the time of onset of a disease or, in the case of a clinical trial, the beginning of treatment or time of randomization. The *repeated (serial) measures* might be quantitative or qualitative, and may also be multivariate. For simplicity, however, the present discussion will be limited to univariate measures. Together, the results of these measurements will form a *response profile* (particular examples being *growth curves* (*see* **Nonlinear Growth Curve**) and time *trends* arising from **pharmacokinetic experiments**). Typically, longitudinal or repeated measures data are collected prospectively, but it is also possible to collect them retrospectively through the use of medical records, for example.

Time series data are similar to those described in this section but, on the whole, they can be distinguished from the latter because they usually arise from a single or, at most, a few extended sequences of observations as opposed to a larger number of shorter sequences. **Survival data**, *event history data* (*see* **Repeated Events**) *multistate* (e.g. states of well-being, morbidity, and death) *transition data* (*see*, **Fix–Neyman Process**) and **competing risks data** are also similar to repeated measures data in that they all involve observation over time. However, instead of enquiring about the state of a patient at each of a series of discrete times, investigators interested in survival times, for example, usually aim to record the exact date of death of each of the patients (i.e. the time of death is, in theory, a continuous rather than a discrete variable). The methods of analysis required for such data are distinctive, often involving time as an explicit response variable, and are covered for example, in **Survival Analysis, Overview**. In practice, many event recording systems do not record continuously but only to within discrete intervals of time, and continuously recorded data can be well approximated by grouping over short intervals of time. Such discrete event history or survival data can then be considered as a series of repeated qualitative measurements on each subject and analyzed using the methods of this article in which time is a covariate or design factor [44].

Studies may involve more than one timescale. For example, treatment studies often consider both time under treatment and subject's age, and multistate transition processes may involve effects due to time since entry to the current state and the cumulative time spent in that and other states. Care may be required to insure that effects on each scale are all **identifiable**. The difficulties posed by **age–period–cohort** effects are a well known example.

Longitudinal studies may be observational (e.g. **cohort studies** epidemiologic surveys) or experimental (e.g. controlled **clinical trials**). In a clinical trial, the treatment might remain constant for any particular cohort or group of patients (with random allocation of patients to the competing treatments) or vary from one occasion to another (with random allocation to groups defined by the order in which the treatments are received). In the case of the latter, the trial is an example of the use of a **crossover design**.

The simplest kind of longitudinal study involves taking measurements on all subjects at the same times: that is, each patient provides exactly the same set of measures. It is possible, however, for both the number and spacing of the repeated measures to vary from one subject to another. The latter may arise from the design of the study, but in addition may be due to unintentionally missing observations. Patients might, for example, fail to keep an appointment on a given date, might be too ill to be interviewed, or might permanently drop out of or be lost from the study through a variety of causes (e.g. death, emigration, or refusal to continue treatment). The various approaches to the statistical analysis of longitudinal data differ in their ability to cope with missing data and in the assumptions made concerning the mechanism by which the missing data might arise. Missing data are a challenge to valid inference from longitudinal studies (*see* **Diggle–Kenward Model for Dropouts**; **Nonignorable Dropout in Longitudinal Studies**). (For details of modeling missing data mechanisms, see, for example, Little [45] or Diggle & Kenward [14].) Investigators should minimize the occurrence of missing values, avoiding them altogether wherever possible. If missing values are inevitable, then investigators should collect as much information as possible about the reasons for the missing data and to try to incorporate this information in their analysis.

Examples of Longitudinal Studies

First, let us consider experimental studies. Frison & Pocock [26] describe a clinical trial in which 152 patients with heart disease were randomly allocated to treatment using an active drug or a placebo during a 12-month follow-up period. The concentration of the liver enzyme creatine phosphokinase (CPK) in the patients' serum was measured as an indicator of liver damage arising as a side effect of the treatment. Each patient had three pretreatment measurements which were taken at 2 months before, 1 month before, and at the time of randomization. They also had eight posttreatment measurements taken every 1.5 months after randomization. An example of a simple crossover trial is provided by Hills & Armitage [32]. The experiment was a comparison of the effects of an active drug and a placebo in the treatment of enuresis. One group of patients received 14 consecutive days of treatment with the active treatment, followed by a similar period of treatment using the placebo. A second group received the treatment combinations in the reverse order: placebo followed by active drug. The response variable was the number of dry nights out of 14: that is, each patient provided two measures – one corresponding to each of the two periods of treatment.

Longitudinal surveys are also common in medical research. Here we describe three longitudinal studies of lung function. Laird & Ware [39], for example, describe a survey in which pulmonary function in about 200 school children was examined under normal conditions, then during an air pollution alert and on three successive weeks following the alert. The main aim of the study was to determine whether the volume of air exhaled in the first second of a forced exhalation (FEV_1) was depressed during the alert. The analysis of repeated **categorical** measures has been illustrated by Ware et al. [55]. Children were assessed annually at ages 9–12 to evaluate the potential effects of air pollution on persistent wheeze. Parents were asked about wheezing by their children during the previous year and responses were grouped into three mutually exclusive categories or states: no wheeze, wheeze with colds, or wheeze apart from colds. Our final example concerns a survey with many missing observations. Lavange & Helms [41] analyzed data from a study of 72 children aged from 3 to 12 years. These data are also discussed by Little [45]. A measure of maximum expiratory flow rate was obtained annually, and differences in the resulting growth curve were related to the sex and race of the children. The number of actual measurements recorded on each child ranged from 1 to 8 (with an average of 4.2). Some values were missing because the child was either older than 3 at the beginning of the study, or younger than 12 at the end of it.

In the analysis of longitudinal data, the critical feature to recognize is that, since sets of measures are obtained from the same subjects, these measures are likely to be **correlated**, and can rarely be considered as independent even after conditioning upon known predictors or **explanatory variables**. How that dependence is dealt with is a principal distinguishing feature of different methods of analysis. However, before outlining these, we now consider more preliminary examination of the data.

Graphical Displays and Data Exploration

Diggle et al. [15] give the following simple guidelines for the exploration of longitudinal data using **graphical displays**:

1. show as much of the relevant data as possible rather than data summaries;
2. highlight aggregate patterns of potential scientific interest;
3. identify both **cross-sectional** and longitudinal patterns in the data;
4. make easy the identification of unusual individuals or unusual observations.

Here we produce a few simple plots for data on salsolinol levels (Table 1). These data were collected during an investigation into the role that the alkaloid salsolinol plays in bodily dependence on alcohol [30]. Fourteen individuals attending an alcohol treatment unit were observed over a period of four days immediately after being admitted to the unit, measurement of salsolinol being made from urine samples taken daily throughout the study period. The individuals were categorized as being in one of two groups: those considered to be severely dependent and those judged to be only moderately dependent. The response variables for the study are the four repeated measurements of urine concentrations of salsolinol. First, box plots (*see* **Graphical Displays**) of the distributions of the measurements at any one time point (see

Table 1 Salsolinol concentrations on four successive days

Obs.	Group	Day 1	Day 2	Day 3	Day 4
1	2	0.64	0.70	1.00	1.40
2	1	0.33	0.70	2.33	3.20
3	2	0.73	1.85	3.60	2.60
4	2	0.70	4.20	7.30	5.40
5	2	0.40	1.60	1.40	7.10
6	2	2.60	1.30	0.70	0.70
7	2	7.80	1.20	2.60	1.80
8	1	5.30	0.90	1.80	0.70
9	1	2.50	2.10	1.12	1.01
10	2	1.90	1.30	4.40	2.80
11	1	0.98	0.32	3.91	0.66
12	1	0.39	0.69	0.73	2.45
13	1	0.31	6.34	0.63	3.86
14	2	0.50	0.40	1.10	8.10

Source: Hand & Taylor [30].

Figure 1) indicates **skewness**. A logarithmic **transformation** (base ten) of the salsolinol concentrations was therefore carried out prior to any further analysis. We next plot the time course for each individual subject, distinguishing the subjects from each of the two alcohol dependency groups (Figure 2).

In Figure 3 is shown a plot of the mean values of the logged salsolinol levels for the two groups, together with their **standard errors**. An alternative would have been to plot a series of box plots, perhaps revealing more information about between-subject differences at each of the time points. Figure 3 highlights the difference between the two groups in rates of change over time, although the main message appears to be that the groups are, in fact, very

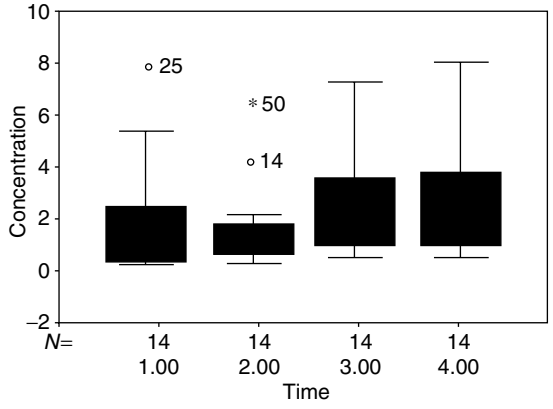

Figure 1 Salsolinol concentrations over time

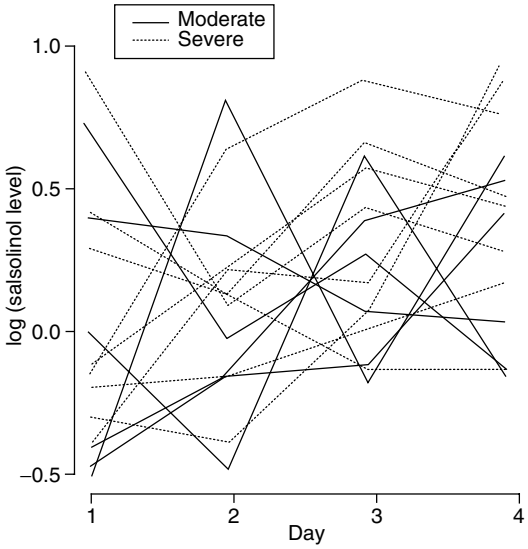

Figure 2 Salsolinol data – individual profiles after log transformation

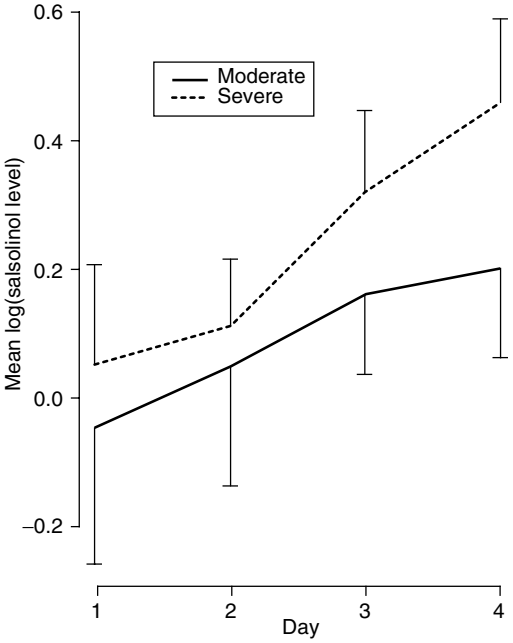

Figure 3 Mean profiles of salsolinol levels after log transformation for severe and moderate alcohol dependent groups

similar. Although graphs of means such as that found in Figure 3 (and, less often, box plots) are much more commonly seen than those for the response

profiles for individual subjects, great care must be taken in their interpretation. Figure 3 hides the pattern of *within-subject* changes. A graph of the latter type might also be very misleading if there were increasing numbers of dropouts over time, with the plotted means being calculated from the survivors at each time. If the dropping out is any way dependent on the present or previous state of the subject then the later means will be **biased**. One way of avoiding this bias is to plot means derived from cases with complete data, but the latter approach might be very inefficient if there are lots of dropouts.

Plots such as those provided in Figures 2 and 3 might indicate how one might extract suitable summary measures for each subject for a subsequent simple analyses of these response features. Visualizing the patterns in the data, and the subsequent extraction of the required response features, might be aided by smoothing each of the individual time courses. A search for the time of maximum response in a pharmacokinetic experiment, for example, might be quite difficult in the presence of considerable within-subject "noise". An example of smoothing in a pharmacokinetic experiment using a **moving average** prior to response feature extraction (the time of maximal response) is provided in Durcan et al. [18]. Other applications of smoothing methods, together with examples of their use, are described in Diggle et al. [15].

Another possibility is a simple multiple scatter plot (ignoring group differences) for the logged salsolinol concentrations. This sort of plot is ideal for the exploration of the correlation structure of repeated measures, although in a more complex data set with greater group differences, it would be preferable to remove the effects of explanatory variables and produce plots using the **residuals**. The results are not presented here, because there seems to be very little evidence of **serial correlation**. Finally, a plot that can be helpful in revealing the relative magnitudes of the sources of variance that give rise to correlations in continuous measures over time is the **variogram**.

Methods of Analysis for Continuous Responses

This section will be concerned with a few of the more commonly used strategies for the analysis of longitudinal data, with particular reference to continuous (usually **normally distributed**) outcome measures. We assume that the primary interest lies in changes in the average response over time at different levels of various explanatory factors, taking into account possible dependencies during **hypothesis testing** and **estimation**. The technical aspects of the methods will not be discussed in any detail but will be covered elsewhere. Methods for the analysis of categorical responses, transitions, and responses in the form of counts will be covered briefly in the next section.

Multivariate Generalizations of Paired t Tests

The **paired *t* test** is one of the basic methods for analyzing a simple two period, pretest/posttest study, comparing an estimate of the simple time 1 − time 2 difference contrast with the variance of this estimate. For greater numbers of measurement occasions, the multivariate generalization of this test is **multivariate analysis of variance** (MANOVA), that extends this approach to various linear contrasts relating to different aspects of change. As the paired *t* test can be generalized to an analysis of change scores, in which differences are regressed against factors and covariates, so too can this be done within MANOVA, in the form of *multivariate analysis of covariance (MANCOVA)*. These procedures enable one to test the differences between vectors of means with an entirely arbitrary pattern of correlations between the repeated measures. This method is therefore not dependent on any unrealistic assumptions concerning the patterns of serial dependencies, but is also likely to be less powerful than more refined methods that explicitly acknowledge the serial nature of the observations and correctly model the dependencies between them. These methods fail altogether when there are more design cells than subjects, a common occurrence where there are numerous measurement occasions.

Autoregression and Ante-Dependence

Another standard method for dealing with the correlation in the responses from a simple two-period study is **analysis of covariance** – analyzing the time-2 response conditional upon the time-1 response and predictors of change. This approach too can be generalized to larger numbers of measurement occasions. The **autocorrelation** or *autocovariance* between responses can either be considered as a nuisance to be allowed for in an analysis or, in

some applications, it can be regarded as the property of particular interest. Consider a possible model for serial dependencies between repeated quantitative measurements. Let $e_1, e_2, \ldots, e_T$ be uncorrelated random variables, where e_t has a mean of zero and variance σ_t^2 for $t = 1, 2, \ldots, T$. Now define a series of measurements $Y_1, Y_2, \ldots, Y_T$ by

$$Y_1 = e_1,$$
$$Y_t = \gamma_t Y_{t-1} + e_t, \quad t = 2, 3, \ldots, T.$$

If the γs are all equal ($\gamma_t = \gamma$ for all t) and so are the variance terms ($\sigma_t^2 = \sigma^2$ for all t), then these equations describe a stationary first-order *autoregressive process* (*see* **ARMA and ARIMA Models**). If, however, these parameters are permitted to vary with time then the equations describe first-order **ante-dependence**. A measurement at time t (that is Y_t) is dependent on the value of Y_{t-1} but, conditional on the value of Y_{t-1}, it is independent of all previous measurements. In general, a set of ordered measurements $Y_1, Y_2, \ldots, Y_T$ is said to have an independence structure of order r if the measurement at time t (with t greater than r), given the preceding r measurements, is independent of all further preceding measurements. Kenward [35] describes a method of assessing the order of a sequence of observations.

A simple unrestricted autoregressive model will involve $T(T+1)/2$ variance and covariance parameters, as would the equivalent MANOVA. An advantage of the ante-dependence approach is that it provides a simple but flexible path for specifying more restrictive dependencies, increasing efficiency for the testing of contrasts of interest and giving the capability of analyzing studies with few subjects and numerous measurement occasions.

In the analysis of a longitudinal data set, one can approach the problem of serial dependencies from several points of view. If we ignore their possible existence the analysis will be simpler, but the resulting inferences are likely to be invalid. If we can replace the response profile for each subject by one or possibly more summary statistics (*derived variables*) which extract the distinct features of interest then the problem is side-stepped. Any resulting analysis of these extracted *response features* will be unaffected by the serial dependencies in the original observations. Again, we might also choose to modify our original approach to analyze the data as if there were no serial dependencies (as in the traditional **analysis of variance** for a *nested* or **split-plot** experiment – the repeated measures being nested within subjects) but then to make adjustments to the resulting test statistics (or their degrees of freedom) to allow for them. This is the rationale for the well-known *Greenhouse–Geisser adjustment* (see [29]) (*see* **Analysis of Variance for Longitudinal Data**).

The more refined methods will be more difficult to carry out and interpret and, more importantly, will not necessarily be robust to an incorrect specification of these serial dependencies. Great care must be exercised in their use.

A related problem concerning the analysis of data from crossover studies is the possible presence of *carryover effects*. In the simplest design – the two-period, two-treatment crossover experiment – this is completely confounded with the treatment by period interaction or order effect. A carryover effect arises when an effect of an early treatment persists in later periods of the trial. This might be due to an inadequate washout period between two periods of chemotherapy, for example, or because the first treatment has induced some permanent change in the patient. Many authors have suggested that this design should only be used when it can be assumed a priori that such carryover effects are absent. Crossover designs should only be used when the short-term relief of chronic symptoms, rather than a cure, is the goal of the trial (examples being the use of lithium in the control of manic symptoms, or the use of insulin to control blood sugar levels).

Time-by-time Analysis

Following the common practice of plotting of group means for each separate time point, it comes as no surprise to find that investigators very frequently carry out separate statistical analyses at each of the time points. If there are n time points being considered, then there will be n separate analyses. On the whole, this is not a method of analysis that should be encouraged – it lacks power and the repeated tests are not statistically independent – although Finney [24] has advocated this **time-by-time analysis** when the number of times is small and the intervals between them large. Quite often, the researcher is interested in the question "At what time do the groups become significantly different?" and this is frequently the motivation for time-by-time analyses. If the latter is the case, then a modification of the approach by Kenward

[35] might be preferred. This is essentially a series of analyses of covariance looking at group differences at any given time point, typically using the previous one or two values as covariates. The method is based on assumptions concerning the ante-dependence structure of the data and the reader is referred to Kenward [35] for technical details. Examples of the use of Kenward's method under the assumption of second-order ante-dependence can be found in Crowder & Hand [13], and for the analysis of the salsolinol data under the assumption of first order ante-dependence in Everitt & Dunn [22].

Derived Variables

Inspection of the individual time courses in Figure 3 leads naturally to two related ideas. The first is to ask what summary statistic or derived variable might be extracted for each case which best describes the main feature of interest in the serial measurements. There may, however, be more than one feature of interest in a series of repeated measures. For our salsolinol data, for example, the two which immediately come to mind are the average of the four measurements for each individual and an overall rate of change (linear trend) for that individual. The second idea is based on fitting a separate **regression** model (growth curve) to each case. One might, for instance, use ordinary **least squares** to fit a straight line to each individual's data. The resulting estimates for the intercept term and slope parameter would, of course, convey the same information as the derived variables from the first approach, but they do suggest that one might extend the idea to the fitting of some sort of **multilevel** or **random effects** model to the data. This will be developed in the following subsection. Here we deal with the analysis of derived variables.

Having obtained the derived variables, we then enter them into a second stage of analysis to estimate their mean for two or more groups and to test for possible differences between these groups. Returning to the salsolinol data, it is in fact possible to derive a mean for the concentrations at the four times and three orthogonal polynomial trends (that is, linear, quadratic, and cubic trends) (*see* **Orthogonality**; **Polynomial Regression**). Differences in the means and trends across groups can be tested using simple t tests (or, in general, using ANOVA models) – each of the four derived variables being analyzed separately. Alternatively, we might wish to test for differences in all three trends simultaneously using **Hotelling's T^2 statistic** or, more generally, through the use of multivariate analysis of variance (MANOVA) procedures. One could, of course, include the average over time in this multivariate test, but this is usually analyzed separately so that one carries out separate analyses for the overall level and for the pattern of temporal change.

Random Effects Models

Returning once more to the salsolinol measurements, let Y_{ijk} represent the logarithm of the salsolinol concentration for the jth subject in the ith group on the kth day. Note that subjects are nested within groups. A possible regression model to describe the whole data set is

$$Y_{ijk} = \beta_0 + \alpha_i + \omega_{ij} + (\beta_i + \beta_{ij})t_k + \varepsilon_{ijk},$$

where t_k is the time to the kth measurement, and the parameters β_0, α_i, and β_i (the so-called fixed effects) correspond to the intercept term, the effect of being in the ith group on the intercept ($i = 1, 2$) and the linear effect of time in the ith group, respectively. The random effects are ω_{ij}, the effect on the intercept of subject j within group i, β_{ij}, the variation of the linear effect of time which is characteristic of subject j within the ith group, and the residual "error" term ε_{ijk}. In terms of the derived variables described above, the linear trend for the ijth individual is equivalent to the estimate of $\beta_i + \beta_{ij}$, but note that we are not now interested in estimating it explicitly – only its variance and possibly covariance with other effects. The random effects are all assumed to have zero expectation and the effects of real interest to the investigator are the β_is and possibly the α_is. Assuming that the responses are conditionally **multivariate normal**, we can then use **maximum likelihood** to estimate the fixed effects and the variances and covariances of the random effects, together with their respective standard errors [39].

Structural Equation and Latent Variable Models

Consider the observed variable, Y, which is now acknowledged to be measured with error. Typically,

$$Y = F + E,$$

where F is a latent variable or factor and E is the corresponding measurement error. If we now consider a series of repeated measures, $Y_t, t = 1, 2, \ldots, T$, with $Y_t = F_t + E_t$, it might be realistic to assume that the Fs are serially correlated, but that the Es are statistically independent. We also usually assume that the Fs and Es are independent. A latent first-order autoregressive model, for example, would have the form

$$Y_1 = F_1 + E_1$$
$$Y_t = \gamma F_{t-1} + E_t, \quad t = 2, 3, \ldots, T.$$

This simple latent variable model might well provide a **parsimonious** description of a series of measures when a similar first-order autoregressive model for the observed measurements would be hopeless. By acknowledging measurement error in this way, we can often considerably simplify the interpretation of the relationships within a set of serial measures. The above model (and any other) implies a particular structure for the covariance matrix of the repeated measures, and the model can therefore be fitted and its **goodness of fit** tested using covariance structure or **structural equation modeling** software (see [17], for example).

Another possibility is a random walk or Wiener model (*see* **Brownian Motion and Diffusion Processes**). Here

$$Y_1 = F_1 + E_1,$$
$$Y_t = F_1 + F_2 + \cdots + F_t + E_t, \quad t = 2, 3, \ldots, T.$$

Here the F_ts are random increments (or decrements) in the response variable Y_t which are "frozen in", accumulating over time. A third possibility is a latent growth curve model of the following form:

$$Y_1 = F_1 + E_1,$$
$$Y_2 = F_1 + F_2 + E_2,$$
$$Y_t = F_1 + \gamma_t F_2 + E_t, \quad t = 3, 4, \ldots, T.$$

In this case the two factors, F_1 and F_2, represent a baseline and a rate of growth (or decline), respectively, and it is quite usual to see that they are correlated; a relatively large child at the start of a longitudinal study, for example, also growing at a rate greater than most of the other children. The reader will note the similarity of this and the random effects model of a previous section.

Quite often, it is of interest to compare growth curves of two or more cohorts of subjects. The covariance structure software can easily deal with this by simultaneously fitting growth curve models to two or more observed **covariance** (or moments) **matrices**. One can then test for the equality of parameters of interest across the groups.

Robust Parameter Covariance Estimates

It will have become apparent that for analyzing longitudinal data, although the main interest may lie in estimating the effects of risk factors and exposures on the expected value of the response, it often seems necessary to expend more effort to ensure that the model for the variances and covariances among the response is correct. Huber [34], and subsequently White [57] and Royall [51], proposed a heteroscedastic consistent "sandwich" estimator for the parameter covariance matrix (*see* **Generalized Estimating Equations**). Variants of this covariance estimator are available among many of the software implementations of procedures described above (e.g. EQS). At the cost of reduced efficiency – often trivial but sometimes large – the use of this method provides some relief from an excessive concern that the random part need be correctly specified in every detail (*see* **Robustness**).

Methods for Responses in the Form of Counts and Categorical Responses

Multivariate distributions for categorical and count data lack the flexibility of the multivariate normal distribution that underlies many of the methods for analyzing repeated continuous responses. In general, choices of distribution that have simple expressions for marginal distributions yield unpleasant expressions for joint or conditional distributions. The statistical literature is awash with models based on various distributions and parameterizations that may cleverly fit the particular needs of the problem illustrated by the authors, but that lack generality. We consider here only methods that we believe have wide scope for application. The principal styles of analysis of repeated count and categorical data tend to focus upon one of two rather different aspects of the overall process. The methods of *survival analysis* tend to focus on issues of timing and on problems, where

the observation scheme is – at least nominally – continuous in time (*see* **Survival Analysis, Overview**). The remaining methods tend to focus rather more on *state occupancy* and *transition*, and often assume a discrete (and often equally spaced) observation scheme shared by all subjects and an analysis in which the treatment of the timescale is often implicit or simply another within subjects factor. These latter are the methods discussed here. This separation in styles of analysis is not always desirable, hampering our ability to generalize conclusions across observation and sample design schemes [4]. Methods that combine these two styles, such as *competing risk models*, are available but are typically cumbersome in use.

Contingency Tables and Loglinear Models

There is an extensive literature examining cross tabulated data from longitudinal studies of discrete outcomes, in particular making use of **loglinear models**. While in general useful as a preliminary tool, for scientific analysis of repeated measures data the interpretation of the parameters presents problems [6]. Kenward & Jones [36], for example, argue that the approach is more suited to "correlation rather than regression analysis". Discrete time Markov transition models (*see* **Markov Chains**) have received considerable attention, even though exactly how results relate to the process measured on a continuous timescale often remains open. Transition tables relating to social and economic mobility have been much studied [5]. Typically, all such tables show strong temporal associations among categories, often largely due to a tendency for simple persistence within the current class (*spurious contagion* [23]; *cumulative inertia* [47]). This has generated more specific forms of **contingency table** test for **quasi-independence** and **quasi-symmetry** (see Everitt [20]).

Latent Class Models

These inertia effects led to the exploration of *mover–stayer* models, a simple form of **latent class model** [42], in which attempts are made to explain a complex temporal association among categories by the admixture of populations each following a more simple temporal process. In this instance, the "stayer" population simply persists in the same category, while the movers might all share a uniform transition rate. Of course, latent class methods have also been applied to longitudinal data to tackle problems of misclassification. The estimation of so-called **hidden Markov chains** is a more recent interest, whether for transitions between states of psychopathology and estimated by maximum likelihood [60], or for repeated screenings for cervical cancer and estimated by Gibbs sampling [52, 58] (*see* **Markov Chain Monte Carlo**). An approach using continuous latent variables for categorical data is discussed below under GEE estimation.

Conditional or Fixed Effects Models

Although the inclusion of subject specific fixed effects as dummy variables into **logistic regression** models for repeated binary measures does not lead directly to a satisfactory form of analysis (due to the *incidental parameter* problem [50]; *see* **Estimating Functions**), an analysis conditioning on the **sufficient statistic** for such a subject-specific effect does. In the simple two-period case without covariates, this corresponds to the **McNemar test** [48]. More generally, it corresponds to a form of **conditional logistic regression** [2, 9, 10], a method familiar to those analyzing matched **case–control studies**. This approach yields estimates only of risk factors or exposures that are time-varying, and can be inefficient where there is substantial use between subject information on effects of interest.

Random Effects or Integrated Likelihood Models

At the cost of assuming subject effects to be uncorrelated with included explanatory variables, a random effects approach provides estimation of time-constant effects and more efficient estimates of time-varying effects. Assuming some distribution for subject-specific effects provides a likelihood for a sequence of discrete outcomes of the form

$$L_i = \int \prod_{j=1}^{T} h^{-1}(Y_{ij}|X_{ij}; \boldsymbol{\beta}, \tau_i)\, \mathrm{d}G(\tau).$$

In general, however, most choices of link function $h(\cdot)$ and parametric distribution $G(\cdot)$ for the subject-specific effects do not lead to an analytically tractable expression, even when the problem is simplified to one of time-constant subjects effects (the so-called

"one factor" model). Choice of the complementary log–log link together with a distribution of subject effects from the Hougaard family [33], for example the **gamma distribution**, offers some possibilities and can be combined with discrete latent classes [53]. Recourse to computational brute force – for example, using quadrature [27] or Monte Carlo methods [52] – allows the use of the potentially more flexible multivariate normal distribution for subject effects. Somewhat curiously, the computational burden becomes little greater if parametric restrictions are eased and instead the **nonparametric maximum likelihood estimator** of the random effects distribution is used [38]. In this case the distribution is represented by mass points with both weights and locations as free parameters, reducing the integration of the above equation to a summation (and almost always over fewer points than that required for "parametric quadrature"). The relationship between the nonparametric and conditional estimators is discussed in Lindsay et al. [43].

Penalized or Predictive Quasi-likelihood and Generalized Linear Mixed Models

An alternative approach to the computation of L_i is through some linearizing approximation, described as *penalized quasi-likelihood* (PQL) by Breslow & Clayton [7]. Essentially, this involves the iteratively reweighted least squares equations of standard GLM estimation [46] being extended to include current estimates of the random effects as well as those for fixed effects. For binary response data and few measurement occasions, this approach does not perform well [8, 16], particularly with respect to estimation of the random effects parameters. However, in many other circumstances it performs much better and offers a flexible and simple approach that yields satisfactory estimates for covariate effects of interest. A similar approach can be used to estimate the parameters of **marginal models** [28] that are considered in more detail in the next two sections.

Empirical Generalized Least Squares

All the preceding approaches have involved specifying some model for the covariance among observations due to the impact of subject-specific effects and past history, and estimating effects of interest conditional upon these effects. An alternative approach is to specify functional forms for the relationships of primary interest – say, the marginal relationship between outcomes and features of the study design – with the rest of the model that deals with covariances being saturated. In the *empirical generalized least squares* approach of Koch et al. [37], implemented in the SAS procedure CATMOD (*see* **Categorical Data Analysis**), the marginal expected proportions are replaced directly by their observed values to provide empirical logits, limiting this method to design matrices involving only discrete variables. These are then linearly related to explanatory variables. Since these proportions are neither independent nor equally variable, ordinary least squares estimation is not appropriate. However, the covariance matrix for these empirical logits will typically be block diagonal with one block for each unique combination of between subjects factors. The ith block is then estimated by $\mathbf{D}_i\mathbf{V}_i\mathbf{D}_i^{\mathrm{T}}$, where $\mathbf{D}_i$ is the matrix of partial derivatives of the logits with respect to the marginal proportions and $\mathbf{V}_i$ is their covariance matrix. With the need to avoid undefined empirical logits and singularities in the estimated covariance matrix, this approach has trouble with sparse data. Kenward & Jones [36] suggest 25–30 responses for each response function for reliable results, typically limiting this method to very few repeated measures. The general approach has been extended to tackle incomplete data and other response functions [40].

Estimation Using a "Working Covariance Matrix" and Generalized Estimating Equations

This powerful and flexible approach is described in the article on **Generalized Estimating Equations**. The approach represents a multivariate generalization of **quasi-likelihood** estimation, allowing a Fisher-scoring method of estimation for models for which a full likelihood may not be known. Although more commonly used to estimate marginal or population-average models, the generalized estimating equations (GEE) approach can also be used to estimate models including subject specific random effects [59].

Muthén [49] presented a related general approach that fits mixed effects and latent variable models to categorical data using a two-stage estimation method. This was based on first estimating fixed effects (thresholds and coefficients), their covariance matrix, a conditional covariance matrix of errors and their covariance matrix, all based on pairwise bivariate

probit. The second stage then fits models to these first-stage estimates. For large samples without complex patterns of missing data and with response measures of mixed type, this is a flexible and powerful method.

Marginal Maximum Likelihood Models

In fact, there are a number of parameterizations that include the marginal means as parameters and that allow closed form likelihood representations for binary sequences. Bahadur [3] described how the joint distribution of a binary sequence could be parameterized in terms of the marginal means and the marginal correlations. Estimation is, however, nonstandard in that the marginal correlations are subject to a reasonably complex set of linear inequality constraints. Fitzmaurice & Laird [25] provided a "mixed" parameterization, one involving the marginal means but parameterizing the association in terms of conditional odds ratios. These latter are unconstrained, and this parameterization also provides orthogonality between the regression and association parameters. It has the disadvantage of conditional parameterizations in that the association measures are specific to a fixed sequence length, and thus this model is not suitable in circumstances involving missing data (or variable length sequences) without further adaptation. Eckholm et al. [19] have provided a third parameterization, this time using the marginal means and the dependence ratio, the first-order dependence ratios being of the form $\mathrm{E}[Y_{ij} = 1, Y_{ik} = 1]/(\mathrm{E}[Y_{ij} = 1]\mathrm{E}[Y_{ik} = 1])$. Within this parameterization, the dependence ratios are subject to relatively simple constraints, the mean and association parameters are not orthogonal, and the model is asymmetric; different results will be obtained depending upon which response is coded 1 or 0.

Time-by-time Analysis

A structured approach to time-by-time analysis, one that provides a rather straightforward method for dealing with missing data, has been provided by Wei & Stram [56]. They provided an estimator for the covariance matrix of the sets of parameters estimated at each time and a method for tackling the problem of multiple testing.

Derived Variable Analysis

The summary measures method can be applied as for continuous data but with the obvious modification of changes in the form and estimation of the derived variables.

Ordinal Data

Extensions from binary to ordinal responses (*see* **Ordered Categorical Data**) are possible for most, though not all, of the methods described [1]. Among random effects approaches, the log-gamma mixed complementary-log–log link models extend directly to ordinal data [12]. The **proportional odds** generalization of the logistic model [11] with random effects can be estimated directly by ML [28] or, if the ordinal response is transformed into a set of binary responses each indicating a response above or below each threshold, then PQL or GEE estimation become easily implemented [7]. In principle, the empirical generalized least squares approach may be applied, but the problems associated with sparse data become still more pressing than with binary data. The multivariate probit-based latent variable approach of Muthén [49] generalizes naturally to the ordinal case.

Count Data

Where the response measure represents an accumulation of discrete events over an interval of time, the **Poisson** likelihood offers a natural starting point. Variable interval lengths are straightforwardly dealt with by means of an *offset*. Extra-variation between subjects beyond that due to the included explanatory variables of the model may be accounted for either by a random effect or by *quasi-likelihood* or robust parameter covariance estimation. Assuming a gamma distribution for the subject-specific variation in rate leads to the well known **negative binomial** model.

Where there have been repeated observation intervals with explanatory variables that vary between intervals, then the approaches available are essentially parallel to those described for repeated binary outcomes. Conditional and parametric random effects estimation are both feasible [31]. Thall & Vail [54] describe a GEE approach. Little progress has been made with latent variable models for count data.

References

[1] Agresti, A. (1989). A survey of models for repeated ordered categorical response data, *Statistics in Medicine* **8**, 1209–1224.

[2] Andersen, E.B. (1980). *Discrete Statistical Models with Social Science Applications*. North-Holland, Amsterdam.

[3] Bahadur, R.R. (1961). A representation of the joint distribution of responses to *n* dischotomous items, in *Studies in Item Analysis and Prediction*, H. Solomon, ed. Stanford University Press, Stanford, pp. 118–168.

[4] Bartholomew, D.J. (1973). *Stochastic Models for Social Processes*. Wiley, New York.

[5] Blumen, J., Koggan, M. & McCarthy, D.J. (1955). *The Industrial Mobility of Labour as a Probability Process*. Cornell University Press, Ithaca, New York.

[6] Bonney, G.E. (1987). Logistic regression for dependent binary observations, *Biometrics* **43**, 951–973.

[7] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**, 9–25.

[8] Breslow, N.E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**, 81–92.

[9] Conoway, M.R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods, *Journal of the American Statistical Association* **84**, 53–62.

[10] Conoway, M.R. (1990). A random effects model for binary data, *Biometrics* **46**, 317–328.

[11] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London.

[12] Crouchley, R. (1995). A random effects model for ordered categorical data, *Journal of the American Statistical Association* **90**, 489–498.

[13] Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*. Chapman & Hall, London.

[14] Diggle, P.J. & Kenward, M.G. (1994). Informative drop-outs in longitudinal data analysis, *Applied Statistics* **43**, 49–93.

[15] Diggle, P.J., Liang, K.-L. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[16] Drum, O. & McCullagh, P. (1993). Comment to Fitzmaurice, G.M., Laird, N. and Rotnitsky, A. Regression models for discrete longitudinal responses, *Statistical Science* **8**, 284–309.

[17] Dunn, G., Everitt, B. & Pickles, A. (1993). *Modelling Covariances and Latent Variables Using EQS*. Chapman & Hall, London.

[18] Durcan, M.J., McWilliam, J.R., Campbell, I.C., Neale, M.C. & Dunn, G. (1988). Chronic antidepressant drug regimes and food and water intake in rats, *Pharmacology Biochemistry & Behavior* **30**, 299–302.

[19] Eckholm, A., Smith, P.W.F. & MacDonald, J.W. (1995). Marginal regression analysis of a multivariate binary response, *Biometrika* **82**, 847–854.

[20] Everitt, B.S. (1992). *The Analysis of Contingency Tables*, 2nd Ed. Chapman & Hall, London.

[21] Everitt, B.S. (1995). The analysis of repeated measures: a practical review with examples, *Statistician* **44**, 113–136.

[22] Everitt, B.S. & Dunn, G. (1993). *Applied Multivariate Data Analysis*. Edward Arnold, London.

[23] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

[24] Finney, D.J. (1990). Repeated measurements: what is measured and what repeats?, *Statistics in Medicine* **9**, 639–644.

[25] Fitzmaurice, G.M. & Laird, N.M. (1993). A likelihood based method for analysing longitudinal binary responses, *Biometrika* **80**, 141–151.

[26] Frison, I. & Pocock, S.J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design, *Statistics in Medicine* **11**, 1685–1704.

[27] Gibbons, R.D. & Hedeker, D.R. (1992). Full information item bi-factor analysis, *Psychometrika* **57**, 423–436.

[28] Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data, *Biometrika* **78**, 45–52.

[29] Hand, D.J. & Crowder, M. (1996). *Practical Longitudinal Data Analysis*. Chapman & Hall, London.

[30] Hand, D.J. & Taylor, C.C. (1987). *Multivariate Analysis of Variance and Repeated Measures*. Chapman & Hall, London.

[31] Hausmann, J., Hall, B. & Grilliches, Z. (1981). *Econometric models for count data with an application to the patent-R&D relationship. Mimeo*, MIT Press, Cambridge, Mass.

[32] Hills, M. & Armitage, P. (1979). The two-period crossover clinical trial, *British Journal of Clinical Pharmacology* **8**, 7–20.

[33] Hougaard, P. (1986). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678; (correction: **75**, (1988). 395).

[34] Huber, P.J. (1967). The behavior of maximum likelihood estimators under non-standard conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 221–233.

[35] Kenward, M.G. (1987). A method for comparing profiles of repeated measurements, *Applied Statistics* **36**, 296–308.

[36] Kenward, M.G. & Jones, B. (1992). Alternative approaches to the analysis of binary and categorical repeated measurements, *Journal of Biopharmaceutical Statistics* **2**, 137–170.

[37] Koch, G.G., Landis, J.R., Freeman, D.H. & Lehnen, R.G. (1977). A general methodology for the analysis of repeated measurement of categorical data, *Biometrics* **33**, 133–158.

[38] Laird, N.M. (1978). Non-parametric maximum likelihood estimation of a mixing distribution, *Journal of the American Statistical Association* **73**, 805–811.

[39] Laird, N.M. & Ware, J.H. (1982). Random effects models for longitudinal data, *Biometrics* **38**, 963–974.
[40] Landis, J.R., Miller, M.E., Davis, C.S. & Koch, G.G. (1988). Some general methods for the analysis of categorical data in longitudinal studies, *Statistics in Medicine* **7**, 109–137.
[41] Lavange, L.M. & Helms, R.W. (1983). *The analysis of incomplete data with modeled covariance structures*. Mimeo 1449, University of North Carolina, Institute of Statistics.
[42] Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.
[43] Lindsay, B.G., Clogg, C.C. & Grego, J. (1991). Semiparametric estimation in the Rasch model and related experimental response models including a simple latent class model for item analysis, *Journal of the American Statistical Association* **86**, 96–107.
[44] Lindsey, J.K. (1993). *Models for Repeated Measurements*. Oxford University Press, Oxford.
[45] Little, R.J. (1995). Modelling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**, 1112–1121.
[46] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.
[47] McGinnis, R. (1968). A stochastic model of social mobility, *American Sociological Review* **23**, 712–722.
[48] McNemar, Q. (1947). A note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–157.
[49] Muthén, B. (1984). A general structural equation model with dichotomous ordered categorical and continuous latent variable indicators, *Psychometrika* **49**, 115–132.
[50] Neyman, J. & Scott, E. (1948). Consistent estimates based on partially consistent observations, *Econometrica* **16**, 1–32.
[51] Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimation, *International Statistical Review* **54**, 221–226.
[52] Spiegelhalter, D.J., Thomas, A., Best, N.G. & Gilks, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling*, Version 5.0. MRC Biostatistics Unit, Cambridge.
[53] Spilerman, S. (1972). Extensions to the mover-stayer model, *American Journal of Sociology* **78**, 599–626.
[54] Thall, P.F. & Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**, 657–671.
[55] Ware, J.H., Lipsitz, S. & Speizer, F.E. (1988). Issues in the analysis of repeated categorical outcomes, *Statistics in Medicine* **7**, 95–107.
[56] Wei, L.J. & Stram, D.O. (1988). Analysing repeated measurements with possibly missing observations by modelling marginal distributions, *Statistics in Medicine* **7**, 139–148.
[57] White, H. (1980). Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1–25.
[58] Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association* **86**, 79–86.
[59] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.
[60] Zoccolillo, M., Pickles, A., Quinton, D. & Rutter, M. (1992). The outcome of childhood conduct disorder: implications for defining adult personality disorder and conduct disorder, *Psychological Medicine* **22**, 971–986.

GRAHAM DUNN & ANDREW PICKLES

Loss Function

The consequences of any decision will depend on the true state of nature, which determines whether the action corresponding to that decision is beneficial or harmful. The statistical formalization of this concept (*see* **Decision Theory**) has the following components:

Θ	is the set of all possible states of nature θ.
D	is the set of all possible decisions (actions) d.
$L(\theta, d)$	is the loss function that expresses the consequences of decision d when the state of nature θ holds.

By convention, the loss function is usually taken to be nonnegative, and is to be minimized. Since losses are to be compared or minimized, only the relative values of the losses of different decisions are important. Some synonyms for loss are "regret" and "cost". Alternatively, consequences are sometimes described in terms of "gain" or **"utility"**, and then maximization is the objective.

The loss function is useful in both statistical theory and in practical applications. A decision theoretic formulation of an existing procedure can clarify its interpretation by identifying its implicit loss function, and suggest generalizations. Practical applications include setting criteria to determine decisions related to disease for **screening**, setting treatment policy, **quality of life** issues, and study design.

Theoretical Uses of Loss Functions

A statistical decision procedure $\delta(X)$ yields a decision $d \in D$ based on the data X. Members of a set Δ of statistical procedures $\delta(X)$ are compared based on the loss function $L[\theta, \delta(X)]$ which, being a function of X, is a **random variable**. The risk of a decision procedure $\delta(X)$ is the **expectation** of the loss function, $R(\theta, \delta) = \mathrm{E}\{L[\theta, \delta(X)]|\theta\}$. Frequentist approaches (*see* **Inference**) compare procedures $\delta_1(X)$ and $\delta_2(X)$ by comparing the **risk** functions $R(\theta, \delta_1)$ and $R(\theta, \delta_2)$. In general, no procedure will minimize $R(\theta, \cdot)$ for all θ, so additional conditions are usually imposed. In ideal cases, the set Δ can be restricted according to a criterion such as **unbiasedness** (*see* **Minimum Variance Unbiased (MVU) Estimator**; **Most Powerful Test**), symmetry, or invariance [4], so that the risk function of some $\delta \in \Delta$ is dominated by all others, making the choice clear. Another approach, **minimax**, chooses the procedure which minimizes the maximum possible expected loss for any θ. **Bayesian** solutions minimize the expectation of $R(\theta, \delta)$ taken with respect to the **prior distribution** of θ.

Hypothesis testing can be viewed in terms of loss functions. Consider a hypothesis test of a **null hypothesis** H_0: $\theta = \theta_0$ vs. an **alternative hypothesis** H_1: $\theta = \theta_1$. Defining our loss to be 0 for a correct conclusion and 1 for an incorrect one gives the loss function shown in Table 1. The test is a decision function defined so that $\delta(x) = 1$ leads to rejection of the null hypothesis and $\delta(x) = 0$ leads to acceptance. The corresponding expected loss is shown in Table 2.

Thus $0-1$ loss leads to expected losses which are type I and type II errors in hypothesis testing. The traditional approach to hypothesis testing restricts Δ so that all tests considered have a fixed type II error. Minimizing the risk then amounts to maximizing the **power** of the test. Structuring the problem in this way aids in formulating more complex problems. For example, Emerson & Tritchler [3] elaborate the decision problem to incorporate a third type of error (Type III error) for a two-sided testing strategy (*see*

Table 1 Loss function for a hypothesis test

True parameter value	Decision: Decide $\theta = \theta_0$	Decision: Decide $\theta = \theta_1$
	Losses	
H_0: θ_0	0	1
H_1: θ_1	1	0

Table 2 Expected losses for a hypothesis test

True parameter value	Risk
θ_0	$0 + 1 \times \Pr(\text{decide } \theta = \theta_1 \mid \theta_0)$
θ_1	$1 \times \Pr(\text{decide } \theta = \theta_0 \mid \theta_1) + 0$

Alternative Hypothesis), concluding that a treatment is beneficial when the null hypothesis is false, but the treatment is actually harmful.

Parameter **estimation** can also be expressed in terms of an underlying loss function. Consider the decision procedure to be the parameter estimate $\hat{\theta}(X)$, which asserts the decision that the true parameter has value $\hat{\theta}(x)$. If $L[\theta, \hat{\theta}(X)] = [\hat{\theta}(X) - \theta]^2$, then $R(\theta, \hat{\theta})$ is **mean square error**.

Practical Applications of Loss Functions

Besides illuminating and guiding statistical theory, loss functions are of use in specific applications. Often the initial step of posing a problem in a decision theoretic framework will help to clarify aspects of a problem, even if the full decision theoretic solution is not required. Formulating loss functions will enable us to incorporate practical considerations into methodology, which are ignored in standard techniques. Some examples follow.

Suppose that the cost of the estimation error can be quantified in monetary units by the loss function $L(\mu, \overline{X}) = \lambda(\overline{X} - \mu)^2$. Then $R(\mu, \overline{x}) = \lambda\sigma_x^2/n$ is a function of only the sample size n, where the risk falls as n grows. We can add a term $C(n)$ to the loss function which states the cost of obtaining a sample of that size, and determine the n which minimizes the resulting loss given assumptions about σ_x^2 [1] (*see* **Sample Size Determination**).

Colton [2] proposed a loss function to guide the design of **clinical trials**. A treatment to be studied will affect two populations of patients: the $2n$ subjects on the two arms of the trial, and the "patient horizon", which consists of the N future patients who will be treated based on the results of that trial until future research provides an even newer treatment. Colton assigns a loss of λ to receiving the inferior treatment and 0 for the better treatment. Then, if the trial results are erroneous and lead to the adoption of the inferior treatment, the loss is $\lambda[n + (N - 2n)]$, since one arm of the trial and the patient horizon will receive the inferior treatment. If the correct treatment is chosen, a loss of λn is incurred by the arm on the inferior treatment. The relevant state of nature is the true treatment difference; a value for this is assumed and λ is taken to be proportional to it. Thus, the risk is $\lambda[n + (N - 2n)\,\Pr(\text{choose inferior})]$ for an assumed value of λ. This risk is a function of the sample size n; the first term of the above sum increases as n, but both factors of the second term decrease. This expresses the tradeoff between the welfare of the trial subjects and the patient horizon. Colton's loss formulation provides an interesting perspective on the impact of clinical trials.

Table 3 Loss function for eligibility screening

	Decision	
	Evaluate	Discard
True state	Losses	
Eligible	1	ϕ
Ineligible	1	0

Shannon et al. [6] consider the screening of patients for clinical trial eligibility. An initial screening of potential trial participants is done to select patients for further evaluation to determine eligibility. Table 3 shows the losses incurred by screening. The cost of evaluation is taken to be 1, and the loss due to discarding an eligible subject is ϕ times that, where $\phi > 1$ is specified subjectively.

In the examples, numerical losses were assigned to outcomes such as losing a trial participant or administering an inferior treatment. Losses of such a nature can be very difficult to quantify. Also, even if only monetary costs are involved, their direct interpretation as loss may be inadequate, especially after expectations are taken. These assessment and representation problems are addressed by utility theory, which derives techniques for quantifying perceptions of losses associated with complex outcomes having many incommensurate attributes [5]. If certain axioms hold, such loss functions accurately reflect preference when expectations are taken.

References

[1] Cochran, W.G. (1977). *Sampling Techniques*. Wiley, New York.

[2] Colton, T. (1963). A model for selecting one of two medical treatments, *Journal of the American Statistical Association* **58**, 388–400.

[3] Emerson, J.D. & Tritchler, D. (1987). The three-decision problem in medical decision making, *Statistics in Medicine* **6**, 101–112.

[4] Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

[5] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York.

[6] Shannon, W.D., Bryant, J., Logan, T.F. & Day, R. (1995). An application of decision theory to patient screening for an autologous tumour vaccine trial, *Statistics in Medicine* **14**, 2099–2110.

D. TRITCHLER

Louis, Pierre–Charles–Alexandre

Born: 1787, in Aï, France.
Died: June 9, 1872, in France.

P.-C.-A. Louis initially studied to be a lawyer; however, he abandoned law for medicine at the age of 20. After completing his initial medical training in Paris in 1813, he traveled throughout Russia for a period of seven years, eventually settling in Odessa. When Louis's medical training proved inadequate to combat an epidemic of diphtheria that occurred in Odessa in 1820, he resolved to return to Paris for additional study; however, he did not find much of use in the lectures of contemporary Parisian physicians.

With his appointment to the hospital, La Charité, in the early 1820s, Louis hoped to forge a more scientific foundation for medicine by collecting extensive records about the patients in the hospital, e.g. their ages, length of residence in Paris, the number who died and recovered from each disease, and the number of days duration of the disease. Louis used these records to determine the **mean** (or average) value for each analytical category and published his findings. In his study of typhoid fever, for example, Louis determined that it was primarily a disease of the young since the mean age of the 50 fatal cases was 23 and the mean age of the 88 who recovered was 21. In his 1835 treatise, *Recherches sur les effets de la saignée*, Louis provided the most famous example of his so-called "numerical method"; he demonstrated that the then common therapeutic practice of bloodletting was not as efficacious as its advocates believed, since 18 patients died out of 47 who had been bled (i.e. 38%) whereas only nine died out of the 36 patients who were not bled (i.e. 25%).

Louis's impact on the Parisian medical scene was most pronounced during the second quarter of the nineteenth century. In 1832, his followers founded the Société Médicale d'Observation to publish findings based on the numerical method. Although the society published three memoirs in 1837, 1844, and 1856, it did not survive after the retirement of Louis from public life in the mid 1850s following the premature death of his only son. Nevertheless, Louis had a long-term impact through the many students that he trained, including such prominent contributors to nineteenth century medicine and public health as the English physician and vital statistician **William Farr** and the American physician, Oliver Wendell Holmes.

J. ROSSER MATTHEWS

Machine Learning

Machine learning is the ability of a machine to recognize **patterns** that have occurred repeatedly and improve its performance based on past experience. According to Mitchell [4], a computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. The underlying idea is to learn theory automatically from data. Machine learning **algorithms** are being applied to practical problems in a wide variety of contexts. Examples include **data mining**, such as searching for irregularities in patient **databases**, **image** recognition problems, and analyses of genomic data (*see* **Human Genome Project**). Machine learning is inherently a multidisciplinary field, drawing on results from **artificial intelligence**, probability and statistics, computational complexity theory, control theory, information theory, philosophy, psychology, neurobiology, and other fields [3]. Developments are occurring in many areas that are familiar to biostatisticians, such as **Bayesian** modeling, graphical models (*see* **Path Analysis**), **Markov Chain Monte-Carlo** methods, **neural networks** and **hidden Markov models**, as well as areas that could be better known such as vector support machines (SVMs). Machine learning tools include methods for **cluster analysis** (unsupervised learning), **discrimination** (supervised learning), **factor analysis**, **regression** and **time series** problems, to name a few. It is generally believed that machine-learning approaches are best suited for areas where there is a large quantity of data but little theory, and so in computational biology such approaches are being widely used; see [2]. Moreover in turn, the explosion of genomic data, and resultant questions and problems, has motivated many advances in machine learning.

The strength of the approach offered by machine learning is the ability to automate the process of fitting very flexible models (which are characterized by large numbers of parameters) to extremely large databases. Currently a weakness is the lack of appropriate techniques for model criticism and validation (*see* **Model Checking**; **Model, Choice of**). The training and validation set approach is often used (*see* **Cross-validation**), but care needs to be taken in its application; see, for example, [1].

Hastie et al. [3] present machine learning from a statistical viewpoint. So, for example, E, T, and P above translate in statistical terms to E: training examples, T: parameterized models, and P: loss/risk measures. Many of the novel ideas in machine learning such as neural networks, boosting, and SVMs have been strengthened by incorporation of statistical rigor and interpretation.

References

[1] Ambroise, C. & McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data, *Proceedings of the National Academy of Sciences of the United States of America* **99**(10), 6562–6566.

[2] Baldi, P. & Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*, 2nd Ed. MIT Press, Cambridge.

[3] Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics); see `http://www-stat.stanford.edu/tibs/ElemStatLearn/`.

[4] Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill, New York; supplementary material at `http://www.cs.cmu.edu/tom/mlbook.html`.

SUSAN R. WILSON

Magic Square Designs

A *magic square* of size n is a set of integers in an $n \times n$ square such that each of the n rows, n columns, and the two main diagonals have the same sum m. A magic square is called *pandiagonal* if all the $2\ (n-1)$ wrap-around diagonals (the combined diagonals that are $+g$ and $-(n-g)$ from a main diagonal, for $g = 1, \ldots, n-1$) also sum to m, and *symmetrical* if all pairs of cells that are symmetrically opposite the center of the square sum to $2m/n$. In the usual case of a *magic square of order n*, the integers used are 1 to n^2, and $m = n(n^2+1)/2$. A pandiagonal and a symmetric magic square of order 4 ($m = 34$, $2m/n = 17$) are, respectively,

15	10	3	6
4	5	16	9
14	11	2	7
1	8	13	12

16	2	3	13
5	11	10	8
9	7	6	12
4	14	15	1

For further discussion and methods of construction, see Dénes & Keedwell [6] and Freeman [8], and the references therein.

Phillips [10] showed how the entries of a magic square of size n with distinct integers can be used to give the times at which the n^2 runs for a **factorial experiment** are made so that the main effects are *linear-trend-free*; that is, **orthogonal** to a straight line trend over time. If n has k factors, $n = n_1 n_2 \ldots n_k$, $n_i > 1$, then a design for (up to) $2k$ factors with levels $n_1, n_2, \ldots, n_k$ (each twice) can be obtained for which the main effects and at least some of the two-factor **interactions** are linear-trend-free (more if the magic square is symmetrical). For example, consider the $n = 4$ pandiagonal magic square above. Letting rows 1 to 4 represent a_0b_0, a_0b_1, a_1b_0, and a_1b_1, respectively, and columns 1 to 4 represent c_0d_0, c_0d_1, c_1d_0, and c_1d_1, respectively, gives, in the usual notation, the following run order for a 2^4 design for which all two-factor interactions are linear-trend-free:

$$(a, abcd, cd, b, bd, c, abc, ad, bc,$$
$$d, abd, ac, acd, ab, 1, bcd).$$

Fewer factors can be used with levels that are products of the n_i, and some n_i can be omitted. For example, a magic square of size 6 can be used for a complete replicate $2^2 \times 3^2$, $2 \times 3 \times 6$ or 6^2 design, or for two replicates of a 2×3^2 or a 3×6 design, or three replicates of a $2^2 \times 3$ or a 2×6 design, etc. If there is more than 1 replicate, then it may be possible to measure order effects within each replicate. A pandiagonal magic square of order n can be used to obtain an n^{3-1} **Latin square** (three factors each at n levels) or, for n odd, an n^{4-2} **Graeco-Latin square** (four factors each at n levels) for which main effects are linear-trend-free. There has been considerable further progress made on trend-free and trend-**robust** designs; see, for example, Bailey et al. [1], Bradley & Yeh [5], and Lin & Dean [9].

There are many connections between magic squares and *Latin squares*; see, for example, Dénes & Keedwell [6]. Amongst these are that if the integers $\{1, \ldots, n\}$ are used n times, then the magic square is a *diagonal Latin square*, and a pandiagonal magic square is a *Knut Vik design*. The Knut Vik design, which generalizes the well-known 5×5 knight's move Latin square, has five orthogonal constraints (block or treatment structures) of size n: rows, columns, the two sets of wrap-around diagonals, and the labels.

Another Latin square design, intended for n treatments in a spatial row–column layout, with one further block structure is the *Magic Latin square* (attributed to **G. M. Cox** – see Federer [7]). This requires a composite $n = n_1 \times n_2$, and forms spatially compact blocks using congruent $n_1 \times n_2$ rectangular blocks formed by the intersection of n_1 adjoining rows and n_2 adjoining columns. The extra set of blocks is not orthogonal to rows and columns, and care is needed in the analysis – see Bailey et al. [2, 3]. An example with $n = 4 = 2 \times 2$, showing the extra block boundaries, is

1	2	3	4
3	4	1	2
2	1	4	3
4	3	2	1

When $n_1 \neq n_2$, a *super magic Latin square* uses both $n_1 \times n_2$ blocks, and $n_2 \times n_1$ blocks to form

two extra sets of blocks. An example with $n = 6 = 2 \times 3 = 3 \times 2$ (with the two blocking structures to the right) is

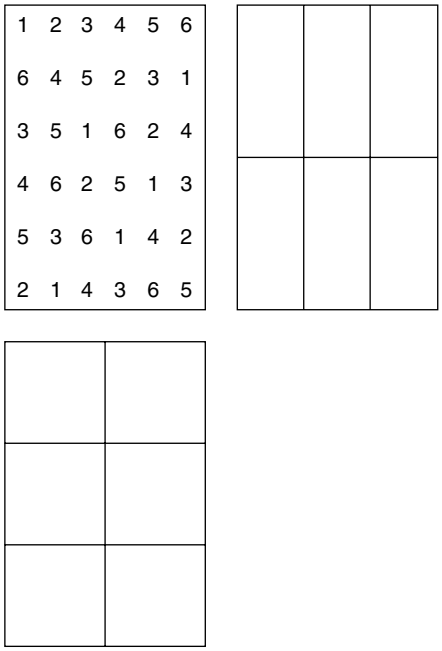

The nonaliased contrasts in the two extra sets of blocks are not orthogonal – see Bailey et al. [2].

The *gerechte designs* introduced by Behrens [4] can be regarded as a generalization of magic Latin squares which use any convenient spatially compact blocks of size n. The block shapes do not need to be congruent, so that gerechte Latin square designs can be obtained for any n. Gerechte designs also exist for rectangular arrays. Careful analysis is required – see Bailey et al. [2, 3].

References

[1] Bailey, R.A., Cheng, C.-S. & Kipnis, P. (1992). Construction of trend-resistant factorial designs, *Statistica Sinica* **2**, 393–411.

[2] Bailey, R.A., Kunert, J. & Martin, R.J. (1990). Some comments on gerechte designs. I. Analysis for uncorrelated errors, *Journal of Agronomy and Crop Science* **165**, 121–130.

[3] Bailey, R.A., Kunert, J. & Martin, R.J. (1991). Some comments on gerechte designs. II. Randomization analysis, and other methods that allow for inter-plot dependence, *Journal of Agronomy and Crop Science* **166**, 101–111.

[4] Behrens, W.U. (1956). Die Eignung verschiedener Feldversuchs-anordnungen zum Ausgleich der Bodenunterschiede, *Zeitschrift für Acker-und Pflanzenbau* **101**, 243–278.

[5] Bradley, R.A. & Yeh, C.-M. (1988). Trend-free block designs, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 324–328.

[6] Dénes, J. & Keedwell, A.D. (1974). *Latin Squares and Their Applications. English* Universities Press, London, Sections 6.1–6.3.

[7] Federer, W.T. (1955). *Experimental Design-Theory and Applications*. Macmillan, New York, Section XV-3.

[8] Freeman, G.H. (1985). Magic square designs, in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 173–174.

[9] Lin, M. & Dean, A.M. (1991). Trend-free block designs for varietal and factorial experiments, *Annals of Statistics* **19**, 1582–1596.

[10] Phillips, J.P.N. (1964). The use of magic squares for balancing and assessing order effects in some analysis of variance designs, *Applied Statistics* **13**, 67–73.

(*See also* **Factorial Designs in Clinical Trials**; **Youden Squares and Row–Column Designs**)

RICHARD J. MARTIN

Mahalanobis Distance

In 1936, **P.C. Mahalanobis** [22] proposed a measure, known as the generalized distance, or Mahalanobis distance, to assess the divergence between two populations based on observations on p characters or variates; the square of this distance is given by

$$\Delta^2 = \frac{1}{p}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors of the p variates in the two populations, and $\boldsymbol{\Sigma}$ is the common **covariance matrix**.

Historical Background

In the 1920s, **Karl Pearson** and his associates considered the problem of "asserting significant resemblance or divergence" between racial groups based on anthropological observations (*see* **Anthropometry**). Following Pearson's suggestion, Tildesley [45] considered a measure, known as the "Coefficient of Racial Likeness" (CRL), given by

$$\frac{1}{p}\sum_{i=1}^{p}\frac{(m_{i1} - m_{i2})^2}{\sigma_{i1}^2/n_{i1} + \sigma_{i2}^2/n_{i2}} - 1,$$

where m_{i1}, σ_{i1}^2, and n_{i1} denote the mean, the variance, and the sample size, respectively, corresponding to the ith variate in the first population, and m_{i2}, σ_{i2}^2, and n_{i2} similarly correspond to the second population. In 1926, Pearson [29] considered only the first term of the above expression. Romanovsky [39] also considered some similar criteria.

Mahalanobis, during his study on caste-groups in India, observed that the coefficient of racial likeness was influenced by sample sizes and it failed to measure the divergence [14]. He [21] suggested a general class of measures, and, in particular, considered the following when homoscedasticity holds:

$$D_0^2 = \frac{1}{p}\sum_{i=1}^{p}\frac{(m_{i1} - m_{i2})^2}{\overline{\sigma}_i^2} - \frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{n_{i1}} + \frac{1}{n_{i2}}\right),$$

where $\overline{\sigma}_i^2$ is a "reliable" value for the common variance of the ith variate. Mahalanobis cited a number of comparisons in which the coefficient of racial likeness and his D_0^2 measure gave widely different results, but he claimed that the values of D_0^2 gave better representation of known anthropological facts. Furthermore, Mahalanobis also proposes measures to assess divergence in variance, **skewness**, and **kurtosis**.

Later, Mahalanobis [22] introduced the correlations among the variates in defining such a measure, and proposed the measure Δ^2 given above. The sample version of Δ^2 for known Σ is given by

$$D_1^2 = \frac{1}{p}(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2}),$$

as well as by

$$D_2^2 = \frac{1}{p}(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2}) - \left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

where $\overline{\mathbf{x}_1}$ and $\overline{\mathbf{x}_2}$ are the sample mean vectors based on samples of sizes n_1 and n_2, respectively. It may be noted that D_2^2 is **unbiased** for estimating Δ^2. For unknown Σ, the sample version of Δ^2 is given by

$$D^2 = \frac{1}{p}(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2})'\mathbf{S}^{-1}(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2}),$$

where $\mathbf{S}$ is the pooled within-group sample covariance matrix with degrees of freedom (df) $n_1 + n_2 - 2$.

Under the assumption that the p variates are distributed as a normal distribution in each of the two populations, the distribution of $pD_1^2 n_1 n_2/(n_1 + n_2)$ is noncentral **chi-square** with p df and noncentrality parameter $p\Delta^2 n_1 n_2/(n_1 + n_2)$. R.C. Bose [3, 4] obtained this result along with the **moments** of D_1^2. S.N. Bose [6, 7] also obtained the moments of D_1^2, but without using its distribution explicitly.

For the problem of testing equality of mean vectors of two p-variate normal distributions with common but unknown covariance matrix $\boldsymbol{\Sigma}$, Hotelling [19] suggested the **Hotelling's** T^2 statistic, given by

$$T^2 = n_1 n_2 (n_1 + n_2)^{-1} pD^2,$$

as the test statistic and also as a modified form of the Coefficient of Racial Likeness. It was shown by Hotelling [19] that the null distribution of

$$\frac{T^2}{n_1 + n_2 - 2}\,\frac{n_1 + n_2 - p - 1}{p}$$

is the ***F* distribution** with p and $n_1 + n_2 - p - 1$ df. This result was also obtained by Fisher [17, 18] in his pioneering papers on **discriminant analysis**; however, Fisher's derivation is not rigorous. Mahalanobis [22] obtained the first four moments of D^2, assuming Σ to be a diagonal matrix. The nonnull distribution of the above statistic is F with df p and $n_1 + n_2 - p - 1$, and noncentrality parameter $n_1 n_2 (n_1 + n_2)^{-1} p\Delta^2$; this was first obtained by Bose & Roy [5]. For a review of the evolution of the D^2-statistic, see DasGupta [14].

Mahalanobis Δ as a Distance

The frame of reference for the work of Mahalanobis was the p-variate normal distribution for the variates under study. It is now known that many standard distance measures, such as Kolmogorov's variational distance, the Hellinger distance, Rao's distance, and so on, are increasing functions of Δ when the two distributions are p-variate normal distributions with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and common covariance matrix $\boldsymbol{\Sigma}$ [27]. This result also holds for a variety of distance measures for elliptic distributions with different locations but common shape parameters [28]. For other related developments on distance functions, see Rao [31, 33, 37], Matusita [26], and Burbea & Rao [8].

Role of Mahalanobis Distance in Discriminatory Analysis

In order to discriminate between two populations based on observations on p characters $\mathbf{X}$, Fisher [17, 18] considered a linear discriminant function $\mathbf{l}'\mathbf{X}$ to maximize $[\mathbf{l}'(\overline{\mathbf{X}_1} - \overline{\mathbf{X}_2})]^2/(\mathbf{l}'\mathbf{S}\mathbf{l})$. The optimal $\mathbf{l}$ turns out to be proportional to $\mathbf{S}^{-1}(\overline{\mathbf{X}_1} - \overline{\mathbf{X}_2})$ and correspondingly, the above ratio becomes pD^2. Fisher then suggested to consider pD^2 as the test statistic to test "significance of the discriminant function", which means testing the equality of the population mean vectors. In this development, Fisher's frame of reference was of course two p-variate normal distributions with common covariance matrix. For the problem of discrimination between two p-variate normal distributions with different covariance matrices, see Anderson [2] and McLachlan [27].

For detailed developments on discriminatory analysis, see Cacoullos [9]. For discrimination of Gaussian processes, see Rao & Varadarajan [38].

Test on Distance

As discussed earlier, Hotelling [19] first proposed a test for $\Delta^2 = 0$ when the underlying distributions are normal with common but unknown covariance matrix. Rao [30, 32] proposed a test for "additional distance", which may be posed as $p\Delta_p^2 = q\Delta_q^2 (q < p)$, where Δ_p^2 denotes the value of Δ^2 based on p variates. DasGupta & Perlman [16] have shown that the power of Hotelling's T^2-test based on p variates may be smaller than the power of the test based on a subset of q variates unless the increase $p\Delta_p^2 - q\Delta_q^2$ is sufficiently large; they have suggested a test based on a preliminary sample so that the effectiveness of inclusion of additional variates could be ascertained.

Rao [35] considered tests for assigned (linear) discriminant functions, as well as for specifications of the ratios of discriminant function coefficients. All of these, in principle, fall into the realm of testing additional distance.

The null distribution of D^2 can be used to obtain simultaneous confidence intervals for $\mathbf{l}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$; see Anderson [2].

Role of Mahalanobis Distance in Classificatory Analysis

The problem of classifying an observation vector $\mathbf{X}$ into one of two p-variate distributions with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and common covariance matrix $\boldsymbol{\Sigma}$, was first posed by Fisher [17] and developed later by Wald [47], Rao [34, 36], and Anderson [1], among many others; see McLachlan [27] for an extensive collection of results on this topic. For reviews of earlier work, see DasGupta [11, 13], Cacoullos [9], and Krishnaiah & Kanal [20].

When the parameters are known, the class of Bayes rules is given by the following: classify $\mathbf{X}$ into the first population if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \left\{ \frac{\mathbf{X} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2} \right\} \geq C;$$

otherwise classify into the second population. The probabilities of misclassification of any such rule are functions of Δ; in particular, if $C = 0$, the probabilities of misclassification are equal and the common value decreases as Δ increases. This result also holds when the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are unknown and they are respectively replaced by $\overline{\mathbf{X}_1}$,

$\overline{\mathbf{X}_2}$, and $\mathbf{S}$ in the above rule, and $n_1 = n_2$; for more detailed results, see DasGupta [12].

DasGupta & Kinderman [15] posed the concept of classifiability which sought condition on the structure of the populations in order to control probabilities of misclassification arbitrarily; for a related development, see Schaafsma & Steerneman [41]. For bounds, approximations, and asymptotic expansions relating to probabilities of misclassification, see McLachlan [27] and DasGupta [12]. For the problem of classification into one of two p-variate normal distributions with different covariance matrices, and the related role of Mahalanobis distance, see McLachlan [27]. Statistical methods for selecting variables in relation to the problem of classification and discriminatory analysis have been discussed in McLachlan [27] and Seber [44] (*see* **Variable Selection**).

Asymptotic Distribution of Δ

For the case of normal distributions, it follows from DasGupta [10] that

$$\mathrm{E}(pD^2) = f(f-p-1)^{-1}p\Delta^2 + (n_1^{-1}+n_2^{-1})f(f-p-1)^{-1}p,$$

where $f = n_1 + n_2 - 2$. Hence

$$\hat{\Delta}^2 = (f-p-1)f^{-1}D^2 - (n_1^{-1}+n_2^{-1})$$

is unbiased for estimating Δ^2. Moreover, the variance of $\hat{\Delta}^2$ is given by

$$(f-p-3)^{-1}\left\{2(p\Delta^2)^2 + 4n_1^{-1}n_2^{-1}(n_1+n_2)(f-1)p\Delta^2 + 2p(f-1)(n_1+n_2)^2n_1^{-2}n_2^{-2}\right\}$$

(see Schaafsma [40]).

It has been shown by Schaafsma & Van Verk [42, 43] that

$$\mathrm{E}(\sqrt{p}D) = \sqrt{p}\Delta + (4f)^{-1}(2p+1)(\sqrt{p}\Delta) + (2f)^{-1}(p-1)\kappa(\sqrt{p}\Delta)^{-1} + O(f^{-2}),$$

and

$$\mathcal{L}\left[f^{1/2}(\sqrt{p}D - \sqrt{p}\Delta)\right] \to \mathrm{N}\left(0, \kappa + \tfrac{1}{2}p\Delta^2\right],$$

as $n_1, n_2 \to \infty$, where $f(n_1+n_2)n_1^{-1}n_2^{-1} \to \kappa \in (0, \infty)$.

Other Applications

The domain of applications of Mahalanobis distance is quite extensive. In particular, the role of Mahalanobis distance in profile analysis (*see* **Summary Measures Analysis of Longitudinal Data**) and cluster analysis is significant (*see* **Cluster Analysis of Subjects, Nonhierarchical Methods**). See Mardia et al. [25], Van Ryzin [46], and Rao [36], in particular. The first application of Mahalanobis distance in cluster analysis is given in Mahalanobis et al. [23]. It may be noted that Mardia [24] has introduced a concept called "Mahalanobis angle", and illustrated its usefulness.

References

[1] Anderson, T.W. (1951). Classification by multivariate analysis, *Psychometrika* **16**, 31–50.

[2] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[3] Bose, R.C. (1936). On the exact distribution and moment coefficients of the D^2-statistic, *Sankhyā* **2**, 143–154.

[4] Bose, R.C. (1936). A note on the distribution of differences in mean values of two samples drawn from two multivariate normally distributed populations and the definition of the D^2-statistic, *Sankhyā* **2**, 379–384.

[5] Bose, R.C. & Roy, S.N. (1938). The distribution of studentized D^2-statistic, *Sankhyā* **4**, 19–38.

[6] Bose, S.N. (1936). On the complete moment coefficients of the D^2-statistic, *Sankhyā* **2**, 385–396.

[7] Bose, S.N. (1937). On the moment coefficients of the D^2-statistic, and certain integral and differential equations connected with the multivariate normal populations, *Sankhyā* **3**, 105–124.

[8] Burbea, J. & Rao, C.R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach, *Journal of Multivariate Analysis* **12**, 575–596.

[9] Cacoullos, T., ed. (1973). *Discriminant Analysis and Applications*. Academic Press, New York.

[10] DasGupta, S. (1968). Some aspects of discrimination function coefficients, *Sankhyā; Series A* **30**, 387–400.

[11] DasGupta, S. (1973). Theories and methods in classification: a review, in *Discriminant Analysis and Applications*, T. Cacoullos, ed. Academic Press, New York, pp. 77–137.

[12] DasGupta, S. (1974). Probability inequalities and errors in classification, *Annals of Statistics* **2**, 751–762.

[13] DasGupta, S. (1982). Optimum rules for classification into two multivariate normal populations with the same covariance matrix, in *Handbook of Statistics*, Vol. 2, P.R. Krishnaiah & L. Kanal, eds. North-Holland, New York, pp. 47–60.

[14] DasGupta, S. (1993). The evolution of the D^2-statistic of Mahalanobis, *Sankhyā, Series A* **55**, 442–459.

[15] DasGupta, S. & Kinderman, A. (1974). Classifiability and designs for sampling, *Sankhyā* **36**, 237–250.

[16] DasGupta, S. & Perlman, M.D. (1974). Power of the noncentral F-test: effect of additional variates on Hotelling's T^2 test, *Journal of the American Statistical Association* **69**, 174–180.

[17] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–188.

[18] Fisher, R.A. (1938). The statistical utilization of multiple measurements, *Annals of Eugenics* **8**, 376–386.

[19] Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**, 360–368.

[20] Krishnaiah, P.R. & Kanal, L., eds (1982). *Handbook of Statistics*, Vol. 2. North-Holland, New York.

[21] Mahalanobis, P.C. (1930). On tests and measures of group divergence, *Journal of the Asiatic Society of Bengal* **26**, 541–588.

[22] Mahalanobis, P.C. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Sciences of India* **2**, 49–55.

[23] Mahalanobis, P.C., Majumder, D.N. & Rao, C.R. (1949). Anthropometric survey of the United Provinces, 1941: a statistical study, *Sankhyā* **9**, 90–234.

[24] Mardia, K.V. (1977). Mahalanobis distance and angles, in *Multivariate Analysis*, Vol. IV, P.R. Krishnaiah, ed. North-Holland, New York pp. 495–511.

[25] Mardia, K.V., Kent, T. & Bibby, M. (1979). *Multivariate Analysis*. Academic Press, New York.

[26] Matusita, K. (1952). Decision rule based on the distance for the classification problem, *Annals of the Institute of Statistical Mathematics* **8**, 67–77.

[27] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

[28] Mitchell, A.F.S. & Krzanowski, W.J. (1985). The Mahalanobis distance and elliptic distributions, *Biometrika* **72**, 464–467.

[29] Pearson, K. (1926). On the coefficient of racial likeness, *Biometrika* **18**, 105–117.

[30] Rao, C.R. (1946). Tests on discriminant functions in multivariate analysis, *Sankhyā* **7**, 407–414.

[31] Rao, C.R. (1949). On the distance between two populations, *Sankhyā* **9**, 246–248.

[32] Rao, C.R. (1949). On the problems arising out of discrimination with multiple characters, *Sankhyā* **9**, 343–366.

[33] Rao, C.R. (1954). On the use and interpretation of distance functions in statistics, *Bulletin of the International Statistical Institute* **34**, 90–97.

[34] Rao, C.R. (1950). Statistical inference applied to classificatory problems, *Sankhyā* **10**, 229–256.

[35] R.C. Bose, Chakravarti, I.M., Mahalanobis, P.C., Rao, C.R. & Smith, J.C. (1970). Inference on discriminant function coefficients, in *Essays in Probability and Statistics*. R.C. Bose et al., eds. University of North Carolina Press, Chapel Hill.

[36] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. Wiley, New York.

[37] Rao, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach, *Journal of Theoretical Population Biology* **21**, 24–43.

[38] Rao, C.R. & Varadarajan, V.S. (1963). Discrimination of Gaussian process, *Sankhyā A* **25**, 303–350.

[39] Romanovsky, V. (1928). On the criteria that two given samples belong to the same normal population (on the different coefficients of racial likeness), *Metron* **7**, 3–46.

[40] Schaafsma, W. (1982). Selecting variables in discriminant analysis for improving upon classical procedures, in *Handbook of Statistics*, Vol. 2, P.R. Krishnaiah & L.N. Kanal, eds. North-Holland, New York, pp. 857–881.

[41] Schaafsma, W. & Steerneman, T. (1981). Discriminant analysis when the number of features is unbounded, *IEEE Transactions on Systems, Man and Cybernetics* **SMC-11**(2), 144–151.

[42] Schaafsma, W. & Van Verk, G.N. (1977). Classification and discrimination problems with applications, part I, *Statistica Neerlandica* **31**, 25–45.

[43] Schaafsma, W. & Van Verk, G.N. (1979). Classification and discrimination problems with applications, part II, *Statistica Neerlandica* **33**, 91–126.

[44] Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

[45] Tildesley, M.L. (1921). A first study of the Burnese skull, *Biometrika* **13**, 247–251.

[46] Van Ryzin, J., ed. (1977). *Classification and Clustering*. Academic Press, New York.

[47] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups, *Annals of Mathematical Statistics* **15**, 145–162.

(*See also* **Classification, Overview**; **Multivariate Analysis, Overview**)

SOMESH DASGUPTA

Mahalanobis, Prasanta Chandra

Born: June 29, 1893, in Calcutta, India.
Died: 1972, in India.

Prasanta Chandra Mahalanobis was educated at Presidency College, Calcutta, and King's College, Cambridge, where he completed the Tripos in Mathematics and Natural Science (Physics). In Part II of the Tripos, he was the only candidate to receive a first class in physics. Cambridge University awarded him a research scholarship. Before starting his research, he traveled to Calcutta for a short vacation, but never returned to England. The war intervened. Also, he had found a teaching job and plenty of other interesting things to do in Calcutta.

Just before Mahalanobis left England for this vacation, his tutor, W.H. Macaulay, drew his attention to the journal ***Biometrika***. Mahalanobis found the articles interesting and purchased an entire set of available volumes and brought these back to Calcutta. A window was opened to a new area of science, permanently changing the direction of his life.

Early on, one of his mentors, Acharya Brojendranath Seal, a philosopher and an encyclopedist who was also interested in statistics, said to him "Prasanta, ... you have to do work in India similar to that of Karl Pearson in England. In today's world, whether it is science or social service, without statistical methods there is no way. This is your job." (Translated from a note in Bengali by P.C. Mahalanobis dated April 17, 1945.) Mahalanobis, who had already begun to read **Karl Pearson's** papers in *Biometrika*, took this challenge seriously. He thus developed an interest in statistical analysis of biological data, which was to last throughout his life and to which he was to make profound contributions.

In 1920, Mahalanobis met the Director of the Zoological and Anthropological Survey of India, Nelson Annandale, who requested Mahalanobis to analyze some **anthropometric** data on a group of Anglo-Indians of Calcutta. Mahalanobis analyzed the data and published his first paper on statistics [1]. He continued to analyze the other anthropometric data in this sample, and presented a synthesis of results in his Presidential Address to the anthropology section of the Indian Science Congress in 1925. In the address, "Analysis of race-mixture in Bengal", Mahalanobis sought to provide answers to several anthropological questions by using statistical methods. (An expanded version of this address was later published by him in 1927 [3].) For example, do Anglo-Indians show a greater affinity with the higher castes of Bengal or with the lower castes? Or, is there any appreciable admixture with aboriginal tribes? To answer such questions, a measure of distance between population groups based on anthropometric measurements was necessary. The only available statistic for comparing resemblance between populations was Pearson's coefficient of racial likeness (CRL) [11, 13]. Mahalanobis realized that the CRL provided a test of divergence between samples drawn from two populations rather than a measure of the actual magnitude of the divergence, because the magnitude of the CRL was dependent on sample sizes. In the study on Anglo-Indians, Mahalanobis proposed and used a measure of the actual magnitude of divergence that he called the "first (provisional) measure of caste distance", D. The resulting inferences derived by Mahalanobis have been found to be largely valid from his own work conducted later in the United Provinces [10] and in Bengal, as well as in later studies of others using more extensive data and more sophist-icated statistical techniques.

During the period 1926–1927, Mahalanobis spent about six months in Karl Pearson's laboratory in the University College, London. During this period, he undertook an extensive analysis of anthropometric data of various European population groups, and closely examined the utility of the CRL for measuring population relationships. In the process, the statistical shortcomings of the CRL became clearer. Upon returning to India, Mahalanobis's ideas on the problem of incorporating the observed **correlations** among anthropometric measurements used in measuring distance took a more concrete form. He published a seminal paper, "On tests and measures of Group Divergence" in 1930, in which the famous D^2-statistic was proposed (*see* **Mahalanobis Distance**) [4]. Based primarily on work done by him in Pearson's laboratory, Mahalanobis published a paper in *Biometrika* in the same year [5]. This paper was the "first application of CRL to the discrimination of racial differences to be ascertained from measurements on the living" (p. 94). It dealt with the populations of Sweden, and Mahalanobis presented an innovative graphical display of anthropometric interrelationships among the populations, taking two

additional extrinsic variables into account, geographical location of habitat and occupation. Thus, the concept of forming clusters of populations began to take shape (*see* **Cluster Analysis, Variables**).

Mahalanobis subsequently proposed the "natural" generalized distance D^2 for correlated variates, as well as its Studentized form using sample values of parameters [8]. In retrospect, it is clear that both measures play a fundamental, important role in statistics and data analysis. The practical impact of the D^2 statistic has been enormous, and continues to be used in many branches of science.

Mahalanobis was apparently not satisfied with simply providing a valuable tool (D^2) for cluster analysis. He began to raise fundamental issues about the application of the D^2 statistic, and argued that inferences on affinities among populations may depend on the number of measurements chosen for assessing distances between populations; in which case, conclusions would not have the desired practical significance. Affinity configurations may change if one set of measurements is replaced by another. Mahalanobis thus laid down an important axiom for the validity of cluster analysis, "dimensional convergence of D^2" [9]. Suppose D_p^2 and D_∞^2 denote, respectively, the distance between a pair of populations based on a set of p measurements and the distance based on all of the measurements. Since it is not possible practically to study all possible measurements, biometrical studies must rely on a finite number, p, of measurements. For affinity relationships to be stable, the distance based on p characters should be a good approximation of that based on the set of all possible characters. For Mahalanobis's distance measure, it can be shown that $D_p^2 \leq D_\infty^2$ and $D_p^2 \to D_\infty^2$ as $p \to \infty$. Mahalanobis's axiom of dimensional convergence states that a suitable choice of p can be made if and only if D^2 is finite. Unfortunately, this important axiom is not mentioned in most textbooks on numerical taxonomy or cluster analysis.

The formulation of the D^2 statistic, derivation of its properties, and its applications are undoubtedly the most profound contributions of Mahalanobis to biostatistics. However, Mahalanobis made many other interesting contributions. Some of the early statistical studies he undertook were on **experimental designs** in agriculture. In 1924, he made some important discoveries pertaining to the probable error of results of agricultural experiments, which put him in touch with **R.A. Fisher**. Later, in 1926, he met Fisher at the Rothamsted Experimental Station and a close personal relationship was immediately established that lasted until Fisher's death. He possessed an uncanny sense of numbers and could quickly point out recording mistakes in data. In two papers entitled, "Revision of Risley's anthropometric data", Mahalanobis [6, 7] reconstructed the large series of anthropometric data, which were earlier condemned as faulty and unsuitable for statistical analysis. This work was highly praised by Sir Ronald Fisher [12]. He also conducted studies on dextrality of snail shells, correlates of disease prevalence in humans and plants, **demography**, and so on. In most of these studies, Mahalanobis developed novel statistical methods or made innovative applications of known methods. For example, in one of his early statistical studies on the **prevalence** of dysentery and its correlates, Mahalanobis [2] developed some useful smoothing techniques for **time-series** data using Fourier series (*see* **Fast Fourier Transform (FFT)**). Such techniques are now commonly used.

Mahalanobis's contributions to large-scale sample surveys, which are among his most significant and lasting gifts to statistics, began with problems of the estimation of area and yield of the jute crop in Bengal in 1937. He was able to demonstrate that estimates based on sample surveys were often more accurate than those based on complete enumeration, and that sample surveys could yield estimates with small margins of error within a short time and at a smaller cost than complete enumeration. He made many methodological contributions to survey sampling that included optimal choice of sampling design (*see* **Optimal Design**) using **variance** and cost functions, and the technique of an **interpenetrating** network of subsamples for assessment and control of errors, especially **nonsampling errors**, in surveys. The concept of pilot surveys was a forerunner of **sequential analysis** developed by **Abraham Wald**, as acknowledged by Wald. In addition to introducing these concepts, Mahalanobis raised important and difficult philosophical questions on the **randomness** and representativeness of a sample, which remain relevant and challenging even today. He was elected Chairman of the United Nations Sub-Commission on Statistical Sampling in 1947, and held this post till 1951. His tireless advocacy of the usefulness of sample surveys resulted in the final recommendation of this Sub-Commission that sampling methods should be extended to all parts of the world. Mahalanobis

received the Weldon Medal from Oxford University in 1944 and was elected a Fellow of The Royal Society, London, in 1945, for his fundamental contributions to statistics, particularly in the area of large-scale sample surveys.

As a scientist, Mahalanobis was, above all, a great applied statistician. Statistics were to be used for a better understanding of scientific data, and for decision-making for the welfare of society. Innovation, systematization and concrete applications are the hallmarks of the applied statistics practiced by Mahalanobis.

References

[1] Mahalanobis, P.C. (1922). Anthropological observations on the Anglo-Indians of Calcutta. Part I: Analysis of male stature, *Records of the Indian Museum* **23**, 1–96.

[2] Mahalanobis, P.C. (1926). Appendicitis, rainfall and bowel complaints. Part II. Scope of the enquiry, *Calcutta Medical Journal* **21**, 151–187.

[3] Mahalanobis, P.C. (1927). Analysis of race-mixture in Bengal, *Journal of Asiatic Society of Bengal* **23**, 301–333.

[4] Mahalanobis, P.C. (1930). On tests and measures of group divergence, *Journal of Asiatic Society of Bengal* **26**, 541–588.

[5] Mahalanobis, P.C. (1930). A statistical study of certain anthropometric measurements from Sweden, *Biometrika* **22**, 94–108.

[6] Mahalanobis, P.C. (1933). Revision of Risley's anthropometric data relating to tribes and castes of Bengal, *Sankhyā* **1**, 76–105.

[7] Mahalanobis, P.C. (1934). Revision of Risley's data relating to Chittagong hill tribes, *Sankhyā* **1**, 267–276.

[8] Mahalanobis, P.C. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Science* **2**, 49–55.

[9] Mahalanobis, P.C., Bose, R.C. & Roy, S.N. (1937). Normalization of statistical variates and the use of rectangular coordinates in the theory of sampling distributions (appendix), *Sankhyā* **3**, 35–40.

[10] Mahalanobis, P.C., Majumder, D.N. & Rao, C.R. (1949). Anthropometric survey of the United Provinces, 1941: a statistical study, *Sankhyā* **9**, 90–324.

[11] Pearson, K. (1936). On the coefficient of racial likeness, *Biometrika* **13**, 105–117.

[12] Rao, C.R. (1974). Prasanta Chandra Mahalanobis, 1893–1972. *Biographical Memories of Fellows of the Royal Society* **19**, 455–492.

[13] Tildesley, M.L. (1921). A first study of the Burmese skull, *Biometrika* **13**, 247–251.

J.K. GHOSH & PARTHA P. MAJUMDER

Mainland, Donald

Born: April 5, 1902.
Died: July 1985 in Kent, Connecticut.

Donald Mainland graduated in medicine at Edinburgh. He later taught anatomy in Edinburgh and received a Doctor of Science degree there for his research in embryology and histology. He moved to Manitoba, Canada, in 1927 and in 1930 became Professor and Chairman of the Department of Anatomy at Dalhousie University.

His early publications showed a concern about measurement issues, and foreshadowed an increasing interest in statistics. In 1936 he wrote on problems of chance in clinical work [2] and the following year he published his first book on statistics in medicine [3]. In 1950 he became Professor of Medical Statistics at New York University and shortly afterwards published his best known book, *Elementary Medical Statistics* [4]. Thereafter, Mainland was a prolific and influential writer on statistical topics.

In addition to his books, Mainland's notable contributions included several series of short essays on statistical topics, most of which were not published in journals but circulated to those "who were lucky enough to learn about 'the Notes', and to satisfy Mainland's hardy standards for the mailing list" [1]. From August 1959 to September 1966 he produced 145 items in the series, *Notes from a Laboratory of Medical Statistics* [5], a further 104 items in the series, *Notes on Biometry in Medical Research* [7], and 16 longer articles as "statistical ward rounds" from 1967 to 1969 in *Clinical Pharmacology and Therapeutics* [6].

After his retirement, Mainland continued to publish occasionally on statistical issues, with two typical outspoken and readable papers published in the *British Medical Journal* when he was in his eighties [8, 9].

The common sense consistently displayed in his writings was undoubtedly greatly aided by his extensive research and teaching in biology – he had also published a textbook on anatomy – and active participation in clinical research.

In 1970, when Mainland ceased writing his series in *Clinical Pharmacology and Therapeutics*, his successor in that role, Alvan Feinstein, described Mainland's contributions to improving the understanding and practice of statistics in medicine [1]. Among his generous comments Feinstein observed, "With his textbook . . . and his many other writings, he has probably contributed as much as any single person to the statistical sensibility of clinical investigators in North America" [1].

References

[1] Feinstein, A.R. (1970). Clinical biostatistics – 1. A new name – and some other changes of the guard, *Clinical Pharmacology and Therapeutics* **11**, 135–148 (reprinted in Feinstein, A.R. (1977). *Clinical Biostatistics*. C.V. Mosby Co., Saint Louis, pp. 1–14).

[2] Mainland, D. (1936). Problems of chance in clinical work, *British Medical Journal* **2**, 221–224.

[3] Mainland, D. (1938). *The Treatment of Clinical and Laboratory Data: An Introduction to Statistical Ideas and Methods for Medical and Dental Workers*. Oliver & Boyd, Edinburgh.

[4] Mainland, D. (1950). *Elementary Medical Statistics: the Principles of Quantitative Medicine* (2nd Ed., 1963). W.B. Saunders, Philadelphia.

[5] Mainland, D. (1959–1966). *Notes from a Laboratory of Medical Statistics* (a series of 145 mimeographed notes distributed by the author).

[6] Mainland, D. (1967–1969). Statistical ward rounds, *Clinical Pharmacology and Therapeutics* **8**, 139–146 to **10**, 576–586 (a series of 16 articles).

[7] Mainland, D. (1967–1970). *Notes on Biometry in Medical Research*. Veterans, Administration Monographs, Washington.

[8] Mainland, D. (1984). Statistical ritual in clinical journals: is there a cure? – I, *British Medical Journal* **288**, 841–843.

[9] Mainland, D. (1984). Statistical ritual in clinical journals: is there a cure? – II, *British Medical Journal* **288**, 920–922.

DOUGLAS G. ALTMAN

Mallows' C_p Statistic

This criterion can be helpful in selecting a biased linear model with fewer parameters and lower **mean square error** (MSE) than one with more parameters and their associated estimation errors. If a p-parameter linear model is fitted by unweighted **least squares** to n observations $y_1, \ldots, y_n$ (supposed uncorrelated and homoscedastic with variance σ^2) giving a residual sum of squares RSS_p (*see* **Analysis of Variance**), then C_p is defined by

$$C_p = \left(\frac{\mathrm{RSS}_p}{s^2}\right) - (n - 2p),$$

where s^2 is a trustworthy estimate of σ^2.

This criterion was introduced by Jones [2] in the equivalent form

$$JC_p(\text{say}) = \frac{\left[\mathrm{RSS}_p - (n - 2p)s^2\right]}{n}.$$

Under the conditions stated and if $\mathrm{E}(s^2) = \sigma^2$, JC_p is an **unbiased** estimate of the MSE, $\mathrm{E}\{\sum \hat{y}_i - \mathrm{E}(y_i)]^2/n\}$, of the model's fitted values as estimates of the true expectations of the observations. (A model with low MSE, as thus defined, may have good performance only for values of the independent variables in the region already observed.)

In order to guide the delicate practical choice of linear model from a number of alternatives, Mallows [3] developed his independent discovery of C_p into a graphical plot of C_p against p on which the line $C_p = p$ is drawn. In this plot the value of p is (roughly) the contribution to C_p from the variance of the estimated parameters, while the remainder $C_p - p$ is (roughly) the contribution from the **bias** of the model. This feature makes the plot a useful device for a broad assessment of the C_p values of a range of models. Its use does not (or at least should not) in itself inhibit choice of the model with the minimum value of C_p. Moreover, if that choice is made, the plot gives no obvious quantitative indication of the extent to which that minimum value, converted to JC_p, underestimates, as a consequence of selection bias, the actually operative MSE. (In [4], Mallows uses asymptotics in which, realistically, p goes to infinity with n – to provide such a quantitative indication, for a range of applications of the plot.)

The numerical comparisons in Burman [1] suggest that the use of just a "one-deep", leave-one-out cross-validatory criterion (*see* **Cross-validation**) may be more **robust** than C_p with respect to that selection bias. However, suggestions like this should be treated cautiously: the whole area abounds in competing and only partially substantiated claims.

For the relationship between C_p and Akaike's AIC, *see* **Akaike's Criteria**.

References

[1] Burman, P. (1996). Model fitting via testing, *Statistica Sinica* **6**, 589–601.

[2] Jones, H.L. (1946). Linear regression functions with neglected variables, *Journal of the American Statistical Association* **41**, 356–369.

[3] Mallows, C.L. (1973). Some comments on C_p, *Technometrics* **15**, 661–675.

[4] Mallows, C.L. (1995). More comments on C_p, *Technometrics* **37**, 362–372.

(*See also* **Diagnostics**; **Goodness of Fit**; **Model, Choice of**; **Multiple Linear Regression**; **Variable Selection**)

M. STONE

Malthus, Thomas Robert

Born: February 17, 1766, in Guildford, UK.
Died: December 23, 1834, in Bath, UK.

After an early education by private tutors, Malthus went to Jesus College, Cambridge, where he studied history, poetry, modern languages, classics, and mathematics. He was elected to a Fellowship at Jesus in 1793, and became a curate in a small town, Albury, in 1798. In that year he published the first version of his celebrated *Essay on the Principle of Population as it affects the Future Improvement of Society*, to be followed in his lifetime by five further editions. Malthus argued that a population would tend to increase geometrically, whereas the means of subsistence would increase only linearly. The consequent pressure caused by increasingly inadequate means of support would be a major determinant of political events and structures. The task of government was to counteract this dire prognosis by measures of population control, such as the encouragement of later marriage, rather than relying on increased poverty and mortality. Malthus's gloomy views ran counter to those of many progressive thinkers, but influenced Darwin's thought.

Malthus was a strong advocate of statistical investigation, and was a founder member of the Statistical Society of London (later the **Royal Statistical Society**) in 1834. His death, only nine months later, led the Society's Council to lament the loss of one "so celebrated in every part of the world where the science of Statistics is cultivated", describing him as "an ardent lover of truth, ... a sedulous investigator of facts, and a generous encourager of all who have followed in the same laborious path" [2] (see also [1, 3]).

The "Malthusian parameter" denotes the rate of increase that would ultimately be achieved by a population with observed age-specific birth and death-rates (*see* **Demography**).

References

[1] Keyfitz, N. (1985). Malthus, Thomas Robert, in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 189–190.

[2] Royal Statistical Society (1934). *Annals of the Royal Statistical Society, 1834–1934*. Royal Statistical Society, London.

[3] Stephen, L. (1896). Malthus, Thomas Robert, *Dictionary of National Biography* **36**, 1–5.

PETER ARMITAGE

Mantel, Nathan

Born: February 16, 1919 in New York City, New York.
Died: May 26, 2002 in Potomac, Maryland.

Nathan Mantel, pioneering biostatistician, and author of more than 380 published articles, was born in New York City to Polish and Hungarian immigrants, Rose Steinberg and Hyman (Nehemiah) Mantel. Nathan was the middle child between two sisters, Ray (Rifka, born 1917) and Anne (Channa, born 1920). Like many Jewish immigrants at the time, Nathan (Naftoolyah) grew up on the lower east side of New York City in tenement housing. He was raised poor, speaking Yiddish at home and at Hebrew School, but speaking and writing in English at his public school [11]. During the great depression, his Judaic studies took a turn away from the Orthodox Judaism of his parents when he and his sisters began residence in the Hebrew Orphan Asylum at 137th Street, Amsterdam Avenue. This unassuming orphanage provided safe harbor for many other future notables. Also in residence were Art Buchwald (syndicated columnist), Aaron L. Jacoby (politician), Dr. Herman Schwartz (biologist), Harold Tovish (sculptor), and many others [2]. During his adolescence, Nathan's maternal grandparents and 11 aunts, who had all remained in Eastern Europe instead of immigrating to the United States, were killed in the holocaust. Many of his paternal relatives immigrated to America successfully.

Nathan's academic training began at New York City's Stuyvesant High School. This premier school for science and mathematics helped to cultivate Nathan's interest and ability in mathematics, though his mature interest did not manifest itself until much later in his life. By most accounts, he was only a mediocre student [11]. However, even in high school, he participated in mathematics competitions. In his senior year, he derived a novel way to solve the Diophantine equation ($ax - by = c$), which was later published in the *American Mathematical Monthly* [7]. This would be the first of his numerous publications. In 1939, he graduated from City College of New York with a major in statistics. At City College, he took courses with other future statisticians, including Marvin **Schneiderman** and Bernard **Greenberg**. He later went on to earn a Master's degree in statistics from American University in 1956, already having published twenty-one articles in the field.

His career as a professional statistician began in 1940, after a series of low-level federal jobs. At this time, Nathan was recruited into what later became the War Production Board, where his skills helped increase the output of the nation's factories. Later, a portion of his World War II military service in the Army Air Force involved statistical analysis of medical research. But, with the war's end, the agency closed down and Nathan, who at that time was living with his family in temporary government housing in what is today the National Park Service's Kenilworth Aquatic Gardens, was jobless.

In 1947, Mantel went for a job interview at the National Cancer Institute (NCI) in the **National Institutes of Health** (NIH). He was quickly hired by Harold **Dorn** as a member of a new biometry group and set to work with such biostatisticians as Jerome **Cornfield**, Samuel **Greenhouse**, Jacob Lieberman, and Marvin Schneiderman (an old college pal of Nathan's). The rest, as they say, is history. Of this time period, Sam Greenhouse wrote, "... among statisticians the world over, we had probably the greatest artist of all – Nathan Mantel. No one could match him in quickly identifying the information in the data related to the questions and the swiftness with which he was able to choose an optimum method of analysis. The statistical procedures which bear his

name are really nothing compared to his ability to analyze data. The former would have eventually been derived by others, but it is doubtful whether anyone else has had his intuition" [5].

In the field of biostatistics, Nathan Mantel has published papers on leukemia, lung cancer, Down's syndrome, chemotherapy, breast cancer, passive smoking, vehicle emissions, and much more. Most notably, he developed the **Mantel–Haenszel** procedure and its extensions. William **Haenszel**, who had been working on interpreting the **case–control studies** of the connection between **smoking** and lung cancer, requested Mantel's assistance on how to analyze the retrospective data. Mantel then collaborated with Haenszel on a paper that aimed to reach the same conclusions "in a retrospective study as would have been obtained from a forward study, if one had been done" [10]. Their highly cited paper, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease", [10] presents the Mantel–Haenszel procedure, which provides a summary estimate of the exposure effect stratified by multiple sources (i.e. different studies) or **confounding** factors (such as age and sex), which is a weighted average of the **odds ratios** across various strata.

The applications and extensions of this procedure are many. Since this test allows for combining data from different sources, it can be used in a variety of contexts: retrospective studies, prospective studies, and laboratory experiments, including those with litter-matched samples. Mantel used the procedure to develop the first version of the **logrank test**, a test that compares **censored** time-to-response distributions [8], and later extended the test to the evaluation of response time data involving transient states [9]. These important applications contributed greatly to the development of **survival analysis** [4]. Mantel notes about his 1959 paper: "It turned out that the procedures in the paper could be extended so that they met perhaps 90 to 95 per cent of the kinds of problems that people were encountering" [6].

In addition, Mantel offered abundant insights to both **epidemiology** and laboratory research: He demonstrated that a prospective logistic risk model can be used to analyze case–control data [6]. He also explored the distribution of cancers among related diseased pairs to test whether the cause of the cancer was due to environmental exposure in addition to hereditary factors [4]. Further, Mantel developed methods to investigate temporal and spatial **clustering** of diseases such as polio, hepatitis, and childhood leukemia [4]. In 1961, Nathan devised the Mantel–Bryan approach to test for safety of carcinogenic agents (*see* **Tumor Incidence Experiments**). His definition of a "virtual safe" dose as a risk of one per 100 million or less was used by the **Food and Drug Administration** for several years before the standard was adapted to a less conservative definition of "safety" at one per million or less [6]. He later commented in an EPA Watch newsletter about the arbitrary nature of the original standard: "We just pulled it out of a hat" [1]. Upon hearing that a bureaucrat had dropped two zeros from his standard of one per 100 million, Nathan is reported to have remarked, "Well, that's government science for you!" Describing his overall approach to problem solving, Mantel wrote, "I generally don't generate ideas of my own. Someone has to come to me with a problem. And, apparently, I'm pretty good at coming up with solutions or ideas for solutions. Identifying problems is what is important – solutions just follow." [3].

A recipient of many professional honors, after retiring from the NCI, Mantel served as a research professor at George Washington University and later at American University. He was a visiting scientist at the New York University School of Medicine, a visiting professor at the University of Tel Aviv, and a visiting professor in neuroepidemiology at Temple University School of Medicine. He was also a lecturer at the China National Center for Preventive Medicine in Beijing. At its 2002 Annual Meeting, the **American Statistical Association** announced the establishment of the Nathan Mantel Lifetime Achievement Award for statisticians who have made significant contributions to the field of biostatistics over their careers. Nathan did not live to see the presentation of this award, as he died in his sleep on May 26, 2002. The epitaph on his gravestone reads, "One in a million", which serves as a concrete reminder of his lasting contributions to statistics and public policy. For additional biographical information and summaries of Nathan's work, refer to [2–6, 11].

References

[1] Anonymous. (1994). Junk science *The National Review* **46**, October 24 p. 22.

[2] Bogen, H. (1992). *The Luckiest Orphans: A History of the Hebrew Orphan Asylum of New York*. University of Illinois Press, Urbana.

[3] Gail, M.H. (1997). A conversation with Nathan Mantel, *Statistical Science* **12**(2), 89–97.

[4] Gail, M.H. (1999). Some of Nathan Mantel's contributions to epidemiology, *Statistics in Medicine* **18**, 3389–3400.

[5] Greenhouse, S.W. (1997). Some reflections on the beginnings and development of statistics in "Your Father's NIH", *Statistical Science* **12**, 82–87.

[6] Greenhouse, S.W. (1999). A selection of Mantel's contributions to laboratory research, *Statistics in Medicine* **18**, 3401–3408.

[7] Mantel, N. (1945). Note on the solution of Diophantine equations, *American Mathematical Monthly* **75**, 318–321.

[8] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemoth Reports* **50**(3), 163–170.

[9] Mantel, N. & Byar, D.P. (1974). Evaluation response-time data involving transient status: an illustration using heart-transplant data, *Journal of the American Statistical Association* **69**, 81–86.

[10] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[11] Wittes, J. (1999). Mantel unhyphenated, *Statistics in Medicine* **18**, 3381–3388.

LAUREN HALE & BENJAMIN HALE

Mapping Disease Patterns

For as long as disease patterns have been mapped there has been skepticism over the value of the pictures which are drawn. For instance, a map of the geography of the 1832 influenza epidemic in Glasgow (Scotland) was produced by the inmates of a lunatic asylum, mainly to occupy their time [1]. Later, in the nineteenth century, the value of mapping disease patterns was recognized as specific epidemiologic breakthroughs were attributed to the insight gained from mapping. Often cited is a map of the distribution of deaths from the 1848 cholera epidemic in London (England) which, so the tale goes, inspired the removal of the handle of the water pump at the center of a cluster of dots on the map, resulting in the curtailing of the epidemic [12].

Maps of diseases are like news pictures of crowd trouble. Viewers should always ask themselves what is not being shown in the map while looking at what is there. In particular, look around the edge of the map. Ask why it ends where it does. For instance, maps of diseases are often centered on the point the author thinks is most important. Figure 1 shows the central section of John Snow's map of deaths from cholera in Soho. Note how the eye is drawn to the pump in the center, particularly by the very high number of deaths at the intersection of Cambridge and Broad Streets. Had Snow drawn his map of all of London he would have discovered a greater density of deaths just south of the river Thames, as shown in Figure 2. This concentration would have changed location again had Snow had recourse to an isodemographic base map, as shown in Figure 3. As our picture of a disease pans out, as we include more cases and as we change the way we view the picture, the patterns on our maps show change too.

Disease mapping has been most strongly influenced by the history of diseases. Figure 4 shows the prevalence of 12 major causes of death in England and Wales since the publication of Snow's map of cholera. Infectious diseases now account for a tiny fraction of deaths in developed countries (which can afford most disease mapping and research). It is causes of death which are not declining, such as suicide, and those which are rising in importance, such as cancers, which increasingly interest researchers. For these causes of illness and death the analysis of point patterns around particular sites is still a major issue, but the patterns are usually far less clearly spatially defined than were outbreaks of cholera.

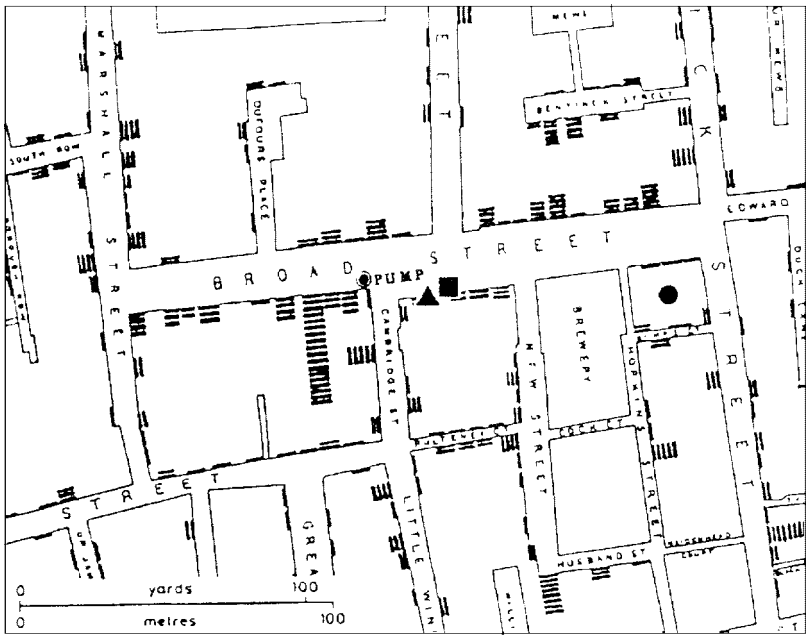

Figure 1 John Snow's map of cholera deaths in Soho, London, 1854 – taken from Cliff & Haggett [1, Figure 1.15D]

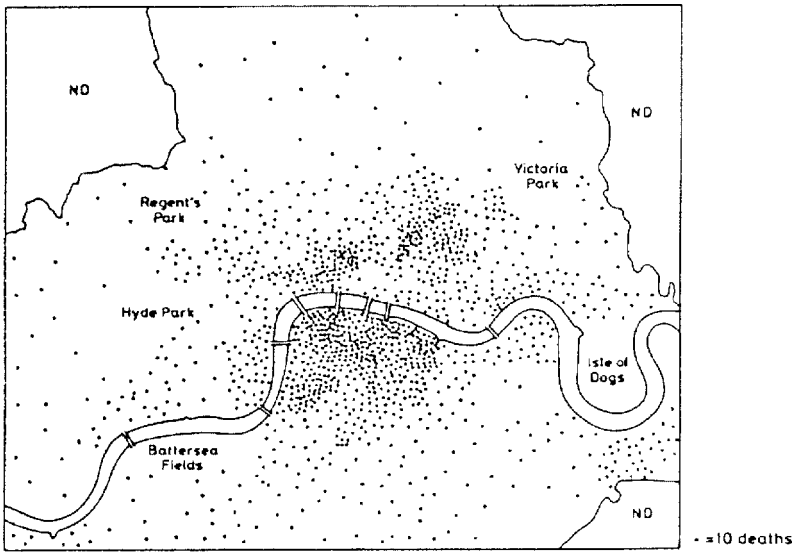

Figure 2 Cholera deaths in London in 1849 – taken from Cliff & Haggett [1, Figure 1.3B]

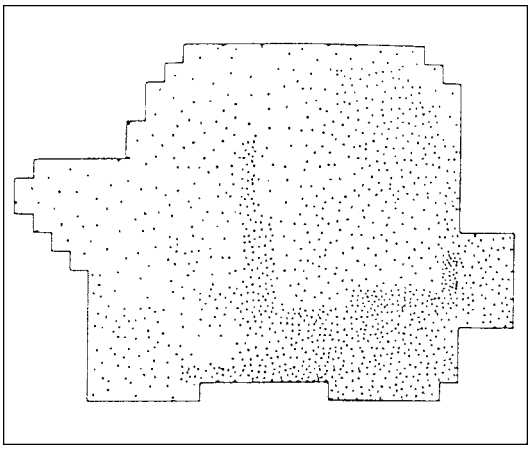

Figure 3 Figure 2 on a population cartogram – taken from Cliff & Haggett [1, Figure 1.18D]

More importantly, it is increasingly being accepted that more abstract factors, such as social inequality, can lie behind particular patterns of disease, and these require more abstract mappings for their study.

There are many different ways of mapping disease but here there is only space to explore one alternative. The alternatives include traditional choropleth mapping, where areas on a map are shaded according to statistics about the population. Most common in epidemiology is the mapping of areas colored by their standardized mortality ratios (*see* **Standardization Methods**). Another common form of mapping is to map points or the incidences of disease, and often color is also used here to highlight different types of disease. Various different point symbols can be used in mapping, particularly common is the use of proportional circles which are colored or segmented to highlight different features of a disease. The size of the circles is often made proportional to the population at risk of contracting a disease, at which point this type of cartography begins to merge into isodemographic mapping [4, 5].

Diseases occur across a population as much as across land. That is not to say that geographic distributions are not important, but that we should take account of the distribution of the population at risk to a particular disease, or cause, before mapping its pattern. One way in which this can be done is to use a map projection which draws every area in proportion to the number of people at risk living in that area – hence the term isodemographic ("equal people"). Isodemographic maps, more commonly called cartograms, are used for many purposes, mostly obviously in mapping the geography of elections. However, their most established use has been in disease mapping. Figure 5 shows one of the earliest examples of a cartogram designed for epidemiologic purposes [15, p. 1023]. Figure 5(a) is the conventional map of the counties of Iowa State, and Figure 5(b) is an equal

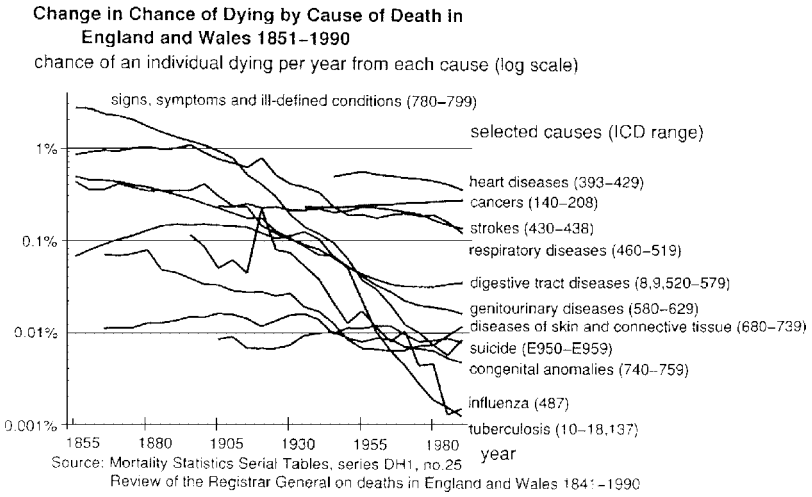

Figure 4 Cause of death 1855–1990 – taken from Dorling [3, Figure 5.21]

population cartogram upon which colored pins were placed to show the locations of reportable diseases. The square in the middle of the cartogram is Des Moines city in Polk County.

The designer of the Iowa cartogram was a doctor working in the state department of health. Many researchers have been struck by the idea that they could learn more about disease through mapping it in unconventional ways. The first cartogram of London was an "epidemiologic map" produced by a doctor working for the then London County Council Department of Public Health [14]. The cartogram (Figure 6) contained crosses drawn in the borough rectangles to show the incidence of polio during the 1947 epidemic. Because the rectangles were each drawn with the same height, their widths are proportional to population as well as their areas. The borough with the highest rate of polio and hence the tallest column of crosses in the Figure was Shoreditch. Almost exactly 100 years separates the two London epidemics, which were first drawn on a map and cartogram, respectively. Cartograms showing distributions within countries came later.

A claim was made to have produced the first cartograms showing national disease distributions only a decade after the crude cartogram of London was first drawn [6]. The nation was Scotland, and a separate cartogram was constructed by hand for each of eight age–sex groups. Figure 7 shows the cartogram being used to study the 1959–1963 mortality of women in Scotland aged 45–54. The author of this cartogram concluded that a national series of cartograms should be produced for each age–sex group for use in epidemiologic studies in Britain. This was never done, and it is debatable whether such an exact mapping base is needed in most studies. A single isodemographic base map of the whole population will usually suffice to uncover all but the most subtle of patterns.

A National Atlas of Disease Mortality in the UK was published in 1963 under the auspices of the Royal Geographical Society; the atlas contained no cartograms. However, a revised edition was published a few years later which made copious use of a "demographic base map" [7]. It is interesting to note that, when the revised edition was being prepared, the president of the Society was Dudley Stamp, who believed that "The fundamental tool for the geographical analysis is undoubtedly the map or, perhaps more correctly, the cartogram" [13, p. 135]. In the cartogram which was used in the revised national atlas (Figure 8), squares were used to represent urban areas, while diamonds were used to show statistics for rural districts. No attempt was made to maintain contiguity, but a stylized coastline was placed around the symbols, which were all drawn with their areas in proportion to the populations at risk from the disease being shown on each particular cartogram.

In the *National Atlas of Disease Mortality in the United Kingdom*, Howe used a national cartogram to display the distribution of standardized mortality between 1959 and 1963 from separate as well as

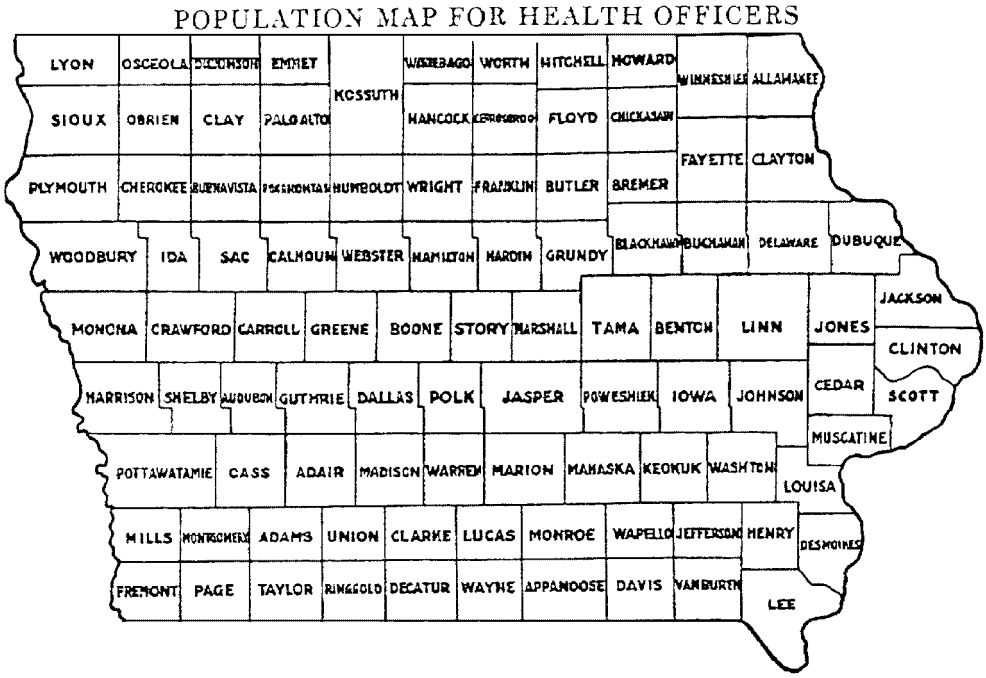

(a)

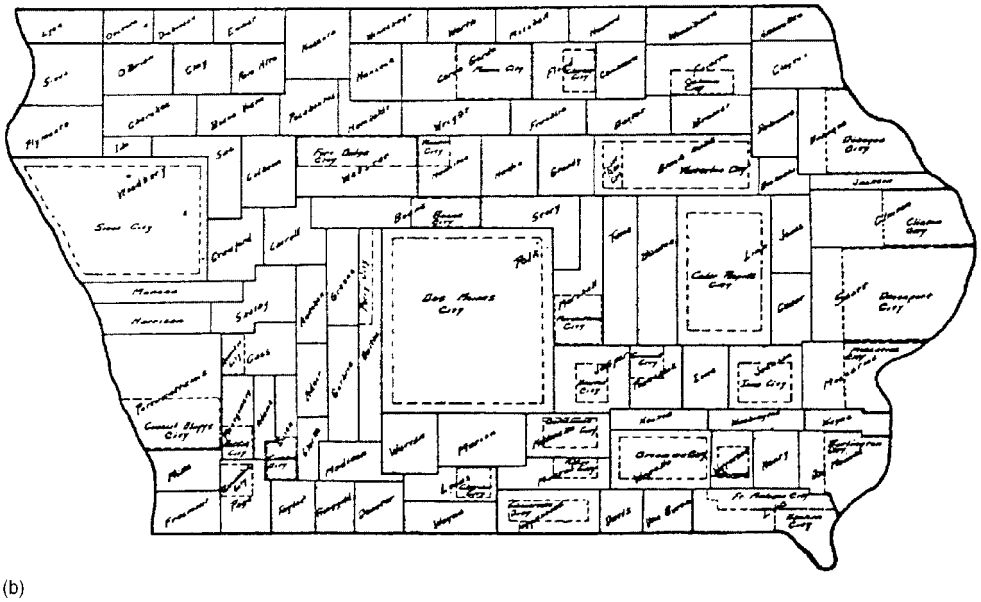

(b)

Figure 5 The use of cold vaccine in Iowa County Area, 1926 – taken from Wallace [15, p. 1023]

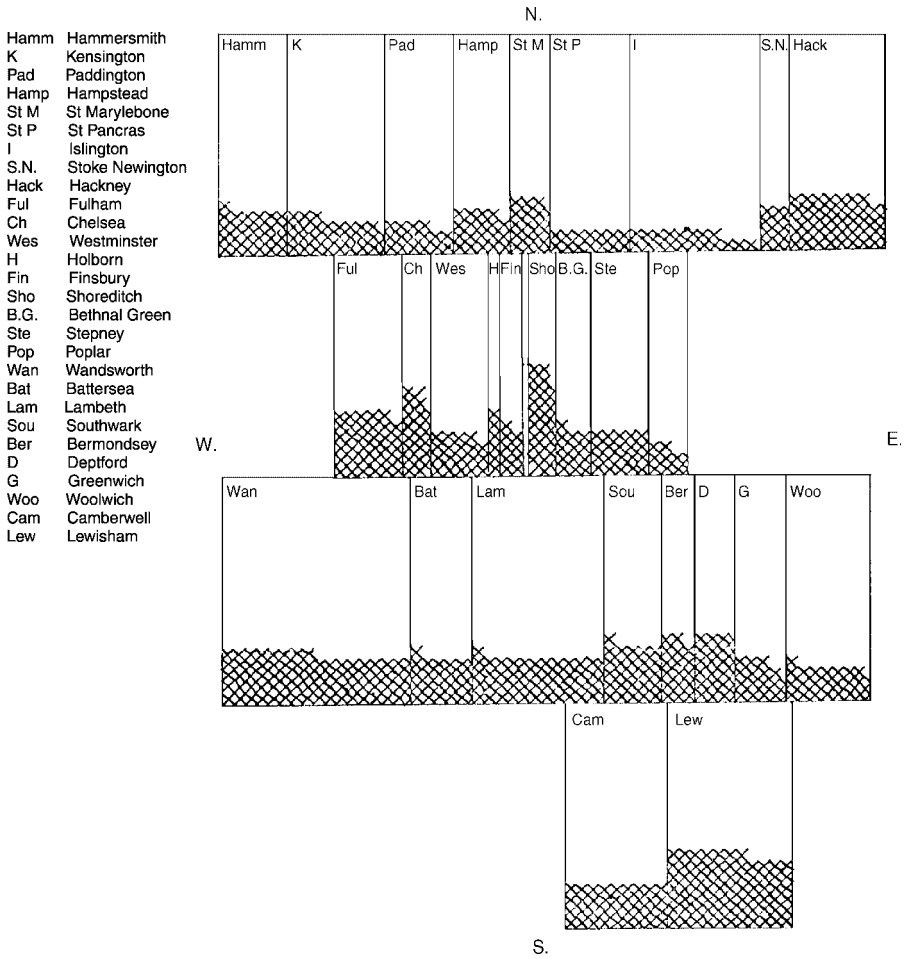

Figure 6 London borough cartogram showing 1947 poliomyelitis notifications – taken from Taylor [14, p. 201]

all causes of death for both men and women. High rates were seen in northern districts and some Inner London boroughs (including Shoreditch, which is also highlighted on one of the earliest cartograms of London; see above). Extremely high rates in central Scotland were particularly noticeable, as were the low rates in districts which surround London. At the extremes the average man living in Salford was 50% more likely to die each year than his counterpart in Bournemouth [7]. Both these areas are shrunk on a "normal" map. The pattern for women was very similar to that for men although, in general, it was less pronounced. However, women did have the highest mortality rate of any area on the map in rural Dunbartonshire, where they were more than twice as likely to die each year than were women nationally (allowing for local age structure). The cartogram highlights this area, but also puts it in the perspective of the populations at risk from the high mortality rates for women in and around the Glasgow area. Questions for investigation are immediately generated by comparing the maps in Howe's atlases with those produced by Forster for a decade earlier (see Figure 7).

Isodemographic mapping is also used to study the prevalence of disease – individual cases of a disease or death which together might possibly be connected. Figure 9 shows the distribution of cases of Wilm's tumor, a childhood cancer, identified in New York State between 1958 and 1962, drawn upon an equal

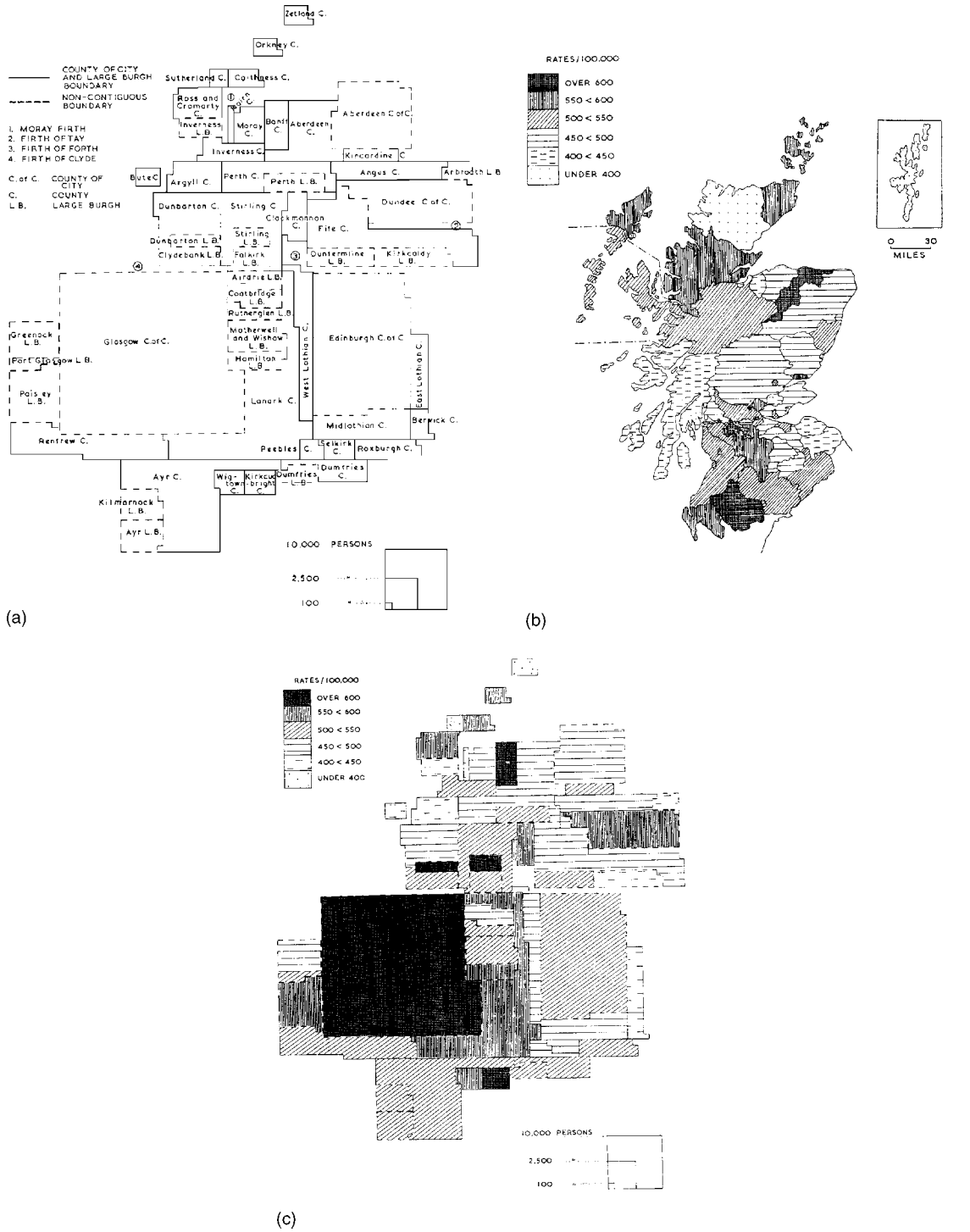

Figure 7 Cartogram and map of Scottish health districts – taken from Forster [6]. (a) Cartogram of females aged 45–54 in 1961 by Scottish health districts; (b) map of 1959–63 mortality rates of females aged 45–54 by district; (c) 1959–63 mortality rates of females aged 45–54 shown in (a)

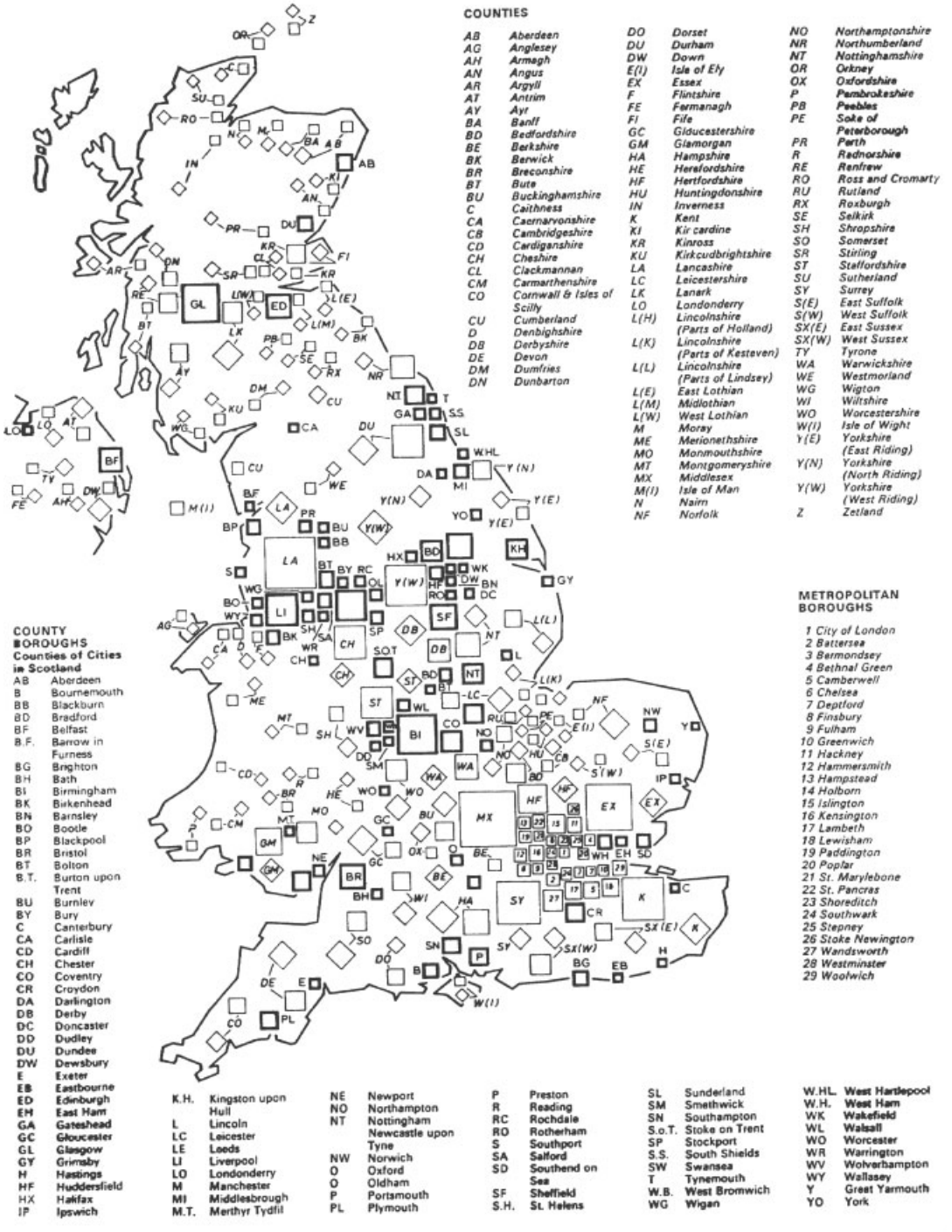

Figure 8 Cartogram of districts of disease mapping in the UK – taken from Howe [7]

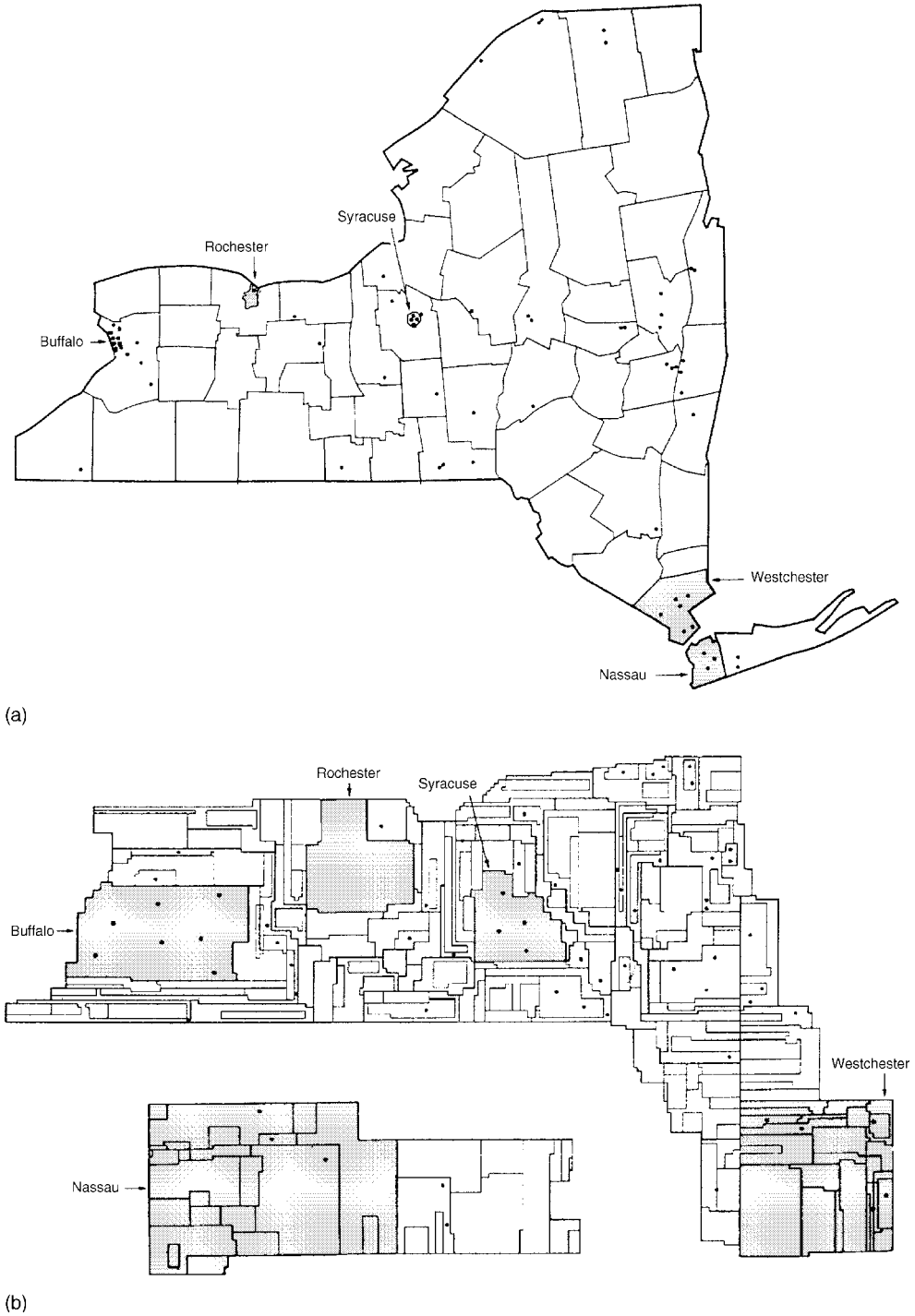

Figure 9 Wilm's tumour cases on (a) map and (b) cartogram in New York State – taken from Levison & Haddon [8]

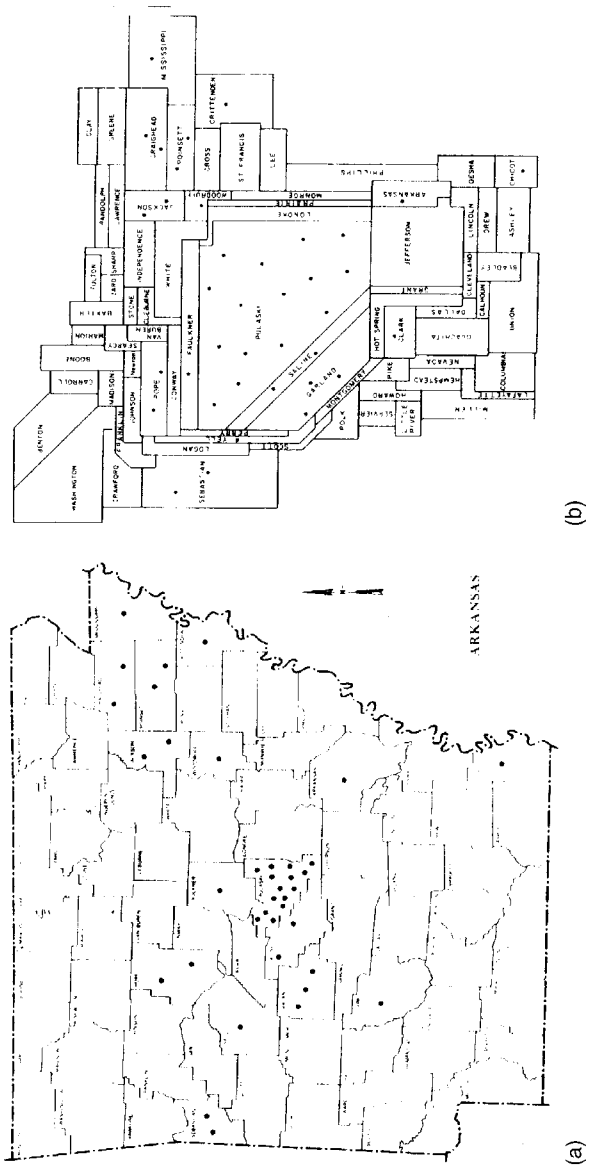

Figure 10 Salmonella Newport cases on (a) map and (b) cartogram in Arkansas State – taken from Dean [2]

land area map. Apparent clusters of cases have been marked on the map [8]. In the second diagram in Figure 9, the same cases are drawn upon an equal population cartogram and the apparent clusters can be seen to have been quite evenly dispersed across the population. The same process has been used in Figure 10 to illustrate how cases of Salmonella food poisoning occurring in Arkansas in 1974 were not unduly clustered in Pulaski county [2].

In recent years researchers have turned their attention to trying to develop cartograms upon which actual, rather than illusory, clusters of disease can be identified (*see* **Clustering**). The major problem with using population cartograms to identify clusters of disease is that the choice of which areas are closest to which on a cartogram can be quite arbitrary. For instance, if the same set of incidences of one particular disease were plotted on three different cartograms, then different parts of the country may appear to have dense clusters of cases depending on which cartogram was chosen. This would be true regardless of whether the clusters were to be identified by eye or by statistical procedures; the different base maps would result in different patterns emerging. The proposition that there is no single "true answer" as to whether a disease is clustered does not go down too well in some circles. Because of this problem a group of researchers at Berkeley developed a computer algorithm for identifying incidences of disease [9]. The algorithm was used to produce the cartogram in Figure 11 of San Francisco county, upon which apparent clusters of disease were shown to be false [11]. However, application of the method to another California county did provide evidence of some clustering of high cancer rates near oil refineries [10].

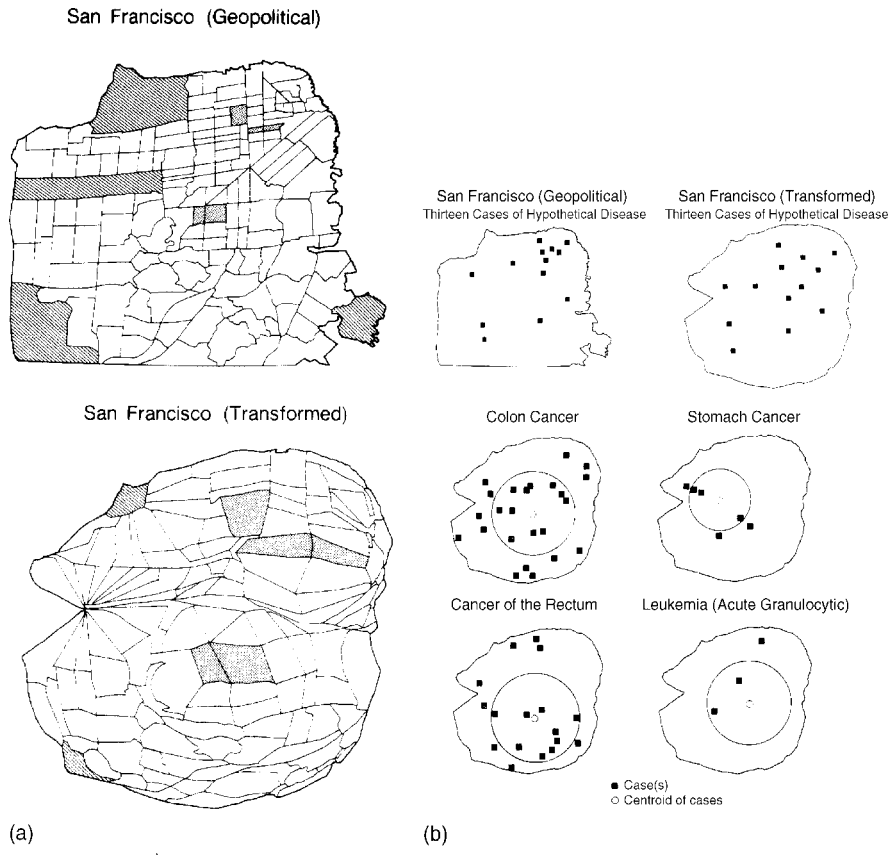

Figure 11 San Francisco map (a) for 1980 census, and cartogram (b) of hypothetical and actual diseases – taken from Selvin et al. [11]

Mapping of disease patterns is becoming increasingly common due to the proliferation of computer mapping. However, many of these programs were designed to produce general maps of any subject and are often most appropriate to show land use or the distribution of points in physical space. Over most of the course of the last century, doctors, public health officials, and researchers have discovered and rediscovered that traditional maps often do not provide the most appropriate projection to look for patterns of disease. Here, a few alternatives have been shown of just one different form of disease mapping to try to explain why it involves more than just sticking pins in paper.

Acknowledgment

The author is grateful to Robert Israel for commenting on a draft of this article and to the following people for permission to reproduce the copyright material shown here: Peter Haggett (*Atlas of Disease Distributions*) for Figures 1–3; Pam Beckley (Her Majesty's Stationery Office) for Figure 5; Michael Plommer (Office for National Statistics) for Figure 6; Carol Torselli (*British Medical Journal*) for Figure 7; Marian Tebben (*Public Health Reports*) for Figure 9; and Mina Chung (American Public Health Association) for Figure 10.

References

[1] Cliff, A.D. & Haggett, P. (1988). *Atlas of Disease Distributions. Analytical Approaches to Epidemiological Data*. Blackwell, Oxford.

[2] Dean, A.G. (1976). Population-based spot maps: an epidemiologic technique, *American Journal of Public Health* **66**, 988–989.

[3] Dorling, D. (1995). *A New Social Atlas of Britain*. Wiley, Chichester.

[4] Dorling, D. (1996). *Area Cartograms: Their Use and Creation*, Concepts and Techniques in Modern Geography (CATMOG) no. 59. School of Environmental Sciences, University of East Anglia, Norwich.

[5] Dorling, D. & Fairbairn, D. (1997). *Mapping: Ways of Representing the World*. Longman, London.

[6] Forster, F. (1966). Use of a demographic base map for the presentation of areal data in epidemiology, *British Journal of Preventive and Social Medicine* **20**, 165–171.

[7] Howe, G.M. (1970). *National Atlas of Disease Mortality in the United Kingdom*, Revised and Enlarged Edition. Nelson, London.

[8] Levison, M.E. & Haddon, W. (1965). The area adjusted map: an epidemiological device, *Public Health Reports* **80**, 55–59.

[9] Selvin, S., Merrill, D., Sacks, S., Wong, L., Bedell, L. & Schulman, J. (1984). *Transformations of Maps to Investigate Clusters of Disease*. Laboratory Report, LBL-18550, Lawrence, Berkeley.

[10] Selvin, S., Shaw, G., Schulman, J. & Merrill, D. (1987). Spatial distribution of disease: three case studies, *Journal of the National Cancer Institute* **79**, 417–423.

[11] Selvin, S., Merrill, D., Schulman, J., Sacks, S., Bedell, L. & Wong, L. (1988). Transformations of maps to investigate clusters of disease, *Social Science and Medicine* **26**, 215–221.

[12] Snow, J. (1854). *On the Mode of Communication of Cholera*. Churchill Livingstone, London.

[13] Stamp, L.D. (1962). A geographer's postscript, in *Taxonomy and Geography*, D. Nichols, ed. The Systematics Association, London, pp. 153–158.

[14] Taylor, I. (1955). An epidemiology map, *Ministry of Health Monthly Bulletin* **14**, 200–201.

[15] Wallace, J.M. (1926). Population map for health officers, *American Journal of Public Health* **16**, 1023.

(*See also* **Geographic Patterns of Disease**; **Geographic Epidemiology**)

DANIEL DORLING

Marginal Likelihood

Suppose that $\mathbf{X} = (X_1, \ldots, X_n)'$ is a vector of random variables whose distribution depends on parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. We suppose that $\boldsymbol{\beta}$ is of primary interest, whereas $\boldsymbol{\lambda}$ is a **nuisance parameter** and typically is of very high dimension. Our aim is to define a derived **likelihood** that would be suitable for inference about $\boldsymbol{\beta}$ when $\boldsymbol{\lambda}$ is unknown.

Let $f(\mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\lambda})$ be the probability density function (pdf) of $\mathbf{X}$; on data $\mathbf{X} = \mathbf{x}$, it defines the joint likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \mathbf{x}) \propto f(\mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\lambda}),$$

which can be used for inference. However, when $\boldsymbol{\lambda}$ is of high dimension, it becomes difficult to interpret the information about $\boldsymbol{\beta}$. In fact, the **maximum likelihood** estimator (MLE) of $\boldsymbol{\beta}$ can have very poor properties, even asymptotically, if the dimension of $\boldsymbol{\lambda}$ increases with that of $\mathbf{X}$. To make inferences about $\boldsymbol{\beta}$ itself without regard to $\boldsymbol{\lambda}$, it is useful to define a derived likelihood which, by some method, eliminates $\boldsymbol{\lambda}$. Marginal likelihood provides one way of doing this.

Suppose that there exists a one-to-one transformation of $\mathbf{X}$ into ($\mathbf{A}$, $\mathbf{T}$) and that the joint pdf of ($\mathbf{A}$, $\mathbf{T}$) factors as

$$f(\mathbf{a}, \mathbf{t}; \boldsymbol{\beta}, \boldsymbol{\lambda}) = f(\mathbf{a}; \boldsymbol{\beta}) f(\mathbf{t}|\mathbf{a}; \boldsymbol{\beta}, \boldsymbol{\lambda}), \qquad (1)$$

where the marginal density of $\mathbf{A}$ does not depend on $\boldsymbol{\lambda}$.

The *marginal likelihood* of $\boldsymbol{\beta}$ based on $\mathbf{A}$ is

$$L_{\mathrm{m}}(\boldsymbol{\beta}; \mathbf{a}) \propto f(\mathbf{a}; \boldsymbol{\beta}), \qquad (2)$$

and this could be used for inference about $\boldsymbol{\beta}$. In general, there is a loss of information in restricting attention to (2) for inference; sometimes, however, invariance or other arguments suggest that A contains the whole of the available information about $\boldsymbol{\beta}$. Even when such arguments do not apply, however, there may still be advantage to using (2) as the basis of inference since it conveniently eliminates $\boldsymbol{\lambda}$.

Two examples serve to illustrate the ideas.

Example 1

Suppose that variability in the measurement of blood glucose is of interest and that pairs of measurements are taken on n independent individuals. Thus, we might assume that X_{1i}, X_{2i} are independent $\mathrm{N}(\lambda_i, \beta^2)$ variates, where λ_i represents the true glucose level for the ith individual and β^2, the variance of the measurement error, is of interest. The maximum likelihood estimate of β^2 is

$$\hat{\beta}^2 = \frac{\sum (x_{1i} - x_{2i})^2}{(4n)},$$

which converges to $\beta^2/2$ in probability as $n \to \infty$. This is an instance in which the mle is inconsistent.

A marginal likelihood, in this problem, is naturally based on the statistics, $A_i = X_{1i} - X_{2i}$, $i = 1, \ldots, n$, which are independent $\mathrm{N}(0, 2\beta^2)$ variates. This gives rise to the marginal likelihood

$$L_{\mathrm{m}}(\beta^2; a) = \beta^{-n} \exp\left[\frac{-\sum a_i^2}{(4\beta^2)}\right].$$

The corresponding marginal mle, $\hat{\beta}^2_{\mathrm{m}} = \sum a_i^2/(2n)$ converges in probability to the correct value β^2 as $n \to \infty$. The choice of the A_is as the basis for inference about β^2 is a natural one; the difference in the measurements for each individual provides the information about β^2 intuitively.

Example 2

A second example arises in Cox's **proportional hazards** model [3] (*see* **Cox Regression Model**). The hazard function for the time to failure T is

$$\lambda(t; z) = \lambda_0(t) \exp(\mathbf{z}'\boldsymbol{\beta}), \qquad (3)$$

where $\mathbf{z} = (z_1, \ldots, z_p)'$ is a vector of fixed covariates. In this model, the baseline hazard rate $\lambda_0(t)$ is left arbitrary and the covariates z are assumed to act multiplicatively on the baseline rate with $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)'$, the vector of regression parameters being of primary interest.

Suppose that $T_1, \ldots, T_n$ is a sample with covariates $\mathbf{z}_1, \ldots, \mathbf{z}_n$. Let $T_{(1)}, \ldots, T_{(n)}$ be the **order statistic** from $T_1, \ldots, T_n$ with corresponding covariates $\mathbf{z}_{(1)}, \ldots, \mathbf{z}_{(n)}$ and let $\mathbf{R} = (R_1, \ldots, R_n)$ be the rank vector. Thus, R_i is the **rank** of the variate T_i among $(T_1, \ldots, T_n)$. The distribution of $\mathbf{R}$ can be shown to be

$$f(\mathbf{r}; \boldsymbol{\beta}) = \Pr(\mathbf{R} = \mathbf{r})$$

$$= \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{z}'_{(i)}\boldsymbol{\beta})}{\sum_{j=1}^{n} \exp(\mathbf{z}'_{(j)}\boldsymbol{\beta})} \right], \qquad (4)$$

and this defines a marginal likelihood for $\boldsymbol{\beta}$. This likelihood (4) is identical to Cox's **partial likelihood** [3]. These arguments can be extended to allow right **censoring** in the data [6].

We conclude with a number of remarks.

1. Marginal likelihood was first introduced by Fraser [4, 5] in the context of the structural model. In his work and in the related work of Kalbfleisch & Sprott [8], the A_is are allowed to depend on the parameter $\boldsymbol{\beta}$.
2. Group invariance arguments can be used in both of the examples given here to justify the use of **A** or **R** as the basis of inference for $\boldsymbol{\beta}$. Barnard [1] describes these arguments in general, and Kalbfleisch & Prentice [6, 7] apply them to the proportional hazards model (3). Other approaches to assessing the "**sufficiency**" of **A** for inference have been discussed by Sprott [9] and Barndorff-Nielsen [2], among others.
3. Marginal likelihood is one of several methods for obtaining derived likelihoods about a parameter of interest. Conditional likelihood (*see* **Conditional Probability**), partial likelihood and **profile likelihood** are other approaches which can apply, depending upon the structure of the statistical problem.

References

[1] Barnard, G.A. (1963). Some aspects of the fiducial argument, *Journal of the Royal Statistical Society, Series B* **25**, 111–114.

[2] Barndorff-Nielsen, O. (1976). Nonformation, *Biometrika* **63**, 567–571.

[3] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] Fraser, D.A.S. (1967). Data transformations and the linear model, *Annals of Mathematical Statistics* **38**, 1456–1465.

[5] Fraser, D.A.S. (1968). *The Structure of Inference*. Wiley, New York.

[6] Kalbfleisch, J.D. & Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model, *Biometrika* **60**, 267–278.

[7] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

[8] Kalbfleisch, J.D. & Sprott, D.A. (1970). Application of likelihood methods to models involving large number of parameters (with discussion), *Journal of the Royal Statistical Society, Series B* **32**, 175–208.

[9] Sprott, D.A. (1975). Marginal and conditional sufficiency, *Biometrika* **62**, 599–605.

JOHN D. KALBFLEISCH

Marginal Models for Multivariate Survival Data

Multivariate survival or failure-time data arise when each study subject may experience several events or when there exists some natural or artificial grouping of subjects which induces dependence among failure times of the same group. Biomedical examples include the sequence of tumor recurrences or infection episodes, the development of physical symptoms or diseases in several organ systems, the occurrence of blindness in the left and right eyes, the onset of a disease among family members, the initiation of cigarette smoking by classmates, and the appearance of tumor in litter-mates exposed to a carcinogen.

Suppose that there are n independent units each of which can potentially experience K types of failures. Let T_{ik} be the time when the kth type of failure occurs on the ith unit, and let C_{ik} be the corresponding **censoring** time. Define $X_{ik} = \min(T_{ik}, C_{ik})$ and $\Delta_{ik} = I(T_{ik} \le C_{ik})$, where $I(\cdot)$ is the indicator function (*see* **Dummy Variables**). Also, let $\mathbf{Z}_{ik}(\cdot) = [Z_{1ik}(\cdot), \ldots, Z_{pik}(\cdot)]'$ denote a p-vector of possibly **time-dependent covariates** for the ith unit with respect to the kth type of failure. The failure time vector $\mathbf{T}_i = (T_{i1}, \ldots, T_{iK})$ and the censoring time vector $\mathbf{C}_i = (C_{i1}, \ldots, C_{iK})$ are assumed to be independent, conditional on the covariate vector $\mathbf{Z}_i = (\mathbf{Z}'_{i1}, \ldots, \mathbf{Z}'_{iK})$, $i = 1, \ldots, n$. The units are allowed to have unequal numbers of failures, which is achieved by setting C_{ik} to zero whenever T_{ik} is missing.

It is natural and convenient to formulate the marginal distribution for each type of failure with a **proportional hazards model**. Depending on whether the baseline hazard functions are different or identical among the K types of failures, the marginal hazard function for the kth type of failure on the ith unit is

$$\boldsymbol{\lambda}_k(t; \mathbf{Z}_{ik}) = \boldsymbol{\lambda}_{0k}(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_{ik}(t)], \tag{1}$$

or

$$\boldsymbol{\lambda}_k(t; \mathbf{Z}_{ik}) = \boldsymbol{\lambda}_0(t) \exp[\boldsymbol{\beta}' \mathbf{Z}_{ik}(t)], \tag{2}$$

where $\boldsymbol{\lambda}_{0k}(t)$, $k = 1, \ldots, K$, and $\boldsymbol{\lambda}_0(t)$ are unspecified baseline hazard functions, and $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters. In some applications it is necessary to allow $\boldsymbol{\lambda}_{0k}(t)$, $k = 1, \ldots, K$, to be different, whereas in others it suffices to assume a common baseline hazard function. In both models (1) and (2), we set $\boldsymbol{\beta}$ to be the same among the K submodels, which entails no loss of generality since this structure can always be achieved by introducing appropriate type-specific covariates.

Inference Procedures

If all the failure times were independent, then the **partial likelihood** functions for $\boldsymbol{\beta}$ would be

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{\exp[\boldsymbol{\beta}' \mathbf{Z}_{ik}(X_{ik})]}{\sum_{j=1}^{n} Y_{jk}(X_{ik}) \exp[\boldsymbol{\beta}' \mathbf{Z}_{jk}(X_{ik})]} \right\}^{\Delta_{ik}}$$

under model (1) and

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{\exp[\boldsymbol{\beta}' \mathbf{Z}_{ik}(X_{ik})]}{\sum_{j=1}^{n} \sum_{l=1}^{K} Y_{jl}(X_{ik}) \exp[\boldsymbol{\beta}' \mathbf{Z}_{jl}(X_{ik})]} \right\}^{\Delta_{ik}}$$

under model (2), where $Y_{ik}(t) = I(X_{ik} \ge t)$. The corresponding score functions would be

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left[\mathbf{Z}_{ik}(X_{ik}) - \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta}, X_{ik})}{S_k^{(0)}(\boldsymbol{\beta}, X_{ik})} \right] \tag{3}$$

and

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left[\mathbf{Z}_{ik}(X_{ik}) - \frac{\overline{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, X_{ik})}{\overline{S}^{(0)}(\boldsymbol{\beta}, X_{ik})} \right], \tag{4}$$

where

$$\mathbf{S}_k^{(0)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_{jk}(t) \exp\left[\boldsymbol{\beta}' \mathbf{Z}_{jk}(t)\right],$$

$$\mathbf{S}_k^{(1)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_{jk}(t) \exp\left[\boldsymbol{\beta}' \mathbf{Z}_{jk}(t)\right] \mathbf{Z}_{jk}(t),$$

$$k = 1, \ldots, K,$$

and

$$\overline{\mathbf{S}}^{(r)}(\boldsymbol{\beta}, t) = \sum_{k=1}^{K} \mathbf{S}_k^{(r)}(\boldsymbol{\beta}, t), \; r = 0, 1.$$

In both cases, the solution to $[\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}]$ is denoted by $\hat{\boldsymbol{\beta}}$.

Although the failure times within the same unit tend to be **correlated**, the estimator $\hat{\boldsymbol{\beta}}$ can be shown to be consistent for $\boldsymbol{\beta}$ and asymptotically p-variate normal provided that the marginal models are correctly specified. However, the conventional **covariance matrix** estimator $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$, where $\mathcal{I}(\boldsymbol{\beta}) = -\partial^2 \log L(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^2$, is no longer valid, the reason being that $\mathcal{I}(\boldsymbol{\beta})$ is not the covariance matrix of $\mathbf{U}(\boldsymbol{\beta})$ in the presence of intraclass dependence. By approximating $\mathbf{U}(\boldsymbol{\beta})$ with a sum of independent and identically distributed zero-mean random vectors, one can show that, for large n and relatively small K, the random vector $\mathbf{U}(\boldsymbol{\beta})$ is approximately zero-mean normal with covariance matrix estimator

$$\mathbf{V}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbf{W}_{ik}(\hat{\boldsymbol{\beta}}) \mathbf{W}_{il}(\hat{\boldsymbol{\beta}})',$$

where

$$\begin{aligned} \mathbf{W}_{ik}(\boldsymbol{\beta}) = {} & \Delta_{ik} \left[\mathbf{Z}_{ik}(X_{ik}) - \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta}, X_{ik})}{S_k^{(0)}(\boldsymbol{\beta}, X_{ik})} \right] \\ & - \sum_{j=1}^{n} \frac{\Delta_{jk} Y_{ik}(X_{jk}) \exp[\boldsymbol{\beta}' \mathbf{Z}_{ik}(X_{jk})]}{S_k^{(0)}(\boldsymbol{\beta}, X_{jk})} \\ & \times \left[\mathbf{Z}_{ik}(X_{jk}) - \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta}, X_{jk})}{S_k^{(0)}(\boldsymbol{\beta}, X_{jk})} \right] \end{aligned}$$

and

$$\begin{aligned} \mathbf{W}_{ik}(\boldsymbol{\beta}) = {} & \Delta_{ik} \left[\mathbf{Z}_{ik}(X_{ik}) - \frac{\overline{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, X_{ik})}{\overline{S}^{(0)}(\boldsymbol{\beta}, X_{ik})} \right] \\ & - \sum_{j=1}^{n} \sum_{l=1}^{K} \frac{\Delta_{jl} Y_{ik}(X_{jl}) \exp[\boldsymbol{\beta}' \mathbf{Z}_{ik}(X_{jl})]}{\overline{S}^{(0)}(\boldsymbol{\beta}, X_{jl})} \\ & \times \left[\mathbf{Z}_{ik}(X_{jl}) - \frac{\overline{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, X_{jl})}{\overline{S}^{(0)}(\boldsymbol{\beta}, X_{jl})} \right] \end{aligned}$$

under models (1) and (2), respectively. Consequently, $\hat{\boldsymbol{\beta}}$ is approximately normal with covariance matrix estimator $\mathbf{D}(\hat{\boldsymbol{\beta}}) = \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}) \mathbf{V}(\hat{\boldsymbol{\beta}}) \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$. We call $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$ and $\mathbf{D}(\hat{\boldsymbol{\beta}})$ the naive and robust estimators, respectively. In the case of $K = 1$, the matrix $\mathbf{D}(\hat{\boldsymbol{\beta}})$ reduces to the Lin–Wei [14] robust covariance matrix estimator for the maximum partial likelihood estimator under **misspecified** proportional hazards models. To test the global hypothesis that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, one may use the chi-square statistic $\mathbf{U}'(\boldsymbol{\beta}_0) \mathbf{V}^{-1}(\boldsymbol{\beta}_0) \mathbf{U}(\boldsymbol{\beta}_0)$ or $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \mathbf{D}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$; to test the general linear hypothesis H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{L}$ is an $r \times p$ matrix of constants and $\mathbf{d}$ is an $r \times 1$ vector of constants, one refers $(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{d})' \{\mathbf{L}\mathbf{D}(\hat{\boldsymbol{\beta}})\mathbf{L}'\}^{-1} (\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{d})$ to the **chi-square distribution** with r **degrees of freedom**.

The above results are analogous to those of the **generalized estimation equations** (GEE) for the analysis of **marginal models** for **longitudinal data** with an independence working assumption. A similar idea can be used to estimate the cumulative baseline hazard functions $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(u) \mathrm{d}u$, $k = 1, \ldots, K$, and $\Lambda_0(t) = \int_0^t \lambda_0(u) \mathrm{d}u$ for models (1) and (2). Specifically, under the independence working assumption, the Aalen–Breslow type estimators for $\Lambda_{0k}(t)$ and $\Lambda_0(t)$ are

$$\hat{\Lambda}_{0k}(t) = \sum_{i=1}^{n} \frac{I(X_{ik} \le t) \Delta_{ik}}{S_k^{(0)}(\hat{\boldsymbol{\beta}}, X_{ik})}, \quad k = 1, \ldots, K, \quad (5)$$

and

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{I(X_{ik} \le t) \Delta_{ik}}{\overline{S}^{(0)}(\hat{\boldsymbol{\beta}}, X_{ik})}. \quad (6)$$

These estimators are consistent and asymptotically normal. In fact, the p-vector of random processes,

$$n^{1/2} [\hat{\Lambda}_{01}(t) - \Lambda_{01}(t), \ldots, \hat{\Lambda}_{0K}(t) - \Lambda_{0K}(t)]',$$

converges weakly to a p-dimensional zero-mean Gaussian random field, and the covariance between $\hat{\Lambda}_{0k}(t)$ and $\hat{\Lambda}_{0l}(s)$ can be estimated by $\sum_{i=1}^{n} \xi_{ik}(t; \hat{\boldsymbol{\beta}}) \xi_{il}(s; \hat{\boldsymbol{\beta}})$, where

$$\begin{aligned} & \xi_{ik}(t; \boldsymbol{\beta}) \\ & = \frac{I(X_{ik} \le t) \Delta_{ik}}{S_k^{(0)}(\boldsymbol{\beta}, X_{ik})} \\ & \quad - \sum_{j=1}^{n} \frac{I(X_{jk} \le t) \Delta_{jk} Y_{ik}(X_{jk}) \exp[\boldsymbol{\beta}' \mathbf{Z}_{ik}(X_{jk})]}{S_k^{(0)}(\boldsymbol{\beta}, X_{jk})^2} \\ & \quad - \left[\sum_{j=1}^{n} \frac{I(X_{jk} \le t) \Delta_{jk} S_k^{(1)}(\boldsymbol{\beta}, X_{jk})}{S_k^{(0)}(\boldsymbol{\beta}, X_{jk})^2} \right]' \\ & \quad \times \mathcal{I}^{-1}(\boldsymbol{\beta}) \sum_{l=1}^{K} \mathbf{W}_{il}(\boldsymbol{\beta}). \end{aligned}$$

In addition, $n^{1/2}[\hat{\Lambda}_0(t) - \Lambda_0(t)]$ converges weakly to a zero-mean Gaussian process, and the covariance between $\hat{\Lambda}_0(t)$ and $\hat{\Lambda}_0(s)$ can be estimated by

$\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{ik}(t;\hat{\boldsymbol{\beta}})\xi_{il}(s;\hat{\boldsymbol{\beta}})$, where

$$\begin{aligned}\xi_{ik}(t;\boldsymbol{\beta}) &= \frac{I(X_{ik}\le t)\Delta_{ik}}{\overline{S}^{(0)}(\boldsymbol{\beta},X_{ik})} \\ &\quad - \sum_{j=1}^{n}\sum_{l=1}^{K}\frac{I(X_{jl}\le t)\Delta_{jl}Y_{ik}(X_{jl})\exp[\boldsymbol{\beta}'\mathbf{Z}_{ik}(X_{jl})]}{\overline{S}^{(0)}(\boldsymbol{\beta},X_{jl})^2} \\ &\quad - \left[\sum_{j=1}^{n}\sum_{l=1}^{K}\frac{I(X_{jl}\le t)\Delta_{jl}\overline{\mathbf{S}}^{(1)}(\boldsymbol{\beta},X_{jl})}{\overline{S}^{(0)}(\boldsymbol{\beta},X_{jl})^2}\right]' \\ &\quad \times \mathcal{I}^{-1}(\boldsymbol{\beta})\mathbf{W}_{ik}(\boldsymbol{\beta}).\end{aligned}$$

The large-sample properties for the corresponding baseline survival function estimators $\exp[-\hat{\Lambda}_{0k}(t)]$, $k=1,\ldots,K$, and $\exp[-\hat{\Lambda}_0(t)]$ follow from the **delta method**. Furthermore, simple modifications can be made to estimate the survival functions associated with specific covariate values.

Software Availability

The estimators $\hat{\boldsymbol{\beta}}$, $\hat{\Lambda}_{0k}, k=1,\ldots,K$, and $\hat{\Lambda}_0$ are constructed under the independence working assumption, and therefore can be obtained from any existing software for the **Cox regression**. The robust covariance matrix estimator for $\hat{\boldsymbol{\beta}}$ is available in **S-PLUS**, SAS, and STATA packages, as well as in a special FORTRAN program [12]. The robust variance–covariance estimators for $\hat{\Lambda}_{0k}, k=1,\ldots,K$, and $\hat{\Lambda}_0$ have not been implemented in commercially available software packages.

An Example

We now provide an illustration with the well-known Diabetic Retinopathy Study [4], which was conducted by the National Eye Institute to evaluate the effectiveness of laser photocoagulation in delaying the onset of blindness in patients with diabetic retinopathy. The study enrolled 1742 patients. One eye of each patient was randomly selected for photocoagulation and the other eye was observed without treatment. The patients were followed over several years for the occurrence of blindness in their left and right eyes.

We confine our attention to a subset of the data with 197 high-risk patients previously analyzed by Huster et al. [7] and Lin [13]. By the end of the study, 54 treated eyes and 101 control eyes in this subsample had developed blindness. In this example, each patient could potentially experience blindness in both eyes; therefore, there are two failure types with $k=1$ and 2, denoting the left and right eyes, respectively. Since there are no biological differences between the left and right eyes, it is natural to assume a common baseline hazard function for the two failure types.

As mentioned above, the primary objective of this study was to assess whether laser photocoagulation delays the occurrence of blindness. Because juvenile and adult diabetes have very different courses, it is desirable to examine how the age at onset of diabetes may affect the time to blindness. Thus, we consider model (2) with $\mathbf{Z}_{ik}=(Z_{1ik},Z_{2ik},Z_{3ik})', i=1,\ldots,197; k=1,2$, where

$$Z_{1ik}=\begin{cases}1, & \text{if the } k\text{th eye of the } i\text{th patient was on treatment,}\\ 0, & \text{otherwise;}\end{cases}$$

$$Z_{2ik}=\begin{cases}1, & \text{if the } i\text{th patient had adult onset diabetes,}\\ 0, & \text{if the } i\text{th patient had juvenile onset diabetes;}\end{cases}$$

and $Z_{3ik}=Z_{1ik}\times Z_{2ik}$. The results for the estimation of the regression parameters are shown in Table 1. The robust standard error estimates are appreciably smaller than the naive estimates, the latter ignoring the dependence between the left and right eyes. The treatment appears to be effective, and this effect is much stronger for adult-onset diabetes than for juvenile-onset diabetes.

Figure 1 displays the estimates and pointwise 95% confidence intervals for the survival functions, namely, the probabilities of retained visual acuity, for adult-onset diabetes, separated by treatment groups. As expected, these probabilities are much higher for the treated eyes than for the untreated ones.

Table 1

Variable	Parameter estimate	Stand. error estimate	
		Naive	Robust
Treatment (Z_1)	−0.425	0.218	0.185
Diabetic type (Z_2)	0.341	0.199	0.196
Interaction ($Z_1\times Z_2$)	−0.846	0.351	0.304

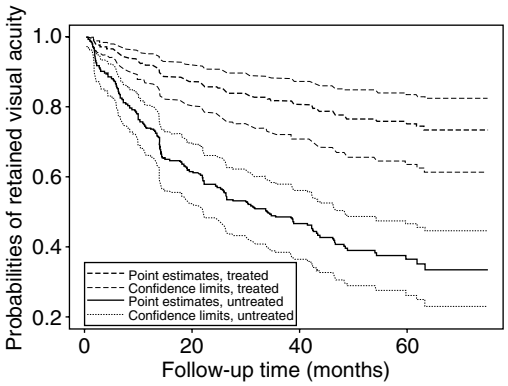

Figure 1 Estimates and pointwise 95% confidence intervals for survival functions

Further Results

The estimation of $\boldsymbol{\beta}$ under models (1) and (2) was first studied by Wei et al. [23] and Lee et al. [9], respectively, and further developed by Lin [13], while the estimation of $\Lambda_{0k}, k = 1, \ldots, K$, and Λ_0 was investigated by Spiekerman & Lin [22]. The latter authors established a rigorous asymptotic theory for the estimation of both the regression parameters and baseline hazard functions under a general marginal model which allows $M, 1 \leq M \leq K$, different baseline hazard functions among the K types of failures. In a separate paper, they [21] developed a class of graphical and numerical techniques for checking the adequacy of models (1) and (2). The readers are referred to the aforementioned papers for further theoretical details as well as additional numerical examples. Incidentally, Huster et al. [7] studied model (2) with a parametric baseline hazard function, while Guo & Lin [6] deal with discrete-time versions of models (1) and (2).

Liang et al. [11] proposed a different procedure for analyzing model (2). Their estimating function is similar to (4), but they replaced $\overline{\mathbf{S}}^{(1)}/\overline{S}^{(0)}$ by an analog which exploits pairwise comparisons of independent observations. The actual form of their **estimating function** is

$$\sum_{i=1}^{n}\sum_{k=1}^{K} I[n_i(X_{ik}) > 0]\Delta_{ik} \times \left[\mathbf{Z}_{ik}(X_{ik}) - n_i^{-1}(X_{ik})\sum_{j\neq i}\sum_{l}\mathbf{e}_{ik,jl}(\boldsymbol{\beta}, X_{ik})\right],$$

where $n_i(t) = \sum_{j\neq i}\sum_l Y_{jl}(t)$ and

$$\mathbf{e}_{ik,jl}(\boldsymbol{\beta}, t) = \frac{Y_{ik}(t)\mathbf{Z}_{ik}(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_{ik}(t)] + Y_{jl}(t)\mathbf{Z}_{jl}(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_{jl}(t)]}{Y_{ik}(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_{ik}(t)] + Y_{jl}(t)\exp[\boldsymbol{\beta}'\mathbf{Z}_{jl}(t)]}.$$

The resultant estimator is **consistent** and asymptotically normal. The relative efficiency of $\hat{\boldsymbol{\beta}}$ vs. the Liang et al. estimator has not been investigated.

Estimating functions (3) and (4) were derived under the independence working assumption. As in the case of **longitudinal data**, it may be more efficient to use estimating functions that take into account the nature of dependence explicitly. This amounts to incorporating certain weight functions into estimating functions (3) and (4). The resultant estimators remain consistent and asymptotically normal with a sandwich-type variance estimator under mild regularity conditions on the weight function. Due to censoring and the nonlinear nature of the proportional hazards model, it is difficult to construct optimal weight functions. Cai & Prentice [2] investigated a weight function that is the inverse of the covariance matrix of the marginal martingales associated with the T_{ik}s. Their theoretical calculations and simulation studies indicated that the efficiency gains in using such weighted estimating functions over estimating functions (3) and (4) are small unless the correlations of failure times are unusually high.

There has been considerable research on semiparametric multivariate failure time distributions which characterize the strength of association among failure time components by a limited number of parameters while leaving the forms of the marginal distributions unspecified (e.g. [3, 18, 1]). One may extend these multivariate distributions by formulating their marginal distributions with model (1) or (2). One may then estimate the marginal regression parameters and baseline hazard functions by (3) and (5) or (4) and (6) and proceed to estimate the association parameters by the pseudo-**maximum likelihood** method [5]. This approach was mentioned by Bandeen-Roche & Liang [1], but its inferential properties have yet to be investigated.

Prentice & Hsu [19] studied simultaneous regression on the marginal hazard ratios and pairwise dependencies, which is analogous to the regression on the means and covariances of noncensored multivariate responses [20]. They used the estimating

function of Cai & Prentice [2] for the marginal hazard ratio parameters and developed a similar *ad hoc* estimating function for the dependence parameters. They showed that the solutions to this pair of estimating functions are consistent and asymptotically normal, with a sandwich-type covariance matrix estimator.

The **accelerated failure-time** and **additive hazards models** are two important alternatives to the proportional hazards model. The former relates the logarithm of the failure time linearly to the covariates [8], while the latter relates the conditional hazard function linearly to the covariates [16]. One may formulate the marginal distributions of multivariate failure time data with accelerated failure time models or additive hazards models rather than proportional hazards models. The corresponding inference procedures were studied, respectively, by Lin & Wei [15] and Lee et al. [10], and by Lin & Ying [17].

References

[1] Bandeen-Roche, K.J. & Liang, K.-Y. (1996). Modelling failure time associations in data with multiple levels of clustering, *Biometrika* **83**, 29–39.

[2] Cai, J. & Prentice, R.L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data, *Biometrika* **82**, 151–164.

[3] Clayton, D.G. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65**, 141–151.

[4] Diabetic Retinopathy Study Research Group (1981). Diabetic retinopathy study, *Investigative Ophthalmology and Visual Science* **21**, 149–226.

[5] Gong, G. & Samaniego, F.J. (1981). Pseudo maximum likelihood, *Annals of Statistics* **9**, 861–869.

[6] Guo, S.W. & Lin, D.Y. (1994). Regression analysis of multivariate grouped survival data, *Biometrics* **50**, 632–639.

[7] Huster, W.J., Brookmeyer, R. & Self, S.G. (1989). Modelling paired survival data with covariates, *Biometrics* **45**, 145–156.

[8] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[9] Lee, E.W., Wei, L.J. & Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Dordrecht, pp. 237–247.

[10] Lee, E.W., Wei, L.J. & Ying, Z. (1993). Linear regression analysis for highly stratified failure time data, *Journal of the American Statistical Association* **88**, 557–565.

[11] Liang, K.-Y., Self, S.G. & Chang, Y.-C. (1993). Modeling marginal hazards in multivariate failure-time data, *Journal of the Royal Statistical Society, Series B* **55**, 441–453.

[12] Lin, D.Y. (1993). MULCOX2: a general computer program for the Cox regression analysis of multivariate failure time data, *Computer Methods and Programs in Biomedicine* **40**, 279–293.

[13] Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach, *Statistics in Medicine* **13**, 2233–2247.

[14] Lin, D.Y. & Wei, L.J. (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association* **84**, 1074–1078.

[15] Lin, J.S. & Wei, L.J. (1992). Linear regression analysis for multivariate failure time observations, *Journal of the American Statistical Association* **87**, 1071–1097.

[16] Lin, D.Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61–71.

[17] Lin, D.Y. & Ying, Z. (1997). Additive hazards regression models for survival data, in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D.Y. Lin & T.R. Fleming, eds. Springer-Verlag, New York, pp. 185–198.

[18] Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**, 487–493.

[19] Prentice, R.L. & Hsu, L. (1997). Regression on hazard ratios and cross-ratios in multivariate failure time analysis, *Biometrika*, **84**, 349–363.

[20] Prentice, R.L. & Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics* **47**, 825–839.

[21] Spiekerman, C.F. & Lin, D.Y. (1996). Checking the marginal Cox model for correlated failure time data, *Biometrika* **83**, 143–156.

[22] Spiekerman, C.F. & Lin, D.Y. (1996). Marginal Regression Models for Multivariate Failure Time Data, *Technical Report 144*. Department of Biostatistics, University of Washington.

[23] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

D.Y. LIN

Marginal Models

When several response variables are simultaneously of interest on each subject in a study, we may expect these responses to be interdependent simply because the same subject is involved for all observations. Because of this dependence among the observations, a **multivariate distribution** will be required to model them adequately. In most cases, repeated measurements will be involved; for simplicity, we shall restrict our examples to such studies. In other words, the *same* response variable will be measured several times on subjects, either because they are found in clusters or because they are observed longitudinally over time [11], (*see* **Longitudinal Data Analysis, Overview**).

The full multivariate distribution will describe the dependence among the responses on each subject. Such distributions can always be factored into a product of univariate marginal and conditional distributions in a number of ways (*see* **Marginal Probability**; **Conditional Probability**). As well, as many univariate marginal distributions exist as the number of dimensions of the response variable. Models based on multivariate distributions can be parameterized in a number of different ways, using various combinations of the conditional and/or marginal distributions, although care must be taken in such constructions [2]. Thus, if a proper probability model is constructed for such data, it necessarily contains both marginal and conditional aspects. A "marginal model" refers to a multivariate model where emphasis is placed on the margins; usually, parameters in the margins are related in simple ways to the **covariates**. As we shall see, this generally induces complex relationships for the conditional distributions.

In certain situations, such as in the experimentation of a **clinical trial**, the dependence relations among responses will be of direct interest. Thus, for example, in a longitudinal setting, we may be concerned with how a response depends on the previous history of a subject, including dependence on previous responses. Then, study of conditional distributions will be appropriate [12]. However, in other situations, such as epidemiological population studies of **prevalence**, we may be interested in the marginal distribution of responses within the population at each point in time or for each member of a cluster or matched group (*see* **Clustering**). The term *marginal* means that the distribution of each response separately is concerned, conditional on covariates but not on any of the other responses. In such a marginal approach, the observations are analyzed as if they were a series of **cross-sectional studies** instead of repeated measurements on the same individuals.

A special kind of conditional distribution involves **random effects**. By conditioning on one or more latent variables accounting for heterogeneity among subjects, a multivariate distribution is induced. In the simplest case, it has uniform dependence among all responses within a cluster. In what follows, the term, conditional distribution, will refer to conditioning directly on the other responses of a subject, and not to random effects.

Clustered and longitudinal studies pose fairly distinct problems. The observations on a cluster are generally not ordered, so that they are interchangeable. The same marginal model will often be appropriate for any member of the cluster, although, in some cases, responses will depend on cluster size. However, in a longitudinal study, observations are ordered in time, so that early observations cannot be made to depend on more recent ones. A model at any time point should be constructed in ignorance of future observations. In most situations, the marginal distribution may be expected to change over time.

To see the relationships between multivariate, conditional, and marginal distributions, consider the simplest example of a repeated categorical response. If we have two observations of the response, the joint or multivariate (here bivariate) distribution can be represented by the probabilities, π_{ij}, where the two indices indicate the combination of categories observed. The marginal distributions are given by the probabilities, $\pi_{i\bullet} = \sum_j \pi_{ij}$ and $\pi_{\bullet j} = \sum_i \pi_{ij}$, and the conditional distributions by $\pi^1_{i|j} = \pi_{ij}/\pi_{\bullet j}$ and $\pi^2_{j|i} = \pi_{ij}/\pi_{i\bullet}$.

To take a concrete example, consider first a clustered, rather than a longitudinal, response. Suppose that we have a sample of subjects for whom we classify each eye as having either good or poor vision. Observations can be represented as a simple **2 × 2 table** with entries being the frequencies, n_{ij}, where the indices refer to the responses on the two eyes. Information on the bivariate distribution, π_{ij}, is contained in the body of the table. A conditional distribution corresponds to fixing the value of one of the two responses, so that information is obtained from the corresponding row or column. We can reconstruct the complete multivariate distribution from a pair of

conditional distributions. Finally, information on the marginal distributions is contained in the marginal totals. This pair of marginal distributions, by itself, does not allow us to reconstruct the complete multivariate distribution. Attempts to do so involve what is called the **ecologic fallacy** [12].

Marginal distributions inform us about the average state of a population. They tell us nothing about the relationships among the responses for individual members of the population. Two responses may have identical marginal distributions without there being a similar dependence between the responses for many, or indeed any, individual subjects. Consider two rather extreme fictitious examples of the joint distributions that might correspond to results under two treatments:

	Left eye			Left eye	
Right eye	Good	Poor	Right eye	Good	Poor
Good	0.01	0.47	Good	0.46	0.02
Poor	0.47	0.05	Poor	0.02	0.50

In the left table, there is a large probability that only one eye of any given individual will be good; the treatment helps only one eye, but for almost all individuals. In the right table, the probability is high that both eyes will be similar; this treatment helps both eyes of half of the individuals. Thus, in the first case, the conditional probability is 0.02 ($= 0.01/0.48$) that one eye will be good, given that the other is, while, in the second case, it is 0.96 ($= 0.46/0.48$). However, marginally, under both treatments, both eyes have exactly the same distribution, known as marginal homogeneity, with the same probability of 0.48 for a good left and a good right eye. Thus, the fact that the marginal probabilities of both left and right eyes being good are the same tells us nothing about whether both eyes of a given subject will be good. Conversely, if the relationship between marginal distributions were different under two treatments, this would not exclude the individual response relationships between eyes being the same for both treatments [1], (*see* **Matched Pairs With Categorical Data**).

For the cluster of two eyes, a reasonable bivariate distribution that allows for interchangeability would set $\pi_{12} = \pi_{21}$, or equivalently, the conditional probabilities, $\pi^1_{1|1} = \pi^2_{1|1}$, to yield a trinomial distribution. From this bivariate distribution, the marginal probabilities are easily obtained as $\pi_{1\bullet} = \pi_{\bullet 1} = \pi_{11} + (\pi_{12} + \pi_{21})/2$.

Consider now a slightly more complex example with a binary response, this time in a longitudinal context [15]. Subjects are followed over time, all beginning in one state, but at some point switching to a second state. This may be represented by a horizontal line, running through time on a graph, starting at level one but jumping vertically down to level zero at the switch point, then continuing horizontally at that level. Each subject may change state at a different time, so that the vertical lines do not coincide. The average or typical individual will have the vertical line situated at the mean of all jump times. However, the marginal model, calculated from the mean number in state one at each time point, will be a sigmoid curve dropping slowly from one and flattening off at zero. Individuals cannot follow such a curve because they must be in one state or the other, not part way in between. The marginal curve gives the average number of subjects in each state at each point in time, but tells us nothing about the trajectory of a typical individual.

From these examples, we can see that marginal probabilities refer to averages, not individuals. Thus, models for such probabilities are sometimes called *population-averaged*, whereas conditional models are called *subject-specific* [16]. The choice between the two types will depend on the question being asked. For example, if the response is the presence or the absence of repeated infections under two treatments, the population-averaged model describes the global difference in infection rates between the treatments, while the subject-specific one looks at the probability of infection of a typical individual, given treatment. Effects in population-averaged models depend on the degree of heterogeneity in the population; the same process in two populations with different heterogeneity will yield different population-averaged effects. The dependence of marginal response on an **explanatory variable** will be smaller than the corresponding individual average dependence in a random effects model, this difference increasing with heterogeneity.

Care must be taken that a marginal model does correspond to a population of interest. For example, in the setting of a clinical trial, the "population" is rather artificial for a number of reasons including the facts that the subjects are volunteers and that all are started on treatments at arbitrary points in time. In

such a context, marginal statements would appear to have limited value.

A marginal model can be constructed in two opposing ways. We may start with some known multivariate distribution (or set of conditional distributions) and construct the marginal distributions by summing or integration. Or we may specify directly the marginal distributions. We have already seen that, in the latter case, the multivariate distribution will not be uniquely defined, so that additional relationships will have to be specified. One common way is by using **copulas** [8]. Each approach has certain advantages and disadvantages.

First, we should note that the **multivariate normal distribution** is quite exceptional and should not be taken as a general example for model construction. Both the conditional and the marginal distributions are easy to derive, and both are also **normal**. The Bernoulli distribution (*see* **Binary Data**) has similar properties, but the **binomial** does not. Both the normal and Bernoulli distributions are very special.

Generally, if the conditional distributions are of a simple known form, the marginal distributions will be complex, usually analytically intractable, and vice versa. This means that if we start with some reasonable conditional distributions that might be appropriate to describe the dependencies of the phenomenon under study at the individual level, the marginal distributions will be complex and difficult to handle. But if we start with some simple and well-known marginal distributions that we find easy to understand at the population level, we are implicitly imposing one of a number of possible conditional distributions implying a complex relationship at the individual level.

Consider an example involving the number of infections in one month in each eye. If we took the conditional distribution, say for the left eye given the condition of the right eye, to be **Poisson**, we would only be assuming that individuals under the same condition of the right eye had the same distribution of responses for the left eye. The marginal distribution would be a weighted average of the two Poisson distributions. On the other hand, if we took the marginal to be Poisson, again for the left eye, we would be assuming that *all* members of the population had on average this same probability distribution.

Scientifically, it should be clear that the first approach is more reasonable. Marginal or population descriptions do not generally have a meaning on their own, but only as built upon acceptable underlying individual dependence relationships. Nevertheless, certain statisticians have argued that the second is justifiable to answer some population questions.

A conditional model generally depends on the number of other responses to which a given response is related (think of a cluster of teeth instead of eyes), while a marginal model does not. The latter is said to be reproducible [10]. Thus, a marginal model can have the same interpretation for clusters of all sizes, while a conditional model may not. Random effects models also have this characteristic. Hence, direct conditioning may not be appropriate for clustered data. On the other hand, reproducibility is generally not a desirable property for longitudinal data. Unequal sized longitudinal "clusters" have histories of different lengths.

For simple categorical data, such as two-way tables, both conditional and marginal models can be constructed fairly easily. Suppose that we want to study the influence of some explanatory variable, x_k, on the two binary responses, $i = 1, 2$ and $j = 1, 2$. The state of two eyes, used above, would be one example, but the study could also be longitudinal. We can construct a model based on a series of **multinomial distributions** such that $\sum_i \sum_j \pi_{ijk} = 1$ for all k. Then, the **loglinear** regression functions for a conditional model, based on a bivariate Bernoulli distribution, are

$$\begin{aligned}
\log\left(\frac{\pi_{11k}\pi_{12k}}{\pi_{21k}\pi_{22k}}\right) &= \log\left(\frac{\pi^1_{1|1,k}\pi^1_{1|2,k}}{\pi^1_{2|1,k}\pi^1_{2|2,k}}\right) \\
&= 2\log\left(\frac{\dot{\pi}_{1\bullet k}}{\dot{\pi}_{2\bullet k}}\right) \\
&= \beta_{10} + \beta_{11}x_k \\
\log\left(\frac{\pi_{11k}\pi_{21k}}{\pi_{12k}\pi_{22k}}\right) &= \log\left(\frac{\pi^2_{1|1,k}\pi^2_{1|2,k}}{\pi^2_{2|1,k}\pi^2_{2|2,k}}\right) \\
&= 2\log\left(\frac{\dot{\pi}_{\bullet 1k}}{\dot{\pi}_{\bullet 2k}}\right) \\
&= \beta_{20} + \beta_{21}x_k \\
\log\left(\frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}\right) &= \beta_{30} + \beta_{31}x_k, \qquad (1)
\end{aligned}$$

where $\dot{\pi}_{\bullet jk}$ and $\dot{\pi}_{i\bullet k}$ are the geometric means. Although the conditional probabilities must be adjusted to follow the linear regression across tables, indexed by k, this allows the marginal frequencies, upon

which they are conditioned, to be held constant at their observed values. Inferences are then independent of these observed marginal totals, and of wide applicability. Although the conditional probability distributions are Bernoulli, the same for all subjects with a given value of x_k, the corresponding marginal distributions are weighted averages over all values of x_k.

Consider now the corresponding regression functions for a marginal model,

$$\begin{aligned}
\log\left(\frac{\pi_{11k}+\pi_{12k}}{\pi_{21k}+\pi_{22k}}\right) &= \log\left(\frac{\pi_{1\bullet k}}{\pi_{2\bullet k}}\right)\\
&= \beta_{10}+\beta_{11}x_k\\
\log\left(\frac{\pi_{11k}+\pi_{21k}}{\pi_{12k}+\pi_{22k}}\right) &= \log\left(\frac{\pi_{\bullet 1k}}{\pi_{\bullet 2k}}\right)\\
&= \beta_{20}+\beta_{21}x_k\\
\log\left(\frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}\right) &= \beta_{30}+\beta_{31}x_k. \qquad (2)
\end{aligned}$$

Here, we have chosen the log **odds ratio** to describe the dependence between the two binary responses (a less elegant solution being to use the **correlation**). The observed marginal frequencies are not held fixed, unless a saturated model is fitted because those in individual tables indexed by k, must be estimated so as to follow the linear regression function. Because inference is not made conditional on the observed marginal frequencies, this limits the applicability of any empirical conclusions concerning dependence, drawn from this model, to tables with the same marginal frequencies.

The conditional distributions obtained from a marginally-specified model or vice versa are complex because the marginal probabilities are sums of multivariate probabilities and probabilities are nonlinearly (here logit) related to the covariates. For example, in the second model, the marginal probabilities are Bernoulli, the same for all subjects with a given value of x_k, whereas the joint and conditional ones vary in some complex way among subjects with the same x_k. In contrast, in the first model, the conditional probabilities are Bernoulli, the same for all subjects with a given value of x_k.

These models have several interesting contrasting characteristics. The first of them, loglinear regression, is a **generalized linear model**, whereas the second is not. This means that the parameter estimates are considerably more difficult to obtain in the latter case. This is further complicated if correlations are used (for more than two responses), because inequality constraints must be applied. At the same time, the conditional model, but not the marginal one, fixes the marginal totals at their observed values, something that has often been considered to be a prerequisite for analyzing a contingency table. Finally, from general properties of the **exponential family**, the dependence parameters, (β_{31}, β_{31}), in the above marginal model are information orthogonal to the marginal regression parameters, $(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})$. In other words, the elements of the **information matrix** relating these parameters together are zero so that their estimates are asymptotically uncorrelated. This is not true in the conditional model. Fitzmaurice and Laird [4] and several other authors have developed more complex probability models based on marginal parameters.

To avoid the complexities of the specification of dependence relationships when marginal distributions are the primary point of interest, one widely promoted approach has been to set up regression equations only describing how the marginal responses are believed to depend on the explanatory variables (similar to the first two of the three equations for the marginal model above). As members of the generalized linear model family, the score equations for estimating the parameters are well understood. But because responses are not independent, some matrix of "working" correlations is introduced into these equations, yielding **generalized estimating equations** (GEE) [10, 16].

Such equations have the property that, if the regression is correctly specified, the point estimates of the regression coefficients will be asymptotically consistent no matter what "working" matrix is chosen (although there is no simple empirical way of checking correctness). However, this is accompanied by at least two major inconveniences. Except in special cases, the GEE corresponds to no probability model in the accepted sense of the term, that is, no model that allows one to calculate the probability of the observed or any future data. Thus, no **likelihood** function is available, singularly complicating the tasks of obtaining useful measures of precision of the point estimates and of comparing "models". Generally, only quasi-standard errors and a quasi-score function are available for making inferences and there has been considerable debate about the choice of the former. Standard errors are well-known to be unreliable in small samples of categorical data [5] so that care must be taken, in the same way as asymptotic

chi-squared tests need to be replaced by **Fisher's exact test** in sparse **contingency tables** (*see* **Exact Inference for Categorical Data**).

Although, the examples given here have involved simple binary responses, extensions to more complex **polytomous**, including ordinal, responses are available (*see* **Ordered Categorical Data**). However, a search of the literature shows that publishing on marginal models has reached a low level over the past five years after considerable activity in the preceding decade. Recent publications on more complex models include [3, 6, 7, 9, 13]. The reader may also wish to consult the review paper by Pendergast et al. [14].

References

[1] Agresti, A. (1989). A survey of models for repeated ordered categorical response data, *Statistics in Medicine* **8**, 1209–1224.

[2] Arnold, B.C., Castillo, E. & Sarabia, J.M. (2001). Quantification of incompatibility of conditional and marginal information, *Communications in Statistics* **A30**, 381–395.

[3] Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures, *Biometrika* **81**, 767–775.

[4] Fitzmaurice, G.M. & Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses, *Biometrika* **80**, 141–151.

[5] Hauck, W.W. & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis, *Journal of the American Statistical Association* **72**, 851–853.

[6] Heagerty, P.J. & Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference, *Statistical Science* **15**, 1–26.

[7] Huang, G.H., Bandeen-Roche, K. & Rubin, G.S. (2002). Building marginal models for multiple ordinal measurements, *Applied Statistics* **51**, 37–57.

[8] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

[9] Lang, J.B., McDonald, J.W. & Smith, P.W.F. (1999). Association-marginal modeling of multivariate categorical responses: a maximum likelihood approach, *Journal of the American Statistical Association* **94**, 1161–1171.

[10] Liang, K.Y., Zeger, S.L. & Qaquish, B. (1992). Multivariate regression for categorical data, *Journal of the Royal Statistical Society* **B54**, 3–40.

[11] Lindsey, J.K. (1999). *Models for Repeated Measurements*, 2nd Ed. Oxford University Press, Oxford.

[12] Lindsey, J.K. & Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials, *Statistics in Medicine* **17**, 447–469.

[13] Molenberghs, G. & Lesaffre, E. (1999). Marginal modelling of multivariate categorical data, *Statistics in Medicine* **18**, 2237–2255.

[14] Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M. & Fisher, M.R. (1996). A survey of methods for analyzing clustered binary response data, *International Statistical Review* **64**, 89–118.

[15] Sheiner, L.B., Beal, S.L. & Sambol, N.C. (1989). Study designs for dose-ranging, *Clinical Pharmacology and Therapeutics* **46**, 63–77.

[16] Zeger, S.L., Liang, K.Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

(*See also* **Correlated Binary Data**; **Linear Mixed Effects Models for Longitudinal Data**; **Marginal Models for Multivariate Survival Data**; **McNemar Test**; **Multilevel Models**; **Nonlinear Mixed Effects Models for Longitudinal Data**)

J.K. LINDSEY

Marginal Probability

In many situations, interest focuses on probability distributions for multiple random variables. For example, one may be studying height, weight, blood pressure, and cholesterol levels in a population, variables likely to be **correlated** with one another. Knowledge of the joint distribution of these variables allows one to calculate the probabilities associated with any particular outcome of interest. Marginal probabilities relate to the univariate distribution, or marginal distribution, associated with any of the variables under consideration.

To fix notation, first consider a **bivariate** model for two random variables X and Y. Let $f_{X,Y}(x, y)$ denote the joint probability mass function if X and Y are discrete or the joint probability density function if X and Y are continuous. If X and Y are discrete, the marginal probability mass functions of X and Y are given by

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

and

$$f_Y(y) = \sum_x f_{X,Y}(x, y),$$

where the summations are taken over all of the values of Y or X. In this case, the joint probability mass function can be written in tabular form, with the columns corresponding to the possible values of X and the rows to the values of Y. Then the marginal distribution of X corresponds to the column sums of the table, and the marginal distribution of Y corresponds to the row sums of the table.

When X and Y are continuous, the marginal probability density functions of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}x.$$

As a special case, when (X, Y) follows a **bivariate normal distribution** with means (μ_X, μ_Y) and **covariance matrix**

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix},$$

then the marginal probability density function of X follows a univariate **normal distribution** with mean μ_X and variance σ_X^2, and similarly for Y.

These marginal distributions can be used to compute probabilities or expectations that involve only X or Y. However, the marginal distributions do not completely describe the joint distribution of X and Y. In fact, many different joint distributions can yield the same marginal distributions. The variables X and Y are independent if and only if the joint distribution of (X, Y) is given by the product of the marginal distributions of X and Y. **Conditional probability** distributions refer to the distribution of one variable for a given value of the other variable.

For multivariate distributions with more than two variables, the corresponding summations or integrals are carried out over the complete range of the other variables under consideration. Extensions to mixtures of discrete and continuous variables are straightforward. For further information, see Casella & Berger [1, Chapter 4]. Regression modeling of multivariate responses sometimes focuses on modeling the marginal mean responses of the observations as a function of covariates, with the correlations between responses possibly being viewed as **nuisance parameters**. Diggle et al.[2, Chapter 8] review marginal modeling of multivariate discrete and continuous responses.

References

[1] Casella, G. & Berger, R.L. (1990). *Statistical Inference*. Wadsworth, Belmont.

[2] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

(*See also* **Longitudinal Data Analysis, Overview**; **Marginal Models**; **Marginal Models for Multivariate Survival Data**)

DAVID WYPIJ

Marker Processes

In many survival studies, individuals give rise to **stochastic processes**, called *marker processes*, that in some way measure the state of "health" of the individuals. Thus, the observed path of the marker process provides information on the propensity of the individual to fail. Such covariates have the potential to be useful in various ways and, in this brief introductory article, we attempt to identify some of these and to give entry points to the developing literature in this area.

Many examples of marker processes arise in survival studies (*see* **Survival Analysis, Overview**). In **clinical trials**, for example, it is common at each follow-up visit to take repeated measures of general health status or of the stage or severity of the disease. In equipment reliability studies, there are often repeated measures of the wear or degradation of the item under study. In the study of the time to breakdown of automobiles, for example, a simple and highly informative marker process is total kilometers traveled. In the study of infection with the Human Immunodeficiency Virus (HIV), much attention has been focused on the estimation of the distribution of "incubation time"; that is, the time from infection with HIV until the diagnosis of AIDS (*see* **AIDS and HIV**). Various marker processes, such as CD4-lymphocyte counts, have been studied and used to provide information both on the time since infection began and on the probability of developing AIDS.

Jewell & Kalbfleisch [5] outline some potential uses of marker processes and their classification forms the basis of the following summary:

1. *Improving estimation of a survival distribution*. In many instances the observed path of the marker provides information on the residual life of the individual under study. The basic idea is that the marker can be utilized to provide an estimate of the residual life for an individual who is **censored**. This can provide more and better information for the estimation of the survivor function (*see* **Survival Distributions and Their Characteristics**). In some instances this allows for adjustment for dependent or informative censoring mechanisms. Taylor et al. [14] and Robins & Rotnitzky [12] give good discussion and examples.
2. *Serving as a surrogate for survival in a comparative trial*. If the time to failure is typically long, then a full survival study may be prohibitively expensive or else require too long for completion. In such instances there may be substantial advantage to using marker processes as **surrogates** for the failure time in investigating the existence and size of treatment effects. In this approach it is required that the survival time be directly related to the marker so that a treatment effect on the marker will have a consequent effect on survival. This potential use of markers was the motivation of Cox [3] in his original paper in this area. Prentice [10] gives a detailed and comprehensive discussion of surrogate endpoints, potential uses, and caveats.
3. *Estimating the time of onset of a disease*. Sometimes the time of onset of a disease is not or cannot be observed. In such instances a marker measured on an affected individual may provide information on the elapsed time since onset. For example, in HIV the time of infection is typically unknown and markers may assist in estimating the time of onset. In a comparative trial, **confounding** of the time of onset with treatment or other comparison has the potential to introduce **bias** into the estimation of effects. The marker's information on time of onset can be used to adjust comparisons in this context and so help to compensate for the onset of confounding. Applications involving unknown onset have been considered by several authors. In the estimation of the incubation period in AIDS, Berman [1] and Munõz et al. [8] provide examples. Rai & Matthews [11] consider the use of tumor size at death in animal carcinogenicity trials (*see* **Tumor Incidence Experiments**) to estimate the unobserved time since tumor onset. Brookmeyer & Gail [2] and Munõz et al. [7] consider some theoretical issues associated with the onset of confounding.

In modeling the relationship between the marker process and the survival probabilities, it is convenient to utilize the hazard function. Let $\{X(t) : t > 0\}$ represent the marker process and let $T > 0$ be the time to failure. Conditionally upon the current and past

values of the marker, the **hazard** function or failure rate is naturally specified as

$$\lambda(t|X(s), 0 \le s \le t) = \lim_{\Delta t \to 0} \frac{\Pr\{T \in [t, t+\Delta t) | T \ge t, X(s),\ 0 \le s \le t\}}{\Delta t}. \quad (1)$$

Various more specific parametric and nonparametric models based on (1) could be specified and considered. In applications, other covariates, either fixed or **time-dependent**, may also be present and one may wish to extend (1) to incorporate them into the model. In a comparative trial, for example, (1) could be extended to include treatment effects.

To utilize the marker to estimate residual life or time since onset, as discussed above, the stochastic laws governing $X(t)$ as well as the relationship between $X(t)$ and the failure rate must both be considered. Jewell & Kalbfleisch [4, 5] have provided one example. They consider an additive model for (1),

$$\lambda(t|X(s) : 0 \le s \le t) = h_0(t) + \beta X(t),$$

where $h_0(t)$ is a baseline hazard and β is a regression parameter relating the nonnegative valued marker $X(t)$ to the failure rate. To complete the model, they make various assumptions about $h_0(t)$ and specify simple Markov models for $X(t)$ (*see* **Markov Processes**). From these specifications the dependence of the distribution of residual and past life on current marker values is investigated. Other approaches to jointly modeling the marker process and the failure mechanism are given in Pawitan & Self [9], Shi et al. [13], Jewell & Nielson [6], and Tsiatis et al. [15].

References

[1] Berman, S.M. (1990). A stochastic model for the distribution of HIV latency time based on T4 counts, *Biometrika* **77**, 733–741.

[2] Brookmeyer, R. & Gail, M. (1987). Biases in prevalent cohorts, *Biometrics* **43**, 739–749.

[3] Cox, D.R. (1983). A remark on censoring and surrogate response variables, *Journal of the Royal Statistical Society, Series B* **45**, 391–393.

[4] Jewell, N.P. & Kalbfleisch, J.D. (1992). Marker processes and applications to AIDS, in *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz & V. Farewell, eds. Birkhauser, Boston.

[5] Jewell, N.P. & Kalbfleisch, J.D. (1996). Marker processes in survival analysis, *Lifetime Data Analysis* **2**, 15–29.

[6] Jewell, N.P. & Nielsen, J.P. (1993). A framework for consistent prediction rules based on markers, *Biometrika* **80**, 153–164.

[7] Munõz, A., Carey, V., Taylor, J.M.G., Chmiel, J.S., Kingsley, L., Raden, M.V. & Hoover, D.R. (1992). Estimation of time since exposure for a prevalent cohort, *Statistics in Medicine* **11**, 939–952.

[8] Munõz, A., Wang, M.-C., Bass, S., Taylor, J.M.G., Kingsley, L.A., Chmiel, J.S., Polk, B.F. & the Multicenter AIDS Cohort Study Group (1989). Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men, *American Journal of Epidemiology* **130**, 530–539.

[9] Pawitan, Y. & Self, S. (1993). Modelling disease marker processes in AIDS, *Journal of the American Statistical Association* **88**, 719–726.

[10] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.

[11] Rai, S. & Matthews, D.E. (1995). The analysis of incomplete data using stochastic covariates, *Canadian Journal of Statistics* **23**, 29–42.

[12] Robins, J.M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, in *AIDS Epidemiology: Methodological Issues*, N.P. Jewell, K. Dietz & V. Farewell, eds. Birkhauser, Boston.

[13] Shi, M., Taylor, J.M.G. & Muñoz, A. (1996). Models for residual time to AIDS, *Lifetime Data Analysis* **2**, 1–14.

[14] Taylor, J.M.G., Munõz, A., Bass, S.M., Saah, A.J., Chmiel, J.S., Kingsley, L.A. & the Multicentre AIDS Cohort Study (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation, *Statistics in Medicine* **9**, 505–514.

[15] Tsiatis, A.A., DeGruttola, V. & Wulfsohn, M.S. (1995). Modelling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS, *Journal of the American Statistical Association* **90**, 27–37.

JOHN D. KALBFLEISCH

Markov Chain Monte Carlo, Recent Developments

Introduction

The preceding article on **Markov chain Monte Carlo** covers developments up to about 1997. There has recently been an explosion of interest in Markov Chain Monte Carlo for biostatistical modeling and analysis, particularly, but not exclusively in a **Bayesian** framework.

Recall that the aim is to estimate **expectations** (expected values) of θ (or functions of θ) from a probability density function $f(\theta)$. If f is nonstandard or high dimensional, MCMC algorithms allow **simulation** of (dependent) values of θ from a **Markov chain** whose stationary distribution is $f(\theta)$. In a Bayesian context, $f(\theta|x)$ is a posterior distribution with parameters θ, given data x.

The preceding entry describes the properties of principal MCMC algorithms, including Metropolis–Hastings, Gibbs, adaptive rejection sampling, and reversible-jump MCMC. It also discusses issues of burn-in, convergence, proposal distributions, and improved mixing through reparametrization. Applications considered there comprise **hierarchical models**, **missing data**, **censored data**, **measurement error**, and temporally or spatially correlated data (*see* **Geographic Epidemiology**; **Time Series**).

The aim of this update is to describe some of the new developments in MCMC methodology and their application in biostatistics.

MCMC Methods

Metropolis–Hastings and Gibbs

Metropolis–Hastings and Gibbs **algorithms** remain popular choices in biostatistical applications. Their appeal lies in the generality of their application, the ability to reduce hierarchical and high-dimensional problems to forms that are amenable to these algorithms, their theoretical properties, and their inclusion as standard tools in the more popular MCMC software (see below).

Many variations on the original Metropolis–Hastings and Gibbs samplers have been developed. The more well known of these include adaptive rejection sampling and adaptive rejection Metropolis sampling (see Gilks' original Entry on MCMC and Gilks et al. [38]) and slice sampling (see below). Other potential *black box* algorithms have been proposed; see, for example, [20].

Hybrid methods, which employ combinations of MCMC algorithms in a single analysis, are also described in **Markov Chain Monte Carlo**. These have continued to grow in popularity because of their flexibility, improved exploratory ability, theoretical validity, and appealing properties of estimation and convergence [99]. Hybrid methods embrace a diversity of constructions, including different Metropolis–Hastings and Gibbs algorithms for different components of θ [7], the insertion of a Metropolis–Hastings step with larger dispersion or probability of acceptance at every nth iteration, mode-jumping proposals [102], the insertion of a Metropolis–Hastings step after each Gibbs cycle [74], and Metropolis-within-Gibbs algorithms [20, 61, 69, 70]; see [83, pp. 319–326] for a formal definition of a hybrid method, comprehensive discussion, and examples. Importantly, building on their results and those of Tierney [99], hybrid methods can be almost automatically constructed to ensure *uniform convergence* to the target distribution.

Other *adaptive algorithms* include mode-jumping proposals [102], methods based on *tempering* (see Celeux et al. [18] and references therein), and approaches based on *regeneration* [38, 46]. See also the algorithms based on sequential Monte Carlo, described below.

Parameterization of the model has an impact on the choice and effectiveness of MCMC methods. The positive impact of reparameterization of hierarchical models is also achieved for other model formulations such as mixtures, $f(\theta) = \Sigma_{j=1}^{k} p_j f(\theta_j)$, where $\Sigma_{j=1}^{k} p_j = 1$, $0 \le p_j \le 1$. Like hierarchical models, mixtures have enjoyed increasing popularity as flexible modeling tools due to the enhanced computational ability afforded by MCMC. However, the application of standard MCMC algorithms to the usual formulation of a mixture density can lead to **identifiability** problems and the possibility of "trapping states" caused by allocation of a small number of observations to a particular component.

Robert and Titterington [90] have proposed effective reparameterizations to overcome these drawbacks. Two such alternatives are discussed by Robert and Casella [83] in the context of a mixture of **normal** densities, $\Sigma_{j=1}^{k} p_j \mathcal{N}(\mu_j, \tau_j^2)$; here, τ^2 is the inverse of the variance. The first, $\Sigma_{j=1}^{k} p_j \mathcal{N}(\mu + \tau\theta_j, \tau^2\sigma_j^2)$, with $\theta_1 = 0, \sigma_1 = 1$, is a simple expression of each component as a perturbation from a global location μ and a global scale τ. An alternative, more stable reparameterization proposed by Robert and Mengersen [84] expresses each component as a perturbation of the previous component. Thus, a two-component normal mixture is expressed as $p\mathcal{N}(\mu, \tau^2) + (1-p)\mathcal{N}(\mu + \tau\sigma, \tau^2\sigma^2)$ and the three-component analog is $p\mathcal{N}(\mu, \tau^2) + (1-p)q\mathcal{N}(\mu + \tau\sigma, \tau^2\sigma^2) + (1-p)(1-q)\mathcal{N}(\mu + \tau\sigma + \tau\sigma\varepsilon, \tau^2\sigma^2\omega^2)$. With an identifiability constraint $\sigma_1 \leq 1, \ldots, \sigma_{k-1} \leq 1$, an improper **prior** for (μ, τ) and a **uniform distribution** on the σ_i's can be adopted and standard MCMC algorithms can be employed.

The desired *acceptance rate* of a Metropolis–Hastings algorithm has also been a matter of recent research. Optimal rates for random walk algorithms have been carefully investigated by Roberts et al. [86] and corresponding guidelines have been suggested. As described and illustrated by Robert and Casella [83, pp. 252–254], high acceptance rates are desirable if the proposal density g approximates the target f such that f/g is bounded for uniform ergodicity. However, low acceptance rates are preferable if a random walk proposal is adopted. These authors also propose the use of the rejected values in a Metropolis–Hastings algorithm through **Rao–Blackwellization** and give references to other acceleration methods.

Active research continues on the *theoretical properties* of Gibbs and Metropolis–Hastings algorithms. In a general state-space context, Tierney [100] identifies necessary and sufficient conditions on the Metropolis–Hastings proposal kernel and the acceptance probability function for the resulting transition kernel and invariant distribution to satisfy the detailed balance conditions. References to recent results on MCMC convergence are given below.

Model Choice

Model choice via MCMC is now a standard practice, as indicated by the recent reviews of the variety of available methods (*see* **Bayesian Methods for Model Comparison**). See, for example, the papers by Han and Carlin [45], Brooks [10] and Dellaportas et al. [28] and the summaries by Carlin and Louis [15, Chapter 5] and Congdon [22, Chapter 10].

As discussed by these authors, methods based on **Bayes factors** include variations on marginal density estimation (see also [21]) and sampling over the model space. Weakliem [104] gives a broad discussion and critique of the Bayes information criterion (BIC) $= \log P(y|\hat{\theta}, M) - p/2 \log n$ for a given model M and sample size n. Bayesian ***P* values**, penalized likelihood methods, and predictive model selection approaches are also available.

Green's [39] *reversible jump Markov Chain Monte Carlo* (RJMCMC) algorithm has recently become a key tool for model choice. Sampling can now be extended to different parameter spaces, such as the dimensions of a model, components of a mixture, or subsets of variables in a regression. This approach has inspired other methodological developments in model selection and a diversity of applications. As conceded by Robert and Casella ([83], pp. 259–264), however, RJMCMC can be difficult to implement because of the requirement for reversible moves and a (differentiable) dimension-matching transform. Moreover, inefficient moves between dimensions can require complicated tuning steps. An alternative to RJMCMC is described by Stephens [97]. Under this method, the parameters of interest are considered as a marked point process.

Instead of choosing a single model based on the above methods, an increasingly common practice is *model averaging*. This is the practice of combining expected values obtained from different models (perhaps describing different dimensions or different combinations of variables) weighted by their corresponding posterior probabilities. Of course, adoption of this approach depends on the aim of the analysis and achieving a balance between improved estimation and easy interpretation.

Slice Sampling

A technique that enables the Gibbs sampler to be used for almost any distribution is the slice sampler. Introduced by Wakefield et al. [103] as a "ratio-of-uniforms" method for generating random variables, developed by Neal [73] as a method for "slicing" distributions, and described by Tierney and Mira

[101] in the context of adaptive MCMC models, the slice sampler enjoys increasing popularity and is part of most modern MCMC texts; see, for example, [14, 83]. The latter authors describe the slice sampler as follows.

If $f(\theta)$ can be written as a product $\Pi_{i=1}^{k} f_i(\theta)$, where the f_i's are positive functions (not necessarily densities), then f can be expressed as $\Pi_{i=1}^{k} \mathcal{I}_{0 \le \omega_i \le f_i(\theta)}$, where $\mathcal{I}$ is the indicator function.

Thus at the tth iteration, $\theta^{(t)}$ is simulated by generating k uniform random variables $\omega_1^{(t)} \sim \mathcal{U}[0, f_1(\theta^{(t-1)})], \ldots, \omega_k^{(t)} \sim \mathcal{U}[0, f_k(\theta^{(t-1)})]$ and taking $\theta^{(t)} \sim U(A^{(t)})$, with $A^{(t)} = \{y; f_i(y) \ge \omega_i^{(t)}, i = 1, \ldots, k\}$.

It can be shown that this chain converges geometrically when f is bounded and converges uniformly when $k = 1$. An upper bound for the rate of convergence has also been established. Moreover, given an independent Metropolis–Hastings algorithm, it is always possible to construct a slice sampler that has a smaller asymptotic variance and smaller second-largest **eigenvalue**, thus ensuring faster convergence to the target distribution. See [68, 88, 101] for a discussion of these and other theoretical properties.

Perfect Simulation

Another development in MCMC that has created its own domain of research is perfect simulation, also known as *exact sampling*. As described in the original paper by Propp and Wilson [79] and subsequently by Kendall [50], the aim of perfect simulation is to sample directly from the stationary distribution $f(\theta)$.

Although this appears to be exactly what MCMC is aiming to avoid, there are several reasons for pursuing the idea. First, independent samples drawn directly from $f(\theta)$ may be preferable to samples obtained from MCMC algorithms, depending on the degree of dependence in the latter and the comparative computational time and complexity. Second, a single sample drawn directly from $f(\theta)$ can be used as a starting point for standard MCMC algorithms. This avoids the well-known problem of *burn-in*, in which the initial value of the chain may induce long-term bias.

For a finite state-space $\mathcal{X}$ of size k, Propp and Wilson [79] proposed an exact sampling algorithm called *coupling from the past* (CFTP). Here, k chains corresponding to all possible starting points in $\mathcal{X}$ are started at time t and run in parallel *back* in time, often in a *coupled manner*, until all the chains *coalesce* (take the same value) at time 0 or earlier. The realizations of the chains at time 0 then form a single $\theta^{(0)}$ from the required distribution.

If the chains have not coalesced by time 0, the chains are run again from time $2t$ and this is continued until the desired result is achieved. It can be shown that coalescence under CFTP will indeed occur in a finite number of backward iterations. In practice, however, the computation time can be unacceptably long. Alternative algorithms have been developed to improve this and other aspects of the original CFTP idea. For example, Fill [35] proposed an *interruptible* algorithm for perfect simulation, in which the chains can be stopped before reaching time 0 but maintain the properties of the CFTP algorithm. As a second example, if a *monotonicity* constraint can be constructed, so that there is stochastically a maximum state x_1 and a minimum state x_0 in $\mathcal{X}$, then CFTP reduces to running only two chains from x_0 and x_1 until they coalesce at time 0, since all the intermediary paths will be between these two extreme cases. As a second example, Kendall and Møller [51] describe extensions to the original CFTP approach, focusing, in particular, on perfect simulation methods based on dominating processes on ordered spaces.

As noted by Robert and Casella [83], the extension of perfect simulation ideas from finite state spaces to the statistical context is an area of current active research. Potential improvements in the speed and control of convergence under this method are mitigated by the concern that the time that is typically required to simulate one realization of $\theta^{(0)}$ under CFTP is much greater (by orders of magnitude) than the computation time of a $\theta^{(t)}$ from a standard MCMC algorithm. Moreover, algorithms for continuous spaces are more difficult to construct. Murdoch and Green [71] originally explored standard statistical examples in a continuous setting. Casella et al. [16] proposed a perfect simulation method for mixture modeling using slice sampling. Perfect slice samplers are also discussed by Mira et al. [66, 67].

Delayed Rejection

Consider a standard Metropolis–Hastings algorithm with target density f, such that if a proposed value y is rejected, the chain remains in the current state x. This preserves the stationary distribution but induces **autocorrelation** in the realized chain. Tierney and

Mira [101] proposed an alternative based on *delayed rejection* or *splitting rejection*. When a proposed value is rejected, a second proposal is made using a different distribution possibly dependent on the previously rejected values and is accepted with probability modified to account for the previous rejection. This is continued until a stopping rule is met, such as a maximum number of attempts or until acceptance is achieved. Thus, a first proposal x is generated from $q_1(x, y)$ and accepted with probability $\min\{1, [f(y)q_1(y, x)]/[f(x)q_1(x, y)]\}$. If y is rejected, a new value z is proposed from $q_2(x, y, z)$ and accepted with probability $\min\{1, [f(z)q_1(z, y) q_2(z, y, x)(1 - \alpha_1(z, y))]/[f(x)q_1(x, y)q_2(x, y, z) (1 - \alpha_1(x, y))]\}$.

Green and Mira [40] proposed a generalization of this idea, extended the method to reversible jump algorithms, and presented a comparison of the performance of their proposed delayed rejection algorithm with other samplers. Performance was defined in terms of the efficiency of estimating the desired expectation on the state space and computed as the product of the running time needed to obtain a fixed number of sweeps and the integrated autocorrelation time.

Sequential Monte Carlo

The term *sequential Monte Carlo* embraces a variety of algorithms, including Langevin or diffusion algorithms, population Monte Carlo or iterated **importance sampling**, and particle filters.

Langevin algorithms, also known as *diffusion algorithms*, were proposed by Grenander and Miller [42] and Phillips and Smith [78] and arise from the discretized solution of a stochastic differential equation, also known as a diffusion equation. Thus, $x^{(t+1)} = x^{(t)} + 0.5\sigma^2 \Delta \log f(x^{(t)}) + \sigma\varepsilon_t$ where $\varepsilon_t \sim \mathcal{N}_p(0, I_p)$ and σ^2 corresponds to the discretization size. Following Besag [6], $x^{(t+1)}$ is then accepted according to a regular Metropolis step.

The theoretical properties of this approach have been investigated by Roberts and Tweedie [91] and Stramer and Tweedie [98], among others. Robert and Casella [83, pp. 264–266] describe and illustrate various diffusion algorithms. Extensions include switching diffusion models and their variations [59] and variations on proposal distributions [95].

Population Monte Carlo, or *iterated importance sampling*, involves simultaneous generation of a vector of random variables $(\theta_1^{(t)}, \ldots, \theta_M^{(t)})$ at each iteration of a Monte Carlo algorithm. Such systems of particles were shown by Mengersen and Robert [64] to be capable of producing i.i.d. samples for a given target distribution, a feature that is only achieved with difficulty via perfect sampling in regular MCMC algorithms.

The increased interest in population Monte Carlo methods is evidenced by the different algorithms proposed by Haario, Sacksman, and Tamminen [44], Mengersen and Robert [64] and others. The latter authors, for example, proposed a "pinball" Metropolis–Hastings algorithm that features "bouncing" via delayed rejection and "repulsion" via an updating mechanism based on a self-avoiding random walk, that is, a standard random walk with corrections to avoid the immediate vicinity of other particles. Paradoxically, the corresponding importance resampling particle system based on the same proposal enjoys poor properties like high degeneracy and low mixing.

Recent interest has also focused on MCMC methods for *particle filters*, which are usually implemented in sequential settings or for processing and analysis of large datasets. A particle filter describes a dynamic state-space model of a process with an underlying state of interest that evolves over time. The posterior distribution of the state is approximated by a set of weighted particles, with the weight of a particle inversely proportional to its probability mass. Numerous algorithms for updating the particles and their weights over time have been proposed. Most of these enjoy rigorous convergence properties [27], and under certain conditions can claim a **Central Limit Theorem** [29].

Most particle filter methods include a resampling step in which particles are replicated in order to convert an unequally weighted set of particles into an equally weighted set with the same distribution. However, this does not prevent the widely acknowledged problem of a continued loss of accuracy in approximation over time, which is reflected in increasing clustering of the particles in a single region of the state-space. This may be resolved by the inclusion of MCMC methods into the particle filter algorithm, as described by Gilks and Berzuini [37] and the references therein. Because the initial set of weighted particles is approximately from the target density, there is no need for a burn-in period

in the MCMC algorithm. Fearnhead [33] proposed the use of MCMC based on **sufficient statistics** as summaries of the trajectory of each particle, thus eliminating the need to store the whole particle history and consequently reducing memory requirements and computational complexity.

Detailed reviews and applications of these methods are given by Liu and Chen [62], Fearnhead [33], and Doucet et al. [32].

MCMC Convergence

Assessment of the convergence of MCMC algorithms remains a major theoretical and practical issue. In fact, it is becoming increasingly important with the development of more complicated models, the construction of new hybrid and dimension-changing algorithms, and the wider uptake of MCMC as a standard statistical tool.

There is now a substantial body of theoretical knowledge about the existence and form of convergence for certain types of target distributions and different MCMC algorithms. These are currently used to build practical bounds on convergence in particular cases, construct Metropolis–Hastings and hybrid algorithms so that geometric or subgeometric (e.g. polynomial, logarithmic, subexponential) rates of convergence are assured, make statements about new algorithms such as Langevin diffusions, and so on; see, for example, [11, 36, 49, 65, 87, 89, 92, 93, 98].

Despite these advances, convergence assessment in most practical setups is based on a (subjective) selection of a subset of the wide array of available empirical diagnostic methods. These methods guide the determination of the length of burn-in, the number of iterations after burn-in, the use of parallel chains and, if necessary, the batch size (also known as the thinning interval or subsampling). Although the latter is inefficient with respect to estimation of expected values, it may be beneficial for other reasons; see [63]. Similarly, debate surrounds the benefits and drawbacks of parallel chains, that is, simulation of independent chains from different initial values, and is discussed by Robert and Casella [83, pp. 365, 366] and references therein.

It is convenient to guide the choice of diagnostic by identifying specific convergence goals. Following Mengersen and Robert [63], these might include convergence of the chain $\theta^{(t)}$ to the stationary distribution f; convergence of the empirical average, $\Sigma_{t=1}^{T} g(\theta^{(t)})/T$ to $E_f(g(\theta))$ for an arbitrary function g; application of the Central Limit Theorem; and generation of an i.i.d. sample $(\theta_1^{(t)}, \ldots, \theta_1^{(t)})$. This identification also assists in the resolution of the often conflicting results from different diagnostics.

In a similar manner, convergence diagnostics can be broadly categorized as methods based on graphical assessment, estimated distance between the empirical and target densities, renewal and regeneration results, variance approximations and comparisons, and discretization of the Markov chain.

Not all of these methods are applicable to convergence assessment of dimension-changing algorithms such as RJMCMC. This special case is discussed by Brooks and Giudici [12].

Detailed reviews of these various approaches have been compiled by Best et al. [8], Cowles and Carlin [26], Mengersen and Robert [63], Brooks and Roberts [13], and Robert and Casella [83, pp. 365–413]. More recently, "movies" for the visualization of MCMC output have been proposed by Lazar and Kadane [57].

Applications of MCMC in Biostatistics

The discussion below highlights a small selection of biostatistics-related areas in which MCMC has become a familiar computational tool.

Classification and regression trees (CART), originally proposed by Breiman et al. [9], are nonparametric description or predictions of a response in the form of binary splits of selected **explanatory variables** (*see* **Tree-structured Statistical Methods**). A Bayesian alternative was independently proposed by Denison et al. [31] based on RJMCMC estimation of the probability distribution over the space of possible trees. Other approaches have also been proposed.

Latent variable models were introduced in **Markov Chain Monte Carlo** as mechanisms for describing complex physical and conceptual systems. Conditional on the latent, or unobserved, variables, hierarchical models can be constructed and corresponding posterior (conditional) distributions can be more simply described. These models can also be considered as *missing data models*.

As an illustration of a popular latent variable approach, Robert [81] and coauthors have described the analysis of mixture models $\Sigma_{j=1}^{k} p_j f(x|\theta_j)$, discussed above. Here, the analysis is considerably

simplified by the introduction of latent variables $z_i, i = 1, \ldots, n$ that indicate the component $j, j \in \{1, \ldots, k\}$, to which each observation x_i belongs. The MCMC algorithm involves estimating z_i given the component parameters, then estimating the component parameters and weights based on the allocation of the z's. The number of components, k, can also be a random variable by extending the MCMC algorithm to have a dimension-jumping step. The exposition of RJMCMC in this context by Richardson and Green [80] has generated a large number of applications. Fernández and Green [34] describe the analysis of spatially correlated **Poisson** data by a Poisson mixture model in which the weights of the mixture vary across locations and the number of components is unknown. RJMCMC is also employed in the context of **nonparametric regression** by Denison et al. [31], Perron and Mengersen [77], and Lindstrom [60].

More generally, the enhanced computational ability afforded by MCMC has led to increased interest in **hidden Markov models** (HMMs); see [85] for extended discussion of these methods.

Time series models continue to attract attention with respect to the development of corresponding MCMC algorithms, the construction of models to represent complex systems, and their application to diverse problems. A comprehensive discussion of developments in this area is given by West and Harrison [105] in the context of *dynamic linear modeling*. See also the recent developments in particle filters, dynamic Bayesian networks, and perfect sampling. MCMC methods for temporal modeling of epidemics and **infectious diseases** have also been widely considered; see, for example, [76].

The analysis of **factorial experiments** using MCMC methods has been investigated by Nobile and Green [75] using mixture models.

MCMC methods for analyzing *spatially correlated* data remain an area of active interest. Knorr–Held and Besag [53], Knorr–Held and Raser [55], Lawson [56], Knorr–Held and Best [54], and Green and Richardson [41] describe Markov random fields [5], RJMCMC and hidden Markov models in the context of disease **mapping**. Various methods for analyzing spatially correlated Poisson data have been proposed by Castelloe [17], Wolpert et al. [106], and Fernandez and Green [34]. As discussed by Anselin and Griffith [2], spatial effects can also describe measurement errors, heteroscedasticity, and unobserved covariates.

Meta-analysis models and corresponding MCMC analysis is now a standard biostatistical tool. Multivariate approaches have been described by Nam et al. [72] and references therein. Wolpert and Mengersen [107] describe adjustments to the **likelihood** in a meta-analysis context that can take account of **misclassification**, **bias**, and other features of individual studies.

Approaches to the design and analysis of **clinical trials** using MCMC are described by Carlin and Louis [15]. Attention is also paid to the analysis of **survival** *models* by these authors and subsequently by Albert and Chib [1].

Bioinformatics is an emerging field that was once considered to be the part of computational biology that explicitly dealt with the development of the increasing number of large **databases**, including methods for data retrieval and analyses, and algorithms for sequence similarity searches, structural predictions, functional predictions and comparisons, and so forth. Very recently, the field has been rapidly evolving, not only because of the impact of the various genome projects, but also because of the development of experimental technologies, particularly microarrays (*see* **Genetic Markers**). Currently bioinformatics is being increasingly widely viewed as a more fundamental discipline that also encompasses mathematics, statistics, physics, and chemistry. A detailed discussion of the current state of bioinformatics is provided by Kenehisa and Bork [52]. MCMC and RJMCMC provide a convenient method of analysis of the complex models and datasets encountered in this field. Very recent applications include evolutionary analysis [24], the analysis of microarray experiments [58] and genome analysis [94]. Baldi and Brunak [3] include a chapter on using MCMC as a "machine learning algorithm" for bioinformatics.

As with any numerical method, sensitivity to the model description and the adopted MCMC algorithm should be an integral part of the analysis. Huelsenbeck et al. [47] discuss this in the context of estimating probabilities of phylogenetic trees.

Software

Gibbs and Metropolis–Hastings algorithms are now standard tools in some statistical packages (*see* **Software, Biostatistical**). See, for example, the SAS procedure for estimating missing data and algorithms available in **S-Plus** and **R**.

BUGS and WinBUGS [96] remain popular specialist software for MCMC analysis. The freely available suite of algorithms has improved and expanded over a number of versions. Congdon [22] almost exclusively uses BUGS to illustrate many applications of Bayesian modeling and provides detailed discussion of its implementation and interpretation.

Other established software packages include a Bayes Linear Programming Package (B/D), Bayesian Knowledge Discover (BKD), a general data analysis tool B-Course, and JavaBayes for Bayesian networks.

CODA [8] remains the most accessible source of generic convergence diagnostics.

Application-specific software is also widely available. For example, S-Plus code for Bayesian model selection is available from Adrian Raftery's website and for particular biostatistics applications from the MD Anderson Cancer Centre. Individual papers also often advertise software.

Books

The popularity of MCMC and its place as a standard statistical tool is evidenced by the recent publication of books devoted to or strongly focused on these methods.

Robert and Casella [83] provide excellent discussion and illustrations of a variety of established and new MCMC approaches and their connections with other **Monte Carlo methods**. They cover theoretical and practical aspects of the methods and give detailed algorithms.

The text by Carlin and Louis [15] introduces Bayes and **empirical Bayes** methods and their applications in a wide variety of settings. Their descriptions are complemented by worked examples using BUGS. Recent developments in MCMC, including RJMCMC, slice sampling, structured MCMC, the computation of MCMC standard errors, and MCMC convergence are also discussed. Case studies include the analysis of longitudinal **AIDS** data, analysis of clinical trials and spatio-temporal modeling of lung cancer rates.

The text by Congdon [22] has become popular among both researchers and practitioners. It gives a very readable account of the general Bayesian method, MCMC algorithms, standard distributions, and a variety of specialist models. The text also details the implementation of the approaches in the software package BUGS. Congdon [23] extends this discussion into more applied Bayesian modeling.

More focused books that describe MCMC algorithms include West and Harrison [105], Robert [82] and Cowell et al. [25], and Chen et al. [19], Ibrahim et al. [48], Banerjee et al. [4], Denison et al. [30], and Gustafson [43].

Published conference proceedings also contain many expositions of recent MCMC advances. See, for example, the "Bayesian Statistics" series arising from the four-yearly Valencia meetings, published by Oxford University Press.

Acknowledgment

Petra Graham assisted with the compiling of references.

References

[1] Albert, J.H. & Chib, S. (2001). Sequential ordinal modeling with applications to survival data, *Biometrics* **57**(3), 829–836.

[2] Anselin, L. & Griffith, D. (1988). Do spatial effects really matter in regression analysis, *Papers of the Regional Science Association* **65**, 11–34.

[3] Baldi, P. & Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, Massachusetts.

[4] Banerjee, S., Gelfand, A.E. & Carlin, B.P. (2003) *Hierarchical Modelling and Analysis for Spatial Data*. Chapman & Hall/CRC, London.

[5] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society Series B – Statistical Methodology* **36**, 192–326.

[6] Besag, J.E. (1994). Discussion of "Markov chains for exploring posterior distributions", *Annals of Statistics* **22**, 1734–1741.

[7] Besag, J., Green, E., Higdon, D. & Mengersen, K.L. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science* **10**, 3–66.

[8] Best, N.G., Cowles, M.K. & Vines, K. (1995). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30, Technical Report, MRC Biostatistics Unit, University of Cambridge. Cambridge, UK.

[9] Breiman, L., Friedman, J.H., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, California.

[10] Brooks, S.P. (2001). On Bayesian analyses and finite mixtures for proportions, *Statistics and Computing* **11**(2), 179–190.

[11] Brooks, S.P., Dellaportas, P. & Roberts, G.O. (1997). A total variation method for diagnosing convergence of MCMC algorithms, *Journal of Computational and Graphical Statistics* **6**, 251–265.

[12] Brooks, S.P. & Giudici, P. (1999). Convergence assessment for reversible jump MCMC simulations in *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, New York, 733–742.

[13] Brooks, S.P. & Roberts, G.O. (1999). On quantile estimation and Markov chain Monte Carlo convergence, *Biometrika* **86**, 710–717.

[14] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

[15] Carlin, B. & Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd Ed. Chapman & Hall, London.

[16] Casella, G., Mengersen, K.L., Robert, C.P. & Titterington, D.M. (2002). Perfect samplers for mixtures of distributions, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **64**, 777–790.

[17] Castelloe, J.M. (1999). Reversible jump Markov chain Monte Carlo analysis of spatial Poisson cluster processes, ASA Proceedings of the Section on Bayesian Statistical Science, pp. 102–107.

[18] Celeux, G., Hurn, M. & Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions, *Journal of the American Statistical Association* **95**, 957–970.

[19] Chen, M.H., Shao, Q.M. & Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

[20] Chen, M.H. & Schmeiser, B.W. (1998). Towards black-box sampling, *Journal of Computational and Graphical Statistics* **7**, 1–22.

[21] Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association* **96**(453), 270–281.

[22] Congdon, P. (2001). *Bayesian Statistical Modelling*. Wiley, London.

[23] Congdon, P. (2003). *Applied Bayesian Modelling*. Wiley, New York.

[24] Corander, J., Waldmann, P. & Sillanpaa, M.J. (2003). Bayesian analysis of genetic differentiation between populations, *Genetics* **163**(1), 367–374.

[25] Cowell, R.G., Dawid, A.P., Lauritzen, S.L. & Spiegelhalter, D.J. (2003). *Probabilistic Networks and Expert Systems. Springer-Verlag Series in Information Science and Statistics*. Springer-Verlag, New York.

[26] Cowles, M.K. & Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative study, *Journal of the American Statistical Association* **91**, 883–904.

[27] Crisan, D. & Doucet, A. (2000). Convergence of sequential Monte Carlo methods, Technical Report CUED/F-INFENG/TR381, Department of Engineering, University of Cambridge.

[28] Dellaportas, P., Forster, J.J. & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC, *Statistics and Computing* **12**(1), 27–36.

[29] Del Moral, P. & Guionnet, A. (1999). Central limit theorem for nonlinear filtering and interacting particle systems, *Annals of Applied Probability* **9**(2), 275–297.

[30] Denison, D.G.T., Holmes, C.C., Mallick, B.K. & Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, New York.

[31] Denison, D.G.T., Mallick, B.K. & Smith, A.F.M. (1998). Automatic Bayesian curve fitting, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **60**, 333–350.

[32] Doucet, A., de Freitas, N. & Gordon, N., eds. (2001). *Sequential Monte Carlo Methods in Practice* Springer-Verlag, New York.

[33] Fearnhead, P. (2002). MCMC, sufficient statistics and particle filter. P. Fearnhead, *Computational and Graphical Statistics* **11**, 848–862.

[34] Fernández, C. & Green, P.J. (2002). Modeling spatially correlated data via mixtures: A Bayesian approach, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **64**(4), 805–826.

[35] Fill, J.A. (1998). An interruptible algorithm for exact sampling via Markov chains, *Annals of Applied Probability* **8**, 131–162.

[36] Fort, G. & Moulines, E. (2000). V-subgeometric ergodicity for a Hastings-Metropolis algorithm, *Statistics and Probability Letters* **49**, 401–410.

[37] Gilks, W.R. & Berzuini, C. (2001). Following a moving target – Monte Carlo inference for dynamic Bayesian models, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **63**, 127–146.

[38] Gilks, W.R., Roberts, G.O. & Sahu, S.K. (1998). Adaptive Markov chain Monte Carlo, *Journal of the American Statistical Association* **93**, 1045–1054.

[39] Green, P.J. (1995). Reversible jump MCMC computation and Bayesian model determination, *Biometrika* **82**(4), 711–732.

[40] Green, P.J. & Mira, A. (2001). Delayed rejection in reversible jump Metropolis-Hastings, *Biometrika* **88**(4), 1035–1053.

[41] Green, P.J. & Richardson, S. (2002). Hidden Markov models and disease mapping, *Journal of the American Statistical Association* **97**(460), 1055–1070.

[42] Grenander, U. & Miller, M. (1994). Representations of knowledge in complex systems (with discussion), *Journal of the Royal Statistical Society Series B – Statistical Methodology* **56**, 549–603.

[43] Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall/ CRC Press, Boca Raton.

[44] Haario, H., Saksman, E. & Tamminen, J. (1999). Adaptive Proposal Distribution for Random Walk Metropolis Algorithm, *Computational Statistics* **14**, 375–395.

[45] Han, C. & Carlin, B.P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: a comparative review, *Journal of the American Statistical Association* **96**(455), 1122–1132.

[46] Hobert, J.P., Jones, G.L., Presnell, B. & Rosenthal, J.S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo, *Biometrika* **89**(4), 731–743.

[47] Huelsenbeck, J.P., Larget, B., Miller, R.E. & Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny, *Systematic Biology* **51**(5), 673–688.

[48] Ibrahim, J.G., Chen, M.H. & Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.

[49] Jarner, S. & Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms, *Stochastic Processes and their Applications* **12**, 224–247.

[50] Kendall, W. (1998). Perfect simulation for the area-interaction point process, in *Probability Towards 2000*, C.C. Heyde, & L. Accardi, eds. Springer-Verlag, New York, 218–234.

[51] Kendall, W.S. & Møller, J. (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes, *Advances in Applied Probability* **32**(3), 844–865.

[52] Kenehisa, M. & Bork, P. (2003). Bioinformatics in the post-sequence era, *Nature Genetics Supplement* **33**, 305–310.

[53] Knorr-Held, L. & Besag, J. (1998). Modelling risk from a disease in time and space, *Statistics in Medicine* **17**, 2045–2060.

[54] Knorr-Held, L. & Best, N.G. (2001). A shared component model for detecting joint and selective clustering of two diseases (Pkg: p49–99), *Journal of the Royal Statistical Society, Series A, General* **164**(1), 73–85.

[55] Knorr-Held, L. & Raser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps, *Biometrics* **56**(1), 13–21.

[56] Lawson, A.B. (2000). Cluster modelling of disease incidence via RJMCMC methods: a comparative evaluation, *Statistics in Medicine* **19**, 2361–2375.

[57] Lazar, N.A. & Kadane, J.B. (2002). Movies for the visualization of MCMC output, *Journal of Computational and Graphical Statistics* **11**(4), 863–874.

[58] Lee, K.E., Sha, N.J., Dougherty, E.R., Vannucci, M. & Mallick, B.K. (2003). Gene selection: a Bayesian variable selection approach, *Bioinformatics* **19**(1), 90–97.

[59] Liechty, J.C. & Roberts, G.O. (2001). Markov chain Monte Carlo methods for switching diffusion models, *Biometrika* **88**(2), 299–315.

[60] Lindstrom, M.J. (2002). Bayesian estimation of free-knot splines using reversible jumps, *Computational Statistics and Data Analysis* **41**(2), 255–269.

[61] Liu, J.S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling, *Statistics and Computing* **6**, 113–119.

[62] Liu, J.S. & Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems, *Journal of the American Statistical Association* **93**, 1032–1044.

[63] Mengersen, K.L. & Robert, C.P. (1999). MCMC convergence diagnostics: a review, in *Bayesian Statistics VI*, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 415–440.

[64] Mengersen, K. & Robert, C.P. (2003). Population-based Markov chain Monte Carlo: the pinball sampler, in *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith & M. West, eds. Oxford University Press, Oxford.

[65] Mira, A. (2001). Ordering and improving the performance of Monte Carlo Markov chains, *Statistical Science* **16**(4), 340–350.

[66] Mira, A., Møller, J. & Roberts, G.O. (2001). Perfect slice samplers, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **63**, 593–606.

[67] Mira, A., Møller, J. & Roberts, G.O. (2002). Correction to "Perfect slice samplers" (2001v63 p593–606), *Journal of the Royal Statistical Society, Series B, Methodological* **64**(3), 581–581.

[68] Mira, A. & Tierney, L. (2002). Efficiency and convergence properties of slice samplers, *Scandinavian Journal of Statistics* **29**(1), 1–12.

[69] Muller, P. (1991). A Generic Approach To Posterior Integration And Gibbs Sampling, Technical Report, **ad no** 91-09, Purdue University, West Lafayette.

[70] Muller, P. (1993). Alternatives to the Gibbs Sampling Scheme, Technical Report, Institute of Statistics and Decision Sciences, Duke University.

[71] Murdoch, D.J. & Green, P.J. (1998). Exact sampling for a continuous state, *Scandinavian Journal of Statistics* **25**(3), 483–502.

[72] Nam, I.-S., Mengersen, K., & Garthwaite, P. (2003). Multivariate meta-analysis, *Statistics in Medicine* **22**, pp. 2309–2333.

[73] Neal, R.M. (2003). Slice sampling (with discussion), *Annals of Statistics*, **31**, 705–767.

[74] Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model, *Statistics and Computing* **8**, 229–242.

[75] Nobile, A. & Green, P.J. (2000). Bayesian analysis of factorial experiments by mixture modelling, *Biometrika* **87**(1), 15–35.

[76] O'Neill, P.D., Balding, D.J., Becker, N.G., Eerola, M. & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods, *Applied Statistics* **49**(4), 517–542.

[77] Perron, F. & Mengersen, K. (2001). Bayesian nonparametric modeling using mixtures of triangular distributions, *Biometrics* **57**(2), 518–528.

[78] Phillips, D.B. & Smith, A.F.M. (1996). Bayesian model comparison via jump diffusions, in *Markov chain Monte Carlo in Practice*, W.R. Gilks, S.T. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, 215–240.

[79] Propp, J.G. & Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics, *Random Structures and Algorithms* **9**, 223–252.

[80] Richardson, S. & Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society Series B – Statistical Methodology* **59**, 731–792.

[81] Robert, C.P. (1996). Inference in mixture models, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, 441–464.

[82] Robert, C.P. ed. (1998). *Discretization and MCMC convergence assessment*. Springer-Verlag, New York.

[83] Robert, C.P. & Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

[84] Robert, C.P. & Mengersen, K.L. (1999). Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler, *Computational Statistics and Data Analysis* **29**, 325–343.

[85] Robert, C.P., Rydén, T. & Titterington, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **62**(1), 57–75.

[86] Roberts, G.O., Gelman, A. & Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms, *Annals of Applied Probability* **7**, 110–120.

[87] Roberts, G.O. & Rosenthal, J.S. (1998). Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion), *Canadian Journal of Statistics* **26**, 5–32.

[88] Roberts, G.O. & Rosenthal, J.S. (1999). Convergence of slice sampler Markov chains, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **61**, 643–660.

[89] Roberts, G.O. & Sahu, S.K. (1997). Updating schemes, covariance structure, blocking and parametrisation for the Gibbs sampler, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **59**, 291–318.

[90] Robert, C.P. & Titterington, M. (1998). Reparametrisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation, *Statistics and Computing* **8**(2), 145–158.

[91] Roberts, G.O. & Tweedie, R.L. (1996). Exponential Convergence of Langevin diffusions and their discrete approximations, *Bernoulli* **2**, 341–364.

[92] Roberts, G.O. & Tweedie, R.L. (1996). Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithms, *Biometrika* **83**, 95–110.

[93] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association* **90**, 558–566.

[94] Salmenkivi, M., Kere, J. & Mannila, H. (2002). Genome segmentation using piecewise constant intensity models and reversible jump MCMC, *Bioinformatics* **18**(Suppl. S), S211–S218.

[95] Skare, O., Benth, F.E. & Frigessi, A. (2000). Smoothed Langevin proposals in Metropolis-Hastings algorithms, *Statistics and Probability Letters* **49**(4), 345–354.

[96] Spiegelhalter, D.J., Thomas, A., Best, N.G. & Lunn, D. (2002). *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge. Available from `www.mrc-bsu.cam.ac.uk/bugs`.

[97] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods, *Annals of Statistics* **28**(1), 40–74.

[98] Stramer, O. & Tweedie, R.L. (1998). Langevin-type models II: self-targeting candidates for MCMC algorithms, *Methodology and Computing in Applied Probability* **1**, 307–328.

[99] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics* **22**, 1701–1786.

[100] Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces, *Annals of Applied Probability* **8**(1), 1–9.

[101] Tierney, L. & Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference, *Statistics in Medicine* **18**, 2507–2515.

[102] Tjelmeland, H. & Hegstad, B.K. (2001). Mode jumping proposals in MCMC, *Scandinavian Journal of Statistics* **28**(1), 205–223.

[103] Wakefield, J.C., Gelfand, A.E. & Smith, A.F.M. (1991). Efficient generation of random variates via the ratio-of-uniform method, *Statistics and Computing* **1**, 129–133.

[104] Weakliem, D.L. (1999). A critique of the Bayesian information criterion for model selection (with discussion), *Sociological Methods and Research* **27**, 359–443.

[105] West, M. & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd Ed.Springer-Verlag, New York.

[106] Wolpert, R., Best, N., Ickstadt, K. & Briggs, D. (2000). Combining models of health and exposure data: the SAVIAH study, in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best & D.J. Briggs, eds. Oxford University Press, Oxford, 393–414.

[107] Wolpert, R. & Mengersen, K. (2004). Adjusted likelihoods for synthesising empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke, *Statistical Science*. To appear.

Further Reading

Besag, J. & Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion), *Journal of the Royal Statistical Society Series B – Statistical Methodology* **61**, 691–746.

K. MENGERSEN

Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a powerful technique for performing integration by **simulation**. In recent years, MCMC has revolutionized the application of **Bayesian** statistics. Many high-dimensional, complex models which were formerly intractable can now be handled routinely. MCMC has also been used in specialized non-Bayesian problems. Introductory material on MCMC methods and biostatistical applications can be found in Gilks et al. [20] and Gelman & Rubin [13].

Suppose that we wish to evaluate the expected value (**expectation**) of some function $g(\boldsymbol{\theta})$ over a probability density function $f(\boldsymbol{\theta}) : \mathrm{E}_f[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta}) f(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$. If we could draw samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}$ independently from $f(\boldsymbol{\theta})$, then we could estimate

$$\hat{\mathrm{E}}_f[g(\boldsymbol{\theta})] = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{\theta}^{(i)}). \tag{1}$$

This technique is called **Monte Carlo** *integration*. We have $\mathrm{var}\{\hat{\mathrm{E}}_f[g(\boldsymbol{\theta})]\} = \mathrm{var}_f[g(\boldsymbol{\theta})]/n$, so the estimate $\hat{\mathrm{E}}_f[g(\boldsymbol{\theta})]$ can be made as accurate as desired by increasing the sample size n. In Bayesian applications, our density $f(\boldsymbol{\theta})$ is a posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$, where $\boldsymbol{\theta}$ is a collection of model unknowns (parameters and missing data), and $\mathbf{x}$ denotes observed data. The function $g(\boldsymbol{\theta})$ might be the kth element of the vector $\boldsymbol{\theta}$, for example, in which case $\mathrm{E}_f[g(\boldsymbol{\theta})]$ would be the posterior expectation of θ_k. Other forms for $g(\cdot)$ could be used to evaluate posterior **variances**, **correlations**, **quantiles**, etc. Note that the accuracy of $\hat{\mathrm{E}}_f[g(\boldsymbol{\theta})]$ is not limited by the amount of data in $\mathbf{x}$.

Typically in Bayesian applications, $\boldsymbol{\theta}$ is high-dimensional and $f(\boldsymbol{\theta}|\mathbf{x})$ has a complicated, nonstandard form. Sampling independently from $f(\boldsymbol{\theta}|\mathbf{x})$ is generally not possible. Therefore we could try to devise sampling schemes which generate *dependent* samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}$, but for which (1) is still a **consistent** estimator of $\mathrm{E}_f[g(\boldsymbol{\theta})]$. One possibility is to use a **Markov chain**: this is then *Markov chain* Monte Carlo. A Markov chain generates each iterate $\boldsymbol{\theta}^{(i)}$, taking into account only the previous value $\boldsymbol{\theta}^{(i-1)}$. Subject to some regularity conditions, a Markov chain will generate samples $\boldsymbol{\theta}^{(i)}$ from its *stationary* distribution, for large i.

In general, it is surprisingly easy to construct a Markov chain the stationary distribution of which is our target distribution $f(\boldsymbol{\theta}|\mathbf{x})$, and for which (1) is a consistent estimator of $\mathrm{E}_f[g(\boldsymbol{\theta})]$. The method was first proposed in 1953 by Metropolis et al. [24], and was generalized in 1970 by Hastings [23]. For many years, the **algorithm** was used mainly in the field of statistical mechanics. In 1984 the *Gibbs sampling* algorithm (later recognized as a special case of the Metropolis–Hastings algorithm) was proposed by Geman & Geman [15] as a tool for image reconstruction (*see* **Image Analysis and Tomography**). In 1990, the considerable potential of the Gibbs sampler was brought to the attention of the wider statistical community by Gelfand & Smith [8]. A generalization of the Metropolis–Hastings algorithm was proposed by Green [22] in 1995.

The Metropolis–Hastings Algorithm

We now describe the Metropolis–Hastings algorithm. For notational convenience we suppress dependence on data $\mathbf{x}$, and for the moment we continue to assume that $\boldsymbol{\theta}$ is a continuous random vector. We begin the chain with an arbitrary starting value, $\boldsymbol{\theta}^{(0)}$, and then produce the chain $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ by iterating around the following two steps. At each iteration $i + 1$:

Step 1: generate a *candidate* value $\boldsymbol{\theta}'$ from a *proposal distribution* $q(\cdot|\boldsymbol{\theta}^{(i)})$;

Step 2: with probability

$$\alpha(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}') = \min\left[1, \frac{f(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}')}{f(\boldsymbol{\theta}^{(i)})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)})}\right] \tag{2}$$

accept the candidate (i.e. set $\boldsymbol{\theta}^{(i+1)}$ equal to $\boldsymbol{\theta}'$); otherwise reject the candidate (i.e. set $\boldsymbol{\theta}^{(i+1)}$ equal to $\boldsymbol{\theta}^{(i)}$).

Heuristically, we aim to generate dependent samples from $f(\cdot)$ by sampling from a more convenient distribution at Step 1, and then correcting for this in a rather unintuitive but appropriate way at Step 2. To implement Step 2, generate a **pseudo-random number** u from a **uniform (0, 1) distribution**. If $u \leq \alpha(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}')$ accept $\boldsymbol{\theta}'$; otherwise reject it. The choice of the proposal density $q(\cdot|\cdot)$ is largely up to the user, the prime considerations being computational convenience and rapid mixing (see below). Note that the target density $f(\cdot)$ need not be normalized to integrate to one, since the normalization constant cancels

in (2). This is particularly convenient for Bayesian analyses, where the posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$ is proportional to the **likelihood** $p(\mathbf{x}|\boldsymbol{\theta})$ times the **prior** $p(\boldsymbol{\theta})$. Thus $p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ can be used in place of $f(\cdot)$ in (2), and there is no need to evaluate the normalization constant $\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$.

For the Metropolis–Hastings chain to be useful, it must be *irreducible*. Informally, irreducibility means that the chain is able to reach anywhere within the domain of $f(\cdot)$ within a finite number of iterations (see [37] for a more careful definition). If the chain is irreducible, it will eventually settle down to produce samples $\boldsymbol{\theta}^{(i)}$ from its *stationary* distribution, which can be shown to be $f(\cdot)$. Thus the choice of starting value $\boldsymbol{\theta}^{(0)}$ is not important, although it is generally advisable to avoid starting values well into the tails of $f(\boldsymbol{\theta})$, which could delay convergence to $f(\cdot)$. There is no particular advantage in starting the chain at the mode of $f(\cdot)$, since the chain must still be run long enough for it to "forget" its starting value. If the chain is reducible, it will never forget its starting value, since the starting value will determine which parts of the space can be reached. Irreducibility is generally easily verified by inspecting the form of $q(\cdot|\cdot)$, although in some genetics applications involving complex pedigrees, establishing irreducibility can be difficult (see [30]).

When applied to output from an irreducible Metropolis–Hastings chain, (1) is a consistent estimator of $\mathrm{E}_f[g(\boldsymbol{\theta})]$. In calculating (1) it is usual to discard the first m iterates of the chain (the *burn-in*), during which the chain exhibits dependence on the starting value $\boldsymbol{\theta}^{(0)}$. Several methods have been developed for diagnosing *convergence* (i.e. determining m) and for determining the run length n. Most are approximate in some way, and the most popular [12, 25] monitor the sample path of only univariate quantities, such as a single element of $\boldsymbol{\theta}$. Some methods rely on output from a single chain, and others require multiple chains to be run. For a recent review of convergence diagnostics, see Cowles & Carlin [5]. There is some debate in the literature regarding the number of chains to run: Gelman & Rubin [11, 12] advocate several long chains, while Geyer [16] recommends one very long chain. There is no justification for running a large number of short chains.

There is no justification for attempting to create pseudo-independent samples for input into (1) by *thinning* the output (i.e. using only every jth iterate from the chain) or, even worse, by running a large number of short chains and using only the last iterate from each. The theory of ergodic Markov chains guarantees the consistency of (1), despite the obvious lack of independence within the chain. The only justification for thinning is to reduce computer storage requirements.

Besides being irreducible, it is also important that the chain is *geometrically ergodic*; in other words, that convergence towards the stationary distribution $f(\cdot)$ proceeds at a geometric rate (see [37]). Unless the chain is geometrically ergodic, it is not possible to say anything useful about the variance of (1), and the chain may be very badly behaved, producing long meanders and erratic behavior. Nongeometrically ergodic chains are therefore effectively useless. Unfortunately, establishing geometric ergodicity in any particular context is not trivial, and theory tends to lag somewhat behind practice, although useful results have been obtained (see [28] and references therein). If the chain is geometrically ergodic, the variance of (1) can be estimated. A popular method is the method of *batch means*: the output is divided into n_1 consecutive batches of size n_2, and the sample mean $\overline{g}_j$ of $g(\cdot)$ within each batch j is calculated. If n_2 is large enough, the batch means will be approximately independent, and the variance of (1) can be estimated as n_1^{-1} times the usual sample variance of $\overline{g}_1, \overline{g}_2, \ldots, \overline{g}_{n_1}$. Often, n_1 is set to about 20. Other methods of variance estimation are given by Geyer [16]. An estimate of the variance of (1) can be used to calculate how much longer to run the chain.

The Metropolis–Hastings algorithm is not limited to situations in which $\boldsymbol{\theta}$ is a continuous random vector, although this is the usual situation in biostatistical applications. Discrete variables occur in genetics applications, where some elements of $\boldsymbol{\theta}$ are unobserved **genotypes**, and in applications where discrete-valued covariates are missing. The same form of acceptance probability (2) applies regardless of whether $f(\cdot)$ is a probability, a probability density, a product of probabilities and densities, or a density with respect to an arbitrary measure.

Proposal Distributions

Considerable freedom can be exercised in the choice of proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)})$, provided that the resulting chain is both irreducible and geometrically ergodic. In a Bayesian context, $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)})$ may also depend on the data $\mathbf{x}$. A *symmetric* proposal,

for which $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)}) = q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}')$ for all $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}^{(i)}$, results in an acceptance probability (2) which does not depend on $q(\cdot|\cdot)$. This is the form described in the original algorithm of Metropolis et al. [24]. An *independence* proposal is one which does not depend on $\boldsymbol{\theta}^{(i)}$, so $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)}) = q(\boldsymbol{\theta}')$. Independence proposals can be either very good (trivially, setting $q(\boldsymbol{\theta}') = f(\boldsymbol{\theta}')$ results in an acceptance probability of 1.0 and independent sampling from $f(\cdot)$), or very bad (if the tails of $q(\cdot)$ are lighter than those of $f(\cdot)$, then the chain will not even be geometrically ergodic [28]).

Many other forms of proposal are possible; see Tierney [36]. Different choices will result in different rates of *mixing*. A rapidly mixing chain will move about the domain of $f(\cdot)$ fluidly, and will quickly converge to $f(\cdot)$. A slow-mixing chain will exhibit significant long-lag **autocorrelations**, and will require a very long run to obtain adequate precision in (1). It is difficult in general to predict the behavior of any particular choice of proposal. A proposal distribution which nearly always results in rejection at Step 2 will be slow-mixing, since the chain will only occasionally move. However, a proposal distribution which nearly always results in acceptance may also be slow mixing, if the high acceptance rate is achieved by proposing only very small steps. Gelman et al. [14] show, for a large class of problems using a symmetric proposal, that one should aim for acceptance rates in the range 0.15–0.5.

Hybrid chains [36] employ a set of proposal distributions $q_1(\cdot|\cdot), q_2(\cdot|\cdot), \ldots$, at each iteration choosing one proposal either randomly, or deterministically by cycling through the set. For example q_1 could be a symmetric proposal, and q_2 an independence proposal. A hybrid chain is often better than the sum of its parts; for example, it may be irreducible and geometrically ergodic even if none of the constituent single-proposal chains are.

Single-component Metropolis–Hastings is a special case of a hybrid chain. Vector $\boldsymbol{\theta}$ is partitioned into k components $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$; for example, each component could be just one element of $\boldsymbol{\theta}$. For each component j, a proposal $q_j(\boldsymbol{\theta}'_j|\boldsymbol{\theta}^{(i)})$ is defined which updates only component j, generating a candidate point $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{j-1}^{(i)}, \boldsymbol{\theta}'_j, \boldsymbol{\theta}_{j+1}^{(i)}, \ldots, \boldsymbol{\theta}_k^{(i)})$. The acceptance probability (2) then becomes

$$\alpha(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}') = \min\left[1, \frac{f(\boldsymbol{\theta}')q_j(\boldsymbol{\theta}_j^{(i)}|\boldsymbol{\theta}')}{f(\boldsymbol{\theta}^{(i)})q_j(\boldsymbol{\theta}'_j|\boldsymbol{\theta}^{(i)})}\right], \qquad (3)$$

assuming that the choice of proposal q_j does not depend on the current $\boldsymbol{\theta}$. Most applications of Metropolis–Hastings use single-component updating, since it is much easier to construct proposals in low dimensions. However, when $f(\cdot)$ specifies high correlations between elements of $\boldsymbol{\theta}$, single-component updating can produce very slow mixing, unless highly correlated elements are blocked into the same component.

The *Gibbs sampler* is a special case of single-component Metropolis–Hastings, in which

$$q_j(\boldsymbol{\theta}'_j|\boldsymbol{\theta}^{(i)}) = f(\boldsymbol{\theta}'_j|\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{j-1}^{(i)}, \boldsymbol{\theta}_{j+1}^{(i)}, \ldots, \boldsymbol{\theta}_k^{(i)}), \qquad (4)$$

where the conditional distribution $f(\boldsymbol{\theta}_j|\cdot)$ is derived from the target joint distribution $f(\boldsymbol{\theta})$, and is called the *full conditional distribution* of $\boldsymbol{\theta}_j$. When $f(\boldsymbol{\theta})$ is a product of terms, as in the applications described below, the full conditional distribution for $\boldsymbol{\theta}_j$ is proportional to the product of those terms containing $\boldsymbol{\theta}_j$. With proposal distributions of the form of (4), the acceptance probability (3) is equal to 1.0, so the chain never rejects. As for generic single-component Metropolis–Hastings, an iteration of the Gibbs sampler involves updating only one $\boldsymbol{\theta}_j$; subsequent iterations may choose j at random, or deterministically by cycling through $j = 1, \ldots, k$. Thus the Gibbs sampler consists entirely in sampling from full conditional distributions, at each iteration updating one parameter, conditioning on the current values of all the other parameters (and data). Note that, in other articles, a Gibbs iteration is sometimes defined as one complete cycle of updating. Below we use $f(\boldsymbol{\theta}'_j|\cdot)$ to denote (4).

Sampling from full conditional distributions can be difficult, but if they are univariate and log-concave (which they often are), *adaptive rejection sampling (ARS)* can be used [19]. See Gilks [17] for further details on constructing and sampling from full conditional distributions. Gibbs sampling and ARS are implemented in the BUGS **software** [33], for general-purpose Bayesian modeling.

Slow mixing in the Gibbs sampler, or any other single-component updating method, can sometimes be resolved by reparameterization. An important example occurs in Bayesian linear models. The linear predictor $\alpha_0 + \alpha_1 x_{1\ell} + \cdots + \alpha_p x_{p\ell}$, where $x_{1\ell}, \ldots x_{p\ell}$ denote **covariates** for individual ℓ, should be reparameterized as $\beta_0 + \alpha_1(x_{1\ell} - \overline{x}_1) + \cdots + \alpha_p(x_{p\ell} - \overline{x}_p)$, where $\overline{x}_j$ denotes the sample mean of $\{x_{j\ell}, \ell = 1, 2, \ldots\}$. Centering the covariates in

this way will reduce posterior correlations between the intercept α_0 and the regression coefficients α_j. Another important example of reparameterization occurs in **hierarchical models**. Consider the Bayesian **random effects model**

$$y_{j\ell} \sim \mathrm{N}(\mu + \alpha_j, \sigma_1^2), \quad \alpha_j \sim \mathrm{N}(0, \sigma_2^2),$$
$$\mu \sim \mathrm{N}(0, \sigma_3^2), \tag{5}$$

where $j = 1, \ldots, m, \ell = 1, \ldots, n$, and the $y_{j\ell}$ are observed data. Gelfand et al. [9] show that the Gibbs sampler mixes poorly for this problem if n is large in relation to m. Thus the mixing rate deteriorates as information on the random effects increases, contradicting a common supposition that mixing is worst when information is scarce. Gelfand et al. [9] suggest a simple reparameterization, which they call *hierarchical centering*:

$$y_{j\ell} \sim \mathrm{N}(\beta_j, \sigma_1^2), \quad \beta_j \sim \mathrm{N}(\mu, \sigma_2^2).$$

With this parameterization, the mixing rate increases with both m and n. The idea extends to more complex hierarchical models with nested random effects. See Roberts & Sahu [29] for a rigorous theoretical evaluation of reparameterization strategies in hierarchical models.

Various other strategies have been devised for improving mixing: see Gilks & Roberts [18] for a review.

Reversible-Jump MCMC

The preceding discussion implicitly assumes that the length of vector $\boldsymbol{\theta}$ is fixed and known. However, Green [22] has recently demonstrated that important classes of models contain a variable number of parameters, and that the Metropolis–Hastings algorithm extends naturally to these situations. Examples of such models include mixture models [27] where the number of mixture components is unknown, and **change-point problems** with an unknown number of changepoints. Such models allow an essentially nonparametric approach to curve-fitting. Another important example concerns **model choice** or model averaging, where several models must be entertained, possibly varying in number of parameters. From a Bayesian perspective, the individual models can be thought of as components of an encompassing model (*see* **Bayesian Methods for Model Comparison**).

The general problem is best conveyed with a toy example. Assume that survival times $\mathbf{x}$ in a clinical trial are **exponentially** distributed, and let θ_1 and θ_2 denote log-mortality rates for each arm of the trial. We consider two models: Model 1 asserts that $\theta_1 = \theta_2$; and Model 2 that $\theta_1 \neq \theta_2$. Let $k = 1, 2$ index the models. We place **priors** on k; on θ_1 given $k = 1$; and on θ_1 and θ_2 given $k = 2$. These ingredients define the posterior distribution $f(k, \theta|\mathbf{x})$, where the length of θ is equal to k. We can consider various types of proposal distribution. For example, proposal type A could change $\boldsymbol{\theta}$ without changing k, and since this does not affect the dimensionality, the usual acceptance formula (2) applies. Proposal type B could change k. For example, if $k^{(i)} = 1$, type B1 could set $k' = 2, \theta_1' = \theta_1^{(i)}$ and sample $\theta_2' \sim \mathrm{N}(\theta_1^{(i)}, \sigma^2)$, where σ^2 is fixed. If $k^{(i)} = 2$, type B2 could set $k' = 1, \theta_2' = \theta_1' = \theta_1^{(i)}$. Note that, in this example, proposal B2 involves no sampling. Assume that, at any iteration, a type B proposal is chosen with probability 0.5. Then the acceptance probability of a B1 move is, from (2),

$$\min\left[1, \frac{f(2, \theta_1', \theta_2'|\mathbf{x})\,\mathrm{d}\theta_1'\,\mathrm{d}\theta_2' \times I(\theta_1' = \theta_1^{(i)})}{\left\{\begin{array}{l} f(1, \theta_1^{(i)}|\mathbf{x})\,\mathrm{d}\theta_1^{(i)} \times (2\pi)^{-1/2}\sigma^{-1} \\ \times \exp\{-[1/(2\sigma^2)](\theta_2' - \theta_1^{(i)})^2\}\,\mathrm{d}\theta_2' \end{array}\right\}}\right]. \tag{6}$$

where $I(\cdot)$ denotes the indicator function (*see* **Dummy Variables**). Notice that each density in (6) is converted into a probability through postmultiplication by dimensional terms $\mathrm{d}\theta_1^{(i)}$, etc. However, these dimensional terms cancel in the numerator and denominator, and so they can be ignored. This is a consequence of *dimension matching* in the proposal distributions. Dimension matching is not automatic. For example, suppose that the B1 proposal samples both $\theta_1' \sim \mathrm{N}(\theta_1^{(i)}, \sigma^2)$ and $\theta_2' \sim \mathrm{N}(\theta_1^{(i)}, \sigma^2)$, but the B2 proposal is as before; then the denominator in (6) would become

$$f(1, \theta_1^{(i)}|\mathbf{x})\,\mathrm{d}\theta_1^{(i)} \times (2\pi)^{-1/2}\sigma^{-1} \exp\left\{-\left[\frac{1}{(2\sigma^2)}\right] \times [(\theta_1' - \theta_1^{(i)})^2 + (\theta_2' - \theta_1^{(i)})^2]\right\} \mathrm{d}\theta_1'\,\mathrm{d}\theta_2'.$$

Now the dimensional terms no longer cancel, so the algorithm is not well-defined.

This example illustrates that the usual Metropolis–Hastings algorithm can be used when the dimensionality of $\boldsymbol{\theta}$ is unknown. Proposal distributions may propose changes to the dimension, but for each such proposal it must be checked that the reverse proposal satisfies the dimension-matching requirement. Of course, all the usual problems of mixing still apply to this more general framework.

Applications

By now, Markov chain Monte Carlo techniques have been applied in most areas of statistics, in particular biostatistics. For example, the book edited by Gilks et al. [20] contains applications in vaccine efficacy (*see* **Vaccine Studies**), clinical monitoring (*see* **Data and Safety Monitoring**), **pharmacokinetics**, disease **mapping**, medical imaging (*see* **Image Analysis and Tomography**), genetics (*see* **Human Genetics, Overview**), and epidemiologic **measurement error**. Also, the book edited by Berry & Stangl [2] includes applications in medical **decision analysis**, **clinical trial** design, **crossover trials**, **meta-analysis** and changepoint analysis of randomized trials, pharmacokinetics, tumor hemodynamics (*see* **Tumor Growth**) and perinatal mortality (*see* **Infant and Perinatal Mortality**). Rather than attempting to review biostatistical applications of MCMC *per se*, we focus on applications of MCMC in modeling situations familiar to biostatisticians; specifically hierarchical models, **missing data**, **censored data**, measurement error, and temporally or spatially correlated data (*see* **Geographic Epidemiology**; **Time Series**).

Hierarchical Models

By far the most common area of application of MCMC has been to hierarchical models, such as (5) (see, for example, [7, 8], and [10]). Most applications employ the Gibbs sampler, since full conditional distributions for the random effect parameters involve only a small subset of the data, and are generally log concave. For example, in the simple hierarchical model (5), assuming for convenience that variance parameters are known, the full conditional distributions are:

$$p(\alpha_j|\cdot) = \mathrm{N}\left(\frac{\overline{y}_{j\cdot} - \mu}{1 + n^{-1}\sigma_1^2\sigma_2^{-2}}, \frac{1}{n\sigma_1^{-2} + \sigma_2^{-2}}\right), \quad (7)$$

$$p(\mu|\cdot) = \mathrm{N}\left(\frac{\overline{y}_{\cdot\cdot} - \overline{\alpha}}{1 + m^{-1}n^{-1}\sigma_1^2\sigma_3^{-2}}, \frac{1}{mn\sigma_1^{-2} + \sigma_3^{-2}}\right), \quad (8)$$

where $\overline{y}_{j\cdot} = \sum_{\ell=1}^{n} y_{j\ell}/n$, $\overline{y}_{\cdot\cdot} = \sum_{j=1}^{m}\sum_{\ell=1}^{n} y_{i\ell}/(mn)$, and $\overline{\alpha}_{\cdot} = \sum_{j=1}^{m}\alpha_j/m$. Running the Gibbs sampler corresponds to sampling from (7) for each j, and from (8), where all variables being conditioned upon take on their most recently sampled values.

Much more elaborate hierarchical models, with covariates, multivariate responses, and more levels in the hierarchy, can also be handled straight forwardly using Gibbs sampling; see for example, Gilks et al. [21]. In most applications, the Gibbs sampler mixes well, and when it does not, reparameterization strategies can be tried (see above). The current popularity of MCMC owes much to its successful application to hierarchical models, which are ubiquitous in biostatistics. In particular, Smith et al. [31] have applied such models to meta-analysis problems.

Imperfect Data

Most, if not all, biostatistical data sets contain imperfections due to missing, censored, or inaccurately measured data. In the pre-Gibbs era, such imperfections were often difficult to handle, and required problem-specific solutions. Here we show that a wide class of data-imperfection problems can be handled in a generic framework, using the Gibbs sampler.

Suppose that a dependent variable, y_ℓ, and a **covariate**, x_ℓ, have been recorded for each individual $\ell = 1, \ldots, n$. Assuming that the y_ℓ are conditionally independent, with probability density specified by $p(y_\ell|x_\ell, \boldsymbol{\theta})$, the posterior distribution of the model parameters $\boldsymbol{\theta}$ is proportional to

$$\prod_{\ell}^{n} p(y_\ell|x_\ell, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}), \quad (9)$$

where $p(\boldsymbol{\theta})$ denotes the prior density for $\boldsymbol{\theta}$.

Now suppose that the data are imperfect in some way. Let z_ℓ denote the observations on individual ℓ, and let $p(z_\ell|x_\ell, y_\ell, \boldsymbol{\phi})$ be a model describing the relationship between the observed, imperfect, data and the ideal data. Assuming that covariates are conditionally independent in the population, with probability specified by $p(x_\ell|\boldsymbol{\psi})$, the joint posterior

distribution of $\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}$ and $\{x_\ell, y_\ell\}$ is proportional to

$$\prod_\ell^n p(x_\ell|\boldsymbol{\psi}) \times p(y_\ell|x_\ell, \boldsymbol{\theta}) \times p(z_\ell|x_\ell, y_\ell, \boldsymbol{\phi}) \times p(\boldsymbol{\psi})p(\boldsymbol{\theta})p(\boldsymbol{\phi}), \quad (10)$$

assuming independent priors.

For example, if the covariates x_ℓ are measured with error, then z_ℓ represents the measured value of x_ℓ and $\boldsymbol{\phi}$ will include parameters specifying the bias and precision of the measurement process (*see* **Errors in Variables**). In many applications, the dependence of z_ℓ on y_ℓ in the measurement model will be dropped. Similarly, if the dependent variable is measured with error, then z_ℓ represents the measured value of y_ℓ, and the measurement model may drop the dependence on x_ℓ. If the covariates are error-free, the first term in (10) may be omitted, as it will not affect inference for the other variables. See Spiegelhalter et al. [32] for an application in which both dependent and independent variables are measured with error. In studies containing substantial measurement error, external validation studies may be performed, which will introduce further multiplicative terms in (10): see Richardson & Gilks [26] for details.

The above set-up also includes missing data as a special case. Suppose that x_ℓ is not recorded for some individuals. Then z_ℓ records whether x_ℓ has been recorded (for example, $z_\ell = 1$ if x_ℓ is missing; $z_\ell = 0$ otherwise), and $p(z_\ell|x_\ell, y_\ell, \boldsymbol{\phi})$ describes the probability that x_ℓ is missing. The set-up allows for the possibility that the missingness of x_ℓ may depend on x_ℓ itself, or on y_ℓ, or both. If z_ℓ does not depend on x_ℓ, the term $p(z_\ell|x_\ell, y_\ell, \boldsymbol{\phi})$ in (10) may be omitted. However, it is important that this term is retained if it does depend on x_ℓ, as this will affect the posterior distribution of x_ℓ, and hence of $\boldsymbol{\theta}$ (i.e. the missingness is informative; *see* **Nonignorable Dropout in Longitudinal Studies**). Similar considerations apply if y_ℓ is missing for some individuals: here z_ℓ indicates the missingness of y_ℓ, which is informative if it depends on y_ℓ. The above set-up also accommodates situations in which both x_ℓ and y_ℓ can be missing. An important class of missing data problems occurs in the field of genetics, where the missing covariate data x_ℓ are unobserved genotypes in a pedigree, and the observed data y_ℓ are phenotypes or marker genotypes. In such problems, the x_ℓ are not conditionally independent given $\boldsymbol{\psi}$, so the analysis framework described here would need to be adapted; see, for example, Thompson & Guo [35] or Thomas & Gauderman [34].

Censored data can also be accommodated in the above framework. If the dependent variable is right-censored at y_ℓ^*, then $z_\ell = (y_\ell^{**}, c_\ell)$, where $y_\ell^{**} = \min(y_\ell, y_\ell^*)$ and $c_\ell = 1$ if $y_\ell > y_\ell^*$, $c_\ell = 0$ otherwise. The model allows for informative censoring; noninformative censoring obtains when y_ℓ^* does not depend on y_ℓ, given x_ℓ. Similarly, censored covariates can also be accommodated; see Gilks et al. [21] for an example.

The Gibbs sampler, applied to the posterior (10), involves the following full conditional distributions:

$$p(\boldsymbol{\theta}|\cdot) \propto \prod_\ell^n p(y_\ell|x_\ell, \boldsymbol{\theta})p(\boldsymbol{\theta}), \quad (11)$$

$$p(\boldsymbol{\psi}|\cdot) \propto \prod_\ell^n p(x_\ell|\boldsymbol{\psi})p(\boldsymbol{\psi}), \quad (12)$$

$$p(\boldsymbol{\phi}|\cdot) \propto \prod_\ell^n p(z_\ell|x_\ell, y_\ell, \boldsymbol{\phi})p(\boldsymbol{\phi}), \quad (13)$$

$$p(x_\ell|\cdot) \propto p(x_\ell|\boldsymbol{\psi}) \times p(y_\ell|x_\ell, \boldsymbol{\theta}) \times p(z_\ell|x_\ell, y_\ell, \boldsymbol{\phi}), \quad (14)$$

$$p(y_\ell|\cdot) \propto p(y_\ell|x_\ell, \boldsymbol{\theta}) \times p(z_\ell|x_\ell, y_\ell, \boldsymbol{\phi}). \quad (15)$$

Running the Gibbs sampler corresponds to sampling from (11)–(13) and, for each ℓ, from (14) and (15), where all variables being conditioned upon take on their most recently sampled values. If any of these full conditional distributions is awkward to sample from directly, it can be replaced by a set of lower-dimensional full conditional distributions, or by a single Metropolis–Hastings step. The point to note about (11) is that it has the same form as the posterior distribution of $\boldsymbol{\theta}$ given full, accurately measured data, as in (9), so the sampling involved in this part of the Markov chain presents no new difficulties. The full conditionals (14) and (15) should be easy to sample from, since they involve only a small subset of the data and would typically be low-dimensional.

A special problem arises when n itself is unknown, due to an unknown number of individuals being selectively lost from the study (for whom, of course, x_ℓ and y_ℓ are unknown). Thus the posterior distribution is variably dimensioned, since it involves an unknown number of missing x_ℓ and y_ℓ variables. For this problem, a reversible-jump Metropolis–Hastings

step would need to be included in the sampling, in which missing individuals are added or removed. De Angelis et al. [6] consider such a problem in AIDS epidemiology, where the missing individuals are those who have not yet been diagnosed with AIDS, but who are infected with the HIV virus (*see* **AIDS and HIV**).

Temporally or Spatially Correlated Data

In many biostatistical applications, it is useful to be able to specify relatedness between data items without attempting to model the causal connections between them. For example, in disease maps (*see* **Mapping Disease Patterns**), disease incidence in one county might be expected to be similar to disease incidence in neighboring counties, but direct causal links between them might not be realistic. Similarly, disease incidence in one calendar year might be expected to be similar to disease incidence in adjacent years, or changes in disease incidence might be similar to changes in adjacent years. Markov random field (MRF) models allow such dependence to be expressed purely descriptively, without causal implications. Four example, suppose that μ_ℓ is the disease incidence rate at time ℓ: then a MRF model might specify

$$p(\mu_\ell|\mu_{-\ell}, \boldsymbol{\theta}) = p(\mu_\ell|\mu_{\ell-1}, \mu_{\ell+1}, \boldsymbol{\theta}), \qquad (16)$$

for $\ell = 2, \ldots, n-1$, with some related form for $\ell = 1, n$, where $\mu_{-\ell}$ denotes $\{\mu_1, \ldots, \mu_{\ell-1}, \mu_{\ell+1}, \ldots, \mu_n\}$, and $\boldsymbol{\theta}$ is a set of parameters specifying the similarity of disease rates in adjacent years. Equation (16) says that μ_ℓ is conditionally independent of $\mu_1, \ldots, \mu_{\ell-2}, \mu_{\ell+2}, \ldots, \mu_n$, given $\mu_{\ell-1}$, $\mu_{\ell+1}$, and $\boldsymbol{\theta}$. This structure could be used to induce some smoothness in the **time series**. Note that a MRF model is nondirected, since, for example, the distribution of μ_ℓ is specified in terms of $\mu_{\ell-1}$, and vice versa. A second-order MRF model might specify

$$p(\delta_\ell|\delta_{-\ell}, \boldsymbol{\theta}) = p(\delta_\ell|\delta_{\ell-1}, \delta_{\ell+1}, \boldsymbol{\theta}), \qquad (17)$$

where $\delta_\ell = \mu_\ell - \mu_{\ell-1}$. This structure could be used to induce smoothness in the gradient of the time series.

Eq. (16) or (17) defines a MRF prior distribution on the unobserved, underlying, rates of disease incidence. For a noncontagious disease, observed disease incidence y_ℓ for each ℓ might be assumed to be independently **Poisson** (μ_ℓ). Bayesian inference for this problem, for known $\boldsymbol{\theta}$, is straightforward using Gibbs sampling, despite the nondirected structure of the MRF prior. Under (16), the full conditional distribution for μ_ℓ is simply

$$p(\mu_\ell|\cdot) \propto p(\mu_\ell|\mu_{-\ell}, \boldsymbol{\theta}) \times p(y_\ell|\mu_\ell), \qquad (18)$$

where $p(y_\ell|\mu_\ell)$ is Poisson (μ_ℓ). The Gibbs sampler simply involves sampling from (18) for each ℓ, always conditioning upon the most recently sampled values in $\mu_{-\ell}$. If $\boldsymbol{\theta}$ is unknown, it should be sampled from its full conditional distribution, but this is generally difficult to derive from (18). Besag et al. [4] discuss MRF prior models which are specified through joint "pairwise-difference" distributions, from which derivations of all full conditional distributions is straightforward. Besag et al. [3] and Bernardinelli & Montomoli [1] discuss Gibbs sampling for disease maps using MRF priors.

References

[1] Bernardinelli, L. & Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk, *Statistics in Medicine* **11**, 983–1007.

[2] Berry, D.A. & Stangl, D.K. (1996). *Bayesian Biostatistics*. Marcel Dekker, New York.

[3] Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**, 1–21.

[4] Besag, J., Green, P.J., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems, *Statistical Science* **10**, 3–41.

[5] Cowles, M.K. & Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association* **91**, 883–904.

[6] De Angelis, D., Gilks, W.R. & Day, N.E. (1998). Bayesian projection of the acquired immune deficiency epidemic, (with discussion), *Applied Statistics* **47**, 449–498.

[7] Dellaportas, P. & Smith, A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling, *Applied Statistics* **42**, 443–460.

[8] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[9] Gelfand, A.E., Sahu, S.K. & Carlin, B.P. (1995). Efficient parameterisations for normal linear mixed models, *Biometrika* **82**, 479–488.

[10] Gelfand, A.E., Hills, S.E., Racine-Poon, A. & Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *Journal of the American Statistical Association* **85**, 972–985.

[11] Gelman, A. (1996). Inference and monitoring convergence, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 131–143.

[12] Gelman, A. & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion), *Statistical Science* **7**, 457–511.

[13] Gelman, A. & Rubin, D.B. (1996). Markov chain Monte Carlo methods in biostatistics, *Statistical Methods in Medical Research* **5**, 339–355.

[14] Gelman, A., Roberts, G.O. & Gilks, W.R. (1996). Efficient Metropolis jumping rules, in *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 599–607.

[15] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[16] Geyer, C.J. (1992). Practical Markov chain Monte Carlo, *Statistical Science* **7**, 473–511.

[17] Gilks, W.R. (1996). Full conditional distributions, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 75–88.

[18] Gilks, W.R. & Roberts, G.O. (1996). Strategies for improving MCMC, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 89–114.

[19] Gilks, W.R. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling, *Applied Statistics* **41**, 337–348.

[20] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[21] Gilks, W.R., Wang, C.C., Yvonnet, B. & Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling, *Biometrics* **49**, 441–453.

[22] Green, P.J. (1995). Reversible jump MCMC computation and Bayesian model determination, *Biometrika* **82**, 711–732.

[23] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.

[24] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equations of state calculations by fast computing machine, *Journal of Chemical Physics* **21**, 1087–1091.

[25] Raftery, A.E. & Lewis, S.M. (1996). Implementing MCMC, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 115–130.

[26] Richardson, S. & Gilks, W.R. (1993). Conditional independence models for epidemiological studies with covariate measurement error, *Statistics in Medicine* **12**, 1703–1722.

[27] Richardson, S. & Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components, (with discussion) *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

[28] Roberts, G.O. (1996). Markov chain concepts related to sampling algorithms, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 45–57.

[29] Roberts, G.O. & Sahu, S.K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler, *Journal of the Royal Statistical Society, Series B* **59**, 291–317.

[30] Sheehan, N. & Thomas, A. (1993). On the irreducibility of a Markov chain defined on a space of genotypic configurations by a sampling scheme, *Biometrics* **49**, 163–175.

[31] Smith, T.C., Spiegelhalter, D.J. & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study, *Statistics in Medicine* **14**, 2685–2699.

[32] Spiegelhalter, D.J., Best, N.G., Gilks, W.R. & Inskip, H. (1996). Hepatitis B: a case study in MCMC methods, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 21–43.

[33] Spiegelhalter, D.J., Thomas, A., Best, N.G. & Gilks, W.R. (1995). *BUGS: Bayesian Inference using Gibbs Sampling*, Version 0.30. Medical Research Council Biostatistics Unit, Cambridge.

[34] Thomas, D.C. & Gauderman, W.J. (1996). Gibbs sampling methods in genetics, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 420–440.

[35] Thompson, E.A. & Guo, S.W. (1991). Evaluation of likelihood ratios for complex genetic models, *IMA Journal of Mathematical Applications in Medicine and Biology* **8**, 149–169.

[36] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics* **22**, 1701–1762.

[37] Tierney, L. (1996). Introduction to general state-space Markov chain theory, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 59–74.

(*See also* **Computer-intensive Methods**; **Markov Chain Monte Carlo, Recent Developments**)

W.R. GILKS

Markov Chains

Markov chains refer to a collection of **random variables** with a special dependency structure. We begin by discussing discrete time Markov chains which are sequences of random variables, denoted by $\{X_n, 0 \leq n < \infty\}$. These stochastic processes are a generalization of a sequence of independent discrete random variables. The random variables, X_n, assume a common discrete set of possible values, $\mathcal{S}$, called the *state space*. Since $\mathcal{S}$ is countable, we can, by relabeling the states, assume that they are labeled by the positive integers. For a finite state space, $\mathcal{S} = \{0, 1, \ldots, K\}$ for some K, while for a countably infinite state space $\mathcal{S} = \{0, 1, \ldots\}$. The dependency structure is defined by the *Markov property*, which is defined by:

$$\Pr(X_{n+1} = j | X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i) = \Pr(X_{n+1} = j | X_n = i).$$

If we allow the index n to represent the present time and $\{1, \ldots, n-1\}$ to represent the past, then the Markov property can be interpreted to imply that future events are conditionally independent of the past given the present. The Markov property implies that the chain, upon entering state i, will stay in that state for a random period governed by a **geometric** probability distribution.

The probability $\Pr(X_{n+1} = j | X_n = i)$ is called a *one-step transition probability*, a transition from state i to state j in one time unit. This probability depends upon three quantities: i, j, and n. If the transition probability is independent of n, then we say the transitions are *time homogeneous* or *stationary*. If, in addition, the transition probabilities are independent of i, then the Markov chain is a sequence of independent, identically distributed (iid) random variables. For the rest of this article, we assume time homogeneous transitions, the most commonly considered case.

It is convenient to represent the transition probabilities $p_{ij} = \Pr(X_1 = j | X_0 = i)$ in matrix form, $\mathbf{P} = (p_{ij})$, a square matrix the dimension of which equals that of $\mathcal{S}$. Each row of $\mathbf{P}$ is a discrete probability distribution, so the rows must satisfy two conditions; (i) $p_{ij} \geq 0$ and (ii) $\sum_{j \in \mathcal{S}} p_{ij} = 1$. A matrix $\mathbf{P}$ satisfying (i) and (ii) is called a one-step transition probability matrix. Such matrices include the case of identical rows where the underlying random variables are independent.

To give a complete description of a time homogeneous Markov chain, one needs to specify three components: (i) the state space $\mathcal{S}$; (ii) the transition probability matrix $\mathbf{P}$; and (iii) the initial probability distribution, $\Pr(X_0 = i), i \in \mathcal{S}$. Given those three quantities, the entire evolution of the Markov chain can be characterized. Often that evolution is described in two ways, the n-step transition probabilities; that is, $\Pr(X_{n+m} = j | X_m = i) = p_{ij}^{(n)}$, and the marginal probability distribution at time n, $\Pr(X_n = j)$. Both can be derived from the *Chapman–Kolmogorov equations*. For n-step transitions write $\Pr(X_{n+m} = j | X_0 = i) = \sum_{k \in \mathcal{S}} \Pr(\{X_m = k\} \cap \{X_{n+m} = j\} | X_0 = i) = \Pr(X_m = k | X_0 = i) \cdot \Pr(X_{m+n} = j | X_m = k) = \sum_{k \in \mathcal{S}} p_{ik}^{(m)} p_{kj}^{(n)}$. These equations can be most conveniently expressed in matrix form: $\mathbf{P}^{(m+n)} = \mathbf{P}^{(m)} \mathbf{P}^{(n)}$, where $\mathbf{P}^{(n)} = (p_{ij}^{(n)})$, the n-step transition probability matrix. By iterating this expression, it is easy to show that $\mathbf{P}^{(n)} = \mathbf{P}^n$; that is, the n-step transition probability matrix is the nth power of the one-step transition matrix.

If the Markov chain is initiated at time 0 with a probability distribution, $\pi^{(0)}$, then it follows that the marginal probability distribution at time n, $\pi^{(n)} = \pi^{(0)} P^{(n)} = \pi^{(0)} P^n$, the product of the initial probability distribution vector with the n-step transition matrix.

An important consideration is the limiting behavior of the Markov chain. Does it **converge** to a single state or does it continue to move through all the states in the state space? To answer this question, a classification of each state and a partitioning of the state space is introduced. This is done by first defining a relation between two states:

Definition. State $i \in \mathcal{S}$ communicates with state $j \in \mathcal{S}$ if and only if there exists an $n \geq 0$ such that $p_{ij}^{(n)} > 0$.

So state i communicates with state j if there is a positive probability that a chain starting in i will reach j in finite time.

Definition. States $i, j \in \mathcal{S}$ intercommunicate if i communicates with j and j communicates with i.

The relation defined by state intercommunication is an equivalence relation and it partitions the state space $\mathcal{S}$ into equivalence classes of intercommunicating states. When a Markov chain has a single

equivalence class of states, that chain is said to be *irreducible*.

The long-run behavior of the Markov chain can be deduced from a very simple idea, whether a Markov chain which starts in a state i is certain to eventually return to that state. If $\Pr(X_n = i \text{ for some } n > 0 | X_0 = i) = 1$, then eventual return is certain. However, because of the Markov property, if we are guaranteed of returning a single time, then once we return, we are guaranteed of a second return, and so on. Consequently, if one return is certain, then an infinite number of returns is also certain, and we say that state i is *recurrent*. However, if the probability of eventual return to i is less than 1, then with each return to i, there is a positive probability of this being the final visit to i. Eventually, no further returns will occur. In this case, the state i is called *transient*, and the number of visits to state i will be a random variable with a geometric distribution. One important result is that for any single equivalence class of states, all of those states are recurrent or all are transient.

If we are interested in the long-run fraction of time that the Markov chain spends in state i, we can restrict attention to recurrent states, because this relative frequency must converge to 0 for the transient states. For an irreducible Markov chain, the vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots)$, where α_i represents the long-run fraction of time spent in state i, satisfies the system of linear equations

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}\mathbf{P}, \qquad \sum_{i \in \mathcal{S}} \alpha_i = 1.$$

The vector $\boldsymbol{\alpha}$ represents a probability distribution. According to the theory of Markov chains, for an irreducible Markov chain these equations will have a unique solution giving a probability distribution or no solution at all. In the former case, the states of the Markov chain are *positive recurrent*, and the mean value of the time to return to state i is $1/\alpha_i$. In the latter case, either all the states are transient, or all the states are recurrent, but the mean value of the time to return to a state is infinite, in which case, the states are said to be *null recurrent*.

In the positive recurrent case, this distribution $\boldsymbol{\alpha}$ is referred to by a variety of names: the "equilibrium", "stationary", or "invariant" distribution; even the "steady state" distribution. The latter is often subject to misinterpretation, since the Markov chain continues to sojourn throughout the state space. The vector $\boldsymbol{\alpha}$ gives the long-run fraction of time spent in any state. It is also the invariant distribution in that if the chain is initiated according to the distribution $\boldsymbol{\alpha}$, then the marginal distribution of the chain at all future times is also $\boldsymbol{\alpha}$.

One can also look at the limiting behavior of the n-step transition probabilities, $\lim_{n\to\infty} p_{ij}^{(n)}$. Of course, this limit will be 0 if i does not communicate with j or if j is transient. However, even if $\alpha_j > 0$, this still does not guarantee that the limit exists. The phenomenon that must be considered is called *periodicity*.

Definition. A state $i \in \mathcal{S}$ is periodic with period k if $p_{ii}^{(n)} > 0$ only for $n = jk$, $j = 1, 2, \ldots$.

In words, a state i is periodic with period k implies that if the Markov chain is initially in state i, it can return to that state only in even multiples of k units of time. A chain with period 1 is *aperiodic*. Periodicity is also a class property; that is, all states in an equivalence class of states have the same period.

Theorem. For an aperiodic, irreducible, positive recurrent Markov chain, $\lim_{n\to\infty} p_{ij}^{(n)} = \alpha_j$.

Thus, for an aperiodic, irreducible, positive recurrent (also called an *ergodic*) Markov chain, the probability vector solution to the system of equilibrium equations $\boldsymbol{\alpha} = \boldsymbol{\alpha}\mathbf{P}$ gives the long-run average fraction of time that the chain spends in each state and is the limiting probability that the chain will be found in each state of $\mathcal{S}$ at a time far in the future. Notice that in the ergodic case, $\lim_{n\to\infty} p_{ij}^{(n)} = \alpha_j$, independent of i, the initial state. In this case, the long-run behavior of the chain is independent of its starting location.

A final concept that is useful in biostatistical applications is that of an *absorbing state*, a state i satisfying $p_{ii} = 1$. Once the chain enters i, it can never leave that state. Such states arise, for example, in cell metastasis models (*see* **Cell Cycle Models**). Starting with a normal cell, the cell may transition through a series of reversible states; however, if it reaches a cancerous state, then it continues to be cancerous for ever after. One can also use the concept as a method to determine the expected amount of time required for the chain, starting in state i, to first reach j. If we let e_{ij} represent the expected value of the first time the chain reaches j, then these quantities satisfy the system of equations: $e_{ij} = 1 + \sum_{k \in \mathcal{S}} p_{ik} e_{kj}$ for $j \neq i$, while $e_{ii} = 0$.

It is straightforward to estimate the elements of the transition matrix, **P**, by **maximum likelihood**. Assume that we are given a single path of a Markov chain that is observed over the time interval $[0, N]$. Each time the chain enters state i, the next transition is independent of the entire transition history, and that step is given by a discrete probability distribution on $\mathcal{S}$ given by $\{p_{ij}, j \in \mathcal{S}\}$. If we observe transitions over N steps, and N_{ij} gives the number of transitions from state i to j, then the maximum likelihood estimator of p_{ij} is given by N_{ij}/N_i, where $N_i = \sum_{j\in\mathcal{S}} N_{ij}$. In some models, the transition probabilities are constrained to have a special form which is a parametric function of some variable θ; that is, $p_{ij} = p_{ij}(\theta)$. Here, one must write the **likelihood** function of θ, an expression which will have the following **multinomial** form: $L(\theta|N_{ij}, i, j \in \mathcal{S}) = p_{X_0}(\theta) \prod_{i\in\mathcal{S}} \prod_{j\in\mathcal{S}} (p_{ij}(\theta))^{N_{ij}}$, where $p_{X_0}(\theta)$ represents the likelihood of the initial state of the Markov chain. This multinomial-like expression must be maximized over θ. The resulting estimators are asymptotically normally distributed, and the methodology is similar to what would be done with data from a **contingency table**. The seminal work on estimation of Markov chains was done by Anderson & Goodman [1]. The reader should also consult Billingsley [2].

Continuous Time Markov Chains

Many Markov chain models are formulated in continuous time, rather than discrete time. In this situation, the Markov chain is represented by $\{X_t, t \geq 0\}$. The state space, $\mathcal{S}$ is also discrete and is again taken to be $\{1, 2, \ldots, N\}$ for a finite Markov chain or $\{1, 2, \ldots, \}$ for the countable state space case.

The Markov property for the continuous time case is expressed by the relation

$$\Pr(X_{t+s} = j | X_s = i, X_u = i_u, 0 \leq u < s) = \Pr(X_{t+s} = j | X_s = i).$$

We consider only the time homogeneous case; that is, $\Pr(X_{t+s} = j|X_s = i) = p_{ij}(t)$, a transition from i to j over t time units which does not depend upon s. Again, the future evolution of the chain is conditionally independent of the past given the present state. Suppose at some time t, the chain is in state i, and we are interested in how much longer it will stay in state i before it jumps to a different state. The Markov property indicates that the remaining sojourn time in i must be independent of the past; hence it must be independent of the amount of time it has already sojourned in state i. This "memoryless" property implies that the sojourn time in each state is governed by an **exponential** distribution.

The basic ideas developed earlier for discrete time Markov chains carry over directly to continuous time Markov chains. For example, the state space can be decomposed into equivalence classes of intercommunicating states, and those classes contain states which are all recurrent or all transient. There is, however, a major difference between the discrete time and continuous time Markov chains. In the discrete time case, there is a smallest increment of time, one time unit. In the continuous time case, there is no smallest unit of time, so one can consider state transitions over arbitrarily small periods of time. Consequently, the concept of a one-step transition probability matrix, **P**, does not apply to the continuous time case. In its place, a *transition rate* matrix, $\mathbf{Q} = (q_{ij})$, is introduced. The individual transition rates are defined by $q_{ij} = p'_{ij}(0)$. Since $\sum_{j\in\mathcal{S}} p_{ij}(t) = 1$ for all t, it follows that $\sum_{j\in\mathcal{S}} q_{ij} = 0$. Consequently, each row of **Q** must sum to 0. The diagonal elements of **Q** are nonpositive, while the off-diagonal elements are nonnegative, since

$$q_{ij} = \begin{cases} \lim_{h\to 0} \dfrac{p_{ii}(h) - 1}{h} \leq 0, & \text{if } i = j, \\ \lim_{h\to 0} \dfrac{p_{ij}(h)}{h} \geq 0, & \text{if } i \neq j. \end{cases}$$

The elements of the **Q** matrix have direct interpretations concerning the behavior of the Markov chain. Recall that the holding time in each state is governed by an exponential distribution. The parameter of that distribution for state i is given by $-q_{ii} = \sum_{j\neq i} q_{ij}$. Once the chain leaves state i, it must jump to another state $j \neq i$. The probability that it jumps to j is given by $q_{ij}/\sum_{k\neq i} q_{ik}$. One could also associate each $\{q_{ij}, j \neq i\}$ with an independent exponential (q_{ij}) random variable, T_{ij}. Suppose that $T_i = \min_{j\neq i} T_{ij}$ and $T_{ij} < T_{ik}, k \neq i, j$. Then, upon entering state i, the chain will stay in state i for T_i time units, then jump to j.

The Chapman–Kolmogorov equations for the continuous time case are $\mathbf{P}(s + t) = \mathbf{P}(s)\mathbf{P}(t)$. These can be rewritten, $\mathbf{P}(t + h) = \mathbf{P}(t)\mathbf{P}(h)$. By subtracting

$\mathbf{P}(h)$ from both sides and taking the limit as $h \to 0$, we obtain the Kolmogorov forward equations,

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}, \qquad \mathbf{P}(0) = \mathbf{I},$$

a system of first-order differential equations with constant coefficients. These equations have a solution,

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) = \mathbf{I} + t\mathbf{Q} + \frac{t^2}{2!}\mathbf{Q}^2 + \cdots.$$

Suppose that one introduces the **eigenvalue** decomposition of $\mathbf{Q}$, $\mathbf{Q} = \mathbf{USV}$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of **eigenvectors**, and $\mathbf{S}$ is a diagonal matrix of eigenvalues of $\mathbf{Q}$. Using this representation of $\mathbf{Q}$, one can write $\mathbf{P}(t) = \mathbf{UD}(t)\mathbf{V}$, where $\mathbf{D}(t)$ is the diagonal matrix $\exp(t\mathbf{S})$. In the case of a positive recurrent, irreducible Markov chain, the transition probabilities converge to a limiting probability distribution, $\boldsymbol{\alpha}$, and this distribution is characterized by the equations

$$\mathbf{O} = \boldsymbol{\alpha}\mathbf{Q}, \sum_{i \in \mathcal{S}} \alpha_i = 1.$$

This equilibrium or stationary vector is the eigenvector corresponding to the eigenvalue 0 of $\mathbf{Q}$. For continuous time Markov chains, the concept of periodicity does not appear; hence, in the positive recurrent case, $\boldsymbol{\alpha}$ represents the long-run fraction of time the Markov chain spends in each of the states in the state space. In addition, $\alpha_j = \lim_{t \to \infty} p_{ij}(t)$.

The most common continuous time Markov chain is the birth–death process (*see* **Stochastic Processes**), a process in which transitions take place only to adjacent states in $\mathcal{S}; q_{ij} = 0$ if $j \neq i-1, i$ or $i+1$. Often, one uses the notation $\lambda_i = q_{ii+1}$ and $\mu_i = q_{ii-1}$, which denote the birth and death rates in state i. In the positive recurrent case, the stationary distribution of the chain is given by

$$\alpha_i = k \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j},$$

where k is a normalization constant to insure that this gives a probability distribution.

The estimation of the parameters of a continuous time Markov chain is similar to estimation in discrete time. We consider the case in which $\mathcal{S} = \{1, \ldots, K\}$. Assume that we are given a single sample path that is observed over the time interval $[0,T]$, and that the initial state is chosen at random. From this sample path, we can reduce to the **sufficient statistics** $\{N_{ij}, 1 \leq i, j, \leq K, i \neq j\}$, the total number of transitions from state i to state j and $\{\tau_i, 1 \leq i \leq K\}$, where τ_i represents the total amount of time spent in state i. The likelihood function is given by

$$L = \left[\prod_{i=1}^{K} \prod_{j=1, j \neq i}^{K} \left(\frac{q_{ij}}{-q_{ii}}\right)^{N_{ij}}\right] \left[\prod_{i=1}^{K} \exp(q_{ii}\tau_i)\right].$$

By taking logarithms, recalling that $q_{ii} = -\sum_{j \neq i} q_{ij}$ and maximizing this expression over q_{ij}, we find the maximum likelihood estimates of q_{ij} to be given by $\hat{q}_{ij} = N_{ij}/\tau_i$, provided that the denominator is positive. If $I_i = 0$, then state i was never entered, and we have no data from which to estimate the transition rates departing from state i. One can also use **Bayesian methods** in this problem. It is possible that a particular model might impose a parametric structure on the transition rates, a situation requiring a different estimation procedure.

Examples of Markov Chains in Biostatistics

We now illustrate the basic concepts discussed above using two classical examples in biostatistics.

Example 1. Radiation Damage

Reid & Landau [4] introduced a Markov chain to model the increase in or recovery from **radiation** damage to an organism. There are $K+1$ states, $\{0, \ldots, K\}$, where 0 denotes no radiation damage, K denotes an absorbing state with perceptible damage, and $\{1, \ldots, K-1\}$ denote intermediate states with increasingly severe states of damage. In discrete time, the one step transition probability matrix is given by the following $(K+1) \times (K+1)$ matrix in which $q_i + p_i = 1, 1 \leq i < K$,

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ q_1 & 0 & p_1 & 0 & \ldots & 0 \\ 0 & q_2 & 0 & p_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & q_{K-1} & 0 & p_{K-1} \\ 0 & 0 & \ldots & 0 & 0 & 1 \end{pmatrix}.$$

In this model, the states 0 and K are absorbing states. If the chain is any intermediate state, $1 \le i < K$, then it will move to an adjacent state. Once the chain hits an absorbing state, it stays there forever, hence the equivalence classes are $\{0\}, \{K\}, \{1, \ldots, K-1\}$. The single state classes are recurrent, while the intermediate class is transient. One might ask for the probability, given the chain is initiated in state i, $1 \le i \le K-1$, that it will reach the healthy state 0 before it reaches the permanently damaged state K. Reid & Landau studied this model under the assumption that $p_i = i/K$ and $q_i = 1 - i/K$. For this particular set of transitions, they showed that if the chain is in state i, then the probability of the chain hitting state K before it returns to the normal state 0 is given by

$$\frac{1}{2^{K-1}} \sum_{j=0}^{i-1} \binom{K-1}{j}, \quad \text{for } 1 \le i \le K.$$

Example 2. Compartment Models

Compartment models are a very large class of models used in **pharmacokinetics** and pharmacodynamics, tracking the flow of substances in the body. The compartments refer to containers such as organs or the blood stream itself. Jacquez [3] gives a comprehensive treatment of these models. While the number of particles of drug or pollutant will be very large, these models often assume independence of movement within the compartments. Thus, these models give the transition structure for one particle, and the behavior of the aggregate can be predicted using the **central limit theorem** and the **law of large numbers**.

A typical compartment model is formulated in continuous time. Consider, for example, a three-state model given by the $\mathbf{Q}$ matrix

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Recall that $q_{11} = -(q_{12} + q_{13})$ and $q_{22} = -q_{21}$. State 1 refers to the bloodstream, 2 refers to the liver, while 3 refers to the bladder, from which the pollutant will be expelled. The drug will reside in the bloodstream for an exponential period, then move either to the bladder or to the liver. The drug will stay in the liver for an exponential period, then move back to the bloodstream. Finally, any drug that reaches the bladder will be removed from the system, so this represents an absorbing state. Again, the bloodstream state and the liver state are transient, while state 3 is absorbing. Consequently, it is of interest to calculate the total amount of time that the pollutant will spend in the liver where damage can occur, and the time it spends in the system before it is removed.

Consider the following numerical example. Suppose that

$$\mathbf{Q} = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of $\mathbf{Q}$ are $(0, -0.382, -2.618)$. If we assume that there is a bolus injection of pollutant at time 0 into the bloodstream, then the transition probabilities for a single particle can be found from the Kolmogorov forward equations. Specifically, we find that

$$\begin{aligned} p_{11}(t) &= 0.276\exp(-0.382t) + 0.724\exp(-2.618t), \\ p_{12}(t) &= 0.448\exp(-0.382t) - 0.448\exp(-2.618t), \\ p_{13}(t) &= -0.724\exp(-0.382t) \\ &\quad -0.276\exp(-2.618t) + 1. \end{aligned}$$

Thus, the pollutant concentration decreases in the bloodstream according to a mixture of exponentials. It increases, then decreases in the liver, and eventually it all resides in the bladder.

References

[1] Anderson, T.W. & Goodman, L. (1957). Statistical inference for Markov chains, *Annals of Mathematical Statistics* **28**, 89–109.

[2] Billingsley, P. (1981). *Statistical Inference for Markov Processes*. Chicago University Press, Chicago.

[3] Jacquez, J.A. (1972). *Compartmental Analysis in Biology and Medicine*. Elsevier, Amsterdam.

[4] Reid, A.T. & Landau, H.G. (1951). A suggested chain process for radiation damage, *Bulletin of Mathematical Biophysics* **13**, 153–163.

J. LEHOCZKY

Markov Processes

A Markov process is often described as a "process without memory": our estimate of the probability of a future event concerning its behavior *given (complete)* information about its present state will not change if we are given in addition any information about its past behavior. See the first section for an illustration of this.

The concept of a "Markov process" is so general as to embrace most models of random systems evolving in time (*see* **Stochastic Processes**): from "classical" *random walks*, **Markov chains**, **branching processes**, *birth-and-death processes, diffusions* and their associated *stochastic differential equations*, to the "postclassical" *branching diffusions, measure-valued diffusions, interacting systems* (which include *contact processes* etc.), which are destined to play an ever more important part in **mathematical biology**. Each of the italicized topics has a huge literature and its own special methods; and it is often better to search the literature for *these* "keywords" rather than the all-embracing "Markov processes". However, Markov-process theory has, of course, unifying themes and methods (*martingale theory, large-deviation theory*, etc.) which pervade all of its branches.

We take a brief tour through the subject, designed to allow glimpses of several topics italicized above.

Simple Random Walk

Suppose that we toss a fair coin just before times 1, 2, 3, ..., and regard the state of our system at time n (which can be 0, 1, 2, ...) as the number of heads minus the number of tails obtained by that time. (We have $X_0 = 0$.) For illustration, regard time 100 as the "present". Suppose that we know that $X_{100} = 6$. *Given* this information and any additional information about the "past" results of the first 99 tosses, X_{101} is either 5 or 7 with probability 1/2 each. The Markov property is obvious here.

We can prove, for example, that the distribution of X_n at time n is approximated by the **normal distribution** of mean 0 and variance n (hence standard deviation $n^{1/2}$), and that, "almost surely" (that is, with probability exactly 1), X_n will fluctuate infinitely, taking every integer (whole-number) value infinitely often.

Markov Chain in Discrete Time

For a Markov chain in discrete time with stationary probabilities, $X_n, n = 0, 1, 2, \ldots$, is a random integer describing the state of the system at time n. The probabilistic law of the system is described by the ("initial") distribution of X_0 and by a "matrix" (or array) of transition probabilities p_{ij}: for each n, the (conditional) probability of the "future" event that $X_{n+1} = j$, given the "present" information that $X_n = i$ and any extra information about the "past" $X_0, X_1, \ldots, X_{n-1}$, is p_{ij}. [Our random walk has $p_{ij} = 1/2$ if $j = i + 1$ or $j = i - 1$, and $p_{ij} = 0$ otherwise.] The sort of questions in which we are interested are the following. Is the system "ergodic" in that, over the long term, it will almost surely share out its time amongst the various states in a predetermined way? At the other extreme, if there is an absorbing state a for which $p_{aa} = 1$, will the system almost surely eventually end up in state a? (For example, is some population almost surely destined to die out?)

Markov Chain in Continuous Time

We modify things so that our system can jump at *any* time, not just at integer-valued times, and this requires us to define *jump-rates* q_{ij} rather than "jump probabilities" p_{ij}. We denote the integer-valued state of our system at time t, where $t \geq 0$, by X_t. Suppose that $j \neq i$. Given the "present" information that $X_t = i$ and any information about the past values X_s for $s < t$, the probability that the system will jump from i to j between times t and $t + h$, where h is a small number, will be $q_{ij}h + o(h)$, where $o(h)$ is a term negligibly small *compared with* h when h tends to 0.

Generalized Birth-and-Death (GBD) Process

This has the property that $q_{ij} = a_i$ for some a_i if $j = i + 1$, $q_{ij} = b_i$ for some b_i if $j = i - 1$, and $q_{ij} = 0$ for other pairs (i, j), where $j \neq i$. Thus X can only jump to a neighboring state. We can answer the analogues of the questions raised in the section "Markov chain in discrete time".

Continuous-Time Random Walk (CTRW)

This is a GBD process with $a_i = b_i = 1/2$ for every i. It will behave rather like the random walk in the first section.

(Standard) Birth-and-Death (BD) Process

This (the simplest type of continuous-time *branching process*) is a GBD process in which only nonnegative integer states 0, 1, 2, ... are allowed, and we have $a_i = \lambda i, b_i = \mu i$ for some constants λ (the "birth rate" per individual) and μ (the "death rate" per individual). Here, we think of X_t as the size of a population at time t. If $X_t = i$, then there are i animals alive at time t, each of which can give birth (to one child) "at rate λ", resulting in a jump rate of $a_i = \lambda i$ for X from i to $i + 1$; if $i > 0$, then each of the i animals can die "at rate μ", resulting in a jump rate of $b_i = \mu i$ for X from i to $i - 1$. State 0 is absorbing. Here are some unsurprising results. In the subcritical case when $\mu > \lambda$, so the death rate exceeds the birth rate, and then the population will almost surely die out. In the supercritical case when $\lambda > \mu$, then, almost surely, the population will either die out or will grow "exponentially" in a sense which can be made precise. In the critical case when $\lambda = \mu$, the population will almost surely die out.

(Mathematical) Brownian Motion (BM) *B*

We now take the first of several steps in building more complex processes from the processes already introduced. If we renormalize CTRW suitably, then we obtain as a limit the most important of all stochastic processes: **Brownian motion**. We take a large number N, and let X be a GBD process with $a_i = N/2$ and $b_i = N/2$. This process is jumping very fast. Consider $Y_t = X_t/N^{1/2}$. Then Y is jumping just as fast as X, but is making only small jumps of size $1/N^{1/2}$. We choose $N^{1/2}$ because of the fact that a standard deviation of $n^{1/2}$ appeared in our discussion in the first section. What happens is that the law of the rescaled process Y converges as N tends to infinity to the law of Brownian motion B. The process B takes real values, not integer values. The Markov property of B is conveyed by the fact that conditional on the "present" information that $B_t = x$ and any information about the past $(B_s : s < t)$, the "future" random variable B_{t+s} has exactly the normal distribution of mean x and variance s. (The fact that the mean value of B_{t+s} given the values $(B_s : s \le t)$ is the value of B_t signifies that B is a *martingale*.)

It is no accident that, if $p(t, x)$ is the density function at x of the law of B_t if $B_0 = 0$, i.e. of the normal distribution with mean 0 and variance t, then

$$p(t, x) = \frac{1}{(2\pi t)^{1/2}} \exp\left(-\frac{x^2}{2t}\right)$$

solves the heat equation

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2 p}{\partial x^2}.$$

This explains the frequent occurrence of second-derivative "diffusion terms" in books on mathematical biology. The function $p(s; x, y) = p(s, y - x)$ is now the transition probability density for B from x to y in an interval of duration s: it plays a role analogous to that of the one-step transition probability p_{ij} for a Markov chain in discrete time.

Brownian motion B is quite remarkable. It approximates many processes. It is amazingly rich and we can "find within B" many other processes including all those we have so far studied. This gives a very illuminating way of proving (rigorously) the celebrated **Central Limit Theorem** on the ubiquitous nature of the normal distribution. We can even "find within B" seemingly much more complicated processes such as the Dawson–Watanabe process described later. We can use B to describe "white-noise perturbations" of ordinary differential equations, turning them into the stochastic differential equations which describe diffusions, and so on.

Diffusions

Brownian motion is the most important diffusion process. We can think of a more general diffusion process on the real line in two ways: either as the limit of a GBD process produced in a way analogous to that in which we obtained BM from CTRW; or as the solution of a *stochastic differential equation (SDE)* of the form

$$\frac{\mathrm{d}X}{\mathrm{d}t} = b(X_t) + \sigma(X_t)\frac{\mathrm{d}B}{\mathrm{d}t},$$

where σ and b are functions. This SDE can be thought of as a random perturbation of the ordinary

differential equation (ODE) $dX/dt = b(X_t)$. (The b here has a quite different connotation from that of the b_i in GBDs.) The Brownian path is nowhere differentiable, so dB/dt is completely meaningless in Newtonian terms. Even so, the great Japanese mathematician, K. Itô, constructed the *stochastic calculus* (nowadays based on *martingale theory*), which allows rigorous formulation and analysis of SDEs. Solutions of SDEs inherit the Markov property from the (particularly strong version of the) Markov property possessed by Brownian motion. Conversely, any (real-valued) Markov process X which fluctuates continuously in time (and which satisfies very mild regularity conditions) is the solution of an SDE as above.

The transition density function $p(s; x, y)$ of the solution X of our *first*-order SDE solves the *second*-order partial differential equation (PDE)

$$\frac{\partial p}{\partial s} = \frac{1}{2}\sigma(x)^2 \frac{\partial^2 p}{\partial x^2} + b(x)\frac{\partial p}{\partial x},$$

which amazing fact allows us to prove even the deepest known theorems on these and certain other PDEs of importance in mathematical biology by probabilistic methods (see below).

Diffusion theory has seen truly spectacular development over the last 40 years. One important way in which diffusions are used is again in approximating other processes: choosing a diffusion with the same "infinitesimal characteristics" as a more complex process can often give a good guide to how that process behaves.

Branching Brownian Motion and the FKPP Equation

The FKPP equation for $u(t, x)$ studied by Fisher and independently by Kolmogorov, Petrovskii, and Piskunov,

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2} + u(1-u),$$
$$u(0, x) = \begin{cases} 1, & \text{if } x < 0, \\ 0, & \text{if } x > 0, \end{cases} \qquad (1)$$

is perhaps the most famous in mathematical biology: it is the simplest *reaction–diffusion equation*. Here, $u(t, x)$ is thought of as describing the density at time t and position x of a population of animals, where there is a logistic constraint on population growth and where the animals diffuse around. (This statement does not in itself specify any random model!) Strange to say: by far the deepest analytic results on the equation have been obtained by H.P. McKean, M. Bramson, J. Neveu, and B. Chauvin & A. Rouault, using probabilistic methods on a stochastic model, branching Brownian motion (BBM), with "free" (rather than logistic) growth. The birthing is exactly as for the pure-birth process, which is a BD process with $\lambda = 1$ and $\mu = 0$. Each child is born at its parent's current position. Once born, animals perform independent Brownian motions. The whole system is Markov, the state of the system at time t summarizing both how many animals are then alive and exactly where they all are. McKean showed that the unique solution $u(t, x)$ of (1) is given by the probability that if we start with one animal born at position x at time 0, then at least one animal is to the left of 0 at time t. If $l(t)$ denotes the leftmost particle position at time t and we start with one animal born at time 0 at position 0, then we have, almost surely, as $t \to \infty$,

$$\frac{l(t)}{t} \to -2^{1/2},$$

and this explains the celebrated "approximate traveling-wave" nature of the solution to (1). If we simulate the situation when at time 0 there is just one animal at position 0, we see that the tracks of the animals almost exactly fill a triangle. All the exact traveling-wave solutions of the FKPP equation have explicit probabilistic representations, even though BBM is the wrong model in biological terms (see the section "MVDs with interaction; improving on FKPP" below).

Measure-Valued Diffusions (MVDs)

The simulation mentioned above is enough to convince one of the good sense of thinking of the random flows of measures ("mass distributions"). Let me explain the most fundamental MVD: the *Dawson–Watanabe (D–W) process*. We take a large number N. We think of each animal as having mass $1/N$, and start with N animals at position 0. Births and deaths are according to a BD process with $\lambda = N/2$ and $\mu = N/2$. Note that this is a critical situation, in which the process will eventually die out. Each particle performs BM, independently of all others. As

N tends to infinity, the law of evolution converges to that of the Dawson–Watanabe process. In one dimension, but only in one dimension, we can think of the value of the (D–W) process at time $t > 0$ as being a positive density function u of a mass distribution. This evolves as the solution of a *stochastic partial differential equation (SPDE)*, a perturbed heat equation

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2} + [u(t,x)]^{1/2} W,$$

where W now denotes a space–time white-noise process derived from a "Brownian motion" (the "Brownian sheet") with two-dimensional "time". Leading experts in this field include D.A. Dawson, E.B. Dynkin, S.N. Evans, J.F. le Gall, E. Perkins, and J.B. Walsh. Important biological applications, especially to genetics, have been given by D.A. Dawson, P. Donnelly, S.N. Ethier, and T.G. Kurtz.

Interacting Systems

Complicated as they are, the above-mentioned models are not complicated enough. Their particles perform their Brownian motions independently of one another: no account is taken of overcrowding or of the interaction between different particles. By contrast, the theory of interacting systems allows more or less anything. One has to be aware of the scope of this theory, even if only a handful of people can currently claim deep understanding. Leading experts include R. Durrett, G. Grimmett, H. Kesten, R. Holley, and T.M. Liggett.

One of the simplest interacting systems is the three-dimensional *contact-process* model for the spread of disease through cells which are considered to be cubes stacked together, and occupying the whole of space. A healthy cell becomes infected at "jump-rate" (a constant) I times the number of infected neighbors while an infected cell recovers at constant "jump-rate" R. (This system is too complicated to be a Markov chain – its state-space is "uncountable" – but it is a Markov process.) We suppose that at time 0 only a finite number of cells are infected. The system exhibits phase transition: if I/R is less than some critical number c (the precise value of which is currently unknown), then the disease will almost surely die out; while if I/R is greater than c, then the disease can (with positive probability) persist for ever, infecting ever more cells. The *time-dependent Ising model* from magnetism, now much used by statisticians in *image processing*, (*see* **Image Analysis and Tomography**), is closely related.

MVDs with Interaction; Improving on FKPP

Some extremely interesting work has been done by C. Mueller, R.B. Sowers and R. Tribe on an interacting system with "logistic" inhibition of **population growth**: an MVD with density satisfying the FKPP equation with an extra term $[u(1-u)]^{1/2} W$ on the right-hand side, with W as before. The system possesses a "coherence" not present in the deterministic model, and can be regarded as superior to it in many respects. Mueller, Sowers, and Tribe study "traveling-wave" aspects.

Self-Organization; Adapting to the Environment

A striking feature of (naturally occurring and man-made) interacting systems is their ability to behave in a pseudo-intelligent way. Brilliant use of this was made in the *Dynamic Alternative Routing* strategy for telephone networks developed by F.P. Kelly and R. Gibbens along with British Telecom. Similar use is made in **neural nets**. Biological "networks" in fungi, ant colonies, anastomosis of blood vessels near tumors (bad!), or in wound healing (good!), are complex interacting systems; at present, models of such systems can only be studied by **simulation**, which can certainly identify bad models even if it cannot conclusively validate good ones.

A Plea

There is a great need to make the more modern material described in this article a lot more accessible to applied workers – and to mathematicians too!

A Few References

The literature is truly vast. Most of it can be discovered via the key names given, in the various databases now available, and in the following. For the first five

sections, see, for example, Feller [6, 7], Grimmett & Stirzacker [8], and Karlin & Taylor [9, 10]. For the next two sections, see, for example, Breiman [1], Ethier & Kurtz [5], and Øksendal [12]. For the following two sections – and things are getting much harder now – see, for example, Dawson [2], Donnelly & Kurtz [3], Durrett & Levin [4], and Mueller & Sowers [11].

References

[1] Breiman, L. (1968). *Probability*. Addison-Wesley, Reading, Mass.

[2] Dawson, D.A. (1993). Measure-Valued Markov Processes, in *Ecole d'Eté de Probabilités de Saint-Flour XXI*, P.L. Henniquin, ed., *Lecture Notes in Mathematics* 1541, Springer-Verlag, Berlin, pp. 2–260.

[3] Donnelly, P. & Kurtz, T.G. (1996). A countable representation of the Fleming-Viot measure-valued diffusion, *Annals of Probability* **24**, 698–742.

[4] Durrett, R. & Levin, S.A. (1994). Stochastic spatial models – a user's guide to ecological applications, *Philosophical Transactions of the Royal Society of London* **343**, 329–350.

[5] Ethier, S.N. & Kurtz, T.G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.

[6] Feller, W. (1957). *Introduction to Probability Theory and its Applications*, Vol. 1, 2nd Ed. Wiley, New York.

[7] Feller, W. (1966). *Introduction to Probability Theory and its Applications*, Vol. 2, Wiley, New York.

[8] Grimmett, G. & Stirzacker, D. (1992). *Probability and Random Processes*, 2nd Ed. Oxford University Press, Oxford.

[9] Karlin, S. & Taylor, H.M. (1975). *A First Course in Stochastic Processes*. Academic Press, New York.

[10] Karlin, S. & Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

[11] Mueller, C. & Sowers, R.B. (1995). Random traveling waves for the KPP equation with noise, *Journal of Functional Analysis* **128**, 439–498.

[12] Øksendal, B. (1992). *Stochastic Differential Equations*, 3rd Ed. Springer-Verlag, Berlin.

(*See also* **Counting Process Methods in Survival Analysis**; **Epidemic Models, Spatial**; **Epidemic Models, Stochastic**; **Migration Processes**; **Queuing Processes**; **Semi-Markov Processes**)

DAVID WILLIAMS

Martini, Paul

Born: January 25, 1889, in Frankenthal, Germany.
Died: September 8, 1964, near Bonn, Germany.

Paul Martini was born in Frankenthal (Palatinate) in the southwestern region of Germany. He studied medicine at the universities of Munich and Kiel. He worked on his thesis in the Institute of Physiology in Munich and obtained his doctorate in medicine in 1917.

He was an assistant in the II. Medizinische Universitätsklinik in Munich which was at that time headed by one of the famous German internists, Friedrich von Mueller. In 1926 he become "extraordinary" professor of medicine. He left the university when he was appointed head physician in the St Hedwigskrankenhaus in Berlin, a large community hospital. In 1932 Martini returned to university life. From that time until he retired he held the chair of internal medicine at the University of Bonn and was director of the Universitätsklinik für Innere Medizin und Nervenkrankheiten.

Among his scientific work, his 1932 Monograph *Methodenlehre der therapeutischen Forschung* [1] is of particular importance. This book contains all the elements of the controlled **clinical trial** and is the first in modern times addressing the problem of a scientific methodology as regards therapeutic research. It is evident that the use of placebo (*see* **Blinding or Masking**) is meant even when this word is not used. Martini wrote: "the medicines have to be given to the patient in a form or in a galenic preparation that their special character or their purpose can not be recognized, they have to be camouflaged. The results have to be evaluated by means of statistics and probability calculus". **Randomization** is not clearly addressed, but subsequent publications make it likely that alternating procedures, for example based on day of birth, were used.

In an article published in 1934, Martini again explained his methodology of therapeutic trials and defended himself against the arguments that had been raised against his ideas [2].

In 1957 he published in the Deutsche Medizinische Wochenschrift his ideas about double-blind trials [3]. He rejected the method of double-blindness as he was not convinced that the results of such trials were superior to the single-blind trials. In 1961 Martini was chairman of an international seminar in the field of drug trials in Berlin.

Martini was a highly esteemed physician and had among his patients many personalities of the political scene in Bonn. During the time of national socialism he was able to avoid involvement. In 1948 he was president of the first postwar "Internistenkongress" (annual meeting of the German Society of Internal Medicine) in Wiesbaden. In 1964 he died in his country house in the Eifel near Bonn.

References

[1] Martini, P. (1932). *Methodenlehre der therapeutischen Forschung*. Julius Springer, Berlin. (Second edition published in 1947 under the title *Methodenlehre der therapeutisch klinischen Forschung*.).

[2] Martini, P. (1934). Rationelle Medizin (Rational medicine), *Muenchner Medizinische Wochenschrift* **81**, 1411–1416.

[3] Martini, P. (1957). Die unwissentliche Versuchsanordnung und der sogenannte doppelte Blindversuch (The unknown experimental design and the so-called double-blind trial), *Deutsche Medizinische Wochenschrift* **82**, 597–602.

(*See also* **Clinical Trials, History of**)

H.J. DENGLER

Matched Analysis

On grounds of both validity and efficiency, the appropriate analysis of data involving category matching mandates the use of stratified analysis methods based on the strata used in the matching process [6] (*see* **Stratification**). Two important methods for analyzing category matched (or, more generally, stratified) data are the **Mantel–Haenszel** procedure [9] and **conditional logistic regression** (see [1], Chapter 7, and [5], Chapter 20).

The Mantel–Haenszel (MH) test statistic [9] is the most widely used and recommended method for testing for overall association in a stratified analysis. And, as we will see, the MH test statistic for stratified data analysis is based on the (central) **hypergeometric distribution**. For dichotomous disease and exposure variables (the setting for this presentation), the MH testing procedure involves a one **degree-of-freedom** (continuity-corrected) **chi-squared statistic** of the general form

$$\chi^2_{\mathrm{MH}} = \frac{[|A - \mathrm{E}_0(A)| - 1/2]^2}{\mathrm{var}_0(A)}, \qquad (1)$$

where A is the random variable denoting the total number (over all strata) of diseased subjects in each stratum who are exposed (i.e. the total number of "exposed cases"), $\mathrm{E}_0(A)$ is the expected total number of exposed cases under the **null hypothesis** of no **association** between exposure and disease, and $\mathrm{var}_0(A)$ is the **variance** of the total number of exposed cases under the same null hypothesis.

Suppose that there are G strata defined by the matching process, with the gth stratum having the structure given in Table 1.

The four marginal frequencies n_{1g}, n_{0g}, m_{1g}, and m_{0g} in the gth stratum convey no information about the strength of the association between exposure and disease in that stratum, but rather indicate only the "amount of information" in that stratum.

Table 1 Data layout for the gth stratum ($g = 1, 2, \ldots, G$)

	E	$\overline{E}$	
D	A_g	B_g	m_{1g}
$\overline{D}$	C_g	D_g	m_{0g}
	n_{1g}	n_{0g}	n_g

Consequently, the four marginal frequencies within each stratum may be assumed (with no compromise to validity) to be "fixed" for analysis purposes, even though the sampling scheme actually used may not have imposed such constraints on the margins of these G **2 × 2 tables**.

Conditional on these fixed margins for all strata, it is sufficient to focus entirely on the "A_g cell", namely, the number of exposed cases in the gth stratum, $g = 1, 2, \ldots, G$. The test statistic (1) is then a conditional test since properties of the **random variable** $A = \sum_{g=1}^{G} A_g$ are based on the condition that the four margins in each stratum are fixed. More specifically, assuming fixed margins and no exposure-disease association, A_g is a (central) hypergeometric random variable, so that

$$\mathrm{E}_0(A) = \sum_{g=1}^{G} \frac{(n_{1g}m_{1g})}{n_g} \quad \text{and}$$

$$\mathrm{var}_0(A) = \sum_{g=1}^{G} \frac{(n_{1g}n_{0g}m_{1g}m_{0g})}{(n_g - 1)n_g^2}.$$

Finally, some algebra can be used to write expression (1) in the form

$$\chi^2_{\mathrm{MH}} = \frac{\left[\left|\sum_{g=1}^{G}(A_g D_g - B_g C_g)/n_g\right| - 1/2\right]^2}{\sum_{g=1}^{G}(n_{1g}n_{0g}m_{1g}m_{0g})/(n_g - 1)n_g^2}; \qquad (2)$$

under the null hypothesis of no exposure–disease association, it can be shown that the test statistic (2) has, for "large samples", an approximate chi-square distribution with 1 df.

It is very important to stress that the "large samples" assumption for the test statistic (2) pertains to the pooled information over all G strata, rather than to stratum-specific numbers. Consequently, in the use of the Mantel–Haenszel test statistic (2), it is permissible to have relatively small numbers in each stratum as long as the total number of subjects on the margins over all strata is sufficiently large. Without going into detail, this form of **robustness** to sparse stratum-specific data accrues due to the assumption of fixed stratum-specific margins, an assumption that maintains validity at only a slight cost in efficiency.

Specific criteria for appropriate sample sizes to maintain the validity of the chi-squared approximation for (2) have been proposed by Mantel & Fleiss [8]. They recommend using (2) provided that the quantities

$$\mathrm{E}_0(A) - \left[\sum_{g=1}^{G} \max(0, m_{1g} - n_{0g})\right] \quad \text{and}$$

$$\left[\sum_{g=1}^{G} \min(n_{1g}, m_{1g})\right] - \mathrm{E}_0(A)$$

both exceed 5 in value.

It is important to mention that the use of the Mantel–Haenszel test statistic (2) should be avoided when there is evidence of strong **effect modification** in the data, as would be reflected by widely varying stratum-specific estimated **odds ratios** $\widehat{\mathrm{OR}}_g = A_g D_g / B_g C_g$, $g = 1, 2, \ldots, G$. Because of the structure of the numerator in (2), the value of (2) can be very small (suggesting no exposure–disease association) when, in fact, some stratum-specific estimated odds ratios are significantly greater than 1 and some are significantly less than 1. Indeed, claims of optimal statistical properties for the Mantel–Haenszel test [11] are valid only in the situation where stratum-specific population odds ratios all have the same value. Tests for lack of uniformity of stratum-specific odds ratios are discussed in Chapter 4 of [1].

Given the assumption of a common population odds ratio for all strata, it makes sense to compute a summary estimator of this common odds ratio; such an estimator is typically a weighted average of the G stratum-specific estimated odds ratios. Mantel & Haenszel [9] proposed several such summary estimators for use in **case–control studies**. The most notable of these is the $\widehat{\mathrm{mOR}}$, which is defined as

$$\widehat{\mathrm{mOR}} = \left[\sum_{g=1}^{G} (A_g D_g)/n_g\right] \Bigg/ \left[\sum_{g=1}^{G} (B_g C_g / n_g)\right]$$

$$= \sum_{g=1}^{G} W_g(\widehat{\mathrm{OR}}_g) \Bigg/ \sum_{g=1}^{G} W_g, \qquad (3)$$

where $W_g = B_g C_g / n_g$. An interesting property of $\widehat{\mathrm{mOR}}$ is that it equals unity only when expression (2) is zero, a property that is not shared by other summary estimators (see [5], Chapter 17). Another advantage of the $\widehat{\mathrm{mOR}}$ over other summary estimators is that it can be used without alteration when there are zero frequencies within the body of some of the stratum-specific tables.

For certain types of matched data, expressions (2) and (3) have simple structures. As one example, for a matched pairs case–control study where each stratum (or pair) consists of one case and one control, then expression (2) reduces to $(b - c)^2/(b + c)$ apart from continuity correction, and expression (3) equals b/c, where b is the number of strata where the case is exposed and the control is not, and c is the number of strata where the control is exposed and the case is not. For the special case of R-to-1 matching, see either Chapter 5 in [1] or Chapter 18 in [5]. For **confidence interval** methods based on (3) in case–control studies with multiple matching, see [3], [12], and [13]. Finally, some generalizations of the Mantel–Haenszel test have been developed for situations where the exposure variable is **nominal** with several categories [9] and where the exposure variable is ordinal (*see* **Measurement Scale**) in nature [2, 7].

A more general and flexible method for the analysis of matched (or, in general, stratified) data is conditional logistic regression (*see* **Logistic Regression, Conditional**). This multivariable modeling procedure is specifically designed to be used when there are small stratum-specific sample sizes. Hence, it is ideally suited for the analysis of matched study designs or to similar situations involving very fine stratification; in fact, its use in these situations is mandatory to avoid **biased** estimates of important odds ratio parameters. In contrast to stratified data analysis methods, conditional logistic regression methods do not require all variables to be categorized; for example, continuous exposure, **confounding**, and effect-modifying variables can be treated as such. In addition, it is theoretically possible to consider simultaneously in one model several exposure variables and to examine potential confounding and effect modification effects due to **covariates** not involved in the matching process.

Suppose we consider the case–control format, with $\mathbf{x}_{1g}, \mathbf{x}_{2g}, \ldots, \mathbf{x}_{m_g g}$ denoting the observed data vectors for the total of $m_g = (m_{1g} + m_{0g})$ cases and controls in the gth stratum, $g = 1, 2, \ldots, G$. Without loss of generality, we arrange these data vectors so that the first m_{1g} vectors belong to the m_{1g} cases in the gth stratum. For a dichotomous response variable

D with $D = 1$ signifying a case and $D = 0$ signifying a control, consider fitting by conditional logistic regression the logistic model

$$\text{logit}[\Pr(D_{lg} = 1)] = \alpha_g + \boldsymbol{\beta}'\mathbf{x}_{lg},$$
$$l = 1, 2, \ldots, m_g \quad \text{and} \quad g = 1, 2, \ldots, G.$$

Then, the contribution from the gth stratum to the full conditional **likelihood** has the structure

$$\text{CL}_g = \prod_{l=1}^{m_{1g}} \exp(\boldsymbol{\beta}'\mathbf{x}_{lg}) \Bigg/ \sum_u \left[\prod_{l=1}^{m_{1g}} \exp(\boldsymbol{\beta}'\mathbf{x}_{ulg}) \right], \tag{4}$$

where the sum $\sum_u$ in the denominator is over all partitions of the set of integers $\{1, 2, \ldots, m_g\}$ into two subsets, the first of which contains m_{1g} elements; there are $m_g!/m_{1g}!m_{0g}!$ such partitions. Thus, CL_g is the conditional probability that the first m_{1g} of the m_g data vectors $\mathbf{x}_{1g}, \mathbf{x}_{2g}, \ldots, \mathbf{x}_{m_g g}$ go with the cases (as they actually do) considering all possible arrangements of these m_g data vectors; in other words, CL_g is the conditional probability of the observed data. The full conditional likelihood CL is then equal to $\text{CL} = \prod_{g=1}^{G} \text{CL}_g$, and standard **maximum likelihood methods** can be used to estimate and to make **inferences** about the elements of $\boldsymbol{\beta}$.

It is important to note that the conditional likelihood CL based on (4) depends only on $\boldsymbol{\beta}$, the parameter vector of interest. The **nuisance parameters** $\alpha_1, \alpha_2, \ldots, \alpha_G$ indexing the matching strata have been eliminated via this permutation procedure, thus precluding the need to estimate unnecessarily an often large number of parameters that provide no information about important exposure–disease odds ratio parameters of interest. In addition, precisely the same likelihood CL is obtained regardless of whether we consider the data to have arisen from a follow-up study or from a case–control study. Also, CL has precisely the structure of Cox's **partial likelihood** [4], based on the **proportional hazards model**, for analyzing follow-up study data. However, an important distinction is that each stratum-specific set in the denominator of CL, instead of involving *all* persons in the study who are disease-free at the time each incident case is identified, consists only of the m_{0g} controls specifically associated with (e.g. sampled at the same time as) the m_{1g} cases.

As an illustration of the conditional likelihood approach for matched data, consider a matched case–control study involving G cases, where the gth case is individually matched to R_g controls on one or more variables. Then, $m_{1g} = 1, m_{0g} = R_g, m_g = (R_g + 1)$, and the conditional likelihood CL takes the specific form

$$\prod_{g=1}^{G} \left[1 + \sum_{l=2}^{R_g+1} \exp[\boldsymbol{\beta}'(\mathbf{x}_{lg} - \mathbf{x}_{1g})] \right]^{-1}. \tag{5}$$

Given the structure of this expression, if any of the elements of $\mathbf{x}$ are matching variables, taking the same value for each member of a matched set, then their contribution to the likelihood is zero and the corresponding elements of $\boldsymbol{\beta}$ cannot be estimated. However, by incorporating such matching variables in the model as **interaction** terms with exposure factors, one can model the variation in odds ratios across matched sets.

Finally, to appreciate that these conditional likelihood methods do, in fact, yield recognizable results in well-known special cases, consider the simple matched pairs case–control study considered earlier, where $R_g = 1$ for all g and where there is a single dichotomous exposure variable. With $e^{\boldsymbol{\beta}} = \text{EOR}$, the exposure odds ratio parameter, it can be shown that (5) is proportional to

$$\left[\frac{\text{EOR}}{(1 + \text{EOR})} \right]^b \left[\frac{1}{(1 + \text{EOR})} \right]^c.$$

By differentiating the logarithm of the above expression with respect to EOR, equating it to zero, and solving, one finds that the conditional maximum likelihood estimator of EOR is $\widehat{\text{mOR}} = b/c$, the ratio of discordant pairs. In contrast, the unconditional maximum likelihood estimator of EOR is $(b/c)^2$, which dramatically illustrates the potential bias associated with the use of unconditional likelihood methods for finely stratified data. While not as extreme as illustrated here, the bias of unconditional likelihood methods is found in many other sparse data situations [10]. These findings emphasize the need to consider the use of conditional likelihood methods when fitting logistic models involving many strata and/or other nuisance parameters to data sets of limited size.

References

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1: *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

[2] Clayton, D.G. (1974). Some odds ratio statistics for the analysis of ordered categorical data, *Biometrika* **61**, 525–531.

[3] Connett, J., Ejigou, A., McHugh, R. & Breslow, N.E. (1982). The precision of the Mantel-Haenszel odds ratio estimator in case-control studies with multiple matching, *American Journal of Epidemiology* **116**, 875–877.

[4] Cox, D.R. (1975). *The Analysis of Binary Data*. Methuen, London.

[5] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont.

[6] Kupper, L.L., Karon, J.M., Kleinbaum, D.G., Morgenstern, H. & Lewis, D.K. (1981). Matching in epidemiologic studies: Validity and efficiency considerations, *Biometrics* **37**, 293–302.

[7] Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure, *Journal of the American Statistical Association* **58**, 690–700.

[8] Mantel, N. & Fleiss, J.L. (1980). Minimum expected cell size requirements for the Mantel-Haenszel one-degree of freedom chi-square test and a related rapid procedure, *American Journal of Epidemiology* **112**, 129–134.

[9] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[10] Pike, M.C., Hill, A.P. & Smith, P.G. (1980). Bias and efficiency in logistic analysis of stratified case-control studies, *International Journal of Epidemiology* **9**, 89–95.

[11] Radhakrishna, S. (1965). Combination of results from several 2×2 contingency tables, *Biometrics* **21**, 86–98.

[12] Robins, J., Breslow, N. & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models, *Biometrics* **42**, 311–323.

[13] Sato, T. (1990). Confidence limits for the common odds ratio based on the asymptotic distribution of the Mantel-Haenszel estimator, *Biometrics* **46**, 71–80.

(*See also* **Confounder**; **Confounder Summary Score**; **Matching**)

LAWRENCE L. KUPPER

Matched Pairs With Categorical Data

Matched pairs with categorical data arise when two measurements of the same categorical variable are obtained from each independent experimental unit. The repeated measurements might be obtained at two time points, for example, if a patient's condition is categorized as "good" or "poor" at diagnosis and then again six months after diagnosis. In other applications, the variable of interest might be measured under two different conditions. As an example, a patient's response to treatment, categorized as satisfactory or unsatisfactory, might be evaluated following treatment with the standard therapy and then again following treatment with a new therapy. The repeated measurements could also be obtained from each member of a matched set. In a matched **case–control study**, for example, each independent experimental unit consists of a case (individual with a specified disease or condition) and a control (individual without the disease or condition) individually matched to the case by factors such as age, sex, residence, employer, etc.

Such data can be displayed in a two-way **contingency table**. Table 1 shows the general layout when two measurements of a categorical response with I categories are obtained from each experimental unit. In this table, n_{ij} is the observed frequency in the ith row and jth column of the table, n_{i+} and n_{+j} denote the row and column marginal frequencies, respectively, and n is the total number of independent experimental units. The data layout displayed in Table 1 is one example of a **square contingency table**. The possible types of response variables include **polytomous data**, **ordered categorical data**, and, for the special case of $I = 2$, **binary data**.

Table 1 Two-way contingency table for matched pairs with categorical data

First measurement of response	Second measurement of response					
	1	...	j	...	I	Total
1	n_{11}	...	n_{1j}	...	n_{1I}	n_{1+}
⋮	⋮	⋱	⋮	⋱	⋮	⋮
i	n_{i1}	...	n_{ij}	...	n_{iI}	n_{i+}
⋮	⋮	⋱	⋮	⋱	⋮	⋮
I	n_{I1}	...	n_{Ij}	...	n_{II}	n_{I+}
Total	n_{+1}	...	n_{+j}	...	n_{+I}	n

Statistical Inference

Statistical inference for matched pairs with categorical data focuses generally on comparing the marginal distributions of the two correlated responses (*see* **Marginal Models**). Let π_{ij} denote the probability of being in the ith row and jth column, for $i, j = 1, \ldots, I$, and let π_{i+} and π_{+j} denote the corresponding row and column marginal probabilities. The hypothesis of marginal homogeneity is

$$\pi_{i+} = \pi_{+i}, \quad i = 1, \ldots, I.$$

The hypothesis of symmetry,

$$\pi_{ij} = \pi_{ji}, \quad i \neq j,$$

is also sometimes of interest. Other hypotheses include **quasi-symmetry** and **quasi-independence**.

Binary Response

When $I = 2$, the hypothesis of symmetry implies marginal homogeneity, and vice versa. In this situation marginal homogeneity is assessed using the **McNemar test**. This test is a special case of the general class of **Mantel–Haenszel methods**. If the sample size n is small, **exact tests for categorical data**, specifically, the exact one-sample test for the success probability of the **binomial distribution**, should be used.

*Polytomous Response (see **Polytomous Data**)*

Let $d_i = n_{i+} - n_{+i}$ and let $\mathbf{d}' = (d_1, \ldots, d_{I-1})$. Stuart [25] proposed a test of marginal homogeneity using the statistic

$$W_0 = \mathbf{d}' \mathbf{V}_0^{-1} \mathbf{d},$$

where $\mathbf{V}_0$, the sample **covariance matrix** under the null hypothesis of marginal homogeneity, has diagonal elements $n_{i+} + n_{+i} - 2n_{ii}$ and offdiagonal elements $-(n_{ij} + n_{ji})$. The asymptotic null distribution of W_0 is χ^2 with $I - 1\,\mathrm{df}\,(\chi^2_{I-1})$. For

2 × 2 tables, the test based on W_0 is identical to McNemar's test. For $I \times I$ tables, Stuart's statistic is the $I - 1$ df general association statistic from the class of Mantel–Haenszel methods.

Bhapkar [4] considered the statistic $W = \mathbf{d}'\mathbf{V}^{-1}\mathbf{d}$, where $\mathbf{V}_0$ is replaced with the unrestricted sample covariance matrix estimator $\mathbf{V}$, which has diagonal elements $n_{i+} + n_{+i} - 2n_{ii} - (n_{i+} - n_{+i})^2$ and off-diagonal elements $-(n_{ij} + n_{ji}) - (n_{i+} - n_{+i})(n_{j+} - n_{+j})$. The statistic W is asymptotically optimal, as shown by Wald [26], and can be computed using weighted **least squares** methodology for the analysis of categorical data [10] (*see* **Categorical Data Analysis**). Ireland et al. [11] noted that $W = W_0/(1 - W_0/n)$.

Although maximum likelihood estimators of the cell probabilities under the hypothesis of marginal homogeneity cannot be expressed in closed form, likelihood methods can also be used to test marginal homogeneity. Madansky [19] gave the generalized maximum **likelihood ratio test** comparing the likelihood maximized under the hypothesis of marginal homogeneity to the likelihood maximized in the unrestricted case. The likelihood ratio test is also presented by Plackett [22, pp. 79–80], who uses the approach of Wedderburn [27]. Firth & Treat [8] and Lipsitz [16] describe how to conduct this test using standard statistical software.

An alternative likelihood-based approach tests marginal homogeneity in the context of the model for quasi symmetry by comparing the maximized likelihoods for the symmetry and quasi-symmetry models (see, for example, Agresti [2, pp. 358–359]). While this test can be carried out using standard software for fitting a **loglinear model**, it is conditional on the model of quasi-symmetry holding.

For polytomous responses ($I > 2$), the hypotheses of marginal homogeneity and symmetry are not equivalent. Under the null hypothesis of symmetry, the expected count in the (i, j) cell, with $i \neq j$, is estimated by $(n_{ij} + n_{ji})/2$. Substituting the estimated expected cell counts into the usual formula for the Pearson χ^2 test, Bowker [5] derived the statistic

$$X^2 = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}.$$

Under the null hypothesis of symmetry, X^2 is approximately $\chi^2_{I(I-1)/2}$. When $I = 2$, this test is identical to McNemar's test.

A likelihood-ratio test can also be used. The test statistic is

$$G^2 = 2 \sum_{i \neq j} n_{ij} \log \left(\frac{2n_{ij}}{n_{ij} + n_{ji}} \right).$$

Since the hypothesis of symmetry has a loglinear model representation, G^2 can be computed using standard software for fitting loglinear models.

*Ordered Categorical Response (see **Ordered Categorical Data**)*

The tests of marginal homogeneity described in the previous section use $I - 1$ df to compare the I pairs of marginal proportions. For ordered categorical variables, alternative tests that use the additional information provided in the ordering of the categories are more powerful for certain types of departures from the null hypothesis.

One approach is to compare marginal mean scores instead of marginal distributions. Given a set of scores that are appropriately assigned to the categories according to the alternative one wishes to detect, 1 df tests of marginal homogeneity analogous to the Stuart and Bhapkar tests can be carried out. The corresponding Stuart-type statistic (using the null covariance matrix estimator $\mathbf{V}_0$) is the Mantel–Haenszel mean score statistic; this test is discussed in White et al. [28]. If marginal rank scores are assigned, the statistic is equivalent to the Friedman [9] test obtained from a two-way rank analysis of variance with subjects as blocks (*see* **Ranks**). Agresti [1, Section 2.3] discusses the corresponding Bhapkar-type statistic (using the unrestricted covariance matrix estimator $\mathbf{V}$) and gives additional references; this test can be computed using weighted least squares methodology for the analysis of categorical data (*see* **Categorical Data Analysis**).

Another approach to testing marginal homogeneity for ordered classifications is based on the conditional symmetry model [20; 2, pp. 361–364]. This loglinear model has only one more parameter than the symmetry model. When conditional symmetry holds, a 1 df chi-square statistic for testing marginal homogeneity is the difference between the likelihood-ratio lack-of-fit statistics for the symmetry and conditional symmetry models. Agresti [1] mentions additional methods useful in the analysis of ordered categorical responses.

Example

Table 2 displays the cross-classification of right eye and left eye unaided distant vision grade in 7477 women employees, aged 30–39 years, in British Royal Ordnance factories during 1943–1946. The outcome variable of interest is an ordered categorical variable with four levels. These data, first quoted by Stuart [24], have been analyzed by numerous authors.

First, treating the categories as nominal rather than ordered, the tests of marginal homogeneity give χ^2 statistics of 11.957, 11.976, and 11.986 for Stuart's W_0, Bhapkar's W, and Madansky's likelihood-ratio statistic, respectively. With respect to the χ^2_3 null distribution, all are significant at $\alpha = 0.01$. The alternative likelihood-based approach of testing marginal homogeneity in the context of the quasi-symmetry model gives a likelihood-ratio statistic of $19.250 - 7.274 = 11.976$, also with 3 df. The values 19.250 and 7.274 are the likelihood-ratio statistics for symmetry (6 df) and quasi-symmetry (3 df), respectively. Bowker's test for symmetry gives $X^2 = 19.107$, which agrees closely with the corresponding value from the likelihood-ratio test.

If one treats the categories as ordered, the use of equally-spaced scores for the levels gives χ^2 statistics of 11.947 (Mantel–Haenszel mean score) and 11.97 (weighted least squares). The Friedman statistic (Mantel–Haenszel mean score test using rank scores) is 11.885. With respect to their asymptotic χ^2_1 null distributions, all three of these criteria are significant at $\alpha = 0.001$.

The likelihood-ratio lack-of-fit statistic from the conditional symmetry model is 7.35 with 5 df. Since this model provides a satisfactory fit to the data ($P = 0.2$), the difference between the likelihood-ratio statistics from the symmetry and conditional symmetry models also tests marginal homogeneity. The value of the statistic is $19.250 - 7.35 = 11.9$, which is also significant at $\alpha = 0.001$.

Table 2 Right eye and left eye unaided distance vision of 7477 women

Grade of right eye	Grade of left eye				
	Highest	Second	Third	Lowest	Total
Highest	1520	266	124	66	1976
Second	234	1512	432	78	2256
Third	117	362	1772	205	2456
Lowest	36	82	179	492	789
Total	1907	2222	2507	841	7477

Related Topics

The above methods are useful in analyzing matched pairs from a single population. In some situations there may be multiple populations defined by additional covariates of interest. If all covariates are categorical with a sufficiently large sample size in each covariate strata, a wide variety of types of regression models can be fitted using the general weighted least squares approach [10], which is described specifically for correlated responses by Koch et al. [13] and others. This methodology is applicable for binary, polytomous, and ordered categorical variables. A major shortcoming, however, is that this method fails if there are continuous covariates and/or small stratum-specific sample sizes.

Maximum likelihood regression models for matched pairs with binary data are discussed by Cox & Snell [7], Lipsitz et al. [18], and other authors. Breslow & Day [6] focus specifically on the use of conditional logistic regression (*see* **Logistic Regression, Conditional**), in the analysis of matched case–control studies. The **generalized estimating equations** (GEE) procedure of Liang & Zeger [14] and its extensions can also be used to analyze matched binary outcomes with covariates. Generalizations of the GEE methodology to polytomous and ordered categorical outcomes have also been studied [3, 12, 15, 17, 21, 23].

References

[1] Agresti, A. (1983). Testing marginal homogeneity for ordinal categorical variables, *Biometrics* **39**, 505–510.

[2] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[3] Agresti, A., Lipsitz, S. & Lang, J.B. (1992). Comparing marginal distributions of large, sparse contingency tables, *Computational Statistics and Data Analysis* **14**, 55–73.

[4] Bhapkar, V.P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data, *Journal of the American Statistical Association* **61**, 228–235.

[5] Bowker, A.H. (1948). A test for symmetry in contingency tables, *Journal of the American Statistical Association* **43**, 572–574.

[6] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. 1: The Analysis of Case-Control, Studies. International Agency for Research on Cancer, Lyon.

[7] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*. Chapman & Hall, London.

[8] Firth, D. & Treat, B.R. (1988). Square contingency tables and GLIM, *GLIM Newsletter* **16**, 16–20.

[9] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.

[10] Grizzle, J.E., Starmer, C.F. & Koch, G.G. (1969). Analysis of categorical data by linear models, *Biometrics* **25**, 489–504.

[11] Ireland, C.T., Ku, H.H. & Kullback, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table, *Journal of the American Statistical Association* **64**, 1323–1341.

[12] Kenward, M.G. & Jones, B. (1992). Alternative approaches to the analysis of binary and categorical repeated measurements, *Journal of Biopharmaceutical Statistics* **2**, 137–170.

[13] Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. & Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data, *Biometrics* **33**, 133–158.

[14] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[15] Liang, K.Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[16] Lipsitz, S.R. (1988). Methods for analyzing repeated categorical outcomes, Unpublished PhD Dissertation. Department of Biostatistics, Harvard University.

[17] Lipsitz, S.R., Kim, K. & Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations, *Statistics in Medicine* **13**, 1149–1163.

[18] Lipsitz, S.R., Laird, N.M. & Harrington, D.P. (1990). Maximum likelihood regression methods for paired binary data, *Statistics in Medicine* **9**, 1517–1525.

[19] Madansky, A. (1963). Tests of homogeneity for correlated samples, *Journal of the American Statistical Association* **58**, 97–119.

[20] McCullagh, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories, *Biometrika* **65**, 413–418.

[21] Miller, M.E., Davis, C.S. & Landis, J.R. (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares, *Biometrics* **49**, 1033–1044.

[22] Plackett, R.L. (1981). *The Analysis of Categorical Data*. Griffin, London.

[23] Stram, D.O., Wei, L.J. & Ware, J.H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates, *Journal of the American Statistical Association* **83**, 631–637.

[24] Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables, *Biometrika* **40**, 105–110.

[25] Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification, *Biometrika* **42**, 412–416.

[26] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society* **54**, 426–482.

[27] Wedderburn, R.W.M. (1974). Generalized linear models specified in terms of constraints, *Journal of the Royal Statistical Society, Series B* **36**, 449–454.

[28] White, A.A., Landis, J.R. & Cooper, M.M. (1982). A note on the equivalence of several marginal homogeneity test criteria for categorical data, *International Statistical Review* **50**, 27–34.

CHARLES S. DAVIS

Matching, Probabilistic

Record linkage is usually concerned with establishing which records relate to the same individual. In the widest sense, record linkage is virtually universal in organizations that process information about people, and is usually achieved via some form of personal identifier, such as a case reference number in a hospital or a national insurance number in a state benefit system. When such unique personal identifiers are not available or feasible, record linkage must make the best possible use of the personal identifiers which are present, such as name, date of birth, area of residence, and so on. Probability matching is the key technique to have been developed to maximize the accuracy of linkage decisions based on the level of agreement and disagreement between the identifiers on different records [9].

Whenever such personal identifiers as name or date of birth are recorded or are transcribed other than electronically, it is possible that different information will be entered on different records relating to the same person (*see* **Data Management and Coordination**). Such discrepancies can arise for a wide range of reasons. Some discrepancies involve error or uncertainty: a name or date may be misheard, the person involved may not be clear about an item of identification, different names may be used in different contexts (formal versus informal), or data may be misread or miskeyed during transfer from one record to another. Other discrepancies reflect changes in circumstances such as a change of name (especially when women get married) or a change of residence.

Whatever the reason for such discrepancies, their presence means that a reliance upon exact correspondence of identifiers to establish that records belong to the same individual will usually lead to a large number of links being missed. For example, in the centrally held linked data set of hospital discharge records in Scotland, there is a discrepancy rate of an order of magnitude of 2%–3% for each of the identifiers first initial; surname; and day, month, and year of birth. Thus an insistence that all five of these identifiers matched exactly would involve losing 10–15% of true matches. Probability matching allows us to link records despite the existence of such discrepancies in identifying information. In Scotland, relatively straightforward methods of probability matching have produced an accuracy of around 99% on such data [6].

Probability matching translates the level of agreement and disagreement between each item of identifying information on two records – for example, both have first initial "J" or there is a discrepancy of one day in the day of birth – into a mathematical score or probability weight, which can be aggregated across all items to produce a relative probability that the two records belong to the same individual. Put more simply, when two items of identifying information are the same on two records, this increases the probability that they belong to the same person. When they are not the same, this decreases the probability. Probability matching puts these almost tautologic observations into systematic form and quantifies their implications [13].

Thus a very simple insight lies at the heart of probability matching. The skill and complexity of the technique lies in adapting the application of this insight to the precise characteristics of the records to be linked. Success in probability matching comes from staying close to the data and being guided by the emergent properties of each linkage.

This is very much the philosophy of Howard Newcombe, who is the founding father of probability matching, having first developed the technique in Canada in the 1950s [2, 14], and who has been involved in its progress ever since. The torch was taken up by the Oxford Record Linkage System in the 1960s [1] and by Scotland from the 1970s onwards [5, 6]. In the past ten years there has been a burgeoning of interest and a wider spread of probability matching techniques.

The statistical formalization of the theory of probability matching has tended to follow in the wake of its practical development. Thus Fellegi & Sunter [3] formalized the theory underlying Newcombe's early work. There is some debate about the importance of statistical formalization for the progress of probability matching [12]. This account, while accepting the importance of statistical formalization in confirming the validity of the technique, follows Newcombe in stressing the empirical and pragmatic aspects. We are primarily interested in the practical application of the technique to achieve the most accurate and efficient results.

The assumption that record linkage is about linking records belonging to the same individuals is

purely for clarity of exposition. The field of application is much wider of course; including, for example, the linkage of mothers to babies or linkages between the records of different members of families in general.

The Elements of Record Linkage Using Probability Matching

Record linkage using probability matching can be seen as involving three elements or stages. The first involves bringing pairs of records together to be compared. The second involves calculating the probability weight for each pair of records. The third involves making the linkage decision based upon the probability weights. Most attention has tended to focus on the second of these phases, the calculation of probability weights. The other elements are just as important and may well be where improvements will be concentrated in the future.

Bringing Records Together

It is normally not feasible to calculate probability weights for all the pairs of records in the data to be linked. For example, the one million Scottish hospital discharge records for a year contain approximately 500 000 000 000 pairs of records. The computing resources that are usually available would not permit probability weights for all of these pairs to be calculated.

We must restrict the number of pairs of records to be compared. This has traditionally been done by using some form of **blocking**, whereby only those pairs of records that share at least one common set of identifying items are compared [4]. By doing this, we run the risk that records which do belong to the same person will not be compared. The trick is to achieve the necessary reduction in pair comparisons while minimizing the number of links missed because the necessary comparisons were not carried out. For example, a common set of blocking criteria is as follows: compare only those records that share the same compressed surname (see below) and first initial, or that share the same date of birth. Pairs of records that differ both in terms of surname or initial and date of birth would not be compared. In a UK context, well under 0.5% of true links would generally be missed by such a blocking configuration.

Traditionally such blocking has been achieved by sorting the files involved to bring together records sharing the specified identifiers. Thus in terms of the blocking criteria outlined above, the files would be sorted by compressed surname and first initial. All records having these identifiers would be compared. Then the files would be sorted by date of birth. All records containing the same date of birth would be compared. Usually, at least one further sort is required to reconcile matches made in the two separate passes through the data.

When data volumes become large, this method has the disadvantage that no matter how few records are being added to an existing linked file, all the records involved must be sorted several times. Thus for example, the main Scottish linked data set contains 14 million linked records. Linking in an additional month's data or even an external data set such as a survey of 10 000 individuals would require sorting the entire file several times. This is extremely time consuming and is not feasible on a routine, frequent basis.

In Scotland, this problem has been solved by indexing incoming or "newcomer" records in memory [7]. The record numbers of all newcomer records sharing the same day and month of birth, for example, are stored in the same row of an array indexed by month and day of birth. The existing or catalog file is read through sequentially. Each catalog record can thus be directed for comparison to the newcomer records sharing its month and day of birth. The same principle can be applied to any numeric element of an identifier, such as the numeric element of a compressed surname. Thus the blocking usually achieved by sorting is mimicked in memory. Each of the newcomer records retains information about the catalog record with which it has achieved the highest probability weight. The major advantage of this technique is that the larger of the two files does not need to be sorted at all. A methodologic implication of the method is that each newcomer record is allowed to link to only one record already in the linked file. This may have advantages (see below).

The development of techniques of rapid direct access by any kind of key (not just numeric) inherent in the development of **database** management software should allow the generalization of such methods of selective comparison. Much more flexibility will be possible in the definition of the subset of pairs of records for which probability weights are to be calculated.

Calculating the Probability Weights

This is the aspect of record linkage which has been best documented, especially by Newcombe [9]. The key to the calculation of probability weights is the **odds ratio**. An odds ratio is calculated for every outcome of the comparison between two identifiers; for example, both records have first initial "J" or the day of birth is one day different. The odds ratio is simply the ratio between the frequency of a given outcome in pairs of records that relate to the same person and the frequency of a given outcome in pairs of records that do not belong to the same person. The odds ratio expresses mathematically how much a given outcome increases or decreases the probability that two records belong to the same person.

Let us take as an example, agreement of month of birth (e.g. both records have month of birth June).

The top line of the odds ratio is the frequency of agreement of month of birth in two records belonging to the same person. This depends upon how much miscoding takes place. Let us say that there is a miscoding of month of birth in one or another record for 3% of the time. Thus the top line of the odds ratio is 97%, meaning that for 97% of the time there is random agreement of month of birth in two records belonging to the same person.

What is the frequency of agreement of month of birth among records not belonging to the same person? This is broadly equivalent to asking how often is there random agreement of month of birth. There are 12 months in the year, so there will random agreement 1/12, or 8.3%, of the time.

The odds ratio for agreement of month of birth is thus 97% divided by 8.3%, or roughly 11.6. Agreement of month of birth thus increases the probability that two records belong to the same person 11.6 times.

This accords with common sense. The more uncommon an identifier is, the greater will be the odds ratio given by agreement. Agreement of the less common first initial "Z" will give a much higher odds ratio than agreement of the more common initial "J". The same principle can be applied to disagreement and levels of disagreement.

The odds ratios from different identifiers can be combined by multiplication to give the overall odds ratio derived from the agreement and disagreement of all identifiers to be compared for two records. Because of the clumsiness involved in multiplying odds, they are usually converted to logarithms to base 2 or binit weights, which can then be added and subtracted.

The most important principle in combining weights for different identifiers is the assumption of independence. Weights must only be given for independent sources of information. It would be illegitimate, for example, to give full weights both for area of residence and general practitioner, since these are highly correlated. Weights can, however, be made conditional; for example, different weights for the initials of men and women [9].

How do we obtain the frequencies of the outcomes in the first place? This tends to be a process combining a priori knowledge, **bootstrapping**, and common sense. A first linkage can be carried out using weights borrowed from previous similar linkages or worked out a priori (e.g. agreement of month of birth at around odds ratio 12). This first linkage can be used to produce files of pairs which do and do not belong to the same person. These "linked and unlinked" files can be used for more precise empirical derivation of odds ratios for a further linkage, and so on. A valuable feature of probability matching is its **robustness** in the face of imperfections in the probability weights assigned to outcomes. As long as weights are broadly correct, the linkage will work.

The final refinement of weights will usually take place after inspection of a sample of the weights close to the decision threshold alongside the pairs of records that generated them. Features of the linkage will usually emerge which could not have been anticipated in advance. To paraphrase Newcombe, linkage is an empirical, iterative and above all common-sense procedure [13].

A Brief Note on Name Compression. Discrepant spelling of surname is one of the most common aspects of mismatching identifiers in two records belonging to the same person. Various methods have been proposed for dealing with this, but a commonly used solution is to combine the method of Soundex codes, whereby similar sounding consonants are brought together and vowels are largely ignored, with the NYSIIS name compression algorithm, whereby other commonly miscoded elements of surnames are brought together. The resulting Soundex/NYSIIS codes are treated like any other identifier [9].

Making the Linkage Decision

The Linkage Threshold. The probability weights calculated by the methods outlined above do not represent absolute odds that the records concerned belong to the same person. They are relative odds that serve to order the pairs in a particular linkage according to the likelihood that they belong to the same person.

A useful diagnostic output in any linkage is a **frequency distribution** of the probability weights achieved by the pair comparisons. This is usually, but not always, bimodal. The bimodal distribution is produced by the superimposition of the weight distribution for pairs of records which do belong to the same person on the distribution for pairs which do not belong to the same person.

The decision threshold for linkage is usually determined by clerical inspection of a sample of pair comparisons across the weight range. The linkage decision can be made either completely automatically or can use supplementary clerical checking. If automatic linkage is chosen, a single threshold is chosen based on the sample checking. If a pair scores above the threshold, the records are linked. If a pair scores below the threshold, the records are not linked. If supplementary clerical checking is involved, then two thresholds are used to define a "gray zone" within which pairs will be clerically checked to make the final decision. Above the higher threshold, links are accepted automatically. Below the lower threshold links are rejected automatically. Within the "gray zone", human judgment is used to make the final decision.

The conversion of relative odds to absolute odds depends upon several factors, including the way in which the linkage is structured. Structuring the linkage to optimize the "terms of conversion" of relative odds to absolute odds is one of the most important aspects of designing a linkage.

The absolute odds required for acceptance that two records belong to the same individual depends upon the purpose of the linkage and the relative cost of a **false positive** link (linking two records which do not belong to the same person) compared with a **false negative** link (failing to link two records which do belong to the same person). For most statistical purposes, a best estimate linkage is required, and the decision threshold will be set at absolute odds of 50/50. For administrative purposes, it may be vitally important that wrong links are not made, while it is less crucial that links are missed. The decision threshold would thus be set at relatively high absolute odds. Where the purpose of the linkage is simply to trawl for potential candidates for linkage – the links themselves being established by other means – a relatively low threshold would be set.

Converting Relative Odds to Absolute Odds: the Importance of Context. As a first step in converting relative odds into absolute odds, Newcombe [9] has proposed numeric rules relating to two files: one a search file and the other a file being searched. Proposition A is that the higher is the proportion of records in the search file for which there exists a linked record in the file being searched, the better will be the conversion factor between relative and absolute odds. Proposition B is that the larger is the file being searched, the worse will be the conversion factor between relative and absolute odds.

These numeric considerations in converting absolute odds to relative odds show that the meaning of a given probability weight depends upon the wider context of the linkage, and in particular on the relationships between the records in the files to be linked [10].

This is a crucial insight, the wider implications of which must be drawn out. In designing a linkage, it is important to structure the linkage so as to take maximum advantage of the structures of the files involved and the relationships between them. For example, are the relationships between the records in two files one-to-one, one-to-many, or many-to-many? How much confidence do we have in previous linkages that may have been carried out on the files involved? How confident are we that a file to be linked already contains only one record per person?

For example, if we want to link to each other a set of hospital discharge records, we have no a priori knowledge of how many records belong to each person. Our best bet is to do a conventional internal linkage and inspect all resulting pairs in setting a threshold. In this case we have relatively little leverage to improve the terms of conversion between relative odds and absolute odds.

However, if we are linking a file of hospital discharge records to a file of death records (*see* **Death Certification**), we can obtain some "structural leverage". Death only occurs once, and assuming that this is reflected in there being only one death record

per person in the file of death records, the linkage becomes many-to-one. Each hospital discharge record should link to only one death record. The terms of conversion from relative to absolute odds can be improved by only retaining, for each hospital discharge record, the best (highest weight) link which is achieved to a death record. Similarly, at the other end of the life cycle, if we are linking babies to mothers, assuming that the mothers' records themselves have been correctly linked, we should allow each baby to link to only one mother.

The most powerful leverage occurs when there is a close to a one-to-one relationship between the files to be linked. For example, as groundwork for the creation of a unique patient identifier for the Scottish population it was necessary to link the regionally operated Community Health Index (CHI) to the National Health Service Central Register (NHSCR) in Scotland. Because both files had close to 100% population coverage, there was a very high a priori probability that there existed an NHSCR record for every CHI record, thus maximizing the conversion factor in terms of Newcombe's rules. Again, by applying the best-link principle, whereby only the highest weighted link for each CHI record was accepted, a massive degree of leverage was obtained. Administratively acceptable links (involving very high absolute odds of linkage) were achieved at low relative odds [8].

Summary

Record linkage using probability matching may be implemented in a very simple and straightforward way or it may involve a highly complex and delicate **algorithm**. However it is done, it is based on simple and initially intuitive insight. Success in record linkage comes from adapting the basic principle of probability matching to the precise characteristics of the data involved – both in terms of making best use of the identifiers available and in terms of structuring the linkage to obtain the greatest leverage from the relationships between the records involved.

Implications

Record linkage using probability matching is an extremely powerful tool, and will become increasingly powerful as techniques improve and computing hardware considerations become less restrictive. Because of its ability to bring together information in ways which were not part of the original intent when the data were first collected, it can be seen as posing a threat in terms of data **confidentiality** and individual privacy.

At all times, the benefits that record linkage using probability matching may bring for patient administration and medical research must be balanced against the dangers that it poses [11]. Careful monitoring and supervision are required.

References

[1] Baldwin, J.A., Acheson, E.D. & Graham, W.J., eds (1987). *Textbook of Record Linkage*. Oxford University Press, Oxford.

[2] Fair, M.E. (1995). An overview of record linkage in Canada, in *American Statistical Association 1995 Proceedings of the Social Statistics Section*. American Statistical Association, Alexandria, pp. 25–33.

[3] Fellegi, I.P. & Sunter, A.B. (1969). *Journal of the American Statistical Association* **64**, 1183–1210.

[4] Gill, L.E. & Baldwin, J.A. (1987). Methods and technology of record linkage: some practical considerations, in *Textbook of Record Linkage*, J.A. Baldwin, E.D. Acheson & W.J. Graham, eds. Oxford University Press, Oxford, pp. 39–54.

[5] Heasman, M.A. & Clarke, J.A. (1979). Medical record linkage in Scotland, *Health Bulletin (Edinburgh)* **37**, 97–103.

[6] Kendrick, S.W. & Clarke, J.A. (1993). The Scottish Record Linkage System, *Health Bulletin (Edinburgh)* **51**, 72–79.

[7] Kendrick, S.W. & McIlroy, R. (1996). One pass linkage: the rapid creation of patient-based data, in *Proceedings of Healthcare Computing 96: Current Perspectives in Healthcare Computing 1996*. British Journal of Healthcare Computing Books, Weybridge, Surrey, pp. 589–598.

[8] Kendrick, S.W., Douglas, M.M., Gardner, D. & Hucker, D. (1997). The best-link principle in the probability matching of population data sets: the Scottish experience in linking the Community Health Index to the National Health Service Central Register, *Methods of Information in Medicine* **37**, 64–68.

[9] Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford University Press, New York.

[10] Newcombe, H.B. (1994). Age-related bias in probabilistic death searches due to neglect of the "prior likelihoods", *Computers and Biomedical Research* **28**, 87–99.

[11] Newcombe, H.B. (1995). When "privacy" threatens public health, *Canadian Journal of Public Health* **86**, 188–192.

[12] Newcombe, H.B., Fair, M.E. & Lalonde, P. (1992). The use of names for linking personal records, *Journal of the American Statistical Association* **87**, 1193–1208.

[13] Newcombe, H.B., Smith, M.E. & Lalonde, P. (1986). Computerized record linkage in health research: an overview, in *Proceedings of the Workshop on Computerized Linkage in Health Research, Ottawa, May 1986*. University of Toronto Press, Toronto, pp. 28–34.

[14] Newcombe, H.B., Kennedy, J.M., Axford, S.J. & James, A.P. (1959). Automatic linkage of vital records, *Science* **130**, 954–959.

STEVE KENDRICK & MARY SMALLS

Matching

Before discussing the procedure known as matching, it is necessary to provide some background and motivation for its use. In epidemiologic studies, it is typically the situation that valid **estimation** of the strength of the relationship between a response variable D of interest (e.g. the presence, $D = 1$, or not, $D = 0$, of some particular disease) and an independent variable E of interest (e.g. the presence, $E = 1$, or not, $E = 0$, of some exposure) necessitates the consideration of so-called **confounding** factors (*see* **Confounder**). Ignoring or inappropriately accounting for the effects of confounding factors can often lead to invalid (i.e. statistically inconsistent) and inefficient estimation of the true exposure–disease association of interest.

As a simple example, suppose that the dichotomous response (or disease) variable D of interest is the presence or absence of lung cancer and that the dichotomous independent (or exposure) variable E of interest is the presence or absence of a history of occupational exposure to asbestos. Then, a dichotomous variable C such as cigarette smoking status (e.g. evidence, $C = 1$, or not, $C = 0$, of a history of smoking), which is an established risk factor for the development of lung cancer, will be a confounder if, *in the data under consideration*, its distribution among the group of study subjects with a history of occupational exposure to asbestos (the "exposed group") is different from its distribution among the group of study subjects who do not have a history of occupationally related asbestos exposure (the "unexposed group"). If C is, in fact, a confounder in the data under consideration, then appropriate adjustment for C *at the analysis stage* (e.g. by **stratification** methods or, equivalently, by multivariable modeling) would be needed. In our particular example involving the three dichotomous variables D, E, and C, one could fit the **logistic regression** model $\text{logit}\,[\Pr(D = 1)] = \beta_0 + \beta_1 E + \gamma_1 C$ by appropriate **likelihood** methods to obtain an adjusted (for C) estimated **odds ratio** $\exp(\hat{\beta}_1)$ and to obtain a corresponding **interval estimator** for the population E–D odds ratio $\exp(\beta_1)$. Here, we are assuming that C is not an **effect modifier** (i.e. there is no **interaction** between E and C), so that it is not necessary to include the product term EC in the above model; we will make this no interaction assumption in our discussion to follow.

However, adjustment for C at the analysis stage can be problematic. For example, if almost all of the study subjects with a history of smoking have lung cancer (i.e. are "cases"), and if a large proportion of the study subjects with no smoking history are "noncases", then such stratum-specific imbalances can lead to poor statistical efficiency in the point and interval estimation of the odds ratio parameter $\exp(\beta_1)$. In more realistic situations where there are typically several confounders to consider simultaneously, distributional imbalances in strata defined by combinations of levels of these confounders can severely compromise the reliability of multivariable modeling analyses.

Design Options: Restriction and Matching

By using appropriate strategies at the *design stage* of a study, it is often possible to avoid many of the confounder-related distributional imbalance problems mentioned earlier. For example, consider a potentially confounding variable such as gender. One way to avoid completely any possible problems associated with an analysis stage adjustment for the variable gender is to decide, at the design stage, to restrict the study so that it involves either only males or only females. This simple study design option is called (*total*) *restriction* because the potential confounder is completely restricted to have exactly the same value for every study subject. Clearly, the disadvantage of (total) restriction is the lack of generalizability of the study results; in our example, by employing (total) restriction with respect to gender, the study conclusions would necessarily only pertain either to males or to females.

Matching, in contrast to total restriction, is a form of *partial restriction* on study subject selection, partial in the sense that only the so-called "referent (or comparison) group", and not the "index group", is chosen subject to certain restrictions. More specifically, for follow-up studies (*see* **Cohort Study**), once the index group of exposed ($E = 1$) subjects is randomly selected from the population of interest, the referent group of unexposed ($E = 0$) subjects is then chosen to be similar to the exposed group with respect to the distributions of one or more potentially confounding factors. For **case–control studies**, once the

index group of diseased ($D = 1$) subjects is chosen at random from the population of interest, the referent group of nondiseased ($D = 0$) subjects is picked to be similar to the cases with respect to the distributions of one or more potentially confounding factors. We use the word "similar", rather than "identical", because the index and matched referent groups will generally not have exactly the same confounder distributions after matching; the degree of similarity will depend on the type and the extent of matching employed.

To discuss types of matching schemes, we need to distinguish between matching on continuous variables (e.g. age, weight, cholesterol level) and matching on categorical variables (e.g. gender, race). Matching on a continuous variable (say, X) necessitates the specification of a rule for deciding when an index subject's value (say, X_1) and a referent subject's value (say, X_0) are "close enough" to declare that the two subjects are "matched" on X. In so-called "caliper matching", one specifies a caliper (or tolerance) value C and declares the index and referent subjects to be matched if $|X_1 - X_0| \leq C$.

The smaller is C, the tighter will be the match on X, but, correspondingly, the harder it will be to find index–referent pairs to satisfy such a stringent matching criterion [8, 9].

Since, in standard epidemiologic practice, variables are generally categorized for matching purposes (e.g. note that caliper matching defines categories of width C), we will henceforth focus on so called *category (or frequency) matching*. In particular, index and referent subjects are said to be matched on a categorized potential confounder if they are in the same category of that variable. In the realistic situation where category matching involves several potential confounding variables, index and referent subjects are said to be matched when they are in the same category for each and every one of the categorized matching variables under consideration. For example, suppose that there are three categorized matching variables of interest: age in four categories (30–39, 40–49, 50–59, and 60–69), race (black, white, and other), and gender (male and female). Then, there will be 24 strata defined by the various combinations of these three matching variables, with, for example, one stratum consisting of black females between the ages of 40 and 49.

In general, then, matching can be considered to be pre (or design stage)-stratification, as opposed to **post** (or analysis stage)-**stratification**, with the goal of such matching being to form strata that are sufficiently balanced to permit valid, stable, and efficient statistical analyses. Once matching is employed at the design stage, it is mandatory at the analysis stage to take the matching into account via the use of appropriate stratified analysis methods [5]. Such **categorical data analysis** procedures include the approach of **Mantel & Haenszel** [6] and the use of conditional logistic regression methods [1, 4] (*see* **Logistic Regression, Conditional**; **Matched Analysis**).

Types of Matching Schemes

There are various types of matching schemes that can be used. One of the more popular matching schemes, especially in case–control studies, is known as *pair matching*. Pair matching refers to the special situation when each stratum is assumed, for analysis purposes, to contain exactly one index subject and one referent subject. However, this assumption will generally lead to an inefficient stratified analysis when the pairing is artificial and unnecessary. For example, for a stratum of cases and controls consisting of black females between the ages of 40 and 49, any case in that stratum could theoretically be paired with any control without altering the basic within-stratum structure. Retaining this "random" pairing in the analysis is clearly unwarranted, and such an "overmatched analysis" generally leads to some loss in statistical efficiency [2]. In contrast, the term "**overmatching**" commonly refers to an undesirable design-stage strategy of matching on variables that make the cases and controls too much alike with respect to exposure status. Such variables are generally of two types, namely, so-called "intervening variables" that are intermediate in the causal pathway between exposure and disease and variables that are (at best) very weak risk factors for the disease in question but are nevertheless highly correlated with exposure status [10]. Such overmatching can sometimes lead to a meaningful loss in statistical efficiency, especially in case–control studies (see "Discussion" below).

A generalization of pair matching is a procedure known as R-to-1 matching, where each stratum is considered to contain one index subject and exactly R referent subjects. Miettinen [7] and others have shown that there is little to gain statistically by taking $R > 4$. For example, when comparing R-to-1

matching with pair matching ($R = 1$) in case–control studies, Ury [11] has shown that the **Pitman efficiency** of the Mantel–Haenszel test for stratified data is $2R/(R + 1)$, so that the Pitman efficiency only increases from 1.600 for $R = 4$ to 1.667 when $R = 5$.

In the most general category matching situation, a particular stratum (say, the gth of G strata) may contain R_g referent subjects and S_g index subjects, giving a *matching ratio* of R_g/S_g (which is not necessarily an integer). If this matching ratio varies with g, then we have a *variable matching ratio plan*. If the matching ratio does not vary over the strata (e.g. as with R-to-1 matching), then we have a *fixed matching ratio plan*. With either plan, the appropriate data analysis would still appropriately accumulate stratum-specific information; and, in terms of statistical efficiency, a fixed matching ratio plan is usually somewhat better.

Advantages and Disadvantages of Category Matching

Some of the *positive aspects* of category matching in epidemiologic studies are as follows:

1. Category matching a set of referent subjects to a **random sample** of index subjects can often lead to a more statistically efficient analysis than can be obtained by choosing the same number of referent subjects by random sampling. This efficiency advantage will tend to occur when the matching variables are well-established determinants of the response variable (e.g. are important risk factors for the disease under study) and are expected to be quite differentially distributed between the exposed and unexposed groups in the observed data (i.e. are anticipated to be strong confounders). For more detailed discussion, see Kupper et al. [5] and Karon & Kupper [3].
2. Matching on a variable like neighborhood of residence can lead to efficient adjustment for the potentially confounding effects of a wide range of social and economic factors that would be difficult, if not impossible, to measure and hence to control.
3. Matching can often lead to savings in time and money. For example, when the cases in a case–control study are chosen from records in different hospitals or in different companies within some industry, it is preferable, for reasons of simplicity and convenience in data collection (and also possibly on validity and efficiency grounds), to choose controls for each case from that same set of hospital or company records.
4. Matching in the selection of the referent group with respect to a given set of potential confounders does not preclude controlling for other nonmatched confounders at the analysis stage via multivariable modeling procedures like conditional logistic regression. In this regard, a recommended strategy would be to match only on important risk factors considered a priori to be highly likely to manifest themselves as strong confounders in the data, and to adjust (if necessary) for other factors at the analysis stage.

Some possible *negative aspects* of category matching in epidemiologic studies are the following:

1. Category matching can be a costly enterprise, both with regard to the *direct* costs of time and labor required to find the appropriate matches and the *indirect* costs (in terms of information loss) owing to the discarding of available referents not able to satisfy possibly stringent matching criteria.
2. When employing category matching, simultaneous recruitment of cases and controls can be problematic since there is no way to know in advance exactly how many controls will be needed to meet sample size requirements in different matching strata defined by the sample of cases. To circumvent this problem, a new "randomized recruitment" method for matching has been developed [12, 13].
3. The referent group chosen by category matching ends up being more like the index group than like the underlying population of referents being sampled. In particular, matching generally precludes the evaluation of the underlying population relationships between the matching variables and exposure status in follow-up studies or between the matching factors and disease status in case–control studies.
4. If the strata defined by the category matching process are wide (so that there is room for the matching factors each to vary sufficiently in value within particular strata), it is possible that stratum-specific residual confounding due to the matching factors can still be present.

Appropriate adjustment for such stratum-specific residual confounding at the analysis stage can be accomplished using multivariable modeling procedures.

Discussion

In summary, category matching on potential confounders can be a fruitful design-based strategy in both follow-up and case–control studies when reliable information, based on knowledge of the disease process under study and previous research findings, indicates that such variables are well-established disease determinants (i.e. are strong risk factors) expected to be quite differentially distributed between exposed and unexposed groups if matching is not employed (e.g. under random sampling of the referent group).

As a word of caution, the use of matching requires more care in case–control studies than in follow-up studies. Since exposure information is collected *after* the occurrence of disease in case–control studies, indiscriminate *overmatching* of controls to cases simultaneously on several factors can lead to a substantial loss in efficiency relative to random sampling of the control group. For example, consider a pair-matched case–control study involving n case–control pairs, where a is the number of pairs where both the case and control are exposed, b is the number of pairs where the case is exposed and the control is not, c is the number of pairs where the control is exposed and the case is not, and d is the number of pairs where neither the case nor the control is exposed. Then, the Mantel–Haenszel test statistic [6] takes the form $(b-c)^2/(b+c)$, and the appropriate odds ratio estimator is b/c (namely, the ratio of discordant pairs). Hence, the effective sample size in such a study is the total number of discordant pairs $(b+c)$, not n. If the matching variables are each correlated with the exposure variable, then overmatching generally increases the number of uninformative pairs in the observed data, namely $(a+d)$, thus leading to a (possibly substantial) loss in efficiency. Thus, in case–control studies especially, the best policy is to consider as candidate matching variables only well-established strong risk factors for the disease in question. As mentioned earlier, matching either on intervening variables or on very weak risk factors highly correlated with exposure status should be avoided.

References

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1: *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

[2] Brookmeyer, R., Liang, K.Y. & Linet, M. (1986). Matched case-control designs and overmatched analyses, *American Journal of Epidemiology* **124**, 693–701.

[3] Karon, J.M. & Kupper, L.L. (1982). In defense of matching, *American Journal of Epidemiology* **116**, 852–866.

[4] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont.

[5] Kupper, L.L., Karon, J.M., Kleinbaum, D.G., Morgenstern, H. & Lewis, D.K. (1981). Matching in epidemiologic studies: validity and efficiency considerations, *Biometrics* **37**, 293–302.

[6] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[7] Miettinen, O.S. (1969). Individual matching with multiple controls in the case of all-or-none responses, *Biometrics* **22**, 339–355.

[8] Raynor, W.J. & Kupper, L.L. (1981). Category matching of continuous variables in case-control studies, *Biometrics* **37**, 811–817.

[9] Rubin, D.R. (1973). Matching to remove bias in observational studies, *Biometrics* **29**, 159–183.

[10] Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.

[11] Ury, H.K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data, *Biometrics* **31**, 643–649.

[12] Weinberg, C.R. & Wacholder, S. (1990). The design and analysis of case-control studies with biased sampling, *Biometrics* **46**, 963–975.

[13] Weinberg, C.R. & Sandler, D.P. (1991). Randomized recruitment in case-control studies, *American Journal of Epidemiology* **134**, 421–431.

LAWRENCE L. KUPPER

Maternal Mortality

Maternal mortality claims the lives of some 585 000 women a year, 99% of them in the developing world. It is the main cause of death among young women aged 15–19, and the third or fourth most common cause in women of childbearing age, generally defined as 15–49. The differences between the developed and the developing world in levels of maternal mortality are greater than for any other indicator of public health: in developed countries a woman has a lifetime risk of maternal death of 1 in 1800: in developing countries this risk is 1 in 48. However, the risks range from 1 in 4000 in the industrialized of countries of northern Europe, to 1 in 12 in eastern and western Africa [1]. Maternal mortality also has severe consequences for the health of children: in developing countries the baby born to a woman who dies in childbirth rarely survives, and her older children face much greater risks of death [7].

Definition of Maternal Mortality

A maternal death is the death of a woman while pregnant, or within 42 days of the termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes [9]. This classification therefore includes deaths from abortion, spontaneous or induced, or from an ectopic pregnancy, but not deaths in pregnancy or the postpartum period caused by violence or accidents.

The distinction between causes related to, or aggravated by, pregnancy or its management gives rise to two other definitions: "direct" and "indirect" obstetric deaths. Direct obstetric deaths are those related to complications of pregnancy, labor or in the 42-day postpartum period (the puerperium), from interventions, or from incorrect treatment or omissions in treatment. Indirect obstetric deaths are those resulting from a pre-existing disease, or one that developed during pregnancy, and that is aggravated by pregnancy. Before 1975, deaths from indirect causes were not classified as maternal deaths.

Causes of Maternal Deaths

On the evidence of a few good community-based studies, direct causes account for the majority–80%–of maternal deaths. In turn, five major causes account for 80% of these direct maternal deaths. Although there is some variation in their relative importance among regions, the distribution of the five causes at global level is as follows: hemorrhage (25%); sepsis (15%); unsafe abortion (13%); eclampsia (8%) and obstructed labor (7%). Indirect causes of maternal death, such as anemia, malaria, cardiovascular disease, hepatitis, and diabetes, account for the remaining 20% of all maternal deaths [1] (*see* **Cause of Death, Underlying and Multiple**).

Sources of Data on Maternal Mortality

Gathering data on maternal mortality is difficult and expensive – and often beyond the resources of the very countries in which the problem is greatest. The chief sources of information are vital registration systems, health services data, and population-based surveys (*see* **Administrative Databases**; **Surveys, Health and Morbidity**; **Vital Statistics, Overview**).

Few developing countries have registration systems to provide information on the numbers of deaths (*see* **Death Certification**). Those that do can rarely provide information on the cause of death, or require that death certificates note pregnancy status. Moreover, in countries where induced abortion is illegal, official statistics seldom fully reflect deaths from this cause. Health service statistics suffer from **selection bias**: women who die in pregnancy or childbirth in health facilities often differ in important health and socioeconomic characteristics from pregnant women in the broader community. It is also difficult to define the appropriate catchment area of a hospital (*see* **Hospital Market Area**) for the derivation of ratios and rates. **Population-based studies**, therefore, have been used increasingly to gather information. Their most important drawback is expense: maternal mortality is a rare event compared with **infant mortality**, for example, and sample sizes need to be very large to obtain reliable estimates. Costs are somewhat lower where questions are added to **censuses** or surveys: the "sisterhood method", for example, has yielded useful

information on maternal mortality by asking adult respondents whether any sisters have died in their childbearing years. The most reliable data, however, where vital registration is incomplete or lacking, are obtained from studies that identify all deaths to women of reproductive age (reproductive age mortality surveys, or RAMOS). Interviewers consult many community sources and then, on the basis of symptoms described by family members and health care providers, classify the deaths as maternal or otherwise. Very few countries have been able to afford these studies.

Given these problems, but faced with the need to measure progress in reducing maternal mortality, the **World Health Organization** (WHO) and the United Nations International Children's Emergency Fund (UNICEF) have recently developed new estimates of maternal mortality [10]. They used country data where available, adjusted for undercount and **misclassification**, and developed a model to predict values for countries with no reliable national data. At the global level, the new estimates represent a significant upward revision of the annual number of maternal deaths – an increase of 80 000 over the figure of just over 500 000 in use for the past 10 years.

Measuring Maternal Mortality

The three most common measures of maternal mortality are the lifetime risk, the maternal mortality rate, and the maternal mortality ratio. The ratio is the number of maternal deaths per 100 000 live births during a certain time period, and is, therefore, a measure of the risks women face when they are pregnant, generally called obstetric risk. However, in order to run this risk, women must be pregnant. The lifetime risk and the maternal mortality rate take account of fertility: they measure both obstetric risk, and the frequency with which women are exposed to that risk through pregnancy. This is seen most easily in the maternal mortality rate: the number of maternal deaths per 100 000 women of reproductive age during a certain time period. The following equation demonstrates the relationship [2]:

$$\text{maternal mortality rate} = \text{maternal mortality ratio} \times \text{general fertility rate}$$

$$\frac{\text{maternal deaths}}{\text{women } 15\text{–}49} = \frac{\text{maternal deaths}}{\text{live births}} \times \frac{\text{live births}}{\text{women } 15\text{–}49}.$$

In using data on maternal mortality it is important to note how the maternal mortality rate is defined. Historically, it was defined as the number of maternal deaths per 100 000 live births, i.e. the definition of the ratio given above, and this is still the definition used in the tenth revision of the **International Classification of Diseases** published by WHO in 1992 in order to provide consistency with previous editions [9]. However, in its analytical work on maternal mortality WHO also distinguishes between the rate and the ratio using the definitions given above [1, 10], as exemplified in Table 1. The distinction is important for directing attention to appropriate interventions.

Table 1 Measures of maternal mortality in developed and developing countries, 1990

	Maternal mortality ratio (maternal deaths per 100 000 live births)	Number of maternal deaths	Lifetime risk[a] of maternal death (1 in:)
World	430	585 000	60
More developed regions	27	4000	1800
Less developed regions	480	582 000	48

Sources: [1, Table 2; 10].

[a]Lifetime risk devised by Roger Rochat, Emory University School of Medicine, USA. Calculated as $1 - (1 - \text{MMR})^{(1.2\text{TFR})}$], where the maternal mortality ratio (MMR) is expressed as a decimal and the total fertility rate (TFR) is adjusted by 1.2 to allow for pregnancies not ending in live births.

The other measure of maternal mortality that also takes into account both the risks within pregnancy and the risks of pregnancy is lifetime risk. In fact, this is the more commonly used indicator of international disparity, conveying graphically the risks of pregnancy in countries with high fertility. Table 1 presents the regional differences in maternal mortality ratios and lifetime risk by region.

Wide though the differences in maternal mortality ratios are, they are much less than the differences in lifetime risk: the risks women face in pregnancy and childbirth are compounded by the frequency with which they face those risks.

Actions to Reduce Maternal Mortality

It follows that maternal mortality can be reduced by interventions that reduce fertility, and that reduce obstetric risk (*see* **Reproduction**). Reductions in the total number of pregnancies result in fewer women at risk of a maternal death: a comparison of maternal mortality in Bali, Indonesia, and Menoufia, Egypt in the 1980s provides a telling example. The maternal mortality ratio in Bali was 718: in Menoufia it was 190–3.8 times as high. Yet the maternal mortality rate of 69 in Bali was only 1.5 times as high as the rate of 45 in Menoufia – because fertility was lower in Bali [3]. Changes in fertility, closely associated with the adoption of family planning, therefore have an important impact on the maternal mortality rate (and on lifetime risk).

Changes in fertility can also affect the maternal mortality ratio by reducing the number of high-risk pregnancies – pregnancies that are unwanted and that may lead women to run the risk of unsafe abortion, or pregnancies in women of older age, or who have had four or more previous births, or whose last birth occurred less than two years previously. These women are often the first to use family planning services, when available. In Bali, contraceptive use was higher, and fertility rates lower among older women than in Menoufia. However, since, in general, most births occur to women at "safe" ages and parities, the majority of maternal deaths do too. Thus, the chief reductions in the maternal mortality ratio are to be achieved by reducing the risks in pregnancy.

Underlying the immediate medical causes of maternal death are many factors contributing to the risk of maternal death. Women's socioeconomic status is a powerful determinant of their health status–and of their access to health services. Socioeconomic status also affects fertility, and the risk of maternal death: women with no, or little, primary education have more children than women with secondary and higher education (*see* **Social Classifications**). Raising women's status is necessary to improving maternal health, but is a long-term objective. In the short term, interventions to reduce the number of obstetric complications, and the number of deaths among women who develop complications, are essential.

Basic Maternal Care

Women need care throughout pregnancy, delivery, and in the postpartum period. Antenatal care is necessary to inform women on how to take care of themselves throughout pregnancy and childbirth, how to recognize danger signals, and what to do should complications arise. It is also necessary to treat conditions that can lead to complications, such as anemia, or which are aggravated by pregnancy, such as malaria and viral hepatitis. Health facilities providing antenatal care, however, need to be linked closely with facilities able to deal with complications that may arise: it is doubtful whether antenatal care that is not part of a more comprehensive system contributes to maternal mortality reduction [6]. Care in delivery should include attendance by trained personnel, in clean conditions to prevent sepsis, and, again, with access to health facilities and providers with the skills, equipment, and drugs to prevent, detect, and manage complications during birth, and during the postpartum period. A recent review and **meta-analysis** (synthesis of findings) of studies of maternal mortality in developing countries indicates that care in the postpartum period is essential: 60% of maternal deaths occurred in the postpartum period [5]. Almost half of postpartum deaths – 45% – occurred in the first 24 hours after delivery, and nearly three-quarters within the first week. The time of death varied according to cause: most postpartum deaths from hemorrhage and pregnancy-induced hypertension (eclampsia) occurred during the first day and week after delivery: most postpartum deaths from sepsis occurred in the second week and later.

Essential Obstetric Care and Emergency Obstetric Care

While basic maternal care meets the needs of all women whose pregnancies, labor, or delivery are uncomplicated, or who are able to return to this care when a complication has been successfully treated, an estimated one-third to one-half of pregnant women develop obstetric complications, and an estimated 15% develop complications that require emergency care. Both essential obstetric care for all complications, and emergency care, have been known by the acronym EOC, giving rise to some confusion. The Inter-Agency Group (IAG) on safe motherhood, comprised of representatives of several of the UN agencies and nongovernmental organizations active in the field of reproductive health, recommends differentiating between these terms by use of the acronym ECOC for Essential Care of Obstetric Complications, and EMCOC for Emergency Care for Obstetric Complications. The obstetric functions provided by ECOC include the functions necessary for EMCOC and, according to the IAG, should be available to all pregnant women with problems, including complications of unsafe abortion. ECOC comprises: surgical obstetrics; anesthesia and medical treatment; blood replacement and manual procedures; labor monitoring, management of problem pregnancies, and neonatal special care (statement developed at the IAG meeting, February 1996).

Maternal Health

Reductions in maternal mortality cannot be equated with improvements in maternal health. They may even be associated with increases in maternal morbidity – the disabilities suffered by women as a result of pregnancy, childbirth, and abortion, or the exacerbation of existing health problems by pregnancy. It has been estimated that for every woman who dies in pregnancy, another 15 survive but suffer long-term consequences [8]. It is also important to recognize that maternal health is part of women's health more broadly defined: this might seem a truism, but some are concerned that emphasis on maternal mortality has stressed women's maternal roles to the exclusion of recognition of other influences on women's health that also, though more indirectly, would improve maternal health [6]. In a slightly different vein, others worry that the drive to measure mortality, and the impact of programs on mortality, may divert resources that would be better deployed in strengthening implementation of those programs. They argue for developing indicators that measure progress in increasing the availability, quality and utilization of maternity services, thus contributing to improved maternal health more generally [4]. The new estimates and methodology developed by WHO and UNICEF, should help to reduce this pressure to produce ratios and **rates**. It is generally agreed, however, that investments in maternal health services, including family planning, provide significant health benefits to women at low cost and are essential components of basic health care. Work [11] on the burden of death and disability caused by various diseases, and on the health benefits and costs of interventions, found prenatal and delivery care and family planning to be among the most cost-effective of health interventions (*see* **Health Economics**).

References

[1] Abou Zahr, C., Wardlaw, T., Stanton, C. & Hill, K. (1996). Maternal mortality, *World Health Statistics Quarterly* **49**, 77–87.

[2] Fortney, J.A. (1987). The importance of family planning in reducing maternal mortality, *Studies in Family Planning* **18**, 109–114.

[3] Fortney, J.A., Susanti, I., Gadalia, S., Saleh, S., Feldblum, P.J. & Potts, M. (1988). Maternal mortality in Indonesia and Egypt, *International Journal of Gynecology and Obstetrics* **26**, 21–32.

[4] Graham, W., Filippi, V.G.A. & Ronsmans, C. (1996). Demonstrating programme impact on maternal mortality, *Health Policy and Planning* **11**, 16–20.

[5] Li, X.F., Fortney, J.A., Kotelchuck, M. & Glover, L.H. (1996). The postpartum period: the key to maternal mortality, *International Journal of Gynecology and Obstetrics* **54**, 1–10.

[6] McDonagh, M. (1996). Is antenatal care effective in reducing maternal morbidity and mortality? *Health Policy and Planning* **11**, 1–15.

[7] Over, M., Ellis, R.P., Huber, J.H. & Solon, O. (1992). The consequences of adult ill-health, in *The Health of Adults in the Developing World*, R.G.A. Feachem, T. Kjellstrom, C.J.L. Murray, M. Over & M.A. Phillips, eds. Oxford University Press for the World Bank, New York, pp. 161–207.

[8] Starrs, A. (1987). Preventing the tragedy of maternal deaths, *Report on the Safe Motherhood Conference 1987*.

[9] WHO (1992). *International Statistical Classification of Diseases and Related Health Problems*, 10th revision. World Health Organization, Geneva.

[10] WHO and UNICEF (1996). Revised 1990 estimates of maternal mortality. A new approach by WHO and UNICEF, *WHO/FRH/MSM96.11* and *UNICEF/PLN/96.1*.

[11] World Bank (1993). Investing in health, *World Development Report*. Published for the World Bank by Oxford University Press, Washington.

(*See also* **Health Services Data Sources in Canada**; **Health Services Data Sources in Europe**; **Health Services Data Sources in the US**)

J. NASSIM

Mathematical Biology, Overview

Mathematical biology has become a flourishing field in which real mathematics combines with real biology. The field is represented by numerous refereed journals, including *Journal of Mathematical Biology, Bulletin of Mathematical Biology, Biomathematics, IMA Journal of Mathematics Applied in Medicine and Biology, Journal of Theoretical Biology, Mathematical Biosciences, Biological Cybernetics*, and *Theoretical Population Biology*. In addition, there are frequent biological articles in *Biophysical Journal* and *SIAM Journal of Applied Mathematics*, and occasional mathematical articles in *Journal of Neurophysiology*, and even *Journal of Molecular Biology*. Lecture note series include *Lecture Notes in Biomathematics* and *Lectures on Mathematics in the Life Sciences*. There are also a number of good recent general textbooks [4, 8, 14], an excellent collection of mathematically sophisticated research papers [12], a survey of applications and unsolved problems in biomedical imaging [3], and important monographs on computational biology [23], stochastic models of carcinogenesis [22], and cardiac arrhythmias [25]. Medical applications include tomographic imaging, genetic linkage analysis, cardiac arrhythmias, epidemic diseases and control strategies, carcinogenesis, and tumor chemotherapy. A recent conference was reviewed in *Science* [6]. This list is necessarily incomplete, but it may be useful for beginning reading in the field.

Mathematical biology consists, not of the mathematics of living things, but of the mathematics of *models* of living things; and choosing the level of simplification and the nature of the abstraction is perhaps the modeler's most distinctive contribution. Mathematical skill and biological knowledge are necessary conditions for success in modeling, but the art of selecting the essential ingredients of complex and elusive phenomena goes beyond them.

> Simplification in mathematical modeling is both a blessing and a curse. The curse is the partial loss of predictive power that comes from whatever lack of correspondence there may be between the model and the real world. The blessing is the insight that comes from the process of pruning away unnecessary detail and leaving behind only what is essential. ... The models presented here are in the nature of metaphors, and these metaphors will have served their purpose if they have helped the reader to see through the bewildering complexity of living systems to the underlying simplicity of certain biological processes and functions (Hoppensteadt and Peskin [8, p. 3]).

It is natural, in this era of fast computation, to want to build as much realism as possible into biological models, and rely on the power of the computer to approximate the complex systems of equations that result and make quantitative predictions that can be tested against experimental data. While not minimizing the practical utility of computer **simulation**, e.g. in cardiac pacemaker design or in predicting the course of epidemics, it may be that the most distinctive contribution of mathematics to biology lies in the opposite direction: simplifying to the point where it is possible to prove *theorems* about the model, and not only to compute with it. The computational route may yield a model that gives good predictions, but is just as impenetrable to understanding as the original biological system, whereas, when one proves a theorem about a model, that theorem is likely to give insight into the underlying biology. Proving nontrivial theorems about models that are adequate "metaphors," to use Hoppensteadt & Peskin's term, is the summit of mathematical biology, but it is ascended only occasionally, and such achievements are worthy to be celebrated. This article delves into the reasoning processes of five good examples, and shows, through them, several different ways that mathematical reasoning can enhance our understanding of biological systems. The five papers are selected as illustrations, without any attempt at a historical review of the place of each in its field.

Neurons with Excitatory Interactions can Oscillate in Phase Opposition

It is well known that neurons with excitatory interactions can synchronize each other, and that neurons with *inhibitory* interactions can entrain each other to oscillate with opposite phases, but what Kopell & Somers show is that excitatory coupling can lead to phase opposition as well [11]. They use singular perturbation theory to study a neuronal model of the relaxation oscillator type (but quite general in form), in which there is a fast-activating current

x and a slow-activating current y, governed by the equations $\varepsilon\dot{x} = F(x, y)$, $\dot{y} = G(x, y)$, the singular limit being taken as $\varepsilon \to 0$. The *nullclines* (loci in the x, y plane of $F = 0$ and $G = 0$) are assumed to be sigmoidal (for the slow current) and cubic (for the fast current). With x on the horizontal axis and y on the vertical axis, the cubic nullcline has three branches: left (descending), middle (ascending), and right (descending), and the two nullclines intersect along the middle branch of the cubic. When ε is small, the neuron's periodic trajectory descends the left branch to its minimum, jumps horizontally to the right branch, ascends the left branch to its maximum, and jumps horizontally back to the left branch. Mutual excitatory coupling between a pair of neurons is modeled by assuming that, when neuron 1 is on its right branch (high x), neuron 2's cubic nullcline is shifted upward on the y axis, and vice versa.

The fundamental condition that must be met for an antiphase solution is that the time to *ascend* the original cubic nullcline is less than the time to *descend* the shifted cubic nullcline. Kopell & Somers reparameterize the slow current (substituting z for y) in such a way that $\mathrm{d}z/\mathrm{d}t$ is constant and positive on the left branch of the unshifted cubic nullcline, $z = 0$ corresponding to the leftward jump point, and $z = 1$ to the minimum. Now suppose that neuron 1 starts just at the point of a jump to the right branch, while neuron 2 is at some $z \in (0, 1)$. Neuron 2 will immediately follow in jumping to the left branch of the shifted cubic nullcline, because neuron 1 is on its right branch. If the fundamental condition is satisfied, then there are z close enough to 0 that neuron 1 will jump back to the left branch before neuron 2 has completed its descent to the minimum, so neuron 2 will end up, after this pair of jumps, back on its original nullcline at a new position $E(z)$. Kopell & Somers prove that, if the fundamental condition holds, and in addition $E(0) > 0$ and $0 \leq E'(z) < 1$, then there is a stable antiphase solution. Furthermore, the antiphase solution persists in the nonsingular case $\varepsilon > 0$.

Kopell & Somers then show that the hypotheses of the theorem hold for several commonly used models of neuronal firing, including the Morris–Lecar equations. This system, which was first analyzed from a qualitative dynamics point of view by Rinzel & Ermentrout [17], is a simplified model of excitable membranes that captures many of their important features. It postulates voltage-gated Ca^{2+} (fast) and K^+ (slow) channels, and has nullclines on the (x, y) phase plane of the cubic and sigmoid types described above. The model has three equilibria [5]: a stable rest point, an unstable node, and a saddle point. The unstable manifold of the saddle point, which contains the rest point, forms an attracting invariant loop. The system generates action potentials in a realistic manner.

It turns out that it is possible for both antiphase and the more typical in-phase oscillations to be stable under the same excitatory coupling conditions. Arrays of bistable oscillators with nearest-neighbor coupling may show "fractured synchrony", i.e. domains within which activity is synchronous, while neighboring domains are in phase opposition [19, Figure 14].

Mutation Rates can be Estimated from Gene Polymorphisms

Mutation rates refer to the history of a population over an extended time, whereas gene **polymorphisms** refer to a cross-section at one time. Using the method of "coalescence", Kimmel & Chakraborty are able to make inferences about genetic history from the present state of the population [9]. Short segments of DNA are often repeated in the genome, and when the units of these "tandem repeats" are 2–6 nucleotides long, they are called "microsatellites". Mutations occur relatively frequently in the repeat length; they can cause expansion or contraction. Kimmel & Chakraborty use the Wright–Fisher model (*see* **Population Genetics**) for genetic evolution without selection in a population of constant size N, a model that ignores diploidy, and treats the $2N$ chromosomes of the $(k + 1)$th generation as if they were sampled uniformly and independently, with replacement, from the $2N$ chromosomes of the kth generation. That is equivalent to the number of copies of each chromosome in the $(k + 1)$th generation having a symmetric **multinomial distribution** [10]. Let each chromosome have a probability ν per generation of a mutation at a given locus (regarded as a **Poisson process**), and let the size U of that mutation (replacing an allele of size X by an allele of size $X + U$) be a **random variable** independent of the time of the mutation.

Under the Wright–Fisher model, any two chromosomes selected from the current generation, if followed backward in their parentage, eventually "coalesce", i.e. have a common parent, and the time T (going backward from the current generation) when coalescence occurs is approximately **exponentially distributed** with parameter $1/2N$, if N is large [10, p. 36].

Now imagine that two chromosomes are drawn at random from the current population, with repeat lengths X_i and X_j at a given locus. Since mutations can occur on either branch of the coalescence tree, the number n of mutations since the two chromosomes had a common ancestor is Poisson-distributed with parameter $2\nu T$. Since the probability **generating function** (pgf) for the **Poisson distribution** with parameter $2\nu t$ is $\exp[2\nu t(s-1)]$ and the exponential density at $T = t$ is $(1/2N)\exp(-t/2N)$, the pgf $\mu(s) = \mathrm{E}(s^n)$ for the number of mutation events is

$$\mu(s) = \frac{1}{2N}\int_0^\infty \exp\left(-\frac{t}{2N}\right)\exp[2\nu t(s-1)]\,\mathrm{d}t = \frac{1}{1-4N\nu(s-1)}.$$

So far, we have taken into account the random processes determining the time to coalescence and the number of mutations since coalescence, but not the random size of the mutation $X \to X + U$. Kimmel & Chakraborty [9] allow an arbitrary function $\varphi(s)$ for the pgf of the mutation size, since its probability distribution is currently an active area of research [18]. Since any mutation has an equal probability of affecting X_i or X_j, its effect on $X_i - X_j$ has pgf $\psi(s) = \frac{1}{2}[\varphi(s) + \varphi(1/s)]$. We now have a compound distribution for $X_i - X_j$, i.e. a random number n of mutations with pgf $\mu(s)$, each step contributing a random amount to the size difference with pgf $\psi(s)$. The pgf of $X_i - X_j$ is, therefore, found by composition:

$$\lambda(s) = \mu \circ \psi = \frac{1}{1-4N\nu[\psi(s)-1]}.$$

The variance of $X_i - X_j$ is found by differentiating $\lambda(s)$ twice and setting $s = 1$; it is $4\nu N\psi''$. This quantity may also be expressed as $4\nu N E(\hat{U}^2)$, by introducing the symmetrized random variable $\hat{U}$, with the probability distribution $\Pr(\hat{U} = n, n \in \mathbb{Z}) = \frac{1}{2}(p_n + p_{-n})$, where $p_n = \Pr(U = n)$. An empirically convenient measure of variability is the probability of homozygosity $\Pr(X_i = X_j)$, which is p_0 in the Laurent expansion of $\lambda(s) = \sum_{k\in\mathbb{Z}} p_k s^k$. p_0 can be evaluated by means of the Cauchy integral formula,

$$p_0 = \frac{1}{2\pi i}\oint_{|s|=1} \frac{\lambda(s)}{s}\,\mathrm{d}s.$$

The SIR Model for Epidemics has Chaotic Solutions

The existence of **chaos** in models for epidemics has been debated many times, and found in some computer simulations [15], but Glendinning & Perry [7] are able to give a definitive answer, at least in a simple case (the SIR model), using Melnikov's method. The SIR model for the spread of diseases is highly simplified, but captures important features of epidemics (*see* **Epidemic Models, Deterministic**). S stands for the proportion of *susceptible* individuals, I for *infected*, and R for *recovered*. The equations of the model are:

$$\dot{S} = -B(I,t)S + \mu - \mu S,$$
$$\dot{I} = B(I,t)S - (\gamma + \mu)I,$$
$$\dot{R} = \gamma I - \mu R,$$

where $S + I + R$, representing the total population, is set equal to 1. μ is the birth(= death) rate (all newborns are assumed susceptible); γ is the rate of recovery (transition to a permanently immune state). Glendinning & Perry [7] take $B(I,t) = \beta(t)I^2$, and assume that $\beta(t)$ has the form $\beta_0(1 + \beta_1 \sin\omega t)$. The sinusoidal term might arise from the annual school calendar, or from long-term cyclical variation in social or environmental factors. Choosing an exponent of $I > 1$ is not implausible, because there might be a threshold for the concentration of viruses in the environment to become infectious, or individuals could harbor low-level infections that increase susceptibility, without becoming infectious [13, p. 200]. The dynamics may be thought of as taking place on the cylinder $(I, R, t) \in \mathbb{R} \times \mathbb{R} \times S^1$, since the forcing term is periodic in t. Consider a Poincaré section at a fixed time (modulo $2\pi/\omega$). If S and R are given at a certain time in one period, then the equations predict what S and R will be at that time in the next period. Thus we have a map f of the (S, R)

plane into itself. "Chaos" can be defined in this way [24, Section 4.11 and Proposition 4.2.7]: there exists a compact and invariant set Λ in the Poincaré section such that:

1. Λ contains periodic points of all orders, and the periodic points are dense in Λ
2. there also exist, in Λ, an uncountably infinite number of points (S, R) such that an orbit started at (S, R) never repeats itself
3. there exists at least one starting point (S', R') in Λ from which the orbit comes arbitrarily close to every point in Λ
4. the Poincaré map has "sensitive dependence on initial conditions" on Λ, which means that $\exists\, \varepsilon > 0$ such that, for every $(S, R) \in \Lambda$, there are points arbitrarily close to (S, R) that eventually separate from (S, R) by at least ε.

This harvest of dynamics is reaped simply by proving that the map f has a saddle point y whose stable and unstable manifolds intersect transversely (i.e. they cross, other than at y) [24, Section 4.4]. (By the *stable manifold* is meant the set of points that approach y under successive iteration by f; by the *unstable manifold* is meant the set of points that approach y under backwards iteration by f. These will both be smooth curves.)

The key step is demonstrating that the stable and unstable manifolds cross. Melnikov's method [24, Section 4.5] deals with the case where the flow is governed by an arbitrarily small periodic perturbation of a Hamiltonian system: $\dot{q} = \partial H/\partial p$, $\dot{p} = -\partial H/\partial q$. The unperturbed system is assumed to have a *homoclinic orbit* $[q^0(t), p^0(t)](-\infty < t < \infty)$ such that $\lim_{t\to\infty} q^0(t) = y$ and $\lim_{t\to\infty} q^0(t) = y$. The interior of the homoclinic orbit is assumed to be filled with a continuous family of periodic orbits. Let $\vec{F}(q, p)$ be the unperturbed flow, and $\vec{G}(q, p, t, \varepsilon)$ be the perturbation, both smoothly varying in all their arguments. The Melnikov function is defined as

$$M(t) = \int_{-\infty}^{\infty} \vec{F}[q^0(s-t), p^0(s-t)] \times \vec{G}[q^0(s-t), p^0(s-t), t]\,\mathrm{d}s,$$

$\times$ standing for the vector cross-product. If $M(t_0) = 0$ for some t_0, and $\mathrm{d}M/\mathrm{d}t|_{t_0} \neq 0$, then the stable and unstable manifolds of y will intersect transversely for sufficiently small ε. Conversely, if $M(t)$ is always $\neq 0$, the stable and unstable manifolds will not intersect. The computational utility of Melnikov's formula is that the integral is carried out on the unperturbed orbits. After elaborate transformations, Glendinning & Perry are able to cast the equations of the SIR model into the form of a Hamiltonian unperturbed system and a perturbation. The Melnikov conditions are indeed satisfied, but only if the periodic term is on a very long time scale, not, for example, the annual scale that we expect in epidemics.

Hydrostatic Forces Determine the Geometry of the Aortic Valve

The objective of this theory was to derive the geometric form of the aortic valve of the heart from the hydrostatic forces acting on it when it balloons out to block the retrograde flow of blood into the heart [16]. The valve consists of three pockets, arranged as 120° sectors of a circle, meeting in the middle when the valve closes under retrograde flow. The fibers of each leaflet are suspended from the two points where the sector boundary meets the circumference, called *commissural points*. The intrinsic coordinates of the valve surface, (u, v), are defined so that the curves $v = \text{const.}$ are the fibers, $v = 0$ being the free edge. u measures arc length along the fibers, $u = 0$ marking the midline. The Cartesian coordinates in space, $\mathbf{X} = (x, y, z)$, are chosen so that the $z = 0$ plane contains the three commissural points (each pair suspending one of the valve leaflets), and $x = y = z = 0$ is the center of the circle through the three points. The equation of the leaflet surface, $\mathbf{X}(u, v)$, is regarded as unknown, to be determined by mechanical equilibrium under hydrostatic forces. $T(u, v)$ is the tension in the fibers, and p_0 is the pressure load applied to the leaflet, assumed uniform. The equations of equilibrium are

$$\frac{\partial}{\partial u}\left(T\frac{\partial \mathbf{X}}{\partial u}\right) + p_0\left(\frac{\partial \mathbf{X}}{\partial u} \times \frac{\partial \mathbf{X}}{\partial v}\right) = 0. \qquad (1)$$

Peskin & McQueen [16] prove that: (i) $T(u, v)$ is independent of u (the tension is constant along each fiber); (ii) the fibers are geodesics on the surface of the valve leaflet; and (iii) the u, v coordinate curves are orthogonal. By a change of variables $\mathrm{d}V/\mathrm{d}v = T(v)/p_0$, equalizing the force per unit of V, (1) transforms into

$$\frac{\partial \mathbf{X}}{\partial V} = \frac{\partial \mathbf{X}}{\partial u} \times \frac{\partial^2 \mathbf{X}}{\partial u^2}. \qquad (2)$$

Peskin & McQueen make use of a remarkable analogy between (2) and the equations of vortex dynamics in moving fluids. They think of V as a *time* variable, so that (2) can be regarded as describing the filling-out of the leaflet by a single fiber, sweeping across the leaflet as V increases, moving in the direction of its binormal at each point. Eq. (2) is the same as the "self-induction approximation" in hydrodynamics for the motion of a line vortex, i.e. its motion under the influence of the velocity field it itself generates. The geometry problem becomes an initial-value problem: starting with $\mathbf{X}(u, 0)$, defined at the free edge of the leaflet, propagate $\mathbf{X}(u, V)$ forward in the "time" variable V, toward the circumference where the leaflet is suspended. The free edge $\mathbf{X}(u, 0)$ looks like a hyperbola when projected onto the x, y plane, and like a cubic when projected onto the x, z plane. To carry out the computations, they use a method developed by Buttke [2] for approximating the motion of a line vortex in a three-dimensional incompressible, isentropic fluid. According to the numerical solution, the fibers are not uniformly spread over the leaflet, but are gathered in bundles formed by the rolling up of the leaflet surface, just as vortex lines tend to kink as they move in a fluid. This result was unexpected, but it agrees very well with the observed anatomy of the aortic valve leaflet. Peskin & McQueen describe their surprise at the degree of agreement between theory and observation:

> When this work was undertaken, our goal was to produce a smooth array of fibers that would function as an aortic valve ... We were aware of the complicated branching structure of the collagen fibers that support the actual valve, but we thought of such a structure as being "too biologic" to be modeled within the present framework...Imagine our astonishment, then, when the result first appeared on the workstation screen! These results show that considerations of mechanical equilibrium determine the anatomy of the aortic valve in a much more detailed way than we had dared to hope [16, p. H326].

Knot Topology Establishes the Mechanism of Tn3 Resolvase

Tn3 resolvase is an enzyme that catalyzes recombination of duplex DNA at specific sites, i.e. the cutting of both strands of two DNA molecules and reattachment of the cut ends of the first molecule to the cut ends of the second, and vice versa [20]. DNA-binding proteins form the template, or *synaptosome*, to which the recombining partners attach. The Tn3 resolvase, a representative example of a major family of these recombinases, acts on closed circular DNA. Its function is in DNA transposition, moving a segment of DNA from one position to another. It is called a *topoisomerase*, because its action changes the topology of the molecule. From the topologic changes it induces, Sumners was able, in a brilliant analysis, to deduce its mechanism of action [21]. He treated the DNA attached to the synaptosome according to Conway's theory of *rational tangles* [1, Section 2.3]. A tangle is a circular region in the projection plane of a knot or link, such that the knot or link crosses the circumference at exactly four points (called NW, NE, SW, SE). A few examples will clarify the idea. The (∞) tangle is just two vertical strings, and the (0) tangle is two horizontal strings. The (3) tangle is made by winding two horizontal strings around each other so that they make three left-handed twists. (In the case of DNA, such twists are "supercoiled", since the primary structure is already coiled.) To make a (3, 2) tangle out of a (3) tangle, first reflect the (3) tangle along the NW–SE diagonal, than make two twists of the free horizontal ends. The process can be continued to make tangles with more indices. The *sum* of two tangles is formed simply by joining the NE end of the first to the NW end of the second, and the SE end of the first to the SW end of the second. From a tangle, a knot or link can be formed by the "numerator construction", which consists of connecting the NW end to the NE end, and the SW end to the SE end. Conway proves that the *continued fraction* derived from a tangle, for example $n + 1/[m + (1/l)]$ from the tangle (l, m, n), characterizes the knot formed by the numerator construction, in the sense that two knots formed from tangles are equivalent if and only if their continued fractions have the same value.

Closely following the biology of the recombination process, Sumners assumes that the DNA strands on the synaptosome form a tangle in Conway's sense, and furthermore that the synaptosome tangle is the sum of two tangles, O_b and P, P representing the two "parental" segments, lying parallel on the synaptosome, that are to be cut and recombined, and O_b representing the rest of the (possibly twisted) DNA that is bound to the synaptosome, but not involved directly in recombination. The actual DNA molecule, before and after recombination, is assumed to be derived by

the numerator construction $N(\cdot)$ from the tangle. The original substrate, in Conway's notation, is $N(O_b + P)$, and it is topologically just a circle. After recombination, the product is $N(O_b + R)$. Whereas P is the (0) tangle (two parallel horizontal strands), R is either the (1) tangle or the (-1) tangle (one twist, formed when the strands are cut and recombine). Sumners takes advantage of the fact that occasionally two or three recombination events occur. Then, according to this model, the products should be $N(O_b + 2R)$ and $N(O_b + 3R)$, respectively. Since the topology of the products in each case is known experimentally, the equations can be solved for the tangles O_b and R. $N(O_b + R)$ is known to be a Hopf link (the simplest two-component link, two circles passed through one another), $N(O_b + 2R)$ is a figure of eight knot, and $N(O_b + 3R)$ is a Whitehead link (two circles joined so that one becomes a figure eight). Sumners proves, on the basis of this information, that R must consist of one left-handed twist, and O_b must consist of three left-handed twists in the vertical direction. In other words, the DNA is supercoiled on the synaptosome, in addition to being prepared for cutting. The *pièce de résistance* of this work is being able to predict, on the basis of the model derived from $N(O_b + jR)$, $j = 1, 2, 3$, what kind of knot $N(O_b + 4R)$ will be. Quadruple recombination is rare, but it does occur; the prediction is a knot called 6_2, a six-crossing composite knot, which agrees with experiment.

Conclusion

Each of these five studies illuminates one of the values of proving theorems in mathematical biology. Kopell & Somers [11] (first section) achieve two things. They predict a new phenomenon – oscillation in phase opposition by neurons coupled through excitatory interactions. They also exhibit the mechanism of that effect, through their analysis of the singular solution. Kimmel & Chakraborty's work [9] on estimation of mutation rate from microsatellite polymorphisms (second section) shows how mathematical reasoning can connect two qualitatively distinct phenomena, in this case one stretched out over the history of the population, and the other observed in a cross-section at a single time. Glendinning & Perry's study [7] of chaos in the SIR model for epidemics (third section) is important, not because the chaotic regime is expected to occur under typical circumstances, but because it resolves a question about dynamics that no amount of computer simulation could ever settle, being necessarily confined to a finite time period, a finite number of starting points, and finite precision. Peskin & McQueen's demonstration [16] that equilibrium under hydrostatic forces can account for the geometry of the aortic valve (fourth section), apart from the fascination of its reasoning, shows how isomorphism in formal structure sometimes makes it possible for a large body of mathematical knowledge to be translated and applied in a new area. In this case, methods from vortex hydrodynamics could be used to study the equations of shape of a body in equilibrium. Finally, Sumners' elegant application [21] of knot theory in DNA biochemistry (fifth section) illustrates how mathematical proof can, at least occasionally, give a definitive answer to a question of mechanism. Each of the selected papers also illustrates the crucial step of formulating a model rich enough to capture the phenomena, yet simple enough to be tractable. Mathematical analysis in biology will remain a subtle art, but when theorems can be proved about realistic models, they are likely to shed valuable light on biologic mechanisms.

References

[1] Adams, C.C. (1994). *The Knot Book*. W.H. Freeman, New York.

[2] Buttke, T.F. (1988). A numerical study of superfluid turbulence in the self-induction approximation, *Journal of Computational Physics* **76**, 301–326.

[3] Committee on the Mathematics and Physics of Emerging Dynamic Biomedical Imaging. (1996). *Mathematics and Physics of Emerging Biomedical Imaging*. National Academy Press, Washington.

[4] Edelstein-Keshet, L. (1988). *Mathematical Models in Biology*. Random House, New York.

[5] Ermentrout, G.B. & Rinzel, J. (1996). Reflected waves in an inhomogeneous excitable medium, *SIAM Journal of Applied Mathematics* **56**, 1107–1128.

[6] Fagerström, T., Jagers, P., Shuster, P. & Szathmary, E. (1996). Biologists put on mathematical glasses, *Science* **274**, 2039–2040.

[7] Glendinning, P. & Perry, L.P. (1997). Melnikov analysis of chaos in a simple epidemiological model, *Journal of Mathematical Biology* **35**, 359–373.

[8] Hoppensteadt, F.C. & Peskin, C.S. (1992). *Mathematics in Medicine and the Life Sciences*. Springer-Verlag, New York.

[9] Kimmel, M. & Chakraborty, R. (1996). Measures of variation at DNA repeat loci under a general stepwise mutation model, *Theoretical Population Biology* **50**, 345–367.

[10] Kingman, J.F.C. (1982). On the genealogy of large populations, *Journal of Applied Probability* **19A**, 27–43.

[11] Kopell, N. & Somers, D. (1995). Anti-phase solutions in relaxation oscillators coupled through excitatory interactions, *Journal of Mathematical Biology* **33**, 261–280.

[12] Lander, E.S. & Waterman, M.S., eds. (1995). *Calculating the Secrets of Life*. National Academy Press, Washington.

[13] Liu, W.-M., Levin, S.A. & Iwasa, Y. (1986). Influence of nonlinear incidence rates upon the behavior of SIRS epidemiological models, *Journal of Mathematical Biology* **23**, 187–204.

[14] Murray, J.D. (1993). *Mathematical Biology*, 2nd Ed. Springer-Verlag, Berlin.

[15] Olsen, L.T. & Schaffer, W.M. (1990). Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics, *Science* **249**, 499–504.

[16] Peskin, C.S. & McQueen, D.M. (1994). Mechanical equilibrium determines the fractal fiber architecture of aortic heart valve leaflets, *American Journal of Physiology* **266**, H319–H328.

[17] Rinzel, J. & Ermentrout, G.B. (1989). Analysis of neural excitability and oscillations, in *Methods in Neuronal Modeling: From Synapses to Networks*, C. Koch & I. Seger, eds. MIT Press, Cambridge, Mass., pp. 135–169.

[18] Rubinsztein, D.C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S.-H., Margolis, R.L., Ross, C.A. & Ferguson-Smith, M.A. (1995). Microsatellite evolution – evidence for directionality and variation in rate between species, *Nature Genetics* **10**, 337–343.

[19] Somers, D. & Kopell, N. (1995). Waves and synchrony in networks of oscillators of relaxation and non-relaxation type, *Physica D* **89**, 169–183.

[20] Stark, W.M., Boocock, M.R. & Sherratt, D.J. (1992). Catalysis by site-specific recombinases, *Trends in Genetics* **8**, 432–439.

[21] Sumners, D.L. (1992). Knot theory and DNA, *Proceedings of Symposia in Applied Mathematics* **45**, 39–72.

[22] Tan, W.-Y. (1991). *Stochastic Models of Carcinogenesis*. Marcel Dekker, New York.

[23] Waterman, M.S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall, London.

[24] Wiggins, S. (1990). *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer-Verlag, New York.

[25] Winfree, A.T. (1987). *When Time Breaks Down*. Princeton University Press, Princeton.

STEVEN MATTHYSSE

Matrix Algebra

The algebra that we learn when teenagers has letters of the alphabet, each representing a number. For example: a father and son are x and y years old, respectively, and their total age is 70. In a decade the father will be twice as old as the son. Hence $x + y = 70$ and $x + 10 = 2(y + 10)$, and so $x = 50$ and $y = 20$.

In contrast, matrix algebra is the algebra of letters each representing many numbers, with those numbers always arrayed in the form of a rectangle (or square). An example is

$$\mathbf{X} = \begin{bmatrix} 9 & 0 & 7 & t \\ u^2 + v & -3 & 6.1 & 5^3 \end{bmatrix}.$$

General Description

A *matrix* is a rectangular array of numbers, which can be any mixture of numbers that are complex, real, zero, positive, negative, decimal, fractions, or algebraic expressions. When none of them is complex (i.e. involving $\sqrt{-1}$), the matrix is said to be *real*. And because statistics deals with data, which are real numbers (especially biological data), almost all of this article applies to real matrices. Each number in a matrix is called an *element*: in being some representation of a single number it is called a *scalar*, to contrast with a matrix which represents many numbers.

Elements are always set out in rows and columns with the number of rows and columns being called the *order* (or *dimension*) of the matrix. Thus, the illustrated $\mathbf{X}$ has order 2×4 ("two by four") with the number of rows being mentioned first. Sometimes the order is used as a subscript to the matrix symbol; for example, $\mathbf{X}_{2\times4}$. In this encyclopedia the widespread custom is used of denoting matrices by bold face, capital, roman letters.

Elements of a matrix can be represented by letters, having subscripts to denote location (row and column) in the matrix. Thus, a matrix $\mathbf{A}$ might be represented as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

The first subscript indicates row, and the second column; for example, a_{23} is in row 2 and column 3. More briefly, we can write

$$\mathbf{A} = \{a_{ij}\} \quad \text{for } i = 1, 2, 3 \text{ and } j = 1, 2, 3.$$

When $\mathbf{B}$ has r rows and c columns,

$$\mathbf{B} = \{b_{ij}\} \quad \text{for } i = 1, 2, \ldots, r \text{ and } j = 1, 2, \ldots, c.$$

A more compact form is

$$\mathbf{B} = \left\{{}_{\mathrm{m}}\ b_{ij}\right\}_{i=1,\,j=1}^{r\quad\ c},$$

the m indicating that it is a matrix. The element in the first row and first column (e.g. a_{11} in $\mathbf{A}$ and the 9 in $\mathbf{X}$) is called the *leading element*.

By virtue of a matrix being a rectangular array there are many special forms, the first two of which are square matrices and vectors.

Square Matrices

1. Square matrices have the same number of rows as columns. $\mathbf{A}$ is an example.
2. Elements on the diagonal from upper left to lower right, those with both subscripts the same, are *diagonal elements*; they constitute *the diagonal of the matrix*.
3. Elements immediately below the diagonal constitute the *sub-diagonal*.
4. Elements not on the diagonal are *off-diagonal elements*.
5. When all off-diagonal elements are zero, and at least some diagonal elements are nonzero, the matrix is a *diagonal matrix*.
6. When all elements below (above) the diagonal are zero the matrix is said to be *upper (lower) triangular*.

Vectors

When a matrix has only one column it is a *column vector* or, more usually, just *vector*; and it shall here be denoted by a bold face, lower case, roman letter, usually from the last part of the alphabet; for example,

$$\mathbf{x} = \begin{bmatrix} 1 \\ 7 \\ -4 \\ 0 \end{bmatrix}.$$

When a matrix has only one row it is called a *row vector*. The notation is similar to that for a column vector, except for a superscript prime:

$$\mathbf{y}' = [0 \quad -4 \quad 9 \quad 12 \quad 37].$$

A column vector is a matrix of order $r \times 1$ when it has r elements; its transpose, a row vector, has order $1 \times r$. For both vectors, r is often called the *order* of the vector.

Basic Operations

A minimal requirement for matrix algebra is to define the arithmetic operations. Moreover, the rectangular nature of matrices begets numerous operations that do not exist for scalars; for example, changing rows into columns, and columns into rows.

The Transpose of a Matrix

Changing $\mathbf{A}$ so that its rows become columns (and hence its columns become rows) gives a matrix called the *transpose* of $\mathbf{A}$, traditionally written as $\mathbf{A}'$ (and sometimes today as $\mathbf{A}^{\mathrm{T}}$). Thus for

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 6 & 1 & -2 & 5 \end{bmatrix}, \qquad \mathbf{A}' = \begin{bmatrix} 1 & 6 \\ 2 & 1 \\ 3 & -2 \\ 4 & 5 \end{bmatrix}.$$

Note that the transpose of $\mathbf{A}'$ is $\mathbf{A}$: $(\mathbf{A}')' = \mathbf{A}$. Also, the transpose of a column vector is a row vector (and vice versa):

$$[1 \quad 2 \quad 3]' = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

This explains the use of $\mathbf{y}'$ at the end of the preceding section.

Partitioned Matrices

The rows and columns of a matrix can be partitioned into a representation that is a matrix of matrices of smaller orders:

$$\mathbf{K} = \left[\begin{array}{cc|cc} 1 & 2 & 3 & 4 \\ 6 & 8 & 4 & 0 \\ 9 & 8 & 1 & 2 \\ \hline 6 & 8 & 3 & 9 \\ 4 & 1 & 6 & 1 \end{array}\right] = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}, \quad \text{for}$$

$$\mathbf{K}_{11} = \begin{bmatrix} 1 & 2 \\ 6 & 8 \\ 9 & 8 \end{bmatrix},$$

and so on. $\mathbf{K}$ is a *partitioned* matrix; the $\mathbf{K}$s with subscripts are *submatrices* of $\mathbf{K}$.

In transposing a partitioned matrix, not only is the matrix of submatrices transposed, but each submatrix is also transposed. Thus

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}' = \begin{bmatrix} \mathbf{A}' & \mathbf{C}' \\ \mathbf{B}' & \mathbf{D}' \end{bmatrix}.$$

A matrix can also be partitioned into its columns (or its rows); for example,

$$\mathbf{K} = [\mathbf{k}_1 \quad \mathbf{k}_2 \quad \mathbf{k}_3 \quad \mathbf{k}_4],$$

where each of the subscripted $\mathbf{k}$s is a column of $\mathbf{K}$.

The Trace of a Matrix

The trace of a matrix is defined only for a square matrix; and *trace* of $\mathbf{A}$ is the sum of the diagonal elements of $\mathbf{A}$, often written as $\mathrm{tr}(\mathbf{A})$. Note that $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}')$, and $\mathrm{tr}(\text{scalar}) = \text{scalar}$.

Addition and Subtraction

Addition and subtraction are defined only for matrices of the same order, whereupon the matrices are said to be *conformable* for addition and subtraction. Then, for $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$,

$$\mathbf{A} + \mathbf{B} = \{a_{ij} + b_{ij}\}.$$

If two matrices do not have the same order their sum and difference do not exist. Note the properties

$$(\mathbf{A} \pm \mathbf{B})' = \mathbf{A}' \pm \mathbf{B}'$$

and

$$\mathrm{tr}(\mathbf{A} \pm \mathbf{B}) = \mathrm{tr}(\mathbf{A}) \pm \mathrm{tr}(\mathbf{B}).$$

Scalar Multiplication

For λ being a scalar, $\lambda\mathbf{A}$ is $\mathbf{A}$ with every element multiplied by λ. Thus, for $\mathbf{A} = \{a_{ij}\}$, $\lambda\mathbf{A} = \{\lambda a_{ij}\}$.

Equality and Null Matrices

Two matrices are equal only when they are equal element by element. Thus, for

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 6 & 8 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 6 & 8 \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} 1 & 2 \\ 5 & 8 \end{bmatrix},$$

$\mathbf{A} = \mathbf{B}$, but $\mathbf{A} \neq \mathbf{C}$. Furthermore,

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} 1-1 & 2-2 \\ 6-6 & 8-8 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{0}.$$

Any matrix having every element zero is a *null matrix*. It is a zero of matrix algebra: note that it is *a* zero not *the* zero, because null matrices can be of any order.

Multiplication

Multiplication of matrices differs greatly from that of scalars. First of all, **AB** and **BA** can, and often do, differ. To distinguish between the two, **AB** is described as **B** *pre-multiplied* by **A** (or as **A** *post-multiplied* by **B**).

The *inner product* of two vectors is a row vector post-multiplied by a column vector, with both vectors having the same number of elements; for example,

$$[1 \quad 7 \quad 2] \begin{bmatrix} 3 \\ 5 \\ 9 \end{bmatrix} = 1(3) + 7(5) + 2(9) = 56.$$

Thus for $\mathbf{x}' = \{x_i\}_{i=1}^{n}$ and $y' = \{y_i\}_{i=1}^{n}$,

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

In contrast, an *outer product* is a column vector post-multiplied by a row vector

$$\mathbf{x}\mathbf{y}' = \{_{\mathrm{m}}\ x_i y_j\}.$$

In this case, the vectors can be of different orders. Note that an inner product is a scalar, whereas an outer product is a matrix.

The product **AB** exists only when **A** has as many columns as **B** has rows; and then **A** and **B** are described as being *conformable for the product* **AB**, whereupon

$$\mathbf{A}_{r\times c}\mathbf{B}_{c\times t} = \mathbf{P}_{r\times t}.$$

In **P**, the element in row i and column j is the inner product of row i of **A** and column j of **B**:

$$\mathbf{P}_{r\times t} = \{p_{ij}\} = \left\{ \sum_{k=1}^{c} a_{ik} b_{kj} \right\} \quad \text{for } i = 1, \ldots, r$$

and $j = 1, \ldots, t$.

A simple numerical example of this is

$$\begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & 4 & 7 \\ -5 & 6 & 8 \end{bmatrix}$$

$$= \begin{bmatrix} [1 \quad 0] \begin{bmatrix} 3 \\ -5 \end{bmatrix} & [1 \quad 0] \begin{bmatrix} 4 \\ 6 \end{bmatrix} & [1 \quad 0] \begin{bmatrix} 7 \\ 8 \end{bmatrix} \\ [2 \quad -1] \begin{bmatrix} 3 \\ -5 \end{bmatrix} & [2 \quad -1] \begin{bmatrix} 4 \\ 6 \end{bmatrix} & [2 \quad -1] \begin{bmatrix} 7 \\ 8 \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} 1(3) + 0(-5) & 1(4) + 0(6) & 1(7) + 0(8) \\ 2(3) + (-1)(-5) & 2(4) + (-1)6 & 2(7) + (-1)8 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 4 & 7 \\ 11 & 2 & 6 \end{bmatrix}.$$

Important consequences of this definition of multiplication are that **AB** exists only for $\mathbf{A}_{r\times c}$ and $\mathbf{B}_{c\times t}$; both **AB** and **BA** exist only for $\mathbf{A}_{r\times c}$ and $\mathbf{B}_{c\times r}$, but they will be of different orders (and so not equal) unless $r = c$. Even then, **AB** and **BA** are not necessarily equal. For example,

$$\begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ -5 & 6 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 11 & 2 \end{bmatrix},$$

$$\text{but } \begin{bmatrix} 3 & 4 \\ -5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix} = \begin{bmatrix} 11 & -4 \\ 7 & -6 \end{bmatrix}.$$

Products with Null Matrices

Every product of a matrix with a null matrix is a null matrix: but those null matrices are not necessarily of the same order. Thus, $\mathbf{0}_{3\times 2}\mathbf{A}_{2\times 5} = \mathbf{0}_{3\times 5}$ and $\mathbf{A}_{2\times 5}\mathbf{0}_{5\times 6} = \mathbf{0}_{2\times 6}$.

Products with Diagonal Matrices

Pre-(post-)multiplying **A** by a diagonal matrix **D** multiplies each row (column) of **A** by the corresponding diagonal element of **D**.

Identity Matrices

If every diagonal element of a diagonal matrix is a one the matrix is called an *identity* matrix, **I**; pre- or

post-multiplication of $\mathbf{A}$ by an identity matrix yields $\mathbf{A}$. Thus $\mathbf{I}$-matrices are the unities of matrix algebra.

Transposing a Product

The transpose of a product is the product of the transposed matrices in reverse order. Thus

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \quad \text{and} \quad (\mathbf{XAY})' = \mathbf{Y}'\mathbf{A}'\mathbf{X}'.$$

Trace of a Product

The trace of a product equals the trace of cyclic permutations of that product: $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ and $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB}) = \text{tr}(\mathbf{BCA})$, but these three do not equal the trace of $\mathbf{ACB}$.

Powers of Matrices

Only square matrices have powers: $\mathbf{A}_{2\times 4}\mathbf{A}_{2\times 4}$ does not exist. $\mathbf{A}_{4\times 4}\mathbf{A}_{4\times 4}$ written as $\mathbf{A}^2_{4\times 4}$ does.

Hadamard Products

The (i, j)th element of $\mathbf{AB}$ is $\sum_k a_{ik}b_{kj}$. But there are other ways of defining a product. One is the *Hadamard product*, defined as

$$\mathbf{A} \cdot \mathbf{B} = \{a_{ij}b_{ij}\}.$$

Thus, the (i, j)th element of the Hadamard product is the product of the (i, j)th elements of $\mathbf{A}$ and $\mathbf{B}$ – which must have the same order.

Direct Products

There is also the direct product

$$\mathbf{A} \otimes \mathbf{B} = \{_{\text{m}}\ a_{ij}\mathbf{B}\}.$$

When $\mathbf{A}$ has order $p \times q$ and $\mathbf{B}$ has order $r \times s$, $\mathbf{A} \otimes \mathbf{B}$ has order $pr \times qs$. For example,

$$\begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} \otimes [6 \quad 7 \quad 8] = \begin{bmatrix} 3[6 & 7 & 8] & 4[6 & 7 & 8] \\ 2[6 & 7 & 8] & 1[6 & 7 & 8] \end{bmatrix}$$
$$= \begin{bmatrix} 18 & 21 & 24 & 24 & 28 & 32 \\ 12 & 14 & 16 & 6 & 7 & 8 \end{bmatrix}.$$

Laws of Algebra

Provided that conformability requirements are met, the following equalities hold:

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + \mathbf{B} + \mathbf{C},$$
$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) = \mathbf{ABC},$$
$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC},$$
$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}.$$

This last equality is the commutative law of addition. In contrast, its mate, the commutative law of multiplication, does not generally hold for matrices; that is, $\mathbf{AB}$ and $\mathbf{BA}$ are not usually equal. Indeed, there are situations in which one exists and the other does not; and when they do both exist they can be of different orders; and even when they both exist and are of the same order (for which $\mathbf{A}$ and $\mathbf{B}$ must be square and of the same order) they are not necessarily equal.

Contrasts with Scalar Algebra

The following results illustrate differences in the algebra of matrices compared with that of scalars:

1. $\mathbf{AX} + \mathbf{BX} = (\mathbf{A} + \mathbf{B})\mathbf{X} \neq \mathbf{X}(\mathbf{A} + \mathbf{B})$;
2. $\mathbf{XP} + \mathbf{QX}$ does *not* have $\mathbf{X}$ as a factor;
3. $\mathbf{AB} = \mathbf{0}$ does not imply that $\mathbf{A}$ or $\mathbf{B}$ are $\mathbf{0}$, nor does it imply that $\mathbf{BA}$ is $\mathbf{0}$;
4. $\mathbf{Y}^2 = \mathbf{0}$ defines $\mathbf{Y}$ as *nilpotent* and does *not* imply that $\mathbf{Y}$ is $\mathbf{0}$
5. $\mathbf{Z}^2 = \mathbf{I}$ does *not* imply that $\mathbf{Z}$ is $\pm\mathbf{I}$;
6. $\mathbf{Q}^2 = \mathbf{Q}$ defines $\mathbf{Q}$ as *idempotent* without implying that $\mathbf{Q} = \mathbf{0}$ or $\mathbf{I}$.

Examples of these last four features are as follows:

$$\mathbf{AB} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} = \mathbf{0},$$
$$\mathbf{BA} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ -2 & -2 \end{bmatrix},$$
$$\mathbf{Y}^2 = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}^2 = \mathbf{0},$$
$$\mathbf{Z}^2 = \begin{bmatrix} 1 & 0 \\ 4 & -1 \end{bmatrix}^2 = \mathbf{I},$$

and

$$\mathbf{Q}^2 = \begin{bmatrix} 3 & -2 \\ 3 & -2 \end{bmatrix}^2 = \mathbf{Q}.$$

One may be tempted to think of these examples as pathologic cases. To some extent they are, born of the need to have illustrations that occupy minimum space; but they serve as stern warnings that what can be done in scalar algebra does not always carry over to matrix algebra.

Special Matrices

Square matrices and vectors have already been mentioned as special forms of matrices. There are many others, some arising from their intrinsic properties, and others from the applications in which they arose. Just a few of the more commonly occurring ones are mentioned here.

Symmetric Matrices

$\mathbf{A}$ is defined as being symmetric when $\mathbf{A}' = \mathbf{A}$. That can occur only when $\mathbf{A}$ is square. Its rows are then mirror images of its columns:

$$\begin{bmatrix} 1 & 7 & 0 \\ 7 & 2 & -3 \\ 0 & -3 & 9 \end{bmatrix}' = \begin{bmatrix} 1 & 7 & 0 \\ 7 & 2 & -3 \\ 0 & -3 & 9 \end{bmatrix};$$

and $a_{ij} = a_{ji}$.

$\mathbf{BB}'$ and $\mathbf{B}'\mathbf{B}$ are both symmetric. This is true for any $\mathbf{B}$. Then $\mathbf{BB}'$ (and $\mathbf{B}'\mathbf{B}$) have diagonal elements that are sums of squares of elements of rows (columns) of $\mathbf{B}$: and

$$\text{tr}(\mathbf{BB}') = \text{tr}(\mathbf{B}'\mathbf{B}) = \sum_i \sum_j b_{ij}^2.$$

When $\mathbf{B}$ is real, $\mathbf{BB}' = \mathbf{0}$ and $\text{tr}(\mathbf{BB}') = 0$ each imply that $\mathbf{B} = \mathbf{0}$.

Elementary Vectors

Columns of identity matrices are *elementary vectors*, represented as $e_i^{(n)}$, the ith column in $\mathbf{I}$ of order n.

Skew-Symmetric Matrices

$\mathbf{A}' = -\mathbf{A}$ defines $\mathbf{A}$ as *skew-symmetric*.

Summing Vectors

A vector having every element a one (1.0) is a *summing vector*, often denoted as $\mathbf{1}$. It is so named because $\mathbf{1}'\mathbf{x}$ is the sum of all elements in $\mathbf{x}$.

Matrices having Every Element Unity

$\mathbf{J}_{p\times k} = \mathbf{1}_p\mathbf{1}_k'$ is a matrix having every element being 1.0. Its most frequent occurrence in statistics is when it is square: $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n'$. A useful variant is $\overline{\mathbf{J}}_\mathbf{n} = (1/n)\mathbf{J}_n$. Then,

$$\mathbf{C}_n = \mathbf{I}_n - \overline{\mathbf{J}}_n$$

is a *centering matrix*, with

$$\mathbf{C}_n\mathbf{x} = \{x_i - \overline{x}\} \quad \text{and} \quad \mathbf{x}'\mathbf{C}_n\mathbf{x} = \sum_{i=1}^{n}(x_i - \overline{x})^2,$$

for $\overline{x} = \sum_{i=1}^n x_i/n = \mathbf{1}'\mathbf{x}/n$.

Probability Transition Matrices

When elements of a matrix $\mathbf{P}$ are probabilities that add to unity over each row, $\mathbf{P1} = \mathbf{1}$. Then $\mathbf{P}^k\mathbf{1} = \mathbf{1}$ for any positive integer k, and $\mathbf{P}$ is called a *probability transition matrix*. It is *doubly stochastic* if $\mathbf{1}'\mathbf{P} = \mathbf{1}'$ (or, equivalently, $\mathbf{P}'\mathbf{1} = \mathbf{1}$), meaning that column sums are also unity.

Idempotent Matrices

$\mathbf{A}$ is *idempotent* when $\mathbf{A}^2 = \mathbf{A}$; then $\mathbf{I} - \mathbf{A}$ is also idempotent (but $\mathbf{A} - \mathbf{I}$ is not).

Orthogonality

1. The *norm* of a real vector $\mathbf{x}$ is $(\mathbf{x}'\mathbf{x})^{1/2}$.
2. $\mathbf{x}$ is a *unit vector* when $\mathbf{x}'\mathbf{x} = 1$.
3. $\mathbf{u} = \mathbf{x}/(\mathbf{x}'\mathbf{x})^{1/2}$, known as normalized $\mathbf{x}$, is always a unit vector when $\mathbf{x}$ is real.

Nonnull vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal vectors* when $\mathbf{x}'\mathbf{y} = 0$ $(= \mathbf{y}'\mathbf{x})$ (*see* **Orthogonality**).

Vectors $\mathbf{v}$ and $\mathbf{w}$ are *orthonormal vectors* when they are orthogonal $(\mathbf{v}'\mathbf{w} = 0)$ and each is a unit vector $(\mathbf{v}'\mathbf{v} = 1$ and $\mathbf{w}'\mathbf{w} = 1)$.

A collection of vectors of the same order is said to be an *orthogonal set of vectors* when they are pairwise orthonormal.

When $\mathbf{P}_{r\times c}$ has rows that are an orthonormal set, $\mathbf{PP}' = \mathbf{I}$. If $\mathbf{P}$ is square with orthonormal rows (columns), then its columns (rows) are also orthonormal, $\mathbf{PP}' = \mathbf{I} = \mathbf{P}'\mathbf{P}$ and $\mathbf{P}$ is an *orthogonal matrix*.

Certain special forms of orthogonal matrices go by the names Helmert, Givens, and Householder. The latter, for example, is $\mathbf{I} - \mathbf{2hh}'$ when $\mathbf{h}'\mathbf{h} = 1$.

Quadratic Forms

$\mathbf{x}'\mathbf{A}\mathbf{x}$ is a *quadratic form*, in which $\mathbf{A}$ can always be (taken as) symmetric. $\mathbf{x}'\mathbf{A}\mathbf{x}$ is a homogeneous second-order function of the elements of $\mathbf{x}$:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_i x_i^2 a_{ii} + \sum x_i x_j a_{ij} = \sum_i x_i^2 a_{ii} + \sum_{j \neq i} x_i x_j (a_{ij} + a_{ji}),$$

and on taking $\mathbf{A} = \mathbf{A}'$, that is, $a_{ij} = a_{ji}$,

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_i x_i^2 a_{ii} + 2 \sum_{j>i} x_i x_j a_{ij}.$$

If $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}'\mathbf{A}\mathbf{x}$ is called a *positive definite (p.d.) quadratic form*, and $\mathbf{A}$ $(= \mathbf{A}')$ is a *p.d. matrix*. If $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$ *and* $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{0}$ for some $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{A}$ are *positive semidefinite* (p.s.d.). The classes of quadratic forms and matrices that include those which are p.d. and p.s.d. are called *nonnegative definite* (n.n.d.).

Determinants

Definition

Associated with any square matrix $\mathbf{A}_{n \times n}$ is its determinant $|\mathbf{A}|$. It is a scalar, an n-order, homogeneous polynomial function of the elements. Two easy examples are for $\mathbf{A}$ of order 2 and 3:

$$|\mathbf{X}| = \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} = a_1 b_2 - a_2 b_1$$

and

$$|\mathbf{Y}| = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = a_1 b_2 c_3 + a_2 b_3 c_1 + a_3 b_1 c_2 - a_3 b_2 c_1 - a_1 b_3 c_2 - a_2 b_1 c_3.$$

For $\mathbf{A}$ of order n, the definition is more difficult: $|\mathbf{A}|$ is the sum of the n different terms that are each a signed product of one element from every row and column of $\mathbf{A}$. In writing $|\mathbf{A}|$ with the rows being $\mathbf{a}', \mathbf{b}', \mathbf{c}', \ldots$, a product written in alphabetic order has sign equal to $(-1)^p$, with p being the sum of the number of reverse sequences of the subscripts. For example, $a_2 b_3 c_1$ in the preceding $|\mathbf{Y}|$ has $p = 2$ because 2, 1 and 3, 1 are reverse sequences; hence, the sign for $a_2 b_3 c_1$ is $(-1)^2 = +1$. For $a_3 b_2 c_1$ there are three reverse sequences, 3, 2 and 3, 1 and 2, 1 and so the sign is $(-1)^3 = -1$.

Minors and Cofactors

Deleting from $|\mathbf{A}|$ the row and column containing a_{ij} leaves a determinant of order $n - 1$ that is called the *minor*, $|\mathbf{M}_{ij}|$, of a_{ij} in $|\mathbf{A}|$. Also $(-1)^{i+j}|\mathbf{M}_{ij}|$, the *signed minor*, is called the *cofactor* of a_{ij} in $|\mathbf{A}| : c_{ij} = (-1)^{i+j}|\mathbf{M}_{ij}|$. Then

$$|\mathbf{A}| = \sum_{i=1}^{n} a_{ij} c_{ij} \quad \text{for all } j \quad \text{and}$$

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij} c_{ij} \quad \text{for all } i,$$

but

$$0 = \sum_{i=1}^{n} a_{ij} c_{ij'} \quad \text{for all } j \neq j' \quad \text{and}$$

$$0 = \sum_{j=1}^{n} a_{ij} c_{i'j} \quad \text{for all } i \neq i'.$$

Calculation

Computers now handle the calculation of determinants. Numerous available shortcuts and associated properties of determinants are detailed in the literature, which was especially rich on this subject up through the 1930s. Searle [4] deals with a few of these topics.

Some Properties Useful for Statistics

1. $|\mathbf{A}'| = |\mathbf{A}|$.
2. $|\mathbf{A}^k| = (|\mathbf{A}|)^k$, for integer k.
3. $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$.
4. $|\mathbf{A}| = +1$, for orthogonal $\mathbf{A}$ (i.e. $\mathbf{A}'\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}'$).
5. $|\mathbf{A}| = 0$, for idempotent $\mathbf{A}$ (i.e. $\mathbf{A}^2 = \mathbf{A}$), except for $\mathbf{A} = \mathbf{I}$.
6. $|\mathbf{I}| = 1$.
7. $|\lambda \mathbf{A}_{n \times n}| = \lambda^n |\mathbf{A}|$, for scalar λ.

Inverse Matrices

Existence

In matrix arithmetic, the very definition of multiplication precludes any obvious definition of division. Indeed, there is no such thing as matrix division; *division by a matrix does not exist*. Instead, multiplication by an inverse matrix is used, similar to the scalar equivalence of dividing by six (for example) being identical to multiplying by $1/6 = 6^{-1}$, the inverse of six: and

$$(6^{-1})6 = 1 = 6(6^{-1}).$$

There is one big difference: whereas every scalar different from zero has an inverse, not every nonnull matrix does.

Suppose that $\mathbf{A}$ has an inverse. Denote it by $\mathbf{A}^{-1}$, as is customary. Then with $\mathbf{I}$ being a "one" of matrix algebra, the matrix analogy of scalars is $(\mathbf{A}^{-1})\mathbf{A} = \mathbf{I} = \mathbf{A}(\mathbf{A}^{-1})$, where the parentheses are solely for emphasis, the usual writing being

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}.$$

This requirement demands that two conditions must be satisfied in order for $\mathbf{A}^{-1}$ to exist:

(i) $\mathbf{A}$ must be square;
(ii) $|\mathbf{A}| \neq 0$.

If either or both (i) and (ii) are not satisfied, $\mathbf{A}$ has no inverse; note, particularly, that every nonsquare matrix has no inverse.

When, for $\mathbf{A}$ square, $|\mathbf{A}| \neq 0$, $\mathbf{A}$ is called *nonsingular*, and if $|\mathbf{A}| = 0$ then $\mathbf{A}$ is called *singular*.

Form

The general form of $\mathbf{A}^{-1}$ is

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\left[\begin{array}{c}\text{the matrix that is } \mathbf{A} \\ \text{with every element} \\ \text{replaced by its cofactor}\end{array}\right]^{\text{transposed}}$$

and $|\mathbf{A}|\mathbf{A}^{-1}$ is called the *adjugate* or *adjoint* of $\mathbf{A}$.

Some Basic Properties

1. $\mathbf{A}^{-1}$ is unique (for given $\mathbf{A}$).
2. $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$.
3. $\mathbf{A}^{-1}$ is nonsingular.
4. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
5. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.
6. $\mathbf{A}' = \mathbf{A} \Rightarrow (\mathbf{A}^{-1})' = \mathbf{A}^{-1}$.
7. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

In all of these results, and whenever an inverse is used, one must always be certain that the matrix satisfies (i) and (ii) above; namely, squareness and nonzero determinant.

Four Special Cases

Denote a diagonal matrix having all its diagonal elements $\lambda_1, \ldots, \lambda_n$ nonzero by

$$\mathbf{D} = \{_{\mathrm{d}}\ \lambda_i\}_{i=1}^{n};$$

then,

$$\mathbf{D}^{-1} = \{_{\mathrm{d}}\ 1/\lambda_i\}_{i=1}^{n}, \quad \mathbf{I}^{-1} = \mathbf{I},$$

$$(a\mathbf{I}_n + b\mathbf{J}_n)^{-1} = \frac{1}{a}\left(\mathbf{I}_n - \frac{b}{a+nb}\mathbf{J}_n\right),$$

$$\mathbf{PP}' = \mathbf{I} = \mathbf{P}'\mathbf{P} \quad \text{implies} \quad \mathbf{P}^{-1} = \mathbf{P}'.$$

Algebra with Inverses

Compared with using division in scalar algebra, one has to be much more careful in using inverses in matrix algebra. This is because one never divides by a matrix; instead, in dealing with equations, one multiplies by an inverse. For example, given $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{AX} = \mathbf{B}$, the equation can be pre-multiplied, on both sides, by $\mathbf{A}^{-1}$ (provided that it exists) to obtain $\mathbf{A}^{-1}\mathbf{AX} = \mathbf{A}^{-1}\mathbf{B}$ and thus $\mathbf{IX} = \mathbf{A}^{-1}\mathbf{B}$ or $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$. Note that $\mathbf{X}$ does not equal $\mathbf{BA}^{-1}$. Provided that conformability is satisfied, one could post-multiply $\mathbf{AX} = \mathbf{B}$ by $\mathbf{A}^{-1}$ and obtain $\mathbf{AXA}^{-1} = \mathbf{BA}^{-1}$; but that is it. No further simplification occurs.

Suppose that we have $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{K}$ such that $\mathbf{PK} = \mathbf{QK}$. This leads to $\mathbf{P} = \mathbf{Q}$ *only* if $\mathbf{K}^{-1}$ exists.

Inverses can also be used in factoring; for example, $\mathbf{R} + \mathbf{RST} = \mathbf{R}(\mathbf{I} + \mathbf{ST}) = \mathbf{R}(\mathbf{T}^{-1} + \mathbf{S})\mathbf{T}$, provided that $\mathbf{T}^{-1}$ exists.

Verifying the form of a particular inverse is often achieved by the following argument. Suppose that it is postulated that $\mathbf{A}$ inverse is $\mathbf{Q}$. Verifying this can be achieved by considering the product $\mathbf{AQ}$. If that can be shown to be equal to $\mathbf{I}$, thus $\mathbf{AQ} = \mathbf{I}$, then

$\mathbf{A}^{-1}\mathbf{AQ} = \mathbf{A}^{-1}\mathbf{I}$; that is, $\mathbf{Q} = \mathbf{A}^{-1}$. For example, suppose that $\mathbf{A}$ is $(\mathbf{I} + \mathbf{XY})$ and $\mathbf{Q}$ is $\mathbf{I} - \mathbf{X}(\mathbf{I} + \mathbf{YX})^{-1}\mathbf{Y}$. Then $\mathbf{A}^{-1}$ is shown to be $\mathbf{Q}$ by considering $\mathbf{AQ}$:

$$\begin{aligned}\mathbf{AQ} &= (\mathbf{I} + \mathbf{XY})[\mathbf{I} - \mathbf{X}(\mathbf{I} + \mathbf{YX})^{-1}\mathbf{Y}] \\ &= \mathbf{I} + \mathbf{XY} - (\mathbf{I} + \mathbf{XY})\mathbf{X}(\mathbf{I} + \mathbf{YX})^{-1}\mathbf{Y} \\ &= \mathbf{I} + \mathbf{XY} - (\mathbf{X} + \mathbf{XYX})(\mathbf{I} + \mathbf{YX})^{-1}\mathbf{Y} \\ &= \mathbf{I} + \mathbf{XY} - \mathbf{X}(\mathbf{I} + \mathbf{YX})(\mathbf{I} + \mathbf{YX})^{-1}\mathbf{Y} \\ &= \mathbf{I} + \mathbf{XY} - \mathbf{XY} \\ &= \mathbf{I},\end{aligned}$$

and so $\mathbf{A}^{-1} = \mathbf{Q}$.

Computers and Inverses

The arithmetic required for calculating an inverse matrix can be voluminous. Fortunately, computers have eased this situation enormously and many **software** packages include reliable routines for doing the arithmetic. Nevertheless, there are cases in which rounding error can lead to erroneous results; thankfully, this occurs very seldom, and software often handles it satisfactorily.

Rank

Linear Dependence and Independence of Vectors

$$\mathbf{Xa} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_c]\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_i \\ \vdots \\ a_c \end{bmatrix} = \sum_{i=1}^{c} a_i \mathbf{x}_i$$

is a vector. It is a *linear combination* of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_c$.

If, for a given $\mathbf{X}$ (with all columns nonnull), a nonnull vector $\mathbf{a}$ exists such that $\mathbf{Xa} = \mathbf{0}$, then the columns of $\mathbf{X}$ are said to be a set of *linearly dependent vectors*. If no such $\mathbf{a}$ exists, the columns are *linearly independent vectors*. These definitions exclude null vectors.

A Definition of Rank

If c columns are linearly dependent, there is always a smaller number of them that are linearly independent. In fact, there may be several sets of less than c columns that are linearly independent, with those sets not necessarily all having the same number of columns. The greatest number of columns in such a set is called the *rank of* $\mathbf{A}$, often denoted $\mathrm{r}(\mathbf{A})$. Thus, $\mathrm{r}(\mathbf{A})$ is the largest number of linearly independent columns available from $\mathbf{A}$. The "largest" is usually omitted. Thus, $\mathrm{r}(\mathbf{A})$ is the number of linearly independent columns in $\mathbf{A}$.

Some Properties and Consequences

Rank is an important and exceedingly useful concept in matrix algebra, with widespread applications. A list of some of the properties of rank follows:

1. The numbers of linearly independent rows and columns in a matrix are the same, $\mathrm{r}(\mathbf{A})$.
2. $\mathrm{r}(\mathbf{0}) = 0$.
3. $\mathrm{r}(\mathbf{A}_{p\times q}) \leq p$ and $\mathrm{r}(\mathbf{A}_{p\times q}) \leq q$.
4. $\mathrm{r}(\mathbf{A}_{n\times n}) \leq n$.
5. $\mathrm{r}(\mathbf{A}_{n\times n}) < n \Leftrightarrow \mathbf{A}$ is singular, $|\mathbf{A}| = \mathbf{0}$, with $\mathbf{A}^{-1}$ not existing.
6. $\mathrm{r}(\mathbf{A}_{n\times n}) = n \Leftrightarrow \mathbf{A}$ is nonsingular, $|\mathbf{A}| \neq \mathbf{0}$, with $\mathbf{A}^{-1}$ existing: $\mathbf{A}$ is said to be of *full rank*.
7. $\mathrm{r}(\mathbf{A}_{p\times q}) = p < q$ means that $\mathbf{A}$ has *full row rank*.
8. $\mathrm{r}(\mathbf{A}_{p\times q}) = q < p$ means that $\mathbf{A}$ has *full column rank*.
9. $\mathbf{A}_{p\times q}$ having rank r can always be expressed as $\mathbf{A}_{p\times q} = \mathbf{K}_{p\times r}\mathbf{L}_{r\times q}$, where $\mathbf{K}$ has full column rank r and $\mathbf{L}$ has full row rank r.
10. $\mathrm{r}(\mathbf{AB}) \leq$ lesser of $\mathrm{r}(\mathbf{A})$ and $\mathrm{r}(\mathbf{B})$.
11. $\mathrm{r}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$ for idempotent $\mathbf{A}$.
12. $\mathrm{r}(\mathbf{A}) = \mathrm{r}(\mathbf{A}')$.
13. $\mathrm{r}(\mathbf{A}) = \mathrm{r}(\mathbf{AA}')$.
14. $\mathrm{r}(\mathbf{A}) = \mathrm{r}(\mathbf{TA})$ for nonsingular $\mathbf{T}$.
15. $\mathrm{r}(\mathbf{A}^{-1}) = \mathrm{r}(\mathbf{A}) = n$ for $\mathbf{A}_{n\times n}$.

Left and Right Inverses

For given $\mathbf{A}_{r\times c}$, there exists:

1. $\mathbf{A}^{-1}$, the inverse of $\mathbf{A}$, such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{AA}^{-1}$ if and only if $\mathbf{A}$ is square, with $|\mathbf{A}| \neq 0$;
2. $\mathbf{L}_{c\times r}$, a *left inverse* of $\mathbf{A}$, such that $\mathbf{LA} = \mathbf{I}_c$ (and $\mathbf{AL} \neq \mathbf{I}_r$) only if $\mathbf{A}$ has full column rank;
3. $\mathbf{R}_{c\times r}$, a *right inverse* of $\mathbf{A}$, such that $\mathbf{AR} = \mathbf{I}_r$ (and $\mathbf{RA} \neq \mathbf{I}_c$) only if $\mathbf{A}$ has full row rank;
4. neither an $\mathbf{A}^{-1}$, $\mathbf{L}$, nor $\mathbf{R}$ of (1), (2), or (3) – for example, any matrix having at least one null row and one null column.

Only when $\mathbf{A}^{-1}$ exists does $\mathbf{A}$ have both an $\mathbf{L}$ and an $\mathbf{R}$; and they both equal $\mathbf{A}^{-1}$. Otherwise, if $\mathbf{A}$ has full column (row) rank it has left (right) inverses of many values.

Vector Spaces

Since a vector of order n has n elements, it can be considered as a point in n-space, which is denoted R^n. Consider a set of vectors S, in R^n. Suppose, for every pair of vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in S, that both the sum $\mathbf{x}_i + \mathbf{x}_j$ and the vectors $a\mathbf{x}_i$ and $b\mathbf{x}_i$ for any scalars a and b are in S; then S is a *vector space*.

Suppose that every vector in the vector space S can be expressed as a linear combination of the set of t vectors, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$. Then that set *spans*, or *generates*, S and is called a *spanning* set of S. If those t vectors are also linearly independent, they are said to be a *basis* for S, and the number of such vectors is the *dimension* of S, $\dim(S)$.

There are many vector spaces of order n; and each of them usually has several bases.

Range and Null Spaces

$\mathbf{A}$ of rank r has r linearly independent columns. All vectors that are linear combinations of those columns form a vector space. It is known as the *column space* of $\mathbf{A}$, the *range* of $\mathbf{A}$, or the *manifold* of $\mathbf{A}$, often denoted by $\mathcal{R}(\mathbf{A})$. Clearly, $r = \mathrm{r}(\mathbf{A}) = \dim[\mathcal{R}(\mathbf{A})]$.

The space defined by the many vectors $\mathbf{x}$ for which $\mathbf{Ax} = \mathbf{0}$ (with $\mathbf{A}$ being rectangular, or square and singular) is the *null space* of $\mathbf{A}$, denoted $\mathcal{N}(\mathbf{A})$. Its dimension is the *nullity* of $\mathbf{A}$: $\mathrm{nullity}(\mathbf{A}) = \dim[\mathcal{N}(\mathbf{A})]$.

Equivalent and Congruent Canonical Forms

Elementary Operators

Three particular adaptations of identity matrices are *elementary operators; each is an identity matrix* with (i) two rows (or columns) interchanged, or (ii) λ in place of a one in the diagonal, or (iii) λ in place of a zero in an off-diagonal element. These and all products of any numbers of them are nonsingular.

Equivalent Canonical Form

For any $\mathbf{A}_{p\times q}$, of rank r, there always exists a $\mathbf{P}$ and a $\mathbf{Q}$, each a product of elementary operators, such that

$$\mathbf{P}_{p\times p}\mathbf{A}_{p\times q}\mathbf{Q}_{q\times q} = \begin{bmatrix} \mathbf{I}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{K}, \text{ say.}$$

$\mathbf{K}$ is the *equivalent canonical form* of $\mathbf{A}$; or the *canonical form under equivalence* of $\mathbf{A}$. Because $\mathbf{P}$ and $\mathbf{Q}$ are products of elementary operators, they are nonsingular, and so the equation $\mathbf{PAQ} = \mathbf{K}$ leads to $\mathbf{A} = \mathbf{P}^{-1}\mathbf{K}\mathbf{Q}^{-1}$. If $\mathbf{A}$ is nonsingular, $\mathbf{K} = \mathbf{I}$ and $\mathbf{A}^{-1} = \mathbf{QP}$.

Congruent Canonical Form

When $\mathbf{A}$ is symmetric (and hence square), the $\mathbf{Q}$ of $\mathbf{PAQ}$ can be $\mathbf{P}'$, giving

$$\mathbf{PAP}' = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \mathbf{C},$$

known as the *congruent canonical form* of $\mathbf{A}$ or the *canonical form under congruence*.

En route to deriving $\mathbf{C}$, one can obtain the form

$$\mathbf{P}_*\mathbf{AP}_*' = \begin{pmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \mathbf{C}_*,$$

where $\mathbf{D}_r$ is a diagonal matrix of order and rank r. For $\mathbf{A}$ being real, $\mathbf{P}_*$ will be real; but if $\mathbf{D}_r$ has negative elements, $\mathbf{P}$ in obtaining $\mathbf{C}$ will be complex. For $\mathbf{A}$ being nonnegative definite, elements of $\mathbf{D}_r$ are always positive and $\mathbf{P}$ is always real.

Utility: Sums of Squares

The utility of these canonical forms is their existence. For each $\mathbf{A}$ there are many values of $\mathbf{P}$ (and $\mathbf{Q}$) but usually not any one of them is of particular interest. It is the fact that they exist that is important, and that provides the means for establishing other useful results. For example, consider the quadratic form $q = \mathbf{x}'\mathbf{Ax}$ with $\mathbf{A} = \mathbf{A}'$ of rank r. Then there is a $\mathbf{P}$ such that $\mathbf{PAP}' = \mathbf{C}$. Thus, $q = \mathbf{x}'\mathbf{P}^{-1}\mathbf{PAP}'(\mathbf{P}')^{-1}\mathbf{x}$, and letting $\mathbf{y} = (\mathbf{P}')^{-1}\mathbf{x}$ gives $q = \mathbf{y}'\mathbf{Cy}$ which, because

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

becomes $q = \sum_{i=1}^{r} y_i^2$. Thus, without knowing $\mathbf{P}$ except for its existence and nonsingularity, we can show that a quadratic form can always be expressed as a sum of r squared terms, where r is the rank of the (symmetric) matrix $\mathbf{A}$ of the quadratic form. That

is a result of great importance in considering the distribution of quadratic forms of normally distributed random variables.

Generalized Inverses

Definition

For any nonnull matrix $\mathbf{A}$, there is a unique matrix $\mathbf{M}$ satisfying:

(i) $\mathbf{AMA} = \mathbf{A}$;
(ii) $\mathbf{MAM} = \mathbf{M}$;
(iii) $(\mathbf{AM})' = \mathbf{AM}$; and
(iv) $(\mathbf{MA})' = \mathbf{MA}$.

These are the *Penrose* conditions and $\mathbf{M}$ is the *Moore–Penrose inverse*. Whereas $\mathbf{M}$ is unique, there are (with one exception) many matrices $\mathbf{G}$ satisfying

$$\mathbf{AGA} = \mathbf{A},$$

which is condition (i). Each matrix $\mathbf{G}$ satisfying $\mathbf{AGA} = \mathbf{A}$ is called a *generalized inverse* of $\mathbf{A}$, and if it also satisfies $\mathbf{GAG} = \mathbf{G}$ it is a *reflexive generalized inverse*. The exception is when $\mathbf{A}$ is nonsingular: there is then only one $\mathbf{G}$; namely, $\mathbf{G} = \mathbf{A}^{-1}$.

Arbitrariness

That there are many matrices $\mathbf{G}$ can be illustrated by showing ways in which from one $\mathbf{G}$ others can be obtained. Thus, if $\mathbf{A}$ is partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11}$ is nonsingular with the same rank as $\mathbf{A}$, then

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_{11}^{-1} - \mathbf{U}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{V} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{W}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{bmatrix}$$

is a generalized inverse of $\mathbf{A}$ for any values of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$. This can be used to show that a generalized inverse of a symmetric matrix is not necessarily symmetric; and that of a singular matrix is not necessarily singular [[4], p. 219].

A simpler illustration of arbitrariness is that if $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$ then so is

$$\mathbf{G}^* = \mathbf{GAG} + (\mathbf{I} - \mathbf{GA})\mathbf{S} + \mathbf{T}(\mathbf{I} - \mathbf{AG}),$$

for any values of $\mathbf{S}$ and $\mathbf{T}$.

Generalized Inverses of $\boldsymbol{X'X}$

The matrix $\mathbf{X}'\mathbf{X}$ plays an important role in statistics, usually involving a generalized inverse thereof, which has several useful properties. Thus, for $\mathbf{G}$ satisfying

$$\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X},$$

$\mathbf{G}'$ is also a generalized inverse of $\mathbf{X}'\mathbf{X}$ (and $\mathbf{G}$ is not necessarily symmetric). Also,

1. $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}$;
2. $\mathbf{X}\mathbf{G}\mathbf{X}'$ is invariant to $\mathbf{G}$;
3. $\mathbf{X}\mathbf{G}\mathbf{X}'$ is symmetric, whether or not $\mathbf{G}$ is;
4. $\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}\mathbf{X}^+$ for $\mathbf{X}^+$ being the Moore–Penrose inverse of $\mathbf{X}$.

Solving Linear Equations

A Single Solution

Given $\mathbf{A}$ and $\mathbf{y}$, the equations $\mathbf{Ax} = \mathbf{y}$ are linear in the unknowns, the elements of $\mathbf{x}$. When $\mathbf{A}$ is nonsingular, the equations are solved uniquely, as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. But for singular or rectangular $\mathbf{A}$, solutions involve using a generalized inverse of $\mathbf{A}$. The following results apply.

First, equations $\mathbf{Ax} = \mathbf{y}$ are said to be *consistent* when any linear relationships existing among rows of $\mathbf{A}$ also exist among elements of $\mathbf{y}$. Only then do solutions exist. Secondly, for singular or rectangular $\mathbf{A}$ there will be many solutions for $\mathbf{x}$, except when $\mathbf{A}$ has full column rank, whereupon there is only one solution, $\mathbf{x} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}$: and this includes, of course, the case of nonsingular $\mathbf{A}$.

Many Solutions

When $\mathbf{A}$ has less than full column rank, there are many solutions. They are characterized as follows, with $\mathbf{G}$ being a generalized inverse satisfying $\mathbf{AGA} = \mathbf{A}$.

1. $\tilde{\mathbf{x}} = \mathbf{G}\mathbf{y}$ is a solution if and only if $\mathbf{AGA} = \mathbf{A}$.
2. Letting $\mathbf{G}$ take all its possible values in $\tilde{\mathbf{x}} = \mathbf{G}\mathbf{y}$ (for $\mathbf{y} \neq \mathbf{0}$) generates all possible solutions.
3. $\tilde{\mathbf{x}} = \mathbf{G}\mathbf{y} + (\mathbf{I} - \mathbf{GA})\mathbf{z}$ is a solution for any arbitrary $\mathbf{z}$ of the same order as $\mathbf{x}$.
4. For a given $\mathbf{G}$, letting $\mathbf{z}$ take all possible values in $\tilde{\mathbf{x}} = \mathbf{G}\mathbf{y} + (\mathbf{I} - \mathbf{GA})\mathbf{z}$ generates all possible solutions.

5. When $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_t$ are any solutions, $\sum_{i=1}^{t} \lambda_i \tilde{\mathbf{x}}_i$ is a solution (with $\mathbf{y} \neq \mathbf{0}$) if and only if $\sum_{i=1}^{t} \lambda_i = 1$; this condition is not needed when $\mathbf{y} = \mathbf{0}$.
6. For $\mathbf{A}_{p\times q}$ and $\mathbf{y} \neq \mathbf{0}$ there are $q - r(\mathbf{A}) + 1 - \delta_{\mathbf{y},\mathbf{0}}$ linearly independent solutions, where $\delta_{\mathbf{y},\mathbf{0}} = 1$ when $\mathbf{y} = \mathbf{0}$ and zero otherwise.
7. The value of $\mathbf{k}'\tilde{\mathbf{x}}$ is invariant to $\tilde{\mathbf{x}}$ if and only if $\mathbf{k}' = \mathbf{k}'\mathbf{G}\mathbf{A}$.
8. When $\mathbf{y} = \mathbf{0}$, solutions are orthogonal to rows of $\mathbf{A}$; and solutions orthogonal to each other can always be derived. The vector space spanned by the solutions, sometimes called the *solution space*, is the *orthogonal complement* of the row space of $\mathbf{A}$.

Partitioned Matrices

Some results for partitioned matrices used in statistics are as follows.

Orthogonality

If $\mathbf{P} = [\mathbf{A} \quad \mathbf{B}]$ is orthogonal,

$$\mathbf{P}\mathbf{P}' = \mathbf{I} \Rightarrow [\mathbf{A} \quad \mathbf{B}]\begin{bmatrix}\mathbf{A}'\\ \mathbf{B}'\end{bmatrix} = \mathbf{I} \Rightarrow \mathbf{A}\mathbf{A}' + \mathbf{B}\mathbf{B}' = \mathbf{I},$$

$$\mathbf{P}'\mathbf{P} = \mathbf{I} \Rightarrow \begin{bmatrix}\mathbf{A}'\\ \mathbf{B}'\end{bmatrix}[\mathbf{A} \quad \mathbf{B}] = \mathbf{I} \Rightarrow \begin{bmatrix}\mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B}\\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B}\end{bmatrix}$$

$$= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \Rightarrow \mathbf{A}'\mathbf{A} = \mathbf{I}, \qquad \mathbf{A}'\mathbf{B} = \mathbf{0}$$

$$\text{and} \quad \mathbf{B}'\mathbf{B} = \mathbf{I}.$$

Note that $\mathbf{A}\mathbf{A}'$ and $\mathbf{B}\mathbf{B}'$ are not identity matrices.

Determinants

$$\begin{vmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{vmatrix} = |\mathbf{A}||\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{D}||\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}|,$$

provided that $\mathbf{A}^{-1}$ and $\mathbf{D}^{-1}$ exist, where needed.

Inverses

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{A}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \begin{bmatrix}-\mathbf{A}^{-1}\mathbf{B}\\ \mathbf{I}\end{bmatrix} \times (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}[-\mathbf{C}\mathbf{A}^{-1} \quad \mathbf{I}]$$

$$= \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}^{-1}\end{bmatrix} + \begin{bmatrix}\mathbf{I}\\ -\mathbf{D}^{-1}\mathbf{C}\end{bmatrix} \times (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}[\mathbf{I} \quad -\mathbf{B}\mathbf{D}^{-1}],$$

again provided that $\mathbf{A}^{-1}$ and $\mathbf{D}^{-1}$ exist as needed.

Schur Complements

In

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}$$

the *Schur complement* of $\mathbf{A}$ is $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ and that of $\mathbf{D}$ is $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$. The inverse of one involves that of the other:

$$(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} = \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C} \times (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}.$$

This result also applies when the two minus signs are changed to plus, and the plus to minus. It also has some useful special cases; for example,

$$(\mathbf{D} \pm \lambda\mathbf{t}\mathbf{t}')^{-1} = \mathbf{D}^{-1} \mp \frac{\mathbf{D}^{-1}\mathbf{t}\mathbf{t}'\mathbf{D}^{-1}}{(\lambda^{-1} \pm \mathbf{t}'\mathbf{D}^{-1}\mathbf{t})}.$$

Generalized Inverses

By analogy with expressions for the inverse, one might expect (with $\mathbf{A}^-$ being a generalized inverse of $\mathbf{A}$)

$$\tilde{\mathbf{Q}} = \begin{bmatrix}\mathbf{A}^- & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \begin{bmatrix}-\mathbf{A}^-\\ \mathbf{I}\end{bmatrix} \times (\mathbf{D} - \mathbf{C}\mathbf{A}^-\mathbf{B})^-[-\mathbf{C}\mathbf{A}^- \quad \mathbf{I}]$$

to be a generalized inverse of

$$\mathbf{Q} = \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}.$$

It is, if and only if $\mathrm{r}(\mathbf{Q}) = \mathrm{r}(\mathbf{A}) + \mathrm{r}(\mathbf{D} - \mathbf{C}\mathbf{A}^-\mathbf{B})$. Satisfying this rank condition depends upon $\mathbf{A}^-$. For some values of $\mathbf{A}^-$ the condition will be satisfied and for others it will not. Only when it is satisfied will $\tilde{\mathbf{Q}}$ be a generalized inverse of $\mathbf{Q}$.

Direct Sums

The direct sum of matrices $\mathbf{A}$ and $\mathbf{B}$, each of any order, is defined as

$$\mathbf{A} \oplus \mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}.$$

Extension to the direct sum of more than two matrices is straightforward.

Provided that the needed conformability requirements are met,

$$(\mathbf{A} \oplus \mathbf{B}) + (\mathbf{C} \oplus \mathbf{D}) = (\mathbf{A} + \mathbf{C}) \oplus (\mathbf{B} + \mathbf{D}),$$

$$(\mathbf{P} \oplus \mathbf{Q})(\mathbf{L} \oplus \mathbf{M}) = \mathbf{PL} \oplus \mathbf{QM},$$

and

$$(\mathbf{X} \oplus \mathbf{Y})^{-1} = \mathbf{X}^{-1} \oplus \mathbf{Y}^{-1}.$$

Direct Products

The *direct product* of two matrices, each of any order, is defined as

$$\mathbf{A}_{p\times q} \otimes \mathbf{B}_{m\times n} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1q}\mathbf{B} \\ \vdots & & \\ a_{p1}\mathbf{B} & \dots & a_{pq}\mathbf{B} \end{bmatrix}_{pm\times qn} = \{_{\mathrm{m}}\ a_{ij}\mathbf{B}\}_{i=1,j=1}^{p\quad q}.$$

It is sometimes called the *Kronecker product*. Some properties follow – assuming that conformability requirements are met:

1. $\mathbf{x}' \otimes \mathbf{y} = \mathbf{y}\mathbf{x}' = \mathbf{y} \otimes \mathbf{x}'$;
2. $\lambda \otimes \mathbf{A} = \lambda\mathbf{A} = \mathbf{A} \otimes \lambda$;
3. $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$, *not* $\mathbf{B}' \otimes \mathbf{A}'$;
4. $(\mathbf{A} \otimes \mathbf{B})(\mathbf{X} \otimes \mathbf{Y}) = \mathbf{AX} \otimes \mathbf{BY}$;
5. $(\mathbf{P} \otimes \mathbf{Q})^{-1} = \mathbf{P}^{-1} \otimes \mathbf{Q}^{-1}$, *not* $\mathbf{Q}^{-1} \otimes \mathbf{P}^{-1}$;
6. $[\mathbf{A}_1 \quad \mathbf{A}_2] \otimes \mathbf{B} = [\mathbf{A}_1 \otimes \mathbf{B} \quad \mathbf{A}_2 \otimes \mathbf{B}]$, $\mathbf{A} \otimes [\mathbf{B}_1 \quad \mathbf{B}_2] \neq [\mathbf{A} \otimes \mathbf{B}_1 \mathbf{A} \otimes \mathbf{B}_2]$;
7. $\mathrm{r}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{r}(\mathbf{A})\mathrm{r}(\mathbf{B})$;
8. $\mathrm{tr}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$;
9. $|\mathbf{A}_{p\times p} \otimes \mathbf{B}_{m\times m}| = |\mathbf{A}|^m |\mathbf{B}|^p$.

Sometimes $\mathbf{A} \otimes \mathbf{B} = \{a_{ij}\mathbf{B}\}$ as defined above is called the *right direct product*, to distinguish it from $\mathbf{B} \otimes \mathbf{A}$, which is then called the *left direct product*; and on rare occasions $\{a_{ij}\mathbf{B}\}$ will be found defined as $\mathbf{B} \otimes \mathbf{A}$.

Eigenvalues and Eigenvectors

The equation

$$\mathbf{Au} = \lambda\mathbf{u}, \text{ i.e. } (\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0},$$

has solutions for $\mathbf{u}$ provided that $\mathbf{A} - \lambda\mathbf{I}$ is singular. This occurs when

$$|\mathbf{A} - \lambda\mathbf{I}| = \mathbf{0}.$$

This is called the *characteristic equation* of $\mathbf{A}$; for $\mathbf{A}_{n\times n}$ it is a polynomial of order n and therefore has n solutions for λ. Those solutions are the **eigenvalues** (or *eigenroots*) of $\mathbf{A}$. They can be real or complex, positive or negative, or zero. For each eigenvalue, λ_* say, a corresponding value of $\mathbf{u}$ can be obtained from solving the equations $(\mathbf{A} - \lambda_*\mathbf{I})\mathbf{u} = \mathbf{0}$, as

$$\mathbf{u}_* = [\mathbf{I} - (\mathbf{A} - \lambda_*\mathbf{I})^-(\mathbf{A} - \lambda_*\mathbf{I})]\mathbf{z}$$

for arbitrary $\mathbf{z}$. (Searle [4, Section 11.4] has details.) $\mathbf{u}_*$ is the **eigenvector** corresponding to λ_*.

Numerical Example

For

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 1 & 1 \\ -7 & 2 & -3 \end{bmatrix},$$

the characteristic equation $|\mathbf{A} - \lambda\mathbf{I}| = 0$ reduces to $(\lambda - 1)(\lambda - 3)(\lambda + 4) = 0$, so that the eigenvalues are 1, 3, and -4. For $\lambda_* = 1$ the eigenvector, from the equation for $\mathbf{u}_*$,

$$\begin{aligned} \mathbf{u}_* &= \left[\mathbf{I} - \begin{pmatrix} 1 & 2 & 0 \\ 2 & 0 & 1 \\ -7 & 2 & -4 \end{pmatrix}^- \begin{pmatrix} 1 & 2 & 0 \\ 2 & 0 & 1 \\ -7 & 2 & -4 \end{pmatrix}\right] \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\ &= \left[\mathbf{I} + \tfrac{1}{4}\begin{pmatrix} 0 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 2 & 0 \\ 2 & 0 & 1 \\ -7 & 2 & -4 \end{pmatrix}\right] \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\ &= \left[\mathbf{I} + \tfrac{1}{4}\begin{pmatrix} -4 & 0 & -2 \\ 0 & -4 & 1 \\ 0 & 0 & 0 \end{pmatrix}\right] \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & -\frac{1}{2} \\ 0 & 0 & \frac{1}{4} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}z_3 \\ \frac{1}{4}z_3 \\ z_3 \end{bmatrix} \quad \text{for any } z_3 \neq 0. \end{aligned}$$

Similarly, for $\lambda_* = 3$,

$$\mathbf{u}_* = \left[\mathbf{I} - \begin{pmatrix} -1 & 2 & 0 \\ 2 & -2 & 1 \\ -7 & 2 & -3 \end{pmatrix}^- \begin{pmatrix} -1 & 2 & 0 \\ 2 & -2 & 1 \\ -7 & 2 & -3 \end{pmatrix}\right] \mathbf{z}$$

$$= \left[\mathbf{I} - \frac{1}{2}\begin{pmatrix} -2 & -2 & 0 \\ -2 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -1 & 2 & 0 \\ 2 & -2 & 1 \\ -7 & 2 & -3 \end{pmatrix}\right]\mathbf{z}$$

$$= \left[\mathbf{I} - \frac{1}{2}\begin{pmatrix} -2 & 0 & -2 \\ 0 & -2 & -1 \\ 0 & 0 & 0 \end{pmatrix}\right]\mathbf{z} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}\mathbf{z}$$

$$= \begin{bmatrix} -z_3 \\ -\frac{1}{2}z_3 \\ z_3 \end{bmatrix}.$$

The case of $\lambda_* = -4$ is left to the reader.

Properties of Eigenvalues

See **Eigenvalue**.

Properties of Eigenvectors

See **Eigenvector**.

Some Summaries

Orthogonal Matrices

Any two of (i) $\mathbf{A}$ being square, (ii) $\mathbf{A}\mathbf{A}' = \mathbf{I}$, and (iii) $\mathbf{A}'\mathbf{A} = \mathbf{I}$ imply the third; and define $\mathbf{A}$ as being orthogonal. The properties of orthogonal $\mathbf{A}$ include the following

1. Rows (columns) are orthonormal;
2. $|\mathbf{A}| = \pm 1$;
3. λ being an eigenroot of $\mathbf{A}$ implies that $1/\lambda$ is also;
4. $\mathbf{AB}$ is orthogonal when $\mathbf{A}$ and $\mathbf{B}$ are.

Idempotent Matrices

Idempotent $\mathbf{A}$ of order n has the following properties:

1. $\mathbf{A}^2 = \mathbf{A}$;
2. $\mathbf{A}$ is singular, unless $\mathbf{A} = \mathbf{I}$;
3. $\mathrm{r}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$;
4. $\mathbf{I} - \mathbf{A}$ is idempotent, with $\mathrm{r}(\mathbf{I} - \mathbf{A}) = n - \mathrm{r}(\mathbf{A})$;
5. If $\mathbf{A}$ is also symmetric (but not $\mathbf{I}$) it is positive semidefinite, and can be expressed as $\mathbf{A} = \mathbf{L}\mathbf{L}'$ for $\mathbf{L}'\mathbf{L} = \mathbf{I}$;
6. For idempotent $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{AB}$ is idempotent if $\mathbf{AB} = \mathbf{BA}$;
7. $\mathrm{r}(\mathbf{A})$ eigenvalues of $\mathbf{A}$ are 1.0, and $n - \mathrm{r}(\mathbf{A})$ are 0;
8. There is a $\mathbf{U}$ such that
$$\mathbf{U}^{-1}\mathbf{A}\mathbf{U} = \begin{bmatrix} \mathbf{I}_{\mathrm{r}(\mathbf{A})} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix};$$
9. $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is idempotent, and is very useful in statistics.

Matrices $a\mathbf{I} + b\mathbf{J}$

The matrix $a\mathbf{I} + b\mathbf{J}$ for $\mathbf{J} = \mathbf{1}\mathbf{1}'$ occurs in a number of analysis of variance situations in statistics. When of order n it has the following properties:

$$(a_1\mathbf{I} + b_1\mathbf{J})(a_2\mathbf{I} + b_2\mathbf{J}) = a_1a_2\mathbf{I} + (a_1b_2 + a_2b_1 + nb_1b_2)\mathbf{J},$$

$$(a\mathbf{I} + b\mathbf{J})^{-1} = \frac{1}{a}\left(\mathbf{I} - \frac{b}{a+nb}\mathbf{J}\right),$$

$$|a\mathbf{I} + b\mathbf{J}| = a^{n-1}(a + nb).$$

Eigenvalues are a, $n-1$ times, and $a + nb$ once.

Nonnegative Definite Matrices

If $\mathbf{A}$ is nonnegative definite (n.n.d.):

1. $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$, for all $\mathbf{x} \neq 0$;
2. $\mathbf{A}$ is assumed to be symmetric, because otherwise it can be replaced by $\frac{1}{2}(\mathbf{A} + \mathbf{A}')$;
3. $|\mathbf{A}| \geq 0$;
4. Diagonal elements of $\mathbf{A}$ are ≥ 0;
5. Principal leading minors are ≥ 0;
6. Eigenvalues are ≥ 0;
7. For $\mathbf{A}$ is positive definite (p.d.), all the above ≥ 0 symbols become > 0;
8. For real $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$ is n.n.d.

For real $\mathbf{X}$ of full column rank:

1. $\mathbf{X}'\mathbf{X}$ is p.d.;
2. $(\mathbf{X}'\mathbf{X})^{-1}$ exists;
3. $\mathbf{X}\mathbf{X}'$ has Moore–Penrose inverse $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'$.

Canonical and Other Forms

For any matrix $\mathbf{A}_{p\times q}$ of rank r:

1. Equivalent canonical form:
$$\mathbf{P}\mathbf{A}\mathbf{Q} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{P} \text{ and } \mathbf{Q} \text{ nonsingular.}$$

2. Similar canonical form:

$$\mathbf{AU} = \mathbf{UD}\{\lambda\},$$

where $\mathbf{D}\{\lambda\}$ is the diagonal matrix of eigenvalues; and $\mathbf{U}$ is the matrix of corresponding eigenvectors. $\mathbf{U}^{-1}$ exists when the diagonability theorem is satisfied (*see* **Eigenvector**), and then

$$\mathbf{U}^{-1}\mathbf{AU} = \mathbf{D}\{\lambda\}.$$

3. Singular-valued decomposition:

$$\mathbf{A} = \mathbf{L}\begin{bmatrix} \boldsymbol{\Delta}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{M}',$$

where $\mathbf{L}$ and $\mathbf{M}$ are each orthogonal, and $\boldsymbol{\Delta}_r = (\boldsymbol{\Delta}^2)^{1/2}$, where

$$\mathbf{L}'\mathbf{AA}'\mathbf{L} = \begin{bmatrix} \boldsymbol{\Delta}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \text{and}$$

$$\mathbf{M}'\mathbf{A}'\mathbf{AM} = \begin{bmatrix} \boldsymbol{\Delta}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

with $\boldsymbol{\Delta}^2$ being the diagonal matrix of the (positive) eigenroots of $\mathbf{A}'\mathbf{A}$ (or, equivalently, of $\mathbf{AA}'$).

For symmetric $\mathbf{A}$ of order p and rank r:

4. Diagonal form:

$$\mathbf{PAP}' = \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \text{with } \mathbf{D}_r \text{ diagonal, order } r.$$

When $\mathbf{A}$ is n.n.d., elements of $\mathbf{D}_r$ are positive.

5. Congruent canonical form:

$$\mathbf{RAR}' = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \text{for } \mathbf{R} \text{ possibly complex.}$$

When $\mathbf{A}$ is n.n.d., $\mathbf{R}$ is real.

6. Orthogonal similar canonical form:

$$\mathbf{U}'\mathbf{AU} = \mathbf{D}\{\lambda\},$$

with $\mathbf{U}$ being orthogonal and $\mathbf{U}^{-1} = \mathbf{U}'$.

7. Spectral decomposition:

$$\mathbf{A} = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i',$$

for λ_i being an eigenvalue and $\mathbf{u}_i$ its corresponding eigenvector.

Solving Equations by Iteration

Current computing facilities provide numerous methods of arithmetically solving equations which cannot be solved algebraically. Matrix notation permits succinct description of one of these methods.

For n equations in n unknowns, represented by $\mathbf{x}$, let the equations be

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}, \tag{1}$$

and define

$$\mathbf{G}(\mathbf{x}) = \{g_{ij}(\mathbf{x})\} = \left\{\frac{\partial}{\partial x_j} f_i(\mathbf{x})\right\} \quad \text{for}$$

$$i, j = 1, \ldots, n. \tag{2}$$

Suppose that $\mathbf{x}_r$ is an approximate solution for $\mathbf{x}$ to $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Then an improved approximation is $\mathbf{x}_{r+1}$ for

$$\mathbf{f}(\mathbf{x}_{r+1}) = \mathbf{f}(\mathbf{x}_r) + \mathbf{G}(\mathbf{x}_r)\boldsymbol{\Delta}_r, \tag{3}$$

where

$$\boldsymbol{\Delta}_r = \mathbf{x}_{r+1} - \mathbf{x}_r. \tag{4}$$

Were $\mathbf{x}_{r+1}$ to be a solution to (1) then $\mathbf{f}(\mathbf{x}_{r+1})$ would be $\mathbf{0}$ and (3) would yield

$$\boldsymbol{\Delta}_r = -[\mathbf{G}(\mathbf{x}_r)]^{-1}\mathbf{f}(\mathbf{x}_r), \tag{5}$$

and with this, (4) gives

$$\mathbf{x}_{r+1} = \boldsymbol{\Delta}_r + \mathbf{x}_r. \tag{6}$$

In this way, (5) and (6) provide an iterative procedure for calculating a solution: for some initial value $\mathbf{x}_0$, use (5) to obtain $\boldsymbol{\Delta}_0$ and then (6) to obtain $\mathbf{x}_1$; and back to (5) to obtain $\boldsymbol{\Delta}_1$, and so on.

Differential Calculus with Matrices

A number of situations in statistics involve maximizing or minimizing a function: for example, maximum likelihood estimation, least squares estimation, minimum variance procedures, minimizing loss functions, and so on. In many cases, differentiation of matrix expressions is involved, for which the following results are often useful.

Differentiating with Respect to a Scalar

Suppose that the elements of $\mathbf{A} = \{a_{ij}\}$ are functions of a scalar x. Then

$$\frac{\partial \mathbf{A}}{\partial x} = \left\{ \frac{\partial a_{ij}}{\partial x} \right\},$$

$$\frac{\partial \mathbf{A}^{-1}}{\partial x} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1},$$

$$\mathbf{A} \frac{\partial \mathbf{A}^{-}}{\partial x} \mathbf{A} = -\mathbf{A}\mathbf{A}^{-} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-}\mathbf{A},$$

where $\mathbf{A}^{-}$ is a generalized inverse of $\mathbf{A}$, satisfying $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$. Also,

$$\mathbf{A} \frac{\partial (\mathbf{A}'\mathbf{A})^{-}}{\partial x} \mathbf{A}' = -\mathbf{A}(\mathbf{A}'\mathbf{A})^{-} \frac{\partial (\mathbf{A}'\mathbf{A})}{\partial x} (\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'.$$

Also, for $\mathbf{A} = \mathbf{A}'$ and elements of $\mathbf{T}$ not involving x,

$$\mathbf{P} = \mathbf{T}(\mathbf{T}'\mathbf{A}\mathbf{T})^{-1}\mathbf{T}' \text{ has } \frac{\partial \mathbf{P}}{\partial x} = -\mathbf{P} \frac{\partial \mathbf{A}}{\partial x} \mathbf{P}.$$

Differentiating with Respect to Elements of a Vector

The basis of differentiating with respect to elements of $\mathbf{x}$ is defining what is meant by $\partial/\partial\mathbf{x}$. This is important, because the definition determines the form of its various applications, and because not all writers use the same definition. Any presentation of this topic should therefore start by defining $\partial/\partial\mathbf{x}$.

A widely used convention is that, for $\mathbf{x}$ being a column vector, $\partial/\partial\mathbf{x}$ is also: thus, for $\mathbf{x} = [x_1 \ldots x_n]'$, we define

$$\mathbf{x} = \{_{\mathrm{c}}\ x_i\}_{i=1}^{n} \quad \text{and} \quad \frac{\partial}{\partial \mathbf{x}} = \left\{ {}_{\mathrm{c}} \frac{\partial}{\partial x_i} \right\}_{i=1}^{n}.$$

Thus $\partial/\partial\mathbf{x}$ is a vector of differential operators. With this definition come the following basic results:

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}'\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{a}) = \mathbf{a} \quad \text{and} \quad \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}) = \mathbf{A}.$$

Then, in order to maintain feasible matrix dimensions, the convention is adopted that

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{A}\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}'),$$

and so

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{A}\mathbf{x}) = \mathbf{A}'.$$

This leads to

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{x} \quad \text{for } \mathbf{A} \text{ not symmetric,}$$

$$= 2\mathbf{A}\mathbf{x} \quad \text{for } \mathbf{A} \text{ symmetric.}$$

Differentiating with Respect to Elements of a Matrix

Again, the basic definition is important: for scalar θ and $\mathbf{X}_{p\times q}$,

$$\frac{\partial \theta}{\partial \mathbf{X}} = \left\{ {}_{\mathrm{m}} \frac{\partial \theta}{\partial x_{ij}} \right\}_{i=1\ j=1}^{p \quad q}.$$

For $\mathbf{X}$ having functionally unrelated elements,

$$\frac{\partial}{\partial \mathbf{X}}[\mathrm{tr}(\mathbf{X}\mathbf{A})] = \mathbf{A}'.$$

But for symmetric $\mathbf{X}$,

$$\frac{\partial}{\partial \mathbf{X}}[\mathrm{tr}(\mathbf{X}\mathbf{A})] = \mathbf{A} + \mathbf{A}' - \mathrm{diag}(\mathbf{A}),$$

where $\mathrm{diag}(\mathbf{A})$ is the diagonal matrix of the diagonal elements of $\mathbf{A}$. Of course, these results also apply to $\mathrm{tr}(\mathbf{A}\mathbf{X}) = \mathrm{tr}(\mathbf{X}\mathbf{A})$.

Differentiating Determinants

Let x_{ij} be the (i, j)th element of $\mathbf{X}$, and let $|\mathbf{X}_{ij}|$ be its cofactor in $|\mathbf{X}|$. Then, for $\mathbf{X}$ having functionally unrelated elements:

$$\frac{\partial |\mathbf{X}|}{\partial x_{ij}} = |\mathbf{X}_{ij}|, \qquad \frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}|(\mathbf{X}^{-1})',$$

and

$$\frac{\partial}{\partial \mathbf{X}} \log |\mathbf{X}| = (\mathbf{X}^{-1})'.$$

For symmetric $\mathbf{X}$, comparable results are

$$\frac{\partial |\mathbf{X}|}{\partial \mathbf{x}_{ij}} = (2 - \delta_{ij})|\mathbf{X}_{ij}|,$$

where $\delta_{ij} = 0$ except when $i = j$, and then $\delta_{ii} = 1$,

$$\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}|[2\mathbf{X}^{-1} - \mathrm{diag}(\mathbf{X}^{-1})],$$

and

$$\frac{\partial}{\partial \mathbf{X}} \log |\mathbf{X}| = 2\mathbf{X}^{-1} - \mathrm{diag}(\mathbf{X}^{-1}).$$

Finally, for any nonsingular $\mathbf{X}$, symmetric or not,

$$\frac{\partial}{\partial \mathbf{y}} \log |\mathbf{X}| = \operatorname{tr}\left(\mathbf{X}^{-1} \frac{\partial \mathbf{X}}{\partial \mathbf{y}}\right).$$

Jacobians

When $\mathbf{y}$ is a vector of n differentiable functions of the n elements of $\mathbf{x}$, such that the transformation of $\mathbf{x}$ to $\mathbf{y}$, to be denoted $\mathbf{x} \to \mathbf{y}$, is 1-to-1, then the matrix

$$\mathbf{J}_{\mathbf{x}\to\mathbf{y}} = \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right) = \left\{ \underset{\mathrm{m}}{\frac{\partial x_j}{\partial y_i}} \right\}_{i=1 \;\; j=1}^{n \quad n}$$

is the *Jacobian matrix* of $\mathbf{x} \to \mathbf{y}$. For example, if $\mathbf{y} = \mathbf{A}\mathbf{x}$,

$$\mathbf{J}_{\mathbf{x}\to\mathbf{y}} = \left[\frac{\partial (\mathbf{A}^{-1}\mathbf{y})'}{\partial \mathbf{y}}\right] = (\mathbf{A}^{-1})'.$$

$||\mathbf{J}_{\mathbf{x}\to\mathbf{y}}||$, the positive value of the determinant of $\mathbf{J}_{\mathbf{x}\to\mathbf{y}}$, is called the *Jacobian* of $\mathbf{x} \to \mathbf{y}$. It is needed when using $\mathbf{x} \to \mathbf{y}$ on an integral such as

$$\varphi = \int \mathrm{f}(\mathbf{x})\, \mathrm{d}\mathbf{x},$$

where $\mathrm{f}(\mathbf{x})$ is a scalar function of elements of $\mathbf{x}$. If $\mathbf{x} \to \mathbf{y}$ is $\mathbf{y} = g(\mathbf{x})$, then

$$\varphi = \int \mathrm{f}(g^{-1}[y])||\mathbf{J}_{\mathbf{x}\to\mathbf{y}}||\, \mathrm{d}\mathbf{y}.$$

With the identity $||\mathbf{J}_{\mathbf{x}\to\mathbf{y}}|| \equiv 1/||\mathbf{J}_{\mathbf{y}\to\mathbf{x}}||$, with elements of $\mathbf{J}_{\mathbf{y}\to\mathbf{x}}$ sometimes being easier to derive than those of $\mathbf{J}_{\mathbf{x}\to\mathbf{y}}$, and when notation other than $\mathbf{x}$ and $\mathbf{y}$ is the context, confusion easily arises as to whether φ involves $\mathbf{J}_{\mathbf{x}\to\mathbf{y}}$ or $\mathbf{J}_{\mathbf{y}\to\mathbf{x}}$. Fortunately, there is a mnemonic that clarifies the situation. Defining the transformation as old $\to$ new, one always uses $\mathbf{J}_{\mathrm{old}\to\mathrm{new}}$, abbreviated to $\mathbf{J}_{\mathrm{o}\to\mathrm{n}}$. In the latter the subscripts are always in the sequence "on", not "no". This always works.

Vec and Vech Operators

Vec and vech are operators that vectorize a matrix. It can be done in various ways, the most useful of which is stacking the columns of a matrix one under the other. For $\mathbf{X}_{p\times q}$, the resulting column is denoted vec $\mathbf{X}$, a column of order pq. For example,

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ a & b & c \end{bmatrix} \text{ gives } \operatorname{vec} \mathbf{X} = \begin{bmatrix} 1 \\ a \\ 2 \\ b \\ 3 \\ c \end{bmatrix}.$$

Three useful properties are as follows:

$$\begin{aligned} \operatorname{vec}(\mathbf{ABC}) &= (\mathbf{C}' \otimes \mathbf{A})\operatorname{vec} \mathbf{B}, \\ \operatorname{tr}(\mathbf{AB}) &= (\operatorname{vec} \mathbf{A}')'\operatorname{vec} \mathbf{B}, \\ \operatorname{tr}(\mathbf{AZ}'\mathbf{BZC}) &= \operatorname{tr}[\mathbf{Z}'(\mathbf{BZCA})] \\ &= (\operatorname{vec} \mathbf{Z})'\operatorname{vec}(\mathbf{BZCA}) \\ &= (\operatorname{vec} \mathbf{Z})'(\mathbf{A}'\mathbf{C}' \otimes \mathbf{B})\operatorname{vec} \mathbf{Z}. \end{aligned}$$

The operator vech $\mathbf{X}$ is defined for $\mathbf{X}$ being symmetric. It has the columns of $\mathbf{X}$, starting at the diagonal elements, stacked one under the other. For example,

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & x & y \\ 3 & y & \alpha \end{bmatrix} \text{ has } \operatorname{vech} \mathbf{X} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ x \\ y \\ \alpha \end{bmatrix}.$$

Henderson & Searle [1, 2] give some history and numerous details.

A particular use of vec and vech is in calculating $||\mathbf{J}_{\mathbf{X}\to\mathbf{Y}}||$. This is the positive value of the determinant of

$$\mathbf{J}_{\mathbf{X}\to\mathbf{Y}} = \frac{\partial (\operatorname{vec} \mathbf{X})'}{\partial (\operatorname{vec} \mathbf{Y})'},$$

and if $\mathbf{X}$ and $\mathbf{Y}$ are both symmetric, vec is replaced by vech.

Matrices having Complex Numbers as Elements

Because statistics almost always deals with real numbers (e.g. data) and not complex numbers that involve $i = \sqrt{-1}$, most of this article deals with real matrices, those having no complex numbers as elements. Nevertheless, since many texts do deal with

matrices of complex numbers, a few basic definitions are given here.

In scalar arithmetic the complex number $a - ib$ is called the *complex conjugate* of $a + ib$, and the two numbers are a *conjugate pair*. Likewise with matrices, $\overline{\mathbf{M}} = \mathbf{A} - i\mathbf{B}$ is the *complex conjugate* of $\mathbf{M} = \mathbf{A} + i\mathbf{B}$, with $\mathbf{M}$ and $\overline{\mathbf{M}}$ being a *conjugate pair*. $\mathbf{M}$ is said to be *Hermitian* when $\overline{\mathbf{M}}' = \mathbf{M}$; and $\mathbf{M}$ is *unitary* if $\overline{\mathbf{M}}'\mathbf{M} = \mathbf{I}$. Thus, being Hermitian is the complex counterpart of being symmetric, as is unitary of orthogonal.

Some Matrix Usage in Statistics

The development and description of statistical methodology benefits enormously from the use of matrices. The following examples briefly illustrate some of the widely used situations in which matrix notation so efficiently encapsulates a multitude of results.

Means and Variances

$\mathbf{x}$ being a vector of random variables with mean $\boldsymbol{\mu}$ implies that $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$, where E represents expectation. Then, because the ith element of $\mathbf{x}$ has a variance, σ_i^2, and each pair of elements, the ith and jth say, have a covariance, σ_{ij}, these variances and covariances can be arrayed in a symmetric matrix, called the **variance–covariance matrix**, for which we use the symbol $\boldsymbol{\Sigma}$. For example, for $\mathbf{x}$ of order 3,

$$\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}.$$

A more general expression is

$$\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{x}) = \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'.$$

For a linear change of variables, from $\mathbf{x}$ to $\mathbf{y} = \mathbf{T}\mathbf{x}$, the mean vector and the variance–covariance matrix are easily established as

$$\mathrm{E}(\mathbf{y}) = \mathrm{E}(\mathbf{T}\mathbf{x}) = \mathbf{T}\mathrm{E}(\mathbf{x}) = \mathbf{T}\boldsymbol{\mu}$$

and

$$\begin{aligned} \mathrm{var}(\mathbf{y}) = \mathrm{var}(\mathbf{T}\mathbf{x}) &= \mathrm{E}(\mathbf{T}\mathbf{x} - \mathbf{T}\boldsymbol{\mu})(\mathbf{T}\mathbf{x} - \mathbf{T}\boldsymbol{\mu})' \\ &= \mathrm{E}\mathbf{T}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\mathbf{T}' = \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}'. \end{aligned}$$

Suppose that $\mathbf{T}$ is a row vector, $\mathbf{t}'$. Then, because a variance is never negative, $\mathrm{var}(\mathbf{t}'\mathbf{y}) = \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} \geq 0$ and so $\boldsymbol{\Sigma}$ is n.n.d.

Correlation

A **correlation** matrix, $\mathbf{P}$ say, is a matrix with 1.0 as its diagonal elements and correlations $\rho_{ij} = \sigma_{ij}/(\sigma_i^2\sigma_j^2)^{1/2}$ (for $i \neq j$) as its off-diagonal elements. Define $\mathbf{D}$ as the diagonal matrix of the σ_i^2 terms. Then $\mathbf{P} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$.

A frequently used form of $\boldsymbol{\Sigma}$ is one which has σ^2 for all diagonal elements (variances) and $\rho\sigma^2$ for all off-diagonal elements (covariances). Then,

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{P} \quad \text{for} \quad \mathbf{P} = (1 - \rho)\mathbf{I} + \rho\mathbf{J},$$

and, for order k,

$$|\boldsymbol{\Sigma}| = \sigma^{2k}(1 - \rho)^{k-1}(1 - \rho + k\rho).$$

Since $\boldsymbol{\Sigma}$ is n.n.d., $|\boldsymbol{\Sigma}| \geq 0$, which implies that $1 + (k - 1)\rho \geq 0$; that is, $\rho \geq -1/(k - 1)$. This is a consequence that one would not be inclined to anticipate on assuming the same covariance, $\rho\,\sigma^2$, between each pair of variables.

Sums of Squares and Products

For a column vector $\mathbf{x}_j$, the jth column of $\mathbf{X}$, the sum of squares of its elements x_{ij} is $\sum_i x_{ij}^2 = \mathbf{x}_j'\mathbf{x}_j$; and $\sum_i x_{ij}x_{ij'} = \mathbf{x}_j'\mathbf{x}_{j'}$. Thus, $\mathbf{X}'\mathbf{X} = \{_{\mathrm{m}}\, \mathbf{x}_j'\mathbf{x}_{j'}\}$ is a matrix of these sums of squares and products.

For $\mathbf{x}_j$ having n elements, define $\mathbf{C}_n = \mathbf{I}_n - \overline{\mathbf{J}}_n$, the centering matrix of order n. Then, $\mathbf{X}'\mathbf{C}_n\mathbf{X}$ has terms $\sum_i (x_{ij} - \overline{x}_{\cdot j})^2$ in its diagonal and terms $\sum_i (x_{ij} - \overline{x}_{\cdot j})(x_{ij'} - \overline{x}_{\cdot j'})$ as its off-diagonal elements, with $\overline{x}_{\cdot j} = 1_n'\mathbf{x}_j/n$. It is the matrix of sums of squares and products corrected for the mean.

If $\boldsymbol{\Delta}$ is defined as the diagonal matrix of the diagonal elements of $\mathbf{X}'\mathbf{C}\mathbf{X}$, then the correlation matrix $\mathbf{P} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$ described earlier is estimated by $\mathbf{R} = \boldsymbol{\Delta}^{-1/2}\mathbf{X}'\mathbf{C}\mathbf{X}\boldsymbol{\Delta}^{-1/2}$.

The Multivariate Normal Distribution

The density function of a normally distributed random variable x having mean μ and variance σ^2 is $\{\exp[-\frac{1}{2}(x - \mu)^2/\sigma^2]\}/(2\pi\sigma^2)^{1/2}$. The counterpart of this for a vector $\mathbf{x}$ of random variables

distributed $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, means that this x has mean $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$, with a **multivariate normal distribution**, which is $\{\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]\}/(|2\pi\boldsymbol{\Sigma}|)^{1/2}$. The **moment generating function** of linear combinations $\mathbf{Kx}$ of $\mathbf{x}$ is $\exp(\mathbf{t}'\mathbf{K}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\mathbf{K}\boldsymbol{\Sigma}\mathbf{K}'\mathbf{t})$.

A very neat consequence of using matrices is the derivation of marginal and conditional distributions in the multivariate normal distribution $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It stems from partitioning $\mathbf{x}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

with $\boldsymbol{\Sigma}_{21} = (\boldsymbol{\Sigma}_{12})'$. Then a marginal distribution is

$$\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

and a conditional distribution is

$$\mathbf{x}_1|\mathbf{x}_2 \sim N[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}].$$

Details are available in Searle [3, Section 2.4f].

Quadratic Forms

Every sum of squares is a homogeneous second-degree function of data. It can therefore be represented as a quadratic form $\mathbf{x}'\mathbf{Ax}$ for $\mathbf{x}$ being the vector of data and $\mathbf{A}$ being symmetric. A variety of properties pertaining to $\mathbf{x}'\mathbf{Ax}$ are then available for whatever sums of squares one is interested in. Some of these properties for $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are as follows:

1. $E(\mathbf{x}'\mathbf{Ax}) = \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$ (normality is not needed for this result);
2. $\text{var}(\mathbf{x}'\mathbf{Ax}) = 2\text{tr}(\mathbf{A}\boldsymbol{\Sigma})^2 + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}$;
3. $\mathbf{x}'\mathbf{Ax}$ has a (noncentral) **chi-square distribution** if and only if $\mathbf{A}\boldsymbol{\Sigma}$ is idempotent;
4. $\mathbf{x}'\mathbf{Ax}$ and $\mathbf{Lx}$ are stochastically independent if and only if $\mathbf{L}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$;
5. $\mathbf{x}'\mathbf{Ax}$ and $\mathbf{x}'\mathbf{Bx}$ are stochastically independent if and only if $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$ or, equivalently, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{0}$.

Regression and Linear Models

There is an enormous volume of literature on these topics, most of it using matrix algebra. Only a minute sampling of it is given here.

Consider a vector of data $\mathbf{y}$, modeled as having expected value $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ with $\mathbf{X}$ being known and $\boldsymbol{\beta}$ being a vector of unknown parameters. Defining $\boldsymbol{\varepsilon}$ as $\mathbf{y} - E(\mathbf{y})$, a vector of residuals leads to modeling $\mathbf{y}$ as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. **Least squares** estimation of $\boldsymbol{\beta}$ dictates minimizing $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and taking the resulting value of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, as the estimator of $\boldsymbol{\beta}$. This leads to the equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$. In regression (*see* **Multiple Linear Regression**) $\mathbf{X}$ almost always has full column rank, so that $\mathbf{X}'\mathbf{X}$ is nonsingular and hence $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. But with many more **general linear models** $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist and a generalized inverse $(\mathbf{X}'\mathbf{X})^{-}$ has to be used. In that case there are many solutions for $\hat{\boldsymbol{\beta}}$, and to indicate this they can be denoted by $\boldsymbol{\beta}^o$. Thus, $\boldsymbol{\beta}^o = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$.

Since $\boldsymbol{\beta}^o$ becomes $\hat{\boldsymbol{\beta}}$ when $(\mathbf{X}'\mathbf{X})^{-1}$ exists, the properties of $\hat{\boldsymbol{\beta}}$ are included among those of $\boldsymbol{\beta}^o$, just a few of which are as follows:

1. There are many solutions, $\boldsymbol{\beta}^o$, but for each of them $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}^o = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ is the same, because $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is invariant to $(\mathbf{X}'\mathbf{X})^{-}$.
2. $E(\boldsymbol{\beta}^o) \neq \boldsymbol{\beta}$, but $E(\mathbf{X}\boldsymbol{\beta}^o) = \mathbf{X}\boldsymbol{\beta}$.
3. The residual sum of squares

$$\text{SSE} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$$

is invariant to $(\mathbf{X}'\mathbf{X})^{-}$. Because it can be expressed as $\text{SSE} = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}']\mathbf{y}$, with the matrix being idempotent, the expected value of SSE for $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$ is $E(\text{SSE}) = [N - r(\mathbf{X})]\sigma^2$. Also, SSE/σ^2 has a chi-square distribution. Moreover, the sum of squares due to fitting the model is $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$; it too has a (noncentral) chi-square distribution, and it is stochastically independent of SSE.

Readers whose appetite has been whetted by this introduction to regression and linear models will find plenty of books and papers to satiate their hunger.

References

[1] Henderson, H.V. & Searle, S.R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics, *Canadian Journal of Statistics* **7**, 65–81.

[2] Henderson, H.V. & Searle, S.R. (1981). The vec permutation matrix, the vec operator and Kronecker products: a review, *Linear and Multilinear Algebra* **9**, 271–288.

[3] Searle, S.R. (1971). *Linear Models*. Wiley, New York.
[4] Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.

(*See also* **Matrix Computations**)

SHAYLE R. SEARLE

Matrix Computations

Much of the development and formulation of the mathematical and statistical models that biostatisticians use relies heavily on matrix notation (*see* **Matrix Algebra**). For example, matrix notation is often the preferred way to describe the mathematics underlying many of the procedures in statistical packages (*see* **Software, Biostatistical**). Routines are now widely available that implement standard matrix operations including matrix multiplication and the solution of systems of linear equations, facilitating computer implementation of matrix formulas presented in the literature. Some knowledge of the alternative available **algorithms** is helpful both for implementers of matrix formulas and for users of existing statistical package implementations. In the following we comment on some alternative widely used approaches to linear **least squares** and related calculations.

Areas where matrix computations have a large place include **regression** methods, **multivariate analysis**, **maximum likelihood** estimation, **robust** estimation, smoothing, and **optimization**. Linear matrix computational methods are more generally important because nonlinear problems are frequently handled by solving a sequence of linearized problems. Numerical linear algebra is, effectively, another name for matrix computations.

Modern numerical matrix algebra gains much of its power from the use of a relatively small number of matrix decompositions, whose numerical properties are well understood. Major aims are guaranteed accuracy, speed of computation (efficiency), and the ability to handle all inputs [6, 7, 9]. The article [9] discusses several topics that we omit or only mention in passing.

Implementing Matrix Computations

Matrix computations must reckon with the finite precision of computer arithmetic. Most common computers now implement the IEEE standard for **floating point arithmetic**, which has around seven decimal digits single-precision and around 16 decimal digit double-precision arithmetic. The double-precision standard is a sound basis on which to build accurate and reliable algorithms.

Technical accuracy and efficiency issues are reasons for providing expert "black box" implementations of what might appear simple calculations such as $||\mathbf{x}|| = (\mathbf{x}'\mathbf{x})^{1/2}$ and matrix multiplication. Specifications for sets of lower-level routines have been established in the **numerical analysis** literature, where they are known as BLAS (basic linear algebra subroutines) [1]. The BLAS, or other such lower-level routines, then make effective building blocks in the creation of higher-level routines.

Understanding Matrix Methods

There are often, in matrix computations just as elsewhere, several different ways to solve the same problem. Knowledge of matrix algorithms may allow the substitution of one algorithm for another when required. For example, a published formula may involve a matrix operation not found in available software. Additional information that is required from a routine may be available, for someone who understands the algorithm, as an adaptation of existing output.

Often it is helpful to know what accuracy can reasonably be expected from a calculation. When results from different algorithms for the same problem differ numerically (perhaps in decimal places after the third or fourth), which is more accurate? What characteristics of input data may lead to such differences in accuracy? Knowledge of the algorithm may be even more important when a calculation fails.

Matrix Inversion

The use of matrix inverses is a convenience in writing down matrix formulas. However, direct implementation of such formulas rarely leads to algorithms that are optimal for practical computation. For example, solving $\mathbf{Sb} = \mathbf{c}$ for $\mathbf{b}$ is preferable to forming $\mathbf{S}^{-1}$ and computing $\mathbf{b} = \mathbf{S}^{-1}\mathbf{c}$. Avoiding unnecessary matrix inversion reduces computational effort, leading to a small improvement in precision. There is a choice of default actions where the inverse does not exist. Later in this article we illustrate approaches which avoid the explicit calculations of matrix inverses.

Linear Least Squares

Linear **least squares** has been the context for much of the discussion of statistical matrix computations. As

well as being important for linear least squares, the matrix computations we describe are important building blocks for many other statistical computations.

We consider the contrived example

$$(\mathbf{X}|\mathbf{y}) = \left[\begin{array}{rrr|r} 1 & 7 & 8 & 6 \\ 1 & -3 & 4 & 4 \\ 1 & 2 & 2 & 0 \\ 1 & 2 & 2 & 6 \\ 1 & 7 & 6 & 5 \\ 1 & 2 & 4 & 7 \\ 1 & -3 & 2 & 3 \\ 1 & 2 & 4 & 1 \\ 1 & 2 & 4 & 4 \end{array}\right].$$

Given $\mathbf{X}(n \times p)$ and $\mathbf{y}(n \times 1)$, least squares calculations determine $\mathbf{b}(p \times 1)$ such that

$$(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = ||\mathbf{y} - \mathbf{Xb}||^2 \qquad (1)$$

is a minimum. In the example above one minimizes the sum of squares $[6 - (b_1 + 7b_2 + 8b_3)]^2 + [4 - (b_1 - 3b_2 + 4b_3)]^2 + \cdots$

Algebraically, the linear least squares problem (1) is equivalent to solving what are called the normal equations, i.e.

$$\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}. \qquad (2)$$

If $\mathbf{S} = \mathbf{X}'\mathbf{X}$ is singular, theoretical arguments show that the normal equations are consistent, but rather than just one solution there are an infinity of solutions. An example appears below in the section on linear dependencies.

We describe and contrast two approaches to the linear least squares problem, one of which forms and solves the normal equations, while the other (the QR method) avoids formation of the normal equations.

A Normal Equation Approach

An effective way to solve the normal equations is to use the Cholesky algorithm, which modifies Gaussian elimination to take advantage of the symmetry of the normal equation matrix of coefficients. Diagrammatically, the steps are

$$(\mathbf{X}|\mathbf{y}) \rightarrow \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{y} \\ (\mathbf{X}'\mathbf{y})' & \mathbf{y}'\mathbf{y} \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{R} & \mathbf{d} \\ \mathbf{0}' & r_{yy} \end{pmatrix}. \qquad (3)$$

The normal equations $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}$ reduce to $\mathbf{Rb} = \mathbf{d}$, where $\mathbf{R}$ is $p \times p$ upper triangular, i.e. below diagonal elements are zero. (It might also be described as right triangular, which perhaps justifies the symbol $\mathbf{R}$.) It is convenient to take $\mathbf{R}$ to be the upper triangular matrix which is formed by the Cholesky decomposition of $\mathbf{X}'\mathbf{X}$, i.e. $\mathbf{R}'\mathbf{R} = \mathbf{X}'\mathbf{X}$. On the right-hand side of (3), $\mathbf{R}$ is augmented with an additional row and column, to form an array which is the Cholesky decomposition of $(\mathbf{X}|\mathbf{y})'(\mathbf{X}|\mathbf{y})$.

For our numerical example the system of equations $\mathbf{Rb} = \mathbf{d}$ is

$$\begin{pmatrix} 3 & 6 & 12 \\ 0 & 10 & 4 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 2 \\ 2 \end{pmatrix}.$$

Calculations are completed by solving first for b_3 $\left(=\frac{1}{2}\right)$, then for $b_2(=0)$ in terms of b_3, and finally for $b_1(=2)$ in terms of b_2 and b_3.

The QR Method for Linear Least Squares

Our description will emphasize points of contact with the normal equations approach. The QR method omits the intermediate step in (3). It determines

$$\mathbf{Q}(\mathbf{X}|\mathbf{y}) = \begin{pmatrix} \mathbf{R} & \mathbf{d} \\ \mathbf{0} & \mathbf{z} \end{pmatrix}, \qquad (4)$$

where $\mathbf{Q}(n \times n)$ is a product of orthogonal matrices and is hence orthogonal, i.e. $\mathbf{Q}'\mathbf{Q} = \mathbf{QQ}' = \mathbf{I} = \text{diag}(1, \ldots, 1)$. The vector $\mathbf{z}$ is $(n - p) \times 1$. If we insist that $\mathbf{R}$ have positive diagonal elements, then it is algebraically identical to the matrix $\mathbf{R}$ formed by the Cholesky decomposition of $\mathbf{X}'\mathbf{X}$. The quantity $\mathbf{z}'\mathbf{z}$ is the sum of squares of residuals from the regression, and equals r_{yy}^2.

Other Methods for Least Squares

Other methods for least squares include the once popular Gauss–Jordan scheme, which calculates $(\mathbf{X}'\mathbf{X})^{-1}$ as well as $\mathbf{b}$. There are in addition a range of iterative methods for least squares, which have found particular application in large sparse problems [3, 6, 9].

Linear Dependencies

In the data set

$$(\mathbf{X}|\mathbf{y}) = \left[\begin{array}{ccc|c} 1 & -2 & -4 & -1 \\ 1 & 1 & -1 & 0 \\ 1 & 2 & 0 & 4 \\ 1 & 5 & 3 & 7 \end{array}\right],$$

the third column is the difference between the second column and twice the first column. Linear dependencies, of which this is a trivial example, arise in least squares problems when one or more variables are a linear combination of earlier variables. The normal equations are

$$\begin{pmatrix} 4 & 6 & -2 \\ 6 & 34 & 22 \\ -2 & 22 & 26 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 10 \\ 45 \\ 25 \end{pmatrix}.$$

The matrix of coefficients $\mathbf{S} = \mathbf{X}'\mathbf{X}$ is, because the coefficients in row 3 are the difference between row 2 and twice row 1, *singular*. Hence $\mathbf{S}^{-1}$ does not exist. Nevertheless the coefficients are, because from normal equations, consistent. With b_3 chosen arbitrarily, $b_2 = 1.2 - b_3$ and $b_1 = 0.7 + 2b_3$. Such nonunique solutions occur when one variable or term in a model is a linear combination of other terms.

In **analysis of variance** applications, dependencies may arise because there are inadequate data to allow the estimation of one or more parameters associated with main effects or interactions. Alternately, one or more **explanatory** variables may be an exact linear combination of other terms in the model, and a decision is needed on which terms to include. Dependencies may be a result of an unanticipated feature of the input data, or of a mistake in the data.

Dependencies are, when working with observational data on a large number of variables, surprisingly common. They may be a huge source of frustration, especially if the program responds by exiting with an uninformative error message. Sensible default actions, and information on the coefficients of the linear relation, may be a huge help. Where column i of $\mathbf{X}$ is a linear combination of earlier columns, an easy device which will allow calculations to continue is to set b_i to zero, effectively deleting column i of $\mathbf{X}$. It would be useful to have criteria for detecting instances where a near singularity may make results nonsensical or hard to interpret. Regrettably, there are no effective simple criteria that will cover all circumstances.

Normal Equations vs. QR

At a fixed level of numerical precision the QR decomposition will solve a wider range of problems than normal equation methods. The difference is marked when there are strong dependencies between the columns of $\mathbf{X}$, leading to a large **standard error** for one or more elements of $\mathbf{b}$. A consequence of large standard error(s) is that the additional numerical precision is unlikely to be statistically meaningful.

The solution of the normal equations retains very nearly the accuracy of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$. Where $\mathbf{X}$ has an initial column of ones, precision may be assisted by expressing values in remaining columns as differences from the column mean, prior to forming $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$. Careful implementations of normal equation methods take this precaution. The precision of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ is then equivalent to that of an accurately formed correlation matrix. In applications in the biological and social sciences, where differences from the mean are rarely accurate to more than two or three significant digits, this seems adequate precision.

Caution may nevertheless advise use of the QR method except in those applications – unbalanced analysis of variance, for example – where columns of $\mathbf{X}$ are unlikely to be highly correlated. There is a helpful discussion in [5] which compares the normal equation method with QR (see also [6]).

QR Algorithms

Another name for the QR method is orthogonal reduction to upper triangular form. Available algorithms for QR include Householder and modified Gram–Schmidt (MGS), which proceed columnwise, and the Givens algorithm, which operates on new rows one at a time to incorporate them into the current version of $\mathbf{R}$. We discuss these in more detail below. Elements of $\mathbf{Q}$ are unlikely to be stored explicitly; instead, key quantities are stored from which $\mathbf{Q}$ can be reconstructed as required.

Algorithms for QR factorization effectively form rows of $\mathbf{R}$ as linear combinations of rows of $\mathbf{X}$ rather than as linear combinations of rows of $\mathbf{X}'\mathbf{X}$. They avoid the loss of accuracy which, in normal equation methods, may occur in the formation of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$.

There is some additional computational cost. When p is much smaller than n, use of QR approximately doubles the number of multiplications and divisions compared with using the normal equations.

Various diagnostic and other information that may be required for least squares modeling may be computed straightforwardly from submatrices of **Q**. Examples include leverage statistics (*see* **Diagnostics**), and the variance–**covariance matrix** of **residuals**. Brief details appear in a later section.

Some Key Matrix Methods

Here we discuss in more detail several algorithms that have major importance in statistical computation, including algorithms mentioned above. We emphasize the connections between algorithms which, to first appearance, are quite different.

Cholesky Decomposition

Given a positive definite matrix **S**, perhaps formed as $\mathbf{X'X}$, the Cholesky decomposition determines an upper triangular matrix **R** such that $\mathbf{S} = \mathbf{R'R}$. Equivalently, one may form $\mathbf{S} = \mathbf{U'DU}$, where **D** is diagonal and **U** is upper triangular with unit diagonal.

Several algorithms are available, which differ in the order in which they form elements of **R**. In the version we now describe, elements in the first row of **R** are formed as

$$r_{11} = \sqrt{s_{11}}, \qquad r_{1j} = r_{11}^{-1} s_{1j}, \quad j = 1, \ldots, p.$$

Then, for $i = 2, \ldots, p$, calculate

$$s_{ij}^{(i-1)} = s_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}, \quad j = i, \ldots, p,$$

and

$$r_{ii} = [s_{ii}^{(i-1)}]^{1/2}, \quad \text{if } (i < p)\ r_{ij} = r_{ii}^{-1} s_{ij}^{(i-1)}, \quad j = i+1, \ldots, p.$$

Note that if **X** has an initial column of ones and remaining columns are centered by expressing values as differences from the column mean, then $1 - s_{ii}^{-1} s_{ii}^{(i-1)}$ is the squared multiple correlation measuring the dependence of column k of **X** on earlier columns. Where $r_{ii} = 0$, all elements in that row may be set to zero.

The Cholesky decomposition may be used in solving the generalized weighted least squares problem, where **W** is a positive-definite symmetric weighting matrix. Observe that if **U** is upper triangular such that $\mathbf{U'U} = \mathbf{W}$, then

$$(\mathbf{y} - \mathbf{Xb})'\mathbf{W}(\mathbf{y} - \mathbf{Xb}) = (\mathbf{y}^* - \mathbf{X}^*\mathbf{b})'(\mathbf{y}^* - \mathbf{X}^*\mathbf{b}),$$

where $\mathbf{y}^* = \mathbf{Uy}, \mathbf{X}^* = \mathbf{UX}$. This is now in the form of (1).

Simulation from a **multivariate normal distribution** with $p \times p$ variance–covariance matrix $\mathbf{\Sigma} = \mathbf{R'R}$ may be handled by setting $\mathbf{u} = \mathbf{R'x}$, where elements of **x** are independent normal random deviates each with mean 0 and variance 1.

The Householder QR Algorithm

The Householder QR algorithm has wide application apart from least squares. It is, for example, used in forming the singular value decomposition, which we describe below. It is usually motivated by describing the matrix **Q** of (4) as a product of Householder *reflections*

$$\mathbf{I} - \frac{2\mathbf{w}_i \mathbf{w}_i'}{\tau_i}, \quad i = 1, \ldots, p,$$

where $\tau_i = ||\mathbf{w}_i||^2$. The first reflection reduces to zero elements all elements except the first in the initial column of **X**, replacing the first row of **X** by the first row of **R**. The second takes the matrix so formed and reduces to zero all elements below the second row in its second column, replacing the second row of this matrix with the second row of **R**. In the adaptation of the Householder method, for which we give algebraic details, one or more rows of **R** may differ from the result of applying Householder reflections by a change of sign of all elements in the row [8]. This simplifies the detailed algebraic description and simplifies the algorithm.

Let $\mathbf{x}_j^{(k-1)}$ $(j \geq k)$ be the result of applying rotations $1, \ldots, k-1$ to column j of **X**, but with elements $1, \ldots, k-1$ set to zero when $k > 1$. Then

$$r_{kk} = \left\| \mathbf{x}_k^{(k-1)} \right\|, \qquad r_{kj} = r_{kk}^{-1} \left(\mathbf{x}_k^{(k-1)} \right)' \mathbf{x}_j^{(k-1)}, \quad j > k. \tag{5}$$

For elements in rows after the kth we use

$$\mathbf{x}_j^{(k)} = \mathbf{x}_j^{(k-1)} - \alpha_k^{-1}\left(x_{kj}^{(k-1)}\mathrm{sgn}\left(x_{kk}^{(k-1)}\right) + r_{kj}\right)\mathbf{x}_k^{(k-1)}, \quad j > k,$$

where $\alpha_k = |x_{kk}| + r_{kk}$.

Where a column is a linear combination of earlier columns, this leads to $r_{ii} = 0$. The easiest way to deal with this is to move any such column to the final column position. More generally, the columns of $\mathbf{X}$ may be permuted so that columns which are highly dependent on earlier columns are taken last – a device known as *pivoting*. The initial order of columns can, if this is required, be restored when calculations are complete. Additional orthogonal rotations may be required to recover the matrix $\mathbf{R}$ that corresponds to the original ordering.

The Modified Gram–Schmidt QR Algorithm

If the variant of Householder just described is applied to a matrix $\mathbf{X}$ which is augmented with p initial rows of zeros, this leads, essentially, to the modified **Gram–Schmidt** (MGS) algorithm [8]. The MGS algorithm may be described in terms of residuals from repeated regressions. This statistical interpretation is a main reason for mentioning it here.

Let $\mathbf{e}_j^{(k-1)}$ be the vector of residuals when the column j, $j \geq k$, of $\mathbf{X}$ is regressed on columns $1, \ldots, k-1$. Then the MGS algorithm forms

$$r_{kk} = \left\|\mathbf{e}_k^{(k-1)}\right\|, \qquad r_{kj} = r_{kk}^{-1}\left(\mathbf{e}_k^{(k-1)}\right)'\mathbf{e}_j^{(k-1)}.$$

Thus the MGS algorithm uses least squares vectors of residuals to form elements of $\mathbf{R}$. For details see [3, 6–8, 10, 11].

The Givens QR Algorithm

This algorithm operates on $\mathbf{X}$ one row at a time, where Householder and modified Gram–Schmidt operate on columns. It is useful where the QR decomposition must from time to time be updated as new data become available.

The matrix $\mathbf{R}$ is filled initially with zeros. Planar rotations,

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{6}$$

then rotate rows of $\mathbf{X}$, one at a time, into the upper triangular array. Thus $\mathbf{R}$ is sequentially updated as each new row of $\mathbf{X}$ is rotated into the upper triangular scheme. The rotations which operate on row k ($k > p$) of $\mathbf{X}$ replace y_k with z_k, where z_k^2 is the increase in the residual sum of squares when row k is included. The planar rotations in the Givens QR algorithm are often called Givens rotations.

Another use for planar rotations is to remove rows that were earlier included, i.e. to *downdate* $\mathbf{R}$. A stable algorithm requires access to the matrix $\mathbf{Q}$ [3, 6]. The algorithm in [4] is as stable as possible when $\mathbf{Q}$ is not available.

Orthogonalization of the Columns of X

One way to view the QR method is that it reduces the problem of minimizing $||\mathbf{y} - \mathbf{X}\mathbf{b}||$ to that of minimizing $||\mathbf{y} - \mathbf{X}^*\mathbf{b}^*||$, where $\mathbf{X}^* = \mathbf{X}\mathbf{R}^{-1}$ and $\mathbf{b}^* = \mathbf{R}\mathbf{b}$. It replaces $\mathbf{X}$ by a matrix $\mathbf{X}^*$ the columns of which are orthogonal, i.e. $(\mathbf{X}^*)'\mathbf{X}^*$ is the matrix $\mathbf{I} = \mathrm{diag}(1, \ldots, 1)$.

Let

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{pmatrix}, \tag{7}$$

where $\mathbf{Q}_1$ is $p \times n$ and $\mathbf{Q}_2$ is $(n-p) \times n$. Then it may be shown that $\mathbf{Q}_1' = \mathbf{X}\mathbf{R}^{-1}$. Thus, if $\mathbf{Q}$ is available, the matrix $\mathbf{X}^* = \mathbf{X}\mathbf{R}^{-1}$ can be extracted as a submatrix.

Orthogonal Polynomials

Low-order **polynomial** functions are frequently used to provide simple curvilinear models for data. If the covariate $\mathbf{x}$ has elements x_i, $i = 1, \ldots, n$, then calculations can in principle be handled as a least squares calculation in which $\mathbf{X}$ has its (i, j)th element equal to x_i^{j-1}, $i = 1, \ldots, n$; $j = 1, \ldots, p$. This natural representation of the problem produces a matrix $\mathbf{X}$ the columns of which are likely to be strongly **correlated**. This gives coefficients which are strongly correlated, with standard errors which are inflated by amounts which depend on the correlations.

The QR method may be used as discussed in (7) to form the matrix $\mathbf{X}^*$ with orthogonal columns. The first column of $\mathbf{X}^*$ is a constant, the second is a multiple of $\mathbf{x} - \overline{x}$, the third involves terms up to degree two in $\mathbf{x}$, and so on. Even better is

to use recurrence formulas for systems of orthogonal polynomials to generate the columns of $\mathbf{X}^*$ (*see* **Orthogonality**). Such use of orthogonal polynomials gives independent and often more interpretable regression coefficients and avoids numerical instability (*see* **Polynomial Approximation**).

The Deletion and Addition of Columns

Removal of a column of $\mathbf{X}$ is achieved by removing the corresponding column from $\mathbf{R}$ and using a series of Givens rotations to reduce the resulting matrix to upper triangular form. The addition of a further column $\mathbf{x}_{p+1}$ to $\mathbf{X}$ is likewise straightforward, providing $\mathbf{Q}$ is available. A further QR reduction is used to reduce $(\mathbf{R}, \mathbf{Q}\mathbf{x}_{p+1})$ to upper triangular form.

Singular Value Decomposition (SVD)

This decomposition finds application in principal components analysis and in many different related multivariate calculations. It offers yet another approach to least squares calculations [6]. Given an $n \times p$ matrix $\mathbf{X}$, it forms

$$\mathbf{X} = \mathbf{UGV}',$$

where $\mathbf{U}$ is $n \times n$ orthogonal, $\mathbf{V}$ is $p \times p$ orthogonal, and $\mathbf{G}$ is $n \times p$ with its only nonzero elements on the uppermost diagonal, namely the singular values.

One or more singular values that are close to zero indicates that $\mathbf{X}$ is near singular, with the relevant linear relations given by the corresponding columns of $\mathbf{V}$. Note that the singular values of $\mathbf{X}'\mathbf{X}$ are the squares of those of $\mathbf{X}$. The Golub–Kahan algorithm for the singular value decomposition first uses Householder reflections to reduce $\mathbf{X}$ to upper bidiagonal form, i.e. all elements are zero except those on the diagonal or in positions immediately above the diagonal. Repeated planar rotations, (6), then reduce the above diagonal elements to zero [6].

Methods for Singular Matrices

Here we examine several technical issues that arise when matrices are singular or close to singular.

Distance from Singularity

Assume that $\mathbf{X}$ has an initial column of ones and that remaining columns are centered. A statistically motivated measure of the distance of column k of $\mathbf{X}$ from a linear combination of all other columns is the inverse $(s_{kk}s^{kk})^{-1}$ of the *variance inflation factor* $s_{kk}s^{kk}$, where s_{kk} and s^{kk} are the kth diagonal elements of $\mathbf{X}'\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^{-1}$, respectively. This variance inflation factor is the amount by which the standard error of b_k is multiplied because of correlation between column k of $\mathbf{X}$ and other columns. Note the relationship

$$s_{kk}s^{kk} = \left(1 - R^2_{k|1,\ldots,k-1,k+1,\ldots,p}\right)^{-1}, \qquad (8)$$

with the squared multiple correlation $R^2_{k|1,\ldots,k-1,k+1,\ldots,p}$ measuring the dependence of explanatory variable k upon other explanatory variables.

Which are the Linear Dependencies?

Let $\mathbf{r}_k^{(k-1)}$ consist of elements 1 to $k-1$ in column k of $\mathbf{R}$. Let $\mathbf{R}_{11}^{(k-1)}$ be the leading $(k-1) \times (k-1)$ submatrix of $\mathbf{R}$. Then the vector of coefficients in the least squares regression of column k of $\mathbf{X}$ on earlier columns is found by solving for $\mathbf{h}$ in

$$\mathbf{R}_{11}^{(k-1)}\mathbf{h} = \mathbf{r}_k^{(k-1)}.$$

(If $r_{ii} = 0$ for one or more $i < k$, then set $h_i = 0$.)

Suppose that m diagonal elements of $\mathbf{R}$ are zero. Then by determining all such vectors $\mathbf{h}$ we can construct a matrix $\mathbf{H}(p \times m)$ with maximum rank m such that

$$\mathbf{XH} = \mathbf{0}. \qquad (9)$$

Columns of $\mathbf{H}$ have the form $(h_1, h_2, \ldots, h_{k-1}, -1, 0, \ldots, 0)'$. The columns of $\mathbf{H}$ are a basis for the orthogonal complement of the row space of $\mathbf{X}$. The general solution to the least squares problem is $\tilde{\mathbf{b}} = \mathbf{b} + \mathbf{Hc}$, where $\mathbf{c}$ is arbitrary. One way to make $\tilde{\mathbf{b}}$ unique is to choose $\mathbf{c}$ so that $\tilde{\mathbf{b}}$ has minimum length, which is itself a least squares problem [8], pp. 106–107, 119. The easiest choice is $\mathbf{c} = \mathbf{0}$.

A Reflexive g-inverse of $\boldsymbol{R}$

Let $\mathbf{R}^-$ be obtained from $\mathbf{R}$ by replacing zero diagonal elements r_{ii} with 1, inverting the resulting matrix, and then placing zeros in the rows and columns where $r_{ii} = 0$. Then

$$\mathbf{RR}^-\mathbf{R} = \mathbf{R}, \mathbf{R}^-\mathbf{RR}^- = \mathbf{R}^-,$$

which are the conditions for $\mathbf{R}^-$ to be a *reflexive g-inverse* of $\mathbf{R}$. The matrix $\mathbf{R}^-$ may be used in the calculation of variances and covariances of regression coefficients that correspond to the choice $\mathbf{c} = \mathbf{0}$ above.

Applications

We give a few examples where an elegant alternative to matrix inversion reduces computational effort. In part our aim is to move away from an exclusive focus on least squares.

Leverages and Standard Errors of Residuals

In least squares, the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ may be calculated as $\mathbf{X}\mathbf{R}^{-1}(\mathbf{X}\mathbf{R}^{-1})'$. If $(\mathbf{X}\mathbf{R}^{-1})'$ is not already available, it may be determined by solving for columns of $\mathbf{Q}_1$ in the lower triangular system of equations $\mathbf{R}'\mathbf{Q}_1 = \mathbf{X}'$ (cf. (7)). The leverage statistic h_i, which is the ith diagonal element of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, may be calculated as the sum of squares of elements of the ith column of $\mathbf{Q}_1$. Note also that $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{Q}_2'\mathbf{Q}_2$. Thus $\mathbf{Q}_2$ may be used in calculating the variance–covariance matrix of residuals.

Partial Sums of Squares and Products

We show how to form partial correlations between columns of $\mathbf{Y}$ $(n \times q)$, conditional on columns of $\mathbf{X}$ $(n \times p)$. Let $\mathbf{Z} = \mathbf{(1, X, Y)}$, where $\mathbf{1}$ is a column of ones. Now use the QR algorithm to form

$$\mathbf{QZ} = \begin{pmatrix} \sqrt{n} & \mathbf{u}_1' & \mathbf{u}_2' \\ \mathbf{0} & \mathbf{R}_{XX} & \mathbf{R}_{XY} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_{YY} \end{pmatrix}, \qquad (10)$$

where $\mathbf{u}_1'$ is $1 \times p$, $\mathbf{u}_2'$ is $1 \times q$, $\mathbf{R}_{XX}$ is $p \times p$, $\mathbf{R}_{XY}$ is $p \times q$, and $\mathbf{R}_{YY}$ is $q \times q$.

Then $\mathbf{R}_{YY}'\mathbf{R}_{YY}$ is the matrix of sums of squares and products of the q vectors of residuals from the regressions of columns of $\mathbf{Y}$ on columns of $\mathbf{X}$. The corresponding matrix of partial correlations is $\mathbf{D}^{-1/2}\mathbf{R}_{YY}'\mathbf{R}_{YY}\mathbf{D}^{-1/2}$, where $\mathbf{D}^{-1/2}$ is the diagonal matrix whose elements are the inverses of the square roots of the diagonal elements of $\mathbf{R}_{YY}'\mathbf{R}_{YY}$.

Canonical Correlation

We assume the orthogonal reduction in (10) above. Then computations may be handled by solving the symmetric eigenproblem

$$|\mathbf{R}_{XY}'\mathbf{R}_{XY} - \lambda\mathbf{R}_{YY}'\mathbf{R}_{YY}| = \mathbf{0}. \qquad (11)$$

This may be rewritten as

$$|(\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1})'\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1} - \lambda\mathbf{1}| = \mathbf{0},$$

which can be solved by finding the singular value decomposition of $\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1}$. The **canonical correlations** ϕ_i, where i runs from 1 to min [rank $(\mathbf{R}_{XX})$, rank $(\mathbf{R}_{YY})$], are given by

$$\phi_i^2 = \frac{\lambda_i}{1 + \lambda_i}$$

([8], pp. 200–202, 206–208; *see* **Eigenvalue**; **Eigenvector**).

Canonical Variate Analysis

Canonical variate analysis provides a perspective on **multivariate analysis of variance**. Let $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, where now columns of $\mathbf{Z}$ specify the groups to which observations belong. Again the orthogonal reduction of $\mathbf{Z}$ to upper triangular form is an effective starting point for further calculations, leading to an eigenproblem of the same form as for canonical correlation [8], pp. 202–203, 208–210.

Matrix Condition Numbers

A matrix *condition number* κ for a matrix $\mathbf{S}$ provides an indication of the relative sensitivity of $\mathbf{Sd}$ or $\mathbf{S}^{-1}\mathbf{d}$ to small relative changes in the elements of $\mathbf{d}$. One possibility is the *spectral condition number* κ_2, which is the ratio $\lambda_{\max}/\lambda_{\min}$ of the largest to smallest eigenvalue of $\mathbf{S}$.

Let $k = \log_{10}\kappa_2(\mathbf{S})$. In general one can expect to lose k digits of accuracy when solving the linear system

$$\mathbf{Sx} = \mathbf{d}$$

for $\mathbf{x}$, or in computing the inverse of $\mathbf{S}$. Note that $\kappa_2(\mathbf{S}) \geq 1$, which means that relative error can never be expected to decrease in solving a linear system. A matrix whose condition number is no more than

10 or 100 is, from a computational perspective, well-conditioned.

Numerical and Statistical Measures of Conditioning

Let κ_2 be the spectral condition number of the correlation matrix derived from $\mathbf{X}'\mathbf{X}$. Then

$$\max_{1\le i\le p}(s_{ii}s^{ii}) \le \kappa_2 \le \sum_{i=1}^{p} s_{ii}s^{ii},$$

where s_{ii} and s^{ii} are defined as in (8). This makes a connection between statistical and numerical measures of conditioning [2, 8], p. 211. The quantity $s_{ii}s^{ii}$ has the benefit that, unlike matrix condition numbers such as κ_2, it is independent of scale.

Note that determination of the minimum value $\min_{\mathbf{b}} ||\mathbf{y} - \mathbf{X}\mathbf{b}||^2$ is well-conditioned, even if $\mathbf{X}$ is singular.

Components of Larger Computations

The notes on computational methods in [5] demonstrate extensive use of matrix calculations as building blocks for a wide variety of other statistical calculations, analysis of variance with multiple error strata (*see* **Multilevel Models**), **generalized linear models** (GLMs), **generalized additive models** (GAMs), local regression smoothing (loess) (*see* **Graphical Displays**), and **nonparametric regression**; see also [11]. New complications are inevitable as matrix computational methods are used to extend the range of models available to statisticians. In models where a variance–covariance structure must be estimated, the notion of a singularity has subtleties beyond those of ordinary least squares.

Software

Many statistical packages allow the user to specify calculations as a sequence of matrix operations. SAS (in the IML Interactive Matrix Language module), SPSS (MATRIX language), STATA, **S-PLUS, R**, and Genstat are some of the statistical systems which have extensive matrix computational abilities. Statistical packages have generally stayed with normal equation methods. S-PLUS and R make extensive use of modern methods such as QR. Note also the extensive modern matrix abilities in the mathematically oriented languages of MATLAB, Gauss, and Mathematica. MATLAB has been used extensively by numerical analysts [7].

The FORTRAN subroutine package LAPACK [1], and earlier packages LINPACK, and EISPACK from which LAPACK is derived, provide high-quality software to perform calculations referred to in this article. These packages are publicly available from the NETLIB online database, and are also part of the NAG and IMSL subroutine libraries.

References

[1] Anderson, E., Bai, Z., Bischof, C., Blaxckfdord, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. & Sorenen, D. (1999). *LAPACK Users' Guide*, 3rd Edn., SIAM, Philadelphia.

[2] Berk, K.N. (1977). Tolerance and condition in regression equations, *Journal of the American Statistical Association* **72**, 863–866.

[3] Björck, A. (1996). *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia.

[4] Bojanczyk, A.W., Brent, R.P., van Dooren, P. & de Hoog, F.R. (1987). A note on downdating the Cholesky factorization, *SIAM Journal of Scientific and Statistical Computation* **8**, 210–221.

[5] Chambers, J.M. & Hastie, T.J. (1991). *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove.

[6] Golub, G.H. & Van Loan, C.F. (1996). *Matrix Computations*, 3rd Ed. Johns Hopkins University Press, Baltimore.

[7] Higham, N.J. (1996). *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia.

[8] Maindonald, J.H. (1984). *Statistical Computation*. Wiley, New York.

[9] Stewart, G.W. (1982). Linear algebra, computational, in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, New York, pp. 5–19.

[10] Stoer, J. & Bulirsch, R. (1992). *Introduction to Numerical Analysis*, 2nd Ed. Springer-Verlag, New York.

[11] Thisted, R.A. (1988). *Elements of Statistical Computation. Numerical Computation*. Chapman & Hall, New York.

JOHN H. MAINDONALD & GORDON K. SMYTH

Maximum Likelihood

The term "maximum likelihood" refers to a general method of **estimation** with important historical and practical significance for biostatistics. Consider a sample $y_1, y_2, \ldots, y_n$ drawn independently from a distribution with density or probability function $f(y;\theta)$ with an unknown vector parameter θ (*see* **Random Variable**). The **likelihood** function is defined as

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta).$$

The maximum likelihood estimate (MLE) is the value $\theta = \hat{\theta}$ which maximizes $L(\theta)$ over the set of all possible values for θ. In practice, it is usually more convenient to maximize the logarithm of the likelihood function, $l(\theta) = \ln L(\theta)$. From calculus, it follows that $\hat{\theta}$ satisfies the score equation $\partial l/\partial\theta = 0$.

Early references to the method of maximum likelihood are attributed to **Gauss**, **Laplace**, and **Edgeworth** [4]. However, the prominent English statistician **R. A. Fisher** is unquestionably responsible for popularizing the technique and for identifying many of its statistical properties [3].

The method has a strong heuristic appeal: choose as your parameter estimate the one which makes the observed data seem most likely. As it turns out, it is often the best estimate possible, particularly in **large samples**. Inference is made easy by the fact that maximum likelihood estimates are **consistent** and asymptotically **normal** under broad regularity assumptions, with a variance that can be estimated from the observed or expected **information matrix**. The MLE is also invariant under one-to-one **transformations** of the parameters, so to obtain the MLE of a transformation of the original parameters, one need only apply the transformation to the original MLE.

Optimal Properties

The following is a list of the most important properties of the MLE in large samples (i.e. $n \to \infty$):

1. *Consistency*. The MLE **converges in probability** to the true value of the parameter.
2. *Asymptotic normality*. The distribution of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to a normal distribution with zero mean and a covariance matrix which is the inverse of the information matrix $\mathbf{I}$. From a practical point of view it is more convenient to say that $\hat{\boldsymbol{\theta}}$ is approximately distributed as $\mathrm{N}(\boldsymbol{\theta}, \mathbf{I}^{-1}/n)$.
3. *Asymptotic efficiency*. The MLE is the best asymptotically normal (**BAN**) estimate in terms of its variance in large samples. Put more precisely, if $\tilde{\boldsymbol{\theta}}$ is another estimate such that $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathscr{L}} \mathrm{N}(\boldsymbol{\theta}, \mathbf{C})$, where $\mathbf{C}$ is a fixed matrix, then $\mathbf{C} \geq \mathbf{I}^{-1}/n$ in the sense that $\mathbf{C} - \mathbf{I}^{-1}/n$ is a positive semidefinite matrix.

Regularity Conditions

It is important to be aware of the general regularity conditions for the optimal properties of the MLE. Most advanced statistics texts have a discussion of these conditions [1, 5], with the classic reference being Cramér [2]. The following three conditions are commonly given:

1. The observed data points $y_1, y_2, \ldots, y_n$ are independently and identically distributed (iid) according to a density or probability function $f(y;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ has finite dimension m. This condition is less restrictive than it may seem, given that y_i may be a vector. In fact, the method of maximum likelihood is popular and appropriate for many regression problems where the distributions of the data points are not identical, and many texts give less restrictive assumptions.
2. The underlying density or probability function is identifiable, i.e. $f(y;\boldsymbol{\theta}_1) = f(y;\boldsymbol{\theta}_2)$ for all y implies that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.
3. The density is "smooth" in the sense that f has derivatives up to the third order with finite expectation, and the information matrix

$$\mathbf{I} = E_\theta\left[\frac{\partial^2 \ln f(y;\theta)}{\partial\theta_j \partial\theta_k}\right], \quad j,k = 1,\ldots,m$$

exists and is nonsingular. The latter conditions ensure that the asymptotic covariance matrix exists.

These conditions are satisfied for many models of interest to biostatisticians, such as the **binomial**, normal, and **Poisson** distributions.

Maximum Likelihood Calculations

Maximum likelihood estimation involves finding a global maximum of a function of one or several parameters. For certain models, the solution can be expressed as a simple function of the data. However, more often the solution must be posed as a nonlinear **optimization** problem. Typically, it involves solving a system of nonlinear equations.

The main methods in use today are the Newton–Raphson (NR) or quasi-Newton (QN) methods, and the Fisher scoring (FS) algorithm [6]. The NR algorithm is based on an approximation of the log likelihood by a quadratic function through a Taylor series expansion of the score functions. To implement the NR algorithm, one has to provide second-order derivatives for the log likelihood function – the so-called Hessian matrix. Initial values for the parameters are updated and the process is repeated until convergence is obtained. If the initial value for parameters is close enough to the maximum, the NR algorithm usually converges quickly. However, if the initial value is poorly chosen it may fail. In particular, the Hessian matrix can become non-positive-definite. Upon convergence, at the final iteration, the inverse of the Hessian matrix provides an approximation to the asymptotic covariance of the MLE. In the QN algorithm only first derivatives are used, and the second derivatives are estimated based on results from previous iterations. The QN algorithm involves line search methods, i.e. maximization of the log likelihood along a given ray in the parameter space.

The difference between the NR and FS algorithms is that the latter uses the expectation of the Hessian matrix rather than the Hessian matrix itself as in the NR algorithm. There are two versions of the FS algorithm. In the first version the expectation of the Hessian is approximated as the sum of cross-products of first derivatives. In the second version, the exact calculated information matrix is used. Thus, to use this version of the FS algorithm one has to have a formula for the information matrix as a function of the parameters calculated prior to the maximization procedure.

Other methods are sometimes used for maximum likelihood estimation. One that deserves special mention is the **EM (expectation–maximization) algorithm**. Certain likelihoods may be thought as involving **missing data**, with the most notable example being **random effects** models. The EM method works in this setting by maximizing the expectation of the log-likelihood iteratively for the complete data. This may aid in difficult maximization problems by taking advantage of simple closed-form solutions available for the "M" stage. Other advantages of the EM algorithm include its natural statistical interpretation and its property of producing an increasing sequence of log likelihood values in the specified parameter space. The principal drawback is that it may be relatively slow.

Examples

Estimation of a Proportion

We observe the occurrence of a certain event for n individuals, where y_i is 1 if the event occurs and 0 otherwise. It can be assumed that events are independent among individuals and have the same probability of occurrence θ. The likelihood can be written as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^m(1-\theta)^{n-m}, \end{aligned}$$

where m is the number of events observed among the n individuals. The log likelihood is

$$l(\theta) = m\ln(\theta) + (n-m)\ln(1-\theta). \qquad (1)$$

To find the maximum, we consider the following score equation:

$$\frac{\mathrm{d}l}{\mathrm{d}\theta} = \frac{m}{\theta} - \frac{n-m}{1-\theta} = 0, \qquad (2)$$

which has the unique solution $\hat{\theta} = m/n$.

Logistic Regression

Logistic regression may be viewed as a continuation of the previous example where the probability of the event occurring depends on some other factor. For instance, y could be an indicator of heart disease and x could denote the weight of an individual. We can model the relationship between y and x as the logistic function of the conditional probability of disease,

$$\Pr(y = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}, \qquad (3)$$

or, equivalently,

$$\ln \frac{\Pr(y = 1|x)}{1 - \Pr(y = 1|x)} = \alpha + \beta x.$$

Here $\theta = (\alpha, \beta)'$. The assumption is that the log **odds ratio** for the occurrence of disease is linear in the covariate, and β is sometimes referred to as the "log odds" parameter.

Now let $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ be values for the disease status and weight of a sample of n individuals. Technically speaking, to write down the likelihood we have to assume that x has a certain distribution which may contain unknown parameters. However, maximum likelihood inference for the logistic regression parameters is not affected by the assumption concerning the distribution of x as long as it does not depend on the parameters α and β. In fact, the values for x_i may be considered as fixed, known constants, and as long as the design matrix (*see* **Experimental Design**) is of full rank, maximum likelihood estimation is valid. The log likelihood is

$$\begin{aligned} l(\alpha, \beta) &= \sum_{i=1}^{n} y_i \ln \Pr(y = 1|x_i) \\ &\quad + \sum_{i=1}^{n} (1 - y_i) \ln[1 - \Pr(y = 1|x_i)] \\ &= \sum_{i=1}^{n} y_i \ln \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \\ &\quad + \sum_{i=1}^{n} (1 - y_i) \ln \frac{1}{1 + \exp(\alpha + \beta x_i)} \\ &= \alpha n + \beta \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \ln[1 + \exp(\alpha + \beta x_i)]. \end{aligned}$$

A typical graph of the likelihood function is shown in Figure 1. The maximum of the log likelihood is found as the solution to the score equations

$$\frac{\partial l}{\partial \alpha} = n - \sum_{i=1}^{n} \frac{1}{1 + \exp(\alpha + \beta x_i)} = 0$$

and

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} x_i = 0.$$

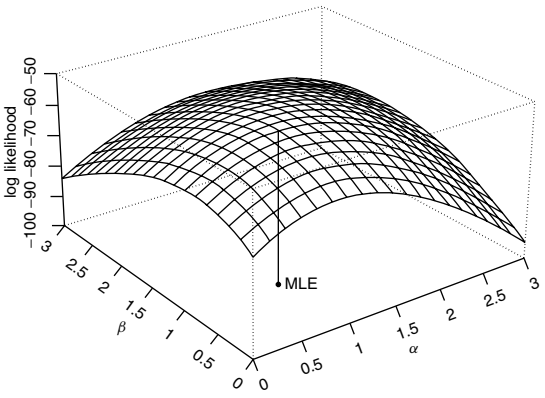

Figure 1 An example of the log likelihood surface and MLE in logistic regression

This system of nonlinear equations must be solved iteratively, e.g. by the Newton–Raphson algorithm.

Discussion and Extensions

In theory, the method of maximum likelihood is very simple: determine an appropriate sampling distribution for the data, write down the likelihood as a function of the unknown parameters, and solve for the estimate. Of course, in practice it is not always so easy. Solutions to the likelihood score equations usually must be arrived at by numerical methods, and may be computationally intensive or inaccurate. For some models and data sets it may be difficult to demonstrate that the likelihood has a unique global maximum. Many small-sample MLEs can be shown to be **biased**. The presence of **nuisance parameters** can exacerbate the computational difficulties. Often the MLE is quite non **robust** to **outliers** and, unlike competing general methods such as **least squares** and the **method of moments**, computation of the MLE requires that the distribution of the data be completely parameterized. Uniformly **minimum variance unbiased** (UMVU) **estimation** theory and **Bayesian methods** compete with maximum likelihood estimation with regard to optimal properties, but also require complete specification of the distribution, and may be even harder to implement.

To overcome some of the difficulties of maximum likelihood estimation, modified methods have been proposed such as **restricted maximum likelihood**, conditional likelihood (*see* **Conditionality Principle**), **pseudo-likelihood, quasi-likelihood**, **partial**

likelihood, and M-estimation (*see* **Robustness**). These methods were all inspired by the powerful heuristic appeal and conceptual simplicity of the original formulation of the method of maximum likelihood.

References

[1] Cox, D.R. & Hinckley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[2] Cramér, H. (1945). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[3] Rao, C.R. (1992). R.A. Fisher: The founder of modern statistics, *Statistical Science* **7**, 34–48.

[4] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, Mass.

[5] Stuart, H. & Ord, J.K. (1991). *Kendall's Advanced Theory of Statistics*, Vol. 2. Oxford University Press, New York.

[6] Thisted, R.A. (1988). *Elements of Statistical Computing*. Chapman & Hall, New York.

TOR D. TOSTESON & EUGENE DEMIDENKO

Maxwell, Albert Ernest

Born: July 7, 1916, in Rockmount, Co. Cavan, Ireland.
Died: 1996, in Leeds, UK.

Albert Ernest Maxwell was educated at the Royal School, Cavan and at Trinity College, Dublin, where he developed interests in psychology and mathematics. After graduating, 'Max' as he was invariably known to his colleagues, became a mathematics teacher at St Patrick's Cathedral School, Dublin. After only three years, at the age of 25, he was appointed Headmaster. His attempt to understand the behavioral problems of some of his pupils renewed his interest in psychology, a subject he eventually pursued more seriously at the University of Edinburgh where he was awarded a doctorate in 1950.

In 1952 Max left schoolteaching to take up the post of lecturer in statistics at the Institute of Psychiatry, a postgraduate school of the University of London. He was to spend the rest of his working life at the Institute, retiring in 1978 as Professor of Psychological Statistics.

For a number of years Max was a member of the Psychology Department of the Institute, collaborating and advising Professor Hans Eysenck, but he was eventually rewarded with the Headship of his own Biometrics Unit, which had responsibility for helping Institute researchers on all aspects of statistical design and analysis.

Max's main area of expertise was in **multivariate analysis**, particularly **factor analysis**, where his collaboration with Dr D. Lawley resulted in an important account of the mathematical theory behind the technique [1].

Max's teaching skills (no doubt learnt whilst a schoolmaster in Co. Cavan) were legendry and many psychologists obtained a firm grasp of statistical methods from his numerous lecture courses and a number of useful textbooks including [2].

References

[1] Lawley, D.N. & Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method*, 2nd Ed. Butterworths, London.

[2] Maxwell, A.E. (1977). *Multivariate Analysis in Behavioural Research*. Chapman & Hall, London.

BRIAN S. EVERITT

McKendrick, Anderson Gray

Born: 1876
Died: 1943

McKendrick made important contributions to the mathematical theory of epidemics. He served as a lieutenant-colonel in the Indian Medical Service, and later became Curator of the College of Physicians at Edinburgh. In 1914 [2], he gave the solution of the general homogeneous birth process (*see* **Stochastic Processes**), and this was followed in 1926 by a major paper [3] on stochastic epidemics (*see* **Epidemic Models, Stochastic**). He then turned to deterministic models (*see* **Epidemic Models, Deterministic**), in a series of papers with W.O. Kermack, which included the celebrated Threshold Theorem (*see* **Epidemic Thresholds**). His work is described in some detail in [1].

References

[1] Irwin, J.O. (1963). The place of mathematics in medical and biological statistics, *Journal of the Royal Statistical Society, Series A* **126**, 1–45.

[2] McKendrick, A.G. (1914). Studies on the theory of continuous probabilities with special reference to its bearing on natural phenomena of a progressive nature, *Proceedings of the London Mathematical Society* **13**, 401–416.

[3] McKendrick, A.G. (1926). Applications of mathematics to medical problems, *Proceedings of the Edinburgh Mathematical Society* **44**, 98–130.

McNemar Test

The McNemar test arose in the context of psychology in which two correlated dichotomous responses were to be compared. One example of this test would be an indication of a response or no response under two experimental conditions. The original paper was by McNemar [7]; see also [8].

Armitage and Berry [1] give an example in which we have two culture media and wish to determine if they are equally effective in detecting tubercle bacilli in sputum specimens. The two media, evaluated for the same sample of 50 specimens, give cell counts of positive and negative results, as given in Table 1.

This can be regarded as a single multinomial table in which the cell probabilities, denoted by $\{\pi_{ij}\}$, add to 1. If the two media have the same ability to detect the bacilli, the null hypothesis takes the form $\pi_{i+} = \pi_{+i}$, i.e. that the marginal proportions are the same. This is easily seen to be equivalent to $\pi_{12} = \pi_{21}$. Since only these probabilities are of interest, the hypothesis can be tested referring only to the counts in these off-diagonal cells. Let the cell counts be denoted by $\{n_{ij}\}$. The (uncorrected) test statistic is given by

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}},$$

which (asymptotically) has a **chi-square distribution** with 1 **degree of freedom** (df). In this case $X^2 = (12 - 2)^2/(12 + 2) = 7.14$. The test is equivalent to the binomial test that the proportion is 0.5, given that the observation lies in one of the off-diagonal cells. It has also been suggested that a continuity corrected test be used. This is equivalent to the continuity correction for the binomial test. The corrected form is $X^2 = (|12 - 2| - 1)^2/(12 + 2) = 5.79$. In this case, both statistics are significant at the $\alpha = 0.05$ level.

Table 1

		Medium B		
		+	−	Total
Medium A	+	20	12	32
	−	2	16	18
Total		22	28	50

Examples of this procedure also arise in diagnostic imaging studies when the investigator wishes to determine if a new method provides better diagnostic **sensitivity** and **specificity** than a standard method. In this situation, there are several options: one may compare only the sensitivity (diagnostic performance when the patients have the condition), or the specificity (diagnostic performance when the patients do not have the condition); it is sometimes suggested that the accuracy (proportion correct) be compared, but this comparison is quite sensitive to the proportion of positive and negative subjects, and is not recommended. For the comparison of sensitivity or specificity, a McNemar test can be formed, and the 1 df test can be made. This provides two tests, which may be contradictory in the sense that one test may have higher sensitivity while the other has higher specificity. The two tests may be combined by adding the χ^2 to get a 2 df test. This gives an overall test of common performance [3].

In diagnostic test comparisons, it is important to note that the investigators may be artificially imposing a two-category result on the data. For example, in diagnosing recurrent cancer, the outcomes might be "negative", "resectable", or "not resectable", and the implications of imposing two categories are unclear. In such a case, it seems relevant to examine the 3×3 matrix of categorizations, and use a procedure which formally tests for marginal symmetry. In addition, some misclassifications are more serious than others. This would imply that a weighted procedure might be used. Tests for this are referenced in [6].

The McNemar test can be generalized to equivalence testing [5]. Conditional logistic regression (*see* **Logistic Regression, Conditional**) can be used to adjust for covariates.

Sample size computations are based on comparing a binomial proportion to 0.5. These calculations give the number of discordant pairs (those not represented on the main diagonal). It is then required to determine what fraction of discordant pairs is likely to arise. Various methods have been proposed recently for this (e.g. [4] and [2]).

References

[1] Armitage, P. & Berry, G. (1994). *Statistical Methods in Medical Research*, 3rd Ed. Blackwell Science, Oxford pp. 127–128.

[2] Connor, R.J. (1987). Sample size for testing differences in proportions for the paired sample design, *Biometrics* **47**, 207–211.

[3] Hamdan, M.A., Pirie, W.R. & Arnold, J.C. (1975). Simultaneous testing of McNemar's problem for several populations, *Psychometrika* **40**, 153–162.

[4] Lachenbruch, P.A. (1992). On the sample size for studies based on McNemar's test, *Statistics in Medicine* **11**, 1521–1527.

[5] Lee, M.L. & Lusher, J.M. (1991). The problem of therapeutic equivalence with paired qualitative data: an example from a clinical trial using haemophiliacs with inhibitors to Factor VIII, *Statistics in Medicine* **10**, 443–451.

[6] Mantel, N. & Fleiss, J.L. (1975). The equivalence of the generalized McNemar's tests for marginal homogeneity in 2^3 and 3^2 tables, *Biometrics* **31**, 727–729.

[7] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–157.

[8] Somes, G. (1985). *McNemar test, Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz & N. Johnson, eds. Wiley, New York, pp. 361–363.

(*See also* **Correlated Binary Data**; **Matched Pairs With Categorical Data**; **Rasch Models**; **Square Contingency Table**)

PETER A. LACHENBRUCH

Mean Deviation

The mean deviation d of a data set $x_1, \ldots, x_n$ is defined as

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - m|, \tag{1}$$

where m is a location measure, most often the arithmetic **mean**, $\overline{x}$, most naturally the sample **median**, med (which minimizes d over all values of m), or, if known, a population location measure, such as the **expectation** or the median of the population.

It is also called mean absolute deviation (or average deviation), and the mean deviation from the mean is also called mean absolute error (MAE).

Its probabilistic counterpart for a random variable X is the first absolute **moment** with respect to (usually) the expectation (if it exists),

$$\delta = \mathrm{E}(|X - \mathrm{E}X|). \tag{2}$$

The mean deviation is thus a measure of scale or dispersion, like the **standard deviation** (sd) or the **range**, or the median (absolute) deviation (MAD) $= \mathrm{med}_i(|x_i - \mathrm{med}_j x_j|)$ (the median, not mean, of the absolute differences from the sample median). In general, they all estimate different quantities; but, for example, for large samples from the normal distribution, one can give conversion factors:

$$d = \left(\frac{2}{\pi}\right)^{1/2} \mathrm{sd} \approx 0.7979 \ \mathrm{sd}$$

and

$$\mathrm{MAD} \approx 0.6745 \ \mathrm{sd} \approx 0.8453 \ d. \tag{3}$$

The mean deviation is the **maximum likelihood estimator** of the scale parameter δ (and is therefore asymptotically **efficient**) for the double-exponential distribution with density $f_{\mu,\delta}(x) = (2\delta)^{-1} \exp(-|x - \mu|/\delta)$ (μ arbitrary, $\delta > 0$). (The corresponding estimator for μ is the median.)

Around 1900, mean deviation and standard deviation were the two most commonly used scale estimators. Both were usually converted into the *probable error* (the error that would be surpassed with probability one-half) according to the above Eqs (3), with MAD replaced by probable error. The MAD, although the direct (and **nonparametric**) estimate of the probable error, apparently was never used then as an estimator (perhaps because of its low efficiency near the normal distribution, and without awareness of its good **robustness** properties).

In 1920, Fisher [2] proved and stressed the optimality of the standard deviation under strictly normal data and showed that the mean deviation in this case has an **asymptotic relative efficiency** of only $1/(\pi - 2) \approx 88\%$, causing the mean deviation to fall into oblivion. The astronomer Eddington [1, p. 147; 2, footnote, p. 762] maintained that the mean deviation was the better (more accurate) scale measure according to (astronomical) practical experience. There is no contradiction – both were right; real data are never exactly normal. In 1960, Tukey ([[6], inserts], with more details and a correction given by Huber [5, p. 3]) showed that less than 0.2% of a very mild form of contamination of a normal distribution suffice to render d better than sd, while for about 5% contamination (a common frequency of gross errors) of the same mild type, d is twice as efficient as sd.

A key to better understanding of these and other facts about the mean deviation is provided by the concepts of robustness theory (the stability theory of statistical procedures). If we add a single observation (e.g. a gross error) in any point x to a large sample with (estimated) parameters m and d, the standardized change of d in the limit of $n \to \infty$ is given by the influence curve or influence function [3]

$$\mathrm{IF}(x) = |x - m| - d, \tag{4}$$

which increases only linearly with x, while the IF for sd increases quadratically, implying a much higher sensitivity to "dirt" in the tails of the observed distribution (*see* **Robustness**). However, both functions are unbounded; in fact, a single **outlier** moving to infinity carries both estimates to infinity. Hence neither should be used. We say, that their *breakdown point* [3] is zero. By contrast, the MAD tolerates about 50% gross errors before it gives arbitrarily false values; its breakdown point is 50%. There are other scale estimators with positive breakdown point and generally higher efficiency than the MAD (e.g. **trimmed** variances, or Huber's scale estimators – cf. [5]). Using sd or d after (some functioning form of) rejection of outliers prevents the worst, but this approach is a complex procedure usually not well understood, and is less efficient than other robust scale estimators (cf. [4]); nevertheless, it might often be the simplest practical solution.

References

[1] Eddington, A.S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.

[2] Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error, *Monthly Notices of the Royal Astronomical Society* **80**, 758–770.

[3] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**, 383–393.

[4] Hampel, F.R. (1985). The breakdown points of the mean combined with some rejection rules, *Technometrics* **27**, 95–107.

[5] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[6] Tukey, J.W. (1960). A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics*, I. Olkin et al., eds. Stanford University Press, Stanford.

FRANK HAMPEL

Mean Square Error

If $\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_n)$ is an estimator of a parameter θ based on a **random sample** of size n, then the mean square(d) error (MSE) of the estimator is defined as the expected value of the squared deviation of the estimator from the true value to be estimated:

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}[(\hat{\theta} - \theta)^2] \\ &= \int \cdots \int [\hat{\theta}(Y_1, \ldots, Y_n) - \theta]^2 \\ &\quad \times f(y_1; \theta) \ldots f(y_n; \theta) \partial y_1 \ldots \partial y_n,\end{aligned}$$

where $f(y;\theta)$ is the density upon which the sample is based. In general, for any estimator $\hat{\theta}$ of a parameter θ, the MSE can equivalently be defined as $\mathrm{E}[(\hat{\theta}) - \theta)^2] = \int (\hat{\theta} - \theta)^2 g(\hat{\theta}; \theta) \partial \hat{\theta}$, where $g(\hat{\theta};\theta)$ is the **sampling distribution** of the estimator $\hat{\theta}$. The MSE is a measure of the closeness of the estimator to the true value. From the following identity:

$$\begin{aligned}\mathrm{E}[(\hat{\theta} - \theta)^2] &= [\mathrm{E}(\hat{\theta}) - \theta]^2 + \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]^2 \\ &= [\mathrm{bias}(\hat{\theta})]^2 + \mathrm{var}(\hat{\theta}),\end{aligned}$$

it is seen that the MSE is the sum of the squared **bias** plus the **variance** of the estimator. Thus, the MSE reflects both the bias of an estimator, i.e. how much its expected value differs systematically from the true value, as well as the precision (variance) of the estimator, which measures how much it varies about its expected value or mean due to sampling variability. A good estimator ideally will have a small MSE, reflecting both small bias and small variance. Choosing an estimator with a small MSE often entails a tradeoff between bias and variance.

Subset Selection in Regression and Prediction

Tradeoffs between bias and variance are illustrated in the problem of choosing the "best" set of predictor variables in multiple linear regression (*see* **Variable Selection**). Let y_i be a response measured on the ith individual in a sample of size n. The objective is to relate y_i to a set of p predictors (**explanatory** or independent variables) and to predict future values of y. The standard model writes $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$, where $x_{i1}, \ldots, x_{ip}$ are p predictors measured on subject i, $\beta_0, \beta_1, \ldots, \beta_p$ are unknown regression coefficients to be estimated, and $e_i, i = 1, \ldots, n$, are independent, identically distributed, residual errors having mean zero and variance σ^2. This model is expressed in matrix notation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{Y}$ is the $n \times 1$ vector of responses, $\mathbf{X}$ is the $n \times (p+1)$ design matrix of rank $(p+1)$ whose ith row is $(1, x_{i1}, \ldots, x_{ip})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, and $\mathbf{e}$ is the $n \times 1$ vector of errors. The **least squares** estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and for a given set of predictors $\mathbf{x} = (1, x_1, \ldots, x_p)'$, the usual predictor of y is $\hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$. This is an **unbiased** predictor, and the *mean squared error of prediction* (MSEP), is defined as $\mathrm{E}(\hat{y} - y)^2 = \sigma^2[1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}]$. The *mean squared error of the regression coefficients* is defined as $\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'$, which equals $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. At times, an objective is to select the "best" subset for predicting y out of a potentially large number p of available predictor variables. Walls & Weeks [5] show that it is possible for the prediction based on a subset of variables to have a smaller MSEP. Partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is the set of variables included in the regression and $\mathbf{X}_2$ are excluded. Similarly, partition $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ and $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$. The least squares estimate of $\boldsymbol{\beta}_1$ based on the subset $\mathbf{X}_1$ is $\tilde{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, with bias $\mathrm{E}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$ and $\mathrm{MSE}(\tilde{\boldsymbol{\beta}}_1) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2\boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}$. The predicted value of y based on $\mathbf{X}_1$, $\tilde{y} = \mathbf{x}_1'\tilde{\boldsymbol{\beta}}_1$, will generally be biased (the bias is nonzero unless $\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{0}$), and $\mathrm{MSEP}(\tilde{y}) = \sigma^2[1 + \mathbf{x}_1'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{x}_1] + \{[\mathbf{x}_2' - \mathbf{x}_1'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2]\boldsymbol{\beta}_2\}^2$, which may be less than $\mathrm{MSEP}(\hat{y})$ based on the full model. In particular, Hocking [2] shows that if $\mathrm{var}(\hat{\boldsymbol{\beta}}_2) - \boldsymbol{\beta}_2\boldsymbol{\beta}_2'$ is positive definite, then (i) $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_1) - \mathrm{MSE}(\tilde{\boldsymbol{\beta}}_1)$ is positive definite, and (ii) $\mathrm{MSEP}(\hat{y}) \geq \mathrm{MSEP}(\tilde{y})$, implying that the reduced model using the subset of variables in $\mathbf{X}_1$ is better in terms of mean squared error both for estimating the regression coefficients $\boldsymbol{\beta}_1$ as well as for predicting y. For example, excluding a single variable X_2 will result in a better estimate of $\boldsymbol{\beta}_1$ and a lower MSEP if $\mathrm{var}(\hat{\boldsymbol{\beta}}_2) > (\boldsymbol{\beta}_2)^2$, i.e. if the variance of the regression coefficient of $\boldsymbol{\beta}_2$ estimated in the full model exceeds the square of the true value of $\boldsymbol{\beta}_2$. Hocking [2] also proposes considering as a criterion the average decrease in predictive mean squared error

over the n points in the sample:

$$\frac{1}{n}\sum_{i=1}^{n}[\text{MSEP}(\hat{y}_i) - \text{MSEP}(\tilde{y}_i)]$$

and discusses how selecting a subset to maximize this criterion is closely related to the use of **Mallows' C_p** and the adjusted R^2 as criteria for subset selection in multiple regression.

The concept that a biased estimator may be preferable in terms of MSE to an unbiased estimator with a large variance also underlies the concept of **ridge regression** [3]. In situations where the p predictor variables are highly intercorrelated (i.e. the problem of **multicollinearity**), they suggest the biased "ridge regression" estimate $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$, where the predictor variables in X have been standardized by subtracting their sample means and dividing by their standard deviations, and the constant k is determined by inspecting the "ridge trace" or plot of $\hat{\boldsymbol{\beta}}_k$ vs. k. For some choice of k, the ridge regression estimate will have smaller MSE than the least squares estimator. **Shrinkage estimators** derived from an **empirical Bayes** approach [1] also typically have a smaller MSE than the usual unbiased estimators. *See also* **James–Stein Estimator** [4].

References

[1] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, New York.

[2] Hocking, R.R. (1974). Misspecification in regression, *American Statistician* **28**, 39–40.

[3] Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: biased estimation for non-orthogonal problems, *Technometrics* **12**, 55–67.

[4] James, W. & Stein, C. (1961). Estimation with quadratic loss, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 361–379.

[5] Walls, R.C. & Weeks, D.L. (1969). A note on the variance of a predicted response in regression, *American Statistician* **23**, 24–26.

MARK D. SCHLUCHTER

Mean

The mean is a central concept in both data analysis and statistical theory. The usual sample mean is an arithmetic average of a set of n numerical observations, $x_1, x_2, x_3, \ldots, x_n$,

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The sample mean is nearly universally denoted by the symbol, $\overline{x}$, and is the most commonly used measure of central tendency of a set of numerical data.

In probability, the population mean, expected value or **expectation** of a **random variable** is the analog of the arithmetic or sample mean, $\overline{x}$. In fact, the **law of large numbers** implies that, for an infinitely large sample of independent random variables drawn from a distribution, the sample mean, $\overline{x}$, is equal to the expected value or mean of the distribution.

In any data analysis the usefulness of a summary statistic, such as the mean, depends on the details of the physical or biological process measured and specific summary information needed from the data. The sample mean can often be modified to provide the appropriate summary information. These modifications and their probability model analogs provide a rich source of both theory and numerical description of data sets.

Often each numerical observation does not have equal weight or importance. For example, the numerical observations themselves may be means of subgroups with different numbers of observations in each subgroup. In this case a *weighted mean* or average is used:

$$\overline{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

where w_i is the weight of the ith numerical observation. In the example where each observation represents a subgroup, the weight should be the number of units in the subgroup. This kind of weighted average is found in **analysis of variance,** sample survey analysis, and other more specialized areas.

Weighting can sometimes correct for sampling bias. A special case is when the probability that a unit is sampled is proportional to the variable of interest – an example of size-biased sampling. In this, weighting each observation by $w_i = 1/x_i$ yields

$$\overline{x}_h = \frac{\sum_{i=1}^{n} \frac{1}{x_i} x_i}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}},$$

which is called the *harmonic mean*. The harmonic mean is found by taking the average value of the reciprocal of the data, $1/x_i$, and then taking the reciprocal of the average value. The harmonic mean is well defined only for positive data, i.e. when all possible values of the data are greater than 0. For positive data the harmonic mean is always less than or equal to the arithmetic mean.

The *geometric mean* can be used when measurements from natural processes have a distribution that is not symmetric and is **skewed** to the right. For skewed data the sample mean is not an adequate description of the center of the data. Both the theory of the **lognormal** distribution and practical data analysis justify the use of a log transformation of the original data,

$$y_i = \ln x_i.$$

Often the log transformed data will have a nearly symmetric distribution. In this case

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

is a good measure of the center of the log transformed data. Transforming $\overline{y}$ back to the original scale of the data yields the geometric mean of x,

$$\mathrm{GM} = \exp \overline{y}.$$

Thus, the geometric mean, GM, is the antilog of the mean of the log transformed data. A mathematically equivalent way of expressing the geometric mean is as the nth root of the product of the n observations,

$$\mathrm{GM} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}.$$

The geometric mean is only well defined for positive data. For positive data the following inequality

holds for geometric, harmonic, and sample (arithmetic) means:

$$\overline{x}_h \leq \mathrm{GM} \leq \overline{x}.$$

One straightforward way to obtain a measure of central tendency that does not depend on the extremes or tails of the data is to first remove or trim the data in both tails of the distribution and then compute the usual arithmetic mean using the remaining data. The result is referred to as a **trimmed** or Winsorized mean.

W. SMITH

Measurement Error in Epidemiologic Studies

This article is concerned with relating a response or outcome to an exposure and **confounders** in the presence of measurement error in one or more of the variables. We focus almost entirely on measurement error in a continuous or measured variable. When categorical variables (exposed or not exposed, case or control, quintiles of fat) are measured with error, they are said to be misclassified (*see* **Misclassification Error**). There are also many links in this topic with methods for handling missing data and with validation studies (*see* **Missing Data in Epidemiologic Studies**; **Validation Study**). For further details and a general overview of the topic, see [20]; [30] should be consulted for the linear model.

Before describing the problem, it is useful first to consider a number of specific examples that have had an impact on the development of the field:

1. Measurement error has long been a concern in relating error-prone predictors such as systolic blood pressure (SBP) to the development of coronary heart disease (CHD). That SBP is measured with error is well known, and estimates [23] suggest that approximately one-third of its observed variability is due to measurement error. The **Framingham Heart Study** is perhaps the best known **cohort study** in which the role of measurement error in SBP has been a concern for many years. MacMahon et al. [44] describe the important public health implications of properly accounting for the measurement error inherent in SBP. In an (as yet) unpublished paper, David Yanez, Richard Kronmal & Lynn Shemanski have discovered an example also in the CHD context where the failure to account properly for measurement error leads to misleading conclusions based on falsely detected statistical significance.
2. In measuring nutrient intake, measurement error has been a long-term concern, as has the impact of this error on the ability to detect nutritional factors leading to cancer, especially breast and colon cancer. Typical cohort studies measure diet by means of food frequency questionnaires which, while related to long-term diet, are known to have biases and measurement errors. Other instruments are in use in this field, including food records (essentially diaries), 24 h recalls and (for a limited number of variables such as total caloric intake) biomarkers. Measurement error in nutrient instruments can be very large, for example because of the daily and seasonal variability of an individual's diet, and the **biases** in and loss of **power** to detect nutrient–cancer relationships can be profound. There is still considerable controversy in this field (see [37], [54], and [41]). Because of the cost of cohort studies in nutrition, **case–control studies** are of considerable interest. However, nutrient intakes in case–control studies are measured after the development of disease in cases, and this might cause differential measurement error, a topic we discuss in some detail below (*see* **Nutritional Exposure Measures**).
3. There are a number of ongoing prospective and case–control studies of disease and serum hormone levels, and this is an area of considerable potential. Measurement error is a major concern here, due to within-individual variation of hormones, as well as various laboratory errors.
4. In measuring environmental risk factors (*see* **Environmental Epidemiology**), measurement error is a common problem. For example, measuring household lead levels is an error-prone process, not only because of laboratory and device error, but also because lead levels are inhomogeneous in both space and time, while measurement methods tend to be in fixed locations at fixed times. Because lead exposure has many possible media (air, dust, soil) with possibly correlated errors, the effects of measurement error can be large and complex.

Outline

This article consists of a series of major Sections, as follows:

1. We first outline the basic concepts of measurement error modeling, making particular distinction between **differential** and **nondifferential measurement error**. We also describe the ideas of functional and structural modeling, as well as indicating how the measurement error problem can be treated as a **missing data** problem.
2. Following the introductory concepts, we discuss the problem of measurement error as it pertains to

the **linear regression** model. Here we introduce the idea of attenuation of regression coefficients, and the biases in parameter estimates caused by measurement error. We also discuss **hypothesis testing**. In the simplest cases, measurement error causes an often large decrease in the power to detect significant effects, while, as indicated above, and as exhibited through the **analysis of covariance**, in an **observational study**, measurement error in a confounder can cause misleading **inferences** about exposure effects.

3. Having described the effects of measurement error on **estimation** and hypothesis testing, we turn to correcting for the effects due to measurement error. We first describe the two most common methods, known as regression calibration and SIMEX, and also a group of techniques called corrected score methods. We also describe the use of instrumental variables.
4. **Maximum likelihood** and **Bayesian** estimation form an important component of the measurement error problem, and are described in some detail. We define the **likelihood** function, and show the crucial difference in the likelihood function between the nondifferential and differential measurement error cases; see (13) and (14).
5. While most of the article is based on measurement error in predictors, there is an important literature on response error, which we also review.
6. Case–control studies are important in epidemiology. A distinguishing feature of case–control studies is that the measurement error may be differential. In the differential measurement error case, we indicate that a specific type of data is required, the validation data sets, in which the true predictor can be observed for a subset of the study participants. If the measurement error is nondifferential, then matters are much easier, and the famous result of Prentice & Pyke [55] on the analysis of case–control studies is shown to have an analogue in the measurement error context.
7. There is a significant and developing literature on measurement error in **survival analysis**, and we indicate two possible approaches to the problem.

Measurement error models have a common structure; we illustrate the terms using a breast cancer and nutrition example:

1. An underlying model for a response in terms of predictors, e.g. linear **regression**, **logistic regression**, **nonlinear regression**; see Carrell & Ruppert [14]. This is the model we would fit if all variables were observed without error. In what follows, we call Y the response. For example, in the breast cancer and nutrition example, Y is breast cancer incidence fit to covariables using logistic regression
2. A variable which is measured subject to error. This could be an exposure or a confounder. We call this variable X. It is often called the *error-prone* predictor or the *latent predictor*. In the breast cancer example, X is long-term nutrient intake
3. The observed value of the mismeasured variable. We call this W, e.g. nutrient intake measured from a food frequency questionnaire
4. Those predictors which for all practical purposes are measured without error, which we call Z, e.g. age, body mass index
5. We are interested in relating the response Y to the true predictors (Z, X). One method, often called the *naive* method, simply replaces the error-prone predictor X with its measured version W. This substitution typically leads to biases in parameter estimates and can lead to misleading inferences
6. The goal of measurement error modeling is to obtain nearly **unbiased** estimates of exposure effects and valid inferences. Attainment of this goal requires careful analysis. Substituting W for X, but making no adjustments in the usual fitting methods for this substitution, leads to estimates that are biased, sometimes seriously. In assessing measurement error, careful attention needs to be given to the type and nature of the error, and the sources of data which allow modeling of this error.

It should be obvious that one should design studies and instruments in such a way as best to lessen or to eliminate measurement error. In this article, we demonstrate some of the impacts of ignoring measurement error, ranging from bias in parameter estimates (Figure 1), to loss of power, requiring therefore much larger sample sizes to detect effects (Figure 2) to cases where the type I errors (*see* **Hypothesis Testing**) occur at higher rates than the usual 5% (Figure 3).

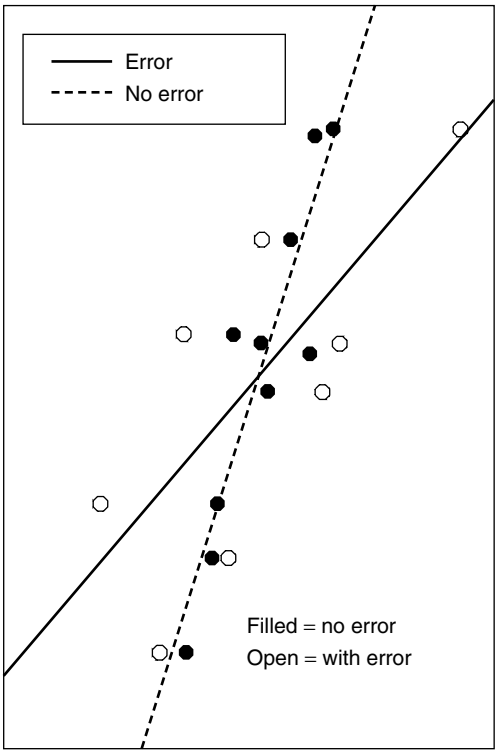

Figure 1 Illustration of additive measurement error model. The filled circles are the true (Y, X) data and the dashed (steeper) line is the least squares fit to these data. The open circles and solid (attenuated) line are the observed (Y, W) data and the associated least squares regression line. For these data $\sigma_x^2 = \delta_u^2 = 1$, $(\beta_0, \beta_x) = (0, 1)$ and $\sigma_\varepsilon^2 = 0.25$

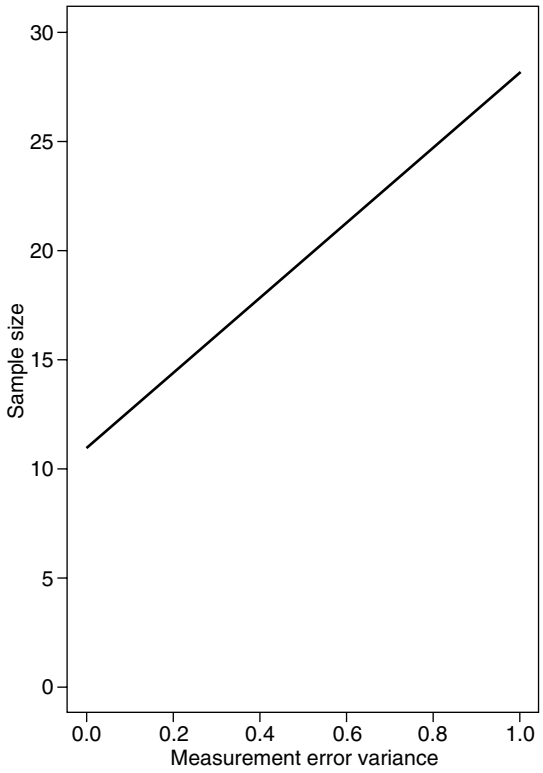

Figure 2 Sample size for 80% power in a one-sided test of level 5% in linear regression, as a function of the measurement error variance. Here the true slope = 0.75, the true variance of X is 1.0, and the true variance about the line is 1.0

Computer Programs

S-PLUS and SAS (*see* **Software, Biostatistical**) computer programs (on Solaris SPARC architecture and for Windows on PCs) which implement many of the methods described in this article (for major **generalized linear models** such as linear, logistic, probit, Poisson and gamma regression) are available at no cost on the World Wide Web at `http://stat.tamu.edu/qvf/qvf./html`.

Bootstrap standard errors are available. They have been developed by Raymond Carroll, Henrik Schmiedieche, and H. Joseph Newton.

A set of programs for **logistic regression** (in SAS and FORTRAN) is available from Professor Donna Spiegelman (e-mail `stdls@gauss.bwh.harvard.edu`). Interested readers should contact her for information concerning extension of these programs to **proportional hazards** and linear regression.

Iowa State University (Department of Statistics, Iowa State University, Ames IA 50011) distributes programs called EV-CARP for linear measurement error models at a cost of $300.

Models for Measurement Error

A fundamental prerequisite for analyzing a measurement error problem is specification of a model for the measurement error process. The *classical error model*, in its simplest form, is appropriate when an attempt is made to determine X directly, but one is unable to do so because of various errors in measurement. For example, consider systolic blood pressure (SBP), which is known to have strong daily

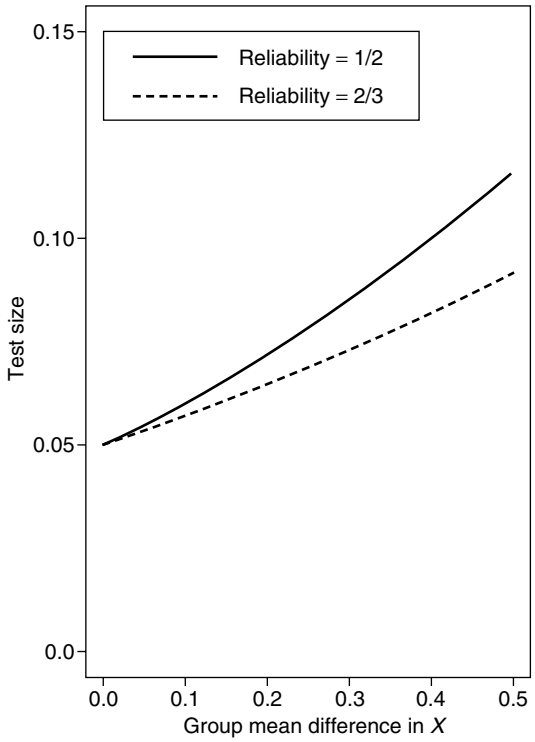

Figure 3 The actual level of a test for exposure effect with a highly predictive covariate measured with error, based on a sample of size $n = 20$. Here the true slope for the covariate $X = 1.0$, the true variance of X is 1.0, the true variance about the line is 1.0, and the reliability is either 2/3 (dashed line) or 1/2 (solid line). The term "Group mean difference in X" is the difference in the mean of X in the exposure group minus the mean of X in the control group

and seasonal variations. In trying to measure SBP, the various sources of error include simple machine recording error, administration error, time of day, and season of the year. In such a circumstance, it sometimes makes sense to hypothesize an unbiased **additive error model**, which we write as

$$\text{(the classic model) } W = X + U, \qquad (1)$$

where U, the error, is assumed to be independent of X. An alternative model, the *controlled variable or Berkson model* [6], is especially applicable to laboratory studies. As an example, consider the herbicide study of Rudemo et al. [61]. In that study, a nominal measured amount W of herbicide was applied to a plant. However, the actual amount X absorbed by the plant differed from W, e.g. because of potential errors in application. In this case,

$$\text{(the Berkson model) } X = W + U, \qquad (2)$$

where U, the error, is assumed to be independent of W.

Determining an appropriate error model to use in the data analysis depends upon the circumstances and the available data. For example, in the herbicide study, the measured concentration W is fixed by design and the true concentration X varies due to error, so that model (2) is appropriate. On the other hand, in the measurement of long-term systolic blood pressure, it is the true long-term blood pressure which is fixed for an individual, and the measured value which is perturbed by error, so model (1) should be used. Estimation and inference procedures have been developed both for error and controlled-variable models.

This hardly exhausts the possible error models. See [20] and [29] for more details and further examples with more complex structure.

Sources of Data

To perform a measurement error analysis, one needs information about the error structure. These data sources can be broken up into two main categories:

1. *internal* subsets of the primary data
2. *external* or independent studies.

Within each of these broad categories, there are three types of data, all of which might be available only in a random subsample of the data set in question:

1. *validation* data, in which X is observable directly
2. *replication* data, in which replicates of W are available
3. *instrumental* data, in which another variable T is observable in addition to W.

An internal validation data set is the ideal, because it can be used with all known analytical techniques, permits direct examination of the error structure and tests of critical error model assumptions, typically leads to much greater precision of estimation and inference, and has strong links to the well-developed

theory of missing data analysis (see below). We cannot express too forcefully that, if at all possible, one should obtain an internal validation data set.

With external validation data, one must assume that the error structure in those data also applies to the primary data (see below).

Replication data are used when it is impossible to measure X exactly, as, for example, when X represents long-term systolic average blood pressure or long-term average nutrient intake. Usually, one would make replicate measurements if there were good reason to believe that the replicated mean is a better estimate of X than a single observation, i.e. the classical error model is the target. In the classical error model (1), replication data can be used to estimate the **variance** of the measurement error, U.

Internal instrumental data sets containing a second measure T are useful for **instrumental variable** analysis, discussed briefly later in this article.

Transportability of Models and Parameters

In some studies, the measurement error process is not assessed directly, but instead is estimated from external data sets. We say that parameters of a model can be transported from one study to another if the model holds with the same parameter values in both studies. Typically, in applications only a subset of the model parameters need be transportable.

In many instances, approximately the same classical error model holds across different populations. For example, consider systolic blood pressure at two different clinical centers. Assuming similar levels of training for technicians making the measurements and a similar measurement protocol, it is reasonable to expect that the distribution of the error in the recorded measure is independent of the clinical center one enters, the technician making the measurement, and the value of X being measured. Thus, in classical error models it is often reasonable to assume that the error distribution is the same across different populations, i.e. transportable.

A common mistake is to transport a correction for measurement error from one study to the next. Such transportation is almost never appropriate. For instance, while the properties of errors of measurement may be reasonably transportable, the distribution of the true (or latent) predictor X is rarely transportable, since it depends so heavily on the population being sampled. Problems arise because corrections for measurement error involve not only the measurement error process but also the distribution of X. For example, systolic blood pressure measurements in the MRFIT study and the Framingham Heart Study may well have the same measurement error variance, but the distribution of true blood pressure X appears to differ substantially in the two studies, and the "correction for attenuation" described below cannot be transported from Framingham to MRFIT (see [17] for further details).

Is there an "Exact" Predictor?

We have based our discussion on the existence of an exact predictor X and measurement error models that provide information about this predictor. However, in practice, it is often the case that the definition of "exact" needs to be carefully considered prior to discussion of error models. In the measurement error literature the term "**gold standard**" is often used for the operationally defined exact predictor, though sometimes this term is used for an exact predictor that cannot be operationally defined. Using an operational definition for an "exact" predictor is often reasonable and justifiable on the grounds that it is the best one could ever possibly hope to accomplish. However, such definitions may be controversial. For example, consider the problem of relating breast cancer risk to the dietary intake of fat. One way to determine whether decreasing one's fat intake lowers the risk of developing breast cancer is to conduct a **clinical trial** in which members of the treatment group are encouraged to reduce fat intakes. If instead one uses observational prospective data, along with an operational definition of long-term intake, one should be aware that the results of a measurement error analysis could be invalid if true long-term intake and operational long-term intake differ in subtle ways.

Differential and Nondifferential Error

It is important to make a distinction between *differential* and *nondifferential* measurement error. Nondifferential measurement error occurs in a broad sense when one would not even bother with W if X were available, i.e. W has no information about the response other than what is available in X.

Nondifferential measurement error typically holds in **cohort studies**, but is often a suspect assumption in **case–control studies**.

Technically, measurement error is nondifferential if the distribution of Y given (X, Z, W) depends only on (X, Z). In this case W is said to be a *surrogate*. Measurement error is *differential* otherwise.

For instance, consider the Framingham example. The predictor of major interest is long-term systolic blood pressure, X, but we can only observe blood pressure on a single day, W. It seems plausible that a single day's blood pressure contributes essentially no information over and above that given by true long-term blood pressure, and hence that measurement error is nondifferential. The same remarks apply to the nutrition examples: measuring diet on a single day should not contribute information not already available in long-term diet.

Many problems can be analyzed plausibly assuming nondifferential measurement error, especially when the **covariate** measurements occur at a fixed point in time, and the response is measured at a later time, as is typical in cohort studies.

There are two exceptions that need to be kept in mind. First, in case–control studies, the disease response is obtained first, and then one measures antecedent exposures and other covariates. In nutrition studies, this ordering of measurement may well cause differential measurement error. For instance, here the true predictor would be long-term dietary intake before diagnosis, but the dietary interview data are obtained only after diagnosis. A woman who develops breast cancer may exaggerate her estimated fat intake, thus introducing **recall bias** (*see* **Bias in Case–Control Studies**). In such circumstances, estimated fat intake will be associated with disease status even after conditioning on true long-term diet before diagnosis.

When measurement error is nondifferential, one can estimate parameters in models for responses given true covariates even when the true covariates are not observable. This is not true when measurement error is differential, except for the linear model. With differential error, one must obtain a validation subsample in which both true covariate measurements and surrogate measurements are available. Most of this article focuses on nondifferential measurement error models. Differential models with a **validation study** are typically best analyzed by techniques for handling missing data (*see* **Missing Data in Epidemiologic Studies**; **Missing Data**; **Multiple Imputation Methods**).

Prediction

Prediction of a response is different from estimation and inference for parameters. If a predictor X is measured with error, and one wants to predict a response *based on the error-prone version* W of X, then, except for an important case discussed below, it makes little sense to worry about measurement error. The reason for this is quite simple. If one has an original set of data (Y, W) then one can fit a convenient model to Y as a function of W. Predicting Y from W is merely a matter of using this model for prediction. There is no need then for measurement error to play a role in the problem.

The one situation requiring that we model the measurement error occurs when we develop a **prediction** model using data from one population but we wish to predict in another population. A naive prediction model that ignores measurement error may not be transportable. This context often becomes quite complex, requiring a combination of missing data and measurement error techniques, and to the best of our knowledge has not been investigated in detail in the literature, an exception being [31].

Is Bias Always Towards the Null?

It is commonly thought that the effect of measurement error is to bias estimates of exposure effects "towards the null" (*see* **Bias Toward the Null**). Hence, one could ignore measurement error when testing the **null hypothesis** of no exposure effect, and one could assume that non-null estimates, if anything, underestimate the effect of exposure. This lovely and appealing folklore is sometimes true but, unfortunately, often wrong. We discuss this point in detail below. A numerical example has recently been provided to us by David Yanez, Richard Kronmal & Lynn Shemanski in a heart disease context with seven covariates and a baseline variable. They found that, while an analysis ignoring measurement error showed highly statistically significant effects in all variables, none of the effects was even close to being statistically significant when the analysis took measurement error into account.

Functional and Structural Models

The words *functional* and *structural* have important places in the area of measurement error models. They act as a shorthand terminology for the basic approach one uses to solve the problem. In *functional modeling* nothing is assumed about the Xs; they could be fixed constants (the usual definition) or random variables. In *structural modeling*, X is assumed to be random, and a parametric distribution (usually the **normal**) is assumed. There has traditionally been considerable concern in the measurement error literature about the **robustness** of estimation and inferences based upon structural models for unobservable variates. Fuller [30, p. 263] discusses this issue briefly in the classical **nonlinear regression** problem, and basically concludes that the results of structural modeling "may depend heavily on the (assumed) form of the X distribution". In probit regression, Carroll et al. [23] report that, if one assumes that X is normally distributed, and it really follows a **chi-square distribution** with one **degree of freedom**, then the effect on the likelihood estimate is "markedly negative"; see also [63]. Essentially all research workers in the measurement error field come to a common conclusion: likelihood methods can be of considerable value, but the possible nonrobustness of inference due to model **misspecification** is a vexing and difficult problem.

The issue of model robustness is hardly limited to measurement error modeling. Indeed, it pervades statistics, and has led to the rise of a variety of **semiparametric** and **nonparametric** techniques. From this general point of view, *functional modeling* may be thought of as a group of semiparametric techniques. Functional modeling uses parametric models for the response, but makes no assumptions about the distribution of the unobserved covariate.

There is no agreement in the statistical literature as to whether functional or structural modeling is more appropriate. Many researchers believe that one should make as few model assumptions as possible and favor functional modeling. The argument is that any extra efficiency gained by structural modeling is more than offset by the need to perform careful and often time-consuming **sensitivity analyses**. Other researchers believe that appropriate statistical analysis requires one to do one's best to model every feature of the data, and thus favor structural modeling.

We take a somewhat more relaxed view of these issues. There are many problems, e.g. linear and logistic regression with additive measurement error, where functional techniques are easily computed and fairly efficient, and we have a strong bias in such circumstances towards functional modeling. In other problems – for example, the segmented regression problem [38] – structural modeling clearly has an important role to play, and should not be neglected.

Measurement Error as a Missing Data Problem

From one perspective, measurement error models are special kinds of missing data problems, because the Xs, being mostly and often entirely unobservable, are obviously missing as well. Readers who are already familiar with linear measurement error models and functional modeling will be struck by the fact that most of the recent missing data literature has pursued likelihood and Bayesian methods, i.e. structural modeling approaches. Readers familiar with missing data analysis will also be interested to know that, in large part, the measurement error model literature has pursued functional modeling approaches. We feel that both functional and structural modeling approaches are useful in the measurement error context, and this article pursues both strategies.

The usual interpretation of the classical missing data problem [42] is that the values of some of the variables of interest may not be observable for all study participants. For example, a variable may be observed for 80% of the study participants, but unobserved for the other 20%. The techniques for analyzing missing data are continually evolving, but it is fair to say that most of the recent advances (**multiple imputation**, data augmentation, etc.) have been based on likelihood (and Bayesian) methods.

The classical measurement error problem discussed to this point is one in which one set of variables, which we call X, is *never* observable, i.e. always missing. As such, the classical measurement error model is an extreme form of a missing data problem, but with *supplemental information* about X in the form of a surrogate, which we call W. Part of the art in measurement error modeling concerns how the supplemental information is related to the unobservable covariate.

Because there is a formal connection between the two fields, and because missing data analysis has become increasingly parametric, it is important to consider likelihood and Bayesian analysis of measurement error models – topics taken up later in this article.

Linear Regression and the Effects of Measurement Error

A comprehensive account of linear measurement error models can be found in Fuller [30].

Many textbooks contain a brief description of measurement error in linear regression, usually focusing on **simple linear regression** and arriving at the conclusion that the effect of measurement error is to bias the slope estimate in the direction of 0. Bias of this nature is commonly referred to as *attenuation* or *attenuation to the null*.

In fact, though, even this simple conclusion has to be qualified, because it depends on the relationship between the measurement, W, and the true predictor, X, and possibly other variables in the regression model as well. In particular, the effect of measurement error depends upon the model under consideration and on the joint distribution of the measurement error and the other variables. In **multiple linear regression**, the effects of measurement error vary, depending on: (i) the regression model, be it additive or multiple regression; (ii) whether or not the predictor measured with error is univariate or **multivariate**; and (iii) the presence of bias in the measurement. The effects can range from the simple attenuation described above to situations where: (i) real effects are hidden; (ii) observed data exhibit relationships that are not present in the error-free data; and (iii) even the signs of estimated coefficients are reversed relative to the case with no measurement error.

The key point is that the measurement error distribution determines the effects of measurement error, and thus appropriate methods for correcting for the effects of measurement error depend on the measurement error distribution.

Simple Linear Regression with Additive Error: Regression to the Mean

We start with the simple linear regression model $Y = \beta_0 + \beta_x X + \varepsilon$, where the scalar X has mean μ_x and variance σ_x^2, and the error in the equation ε is independent of X, has mean zero, and variance σ_ε^2. The error model is additive as in (1). In this classical additive measurement error model, it is well known that an ordinary **least squares** regression ignoring measurement error produces an estimate not of β_x, but instead of $\beta_{x*} = \lambda\beta_x$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1. \tag{3}$$

Thus ordinary least squares regression of Y on W produces an estimator that is attenuated to 0. The attenuating factor, λ, is called the *reliability ratio*.

One would expect that, because W is an error-prone predictor, it has a weaker relationship with the response than does X. This can be seen both by the attenuation and also by the fact that the residual variance of this regression is increased, being not σ_ε^2 but instead

$$\begin{aligned}\mathrm{var}(Y|W) &= \text{residual variance of observed data} \\ &= \sigma_\varepsilon^2 + \lambda\beta_x^2\sigma_u^2.\end{aligned}$$

This facet of the problem is often ignored, but it is important. *Measurement error causes a double-whammy*: not only is the slope attenuated, but the data are more noisy, with an increased error about the line.

To illustrate the attenuation associated with the classical additive measurement error, the results of a small simulation are displayed in Figure 1.

Ten observations were generated with $\sigma_x^2 = \sigma_u^2 = 1$, $(\beta_0, \beta_x) = (0, 1)$, and $\sigma_\varepsilon^2 = 0.25$. The filled circles and steeper line depict the true but unobservable data (Y, X) and the regression line of Y on X. The empty circles and attenuated line depict the observed (Y, W) data and the linear regression of Y on W.

Figure 1 is indicative of a phenomenon called **regression to the mean**. Intuitively, what this means is that the extremes in the observed (W) data are *too* extreme, and that the true X is closer to the mean of the data. In fact, in normally distributed data, if X has a population mean μ_x, then having observed the fallible instrument, the best prediction of X is $\mu_x(1 - \lambda) + \lambda W$, where $\lambda < 1$ is defined in (3). The net effect is that the best (linear) predictor of X is always closer to the overall mean than any observed W.

The foregoing is one facet of regression to the mean. A more common definition is complementary. In a study participant with an unusually large observed W, if one repeats the measurement and obtains a second (replicated) measure, then this replicate is generally less (and often much less) than the original extreme value.

For instance, in a study of true long-term fat intake (X) using a 24 h recall instrument (W), if one focuses on the person with the highest reported fat intake, then (i) that person's true fat intake is most likely less than the observed intake, and (ii) if one repeats the 24 h recall instrument, then the new reported fat intake is likely to be less than the original reported fat intake.

The second part of the "double-whammy" is a loss of **power**. The following example is meant to illustrate this loss of power, and it is easiest to do this illustration in the special case that all variances are known. Suppose that one wants to test the null hypothesis H_0: $\beta_x = 0$ of zero slope against the one-sided alternative H_1: $\beta_x > 0$, using a test with a 5% level (type I error) which has power 80% to detect that the slope $\beta_x = 0.75$. With known variances, in the absence of measurement error, the required sample size is

$$n = \frac{(z_{0.95} + z_{0.80})^2 \sigma_\varepsilon^2}{\sigma_x^2 \beta_x^2},$$

where z_α is the usual α percentile of the normal distribution. With measurement error, the same formula applies, except that, with $\beta_x = 0.75$, one replaces σ_ε^2 by $\sigma_\varepsilon^2 + \lambda\beta_x^2\sigma_u^2$, σ_x^2 by $\sigma_x^2 + \sigma_u^2$, and β_x by $\lambda\beta_x$. In Figure 2, we plot the sample sizes as a function of the measurement error variance in the case that X has variance $\sigma_x^2 = 1$ and the error about the line has variance $\sigma_\varepsilon^2 = 1$. In the absence of measurement error, approximately 10 observations are required to obtain the desired power. However, if the measurement error variance $\sigma_u^2 = 1$ and thus the reliability $= 1/2$, then approximately 30 observations are required. Thus, measurement error causes a loss of power. In planning a study with a large measurement error in a covariate, one will typically require a much larger sample size to meet power goals than if there were no measurement error.

It is a common belief that the effect of measurement error is always to attenuate the slope of the regression line, but in fact attenuation depends critically on the assumed classical additive measurement error model. Very different results are obtained if measurement errors are differential. One example where this problem may arise is in dietary calibration studies. In a typical dietary calibration study, one is interested in the relationship between a self-administered food frequency questionnaire (FFQ, the value of Y) and usual (or long-term) dietary intake (the value of X) as measures of, for example, the percentage of calories from fat in a person's diet. FFQs are thought to be biased for usual intake, and in a calibration study researchers will obtain a second measure (the value of W), typically from a food diary, a 24h recall, or a short-term biomarker. In this context, it is often assumed that the diary, recall, or biomarker is unbiased for usual intake. If, as sometimes occurs, the FFQ and the diary/recall are given very nearly contemporaneously, it is unreasonable to assume that the error in the relationship between the FFQ and usual intake is uncorrelated with the error in the relationship between a diary–recall–biomarker and usual intake. This correlation has been demonstrated [29], and gives rise to differential error. It can be shown [20] that, if there is significant **correlation** between the measurement error and the error about the true line, then the regression of Y on W can have a slope biased away from the null. Thus, correction for bias induced by measurement error clearly depends on the nature, as well as the extent, of the measurement error.

Multiple Regression: Single Covariate Measured with Error

In multiple linear regression the effects of measurement error are more complicated, even for the classical additive error model.

We now consider the case where X is scalar, but there are additional covariates Z measured without error. In the linear model the mean is $\beta_0 + \beta_x X + \beta_z Z$. Under the usual conditions of independence of errors, the least squares regression estimator of the coefficient of W consistently estimates $\lambda_1\beta_x$, where

$$\lambda_1 = \frac{\sigma_{x|z}^2}{\sigma_{w|z}^2} = \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}, \tag{4}$$

and $\sigma_{w|z}^2$ and $\sigma_{x|z}^2$ are the (residual) variances of the regressions of W on Z and X on Z, respectively. Note that λ_1 is equal to the simple linear regression attenuation $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ only when X and Z are uncorrelated. *The basic point is that the attenuation depends on the relationships among the covariates.*

The problem of measurement-error-induced bias is not restricted to the regression coefficient of X. The coefficient of Z is also biased in general, unless Z is independent of X [19]. In fact, naive ordinary least

squares estimates not β_z but rather

$$\beta_{z*} = \beta_z + \beta_x(1 - \lambda_1)\gamma_z, \tag{5}$$

where γ_z is the coefficient of Z in the regression of X on Z.

This result has important consequences in epidemiology when interest centers on the effects of covariates measured without error. For example, consider the case that Z is a **binary** exposure variable (exposed or not) which is classified correctly, and X is an important confounder measured with significant error. Then Carroll et al. [19] show that ignoring measurement error produces a **consistent** estimate of the exposure effect only if the design is balanced, i.e. X has the same mean in both groups and is independent of treatment. With considerable imbalance, the naive analysis may lead to the conclusion that: (i) there is a treatment effect when none actually exists; and (ii) the effects are negative when they are actually positive, and vice versa. In most observational studies the confounder and the exposure are correlated (see [34] and [35]). Errors in measuring the confounders can produce very misleading results.

Multiple Covariates Measured with Error

If multiple covariates are measured with error, then the direction of the bias induced by this error does not follow any simple pattern. One may have attenuation, reverse-attenuation, changes of sign, or an observed positive effect even at a true null model. This is especially the case when the predictors measured with error are correlated or their errors are correlated. In such a problem, there really seems to be no substitute for a careful measurement error analysis.

Correcting for Bias

As we have just seen, the ordinary least squares estimator is typically biased under measurement error, and the direction and magnitude of the bias depends on the regression model and the measurement error distribution. We next describe two commonly used methods for eliminating bias.

In simple linear regression with the classical additive error model, we have seen in (3) that ordinary least squares is an estimate of $\lambda\beta_x$; recall that λ is called the reliability ratio. If the reliability ratio were known, then one could obtain a proper estimate of β_x simply by dividing the ordinary least squares slope by the reliability ratio.

Of course, the reliability ratio is rarely known in practice, and one has to estimate it. If $\hat{\sigma}_u^2$ is an estimate of the measurement error variance (this is discussed below), and if $\hat{\sigma}_w^2$ is the sample variance of the Ws, then a consistent estimate of the reliability ratio is $\hat{\lambda} = (\hat{\sigma}_w^2 - \hat{\sigma}_u^2)/\hat{\sigma}_w^2$. The resulting estimate is $\beta_{x*}/\hat{\lambda}$. In small samples the sampling distribution of this estimate is highly skewed, and in such cases a modified version of the **method of moments** estimator is recommended [30].

The **algorithm** described above is called the *method-of-moments* estimator. The terminology is apt, because ordinary least squares and the reliability ratio depend only on moments of the observed data.

The method-of-moments estimator can be constructed for the **general linear model**, as well as for simple linear regression. Consult the book by Fuller [30], especially Chapter 2.

Another well publicized method for linear regression in the presence of measurement error is *orthogonal regression*. It is fairly rare in epidemiologic situations that the model underlying orthogonal regression holds [15], and we will not discuss the method any further.

Bias vs. Variance

Estimates that do not account for measurement error are typically biased. Unfortunately, correcting for this bias often has a price. In particular, the resulting corrected estimator will be more variable than the biased estimator, and wider **confidence intervals** result. For example, Rosner et al. [60] describe a problem in logistic regression, where the response is the development of breast cancer, and the predictor measured with error is daily saturated fat intake. Ignoring measurement error, they obtained an estimated **odds ratio** for saturated fat of 0.92, with a 95% confidence interval from 0.80 to 1.05. The corrected estimated odds ratio was 0.83 with a confidence interval from 0.61 to 1.12, which is twice as wide as the previous interval.

Attenuation in General Problems

We have already seen that, with multiple covariates, even in linear regression the effects of measurement

error are complex, and not easily described. In this Section, we provide a brief overview of what happens in nonlinear models.

Consider a scalar covariate X measured with error, and suppose that there are no other covariates. In the classical error model for simple linear regression we have seen that the bias caused by measurement error is always in the form of attenuation, so that ordinary least squares preserves the sign of the regression coefficient asymptotically, but is biased towards zero. Attenuation is a consequence then of (i) the simple linear regression model and (ii) the classical additive error model. Without (i) and (ii), the effects of measurement error are more complex; we have already seen that attenuation may not hold if (ii) is violated.

In logistic regression, when X is measured with additive error, attenuation does not always occur, but it is typical and generally much like that of linear regression.

Dosemeci et al. [28] give an example of **misclassification error** that shows that trends are not always preserved under nondifferential measurement error. Suppose that 1348 subjects are exposed at no ($X = 0$), low ($X = 1$), and high ($X = 2$) levels to a harmful substance. Suppose that the chance of an adverse outcome is 1/2, 2/3, and 6/7 for no, low, and high exposures, while the chances of the exposures themselves are 0.0059347, 0.8902077, and 0.1038576, respectively. If true exposure could be ascertained, then the expected outcomes would be as in the section of Table 1 labeled "true". If we were to regress Y on the **dummy variables** X_1 indicating low exposure ($X_1 = 1$), and X_2 indicating high exposure ($X_2 = 1$), then the true logistic regression parameters for X_1 and X_2 would be $\log 2 = 0.69$ and $\log 6 = 1.79$, respectively, indicating that the two higher exposure levels have response rates higher than the response rate associated with the no-exposure level. The true data clearly indicate a harmful effect due to exposure.

Table 1 A hypothetical logistic regression example with nondifferential measurement error. The entries are the expected counts. The true logistic parameters for dummy variables low and high exposure are $\log 2$ and $\log 6$, respectively, while the observed coefficients for the error prone data are $\log 0.46$ and $\log 0.53$, respectively

Disease status	Exposure = none	Exposure = low	Exposure = high
True			
$Y = 1$	4	800	120
$Y = 0$	4	400	20
Observed			
$Y = 1$	52	480	392
$Y = 0$	12	240	172

Now suppose, however, that measurement error (in this case misclassification) occurs, so that 40% of those truly at high exposure are misclassified into the no-exposure group, and 40% of those truly at low exposure are misclassified into the high-exposure group. Let W be the resulting variable taking on the three observed levels of exposure, with corresponding dummy variables W_1 and W_2. This is a theoretical example, of course, and one can criticize it for not being particularly realistic, but it is an example of nondifferential measurement error. The observed data we expect to see using the surrogates W_1 and W_2 are also given in Table 1.

The observed logistic regression parameters for W_1 and W_2 are $\log 0.46 = -0.78$ and $\log 0.53 = -0.63$, respectively, indicating that the two higher exposure levels have response rates lower than the response rate associated with the no-exposure level. The observed data suggest a beneficial effect due to exposure, even though the exposure is harmful!

Hypothesis Testing

In this section, we discuss hypothesis tests concerning regression parameters. To keep the exposition simple, we focus on linear regression. However, the results hold in some generality, especially for logistic and **Poisson regression**. We assume nondifferential measurement error and the classical additive error model.

The simplest approach to hypothesis testing calculates the required test statistic from the parameter estimates obtained from a measurement error analysis and their estimated standard errors. Such tests are justified whenever the estimators themselves are justified. However, this approach to testing is only possible when the indicated methods of estimation are possible, and thus requires either knowledge of the measurement error variance, or the presence of validation data, or replicate measurements, or instrumental variables.

There are certain situations in which naive hypothesis tests are justified and thus can be performed

without additional data or information of any kind. Here "naive" means that we ignore measurement error and substitute W for X in a test that is valid when X is observed. This Section studies naive tests, describing when they are and are not acceptable.

We use the criterion of asymptotic validity to distinguish between acceptable and nonacceptable tests. We say a test is asymptotically valid if its type I error rate approaches its nominal level as the sample size increases. Asymptotic validity (which we shorten to validity) of a test is a minimal requirement for acceptability.

The main results on the validity of naive tests under nondifferential measurement error are as follows:

1. The naive test of no effects due to X is valid. This means that if one wants to test whether *all* components of X together have no effect, then it is valid to ignore nondifferential measurement error. Thus, for example, if X is the exposure, then a valid test of the null hypothesis for X is obtained by ignoring measurement error and performing the standard test for the problem at hand.
2. The naive test described above is also fully efficient if X is linearly related to W and Z, but not otherwise [79]. Thus, while in principle one can obtain additional power by a measurement error analysis, many times in practice the naive test of the null hypothesis for X is reasonably efficient.
3. In many problems, more than one covariate is measured with error. For example, suppose that the exposure and one of the confounders are measured with error. Generally, the naive test of the null hypothesis for the exposure is invalid, except under special circumstances, e.g. the exposure and confounder are statistically independent, as are their measurement errors.
4. In general, naive tests for Z are invalid, except possibly if Z is uncorrelated with X. Thus, if X is the exposure and Z is a confounder, then naive tests for significance of the exposure are valid, but they are not valid for testing the significance of the confounder. Somewhat more troubling, though, is the case when X is a confounder related to the exposure Z; here the naive test for the exposure is generally invalid, *even if exposure is measured without error*. We have mentioned this example previously in the case that the exposure is binary (see [19]).

The last point can be demonstrated in the **analysis of covariance**, in which Z is a binary exposure variable and X is a confounder with strong predictive ability. In the analysis of covariance, the model is

$$Y = \beta_0 + \beta_z Z + \beta_x X + \varepsilon,$$

where ε is the error about the line, with variance σ_ε^2. The binary indicator Z takes on the values ± 1, with 50% of the data being unexposed ($Z = -1$) and 50% of the data being exposed ($Z = 1$). Within the unexposed group, X has mean $-\theta/2$ and variance σ_x^2, while, within the exposed group, X has mean $\theta/2$ and variance σ_x^2. The difference between the means for X in the two groups is θ. In a randomized **clinical trial**, one would expect that $\theta = 0$, since **randomization** ensures that the population means of X are the same in the exposed and unexposed groups. In nonrandomized studies, one would expect that $\theta \neq 0$. Thus, the larger the value of θ, the more unbalanced is the study. In Figure 3, we plot the level (type I error) of the test for the exposure effect which ignores measurement error as a function of the difference in group means θ. This calculation is done for the case that $n = 20$ (10 exposed and 10 unexposed), $\sigma_\varepsilon^2 = 1, \sigma_x^2 = 1$, and $\beta_x = 1$, for reliability ratios $\dot{\lambda} = 1/2$ and $= 2/3$. The graph shows that if the means of the confounders are sufficiently different, then, instead of a type I error of 5%, the test for exposure effect which ignores measurement error in the *confounder* can have type I error rates higher than 10%, even for such small sample sizes.

Regression Calibration and SIMEX

We now describe two simple, generally applicable approaches to nondifferential measurement error analysis, regression calibration, and simulation extrapolation (SIMEX).

Regression Calibration

The basis of regression calibration is the replacement of X by the regression of X on (Z,W). After this approximation, one performs a standard analysis. This *regression calibration* algorithm was suggested as a general approach by Carroll & Stefanski [16] and

Gleser [33]. Prentice [52] pioneered the idea for the **proportional hazard** model, and a modification of it has been suggested for this topic by Clayton [25]; see below. Armstrong [4] suggests regression calibration for **generalized linear models**, and Fuller [30, pp. 261–262] briefly mentions the idea. Rosner et al. [59, 60] have developed the idea for **logistic regression** into a workable and popular methodology, complete with a good computer program. Because of the importance of their contribution to epidemiologic applications, regression calibration is often referred to as "Rosner's Method". Other interesting and important applications and methodology related to regression calibration include work by Whittemore [83], Pierce et al. [51], Liu & Liang [43], and Kuha [39]. In some special cases, regression calibration is equivalent to the classical method of moments bias correction.

The main justifications of the regression calibration approximation are that, for some models, e.g. **loglinear** mean models and **linear regression**, the regression calibration approximation is often exact except for a change in the intercept parameter. For logistic regression, in many cases the approximation is almost exact.

The Regression Calibration Algorithm. The regression calibration algorithm is what Pierce et al. [51] call a "replacement method":

1. Using replication, validation or instrumental data, estimate the regression of X on (Z,W) (see below). This is called the *calibration function*.
2. Replace the unobserved X by its estimate from the regression model, and then run a standard analysis to obtain parameter estimates.
3. Adjust the resulting standard errors to account for the estimation at the first step, using either the bootstrap or asymptotic methods [20].

The simplest form of regression calibration is the "correction for attenuation" used in linear regression. It is easiest to describe in the following situation:

1. X is a scalar.
2. The measurement error is additive, with estimated error variance $\hat{\sigma}_u^2$.

For estimating the effect of X, the regression calibration estimator is formed by three steps: (i) form the naive estimator by ignoring measurement error; (ii) let $\hat{\sigma}_{w|z}^2$ be the regression **mean square error** from a linear regression of W on Z (this is the sample variance of the Ws if there are no other covariates Z); (iii) the regression calibration estimator is defined by multiplying the naive estimator by $\hat{\sigma}_{w|z}^2/(\hat{\sigma}_{w|z}^2 - \hat{\sigma}_u^2)$.

Estimating the Calibration Function Parameters. With *internal validation data*, the simplest approach is to regress X on the other covariates (Z, W) in the validation data. While linear regression will be typical, it is not required.

In some problems, an *unbiased second instrument* T is available for a subset of the study participants. For instance, one might be interested in $X =$ caloric intake over a year, but have available only $T =$ the result of a biomarker experiment using a technique known as doubly labeled water over a 2 week period, which does not equal X because it does not take into account the variability of diet over a year. In this case one uses the regression of T on (Z,W) as the calibration function. This is the method used by Rosner et al. [59] in their analysis of the Nurses' Health Study.

Finally, in the classical additive error model, one often has merely a second measurement (a replicate) for a subset of the study population. One could treat this replicate as an unbiased second instrument and apply the method described in the previous paragraph. If the Ws are not too far from normally distributed, a more efficient method is to use the so-called best linear approximation to the calibration function (see [20, pp. 47–48]). This takes into account that some of the study participants do have a replicated W and hence use the data in a reasonably efficient fashion.

Suppose there are k_i replicate measurements of X_i, and that $\overline{W}_i$ is their mean. Replication enables us to estimate the measurement error **covariance matrix** σ_u^2 by the usual **variance components** analysis, as follows:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(W_{ij} - \overline{W}_{i\cdot})^2}{\sum_{i=1}^{n}(k_i - 1)}. \qquad (6)$$

The calibration function is defined as follows. Suppose the observations are $(Z_i, \overline{W}_{i\cdot})$, where $\overline{W}_{i\cdot}$ is the

mean of k_i replicates. We use **analysis of variance** formulas. Let

$$\hat{\mu}_x = \hat{\mu}_w = \frac{\sum_{i=1}^{n} k_i \overline{W}_{i\cdot}}{\sum_{i=1}^{n} k_i}, \qquad \hat{\mu}_z = \overline{Z}_{\cdot},$$

$$\nu = \frac{\sum_{i=1}^{n} k_i - \sum_{i=1}^{n} k_i^2}{\sum_{i=1}^{n} k_i},$$

$$\hat{\sigma}_z^2 = (n-1)^{-1} \sum_{i=1}^{n} (Z_i - \overline{Z}_{\cdot})^2,$$

$$\hat{\sigma}_{xz} = \frac{\sum_{i=1}^{n} k_i (\overline{W}_{i\cdot} - \hat{\mu}_w)(Z_i - \overline{Z}_{\cdot})}{\nu},$$

$$\hat{\sigma}_x^2 = \frac{\left\{ \left[\sum_{i=1}^{n} k_i (\overline{W}_{i\cdot} - \hat{\mu}_w)^2 \right] - (n-1)\hat{\sigma}_u^2 \right\}}{\nu}.$$

The resulting estimated calibration function which is used to replace $\mathbf{X}$ in the standard analysis is

$$\hat{\mu}_w + (\hat{\sigma}_x^2, \hat{\sigma}_{xz}) \begin{bmatrix} \hat{\sigma}_x^2 + \hat{\sigma}_u^2/k_i & \hat{\sigma}_{xz} \\ \hat{\sigma}_{xz} & \hat{\sigma}_z^2 \end{bmatrix}^{-1} \begin{pmatrix} \overline{W}_{i\cdot} - \hat{\mu}_w \\ Z_i - \overline{Z}_{\cdot} \end{pmatrix}. \tag{7}$$

Expanded Regression Calibration Models. Rudemo et al. [61], Carroll & Stefanski [16] and Carroll et al. [20] all describe refinements to the regression calibration algorithm. Rudemo et al. [61] describe a bioassay problem (*see* **Biological Assay, Overview**) with a heteroscedastic Berkson error model. Racine-Poon et al. [56] describe a similar problem.

There is a long history of approximately consistent estimates in nonlinear problems, of which regression calibration and the SIMEX method are the most recent such methods. Readers should also consult Stefanski & Carroll [70], Stefanski [67], Amemiya & Fuller [3], and Whittemore & Keller [85] for other approaches.

The SIMEX Method

We now describe a method that shares the simplicity of regression calibration and is well suited to problems with additive or multiplicative measurement error. Simulation extrapolation (SIMEX) is a **simulation**-based method of estimating and reducing bias due to measurement error. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error-induced bias vs. the variance of the added measurement error, and **extrapolating** this trend back to the case of no measurement error. The technique was proposed by Cook & Stefanski [26], and further developed by Carroll et al. [22] and Stefanski & Cook [73]. See also Stefanski, [68].

An integral component of SIMEX is a self-contained simulation study resulting in **graphical displays** that illustrate the effect of measurement error on parameter estimates and the need for bias correction. The graphical displays are especially useful when it is necessary to motivate or explain a measurement error model analysis.

This Section describes the basic idea of SIMEX, focusing on linear regression with additive measurement error. For this simple model the effect of measurement error on the least squares estimator is easily determined mathematically, as we have shown. *The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can also be determined experimentally via simulation*. If we regard measurement error as a factor whose influence on an estimator is to be determined, we are naturally led to consider simulation experiments in which the level of the measurement error, i.e. its variance, is varied intentionally.

The SIMEX Algorithm

Suppose that, in addition to the original data used to calculate the naive estimate $\hat{\beta}_{x,\text{naive}}$, there are $M-1$ additional data sets available, each with successively larger measurement error variances, say $(1+\zeta_m)\sigma_u^2$, where $0 = \zeta_1 < \zeta_2 < \cdots < \zeta_M$. Of course, the least squares estimate of slope from the mth data set ignoring measurement error, $\hat{\beta}_{x,m}$, consistently estimates $\beta_x \sigma_x^2 / [\sigma_x^2 + (1+\zeta_m)\sigma_u^2]$.

We can think of this problem as a nonlinear regression model, with dependent variable $\hat{\beta}_{x,m}$ and

independent variable ζ_m, having a mean function of the form

$$\mathcal{G}(\zeta) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1+\zeta)\sigma_u^2}, \zeta \geq 0.$$

The parameter of interest, β_x, is obtained from $\mathcal{G}(\zeta)$ by extrapolation to $\zeta = -1$. We describe the process schematically in Figure 4.

SIMEX imitates the procedure just described. In the *simulation step*, additional independent measurement errors with variance $\zeta_m \sigma_u^2$ are generated and added to the original data, thereby creating data sets with successively larger measurement error variances. For the mth data set, the total measurement error variance is $\sigma_u^2 + \zeta_m \sigma_u^2 = (1+\zeta_m)\sigma_u^2$. Next, estimates are obtained from each of the resulting contaminated data sets. The simulation and reestimation step is repeated a large number of times (to remove simulation variability) and the average value of the estimate for each level of contamination is calculated. These averages are plotted against the ζ values, and regression techniques are used to fit an extrapolant function to the averaged, error-contaminated estimates. Extrapolation back to the ideal case of no measurement error ($\zeta = -1$) yields the SIMEX estimate.

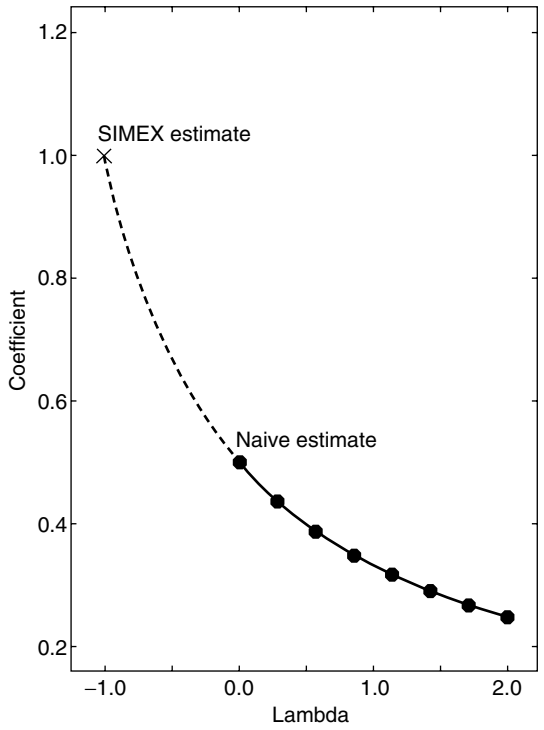

Figure 4 A generic plot of the effect of measurement error of size $(1+\zeta)\sigma_u^2$ on parameter estimates. The value of ζ is on the x-axis, while the value of the estimated coefficient is on the y-axis. The SIMEX estimate is an extrapolation to $\zeta = -1$. The naive estimate occurs at $\zeta = 0$

The first part of the algorithm is the simulation step. As described above, this involves using simulation to create additional data sets with increasingly large measurement error $(1+\zeta)\sigma_u^2$. For any $\zeta \geq 0$, define

$$W_{b,i}(\zeta) = W_i + \zeta^{1/2} U_{b,i}, \quad i = 1, \ldots, n, \quad b = 1, \ldots, B, \qquad (8)$$

where the computer-generated *pseudo-errors*, $\{U_{b,i}\}_{i=1}^n$, are mutually independent, independent of all the observed data, and identically distributed, normal random variables with mean 0 and variance σ_u^2.

Having generated the new predictors, we compute the resulting naive estimates, component by component. For each ζ, do this B times ($B = 100$ usually works fine) and compute their average, $\hat{\beta}(\zeta)$. It is the points $\{\hat{\beta}(\zeta_m), \zeta_m\}_1^M$ that are plotted as filled circles in Figure 4. This is the simulation component of SIMEX.

The extrapolation step of the proposal entails modeling each of the components of $\hat{\beta}(\zeta)$ as functions of ζ for $\zeta \geq 0$, and extrapolating the fitted models back to $\zeta = -1$. In Figure 4 the extrapolation is indicated by the dashed line and the SIMEX estimate is plotted as a cross. Carroll et al. [20] describe practical modifications of the algorithm, and how to estimate variances of parameters. Inference for SIMEX estimators can also be performed via the **bootstrap**. Because of the computational burden of the SIMEX estimator, the bootstrap requires considerably more computing time than do other methods. Without efficient implementation of the estimation scheme at each step, the SIMEX bootstrap may take an inconveniently long time to compute. On my computing system for measurement error models the implementation is efficient, and most bootstrap applications take little time.

We have described the SIMEX algorithm in terms of the additive measurement error model. However, SIMEX applies more generally.

For example, consider multiplicative error. Taking logarithms transforms the **multiplicative model** to the **additive model**. SIMEX works naturally here,

in that one performs the simulation step (8) on the logarithms of the Ws and not on the Ws themselves.

With replicates, one can also investigate the appropriateness of different **transformations**. For example, after transformation, the **standard deviation** of the intra-individual replicates should be uncorrelated with their mean, and one can find the transformation (logarithm, square root, etc.) which makes the two uncorrelated.

Example

To illustrate SIMEX, we use data from the Framingham Heart Study, correcting for bias due to measurement error in systolic blood pressure measurements. The Framingham study consists of a series of exams taken two years apart. We use Exam #3 as the baseline. There are 1615 men aged 31–65 in this data set, with the outcome, Y, indicating the occurrence of coronary heart disease (CHD) within an 8-year period following Exam #3; there were 128 such cases of CHD. Predictors employed in this example are the patient's age at Exam #2, smoking status at Exam #1, and serum cholesterol at Exams #2 and #3, in addition to systolic blood pressure (SBP) at Exam #3, the latter being the average of two measurements taken by different examiners during the same visit. In addition to the measurement error in SBP measurements, there is also measurement error in the cholesterol measurements. However, for this example we ignore the latter source of measurement error and illustrate the methods under the assumption that only SBP is measured with error.

The covariates measured without error, Z, are age, smoking status, and serum cholesterol, with $W = \log(\text{SBP} - 50)$. Implicitly, we are defining X as the long-term average of W. We illustrate the analyses for the case where W is the mean of the two transformed SBPs, and σ_u^2 is estimated using (6). The estimated linear model correction for attenuation, or inverse of the reliability ratio, is 1.16; if only one SBP measurement were used, the correction would be 1.33.

Figure 5 contains plots of the logistic regression coefficients $\hat{\Theta}(\zeta)$ for eight equally spaced values of ζ spanning [0, 2] (solid circles). For this example $B = 2000$. The points plotted at $\zeta = 0$ are the naive estimates $\hat{\Theta}_{\text{naive}}$. The nonlinear least-squares fits of $\mathcal{G}_{\text{RL}}(\lambda, \Gamma)$ to the components of $\{\hat{\Theta}(\zeta_m), \zeta_m\}_1^8$ (solid curves) are extrapolated to $\zeta = -1$ (dashed curves), resulting in the SIMEX estimators (crosses). The open circles are the SIMEX estimators that result from fitting quadratic extrapolants. To preserve clarity the quadratic extrapolants were not plotted. Note that the quadratic-extrapolant estimates are conservative relative to the rational linear-extrapolant estimates in the sense that they fall between the rational linear-extrapolant estimates and the naive estimates.

We have stated previously that the SIMEX plot displays the effect of measurement error on parameter estimates. This is especially noticeable in Figure 5. In each of the four graphs in Figure 5, the range of the ordinate corresponds to a one-standard-error confidence interval for the naive estimate constructed using the information standard errors. Thus Figure 5 illustrates the effect of measurement error relative to the variability in the naive estimate. It is apparent that the effect of measurement error is of practical importance only on the coefficient of $\log(\text{SBP} - 50)$.

Conditional and Corrected Scores for Functional Modeling

Regression calibration and SIMEX are easily applied general methods for nondifferential error. Although the resulting estimators are **consistent** in important special cases such as linear regression and loglinear mean models, they are only approximately consistent in general.

For certain generalized linear models and measurement error distributions there are easily applied functional methods that are fully (and not just approximately) consistent, and make no assumptions about the distribution of X.

We focus on the case of additive normally distributed measurement error with measurement error variance σ_u^2. Although the problem has this parametric error assumption, it also has a nonparametric component: no assumptions are made about the true predictors X.

Suppose for the sake of discussion that the measurement error variance σ_u^2 is known. In the functional model, the unobservable Xs are fixed constants, and hence the unknown parameters include the Xs. With additive normally distributed measurement error, one strategy is to maximize the joint density of the observed data with respect to all of the unknown parameters including the Xs. While this works for linear regression [32], it fails for more complex models

such as logistic regression. Indeed, the logistic regression functional maximum likelihood estimator is both inconsistent and difficult to compute [70]. An alternative approach is to change to the structural model and apply likelihood techniques (see below).

In this Section, we consider two functional methods, the conditional-score and corrected-score methods. We start with logistic and gamma–loglinear modeling as important examples for which these techniques apply. The conditional methods exploit

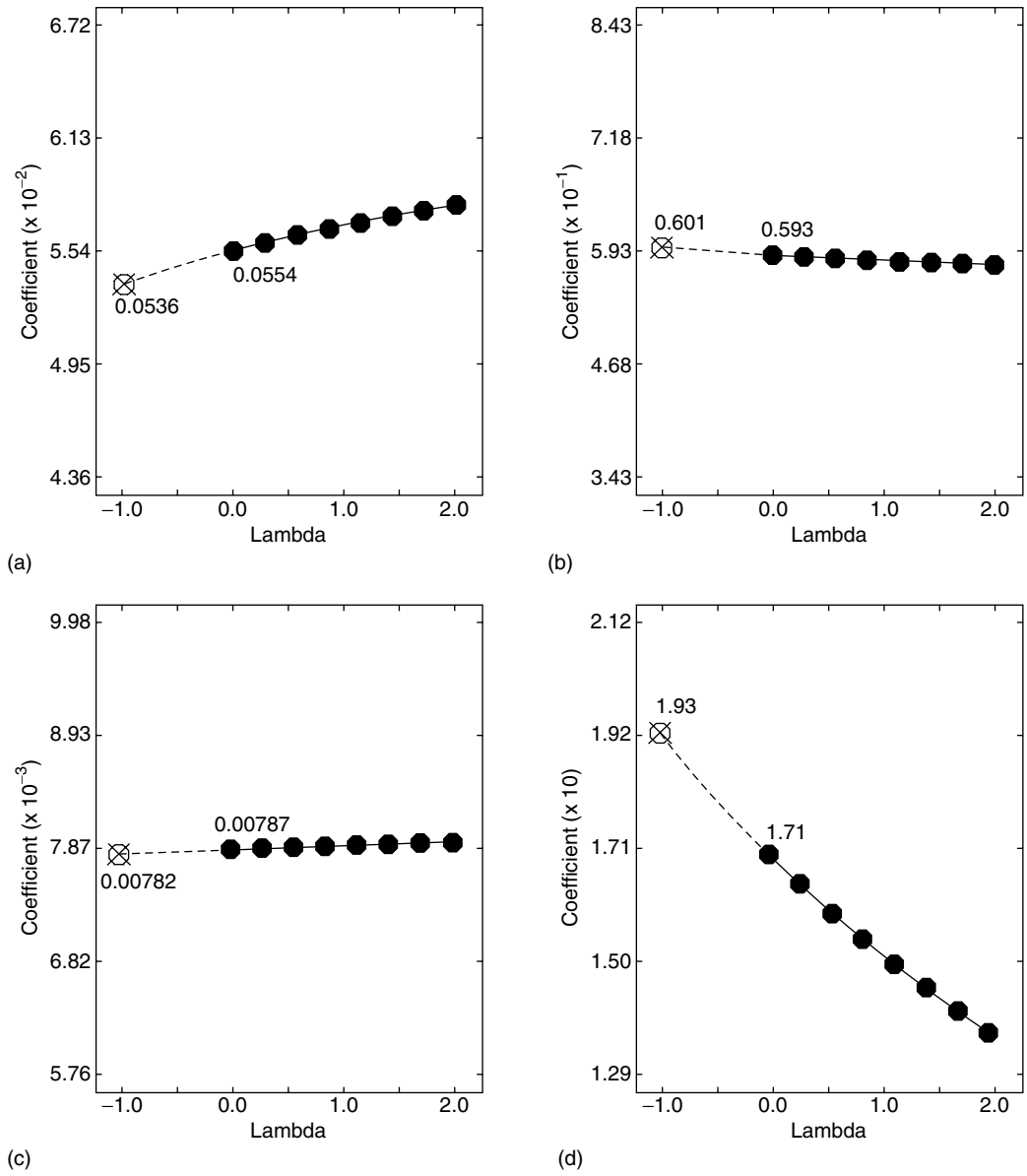

Figure 5 Coefficient extrapolation functions for the Framingham logistic regression modeling. The simulated estimates $\{\hat{\Theta}(\zeta_m), \zeta_m\}_1^8$ are plotted (solid circles) and the fitted rational linear extrapolant (solid line) is extrapolated to $\zeta = -1$ (dashed line), resulting in the SIMEX estimate (cross). Open circles indicate SIMEX estimates obtained with the quadratic extrapolant. (a) Age; (b) Smoking; (c) Cholesterol; (d) log(SBP − 50)

special structures in important models such as linear, logistic, Poisson loglinear, and gamma-inverse, and then use a traditional statistical device – conditioning on **sufficient statistics** – to obtain estimators. The corrected-score method effectively estimates the estimator one would use if there were no measurement error.

First consider the **multiple linear regression** model with mean $\beta_0 + \beta_x X + \beta_z Z$, and write the unknown regression parameter as $\Theta = (\beta_0, \beta_x, \beta_z)$. When the measurement error is additive with nondifferential measurement error variance Σ_{uu}, the usual **method-of-moments** regression estimator can be derived as the solution to the equation

$$\sum_{i=1}^{n} \psi_* (Y_i, Z_i, W_i, \Theta, \Sigma_{uu}) = 0, \qquad (9)$$

where

$$\psi_* (Y, Z, W, \Theta, \Sigma_{uu}) = (Y - \beta_0 - \beta_x^t X - \beta_z^t Z) \times \begin{pmatrix} 1 \\ Z \\ W \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \Sigma_{uu}\beta_x \end{pmatrix}$$

is the *corrected score* for linear regression. If Σ_{uu} is unknown, then one substitutes an estimate of it into (9) and solves for the regression parameters.

The key point to note here is that, in solving (9), we need know nothing about the Xs. This feature is common to all the methods in this Section.

Eq. (9) is an example of an **estimating equation** approach for estimating a set of unknown parameters. The reader can consult the Appendix of [20] for an overview of estimating equations, although this is unnecessary for the purpose of using the methods. Asymptotic standard errors for the estimators can be derived using either the bootstrap or the sandwich formula.

Logistic regression is best handled using the conditional-score method. For example, consider the usual linear-logistic model, where Y is binary and has success probability following the logistic model $H(\beta_0 + \beta_x X + \beta_z Z)$. The conditional score is

$$\psi_*(Y, Z, W, \Theta, \sigma_u^2) = \{Y - H[\beta_0 - \beta_x^t \Delta(\cdot) - 0.5\beta_x^t \sigma_u^2 \beta_x - \beta_z^t Z]\} \begin{pmatrix} 1 \\ Z \\ \Delta(\cdot) \end{pmatrix}, \qquad (10)$$

where $\Delta(\cdot) = \Delta(Y, W, \beta_x, \sigma_u^2) = W + Y\sigma_u^2\beta_x$. Eq. (10) is substituted into (9), and the resulting equation is solved numerically.

When Y has a **gamma distribution** with loglinear mean $\exp(\beta_0 + \beta_x X + \beta_z Z)$, it has a variance which is ϕ times the square of the mean. For this important example, the corrected-score estimator is obtained from the corrected score

$$\psi_*(Y, Z, W, \Theta, \sigma_u^2) = \begin{pmatrix} 1 \\ Z \\ W \end{pmatrix} - \exp[\Delta(Z, W, \Theta, \sigma_u^2)] \times \begin{pmatrix} Y \\ ZY \\ Y(W + 0.5\sigma_u^2\beta_x) \end{pmatrix}, \qquad (11)$$

where $\Delta(Z, W, \Theta, \sigma_u^2) = -\beta_0 - \beta_x^t W - \beta_z^t Z - 0.5\beta_x^t \sigma_u^2 \beta_x$.

Unbiased Score Functions via Conditioning

The conditional estimators of Stefanski & Carroll [71] and Nakamura [48] are discussed in detail in Carroll et al. [20, Chapter 6]. They apply to linear, logistic, Poisson loglinear, and gamma inverse regression [the mean is $1/(\beta_0 + \beta_x X + \beta_z Z)$]. Their methods have simple formulas for standard errors, although, of course, as usual, the bootstrap applies.

Exact Corrected Estimating Equations

Suppose that it is possible to find a function of the observed data, say $\psi_*(Y, Z, W, \Theta)$, having the property that

$$\mathrm{E}[\psi_*(Y, Z, W, \Theta)|Y, Z, X] = \psi(Y, Z, X, \Theta), \qquad (12)$$

for all Y, Z, X, and Θ. Then corrected score function estimators simply replace ψ by ψ_*. Corrected score functions satisfying (12) do not always exist, and finding them when they do is not always easy.

One useful class of models that admits corrected functions contains those models with log likelihoods of the form

$$\log[f(y|z, x, \Theta)] = \sum_{k=0}^{2} [c_k(y, z, \Theta)(\beta_x^t x)^k] + c_3(y, z, \Theta)\exp(\beta_x^t x);$$

see the examples given below. Then, using normal distribution **moment generating function** identities, the required function is

$$\psi_*(y, z, w, \Theta, \sigma_u^2) = \frac{\partial}{\partial \Theta^t}\left[\sum_{k=0}^{2}[c_k(y, z, \Theta)(\beta_x^t w)^k] - c_2(y, z, \Theta) \times \beta_x^t \sigma_u^2 \beta_x + c_3(y, z, \Theta)\exp(\beta_x^t w - 0.5\beta_x^t \sigma_u^2 \beta_x)\right].$$

Regression models in this class include:

1. normal linear with mean $= \eta$, variance $= \phi$, $c_0 = -(y - \beta_0 - \beta_z^t z)^2/(2\phi) - \log(\phi^{1/2})$, $c_1 = (y - \beta_0 - \beta_z^t z)/\phi$, $c_2 = -(2\phi)^{-1}$, $c_3 = 0$
2. Poisson with mean $= \exp(\eta)$, variance $= \exp(\eta)$, $c_0 = y(\beta_0 + \beta_z^t z) - \log y!$, $c_1 = y$, $c_2 = 0$, $c_3 = -\exp(\beta_0 + \beta_z^t z)$
3. gamma with mean $= \exp(\eta)$, variance $= \phi \exp(2\eta)$, $c_0 = -\phi^{-1}(\beta_0 + \beta_z^t z) + (\phi^{-1} - 1)\log y + \phi^{-1}\log(\phi^{-1}) - \log[\Gamma(\phi^{-1})]$, $c_1 = \phi^{-1}$, $c_2 = 0$, $c_3 = -\phi^{-1} y \exp(-\beta_0 - \beta_z^t z)$.

Comparison of Methods

The methods are applicable at the same time only in linear regression (where they are identical) and Poisson regression. For Poisson regression the corrected estimating equations are more convenient because they are explicit, whereas the conditional estimator involves numerical summation. For Poisson regression the conditional-score estimator is more efficient than the corrected-score estimator in some practical cases.

Instrumental Variables

We have assumed that it was possible to estimate the measurement error variance, say with replicate measurements or validation data. However, it is not always possible to obtain replicates or validation data, and thus direct estimation of the measurement error variance is sometimes impossible. In the absence of information about the measurement error variance, estimation of the regression model parameters is still possible provided the data contain an *instrumental variable* T, in addition to the unbiased measurement $W = X + U$.

There are three basic requirements that an **instrumental variable** must satisfy: (i) it must be correlated with X; (ii) it must be independent of $W - X$; and (iii) it must be a surrogate, i.e. subject to nondifferential measurement error.

One possible source of an instrumental variable is a second measurement of X obtained by an independent method. This second measurement need not be unbiased for X. Thus the assumption that a variable is an instrument is weaker than the assumption that it follows the classical additive error model.

Instrumental variable estimation in linear models is covered in depth by Fuller [30]. The work described here, outside the linear model, is based on that of Carroll & Stefanski [17] and Stefanski & Buzas [69]. Other pertinent references include [1], [2], and [13].

We have found that instrumental variables require a slightly different notation. For example, $\beta_{Y|\mathbf{1}ZX}$ is the coefficient of $\mathbf{1}$, i.e. the intercept, in the regression of Y on $\mathbf{1}$, Z, and X; $\beta_{Y|\mathbf{1}Z\mathbf{X}}$ is the coefficient of Z in the regression of Y on $\mathbf{1}$, Z, and X. This notation allows representation of subsets of coefficient vectors, e.g. $\boldsymbol{\beta}_{Y|\mathbf{1}Z\mathbf{X}} = (\beta_{Y|\mathbf{1}ZX}, \beta_{Y|\mathbf{1}Z\mathbf{X}})$ and $\boldsymbol{\beta}_{X|\mathbf{1}Z\mathbf{T}} = (\beta_{X|\mathbf{1}ZT}, \beta_{X|\mathbf{1}Z\mathbf{T}}, \beta_{X|\mathbf{1}Z\mathbf{T}})$.

Our analysis is based upon regression calibration in generalized linear models, e.g. linear, logistic, and Poisson regression. It might be useful simply to think of this Section as dealing with a class of important models, whose details of fitting are standard in many computer programs.

The approximate models and estimation algorithms are best described in terms of the composite vectors

$$\mathbf{X} = (\mathbf{1}, Z, X), \qquad \mathbf{W} = (\mathbf{1}, Z, W),$$
$$\mathbf{T} = (\mathbf{1}, Z, T).$$

Define $\boldsymbol{\beta}_{Y|\tilde{\mathbf{X}}} = (\beta_{Y|\mathbf{1}ZX}, \beta_{Y|\mathbf{1}Z\mathbf{X}}, \beta_{Y|\mathbf{1}Z\mathbf{X}})$.

We note here that, in addition to the assumptions stated previously, we will also assume that the regression of X on (Z, T, W) is approximately linear. This restricts the applicability of our methods somewhat, but is sufficiently general to encompass many potential applications.

The simplest instrumental variables estimator starts with a (possibly multivariate) regression of $\mathbf{W}$ on $\mathbf{T}$ to obtain $\hat{\beta}_{\mathbf{W}|\mathbf{T}}$. Then Y is regressed on the predicted values $\hat{\beta}_{\mathbf{W}|\mathbf{T}}\mathbf{T}$, which results in an estimator of $\boldsymbol{\beta}_{Y|\mathbf{X}}$.

This estimator is easily computed as it requires only linear regression of the components of **W** on **T**, and then the use of standard regression programs to regress Y on the "predictors" $\hat{\beta}_{\mathbf{W}|\mathbf{T}}\mathbf{T}$.

Carroll et al. [20] describe somewhat more elaborate methods of instrumental variable estimation, which can be more efficient than this simple method, especially if the number of components of T differs from the number of components of W.

This Section describes the use of likelihood methods in measurement error models. There have been a few examples in the literature based on likelihood. See [23], [63], [64], and [78] for probit regression, [84] for a Poisson model, [27], [62] and [81] in logistic regression, and [38] in a **change-point problem**. The relatively small literature belies the importance of the topic and the potential for further applications.

There are a number of important differences between likelihood methods and the methods described in previous Sections:

1. The previous methods are based on additive or multiplicative measurement error models, possibly after a transformation. Typically, few, if any, distributional assumptions are required. Likelihood methods require stronger distributional assumptions, but they can be applied to more general problems, including those with discrete covariates subject to **misclassification error**.
2. The likelihood for a fully specified parametric model can be used to obtain **likelihood ratio** confidence intervals. In methods not based on likelihoods, inference is based on bootstrapping or on normal approximations. In highly nonlinear problems, likelihood-based confidence intervals are generally more reliable than those derived from normal approximations.
3. Likelihood methods are often computationally more demanding, whereas the previous methods require little more than the use of standard statistical packages.
4. **Robustness** to modeling assumptions is a concern for both approaches, but is generally more difficult to understand with likelihood methods.
5. There is a belief that the simpler methods described previously perform just as well as likelihood methods for many statistical models, including the most common generalized linear models. There is little documentation as to whether the folklore is realistic. The only evidence that we know of is given for logistic regression by Stefanski & Carroll [72], who contrast the maximum likelihood estimate and a particular functional estimate. They find that the functional estimate is fairly efficient relative to the maximum likelihood estimate unless the measurement error is "large" or the logistic coefficient is "large". One should be aware, however, that their calculations indicate that there are situations where *properly parameterized* maximum likelihood estimates are considerably more efficient than estimates derived from functional modeling.

Likelihood Specification: Differential and Nondifferential Error

We consider here only the simplest problem in which X is not observable for all subjects, but there are sufficient data, either internal or external, to characterize the distribution of W given (X, Z) (with validation data, we are in the realm of missing data). To perform a likelihood analysis, one must specify a parametric model for every component of the data. Likelihood analysis starts with a model for the distribution of the response given the true predictors. The likelihood (density or mass) function of Y given (Z, X) will be called $f_{Y|Z,X}(y|z, x, \mathcal{B})$ here, and interest lies in estimating $\mathcal{B}$. For example, if Y is normally distributed with mean $\beta_0 + \beta_x X + \beta_z Z$ and variance σ^2, then $\mathcal{B} = (\beta_0, \beta_x, \beta_z, \sigma^2)$ and

$$f_{Y|Z,X}(y|z, x, \mathcal{B}) = \sigma^{-1}\phi[(y - \beta_0 + \beta_x x + \beta_z z)/\sigma],$$

where $\phi(v) = (2\pi)^{-1/2}\exp(-0.5v^2)$ is the standard normal density function. If Y follows a logistic regression model with mean $H(\beta_0 + \beta_x X + \beta_z Z)$, then $\mathcal{B} = (\beta_0, \beta_x, \beta_z)$ and

$$f_{Y|Z,X}(y|z, x, \mathcal{B}) = H^y(\beta_0 + \beta_x x + \beta_z z) \times [1 - H(\beta_0 + \beta_x x + \beta_z z)]^{1-y}.$$

A likelihood analysis starts with determination of the joint distribution of Y and W given Z, as these are the observed variates. There are three components required:

1. A model relating the response to the "true" covariates, see just above.

2. An error model, here called $f_{W|Z,X}(w|z, x, \tilde{\alpha}_1)$. In many applications, the error model does not depend on Z. For example, in the classical additive measurement error model (1) with normally distributed measurement error, σ_u^2 is the only component of $\tilde{\alpha}_1$, and the error model density is $\sigma_u^{-1}\phi[(w - x)/\sigma_u]$, where $\phi(\cdot)$ is the standard normal density function. In the classical error model with independent replicates, W consists of the k replicates, and $f_{W|Z,X}$ is the k-variate normal density function with mean zero, common variance σ_u^2, and zero correlation. A generalization of this error model that allows for correlations among the replicates has been studied [81]. In some application areas, error model structures are studied independently of their role in measurement error modeling, and one can use this research to estimate error models for the problem at hand.
3. A model for the distribution of the latent variable, here called $f_{X|Z}(x|z, \tilde{\alpha}_2)$. Specifying a model for the distribution of the true covariate X given all the other covariates, Z is more difficult. Difficulties arise because: (i) the distribution is usually not transportable, so that different studies yield very different models; and (ii) X is not observed.

Having hypothesized the various models, the likelihood of the observed data under nondifferential measurement error is

$$\begin{aligned} f_{Y,W|Z}(y, w|z, \mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2) &= \int f_{Y|Z,X}(y|z, x, \mathcal{B}) f_{W|Z,X}(w|z, x, \tilde{\alpha}_1) \\ &\quad \times f_{X|Z}(x|z, \tilde{\alpha}_2)\, \mathrm{d}\mu(x). \end{aligned} \tag{13}$$

The notation $\mathrm{d}\mu(x)$ indicates that the integrals are sums if X is discrete and integrals if X is continuous. The likelihood for the problem is just the product over the sample of these terms.

There is a significant difference between the likelihood function in the differential and nondifferential cases. This can be expressed in various ways, but the simplest is as follows. In general, and dropping parameters, the likelihood of the observed data is

$$f_{Y,W|Z}(y, w|z) = \int f_{Y,W,X|Z}(y, w, x|z)\, \mathrm{d}\mu(x).$$

Using standard conditioning arguments, this becomes

$$\begin{aligned} f_{Y,W|Z}(y, w|z) &= \int f_{W|Y,Z,X}(w|y, z, x) f_{Y|Z,X}(y|z, x) \\ &\quad \times f_{X|Z}(x|z)\, \mathrm{d}\mu(x) \\ &= \int f_{Y|Z,X}(y|z, x) f_{W|Y,Z,X}(w|y, z, x) \\ &\quad \times f_{X|Z}(x|z)\, \mathrm{d}\mu(x). \end{aligned} \tag{14}$$

Note that the only difference between (13) and (14) is in the error term. In the former, under nondifferential measurement error, W and Y are independent, so that $f_{W|Y,Z,X}(w|y, z, x) = f_{W|Z,X}(w|z, x)$.

What makes differential error so difficult is that, under differential measurement error, we must ascertain the distribution of W given the other covariates *and the response* Y. This is essentially impossible to do in practice unless one has a subset of the data in which all of (Y, Z, X, W) are observed, i.e. a *validation* data set (*see* **Validation Study**).

Numerical Computation of Likelihoods

Typically one maximizes the logarithm of the overall likelihood in the unknown parameters. There are two ways one can maximize the likelihood function. The most direct is to compute the likelihood function itself, and then use numerical optimization techniques to maximize the likelihood. Below we provide a few details about computing the likelihood function. The second general approach is to view the problem as a missing data problem, and then use missing data techniques (*see* **Missing Data**); see, for example, [42] and [75].

Computing the likelihood analytically is easy if X is discrete, as the conditional expectations are simply sums of terms. Likelihoods in which X has some continuous components can be computed using a number of different approaches. In some problems the log likelihood can be computed or very well approximated analytically. In most problems that we have encountered, X is a scalar or a 2×1 vector. In these cases, standard numerical methods such as Gaussian quadrature can be applied, although they are not always very good. When sufficient computing resources are available, the likelihood can be computed using **Monte Carlo** techniques.

Bayesian Methods

Bayesian estimation and inference in the measurement error problem is a promising approach under active development (*see* **Bayesian Methods**). Examples of this approach are given by Schmid &

Rosner [65], Richardson & Gilks [57], Stephens & Dellaportas [74], Müller & Roeder [47], Mallick & Gelfand [45], and Kuha [40].

Bayesian analysis of parametric models requires specifying a likelihood (as described above) and a **prior distribution** for the parameters, the latter representing knowledge about the parameters prior to data collection. The product of the prior and likelihood is the joint density of the data and the parameters. Using **Bayes' Theorem**, one can in principle obtain the posterior density, i.e. the conditional density of the parameters given the data. The posterior summarizes all of the information about the values of the parameters and is the basis for all Bayesian inference. For example, the mean, median, or mode of the posterior density are all suitable point estimators. A region with probability $1 - \alpha$ under the posterior is called a "credible set", and is a Bayesian analog to a confidence region.

Computing the posterior distribution is often a nontrivial problem, because it usually requires high-dimensional numerical integration. This computational problem is the subject of much recent research, with many major advances. The method currently receiving the most attention in the literature is the Gibbs sampler (see [66] and [24]; *see* **Markov Chain Monte Carlo**). Also, see Tanner [75] for a book-length introduction to modern methods for computing posterior distributions.

In the Bayesian approach with Gibbs sampling, the Xs are treated as "missing data" (they just happen to be missing for all study subjects unless there is a validation study!). The approach for the classical additive error model is:

1. Assuming nondifferential error, write the likelihood of Y given (X, Z), the likelihood of W given (X, Z), and the likelihood of X given Z depending on parameters, just as in a regular likelihood problem.
2. If X were observable, then the likelihood would be the product of the three terms given above.
3. Select a starting value for the parameters, e.g. from SIMEX.
4. Use a simulation approach to fill in the "missing" Xs, i.e. from the posterior distribution of X given the observed data and the current values of the parameters. In this step, it is rare that the posterior distribution is known exactly, and so one has to use a device such as the Metropolis–Hastings algorithm.
5. Now one has complete data, with Xs all filled in, and one uses simulation to draw a sample of parameters from the posterior distribution of the parameters given the observed data and the current Xs.
6. Repeat the process of generating X and the parameters. These multiple samples of parameters are used to evaluate features of the posterior distribution.

While the procedure is easy to write down, the computations may be difficult.

More importantly, though, is the need to consider the distribution of X given Z. As we emphasized above, the simplest structural approach assumes that X is normally distributed, but this is often a strong assumption. The popularity of functional methods lies in the fact that such methods require no distributional assumptions about the Xs. There is considerable current effort being made to circumvent the problem of model robustness by specifying a flexible distribution for X.

Mixture Modeling

When there are no covariates measured without error, the nonlinear measurement error problem can be viewed as a special case of what are called mixture problems (see [77]). The idea is to pretend that X has a distribution, but to estimate this distribution nonparametrically. Applications of nonparametric mixture methods to nonlinear measurement error models have only recently been described by Thomas et al. [76] and Roeder et al. [58].

An alternative formulation is to let X have a flexible distribution, which covers a wide range of possibilities including the normal distribution. The simplest such model is the mixture of normals, which has been applied by Wang et al. [81] and by Küchenhoff & Carroll [38].

Response Error

In preceding Sections we have focused exclusively on problems associated with measurement error in predictor variables. Here we consider problems that arise when a true response is measured with error. For example, in a study of factors affecting dietary intake of fat, e.g. sex, race, age, socioeconomic status,

etc., true long-term dietary intake is impossible to determine and instead it is necessary to use error-prone measures of long-term dietary intake. Wittes et al. [86] describe another example in which damage to the heart muscle caused by a myocardial infarction can be assessed accurately, but the procedure is expensive and invasive, and instead it is common practice to use the peak cardiac enzyme level in the bloodstream as a proxy for the true response.

For a binary response (case or control), see **Misclassification Error**.

The exclusive attention paid to predictor measurement error earlier in this article is explained by the fact that predictor measurement error is seldom ignorable, by which is meant that the usual method of analysis is statistically valid, whereas response measurement error is often ignorable when the response is continuous. Here, "ignorable" means that the model holding for the true response holds also for the proxy response with parameters unchanged, except that a measurement error variance component is added to the response variance. For example, in linear regression models with simple types of response measurement error, the response measurement error is **confounded** with equation error and the effect is simply to increase the variability of parameter estimates. Thus, response error is ignorable in these cases, although of course power will be lost. However, in more complicated regression models, certain types of response error are not ignorable and it is important to account for the response error explicitly in the regression analysis.

Although the details differ between methods for predictor error and response error, many of the basic ideas are similar. Throughout this section, the response proxy is denoted by S. We consider only the case of measurement error in the response, and not the more complex problem where both the response and some of the predictors are measured with error.

We first consider the analysis of the observed data when the response is subject to independent additive or multiplicative measurement error. Suppose that the proxy response S is unbiased for the true response. Then, in either case, the proxy response has the same mean (as a function of exposure and confounders) as the true response, although the variance structure differs. In models such as linear regression, or more generally for **quasi-likelihood** estimation, this means that the parameter estimates are consistent, but inferences may be affected. For example, in linear regression, additive, unbiased response error does not change the mean and simply increases the variance by a constant, so that there is no effect of measurement error other than loss of power. However, for multiplicative, unbiased response error, while the mean remains unchanged, the variances now are no longer constant, and hence inferences which pretend that the variances are constant would be affected. The usual solution is to use a robust covariance estimator, also known as the sandwich estimator (*see* **Generalized Estimating Equations**).

If the proxy response **S** is not unbiased for the true response, then a validation study is required to understand the nature of the bias and to correct for it. In a series of papers, Buonaccorsi [8, 9, 11] and Buonaccorsi & Tosteson [12] discuss the use of adjustments for a biased response. See Carroll et al. [20] for further details.

We call **S** a *surrogate response* if its distribution depends only on the true response and not otherwise on the covariates, i.e. the information about the surrogate response contained in the true response is the same no matter what the values of the covariates. In symbols, if $f_{S|Y,Z,X}(s|y,z,x,\gamma)$ denotes the density or mass function for S given (Y,Z,X), then $f_{S|Y,Z,X}(s|y,z,x,\gamma) = f_{S|Y}(s|y,\gamma)$. In both the additive and multiplicative error models, S is a surrogate. This definition of a surrogate response is the natural counterpart to a surrogate predictor, because it implies that all the information in the relationship between S and the predictors is explained by the underlying response. See Prentice [53] and Carroll et al. [20] for further details.

In general, i.e. for a possibly nonsurrogate response, the likelihood function for the observed response is

$$f_{S|Z,X}(s|z,x,\mathcal{B},\gamma) = \int f_{Y|Z,X}(y|z,x,\mathcal{B}) f_{S|Y,Z,X}(s|y,z,x,\gamma)\,\mathrm{d}\mu(y). \tag{15}$$

There are a number of implications of this formula:

1. If S is a surrogate, and if there is no relationship between the true response and the predictors, then neither is there one between the observed response and the predictors. Hence, if interest lies in determining whether *any of the predictors* contains any information about the response, then one can use naive hypothesis tests and

ignore response error. The resulting tests have an asymptotically correct level, but a decreased power relative to tests derived from true response data. This property of a surrogate is important in clinical trials; see Prentice [53] (*see* **Surrogate Endpoints**).

2. If S is *not* a surrogate, then there may be no relationship between the true response and the covariates, but the observed response may be related to the predictors. Hence, naive tests will not be valid in general if **S** is not a surrogate.

Note that one implication of (15) is that a likelihood analysis with mismeasured responses requires a model for the distribution of response error. Except for additive and multiplicative error, understanding such a model requires a validation study.

Case–Control Studies

A *case–control study* is one in which sampling is conditioned on the disease response; it is useful to think that the response is first observed and only later are the predictors observed. A similar design, *choice-based sampling*, is used in econometrics. We use case–control terminology and concentrate on logistic regression models. A distinguishing feature of case–control studies is that the measurement error may be differential.

Two-phase case–control designs, where X is observed on a subset of the data, have been studied by Breslow & Cain [7], Zhao & Lipsitz [87], Tosteson & Ware [80], and Carroll et al. [18], among others. These designs are significant because the validation, if done on both cases and controls, frees us from the nondifferential error assumption.

We assume that the data follow a logistic model in the underlying source population, although the results apply equally well to the more general models described by Weinberg & Wacholder [82]. For such models, Prentice & Pyke [55] and Weinberg & Wacholder [82] show that when analyzing a classical case–control study one can ignore the case–control sampling scheme entirely, at least for the purpose of estimating **relative risk**. Furthermore, these authors show that, if one *ignores the case–control sampling scheme and runs an ordinary logistic regression*, then the resulting relative risk estimates are consistent and the standard errors are asymptotically correct.

The effect of measurement error in logistic case–control studies is to bias the estimates. Carroll et al. [21] show that, for many problems, one can ignore the case–control study design and proceed to correct for the bias from measurement error as if one were analyzing a random sample from the source population. With nondifferential measurement error, this result applies to the methods we have described previously for prospective studies. Regression calibration needs a slight modification, namely that the regression calibration function should be estimated using the controls only.

Michalek & Tripathi [46], Armstrong et al. [5], and Buonaccorsi [10] consider the normal **discriminant** model. Satten & Kupper [62] have an interesting example of likelihood analysis for nondifferential error validation studies when the validation sampling is in the controls.

Survival Analysis

One of the earliest applications of the regression calibration method was discussed by Prentice [52] in the context of **survival analysis**. Further results in survival analysis were obtained by Pepe et al. [50], Clayton [25], Nakamura [49], and Hughes [36]. While the details differ in substantive ways, the ideas are the same as put forward in the rest of this article, and here we provide only a very brief overview in the case of covariates which do not depend on time.

Suppose that the instantaneous risk that the time T of an event equals t conditional on no events prior to time t and conditional on the true covariate X is denoted by

$$\psi(t, X) = \psi_0(t) \exp(\beta_x X), \tag{16}$$

where $\psi_0(t)$ is the baseline **hazard** function. When the baseline hazard is not specified, (16) is commonly called the **proportional hazards** assumption. When X is observable, it is well known that estimation of β_x is possible without specifying the form of the baseline hazard function.

If X is unobservable and instead we observe a surrogate W, then the induced hazard function is

$$\psi^*(t, W, \beta_x) = \psi_0(t)\mathrm{E}[\exp(\beta_x X)|T \geq t, W]. \tag{17}$$

The difficulty is that the expectation in (17) for the observed data depends upon the unknown baseline

hazard function ψ_0. Thus, the hazard function does not factor into a product of an arbitrary baseline hazard times a term that depends only on observed data and an unknown parameter, and the technology for proportional hazards regression cannot be applied without modification.

The problem simplifies when the event is rare, so that $T \geq t$ occurs with high probability for all t under consideration. As shown by Prentice [53] and others, under certain circumstances this leads to the regression calibration algorithm. The rare event assumption allows the hazard of the observed data to be approximated by

$$\psi^*(t, W, \beta_x) = \psi_0(t)\mathrm{E}[\exp(\beta_x X)|W]. \qquad (18)$$

The hazard function (18) requires a regression calibration formulation! If one specifies a model for the distribution of X given W, then (18) is in the form of a proportional hazards model (16), but with $\beta_x X$ replaced by $\log\{\mathrm{E}[\exp(\beta_x X)|W]\}$. An important special case leads directly to the standard regression calibration model, namely when X given W is normally distributed.

Clayton [25] proposed a modification of regression calibration which does not require events to be rare. At each time t_i, $i = 1, \ldots, k$, for which an event occurs, define the risk set $R_i \subseteq \{1, \ldots, n\}$ as the case numbers of those members of the study cohort for whom an event has not occurred and who were still under study just prior to t_i. If the Xs were observable, and if X_i is the covariate associated with the ith event, in the absence of ties the usual proportional hazards regression would maximize

$$\prod_{i=1}^{k} \frac{\exp(\beta_x X_i)}{\sum_{j \in R_i} \exp(\beta_x X_j)}.$$

Clayton basically suggests using regression calibration within each risk set. He assumes that the true values X within the ith risk set are normally distributed with mean μ_i and variance σ_x^2, and that within this risk set $W = X + U$, where U is normally distributed with mean zero and variance σ_u^2. Neither σ_x^2 nor σ_u^2 depend upon the risk set in his formulation.

Given an estimate $\hat{\sigma}_u^2$, one applies the usual regression calibration calculations to construct an estimate of $\hat{\sigma}_x^2$.

Clayton modifies regression calibration by using it within each risk set. Within each risk set, he applies the formula (7) for the best unbiased estimate of the Xs. Specifically, in the absence of replication, for any member of the ith risk set, the estimate of the true covariate X from an observed covariate W is

$$\hat{X} = \hat{\mu}_i + \frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_u^2}(W - \hat{\mu}_i),$$

where $\hat{\mu}_i$ is the sample mean of the Ws in the ith risk set.

As with regression calibration in general, the advantage of Clayton's method is that no new software need be developed, other than to calculate the means within risk sets.

Acknowledgment

This work was supported by a grant from the National Cancer Institute (CA-57030).

References

[1] Amemiya, Y. (1985). Instrumental variable estimator for the nonlinear errors in variables model, *Journal of Econometrics* **28**, 273–289.

[2] Amemiya, Y. (1990). Instrumental variable estimation of the nonlinear measurement error model, in *Statistical Analysis of Measurement Error Models and Application*, P.J. Brown & W.A. Fuller, eds. American Mathematics Society, Providence.

[3] Amemiya, Y. & Fuller, W.A. (1988). Estimation for the nonlinear functional relationship, *Annals of Statistics* **16**, 147–160.

[4] Armstrong, B. (1985). Measurement error in generalized linear models, *Communications in Statistics, Part B – Simulation and Computation* **14**, 529–544.

[5] Armstrong, B.G., Whittemore, A.S. & Howe, G.R. (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer, *Statistics in Medicine* **8**, 1151–1163.

[6] Berkson, J. (1950). Are there two regressions?, *Journal of the American Statistical Association* **45**, 164–180.

[7] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case-control data, *Biometrika* **75**, 11–20.

[8] Buonaccorsi, J.P. (1988). Errors in variables with systematic biases, *Communications in Statistics – Theory and Methods* **18**, 1001–1021.

[9] Buonaccorsi, J.P. (1990). Double sampling for exact values in some multivariate measurement error problems, *Journal of the American Statistical Association* **85**, 1075–1082.

[10] Buonaccorsi, J.P. (1990). Double sampling for exact values in the normal discriminant model with application

to binary regression, *Communications in Statistics – Theory and Methods* **19**, 4569–4586.

[11] Buonaccorsi, J.P. (1991). Measurement error, linear calibration and inferences for means, *Computational Statistics and Data Analysis* **11**, 239–257.

[12] Buonaccorsi, J.P. & Tosteson, T. (1993). Correcting for nonlinear measurement error in the dependent variable in the general linear model, *Communications in Statistics – Theory and Methods* **22**, 2687–2702.

[13] Buzas, J.S. & Stefanski, L.A. (1995). A note on corrected score estimation, *Statistics and Probability Letters* **28**, 1–8.

[14] Carroll, R.J. & Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, London.

[15] Carroll, R.J. & Ruppert, D. (1996). The use and misuse of orthogonal regression in measurement error models, *American Statistician* **50**, 1–6.

[16] Carroll, R.J. & Stefanski, L.A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors, *Journal of the American Statistical Association* **85**, 652–663.

[17] Carroll, R.J. & Stefanski, L.A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses, *Statistics in Medicine* **13**, 1265–1282.

[18] Carroll, R.J., Gail, M.H. & Lubin, J.H. (1993). Case-control studies with errors in predictors, *Journal of the American Statistical Association* **88**, 177–191.

[19] Carroll, R.J., Gallo, P.P. & Gleser, L.J. (1985). Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance, *Journal of the American Statistical Association* **80**, 929–932.

[20] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

[21] Carroll, R.J., Wang, S. & Wang, C.Y. (1995). Asymptotics for prospective analysis of stratified logistic case-control studies, *Journal of the American Statistical Association* **90**, 157–169.

[22] Carroll, R.J., Küchenhoff, H., Lombard, F. & Stefanski, L.A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models, *Journal of the American Statistical Association* **91**, 242–250.

[23] Carroll, R.J., Spiegelman, C., Lan, K.K., Bailey, K.T. & Abbott, R.D. (1984). On errors-in-variables for binary regression models, *Biometrika* **71**, 19–26.

[24] Casella, G. & George, E.I. (1992). Explaining the Gibbs sampler, *American Statistician* **46**, 167–174.

[25] Clayton, D.G. (1991). Models for the analysis of cohort and case-control studies with inaccurately measured exposures, in *Statistical Models for Longitudinal Studies of Health*, J.H. Dwyer, M. Feinleib, P. Lipsert et al., eds. Oxford University Press, New York, pp. 301–331.

[26] Cook, J. & Stefanski, L.A. (1995). A simulation extrapolation method for parametric measurement error models, *Journal of the American Statistical Association* **89**, 1314–1328.

[27] Crouch, E.A. & Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{\infty}^{\infty} f(t) \exp(-t^2)dt$: applications to logistic-normal models, *Journal of the American Statistical Association* **85**, 464–467.

[28] Dosemeci, M., Wacholder, S. & Lubin, J.H. (1990). Does non-differential misclassification of exposure always bias a true effect towards the null value?, *American Journal of Epidemiology* **132**, 746–748.

[29] Freedman, L.S., Carroll, R.J. & Wax, Y. (1991). Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake, *American Journal of Epidemiology* **134**, 510–520.

[30] Fuller, W.A. (1987). *Measurement Error Models*. Wiley, New York.

[31] Ganse, R.A., Amemiya, Y. & Fuller, W.A. (1983). Prediction when both variables are subject to error, with application to earthquake magnitude, *Journal of the American Statistical Association* **78**, 761–765.

[32] Gleser, L.J. (1981). Estimation in multivariate errors in variables regression model: large sample results, *Annals of Statistics* **9**, 24–44.

[33] Gleser, L.J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models, in *Statistical Analysis of Measurement Error Models and Application*, P.J. Brown & W.A. Fuller, eds. American Mathematical Society, Providence.

[34] Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564–569.

[35] Greenland, S. & Robins, J.M. (1985). Confounding and misclassification, *American Journal of Epidemiology* **122**, 495–506.

[36] Hughes, M.D. (1993). Regression dilution in the proportional hazards model, *Biometrics* **49**, 1056–1066.

[37] Hunter, D.J., Spiegelman, D., Adami, H.-O., Beeson, L., van der Brandt, P.A., Folsom, A.R., Fraser, G.E., Goldbohm, A., Graham, S., Howe, G.R., Kushi, L.H., Marshall, J.R., McDermott, A., Miller, A.B., Speizer, F.E., Wolk, A., Yaun, S.S. & Willett, W. (1996). Cohort studies of fat intake and the risk of breast cancer-a pooled analysis, *New England Journal of Medicine* **334**, 356–361.

[38] Küchenhoff, H. & Carroll, R.J. (1997). Segmented regression with errors in predictors: semiparametric and parametric methods, *Statistics in Medicine* **16**, 169–188.

[39] Kuha, J. (1994). Corrections for exposure measurement error in logistic regression models with an application to nutritional data, *Statistics in Medicine* **13**, 1135–1148.

[40] Kuha, J. (1997). Estimation by data augmentation in regression models with continuous and discrete covariates measured with error, *Statistics in Medicine* **16**, 189–201.

[41] Li, L., Freedman, L., Kipnis, V. & Carroll, R.J. (1997). Effects of bias and correlated measurement errors in the validation of food frequency questionnaires. Preprint.

[42] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[43] Liu, X. & Liang, K.Y. (1992). Efficacy of repeated measures in regression models with measurement error, *Biometrics* **48**, 645–654.

[44] MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A. & Stamler, J. (1990). Blood pressure, stroke and coronary heart disease: Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias, *Lancet* **335**, 765–774.

[45] Mallick, B.K. & Gelfand, A.E. (1996). Semiparametric errors-in-variables models: a Bayesian approach, *Journal of Statistical Planning and Inference* **52**, 307–322.

[46] Michalek, J.E. & Tripathi, R.C. (1980). The effect of errors in diagnosis and measurement on the probability of an event, *Journal of the American Statistical Association* **75**, 713–721.

[47] Müller, P. & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables, *Biometrika* **84**, 523–537.

[48] Nakamura, T. (1990). Corrected score functions for errors-in-variables models: methodology and application to generalized linear models, *Biometrika* **77**, 127–137.

[49] Nakamura, T. (1992). Proportional hazards models with covariates subject to measurement error, *Biometrics* **48**, 829–838.

[50] Pepe, M.S., Self, S.G. & Prentice, R.L. (1989). Further results in covariate measurement errors in cohort studies with time to response data, *Statistics in Medicine* **8**, 1167–1178.

[51] Pierce, D.A., Stram, D.O., Vaeth, M. & Schafer, D. (1992). Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors, *Journal of the American Statistical Association* **87**, 351–359.

[52] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[53] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.

[54] Prentice, R.L. (1996). Dietary fat and breast cancer: measurement error and results from analytic epidemiology, *Journal of the National Cancer Institute* **88**, 1738–1747.

[55] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika* **66**, 403–411.

[56] Racine-Poon, A., Weihs, C. & Smith, A.F.M. (1991). Estimation of relative potency with sequential dilution errors in radioimmunoassay, *Biometrics* **47**, 1235–1246.

[57] Richardson, S. & Gilks, W.R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models, *American Journal of Epidemiology* **138**, 430–442.

[58] Roeder, K., Carroll, R.J. & Lindsay, B.G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables, *Journal of the American Statistical Association* **91**, 722–732.

[59] Rosner, B., Spiegelman, D. & Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error, *American Journal of Epidemiology* **132**, 734–745.

[60] Rosner, B., Willett, W.C. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error, *Statistics in Medicine* **8**, 1051–1070.

[61] Rudemo, M., Ruppert, D. & Streibig, J.C. (1989). Random effect models in nonlinear regression with applications to bioassay, *Biometrics* **45**, 349–362.

[62] Satten, G.A. & Kupper, L.L. (1993). Inferences about exposure-disease association using probability of exposure information, *Journal of the American Statistical Association* **88**, 200–208.

[63] Schafer, D. (1987). Covariate measurement error in generalized linear models, *Biometrika* **74**, 385–391.

[64] Schafer, D. (1993). Likelihood analysis for probit regression with measurement errors, *Biometrika* **80**, 899–904.

[65] Schmid, C.H. & Rosner, B. (1993). A Bayesian approach to logistic regression models having measurement error following a mixture distribution, *Statistics in Medicine* **12**, 1141–1153.

[66] Smith, A.F.M. & Gelfand, A.E. (1992). Bayesian statistics without tears: a sampling-resampling perspective, *American Statistician* **46**, 84–88.

[67] Stefanski, L.A. (1985). The effects of measurement error on parameter estimation, *Biometrika* **72**, 583–592.

[68] Stefanski, L.A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models, *Communications in Statistics – Theory and Methods* **18**, 4335–4358.

[69] Stefanski, L.A. & Buzas, J.S. (1995). Instrumental variable estimation in binary regression measurement error models, *Journal of the American Statistical Association* **90**, 541–549.

[70] Stefanski, L.A. & Carroll, R.J. (1985). Covariate measurement error in logistic regression, *Annals of Statistics* **13**, 1335–1351.

[71] Stefanski, L.A. & Carroll, R.J. (1987). Conditional scores and optimal scores in generalized linear measurement error models, *Biometrika* **74**, 703–716.

[72] Stefanski, L.A. & Carroll, R.J. (1990). Structural logistic regression measurement error models, in *Proceedings of the Conference on Measurement Error Models*, P.J. Brown & W.A. Fuller, eds. Wiley, New York.

[73] Stefanski, L.A. & Cook, J. (1995). Simulation extrapolation: the measurement error jackknife, *Journal of the American Statistical Association* **90**, 1247–1256.

[74] Stephens, D.A. & Dellaportas, P. (1992). Bayesian analysis of generalized linear models with covariate measurement error, in *Bayesian Statistics 4*, J.M. Bernado, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 813–820.

[75] Tanner, M.A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions*

and Likelihood Functions, 2nd Ed. Springer-Verlag, New York.

[76] Thomas, D., Stram, D. & Dwyer, J. (1993). Exposure measurement error: influence on exposure-disease relationships and methods of correction, *Annual Review of Public Health* **14**, 69–93.

[77] Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

[78] Tosteson, T., Stefanski, L.A. & Schafer D.W. (1989). A measurement error model for binary and ordinal regression, *Statistics in Medicine* **8**, 1139–1147.

[79] Tosteson, T. & Tsiatis, A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates, *Biometrika* **75**, 507–514.

[80] Tosteson, T.D. & Ware, J.H. (1990). Designing a logistic regression study using surrogate measures of exposure and outcome, *Biometrika* **77**, 11–20.

[81] Wang, N., Carroll, R.J. & Liang, K.Y. (1996). Quasi-likelihood and variance functions in measurement error models with replicates, *Biometrics* **52**, 401–411.

[82] Weinberg, C.R. & Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept models, *Biometrika* **80**, 461–465.

[83] Whittemore, A.S. (1989). Errors in variables regression using Stein estimates, *American Statistician* **43**, 226–228.

[84] Whittemore, A.S. & Gong, G. (1991). Poisson regression with misclassified counts: application to cervical cancer mortality rates, *Applied Statistics* **40**, 81–93.

[85] Whittemore, A.S. & Keller, J.B. (1988). Approximations for regression with covariate measurement error, *Journal of the American Statistical Association* **83**, 1057–1066.

[86] Wittes, J., Lakatos, E. & Probstfield, J. (1989). Surrogate endpoints in clinical trials: cardiovascular trials, *Statistics in Medicine* **8**, 415–425.

[87] Zhao, L.P. & Lipsitz, S. (1992). Designs and analysis of two-stage studies, *Statistics in Medicine* **11**, 769–782.

RAYMOND J. CARROLL

Measurement Error in Survival Analysis

Introduction

Let $T \geq 0$ be a failure time variate (*see* **Survival Distributions and Their Characteristics**) and $Z = (Z_1', Z_2')'$, a corresponding **covariate** vector. Suppose first that the **hazard rate** for T given Z follows a **Cox regression** [9] model $\lambda(t; Z) = \lambda_0(t) \exp(Z'\beta)$, where λ_0 is an unspecified baseline hazard function and $\beta = (\beta_1', \beta_2')'$ is a corresponding parameter to be estimated. Failure times may be subject to right **censoring** by a variate C that is assumed to be independent of T given Z, so that one observes $X = T \wedge C$ and $\delta = I[X = T]$. On the basis of an independent random sample $(X_i, \delta_i, Z_i), i = 1, \ldots, n$ the standard "**partial likelihood**" estimator $\hat{\beta}$ solves $U(\hat{\beta}) = 0$ where

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \{Z_i - \overline{Z}(\beta, t)\} \, \mathrm{d}N_i(t). \qquad (1)$$

In (1) $\mathrm{d}N_i(t) = 1$ if $(X_i = t, \delta_i = 1)$ and is zero otherwise, the covariate "average" $\overline{Z}$ is given by $\overline{Z}(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$ where

$$S^{(j)}(\beta, t) = \sum_{i=1}^{n} Y_i(t) Z_i^j \exp(Z_i'\beta), \text{ for } j = 0, 1 \qquad (2)$$

and the "at risk" process Y_i is given by $Y_i(t) = I(X_i \geq t)$; see, for example, [3, 4, 13] for development of the asymptotic distribution theory for $\hat{\beta}$.

Suppose now that the component Z_1 of the regression vector is unavailable for some or all of the study population, whereas Z_2 and an error prone estimate W of Z_1 is routinely available. How then can the **relative risk** (hazard ratio) parameter β be estimated? This measurement error, or errors-in-variables problem arises in many application areas, in conjunction with failure time and other types of response variables. For example, in a **nutritional epidemiology** application, Z_1 may be comprised of an individual's long-term (e.g. 10 or 20 years) average daily intake of fat, along with dietary and nondietary **confounding** factors, while T is the time from entry into a **cohort study** until the diagnosis of a specific disease, such as coronary heart disease or colon cancer. The measured fat intake, a component of W in this context, may derive from self-reported food consumption over a short period of time (e.g. a few days or months) in which case the measured fat intake may differ from the theoretical quantity due to day-to-day or month-to-month variations in actual consumption, because of errors in dietary recording or recall, because of inaccuracies in the nutrient database used to estimate nutrient consumption from food consumption, or due to systematic self-report bias that may, for example, relate to "social desirability" characteristics.

If the variation in $W - Z_1$ is small compared to that for Z_1, one may simply be able to replace missing Z_1 values by corresponding W values in (1) and obtain estimates of β having little bias. See [12] for a study of bias in this context. Otherwise, a more careful approach is required. In the best of situations, a **validation** subsample can be obtained that includes (X, δ, Z, W), while only (X, δ, Z_2, W) is available on the remainder of the sample. Very commonly, however, only a **reliability** subsample consisting of repeat measures $W_1, W_2, \ldots$ of W can be obtained so that the data consist of $(X, \delta_1, Z_2, W_1, W_2, \ldots)$ on the reliability subsample and (X, δ, Z_2, W_1) on the remainder of the study cohort. It is often assumed that reliability sample measurement errors have mean zero and are independent of Z, each other, and other study subject characteristics, assumptions that are likely violated in the above nutritional epidemiology context.

While the focus of this entry is on estimation in the Cox model [9], there is also some work on estimation in an **additive hazards model** with covariate measurement error. This latter work will be described in the section "Measurement Error in the Additive Hazards Model".

Relative Risk Parameter Estimation with a Validation Subsample

Estimation of the Cox model regression parameter β in the presence of measurement error and a validation subsample is very closely connected to that of estimation with missing covariate data (*see* **Missing Data**). Specifically, W is said to be a **surrogate** for Z_1 if $\lambda(t; Z, W) = \lambda(t, Z) = \lambda_0(t) \exp(Z'\beta)$. Hence β may be estimated by Cox regression on (Z_1', Z_2', W') while regarding Z_1 to be sometimes missing, and

while requiring the regression coefficient for W to be zero. In fact, the surrogacy assumption can be checked by testing for a zero coefficient for W in such an analysis.

The hazard rate at time t for an "at risk" individual having Z_1 missing is given [21] by

$$\lambda_0(t)\mathrm{E}[\exp(Z'\beta)], \tag{3}$$

where the **expectation** is **conditional** on $(Z_2, W, T \geq t)$. The inclusion of $T \geq t$ in the conditioning event implies that this expectation is typically a complicated function of β and $\lambda_0(\cdot)$. If the probability that Z_1 is missing depends only on Z but not on (X, δ), then a "complete case" analysis that simply drops the observations having missing Z_1-values from (1) will typically yield **consistent estimators** of β, but may be quite inefficient. Toward more efficient estimation Zhou and Pepe [30] propose an "estimated partial likelihood" procedure wherein if Z_1, assumed to be discrete, is missing the expectation in (3) is replaced by a **nonparametric** estimate. This method allows the missingness rate to depend on (Z_2, W) but not on (X, δ, Z_1). Zhou and Wang [31] extended this method to continuous covariates using kernel estimation, though this extension may not be practical if the dimension of $\{Z_2, W\}$ is at all large. Lin and Ying [15] proposed an "approximate partial likelihood" procedure in which the summations leading to $\overline{Z}$ were restricted to individuals having Z_1 available. For components of $U(\beta)$ corresponding to Z_1, the overall summation was also restricted to individuals having known Z_1 values. This estimator can, but need not, improve upon the efficiency of the complete case estimator, and it requires a missing completely at random assumption (e.g. [17]); that is, the missingness probability is not allowed to depend on any aspect of (X, δ, Z, W).

An alternate simple procedure, referred to as regression **calibration** [6, 21], approximates the expectation in (3) by $\exp\{\mathrm{E}(Z_1|Z_2, W)'\beta_1 + Z_2'\beta_2\}$ and estimates $\mathrm{E}(Z_1|Z_2, W)$, typically using a simple **least-squares** procedure, using the validation subsample. This method is applicable if missingness depends only on $\{Z_2, W\}$. Because of the relative risk approximation, the resulting regression parameter estimates typically have some asymptotic bias. Wang et al. [27] develop the asymptotic theory for this estimator, along with a suitable variance estimator. In extensive **simulations**, they showed the bias to be surprisingly modest in situations of practical interest. The bias can be substantial, however, for large β values, depending somewhat on the censorship pattern. See [10] for some additional related approaches.

More comprehensive estimators of β generally require further modeling assumptions, either concerning the missingness probabilities, or the probability distribution for the covariates, or both. Chen and Little [7] consider a **nonparametric maximum likelihood** (NPML) procedure under which β and $\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$ are chosen to maximize

$$L = \prod_{i=1}^{n} \left\{ \lambda_0(X_i)^{\delta_i} \int \exp(Z_i'\beta) \times \exp[-\exp(Z_i'\beta)\Lambda_0(X_i)]F(Z_i;\theta)\,\mathrm{d}Z_i \right\}, \tag{4}$$

where the integral for the ith term is over Z_{1i} given Z_{2i} if Z_{1i} is missing, and reduces to the integrand at Z_i if Z_{1i} is available, and F denotes the probability density for Z, which is allowed to depend on a fixed length parameter vector θ. These authors maximize L after approximating Λ_0 by a step function with jumps at the uncensored X values, and obtain asymptotic distribution theory building on the work of Murphy et al. [18]. This NPML method allows missingness rates to depend on observed data (X, δ, Z_2), though a stronger than usual independent censoring assumption is required in that the censoring time is required to be independent of T given the observed covariate, and hence is not allowed to depend on the potentially unavailable Z_1 value; see [29] for earlier related work.

The augmented **inverse probability weighted estimator** (AIPW) of Wang and Chen [26] avoids this stronger censorship condition through inverse probability weighting. Their estimator of β, building on the work of Robins et al. [23], solves an estimating equation (*see* **Estimating Functions**) of the form

$$\sum_{i=1}^{n} \left\{ \eta_i \hat{\pi}_i^{-1} \int_0^{\infty} [Z_i - \hat{Z}(\beta, t)]\,\mathrm{d}N_i(t) + B_i\beta) \right\} = 0, \tag{5}$$

where $\eta_i = 0$ if Z_{1i} is missing and $\eta_i = 1$ otherwise, $\hat{\pi}_i$ is an estimate of the probability that Z_{1i} is missing, derived from a separate modeling exercise, and both the estimated covariate averages $\hat{Z}$ and the "augmentation term" B_i involve expectations

over the distribution of Z_1 given (X, δ, Z_2, W). This procedure has a nice **robustness** property in that it will generally provide consistent estimators of β under a missing at random [17] assumption even if the missingness model is not correctly specified, provided the distribution of Z_1 given (X, δ, Z_2, W) is correctly specified; and it will generally provide consistent estimates of β even if the covariate distribution is misspecified, provided the missingness rates are correctly modeled as a function of (X, δ, Z, W).

Further study of the properties of the NPML and AIPW estimators is needed, but simulation studies to date [7, 26] suggest good efficiency relative to the other estimators mentioned. The efficiency of the NPML estimator seems particularly good, though a noticeable bias was detected in simulations [26] if censoring rates depend strongly on the missing Z_1.

Relative Risk Parameter Estimation with a Reliability Subsample

As noted above, it quite often happens that measurements of Z_1 are unavailable for the entire sampled cohort; that is, there is no validation subsample. In these circumstances, it is necessary to make error model assumptions to connect the covariate measurement W to the "true", but missing, covariate Z_1. Often a classical measurement model assumption

$$W = Z_1 + \varepsilon, \tag{6}$$

is made, where the additive error ε is assumed to be independent of Z and of the corresponding failure and censoring times, and ε is assumed to have mean zero and a variance σ^2. Repeat measurements $W_1, W_2, \ldots$ on some study subjects are needed to estimate σ^2, or other aspects of the error distribution. The error variates corresponding to multiple measurements on a study subject are usually assumed to be independent, as are the error variates across study subjects. The measurements $\{W_1, W_2, \ldots\}$ meeting these conditions are said to constitute a covariate reliability sample.

Before proceeding, it is important to note that (6) and its attendant assumptions may be oversimplified or inappropriate in many important circumstances. For example, in the nutritional epidemiology setting mentioned above, one might expect nutrient consumption from repeat self-reports of dietary intakes to include systematic bias (e.g. errors having different distributions depending on such characteristics as body mass, age, and ethnicity) as well as positive within-person **correlations**. In such settings it may be crucial to identify biomarkers, or other objective measures that plausibly adhere to (6). Such objective measures are likely too expensive to be practical in the entire cohort in an epidemiologic study, so that a more comprehensive measurement model may be needed that assumes (6) for the objective measure on a subset, along with a more flexible model for the self-report data. In effect, the objective measures data can then be used to calibrate the self-report data on the entire cohort; see [20, 22, 24] for related discussion.

Assuming a reliability sample adhering to (6) to be available Xie et al. [28] adapt the regression calibration approach to reliability sample data, and extend it by improving the approximation to the relative risk in (3) to $\exp\{\mathrm{E}(Z_1|Z_2, W_1, W_2, \ldots; T \geq t)'\beta_1 + Z_2'\beta_2\}$ leading to a recalibration within each risk set, prior to applying a partial likelihood estimation procedure. More specifically, simple **variance component** arguments lead to estimates of the mean and covariance of $(Z_1, \overline{W}, Z_2)$ at each failure time, where $\overline{W}$ is the average of W values available for an individual. A joint normality assumption for $(Z_1, \overline{W}, Z_2)$ then leads to an approximate estimator of $\mathrm{E}(Z_1|\overline{W}, Z_2, X \geq t)$ for use at time t in the partial likelihood function. Asymptotic distribution theory and a variance estimator were provided [28] for the "ordinary" regression calibration, and "risk set" regression calibration estimators under reliability sampling. Both estimators typically incorporate some asymptotic bias, but the recalibration within risk sets extends the set of configurations where the bias will be negligible; see [8] for related work.

Consistent estimation of Cox model parameters, if only a reliability sample adhering to (6) is available, is possible using a corrected score function approach (*see* **Likelihood**). This approach (e.g. [19]) involves replacing the terms in the (standardized) score function (1) by consistent estimates based on the reliability sample; see also [5]. Some corrected score proposals require distributional assumptions on Z_1 or ε to hold, but recent work by Huang and Wang [11] avoids distributional assumptions in either the true covariate, or the error variate for consistent estimation of β, assuming two or more W-values are available for each individual. Briefly, $n^{-1}U(\beta)$ from

(1) can be consistently estimated by

$$n^{-1}\sum_{i=1}^{n}\int_0^\infty \left\{\begin{pmatrix}\overline{W}_i\\ Z_{2i}\end{pmatrix} - \frac{\hat{S}^{(1)}(\beta,t)}{\hat{S}^{(0)}(\beta,t)}\right\} \mathrm{d}N_i(t), \quad (7)$$

where

$$\hat{S}^{(j)}(\beta,t) = \sum_{i=1}^{n} Y_i(t)\mathcal{A}_i\left[\begin{pmatrix}\tilde{W}_{1i}\\ Z_{2i}\end{pmatrix} \times \exp\left\{\begin{pmatrix}\overline{W}_{2i}\\ Z_{2i}\end{pmatrix}'\beta\right\}\right], \quad j=0,1$$

and the operator $\mathcal{A}_i$ is a summation over all distinct pairs of $\overline{W}_{1i}$ and $\tilde{W}_{2i}$ selected from the set of $\{W_{1i}, W_{2i}, \ldots\}$ of W-values available on the ith study subject. The independence of the error terms in (6) for these W-values implies that $n^{-1}\hat{S}^{(j)}(\beta,t)$ estimates the corresponding $n^{-1}S_j(\beta,t)$ aside from the factor $E(\mathrm{e}^{\varepsilon\beta_1})$, which cancels out of the ratio in (7). This factor, needs to be estimated to obtain a **cumulative hazard** estimator, requiring some further assumption. For example, an assumption of symmetry of the error distribution is sufficient for this purpose.

The corrected score regression parameter estimator of Huang and Wang [11] performed well in simulations reported by these authors. The lack of monotonicity of the **estimating function** (7), however, presents some numerical challenges that have yet to be fully addressed.

The methods described here can be generalized to allow time-varying regression coefficients; see [25] for an interesting illustration based on repeat measurements on a **time-dependent covariate**.

Measurement Error in the Additive Hazards Model

A fixed covariate form of the additive hazards model (Aalen [1, 2], Andersen et al. [3]) can be written

$$\lambda(t;Z) = \lambda_0(t) + Z'\beta, \quad (8)$$

with restriction to assure nonnegative hazard rates. This model also has substantial applied potential, especially if extended to allow time-varying covariates. It's linear form admits some convenient, noniterative procedures for β-estimation in the presence of covariate measurement error.

Aalen [1] notes that the form of (8) is retained, with attenuated regression coefficient, under a classical measurement error model (5). Kulich and Lin [14] consider a corrected score approach to estimation in (8), assuming a validation subsample to be available. If Z is always available, Lin and Ying [16] propose that β in (8) be estimated by the explicit quantity

$$\hat{\beta} = \left[\sum_{i=1}^{n}\int_0^\infty \{Z_i - \overline{Z}(0,t)\}^{\otimes 2} Y_i(t)\,\mathrm{d}t\right]^{-1} \times \sum_{i=1}^{n}\int_0^\infty \{Z_i - \overline{Z}(0,t)\}\,\mathrm{d}N_i(t), \quad (9)$$

where $a^{\otimes 2} = aa'$. Kulich and Lin [14] derive a corresponding class of explicit estimators of β when Z_1 is available only in a validation sample, while W and Z_2 are always available. Under assumptions on the form of the mean and variance of W given Z these authors develop an unbiased score contribution for each study subject, and a corresponding explicit class of regression parameter estimates that is indexed by a downweighting parameter for the nonvalidation subsample. At least in important special cases this downweighting parameter can be estimated in an optimal fashion, yielding an estimator that necessarily improves on the efficiency of the complete case estimator, often substantially. Estimation procedures under (8) with only a reliability sample evidently have yet to be presented.

Acknowledgment

This work was supported by grant CA-53996 from the US National Institutes of Health. The author would like to thank Drs. C.Y. Wang and Sharon Xie who contributed to an earlier entry on this topic, and Dr. Jack Kalbfleisch for contributing to related material in [13].

References

[1] Aalen, O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[2] Aalen, O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**, 907–925.

[3] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[5] Buzas, J.S. (1998). Unbiased scores in proportional hazards regression with covariate measurement error, *Journal of Statistical Planning and Inference* **67**, 247–257.

[6] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Nonlinear Measurement Error Models*. Chapman and Hall, London.

[7] Chen, H.Y. & Little, R.J.A. (1999). Proportional hazards regression with missing covariates, *Journal of the American Statistical Association* **94**, 896–908.

[8] Clayton, D.G. (1991). Models for the analysis of cohort and case-control studies with inaccurately measured exposures, in *Statistical Models for Longitudinal Studies of Health*, J.H. Dwyer, M. Feinleib, H. Lippert eds. Oxford University Press, New York, pp. 301–333.

[9] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society Series B* **34**, 187–220.

[10] Hu, P., Tsiatis, A.A. & Davidian, M. (1988). Estimating the parameters in the Cox model when covariate variables are measured with error, *Biometrics* **54**, 1407–1419.

[11] Huang, Y. & Wang, C.Y. (2000). Cox regression with accurate covariates unascertainable: a nonparametric correction approach, *Journal of the American Statistical Association* **45**, 1209–1219.

[12] Hughes, M.D. (1993). Regression dilution in the proportional hazards model, *Biometrics* **49**, 1056–1066.

[13] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. Wiley, New York.

[14] Kulich, M. & Lin, D.Y. (2000). Additive hazards regression with covariate measurement error, *Journal of the American Statistical Association* **95**, 238–248.

[15] Lin, D.Y. & Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association* **88**, 1341–1349.

[16] Lin, D.Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61–71.

[17] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis of Missing Data*. Wiley, New York.

[18] Murphy, S.A., Rossini, A.J. & Van der Vaart, A.W. (1997). Maximum likelihood estimation in the proportional odds model, *Journal of the American Statistical Association* **92**, 968–976.

[19] Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error, *Biometrics* **48**, 829–838.

[20] Plummer, M. & Clayton, D. (1993). Measurement error in dietary assessment: an assessment using covariance structured models, part II, *Statistics in Medicine* **12**, 937–948.

[21] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[22] Prentice, R.L., Sugar, E., Wang, C.Y., Neuhouser, M. & Patterson, R.E. (2002). Research strategies and the use of biomarkers in studies of diet and chronic disease, *Public Health Nutrition* **5**, 977–984.

[23] Robins, J.M., Rotnitsky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* **89**, 846–866.

[24] Sawaya, A.L., Tucker, K., Tsay, R., Willett, W., Salzman, E., Dallal, G.E. & Roberts, S.B. (1996). Evaluation of four methods for determining energy intake in young and older women: comparison with double labeled water measurements of total energy expenditure, *American Journal of Clinical Nutrition* **63**, 491–499.

[25] Tsiatis, A.A., DeGruttola, V. & Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 count in patients with AIDS, *Journal of the American Statistical Association* **90**, 27–37.

[26] Wang, C.Y. & Chen, H.Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression, *Biometrics* **57**, 414–419.

[27] Wang, C.Y., Hsu, L., Feng, Z.D. & Prentice, R.L. (1997). Regression calibration in failure time regression, *Biometrics* **53**, 131–145.

[28] Xie, S.X., Wang, C.Y. & Prentice, R.L. (2001). A risk set calibration method for failure time regression by using a covariate reliability sample, *Journal of the Royal Statistical Society Series B* **63**, 855–870.

[29] Zhong, M., Sen, P.K. & Cai, J. (1996). Cox regression model with mismeasured covariate or missing covariate, in ASA Proceedings of the Biometrics Section, Washington, D.C., pp. 323–328.

[30] Zhou, H. & Pepe, M. (1995). Auxiliary covariate data in failure time regression analysis, *Biometrika* **82**, 139–149.

[31] Zhou, H. & Wang, C.Y. (2000). Failure time regression with continuous covariates measured with error, *Journal of the Royal Statistical Society Series B* **62**, 657–665.

(*See also* **Measurement Error in Epidemiologic Studies**; **Survival Analysis, Overview**)

ROSS L. PRENTICE

Measurement Scale

Many different types of variables occur in statistical investigations. Some are merely classifications, such as a diagnosis into one of three unrelated diseases. Sometimes the classifications have an order, for example the stages of cancer or the common categorization into mild, moderate, and severe. For other variables the order is everything – "preference" recorded on a visual analog scale is an example. Here there are no categorical groups, and yet one can say that one score corresponds to a greater preference than another. Yet other variables seem to impose more sophisticated mathematical relationships between the possible values. With temperature, for example, we can say not only that one temperature is larger than another, but also that the difference between a given pair of temperatures is larger than the difference between some other pair. And, for other variables, we can go even further: for weight or concentration or height we can say that one is twice the other, or half again as large as the other. And, of course, yet other variables are simply counts – the number of cells on a plate, for example.

In fact, many different classifications of variable types have been proposed, often motivated from the perspective of statistical analysis: the set of techniques needed to analyze one type of variable often differs from that needed to analyze another type. However, one classification in particular has had a major impact on statistics. This is the classification into *nominal, ordinal, interval*, and *ratio* scales proposed by the psychophysicist Stevens [18, 19].

Prior to Stevens' work, the emphasis in understanding measurement had been in the physical sciences. The problems there, at least at that stage and at least superficially, seemed more straightforward. Measurement involved assigning numbers to represent the properties of objects, where the objects satisfied (i) an order relationship and (ii) a physical process of addition. The latter is illustrated by the placing of two objects in one pan of a weighing scales, and balancing them by a third object in the other pan. In terms of weight, this third object corresponds to the "physical addition" of the first two. Such a physical addition process is nowadays called *concatenation*. In such situations the notion of "quantity" seems relatively straightforward. Axiom systems describing (i) and (ii), which the objects must satisfy in order for the relationships between them to be representable by order and addition of numbers, were developed by von Helmholtz [23] and Hölder [7]. Campbell [2], in particular, adopted this approach. Of course, even here things are not completely simple: physical concepts such as density are defined in terms of other concepts. They have thus been called "derived" measurements, with the directly measured ones being called "fundamental measurements". Moreover, and more importantly, densities do not physically add in the same way as weight: combine two samples of gas with different densities and the result is something with an intermediate density, not the sum of the densities.

However, in other scientific areas, notably psychology, things are even less clear. In particular, there is often no "physical addition operation" evident. Because of this, notions of measurement in psychology came under much criticism [3]. Stevens tackled these criticisms by pointing out that the numerical representation preserving the relationships between a set of empirical objects was not unique and that the alternative numerical representations are obviously related since they represent the same empirical system. To get from one legitimate numerical representation to another, some transformation or mapping is involved. Stevens suggested that different such mappings defined different *types* or *scales* of measurement. Thus, if any one-to-one mapping was allowed (that is, any mapping which preserved the unique identity of the classes of objects), then the measurement was on a *nominal* scale. For example, the appearance of a lesion may be classified into one of three distinct types. If any order-preserving mapping was allowed (that is, if any numerical representation which assigned numbers in the same order to the objects was allowed), then measurement was on an *ordinal* scale. For example, "severity" might be encoded so that any alternative encoding would be equally legitimate, provided it had the same order. If any linear transformation was allowed (that is, if rescaling the numbers by changing the units and then adding some constant resulted in an alternative numerical assignment which still preserved the relationships between objects), then the scale was *interval*. Body temperature is an example: this might be measured in degrees centigrade or degrees Fahrenheit, the two being related by a linear transformation. And, finally, if any change of

the units yielded an alternative, equally legitimate numerical assignment, then the scale was *ratio*, for example, changing the units of length from inches to centimeters. The physical measurements with which Campbell was concerned were of this last type; Stevens thus generalized the notions of measurement. In each case the class of transformations, which lead to another, equally legitimate, representation of the empirical system being modeled is called the class of *admissible* or *permissible* transformations. Mathematically, this class defines the scale being studied.

The notion of admissible transformations has implications for what statistical statements may sensibly be made using the data. If, for example, any numerical assignment which preserves the order of a set of values is equally legitimate, then comparing the arithmetic means of two groups is of dubious value: it may be possible to invert the relative order of the two means by a suitable choice of transformation. To illustrate, suppose that one numerical representation has the values $\{1, 5\}$ for the two objects in one group and $\{3, 4\}$ for the two objects in the other group. Then the mean, 3, of the first group is smaller than the mean, $3\frac{1}{2}$, of the second group. However, consider the alternative numerical assignment $\{1, 7\}$ for the members of the first group and $\{3, 4\}$ for the members of the second. This preserves the order of the numbers – the object which was previously assigned the smallest number has still been assigned the smallest number, and so on. But this new numerical assignment yields respective means of 4 and $3\frac{1}{2}$. Now the mean of the first group is larger than the mean of the second. In general, one's conclusions will be an artifact of one's choice of numerical assignment, and will not reflect any truth about the empirical reality. Thus it seems that statistical arguments must take account of scale type. However, the assumptions made by inferential statistical arguments are distributional, and not scale-specific. Indeed, given a set of numbers, no matter what their scale type, arbitrary statistical statements may be made about those numbers – one is able to compute a t statistic and carry out a t test whatever the scale type, and this might even seem a sensible thing to do if suitable distributional assumptions are satisfied. The tension between these two viewpoints has stimulated a major debate, running from the time of Campbell and Stevens right up to the present [5, 6, 15, 21, 22]. Its resolution is subtle and lies in awareness of the fact that the objective of statistical analysis is ultimately to make a statement about the empirical system being studied, and not simply about the numbers being used to represent that system. However, this needs to be tempered by the fact that statistical statements applied to what are apparently impermissible transformations of the data may lead to the detection of hitherto unsuspected structures in those data. It seems that if one wants to test strong theories, described in terms of numbers derived by a well-defined mapping from a well-understood empirical system, then the strictures imposed by the theory of scale types should be adhered to. But if one's theories are less stringently formalized, then the constraints of scale type are less important and, indeed, adhering to them may risk missing important discoveries (see [6] and the ensuing discussion).

The nominal, ordinal, interval, and ratio typology is an old one. Since its formulation a huge amount of work has been carried out, partly philosophical, concerned with relating measurement activities to scientific questions, and partly mathematical, concerned with developing axiom systems which an empirical system must obey if it is to be representable by a given numerical system. For reviews of this work see [8–10, 14, 17, 20]. One conclusion of this work has been to show that, for mappings to the real numbers, only certain types of scales can exist, and that Stevens' ordinal, interval, and ratio classification is closely related to this set. However, an interesting anomaly is that for mappings to rational numbers there is an infinite variety of scale types [1]. Since all data are recorded to only a finite number of digits, scientific mappings are in fact to only a subset of the rationals. Quite what the implication of this is, if any, remains to be seen.

Clearly, physical addition operations hold a central place in measurement theory. Such operations can be mapped to addition, so that very familiar numerical operations can be used. But they are not the only empirical relationships which can be mapped to addition. A completely different class of relationships arises in *conjoint measurement*. Suppose that the objects in the empirical system can be ordered according to attribute A, that each object can be described in terms of a pair of attributes (B, C), and that each of B and C can be ordered. Then, subject to certain conditions, it is possible to find numerical assignments such that the relationships between

objects can be represented by addition between the assigned numbers (see, for example, [11] and [13]).

So far we have described the aim of measurement as being to assign numbers according to a numerical system within which the relationships correspond to those between the empirical objects. This approach is termed *representational measurement theory*, and is by far the best developed. However, it is not adequate for all situations in which measurement is used. In particular, in many areas of psychology this approach seems to be inadequate, chiefly because it is not clear precisely what empirical system is being modeled. As a consequence, alternative theories have been developed. Chief amongst these is the *operational* approach. This takes the measurement operation as *defining* the attribute being measured. As a consequence, no notion of permissible transformations, and consequently of scale types, can arise. Yet a third theory, termed by Michell [12] the *classical* approach, takes as its starting point that numerical quantities of attributes exist, with the objective of measurement being the determination of the magnitude of these quantities. A key driving force for both the representational and classical approaches is the desire to characterize the relationship between the empirical system and the numerical system. In the operational approach, however, with the assumption of an underlying empirical reality not being necessary, the emphasis is on internal consistency and reproducibility. One might describe the aims of the representational, operational, and classical approaches as being, respectively, to assign, define, and discover numerical representations.

As will be apparent from the above and from the reference list at the end of this article, much of the debate about the fundamental concepts of measurement has occurred in the psychological literature. This is not surprising: in psychology of all disciplines, measurement is difficult. Rarely can the attributes being studied be directly observed, so that subtle indirect measurement procedures have to be devised. Naturally this stimulates debate about the precise nature of the measurement activity and what the resulting numbers actually mean. Earlier manifestations of measurement theory in physics, most notably in *dimensional analysis* [16], although useful, stimulated little debate about underlying principles. Measurement procedures in psychology involve constructing complex instruments which often require collecting many scores or numbers which need to be combined using sophisticated statistical techniques to yield a final measurement. Examples of such methods include **paired comparisons**, Guttman, Likert, and Thurstone scaling factor analysis, and item response theory [4] (*see* **Psychometrics, Overview**). Measurement procedures in medicine can be equally complex: a prime example being attempts to formulate **quality of life** scales.

References

[1] Cameron, P. (1989). Groups of all order-preserving homeomorphisms of the reals that satisfy finite uniqueness, *Journal of Mathematical Psychology* **31**, 135–154.

[2] Campbell, N.R. (1920). *Physics: The Elements*. Cambridge University Press, Cambridge.

[3] Ferguson, A., Meyers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., Campbell, N.R., Craik, K.J.W., Drever, J., Guild, J., Houston, R.A., Irwin, J.O., Kaye, G.W.C., Philpott, S.J.F., Richardson, L.F., Shaxby, J.H., Smith, T., Thouless, R.H. & Tucker W.S. (1940). Quantitative estimates of sensory events, *Report of the British Association for the Advancement of Science* **2**, 331–349.

[4] Hambleton, R.K., Swaminathan, H. & Rogers H.J. (1991). *Fundamentals of Item Response Theory*. Sage, Newbury Park.

[5] Hand, D.J. (1993). Comment on "Nominal, ordinal, and ratio scales typologies are misleading", *American Statistician* **47**, 314–315.

[6] Hand, D.J. (1996). Statistics and the theory of measurement (with discussion), *Journal of the Royal Statistical Society, Series A* **159**, 445–492.

[7] Hölder O. (1901). Die Axiome der Quantitat und die Lehre vom Mass, Berichte über die Verhandlungen der königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, *Mathematische-Physiche Klasse* **53**, 1–64.

[8] Krantz, D.H., Luce, R.D., Suppes, P. & Tversky P. (1971). *Foundations of Measurement*, Vol. 1: Additive and Polynomial Representations. Academic Press, New York.

[9] Luce, R.D. (1996). The ongoing dialog between empirical science and measurement theory, *Journal of Mathematical Psychology* **40**, 78–98.

[10] Luce, R.D., Krantz, D.H., Suppes, P. & Tversky A. (1990). *Foundations of Measurement*, Vol. **3**: Representation, Axiomatization, and Invariance. Academic Press, San Diego.

[11] Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement, *Journal of Mathematical Psychology* **1**, 1–27.

[12] Michell, J. (1986). Measurement scales and statistics: a clash of paradigms, *Psychological Bulletin* **100**, 398–407.

[13] Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. Lawrence Erlbaum, Hillsdale.
[14] Narens, L. (2002). *Theories of meaningfulness*, Lawrence Erlbaum, Mahwah, NJ.
[15] Niederée, R. (1994). There is more to measurement than just measurement: measurement theory, symmetry, and substantive theorizing, *Journal of Mathematical Psychology* **38**, 527–594.
[16] Porter, A.W. (1933). *The Method of Dimensions*. Methuen, London.
[17] Roberts, F.S. (1979). *Measurement Theory*. Addison-Wesley, Reading.
[18] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.
[19] Stevens, S.S. (1951). Mathematics, measurement, and psychophysics, in *Handbook of Experimental Psychology*, S.S. Stevens, ed. Wiley, New York.
[20] Suppes, P., Krantz, D.H., Luce, R.D. & Tversky, A. (1989). *Foundations of Measurement*, Vol. 2: Geometrical, Threshold, and Probabilistic Representations. Academic Press, San Diego.
[21] Velleman, P.F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio scales typologies are misleading, *American Statistician* **47**, 65–72.
[22] Velleman, P.F. & Wilkinson, L. (1993). Reply to comments on Velleman and Wilkinson (1993), *American Statistician* **47**, 315–316.
[23] von Helmholtz, H. (1887). Zählen und Messen erkenntnis-theoretisch betrachet, *Philosophische Aufsätze Eduard Zeller gewidmet*. Leipzig; English translation by C.L. Bryan, *Counting and Measuring*. van Nostrand, Princeton, 1930.

(*See also* **Nominal Data**; **Ordered Categorical Data**)

DAVID J. HAND

Median Effective Dose

In biological experiments, the primary goal is to investigate the responses of biologic subjects to a stimulus administered at various levels (*see* **Stimulus–Response Studies**). Drugs, chemical compounds, toxicants, food preservatives, radiation, or specific environmental conditions are typical examples of stimuli. Biologic subjects can be human volunteers, animals, insects, microorganisms, or living tissue. In practice, however, it is impossible and difficult to measure the responses quantitatively. However, the responses of the units to the stimulus can be easily documented by the occurrence of some meaningful and well-defined events such as death, survival, convulsion, infection, eradication of infection, induction of estrus, or cure of disease. These responses are known as *quantal responses* (*see* **Binary Data**). As indicated by Ashford [4] and Morgan [48], a quantal response usually involves an irreversible process in organisms that either respond or do not respond. The corresponding experiments are called quantal biological assays or quantal bioassays [22, 23, 33, 70, 72, 73], (*see* **Biological Assay, Overview**). Biological assay not only plays an important role in the evaluation of pharmacological and toxicological effects of a chemical compound, but is also crucial for screening possible drug entities. Note that a quantal bioassay is closer to the *direct* bioassay than it is to the *indirect* quantitative bioassay.

Finney [23] defined the tolerance of a biological subject as the dose level just sufficient to produce predefined events (*see* **Quantal Response Models**). If the dose level given is lower than the tolerance of the subject, then the event will not occur. The subject will respond if the dose level administered is higher than the tolerance. However, the tolerance to a particular stimulus may be different from subject to subject owing to genetic, environmental, and other unknown factors. The resulting distribution of dose levels of a stimulus with respect to a well-defined event is referred to as the *tolerance distribution*. In a traditional fixed indirect quantal bioassay, a number of subjects are randomly assigned to receive one of several preselected dose levels of the stimulus under investigation. The occurrence of the event of interest is then documented for each subject according to a prespecified time schedule. Therefore, a quantal bioassay is an indirect bioassay since it only records whether the tolerance of a subject is higher or lower than the given dose levels. The proportions of subjects who respond at each dose level are then used to describe the tolerance distribution, which provides an estimation of the dose–response relationship of the stimulus. The most commonly employed measure to characterize the tolerance distribution is the median effective dose (MED). The MED is the **median** of the tolerance distribution, which is the dose of a stimulus that generates, on average, a predefined response in 50% of subjects. Hence, the MED is also referred to as the ED_{50} [22]. The concept of an MED was first introduced by Trevan [69] as the median lethal dose (MLD) to describe the potency of a test stimulus. In general, let p be a real number between 0 and 1, ED_{100p} (or LD_{100p}) is then the dose of a stimulus that produces an average effect in $100p\%$ of subjects. Other related measures are $\mathrm{ED}_{100p(t)}$ and $\mathrm{ET}_{100p(d)}$, where $\mathrm{ED}_{100p(t)}$ is the MED at which on average $100p\%$ of subjects will respond by time t, and $\mathrm{ET}_{100p(d)}$ is the median effect time at dose level d by which the occurrence of an event is observed in $100p\%$ of subjects [51]. Finney [22, 23] provided a comprehensive review of the design and estimation of MED and related topics. More detailed developments can be found in [33] and [48]. Wu [77] and Ashford [4] indicated that the concept of quantal responses and ED_{50} is also useful in the areas of education, economics, energy, transportation, criminology, legislation, and psychology.

Tolerance Distribution

Consider an indirect quantal bioassay in which r_i of n_i subjects respond to the ith dose level, denoted by x_i, which is often expressed on the basis of a logarithmic or other **transformed** scale of a test stimulus, where $i = 1, \ldots, k$. Let $P_i = F(x_i, \overline{\omega})$ be the probability of the response of a subject receiving the stimulus at dose levels x_i; here, $F(\cdot)$ denotes the cumulative distribution function (cdf), and $\overline{\omega}$ is a q-dimensional vector of unknown parameters to characterize the relationship between the tolerance distribution and dose levels. Given that $n_1, \ldots, n_k$ and $r_1, \ldots, r_k$ are mutually independent **binomial random variables**, that is, $r_i \sim \mathrm{B}(n_i, P_i)$, $i = 1, \ldots, K$,

the **likelihood** is given as

$$L(\overline{\omega}) = \prod_{i=1}^{k} C_i P_i^{r_i} (1 - P_i)^{n_i - r_i} \tag{1}$$

where $C_i = n_i!/[r_i!(n_i - r_i)!]$.

A commonly employed model to describe the dose–response relationship is to fit a simple **linear regression** to dose levels, with intercept α and slope β. That is,

$$\begin{aligned} P_i &= F(x_i; \overline{\omega}) \\ &= F(x_i; \alpha, \beta) \\ &= F(\alpha + \beta x_i) \end{aligned} \tag{2}$$

This formulation of the dose–response relationship, in fact, assumes a **location/scale** model for the tolerance distribution, with location parameter α and scale parameter β. One of the most commonly employed tolerance distribution in indirect quantal bioassays is the **normal distribution** for logarithmic tolerances whose probability density function (pdf) is given as

$$f(x) = \left\{ \frac{1}{\sigma(2\pi)^{1/2}} \right\} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2} \right], \quad -\infty < x < \infty \tag{3}$$

Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cdf and pdf. It follows from (2) that

$$\begin{aligned} Y_i &= \Phi^{-1}(P_i) \\ &= \alpha + \beta x_i. \end{aligned} \tag{4}$$

The inverse normal cdf applied to the probability of response is called the *probit transformation* [22]. Gaddum [24] referred to Y as the normal equivalent deviate (NED) of P_i. Since the standard normal distribution is symmetric about 0, and since ED_{50} is the dose level where the predefined response occurs in 50% of subjects, it follows from (3) that when $x = \mathrm{ED}_{50}$ $P = 0.5$, $Y = 0$ and

$$\begin{aligned} \mathrm{ED}_{50} &= \theta \\ &= -\frac{\alpha}{\beta}. \end{aligned} \tag{5}$$

A comparison between the normal pdf in (3) and θ reveals that the normal distribution in (3) has a mean,

$$\mu = -\frac{\alpha}{\beta}, \tag{6}$$

with standard deviation

$$\sigma = \frac{1}{\beta}. \tag{7}$$

In other words, one can always reparameterize the mean and standard deviation directly in terms of ED_{50} and the slope β as follows:

$$\begin{aligned} Y_i &= \Phi^{-1}(P_i) \\ &= \beta(x_i - \theta). \end{aligned} \tag{8}$$

This result is true for any distribution such that $F(-t) = F(t)$, which also includes the **logistic distribution**, with the cdf given as

$$F(\alpha + \beta x_i) = \{1 + \exp[-(\alpha + \beta x_i)]\}^{-1}. \tag{9}$$

Berkson [7] first referred to the inverse transformation of the probability from a logistic distribution as the logit, which is given as

$$Y_i = \ln\left[\frac{P_i}{1 - P_i} \right]. \tag{10}$$

In (4) and (10), both responses Y_i and stimulus are subject to measurement errors. Ashford [4] and Morgan [48] indicated that the effect of measurement errors is to decrease the slope in (4) and (10), between the inverse transformation of the cdf of the tolerance distribution and the stimulus. However, the point estimator of the MED remains unchanged. Other location/scale families include angular (*see* **Delta Method**), **uniform**, **Cauchy**, and **extreme-value** distributions. For more details about these distributions, see [22, 23, 48].

Finney [23] indicated that if a response rate is between 0.05 and 0.95, then it is difficult to distinguish between the normal, logistic, angular, and uniform distributions. In fact, one will not be able to discriminate the normal distribution from the logistic distribution for P_i, between 0.01 and 0.99. In practice, it is almost impossible to identify the correct transformation from the data. However, the impact on estimation of ED_{50} for an incorrectly selected tolerance distribution is negligible. To estimate ED_{50}, the logistic distribution is preferred owing to its nice statistical properties and the simplicity of computation. Since the difference among the normal, logistic, angular, and uniform distributions occurs only in the extreme tails of the distributions, the selection of models *(see* **Model, Choice of**) and the **goodness-of-fit** of models

are extremely important in the estimation of extreme dose levels such as ED_{05} or ED_{90}. Consequently, we may consider alternative three-parameter models such as the Aranda–Oradz model [48] or the quantity obtained from the omega distribution suggested by Copenhaver & Mielke [13] for statistical inference of ED_{05} or ED_{90}.

Estimation Procedures

We first restrict our discussion on the estimation of the MED to the simple two-parameter model. Principles and extensions to more complicated models are straightforward. Even for the simpler location/scale model, it is a nonlinear model. The **maximum likelihood** method is used to estimate unknown parameters. Let a and b denote the maximum likelihood estimates (MLE) of α and β, respectively. The MLE of ED_{50} can then be obtained either as $\hat{\theta} = [F^{-1}(0.5) - a]/b$ under model (2) or as the MLE of θ derived under model (8). If the assumed tolerance distribution, for example, the normal or logistic distributions, is symmetric about 0, then the MLE of ED_{50} is simply equal to $\hat{\theta} = -(a/b)$.

Various methods for numerical optimization can be applied to find the MLE. These methods usually involve the technique of iterative reweighted **least squares** (IRLS), which can be found in most commercial statistical computer packages such as SAS, BMDP, GLIM, IMSL, and others (*see* **Software, Biostatistical**). The Newton–Raphson method uses the Fisher **information matrix** as the weighting matrix during the iterative process, while the method of scoring employs the information matrix. Morgan [48] indicated that it is more convenient to use the method of scoring than the Newton–Raphson method, although for the logit model, both methods are identical (*see* **Optimization and Nonlinear Equations**).

Under appropriate regularity conditions [60], the vector of MLEs is asymptotically normal with mean zero and **covariance matrix** V, where V is the limit of the inverse of the information matrix, as the sample size goes to infinity and is referred to as the asymptotic covariance matrix of a and b. For the probit model, Griffiths et al. [28] compared three estimators of the covariance matrix for the MLEs of α and β. They are the inverse of the negative of the Hessian matrix, the inverse of the information matrix, and the inverse of the outer product of the first-order partial derivative of the log-likelihood function proposed by Berndt et al. [9]. These estimators of the covariance matrix V are asymptotically equivalent. However, Griffiths et al. [28] showed through a **simulation** study that in small samples, on average, both the information matrix and the Hessian matrix provide almost identical results and more accurate estimates of the asymptotic covariance matrix than the estimate proposed by Berndt et al. [9]. The **mean squared error** of the MLE is considerably larger than the asymptotic covariance matrix. As a result, the bias in finite samples could potentially be large.

After the MLEs and their corresponding estimated asymptotic covariance matrices are obtained, the following goodness-of-fit test statistic can be used to verify the underlying assumptions of the model:

$$X^2 = \sum \frac{(r_i - n_i \hat{P}_i)^2}{n_i \hat{P}_i (1 - \hat{P}_i)}, \tag{11}$$

where $n_i \hat{P}_i$ is the number of responses predicted by the fitted model.

If the model is adequate, then asymptotically, X^2 follows a central χ^2 with $K - 2$ df (*see* **Chi-square Distribution**). If X^2 indicates a significant lack-of-fit, then the possible causes should be carefully and thoroughly investigated. The possible causes of lack-of-fit include a poorly fitting model, the violation of binomial assumption due to a possible **correlation** between responses of subjects, or the existence of heterogeneity among responses of different subjects (*see* **Overdispersion**). Finney [22, 23] suggested the use of $X^2/(K - 2)$ as a heterogeneity factor for scaling up all variances. Since X^2 is used as a test statistic for goodness-of-fit, an intuitive alternative approach to the estimation of α and β is to find the estimates to minimize this quantity. The resulting method is the noniterative minimum X^2 method advocated by Berkson [8]. Taylor [68] showed that the minimum X^2 method is a regular best asymptotically normal **BAN** procedure. Therefore, one of advantages of the minimum X^2 method is that it provides noniterative explicit estimates that are asymptotically equivalent to the maximum likelihood procedure. Morgan [48] indicated that caution is required in the interpretation of the asymptotic optimality for the two methods. The optimality of the maximum likelihood method is achieved when the number of doses goes to infinity, while the minimum X^2 method reaches its optimality

when the number of subjects in each group goes to infinity [42, 61].

Confidence Intervals

Let v_{11}, v_{22}, and v_{12} be the estimated asymptotic variances and covariances of the MLEs a and b obtained from either the Hessian or the information matrices by replacing α and β with a and b. The estimated asymptotic variance of the MLE for θ is then given by

$$v(\hat{\theta}) = \frac{v_{11} + 2\hat{\theta}v_{22} + \hat{\theta}^2 v_{22}}{b^2}. \tag{12}$$

Since $(\hat{\theta} - \theta)/[v(\hat{\theta})]^{1/2}$ converges in distribution to the standard normal distribution, the $(1-\alpha)100\%$ **confidence interval** for θ can be obtained by the **delta method** as follows:

$$\hat{\theta} \pm z_{\alpha/2}[v(\hat{\theta})]^{1/2}, \tag{13}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution.

However, the $(1-\alpha)100\%$ confidence interval for θ based on **Fieller's theorem** [20, 21] is the set of θ values such that

$$\left\{\theta \left| \frac{(a+\theta b)^2}{b^2 v(\theta)} < z_{\alpha/2}^2 \right.\right\} \tag{14}$$

Let g be $z_{\alpha/2}^2$ times the inverse of the square of the statistic for testing whether the slope is different from zero. Failure to reject the null hypothesis of the zero slope implies that $g \geq 1$, and hence, the resulting $(1-\alpha)100\%$ confidence interval for the MED by Fieller's theorem will either be the entire real line or the union of two disjoint open intervals.

Consider the asymptotic **likelihood ratio test** to test the **null hypothesis**, $\theta = \theta_0$, against the **alternative hypothesis**, $\theta \neq \theta_0$. Let $l(\theta_0)$ be the value of the log-likelihood maximized with respect to β under the null hypothesis that $\theta = \theta_0$, and let $l(\theta, \beta)$ be the value of the log-likelihood maximized with respect to both β and θ under the alternative hypothesis. It follows that the $(1-\alpha)100\%$ confidence interval for ED_{50} by the likelihood ratio method is given by the following set:

$$\{\theta | 2[l(\theta, \beta) - l(\theta_0)] < z_{\alpha/2}^2\}. \tag{15}$$

Asymptotically, the $(1-\alpha)100\%$ confidence intervals for the MED derived by the delta method, Fieller's theorem, and the likelihood ratio method are all equivalent. However, the interval by the delta method always exists and is finite. On the other hand, the confidence interval produced by Fieller's theorem and the likelihood ratio test may not be of finite length. Williams [74] reported that overwhelming evidence exists that the confidence interval by Fieller's theorem is conservative.

Because Fieller's theorem is an exact method under the normality assumption, the only source of error for its corresponding confidence interval involves the normal approximation to p_i. If g is small, then both the delta and Fieller's intervals will be virtually the same. As a result, Finney [22, 23] recommended the use of Fieller's interval only if g is <0.05. However, limited simulation performed by Abdelbasit & Plackett [2] failed to support the use of Fieller's interval as advocated by Finney [22, 23]. Cox [14] found that the delta method provides a useful alternative to Fieller's theorem. Sitter & Wu [64] provided both theoretically and empirically, the most comprehensive comparison between the intervals derived by the delta method and Fieller's theorem. They showed that Fieller's interval and the delta interval differ only by an inflation factor and a shift factor, with an asymptotic order of $O_p(1/n)$. The inflation factor is 1 (<1) when the design is symmetric (asymmetric). However, the shift factor vanishes when the design is symmetric.

The simulation results of Sitter & Wu [64] provide convincing evidence to support the use of Fieller's interval rather than the delta interval for ED_{50}. Fieller's theorem will generate a confidence interval for ED_{50} with an infinite length when the slope is not statistically different from 0 (i.e. $g \geq 1$). In this situation, the relationship between the response probability and the dose levels cannot be adequately and satisfactorily established through the assumed model. Consequently, any inference, including the construction of the confidence interval for ED_{50} by Fieller's theorem based on the assumed model, is meaningless.

For responses with more than one outcome (*see* **Polytomous Data**), one possible model for the estimation of ED_{50} is the **proportional-odds model**. Let P_{ij} be the probability of observing the jth outcome of a total of J possible outcome categories at the ith dose level. Then, a **multinomial** model can be used

to describe the data using the cumulative logits

$$\begin{aligned}&\log\text{it}(P_{i1}+\cdots+P_{ij})\\&=\ln\left\{\frac{P_{i1}+\cdots+P_{ij}}{1-(P_{i1}+\cdots+P_{ij})}\right\}\\&=\alpha_j+\beta x_i,\quad j=1,\ldots,J-1;\ i=1,\ldots,k,\end{aligned}\tag{16}$$

where $\alpha_1<\cdots<\alpha_J$. The ED_{50} can be estimated in the usual way for the cumulative collapsed categories. However, the assumption of parallelism, of the lines for different values of j, must be verified before the estimates of α_j and β can be used for inferences about the ED_{50}.

Mixture Models

Similar to the "**placebo** effect" in clinical trials, the natural response in bioassay or toxicology should be also taken into account. The natural response is the response separated from that attributed to the administered stimulus. An approach to incorporating the natural response into the estimation of ED_{50} is to employ the model suggested by Abbott [1] for the following response probability:

$$F(x_i)=\lambda+(1-\lambda)F^*(x_i),\tag{17}$$

where $0\le\lambda\le1$ is the probability of natural response and $F^*(x_i)$ is the probability for dose level x_i whose occurrence of the defined event is not due to natural causes. As discussed earlier, a significant heterogeneity factor X^2 may indicate that (i) the population under investigation may not be homogeneous or (ii) the observed tolerance distribution is not unimodal. In this case, the underlying population may consist of a mixture of homogeneous subpopulations. The response probability of each subpopulation is still related to the dose by some individual tolerance distribution.

Suppose that there are a total of h subpopulations, each with a tolerance distribution F_j. For the ith dose level, n_{ij} of n_i subjects come from the jth population with probability δ_j and m_{ij} subjects respond, where $j=1,\ldots,h; I=1,\ldots,k$. Therefore, for a fixed n_i, $(n_{i1},\ldots,n_{ih})$ are jointly distributed as a multinomial distribution with parameter $(n_i,\delta_1,\ldots,\delta_h)$. In addition, the sum of m_{ij} is r_i. For fixed $(n_{i1},\ldots,n_{ih})$, m_{ij} are independently distributed as binomial random variables with parameters $[n_{ij}, F_j(d_i)]$. If a control group is also included in the study to account for natural mortality, then the likelihood may be formulated as [41].

$$\begin{aligned}L(\overline{\omega})&=C_0\lambda^{r_0}(1-\lambda)^{(n_0-r_0)}\\&\quad\times\prod_{i=1}^{k}C_iP_i^{r_i}(1-P_i)^{(n_i-r_i)},\end{aligned}\tag{18}$$

where $C_i=n_i!/[r_i!(n_i-r_i)!]$, $i=0,\ldots,k$; $P_i=\sum\delta_jF_j(d_i)$.

For the location/scale family, the parameters in $\overline{\omega}$ consist of $\lambda,\delta_1,\ldots,\delta_{h-1};\ \alpha_1,\ldots,\alpha_h;\beta_1,\ldots,\beta_h$; and if $x_i=\ln d_i$, then

$$\begin{aligned}F_j(x_i)&=\lambda+(1-\lambda)F^*(\alpha_j+\beta_jx_i),\\&\quad j=1,\ldots,h;\ i=1,\ldots,k.\end{aligned}\tag{19}$$

Lwin & Martin [41] proved that the MLEs exist under model (18). They suggested that the MLEs may be obtained either by the usual Newton–Raphson or scoring methods or the Nelder–Mead simplex algorithm [50]. In addition, they also pointed out that the **EM algorithm** [16] can be used to solve the likelihood functions for MLEs.

Overdispersion

Statistical inference about the ED_{50} as described above, assumes independence of individual subjects. In toxicology or **teratology** experiments, subjects are sampling units nested within the experimental units, such as litters. Therefore, responses observed from the subjects within the same litter are not independent. Hence, the actual variability exhibited by the data is larger than that expected under the independent binomial assumption. This phenomenon is referred to as **overdispersion** or extra-binomial variation caused by the litter effect. Other examples of overdispersion can be found in [15, 48, 75, 76].

Consider an experiment in which m_i litters receive the ith dose level of the stimulus. Let r_{ij} be the number of responses in n_{ij} subjects in the jth litter of the ith dose group, $j=1,\ldots m_i$, $i=1,\ldots,k$. Denote the corresponding response probability by P_{ij}. Then, given n_{ij} and P_{ij}, the **conditional** distribution of r_{ij} follows the binomial distribution given in (1). If we further assume that P_{ij} follows a **beta distribution** with parameters η_{1i} and η_{2i}, the marginal distribution

of r_{ij} follows a **beta-binomial distribution** with the following log-likelihood function as shown by Segreti & Munson [59].

$$L = \text{constant} + \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left\{\sum_{r=0}^{r_{ij}-1}\log(\gamma_{1i}+r\gamma_{2i}) + \sum_{r=0}^{n_{ij}-r_{ij}-1}\log(1-\gamma_{1i}+r\gamma_{2i}) - \sum_{r=0}^{n_{ij}-1}\log(1+r\gamma_{2i})\right\}, \quad (20)$$

where $\gamma_{1i} = \eta_{1i}/(\eta_{1i}+\eta_{2i})$ and $\gamma_{2i} = (\eta_{1i}+\eta_{2i})^{-1}$.

Under the assumption of a common logistic distribution for the tolerance distributions for all subpopulations, the model to incorporate the natural response in (19) may describe the relationship between the response probability and dose levels of the stimulus as

$$\gamma_{1i} = \lambda + (1-\lambda)[1+\exp(\alpha+\beta x_i)]^{-1} \quad i = 1, \ldots, k. \quad (21)$$

The point and interval estimation for ED_{50} are straightforward from (21), after the MLEs of γ_{1i} and γ_{2i} are obtained by either the Newton–Raphson method or the Nelder–Mead algorithm method (*see* **Optimization and Nonlinear Equations**). The beta-binomial model assumes that the response probability from different litters follows a beta distribution. However, one can relax this strong assumption by specifying only the mean as γ_{1i} and variance as $\gamma_{1i}(1-\gamma_{1i})\rho$, where $\rho = \gamma_{2i}/(1+\gamma_{2i})$, $i = 1, \ldots, k$. It turns out to be the robust procedure of **quasi-likelihood** [43], which can be readily carried out to estimate ED_{50} by commercial statistical software packages such as GLIM (*see* **Software, Biostatistical**).

Other approaches to the extra-binomial variation for inference of the MED include the Poisson-gamma model employed by O'Neill & O'Neill [52], and the correlated-binomial model, the logistic-normalbinomial, and the probit-normal binomial models suggested by Morgan [48].

Quantal Response over Time

In a bioassay, it is not uncommon to observe responses at some prescheduled discrete time points. These time points may be selected as design points in addition to the dose of the stimulus. Time, therefore, is another classification factor [10]. If these time points are not design points but some prescheduled observing time points, then the ED_{50} is a function of time. Suppose that an experiment consists of the administration of each of k dose levels to a group of n_i subjects whose responses are observed at J time points, $0 = t_0 < t_1 < \cdots < t_J < t_{J+1} = \infty$. Let r_{ij} be the number of subjects responding in the time interval (t_{j-1}, t_j) and let $F(t_j|d_i)$ be the probability of observing a response by time t_j at the ith dose level, $j = 1, \ldots, J+1$, $i = 1, \ldots, k$. Pack & Morgan [53] showed that the likelihood can then be written as

$$L = \prod_{i=1}^{k}\left\{\left\{\prod_{j=1}^{J}[F(t_j|d_i) - F(t_{j-1}|d_i)]^{r_{ij}}\right\} \times [1-F(t_J|d_i)]^{n_i-r_i}\right\}, \quad (22)$$

where, at the ith dose level, $F(t_j|d_i) - F(t_{j-1}|d_i)$ represents the probability of observing a response during the time interval (t_{j-1}, t_j), and $1 - F(t_J|d_i)$ is the probability of no response at the end of the experiment.

Since there is a possibility that the subjects may not be affected by the stimulus, one may choose a mixed distribution with the logistic function as the mixing proportion:

$$A(x_i) = \{1+\exp[-(\alpha+\beta x_i)]\}^{-1}. \quad (23)$$

Let F_1 denote the response-time distribution for subjects who respond to the stimulus, and F_2 be that for those unaffected by the stimulus. The response-time distribution for all subjects at the ith dose level can be expressed as

$$F(t_j|d_i) = A(x_i)F_1(t_j|d_i) + [1-A(x_i)]F_2(t_j). \quad (24)$$

Note that $F_2(t_j)$ is independent of dose. Although various forms can be chosen for F_1, Pack & Morgan [53] suggested using

$$F_1(t|d_i) = \begin{cases} 1-\{1+\lambda\exp[\psi\ln(t)-\eta_i]\}^{-1/\lambda}, & \lambda \neq 0 \\ 1-\exp\{-\exp[\psi\ln(t)-\eta_i]\}, & \lambda = 0 \end{cases} \quad (25)$$

If the model ignores the part of response-time distribution for those who are not affected by the stimulus, then the inference for ED_{50} at the end of the study is the same as that for $-(\alpha/\beta)$.

The problem of overdispersion also occurs for quantal response experiments over time; see, for example, [48, 51, 54]. Petkau & Sitter [54] suggested the use of the Dirichlet-multinomial model as an alternative approach to handling the extra variation. Suppose that there are a total of L replications of the experiment. Let R_{ijl} be the cumulative number of responses observed up to time t_j for n_{ijl} subjects in the lth replicate, receiving the ith dose level. Then, given n_{ijl}, $(r_{i1l,\ldots,}r_{iJl})$ follows a multinomial distribution with probabilities $(p_{i1l,\ldots,}p_{iJl})$, where r_{ijl} and p_{ijl} are similarly defined for the lth replicates. Furthermore, if the vector of probabilities $(p_{i1l,\ldots,}p_{iJl})$ is randomly distributed as the Dirichlet distribution, then the marginal distribution of $(r_{i1l,\ldots,}r_{iJl})$ follows the Dirichlet-multinomial. As a result, Petkau & Sitter [54] suggested that the inference for ED_{50} may be obtained from the following **Weibull** response-time distribution in the absence of a replication effect.

$$F(t_j, d_i) = 1 - \exp[-\exp(\alpha + \beta x_i)t_j\gamma] \qquad (26)$$

or

$$\theta = \frac{\ln[-\ln(0.5)] - \alpha - \gamma \ln t_j}{\beta}. \qquad (27)$$

Hence, the estimated ED_{50} is a function of time.

O'Hara Hines & Lawless [51] considered a number of models for overdispersion, which can be incorporated into the **generalized linear model** framework for multinomial data. The idea is to incorporate various random components into the link function and use the quasi-likelihood or generalized least squares estimating equation to estimate the unknown parameters. O'Hara Hines & Lawless [51] found that, on average, the robust covariance matrix estimator based on the **generalized estimating equations** (GEEs) proposed by Liang & Zeger [40] performed well. As a result, it is recommended that the multinomial estimating equations with the robust variance estimate in the presence of extra-multinomial variation be used for quantal response over time. Alho & Valtonen [3] extended the results of the likelihood confidence interval for ED_{50} by Williams [74] to a generalized linear model with a known scale parameter, which permits **explanatory variables** other than those related to stimulus. Laurence & Morgan [39] and Morgan [48] provided a complete review regarding the advantages and drawbacks of the stochastic model for analysis of quantal response over time proposed by Puri & Senturia [55], which was subsequently extended by Diggle & Gratton [17].

Bayesian Approaches

Racine et al. [56] also proposed a number of **Bayesian** approaches for the estimation of ED_{50}. They suggested that either a **bivariate normal distribution** or independent beta distributions be considered as the possible choices for the **prior distribution** of the underlying parameters. For the probit model, both prior distributions yield a bivariate normal posterior distribution. However, **numerical integration** is usually required for the evaluation of the integral for the posterior distribution of ED_{50}. In addition, under the uniform and normal prior distribution, Grieve [26] derived the posterior probabilities of a substance belonging to a predetermined toxicity classes.

Grieve [27] further examined the relationship between Fieller's theorem, likelihood methods, and Bayesian methods for interval estimation of the LD_{50}. In particular, the quadratic equation of θ has its minimum at the MLE, $-(a/b)$, its maximum at $(av_{12} - bv_{11})/(bv_{12} - av_{22})$, and has an asymptote with a value, b^2/v_{22}. Therefore, if $z^2_{\alpha/2}$ is between the asymptote and its maximum, the $(1-\alpha)100\%$ confidence interval for θ is in the form of two disjoints intervals. On the other hand, if $z^2_{\alpha/2}$ is greater than the maximum, the $(1-\alpha)100\%$ confidence interval for θ is the whole real line. For the likelihood ratio interval estimate of LD_{50}, Grieve [27] showed that the likelihood function for θ always has both a minimum and a minimum. Consequently, the log-likelihood function of θ has exactly the same characteristic form as does Fieller's quadratic function shown above. Hence, there will always be a chance of α that the $(1-\alpha)100\%$ confidence interval for θ by the likelihood method will comprise the whole real line. To overcome this common problem shared by both Fieller's theorem and the likelihood ratio method, under the assumption that θ and β are **orthogonal** Grieve proposed to use the conditional **profile likelihood** (CPL) approach to interval estimation of LD_{50}. In other words, a likelihood ratio statistic is constructed from the conditional distribution of the data, given the MLE of β. The resulting CPL has the form

of the original likelihood, modified by the observed information for β, given θ. It follows that the CPL method takes into account the uncertainty regarding β, which is ignored in the likelihood approach. In addition, if the prior distribution for θ and β has the form

$$p(\theta, \beta) = \beta, -\infty < \theta < \infty, 0 < \beta, \tag{28}$$

the posterior distribution of θ derived from the Bayesian method developed in [26] is approximately the same as the conditional profile likelihood if θ and β are *a priori* independent as demonstrated by Grieve [27]. Grieve claimed that from two distinct perspectives, attempts to overcome the difficulties associated with interval estimation of LD_{50} lead to Bayesian solutions. More literature on Bayesian inference of ED_{50} can be found in [25, 45, 66, 67].

Nonparametric and Robust Methods

In the **pharmaceutical industry**, quantal response bioassays are routinely performed to examine the efficacy and safety of new drugs in animals during the early stage of drug development. It is then necessary to obtain an initial estimate of ED_{50} and its standard error with reasonable accuracy and precision. For this purpose, **nonparametric methods** are often employed for pilot bioassays to provide **robust** estimates to plan subsequent designs in the establishment of the final estimate of ED_{50}.

Under the independent binomial model (1), the unrestricted MLE for P_i is the observed proportion of the number of subjects at the ith dose level, $p_i = r_i/n_i$. To obtain nonparametric and robust estimates for ED_{50}, we may start with the MLEs of P_i under the order restriction, $P_1 \le P_2 \le \cdots \le P_k$ (*see* **Isotonic Regression**). Barlow et al. [5] gave the distribution-free MLEs of P_i under the order restriction as follows:

$$\tilde{P}_i = \max_{1<u<i} \min_{i<v<k} \left(\frac{\sum_{j=u}^{v} r_j}{\sum_{j=u}^{v} n_j} \right). \tag{29}$$

Spearman–Kärber Estimator

In 1908, Spearman [66] first proposed a simple and yet easily understood method for the estimation of ED_{50}, which was reintroduced by Kärber [38] in 1931. Let Δ_i denote the increment of the dose from dose level x_i to dose level x_{i+1}, then the Spearman–Kärber (S–K) estimator for ED_{50} is given as

$$\hat{\theta} = \left(p_1 - \frac{\Delta_i}{2}\right) + \sum_{i=1}^{k-1} (p_{i+1} - p_i) \left(x_i + \frac{\Delta_i}{2}\right) + (1 - p_k)\left(x_k + \frac{\Delta_{k-1}}{2}\right), \tag{30}$$

where $\Delta_i = x_{i+1} - x_i$. If $p_1 = 0$ and $p_k = 1$, then the Spearman–Kärber estimator in (30) reduced to its usual form, that is,

$$\hat{\theta} = \sum_{i=1}^{k-1} (p_{i+1} - p_i) \left[\left(\frac{x_{i+1} + x_i}{2}\right)\right]. \tag{31}$$

The S–K estimator for ED_{50} is in fact the area under the response probability-time curve calculated by the usual trapezoidal rule. Since the unrestricted MLEs, p_i, may not be monotonically increasing, we need to smooth p_i by using the MLEs given in (29), obtained under the order restriction before calculation of the S–K estimate of ED_{50}. The variance of the S–K estimator is

$$V(\hat{\theta}) = \sum_{i=2}^{k-1} \frac{P_i(1 - P_i)(x_{i+1} - x_i)^2}{4n_i}, \tag{32}$$

with an **unbiased** estimator,

$$v(\hat{\theta}) = \sum_{i=2}^{k-1} \frac{p_i(1 - p_i)(x_{i+1} - x_i)^2}{4(n_i - 1)}. \tag{33}$$

Morgan [48] indicated that although the S–K estimator is simple with an explicit expression, and is a function of the **sufficient statistics** $(r_1, \ldots, r_k)$, it is not unbiased for ED_{50}.

The robustness of the S–K estimator can be further improved by the use of **trimming** suggested by Hamilton in [29] and [31]. Sanathanan et al. [58] introduced the use of trimming in the logit and probit models. They suggested that the trimming χ^2 criterion be the heterogeneity factor, X^2/df, where $X^2$is the test statistic for the goodness-of-fit defined in (11). They also recommended minimizing the heterogeneity factor iteratively over the range of dose levels such that the fitted proportion is within the range of 0.0001 and 0.999 for the calculation of the information matrix. However, Morgan [48] indicated that for

quantal response data with few occurrences of 0% or 100% response, the trimmed logit procedure suggested by Sanathanan et al. [58] has very little effect. For other evaluations of S–K estimators and trimming procedures, see [29–32, 35, 48, 58].

For reference to other simple estimators of the ED_{50}; *see* **Biological Assay, Overview**.

L-, M-, and R-Estimators

For the discussion in this section, following James et al. [35], we consider a bioassay experiment in which n subjects are randomly selected to receive each of the $2k+1$ equally spaced dose levels, $x_{-k}, \ldots, x_{-1}, x_0, x_1, \ldots, x_k$. Let P_i and p_i be defined as in the previous subsection, and let Δ denote the common dose increment. Under the convention that $p_{-k-1} = 0$ and $p_{k+1} = 1$, the empirical tolerance distribution is the following piecewise linear function:

$$F(x) = \begin{cases} p_i, & \text{if } x = x_i, -k \le i \le k, \\ 0, & \text{if } x \le x_{-k-1} = x_0 - (k+1)\Delta, \\ 1, & \text{if } x \ge x_{k+1} = x_0 + (k+1)\Delta, \end{cases} \tag{34}$$

and $F(x)$is linear and continuous in $[x_i, x_{i+1}]$, for all i.

Let $J(u)$ be a nonnegative function defined on the interval [0, 1], which is symmetric about 0 and let $\int J(u)du = 1$. The L-estimator of ED_{50}, a linear combination of **order statistics**, is given as

$$\hat{\theta} = x_0 + \Delta \sum_{i=-k}^{k+1} i\, J(p_i)(p_i - p_{i-1}) \tag{35}$$

The S–K estimator is a special example of the L-estimator because (i) when $J(.) = 1$, the L-estimator is the untrimmed S–K estimator and (ii) when $J(u) = 1/(1-2a)$, where $a \le u \le (1-a)$, the L-estimator is the $100a\%$ trimmed S–K estimator.

The M-estimator of ED_{50} is the root of the following equation [45].

$$\sum_{i=-k}^{k+1} \psi\left[\frac{d_i - \theta}{s}\right][p_i - p_{i-1}] = 0, \tag{36}$$

where $\psi(.)$ is a suitable function and s is a scaling factor.

If $\psi(x) = x(1-x^2)^2$, 0 for $|x| \le 1$ or $|x| > 1$, respectively, then we have Tukey's biweight M-estimator of ED_{50}.

Let $G(u)$ be a nondecreasing integrable score function for $0 < u < 1$ and $G(1-u) = -G(u)$. The R-estimator for ED_{50} [35] is the solution to the following equation:

$$h(F,\theta) = \int G\left\{\frac{[F(x) + 1 - F(2\theta - 1)]}{2}\right\} \mathrm{d}F(x) = 0 \tag{37}$$

if a unique solution exists. If not, define

$$\hat{\theta} = \frac{\{\sup[\theta : h(F,\theta) > 0] + \inf[\theta : h(F,\theta) < 0]\}}{2}. \tag{38}$$

The **sign test** score function ($G(u) = -1$, for $u < 0.5$ and $G(u) = +1$ for $u > 0.5$), the Wilcoxon test score function ($G(u) = (u - 1/2)$ (*see* **Wilcoxon Signed-rank Test**), and the Van der Waerden score function ($G(u) = \Phi^{-1}(u)$) will give, respectively, the sample median, the Hodges–Lehmann estimator, and the normal score estimator. If $G(u) = \ln[u/(1-u)]$, then, according to James et al. [35], the resulting R-estimator is called the *logistic score estimator*. However, the R-estimator is not a **consistent estimator** for ED_{50}, and hence it is asymptotically biased.

James & James [34] defined the influence curve for the estimators of the MED in quantal bioassay under the assumption of a symmetric tolerance distribution. They also obtained the influence curve for L-, M-, and R-estimators as well as the logistic score estimator. In general, the logistic score estimator and the MLE under the logit model are not robust, while the Tukey biweight M-estimator and the Hodges–Lehmann R-estimator are robust. James et al. [35] gave a complete review of **asymptotic relative efficiency**, for different estimators of ED_{50}, under various tolerance distributions. From their paper, and from [29, 37, 48,], a moderately trimmed S–K estimator, say by 5%, may be the recommended estimator (*see* **Robustness**).

Study Design

The basic issues for designing a nonsequential bioassay to estimate the MED involve (i) selection of the dose levels, (ii) determination of the number of dose levels, (iii) estimation of the number of the subjects, and (iv) distribution of the number of subjects to a

fixed number of dose levels. Most of research concentrates on (i), (ii), and (iv), summarized below. For details on design issues about quantal response, see [2, 12, 22, 23, 36, 44, 46, 48, 49, 63, 65, 71, 77].

With respect to the estimation of ED_{50} for a symmetric tolerance distribution, various criteria can be obtained under the requirements for D-, A-, E-, G-,and F-optimality [19, 62] (*see* **Optimal Design**). All these criteria depend on the asymptotic variance of the MLEs, a and b, and hence on the information matrix of *a and b*. For example, for the D-optimality, the design is, in fact, chosen at the dose level to maximize the determinant of the inverse of information matrix, that is, to select x to maximize

$$w_D(x) = \left\{ \frac{xf^2(x)}{F(x)[1 - F(x)]} \right\}^2. \qquad (39)$$

For a logistic tolerance distribution and a two-point design, w has its maxima at $x = \pm 1.5434$. These values correspond to the response probability of 0.176 and 0.824. Consequently, the D-optimal design for a logistic distribution allocates half the subjects separately to the dose levels $ED_{17.6}$, and the other half to $ED_{82.4}$. Similarly, for the probit model, the D-optimal doses are $ED_{12.8}$ and $ED_{87.2}$.

The A-optimal design requires minimizing the trace of the asymptotic covariance matrix of a and b. The F-optimality minimizes the squared half-length of the $100(l - \alpha)\%$ confidence interval for ED_{50} on the basis of Fieller's theorem [2, 22, 23, 65]. Sitter & Wu [65] provided some numerical results for design points of dose levels for two-point and three-point designs on the basis of these criteria.

Kalish [36] considered D-optimality with the use of a second-order approximation of the order of $1/n^2$ for estimating the variances of MLEs of the MED and the slope under the formulation of the tolerance distribution in (8). She referred to these two designs as LD_{50}-optimal and slope-optimal designs. Kalish [36] recommended the equally weighted three-point design, with design points at LD_{20}, LD_{50}, and LD_{80}. From theoretical and simulation results, this design is very efficient for estimating LD_{50} and for global estimation of the dose–response curve. Sun & Tsutakawa [67] proposed Bayesian designs that avoid experiment results with little information at the expense of a small sacrifice in the Bayes risk. On the other hand, Minkin & Kundhal [47] considered the length of likelihood-based confidence interval as a criterion for the dose allocation.

From a practical viewpoint, a useful optimal design should provide the minimal number of doses and the minimal number of subjects required at each level. However, the design points derived under various optimal criteria are functions of unknown parameters. Consequently, in practice, it is extremely difficult to implement these optimal designs, not only because the number of dose levels might be unknown, but also because reliable and good initial estimates for the unknown parameters are usually unavailable during the planning stage of experiments. In this situation, Müller & Schmitt [49] recommended, for a symmetric distribution, to choose as many dose levels as possible with the allocation of one subject per dose level.

Sitter [63] used the **minimax** principle to find robust designs by minimizing the maximum of some criteria over a bounded rectangular region of possible parameters, which have the form

$$P = \{(\theta, \beta) : |\theta - \theta_0| \le \theta_\Delta, \beta_L < \beta < \beta_U\} \qquad (40)$$

where θ_Δ and (β_L, β_U) define the possible ranges, respectively, for ED_{50} and slope. Suppose that the dose levels are symmetric about the location of the tolerance distribution, θ_0, then the designs are completely specified by the number of dose levels k and the common increment between adjacent dose levels, d. Let

$$\begin{aligned} d_i &= d\beta_L \left[\frac{i - (k+1)}{2} \right], \\ x_i &= \frac{d_i}{\beta_L + \theta_o}, \\ a &= \beta_L(\theta - \theta_0), \\ b &= \frac{\beta}{\beta_L}, \\ d' &= \beta_L d. \end{aligned} \qquad (41)$$

Sitter [63] suggests a design such that

$$W(\delta) = \min_{k=2,3,\ldots} \min_{-\infty < d' < \infty} \max_S R(k, d', a, b), \qquad (42)$$

where $R(.)$ can be the criterion based on either D-optimality or F-optimality, and

$$S = \left[\frac{(a, b) : 0 \le a \le \theta_\Delta; 1 \le b \le \beta_U}{\beta_L} \right].$$

Sitter [63] provides tables for the number of dose levels and dose increments, for a number of various

combinations of θ_Δ and β_U/β_L under a logit model, and for robust designs using D-optimality and F-optimality. These proposed designs are robust to poor initial estimates of the unknown parameters and only require prior information about the possible ranges for the location and scale parameters. The minimax procedure provides the number of doses as well as the dose increment for an easy and practicable implementation of the design.

Sequential Procedures

The results of the classical fixed bioassays might either be that all subjects respond or that none responds, if the investigators misjudge the location and/or the scale of the tolerance distribution. Consequently, little information about ED_{50} can be gained and all resources are wasted. Therefore, sequential procedures provide attractive alternatives to the classical designs. The use of sequential procedures in estimating the ED_{50} goes back to Bartlett [6]. However, Dixon & Mood [18] first proposed the **up-and-down method** for the estimation of ED_{50}.

For an up-and-down experiment, after a series of equally spaced dose levels have been selected, the first subject is tested at the best guessed dose level for ED_{50}. Each subsequent test is performed at the next lower or the next higher dose level according to which the predefined response is or is not observed in the previous subject. Let x_i, where $i = 0, 1, \ldots, n$, be the dose level used in the ith trial of an up-and-down experiment, with x_0 being the initial dose level and d being the common fixed spacing between adjacent dose levels. Dixon & Mood [18] gave the following estimator:

$$\hat{\theta}_M = \sum_{i=0}^{k} \frac{x_i}{k+1}, \tag{43}$$

which was modified by Brownlee et al. [11] as

$$\hat{\theta}_B = \sum_{i=1}^{k+1} \frac{x_i}{k+1}. \tag{44}$$

Choi [12] further provided a different estimator for ED_{50} on the basis of the up-and-down experiment as

$$\hat{\theta}_C = \sum_{S} \frac{w_i}{t-1}, \tag{45}$$

where S denotes the t dose levels where a change in response at x_i occurs and

$$w_i = \begin{cases} \dfrac{x_i + \mathrm{d}}{2}, & x_i \text{ is a trough} \\ \dfrac{x_i - \mathrm{d}}{2}, & x_i \text{ is a peak} \end{cases}$$

Choi [12] also used **Markov-chain** theory for estimation of the dispersion matrix for confidence intervals of ED_{50} based on the up-and-down experiment.

In general, the $n+1$ dose level for the up- and-down experiment can be expressed as

$$x_{n+1} = x_n - 2d(I_n - 0.5), \tag{46}$$

where I_n is 1 if the response is observed at the nth dose level and is 0 otherwise. This leads to the sequential Robbins–Monro [57] method for estimating the ED_{50}, which is given as

$$\hat{\theta}_{RM} = \frac{x_n - c}{n(I_n - 0.5)} \tag{47}$$

(*see* **Stochastic Approximation**). In (47), the optimal value of c is chosen to be $[f(\theta)]^{-1}$ to minimize the asymptotic variance of $[n(I_i - \theta)]^{1/2}$. Let b_n be the estimated slope resulting from fitting a linear regression of I_i on x_i with the available results based on the current n subjects. Then, the adaptive Robbins–Monro estimator after the nth trial is given as

$$\hat{\theta}_{ARM} = x_n - \left(\frac{1}{nb_n}\right)(I_n - 0.5). \tag{48}$$

Wu [77] proposed a logit-MLE method to use all available information to date for determination of the $n+1$ dose level. Suppose that the study has completed an initial design, either fixed or sequential, with n subjects at n dose levels. The next dose level is then given as

$$x_{n+1} = x_n - \left(\frac{k_n^*}{n}\right)(I_n - 0.5), \tag{49}$$

where $k_n^* = \max[c_1, \min(k_n, c_2)]$, and c_1 and c_2 are some truncation constants. Wu [77] suggested the choice of $c_1 = 0$ to avoid possible large changes in dose levels. He reported that the relative performance of these methods heavily depends on the choice of the initial design. If n is rather large, then MLE or $\hat{\theta}_{ARM}$ should be used to take advantage of the asymptotic optimality. A large value of the constant c should be selected if the initial guess of ED_{50} is poor. If

only scant information about θ and $f(\theta)$ is available at the planning stage of the experiment for the initial design, then one needs to use a wide range of dose levels that should be evenly placed.

McLeish & Tosh [44] extended Wu's logit-MLE method in a manner such that, after k dose levels in n subjects have been tested, the next dose level is chosen so as to minimize the asymptotic variance of MLE for the θ obtained from the updated information matrix by the delta method. This method not only allows flexibility of the incorporation of cost but is also more robust against misspecification of the parameters than Wu's logit-MLE procedure. In addition, McLeish and Tosh's method for allocating dose levels seems to achieve full asymptotic efficacy of estimation.

References

[1] Abbott, W.S. (1925). A method of computing the effectiveness of an insecticide, *Journal of Economic Entomology* **18**, 265–267.

[2] Abdelbasit, K.M. & Plackett, R.L. (1983). Experimental design for binary data, *Journal of the American Statistical Association* **78**, 90–98.

[3] Alho, J.M. & Valtonen, E. (1995). Interval estimation of inverse dose-response, *Biometrics* **51**, 491–501.

[4] Ashford, J.R. (1985). Quantal response analysis, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N. Johnson, eds. Wiley, New York, pp. 402–406.

[5] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, London.

[6] Bartlett, M.S. (1946). A modified probit technique for small probabilities, *Journal of the Royal Statistical Society, Supplement* **8**, 113–117.

[7] Berkson, J. (1944). Application of the logistic function to bioassay. *Journal of the American Statistical Association* **39**, 357–369.

[8] Berkson, J. (1955). Maximum likelihood and minimum likelihood; f2 estimates of the logistic function, *Journal of the American Statistical Association* **50**, 130–162.

[9] Berndt, E.R., Hall. B.H., Hall, R.E. & Hausman, J.A. (1974). Estimation and inference in non-linear structural models, *Annals of Economics and Social Management* **3**, 653–665.

[10] Boyce, C.B.C. & Willaims D.A. (1967). The influence of exposure time on the susceptibility of Australorbis glabratus to N-tritylmorpholine, *Annals of Tropical Medicine and Parasitology* **61**, 15–20.

[11] Brownlee, K.A., Hodges, J.L. & Rosenblatt, M. (1953). The up-and-down method with small samples, *Journal of the American Statistical Association* **48**, 262–277.

[12] Choi, S.C. (1990). Interval estimation of the LD_{50} based on an up-and-down experiment, *Biometrics* **46**, 485–492.

[13] Copenhaver, T.W. & Mielke, P.W. (1977). Quantit analysis: a quantal assay refinement, *Biometrics* **33**, 175–186.

[14] Cox, C. (1990). Fieller's theorem, the likelihood and the delta method, *Biometrics* **46**, 709–718.

[15] Crowder, M.J. (1978). Beta-binomial ANOVA for proportions, *Applied Statistics* **27**, 34–37.

[16] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[17] Diggle, P.J. & Gratton, R.J. (1984). Monte Carlo methods of inference for implicit statistical methods, *Journal of the Royal Statistical Society, Series B* **46**, 193–227.

[18] Dixon, W.J. & Mood, A.M. (1948). A method for obtaining and analyzing sensitivity data, *Journal of the American Statistical Association* **43**, 109–126.

[19] Fedorov, V.V. (1972). *The Theory of Optimum Experiments*. Academic Press, New York.

[20] Fieller, E.C. (1944). A fundamental formula in the statistics of biological assay, and some applications, *Quarterly Journal of Pharmacology* **17**, 117–123.

[21] Fieller, E.C. (1954). Some problems in interval estimation, *Journal of the Royal Statistical Society, Series B* **16**, 175–185.

[22] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, London.

[23] Finney, D.J. (1978). *Statistical Method in Biological Assay*, 3rd Ed. Griffin, London.

[24] Gaddum, J.H. (1933). Reports on Biological Standards. III. Methods of Biological Assay Depending on a Quantal Response. Medical Research Council, Special Report Series, no. 183.

[25] Govindarajulu, Z. (1988). *Statistical Techniques in Bioassays*. Karger, Basel.

[26] Grieve, A.P. (1988). A Bayesian approach to the analysis of LD50 experiments, in *Bayesian Analysis*, 3rd Ed., Bernardo, J.M., deGroot, M.H., Lindley, D.V. & Smith, A.F.M., eds. Oxford University Press, Oxford.

[27] Grieve, A.P. (1996). On likelihood and Bayesian methods for interval estimation of the LD_{50}, in *Statistics in Toxicology – A Volume in Memory of David A. William*, B.J.T. Morgan, ed. Oxford University Press, Oxford, pp. 87–100.

[28] Griffiths, W.E., Hill, R.C. & Poper, P.J. (1987). Small sample properties of probit model estimators, *Journal of the American Statistical Association* **82**, 929–937.

[29] Hamilton, M.A. (1979). Robust estimates of the ED50, *Journal of the American Statistical Association* **74**, 344–354.

[30] Hamilton, M.A. (1980). Inference about, the ED_{50} using the trimmed Spearman-Karber procedure - a Monte Carlo investigation, *Communications in Statistics B***9**, 235–245.

[31] Hamilton, M.A., Russo, P.C. & Thurston, R.V. (1977). Trimmed Spearman-Karber method for estimating median lethal concentrations in toxicity bioassays, *Environmental Science and Technology* **11**, 714–719. (Correction, **12** (1978), 417).

[32] Hoekstra, J.A. (1989). Estimation of the ED_{50}, *Biometrics* **45**, 337–338.

[33] Huber, J.J. (1992). *Bioassay*, 3rd Ed. Kendall/Hall, Dubuque.

[34] James, B.R. & James, K.L. (1983). On the influence curve for quantal bioassay, *Journal of Statistical Planning and Inference* **8**, 331–345.

[35] James, B.R., James, K.L. & Wastenberger, H. (1984). An efficient R-estimator for the ED_{50}, *Journal of the American Statistical Association* **79**, 164–173.

[36] Kalish, L.A. (1990). Efficient design for estimation of median lethal dose and quantal dose-response curves, *Biometrics* **46**, 737–748.

[37] Kappenman, R.F. (1987). Nonparametric estimation of dose-response curves with application to ED_{50} estimation, *Journal of Statistical Computation and Simulation* **28**, 1–13.

[38] Karber, G. (1931). Beitrag zur kollektieven Behandlung pharmakologischer Reihenversuche, *Archiv fur Experimentelle Pathologic and Pharmakologie* **162**, 480–487.

[39] Laurence, A.F. & Morgan, B.J.T. (1989). Observations on a stochastic model for quantal assay data, *Biometrics* **45**, 733–744.

[40] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[41] Lwin, T. & Martin, P.J. (1989). Probits of mixtures, *Biometrics* **45**, 721–732.

[42] Mantel, N. (1985). Reader reaction: maximum likelihood vs. minimum chi-square, *Biometrics* **41**, 777–780.

[43] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[44] McLeish, D.L. & Tosh, D.H. (1990). Sequential designs in bioassay, *Biometrics* **46**, 103–116.

[45] Miller, R.G. & Halpern, J.W. (1980). Robust estimators for quanta! bioassay, *Biometrika* **67**, 103–110.

[46] Minkin, S. (1987). On optimal design for binary data, *Journal of the American Statistical Association* **82**, 1098–1103.

[47] Minkin, S. & Kundhal, K. (1999). Likelihood-based experimental design for estimation of ED_{50}, *Biometrics* **55**, 1030–1037.

[48] Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. Chapman & Hall, London.

[49] Müller, H.G. & Schmitt, T. (1990). Choice of number of doses for maximum likelihood estimation of the ED_{50} for quantal dose-response data, *Biometrics* **46**, 117–130.

[50] Nelder, J.A. & Mead, R. (1965). A simplex method for function minimisation, *Computation Journal* **7**, 308–313.

[51] O'Hara Hines, R.J. & Lawless, J.F. (1993). Modeling overdispersion in toxicological mortality data groups over time, *Biometrics* **49**, 107–122.

[52] O'Neill, T.J. & O'Neill, H.C. (1993). A gamma model for extra-binomial variation in dilution assays, *Biometrics* **49**, 237–242.

[53] Pack, S.E. & Morgan, B.J.T. (1990). A mixture model for interval-censored time-to-response quantal assay data, *Biometrics* **46**, 749–758.

[54] Petkau, A.J. & Sitter, R.R. (1989). Models for quantal response experiments over time, *Biometrics* **45**, 1299–1306.

[55] Pun, P.S. & Senturia, J. (1972). On a mathematical theory of quantal response assays, *Proceeding of 6th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. University of California Press, Los Angeles, pp. 231–247.

[56] Racine, A., Grieve, A., Flühler, H. & Smith, A.F.M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry, *Applied Statistics* **35**, 93–150.

[57] Robbins, H. & Monro, S. (1951). A stochastic approximation method, *Annals of Mathematical Statistics* **22**, 400–407.

[58] Sanathanan, L.P., Gade, E.T. & Skipkowitz, N.L. (1987). Trimmed logit method for estimating the ED50 in quantal bioassay, *Biometrics* **43**, 825–832.

[59] Segreti, A.C. & Munson, A.E. (1981). Estimation of the median lethal dose when responses within a litter are correlated, *Biometrics* **37**, 153–154.

[60] Silvapulle, M.J. (1981). On the existence of maximum likelihood estimators for the binomial response model, *Journal of the Royal Statistical Society, Series B* **43**, 310–313.

[61] Silverstone, H. (1957). Estimating the logistic curve, *Journal of the American Statistical Association* **52**, 567–577.

[62] Silvey, S.D. (1980). *Optimal Design*. Chapman & Hall, London.

[63] Sitter. R.R. (1992). Robust designs for binary data, *Biometrics* **48**, 1145–1176.

[64] Sitter, R.R. & Wu, C.F.J. (1993). On the accuracy of Fieller intervals for binary response data, *Journal of the American Statistical Association* **88**, 1021–1025.

[65] Sitter, R.R. & Wu, C.F.J. (1993). Optimal designs for binary response experiments; Fieller, D, and A criteria, *Scandinavian Journal of Statistics* **20**, 329–341.

[66] Spearman, C. (1908). The method of "right and wrong cases" ("constant stimuli") without Gauss's formulae, *British Journal of Psychology* **2**, 227–242.

[67] Sun, D. & Tsutakawa, R.K. (1997). Bayesian design for dose-response curves with penalized risk, *Biometrics* **53**, 1262–1273.

[68] Taylor, W.F. (1953). Distance functions and regular best asymptotically normal estimates, *Annals of Mathematical Statistics* **24**, 85–92.

[69] Trevan, J.W. (1927). The error of determination of toxicity, *Proceedings of the Royal Society of London, Series B* **101**, 483–514.

[70] Tsutakawa, R.K. (1972). Design of an experiment for bioassay, *Journal of the American Statistical Association* **67**, 584–590.

[71] Tsutakawa, R.K. (1980). Selection of dose levels for estimating a percentage point of a logistic quantal response curve, *Applied Statistics* **29**, 25–33.

[72] Tsutakawa, R.K. (1985). Bioassay, statistical methods, in *Encyclopedia of Statistical Sciences*, Vol. 1, S. Kotz& N. Johnson, eds. Wiley, New York, pp. 236–243.

[73] USP/NF. (1990). *The United States Pharmacopeia XXII and the National Formulary XVII*. The United States Pharmacopeial Convention, Rockville.

[74] Williams, D.A. (1986). Interval estimation of the median lethal dose, *Biometrics* **42**, 641–646.

[75] Williams, D.A. (1988). Reader reaction: estimation bias using the beta-binomial distribution in teratology, *Biometrics* **44**, 305–307.

[76] Williams, D.A. (1989). Hypothesis tests for overdispersed generalised linear models, *GLIM Newsletter* **37**, 29–39.

[77] Wu, C.F.J. (1985). Efficient sequential designs with binary data, *Journal of the American Statistical Association* **80**, 974–984.

JEN-PEI LIU & SHEIN-CHUNG CHOW

Median Survival Time

A useful summary of a survival curve is the median survival time. Approximately 50% of the population under study could be expected to survive beyond the **median**. More formally, the median survival time is defined as $M = F^{-1}(1/2) = \inf(t : F(t) \geq 0.5)$, where $F(t)$, the cumulative distribution function, is the probability of surviving less than or equal to time t. The median survival is used considerably more frequently than the mean as a summary statistic because of the difficulties in estimating means from heavily right-censored data without strong parametric assumptions.

The median is nonparametrically estimated from right-censored survival data by first calculating the Kaplan–Meier [8] survival curve $\hat{S}(t) = 1 - \hat{F}(t)$ (*see* **Kaplan–Meier Estimator**), which is an estimate of the probability of surviving beyond t. The estimated median survival time is $\hat{M} = \hat{F}^{-1}(1/2)$ which is the smallest observed event (uncensored) time where the Kaplan–Meier estimate is not greater than 1/2. There can be a great deal of variability in the estimated median if the survival curve is relatively flat near 0.5.

Several approaches have been proposed to obtain confidence intervals for the median survival. Brookmeyer & Crowley [3] suggested inverting a generalized sign test for right-censored data for testing the null hypothesis H_0: median survival time $= t$, versus H_1: median survival time $\neq t$. The generalized sign test statistic, $\hat{S}(t)$, is the Kaplan–Meier estimate evaluated at t. One does not reject the null hypothesis at level α if

$$[\hat{S}(t) - 1/2]^2 \leq \chi^2_\alpha \widehat{\text{var}}[\hat{S}(t)],$$

where χ^2_α is the α critical value of a χ^2 with one degree of freedom and where $\widehat{\text{var}}(\hat{S}(t))$ is Greenwood's estimate of the variance of $\hat{S}(t)$,

$$\widehat{\text{var}}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{\{x_i \leq t\}} \frac{d_i}{n_i(n_i - d_i)},$$

where d_i and n_i are the number of uncensored events and number at risk, respectively, at distinct event times x_i. The $(1 - \alpha)$ 100% confidence interval is defined as the interval $I_\alpha = [t_l, t_u)$, where t_l is the smallest event (uncensored) time with $\hat{S}(t) \geq 0.5$ that is not rejected by the generalized sign test at level α and t_u is the smallest observed event (uncensored) time with $\hat{S}(t) < 0.5$ that is rejected at level α. Occasionally it happens that an upper confidence limit cannot be obtained because the last observed event time is in I_α, then I_α becomes a one-sided confidence interval of the form $[t_l, \infty]$. We also obtain a one-sided confidence interval if the Kaplan–Meier survival curve does not reach the median because of extensive censoring.

This confidence interval method can also be illustrated by the following approach [1]. First, calculate pointwise $(1 - \alpha)$ 100% confidence intervals for the entire survival curve using the Kaplan–Meier estimate and Greenwood's formula, $\hat{S}(t) \pm Z_{\alpha/2}[\widehat{\text{var}}\hat{S}(t)]^{1/2}$. Then the confidence interval for the median is defined by the times where these upper and lower confidence limits equal 1/2. This is shown graphically in Figure 1 for survival data from a colorectal cancer clinical trial consisting of 53 patients of whom 16 were censored [3]. The figure shows the Kaplan–Meier estimate along with the upper and lower 95% confidence intervals for $S(t)$. The median survival time was $\hat{M} = 61$ weeks and the 95% confidence interval was [38, 73].

Several other related confidence interval approaches have also been proposed (see, for example, [6, 12, 13], and [7]). Several researchers have suggested replacing Greenwood's estimate $\widehat{\text{var}}\hat{S}(t)$ by various estimators of the null variance of $\hat{S}(t)$ under the null hypothesis that t is the true median [3, 7, 13]. Extensive simulation studies to investigate the performance of these confidence interval procedures indicate that, in small samples ($N \leq 20$), use of

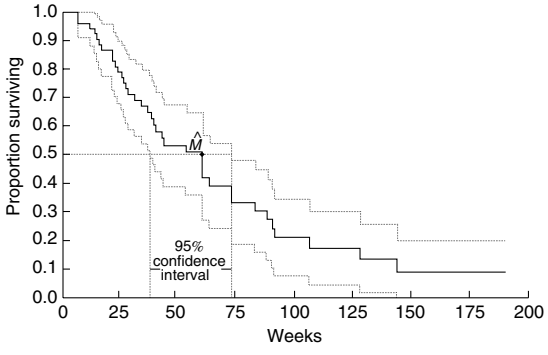

Figure 1 Illustration of confidence interval procedure for the median survival time using survival data from a colorectal cancer clinical trial [3]

the null variance may perform better than use of Greenwood's estimate with respect to coverage probabilities [7]. Pointwise confidence intervals for the survival curve based on $\log[-\log \hat{S}(t)]$ or $\arcsin\{[\hat{S}(t)]^{1/2}\}$ **transformations** have been found to perform well [2, 14]. This suggests that transformation of $\hat{S}$ may also improve the small-sample performance of confidence intervals for the median. Bootstrapped confidence interval procedures have also been described [5, 11] and give similar results. Confidence interval procedures for the difference between two medians from right-censored survival data have been described in [15]. Repeated confidence intervals for the median survival time with accumulating data have been proposed by Jennison & Turnbull [7].

To test the equality of k medians from uncensored data, the classical median test can be used [9, 16]. The median test for uncensored data consists of pooling the observations from k samples, determining the median of the pooled sample, then counting the number A_i in the ith sample that exceed the pooled median. The test statistic is

$$4\sum_{i=1}^{k}\frac{(A_i - n_i/2)^2}{n_i},$$

where n_i is the number of observations in the ith sample, and asymptotically has a χ^2 distribution with $k-1$ degrees of freedom. One generalization to right-censored data [4], involves defining the "pooled sample median" $\hat{M}$ as the median of a weighted average of the individual Kaplan–Meier estimates from each sample that is weighted by the relative sample sizes. The statistic is based on the deviation from 1/2 of each individual Kaplan–Meier estimate evaluated at the pooled median $\hat{M}$. We define $\mathbf{z} = \{\sqrt{N}[\hat{F}_i(\hat{M}) - 1/2]\}$, where $N = \Sigma n_i$, then the median test statistic for censored data is of the form $\mathbf{z}' = \hat{\mathbf{\Sigma}}^{-1}\mathbf{z}$, where $\hat{\mathbf{\Sigma}}^{-1}$ is a generalized inverse [4]. Under H_0, the test statistic is $\chi^2(k-1)$. An alternative generalization, proposed by Prentice [10], was based on a general family of linear rank tests for censored data and is particularly powerful for detecting location shifts in the double exponential distributions. Either version of the median test for censored survival data is especially sensitive for detecting differences in medians in location shifts. Other test statistics, such as the **logrank test**, will be more powerful against other alternatives such as the proportional hazards family.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Borgan, Ø. & Liestøl, K. (1990). A note on confidence intervals and bands for the survival function based on transformations, *Scandinavian Journal of Statistics* **17**, 35–41.

[3] Brookmeyer, R. & Crowley, J. (1982). A confidence interval for the median survival time, *Biometrics* **38**, 29–41.

[4] Brookmeyer, R. & Crowley, J. (1982). A k-sample median test for censored data, *Journal of the American Statistical Association* **77**, 433–440.

[5] Efron, B. (1981). Censored data and the bootstrap, *Journal of the American Statistical Association* **76**, 312–319.

[6] Emerson, J. (1982). Nonparametric confidence intervals for the median in the presence of right censoring, *Biometrics* **38**, 17–27.

[7] Jennison, C. & Turnbull, B.W. (1985). Repeated confidence intervals for the median survival time, *Biometrika* **72**, 619–625.

[8] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[9] Mood, A.M. (1950). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

[10] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika* **65**, 167–179.

[11] Reid, N. (1981). Estimating the median survival time, *Biometrika* **68**, 601–608.

[12] Simon, R. & Lee, Y.J. (1982). Nonparametric confidence limits for survival probabilities and median survival time, *Cancer Treatment Reports* **66**, 37–42.

[13] Slud, E.V., Byar, D.P. & Green, S.B. (1984). A comparison of reflected versus test-based confidence intervals for the median survival time based on censored data, *Biometrics* **40**, 587–600.

[14] Thomas, D.R. & Grunkemeier, G.L. (1975). Confidence interval estimates of survival probabilities for censored data, *Journal of the American Statistical Association* **70**, 865–871.

[15] Wang, J.-L. & Hettmansperger, T.P. (1990). Two-sample inference for median survival times based on one-sample procedures for censored survival data, *Journal of the American Statistical Association* **85**, 529–536.

[16] Westenberg, J. (1948). Significance test for median and interquartile range, *Nederlandsche Koninkliijke Akademie Van Wetenschappen* **51**, 252–261.

RON BROOKMEYER

Median

The sample median of a set of ordered data is the middle observation if the sample size is odd and the average of the two middle observations if the sample size is even. In the case of even sample size, any value in the interval between the middle two data points will serve as a median, but the midpoint of the interval is generally chosen by convention.

The median of a probability distribution is a point that divides the probability distribution into two equal parts. Let $F(x)$ represent a cumulative distribution function (cdf) (*see* **Random Variable**). If F is strictly increasing at m and $F(m) = 1/2$, so $m = F^{-1}(1/2)$, then m is the unique median. If a cdf is not strictly increasing at the prospective median, then more care is required for an analytical definition. Define the inverse of a cdf as $F^{-1}(t) = \inf[x : F(x) \geq t]$. Then $m = F^{-1}(1/2)$ is the unique median or is the left endpoint of the interval of medians. When F is the empirical cdf (*see* **Goodness of Fit**), m is the middle data value or is the first of the two middle data values.

The sample median is a **robust** estimate of the population median. It is robust in the sense that it is not affected by **outliers** (bounded influence function), and it takes roughly 50% contamination of the data to ruin it (50% breakdown point). The median is approximately normally distributed in large samples and the asymptotic variance (standard error) is $1/[4f^2(\mu)n]$, where f is the probability density function (pdf) of the population, μ is the population median, and n is the sample size. Sheather [3] reviews methods of estimation of the finite sample variance and the asymptotic variance of the sample median. It is interesting that the standard error of the median can be estimated using the **bootstrap** but not the **jackknife**. In fact, the bootstrap distribution of the median can be found in closed form and does not have to be **simulated** (see Efron & Tibshirani [1]).

The sample median solves the L_1 minimization problem; that is, it minimizes $\Sigma|x_i - t|$ as a function of t (*see* **Mean Deviation**). It is the **maximum likelihood estimator** for the center of a Laplace or double exponential distribution. Hence, for the Laplace distribution, the median is asymptotically optimal. On the other hand, when the underlying distribution is normal, the **asymptotic relative efficiency** of the median relative to the mean is only 0.637. The median shares efficiency properties with the simple **sign test**.

In biostatistics it may be of interest to estimate the median of a survival distribution in a **censored data** setting (*see* **Median Survival Time**). In this situation the empirical cdf is replaced by the product limit estimate (PLE) $\hat{F}$ due to **Kaplan & Meier** [2]. Then the estimate is based on $\hat{F}^{-1}(1/2)$; see Reid [5] for a nice discussion.

The median can be defined for multivariate distributions, and then estimates can be developed from the multivariate data. Small [4] surveys the various ways that the idea of median has been extended into high dimensions under various types of equivariance requirement.

References

[1] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[2] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[3] Sheather, S.J. (1987). Assessing the accuracy of the sample median: Estimated standard errors versus interpolated confidence intervals, in *Statistical Data Analysis Based on the L_1-Norm and Related Methods* Y. Dodge, ed. North-Holland, Amsterdam, pp. 203–215.

[4] Small, C.G. (1990). A survey of multidimensional medians, *International Statistical Review* **58**, 263–277.

[5] Reid, N. (1981). Estimating the median survival time, *Biometrika* **68**, 601–608.

T. HETTMANSPERGER

Medical Devices

Introduction

A medical device is any item that treats or diagnoses a health condition but whose action is not primarily chemical or biological. Simply defined, it is any product used in medicine that is not a drug or a biological agent. The array of products that fall within the medical device category is extremely broad; they range from tongue depressors, syringes, and wheelchairs to coronary stents, DNA microarrays, and CT scanners. The diversity of the industry makes medical devices an exciting arena in which advances in a host of basic science and engineering disciplines can be brought to bear on the improvement of human health. It is also the source of many interesting research areas in the statistical sciences.

This article will focus initially on the design and analysis of therapeutic devices and implants (nondiagnostic devices), with some discussion of sham devices and **blinding**, noninferiority (*see* **Equivalence Trials**) and active control studies, **survival analysis** and repeated measures (*see* **Multiplicity in Clinical Trials**), and the use of historical controls (*see* **Bias from Historical Controls**). Attention will then turn to diagnostic devices, with emphasis on **microarrays**, and Bayesian approaches to medical device studies (*see* **Bayesian Methods in Clinical Trials**). It will conclude with some discussion of surveillance of medical devices (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**).

The Nature of Nondiagnostic Medical Device Studies

In many cases, the mechanism of action of a medical device is well understood and is local as opposed to systemic. Compared to other medical products, such as pharmaceutical drugs, medical devices tend to have a much shorter commercial life cycle. Typically, it may take only two years for a medical device to become obsolete after its first use as it is often supplanted by a newer model. In contrast, drug product lines can last 10 to 20 years or more. Devices usually evolve by a series of small changes, and the pace of invention can be very fast. Often, there is almost constant tinkering with the design and manufacturing of a medical device. Consequently, in some cases, there are a large number of models for the same, or only slightly different, indications. The fast evolution of medical devices, coupled with a built-in expectation by the public that newer is better, presents unique challenges in designing and evaluating medical device studies [92]. For some devices, a randomized **clinical trial** may not be feasible because of the difficulty in recruiting patients. Statistical challenges include coping with changes to the protocol (*see* **Clinical Trials Protocols**) and perhaps even to the new device during the course of a clinical trial to evaluate the effectiveness and safety of the device.

Placebo Effect and Sham Controls

In studies of medical devices, there are many instances in which evidence for a **placebo** effect exists, that is to say, people react differently if they are in a clinical trial, regardless of whether they know (or even strongly suspect) which treatment they are receiving. This is especially worrisome when the primary outcome measured is something like pain or function as subjectively assessed by the patient and sometimes by the physician or health care worker. The statistical issue here is, of course, bias (*see* **Bias, Overview**). One way to remove bias due to the placebo effect is to have some sort of sham (or placebo) control as a second arm in a randomized clinical trial. Therefore, the sham as a control plays an important role in the evaluation of the safety and effectiveness of medical devices, especially those devices for the treatment of pain or function [46, 61, 65, 90]. In the interpretation of study results, in addition to placebo effects, many factors could account for the apparent effect of the sham device: natural course of disease, fluctuation of symptoms, **regression to the mean**, patient bias, and physician bias.

However, compared to drug evaluations, the well-designed placebo-controlled clinical trial design is used less frequently in medical device applications because of ethical or practical reasons (*see* **Ethics of Randomized Trials**). Randomizing patients into the sham control arm may raise serious ethical concerns, especially when considerable risk is involved, such as in the case of sham surgery. As an alternative, an active control is widely used, and even a historical control (*see* **Bias from Historical Controls**) is employed when indeed appropriate. Of

course, in an active-controlled device trial, two active treatments might have different amounts of placebo effects so that an observed treatment difference might result from differences in their placebo effects rather than differences in the treatments themselves, which makes study design and interpretation more challenging.

Blinding or Masking

In a medical device study, masking (or blinding) the subjects to which treatment they are receiving is essential to control for the bias associated with knowing what treatment is being administered. Piantadosi [74] observes that **blinding or masking** helps to control the Type I error (*see* **Hypothesis Testing**). Unfortunately, in some instances involving medical devices, it may be impossible to mask the patient or the investigator/surgeon as to who receives which treatment or implant. A third party evaluator who is masked to the treatment assignment is often employed in such cases.

An interesting situation occurs when the trial is blinded, but the patients at the end of the trial are asked to guess the treatment assignment that they have received. In essence, this is a possible way to check whether the blind has been maintained. While it is possible to correctly guess the treatment in pharmaceutical trials, it is more likely in many device trials. Sometimes the symptoms one experiences are dead giveaways as to which treatment has been received. Informed consent (*see* **Ethics of Randomized Trials**) often guarantees that patients have this information. An interesting statistical problem is to figure out how to use data on patient perception of treatment assignment to assess the possible bias and, more interestingly, how to correct for it.

Comparisons of Statistical Issues for Device and Drug Trials

Medical device studies share many of the same issues that drug trials do [15]. These range from **missing data**, adjusting for interim looks, and **Data and Safety Monitoring Boards** (DSMBs) (also called Data Monitoring Committees (DMCs)) to **intention-to-treat**, **multiplicity** (of tests and endpoints), and **time-dependent covariates**. Since often there are difficulties in carrying out long-term device studies, there tend to be more missing data than one might expect for most pharmaceutical trials. In the context of medical devices, there are publications that address **surrogates** [20] and subgroup analysis (*see* **Treatment-covariate Interaction**) [86]. One topic that is of more interest perhaps in medical device clinical trials is that of the **interaction** between treatment effect (control versus new device) and center. In medical device studies, such interactions are of interest since they may be an indication that there are significant differences in how the devices are used from center to center, requiring a protocol revision or better training concerning the use of the device. Another challenging area concerns changes to the protocol or to the device during the course of the study.

Implants

Implanted medical devices pose unique challenges to the evaluation of their performance. As with most surgical procedures, there is often a learning curve; namely, the ability to implant or to use the device improves with familiarity. Implants are usually designed to be in the body for a long period of time. Therefore, there is increased importance for the statistical tools of **survival analysis** and repeated-measures analysis (*see* **Multiplicity in Clinical Trials**). Whereas the administration of drugs can be discontinued because of serious adverse events, a problematic implant may need to be explanted.

Survival Analysis

Survival analysis is frequently applied to time-to-event data in clinical studies of medical devices. For devices requiring long-term follow-up, such as implanted devices, outcomes are often expressed as probabilities or rates. For example, one might assess the 24-month fusion success/failure probabilities for spinal fusion devices and for hip and knee implants or the 6-month major adverse cardiac event rate for coronary stents. Statistical issues frequently encountered in medical device survival analysis include small sample adjustments for estimating 95% **confidence intervals** of cumulative survival probability, comparison of crude event probabilities, linearized event rates or **incidence densities**, **life-table** probability, survival experience extrapolation

and prediction beyond the last observed follow-up [50], **matched**-pair and **multivariate survival analyses**, recurrent events (*see* **Repeated Events**), and **random-effect** survival analysis with **frailty** parameters.

Since the sample sizes encountered in medical device follow-up studies are often smaller than those in drug trials, there may be fewer observed events at later, but clinically important, follow-up times. In those situations, the routinely used Greenwood formula (*see* **Median Survival Time**) available in standard statistical software underestimates the **standard error**. In [50], several statistical approaches are proposed to improve estimation of standard error and hence the 95% confidence interval.

Investigators sometimes wish to **extrapolate** or predict future survival experiences beyond the last available observation in the follow-up. For example, a two-year implanted-device survival is sometimes predicted from observed one-year data. The frequentist approach generally requires the following statistical validation procedures: parametric **model** building based on a sufficiently large number of events, parameter **estimation**, **diagnostic** checking, **prediction** or **forecasting**, and periodic model validation (*see* **Model Checking**). Clinically important **covariates** should also be investigated during parametric model building [50].

In medical device trials using the **matched-pair** design, experimental and control treatments are randomly assigned to different locations or sites within a patient, such as bilateral knees, hips, or eyes, or multiple teeth or skin locations. Owing to the correlation between two or more locations within the same patient, nonstandard survival analysis methods are required to analyze such matched-pair time-to-event data. If no censoring occurs (*see* **Censored Data**), then one may transform the time-to-event into an occurrence of an event prior to a prespecified follow-up time and analyze these data as correlated binary data; such data have been called current status data. For matched-pair survival data with censoring, which is often seen in device trials, some methods can be found in [36, 37, 56, 68].

In medical device trials, **survival analysis** is commonly applied to time-to-first-event data. However, in some cases, repeated events of the same type or different types are frequently seen. Examples are catheter restenosis for kidney dialysis patients, restenosis for cardiac patients with balloon catheterization or stent implant after catheterization, repeated air leaks for pulmonary patch, repeated device migration or infection for hip and knee implants, and repeated infections for cochlear implants. Survival analysis applied to repeated events, in addition to the first event, provides useful clinical insight into device performance over long-term follow-up. Statistical models for such data include **marginal models** [88].

Other issues, such as **time-dependent covariates**, **multivariate survival analysis**, and random-effect **frailty** models, are also relevant in medical device trials.

Repeated Measures

The repeated-measures design is common in medical device clinical studies, particularly with implanted devices for which study patients are observed at various follow-up times after initial implant, diagnosis, or randomization (*see* **Multiplicity in Clinical Trials**). Such design could occur in either a two-arm, parallel, randomized, prospective **multicenter trial** or a single-arm, prospective study. Most scheduled follow-up times are unequally spaced, such as, at baseline, 1, 3, 6, 12, and 24 months posttreatment. Types of clinical data include continuous, ordinal (such as pain score by visual analog scale), binary (such as implant success or failure), or **Poisson** count data (such as number of epileptic seizures or headaches after neurological device treatment). A repeated-measures design can also be employed in conjunction with paired data, such as those seen in ophthalmic, dental, ear, and orthopedic devices. Various statistical methods used in a repeated-measures design include repeated-measures **analysis of variance** (RMANOVA), **multivariate analysis of variance** (MANOVA), time-by-time comparisons by analysis of variance (ANOVA), general mean response or **mixed model**, **generalized estimating equations** (GEE), and profile analysis (*see* **Summary Measures Analysis of Longitudinal Data**). **Analysis of covariance** (ANCOVA) is often used to adjust treatment effects for baseline differences. Sometimes analyses are based on derived variables such as change from baseline, percentage change from baseline, binary outcome based on a clinically acceptable cutoff point such as 50% change from baseline, summary statistics such as slopes, and others. In these cases, the correlation between the individual patient baseline value

and the derived response variables should be carefully evaluated in order to choose derived response variables appropriately. The advantages and disadvantages of each of these statistical methods in the context of medical device applications are discussed in [51].

Subjects with **missing data** are a frequent occurrence in medical device repeated-measures follow-up studies. Various statistical assumptions have been made in handling missing data, including missing at random (MAR) and missing completely at random (MCAR). Those assumptions are relied upon to various degrees in mixed models, GEE, last value carried forward (LVCF) (*see* **Clinical Trials of Antibacterial Agents**), **multiple imputation methods**, and others. Both **intent-to-treat** and per-protocol (*see* **Clinical Trials of Antibacterial Agents**) approaches are often evaluated for medical device trials.

Observational Studies and Causal Inference in Medical Device Evaluation

Just as in any other area of medicine, for medical devices, the randomized clinical trial (RCT) is the most rigorous empirical tool for the investigation of treatment effects. Nevertheless, sometimes studies that do not involve explicit randomization in their design, that is, **observational studies**, are proposed as a supplement, or even a substitute, for a randomized clinical trial. A typical example in the context of medical device evaluation is using for comparison the treatment (or control) arms from previous clinical trials. Those comparison groups are usually called historical controls (*see* **Bias from Historical Controls**). It is often the case that observational studies compare favorably with RCT in terms of cost, making them an attractive alternative. Of course, any cost advantage of an observational study always comes with a major shortcoming, namely, vulnerability to potential bias. The problem of **bias in observational studies** is a challenging subject in statistics, for which systematic methodologies have been developed only relatively recently. The medical devices arena is one of the areas in which those methodologies are already beginning to have an important and fast-growing impact.

The concerns over bias in an observational study arise most naturally when the treatment group differs systematically from the historical control in the distributions of observed **covariates**, characteristics that are currently believed to be possibly related to the outcome; this situation is referred to as overt bias [76]. One way of dealing with the problem of overt bias is by dividing subjects into subclasses within which the distributions of observed covariates are similar among the treatment groups. Treatment comparison within each subclass would then have less bias due to observed covariates. Such subclasses are referred to as either strata or matched sets and their construction as **stratification** or matching. At first sight, one would expect that such a scheme of covariate balance would suffer from the usual curse of dimensionality, that is, it would be infeasible when more than a few covariates are to be balanced. The statistical theory of **propensity scores** [77] reassures us that this need not always be the case. In fact, for comparing two treatment groups, the task of covariate balance is essentially a one-dimensional problem involving a function of covariates called the propensity score, defined as the **conditional probability** of being in one of the treatment groups given the value of the covariate vector. Covariate balance using estimated propensity scores can be very effective in practice, as shown in the classic example in [78], involving 74 covariates and 2 treatment groups of sizes 590 and 925, respectively. The practical utility of the propensity-score approach has become more and more widely recognized, and there is now an emerging literature on its application in studies involving medical devices (e.g. [32, 67]).

While overt bias can be addressed by covariate balance, hidden bias is more difficult to deal with directly. To see the distinction between these two kinds of bias, we first need to define them more precisely. Within modern statistics, the notion of bias is inextricably linked to that of causal inference (*see* **Causation**), which is formulated on the basis of the two key concepts of potential outcomes and assignment mechanism [81]. According to the currently widely accepted statistical theory of causality [79], the causal effect of treatment A on a unit relative to treatment B (e.g. control) for an outcome variable is the difference between the two potential values of the outcome variable of the unit under the two treatments. A causal estimand is a parameter that compares the distribution over a set of units of the potential outcome under treatment A to that over the same set of units under treatment B (e.g. [27]). By definition, the potential outcomes of a unit under treatments

A and B can never both be observed. The task of statistical causal inference is to obtain valid estimates for causal estimands under the above constraint. This can be achieved by design, as in RCT, by assigning treatments to units according to an explicit and known probability model. In observational studies, causal inference can be conducted through the postulation of plausible assignment mechanisms. If the postulated plausible assignment mechanism is ignorable [80], that is, if it follows a probability model in which the probability of a unit being assigned to a treatment is independent of the unobserved potential outcomes, given observed covariates, then there is only overt bias, which can be addressed directly using covariate balance (*see* **Covariate Imbalance, Adjustment for**).

If the assignment mechanism is nonignorable, there will be hidden bias. Hidden bias can be modeled via a hypothetical unobserved covariate that is related to both treatment assignment and outcome. Such models may be used to perform what is called **sensitivity analysis** to address the issue of hidden bias indirectly, by asking how the causal inference would be altered under various assumptions about the behavior of a hypothetical unobserved covariate. For example, we may specify a class of probability models for the dependence of treatment assignment on the unobserved covariate and calculate extreme distributions of some test statistic under the class of models and the null hypothesis. This would give us a range for quantities determined by the null distribution, such as ***P* values** and estimates of an additive effect. If under the relatively severe dependence of treatment assignment on the unobserved covariate, the range of P values remain significant and the range of estimates point to effects in the same direction, then the conclusion of the study is said to be insensitive to hidden bias. Extensive discussions on this subject can be found in [76].

Observational studies are important to medical device evaluation and not just because they can be attractive alternatives to randomized clinical trials. It is not uncommon for studies designed as RCTs to eventually acquire features of observational studies because of complications such as noncompliance (*see* **Compliance Assessment in Clinical Trials**). Indeed, a continuum is thought to exist between an ideal RCT and a typical observational study. We often find ourselves to be at neither of the extremes and would thus benefit from statistical causal models that deal with situations across the whole continuum [3, 38]. Such statistical causal models also tend to provide a framework within which conventional statistical thinking can be improved upon. For example, the unifying theory on causal modeling [27] has proved to be helpful in giving a clearer, scientifically more meaningful definition for the concept of surrogacy (*see* **Surrogate Endpoints**). Of course, the availability of statistical techniques that address deviations from an ideal RCT does not in any way free us from the responsibility of trying our best to follow the protocol of an RCT as closely as possible.

The medical devices arena provides a fertile ground for the application of causal models. It has a large demand for empirical investigations of causal effects, and, moreover, the nature of the investigation often imposes special structural features and constraints that may present interesting and challenging cases for study design and causal modeling. The practice of double blinding, which is standard in many areas of medicine, is sometimes quite difficult to implement for medical devices. Treatments involving medical devices also tend to be more complex in structure, often consisting of multifaceted components. Devices of the same model may differ in engineering perfection, and medical professionals who are responsible for deploying them may vary in skill levels. Therefore, in conceiving a treatment with a medical device as homogeneous, some idealization is inevitable. Causal modeling can also play an important role in the evaluation of combination products, those involving both drugs and devices.

Noninferiority and Active Control Trials

A noninferiority active control clinical trial design is an increasingly popular approach for evaluating medical devices (*see* **Equivalence Trials**). The primary objective of such a trial is to demonstrate that a new (experimental) device performs as well as an existing one (active control). The design is useful when the new device is preferable for reasons such as a longer lifetime or a superior safety profile. The fundamental principles of these trials and the statistical methodologies applied in the area of therapeutic devices are the same as those in the area of drug development [10, 26, 87]. However, some statistical issues and design concerns have been encountered in medical device studies more frequently than in other

studies [96]. Some typical design and data analysis issues include the formulation of study hypotheses, the selection of active controls, the requirement of prespecified noninferiority margins δ (>0), the possibility of multiple testing procedures for different claims, and the need for appropriate data analysis for different study designs.

Blackwelder-type hypothesis testing is usually employed in medical device studies; the **alternative hypothesis** would be that the new treatment is not worse than active control by more than δ, with respect to some parameter of interest, such as proportion or mean. Unfortunately, some investigators attempt to use a conventional superiority alternative hypothesis to investigate noninferiority in some device studies. It is pointed out [10] that in determining whether a new device is as effective as an active control, the test of the conventional null hypothesis of equal effects is inappropriate and leads to logical difficulties. In particular, failure to reject the alternative hypothesis in the conventional superiority testing does not establish noninferiority [2].

An active control could be another effective device or standard of care. However, there is the danger that a series of active control trials might push the general treatment in the wrong direction by testing devices that are progressively inferior to previously active controls, which is called device creep. Also, it would be inappropriate to select as an active control a device that is out of date due to rapidly developed new technology.

A noninferiority active control study should have the ability to distinguish the new device from ineffective products; in pharmaceutical circles, this is referred to as assay sensitivity. The effectiveness of an active control can be demonstrated through its effect size, which is the treatment difference between the active control and the sham control. (A comparison to the no-treatment control instead of the sham could lead to bias and consequently a misestimated effect size.) However, the active control effect size cannot be measured in the current study since there is no sham arm; consequently, this effect size has to be deduced on the basis of historical experience showing the superiority of the active control over the sham. Some choices for estimated effect size include a point estimate of effect size from one large historical sham-controlled study or from multiple such studies. But the use of a point estimate is of concern since it disregards the uncertainty associated with the point estimate. To account for this uncertainty, a **Bayesian** approach could be employed [87].

Another crucial consideration concerns the choice of the reasonable noninferiority margin δ. This quantity should be sufficiently smaller than the effect size of active control so that from a clinical point of view a new device can be considered effective when a noninferiority claim is confirmed. In particular, the choice of δ should not lead to a situation where the new device is essentially equivalent or worse than the sham or no treatment, yet the null hypothesis of inferiority by more than δ is rejected.

In recent medical device active control studies, there is an increasing interest in hypothesis-testing procedures that simultaneously test for superiority and equivalence [64]. A test of conventional superiority (new device is better than active control) after claiming noninferiority or a test of noninferiority after failure to claim superiority reduces to simultaneous testing for noninferiority and superiority using a one-sided $(1-\alpha)100\%$ confidence interval. It is known that, given a prespecified noninferiority margin δ prior to analysis of trial data, the procedures need no adjustment for **multiplicity** by the closed testing procedure. However, it is crucial that the noninferiority margin δ be predetermined and described in a protocol; otherwise, a data analysis can always find the minimum value of δ, leading to a claim of noninferiority of the new device.

An interesting question is, given that one is willing to accept some decrease in the effectiveness up to δ for the new device and still consider it noninferior to the active control, should the device, by symmetry, be not just better by any amount but superior by a specified amount δ than the active control to qualify for superiority. That is, if the roles of the new treatment and active control are allowed to reverse, can one conclude that the active control is noninferior to the new device as long as the active control is not worse than the new device by δ? It implies that the new device is not superior to the active control if the magnitude of superiority of the new device over the active control is between 0 and δ, which conflicts with conventional superiority testing. In statistical terms, it is in fact a question of which superiority hypothesis should be tested following the noninferiority conclusion, the conventional superiority testing for any treatment difference or the one that confirms a difference of a prespecified magnitude δ [17].

Diagnostic Devices

Diagnostic devices are fundamentally different from therapeutic devices in that they are intended for detecting a condition rather than treating it. They can be broadly classified into two categories: *in vitro* and *in vivo*. Genetic tests and tests for blood glucose, cholesterol, hepatitis, and HIV are all laboratory tests based on tissue or blood specimens sampled from patients and therefore belong to the first category. Implanted glucose meters, devices using autofluorescence to detect disease, apnea monitors, and all sorts of diagnostic imaging (e.g. magnetic resonance imaging, ultrasound, mammography) are in the second category because all these involve test procedures performed directly on the patient. (*See also* **Diagnostic Tests, Evaluation of**.)

Diagnostic devices may generate quantitative or qualitative results. Blood glucose levels and cholesterol levels are examples of quantitative measures; in contrast, the results of influenza and pregnancy tests often take only two values, labeled positive and negative, and are qualitative measures. Quantitative measures can be transformed into dichotomous (qualitative) results via a threshold or cutoff value. For example, we may encode a quantitative prostate screening antigen (PSA) test result as positive or negative depending on whether the value is larger than or smaller than the threshold of, say, 4 ng/mL. Multiple cutoff points may be applied to a quantitative test to generate ordinal categories, for example, urinalysis of glucose with test results of negative, trace, 1+, 2+, and 3+ (*see* **Ordered Categorical Data**). Ordinal categories can also arise directly as in the case of the breast imaging recording and data system (BI-RADS®) scale of 0 to 5 for mammography.

If multiple tests can be applied to the same subject or specimen, it is frequently argued that the most efficient design is to have each person serve as his or her own control. However, without randomization, one must be on constant guard to control bias [4]. For large **screening trials**, randomization to one of two independent arms is often employed to compare different screening strategies. In the case of mammography, digital versus analog, the additional radiation exposure of two mammograms is a consideration for a single-arm diagnostic trial.

The performance of a test producing a dichotomous measure (positive and negative) is measured by its **sensitivity** and **specificity** when there is a truth standard (a **gold standard test**). The performance may also be evaluated on the basis of (positive and negative) **predictive values**; however, these latter values depend explicitly on **prevalence** as well as sensitivity and specificity. The comparison of two dichotomous tests is often problematic unless one test is simultaneously statistically superior to the other in sensitivity and specificity; more often than not, one test has better sensitivity and the other has better specificity. (For tests that are inherently continuous, it is usually preferable to perform the comparison before dichotomization, as discussed below.) Several references for the statistical methodologies for diagnostic tests are [71, 97]. A recent recommendation on the reporting of diagnostic accuracy studies is given in [12].

The absence of a truth standard presents challenges to the evaluation of qualitative test performance. One possibility is to evaluate a test on the basis of its agreement with another test. But two tests can agree and both can be incorrect. While it is often helpful to report the entire **2 by 2 table**, measures of agreement called positive and negative percent agreement are sometimes quite helpful. For example, positive agreement of a new test to its comparator is the percent that the new test is positive, given that the comparator is positive. Agreement can also be extended to ordered categories using the weighted **kappa** statistic. However, the comparison of two tests in the absence of truth can at most only establish equivalence but never the superiority of one test over the other.

Sometimes, the gold standard is available but cannot be applied to all study subjects for practical reasons. For example, in a study to evaluate the performance of the prostate screening antigen (PSA) test and the Digital Rectal Examination (DRE), the gold standard (biopsy) is usually not applied to the subjects with negative DRE results and negative PSA results. If the gold standard is applied when either test is positive, one could compare the performance of the two tests using the ratio of sensitivities and (1-specificities) [72]. However, sometimes this is not possible since for a new unproven test it may be very difficult to argue for the invasive biopsy. A more valid comparison in this case is to ask whether PSA adds any diagnostic capability to DRE since the PSA test is viewed as adjunctive and not "stand alone". When the gold standard is more likely to be applied to subjects who appear to be at high risk

for the disease, there can be bias in the estimation of the diagnostic accuracy of the test (verification bias). Here "intention-to-diagnose", a term coined by statisticians at the Food and Drug Administration's Center for Biologic Evaluation and Research, and not **intention-to-treat** (ITT), is the principle at work. As in ITT, it is very risky to ignore such missingness. One approach is to use imputation (*see* **Multiple Imputation Methods**) to develop unbiased estimates of sensitivities and specificities and then to inflate the nominal variances of these estimates to account for the imputed values, or alternatively, to use a **bootstrap** approach.

Sometimes, a random sample of subjects who test negative is assessed with a gold standard test. For example, in a study to evaluate the performance of the Human Papilloma Virus (HPV) test and the Pap test for the detection of cervical cancer, all women with either HPV positive results or abnormal Pap results were referred to the gold standard of colposcopy, and 10% of all women with negative HPV results and normal Pap results were also referred to colposcopy. The results of colposcopy of other subjects with negative HPV results and normal Pap tests can be considered a **missing-data** problem. With such data, it is possible to estimate sensitivity and specificity with confidence intervals [35].

A problematic but still common method of evaluating two tests is called discrepant (or discrepancy) resolution. The idea is to subject the off-diagonal entries (entries where the two diagnostic tests disagree) to further testing. The use of these results to modify the original table to produce estimates directly leads to bias in the estimation of performance [33, 63].

The diagnostic accuracy of a quantitative test can be evaluated by **Receiver Operating Characteristic** (ROC) analysis. The ROC plot is a graph of the observed sensitivity versus the 1-observed specificity of the diagnostic test, evaluated at all possible thresholds that one could use to dichotomize the diagnostic test [98]; the ROC plot is the empirical version of the ROC curve. One global measure of the diagnostic capability of the test is the area under the (empirical) ROC plot. Typically, in a study comparing a new device to an established one (the predicate device), each study specimen is tested with both the tests. The bootstrap method is quite helpful in reflecting the variability of the ROC plot and in comparative analysis [14]. In addition, if one wished to identify the threshold associated with say 90% observed sensitivity, one could reflect the variability associated with the particular derived threshold by a confidence interval based on bootstrapping.

Three measures of analytical performance of the tests are systematic bias, precision, and limit of detection. Systematic bias, the difference between the mean of the results of measurements and a true value, is one important characteristic. For the *in vitro* laboratory test, this true value could be that of an analyte, and for *in vivo* tests, an example would be temperature from an ear thermometer. Usually, systematic bias of a new test can be evaluated by a **regression** analysis of the new test on the reference method. If it can be assumed that the reference test measures the true value with no error, the performance can be determined using ordinary **least squares** regression, comparing the deviation of the slope from 1 and the intercept from 0. The amount of **random error** manifests itself through agreement between results of independent measurements under stipulated conditions. Variability is usually expressed numerically by **standard deviation** and its inverse, precision, or by the coefficient of variation (*see* **Standard Deviation**). For an analyte, as the concentration being measured gets smaller and smaller, its presence is harder to ascertain. The limit of detection is the lowest concentration that can be reliably distinguished from zero, with α and β for the two types of error. For a sample containing no analyte (blank sample), the $(1-\alpha)$-quantile of the distribution of blank values indicates a limit that for the blank sample is only exceeded by a probability of α. The samples that provide values exceeding this limit may be declared to contain nonzero analyte (type I error α). At the same time, some of the measurements of the sample with a low amount of analyte fall below this limit and hence are declared to be zero analytes (type II error β). One statistical approach to this problem that assumes Gaussian error distributions (*see* **Normal Distribution**) is given in [19].

The absence of a gold standard in the case of continuous data creates a challenge. For two tests, it may be sufficient to investigate how similar they are; this is often called method comparison. In the case in which there are **errors in variables** (or measurement error in both tests), one can employ measurement error models [28]. In the medical device literature, this approach may be remedied

by orthogonal regression (*see* **Orthogonality**), Deming regression, or Passing–Bablock regression [70]. The Bland–Altman plot is especially helpful [11] for investigating not just relative bias but also heteroscedasticity (*see* **Scedasticity**). Hawkins extends this to a more formal regression approach with regression diagnostics to examine model assumptions [34]. The notion of equivalence, especially individual equivalence, can depend not just on the slope and intercept but also on the variability about the line [75] (*See also* **Diagnostic Test Evaluation Without a Gold Standard**).

Some diagnostic devices are used for special purposes. There are qualitative tests used for triage to avoid more difficult, expensive, or invasive procedures. For example, in some age groups, a test for HPV is used to decide whether to send patients with an abnormal Pap smear to colposcopy. Monitoring devices track certain medical conditions over time to detect changes from baseline or normal or to detect changes in dose response to medications. Blood glucose meters, prothrombin time tests, and pulse oximeters are examples of such devices. For these devices, it is usually more appropriate to use equivalence based on the individual rather than on the population because the absolute difference between the measure and the true value may cause danger to patients.

Repeated measures or **longitudinal data analysis** are frequently applied to diagnostic medical devices, such as in an equivalence study via matched-pair comparison between the gold standard and the test kit for glucose measurements, taken at various follow-up times. The repeated measures can be equally spaced or unequally spaced, balanced or unbalanced (each patient may have different numbers of repeated observations), with or without missing data. Statistical methods used to evaluate equivalence of two-test methods in matched-pair repeated-measure or longitudinal data analysis include simple ordinary least squares, generalized least squares, GEE, overall mean paired difference with 95% confidence interval, random-effect regression model, repeatability and reproducibility studies, and concordance correlation [53].

Diagnostic Imaging

A particular application of *in vivo* tests is diagnostic imaging (*see* **Image Analysis and Tomography**). The data can be quantitative or qualitative and for the latter tend to be ordinal. An example of an ordinal scale is the BI-RADS scale that is employed in mammography. There are advantages to having finer ordinal scales where possible; for breast cancer, some advocate a 100-point scale corresponding to the estimated probability of malignancy rather than the coarse BI-RADS scale. ROC methodology can be used to determine whether an imaging technology has any diagnostic capability or whether a new diagnostic test has any adjunctive capability relative to other known test(s). Examples include mammography and chest X rays.

In many diagnostic-imaging situations, there is considerable variability in the readers of the images. This can be an impediment to investigating whether a new imaging system is superior to others as well as in investigating whether it is noninferior to an existing system. An example of the latter is digital mammography in comparison to analog mammography. One might expect the two to be very similar, but the reader-to-reader variability makes this difficult to assess. Beiden, Wagner, and Campbell [5] have employed **bootstrapping** to study the **components of variance** in such comparisons and to model explicitly the reader variability and the case variability in order to compare two modalities. In the case of digital mammography, this approach has been used in a study design to figure out how to trade off cases and readers.

For a recent discussion of the evaluation of medical imaging and **computer-aided diagnosis** (CAD) systems, see [91]. Diagnostic-imaging tests present unique study design problems. In the case of readers, it is often a question of what information is available and when. A reasonable design is to gradually provide more information and observe the ratings and how they change. For example, the diagnostic image may first be read with no other information and the rating recorded and then read with clinical information to see whether the rating changes. A further design difficulty in some studies that rely on the same readers for different modalities is that some readers claim to remember the films and the ratings from a first reading, even if one month or more time has elapsed. In the case of CAD, the presence of the technology may affect the performance of the reader either by increasing his or her vigilance before the application of CAD or, retarding it, by anticipating that the CAD system will pick up any oversights. There are also issues

associated with satisfaction of search after obtaining the CAD results.

Microarrays

While most diagnostic *in vitro* medical devices have been designed to test for a single analyte that is associated with disease, **microarrays** are a revolutionary medical device technology that can be used to study thousands of analytes simultaneously and, by doing so, have the potential to advance dramatically the diagnosis, treatment, and prevention of disease. Microarrays can be used to diagnose disease, screen for the mutations or single nucleotide **polymorphisms** (SNPs) associated with increased risk of disease, develop new classification systems for stages or subclasses of disease, identify drug or gene therapy targets (e.g. genes or proteins) in the treatment of disease, and predict drug response and drug toxicity in individual patients. Microarray data can also be used to study interactions on phenotypes among genes and between genes and environmental exposures. In short, microarrays enable the genomewide study of biomarkers, genes, and proteins, which promises to produce dramatic advances in our understanding of molecular variations among normal and diseased populations. A very good introduction to the relevant biology, chemistry, and technologies is given in [85]. Overviews of statistical methods for microarrays are given in [83, 85].

A DNA microarray is a glass or other surface onto which many thousands of individual **DNA sequences** are printed. The DNA sequences are called probes and are typically either cDNA, single-stranded clones of entire DNA sequences of genes, obtained by reverse transcription of mRNA, or oligonucleotides (oligos), synthetic DNA representing short specific segments of DNA sequences. In microarray experiments, mRNA from cells of a sample are extracted, labeled with fluorescent dye, and then allowed to combine with or *hybridize* to probes on the array. An mRNA molecule will, in principle, only hybridize to the complementary probe on the array. The relative expression of the gene (mRNA abundance) can be quantified by measuring the fluorescence intensity at the spot of the probe.

The large amount of information generated by gene expression microarray experiments is unfortunately accompanied by complex questions relating to quality control [49, 73]. Fluorescence measurements of gene expression are obtained by a sequence of image analysis steps that include locating the spots within grids (addressing), distinguishing spots from the background (segmentation), measuring fluorescence intensity via laser scanning, and correcting spot intensities for background intensity (cf. [94]). The intensity at the spot may have to be adjusted for the shape of the spot and the amount of probe that has been laid down. Reproducibility of intensity measurements is key to validating microarray data. Recently, guidelines on the minimum information about a microarray experiment (MIAME) have been published [13].

A major statistical issue in the design of microarray experiments is that they typically involve the study of a very large number of variables (e.g. thousands of genotypes or gene expressions) on a relatively very small number of experimental units (e.g. tumor samples). The obvious problem is that outcomes such as disease status are overfit by statistical models using as many variables as samples.

The designs of cDNA and oligo arrays are different [84]. cDNA microarrays are in an **incomplete block design** in that they involve a competitive hybridization of mRNA from two samples to the cDNA probes. The mRNA from the two samples are labeled with different color dyes (e.g. green and red fluorescent dyes Cy3 and Cy5) and mixed before hybridization. The intensities in both the red and green channels can be measured separately with the laser scanner, and a comparison is usually based on their log ratio. Early experiments compared each sample with a reference, with the same dye always being used for the reference (e.g. Cy3); statistically, this design is not desirable because it does not control for dye effects (differential dye efficiencies) and gives too much information on the reference, which is not of interest. The *dye swap design* adjusts for dye effects by comparing the sample with the reference on two arrays, with the dyes being swapped for the second array. The dye swap design is a **Latin square design**, where array and dye are **blocking** factors with two levels each. The *loop design* forms a cycle of comparisons of reference to sample 1, sample 1 to sample 2, and so on up to sample n to reference, with the first in the pair always labeled with, say, green dye [48]. This design could be considered an incomplete Latin square in that each sample appears once each in a pair of the n+1 arrays

and once for each dye. The loop design eliminates dye effects and is efficient at obtaining information on samples, but if one array is unusable, then the optimal properties of the design are destroyed. Designs that are more **robust** to unusable arrays are being considered.

In oligo arrays, each gene is represented by 16 to 20 probe pairs of oligos. Each pair consists of a perfect match (PM) probe and a mismatch (MM) probe, in which the central nucleotide base is inverted. The probes are scattered across the array. One sample labeled with fluorescent dye is allowed to hybridize to probes on the array. The intensity of the sample is measured by a summary of the PM intensities corrected for a summary of the MM intensities, which serves as a control. Because only one sample is hybridized per array, comparisons of samples are completely **confounded** by array effects unless the arrays are replicated for the sample.

Normalization refers to adjustment of the intensities for systematic biases, such as dye effects in cDNA arrays, array effects, and spatial effects within arrays. Regression models of the intensities can be used to adjust for confounding effects [16, 47, 89, 93]. Such models typically include main effects for genes, main effects for varieties of samples (e.g. tumorous versus non-tumorous), and gene by variety **interactions**; here the interactions represent the effects of interest, that is, the differential gene expressions among the varieties. Confounding effect that can be included in the model are array effects, spatial effects within arrays (e.g. spot effects), and dye effects (for cDNA arrays). Array effects are random, but nonetheless, they are sometimes considered to be fixed because this assumption simplifies the analysis considerably. The number of confounding effects can be considerable, which suggests that **propensity-score** methods [77, 78] that are used to adjust observational studies for numerous confounding effects might be useful in microarray experiments. For cDNA arrays, dye effects depend on the mean intensity level. These effects are commonly examined with a so-called M−A scatterplot, where M is the log ratio of intensities for two samples and A is the log of the geometric mean of the intensities. The intensity-specific effects are commonly adjusted out by fitting a lowess curve [18] to the scatterplot. Assuming only a fraction of the genes are differentially expressed between the two samples, the lowess curve can be considered to represent the average log ratio at a given intensity level when there is no differential expression at that level [24].

One goal of microarray experiments is to find all the genes that are differentially expressed between two varieties of samples (e.g. tumorous versus non-tumorous). These genes can then be used to create a specialized microarray for diagnostic testing, for example. When there are thousands of genes, this is a huge **simultaneous-testing** problem in need of a multiplicity adjustment. For example, if 5000 genes are studied and none are differentially expressed, then a nominal 5% level testing is expected to falsely detect differential expression in 250 genes. The adjustment need not be as severe as the **Bonferroni** correction when correlation between the tests is considered. The tests are correlated because the same samples are used to test each gene, and groups of genes with similar function can be upregulated or downregulated in tandem across samples. Permutation of the varieties of the samples can be used to approximate the joint null distribution of the P values, which can then be used to obtain an adjusted P value for a gene, that is, the familywise level of the tests at which the particular test for the gene would reject the null hypothesis of no differential expression [24]. Alternatively, a **Bayesian approach** to multiplicity considers gene effects (and/or gene by variety interactions) as random, which induces a multiplicity adjustment that is inversely related to the ratio of the variance between genes to the variance within genes [66]. A noninformative, nonparametric Bayesian approach estimates the mixture distribution of differentially expressed and nondifferentially expressed genes with the empirical distribution, estimates the null distribution for nondifferentially expressed genes via permutation, estimates the mixing proportion by exploiting bounds for it on the basis of the two distributions, and applies **Bayes' theorem** to these estimated quantities to obtain the posterior probabilities of differential expression for genes [25]. In the same paper, a connection is made between the posterior probability and the false discovery rate, the expected proportion of Type I errors, which is increasingly being controlled in large multiplicity problems, as opposed to the more traditional and more conservative familywise Type I error rate.

Microarray experiments can also be used to classify samples into defined groups, such as tumorous or nontumorous tissue, or stages of disease such as metastatic versus *in situ* cancer. Discrimination

methods have been compared, including Fisher's linear discrimination (*see* **Discriminant Analysis, Linear**), **maximum likelihood** (quadratic) discrimination, nearest neighbor, classification and regression trees (CART) (*see* **Tree-structured Statistical Methods**), bootstrap aggregating (bagging) procedures, and boosting procedures, with the simple procedures of linear discrimination and nearest neighbors appearing to work relatively well [23].

These experiments have also been employed to discover new subclasses of disease. For example, Alizadeh et al. [1] discovered two subclasses of diffuse large B-cell lymphoma, one of which did not respond well to standard treatment. Because the classification variable is completely unknown, methods for discovering new classes are termed unsupervised or clustering methods. These methods include hierarchical clustering, k-means clustering, self-organizing maps, mixture model approaches [95], factor analysis, and plaid models [55]. The latter is distinguished in that it simultaneously clusters genes as well as samples. All clustering methods are based on a summary measure of the similarity of expression profiles. The Pearson correlation is usually used as the measure of similarity modulo some transformations of the raw intensities.

Genomewide scans for SNPs can be used to rapidly identify an SNP associated with disease. Interestingly, genomewide linkage disequilibrium studies of SNPs do not have to be adjusted for multiplicity when the disease of interest is monogenic; a confidence-based approach can be used to bound the location of the disease gene to within neighborhoods of single nucleotide polymorphisms [57]. Alternatively, each SNP can be tested directly for association with disease; in contrast with linkage disequilibrium studies, association studies need to be adjusted for multiplicity [62].

Bayesian Statistics and Medical Devices

The fundamental idea of the Bayesian approach to statistics lies in representing one's uncertainty about an unknown quantity of interest as a probability distribution. In medical device clinical studies, examples of unknown quantities of interest are safety and effectiveness endpoints, a patient's outcome to be observed in the future, or even missing observations. Before medical device trials are conducted, the probability distributions assigned to the quantities of interest are called **prior distributions**. After data are gathered and new information becomes available, the prior distribution is mathematically updated according to **Bayes' theorem**, becoming a posterior distribution. This approach provides a mathematically valid way of combining previous information (the prior distribution) with current data, adjusting to changing levels of evidence, and working as Science works: today's posterior distribution becomes tomorrow's prior distribution. This approach also allows for the derivation of predictive probability, a special type of posterior probability, namely, the probability of a future observation given outcomes that have already been observed.

The incremental steps in which improvements are made in device development make the Bayesian approach particularly suitable. Good prior information is often available from, for example, trials in other countries, earlier trials on previous device versions, or possibly bench tests or animal studies. Lack of prior information is represented by a noninformative prior distribution and does not thwart the use of **Bayesian methods**.

The use of Bayesian statistics in planning and analyzing clinical trials [44, 59], especially for medical devices [9], has increased dramatically in recent years (*see* **Bayesian Methods in Clinical Trials**). The Bayesian paradigm, which serves as a basis for formal decision theory, offers an ideal framework for investigational or regulatory settings in which decisions for developing, approving, not approving, or improving medical devices are made on a daily basis. A Bayesian initiative was begun in the Division of Biostatistics at the Food and Drug Administration's Center for Devices and Radiological Health [41]. The medical device community has begun to pioneer using Bayesian methods in planning and analyzing confirmatory clinical trials in recent years.

A Bayesian clinical trial for a medical device may include prior information for the new device, for the control device, or for both the new and control devices. Previous device studies used as sources of prior information should be similar to current studies in terms of devices used, objectives, endpoints studied, protocol, patient population, investigational sites, physician training, patient management, and proximity in time. Covariates such as demographics and prognostic variables can be used to calibrate previous studies to the current study. The use of prior

information often leads to more precise estimates enabling decision-makers to reach a decision on a device with smaller and shorter trials.

Bayesian methods can play a particularly useful role in diagnostic devices studies. Recent research suggests that it may provide an analytical framework capable of estimating and comparing sensitivities and specificities of diagnostic tests in the absence of a perfect test (truth standard) [21, 30].

Although the most direct way of encoding prior information is via a probability distribution assigned to quantities of interest, in many cases, an attractive approach to incorporating information from previous device studies into the current study is by using Bayesian hierarchical modeling [29]. The **hierarchical model** assumes that study-specific parameters are realizations or values of random variables from a common distribution. This usually (but not always) results in more precise estimates of the current study parameters [60]. Thus, we often say that a hierarchical model enables the current study to borrow strength from previous studies. Hierarchical models are self-correcting in that the current study borrows less strength from the previous studies as variation between studies increases, protecting against overreliance on inadequate prior information. One way of using a Bayesian hierarchical model in a clinical trial is to supplement the concurrent control with historical controls; this can allow fewer patients in the concurrent control arm, a larger proportion of patients being allocated to the experimental device, and overall, a smaller trial. Hierarchical models can also be used to adjust subgroup analyses (*see* **Treatment-covariate Interaction**) for effects by allowing a subgroup to borrow strength from related subgroups [22]. For an application of Bayesian hierarchical modeling to coronary artery stents, see [69].

Exchangeability is a key idea in Bayesian inference. Two patients are exchangeable if their roles can be switched without affecting the inference. For technical definitions of exchangeability, see [8]. In medical device clinical trials, exchangeability may be thought of at different levels; there can be exchangeability of studies, of centers, and of patients. In the case of a single **multicenter trial**, centers are considered exchangeable if their roles can be switched without affecting the inference, such as when they are considered as elements of the same superpopulation of centers. But the exchangeability of centers does not imply the exchangeability of patients among centers. If patients treated in different centers are considered exchangeable, then they are said to be "poolable" for making inference. However, as long as the centers are exchangeable, the patients in the centers can be combined by the use of a Bayesian hierarchical model. Between-center variability, which can be large in device trials, is accounted for by assuming that center-specific device effects are random (*see* **Random Effects**). Besides device effects, there are center effects; consequently, the patients in different centers will not be completely pooled together but may be pooled to a certain degree. The degree of pooling will depend on the variability among the centers as compared to the variability among the patients within each center.

The Bayesian approach does not require the sample size to be determined in advance because inferences are based on the parameter space rather than on the sample space, as in the frequentist approach. The sample size required for a Bayesian medical device trial depends on the amount of information necessary to reach a decision, on the amount of prior information available, and on the variability of the data. It can be revised at any point in the study by considering the current posterior distribution instead of the prior distribution. The underlying idea is to gather just enough information to make a decision about the device and therefore not expose patients to unnecessary risk.

In practice, however, particularly in the medical device arena, a minimum sample may be needed to evaluate fairly rare outcomes such as complications or device malfunctions or failures, to verify model assumptions, or to ensure that the prior information does not overwhelm the clinical data of the current study. A maximum sample size may also be useful for economical or ethical reasons or to plan the logistics of the trial. Bayesian approaches to sample size determination include those based on power to make decisions [82], interval length and coverage probability [42, 43], and decision theory [58].

The Bayesian decision-making process is based on posterior probabilities that most of the time do not depend on the experimental design (the **Likelihood Principle** [7]). In contrast, in the frequentist approach, decisions are often based on P values, which strongly depend on the experimental design. As a result, the Bayesian approach can provide more flexibility in both design and analysis [6, 40]. Experiments can be altered in midcourse, arms may be dropped, the sample size may be reduced or

augmented, and interim analyses may require no adjustments. None of these modifications interfere with posterior or predictive probabilities and, therefore, are easily accommodated in the Bayesian framework. When interim looks or other modifications are performed, one may assess probabilities of type I and type II errors through simulations.

In some medical device clinical trials, it is more ethical to change the randomization rate during the course of the trial, thus increasing the chance of a patient being randomized to the "winning" arm. Here, too, the Bayesian approach is amenable to changes in the randomization rates during the course of the trial [45].

Predictive distributions are widely used in medical device clinical trials. For example, by using Bayesian predictive probabilities, one may stop a trial early on the basis of results obtained at an interim point. If the predictive probability of the trial success is sufficiently high (or low), one may stop the trial and declare success (or failure) early. Predictions can be made only if the patients yet to be observed are exchangeable with the patients already observed. In device trials, unobserved patients enrolled later in the study may not be exchangeable with patients enrolled earlier if there is a learning curve associated with the device. In addition, one can also calculate the predictive probability of the outcome of a future patient given the observed outcomes of the patients in a clinical trial, provided that the patient is exchangeable with the patients in the trial. Predictive distributions can also be used when important data are missing. In this case, missing data can be predicted given the observed data, and trial results can be adjusted accordingly. However, the adjustment strongly depends on assumptions about patterns of "missingness". A particular case occurs when patients have two measurements, the first at an earlier follow-up visit and the second, sometimes missing, at a later follow-up visit. Then predictions for the later follow-up visit may be made (even before the follow-up time has elapsed) on the basis of measures at the early follow-up visit, provided there are some patients that have results from both follow-up visits and there is a strong correlation between the early and the later measurement.

The Bayesian approach can be used for hypothesis testing and interval estimation. Bayesian hypothesis testing in medical device studies uses the posterior distribution to calculate the probability that a particular hypothesis, either null or alternative, is true, given the observed data. An alternative Bayesian approach is based not only on the posterior probabilities of the hypotheses but also on the costs of making decision errors. Bayesian interval estimates are based on the posterior distribution. If the posterior probability that an endpoint lies in an interval is 0.95, then this interval is called a 95% credible interval. For constructing credible intervals, see [39].

Bayesian calculations require integration. For example, the posterior probability of a hypothesis is expressed as an integral involving the posterior distribution. When, as is often the case, the dimensionality of these integrals is high, Bayesian calculations are made with **Markov Chain Monte Carlo** (MCMC) sampling methods, such as Gibbs sampling. The Gibbs sampling method creates a **Markov Chain** by sampling sequentially from the posterior distribution for each parameter, conditional on the last sampled values of the other parameters and the data. These distributions are called full conditional distributions. In contrast with the (joint) posterior distribution of the parameters, the full conditional distributions can often be written in closed form, making sampling from them straightforward. Under mild regularity conditions, the sampling distribution converges to the posterior distribution. Subsequent samples can then be used to compute the desired Bayesian integrals. Because the samples taken are not independent (they form a Markov Chain), a Bayesian integral is correctly computed only if enough samples are taken such that the support of the posterior distribution has been completely explored. A Bayesian statistics program that uses Gibbs sampling is Bayesian inference Using Gibbs Sampling, or BUGS [31].

In summary, the Bayesian approach is playing an increasing role in medical device clinical trials because it is compatible with the process of research and development of such products. Bayesian methods require special technical expertise and are often computer intensive. However, the savings of flexible, smaller, and, often, shorter trials usually offset the higher technical and statistical complexity.

Surveillance of Devices

The surveillance of medical devices after marketing permission, which is of concern to regulatory bodies throughout the world, generates interesting statistical problems. Postmarket surveillance is as important

in medical devices as it is in pharmaceutical products (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**). Usually, the reporting systems suffer from severe underreporting as well as the inability oftentimes to definitely link the adverse event to the use of the device. Device-specific statistical efforts to identify a signal in a database of medical device adverse event reports include [52, 54].

The assistance of Marina Kondratovich, Ph.D., and Kay Barrick in the preparation of this entry is gratefully acknowledged.

References

[1] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. & Staudt, L.M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**(6769), 503–511.

[2] Altman, D.G. & Bland, J.M. (1995). Absence of evidence is not evidence of absence, *British Medical Journal* **311**, 485.

[3] Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with discussion), *Journal of the American Statistical Association* **91**, 444–472.

[4] Begg, C.G. (1987). Biases in the assessment of diagnostic tests, *Statistics in Medicine* **6**, 411–423.

[5] Beiden, S.V., Wagner, R.F. & Campbell, G. (2000). Components-of-variance models and multiple bootstrap experiments: an alternative methodology for random-effects, receiver operating characteristic analysis, *Academic Radiology* **7**, 341–349.

[6] Berger, J.O. & Berry, D.A. (1988). The relevance of stopping rules in statistical inference, in *Statistics, Decision Theory, and Related Topics, IV*, S.S. Gupta & J.O. Berger, eds. Springer-Verlag, New York, pp. 29–72 (with discussion).

[7] Berger, J.O. & Wolpert, R.L. (1988). *The Likelihood Principle*, 2nd Ed. IMS, Hayward.

[8] Bernardo, J. & Smith, A. (1993). *Bayesian Theory*. John Wiley & Sons, New York.

[9] Berry, D. (1997). *Using a Bayesian Approach in Medical Device Development*, Technical Report 97–21, Duke University Institute of Statistics and Decision Sciences, Durham. Web address `http://www3.mdanderson.org/depts/biostatistics/people/dberry/CDRH.doc` (accessed December, 2002).

[10] Blackwelder, W.C. (1982). "Proving the null hypothesis" in clinical trials, *Controlled Clinical Trials* **3**, 345–353.

[11] Bland, J.M. & Altman, D.G. (1999). Measuring agreement in method comparison studies, *Statistical Methods in Medical Research* **8**, 135–160.

[12] Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.R., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D. & de Vet, H.C.W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative, *Clinical Chemistry* **49**, 1–6.

[13] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C.P., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. & Vingron, M. (2001). Minimum information about a microarray experiment (MIAME): Toward standards for microarray data, *Nature Genetics* **29**, 365–371.

[14] Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests, *Statistics in Medicine* **13**, 499–508.

[15] Campbell, G. (1996). Statistical issues in medical devices: a regulatory perspective, *American Statistical Association 1996 Proceedings of the Biopharmaceutical Section*. American Statistical Association, Alexandria, pp. 224–229.

[16] Chu, T.-M., Weir, B. & Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments, *Mathematical Biosciences* **176**, 35–51.

[17] Chuang-Stein, C. (2001). Testing for superiority or inferiority after concluding equivalence? *Drug Information Journal* **35**, 141–143.

[18] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

[19] Davenport, J.M. & Schlain, B. (2000). Testing claimed minimal detectable concentration of in vitro medical diagnostic devices, *Clinical Chemistry* **46**, 1669–1680.

[20] DeMets, D.L. (2000). The role of surrogate outcome measures in evaluating medical devices, *Surgery* **128**, 379–385.

[21] Dendukuri, N. & Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests, *Biometrics* **57**, 158–167.

[22] Dixon, D.O. & Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial, *Statistics in Medicine* **11**, 13–22.

[23] Dudoit, S., Fridlyand, J. & Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association* **97**, 77–87.

[24] Dudoit, S., Yang, Y.H., Callow, M.J. & Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* **12**, 111–139.

[25] Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* **96**, 1151–1160.

[26] Farrington, C.P. & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, *Statistics in Medicine* **9**, 1447–1454.

[27] Frangakis, C.E. & Rubin, D.B. (2002). Principal stratification in causal inference, *Biometrics* **58**, 21–29.

[28] Fuller, W.A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

[29] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

[30] Georgiadis, M.P., Johnson, W.O., Singh, R. & Gardner, I.A. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests, *Applied Statistics* **52**, 63–76.

[31] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[32] Grunkemeier, G.L., Payne, N., Jin, R. & Handy, J.R. Jr. (2002). Propensity score analysis of stroke after off-pump coronary artery bypass grafting, *Annals of Thoracic Surgery* **74**, 301–305.

[33] Hagdu, A. (1997). Bias in the evaluation of DNA-amplification tests for detecting chlamydia trachomatis, *Statistics in Medicine* **16**, 13912–1399.

[34] Hawkins, D.M. (2002). Diagnostics for conformity of paired quantitative measurements, *Statistics in Medicine* **21**, 1913–1935.

[35] Hawkins, D.M., Garrett, J.A. & Stephenson, B. (2001). Some issues in resolution of diagnostic tests using an imperfect gold standard, *Statistics in Medicine* **20**, 1987–2001.

[36] Holt, J.D. & Prentice, R.L. (1974). Survival analysis in twin and matched pair experiments, *Biometrics* **61**, 1–30.

[37] Huster, W.J., Brookmeyer, R. & Self, S.G. (1989). Modeling paired survival data with covariates, *Biometrics* **45**, 245–156.

[38] Imbens, G.W. & Rubin, D.B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance, *Annals of Statistics* **25**, 305–327.

[39] Irony, T.Z. (1992). Bayesian estimation for discrete distributions, *Journal of Applied Statistics* **19**, 533–549.

[40] Irony, T.Z. (1993). Information in sampling rules, *Journal of Statistical Planning and Inference* **36**, 27–38.

[41] Irony, T.Z. & Pennello, G.A. (2001). Choosing an appropriate prior for Bayesian medical device trials in the regulatory setting, *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*, Vol. M, #85. American Statistical Association, Alexandria.

[42] Joseph, L., Wolfson, D.B. & Du Berger, R. (1995a). Sample size calculations for binomial proportions via highest posterior density intervals, *The Statistician* **46**, 139–144.

[43] Joseph, L., Wolfson, D.B. & Du Berger, R. (1995b). Some comments on Bayesian sample size determination, *The Statistician* **44**, 167–171.

[44] Kadane, J.B. (1995). Prime time for Bayes, *Controlled Clinical Trials* **16**, 313–318.

[45] Kadane, J.B. (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*. John Wiley & Sons, New York.

[46] Kaptchuk, T.J., Goldman, P., Stone, D.A. & Stason, W.B. (2000). Do medical devices have enhanced placebo effects? *Journal of Clinical Epidemiology* **53**(8), 786–792.

[47] Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J. & Churchill, G.A. (2002). Statistical analysis of a gene expression microarray experiment with replication, *Statistica Sinica* **12**, 203–217.

[48] Kerr, M.K. & Churchill, G.A. (2001). Experimental design for gene expression microarrays, *Biostatistics* **2**, 183–201.

[49] King, H.C. & Sinha, A.A. (2001). Gene expression profile analysis by DNA microarrays: promise and pitfalls, *Journal of the American Medical Association* **286**, 2280–2288.

[50] Lao, C.S. (1995). Statistical consideration for survival analysis from medical device clinical studies, *Journal of Biopharmaceutical Statistics* **5**, 159–170.

[51] Lao, C.S. (1998). The repeated-measure matched pair design in medical device clinical studies, in American Statistical Association 1998 Proceedings of the Biopharmaceutical Section, Alexandria, pp. 75–80.

[52] Lao, C.S. (2000a). Statistical issues involved in medical device post-marketing surveillance, *Drug Information Journal* **34**, 483–493.

[53] Lao, C.S. (2000b). Equivalence in test assay method comparison for the repeated-measure, matched-pair design in medical device studies: Statistical considerations, *Journal of Biopharmaceutical Statistics* **10**, 433–445.

[54] Lao, C.S., Kessler, L.G. & Gross, T. (1998). Proposed statistical methods for signal detection of adverse medical device events, *Drug Information Journal* **32**, 183–191.

[55] Lazzeroni, L. & Owen, A. (2002). Plaid models for gene expression data, *Statistica Sinica* **12**, 61–86.

[56] Lee, E.W., Wei, L.J. & Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in *Survival Analysis: State of the Art*, J.P. Klein & P. Goel, eds. Kluwer Academic Publishers, Boston, pp. 237–248.

[57] Lin, S., Rogers, J.A. & Hsu, J.C. (2001). A confidence-set approach for finding tightly linked genomic regions, *American Journal of Human Genetics* **68**, 1219–1228.

[58] Lindley, D.V. (1997). The choice of sample size, *The Statistician* **46**, 129–138.

[59] Malakoff, D. (1999). Bayes offers a "new" way to make sense of numbers, *Science* **286**, 1460–1464.

[60] Malec, D. (2001). A closer look at combining data among a small number of binomial experiments, *Statistics in Medicine* **20**, 1811–1824.

[61] Margo, C.E. (1999). The placebo effect, *Survey of Ophthalmology* **44**, 31–44.

[62] McCarthy, J.J. & Hilfiker, R. (2000). The use of single-nucleotide polymorphism maps in pharmacogenomics, *Nature Biotechnology* **18**, 505–508.

[63] Meier, K. (2002). Reporting results from studies evaluating diagnostic tests, *Clinical Microbiology Newsletter* **24**, 60–63.

[64] Morikawa, T. & Yoshida, M. (1995). A useful testing strategy in Phase III trials: combined test of superiority and test of equivalence, *Journal of Biopharmaceutical Statistics* **5**, 297–306.

[65] Moseley, J.B., O'Malley, K., Petersen, N.J., Menke, T.J., Brody, B.A., Kuykendall, D.H., Hollingsworth, J.C., Ashton, C.M. & Wray, N.P. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee, *New England Journal of Medicine* **347**, 81–88.

[66] Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. & Tsui, K.W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology* **8**(1), 37–52.

[67] Normand, S.T., Landrum, M.B., Guadagnoli, E., Ayanian, J.Z., Ryan, T.J., Cleary, P.D. & McNeil, B.J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores, *Journal of Clinical Epidemiology* **54**, 387–398.

[68] O'Brien, P.C. & Fleming, T.R. (1987). A paired-Wilcoxon test for censored paired data, *Biometrics* **43**, 169–180.

[69] O'Malley, A.J., Normand, S.T. & Kuntz, R.E. (2003). Application of models for multivariate mixed outcomes to medical device trials: coronary artery stenting, *Statistics in Medicine* **22**, 313–336.

[70] Passing, H. & Bablock, W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods, *Journal of Clinical Chemistry and Clinical Biochemistry* **21**, 709–720.

[71] Pepe, M.S. (2003). *The Evaluation of Diagnostic Tests and Biomarkers*. Oxford Press, London.

[72] Pepe, M.S. & Alonzo, T.A. (2001). Comparing disease screening tests when true disease status is ascertained only for screen positives, *Biostatistics* **2**, 249–260.

[73] Petricoin III, E.F., Hackett, J.L., Lesko, L.J., Puri, R.K., Gutman, S.I., Chumakov, K., Woodcock, J., Feigal, D.W., Zoon, K.C. & Sistare, F.D. (2002). Medical applications of microarray technologies: a regulatory science perspective, *Nature Genetics Supplement* **32**, 474–479.

[74] Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*. John Wiley & Sons, New York.

[75] Ponnapalli, R.M., Vishnuvajjala, R.L. & Campbell, G. (1999). On equivalence trials with medical devices, *American Statistical Association 1999 Proceedings of the Biopharmaceutical Section*. American Statistical Association, Alexandria, pp. 224–226.

[76] Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Ed. Springer-Verlag, New York.

[77] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

[78] Rosenbaum, P.R. & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association* **79**, 516–524.

[79] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.

[80] Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization, *Annals of Statistics* **6**, 34–58.

[81] Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies, *Statistical Science* **5**, 472–480.

[82] Rubin, D.B. & Stern, H.S. (1998). Sample size determination using posterior predictive distributions, *Sankhya* **60**, 161–175.

[83] Satagopan, J.M. & Panageas, K.S. (2003). Tutorial in biostatistics: a statistical perspective on gene expression data analysis. *Statistics in Medicine* **22**, 481–499.

[84] Sebastiani, P. & Ramoni, M. (2002). Statistical challenges in functional genomics, Submitted to *Statistical Science*. Text available at URL: `http://www.genomethods.org`.

[85] Schena, M. (2003). *Microarray Analysis*. Wiley, New York.

[86] Scott, P.E. & Campbell, G. (1997). Interpretation of subgroup analyses in medical device clinical trials, *Drug Information Journal* **32**, 213–220.

[87] Simon, R. (1999). Bayesian design and analysis of active control clinical trials, *Biometrics* **55**, 484–487.

[88] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

[89] Tseng, G.C., Oh, M.-K., Rohlin, L., Liao, J.C. & Wong, W.H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Research* **29**, 2549–2557.

[90] Turner, J.A., Deyo, R.A., Loeser, J.D., Korff, M.V. & Fordyce, W.E. (1994). The importance of placebo effects in pain treatment and research, *Journal of the American Medical Association* **271**, 1609–1614.

[91] Wagner, R.F., Beiden, S.V. & Campbell, G., Metz, C. & Sacks, W.M. (2002). Assessment of medical imaging and computer-assist systems: lessons from recent experience, *Academic Radiology* **9**, 1264–1277.

[92] Wittes, J. (2001). Clinical trials of the effectiveness of devices: An analogy with drugs, *Surgery* **129**, 517–523.

[93] Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P.R., Afshari, C.A. & Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models, *Journal of Computational Biology* **8**, 625–637.

[94] Yang, Y.H., Buckley, M.J., Dudoit, S. & Speed, T.P. (2002). Comparison of methods for image analysis on cDNA microarray data, *Journal of Computational and Graphical Statistics*, **11**, 108–136.

[95] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. & Ruzzo, W.L. (2001). Model-based clustering and data transformations for gene expression data, *Bioinformatics* **17**, 977–987.

[96] Yue, L.Q. (2001). Design issues in non-inferiority medical device clinical trials, *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*, Vol. M, #451. American Statistical Association, Alexandria.

[97] Zhou, X.-H., Obuchowski, N.A. & McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, New York.

[98] Zweig, M.H. & Campbell, G. (1993). Receiver operating characteristic plots: a fundamental evaluation tool in clinical medicine, *Clinical Chemistry* **39**, 561–577.

GREGORY CAMPBELL, TELBA Z. IRONY, CHANG S. LAO, HENG LI, GENE PENNELLO, R. LAKSHMI VISHNUVAJJALA & LILLY Q. YUE

Medical Ethics and Statistics

Although the first topic most people consider when ethics and statistics are mentioned is **clinical trials** (*see* **Ethics of Randomized Trials**), there are several other areas where uncertainty raises ethical difficulties. Professional codes of conduct, and the foundations of medical ethics, discuss avoiding harm and providing benefit. The existence of sound knowledge of what is beneficial or harmful is largely assumed, but there is often only limited information on the effects of a medical intervention. Issues of uncertainty cannot be avoided by professionals' claiming to treat individuals, as knowledge based on some group of people must be used in deciding what will be of benefit to the person currently seeking treatment. Doctors cannot behave ethically entirely individually, nor can they treat patients as isolated individuals, except in a limited sense. Medicine and health care are too complex to be within the understanding and control of one doctor or one profession.

If health care professionals are to be able to behave ethically, they must discover what effects their interventions have on a variety of people. Uncertainty, and summarizing the characteristics of populations, are the territory of statistics. Statistics provides the optimal methods for designing studies to gain knowledge, and for making **inferences** from limited empirical data [30]. So doctors and others need to work with statisticians, and professional societies need to ensure that some of their members have a fairly good understanding of statistics.

Various questions are raised by basing health care ethics on the seemingly uncontroversial statement that "the professional must do what is best for the patient". The criteria for judging what is "best" in routine medical practice, normality, screening for disease, epidemiological research, and the communication of uncertainty are discussed. Ethics of medical research in developing countries, popular versions of uninformed consent, and cluster randomized trials, are also addressed. Areas of philosophy other than ethics are also referred to, as the moral status of many decisions and actions in health care will be related to political philosophy, logic, and the theory of knowledge. The article concludes by relating these concerns to the guidelines or codes of the **Royal Statistical Society (RSS)**, the **American Statistical Association**, and the **International Statistical Institute** [4, 36, 48].

Criteria of Excellence

Attempts to measure and define **quality of life** highlight the complexity contained in the simple moral command "do what is best". Statisticians are familiar with having to choose from several possible criteria: for example, an estimator might be **unbiased**, but less precise than a biased estimator. Health care requires consideration of what one means by doing the "best" for a patient, as one will have to choose between methods that are most satisfying for the clinician to use, economical enough to appeal to a manager, or provide the patient with the best quality of life available to them. Mediating between different interest groups means that it will often not be trivial to decide what is best, and what data a statistician should collect or analyze.

The attempt to avoid this complexity, by considering only how to treat the person now in the room with me, fails, because few doctors would insist on continuing to remove a splinter from a toe when a person outside is choking to death. Further, most people are not isolated individuals with respect to health care, because the cost of care is shared either with a political unit such as a country, or by a group defined by membership of a particular insurance scheme. Often there will be environmental influences on health, such as malnutrition, or exposure to infection or pollution. Expenditure of effort, time, or money on one person will almost always limit expenditure on another; skills acquired in treating one patient might benefit another. Finding out what is best for patients will always involve observing relevant populations. The definitions of "relevant" and "population" are often not made explicit in medical practice, nor consciously related to the particular economy.

Another potential source of confusion and conflict is that the measurement of variables such as length and quality of life might rely on different theories of measurement [26] (*see* **Measurement Scale**), and be affected by beliefs about some populations being "less worthy" than others – consider the idea of restricting health care for smokers. Most statisticians might agree that it is wise to keep measures of quality of life simple, so that the comparison of values is

explicit, rather than averaging across incommensurate and contradictory variables [18].

Normal: The Exaltation of Mediocrity

It is not clear whether statistics should be credited, or blamed, for the development of the concept of "normal", and its use as a moral and evaluative category. Social scientists began to use the **normal** (Gaussian) distribution in the twentieth century, initially as a model for the behavior of measurement errors, but then to summarize measurements on living creatures [24]. Most biostatisticians will have had requests for help with establishing a "normal range" (*see* **Normal Values of Biological Characteristics**). The various problems with reference ranges, a more appropriate name, such as the choice of subjects used to choose the range, are discussed in the textbook by Altman [3]. If sufficient caution is not exercised in establishing reference ranges, considerable numbers of people might be inappropriately treated, or alarmed. Further, the naive use of such ranges might lead to people almost always being found to "need" treatment: if a reasonably large number of tests are done, there will be a high probability that at least one result will fall outside a reference range.

Very often, the major flaw in the use of reference ranges is that the wrong question is being addressed. Reference ranges are used to justify some action, such as prescribing cholesterol lowering diets or drugs. However, the salient issue is at what level of a variable there is good reason to intervene on the basis of adequate evidence that the intervention will achieve a worthwhile aim. The relevant goal must be reduction in mortality or morbidity. To define high cholesterol as such as an indicator of morbidity is to use a circular definition, which might create work and wealth for doctors and industries, but will not necessarily improve health. In the case of cholesterol, it is not clear that attempting to make a person's blood levels "normal" benefits them [23]. Attempting to achieve "normality" might amount to maltreatment.

In mental health, the issue of appropriate reference populations is particularly important. Many psychological tests are "standardized" using "normal" people who are undergraduates, patients, or nurses. One "evaluation" of a diagnostic questionnaire for depression claims that it is normal for women to be more depressed than men [37]. Who would claim that it is "normal" for more men than women to get lung cancer? Within a science, which is attempting to measure, say, the gravitational constant, it is reasonable to wish to have one's measurements clustering about the **mean**. There is no obvious reason why the general population should conform to some average, particularly not to the tastes and habits of students or nurses. To require this is at best a recipe for mediocrity, and at worst, a totalitarian definition of what ought to be.

Ethical and political philosophy include debate about what criteria should be used to judge the morality of actions. The consequences of careless thinking, and use of some, perhaps vague, idea of "normal" as "what ought to be" is nicely illustrated by a satirical proposal to classify happiness as a psychiatric illness [10]. Some of this careless thinking might be helped by an unnoticed shift in the theory of measurement from a representational theory in physics to an operational theory in psychology [26].

Another aspect of treating a patient as a "normal" member of a particular population which has disturbing consequences arises from differences in compliance (*see* **Compliance Assessment in Clinical Trials**). As there is evidence that about 10% of patients are excellent compliers, dose levels based on the **"intention to treat" analysis** of trials will be levels such that most patients will actually take a therapeutic dose. Excellent compliers will take a higher dose, and therefore be at greater risk of side effects. Although a doctor might recognize a need to warn patients of side effects, it is difficult to predict which people will be good compliers, and, simplistically, it is not normal to take medication as prescribed. The dose level set is thus, to some extent, set to benefit the majority at the expense of the "best" patients. Any claim to be treating a patient as an individual will have limited accuracy. As a doctor will never know everything, and is unlikely to know about a patient's compliance, she must decide on a course of action likely to benefit patients belonging to some type or subgroup.

Observational and Epidemiological Research

The author wishes it were unnecessary to state that a study must be scientifically sound in order to be ethical. Altman claims that the ethical implications

of unscientific studies include the misuse of patients and resources, and the consequences of publishing misleading results [2]. Although the suggestion that the misuse of statistics is unethical has not been challenged, substandard design and incorrect analysis can be seen in almost any issue of any medical journal. If Leaning's claim, that editors are forced by the Nuremberg code (the explicit statement occurs in the Helsinki code) not to publish information that has been unethically obtained, were true [38] and rigidly enforced, the size and number of biomedical journals would be greatly reduced.

Informed consent is usually required to do the main ethical work in clinical trials. The Nuremberg code states that for research involving human *experimentation* to be ethical, "the voluntary consent of the human subject is absolutely essential" [52]. The declaration of Helsinki also concentrates on experimental research [61]. As **observational** and epidemiological research are not experimental, in that the subjects' environment or treatments are not manipulated by the researchers, the Nuremberg code and its derivatives might be deemed irrelevant. However, the RSS Code of Conduct asks that informed consent be obtained, if possible, for all enquiries involving human subjects.

Informed consent for epidemiological research, including the use of routine data, has been introduced in the European Community. Many UK biostatisticians and health care professionals are opposed to this, as being inappropriate given the risk and benefits of such research. To require individual consent for each study would introduce sufficient bias due to **nonresponse** to render the studies almost useless. The main argument for informed consent regards individual privacy as the supreme right (*see* **Confidentiality**). However, the community, including the individual, is likely to benefit considerably from epidemiological studies of the origin and course of diseases [44].

Observational studies of the nonrandom introduction of innovative treatments, which some **Bayesian** philosophers advocate, [57], evade, rather than avoid, informed consent. The use of routinely collected data (*see* **Administrative Databases**) relies on people's consent to be governed. For example, in the United Kingdom, people who have cancer could not refuse to have data on their illness sent to Cancer Registries (*see* **Disease Registers**). Consent is now required for much health data, but participation in the **census** is still mandatory. In many instances, some of the subjects on whom data are collected are dead. One cannot get informed consent, and does not need it, as there is no person whose privacy is invaded. Is there even an ethical worry at all? Privacy is safeguarded through anonymity.

Screening

The moral issues raised by population **screening** are also inherently statistical. In order to assess whether the **incidence** of a disorder is being reduced by a screening programme, adequate statistical records of the population must be kept. This raises questions about what moral framework one claims in order to justify **surveillance** of a population, and thus takes the debate into political philosophy. There is also uncertainty for the individual because of the possibility of **false positives** and **false negatives**.

The justification for screening depends on the political framework one chooses to use, such as liberal, paternalist, or statist. Doctors usually use a paternalistic definition. A liberal definition, based on a right to the knowledge needed for self-determination, is sometimes used by patients. As one is considering people who are well, the justification must rely on people's consent to be governed; it cannot rely on their becoming patients by asking for assistance. Perhaps we think that people have a duty to be enquiring both medically and socially. It is therefore legitimate to require people to think of themselves as "patients-in-waiting", who must find out their state of health, and make decisions on the basis of that information. As mentioned above, the cost of health care is usually shared by a group. The person who chooses not to have screening, and later requires lengthy and expensive medical treatment for a disease for which screening was offered, arguably has caused their group (society) to incur unnecessary costs. An insurance-based system could demand submission to screening, but a liberal democratic state should only ask it. Financial inducements to doctors to increase screening rates leads to coercion of people. For this argument to have real force, one would need to have good estimates of the **sensitivity** and **specificity**, and of all the costs, so that the scope and frequency of the programme could be decided (*see* **Diagnostic Tests, Evaluation of**).

Instead of taking a statist view, one might take a liberal or modernist stance, and argue that screening

does not give people more decisions than they can make, but enhances their autonomy and makes them better (happier rather than morally better) adults by giving them more choices to make. However, people who are very concerned about their health are often labeled as "hypochondriacs", that is, ill. Of course, if one uses "normal" as one's criterion, then by requiring most people to be anxious about their health, one could almost eliminate the disease of hypochondria! Most liberal philosophies consider society to consist of rational people, and so would require a different justification for screening children, or those whose reason is incapacitated.

A person accepting screening will have to make decisions based on uncertain results. A Health Education Authority leaflet on breast screening states "... mammography is not 100% reliable". In prenatal screening, false positives sometimes vanish magically. The standard leaflet given out by general practitioners in the United Kingdom mentions false negatives, and quotes rates for Down's syndrome and neural tube defect (NTD). False positives are mentioned for the first screening test, but not for the second. A recent reference book [39] has a careful discussion of the need to consider **prognosis** as well as diagnosis. However, in the section on NTD (pp. 10–16), the data sets mentioned are rather small, and the prognosis is not uniformly poor. After stating that ultrasound scanning is preferable, because the sensitivity is "between 60 and 90%" and "specificity is much higher", a figure illustrating a screening programme is given, which has no false positives. Even with a sensitivity of 90%, given the **prevalence** used, 0.2%, to get a ratio of one unaffected to one affected fetus among those testing positive, one requires a specificity of 99.8%. People who debate the rights or wrongs of abortion for particular conditions rarely frame the question "at what level of probability is one justified in terminating a fetus?" Given the inherent uncertainty of prenatal diagnosis, those who wish to endorse abortion in these circumstances should be willing to make statements such as "if there is a 1 in 5 chance that the infant has NTD, then it is acceptable to terminate her".

For screening programmes to be even minimally ethical, under almost any political philosophy, the statistics of screening must be available. In all cases, sensitivity and specificity rates have to be *estimated*. The popular assumption that the results of tests, and hence diagnoses, are certain has to be challenged. Careful research can reduce, but never eliminate uncertainty. The well-rehearsed recommendations about careful design as well as correct analyses apply also to screening.

Conveying Uncertainty in Diagnosis and Prognosis

As uncertainty is unavoidable and decisions about care have to be made, health care should be a fertile field for **decision analysis** [28]. Uncertainty, in diagnosis or prognosis, is often dealt with by failing or refusing to provide any information. Until quite recently, parents whose child was diagnosed as having cerebral palsy would not have been given accurate information about the child's likely survival, as no such information existed [22]. Anecdotal evidence suggests that parents with severely handicapped children would have been told not to expect the child to live much beyond the age of 10, although the **median survival** is about 20 [35].

To a patient with lung cancer or colonic cancer, the information about the median survival ("half of people with the type of cancer you have live more than x months") is potentially very important in his decisions. A study of gastroenterologists showed that there is considerable variation across Europe in honesty and respect of confidentiality with regard to cancer [53]. The authors describe deciding whether to be dishonest as a "typical ethical dilemma", but deciding whether to be dishonest requires a decision about a temptation, not a decision about a principle, rather seeing lying as a temptation. Of course, the medical profession can be characterized as the only profession that debates whether to tell the truth. To deny information, which a consultant should either know, or be able to access, is irresponsible. The UK General Medical Council's guidance requires doctors to "give patients the information they ask for or need about their condition, its treatment and prognosis". It is not clear how such a consultant could justify a claim that the patient did not want the information, which they had requested. Attempted justifications include a concern to protect the image of the profession, and a belief that patients cannot cope with truth, or uncertainty: "professed uncertainty is in itself generally undesirable" [27]. In contrast, Bursztajn et al. [12] give an enthusiastic account of the benefits of realizing that we live in an uncertain world, and

of accepting that medical choices should explicitly acknowledge this. Uncertainty is often regarded as a justification for randomized controlled trials (RCT).

In communicating risk, "nondirective" counseling is commonly held to be desirable, perhaps ethically preferable to "directive" counseling or to the doctor stating her own beliefs. A moral doctor should convey uncertainty as clearly and impartially as possible. Observations of obstetricians who are responsible for counseling parents who require information about amniocentesis [41] revealed that the same **risk** was described as high or low, depending on the topic: miscarriage or Down's syndrome. If the same risk is described in contradictory ways in a session, which is claimed to be nondirective, then there is clear dishonesty. The Chief Medical Officer has also recommended establishing a scale of uncertainty to which the general public can relate. This was, in part, because a report of increased risk of blood clots for women taking particular contraceptive pills received wide coverage in the media; the risks associated with pregnancy were not given the same prominence. In his presidential address to the RSS, Smith suggested establishing a scale of risks [50] (*see* **Risk Assessment**).

In the debate about the moral superiority of Bayes inference in assessing whether new treatments are effective, philosophers on both sides show a reluctance to acknowledge the reality of uncertainty. In an article extolling **randomization**, Papineau [46] makes a breathtaking leap from probability to certainty: "... if it turns out that T makes no probabilistic difference to R either among young people, or among old people – then we can conclude that T *doesn't* cause R, ...". The emphasis is the author's, who appears to be unaware of the concept of **power**. In arguing that randomization is unnecessary because Bayes inference can adjust for all possible **confounding** factors, Urbach [56] indicates his belief in the lack of uncertainty in medicine, which no one who is aware of, for example, the inaccuracy of **death certificates**, or the limited accuracy of prediction for survival, could espouse [16].

Medical Research Ethics in Developing Countries

The guidelines published by the **World Health Organisation** (WHO) and the Council for International Organisations of Medical Sciences (WHO/CIOMS), [17], were framed with particular concern for developing countries [29], not to duplicate the principles already established, but to suggest how these principles might be applied. Much of this is useful, but some principles verge on racism.

> *"Subjects in developing countries*
>
> 14. Rural communities in developing countries may not be conversant with the concepts and techniques of experimental medicine. It is in these communities that diseases not endemic in developed countries exact a heavy toll of illness, incapacity and death. Research on the prophylaxis and treatment of such diseases is urgently required and can be finally carried out only within the community of risk.
> 15. Where individual members of a community do not have the necessary awareness of the implications of participation in an experiment to give adequately informed consent directly to the investigators, it is desirable that the decision whether or not to participate should be elicited through the intermediary of a trusted community leader. The intermediary should make it clear that participation is entirely voluntary, and that any participant is entirely free to abstain or withdraw at any time from the experiment."

This singling out of rural communities is offensive, as the assumption that rural people in developing countries are less able to give informed consent is not justified [19]. A review of the ethics of clinical trials, and the sociocultural contexts, found no evidence of cultural objections or obstacles to voluntary consent [7].

Remarks CIOMS.18 and 19 on review procedures mention the role of statisticians, but recognize practical and political realities. As statisticians, we can welcome the recognition of our potential contribution, but even "highly developed" countries do not have an adequate supply of statisticians to support ethics committees [59]. The requirements CIOMS.27 and 28 for ethical review in both host and "external" countries are noteworthy. As recognized under review procedures, different countries have different resources for ethics committees, which require considerable expenditure, at least of time [60]. Concern was expressed about the independence of ethics committees in some countries at the international school from which the proceedings arise. A longer term view of the impact of research interventions is required by CIOMS.29, and the wider social context is stressed by CIOMS.32.

A heated debate about research in developing countries started by trials of interventions to reduce transmission of HIV from mothers to their infants, before and after birth (*see* **AIDS and HIV**). Lurie and Wolf [40] noted that an intervention, the AIDS Clinical Trials Group study 076 (ACTG076) regimen of an antiretroviral drug, AZT, had been shown to be effective in 1994, but, despite this, in many later trials of vertical HIV transmission, some or all patients were not provided with antiretroviral drugs. Of 18 studies identified, 15 used **placebo** controls. The two USA trials provided antiretroviral drugs for all patients.

Bayer makes the important point that the real ethical problem is not whether to use placebos, but the immorality of the world economic order [8]. The "maldistribution of wealth and resources" makes the vertical HIV transmission trials a focus of (emotional) outrage. The goal of reducing HIV transmission in Africa requires information on affordable, implementable interventions, which will be the basis of health care policy. Matchada cites the failure of to eradicate tuberculosis, despite "free drugs", because of infrastructure barriers [42]. He compares expenditure by African nations on war and debt servicing with that of health care. The effect of exploiting medical care for commercial gain, on a worldwide scale, is investigated by Benatar [9]. Theological equality of persons exists, but not socioeconomic equality.

Wider issues are raised by Annas and Grodin [5], who place the debate in the context of the UN Declaration of Human Rights (DH). The goal of slowing the HIV epidemic might not be most sensibly achieved by addressing vertical transmission. A fuller review of the background and arguments, and discussion of the role of statisticians is given in [32]. A useful critical overview of the philosophical debate is given by Schüklenk and Ashcroft [49]. They point out that it is doubtful that an identifiable local standard of care exists, because the standard of care in, say, Ivory Coast, depends on prices set by Western manufacturers of drugs and equipment. In order to argue that particular principles are inappropriate, one has to present a case showing, for example, that protection of profits by patent and monopoly is more important than limiting the course of an epidemic.

Although health professionals might prefer to own "their" codes, the revisions of DH were not proposed for medical reasons, and the debate is between ethics, including distributive justice, and economic prudence [5, 49]. Schüklenk and Ashcroft suggest that ownership of DH might be passed to the United Nations [49]. Even if the revision of DH were simply a matter of ethics, the application and interpretations of human rights requires law, policy, and hence politics. Adjudication among groups with different interests is a political responsibility, and ways of including the public in the debate are required [49].

The 2000 revision of DH illustrated the need for collaboration. The sentence "This does not exclude the use of placebo, or no treatment, in studies where no proven prophylactic, diagnostic or therapeutic method exists" [27, 61] was promptly criticized by those who carry out trials. It is precisely when proven therapies exist that placebo controls are used. In 2002, a footnote was added to DH.

Informed and Uninformed Consent

It is valuable to realize that there are various aspects and interpretations of informed consent [1]. Guidelines or codes of conduct usually focus on consent as involving delivery of information in a manner that respects the rights of the person. At one extreme, this is viewed as a polite ceremony that is not essential, as doctors always have their patient's best interests to the fore. Consent is convenient because it transfers responsibility from the doctor or researcher to the patient or subject. The other extreme views consent as necessary protection against useless, dangerous, or unwelcome interventions imposed by a powerful profession. Some people argue that requesting consent changes patients from research objects to research subjects [60], but others, both patients and doctors, feel that the process can be detrimental to the patient [27, 55].

Two systematic reviews of reasoning [6] and empirical data [20] on patients' understanding of informed consent noted that there are difficulties. However, the authors recommend that informed consent remains essential [6] or that "... the spirit of informed consent" be retained and seriously attempted, with ethics committees for further protection [20]. It is possible that the protection of patients that informed consent is intended to provide [25, 52], might be achieved by alternative forms of consent. Alternative proposals that might give such critics of informed consent and randomized trials what they say they want have been examined [34].

It is important to consider carefully what information should be provided in order to allow a patient to make an informed choice. It is arguably unethical to impose one's own standards of understanding on a patient. If a patient refuses to participate in a particular trial, we do not require this refusal to be informed, as this could be regarded as coercive. There is a dramatic contrast between the standards of ordinary medical practice and randomized trials [15].

The invitation, at a time when the patient is already under stress, to participate in a clinical trial can be upsetting [54]. It is worth considering patients' informed response to trials in general. Those who realize that trials might well be beneficial, irrespective of treatment received, are likely to want to be offered enrollment in any trial for which they are eligible [15, 20]. The idea of general prior refusal is natural enough, if a person believes that randomized trials are wrong. If the alternative to randomized trials is uncontrolled, unreliable experimentation, rather than no experimentation [14], then it is difficult to make general prior refusal entirely coherent. One can sketch the justification for having "trial-free" doctors under various constructions of professionalism [34].

Pre-exclusion of patients by doctors on grounds of guesswork about patient preferences already exists [51]. If a doctor thinks that "professed uncertainty is of itself undesirable" [27], they might not offer patients the opportunity to participate in a trial. Alternatively, if a doctor takes views the trial as the treatment, they might (logically, though not legally) use the same standard of lack of consent used in routine treatment, and enroll the patient without drawing attention to randomization, as probabilistic choices are common in medical practice.

The irony of the requests for alternatives to informed consent is that to grant them requires restrictions on patients' knowledge, personal responsibility, and freedom of choice. This implies inadequate patient protection, free-riding, unreasonable avoidance of decision-making, increased decision-making by doctors, and (self-)exclusion from optimal treatment.

Two moral concerns are addressed in informed consent: individual self-determination, or autonomy, and justice, particularly justice for groups of people such as Jewish or black people. Randomized trials do not sacrifice present patients for future patients [31]. In contrast, to insist that obtaining informed consent is more important than providing the therapeutic package most likely to lead to a better outcome for this patient, is to sacrifice this patient for the protection of other citizens. We insist on subjecting patients to the experience of being asked for informed consent, because we know that it is very dangerous to allow doctors to decide whether an experiment on people is justified [25]. If we dispense with informed consent, some group of people will suffer abuse. Even with official recognition of the need for informed consent, groups of patients sometimes suffer [58].

Cluster Randomized Trials

Cluster randomized designs (CRDs) are increasingly used in research into health care and health services (*see* **Group-randomization Designs**). Ethics of individual patient randomized trials have been elucidated in a number of different codes, but less attention has been given to the ethical issues raised by cluster randomized trials. The challenges raised by cluster randomized controlled trials are evaluated by considering the essential elements of ethical medical research, particularly experiments on people, and the distinguishing features of cluster randomized controlled trials from ordinary RCTs [33].

Cluster-randomization designs are experiments in which intact social units are randomly allocated to one of two or more intervention or treatment strategies. There are scientific and practical reasons that can be given to justify the use of CRDs. The scientific reasons are that intervention might act at cluster level, (e.g. a vaccine) or be carried out at cluster level (e.g. guidelines for medical doctors), that there might be treatment contamination if participants can exchange information, or that subject compliance can be enhanced by discussions. Logistical and political constraints include administrative convenience, political necessity for permission to be obtained and requirements of access to routine data.

The fact that the unit of randomization includes several patients or participants has implications for both consent and the science of CRDs.

Protocols must make adequate provision for professional statistical input in order to be scientifically sound, and hence ethical in the light of the Nuremberg principles 2, 3, 6 & 8. The need for appropriate methods of analysis is relevant both to the design, in assessing previous knowledge, and the analysis of the results. An important issue in ensuring that a

study will be able to yield useful information (Nuremberg 2) is the sample size, and the effective size of a CRD is smaller than the total number of individuals studied. Few CRDs allow for between-cluster variation when estimating the power of trials, and not all reports of CRDs allow for the clustering in the analysis. Thus, any summary of previous knowledge might be affected by previous errors. The funding of a new CRD obviously must include provision for specialist skills to ensure correct analyses of past and current data and a thorough analysis of the risks and benefits (Nuremberg 6). Resource use requires social and political categories of thought.

Decisions about early stopping are dependent on planned and unplanned interim analyses, in CRDs as in RCTs (Nuremberg 10). The decisions might well be more complicated in CRDs, as there might be a need to accrue sufficient numbers in each cluster, and the point at which there is convincing evidence of benefit or harm might be very difficult to discern (*see* **Data and Safety Monitoring**). Disadvantages associated with early stopping of a RCT, such as lack of credibility and realism, imprecision, and bias [47] will be accentuated for CRDs.

The important structural features of CRDs for consideration of informed consent are the units of randomization or allocation, units of intervention, and units of observation. The various units are not necessarily the same people, as patients might receive treatment, but the conduct of nurses be the focus of observation and a general practice the unit of randomization. There might be gate-keepers; alternative interventions might not be easily available and in some instances, a participant cannot easily withdraw, for example, if an insecticide is sprayed throughout a village. In the case in which a professional is the primary experimental subject, if she chooses to leave a trial early, she will effectively remove all her patients also (cf Nuremberg 9).

The different levels of randomization and intervention mean that there are potentially various types of consent and levels at which it can be sought. Consent might be sought, or not sought, at some of the various levels, for use of routinely held data, for collection of additional data, with or without the use of invasive procedures, or for the offer, or administration, of an intervention. We usually think of consent as operating at the level of the individual person: with CRDs, there are further levels to consider. The definition of a community, and how its representatives are chosen, takes the debate on ethics into political philosophy. A Nigerian study indicates that the leaders' views, which are cheaper and easier to obtain, cannot be relied on as proxies for the opinions of heads of households [45]. There is no firm evidence that in any cultures heads of household might give or withhold consent for adults in their household [6].

Guidelines of WHO/CIOMS accept the possibility that individual consent is not feasible. The decision to undertake the research is then given to a "public health authority", with attention given to providing the community with information on the research [29]. This authority must therefore take responsibility for the consequences of the research, although the people who are such authorities might not themselves be directly exposed to the interventions. However, it is not obvious that there will be only one authority which can, or should, take such decisions and responsibility. Here again the political dimension of research ethics is clear. The Nuremberg code was drawn up as a result of the failure of political authorities to protect all groups of citizens [58]. Mere feasibility is not a strong reason to fail to request individual consent. The fact that subjects might not be able to avoid the treatment, although they do not consent to it, or might not otherwise have access to treatment, is a reason not to impose the treatment, not a reason to evade individual voluntary informed consent.

Consent should be obtained before any intervention, but it is not ethically essential that it is obtained before decision as to what intervention would be offered if the person were to agree to enter the trial. One seeks consent to be in an experiment, not simply consent to a particular treatment. The primary reason for obtaining consent before randomization is a scientific one: it reduces the possible bias arising from different patterns of consent in the various treatment arms. Scientific and logistical constraints associated with CRDs imply that consent cannot necessarily be requested before an intervention is assigned to a person. This is not a problem, as trial entry and treatment assignment are not equivalent. For example, a person could refuse the vaccination, after their community had been assigned this intervention, which operates at both community and individual level.

Guidelines on the scientific and ethical conduct of CRDs are provided by the UK Medical Research Council [43].

Statistical Ethics

The *International Statistical Institute Declaration on Professional Ethics* [36] recognizes the variety of settings in which statisticians work, and the many branches of the discipline. The Declaration is therefore an informative framework of principles, not a set of regulations. Each principle is followed by a commentary and bibliography. The intention is that statisticians who consider departing from the principles do so as a result of deliberation, not ignorance. The principles are grouped into four categories, with no category having priority: Obligations to society, funders and employers, colleague, and subjects are considered.

Social obligations comprise considering conflicting interests, including guarding against misuse and misinterpretation of statistics; extending the scope of statistics to benefit as large a community as possible; and pursuing objectivity with openness about limitations. Obligations to paymasters require clarity about roles and responsibilities; impartial assessment of alternative statistical methods; no pre-emption of outcomes, and safeguarding privileged information, while revealing the statistical methods and techniques used. In return for respect for exclusive technical and professional knowledge, statisticians must be honest about the limits of their expertise.

The three obligations to colleagues described are maintaining confidence in statistics, transparency of methods, and knowing one's own ethical principles as well as those of one's collaborators. These principles arise from the value of professional citizenship, which confers privileges of access to data, and the dependence of the reputation of statistics on the conduct of individual statisticians. A difficult responsibility is neither ". . . overstating or understating the validity or generalizability of data . . ." (ISI.3.3).

Animals as subjects are acknowledged, but the subjects to whom obligations are described are individual people, households, and corporate entities. Statisticians should avoid intrusion. There is an excellent discussion of the implications of the obligation to obtain informed consent, in terms of adequacy of information and of consent. Statisticians are expected to ". . . adhere to the principle of obtaining informed consent directly from subjects" even if they first have to negotiate with a "gate-keeper" who is blocking access. Careful consideration of modifications to informed consent addresses observation studies, dealing with proxies, secondary use of records and misleading potential subjects. With respect to the last, withholding information is deceitful, and instances when legitimate censure can be avoided because of special research requirements are rare, and difficult to justify. In such cases, posthoc consent should be considered.

The interests of subjects must be protected, not merely within the study, but also with regard to subjects' relationships with their environment. Social position can hold risks: "The interests of subjects may also be harmed by virtue of their membership of a group or section of society (see Clause 1.1). So statisticians can rarely claim that a prospective inquiry is devoid of possible harm to subjects. They may be able to claim that, as individuals, subjects will be protected by the device of anonymity. But, as members of a group or indeed as members of society itself, no subject can be exempted from the possible effects of decisions based on statistical findings." Confidentiality of records is essential, as is inhibiting disclosure of identity by providing configurations of attributes, which are distinctive.

The *RSS Code of Conduct* [48] primarily describes the professional duties of a statistician, with a view to upholding the reputation of the profession. Nevertheless, some of the rules do address ethical matters. Part of the context for the rules is the recognition that the general public have no easy way of judging the quality of statistical work.

The "Public Interest" is the focus of the first two rules. Fellows of the RSS are required to have knowledge of, and comply with, the legislation, regulations, and standards "relevant to their chosen field". Hence any statistician involved in refereeing articles or grants that have a component of medical research should be familiar with the Nuremberg code [52] and the Declaration of Helsinki [61]. The second rule has wide implications, as it expects Fellows to avoid any actions that damage basic human rights:

> Fellows shall in their professional practice have regard to basic human rights and shall avoid any actions that adversely affect such rights. Enquiries involving human subjects should, as far as practicable, be based on the freely given informed consent of subjects.

Although the adjectives qualifying "informed consent" shows more imagination than some codes, the subordinate clause recognizes that there can be

studies for which it is not practical to obtain informed consent. Confidentiality should always be respected.

With respect to statisticians' duties to employers or clients, the RSS Code acknowledges that the professional judgment of statisticians may be overruled. Most medical statisticians will have experienced this, in collaboration, consultancy, refereeing, or as a member of an ethics committee or regulatory body. In such circumstances, the RSS code (RSS.3) requires the Fellow to indicate the likely consequences of ignoring their judgments. It is possible to understand why point 4 of the RSS code asks fellows to try to avoid becoming party to activities that conflict with the basic public interest. It is not clear why fellows should avoid becoming "privy to information" concerning activities that would conflict with their public responsibilities. If a statistician can discern that a study ignores basic human rights because it is poorly planned, and is using people and resources in an endeavor that cannot result in any useful information, rather than trying to avoid such information, finding it and advising appropriate authorities of it so that the misuse of resources can be terminated would be acting in the public interest [13, 21]. Statisticians are prohibited (RSS.6) from allowing their name be associated with "any misleading summary of data", with particular attention to "the way the data were selected". Two other common concerns are: accurate description of reasons and assumptions behind the method of analysis, and giving opinions not supported directly by the data reported.

The American Statistical Association's *Ethical Guidelines for Statistical Practice* [4] emphasize the duty of statisticians to maintain professional integrity. In particular, they should provide honest and objective interpretation, based on evidence, with disclosure of any special interests. Statisticians have a responsibility to respondents who provide data, especially with respect to privacy and confidentiality. This includes establishing informed consent, and detailed concerns about offering and ensuring confidentiality. Statistical work must be open to assessment, with the limits and source of data made clear, and the role of statistical analysis, including choice of procedures visible. Data should be available for analysis by appropriate others. As users of statistics may be dependent on expert advice, good conduct and good communication are essential. The guideline on collecting "only the data needed for the purpose of their inquiry" is worth bearing in mind, as accuracy can be discouraged by excessive requests. The tendency in medical research to request information "while we are there" can add unnecessarily to the cost in time and effort of clinical research.

Conclusions

Although most of us use "Nazi medicine" as the epitome of unethical research, it is not trivial to explain exactly what was wrong [11], why the atrocities occurred in the one country that had a legal doctrine of informed consent and medical ethics [58] and why no use should be made of data that are improperly collected [38]. Doctors cannot easily deny the willing involvement of their profession in **eugenics**. Eugenics was widely supported in the 1920s and 1930s; new eugenics are popular now in antenatal screening. Statisticians cannot thank God that they are not like other men: **Galton** coined the term "eugenics".

The most interesting ethical problems arise from the use to which the statistics will be put. Any reasonable code of conduct will require statisticians to exercise competence and diligence, to be willing to consider their fallibility, and to be vigilant in the communication of the results of analyses and their interpretation, regardless of the area of application.

References

[1] Alderson, P. & Goodey, C. (1998). Theories of consent, *British Medical Journal* **317**, 1313–1315.

[2] Altman, D.G. (1982). Misuse of statistics is unethical, in *Statistics in Practice*, S.M. Gore & D.G. Altman, eds. British Medical Association, London, pp. 1–2.

[3] Altman, D.G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall, London.

[4] American Statistical Association Committee on Professional Ethics. (1983). *Ethical Guidelines for Statistical Practice*. American Statistical Association, Alexandria.

[5] Annas, G.J. & Grodin, M.A. (1998). Human rights and maternal-fetal HIV transmission prevention trials in Africa, *American Journal of Public Health* **88**, 560–563.

[6] Ashcroft, R.E., Chadwick, D.W., L, S.R., T, R.H., Frith, L. & Hutton, J.L. (1997). Implication of sociocultural contexts for the ethics of clinical trials, *Health Technology Assessment* **1**(9), iv+65.

[7] Ashcroft, R.E., Chadwick, D.W., L, S.R., T, R.H., Frith, L. & Hutton, J.L. (1998). Implication of sociocultural contexts for the ethics of clinical trials, in *Health Services Research Methods: A guide to best practice*, N. Black, J. Brazier, R. Fitzpatrick & B. Reeves, eds. *British Medical Journal*, London, pp. 108–116.

[8] Bayer, R. (1998). The debate over maternal-fetal HIV transmission prevention trials in Africa, Asia, and the Caribbean: racist exploitation or exploitation of racism? *American Journal of Public Health* **88**, 567–570.

[9] Benatar, S.R. (1998). Global disparities in health and human rights: a critical commentary, *American Journal of Public Health* **88**, 295–300.

[10] Bentall, R.P. (1992). A proposal to classify happiness as a psychiatric disorder, *Medical Ethics* **18**, 94–98.

[11] Biagioli, M. (1992). Science, modernity and the "Final Solution", in *Probing the Limit of Representation: Nazism and the "Final Solution"*, S. Friedlander, ed. 371–377, Harvard University Press, Cambridge, pp. 185–205.

[12] Bursztajn, H.J., Feinbloom, R.I., Hamm, R.M. & Brodsky, A. (1990). *Medical Choices, Medical Chances*, 2nd Ed. Routledge, New York and London.

[13] Buyse, M., George, S.L., Evans, S., Geller, N.L., Ranstam, J., Scherrer, B., Lesaffre, E., Murray, G., Edler, L., Hutton, J.L., Colton, T., Lachenbruch, P. & Verma, B.L. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials, *Statistics in Medicine* **18**, 3435–3451.

[14] Chalmers, I. (1983). Scientific inquiry and authoritarianism in perinatal care and education, *Birth* **10**, 151–164.

[15] Chalmers, I. & Lindley, R. (2000). Double standards on informed consent to treatment: ignored for a quarter of a century by most professional medical ethicists, in *Informed Consent in Medical Research Respecting Patients' Rights in Research and Practice*, L. Doyal & J.S. Tobias, eds. BMJ Publications, London, pp. 113–125.

[16] Christakis, N.A. & Lamont, E.B. (2000). Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study, *British Medical Journal* **320**, 469–473.

[17] Council of the International Organisation of Medical Sciences. (1993). *International Guidelines for Biomedical Research Involving Human Subjects*. esp. Guidelines 1–3.

[18] Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Spiegelhalter, D.J. & Jones, D.R. (1992). Quality of Life: Can we keep it simple? (with discussion), *Journal of the Royal Statistical Society, Series A* **155**, 353–393.

[19] Edwards, S.J.L., Braunholtz, D.A., Lilford, R.J. & Stevens, A.J. (1999). Ethical issues in the design and conduct of cluster randomised controlled trials, *British Medical Journal* **318**, 1407–1409.

[20] Edwards, S.J.L., Lilford, R.J., Braunholtz, D.A., Jackson, J., Hewison, J. & Thornton, J. (1998). Ethical issues in the design and conduct of randomised controlled trials, *Health Technology Assessment* **2**(15), vi+128.

[21] Evans, S. (1993). Statistical aspects of the detection of fraud, in *Fraud and Misconduct in Medical Research*, S. Lock & F. Wells, eds. British Medical Journal, London, pp. 61–74.

[22] Evans, P.M., W, S.J. & Alberman, E. (1990). Cerebral palsy: why we must plan for survival, *Archives of Disease in Childhood* **65**, 1329–1333.

[23] Gallerani, M., Manfredini, R., Caracciolo, S., Scapoli, C., Molinari, S. & Fersini, C. (1995). Serum cholesterol concentrations in parasuicide, *British Medical Journal* **310**, 1632–1636.

[24] Hacking, I. (1975). *The Emergence of Probability*. Cambridge University Press, London.

[25] Hanauske-Abel, H.M. (1996). Not a slippery slope or sudden subversion: German medicine and national socialism in 1993, *British Medical Journal* **313**, 1453–1463.

[26] Hand, D.J. (1996). Statistics and the theory of measurement, *Journal of the Royal Statistical Society, Series B* **159**, 445–492.

[27] Hilden, J. & Habbema, J.D.F. (1987). Prognosis in medicine: an analysis of its meaning and roles, *Theoretical Medicine* **8**, 349–365.

[28] Hilden, J. & Habbema, J.D.F. (1990). The marriage of clinical trials and clinical decision science, *Statistics in Medicine* **9**, 1243–1257.

[29] Howard-Jones, N. (1981). Human experimentation in historical and ethical perspectives, *Social Science and Medicine* **16**, 1429–1448.

[30] Hutton, J.L. (1995). Statistics is essential for professional ethics, *Journal of Applied Philosophy* **12**, 253–261.

[31] Hutton, J.L. (1996). The ethics of randomised controlled trials: a matter of statistical belief, *Health Care Analysis* **4**, 95–102.

[32] Hutton, J.L. (2000). Ethics of medical research in developing countries: the role of international codes of conduct, *Statistical Methods in Medical Research* **9**, 185–206.

[33] Hutton, J.L. (2001). Are distinctive ethical principles required for cluster randomised controlled trials? *Statistics in Medicine* **20**, 473–488.

[34] Hutton, J.L. & Ashcroft, R.E. (2000). Some popular versions of uninformed consent, *Health Care Analysis* **8**, 41–52.

[35] Hutton, J.L., Cooke, T. & D, P.O. (1994). Life expectancy in children with cerebral palsy, *British Medical Journal* **309**, 431–435.

[36] International Statistical Institute. (1985). *Declaration of Professional Ethics*. International Statistical Institute, Voorburg.

[37] Knight, R.G., Waal-Manning, H.J. & Spears, G.F. (1983). Some norms and reliability data for the State-trait anxiety inventory and the Zung Self-rating depression scale. *British Journal of Clinical Psychology* **22**, 245–249.

[38] Leaning, J. (1996). War crimes and medical science, *British Medical Journal* **313**, 1413–1415.

[39] Lilford, R.J. (1990). *Prenatal Diagnosis and Prognosis*. Butterworths, London.

[40] Lurie, P. & Wolfe, S.M. (1997). Unethical trials of interventions to reduce perinatal transmission of the

human immunodeficiency virus in developing countries, *New England Journal of Medicine* **337**, 853–856.

[41] Marteau, T.M., Plenicar, M. & Kidd, Jane. (1993). Obstetricians presenting amniocentesis to pregnant women:practice observed, *Journal of Reproductive and Infant Psychology* **11**, 3–10.

[42] Matchaba, P. (1999). Are African governments allocating sufficient resources tofight AIDS? *British Medical Journal* **318**, 1351.

[43] MRC clinical trial series. (2002). *Cluster Randomised Trials: Methodological and Ethical Considerations*. Medical Research Council, London.

[44] MRC ethics series. (2000). *Personal Information in Medical Research*. Medical Research Council, London, New guidance on Health and Social Care Act 2001: "Section 60" added – January 2003.

[45] Onwujekwe, O., Shu, E. & Okonkwo, P. (1999). Can community leaders' preferences be used to proxy those of the community as a whole? *Journal of Health Services Research & Policy* **4**, 133–138.

[46] Papineau, D. (1994). The virtues of randomization, *British Journal for the Philosophy of Science* **45**, 437–450.

[47] Pocock, S.J. (1992). When to stop a clinical trial, *British Medical Journal* **305**, 235–240.

[48] Royal Statistical Society. (1993). *Code of Conduct*. Royal Statistical Society, London.

[49] Schüklenk, U. & Ashcroft, R.E. (2000). International research ethics, *Bioethics* **14**, 158–172.

[50] Smith, A.F.M. (1996). Mad cows and ecstasy: chance and choice in an evidence-based society (the address of the president), *Journal of the Royal Statistical Society, Series A* **159**, 367–383.

[51] Taylor, K.M., Margolese, R.G. & Soskolne, C.L. (1984). Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer, *New England Journal of Medicine* **310**, 1363–1367.

[52] The Nuremberg Code. (1947). *British Medical Journal*, **313**, 1449, 1996.

[53] Thomsen, O.O., Wulff, H.R., Martin, A. & Singer, P.A. (1993). What do gastroenterologists in Europe tell cancer patients? *Lancet* **341**, 473–476.

[54] Thornton, H. (1994). Clinical trials – a brave new partnership, *Journal of Medical Ethics* **20**, 19–22.

[55] Toynbee, P. (1996). No one really wins in this life-and-death lottery, *The Independent* 25 May, p.15 (cols 1–7).

[56] Urbach, P. (1985). Randomization and the design of experiments, *Philosophy of Science* **52**, 256–273.

[57] Urbach, P. (1993). The value of randomization and control in clinical trials, *Statistics in Medicine* **12**, 1421–1432.

[58] Vollmann, J. & Winau, R. (1996). Informed consent in human experimentation before the Nuremberg code, *British Medical Journal* **313**, 1445–1450.

[59] Williamson, P.R., Hutton, J.L., Bliss, J., Blunt, J., Campbell, M.J. & Nicholson, R. (2000). Statistical review by research ethics committees, *Journal of the Royal Statistical Society, Series A* **163**, 5–13.

[60] Woodward, B. (1999). Challenges to human subject protections in US medical research, *Journal of the American Medical Association* **282**, 1947–1952.

[61] World Medical Association Declaration of Helsinki. (1964, 2000). Ethical principles for medical research involving human subjects. `http://www.wma.net/`.

Further Reading

Irwin, A. (1995). *Citizen Science*. Routledge, London.

J.L. HUTTON

Medical Expenditure Panel Survey (MEPS)

The Medical Expenditure Panel Survey (MEPS) provides nationally representative data on **health care utilization**, expenditures, insurance coverage, sources of payment, and access-to-care measures at the individual and family level. The survey was designed to facilitate analyses of how individual characteristics, behavioral factors, and financial arrangements affect health care utilization and expenditures (*see* **Surveys, Health and Morbidity**). MEPS is sponsored by the Agency for Health care Research and Quality (AHRQ) and cosponsored by the **National Center for Health Statistics**, **Centers for Disease Control** and Prevention (NCHS/CDC).

Since its inception in 1996, MEPS has been a continuous ongoing survey of the US civilian noninstitutionalized population. Predecessors were once-a-decade surveys: the 1977 National Medical Care Expenditure Survey and 1987 National Medical Expenditure Survey.

Survey Design and Content

The MEPS is a family of three-interrelated surveys of the US civilian noninstitutionalized population: the Household Component, the Medical Provider Component, and the Insurance Component. The MEPS Household Component contains data on health care use, medical expenditures, sources of payment and insurance coverage, **health status**, demographics, employment, and access to health care. Households are selected for the annual Household Component from those participating in the previous year's National Health Interview Survey (NHIS), an ongoing annual household survey of approximately 42 000 households (109 000 individuals) conducted by NCHS/CDC to obtain national estimates of health care utilization, health conditions, health status, insurance coverage, and access. Combined use of NHIS and MEPS data adds capacity for longitudinal analyses. MEPS has a multistage, clustered sample design (*see* **Multistage Sampling**) with 195 primary sampling units (PSUs) [3]. Sampling weights are used to produce population estimates for individuals, families and population subgroups, such as the elderly and children.

The survey employs an overlapping **panel study** design in which any given sample panel is interviewed in person 5 times over 30 months to yield annual use and expenditure data for two calendar years. **Computer-assisted** personal interview (CAPI) is used in an interview with a family respondent who reports for him/herself and for other family members [5]. In the initial year of the survey (1996), the household sample consisted of 8655 families and 21 571 individuals with calendar year data. In 1997, the MEPS sampled 13 087 families and 32 626 individuals, with oversampling of: Hispanics, blacks, adults with functional impairments, children with limitations in activities, individuals predicted to incur high-levels of medical expenditures, and low-income households. Since 1997, data from two panels are combined to produce estimates for each calendar year. Since 2002, the MEPS sample seeks 15 000 families and 40 000 individuals yearly.

The MEPS Medical Provider Component collects detailed data on expenditures and sources of payment from the medical providers, facilities, and pharmacies that serve individuals surveyed for the Household Component. These data are the primary source for imputing medical expenditure data to correct for item **nonresponse** by the MEPS household sample participants [8].

Medical providers (MD/DO) for households where expenditure data was expected to be particularly insufficient were sampled at higher rates, for example, households with any Medicaid enrollees or with HMO enrollees. All hospitals providing inpatient and/or outpatient services to household members are contacted. The data from medical providers include: the dates, medical content, and charges and payment sources associated with each encounter. Data from pharmacies describe each prescription: fill date; drug name, dose and NDC code; charges and payments by source. Hospitals self-administer their data collection; physician offices are contacted by telephone; pharmacies receive a mail survey with telephone follow-up. Since 2002, each annual Medical Provider Survey involves interviews with more than 4000 hospitals and related outpatient facilities, 22 000 office-based providers, 11 000 hospital-identified physicians, 800 home health providers and 9000 pharmacies.

The MEPS Insurance Component was designed to produce national, regional, and state estimates of the amount, types, and costs of job-related health insurance. Mail interviews are conducted annually with

30 000 establishments to support estimates of health insurance availability at the workplace, to describe the types of employer-sponsored coverage and their associated costs. Establishment survey data include: size, the type of workforce employed, aggregate data on payroll and available fringe benefits, industrial classification, corporate status, and the number and characteristics of health plans offered [9]. Data are collected for each plan: its scope and breadth of benefits; copayments; number of current workers and retirees enrolled; and whether it is fully or self-insured. Since 2000, the Bureau of Economic Analysis has used MEPS health insurance premium cost data to estimate the health component of the Gross Domestic Product. States also use the data to assess time trends in employer-provided health benefits, and to compare their employers' health insurance cost experience to national, regional, and other states' profiles.

National Estimates of Health Care Expenditures and Coverage

Health care expenditures representing over one-seventh of the US Gross Domestic Product are growing faster than other sectors of the economy and consume much of the Federal and states' budgets. Researchers have used the MEPS data to determine the direction of the association between the use of newer drugs and all other types of nondrug medical spending [7], and to identify inappropriate medication use, a major patient safety concern with significant cost consequences [10].

Health care spending is highly concentrated. The 1996 MEPS found the top one percent of the population accounting for 27% of the total health care expenditures incurred by the civilian noninstitutionalized population, and the top five percent of people, accounting for 55% [2]. Consequently, the MEPS uses oversampling and **poststratification** (to conform to more accurate population estimates of decedents) to improve the quality of survey estimates for this policy-relevant population subgroup.

Access to health insurance coverage is a critical public policy issue. The MEPS data support estimates of the size and composition of the insured and uninsured populations, and reveal how demographic characteristics, economic factors, and health status affect health plan eligibility and decisions to enroll in health insurance plans. In addition to providing cross-sectional estimates of health insurance coverage each year, the MEPS data can identify individuals with gaps in coverage over a calendar year as well as the duration of gaps for up to 24 months. From 1996 to 1999, between 59 million and 62 million Americans were uninsured at some point each year [6]. The MEPS data support estimates of out-of-pocket health care burdens and the extent of underinsurance in the United States.

Recent Design Enhancements

Beginning in 2000, a self-administered questionnaire was added to enhance the value of MEPS for exploring a range of issues relating to access to care, health care quality, health status, and patient satisfaction. Questions include a subset of those developed for the Consumer Assessments of Health Plans Study (CAHPS®), all questions from the SF-12 (Medical Outcomes Study, Short Form) [1], and, to facilitate international comparisons on health status and quality measurement, the questions that comprise the EuroQol 5D (EQ-5D) [4]. The MEPS is further supplemented by "provider accountability" measures for individuals with high prevalence, serious medical conditions, such as diabetes, asthma, and hypertension. For example, a self-administered questionnaire for diabetics obtains yearly information on the frequency of health professional examinations for hemoglobin A-1-C, foot sores or irritations, and eye examinations with pupils dilated.

Data Products

MEPS releases person level, medical event level, condition level, and job level data. Each year, MEPS releases eight specific event files: dental, emergency room, home health, hospital stays, medical visits, outpatient stays, and other medical and prescribed medicines. Each record in a condition file represents a health condition reported by a person in a survey household. Each record in a job file describes a job held by a surveyed person, including wages, benefits, and industry type (e.g. service or manufacturing).

MEPS public use data files on health care utilization, medical expenditures, insurance coverage, and sources of payment are produced annually. Each file includes information from several rounds of data

collection that together comprise a complete calendar year's worth of information. MEPS also releases annual point-in-time ("snapshot") files, for example, the health insurance file that describes coverage for the first part of each calendar year. At this time, annual public use files are typically released within 12 months. For example, most 2001 files were available by the beginning of 2003.

Many MEPS databases contain detailed personal information. To maintain respondent confidentiality while enabling valuable research that could not be accomplished without such data, AHRQ maintains a closely monitored Data Center where researchers can access the protected information needed for approved projects.

Summary

The MEPS has become more comprehensive over time through design and content enhancements, including its greater flexibility to permit sample size enhancements for new initiatives and to facilitate oversampling of policy relevant population subgroups. The survey continues to serve as a national resource for examining the dynamics of recent patterns in health care utilization, expenditures, coverage, and access to care at the national level. The public sector (e.g. Office of Management and Budget (OMB), Congressional Budget Office (CBO), Medicare Payment Advisory Commission (MedPAC), and Treasury Department) uses the MEPS to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs, such as Medicare. Since 2000, data on premium costs from the MEPS Insurance Component have been used by the Bureau of Economic Analysis to produce estimates of the GDP for the nation. The MEPS Insurance Component establishment surveys have been coordinated with the National Compensation Survey conducted by the Bureau of Labor Statistics through participation in the Interdepartmental Work Group on Surveys to minimize overlap in content. Private businesses, foundations, academic institutions, and the health services researchers also use these data for a wide range of purposes.

The MEPS website (`www.meps.ahrq.gov`) provides more detailed information on data availability and research summaries that illustrate the analytical breadth and utility of the MEPS to measure health care trends and inform health policy and practice.

The views expressed in this chapter are those of the author and no official endorsement by the Department of Health and Human Services or the Agency for Health care Research and Quality is intended or should be inferred. The author wishes to thank Dr. Arlene Ash, Dr. Paula Diehr, Ms. Trena Ezzati-Rice and Ms. Elizabeth Conklin for their careful review of the article and for their helpful suggestions.

References

[1] Agency for Healthcare Research and Quality (2004). Your health and health opinions, a self administered supplement to the MEPS. Available at `http://www.meps.ahrq.gov/` Last accessed February 6, 2004.

[2] Berk, M. & Monheit, A. (2001). The concentration of health care expenditures, revisited, *Health Affairs* **20**, 9–18.

[3] Cohen, S.B. (1997). *Sample Design of the 1997 Medical Expenditure Panel Survey Household Component*, AHRQ Pub. 1997–01.

[4] Cohen, S.B. (2003). Design strategies and innovations in the medical expenditure panel survey, *Medical Care* **4**(7), 5–12.

[5] Cohen, J.W. (1997). *Design and Methods of the Medical Expenditure Panel Survey Household Component*, AHRQ Pub. 1997:26.

[6] Institute of Medicine (2002). *Health Insurance is a Family Matter*. The National Academies Press, Washington, DC.

[7] Lichtenberg, F. (2001). Are the Benefits of Newer Drugs Worth their Cost? Evidence from the 1996 MEPS, *Health Affairs* **20**(5), 241–251.

[8] Machlin, S.R. & Taylor, A.K. (2000). *Design, Methods, and Field Results of the 1996 Medical Expenditure Panel Survey Medical Provider Component*, AHRQ Pub. 2000: 28.

[9] Sommers, J. (1999). *List Sample Design of the 1996 Medical Expenditure Panel Survey Insurance Component*, AHRQ Pub. 1999:6.

[10] Zhan, C., Sangl, J., Bierman, A.S., Miller, M.R., Friedman, B., Wickizer, S.W. & Meyer, G.S. (2001). Potentially inappropriate mediation use in the community-dwelling elderly: findings from the 1996 medical expenditure panel survey, *JAMA* **286**(22), 2823–2829.

STEVEN B. COHEN

Medical Journals, Statistical Articles in

There are thousands of publications dealing with aspects of statistics throughout the medical literature. Many of these are "Letters to the Editor", that are critical of some statistical aspect of an already published article; they appear frequently with a response, usually defensive, from the author(s) of the original article. This simple exchange of views is rarely extended in print, and the "response" from authors automatically gives the final word. The adversarial format of these exchanges forestalls scientific resolution of contentious issues and consequently does not provide a satisfactory solution. Less often, but with higher profile, statistical criticism sometimes appears in focused editorials that are invited by editors (sometimes from anonymous authors), and that are published in the same issue of journals as the study (or paper) they comment on. Such editorials can influence the scientific credibility of important studies, because they are published at the same time as the study they focus on and are given some prominence. However, they are usually not peer-reviewed and the statistical criticisms they make may be inaccurate.

Many articles discussing statistical issues appear, as would be expected, in journals devoted to methodologic aspects of epidemiology (notably ***Journal of Clinical Epidemiology***, *American Journal of Epidemiology* and the online BioMed Central (BMC) *medical Research Methodology*), **clinical trials** (***Controlled Clinical Trials*** and, *Clinical Trials*), and diagnosis (*Medical Decision Making*); they have also been published for many years in journals devoted to psychology, especially *Psychological Bulletin, Applied Psychological Measurement*, and *Educational and Psychological Measurement*; indeed, psychologists and psychiatrists have their own sophisticated journals of methodology, *The British Journal of Mathematical and Statistical Psychology*, *Psychometrika*, *Multivariate Behavioural Research* and the *International Journal or methods in psychiatric Research*. The papers in these journals are not discussed further; the remainder of this article will concentrate upon the general medical and medical specialty journals.

Apart from letters, editorials, and papers in the journals mentioned above, there are still many articles each year that deal with statistical issues in an elaborate way; they may be categorized broadly under the following headings:

1. isolated papers on a particular statistical issue
2. series of thematic papers dedicated to a narrow statistical area
3. series of papers covering broad areas of medical statistics
4. guidelines
5. surveys of published papers reporting the frequency of usage of statistical techniques
6. reviews of published papers examining critically aspects of design, analysis, conduct, presentation, and summary
7. systematic reviews (meta-analysis) incorporating assessment of methodologic quality.

Each of these types is described and discussed briefly with some examples. Those chosen are illustrative, if not fully representative, of each category, and have not necessarily been selected as the best or the most comprehensive. Some papers bridge two categories. Some articles on statistics are accompanied by editorials, particularly when they have been commissioned by a journal; some attract criticism in published letters. It is also notable that some of the papers focusing on statistical issues, particularly those in the fifth and sixth categories, do not appear to include a statistician either among the authors or among the acknowledgments.

Isolated Papers on a Particular Statistical Issue

Isolated papers that discuss one particular statistical issue occur sporadically throughout the worldwide medical literature. Some are written to provide examples of the correct application of a statistical technique stimulated by obvious misapplications within a particular specialty of medicine or more widely; some are written to explain the basis of more complicated statistical methods; others are more provocative and written to stimulate debate about controversial issues. The standard of presentation varies widely: some are regarded as "classics", while others, written by authors who do not fully appreciate the complexities of the subject they are writing about, are inaccurate.

Much has been written about all aspects of clinical trials within many specialties of medicine; such papers covering design, conduct, analysis, and

reporting inevitably range across many statistical features. In particular, the small size of many trials inevitably led to the appearance of papers, explaining the need for "proper" **power** calculation during design (*see* **Sample Size Determination**). More controversial issues, particularly analysis under the paradigm of "**intention to treat**", produced, and indeed continues to generate, discussion papers within specialties. Wider application of **Bayesian** analysis and interpretation, realized through massive increases in computer processing power and storage, has resulted in papers explaining why it is needed and how it works [51, 52]. Some papers meet a specific need, for example setting out the classical analysis for both continuous and discrete variables in **crossover trials** [48]; others explain and reexplain concepts such as **regression to the mean** that continue to confuse medical researchers [59, 60, 88].

The need for better understanding of elementary techniques is illustrated: first by Godfrey [38], who sets out the basics of **linear regression** analysis with examples from 36 papers published in *New England Journal of Medicine (NEJM)* over two years; secondly by Hoffman [49], who points out the problems of applying the standard **chi-square test** to paired data; thirdly by Brown [15], who seeks to dispel confusion between **standard deviation** and **standard error**; and finally by Elashoff, commenting on multiple *t*-tests (*see* **Multiple Comparisons**) [26].

Of course, it is not just the application of elementary techniques that requires care. The introduction of more complicated methodology usually found in the pages of statistical journals demands more ready explanation and illustration in the medical journals, where it will need to be interpreted. For example, modeling techniques such as **logistic regression** [33] and the **Cox regression model** [27, 78] were described in the medical literature long before simple exposition in textbooks of medical statistics. Recently techniques for handling "missing" data have been developed in the statistical literature, and are now extending rapidly into specialised medical areas, for example, psychiatry [81], and obesity [35]. Some papers develop links between apparently disparate areas to develop clearer interpretation – for example, Hanley & McNeil [43] – drawing on the association between the area under the **receiver operating characteristic (ROC) curve** and the **Wilcoxon–Mann–Whitney** statistic.

A more controversial issue was the introduction of the randomized consent design [89], a new method for planning clinical trials, published in *NEJM* as a "Special Article" and accompanied by three editorials discussing its merits and demerits (*see* **Ethics of Randomized Trials**). This design, which requires the **randomization** of patients without their prior knowledge and consent, raises difficult ethical and legal issues. A further commentary on this design, summarizing experience of its use, appeared in the same journal five years later [28].

Fourteen papers that appeared originally in *NEJM* were updated and published collectively as a book [9]; an additional six articles were written specifically for the book itself. A second edition with substantive changes followed six years later in 1992 (some chapters were removed, some added, and some updated); they covered fundamental statistical concepts, use of statistical analysis, design, **controls**, series of consecutive patients, classification of research reports, decision analysis, linear **regression**, comparing multiple means, ordered categories (*see* **Ordered Categorical Data**), reporting methods in clinical trials, power and sample size, and **meta-analysis**; another article on statistical reporting in medical journals was drawn from *Annals of Internal Medicine*. Compendia of this type constitute a valuable resource; to our knowledge no others have been published.

Series of Thematic Papers Dedicated to a Narrow Statistical Area

There are many series of published papers that consist of up to three or four articles, occasionally more, focused either on the application of a particular statistical technique in medicine or on one particular area. One example of the former is the series of eight papers in *British Medical Journal (BMJ)* on systematic reviews (overviews or meta-analysis) produced in 1994 as a consequence of the huge increase in the number of such studies and the need to explain their rationale and methodology in greater detail; the series was later published in a book with an extensive bibliography [17] that expanded with the second edition published six years [25]. Another example was the ambitious trio of papers published by the *Lancet* introducing and explaining the ideas and concepts of **neural network** techniques (first paper [21]). Occasionally whole issues of journals have been devoted

to statistical issues, for example meta-analysis [65] and design and analysis of studies of gingivitis and periodontitis [77].

Several examples are needed to illustrate the diversity of applications to a particular area. First, the move away from ***P* values** and **hypothesis testing** towards **estimation** in the mid-1980s would only be realized if data analysts could readily calculate **confidence intervals** for a range of summary statistics. Since the literature on this was widely scattered and remote from the medical literature, *BMJ* responded by publishing a series of four papers in 1988 that both explained the reasons for interval estimation (*see* **Estimation, Interval**) and provided the methodology. The series was later collected in a book [36], that included both an earlier motivating paper and statistical guidelines [5, 37]; a second edition with further additions appeared eleven years later [6]. A second series of three papers set out the basic principles of good form design (*see* **Questionnaire Design**). Although not necessarily regarded as a subcomponent of "statistics", this is an area that is frequently discussed with statisticians to control both the extent and quality of data to be collected [85]. As a third example there are the two seminal papers by Peto et al. [62] that had a great influence, not just in cancer but far beyond, on the analysis and reporting of pragmatic clinical trials that followed patients to a specified event such as death. These two papers discussed in great detail not just the methods of analysis but also how to handle many of the problems encountered in such trials. The final examples illustrate medical journals responding to topical issues with a need for better understanding of contemporary techniques in assessing **quality of life** (three papers) [32], in estimating and interpreting costs (*see* **Health Economics**) (six papers) (first paper [69]); in better management (15 papers including **decision theory**) (first paper [74]), and in "qualitative" research (seven papers) (first paper [64]); or discussing perennial favorites like placebos (*see* **Blinding or Masking**) (seven papers) (first paper [40]), and epidemiology and clinical trials (eleven papers) (first paper [41]).

Of course, series of papers on a specific statistical theme are not restricted to general medical or cancer journals. In particular, in **psychiatry** the importance of statistics has been debated and acknowledged for many years. For example, May et al. [56], discuss the assessment of psychiatric outcome in both **cross-sectional** and follow-up studies, while Streiner spans a wide range of research methods in a series of 23 articles that started in 1990, and continues today (first paper [75]).

Series of Papers Covering Broad Areas of Medical Statistics

There have been several series of papers covering broad areas of medical statistics. The earliest and best known is, of course, the one by **Bradford Hill** that was published weekly in the *Lancet* from 2 January to 24 April 1937. The first article was prefaced by an editorial entitled "Mathematics and Medicine", that opened with the words

> Statistics are curious things. They afford one of the few examples in which the use, or abuse, of mathematical methods tends to induce a strong emotional reaction in non-mathematical minds. This is because statisticians apply, to problems in which we are interested, a technique which we do not understand. It is exasperating, when we have studied a problem by methods that we have spent laborious years in mastering, to find our conclusions questioned, and perhaps refuted, by someone who could not have made the observations himself. It requires more equanimity than most of us possess to acknowledge that the fault is in ourselves.

This series of 17 3–4-page articles which covered the aims of statistics, selection, presentation, variation, averages, proportions, differences, **chi-squared**, **correlation**, **life tables**, and **survival**, common fallacies and difficulties, proportional rates, crude rates, and the calculation of standard deviation and correlation coefficient, was quickly published collectively as the celebrated book, *The Principles of Medical Statistics* [46] in the same year, and in 11 subsequent revised and expanded editions spanning 54 years [47].

One of the longest series was the extensive collection written by Feinstein under the heading *Clinical Biostatistics* and published in *Clinical Pharmacology and Therapeutics* from 1970 onwards. (He also wrote three other much shorter series for *Archives of Internal Medicine, Annals of Internal Medicine*, and *Yale Journal of Biology and Medicine*.) Twenty-nine articles selected from the first 40 were published collectively as a book with the same title [31]. This extensive series of over 50 articles built on and was

inspired by an earlier series in the same journal by **Mainland** and an unpublished collection of 145 "Notes from a laboratory of medical statistics", also by Mainland. These long articles discussed in detail not just the features of design, analysis, presentation, and interpretation, but also included quantitative surveys of the medical and statistical literature, critiques of individual studies, and discussion of the **teaching of statistics** and the ethics of research.

By contrast to the series above, that taught much about concept and methodology, a later series of short appealing articles published in *BMJ* in 1976 by Swinscow reflected the need of both the medical profession and journals at the time by presenting a very practical approach to the application of statistics through simple calculation of summary statistics (**means**, standard deviations, standard errors, proportions, correlation, and regression) and immediate application of significance tests (t-test (*see* **Student's *t* Statistics**), chi-square, **Fisher's exact**, rank sum). This series was also published as a book [76], the immense popularity of which can be gauged from the ten editions, and many reprintings, published over 26 years. This success was repeated several years later with a series on the basics of epidemiology [70].

As a follow-up to the papers by Swinscow and to remind researchers and doctors that there is more to medical statistics than the calculation of summary measures and hypothesis tests, *BMJ* commissioned two further series that were published contiguously in 1982. The specific aims were to remind researchers and authors that "they need statistical advice before starting a project and not at its end", to make medical statisticians "appreciate that most doctors are still bewildered by statistical jargon and too often react by ignoring the more important aspects of logic and correctness of argument", and finally to educate editors of journals (and their advisers) "to be on the lookout for pitfalls and use expert statistical advisers more frequently than they do". The two series were *Statistics and Ethics in Medical Research*, that covered misuse, design, sample size, data collection, analysis, presentation, interpretation, and how to improve the quality of statistics in medical journals, and *Statistics in Question*, which in two parts covered many aspects of clinical trials (13 articles), and the principles of data display, presentation, summary, and interpretation (10 articles). Following the precedent established with the earlier series, these two series were also published as a book [39].

Starting in 1994, and perhaps harking back to the seminal *Notes* of Mainland, *BMJ* has introduced a longer series of very brief occasional *Statistics Notes*, not just to remind readers (yet again!) about basics, but also, and importantly, "to keep them up to date with more complex techniques that are finding their way into medical studies" [13]. Each note occupies at most one page, dwells on a single topic, and is self-contained. By 2003, the series extended to 47 *Notes* and included some topics that are needed frequently in practice but are not easy to locate – for example, regression to the mean, quantiles (*see* **Quantiles**), multiple significance tests, correlation with repeat observations on the same subject (*see* **Longitudinal Data Analysis, Overview**), data summary after **transformation, measurement error**, and **Cronbach's alpha**.

Another long series (17 papers under the heading *Statistics from the Inside*) by Healy (first paper [44]) appeared in *Archives of Diseases in Childhood* over the period from 1991 to 1995; as well as the usual basic statistical concepts, it also covered diagnostic and **screening** tests as well as reference values. Somewhat shorter series are more common. An example is a series of four basic articles on general statistical principles under the heading *Basic Statistics for Clinicians* in *Canadian Medical Association Journal* [42].

Guidelines

Guidelines are intended to present a succinct summary (sometimes highly detailed) of procedures or standards that should be followed in performing set tasks; several have been written for performing or assessing research. A few examples are guidelines for performing clinical trials in particular specialties of medicine [10, 66], for evaluating the quality of clinical trials generally [18], for reporting epidemiologic studies [14, 31] and studies of screening tests [79], for structuring reports [54], and for evaluating them [67, 87]. The Evidenced Based Medicine Working Group in Canada has produced several guides for assessing and interpreting reviews [61]. There are many others. There are also more general statistical guidelines [5] that some journals draw attention to in their advice to authors. Other guidelines concern licensing applications to Regulatory Authorities [20]. Abbreviated guidelines in the form of checklists have also been produced,

for example for designing clinical trials [19] and for statistical review [37]. Among more recent guidelines is the CONSORT statement [11], that was intended to improve and standardize the presentation of results from randomized clinical trials. Unusually, the authors of these guidelines included some journal editors, which may explain why the CONSORT statement was adopted by many leading journals; with a consequent improvement in publication standards [23]. Further improvement is expected following publication of the revised CONSORT statement [58], accompanied by a detailed explanation of its use [7], and the recent extension to cluster randomised trials [16]. Similar initiatives have been instigated to improve reporting in other areas of medical research.

Surveys Reporting the Frequency of Usage of Statistical Techniques

There have been several surveys of the use of statistical techniques in the biomedical literature. Although these "content analyses" are sometimes combined with the assessment of correct usage, as discussed in the next category, their aim is generally to review the knowledge required to understand published papers rather than whether or not the reported techniques are used correctly. For example, Hokanson et al. [50], adopting categories established by previous investigators [29], estimated the frequencies with which various statistical techniques were reported in almost 5000 papers published in five major American **oncology** journals in 1983 and 1984; they concluded that readers familiar with about a dozen techniques could expect to understand approximately 90% of quantitative concepts in those journals; not surprisingly the techniques were mostly those covered in many elementary texts about medical statistics. These results were supported by those of Marsh & Hawkins [55], who used similar techniques to survey 44 publications from multicenter randomized clinical trials sponsored by the National Eye Institute or the National Heart, Lung, and Blood Institute. They found that knowledge of 12 techniques would be sufficient to understand 90% of the published analyses. However, they also showed that no publication from the set of 44 was fully accessible with knowledge of just the six most frequently used statistical techniques (descriptive statistics, **contingency tables**, *t*-test, power, life tables, and regression for survival).

However, there have been major changes in more recent years. Later surveys of *NEJM* [2, 30] found great increases in the use of more complex methods (notably logistic regression and survival analysis) and also an increase in the average number of techniques used in each paper. They provide a rare example of longitudinal content analysis within one journal. The contrast between two time periods is also discussed in [4].

Critical Reviews of Published Papers

Since 1920 there have been several hundred reviews of the published medical literature that have examined in some detail various aspects of design, data collection, analysis, presentation, and summary. Their objective is to report the frequency of statistical misuse or bad practice. Many have focused on clinical trials and/or epidemiologic studies within specific subspecialties of medicine, and others on specific statistical techniques; some were restricted to specific journals, while others were very broadly based. The earliest we know of was published in 1929 [24], and consists of a survey of the extent to which "statistical logic" is used in a sample of 200 medical–physiologic papers from then current American periodicals. This is a long paper and, at over 120 pages, is longer than many of the series covered in the third category above; the second page reports the results of the survey and the remainder essentially form a textbook of medical statistics for physiologists. Although Bradford Hill reviewed papers published in the *Lancet* when preparing his famous series in 1937, the next review we know of appeared in 1951 [71], and looked at the use of **controls** in papers that appeared in five leading American periodicals during the first half of 1950. From then the increase has been seemingly exponential. Indeed, reviews of this type have become an industry of their own, with over 120 journals having had their content subjected to statistical scrutiny. The surveys range from the simple bald summary that "we conducted a review of the literature" and found that "not more than a dozen adequately designed long-term follow-up studies are available" to far more elaborate studies examining in excess of a thousand publications and reporting the findings in detail.

One of the first large studies, and certainly one of the most influential, was the two-part review by

Schor & Karten [72]. They chose ten from among the 67 most frequently read medical journals, randomly selected three issues of each from the first three months of 1964, and then reviewed each article to investigate whether or not the conclusions that were drawn "were valid in terms of the design of the experiment, the type of analysis performed, and the applicability of the statistical tests used or not used". Their conclusions suggested that "none of the ten journals had more than 40% of its analytical studies considered acceptable; two of the ten had no acceptable reports". The second part of this review went a stage further when, later in 1964, one of the ten journals instituted statistical review of submitted papers that were judged medically acceptable for publication. In the next 18 months 514 manuscripts were submitted to statistical evaluation, of which 133 (26%) were considered acceptable and 34 (7%) were so poor as to be considered unsalvageable. Amongst other recommendations, the authors suggested that a statistician "either be part of the research team or be consulted before a study is attempted", a cry that has been repeated many times over the subsequent 30 years (*see* **Statistical Review for Medical Journals, Guidelines for Authors**; **Statistical Review for Medical Journals, Journal's Perspective**).

One of the best known reviews is that of Freiman et al. [34], who looked at the results from 71 randomized clinical trials that were "negative" in the sense that the comparison of control and experimental therapies was not statistically significant at the 5% level ($P > 0.05$); the trials were reported in 20 different journals over the period from 1960 to 1977. The authors observed that 67 (94%) of the trials had a greater than 10% risk of missing a true 25% therapeutic improvement and that 50 (70%) had a similar risk of missing a 50% improvement. Many other surveys have looked at statistical power, with similar findings, a recent one actually appearing in a new journal publishing negative results [45].

A much broader review is that of McGuigan [57], who looked for statistical errors in all papers (164 in total) reporting numerical results published in the *British Journal of Psychiatry* during 1993. Using the methods established in an earlier survey of the same journal by White [82] (covering the period 1977–1978), McGuigan reported an overall error rate of 40% (compared with White's 45%); individual types of error (rate) were characterized as: description of randomization or control selection (43%); measures of location (27%); measures of dispersion (27%); Student's t-test (80%); chi-square test (15%); **null hypothesis** description (5%); description of methods (16%); description of statistic (1%); statement of results (17%); interpretation of P values (2%); and incorrect or inadequate analysis (27%). McGuigan also plotted the rates from 14 surveys of statistical errors in the medical literature and published between 1960 and 1993; the median was in excess of 50% (although the definition of statistical error was not the same in each study (cf. [1]). Such surveys now extend more widely, for example, to Chinese [80] and Czech [63] biomedical journals.

Three further examples will suffice to demonstrate the character and range of these surveys. Badgley [8] surveyed all articles published in two Canadian journals in the first half of 1960, looking for those (103) that used "group data" (i.e. epidemiologic surveys and clinical trials and excluding case reports (*see* **Case Series, Case Reports**), reviews, descriptive papers, and articles providing a survey of the literature on a given topic); he reviewed five aspects of each article: the definition of terms; the selection of a population or sample; the use of controls; types of statistical analysis; and the derivation of conclusions. His summary reported that the "assessment revealed the need for greater precision in the design of many studies using group data and for caution in the interpretation of results". Ried & Hall (in a letter [68]) reported a survey of 569 papers published between 1980 and 1983 in the *American Journal of Clinical Nutrition*, looking for multiple significance testing within a single dataset. They found a median of 21 tests per paper (quartiles: 6 and 48) and commented "13% of the papers failed to state the type of statistical test that was used, only four papers made an allowance for the multiple use of statistical tests and 427 of the papers (84%) failed to specify predetermined levels of statistical significance". Linnet [53] reviewed assessments of diagnostic tests published in *Clinical Chemistry* during the period 1979–1983, looking at sample sizes and the statistical confidence of **sensitivity** estimates. He found 84 relevant papers and concluded from his review that "the precision of sensitivity estimates of tests was seldom considered by the investigators, and the significance of differences of sensitivity estimates was, with a single exception, not tested". Others were summarized by Altman [1, 2].

Interpretation of the results from the reviews themselves is not without difficulties. Williamson et al. [84] reported in 1986 an analysis of 33 surveys of the quality of the medical literature. Three experienced medical statisticians independently evaluated the quality of these review articles using a checklist of 40 items indicating the extent to which the assessment methods reported in an article substantiated the authors' results. While none was entirely unsubstantiated or contradicted, 15% were only weakly substantiated.

Systematic Reviews (Meta-analysis)

The last 15 years have seen a huge increase in the use of systematic reviews (meta-analysis) in medicine, especially of therapeutic trials [25]. This has led to much greater scrutiny of the statistical techniques employed in the design, analysis, and reporting of clinical trials, and to empirical studies of the relation between study features and the results of trials [73]. Meta-analyses quite often report the credibility of individual trials using assessment schedules that focus on key features. Such information can be used to determine which studies are included in a meta-analysis, or may be used as part of a **sensitivity analysis**. It is also possible to generate a quality score [18] that can be used to weight the results of individual trials before combination in the overview [22], although this is not a generally accepted approach. Systematic reviews are now more common in various nonexperimental situations, for example, epidemiologic studies and diagnostic tests, and concerns about the quality of primary studies has already resulted in quality assessment of the Catter [83].

Discussion

It is unfortunate that to date no annotated bibliography or catalog of all the articles in the first six categories listed above has been constructed. This would be an extremely valuable scientific resource both for research and for teaching and, further, would prevent unnecessary duplication among journals. There are no useful combinations of keywords that can be used in a search of electronic databases to identify the great majority of articles within the categories above, except perhaps for guidelines and meta-analyses. Some papers have titles that give no clue to the inclusion of a review of methodology [38, 53]. As we have noted, there is a considerable body of statistical articles within the medical literature; papers on statistics can appear in any medical journal. In particular, all the leading general medical journals have published important papers on aspects of statistics reflecting their appreciation of the importance of the sound application of the statistical components of any research article. The BMJ has been foremost in the drive to achieve higher standards, first by publishing far more such papers than any other journal, second by commissioning both individual papers and thematic series that can be readily understood and applied by physicians, and third, by republishing some articles collectively in compact, low-priced books.

Didactic articles have a long history and seem to be especially valued by journals and readers. Such articles may describe standard methodology or may introduce new methods in a doctor-friendly manner. Expository statistical papers (e.g. [12, 62] – category 1 above) can reach 500 citations within 4–5 years [4].

Perhaps second in impact are those studies of the quality of the statistical aspects of published research. Over a period of more than 70 years such reviews have consistently shown that many published papers are flawed. In some cases such reviews have led directly to changes in editorial policy, especially regarding increased statistical review of manuscripts. The main impact of such studies is probably slowly cumulative over time, and permeates gradually across the literature from the general journals to the specialist journals.

Statistical errors can occur at every stage of a research project, although reviews of the literature have been mainly concerned with methods of analysis. The underlying reason for the plethora of statistical errors is that the majority of statistical analyses are performed by people with an inadequate understanding of statistical methods [86]. They are then peer-reviewed by people who are generally no more knowledgeable [3]. There are other contributory reasons, such as the fact that several introductory textbooks in statistics are themselves full of errors [3]. Another problem is the copying of incorrect methodology from one study to another – for example, use of the correlation coefficient for comparing two methods of measurement [12].

Altman & Goodman [4] studied citations of 18 important statistical publications and also some content analyses to evaluate the speed with which new statistical methods infiltrate medical journals. They suggested several possible reasons for the apparent increased speed of diffusion: the increasing number of statisticians working in medicine, the wide accessibility of powerful desktop computers, and the more rapid development and dissemination of software to implement new statistical methods (*see* **Software, Biostatistical**). It is likely that general understanding of basic statistical methods (*t* and chi-square tests, for example) has improved, but there is ample evidence that many errors still occur in the use of these simple methods [57]. The increased use of more complex methods [2, 30], aided by easy access to powerful computers, has led to new problems, many of which cannot be detected in published papers. Several more complex statistical methods introduced in the 1980s are beginning to be seen more frequently – examples include neural networks, **multilevel models**, and Gibbs sampling (*see* **Markov Chain Monte Carlo**) [4]. Journals should expect to see growing numbers of papers using them, and doubtless a cluster of new didactic articles describing them. Nevertheless, the speed with which new methods are introduced may pose problems for statistical referees, for the physicians who read the published work, and for the journals themselves.

References

[1] Altman, D.G. (1982). Statistics in medical journals, *Statistics in Medicine* **1**, 59–71.

[2] Altman, D.G. (1991). Statistics in medical journals: developments in the 1980s, *Statistics in Medicine* **10**, 1897–1913.

[3] Altman, D.G. & Bland, J.M. (1991). Improving doctors' understanding of statistics (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 223–267.

[4] Altman, D.G. & Goodman, S.N. (1994). Transfer of technology from statistical journals to the biomedical literature: past trends and future predictions, *Journal of the American Medical Association* **272**, 129–132.

[5] Altman, D.G., Gore, S.M., Gardner, M.J. & Pocock, S.J. (1983). Statistical guidelines for contributors to medical journals, *British Medical Journal* **i**, 1489–1493.

[6] Ahman, D.G., Machin, D., Bryart, T.N. & Gardner, M.J. (2000). *Statistical with Confidence: Confidence Interior and Statistical Guidelines*, 2nd Ed. British Medical Journal, London.

[7] Ahman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzche, P.C. & Lang, T., for the CONSORT samp. (2001). The revised CONSORT statement for reprinting randomized trials: explanation and elaboration, *Annals of internal medicine* **134**, 663–694.

[8] Badgley, R.F. (1961). An assessment of research methods reported in 103 scientific articles from two Canadian medical journals, *Canadian Medical Association Journal* **85**, 246–250.

[9] Bailar, J.C. & Mosteller, F. (1986). *Medical Uses of Statistics*, 1st Ed. NEJM Books, Boston.

[10] Beam, T.R., Gilbert, D.N. & Kunin, C.M. eds. (1992). Guidelines for the evaluation of anti-infective drug products, *Clinical Infectious Diseases* **15**, Supplement 1.

[11] Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. & Stroup, D.F. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement, *Journal of the American Medical Association* **276**, 637–638.

[12] Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **i**, 307–310.

[13] Bland, J.M. & Altman, D.G. (1994). Statistics Notes: Correlation, regression, and repeated data, *British Medical Journal* **308**, 596.

[14] Bracken, M.B. (1989). Reporting observational studies, *British Journal of Obstetrics and Gynaecology* **96**, 383–388.

[15] Brown, G.W. (1982). Standard deviation, standard error: which "standard" should we use? *American Journal of Diseases of Childhood* **136**, 937–941.

[16] Campbell, M.K., Elbourne, D.R. & Ahman, D.G., for the CONSORT Group. (2004). CONSORT statement: exterior to charter randomized trials. *British Medical Journal* **328**, 702–708.

[17] Chalmers, I. & Altman, D.G. eds. (1995). *Systematic Reviews*. BMJ Publishing Group, London.

[18] Chalmers, T.C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial, *Controlled Clinical Trials* **2**, 31–49.

[19] Chaput de Saintonge, D.M. (1977). Aide-mémoire for preparing clinical trial protocols, *British Medical Journal* **i**, 1323–1324.

[20] CPMP Working Party on Efficacy of Medicinal Products (1995). Statistical methodology in clinical trials in applications for marketing authorizations for medicinal products, Note for Guidance iii/3630/92-EN, *Statistics in Medicine* **14**, 1659–1682.

[21] Cross, S.S., Harrison, R.F. & Kennedy, R.L. (1995). Introduction to neural networks, *Lancet* **346**, 1075–1079.

[22] Detsky, A.S., Naylor, C.D., O'Rourke, K., McGeer, A.J. & L'Abbé, K.A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis, *Journal of Clinical Epidemiology* **45**, 255–265.

[23] Deveraux, P.J., Manns, B.J., Ghali, W.A., Quan, H., & Guyatt, G.H. (2002). The reporting of methodological factor in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated standards of Reporting Trials (CONSORT) checklist, *Controlled Clinical Trials* **23**, 380–388.

[24] Dunn, H.L. (1929). Application of statistical methods in physiology, *Physiological Review* **9**, 275–398.

[25] Egger, M., Davey Smith, G. & Ahman, D.G. (2001). *Systematic Reviews in Health care: meta-analysis in context*. Second edition. BMJ publishing Group, London.

[26] Elashoff, J.D. (1981). Down with multiple *t*-tests, *Gastroenterology* **80**, 615–620.

[27] Elashoff, J.D. (1983). Surviving proportional hazards, *Hepatology* **3**, 1031–1035.

[28] Ellenberg, S.S. (1984). Special Report: Randomization designs in comparative clinical trials, *New England Journal of Medicine* **310**, 1404–1408.

[29] Emerson, J.D. & Colditz, G.A. (1983). Use of statistical analysis in the *New England Journal of Medicine*, *New England Journal of Medicine* **309**, 709–714.

[30] Emerson, J.D. & Colditz, G. (1992). e of statistical analysis in the *New England Journal of Medicine, Medical Uses of Statistics*, 2nd Ed. J.C. Bailar, III & F. Mosteller, eds. NEJM Books, Boston, pp. 45–57.

[31] Feinstein, A. (1977). *Clinical Biostatistics*. C.V. Mosby Company, St Louis.

[32] Fitzpatrick, R., Fletcher, A.F., Gore, S., Jones, D., Spiegelhalter, D. & Cox, D. (1992). Quality of life measures in health care, *British Medical Journal* **305**, 1074–1077, 1145–1148, 1205–1209.

[33] Fleiss, J.L., Williams, J.B.W. & Dubro, A.F. (1986). The logistic regression analysis of psychiatric data, *Journal of Psychiatric Research* **20**, 195–209.

[34] Freiman, J.A., Chalmers, T.C., Smith, H. & Kuebler, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial, *New England Journal of Medicine* **299**, 690–694.

[35] Gadbury, G.L., Coffey, C.S. & Allison, D.B. (2003). Modern statistics methods for handling missing repeated measurements in obesity trial data: beyond LOCF. *Obesity Reviews* **4**, 175–184.

[36] Gardner, M.J. & Altman, D.G. (1989). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. British Medical Journal, London.

[37] Gardner, M.J., Machin, D. & Campbell, M.J. (1986). Use of checklists in assessing the statistical content of medical studies, *British Medical Journal* **i**, 810–812.

[38] Godfrey, K. (1985). Simple linear regression in medical research, *New England Journal of Medicine* **313**, 1629–1636.

[39] Gore, S.M. & Altman, D.G. (1982). *Statistics in Practice*. British Medical Association, London.

[40] Gøtzsche, P.C. (1994). Is there logic in the placebo? *Lancet* **344**, 925–926.

[41] Grimes, D.A. & Schulz, K.F. (2002). An review of clinical research: the lay of the land. *Lancet* **359**, 57–61.

[42] Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H. & Walter, S. (1995). Basic statistics for clinicians. 1. Hypothesis-testing; 2. Interpreting study results – confidence intervals; 3. Assessing the effects of treatment – measures of association; 4. Correlation and regression, *Canadian Medical Association Journal* **152**, 27–32, 169–173, 351–357, 497–504.

[43] Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143**, 29–36.

[44] Healy, M.J.R. (1991). Statistics from the Inside: Populations and samples, *Archives of Diseases in Childhood* **66**, 1355–1361.

[45] Hebert, R.S., Wright, S.M., Dittus, R.S. & Elasy, T.A. (2002). Prominent medical journals often provide insufficient information to assess the validity of studies with negative results. *BMC Journal of Negative Results in Biomechanic* **1**, 1.

[46] Hill, A.B. (1937). *Principles of Medical Statistics*. Lancet, London.

[47] Hill, A.B. & Hill, I.D. (1991). *Bradford Hill's Principles of Medical Statistics*, 12th Ed. Edward Arnold, London.

[48] Hills, M. & Armitage, P. (1979). The two-period crossover clinical trial, *British Journal of Clinical Pharmacology* **8**, 7–20.

[49] Hoffman, J.I.E. (1976). The incorrect use of chi-square analysis for paired data, *Clinical and Experimental Immunology* **24**, 227–229.

[50] Hokanson, J.A., Luftman, D.J. & Weiss, G.B. (1986). Frequency and diversity of use of statistical techniques in oncology journals, *Cancer Treatment Reports* **70**, 589–594.

[51] Lilford, R.J. & Braunholtz, D. (1996). The statistical basis of public policy: a paradigm shift is overdue, *British Medical Journal* **313**, 603–607.

[52] Lilford, R.J., Thornton, J.G. & Braunholtz, D. (1995). Clinical trials and rare diseases: a way out of a conundrum, *British Medical Journal* **311**, 1621–1625.

[53] Linnet, K. (1985). Precision of sensitivity estimations in diagnostic test evaluations. Power functions for comparisons of sensitivities of two tests, *Clinical Chemistry* **31**, 574–580.

[54] Makuch, R.W. (1982). Statistical guidelines for medical research reports, *Cancer Treatment Reports* **66**, 217–219.

[55] Marsh, M.J. & Hawkins, B.S. (1994). Publications from multicentre clinical trials: statistical techniques and accessibility to the reader, *Statistics in Medicine* **13**, 2393–2406.

[56] May, P.R.A., Yale, C. & Dixon, W.J. (1972; 1973). Assessment of psychiatric outcome: 1. Cross-section analysis; II. Simple Simon analysis; III. Process analysis, *Journal of Psychiatric Research* **9**, 271–284, 285–292; **10**, 31–42.

[57] McGuigan, S.M. (1995). The use of statistics in the *British Journal of Psychiatry*, *British Journal of Psychiatry* **167**, 683–688.

[58] Moher, D., Schulz, K.F. & Ahman, D.G., for the CONSORT Group. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Annals of Internal Medicine* **134**, 657–662.

[59] Newell, D.J. (1991). A seven-point plan to explain regression to the mean, *Australian Journal of Public Health* **15**, 151.

[60] Newell, D. & Simpson, J. (1990). Regression to the mean, *Medical Journal of Australia* **153**, 166–168.

[61] Oxman, A.D., Davis, D.A., Feightner, J.W., Finnie, N.V., Hutchison, B.G., Lusk, S., Macdonald, P.J., Mcauley, R.G. & Sellors, J.W. (1994). Evidence-based care: 1. Setting priorities – how important is this problem; 2. Setting guidelines – how should we manage this problem; 3. Measuring performance – how are we managing this problem; 4. Improving performance – how can we improve the way we manage this problem; 5. Lifelong learning – how can we learn to be more effective, *Canadian Medical Association Journal* **150**, 1249–1254, 1417–1423, 1575–1579, 1793–1796, 1971–1973.

[62] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1976; 1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and Design; II. Analysis and examples, *British Journal of Cancer* **34**, 585–612; **35**, 1–39.

[63] Pilčík, T. (2003). Statistics in three biomedical journals. *Physiological Research* **52**, 39–43.

[64] Pope, C. & Mays, N. (1995). Reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research, *British Medical Journal* **311**, 42–45.

[65] Potsdam International Consultation on Meta-Analysis, (1995). *Journal of Clinical Epidemiology* **48**, 1–172.

[66] Raskob, G.E., Lofthouse, R.N. & Hull, R.D. (1985). Methodological guidelines for clinical trials evaluating new therapeutic approaches in bone and joint surgery, *Journal of Bone and Joint Surgery* **67**, 1294–1297.

[67] Riddell, B.C., Walter, D.E. & Wells, J.M. (1979). An approach to evaluating reports of research studies, *Canadian Journal of Hospital Pharmacy* **32**, 69–70.

[68] Ried, M. & Hall, J.C. (1984). Multiple statistical comparisons in nutritional research, *American Journal of Clinical Nutrition* **40**, 183–184.

[69] Robinson, R. (1993). Economic evaluation and health care: what does it mean? *British Medical Journal* **307**, 670–673.

[70] Rose, G. & Barker, D.J.P. (1979). *Epidemiology for the Uninitiated*. British Medical Association, London.

[71] Ross, O.B. (1951). Use of controls in medical research, *Journal of the American Medical Association* **145**, 72–75.

[72] Schor, S. & Karten, I. (1965). Statistical evaluation of medical journal manuscripts, *Journal of the American Medical Association* **195**, 145–150.

[73] Schulz, K.F., Chalmers, I., Hayes, R.J. & Altman, D.G. (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association* **273**, 408–412.

[74] Simpson, J. (1994). Management for Doctors: Doctors and management – why bother?, *British Medical Journal* **309**, 1505–1508.

[75] Streiner, D.L. (1990). Sample size and power in psychiatric research. *Canadian Journal of Psychiatry* **35**, 616–620.

[76] Swinscow, T.D.V. (1976). *Statistics at Square One*. British Medical Association, London.

[77] Task-Force-on-Design-and-Analysis-Inc and American Dental Association Conference on equivalency and superiority claims for products for gingivitis and periodontitis (1992). *Journal of Periodontal Research* **27**(4), Part 2.

[78] Tibshirani, R. (1982). A plain man's guide to the proportional hazards model, *Clinical and Investigative Medicine* **5**, 63–68.

[79] Wald, N. & Cuckle, H. (1989). Reporting the assessment of screening and diagnostic tests, *British Journal of Obstetrics and Gynaecology* **96**, 389–396.

[80] Wang Q, & Zhang, B. (1998). Research design and statistical methods in chinese medical journals. *Journal of the American Medical Association* **280**, 283–285.

[81] White, I.R., Moodie, E., Thompson, S.G. & Croudace, T. (2003). A modelling strategy for the analysis of clinical trials with partly missing longitudinal data. *International Journal of Methods in Psychiatric Research* **12**, 139–150.

[82] White, S.J. (1979). Statistical errors in papers in the *British Journal of Psychiatry*, *British Journal of Psychiatry* **135**, 336–342.

[83] Whiting, P., Rutjes, A.W., Reitsma, J.B., Bossuyt, P.M. & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* **3**(1), 25.

[84] Williamson, J.W., Goldsmith, P.G. & Colton, T. (1986). The quality of medical literature: an analysis of validation assessments, *Medical Uses of Statistics*, J.C. Bailar & F. Mosteller eds. NEJM Books, Boston, Chapter 19, pp. 370–391.

[85] Wright, P. & Haybittle, J. (1979). Design of forms for clinical trials, *British Medical Journal* **ii**, 529–530, 590–592, 650–651.

[86] Wulff, H.R., Andersen, B., Brandenhoff, P. & Guttler, F. (1987). What do doctors know about statistics?, *Statistics in Medicine* **6**, 3–10.

[87] Yancey, J.M. (1990). Ten rules for reading clinical research reports, *American Journal of Surgery* **159**, 533–539.

[88] Yudkin, P.L. & Stratton, I.M. (1996). How to deal with regression to the mean in intervention studies, *Lancet* **347**, 241–243.

[89] Zelen, M. (1979). A new design for randomized clinical trials, *New England Journal of Medicine* **300**, 1242–1245.

Further Reading

Epidemiology Work Group of the Interagency Regulatory Liaison Group (1981). Guidelines for documentation of epidemiologic studies, *American Journal of Epidemiology* **114**, 609–618.

ANTHONY L. JOHNSON &
DOUGLAS G. ALTMAN

Medical Research Council (MRC)

The Medical Research Committee was set up in the UK in 1913 by the Government. The name Medical Research Council (MRC) was established by Royal Charter in 1920. At the outset, tuberculosis was a particular research target, but the scope of the MRC soon broadened to most areas of medical research.

The central institute of the Medical Research Committee was set up in Hampstead (London) in 1914, becoming the National Institute for Medical Research (NIMR) in 1920. The founding departments were bacteriology, applied physiology, biochemistry and pharmacology, and medical statistics. The inclusion of a department of medical statistics demonstrates the MRC's early recognition of the importance of statistics in medical research.

In fact, the Medical Research Committee had a Department of Medical Statistics since 1914, under the leadership of **John Brownlee**. This Department carried out original statistical research, but also undertook much routine work such as the sorting and classification of medical records. This latter activity was especially heavy during the war years, at which time more than 100 clerical staff were employed. In 1920 **Major Greenwood**, who was Statistical Medical Officer on the staff of the Ministry of Health, moved to NIMR. Greenwood became chair of the Industrial Health Statistics Committee and also a new Nutrition Committee. The latter, rather than Brownlee's department, became an advisory committee on statistical matters for the whole of the Council's work, and in 1925 its name was changed to the Statistical Committee.

During the 1920s it became clear that the MRC needed to reconsider the organization of its statistical work. In 1927 Greenwood was appointed to direct the Department of Epidemiology and Vital Statistics at the London School of Hygiene and Tropical Medicine (LSHTM). Brownlee's sudden death allowed the MRC to transfer all its statistical activity to the LSHTM. From 1928 the staff of the NIMR Department, Greenwood's Ministry of Health Department, and the Statistical Committee were merged into a single unit at the LSHTM.

In 1931 the MRC set up a Therapeutic Trials Committee (TTC) to oversee **clinical trials** in many areas, especially of potential new remedies. Surprisingly there were no statisticians on this committee, although Greenwood and **Bradford Hill** could be called upon to give statistical input where necessary. Although these early trials were controlled, it was not until 1946 that **randomization** was used. The concept of randomization had been introduced in agricultural research by **R.A. Fisher**. Despite his involvement with other MRC committees, such as the Human Genetics Committee, it seems that Fisher did not have any involvement with the TTC, but he did direct an MRC program of research into the genetic study of **blood groups**.

Greenwood retired in 1945 and was succeeded by Bradford Hill, under whose leadership the MRC group became the Statistical Research Unit. Under Hill the Unit developed an increasing reputation, notably becoming heavily involved in the MRC's clinical trials program. The introduction of randomized controlled trials was clearly a major event; the most famous of these was the **Medical Research Council Streptomycin Trial**, which was the first randomized trial to be published. Hill also made crucial contributions to epidemiologic research. In particular, in 1947 he embarked with Richard Doll on a series of famous studies of smoking (*see* **Smoking and Health**). This work led not only to clear evidence of a causal link with lung cancer, but it also stimulated Hill's development of the underlying principles of **observational studies** and the criteria for establishing causality from such studies (*see* **Hill's Criteria for Causality**).

In 1960 Bradford Hill was succeeded by Richard Doll, who directed the Unit until 1969. Doll was appointed deputy director of the MRC's new Clinical Research Centre (CRC) at Northwick Park (Harrow) but shortly afterwards moved to Oxford as Regius Professor of Medicine. The double move led to the disbandment of the Statistical Research Unit and many of the staff left the MRC and joined Doll in Oxford. The MRC set up a new small group, the Statistical Research and Services Unit headed by Ian Sutherland. In 1980 the Unit relocated to Cambridge, at which time it became the MRC Biostatistics Unit. As Sutherland's retirement approached, there was some uncertainty about the future of the Unit owing to the MRC's financial crisis, but in 1986 Nicholas Day was appointed Director. In recent years the Unit has increased in size to the point where it is one of

the largest biostatistics groups in the UK. Day retired in 1999 and was succeeded by Simon Thompson.

As the MRC's activities have diversified, and the need for statistical input was recognized increasingly, other statistical groups have been set up within the MRC. Following precedent, a medical statistics group was one of the first to be set up in CRC in 1970, initially under the leadership of Michael Healy. That group ceased to exist in 1993 as a consequence of the closure of the CRC.

Another Unit with considerable statistical expertise is the Clinical Trials Unit. This was founded in 1998 and consists of three Divisions. The Cancer Division was formerly the Cancer Trials Office (CTO), founded in Cambridge in 1977, and coordinates clinical trials in cancer. A meta-analysis group, originally set up by the CTO, continues to coordinate international collaborations on **meta-analysis** of cancer trials and has extended its remit to other disease areas. The HIV Division, formerly the HIV Clinical Trials Centre, conducts and coordinates trials in HIV (*see* **AIDS and HIV**), many of which are international studies. The Division Without Portfolio initiates trials in other areas where there are important questions but either insufficient infrastructure or few clinical trials at present.

While the MRC supports medical research by giving grants to individual scientists, it also provides long-term one of its main means of providing long-term support for research is through its establishments where it employs its own staff, currently three institutes and approximately 50 units. Statisticians are employed in many of these units. Two of the most prominent over the years have been the MRC Pneumoconiosis Unit near Cardiff, once headed by **Archie Cochrane**, and the Environmental Epidemiology Unit in Southampton, both now closed. An Epidemiology Unit in Cambridge has recently been established.

The MRC receives an annual grant to support research from Parliament via the Department of Trade and Industry. It also receives funding for some projects from other government sources including the Health Departments, the Overseas Development Administration, and the Ministry of Defence. Other sources of funding include industry and international agencies such as the **World Health Organization** and the European Commission. The MRC supports some research jointly with medical charities. The MRC is independent in its choice of what research to support.

The focus of this discussion has been on statistics within the MRC. In this regard the MRC has consistently been a major player within the UK, recognizing at a very early stage the essential importance of statistics to sound research, especially in clinical areas. The wider history of the MRC is considered at length in [1–3].

References

[1] Austoker, J. & Bryder, L., eds (1989). *Historical Perspectives on the Role of the MRC*. Oxford University Press, Oxford.

[2] Thomson, A.L. (1973). *Half a Century of Medical Research*. Vol. I. *The Origins and Policy of the Medical Research Council (UK)*. HMSO, London.

[3] Thomson, A.L. (1975). *Half a Century of Medical Research*. Vol. II. *The Programme of the Medical Research Council (UK)*. HMSO, London.

JOAN AUSTOKER & DOUGLAS G. ALTMAN

Medical Research Council Streptomycin Trial

The first **Medical Research Council** (MRC) trial of streptomycin in the treatment of pulmonary tuberculosis [10] occupies a special place in the history of medical statistics. This is not primarily for the findings, although these had important implications for subsequent tuberculosis research, but because the trial provided an explicit model for that research and served as a catalyst for the scientific investigation in man of other treatments and interventions throughout clinical and preventive medicine.

At the time, just after World War II, tuberculosis was the principal medical cause of death among young adults in Europe and the US, and streptomycin was the first drug to offer real promise of effective treatment. The antituberculosis activity of streptomycin was discovered in the US in 1944. Although suggestions for adequately controlled studies were put forward [8, quoted in 12], no controlled assessment of the efficacy of streptomycin in man had been undertaken by 1946. Limited supplies of the drug were made available to the MRC for such an assessment in the UK. The MRC set up a research team, their Tuberculosis Research Unit, with Dr Philip D'Arcy Hart, already deeply involved in tuberculosis research, as its director and Dr Marc Daniels as the clinical coordinator for the study. General responsibility for the planning, direction, and reporting of the study lay, from September 1946, with a special MRC committee with Dr Geoffrey Marshall, a leading tuberculosis physician, as chairman, D'Arcy Hart as secretary, and a membership which included **Professor Austin Bradford Hill**. The principal credit for the study that ensued belongs jointly to D'Arcy Hart, Daniels, and Bradford Hill [7]. All three were well prepared, indeed poised, to undertake the "rigorously planned investigation with concurrent controls" that was needed [10]. Bradford Hill had set down the principles of clinical experimentation in man in 1937 [4], including "random allotment" achieved by strict alternation or (in later editions of [5]), by using random sampling numbers (*see* **Randomization**). D'Arcy Hart [9] and Bradford Hill [13] had separately been involved in planning and executing rigorously controlled MRC trials – the one using alternation and the other random sequences. Daniels had been a principal investigator in the Royal College of Physicians' epidemiologic survey of tuberculosis in young adults [2].

The main aim was to assess the effect of the drug in pulmonary tuberculosis, carefully defined as "acute progressive bilateral pulmonary tuberculosis of presumably recent origin, bacteriologically proved, unsuitable for collapse therapy, age-group 15 to 30". At the time the only treatment for such patients was bed rest, and this fully justified treating the parallel **control** group in the trial with bed rest alone, especially as the available supply of streptomycin was insufficient to treat all such patients. Between January and September 1947, 109 patients, assessed as suitable by a central panel, were admitted to the trial from seven centers in England, Wales, and Scotland. Two patients died within a week, leaving 55 allocated to streptomycin and bed rest (S) and 52 to bed rest alone (C).

The allocation "was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each center by Professor Bradford Hill" [10, reprinted in 6] and contained in a numbered set of sealed envelopes held by the coordinator, the appropriate envelope being opened as each patient was admitted. The details of the control scheme remained unknown to the coordinator and all of the investigators throughout. C patients were not informed they were part of a special study and "usually they were not in the same wards as S patients, but the same régime was maintained". All patients were treated with bed rest for 6 months, and their condition was assessed on admission and monthly thereafter. S patients received in addition 2 g streptomycin daily in four injections for 4 months.

Randomization was shown to have "equalized the groups; if anything, there are more severe cases in the S group". Changes in the radiologic picture were regarded as the most important single measure of response, and were assessed by three specialists reading the films independently, not knowing whether they were of C or S cases. There was fair agreement, any differences being readily resolved at a joint session. The changes during the 6 months are shown in Table 1. The difference in the percentages showing considerable improvement is significant (chance probability less than one in a million; *see* ***P* Value**).

There were correspondingly large differences between the S and C series in other measures and

Table 1 Assessment of radiologic appearance at 6 months as compared with appearance on admission

	Streptomycin group		Control group	
Radiologic assessment	Number	Percentage	Number	Percentage
Considerable improvement	28	51	4	8
Moderate or slight improvement	10	18	13	25
No material change	2	4	3	6
Moderate or slight deterioration	5	9	12	23
Considerable deterioration	6	11	6	11
Deaths	4	7	14	27
Total	55	100	52	100

assessments of response at 6 months. The greatest benefits from streptomycin were among the most acutely ill patients. Most of the improvement in the S cases occurred during the first 2–3 months, and thereafter many patients began to deteriorate. The short-lived benefit from streptomycin was shown to be related to the rapid emergence of strains of tubercle bacilli resistant to high concentrations of the drug. In addition, vestibular toxicity occurred frequently in the S series. The report [10] contained many illustrative case histories with radiographs; an addendum showed a significant difference in mortality between the two series at the end of one year.

The trial was designed to answer a specific group of questions concerning the effects of streptomycin on the progress of tuberculosis in man. The restriction of the intake to pulmonary tuberculosis hitherto treatable only by rest in bed, the random selection of those to receive streptomycin, and the precautions to avoid **bias** in management and assessment of the patients enabled the effects of the added streptomycin to be separated from those of the natural course of the disease. The answers for that type of tuberculosis, in a particularly lucid and detailed report, were unequivocal, on efficacy and its duration, on bacterial resistance, and on drug toxicity. The limited amount of streptomycin available to the MRC greatly facilitated the introduction of the random allocation scheme that was crucial to the reliability of the trial findings. Many years later, Bradford Hill [7] expressed doubts whether the random allocation would have been achieved if supplies had been greater.

The clear demonstration in the report of the advantages and disadvantages of the first effective drug treatment for a widespread and lethal infectious disease made a considerable impact on clinicians and statisticians. In the field of tuberculosis, the trial initiated a 40-year series of mostly multicenter controlled chemotherapy trials [3] (*see* **Multicenter Trials**) under MRC auspices in Britain, East Africa, India, and Hong Kong: in different types of tuberculosis; assessing new drugs in combinations to combat bacterial resistance; comparing different durations of treatment for the prevention of subsequent relapse; comparing treatment at home with treatment in a sanatorium [1, 11]; investigating alternative dosages and rhythms of administration to reduce toxicity; and comparing supervised with unsupervised regimens to improve compliance.

The use of random allocation, and the attention paid throughout all aspects of the trial design to obtaining an **unbiased** comparison, also provided a major stimulus to the postwar development of the **clinical trial** in other diseases. Although there is still a place for systematic allocation schemes that are effectively random, e.g. [9], the introduction of allocation schemes based on random sequences is undoubtedly the most important single advance in the evolution of the clinical trial [12] during the twentieth century (*see* **Randomized Treatment Assignment**).

References

[1] Andrews, R.H., Devadatta, S., Fox, W., Radhakrishna, S., Ramakrishnan, C.V. & Velu, S. (1960). Prevalence of tuberculosis among close family contacts of tuberculous patients in South India, and influence of segregation of the patient on the early attack rate, *Bulletin of the World Health Organisation* **23**, 463–510.

[2] Daniels, M., Ridehalgh, F. & Springett, V.H. (1948). Tuberculosis in Young Adults, *Report on the Prophit Tuberculosis Survey 1935–1944*. H.K. Lewis, London.

[3] Fox, W., Ellard, G.A. & Mitchison, D.A. (1999). Studies on the treatment of tuberculosis undertaken

by the British Medical Research Council Tuberculosis Units, 1946–1986, with relevant subsequent publications, *International Journal of Tuberculosis and Lung Disease* **3**(10) Supplement, S231–S279.

[4] Hill, A.B. (1937). Principles of medical statistics I. The aim of the statistical method, *Lancet* **i**, 41–43. Reprinted in [3].

[5] Hill, A.B. (1937). *The Principles of Medical Statistics*, 1st Ed. Lancet, London.

[6] Hill, A.B. (1962). *Statistical Methods in Clinical and Preventive Medicine*. Livingstone, Edinburgh.

[7] Hill, A.B. (1990). Memories of the British Streptomycin Trial in Tuberculosis. The first randomized clinical trial, *Controlled Clinical Trials* **11**, 77–79.

[8] Hinshaw, H.C. & Feldman, W.H. (1944). Evaluation of chemotherapeutic agents in clinical tuberculosis: a suggested procedure, *American Review of Tuberculosis* **50**, 202–213.

[9] Patulin Clinical Trials Committee, Medical Research Council (1944). Clinical Trial of Patulin in the Common Cold, *Lancet* **ii**, 373–375.

[10] Streptomycin in Tuberculosis Trials Committee (1948). Streptomycin treatment of pulmonary tuberculosis, *British Medical Journal* **ii**, 769–782. Reprinted in [4].

[11] Tuberculosis Chemotherapy Centre Madras (1959). A concurrent comparison of home and sanatorium treatment of pulmonary tuberculosis in South India, *Bulletin of the World Health Organization* **21**, 51–144.

[12] Vandenbroucke, J.P. (1987). A short note on the history of the randomized controlled trial, *Journal of Chronic Diseases* **40**, 985–987.

[13] Whooping-Cough Immunization Committee (1951). The prevention of whooping-cough by vaccination. A Medical Research Council investigation, *British Medical Journal* **i**, 1463–1471.

I. SUTHERLAND

Medicare Data

A portion of this work was supported by the ResDAC contract from CMS.

Medicare is a federal health insurance program for nearly all Americans aged 65 and over, for younger persons entitled to Social Security disability payments for at least 2 years, and for people with end-stage renal disease (ESRD). It is administered by the Center for Medicare and Medicaid Services (CMS, formerly the Health Care Financing Administration, HCFA) in the US Department of Health and Human Services. In 2003, persons with entitlement due to these mandates numbered approximately 34.7 million, 5.8 million and 300 000 respectively. By 2020, total enrollment is expected to exceed 60 million. Medicare data are a notable research resource because of the program's huge size and national scope, the range and quality of information captured, the fact that each beneficiary's eligibility is known in each month, and because most beneficiaries remain continuously enrolled until their death. Medicare is the largest single purchaser of health care services in the United States, spending over $260 billion dollars in 2003, including three-quarters of costs for patients with ESRD, nearly half of all hospital revenues and over 70% of hospice spending.

Medicare files contain information on all beneficiaries enrolled and the nature of their benefit, medical problems, (diagnoses) and utilization (claims and costs) for the approximately 87% of Medicare beneficiaries enrolled in the "traditional" fee-for-service (FFS) plan, as well as financial and other descriptive data on licensed providers (e.g. hospitals and physicians).

Eligibility data, including extensive demographics and enrollment information, are available in the annual Denominator files, with one record per beneficiary eligible to receive Medicare benefits during any part of that year. Medicare enrollment is by calendar month, with 12 monthly indicators of when each person is enrolled for Part A (Hospital Insurance, HI) and Part B (Supplemental Medical Insurance, SMI, covering physician and ambulatory care) benefits, and whether they are receiving the FFS or managed care (Medicare Advantage, formerly Medicare + Choice) option.

The valuable annual Medicare Provider and Review (MedPAR) research file, containing one record per hospital or skilled nursing facility discharge, has been available since 1984. Information includes: a hospital identifier, admission and discharge dates, up to 10 diagnoses and 10 procedures, and the **diagnosis related group (DRG)**, a classification that largely determines Medicare's payment. The first listed, or "principal", diagnosis conveys the medical condition determined (at the time of discharge) to have been the principal cause for hospitalization. Charges across several categories of services (e.g. bed charges and intensive care unit utilization), allowed charges, and payments (reimbursements) by CMS and others are also noted.

Research opportunities expanded considerably in 1992 when "final action" (fully adjudicated billing) data from skilled nursing facilities, home health care, hospice, hospital outpatient and physicians, other suppliers and health professionals became available in standard analytical files (SAFs). All SAFs contain provider identifiers, service date(s), diagnoses, services provided, and CMS payments.

Methodological studies validating the Medicare claims information, the demonstrated value of claims data in predicting health care outcomes, such as cost, hospitalization and death, and the recorded experience of nearly all Americans over the age of 65, make Medicare data an important source of information on health care delivery and its consequences in the United States. The data, of course, only relate to covered benefits. Thus, at present they contain information on drugs only when administered in facilities (such as chemotherapy), and most nursing home utilization (because it is covered by Medicaid, rather than Medicare) is missing. Also, the data indicate when a test was done, but not the result; the absence of follow-up treatment could equally well indicate a negative finding or a lapse in patient management. While dates of death are accurate and complete, the data contain neither cause nor place of death. (This information can be merged on from the National Death Index, maintained by the **Centers for Disease Control and Prevention (CDC)**.) Comprehensive profiles of a person's medical problems can be extracted from diagnoses coded on claims using the **World Health Organization's international classification of diseases** (ICD) coding system [2, 6]. A "clinically modified" ICD version, ICD-9-CM, maintained by the National Center for Health Statistics

(NCHS), has been in use in the United States since 1980. However, detailed physiologic information, such as blood pressure or hematocrit readings, cancer stage or cardiac ejection fraction, is not available.

Data Completeness

The Medicare program is complex. Knowing what can be found in the claims files (and where) requires understanding the program's benefits and how they are paid. For example, the managed care option (called Medicare Advantage, formerly Medicare + Choice) allows beneficiaries to choose a commercial managed care plan to coordinate their care (*see* **Health Services Organization in the US**). For the beneficiary, this can provide reduced or waived copayments and deductibles or pharmacy benefits. However, these arrangements create data gaps, since CMS does not currently require Medicare Advantage plans to submit encounter records (dummy claims). Between 2000 and 2003, CMS used diagnoses from Medicare Advantage hospitalization records to calculate payments to plans based on enrollees' expected future health care use [4]. The hospitalization records used in these calculations, however, are not systematically available in research files, although some Medicare Advantage hospitalizations appear in MedPAR. Medicare Advantage enrollment of the beneficiary and readmission within two weeks of a previous discharge are the main reasons for zero-reimbursement MedPAR hospitalizations. Beginning in 2004, CMS will calculate payments to plans based on "expected future need" as calculated from submitted lists of their enrollees' medical problems. It will no longer require encounter records (either hospital based or ambulatory) from Medicare Advantage plans [7].

The claims data can also be incomplete for the 3 to 4% of beneficiaries who do not take part in both parts of the Medicare benefit. Medicare Part A (hospital, skilled nursing facility and hospice) coverage is automatically received by almost all Medicare recipients, while Part B (principally physician and ambulatory care services) coverage is available for a monthly fee ($58.70 in 2003). Persons with both Parts A and B coverage use more Part A services than those with Part A coverage only. Why? Part B coverage provides such good value, that people who do not purchase it often have generous alternative insurance that pays for their hospitalizations as well. Because billing data provide an incomplete record of care for both M + C enrollees and those missing either Part A or Part B entitlement, studies of service gaps (such as failure to receive appropriate surgical follow-up) typically exclude such enrollees.

Other data gaps are less easy to remedy. For example, although the veterans health affairs system maintains merged Medicare/VHA files and has documented the substantial size and nature of overlapping health care use for over 1 million Medicare beneficiaries who use the VHA [8], Medicare data cannot identify VHA users, and thus, (unknowingly) views their Medicare utilization as complete.

Another form of data incompleteness occurs for surgical care when multiple provider/patient encounters are "bundled" into a single global payment. Suppose, for example, that a surgeon receives a single payment for all the care surrounding a mastectomy: preoperative evaluation, the surgery itself, and routine postoperative care, including routine follow-up. No individual bills are submitted for bundled surgical care (and if they were, they would be rejected), so the claims data provide no evidence as to the nature or timing of the follow-up services actually delivered.

Demographic Data

Medicare enrollment files contain dates of birth and death, gender, and race. Since 1994, every beneficiary is assigned to one of seven race codes: white, black, Asian, Hispanic, North American Native, other, and unknown. These codes are the basis for numerous studies examining whether patterns of care and care outcomes vary across racial groups. A study comparing self-reported race with Medicare's race variable, found good accuracy for blacks, but other racial groups were often misidentified as white [1] probably because an earlier form of the variable offered only four choices: white, black, others, and unknown. Undercounts are most problematic for Native Americans and Hispanics. The undercount of Native Americans, a not-very-populous group that may be receiving particularly poor **quality care**, makes studying their care particularly difficult. Even the "new" Medicare race categories pose problems, since most federal race coding schemes (including the National Death Index and the Census) distinguish between race (white, black, Asian, North American Native) and ethnicity (Hispanic heritage or not). Thus, people can be white Hispanic, white non-Hispanic, and

so on. In Medicare, a person of Hispanic heritage is so classified, regardless of race. Despite these problems, Medicare data are crucial for studying racial disparities in health care because most commercial insurance databases do not record race.

Although Medicare knows each beneficiary's residential address, standard research files indicate, at most, state, country, and zip code of residence. Socioeconomic status (SES) indicators, such as income and education, are not directly available, but can be proxied from census data (such as median income or percent with less than a high school diploma) merged in at the zip code or census-tract level.

Nearly one-fifth of Medicare FFS beneficiaries are also dually entitled to Medicare and some form of Medicaid benefit. Medicaid benefits, such as pharmaceuticals and nursing home care, frequently supplement, rather than substitute for, Medicare utilization, and most, but not all, Medicaid beneficiaries can be identified in Medicare through the denominator file variable "state buy-in", an indicator of participation in one of eight [9] programs. These programs provide a variety of benefits ranging from help in paying Medicare premiums through full Medicaid coverage. The Medicaid eligibility variable ("state-buy-in") is often treated as an indicator of poverty. Researchers write, for example, that "the effect of factor X disappears 'after controlling for' Medicaid enrollment (or poverty)" However, "state-buy-in" does not identify all Medicaid enrollees, nor does it coincide with a coherent definition of poverty. While all states' Medicaid programs require income less than twice the federal poverty level, program income thresholds differ. Further, even for those entitled to Medicaid, enrollment requires a request. Thus, this variable identifies some, but not all, low-income elderly, and some, but not all, Medicare beneficiaries who also receive Medicaid benefits.

The group health plan master file contains data on beneficiaries who have ever been enrolled in a Medicare Advantage (managed care) organization, including dates of enrollment and changes in enrollment, as well as the specific plan(s).

Utilization Data

One of the most challenging aspects of studying Medicare utilization is figuring out how health care is divided across files. For example, emergency room (ER) care is stored in inpatient files (such as the MedPAR) if the ER use results in a hospitalization, but is stored in the "outpatient" file if it is not followed by a hospitalization. Likewise, the meaning of "outpatient" is not the opposite of "inpatient". Medicare files are divided by type of care and billing form. Claims from providers that bill using the UB-92 will be in different files than claims from providers that bill on CMS-1500 forms. Some patients are treated on an ambulatory basis by facilities that bill using the UB-92. Those bills will be found in the Outpatient file, whose name refers to hospital outpatient departments. Other providers that treat ambulatory patients use the CMS-1500 form for billing. The Carrier file, formerly called the *physician/supplier Part B file* (also, the *national claims history* file, or *NCH*) contains bills from physicians for care provided in any setting, facility bills for care received in freestanding ambulatory surgical centers, and bills from other providers, such as nurse practitioners, ambulances, and freestanding laboratories. Thus, the same type of procedure (such as a cataract excision) resides in the Outpatient file if done in a hospital outpatient department, or in the Carrier file if done in a freestanding ambulatory surgical center.

Finding all procedures or health care system encounters, requires examining both "facility" and "physician/provider" bills. And, payments for a single procedure often generate more than one bill. For example, there will be a facility or technical charge for an x-ray (for using the x-ray machine, for the machine technician, the film, etc.) and a professional charge for the radiologist who reads the x-ray. We are most familiar with this pattern in the context of inpatient hospital stays. The MedPAR contains facility bills for an inpatient hospital stay (the technical charges) and the Carrier file will contain bills from physicians who care for the patient during the stay – emergency room physicians, radiologists, surgeons, anesthesiologists, cardiologists, and so on. These physician bills are the professional components. If users of **administrative data** are not careful, they can overcount services by counting technical and professional bills separately rather than combining them into a single service.

Coding protocols for procedures are entirely different depending on whether they appear in facility bills (ICD-9) or are coded from doctors' offices (current procedural terminology, CPT-4). While the

ICD system is public and can be obtained from NCHS [10], CPT is copyrighted by the American Medical Association and its use in any product or publication requires a license [11]. The CMS-1500 bill, used both for some facility and all physician bills, codes services (procedures) in three components: the code, the modifier, and the units. Sometimes both physician and facility bills for the same service are submitted on the same CMS-1500 form. The code modifier provides further details on the service provided, or by whom. Modifiers answer questions such as: was the surgery on the left or right eye? Is the bill for a facility, a physician, or both? Is it for a solo surgeon or an assisting surgeon? Information in the units field on the CMS-1500 varies by provider type, such as: ambulances are paid by the mile, anesthesiologists by the minute, and radiation oncology by number of treatments. For people in special circumstances, additional information is often available in separate files. All Medicare-certified nursing homes must collect extensive data on the health status on admission, at least quarterly, and with a significant change in status, for all patients. These data are available to researchers in the minimum data set (MDS) [12]. Another interesting Medicare data source is the outcome and assessment information set (OASIS) data for home health care. OASIS records the problems being addressed, and **comorbidities** and functional impairments used to determine a Home Health Resource Group (HHRG) global payment for all home health services during a 60-day "episode" [5].

Some information used to justify submitted bills (such as proof of an inconclusive CT scan prior to PET scan) is not available in Medicare research files.

Provider Files

Medicare also maintains substantial information about the providers with which it contracts, such as physicians, hospitals, skilled nursing facilities, providers of durable medical equipment. The provider of services file (POS) contains information about institutional providers including freestanding ambulatory surgical centers. Cost report files contain detailed information about the costs to institutions of providing care for Medicare beneficiaries and how that care relates to payments received. The Unique Physician Identification Number (UPIN) master file contains information about physicians such as their specialty and practice location. These files can be ordered directly from CMS [13].

Special Linked Data Files

The Medicare Current Beneficiary Survey (MCBS) is a rolling **panel survey** of a nationally representative sample of about 18 000 beneficiaries, living independently or in long-term care facilities. This CMS survey provides longitudinal information on health service utilization, insurance, and expenses (by all payers, not just Medicare); health and functional status; income, assets, living and care arrangements, and **quality of life**. These data are linked to Medicare's usual files and provide a unique source of data relating to the total health and health spending profile of Medicare beneficiaries [14].

The Surveillance, Epidemiology, and End Results Program (SEER), run by the National Cancer Institute (NCI), collates data from participating cancer registries across the United States (*see* **Cancer Registries**). These registries have been linked with Medicare data in a data system called "SEER-Medicare", maintained by the NCI [3] or [15]. The linked SEER-Medicare files contain all incident cancer cases identified in SEER registries and all Medicare claims for these cases from 1986 forward (regardless of when they were diagnosed with cancer). A comparison file is also available, containing Medicare claims for a similar period of time for a 5% **random sample** of Medicare beneficiaries living in cancer registry areas.

The United States Renal Data System (USRDS), begun in 1988 and jointly funded by the National Institute of Diabetes and Digestive and Kidney Diseases and CMS, collects, analyzes, and makes available for extramural research, substantial information in addition to Medicare's usual data. USRDS data describe entry into the program (date, initial treatment modality, cause of renal failure, and other health factors) and follow-up information (such as wait-list status for transplant, changes in treatment modality, and the date and cause of death) [16].

Data Artifacts

Observed changes in patient care data over time can be artifacts. Across the history of the Medicare program, there have been numerous changes in coding,

payment, and recording of services. It is important to distinguish true secular changes from mere data changes. For example, since 1991, the number of preventive services covered by the Medicare program has steadily increased. Colorectal cancer **screening** became a covered benefit in 1998 and prostate cancer screening in 2000. Previously, when only diagnostic testing (but not screening) was covered, some people received screening coded as a diagnostic work-up, some paid for screening themselves (and the bills did not show up in Medicare), and many simply were not screened. Thus, it is hard to disentangle changes in screening and diagnostic testing from changes in how these tests appear in Medicare's bills. Secular trends can be subtle. Both absolute and relative payment rates have changed during the 1990s, particularly for physicians. The prospective payment system, applied to hospitals since the 1980s, has now been expanded to include hospital outpatient services, home health, and skilled nursing facility care. Over this same period, Medicare managed care expanded considerably and then retracted slightly. Racial coding shifted from four levels to seven. Diagnostic codes, presently using the ICD-9-CM coding system, and becoming richer every year, will eventually convert to the ICD-10 formats already in use in other countries. The lack of a simple map between ICD-9 and 10 will complicate **longitudinal analyses**, just as other changes in the population and their data always have.

Although beneficiaries receive their care under a unique health insurance claim (HIC) identifier, a small percentage (no more that 2% per year) change HICs (most commonly due to a change in marital status for people with Medicare entitlement via their spouse's work history). In longitudinal studies, researchers should request files from CMS that substitute a single identifier for each person.

With over 40 million Medicare beneficiaries, 10 million hospitalizations and more than a billion total annual claims for Medicare enrollees, researchers will inevitably find data problems, such as an occasional person whose age exceeds that of the oldest living person in the Guinness Book of World Records, or service care dates that extend beyond the date of death. Likewise, while problems are occasionally found, zip codes of residence are accurate for 99.9% of beneficiaries.

Data Availability for Research

Although CMS's primary obligation is to serve and protect beneficiaries, it also invests in making Medicare data available to researchers. Given the huge scope of data, it is often adequate (and prudent) to conduct research on Medicare's 5% research files. These files consist of all the data for a random 5% sample of Medicare enrollees.

Most academic researchers use research identifiable files (RIFs). Such data are not sold, but only loaned (on a need-to-know basis) for use in answering a specific question via an approved research protocol. Additional questions may be pursued only with explicit permission. Files have a data retention date, after which they must be returned or destroyed. The data may not be used for direct marketing, to make unauthorized patient contact or to identify individuals. CMS reserves the right to preview presentations and publications, with the key concern that no publication release information on a subgroup with 10 or fewer people. CMS funds a Research Data Assistance Center (ResDAC) to help researchers obtain and appropriately use CMS data [17]. In the past, beneficiary data were also available as encrypted files for use by researchers or research projects that could not meet the threshold set by CMS for using identifiable data. Owing to privacy rules of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), CMS no longer sells the beneficiary encrypted files, substituting limited data sets (LDS).

In summary, Medicare data is a unique source of data for **health services research**, due to its richness, size, national representativeness, continuity of enrollment, presence of race and date-of-death data, well-documented history of use in research, and the help available for researchers wishing to use it for studies of potential benefit to Medicare beneficiaries.

References

[1] Arday, S.L., Arday, D.R., Monroe, S. & Zhang, J. (2000). HCFA's racial and ethnic data: current accuracy and recent improvements, *Health Care Financing Review* **21**(4), 107–116.

[2] Ash, A.S., Ellis, R.P., Pope, G.C., Ayanian, J.Z., Bates, D.W., Burstin, H., Iezzoni, L.I., MacKay, E. & Yu, W. (2000). Using diagnoses to describe populations and predict costs, *Health Care Financing Review* **21**(3), 7–28.

[3] Ashton, C. & McHorney, C. *Medical Care* **40**(8: suppl), (2002).
[4] McCall, N., Harlow, J. & Dayhoff, D. (2001). Rates of hospitalization for ambulatory care sensitive conditions in the Medicare + Choice population, *Health Care Financing Review* **22**(3), 127–145.
[5] OASIS and Outcome-Based Quality Improvement in Home Health Care: Research and Demonstration Findings, Policy Implications, and Considerations for Future Change. By Peter W. Shaughnessy, PhD; Kathryn S. Crisler, MS, RN; David F. Hittle, PhD; Robert E. Schlenker, PhD.
[6] http://www.cdc.gov/nchs/icd9.htm.
[7] http://www.cms.hhs.gov/healthplans/rates/2004/cover.asp.
[8] http://www.virec.research.med.va.gov.
[9] http://www.cms.hhs.gov/dualeligibles/bbadedef.asp.
[10] http://www.cdc.gov/nchs/products/elec_prods/subject/icd96ed.htm.
[11] http://www.ama-assn.org/ama/pub/category/3657.html.
[12] http://www.cms.hhs.gov/quality/mds30/.
[13] http://www.cms.gov/data/purchase/default.asp.
[14] http://www.cms.hhs.gov/MCBS/default.asp.
[15] http://www.healthservices.cancer.gov/seermedicare/.
[16] http://www.usrds.org.
[17] http://www.resdac.umn.edu.

Further Reading

Claims data sets generally and Medicare in particular:

Iezzoni, L.I. ed. (2003). *Risk Adjustment for Measuring Health Care Outcomes*, 3rd Ed. Health Administration Press, Ann Arbor, pp. 83–138. Medicare and other important claims data sets (VA, Medicaid), their strengths and weaknesses are discussed extensively in Chapter 5 "Coded Data from Administrative Sources".

Health, United States, (2003). US Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. Hyattsville. Appendix I, Data Sources.

Center for Medicare and Medicaid Services websites:

www.cms.hhs.gov/data/default.asp
www.cms.hhs.gov/researchers/

ARLENE S. ASH & BETH VIRNIG

Medicines and Healthcare Products Regulatory Agency (MHRA) (Formerly MCA)

The UK Medicines and Healthcare products Regulatory Agency (MHRA) was established in April 2003 by the merger of the previously existing Medicines Control Agency (MCA) and Medical Devices Agency (MDA). The MHRA has responsibility for the protection and promotion of public health through the regulation of the safety, quality, and efficacy of human medicines (*see* **Drug Approval and Regulation**) and for monitoring medical devices. The Agency has a Medicines and a Devices Sector – most of the statistical involvement in terms of professional statisticians within the Agency comes from within the medicines sector and this article focuses on those activities. Before the merger, the Devices Agency used outside statisticians for some assessments of applications for new products.

The MHRA safeguards public health by ensuring that all medicines on the UK market meet the appropriate standards of safety, quality, and efficacy. It does so using duties and powers in the Medicines Act 1968, as well as with more recent UK and European legislation.

The principal functions of MHRA are the assessment of applications from pharmaceutical companies for marketing authorizations, surveillance of medicines after they are on the market (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**), inspection of manufacturing sites, enforcement of the Medicines Act, and the setting of public health standards for pharmaceutical products.

The regulatory framework within which the MHRA operates includes:

1. The *Licensing Authority* is stipulated in the Medicines Act as the Ministers responsible for Health in England, Scotland, Wales, and Northern Ireland, together with the Agriculture Ministers for veterinary medicines. This is invariably unified through decisions and actions taken by the Secretary of State for Health in England, acting through the MHRA for the entire United Kingdom. The MHRA is the executive body that carries out the decisions of the Licensing Authority. In coming to decisions as to the granting of a marketing authorization or action on changes to current authorizations, Ministers may take expert advice from the *Committee on Safety of Medicines* (CSM) and from the *Medicines Commission* (MC).
2. The *Medicines Commission*, an independent body with academic, consumer, and industry representation, also advises the Secretary of State on a wide range of policy matters related to the regulation of medicines, including veterinary medicines. It also hears representation from applicant companies appealing against advice from the CSM.

Each year around 20 to 30 new biotechnologic and other products receive a centralized European authorization, valid in all European Union (EU) Member States. Assessment is coordinated by the European Medicines Agency (EMEA), but professional staff of the EU Member States carry out the scientific evaluation. The MHRA is one of the leading evaluators in this European system. For these products, the European Community is the Licensing Authority. Their licensing decisions are based on recommendations from the Committee for Human Medicinal Products (CHMP), which is the expert advisory group to the EMEA and comprises experts from each Member State. However, the MHRA acts on behalf of the UK licensing authority to undertake postmarketing surveillance for products marketed in the United Kingdom, as this is a national responsibility.

In summary, the key activities of the MHRA (Medicines Sector) are to implement and enforce domestic European legislation, by

1. issuing and maintaining marketing authorizations, which allow drugs to be marketed in the United Kingdom, thus protecting the health of UK citizens; this involves issuing national licenses, and participating in the centralized and mutual recognition of European licensing systems;
2. monitoring medicines on the UK market for adverse reactions and taking any necessary action in domestic and international contexts to protect UK public health;
3. controlling the sale and supply of medicines;

4. inspecting and licensing manufacturers and wholesalers to ensure the quality of medicines on the UK market;
5. enforcing Medicines Law;
6. advising Ministers on medicines regulation in both the domestic and European as well as wider international context;
7. publishing, through the *British Pharmacopoeia,* standards relating to the quality of medicines.

The Agency has approximately 750 staff, which as of 2004 includes four professional medical statisticians working on licensing issues and one working on pharmacovigilance. There are additional professional academic statisticians who sit on advisory boards such as the Committee on Safety of Medicines and the Medicines Commission. In 2003, the Agency issued 24 licenses for new drugs, 1500 abridged licenses, and 21 000 variations to licenses, as well as receiving and analyzing nearly 17 000 adverse drug reaction reports from the United Kingdom and 44 000 from outside the United Kingdom. All of this work is in the context of increasing complexity in the scientific assessment work that needs to be undertaken and in the legal and administrative environment.

Licensing work also includes control of **clinical trials**, parallel imports, and registration of homeopathic medicines.

Postmarketing surveillance is a particularly important area for protecting public health and is an essential part of the regulatory process. New medicines inevitably come to market with limited experience. After licensing, safety issues not identified from clinical trials are recognized and acted upon. Statistical work has been done to improve the processes of analyzing these adverse reaction reports so that the UK system for surveillance tends to be more developed than those of most other Member States, and hence, it tends to be at the forefront of action required to protect public health. The Licensing Authority has powers to revoke or suspend licenses compulsorily. These issues tend to be controversial and complex in nature, and attract considerable media attention.

Further information about MHRA is available from their web site at `www.mhra.gov.uk`.

STEPHEN J.W. EVANS & SIMON J. DAY

Medico–Legal Cases and Statistics

Medico–legal statistics concerns the use of biostatistical and epidemiologic data and methods in the context of law and government regulatory policy. Statistical reasoning has an important role in legal cases concerning product liability, compliance with environmental or **occupational health** and safety laws, the safety and efficacy of drugs and medical devices, medical malpractice, and the determination of safe levels of exposure to potentially toxic chemicals (*see* **Risk Assessment for Environmental Chemicals**). Recently, several Lanham Act cases concerned with commercial issues such as fair advertising, validity of patents, and intellectual property have involved biostatistical evidence. An introduction to the area is given here with citations to relevant cases and references.

Malpractice

An illustrative example is the *Brochner* malpractice case, 724 P2d 1293 S.Ct CO (1986). In 1964 a hospital granted Dr Brochner staff privileges and over the next few months he performed numerous craniotomies. The hospital noticed that tissue samples from many of his patients appeared normal and in 1966 required that he obtain outside consultation. In March of 1968 the hospital learned that 14 of 28 tissue samples of his neurosurgery patients were normal. In November 1968 Dr Brochner performed a craniotomy on Ms Cortez, who was injured. She filed suit against both the doctor and the hospital. At the trial her expert testified that an acceptable rate of normal tissues amongst patients who were operated on was one in 100 and that two of 28 should require investigation. Essentially, the expert used the **Poisson** approximation to the **binomial** to conclude that the probability of observing two or more normal tissues in a sample of 28 when each sample has probability 0.01 of being normal is about 0.001. Similarly, the probability of observing 14 or more normal tissues assuming Dr Brochner was selecting patients according to the standard medical criteria was less than one in a billion. Thus, the statistical evidence strongly supported the plaintiff's claim that Dr Brochner was subjecting healthy patients to the risk of surgery, and the time when the information was known showed that the hospital was aware of it but did not revoke his right to operate. Both parties settled with the plaintiff and the lawsuit concerned whether the doctor was liable to the hospital's insurer. As both parties were at fault, he did not need to indemnify the company. McClellan [28] discusses recent developments that have eased plaintiffs' burden of proving that they were harmed by a doctor's or hospital's negligence. Early cases often required that an expert state that had the negligent act not occurred to a "reasonable medical certainty" the patient would not have been harmed. In *Hamil vs. Bashline*, 392 A.2d 1280, S.Ct PA (1978) the court awarded damages when an expert testified that the patient would have had a 75% chance of survival had the negligent act not occurred. Thus, probabilistic estimates of the survival probabilities assuming the appropriate medical treatment should play a role in determining liability and in the award of damages.

Food and Drug Law

In the United States new drugs have to be approved for safety and efficacy by the **Food and Drug Administration (FDA)**. To demonstrate efficacy the FDA requires companies to submit two well-controlled studies establishing that the drug will have the effect it purports to have (*see* **Drug Approval and Regulation**). The formal regulations, approved of by the Supreme Court in *Weinberger*, 412 US 609 (1982), state that the study plan should describe the selection of subjects, and use **randomized treatment assignment** to assure the groups are balanced with respect to relevant **covariates** such as age, prior health, etc. Usually, the efficacy requirement is satisfied by a placebo-control double-blind (*see* **Blinding or Masking**) study (*see* **Clinical Trials, Overview**) [38]. Both the US and UK have set up a multiphase process (21 CFR Sec. 314.126; [32]), first to determine the pharmacodynamic properties and safe dose level of a new drug and finally to assess its efficacy. In response to the rapid increase in **AIDS** cases, the FDA modified some of its previous procedures to bring new therapies to patients and clinicians [34], and may place greater weight on the preferences of the diseased population in weighing the factors of safety and efficacy in its future risk–benefit decisions. The standards

of the countries in Europe are being unified by the European Community. At least one clinical trial demonstrating efficacy is required, and recent developments are reviewed by Kingham et al. [21] and Lewis et al. [26]. The standards for approving medical devices are described by Horton [20] and Munsey [31].

Statistical issues often arise in assessing clinical trials. There has been a substantial literature concerning whether one-sided or two-sided tests should be used [9, 11, 22, 33] (*see* **Alternative Hypothesis**). The FDA typically requires two-sided tests when the **controls** receive a standard regimen and the new drug is given to the experimental group. Even with placebo controls, two-sided tests are used to ensure that any observed effect in the treatment arm exceeds the "placebo effect". One-sided procedures have been developed for **bioequivalence** studies (*see* **Equivalence Trials**) where the **null hypothesis** is nonequivalence and the alternative is equivalence [8]. Often proponents of a new drug or procedure assert that studies that do not show an effect should not detract from studies showing a statistically significant effect or that the drug is effective in several subgroups of patients and should be approved for them. The appropriate approach is to account for the number of tests made using **multiple comparison** methods [18, 30] and to use combination methods, such as Fisher's summary chi-square, to pool the results of several studies (*see* **Meta-analysis of Clinical Trials**).

Generic drugs are regulated under different laws from those for new drugs [12]. Typically, a manufacturer needs to show that its product is bioequivalent to one which was approved as a new drug previously. This means that the drugs deliver the same total dosage over a reasonable time period, e.g. 24 h, and reach similar maximum dose levels at about the same time. Both the FDA and the producer of the original older drug may challenge the data used to support the claim of bioequivalence. The FDA requires a generic equivalent to carry out full testing if its excipients differ from the already approved drug. In *Premo vs. US*, 475 F.Supp. 52 (SD NY, 1979) the district court accepted the Premo's equivalence data (reproduced in the opinion and in Gastwirth [14, p. 775]); however, the appellate court reversed this decision and held that the lower court wrongly substituted its opinion on safety and efficacy for that of the FDAs. The data showed that the established drug reached a higher dose level than the generic one and reached that level faster.

In addition to approving drugs, in the US the FDA also monitors adverse effects by requiring manufacturers to report them. As the ascertainment of cases and the follow-up information on the patient's status are often incomplete, standard methods for estimating **relative risks** are inappropriate. Brookmeyer & Yasui [4] discuss the use of passive surveillance (*see* **Follow-up, Active Versus Passive**) **disease registry** data and show that it is a useful supplement to **cohort** data.

Unfair or False Advertising Cases

The Lanham Act is designed to ensure that consumers receive reasonably accurate information about products and to assure manufacturers who produce reliable products that others will not unfairly infringe on the good reputation of those products. A number of major cases have dealt with the fairness or accuracy of advertisements, especially comparative advertisements where one firm asserts that its product is superior to that of a named competitor. There are two types of cases. In the first case firm A claims that its product is superior to firm B's but does *not* assert that this claim is supported by scientific studies. The second type deals with cases in which firm A has stated or implied that the claim of superiority is supported by scientific studies. As the law recognizes that the public takes a skeptical view of unsupported claims, the burden on the complaining party, firm B, differs in the two types of cases.

In the first type, to challenge successfully an unsupported claim that A's product is better, firm B needs to show that the claim is false. This is usually accomplished by firm B carrying out an appropriate study demonstrating that the two products are equivalent, i.e. it is not necessary for firm B to show that its product is better – only that A's product is not superior. In the second class of cases, firm B needs only prove that the studies firm A relied upon to support its claim were not sufficiently reliable to permit one to conclude, with reasonable certainty, that the claim of superiority was established. Sometimes a suit is brought by a competitor, but the Federal Trade Commission (FTC) and FDA also regulate health claims [36]. In the case of *Stirling Drug Inc. vs. FTC*,

741 F.2d 1146 (9th Cir., 1984), the firm's advertisement claimed that Bayer aspirin was pharmaceutically superior to its competitors with regard to purity, freshness, and speed of disintegration, and indicated that these claims had been established by scientific means. Stirling Drug had based its claim on one in-house study, and the FTC's analysis indicated that the claims of superiority regarding the named attributes were not supported. The opinion noted that the FTC requires a well-controlled clinical trial where real patients having actual symptoms are studied. There should be a written protocol (*see* **Clinical Trials Protocols**), double-blinding, and a placebo control, and the data should be analyzed by established statistical techniques and indicate both statistical and clinical significance (*see* **Clinical Significance Versus Statistical Significance**). In *Proctor & Gamble* vs. *Chesebrough Pond Inc.*, 747 F.2d 114 (2d. Cir., 1984) the first firm claimed its hand lotion was superior to the product made by its competitor, while the second firm only advertised that no other lotion was better. After describing the studies [14, pp. 777–780], the court decided that neither company's evidence was sufficiently strong to justify a preliminary injunction, which would have stopped the advertisements of the party with the much weaker case until the full trial was held. Related cases involving the soundness of the clinical studies supporting advertising claims are *Thompson Medical Co. vs. Ciba Geigy*, 672 F. Supp. 679 (SD NY, 1985), *ALPO Pet Food Inc. vs. Ralston Purina*, 913 F.2d 958 (DC, 1990), *McNeill-P.C.C. vs. Bristol Myers Squibb Co*. 938 F.2d 1544 (2d Cir., 1991), and *Mylan Labs. vs. Metkari* 7 F.3d 1130 (4th Cir., 1993). The *Rhone-Poulenc Rorer Pharm. Inc. vs. Marion Merrell Dow* Inc. (8th Cir., 1996) case discussed the **bioavailability** studies of their drugs for treating hypertension and angina. RPR launched its drug, Dilacor, by advertising that it was the same but cheaper than MMD's older drug, Cardizem. MMD's counteradvertising campaign first said that Dilacor had only 50% the bioavailability of Cardizem, but later advertisements based on a better study claimed that Dilacor delivered only 74%–81% of the relative doses of Cardizem. RPR claimed that its own studies refuted the study of MMD; however, the court found that even RPR's study showed a reduced bioavailability of its drug. The court found that RPR's original advertisement falsely represented that the two drugs were interchangeable. It also noted that the early MMD advertisement violated the Lanham Act as it exaggerated the shortfall in bioavailability; however, it found that the advertisements based on the later study were not false and declined to award RPR monetary damages because it had not demonstrated injury. In addition to the biostatistical evidence, MMD noted that surveys showed that consumers viewed the products as similar. When one product infringes on another, it is typical for the opposing party to introduce **survey** evidence demonstrating that consumers are confused by an alleged similarity [7; 14, Chapter 9; 16]. Such evidence is also helpful in estimating monetary damage, as the fraction of consumers who are misled may be regarded as the potential market share the infringer unfairly gained.

Environmental and Occupational Health

To protect public health, governments regulate the amount of potentially toxic chemicals that workers can be exposed to [37], that can remain in foods treated by pesticides [6] or that can escape into the atmosphere or water supply. Reporting and self-monitoring requirements are used to ensure the safety of the public [35, 41]. The responsible government agencies prescribe the sampling procedures that need to be carried out by the producers in great detail. The decision in *US v. Marine Shale Processors*, 81 F.3d 1329 (5th Cir., 1996) shows that courts are not sympathetic to a firm that does not carry out sampling according to the specified procedures and later claims that it was unable to carry out the appropriate statistical tests. In that case the firm's own daily data indicated that it had violated a regulation on 27 occasions.

Many cases concern the scientific underpinning of a mandated exposure limit [13]. In the US, when human data as well as animal experiments indicate an increased **risk** of a serious illness, e.g. cancer, courts typically uphold a regulation [as in *Society of Plastics Industry Inc. vs. OSHA*, 509 F.2d 1301 (2nd Cir., 1975)]. Much more controversy arises when there are little data on humans demonstrating that reducing exposure to the proposed level will actually lower the expected number of cases. In particular, whether or not one assumes a threshold below which no harm is expected to occur can lead to very different estimates of toxicity and corresponding safe dose levels [40] (*see* **Dose–Response Models in Risk Analysis**; **Extrapolation, Low Dose**). The basic literature is reviewed by Armitage [2], Krewski & Brown [23],

and Leape [25]. Recent developments are summarized in Lin et al. [27], who reiterate the fact that many models approximately fit the data and that current bioassay experiments (*see* **Biological Assay, Overview**) do not enable one to estimate reliably the interspecies concordance – the percentage of chemicals classified the same way in tests on both species. The problem of unchecked or unverifiable assumptions arises in other applications of risk assessment in environmental safety [15].

In addition to regulation, members of the public who are exposed to toxic chemicals due to the negligence of a producer or user of the agent can file a tort law claim. In these cases plaintiffs may rely on epidemiologic studies demonstrating that the chemical in question increases the risk of a serious disease. Data from several such cases are reported in Gastwirth [14, Chapter 14]. Harr [17] describes the litigation arising from contamination of well water in Woburn, MA, and Finkelstein & Levin [10, p. 298] discuss the data from the original **case–control study** [24]. The current status of the law is comprehensively treated by Boston [3].

Related Developments

The use of statistical evidence in legal cases has affected the patients or respondents in surveys and case–control studies. The protection of the **confidentiality** that investigators can provide participants has been questioned as litigants have sought to subpoena the raw data underlying a publication. In *Lampshire vs. Proctor & Gamble*, 94 FRD 58 (ND GA, 1982), the district court did not allow access to the patients who served as cases and controls in a study of Toxic Shock Syndrome. The **Centers for Disease Control (CDC)** argued that allowing the defendant firm to reinterview would diminish the likelihood of participation in future studies. Indeed, the CDC submitted a survey of the respondents; a high percentage did not wish to be interviewed by the defendant. The court noted also that there was no reason for respondents to give biased answers, as the cases desired the best treatment, and even the controls would only benefit from good science. For further examples and discussion see Cecil & Boruch [5] and McHale [29], where cases arising from the lack of informed consent in clinical trials are cited.

Outside of violations of informed consent rules, there have been few cases involving liability for injury or death of patients in clinical trials of new drugs. The death of five patients in the clinical trial of the experimental drug fialuridine, however, led to the case *In re* Fialuridine Products Liability Litigation, 163 FRD 386 (DC, 1995), which was settled. Traynor [39] summarizes the principles of products liability law and notes that drug manufacturers of prescription drugs have the benefit of the "learned intermediary rule" that assumes that once the appropriate health care provider has been notified of the potential risks as well as benefits of the drug, the patient also has been properly warned. The rationale is that the physician will take into account the situation of the specific patient. The rule was developed for prescription drugs but may not be relevant for clinical trials as double-blinding means that the treating physician does not know whether the patient is receiving the drug or the placebo. Hence, they can no longer be considered a "learned intermediary" who is in a position to explain the precise risks of the treatment to the patient. Disentangling the effect of the drug from that of the illness may prove difficult, so patients may have a problem in showing that the experimental drug was the legal cause of their injury [39]. Ethically, patients should be compensated for injuries they sustain in clinical trials, but, pragmatically, payment for the costs of medical treatment by the sponsor of the investigational drug may deter claims.

Several cases have been filed alleging that enrollees were not informed of the nature of the risks involved, the financial interests of the sponsors and investigators, or of adverse effects suffered by other patients in the clinical trial. One such case, *Wright versus Fred Hutchinson Cancer Research Center*, described in Jedrey and Feltz [19], is still pending. Subsequent to the filing of the case in 2001 the Center, however, announced changes in some of its policies. Now researchers are prohibited from stock ownership or royalties on patents that are directly and significantly related to their research. Moreover, they are required to disclose to participants and in publications any financial interests have in for-profit companies sponsoring clinical trials. Another pending case, *Grimes versus Kennedy Krieger Institute*, 782 A. 2d 807 (Md. 2001), an affiliate of Johns Hopkins, deals with the legal duty or responsibility researchers owe to participants. The investigators were examining techniques for removing lead paint from residential buildings and planned to measure the amount of lead in the blood of children to assess the effectiveness of

partial lead-paint removal. Allegedly, the researchers encouraged participating landlords to rent to families with small children even though they were aware that they knew of the hazards of lead dust exposure to children. The informed consent document, which simply stated that "lead poisoning in children is a problem" and that exposure to lead in paint, dust, and soil is a major source of exposure in children allegedly failed to mention the known harmful effects of lead. The trial court dismissed the complaint but the appeals court reinstated the negligence suits. It held that the consent agreement imposed special duties on the researcher to ensure that participants are given appropriate warnings of the risks involved.

Recent Developments

The requirement in the U.S. that companies inform employees of medical and scientific literature showing a link between disease and exposure to a toxic chemical was adopted in a recent French case, *Garafalo versus IBM France*, Tribunal de Grande Instance de Nanterre, 11/25/03). The court appointed a medical expert to assess claims that several of the plaintiff's ailments were due to his exposure to unacceptably high levels of ethylene glycol and other toxic solvents in "clean rooms" designed for dust-free manufacture of semiconductors.

While most *false advertising* cases are brought by producers of competing products, governmental agencies can also sue manufacturers who make false safety claims. The case, *Spitzer versus Dow AgroSciences LLC*, N.Y. Sup. Ct. No. 403920/03 was settled after the firm agreed to pay $2 million for violating an earlier consent decree prohibiting it from claiming that its insecticide containing chlorypyrifos was safe. In 2000, the U.S. Environmental Agency had arranged for a phasing out of the use of the chemical in homes and gardens. Scientific studies that showed the chemical was toxic to the brain and nervous system, especially to the development of infants, played a role in establishing that the chemical was harmful.

The importance of proper statistical analysis of data will play a role in a case [1] currently being litigated in which a generic drug company, Ivax Corp. is challenging the validity of a clinical study used by Eli Lilly & Co. when it obtained its patent on Zyprexa, a drug used to treat schizophrenia. In the clinical study of 40 beagles of both sexes were randomly assigned to be given Zyprexa, the current drug (called molecule 222) or a placebo. Lilly used the study to show that its drug was significantly different from 222. The study was originally designed to assess whether the new drug had lower toxicities for blood disorders as the company thought that 222 would cause them. It turned out that neither drug nor placebo caused them; however, a Lilly scientist noted that some of the female dogs given high doses of 222 had higher cholesterol levels. Allegedly, Lilly reanalyzed the data comparing these eight females obtaining a statistically significant difference in cholesterol levels between them and the other females. This result was used to demonstrate that its drug differed from 222. The generic companies claim that Lilly did not inform the FDA that the effect was not seen in male dogs, indicating that the effect might not translate to humans. Lilly argues that it gave all the information to the FDA and that its drug was a major improvement in treatment. From a statistical view, the propriety of examining the data to find a hypothesis to test and testing it on that very data may be an issue. The number of other effects tested, the **multiple comparisons** problem, may deserve consideration by the court. From a legal viewpoint, it will be interesting to compare the court's ultimate decision with the decision in *Warner-Lambert versus Hechler*, 787 F.2d 147 (3rd Cir. 1986). In that case, the court rejected the claim that a drug was effective based on statistically significant results of tests on six subgroups as the firm had conducted 240 comparisons.

References

[1] Abboud, L. (2004). Briefs claim Lilly used flawed study to get patent, *Wall Street Journal* Jan. 26th, B4.

[2] Armitage, P. (1982). The assessment of low dose carcinogenicity, *Biometrics* **38**, *Supplement*, 119–129.

[3] Boston, G.W. (1995). Toxic apportionment: A causation and risk contribution model, *Environmental Law* **25**, 549–649.

[4] Brookmeyer, R. & Yasui, Y. (1995). Statistical analysis of passive surveillance disease registry data, *Biometrics* **51**, 831–842.

[5] Cecil, J.S. & Boruch, R. (1988). Compelled disclosure of research data, *Law and Human Behavior* **12**, 181–189.

[6] Curme, C.S. (1994). Regulation of pesticide residues in foods: Proposed solutions to current inadequacies under FFDLA and FIFRA, *Food and Drug Law Journal* **49**, 609–648.

[7] Diamond, S.S. (1994). Reference guide on survey research, in *Reference Manual on Scientific Evidence*. Federal Judicial Center, Washington.

[8] Dubey, S.D. (1991). Some thoughts on the one-sided and two-sided tests, *Journal of Biopharmaceutical Statistics* **1**, 139–150.

[9] Dunnett, C.W. & Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials, *Statistics in Medicine* **15**, 1729–1738.

[10] Finkelstein, M.O. & Levin, B. (1990). *Statistics for Lawyers*. Springer-Verlag, New York.

[11] Fisher, L.D. (1991). The use of one-sided tests in drug trials: an FDA Advisory Committee member's perspective, *Journal of Biopharmaceutical Statistics* **1**, 151–156.

[12] Fleder, J.R. (1994). The history, provisions, and implementation of the Generic Drug Enforcement Act of 1992, *Food and Drug Law Journal* **49**, 89–107.

[13] Flournoy, A.L. (1991). Legislating inaction: asking the wrong questions in protective environmental decision-making, *Harvard Environmental Law Review* **15**, 327–391.

[14] Gastwirth, J.L. (1988). *Statistical Reasoning in Law and Public Policy*. Academic Press, San Diego.

[15] Gastwirth, J.L. (1989). The potential effect of unchecked statistical assumptions, *Journal of Energy Law and Policy* **9**, 177–194.

[16] Gastwirth, J.L. (1996). Review of reference guide on survey research by Diamond, *Jurimetrics* **36**, 181–191.

[17] Harr, J. (1994). *A Civil Action*. Random House, New York.

[18] Hochberg, Y. & Tamane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[19] Horton, L.R. (1995). Medical device regulation in the European Union, *Food and Drug Law Journal* **50**, 464–476.

[20] Jedry, C.M. and Feltz, M.K. (2002). Clinical trials on trial: Investigator and Institutional Review Board legal risk areas in recent cases, *Toxics Law Reporter*, **17**, 637–642.

[21] Kingham, R.F., Bogaert, P.W.L. & Eddy, P.S. (1994). The New European Medicines Agency, *Food and Drug Law Journal* **49**, 301–321.

[22] Koch, G.G. (1991). One-sided and two-sided tests and *p* values, *Journal of Biopharmaceutical Statistics* **1**, 161–170.

[23] Krewski, D. & Brown, C. (1981). Carcinogenic risk assessment: a guide to the literature, *Biometrics* **37**, 353–366.

[24] Lagakos, S.W., Wessen, B.J. & Zelen, M. (1986). An analysis of contaminated water and health effects in Woburn, Massachusetts, *Journal of the American Statistical Association* **81**, 583–596.

[25] Leape, J.P. (1980). Quantitative risk assessment in regulation of environmental carcinogens, *Harvard Environmental Law Review* **4**, 86–116.

[26] Lewis, J.A., Jones, D.R. & Rohmel, J. (1995). Biostatistical methodology in clinical trials – a European guideline, *Statistics in Medicine* **14**, 1655–1682.

[27] Lin, T., Gold, L.S. & Freedman, D. (1995). Carcinogenicity tests and interspecies concordance, *Statistical Science* **10**, 337–353.

[28] McClellan, F.M. (1994). *Medical Malpractice*. Temple University Press, Philadelphia.

[29] McHale, J.V. (1993). Guidelines for medical research – some ethical and legal problems, *Medical Law Review* **1**, 160–185.

[30] Miller, R.G. Jr. (1981). *Simultaneous Statistical Inference*, 2nd Ed. Springer-Verlag, New York.

[31] Munsey, R.R. (1995). Trends and events in FDA Regulation of medical devices over the last fifty years, *Food and Drug Law Journal* **50**, 163–177.

[32] Newdick, C. (1992). The impact of Licensing Authority approval on pharmaceutical product liability: a survey of American and U.K. law, *Food and Drug Law Journal* **47**, 41–57.

[33] Peace, K.E. (1991). One-sided or two-sided *p* values: which most appropriately address the question of drug efficacy? *Journal of Biopharmaceutical Statistics* **1**, 133–138.

[34] Podraza, R. (1993). The FDA's response to AIDS: Paradigm shift in new drug policy?, *Food and Drug Law Journal* **48**, 351–376.

[35] Reitze, A.W., Jr & Hoffman, L.D. (1996). Self-reporting and self-monitoring requirements under environmental laws, *Environmental Lawyer* **1**, 681–745.

[36] Sachs, E.A. (1993). Health claims in the marketplace: the future of the FDA and FTC's regulatory split, *Food and Drug Law Journal* **48**, 263–283.

[37] Schroeder, E.P. & Shapiro, S.A. (1984). Responses to occupational disease: the role of markets, regulation and information, *Georgetown Law Journal* **72**, 1231–1309.

[38] Smith, J.J. (1992). Science, politics, and policy: the tacrine debate, *Food and Drug Law Journal* **47**, 511–532.

[39] Traynor, M. (1996). Products liability, *National Law Journal*, **Nov. 18**, B6.

[40] Van Stackelberg, K. & Burmaster, D.E. (1994). A discussion on the use of probabilistic risk assessment in human health impact assessment, *Environmental Impact Review* **14**, 385–401.

[41] Whiteman, M.E. (1995). Complying with chemical regulations and new chemical notifications in Europe, *Environmental Claims Journal* **7**, 85–115.

JOSEPH L. GASTWIRTH

Mendel's Laws

The beginning of the science of genetics is usually credited to the experiments of Gregor Mendel, first announced in 1865 [2, 3]. However, while Mendel understood the importance of his own discoveries [1], the rest of the scientific community at the time did not, even though Mendel's experiments were beautifully designed, and his interpretation was clear and insightful. Mendel's work was virtually ignored until 1900 when it was finally rediscovered. The reasons for this initial neglect were twofold. First, Mendel was the first to focus on the numerical relationships among traits appearing in the offspring of carefully controlled matings. This quantitative approach, with its emphasis on probability and ratios, was unfamiliar in biology, which was at the time thought to be too complicated to be understood with such simple models. Secondly, Mendel used discrete instead of continuous characters for his experiments, and introduced the important concept that the hereditary material was particulate. Because the emphasis in discussions of heredity at the time was focused on interspecific differences, and because the prevailing view of heredity was that of continuous variation which blended from one generation to the next, it was difficult for scientists of the time to understand the relevance and importance of the results of Mendel's experiments on a series of simple, discrete traits.

Mendel performed his experiments with seven discrete traits in the garden pea, each trait consisting of a pair of alternative, visible, and highly contrasting characteristics. For example, the seed color could be green or yellow, or the flowers white or violet, with each plant producing just one type of seed or flower color. By focusing on single, dichotomous traits and controlled matings, Mendel was able to use these experiments to determine the underlying, discrete basis of the hereditary system in the pea, and by extension, ultimately the fundamental hereditary system of most organisms. He postulated two laws to explain his results.

Mendel's first law, or the *law of segregation*, describes the inheritance of a single trait, such as seed color. On the basis of the results of his experiments, Mendel postulated the existence of discrete factors that are transmitted from parent to offspring through the gametes (the egg or ovum, sperm, or pollen), with different factors determining the characteristics of different traits. These inherited factors are now called **genes**. Mendel also hypothesized that these factors occur in pairs in individuals, and that when a gamete is formed, only one of the two factors is included in the gamete. Furthermore, on the basis of his numerical results, he postulated that each of these two factors has an equal chance of being passed to an offspring, that is, the factors *segregate* during the production of gametes (meiosis), and that gametes thus have only one of each pair of factors. In addition, he proposed that these particulate factors persist in an unchanged state through successive generations of hereditary transmission. Conversely, when the gametes from the male and female parents fuse to form a zygote, the doubling of factors is restored.

Mendel's second law, or the *law of independence*, describes the joint behavior of loci controlling two different traits. The principle is simple: it states that the alleles at one locus segregate independently of the alleles of other loci. This proposition is now known to be strictly true only for loci that are on different chromosomes. Because the seven traits that Mendel used were, in fact, on different chromosomes, the traits Mendel used behaved in this fashion. Genes that are on the same chromosome tend to be inherited together (*see* **Linkage Analysis, Model-based**), in which case this principle will not necessarily hold.

One of the major differences between Mendel's two laws and many other quantitative models in the biological sciences is our understanding of the underlying processes leading to the observed data, and our resulting confidence in our ability to predict numerical outcomes. While Mendel did not yet know about chromosomes and the mechanism by which chromosomes are packaged into gametes (gametogenesis), this mechanism is now known to provide the basis behind the equal probability of transmission to a child of either member of a pair of genes. Gametogenesis involves the segregation of the two different members of a pair of chromosomes (carrying the two different genes for a particular trait) into two daughter cells, each with only one chromosome. Gametes are formed after one more round of duplication of these daughter cells, but the resulting ratio of gametes carrying each of the two original genes remains at one-to-one, resulting in the 50% probability in Mendel's first law that a particular gamete contains either of the two original genes. In addition, different pairs of chromosomes segregate independently in gametogenesis,

leading to the independence of inheritance of genes on different chromosomes.

To this day, Mendel's first law still forms the fundamental basis behind much of genetics. Although the law is elegant and simple, its consequences are not simple and can lead to complicated distributions of traits in populations and pedigrees. Mendel's second law is slightly less general since it is not applicable to genes that are on the same chromosome, or are at least not far apart on the same chromosome. However, because of the large size of the genome in most organisms, the second law usually adequately describes the behavior of the inheritance of multiple genes, although there are exceptions for linked genes (*see* **Linkage Analysis, Model-free**). This second law forms the basis of important assumptions made in the methodology behind many areas of genetic analysis (*see* **Polygenic Inheritance**; **Segregation Analysis, Classical**; **Twin Analysis**).

References

[1] Hartl, D.L. & Orel, V. (1992). What did Gregor Mendel think he discovered?, *Genetics* **131**, 245–253.

[2] Mendel, J.G. (1865). Verhandlungen des naturforschenden Vereines in Brünn, *Abhandlungen* **4**, 3–47.

[3] Mendel, J.G. (1966). (translation). *The Origins of Genetics: A Mendel Source Book*, C. Stern & E. Sherwood, eds. Freeman, San Francisco, pp. 1–48.

ELLEN M. WIJSMAN

Merrell, Margaret

Born: December 3, 1900, in LaGrange, Illinois.
Died: December 21, 1995, in Shelburne, New Hampshire.

Margaret Merrell is considered to be among the most admired of the early teachers of biostatistics. After graduating from Wellesley College in 1922, she taught mathematics at the Bryn Mawr School in Baltimore until 1925. She then entered the Johns Hopkins University School of Hygiene and Public Health as one of the first graduate students in biostatistics, at the same time beginning a career as member of the faculty in the department that would continue until she retired in 1959. She completed the requirements for a doctorate in biostatistics in 1930.

Merrell made significant contributions both as a developer of new methodology and as a consulting biostatistician for the conduct of infectious disease **clinical trials**. Her great legacy, however, was the standard she set for the **teaching of biostatistics** to health professionals. The Johns Hopkins University awarded her an honorary doctorate in 1981. She was a Fellow of the **American Statistical Association**, the American Association for the Advancement of Science, and Sigma Chi.

DENNIS O. DIXON

Meta-analysis in Epidemiology

Because of the pressure for timely and informed decisions in public health and clinical practice and because of the explosion of information in the scientific literature, research results must be synthesized to answer urgent questions [2, 25, 73]. Principles of evidence-based methods to assess the effectiveness of health care interventions and to set policy are cited increasingly [17]. Approaches to summarizing evidence include narrative reviews, systematic reviews and meta-analysis.

In general, randomized (controlled) clinical trials (RCTs) provide more useful evidence than do cohort studies, and cohort studies often provide better evidence than do case–control studies [137]. Cross-sectional studies and case series provide a weaker basis for etiologic reasoning. Because there are usually too few RCTs available to test clinical hypotheses, and because RCTs are rarely available to test etiologic hypotheses, particularly for chronic conditions, combining data from observational (cohort and case–control) studies is often necessary [10, 35, 131, 152, 165] (*see* **Case–Control Study**; **Cohort Study**).

Scientific Synthesis

In a traditional narrative review of the epidemiologic or medical literature, subject-matter experts review studies, decide which are relevant to the particular topic, and highlight their findings in terms of results and, to a lesser degree, methodology. The limitations of this or any approach to a literature review include: (a) biases in the original studies, reporting and publication policies; (b) absence in reported studies of specific data needed for the review; (c) investigator bias caused by subjective inclusion of studies; (d) uneven quality of the primary data; and (e) biased interpretation of outcome [156]. Such limitations have caused some authors to disregard the results of such reviews as having been prepared "with disregard for scientific principles" and therefore resulting in misleading decisions with serious consequences, often affecting health and quality of life [25, 111].

Systematic review methods have been adopted to address these problems. Systematic rules for conducting a synthesis include an explicit description of methodology so that results can be interpreted in light of biases and limitations [41]. Use of such systematic rules enables the investigator to refine large amounts of information, provide estimates of variables needed for economic and decision analysts, provide an efficient scientific technique, establish generalizability of scientific findings, assess consistency of relationships, explain data inconsistencies, increase statistical power and increase the precision of estimates [14, 25]. The techniques of meta-analysis use all of the steps of a systematic review, but in addition include a statistical combination of the results of previous studies to arrive at conclusions about a body of research (e.g. [150, 156]). Although systematic reviews in general, and meta-analyses specifically, are not immune to the potential pitfalls of a narrative review, the technique reduces the possibility of such errors and explicitly describes potential limitations (e.g. bias) of the results and interpretation [3, 112].

The statistical roots of systematic reviews can be found as early as the beginning of the twentieth century [127]. The term *meta-analysis* was first used in 1976 in the educational literature [69]: "the statistical analysis of a large collection of results from individual literature, for the purpose of integrating the findings". Since the method usually uses as "data" summary statistics derived from published reports of original studies, it is an *analysis of a statistical analysis* (thus, *meta*-analysis). Meta-analysis is most useful when individual study results are inconsistent and primary study sizes are small [97, 102, 128], since combining studies increases power. Meta-analysis is often recommended before undertaking a new study, to learn from earlier studies and to determine whether a new study will add substantially to what is already known about the topic [76, 95, 96].

While systematic reviews of evidence are usually desirable, meta-analysis (the statistical synthesis of results) should not be used indiscriminately [39, 42, 43, 62, 77]. In fact, legal suits have been brought by industry against researchers, charging "negligent misleading" when disparate findings are summarized by a single "class effect" [82]. The method is inappropriate if the number of studies is small or if there are large differences among the studies in study populations, interventions or effect measures [57, 92].

Approaches to summarizing data other than meta-analysis include vote counting and pooling [121,

122]. Vote counting relies only on the statistical significance of results [105], and may tend to indicate the wrong decision more often as the amount of evidence (number of studies) increases and does not incorporate characteristics of the original studies [80]; thus, we do not discuss that method here. The use of pooling, or combining original data, may be limited by the availability of data from primary authors [146]. When feasible, however, pooling original data offers definite advantages. Measures of exposure and outcome can be standardized, and adjustment for confounders can be done in a consistent manner across studies if the original data are available. Thus, with pooling, preliminary analyses yielding summary estimates of exposure effect for each study may be rendered more homogeneous than would be the case in more conventional meta-analyses. Once these study-specific estimates are obtained, standard meta-analytic techniques may be used to combine them. Lubin et al. [110] used pooling to combine data from cohort studies of underground miners to estimate the effect of radon exposure on risk of developing lung cancer.

Uses of Meta-analysis in Epidemiology

Meta-analyses were first used in clinical studies to combine results from RCTs [19, 27, 119, 133]. For example, a meta-analysis of 33 trials that compared treatment using intravenous streptokinase with a placebo in patients hospitalized for acute myocardial infarction showed a favorable effect of treatment, whereas only six of the 33 primary trials showed a statistically significant effect [100]. Continuously updated reviews, such as those provided by the Cochrane Database of Systematic Reviews [26, 32] facilitate proper conduct of meta-analyses of RCTs. Guidelines for publication of RCTs have been developed to encourage the inclusion of sufficient information in the published report to properly analyze the data using a meta-analysis [37, 115] and guidelines for the reporting of meta-analysis of RCTs have been developed [114].

Meta-analysis is being used increasingly to combine results from observational studies when randomized controlled designs are not available or not feasible [153]. Here, we define an observational study as an etiologic or effectiveness study using an analysis from an existing database, a cross-sectional study, a case series, a case–control design, a design with historical controls or a cohort design [71, 126]. Observational designs lack the experimental element of a random allocation to an intervention and rely on studies of association between changes or differences in one characteristic (e.g. an exposure or intervention), and changes or differences in an outcome of interest. These designs have long been recommended and used in the evaluation of educational programs [40] and of exposures that might cause disease [92]. For example, studies of risk factors generally cannot be randomized, because they relate to inherent human characteristics or practices, or because such a randomization might be unethical [106, 126]. At times, clinical data on treatments may be summarized in order to design a randomized treatment comparison [89, 161]. Observational data may also be needed to assess how well an intervention works in a community as opposed to the special setting of a controlled trial [111].

Meta-analyses of observational studies present particular challenges because of inherent biases and differences in study designs [65, 93]. Nonrandomized comparisons are subject to both selection bias and other types of confounding, and combining several studies all subject to the same bias will only reinforce that bias [144] (*see* **Bias in Case–Control Studies**; **Bias in Cohort Studies**). In addition, observational studies may lack some of the elements of a well-designed clinical trial, such as careful definition of endpoints, interventions and study population [119]. Also, observational studies may have different exposed populations and control groups, may suffer from measurement error of exposures, and may explore only varying outcome measures. Because such factors may influence various studies differently, a single summary measure for exposure effect may be misleading. A more important use of meta-analysis of observational studies may well be as a tool for understanding and quantifying sources of heterogeneity in results across studies [113, 120].

Although meta-analysis of observational studies may not always be appropriate (e.g. [74, 139]) – particularly if the goal is to produce a single summary estimate of an association [101] – the number of published meta-analyses concerning health issues has increased substantially during the past four decades: from 678 before 1992; to 525 from 1992 through 1995; to approximately 400 in 1996 alone. Furthermore, a 1997 study of published

meta-analyses documented that 86% of authors were the first or second author of only a single meta-analysis before the one used in the study [140], indicating the broadening use of this method.

Steps in a Meta-analysis

The basic steps in a meta-analysis include: (a) a clear statement of the problem and hypothesis to be tested; (b) a clearly defined statement of inclusion and exclusion criteria for admission of studies; (c) a methodology for locating research studies; (d) the classification and coding of study characteristics to be combined in the meta-analysis and a quantitative measurement of study characteristics and of the effect of the exposure on outcome; (e) an assessment of the quality of the methods used in the studies; (f) a statistical analysis that includes methods for combining study results when appropriate and determining the sources of heterogeneity of the data; and (g) interpretation of results, including an assessment of bias of individual studies, a discussion of heterogeneity, and identification of areas for further research [128, 156].

Statement of the Problem and Hypothesis to be Tested

Problem formulation is critical and includes the explicit definition of both outcomes and potential confounding variables. This step enables the investigator to abstract accurate and consistent data from reports of studies and to choose appropriate statistical models for the analysis. This step is especially critical for meta-analyses of observational studies. For example, suppose the major task is the exploration of evidence for a theory: Does a high level of homocysteine contribute to increased risk for cardiovascular disease [20]? Such a statement of the problem allows exclusion of studies using patients with competing conditions, which may preclude a clear evaluation of outcome.

As for any study, the protocol is the blueprint for the conduct of the meta-analysis. The protocol should contain a clear statement of the problem, objectives, hypotheses to be tested, background and specifications for information retrieval, data collection and analysis.

Establishing Inclusion and Exclusion Criteria

As in any statistical study, sample design is an important determinant of the utility and scientific validity of results. In a meta-analysis, the sampling units are the results of published or unpublished studies. The study inclusion/exclusion criteria provide the "case definition" for results to be used in the synthesis. Objective exclusion criteria should be determined a priori to meet scientific criteria and not as a matter of convenience. For example, a decision to exclude studies published before a specific date should be based on evidence that a technology or therapy changed at that time (and therefore historical results may not be comparable with more recent results), and not on the fact that earlier studies may not be cataloged electronically. "Fugitive literature" refers to studies that may be published in documents that are difficult to locate, because they are not published, are published but not abstracted, or have limited circulation (e.g. dissertations, conference abstracts and proceedings or government reports). If studies in the "fugitive literature" are to be excluded, a rationale should be given. Similarly, a decision to exclude foreign language studies should follow a determination that studies published in a language other than English are different in some substantive way, and not be made merely on the basis of a lack of translation capability [60, 75, 116]. Although some control over heterogeneity of design may be accomplished through the use of exclusion rules, a more informative approach could be to use broad inclusion criteria for studies, and then to perform analyses to determine whether measured design features influence measurements of exposure effect on outcome [11].

Locating Research Studies

The goal of the search process is to (a) identify all relevant primary studies (published and unpublished) for potential inclusion in the meta-analysis; and (b) to determine which studies are to be included. Accurate and thorough specifications of the search strategy will allow for the replication of the meta-analysis and permit others to evaluate the external validity of the findings. Examining the search strategy in depth can help to explain different conclusions from different meta-analyses on the same topic [157]. The term "search" implies the entire process, which is usually composed of several different methods of searching. A 1996 review of 103 meta-analyses in education documented that search procedures were described inadequately, fewer than half of the meta-analyses

reported details of classifying and coding the primary study data and only 22% assessed quality of the primary studies [140].

The literature search should be systematic and comprehensive. The researcher uses several sources of information to locate data for retrieval, including written indices; computerized searches; bibliographies of published papers; and unreferenced and sometimes unpublished data from academic, private and governmental researchers [23, 38]. Complete searches should go beyond computerized indices, which have been shown to have a sensitivity as low at 50% in some examples [52]. Researchers estimate that 25%–50% of all initiated randomized controlled trials are never published, and excluding data from unpublished studies may result in bias or loss of precision in estimation of effect size [51].

Several computerized databases exist, many operated by the National Library of Medicine and included in the Medical Library Information Retrieval System (MEDLARS). These include AIDSLINE (for AIDS-related citations, 1980 to the present), CANCERLIT (containing cancer literature from journals, government reports and conferences, 1963 to the present), TOXLINE (containing citations on the effects of drugs and other chemicals), and Dissertation Abstracts (containing abstracts of American and Canadian doctoral dissertations, 1861 to the present). Registries of randomized trials [32, 55] provide information prior to publication of RCTs for topics of interest to the collaborators. For other study designs, or for situations in which registries do not exist, contact with experts in the field may yield more complete ascertainment. It is important to identify and remove redundant reports for the same study [90]. When computerized indices are used, the search strategy should be specified completely to allow replication. The description should include keywords used, fields searched (e.g. whether the search was by text word, title or subject), software used for searching (e.g. OVID), and any software-specific functions (e.g. "explosion" of terms).

Classification, Coding and Measurement of Study Characteristics and Measurement of Exposure Effect on Outcome

The classification and coding of study characteristics follows directly from problem formulation [156]. This step can consume the majority of time invested in a meta-analysis – approximately 90% [86, p. 85]. In addition to increasing the time required for a synthesis, adding characteristics of the studies included increases the probability of finding at least one chance association as significant. Thus, many meta-analysts recommend coding a study characteristic only when theoretical justification exists [149]. This recommendation is controversial, however, since additional information about study characteristics can provide documentation for findings, can assist analysis of sources of heterogeneity and can provide areas for additional research. Furthermore, the requirement for formal theoretical justification can restrict creative hypothesis generation.

Blinding (masking) readers to identifying information about papers (e.g. author's affiliation) has been advocated. In a study of blinding, five meta-analyses of RCTs were conducted in parallel by two groups randomly assigned to read papers that either had or had not been masked as to the identity of the authors and institutions producing the original papers, and as to which treatment group was which. Although the unmasked readers assigned higher quality scores on average than the masked readers, masking made little difference in the summary odds ratios [7, 16]. We are unaware of any published studies of the effects of blinding on meta-analyses of observational data.

Assessment of Quality of Included Studies

The use of quality scoring in meta-analysis is controversial [57, 61]. One potential use of quality scores is to assign greater weight to some studies than others when combining results. A second use is for grouping studies according to quality to determine whether estimates of exposure effect depend on study quality [67, 77]. Quality scores constructed in an *ad hoc* fashion, however, may lack demonstrated validity. Furthermore, examples indicate that estimates of exposure effect are not always associated with quality [74]. Nevertheless, some *particular* aspects of study quality, such as adherence to the randomization scheme in RCTs, have been shown to be associated with effect size [13, 118, 138].

Statistical Methods for Combining Study Results and Analyzing Heterogeneity

Statistical issues related to combining data from multiple sources in meta-analysis are the subject of ongoing research [61, 104]. The most important statistical

issues concern which studies should be combined. When feasible, meta-analysis should be restricted to RCTs, the study design that provides most useful evidence. When too few RCTs are available, combining data from observational (cohort and case–control) studies is necessary.

Beyond the determination of studies to be combined, a central question is whether variations in research studies (methodologic or contextual) are related to variations in effect size. This question can be approached by the use of *fixed-* or *random-effects models*. The simplest fixed-effects models assume that the exposure effect is constant across studies and that variation from one study to the next is due solely to within-study random variation. More elaborate fixed-effects models may allow the outcome to depend on several fixed effects, corresponding to several variables that characterize the studies [79]. Random-effects models allow the intrinsic exposure effect to vary from study to study so that variation in the estimated exposure effects reflects both sampling error within studies and effect variation across studies [81]. Random-effects models assume that the studies included are selected randomly from a population of studies with varying exposure effects and are used to accommodate unexplained heterogeneity of exposure effects [87]. Random-effects models increase the estimated variance around estimates of associations and produce different point estimates than fixed-effects models. It should be emphasized, however, that using random-effects models to account for unexplained heterogeneity should not substitute for a thorough exploratory (fixed-effects) analysis of how study design and population characteristics affect estimates of exposure effects. If random-effects models are used, then the rationale for model selection should be given, and estimates of among-study variation should be reported [53].

Epidemiologic studies undertaken to establish causal explanations of disease–exposure association frequently estimate risks at different levels of exposure [84]. Such dose–response studies yield study-specific slopes that may be combined using meta-analytic techniques [141, 160]. One meta-analytic method for estimating a combined dose–response effect from case–control and cohort studies incorporates the same dose–response model in each component of the likelihood, which is the product of the study-specific likelihoods [15]. Brumback [22] used the estimation–minimization (EM) algorithm (termed the "method of weights") to maximize this likelihood; the calculations use standard weighted regression software.

Regardless of the statistical measure used to combine data, a meta-analysis to combine results across studies should include: (a) presenting the study-specific exposure estimates with estimates of study-specific random error; (b) presenting summary estimates of exposure effect across studies, with estimates of variability; (c) testing for heterogeneity of exposure effects across studies, and, if present, investigating possible causes of heterogeneity; and (d) providing quantitative support for interpreting the results.

Interpretation of Results

The interpretation should focus on a discussion of the strengths and weaknesses of the evidence, including possible biases, and on the justification for combining estimates in the presence of potential heterogeneity of study results. Bias has been defined as any systematic error that leads to the distortion of accurate results (e.g. [128]). In the original studies, bias can result from flaws in the study design that tend to distort the magnitude or direction of associations in the data. In meta-analyses, additional bias can result from the way in which studies are selected for inclusion and from the way in which data are gathered and analyzed (*see also* **Bias, Overview**).

Assessment of Bias in Individual Studies. One approach to assessing the research quality of observational studies is based on "threats to validity" [40]. Thirty-three independent threats to validity are categorized into four groups – internal, external, statistical conclusion and construct validity. Internal validity is "the truthfulness with which statements can be made about whether there is a causal relationship from one variable to another in the form in which the variables were manipulated or measured" [40, p. 38]. External validity reflects the extent to which the relation can be generalized across other populations. Statistical conclusion validity refers to the quality of the statistical analysis and inference. Construct validity concerns threats that may confound cause and effect measurement.

In general, quality assessment indicates that not all studies retrieved should be included in the meta-analysis. Construct and external validity can be used

to determine whether a study addresses the hypothesis of interest, participants, time period and location; i.e. the relevance of the study to the meta-analysis [135]. Construct validity can also be used in meta-analyses assessing theories [109]. Studies with fewest threats to internal validity should be considered highest quality [167]. Improper statistical techniques may also render a study unacceptable for inclusion (e.g. inappropriate statistical tests, incorrect grouping of values, inappropriate conclusions or the absence of information needed to calculate effect estimates or variability). Other scientific methods may also yield problems for quantitative synthesis. Toward this end, guidelines have been developed for assessing the methods in data on gene-disease associations [108].

Publication Bias. One reason this occurs is that authors are less likely to submit manuscripts reporting negative results to journals. There is no evidence that publication bias occurs once manuscripts have been submitted to a medical journal [124]. Publication bias, i.e. the selective publication of studies on the basis of the magnitude and direction of their findings, represents a particular threat to the validity of any meta-analysis [51, 56, 132]. Statistical methods assist in the assessment of publication bias and in correcting for this problem [5, 6, 46]. Methods for detecting publication bias include correlating the observed exposure effect size with design features of the studies that might be "risk factors" for publication bias (such as sample size, presence or absence of randomization and prospective vs. retrospective design) [12, 50, 70, 78]. For example, for a meta-analysis using a mixture of randomized and nonrandomized studies, if randomization status appears to be associated with size of risk estimates, then one might eliminate (or analyze separately) the nonrandomized studies [134].

The effect of sample size on publication bias can be assessed graphically by a "funnel plot" of sample size vs. effect size [104]. In the absence of publication bias caused by sample size, the plot should appear as an inverted funnel; that is, large variability will be shown with small studies and decreasing spread as the sample size increases, with constant mean effect size regardless of sample size. If this shape is not apparent, then publication bias should be suspected. For example, if large studies are clustered around the null value with smaller studies skewed around a positive effect, then one suspects that some small negative studies were not included, revealing publication bias [93].

The use of the funnel plot should not be substituted from a careful examination of the literature search [142]. Formal statistical significance tests can be used to determine whether estimates of intervention effects are correlated with sample size [8]. A rank correlation based on Kendall's tau [1] requires no underlying assumptions, but may lack power. Alternatively, a test based on Spearman's rho is more tractable computationally [36]. A formal sensitivity analysis may be a more robust approach to assessing publication bias [44].

Statistical methods for correcting for publication bias include sampling methods and analytic methods. Sampling methods are based on the following logic: if publication bias is caused by preferential publication based on study results, then this problem can be prevented by restricting the search to a sampling frame that cannot be influenced by study results, such as registries of prospective trials [55]. There are few registries of observational studies [30]. Attempting to locate all relevant observational studies through contact with experts and use of sources of unpublished reports to supplement standard computerized searches remains an option [45].

Analytic methods for addressing publication bias include the "file drawer" method [132], which addresses the question: If a combined estimate indicates a statistically significant exposure effect, then how many studies with null results must exist somewhere (the file drawer) to overturn the results? Hedges & Olkin [81] answer a similar question for methods that combine evidence from the individual studies' significance levels, rather than methods that obtain a combined estimate of exposure effect. Some methods to correct for publication bias are based on the assumption that a study is included in the meta-analysis with probability proportional to the estimated effect size [29, 125].

Heterogeneity. Heterogeneity of populations (e.g. study subjects), study designs (e.g. case–control vs. cohort studies), methods for measuring exposure and outcomes, analytical approaches to confounding and other issues must be recognized, reported, and, when possible, addressed in the analysis of the data by exploring associations between variation in study design and analysis and variation in estimated exposure effects [11]. In cases where heterogeneity of

exposure effects is large, a single summary measure of exposure effect may be inappropriate. Analyses that stratify by study feature or regression analysis with design features as predictors can be useful in assessing whether these features influence estimates of exposure effect [158]. Studies should never be discarded solely on the basis of having results that disagree with those of the majority of other studies [34], but rather should be examined for underlying characteristics of design or analysis that may have led to the discrepant results.

Statistical tests for heterogeneity include DerSimonian & Laird's Q-test [31, 47], an approach based on weighted least squares [107], an application of the likelihood ratio test [151], and measures developed by Higgins and Thompson, from mathematical criteria, that are independent of the number of studies and the treatment effect metric [83] and methods based on a bootstrap approach [155]. The power of these tests to detect heterogeneity is often small, however, because the number of studies (the effective sample size) is typically less than 30 [76]. Empirical evaluation shows that the DerSimonian and Laird Q-statistic is preferable from the point of view of validity, power and computational ease [155]. Graphical displays, stratification and regression analysis are useful methods to address criticisms of summarization in the presence of heterogeneity [74]. In the absence of absence of an a priori hypothesis about the source of heterogeneity, classic methods such as regression may prove difficult. In such cases, a graphical method to identify sources of heterogeneity may be preferable [4]. For the specific case of combining comparative trials with uncontrolled historical studies, Begg & Pilote have proposed a method using a random effects approach [9]. Investigating heterogeneity was a key feature of a meta-analysis of asbestos exposure and risk of gastrointestinal cancer [68]. This example shows that sources of bias and heterogeneity can be hypothesized before analysis and subsequently confirmed by the analysis.

Finally, sensitivity analysis can permit exploration of sources of heterogeneity and can suggest future research directions. Sensitivity analysis can be used in each step of a meta-analysis. During the search and citation retrieval step, use of more than one investigator is helpful when the research question spans disciplines and requires different subject-matter expertise. Exploratory data analysis [72] can be used to investigate features of articles retrieved in the search and reveal factors that may have impact on the choice of more formal statistical procedures. Combining P-values and regression analysis [81, Chapter 12] aids in sensitivity analysis.

Schlesselman [136] used the possible association between endometrial cancer and oral contraceptives to comment on issues related to potential bias in the studies of this association. His meta-analysis combined both cohort and case–control studies and used a sensitivity analysis to illustrate the effect of omitting specific studies. He addressed possible bias caused by restriction to English language articles by performing analyses limited to such studies.

In summary, interpretation should include assessing the internal validity of component studies; namely, whether they were well-designed, executed and analyzed. Discussion of whether the studies included are appropriate for answering the meta-analytic question should include efforts taken to avoid publication bias. For example, do funnel plots support the contention that publication bias has been minimized? If not, how much bias can be anticipated? To what extent should heterogeneity of effect estimates be related either to systematic features of various studies (such as sample size, type of study or study quality) or to nonidentifiable random variation? In view of any heterogeneity of results, is it reasonable to summarize the results in a single measure of exposure effect (with an estimated confidence interval), or rather to state that a single summary is not appropriate and that further research is needed to define the sources of variation and extent of the association?

A Case Study

A 1991 study of the effects of duration of estrogen use on breast cancer risk [148] illustrates many of the decisions made in performing a meta-analysis in epidemiology. Published reports agreed on the risk associated with ever-use of estrogen replacement therapy, but little evidence was shown for increases in risk due to short-term use (less than five years), and there was less agreement on the effect of long-term use. The authors located case–control and cohort studies; among the case–control studies, designs were heterogeneous in choice of controls (e.g. hospital or community). Heterogeneity testing

was used to determine criteria for subgroup analysis. Estimated dose–response slopes from primary studies were used for the meta-analysis to account for inter-study variability. Fixed- and random-effects models were computed, and assumptions such as equal baseline risk were evaluated by sensitivity testing. Because tests of homogeneity were significant, the authors analyzed data from studies that used community controls separately from those that used hospital controls [147].

Increased risk of breast cancer with duration of estrogen use was found among studies with community controls (risk of breast cancer after 10 years of estrogen use increased by at least 15%); studies with hospital controls showed a similar increase when a single outlier study conducted in Europe was excluded. Differences in results of fixed- and random-effects models may have been due to this source of variation. Other sources of heterogeneity explored included study design, location of study population (US vs. Europe) indicating differences in estrogen preparation; the two European studies using hospital controls showed an increased risk, while studies in the US showed a small decrease in risk with duration of estrogen use.

This example indicates the role of heterogeneity of study design in the interpretation of results of a meta-analysis. In this case, many health conditions are associated with estrogen use and a woman's decision to use estrogen. Thus, the choice of controls may have been a critical determinant of heterogeneity. For example, greater use of estrogen among women who receive acute care in hospitals might explain the apparent decrease in breast cancer risk shown by studies that used hospital controls.

Discussion

Taking stock of what is known in epidemiology involves reviewing the existing literature, summarizing it in appropriate ways, exploring the implications of heterogeneity of study designs, and determining how heterogeneity of design might relate to heterogeneity of study results. Meta-analysis provides a systematic way of performing this research synthesis, while indicating when more research is necessary and is a widely used and increasingly popular technique. Nevertheless, some criticisms and caveats are important for using the results of meta-analyses of observational studies [54, 64, 66, 143].

Criticisms of Meta-analysis in Epidemiology

The use of meta-analysis in epidemiology is not universally accepted due to several limitations [162]. First, bias can occur in the original studies (resulting from flaws in the study design that tend to distort the magnitude or direction of associations in the data), or from the way in which studies are selected for inclusion [18, 63]. Methods have been developed to aid in the detection of publication bias, a particular threat to the validity of meta-analysis of observational studies [56, 132]. In addition, funding source can be an important source of bias affecting results [94].

Secondly, when combining observational studies, heterogeneity of populations (e.g. US vs. international), design (e.g. case–control vs. cohort studies), and outcome (e.g. different studies yielding different relative risks that cannot be accounted for by sampling variation) is expected [11, 159]. In cases where heterogeneity of outcomes is particularly problematic, a single summary measure may be inappropriate. Analyses that stratify by study feature or regression analysis with design features as predictors can be useful in assessing whether study outcomes indeed vary systematically with these features [34, 154].

Thirdly, the use of quality scoring in meta-analysis is controversial [61] because scores constructed in an *ad hoc* fashion may lack demonstrated validity, and results may not be associated with quality [21].

Fourthly, a statistical summary of evidence may be misused to obscure important variations in exposure effects [57]. Meta-analyses have been criticized because of discrepancies between their results and those from large randomized trials [24, 28, 48, 91, 103, 130].

Extensions and Related Areas

Cumulative Meta-Analysis. The concept of the cumulative meta-analysis was introduced in 1993 [27]. In a cumulative meta-analysis, studies are combined and data synthesized on an ongoing basis. As soon as a study relevant to the particular topic is published, its results are entered into the meta-analysis and estimates of effect or relative risk are then updated. Adaptation of methods from interim analyses of clinical trials has been used in this situation [58, 129, 164]. This updating enables the most current estimation of the effect of a particular intervention or particular risk factor. Retrospective analysis of clinical trials for myocardial infarction indicates

that a cumulative meta-analysis of the effects of thrombolytic therapy and lidocaine on cardiovascular mortality could have changed clinical practice as much as 10–15 years earlier had such analyses been conducted, published and disseminated adequately [100]. Alternatively, one may perform the cumulative analysis by adding sequentially studies with increasing quality score or increasing study size to investigate the effects of these factors. Attention should be paid to the results of repeated testing by adjusting significance levels.

Cochrane Collaboration, investigators of a particular subject-matter area, aggregate data on an ongoing basis, conduct cumulative meta-analyses and make the results available to clinical researchers and others interested in clinical and public health policy [18, 33, 166].

Statistical Software. Egger et al. [59] attribute part of the growth in the number of meta-analyses to the recent appearance of software packages that implement the methods. Available packages vary in their provision of tutorials, graphical features and flexibility of modeling. Almost all packages include the capability for fixed- and random-effects modeling, tests of homogeneity and ability to handle multiple types of response. Some, such as RevMan© produced by the Cochrane Collaboration, are available via the Internet. For many applications (e.g. dose–response analyses, or meta-regressions), one must still use preliminary programs to estimate inputs, such as slopes, for use in meta-analytic programs.

Conclusions: The Role of Meta-analysis in Epidemiology

Regardless of the problems and technical solutions, an analytic synthesis of evidence is critical for policy in epidemiology and public health [163]. When policy-makers attempted to study the effect of passive smoking in public places (inhalation of others' smoke), tobacco-industry lobbyists presented competent studies showing little evidence of harm, and antismoking activists presented studies that showed passive smoking to be a cause of lung cancer [85].

The systematic evaluation of any health topic is essential to excellent clinical or public health practice [98]. The Guide to Community Preventive Services tries to determine what works in community health using systematic reviews on a variety of health topics important to communities, public health agencies, and health care systems. Systematic reviews evaluate the evidence of effectiveness, which is then translated into a recommendation or a finding of insufficient evidence. For those interventions where there is insufficient evidence of effectiveness, this process provides guidance for further prevention research [168]. The results of such evaluations also help define priorities in research. The conduct of meta-analyses, therefore, warrants rigorous implementation. The introduction of systematic reviews and meta-analyses has fostered controversy [49], but it has also initiated a critical examination of the process of research synthesis. This process must continue with careful consideration given to both epidemiologic and statistical issues and appropriate examination of the impact on health and quality of life. In addition to the specific methodologic issues mentioned above, more effort must continue in the design and implementation of primary studies and the reporting of such studies, including methods as well as results [7, 42, 99, 117, 123, 145]. Efforts such as the Cochrane Collaboration have demonstrated the value of careful collection and storage of information in readily accessible computer data banks. The development of computer software both for data manipulation and transport, as well as statistical analysis, will continue to reap rewards for researchers and practitioners. Improved accessibility of the results of meta-analyses is another potential benefit of the computerization of data. At the same time, researchers and practitioners must be trained to understand the benefits and limitations of alternative methods of data synthesis, as should those who make public policy decisions and influence medical care and public health practice. Statisticians have a critical role in developing methods for special problems and assisting in the design, execution and analysis of these studies.

References

[1] Armitage, P. & Berry, G. (1987). *Statistical Methods in Medical Research*, 2nd Ed. Blackwell, Oxford.

[2] Badgett, R.G., O'Keefe, M. & Henderson, M.C. (1997). Using systematic reviews in clinical education, *Annals of Internal Medicine* **126**, 886–891.

[3] Bailar, J.C. (1997). The promise and problems of meta-analysis, *New England Journal of Medicine* **337**, 559–560.

[4] Baujat, B., Mahe, C., Pignon, JP. & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Statistics in Medicine*. **21**(18), 2641–2652.

[5] Begg, C. (1994). Publication bias, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds. Russel Sage Foundation, New York, pp. 399–409.

[6] Begg, C.B. & Berlin, J.A. (1989). Publication bias and dissemination of clinical research, *Journal of the National Cancer Institute* **81**, 107–115.

[7] Begg, C.B., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. & Stroup, D.F. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement, *Journal of the American Medical Association* **276**, 637–639.

[8] Begg, C.B. & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias, *Biometrics*, 1088–1101.

[9] Begg, C.B. & Pilote, L. (1991). A model for incorporating historical controls into a meta-analysis, *Biometrics* **47**, 899–906.

[10] Beral, V. (1995). The practice of meta-analysis: discussion. Meta-analysis of observational studies: a case study of work in progress, *Journal of Clinical Epidemiology* **48**, 165–166.

[11] Berlin, J.A. (1995). Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies, *American Journal of Epidemiology* **142**, 383–387.

[12] Berlin, J.A., Begg, C.B. & Louis, T.N. (1989). An assessment of publication bias using a sample of published clinical trials, *Journal of the American Statistical Association* **84**, 381–392.

[13] Berlin, J.A. & Colditz, G.A. (1999). The role of meta-analysis in the regulatory process for foods, drugs, and devices, *Journal of the American Medical Association* **281**, 830–834.

[14] Berlin, J.A., Longnecker, M.P. & Greenland, S. (1993). Meta-analysis of epidemiologic dose–response data, *Epidemiology* **4**, 218–228.

[15] Berlin, J.A., Miles, C.G. & Cirigliano, M.D. (1997). Does blinding of readers affect the results of meta-analyses? Results of a randomized trial, *Journal of Current Clinical Trials*, `doc no. 205`, online.

[16] Berlin, J.A. & University of Pennsylvania Meta-analysis Blinding Study Group (1997). Does blinding of readers affect the results of meta-analyses? *Lancet* **350**, 185–186.

[17] Bero, L.A. & Jadad, A.R. (1997). How consumers and policymakers can use systematic reviews for decision making, *Annals of Internal Medicine* **127**, 37–42.

[18] Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflub, T. & Friedenreich, C. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology, *International Journal of Epidemiology* **28**, 1–9.

[19] Boissel, J.P., Blanchard, J., Panak, E., Peyrieus, J.C. & Sacks, H. (1989). Considerations for the meta-analysis of randomized clinical trials, *Controlled Clinical Trials* **10**, 254–281.

[20] Boushey, C.J., Beresford, S.A., Omenn, G.S. & Motulsky, A.G. (1995). A quantitative assessment of plasma homocysteine as a risk factor for vascular disease: probable benefits of increasing folic acid intakes, *Journal of the American Medical Association* **274**, 1049–1057.

[21] Breslow, R.A., Ross, S.A. & Weed, D.L. (1998). Quality of reviews in epidemiology, *American Journal of Public Health* **88**, 475–477.

[22] Brumback, B.A., Holmes, L.B. & Ryan, L.M. (1999). Adverse effects of chorionic villus sampling: a meta-analysis, *Statistics in Medicine* **18**, 2163–2175.

[23] Byars, D.P. (1988). The use of data bases and historical controls in treatment comparison, in *On Combining Information: Historical Controls, Overviews, and Comprehensive Cohort Studies*. *Recent Results in Cancer Research*, Vol. 111, Springer-Verlag, New York.

[24] Cappelleri, J., Ioannidis, J., Schmid, C., de Ferranti, S., Aubert, M., Chalmers, T.C. & Lau, J. (1996). Large trials vs. meta-analysis of smaller trials, *Journal of the American Medical Association* **276**, 1332–1338.

[25] Chalmers, I. & Altman, D.G., eds. (1995). *Systematic Reviews*. British Medical Journal Publishing Groups, London.

[26] Chalmers, I., Dickersin, K. & Chalmers, T.C. (1992). Getting to grips with Archie Cochrane's agenda, *British Medical Journal* **304**, 768–786.

[27] Chalmers, T.C. & Lau, J. (1993). Meta-analytic stimulus for changes in clinical trials, *Statistical Methods in Medical Research* **2**, 161–172.

[28] Chalmers, T.C., Levin, H., Sacks, H.S., Reitman, D., Berrier, J. & Nagalingam, R. (1987). Meta-analysis of clinical trials as a scientific discipline I: Control of bias and comparison with large co-operative trials, *Statistics in Medicine* **6**, 315–325.

[29] Chalmers, T.C., Smith, H., Jr, Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial, *Controlled Clinical Trials* **2**, 31–49.

[30] Chollar, S. (1998). A registry for clinical trials, *Annals of Internal Medicine* **128**, 701–702.

[31] Cochran, W.G. (1954). The combination of estimates from different experiments, *Biometrics* **10**, 101–129.

[32] Cochrane Collaboration. Handbook. INTERNET: `http://www.cochrane.co.uk`.

[33] Colditz, G.A., Brewer, T.F., Berkey, C.S., Wilson, M.E., Burdick, E., Fineberg, H.V. & Mosteller, F. (1994). Efficacy of BCG vaccine in the prevention of tuberculosis: meta analysis of the published literature, *Journal of the American Medical Association* **271**, 696–702.

[34] Colditz, G.A., Burdick, E. & Mosteller, F. (1995). Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary, *American Journal of Epidemiology* **142**, 371–382.

[35] Cole, M.G. & Bellavance, F. (1997). Depression in elderly medical inpatients: a meta-analysis of outcomes, *Canadian Medical Association Journal* **157**, 1055–1060.

[36] Colton, T. (1974). *Statistics in Medicine*. Little, Brown, Boston, Mass.

[37] The CONSORT Working Group. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* **276**(8), 637–639.

[38] Cook, D.J., Guyatt, G.H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., Mcllroy, W. & Oxman, A.D. (1993). Should unpublished data be included in meta-analyses? Current conflictions and controversies, *Journal of the American Medical Association* **21**, 2749–2753.

[39] Cook, D.J. & Mulrow, C.D. (1997). Systematic reviews: synthesis of best evidence for clinical decisions, *Annals of Internal Medicine* **126**, 376–380.

[40] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton-Mifflin, Boston, Mass.

[41] Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A. & Mosteller, F. (1992). *Meta-Analysis for Explanation: A Casebook*. Russel Sage Foundation, New York.

[42] Cook, T.D. & Leviton, L.C. (1980). Reviewing the literature: a comparison of traditional methods with meta-analysis, *Journal of Personality* **48**, 449–472.

[43] Cooper, H. & Hedges, L.V., eds. (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York.

[44] Copas, J.B. & Shi, J.Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research.* **10**(4), 251–265.

[45] Counsell, C. (1997). Formulating questions and locating primary studies for inclusion in systematic reviews, *Annals of Internal Medicine* **127**, 380–387.

[46] Dear, K.B. & Begg, C.B. (1992). An approach to assessing publication bias prior to performing a meta analysis, *Statistical Science* **7**, 237–245.

[47] DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials, *Controlled Clinical Trials* **7**, 177–188.

[48] DerSimonian, R. & Levine, R.J. (1986). Resolving discrepancies between meta-analysis and a subsequent large controlled trial, *Journal of the American Medical Association* **282**, 664–670.

[49] Dickersin, K. & Berlin, J.A. (1992). Meta-analysis: state of the science, *Epidemiologic Reviews* **14**, 54–76.

[50] Dickersin, K., Chan, S., Chalmers, T.C., Sacks, H.S. & Smith, H., Jr (1987). Publication bias and clinical trials, *Controlled Clinical Trials* **8**, 343–353.

[51] Dickersin, K. & Min, Y.I. (1993). NIH clinical trials and publication bias, *Online Journal of Current Clinical Trials*, `doc. no. 50`, online

[52] Dickersin, K., Scherer, R. & Lefebre, C. (1995). Identifying relevant studies for systematic reviews, in *Systematic Reviews*, I. Chalmers & D.G. Altman, eds. British Medical Journal Publishing Group, London.

[53] DuMouchel, W. (1994). *Hierarchical Bayes Linear Models for Meta-analysis*. National Institute of Statistical Sciences, Research Triangle Park, North Carolina.

[54] Dyer, A.R. (1986). A method for combining results from several prospective epidemiologic studies, *Statistics in Medicine* **5**, 303–317.

[55] Easterbrook, P.J. (1992). Directory of registries of clinical trials, *Statistics in Medicine* **11**, 345–423.

[56] Easterbrook, P.J., Berlin, J.A., Gopalan, R. & Matthews, D.R. (1991). Publication bias in clinical research, *Lancet* **337**, 867–872.

[57] Egger, M., Scheider, M. & Davey-Smith, G. (1998). Meta-analysis: spurious precision? Meta-analysis of observational studies, *British Medical Journal* **316**, 140–144.

[58] Egger, M., Smith, G.D. & Sterne, J.A. (1998). Meta-analysis: is moving the goal post the answer? *Lancet* **351**, 1517.

[59] Egger, M. & Sterne, J.A. (1998). Software for meta-analysis, *British Medical Journal* **316**, online.

[60] Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C. & Antes, G. (1997). Language bias in randomised controlled trials published in English and German, *Lancet* **350**, 326–329.

[61] Emerson, J.D., Burdick, E., Hoaglin, D.C., Mosteller, F. & Chalmers, T.C. (1990). An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials, *Controlled Clinical Trials* **11**, 339–352.

[62] Eysenck, H.J. (1984). Meta-analysis: an abuse of research integration, *Journal of Special Education* **18**, 41–59.

[63] Felson, D.T. (1992). Bias in meta-analytic research, *Journal of Clinical Epidemiology* **45**, 885–892.

[64] Fleiss, J.L. & Gross, A.J. (1999). Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique, *Journal of Clinical Epidemiology* **44**, 127–139.

[65] Ford, E.S., Smith, S.J., Stroup, D.F., Steinberg, K.K., Mueller, P.W., Thacker, S.B. & Weitman, E.A. (2002). Homocyst(e)ine and cardiovascular disease: a review of the evidence with special emphasis on case–control studies and nested case–control studies. *International Journal of Epidemiology* **31**(1), 59–70.

[66] Friedenreich, C. (1993). Methods for pooled analyses of epidemiologic studies, *Epidemiology* **4**, 295–302.

[67] Friedenreich, C. (1994). Influence of methodologic factors in a pooled analysis of 13 case–control studies of colorectal cancer and dietary fiber, *Epidemiology* **5**, 66–67.

[68] Frumkin, H. & Berlin, J. (1988). Asbestos exposure and gastrointestinal malignancy review and meta-analysis [published erratum appears in *American Journal of Industrial Medicine* 1988, 14(4), 493], *American Journal of Industrial Medicine* **14**, 79–95.

[69] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Researcher* **5**, 3–8.

[70] Gleser, L.J. & Olkin, I. (1996). Models for estimating the number of unpublished studies, *Statistics in Medicine* **15**, 2493–2507.

[71] Green, S.B. & Byars, D.P. (1984). Using observational data from registries to compare treatments: the fallacy of omnimetrics, *Statistics in Medicine* **3**, 361–370.

[72] Greenhouse, J.B. & Iyengar, S. (1994). Sensitivity analysis and diagnostics, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds. Russell Sage Foundation, New York, pp. 383–398.

[73] Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature, *Epidemiological Reviews* **9**, 1–30.

[74] Greenland, S. (1994). Invited commentary: a critical look at some popular meta-analytic methods (Comment in *Am J Epidemiol*, 1994, 1 Aug, **140**(3), 297–299; discussion 300–301, *Am J Epidemiol*, 1995, 1 Nov, **142**(9), 1007–1009), *American Journal of Epidemiology* **140**, 290–296.

[75] Gregoire, G., Derderian, F. & LeLorier, J. (1995). Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias?, *Journal of Clinical Epidemiology* **48**, 159–163.

[76] Harwell, M. (1997). An empirical study of Hedges' homogeneity test, *Psychological Methods* **2**, 219–231.

[77] Hasselblad, V., Eddy, D.M. & Kotchmar, D.J. (1992). Synthesis of environmental evidence: nitrogen dioxide epidemiology studies, *Journal of the Air and Waste Management Association* **42**, 662–671.

[78] Hedges, L.V. (1992). Modeling publication selection effects in meta analysis, *Statistical Science* **2**, 246–255.

[79] Hedges, L.V. (1994). Fixed effect models, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds. Russell Sage Foundation, New York.

[80] Hedges, L.V. & Olkin, I. (1980). Vote-counting methods in research synthesis, *Psychological Bulletin* **88**, 359–369.

[81] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Boston, Mass.

[82] Hemminki, E., Hailey, D. & Koivusalo, M. (1999). The courts – a challenge to health technology assessment, *Science* **285**, 203–204.

[83] Higgins, J.P. & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*. **21**(11), 1539–1558.

[84] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[85] Hunt, M. (1997). *How Science Takes Stock: The Story of Meta-Analysis*. Russell Sage Foundation, New York.

[86] Hunter, J.D. & Schmidt, F.L. (1995). *Methods of Meta-Analysis*. Sage Publications, London.

[87] Hunter, J.E. & Schmidt, F.L. (1994). Correcting for sources of artifactual variation across studies, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds. The Russell Sage Foundation, New York.

[88] Huston, P. (1996). Cochrane Collaboration helping unravel tangled web woven by international research, *Canadian Medical Association Journal* **154**, 1389–1392.

[89] Huston, P. (1996). Health services research: reporting on studies using secondary data sources, *Canadian Medical Association Journal* **155**, 1697–1702.

[90] Huston, P. & Moher, D. (1996). Redundancy, disaggregation, and the integrity of medical research, *Lancet* **347**, 1024–1026.

[91] Ioannidis, J.P., Cappelleri, J.C. & Lau, J. (1998). Issues in the comparison of meta-analysis and large trials, *Journal of the American Medical Association* **279**, 1089–1093.

[92] Ioannidis, J.P. & Lau, J. (1999). Pooling research results: benefits and limitations of meta-analysis, *Journal of Quality Improvement* **25**, 462–469.

[93] Iyengar, S. & Greenhouse, J.B. (1988). Selection models and the file drawer problem, *Statistical Sciences* **3**, 109–135.

[94] Jadad, A.R., Sullivan, C., Luo, D., Allen, I.E., Ross, S.D. & Sheinhait, I.A. (1999). Patients' preferences during the treatment of obstructive airway disease: a systematic review of studies comparing Turbuhaler with pressurized metered dose inhalers, *Annals of Allergy, Asthma, and Immunology*, in press.

[95] Kheifets, L.I., Afifi, A.A., Buffler, P.A., Zhang, Z.W. & Matkin, C.C. (1997). Occupational electric and magnetic field exposure and leukemia. A meta-analysis, *Journal of Occupational and Environmental Medicine* **39**, 1074–1091.

[96] Kleijinen, J., ter Riet, G. & Knipschild, P. (1990). Vitamin B-6 in the treatment of the premenstrual syndrome – a review, *British Journal of Obstetrics and Gynecology* **97**, 847–852.

[97] Krutan, B.M., Taylor, M.L. & Freeman, E. (1990). Vitamin B-6 in the treatment of the premenstrual syndrome, *Journal of the American Dietary Association* **90**, 859–861.

[98] L'Abbe, K.A., Detsky, A.S. & O'Rourke, K. (1987). Meta-analysis in clinical research, *Annals of Internal Medicine* **107**, 224–233.

[99] Lang, T.A. & Secic, M. (1997). *How to Report Statistics in Medicine*. American College of Physicians, New York.

[100] Lau, J., Antman, E.M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F. & Chalmers, T.C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction, *New England Journal of Medicine* **327**, 248–254.

[101] Lau, J., Ioannidis, J.P. & Schmid, C.H. (1998). Summing up evidence: one answer is not always enough, *Lancet* **351**, 123–127.

[102] Law, M.R., Morris, J.K. & Wald, N.J. (1997). Environmental tobacco smoke exposure and ischaemic heart disease: an evaluation of the evidence, *British Medical Journal* **315**, 973–980.

[103] LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J. & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials, *New England Journal of Medicine* **337**, 536–542.

[104] Light, R.J. & Pillemer, D.B. (1984). *Summing Up: The Science of Reviewing Research*. Harvard University Press, Boston, Mass.

[105] Light, R.J. & Smith, P.B. (1971). Accumulating evidence; procedures for resolving contradictions among different research studies, *Harvard Educational Review* **41**, 429–471.

[106] Lipsett, M. & Campleman, S. (1999). Occupational exposure to diesel exhaust and lung cancer: a meta-analysis, *American Journal of Public Health* **89**, 1009–1017.

[107] Lipsitz, S.R., Dear, K.B., Laird, N.M. & Molenberghs, G. (1998). Tests for homogeneity of the risk difference when data are sparse, *Biometrics* **54**, 148–160.

[108] Little, J., Bradley, Bray, M.S., Clyne, M., Dorman, J., Ellsworth, D.L., Hanson, J., Khoury, M., Lau, J., O'Brien, T.R., Rothman, N., Stroup, D.F., et al. (2002). Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am J Epidemiol* **156**, 300–310.

[109] Lohrer, B.T., Noe, R.A., Moeller, N.L. & Fitzgerald, M.P. (1985). A meta-analysis of the relation of job characteristics to job satisfaction, *Journal of Applied Psychology* **79**, 280–289.

[110] Lubin, J.H., Boice, J.D., Jr, Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P. & Samet, J.M. (1995). Lung cancer in radon-exposed miners and estimation of risk from indoor exposure, *Journal of the National Cancer Institute* **87**, 817–827.

[111] Mann, C.C. (1994). Can meta-analysis make policy?, *Science* **266**, 960–962.

[112] Meinert, C.L. (1989). Meta-analysis: science or religion?, *Controlled Clinical Trials* **10**, 257S–263S.

[113] Miller, W.C., Koceja, D.M. & Hamilton, E.J. (1997). A meta-analysis of the past 25 years of weight loss research using diet, exercise or diet plus exercise intervention, *International Journal of Obesity and Related Metabolic Disorders* **21**, 941–947.

[114] Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D. & Stroup, D., for the QUOROM Group. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement Lancet, (1999). 1896–1900.

[115] Moher, D., Schulz, K.F., Altman, D., for the CONSORT Group. (2001). The CONSORT Statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* **285**, 1987–2007.

[116] Moher, D., Fortin, P., Jadad, A.R., Juni, P., Klassen, T., Le Lorier, J., Liberati, A., Linde, K. & Penna, A. (1996). Completeness of reporting of trials published in languages other than English: implication for conduct and reporting of systematic reviews, *Lancet* **347**, 363–366.

[117] Moher, D. & Olkin, I. (1995). Meta-analysis of randomized controlled trials. A concern for standards, *Journal of the American Medical Association* **274**, 1942–1948.

[118] Moher, D., Pham, B., Jones, A., Cook, D.J., Jadad, A.R., Moher, M., Tugwell, P. & Klassen, T.P. (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analysis?, *Lancet* **352**, 609–613.

[119] Mosteller, F. & Chalmers, T.C. (1992). Some progress and problems in meta-analysis of clinical trials, *Statistical Science* **7**, 227–236.

[120] Naylor, C.D. (1997). Meta-analysis and the meta-epidemiology of clinical research, *British Medical Journal* **315**, 617–619.

[121] Ohlsson, A. (1994). Systematic reviews – theory and practice, *Scandinavian Journal of Clinical Laboratory Investigations* **219**, 25–32.

[122] Olkin, I. (1996). Meta-analysis: current issues in research synthesis, *Statistics in Medicine* **15**, 1253–1257.

[123] Olson, C.M. (1994). Understanding and evaluating a meta-analysis, *Academic Emergency Medicine* **1**, 392–398.

[124] Olson, C.M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, JW., Zhu, Q., Reiling, J. & Pace, B. (2002). Publication bias in editorial decision making. *JAMA.* **287**, 2825–2828.

[125] Patil, G.P. & Rao, C.R. (1977). The weighted distributions: a survey of their applications, in *Applications of Statistics*, P.R. Krishiaiah, ed. North-Holland, Amsterdam.

[126] Peipert, J.F. & Phipps, M.G. (1998). Observational studies, *Clinical Obstetrics and Gynecology* **41**, 235–244.

[127] Person, K. (1904). Report on certain enteric fever inoculations, *British Medical Journal* **2**, 1243–1246.

[128] Petitti, D. (1994). *Meta-Analysis, Decision Analysis, and Cost Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. Oxford University Press, New York.

[129] Pogue, J.M. & Yusuf, S. (1997). Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis, *Controlled Clinical Trials* **18**, 580–593.

[130] Pogue, J.M. & Yusuf, S. (1998). Overcoming the limitations of current meta-analysis of randomized controlled trials, *Lancet* **351**, 47–52.

[131] Realini, J.P. & Goldzieher, J.W. (1985). Oral contraceptives and cardiovascular disease: a critique of the epidemiologic studies, *American Journal of Obstetrics and Gynecology* **152**, 729–798.

[132] Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results, *Psychological Bulletin* **86**, 638–641.

[133] Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A. & Chalmers, T.C. (1983). Meta-analyses

of randomized controlled trials, *New England Journal of Medicine* **316**, 450–455.

[134] Sacks, H.S., Chalmers, T.C. & Smith, H. (1983). Sensitivity and specificity of clinical trials: randomized versus historical controls, *Archives of Internal Medicine* **143**, 753–755.

[135] Sacks, H.S., Reitman, D., Pagano, D. & Kupelnick, B. (1996). Meta-analysis: an update, *Mount Sinai Journal of Medicine* **63**, 216–224.

[136] Schlesselman, J.J. (1997). Risk of endometrial cancer in relation to use of combined oral contraceptives. A practitioner's guide to meta-analysis, *Human Reproduction* **12**, 1851–1863.

[137] Schulz, K.F. (1998). Randomized controlled trials, *Clinical Obstetrics and Gynecology* **41**, 245–256.

[138] Schulz, K.F., Chalmers, I., Hayes, R.J. & Altman, D.G. (1995). Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association* **273**, 408–412.

[139] Shapiro, S. (1994). Meta-analysis/Shmeta-analysis, *American Journal of Epidemiology* **140**, 771–778.

[140] Sipe, T.A. & Curlette, W.L. (1997). A meta-synthesis of factors related to educational achievement: a methodological approach to summarizing and synthesizing meta-analyses, *International Journal of Educational Research* **25**, 583–698.

[141] Smith, S.J., Caudill, S.P., Steinberg, K. & Thacker, S.B. (1995). On combining dose–response data from epidemiologic studies by meta-analysis, *Statistics in Medicine* **14**, 531–544.

[142] Song, F., Khan, K.S., Dinnes, J. & Sutton, A.J. (2002). Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *International Journal of Epidemiology*. **31**(1), 88–95.

[143] Spector, T.D. & Thompson, S.G. (1991). The potential and limitations of meta-analysis, *Journal of Epidemiology and Community Health* **45**, 89–92.

[144] Spitzer, W.O. (1995). The challenge of meta-analysis, *Journal of Clinical Epidemiology* **48**, 1–4.

[145] Standards of Reporting Trials Group (1994). A proposal for structured reporting of randomized controlled trials, *Journal of the American Medical Association* **272**, 1926–1931.

[146] Steinberg, K.K., Smith, S.F., Lee, N., Stroup, D.F., Olkin, I. & Williamson, G.D. (1997). A comparison of meta analysis to pooled analysis: an application to ovarian cancer, *American Journal of Epidemiology* **145**, 1917–1925.

[147] Steinberg, K.K., Smith, S.J., Thacker, S.B. & Stroup, D.F. (1994). Breast cancer risk and duration of estrogen use: the role of study design in meta-analysis, *Epidemiology* **5**, 415–421.

[148] Steinberg, K.K., Thacker, S.B., Smith, S.J., Stroup, D.F., Zack, M.M., Flanders, W.D. & Berkelman, R.L. (1991). A meta-analysis of the effect of estrogen replacement therapy on the risk of breast cancer, *Journal of the American Medical Association* **265**, 1985–1990.

[149] Stock, W.A. (1995). Systematic coding for research synthesis, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds. The Russell Sage Foundation, New York.

[150] Stoto, M. (1995). Research synthesis for public health policy: experience of the Institute of Medicine, in *Evaluation for the 21st Century: A Handbook*, W. Shadish & E. Chelimsley, eds. Sage Publications, New York.

[151] Stram, D.O. & Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model, *Biometrics* **50**, 1171–1177.

[152] Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., Rennie, D., Moher, D., Becker, B.J., Sipe, T.A. & Thacker, S.B. (2002). Meta-Analysis Of Observational Studies in Epidemiology: A Proposal for Reporting. *JAMA* **283**, 2008–2012.

[153] Stroup, D.F., Thacker, S.B. & Olson, C.M. (2001). Characteristics of Meta-analyses related to acceptance for publication in a medical journal. *Journal of Clinical Epidemiology* **54**, 655–660.

[154] Sutton, A.J., Jones, D.R., Abrams, K.R., Sheldon, T.A. & Song, F. (1999). Systematic reviews and meta-analysis: a structured review of the methodological literature, *Journal of Health Services Research and Policy* **4**, 49–55.

[155] Takkouche, B., Cadarso-Suárez, C. & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis, *American Journal of Epidemiology* **150**, 206–215.

[156] Thacker, S.B. (1988). Meta-analysis: a quantitative approach to research integration, *Journal of the American Medical Association* **259**, 1685–1689.

[157] Thacker, S.B. & Stroup, D.F. (2002). Methods and interpretation in systematic reviews: Commentary on two parallel reviews of epidural analgesia during labor. *Am J Obstet Gynecol* **186**, S78–S80.

[158] Thacker, S.B., Stroup, D.F., Branche, C.M., Gilchrist, J., Goodman, R.A. & Weitman, E.A. (1999). The prevention of ankle sprains in sports: a systematic review of the literature. *Am J Sports Med*, **27**, 753–760.

[159] Thompson, S.G. (1994). Why sources of heterogeneity in meta-analysis should be investigated, *British Medical Journal* **309**, 1351–1355.

[160] Vanhonacker, W.R. (1996). Meta-analysis and response surface extrapolation: a least squares approach, *American Statistician* **50**, 294–299.

[161] Vickers, A., Cassileth, B., Ernst, E., Fisher, P., Goldman, P., Jonas, W., Kang, S.K., Lewith, G., Schulz, K. & Silagy, C. (1997). How should we research unconventional therapies? A panel report from the Conference on Complementary and Alternative Medicine Research Methodology, National Institutes of Health, *International Journal of Technology Assessment in Health Care* **13**, 111–121.

[162] Wachter, K.W. (1998). Disturbed by meta-analysis? *Science* **241**, 1407–1408.

[163] Wachter, K.W. & Straf, M. (1990). *The Future of Meta-analysis*. Russell Sage Foundation, New York.

[164] Ware, J.H., Muller, J.E. & Braunwald, E. (1985). The futility index: an approach to the cost-effective termination of randomized clinical trials, *American Journal of Medicine* **78**, 635–643.

[165] Wells, A.J. (1998). Heart disease from passive smoking in the workplace, *Journal of the American College of Cardiology* **31**, 1–9.

[166] Wilson, M.E., Fineberg, H.V. & Colditz, G.A. (1995). Geographic latitude and the efficacy of bacillus Calmette–Guerin vaccine, *Clinical Infectious Diseases* **20**, 982–991.

[167] Wortman, P.M. (1994). Judging research quality, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds. Russell Sage Foundation, New York.

[168] Zaza, S., Lawrence, R.S., Mahan, C.S., Fullilove, M., Fleming, D. Isham, G.J. & Pappaioanou, M. (2000). Scope and organization of the Guide to Community Preventive Services. *American Journal of Preventive Medicine*. **18**(1 Suppl), 27–34.

Further Reading

Berlin, J.A., Laird, N.M., Sacks, H.S. & Chalmers, T.C. (1989). A comparison of statistical methods for combining event rates from clinical trails, *Statistics in Medicine* **8**, 141–151.

Cook, D.J., Sackett, D.L. & Spitzer, W. (1995). Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation on meta-analysis, *Journal of Clinical Epidemiology* **48**, 167–171.

Fleiss, J.L. (1993). The statistical basis of meta-analysis, *Statistical Methods of Medical Research* **2**, 121–145.

Hasselblad, V. (1995). Meta-analysis of environmental health data, *Science of the Total Environment* **160–161**, 545–558.

Hasselblad, V., Mosteller, F., Littenberg, B., Chalmers, T.C., Hunink, M.G., Turner, J.A., Morton, S.C., Diehr, P., Wong, J.B. & Powe, N.R. (1995). A survey of current problems in meta-analysis. Discussion from the Agency for Health Care Policy and Research inter-PORT Work Group on Literature Review/Meta-Analysis, *Medical Care* **33**, 202–220.

Miller, N. & Pollock, V.E. (1994). Meta-analytic synthesis for theory development, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds. The Russell Sage Foundation, New York.

Mosteller, F. & Colditz, G.A. (1996). Understanding research synthesis (meta-analysis), *Annals of Review of Public Health* **17**, 1–23.

(*See also* **Combining *P* Values**)

DONNA F. STROUP & STEPHEN B. THACKER

Meta-analysis in Human Genetics

Replication or confirmation of scientific findings is a hallmark of good science. Laboratory experiments may be relatively easy to replicate, but field studies or observational studies can be quite challenging. Especially difficult are studies involving human populations. Indeed, most studies reporting new genetic linkage findings have had little success in being replicated [3, 24, 39]. In some cases, however, the effects of trait loci (*see* **Gene**) are so strong that replication is easily observed. Mendelian traits (*see* **Mendel's Laws**) fall into this category and include diseases such as cystic fibrosis and hypercholesterolemia. **Complex diseases** (e.g. heart disease, diabetes, schizophrenia, asthma) and quantitative traits (e.g. height, low density lipoprotein cholesterol), which are believed to be influenced by any number of genetic loci (*see* **Genotype**), require relatively much larger sample sizes or special sampling designs to attain sufficient statistical power to detect and locate the underlying **genes**. Replication of genetic findings for these non-Mendelian traits appears to be very difficult. The reasons are many, including **genetic heterogeneity**, sampling designs and **ascertainment** rules, environmental factors, and **covariates** such as age and sex. It may be possible to statistically adjust for some of these factors, for example by **regression** methods or stratification. Even after statistical adjustment, however, differences may remain among the primary studies. These differences may be real or due to chance. Combining results across studies may aid in assessing the overall genetic findings, the results of which can have important ramifications. For example, since it is now possible to develop molecular-based pharmaceutical treatments, it is vital to have an understanding of the genetic mechanisms underlying a given human trait. The synthesis of findings from similar studies forms the basis of much scientific understanding and meta-analysis is one approach to this end.

Basic Ideas and Methods of Meta-Analysis

The standard scientific review of a collection of studies addressing the same question often takes the form of a subjective qualitative assessment. Meta-analysis is a term used to describe *quantitative* methods to integrate or pool the numerical results from a collection of similar studies. Hedges & Olkin [23] narrow the definition by noting that "because meta-analysis usually relies on 'data' in the form of summary statistics derived from the primary analyses of studies, it is truly an analysis of the results of statistical analyses". It is in the spirit of this definition that meta-analysis is used in genetic studies.

Pooling information across replicate or similar studies is a common problem. If the studies are sufficiently similar, then combining the raw data from the different studies and analyzing this combined data set is the best one can do. More often, however, only published results (summary statistics, ***P* values**) are available for analysis. A formal approach for analyzing P values was proposed by Fisher [12] and Pitman [33], both of whom wrote about summarizing statistical findings across a collection of analogous studies. Assume that each of k independent studies tests a common **null hypothesis** and that each yields a one-tailed P value in the same direction. Under the omnibus null hypothesis that each of the k null hypotheses holds true, the test statistic

$$X^2 = -2\sum_{i=1}^{k} \log(P_i)$$

follows a χ^2 distribution with $2k$ degrees of freedom. Therefore, it is possible to jointly evaluate the evidence in favor of the alternative hypothesis by using a single test and all that is required are the published P values from the individual studies. Variations on this approach include analysis of the P value order statistics and a weighted linear combination of the P values to account for varying sample sizes across the studies [23]. This method is still widely used (e.g. in epidemiologic studies and clinical trials) as an approach to meta-analysis despite known limitations. It is well known, for example, that the X^2 statistic can be highly influenced by a single highly significant P value, which in turn may be due to a large sample size. Another concern is the simple fact that a single P value is used to summarize each (possibly complex) study as a basis for meta-analysis.

A more informative approach to meta-analysis is based on combining parameter estimates that reflect the treatment effect. For example, in clinical trials one may be interested in knowing the difference in the proportion of treatment and control subjects

$(p_T - p_C)$ that experience relief. A P value analysis will indicate whether there is consensus across the studies, but it may be more important to know the overall strength of the deviation from no treatment effect. This may not always be possible in practice since different studies addressing the same null hypothesis may use different statistics. When it seems reasonable to provide an overall estimate of a common parameter, it is important to make a careful judgment about the similarity of the studies. If one concludes that the studies are sufficiently homogeneous in design and sampling scheme that they are measuring or estimating the same quantity (δ), then it is appropriate to "pool information across studies" using a **fixed effect** model,

$$d_i = \delta + e_i,$$

where d_i is the observed value for the fixed effect δ from the ith study, and e_i is viewed as a random error term, independent and identically distributed across studies. The pooled estimator of δ is a weighted (least squares) average of the individual d_i:

$$\hat{\delta}_w = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i},$$

where the weights are inversely proportional to the (estimated) **variances** of e_i, $w_i = \text{var}^{-1}(e_i)$. Genetic studies, even when "replication" is intended, in general tend not to be sufficiently homogeneous that it is believed the same quantity is being estimated from study to study. Indeed, population **admixture**, ascertainment, and **marker** maps, to name a few factors, vary enough so that the fixed effect model does not adequately reflect the study heterogeneity. A standard test for homogeneity makes use of a χ^2 statistic:

$$Q_w = \sum_{i=1}^{k} w_i (d_i - \hat{\delta}_w)^2,$$

which is approximately distributed as a χ^2 variable with $k - 1$ degrees of freedom.

If one judges that the differences in the d_i are due to both within-study variation (e_i) and among-study variation, then the model for d_i may be generalized to a **random effect** model to account for the two sources of variation. In this case it is useful to view the actual conducted studies as being a sample from a larger population of similar studies, each with an underlying (but unknown) δ that reflects the individual characteristics of its study. Thus, there is a corresponding population $\{\delta_j\}$, with mean δ and variance σ_δ^2. A two-stage or **hierarchical model** describes this situation:

$$d_i = \delta_i + e_i;$$

$$\delta_i = \delta + \varepsilon_i,$$

where $\text{var}(\varepsilon_i) = \sigma_\delta^2$, e_i is as above, and all the error terms are independent. In other words,

$$d_i = \delta + \varepsilon_i + e_i, \quad i = 1, \ldots, k.$$

From the hierarchical modeling perspective it can be seen that there is information about δ_i in all the studies, since they all vary about a common mean (δ). In this case, one can "borrow strength across studies" to estimate δ_i. Note that if there is no study-to-study variation [$\text{var}(\varepsilon_i) = \sigma_\delta^2 = 0$], then the random effect model coincides with the fixed effect model. If some of the study-to-study variation can be explained, then further modeling of δ_i may be incorporated in a mixed effect model. As with the fixed effect model, an estimate for the common δ is a weighted average:

$$\hat{\delta}_w = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i},$$

where now the weights reflect both within-study variation and among-study variation, $w_i = [\text{var}(e_i) + \text{var}(\varepsilon_i)]^{-1}$. In practice, the variances (and, hence, the weights) are almost always estimated [8]. It is important that the hypothesis of homogeneity be investigated since the standard errors based on the two modeling approaches may differ substantially. The fixed effect standard error tends to be optimistically too small when sources of among-study variation are ignored.

Meta-Analysis Applied to Statistical Genetics

Application of the above methods in **human genetics** involves the common null hypothesis of no

association or linkage at a given marker locus or along some chromosomal region. Lack of consensus across studies may be due to having some studies supporting linkage and others supporting no linkage, or having most studies support only marginal significance, but none or very few showing strong evidence of linkage. The former may likely be due to substantial study heterogeneity, while the latter may arise because of low statistical power at the level of the individual studies. The main objective of a meta-analysis is to combine the results from the different studies to obtain an overall assessment of association or linkage in a given chromosomal region. The idea for meta-analysis in genetic studies is not exactly new. When the recombination fraction (θ) is the appropriate genetic parameter of interest [38], the lod function, $Z(\theta) = \log_{10}[L(\theta)/L(\theta = 0.5)]$, for each study is maximized to obtain independent estimates of θ [32]. An overall estimate can be determined from the sum of the individual lod functions and a test of homogeneity can subsequently be carried out.

As an illustration of meta-analysis involving modern genetic linkage studies, Allison & Heo [2] use Fisher's P value method in a meta-analysis of six independent samples, each concerning linkage of body mass index to the region of the human genome containing the OB gene. In the primary meta-analysis, Allison & Heo show three samples with evidence of linkage at the 0.05 significance level, while three samples do not show significant linkage. The X^2 meta-analysis statistic yields a P value of 1.5×10^{-5}, providing strong overall evidence of linkage. One of the main problems with such standard meta-analytic methods is the level of study-to-study heterogeneity, especially **genetic heterogeneity**, as well as sample size, ascertainment of families, marker maps (marker distribution, average intermarker distance, marker heterozygosity), disease definition, and statistical analysis [34]. As such, it is not always a simple matter to extract a single P value from each of the primary studies, as demonstrated by Allison & Heo. Even if this can be accomplished, it should be noted that this approach to meta-analysis is unequivocally valid only if the P value is taken at the same marker locus from each study. In practice, however, the individual study P values are associated with the most significant marker within a given chromosomal region, which may contain more than a handful of markers. In practice, then, we may expect an increased false-positive rate. Additionally, it has been observed [35] that there is a **bias** in Fisher's X^2 test when it is applied to model-free linkage methods that constrain lod scores to nonnegative values. The bias is especially acute in the context of genome scans, but an adjustment is available. Guerra et al. [20] discuss Fisher's P value method in the context of genome scans.

In those situations where a common parameter estimate of linkage or association is available, a goal of the meta-analysis is to obtain a combined estimate to indicate the overall strength of the **correlation** across studies. Thus far, two common situations arise in practice. The first deals with population-based studies of associations between marker **polymorphism** and disease status. The standard study design is a **case–control study** where the cases and controls are affected and unaffected subjects, respectively, and the "treatment factor" is a specified marker **genotype**. Thus, if there is interest in the association between the AA genotype at a given locus and disease status, then the data for each individual study are typically shown as a table of counts (Table 1), and relevant differences in proportions, **relative risks**, or **odds ratios** [1] are analyzed. In genetic studies [27, 41] it is common to work with the observed odds ratio ($OR = ad/bc$). In this case, a test of homogeneity for a common OR is performed and, if not rejected, a common OR is estimated and used as an overall estimate of the genotype–phenotype association. The overall estimate can be calculated as a crude OR based on pooled counts (which are typically available in the primary publications), Mantel–Haenzsel estimator, or Woolf estimator. The latter appears to be a popular choice among genetic studies, perhaps due to its early appearance in the genetics literature, where it was introduced in the context of correlating **blood groups** with disease. Instead of using the ORs themselves, Woolf [46] proposed using the logarithms of the ORs since on this scale they are approximately normally distributed. For k studies he proposed using a weighted average

Table 1

	Affected	Unaffected
AA genotype	a	b
Not AA genotype	c	d

of these logarithms,

$$\log(OR_{\mathrm{W}}) = \frac{\sum_{i=1}^{k} w_i \log(OR_i)}{\sum_{i=1}^{k} w_i},$$

where the weights are given by the inverse of the variance of the log *OR*:

$$w_i = \mathrm{var}^{-1}[\log(OR_i)] = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{-1}.$$

Woolf's test for homogeneity is based on a χ^2 statistic,

$$X^2 = \sum_{i=1}^{k} w_i \left[\log(OR_i) - \log(OR_{\mathrm{W}})\right]^2,$$

which, under the null hypothesis of a common *OR*, approximately follows a χ^2 distribution with $k-1$ degrees of freedom. Case–control studies in genetics have been used to evaluate the total evidence for population associations between marker genotypes and a variety of genetic traits, including schizophrenia [9], hypertension [26], and cleft lip [31]. To investigate various genotype effects (e.g. homozygosity, additivity), as well as to account for **covariates**, the meta-analysis can also be carried out by **logistic regression** with case–control status as the dependent variable [40, 41].

Analogous to the case–control analysis for disease data one can analyze an overall mean difference of a quantitative trait between two groups defined by genotypes. For example, Juo et al. [25] conducted a meta-analysis of the mean difference in apolipoprotein A-I levels between two genotype groups defined by an A/G **polymorphism** in the promoter of the apolipoprotein A-I gene. The overall mean difference is a weighted average of the primary study mean differences. A test of homogeneity and a confidence interval for the overall difference form the basis of the analysis [16]. Juo et al. [25] include a good discussion of **confounding** factors and stratification as they pertain to meta-analysis.

Genetic linkage studies allow for other measures of genetic effect size that can be defined and pooled across a collection of genetic studies. For model-free methods that do not depend on a parametric genetic model for analysis, allele-sharing data may be used [18, 19] to define a genetic effect size. Affected sib pairs, for example, are expected to share a higher proportion (π) of alleles than that expected ($\pi = 0.5$) under no linkage. The mean proportion of identity-by-descent (ibd) (*see* **Identity Coefficients**) allele-sharing at a specified locus can therefore be taken as an interpretable parameter to be combined across studies. Let $\overline{\pi}_i, i = 1, 2, \ldots, k$, denote the average proportion of ibd allele-sharing among n_i affected sib pairs in study i. A combined estimate of mean allele-sharing across homogeneous studies can be constructed by a weighted least squares estimate based on a fixed effect model as discussed above:

$$\overline{\pi}_w = \frac{\sum_{i=1}^{k} w_i \overline{\pi}_i}{\sum_{i=1}^{k} w_i},$$

where $w_i = 1/\hat{\sigma}_i^2$, the reciprocal of the estimated variance of $\overline{\pi}_i$. A standard error for the combined estimate can be easily calculated according to the result $\mathrm{var}(\overline{\pi}_w) = 1/\sum w_i$. The homogeneity hypothesis can be tested with the following statistic:

$$Q_w = \sum_{i=1}^{k} \frac{(\overline{\pi}_i - \overline{\pi}_w)^2}{\hat{\sigma}_i^2},$$

which asymptotically follows a χ^2 distribution with $(k-1)$ degrees of freedom under the null hypothesis. Under the more realistic scenario of study heterogeneity, Gu et al. [19] have proposed a random effect modeling approach where the observed proportion ($\overline{\pi}_i$) of ibd sharing in study i reflects a random effect due to study-to-study variation and a random effect due to within-study variation. Following their notation, the model for $\overline{\pi}_i$ is

$$\overline{\pi}_i = \tau + \delta_i + \varepsilon_i,$$

where the within-study variation is $\mathrm{var}(\varepsilon_i)$ and the among-study variation is $\mathrm{var}(\delta_i)$. Estimating the two variance components and the overall measure of allele-sharing (τ) follows the general approach discussed above and is detailed in Gu et al. [19], while Goldstein et al. [15] investigate both fixed effect and random effect models for $\overline{\pi}_i$ in the context of genome scans. The random effect model can be extended to include study-specific covariates [17].

There are other types of studies that make use of allele-sharing data, but which may not explicitly provide observed values of $\overline{\pi}_i$. Nevertheless, it is still possible to extract this information from a variety of studies [18]. Once such information is obtained, the above models may be used with the derived or estimated proportions of allele-sharing. Similarly, random effect and Bayesian hierarchical models have been proposed for combining Haseman–Elston slope estimates (see below) from independent studies investigating linkage to a common quantitative trait locus using the same marker.

The above meta-analysis methods for linkage have not been specifically designed for genome scans; they are clearly applicable for investigating linkage at a single locus or very small chromosomal region. However, as with previous statistical methods developed for genetic linkage at a single marker, one can obtain a meta-analysis test statistic score map along the genome and assess significance through a multiple testing procedure. The standard guidelines for **genome-wide significance** levels were given by Lander & Kruglyak [28], but it is unclear how they may apply to meta-analysis. It is thought that less stringent significance levels than those proposed by Kruglyak & Lander may be appropriate for meta-analyses of genome scans, but a definitive theory or strong empirical evidence does not exist at this time. Badner & Goldin [4] depart from the standard approach of generating and analyzing a meta-analysis version of the underlying test statistic. They seek to answer the natural question, "How often can we expect more than one genome scan to exceed a given threshold in a defined region if, in fact, linkage is absent in the region?" A binomial probability model is applicable and used to find theoretical answers to the question. For a single-point linkage analysis in the primary studies, the basis of their meta-analysis method [the multiple genome test (MGT)] is to calculate the chance of having observed r_k out of s studies with significant linkage in each of k chromosome regions, each defined by a range of consecutive markers (e.g. four). Meta-analysis based on **multipoint linkage** is similarly defined by considering 30 cM contiguous regions. By varying the thresholds in the binomial model, Badner & Goldin demonstrate that less stringent thresholds at the primary study level may yield highly significant results at the meta-analysis level. **Simulation** studies show good agreement between theoretical and empirical estimates of false-positive rates for a single-point analysis, but not a multipoint analysis. On balance, for the simulations investigated, the MGTs did not show a significant advantage in power over the Kruglyak & Lander guidelines. Although several limitations are discussed, their method is potentially quite useful as it directly addresses the question of consensus. For related issues, see the variety of significance thresholds used for meta-analysis in Genetic Analysis Workshop 11 [4, 15, 17, 20, 44, 47].

Relatively few meta-analytic methods specifically designed for genome scans currently exist. Each is based on obtaining a meta-analysis test statistic value in each of k_c user-defined segments per chromosome (C) and then testing for significance within each segment. Common marker maps are not assumed and each method proposes an approach to control the genome-wide type I error probability that results from multiple testing (*see* **Multiple Comparisons**) across the segments. A nonparametric rank-based method has been proposed by Wise et al. [45], who apply it to four screens for multiple sclerosis and 11 screens for autoimmune disorders; Wise & Lewis [44] investigate the type I error and the power of the method in a simulation study. Within each of m studies, each chromosome is divided into segments of equal length and the most significant result from each segment is ranked among all segments covering the entire genome; the most extreme results have the higher ranks. Under the null hypothesis of no linkage in any segment, the ranks within each primary study will be uniformly distributed throughout the genome. Within each defined segment, the primary study ranks are summed across the m genome scans, and for n segments across the genome, the chance that the ranks, $X_i, i = 1, \ldots, m$, from a fixed segment sum to R is

$$\Pr\left(\sum_{i=1}^{m} X_i = R\right) = \frac{1}{n^m}\sum_{k=0}^{d}(-1^k)\binom{R-kn-1}{m-1} \times \binom{m}{k}, \quad m \le R \le mn,$$

and 0 if $R > mn$. In the formula, d is the integer part of $(R-m)/n$. This method, like most developed thus far, treats each genome scan equally and does not incorporate strategies to allow for study differences such as may be due to marker maps and sample sizes. However, it is broadly applicable as it can be used across studies that use different analyses (e.g.

model-based and model-free), disease definitions, and even different diseases (e.g. different psychiatric disorders).

Meta-analytic approaches for multiple genome scans based on a genetic effect size *per se*, instead of significance results, are also being developed. We consider a method proposed by Etzel [10] and further developed by Etzel & Guerra [11]. To fix ideas, assume that there exists at most a single trait locus and that each of k studies have used the Haseman–Elston robust sib-pair method [22] to search for the gene in a chromosome scan of length L cM. This situation will illustrate some of the issues that arise in developing a meta-analytic procedure for genome scans. The approach is not unique to the Haseman–Elston method and is generally applicable to other genetic effect sizes or test statistics. Assume that within each of m nonoverlapping, contiguous segments of equal length covering the chromosome, each study has exactly one marker $(M_{ij}, i = 1, \ldots, k; j = 1, \ldots, m)$ at distinct locations. The Haseman–Elston method regresses the square of sib-pair phenotypic differences on the proportion of shared ibd (marker) alleles and a significant negative slope gives evidence of linkage. The expected value of the slope estimate, $\hat{\beta}$, is a function of the recombination fraction (θ) between the marker and trait loci and the variance component (σ_g^2) due to polymorphism at the trait locus:

$$\mathrm{E}(\hat{\beta}) = -2(1 - 2\theta)^2\sigma_g^2.$$

Therefore, assume that each study provides a summary statistic $(\hat{\beta}_{ij}, S_{ij}^2), i = 1, \ldots, k, j = 1, \ldots, m$, at each marker locus, where S^2 is the estimated variance of $\hat{\beta}$. The meta-analytic method is defined as follows. At each of q test points along the chromosome, an overall estimate of the Haseman–Elston slope is calculated as a weighted average of normalized $\hat{\beta}$s that correspond to the segment in which the test point is located. The weights correspond to a random effect estimator that reflects both within-study and among-study variation. The test point at which the overall estimate is most significant is taken as the most likely location of the trait locus. The normalization is required to adjust for the fact that within each chromosomal segment the study markers are at different loci and therefore the recombination fractions implicit in the slope estimates are different. That is, the method does not require studies to use the same marker maps. The weighted averages are calculated from the following normalized statistics:

$$\hat{\beta}_{ijq} = \frac{\hat{\beta}_{ij}}{(1 - 2\theta_{ijq})^2} \quad \text{and} \quad S_{ijq}^2 = \frac{S_{ij}^2}{(1 - 2\theta_{ijq})^4},$$

where θ_{ijq} is the recombination fraction between the marker M_{ij} and test point q, calculated by applying a mapping function (*see* **Linkage Analysis, Multivariate**). The overall estimate of the Haseman–Elston slope is calculated as:

$$\hat{\beta}_q = \frac{\sum_{i=1}^{k} w_i \hat{\beta}_{ijq}}{\sum_{i=1}^{k} w_i}; \quad w_i = \frac{1}{\hat{\sigma}_q^2 + S_{ijq}^2},$$

where $\hat{\sigma}_q^2$ estimates the study-to-study variation at test point q:

$$\hat{\sigma}_q^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\hat{\beta}_{ijq} - \overline{\beta}_q)^2 - \frac{1}{k}\sum_{i=1}^{k} S_{ijq}^2,$$

with $\overline{\beta}_q$ the average of the $\hat{\beta}_{ijq}$. The overall slope estimate is standardized by its standard error (se) to assess its significance with reference to a standard normal distribution:

$$z_q = \frac{\hat{\beta}_q}{\mathrm{se}(\hat{\beta}_q)}; \quad \mathrm{se}(\hat{\beta}_q) = \left(\sum_{i=1}^{k} w_i\right)^{-1/2}.$$

Denoting the most significant overall slope by β_q^*, the genetic variance component can be estimated as $\hat{\sigma}_g^2 = -\beta_q^*/2$. Simulations have been carried out [10] to determine if the meta-analysis method increases power to detect linkage over the pointwise (Haseman–Elston) method at the primary study level, compare its performance under various configurations of primary study evidence, and assess its accuracy in locating the trait locus and estimating the component of genetic variance σ_g^2. Since it is unknown how to select significance thresholds for a meta-analysis of genome scans, two approaches were taken. First, z_q was referred to the normal quantile (-2.633) corresponding to a chromosome-wide significance level of 0.05, which in turn corresponds to a pointwise level of 0.004 by the method of Lander & Kruglyak [28]. Secondly, the standard 0.05 threshold at the pointwise level was used, ignoring multiple

testing. The Haseman–Elston slope estimate at the primary study level was evaluated using a Bonferroni correction ($0.05/m$) for multiple testing across a map of m markers. Simulation configurations varied over marker density, sib-pair sample size, and trait gene location. The main findings show that:

1. Both meta-analysis methods provided substantially better power than the individual studies.
2. The type I error probability of the meta-analysis method using the standard pointwise 0.05 threshold was consistently higher than that based on the Lander–Kruglyak (LK) genome-wide threshold.
3. The power of the meta-analysis using the LK threshold was slightly less than the meta-analysis based on the 0.05 pointwise threshold. However, the pointwise method showed a high false-positive rate outside a relatively small neighborhood (± 10 cM) of the trait locus. In practice, the size of the neighborhood can be expected to be a function of the marker density.
4. In all cases, the meta-analysis based on the LK threshold located the trait locus within 2.5 cM.
5. Meta-analysis with the LK threshold consistently provided unbiased estimates of the genetic variance, σ_g^2. It is unknown whether the Lander–Kruglyak threshold is appropriate for this meta-analytic approach; in particular, the Lander–Kruglyak guidelines assume a dense marker map, which is not the case in the meta-analysis simulation or in most real genome scans.

To minimize statistical parametric assumptions (e.g. normality of a test statistic) and to reflect the inherent characteristics (e.g. marker map) of different genetic studies, randomization (permutation) or Monte-Carlo methods [5, 7, 21, 42] can be use to construct nonparametric null distributions for the overall slope estimate. In this case, a permutation approach resulted in an improvement in the meta-analysis type I error probability and therefore increased the precision in locating the gene locus [10]. The power of the permutation test to locate the trait locus within 10 cM was, on average, slightly less than the power of the meta-analysis method that used the Lander–Kruglyak normal critical value. In the simulations considered, the permutation test is superior to the parametric meta-analysis method when the trait locus is located near either end of the chromosome. Inspection of the permutation thresholds also shows that the thresholds vary markedly with the position of the trait locus, confirming that the method adjusts to the specific characteristics of the study [5].

Meta-analysis methodology is just emerging as a tool for modern genetic analysis, but its usefulness and importance is quickly being recognized [28, 36] as we learn how difficult it is to replicate initial linkage or association findings. In addition to the technical methodology, there are practical issues that must be addressed, especially as more genetic studies on the same trait become available. The planning of a meta-analysis requires great care. Some studies will be useful and others will not. The following steps summarize what must be done to increase the chances of a successful and informative meta-analysis:

1. Formulate a specific purpose and explicitly define the outcome to be extracted from each study. Bear in mind that genetic effect sizes are more informative than significance results.
2. Identify relevant primary studies.
3. Establish inclusion/exclusion criteria for primary studies.
4. Detail data abstraction and acquisition.
5. Decide on data analysis method(s) and understand their limitations. Carefully investigate study-to-study heterogeneity.
6. Give a careful interpretation, especially as it applies to extrapolation of findings.

Further Reading

A basic reference for meta-analytic ideas, concepts, and methods is Hedges & Olkin [23], while Draper et al. [8] provide a general discussion on combining information. Mann [30] gives an overview of meta-analysis in medical research. Rao & Province [37] cover statistical genetics, including a chapter on meta-analysis; see also Cooper & Hedges [6]. The meta-analytic methods detailed in this article cover current methods that appear to be the most common in human genetics studies. Other approaches include a Bayesian method [43], Morton's β model [29], and model selection [14]. The term meta-analysis was coined by Glass [13].

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Allison, D.B. & Heo, M. (1998). Meta-analysis of linkage data under worst-case conditions: a demonstration using the human OB region, *Genetics* **148**, 859–865.

[3] Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H. & Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find, *American Journal of Human Genetics* **69**, 936–950.

[4] Badner, J.A. & Goldin, L.R. (1999). Meta-analysis of linkage studies, *Genetic Epidemiology* **17**, Supplement 1, S485–S490.

[5] Churchill, G.A. & Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping, *Genetics* **194**, 963–971.

[6] Cooper, H. & Hedges, L.V. (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York.

[7] Doerge, R.W. & Churchill, G.A. (1996). Permutation test for multiple loci affecting a quantitative character, *Genetics* **142**, 285–294.

[8] Draper, D., Gaver, D.P., Goel, P., Greenhouse, J., Hedges, L., Morris, C.N., Tucker, J.R. & Waternaux, C.M. (1992). *Combining Information*, *Contemporary Statistics*, Number 1. National Academy Press, Washington.

[9] Dubertret, C., Gorwood, P., Ades, J., Feingold, J., Schwartz, J.-C. & Sokoloff, P. (1998). Meta-analysis of DRD3 gene and schizophrenia: ethnic heterogeneity and significant association in Caucasians, *American Journal of Medical Genetics* **81**, 318–322.

[10] Etzel, C.J. (1999). Meta-analysis for genetic linkage studies. Ph.D. dissertation, Southern Methodist University.

[11] Etzel, C.J. & Guerra, R. (2001). Meta-analysis of genetic linkage studies of quantitative trait loci, Technical Report no. TR01-2, Department of Statistics, Rice University, submitted for publication.

[12] Fisher, R.A. (1954). *Statistical Methods for Research Workers*, 12th Ed. Hafner, New York.

[13] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Researcher* **5**, 3–8.

[14] Goffinet, B. & Gerber, S. (2000). Quantitative trait loci: a meta-analysis, *Genetics* **155**, 463–473.

[15] Goldstein, D.R., Sain, S.R., Guerra, R. & Etzel, C.J. (1999). Meta-analysis by combining parameter estimates: simulated linkage studies, *Genetic Epidemiology* **17**, Supplement 1, S581–S586.

[16] Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature, *Epidemiology Review* **9**, 1–30.

[17] Gu, C., Province, M. & Rao, D.C. (1999). Meta-analysis of genetic linkage to quantitative trait loci with study-specific covariates: a mixed-effects model, *Genetic Epidemiology* **17**, Supplement 1, S599–S604.

[18] Gu, C., Province, M. & Rao, D.C. (2001). Meta-analysis for model-free methods, in *Genetic Dissection of Complex Traits*, D.C. Rao & M. Province, eds. Academic Press, New York, pp. 255–272.

[19] Gu, C., Province, M., Todorov, A. & Rao, D.C. (1998). Meta-analysis methodology for combining non-parametric sibpair linkage results: genetic heterogeneity and identical markers, *Genetic Epidemiology* **15**, 609–626.

[20] Guerra, R., Etzel, C., Goldstein, D. & Sain, S. (1999). Meta-analysis by combining p-values: simulated linkage studies, *Genetic Epidemiology* **17**, Supplement 1, S605–S609.

[21] Guerra, R., Wan, Y., Jia, A., Amos, C.I. & Cohen, J.C. (1999). Testing for linkage under robust genetic models, *Human Heredity* **49**, 146–153.

[22] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[23] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York.

[24] Johnson, G.C.L. & Todd, J.A. (2000). Strategies in complex disease mapping, *Current Opinion in Genetics and Development* **10**, 330–334.

[25] Juo, S.-H.H., Wyszynski, D.F., Beaty, T.H., Huang, H.-Y. & Bailey-Wilson, J.E. (1999). Mild association between the A/G polymorphism in the promoter of the apolipoprotein A-I gene and apolipoprotein A-I levels: a meta-analysis, *American Journal of Medical Genetics* **82**, 235–241.

[26] Kato, N., Sugiyama, T., Morita, H., Kurihara, H., Yamori, Y. & Yazaki, Y. (1999). Angiotensinogen gene and essential hypertension in the Japanese: extensive association study and meta-analysis on six reported studies, *Journal of Hypertension* **17**, 757–763.

[27] Khoury, M.J., Beaty, T.H. & Cohen, B.H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press, New York.

[28] Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**, 241–247.

[29] Liò, P. & Morton, N. (1997). Comparison of parametric and nonparametric methods to map oligogenes by linkage, *Proceedings of the National Academy of Sciences* **94**, 5344–5348.

[30] Mann, C. (1990). Meta-analysis in the breech, *Science* **249**, 476–480.

[31] Mitchell. L.E. (1996). Transforming growth factor α locus and nonsyndromic cleft lip with or without cleft palate: a reappraisal, *Genetic Epidemiology* **14**, 231–240.

[32] Ott, J. (1991). *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

[33] Pitman, E.J. (1937). Significance tests which may be applied to samples from any populations, *Journal of the Royal Statistical Society, Series B* **4**, 119–131.

[34] Prathikanti, S. & McMahon, F.J. (2001). Genome scans for susceptibility genes in bipolar affective disorder, *Annals of Medicine* **33**, 257–262.

[35] Province, M.A. (2001). The significance of not finding a gene, *American Journal of Human Genetics* **69**, 660–663.

[36] Rao, D.C. (1998). CAT scans, PET scans, and genomic scans, *Genetic Epidemiology* **15**, 1–18.
[37] Rao, D.C. & Province, M. (2001). *Genetic Dissection of Complex Traits*, Academic Press, New York, pp. 255–272.
[38] Rice, J.P. (1997). The role of meta-analysis in linkage studies of complex traits, *American Journal of Medical Genetics* **74**, 112–114.
[39] Roberts, S.B., MacLean, C.L., Neale, M.C., Eaves, L.J. & Kendler, K.S. (1999). Replication of linkage studies of complex traits: an examination of variation in location estimates, *American Journal of Human Genetics* **65**, 876–884.
[40] Selvin, S. (1996). *Statistical Analysis of Epidemiologic Data*, 2nd Ed. Oxford University Press, New York.
[41] Sham, P. (1998). *Statistics in Human Genetics*, Arnold, New York.
[42] Wan, Y., Guerra, R. & Cohen, J.C. (1997). A permutation method for the robust sib-pair linkage method, *Annals of Human Genetics* **61**, 79–87.
[43] Wang, K., Vieland, V. & Huang, J. (1999). A Bayesian approach to replication of linkage findings, *Genetic Epidemiology* **17**, Supplement 1, S749–S754.
[44] Wise, L.H. & Lewis, C.M. (1999). A method for meta-analysis of genome searches: application to simulated data, *Genetic Epidemiology* **17**, Supplement 1, S767–S771.
[45] Wise, L.H., Lanchbury, J.S. & Lewis, C.M. (1999). Meta-analysis of genome scans, *Annals of Human Genetics* **63**, 263–272.
[46] Woolf, B. (1955). On estimating the relation between blood group and disease, *Annals of Human Genetics* **19**, 251–253.
[47] Xu J., Wiesch, D.G., Taylor, E.W. & Meyers, D.A. (1999). Evaluation of replication studies, combined data analysis, and analytical methods in complex diseases, *Genetic Epidemiology* **17**, Supplement 1, S773–S778.

RUDY GUERRA

Meta-analysis of Clinical Trials

Meta-analysis is the systematic and quantitative review of the results of a set of individual studies, intended to integrate their findings [11]. Informal synthesis of evidence has always been practiced, and even the idea of combining results quantitatively across research studies can be traced back to the early 1900s. The basis of the statistical methods now generally used is also long established [6]. Meta-analysis as a specific technique was developed in the **social sciences**, but soon became adopted as a fundamental tool in medical research, especially as a way of reviewing and combining the evidence from **clinical trials**.

There are a number of reasons why meta-analysis is such an important technique in clinical trials research [22, 26]. It is now recognized that narrative reviews of a set of clinical trials can be very misleading, being potentially distorted by the selection of evidence, the emphasis placed on its components, and the personal opinion of the reviewer. Secondly, the explosion of research evidence, in the form of published trials, often cannot be easily assimilated without formal review. Thirdly, in assessing the benefits of a particular medical treatment, judgments should be based on the totality of reliable evidence, for example from all relevant well-conducted randomized clinical trials. Fourthly, given that sample sizes of individual clinical trials are often too small to detect clinically important effects reliably, synthesis of evidence across trials is necessary.

Meta-analysis therefore has a number of aims: to review systematically the available clinical trial evidence; to provide quantitative summaries of the results from each study; to combine these results across studies, if appropriate; and to provide an overall interpretation. By combining results, more statistical **power** for detecting treatment effects is available and the precision of estimated treatment effects is enhanced. Meta-analysis discourages the common simplistic and misleading interpretation that the results of individual clinical trials are in conflict because some are labeled "positive" (i.e. statistically significant) and others "negative" (i.e. statistically nonsignificant). Yet it allows the investigation of possible reasons for real differences between the treatment effects in different clinical trials, i.e. sources of heterogeneity.

Meta-analysis has both qualitative and quantitative components. For example, the description of the available trials, in terms of their relevance and methodologic strengths and weaknesses, plays a crucial part in providing a meaningful interpretation, but is essentially qualitative. Summarizing the trials' results and their combination is clearly quantitative. Various synonymous terms are used to describe meta-analysis. The term "overview" has been used by some authors, and "pooling" by others. Because the first fails to indicate the quantitative aspect of the analysis, and the second gives the unfortunate (and incorrect) impression that the data from each trial are simply lumped together, these now tend to be avoided. More recently, the term "systematic review" has been coined, and some authors use this to refer to the whole process of qualitative and quantitative review, restricting the term meta-analysis to the quantitative aspects.

In this article, we start by discussing the nature of practical meta-analyses, then focus on the quantitative and statistical aspects, and conclude by considering interpretational issues.

Types of Meta-Analysis

In principle, the trials included in a meta-analysis might be clinically homogeneous. For example, they might all study a similar type of patient for a similar duration with the same treatments in the two groups of each trial. In practice, however, the trials included are usually heterogeneous. For example, the **eligibility criteria** may differ between trials, the treatments used may not be identical, the duration of treatment and length of follow-up may differ, and the use of ancillary treatments or care may not be the same. Hence, in most situations, the objective of a meta-analysis cannot be equated with that of a single large trial, even if that trial has wide eligibility criteria. While a single trial focuses on the effect of a specific treatment in specific circumstances, a meta-analysis aims to obtain a more generalizable conclusion about the effect of a generic treatment policy in a wider range of situations. For example, a meta-analysis of blood cholesterol lowering trials included trials using

drugs, dietary intervention, and even surgery; the focus was on the effect of cholesterol reduction itself, rather than on the specific regimens used to lower cholesterol [31].

There has been enormous growth in the number of meta-analyses published in recent years [11] over all fields of medicine, so that examples can currently be found in almost any year's issue of a general or specialist medical journal. Some are investigations of published trials (meta-analysis of the literature), others use summary statistics or tabulated data obtained from the individual trialists (meta-analysis of summary data), while others are based on the individual patient data (IPD) from each trial (meta-analysis of patient data). The latter usually require the formation of international collaborative groups and long-term resource commitment; for example, see [1, 12].

While the majority of meta-analyses currently rely on published data, sometimes supplemented by additional summary data for certain trials, the proportion of those based on IPD is increasing. The use of IPD allows checking of the original data (whereas published information can be wrong or misleading), permits the updating of follow-up information (which is particularly useful in **survival** studies), and has much greater potential than summary data for investigating which patients may benefit most from treatment. For these reasons, IPD meta-analyses must be regarded as the gold standard [25]. However, they cannot always be carried out, either because individual patient data are not available (whether for practical, scientific, or political reasons) or because the exercise is too costly.

The worldwide **Cochrane Collaboration** has been a recent and important development promoting the production and dissemination of systematic reviews of high quality, in an effort to achieve **evidence-based medical practice**. The aims of the Collaboration are to facilitate meta-analyses of randomized clinical trials in all areas of medicine, to disseminate their results effectively, and to update them regularly. To this end, a number of Cochrane Centres have been established in different parts of the world, as well as Review Groups with responsibility for coordinating the work in specific medical areas. The meta-analyses are performed according to certain methodologic guidelines [24] and checked by editorial teams. They are then released in a standard format on compact disc together with searching software (the Cochrane database of systematic reviews – CDSR). Many meta-analyses now being undertaken contribute to the Cochrane Collaboration and the CDSR.

Preliminaries

A meta-analysis, like other scientific research, requires a written structure or protocol which defines its objective and scope. The identification of trials for inclusion in a meta-analysis is the next, but difficult, stage. Simply identifying relevant randomized trials through computerized bibliographic searches such as Medline is usually inadequate; it is often necessary to search reference lists and citations and to communicate with specialists in the area. One also has to decide whether to include abstracts, how to include data from trials in progress, and how to identify relevant non-English publications; all are difficult issues [24]. The decision as to whether particular trials are relevant to the objective of the meta-analysis can be somewhat subjective, and so one needs to be explicit about the criteria for including trials and to give a list of excluded trials.

A qualitative review of the individual trials, for example delineating their methodologic weaknesses, is a necessary step for a later overall interpretation of results. Failure to undertake proper **randomization**, to maintain follow-up of all patients, or in some circumstances to preserve blinding of treatments (*see* **Blinding or Masking**), can lead to **bias** in individual trials and hence in the overall meta-analysis [27]. Unfortunately, these weaknesses are not always clear from publications. One may even have to decide that the trials are not combinable, either because the quality of the reported results is poor or because the treatments or patients are not similar enough. The quality of a meta-analysis is only as good as the quality of the component trials which are included in it (except with regard to sample size): if the trials are biased, so will be the meta-analysis.

The essential quantitative information needed is the estimated treatment effect, measured on the same scale in each trial, and its **variance**. One common problem in undertaking a meta-analysis of published trials is that this information is not available. For example, treatment effects may be presented in different ways in different trials, particularly in survival studies, or their **standard errors** may not be available. Sometimes these can be derived from other published information, such as the raw numbers or

***P* values**, but often the original trialists have to be contacted to supply the necessary information.

Scale of Treatment Effects

Many trials have **binary** event data, such as death, as the outcome. Treatment effects can then be expressed as **risk** differences, risk ratios (*see* **Relative Risk**), or **odds ratios**. Any of these can be used as a basis for comparing results across trials, but because **absolute risks** will be very dependent on the underlying risk of the patients included and the duration of follow-up, risk differences are often severely heterogeneous across trials. Hence meta-analyses are usually based on relative measures, odds ratios being the most commonly used in practice because of the variety of statistical methods that are available. For similar reasons, hazard ratios (*see* **Hazard Rate**) are most often used for survival studies. For ordinal data a **proportional odds** method of analysis can sometimes be applied. A logarithmic **transformation** (log odds ratio, or log hazard ratio) is generally used, since this improves the **normality** of the estimate's **sampling distribution**.

In trials where the outcome is continuous, for example blood pressure, treatment effects are usually expressed as **mean** differences. However, some trials may express their results using change from baseline, proportionate change from baseline, baseline adjusted levels, a data transformation (such as logarithm), or **medians**. Not all of these methods are compatible, but a consistent manner of expressing results is required for meta-analysis. In some fields, such as psychology, it is common to express results as effect sizes, by dividing means by the overall **standard deviation** of the measurements. This is especially so when somewhat different outcome measures have been used to assess the same underlying quantity (*see* **Normal Scores**).

Statistical Tests Used in Meta-Analysis

It is now useful to introduce some notation. We suppose there are k trials, $i = 1, \ldots, k$, each with a treatment group and a **control** group, and that

θ_i = true treatment effect in trial i,

$\hat{\theta}_i$ = estimated treatment effect in trial i,

v_i = variance of $\hat{\theta}_i$,

$w_i = 1/v_i$.

For example, $\hat{\theta}_i$ could be the observed log odds ratio in a trial with binary outcomes, or the observed mean difference in a trial with continuous outcomes, and is an estimate of the true but unknown θ_i. In practice, v_i is an estimated quantity based on the data in each trial. In what follows, the summation sign always refers to summation over all trials $i = 1, \ldots, k$.

An overall **null hypothesis** that the treatment effect is zero in every trial is $H_0 : \theta_i = 0$ for all i. Two tests of this null hypothesis are [31]:

1. $\sum w_i \hat{\theta}_i^2$ referred to a $\boldsymbol{\chi_k^2}$ **distribution**; this is a general test similar to **Hotelling's $\boldsymbol{T^2}$**, which has no particular alternative hypothesis in mind.
2. $(\sum w_i \hat{\theta}_i)^2 / \sum w_i$ referred to a χ_1^2 distribution; this is a "directional" test more powerful than the general test above against alternatives of the form $H_1 : \theta_i < 0$ for all i (or $H_1 : \theta_i > 0$ for all i).

In the context of meta-analysis, the directional alternatives are those of most medical interest (and indeed are generally the most plausible) and so it is the directional test that is used. Moreover, the directional test is particularly powerful against the alternative $H_1 : \theta_i = \theta (\neq 0)$ for all i, that the true treatment effect in each trial is the same nonzero quantity. Using χ^2 distributions assumes that the estimated treatment effects are normally distributed.

For binary outcomes, each trial's results can be summarized as a **2 × 2 table** of counts. The **Mantel–Haenszel** test is in fact a particular example of the directional test. It has been rephrased by Peto [39] in terms of comparing the observed (O_i) number of events in the treated group of trial i with that expected (E_i) under the hypothesis of no treatment effect $(\theta_i = 0)$. Using the variance V_i of O_i from the **hypergeometric distribution** for a 2×2 table, the Mantel–Haenszel test of H_0 refers $[\sum (O_i - E_i)]^2 / \sum V_i$ (if a continuity correction (*see* **Yates's Continuity Correction**) is not used) to a χ_1^2 distribution. This can be seen to be (asymptotically) equivalent to the directional test by noting that $(O_i - E_i)/V_i$ is an approximation to the log odds ratio estimate (that is $\hat{\theta}_i$) with variance $1/V_i \approx v_i$ [38].

If this overall null hypothesis is rejected, an assumption often made in meta-analysis is that the

true treatment effects in each trial are the same nonzero quantity, θ (a **fixed effect** model). A test of the hypothesis $H_0 : \theta_i = \theta$ for all i is a test of homogeneity (or test for heterogeneity), achieved by referring $Q = \sum w_i(\hat{\theta}_i - \hat{\theta})^2$, where $\hat{\theta} = \sum w_i\hat{\theta}_i / \sum w_i$, to a χ^2_{k-1} distribution. This is a test of trial by treatment interaction, and like other tests of **interaction** lacks power in many practical situations. Hence, in particular, nonsignificance of the test cannot be taken as evidence in favor of treatment effect homogeneity, and so the test has somewhat limited usefulness in the context of meta-analysis.

If the assumption of homogeneity of true treatment effects (θ_i) is not accepted, either because of the above test or for reasons of principle, variation in the θ_i between trials must be accommodated (a **random effects** model). If it is assumed that the θ_i have some distribution across trials, with mean θ^* and variance σ^2, then we have a **hierarchical model** for the observed data:

$$\hat{\theta}_i \sim \text{mean } \theta_i, \text{ variance } v_i,$$
$$\theta_i \sim \text{mean } \theta^*, \text{ variance } \sigma^2.$$

A test of whether the "average" treatment effect θ^* is zero is similar to the directional test above, except that different weights are used. Now $(\sum w_i^*\hat{\theta}_i)^2 / \sum w_i^*$ is referred to a χ^2_1 distribution, where $w_i^* = 1/(v_i + \sigma^2)$. In using the χ^2 distribution, normality of both the estimated treatment effects within trials and the true treatment effects across trials is assumed. To carry out this test, the between-trial component of variance must be estimated (see below).

Table 1 shows the data from a meta-analysis of nine controlled trials of the use of diuretics in pregnancy to prevent preeclampsia [7, 33]. The estimated odds ratios and 95% **confidence intervals** for each trial are shown in Figure 1. Using a log odds ratio scale for the preeclampsia outcome, the directional test of $H_0 : \theta_i = 0$ for all i yielded $\chi^2_1 = 21.6 (P < 0.001)$. The conclusion is not that the treatment worked in all the trials, but only that at least one of the true treatment effects is nonzero. The test for heterogeneity (of $H_0 : \theta_i = \theta$ for all i) yielded $\chi^2_8 = 27.3 (P < 0.001)$, indicating that the true treatment effects were not the same in all trials. If one was prepared to use the hierarchical random effects model above, the test of $H_0 : \theta^* = 0$ yielded $\chi^2_1 = 6.4 (P = 0.01)$. Hence, allowing for the apparent heterogeneity of true treatment effects between trials reduced the evidence for an overall treatment effect. This emphasizes that a test of an assumed common treatment effect is different, both logically and in practice, from a test of an average treatment effect.

Table 1 Incidence of preeclampsia in nine randomized trials of diuretics, and odds ratios

Trial[a]	Incidence of preeclampsia (number of patients)		OR
	Diuretic	Control	
Weseley	11% (14/131)	10% (14/136)	1 · 04
Flowers	5% (21/385)	13% (17/134)	0 · 40
Menzies	25% (14/57)	50% (24/48)	0 · 33
Fallis	16% (6/38)	45% (18/40)	0 · 23
Cuadros	1% (12/1011)	5% (35/760)	0 · 25
Landesman	10% (138/1370)	13% (175/1336)	0 · 74
Krans	3% (15/506)	4% (20/524)	0 · 77
Tervila	6% (6/108)	2% (2/103)	2 · 97
Campbell	42% (65/153)	39% (40/102)	1 · 14

[a]Principal author, referenced by Collins et al. [7].

Fixed Effect Methods of Estimation

It is more informative to provide overall estimates (*see* **Estimation**) of treatment effect, together with confidence intervals, than simply to test hypotheses. The vast majority of published meta-analyses use "fixed effect" estimates of treatment effect, making the assumption of homogeneity ($\theta_i = \theta$ for all i). Whether this is reasonable is discussed below, but for the moment we pursue this analysis.

In estimating an assumed common θ, the weighted average $\hat{\theta} = \sum w_i\hat{\theta}_i / \sum w_i$ is an **unbiased** estimate, and has the smallest variance, namely $1/\sum w_i$, amongst weighted averages of the $\hat{\theta}_i$ (*see* **Minimum Variance Unbiased (MVU) Estimator**). Assuming normality allows the calculation of a 95% confidence interval for θ, being $\hat{\theta} \pm 1.96(1/\sum w_i)^{1/2}$. In this "inverse-variance" weighting, the greatest weight is given to the largest and most informative trials, which have more precisely estimated treatment

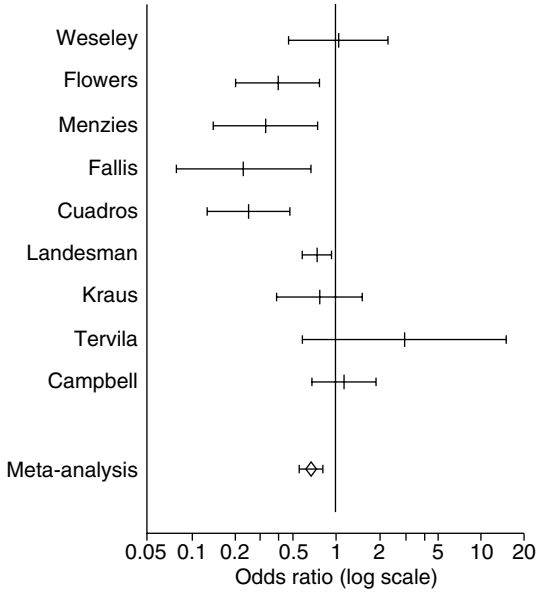

Figure 1 ORs for preeclampsia and 95% confidence intervals in nine trials of diuretics (ORs less than unity represent beneficial effects of diuretics; meta-analysis based on fixed-effect assumption). Reproduced from Thompson & Pocock, *Lancet*, vol. 338, pp. 1127–1130 [33]. © The Lancet Ltd, 1991.

effects (small v_i, so large w_i). The estimate is not derived simply from lumping together the data from each trial, but appropriately from a stratified analysis (*see* **Stratification**) by combining trial-specific estimates. For trials with binary outcomes, this method can be used directly for log odds ratios; the use of empirical logits, for example adding 0.5 to the cells of 2×2 tables which contain a zero, is then necessary. Asymptotically equivalent results can be obtained using **logistic regression** [31] or Peto's method based on $(O - E)/V$ as an approximation to the log odds ratio [7, 39], although the latter can be biased in some extreme situations [16]. Each of these methods gives an overall log odds ratio and confidence interval, which can be exponentiated to provide results on the untransformed odds ratio scale. The Mantel–Haenszel estimator weights the untransformed odds ratios approximately inversely proportional to their variances. In numerical examples all these methods produce nearly identical results [13, 31]. A general formulation of these and similar methods can be made in terms of score statistics and Fisher **information** [38].

The main practical issue is not the choice of particular method, but whether the assumption of homogeneity of true treatment effects on which these methods are all based is justified. It is also informative to calculate the proportion of weight allocated to each trial ($w_j / \sum w_i$ for trial j). Often it is just one or a few trials that dominate the overall results, in which case one may be justly concerned about the generalizability of the results [33].

Random Effects Methods of Estimation

The homogeneity assumption can be relaxed by using the hierarchical model given above, which incorporates a between-trial component of variance σ^2. This leads to estimation of the overall average treatment effect θ^* around which, it is assumed, the true treatment effects in each trial are randomly distributed. The variance of each $\hat{\theta}_i$ is now $(v_i + \sigma^2)$ and the appropriate weights w_i^* are put equal to the reciprocal of these variances. The average effect of treatment is estimated as $\hat{\theta}^* = \sum w_i^* \hat{\theta}_i / \sum w_i^*$, with variance $1/\sum w_i^*$. Assuming normality both within and between trials allows the calculation of a confidence interval for θ^*.

To carry out these calculations, σ^2 must be estimated. Usually a **moment** estimator is used since it can be derived straightforwardly from the heterogeneity statistic Q [10]. In principle, **maximum likelihood** estimation is preferable but this requires an iterative solution [10, 17]. In most practical situations, the choice between estimators is unimportant. The important aspect is that σ^2 is now allowed to be positive, permitting between-trial heterogeneity of true treatment effects; when σ^2 is zero the random effects method becomes identical to the fixed effect method. In moving from a fixed effect analysis to a random effect analysis, the weights given to each trial become more evenly distributed, so that in a random effects analysis small trials receive relatively more weight.

Related Methods

Use of $1/\sum w_i$ as the variance of $\hat{\theta}$ assumes that the individual v_i are known rather than estimated. Use of $1/\sum w_i^*$ as the variance of $\hat{\theta}^*$ assumes in addition that σ^2 is known. Allowance for the imprecision in the estimate of σ^2 can be made using a marginal

profile method [17], but allowing also for the imprecision in v_i requires exact methods for fixed effect analyses [21] or a full **likelihood** approach for random effect analyses [36]. These extensions can also be used in deriving significance tests. The construction of confidence intervals for θ and θ^* requires assumptions of normality, which can be investigated using normal plots of $w_i^{1/2}(\hat{\theta}_i - \hat{\theta})$ or $w_i^{*1/2}(\hat{\theta}_i^* - \hat{\theta}^*)$, respectively. However, such techniques are of limited usefulness when only a few trials are included in the meta-analysis.

As well as estimating an overall treatment effect, there can often be interest in estimating the true treatment effect θ_i in each trial. These are most easily obtained as **empirical Bayes** estimates, which are shrunk (*see* **Shrinkage**) from the usual (maximum likelihood) estimate towards the overall average treatment effect by a factor which depends on the ratio of within-trial to between-trial variances [30]. Confidence intervals for these estimates can also be calculated. In the case of the fixed effect model, the estimates for each trial equal the overall estimated treatment effect, so that there is complete shrinkage. As σ^2 increases, the posterior empirical Bayes estimates of the true treatment effects become more widely spread.

Fully **Bayesian** approaches for random effects meta-analysis are now also computationally feasible [30]. These can be viewed as hierarchical models, where the conditional independence between parameters is used in deriving an appropriate graphical model. **Priors** have to be set on certain parameters, such as θ^* and σ^2. These can be uninformative, but some authors argue for the use of informative priors for σ^2 derived on the basis of other related meta-analyses [18]. Fully Bayesian approaches utilizing **Markov chain Monte Carlo methods** (Gibbs sampling) can easily extend the usual normality assumptions for true treatment effects to allow for a heavier tailed distribution, such as a **Student's *t* distribution**.

Methods have been developed for more complex situations, for example to include trials with more than two treatment groups [18, 31] or to combine randomized clinical trial evidence with that from **non-randomized studies** [23]. Methods for simultaneous consideration of multiple outcomes (*see* **Multiple Endpoints, *P* Level Procedures**) have been proposed [2], as have techniques for cumulative meta-analysis [19]. The latter have been used to show how appropriate changes in medical practice could have occurred earlier if meta-analyses had been performed and their conclusions implemented. Meta-analysis techniques have also been applied to the analysis of **multicenter trials** [17], to the analysis of paired cluster randomized trials [34] (*see* **Group-randomization Designs**), and to the evaluation of **surrogate endpoints** [9].

Interpretation

The fixed effect method of estimation of a common treatment effect yields a narrower confidence interval than the random effects estimation of an average treatment effect when there is heterogeneity in results between trials. For example, in the nine diuretic trials to prevent preeclampsia in pregnancy (Table 1, Figure 1), the estimated overall odds ratios and 95% confidence intervals were 0.66 (0.57–0.79) and 0.60 (0.40–0.89), respectively, using the two methods. By incorporating the between-trial component of variability into the analysis, the inference about the magnitude of the treatment effect is appropriately less certain. The simplistic assumption of a common treatment effect in all trials used in a fixed effect analysis ignores this potential extra source of variability and can lead to overdogmatic interpretation [33]. Since the trials in a meta-analysis are almost always clinically heterogeneous, it is to be anticipated that to some extent their quantitative results will be statistically heterogeneous [32]: there will tend to be more variation between the results of the trials than is simply compatible with chance, even though a test for heterogeneity may not be formally statistically significant. Hence a random effects model appears more justified in practice than a fixed effect model in terms of making **inferences** which apply to future trials or patients.

Not all agree with this argument, however, and Peto, in particular, has argued strongly for a fixed effect approach [12]. His view appears to be based on hypothesis testing rather than estimation, with the ***P*-value** indicating the extent to which the results could have arisen simply through the play of chance with respect to the randomization process in each trial. Combined with the common-sense and empirical view that qualitative trial by treatment interactions (where the direction of true treatment effects differs across trials) are unlikely, the fixed effect

estimate of treatment effect is interpreted only as some "typical" treatment effect for the trials that have been performed, rather than more formally as a basis for inference to future trials or patients. Indeed, how the results of a meta-analysis can be used to inform decisions about treating individual patients in clinical practice is not simple, usually requiring assumptions about generalizability or **extrapolation** which are based on very little evidence.

The random effects method does not completely solve the problem of heterogeneity. It relies, for example, on the simplistic assumption that the heterogeneity between trials can be represented by a single variance. Moreover, the idea that the available trials have true treatment effects which are drawn from some distribution is unrealistic, and provides only a convenient way of allowing for unexplained variation between them. However, heterogeneity can be regarded as an asset rather than a problem. It allows clinically and scientifically more useful approaches attempting to investigate how potential sources of heterogeneity impact on the overall treatment effect [32], as discussed below.

Sources of Heterogeneity

Rather than estimate a single treatment effect, it is possible to investigate whether certain trial characteristics, such as drug dose, duration of treatment, or length of follow-up are related to the treatment effects observed. For example, in analyzing the effects of BCG vaccination on the risk of tuberculosis, the effect of the geographic latitude of each trial was investigated [3]. In the blood cholesterol lowering trials, the reduction in heart disease risk was related to the extent and duration of cholesterol reduction [32]. In **AIDS** trials, the reduction in mortality was related to the change in CD4 counts [9]. Ideally the **covariates** used in such analyses should be specified in advance to reduce the risk of post hoc conclusions prompted by inspecting the available data; as in subgroup analyses for individual clinical trials (*see* **Treatment-covariate Interaction**), there is a danger of **false positive** results. Similar analyses are also possible for patient characteristics, provided that individual patient data are available. The statistical purpose of such analyses is to see to what extent, and with what certainty, covariates can explain the between-trial component of variance. Hence a mixed effects model is obtained, where some or possibly all of the "random" between-trial variation is explained by fixed covariates. The medical outcome should be a better scientific understanding of the data and more useful clinical conclusions on which to base decisions about medical interventions [32].

It can be useful to calculate the contribution of each trial to the heterogeneity statistic Q, i.e. $w_i(\hat{\theta}_i - \hat{\theta})^2$ for trial i, which under the assumption of homogeneity should have approximate χ^2_1 distributions [31] (*see* **Agreement, Measurement of**). This can show that the results of just one or a few trials are anomalous, and suggest that either particular clinical aspects or methodologic biases may be the cause. In analyses which relate the observed treatment effects to covariates, weighted **regression** is appropriate. The weights used must, however, reflect both the within-trial variances and the residual heterogeneity, i.e. variability between trial results that is not explained by the covariates [3].

One particular potential source of heterogeneity, the underlying risk of the patients in the trial, has often been investigated recently, for a number of reasons. That the overall treatment effect, for example the odds ratio, does not depend on this underlying risk is a strong assumption [4]. Moreover, if there were a relationship with underlying risk, such an analysis would help identify the patients who were likely to benefit most from a treatment, or whether some groups of patients might even be disadvantaged. Such analyses also have **health economic** consequences for the appropriate use of the treatment. However, in most meta-analyses, the only available measure of underlying risk is the observed rate of events in the control group of each trial. In relating this to the observed treatment effect in each trial, there can be a severe **bias** stemming from **regression toward the mean** [28]. Correct analyses are more difficult and are based on certain assumptions, but take into account the sampling variation in the observed control group risk estimates [20, 35]. The future of such analysis is to use individual patient data to relate treatment benefit to measured patient covariates (or a prognostic score) rather than the unmeasured "underlying risk", so that clinically useful results can be obtained.

Presentation

There is a move towards ensuring that the quality of reported meta-analyses is as high as possible [8], especially since they underpin the practice of evidence-based medicine. Presentation is one component of this. Meta-analysis in medical journals usually contains a brief tabular description of the characteristics of the trials included, and of their principal quantitative results (such as Table 1). In addition, figures have been developed to show the main results diagrammatically. Typically these comprise the estimates and 95% confidence intervals for each trial, together with an overall estimate and confidence interval (such as Figure 1). Such diagrams are sometimes called "forest plots". Some authors use 99% confidence intervals for individual trials to offset the fact that many trials are being displayed simultaneously. Others use filled squares or circles to represent each estimate, with areas proportional to the inverse of the variance, so that the eye is drawn not to the trials with wide confidence intervals which are least informative but towards those with greatest precision (see, for example, [1]). Overall effects from survival studies, analyzed as log hazard ratios, may be transformed to approximate overall survival curves to convey the clinically important messages clearly [12].

Such diagrams are not very useful for revealing heterogeneity, and "radial plots" of the standardized treatment effect $\hat{\theta}_i / v_i^{1/2}$ against reciprocal standard error $1/v_i^{1/2}$ have been advocated instead [31]. In the presence of heterogeneity, separate results according to the relevant covariates need to be shown. In the case of a continuous covariate, this can also be represented diagrammatically (see, for example, [32]).

Biases and Sensitivity Analyses

Not all trials are free from bias, nor are all trials analyzed on an **intention to treat** basis. In the absence of properly conducted randomization, adequate blinding, or complete follow-up, the bias in individual trial results can lead to bias in the overall results of a meta-analysis. The problems are even more severe in the meta-analysis of epidemiologic studies, since **confounding** factors may be a major issue. Even with full intention to treat analysis of clinical trials, the degree of patient **compliance** may have differed across trials, leading to problems in the overall interpretation of a meta-analysis.

The possible consequences of such biases need to be addressed, albeit imperfectly in practice. For example, the effect of randomization concealment can be investigated as a potential source of heterogeneity [27]. Alternatively, the methodologic quality of the trials can be rated, and this used as a rationale for excluding the weakest trials, or as a covariate in analysis [15]. Individual patient data allow much more scope than published data or summary data for the more definitive identification of poor quality trials, or the correction of errors in the publication of trial results [25].

Publication bias has become a particular concern in meta-analysis, since the aim is to obtain a summary of the totality of evidence. Even though large **multicenter trials** will tend to be published whatever their results, small studies may be published only if they have impressive observed results. Hence the published literature is potentially biased. Empirically, a "funnel plot" of the observed treatment effects against sample size (or precision) may show some evidence that small negative studies are missing from the published literature [37]. In the presence of such evidence, a random effects model is not necessarily desirable since it puts more weight than a fixed effect model on the small studies which are available. Various methods have been proposed to calculate the number of negative unpublished studies that would be needed to overturn the qualitative conclusion of a meta-analysis [14]. More constructively for the future, registers of clinical trials which are in progress, or are funded or have been approved by ethical committees, are being kept so that the completeness of the published literature can be assessed directly, and unpublished information sought [29].

In the presence of imperfect information, **sensitivity analysis** is a useful tool, for example to investigate the extent to which conclusions change if methodologically weak trials are excluded. They can also be used to investigate the **robustness** of conclusions to the inclusion criteria for trials originally adopted. Where one or just a few trials dominate the overall results of a meta-analysis, sensitivity analyses can be conducted to see how the conclusions change when these trials are omitted. Indeed, the random effects method can be viewed as a sensitivity analysis for the failure of the assumption of

homogeneity [31]. While these analyses are useful in judging the effects of the more subjective decisions taken in conducting a meta-analysis, it can sometimes be difficult to form an overall interpretation from the results of many sensitivity analyses, especially when the number of trials included in a meta-analysis is small.

The Future of Meta-Analysis

Meta-analysis has rightly had a major impact on medical science in the past 10 years, and has formed the basis for the development of evidence-based medical practice. The statistical basis of meta-analysis is well developed. Although the fixed vs. random effects discussion will linger on, it seems likely to be superseded by greater use of mixed models in which potential sources of heterogeneity are directly investigated. Although there will continue to be development of statistical methods at the margin, the main need is now more practical. The majority of meta-analyses are currently still based on published data. Hence, for example, empirical work is required on the practical problems in meta-analyses of imperfect published data, and on delineating appropriate sensitivity analyses that will aid interpretation. Statistical support for the many meta-analyses undertaken is woefully limited, resulting in poor quality in some cases and consequent but unjustified disillusionment with the underlying idea. Alternative methods for the dissemination of meta-analytic results in order to affect clinical practice need to be assessed [5]. The use of individual patient data meta-analyses is likely to expand in the future, since it gives more scope both for including only reliable evidence and for allowing more detailed investigation of results, especially with regard to heterogeneity of treatment effects according to patient characteristics.

References

[1] Antiplatelet Trialists' Collaboration (1994). Collaborative overview of randomized trials of antiplatelet therapy. I. Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients, *British Medical Journal* **308**, 81–106.

[2] Berkey, C.S., Anderson, J.J. & Hoaglin, D.C. (1996). Multiple-outcome meta-analysis of clinical trials, *Statistics in Medicine* **15**, 537–557.

[3] Berkey, C.S., Hoaglin, D.C., Mosteller, F. & Colditz, G.A. (1995). A random-effects regression model for meta-analysis, *Statistics in Medicine* **14**, 395–411.

[4] Brand, R. & Kragt, H. (1992). The importance of trends in the interpretation of an overall odds ratio in a meta-analysis of clinical trials, *Statistics in Medicine* **11**, 2077–2082.

[5] Chalmers, I. (1991). Improving the quality and dissemination of review of clinical research, in *The Future of Medical Journals: in Commemoration of 150 years of the British Medical Journal*, S. Lock, ed. British Medical Society, London, pp. 127–146.

[6] Cochran, W.G. (1954). The combination of estimates from different experiments, *Biometrics* **10**, 101–129.

[7] Collins, R., Yusuf, S. & Peto, R. (1985). Overview of randomized trials of diuretics in pregnancy, *British Medical Journal* **290**, 17–23.

[8] Cook, D.J., Sackett, D.L. & Spitzer, W.O. (1995). Methodologic guidelines for systematic reviews of randomized controlled trials in health care evaluation from the Potsdam consultation of meta-analysis, *Journal of Clinical Epidemiology* **48**, 167–171.

[9] Daniels, M.J. & Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers, *Statistics in Medicine* **16**, 1965–1982.

[10] DerSimonian, R. & Laird, N.M. (1986). Meta-analysis in clinical trials, *Controlled Clinical Trials* **7**, 177–188.

[11] Dickersin, K. & Berlin, J.A. (1992). Meta-analysis: state-of-the-science, *Epidemiologic Reviews* **14**, 154–176.

[12] Early Breast Cancer Trialists' Collaborative Group. (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy, *Lancet* **339**, 1–15, 71–85.

[13] Fleiss, J.L. (1993). The statistical basis of meta-analysis, *Statistical Methods in Medical Research* **2**, 121–145.

[14] Gleser, L.J. & Olkin, I. (1996). Models for estimating the number of unpublished studies, *Statistics in Medicine* **15**, 2493–2507.

[15] Greenland, S. (1994). A critical look at some popular meta-analytic methods, *American Journal of Epidemiology* **140**, 290–296.

[16] Greenland, S. & Salvan, A. (1990). Bias in the one-step (Peto) method for pooling study results, *Statistics in Medicine* **9**, 247–252.

[17] Hardy, R.J. & Thompson, S.G. (1996). A likelihood approach to meta-analysis with random effects, *Statistics in Medicine* **15**, 619–629.

[18] Higgins, J.P.T. & Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis, *Statistics in Medicine* **15**, 2733–2749.

[19] Lau, J., Antman, E.M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F. & Chalmers, T.C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction, *New England Journal of Medicine* **327**, 248–254.

[20] McIntosh, M.W. (1996). The population risk as an explanatory variable in research synthesis of clinical trials, *Statistics in Medicine* **15**, 1713–1728.

[21] Mehta, C.R. & Walsh, S.J. (1992). Comparison of exact, mid-P, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables, *American Statistician* **46**, 146–50.

[22] Mulrow, C.D. (1994). Rationale for systematic reviews, *British Medical Journal* **309**, 597–599.

[23] Olschewski, M., Schumacher, M. & Davis, K.B. (1992). Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design, *Controlled Clinical Trials* **13**, 226–239.

[24] Oxman, A. (1994). Preparing and maintaining systematic reviews, in *Cochrane Collaboration Handbook*, Section VI, D. Sackett, ed. Cochrane Collaboration, Oxford.

[25] Oxman, A.D., Clarke, M.J. & Stewart, L.A. (1995). From science to practice: meta-analyses using individual patient data are needed, *Journal of the American Medical Association* **274**, 845–846.

[26] Peto, R. (1987). Why do we need systematic overviews of randomized trials?, *Statistics in Medicine* **6**, 233–240.

[27] Schulz, K.F., Chalmers, I., Hayes, R. & Altman, D.G. (1995). Empirical evidence of bias: dimensions of methodologic quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association* **273**, 408–412.

[28] Sharp, S.J., Thompson, S.G. & Altman, D.G. (1996). The relation between treatment benefit and underlying risk in meta-analysis, *British Medical Journal* **313**, 735–738.

[29] Simes, R.J. (1986). Publication bias: the case for an international registry of clinical trials, *Journal of Clinical Oncology* **4**, 1529–1541.

[30] Smith, T.C., Spiegelhalter, D.J. & Thomas, A. (1995). Bayesian approaches to random effects meta-analyses: a comparative study, *Statistics in Medicine* **14**, 2685–2699.

[31] Thompson, S.G. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction, *Statistical Methods in Medical Research* **2**, 173–192.

[32] Thompson, S.G. (1994). Why sources of heterogeneity in meta-analysis should be investigated, *British Medical Journal* **309**, 1351–1355.

[33] Thompson, S.G. & Pocock, S.J. (1991). Can meta-analysis be trusted?, *Lancet* **338**, 1127–1130.

[34] Thompson, S.G., Pyke, S.D.M. & Hardy, R.J. (1997). The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques, *Statistics in Medicine* **16**, 2063–2077.

[35] Thompson, S.G., Smith, T.C. & Sharp, S.J. (1998). Investigating underlying risk as a source of heterogeneity in meta-analysis, *Statistics in Medicine*, to appear.

[36] Van Houwelingen, H.C., Zwinderman, K.H. & Stijnen, T. (1993). A bivariate approach to meta-analysis, *Statistics in Medicine* **12**, 2273–2284.

[37] Vandenbroucke, J.P. (1988). Passive smoking and lung cancer: a publication bias?, *British Medical Journal* **296**, 391–392.

[38] Whitehead, A. & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials, *Statistics in Medicine* **10**, 1665–1677.

[39] Yusuf, S., Peto, R., Lewis, J., Collins, R. & Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials, *Progress in Cardiovascular Diseases* **27**, 335–371.

Bibliography

Chalmers, I. & Altman, D.G., eds (1995). *Systematic Reviews*. BMJ Publishing Group, London.

Cooper, H. & Hedges, L.V., eds (1994). *The Handbook of Research Synthesis*. Sage, Newbury Park.

Eddy, D.M., Hasselblad, V. & Shachter, R. (1992). *Meta-Analysis by the Confidence Profile Method*. Academic Press, San Diego.

Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, London.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis*. Sage, Newbury Park.

Light, R.J. & Pillemer, D.B. (1984). *Summing Up: the Science of Reviewing Research*. Harvard University Press, Cambridge, Mass.

Pettiti, D. (1994). *Meta-Analysis, Decision Analysis and Cost Effectiveness Analysis*. Oxford University Press, Oxford.

Statistical Methods in Medical Research (1993). **2**, 117–192.

Statistics in Medicine. (1987). **6**, 217–409.

Wolf, F.M. (1985). *Meta-analysis: Quantitative Methods for Research Synthesis*. Sage, Beverly Hills.

(*See also* **Meta-analysis of Diagnostic Tests**; **Combining *P* Values**)

SIMON G. THOMPSON

Meta-analysis of Diagnostic Tests

If we wish to know the accuracy (e.g. **sensitivity** and **specificity**) of a diagnostic test, there may be several appropriate studies from which to derive estimates. First, we need to identify the pool of relevant studies. Secondly, we need to appraise the methodological quality of each of the studies and select those which reach a minimum standard, or use the quality items as predictors in a **regression** model. Thirdly, we need to derive a summary estimate from these selected studies. Such a meta-analysis will have two advantages: (i) it will provide greater precision in the estimates of accuracy (sensitivity and specificity) than from a single study, and (ii) it will allow an examination of the existence and the causes of heterogeneity between studies, for example, because of differing patient populations, test methods, or quality [9].

In this article, we deal with these three steps in performing such a systematic review.

Finding the Studies

Finding all of the relevant primary studies is not a trivial task. A standard strategy would include the search of a computerized database (including the identification of any review articles that might identify studies not picked up in the searches), hand searching of selected journals, and checking of the references of all of the relevant studies identified [4]. Electronic databases can be searched to find studies in which both the name of the tests and the disease of interest appear in the title or abstract. If this process results in a large number of articles that are not relevant, some means is needed of confining the search to those studies looking at diagnostic accuracy. This can be achieved, though at the risk of missing some relevant papers, by limiting the search to papers that contain or are indexed using methodological terms that identify diagnostic test studies. For example, in MEDLINE, the library database of the National Library of Medicine, it has been shown that most good diagnostic test articles would be detected by a search of the terms accuracy (appearing in the title or abstract) OR sensitiv* OR diagnos* appearing in the title, abstract or MEDLINE-indexed keywords [6]. The * indicates a wildcard so that it includes, for example, sensitivity and sensitivities. How to specify the wild card "*" varies with the MEDLINE interface. A set of these methodological filters is described and available for use electronically in the PUBMED version of MEDLINE (`http://www.pubmed.gov/query/static/clinical.html`).

Because studies of diagnostic tests may sometimes be done using routinely collected data, and are therefore less resource intensive than clinical trials, the likelihood of many unpublished studies, and hence of significant publication bias, is higher. (*see* **Meta-analysis of Clinical Trials**)

Quality Appraisal

The assessment of study quality can be used either to limit the meta-analysis to those studies of better quality or to explore the extent to which elements of study quality are related to the results. The assessment of methodological quality can be biased by the reviewer's preconceptions about what the results of the study should show. Therefore, the appraisal of primary studies should be based on a prespecified set of criteria that indicate objective standards of quality and be based on assessment by two independent reviewers with resolution of disagreements by consensus or the use of a third reviewer. On the basis of conceptual considerations and empirical evidence about the effect of elements of study quality on study results [13], several sets of criteria are now available [1, 25]. Common important elements in the criteria are:

1. Were the tests compared with a valid reference standard?
2. Were the test and reference standard measured independently (blind) of each other? Categories consist of:
   (i) test measured independently of reference standard and reference standard independently of test (MOST VALID);
   (ii) test measured independently of reference standard but not vice versa;
   (iii) reference standard measured independently of test but not vice versa;
   (iv) test and reference standard not measured independently of each other (LEAST VALID).

3. Was the choice of patients who were assessed by the reference standard independent of the test's results? (Avoidance of verification bias – this occurs when, for different test outcomes, the fraction subjected to the reference standard varies.)
4. If tests were compared, were they read independently on all individuals, or were different tests randomly allocated to study participants?

Issues to Consider for Applicability (Exploring Heterogeneity)

To help decide which studies to include and the extent to which heterogeneity in test accuracy should be explored, meta-analysts should also consider the potential sources of heterogeneity [7], of which the following are some of the major issues:

sequence: how the test(s) of interest are being used in the sequence of available tests;

role: whether tests are being evaluated as replacements for existing tests or additional tests;

test type: where there are several tests, ascertaining differences in accuracy may be the objective of the meta-analysis [21];

population and setting: for example, whether the test is being performed in primary care setting on people who present for the first time with a set of symptoms, or in a hospital setting after having been through a "referral filter", which would have excluded people with milder forms of disease, people who responded to first-line treatment, or people who were easier to diagnose.

Combining Studies

To combine studies, we must choose summary measures to represent each study, with sensitivity and specificity being commonly used. However, studies differ in their "threshold" for calling a test positive. For example, if the test-reader is very concerned about missing a disease case, they may choose a very low threshold point along the spectrum from negative to positive, and thus favor high sensitivity over high specificity. The **Receiver Operator Characteristic (ROC) curve** plot (see Figure 1) shows this variation in threshold by plotting sensitivity versus specificity (or more precisely, 1 – specificity). Thus, plotting the data in ROC space should be used as an initial

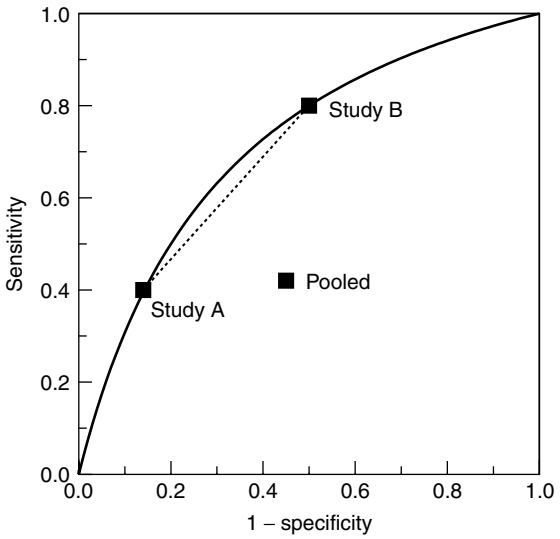

Figure 1 The receiver operating curve of two hypothetical studies, *A* and *B*

summary of all studies. We then require a method of combining studies, which accounts for both the discrimination ability of the test and this variation in threshold.

There are two widely used methods of combining the results of studies of diagnostic accuracy [8]. The first and worse of these is direct pooling. In Table 1, study A has a high specificity, but few nondiseased cases; study B has a high sensitivity but few diseased cases. Although the **odds ratio** in each individual study is 4 (indicating a reasonable discrimination ability), the overall odds ratio is 0.87 (indicating a worse-than-useless test). As illustrated in Table 1 and Figure 1, direct pooling can result in large distortions of the true accuracy of a diagnostic test. This is because of **confounding** from different thresholds being applied to studies in which there are different disease prevalences. The size and direction of this distortion is unpredictable, but it is particularly likely to be a problem if there is a wide range of prevalences of disease across the different diagnostic studies.

An alternate method that has been suggested to avoid this problem is to calculate the sensitivity and specificity within each study first, then calculate a (weighted) average of the sensitivities and, separately, calculate a (weighted) average of the specificities [2, 17]. This avoids the confounding problem associated with direct pooling, but may still lead to

Table 1 Results from two hypothetical studies (A and B), plus pooled results

Test	D	ND	Sensitivity	Specificity	Odds ratio	Youden
+	200	10				
−	300	60				
Total A	500	70	0.4	0.86	4	0.26
+	20	200				
−	5	200				
Total B	25	400	0.8	0.5	4	0.3
+	220	210				
−	305	260				
Total A + B	525	470	0.42	0.55	0.89	−0.03

Youden = Sensitivity + Specificity − 1.

an underestimation of the true accuracy if there is variation in the threshold used by different studies, that is, there is evidence of an association between sensitivity and specificity across studies. For example, Figure 1 shows the two studies from Table 1; any averaging of the sensitivity or specificity will lie along the dotted line joining A and B, with the location depending on the relative weighing used. If the same weights are used for sensitivity and specificity (e.g. equal weights or weights based on study size), then this is one point on the line. If different weights are used to combine sensitivity and specificity (e.g. based on disease numbers for sensitivity and nondiseased numbers for specificity), then these may be read from different points on the line, and hence the joint estimate is not confined to the line.

Combining Dichotomous Test Results Via SROC

There are a number of appropriate meta-analytic techniques for diagnostic tests, most of which plot the sensitivity against specificity for each of the primary studies, and then attempt to construct a summary Receiver Operator Characteristic curve (SROC) through these data points [8]. We will deal first with dichotomous tests, then outline briefly methods for tests with continuous or ordinal results.

For dichotomous tests, one convenient presentation is to use the odds ratio as a summary measure of the discrimination ability of a test. If the odds ratio is constant across different thresholds, then this will lead to a symmetric SROC curve. However, this is an assumption that can and should be tested. Before going into the details of these methods, to understand their interpretation better, we first look at the relationship between odds ratios, **likelihood ratios**, and the sensitivity and specificity.

Besides characterizing the accuracy of a test, the other use of sensitivity and specificity is in the calculation of **predictive values** via **Bayes' Theorem**. One convenient formulation of this is the odds–likelihood ratio version of Bayes' Theorem, namely,

$$\text{posttest odds} = \text{pretest odds} \times \text{likelihood ratio}, \quad (1)$$

where the likelihood ratio (LR) is $\Pr(result|disease)/\Pr(result|non\ disease)$. Thus, for a positive result, the $LR^+ = \Pr(+ve|D)/\Pr(+ve|non\ D) = sensitivity/(1 - specificity)$; similarly, for the negative result the $LR^- = \Pr(-ve|D)/\Pr(-ve|non\ D) = (1 - sensitivity)/specificity$. Hence, the odds ratio, OR, can be expressed as

$$\begin{aligned} OR &= \frac{sensitivity}{(1 - sensitivity)} \Big/ \frac{(1 - specificity)}{specificity} \\ &= \frac{LR^+}{LR^-} \end{aligned} \quad (2)$$

The log-odds ratio (using natural logarithm) can be expressed as

$$\begin{aligned} A &= \text{logit}(sensitivity) - \text{logit}(1 - specificity) \\ &= \log OR = \log LR^+ - \log LR^- \\ &= \log LR^+ + \frac{1}{\log LR^-} \end{aligned} \quad (3)$$

Since the LRs are measures of the **power** of the test to change the pretest odds, the OR can be viewed as summarizing the total discrimination ability of the test. Thus, A ($= \log OR$) is a measure of a test's discrimination ability (note that the other common measure is the area under the ROC). If the positive result and negative result change the odds equally, that is, $LR^+ = 1/LR^-$, then the threshold

is not skewed (or "biased" in ROC jargon) toward either diagnosis. The implicit test threshold may be represented by:

$$S = \text{logit}(\textit{sensitivity}) + \text{logit}(1 - \textit{specificity}) \quad (4)$$

Note that if sensitivity = specificity, then $S = 0$ and also $LR^+ = 1/LR^-$, indicating no skew in favor of either false positive or false negative rates. Thus for each study, we can obtain a measure of the discrimination ability, A (as measured by the log-odds ratio) and the threshold (as measured by S). S can be regarded as a proxy for test threshold, since it is the sum of the log odds of a positive test result in the diseased and the log odds of a positive test result in the nondiseased groups. Hence, S will increase as the criterion for a positive test becomes less stringent.

The *OR* may increase or decrease with the threshold, and hence lead to an asymmetric SROC curve. Therefore, Moses and coworkers [14, 18] suggest plotting A against S to check whether the discrimination ability varies with the threshold. Regression lines may be estimated in several ways, usually by weighted or unweighted **least squares**. The weighted analysis weights each study by the inverse variance of the log *OR* for that study ($\text{var}(\log OR) \approx 1/a + 1/b + 1/c + 1/d$, where $a - d$ are the observed counts in the four cells of the **2 x 2 table**); 0.5 is added routinely to all cell counts for all tables when computing the log *OR* and corresponding weight. A **robust regression** may also be used when assumptions are not met for the least squares analysis.

If the regression of A on S shows that the slope is not significantly different from zero, then the SROC may be considered symmetric, the constant *Bo* represents log *OR*, and hence exp *(Bo)* represents the odds ratio. If A does not depend on S, then any of the standard techniques for combining odds ratios can be used, for example, **Mantel–Haenzsel** if a fixed effect is assumed for test accuracy across studies or DerSimonian–Laird if **random effects** are assumed [8]. However, if the slope (coefficient of S) is significantly different from zero; the discrimination ability of the test varies with test threshold, and thus, the SROC is asymmetric. Model coefficients can be used to estimate the area under the SROC [24].

The Moses–Littenberg and SROC plots are illustrated in Figures 2 and 3, respectively, for ultrasound of the carotid arteries (data taken from Table 2 of Hasselblad & Hedges [5]). In Figure 2 is shown a log *OR* ranging between approximately 3 and 7.

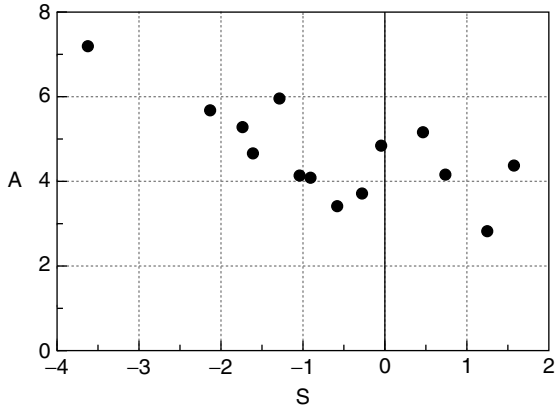

Figure 2 A plot of discrimination ability A [log (odds ratio)] versus S (a measure of threshold) for studies of ultrasound for carotid artery stenosis

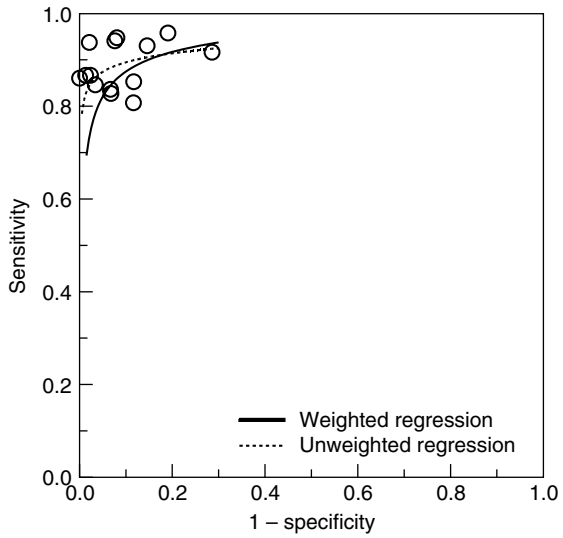

Figure 3 An SROC plot for studies of ultrasound for carotid artery stenosis

An unweighted regression gave an estimated intercept of 4.3, which indicates a good test with an intercept odds ratio of 74, but the negative slope of −0.57, indicates that the *OR* decreases with increasing S (threshold). The corresponding estimates for the weighted regression are 4.1 and −0.28 for the intercept and slope respectively. Figure 3 shows the corresponding SROC plots. The cluster of studies in the top-left corner indicates generally good discrimination ability (a perfect test would include the 100%

sensitivity, 0% 1 – specificity point, whereas a worthless test would be a straight diagonal from the 0–0% to 100–100% points).

The summary ROC curves shown in Figure 3 are obtained by computing the expected sensitivity for chosen values of 1 – specificity that lie within the range observed for the studies included in the analysis. Back transformation is used to express the expected sensitivity as a function of 1 – specificity and the parameter estimates of the model [18]. In the example, both of the resulting SROCs are moderately asymmetric (closer to the left-vertical axis than the top axis), particularly for the unweighted analysis.

Exploring Heterogeneity

If the ROC plot shows important heterogeneity in threshold, discrimination ability or both, the analyst should look to explain this. The causes may be categorized as (i) study design features, (ii) population differences, and (iii) test differences. A useful tool for exploring this is meta-regression: a regression analysis over all studies with study features used as predictors [12]. For example, in a meta-analysis of Thallium Scintigrams for coronary artery disease, Irwig et al. [9] add to their regression of A on S an indicator variable (*see* **Dummy Variables**) for whether the reading technique was computerized. This showed no difference in A for computerized versus noncomputerized reading, although computerized readings had a significantly lower threshold (more sensitive but less specific) as shown by a t test on S.

Further Methods

Alternatives for Combining Dichotomous Test Studies

The technique described above is the most commonly used in practice, but it is only one of several ways of combining studies with dichotomous test results. The first method developed [10] plotted and regressed logit(sensitivity) against logit(specificity); if the slope is 1, this implies a constant odds ratio but with shifting threshold between studies. This is conceptually equivalent to the method described above, but with a different parameterization; the authors also used the **profile likelihood** method rather than the linear and robust regression used for estimation by Moses and Littenberg. Both methods assume that the model parameters are fixed effects.

An alternative approach, based on a latent scale **logistic regression** model, provides a more flexible framework for SROC modeling. The hierarchical summary Receiver Operator Characteristic (HSROC) approach developed by Rutter and Gatsonis [19, 20] allows test threshold (cut-point), test accuracy (location), and the dependence of test accuracy on threshold (scale) to be modeled. Under this model, the probability of a positive test result π_{ij} in study i and disease group j ($1 =$ diseased, $2 =$ nondiseased) is assumed to follow a **binomial distribution**. The model takes the form $\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i\, dis_{ij}) \exp(-\beta\, dis_{ij})$ where dis_{ij} represents the "true" disease status (coded as -0.5 for the nondiseased and 0.5 for the diseased). Each study has its own implicit threshold (θ_i, equivalent to $S_i/2$) and diagnostic accuracy (α_i, log-odds ratio), both specified as random effects. The scale parameter (β) provides for asymmetry in the SROC by allowing accuracy to vary with implicit threshold. This parameter is assumed to be fixed as no single study can provide an estimate of the shape of the SROC. The random effects are assumed to be independent and normally distributed with $\theta_i \sim N(\Theta, \tau_\theta^2)$ and $\alpha_i \sim N(\Lambda, \tau_\alpha^2)$. The parameter estimates can be used to estimate the summary ROC, the expected operating point (1 – specificity, sensitivity) and corresponding likelihood ratios for a diagnostic test. This model also has the advantage that it takes into account both within study variability and heterogeneity between studies. Covariates may be added to the model to assess whether test threshold, accuracy, and/or SROC shape vary with study or patient characteristics. Rutter and Gatsonis outline how the model may be fitted using a fully Bayesian analysis using MCMC estimation [20]. Empirical Bayes estimates of the model parameters can also be obtained using Proc NLMIXED in SAS [15].

Combining tests that are Continuous variables

Hasselblad & Hedges [5] suggest another alternative that provides a connection with continuously valued diagnostic tests. This is based on the finding that, if the diseased and nondiseased populations have **normally distributed** test results, then log OR is approximately constant over the range of possible threshold choices. If the distributions of test results

are logistic and have equal variances, then log *OR* constancy is exact rather than approximate. Under these circumstances, the log-odds ratio is simply a constant multiplied by the standardized difference between the two means

$$\Delta = \left(\frac{\sqrt{3}}{\pi}\right) \log OR \tag{5}$$

where Δ is the standardized difference between the two means (a commonly used alternative measure of discrimination ability), and the two constants come from the logistic density function. Hasselblad & Hedges [5] also show that this method is relatively robust to violations of the assumptions of equal variance and **logistic distributions**.

This relationship leads to a simple approach to providing summary estimates for several studies with continuous test results, or for a mixture of dichotomous and continuous results. In either case, the discrimination ability of each test can be represented by a summary estimate ($\log OR$ or Δ) and these combined as a weighted mean, with the inverse of the variance used as the weights:

$$\log \overline{OR} = \frac{\sum w_i \log OR_i}{\sum w_i} \tag{6}$$

where w_i is the inverse variance of log *OR*, and the variance of the weighted mean is the inverse of the denominator in the above equation.

Combining Tests that are Ordinally Valued

For ordinally valued test studies, ordinal regression methods have been suggested (*see* **Polytomous Data**). The simplest of these assumes a constant odds ratio and a fixed number of result categories [22]:

$$\text{logit}[\Pr(Y \leq j | x_1, \ldots, x_k)] = \theta_j + (\alpha_1 x_1 + \cdots + \alpha_k x_k), \tag{7}$$

where θ_j is a separate constant for each category and $x_1, \ldots, x_k$ are a series of **explanatory variables**. This approach is only valid if it is reasonable to assume that the SROC is symmetric.

An alternative, computer-intensive method used by Kester and Buntinx fits the Moses model to estimate an ROC for each study [11]. The **bootstrap method** is used to obtain a valid estimate of the constant and coefficient of *S* for each study that take into account correlation between multiple estimates of *A* and also *S* within the same study. Summary estimates across studies are then obtained using the random effects bivariate regression method of van Houwelingen and Zwinderman [23]. Dukic and Gatsonis describe a Bayesian **hierarchical model**, which assumes a study-specific ROC that is sampled from a population of ROC curves for such studies [3]. Studies are not constrained to have the same set or number of categories. SROC curves and corresponding credible regions can be constructed using this approach. Dukic and Gatsonis also describe a simpler, **fixed effects** model that assumes the ROC curves for all studies have the same location and scale parameters, that is, the same accuracy and shape. However, thresholds may again vary across studies.

Combining Areas under ROCs

We have focused on methods that use the odds ratio as the summary measure of discrimination, and then used this to derive the SROC. An alternative is to use the area under the ROC for each study [16, 26], and combine areas. This can be used for dichotomous, ordinal, or continuous data, but has a clear advantage for ordinal data where studies can be combined readily even if the number of categories differs between studies. The disadvantage is the loss of ability to explore threshold variation and shape of the ROC curve.

Meta-analytic methods for diagnostic tests are less well developed than for clinical trials. The development of a mixed model for fitting SROC curves to dichotomous test results and the availability of software for fitting the model provides a general approach for the meta-analysis of diagnostic studies that takes into account within and between study variability. Both the SROC and HSROC methods provide a means of exploring heterogeneity between studies. However, the HSROC model has not been widely used and is not well tested in practice. Current methods for the meta-analysis of ordinal test results are either very complex to fit, or make simplifying assumptions that may be inappropriate. The potential impact of publication bias, and other problems of meta-analysis have been little explored.

While a number of analytic methods have now been developed for different types of tests, presentation of these results for clinical use has been less well explored. An SROC alone is insufficient to allow application in a clinical setting. A minimum requirement is back transformation to the sensitivities and

specificities for particular cutpoints. However, this raises the problem of whether the scaling of measurements is comparable between studies. Further attention is needed to the clinical application of the meta-analytic results.

References

[1] Bossuyt, P.M., Irwig, L. & Glasziou, P.P. *Diagnostic Test Appraisal Form*, Available at `http://www.health.usyd.edu.au/step/about/appraisal`.

[2] Deeks, J. (2001). Systematic reviews of evaluations of diagnostic and screening tests, in *Systematic Reviews in Health Care: Meta-analysis in Context*, 2nd Ed., M. Egger, G. Davey-Smith & D. Altman eds. BMJ Publishing Group, London.

[3] Dukic, V. & Gatsonis, C. (2003). Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds, *Biometrics* **59**, 936–946.

[4] Glasziou, P.P., Irwig, L., Bain, C. & Colditz, G. (2001). *Systematic Reviews in Health Care*. Cambridge University Press, Cambridge.

[5] Hasselblad, V. & Hedges, L.V. (1995). Meta-analysis of screening and diagnostic tests, *Psychological Bulletin* **117**, 167–178.

[6] Haynes, R.B. & Wilczynski, N. Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: an analytic survey, *BMJ* **328**(7447), 1040.

[7] Irwig, L., Bossuyt, P., Glasziou, P., Gatsonis, C. & Lijmer, J. (2002). Designing studies to ensure that estimates of test accuracy are transferable, *BMJ* **324**, 669–671.

[8] Irwig, L., Macaskill, P., Glasziou, P. & Fahey, M. (1995). Meta-analytic methods for diagnostic test accuracy, *Journal of Clinical Epidemiology* **48**, 119–130.

[9] Irwig, L., Tosteson, A.N.A., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T.C. & Mosteller, F. (1994). Guidelines for meta-analyses evaluating diagnostic tests, *Annals of Internal Medicine* **120**, 667–676.

[10] Kardaun, J.W.P.F. & Kardaun, O.J.W.F. (1990). Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation, *Methods of Information in Medicine* **29**, 12–22.

[11] Kester, A.M. & Buntinx, F. (2000). Meta-analysis of ROC curves, *Medical Decision Making* **20**, 430–439.

[12] Lijmer, J.G., Bossuyt, P.M. & Heisterkamp, S.H. (2002). Exploring sources of heterogeneity in systematic reviews of diagnostic tests, *Statistics in Medicine* **21**, 1525–1537.

[13] Lijmer, J.G., Mol, B.W., Heisterkamp, S., Bonsel, G.J., Prins, M.H., van der Meulen, J.H.P. & Bossuyt, P.M. (1999). Empirical evidence of design related bias in studies of diagnostic tests, *JAMA* **282**, 1061–1066.

[14] Littenberg, B. & Moses, L.E. (1993). Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method, *Medical Decision Making* **13**, 313–321.

[15] Macaskill, P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis, *Journal of Clinical Epidemiology*; in press.

[16] McClish, D.K. (1992). Combining and comparing area estimates across studies or strata, *Medical Decision Making* **12**, 274–279.

[17] Midgette, A.S., Stukel, T.A. & Littenberg, B. (1993). A Meta-analytic method for summarising diagnostic test performance: receiver operating characteristic summary point estimates, *Medical Decision Making* **13**, 253–257.

[18] Moses, L.E., Shapiro, D. & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations, *Statistics in Medicine* **12**, 1293–1316.

[19] Rutter, C.M. & Gatsonis, C.A. (1995). Regression methods for meta-analysis of diagnostic test data, *Academic Radiology* **2**, S48–S56.

[20] Rutter, C. & Gatsonis, C. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations, *Statistics in Medicine* **20**, 2865–2884.

[21] Scouller, K., Conigrave, K., Macaskill, P., Irwig, L. & Whitfield, J. (2000). Should we use Carbohydrate-deficient transferrin instead of gamma-glutamyl-transferase for detecting problem drinkers? A systematic review and meta-analysis, *Clinical Chemistry* **46**, 1894–1902.

[22] Tosteson, A.N.A. & Begg, C.B. (1988). A general regression methodology for ROC curve estimation, *Medical Decision Making* **8**, 204–215.

[23] van Houwelingen, H.C., Zwinderman, K.H. & Stijnen, T. (1993). A bivariate approach to meta-analysis, *Statistics in Medicine* **12**, 2273–2284.

[24] Walter, S.D. (2002). Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data, *Statistics in Medicine* **21**, 1237–1256.

[25] Whiting, P., Rutjes, A.W., Reitsma, J.B., Bossuyt, P.M. & Kleinjen, J. (2003). *The Development of QUADAS: A tool for the Quality Assessment of Studies of Diagnostic Accuracy Included in Systematic Reviews*, BMC Medical Research Methodology, 3:25. Available at `http://www.biomedcentral.com/1471-2288/3/25`.

[26] Zhou, X. (1996). Empirical Bayes combination of estimated areas under ROC curves using estimating equations, *Medical Decision Making* **16**, 24–28.

(*See also* **Combining *P* Values**)

PETRA MACASKILL, PAUL GLASZIOU & LES IRWIG

Method of Moments

The *method of moments* is a straightforward statistical technique for constructing point estimators of the parameters in a statistical model. In the early days of statistics it was a fundamental statistical tool; in particular, **Karl Pearson** encouraged its use as a method of fitting frequency curves that deviated from the normal. However, Fisher [4] established that it could be highly inefficient, and since that time it has been largely displaced by methods known to be more statistically efficient, primarily **maximum likelihood**. Even so, the method continues to play an important auxiliary role in estimation owing to its flexibility, because one can often derive computationally simple methods in otherwise difficult problems. The method is nearly foolproof in the sense of providing consistent estimators with easily derived standard errors. An additional motivation for the use of the method is that its validity depends only upon a small number of easily stated and easily checked assumptions about the structure of the statistical problem.

The Basic Method

Suppose we have a statistical model for a univariate variable X, with a d-dimensional parameter $\boldsymbol{\theta}$. In its most basic form, the method of moments requires setting up a system of d equations for the d unknown θs by equating the first d sample **moments** of the variable X with their expectations under the model, then solving that system for the method of moments estimator $\hat{\boldsymbol{\theta}}$. That is, if the data $X_1, X_2, \ldots, X_n$ form a random sample of univariate observations, then let the sample moments be denoted by $m_r = n^{-1}\sum_{i=1}^{n} X_i^r$, and let the theoretical moments be denoted by $\mu_r(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[X^r]$, where $\mathrm{E}_{\boldsymbol{\theta}}$ indicates the operation of **expectation** under the proposed probability model. The basic *method of moments estimator* would then be the value of $\boldsymbol{\theta}$ that solves

$$m_r = \mu_r(\boldsymbol{\theta}) \quad \text{for } r = 1, 2, \ldots, d. \tag{1}$$

Ideally, this system can be solved explicitly for the estimators, or the solution can be obtained with low computational difficulty. The question of the existence and uniqueness of the solutions is not always elementary; see [9] and [10] for a careful analysis in the case of the *mixture model*.

It should be noted that a single problem has more than one method of moments because one can transform the data, say by $y = g(x)$, and use the moment system determined by the new variable y. The transformation could be chosen either to simplify the equations or for a gain in efficiency.

Elementary Examples

If the parameter dimension is two, then the first two moments are matched, which is equivalent to matching the sample mean and variance of X to the theoretical mean and variance. Thus in the **normal** model with $X \sim \mathrm{N}(\mu, \sigma^2)$, the method of moments estimator for μ is the sample mean, and for σ^2 it is the sample variance. If the model is **gamma** with parameters (α, β), then the first two moments of the gamma distribution are $\mu_1 = \alpha\beta$ and $\mu_2 = \alpha\beta^2 + \alpha^2\beta^2$. Equating these quantities to m_1 and m_2 yields the solutions $\hat{\alpha} = m_1^2/(m_2 - m_1^2)$ and $\hat{\beta} = (m_2 - m_1^2)/m_1$. Thus, unlike the maximum likelihood equations, there is a solution that does not require an algorithmic method.

Multivariate Data

The extension of the basic method of moments to multivariate data is not elementary. The reason is that if the variable X is of dimension p, then there are p first moments, $p(p-1)/2$ second moments, $p(p-1)(p-2)/6$ third moments, and so forth. It follows that setting up a system of exactly d equations will typically be impossible without the careful selection of a subset of higher-order moments. For more on this, see Lindsay & Basak [11], who created a system of multivariate moment equations for the normal mixture model based on the criteria of having unique and simple-to-compute estimators.

Generalized Method of Moments

There are some natural extensions of the method of moments that enrich the approach, allowing the user to create equations that are either easier to solve or more theoretically efficient. For example, one could choose a set of d functions of the variable x, say $g_1(x), \ldots, g_d(x)$, and then find the solution in $\boldsymbol{\theta}$ to a set of *generalized moment equations* that equate

the sample average of these variables with their theoretical expectations:

$$\overline{g}_r = n^{-1} \sum_{i=1}^{n} g_r(X_i) = \mathrm{E}_{\theta}[g_r(X)], \quad \text{for } r = 1, \ldots, d. \tag{2}$$

Note that if $g_r(X) = x^r$, then this is the basic method of moments. If one sets $g_r(x) = \exp(t_r x)$ for a set of grid values $t_1, \ldots, t_d$, then one is matching the empirical **moment-generating function** $n^{-1} \sum \exp(tX)$ to its theoretical value $\mathrm{E}_{\theta}[\exp(tX)]$ along the grid values. One can use generalized moment equations in the case of multivariate X by using scalar-valued functions $g_r(\mathbf{x})$ in the above equations.

It should be noted that in the **exponential family of distributions**, the maximum likelihood equations are generalized moment equations for a set of appropriately chosen functions g_r. In this case, and this case only, the method of moments yields fully efficient estimators. For example, in the normal model the functions $g_1(x) = x$ and $g_2(x) = x^2$ give full efficiency, while in the gamma model the necessary functions are $g_1(x) = x$ and $g_2(x) = \ln x$. It follows that if one uses x and x^2 in the gamma, as was done above, then the estimators do not have optimal efficiency.

Least Squares Method of Moments

A further extension of the method allows one to use a set of functions $[g_r(X) : r = 1, \ldots, R]$ whose cardinality R is greater than the dimension d of the parameter vector. In parallel with **least squares** estimation, one establishes an objective function of the form

$$S(\boldsymbol{\theta}) = \sum_{r=1}^{R} \{\overline{g}_r - \mathrm{E}_{\theta}[g_r(X)]\}^2. \tag{3}$$

The *least squares method of moments estimator* is the value of $\boldsymbol{\theta}$ that minimizes $S(\boldsymbol{\theta})$. This method was used by Quandt & Ramsey [16] to determine estimators in the mixture-of-normals problem. They used the system of functions $g_r(X) = \exp(t_r X)$ for a grid of values t_r, so that the method was based on finding a close fit of the fitted moment-generating function to the empirical one.

Distributional Theory

The asymptotic distribution theory of the method of moments estimators is easily derived using the **delta method**. This leads in a ready fashion to the asymptotic normality of the estimators as well as to a formula for the asymptotic **covariance matrix**. The most natural approach to the statistical theory is through the wider subject of **estimating functions** [6, pp. 3–20]. This theory deals with estimators that are derived as the solutions to a set of equations of the form:

$$h_r(X, \boldsymbol{\theta}) = 0, \quad \text{for } r = 1, \ldots, d, \tag{4}$$

where the functions $h_r(X, \boldsymbol{\theta})$, $r = 1, \ldots, d$, have mean zero under the probability model: $\mathrm{E}_{\theta}[h_r(X; \boldsymbol{\theta})] = 0$. The method of moments simply corresponds to the use of functions h_r of the special form $\overline{g}_r - \mathrm{E}_{\theta}[g_r(X)]$.

If one wishes to use a set of $R > d$ moment equations, as in the least squares method of moments, there is an optimal way to combine the moment equations linearly to construct a set of d equations. Hansen [7] and Lindsay [8] showed that the highest efficiency for the resulting estimators arises from solving

$$\{\nabla \mathrm{E}_{\theta}[\mathbf{g}(X)]\}^{\mathrm{T}} V(\boldsymbol{\theta}) \{\overline{\boldsymbol{g}} - \mathrm{E}_{\theta}[\mathbf{g}(X)]\} = 0, \tag{5}$$

where $V(\boldsymbol{\theta})$ is the inverse of the theoretical covariance matrix of $\{\mathbf{g} - \mathrm{E}_{\theta}[\mathbf{g}(X)]\}$. Under some regularity assumptions, the resulting estimators are asymptotically equivalent to those arising from minimizing

$$S^*(\boldsymbol{\theta}) = \{\overline{\boldsymbol{g}} - \mathrm{E}_{\theta}[\mathbf{g}(X)]\}' V(\boldsymbol{\theta}) \{\overline{\boldsymbol{g}} - \mathrm{E}_{\theta}[\mathbf{g}(X)]\}, \tag{6}$$

a generalized least squares criterion. If the covariance matrix is a constant multiple of the identity, then minimizing $S^*(\boldsymbol{\theta})$ is equivalent to minimizing $S(\boldsymbol{\theta})$, but otherwise efficiency is improved.

A review of the modern literature shows a wide range of continuing applications of the method of moments. Some of the more important ones include the following:

1. *Regression modeling*. Hansen [7] introduced a generalized method of moments estimator for the **regression** problem as follows. Suppose $\mathbf{u}_r$, $r = 1, \ldots, R$, is a system of n-dimensional vectors, each of which is uncorrelated with the residual

vector $\mathbf{Y} - Z\boldsymbol{\beta}$, so that $\mathrm{E}_\beta[\mathbf{u}'_r(\mathbf{Y} - Z\boldsymbol{\beta})] = 0$. Then it is clear that one can use the functions $g_r(\mathbf{Y}) = \mathbf{u}'_r\mathbf{Y}$ together with the aforementioned optimal linear combination criterion to perform generalized method of moments.

2. *A binomial example*. O'Quigley [15] considered a problem in which the goal was to estimate the **binomial** parameter n based on a sequence of independent data of the form $(X_i, p_i), i = 1, \ldots, N$, where the variable X_i was bin (n, p_i). The estimation problem arose in the context of estimating the number of stem cells involved in repopulating the marrow following allogenic bone marrow transplantation. It was shown that one could generate a method of moments estimator for n based on asymptotic moments that is similar in efficiency to maximum likelihood, and an improvement upon the moment estimator that had been used previously.
3. *Mixture models*. Finding the maximum likelihood estimators in the mixture model is quite computationally onerous. There are typically multiple solutions to the **likelihood** equations. A computationally intensive approach to this problem is to try to determine all the solutions, then use the one with the highest likelihood. The problem is aggravated by the fact that the commonly used algorithms are either unreliable or slow at finding roots. This creates a situation where the method of moments provides a useful tool, even though the estimators are not highly efficient. Instead of searching for the maximum of the likelihood, one uses the solution to the likelihood equations found by searching algorithmically from a moment-based estimator. The consistency of the moment estimator ensures that the resulting solution has good theoretical properties. Furman & Lindsay [5] and Lindsay & Basak [11] show that such moment-based estimators can be very effective when used as starting values for the **EM algorithm**.
4. *Supplementary moment estimators*. Another important modern use of the method of moments idea is as a supplement to another system of estimation. For example, if one is using the **quasi-likelihood** approach described in [13, p. 325], then the regression parameters are found from a system of estimating equations, while the dispersion parameter is fit by equating a theoretical and observed moment [12, 17]. The advantage to this methodology is that one can form a simple consistent estimator with a minimum of additional assumptions, satisfying a basic goal of quasi-likelihood theory. Williams [18], Breslow [3], and Moore [14] have extended this approach in the context of regression analysis in overdispersed binomial and **Poisson regression** problems.

Beal [2] and Altman [1] provide further examples of the use of the supplementary method of moments to handle otherwise difficult problems in the estimation of variances and correlations.

References

[1] Altman, N.S. (1993). Estimating error correlation in nonparametric regression. *Statistics and Probability Letters* **18**, 213–218.

[2] Beal, S.L. (1991). Computing initial estimates with mixed effects models: a general method of moments, *Biometrika* **78**, 217–220.

[3] Breslow, N. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics* **33**, 38–44.

[4] Fisher, R.A. (1921). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **222**, 309.

[5] Furman, W.D. & Lindsay, B.G. (1994). Measuring the relative effectiveness of moment estimators as starting values in maximizing mixture likelihoods, *Computational Statistics and Data Analysis* **17**, 493–507.

[6] Godambe, V.P. (1991). *Estimating Functions*, Oxford Statistical Science Series, Vol. 7. Oxford University Press, Oxford.

[7] Hansen, L. (1982). Large sample properties of generalized method of moments estimators, *Econometrica* **50**, 1029–1054.

[8] Lindsay, B.G. (1982). Conditional score functions: some optimality results, *Biometrika* **69**, 503–512.

[9] Lindsay, B.G. (1989). Moment matrices: applications in mixtures, *Annals of Statistics* **17**, 722–740.

[10] Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol. 5. Institute of Mathematical Statistics, California.

[11] Lindsay, B.G. & Basak P. (1993). Multivariate normal mixtures: a fast consistent method of moments, *Journal of the American Statistical Association* **88**, 468–476.

[12] McCullagh, P. (1983). Quasi-likelihood functions, *Annals of Statistics* **11**, 59–67.

[13] McCullagh, P. & Nelder, J.A. (1995). *Generalized Linear Models*, 2nd Ed. *Monographs on Statistics and Applied Probability*, Vol. 37. Chapman & Hall, New York.

[14] Moore, D.F. (1986). Asymptotic properties of moment estimators for over dispersed counts and proportions, *Biometrika* **73**, 583–588.

[15] O'Quigley, J. (1992). Estimating the binomial parameter n on the basis of pairs of known and observed proportions, *Applied Statistics* **41**, 173–180.

[16] Quandt, R.E. & Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions, *Journal of the American Statistical Association* **73**, 730–752.

[17] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**, 439–447.

[18] Williams, D.A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics* **31**, 144–148.

(*See also* **Estimation**)

BRUCE G. LINDSAY

Michaelis–Menten Equation

The *Michaelis–Menten equation* describes the theoretic relationship between the initial velocity, v, of a simple enzymatically catalyzed reaction and the substrate concentration, s. It has the following form:

$$v = \frac{Vs}{(K_m + s)}, \tag{1}$$

where the constants V and K_m are the *maximum velocity* and the *Michaelis constant*, respectively. The curve described by (1) is a rectangular hyperbola through the origin, with asymptotes $s = -K_m$ and $v = V$. Substituting $v = V/2$ in the above expression, it can be seen that the value of the Michaelis constant is, in fact, the substrate concentration at half-maximal velocity. V is not a fundamental property of an enzyme, but it depends on the enzyme concentration.

The detailed properties of enzyme systems that can be described by the Michaelis–Menten equation are given in Cornish-Bowden [1], as is the history of the development of enzyme kinetics. Michaelis & Menten published details of their equation in 1913, and are regarded as founders of modern enzymology, although their equation had been derived earlier by Henri (see [1]).

Graphical Representations

If a series of initial velocities is measured at different substrate concentrations, it is natural to examine the relationship between them through the use of simple plots. The most obvious starting point is to plot v vs. s. Much more commonly, however, the investigators take reciprocals of both sides of (1) to produce

$$\frac{1}{v} = \frac{1}{V} + \frac{K_m}{Vs}. \tag{2}$$

A plot of $1/v$ vs. $1/s$ will be a straight line with slope K_m/V and intercept $1/V$ on the $1/v$ axis. This is known as the *double-reciprocal* or *Lineweaver–Burk plot*. Multiplying both sides of (2) by s yields

$$\frac{s}{v} = \frac{K_m}{V} + \frac{s}{V}. \tag{3}$$

This indicates that a plot of s/v vs. s should be a straight line, with slope $1/V$ and intercepts K_m/V on the s/v axis and $-K_m$ on the s axis. This is known as the *Hanes plot*. Finally, the *Eadie–Hofstee plot* is a graph of v vs. v/s and should be a straight line with slope $-K_m$ and intercepts V on the v axis and V/K_m on the v/s axis. This relationship is obtained by multiplying both sides of (2) by vV and rearranging to give

$$v = V - \frac{K_m v}{s}. \tag{4}$$

All three of these straight-line relationships can be used for diagnostic purposes and for obtaining preliminary parameter estimates. They should not, however, be used with simple **linear regression** programs for any formal approach to the analysis of data of this type. If the investigator wishes to affect the activity of an enzyme system with some sort of inhibitor, then these diagnostic plots can be very useful in indicating the type of inhibition. A *competitive inhibitor* (the most common type), will affect K_m but not V; a series of plots with differing concentrations of the inhibitor would be expected to yield a set of double-reciprocal plots passing through a common intercept on the $1/v$ axis.

Curve-Fitting and Parameter Estimation

It is straightforward to fit data directly to the Michaelis–Menten equation using **least squares** (either unweighted or weighted to allow for heteroscedasticity) or **maximum likelihood** criteria. An early paper by Wilkinson [9] showed how one might adapt a simple linear regression program to fit data to the Michaelis–Menten equation; and McCullagh & Nelder [4] discuss techniques for adapting GLIM procedures (*see* **Software, Biostatistical**). Following Nelder [5, 6], examination of the double-reciprocal plot in (2) shows that the Michaelis–Menten equation is an example of a **generalized linear model**. The link function relating the response (initial velocity) to the linear predictor is the reciprocal; and the linear predictor is given by the right-hand side of the equation. Errors could be assumed to be normal with a constant variance but, more naturally, they might be regarded as **gamma** variates.

Robust and Distribution-Free Estimation

The most popular distribution-free method of estimating the parameters of the Michaelis–Menten equation is based on the method of Theil [8]. This is the method that biochemists call the *direct linear plot* [2, 3]. Distribution-free methods (*see* **Nonparametric Methods**) have been used by biochemists because of their **robustness** to departures from the statistical assumptions usually made about the measurement errors: the direct linear plot method is insensitive to the occasional laboratory blunder. Consider any two pairs of observations, (s_i, v_i) and (s_j, v_j), with s_j assumed to be greater than s_i. Using these two pairs of observations, estimates of the Michaelis–Menten parameters can be obtained from the following [2]:

$$\left(\frac{1}{V}\right)_{ij} = \frac{(s_j/v_j) - (s_i/v_i)}{s_j - s_i}$$

and

$$\left(\frac{K_m}{V}\right)_{ij} = \frac{(1/v_i) - (1/v_j)}{(1/s_i) - (1/s_j)}. \qquad (5)$$

Note that these are the estimates of the intercept and the slope of the double reciprocal plot (2). In all, there are N pairs of observations, and there are $N(N-1)/2$ different possibilities for the parameter estimates, provided that the s_i are all different. The direct linear plot estimates for $1/V$ and for K_m/V are given by the **medians** of the $N(N-1)/2$ solutions provided by (5). Estimates of V and K_m are then calculated from these two medians. Although this estimation method is easily programmed for use on a personal computer, it has traditionally been used as a graphical estimation method – hence its name [3]. An alternative estimation procedure is provided by the *repeated median estimator* [7] (*see* **Robust Regression**).

References

[1] Cornish-Bowden, A. (1979). *Fundamentals of Enzyme Kinetics*. Butterworth, London.

[2] Cornish-Bowden, A. & Eisenthal, R. (1978). Estimation of Michaelis constant and maximum velocity from the direct linear plot, *Biochimica et Biophysica Acta* **523**, 268–272.

[3] Eisenthal, R. & Cornish-Bowden, A. (1974). The direct linear plot: a new graphical procedure for estimating enzyme kinetic parameters, *Biochemical Journal* **139**, 715–720.

[4] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[5] Nelder, J.A. (1966). Inverse polynomials, a useful group of multi-factor response functions, *Biometrics* **22**, 128–141.

[6] Nelder, J.A. (1968). Weighted regression, quantal response data, and inverse polynomials, *Biometrics* **24**, 979–985.

[7] Siegel, A.F. (1982). Robust regression using repeated medians, *Biometrika* **69**, 242–244.

[8] Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, I, II and III, *Proceedings of Koninklijke Nederlandsche Akadamie van Wetenschappen* **53**, 386–392, 521–525, and 1397–1412.

[9] Wilkinson, G.N. (1961). Statistical estimations in enzyme kinetics, *Biochemical Journal* **80**, 324–332.

(*See also* **Graphical Displays**; **Pharmacokinetics and Pharmacodynamics**)

GRAHAM DUNN

Midwifery, Obstetrics, and Neonatology

Midwifery is undoubtedly the oldest of the professions concerned with caring for women and their babies before, during, and after birth. The word "midwife" comes from Middle English and means "with the woman", while the French word "sage-femme" (wise woman), reflects the way that the women who cared for other women during labor and childbirth might also be traditional healers.

The word "obstetrician" comes from "obstetrix", the Latin word for midwife, itself derived from the verb "obstare", meaning stand at, before, or against. In Europe, male doctors' involvement with childbirth as "man midwives" or obstetricians dates back to about the middle of the eighteenth century, when they began to establish themselves as the practitioners for complicated deliveries, and their knowledge and influence grew [44]. Obstetricians had relatively low esteem during the latter half of the nineteenth century when most women were delivered by midwives or general practitioners.

A feature of maternity care over the past two centuries has been interprofessional rivalry between midwives and specialist and generalist doctors. Controversy has raged not only about who should do what tasks – for example, whether midwives could do forceps deliveries – but also about who is entitled to conduct deliveries at all. The outcome has varied widely between developed countries. Thus, in the US, certified midwives conducted just 8.0 of live births in 2001 [51]. For many years, midwifery was actually illegal in most Canadian provinces.

In contrast, midwives conduct about two-thirds of deliveries in the UK. Even where midwives are the most usual birth attendants and are officially recognized as the independent practitioners responsible for supervising normal pregnancy and birth, the rise in obstetric technology led to a downgrading of their role. In reaction to this there has been a reassertion of the role of midwives in the UK and elsewhere since the early 1980s, and a growing interest in basing practice on research evidence and developing midwifery research [54]. This has coincided with a reawakening of interest in midwifery in the US and moves to legalize midwifery in some Canadian provinces.

Some international agencies concerned with developing services for women in less developed countries have been active in training existing traditional attendants in appropriate practices and in training midwives, but other agencies are dominated by North American views that every woman should be delivered by an obstetrician.

Special care for immature newborn babies dates back at least to the 1890s and developed considerably in the 1920s and 1930s [56]. Nevertheless, the most widespread developments took place from the 1960s and 1970s. Since then, there have been parallel developments of neonatology (meaning "the science of the care of the newborn") as it is now known, as a separate subspecialty within pediatrics, and of neonatal nursing as a specialism within nursing.

What are the Questions?

In this context, it is inevitable that key issues from the nineteenth century onwards have related to the relative merits of different settings for birth and of the types of practitioners working in them. Attention has also focused on geographic variations and trends over time in the outcome of pregnancy and the relative strength of **association** between the socioeconomic circumstances of the child-bearing population, genetic factors, and the quality of the services available to women giving birth. Although questions have been raised for centuries about specific practices, it is only in the past 20 or 30 years that systematic attempts have been made to evaluate them.

What Methods have been Used?

As in many other areas, there is a long history of descriptive studies based on case series (*see* **Case Series, Case Reports**) and population-based data. These were given added impetus by the introduction of civil registration in many countries during the nineteenth century. Inevitably some analyses were hospital-based and the **selection biases** inherent in this were recognized in the latter half of the nineteenth century. During the first half of the twentieth century, **correlation** and **regression** were introduced into descriptive studies and **multivariate analysis** became more widespread with the availability of

electronic computers in the second half of the century. These greatly increased the potential for **record linkage** and follow-up studies, including those which follow-up **cohorts** of babies for a year or so, into childhood, or in some cases for the rest of their lives. **Case–control** studies were first used in this field at the beginning of the twentieth century and became much more widespread towards the end of the century. **Quasi-experimental methods**, notably "natural experiments", have a long history in this field, and the first controlled trials (*see* **Clinical Trials, Overview**) were done in the 1920s.

What is Measured?

Up to the mid-twentieth century the overriding concern has been to prevent death or severe morbidity in the mother. Concern about the health of children in the nineteenth century related more to public health and the living circumstances of young children than to conditions at birth. Although there was concern about **infant mortality** at the beginning of the twentieth century, this became the major outcome only after the massive decline in **maternal mortality** in the mid-twentieth century. Towards the end of the twentieth century, the increasing survival rates of very immature babies, largely as a consequence of developments in neonatal intensive care, has led to concern about monitoring morbidity in the survivors. Costs have always affected women's choice of maternity care, but the increasing extent to which this is funded either by public funds or by major institutions, such as insurance companies, has led to increasing concentration on measuring the costs of care. Reaction to the increasing use of technology at birth and the rise of consumerism has led to work to obtain women's views of the care they receive.

Measuring any of these is far from straightforward, and some of the problems were already recognized by the mid-nineteenth century [55] (*see* **Nightingale, Florence**). An advantage when doing **population-based studies** in this field, is that the usual denominators, the numbers of women giving birth and the numbers of babies born, can usually be assessed fairly accurately. On the other hand, the fact that a pregnancy can result in one, two, or occasionally three or more babies presents problems when doing analyses in terms of outcomes for the babies.

Time Trends and Geographic Variations in Infant Mortality – "Nature", "Nurture", or "Quality of Assistance"?

With the development in the nineteenth century of birth and death registration and publication of statistics derived from them, geographic variations in infant mortality became visible (*see* **Vital Statistics, Overview**). In England and Wales, **William Farr** used the "healthy districts", with the lowest mortality, as a yardstick with which to compare the others and make the case for improving sanitation and public health. For example, he showed that the aggregated mortality rate for 1861–1870 among children aged under 5 in Liverpool was more than three times higher than in the "healthy districts" [26].

In the last quarter of the nineteenth century, a rise in infant mortality in England and Wales coincided with decreases in both general mortality and the birth rate. This came at a time when controversies about Charles Darwin's theories of evolution were leading to debate about whether infant mortality was the consequence of poor living conditions, lack of maternity care, or a beneficial culling of potentially unfit members of the population. The debate was further fuelled by the discovery that many potential recruits for the Boer War were unfit. This led to a much closer scrutiny of infant mortality rates and associated factors and to developments in statistical methods [21, 50].

In the first volume of ***Biometrika***, founded "especially for those who are interested in the application to biology of the modern methods of statistics" [28], **George Udny Yule** asked, "Would it not be worth while for an evolutionist statistician to give some attention to the mass of material accumulated in the Decennial Supplement to the reports of the Registrar General for England and Wales?" [78]. He suggested calculating correlation coefficients between childhood and adult death rates.

At the General Register Office, John Tatham compared infant mortality rates for urban and rural counties in 1873–1877 with those in 1898–1902. He found no change in the rural rate, but a rise in the urban rate [63]. The extent to which this may reflect selective migration as well as differences in conditions is unclear. A series of reports from the Local Government Board, whose responsibilities included public health, looked in detail at geographic differences in infant mortality and the mortality of children

under the age of 5 and the factors associated with it. The first report used rankings of counties and simple descriptive techniques [42]. It discussed associations between infant mortality and sex of babies, "legitimacy", family size, stillbirths, quality of help available, age of the mother, nondomestic employment, overcrowding, sanitation, "ignorance", and "fecklessness". The introduction by the Board's Chief Medical Officer, Arthur Newsholme, included a discussion of possible associations with death in later life. In an Appendix, "On the possible selective influences of mortality on the mortality in the next four years of life", G. Udny Yule calculated correlation and regression coefficients between local infant and childhood mortality rates in successive years and found little evidence of negative correlation between infant and childhood mortality beyond the second year of life.

Karl Pearson's response, "The intensity of natural selection in man", used **life tables** for the periods 1838–1854 to 1891–1900 for England and Wales as a whole and for the "healthy districts" [64]. He found negative correlations between rising infant mortality and falling childhood mortality. He interpreted this as showing the survival of the fittest [57]. In a later analysis at a time of falling infant mortality, Karl Pearson and Ethel Elderton used the method of finite differences to try to adjust for environmental influences [24]. The results still supported their view of natural selection. This work was heavily criticized, notably by **John Brownlee**, who criticized their analyses on the grounds that they did not relate deaths to their corresponding cohorts of births and took no account of the periodicity of epidemic diseases, which were a common cause of death at the time [9].

These analyses were based on aggregated data for geographic areas (*see* **Ecologic Study**). The introduction of the Hollerith counter sorter into US vital statistics offices from the 1890s and the English General Register Office from 1911 [37, 62] extended the range of analyses which were feasible and made it possible to relate births and deaths to local government districts in which people lived.

The Local Government Board's third report focused on Lancashire, and five towns in particular [43]. The comparison between three of these – Burnley, Nelson and Colne, which had infant mortality rates of 176, 130, and 87, respectively, in 1911–1913 – has much more recently given impetus to a succession of studies comparing the health of adults in the 1980s and 1990s with their circumstances at birth [5, 6, 41], although similar studies had already been done in the 1970s [27]. Such studies have used either longitudinal **birth cohort studies** or a variety of other techniques for locating people and following them up.

In the US, birth and death registration was still incomplete in some states in the early years of the twentieth century. Nevertheless, analyses of infant mortality showed wide variations and prompted local investigations and the establishment of the Children's Bureau [76]. The Bureau's investigation of infant mortality in eight cities, directed by Robert Morse Woodbury, took a cohort approach [73]. This involved ascertaining all the births in the cities from a variety of sources and following them through the first year of life. Analysis of data about 2555 infant deaths and 22 977 live births on a relatively primitive punched card system was a challenge. Lacking techniques for multivariate analysis, he used a method of standardization, the "method of expected deaths" (*see* **Standardization Methods**). Using this, he concluded that the inverse association between infant mortality and the babies' fathers' earnings was much stronger than that with other factors, including family size, mothers' age, parity, birth interval, and type of feeding.

In an extensive analysis of data from England and Wales published in 1929, Peter McKinlay suggested that factors associated with infant mortality could be grouped into "(1) the quality of obstetric assistance in childbed (2) the health of the mother (3) social and environmental conditions". He used correlation coefficients, including multiple and partial correlation, to investigate associations with stillbirths and infant mortality, which he divided into "antenatal", "neonatal", and "postnatal" deaths [52]. He found the strongest associations with **quality of care** in the antenatal and neonatal periods, although environmental factors were also important in the latter, while the health of the mother and environmental factors were prominent in the postnatal period (*see* **Environmental Epidemiology**).

In the 1940s, Barnet Woolf & John Waterhouse used **multiple regression** in two papers on infant mortality in county boroughs of England and Wales. Motivated by a desire to counter what they felt to be insidious influences of the **eugenicists** in hindering social reform, they started from the standpoint that "a large proportion of infant deaths are preventable and no other vital index approaches infant mortality

in variability or sensitivity to social conditions" [74, 75]. They found it was strongly associated with overcrowding, male unemployment, percentages of males in social classes IV and V, percentages of women employed on manufacturing processes, and latitude.

The advent of computers considerably reduced the laboriousness of such analyses. There were many similar analyses in the 1960s and 1970s using multiple regression and **principal components analysis**, which found stronger associations with social and environmental factors than with the availability of health care [3, 48, 49]. In general, these treated proportions and rates as continuous variables, while further analyses in the 1980s and 1990s have tended to use **logistic regression**.

Computers have also increased the potential for **record linkage**, although the first study to link records of deaths of babies under the age of 1 year to their birth records was done in the punched-card era. This was a study of the 44 000 deaths in the first year of life in 1949 and 1950, together with the 1.5 million live births and 33 000 stillbirths [35]. The 1600 deaths in the second year of life among babies born in 1949 were also studied. The aim was to use data collected at birth registration to identify the categories of women with the highest mortality, with the aim of giving them priority for specialist maternity care. These analyses and others which followed suggested that the association between mortality of fetuses, babies, and mothers and mothers' ages and their parity, the number of previous births, was U-shaped, with the highest mortality being amongst the youngest and oldest women and their babies. In contrast, a study which followed up successive pregnancies to women in Aberdeen, Scotland, who had their first pregnancy during the years 1949–1954 suggested that perinatal mortality rates declined during women's reproductive career [7].

In the 1970s, such record linkage became routine in England and Wales and in an increasing number of US states. By the 1990s, most Nordic countries, Scotland, England and Wales, Israel, and some states of the US and Australia had taken this further and were able to link together successive pregnancies to the same woman. Some are able to link other records, such as **census** returns, hospital admissions, abortions, and cancer registrations (*see* **Disease Registers**) [1]. This makes it possible to analyze the outcomes of successive pregnancies according to the characteristics of both women and their babies. One of the first of these, based on all births in Norway from 1967 to 1973, reached similar conclusions to those of the earlier analysis from Aberdeen [4]. This gave rise to considerable debate about how to account for effects of self-selection for further pregnancies.

Where to be Born?

In the mid-nineteenth century, it was unusual to give birth in a hospital or lying-in institution, for very good reasons. William Farr was alluding to the results of numerous descriptive analyses when he wrote, "Contrary to expectations the advantages these institutions offered were overbalanced by one dread drawback; the mortality of mothers was not diminished; nay it became in some instances excessive; in others appalling" [25]. Much of this high mortality was due to puerperal fever – a major cause of maternal mortality up to the 1930s.

The discovery of the cause and contagiousness of puerperal fever is widely and incorrectly attributed to Ignaz Semmelweiss, whose famous treatise, "The aetiology, concept and prophylaxis of childbed", was published in 1861 [58]. In fact this had already been demonstrated in a descriptive account at least 50 years earlier by an Aberdeen obstetrician, Alexander Gordon [31], and other descriptive accounts were published during the first half of the nineteenth century [44], including a review by the American physician and poet, Oliver Wendell Holmes [19].

In this work on puerperal fever, Semmelweiss was able to use data from a "natural experiment". At the time of his appointment to the Vienna Maternity Hospital in 1846, it was divided into two clinics, and women were admitted to the two clinics on alternate days. One was for the instruction of medical students and doctors, and mortality in this clinic was much higher than in the second clinic, which was used for the instruction of midwives. Semmelweiss found that medical students would attend post-mortems of women who died of puerperal fever before going over to the clinic and doing vaginal examinations, without first changing their clothes or washing their hands. After he insisted that medical students wash their hands in disinfectant, mortality in their clinic fell [44].

Such comparisons of mortality have never been straightforward, however. First there are problems of definition, as Florence Nightingale found out when she tried to compare mortality rates for different places of birth: "Midwifery statistics are in an unsatisfactory condition" and "... there appears to have been no uniform system of records of deaths or the causes of deaths, in many institutions ..." [55]. Despite the introduction of antiseptic and aseptic techniques, which lowered mortality rates in some hospitals towards the end of the century, rates were higher than at home. The question of selection bias was raised, however. In 1904, William Williams suggested that "the majority of the worst cases are brought to hospital after labor commences and that a large number undergo some of the major operations" [72]. The high mortality in workhouse infirmaries which catered for destitute women was often attributed solely to their resulting poor health. This was questioned by Sidney & Beatrice Webb, who pointed to the inadequate care and unsavory conditions in some infirmaries [69].

The issue of selection bias is one which has always dogged the debate about the relative risks and merits of different settings for birth. In England a policy of universal hospital birth was adopted in the light of the observation that perinatal mortality had fallen at a time when the percentage of births in hospital had fallen [61]. This was parodied by **Archie Cochrane**, who pointed out that the length of postnatal stay after childbirth had fallen over the same period [15], and challenged extensively by Marjorie Tew [65]. Although it is possible to do randomized trials to compare other outcomes of care in different settings, the key issue is still safety [11].

Evaluation of Specific Aspects of Care

Although singled out by Archie Cochrane for the "wooden spoon", obstetrics, together with midwifery and neonatology, have been well ahead of other clinical professions in evaluating the care provided, particularly in the use of randomized trials. The numbers of trials published rose from fewer than 20 in 1950 to over 400 in 1990 [32]. By 2003, the Cochrane Pregnancy and Childbirth Group had 9129 trial reports in its register [23] and the Cochrane Neonetal Group had about 3000 [22].

Trials were being done much earlier, however. As early as the 1920s, nearly 400 women took part in a trial of the effect of shaving women's perineal hair on admission to hospital on the incidence of puerperal fever [39]. Neither this trial nor any later research showed shaving to be beneficial, but it was still common practice in England and Wales in the early 1980s [29].

Five trials published in the 1950s with sample sizes ranging from less than 100 to over 1600 compared diethylstilbestrol (DES), a drug thought to prevent miscarriage, with concurrent controls. Three of the trials were double-blind (*see* **Blinding or Masking**). None showed a difference in rates of miscarriage, stillbirth, and neonatal death, although in the 1980s a reanalysis of one trial suggested that DES was harmful [30]. DES continued in use until a case–control study published in 1970 [36] showed that between 1.4 and 14 per 10 000 women exposed to it as fetuses were likely to develop a very rare cancer – cancer of the vagina.

The rise in the numbers of neonatal trials is partly a reflection of the growth of the specialty. Some of the key research relates to the use of supplemental oxygen for babies born preterm. Trials compared giving preterm babies air with more than 50% oxygen with much lower concentrations of oxygen. These showed higher rates of retrolental fibroplasia, now known as retinopathy of prematurity, which can lead to blindness among surviving babies who received the higher concentrations [59], and led to policies of keeping the oxygen concentrations below 40%.

Inappropriate subgroup analyses have been a problem in this field, as in others. For example, a structured review of the relevant trials showed that administering corticosteroids to women who are about to deliver preterm can help their babies' lungs mature [18]. Nevertheless, this practice was not adopted for some time after the early trials were published, as a subgroup analysis suggested that the result applied only to black female babies [16].

Some of the earlier trials in this field were dismissed as they were not large enough to detect small or moderate treatment effects. For example, two small trials involving 350 and 462 women comparing the rates of seizures among babies whose mother had electronic fetal monitoring in labor with those who had intermittent auscultation were unable to detect a difference, although in both cases the rate appeared to be lower in the monitored group. A larger trial of nearly 13 000 women was able to detect a difference in the rate of seizures in the neonatal period but no

difference in cerebral palsy rates when the children were 4 years old [46].

One response to the problem of small sample size is to pool the results of trials formally through systematic review and **meta-analysis**. The first attempt to do this consistently for a given area of health care was in the Oxford Database of Perinatal Trials, the first release of which was published in 1988 [13]. It was used for the two volumes of *Effective Care in Pregnancy and Childbirth* [14] and a third book, *Effective Care of the Newborn Infant* [60]. The database also acted as the prototype for the Cochrane Database of Systematic Reviews (*see* **Cochrane Collaboration**).

As in other fields, there are important questions which randomized trials are unable to answer. One example is the relative safety of different settings for birth. Mortality rates are now so low that it is impossible to do a large enough trial to compare mortality for women at low risk of complications without completely reorganizing the maternity services in a wide geographic area and thus interfering with the types of service being compared. Furthermore it is impossible to blind participants or caregivers to the form of care being given [11].

Other Issues Related to the Care of Individuals

There is a long history in this field of "confidential enquiries" which examines the circumstances and pathology of individual deaths of women and babies at or around the time of birth. All too often this is done without reference to any comparison group, even though a study which compared 2527 maternal deaths in Scotland in the years 1927–1932 with all women who gave birth in a 6 month period during this time was published as long ago as 1935 [20]. The Confidental Enquiry into Stillbirths and Deaths in Infancy in England, Wales, and Northern Ireland used controls in a study of deaths at 27 to 28 weeks of gestational age in [17].

Caregivers have always wanted to identify the women who are likely to experience problems. Since the beginning of the twentieth century, obstetric textbooks have listed social and physical "risk factors", associated with complications of pregnancy and poor outcomes. The development of computerized databases in the 1960s made it possible to use multivariate analysis to develop more formal scoring systems aimed at predicting a variety of complications and adverse outcomes. One of the first of these was derived by Harvey Goldstein from his analyses of the survey of deaths among babies born in Great Britain during a week in 1958 [10].

In general, these scoring systems have had poor predictive value for the individual and often do not apply outside the populations in which they were developed. A review in 1989 commented that

> When risk scoring is applied in clinical practice, there is a very real danger that a potential but highly imprecise risk of adverse outcome becomes replaced by the certain risk of dubious treatments and interventions whose benefits have not been demonstrated and whose hazards are largely unknown [2].

Since the beginning of the twentieth century, statisticians have been interested in the statistical properties of the distribution of babies' **birthweights**, particularly among the smallest babies, who have the highest mortality rates [77]. As a result, they have been of greatest concern to clinicians as methods of neonatal care developed. By the middle of the century, it was recognized that birthweight alone was inadequate as a measure of maturity as there is a difference between preterm babies who are small because they are born too early and growth-retarded babies who are born later. This led to the construction of "standard" charts giving centiles of the birthweight distribution at each of the gestational ages (*see* **Quantiles**) [45, 66]. There are two sets of statistical problems involved – those of selecting the "standard" population and the choice of methods for fitting curves to data.

As birthweight means and distributions have been shown to vary according to the baby's sex and multiplicity, the mother's age and parity, the parents' racial and socioeconomic characteristics, and the altitude at which the mother lives, the question of choosing an appropriate population is far from trivial [40, 53]. In addition, early data came from hospitals at a time when many women gave birth at home, so they may not have been representative of the surrounding population [49].

The data used to construct the first set of standards widely used in the US came from a hospital and also related to women living at a high altitude in the state of Colorado [45, 47]. The first set widely used in the UK was based on 52 004 singleton births

within marriage which took place in the city of Aberdeen, Scotland, in the years 1948–1964 [66]. As well as being restricted to births within marriage, the population was ethnically homogeneous, even though socially varied. Since the 1970s, the tendency to induce babies or deliver them by elective Caesarean section before term has foreshortened gestational age and affected its distribution. The way it is estimated has changed with the development of ultrasound scanning. Furthermore, for very preterm babies, the numbers of babies in any data set are likely to be small and the question of whether or not to restrict the tables to live births or to add in fetal deaths becomes increasingly crucial [67]. The question of how to deal with **extreme values** is an issue at all gestations.

The methods used to fit curves to the distributions have inevitably developed over time in response to the availability of computing power and statistical methods. An important factor is the extent to which the non-normality of the birthweight distribution is taken into account [12]. In some populations, it is bimodal, particularly at low gestational ages [66]. Approaches which have been used range from a simple step function [66], using bivariate elastic **spline** interpolation to fit contours to distributions of birthweight and gestational age [38], using **nonparametric methods** to smooth empirical centiles (*see* **Nonparametric Regression**) and fitting **polynomial regression** curves to birthweight at each gestational age, assuming a **normal distribution** [8].

The differences in birthweight distributions for different populations also make it difficult to compare their mortality rates. Standardization of birthweight-specific mortality rates has been shown to be **biased**, in particular against populations with heavier birthweights [70]. An alternative method aims to eliminate this bias by using the frequency distribution of birthweight and the curve of weight-specific mortality to describe the **excess mortality** in one population compared to another [71]. So far, it has not been widely used in routine practice, probably because of its relative complexity.

Future Questions – Heredity, Environment, Clinical Effectiveness, and Quality of Care

So far, the design of randomized trials in this area has been relatively basic, in comparison with those used in agriculture or psychology. A relatively recent development, from the late 1980s onwards, is the use of **split plot designs**, described as "cluster randomization" (*see* **Randomized Treatment Assignment**) in this context, especially when comparing programmes of care, such as antenatal care [32, 33, 68]. The introduction of **Bayesian methods** into trials has in this field begun [34]. In descriptive studies, **multilevel modeling** is now being used to bring together analytically the characteristics of parents and the areas in which they live in studies of geographic variation (*see* **Geographic Epidemiology**).

As described earlier, statistical techniques developed at the beginning of the twentieth century were used in the heated debate about whether differences in mortality were related to "heredity", "environment", or "quality of assistance in childbed". As we begin the twenty-first century, it is increasingly recognized that genetic factors (*see* **Genetic Epidemiology**), clinical effectiveness, **quality of care**, and socioeconomic and environmental factors may all be related to a variety of measures of the outcome of pregnancy. The challenge for statistics is to develop appropriate quantitative techniques for describing and assessing pregnancy and its outcome.

References

[1] Adams, M.M., Herman, A.A. & Notzon, F.C., eds (1977). International symposium on maternally linked pregnancy outcomes, *Paediatric and Perinatal Epidemiology* **11**(Supplement 1), 1–150.

[2] Alexander, S. & Keirse, M.J.N.C. (1989). Formal risk scoring during pregnancy, in *Effective Care in Pregnancy and Childbirth*, I. Chalmers, M. Enkin & M.J.N.C. Keirse, eds. Oxford University Press, Oxford.

[3] Ashford, J. (1981). Trends in maternity care in England and Wales 1963–1977, in *Matters of Moment*, G. McLachlar, ed. Oxford University Press, Oxford.

[4] Bakketeig, L.S. & Hoffman, H.J. (1979). Perinatal mortality by birth order within cohorts based on sibship size, *British Medical Journal* **2**, 693–696.

[5] Barker, D.J.P., ed. (1992). *The Fetal and Infant Origins of Adult Disease*. British Medical Journal, London.

[6] Barker, D.J.P. & Osmond, C. (1987). Inequalities in health: specific explanations in three Lancashire towns, *British Medical Journal* **194**, 749–752.

[7] Billewicz, W.Z. (1973). Some implications of self-selection for pregnancy, *British Journal of Preventive and Social Medicine* **27**, 49–52.

[8] Bonellie, S.R. & Raab, G.M. (1996). A comparison of different approaches for fitting centile curves to birthweight data, *Statistics in Medicine* **15**, 2657–2667.
[9] Brownlee, J. (1917). The relation of infant mortality to mortality in subsequent life, *Journal of the Royal Statistical Society* **80**, 222–248.
[10] Butler, N. & Alberman, E.D., eds (1969). *Perinatal Problems*. E & S Livingstone, Edinburgh.
[11] Campbell, R. & Macfarlane, A.J. (1994). *Where to be Born? The Debate and the Evidence*, 2nd Ed. National Perinatal Epidemiology Unit, Oxford.
[12] Carr-Hill, R.A. & Pritchard, C.W. (1983). Reviewing birthweight standards, *British Journal of Obstetrics and Gynaecology* **90**, 718–725.
[13] Chalmers, I., ed. (1988). *The Oxford Database of Perinatal Trials*. Oxford University Press, Oxford.
[14] Chalmers, I., Enkin, M. & Keirse, M.J.N.C., eds (1989). *Effective Care in Pregnancy and Childbirth*. Oxford University Press, Oxford, pp. 612–623.
[15] Cochrane, A.L. (1972). *Effectiveness and Efficiency, Random Reflections on the Health Service*. Nuffield Provincial Hospitals Trust, London.
[16] Collaborative Group on Antenatal Steroid Therapy (1984). Effect of antenatal steroid administration on the infant, long term follow up, *Journal of Pediatrics* **104**, 259–267.
[17] Confidential Enquiry into Stillbirths and Deaths in Infancy. (2003). Project 27/28. An enquiry into quality of care and its effect on the survival of babies born at 27–28 weeks. Macintosh 17, ed London, TSO.
[18] Crowley, P. (1997). Corticosteroids prior to preterm delivery, Pregnancy and Childbirth Module of the *Cochrane Database of Systematic Reviews*, J.P. Neilson, C.A. Crowther, E.D. Hodnett, G.J. Hofmeyr & M.J.N.C. Keirse, eds. *The Cochrane Collaboration*, Issue 1. Update Software, Oxford.
[19] Cullingworth, C.J. (1906). *Oliver Wendell Holmes and the Contagiousness of Puerperal Fever*. Henry Glaisher, London.
[20] Douglas, C.A. & McKinlay, P.L. (1935). *Report on Maternal Morbidity and Mortality in Scotland*. Department of Health for Scotland, Edinburgh.
[21] Dwork, D. (1987). *War is Good for Babies and Other Young Children*. Tavistock Publications, London.
[22] Editorial Team Cochrane Neonatal Review Group. (2003). In: *The Cochrane Library*; Issue 2. Update Software, Oxford.
[23] Editorial Team Cochrane Pregnancy and Childbirth Group. (2003). In: *The Cochrane Library*, Issue 2. Update Software, Oxford.
[24] Elderton, E.M. & Pearson, K. (1915). Further evidence of natural selection in man, *Biometrika* **10**, 488–506.
[25] Farr, W. (1872). Letter to the Registrar General, in *General Register Office. Thirty Third Annual Report of the Registrar General for the Year 1870*. HMSO, London.
[26] Farr, W. (1875). Letter to the Registrar General, in *Supplement to the Thirty Fifth Annual Report of the Registrar General. Births, Deaths and Marriages in England 1861–70*. HMSO, London.
[27] Forsdahl, A. (1977). Are poor living conditions in childhood and adolescence an important risk factor for arteriosclerotic heart disease?, *British Journal of Preventive and Social Medicine* **31**, 91–95.
[28] Galton, F. (1901). Biometry, *Biometrika* **1**, 7–10.
[29] Garcia, J. & Garforth, S. (1989). Labour and delivery routines in English consultant maternity units, *Midwifery* **5**, 155–162.
[30] Goldstein, P.A., Sacks, H.S. & Chalmers, T.C. (1989). Hormone administration for the maintenance of pregnancy, in *Effective Care in Pregnancy and Childbirth*, I. Chalmers, M. Enkin & M.J.N.C. Keirse, eds. Oxford University Press, Oxford, pp. 612–623.
[31] Gordon, A. (1795). *A Treatise on the Epidemic Puerperal Fever of Aberdeen*. G., G. & J. Robinson, London.
[32] Grant, A.M. (1993). Randomized trials in perinatology, major achievements and future potential, *Annals of the New York Academy of Sciences* **703**, 107–117.
[33] Grant, A.M., Elbourne, D.R., Valentin, L. & Alexander, S. (1989). Routine formal fetal movement counting and the risk of antepartum late death in normally formed singletons, *Lancet* **ii**, 345–349.
[34] GRIT Study Group and participants. (2003). A randomised trial of timed delivery for the compromised. Preterm fetus: short term out comes and Bayesian interpretator *BJOG* **110**, 27–32.
[35] Heady, J.A. & Heasman, M.A. (1959). Social and Biological Factors in Infant Mortality, *Studies on Medical and Population Subjects, No. 15*. HMSO, London.
[36] Herbst, A.L. & Scully, R.E. (1970). Adenocarcinoma of the vagina in adolescence. A report of 7 cases including 6 clear-cell carcinomas (so called mesonephromas), *Cancer* **25**, 745–757.
[37] Higgs, E. (1996). The statistical big bang of 1911: ideology, technological innovation and the production of medical statistics, *Social History of Medicine* **9**, 409–426.
[38] Hoffman, H.J., Stark, C.R., Lundin, F.E. & Ashbrook, J.D. (1974). Analysis of birth weight, gestational age, and fetal viability, US births, 1968, *Obstetrical and Gynecological Survey* **29**, 651–681.
[39] Johnson, R.A. & Sidall, R.S. (1922). Is the usual method of preparing patients for delivery beneficial or necessary?, *American Journal of Obstetrics and Gynecology* **4**, 645–650.
[40] Kline, J., Stein, S. & Susser, M. (1989). *From Conception to Birth. Epidemiology of Prenatal Development*. Oxford University Press, Oxford.
[41] Kuh, D. & Davey Smith, G. (1993). When is mortality risk determined? Historical insights into a current debate, *Social History of Medicine* **6**, 101–123.
[42] Local Government Board. (1910). *Report by the Medical Officer on Infant and Child Mortality. Supplement to the Thirty Ninth Annual Report of the Local Government Board, 1909–10, Cd 5312*. HMSO, London.

[43] Local Government Board (1914). *Third Report by the Medical Officer on Infant and Child Mortality in Lancashire. Supplement to the Forty Third Annual Report of the Local Government Board, 1913–14, Cd 7511*. HMSO, London.

[44] Loudon, I. (1992). *Death in Childbirth*. Clarendon Press, Oxford.

[45] Lubchenco, L.O., Hansman, C., Dressler, M. & Boyd, E. (1963). Intrauterine growth as estimated from liveborn birthweight data at 24 to 42 weeks of gestation, *Pediatrics* **32**, 793–800.

[46] MacDonald, D., Grant, A.M., Sheridan Pereira, M., Boylan, P. & Chalmers, I. (1995). The Dublin randomized controlled trial of intrapartum fetal heart rate monitoring, *American Journal of Obstetrics and Gynecology* **152**, 524–539.

[47] Macfarlane, A.J. (1987). Altitude and birthweight: commentary, *Journal of Pediatrics* **111**, 842–844.

[48] Macfarlane, A.J. & Cole, T. (1985). From depression to recession – evidence about the effects of unemployment on mothers' and babies' health 1930s to 1980s, in *Born Unequal: Perspectives on Pregnancy and Childbearing in Unemployed Families*. Maternity Alliance, London, pp. 38–57.

[49] Macfarlane, A.J. & Mugford, M. (1984). *Birth Counts: Statistics of Pregnancy and Childbirth*. HMSO, London.

[50] MacKenzie, D.A. (1981). *Statistics in Britain, 1865–1930. The Social Construction of Scientific Knowledge*. Edinburgh University, Edinburgh.

[51] Martin, J.A., Hamiltan, B.E., Ventera, S.J., Menacker, F., Park, M.M., Sulton, P.D. (2002). Births: final data for 2001. National Vital Statistics Reports, Vol. 51, no. 2. National Center for Health Statistics, Hyattsville.

[52] McKinlay, P.L. (1929). Some statistical aspects of infant mortality, *Journal of Hygiene* **28**, 394–417.

[53] McKeown, T. & Gibson, J.T. (1951). Observations on all births (23,970) in Birmingham, 1947. II: Birthweight, *British Journal of Preventive and Social Medicine* **5**, 98–112.

[54] National Electronic Library for Health: midirs Informed Choice. (2003). `http://www.midirs.org/nelh/nelh.nsf/welcome?openform`.

[55] Nightingale, F. (1871). *Notes on Lying-in Institutions, Together with a Proposal for Organising an Institution for Training Midwives and Midwifery Nurses*. Longmans, Green & Company, London.

[56] Oppenheimer, G.M. (1996). Prematurity as a public health problem: US policy from the 1920s to the 1960s, *American Journal of Public Health* **86**, 870–878.

[57] Pearson, K. (1912). On the intensity of natural selection in man, *Proceedings of the Royal Society, Series B* **85**, 469–476.

[58] Semmelweiss, I. (1861). *The Aetiology, Concept and Prophylaxis of Childbed Fever*. C. Carter, trans., ed. Wisconsin Publications in the History of Science and Medicine, 1983.

[59] Silverman, W.A. (1980). *Retrolental Fibroplasia, A Modern Parable*. Academic Press, London.

[60] Sinclair, J.C. & Bracken, M.B., eds (1992). *Effective Care of the Newborn Infant*. Oxford University Press, Oxford.

[61] Standing Committee Maternity and Midwifery Advisory Committee (Chairman, J. Peel) (1970). *Domiciliary and Maternity Bed Needs*. HMSO, London.

[62] Stevenson, T.H.C. (1910). Suggested lines of advance in English vital statistics, *Journal of the Royal Statistical Society* **73**, 685–702.

[63] Tatham, J. (1904). English mortality among infants under one year of age, in *Report of the Interdepartmental Committee on Physical Deterioration, Vol. 1. Cmnd 2175*. HMSO, London.

[64] Tatham, J. (1907). Letter to the Registrar General, in *Supplement to the Sixty-Fifth Report of the Registrar General 1891–1900, Part I. Cd 2618*. HMSO, London.

[65] Tew, M. (1995). *Safer Childbirth?* 2nd Ed. Chapman & Hall, London.

[66] Thomson, A.M., Billewicz, W.Z. & Hytten, F.E. (1968). The assessment of fetal growth, *Journal of Obstetrics and Gynaecology of the British Commonwealth* **75**, 903–916.

[67] Tin, W., Wariyar, U.K. & Hey, E.N. on behalf of the Northern Neonatal Network (1997). Selection biases invalidate current low birthweight weight-for-gestation standards, *British Journal of Obstetrics and Gynaecology* **104**, 180–185.

[68] Vilar, J., Carroli, G., Khan-Neelofur, D., Piaġgio, G., Gilmezoglu, M. (2003). Patterns of routine antenatal care for low-risk pregnancy In: *The Cochrans Library*, Issue 2, Updates software, Oxford.

[69] Webb, S. & Webb, B., eds (1909). *The Break-Up of the Poor Law, being Part One of the Minority Report of the Poor Law Commission*. Longman, Green and Co., London.

[70] Wilcox, A.J. & Russell, I.T. (1983). Perinatal mortality: standardizing for birthweight is biased, *American Journal of Epidemiology* **118**, 857–864.

[71] Wilcox, A.J. & Russell, I.T. (1986). Birthweight and perinatal mortality. III: Towards a new method of analysis, *International Journal of Epidemiology* **15**, 188–196.

[72] Williams, W. (1904). *Deaths in Childbed, a Preventable Mortality*. H.K. Lewis, London.

[73] Woodbury, R.M. (1925). *Causal Factors in Infant Mortality. A Statistical Study Based on Investigations in Eight Cities*. Government Printing Office, Washington.

[74] Woolf, B. (1947). Studies on infant mortality. Part II: Social aetiology of stillbirths and infant deaths in county boroughs of England and Wales, *British Journal of Social Medicine* **2**, 73–125.

[75] Woolf, B. & Waterhouse, J. (1945). Studies on infant mortality. Part I: Influence of social conditions in county boroughs of England and Wales, *Journal of Hygiene* **44**, 67–98.

[76] Yankauer, A. (1994). A classic study of infant mortality - 1911–1915, *Pediatrics* **94**, 874–877.

[77] Ylppo, A. (1919). Zur physiologie, klinik und zum schicksal der Frühgeborenen, *Zeitschrift für Kinderheilkunde* **24**, 1–110.

[78] Yule, G.U. (1902). Local death rates, *Biometrika* **1**, 384.

ALISON MACFARLANE & DIANA ELBOURNE

Migrant Studies

Studies on migrant populations are based on the assumption that migrants carry a risk that to some extent reflects that of their country of origin rather than the host country. Migrant studies are, therefore, a category of **ecologic studies** in which geographic differences in **risk** are replaced by risk differences among population groups (immigrants vs. host population and vs. population of origin).

Migration can be studied for three main reasons:

1. The study of migrant populations can be used for generating (as opposed to "testing") or confirming hypotheses derived from etiologic studies of environmental risk factors associated with disease occurrence.
2. The study of the health status of minorities who have emigrated from abroad has recently acquired public health significance, because of the effect of migration from the underdeveloped to the developed world on the occurrence of acute and socially relevant diseases in the host country.
3. Migrant status may be used in **case–control** and **cohort** studies as a variable representing possible **confounding** exposures.

Definition of Migrant Status

There are several ways of defining a subject as a migrant.

Place of birth Subjects born abroad are considered to be immigrants. This definition is the most widely used in epidemiologic studies. For diseased subjects, the information is either obtained from **death certificates**, which report the country of birth, or from **disease registries**, such as cancer registries. Furthermore, place of birth is usually enumerated in population **censuses**, but sometimes only for a subset of the population.

Citizenship Subjects with foreign citizenship are considered to be immigrants. The original citizenship may be retained by residents of foreign countries or obtained by the spouse of a migrant. The foreign offices of some countries provide periodic information on these persons.

Ethnic origin Subjects may be considered as migrants if both parents were born abroad or if they answer positively to the question: "Are you a migrant?" (*see* **Ethnic Groups**).

Each of the above-mentioned definitions identifies a population group of different size. For example, Table 1 shows the number of Italian migrants in the US in the decade 1970–1980 according to each definition [8].

Sources of Information for Migrant Studies

Information on Diseased Subjects

First-Generation Migrants. In the majority of migrant studies, information on diseased subjects is derived from routine **surveillance** systems. These are mortality statistics, when the content of the death certificates allows it, or cancer registries statistics, when the interest is focused on cancer risk. Other pathology reporting systems have begun to include information on migrant status in order to evaluate the effect of migration on the epidemiology of some diseases of emerging interest, such as tuberculosis and **AIDS**.

If information on the date of migration is recorded individually, the duration of stay and the age at migration to the host country can also be computed [4, 36]. This information is, however, seldom routinely available. In the US, the Social Security Number (SSN), which is assigned sequentially to all residents, has been used as a proxy of age at migration. If the SSN is assigned to a migrant after the usual age of entry to work, it is considered most likely that he/she migrated as an adult (late migrant). In contrast, those whose

Table 1 Number of Italian migrants in the US in the decade 1970–1980

Italian-born	831 000
Italian citizens	230 000
First- and second-generation immigrants	5 000 000
Italian origin	8 800 000
Italian origin identified as one of the subject's roots	12 180 000

Source: [4]. Reproduced with permission of IARC.

SSN was assigned at the usual age of initial employment are considered to have migrated in childhood (early migrant). Unfortunately, such information can be used only for cases and not for the general population, thus limiting the choice of study design [24].

Second-Generation Migrants. Parents' birthplace is routinely recorded on death certificates in some countries and by some cancer registries, allowing for the identification of second-generation migrants [9, 35]. Alternatively, studies on second-generation migrants may be based on information on both ethnicity and birthplace. Members of ethnic groups born locally are considered to be second-generation migrants, while those born abroad are first-generation migrants [39].

If second-generation migrants can be identified, the modification of risk between first-generation migrants and their descendants can be estimated. Moreover, disease risk can be studied in individuals of mixed parentage, in whom the genetic susceptibly may be intermediate between those of the two populations; furthermore, environmental exposures, such as lifestyle habits, may be influenced to different extents by the origin of the father and the mother [9, 35].

Information on the Population

First-Generation Migrants. To estimate incidence or mortality rates by migrant status, information is required on the population at risk of developing the disease under study. This may be provided by censuses, as long as the definition of migrant status in the denominator is the same as that in the numerator.

The use of censuses as a source of information for the denominator tends to limit the number of variables that can be considered in the study, as time since migration, age at migration, and other variables for diseased or deceased subjects, are seldom available for the population at risk.

Second-Generation Migrants. Only some population censuses include the country of birth of the parents of the enumerated subjects [35]. When available, this allows estimation of disease and death rates for second-generations migrants.

Use of Census Information for Longitudinal Migrant Studies

The identification of a cohort of migrants (first- or second-generation) through censuses may allow follow-up and cross-linking with routine information on diseases and deaths [15] (*see* **Record Linkage**). The number of persons "lost to follow-up" in these cohorts, however, tends to be high, because of the tendency of people who migrate once to move again, often back to the country of origin.

Information on Other Variables: Socioeconomic Level and Lifestyle

In some countries, routine sources of numerators and denominators provide some information on socioeconomic level, usually approximated by occupation, educational level, or a combination of the two [6] (*see* **Social Classifications**).

Surveys of the frequency of exposure to disease determinants (tobacco, alcohol, dietary habits) at a population level seldom include information on migrant status [22, 25]. When this is not available, exposure **prevalence** derived from population-based surveys in the country of origin, and in the host country, may be used to interpret differences in the disease risk of migrants [28].

Information on the prevalence of lifestyle habits of migrants may be derived from **control** groups in case–control studies in which migrant status is considered [38].

Exposure to lifestyle, environmental, and other risk factors in migrants, and in control groups, has been determined directly in only a few studies. It can be done through the use of questionnaires in cohort or case–control studies and makes it possible to disentangle the roles of different exposures in determining the risk pattern related to migrant status. The exposures of interest tend to vary widely among first- and second-generation migrants and in relation to duration of stay, providing greater **power** to detect associations and therefore smaller study size in comparison with populations in which the level of exposure is more homogeneous. Furthermore, direct measurement of exposure makes it possible to study the relationship between time variables, such as age and time at migration, and lifestyle changes. Most studies [40] have addressed the role of diet in determining the risk for cancers at various sites. In some

cases, blood samples were obtained so that internal doses of nutrients and micronutrients could be measured in a prospective **cohort** design [30].

Sources of Bias in Migrant Studies

If a different definition of migrant status is used for the denominator than for the numerator in computing mortality or morbidity rates, erroneous estimates of the rates in an unpredictable direction may result (*see* **Denominator Difficulties**).

Because of the infrequency of censuses (commonly at 10-year intervals), censuses must often be interpolated to estimate appropriate denominators for migrants. This may introduce an additional source of **bias**, due to an underestimate of the denominator, when active migration is still occurring during the period of interpolation.

The accuracy of diagnostic information and of the coding of diseases changes geographically. This may lead to artifactual differences when the rates in one country (local-born and migrants) are compared with those in another (country of origin), due to information bias on disease status. If diagnostic procedures and coding practices are not selective by migrant status, this source of bias does not affect comparisons between migrants and locally born people within the host country. It can be hypothesized, however, that the access to certain diagnostic procedures may be different for migrants within the same country, especially for those of low socioeconomic status, owing to communication problems or legal status. This will introduce bias in disease status, partially hampering comparisons of rates with those of locally born persons.

Furthermore, a possible selection of subjects who migrate, in contrast to the population of the country of origin, must be considered.

First, migration may be selective by subarea within the country of origin [8]. If there are different patterns of risk in the population of origin by subarea, any comparison between migrants and the population of country of origin as a whole will be incorrect. If information on the subarea of origin is available, however, a specific comparison with the subarea may be accomplished.

Secondly, subjects who decide to migrate may have a different disease occurrence pattern from that of the population as a whole, as health status is related to the opportunity to migrate. This **selection bias** usually leads in the direction of a lower disease risk among migrants (healthy migrant effect) [26]; however, it may be associated with a higher disease risk if diseased subjects tend to join a family that has previously migrated, or tend to migrate to another country for retirement or care (unhealthy migrant effect). To evaluate the relevance of this selection bias, the disease experience in the first period after migration is sometimes considered separately, when information on the date of migration is available [36] (*see* **Bias, Overview**).

Finally, if migrants are reluctant to use unfamiliar medical services or unable to afford to do so, they may "go back home" when severely ill, thus disappearing from the numerator while still contributing to the denominator [29]. This will lead to an underestimate of mortality and incidence rates.

Statistical Methods

All classical **descriptive epidemiological** methods can be used in migrant studies.

Variables Under Study

Migrant status is the exposure variable for which disease risk is estimated: the reference category is represented by nonmigrants in the host country or by subjects resident in the country or origin. Time variables, when available, can be investigated.

Age. The first time variable to be considered is age, because changes in risk result from aging and because of the peculiar age structure of migrant populations, which tend to an overrepresentation of young adults, especially in recently migrated groups. Age is, therefore, associated with both disease and migrant status (*see* **Confounding**).

Calendar Time. Disease rates are likely to change over calendar time. If this happens differentially in the country of origin, in the host country, and in the migrant population, any comparison should take into account the effect of such changes.

Duration of Stay and Age at Migration. Duration of stay is an index of duration of exposure to determinants of the disease under study during the stay in

the host country, and therefore represents a proxy of cumulative exposure levels. If differences in disease risk between migrants and locally born people, and between migrants and the population of origin, are related to differential exposure levels, these should be found to be associated with the duration of stay. Furthermore the speed with which the disease pattern in migrants changes by duration of stay can be interpreted in terms of duration of exposure (or non-exposure) to etiologic agents in the host country.

Age at migration is strictly related to duration of stay, as subjects migrating at younger ages tend to stay longer. In terms of the natural history of the disease, this variable is potentially informative for latency, estimated as the interval between the beginning of exposure and disease occurrence (*see* **Latent Period**). For example, the finding that subjects migrating from a high-risk to a low-risk country at a young age retain a life-long higher risk than that in the host country, and than that of people who migrated during adulthood, indicates that exposure during childhood is relevant for the disease. In cancer studies, and in general for diseases with a multistep etiologic process, it suggests that the determinants involved in migration act as initiators of the disease.

Ideally, one would examine simultaneously the effect of age on arrival and duration of stay in the host country, controlling for the other relevant temporal variables (age and period of occurrence of the disease). This is not feasible, however, as age, duration of stay, and age on arrival are not independent: the definition of two of them implies knowledge of the third. The situation is similar to that of the **age–period–cohort** problem framework. To overcome this difficulty, separate analyses sequentially ignoring one of the two variables (duration of stay or age on arrival) are performed in migrant studies [4].

Other Variables. Information on other variables related to both disease and exposure may be considered when available. As mentioned above, this is possible in case–control and cohort studies in which individual questionnaires are used. Additional variables can be derived from routine surveys, such as on occupation and education, as a proxy for socioeconomic status, and on place of residence in the host country as a proxy for access to diagnostic procedures and care (*see* **Surveys, Health and Morbidity**).

Statistical Analysis

The statistical analysis of migrant studies depends upon the data sources available.

Denominator-Based Analysis. *Age-Standardized Rates*. Iuf the information on the denominator is reliable, incidence or mortality rates for migrants can be calculated and compared with those of residents of the host country and/or the country of origin. Direct standardization is used to adjust for age distribution [32]. Standardized rate ratios (SRRs) are generally computed, since interest is focused on the magnitude of the difference between the two rates (migrants vs. locally born and/or migrants vs. country of origin) [3]. The choice of a common standard population (e.g. locally born in the host country) allows the use of rate ratios.

Direct standardization, however, is sensitive to small numbers of events in the study population [32]. For migrants, some age-specific rates may be based on very few cases. The result is unstable rates and large **confidence intervals** around the rate ratios.

To increase the precision of the measure, indirect standardization, from which a smaller **standard error** is expected, is used for rare diseases and for small migrant groups, with the estimation of Standardized Mortality or Incidence Ratios (SMRs, SIRs). The standard set of rates are the age-specific rates of the host country as a whole or, more properly, of the local-born, from which migrants have been excluded [42]. It must be considered, however, that SMRs and SIRs are internally standardized and not mutually comparable [32] (*see* **Standardization Methods**).

Other Methods of Adjustment. When variables other than age are considered in a study, the **Mantel–Haenszel** estimator can be used to obtain a summary estimate of risk, adjusted by age and by the other variables. **Loglinear models**, however, are currently preferred for migrant studies, when several variables are available for analysis and **stratification** would fail because of insufficient numbers.

Recently, loglinear modeling based on the **Poisson distribution** has been applied to migrant studies in order to control simultaneously for a number of confounding factors [17]. This application is based on the following two assumptions:

1. The number of cases per cell is assumed to follow a Poisson distribution, with a **mean** value proportional to the number of **person-years at risk**.
2. The logarithm of the rate is assumed to be a linear function of the combination of classification variables that best describe the disease risk in the migrant population.

The **relative risks** obtained by model fitting are expected to show a greater numerical stability in comparison with those computed by traditional standardization methods.

A comparison of risk estimates with 95% confidence intervals obtained when different methods were applied to a large set of mortality data for Italian migrants to Canada is shown in Table 2 for deaths from selected cancers [19].

Numerator-Based Analysis. When the denominator is not available, or the population at risk cannot be cross-classified by the variables of interest, the analysis is based on diseased/deceased subjects only. The **proportional mortality** or incidence ratio (PMR or PIR) is the measure often used in such cases [36], and the relative proportion of diseases in the locally born population other than the one of interest is taken as the standard to adjust by age.

A proportional mortality or incidence study can be classified as a variant of a case–control study, where the cases are deaths or incident events classified by migrant status and the controls are other deaths or incident events of a different disease occurring in the same base population. This study design is based on the assumption that the migrant status among the controls has the same distribution as in the base population, i.e. that the overall rate of the disease/s in the controls is not related to migration.

Instead of PMRs or PIRs, a Mantel–Haenszel or, more frequently, a loglinear modeling approach is used in numerator-based studies when variables of interest other than age are considered. The assumptions described for PMR and PIR studies are used. When **logistic regression** models are applied, the cases in the cells are assumed to follow a **binomial distribution**, and the logit **transformation** of the disease probability is considered to be a linear function of the classification variables. If the assumption that the disease or death risk in controls is not related to migration is true, the estimates from the logistic model approximate those derived from the Poisson regression. The choice of appropriate controls is therefore crucial in this study design.

A comparison of the results obtained from the same set of cancer deaths among Italian migrants to Australia using Poisson and logistic regression (the latter using three sets of controls) is shown in Table 3 [19]. The risk estimates obtained using non-cancer or all other deaths as controls are consistently greater than those obtained using cancer controls or Poisson regression. This result is due to a lower risk of death from all causes and from causes other than cancer in migrants than among locally born persons. The results, therefore, do not confirm the assumption that the disease in controls are unrelated to migration when these two sets of controls are considered.

Evaluation of Goodness of Fit of Regression Models – the "Overdispersion" Phenomenon. When the analysis is based on modeling, **Goodness of**

Table 2 Comparison of age-adjusted estimates of risks and their 95% confidence intervals obtained by different methods, for male Italian migrants relative to locally born, Canada, 1964–1985

Cancer site	SRR[a]	SMR[b]	RR[c]	RR[d]
Esophagus	0.72(0.60–0.87)	0.69(0.56–0.82)	0.69(0.58–0.83)	0.69(0.57–0.83)
Stomach	1.28(1.18–1.39)	1.30(1.20–1.40)	1.30(1.20–1.40)	1.30(1.20–1.40)
Lung	0.76(0.72–0.80)	0.74(0.70–0.78)	0.74(0.70–0.78)	0.74(0.70–0.78)
Melanoma	0.96(0.71–1.29)	0.86(0.63–1.09)	0.86(0.65–1.14)	0.86(0.65–1.14)
Leukemia	1.18(1.05–1.33)	1.18(1.05–1.31)	1.18(1.05–1.32)	1.18(1.05–1.32)

[a]Standardized rate ratio (direct standardization).
[b]Standardized mortality ratio (indirect standardization).
[c]Relative risk estimates according to the Mantel–Haenszel procedure.
[d]Relative risk estimates according to the Poisson regression procedure.
Source: [4]. Reproduced with permission of IARC.

Table 3 Comparison of age-adjusted estimates of risks and their 95% confidence intervals obtained by Poisson regression and logistic regression, with different choices of controls, for male Italian migrants relative to locally born, Australia, 1964–1985

Cancer site	Relative risk (Poisson regression)	Relative risk (Logistic regression)		
		Controls		
		Other-cancer deaths	Noncancer deaths	Noncancer and other-cancer deaths
Stomach	1.45(1.16–1.82)	1.75(1.61–1.91)	2.39(2.20–2.60)	2.23(2.05–2.42)
Lung	0.95(0.85–1.07)	1.19(1.12–1.26)	1.61(1.52–1.70)	1.51(1.44–1.60)
Melanoma	0.27(0.18–0.40)	0.32(0.24–0.42)	0.49(0.37–0.64)	0.44(0.34–0.57)

Source: [4]. Reproduced with permission of IARC.

fit can be assessed with the log **likelihood ratio** statistic [17]. Provided that the Poisson or binomial assumptions hold, and the regression model is correctly specified, this statistic is of the same magnitude as the **degrees of freedom**, or smaller for small cell sample size. However, especially when a large data set is used and the **contingency table** is not classified by factors that are relevant to the response, the phenomenon of **overdispersion** may occur, reflected in a log-likelihood ratio greater than predicted.

This case occurs frequently in migrant studies, especially when the comparison is between those born in the host country and the general population in the country of origin, thus involving very large data sets with few **explanatory variables** available for the analysis. The problem of overdispersion can be addressed in the analysis by using a conservative approach in estimating the confidence intervals of the effect parameters [1].

Contribution of Migrant Studies to Insight into Disease Etiology

Most studies of migrant populations address cancer incidence or mortality (for some relevant references on this issue, see Geddes et al. [10], Steinitz et al. [36], Haenszel [13], Haenszel et al. [14], and Thomas & Karagas [40]). The published studies refer to migration from high- to low-risk countries for some cancer sites (e.g. stomach cancer in migrants from Japan and Italy to the US and Australia) or from low- to high-risk countries (e.g. breast cancer in migrants from Japan and China to the US). The analysis of temporal variables, such as duration of stay in the host country, and age at migration, and the study of second-generation migrants have provided valuable information on the size and timing of changes in cancer risk in response to changes in the external environment and/or lifestyle [9, 20, 35].

Another result of migrant studies is information on cancer rates in the migrant's country of origin, when these are not currently recorded or reasonably valid. This is of particular value for migrants from the underdeveloped world who have recently migrated to developed countries. Such estimates should, however, be considered with caution because of the possible **selection bias** of migrants (see above).

The relation of risk to the frequency of exposure to dietary factors, as derived from **cross-sectional studies** of migrants, their offspring, host countries, and countries of origin, has been highlighted in some studies [22, 23, 28].

The results of studies in which dietary and other variables were considered have been used to infer possible environmental factors in the geographic differences in cancer rates. Although such studies are not common, they represent a potential field of development in migrant studies.

Other migrant studies, mainly based on mortality data, address cardiovascular disease and stroke [15, 16, 37]. In these studies, temporal variables and measurement of risk are treated by methods similar to those used in cancer studies.

Another developing field of interest is of diseases and accidents that are suspected of being determined by migration itself, or which are prominent in the migrant population, thus affecting the rates in the host country. This is the case of studies on suicide [21],

homicide [34], work-related fatalities [7], tuberculosis [5, 27], birth outcomes [2, 12, 33], psychiatric disorders [31], hepatitis [18, 41] and HIV/**AIDS** [11].

Most of these studies, however, do not involve use of the methods described above, and risk estimates are not provided for the migrant population in comparison with the locally born population or with the country of origin. This may be due to the lack of population-based registries for the diseases under study and to the relatively small groups of subjects involved.

References

[1] Aitkin, M., Anderson, D., Francis, B. & Hinde, J. (1989). *Statistical Modelling in Glim.* Clarendon Press, Oxford.

[2] Alexander, G.R., Mor, J.M., Kogan, M.D., Leland, N.L. & Kieffer, E. (1996). Pregnancy outcomes of US-born and Japanese Americans, *American Journal of Public Health* **86**, 820–824.

[3] Armstrong, B.K., Woodings, T.L., Stenhouse, N.S. & McCall, M.G. (1983). *Mortality from Cancer in Migrants to Australia 1962 to 1971*. NH & MCR Research Unit in Epidemiology and Preventive Medicine, Raine Medical Statistics Unit, Perth, University of Western Australia.

[4] Balzi, D., Khlat, M. & Matos, E. (1993). Australia: mortality data, in *Cancer in Italian Migrant Populations*, M. Geddes, D.M. Parkin, M. Khlat, D. Balzi, & E. Buiatti, eds. IARC Scientific Publications, Lyon, pp. 125–137.

[5] Bhatti, N., Law, M.R., Morris, J.K., Halliday, R. & Moore-Gillon, J. (1995). Increasing incidence of tuberculosis in England and Wales: a study of the likely causes, *British Medical Journal* **310**, 976–979.

[6] Bouchardy, C. (1993). France, in *Cancer in Italian Migrant Populations*, M. Geddes, D.M. Parkin, M. Khlat, D. Balzi, & E. Buiatti, eds. IARC Scientific Publications, Lyon, pp. 149–159.

[7] Corvalan, C.F., Driscoll, T.R. & Harrison, J.E. (1994). Role of migrant factors in work-related fatalities in Australia, *Scandinavian Journal of Work and Environmental Health* **20**, 364–370.

[8] Geddes, M. (1993). Italian migration: an overview, in *Cancer in Italian Migrant Populations*, M. Geddes, D.M. Parkin, M. Khlat, D. Balzi & E. Buiatti, eds. IARC Scientific Publications, Lyon, pp. 11–19.

[9] Geddes, M., Balzi, D., Buiatti, E., Brancker, A. & Parkin, D.M. (1994). Cancer mortality in Italian migrants to Canada, *Tumori* **80**, 19–23.

[10] Geddes, M., Parkin, D.M., Khlat, M., Balzi, D. & Buiatti, E. (1993). *Cancer in Italian Migrant Populations*. IARC Scientific Publications, Lyon.

[11] Gellert, G.A., Maxwell, R.M., Higgins, K.V., Mai, K.K., Lowery, R. & Doll, L. (1995). HIV/AIDS knowledge and high risk sexual practices among southern California Vietnamese, *Genitourinarian Medicine* **71**, 216–223.

[12] Guendelman, S. & English, P.B. (1995). Effect of United States residence on birth outcomes among Mexicans immigrants: an exploratory study, *American Journal of Public Health* **142**, S30–S38.

[13] Haenszel, W. (1961). Cancer mortality among the foreign-born in the United States, *Journal of the National Cancer Institute* **26**, 37–132.

[14] Haenszel, W., Kurihara, M., Segi, M. & Lee R.K.C. (1972). Stomach cancer among Japanese in Hawaii, *Journal of the National Cancer Institute* **49**, 969–988.

[15] Harding, S. & Balarajan, R. (1996). Pattern of mortality in second generation Irish living in England and Wales: longitudinal study, *British Medical Journal* **312**, 1389–1392.

[16] Kagan, A., Harris, B.R., Winkelstein, W., Johnson, K.G., Kato, H., Syme, S.L., Rhoods, G.G., Gay, M.L., Nichaman, M.Z., Hamilton, H.B. & Tillotson, J. (1974). Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: demographic, physical, dietary and biochemical characteristics, *Journal of Chronic Disease* **27**, 345–364.

[17] Kaldor, J., Khlat, M., Parkin, D.M., Shiboski, S. & Steinitz, R. (1990). Log-linear models for cancer risk among migrants, *International Journal of Epidemiology* **19**, 233–239.

[18] Karetnyi, Y.V., Mendelson, E., Shlyakhov, E., Rubinstein, E., Golubev, N., Levin, R., Sandler, M., Schreiber, M., Rubinstein, U., Shif, I., Handsher, R., Varsano, N. & Modan, B. (1995). Prevalence of antibodies against hepatitis A virus among new immigrants in Israel, *Journal of Medical Virology* **46**, 61–65.

[19] Khlat, M. & Balzi, D. (1993). Statistical methods, in *Cancer in Italian Migrant Populations*, M. Geddes, D.M. Parkin, M. Khlat, D. Balzi & E. Buiatti, eds. IARC Scientific Publications, Lyon, pp. 37–47.

[20] Khlat, M., Vail, A., Parkin, D.M. & Green, A. (1992). Mortality from melanoma in migrants to Australia: variation by age at arrival and duration of stay, *American Journal of Epidemiology* **135**, 1103–1113.

[21] Kliewer, E.V. & Ward, R.H. (1988). Convergence of immigrant suicide rates to those in the destination country, *American Journal of Epidemiology* **127**, 640–648.

[22] Kolonel, L., Hinds, W. & Hankin, J. (1980). Cancer patterns among migrant and native-born Japanese in Hawaii in relation to smoking, drinking and dietary habits, in *Genetic and Environmental Factors in Experimental and Human Cancer*, H.V. Gelboin, B. MacMahon, T. Matsushima, T. Sugimura, S. Takayama & H. Takabe, eds. Japanese Scientific Societies Press, Tokyo, pp. 327–340.

[23] Kolonel, L.N., Nomura, A.M.Y., Hirohata, T., Hankin, J.H. & Hinds, M.W. (1981). Association of diet and place of birth with stomach cancer incidence in Hawaii Japanese and Caucasians, *American Journal of Clinical Nutrition* **34**, 2478–2485.

[24] Mack, T.M., Walker, A., Mack, W. & Bernstein, L. (1985). Cancer in Hispanics in Los Angeles County, *National Cancer Institute Monograph* **69**, 99–104.

[25] Margetts, B.M., Hopkins, S.M., Binns, C.W., Miller, M.R. & Armstrong, B.K. (1981). Nutrient intakes in Italian migrants and Australians in Perth, *Food and Nutrition* **31**, 7–10.

[26] Marmot, M.G., Adelstein, A.M. & Bulusu, L. (1984). Immigrant mortality in England and Wales 1970–78: causes of death by country of birth, *Studies on Medical and Population Subjects*. HMSO, London.

[27] McKenna, M.T., McCray, E. & Onorato, I. (1995). The epidemiology of tuberculosis among foreign-born persons in the United States, 1986 to 1993, *New England Journal of Medicine* **332**, 1071–1076.

[28] McMichael, A.J., McCall, M.G., Hartshorne, J.M. & Woodings, T.L. (1980). Patterns of gastro-intestinal cancer in European migrants to Australia: the role of dietary change, *International Journal of Cancer* **25**, 431–437.

[29] Muir, C.S. & Staszewski, J. (1986). Geographical epidemiology and migrant studies, in *Biochemical and Molecular Epidemiology of Cancer*, C. Harris, ed. Liss, New York, pp. 135–148.

[30] Nomura, A.M., Stemmermann, G.N., Heilbrun, L.K., Salked, R.M. & Vuilleumier, J.P. (1985). Serum vitamin levels and the risk of cancer of specific sites in men of Japanese ancestry in Hawaii, *Cancer Research* **45**, 2369–2372.

[31] Roberts, N. & Cawthorpe, D. (1995). Immigrant child and adolescent psychiatric referrals: a five-year retrospective study of Asian and Caucasian families, *Canadian Journal of Psychiatry* **40**, 252–256.

[32] Rothman, K.J. (1986). *Modern Epidemiology*, Little, Brown, & Company, Boston.

[33] Singh, G.K. & Yu, S.M. (1996). Adverse pregnancy outcomes: differences between US and foreign-born women in major US racial and ethnic groups, *American Journal of Public Health* **86**, 837–843.

[34] Sorenson, S.B. & Shen, H. (1996). Homicide risk among immigrants in California, 1970 through 1992, *American Journal of Public Health* **86**, 97–100.

[35] Steinitz, R., Iscovich, J.N. & Katz, L. (1990). Cancer incidence in young offspring of Jewish immigrants to Israel: a methodological study, I: Nasopharyngeal malignancies and Ewing sarcoma, *Cancer Detection and Prevention* **14**, 547–553.

[36] Steinitz, R., Parkin, D.M., Young, J.L., Bieber, C.A. & Katz, L. (1989). *Cancer Incidence in Jewish Migrants to Israel 1961–1981*. IARC Scientific Publications, Lyon.

[37] Stenhouse, N.S. & McCall, M.G. (1970). Differential mortality from cardiovascular disease in migrants from England and Wales, Scotland and Italy, and native-born Australians, *Journal of Chronic Diseases* **23**, 423–431.

[38] Terracini, B., Siemiatycki, J. & Richardson, L. (1990). Cancer incidence and risk factors among Montreal residents of Italian origin, *International Journal of Epidemiology* **19**, 491–497.

[39] Thomas, D.B. & Karagas, M.R. (1987). Cancer in first and second generation Americans, *Cancer Research* **47**, 5771–5776.

[40] Thomas, D.B. & Karagas, M.R. (1996). Migrant studies, in *Cancer Epidemiology and Prevention*, D. Schottenfeld & J.F. Fraumeni Jr, eds. Oxford, University Press, Oxford, pp. 236–254.

[41] Trautwein, C., Kiral, G., Tillmann, H.L., Witteler, H., Michel, G. & Manns, M.P. (1995). Risk factors and prevalence of hepatitis E in German immigrants from the former Soviet Union, *Journal of Medical Virology* **45**, 429–434.

[42] Wang, Z.J., Ramcharan, S. & Love, E.J. (1989). Cancer mortality of Chinese in Canada, *International Journal of Epidemiology* **18**, 17–21.

E. BUIATTI & D. BALZI

Migration Processes

In a migration process, both immigration and emigration take place. Immigration causes a population size to increase, while emigration causes a population size to decrease, just as in a birth–death process (*see* **Stochastic Processes**). However, there is a fundamental difference between the two processes. In a migration process the immigration rate (immigration intensity) is independent of the population size, whereas in a birth–death process the birth rate (birth intensity) is a function of the population size at the time of birth. This difference affects the complexity of the two processes. While the formulas in a birth–death process usually are complicated, especially when the birth intensity and the death intensity are functions of time, the formulas in a migration process are relatively simple. Generally, in a single colony migration process, the population size distribution is a combination of a **binomial** distribution and a **Poisson distribution**. The binomial distribution is related to the initial population size, and the Poisson distribution is associated with migration. The two distributions are independent of each other. If the initial population size is zero, then the population size distribution is Poisson. If there is no immigration, then the population size distribution reduces to a binomial distribution, as shown below.

A Simple Migration Process

Suppose that a population's growth is subject to a migration process with immigration intensity η and an emigration intensity μ. Let $X(t)$ be the population size at time t, with the initial population size at $t = 0$, $X(0) = i$. Let the probability distribution of $X(t)$ be denoted by

$$P_{ik}(0,t) = \Pr[X(t) = k | x(0) = i], \quad k = 0, 1, \ldots. \tag{1}$$

We derive a system of differential equations for $P_{ik}(0,t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{i0}(0,t) = -\eta P_{i0}(0,t) + \mu P_{i1}(0,t) \tag{2}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{ik}(0,t) = -(\eta + k\mu) P_{ik}(0,t) + \eta P_{i,k-1}(0,t) + (k+1)\mu P_{i,k+1}(0,t), \tag{3}$$

with the initial conditions at $t = 0$:

$$P_{ii}(0,0) = 1 \quad \text{and} \quad P_{ik}(0,0) = 0, \quad \text{for } k \neq i. \tag{4}$$

Each of the differential equations in (3) contains three unknown probabilities – $P_{ik}(0,t)$, $P_{i,k-1}(0,t)$, and $P_{i,k+1}(0,t)$ – and cannot be solved directly. We resort to the method of probability **generating functions** [4].

Let the probability generating function of $X(t)$ be denoted by $G_x(s;t)$, so that

$$G_X(s;t) = \sum_{k=0}^{\infty} s^k P_{ik}(0,t), \tag{5}$$

with the initial condition at $t = 0$:

$$G_X(s;0) = s^i. \tag{6}$$

Taking the derivatives of (5) with respect to t, we find a partial differential equation for the probability generating function:

$$\frac{\partial}{\partial t} G_X(s;t) = \mu(1-s)\frac{\partial}{\partial s} G_X(s;t) - \eta(1-s) G_X(s;t). \tag{7}$$

Solving (7), with the initial condition (6), we find

$$G_X(s;t) = [1 - (1-s)\exp(-\mu t)]^i \times \exp\left\{-(1-s)\frac{\eta}{\mu}[1 - \exp(-\mu t)]\right\}. \tag{8}$$

When $s = 1$, $G_x(1,t) = 1$, so the distribution of $X(t)$ is proper. For each $t > 0$, we can find the probability $P_{ik}(0,t)$, for each k, the expectation $\mathrm{E}[X(t)]$, and the variance of $X(t)$, by taking appropriate derivatives of $G_x(s;t)$, although the computations are quite involved. However, these quantities can be derived directly with a different interpretation of (8).

Formula (8) is a product of two factors:

$$G_X(s;t) = g_Y(s;t) \times g_Z(s;t),$$

where

$$g_Y(s;t) = [1 - (1-s)\exp(-\mu t)]^i \tag{9}$$

and

$$g_Z(s;t) = \exp\left\{-(1-s)\frac{\eta}{\mu}[1 - \exp(-\mu t)]\right\}. \tag{10}$$

Formula (9) is the probability generating function of a binomial random variable, say $Y(t)$, with parameters i and $\exp(-\mu t)$; while formula (10) is the probability generating function of a **Poisson** random variable, say $Z(t)$, with a parameter function

$$\left(\frac{\eta}{\mu}\right)[1-\exp(-\mu t)]. \tag{11}$$

According to a theorem in probability generating functions, $X(t)$ is the sum of two independently distributed random variables,

$$X(t) = Y(t) + Z(t),$$

with formulas (9) and (10) as their respective probability generating functions. It follows that the distribution of $X(t)$ is a convolution of the distributions of $Y(t)$ and $Z(t)$, and

$$P_{ik}(0,t) = \sum_{j=0}^{\min[i,k]} \Pr[Y(t)=j] \times \Pr[Z(t)=k-j],$$

where $\min[i,k]$ stands for the smaller of i and k,

$$\Pr[Y(t)=j] = \binom{i}{j} \exp(-j\mu t)[1-\exp(-\mu t)]^{i-j}$$

and

$$\Pr[Z(t)=j] = \frac{\left(\frac{\eta}{\mu}\right)^j [1-\exp(-\mu t)]^j}{j!} \times \exp\left\{-\frac{\eta}{\mu}[1-\exp(-\mu t)]\right\}.$$

The expectation of $X(t)$ is

$$\mathrm{E}[X(t)] = \mathrm{E}[Y(t)] + \mathrm{E}[Z(t)] = i\exp(-\mu t) + \frac{\eta}{\mu}[1-\exp(-\mu t)], \tag{12}$$

and the variance of $X(t)$ is

$$\mathrm{var}[X(t)] = i\exp(-\mu t)[1-\exp(-\mu t)] + \frac{\eta}{\mu}[1-\exp(-\mu t)].$$

If the initial population size is zero, $i = 0$, then $X(t) = Z(t)$ has a Poisson distribution; if there is no immigration, $\eta = 0$, then $X(t) = Y(t)$ has a binomial distribution, as noted earlier. In general, $Y(t)$ and $Z(t)$ correspond, respectively, to individuals that were and were not present in the population at time $t = 0$.

In the above discussion we have assumed that the migration intensities were constant. When they are functions of time, $\eta(t)$ and $\mu(t)$, the binomial probability becomes

$$\exp\left[-\int_0^t \mu(\tau)\,\mathrm{d}\tau\right] \tag{13}$$

and the Poisson parameter in (11) becomes

$$\int_0^t \eta(\tau)\exp\left[-\int_\tau^t \mu(\xi)\,\mathrm{d}\xi\right]\mathrm{d}\tau. \tag{14}$$

With the substitutions of (13) and (14), we will have the same formulas as before for the generating functions, the probabilities, and the expectations.

A Survival Distribution

In this distribution we assume that there are two forces continuously acting on an individual to influence his survival and death. One force causes the mortality intensity function to increase, while the other causes the mortality intensity function to decrease. As a concrete example, consider an individual who is continuously exposed to a low level of radiation and other toxic material in the environment. During a time interval $(\tau, \tau + \mathrm{d}\tau)$, for $0 < \tau < t$, there is a probability $\eta\,\mathrm{d}\tau + o(\mathrm{d}\tau)$ that the individual will absorb a unit of toxic material, and a probability $\nu\,\mathrm{d}\tau + o(\mathrm{d}\tau)$ that the biological reaction inside the human body will cause a unit of toxic material in the body to be discharged. The units thus follow a migration process with initial population size $i = 0$, and, from (12), the expected number of units at time t is

$$\frac{\eta}{\nu}[1-\exp(-\nu t)]. \tag{15}$$

This is the expected total amount of toxic material absorbed during the interval $(0, t)$ and is present at time t. Since the toxic material is supposed to be harmful to an individual, a reasonable assumption is that the force of mortality at time t, $\mu(t)$, should be a function of the quantity in (15). The simplest

function is proportional, so that the force of mortality (*see* **Hazard Rate**) at time t is

$$\mu(t) = \frac{\beta}{\nu}[1 - \exp(-\nu t)],$$

where $\beta = b\eta$, b being the proportionality coefficient. It follows that the cumulative survival function is

$$\int_0^t \mu(\tau)\,\mathrm{d}\tau = \frac{\beta}{\nu}\left\{t - \frac{1}{\nu}[1 - \exp(-\nu t)]\right\}.$$

(*See* **Survival Distributions and Their Characteristics**).

Let T be the survival time of an individual. Then the distribution of T is given by

$$F_T(t) = 1 - \exp\left\{-\frac{\beta}{\nu}\left[t - \frac{1}{\nu}[1 - \exp(-\nu t)]\right]\right\}.$$

As $t \to \infty$, the distribution function tends to 1, and therefore the distribution of the survival time T is proper. The expectation of T is

$$\mathrm{E}[T] = \frac{1}{\nu}c^{-c}\mathrm{e}^{c}\Gamma(c, c),$$

where $\Gamma(c, c)$ is an incomplete **gamma** function

$$\Gamma(c, c) = \int_0^c y^{c-1}\mathrm{e}^{-y}\,\mathrm{d}y,$$

$$y = \frac{\beta}{\nu^2}\exp(-\nu t) \quad \text{and} \quad c = \frac{\beta}{\nu^2}.$$

This distribution was proposed in Chiang & Conforti [5] and was useful in the estimation of the time to tumor.

A Multi-colony Migration Process

Suppose now that the population consists of m colonies, labeled $1, 2, \ldots, m$. For $i = 1, 2, \ldots, m$, let η_i and μ_i denote, respectively, the immigration and emigration intensities for colony i. Thus, the probability that an individual immigrates into the population via colony i in the time interval $(t, t + \Delta t)$ is $\eta_i \Delta t + o(\Delta t)$, and the probability that an individual, who is in colony i at time t, emigrates from the population during $(t, t + \Delta t)$ is $\mu_i \Delta t + o(\Delta t)$. For $i \neq j$, let v_{ij} denote the migration intensity from colony i to colony j, so the probability that an individual, who is in colony i at time t, migrates to colony j during $(t, t + \Delta t)$ is $v_{ij}\Delta t + o(\Delta t)$. Suppose that at time $t = 0$ there are n_i individuals in colony i ($i = 1, 2, \ldots, m$). This model is fairly straightforward to analyze, since the behaviors of distinct individuals follow independent Markov processes. Here, we outline the key results. Further details may be found in [6].

For $i = 1, 2, \ldots, m$ and $t \geq 0$, let $X_i(t)$ denote the number of individuals in colony i at time t, and suppose that of those $Y_i(t)$ are *original* (i.e. in the population at time $t = 0$) and $Z_i(t)$ are *new* (i.e. have immigrated into the population during $(0, t]$.) Let $\boldsymbol{X}(t) = (X_1(t), X_2(t), \ldots, X_m(t))^{\mathrm{T}}$, where T denotes transpose, and define $\boldsymbol{Y}(t)$ and $\boldsymbol{Z}(t)$ similarly. Then,

$$\boldsymbol{X}(t) = \boldsymbol{Y}(t) + \boldsymbol{Z}(t) \quad (t \geq 0), \tag{16}$$

and the independence of individuals implies that $\boldsymbol{Y}(t)$ and $\boldsymbol{Z}(t)$ are also independent.

The distributions of $\boldsymbol{Y}(t)$ and $\boldsymbol{Z}(t)$ can be described as follows. For $t \geq 0$, let $P(t) = [p_{ij}(t)]$, where $p_{ij}(t)$ is the probability that an individual is in colony j at time t given that it was in colony i at time 0 ($i, j = 1, 2, \ldots, m$). The movement of an individual among the colonies follows a continuous time Markov chain, that may be transient due to emigration. Standard theory for such processes implies that

$$P(t) = \exp(Vt), \tag{17}$$

where V is the $m \times m$ matrix with elements v_{ij}, if $i \neq j$, and $v_{ii} = -\left(\mu_i + \sum_{k \neq i} v_{ik}\right)$, and $\exp(Vt) = \sum_{k=0}^{\infty} t^k V^k / k!$ is the usual matrix exponential (see e.g. Bellman [1], page 169). For the formulae of the individual $p_{ij}(t)$, reference may be made to Chiang [3, 4].

For $i = 1, 2, \ldots, m$, let $\boldsymbol{Y}_i^*(t) = (Y_{i1}^*(t), Y_{i2}^*(t), \ldots, Y_{im}^*(t))^{\mathrm{T}}$, where $Y_{ij}^*(t)$ is the number of original colony-i individuals that are in colony j at time t. The independence of individuals implies that $\boldsymbol{Y}_i^*(t)$ follows the defective (owing to emigration) multinomial distribution given by

$$\Pr[Y_{i1}^*(t) = r_1, Y_{i2}^*(t) = r_2, \ldots, Y_{im}^*(t) = r_m] = n_i! \prod_{j=0}^{m} \frac{\{p_{ij}(t)\}^{r_j}}{r_j!} \left(r_j \geq 0\ (j = 1, 2, \ldots, m), \sum_{j=1}^{m} r_j \leq n_i\right), \tag{18}$$

where $r_0 = n_i - \sum_{j=1}^{m} r_j$ and $p_{io}(t) = 1 - \sum_{j=1}^{m} p_{ij}(t)$. Moreover, $\boldsymbol{Y}_1^*(t), \boldsymbol{Y}_2^*(t), \ldots, \boldsymbol{Y}_m^*(t)$ are independent. Now

$$Y_i(t) = \sum_{j=1}^{m} Y_{ji}^*(t) \quad (i = 1, 2, \ldots, m), \tag{19}$$

so the joint distribution of $\boldsymbol{Y}(t)$ is completely described.

Turning to the distribution of the numbers of new individuals $\boldsymbol{Z}(t)$, since individuals immigrate into the m colonies at the points of independent Poisson processes and they behave independently, $Z_1(t), Z_2(t), \ldots, Z_m(t)$ follow independent Poisson distributions, with means given by

$$E[Z_i(t)] = \sum_{j=1}^{m} \eta_j \int_0^t p_{ji}(t-u)\,\mathrm{d}u \qquad (t \geq 0; \quad i = 1, 2, \ldots, m). \tag{20}$$

Further, provided the immigration, emigration, and migration rates are such that any individual ultimately leaves the population with probability one, the integral in (20) can be obtained using

$$\int_0^t P(t-u)\,\mathrm{d}u = V^{-1}(\exp(Vt) - I). \tag{21}$$

Equations (16) to (21) specify completely the distribution of the population at any time $t > 0$. The matrix exponential $\exp(Vt)$ can be found in terms of the eigenvalues and eigenvectors of V, provided V admits a spectral decomposition (see e.g. [8]). Note that in the case when any individual ultimately leaves the population, in the limit as $t \to \infty$ the population consists entirely of new individuals and the sizes of the m colonies follow independent Poisson distributions, with means given by $-\boldsymbol{\eta}^{\mathrm{T}} V^{-1}$, where $\boldsymbol{\eta}^{\mathrm{T}} = (\eta_1, \eta_2, \ldots, \eta_m)$.

An extension of the above model, in which the immigration, emigration, and migration intensities are time-dependent is considered by Faddy [6]. The independent decomposition (16) into original and new individuals still holds, and the distributions of $\boldsymbol{Y}(t)$ and $\boldsymbol{Z}(t)$ can still be described in terms of independent multinomial and Poisson distributions. However, the time-dependent nature of the intensities implies that $p_{ij}(t)$ has to be replaced by $p_{ij}(s, t)$ $(t > s)$, where $p_{ij}(s, t)$ is the probability that an individual that is in colony i at time s is in colony j at time t. Moreover, $p_{ij}(s, t)$ can be found explicitly only in a few special cases (see e.g. [2] and [9]).

Multi-colony migration processes also find application in other settings such as **compartment models** and queueing networks. In the latter, the intensities are often functions of the population state and interest usually focuses on the equilibrium distribution of the system (see e.g. [7, Chapter 2]). The single colony model discussed earlier also describes a queue with infinitely many servers (*see* **Queuing Processes**).

References

[1] Bellman, R. (1970). *Introduction to Matrix Analysis*. 2nd edn. McCraw-Hill, New York.

[2] Cardenas, M. & Matis, J.H. (1975). On the time-dependent reversible stochastic compartmental model – II. A class of n-compartment systems. *Bull Math Biol* **37**, 555–564.

[3] Chiang, C.L. (1964). A stochastic model of competing risks of illness and competing risks of death. *Stochastic Models in Medicine and Biology* (J. Gurland, editor), University of Wisconsin Press, Madison, 323–354.

[4] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Kreiger, New York.

[5] Chiang, C.L. & Conforti, P. (1989). A survival model and estimation of time to tumor. *Mathematical Biosciences* **94**, 1–29.

[6] Faddy, M.J. (1977). Stochastic compartmental models as approximations to more general stochastic systems with the general stochastic epidemic as an example. *Adv Appl Prob* **9**, 448–461.

[7] Kelly, F.P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester.

[8] McClean, S.I. (1976). A continuous-time population model with Poisson recruitment. *J Appl Prob* **13**, 348–354.

[9] Raman, S. & Chiang, C.L. (1973). On a solution of the migration process and the application to a problem in epidemiology. *J Appl Prob* **10**, 718–727.

CHIN LONG CHIANG & FRANK BALL

Minimax Theory

Minimax strategies are a pivotal concept in the theory of games. They were independently introduced in this context in the 1920s by Borel [2, 3] and von Neumann [23]. The 1944 monograph by von Neumann & Morgenstern [24] was very influential in fostering a coherent formulation and in explaining their role in game theory and particularly in the theory of two-person, zero-sum games. For a thorough treatment of game theory and of the place of minimaxity therein, see Luce & Raiffa [18].

In statistical theory minimaxity first appears in the seminal 1939 paper of Wald [25]. This paper laid the foundations of statistical **decision theory**, and the concept of minimaxity flowed naturally after the definitions there of **loss** and **risk**. See Brown [6] for discussion of the extent this development was influenced by Wald's earlier contacts with Morgenstern.

Minimax considerations are based on the risk function. Let $\mathscr{F}$ denote the set of possible distributions in a statistical decision problem and let Δ denote the set of randomized decision functions. Then the risk, $R(F, \delta)$, is the expected loss to the statistician who uses the procedure $\delta \in \Delta$ when the true distribution is $F \in \mathscr{F}$. Conventionally, one assumes the loss is nonnegative, and hence $R \geq 0$. Small values of R are desirable to the statistician, but since F is unknown these cannot necessarily be obtained.

The minimax risk is defined to be $M = \inf_{\delta \in \Delta} \sup_{F \in \mathscr{F}} R(F, \delta)$. A procedure δ_ε with $\sup_{F \in \mathscr{F}} R(F, \delta_\varepsilon) \leq M + \varepsilon$ is called ε-minimax ($\varepsilon \geq 0$). When $\varepsilon = 0$, the corresponding δ_0 is called minimax.

Various mathematical results characterizing existence and structure of minimax procedures involve putative **prior distributions** on $\mathscr{F}$ (endowed with a suitable σ-field). Let $\mathcal{P}$ denote the class of prior distributions, let $R^*(P, \delta) = \mathrm{E}_p(R(F, \delta))$ denote the expected risk under $P \in \mathcal{P}$, and let δ_p denote a corresponding Bayes procedure, i.e. one for which $R^*(P, \delta_p) = \inf_{\delta \in \Delta} R^*(P, \delta)$.

The fundamental minimax theorem is valid under certain important regularity conditions. When valid it asserts the existence of a minimax procedure which is Bayes for a corresponding prior, P_0, called a least favorable prior. This yields the following important string of equalities:

$$\sup_{p \in \mathcal{P}} \inf_{\delta \in \Delta} R^*(P, \delta) = M = R^*(P_0, \delta_0) = \inf_{\delta \in \Delta} \sup_{p \in \mathcal{P}} R^*(P, \delta_0). \qquad (1)$$

Regularity conditions implying (1) and (2) can be found in Wald [26], LeCam [16], and Brown [4, 5]. Much milder conditions imply the existence of a minimax procedure and a least favorable sequence of priors, $\{P_i : i = 1, \ldots\}$, such that

$$\sup_i R^*(P_i, \delta_{p_i}) = M. \qquad (2)$$

Virtually all statisticians have agreed that in the presence of confidently held prior probability beliefs a Bayes procedure should be used. Of course, there has been and continues to be considerable disagreement as to how strongly and how universally held prior belief needs to be, and as to the meaning of prior probability itself.

The key equalities (1) or (2) suggest an alternative approach to the choice-of-procedure dilemma which may be appropriate in the absence of confidently held prior probability beliefs. They imply that the minimax procedure is an optimal approach against a malevolent "nature". Such a "nature" is one who could either divine the statistician's intended procedure and then pick an $F \in \mathscr{F}$ so as to maximize the statistician's risk, or one who could merely arrange always to choose F by the least favorable prior when (1) holds and by a nearly least favorable one when only (2) holds.

Because of this, some statisticians have suggested that minimaxity provides an objective criterion leading to a unique and satisfactory choice of decision procedure for each non-Bayesian decision problem. Wald may himself have felt this way in the 1940s but apparently abandoned this belief by the time of his sudden death in 1950. For further discussion, consult Savage [20, 21] and Brown [6]. One reason for the abandonment of this idea was the early realization that there are situations where an ε-minimax procedure has a risk function which would be preferred by all but the most pessimistic statistician over that of the minimax procedure. See, for example, Hodges & Lehmann [9], Robbins [19], and Wolfowitz [27].

The discovery by Stein [12, 22] that **shrinkage estimators** may dominate the usual minimax estimator of a multivariate mean is definitive evidence that, at best, the minimax principle does not provide a *unique* objective solution in many common statistical settings.

Minimaxity has been – and continues to be – an important and stimulating concept for statistics, in spite of its failure to provide universally appropriate procedures. It has played an essential role as a motivation and as an organizing principle in many important statistical areas. It often also provides a benchmark against which other proposed procedures can be measured.

Areas where minimaxity plays a key role include asymptotic analysis (via the concept of local asymptotic minimaxity) as in LeCam [15] or Lehmann [17], **robust** estimation theory as in Huber [10, 11], robust Bayesian methodology (via the notion of Γ-minimaxity) as in Berger [1], **optimal design** of experiments as in Kiefer & Wolfowitz [14] and Kiefer [13], and nonparametric function estimation as in Donoho & Liu [7] and Donoho et al. [8]. See Brown [6] as well as other references cited there for more details.

References

[1] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed. Springer-Verlag, New York.

[2] Borel, E. (1921). La théorie de jeu et les équations intégrales à noyan symmetrique, *Comptes Rendus de l'Académie des Sciences* **173**, 1304–1308.

[3] Borel, E. (1924). Sur les jeux ou interviennent l'hasard et l'habileté des jouers, in *Elements de la Théorie des Probabilités*, 3ed Ed. Librairié Scientifique, Paris, pp. 204–221.

[4] Brown, L.D. (1977). Closure theorems for sequential-design processes, in *Statistical Decision Theory and Related Topics*, Vol. 2, S.S. Gupta & D.S. Moore, eds. Academic Press, New York, pp. 57–91.

[5] Brown, L.D. (1980). A necessary condition for admissibility, *Annals of Statistics* **8**, 540–545.

[6] Brown, L.D. (1993). Minimaxity, more or less, in *Statistical Decision Theory and Related Topics*, Vol. 5, S.S. Gupta & J.O. Berger, eds. Springer-Verlag, New York, pp. 1–18.

[7] Donoho, D.L. & Liu, R.C. (1991). Geometrizing rates of convergence, III, *Annals of Statistics* **19**, 668–701.

[8] Donoho, D.L., Liu, R.C. & MacGibbon, B. (1990). Minimax rates for hyperrectangles and implications, *Annals of Statistics* **18**, 1416–1437.

[9] Hodges, J.L., Jr & Lehmann, E.L. (1950). Some problems in minimax point estimation, *Annals of Mathematical Statistics* **21**, 182–197.

[10] Huber, P.J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**, 73–101.

[11] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[12] James, W. & Stein, C. (1961). Estimation with quadratic loss, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, J. Neyman, ed. University of California Press, Berkeley, pp. 311–319.

[13] Kiefer, J.C. (1974). General equivalence theory for optimum designs (approximate theory), *Annals of Statistics* **2**, 849–879.

[14] Kiefer, J.C. & Wolfowitz, J. (1960). The equivalence of two extremum problems, *Canadian Journal of Mathematics* **12**, 363–366.

[15] LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, in *University of California Publication in Statistics*, Vol. 1, no. 11, University of California Press, Berkeley, pp. 277–330.

[16] LeCam, L. (1955). An extension of Wald's theory of statistical decision functions, *Annals of Mathematical Statistics* **26**, 69–81.

[17] Lehmann, E.L. (1997). *Theory of Point Estimation*. 2nd Ed. Wiley, New York.

[18] Luce, R.D. & Raiffa, H. (1957). *Games and Decisions, Introduction and Critical Survey*. Wiley, New York (republished in a Dover edition, 1989).

[19] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 131–148.

[20] Savage, L.J. (1954). *The Foundations of Statistics*. Wiley, New York (second revised edition, published by Dover, 1972).

[21] Savage, L.J. (1961). The foundations of statistics reconsidered, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 575–586.

[22] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, J. Neyman, ed. University of California Press, Berkeley, pp. 197–206.

[23] von Neumann, J. (1928). Zur theorie der gesellschaftsspielen, *Mathematische Annalen* **100**, 295–320.

[24] von Neumann, J. & Morgenstern, C. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

[25] Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses, *Annals of Mathematical Statistics* **10**, 299–326.

[26] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

[27] Wolfowitz, J. (1951). On ε-complete classes of decision functions, *Annals of Mathematical Statistics* **22**, 461–465.

(*See also* **Bayesian Methods**; **Foundations of Probability**; **Subjective Probability**)

L.D. BROWN

Minimum Therapeutically Effective Dose

Moore [10] indicates that the difference between a drug and a poison is the dose. Hence, it is extremely important to identify the dose range of a drug product that provides effective and safe treatment of a certain disease. The lower limit of this dose range is usually referred to as the minimum therapeutically effective dose (MTED). As a result, the MTED is defined as the lowest dose level of a drug product yielding a therapeutically significant response in average efficacy that is also statistically significantly superior to the response provided by the placebo [6, 11, 13]. According to this definition, the MTED must produce a response with a magnitude of clinical superiority over the placebo, since a small but statistically significant response resulting from either large sample sizes or small variability can be of no real therapeutical meaning (*see* **Clinical Significance Versus Statistical Significance**). Furthermore, the MTED must yield a statistically significant clinical response. This is because a large response produced at a certain dose level, if it is statistically insignificant from the placebo response, fails to establish the scientific evidence of effectiveness for that dose level. Similarly, the maximum tolerable dose (MTD) is the highest possible, but still tolerable, dose level with respect to a prespecified clinical limiting toxicity [9, 15]. The maximum effective dose (MED) is the highest dose level beyond which no additional therapeutically meaningful improvement in average efficacy can be achieved. The therapeutic range (window) is then defined as the range of the dose levels from the MTED, denoted by d_{L}, to the minimum of MTD and MED, denoted by d_{U}. If d_{L} is much smaller than d_{U}, then the corresponding drug product is said to have a wide therapeutic window. However, if the MTD is very close to, or even smaller than, the MTED, then the drug product is of no practical therapeutical use. The definition discussed above focuses on the average efficacy of a patient population. Fillon [6] gives a definition of the MTED for a particular patient as the lowest dose level that provides a prespecified therapeutical effectiveness above a predetermined percentage of patients. We refer to the first traditional definition as the population MTED (PMTED) and the second definition as the individual MTED (IMTED).

Study Design

The MTED is usually estimated from the data of primary efficacy endpoints from dose-ranging or **dose-response** clinical trials conducted during phase II clinical development of a drug product. These dose–response studies are usually randomized, double-blind, parallel-group designs with inclusion of a concurrent placebo group. Occasionally, a **crossover design** such as Williams' design [4] is employed. Many clinicians, however, find that a variety of titration designs, either with or without a concurrent parallel placebo group for dose-ranging studies, are useful because they mimic clinical practices in the real world. For details on designs for dose-response trials, see ICH E4 guideline [8].

Inclusion of a placebo is essential for estimation of the MTED, because a dose of any drug product cannot establish its therapeutical effectiveness without comparison with a placebo. In addition to choosing an appropriate statistical design, the selection of dose levels, the number of dose levels, and sample sizes for each dose group are crucial for estimation of the MTED. These issues are not only related to each other, but are also very difficult to deal with. The dose range should be chosen as wide as possible within the safety limit so that the dose–response relationship can be adequately characterized. For the same reason, the number of dose levels, including the active agent and the placebo, should be at least three. It is also preferable to select a dose level whose response is not expected to be statistically different from the placebo response, so that the MTED can be estimated more precisely. **Sample size determination** includes estimation of the total sample size and its distribution across different dose groups. As indicated by Ruberg [13], sample size should be determined on the basis of statistical tests for the hypotheses of interest for primary efficacy clinical endpoints.

Statistical Analysis

Basically, there are two commonly used approaches for estimation of the MTED. One is the method of **hypothesis testing** based on the **analysis of variance** (ANOVA). It consists of step-down, step-up, and single-step procedures. The step-down procedures include Dunnett's step-down procedure, Williams' test for ordered alternatives [17], and the step-down

linear contrasts [1, 16]. The **Bonferroni** procedure proposed by Hochberg [7], and its later refinement by Dunnett & Tamhane [5], in conjunction with the application of Helmert contrasts, are typical step-up procedures for estimation of the MTED. Ruberg [12] introduced the application of the step contrasts and basin steps as single-step procedures to estimating the MTED. The MTED estimated by the ANOVA will be one of the dose levels evaluated in the trial. An interval estimation for the MTED has not yet been developed for the ANOVA approach. Although a functional form such as the four-parameter logistic function is generally required, both point and interval estimates can be obtained by the model-based approach [14]. In addition, one can incorporate more directly the information of the therapeutically effective response into the assumed model for estimation of the MTED. Nonparametric methods for identification of MTED are also proposed [2, 3]. Ruberg [13, 14] gives a comprehensive review of the current state-of-the-art in design and estimation of the MTED.

Discussion

The therapeutically meaningful response can be expressed either as the original actual response at a particular dose level or as an additional clinically meaningful improvement over the placebo response. All approaches described above assume the therapeutically significant response a priori as a known constant. This assumption relies on the external validity of the past history, and previous experience of the disease under study, for the assumed clinically meaningful responses. This assumption is reasonable, provided that the placebo response has been well established by adequate, well-controlled studies and the medical condition is well understood for evaluation of the drug product. For internal validity, however, the therapeutically significant response should be determined by the response provided by a current placebo control group.

References

[1] Capizzi, T., Survill, T.T. & Heyse, J.F. (1992). An empirical and simulated comparison of some tests for detecting progress of response with increasing doses of a compound, *Biometrical Journal* **34**, 275–289.

[2] Chen, Y.I. (1999a). Nonparametric identification of the minimum effective dose, *Biometrics* **55**, 1236–1240.

[3] Chen, Y.I. (1999b). Rank-based tests for dose-finding in nonmonotonic dose-response settings, *Biometrics* **55**, 1258–1262.

[4] Chow, S.C. & Liu, J.P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker, New York.

[5] Dunnett, C.W. & Tamhane, A.C. (1992). A step-up multiple test procedure, *Journal of the American Statistical Association* **87**, 162–170.

[6] Fillon, T.G. (1995). Estimating the minimum therapeutically effective dose of a compound via regression modelling and percentile estimation, *Statistics in Medicine* **14**, 925–932.

[7] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

[8] ICH E4 Guideline (1994). Dose-Response Information to Support Registration, International Conference on Harmonisation.

[9] Korn, E.L., Midthune, D., Chen, T.T., Rubinstein, L.V., Christian, M.C. & Simon, R. (1994). A comparison of two phase I trial designs, *Statistics in Medicine* **13**, 1799–1806.

[10] Moore, T.J. (1995). *Deadly Medicine*, Simon & Schuster, New York.

[11] Rodda, B.E., Tsianco, M.C., Bolognese, J.A. & Kersten, M.K. (1988). Clinical development, in *Biopharmaceutical Statistics for Drug Development*, K.E. Peace, ed. Marcel Dekker, New York.

[12] Ruberg, S.J. (1989). Contrasts for identifying the minimum effective dose, *Journal of the American Statistical Association* **84**, 816–822.

[13] Ruberg, S.J. (1995). Dose response studies. I. Some design considerations, *Journal of Biopharmaceutical Statistics* **5**, 1–14.

[14] Ruberg, S.J. (1995). Dose response studies. II. Analysis and interpretation, *Journal of Biopharmaceutical Statistics* **5**, 15–42.

[15] Storer, B.E. (1989). Design and analysis of phase I clinical trials, *Biometrics* **45**, 925–937.

[16] Tukey, J.W., Ciminera, J.L. & Heyes, J.F. (1985). Testing the statistical certainty of a response to increasing doses of a drug, *Biometrics* **41**, 295–301.

[17] Williams, D.A. (1972). The comparison of several dose levels with a zero dose control, *Biometrics* **28**, 519–531.

(*See also* **Phase II Trials**)

JEN-PEI LIU & SHEIN-CHUNG CHOW

Minimum Variance Unbiased (MVU) Estimator

Methods for determining the minimum possible variance of an **unbiased** estimator constitute a fundamental topic in mathematical statistics. An estimator T^* is minimum variance unbiased (MVU) for $g(\theta)$ when

$$\mathrm{E}_\theta(T^*) = g(\theta)$$

and

$$\mathrm{var}_\theta(T^*) \leq \mathrm{var}_\theta(T), \quad \text{for all } T \text{ such that } \mathrm{E}_\theta(T) = g(\theta). \tag{1}$$

Two primary results that relate to MVU estimators are the **Cramér–Rao Inequality** [1, 6, 13], and the **Rao–Blackwell theorem** [3, 17].

Information Inequality

Assume that a statistic $T_n = t(X_1, X_2, \ldots, X_n)$ is an estimator of $g(\theta)$, where $X_1, X_2, \ldots$ is a sequence of independent, identically distributed **random variables** with probability density function $f(x;\theta), \theta \in \Theta$. Under certain regularity conditions (*see* **Cramér–Rao Inequality**), if $\mathrm{E}(T_n) = g(\theta)$, then

$$\mathrm{var}_\theta(T_n) \geq \frac{[g'(\theta)]^2}{n\ \mathrm{E}_\theta\left\{\left[\dfrac{\partial}{\partial\theta}\ln f(X;\theta)\right]^2\right\}}. \tag{2}$$

This result is known as the Cramér–Rao inequality (or the **information** inequality). The right-hand side (RHS) of (1) is called the Cramér–Rao (or information) lower bound.

Clearly, whenever the variance of an unbiased estimator is equal to the RHS, that estimator is MVU. The lower bound is attained if and only if

$$\frac{\partial}{\partial\theta}\ln f(X;\theta) = K(\theta, n)[T_n - g(\theta)]. \tag{3}$$

It follows from (3) that if T is an unbiased estimator of a function $g(\theta)$, and $\mathrm{var}_\theta(T_n)$ attains the lower bound in the RHS of (1), then $f(x;\theta)$ belongs to an **exponential class**:

$$f(x;\theta) = \exp[A(\theta)B(x) + C(x) + D(\theta)] \tag{4}$$

for appropriate functions A, B, C, and D. In this case, the estimator T_n is given by $\sum_{i=1}^n B(x_i)$. We see below that (4) is also a sufficient condition for MVU estimators.

In general, the Cramér–Rao inequality is not sharp, so that other methods are necessary to find lower bounds for MVU estimators. When the lower bound cannot be obtained, "better" (i.e. greater) lower bounds than (2) can be obtained. For example, Bhattacharya [2] obtains better lower bounds using higher-order derivatives of the score function (*see* **Likelihood**). However, improvements in the Cramér–Rao lower bound are only of order $O(1/n^2)$ [8] (*see* **Orders of Magnitude**). Using other methods, Kiefer [9] and Chapman & Robbins [5] derive minimum variance bounds that are better than (2) and avoid regularity conditions.

Rao–Blackwell Theorem

A statistic $S = s(X_1, X_2, \ldots, X_n)$ is **sufficient** for θ if the **conditional probability** distribution of X given S does not depend on θ for any s. If $S_1, S_2, \ldots, S_k$ are sufficient for θ and T^* is unbiased for $g(\theta)$, then the following hold:

1. Let $T^* = \mathrm{E}_\theta(T|S_1, S_2, \ldots, S_k)$. Then $\mathrm{E}(T^*) = g(\theta)$.
2. $\mathrm{var}_\theta(T^*) \leq \mathrm{var}_\theta(T)$ for all θ.
3. $\mathrm{var}_\theta(T^*) < \mathrm{var}_\theta(T)$ for some $\theta \in \Theta$ unless $T = T^*$ with probability 1.

This result is due to Rao [14] and Blackwell [3]. It supplies a method for improving the variance of any unbiased estimator of $g(\theta)$ that is not a function of a sufficient statistic.

If the family $f(x;\theta)$ is "complete", then further results are possible. A family of densities $f(x;\theta)$ is said to be complete when $\mathrm{E}_\theta[z(T)] = 0$ for all $\theta \in \Theta$ implies $\Pr_\theta[z(T) = 0] = 1$ for all $\theta \in \Theta$. If S is a complete, sufficient statistic and $\mathrm{E}_\theta[T(S)] = g(\theta)$, then $T(S)$ is MVU for $g(\theta)$ [11].

When $f(x;\theta)$ belongs to an exponential family and has the form in (4), then $\sum_{i=1}^n B(x_i)$ is a complete sufficient statistic. Thus, for exponential families, if $T_n = \sum_{i=1}^n B(x_i)$ and $\mathrm{E}_\theta(T_n) = g(\theta)$, then T_n

is MVU. In this case, however, T_n satisfies (3), so that T_n achieves the Cramér–Rao lower bound.

Large-Sample Results

For any given n, a function of the sufficient statistic will have minimum variance for estimating its expected value. For large samples (*see* **Large-sample Theory**), *any* function of the sufficient statistic will estimate its expected value at the Cramér–Rao lower bound [8]. Finally, under slightly stronger assumptions (which hold for exponential families), all **maximum likelihood** estimates asymptotically approach the lower bound as $n \to \infty$ (Serfling [17], after Cramér [7]).

References

[1] Aitken, A.C. & Silverstone, H. (1942). On the estimation of statistical parameters, *Proceedings of the Royal Society Edinburgh, Series A* **61**, 186–194.

[2] Bhattacharya, A. (1946). On some analogues of the amount of information and their uses in statistical estimation, *Sankhyā* **8**, 1–14, 201–218, 315–328.

[3] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation, *Annals of Mathematical Statistics* **18**, 105–110.

[4] Casella, G. & Berger, R.L. (2002). *Statistical Inference*, 2nd Ed. Duxbury Press, North Scituate, MA.

[5] Chapman, D.G. & Robbins, H. (1951). Minimum variance estimation without regularity assumptions, *Annals of Mathematical Statistics* **22**, 581–586.

[6] Cramér, H. (1946a). A contribution to the theory of statistical estimation, *Skandinavisk Aktuarietidskrift* **29**, 85–94.

[7] Cramér, H. (1946b). *Mathematical Methods in Statistics*. Princeton University Press, Princeton.

[8] Stuart, A. & Ord, J.K. (1994). *Kendall's advanced theory of statistics*. Volume 1. Distribution theory (Sixth edition), Edward Arnold Publishers Ltd., London; Baltimore.

[9] Kiefer, J. (1952). On minimum variance estimators, *Annals of Mathematical Statistics* **23**, 627–629.

[10] Lehmann, E.L. & Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag Inc, Berlin; New York.

[11] Lehmann, E.L. & Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation, *Sankhyā, Series A* **10**, 305–340.

[12] Mood, A.M., Graybill, F.A. & Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd Ed. McGraw-Hill, New York.

[13] Rao, C.R. (1945). Information and the accuracy attainable in the estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society* **37**, 81–91.

[14] Rao, C.R. (1949). Sufficient statistics and minimum variance unbiased estimates, *Proceedings of the Cambridge Philosophical Society* **45**, 213–218.

[15] Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd Ed. Wiley, New York.

[16] Rohatgi, V.K. & Saleh, A.K. (2000). *An Introduction to Probability and Statistics*. John Wiley & Sons.

[17] Serfling, R.J. (1980). *Approximation Theorems in Mathematical Statistics*. Wiley, New York.

Bibliography

These topics are covered to some extent in any book on mathematical statistics. Mood et al. [12] and Casella & Berger [4] both give thoughtful and cohesive discussions of MVU estimators at an undergraduate mathematical level. The presentation by Stuart & Ord [8] is more complete and more technical, but equally useful in its synthesis of the various theoretical components. Lehmann and Casella [10], Rao [15], and Rohatgi and Saleh [16] also give extensive and rigorous treatments of the topic.

(*See also* **Estimation**)

J. BETHEL

Mining Time Series Data

Time series data is ubiquitous; large volumes of time series data are routinely created in medical and biological domains, examples of which include **gene expression** data [1], electrocardiograms, electroencephalograms, (*see* **Clinical Signals**), gait analysis, **growth development** charts, and so on. Although statisticians have worked with time series for more than a century, many of their techniques hold little utility for researchers working with massive time series **databases** (for reasons discussed below).

The major tasks considered by the time series data mining community are as follows:

- *Indexing* (Query by Content): Given a query time series Q, and some **similarity/dissimilarity** measure D(Q, C), find the most similar time series in database *DB* [3, 6, 10, 15].
- ***Clustering***: Find natural groupings of the time series in database *DB* under some similarity/dissimilarity measure D(Q, C) [1, 5, 11, 13].
- ***Classification***: Given an unlabeled time series Q, assign it to one of two or more predefined classes [7, 13].
- ***Prediction*** (**Forecasting**): Given a time series Q containing n datapoints, predict the value at time $n + 1$.
- ***Association*** *Detection*: Given two or more time series, find relationships between them. Such relationships may or may not be **causal** and may or may not exist for the entire duration of the time series [4].
- *Summarization*: Given a time series Q containing n datapoints, where n is an extremely large number, create an (possibly graphic) approximation of Q, which retains its essential features but fits on a single page, computer screen, and so on [9, 18].
- *Anomaly detection* (interestingness detection): Given a time series Q, assumed to be normal, and a unannotated time series R, find all sections of R, which contain anomalies or "surprising/interesting/unexpected" occurrences [8, 12, 17].
- *Segmentation*: Given a time series Q containing n datapoints, construct a model $\overline{Q}$, from K piecewise segments ($K \ll$ n) such that $\overline{Q}$ closely approximates Q [13].

Note that indexing and clustering make *explicit* use of a distance measure, and many approaches to classification, prediction, association detection, summarization, and anomaly detection make *implicit* use of a distance measure. In this article, we will not consider distance measures in depth, instead we refer the reader to **Time Series Similarity Measures**.

It is interesting to note that with the exception of indexing, research into the tasks enumerated above predate not only the decade old interest in **data mining**, but in computing itself. What then are the essential differences between the classic versions and the data mining versions of these problems? The key difference is simply one of size and scalability; time series data miners routinely encounter datasets that are gigabytes in size. As a simple motivating example, consider hierarchical clustering. The technique has a long history and a well-documented utility. If, however, we wish to hierarchically cluster a mere million items, we would need to construct a matrix with 10^{12} cells, well beyond the abilities of the average computer for many years to come. A data mining approach to clustering time series, in contrast, must explicitly consider the scalability of the **algorithm** [11].

In addition to the large volume of data, it is often the case that each individual time series has a very high dimensionality [3]. Whereas classic algorithms assume a relatively low dimensionality (for example, a few measurements such as "height, weight, blood sugar etc."), time series data mining algorithms must be able to deal with dimensionalities in the hundreds and thousands. The problems created by high dimensional data are more than mere computation time considerations; the very meanings of normally intuitive terms, such as "similar to" and "cluster forming" become unclear in high-dimensional space. The reason is that, as dimensionality increases, all objects become essentially equidistant to each other, and thus classification and clustering lose their meaning. This surprising result is known as the "curse of dimensionality" and has been the subject of extensive research [2]. The key insight that allows meaningful time series data mining is that although the actual dimensionality may be high, the *intrinsic* dimensionality is typically much lower. For this reason, virtually all time series data mining algorithms avoid operating on the original "raw" data; instead, they consider some higher-level representation or abstraction of the data.

Time Series Representations

As noted above, time series datasets are typically very large; for example, just eight hours of electroencephalogram data can require in excess of a gigabyte of storage. This is a problem because for almost all data mining tasks, most of the execution time spent by algorithm is used simply to move data from disk into main memory. This is acknowledged as the major bottleneck in data mining, because many naïve algorithms require multiple accesses of the data. As a simple example, imagine we are attempting to do k-means clustering of a dataset that does not fit into main memory. In this case, every iteration of the algorithm will require that data in main memory be swapped. This will result in an algorithm that is thousands of times slower than the main memory case.

With this in mind, a generic framework for time series data mining has emerged. The basic idea can be summarized as follows.

It should be clear that the utility of this framework depends heavily on the quality of the approximation created in step 1. If the approximation is very faithful to the original data, then the solution obtained in main memory is likely to be the same or very close to the solution we would have obtained on the original data. The handful of disk accesses made in step 2 to confirm or slightly modify the solution will be inconsequential compared to the number of disk accesses required if we had worked on the original data. With this in mind, there has been a huge interest in approximate representation of time series. Figure 1 illustrates a hierarchy of every representation proposed in the literature.

To develop the reader's intuition about the various time series representations, we have illustrated four of the most popular representations in Figure 2 (*see* **Spectral Analysis**; **Wavelet Analysis**).

Given the plethora of different representations, it is natural to ask which is best. Recall that the more faithful the approximation, the less clarification disk accesses we will need to make in step 3 of Table 1. In the example shown in Figure 2, the discrete Fourier approach seems to model the original data the best; however, it is easy to imagine other time series where another approach might work better. There have been many attempts to answer the question of which is the best representation, with proponents advocating their favorite technique [3, 6, 15, 16]. The literature abounds with mutually contradictory statements, such as "Several wavelets outperform the ... DFT" [15], "DFT-based and DWT-based techniques

Table 1 A generic time series data mining approach

1)	Create an approximation of the data, which will fit in main memory, yet retains the essential features of interest.
2)	Approximately solve the problem at hand in main memory.
3)	Make (hopefully very few) accesses to the original data on disk to confirm the solution obtained in step 2, or to modify the solution so it agrees with the solution we would have obtained on the original data.

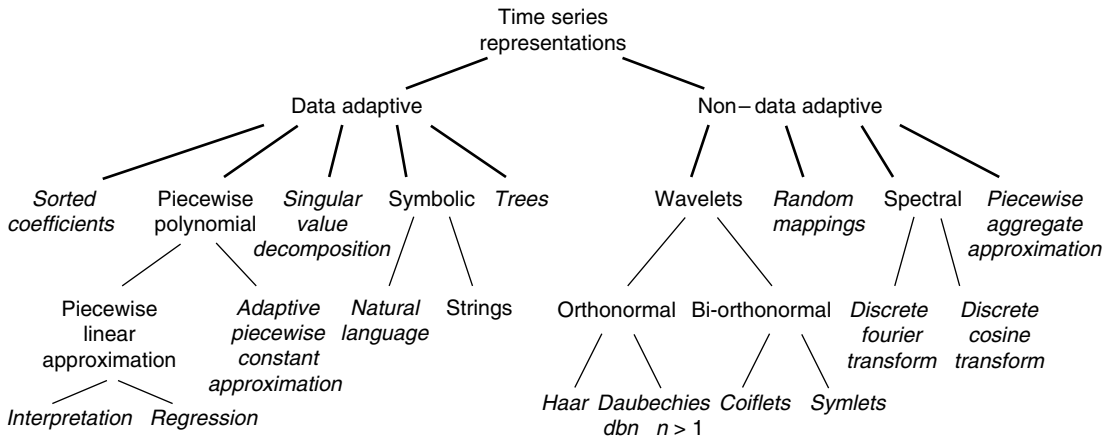

Figure 1 A hierarchy of time series representations

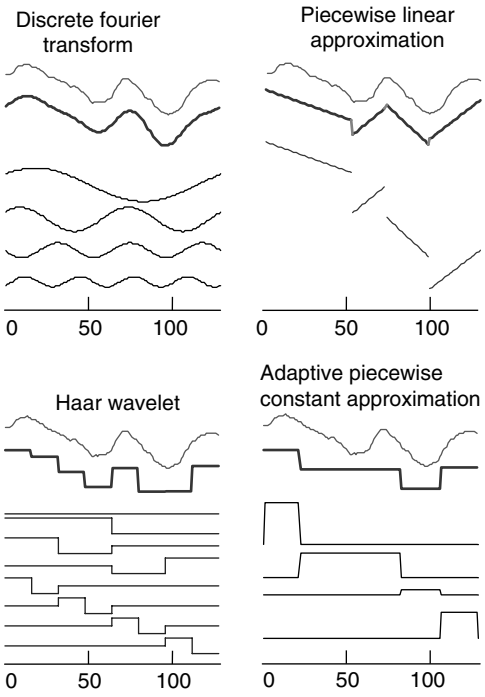

Figure 2 Four popular representations of time series. For each graphic, we see a raw time series of length 128 datapoints. Below it we see an approximation using 1/8 of the original space. In each case, the representation can be seen as a linear combination of basis functions. For example, the discrete Fourier representation can be seen as a linear combination of the four sine/cosine waves shown at the bottom of the graphic

yield comparable results" [19], "Haar wavelets perform. . . better that DFT" [10]. However an extensive empirical comparison on 50 diverse datasets suggests that while some datasets favor a particular approach, overall there is little difference between the various approaches in terms of their ability to approximate the data [14]. There are, however, other important differences in the usability of each approach [3]. We will consider some representative examples of strengths and weaknesses below.

The wavelet transform is often touted as an ideal representation for time series data mining because the first few wavelet coefficients contain information about the overall shape of the sequence, while the higher-order coefficients contain information about localized trends [15, 17]. This multiresolution property can be exploited by some algorithms, and contrasts with the Fourier representation in which every coefficient represents a contribution to the global trend [6, 16]. However, wavelets do have several drawbacks as a data mining representation. They are only defined for data whose length is an integer power of two. In contrast, the piecewise constant approximation suggested by [20], has exactly the same fidelity of resolution as the Haar wavelet, but is defined for arbitrary length time series. In addition, it has several other useful properties, such as the ability to support several different distance measures [20], and the ability to be calculated in an incremental fashion as the data arrives [3]. Choosing the right representation for the task is the key step in any time series data mining endeavor. The points above only serve as a sample of the issues that must be addressed.

Readings

The field of time series data mining is relatively new and ever changing. Because of the length of journal publication delays, the most interesting and useful work tends to appear in top-tier conference proceedings. Interested readers are urged to consult the latest proceedings of the major conferences in the field. These include the ACM Knowledge Discovery in Data and Data Mining, IEEE International Conference on Data Mining, and the IEEE International Conference on Data Engineering.

References

[1] Aach, J. & Church, G. (2001). Aligning gene expression time series with time warping algorithms, *Bioinformatics* **17**, 495–508.

[2] Aggarwal, C., Hinneburg, A. & Keim, D.A. (2001). On the surprising behavior of distance metrics in high dimensional space, in *Proceedings of the 8th International Conference on Database Theory*. London, Jan 4–6, pp. 420–434.

[3] Chakrabarti, K., Keogh, E., Pazzani, M. & Mehrotra, S. (2002). Locally adaptive dimensionality reduction for indexing large time series databases, *ACM Transactions on Database Systems* **27**(2), 188–228.

[4] Das, G., Lin, K., Mannila, H., Renganathan, G. & Smyth, P. (1998). Rule discovery from time series, in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. New York, Aug 27–31, pp. 16–22.

[5] Debregeas, A. & Hebrail, G. (1998). Interactive interpretation of kohonen maps applied to curves, in *Proceedings of the 4th International Conference of Knowledge*

Discovery and Data Mining. New York, Aug 27–31, pp. 179–183.

[6] Faloutsos, C., Ranganathan, M. & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases, in *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Minneapolis, May 25–27, pp. 419–429.

[7] Geurts, P. (2001). Pattern extraction for time series classification, in *Proceedings of Principles of Data Mining and Knowledge Discovery, 5th European Conference*. Freiburg, Sept 3–5, pp. 115–127.

[8] Guralnik, V. & Srivastava, J. (1999). Event detection from time series data, in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, Aug 15–18, pp. 33–42.

[9] Indyk, P., Koudas, N. & Muthukrishnan, S. (2000). Identifying representative trends in massive time series data sets using sketches, in *Proceedings of the 26th International Conference on Very Large Data Bases*. Cairo, Sept 10–14, pp. 363–372.

[10] Kahveci, T. & Singh, A. (2001). Variable length queries for time series data, in *Proceedings of the 17th International Conference on Data Engineering*. Heidelberg, Apr 2–6, pp. 273–282.

[11] Kalpakis, K., Gada, D. & Puttagunta, V. (2001). Distance measures for effective clustering of ARIMA time-series, in *Proceedings of the IEEE International Conference on Data Mining*. San Jose, Nov 29–Dec 2, pp. 273–280.

[12] Keogh, E., Lonardi, S. & Chiu, W. (2002). Finding surprising patterns in a time series database in linear time and space, in *the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, July 23–26, pp. 550–556.

[13] Keogh, E. & Pazzani, M. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback, in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. New York, Aug 27–31, pp. 239–241.

[14] Keogh, E. & Kasetty, S. (2002). On the need for time series data mining benchmarks: a survey and empirical demonstration, in *the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, July 23–26, pp. 102–111.

[15] Popivanov, I. & Miller, R.J. (2002). Similarity search over time series data using wavelets, in *Proceedings of the 18th International Conference on Data Engineering*. San Jose, Feb 26–Mar 1, pp. 212–221.

[16] Rafiei, D. & Mendelzon, A.O. (1998). Efficient retrieval of similar time sequences using DFT, in *Proceedings of the 5th International Conference on Foundations of Data Organization and Algorithms*. Kobe, Nov 12–13.

[17] Shahabi, C., Tian, X. & Zhao, W. (2000). TSA-tree: a wavelet based approach to improve the efficiency of multi-level surprise and trend queries, in *Proceedings of the 12th International Conference on Scientific and Statistical Database Management*. Berlin, July 26–28, pp. 55–68.

[18] Wijk, J.J. van & van Selow, E. (1999). Cluster and calendar-based visualization of time series data, *Proceedings of 1999 IEEE Symposium on Information Visualization*, Oct 25–26, IEEE Computer Society, San Francisco, CA, pp. 4–9.

[19] Wu, Y., Agrawal, D. & El Abbadi, A. (2000). A comparison of DFT and DWT-based similarity search in time-series databases, in *Proceedings of the 9th ACM CIKM International Conference on Information and Knowledge Management*. McLean, Nov 6–11, pp. 488–495.

[20] Yi, B. & Faloutsos, C. (2000). Fast time sequence indexing for arbitrary lp norms, in *Proceedings of the 26th International Conference on Very Large Databases*. Cairo, Sept 10–14, pp. 385–394.

EAMONN J. KEOGH

Misclassification Error

It was recognized early [3] that misclassification of categorical variables induces problems of analysis and interpretation. In epidemiology there has been continuing interest in assessing effects of misclassification on exposure–disease **associations**. More recent attention has been paid to methodology for estimating the misclassification structure and adjusting for resulting **biases**. This involves gathering auxiliary data through validation samples (*see* **Validation Study**) and repeated measurements. Although there are immediate parallels between the rationale for handling misclassification and the discussion on measurement errors in continuous exposure variables, the two topics have different historical paths and the statistical techniques differ in technical detail. **Measurement Error in Epidemiologic Studies** deals with the case of continuous covariates. Early reviews on the effects of misclassification include a bibliography by Dalenius [8] and a paper by Chen [4]. Kuha & Skinner [26] offer a more recent account. Here we describe effects caused by misclassification and present some of the methodology for adjustment.

Effects of Misclassification

Univariate Analyses

Let A^* denote the classification variable subject to error and A the true variable that the classification variable is intended to measure. We refer to A^* as a surrogate for A. For each unit (individual) the outcome of A, and A^*, falls into one of m mutually exclusive categories. Independence between units is assumed. We write the misclassification probabilities

$$\Pr(A^* = j|A = k) = \theta_{jk}, \quad j, k = 1, \ldots m.$$

The parameters θ_{jk} governing the misclassification structure may be collected into an $m \times m$ misclassification matrix $\Theta = [\theta_{jk}]$ with nonnegative elements and columns that sum to one. For a **binary** response, where $m = 2$ and the categories indicate the presence $(A = 2)$ or absence $(A = 1)$ of disease, the misclassification matrix involves only two parameters

$$\Theta = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} = \begin{pmatrix} \beta & 1-\alpha \\ 1-\beta & \alpha \end{pmatrix}. \quad (1)$$

The parameter α in (1) is called the **sensitivity** of the measuring instrument and β the **specificity**.

The effect of using the surrogate classification A^* may be summarized by

$$\pi_{A^*} = \Theta\pi_A,$$

where $\pi_{A*} = (\pi_{A*}(1), \ldots, \pi_{A*}(m))'$ and $\pi_A = (\pi_A(1), \ldots, \pi_A(m))'$ are the population proportions in the categories of the surrogate variable and the true variable, respectively. Sample proportions of A^* are thus biased estimates of π_A. The nature of this bias is most easily described in the binary case, where

$$\pi_{A*}(2) = (1-\beta)\pi_A(1) + \alpha\pi_A(2) \quad (2)$$

(for example [6]). Even when the misclassification matrix differs from the identity matrix it is clear from (2) that the two errors are mutually compensating if $(1-\beta)\pi_A(1) = (1-\alpha)\pi_A(2)$. The degree of compensation depends on the true proportions $\pi_A(1)$ and $\pi_A(2)$. An instrument with given misclassification matrix can thus induce different degrees of bias in different populations.

Bivariate Analyses

2 × 2 Tables. In a **two-by-two table**, let A define the presence or absence of disease and B two exposure groups, e.g. smokers and nonsmokers. We first consider the case where the response variable is subject to misclassification, i.e. A is measured by the surrogate A^*. Let $\pi_{A|B}(j|l)$ denote the proportion of units in the population for which $A = j$ in exposure group $B = l$, and let $\pi_{A*|B}(j|l)$ denote the corresponding proportion for A^*.

When focus is on the difference in the response proportions between the two exposure groups, then an **unbiased** estimator of the surrogate difference $\pi_{A*|B}(2|2) - \pi_{A*|B}(2|1)$ will in general be biased for the true difference $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$. The argument simplifies if both exposure groups have the same sensitivity α and specificity β. In this case the misclassification mechanism for A is said to be **nondifferential** with respect to B. It follows from (2) that under nondifferential misclassification

$$\pi_{A*|B}(2|2) - \pi_{A*|B}(2|1) = (\alpha + \beta - 1)[\pi_{A|B}(2|2) - \pi_{A|B}(2|1)]. \quad (3)$$

It is reasonable to expect each of the misclassification probabilities $1-\alpha$ and $1-\beta$ to be less than 0.5, in which case the factor $(\alpha+\beta-1)$ in (3) takes values between 0 and 1. The difference measured by the surrogate A^* is thus always smaller than the true difference based on A. The effect of nondifferential misclassification is to *attenuate*, i.e. "to make seem smaller", the difference in subclass proportions. This was noted by Rubin et al. [30] as early as 1956. Nondifferential misclassification similarly attenuates the ratio $\pi_{A|B}(2|2)/\pi_{A|B}(2|1)$ toward the null value of one (see, for example, [7]) (*see* **Bias Toward the Null**).

If, instead, the response variable A is correctly classified while the exposure B is misclassified as B^*, and if the misclassification of B is nondifferential with respect to A, then

$$\begin{aligned}\pi_{A|B*}(2|2)-\pi_{A|B*}(2|1) &= \frac{(\alpha_B+\beta_B-1)\pi_B(2)\pi_B(1)}{\pi_{B*}(2)\pi_{B*}(1)} \\ &\quad\times[\pi_{A|B}(2|2)-\pi_{A|B}(2|1)], \qquad (4)\end{aligned}$$

where $\pi_B=(\pi_B(1),\pi_B(2))'$ and $\pi_{B*}=(\pi_{B*}(1),\pi_{B*}(2))'$ are the population proportions of B and B^* respectively, and α_B and β_B are the sensitivity and specificity of the classification of B. The factor multiplying $\pi_{A|B}(2|2)-\pi_{A|B}(2|1)$ in (4) is again between 0 and 1, when $0<\alpha_B+\beta_B-1\le 1$, so that the effect is a similar type of attenuation as described above.

If both the response A and the exposure B are subject to misclassification, and if the surrogate pair (A^*, B^*) is jointly determined by the pair (A, B) through the misclassification probabilities $\Pr(A^*=j^*, B^*=k^*|A=j, B=k)$, then misclassification of A and B is said to be *independent* if

$$\begin{aligned}\Pr(A^*=j^*, B^*=k^*|A=j, B=k) &= \Pr(A^*=j^*|A=j, B=k) \\ &\quad\times\Pr(B^*=k^*|A=j, B=k),\end{aligned}$$

and it is *nondifferential* if

$$\Pr(A^*=j^*|A=j, B=k)=\Pr(A^*=j^*|A=j)$$

and

$$\Pr(B^*=k^*|A=j, B=k)=\Pr(B^*=k^*|B=k).$$

Under the condition of independent and nondifferential misclassification in A and B, Gullen et al. [20] show that an unbiased estimator of $\pi_{A*|B*}(2|2)-\pi_{A*|B*}(2|1)$ again attenuates the true difference. If on the other hand A, or B, or both, are subject to *differential* misclassification, then the bias inherent in $\pi_{A*|B*}(2|2)-\pi_{A*|B*}(2|1)$ can take any arbitrary form. A clear account of the possible effects of differential misclassification is presented by Goldberg [16] (*see* **Differential Error**).

Note that changes in the categorization of a misclassified variable may turn a nondifferential misclassification into a differential one. Wachholder et al. [34] discuss the situation in which A has three categories and is subject to nondifferential misclassification with respect to B. They show that combining two of the categories of A induces differential misclassification with respect to B. On a similar note, Flegal et al. [13] show that if a nondifferentially mismeasured continuous variable is dichotomized, this may induce differential error.

$2\times m$ Tables. When comparing proportions defined by a binary response A in three or more exposure subgroups ($m>2$) defined by B, the result in (3) holds when the response A is nondifferentially misclassified with respect to B. The ordering of the response proportions over exposure subgroups is thus preserved, but the differences are attenuated.

If, instead, the response A is correctly classified but the exposure B is subject to nondifferential misclassification, then (i) measures of association for response proportions between the two extreme exposure subgroups $B=1$ and $B=m$ are again attenuated, but (ii) associations between other exposure subgroups may be biased either away from or towards null [2, 15]. Exposure misclassification can even change the ordering of the response proportions in the intermediate subgroups and distort trends. However, if misclassification is confined to adjacent exposure subgroups, then attenuation occurs, but the ordering of the response proportions is retained [28].

Hypothesis Testing for Two-Way Tables. An important consequence of the attenuation results is that if there is no association between A and B, then there will be no association between the surrogates A^* and B^* under nondifferential misclassification in one variable [29] or under nondifferential and

independent misclassification in both variables [1]. The test of no association between A and B based on A^* and B^* will thus have the correct significance level (*see* **Hypothesis Testing**), but the **power** is in general reduced. Marshall et al. [28] show similar results for a test of no trend in subclass proportions for tables where the outcome A is binary and a polytomous subgroup variable B is nondifferentially misclassified.

Multivariate Analyses

The simplest **multivariate** case involves a $2 \times 2 \times 2$ table. Let A be a binary response variable and B and C binary variables defining subgroups of the population. In particular, B may refer to an exposure and C to a potential **confounder**.

If C is correctly classified, then we can consider the two-way tables between A and B separately for the two levels of C, and apply the bivariate results of the previous section. If only A is subject to nondifferential misclassification, then the difference in response proportions between the two exposure groups is attenuated by the factor given in (3). If, however, the exposure B is nondifferentially misclassified, we have from (4) that the degree of attenuation depends on the true proportions, in this case on $\pi_{B|C}(2|2)$ for $C = 2$ and on $\pi_{B|C}(2|1)$ for $C = 1$. If there is association between B and C, then the difference in response proportions between the two exposure groups may be attenuated to a different degree in the two categories of C. Nondifferential misclassification in the exposure may thus induce spurious heterogeneity (or mask true heterogeneity) in the exposure–disease association for different levels of the confounder [17].

If C is subject to nondifferential misclassification (with $0 < \alpha_C + \beta_C - 1 \leq 1$) with respect to A and B, which are both classified without error, then $\pi_{A|B,C*}(2|k, l)$ lies between the true proportions $\pi_{A|B,C}(2|k, l)$ and $\pi_{A|B}(2|k)$ for any $k, l = 1, 2$ [17, 26]. The proportions $\pi_{A|B}(2|k)$ are obtained by summing the data over the levels of C. This form of bias is known as residual confounding (for example [31]). It occurs because the analysis is restricted to the wrong levels of the confounder C, and thus the heterogeneity in the proportions due to confounding is not fully controlled for. The bias due to residual confounding may be either away from or toward the null value, and it can even induce a exposure–response association with the wrong sign. Both the size and power of a test of no association between A and B adjusted for C are thus incorrect when C is subject to nondifferential misclassification.

The above example of misclassification in a $2 \times 2 \times 2$ table may be extended in various ways. Some of the variables in a three-way table may be **polytomous** and there may be independent and nondifferential misclassification in more than one variable. In this case the effect of misclassification is a combination of attenuation and residual confounding [14]. Misclassification that is not both independent and nondifferential can produce any kind of biases (some examples are given by Greenland & Robins [19]).

For tables involving more than three variables it is not in general possible to give even qualitative statements about how misclassification distorts the analysis. One exception is a useful result due to Korn [24]: if there is independent and nondifferential misclassification in several variables in a multiway table, and if each of these misclassified variables appears in only one term of a **hierarchical loglinear model** specifying the association structure of the table, then this association structure is preserved under the misclassification. A test of **goodness of fit** for this loglinear model will have the correct significance level but reduced power. Korn [25] evaluates the loss of power due to misclassification when using a **likelihood ratio test** for comparing two nested models where the association structure is preserved.

Auxiliary Data on Misclassification

Adjustment for potential bias due to misclassification requires some information on the misclassification structure. If the structure is known, either through prior information or by assumption, then adjustment is straightforward. In general, however, the misclassification structure is unknown and estimated from a suitable set of *auxiliary data*, which are assumed to have the same misclassification parameters as the *primary data*. We briefly describe the two main types of auxiliary data: *validation samples* and *repeated measurements*.

Validation Samples

Both the true variable A, say, and the surrogate variable A^*, possibly together with other variables, are measured on each unit in a validation sample.

This raises two important questions: (i) How does one measure the true value A? (ii) How should the units in the validation sample be selected?

A measurement of A may be possible using an instrument referred to as the **gold standard**. It may be too expensive, however, to use the gold standard on all units in the primary study, or the gold standard may be available only for a subset of the units. The gold standard is a key concept in validation studies and the assumption that it measures A accurately, or with negligible error, is crucial (cf. [33], for a discussion of bias induced by using an erroneous or "alloyed" gold standard).

Ideally the validation sample should be a subsample of the primary data, obtained by a known randomized **double sampling** scheme. **Simple random sampling** from the primary data gives an *internal validation sample*, where the proportions in the categories of A in the validation sample are **unbiased** estimates of the corresponding population proportions. Both the misclassification probabilities $\Pr(A^*|A)$ and the **predictive values** $\Pr(A|A^*)$ in the population of interest can thus be consistently estimated from internal validation data. There may, however, be practical reasons that prevent such double sampling. If validation data from an earlier study are used, or if the gold standard is available only in a specific subpopulation, or if validation data are collected after the primary data are in hand, then it may be unreasonable to assume that the distribution over the categories of A are the same for units in the validation sample and in the primary study population. We then say that the validation data are *external*, and only the misclassification probabilities are assumed to be transportable between data sets.

Instead of using simple random sampling it may be useful to draw a prespecified proportion of the validation sample units within each category of A^*. This increases the efficiency in estimating the predictive values $\Pr(A|A^*)$, and is thus useful in internal validation studies [21].

Repeated Measurements

Even without a gold standard it may be possible to estimate misclassification parameters from repeated measurements of the surrogate. The measurements may be replicates using the same instrument or they may be obtained using different instruments. The distinction between internal and external data is relevant also for repeated measures, and which of these is in question depends on how the distributions over categories of A are related in the auxiliary and primary data sets.

For models based on repeated surrogate measures to be **identifiable**, a sufficient number of the measurements should be conditionally independent given the true value A. The required number depends on the model; some simple models are identifiable from just two measurements, while three measurements are sufficient for most models (see [35] and [27] for general identifiability conditions).

Adjusting for Effects of Misclassification

Misclassification parameters estimated from auxiliary data may be used to estimate parameters of interest adjusting for the biases induced by misclassification. Here we describe three classes of adjustment methods: simple matrix methods and model-based methods using either validation data or repeated measurements.

Matrix Methods

The most straightforward way to adjust for misclassification is via simple **back-calculation**. We refer to this as the *matrix method* of adjustment. The aim is to estimate the vector of cell proportions π_A for variable A. Here A may represent one variable or the cross-classification of several variables. Suppose that a primary data set and a validation data set are available, with n_p and n_v observations, respectively. The validation data provide an estimate, denoted by $\hat{\Theta}(A^*|A)$, for the matrix of misclassification probabilities $\theta_{jk} = \Pr(A^* = j|A = k)$. A matrix estimate of π_A is given by

$$\hat{\pi}_A^m = \{\hat{\Theta}(A^*|A)\}^{-1}\hat{\pi}_{A^*}, \tag{5}$$

with $\hat{\pi}_{A^*}$ the vector of observed cell proportions for the surrogate A^* in the primary data set. The analysis of interest is performed on the transformed table $\hat{\pi}_A^m$. An estimated variance matrix for $\hat{\pi}_A^m$, or any quantities derived from it, such as **odds ratios**, can be obtained using the **delta method** [18]. The simple matrix estimator (5) is well known in the epidemiologic literature. It is straightforward to compute, but has the drawbacks that the estimated probabilities are not constrained to lie between 0 and 1, and its small

sample properties may be poor due to the matrix inversion.

The estimator (5) can also be motivated as a **maximum likelihood** estimator (MLE) of π_A under a model where (i) π_A is unrestricted, (ii) the model for the misclassification probabilities is the one under which $\hat{\Theta}(A^*|A)$ was estimated (or taken as known), and (iii) the validation data are external. For most other models the MLE needs to be computed using iterative methods described in the next section. An important exception is a case where the validation data are internal and the model structure is such that there exists a one-to-one transformation from π_A and $\Theta(A^*|A)$ to π_{A^*} and $\Lambda(A|A^*)$, where $\Lambda(A|A^*)$ denotes the $m \times m$ matrix of predictive values $\Pr(A = i|A^* = j)$. The MLE of π_A is then also a closed-form matrix estimate, given by

$$\hat{\pi}_A^c = f_p \hat{\Lambda}(A|A^*)\hat{\pi}_{A^*} + (1 - f_p)\hat{\pi}_A^{(v)} \qquad (6)$$

where $f_p = n_p/(n_p + n_v)$, $\hat{\pi}_A^{(v)}$ is the vector of observed cell proportions of A in the validation data set, and $\hat{\Lambda}(A|A^*)$ is the matrix of predictive values estimated from the validation data. Estimates of this type were proposed by Tenenbein [32] for estimating cell probabilities of a single variable when both π_A and the misclassification probabilities are unrestricted. Tenenbein also gave formulas for the **variance** of $\hat{\pi}_A^c$.

In cases where (6) is the MLE of π_A, the external validation MLE $\hat{\pi}_A^m$ in (5) is also **consistent**, but not fully **efficient**. It may even have a higher variance than $\hat{\pi}_A^{(v)}$ alone. Surprisingly, the same is also true for the estimate $\tilde{\pi}_A^c = f_p\{\hat{\Theta}(A^*|A)\}^{-1}\hat{\pi}_{A^*} + (1 - f_p)\hat{\pi}_A^{(v)}$, which appears to be a compromise between (5) and (6). This estimate should not be used, because it is inconsistent when the validation data are external and less efficient than $\hat{\pi}_A^c$ when they are internal. Its variance may even *increase* with increasing n_p [26].

Matrix adjustments for misclassification are most useful in fairly simple problems with a small number of variables and few categories per variable. The estimates imply a model where the cell probabilities of the true variables are unrestricted and the model for the misclassification structure is either saturated (*see* **Generalized Linear Model**) or has a special form such as independent and nondifferential misclassification for all variables. In large problems this may lead to sparse tables and imprecise estimates for the many parameters. It is then desirable to consider more **parsimonious** models, especially when the focus is on **inference** about the association structure between the true variables. This can be done, at the expense of further model assumptions and some extra computing, by using model-based adjustment procedures described in the next section.

Modeling

Let A denote a set of variables subject to misclassification and A^* the corresponding set of surrogate variables, and let C be variables classified without error. It is also useful to define a sample indicator variable L which identifies the data set to which a unit belongs. L is binary when there is one primary sample and one validation sample, but other study designs can also be incorporated in this framework. The joint distribution of (A, A^*, C, L) may be specified through two submodels (cf. Espeland & Odoroff [12], who consider a slightly different set of models):

1. A model for the true variables (A, C, L). **Interactions** between L and (A, C) indicate differences in the distribution of the true variables between samples such as when a validation sample is external. The model of interest is the model for (A, C) in the primary sample.
2. A model for the misclassification probabilities, specified by interactions within A^* and between (A, C) and A^*. The model is saturated with respect to the true variables (A, C). Because the misclassification probabilities are assumed to be transportable between data sets, there should not be any interaction terms between L and A^*.

Both submodels are usually taken to be hierarchical loglinear models, but the joint model generated by them will not in general be loglinear [12].

The misclassification problem may be treated as one of incomplete **contingency tables**, collapsed over the margins corresponding to the unobserved variables. Suppose that there is one primary data set and one validation data set. The log **likelihood** function for the observed variables can be written as

$$L = \sum_{\text{prim}} n_{A^*C} \log \pi_{A^*,C} + \sum_{\text{val}} n_{AA^*C} \log \pi^{(v)}_{A,A^*,C} \qquad (7)$$

where n_{A^*C} are the observed cell counts for (A^*, C) in the primary data and $\pi_{A^*,C} = \sum_A \pi_{A,A^*,C}$ are the corresponding cell probabilities satisfying the specified model, and n_{AA^*C} and $\pi^{(v)}_{A,A^*,C}$ are the cell counts and probabilities for (A, A^*, C) in the validation data. The models may be fitted by maximizing (7) using iterative techniques, especially the **EM algorithm**. At the E step of the algorithm, observations from the observed (A^*, C) table in the primary data are allocated values of A to create a notionally complete (A, A^*, C) table. This is used at the M step, together with the validation sample, to fit the required joint model, and the process is iterated until convergence. Different versions of the EM algorithm for misclassification problems have been proposed by Chen et al. [5] and Espeland & Odoroff [12], who also consider the estimation of **standard errors** for the resulting estimates. The joint likelihood can also be maximized using other **algorithms** such as direct Newton–Raphson maximization [10, 11].

Methods Using Repeated Measurements

When the misclassification parameters are estimated from repeated measurements, the true values of the misclassified variables are latent variables (*see* **Path Analysis**) which are never observed. The misclassification probabilities and models of interest can be estimated from such data subject to appropriate identifiability assumptions. The analysis proceeds by specifying models for the true variables and misclassification as above and obtaining MLEs for their parameters. In some very simple cases, such as when estimating the proportions of a single binary misclassified variable, estimates are available in a closed form [22]. It is then also possible to use external repeated measurements to estimate the misclassification matrix in the matrix estimate (5) [9]. For most models, however, estimates have to be obtained iteratively, using general techniques of **latent class** modeling (see, for example, [23]). The calculations may again be conveniently carried out using the EM algorithm.

Conclusions

Misclassification induces bias in the estimates of quantities of interest obtained from observed surrogate variables. In some special cases it is possible to characterize qualitatively the nature of the bias, such as when measures of association are attenuated. In many situations, however, biases in any direction are possible. It is then desirable to collect auxiliary data such as validation data or repeated measurements from which the misclassification probabilities can be estimated, and to use these estimates to adjust analyses explicitly for the effects of misclassification. The most straightforward adjustment methods are simple matrix methods, which may, however, be unsatisfactory in larger models. Model-based adjustment methods may then be used for **estimation**.

References

[1] Assakul, K. & Proctor, C.H. (1967). Testing independence in two-way contingency tables with data subject to misclassification, *Psychometrika* **32**, 67–76.

[2] Birkett, N.J. (1992). Effects of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure, *American Journal of Epidemiology* **136**, 356–362.

[3] Bross, I. (1954). Misclassification in 2×2 tables, *Biometrics* **10**, 488–495.

[4] Chen, T.T. (1989). A review of methods for misclassified categorical data in epidemiology, *Statistics in Medicine* **8**, 1095–1106.

[5] Chen, T.T., Hochberg, Y. & Tenenbein, A. (1984). Analysis of multivariate categorical data with misclassification errors by triple sampling schemes, *Journal of Statistical Planning and Inference* **9**, 177–184.

[6] Cochran, W.G. (1968). Errors of measurement in statistics, *Technometrics* **10**, 637–666.

[7] Copeland, K.T., Checkoway, H., McMichael, A.J. & Holbrook, R. (1977). Bias due to misclassification in the estimation of relative risk, *American Journal of Epidemiology* **105**, 488–495.

[8] Dalenius, T. (1977). Bibliography of non-sampling errors in surveys, *International Statistical Review* **45**, 71–89, 181–197, 303–317.

[9] Duffy, S.W., Rohan, T.E. & Day, N.E. (1989). Misclassification in more than one factor in a case-control study: A combination of Mantel-Haenszel and maximum likelihood approaches, *Statistics in Medicine* **8**, 1529–1536.

[10] Ekholm, A. & Palmgren, J. (1987). Correction for misclassification using doubly sampled data, *Journal of Official Statistics* **3**, 419–429.

[11] Espeland, M.A. & Hui, S.L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors, *Biometrics* **43**, 1001–1012.

[12] Espeland, M.A. & Odoroff, C.L. (1985). Log-linear models for doubly sampled categorical data fitted by the EM algorithm, *Journal of the American Statistical Association* **80**, 663–670.

[13] Flegal, K.M., Keyl, P.M. & Nieto, F.J. (1991). Differential misclassification arising from nondifferential errors

in exposure measurement, *American Journal of Epidemiology* **134**, 1233–1244.

[14] Fung, K.Y. & Howe, G.R. (1984). Methodological issues in case-control studies. III: The effect of joint misclassification of risk factors and confounding factors upon estimation and power, *International Journal of Epidemiology* **13**, 366–370.

[15] Gladen, B. & Rogan, W.J. (1979). Misclassification and the design of environmental studies, *American Journal of Epidemiology* **109**, 607–616.

[16] Goldberg, J.D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table, *Journal of the American Statistical Association* **70**, 561–567.

[17] Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564–569.

[18] Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification, *Statistics in Medicine* **7**, 745–757.

[19] Greenland, S. & Robins, J.M. (1985). Confounding and misclassification, *American Journal of Epidemiology* **122**, 495–506.

[20] Gullen, W.H., Bearman, J.E. & Johnson, E.A. (1968). Effects of misclassification in epidemiologic studies, *Public Health Reports* **83**, 914–918.

[21] Haitovsky, Y. & Rapp, J. (1992). Conditional resampling for misclassified multinomial data with applications to sampling inspection, *Technometrics* **34**, 473–483.

[22] Harper, D. (1964). Misclassification in epidemiological surveys, *American Journal of Public Health* **54**, 1882–1886.

[23] Kaldor, J. & Clayton, D. (1985). Latent class analysis in chronic disease epidemiology, *Statistics in Medicine* **4**, 327–335.

[24] Korn, E.L. (1981). Hierarchical log-linear models not preserved by classification error, *Journal of the American Statistical Association* **76**, 110–113.

[25] Korn, E.L. (1982). The asymptotic efficiency of tests using misclassified data in contingency tables, *Biometrics* **38**, 445–450.

[26] Kuha, J. & Skinner, C. (1997). Categorical data analysis and misclassification, in *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz & D. Trewin, eds. Wiley, New York, pp. 633–670.

[27] Liu, X. & Liang, K.-Y. (1991). Adjustment for non-differential misclassification error in the generalized linear model, *Statistics in Medicine* **10**, 1197–1211.

[28] Marshall, J.R., Priore, R., Graham, S. & Brasure, J. (1981). On the distortion of risk estimates in multiple exposure level case-control studies, *American Journal of Epidemiology* **113**, 464–473.

[29] Mote, V.L. & Anderson, R.L. (1965). An investigation of the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data, *Biometrika* **52**, 95–109.

[30] Rubin, T., Rosenbaum, A.B. & Cobb, S. (1956). The use of interview data for the detection of association in field studies, *Journal of Chronic Diseases* **4**, 253–266.

[31] Savitz, D.A. & Barón, A.E. (1989). Estimating and correcting for confounder misclassification, *American Journal of Epidemiology* **129**, 1062–1071.

[32] Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection, *Technometrics* **14**, 187–202.

[33] Wachholder, S., Armstrong, B. & Hartge, P. (1993). Validation studies using an alloyed gold standard, *American Journal of Epidemiology* **137**, 1251–1258.

[34] Wachholder, S., Dosemeci, M. & Lubin, J.H. (1991). Blind assignment of exposure does not always prevent differential misclassification, *American Journal of Epidemiology* **134**, 433–437.

[35] Walter, S.D. & Irwig, L.M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, *Journal of Clinical Epidemiology* **41**, 923–937.

JOUNI KUHA, CHRIS SKINNER & JUNI PALMGREN

Misclassification Models

Classification errors in **categorical data** may distort results of statistical analyses (*see* **Misclassification Error**). Models for misclassification processes have been developed to study and compensate for the effects of such errors, and hence protect the validity of data analyses. Such models have been applied in a number of biostatistical contexts including modeling the natural history of a disease or growth process (*see* **Growth and Development**), evaluation of **diagnostic tests**, and analytic **epidemiology**. Specific scientific questions in these contexts may require inferences about (i) an underlying biological process, reflected in statistical **associations** that might be obscured or distorted by nuisance misclassification; (ii) the misclassification process itself; or (iii) both the underlying and misclassification processes.

The true underlying process involves a categorical response that may be univariate binary, nominal, ordinal, or multivariate with any of these components. In most applications, misclassification is represented by a Bernoulli or multinomial **random variable**. Commonly, the misclassification rate depends on the true underlying response. Various approaches that tie together misclassification and the true response process have been proposed. Different applications require different statistical methods, each with specific advantages and limitations. Depending on the application, the parameters of the misclassification model are estimated using data from either the current or an additional, supplementary study.

In modeling the natural history of a disease, interest usually focuses on understanding the stochastic changes in the disease process over time. The choice of models for the true disease or growth process depends on various factors including the **measurement scale** of the process and whether observations are taken at regular or irregularly spaced intervals. Unfortunately, classifications of disease severity are often subject to error, and analytic models that ignore misclassification are prone to biased inferences, particularly, if misclassification is related to the underlying disease process. Approaches to account for misclassification have been developed for modeling disease and growth processes including HIV/AIDS (*see* **AIDS and HIV**), hypertension, parasitic infection, and sexual maturation.

Investigators often wish to estimate the accuracy or, equivalently, the error rate, of a diagnostic test (*see* **Diagnostic Test Accuracy**). In this context, diagnostic error is simply another name for misclassification. For a binary disease status, error can be characterized jointly by the **false negative** and **false positive rates**, and accuracy by their respective complements: **sensitivity**, the probability of testing positive when the disease is present, and **specificity**, the probability of testing negative when the disease is absent. Sensitivity and specificity are simple to estimate when a definitive **gold standard test** exists, but not when a gold standard is nonexistent or too costly to obtain (*see* **Diagnostic Test Evaluation Without a Gold Standard**). Various **latent class** modeling approaches, in which multiple tests are used to determine a model-based consensus estimate of true disease status, have been proposed to estimate diagnostic accuracy and error rates for cancer biomarkers, diagnostic imaging, dental examinations, and pathological classification (*see* **Diagnostic Tests, Multiple**).

In analytic epidemiology, interest typically centers on assessing associations among categorical variables, of which one or more are subject to misclassification. Methods will be discussed to correct for such classification errors using error rates estimated from a subgroup or other population. For example, a **case–control study** designed to relate cervical cancer to sexual partner's circumcision status might elicit the latter by proxy report of the woman, subject to error, rather than by interview or physical examination of the male partner. Data on both proxy-reported and actual circumcision status might be obtained from a subsample or a small external population, and the observed error rates used to adjust inferences on the cancer-circumcision relationship.

Modeling Growth or the Natural History of a Disease

Let n_i be the number of observations made on the ith of N individuals, and let $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})'$ and $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ be random vectors respectively of observed and true disease states at each observation time. The Y_{ij} are assumed to be values of a **binary**, nominal, or ordinal variable describing the true underlying disease course or stage of growth, and the X_{ij} are the corresponding observed states after possible misclassification. The joint probabilities of

$\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ can be written as

$$P(\boldsymbol{Y}_i, \boldsymbol{X}_i) = P(\boldsymbol{X}_i|\boldsymbol{Y}_i)P(\boldsymbol{Y}_i), \tag{1}$$

where $P(\boldsymbol{Y}_i)$ is the probability of the true underlying disease or growth trajectory and $P(\boldsymbol{X}_i|\boldsymbol{Y}_i)$, which represents the misclassification process, is the **conditional probability** of the observed given the true trajectory.

Most approaches for modeling the natural history of a disease assume that the misclassification process at time t depends on the true disease process only through Y_{it} and not on the preceding path of true $(Y_{i1}, \ldots, Y_{i,t-1})$ or observed $(X_{i1}, \ldots, X_{i,t-1})$ disease stages, and thus that

$$P(\boldsymbol{X}_i|\boldsymbol{Y}_i) = \prod_{j=1}^{n_i} P(X_{ij}|Y_{ij}). \tag{2}$$

The misclassification process $P(\boldsymbol{X}|\boldsymbol{Y})$ and disease process $P(\boldsymbol{Y})$ are tied together through shared **covariates**.

Espeland et al. [14], Nagelkerke et al. [18], and Rosychuk and Thompson [22] have proposed models for longitudinal binary data subject to misclassification (*see* **Longitudinal Data Analysis, Overview**). Espeland et al. [14] focused on modeling an underlying progressive process with an absorbing state (i.e. one that individuals may enter but not leave). The authors use an **EM algorithm** [10] for parameter estimation and apply their methodology to dichotomous maturation data. Nagelkerke et al. [18] and Rosychuk and Thompson [22] respectively proposed Markov and semi-Markov models for modeling parasitic infection with an alternating binary process (*see* **Markov Chains**; **Markov Processes**; **Transition Models for Longitudinal Data**). Applications to other chronic diseases have also been proposed [7].

Albert et al. [3] proposed a model for longitudinal ordinal data (*see* **Ordered Categorical Data**) with misclassification. Similar to Espeland et al. [14], their approach is for underlying progressive processes and is applied to longitudinal sexual maturation data. We elaborate on this example to illustrate modeling of longitudinal categorical data with misclassification. Tanner staging is an ordinal rating scale for gender-specific sexual maturation that ranges from one (no sexual development) to five (full development). The National Growth and Health Study (NGHS) followed 1155 girls from age 9 or 10 to age 16 at yearly intervals, recording growth-related measures including Tanner staging. Tanner staging is prone to misclassification; biologically impossible decreases in maturation were reported during follow-up for 45% of the 1155 girls in the NGHS data. Other errors, such as recording an increase in sexual maturation stage when a girl has not in fact progressed, are not obvious but probably occur as frequently as more blatant misclassification. Here, Y_{ij} is a latent random variable reflecting the true Tanner stage, and X_{ij} is the recorded Tanner stage measurement. Scientific interest focused on examining the effect of age at a given stage of maturation on the rate of progressing to subsequent stages, and on comparing both true sexual development and Tanner stage misclassification across racial groups.

Assuming that diagnostic error is independent across visits (i.e., $X_{ij}|Y_{ij}$ is independent of $X_{ij'}|Y_{ij'}$) and that the underlying maturation process follows a first-order Markov chain, the joint probability of $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ can be written as

$$P(\boldsymbol{X}_i, \boldsymbol{Y}_i) = P(\boldsymbol{X}_i|\boldsymbol{Y}_i)P(\boldsymbol{Y}_i) = \left(\prod_{j=1}^{n_i} P(X_{ij}|Y_{ij})\right) \times \left(P(Y_{i1}) \prod_{j=2}^{n_i} P(Y_{ij}|Y_{i,j-1})\right). \tag{3}$$

Denote the probabilities governing the underlying maturation process as the initial state probabilities $p_l = P(Y_{i1} = l)$ and the transition probabilities from state l to state m as $p_{lm} = P(Y_{ij} = m|Y_{i,j-1} = l)$.

To exploit ordinality of the sexual development stages and obtain a parsimonious and interpretable parameterization, Albert et al. [3] used **proportional odds** parameterizations for the underlying maturation process (*see* **Proportional-odds Regression**). First, the p_l were reexpressed in cumulative form as

$$\gamma_l = \text{logit } P(Y_{i1} \le l) = \text{logit } \left(\sum_{u=1}^{l} p_u\right), \tag{4}$$

where $-\infty \le \gamma_1 \le \gamma_2 \le \cdots \le \gamma_{k-1} < \gamma_k = \infty$. Second, the transition probabilities p_{lm} were similarly reparameterized as

$$\theta_{lm} = \text{logit } P(Y_{ij} \le m|Y_{ij-1} = l) = \text{logit}\left(\sum_{i=u}^{m} p_{lu}\right), \tag{5}$$

where $-\infty = \theta_{l1} = \theta_{l2} = \cdots = \theta_{l,l-1} \leq \theta_{ll} \leq \theta_{l,l+1} \leq \cdots \leq \theta_{l,k-1} \leq \theta_{lk} = \infty$ are restrictions on the parameters such that only monotonic increases in the underlying process are possible. Further, parameter reduction was achieved by presuming the θ_{lm} for each l to be linear in stages traversed, that is,

$$\theta_{lm} = \theta_l + \alpha_l(m - l) \tag{6}$$

for $l \leq m \leq k-1$ with $\alpha_l > 0, l = 1, \ldots, k-2$. Note that $\alpha_l \geq 0$ ensures that all the transition probabilities are nonnegative, as required by $P(Y_{ij} \leq m'|Y_{i,j-1} = l) \geq P(Y_{ij} \leq m|Y_{i,j-1} = l)$ if $m' \geq m$. For each maturation level l, large θ_l corresponds to high probability of remaining in state l, while α_l determines, given θ_l, the relative probabilities that girls who mature in a given year will mature one Tanner stage, or more than one stage.

Albert et al. additionally modeled the misclassification mechanism as

$$P(X_{ij} = m|Y_{ij} = l) = \begin{cases} \dfrac{\psi(l,m)}{1 + \sum_{\substack{\omega=1 \\ \omega \neq l}}^{k} \psi(l,\omega)} & l \neq m \\ \dfrac{1}{1 + \sum_{\substack{\omega=1 \\ \omega \neq l}}^{k} \psi(l,\omega)} & l = m \end{cases}, \tag{7}$$

where $\psi(l,m) = \exp(\lambda_l + \eta_l|l-m|)$. This parameterization specifies symmetric misclassification around the true state; for example, one is just as likely to underestimate as overestimate a child's true state by one. Large negative values of λ_l reflect high probability of correctly classifying the lth state, and of η_l reflect low chance of misclassifying by more than one state. An elaboration allowing asymmetric misclassification is

$$\psi(l,m) = \exp(\lambda_l + \eta_{1l}(l-m)I_{(l>m)} + \eta_{2l}(m-l)I_{(l<m)}), \tag{8}$$

where $I_{(x)} = 1$ when x is true and 0 otherwise.

Maximum likelihood estimation may be based on the marginal distribution (*see* **Marginal Probability**) of the observed $\boldsymbol{X}_i$'s, obtained as

$$L = \prod_{i=1}^{N} P(\boldsymbol{X}_i) = \prod_{i=1}^{N} \Bigg(\sum_{i_1=0}^{k} \sum_{i_2=i_1}^{k} \cdots \sum_{i_n=i_{n-1}}^{k} P(\boldsymbol{X}_i|\boldsymbol{Y}_i = (i_1, i_2, \ldots, i_n)) P(\boldsymbol{Y}_i = (i_1, i_2, \ldots, i_n)) \Bigg). \tag{9}$$

Direct maximization of this likelihood is computationally infeasible with five categories and eight follow-up time points, as in the Tanner staging example. An EM algorithm [10], incorporating a backward–forward algorithm (originally developed for fitting **hidden Markov models**) for evaluating the E-step [6], was used. Standard errors of model parameters were estimated using the nonparametric **bootstrap**. Linear terms were added to (6) (i.e. $\beta_l \text{Age}_{ij}, l = 1, 2, 3$, and 4, where Age_{ij} is the age of the ith child at the jth follow-up time), to allow the maturation process to depend on age. Separate models fit to white and black girls showed racial differences in both maturation and misclassification processes. Relative to their white counterparts, black girls were more mature at ages 9 to 10, passed more rapidly through Tanner stages (1–2), and were misclassified more frequently (tending to be classified at higher than actual maturity).

In some settings, disease states are defined by ranges of continuous measurements. For example, various stages in the progression of HIV/AIDS are defined by intervals of CD4 counts [23] (*see* **AIDS and HIV**). Errors in the continuous measurements then induce misclassification of disease state in a straightforward fashion. Satten et al. [23] proposed a misclassification model that exploits this structure. They model both the latent true value of the continuous variable conditional on the disease stage, and error in the continuous measurement. For the jth observation on individual i, let S_{ij} be the true disease state (e.g. disease stage for HIV/AIDS by intervals of CD4 counts), Y_{ij} be the continuous observation measured without error (e.g. the true CD4 count), and X_{ij} be the continuous observation measured with error (e.g. the observed CD4 count). Satten et al. [23] modeled the $\boldsymbol{S}_i = (S_{i1}, S_{i2}, \ldots, S_{in_i})'$ using a Markov process, and the Y_{ij} given S_{ij} as uniform over each interval for all stages but the highest unbounded interval, for which they used the upper half of a **log-normal distribution**. The Y_{ij} were assumed to be

observed only after perturbation by Gaussian measurement errors ε_{ij} with mean 0 and variance σ^2: $X_{ij} = Y_{ij} + \varepsilon_{ij}$. Estimation is based on maximizing the marginal likelihood of the observed X_{ij}'s, through an EM algorithm [10] with E-step using a backward–forward algorithm [6].

Motivated by a study of hypertension using data from the **Framingham** Heart Study, Albert [1] proposed a similar model for population-based follow-up where a binary disease status variable is defined by dichotomizing a continuous variable measured with error (*see* **Categorizing Continuous Variables**). At any follow-up time, an individual is assumed to be in the disease state (e.g. hypertensive) if the continuous variable (e.g. true diastolic blood pressure, DBP) exceeds a threshold (e.g. 95 mmHg), and disease-free otherwise. The population is assumed to be comprised of three types of individuals: (i) those always in the disease state, (ii) those never in the disease state, and (iii) those who migrate between states according to a two state Markov chain; that is, the population follows a "mover–stayer" model. The density of the continuous variable Y_{ij}, conditional on the true binary disease state S_{ij}, is modeled as

$$f(y|S_{ij}) = \begin{cases} 2\phi\left(\frac{y-c}{\sigma_1}\right) & \text{if } S_{ij} = 1 \text{ and } y > c \\ 2\phi\left(\frac{c-y}{\sigma_0}\right) & \text{if } S_{ij} = 0 \text{ and } y \le c \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where $\phi(z)$ denotes the normal density at z (*see* **Normal Distribution**), c is the established cutoff for defining disease in the true continuous measurement, and σ_0^2 and σ_1^2 characterize the variability in the true continuous measurement given the true disease status 0 or 1. Y_{ij} is assumed to be measured with Gaussian $(0, \sigma^2)$ error as described above (i.e. we observe X_{ij}, where $X_{ij} = Y_{ij} + \varepsilon_{ij}$ and where ε_{ij} is Gaussian $(0, \sigma^2)$ error). Disease incidence as well as both point and period prevalence may be obtained as functions of model parameters, and estimated after maximizing the **marginal likelihood** of the X_{ij}'s using an **algorithm** similar to that of Satten et al. [23].

Modeling Diagnostic Error without a Gold Standard

In contrast to modeling natural history of a disease or growth process, interest in this application usually focuses on estimating the diagnostic error or misclassification and not the underlying true response. Let Y_{ij} be the test result for the j^{th} of n dichotomous tests on individual i, whose true disease status is d_i, and let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$. All approaches assume a latent class model, where d_i is the latent class. The elements of $\boldsymbol{Y}_i$ have joint distribution

$$P(Y_{i1}, Y_{i2}, \ldots, Y_{in}) = \sum_{l=0}^{1} P(Y_{i1}, Y_{i2}, \ldots, Y_{in}|d_i = l)P(d_i = l), \tag{11}$$

where $P(d_i = 1)$ is the **prevalence** of disease in the population. The initial work in this area [9, 15] assumed that test responses are conditionally independent given true disease status, namely,

$$P(Y_{i1}, Y_{i2}, \ldots, Y_{in}|d_i) = \prod_{j=1}^{n} P(Y_{ij}|d_i). \tag{12}$$

The parameters of the model are the true **prevalence** $P(d_i = l)$ and the sensitivities and specificities, respectively $P(Y_{ij} = 1|d_i = 1)$ and $1 - P(Y_{ij} = 1|d_i = 0)$, for each of the n tests. More than 2 tests are required (i.e. $n \ge 3$) to identify these parameters. Maximum likelihood has been proposed for parameter estimation. Vacek [27] and Torrance-Rynard and Walter [25] investigated the effect of conditional independence on estimating diagnostic error, and showed that parameter estimators of sensitivity and specificity are usually biased when conditional independence is falsely assumed.

A number of papers have incorporated dependence between tests [16]. Espeland and Handelman [11] proposed **loglinear** models with higher order interaction terms to represent associations between tests. Models that incorporate conditional dependence through the introduction of **random effects** have also been proposed [21, 26]. In one random effects formulation [21],

$$P(Y_{i1}, Y_{i2}, \ldots, Y_{in}|d_i) = \int \left(\prod_{j=1}^{n} P(Y_{ij}|d_i, b) \right) \phi(b)\, \mathrm{d}b, \tag{13}$$

where $P(Y_{ij}|d_i, b) = \Phi(\beta_{jd_i} + \sigma_{d_i} b)$ and where $\phi(x)$ and $\Phi(x)$ are the standard normal density and cumulative distribution function, respectively. This integral can be numerically evaluated using Gaussian quadrature (*see* **Numerical Integration**). The sensitivities and specificities for each test can be evaluated by marginalizing over the random effect, and prove to be $\Phi\left(\beta_{j1}/\sqrt{(1+\sigma_1^2)}\right)$ and $1 - \Phi\left(\beta_{j0}/\sqrt{(1+\sigma_0^2)}\right)$, respectively.

Albert et al. [4] proposed a finite mixture model in which some individuals are always classified correctly by any test, while others are subject to diagnostic error. Let l_{id_i} be an indicator of whether the ith individual, given disease status d_i, is always classified correctly (i.e. $l_{i1} = 1$ when a truly positive specimen always tests positive and $l_{i0} = 1$ when a truly negative test always is rated negative), where $P_0 = P(l_{i0} = 1)$ and $P_1 = P(l_{i1} = 1)$. The probabilities of testing positive given d_i and l_{id_i} are

$$P(Y_{ij} = 1|d_i, l_{id_i}) = \begin{cases} 1 & \text{if } d_i = 1 \text{ and } l_{i1} = 1 \\ 0 & \text{if } d_i = 0 \text{ and } l_{i0} = 1 \\ \rho_j(1) & \text{if } d_i = 1 \text{ and } l_{i1} = 0 \\ 1 - \rho_j(0) & \text{if } d_i = 0 \text{ and } l_{i0} = 0 \end{cases}, \quad (14)$$

where $\rho_j(d_i)$ is the probability of the jth test making a correct diagnosis given that the individual is subject to diagnostic error ($l_{i1} = 0$ or $l_{i0} = 0$). The sensitivity and specificity for the jth test are $P(Y_{ij} = 1|d_i = 1) = P_1 + (1 - P_1)\rho_j(1)$ and $P(Y_{ij} = 0|d_i = 0) = P_0 + (1 - P_0)\rho_j(0)$, respectively. This finite mixture model is closely related to a latent class model of Espeland and Handelman [11], in which latent classes corresponding to unambiguously positive and negative cases are incorporated.

In a fourth approach to incorporating conditional dependence, Yang and Becker [28] proposed a **marginal model** that assumes only second-order interactions between tests conditional on the true disease status d_i (i.e. no third- or higher-order interactions). The major advantage of their approach is that estimates of sensitivity and specificity are simple functions of estimated model parameters. Yang and Becker parameterize $P(Y_{i1}, Y_{i2}, \ldots, Y_{in}|d_i)$ in (11) in terms of $\theta_j = \text{logit } P(Y_{ij} = 1|d_i)$, $j = 1, 2, \ldots, n$, as well as in terms of all pairwise log-**odds ratio** associations between tests

$$\gamma_{j,j'}(d_i) = \log\left[\frac{P(Y_{ij} = 0, Y_{ij'} = 0|d_i)P(Y_{ij} = 1, Y_{ij'} = 1|d_i)}{P(Y_{ij} = 0, Y_{ij'} = 1|d_i)P(Y_{ij} = 1, Y_{ij'} = 0|d_i)}\right]. \quad (15)$$

Yang and Becker propose an EM algorithm for parameter estimation and illustrate their methodology with 4 different tests of HIV given to 428 patients.

Shih and Albert [24] proposed a general methodology for analyzing **correlated binary data** subject to misclassification or diagnostic error. Their approach can be applied in situations in which a binary outcome is measured repeatedly in time or space and/or evaluated by several tests or raters, and in which the true disease status may change over time or space. Let y_{ijk} denote the observed binary response of the ith individual evaluated at the jth time point or spatial location by the kth test, and let d_{ij} be the true disease status of individual i at the jth time point or spatial location. Also, let $\boldsymbol{x}_{ij}$ denote a vector of covariates that may change across individuals and time/space. The following **generalized linear mixed model** was proposed for the true, correlated disease states within individuals:

$$\text{logit}\{P(d_{ij} = 1|\boldsymbol{x}_{ij}, b_i)\} = \boldsymbol{\beta}'\boldsymbol{x}_{ij} + b_i, \quad (16)$$

where b_i is a Gaussian random effect with mean 0 and variance σ^2. The misclassification model is

$$\text{logit}\{P(y_{ijk} = 1|d_{ij}, b_i)\} = \gamma_1 d_{ij} + \gamma_2(1 - d_{ij}) + \gamma_3 b_i, \quad (17)$$

where γ_1 and γ_2 govern the false negative and false positive rates, respectively. The fact that b_i is shared between the response and misclassification models induces a relationship between the probability of a true response and the probability of a misclassification. Specifically, the probability of a misclassification conditional on b_i is given by

$$P(d_{ij} = 0|b_i)P(y_{ijk} = 1|d_{ij} = 0, b_i) + P(d_{ij} = 1|b_i)P(Y_{ijk} = 0|d_{ij} = 1, b_i). \quad (18)$$

After substitution of (16) and (17), careful inspection of (18) reveals that the misclassification probability is smallest when b_i is positive and large, corresponding to high probability of disease, and when b_i is negative and large, corresponding to low probability of disease. This feature is attractive, since we would expect all tests to agree in unambiguous cases in which the probabilities of disease are either very close to one or zero.

Cook et al. [8] proposed another model in which multiple raters or tests assess the status of a binary disease process at multiple time points. The disease process is modeled with a first-order Markov chain but, as the authors note, extension to a Markov process for irregularly spaced binary observations is straightforward. The model allows for conditional dependence between raters at a given time point with a loglinear formulation similar to that of Espeland and Handelman [11].

There has been some criticism of these models. Alonzo and Pepe [5] criticize latent class modeling approaches for estimating diagnostic error without a gold standard, noting identifiability problems when using a small number of tests, and questioning whether a model-based consensus of truth makes biological sense.

Epidemiology

In this context, interest centers on estimating associations between categorical variables when some or all of these variables are subject to misclassification. There is a substantial literature on the effects of misclassification in association models, such as loglinear models. Ignoring random classification errors attenuates estimates of association, and results in loss of **power** for testing statistical significance. Estimators and tests of association may be even more severely compromised when misclassification depends on the variables involved in the associations.

For instance, **logistic regression** when the outcome variable is subject to misclassification has been discussed by various authors. Neuhaus [19] showed that ignoring misclassification in the binary responses can lead to highly biased estimates of model parameters (*see* **Unbiasedness**), while analysis accounting for misclassification may sacrifice substantial **efficiency** relative to analysis of the true responses. Neuhaus [20] presents analytic expressions demonstrating bias in parameter estimates for correlated binary models. He focuses on examining bias for a generalized linear mixed model with binary response, and for marginal approaches such as **generalized estimating equations**.

Since the categories of a binary observation may always be labelled as presence and absence of an event or characteristic, it is reasonable to adopt the terminology of diagnostic testing to describe misclassification in this context. Thus, letting $Y_i = 1$ if the disease or other characteristic is truly present for the ith individual and $Y_i = 0$ otherwise, and $T_i = 1$ if the ith subject is classified as diseased and $T_i = 0$ otherwise, we define the sensitivity and specificity of the observational process as respectively $sens = \text{Prob}(T_i = 1|Y_i = 1)$ and $spec = \text{Prob}(T_i = 0|Y_i = 0)$.

Magder and Hughes [17] proposed an EM algorithm for fitting logistic regression models when sensitivity is known or can be estimated from previous studies. They showed how the approach can be implemented with standard **software**, and discussed its extension to situations in which sensitivity and specificity are both unknown, emphasizing that estimation when diagnostic error is unknown may be sensitive to modeling assumptions. Given a set of known covariates $\boldsymbol{X}_i$, Magder and Hughes proposed a latent logistic regression model

$$\text{Prob}(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_i'\boldsymbol{\beta})}, \qquad (19)$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. Model (19) can be fit with standard logistic regression by including each individual as both diseased and nondiseased with weights equal to $\hat{Y}_i$ and $(1 - \hat{Y}_i)$, respectively, where $\hat{Y}_i$ is the probability that the ith individual is truly diseased given the values of T_i, $\boldsymbol{X}_i$, and $\boldsymbol{\beta}$. In the diagnostic testing context, $\hat{Y}_i$ is known as the "positive **predictive value**" when $T_i = 1$, and $(1 - \hat{Y}_i)$ is known as the "negative predictive value" when $T_i = 0$. Using **Bayes Theorem**, when $T_i = 1$,

$$\hat{Y}_i = \frac{\text{Prob}(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta})(sens)}{\begin{array}{c}\text{Prob}(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta})(sens) + \\ \text{Prob}(Y_i = 0|\boldsymbol{X}_i, \boldsymbol{\beta})(1 - spec)\end{array}}. \qquad (20)$$

Similarly, when $T_i = 0$,

$$\hat{Y}_i = \frac{\text{Prob}(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta})(1 - sens)}{\begin{array}{c}\text{Prob}(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta})(1 - sens) + \\ \text{Prob}(Y_i = 0|\boldsymbol{X}_i, \boldsymbol{\beta})(spec)\end{array}}. \qquad (21)$$

When *sens* and *spec* are known, Magder and Hughes proposed computing maximum likelihood estimates of $\boldsymbol{\beta}$ by iterating between fitting (19) using standard logistic regression software (again by including each subject as both a diseased and nondiseased observation with weights $\hat{Y}_i$ and $1 - \hat{Y}_i$, respectively) and reestimating the weights $\hat{Y}_i$ using (20) and (21) until the parameter values converge. When either or both of *sens* and *spec* are unknown, Magder and Hughes proposed an additional step in the iteration, at which the unknown quantities are estimated by

$$\widehat{sens} = \frac{\sum_i \hat{Y}_i T_i}{\sum_i \hat{Y}_i}$$

$$\widehat{spec} = \frac{\sum_i (1 - \hat{Y}_i)(1 - T_i)}{\sum_i (1 - \hat{Y}_i)}. \qquad (22)$$

However, Magder and Hughes mentioned that estimating sensitivity and specificity may be problematic for a number of reasons. First, for saturated models, parameter estimates may not be identifiable. Second, estimates may strongly depend on parametric assumptions in the logistic regression model.

There is also a large literature on modeling categorical data with misclassification under **double sampling**. In this situation, expensive and presumably error free methods are used in a smaller subsample, to assess misclassification in more extensive data obtained with comparatively economical but error prone procedures. For instance, Espeland and Odoroff [13] proposed a loglinear modeling approach allowing maximum likelihood estimation via an EM algorithm, and yielding straightforward expressions for variance estimates. Specifically, let A, B, and C be categorical factors observed on a large sample of N individuals and subject to misclassification. Denote A^*, B^*, and C^* as categorical factors not subject to error measured on a subsample of n of the original N individuals. Also, denote L as a binary factor indicating whether an individual is in the subsample in which more expensive confirmatory testing is undertaken. Espeland and Odoroff parameterized three distinct components required for double sampling models:

1. The model for the subsampling process, which describes the relationship between L and the other factors, is called the *sampling model*.
2. The model for the relationship between A and A^*, B and B^*, C and C^* is called the *misclassification model*.
3. The model for the relationship between A, B, and C is referred to as the *experimental model*.

These authors proposed a loglinear model for parameter estimation and impute missing cell counts (e.g. owing to the omission of more expensive and accurate measurements for individuals outside the subsample; *see* **Missing Data**) using an EM algorithm.

Espeland and Hui [12] illustrated the loglinear modeling approach with numerous epidemiologic examples including (i) a study of the association between cervical cancer and sexual partner's circumcision status, where circumcision status is self-reported for the entire sample of male partners, and a physical exam is done on a small subsample of these partners; (ii) a case–control study of the association between peptic ulcers and stomach cancer, in which self-reported history of peptic ulcer is obtained on all subjects, and a more careful peptic ulcer history is performed on a subset of subjects in the original sample.

Conclusions

Misclassification models have been applied in various application areas; we have discussed the modeling of a disease process, diagnostic error modeling, and association modeling in epidemiologic studies. When misclassification error rates are unknown and cannot or have not been estimated in previous studies, various methods for simultaneously estimating misclassification error and true response model parameters have been proposed. Although **identifiability** of parameters has been established for many of these models, the **robustness** of inference to misclassification modeling assumptions has received little examination. However, recent work of Albert and Dodd [2] on estimating diagnostic error without a gold standard demonstrates that estimators of diagnostic error may be asymptotically biased if the conditional dependence structure is misspecified. To complicate the situation, when the number of tests is small, the expected log-**likelihoods** for models with

different conditional dependence structures may be nearly identical [2], making model selection based on likelihood comparisons difficult or impossible. Thus, unverifiable modeling assumptions may be required to make inference on diagnostic errors without a gold standard. This suggests that robustness of misclassification error estimation to modeling assumptions in other contexts is an important area for future research.

References

[1] Albert, P.S. (1999). A mover-stayer model for longitudinal marker data, *Biometrics* **55**, 1252–1257.

[2] Albert, P.S. and Dodd, L. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard, *Biometrics* **60**, 427–435. submitted.

[3] Albert, P.S., Hunsberger, S.A. & Biro, F.M. (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation, *Journal of the American Statistical Association* **92**, 1304–1311.

[4] Albert, P.S., McShane, L.M., Shih, J.H. & the U.S. National Cancer Institute Bladder Tumor Marker Network. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors, *Biometrics* **57**, 610–619.

[5] Alonzo, A. & Pepe, M. (2001). Using a combination of reference tests to assess the accuracy of a diagnostic test, *Statistics in Medicine* **18**, 2987–3003.

[6] Baum, L.E., Petrie, T., Soules, G. & Weiss, N.A. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics.* **41**, 164–171.

[7] Bureau, A., Shiboski, S. & Hughes, J.P. (2003). Applications of continuous time hidden Markov models to the study of misclassified disease outcomes, *Statistics in Medicine* **57**, 610–619.

[8] Cook, R.J., Ng, E.T.M. & Meade, M.O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models, *Biometrics* **56**, 1109–1117.

[9] Dawid, A.P. & Skene, A.M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied Statistics* **28**, 20–28.

[10] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM-algorithm, *Journal of the Royal Statistical Society. Series B* **39**, 1–38.

[11] Espeland, M.A. & Handelman, S.L. (1989). Using latent class models to characterize and assess relative error in discrete measurements, *Biometrics* **45**, 587–599.

[12] Espeland, M.A. & Hui, S.L. (1987). A general approach to analyzing epidemiologic data than contain misclassification errors, *Biometrics* **43**, 1001–1012.

[13] Espeland, M.A. & Odoroff, C.L. (1985). Log-linear models for doubly sampled categorical data fitted by the EM algorithm, *Journal of the American Statistical Association* **80**, 663–670.

[14] Espeland, M.A., Platt, O.S. & Gallagher, D. (1989). Joint estimation of incidence and diagnostic error rates from irregular longitudinal data, *Journal of the American Statistical Association* **84**, 972–979.

[15] Hui, S.L. & Walters, S.D. (1980). Estimating the error rates of diagnostic tests, *Biometrics* **36**, 167–171.

[16] Hui, S.L. & Zhou, X.H. (1998). Evaluation of diagnostic tests without gold standards, *Statistical Methods in Medical Research* **7**, 354–370.

[17] Magder, L.S. & Hughes, J.P. (1997). Logistic regression when the outcome is measured with uncertainty, *American Journal of Epidemiology* **146**, 195–203.

[18] Nagelkerke, N.J.D., Chunge, R.N. & Kinoti, S.N. (1990). Estimation of parasitic infection dynamics when detectability is imperfect, *Statistics in Medicine* **9**, 1211–1219.

[19] Neuhaus, J.M. (1999). Bias and efficiency loss due to misclassified responses in binary regression, *Biometrika* **86**, 843–855.

[20] Neuhaus, J.M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification, *Biometrics* **58**, 675–683.

[21] Qu, Y., Tan, M. & Kutner, M.H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests, *Biometrics* **52**, 797–810.

[22] Rosychuk, R.J. & Thompson, M.E. (2001). A semi-Markov model for binary longitudinal responses subject to misclassification, *The Canadian Journal of Statistics* **29**, 395–404.

[23] Satten, G.A. & Longini, I.M. (1996). Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease, *Applied Statistics* **45**, 275–309.

[24] Shih, J.H. & Albert, P.S. (1999). A latent model for correlated binary data with diagnostic error, *Biometrics* **55**, 1232–1235.

[25] Torrance-Rynard, V.L. & Walter, S.D. (1997). Effects of dependent errors in the assessment of diagnostic test performance, *Statistics in Medicine* **97**, 2157–2175.

[26] Uebersax, J.S. & Grove, W.M. (1993). A latent trait finite mixture model for the analysis of rating agreement, *Biometrics* **49**, 823–835.

[27] Vacek, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests, *Biometrics* **41**, 959–968.

[28] Yang, I. & Becker, M.P. (1997). Latent variable modeling of diagnostic accuracy, *Biometrics* **53**, 948–958.

PAUL S. ALBERT

Missing Data Estimation, "Hot Deck" and "Cold Deck"

The terms *hot deck* and *cold deck* refer to two related classes of methods used for imputation of missing values in sample surveys (*see* **Multiple Imputation Methods**; **Missing Data**). Many methods of imputation, especially those developed and used before the 1980s, involve direct "donation" of the value of a specific item from a record that has a measured value for the item to a record having nonresponse on the particular item. In his monograph entitled *Compensating for Missing Data in Surveys*, Kalton [2] defines the class of *hot deck* methods as any method involving such "donation" that takes the "donors" from the same sample survey as the recipients. The class of *cold deck* methods refers to analogous methods that take the donors from past data. As noted by Kalton [2], however, the term *hot deck* more often refers to a more specific set of sequential methods that has been used extensively for the imputation of items in the *Current Population Survey* [1]. We use it in this more restrictive sense and describe one variation of it below.

Description of the Hot Deck Method

In the hot deck method, cells are defined on the basis of variables that are considered "important" for imputation. These are generally variables that relate to the particular sample design used (e.g. **cluster**, **stratified**, etc.) or to demographic or other variables. The data are then sorted first according to these defined cells and secondly by other variables that are considered relevant for imputation. Following this, a register is then defined for each of the defined cells consisting of values of each variable that is to be imputed. The initial value of the register would consist of the value of each variable to be imputed for the first record in each cell that has recorded values for each of these variables. In a single pass through the data, the cell of each record is identified and, if a variable is missing, then it is given the value of that variable that is in the register for that cell. If, however, the record is complete on all variables to be imputed, then the values of the variables for this record replace the previous values in the registry for that cell. This process is repeated until all missing values are imputed. The process is the same for the cold deck method with the exception that the "donor" data are from a past survey.

Example

Let us consider a sample survey conducted on males 65 years of age and older in a large city and based on a two-stage cluster sample (*see* **Multistage Sampling**) in which the primary sampling units (PSUs) are blocks and the second stage units are households. The data consist of information collected from all males 65 years of age and older in 10 sample households within three sample blocks. The cell to be used in imputation are block (three categories) and age group (65–74 years/75 years and older). Data to be imputed are history of stroke ($1 = \text{yes}/2 = \text{no}$) and history of hypertension ($1 = \text{yes}/2 = \text{no}$). The data for each person, sorted by block and age group and listed by household within each of the six block–age group cells, are shown in Table 1. The records comprising the initial values of the register for each of the six cells are highlighted in boldface (records 2, 9, 14, 23, 26, and 34).

As one moves down the table, the first missing value is history of stroke in record 1. That value "2" is donated from record 2 (the record in the initial register for the block 1 – age group 65–74 cell). The next record having a missing value is record 5, which is also in the block 1 – age group 65–74 cell, and that record has missing values for both history of hypertension and history of stroke. The value "1" for history of hypertension and "1" for history of stroke is donated to record 5 from record 4 which comprises the registry at that point in the process. The entire imputation process proceeds in this fashion and is summarized in Table 2.

Comment

The above example illustrates one of a variety of similar methods that are generally discussed under the rubric *hot deck*. The major strengths of this category of methods are: (i) that data can be imputed very simply in a single "pass" of the data file; (ii) it always

Table 1 Data from sample survey of males 65 years of age and above

Record	Block	Age group "1" = 65–74 "2" = 75+	Household	History of hypertension "1" = Yes "2" = No	History of stroke "1" = Yes "2" = No
1	1	1	1	1	–
2	**1**	**1**	**2**	**1**	**2**
3	1	1	2	2	2
4	1	1	3	1	1
5	1	1	4	–	–
6	1	1	5	1	2
7	1	1	6	2	1
8	1	1	10	–	–
9	**1**	**2**	**1**	**1**	**1**
10	1	2	7	2	2
11	1	2	8	2	–
12	1	2	9	1	1
13	2	1	1	2	–
14	**2**	**1**	**3**	**2**	**2**
15	2	1	4	–	2
16	2	1	5	2	2
17	2	1	5	–	–
18	2	1	7	2	2
19	2	1	8	1	1
20	2	1	8	1	2
21	2	1	9	–	1
22	2	1	10	2	2
23	**2**	**2**	**2**	**1**	**1**
24	2	2	6	1	1
25	2	2	10	1	–
26	**3**	**1**	**1**	**1**	**1**
27	3	1	2	2	1
28	3	1	3	2	2
29	3	1	6	1	2
30	3	1	7	1	1
31	3	1	8	2	2
32	3	1	8	1	1
33	3	1	9	1	2
34	**3**	**2**	**4**	**1**	**2**
35	3	2	5	1	–
36	3	2	10	1	1
37	3	2	10	1	–

uses data that have been observed in the same data set; and (iii) imputation can generally be performed on records that have missing values for several variables more easily by this method than by methods based on more explicit **regression** models. The major weaknesses are: (i) it is not based on an explicit model; (ii) the values that are imputed depend on the ordering of records within cells, which gives them an aura of being somewhat arbitrary; and (iii) in subsequent analyses, the imputed values are considered as "real", and this often has an unpredictable effect on the estimated **standard errors** of statistics generated from the sample survey.

In summary, the class of hot deck methods is of importance not only historically but also in situations where resources or time are not available to consider more rigorous contemporary methods such as multiple imputation. Further discussion of this method, especially as used in conjunction with more contemporary methods, is presented in texts by Little & Rubin [3], Rubin [5], and in a chapter by Little & Schenker [4].

Table 2 Hot deck imputation process for data in Table 1

Record having missing value ("recipient")	Cell	Record comprising register ("donor")	Imputed value(s)
1	block 1, age group 1	2	history of stroke = 2
5	block 1, age group 1	4	history of hypertension = 1 history of stroke = 1
8	block 1, age group 1	7	history of hypertension = 2 history of stroke = 1
11	block 1, age group 2	10	history of stroke = 2
13	block 2, age group 1	14	history of stroke = 2
15	block 2, age group 1	14	history of hypertension = 2
17	block 2, age group 1	16	history of hypertension = 2 history of stroke = 2
21	block 2, age group 1	20	history of hypertension = 1
25	block 2, age group 2	24	history of stroke = 1
35	block 3, age group 2	34	history of stroke = 2
37	block 3, age group 1	36	history of stroke = 1

References

[1] Hansen, R.H. (1978). *The Current Population Survey: Design and Methodology*, Technical Paper No. 40. US Bureau of the Census, Washington.

[2] Kalton, G. (1983). *Compensating for Missing Survey Data*. Survey Research Center, Institute of Social Research, The University of Michigan, Ann Arbor.

[3] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[4] Little, R.J.A. & Schenker, N. (1994). Missing data, in *Handbook for Modeling*, G. Arminger, C. Clogg & M.E. Sobel, eds. Plenum, New York, Chapter 2, pp. 39–75.

[5] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

PAUL S. LEVY

Missing Data in Clinical Trials

Missing data are a fact of life in **clinical trials**, and it is important that we have statistical techniques that can accommodate incomplete data. Missing data arise for a multiplicity of reasons, and indeed are so common that one should be highly suspicious of any clinical trial report that claims to have no missing data! An unusual reason for missing data occurred when a set of case record forms was destroyed in an explosion [1]. More commonly, a patient fails to attend a scheduled study visit, or a blood sample goes astray or is inadequately refrigerated. Sometimes, patients retrospectively withdraw their consent to participate in a trial. Equally, since **clinical trials protocols** tend to evolve over time, a new baseline **covariate** may be added to the record form, or a simplified **quality of life** questionnaire replaces a version that was unacceptable to the patients. Of course, such problems should be identified during pilot studies, but time constraints are such that it is not always possible to conduct adequate pilot tests. If there are changes to data collection during the course of a trial, then relevant data may not be recorded for the early patients. Another very common example of missing data is in trials where the outcome measure is a survival time. It would be unusual and generally rather inefficient to continue a trial until all of the recruited patients had died. Instead, we fix a final closing date, and analyze the data, accepting that the actual survival times are missing for the patients who are alive at the close of study (*see* **Censored Data**).

Types of Missing Data

There is no single approach that can cope with this variety of missing data. Before attempting to discuss solutions to the handling of missing data, it is helpful to classify different types of "missingness". Rubin [13] gives a useful taxonomy, and his terminology for the different mechanisms that can generate missing data is widely accepted (*see* **Missing Data**).

The first class comprises data that are "missing completely at random" (MCAR). Here, the presence or absence of an observation is completely unrelated to the value that might have been observed. For example, if a new question is added to a record form, then the data for early patients are expected to be MCAR. However, even in this simple situation, the missing data could fail to be MCAR if there is some underlying secular trend in the prognosis of the condition being studied. Similarly, if a laboratory test is not performed because equipment is out of order, then the data are most likely MCAR. In contrast, if a patient fails to attend an appointment, this might be, for example, because the patient feels too ill to travel, and since this reason is likely to be related to the outcome of the missing examination, it would be inappropriate to regard the data as MCAR.

Note that one must be very careful in interpreting the term MCAR. Even when the data are MCAR, the pattern of missing data can be systematic, as in the above example of a new question being added to a record form.

Data that are MCAR are generally straightforward to handle. For example, it would be valid to exclude the patients with incomplete data from analysis. This would not necessarily be efficient, but the approach would not introduce a systematic **bias** into the comparison of the treatments.

The second class comprises data that are "missing at random" (MAR). In this class, the fact that an observation is missing, after conditioning on the observed data, provides no further information. Murray & Findlay [11] give an example of a study of hypertension, in which patients were withdrawn from the study if, at a study assessment, their diastolic blood pressure (DBP) exceeded 110 mmHg. Thus, the fact that an individual's DBP was not recorded at the eight-week assessment provided no additional information to the observation that, at the four-week assessment, the DBP exceeded the threshold of 110 mmHg. Unlike for data that are MCAR, if the data are MAR, then it is not generally valid to exclude from the analysis the patients with incomplete data.

The key point when data are MAR is that there is no need to model the "missingness" mechanism explicitly. Provided that an appropriate statistical model can be developed to describe the observed data, then valid **inferences** can be made using **maximum likelihood** methods applied to the observed, incomplete data.

The final class comprises "nonignorable missing data". The most familiar example here is censored data, where it would be invalid to base inferences on

a **likelihood** function that incorporated data only from the uncensored observations. In the case of survival data, special techniques such as the **Kaplan–Meier estimator** have been developed which take due account of censoring, but any more advanced **parametric model** or semi-parametric model must incorporate a factor in the likelihood function which derives from the censored observations.

The fundamental point about nonignorable missing data is that it is not sufficient simply to model the observed data, but rather that the statistical model must incorporate the "missingness" mechanism.

Testing for Type of "Missingness"

In general, it is not possible to perform a formal test of the assumption that data are MAR, and instead one must rely on information that is external to the observed data [14]. Diggle & Kenward [4] give a special case of a model for **longitudinal data** with drop-outs, which allows one to test whether missing data are MCAR or MAR, but the method is heavily dependent on the validity of the general model within which the special cases of MCAR and MAR are nested. Thus the situation is circular, as the observed, incomplete data are insufficient to test the general model [14]. Nonignorable missing data mechanisms, and tests for types of "missingness", are areas of very active research. See, for example, Baker [2], and the references mentioned in that publication, for a lead into the current research.

Approaches to Analyzing Incomplete Data

A large number of techniques are available for analyzing incomplete data, ranging from *ad hoc* "fixes" to sophisticated modeling. The techniques can be grouped under three headings: analysis of complete cases, imputation, and modeling.

Analysis of Complete Cases

The simplest approach to handling missing data is to analyze only those individuals with complete data. Disadvantages of this approach are, first, that it is inefficient because we discard potentially useful information and, secondly, that we potentially introduce a bias if the incomplete cases are atypical of the study population. This second point cannot occur, by definition, if the data are MCAR; but in other situations, including when the data are MAR, an analysis based solely on complete cases is likely to be misleading.

Imputation

"Imputation" is the general term given to the process of filling in missing values. This approach can be very valuable, especially when the proportion of missing values is very low but, as with all missing data methods, should be used with care. Little & Rubin [8] give a good account of the different procedures and their limitations. A missing value can be replaced by an overall **mean** of the missing variable, or, using a **regression** approach, by its conditional **expectation** given the data observed for that individual. A further approach, popular within the pharmaceutical industry, is "last value" imputation [7]. This is generally used when there are drop-outs from longitudinal trials. In this approach, a missing value is replaced by the last observed value of that variable for the individual in question.

Imputations generally lead to a systematic underestimation of variability, if we proceed to analyze the completed data set as if all the data, including those imputed, were actually observed. The effects of this problem can be minimized by replacing each missing value with a value simulated from its conditional distribution, or by using the idea of **multiple imputation** [8]. With multiple imputation, the data set is completed several times over, and each completed data set is analyzed using standard complete-data techniques. The results of all of these analyses are then merged to give a final inference. If the imputed values are all derived from a single model of "missingness", the final inference incorporates the impact of the incomplete data under this model. If the completed data sets derive from different possible models of "missingness", then the final inference also reflects the uncertainty associated with model selection.

Modeling

Imputation procedures generally suffer from the lack of a sound theoretic basis. If, instead, we have a suitable statistical model, then inferences can be made using the maximum likelihood method. Indeed, if we have data which are MAR, then we need only

model the observed data. In this situation, the very flexible **EM algorithm** [3] allows maximum likelihood estimates to be derived. However, even in the most regular of situations, likelihood surfaces can be "badly behaved" when data are incomplete. Murray [9] gives an example with **bivariate normal** data, in which the likelihood surface has a saddle point.

With nonignorable missing data mechanisms, the statistical model must go further and include explicit assumptions about these mechanisms. As already mentioned, survival models that can incorporate censored observations are a good example of this approach. All such analyses are likely to be based on assumptions that can only be assessed using external information. A crucial part of any such analysis is a detailed **sensitivity analysis**, exploring the impact of departures from these assumptions [12].

As illustrated by Murray & Findlay [12], the appropriate handling of missing data can have a profound impact on the results of an analysis. The paper reports the results of a comparative trial in hypertension, where 131 of 429 (31%) patients had incomplete data, usually the result of being withdrawn according to the protocol when their blood pressure was inadequately controlled. In an analysis comparing the mean diastolic blood pressure in the two treatment groups at the end of the study, an inappropriate assumption of MCAR gave a difference of 1.2 mmHg in one direction, when the more appropriate assumption of MAR gave a difference of 1.2 mmHg in the other direction. The **standard error** was 1.1 mmHg, so the **bias** arising from the inappropriate assumption of MCAR was in excess of two standard errors.

Formulating the Question

Within the problem of identifying a suitable statistical approach to the analysis there often lies an issue of problem formulation [6]. What precise question is the analysis trying to address? The key distinction is between pragmatic and explanatory questions [10, 15] which, in the context of a clinical trial, is reflected in the distinction between "**intention-to-treat**" and "as-treated" analyses.

The explanatory approach asks a "what if" question. What would have been observed if patients had not withdrawn with side effects? What would have been observed if we had kept patients in the trial, even though their blood pressure was not controlled? The analysis referred to above [11], in which patients were withdrawn from a hypertension trial if their blood pressure was uncontrolled, leading to the data being MAR, addressed an explanatory "what if" question.

The pragmatic approach asks instead about the overall impact of a treatment on a patient. The fact that a patient is withdrawn from a trial is regarded as important in itself. The key to handling missing data in a pragmatic trial is somehow to include unfavorable events leading to withdrawals as part of the primary endpoint. For example, a patient who withdraws with side effects might be classified as a failure, whereas a patient who withdraws because all symptoms have resolved might be classified as a success.

An example of this approach is a trial of exercise tolerance in patients with heart failure following myocardial infarction. This was an ancillary protocol to the AIRE Study [16], a large mortality study in which patients were randomized to receive either placebo or an angiotensin-converting enzyme inhibitor. In the exercise substudy, patients undertook an exercise test six months after starting their randomized treatment. The active treatment proved very effective in reducing mortality (a reduction in hazard of 27%), which could potentially have invalidated the analysis of the exercise data. In theory, the least fit placebo patients, having died, may have left a selected healthy subset to undertake the exercise tests; whereas in the active group, more of the acutely ill patients may have survived, albeit with a limited ability to exercise. A simple comparison of average exercise duration in the two groups would then be misleading, as it would not compare like with like. The missing data process is nonignorable, since a missing value tells us that a patient was too unwell to exercise, or had already died. The analysis was therefore not based on average exercise duration. Instead, a threshold was set which defined a successful exercise test. A patient who died, or was unable to achieve the threshold exercise duration, was classified as a failure. This newly defined endpoint lost some efficiency, moving from a continuous outcome (exercise duration) to a **binary** one (success/failure), but assured no missing data and an **unbiased** analysis. Another example of this approach, which is close to a proposal of Gould [5], is reported by the Xamoterol in Severe Heart Failure Study Group [17].

References

[1] Bailey, I., Bell, A., Gray, J., Gullan, R., Heiskannan, O., Marks, P.V., Marsh, H., Mendelow, D.A., Murray, G., Ohman, J., Quaghebeur, G., Sinar, J., Skene, A., Teasdale, G. & Waters, A. (1991). A trial of the effect of nimodipine on outcome after head injury, *Acta Neurochirurgica* **110**, 97–105.

[2] Baker, S.G. (1996). The analysis of categorical case-control data subject to nonignorable nonresponse, *Biometrics* **52**, 362–369.

[3] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[4] Diggle, P. & Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion), *Applied Statistics* **43**, 49–93.

[5] Gould, A.L. (1980). A new approach to the analysis of clinical drug trials with withdrawals, *Biometrics* **36**, 721–727.

[6] Hand, D.J. (1994). Deconstructing statistical questions (with discussion), *Journal of the Royal Statistical Society, Series A* **157**, 317–356.

[7] Lewis, J.A. (1989). Correcting for the bias caused by dropouts in hypertension trials (letter), *Statistics in Medicine* **8**, 1302–1303.

[8] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[9] Murray, G.D. (1977). Contribution to the discussion of A.P. Dempster, N.M. Laird, and D.B. Rubin, *Journal of the Royal Statistical Society, Series B* **39**, 27–28.

[10] Murray, G.D. (1991). Statistical aspects of research methodology, *British Journal of Surgery* **78**, 777–781.

[11] Murray, G.D. & Findlay, J.G. (1988). Correcting for the bias caused by drop-outs in hypertension trials, *Statistics in Medicine* **7**, 941–946.

[12] Murray, G.D. & Findlay, J.G. (1989). Correcting for the bias caused by dropouts in hypertension trials (letter), *Statistics in Medicine* **8**, 1303–1304.

[13] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

[14] Rubin, D.B. (1994). Contribution to the discussion of P. Diggle and M.G. Kenward, *Applied Statistics* **43**, 80–82.

[15] Schwartz, D., Flamant, R. & Lellouch, J. (1980). *Clinical Trials*. Academic Press, London.

[16] The Acute Infarction Ramipril Efficacy (AIRE) Study Investigators (1993). Effect of ramipril on mortality and morbidity of survivors of acute myocardial infarction with clinical evidence of heart failure, *Lancet* **342**, 821–828.

[17] Xamoterol in Severe Heart Failure Study Group (1990). Xamoterol in severe heart failure, *Lancet* **336**, 1–6.

GORDON D. MURRAY

Missing Data in Epidemiologic Studies

In analytic epidemiologic studies, such as **case–control studies**, **cohort studies** and related designs, data are usually collected by questionnaire or interview, or are abstracted from existing records, such as hospital records on treatment or diagnosis, personnel employment records on occupational exposures, or death certificates. Except in studies with a two-stage design (see below), complete information is sought for all subjects included in the study.

In case–control studies, one requires data on previous exposures that may have occurred long before the study. Adequate planning and organization are required to try to ensure that data are collected in an identical way for diseased persons (cases) and for healthy subjects (**controls**). Data are also collected on known or suspected **confounding** variables in order to adjust for these variables in the analysis. Preliminary data on matching variables, such as information on sex and age, are needed for matched case–control designs (*see* **Matching**). In cohort studies personal interviews are carried out infrequently, but data are often abstracted from existing files or records. In occupational cohort studies one can use personnel records to obtain data on the occupational history and sometimes on specific exposures; sometimes records from the office of the occupational hygienist or routinely collected data from the medical officer will be useful (*see* **Occupational Epidemiology**). The quality and completeness of such data may differ substantially between companies or even departments of the same company. Data quality may also differ for different job categories and could therefore depend on the exposure of interest. Disease information in cohort studies is sometimes abstracted from hospital records or from cancer registry files (*see* **Disease Registers**). In mortality studies, the date and **cause of death** are abstracted from official death certificates or from other sources. An important issue in planning and organizing cohort studies is to try to guarantee a nonselective retrieval of information for the personal history (occupational history, lifestyle, residential history). It is also important to avoid any selective follow-up to obtain the date of diagnosis or date of death. The diagnosis and/or the causes of death should be assessed in a comparable way for exposed and nonexposed subjects (*see* **Bias in Case–Control Studies**; **Bias in Cohort Studies**).

Sources of Missing Values in Epidemiologic Research

Unplanned Missing Values

Despite well-organized data collection efforts, data may contain errors, the data collection is sometimes incomplete, and missing values occur. Missing data can arise as total **nonresponse** or as item nonresponse. Total nonresponse results from refusal of subjects to participate in the study or from inability to locate the selected subjects. For example, in **population-based case–control studies,** controls may have been selected but are not accessible because they have recently moved. Total nonresponse is a frequent source of **selection bias**. In this article we restrict ourselves to item nonresponse, which refers to the lack of data on one or several items from a study participant but not on all items.

Item nonresponse may arise because a person refuses to answer certain questions. For example, if the question is too sensitive (e.g. alcohol consumption, sexual behavior, income, or health-related questions), a study participant may refuse to answer that item. What is regarded as sensitive may differ from one person to the next, and it may vary with personal behavior and/or depend on the answer to the sensitive question. Older people may be more willing to answer a certain question than younger people. Persons with a very high or very low income may not be willing to report it. Another reason for missing values is that subjects do not know the answer because they are unable to recall certain events. It also happens that a given answer is inconsistent with other answers and can therefore not be used in the analysis, as when a person says on one part of the questionnaire that she never smoked but later reports a consumption of 20 cigarettes daily. Missing values can also occur if the interviewer fails to ask all questions, as may happen if the interview is interrupted. Parts of the questionnaire may not be readable or may be destroyed during the process of editing. If data are abstracted from records, these records may be incomplete, illegible, or simply missing. Operating procedures in some departments of an industrial setting or a hospital may require records to be destroyed. In many situations, records include gaps or insufficient or uninterpretable

information, resulting in missing values. Similarly, measures based on chemical or physical procedures may fail to produce a value because required amounts of blood or tissue are not available, or because of a laboratory accident or technical failure. In all these cases the missing values are unplanned, and we usually have limited information on why the data are missing. This lack of information on the mechanism of missingness makes this type of missing value problematic during an analysis.

Planned Missing Values

Because epidemiologic studies may require thecollection of data on many variables for many subjects some sampling strategies have been developed that require less data. A two-stage design (*see* **Case–Control Study, Two-phase**) may be performed in which first-stage data on the disease and crude exposure status are collected for many subjects, but additional information on detailed exposure or on confounding variables is collected only for a subsample in a second stage. The second stage may include equal numbers of crudely exposed and unexposed subjects. In a two-stage design, a large amount of data can be missing, but the missingness is understood and under the control of the investigator. The probability that a value is missing is known or can be calculated easily and can be used for the analysis. Simple and efficient procedures to estimate exposure effects for such designs were proposed by White [60] already in 1982. Closely related ideas of planned missingness are found in **validation studies** in which an easy-to-measure surrogate variable is collected for all subjects, and "**gold standard**" measurements are made only for a subsample.

Missing Value Mechanisms

Whenever we want to handle a data set with missing values appropriately, the probability law generating the missing values will be of importance. Formally, this law, usually called the missing value mechanism, is the conditional distribution of the missing indicators, given all variables considered. To facilitate the discussion, we introduce some notation and consider here the situation with one exposure and one **confounder** variable, where only the confounder variable may be missing. Hence we consider for each subject four variables: the disease status D, the exposure E, the confounder C, and the response indicator R, such that we actually observe C if and only if $R = 1$. This situation is complex enough to explain most problems and the basic approach to solutions. Some solutions, however, do not generalize to settings with several exposures and/or confounders, especially in the case of arbitrary missing patterns; we will point this out where it is necessary. Also, one can exchange the role of E and C.

Now, the missing value mechanism is given by the **conditional probabilities** of observing C, that is, by

$$q(d, e, c) := \Pr(R = 1 | D = d, E = e, C = c).$$

To understand the possible dependencies of the observability of C on D, E and C, we shall discuss some specific situations. In case–control studies, missingness often depends on the disease status, as cases and controls may differ in their behavior and willingness to participate in the investigation and to respond to specific questions. For example, Schlehofer et al. [49] report results of a case–control study on risk factors for brain tumor, including blood group among other factors. For controls, only interview data were available, but for cases hospital records could be used in addition. This results in missing rates of 9% for cases, but of 46% for controls. By contrast, in a prospective cohort study one can usually exclude a dependence of the response probabilities on the disease status, if all **covariate** data are collected at the start of the study. Retrospective cohort studies (*see* **Cohort Study, Historical**) and most hybrid designs, such as **nested case–control studies**, often exhibit a dependence of missingness probabilities on the disease status.

Also, the exposure variable may have an influence on observability of the confounder. In an investigation of the risk of radiation therapy, a given therapy may be associated with hospital records containing detailed information on potential confounders. In studies of exposure in nuclear plants, higher exposure levels may be associated with frequent medical examinations and increased chance to assess information on confounders.

There exist a variety of settings in which the probability of observing a variable depends on the value of the variable itself. In interviews or studies by questionnaire, heavy drinkers or smokers may refuse to

answer questions about such behavior, and very poor or very rich people may refuse to give information on their income. Likewise, long-term unemployed subjects may refuse to give information on their working history. Often the value of a variable may influence the probability of knowing or remembering it. For example, if we ask subjects to recall whether there is any case of a disease among their first and second degree relatives, and if there is no such case, he or she will often answer "I don't know", because he or she does not know all the relatives. But if there is one case, it suffices to know this one to give an answer. Even "objective" sources like hospital records do not guarantee that there is no dependence on the true value. In looking for exposure to a specific therapy, it is often easy to detect such a treatment if it has been given, but to assert that the treatment has not been given requires a complete search of hospital records over the time period of interest.

In epidemiology we may often have rather complicated missing mechanisms. For example, in case–control studies, cases may refuse more often to admit an unhealthy lifestyle than controls, because they feel guilty. On the other hand, they may remember previous exposures better because they have sought reasons for their illness. Similarly, the willingness to admit to specific sexual behaviors may differ among sex and age groups. As another example, the availability of information on confounder variables may depend both on the disease status and the exposure level. If exposed subjects and cases are willing subjects, only unexposed controls may yield missing values. These possible interactions make handling of incomplete data especially difficult.

So far, we have described possible scenarios. Some of them are more dangerous than others, however, depending on the type of analysis. If one wants to make efficient use of subjects with incomplete confounder information, the missing at random (MAR) assumption is of central importance. In our context, the MAR assumption is

$$q(d, e, c) = q(d, e),$$

namely that the true value of C is conditionally independent of R given D and E. This assumption allows one to estimate the conditional distribution of C, given D, E and $R = 0$ from those subjects with $R = 1$, which is the key to efficient use of all data. Note that the MAR assumption allows a dependence of the occurrence of missing values on D and E. In two-stage designs we can exclude a dependence on C by design, but sampling fractions typically depend on D and E. In the literature on missing values, one sometimes finds the missing completely at random (MCAR) assumption, $q(d, e, c) = q$, but this is seldom realistic in epidemiology.

If one wants to ignore the subjects with incomplete covariate data in the analysis, it is essential to assume that the selection of such subjects introduces no **bias**, which leads to different requirements as discussed later. We should finally mention that in a case–control study the definition of $q(d, c, e)$ refers to the selected subjects, but it coincides with the values in the total population, provided that selection probabilities really depend only on the case–control status, and not on other information which is a requirement for any well-conducted case–control study.

Fitting Logistic Regression Models with Incomplete Covariate Data

For epidemiologic investigations, **logistic regression** is an important tool to analyze the joint effect of one or several exposure variables on the disease risk adjusted for one or several confounding variables. In the case of one exposure and one confounder variable, the logistic model for risk in the underlying population assumes that the conditional probability of disease given the exposure value e and the confounder value c is given as

$$\begin{aligned}\Pr(D = 1 | E = e, C = c) &= \Lambda(\beta_0 + \beta_E e + \beta_C c) \\ &=: p_\beta(e, c),\end{aligned}$$

with $\Lambda(t) = 1/[1 + \exp(-t)]$. As suggested by this formula, E and C may be **binary** or continuous variables, extensions to **polytomous** variables are straightforward, and most statements in this article are valid for any type of covariates.

With complete data we can estimate the parameters β_0, β_E, and β_C by the **maximum likelihood** principle. There are several proposals of different quality to cope with incomplete data. To understand the behavior of most simple methods for handling incomplete covariate data, we examine the conditional probabilities of the disease status given the

actual information we observed. Considering only subjects in the cohort with complete data, we have

$$\Pr(D = 1|E = e, C = c, R = 1) = \Lambda\left(\beta_0 + \log\frac{q(1, e, c)}{q(0, e, c)} + \beta_E e + \beta_C c\right), \quad (1)$$

which can be derived by analogy with the justification of logistic regression models for case-control data, as given by Breslow & Day [5, p. 203]. This result follows by noting that $q(d, e, c)$ are nothing other than the probabilities of selecting these subjects, just as cases and controls have selection probabilities from the base population in a case-control study. Eq. (1) implies that fitting a logistic regression model to the subjects with complete data only will give valid estimates for β_E and β_C, provided $q(d, e, c)$ can be decomposed into $q(d) \cdot q(e, c)$. For subjects with a missing confounder value we have

$$\Pr(D = 1|E = e, R = 0) = \int \Lambda\left(\beta_0 + \log\frac{1 - q(1, e, c)}{1 - q(0, e, c)} + \beta_E e + \beta_C c\right) \mathrm{d}F(c|E = e, R = 0). \quad (2)$$

Most simple methods to handle incomplete covariate data try to approximate (1) and (2) by simple logistic models, and the resulting **misspecification** can cause serious bias. In contrast, methods relying on the **likelihood** or on appropriately chosen **estimating** equations have the potential to produce **consistent** estimates. Hence we now consider the likelihood in the incomplete data case. From the joint distribution of the observed variables, subjects without a missing value contribute

$$q(d, e, c) \times p_\beta(e, c)^d \times [1 - p_\beta(e, c)]^{1-d} \times \Pr(C = c|E = e) \times \Pr(E = e),$$

and subjects with a missing value contribute

$$\int [1 - q(d, e, c)] \times p_\beta(e, c)^d \times [1 - p_\beta(e, c)]^{1-d} \times \Pr(C = c|E = e) \times \Pr(E = e)\,\mathrm{d}c.$$

If the MAR assumption $q(d, e, c) = q(d, e)$ holds, not only $\Pr(E = e)$ but also the terms involving q can be removed from the likelihood. However, the likelihood still depends on $\Pr(C = c|E = e)$; hence the classical maximum likelihood principle requires specifying the distribution of the covariates, at least in part, which is fundamentally unlike the complete data case. Trying to avoid these difficulties leads to **semiparametric** approaches. The likelihood presented above is based on a prospective sampling scheme. In the case of complete data, it is well known that such a likelihood also yields valid estimates of β_E and β_C in the analysis of case–control studies [32]. This is also true for incomplete data, as shown by Carroll et al. [9].

In the following we outline the main simple and sophisticated methods for handling incomplete covariate data.

Complete Case Analysis

In a complete case analysis all subjects with a missing value are omitted from the analysis. The validity of this approach is based on the implicit assumption that the **regression** model for the subjects with complete data is identical to the model for all subjects, i.e. that

$$\Pr(D = 1|E = e, C = c, R = 1) = \Pr(D = 1|E = e, C = c)$$

holds. Using (1), this is true, if $q(d, e, c) = q(e, c)$, i.e. if missing probabilities do not depend on the disease status. This is also intuitively clear; if missing probabilities depend only on the covariate values, restriction to subjects without missing values changes only the population, but not the regression model, whereas missing probabilities depending additionally on the outcome introduce some type of **selection bias**. An isolated difference between the missingness probabilities for cases and controls affects the estimation of the intercept but does not affect the estimation of β_E and β_C; in general consistent estimation of the latter is guaranteed if $q(d, e, c) = q(d) \cdot q(e, c)$, which follows directly from (1) [19].

Therefore a complete case analysis has the favorable property that it yields consistent estimates of the regression parameters, even if the MAR assumption is violated. It has the unfavorable property that consistency of parameter estimates depends on the assumption that missingness probabilities do not depend jointly on the disease status and the covariate values. The latter is however often questionable for case–control studies (cf. final section). The bias of the **odds ratio** based on a complete case analysis

can be easily computed [55], and it can be shown that realistic differences in the missingness probabilities can lead to substantial bias. For example, if exposed cases are better documented than controls and unexposed cases such that the missingness probability for the exposed cases is 10% and 40% for the other groups, then the odds ratio for exposure is overestimated by a factor of 1.5.

Additional Category or Missing Indicator Method

Epidemiologists often work with categorical variables and sometimes define an additional category for missing observations. Such coding suggests that we analyze the data under the implicit assumption that

$$\Pr(D = 1|E = e, C = c, R = 1) = \Lambda(\beta_0 + \beta_E e + \beta_C c)$$

and

$$\Pr(D = 1|E = e, R = 0) = \Lambda(\beta_0 + \beta_E e + \beta^*).$$

Equivalently we can impute for the missing values of C the value 0 and add the missing indicator $M = 1 - R$ to the regression model. This "missing indicator method", which is also applied to continuous covariates, results in the same specification and hence the same estimates. The approach is inappropriate, as one cannot expect to achieve good estimates for the adjusted risk β_E if adjustment for the unobserved values of the confounding variable is attempted by introducing the additional parameter β^* in the second equation above. To see this, let us assume that $q(d, e, c) \equiv q$, i.e. MCAR, such that the subjects with and without missing values form two random subsamples. Then in the first equation above β_E corresponds to the adjusted log-odds ratio (OR) of the exposure, whereas in the second line β_E corresponds to the unadjusted log-OR, because $\beta_0 + \beta^*$ can be regarded as one intercept. Consequently, the quantity $\exp(\hat{\beta}_E)$ estimates a quantity between the adjusted and unadjusted odds ratio. Hence the goal of obtaining realistic odds ratios that describe the effect of exposure adjusted for confounding variables cannot be achieved if missing values in the confounding variables are regarded as an additional category. Moreover, if the missingness probabilities are allowed to depend on the disease status and/or exposure status, then $\exp(\hat{\beta}_E)$ can lead to values outside the range between the adjusted and unadjusted odds ratio. The bias is often accompanied by underestimation of the variability; Greenland & Finkle [20] report the results of a **simulation** study with two Gaussian (*see* **Normal Distribution**) covariates, where the missing indicator method results in true coverage probabilities of 55% for nominal 95% **confidence intervals**.

So far we have considered the effect of coding missing values as an additional category on the estimation of β_E. In the epidemiologic literature the estimate of β^* is often reported, too, and compared to the value of $\hat{\beta}_C$. Often there is an implicit assumption that $\hat{\beta}^*$ has to be between 0 and $\hat{\beta}_C$, or, in the case of several categories, within the range of the effect estimates (including 0 for the baseline category). If missing probabilities depend only on the exposure, and the degree of **correlation** between confounder and exposure is small, then this is approximately true, as can be shown using the approximation discussed in the next section. However, if missingness probabilities depend on the disease status, the relative disease frequency among subjects with complete data differs from the relative disease frequency among subjects with incomplete data, and β^* mainly reflects this difference.

Although regarding missing values as an additional category cannot be recommended in general, it can be appropriate in special settings, where missing values characterize a meaningful subset of all individuals. For example, Commenges et al. [11] report a study comparing different procedures to diagnose dementia in a screening setting. They found missing values in those variables corresponding to the results of two tests to be highly predictive, because the missing values reflected a subject's failure to comprehend the test.

Single-Imputation Methods

This class of methods is characterized by imputing for each missing value a single value and analyzing the completed data set. If the confounder C is continuous, the simplest choice is to replace each missing value by the overall mean $\overline{C}$ of the observed values of the confounding variable. Instead of using an estimate for the overall expectation of C, one may use estimates of the conditional expectations: if E is categorical, then we can impute the mean of the observed values of C within each category of E; if E is continuous, then

we can compute a **regression** of the observed values of C on E. If C is binary, then relative frequencies replace the means, and Schemper & Smith [46] proposed the term probability imputation. The imputation of estimates for the conditional expectations yields an approximately valid **inference**, if missing probabilities do not depend on the disease state and the true, unobserved value, i.e. if $q(d, e, c) = q(e)$. In this situation, we have

$$\text{by (1)} \quad \Pr(D = 1|E = e, C = c, R = 1) = p_\beta(e, c)$$

and

$$\text{by (2)} \quad \Pr(D = 1|E = e, R = 0) = \int \Lambda(\beta_0 + \beta_E e + \beta_C c)\, \mathrm{d}F^{C|E=e}(c).$$

If we regard Λ as an approximately linear function, then we have

$$\Pr(D = 1|E = e, R = 0) \approx \Lambda(\beta_0 + \beta_E e + \beta_C \cdot E[C|E = e]).$$

Hence imputing estimates for the conditional expectation results in an approximately correct specification of the conditional disease probabilities, and hence the resulting bias of the parameter estimates is often small. One has to expect, however, that **variance** estimates tend to be too small, because the imputed values are treated as true ones and no adjustment is made for the additional variability introduced by imputing estimates. Results of simulation studies [45, 46, 53, 58] suggest that both bias and underestimation of the variance are only problematic for extreme parameter constellations with high missingness rates and very influential confounding variables.

The justification so far depends on the assumption that missingness probabilities do not depend on the disease status. This is not necessary, because imputation of conditional expectations can always be regarded as an approximation to simple semiparametric approaches [58]. However, some care is necessary; if missingness probabilities depend on the disease status, then naive estimates for conditional expectations are wrong; it is necessary to estimate the conditional expectations separately within diseased and undiseased subjects and then to form a weighted average [58]. Moreover, for extreme parameter constellations the bias can be still substantial [53].

Generalizations to several covariates with arbitrary missing patterns are straightforward, as long as there are enough subjects with complete information. But many auxiliary regression models may be required. In general, misspecification of these auxiliary regression models can be a source of additional bias in the parameter estimates, but little is known about this problem.

Modifying the Complete Case Estimates

Under the MAR assumption, the response probabilities $q(d, e)$ can be estimated from the observed data, for example by fitting a logistic regression model with outcome variable R and covariates D and E. The bias of the complete case estimates can be expressed as a function of q, and hence we can correct the bias [53, 55]. Alternatively, one may fit a logistic regression model with estimated offsets in (1) to the subjects with complete covariate data [4]. If E is categoric and a saturated model (*see* **Generalized Linear Model**) is used in estimating q, then both approaches coincide and are identical to maximum likelihood estimates [57]. As simple expressions for the corresponding asymptotic variances can be provided [7], this is a simple method to achieve **consistent** and **efficient estimates** in this special setting if the MAR assumption is tenable. Unfortunately there is no simple generalization for arbitrary missingness patterns.

Estimation of the Score Function: Weighting, Filling, and the Mean Score Method

In the complete data case, maximization of the likelihood is equivalent to finding a root of the **score** function

$$S_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} S_\beta(D_i, E_i, C_i),$$

with

$$S_\beta(d, e, c) = \frac{\partial}{\partial \beta} \{d \log p_\beta(e, c) + (1 - d) \log(1 - p_\beta(e, c))\}.$$

In the incomplete data case the contribution to the score function is unknown for subjects with a missing value. Nevertheless, one can try to estimate $S_n(\beta)$. A

first approach is to regard the subjects with complete covariate information as a subsample with selection probabilities $q(d, e, c)$ and to try to estimate the "population average" $ES_\beta(D, E, C)$. The classical **Horvitz–Thompson estimator** satisfies this task by weighting each contribution of the subsample with $q(d, e, c)^{-1}$. However, $q(d, e, c)$ is unknown, and only under the MAR assumption can we arrive at estimates $\hat{q}(d, e)$ and at a weighted score function

$$\tilde{S}_n(\beta) = \frac{1}{n} \sum_{\substack{i=1 \\ R_i=1}}^{n} \frac{S_\beta(D_i, E_i, C_i)}{\hat{q}(D_i, E_i)}.$$

Solving $\tilde{S}_n(\beta) = 0$ results in consistent estimates of β. Solving $\tilde{S}_n(\beta) = 0$ can be done by any software package for logistic regression that allows arbitrary weights (*see* **Software, Biostatistical**). However, variance estimates obtained this way are invalid, and can be much too small [53, Section 5.11]. If a parametric model $q_\alpha(d, e)$ is used in estimating the response probabilities, explicit estimates of the variance can be provided [33, [53], p. 17], but they cannot be computed with standard software. If E and C are both categorical, then the approach is equivalent to distributing subjects with a missing value to the cells of the **contingency table** of subjects without a missing value with fractions equal to estimates of the conditional probability for the true value. This intuitive method was called "filling" by Vach & Blettner [55]. The idea to weight contributions to the score function reciprocally to the response probabilities was also used by Flanders & Greenland [15] and Zhao & Lipsitz [61]. However, they consider the analysis of designs for which the response probabilities were known.

An alternative approach to estimating S_β is to replace each unknown contribution $S_\beta(D_i, E_i, C_i)$ for subjects with unknown C_i by an estimate for $\mathrm{E}[S_\beta(D_i, E_i, C_i)|D_i, E_i]$, i.e. an estimate for the conditional expectation of the score function given the observed variables. Reilly & Pepe [34] investigate this approach in detail for the special case where E is categorical. In that case, estimates of the conditional expectations are simple averages over the subjects without missing values, and the approach is equivalent to weighting. However, whereas the weighting approach is difficult to generalize to the case of several covariates with arbitrary missingness patterns, this is in principle possible for the individual estimation of the conditional expectations by **nonparametric regression**.

Finally, estimates based on the weighting or the mean score approach are consistent under the MAR assumption but not always efficient. Especially if missingness rates are large, there can be a substantial loss in comparison to efficient approaches [38; 53, Section 5.2; 61].

Maximum Likelihood Estimation

Application of the maximum likelihood (ML) principle requires a parametric specification $f_\alpha(c|e)$ for the conditional distributions $\Pr(C = c|E = e)$ (cf. above). Then under the MAR assumption the contributions to the likelihood are given by

$$p_\beta(e, c)^d (1 - p_\beta(e, c))^{1-d} f_\alpha(c|e), \text{if } R = 1,$$
$$\int p_\beta(e, c)(1 - p_\beta(e, c))^{1-d} f_\alpha(c|e)\, \mathrm{d}c, \text{if } R = 0.$$

The integral in the likelihood makes maximization a little bit cumbersome. The **EM algorithm** [12] is a standard tool to maximize the likelihood in incomplete data problems. However, if C is continuous, even the EM algorithm may require numerical integration. If C is categorical, integration reduces to summation, and both the EM algorithm [24] or a direct maximization using the Newton–Raphson method are feasible. The latter has the advantage of automatically computing the quantities necessary to estimate the variance of the parameter estimates, whereas use of the EM algorithm requires additional effort [30, 52]. The ML principle is also applicable in the general setting with several covariates and arbitrary missingness patterns, as long as we are able to specify a parametric family for the conditional distribution of the covariates affected by missing values given the unaffected covariates.

The ML estimates are consistent and efficient as long as the MAR assumption is valid and the true distribution of the covariates is within the specified family. This specification is a crucial point of the ML approach, because this requirement is not necessary in the complete data case, and our knowledge about the distributions of and dependencies between the covariates is usually limited. Misspecification of the distribution of the covariates, however, can induce a bias in the regression parameter estimates. Thus it becomes necessary to model nuisance features of

the problem carefully. If all covariates are categorical, loglinear models provide a simple framework to describe the joint distribution [56], but if continuous covariates are involved, parametric models flexible enough seem to be hard to specify.

If all covariates are categorical, one can also fit a **loglinear model** to the joint distribution of all variables [16, 59] and use relationships between loglinear and logistic models.

Semiparametric Maximum Likelihood Estimation

We have seen in the last section that maximum likelihood estimation requires specification of a parametric family for the conditional distribution of C given E. It is an appealing idea to avoid this unpleasant task by replacing $f(c|e)$ by a **nonparametric** estimate. Pepe & Fleming [31] consider the case of a categorical exposure, such that the empirical distribution within each exposure stratum can be used; Carroll & Wand [8] consider a continuous exposure and use kernel estimates (*see* **Density Estimation**). Both approaches rely on the assumption that missingness probabilities do not depend on the disease status, but they can be generalized to this setting (Vach & Schumacher [58]). Computations of the resulting estimates of β require special software, as does estimation of the variance. The resulting estimates are not fully efficient in comparison to the estimates of the next section. It is also difficult to generalize these approaches to settings with several covariates and arbitrary missingness patterns, because this requires nonparametric estimation of high-dimensional **multivariate** conditional distributions.

Semiparametric Efficient Estimation

The last two sections have suggested that the handling of incomplete covariate data is ideally a **semiparametric** problem; we are interested in the parameters of the regression model describing the conditional distribution of disease status given all exposure and confounding covariates, but the distribution of the covariates, in spite of being essential for the likelihood, should be left unspecified. In recent years there has been substantial progress (for example [3]) in the general field of efficient semiparametric estimation. Robins et al. [38] used this theory to fit generalized linear models with incomplete covariate data. They showed that roughly any consistent estimator for β is asymptotically equivalent to one defined as the solution of an estimating equation $\sum_{i=1}^{n} S_\beta(D_i, E_i, C_i) = 0$, where

$$S_\beta(D, E, C) = R\frac{h(E, C)(D - p_\beta(E, C))}{q(D, E)} - \frac{\varphi(D, E)(R - q(D, E))}{q(D, E)}.$$

They were also able to characterize functions h_{opt} and φ_{opt} which lead to a semiparametric efficient estimate, i.e. the asymptotic variance of this estimate is exactly the supremum of the asymptotic variances of all maximum likelihood estimators based on parametric families $f_\alpha(c|e)$ covering the true $f(c|e)$. Of course, this is the best we can expect without imposing parametric assumptions. Unfortunately h_{opt} and φ_{opt} depend on the true values of β and the true distribution of C given E and are moreover not available in closed form.

However, an adaptive procedure is possible which starts with a parametric assumption on the distribution of the covariates, then estimates all parameters, uses an iterative procedure to compute $\hat{h}_{\text{opt}}$ and $\hat{\varphi}_{\text{opt}}$ based on the assumption that the estimates correspond to the true parameters, and finally solves the estimating equations with h and φ replaced by $\hat{h}_{\text{opt}}$ and $\hat{\varphi}_{\text{opt}}$, and q replaced by an appropriate estimate. In contrast to ML estimation, a misspecification of the covariate distribution does not result in inconsistent estimates, and, in spite of the adaptive steps, the estimates are efficient, if the specification of the covariate distribution was correct. Details of this adaptive procedure can be found in Robins et al. [38] and Rotnitzky & Robins [40]. The approach can be also generalized to several covariates with arbitrary missingness patterns; however, here the computation of $\hat{h}_{\text{opt}}$ and $\hat{\phi}_{\text{opt}}$ is more difficult.

Multiple Imputation

Multiple imputation is a general technique for statistical inference with incomplete data. The basic idea is to create several data sets with different values imputed for the missing values, and to analyze each data set by standard software, such as software for logistic regression. If the imputations are generated in a so-called "proper" manner, the average of the parameter estimates provides a consistent estimate.

Furthermore, the average of the variance estimates and the empirical variance of the multiple parameter estimates can be combined to form a total variance estimate; confidence intervals and ***P* values** can be computed, too. Rubin & Schenker [44] present an overview of the basic techniques.

It seems reasonable to generate imputations from estimates of the conditional distribution of the unobserved values. However, this is an improper method in the sense that variance estimates tend to be too small, because they do not take into account the variance due to estimating the conditional distribution. Proper methods can be defined by additionally estimating the conditional distribution in each imputation step based on a **random sample** with replacement of the subjects without missing values [14, 42, 43]. Of course, any attempt to estimate the conditional distribution of the missing values from the observed values depends on the MAR assumption.

With respect to our setting, Reilly & Pepe [34, 35] have considered the special case where E is categoric. Values to be imputed for missing values in C are drawn from the empirical distributions of C within the strata defined by D and E. This hot-deck imputation method is improper, although Reilly & Pepe [35] provided a valid variance estimator. Moreover they showed that hot-deck multiple imputation with infinite imputations is asymptotically equivalent to the mean-score method. In particular, this implies that the hot-deck method has the same deficiencies with respect to efficiency. Greenland & Finkle [20] report results of a simulation study with E and C both continuous and affected by missing values. Imputations were drawn from estimated conditional distributions resulting from fitting **bivariate normal distributions** within the diseased and undiseased subjects. Although this is an improper method, they observed that confidence intervals keep their nominal level. They also observed a loss of efficiency in comparison to maximum likelihood estimation (*see* **Missing Data Estimation, "Hot Deck" and "Cold Deck"**).

Multiple imputation can be also applied in general settings with arbitrary missingness patterns. The crucial point is the choice of the procedure to estimate the necessary conditional distribution. If we rely on parametric assumptions on the distribution of the covariates, we have the same unpleasant situation as with ML estimation. However, one can alternatively draw imputations from a set of nearest neighbors, i.e. subjects with complete information and similar values with respect to the observed variables. The choice of an appropriate distance measure requires some knowledge about the distribution of the covariates, but not necessarily an explicit model. Heitjan & Little [22] give an illuminating example.

Methods Based on the Retrospective Likelihood

The methods considered so far rely on a prospective sampling scheme implying independence of the disease status among different subjects. In case–control studies this assumption is violated. However, in incomplete data problems the use of the prospective likelihood can also be justified for retrospective data [9]. The resulting estimates are consistent, the estimated **standard errors** are never too small, and they are correct if we make no assumptions on the distribution of the covariates. Nevertheless, methods based on the retrospective likelihood are of interest, especially for the analysis of two-stage designs. In such a design, the number of subjects with complete data is fixed in advance, and hence missingness indicators are not independent, which is a further violation of the prospective sampling scheme.

Maximum likelihood estimation with respect to the retrospective likelihood is considered by Scott & Wild [51] and Breslow & Holubkov [6]. Two different **pseudo-likelihood** approaches, in which some parameters are pre-estimated in a naive manner, are considered by Breslow & Cain [4] and Schill et al. [48]. A weighting approach is due to Flanders & Greenland [15]. Comparisons with respect to the **asymptotic relative efficiency** and simulation studies [6, 47, 61] often reveal large inefficiencies of the weighting approach and some inefficiencies of the two pseudo-maximum likelihood approaches.

Handling of a Questionable MAR Assumption

All sophisticated, and especially all efficient, approaches to handle incomplete covariate data rely on the MAR assumption. In many applications this assumption is questionable, but one may still want to use methods relying on it. In that case, it is necessary to think about or investigate the possible impact of a violation. One may argue that if there is a pure violation, in the sense that missingness depends only

on the true value of the covariate, then the impact must be small, because the association between the covariates and the outcome is not changed. Schemper & Smith [46] provide an informal argument for this conjecture. Investigations for the special case of categorical C and E [57] corroborate the conjecture. These studies further demonstrate that the impact on the exposure effect estimate can be substantial if there are differences in the degree of violation between diseased and undiseased or between exposed and unexposed subjects, which is also intuitively clear, because such differences change the observed association.

If one does not want to rely on such general, theoretical considerations, one may try to investigate the impact of an invalid MAR assumption for a particular data set. This can be easily done within the multiple imputation framework, for example by drawing more larger values for a variable or more values from a specific category (cf. [44]). Vach & Blettner [56] present a framework to specify violations within the framework of ML estimation and perform a **sensitivity analysis** for two case–control studies. Baker [2] takes an additional step and does not specify, but tries to estimate, the parameters of the non-MAR mechanism. Rotnitzky & Robins [40] consider this step within the framework of semiparametric efficient estimation. However, a (saturated) logistic model and a (saturated) non-MAR model are in general not jointly **identifiable**; hence any attempt to estimate non-MAR mechanisms relies on restrictions of the two models allowing identifiability. This alone, however, is not enough, as identifiability does not imply reasonable properties of resulting estimates in this setting; Rotnitzky & Robins [40] show in the semiparametric setting that in spite of identifiability there need not exist a $\sqrt{n}$-consistent estimator. Hence, the usefulness of these approaches has to be investigated further before recommendations can be made.

Robins & Gill [37] point out that in settings with arbitrary missingness patterns, the MAR assumption as defined by Rubin [41] allows some configurations of no practical relevance. This fact can be used to change the MAR assumption, allowing some special non-MAR mechanisms to be estimated without problems of identifiability. Robins & Gill [37] and Robins [36] present two examples of this kind.

Handling of Incomplete Data in other Models Used in Analytic Epidemiology

Poisson Regression, Gaussian Regression, and Generalized Linear Models

Nearly everything we have said in the last section with respect to logistic regression is also valid for other regression models where parameters are estimated by maximum likelihood. In particular, the difficulties with maximum likelihood estimation in the incomplete data case are the same, and the semiparametric approaches work in the general setting of generalized linear models. With respect to the simple methods, there are two differences. First, there is no general analogy to the modifications of the complete case estimates. Second, the single imputation methods need more care. We can expect nearly unbiased estimates of the regression parameters after imputation of conditional means, as this implies a roughly correct specification of the conditional expectation of the outcome variable. Indeed, in the case of Gaussian regression one can prove consistency [18]. However, only in binary regression models does correct specification of the conditional mean imply correct specification of the conditional variance. In general, the conditional variance of the outcome increases if some covariate values are missing; hence, after the imputation of conditional means, a further analysis should be based on a heteroscedastic model. For this reason, weighted **least squares** estimates are advocated in Gaussian regression after imputation of conditional means. An overview of this and other techniques suitable for Gaussian regression models is given by Little [27]. Some of the proposals depend on the assumption of a **multivariate normal distribution** of all variables and hence have limited application in epidemiology. The impact of the variance heterogeneity for other types of regression models, especially **Poisson regression**, has not been investigated. Thus we recommend that the single imputation method should be used with caution.

Cox Regression with Incomplete Covariate Data

For the analysis of (censored) survival times the **proportional hazard** model [10] is widely used in epidemiology. Simple methods to handle incomplete covariate data are subject to the same criticism as for logistic regression, with the additional difficulty

that, especially in **retrospective studies**, censoring may be associated with missingness in covariates. Even in a complete case analysis, the assumption of noninformative **censoring** can be violated. With respect to more sophisticated approaches, it is difficult to extend the maximum likelihood approach to survival models (*see* **Survival Analysis, Overview**), as the **nuisance parameter** involves the baseline hazard, although a semiparametric partial maximum likelihood approach is possible [62]. A weighting approach has been proposed by Pugh et al. [33], and Lin & Ying [26] consider an appropriately modified score function, but their approach requires MCAR. None of these approaches can be easily generalized to situations with general missingness patterns. Robins et al. [38] also point out the difficulty of obtaining a feasible solution from the theory of semiparametric efficient estimation. In the face of this problem, one may be willing to use alternative fully parametric regression models for survival data, such that, especially in the case of categorical covariates, the ML principle can be used. In this spirit, Schluchter & Jackson [50], Baker [1] and Vach [54] suggest approximating the **Cox regression model** by a logistic model for grouped survival data, and Lipsitz & Ibrahim [29] consider **Weibull** models. The use of single imputation methods has been considered by Schemper & Smith [46].

Analysis of Matched Case–Control Studies

The handling of incomplete covariate data in matched case–control studies has received little attention. Haber & Chen [21] consider the case of a single exposure variable as the only covariate and compare the matched and unmatched odds ratio estimator. They conclude that in the case of missing exposure information for some cases and controls, the advantages of the unmatched estimator increase in comparison to the complete data case. **Conditional logistic regression** (*see* **Matched Analysis**) is a standard tool for the analysis of matched case–control studies. Missing values in the covariates constitute an even greater problem with **conditional logistic regression** than with ordinary logistic regression, as a complete case analysis with one-to-one-matching causes loss of the complete pair if the covariate is missing in either the case or the control. Despite a small simulation study [17], a systematic investigation is still needed.

Regression Models for Longitudinal and Multivariate Data

Regression models for longitudinal or clustered data (*see* **Clustering**), especially **marginal models**, are proving useful in epidemiology for the analysis of familial aggregation and of environmental studies (*see* **Environmental Epidemiology**) as well as for studies of biochemical markers. With respect to incomplete covariate data, there is little to add to what we have said previously. However, in these applications outcome variables may also be missing, especially from drop outs in longitudinal studies. We want to restrict ourselves to some basic comments, especially on the differences with the incomplete covariate problem.

First, the MAR assumption is again of central importance. In the case of drop outs, the question is whether we are able to observe the crucial event causing the drop out, or whether the drop out hides this event. Secondly, if the MAR assumption is tenable and if we consider regression models specifying the joint conditional distribution of the outcome variables and allowing the use of the ML principle with complete data case, then the ML principle can also be used in the presence of missing values in the outcome variables and reduces usually to an analysis of all units with measured outcome. Thirdly, the popular marginal models [25] do not belong to this class, and for them the MAR assumption is not sufficient to exclude a bias due to missing values, if only the available units are used; a solution has been provided by Robins et al. [39]. Finally, if the MAR assumption is violated, we have often some rather precise ideas on the drop out mechanism, which may permit adjustment by choosing an appropriate model [13, 23, 28].

Strategies to Cope with Incomplete Data

The best advice is to minimize the possibility for missing values. We should plan appropriate data collection procedures and design interviews and questionnaires so that subjects have little reason to refuse an answer. Adequate planning can also help to avoid differential missingness with respect to disease status or exposure status. The same data collection procedures should be used for cases and controls in case–control studies, and exposed and unexposed subjects should be followed using similar procedures

and effort in cohort studies. Usually one knows in advance which variables are most likely to have missing values. Then a fruitful strategy can be to collect data on a surrogate variable that is available on most subjects and to collect the variable of interest with additional effort only in a randomly selected subsample, assuring that the MAR assumption holds. Then it is possible to use statistical methods very similar to the sophisticated methods discussed earlier, except that the surrogate variable is not included in the regression model (*see* **Validation Study**). A general idea is to collect additional data to predict missingness. By incorporating such variables in the analysis, the MAR assumption may become more reliable. Finally, one can try to recontact a representative sample of the nonresponders, and try to collect the missing data. If this succeeds, a valid analysis becomes possible in principle.

If these approaches are infeasible or unsuccessful, then one should at least discuss the possible impact of the missing values on the analysis. The first step is to report the missing rates for all variables, stratified by disease status and exposure levels, and to summarize major associations of missingness with other variables. The second step is to justify the analytical approach. If a complete case analysis is applied in a case–control study, then one should give arguments to exclude an important difference in the missing value mechanism between cases and controls. If one uses methods relying on the MAR assumption, then the latter must be justified or a sensitivity analysis should be conducted.

Conclusions

Missing values are a common problem in the analysis of epidemiologic studies. The problem should be addressed in planning the study so as to minimize their occurrence. Careful planning may also allow one to control or to understand the missingness mechanism and thereby to facilitate valid inference. If one has sufficient insight into the missingness mechanism, then one can take advantage of efficient statistical methods, although there remains a need for more practical experience with these techniques and improved availability of software. Such analytical methods cannot salvage a poorly planned and executed study, however, that has many missing values and offers little insight into the missingness mechanism.

References

[1] Baker, S.G. (1994). Regression analysis of grouped survival data with incomplete covariates: Nonignorable missing-data and censoring mechanisms, *Biometrics* **50**, 821–826.

[2] Baker, S.G. (1996). Reader reaction: The analysis of categorical case–control data subject to nonignorable nonresponse, *Biometrics* **52**, 362–369.

[3] Bickel, P.J., Klaassen, C.A., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. John Hopkins University Press, Baltimore.

[4] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case–control data, *Biometrika* **75**, 11–20.

[5] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1: *The Analysis of Case–Control Studies*. IARC Scientific Publications, Lyon, No. 32.

[6] Breslow, N.E. & Holubkov, R. (1997). Weighted likelihood, pseudolikelihood and maximum likelihood methods for logistic regression two-stage data, *Statistics in Medicine* **16**, 103–116.

[7] Cain, K.C. & Breslow, N.E. (1988). Logistic regression analysis and efficient design for two-stage studies, *American Journal of Epidemiology* **128**, 1198–1206.

[8] Carroll, R.J. & Wand, M.P. (1991). Semiparametric estimation in logistic measurement error models, *Journal of the Royal Statistical Society, Series B* **53**, 573–585.

[9] Carroll, R.J., Wang, S. & Wang, C.Y. (1995). Prospective analysis of logistic case–control studies, *Journal of the American Statistical Association* **90**, 157–169.

[10] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[11] Commenges, D., Gagnon M., Letenneur, L., Dartigues, J.F., Barbarger-Gateau, P. & Salamon R. (1992). Improving screening for dementia in the elderly using mini-mental state examination subscores, Benton's visual retention test, and Isaacs' set test, *Epidemiology* **3**, 185–188.

[12] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[13] Diggle, P. & Kenward, M.G. (1994). Informative dropout in longitudinal data analysis, *Applied Statistics* **43**, 49–93.

[14] Efron, B. (1994). Missing data, imputation and the bootstrap (with discussion), *Journal of the American Statistical Association* **89**, 463–479.

[15] Flanders, W.D. & Greenland, S. (1991). Analytical methods for two-stage case–control studies and other stratified designs, *Statistics in Medicine* **10**, 739–747.

[16] Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing

data, *Journal of the American Statistical Association* **77**, 270–278.

[17] Gibbons, L.E. & Hosmer, D.W. (1991). Conditional logistic regression with missing data, *Communications in Statistics – Simulation and Computation* **20**, 109–119.

[18] Gill, R.D. (1986). A note on some methods for regression analysis with incomplete observations, *Sankhyā, Series B* **48**, 19–30.

[19] Glynn, R.J. & Laird, N.M. (1983). Regression estimates and missing data: Complete case analysis. *Unpublished manuscript*, Department of Biostatistics, Harvard University.

[20] Greenland, S. & Finkle, W.D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analysis, *American Journal of Epidemiology* **142**, 1255–1264.

[21] Haber, M. & Chen, C.C.H. (1991). Estimation of odds ratios from matched case–control studies with incomplete data, *Biometrical Journal* **33**, 673–682.

[22] Heitjan, D.F. & Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System, *Applied Statistics* **40**, 13–29.

[23] Hogan, J.W. & Laird, N.M. (1997). Model-based approaches to analyzing incomplete longitudinal and failure time data, *Statistics in Medicine* **16**, 259–284.

[24] Ibrahim, J.G. (1990). Incomplete data in generalized linear models, *Journal of the American Statistical Association* **85**, 765–769.

[25] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[26] Lin, D.Y. & Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association* **88**, 1341–1349.

[27] Little, R.J.A. (1992). Regression with missing X's: A review, *Journal of the American Statistical Association* **87**, 1227–1237.

[28] Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**, 1112–1121.

[29] Lipsitz, S.R. & Ibrahim, J.G. (1996). Using the EM-algorithm for survival data with incomplete categorical covariates, *Lifetime Data Analysis* **2**, 5–14.

[30] Louis, T.A. (1982). Finding the observed information when using the EM algorithm, *Journal of the Royal Statistical Society, Series B* **44**, 226–233.

[31] Pepe, M.S. & Fleming, T.R. (1991). A nonparametric method for dealing with missing covariate data, *Journal of the American Statistical Association* **86**, 108–113.

[32] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case–control studies, *Biometrika* **66**, 403–412.

[33] Pugh, M., Robins, J., Lipsitz, S. & Harrington, D. (1993). Inference in the Cox Proportional Hazards Model with Missing Covariate Data, *Technical Report 758Z*. Division of Biostatistics, Dana-Farber Cancer Institute, Boston.

[34] Reilly, M. & Pepe, M. (1995). A mean score method for missing and auxiliary covariate data in regression models, *Biometrika* **82**, 299–314.

[35] Reilly, M. & Pepe, M. (1997). The relationship between hot-deck multiple imputation and weighted likelihood, *Statistics in Medicine* **16**, 5–19.

[36] Robins, J.M. (1997). Non-response models for the analysis of non-ignorable missing data, *Statistics in Medicine* **16**, 21–37.

[37] Robins, J.M. & Gill, R. (1997). Non-response models for the analysis of non-monotone ignorable missing data, *Statistics in Medicine* **16**, 39–56.

[38] Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* **89**, 846–866.

[39] Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**, 106–121.

[40] Rotnitzky, A. & Robins, J.M. (1997). Analysis of semiparametric regression models with non-ignorable non-response, *Statistics in Medicine* **16**, 81–102.

[41] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

[42] Rubin, D.B. (1981). The Bayesian bootstrap, *Annals of Statistics* **9**, 130–134.

[43] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[44] Rubin, D.B. & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications, *Statistics in Medicine* **10**, 585–598.

[45] Schemper, M. & Heinze, G. (1997). Probability imputation revisited for prognostic factor studies, *Statistics in Medicine* **16**, 73–80.

[46] Schemper, M. & Smith, T.L. (1990). Efficient evaluation of treatment effects in the presence of missing covariate values, *Statistics in Medicine* **9**, 777–784.

[47] Schill, W. & Drescher, K. (1997). Logistic analysis of studies with two-stage sampling: A comparison of four approaches, *Statistics in Medicine* **16**, 117–132.

[48] Schill, W., Jöckel, K.H., Drescher, K. & Timm, J. (1993). Logistic analysis in case–control studies under validation sampling, *Biometrika* **80**, 339–352.

[49] Schlehofer, B., Blettner, M., Becker, N., Martinsohn, C. & Wahrendorf, J. (1992). Medical risk factors and the development of brain tumor, *Cancer* **69**, 2541–2547.

[50] Schluchter, M.D. & Jackson, K.L. (1989). Log-linear analysis of survival data with partially observed covariates, *Journal of the American Statistical Association* **79**, 772–780.

[51] Scott, A.J. & Wild, C.J. (1991). Fitting logistic regression models in stratified case–control studies, *Biometrics* **47**, 497–510.

[52] Tanner, M. (1994). *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer-Verlag, New York.

[53] Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*, Lecture Notes in Statistics 86. Springer-Verlag, New York.

[54] Vach, W. (1997). Some issues in estimating the effect of prognostic factors from incomplete covariate data, *Statistics in Medicine* **16**, 57–72.

[55] Vach, W. & Blettner, M. (1991). Biased estimation of the odds ratio in case–control studies due to the use of ad-hoc methods of correcting for missing values for confounding variables, *American Journal of Epidemiology* **134**, 895–907.

[56] Vach, W. & Blettner, M. (1995). Logistic regression with incompletely observed categorical covariates – Investigating the sensitivity against violation of the missing at random assumption, *Statistics in Medicine* **14**, 1315–1329.

[57] Vach, W. & Illi, S. (1997). Biased estimation of adjusted odds ratios from incomplete covariate data due to violation of the MAR assumption, *Biometrical Journal* **39**, 13–28.

[58] Vach, W. & Schumacher, M. (1993). Logistic regression with incompletely observed categorical covariates – a comparison of three approaches, *Biometrika* **80**, 353–362.

[59] Williamson, G.D. & Haber, M. (1994). Models for three-dimensional contingency tables with completely and partially cross-classified data, *Biometrics* **50**, 194–203.

[60] White, J.E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

[61] Zhao, L.P. & Lipsitz, S. (1992). Designs and analysis of two-stage designs, *Statistics in Medicine* **11**, 769–782.

[62] Zhou, H. & Pepe, M.S. (1995). Auxiliary covariate data in failure time regression, *Biometrika* **82**, 139–149.

(*See also* **Missing Data**)

WERNER VACH & MARIA BLETTNER

Missing Data

This article concerns the analysis of biostatistical data that are subject to missing values. It builds on earlier research [63, 65–67]. Missing values arise in biostatistics for many reasons. For example:

1. In longitudinal studies, data are missing because of *attrition*, i.e. subjects drop out prior to the end of the study (*see* **Longitudinal Data Analysis, Overview**).
2. In sample surveys, some individuals provide no information because of noncontact or refusal to respond (*unit* **nonresponse**). Other individuals are contacted and provide some information, but fail to answer some of the questions (*item* nonresponse). For example, the National Health and Nutrition Examination Survey (NHANES) includes data from an individual interview and a health examination. Some survey respondents miss particular variables because they refused to answer sensitive questions, or measurements were not carried out or were incorrectly recorded, for example they lie outside allowable ranges. Other individuals are missing all the recordings from the health examination since they failed to show up, but have information from the individual interview recorded.
3. Information about a variable is partially recorded. A common example in biostatistics is *right censoring* (*see* **Censored Data**), where times to an event (death, progression of disease) are being recorded, and for some individuals the event has still not taken place when the study is terminated. The times for these subjects are known to be greater than that corresponding to the latest time of observation, but the actual time is unknown. Another example of partial information is **interval censoring**, where it is known that the time to an event lies in an interval. For example, in a longitudinal study of a chronic disease, it may be established that some event (such as reinjury of hip after hip replacement surgery) took place some time between two visits to the doctor for checkups. The time to reinjury is then known to lie in an interval determined by the two checkups. If the interval is narrow compared with the distribution of event times themselves, then the simple approach of locating the event at the midpoint of the interval may be a good approximation, but otherwise methods that treat the event time as partially missing data may be important [13, 19, 38].
4. In clinical studies that involve chart review, charts are often incomplete or lacking in sufficient detail to determine particular items. Often indices are constructed by summing values of particular items, and if any of the items that form the index are missing, then some procedure is needed to deal with the missing data.
5. Missing data can arise by design. For example, suppose one objective in a study of obesity is to estimate the distribution of a measure Y_1 of body fat in the population, and correlate it with other factors. Since Y_1 is expensive to measure, it can only be obtained for a limited sample, but a crude proxy measure Y_2, such as body mass index, can be obtained for a much larger sample. A useful design is to measure Y_2 and **covariates** for a large sample and Y_1, Y_2, and covariates for a smaller subsample. The subsample allows predictions of the missing values of Y_1 to be generated for the larger sample, using one of the methods of analysis described below, yielding more efficient estimates than are possible from the subsample alone.

Unless missing data are deliberately incorporated by design, the most important step in dealing with missing data is to try to avoid it during the data-collection stage. Given that data are likely to be missing after data collection, however, it is also useful to try to collect covariates that are likely to be predictive of the missing values, so that an adequate adjustment can be made. In addition, the process that leads to missing values should be determined during the collection of data if possible. This assists in modeling the missing-data mechanism when an adjustment for the missing values is performed [62].

Three major approaches to the analysis of missing data can be distinguished:

1. discard incomplete cases and analyze the remainder (complete-case analysis):
2. impute or fill in the missing values and then analyze the filled-in data; and
3. analyze the incomplete data by a method that does not require a complete (that is, a rectangular) data set.

With regard to point 3, I focus on powerful **likelihood**-based methods, specifically **maximum likelihood** (ML) and **Bayesian** simulation. The latter is closely related to **multiple imputation** [96], an extension of single imputation that allows uncertainty in the imputations to be reflected appropriately in the analysis. Approaches to longitudinal data with missing values are considered in the concluding section.

A basic assumption in all our methods is that missingness of a particular value hides a true underlying value that is meaningful for analysis. This may seem obvious but is not always the case. For example, consider a longitudinal analysis of CD4 counts in a **clinical trial** for **AIDS** [9]. For subjects who leave the study because they move to a different location, it makes sense to consider the CD4 counts that would have been recorded if they had remained in the study. For subjects who die during the course of the study, it is less clear whether it is reasonable to consider CD4 counts after time of death as missing values. Rather, it may be preferable to treat death as a primary outcome and restrict the analysis of CD4 counts to individuals who are alive. A more complex missing data problem arises when individuals leave the study for unknown reasons, which may include relocation or death.

Pattern and Mechanism of Missing Data

It is useful to distinguish the *pattern* of the missing data and the missing data *mechanism*. The pattern simply defines which values in the data set are observed and which are missing. Specifically, let $Y = y_{ij}$ denote an $n \times p$ rectangular data set without missing values, with ith row $y_i = y_{i1}, \ldots, y_{ip}$, where y_{ij} is the value of variable Y_j for subject i. With missing data, define the *missing-data indicator matrix* $M = m_{ij}$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. The matrix M then defines the pattern of missing data.

When a data set contains missing values, it is important that information is coded so that M can be determined, even if it is not specifically created. This is usually done by designating a special missing-value code for missing values (such as 9999) that lies outside the allowable range for the variable. It is important to distinguish between zero values and missing values, since failure to do this creates considerable problems in analysis.

Some methods for handling missing data apply to any pattern of missing data, whereas other methods assume a special pattern. An important example of a special pattern is *univariate* nonresponse, where missingness is confined to a single variable. Another is *monotone* missing data, where the variables can be arranged so that $Y_{j+1}, \ldots, Y_p$ is missing for all cases where Y_j is missing, for all $j = 1, \ldots, p - 1$ (see Figure 1). This pattern arises commonly in longitudinal data subject to attrition.

The missing-data mechanism concerns the reasons why values are missing, and in particular whether these reasons relate to values in the data set. For example, a subject in a longitudinal study may be more likely to avoid a treatment and drop out of a study because (s)he felt the treatment was ineffective, which might be related to a poor value of an outcome measure. Rubin [93] treated M as a random matrix, and characterized the missing-data mechanism by the conditional distribution of M given Y, say $f(M|Y, \phi)$, where ϕ denotes unknown parameters. If missingness does not depend on the values of the data Y, missing or observed, that is:

$$f(M|Y, \phi) = f(M|\phi), \quad \text{for all } Y, \phi,$$

then the data are called missing completely at random (MCAR) – note that this assumption does not mean that the pattern itself is random, but rather that missingness does not depend on the data values. An MCAR mechanism is plausible in planned missing-data designs as in example 5 above, but is a strong assumption when missing data do not occur by design because missingness usually depends on recorded variables. Let Y_{obs} denote the observed values of Y and Y_{mis} the missing values. A less restrictive

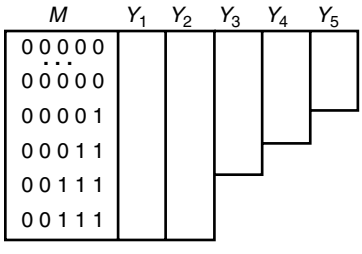

Figure 1 Schematic of a monotone missing data pattern, with rows representing cases, $Y_1, \ldots, Y_5$ repeated measures at five time points, and blocks representing data. M is the missing-data indicator matrix

assumption is that missingness depends only on values Y_{obs} that are observed, and not on values Y_{mis} that are missing. That is:

$$f(M|Y, \phi) = f(M|Y_{\text{obs}}, \phi), \quad \text{for all } Y_{\text{mis}}, \phi.$$

The missing data mechanism is then called missing at random (MAR). Murray & Findlay [80] provided an instructive example of MAR for data from a study of hypertensive drugs where the outcome was diastolic blood pressure. By protocol, the subject was no longer included in the study when the diastolic blood pressure got too large. This mechanism is not MCAR, since it depends on the values of blood pressure. But blood pressure at the time of drop out was observed before the subject dropped out. Hence the mechanism is MAR, because drop out only depends on the observed part of Y. Many methods for handling missing data assume the mechanism is MCAR or MAR, and yield **biased** estimates when the data are not MAR.

Complete-Case Analysis

A common and simple method is complete-case (CC) analysis, also known as *listwise* deletion, where incomplete cases are discarded and standard analysis methods applied to the complete cases. In many statistical packages (*see* **Software, Biostatistical**) this is the default analysis. Valid (but often suboptimal) inferences are obtained when the missing data are MCAR, since then the complete cases are a random subsample of the original sample with respect to all variables. However, even when MCAR holds, the rejection of incomplete cases seems an unnecessary waste of information: if the number of variables is large, then even a sparse pattern of missing values can result in a substantial number of incomplete cases. One approach to incorporating the incomplete cases is to drop variables with high levels of nonresponse; Rubin [92] provides systematic methods in the **regression** context.

Aside from efficiency considerations, a serious problem with dropping incomplete cases is that the complete cases are often a biased sample, i.e., the missing data are not MCAR. The size of the resulting bias depends on the degree of deviation from MCAR, the amount of missing data, and the specifics of the analysis. In particular, the bias in estimating the **mean** of a variable is easily shown to be the difference in the means for complete and incomplete cases multiplied by the fraction of incomplete cases. Thus, the potential for bias increases with the fraction of missing data (*see* **Bias from Nonresponse**). In sample surveys this motivates strenuous attempts to limit unit nonresponse through multiple follow-ups, and surveys with high rates of unit nonresponse (say 30% or more) are often considered unreliable for making inferences to the whole population. For comparisons of means [69] and more generally regression analysis [59], the bias from CC analysis is often smaller. Specifically, it yields valid **inferences** in regression provided the model is correctly specified and missingness depends on the predictor variables, observed or missing, but not on the outcome.

When data are MAR but not MCAR, a useful modification of CC analysis is to assign a nonresponse weight to the respondents to remove or reduce nonresponse bias. In **probability sampling**, a sampling weight inversely proportional to the probability of selection is often used to adjust for differential selection probabilities. If nonresponse is viewed as another stage of probabilistic selection of units, then the product of the probability of selection by design and the probability of response given selection is the probability of being observed, and the inverse of this can be used as a weight in the analysis. Whereas sample design probabilities are known, nonresponse probabilities are unknown and need to be estimated from the data. A standard approach is to form adjustment cells (or subclasses) on the basis of background variables measured for respondents and nonrespondents; for unit nonresponse adjustment these are often based on geographical areas or groupings of similar areas based on aggregate socioeconomic data. All nonrespondents are given zero weight and the nonresponse weight for all respondents in an adjustment cell is then the inverse of the response rate in that cell. If more than one background variable is measured, then adjustment cells can be based on a joint classification, collapsing small cells as necessary. For a health survey application (*see* **Surveys, Health and Morbidity**), see Ezzati & Khare [15]. This method removes the component of nonresponse bias attributable to differential nonresponse rates across the adjustment cells, and eliminates bias if within each adjustment cell respondents can be regarded as a random subsample of the original sample within that cell (i.e. the data are MAR given indicators for the adjustment cells).

A useful alternative approach with more extensive background information is *response propensity* **stratification**, where (i) the indicator for unit nonresponse is regressed on the background variables, using the combined data for respondents and nonrespondents and a method such as **logistic regression** appropriate for a **binary** outcome; (ii) a predicted response probability is computed for each respondent based on the regression in (i); and (iii) adjustment cells are formed on the basis of a categorized version of the predicted response probability. Theory [56, 90] suggests that this is an effective method for removing nonresponse bias attributable to the background variables. For a health survey application, see [30]. Robins et al. [89] and Robins & Rotnitsky [88] apply a similar weighting approach in the more general settings of **generalized estimating equations** for repeated measures analysis (*see* **Longitudinal Data Analysis, Overview**) and **multivariate regression**.

Weighting methods can be useful for removing or reducing nonresponse bias, but they do have serious limitations. First, information in the incomplete cases is still discarded, so the method is inefficient. Weighted estimates can have unacceptably high **variance**, as when outlying values of a variable are given large weights. Secondly, variance estimation for weighted estimates with estimated weights is problematic. Explicit formulas are available for simple estimators such as means under **simple random sampling** [82], but methods are not well developed for more complex problems, and often ignore the component of variability from estimating the weight from the data. Bias and variance considerations aside, statisticians rightly resist attempts to analyze data selectively, and hence aim to analyze all the data to the extent possible; alternatives to CC analysis that incorporate the incomplete cases in a satisfactory way are recommended unless the fraction of incomplete cases is very small, say 10% or less.

Available-case (AC) analysis [65, section 3.3] is a straightforward attempt to exploit the incomplete information by using all the cases available to estimate each individual parameter. For example, suppose the objective is to estimate the correlation matrix of a set of continuous variables $Y_1, \ldots, Y_p$. Complete-case analysis uses the set of complete cases to estimate all the **correlations**; AC analysis uses all the cases with both Y_j and Y_k observed to estimate the correlation of Y_j and Y_k, $1 \leq j, k \leq p$. Since the sample base of available cases for measuring each correlation includes the set of complete cases, the AC method appears to make better use of available information. The sample base changes from correlation to correlation, however, creating potential problems when the missing data are not MCAR or variables are highly correlated. In the presence of high correlations, there is no guarantee that the AC correlation matrix is even positive definite. Haitovsky's [32] simulations concerning regression with highly-correlated continuous data found AC markedly inferior to CC. However, Kim & Curry [44] found AC superior to CC in **simulations** based on weakly correlated data. Simulation studies comparing AC regression estimates with maximum likelihood (ML) under **normality** suggest that ML is superior even when underlying normality assumptions are violated [4, 58, 81]. Although AC estimates are easy to compute **standard errors** are more complex [109]. The method cannot be generally recommended.

Imputation

Methods that impute, or fill in, the missing values have the advantage that, unlike CC analysis, observed values in the incomplete cases are retained. A common naive approach imputes missing values by their simple unconditional sample means (i.e. marginal means). Wilks [111] and Afifi & Elashoff [1] discussed this method in **bivariate** settings. Unconditional mean imputation can yield satisfactory point estimates of some parameters such as unconditional means and totals, but it yields inconsistent estimates of other parameters, even if the data are MCAR. In particular, sample variances from the data filled in by means clearly underestimate actual variances, since the imputed cases contribute zero to the sum of squared deviations from the sample mean. Unconditional mean imputation yields an inconsistent estimate of the **covariance matrix** and distorted estimates of **association** [65, Chapter 3]. **Inferences** (tests and **confidence intervals**) based on the filled-in data are seriously distorted by bias and overstated precision. Thus, unconditional mean imputation cannot be generally recommended.

An improvement over unconditional mean imputation is *conditional mean* imputation, in which each missing value is replaced by an estimate of its conditional mean given the values of the observed

values. For example, in the case of univariate nonresponse with $Y_1, \ldots, Y_{p-1}$ fully observed and Y_p sometimes missing, one approach is to classify cases into cells on the basis of similar values of observed variables, and then to impute missing values of Y_p by the within-cell mean from the complete cases in that cell. A more general approach is regression imputation, in which the regression of Y_p on $Y_1, \ldots, Y_{p-1}$ is estimated from the complete cases, including **interactions** as needed, and the resulting prediction equation is used to impute the estimated conditional mean for each missing value of Y_p. For a general pattern of missing data, the missing values for each case can be imputed from the regression of the missing variables on the observed variables, computed using the set of complete cases. Iterative versions of this method lead (with some important adjustments) to ML estimates under **multivariate normality** [6, 84].

Although conditional mean imputation incorporates information from the observed variables and yields best predictions of the missing values in the sense of **mean square error**, imputations should be judged in terms of the quality of inferences about population parameters from the filled-in data. From this perspective, conditional mean imputation leads to distorted estimates of quantities that are not linear in the data, such as percentiles (*see* **Quantiles**), correlations and other measures of association, and variances and other measures of variability. A solution to this problem is to use random draws rather than best predictions to preserve the distribution of variables in the filled-in data set. An example is *stochastic regression* imputation, in which each missing value is replaced by its regression prediction plus a **random error** with variance equal to the estimated residual variance.

In other approaches, imputations are drawn from the actual values in the data set. A common version of this method in longitudinal studies subject to attrition is to carry the last observation forward in time to fill out the dataset [87]. Clearly, this method is making a very strong assumption about missing data: that the missing values in a case are all identical to the last observed value. Even if we accept the notion that the average level of the variable does not change after drop out, there is no fluctuation about that average. Little & Su [69] suggested better methods for longitudinal imputation based on simple row and column fits. Another method that imputes respondent values is the *hot deck*, as used by the Census Bureau for imputing income in the Current Population Survey (CPS) [33]. For each nonrespondent on one or more income items, the CPS hot deck finds a matching respondent on the basis of variables that are observed for both; the missing items for the nonrespondent are then replaced by the respondent's values. For matching purposes in the CPS, all variables are categorized, and the number of variables used to define matches is large. When no match can be found for a nonrespondent based on all of the variables, the CPS hot deck searches for a match at a lower level of detail, obtained by omitting some variables and collapsing the categories of others. David et al. [8] compared imputations from the CPS hot deck with imputations using a more parsimonious regression model for income.

A more general approach to hot deck imputation is to define a distance function on the basis of the variables that are observed for both nonrespondents and respondents. The missing values for each nonrespondent are then imputed from a respondent that is close to the nonrespondent in terms of the distance function. One such method is *predictive mean matching* [57, 95]. Consider, for simplicity, univariate nonresponse, and suppose that a model predicting Y_p from the other variables $Y_1, \ldots, Y_{p-1}$ has been estimated using the complete cases. For each nonrespondent, predictive mean matching finds a respondent whose predicted value of Y_p is close to the predicted value of the nonrespondent. The respondent's observed value of Y_p is then imputed to the nonrespondent. Lazzeroni et al. [51] showed in simulations that this method in somewhat **robust** to **misspecification** of the model used for matching.

The imputation methods discussed so far assume the missing data are MAR. In contrast, models that are not missing at random (NMAR) assert that even if a respondent and nonrespondent to Y_p appear identical with respect to observed variables $Y_1, \ldots, Y_{p-1}$, their Y_p values differ systematically. Greenlees et al. [31] and Lillard et al. [54] discussed how imputations for missing CPS data can be based on NMAR models. It is also possible to create an NMAR hot deck procedure; for example, respondents' values that are to be imputed to nonrespondents could be multiplied by an inflation or deflation factor that depends on the variables that are observed. A crucial point about the use of NMAR models is that often there is no direct evidence in the data to address the validity of their underlying assumptions. Thus, whenever NMAR models are being considered

it is prudent to consider several NMAR models and explore the sensitivity of analyses to the choice of model [94] (*see* **Sensitivity Analysis**). See Little & Wang [70] for an application of this idea to a longitudinal study of treatments of schizophrenia.

A serious defect with imputation is that it seems to be inventing data. More specifically, a single imputed value cannot represent all the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modeled correctly and random imputations are created. **Large-sample** results [97] show that for simple situations with 30% of the data missing, single imputation under the correct model results in nominal 90% confidence intervals having actual coverages below 80%. The inaccuracy of nominal levels is even more extreme in multiparameter testing problems.

A modification of imputation that fixes this problem is **multiple imputation** (MI) [96, 98]. Instead of imputing a single set of draws for the missing values, a set of M (say $M = 5$) data sets are created, each containing different sets of draws of the missing values from their predictive distribution. We then apply the analysis to each of the M data sets and combine the results in a simple way. In particular for scalar estimands, the MI estimate is the average of the estimates from the M data sets, and the variance of the estimate is the average of the variances from the five data sets plus $1 + 1/M$ times the sample variance of the estimates over the M data sets (the factor $1 + 1/M$ is a small-M correction). The last quantity here estimates the contribution to the variance from imputation uncertainty, missed by single imputation methods. Another benefit of multiple imputation is that the averaging over data sets results in more efficient point estimates than does single random imputation. Often MI is not much more difficult than doing a single imputation – the additional computing from repeating an analysis M times is not a major burden and methods for combining inferences are straightforward. Most of the work is in generating good predictive distributions for the missing values.

Maximum Likelihood for Ignorable Models

Complete-case analysis and imputation both result in rectangular data sets for analysis. But there are statistical methods that let us analyze a nonrectangular data set without having to impute the missing values. One such approach is the method of maximum likelihood (ML) with associated **large-sample** standard errors based on the **information matrix**.

The ML approach avoids imputation by formulating a statistical model and basing inference on the likelihood function of the incomplete data. Define Y and M as above, and let $X = x_{ij}$ denote an $n \times q$ matrix of fixed covariates, assumed fully observed, with the ith row $x_i = x_{i1}, \ldots, x_{iq}$, where x_{ij} is the value of covariate X_j for subject i. Covariates that are not fully observed should be treated as **random variables** and modeled with the set of Y_js [64]. The data and missing-data mechanism are modeled in terms of a joint distribution for Y and M given X. *Selection models* specify this distribution as

$$f(Y, M|X, \theta, \Psi) = f(Y|X, \theta) f(M|Y, X, \Psi), \quad (1)$$

where $f(Y|X, \theta)$ is the model in the absence of missing values, $f(M|Y, X, \Psi)$ is the model for the missing-data mechanism, and θ and Ψ are unknown parameters. The likelihood of θ and Ψ given the data Y_{obs}, M, and X is then proportional to the density of Y_{obs} and M given X regarded as a function of the parameters θ and Ψ, and is obtained by integrating out the missing data Y_{mis} from (1), i.e.

$$L(\theta, \Psi|Y_{\text{obs}}, M, X) = \text{const} \times \int f(Y, M|X, \theta, \Psi)\, dY_{\text{mis}}. \quad (2)$$

The likelihood of θ *ignoring the missing-data mechanism* is obtained by integrating the missing data from the marginal distribution of Y given X, i.e.

$$L(\theta|Y_{\text{obs}}, X) = \text{const} \times \int f(Y|X, \theta)\, dY_{\text{mis}}. \quad (3)$$

The likelihood (3) is easier to work with than (2) since it is computationally simpler and, more importantly, avoids the need to specify a model for the missing-data mechanism, about which little is known in many situations. Hence it is important to determine when valid likelihood inferences are obtained from (3) instead of the full likelihood (2). Rubin [93] showed that valid inferences about θ are obtained from (3) when the data are MAR, i.e.

$$p(M|X, Y, \Psi) = p(M|X, Y_{\text{obs}}, \Psi), \quad \text{for all } Y_{\text{mis}} \text{ and } \Psi.$$

If, in addition, θ and Ψ are distinct in the sense that they have disjoint sample spaces, then likelihood inferences about θ based on (3) are equivalent to inferences based on (2); the missing-data mechanism is then called *ignorable* for likelihood inferences. Large-sample inferences about θ for an ignorable model are based on ML theory, which states that under regularity conditions

$$\theta - \hat{\theta} \sim \mathrm{N}_k(0, C), \tag{4}$$

where $\hat{\theta}$ is the value of θ that maximizes (3), and $\mathrm{N}_k(0, C)$ is the k-variate normal distribution with mean zero and covariance matrix C given by the inverse of an information matrix; for example, $C = I^{-1}(\hat{\theta})$, where I is the observed information matrix $I(\theta) = -\partial^2 \log L(\theta|Y_{\mathrm{obs}}, X)/\partial\theta\,\partial\theta^{\mathrm{T}}$, or $C = J^{-1}(\hat{\theta})$, where $J(\theta)$ is the expected value of $I(\theta)$. As in [65], (4) is written to be open to a frequentist interpretation if $\hat{\theta}$ is regarded as random and θ fixed, or a **Bayesian** interpretation if θ is regarded as random and $\hat{\theta}$ fixed. Thus, if the data are MAR, the likelihood approach reduces to developing a suitable model for the data and computing $\hat{\theta}$ and C.

Likelihoods based on incomplete data often have complicated forms and require iterative maximization algorithms. In some situations the method of *factored likelihoods*, first described by Anderson [3], yields explicit ML estimates. The idea is to **transform** θ to $\phi(\theta) = [\phi_i(\theta), \ldots, \phi_Q(\theta)]$, where the components $\phi_1, \ldots, \phi_Q$ are distinct, and the likelihood of ϕ factors into the product $L(\phi|Y_{\mathrm{obs}}, X) = \prod_{q=1}^{Q} L_q(\phi_q|Y_{\mathrm{obs}}, X)$, where each factor $L_q(\phi_q|Y_{\mathrm{obs}}, X)$ corresponds to a complete-data problem or a simpler incomplete-data problem. The ML estimate $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_Q)$ of ϕ is found by maximizing each factor $L_q(\phi_q|Y_{\mathrm{obs}}, X)$ separately, and the ML estimate of θ is then $\hat{\theta} = \theta(\hat{\phi})$, where $\theta(\phi)$ is the inverse transformation from ϕ to θ. Consider, for example, **bivariate normal** data:

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \sim_{\mathrm{ind}} \mathrm{N}_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right),$$

and a monotone pattern with m complete cases $\{(y_{i1}, y_{i2}) : i = 1, \ldots, m\}$ and $n - m$ incomplete cases $\{y_{i1} : i = m + 1, \ldots, n\}$ with Y_2 missing. Let $\theta = (\mu_1, \sigma_{11}, \mu_2, \sigma_{22}, \sigma_{12})$ and $\phi = (\phi_1, \phi_2)$, where $\phi_1 = (\mu_1, \sigma_{11})$, $\phi_2 = (\beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$, and $\beta_{21.1} = \sigma_{12}/\sigma_{11}$, $\beta_{20.1} = \mu_2 - \beta_{21.1}\mu_1$, $\sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$ are, respectively, the slope, intercept, and residual variance of the regression of Y_2 on Y_1. The ignorable model likelihood of ϕ based on Y_{obs} then factorizes into the complete-data likelihood of ϕ_1 based on the n observations $\{y_{i1} : i = 1, \ldots, n\}$ and the complete-data likelihood of ϕ_2 for the regression of Y_2 on Y_1 based on the m complete cases $\{(y_{i1}, y_{i2}) : i = 1, \ldots, m\}$. Explicit expressions for the ML estimates of ϕ and hence θ are readily obtained. In particular:

$$\begin{aligned} \hat{\mu}_2 &= \hat{\beta}_{20.1} + \hat{\beta}_{21.1}\hat{\mu}_1 = \overline{y}_2 - b_{21}\overline{y}_1 + b_{21}\hat{\mu}_1 \\ &= \overline{y}_2 + b_{21}(\hat{\mu}_1 - \overline{y}_1), \end{aligned}$$

where $\overline{y}_1$, $\overline{y}_2$ and b_{21} are the sample means of Y_1 and Y_2 and **least squares** slope of Y_2 on Y_1 based on the m complete cases and $\hat{\mu}_1$ is the sample mean of Y_1 based on all n cases. This is known as the regression estimate (*see* **Ratio and Regression Estimates**) of the mean of Y_2, and is a well-known estimator from **double sampling** in sample surveys. It is also the average of observed and imputed values from regression imputation, discussed before. For further details on this example, and applications to multivariate normal and **multinomial** data with a monotone missing data pattern, see [91] or [65, Chapter 6].

The factored likelihood method does not work in the above problem if incomplete cases on Y_2 are also available. Here, and in many other problems, maximization of the likelihood requires numerical methods. Standard optimization methods such as Newton–Raphson or Scoring can be applied; for example, Hartley & Hocking [34] applied a scoring algorithm to multivariate normal data with missing values, and Jennrich & Schluchter [42] applied modified scoring to unbalanced repeated-measures data. Alternatively, the **EM algorithm** [10] can be applied, a general algorithm for incomplete data problems that provides an interesting link with imputation methods. The history of EM, which dates back at least to McKendrick [73] for particular problems, is sketched in [65]. For more recent work on extensions, see [77].

For ignorable models, let $L(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, X)$ denote the likelihood of θ based on the hypothetical complete data $Y = (Y_{\mathrm{obs}}, Y_{\mathrm{mis}})$ and covariates X. Let $\theta^{(t)}$ denote an estimate of θ at iteration t of EM. Iteration $t + 1$ consists of an E-step and an M-step. The E-step consists of taking the **expectation** of $\log L(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, X)$ over the conditional distribution of Y_{mis} given Y_{obs} and X, evaluated at $\theta = \theta^{(t)}$. That is, the expected log likelihood $Q(\theta|\theta^{(t)}) = \int \log L(\theta|Y_{\mathrm{obs}},\ Y_{\mathrm{mis}}, X)\, f(Y_{\mathrm{mis}}|Y_{\mathrm{obs}},\ X,\ \theta^{(t)})\, \mathrm{d}Y_{\mathrm{mis}}$.

is formed. When the complete data belong to an **exponential family** with complete-data **sufficient statistics** S, the E-step simplifies to computing expected values of these statistics given the observed data and $\theta = \theta^{(t)}$, thus in a sense "imputing" the sufficient statistics [105].

The M-step determines $\theta^{(t+1)}$ to maximize $Q(\theta|\theta^{(1)})$ with respect to θ. In exponential family cases this step is the same as for complete data, except that the complete-data sufficient statistics S are replaced by their estimates from the E-step. Thus, the M-step is often easy or available with existing software, and the programming work is mainly confined to E-step computations. Under very general conditions, each iteration of EM increases the log likelihood, and under more restrictive but still general conditions EM converges to a maximum of the likelihood function [112]. If a unique finite ML estimate of θ exists, then EM will find it.

Little & Rubin [65] provided many applications of EM to particular models, including: (i) multivariate normal data with a general pattern of missing values and the related problem of multivariate linear regression with missing data [6, 84]; (ii) robust inference based on **multivariate *t*** models [50, 58]; (iii) **loglinear models** for multiway **contingency tables** with missing data [21]; and (iv) the general location model for mixtures of continuous and categorical variables [68, 83], which yields ML algorithms for logistic regression with missing covariates [108]. Schluchter & Jackson [102] provided an EM algorithm for **survival analysis** with missing covariates. An extensive bibliography of the myriad of EM applications is given in [74].

The EM algorithm is reliable, but has a linear convergence rate determined by the fraction of missing information, as defined in [10]. When the fraction of missing information is large, convergence can be painfully slow. Meng & Van Dyk [77] showed how the clear choice of the missing data can be used to speed convergence. There is an extensive literature on extensions and enhancements of EM for cases where the E- or M-step is hard or slow [17, 18, 41, 47, 48, 71, 76, 77]. EM does not involve computation and inversion of an information matrix based on the observed data. This makes the algorithm particularly attractive in problems where the number of parameters is large, as in ML algorithms for biomedical imaging, such as positron emission tomography [17, 18, 49, 103] (*see* **Image Analysis and Tomography**). This feature of EM has the disadvantage that asymptotic standard errors based on the inverse of the information matrix are not an output. An information matrix can be computed and inverted separately. Alternative approaches to computing standard errors are to use the formulas in [72], to build an information matrix from supplemental EM steps [75], to use **bootstrap methods** [14, 58], or to switch to a Bayesian simulation method that simulates the posterior distribution of θ (see below).

Maximum Likelihood for Nonignorable Models

Ignorable ML is appropriate when the data are MAR. Nonignorable, non-MAR models apply when missingness depends on the missing values. For example, if a subject dropped out of the longitudinal study when his/her blood pressure got too high and we did not observe that blood pressure, or if in an analgesic study measuring pain, the subject dropped out when the pain was high and we did not observe that pain value, missingness depends on the missing value. A correct likelihood analysis must be based on the full likelihood from a model for the joint distribution of Y and M. The standard likelihood asymptotics apply to nonignorable models provided the parameters are identified, and computational tools such as EM also apply to this more general class of models. However, often information to estimate simultaneously the parameters of the missing-data mechanism and the parameters of the complete-data model is limited, and estimates are sensitive to **misspecification** of the model. Often a **sensitivity analysis** is needed to see how much the answers change for various assumptions about the missing-data mechanism.

There are two broad classes of models for the joint distribution of Y and M. *Selection* models model the joint distribution as in (1). *Pattern-mixture* models specify

$$f(Y, M|X, \pi, \phi) = f(Y|X, M, \phi) f(M|X, \pi), \quad (5)$$

where ϕ and π are unknown parameters and now the distribution of Y is conditioned on the missing-data pattern M [29, 60, 65, 94]. Eqs (1) and (5) are simply two different ways of factoring the joint distribution of Y and M. When M is independent of Y the two specifications are equivalent with $\theta = \phi$

and $\psi = \pi$. Otherwise (1) and (5) generally yield different models.

Most of the literature on missing data has concerned selection models of the form (1) for univariate nonresponse. Examples are the probit selection model [2, 36], and the closely related logit model of Greenlees et al. [31], extended to repeated-measures data in [11]. The sensitivity of answers to model misspecification is discussed in [[29; 55; 65], Chapter 11; [104]] and the discussion in [11]. Nonignorable models for contingency tables are discussed in [5; 16; 65, Chapter 9; [86]].

Pattern-mixture models seem more natural when missingness defines a distinct stratum of the population of intrinsic interest, such as individuals reporting "don't know" in an opinion survey. However, pattern-mixture models can also provide inferences for parameters θ of the complete-data distribution by expressing the parameters of interest as functions of the pattern-mixture model parameters ϕ and π. An advantage of the pattern-mixture modeling approach over selection models is that assumptions about the form of the missing-data mechanism are sometimes less specific in their parametric form, since they are incorporated in the model via parameter restrictions. This idea is explained in specific normal models in [61] and [70].

Heitjan & Rubin [40] and Heitjan [39] extended the formation of missing-data problems via the joint distribution of Y and M to more general incomplete data problems involving coarsened data. The idea is to replace the binary missing-data indicators $M = \{m_{ij}\}$ by random coarsening (*see* **Coarsening at Random**) variables $G = \{g_{ij}\}$, which map the y_{ij} values to coarsened versions $z_{ij}(y_{ij})$. Particular values of g_{ij} could map y_{ij} to "completely observed" and "completely missing", as with m_{ij} above, but other values of g_{ij} might map y_{ij} into other sets, for example a finite interval would correspond to interval censoring. The data are then defined as *coarsened completely at random* or *coarsened at random* depending on whether the distribution of G is independent of Y, or depends on Y only through observed data. Full and ignorable likelihoods can be defined for this more general setting. This theory provides a bridge between missing-data theory and theories of censoring (*see* **Censored Data**) in the survival analysis literature. For biomedical applications, see [38].

Bayesian Simulation Methods

Maximum likelihood is most useful when sample sizes are large, since then the log likelihood is nearly quadratic and can be summarized well using the ML estimate θ and its large sample variance–covariance matrix. When sample sizes are small, a useful alternative approach is to add a **prior distribution** for the parameters and compute the posterior distribution of the parameters of interest. For ignorable models this posterior is

$$p(\theta|Y_{\text{obs}}, M, X) \equiv p(\theta|Y_{\text{obs}}, X) = \text{const} p(\theta|X) \times f(Y_{\text{obs}}|X, \theta),$$

where $p(\theta|X)$ is the prior and $f(Y_{\text{obs}}|X, \theta)$ is the density of the observed data. Since the posterior distribution rarely has a simple analytic form for incomplete-data problems, **simulation** methods are often used to generate draws of θ from the posterior distribution $p(\theta|Y_{\text{obs}}, M, X)$. I outline two of these simulation methods for the ignorable case, although the techniques can also be applied to nonignorable models.

For missing data problems where the likelihood can be factored into complete-data components, $L(\phi|Y_{\text{obs}}, X) = \prod_{q=1}^{Q} L_q(\phi_q|Y_{\text{obs}}, X)$ and the parameters $\phi_1, \ldots, \phi_Q$ are also a priori independent, the posteriors of $\phi_1, \ldots, \phi_Q$ are also independent, and draws $\phi^{(d)} = (\phi_1^{(d)}, \ldots, \phi_Q^{(d)})$ can be obtained directly from the complete-data posterior distributions. Draws of θ are then obtained as $\theta^{(d)} = \theta(\phi^{(d)})$, where $\theta(\phi)$ is the inverse **transformation** from ϕ to θ. This method is analogous to the factored likelihood method for ML estimation described above. For an application to normal data see [65, Chapter 6].

Data augmentation [107] is an iterative method for simulating the posterior distribution of θ that combines features of the EM algorithm and multiple imputation, with M imputations of each missing value at each iteration. It can be thought of as a small-sample refinement of the EM algorithm using simulation, with the imputation step corresponding to the E-step and the posterior step corresponding to the M-step. An important special case of data augmentation arises when M is set equal to one, yielding the following special case of the Gibbs' sampler [22, 25] (*see* **Markov Chain Monte Carlo**). Start with an initial draw $\theta^{(0)}$ from an approximation

to the posterior distribution of θ. Given a value $\theta^{(t)}$ of θ drawn at iteration t:

1. draw $Y_{\text{mis}}^{(t+1)}$ with density $p(Y_{\text{mis}}|Y_{\text{obs}}, X, \theta^{(t)})$;
2. draw $\theta^{(t+1)}$ with density $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)}, X)$.

The procedure is motivated by the fact that the distributions in 1 and 2 are often much easier to draw from than the correct posterior distributions, $p(Y_{\text{mis}}|Y_{\text{obs}}, X)$ and $p(\theta|Y_{\text{obs}}, X)$. The iterative procedure can be shown in the limit to yield a draw from the joint posterior distribution of Y_{mis} and θ given Y_{obs} and X. The algorithm was termed *chained data augmentation* in [106]. This algorithm can be run independently K times to generate K i.i.d. draws from the approximate joint posterior distribution of θ and Y_{mis}. A number of articles [23, 24, 26, 107] have discussed techniques for monitoring the convergence of the algorithms. Schafer [100] developed algorithms that use iterative Bayesian simulation to multiply impute rectangular data sets with arbitrary patterns of missing values when the missing-data mechanism is ignorable. The methods are applicable when the rows of the complete-data matrix can be modeled as i.i.d. observations from the multivariate normal, multinomial loglinear, and general location models.

Methods for Unbalanced Repeated-Measures Data

I conclude by reviewing methods for longitudinal data with unequal numbers of measurements between subjects. For normal outcomes and ignorable missing data, a wide range of problems can be tackled using the random-effects model:

$$(y_i|X_i, \beta_i) \sim_{\text{iid}} \text{N}_k(X_{1i}\alpha + X_{2i}\beta_i, \Sigma),$$
$$\beta_i|X_i \sim_{\text{iid}} \text{N}_q(0, \Gamma),$$

where $\text{N}_p(\alpha, B)$ denotes the p-variate normal distribution with mean α covariance matrix B; X_{1i} is a known $(K \times p)$ design matrix containing fixed within-subject and between-subject covariates, with associated unknown $(p \times 1)$ parameter vector α; β_i is an unknown $(q \times 1)$ random-coefficient vector; and X_{2i} is a known $(K \times q)$ matrix for modeling the random effects. Estimation for this model is discussed in [35, 42, 46], and ML estimation is currently available in SAS Proc Mixed [99], or the BMDP program BMDP5V [12] (*see* **Software, Biostatistical**). For Bayesian inference using the Gibbs' sampler, see [28].

For longitudinal categoric data with unequal numbers of measurements, standard loglinear models are unsatisfactory because of the conditional interpretation of the parameters. ML methods have been proposed on the basis of marginal multinomial models (*see* **Marginal Models**) [43, 115]. An alternative approach is to assume categoric outcomes are indicators for underlying continuous outcomes that follow a normal model [27, 37]. Nonlikelihood approaches include the weighted **least squares** methods [45], and iterative methods based on generalized estimating equations [20, 52, 53, 85, 89]. Another approach is analysis by summary measures, in which we obtain a summary measure for each individual and then analyze it across the subjects.

A variety of nonignorable pattern-mixture and selection models for drop outs in longitudinal data have been proposed, including models for informative drop out where drop out depends on underlying unobserved slopes characterizing a patient's decline [9, 62, 79, 101, 110, 113, 114]. An advantage of pattern-mixture models in this setting is that this part of the model can usually be fit using standard software such as PROC MIXED [99] by simply including the drop out indicator as a covariate in the model for the distribution of the y_is. Nonignorable models for repeated-measures categorical data are considered in [7] and [78].

Acknowledgments

This research was supported by National Science Foundation Grant DMS 9408837. Donald Rubin and Nathaniel Schenker's important influences on this review are gratefully acknowledged.

References

[1] Afifi, A.A. & Elashoff, R.M. (1967). Missing observations in multivariate statistics II: point estimation in simple linear regression, *Journal of the American Statistical Association* **62**, 595–604.

[2] Amemiya, T. (1984). Tobit models: a survey, *Journal of Econometrics* **24**, 3–61.

[3] Anderson, T.W. (1957). Maximum likelihood estimation for the multivariate normal distribution when some observations are missing, *Journal of the American Statistical Association* **52**, 200–203.

[4] Azen, S.P., Van Guilder, M. & Hill, M.A. (1989). Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomly incomplete data, *Statistics in Medicine* **8**, 217–228.

[5] Baker, S.G. & Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonresponse, *Journal of the American Statistical Association* **83**, 62–69.

[6] Beale, E.M.L. & Little, R.J.A. (1975). Missing values in multivariate analysis, *Journal of the Royal Statistical Society, Series B* **37**, 129–145.

[7] Conaway, M.R. (1992). The analysis of repeated categorical measurements subject to nonignorable nonresponse, *Journal of the American Statistical Association* **87**, 817–824.

[8] David, M.H., Little, R.J.A., Samuhel, M.E. & Triest, R.K. (1986). Alternative methods for CPS income imputation, *Journal of the American Statistical Association* **81**, 29–41.

[9] DeGruttola, V. & Tu, X.M. (1994). Modeling progression of CD4-lymphocyte count and its relationship to survival time, *Biometrics* **50**, 1003–1014.

[10] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[11] Diggle, P. & Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion), *Applied Statistics* **43**, 49–94.

[12] Dixon, W.J. (1988). *BMDP Statistical Software*. University of California Press, Berkeley.

[13] Dorey, F.J., Little, R.J.A. & Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring, *Statistics in Medicine* **12**, 1589–1603.

[14] Efron, B. (1994). Missing data, imputation and the bootstrap, *Journal of the American Statistical Association* **89**, 463–479.

[15] Ezzati, T. & Khare, M. (1992). Nonresponse adjustments in a National Health Survey, in *American Statistical Association 1992 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 339–344.

[16] Fay, R.E. (1986). Causal models for patterns of nonresponse, *Journal of the American Statistical Association* **81**, 354–365.

[17] Fessler, J.A. & Hero, A.O. (1994). Space-alternating generalized expectation-maximization algorithm, *IEEE Transactions on Signal Processing* **42**, 2664–2677.

[18] Fessler, J.A. & Hero, A.O. (1995). Penalized maximum-likelihood image reconstruction using space-alternating generalized expectation-maximization algorithm, *IEEE Transactions on Image Processing* **4**, 1417–1438.

[19] Finkelstein, D.M. & Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data, *Biometrics* **41**, 933–945.

[20] Fitzmaurice, G.M., Laird, N.M. & Rotnitzky, A.G. (1993). Regression models for discrete longitudinal responses, *Statistical Science* **8**, 284–309.

[21] Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data, *Journal of the American Statistical Association* **77**, 270–278.

[22] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[23] Gelfand, A.E., Hills, S.E., Racine-Poon, A. & Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *Journal of the American Statistical Association* **85**, 972–985.

[24] Gelman, A. & Rubin, D.B. (1992). Honest inferences from iterative simulation (with discussion), *Statistical Science* **4**, 457–511.

[25] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs' distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[26] Geyer, C.J. (1992). Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **4**, 473–511.

[27] Gibbons, R.D. & Bock, R.D. (1987). Trend in correlated proportions, *Psychometrika* **52**, 113–124.

[28] Gilks, W.R., Wang, C.C., Yvonnet, B. & Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs' sampling, *Biometrics* **49**, 441–453.

[29] Glynn, R., Laird, N.M. & Rubin, D.B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse, in *Drawing Inferences from Self-Selected Samples*, H. Wainer, ed. Springer-Verlag, New York, pp. 119–146.

[30] Goksel, H., Judkins, D.R. & Mosher, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey, in *American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 581–586.

[31] Greenlees, W.S., Reece, J.S. & Zieschang, K.D. (1982). Imputation of missing values when the probability of nonresponse depends on the variable being imputed, *Journal of the American Statistical Association* **77**, 251–261.

[32] Haitovsky, Y. (1968). Missing data in regression analysis, *Journal of the Royal Statistical Society, Series B* **30**, 67–81.

[33] Hanson, R.H. (1978). The current population survey: design and methodology, *Technical Paper*, No. 40, US Bureau of the Census, Washington.

[34] Hartley, H.O. & Hocking, R.R. (1971). The analysis of incomplete data, *Biometrics* **14**, 174–194.

[35] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems (with discussion), *Journal of the American Statistical Association* **72**, 320–340.

[36] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models, *Annals of Economic and Social Measurement* **5**, 475–492.

[37] Hedeker, D. (1993). *MIXOR: A Fortran Program for Mixed-Effects Ordinal Probit and Logistic Regression*. Prevention Research Center, University of Illinois at Chicago, Chicago.

[38] Heitjan, D.F. (1993). Ignorability and coarse data: some biomedical examples, *Biometrics* **49**, 1099–1109.

[39] Heitjan, D.F. (1993). Ignorability in general complete-data models, *Biometrika* **81**, 701–708.

[40] Heitjan, D. & Rubin (1991). Ignorability and coarse data, *Annals of Statistics* **19**, 2244–2253.

[41] Jamshidian, M. & Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm, *Journal of the American Statistical Association* **88**, 221–228.

[42] Jennrich, R.I. & Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics* **42**, 805–820.

[43] Kenward, M.G., Lesaffre, E. & Molenberghs, G. (1994). An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random, *Biometrics* **50**, 945–953.

[44] Kim, J.O. & Curry, J. (1977). The treatment of missing data in multivariate analysis, *Sociological Methods and Research* **6**, 215–240.

[45] Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. & Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data, *Biometrics* **33**, 133–158.

[46] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[47] Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm, *Journal of the Royal Statistical Society, Series B* **57**, 425–437.

[48] Lange, K. (1995). A quasi-Newtonian acceleration of the EM algorithm, *Statistica Sinica* **5**, 1–18.

[49] Lange, F. & Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography, *Journal of Computer-Assisted Tomography* **8**, 306–316.

[50] Lange, K., Little, R.J.A. & Taylor, J. (1989). Robust statistical inference using the T distribution, *Journal of the American Statistical Association* **84**, 881–896.

[51] Lazzeroni, L.C., Schenker, N. & Taylor, J.M.G. (1990). Robustness of multiple imputation techniques to model specification, in *American Statistical Association 1990 Proceedings of the Section on Survey Reasearch Methods*. American Statistical Association, Alexandria, pp. 260–265.

[52] Liang, K-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[53] Liang, K-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[54] Lillard, L., Smith, J.P. & Welch, F. (1986). What do we really know about wages: the importance of nonreporting and census imputation, *Journal of Political Economy* **94**, 489–506.

[55] Little, R.J.A. (1985). A note about models for selectivity bias, *Econometrica* **53**, 1469–1474.

[56] Little, R.J.A. (1986). Survey nonresponse adjustments, *International Statistical Review* **54**, 139–157.

[57] Little, R.J.A. (1988). Missing data adjustments in large surveys, *Journal of Business and Economic Statistics* **6**, 287–301.

[58] Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values, *Applied Statistics* **37**, 23–38.

[59] Little, R.J.A. (1992). Regression with incomplete X's; a review, *Journal of the American Statistical Association* **87**, 1227–1237.

[60] Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association* **88**, 125–134.

[61] Little, R.J.A. (1994). A class of pattern-mixture models for normal missing data, *Biometrika* **81**, 471–483.

[62] Little, R.J.A. (1995). Modeling the drop-out mechanism in longitudinal studies, *Journal of the American Statistical Association* **90**, 1112–1121.

[63] Little, R.J.A. & Rubin, D.B. (1983). Incomplete data, in *Encyclopedia of the Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 46–53.

[64] Little, R.J.A. & Rubin, D.B. (1983). On jointly estimating parameters and missing data by maximizing the complete data likelihood, *American Statistician* **37**, 218–220.

[65] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[66] Little, R.J.A. & Rubin, D.B. (1989). Missing data in social science data sets, *Sociological Methods and Research* **18**, 292–326.

[67] Little, R.J.A. & Schenker, N. (1994). Missing data, in *Handbook for Statistical Modeling in the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg & M.E. Sobel, eds. Plenum Press, New York, pp. 39–75.

[68] Little, R.J.A. & Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika* **72**, 497–512.

[69] Little, R.J.A. & Su, H.L. (1989). Item nonresponse in panel surveys, in *Panel Surveys*, D. Kasprzyk, G. Duncan, G. Kalton & M.P. Singh, eds. Wiley, New York, pp. 400–425.

[70] Little, R.J.A. & Wang, Y.-X. (1996). Pattern-mixture models for multivariate incomplete data with covariates, *Biometrics* **52**, 98–111.

[71] Liu, C. & Rubin, D.B. (1994). The ECME algorithm: a simple extension of EM and ECM with fast monotone convergence, *Biometrika* **81**, 633–648.

[72] Louis, T.A. (1982). Finding the observed information matrix using the EM algorithm, *Journal of the Royal Statistical Society, Series B* **44**, 226–233.

[73] McKendrick, A.G. (1926). Applications of mathematics to medical problems, *Proceedings of the Edinburgh Mathematics Society* **44**, 98–130.

[74] Meng, X.L. & Pedlow, S. (1992). EM: a bibliographic review with missing articles, in *American Statistical Association 1992 Proceedings of the Section on Computing*. American Statistical Association, Alexandria, pp. 24–27.

[75] Meng, X.L. & Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *Journal of the American Statistical Association* **86**, 899–909.

[76] Meng, X.L. & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika* **80**, 267–278.

[77] Meng, X.L. & Van Dyk (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion), *Journal of the Royal Statistical Society, Series B* **59**, 511–567.

[78] Molenberghs, G., Kenward, M.G. & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with informative dropout, *Biometrika* **84**, 33–44.

[79] Mori, M., Woolson, R.F. & Woodsworth, G.G. (1994). Slope estimation in the presence of informative censoring: modeling the number of observations as a geometric random variable, *Biometrics* **50**, 39–50.

[80] Murray, G.D. & Findlay, J.G. (1988). Correcting for the bias caused by drop-outs in hypertension trials, *Statistics in Medicine* **7**, 941–946.

[81] Muthen, B., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random, *Psychometrika* **52**, 431–462.

[82] Oh, H.L. & Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse, in *Incomplete Data in Sample Surveys*, Vol. 2: *Theory and Bibliographies*, W.G. Madow, I. Olkin & D.B. Rubin, eds. Academic Press, New York, pp. 143–184.

[83] Olkin, I. & Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables, *Annals of Mathematical Statistics* **32**, 448–465.

[84] Orchard, T. & Woodbury, M.A. (1972). A missing information principle: theory and applications, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 697–715.

[85] Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements, *Statistics in Medicine* **12**, 1723–1732.

[86] Park, T. & Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse, *Journal of the American Statistical Association* **89**, 44–52.

[87] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, New York.

[88] Robins, J. & Rotnitsky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association* **90**, 122–129.

[89] Robins, J., Rotnitsky, A. & Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**, 106–121.

[90] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

[91] Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems, *Journal of the American Statistical Association* **69**, 467–474.

[92] Rubin, D.B. (1976). Comparing regressions when some predictor values are missing, *Technometrics* **18**, 201–205.

[93] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

[94] Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys, *Journal of the American Statistical Association* **72**, 538–543.

[95] Rubin, D.B. (1986). Statistical matching and file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics* **4**, 87–94.

[96] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[97] Rubin, D.B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *Journal of the American Statistical Association* **81**, 366–374.

[98] Rubin, D.B. & Schenker, N. (1991). Multiple imputation in health-care databases: an overview and some applications, *Statistics in Medicine* **10**, 585–598.

[99] SAS (1992). The mixed procedure, in *SAS/STAT Software: Changes and Enhancements*, Release 6.07, Technical Report P-229. SAS Institute, Inc., Cary.

[100] Schafer, J.L. (1996). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

[101] Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data, *Statistics in Medicine* **11**, 1861–1870.

[102] Schluchter, M.D. & Jackson, K.L. (1989). Loglinear analysis of censored survival data with partially observed covariates, *Journal of the American Statistical Association* **84**, 42–52.

[103] Shepp, L.A. & Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography, *IEEE Transactions on Image Processing* **2**, 113–122.

[104] Stolzenberg, R.M. & Relles, D.A. (1990). Theory testing in a world of constrained research design – the significance of Heckman's censored sampling bias correction for nonexperimental research, *Sociological Methods and Research* **18**, 395–415.

[105] Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family, *Scandinavian Journal of Statistics* **1**, 49–58.

[106] Tanner, M.A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. Springer-Verlag, New York.

[107] Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association* **82**, 528–550.

[108] Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*. Springer-Verlag, New York.

[109] Van Praag, B.M.S., Dijkstra, T.K. & Van Velzen, J. (1985). Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem, *Psychometrika* **50**, 25–36.

[110] Wang-Clow, F., Lange, M., Laird, N.M. & Ware, J.H. (1995). Simulation study of estimators for rate of change in longitudinal studies with attrition, *Statistics in Medicine* **14**, 283–297.

[111] Wilks, S.S. (1932). Moments and distribution of estimates of population parameters from fragmentary samples, *Annals of Mathematical Statistics* **3**, 163–195.

[112] Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Annals of Statistics* **11**, 95–103.

[113] Wu, M.C. & Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model, *Biometrics* **45**, 939–955.

[114] Wu, M.C. & Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics* **44**, 175–188.

[115] Zhao, L.P. & Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika* **77**, 642–648.

RODERICK J. LITTLE

Misspecification

Statistical models specify the density of a **response random variable** Y. This density depends on parameters which often relate to a specified function of **explanatory variables**, or **covariates** X. The parameters of the density and of the function of covariates are typically unknown and the statistical problem is to use a sample of data to calculate accurate estimates of the unknown parameters along with measures of uncertainty associated with the estimates.

In practice, data analysts often choose a particular statistical model because it is easy to work with mathematically or is simple to fit using easily available **software**. For example, when one wants to assess the association of an explanatory variable with an outcome, it is tempting to fit a **linear regression** model, since this model is easy to fit and describe.

However, there may exist a more accurate specification of a statistical model than that chosen by an analyst. That is, there may be densities that fit the data better and functions of the covariates that better describe their relationship with the response. When this occurs, we have misspecified the statistical model. This misspecification may be due to either an incorrect conceptual understanding of the phenomenon under study or an inability to collect data on all the relevant factors related to the outcome under study. Model misspecifications include choosing the incorrect link function or omitting important covariates with **generalized linear models**, incorrectly assuming independence of observations or misspecifying the within-cluster dependence structure with clustered data (*see* **Cluster Analysis of Subjects, Hierarchical Methods**), misspecifying the mixing distribution in generalized linear mixed models, and wrongly assuming **proportional hazards** with survival data.

Effects of Model Misspecification

Model misspecification can produce **biased** or inefficient estimates (*see* **Efficiency and Efficient Estimators**) of the associations of covariates with the response, invalid variance estimates, and less powerful tests of hypotheses concerning covariate–response associations (*see* **Power**). To examine effects of model misspecification such as bias and loss of efficiency, we must calculate the expected value and variance of estimators obtained from the misspecified model with respect to the true, underlying density of the responses. When we obtain estimators by maximizing an assumed likelihood, further theory is available. Suppose that we use **maximum likelihood** to fit a model that assumes that the response Y follows a distribution F with parameter vector $\boldsymbol{\xi}^*$ and covariates X_F, while, in truth, Y follows a distribution G with parameter vector $\boldsymbol{\xi}$ and covariates X_G such that X_F is a subset of X_G. The work of Huber [3], Akaike [1] and White [10, 11] shows that the "maximum likelihood" estimator, $\hat{\boldsymbol{\xi}}^*$, under the false model **converges** to the value $\boldsymbol{\xi}^*$ that minimizes the **Kullback–Leibler** divergence [4] between the true and misspecified models. That is, $\boldsymbol{\xi}^*$ minimizes

$$\mathrm{E}_{X_G}\mathrm{E}_{Y|X_G} \log\left\{\frac{g(y|\boldsymbol{\xi}, X_G)}{f(y|\boldsymbol{\xi}^*, X_F)}\right\}, \tag{1}$$

where g and f are the true and misspecified response densities, and one takes the expectation with respect to the true model. Thus, when we maximize the incorrect likelihood, we obtain the parameters of the misspecified model that minimize the average difference between the logarithms of the true and misspecified densities. White [10, 11] further showed that $\hat{\boldsymbol{\xi}}^*$ has an asymptotic normal distribution with $\mathrm{var}(\hat{\boldsymbol{\xi}}^*)$ given by a matrix product of the form

$$\mathrm{var}(\hat{\boldsymbol{\xi}}^*) = \mathbf{A}^{-1}(\boldsymbol{\xi}^*)\mathbf{B}(\boldsymbol{\xi}^*)\mathbf{A}^{-1}(\boldsymbol{\xi}^*), \tag{2}$$

where

$$\mathbf{A}(\xi^*) = \mathrm{E}_{X,Z}\mathrm{E}_{Y|X,Z}\left[\frac{\partial^2 \log P_F(Y = y|\xi^*, X)}{\partial\xi_i^*\partial\xi_j^*}\right],$$

$$\mathbf{B}(\xi^*) = \mathrm{E}_{X,Z}\mathrm{E}_{Y|X,Z}\left[\frac{\partial \log P_F(Y = y|\xi^*, X)}{\partial\xi_i^*} \times \frac{\partial \log P_F(Y = y|\xi^*, X)}{\partial\xi_j^*}\right],$$

and expectations are with respect to the true model with G. The matrix $\mathbf{A}$ involves the **information** from the misspecified likelihood, while the matrix $\mathbf{B}$ involves the true variance–covariance structure of the responses. Under correct model specification, i.e. $f = g$, $\mathbf{A} = \mathbf{B}$.

Further results are available in other settings. For example, Li & Duan [5] considered the case where

the responses follow a generalized linear model with slope parameter vector $\boldsymbol{\gamma}$, but one fits a model that misspecifies the link function. Li & Duan showed that the estimated slope $\hat{\boldsymbol{\gamma}}^*$ from the misspecified model typically **consistently** estimates $\boldsymbol{\gamma}$ up to a scale factor. That is, $\mathrm{E}(\hat{\boldsymbol{\gamma}}^*) = c\boldsymbol{\gamma}$, for a constant c. Thus, one can still consistently estimate ratios of slope coefficients, γ_i/γ_j with a misspecified link function.

Solomon [8] and Struthers & Kalbfleisch [9] derived analogous results for survival models. These authors considered the effect of fitting a proportional hazards model when the data actually follow an **accelerated failure–time model** and vice versa, and found that the asymptotic expectation of the estimated regression coefficients of the misspecified models was approximately proportional to the true coefficients.

An Example: Omitted Covariates

We illustrate these results by examining the effects of omitted covariates in **logistic regression** models. We suppose that Y is a **binary** outcome, X and Z are covariates, and that the true model for the probability of response is

$$\operatorname{logit}\Pr(Y = 1|X, Z) = \mu + \beta X + \gamma Z. \quad (3)$$

To these data we fit the misspecified model

$$\operatorname{logit}\Pr(Y = 1|X) = \mu^* + \beta^* X. \quad (4)$$

Minimizing the Kullback–Leibler divergence, given in (1), in μ^* and β^*, we obtain

$$\mathrm{E}_{X,Z}\mathrm{E}_{Y|X,Z}\left\{\frac{1}{[1+\exp(-\mu-\beta X-\gamma Z)]} - \frac{1}{[1+\exp(-\mu^*-\beta^* X)]}\right\} = 0,$$

$$\mathrm{E}_{X,Z}\mathrm{E}_{Y|X,Z}\left\{\frac{X}{[1+\exp(-\mu-\beta X-\gamma Z)]} - \frac{X}{[1+\exp(-\mu^*-\beta^* X)]}\right\} = 0.$$

Thus, μ^* and β^* will depend on the true values μ and β and the joint distribution of X and Z. If X and Z are independent, then it follows that

$$[1+\exp(-\mu^*-\beta^* X)]^{-1} = \mathrm{E}_Z[1+\exp(-\mu-\beta X-\gamma Z)]^{-1} \quad (5)$$

and that $\beta^* = 0$ solves (5) when $\beta = 0$. Expanding the logit of (5) in a Taylor series about $\beta = 0$ yields

$$\mu^* + \beta^* X \approx \log\left[\frac{\mathrm{E}(p)}{\mathrm{E}(q)}\right] + \beta X\left[1 - \frac{\mathrm{var}(p)}{\mathrm{E}(p)\mathrm{E}(q)}\right], \quad (6)$$

where $\operatorname{logit}(p) = \mu + \gamma Z$ and $q = 1 - p$. Since $\mathrm{var}(p) \le \mathrm{E}(p)\mathrm{E}(q)$, we have $|\beta^*| \le |\beta|$, so omitting the covariate Z leads to attenuated estimates of the effect of X (*see* **Shrinkage**).

We compute variances of the estimators obtained from the misspecified, omitted covariate model using (2) and calculate the **asymptotic relative efficiency** (ARE) of $\hat{\beta}^*$ to $\hat{\beta}$ as

$$\mathrm{ARE}(\hat{\beta}^* \text{ to } \hat{\beta} \text{ at } \beta = 0) = \left[\lim_{\beta\to 0}\left\{\frac{\partial}{\partial\beta}\beta^*\right\}\left\{\frac{\partial}{\partial\beta}\beta\right\}^{-1}\right]^2\left[\lim_{\beta\to 0}\frac{\mathrm{var}(\hat{\beta})}{\mathrm{var}(\hat{\beta}^*)}\right].$$

The ARE involves the variances of $\hat{\beta}$ and $\hat{\beta}^*$ as well as the relationship between the parameters that these estimators estimate. Since $\beta^* = 0$ when $\beta = 0$, it is appropriate to compare variances and estimation efficiency at this value. When X and Z are independent, these calculations yield

$$\mathrm{ARE}(\hat{\beta}^* \text{ to } \hat{\beta} \text{ at } \beta = 0) = 1 - \frac{\mathrm{var}(p)}{\mathrm{E}(p)\mathrm{E}(q)}. \quad (7)$$

Thus, misspecifying a logistic regression model by omitting a covariate Z, that is independent of the included X, leads to attenuated estimates of the association of X with the response and less powerful tests of the hypothesis that X is not associated with the response.

The **two-by-two tables** in Tables 1 and 2 further illustrate this special form of model misspecification. Suppose that Y is a binary outcome, X and Z are binary covariates, and that the true model for the

Table 1 Hypothetical data from a logistic model

	$Z=0$			$Z=1$			
	$X=1$	$X=0$			$X=1$	$X=0$	
$Y=1$	90	75	165	$Y=1$	50	25	75
$Y=0$	10	25	35	$Y=0$	50	75	125
	100	100	200		100	100	200

Table 2 Data from Table 1 combined over levels of Z

	$X = 1$	$X = 0$	
$Y = 1$	140	100	240
$Y = 0$	60	100	160
	200	200	400

probability of response follows (3). Suppose, furthermore, that the observations arising from this model are as given in Table 1.

Each two-by-two table in Table 1 leads to an estimate $\hat{\beta} = \log 3$ and a fitted logistic regression model using both X and Z as covariates yields $\hat{\beta} = \log 3$ with associated standard error $\text{se}(\hat{\beta}) = 0.244$ and statistic for a Wald test (*see* **Likelihood**) of $H_0 : \beta = 0$ of $[\hat{\beta}/\text{se}(\hat{\beta})]^2 = 20.29$. Note that $\Pr(X = 1)$ is the same for $Z = 0$ and $Z = 1$, indicating that X and Z are independent. Combining the data over Z, that is, omitting Z and fitting the model given in (4) using the combined table, as given in Table 2, leads to an attenuated estimate of $\hat{\beta}^* = \log 7/3$ with associated standard error $\text{se}(\hat{\beta}^*) = 0.209$ and Wald test of $H_0 : \beta = 0$ of $[\hat{\beta}^*/\text{se}(\hat{\beta}^*)]^2 = 16.44$. Omitting Z not only leads to an attenuated estimate of the effect of X on Y but also to a less powerful test of the hypothesis that X has no effect on Y. To illustrate the quality of the approximation given in (6), applying the Taylor approximation in (6) to the data from Tables 1 and 2 suggests that $\beta^*/\beta = 0.75$, which corresponds closely to the observed value of 0.77. Applying the ARE formula given in (7) to Tables 1 and 2 yields an ARE value of 0.79. This closely corresponds to the observed ratio of the Wald test based on $\hat{\beta}^*$ to that based on $\hat{\beta}$ of 0.81. The analogous ratio of **likelihood ratio test** statistics is 0.78.

Detection of Model Misspecification

One uses **diagnostic** methods such as **residual** plots to examine whether a statistical model adequately describes a given data set; that is, whether one has correctly specified the model. Chapter 12 of McCullagh & Nelder [6] and Chapter 4 of Fahrmeir & Tutz [2] describe many such approaches. These diagnostic procedures include plots to examine the adequacy of assumed link and variance functions and the scale of the model covariates, as well as methods in which one embeds the chosen model within a larger class and tests to see whether a model in the larger class provides a much more accurate description of the data (*see* **Goodness of Fit**).

Since the matrices $\mathbf{A}$ and $\mathbf{B}$ in the variance formula given in (2) are equal with correct model specification, one could compare estimates of these two matrices to examine model adequacy. Indeed, Royall [7] recommends estimating variances of estimated model parameters by plugging estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ into (2) to provide inference that is **robust** to model misspecification.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, B.N. Petrov & F. Czáki, eds. Akademiai Kiadó, Budapest, pp. 267–281.

[2] Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

[3] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics*, Vol. 1. University of California Press, Berkeley, pp. 221–233.

[4] Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[5] Li, K.-C. & Duan, N. (1989). Regression analysis under link violation, *Annals of Statistics* **17**, 1009–1052.

[6] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[7] Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimates, *International Statistical Review* **54**, 221–226.

[8] Solomon, P.J. (1984). Effect of misspecification of regression models in the analysis of survival data, *Biometrika* **71**, 291–298.

[9] Struthers, C.A. & Kalbfleisch, J.D. (1986). Misspecified proportional hazards models, *Biometrika* **73**, 363–369.

[10] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1–25.

[11] White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press, Cambridge.

JOHN M. NEUHAUS

Mode

The population mode of a variable is that value of the variable which is possessed by the greatest number of members of the population. Empirically, the mode may be described as the most frequently occurring value in a sample of observations. It is much used to summarize **nominal data**; for example, the most common blood group, most common eye color, most common type of operation carried out in a particular hospital, and so on. It is occasionally used as a measure of location for continuous variables, but is only really suitable for those with symmetric **unimodal** distributions. For a symmetric distribution, the **mean**, **median**, and mode coincide.

In more formal terms, if $f(x)$ is a **probability** density function with continuous first derivative, a mode is a value of x for which

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = 0, \qquad \frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} < 0.$$

Thus there may be more than one mode of a distribution. An example of a *bimodal distribution* is shown in Figure 1. Empirical examples of such distributions are not common, but an example of a histogram with two distinct modes is shown in Figure 2. The data here correspond to the sizes of myelinated lumbosacral ventral root fibers taken from a kitten of a particular age. The first mode is associated with axons of gamma neurons and the second with alpha neurons.

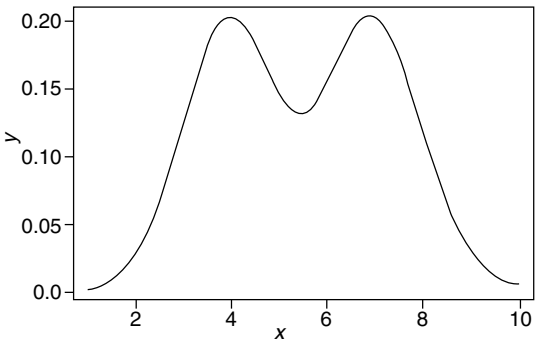

Figure 1 A bimodal density function

Tests for modes and for possible multimodality of density functions are important in, for example,

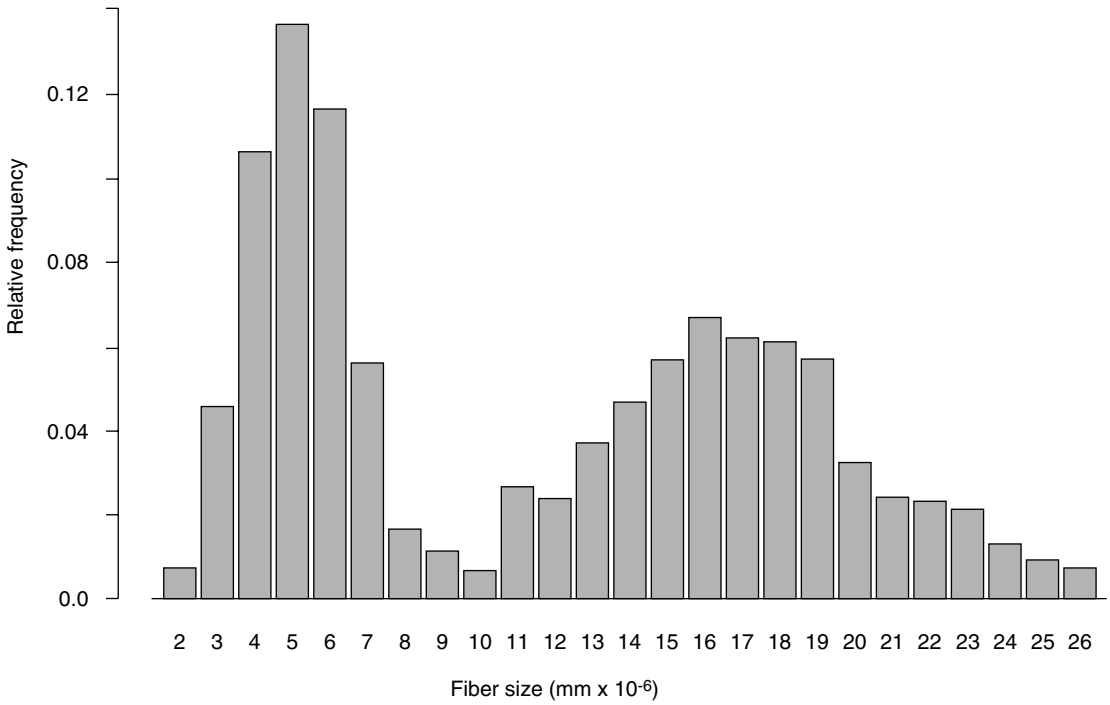

Figure 2 A histogram of myelinated lumbosacral ventral root fiber sizes from a kitten of a particular age

cluster analysis; a number of such tests are described in [1, 2], and [3].

References

[1] Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society, Series B* **43**, 97–99.

[2] Silverman, B.W. (1983). Some properties of a test for multimodality based on kernel density estimates, in *Probability, Statistics and Analysis*, J.F.C. Kingman & G.E.H. Reuter, eds. Cambridge University Press, Cambridge, pp. 248–259.

[3] Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

BRIAN S. EVERITT

Model Checking

A statistical model is an artificial construction and, as such, must be tested before use – indeed, it is often the deficiencies of a "first-guess" model which point the way towards an increased understanding of the behavior of a system. Modeling is (or should be) an iterative process, successively updating and checking a model until it is deemed to be adequate. Note that the adequacy of a model generally depends on its intended use – the question is not, "Is the model correct?" but rather, "Will it do the job?".

All models aim to describe, or explain, the behavior of some quantities of interest (the *outcome* or **response variable** in a statistical context), usually in terms of their relationship to other quantities (*predictor* or **explanatory variables**). Checking a model, therefore, involves asking two questions:

1. Is the model structure realistic (in the sense that it either reflects prior knowledge of the system being studied, or represents relationships that are empirically observed)?
2. How closely does it describe the observed behavior of the response variables?

Answering the first of these questions is largely a subjective affair requiring expert knowledge of the system being modeled; however, there are simple basic checks that can be made – for example, if one of the response variables is a proportion, does the model always yield values between 0 and 1? A good model should have such features built into it by design (*see* **Model, Choice of**), so this will not be discussed further here.

Answering the second question requires data analysis. In ideal circumstances, it may be possible to fit a model using one set of data and calibrate it using another. However, in biostatistical applications data are usually sufficiently scarce that this is not possible, and model fitting and checking are carried out using the same data. Techniques for assessing model adequacy are wide-ranging, and may be formal or informal. They all involve some measure of (dis)agreement between data and model – whether through some sort of residual analysis (*see* **Residuals**), where agreement equates to having observed data values which are similar to those predicted by the model, or through **likelihood**-based methods, where agreement is synonymous with the data values having high plausibility under the model. Departures from the model may be either *isolated*, where individual data points fall outside the general pattern, or *systematic*. We briefly outline techniques which may be appropriate for detecting these departures from the fitted model. Good accounts of the general ideas may be found in [6] and in [11, Chapter 12].

General Methods

When the response variable(s) are continuous (or ordinal with several categories; *see* **Ordered Categorical Data**), perhaps the simplest informal checks of a model are provided by visual inspection of residual plots. The most common plots show appropriately standardized residuals as some function of the model fitted values, or of explanatory variables. Isolated departures from the model are easily spotted using such plots; systematic departures may be indicated if any pattern is discernible, although a merely visual assessment can be misleading. A thorough treatment is given in [2].

Closely related to the idea of residual plots is that of comparing predicted responses from the fitted model with those obtained from the same data using nonparametric methods. This approach allows formal diagnostic procedures (*see* **Diagnostics**) to be established for checking the validity of a parametric model. Examples of these ideas may be found in [3, 4], and [7]; a straightforward overview is given in [1].

When a response variable is discrete, graphical methods of model checking are usually less straightforward. If the predictor variables are also discrete so that the data arise as a contingency table, standardized residuals may be computed for each cell in the table, and the resulting table of residuals inspected visually. The model may be inadequate if any residuals deviate significantly from zero, or if there appears to be a pattern to the residuals in some part of the table (for example, if there is a block of cells where all the residuals are negative). In cases where the predictor variables are continuous (for example, in **logistic regression**) residuals are harder to define – some techniques are given in [9].

Finally, goodness-of-fit tests (*see* **Goodness of Fit**) provide a quick and simple method of checking the fit of a model, although it should be stressed that these tests are usually quite general and may lack

power; they should always be supplemented by an examination of residuals.

Isolated Departures

Isolated departures from a model may not, in themselves, indicate that the model is a poor one for the purposes for which it was intended: it may be, for example, that data values have been wrongly recorded, or that the discrepancies occur at the extremes of the available data where the model is not expected to apply. It is always worth examining the original data record in such cases, as this can often provide an insight into why the model is failing. Many of the considerations which arise in the detection and assessment of **outliers** apply equally well in this context.

When an individual data point is suspect, yet is not obviously wrong, its effect upon the fitted model must be examined. In general, techniques for determining the influence of an individual point involve removing it and refitting the model. The *deletion residual* for a case is defined as the difference between the observed data value and that predicted by the refitted model, and is a measure of the *consistency* of that case with the rest of the data. Other statistics, such as the Cook statistic [11, p. 406], seek to quantify the effect of a point upon the estimated model parameters.

If individual cases are found to have a significant effect upon the fitted model, it may be worth refitting the model using some robust estimation procedure (*see* **Robustness**; **Robust Regression**) so as to downweight the contribution of the suspect points. This usually entails some modification of the classical distribution theory used in procedures such as goodness-of-fit testing; discussion of these issues may be found in [5, 12, 14].

Systematic Departures

Plotting residuals against predictor variables can be very helpful in identifying systematic discrepancies between model and data – if there is pattern in the plot of residuals against a predictor X, this implies that it is some function of X, rather than X itself, which should be used as a predictor.

Systematic discrepancies generally indicate that the overall pattern of a system's behavior is not adequately represented by the modeled relationships between the various components of that system. To check formally for such discrepancies requires some intuition regarding directions in which the model may be improved. Formal methods for checking the adequacy of a model involve embedding it in a wider class of models, and carrying out hypothesis tests to determine whether or not anything is to be gained from extending the model. Clearly, the wider class of models needs to be chosen with some care to yield informative results.

The references below describe techniques for model checking across a broad range of subject areas. In addition to those already cited, [15] is included for its comprehensive survey of the literature in a **time series** context, [10] contains material that is of use in survival analysis (*see* **Survival Analysis, Overview**), and [8] and [13] give a Bayesian perspective (*see* **Bayesian Methods**).

References

[1] Altman, N.S. (1992). An introduction to kernel and nearest-neighbour parametric regression, *American Statistician* **46**, 175–185.

[2] Atkinson, A.C. (1985). *Plots, Transformations and Regression*. Clarendon Press, Oxford.

[3] Azzalini, A. & Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships, *Journal of the Royal Statistical Society, Series B* **55**, 549–557.

[4] Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Clarendon Press, Oxford.

[5] Copas, J.B. (1988). Binary regression models for contaminated data, *Journal of the Royal Statistical Society, Series B* **50**, 225–265.

[6] Davison, A.C. & Tsai, C.L. (1992). Regression model diagnostics, *International Statistical Review* **60**, 337–353.

[7] Firth, D., Glosup, J. & Hinkley, D.V. (1991). Model checking with nonparametric curves, *Biometrika* **78**, 245–252.

[8] Gelman, A.E., Carlin, J.S., Stem, H.S., & Rubin, O.B. (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton.

[9] Landwehr, J.M., Pregibon, D. & Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models, *Journal of the American Statistical Association* **79**, 61–71.

[10] Lin, D.Y. & Spiekerman, C.F. (1996). Model checking techniques for parametric regression with censored data, *Scandinavian Journal of Statistics* **23**, 157–177.

[11] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[12] O'Hara Hines, R.J. & Carter, E.M. (1993). Improved added variable and partial residual plots for the detection of influential observations in generalized linear models, *Applied Statistics* **42**, 3–20.

[13] Pettit, L.I. (1986). Diagnostics in Bayesian model choice, *Statistician* **35**, 183–190.

[14] Simonoff, J.S. & Tsai, C.L. (1991). Assessing the influence of individual observations on a goodness-of-fit test based on nonparametric regression, *Statistics and Probability Letters* **12**, 9–17.

[15] Tsay, R.S. (1992). Model checking via parametric bootstraps in time-series analysis, *Applied Statistics* **41**, 1–15.

(*See also* **Cross-validation**, **Nonparametric Regression**)

R.E. CHANDLER

Model, Choice of

Choosing and fitting statistical models to data occupies a large proportion of the time of many medical statisticians. It is an activity that requires the use of many specific and well-defined techniques, such as using a certain **algorithm** to fit a model or calculating a particular **diagnostic**. However, the sequence of techniques used to select a model and decide its suitability will owe more to experience and judgment than technical expertise. Statistical modeling provides one of the best illustrations of Healy's remark that the practice of statistics is not a science but "that blend of knowledge and practical know-how" that he describes as a technology [15]; in this respect statistics has much more in common with medicine than is often supposed.

The present article describes, in broad outline, the issues that surround the choice of a statistical model. It has already been mentioned that experience and judgment play a greater role in this area of statistics than in many others, so many of the views expressed will inevitably have a marked personal component. On the other hand, while all the examples are medical, many of the issues discussed will be familiar to statisticians, whatever their area of application. Cox [10] provides a more general view of statistical modeling.

Constraints on space mean that most of the article will concentrate on the position when the data have been collected. The reader should, however, be aware that the design of a study can be greatly enhanced if the nature of the model that will be used is borne in mind at that stage (*see* **Experimental Design**).

In practical terms, a statistical model can be thought of as a tool that allows the statistician to determine and describe the relationships between variables in the data, and to quantify the variation present. At a deeper level there remains some uncertainty surrounding modeling [20]. The choice of a model depends on many things, but in most instances, the purpose of the study, the existing knowledge about the system under investigation, and how well the model fits the data are likely to be paramount.

Modeling in Clinical Trials and Epidemiology

Randomized Controlled Trials

Randomized controlled trials do not, as a rule, require much by way of statistical modeling (*see* **Clinical Trials, Overview**). The comparison of randomized groups can be based on a very general model of unit treatment additivity and the act of **randomization**. **Crossover** trials are not considered because they have special problems; similar concerns apply to cluster-randomized studies (*see* **Group-randomization Designs**), which are also excluded.

Some aspects of model choice do arise in connection with the outcome variable. For example, is a **proportional hazards** model suitable for an outcome that is a survival time, and, if so, is a fully parametric specification such as a **Weibull distribution** to be preferred to a semiparametric **Cox regression model**? The choice in this example may be based on how well the Weibull model fits, although other considerations, such as whether **predictions** of survival times are required, may also play an important role.

In smaller trials, the investigators may be reluctant to rely wholly on the randomization to produce comparable treatment groups. In these circumstances important prognostic variables can be used in an **analysis of covariance**. Decisions on which variables to include as **covariates** are usually not based on the data but have been identified a priori, in the trial protocol, where the considerations will have had a medical rather than statistical emphasis. This approach simplifies and strengthens the statistical analysis but it does not settle the matter entirely; for example, it may not be practical to specify, before the data are collected, the form in which the prognostic variables should enter the model. However, even in small studies, randomization should have produced groups that are fairly closely matched, and it is probably reasonable to view the model as providing a correction to a small imbalance, and the form of the variables is unlikely to be a major problem.

Although there are aspects of randomized controlled trials that may require some statistical modeling, it is of the utmost importance to remember that these are secondary considerations. The primary aim of the trial is to provide an estimate of the treatment effect and there is no direct interest in the model. If a model has been used in the analysis, then it is likely that some aspects of the model, such as the

inclusion of a variable the prognostic value of which is widely accepted, will be unchallenged, whereas other aspects, such as the inclusion of other variables, the form in which they are included (e.g. linear or quadratic), or the use of a Weibull distribution, may be less widely accepted. The trial will be weakened if the size of the estimate is crucially dependent on less well-founded aspects of the model; if the direction of the effect is dependent on these aspects, then the trial will be seriously undermined. In practical terms, the analyst can simply try a range of plausible models and observe their effect on the size and precision of the estimate of the treatment effect.

Modeling in trials is relatively straightforward, because the primary aim of the analysis is clear and because randomization should have relegated modeling to a subsidiary role.

Epidemiologic Studies

The discussion in this section is restricted to various types of **case–control** and **cohort studies**. Other forms of study could, undoubtedly, qualify under this heading and some of these, for example **screening** programs and **projections** of the number of AIDS cases, are considered in later sections.

In common with clinical trials, these studies usually have a clear aim that is readily summarized, usually by a **relative risk** or **odds ratio**. For example, they may be concerned with quantifying the risk of contracting some disease in those exposed to some hazard, relative to those not exposed. Unlike the situation in clinical trials, randomization cannot be used to produce groups that are comparable in all respects except exposure. Consequently, the investigator must rely on **matching** and on statistical modeling to adjust for differences in **confounding** variables. Deciding which variables to include in the model can be problematic, and some relevant issues of more general application will be touched on later. However, attempts to specify as much as possible of the model a priori can help. Also, considering the effect on important relative risks of using different plausible models is useful.

In a particular application it is likely that the problems of **variable selection** will be the main concern of the statistician. Apart, perhaps, from paying some attention to **additive** and **multiplicative** risks, the underlying form of the models will not be seriously questioned. This is because the considerations which have shaped these models tend to occur at a more general level than the individual study. **Poisson regression** or Cox regression for cohort studies and conditional or unconditional **logistic regression** for case–control studies arise because of the sampling schemes, general properties of **binary data** and statistical theory [5, 6]. In some areas of application, deeper, subject-specific justification of statistical models may be available, such as cancer epidemiology, in which **multistage carcinogenesis models** can provide guidance.

Modeling in General Biostatistics

The primary aims of randomized trials and epidemiologic investigations differ little from study to study. Other kinds of medical investigation, or even secondary analyses of trials or epidemiologic studies, can have much more diverse aims and, accordingly, the way in which models are chosen is much more varied. In choosing a model the analyst is guided by two distinct sets of considerations; namely, the purpose to which the model will be put, and the amount of quantitative information about the structure of the system being modeled. It is possible that quite different models of the same system may be needed for different purposes.

If we are trying to understand the relationship between some outcome, y, and other variables, written as a vector $\mathbf{x}$, then a common approach is to assume that

$$\mathrm{E}(y) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. This may be adequate, but there is usually no reason why this form of relationship should hold. Such empirical models inevitably fail to carry much conviction; models which are based on a deeper consideration of the underlying system are intrinsically more compelling. For example, predictions based on such models will be grounded not only in the data used to fit the model, but also in the theory which gave rise to it. Likewise, associations between y and an element of $\mathbf{x}$ may be missed by (1) because of its linear form. Even if an association is found using (1), it may have limited value because the mechanism of the association has not been elucidated. Finally, fitting a theoretically based model allows the theory to be examined and possibly extended.

The main obstacle to the high-sounding sentiments just expressed is that in most areas of biostatistics

there is no suitable theory, and even when a semblance of one does exist, the guidance it offers is often partial. To illustrate this, three examples of models that are, to varying extents, based in theory are presented.

Example 1: Fetal Mandible Length

Data relating fetal mandible length to gestational age [8] were analyzed by Appleton [3] using a model that related mandible size to the process of cell proliferation that gave rise to its growth. He took the length to be $k(P + Q)^{1/3}$, where P and Q represent the number of proliferative and nonproliferative cells, and which obey

$$\frac{\mathrm{d}P}{\mathrm{d}t} = 2\alpha\eta P, \qquad \frac{\mathrm{d}Q}{\mathrm{d}t} = 2\alpha(1-\eta)P,$$

where α is the cell birth rate and η is a decreasing function of Q. The biology has brought us so far, but the precise form of $\eta(Q)$ is not specified: this type of shortfall is not untypical. Nevertheless, the form for $\eta(Q)$ may not be crucial (Appleton used a negative exponential) and the model certainly provides insight into how parameters, *the biological meaning of which is clear*, affect the data.

However, from a statistical point of view the approach is difficult; parameters were estimated using an *ad hoc* procedure and centiles (*see* **Quantiles**) would be awkward to calculate. One of the aims of the initial investigators was to construct centile charts, so for this purpose the simpler, empirical approach of Royston & Altman [25] may be both adequate and preferable.

The above example illustrates clearly how model choice needs to be related to the purpose of the study, as well as to the data. Of course, there would be no objection to the use of the cell-proliferative model to produce centiles, had it been straightforward to do so. However, it is doubtful whether the analyst would wish to embark on an elaborate analysis, whatever its advantages, if a more familiar and simpler approach could provide the required solution. More generally, it almost goes without saying that model choice must acknowledge what is feasible: much of the **mathematical biology** that might be called upon to provide the theory to underpin models is cast in terms of differential equations, as this example illustrates. Unless these equations have closed-form solutions, present statistical theory is not well equipped to take advantage of the insights they may offer.

Example 2: Compartmental Models

The class of compartmental models provides a good illustration of the utility of a theory based on differential equations with a closed-form solution. Many biological systems, mostly in pharmacology and drug metabolism but also in areas such as dialysis medicine, can be represented as several notional compartments, with a substance of interest (e.g. a drug or metabolite concentration) diffusing between the compartments at rates proportional to the existing concentrations. This gives rise to linear differential equations with fixed coefficients the solutions of which are of the form

$$\text{concentration} = A\exp(-\alpha t) + B\exp(-\beta t) + \cdots,$$

where the number of exponentials is equal to the number of compartments in the model. There is no need to use the differential equations directly in the analysis of the data – the theory has been used to indicate that a model of the above form should be fitted. Moreover, the theory shows how biological meaning can be ascribed to functions of the parameters (*see* **Pharmacokinetics and Pharmacodynamics**). Although this kind of model has been very successful, it may represent a theory that is further from reality than the model in the preceding example. The compartments are usually notional parts of the anatomy of a patient, and often give rise to volumes of distributions that greatly exceed the volume of the patient. This illustrates that even successful models which possess a biological foundation may need some tolerance in their interpretation.

From a statistical viewpoint, compartmental models are susceptible to theoretical ambition: it is often much easier to extend a model by adding extra compartments than it is to collect data to sustain the resulting model. An example arises in dialysis medicine, where it is thought that creatinine clearance follows a two-compartment model, $A\exp(-\alpha t) + B\exp(-\beta t)$, with $\alpha \gg \beta > 0$. Any information on α is in the creatinine concentrations measured in the first 10–20 minutes of dialysis, but it is precisely in this period that measured creatinine concentrations are completely unreliable, due to the conditions of mixing that prevail in the early stages of dialysis. It is thus virtually impossible to fit the model that theory indicates is appropriate.

Example 3: Measurement of Cerebral Blood Flow

A more esoteric example concerns the measurement of cerebral blood flow rate using the Kety–Schmidt technique [17]. A low concentration of an inert gas (nitrous oxide) is introduced into the breathing mixture of the subject and the arterial and venous concentrations of the gas are each measured several times (about eight times) over the next 20–30 minutes. Both concentrations rise to the same equilibrium level, A, but the arterial level, $C_a(t)$ rises faster than the venous level, $C_v(t)$; this is illustrated in Figure 1, in which is shown both the fitted model and a typical set of data. Fick's principle of diffusion indicates that the cerebral blood flow rate can be calculated from these quantities as

$$K \frac{A}{\int_0^\infty C_a(t) - C_v(t)\,\mathrm{d}t},$$

where K is a known constant. A theory has led thus far, but more detailed specification is not provided. We know that $C_a(t) \geq C_v(t) > 0$, $C_a(0) = C_v(0) = 0$, and that both curves increase monotonically to a limit of A, but many functional forms are still open; for example,

$$C_x(t) = A[1 - \exp(-k_x t)] \qquad \text{or}$$

$$C_x(t) = A\left(1 - \frac{k_x^2}{(t + k_x)^2}\right).$$

In the absence of a specific prescription, choosing between these, or other models, must be based on whatever guidance – biological, practical, or statistical – is available.

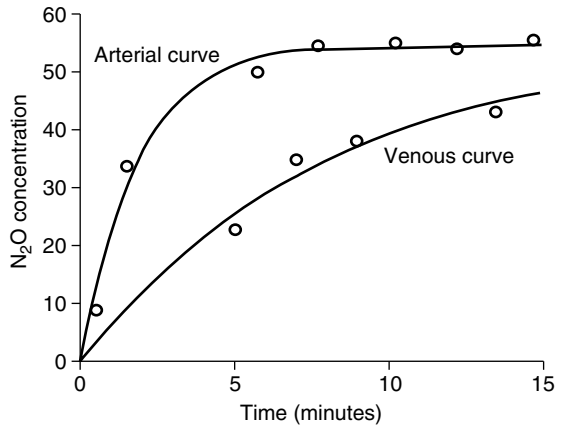

Figure 1

Kety [16] provides a detailed description of diffusion processes between different organs in the body that are related to, but not identical with, those involved in producing $C_a(t)$ and $C_v(t)$, and all of these use exponential functions. Although the arguments are not directly relevant, in the absence of other guidance, expressions for $C_a(t)$ and $C_v(t)$ based on exponential functions seem preferable.

Of course, whatever the nature of the background information on which a model is based, there is still a need to assess the performance of a model in practice. It is shown in Figure 1 that $C_a(t) = A[1 - \exp(-k_a t)]$ and $C_v(t) = A[1 - \exp(-k_v t)]$ provides a reasonable fit, so the estimated cerebral blood flow rate is found to be $K[k_a k_v/(k_a - k_v)]$, and this can be estimated by substituting estimates for the ks. When assessed over many more determinations, this model performs well, giving an example of a model that is specified by clear but incomplete theory and then supplemented by a number of *ad hoc* steps that are typical of the statistician's task in this type of analysis. A more detailed discussion of this example is available in Matthews et al. [21].

It is shown in Figure 1 that the model chosen in the final example fits the data well. Clearly, a model will not be acceptable unless it fits the data adequately, even if the model has a sound theoretical basis. However, how well a statistical model fits is often a relative matter, and it may well be sensible to use a model which has a sound basis and appears to offer a reasonable fit, in preference to an unfounded model which happens to offer a better fit to a particular dataset. Moreover, what constitutes an adequate fit is likely to be intimately bound up with the purpose of the analysis, and is an issue which can require careful judgment. Of course, determining *why* a theoretically based model does not fit can be a very informative exercise.

Models for Imputation

A rather different use of a model is to attempt to estimate some quantity which is not directly estimable on the basis of the collected, or collectable, data; three examples serve to illustrate this type of model.

Example 1: Lead Time in a Screening Program

One of the aims of screening a well population is to identify individuals with pre-clinical disease so that treatment can be instituted earlier; the time gained by use of a screening program (*see* **Screening, Overview**) is known as the *lead time*. Clearly, for any individual identified by the program as having the disease, the lead time cannot be observed. However, the mean lead time achieved may be of importance for assessing the value of the program.

Walter & Day [26] derive a method for estimating the mean lead time which involves postulating the density $f(\cdot)$ for the *sojourn time* of the disease in the population, which is the time during which the pre-clinical disease is potentially detectable by the program. An important component in the final choice of density is how well it fits the data, but it must also be recognized that the data may not be able to provide precise information on the form of f. As no fundamental guidance on the form of f is likely to be available, a choice that reflects features such as the generally skewed nature of waiting times must be made; Walter & Day considered **exponential**, **lognormal** and step functions. However, it is the mean lead time, not the form of f, that is of primary interest, and it is the sensitivity of this quantity, amongst sensible choices of f, that will determine how much reliance can be placed on the results of this analysis.

Example 2: Reporting Delays in Projection of Numbers of AIDS Cases

Amongst other things, the Cox Report [9] attempted to predict the number of **AIDS and HIV** cases likely to arise in the UK over a 2–5 year period, starting in 1988 (*see* **Projections: AIDS, Cancer, Smoking**). The prediction turned out to be highly influenced by the numbers of recently reported new cases. However, these cases were the ones that were most affected by delays in reporting to the Communicable Disease Surveillance Centre. To examine the effect of the delay, simple models of the delay process were constructed (see Appendices 7 & 8 of the Cox Report) and predictions on the basis of models incorporating this feature were proposed.

In both of these examples, calculations to permit the estimation of quantities of interest require the specification of the distribution $f(t)$ of some other quantity, such as sojourn time or reporting delay. In neither case is this distribution itself of primary interest nor is there any substantial guidance from the context of the application on the appropriate form of $f(t)$. Consequently the prudent analyst would pay attention to the effect of perturbations in the form of $f(t)$ on the quantities of direct interest.

Example 3: Proportion of Divided Cells in Thin Section

A slightly different form of imputation is reported by Wheeler and her colleagues [27]. Here interest centers on the proportion of labeled cells that divide at different times after labeling (*see* **Cell Cycle Models**). Labeled cells are visible under the microscope and it might be thought that cells which divide would appear as pairs of cells. However, counting such pairs suggested that only about 30% of all labeled cells eventually divided. It was thought that this unexpectedly low value (100% was expected) might be due to a geometrical artifact, namely that some cells divided in such a way that only one of the daughter cells was visible in the thin section of tissue which was scrutinized; see Figure 2.

A model was derived to take account of this feature; the model makes simple assumptions about the shapes of the cells and their relative positions after division and proceeds to use geometrical arguments to find an expression for

$$\Pr(\text{both daughter cells visible in section} \mid \text{cell has divided}).$$

Application of this formula indicated that the proportion of cells dividing was close to 100%. This model is, perhaps, less abstract than those in the previous examples but it does make several important simplifying assumptions. However, the simple model is probably quite sufficient. The main aim here is to investigate whether the artifact illustrated in Figure 2 is of sufficient magnitude to explain the discrepancy between observed and expected proliferation rates of

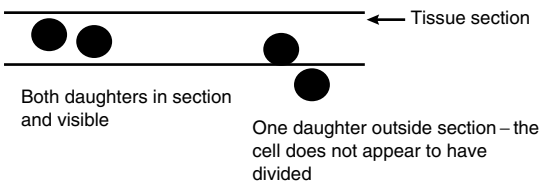

Figure 2

30% and 100%. Once a plausible model has shown that this is indeed possible, the histologist is unlikely to require information from more refined models.

Empirical Models

The preceding sections have discussed the way in which modeling can proceed when substantial guidance exists about the way the process being modeled "works". In medicine this is very much the exception rather than the rule. It is far more common that the statistician must assume that some quantity related to the outcome y and covariates x, usually $\mathrm{E}(y|x)$ or some function thereof, can be modeled as a smooth function of x. The rest of this section concerns models where the dependence on covariates is through a linear predictor, $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ represents unknown parameters that usually have to be estimated. Models of this form have proved very useful over the years, but their form can be rather restrictive, and alternative approaches, such as graph-theoretic models [11], **neural networks** and **tree-structured statistical methods** are becoming more widespread.

When constructing a model empirically, the form of the model is often largely dictated by whether the outcome is continuous, categorical, ordinal, and so on. Of course, many matters of detail arise, such as the need to transform a continuous variable or whether the variance or link function of a **generalized linear model** is correct. However, the main problem faced by the analyst is how to construct $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. It should be noted that this choice will generally be influenced by decisions made on such matters as **transformations** or link functions, and both these different facets of the model need to be considered and reconsidered as the modeling progresses.

Many studies in which an empirical approach to modeling is used are concerned with identifying associations between an outcome and a series of variables that are thought to affect the outcome. Other studies aim to build a model that is subsequently used to make predictions for new patients. Other kinds of study are possible and, because empirical modeling is used so widely, there are many approaches to the problems of **variable selection**: only a few general points are made here.

Studies seeking associations are very far from the well-specified models described in the earlier sections. Here it is likely that the investigator will have collected many covariates, with up to 50 or even 100 variables being commonplace. Such studies must be approached with caution, as many models will be possible, and little guidance is available as to when a satisfactory model has been found. In many cases it is prudent to await the collection of new data in order to assess the model; if the initial data set is sufficiently large, then the model might be determined on a random subset of the data and assessed on the remaining cases. This is especially important if the model is to be used for prediction [23].

Stepwise methods (*see* **Variable Selection**), such as forward selection or backwards elimination, are widely used, and have their uses. They aim to select the most important variables, with respect to some purely statistical criterion such as "proportion of variance explained", provided that the contribution of the variable can be distinguished from background noise. However, being chosen on solely statistical criteria, the resulting models may be difficult to interpret and may be at odds with existing knowledge. Miller [22] gives valuable information on the limitations of this approach; in particular, he points out that the estimates of $\boldsymbol{\beta}$ obtained after stepwise selection will be biased away from $\mathbf{0}$, sometimes by very substantial amounts.

The form of the model may be constrained by sensible pre-selection, perhaps by grouping variables according to type; for example, hematological, biochemical, performance status, and so on. Only simple summaries or representative members of each group are then entered into the model selection. Other *ad hoc* measures can be useful: if they have a common scale, highly **correlated** variables x_1 and x_2 might more profitably be entered as $\frac{1}{2}(x_1 + x_2)$, $x_1 - x_2$ and so on (*see* **Collinearity**).

Some attention needs to be given to the form in which variables are entered into the linear predictor. Categorical variables will be entered as **dummy variables**, as will ordinal variables (*see* **Ordered Categorical Data**), although in the latter case it will often be sensible to extract a trend that reflects the ordered nature of the categories (*see* **Trend Test for Counts and Proportions**). Continuous variables are usually entered in $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ as a linear term, but this may not always be appropriate. Non-linearity in the effect of a variable can be accommodated by first turning it into a categorical variable defined by suitably chosen "cut-points" (*see* **Categorizing Continuous Variables**). The following is a good example of how

such effects might be anticipated: the risk to a baby of infection with respiratory syncytial virus is affected by the age of the mother [24], but it is clear that the effect of a change from 16 to 20 is likely to be much greater than a change from 24 to 28, and a categorized version of maternal age may well be the most sensible approach. However, in general the pre-specification of cut-points is difficult, and there are considerable dangers with data-dependent specification [2]. Adding nonlinear terms is certainly a possibility, but polynomial terms often have undesirable properties and more imaginative functions, such as fractional polynomials [25], may be needed (*see* **Polynomial Regression**). Another approach is to use the data to determine the form of dependence: **generalized additive models** replace $x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p$ with $f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi})$, where the f_js are data-derived smooth functions [14].

In empirical modeling generally, and stepwise selection methods in particular, little attention is paid to the possibility of **interactions** between variables. The problems of selecting variables have already been stressed, and to entertain the possibility of interactions inevitably complicates matters. However, if attention is restricted to two-way interactions, and then only to those that a priori are strongly plausible, considerable improvements in the fit and realism of models can be achieved.

Modeling Variance

The modeling described thus far usually relates, either explicitly or implicitly, to the mean response, but modeling of other aspects of the response, in particular its **variance**, is certainly possible. Explicit modeling of both mean and variance, by linear predictors in two sets of covariates, namely $\mathbf{x}_i$ and $\mathbf{z}_i$, as

$$\mathrm{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \qquad \mathrm{var}(y_i|\mathbf{z}_i) = \exp(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\lambda})$$

is described by Aitkin [1]. Such models can be useful, for example when deriving centile charts, but are not widely used. In practice, many nontrivial variance structures do not arise because of explicit modeling but, rather, as an implicit consequence of the need to take account of some aspect of the structure of the data. A simple example of this is a generalized linear model, where the variance function reflects the choice of outcome distribution.

More complicated examples can be found in spatial statistics (*see* **Epidemic Models, Spatial**) and data from complex surveys (*see* **Sample Surveys in the Health Sciences**). Perhaps the most common example that arises in medical statistics is **longitudinal data analysis**, in which an individual is measured on successive occasions. The measurements on an individual are likely to be correlated, and any analysis which does not acknowledge this aspect of the data will be flawed; in particular, it is likely to exaggerate the amount of information in the data. To overcome this, the outcomes on an individual can be modeled by $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\mu}$ is the mean response, usually modeled in terms of some covariates, and the residuals $\boldsymbol{\varepsilon}$ have a dispersion matrix $\mathbf{V}$ which reflects the nonindependence of the responses. This approach may seem attractive, but can have practical drawbacks: if $\mathbf{V}$ is unstructured it will depend on many parameters, or a more **parsimonious** model, such as an **ante-dependence model**, will need to be chosen. In many applications the number of observations per individual may make this level of modeling difficult.

An indirect way of introducing dependence within an individual which has some intrinsic appeal is through **random coefficient** models (*see* **Random Effects**; **Multilevel Models**). Suppose that the outcome being measured is fetal heart-rate during the second stage of labor, and it is thought that this increases with time. A possible model is

$$y_{ik} = \alpha_i + \beta_i t_{ik} + \varepsilon_{ik},$$

where y_{ik} is the heart-rate of the ith fetus on the kth occasion that the fetus was measured, being at time t_{ik}. The residual terms ε_{ik} are assumed to be *independent* with variance σ^2. Rather than taking the parameters α_i and β_i to be fixed, they are now assumed to be realizations from a **bivariate distribution** with mean α and β and dispersion matrix

$$\begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}.$$

This model induces a covariance of $\sigma_\alpha^2 + \sigma_\beta^2 t_{ik} t_{ik'} + \sigma_{\alpha\beta}(t_{ik} + t_{ik'}) + \sigma^2$ between the kth and k'th outcomes on an individual, so a nontrivial dependence structure has emerged simply from the random coefficients. Models of this type have been used widely under a variety of names, but perhaps their most unified exposition and most powerful advocacy is given

by Goldstein [13], who calls them **multilevel models**. The rather specific forms of dispersion induced by these techniques may not be suitable for all applications but with sufficient ingenuity these can be amended, and in practice they have been found to be very flexible.

Multilevel models can also be adapted for use with noncontinuous outcomes. Hierarchical generalized linear models constitute an alternative line of development, which can accommodate a variety of dispersion structures and types of outcome. Further details of this class of models can be found in [19] and in the references therein.

The analysis of longitudinal data illustrates the following general point about the use of more elaborate models. It is important that the statistician should be convinced of the necessity of using more complicated models: simpler approaches will generally be easier to implement and, crucially, they may be much easier to explain to medical colleagues. The use of **summary measures** is a simple and readily understood approach to the analysis of longitudinal data, which will often circumvent the need for more complicated models of any description.

Some Modeling Pitfalls

Although statistical modeling has been very useful in many areas of medical statistics, the foregoing discussion has shown that selecting a model is far from an exact and precisely defined procedure. As such, it is important to keep in mind ways in which the process can go astray.

An obvious way in which things can go wrong is by selecting the wrong model: for example, if the data really follow the model $\mathrm{E}(y|x_1, x_2) = x_1^{\mathrm{T}}\beta_1 + x_2^T\beta_2$ and $\mathrm{E}(y|x_1) = x_1^{\mathrm{T}}\beta_1$ is fitted, then it is well known that the resulting parameter estimates will be **biased** (*see* **Misspecification**). Of course, there probably is no model which the data *really follow* and a good discussion of the practical attitude to a 'true model' is provided by Chatfield [7]. In most cases, the best we can hope for is a model that is consistent with what is already known and is adequate for the purpose to hand. This view of what forms an acceptable model is very much in keeping with Healy's notion of technology, where the practitioner is less concerned with some notion of "absolute truth" than with making the most of what truth is available.

Other, less philosophical, problems abound, and it would be impossible to give a complete catalog. However, two problems – namely, model uncertainty and mathematical coupling – are sufficiently general that they deserve special mention.

It will be clear from the preceding discussions that a good deal of uncertainty attends the process of deciding on the model, M, that will ultimately be used to fit the data. However, when presenting the results of model-fitting, it is customary to ignore this source of variability: so, for example, when quoting the uncertainty of parameter estimates, it is $\mathrm{var}(\hat{\beta}|M)$, not $\mathrm{var}(\hat{\beta})$ that is used. Ignoring the uncertainty in model selection can be very misleading, and there is much interest in ways in which this source of variation can be acknowledged [7, 12].

However, the appropriate level of formality necessary for this task needs careful consideration. Appleton [3] presents an example which is concerned with the estimation of cell growth rates at a given time; cross sectional data on numbers of cells at different times were available and several plausible growth curves were fitted. Although all curves appeared to provide reasonable fits to the data, the rate of growth differed substantially between the curves. A formal approach might attempt to amalgamate these estimates, and include the inter-model variation in the final interval estimate. However, presenting all the individual fits and rates may be more informative; this approach could lead the investigators to the disappointing but possibly sensible conclusion that their data were unable to answer their question. Both approaches are far preferable to simply selecting a model and providing an estimate with no further comment.

Another potential pitfall for the statistical modeler is mathematical coupling [4]. Essentially, it arises when attempts are made to relate an outcome variable to a covariate used in the definition of the outcome. For example, suppose that blood glucose concentration is measured at midnight (x) and at 0600 (y), and it is required to relate the change in concentration, $y - x$, to the initial value x: a spurious relationship can be caused by the presence of x on both sides of the equation. Related problems abound in analyses involving ratios, and are particularly hazardous for the unwary analyst [18] (*see* **Baseline Adjustment in Longitudinal Studies**).

No general solutions to either of these problems exist, but it is important that the statistician should

be aware of these and many other subtle traps that abound, and which can only be avoided by the exercise of shrewd judgment, as well as technical expertise. It is perhaps appropriate to end on this note, because much of this article has illustrated that many of the difficult problems encountered in modeling call on the statistician's judgment, common sense, and knowledge of the scientific discipline in which he works, and not just on his statistical expertise.

References

[1] Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM, *Applied Statistics* **36**, 332–339.

[2] Altman, D.G., Lausen, B., Sauerbrei, W. & Schumacher, M. (1994). Dangers of using optimal cutpoints in the evaluation of prognostic factors, *Journal of the National Cancer Institute* **86**, 829–835.

[3] Appleton, D.R. (1995). What do we mean by a statistical model?, *Statistics in Medicine* **14**, 185–197.

[4] Archie, J.P. (1981). Mathematic coupling of data, *Annals of Surgery* **193**, 296–303.

[5] Breslow, N.E. & Day, N.E. (1980). *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

[6] Breslow, N.E. & Day, N.E. (1987). *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[7] Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion), *Journal of the Royal Statistical Society, Series A* **158**, 419–466.

[8] Chitty, L.S., Campbell, S. & Altman, D.G. (1993). Measurement of the fetal mandible – feasibility and construction of a centile chart, *Prenatal Diagnosis* **13**, 749–756.

[9] Cox, D.R. (1988). *Short-term Prediction of HIV Infection and AIDS in England and Wales: Report of a Working Group*. HMSO, London.

[10] Cox, D.R. (1990). Role of models in statistical analysis, *Statistical Science* **5**, 169–174.

[11] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies – Models, Analysis and Interpretation*. Chapman & Hall, London.

[12] Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion), *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

[13] Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd Ed. Edward Arnold, London.

[14] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[15] Healy, M.J.R. (1978). Is statistics a science? *Journal of the Royal Statistical Society, Series A* **141**, 385–393.

[16] Kety, S.S. (1951). The theory and applications of the exchange of inert gas at the lungs and tissues, *Pharmacological Review* **3**, 1–41.

[17] Kety, S.S. & Schmidt, C.F. (1948). The nitrous oxide method for the determination of cerebral blood flow in man: theory, procedure and normal values, *Journal of Clinical Investigation* **27**, 476–483.

[18] Kronmal, R.A. (1993). Spurious correlation and the fallacy of the ratio standard revisited, *Journal of the Royal Statistical Society, Series A* **156**, 379–392.

[19] Lee, Y. & Nelder, J.A. (2001). Hierarchical generalized liner models: A synthesis of generalized linear models, random-effects models and structured dispersions, *Biometrika* **88**, 987–1006.

[20] Lehmann, E.L. (1990). Model specification: the views of Fisher and Neyman, and later developments, *Statistical Science* **5**, 160–168.

[21] Matthews, J.N.S., Matthews, D.S.F. & Eyre, J.A. (1999). A statistical method for the estimation of cerebral blood flow using the Kety-Schmidt technique, *Clinical Science* **97**, 485–492.

[22] Miller, A.J. (1990). *Subset Selection in Regression*. Chapman & Hall, London.

[23] Phillips, A.N., Thompson, S.G. & Pocock, S.J. (1990). Prognostic scores for detecting a high risk group: estimating the sensitivity when applied to new data, *Statistics in Medicine* **9**, 1189–1198.

[24] Pullan, C.R., Toms, G.L., Martin, A.J., Gardner, P.S., Webb, J.K.G. & Appleton, D.R. (1980). Breast-feeding and respiratory syncytial virus infection, *British Medical Journal* **281**, 1034–1036.

[25] Royston, P. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[26] Walter, S.D. & Day, N.E. (1983). Estimation of the duration of a pre-clinical disease state using screening data, *American Journal of Epidemiology* **118**, 865–886.

[27] Wheeler, J., Matthews, J.N.S. & Morley, A.R. (1991). Corrections in counting paired nuclei labelled with bromodeoxyuridine in tissue sections, *Cell Proliferation* **24**, 143–157.

(*See also* **Goodness of Fit**; **Hierarchical Models**; **Model Checking**; **Separate Families of Hypotheses**)

JOHN N.S. MATTHEWS

Molecular Epidemiology

Molecular epidemiology (ME) refers to the use of biomarkers in epidemiologic study designs, with emphasis on markers designed to measure exposure, to characterize host susceptibility, and to measure disease. Other terms involving closely related types of studies include biochemical epidemiology [16], pharmacogenetics [5], ecogenetics [10], and transitional studies [4], the last term referring to studies designed to bridge the gap between laboratory investigations and population studies. The first systematic description of the approach was provided by Perera & Weinstein [11]. They defined "molecular cancer epidemiology" as "advanced laboratory methods in combination with analytical epidemiology to identify at the biochemical or molecular level specific exogenous agents and/or host factors that play a role in human cancer causation". Consistent with the broad contribution of molecular biology to a more profound understanding of human disease, laboratory markers have been increasingly integrated into epidemiologic studies.

Objectives

Molecular epidemiology can be viewed as a synthesis involving the application of the methods of molecular biology to the study of disease on the population level. The contribution of molecular biology during the last third of the twentieth century has resulted in a redefinition of our basic understanding of human disease. **Population-based studies** have also grown in size and sophistication as the need to understand the cause of disease has been better appreciated, especially in an era where high costs for medical care indicate a need for expanded efforts at disease prevention.

Biological markers used in classical epidemiologic study designs can contribute to understanding **dose–response** relationships by assessing biologically effective dose, making interspecies comparisons, quantifying human interindividual variability, and identifying subsets at altered risk [13, 16]. In addition, biomarkers may provide more sensitive, specific, quantitative, or reproducible indications of study endpoints than traditional approaches, and therefore may in theory improve both study efficiency and validity. Such markers might provide early or specific indications of disease, and thereby identify a cancer at an earlier more treatable stage, or even in time for a preventive intervention. The mechanistic insight gained from biomarkers study may enhance disease understanding in profound ways.

Types of Biomarkers

Biomarkers are used to make three general types of measurements: (i) internal exposure, often measuring a compound of interest bonded to a macromolecule (i.e. hemoglobin or DNA, a critical "target"), but also including substances or their metabolites such as nicotine or its metabolite cotinine as a marker for exposure to tobacco smoke; (ii) host susceptibility factors, typically metabolic traits that are due to hereditary variation; and (iii) early biologic effects, mutations or cytogenetic damage – these are "effect" markers, that is indicators of disease or biological effects of pathologic significance.

Although for discussion purposes these categories are considered distinct, there is overlap. For example, detection of nicotine in the blood might be considered a marker of smoking (exposure), or an indicator of potential pathologic effects (early effect or disease marker), or might comprise part of a phenotype (e.g. ratio of nicotine to its metabolites) reflecting activity of a metabolizing enzyme (susceptibility factor).

Exposure Markers

The first category, exposure markers, offers the possibility of extending the reach of classic epidemiology beyond traditional questionnaire or external exposure monitoring. The actual dose of a compound of interest in the organism is assessed by the biomarker. Much attention has focused on measurement of adducts to DNA and other macromolecules, as it has been hypothesized that these might reflect both the relevant exposure, metabolic activation (generally considered an obligate step in carcinogenesis), and the actual quantity of compound that has reached a critical cellular target. In light of the target (i.e. DNA) involved, it is plausible that these compounds reflect a biologically relevant measure on the pathway to malignancy. The use of biomarkers is of course not new in the history of epidemiology. Examples include the use of polio antibody patterns to detect

Table 1 Conventional and molecular epidemiology

Feature	Conventional	Molecular
Use of biomarkers	Incidental	Systematic
Size	Varies, but often large	Varies, but typically small
Type of biomarker	Well established measures of exposure or disease	New or investigational markers of exposure, effect, or susceptibility
Cost	Low per subject	High per subject
Goals	Identify relationship between exposure and disease	Clarify mechanism and identify high-risk subgroups
Advantages	Historically the major method used to identify external cancer causes Public health orientation	Enter the "black box", i.e. elucidate mechanism Reduce interindividual variation Identify subsets at risk Potentially refine risk estimates Facile link to animal studies
Disadvantages	For certain cancers no cause is known, in spite of study Variable susceptibility is poorly understood Poor record at identifying reasons for individual susceptibility	Costly and complex Misclassification increased due to laboratory error Biomarker collection compromises validity Little public health benefit has resulted Ethical concerns

immunity. Some of the features of this approach that contrast with conventional epidemiology are indicated in Table 1.

"Susceptibility Markers" and Genetic Studies

The second category of markers involves host susceptibility factors, typically but not always genetic (hereditary) traits that control the metabolism of substances involved directly or indirectly with human disease. A susceptibility factor modifies the disease risk conferred by a specific exposure. It is further distinguished from exposure because susceptibility is generally preexisting and nontransitory. An early focus of molecular epidemiologic studies was a hypothesized genetic component in cancer. In contrast to earlier studies that searched for an obvious hereditary factor that accounted for the aggregation of specific cancers in families, these studies hypothesized that an influence of certain **genes** would exist for common apparently sporadic cancers in the general population as well.

Four studies in the 1980s laid the groundwork for the study of genetic susceptibility using epidemiologic study designs. This first generation of studies all involved a determination of a "metabolic phenotype", i.e. a pattern of metabolism of a test substrate that would reflect the underlying inherited genetic trait. In 1979, Lower et al. reported a relationship between the acetylation phenotype and incidence of urinary bladder cancer [8]. This early report was an early example of a study involving a relationship between a common genetic factor that controls the metabolism of an environmental contaminant and a disease in a case and control population. The term "molecular epidemiology" appeared in the title. The hypothesis was that "slow acetylators", a group comprising 50% of Western populations, would be less able to acetylate and thereby inactivate carcinogenic aromatic amine carcinogens. The second study examined aryl hydrocarbon hydroxylase (AHH) inducibility in relation to lung cancer [6]. The 20% of the population with the high inducibility phenotype were thought to convert carcinogens in tobacco to their active form at an accelerated rate, accounting for their increased risk of lung cancer. Ayesh et al.studied the relationship of debrisoquine metabolism (*CYP2D6*), to lung cancer [1]. A fourth genetic factor, glutathione *S*-transferase (*GSTM1*) was studied in relation to lung cancer. Subjects without this activity, who were presumably deficient in ability to eliminate carcinogenic epoxides from cigarette smoke, exhibited excess lung cancer risk [14]. All of these studies relied on a

laboratory measurement of a phenotype to infer the genetic susceptibility factor. In contrast, more recent studies directly identify the **genotype** through the study of an individual's DNA. Some of the advantages and disadvantages of the phenotype and genotype approaches are indicated in Table 2. Each of the four genetic traits initially approached with phenotypic measurements is under continued study today.

A major evolution over this period is the superseding of phenotype probe drug approaches by direct genotype assays. The genotype approach has distinct advantages and drawbacks (summarized in Table 2), but generally genotyping is increasingly the study approach of choice. Acetylation phenotyping, originally accomplished by administering sulfa drugs or caffeine, is now performed using a direct genotyping approach that detects the major mutations in the *NAT2* gene. The association of slow acetylators (*NAT2*-deficient subjects) with bladder cancer has been repeatedly observed in subjects with occupational exposure to arylamines. The debrisoquine phenotype can likewise be accurately detected by genotyping of the *CYP2D6* gene. The genotype determination is more complex since partially activating and inactivating mutations exist, and many minor variants must be tested. The degree of **association** of this trait with lung cancer is controversial. It appears most likely that elevated risk for smoking-related cancer in extensive metabolizers is limited to subjects with heavy smoking histories [2]. Investigators have attempted to understand the relationship of AHH activity to **polymorphisms** of both the *CYP1A1* and *Ah receptor* genes. The precise relationship of the gene polymorphisms to enzyme activity (and presumably to the ability to activate carcinogens in tobacco, thereby accounting for lung cancer risk) is incompletely understood, and may involve other genes such as the Ah receptor. Studies of the phenotype are subject to **bias** (see Table 2) but have often shown an effect, while genotype studies have been negative in Western studies, but positive in Japan. These differences may reflect the fact that both the mutation frequency (i.e. gene frequency) and mutation type (allelic heterogeneity) exhibit ethnic variation. Finally, a summary of the available literature suggests that individuals that lack *GSTM1* activity (i.e. without at least one functional *GSTM1* allele) exhibit consistently increased risk of both lung and bladder cancer (relative risk 1.2–1.6) [3] (*see* **Genetic Epidemiology**).

Effect Markers

Effect markers comprise the third category. These include nonspecific markers such as mutations in *Salmonella typhimurium* detected in urine or feces, various assays for chromosomal abnormalities (e.g. micronuclei, sister-chromatid exchange), as well as more specific findings, such as the specific chromosome translocations that characterize hematologic malignancies. These markers complement better

Table 2 The phenotype and genotype approaches in molecular epidemiology

Consideration	Phenotype	Genotype
Advantage	Approach is historically tested Reflects physiologic, *in vivo* disposition of drug Inducers, inhibitors, substrates all combine to reflect physiology	Identifies heterozygotes Simple, requires only germline DNA sample Unaffected by illness, diet, medications, etc. Can be performed with microquantities Noninvasive samples (i.e. DNA obtained from mouth wash, hair follicles, or standard blood sample)
Disadvantage	Numerous factors may distort measurements, e.g. drug–drug interaction More complex analysis Patient cooperation required Phenotyping protocols difficult to adapt to field study Time-consuming nature of test results in refusal to participate	Functional status of mutations may be unknown Risk of exposure to blood-borne pathogens Ethical questions arise since DNA may be used for other tests Allelic heterogeneity Ethnic differences

known histologic and cytologic markers such as dysplasia or increased mitotic frequency.

Related Approaches

Inborn errors of metabolism were described by Garrod in the early twentieth century and are distinct from the genetic traits of interest to ME in that they invariably produced phenotypic manifestations and clinical consequences. In genetic terms, the associated condition was fully **penetrant** given the genotype. Most of the genetic traits of interest in ME are not highly penetrant. In pharmacogenetics, an area closely related to ME, the phenotype is detected with laboratory probes and does not always result in clinical sequelae, i.e. the condition is not fully penetrant. Sequelae result after specific exposures, typically to pharmaceutical agents (but also xenobiotics, carcinogens, or endogenous compounds) whose metabolism is dependent upon the enzyme (or receptor, immune factor, or other element) that is subject to pharmacogenetic variability. Chronic conditions are thought to occur with altered frequency based on long-term exposures to specific agents subject to this type of variability.

Criticism

The field has attracted much attention and methodological critique. First, there are those who have questioned whether ME is a true subdiscipline with substantive new content [9]. While no one distinguishing factor can uniquely identify ME studies, the alteration of study design to allow the use of biomarkers is probably characteristic. A second area that has generated negative comment is the small size and inadequate attention to design issues in some studies. The size of certain studies has been constrained by the cost of specific assays. For example, a newly developed or expensive gas chromatography/mass spectroscopy assay could cost over $1000 per sample. Given this limitation it is axiomatic that proper design of a study should be a major concern in order to achieve maximum efficiency. Certain goals of ME such as identifying subsets of the population at elevated risk will necessitate large sample size. Some studies have placed emphasis on the "molecular" aspects but with minimal "epidemiology" or sophistication in the statistical treatment. Both the quality and the co-opting of the "epidemiology" label to cover such studies are unfortunate, but some bad studies do not invalidate the proper use of the approach. Some have also dismissed the "molecular" modifier for epidemiology as unnecessary, stating that it is improper to identify a component of epidemiology based on a measurement technique, compared with recognized fields of epidemiology such as "pediatric", "infectious", "occupational", or "clinical". Such a complaint seems churlish given that the growth of scientific inquiry does not follow a set pattern and nomenclature is anything but consistent.

A more basic issue to emerge is that some goals of ME may not be attainable with the designs being used. For example, some proponents of ME would like to estimate the **absolute risk** of cancer in an individual associated with a specific set of genetic factors. While the general direction of inquiry is laudable, the hospital-based case–control approach (*see* **Case–Control Study, Hospital-based**) often advocated is incompatible with this goal, and population-based case–control (*see* **Case–Control Study, Population-based**) or **cohort** designs will be required. In addition, much larger studies (i.e. thousands of subjects rather than hundreds) will be required to detect **gene–environment interactions** of medical interest. The idea that combinations of factors (i.e. genetic factors plus mutation load) will refine individual risk [15] challenges the traditional public health advocacy of epidemiology and refocuses emphasis on "individual" clinical risk in a way that is disturbing to many [7, 9]. The implicit reductionism of the approach raises worrisome ethical issues. In particular, the improper use of genetic information derived from these studies in an increasing concern.

Two critiques finally emerge as central. The first is that historically, all the major etiologic environmental factors known to cause cancer have been identified not through mechanistic or animal studies, but through **observational studies** in humans. Smoking and lung cancer is a paradigmatic example. Case–control and cohort studies unequivocally demonstrated the association of tobacco use and lung cancer a least a decade before Aurbach's smoking beagles were shown to exhibit characteristic preneoplastic changes in the respiratory tract (*see* **Smoking and Health**). Moreover, the history of epidemiology demonstrates that public health and prevention can be accomplished in the absence of detailed mechanistic understanding.

Secondly, it is difficult to demonstrate that specific biomarkers are superior to properly designed traditional questionnaire approaches to exposure assessment. A carefully designed smoking questionnaire (*see* **Questionnaire Design**) will demonstrate a stronger association of tobacco use with lung cancer then either serum cotinine (nicotine metabolite marker of recent smoking) or smoking-related carcinogen adducts of hemoglobin (e.g. 4-aminobiphenyl, a marker of intermediate exposure). It can be argued that these markers relate to recent or intermediate time periods, and the exposure that caused the disease is remote. Nevertheless, without a clear public health benefit, one might ask whether ME studies are worth the cost in time, resources, lost eligible subjects (some will refuse a biospecimen request), and new sources of bias. The roles of questionnaire information and biomarkers are complementary, and there are clearly settings where questionnaire approaches are not suitable. Nevertheless, the tacit assumption that biomarkers are universally superior requires critical scrutiny [12].

Summary

The advances of molecular biology have transformed science in the late twentieth century and may ultimately be more far-reaching than the revolutionary advances of physics in the first half of the century. It is inevitable that this understanding must permeate scientific investigation involving human disease, including the study of disease on the population level. Epidemiology provides the tools for such study, and it is fitting that these tools will be transformed and adapted to optimize the use of biomarkers. The trend towards increasing emphasis on laboratory methods in clinical medicine shows no signs of declining, and such approaches will find application in the studies that define both the diseases and the factors that cause them. Seen in this context, the growth of molecular epidemiology is both natural and inevitable. To achieve the full potential of the ME approach, however, investigators will need to pay close attention to issues of study design and quality control.

References

[1] Ayesh, R., Idle, J.R., Richie, J.C., Crothers, M.J. & Hetzel, M.R. (1984). Metabolic oxidation phenotypes as markers for susceptibility to lung cancer, *Nature* **312**, 169–170.

[2] Bouchardy, C., Benhamou, S. & Dayer, P. (1996). The effect of tobacco on lung cancer risk, *Cancer Research* **56**, 251–253.

[3] D'Errico, A., Taioli, E., Xiang, C. & Vineis, P. (1996). Genetic metabolic polymorphisms and the risk of cancer: A review of the literature, *Biomarkers* **1**, 174–177.

[4] Hulka, B.S. (1991). Epidemiological studies using biomarkers: issues for epidemiologists, *Cancer Epidemiology Biomarkers and Prevention* **1**, 13–19.

[5] Kalow, W. (1962). *Pharmacogenetics: Heredity and Response to Drugs*. W.B. Saunders, Philadelphia.

[6] Kellerman, G., Shaw, C.R. & Luyten-Kellerman, M. (1973). Aryl hydrocarbon hydroxylase inducibility and bronchogenic carcinoma, *New England Journal of Medicine* **289**, 934–937.

[7] Loomis, D. & Wing, S. (1990). Is molecular epidemiology a Germ Theory for the end of the twentieth century?, *International Journal of Epidemiology* **19**, 1–3.

[8] Lower, G.M., Nilsson, T., Nelson, C.E., Wolf, H., Gamsky, T.E. & Bryan, G.T. (1979). *N*-Acetyltransferase phenotype and risk in urinary bladder cancer: Approaches in molecular epidemiology. Preliminary results in Sweden and Denmark, *Environmental Health Perspectives* **29**, 71–79.

[9] McMichael, A.J. (1994). Invited commentary–molecular epidemiology: new pathway or new traveling companion?, *American Journal of Epidemiology* **140**, 1–11.

[10] Mulvihill, J.J. (1976). Host factors in human lung tumors, an example of ecogenetics in human oncology, *Journal of the National Cancer Institute* **57**, 3–7.

[11] Perera, F.P. & Weinstein, I.B. (1982). Molecular epidemiology and carcinogen-DNA adduct detection: New approaches to studies of human cancer detection, *Journal of Chronic Diseases* **35**, 581–600.

[12] Rothman, N., Stewart, W.F. & Shulte, P.A. (1995). Incorporating biomarkers into cancer epidemiology: a matrix of biomarker and study design categories, *Cancer Epidemiology Biomarkers and Prevention* **4**, 301–311.

[13] Schulte, P.A. & Mazzuckelli, L.F. (1991). Validation of biological markers for quantitative risk assessment, *Environmental Health Perspectives* **90**, 239–246.

[14] Seidegard, J., Pero, R.W., Miller, D. & Beattie, E.J. (1986). A glutathione transferase in human leukocytes as a marker for susceptibility to lung cancer, *Carcinogenesis* **7**, 751–753.

[15] Shields, P.G. & Harris, C.C. (1991). Molecular epidemiology and the genetics of environmental cancer, *Journal of the American Medical Association* **266**, 681–687.

[16] Vineis, P. & Caporaso, N. (1988). Applications of biochemical epidemiology in the study of human carcinogenesis, *Tumori* **74**, 19–26.

NEIL CAPORASO

Moment Generating Function

Moment generating functions are used to derive **moments** of distributions, establish the distributions of sums and differences of independent **random variables**, and derive limiting distributions of sequences of random variables. We say that a variable X with distribution function $F(x) = \Pr(X \leq x)$ has a moment generating function $M_X(t)$, i.e. the moment generating function exists if

$$M_X(t) = \mathrm{E}[\exp(tX)] = \int_{-\infty}^{\infty} \exp(tx)\, \mathrm{d}F(x) \quad (1)$$

is finite for any real number t in some open interval $-T < t < T$. For a continuous distribution with density function $f(x)$,

$$M_X(t) = \int_{-\infty}^{\infty} \exp(tx) f(x)\, \mathrm{d}x, \quad (2)$$

and for a discrete random variable taking values $a_1, a_2, \ldots, a_m$ with probabilities $p_1, p_2, \ldots, p_m$, respectively,

$$M_X(t) = \sum_{j=1}^{m} \exp(ta_j) p_j. \quad (3)$$

When it exists, $M_X(t)$ is a strictly positive and continuously differentiable function of t, for $|t| < T$. Moreover, $M_X(0) = 1$ and moments of any order exist. Using $M_X^{(r)}(t)$ to denote the rth derivative of $M_X(t)$ with respect to t, we may obtain the rth moment of X about the origin from

$$\mathrm{E}(X^r) = M_X^{(r)}(0) \quad \text{for } r = 1, 2, \ldots, \quad (4)$$

and a Taylor series expansion about the origin yields

$$M_X(t) = 1 + \sum_{r=1}^{\infty} \frac{\mathrm{E}(X^r) t^r}{r!}. \quad (5)$$

Moments of positive random variables involving non-integer powers may be obtained from a corresponding result derived by Cressie & Borkent [1]. Central moments, $\mu_r = \mathrm{E}[(X - \mu)^r]$, $r = 1, 2, \ldots$, may be obtained by evaluating derivatives of the central moment generating function,

$$\mathrm{E}\{\exp[(X - \mu)t]\} = \exp(-\mu t) M_X(t) = 1 + \sum_{r=1}^{\infty} \frac{\mu_r t^r}{r!}$$

at $t = 0$.

The existence of the moment generating function uniquely determines a distribution. If X and Y are random variables with respective distribution functions $F(x)$ and $G(y)$ and moment generating functions $M_X(t)$ and $M_Y(t)$, then $F(x) = G(x)$ for all real x if and only if $M_X(t) = M_Y(t)$ for all t in some open interval $-T < t < T$.

Another important use of the moment generating function is to establish the limiting distribution for a sequence of random variables. Let $X_1, X_2, \ldots$, denote a sequence of random variables with distribution functions $F_1(x_1), F_2(x_2), \ldots$, respectively, and suppose the moment generating function $M_{X_n}(t)$ exists for each X_n. Then, the pointwise **convergence** of $M_{X_n}(t)$ to some function $M_\infty(t)$ for all $|t| < T$ implies $F_n(x)$ converges to $F_\infty(x)$, the distribution function corresponding to $M_\infty(t)$, as $n \to \infty$, for all points x for which $F_\infty(x)$ is continuous. $F_\infty(x)$ will be a proper distribution, however, if and only if $M_\infty(0) = 1$. Conversely, the pointwise convergence of $F_n(x)$ to some limit $F_\infty(x)$ implies the pointwise convergence of the corresponding sequence of moment generating functions to the moment generating function for $F_\infty(x)$ for any t in any open subset of R in which the $M_{X_n}(t)$ are uniformly bounded.

A multiplicative property provides convenient derivations of sums or differences of independent random variables. If $X_1, X_2, \ldots, X_n$ are independent random variables, for example, then

$$M_{X_i - X_j}(t) = M_{X_i}(t) M_{X_j}(-t) \quad (6)$$

and

$$M_{X_1 + X_2 + \cdots + X_n}(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t). \quad (7)$$

We can establish the distribution of a difference or sum by recognizing the product on the right of (6) or (7) as the moment generating function for a specific distribution.

Koopmans [3] derived a basic probability inequality from the existence of the moment generating

function. For independent and identically distributed random variables $X_1, X_2, \ldots, X_n$, where the common moment generating function $M_X(t)$ exists for $|t| < T$, we have

$$\Pr(X_1 + X_2 + \cdots + X_n > 0) \leq [M_X(t)]^n$$

for any $0 < t < T$. Koopmans used this result to investigate **laws of large numbers**.

The joint moment generating function of a multivariate random variable $\mathbf{X} = (X_1, \ldots, X_k)$, defined as

$$M_{\mathbf{X}}(t) = \mathrm{E}\left[\exp\left(\sum_{j=1}^{k} t_j X_j\right)\right], \qquad (8)$$

is said to exist if it is finite for all $-T < t_j < T$, $j = 1, \ldots, k$. Mixed moments such as $\mathrm{E}(X_i^r X_j^s)$ may be obtained by differentiating (8) r times with respect to t_i and s times with respect to t_j, and evaluating the resulting derivative at the origin. Moment generating functions for marginal distributions are obtained by setting appropriate t_js equal to zero; for example,

$$M_{X_2}(t_2) = M_{X_1, X_2, X_3}(0, t_2, 0).$$

Random variables $X_1, X_2, \ldots, X_k$ are mutually independent if and only if

$$M_{\mathbf{X}}(t) = M_{X_1}(t_1) M_{X_2}(t_2) \cdots M_{X_k}(t_k).$$

For nonnegative integer-valued random variables, it is sometimes convenient to derive factorial moments

$$\mu_r{}^* = \mathrm{E}[X(X-1)\cdots(X-r+1)], \quad r = 1, 2, \ldots$$

from the derivatives of the factorial moment **generating function**

$$G_X(t) = \mathrm{E}(t^X) = \mathrm{E}\{\exp[X \ln(t)]\} = M_X[\ln(t)] \quad (9)$$

evaluated at $t = 1$. $G_X(t)$ is also called the probability generating function because it uniquely determines the probability function for X through the relationship

$$\Pr(X = r) = \frac{G_X^{(r)}(0)}{r!}.$$

Johnson et al. [2] and Smith [4] provide more extensive reviews of the uses and properties of various types of generating functions.

References

[1] Cressie, N. & Borkent, M. (1986). The moment generating function has its moments, *Journal of Statistical Planning and Inference* **13**, 337–344.

[2] Johnson, N.L., Kotz, S. & Kemp, A.W. (1992). *Univariate Discrete Distributions*, 2nd Ed. Wiley, New York.

[3] Koopmans, L.H. (1993). A note on using the moment generating functions to teach the laws of large numbers, *American Statistician* **47**, 199–202.

[4] Smith, W.L. (1992). Moment generating function, in *Encyclopedia of Statistical Sciences*, Vol. 3, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 372–376.

(*See also* **Characteristic Function**)

KENNETH KOEHLER

Moments

Moments are used to quantify and summarize the **mean** value, level of dispersion, and other features of a **probability** distribution. The rth moment of a **random variable** X about zero is simply the average value (or expected value) of X^r. It is defined by the Riemann–Stieltjes integral

$$\mathrm{E}(X^r) = \int_{-\infty}^{\infty} x^r \, \mathrm{d}F(x), \tag{1}$$

where $F(x) = \Pr(X \leq x)$ is the distribution function for X. We take r to be a positive integer, but this is not a requirement. When X is a continuous random variable with density function $f(x)$, we have

$$\mathrm{E}(X^r) = \int_{-\infty}^{\infty} x^r f(x) \, \mathrm{d}x.$$

For a discrete distribution on the nonnegative integers,

$$\mathrm{E}(X^r) = \sum_{k=0}^{\infty} k^r \Pr(X = k).$$

Moment formulas may also be obtained from derivatives of **moment generating functions** or from derivatives of **characteristic functions**.

The first moment about zero ($r = 1$) is the mean of the distribution of possible values for X. It is also called the expected value (or **expectation**) of X. We use the symbol μ to denote the mean.

The rth moment about a constant c is $\mathrm{E}[(X - c)^r]$. Taking $c = \mu$, we have the rth central moment

$$\mu_r = \mathrm{E}[(X - \mu)^r] = \int_{-\infty}^{\infty} (x - \mu)^r \, \mathrm{d}F(x), \tag{2}$$

which is also called the rth moment about the mean. For a discrete distribution on the nonnegative integers,

$$\mu_r = \sum_{k=0}^{\infty} (k - \mu)^r \Pr(X = k),$$

and for a continuous distribution with density function $f(x)$,

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) \, \mathrm{d}x.$$

The first central moment is zero by definition. The second central moment, usually denoted by σ^2, is called the **variance** of X, and its positive square root, $\sigma = (\mu_2)^{1/2}$, is called the **standard deviation**. As the square root of the average squared deviation from the mean, the standard deviation provides a measure of the level of dispersion in the distribution of the possible values of X. It can also be given a probability interpretation. When X has any **normal** (Gaussian) distribution, for example, the probability that an observed value for X lies in the interval $(\mu - \sigma, \mu + \sigma)$ is approximately 0.68 and the probability that it lies in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is approximately 0.95.

Central moments are obtained from moments about zero by the formula

$$\mu_r = \sum_{j=0}^{r} (-1)^j \binom{r}{j} \mu^j \mathrm{E}(X^{r-j}), \tag{3}$$

where $\binom{r}{j}$ is a binomial coefficient. For example,

$$\begin{aligned}
\mu_2 &= \mathrm{E}(X^2) - \mu^2, \\
\mu_3 &= \mathrm{E}(X^3) - 3\mu\mathrm{E}(X^2) + 2\mu^3, \\
\mu_4 &= \mathrm{E}(X^4) - 4\mu\mathrm{E}(X^3) + 6\mu^2\mathrm{E}(X^2) - 3\mu^4.
\end{aligned}$$

Inverse formulas are

$$\begin{aligned}
\mathrm{E}(X^2) &= \mu_2 + \mu^2, \\
\mathrm{E}(X^3) &= \mu_3 + 3\mu_2\mu + \mu^3, \\
\mathrm{E}(X^4) &= \mu_4 + 4\mu_3\mu + 6\mu_2\mu^2 + \mu^4.
\end{aligned}$$

Ratios of moments, or ratios of functions of moments, are also used to help characterize the level of dispersion and the shape of a probability distribution. The coefficient of variation, σ/μ, may be used to quantify variation in an assay method or some other measurement technique relative to the mean size of the quantity to be measured. It is often expressed as percentage by multiplying by 100%. We may describe the shape of a distribution with the **skewness** index

$$[\beta_1(X)]^{1/2} = \mu_3(\mu_2)^{-3/2} \tag{4}$$

and the **kurtosis** index

$$\beta_2(X) = \mu_4(\mu_2)^{-2} \tag{5}$$

Although these quantities may not uniquely determine the shape of a distribution, moments above the fourth order are rarely used to summarize properties of distributions.

If a random variable can only take values in a finite interval, then knowledge of all of the moments completely determines the distribution. This may not be true for some unbounded distributions, however, without imposing further constraints, but the most commonly used families of distributions, such as **Pearson distributions**, are characterized by no more than four parameters, which are usually uniquely determined by no more than four moments. We refer the reader to Johnson et al. [1, 2] for reviews of such families. We may estimate parameters from observed data by deriving a formula for each parameter as a function of the moments and substituting estimates of moments into the formulas (*see* **Method of Moments** estimation).

We define joint moments for an n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)$ as expectations of products. Quantities like $\mathrm{E}\left(\prod_{j=1}^{n} X_j^{r_j}\right)$, for example, are called product moments about zero. Central product moments (also called central mixed moments) are defined by

$$\mu_{r_1 r_2 \ldots r_n} = \mathrm{E}\left[\prod_{j=1}^{n} (X_j - \mu_j)^{r_j}\right], \tag{6}$$

where $\mu_j = \mathrm{E}(X_j)$. A special case,

$$\mathrm{cov}(X_j, X_k) = \mathrm{E}[(X_j - \mu_j)(X_k - \mu_k)] \tag{7}$$

is called the covariance of X_j and X_k (*see* **Covariance Matrix**). We can characterize the strength of the linear relationship between X_j and X_k with the **correlation** coefficient

$$\rho_{jk} = \frac{\mathrm{cov}(X_j, X_k)}{\{\mathrm{E}[(X_j - \mu_j)^2]\,\mathrm{E}[(X_k - \mu_k)^2]\}^{1/2}}, \tag{8}$$

which achieves its bounds of 1 (or -1) when all possible values for (X_j, X_k) lie on a straight line with a positive (or negative) slope. When X_j and X_k are mutually independent, then $\mathrm{cov}(X_j, X_k) = \rho_{jk} = 0$, but the converse is not always true.

Other types of moments are sometimes considered. The rth *absolute moment* about zero, $\mathrm{E}(|X|^r)$, and the rth absolute central moment, $\mathrm{E}(|X - \mu|^r)$, are examples. For integer valued discrete random variables, it may be convenient to derive *factorial moments* instead of central moments or moments about zero. The rth descending factorial moment of a random variable X is defined as

$$\mu_{(r)} = \mathrm{E}[X(X-1)\ldots(X-r+1)], \tag{9}$$

the expectation of a product of r incrementally decreasing terms, for $r = 1, 2, 3, \ldots$. Moments about zero are obtained from descending factorial moments through the formula

$$\mathrm{E}(X^r) = \sum_{k=1}^{r} \frac{\mu_{(k)}}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^r. \tag{10}$$

In particular,

$$\begin{aligned}
\mathrm{E}(X) &= \mu_{(1)}, \\
\mathrm{E}(X^2) &= \mu_{(2)} + \mu_{(1)}, \\
\mathrm{E}(X^3) &= \mu_{(3)} + 3\mu_{(2)} + \mu_{(1)}, \\
\mathrm{E}(X^4) &= \mu_{(4)} + 6\mu_{(3)} + 7\mu_{(2)} + \mu_{(1)}.
\end{aligned}$$

Cumulants are an alternative to moments that have historically played a prominent role in the study of probability distributions. While moments about zero are obtained from coefficients in a power series expansion of the characteristic function of a distribution, cumulants are obtained from coefficients in a power series expansion of the natural logarithm of the characteristic function. Cumulants have some convenient theoretical properties, but moments are more frequently used in practical applications. We refer the reader to Stuart & Ord [3] for additional definitions and formulas and a good account of the relationships between various types of moments and cumulants.

References

[1] Johnson, N.L., Kotz, S. & Balakrishnan N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd Ed. Wiley, New York.

[2] Johnson, N.L., Kotz, S. & Kemp, A.W. (1992). *Univariate Discrete Distributions*, 2nd Ed. Wiley, New York.

[3] Stuart, A. & Ord, J.K. (1987). *Kendall's Advanced Theory of Statistics*, Vol. 1, 5th Ed. Griffin, London.

KENNETH KOEHLER

Monte Carlo Methods

Monte Carlo techniques are used in situations in which the analytic solution of the problem is either intractable or time consuming. Instead of calculating exact quantities, simulation is used to produce stochastic approximations to the solution.

In practice, Monte Carlo methods are discussed interchangeably with **simulation**. Useful discussions can be found in a variety of monographs, such as [7, 8], and [9], to which the reader is referred for a thorough review.

Monte Carlo techniques are used in a wide range of problems. Common uses are the construction of Monte Carlo **hypothesis tests**, **bootstrap** distributions, and for **numerical integration** in **Bayesian** calculations. More specialized uses are the analysis of complex stochastic systems.

Monte Carlo techniques have a long history in mathematics. An early example is Buffon's Needle dating from 1733, described in [7]. Realistically, the widespread application of Monte Carlo techniques has only become feasible with the availability of cheap and efficient computing since the 1970s. In fact, the exponential increase in available computing power has led to Monte Carlo techniques facilitating major statistical advances.

As a trivial example of a Monte Carlo method we consider the calculation of the mean μ of a density $g(x)$. The analytic solution is

$$\mu = \int x g(x)\, \mathrm{d}x,$$

which may be difficult to evaluate. In the Monte Carlo approach, we sample k observations, $X_1, \ldots, X_k$, from $g(x)$ and form the Monte Carlo estimate

$$\hat{\mu} = \frac{\sum_{i=1}^{k} X_i}{k}.$$

A **law of large numbers** can be used to show that this estimate will converge strongly to the true value, provided the integral exists (*see* **Convergence in Distribution and in Probability**). Thus, we can produce an estimate of a required accuracy by manipulating k.

An important point to note is that the term Monte Carlo does not refer to a particular stochastic **algorithm**, only to the fact that a stochastic algorithm has been used. For instance, in the previous example there is an infinite number of different stochastic algorithms that we could use to sample the observations and estimate μ.

The choice of algorithm depends on a variety of factors. In many problems there is a "natural" formulation. For example, we estimate the mean of a distribution by taking the mean of a sample from the distribution. Another criterion that should be considered when choosing an algorithm is the **efficiency** of the method. This leads to the contemplation of variance-reduction techniques, which are used to produce more accurate estimates for the same computational effort (*see* **Simulation**).

In practice there is a tradeoff between the use of an inefficient algorithm that is easy to design and implement and an efficient algorithm that could take detailed analysis to design. Provided that the inefficient algorithm is not pathological, it is sufficient merely to run the algorithm for an appropriately longer period of time, to obtain results comparable to those with the efficient algorithm. An important caveat is that great care must be taken to ensure that satisfactory convergence has occurred. In any application the relevant literature should be consulted.

We now briefly describe several important examples of the use of Monte Carlo techniques.

Monte Carlo Hypothesis Testing

The first technique considered is that of Monte Carlo **hypothesis testing**. Assume we have a statistic T, and a simple null hypothesis H_0 and composite alternative H_a. Consider constructing a test of size $1 - \alpha$. If the sampling distribution of $T|\mathrm{H}_0$ is known and analytically tractable, then we can use standard calculus methods to construct a **critical region** for rejecting H_0.

Alternatively, we consider constructing a Monte Carlo test of H_0. Specification of the rejection region involves the calculation of a **quantile** of the null distribution of T. In the Monte Carlo approach we estimate the required quantile stochastically. To do this we sample from the distribution $h(t)$ of $T|\mathrm{H}_0$, to produce $T_1, \ldots, T_k$. Depending on the alternative, we reject H_0 if the observed t is more "extreme" than

$100(1-\alpha)\%$ of the combined simulated values and the observed t.

As a simple example we consider the one sample t test that the mean of a normal distribution is μ against the alternative that it is less than μ. If we observe a sample $X_1, \ldots, X_n$, then our statistic is

$$T = \frac{\overline{X} - \mu}{\text{se}},$$

with

$$\text{se} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n(n-1)},$$

the standard error of the mean. Using the classical approach, we calculate that the rejection region under the null is $t < t_{\alpha,n-1}$, where $t_{\alpha,n-1}$ is the α quantile of **Student's *t* distribution**, with $n-1$ degrees of freedom.

Under a Monte Carlo approach we could sample $T_1, \ldots, T_k$ from a t distribution with $n-1$ degrees of freedom. We would then reject H_0 if T was less than $T_{[\alpha]}$, where $T_{[\alpha]}$ is the α quantile of $(T_1, \ldots, T_k, T)$. Thus we use the sample from the t distribution to estimate the α quantile.

Alternatively, we can consider the problem of estimating the P value of the test. In the example given, this is equivalent to finding the value of the integral

$$\int_{-\infty}^{t} h(x)\,\mathrm{d}x,$$

which is consistently estimated by

$$\hat{P} = \frac{\sum_{i=1}^{k} I(T_i \le t)}{k+1},$$

where $I(T_i \le t)$ is the indicator function for the event $T_i \le t$, and $\hat{P}$ is the Monte Carlo estimate.

In practice, the Monte Carlo approach to hypothesis testing is advantageous if the distribution of the chosen test statistic is intractable but it is convenient to sample from this distribution. Some practical examples can be found in [1].

Jöckel [6] has investigated the properties of Monte Carlo hypothesis tests under quite general conditions and produces the following conclusions. First, if we choose large enough samples (k) to produce our Monte Carlo quantities, then the results will be equivalent to the analytic results, with high probability. Secondly, the power of the Monte Carlo test is an increasing function of k.

Bootstrap

The second example we will consider is the bootstrap [2]. The bootstrap is a simple but powerful idea that is applicable to a wide range of problems. The central idea of the bootstrap is the following [3]. Suppose we wish to estimate some functional θ of a distribution $F(\beta; x)$,

$$\theta = \int g(x)\,\mathrm{d}F(\beta; x).$$

The bootstrap estimate is found by replacing the unknown population distribution in the integral with the sample's empirical distribution, as follows:

$$\hat{\theta} = \int g(x)\,\mathrm{d}\hat{F}(x),$$

where $\hat{F}(x)$ is the empirical distribution of the observed sample.

Now consider estimating the variability of this statistic. To estimate the variance of the sampling distribution of our estimate, we again replace the population distribution with the sample distribution, to give

$$\hat{\sigma}^2 = \mathrm{E}_{\hat{F}}[\hat{\theta} - \mathrm{E}_{\hat{F}}(\hat{\theta})]^2,$$

where $\hat{\sigma}^2$ is called the bootstrap estimate of the sampling variation of $\hat{\theta}$, and $\mathrm{E}_{\hat{F}}$ denotes expectation over the empirical distribution of the sample. This calculation could be performed analytically, and requires tabulating a complicated set of permutations. Instead, we can form a Monte Carlo estimate by generating draws of the same size as the original sample from the empirical distribution of the sample. For each such draw we calculate this statistic. We iterate this procedure k times to obtain $\theta_1^*, \ldots, \theta_k^*$, and then calculate

$$\hat{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k}(\theta_i^* - \overline{\theta}^*)^2,$$

where $\overline{\theta}^*$ is the mean of the bootstrap samples.

This is equivalent to drawing with replacement from the original sample, which is how the bootstrap is sometimes described.

Complex Systems

Another example of the use of Monte Carlo methods is in the analysis of complicated systems. Simulation methods are often used to explore the properties of complex systems, **likelihoods**, or estimators. Deterministic simulation systems are frequently used in areas such as finance where the impact of varying input parameters is assessed by changing the values in a **spreadsheet** [10].

Simulations involving stochastic components are often called Monte Carlo simulations. In many situations, uncertainty or randomness plays a key part in the dynamics of the system. Examples include queuing for operations and the demand for inventory items. Mathematical models can be proposed and analytical solutions can be sought, but in any realistic formulation the mathematics quickly becomes intractable.

Epidemic theory has been a fertile area for Monte Carlo simulations (*see* **Epidemic Models, Stochastic**). A recent epidemic that has received considerable attention is the HIV/AIDS (*see* **AIDS and HIV**) epidemic. Monte Carlo methods have been used to build realistic behavior patterns into a model for this epidemic and to explore its evolution based on those assumptions. Monte Carlo has also been used as an estimation technique. Often, good information exists for some aspects of the model. The other parameters of the model are varied until the results of the simulation agree best with the observed data.

An example of this methodology was given in [4] and [5]. Data from homosexual men in San Francisco provided input parameters for several aspects of the model. The model was a compartmental one with different levels of sexual activity, immigration, emigration, and death. Data on the probability of transmission related to sexual activity collected from several other studies were used to complement the primary data set.

The simulation study generated incidence projections for HIV and AIDS. A **sensitivity analysis** was conducted to see the range of behavior variables that were consistent with the observed incidence patterns. The conclusion was that there was a range of parameter values which could generate the observed data. The most important outcome of the analysis was the flagging of the most influential parameters for further collection of data.

Markov Chain Monte Carlo

The final example we will consider is the use of **Markov chain Monte Carlo** (MCMC) techniques such as Gibbs sampling. These techniques use stochastic simulation to explore and summarize a multivariate density which may not be analytically tractable. This problem arises often in the area of Bayesian statistics.

With MCMC techniques, standard results [9] are used to construct **Markov chains** with the required stationary distributions. Popular choices are variants on the Metropolis–Hastings algorithm. From an initial state vector, the chain is simulated until the sequence has reached approximate stationarity. This is the Monte Carlo part of the algorithm. The resulting sequence is used to make stochastic approximations to any required functional of the multivariate distribution.

References

[1] Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, New York.

[2] Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Annals of Statistics* **7**, 1–26.

[3] Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

[4] Hethcote, H.W., Van Ark, J.W. & Karon, J.M. (1991). A simulation model for AIDS in San Francisco: II. Simulations, therapy and sensitivity analysis, *Mathematical Biosciences* **106**, 223–247.

[5] Hethcote, H.W., Van Ark, J.W. & Longini, I.M. (1991). A simulation model for AIDS in San Francisco: I. Model formulation and parameter estimation, *Mathematical Biosciences* **106**, 203–222.

[6] Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests, *Annals of Statistics* **14**, 336–347.

[7] Morgan, B.J.T. (1984). *Elements of Simulation*. Chapman & Hall, London.

[8] Ripley, B.D. (1987). *Stochastic Simulation*. Wiley, New York.

[9] Tanner, M. (1993). *Tools for Statistical Inference*. Springer-Verlag, New York.

[10] Winston, W.L. (1991). *Operations Research Applications and Algorithms*, 2nd Ed. PWS-Kent, Boston.

(*See also* **Computer-intensive Methods**)

T.J. O'NEILL, S.C. BARRY & BOREK PUZA

Moran, Patrick Alfred Pierce

Born: July 14, 1917, in Sydney, Australia.
Died: September 19, 1988, in Canberra, Australia.

Pat Moran was an Australian statistical scientist with an enormous breadth and depth to his research interests. During his prolific career, Moran made substantial contributions to **population genetics**, medical statistics (particularly **psychiatry**), geometric probability, mathematical statistics, **time series**, and applied probability. He received numerous scientific honours, and now is commemorated by two Australian awards for young research statisticians.

After graduating in 1937 with First Class Honours in Mathematics from Sydney University, Moran went to Cambridge, England, where in 1939 he took Part III of the Mathematics Tripos. World War II interrupted further studies, and initially Moran had a job with the Ministry of Supply where he worked with D.G. Kendall and M.S. Bartlett. Then he joined the Australian Scientific Liaison Office in London, where reporting on all manner of wartime matters developed his clear and concise writing style as well as his breadth of scientific interests. Over these years he came to appreciate the importance of statistical methods; in fact he began reading **M.G. Kendall's** *Advanced Theory of Statistics* during the flying bomb raids on London. At the end of the war, Moran accepted the Baylis research studentship at St Johns College, Cambridge, to do a Ph.D. in Mathematics. He never did complete a Ph.D. (a fact which later on he would relate with pride), instead preferring a reasonable income to support his wife and then their children. From 1946 to 1951 Moran was a Senior Research Officer at the Institute of Statistics, Oxford, and was attached to Balliol College. He lectured at Trinity College (1949–1951) and was made a University Lecturer in 1951.

In 1952, Moran was appointed to the first Chair of Statistics at the Australian National University (ANU), which was established postwar as a research and graduate training institution in Canberra. He held this position until his retirement thirty years later. Although Moran preferred to maintain a small department, nevertheless he attracted first-rate faculty and visitors, as well as many graduate students, and the Department became "the cradle of modern Australian Statistics" [2]. From this position, Moran's influence on statistical science in Australia was profound: in an obituary Hall [2] noted that "ten out of the present seventeen professors of Statistics in Australia have been associated with the Department as either students or staff".

Moran's first papers written during the war were on Hausdorff measure (from his earlier research at Cambridge) and on convex sets. One of these was motivated by a problem posed by the Bomb Fragmentation Committee, and the resulting principle was used much later with a scanning beam electron microscope [2]. After the war, Moran published papers on **rank correlation**, and at Oxford became interested in the analysis of animal populations. In Canberra, Moran initiated the study of dam and storage system theory, publishing a monograph in 1959 which later was translated into Russian and Czech. Geometrical probability was an enduring interest from his war years, and his 1963 monograph with M.G. Kendall [3] was later translated into Russian. Some of his papers in this area were stimulated by problems posed by immunologists at the John Curtin School of Medical Research, ANU, such as the determination of the random pattern of antibodies attached to a spherical virus. Moran did not begin his influential research in genetics until the late 1950s, producing a major monograph [4] in 1962, giving a systematic account of the mathematical aspects of the genetics of natural populations (also translated into Russian) (*see* **Genetic Epidemiology**). Moran's 1967 [5] **probability** text was written to provide an outline of probability theory which can be both "generalized in a highly abstract manner" and can be used to describe many "complicated phenomena in natural science". This treatise lies usefully between an elementary introduction and deep abstract analysis, and was reprinted in 1984. Moran's list of over 180 publications can be found at [1], supplemented at [2].

Moran was a Roman Catholic, and much enjoyed turning his intellectual powers to theological discussions with clergy. He was a modest person who thought deeply about many matters, with his familiar pipe never far away. His criticism was always direct, and his insightful comments were a source of stimulation for many researchers, in statistics and mathematics, biology, and medicine.

References

[1] Gani, J. & Hannan, E.J., eds (1982). Essays in Statistical Science, Special Volume 19A of *Journal of Applied Probability*, pp. 1–6.

[2] Hall, P.G. (1989). Obituary: Patrick Alfred Pierce Moran, *Biometrics* **45**, 687–692.

[3] Kendall, M.G. & Moran, P.A.P. (1963). *Geometrical Probability*. Griffin, London.

[4] Moran, P.A.P. (1962). *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

[5] Moran, P.A.P. (1967). *An Introduction to Probability Theory*. Clarendon Press, Oxford.

SUSAN R. WILSON

Morbidity and Mortality, Changing Patterns in the Twentieth Century

The twentieth century has been a period of unprecedented gains in longevity and health status. At the turn of the twentieth century, **life expectancy** at birth in Europe, North America, and Australia and New Zealand was typically around 45–50 years, similar to levels prevailing in Africa today, and not much greater than the levels of 35–40 years which had prevailed in Europe for centuries. As the twentieth century draws to a close, life expectancy in most industrialized countries is of the order of 75–80 years, or even higher for females in some countries. In other words, life expectancy has increased by more than 50% over the last 100 years or so in the industrialized world, but these gains have not been enjoyed equally by all population groups. The twentieth century has seen the emergence of dramatic inequalities in survival, notably between men and women, but also between the better educated and poorer sectors of society.

In this brief review of 100 years of epidemiologic history, trends and differentials in mortality will be presented to the extent that data are available to document them, and the emergence (or decline) of major epidemics and endemic conditions will be discussed. Much of the analysis will be limited to mortality data since this is the most comprehensive, comparable, and unambiguous source of information on health status. With the exception of cancer registries (*see* **Disease Registers**) for some (generally subnational) populations, and **surveillance** sites for vascular events included under the **World Health Organization** (WHO) MONICA project, there are no comparable, standardized data on morbidity from which comparative trend analyses can be made. (MONICA is a 10-year (1984–94) epidemiologic surveillance system established in 35 countries, mostly industrialized, to monitor vascular disease incidence and mortality in defined populations, hence the name *Moni*toring of Cardiovascular Diseases and Risk Factors.) Equally importantly, there are no comparable data to investigate whether the extra years of life gained, particularly at older ages, have been accompanied by a rise, or fall, in disability. Data sources for assessing disability vary among countries and even within a population over time. Therefore little can be said, with any confidence, about changing patterns of disability, although this information is clearly required to assist policy and program formulation.

Data Sources for Mortality

Vital registration of births and deaths is the most useful source of mortality data for populations where complete recording of events has been achieved (*see* **Vital Statistics, Overview**). Where the death certificate includes diagnosis of the underlying **cause of death**, certified by a registered medical practitioner in accordance with the principles and procedures of the revision of the **International Classification of Diseases (ICD)** currently in force (*see* **Death Certification**), the information can also be used for epidemiologic assessments. Complete (or virtually complete) vital registration exists for industrialized countries, including Eastern Europe and the former USSR. In addition, several countries in developing regions of the world, including Argentina, Chile, Cuba, Mexico, Singapore, Uruguay, and some countries in the Caribbean have virtually complete registration and medical certification of deaths. Of the developing regions, medical certification of deaths is most advanced for Latin America and the Caribbean (43% of deaths), and least advanced in Sub-Saharan Africa (1% of deaths) [2].

Even in the absence of reliable vital registration data, patterns and levels of mortality can still be usefully ascertained through less expensive systems covering a sample of the population. For example, China has established a network of 145 Disease Surveillance Points (DSP) which record over 50 000 deaths annually in a population of 10 million people, representative of mortality conditions throughout China. Causes of death among the rural population in India are assessed via a "verbal autopsy" system operating from 1300 primary health care centers throughout the country. ("Verbal autopsy" is a method for diagnosing the approximate cause of death through a structured interview with relatives of the deceased. The interview is usually administered by a nonmedical person some weeks or months after death. Relatives are asked about a series of symptoms prior to death from which a diagnosis of the cause of death is made, preferably by a qualified physician.) While not as

reliable as the DSP system in China, the Indian data are nonetheless useful for delineating broad cause of death patterns throughout the country [2].

Issues of Comparability

The interpretation of analyses of vital registration data on causes of death must be made with caution since the comparability of data is undoubtedly affected by many factors (*see* **Mortality, International Comparisons**). Even among the developed nations, where causes of death over the course of the twentieth century have generally been classified according to standards and principles agreed upon by various international committees (since 1948, under the auspices of WHO), variations in diagnostic practice among countries affect data comparisons. A WHO international comparative study carried out in the 1960s reported significant variations in certification practices among six European countries with major differences in the proportion of deaths which were certified by pathologists [10]. No doubt these differences have diminished in recent decades, but the practice of autopsy still varies substantially among industrialized countries, with implications for data comparability [11].

Cultural differences also no doubt are a significant factor in the coding of injuries. Suicides in particular are undoubtedly underreported in some countries owing to the social or religious stigma associated with the act, with the death in such cases usually being coded to accidental injuries [6]. Studies have also revealed "diagnostic preferences" for chronic diseases. For example, in the 1950s and early 1960s, deaths which were coded to chronic respiratory diseases in the UK, Australia, and New Zealand may well have been coded to a cardiovascular disease in the US [5].

The statistical comparability of cause of death data has certainly been affected by the successive revisions of the ICD. The most profound change occurred with the introduction of the Sixth Revision around 1950, in which major alterations to the format of the list of causes were made to accommodate the sweeping changes in the principles of cause of death classification introduced with that revision. The introduction of the Eighth Revision in 1968 substantially affected the **time series** comparability of certain major causes of death, particularly ischemic heart disease (IHD), with up to 15% more deaths being coded to IHD than to the most comparable cause in the Seventh Revision [9].

The comparability of epidemiologic analyses of mortality data is also very much affected by the extent to which deaths are coded to ill-defined and unspecified diagnoses. This affects both comparisons among countries as well as trends within a country over time. For example, around 1950, approximately 15%–20% of all deaths in Belgium, France, Greece, Poland, and Spain were coded to ill-defined causes, compared with 1%–2% in Australia, Austria, Canada, Denmark, New Zealand, Switzerland, the UK, and the US. Without adjustment for diseases coded to ill-defined conditions, cross-national comparisons might be very misleading [4]. By the early 1990s, ill-defined causes throughout the industrialized world had declined to about 1%–3% of all deaths. For countries where this practice was common four decades ago, time series analyses for specific causes (especially cardiovascular diseases) must be interpreted with great prudence.

Trends in Life Expectancy and Age-Specific Mortality Rates

Life Expectancy

Life expectancy at birth is a convenient summary index of prevailing mortality conditions at each age. The measure is not a linear function of age-specific death rates and hence equal reductions in mortality at different ages will not have an identical impact on life expectancy – a reduction in death rates at younger ages will result in a larger gain in life expectancy at birth than a similar reduction in death rates at older ages. Despite this feature, life expectancy is widely understood and is perhaps still the most commonly used indicator to summarize overall mortality levels in a population.

Table 1 provides an overview of the gains in life expectancy in selected countries over the course of the twentieth century. By far the largest absolute increase has been enjoyed in Japan (34 years for males, 39 years for females), followed by Italy. Much less progress has been registered in Eastern Europe although national trends are not strictly comparable owing to differences in time period, **life table** methodologies, and population coverage around the turn of the century. What the table does suggest, however, is that the pattern of mortality reduction

Table 1 Life expectancy at birth, 1900–95, selected countries

(a) Males

Country	Life expectancy at birth in				
	1900–10	1930–40	1950–55[a]	1970–75[a]	1990–95[a]
Japan	42.4	47.9	62.1	70.6	76.4
Sweden	56.6	62.1	70.4	72.1	75.4
Australia	57.6	63.6	69.9	68.4	74.7
Spain		47.2	61.6	70.2	74.6
Netherlands		63.7	70.9	71.1	74.4
Italy	43.0	53.4	64.3	69.2	74.2
Norway	56.4	62.6	70.9	71.4	73.6
UK	45.3		66.7	69.0	73.6
France	43.4	55.2	63.7	68.6	73.0
Denmark		61.4	69.6	70.9	72.5
New Zealand	58.0	63.3	67.5	68.7	72.5
US	45.6	57.6	66.2	67.5	72.5
Poland			58.6	67.0	66.7
Hungary			61.5	66.5	64.5
Russian Federation	30.9	40.4	62.5	63.1	61.7

(b) Females

Country	Life expectancy at birth in				
	1900–10	1930–40	1950–55[a]	1970–75[a]	1990–95[a]
Japan	43.7	50.7	65.9	76.2	82.5
Sweden	59.5	64.2	73.3	77.5	81.1
France	47.0	59.8	69.5	76.3	80.8
Australia	61.4	67.3	72.4	75.2	80.6
Italy	43.7	55.5	67.8	75.2	80.6
Spain		50.8	66.3	75.7	80.5
Netherlands		65.0	73.4	77.0	80.4
Norway	59.3	65.8	74.5	77.6	80.3
US	48.3	61.0	72.0	75.3	79.3
UK	49.3		71.8	75.2	78.7
New Zealand	59.9	66.0	71.8	74.8	78.6
Denmark		63.3	72.4	76.4	78.2
Poland			64.2	74.1	75.7
Hungary			65.8	72.4	73.8
Russian Federation	33.0	46.7	70.5	73.5	73.6

[a]Annual average.

among the industrialized countries is extremely heterogeneous and that, while significant progress has been achieved, there has been more divergence than convergence among the industrialized countries.

Viewing trends in life expectancy over almost a century can conceal significant time trends that have characterized this century of mortality change. Life expectancy at birth has not increased monotonically since the early 1900s. Rather, significant gains were achieved virtually everywhere until the beginning of the 1950s. From the mid-1950s, male life expectancy stagnated, or even declined modestly in some Western European countries, as well as in Australia and North America (see Figure 1). Meanwhile, female life expectancy continued to rise. From the mid-to-late 1960s, male life expectancy began to rise again, quite sharply in some countries. At about the same time, one of the most remarkable reversals in life expectancy began throughout Eastern Europe with male life expectancy declining by up to 4–5 years in most countries, and by even more (7 years) in Russia [8]. Even this general trend has been accompanied by significant national variations with male life expectancy beginning to show signs of a renewed rise in the Czech Republic, Hungary, and Poland, but deteriorating markedly in Russia since 1988, having risen sharply in the former USSR between 1980 and 1987 [8].

Demographers and epidemiologists have been studying these remarkable changes in Eastern Europe for more than two decades. Most would agree that the trends are real and not due to sudden changes in the completeness or accuracy of reporting of deaths following widespread social and political change in the late 1980s. Epidemiologic research suggests that much of the pervasive increase in male mortality in Eastern Europe since the late 1960s is due to tobacco usage [3] or, in the case of the recent dramatic mortality increases in Russia, alcohol abuse [7].

Age-Specific Mortality Change

Given the very close relationship between the age-pattern of mortality and the prevailing cause of death structure – infectious diseases tend to kill many more children than adults, while chronic diseases do the converse – the conquest of the communicable diseases such as tuberculosis, measles, malaria, diarrheal diseases, and acute respiratory infections, which had largely been completed in the industrialized countries by 1950 or thereabouts, has resulted in massive declines in **infant mortality** and child mortality, and a significant reduction of death rates among young adults.

Around 1900, infant mortality rates in Australasia, Europe, or North America typically hovered around 100–150 deaths per 1000 live births, similar to levels currently prevailing in many African countries (see Table 2). Overall child mortality rates (measured as the probability of a newborn infant dying before age 5) were of the order of 200 per 1000 live births. In 1995, infant mortality rates in the industrialized countries are typically around 10 or less per 1000 live births, reaching as low as 4–5 per 1000 in Japan and parts of Scandinavia.

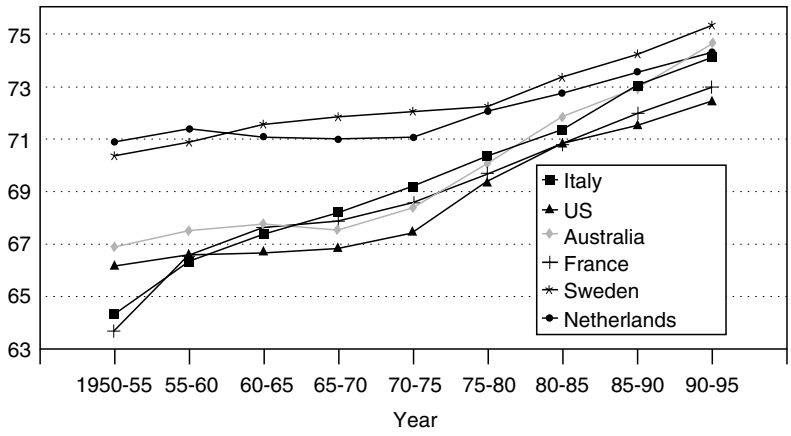

Figure 1 Trends in life expectancy at birth in selected countries, 1950–95, males

Table 2 Infant mortality rate (per 1000 live births)

Country	1900	1930	1950–55	1970–75	1990–95
Japan	168M/147F		51	12	4
Finland			34	12	5
Iceland			21	12	5
Sweden	78M/63F	62M/47F	20	10	5
Belgium			45	19	6
Germany			51	21	6
Switzerland		61M/48F	29	13	6
Australia	74M/59F	45M/36F	24	17	7
Austria			53	24	7
Canada		105M/82F	36	16	7
Denmark		89M/69F	28	12	7
France		80M/63F	45	16	7
Ireland			41	18	7
Luxembourg			44	16	7
Netherlands		54M/41F	24	12	7
Spain		150M/130F	62	21	7
UK			29	17	7
Italy	176M/158F	112M/100F	60	26	8
Norway	75M/61F	50M/41F	23	12	8
Czech Republic			43	20	9
Israel			41	23	9
Malta			75	22	9
New Zealand	80M/69F	44M/36F	26	16	9
USA	162M/133F	73M/58F	28	18	9
Greece		98M/97F	60	34	10
Portugal		196M/166F	91	45	10
Slovokia			73	24	12
Bulgaria			92	26	14
Hungary			71	34	15
Poland			95	27	15
Yugoslavia			110	47	20
Russian Federation			98	28	21
Romania			101	40	23

Although less dramatic, and with some interruptions for men as noted earlier, adult mortality levels have also declined more or less continuously since the beginning of the twentieth century. This decline has been much more evident for women, although evidence of a stagnation in mortality rates for women in some Eastern European countries first became evident in the early 1970s. Since the late 1970s, several countries, including Australia, the Netherlands, the UK, and the US, have seen further substantial reductions in male (and female) mortality, both in middle and old age. Much of this recent decline in death rates can be attributed to further declines in ischemic heart disease and stroke mortality, continuing a trend which began in the late 1960s.

From this brief analysis, one may conclude that adult mortality rates tend to be higher in populations with higher overall mortality and tend to decline (more or less monotonically) with declines in general mortality levels. This is perhaps counterintuitive, but is confirmed by recent global mortality analyses which suggest that the risk of adult death throughout the developing world is substantially higher than in Australasia, North America, Japan, and Western Europe [2].

Demographers have developed methods to decompose or disaggregate changes in life expectancy at birth into contributions due to changes in mortality at different ages. These can be either positive contributions (in which case mortality rates in the age group have declined), or negative, whereby life expectancy has increased despite an increase in mortality rates in a given age group. To illustrate the utility of these methods, the age pattern of contributions to life expectancy trends for males in three populations, Australia, England and Wales, and Hungary, are shown in Table 3 for the period 1950–79.

The very substantial contribution (in years) from post-1950 declines in infant and child mortality is evident in all three populations (1.37 years out of a 3.0 year increase in life expectancy in England and Wales, 4.32 years of a 4.9 year increase in Hungary, and 1.48 years out of a 4.2 year gain in Australia) [1]. Since mortality rates at ages 15–34 years were already comparatively low in the early 1950s, further declines at these ages did not contribute greatly to changes in life expectancy. Reductions in mortality for higher age-groups resulted in similar absolute contributions to increasing life expectancy in both England and Wales, and Hungary, at least until the mid-1960s.

Given the abrupt cessation of overall male mortality decline in Hungary and neighboring countries from the mid-1960s and the rapid increase in male life expectancy in Australia (and other Western countries) since the early 1970s, analyses for these subperiods are also presented in the table.

The complexity of age patterns of mortality change and their influence on overall life expectancy is well illustrated by these two examples which, in many respects, are representative of recent mortality trends in the industrialized countries. In Hungary, male life expectancy remained unchanged between 1960–64 and 1975–78 (but declined subsequently). This was due to rises in mortality (i.e. negative contributions to life expectancy) at all ages 25 years and older, and particularly at ages 45–54 years. Conversely, further declines in infant and child mortality, and, to a much lesser extent, at ages 15–24 years, acted to increase life expectancy but were exactly counteracted by rising death rates at older ages. The pattern of mortality change in Australia over the same period was exactly the reverse. Between 1950–54 and 1970–76, male life expectancy hardly changed at all (up by 1.3 years), almost all of which (0.9 years) was due to declines in infant and child death rates, with only small (positive or negative) contributions from relatively stable adult death rates. During the 1970s, however, male death rates at ages 45 and over declined dramatically in Australia, accounting for 1.9 years of the 2.9 year increase in life expectancy at birth, with much of the remainder (0.6 years) being due to continued declines in infant and child mortality.

Sex Differentials in Mortality

One of the most remarkable features of twentieth century mortality decline in the industrialized countries has been the dramatic widening of male–female differentials in mortality. Around the turn of the century, life expectancy for females was typically 2–3 years higher than for males, and in some countries, such as Ireland and Italy, the gap was less than 1 year. Female mortality rates exceeded those of males in many countries at various ages up to the end of the childbearing period. Indeed, the contribution of **maternal mortality** to the sex mortality differential around the turn of the century was sufficiently high

Table 3 Age components of changing life expectancy at birth, selected nations, 1950–54 to 1979

Country	Life table periods	Sex	Contribution (in years) to change in life expectancy at birth due to mortality trends at ages								Interaction	Total increase in life expectancy (in years)[a]
			0–14	15–24	25–34	35–44	45–54	55–64	65–74	75+		
UK: England	1950–54	M	1.37	0.06	0.21	0.20	0.25	0.42	0.24	0.19	0.09	3.0
and Wales	to	F	1.16	0.16	0.29	0.21	0.21	0.35	0.64	0.77	0.20	4.0
	1975–78	F–M	−0.21	0.10	0.08	−0.01	−0.04	−0.07	0.40	0.58	0.11	1.0
Hungary	1950–54	M	2.83	0.33	0.29	0.30	0.41	0.27	0.18	0.18	0.16	4.9
	to	F	2.50	0.37	0.38	0.29	0.33	0.44	0.42	0.30	0.22	5.3
	1960–64	F–M	−0.33	0.04	0.09	−0.01	−0.08	0.17	0.24	0.12	0.06	0.4
	1960–64	M	1.49	0.08	−0.02	−0.30	−0.54	−0.27	−0.31	−0.12	0.01	0.0
	to	F	1.33	0.07	0.07	0.00	−0.06	−0.01	0.17	0.18	0.02	1.8
	1975–78	F–M	−0.16	−0.01	0.09	0.30	0.48	0.26	0.48	0.30	0.01	1.8
Australia	1950–54	M	0.87	0.05	0.13	0.10	0.08	0.10	−0.01	−0.01	0.01	1.3
	to	F	0.79	0.06	0.17	0.20	0.25	0.27	0.41	0.49	0.09	2.7
	1970–74	F–M	−0.08	0.01	0.04	0.10	0.17	0.17	0.42	0.50	0.08	1.4
	1970–74	M	0.61	0.07	0.02	0.19	0.30	0.52	0.55	0.48	0.17	2.9
	to	F	0.35	0.03	0.07	0.18	0.28	0.39	0.60	0.95	0.18	3.0
	1979	F–M	−0.26	−0.04	0.05	−0.01	−0.02	−0.13	0.05	0.47	0.01	0.1

Source: [1]

[a]The age components and the interaction contribution may not sum exactly to the total due to rounding.

to reduce the female advantage in life expectancy over males by 0.3 to 0.5 years [1]. Conversely, accidents and violence were a major cause of male excess mortality, typically accounting for about half of the female advantage in life expectancy at birth.

By far the largest contribution to the increase in male excess mortality by the mid-1960s was the diverging mortality trends for men and women from cardiovascular diseases, at least in Australia and the US. A similar pattern is also evident in the Scandinavian countries (and indeed in most other industrialized nations) after 1950.

The contribution of cancer to widening sex mortality differentials is rather more complex. Prior to about 1930, female mortality from cancer exceeded that for men, largely due to cancers of the genital tract. By the mid-1960s, about half a year had been added to the gap in life expectancy in Australia and Scandinavia due to differential male–female trends from cancer, and almost a full year in the US. Much of this, and subsequent increases after 1960, can be attributed to the massive increase in male lung cancer mortality (see next section). Finally, it is also interesting to observe the growth in the contribution of male excess mortality from motor vehicle accidents. Around 1910, there were too few cars for this to be a significant cause of death. Since then death rates from car crashes rose dramatically, especially for males, so that by 1964, this cause alone contributed about half a year to the gap in life expectancy between the sexes.

Sex differentials in mortality have continued to widen in recent decades with the result that average life expectancy at birth for females is currently typically about 6–7 years higher than for males, and in some countries (e.g. Hungary and France) the gap is closer to 9 years. In others, e.g. Australia and the UK, there is evidence that the sex differential in mortality is no longer increasing. This is due to the very substantial declines in male mortality from lung cancer, ischemic heart disease, and stroke in these countries, following widespread reductions in smoking by men which began several decades ago.

Cause of Death Trends

The twentieth century has been characterized by a massive decline in **communicable diseases** and maternal and perinatal causes in industrialized countries, and increasingly in many developing countries as well. The extent of this reduction is well illustrated in Figure 2, which shows the **proportionate mortality** for males in selected countries from various broad causes over the last 100 years or so. The pattern for females is broadly similar, with the added feature that maternal deaths have declined from around 2%–5% of female deaths in the early 1900s to less than one-tenth of 1% today. From causing about 30% of deaths around 1900, infectious and parasitic diseases now cause less than 5% of deaths in the industrialized countries, and this figure would be even lower were it not for the **AIDS** epidemic. Noncommunicable diseases have emerged as the leading causes of death by far as the twentieth century draws to a close, despite the very substantial reductions in vascular disease mortality in some developed countries in recent decades.

The trends in accidents and violence (external causes) are particularly interesting. Although the proportionate contribution of these nonmedical causes to overall mortality has remained relatively constant at around 6%–8%, the composition of causes within the category of accidents and violence has changed dramatically. For example, earlier in the century, industrial accidents were the leading cause of male deaths from external causes, reflecting the risks associated with several occupations commonly practiced at that time (*see* **Occupational Mortality**). Subsequently, with the modernization of the labor market and legislative reform for occupational safety, these accidents have greatly diminished in frequency. For males, at least, motor vehicle accidents have emerged as the principal cause of death from nonmedical causes, rising from virtually zero around 1900 to account for about 30% of violent deaths among males in many industrialized countries, and an even higher proportion (50% or so) of violent deaths at the young adult ages (15–34 years) (see Figure 3).

An overview of mortality change in the industrialized countries since 1950 is given in Figure 4, which shows the *relative* change (death rate in 1950–54 = 100) in age-standardized death rates from selected leading causes of death for men and women separately (*see* **Standardization Methods**). The graph shows the average experience of 22 industrialized countries and demonstrates the varied epidemiologic history of the world's richest countries over the last

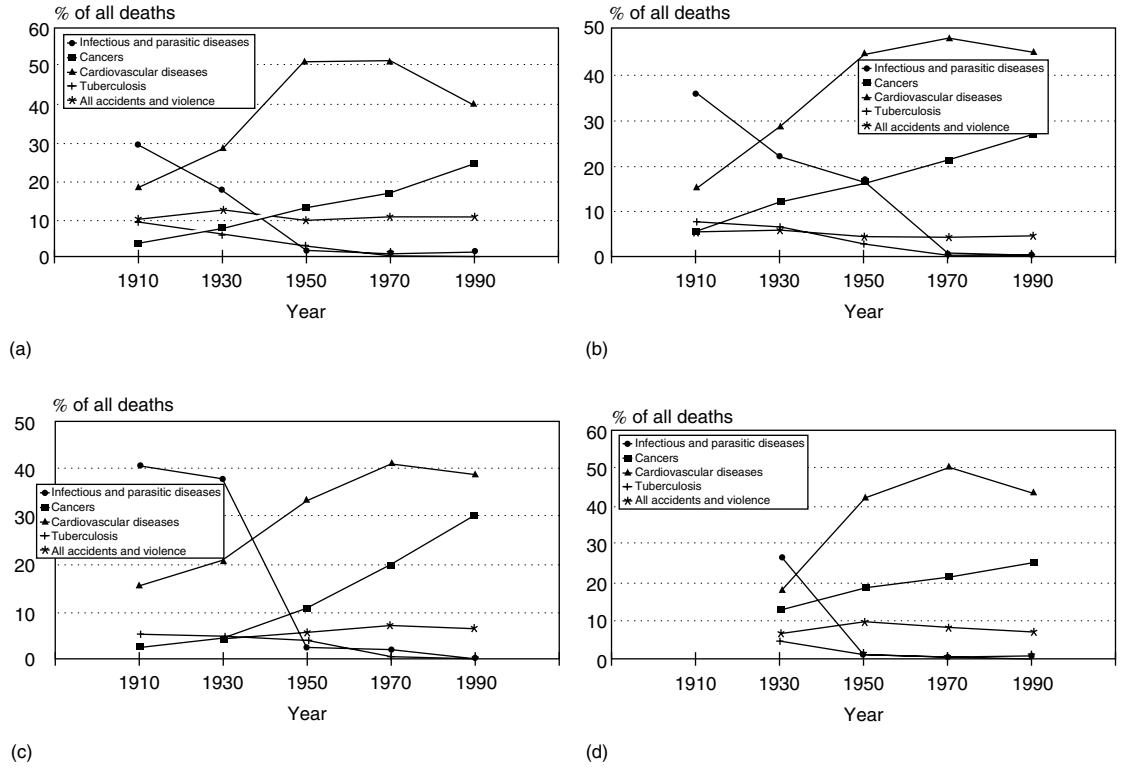

Figure 2 Proportionate mortality (in %) from broad causes, selected countries, males, 1910–90. (a) US, (b) England and Wales, (c) Italy, and (d) Denmark

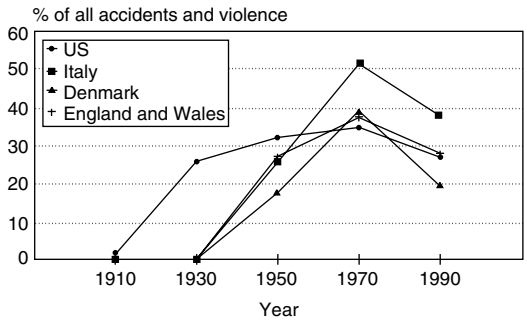

Figure 3 Proportion of all violent deaths due to motor vehicle accidents, selected countries, males, 1910–90

few decades. The rise (for men) and then steady decline in ischemic heart disease and stroke mortality is clear, as is the peak in motor vehicle accident mortality in the early 1970s. Since then, death rates from traffic crashes have returned to levels last seen in the 1950s for women, and 25% lower than the 1950–54 level for males. This has occurred despite a dramatic increase in the number of motor cars. This remarkable reversal is due to a number of factors, including improved highway conditions and stricter measures to control drunken driving in these countries.

But perhaps the most dramatic change in mortality since the middle of the century has been the extraordinary growth in lung cancer mortality, for both males and females. Male lung cancer rates have increased, on average, by almost 200% since 1950–54, while for females, the rise, in relative terms at least, has been even greater (more than 300%). Even though the relative increase in lung cancer rates has been higher for females, the absolute level of rates is still much higher in males owing to their longer smoking history. The enormity of the lung cancer epidemic in the industrialized countries during the course of the twentieth century is perhaps best summarized by the trends for the US (see Figure 5).

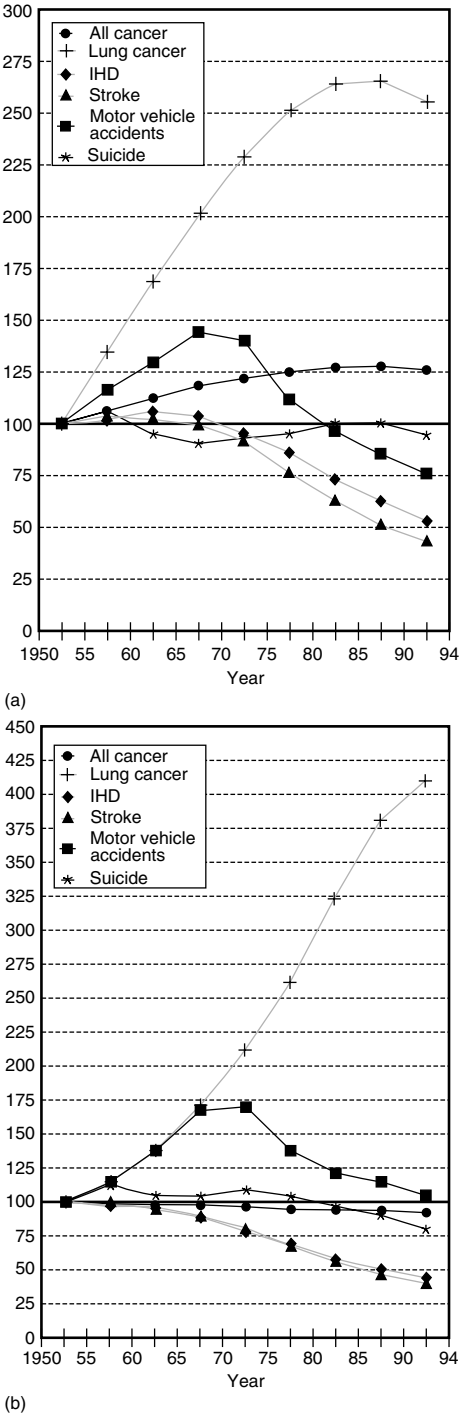

Figure 4 Relative change in mortality (1950–54 = 100) for selected causes of death in 22 industrialized countries, 1950–54 to 1990–94, (a) males, (b) females

From a level of around five deaths per 100 000 in 1930, US male lung cancer rates have risen about 15-fold to peak in the early 1990s. Other cancers have remained relatively stable, or, in the case of stomach cancer, declined substantially. For women, the rise in lung cancer only began in the early 1960s, some decades after American women began to smoke in large numbers.

There have been notable successes in reducing lung cancer mortality, particularly in Australia, Finland, the Netherlands, and the UK, where death rates from the disease have been steadily declining and are now at levels 20%–40% below their peaks. Male lung cancer rates in some industrialized countries are still rising, most notably in Japan (an increase of over 1000% since 1950), but also in Greece, Hungary, Portugal, Poland, and Spain. Indeed, lung cancer mortality in Hungary in 1994 reached 122 deaths/100 000 population (age-standardized), exceeding even the highest level reported for UK men (111/100 000) at the height of their epidemic (1974) [11].

Lung cancer death rates for women are rising everywhere except in Australia, New Zealand, and the UK, where death rates appear to have stabilized. The highest mortality rate for women in the early 1990s is reported for American women (38/100 000), closely followed by Denmark (36/100 000). Indeed, in several populations (see Figure 5), lung cancer now exceeds breast cancer as the leading site for mortality from the disease (*see* **Smoking and Health**).

Along with the reversal in lung cancer rates for men in some countries, the other great public health success of the second half of the twentieth century has been the extraordinary decline in ischemic heart disease mortality and stroke (cerebrovascular disease). Beginning in the mid-to-late 1960s, death rates from these diseases began to decline following a decade or more of rising rates in many countries. Death rates are now less than half their post World War II peak levels and are still declining. Largely as a result of these declines in major vascular diseases, overall mortality levels have fallen by up to 40% in many Western countries.

The other major disease for which significant progress in reducing mortality has been achieved is cirrhosis of the liver. In countries such as Australia, France, Germany, Portugal, and Spain, male death rates from the disease rose steadily during the 1950s and 1960s and reached a peak level of 45–55 deaths/100 000 population in the mid-1970s. Since

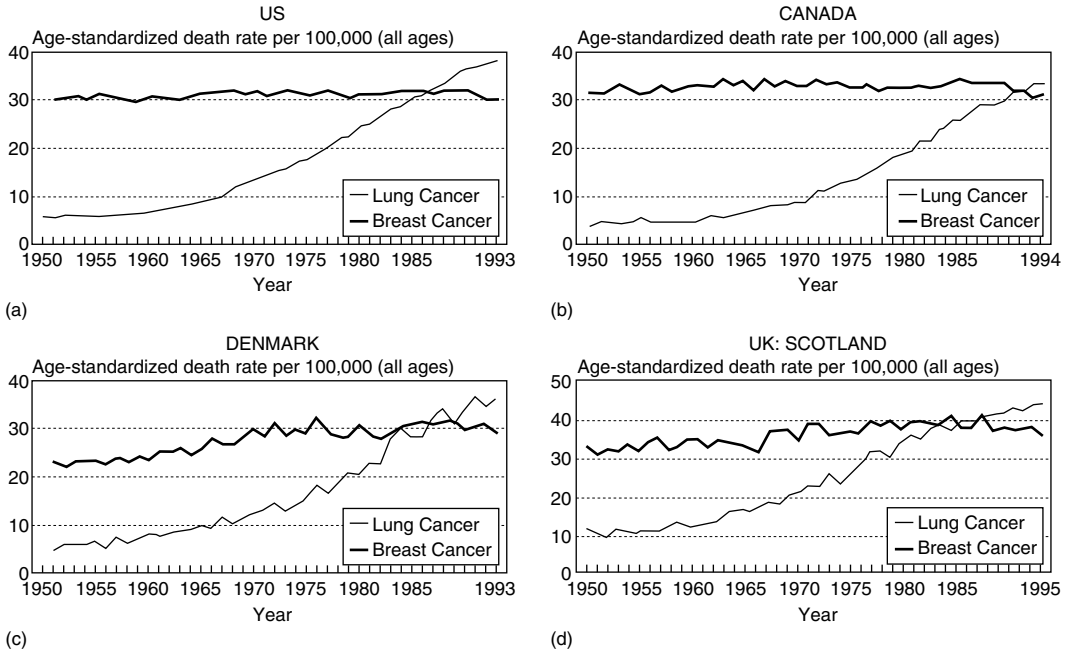

Figure 5 Trends in breast and lung cancer mortality among women, selected countries, 1950–95. (a) US, (b) Canada, (c) Denmark, and (d) UK (Scotland)

then, death rates have halved in France and Portugal, and have declined by 20%–30% in the other countries where death rates from the disease have been comparatively high. On the contrary, there is no evidence that mortality has declined among men in Eastern Europe, and indeed it appears to be rising in several of these countries.

Whither the Future: Mortality and Causes of Death in the Twenty-First Century

As the twentieth century draws to a close, it is perhaps important to reflect briefly on major threats to health in the first decades of the twenty-first century. Unquestionably, the two epidemics of greatest public health concern must be use of *tobacco* and *HIV* infection. Between 1950 and 2000, tobacco will have caused over 60 million deaths in the developed countries of the world, more than 50 million men and about 10 million women [3]. In 1995, tobacco was estimated to have caused about 3 million deaths globally, about 2 million in the developed countries and about a million, but with substantial uncertainty, in less developed countries. In the twentieth century, most of the deaths from tobacco have been in developed populations, but in the twenty-first century the opposite will be true. The annual numbers of deaths are still increasing in developed populations but they are increasing even faster elsewhere. Over the past few decades there has been a massive rise in global cigarette consumption, particularly in developing countries, where 50% of men smoke. On current trends, annual global tobacco deaths are likely to reach 10 million in the 2020s or early 2030s. The chief uncertainty is not whether, but when, annual mortality will reach this level. On present smoking patterns, half a billion of the world's current population will eventually be killed by tobacco. These predictions will be substantially wrong only if there are substantial changes in global smoking patterns.

AIDS caused by the human immunodefeciency virus (HIV), is the only other major cause of death that is rising rapidly. First diagnosed in the early 1980s, the disease is estimated to have caused about 400 000 deaths in 1990, the majority in Sub-Saharan Africa [2]. **Epidemic modeling** of the disease suggests that the peak in global mortality will be attained

sometime between 2005 and 2010, when annual deaths are predicted to reach about 1.7 to 1.8 million a year [2]. Beyond then, the epidemic is expected to decline slowly due to the past (and projected) efforts at prevention. As with tobacco, these projections could be gross underestimates if HIV incidence were to increase rapidly in some large population groups. If this were to happen, it would most probably occur in Asia where seroprevalence has been increasing dramatically in some high-risk populations (*see* **Projections: AIDS, Cancer, Smoking**).

The third area for concern is the emergence, or reemergence, of various infectious diseases which, if uncontrolled, could cause a substantial number of deaths in the future. The reemergence of tuberculosis as a significant health issue in the developed countries is an object lesson for the public health profession not to become complacent about past successes in disease control. Equally, the ebola virus as well as significant cholera outbreaks attest to the need for continual vigilance in **surveillance of diseases**. Finally, the very large unfinished agenda of controlling the leading causes of child mortality in developing countries, particularly diarrheal diseases, acute respiratory infections, the vaccine-preventable diseases, and malaria, which each year collectively kill more than 12 million infants and young children [2], must remain a major global public health priority into the next century.

References

[1] Lopez, A.D. (1983). The sex mortality differential in developed countries, in *Sex Differentials in Mortality: Trends, Determinants and Consequences*, A.D. Lopez & L.T. Ruzicka, eds. Australian National University Press, Canberra, pp. 53–120.

[2] Murray, C.J.L. & Lopez, A.D. (1996). Estimating causes of death: new methods and global and regional applications in 1990, in *The Global Burden of Disease*, C.J.L. Murray & A.D. Lopez, eds. Harvard University Press on behalf of the World Health Organization and the World Bank, Cambridge, Mass, pp. 117–200.

[3] Peto, R., Lopez, A.D., Boreham, J., Thun, M. & Heath, C. (1994). *Mortality from Smoking in Developed Countries, 1950–2000*. Oxford University Press, Oxford.

[4] Preston, S.N. (1976). *Mortality Patterns in National Populations*. Academic Press, New York.

[5] Reid, D.D. & Rose, G.A. (1964). Assessing the comparability of mortality statistics, *British Medical Journal* **2**, 1437–1439.

[6] Ruzicka, L.T. (1995). Suicide mortality in developed countries, in *Adult Mortality in Developed Countries: From Description to Explanation*, A.D. Lopez, G. Caselli & T. Valkonen, eds. Clarendon Press, Oxford, pp. 83–110.

[7] Shkolnikov, V. & Nemstov, A. (1997). *The Anti-Alcohol Campaign and Variations in Russian Mortality*. National Academy of Sciences, Washington.

[8] Shkolnikov, V., Meste, F. & Vallin, J. (1996). Health crisis in Russia II. Changes in causes of death: a comparison with France and England and Wales (1970 to 1993), *Population* **8**, 155–189.

[9] United States Department of Health, Education and Welfare (1975). *Comparability of Mortality Statistics for the Seventh and Eighth Revisions of the International Classification of Diseases, United States*. Public Health Service, Series 2, No. 66.

[10] World Health Organization (1967). The accuracy and comparability of death statistics, *World Health Organization Chronicle* **21**, 11–17.

[11] World Health Organization (1996). *World Health Statistics Annual 1995*. World Health Organization, Geneva.

ALAN D. LOPEZ

Mortality, International Comparisons

The international comparison of mortality or other health-related statistics is probably the most useful, simple, and widely used method to assess the health status of the population of a particular country. In most cases health can be measured only in relative terms, by placing the country on a scale between the best and worst achievements being observed in other countries. Most often, comparisons are made between countries in a specific geographic region or with similar level of socioeconomic development. Such comparisons form an important part of national public health reports or documents on national health policies. Publications and reports of international organizations active in the field of health are usually also largely based on international comparisons of health statistics including mortality data. International mortality comparisons are often the subject of research papers.

To make international comparisons of health data possible, there are several essential and obvious conditions. Data have to be available from a sufficient number of countries and they must be based on the same definitions in order to be comparable. In this respect, mortality statistics are probably the best presently available health data for international comparisons. The **World Health Organization** has been collecting mortality information since the early 1950s, just after the establishment of the Organization. Currently, about 70 countries are regularly reporting detailed data to the WHO on an annual basis. These statistics are based on the concept of the underlying cause of death (*see* **Death Certification**) and are usually coded using the **International Classification of Diseases** (ICD). Generally, these data can be estimated as being of good quality (accuracy) particularly in developed countries with well established and functioning systems of **vital statistics**. However, there are many potential methodologic problems limiting the comparability of mortality data even among developed countries. These problems are mostly related to the coding of the underlying cause of death (*see* **Cause of Death, Underlying and Multiple**). However, the impact of variations in coding procedures on actual cause-specific mortality statistics is very difficult to measure regularly in quantitative terms, as it requires special studies to compare actual methods and practices of coding death certificates between countries. Studies which have been carried out so far have confirmed the perception that in some cases differences in the coding methods and practices may cause significant **bias** in the number of deaths from specific diseases. There are several elements in death registration which may have an influence on the international comparability or may cause an artifact in the trend of particular cause-specific mortality within the country. At least the following could be mentioned:

1. the level of training and corresponding practices in filling in death certificates by health professionals;
2. the form of the certificate itself – for example, the number of lines provided to list underlying and intermediate causes of death;
3. the regulations and administrative structures defining further transfer and processing of death certificates – for example, whether completeness and quality are controlled locally;
4. the coding of the cause of death from the written textual form into the ICD code – for example, whether it is done locally or centrally, manually or automated.

Another factor that is often forgotten, but which may cause significant bias in mortality rates used for international comparisons, is related to the population estimates used as a **denominator** to calculate mortality rates; for example, the number of deaths per 100 000 population. The total resident population in a given country is counted more or less accurately only during population **censuses**, which in most countries are carried out once every 10 years. In between census periods, population estimates are calculated on the basis of births, deaths, immigration, emigration, and aging of the population (*see* **Demography**). In practice, these estimates may not be accurate enough. When such estimates are used to calculate mortality rates, these inaccuracies can cause distortion in mortality trends and, correspondingly, in international comparisons.

Usually it is difficult to detect whether there is any bias in the mortality data of a particular country as compared to other countries. However, one has to keep in mind this possibility while making international comparisons.

There are also several statistical aspects that have to be taken into account in order to avoid the possibility of misleading conclusions based on international comparisons. First of all, the absolute number of deaths, without taking into account the size of the population, should not be used. Mortality rates or other indices should normally be used. In cases in which mortality for all ages or for a wide age band is compared, appropriate mortality rates have to be age-standardized beforehand (*see* **Standardization Methods**). Comparisons of crude death rate (i.e. a simple ratio of the number of deaths to the population size) are often misleading, particularly when one compares countries with different population age structures. For example, the crude death rate is usually higher in developed countries compared to developing ones, although an opposite situation should be expected when considering the health of the population in general. This happens purely because of differences in population structure; that is, developed countries have a much higher proportion of older people with, naturally, high mortality. There are two methods (direct and indirect) to age-standardize mortality rates in order to eliminate the influence of differences in population age structure between countries. If there is a sufficient amount of data for each age group, usually the direct method is used. Mortality rates are calculated for each age group and then are combined into the one index, assuming that the given country has the "standard" population structure. There are two commonly used standards for international comparisons: the world and the European standard populations (see Table 1).

The indirect method of standardization is usually used in cases in which relatively rare causes of deaths are compared, or there are not enough data due to other reasons, to estimate mortality in each age group. This standardization is based on the assumption that the age-specific mortality is the same – that is, "standard" – in each country. These "standard" age-specific mortality rates are usually calculated using combined data from all countries included in the comparisons. The expected number of deaths is calculated on the basis of the above "standard" mortality rates and the actual age distribution of the population in a given country. The ratio of actually observed and calculated expected cases is used as the standardized mortality ratio.

One also has to be careful when comparing countries with small populations. Mortality indices for these countries are less stable, and the position of such countries may change significantly from one year to the next because of random variations.

Table 1 Standard populations (world and European)

Age group (years)	World	European
0	2 400	1 600
1–4	9 600	6 400
5–9	10 000	7 000
10–14	9 000	7 000
15–19	9 000	7 000
20–24	8 000	7 000
25–29	8 000	7 000
30–34	6 000	7 000
35–39	6 000	7 000
40–44	6 000	7 000
45–49	6 000	7 000
50–54	5 000	7 000
55–59	4 000	6 000
60–64	4 000	5 000
65–69	3 000	4 000
70–74	2 000	3 000
75–79	1 000	2 000
80–84	500	1 000
85+	500	1 000
Total	100 000	100 000

Sources: (a) Waterhouse et al. [1]; (b) *World Health Statistics Annual*, Geneva, WHO (any issue). Reproduced by permission of the IARC and the WHO.

For international comparisons, it is preferable to use mortality rates or other mortality based indices which are calculated centrally; for example, by the World Health Organization. Indices calculated individually by each country may have some bias due to the different calculation methods and **software** used in each country. This may happen particularly in the case of **life expectancy** as there are several different mathematical methods and software packages to calculate this index from the raw, age-disaggregated mortality data.

Mortality data are collected by and are available from several international organizations and agencies (e.g. the United Nations, the World Health Organization, the Statistical Office of the European Communities, and the Organization for Economic Cooperation and Development). The database maintained by the WHO is probably the most comprehensive and widely used. Detailed mortality data are published yearly in the *World Health Statistics Annual* [2]. Copies of the database with raw mortality data are available on request in computer-readable form from

the WHO headquarters [3]. For European countries, this information in the form of age-standardized mortality rates is also available as a part of the "Health for All" statistical data base maintained by the WHO Regional Office for Europe in Copenhagen. These data, together with user-friendly data presentation software which facilitates international comparisons, can be downloaded from the **Internet** (`www.who.dk`).

References

[1] Waterhouse, J., Muir, C., Correa, P. & Powell, J., eds (1976). *Cancer Incidence in Five Continents*, Vol. 3. IARC, Lyon, p. 456.

[2] World Health Organization (1995). *World Health Statistics Annual 1994*. WHO, Geneva.

[3] World Health Organization, Division of Health Situation and Trend Assessment, 20 Avenue Appia, CH-1211 Geneva 27, Switzerland.

(*See also* **Data Access, National and International**; **Geographic Patterns of Disease**)

R. PROKHORSKAS

Most Powerful Test

Random phenomena abound in biological and medical studies. When a biological researcher monitors whether an experimental unit in a tumorgenicity study will develop a tumor, the researcher will not be certain whether such an event will occur. Rather, his/her knowledge will be represented by the probability that the event will occur. Let $X = 1(0)$ whenever the event occurs (does not occur). Then the researcher's knowledge concerning the occurrence of a tumor will be represented by the Bernoulli probability mass function (pmf)

$$f_X(x|\theta) = \Pr_{\theta}(X = x) = \theta^x(1-\theta)^{1-x}, x = 0, 1, \tag{1}$$

where $\theta \in \Theta = [0, 1]$ is the probability of a tumor occurring (*see* **Binary Data**). The extreme values $\{0, 1\}$ of θ represent certain knowledge, while $\theta = 1/2$ represents the least amount of knowledge concerning the event. In this example, X is the variable of interest, $\mathcal{X} = \{0, 1\}$ is the range space of X, θ is the relevant parameter, $\Theta = [0, 1]$ is the parameter space, and $f_X(\cdot|\theta)$ is the pmf of X for the parameter value θ.

Another situation is when a medical researcher observes the time of occurrence X of some event such as the onset of AIDS for HIV-infected individuals. In contrast to the first example, the variable X takes values in an uncountable range space $\mathcal{X} = (0, \infty)$, and knowledge of this event is specified by a probability density function (pdf) $f_X(\cdot|\theta)$, where θ is a parameter. As h approaches 0, the pdf has the interpretation

$$f_X(x|\theta)h = \Pr_{\theta}(x < X \leq x + h) + o(h).$$

A specification that arises in many survival time studies is provided by the **exponential** density function

$$f_X(x|\theta) = \theta \exp(-\theta x), x > 0.$$

Thus, generally, a researcher will be interested in some characteristic represented by a variable X. Uncertain knowledge concerning this characteristic is represented by a family of pmfs or pdfs given by $\mathcal{P} = [f_X(\cdot|\theta) : \theta \in \Theta]$, where θ is a parameter taking values in a parameter space Θ. Summary measures about X needed for making important decisions, such as the mean, standard deviation, median, or quartiles, are functions of θ. The true value of θ, denoted by θ_0, is unknown, and it is a goal of the researcher to gain knowledge concerning this value to further his/her knowledge concerning X. To achieve this goal, either through scientific experimentation, clinical trials, etc., he/she observes the **random variables** $X_1, X_2, \ldots, X_n$ which are identically and independently distributed (iid) from $f_X(\cdot|\theta_0)$. The joint pmf or pdf of $(X_1, X_2, \ldots, X_n)$ is therefore

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

Statistical Hypotheses and Tests

The problem of statistical **hypothesis testing** is to decide, on the basis of a realization $(x_1, x_2, \ldots, x_n)$ of $(X_1, X_2, \ldots, X_n)$, whether to reject a **null hypothesis** in favor of an **alternative hypothesis**, or fail to reject it. These hypotheses, which are statements concerning the value of θ, are generally chosen so that the alternative represents change. They are written symbolically as

$$\text{(null)}\mathrm{H}_0 : \theta \in \Theta_0 \text{ and } \text{(alternative)}\ \mathrm{H}_1 : \theta \in \Theta_1,$$

where $\{\Theta_0, \Theta_1\}$ is a partition of Θ. A statistical test of H_0 vs. H_1 requires a statistic $\delta(X_1, \ldots, X_n)$, i.e. a function of $(X_1, \ldots, X_n)$ and possibly other known constants, taking values in [0, 1]. For a realization $(x_1, \ldots, x_n)$, $\delta(x_1, \ldots, x_n)$ is the probability of rejecting H_0 given $(x_1, \ldots, x_n)$. For a test δ, its **power** function $\pi_\delta : \Theta \to [0, 1]$ is

$$\pi_\delta(\theta) = \mathbf{E}_\theta[\delta(X_1, \ldots, X_n)],$$

where $\mathrm{E}_\theta[\cdot]$ represents expectation with respect to the joint pmf or pdf $f_{(X_1,\ldots,X_n)}(\cdot, \ldots, \cdot|\theta)$. The power function is the expected probability of rejecting H_0 when the parameter value is θ. Ideally, we would want $\pi_\delta(\theta) = 0$ whenever $\theta \in \Theta_0$, and $\pi_\delta(\theta) = 1$ whenever $\theta \in \Theta_1$; however, such an ideal situation is seldom achieved except in artificial and/or trivial problems. The size of a test δ is

$$\text{size}(\delta) = \sup_{\theta \in \Theta_0} \pi_\delta(\theta),$$

and δ is of **level** $\alpha (0 \leq \alpha \leq 1)$ if $\text{size}(\delta) \leq \alpha$. $\text{Size}(\delta)$ can be interpreted as the maximum expected probability of committing an *error of type I*, which is committed when the test rejects H_0 when in reality H_0

is true. On the other hand, when $\theta \in \Theta_1$, $1 - \pi_\delta(\theta)$ represents the average probability of committing an *error of type II*, which is committed when the test fails to reject H_0 – a wrong decision, since in such a case H_1 is true.

If $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, we say that Θ_0 and Θ_1 are simple hypotheses. In such a situation, given an $\alpha \in [0, 1]$, a test δ^* is a most powerful α-level (MP-α) if

1. $\text{size}(\delta^*) = \pi_{\delta^*}(\theta_0) \le \alpha$, and
2. for any other test δ with size $(\delta) = \pi_\delta(\theta_0) \le \alpha$, $\pi_{\delta^*}(\theta_1) \ge \pi_\delta(\theta_1)$.

If either Θ_0 or Θ_1 is not simple, the hypothesis is composite. When the null or the alternative hypothesis is composite, a test δ^* is a uniformly most powerful test of level α (UMP-α) if

1. size $(\delta^*) = \sup_{\theta \in \Theta_0} \pi_{\delta^*}(\theta) \le \alpha$, and
2. for any other test δ with size $(\delta) = \sup_{\theta \in \Theta_0} \pi_\delta(\theta) \le \alpha$, $\pi_{\delta^*}(\theta) \ge \pi_\delta(\theta)$ for every $\theta \in \Theta_1$.

Neyman–Pearson Fundamental Lemma

For testing a simple $H_0 : \theta = \theta_0$ vs. a simple $H_1 : \theta = \theta_1$, the **Neyman–Pearson** fundamental lemma (Neyman & Pearson [4]; see also Lehmann [3]) guarantees the existence of an MP-α test which is of the form

$$\delta^*(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } L(\theta_1|x_1, \ldots, x_n) \\ & \quad > cL(\theta_0|x_1, \ldots, x_n), \\ \gamma, & \text{if } L(\theta_1|x_1, \ldots, x_n) \\ & \quad = cL(\theta_0|x_1, \ldots, x_n), \\ 0, & \text{if } L(\theta_1|x_1, \ldots, x_n) \\ & \quad < cL(\theta_0|x_1, \ldots, x_n), \end{cases}$$

where $L(\theta|x_1, \ldots, x_n) = f_{(X_1, \ldots, X_n)}(x_1, \ldots, x_n|\theta)$ is the **likelihood** function, and $c \in [0, \infty)$ and $\gamma \in [0, 1]$ are some constants which are chosen so that $\text{size}(\delta^*) = \alpha$. The fundamental lemma furthermore guarantees that if δ^{**} is an MP-α test for H_0 vs. H_1, then for some $c \in [0, \infty)$, it is of the form

$$\delta^{**}(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } L(\theta_1|x_1, \ldots, x_n) \\ & \quad > cL(\theta_0|x_1, \ldots, x_n), \\ 0, & \text{if } L(\theta_1|x_1, \ldots, x_n) \\ & \quad < cL(\theta_0|x_1, \ldots, x_n). \end{cases}$$

To illustrate, consider the Bernoulli example above, and let $X_1, \ldots, X_n$ be iid from $f_X(x|\theta) = \theta^x(1-\theta)^{1-x}$, $x \in \{0, 1\}$. Suppose interest is in testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where $\theta_0 < \theta_1$. The likelihood function, given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, is $L(\theta|x_1, \ldots, x_n) = \theta^{\Sigma x_i}(1-\theta)^{n - \Sigma x_i}$. Since

$$\log\left[\frac{L(\theta_1|x_1, \ldots, x_n)}{L(\theta_0|x_1, \ldots, x_n)}\right] = \left(\sum_{i=1}^n x_i\right) \log\left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right] + n \log\left(\frac{1-\theta_1}{1-\theta_0}\right),$$

and noting that $\theta_1(1-\theta_0)/\theta_0(1-\theta_1) > 1$, then the MP-$\alpha$ test is of the form

$$\delta^*(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } \sum x_i > k, \\ \gamma, & \text{if } \sum x_i = k, \\ 0, & \text{if } \sum x_i < k. \end{cases}$$

If the true value of the parameter is θ, the statistic $T = \sum X_i$ has a **binomial distribution** with parameters n and θ, so

$$\Pr_\theta(T = j) = b(j; n, \theta) \equiv \binom{n}{j} \theta^j (1-\theta)^{n-j}, \quad j = 0, 1, \ldots, n.$$

To satisfy the requirement that the size of the test is α, one could take

$$k = \min\left[j \in (0, 1, \ldots, n) : \sum_{i=j+1}^n b(i; n, \theta_0) \le \alpha\right]$$

and

$$\gamma = \frac{\alpha - \sum_{i=k+1}^n b(i; n, \theta_0)}{b(k; n, \theta_0)},$$

so the MP-α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ becomes

$$\delta^*(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } \sum x_i > k, \\ \gamma, & \text{if } \sum x_i = k, \\ 0, & \text{if } \sum x_i < k. \end{cases}$$

Since the test δ^* does not depend on θ_1, it is also an MP-α test for $H_0 : \theta = \theta_0$ vs. $H_1' : \theta = \theta_1'$ provided that $\theta_1' > \theta_0$. Consequently, it is a UMP-α test for testing H_0 vs. $H_1'' : \theta > \theta_0$. Furthermore, since $\pi_{\delta^*}(\theta_0') \le \pi_{\delta^*}(\theta_0)$ for every $\theta_0' \le \theta_0$, then, as a test

for $H_0' : \theta \leq \theta_0$ vs. H_1'', δ^* is of level α and hence is UMP-α for testing H_0' vs. H_1''.

For the exponentially distributed time-to-event example at the beginning of this article suppose one wants to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ based on the values of $X_1, \ldots, X_n$ which are iid from $f_X(x|\theta) = \theta \exp(-\theta x)$, $x > 0$. The likelihood function is $L(\theta|x_1, \ldots, x_n) = \theta^n \exp(-\theta \sum x_i)$. By the fundamental lemma the MP-α test is

$$\delta^*(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } \sum x_i < c', \\ 0, & \text{if } \sum x_i > c', \end{cases}$$

since

$$[(x_1, \ldots, x_n) : L(\theta_1|x_1, \ldots, x_n) > cL(\theta_0|x_1, \ldots, x_n)]$$

is equivalent to

$$\left[(x_1, \ldots, x_n) : \sum_{i=1}^{n} x_i < c'\right]$$

for some c', and with the change in direction of the inequality due to the inequality $\theta_0 - \theta_1 < 0$. Under $H_0 : \theta = \theta_0$, the statistic $2\theta_0 \sum_{i=1}^{n} X_i$ has a central **chi-square distribution** with $2n$ **degrees of freedom**, whose quantiles are well-tabulated for small to moderate values of $2n$ (cf. [1] and [5]) or are obtained easily using a computer. Consequently, if one sets $c' = \chi^2_{2n;1-\alpha}$, where $\Pr(\chi^2_{2n} \geq \chi^2_{2n;1-\alpha}) = 1 - \alpha$, where χ^2_k is a central chi-square distributed variable with k degrees of freedom, then the test

$$\delta^*(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } 2\theta_0 \sum x_i \leq \chi^2_{2n;1-\alpha}, \\ 0, & \text{if } 2\theta_0 \sum x_i > \chi^2_{2n;1-\alpha}, \end{cases} \tag{2}$$

is an MP-α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ with $\theta_1 > \theta_0$. Again, since δ^* does not depend on θ_1, it is an MP-α test for H_0 vs. $H_1' : \theta = \theta'$ with $\theta' > \theta_0$. Consequently, it is a UMP-α test for H_0 vs. $H_1'' : \theta > \theta_0$. Furthermore, since for any other $\theta_0' \leq \theta_0$

$$\begin{aligned} \pi_{\delta^*}(\theta_0') &= \mathrm{E}_{\theta_0'}[\delta^*(X_1, \ldots, X_n)] \\ &= \Pr_{\theta_0'}\left(2\theta_0 \sum X_i \leq \chi^2_{2n;1-\alpha}\right) \\ &= \Pr\left\{\chi^2_{2n} \leq \frac{\theta_0'}{\theta_0}\chi^2_{2n;1-\alpha}\right\} \\ &\leq \Pr\left\{\chi^2_{2n} \leq \chi^2_{2n;1-\alpha}\right\} \text{ since } \theta_0' \leq \theta_0 \\ &= \alpha, \end{aligned}$$

so, as a test for $H_0' : \theta \leq \theta_0$ vs. H_1'', it is therefore of level α. Consequently, δ^* in (2) is a UMP-α test for H_0' vs. H_1''.

In the above examples, the MP-α Neyman–Pearson tests turn out to be UMP-α tests for testing one-sided alternatives. These are particular cases of testing problems where the monotone likelihood ratio (MLR) property holds. When the parameter $\theta \in \Theta \subseteq \Re = (-\infty, \infty)$ is one-dimensional, the family $\mathcal{P}$ possesses the MLR property in a (one-dimensional) statistic $S(X_1, \ldots, X_n)$ if, for every $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$:

.

1. $L(\theta_2|x_1, \ldots, x_n)/L(\theta_1|x_1, \ldots, x_n) = h[S(x_1, \ldots, x_n); \theta_1, \theta_2]$ for some function $h(\cdot; \cdot, \cdot)$; and
2. $h(s; \theta_1, \theta_2)$ is a monotone nondecreasing function in s.

Under this situation, the MP-α test for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where $\theta_0(<, >)\theta_1$, is also the UMP-α test for $H_0' : \theta(\leq, \geq)\theta_0$ vs. $H_1' : \theta(>, <)\theta_0$, and it is of the form

$$\delta^*(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } S(x_1, \ldots, x_n)(>, <)c, \\ \gamma, & \text{if } S(x_1, \ldots, x_n) = c, \\ 0, & \text{if } S(x_1, \ldots, x_n)(<, >)c, \end{cases}$$

where constants c and γ are chosen in order for $\pi_{\delta^*}(\theta_0) = \alpha$. For more details concerning this MLR property and its consequences, we refer the reader to Lehmann [3], pp. 78–86. Finally, we remark that, for two-sided types of alternatives, e.g. $H_1 : \theta \neq \theta_0$, UMP-$\alpha$ tests generally do not exist, and hence there is usually a need to restrict the search for a "best" test to a smaller class, such as the class of **unbiased** tests and/or invariant tests. An in-depth treatment of such tests can be found in Lehmann [3], Chapters 4–6. Some other references which discuss optimal hypothesis tests and provide numerous examples are the books by Bickel & Doksum [1], Casella & Berger [2], and Rohatgi [5].

References

[1] Bickel, P. & Doksum, K. (2000). *Mathematical Statistics: Basic Ideas and Selected Topics*, 2nd Ed. Prentice-Hall PTR.

[2] Casella, G. & Berger, R. (2001). *Statistical Inference*. 2nd Ed. Wadsworth.

[3] Lehmann, E. (1997). *Testing Statistical Hypotheses*, 2nd Ed. Springer-Verlag, New York.

[4] Neyman, J. & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society, Series A* **231**, 289–337.

[5] Rohatgi, V. & Saleh, A.K. (2001). *An Introduction to Probability Theory and Mathematical Statistics*. 2nd Ed. Wiley, New York.

EDSEL A. PEÑA & VIJAY K. ROHATGI

Moving Average

Moving averages are operations applied to **time series** to achieve smoothing. The aim is usually to smooth the series enough to distinguish particular features of interest. A simple but effective model of a time series is to regard it as being made up of a long-term trend, a cycle and random effects. Symbolically,

$$X_t = m(t) + g(t) + e_t,$$

that is,

$$\text{observation} = \text{trend} + \text{cycle} + \text{random variation}.$$

Moving averages are often used to eliminate the seasonal or cyclic effects (*see* **Circadian Variation**) and hence to emphasize the trend terms, but this is not their only use.

Suppose we have a time series, i.e. a series of measurements taken over time, say

$$X_1, X_2, X_3, \ldots, X_N.$$

Define a new series by averaging the first s values, then deleting the first value from this group, adding the $(s+1)$th observation to the group and averaging, then deleting the first value from this group, adding the $(s+2)$th and averaging, then ... and so on. Symbolically,

$$\begin{aligned}
X^*_{(s+1)/2} &= \frac{(X_1 + X_2 + \cdots + X_s)}{s},\\
X^*_{(s+3)/2} &= \frac{(X_2 + X_3 + \cdots + X_{(s+1)})}{s},\\
X^*_{(s+5)/2} &= \frac{(X_3 + X_4 + \cdots + X_{(s+2)})}{s},\\
&\cdots\\
X^*_{N-(s-1)/2} &= \frac{(X_{N-s+1} + \cdots X_{N-1} + X_N)}{s}.
\end{aligned}$$

The new series, X^*_t, has been obtained from the original by applying a *moving average of length s* to the original series. All that we have done is to average successive blocks of s terms. This is easier to see for specific values of s; for $s = 3$ and 4 we have

$$\begin{aligned}
&s = 3\\
&X^*_2 = \frac{(X_1 + X_2 + X_3)}{3},\\
&X^*_3 = \frac{(X_2 + X_3 + X_4)}{3},\\
&X^*_4 = \frac{(X_3 + X_4 + X_5)}{3},\\
&\cdots.
\end{aligned}$$

$$\begin{aligned}
&s = 4\\
&X^*_{5/2} = \frac{(X_1 + X_2 + X_3 + X_4)}{4},\\
&X^*_{7/2} = \frac{(X_2 + X_3 + X_4 + X_5)}{4},\\
&X^*_{9/2} = \frac{(X_3 + X_4 + X_5 + X_6)}{4},\\
&\cdots.
\end{aligned}$$

The smoothed values, X^*_t, are assumed to be at the mean of the time values in each block. When s is even we see that the smoothed series lies between the original time points, which may be inconvenient. We can overcome this by applying a further two-point average to the "badly sited" series to give a new smoothed version called the "centered" moving average. The numerical example below uses this method.

It is straightforward to show that if a series has a cyclic effect of period s, then the application of an s-term moving average will remove it and will leave locally linear effects unchanged. Suppose in our model above that $X_t = g(t) + m(t) + e_t$, and $g(t)$ is periodic with period s. We assume that $\sum_{k=1}^{s} g(k) = 0$, since we can adjust any constant terms in the trend function. Then it follows that $\sum_{k=1}^{s} g(t+k) = 0$, so the effect of averaging s successive terms is to remove the $g(t)$ terms. For details, see [3] or [1].

As an example of the kind of calculation in Table 1, we demonstrate the effect of taking a four and then a two-point moving average of some the mortality rates in Baltimore, USA. The data are from Bliss [2]. Only part of the calculation is displayed.

The whole series is given in Figure 1. As we can see from the figure, the quarterly oscillations in the original data are suppressed by the moving average to give the smoothed series. The application of the quarterly average of length four, coefficients [1/4, 1/4, 1/4, 1/4] followed by the application of a moving average [1/2, 1/2] is equivalent to a five-point moving average with coefficients [1/8, 1/4, 1/4, 1/4, 1/8].

Table 1 Quarterly log (death rates per 100 000) in Baltimore

X_t	Four-point MA	Two-point MA
0.8597		
0.8419		
	0.84261	
0.8300		0.8412
	0.8398	
0.8389		0.8398
	0.8399	
0.8484		0.8392
	0.8385	
0.8424		0.8379
	0.8372	
0.82443		0.8374
	0.8376	
0.8336		0.8372
	0.8368	
0.8501		0.8368
	0.8368	
0.8392		0.8360
	0.8352	
0.8245		0.8342
	0.8332	
0.8270		0.8327

The bracket notation here is a useful one in that it gives the number of terms in the average and the individual coefficients. The terms in the bracket are used to give a weighted average.

Once the cyclic effects are removed, the trend is easily seen. The cyclic component is evaluated by taking the cycle-free smoothed series from the original X_t. Of course, we are assuming that the effects are additive, that is,

$$X_t = \text{cyclic term} + \text{trend} + \text{other terms}.$$

However, if we have multiplicative effects, say

$$X_t = \text{cyclic term} \times \text{trend} \times \text{others},$$

then it is clearly necessary to take logs. We can then apply the moving average to the log series.

In addition to the removal of cyclic effects, moving averages arise in a natural way in the piecewise fitting of polynomials to time series. Suppose we decide to fit a polynomial of degree 3, say, $a_0 + a_1 t + a_2 t^2 + a_3 t^3$, to five points of our series. We know it is very much simpler to transform our time scale so that our points are X_{-2}, X_{-1}, X_0, X_1, X_2 at times -2, -1, 0, 1, 2.

The usual regression normal equations are

$$\sum_{t=-2}^{2} X_t = 5a_0 + a_1 \sum_{t=-2}^{2} t + a_2 \sum_{t=-2}^{2} t^2 + a_3 \sum_{t=-2}^{2} t^3,$$

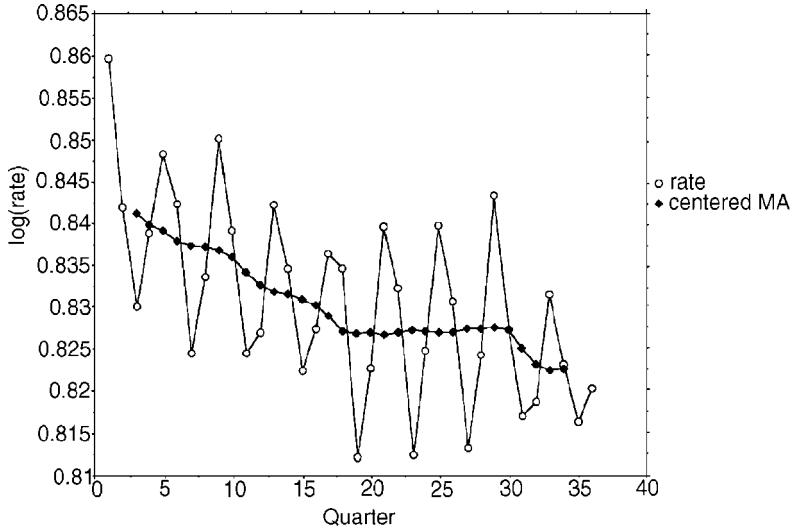

Figure 1 Log death rates and a moving average

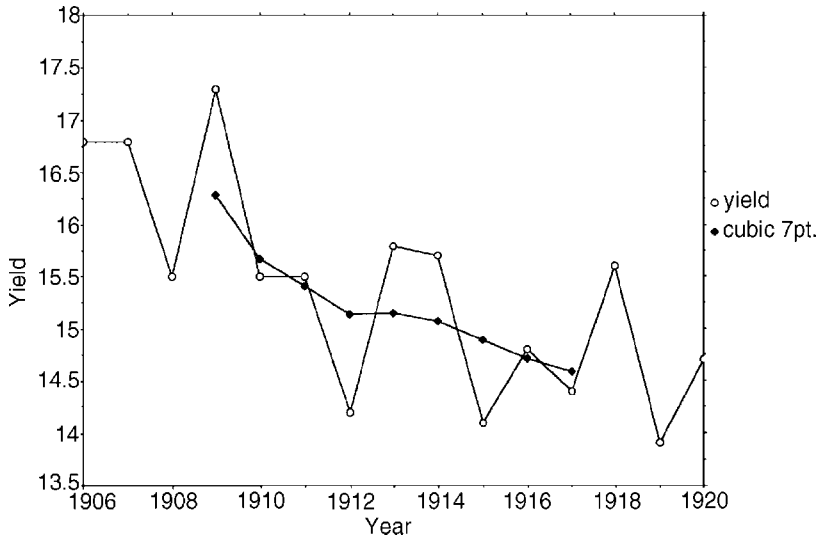

Figure 2 Barley yield in the UK, 1906–1920 and seven-point cubic

$$\sum_{t=-2}^{2} tX_t = a_0 \sum_{t=-2}^{2} t + a_1 \sum_{t=-2}^{2} t^2 + a_2 \sum_{t=-2}^{2} t^3 + a_3 \sum_{t=-2}^{2} t^4,$$

$$\sum_{t=-2}^{2} t^2X_t = a_0 \sum_{t=-2}^{2} t^2 + a_1 \sum_{t=-2}^{2} t^3 + a_2 \sum_{t=-2}^{2} t^4 + a_3 \sum_{t=-2}^{2} t^5,$$

$$\sum_{t=-2}^{2} t^3X_t = a_0 \sum_{t=-2}^{2} t^3 + a_1 \sum_{t=-2}^{2} t^4 + a_2 \sum_{t=-2}^{2} t^5 + a_3 \sum_{t=-2}^{2} t^6,$$

but

$$\sum_{t=-2}^{2} t = \sum_{t=-2}^{2} t^3 = \sum_{t=-2}^{2} t^5 = 0 \text{ and}$$

$$\sum_{t=-2}^{2} t^2 = 10, \qquad \sum_{t=-2}^{2} t^4 = 34, \qquad \sum_{t=-2}^{2} t^6 = 130,$$

so we have

$$\begin{aligned}
\{X_{-2} + X_{-1} + X_0 + X_1 - X_2\} &= 5a_0 \qquad 10a_2\\
\{-2X_{-2} - X_{-1} \quad + X_1 - 2X_2\} &= \qquad 10a_1 \qquad 34a_3\\
\{4X_{-2} + X_{-1} \quad + X_1 + 4X_2\} &= 10a_0 \qquad 34a_2\\
\{-8X_{-2} - X_{-1} \quad + X_1 - 8X_2\} &= \qquad 34a_1 \qquad 130a_3,
\end{aligned}$$

giving $a_0 = (1/35)[-3X_{-2} + 12X_{-1} + 17X_0 + 12X_1 - 3X_2]$.

If we fit this polynomial in piecewise segments and use the constant term as the smoothed values, this is equivalent to a moving average of length 5 with unequal weights that we write as (1/35) [−3, 12, 17, 12, −3].

This is sometimes abbreviated to (1/35) [−3, 12, **17**] where only half the bracket is given. When in doubt, note that the terms in the bracket must sum to the denominator of the divisor. So, in (1/35) [−3, 12, 17, 12, −3], $-3 + 12 + 17 + 12 - 3 = 35$.

We can find moving averages for all such polynomials; thus, for cubics:

5 points	(1/35) [−3, 12, 17, 12, −3],
7 points	(1/21) [−2, 3, 6, 7, 6, 3, −2],
9 points	(1/231) [−21, 14, 39, 54, 59, 54, 39, 14, −21],
11 points	(1/429) [−36, 9, 44, 69, 84, 89, 84, 69, 44, 9, −36],
13 points	(1/143) [−11, 0, 9, 16, 21, 24, 25, 24, 21, 16, 9, 0, −11],

15 points (1/1105) [−78, −13, 42, 87, 122, 147, 162, 167, 162, 147, 122, 87, 42, −13, −78],

while similar expressions are available for quintic curves. The effect of the seven-point cubic is shown in Figure 2.

Such averages are useful in giving a way of smoothing series that requires very few assumptions and that is reasonably simple to use, especially with a spreadsheet. The drawbacks are precisely the same; the informal procedure does not lend itself to testing or model building. It is difficult to make a rational choice of polynomial and moving-average length. Some theory is possible, see [1]. Thus, for a fixed-length moving average the variance increases with polynomial order, while for fixed order the bias decreases with length, but this result is of limited use. Also, one needs to take care, since cavalier use of successive moving averages may well lead to induced cyclic effects in the smoothed series (*see* **Slutzky–Yule Effect**).

Moving averages have a long history linked to interpolation from tables, especially actuarial tables. Many interpolation formulas involve differences, and smoothing formulas are used to give a graduation with smooth successive differences. A moving average is said to be correct to order q if differences of polynomials of this order remain unaffected. One popular example is Spencer's 15-point formula, which consists of applying (1/4) [−3, 3, 4, 3, −3], then (1/5) [1, 1, 1, 1, 1], then (1/4) [1, 1, 1, 1], followed by (1/4) [1, 1, 1, 1].

References

[1] Anderson T.W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.

[2] Bliss. C.I. (1970). *Statistics in Biology*, Vol. 2. McGraw-Hill, New York.

[3] Janacek, G. & Swift, A. (1993). *Time Series*. Ellis Horwood, Chichester.

G.J. JANACEK

Multicenter Trials

A multicenter **clinical trial** is defined in this article to be a study involving two or more field sites that is conducted according to a common protocol (*see* **Clinical Trials Protocols**) and that uses a single data coordinating center to receive, process and analyze study data. In conformity with terminology adopted in [1], field sites are the primary sites of participant accrual, intervention, data collection and follow-up. For example, in trials comparing alternative treatments for a specific disease, the field sites may consist of hospitals, clinics or other locations that provide patient care. The number of field sites varies greatly across studies according to their size, complexity and purpose. For example, the Beta-Carotene and Retinol Efficacy Trial (CARET), a **prevention trial** in men and women at high risk for lung cancer, involves six field sites, whereas, in the cooperative oncology group setting, it is common for trials to encompass hundreds of sites.

In this article we provide several examples of multicenter trials, and discuss the rationale for conducting them. We also describe organizational, structural, and analytic aspects of multicenter trials. Finally, we suggest guidelines for the conduct of such studies.

More comprehensive discussions can be found in the books by Meinert [7] and Pocock [11], and in a special edition of *Controlled Clinical Trials* edited by Wittes [15].

Examples of Multicenter Trials

There are currently thousands of active multicenter trials designed to evaluate treatment or prevention strategies for every major disease. Examples from three different areas are discussed below to indicate the diverse nature of multicenter trials, and their impact on the practice of medicine.

Breast Cancer

In the past quarter-century there has been a profound change in the surgical treatment of breast cancer. Radical (Halstedian) mastectomy has been replaced by less extensive surgery, and, today, lumpectomy (local excision), together with radiation, may be recommended for the majority of women with early-stage breast cancer [10]. This shift in treatment strategy was supported by a series of multicenter randomized (*see* **Randomization**) trials comparing less-extensive to more-extensive surgery, which found no survival benefit to the more extensive forms of surgery [2]. One of the early studies in this series was conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP) and compared radical mastectomy to total mastectomy with and without post-operative breast irradiation [3]. Between July 1971 and September 1974, 1765 women with operable breast cancer entered this study at 34 institutions throughout the US and Canada. After publication of the five-year results, which demonstrated no difference in survival between treatments, the percentage of radical mastectomies done as the surgical procedure for primary operable breast cancer in the US dropped from 75% to less than 5%.

Coronary Disease

Not all multicenter trials have survival or disease-free survival (*see* **Survival Analysis, Overview**) as their primary endpoint (*see* **Outcome Measures in Clinical Trials**). An example is given by a multicenter, randomized placebo-controlled trial of the angiotensin-converting enzyme inhibitor cilazapril, designed to assess its efficacy in reducing the rate of restenosis following percutaneous transluminal coronary angioplasty [9]. Seven hundred and thirty-five patients were randomized to receive either 2.5 mg cilazapril in the evening following angioplasty and 5 mg b.i.d. for six months, or placebo. Patients also received aspirin for six months. The primary endpoint was defined to be the difference in minimal coronary lumen diameter, post-angioplasty to six months follow-up. The average change in minimal coronary lumen diameter did not differ between controls and patients treated with cilazapril, nor did the frequency of serious clinical events (death, myocardial infarction, coronary revascularization, recurrent angina). These findings were confirmed in a subsequent, larger study [8].

Prevention

Two trials that were of importance in assessing the efficacy of beta-carotene as a preventive agent against lung cancer in high-risk participants were the Alpha-Tocopherol, Beta-Carotene Cancer Prevention

Trial (ATBC) [14] and the Beta-Carotene and Retinol Efficacy Trial (CARET) [13]. In the ATBC trial, 29 133 male smokers in Finland were randomized to receive daily 50 mg alpha tocopherol (vitamin E), 20 mg beta-carotene, both drugs or placebo. Participants remained on treatment for five to eight years. In the CARET trial, 18 314 high-risk participants were randomized to receive either daily beta-carotene (30 mg) and vitamin A (retinyl palmitate 25 000 IU) or placebo. This study was terminated ahead of schedule in January 1996 after about four years of treatment. In both studies, more lung cancers were diagnosed and there were more deaths among participants receiving beta-carotene than among those participants not receiving the drug. The two multicenter trials illustrate the not uncommon occurrence of large randomized trials refuting an effect suggested by **nonrandomized** studies.

Rationale for Conducting Multicenter Trials

Adequacy of Accrual

The decision to mount a large-scale multicenter study is driven by the need to recruit subjects at a faster rate than can be accomplished at a single center. This will be the case when anticipated treatment differences are important but relatively small, or only a minority of participants are likely to experience the outcome(s) in question. Thus, prevention studies are almost always organized as multicenter trials (*see* **Prevention Trials**). In trials investigating the efficacy of treatments against established disease, the multicenter approach is common in the design of Phase III trials, which involve the randomized comparison of new treatments against a currently accepted standard. **Phase I trials** (assessment of toxicity) and **Phase II trials** (preliminary establishment of clinical activity) require many fewer patients, and are often carried out within a single institution. However, in rare diseases the multicenter approach may be necessary, even for Phase I and II trials.

Increased Generalizability of Study Conclusions

If the protocol is crafted to define clearly the patient population of interest without arbitrary exclusion of potential participants (*see* **Eligibility and Exclusion Criteria**), then the use of multiple field sites will normally result in a more heterogeneous population of participants (both patients and physicians) and will enhance the generalizability of the results. Of equal importance is the fact that widespread participation in the trial will ensure greater acceptance of the study results in the community, without which the conclusions of the study cannot be translated into standard medical practice.

Broad-Based Clinical Trials Bring State-of-the-Art Treatment to the Community

Most current therapies, at least in cancer treatment, are first made available to patients through the clinical trials process. In the Phase III setting, it is an ethical requirement that unproven treatments have a reasonable probability of improving on the current standard of care, and a low probability of being materially inferior (*see* **Ethics of Randomized Trials**). Furthermore, the care of patients enrolled on multicenter trials is governed by a carefully designed protocol, and standards of patient management, testing, and follow-up are arguably better than might be expected outside the clinical trials process. Additionally, participation in clinical trials gives community-based physicians and other care-givers the opportunity to learn about newly evolving strategies of treatment and patient management. This should have a positive effect on the care received by all patients, including those who choose not to participate in clinical trials.

A Caveat

A community-based multicenter trial is not the optimal setting to compare complicated regimens that require intensive training on the part of the treating physicians. Accrual to such studies may be difficult, and it is possible that patient care could be compromised. Furthermore, if a complicated treatment appears to provide no benefit on the basis of the trial data, the interpretation of these results will be ambiguous if there is any question regarding protocol adherence (*see* **Compliance Assessment in Clinical Trials**).

Organizational Structure and Personnel

All multicenter trials share a common structure involving distributed field sites responsible for participant accrual, intervention, primary data collection, and follow-up. These activities are directed through a coordinating center, and overall

responsibility for the conduct of the trial is generally assumed by a study chairman. The coordinating center serves numerous critical functions, as detailed by Meinert [7]; these include study design; the development of a protocol document; recruitment, training, and coordination of accrual sites; patient randomization; data entry and processing (*see* **Data Management and Coordination**); ongoing monitoring of toxicity data; periodic interim analyses of study endpoints (*see* **Data and Safety Monitoring**); auditing of field sites (*see* **Clinical Trials Audit and Quality Control**); regulatory reporting; final data analyses; and preparation of abstracts and manuscripts. While there is considerable efficiency to be gained by having all components of the coordinating center in one place, it may be the case that an institution that has expertise in one function fails to have expertise in another. Thus, coordinating center staff will not always be located at the same institution. In fact, a common model is to divide the coordinating center into an operations office having responsibility for logistical aspects of the trial and a data center responsible for data management and statistical reporting. Often, these two functions are physically separated.

The major personnel involved in conducting a multicenter trial are listed below, together with brief summaries of their functions:

Study Chairman

The study chairman is responsible for the overall project. Ideally, the individual will be involved from the time the study is first conceptualized until the final analyses are performed and results reported. This individual should be an expert in the disease being studied, should have previous clinical trial experience, and must be strongly committed to the success of the study. He or she must have the time and commitment to address key issues, particularly regarding recruitment and compliance. For large studies, a steering committee may assist the study chairman in overseeing the design and conduct of the trial.

Trial Statistician

A statistician should be identified to participate in and be responsible for the statistical design, monitoring, and analysis of the study. Responsibilities include preparation and presentation of interim endpoint analyses, preparation of toxicity tables for ongoing review, tracking of accrual, monitoring of follow-up to assure no **biases** are occurring and, ultimately, collaboration in the preparation of manuscripts summarizing trial results. The trial statistician also has general responsibility for the statistical design of ancillary studies and any related data transfers.

Operations Officer

For large trials, an operations officer should be appointed to oversee the study logistics, often with the assistance of the study chair and coordinating center support staff. The logistical considerations to be addressed include protocol development, regulatory compliance, communication with field sites, drug distribution, trial participant meetings, medical review, and interactions with the steering committee, trial sponsors, and vendors. For small studies, many of these duties may be assumed by the study chairman or trial statistician.

Data Manager

The data manager assures that quality data are entered into the research database in a timely manner. He or she participates in the design of forms (*see* **Questionnaire Design**), assists in determining which data are feasible to collect and the best procedures for doing so, sets and adheres to timeliness goals for data submission and entry, enters and/or checks data, queries field sites about missing or unclear data, abstracts data from medical reports, identifies items required for further medical review, and assists in the training of field site personnel.

Randomization Specialist

The randomization specialist reviews patient demographic and clinical data, verifies patient eligibility (*see* **Eligibility and Exclusion Criteria**), checks that informed consent has been properly documented, and finalizes treatment assignments (*see* **Randomized Treatment Assignment**).

Quality Assurance Officer

The quality assurance officer implements procedures to verify that field sites are in compliance with the protocol and that data processing and monitoring at the coordinating center is accurate and

timely. These procedures include centralized medical review, on-site audits, data entry verification, data flow monitoring, computerized edit checks for data completeness and consistency, and procedures for assuring data **confidentiality** and security.

Computer Support Personnel

This function may be divided into three areas. The first is **database** management, including maintenance of the research database and support of computerized aspects of the data quality assurance program. A second area is applications programming in support of statistical reports, and operational functions such as drug distribution, Institutional Review Board (IRB) approvals, and field site performance monitoring. A third area is systems management, which includes the development and maintenance of internal hardware and **software** systems, intra-office communications and support systems, and the development of capabilities for communicating with field sites, such as e-mail and **Internet**.

Resource Center Directors

Many trials have centers that perform special functions, such as reading pathology slides, serum banking, or reviewing unusual toxicities. A director is required to assure quality control and to facilitate communication and data transfer with headquarters.

Training Director

An individual should be charged with the responsibility for training investigators and support personnel at the participating field sites. This function includes on-site training, workshops at one or more central locations, and preparing data management and/or treatment handbooks, videos, and centralized resources accessible by phone, e-mail or Internet. This individual may also assume responsibility for responding to questions from the field sites and soliciting input from investigators and support personnel to help identify problems with the protocol conduct and methods for reducing an administrative burden.

Field Site Personnel

Contact people should be identified at each field site. These may include both the treating physician and an individual who has primary responsibility for data submission (a clinical research associate, nurse coordinator or data manager).

Data Monitoring Board

Most trials have **data and safety monitoring boards** that review interim study data. Some are independent of the study investigators, others include both study investigators and independent members, and some boards comprise study investigators and coordinating center staff. The board reviews the progress of the trial to assure that the statistical monitoring plan is followed, that there are no dangerous trends in the toxicity or outcome data, and generally to assure that the trial is conducted according to protocol (*see* **Data and Safety Monitoring**).

Guidelines for the Conduct of Multicenter Trials

It is imperative in conducting any trial that all personnel share a common understanding of the aims, operational details, and reporting requirements. Six specific areas are summarized below.

Study Protocol

Because of the dispersed nature of a multicenter trial, it is important that a written protocol be developed that addresses all aspects of the study (*see* **Clinical Trials Protocols**). The protocol is an instrument for communicating the study requirements to the investigators; hence, must be a clear, concise document that minimizes individual interpretations. Topics to be addressed include:

1. Rationale for and specific aims of the study.
2. Patient eligibility requirements and entry procedures.
3. Study endpoint definitions.
4. Required procedures for treatment administration, including precise rules for dose determinations.
5. Patient management guidelines, including specifications for dose reductions, treatment delays and treatment terminations.
6. Schedules of required clinical tests and assessments.
7. Schedule for submission of required materials and data, including long-term follow-up.
8. Data and materials submission procedures.

9. Regulatory obligations, including informed consent and reporting of adverse events.
10. Statistical considerations, to include: method of treatment assignment; anticipated accrual pattern; **power** analysis justifying sample size requirements (*see* **Sample Size Determination for Clinical Trials**); interim monitoring and analysis plans; and planned time and methodology of final analyses. The section should also describe methods to be used to address secondary aims of the study, compare toxicities, and analyze data from any ancillary laboratory studies.

Data Requirements

Data submission requirements should be examined to eliminate unnecessary data items. Generally, data should not be required for submission unless they are required to (i) address the specific aims of the study, (ii) assure that the study is carried out in compliance with its protocol, (iii) monitor the progress of the study both with respect to toxicity and clinical outcome, or (iv) fulfill regulatory reporting requirements (*see* **Drug Approval and Regulation**). Likewise, data collection instruments should be made as simple as possible (*see* **Questionnaire Design**). Their design should consider the perspective of the clinical research associates, nurses and data managers who will be responsible for their completion (*see* **Data Management and Coordination**).

Submission of Materials

Requirements for the submission of materials such as fresh tissues or serum samples must be considered carefully. Such requirements may place undue burdens on community-based hospitals and may seriously impact patient recruitment. When designing ancillary studies requiring such materials, consideration should be given to restricting sample collection to a subset of sites experienced in processing the required material.

Communication

It is essential that channels of communication be kept open between the field sites and the coordinating center. Provision should be made for regular meetings at which investigators are apprised of the current status of the trial, and are encouraged to provide feedback to the scientific leadership regarding problems encountered at the local level. Between meetings there should be frequent communication via mail, electronic mail, fax, newsletters or other media. Field investigators should be encouraged to contribute to the scientific program of the group. Some clinical trials have benefited from encouraging the involvement of trial participants on advisory committees, in discussion groups, and in the preparation of newsletters.

Site Performance Monitoring

It is important that field site investigators have an a priori understanding of the performance standards expected of them. Acceptable standards for performance with regard to on-site audits, accrual, data submission delinquency, protocol compliance, and patient eligibility should be explicit. Standards should be realistic, so that there is an appropriate balance between quality assurance and feasibility of participation.

Special Statistical Requirements

In multicenter trials it is common practice to stratify treatment assignment by field site in addition to other relevant stratification variables. If there are only a few field sites, full stratification may be used to achieve a balanced treatment assignment within each combination of **stratification** variables. In the cooperative group setting where the number of field sites may be very large, a dynamic allocation method may be required, since stratum sizes would typically be too small to achieve treatment balance within each cell. This ensures that treatment assignments will be balanced with respect to each individual stratification variable (*see* **Adaptive and Dynamic Methods of Treatment Assignment**).

If there are only a few field sites, one should include the sites in analyses, either as stratification factors or as independent variables in a model. One might also wish to look for important **interactions** of field site with treatment (*see* **Treatment-covariate Interaction**). If there are many field sites, it may be impractical to use field sites as stratification variables in analysis. However, when there are many field sites, no one field site will be likely to influence greatly the results of the analyses. Application of **Bayesian methods** to the analysis of institutional effects in multicenter trials may provide new insight. Three fairly recent publications by Skene &

Wakefield [12], Gray [4], and Gustafson [5] describe Bayesian approaches.

Interim monitoring of toxicities and events, often done by independent data and safety monitoring boards, requires careful planning in the statistics section of the protocol. Board meetings must be planned in advance, so it is not practical to schedule interim analyses at a fixed number of events. Flexible monitoring rules like those introduced by Lan & DeMets [6] are required to adjust for the fact that analyses are based on fixed calendar times rather than a fixed number of events.

References

[1] Blumenstein, B.A., James, K.E., Lind, B.K. & Mitchell, H.E. (1995). Functions and organization of coordinating centers for multicenter studies, *Controlled Clinical Trials* **16**, 4S–29S.

[2] Early Breast Cancer Trialists' Collaborative Group (1993). Effects of radiotherapy and surgery in early breast cancer (an overview of the randomized trials), *New England Journal of Medicine* **33**, 1444–1455.

[3] Fisher, B., Montague, E. & Redmond, C., Bartou, B., Borland, D., Fisher, E.R., Deutsch, M., Schwarz, G., Margolese, R., Donegan, W., Volk, H., Konvolinka, C., Gardner, B., Cohn, I. Jr, Lesnick, G., Cruz, A.B., Lawrence, W., Nealon, T., Butcher, H. & Lawton, R. (1977). Comparison of radical mastectomy with alternative treatments for primary breast cancer. A first report of results from a prospective randomized clinical trial, *Cancer* **39**, Supplement, 2827–2839.

[4] Gray, R.J. (1994). A Bayesian analysis of institutional effects in a multicenter cancer clinical trial, *Biometrics* **50**, 244–253.

[5] Gustafson, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data, *Biometrics* **53**, 230–242.

[6] Lan, K.K.G. & DeMets, D.L. (1989). Group sequential procedures: calendar versus information, *Statistics in Medicine* **8**, 1191–1198.

[7] Meinert, C.L. (1986). *Clinical Trials Design, Conduct, and Analysis*. Oxford University Press, Oxford.

[8] Multicenter American Research Trial With Cilazapril after Angioplasty to Prevent Transluminal Coronary Obstruction and Restenosis (MARCATOR) Study Group (1995). Effect of high dose angiotensin-converting enzyme inhibition on restenosis: final results of the MARCATOR study, a multicenter, double-blind, placebo-controlled trial of cilazapril, *Journal of the American College of Cardiology* **25**, 362–369.

[9] Multicenter European Research Trial with Cilazapril after Angioplasty to Prevent Transluminal Coronary Obstruction and Restenosis (MERCATOR) Study Group (1992). Does the new angiotensin-converting enzyme inhibitor cilazapril prevent restenosis after percutaneous transluminal coronary angioplasty? Results of the MERCATOR study: a multicenter, randomized, double-blind, placebo-controlled trial, *Circulation* **86**, 100–110.

[10] National Institutes of Health (1990). Consensus statement: treatment of early-stage breast cancer, in *NIH Consensus Development Conference* June 18–21, 1990, Vol. 8, No. 6. National Institutes of Health, Bethesda, pp. 1–19.

[11] Pocock, S.J. (1984). *Clinical Trials, A Practical Approach*. Wiley, New York.

[12] Skene, A.M. & Wakefield, J.C. (1990). Hierarchical models for multicenter binary response studies, *Statistics in Medicine* **9**, 919–929.

[13] Smigel, K. (1996). Beta-carotene fails to prevent cancer in two major studies; CARET intervention stopped [news]. [Clinical Trials. News], *Journal of the National Cancer Institute* **88**, 145.

[14] The Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study Group (1994). The effect of vitamin E and beta-carotene on the incidence of lung cancer and other cancers in male smokers, *New England Journal of Medicine* **330**, 1029–1035.

[15] Wittes, J. (ed.) (1995). Data management for multicenter studies: methods and guidelines, *Controlled Clinical Trials* **16**, 1S–178S.

JOHN BRYANT, WALTER CRONIN & SAM WIEAND

Multidimensional Scaling

Multidimensional scaling comprises a set of models and associated methods for constructing a geometrical representation of proximity or dominance relationships between elements in one or more sets of entities. While this characterization of multidimensional scaling may seem rather abstract, any further specification would unnecessarily narrow it to certain specific types of data or representation models.

The types of data to which multidimensional scaling can be applied can be categorized in terms of three attributes: the number of ways of the data array, the number of modes, and the type of relationship expressed by the data. The *number of ways* of a data set refers to the number of dimensions or factors of the data array. When the data can be arranged in a matrix, they are called two-way data. One way corresponds to the rows of the matrix and the other way corresponds to the columns. Three-way data are arranged in a three-dimensional array – the third way referring to the slices of the array. The *number of modes* indicates the number of different sets of entities to which the ways of the data array refer. If both the rows and the columns of a two-way data matrix index the same set of entities (such as objects, subjects, etc.), then the data are called *two-way one-mode data*. A typical example of a two-way one-mode data set is a **correlation** matrix. When the rows and columns of the data matrix refer to two different sets of entities (e.g. subjects and objects), the data are called *two-way two-mode data*. A rectangular matrix with ratings indicating the extent to which certain symptoms pertain to certain diseases is an example of such a data set. The set of symptoms constitutes one mode while the set of diseases constitutes the other mode. Finally, three-way data of which the ways index one, two, or three different sets of entities, are respectively referred to as *three-way one-mode, three-way two-mode*, or *three-way three-mode data*. A set of square symmetric matrices containing the correlations between the same set of variables on a number of different occasions constitutes an example of a three-way two-mode data set. Finally, a three-way three-mode data array can, for instance, be obtained by having N_1 physicians rate the extent to which N_2 patients exhibit a set of N_3 symptoms. Each of the ways of the resulting $N_1 \times N_2 \times N_3$ data array corresponds to a different set of entities, namely physicians, patients, and symptoms.

Multidimensional scaling can be applied to data that express two *types of relationship*: proximity relations and dominance relations. In proximity data the data values indicate the proximity (similarity or dissimilarity) between the entities to which their indices refer (*see* **Similarity, Dissimilarity, and Distance Measure**). If larger values indicate a greater proximity, then the data are called *similarity data*. If larger values indicate a smaller degree of proximity, then the data are *dissimilarity data*. (Note that similarity data can always be converted into dissimilarities by subtracting all values from a suitably large constant.) Thus, a correlation matrix constitutes a two-way one-mode set of similarity data. The other type of data relationships to which multidimensional scaling can be applied are dominance relationships. In dominance data, the data values indicate how strongly one entity dominates the other. A paired comparisons matrix, where each entry shows the percentage of times the row element is preferred to the column element, is an example of a two-way one-mode dominance matrix.

Historically, multidimensional scaling was developed for constructing a spatial representation of two-way one-mode proximity data. Later, it was extended to other types of data (such as three-way proximity data and dominance data) and to other types of models (such as nonspatial models). In the next section we discuss its most common form namely multidimensional scaling of two-way one-mode proximity data. In the subsequent sections we present some of the extensions to three-way two-mode data and to nonspatial models.

Two-Way Multidimensional Scaling

Starting from two-way one-mode symmetric proximity data, two-way multidimensional scaling attempts to represent the objects indexed by the rows and columns of the data matrix by points in a multidimensional space such that the interpoint distances correspond as well as possible to the observed proximity data in some well-defined sense. Being one-mode data, the rows and columns of the data matrix refer to a common set of N entities. These entities will be generically referred to as "objects". It will be assumed that the data are dissimilarities. The $N \times N$ data matrix will be denoted $\mathbf{\Delta} = ((\delta_{ij}))$, where δ_{ij}

indicates the observed dissimilarity between object i and object j. $\mathbf{\Delta}$ is assumed to be symmetric, i.e. $\delta_{ij} = \delta_{ji}$. Usually the entries on the main diagonal of $\mathbf{\Delta}$ are not observed and hence are undefined.

Such two-way symmetric proximity data can be obtained in several ways. In the behavioral and social sciences, it is common to have subjects judge the degree of similarity or dissimilarity between all $N(N-1)/2$ pairs of distinct objects on a numerical scale, yielding so-called *direct ratings*. Ordinal proximity data can be obtained, for example, by having subjects arrange the objects in **rank** order according to their proximity to a reference object. Or, subjects can compare two pairs of objects at a time, and indicate for each couple of pairs which pair contains the most similar (or dissimilar) objects. Alternately, proximities can be derived from *co-occurrence data*. In such a case the proximity measure is based on the number of occasions on which two objects co-occur. In scientometrics, for instance, the similarity between two journals can be defined as the number of times the two journals are cited in the same list of references. Or, when subjects were asked to sort a (large) set of objects into a number of mutually exclusive and exhaustive categories such that similar objects are put in the same category, a pairwise proximity measure can be derived from the number of subjects who put the two objects in the same category. Two-way symmetric proximity data can also be computed from *confusion* or *transition frequencies*. In this case the proximity is based on the number of times one object is confused or succeeded by another object. In sociology, for instance, the proximity between two professions could be computed from social mobility data indicating the number of parents with profession i who have a child with profession j. Finally, multidimensional scaling is often applied to *derived* or so-called *second-order proximities*. Starting from multivariate data (e.g. measurements of N objects on a number of variables), correlations, profile distances, or other derived proximity measures are computed to quantify the degree of association between the objects.

In multidimensional scaling the N objects will be represented as points in a multidimensional space, such that the interpoint distances approximate the observed dissimilarities δ_{ij} as well as possible. While extensions to other kinds of spaces exist, multidimensional scaling usually constructs a representation in a *Euclidean* space. The dimensionality of the space will be indicated by R. If x_{ir} denotes the coordinate of the point representing object i on the rth dimension, the Euclidean distance between the points representing objects i and j can be written as

$$d_{ij} = \left[\sum_{r=1}^{R}(x_{ir} - x_{jr})^2\right]^{1/2} = [(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)]^{1/2},$$

where $\mathbf{x}_i$ is an R-component column vector defined as $\mathbf{x}_i = (x_{i1}, \ldots, x_{iR})'$, The purpose of multidimensional scaling is to represent the N objects in an R-dimensional Euclidean space such that the distances d_{ij} are close to the observed δ_{ij}. However, the Euclidean distances d_{ij} do not approximate the observed dissimilarities δ_{ij} directly, but approximate some permissible transformation f of the observed dissimilarities:

$$d_{ij} \approx f(\delta_{ij}).$$

The type of transformation f that is applied depends on the measurement level of the data (*see* **Measurement Scale**). If the data constitute a ratio scale (i.e. if the dissimilarities are unique up to a positive similarity transformation), f is defined as

$$f(\delta_{ij}) = b\delta_{ij}, \quad b > 0.$$

Owing to the scale indeterminacy of the representation discussed later in the article, b can be set equal to 1 without loss of generality. When the data constitute an interval scale (i.e. when the data are unique up to a positive linear transformation), f is defined as

$$f(\delta_{ij}) = a + b\delta_{ij}, \quad b > 0.$$

Finally, when the proximities are ordinal data (e.g. rank order data), f is a weak monotone transformation such that

$$f(\delta_{ij}) \geq f(\delta_{kl}), \quad \text{if } \delta_{ij} > \delta_{kl}.$$

In the case of a tie, i.e. if $\delta_{ij} = \delta_{kl}$, either no restriction is imposed on the relationship between $f(\delta_{ij})$ and $f(\delta_{kl})$ or the two transformed dissimilarities are required to be equal, i.e. $f(\delta_{ij}) = f(\delta_{kl})$. The former approach has been labeled by Kruskal [16, 17], in his breakthrough work formulating a rigorous mathematical and numerical approach to two-way nonmetric multidimensional scaling, "the primary approach to ties", while the latter is known as the "secondary

approach to ties". The transformed dissimilarities $f(\delta_{ij})$ are sometimes called *target distances* or *optimally scaled data*. Multidimensional scaling thus attempts to find coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_N$ such that

$$d_{ij} \approx f(\delta_{ij}).$$

When the data are ordinal and a monotone transformation is allowed, the procedure is called *nonmetric* multidimensional scaling. When the proximities are interval or ratio level data, the procedure is referred to as *metric* multidimensional scaling.

The **goodness of fit** of a multidimensional scaling representation is based on a normalized sum of squared deviations between the transformed dissimilarities and the derived Euclidean distances:

$$\mathcal{L}_f(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{\sum_{i<j}^{N} [f(\delta_{ij}) - d_{ij}]^2}{\sum_{i<j}^{N} d_{ij}^2}.$$

The normalization is necessary to make the **loss function** independent of the scale of the derived space. Other normalization factors are sometimes used and may be preferable in certain cases. Note that $\mathcal{L}_f$ depends on the transformation f and that the optimal transformation f is unknown. To make the loss function independent of f, it is defined as

$$\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \underset{\text{all } f}{\text{minimum}} \mathcal{L}_f(\mathbf{x}_1, \ldots, \mathbf{x}_N).$$

The square root of $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is known in the literature as the "stress" function [16, 17]. In two-way multidimensional scaling, given the data $\mathbf{\Delta}$ and given a dimensionality R, coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_R$ are sought that minimize $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$. This optimization problem involves minimizing a nonlinear function in many variables and cannot be solved analytically. Instead iterative procedures are used. Such a procedure starts from some initial estimates of the coordinates and iteratively improves these estimates until no further decrease in $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is possible. While the solution obtained through such a procedure is known to be locally optimal, it is not guaranteed to be a global minimum $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$. The best way to safeguard against local optima is to carry out the analysis several times, starting each run from different initial estimates. If the same optimal solution is found repeatedly, then it is likely to constitute a global optimum.

It should be noted that $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is not affected by certain transformations of $\mathbf{x}_1, \ldots, \mathbf{x}_N$. More specifically, shifting the origin of the coordinates (i.e. replacing $\mathbf{x}_i$ by $\mathbf{x}_i + \mathbf{c}$, $i = 1, \ldots, N$, where $\mathbf{c}$ is an R-component constant vector) does not affect the goodness of fit of the solution. Likewise, any transformation of the form $\mathbf{x}_i \rightarrow \mathbf{T}\mathbf{x}_i$, $i = 1, \ldots, N$, with $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$, leaves the Euclidean distances unchanged and so does not affect the goodness of fit. This family of transformations includes permutations, reflections, and orthogonal rotations of the configuration. Finally, owing to the normalization factor, changing the scale of the configuration (i.e. replacing all $\mathbf{x}_i$ by $\alpha\mathbf{x}_i, \alpha \neq 0$) does not alter the goodness of fit. Therefore, a solution that minimizes $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is only unique up to the similarity transformations mentioned above and the dimensions of such a configuration can be freely translated, permuted, reflected, orthogonally rotated, and uniformly rescaled to facilitate the interpretation.

To evaluate the goodness of fit, the minimum value obtained for the loss function should be inspected as well as scatter plots of δ_{ij} vs. d_{ij}, $f(\delta_{ij})$ vs. d_{ij} and δ_{ij} vs. $f(\delta_{ij})$. As mentioned above, the loss function $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is minimized for a given dimensionality R. The appropriate dimensionality is usually not known a priori. Of course, the larger R, the better the goodness of fit that can be obtained, but the less data reduction will occur and the more complicated the solution will be. To select an appropriate dimensionality, the analysis is usually carried out for decreasing values of R. From a plot of the goodness of fit vs. the dimensionality, an appropriate value for R is selected (usually at the value where an "elbow" occurs).

Several software programs exist that implement two-way multidimensional scaling. One of the more often used programs is KYST-2A [19] that is fully documented in Kruskal & Wish [18].

To illustrate a two-way analysis, we apply KYST-2A to some data collected by Ekman [14], and originally analyzed by Shepard [24, 25] in his pioneering work on "analysis of proximities" describing the earliest approach to what later developed into modern two-way nonmetric multidimensional scaling. Ekman obtained similarity judgments from human subjects about 14 colors varying in wavelength from 434 mμ to 674 mμ. The colors were projected two at a time on a screen and the subjects were instructed to rate

the "qualitative similarity" on a five-point scale. The average similarity ratings are presented in Table 1. KYST-2A was applied to these data with R varying from 5 to 1, treating the data as ordinal data. The minimum values obtained for the stress function are plotted as a function of R in Figure 1. This figure clearly exhibits an "elbow" at $R = 2$. The associated two-dimensional configuration is presented in Figure 2. In the figure, the objects are labeled by their wavelength in mμ. The figure clearly reveals the well-known color circle ranging from violet (434 mμ) over blue (472 mμ), green (504 mμ), and yellow (584 mμ) to red (674 mμ). The input similarities are plotted against the resulting Euclidean distances in Figure 3. The figure clearly shows that the Euclidean distances approximate a monotonic transformation of the observed proximity data.

There is also an older (now called the "classical") approach to two-way metric multidimensional scaling, best described in Torgerson [26], that relies on a singular value decomposition (*see* **Correspondence Analysis**) of a symmetric estimated "scalar products" matrix derived via some preprocessing of the original proximity data. While this procedure, being based on the singular value decomposition, does not suffer from a local optimum problem, the loss function that it is optimizing does not have as clearly defined properties as $\mathcal{L}$ does.

The basic two-way multidimensional scaling procedure described above has been extended in many respects. Statistical formulations enabling **maximum likelihood** estimation have been developed (see [21] for a survey), as well as extensions to non-Euclidean metric (see [1] for a review of extensions to non-Euclidean Minkowski metrics). The families of permissible data transformations have been extended (e.g. [27]) and constrained procedures have been devised (e.g. [10]).

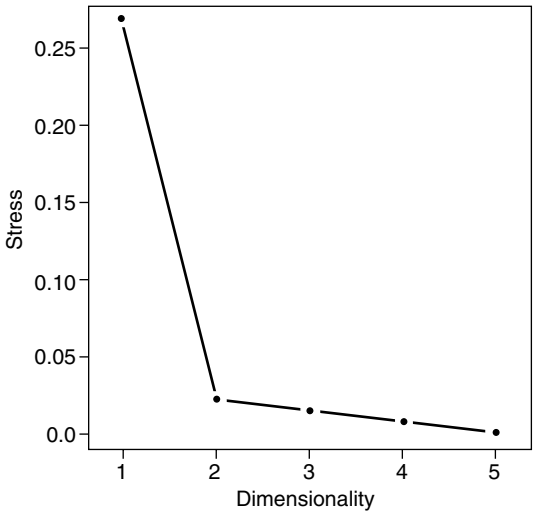

Figure 1 Minimum stress values obtained for the Ekman [14] data

One of the more important extensions of two-way multidimensional scaling has been to the case of two-way two-mode data (see [4] for a review). Such data can arise in a number of different ways; for instance, when proximity judgments are made between pairs of objects, where one object belongs to a set A and the other to a set B (with A and B disjoint sets), but

Table 1 Mean judged perceptual similarity between 14 spectral colors (data obtained by Ekman [14])

	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434	–	0.86	0.42	0.42	0.18	0.06	0.07	0.04	0.02	0.07	0.09	0.12	0.13	0.16
445	0.86	–	0.50	0.44	0.22	0.09	0.07	0.07	0.02	0.04	0.07	0.11	0.13	0.14
465	0.42	0.50	–	0.81	0.47	0.17	0.10	0.08	0.02	0.01	0.02	0.01	0.05	0.03
472	0.42	0.44	0.81	–	0.54	0.25	0.10	0.09	0.02	0.01	0.00	0.01	0.02	0.04
490	0.18	0.22	0.47	0.54	–	0.61	0.31	0.26	0.07	0.02	0.02	0.01	0.02	0.00
504	0.06	0.09	0.17	0.25	0.61	–	0.62	0.45	0.14	0.08	0.02	0.02	0.02	0.01
537	0.07	0.07	0.10	0.10	0.31	0.62	–	0.73	0.22	0.14	0.05	0.02	0.02	0.00
555	0.04	0.07	0.08	0.09	0.26	0.45	0.73	–	0.33	0.19	0.04	0.03	0.02	0.02
584	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	–	0.58	0.37	0.27	0.20	0.23
600	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	–	0.74	0.50	0.41	0.28
610	0.09	0.07	0.02	0.00	0.02	0.02	0.05	0.04	0.37	0.74	–	0.76	0.62	0.55
628	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.50	0.76	–	0.85	0.68
651	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.20	0.41	0.62	0.85	–	0.76
674	0.16	0.14	0.03	0.04	0.00	0.01	0.00	0.02	0.23	0.28	0.55	0.68	0.76	–

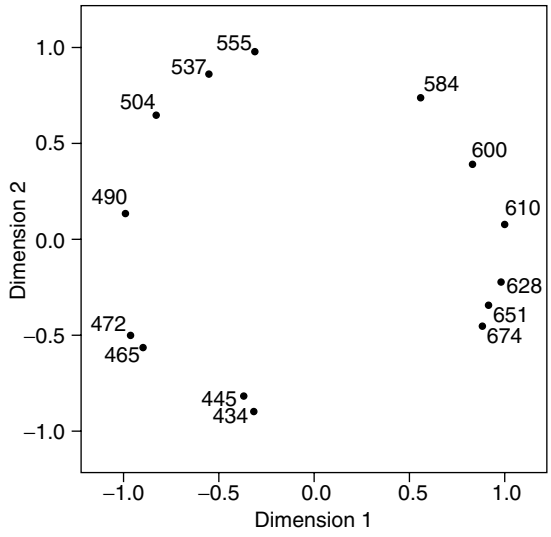

Figure 2 Two-dimensional representation of the Ekman [14] data

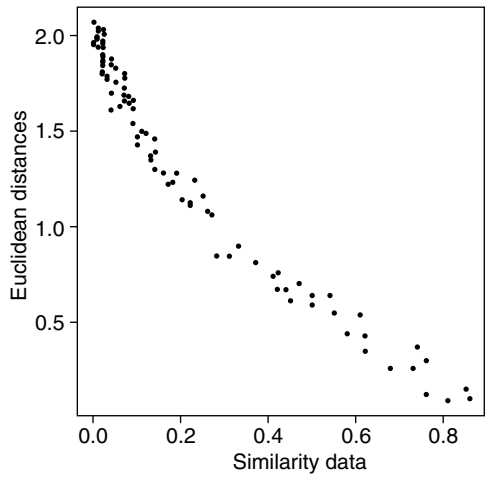

Figure 3 Plot of the similarities vs. the obtained two-dimensional Euclidean distances for the Ekman [14] data

no such judgments are made about pairs of objects both belonging to the same set. *A* and *B* may be sets of entities of basically the same kind, such as two different classes of diseases, or sets of very different types, such as a set of diseases, on the one hand, and a set of treatments on the other (where the "proximity" relation might be one of "effectiveness" of the treatment for the disease). One especially important class of two-way two-mode data to which this variant is often applied is a subjects (e.g. patients) by objects (e.g. medical practitioners) matrix with ratings or rankings of the preference for the objects by each of the subjects. This latter type of data – often called individual differences preferential choice data – can be viewed as a type of *dominance* data, measuring the relative dominance (tendency to be preferred to, or chosen over, other objects) of each of the objects, for each of the subjects. Coombs [9] has pointed out that such data can also be interpreted as *conditional* proximity data – conditional, since the ratings or rank orders of preference are comparable only *within* an individual subject (so only within the rows, of a subjects × object data matrix), but not *between* rows corresponding to different subjects. Coombs interprets preference (or other types of dominance) data as (conditional) proximities based on a very general model that assumes that a subject's preference for an object is inversely monotonically related to the distance of that object (in a subjective multidimensional space) from the subject's "ideal" (or most preferred) point. So the closer an object is to the ideal point, the more it is preferred. These are *conditional* proximities because one subject's ratings or rankings of preference (or other dominance relation) cannot be compared with those of a different subject since they are not measured on the same scale. This model is known as the *unfolding model* or *ideal point model*. Individual differences among subjects are represented in this general model by allowing for a different ideal point for each subject. Some specialized options are available in KYST-2A, entailing a different normalization or changes to the stress loss function (depending on the specific type of data being analyzed), for carrying out such an unfolding analysis. As emphasized by Carroll [4] and Kruskal et al. [19], it is important to use the *correct* options when analyzing such data, since the use of incorrect ones can lead to "degenerate" solutions that convey little or no information about the data, even though they may have a stress value indicating perfect or near-perfect fit.

It should be noted that many types of data, other than individual differences preference data, may be viewed as conditional proximities. For example, the disease × treatment matrix mentioned earlier would be of this same form if the treatments were, say, rank ordered for effectiveness separately for each disease. For reasons that will not be discussed here (but

see [4]) such analysis of two-way two-mode proximity data is often called *multidimensional unfolding*.

Three-Way Multidimensional Scaling

The extension of the basic two-way procedure that had the largest impact is the extension to three-way two-mode data. In three-way multidimensional scaling the input data consist not of a single square symmetric proximity matrix $\mathbf{\Delta}$, but of a series of matrices $\mathbf{\Delta}^{(1)}, \ldots, \mathbf{\Delta}^{(M)}$ containing dissimilarities about the same set of objects, but obtained from M different sources. The sources can be, for instance, different subjects, different occasions, etc. The INDSCAL model [6] – the most common three-way model – assumes that each source weights the dimensions of a common object space $\mathbf{x}_1, \ldots, \mathbf{x}_N$ idiosyncratically. The (nonnegative) weight that source i attaches to dimension r will be denoted w_{ir}. In the INDSCAL model, $\delta_{jk}^{(i)}$, the dissimilarity between objects j and k observed from source i, is represented by the weighted distance $d_{jk}^{(i)}$:

$$d_{jk}^{(i)} = \left[\sum_{r=1}^{R} w_{ir}(x_{jr} - x_{kr})^2\right]^{1/2}, \quad w_{ir} \geq 0,$$
$$= \left[\sum_{r=1}^{R} (y_{jr}^{(i)} - y_{kr}^{(i)})^2\right]^{1/2},$$

with

$$y_{jr}^{(i)} = (w_{ir})^{1/2} x_{jr}.$$

Thus the weighted distance $d_{jk}^{(i)}$ is equivalent to an ordinary Euclidean distance in a space $\mathbf{Y}^{(i)} = ((y_{jr}^{(i)}))$, where each dimension is rescaled by the square root of the corresponding source weight. The space defined by $\mathbf{X} = ((x_{jr}))$ is often called the *common space* or the *object space*, while $\mathbf{W} = ((w_{ir}))$ is referred to as the *source space*. $\mathbf{X}$ and $\mathbf{W}$ thus define two disjoint spaces, both having the same dimensionality R, one space (the common space) representing the objects and one space (the source space) representing the sources. The weight that source i applies to the rth dimension of the common space, w_{ir}, can be derived from the source space by simply projecting the point representing source i onto the rth coordinate axis. Finally, $\mathbf{Y}^{(i)}$, the common space rescaled for source i, is called the *private space* for source i (see, for example, Arabie et al. [2]).

Contrary to the distances obtained in two-way multidimensional scaling, the distances $d_{jk}^{(i)}$ are *not* invariant under orthogonal rotations of the object space (unless the ratio of the weights applied to a pair of dimensions is identical for all sources, in which case there exists an indeterminacy involving a generally nonorthogonal rotation in the plane defined by these two dimensions). The rotational uniqueness of the INDSCAL model often leads to representations that are easier to interpret. It is primarily this rotational uniqueness property that has made INDSCAL so popular. There are, however, some transformations of the parameters $\mathbf{X}$ and $\mathbf{W}$ that do not affect the weighted distances $d_{jk}^{(i)}$, namely permutations and reflections of the dimensions, a translation of the object space (i.e. replacing $\mathbf{x}_j$ by $\mathbf{x}_j + \mathbf{c}$ for $j = 1, \ldots, N$), and a joint transformation of the type

$$x_{jr} \rightarrow \alpha_r x_{jr}, \qquad w_{ir} \rightarrow \frac{w_{ir}}{\alpha_r^2},$$

with $\alpha_r \neq 0$ and $r = 1, \ldots, R$.

Several numerical procedures have been developed for fitting the INDSCAL model to metric or nonmetric three-way two-mode proximity data. The most widely used computer program for fitting the INDSCAL model to interval or ratio level proximity data is SINDSCAL [20], while probably the most common procedure to fitting INDSCAL to nonmetric (ordinal) data is the ALSCAL program [29]).

To illustrate the INDSCAL model, we present the results of an INDSCAL analysis originally carried out by Carroll & Chang [7] on some color perception data collected by Helm [15]. Helm [15] had 14 subjects (10 with normal color vision and four with a red–green color deficiency) judge the dissimilarity of ten colors that were approximately equally spaced on the color circle. Figure 4 displays the common space as well as the source space. The two dimensions of the common space can be interpreted as respectively a yellow–blue and a red–green dimension. As can be expected, the four color deficient subjects (labeled CD1–CD4 in the figure) clearly weight the yellow–blue dimension more heavily than the red–green dimension. Figure 5 presents the private spaces for two typical subjects, a subject with normal color vision (subject N7) and a color deficient subject (subject CD1). This figure illustrates how the private spaces are derived from the common space by idiosyncratically weighting the dimensions.

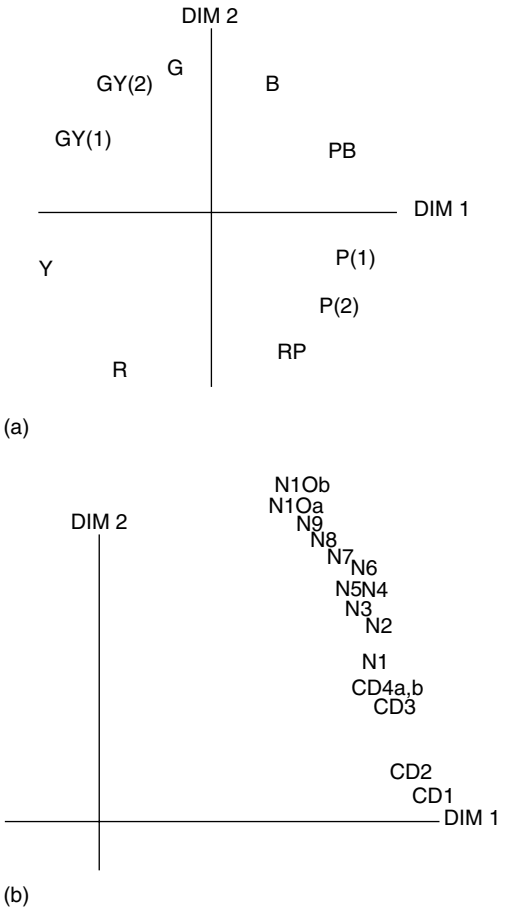

Figure 4 Two-dimensional common space (a) and source space (b) for the Helm [15] data. The objects are labeled as follows: R = red, Y = yellow, GY(1) = green yellow, GY(2) = green yellow with more green than GY(1), G = green, B = blue, PB = purple blue, P(1) = purple, P(2) = purple with more red than P(1), and RP = red purple. The subjects with normal color vision are labeled N1–N10, while the color-deficient subjects are labeled CD1–CD4

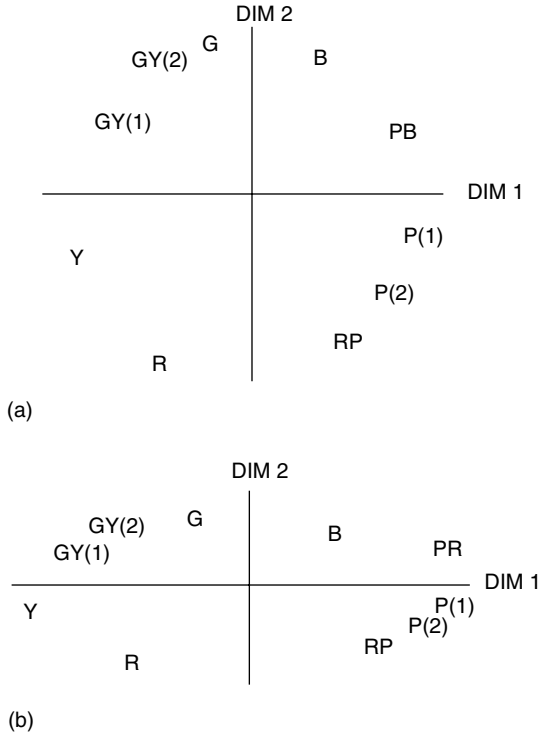

Figure 5 Private spaces for a subject with (a) normal color vision (N7) and (b) a color-deficient subject (CD1)

The basic three-way model has been extended in several ways, among others to allow for more elaborate idiosyncratic transformations of a common space (for a recent review, see [5]), to fit an "extended INDSCAL" model including "specific" as well as "common" dimensions, while allowing either metric or a special form of nonmetric fitting using a maximum likelihood criterion of fit [8], to handle individual differences in a more **parsimonious** way (for instance, through a **latent class** formulation [28]), or to accommodate other types of data such as three-way three-mode data or dominance data (for a review of three-way multidimensional scaling models for paired comparisons data, see [12]). Analysis of three-way three-mode proximity data (e.g. judgments of efficacy of each of several treatments for each of a number of diseases by different physicians), interpreted in terms of a generalization of the two-way two-mode ideal point model, is sometimes called *three-way unfolding* (see [11]).

Nonspatial Models

The multidimensional scaling methods discussed in the previous sections attempt a representation in a Euclidean or weighted Euclidean space. In addition, methods have been developed that arrive at a nonspatial representation using, instead of a Euclidean distance model, a tree or other kind of network model. The simplest and at the same time most often used nonspatial model is an ultrametric tree (*see* **Classification, Overview**). An ultrametric tree

is a rooted tree in which a nonnegative weight is attached to each node of the tree such that (i) the terminal nodes have zero weight, (ii) the largest weight is attached to the root, and (iii) the weights associated with the nodes on the path from any terminal node to the root constitute a strictly increasing sequence. In an ultrametric tree the distance between any two terminal nodes i and j, denoted d_{ij}, is defined as the maximum of the weights attached to the nodes on the path connecting i and j. These distances satisfy the so-called ultrametric inequality:

$$d_{ij} \leq \max(d_{ik}, d_{jk}),$$

for all i, j, k. Or, equivalently, the largest two of d_{ij}, d_{ik}, and d_{jk} are equal for all i, j, and k.

In an ultrametric tree representation of a set of two-way one-mode proximity data $\mathbf{\Delta}$, the objects indexed by the rows and columns of $\mathbf{\Delta}$ are represented by the terminal nodes of an ultrametric tree. The topology of the tree and the weights attached to the nodes of the tree are chosen so that the resulting ultrametric tree distances correspond as closely as possible to the observed dissimilarities. An ultrametric tree representation of $\mathbf{\Delta}$ defines a hierarchical clustering on the set of objects. Each internal node of the tree defines a partitioning of the objects represented by the terminal nodes of the subtree below that internal node. The weight attached to the internal node indicates the strength of this partitioning. The successive partitionings defined by the internal nodes are nested, thus yielding a hierarchical clustering.

This is illustrated in Figure 6, which presents the least squares ultrametric tree representation obtained by De Soete & Carroll [13] of some data collected by Rosenberg & Kim [23] concerning the similarity between kinship terms. Subjects were asked to group 15 kinship terms on the basis of their similarities in minimally two and maximally 15 categories. Dissimilarity data were derived by counting for each pair of kinship terms the number of subjects who put the two terms in different categories. Figure 6 displays the least squares ultrametric tree representation of the data obtained from the female subjects (listed in [22, Table 7.2]). As can be seen from the figure, the root node n_1 separates the direct kin (grandparents, grandchildren, parents, brother, sister) from the collaterals (uncle, aunt, cousin, nephew, niece). Within the cluster of the direct kin, node n_2 distinguishes the nuclear family from the kin that are two generations away from the ego. Within the last group,

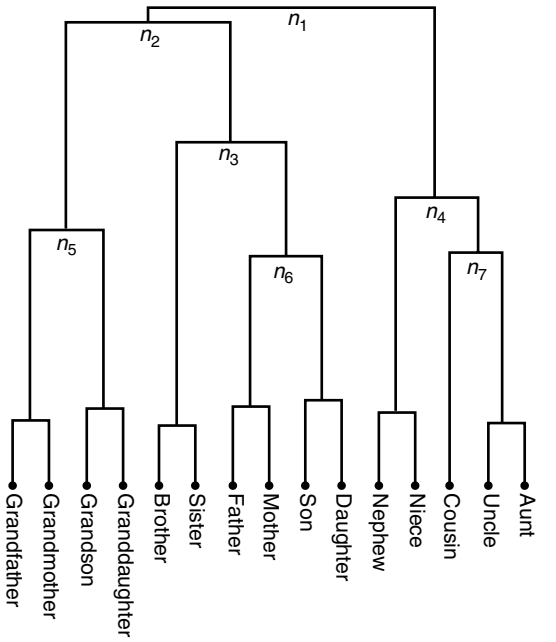

Figure 6 Ultrametric tree representation of the kinship data

node n_5 distinguishes those that are $+2$ generations (grandparents) from those that are -2 generations (grandchildren) away from the ego. Node n_3 separates the members of the nuclear family from the same generation (brother, sister) from those that are one generation apart. Within this last group, node n_6 distinguishes those that are $+1$ generation away from those that are -1 generation away from the ego. Among the collaterals, the same generation distinctions appear (nodes n_4 and n_7). At the lowest level, the kinship terms are distinguished on the basis of gender.

Besides ultrametric trees, other types of tree and network models have been used to represent proximity or dominance data. This includes nonhierarchical trees where the ultrametric is replaced by what is variously called an "additive", "path-length", or "four-point" metric (in such a tree a length or weight is associated with each branch or link and the distance between two objects is defined as the length of the unique path connecting the terminal nodes representing the two objects), as well as "multiple tree" models in which proximities are modeled via sums of distances associated with two or more distinct trees.

Also included are other network (or discrete) models such as general graph structures and overlapping or nonoverlapping clustering models. Furthermore, another class of representations of proximity data, seldom used to date, but having considerable potential for effective application, are "hybrid models" combining tree or other (discrete) network models with the (continuous) spatial dimensional models most typically associated with multidimensional scaling. Finally, tree models have been devised for analyzing various types of three-way data (for a comprehensive review, see [13]).

Concluding Comments

In this article, multidimensional scaling was introduced and its most common techniques have been described and illustrated. The reader wanting more detailed information is referred to the recent chapter by Carroll and Arabie ([5]) and to various chapters in Arabie et al. [3].

References

[1] Arabie, P. (1991). Was Euclid an unnecessarily sophisticated psychologist?, *Psychometrika* **56**, 567–587.

[2] Arabie, P., Carroll, J.D. & DeSarbo, W.S. (1987). *Three-Way Scaling and Clustering*. Sage, Newbury Park.

[3] Arabie, P., Hubert, L.J. & De Soete, G., eds (1996). *Classification and Clustering*. World Scientific, River Edge.

[4] Carroll, J.D. (1980). Models and methods for multidimensional analysis of preferential choice (or other dominance) data, in *Similarity and Choice*, E.D. Lantermann & H. Feger, eds. Hans Huber, Bern, pp. 234–289.

[5] Carroll, J.D. & Arabie, P. (1997). Multidimensional scaling, in *Handbook of Perception and Cognition*, Vol. 9, M. Birnbaum, ed. Academic Press, San Diego.

[6] Carroll, J.D. & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart–Young" decomposition, *Psychometrika* **35**, 283–319.

[7] Carroll, J.D. & Chang, J.-J. (1970). Reanalysis of some color data of Helm's by INDSCAL procedure for individual differences multidimensional scaling, in *Proceedings of the 78th Annual Convention*, American Psychological Association, Washington, pp. 137–138.

[8] Carroll, J.D. & Winsberg, S. (1995). Fitting an extended INDSCAL model to three-way proximity data, *Journal of Classification* **12**, 57–71.

[9] Coombs, C.H. (1964). *A Theory of Data*. Wiley, New York.

[10] de Leeuw, J. & Heiser, W. (1980). Multidimensional scaling with restrictions on the configuration, in *Multivariate Analysis*, Vol. 5, P.R. Krishnaiah, ed. North-Holland, Amsterdam, pp. 501–522.

[11] DeSarbo, W.S. & Carroll, J.D. (1985). Three-way metric unfolding via alternating weighted least squares, *Psychometrika* **50**, 275–300.

[12] De Soete, G. & Carroll, J.D. (1992). Probabilistic multidimensional models of pairwise choice data, in *Multidimensional Models of Perception and Cognition*, F.G. Ashby, ed. Erlbaum, Hillsdale, pp. 61–88.

[13] De Soete, G. & Carroll, J.D. (1996). Tree and other network models for representing proximity data, in *Clustering and Classification*, P. Arabie, L.J. Hubert, & G. De Soete, eds. World Scientific, River Edge, pp. 157–197.

[14] Ekman, G. (1954). Dimensions of color vision, *Journal of Psychology* **38**, 467–474.

[15] Helm, C.E. (1964). A multidimensional ratio scaling analysis of perceived color relations, *Journal of the Optical Society of America* **54**, 256–262.

[16] Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**, 1–27.

[17] Kruskal, J.B. (1964). Nonmetric multidimensional scaling: a numerical method, *Psychometrika* **29**, 115–129.

[18] Kruskal, J.B. & Wish, M. (1978). *Multidimensional Scaling*. Sage, Newbury Park.

[19] Kruskal, J.B., Young, F.W. & Seery, J.B. (1977). *How to use KYST-2A: A Very Flexible Program to do Multidimensional Scaling and Unfolding*. Unpublished User Manual. AT&T Bell Laboratories, Murray Hill.

[20] Pruzansky, S. (1975). *How to Use SINDSCAL: A Computer Program for Individual Differences in Multidimensional Scaling*. Unpublished User Manual. AT&T Bell Laboratories, Murray Hill.

[21] Ramsay, J.O. (1982). Some statistical approaches to multidimensional scaling data, *Journal of the Royal Statistical Society, Series A* **145**, 285–312.

[22] Rosenberg, S. (1982). The method of sorting in multivariate research with applications selected from cognitive psychology and person perception, in *Multivariate Applications in the Social Sciences*, N. Hirschberg & L.G. Humphreys, eds. Erlbaum, New York, pp. 117–142.

[23] Rosenberg, S. & Kim, M.P. (1975). The method of sorting as a data-gathering procedure in multivariate research, *Multivariate Behavioral Research* **10**, 489–502.

[24] Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. I, *Psychometrika* **27**, 125–140.

[25] Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. II, *Psychometrika* **27**, 219–246.

[26] Torgerson, W.S. (1958). *Theory and Methods of Scaling*. Wiley, New York.

[27] Winsberg, S. & Carroll, J.D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model, *Psychometrika* **54**, 217–229.
[28] Winsberg, S. & De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL, *Psychometrika* **58**, 315–330.
[29] Young, F.W. & Lewyckyj, R. (1981). *ALSCAL4 User Guide*. Unpublished User Manual. L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill.

(*See also* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis, Variables**)

GEERT DE SOETE & J. DOUGLAS CARROLL

Multilevel Models

Biostatistical data often have a hierarchical structure. Typically these structures are naturally occurring ones: animal populations are characterized by individuals nested within parents, themselves often nested within groups or herds which may also be nested within spatial entities. In other cases the structure may result from research designs, as in multicenter clinical trials (*see* **Multicenter Trials**) where patients are nested within clinics. In yet other cases the data may not obviously seem to be nested, yet viewing them as such may yield new insights or more efficient analysis techniques. Examples are repeated measure designs, where measurements are "nested" within individual subjects (*see* **Longitudinal Data Analysis, Overview**), and multivariate response data, where measurements are "nested" within individuals.

In addition to nesting relationships among data units we may also have cross-classifications. For example, an individual cow may be nested within a herd of cattle, but also be the offspring of parent stock, where any parent may contribute to several herds: individual cows are thus cross-classified by parents as well as nested within their herds. A further complexity is also often present whereby individual units at one level of a data hierarchy may be nested within more than one higher-level unit. An example is spatial data, where each individual person can be classified by the geographical locality where they live, but will also be influenced in terms, say, of their health or behavior, by surrounding localities. In this case we regard them as belonging to a primary unit plus a number of secondary units.

In the following sections I develop a set of models for describing such data, increasing in complexity as they move from simple hierarchies with continuously distributed responses, to cross-classifications and multivariate data and to discrete responses. Various extensions and special cases will also be considered. The emphasis is on model specification rather than estimation, although there is a brief section on the latter.

The Basic Multilevel Model

For simplicity consider a simple data structure where an outcome is measured on patients in a number of centers, together with one or more treatments or covariates. We wish to model a relationship between the outcome and the **explanatory variables**, taking into account the possibility that this relationship may vary across centers. We shall refer to the centers as higher-level units and patients as lower-level units. In the present case we just have two levels with centers as level 2 units and patients as level 1 units. A simple such model can be written as follows:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{ij}, \\ \text{var}(e_{ij}) &= \sigma^2_{e0}, \\ \text{var}(u_j) &= \sigma^2_{u0}, \end{aligned} \tag{1}$$

where y_{ij} is the response and x_{ij} the value of a single explanatory variable for the ith patient in the jth center. The slope coefficient β_1 is for the present assumed to be the same at all centers, while the random variable u_{0j} represents the departure of the jth clinic's intercept from the overall population intercept term β_0. The first two terms on the right-hand side of (1) constitute the fixed part of the model and the last two terms describe the random variation. We develop the model initially assuming that the random variables have a (multivariate) normal distribution, and discuss the nonnormal case later. This model could be viewed as a standard **analysis of covariance** if we treated each u_{0j} as a fixed parameter to be estimated. Such a model, however, often will be inappropriate, for the following reasons.

First, we may have a very large number of centers, leading to a very large number of separate parameters to estimate. Secondly, some of the clinics may have very few patients, so that their individual departures will be poorly estimated. Most importantly, we may be interested in treating the centers as a sample from a *population* of centers and wish to make general inferences about the likely behavior of other centers in this population rather than, or in addition to, providing separate estimates for each center in the sample. For all these reasons it will usually be more appropriate to regard u_{0j} as random and to write

$$u_{0j} \sim \text{N}(0, \sigma^2_{u0}), \qquad e_{0ij} \sim \text{N}(0, \sigma^2_{e0}).$$

We can also elaborate (1) by allowing the coefficient β_1 to vary across centers and rewrite the model in the more compact form

$$y_{ij} = \beta_{0ij} x_0 + \beta_{1j} x_{1ij},$$

$$\begin{aligned}
\beta_{0ij} &= \beta_0 + u_{0j} + e_{ij},\\
\beta_{1j} &= \beta_1 + u_{1j}, \qquad (2)\\
\mathbf{U} &= \{u_{0j}, u_{1j}\}, \qquad \mathrm{E}(\mathbf{U}) = \mathbf{0},\\
\mathrm{cov}(\mathbf{U}) &= \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}, \qquad \mathrm{var}(e_{ij}) = \sigma^2_e.
\end{aligned}$$

This model is often referred to as a "random coefficient model" by virtue of the fact that the coefficients β_{0ij} and β_{1j} in the first equation of (2) are random quantities (*see* **Random Effects**). It is possible, however, to have random coefficient models that are only single level (see below); we thus drop this term in order to emphasize the hierarchical data structure.

As more explanatory variables are introduced into the model we can choose to allow them random coefficients at the center level, thereby introducing further covariances as well as variances at level 2. This will lead to models with complex covariance structures. One of the aims of multilevel modeling is to explore such potential structures and also to attempt to explain them in terms of further variables. Having fitted such a model we can obtain posterior estimates for the individual "residuals" $(u_{0j}, u_{1j}, e_{0ij})$ at either level by estimating their expected values (or other functions of their distributions), given the data and model estimates. Thus, for example, we can estimate $\mathrm{E}(u_{0j}|Y, \beta, \theta)$, where

$$\boldsymbol{\beta}^{\mathrm{T}} = \{\beta_1, \beta_2\}, \qquad \boldsymbol{\theta} = \{\sigma^2_{u0}, \sigma_{u01}, \sigma^2_{u1}, \sigma^2_{e0}\}. \qquad (3)$$

The multilevel model is here described in non-Bayesian terms. For a full Bayesian specification of this model we would need to add prior distribution assumptions for the parameters in (3). The interested reader is referred, for example, to [4, 5] for details with examples.

In the next section we look at a general formulation and then some important special cases. A fully detailed treatment of the topics is not possible here and the reader is referred to [6] and [5] for details of methodology with examples and a discussion of computer software. A World Wide Web site is available which contains information about current developments, references, software, and so on, at `http://multilevel.ioe.ac.uk` (*see* **Internet**).

Cross-Classifications

Many data structures are not purely hierarchical, but mixtures of hierarchies and cross-classifications. For example, in a school health survey children may be assessed by raters, each school having just one rater. Thus we have a structure where children are grouped within cells defined by the cross-classification of raters by schools, and we wish to model the level-2 variation as a function of both the between-rater and between-school variation. If the design were changed, so that a separate team of raters visited each school and each child was measured by a single rater, then the cross-classification would be that of raters by children nested within schools. If, again, there was a single team of raters who visited every school, then the cross-classification would be of raters by children across the whole sample. In this case we have no separable hierarchy and we would wish to model the total response variation as a function of the between-child, between-school, and between-rater variation.

Rasbash & Goldstein [13] and Browne et al. [1] discuss various examples of this kind and set out the appropriate models together with procedures for efficient estimation. Corresponding to the first and second examples given above we can write the following models, using a more general notation for the fixed part of the model, where i indexes children, j_1 indexes schools, and j_2 indexes raters.

We write

$$y_{i(j_1j_2)} = X_{i(j_1j_2)}\beta + u_{j_1} + u_{j_2} + e_{i(j_1j_2)} \qquad (4)$$

for the first model with children nested within the level-2 cross-classification and with the following level-2 covariance structure

$$\begin{aligned}
\mathrm{cov}(y_{i(j_1j_2)}, y_{i'(j_1j'_2)}) &= \sigma^2_{u_1},\\
\mathrm{cov}(y_{i(j_1j_2)}, y_{i'(j'_1j_2)}) &= \sigma^2_{u_2}, \qquad (5)\\
\mathrm{var}(y_{i(j_1j_2)}) &= \mathrm{cov}(y_{i(j_1j_2)}, y_{i'(j_1j_2)})\\
&= \sigma^2_{u_1} + \sigma^2_{u_2}.
\end{aligned}$$

The second model is written as

$$y_{(i_1i_2)j} = X_{(i_1i_2)j}\beta + u_j + e_{i_1j} + e_{i_2j}. \qquad (6)$$

In both (4) and (6) we have assumed an **"additive" model** for the variance contributions, and the

adequacy of this can be tested against a model which includes an interaction term, e.g.

$$y_{(i_1 i_2)j} = X_{(i_1 i_2)j}\beta + u_j + e_{i_1 j} + e_{i_2 j} + e_{(i_1 i_2)j}. \quad (7)$$

In addition, we can have further random coefficients and levels of nesting or crossing.

Multiple Unit Membership

We have assumed so far that each lower-level unit, such as a school student or patient, belongs to just one higher-level unit of a particular kind. In many cases, however, such units may belong to more than one higher-level unit. For example, in a child growth study, children may change schools from one occasion to the next, and a particular case is that of spatial data where an individual is influenced by the geographical unit where she lives and also (with differing weights) by neighboring areas. We can write a simple two-level model of this kind as follows where, for simplicity, we suppose the maximum number of level-2 units to which a level-1 unit may belong is two:

$$\begin{aligned} y_{i(j_1 j_2)} &= X_{i(j_1 j_2)}\beta + w_{1ij_1}u_{j_1} \\ &\quad + w_{2ij_2}u_{j_2} + e_{i(j_1 j_2)}, \\ w_{1ij_1} + w_{2ij_2} &= 1, \\ \mathrm{var}(y_{i(j_1 j_2)}) &= (w_{1ij_1}^2 + w_{2ij_2}^2)\sigma_u^2 + \sigma_e^2, \quad (8) \\ \mathrm{cov}(y_{i(j_1 j_2)}, y_{i'(j_1 j_2)}) &= (w_{1ij_1}w_{1i'j_1} \\ &\quad + w_{2ij_2}w_{2i'j_2})\sigma_u^2, \\ \mathrm{cov}(y_{i(j_1 j_2)}, y_{i'(j_1' j_2)}) &= w_{2ij_2}w_{2i'j_2}\sigma_u^2. \end{aligned}$$

As before, we can further elaborate this model by allowing random coefficients, further hierarchical levels, and further crossing factors. For example, in the example of children changing schools we may cross-classify the schools by the neighborhoods where the children live with the possibility of multiple neighborhood membership in the above sense and across time. Browne et al. [1] discuss such models using MCMC estimation, with examples.

Repeated Measures Data and Multivariate Data

An interesting special case of a two-level structure is that of repeated measures models such as the following:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}, \quad (9)$$

where the response, say, is the weight of an animal related to a linear function of age, x, with the intercept and slope varying across animals (*see* **Random Coefficient Repeated Measures Model**).

Another important special case is that of multivariate data, where the response is a vector. Consider first a "single-level" multivariate linear model, with two responses, height and weight, measured on a sample of males and females. For the jth variable ($j = 0$ for height, $j = 1$ for weight) measured on the ith subject we have the following model equation:

$$\begin{aligned} y_{ij} &= \beta_{01}z_{1ij} + \beta_{02}z_{2ij} + \beta_{11}z_{1ij}x_j \\ &\quad + \beta_{12}z_{2ij}x_j + u_{1j} + u_{2j} \\ z_{1ij} &= \begin{cases} 1, & \text{if height,} \\ 0, & \text{if weight,} \end{cases} \\ z_{2ij} &= 1 - z_{1ij}, \\ x_j &= \begin{cases} 1, & \text{if female,} \\ 0, & \text{if male,} \end{cases} \quad (10) \\ \mathrm{var}(u_{1j}) &= \sigma_{u1}^2, \\ \mathrm{var}(u_{2j}) &= \sigma_{u2}^2, \\ \mathrm{cov}(u_{1j}, u_{2j}) &= \sigma_{u12}. \end{aligned}$$

A part of the data matrix for this structure might be as given in Table 1, so that at level 2 we have the variances and covariance of height and weight while there is no variation at level 1, and the fixed part of the model is defined using the relevant dummy variables associated with each response. Notice that in the data matrix the third individual has no weight measurement. By specifying the multivariate model as in (10) we can implicitly fit data where some responses are missing: we simply omit the relevant

Table 1 Example data for a repeated measures design

		Intercepts (z)		
Individual	Response	Height	Weight	Gender (x)
1 (female)	y_{11}	1	0	1
1	y_{12}	0	1	1
2 (male)	y_{21}	1	0	0
2	y_{22}	0	1	0
3 (female)	y_{31}	1	0	1

level-1 unit corresponding to the missing observation. The model can be generalized readily in the ways already discussed by allowing random coefficients, cross-classifications, etc. and further levels of nesting. An example of a multivariate model analysis will be given later.

Modeling Variances

In addition to specifying the average response as modeled in the fixed part of the model, we have discussed modeling the covariance structure at level 2 (and higher levels) by introducing random coefficients. We may also introduce random coefficients which vary across level-1 units and this provides a flexible general procedure for variance modeling. Consider the following model:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + (u_j + e_{0ij} + e_{1ij} x_{ij}), \\ \mathrm{var}(e_{0ij}) &= \sigma^2_{e0}, \qquad \mathrm{var}(e_{1ij}) = 0, \\ \mathrm{cov}(e_{0ij}, e_{1ij}) &= \sigma_{e01}, \end{aligned} \tag{11}$$

so that the level-1 contribution to the overall variance is the linear function

$$\sigma^2_{e0} + 2\sigma_{e01} x_{ij}.$$

Note that we have constrained one of the "variances" at level 1 to be zero in order to give a linear rather than a quadratic variance function. In fact, the parameters $\sigma^2_{e0}, \sigma^2_{e1}$, and σ_{e01} are not to be interpreted as separate variances and covariances, but simply as parameters defining the variance structure. The variable, x, may be any kind of explanatory variable. For example, if it were a dummy variable for gender, then the model would allow a separate level-1 variance for males and females. In this way it is possible to model the variance, as well as the mean, as functions of explanatory variables. Examples are given in Goldstein [5, Chapter 3].

In some circumstances, linear models for a variance, such as implied by (11), are inappropriate because they may predict an overall level-1 variance which is negative for part of the range. In this case we can consider alternative models where the level-1 variance has the form, for example,

$$\mathrm{var}(e_{ij}) = \exp(\beta^*_0 - \beta^*_1 x_{ij}), \tag{12}$$

which is nonnegative and where we require estimates of the β^*_0 and β^*_1. Goldstein [5] shows how maximum likelihood estimates for such models can be obtained.

Nonlinear and Generalized Linear Models

We can write a two-level **generalized linear model** in the form

$$\pi_{ij} = f(X_{ij}\beta_j), \tag{13}$$

where π_{ij} is the expected value of the response for the ijth level-1 unit and f is a nonlinear function of the "linear predictor" $X_{ij}\beta_j$, where we can have random coefficients at level 2. We need to specify a distribution for the *observed* response $y_{ij}|\pi_{ij}$: where the response is a proportion this is typically taken to be binomial, and where the response is a count taken to be Poisson (*see* **Poisson Regression**). Eq. (13) is a special case of a **nonlinear regression** model which is completed by specifying a suitable link function $f(\cdot)$. Thus, for binary response data we might have a simple model:

$$\begin{aligned} \mathrm{logit}(\pi_{ij}) &= \beta_0 + \beta_1 x_{1ij} + u_{0j}, \\ y_{ij} &\sim \mathrm{bin}(1, \pi_{ij}), \end{aligned} \tag{14}$$

with a corresponding model for counts using a log link function. The random part of (14) can be elaborated with further random coefficients, cross-classifications, etc.

These models can be extended to **multinomial** (ordered or unordered) responses [5, Chapter 7].

Survival Models

Survival time data (*see* **Survival Analysis, Overview**) will often have a multilevel structure: for example we may measure illness durations within centers or waiting times in hospitals with variation across centers and hospitals. We may also have repeated duration episodes *within individuals*, for example repeated periods of disease and remission, where different kinds of episode also may exist. We briefly mention here three common types of model and their multilevel specification. Further details are given by Goldstein [5].

The first type is the extension of the semiparametric **Cox regression model**, often referred to as a **frailty** model. When defining risk sets for this model we can choose to order our failure times across the

whole data set or *within* level-2 units, say hospitals. In the former case the marginal relationship between the hazard and the covariates is not generally proportional, and in the latter case it is proportional within level-2 units.

At each failure time l we define a response variate for each member of the risk set

$$y_{ijk(l)} = \begin{cases} 1, & \text{if } i \text{ is the observed failure,} \\ 0, & \text{if not,} \end{cases}$$

where i indexes the members of the risk set, and j and k level-1 and level-2 units, respectively. The response is treated as a Poisson variate with mean function for a simple **variance components** model given by

$$\pi_{jk(l)} = \exp(\alpha_l + X_{jk}\beta + u_k), \tag{15}$$

where there is a "blocking factor" α_l for each failure time. The second type of model is a "log duration" or **accelerated failure time** model which can be written as

$$l_{ij} = \ln(t_{ij}) = X_{ij}\beta_j + e_{ij}, \tag{16}$$

for the failure times t_{ij}. This is in the standard form for a two-level random coefficient model. A complication is that we may have (level-1) censored observations, and this implies that we need a careful specification of the level-1 distribution to incorporate **censoring** information in the estimation. Some common choices are the normal, **extreme value**, and log-**gamma** distributions (*see* **Parametric Models in Survival Analysis**).

The third type of model, which leads to a particularly simple form, is the discrete time **proportional hazards model**. For a two-level model we write

$$\log[-\log(1 - \pi_{jk(l)})] = X_{jk}\beta_k + \alpha_{(l)}, \tag{17}$$

where, as before, the $\alpha_{(l)}$ are constants to be estimated, one for each time interval. This leads to a model where the response is a binomial variate, being the number of deaths divided by the number in the risk set at the start of the interval. As with the first type, any censored observations in an interval are excluded from the risk set.

Estimation

The basic model assumes multivariate normality and standard (as well as restricted) maximum likelihood methods are available using Fisher scoring, iterative generalized least squares (*see* **Generalized Linear Model**) or the **EM algorithm**. Bayesian estimation is available using **Markov chain Monte Carlo** (MCMC) methods such as Gibbs sampling [4], which is also available for generalized linear models with the appropriate distributional assumptions. An alternative in this case is to use **quasi-likelihood** estimation together with appropriate bias correction procedures [7], or the related **generalized estimating equation** (GEE) procedure, [10]. For inference, interval estimates are obtained directly from MCMC and via large sample deviance statistics or bootstrapping for likelihood estimation. Maximum likelihood procedures are also available [5].

An Example

To illustrate the flexibility of multilevel models we fit a bivariate two-level model where one response is normal and the other is binary.

The data are part of the "Health and Lifestyle Survey", a sample of 9003 individuals within households nested within 396 electoral wards in Britain and carried out in 1984/85. For present purposes data on smoking habits are analyzed using information about gender and age. Further details are given in [3]. The information about smoking behavior consists of whether or not the respondent smoked cigarettes and if they did, how many per day. Sixty-five percent did not smoke and the mean number smoked for those who did is 15.2 with a standard deviation of 9.3. The distribution of the number smoked is positively skewed which suggests a normalizing transformation. The use of this, however, does not substantially alter the results and the analysis is presented in terms of the actual number smoked.

One aim of the analysis is to ascertain how the probability of smoking and the number smoked each relate to the explanatory variables. The other is to estimate the between-area variation, and in particular to see whether areas where the proportion of nonsmokers is high are also the areas where smokers tend to smoke greater numbers of cigarettes. We write the model in two parts.

For the binary response probability

$$\begin{aligned} \text{logit}\,(\pi_{ij}) &= (x_1\beta_1)_{ij} + u_{1j} \\ y_{ij} &\sim \text{binomial}\,(1, \pi_{ij}) \end{aligned} \tag{18}$$

For the continuous response

$$y_{ij} = (x_1\beta_1)_{ij} + u_{2j} + e_{ij}$$

with

$$e_{ij} \sim N(O, \sigma_e^2), \begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} \sim N(O, \Omega_u),$$

$$\Omega_u = \begin{pmatrix} \sigma_{u1}^2 & \\ \sigma_{u12} & \sigma_{u2}^2 \end{pmatrix}$$

where u_{1k} and u_{2k}, respectively, refer to the ward-level contributions t o the discrete and continuous parts of the model. This model combines a model for smokers where the response is the number of cigarettes smoked and a model with a binary response which is whether or not the subject smoked. Thus, each smoker will have two responses, a "1" for the binary response variable and the number smoked for the continuous response. Each nonsmoker will have just one response, a "0" indicating that they are a nonsmoker.

This model can be fitted with the MLwiN software package [14]. The bivariate structure is modeled as level 1, where there is no random variation, so that the full model is three-level. The results are presented in Table 2.

At the electoral ward level there is a high correlation (0.81) between the proportion of smokers and the number smoked. Men are more likely to be smokers and to smoke more and there is an age effect for the number smoked, with a maximum among 50 year olds, and declining thereafter. The relationship is weaker for the probability of smoking. A model that allowed gender to have a random coefficient at level 2 was fitted, but a large sample test for the extra variance and two covariance terms gave a χ^2 value of 6.8 on three degrees of freedom ($P = 0.08$). Attempting to fit the age coefficient as random at level 2 produces a zero estimated variance. We can also test the assumption of binomial variation for the smoking response by fitting extra binomial variation. This is estimated as 0.98, where a value of 1.0 corresponds to binomial variation with a standard error of 0.015, providing little evidence of extra binomial variation (*see* **Overdispersion**).

Table 2 Bivariate model for smoking/nonsmoking and number smoked. Gender is coded 1 for male and 0 for female: age is measured about the mean of 45.9 years. The level-1 variance is constrained to 1.0 which corresponds to binomial variation

	Response	
Parameter	Binary (se)	Continuous (se)
Fixed		
Intercept	−0.54	15.7
Gender	0.14 (0.05)	2.82 (0.32)
Age	−0.03 (0.03)	1.22 (0.21)
$(\text{Age})^2$	0.0011 (0.0007)	−0.02 (0.005)
$(\text{Age})^3$	−0.000012 (0.000005)	0.00009 (0.00003)
Random		
Level 2:		
Intercept variance	0.17 (0.03)	1.45 (0.81)
covariance	0.40 (0.11)	
Level 1:		
Intercept		79.2 (2.1)

Further Topics

Finally, we mention briefly some further topics, most of which are currently the subject of methodological research.

The standard **meta-analysis** model can be viewed as a special case of a general multilevel model. For the jth study in such an analysis we can define the standardized effect d_j where this is a dimensionless quantity. It may, for example, be a correlation coefficient, a standardized regression coefficient, group difference, or weighted group difference. We can write a simple model as follows:

$$d_j = \delta + v_j + u_j, \qquad \text{var}(u_j) = \sigma_j^2,$$
$$\text{var}(v_j) = \sigma_v^2, \tag{19}$$

where in the usual case σ_j^2 is assumed known and is treated as an offset in the random part of the model, but may also in some circumstances be estimated. The parameter δ is the population parameter of interest and σ_v^2 is the between-study (level-2) variance of the standardized effect. We can add random coefficients and covariates representing study factors to (19) in an attempt to explain between-study differences, which is a further aim of meta-analysis studies. Goldstein et al. [8] give a detailed discussion.

As in single-level models, **diagnostics** are important. We can estimate standardized **residuals** at any level of a data hierarchy and study these together with looking for influential units. A detailed discussion is given in [9].

Further important issues are those concerned with missing units and missing data generally, especially where the missingness is informative, and research is being conducted in this area (*see* **Nonignorable Dropout in Longitudinal Studies**). Another topic which is actively being researched is that of multilevel structural equation modeling [11, 12, 6].

Software

Some of the major software packages, for example SAS, STATA, and GENSTAT, can handle many, although not all, of the models described in this article (*see* **Software, Biostatistical**). Several general-purpose software packages have been written, e.g. HLM [2] and MLwiN [14]. A review of these packages has been carried out by [15].

References

[1] Browne, W., Goldstein, H. & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models, *Statistical Modelling* **1**, 103–124.

[2] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Sage, Newbury Park.

[3] Duncan, C., Jones, K. & Moon, G. (1996). Health-related behaviour in context: a multilevel modeling approach, *Social Science and Medicine* **42**, 817–830.

[4] Gilks, W., Richardson, S. & Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[5] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Ed. Edward Arnold, London, Wiley, New York.

[6] Goldstein, H. & Browne, W. (2002). Multilevel factor analysis modelling using Markov Chain Morte Carlo estimation, in G. Marcoulides & I. Moustaki (eds). *Latent Variable and Latent Structure Models*, Lawrence Erlbaum, London.

[7] Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses, *Journal of the Royal Statistical Society, Series A* **159**, 505–513.

[8] Goldstein, H., Yang, M., Omar, R., Turner, R. & Thompson, S. (2000). Meta analysis using multilevel models with an application to the study of class size effects, *Journal of Royal Statistical Society, C* **49**, 399–412.

[9] Langford, I. & Lewis, T. (1998). Outliers in multilevel data, *Journal of Royal Statistical Society, A* **161**, 121–160.

[10] Liang, K.-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[11] McDonald, R.P. & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data, *British Journal of Mathematical and Statistical Psychology* **42**, 215–232.

[12] Muthen, B.O. (1994). Multilevel covariance structure analysis, *Sociological Methods and Research* **22**, 376–398.

[13] Rasbash, J. & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model, *Journal of Educational and Behavioural Statistics* **19**, 337–350.

[14] Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., Lewis, T. (2000). *A User's Guide to MLwiN*, (2nd edn), Institute of Education, London.

[15] Zhou X., Perkins, A.J. & Hui, S.L. (1999). Comparisons of software packages for generalized linear multilevel models, *American Statistician* **53**, 282–290.

(*See also* **Nonlinear Mixed Effects Models for Longitudinal Data**; **Random Coefficient Repeated Measures Model**)

HARVEY GOLDSTEIN

Multilocus (Gene × Gene Interaction)

The concept of multilocus (inter-locus) or **gene** × gene **interaction** is often used without being precisely defined. In essence, gene × gene interaction refers to departure from "independence" of the effects of different genetic loci in the way that they combine to cause disease. However, this concept has been confused by the fact that what is meant by independence and the precise definitions of interaction used by biologists, epidemiologists, statisticians and human and quantitative geneticists have often differed, even when using identical terminology.

Interactions between loci are sometimes referred to as *epistatic* interactions or *epistasis*. This term was first used by Bateson [1] to describe a masking effect in which a variant at one locus prevents the variant at another locus from manifesting its effect. This concept of gene × gene interaction is often employed by a biologist or biochemist when investigating biologic interaction between proteins. In quantitative genetics, however, the term *epistatic* has classically been used to refer to a deviation from additivity in the effects of alleles at different loci with respect to prediction of a quantitative phenotype, in particular by Fisher [7]. Note that this definition is NOT equivalent to the Bateson [1] definition. Epistasis in the Fisher [7] sense is closer to the usual concept of statistical interaction: departure from a (specific) linear model describing the relationship between predictive factors (here assumed to be alleles at different genetic loci) and an outcome or phenotype of interest. Note that with this definition, the choice of scale becomes important since factors that are additive with respect to an outcome measured on one scale may exhibit interaction when a different, transformed scale is used [10, 11].

Mathematically, the quantitative genetic concept of gene × gene interaction may be represented for two loci by the linear model

$$y = \beta_0 + \beta_{a_1} x_1 + \beta_{d_1} z_1 + \beta_{a_2} x_2 + \beta_{d_2} z_2 + \beta_{i_{aa}} x_1 x_2 + \beta_{i_{ad}} x_1 z_2 + \beta_{i_{da}} z_1 x_2 + \beta_{i_{dd}} z_1 z_2, \quad (1)$$

where y is a quantitative phenotype, and x_i and z_i are dummy variables related to the underlying **genotype** at locus i. For example, for a diallelic locus with alleles denoted 1 and 2, we might set $x_i = 1$ and $z_i = -0.5$ for a 1/1 homozygote, $x_i = 0$ and $z_i = 0.5$ for a heterozygote and $x_i = -1$ and $z_i = -0.5$ for a 2/2 homozygote. The coefficients β_0, β_{a_1}, β_{d_1}, β_{a_2} and β_{d_2} represent genetic parameters to be estimated corresponding to the mean effect and additive and dominance effects at loci 1 and 2; $\beta_{i_{aa}}$, $\beta_{i_{ad}}$, $\beta_{i_{da}}$ and $\beta_{i_{dd}}$ correspond to the interaction effects. The lack of a gene × gene interaction in this scenario corresponds to all interaction coefficients being equal to 0.

In **human genetics**, the phenotype of interest is often qualitative and usually dichotomous, denoting presence or absence of disease. Models for the joint action of, and interaction between, loci have typically focused on the **penetrance**, the probability of developing disease given genotype. Let p_{ij} be the probability of developing disease given that there is genotype i at locus 1 and j at locus 2. Three common models have been considered [14]: an additive model in which p_{ij} may be written as $p_{ij} = \alpha_i + \beta_j$, where α_i and β_j are parameters representing the contributions of the different genotypes at locus 1 and 2, respectively; a heterogeneity model in which p_{ij} may be written as $p_{ij} = \alpha_i + \beta_j - \alpha_i \beta_j$; and a multiplicative model in which p_{ij} may be written as $p_{ij} = \alpha_i \beta_j$. The additive and heterogeneity models are usually assumed to represent nonepistatic models and to correspond to a situation in which the biologic pathways involved in disease are at some level separate or independent. The multiplicative model is usually considered to be an epistatic model in which the loci and pathways involved are not independent. Note, however, that a multiplicative model can be considered to be an additive model when transformed to the logarithmic scale. In a statistical sense, therefore, the multiplicative model signifies independent additive effects of the loci on a logarithmic scale.

Two other models are commonly used in human genetics. A model popular with epidemiologists is an additive model for the logit or the logarithm of the odds, in which $\ln[p_{ij}/(1 - p_{ij})]$ may be written as $\ln[p_{ij}/(1 - p_{ij})] = \alpha_i + \beta_j$. The lack of an interaction term γ_{ij} (i.e. the fact that when the model is expressed as $\ln[p_{ij}/(1 - p_{ij})] = \alpha_i + \beta_j + \gamma_{ij}$, the γ_{ij} parameter equals 0) signifies independence of the locus effects on the logit scale. Another popular model from classical genetics is a threshold model, in which the loci are assumed to contribute to an underlying, unobserved, continuous trait in an additive fashion and development of disease occurs if

this trait exceeds a certain threshold [8, 13, 18, 19]. Note that both of these models, although additive and therefore expressible without interaction effects as defined on their original scales, correspond to models with interactive effects (epistasis) when transformed to the penetrance scale.

Although in human genetics the penetrance or some function of the penetrance is often used as a surrogate for a quantitative phenotype of interest, treatment of gene × gene interactions is generally easiest when the phenotype has a genuine quantitative scale of measurement. Two popular methods for analysis of quantitative traits in families using pedigree data are the *variance component method* and the *Haseman–Elston method* (*see* **Linkage Analysis, Model-free**). The variance component method models the phenotypic covariance between relatives in terms of the underlying identity-by-descent (ibd) sharing probabilities at one or more genetic loci (*see* **Identity Coefficients**). This contrasts with the Haseman–Elston method [12] and subsequent extensions [6, 9, 20] in which some function (such as the squared difference or product) of the phenotype values for a pair of relatives is modeled in a **regression** framework in terms of the underlying ibd-sharing probabilities. Both of these methods generalize quite easily to account for epistatic interactions between loci. For the Haseman–Elston method in particular, all that is required is to include products of ibd-sharing probabilities at different loci as predictors in the regression (6). Although epistatic components of variance are often ignored in initial studies of linkage to quantitative traits, in some cases these components can be relatively large, and detecting these inter-locus interactions may in fact prove to be a more powerful strategy for the detection of genetic effects than concentrating solely on the independent effects of the individual loci [17].

The relationship between the differing definitions of gene × gene interaction is quite complex, and hence the degree to which statistical modeling can elucidate underlying biologic mechanisms may be limited. The problem is one that has been recognized for some time in epidemiology, namely, that any given data pattern and statistical model can usually be obtained from a number of different underlying mechanisms or models for disease development [15, 16]. This makes biologic inference from the results of a statistical interaction test very difficult [4]. Only if a prior biologic model can be postulated in detail is it likely that statistical modeling will allow insight into the underlying biologic mechanisms. Although direct biologic inference may be limited, identification of the most parsimonious statistical model for the joint effects of several loci, including interactions, can be useful for prediction of phenotype and targeting of interventions. Moreover, the increase in power that in some cases is obtained by allowing for different modes of interaction between potential disease loci can lead to identification of disease loci that might otherwise remain undetected [2, 3, 5].

References

[1] Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge, p. 79.

[2] Cho, J.H., Nicolae, D.L., Gold, L.H., Fields, C.T., LaBuda, M.C., Rohal, P.M. et al. (1998). Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1, *Proceedings of the National Academy of Sciences* **95**, 7502–7507.

[3] Cordell, H.J., Wedig, G.C., Jacobs, K.B. & Elston, R.C. (2000). Multilocus linkage tests based on affected relative pairs, *American Journal of Human Genetics* **66**, 1273–1286.

[4] Cordell, H.J., Todd, J.A., Hill, N.J., Lord, C.J., Lyons, P.A., Peterson, L.B., Wicker, L.S. & Clayton, D.G. (2001). Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes, *Genetics* **158**, 357–367.

[5] Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. & Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans, *Nature Genetics* **21**, 213–215.

[6] Elston, R.C., Buxbaum, S., Jacobs, K.B. & Olson, J.M. (2000). Haseman and Elston revisited, *Genetic Epidemiology* **19**, 1–17.

[7] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[8] Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, London/New York/Oxford, p. 111.

[9] Forrest, W.F. (2001). Weighting improves the "new Haseman–Elston" method, *Human Heredity* **52**, 47–54.

[10] Frankel, W.N. & Schork, N. (1996). Who's afraid of epistasis?, *Nature Genetics* **14**, 371–373.

[11] Greenland, S. & Rothman, K.J. (1998). Concepts of interaction, in *Modern Epidemiology*, 2nd Ed., R. Winters & E. O' Connor, eds. Lippincott–Raven, Philadelphia, pp. 329–342.

[12] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[13] Pearson, K. (1900). Mathematical contributions to the theory of evolution VIII: on the inheritance of characters not capable of exact quantitative measurement, *Philosophical Transactions, Series A* **195**, 79–150.

[14] Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models, *American Journal of Human Genetics* **46**, 222–228.

[15] Siemiatycki, J. & Thomas, D.C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis, *International Journal of Epidemiology* **10**, 383–387.

[16] Thompson, W.D. (1991). Effect modification and the limits of biological inference from epidemiologic data, *Journal of Clinical Epidemiology* **44**, 221–232.

[17] Tiwari, H.K. & Elston, R.C. (1998). Restrictions on components of variance for epistatic models, *Theoretical Population Biology* **54**, 161–174.

[18] Wright, S. (1934). An analysis of variability in number of digits in an inbred strain of guinea pigs, *Genetics* **19**, 506–536.

[19] Wright, S. (1934). The results of crosses between inbred strains of guinea pigs, differing in number of digits, *Genetics* **19**, 537–551.

[20] Xu, X., Weiss, S., Xu, X. & Wei, L.J. (2000). A unified Haseman–Elston method for testing linkage with quantitative traits, *American Journal of Human Genetics* **67**, 1025–1028.

HEATHER J. CORDELL

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution to more than two possible discrete outcomes. The roll of a die, for example, can give rise to one of six possible results. More generally, consider an experiment or survey in which there are k (≥ 2) distinct outcomes. Every realization of the experiment results in one of these k possible outcomes. The probability that each experiment results in the jth outcome ($j = 1, \ldots, k$) is denoted by p_j. The probabilities $p_1, \ldots, p_k$ are nonnegative and sum to one. In the example of the roll of a fair die, p_1 through p_6 are all equal to 1/6.

The multinomial distribution describes the joint distribution of a collection of frequencies of the k outcomes arising from n independent replications of this experiment. After n (≥ 1) independent experiments, let N_j denote the random variable counting the number of experiments that result in the jth outcome. The probability of jointly witnessing the collection of N_1 occurrences of the first categorical outcome, N_2 occurrences of the second, and so on through N_k occurrences of the kth outcome is

$$\Pr(N_1, \ldots, N_k) = \frac{n!}{N_1!N_2!\cdots N_k!} p_1^{N_1} p_2^{N_2} \cdots p_k^{N_k}. \tag{1}$$

The collection of counts $N_1, \ldots, N_k$ in (1) are said to follow a multinomial distribution. Each N_j can take on the value $0, 1, \ldots, n$ subject to the constraint that the sum of all counts $N_1 + \cdots + N_k$ is always equal to n, the number of experiments. The parameter n is referred to as the sample size or index of the distribution. When k is equal to two, (1) reduces to the probability mass function of the **binomial distribution**. There are

$$\binom{n+k-1}{k-1}$$

distinct outcomes of $N_1, \ldots, N_k$ in the multinomial distribution given by (1).

As an example of multinomial data, Robertson [5] describes the F_2 progeny of a cross between hybrid barley plants. There are four possible phenotypes (visible outcomes) from this cross: green non-two-row; green two-row; chlorina non-two-row; and chlorina two-row. According to Mendelian inheritance, these should occur in the ratio of 9 : 3 : 3 : 1 (*see* **Mendel's Laws**). The $k = 4$ probabilities $p_1, \ldots, p_4$ are then 9/16, 3/16, 3/16, and 1/16. In a total of $n = 1898$ barley plants, the counts of the four phenotypes $N_1, \ldots, N_4$ reported by Robertson [5] are 1178, 291, 273, and 156, respectively.

All marginal and conditional multinomial distributions are also multinomial. Any count N_j taken alone has a binomial marginal distribution with parameters n and p_j (*see* **Marginal Probability**). More generally, any subset of $N_1, \ldots, N_k$ also follows the multinomial distribution. The conditional distribution of the counts $N_2, \ldots, N_k$ given the fixed value of N_1 is a $k-1$ category multinomial distribution with sample size $n - N_1$ and probabilities $p_2/(1-p_1), \ldots, p_k/(1-p_1)$ (*see* **Conditional Probability**). In other words, all subsets of $N_1, \ldots, N_k$ follow a multinomial distribution. Fixing the values of a subset of $N_1, \ldots, N_k$ results in a multinomial distribution for the remaining counts.

The moments of the multinomial distribution are as follows. The mean of N_j is

$$\mathrm{E}(N_j) = np_j$$

and its variance is

$$\mathrm{var}(N_j) = np_j(1-p_j).$$

These are the same as the corresponding moments of the binomial distribution. The covariance of N_i and $N_j (i \neq j)$,

$$\mathrm{cov}(N_i, N_j) = -np_i p_j,$$

and their **correlation**

$$\mathrm{corr}(N_i, N_j) = -\left[\frac{p_i p_j}{(1-p_i)(1-p_j)}\right]^{1/2},$$

are both negative, because the frequencies $N_1, \ldots, N_k$ are constrained to sum to n. Intuitively, these negative correlations appear because increasing the count in any one of the k categories will reduce the counts in all other categories.

There is a close connection between the multinomial distribution and the **Poisson distribution**. Let $X_1, \ldots, X_k$ denote independent Poisson distributed counts with respective means $\lambda_1, \ldots, \lambda_k$. Given the value of the sum $\sum X_i$, the conditional distribution of $X_1, \ldots, X_k$ is multinomial with index $n = \sum X_i$ and probabilities $\lambda_1/\sum \lambda_i, \ldots, \lambda_k/\sum \lambda_i$.

This close connection between the Poisson and multinomial distributions is often confusing, and led to a debate between Haldane [3] and Cochran [1] over which distribution was more appropriate in determining the proper degrees of freedom for the **chi-square test**. In the barley example cited above, should the observed frequencies $N_1, \ldots, N_4$ be treated as four independent Poisson counts or as a single multinomial sample? The answer is that these counts most probably came about as independent Poisson observations. However, any inference such as estimating the p_js will always be done conditional on having observed $n = 1898$ plants. In other words, any statistical inference to be drawn from these data will be the same whether the figure of 1898 barley plants was determined before the experiment was conducted or if, in fact, this number was determined by some random process, as is more likely the case.

There are two important approximations to the multinomial distribution that are useful when the index n is large. The first of these is the joint multivariate normal distribution of the standardized counts

$$\frac{N_j - np_j}{[np_j(1-p_j)]^{1/2}}$$

obtained when $p_1, \ldots, p_k$ are held fixed and n is allowed to grow.

The second approximation points out another connection between the multinomial and Poisson distributions. This comes about in the same manner as the Poisson approximation to the binomial distribution. This approximation is obtained when n is large and a subset of the probabilities $p_1, \ldots, p_j$ $(j < k)$ become small at such a rate that the multinomial means np_i $(i = 1, \ldots, j)$ have finite, nonzero limits. In this case the frequencies $N_1, \ldots, N_j$ behave approximately as independent Poisson counts.

Estimation

The most common situation encountered in practice is when the sample size n is known and $p_1, \ldots, p_k$ must be estimated from the observed counts $N_1, \ldots, N_k$. In certain settings, however, not all of the $N_1, \ldots, N_k$ are observable and n is unknown. A problem in which both n and $p_1, \ldots, p_k$ are to be estimated often appears as **capture–recapture** surveys, in which the goal is to determine the size of a closed population. These surveys arise in wildlife management or epidemiologic studies in which there is a need to estimate the number of animals in a region or the number of diseased, but not yet diagnosed, individuals in the population. In the example of the barley data, there may be a fifth category, the genetic composition of which is always fatal to the plant resulting in an unobservable phenotype.

The usual problem assumes that n is known and estimates of $p_1, \ldots, p_k$ are needed. If nothing is known about the structure of $p_1, \ldots, p_k$, then the empirical frequencies

$$\hat{p}_i = \frac{N_i}{n}$$

are commonly used to estimate $p_1, \ldots, p_k$. The $\hat{p}_1, \ldots, \hat{p}_k$ are the maximum likelihood estimates of $p_1, \ldots, p_k$ because they maximize the probability given in (1). The $\hat{p}_1, \ldots, \hat{p}_k$ are unbiased (*see* **Unbiasedness**) for $p_1, \ldots, p_k$ and have the smallest variances of all unbiased estimators of $p_1, \ldots, p_k$.

There are many settings in which something is known about the mathematical structure of $p_1, \ldots, p_k$. This is usually expressed as a **loglinear model**. Loglinear models are multiplicative models and are so named because they are linear after taking the logarithm. These models are most useful in describing interactions of the various factors in multidimensional count data.

In the barley data, an example of a loglinear model is the model of independent linkage between color (green or chlorina) and the two-row phenotype. This model is suggested after writing the data as the two-by-two table given in Table 1. In the model of independence, $p_1, \ldots, p_4$ are estimated such that their odds ratio

$$\psi = \frac{p_1 p_4}{p_3 p_4}$$

is equal to one. The loglinear model $\psi = 1$ describes a multiplicative relationship between $p_1, \ldots, p_4$. This relationship and the corresponding estimates of

Table 1 The barley data arranged as a two-by-two table

		Two-row phenotype	
		No	Yes
Color	Green	1178	291
	Chlorina	273	156

$p_1, \ldots, p_4$ are familiar as the model of independence of rows and columns in the two-by-two table.

Bayesian methods for estimating the $p_1, \ldots, p_k$ parameters are subjective and incorporate prior knowledge. Bayesian methods require that we quantify any uncertainty the investigator may have about this knowledge before the data are observed. A general reference for Bayesian methods in the context of discrete data analysis is [4]. For the barley data, Robertson might have expected the Mendelian ratio of 9 : 3 : 3 : 1, but would have accepted another model if the data provided a large amount of evidence otherwise. A reasonable estimate for him to consider is a weighted average of the (prior) Mendelian rates p_i^{M} and the empirical frequencies $\hat{p}_i$ obtained from the data. That is, a Bayesian estimate p_i^{B} of p_i takes the form

$$p_i^{\mathrm{B}} = \left(\frac{c}{c+n}\right) p_i^{\mathrm{M}} + \left(\frac{n}{c+n}\right) \hat{p}_i \qquad (2)$$

where p_i^{M} is the Mendelian frequency, known prior to conducting the experiment. For all values of $c > 0$ and sample sizes $n > 0$, the Bayesian estimate p_i^{B} always lies between the prior (Mendelian) value p_i^{M} and the empirical fraction $\hat{p}_i$. A useful feature of the Bayes estimate given in (2) is the ability to produce nonzero estimates of p_i when the corresponding observed frequency N_i is zero.

The parameter $c > 0$ in (2) quantifies the level of certainty in the Mendelian model for the current problem. A large value of c indicates a lot of faith in this model, for which only a small amount of weight should be given to the data in the form of the empirical frequencies $\hat{p}_1, \ldots, \hat{p}_k$. On the other hand, either a large degree of uncertainty (small c) or a lot of data (large n) results in a heavy reliance on the data through $\hat{p}_1, \ldots, \hat{p}_k$ and relatively little to the prior model. A philosophical difference that many have with Bayesian methods is the subjective and seemingly arbitrary way that a value for c is assigned in (2). Different investigators will have different degrees of certainty and arrive at different Bayesian estimates of $p_1, \ldots, p_k$.

Four different estimates of the barley frequencies are summarized in Table 2. Mendel's theory of inheritance suggests that these four probabilities $p_1^{\mathrm{M}}, \ldots, p_4^{\mathrm{M}}$ should be in the ratio of 9 : 3 : 3 : 1. The maximum likelihood estimates $\hat{p}_1, \ldots, \hat{p}_4$ are the empirical frequencies N_i/n. The Bayesian estimates p_i^{B} given in (2) are a weighted average of the empirical ($\hat{p}_i$) and Mendelian (p_i^{M}) rates. The value of $c = 1000$ is used here as an illustration. The fitted loglinear model specifies that color and two-row phenotypes are not genetically linked and appear as independent characteristics.

Testing Fit

There has been a large amount of attention paid to testing the fit of the multinomial model (*see* **Goodness of Fit**). Relatively less is known about examining the adequacy of the multinomial distribution to explain the data. Instead, virtually all of this work has been directed at testing whether the $p_1, \ldots, p_k$ have been correctly estimated and appropriately explain the data. Best known in this area is the Pearson χ^2 statistic,

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i},$$

which dates back to the year 1900. Large values of χ^2 relative to its reference distribution indicate a significant lack-of-fit in the $p_1, \ldots, p_k$ parameters.

In more recent years, the χ^2 statistic has also been used to measure **overdispersion** of the multinomial model. Overdispersion refers to the situation in which the variances of the counts $N_1, \ldots, N_k$ are greater

Table 2 Four different estimates of the probabilities for the barley data

Phenotype		Observed count,	Mendelian probability,	Maximum likelihood,	Fitted loglinear	Bayesian estimate,
Color	Two-row	n_i	p_i^{M}	$\hat{p}_i$	model	p_i^{B}
Green	No	1178	0.5625	0.6207	0.5917	0.6006
Green	Yes	291	0.1875	0.1533	0.1823	0.1651
Chlorina	No	273	0.1875	0.1438	0.1728	0.1589
Chlorina	Yes	156	0.0625	0.0822	0.0532	0.0754

than predicted by the multinomial model. It is impossible to distinguish between the case of overdispersion and misspecification of the $p_1, \ldots, p_k$ parameters unless multiple samples are available.

When n is large and the $p_1, \ldots, p_k$ are replaced in χ^2 by appropriate estimates, then χ^2 behaves as χ^2 with $k - t - 1$ degrees of freedom, where t is the number of parameters estimated in $p_1, \ldots, p_k$. Under these conditions, it has long been known that there are many other statistics the behavior of which is closely tied to the χ^2 statistic. Among these test statistics are the generalized likelihood ratio

$$G^2 = 2 \sum_i N_i \log \frac{N_i}{np_i},$$

(*see* **Likelihood Ratio Tests**), the Freeman–Tukey χ^2

$$Z^2 = 4 \sum_i [N_i^{1/2} - (np_i)^{1/2}]^2,$$

and the Neyman χ^2

$$N^2 = \sum_i \frac{(N_i - np_i)^2}{N_i}.$$

All of these statistics, others not mentioned, and many not yet fully documented are contained in a family of statistics first described by Cressie & Read [2]. The power divergence statistics are of the form

$$2nI^\lambda = \frac{2}{\lambda(\lambda + 1)} \sum N_i \left[\left(\frac{N_i}{np_i} \right)^\lambda - 1 \right]$$

and are indexed by the parameter λ. Different values of λ result in many well known χ^2 equivalent statistics as special cases or as limits. For examples, the value of $\lambda = 1$ yields the χ^2 statistic; $\lambda = -1/2$ gives the Z^2 statistic; and as λ gets close to zero, $2nI^\lambda$ behaves as G^2. If n is very large and all p_i are replaced by their estimates using the correct model, then $2nI^\lambda$ should be close in value to χ^2 regardless of the value of λ.

A final method for examining goodness of fit is the use of exact tests. These computer-intensive methods involve a complete enumeration of all possible distinct outcomes of the multinomial likelihood function given in (1). This results in an exact test of significance in the sense that significance levels are determined exactly and no asymptotic approximations are needed. As an example of an exact test of significance, let us examine the fit of the barley data to the rates predicted by Mendel's model. The probability of observing the barley data assuming the Mendelian probabilities is 2.177×10^{-17}. This probability is not the significance level for testing Mendel's model. The exact significance level is obtained by enumerating all possible outcomes of the $n = 1898$ plants into the four distinct phenotypes. There are

$$\binom{n + k - 1}{k - 1} = \binom{1901}{3} = 1.143 \times 10^9$$

of these outcomes. The exact probability of an outcome with a likelihood of 2.177×10^{-17} or less is equal to 5.883×10^{-12}. This latter figure is the exact significance level and indicates a poor fit to the model of Mendelian inheritance for this data.

References

[1] Cochran, W.G. (1937). Note on J.B.S. Haldane's paper: "The exact value of moments of the distribution of χ^2", *Biometrika* **29**, 407.

[2] Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

[3] Haldane, J.B.S. (1937). The exact value of the moments of the distribution of χ^2, used as a test of goodness of fit, when expectations are small, *Biometrika* **29**, 133–143.

[4] Lindley, D.V. (1964). The Bayesian analysis of contingency tables, *Annals of Mathematical Statistics* **35**, 1622–1634.

[5] Robertson, D.W. (1937). Maternal inheritance in barley, *Genetics* **22**, 104–113.

(*See also* **Categorical Data Analysis**; **Exact Inference for Categorical Data**)

DANIEL ZELTERMAN

Multiple Comparisons

Multiplicity considerations arise in experimental research when it is desired to make inferences about several aspects of a problem simultaneously, while controlling some aspect of the frequency properties of the statistical procedure (*see* **Simultaneous Inference**). For example, observational units may generate multivariate responses and it may be of interest to examine **covariate** effects on each response (*see* **Multiplicity in Clinical Trials**) Alternatively, interest may lie in carrying out multiple analyses on the basis of subgroups of patients defined a priori. Repeated significance tests, often carried out to ensure early detection of effective treatments in **clinical trials**, also raise multiplicity issues (*see* **Data and Safety Monitoring**). In all of these cases, several tests of significance (*see* **Hypothesis Testing**) are typically carried out. If carried out naively, then the probability of making one or more false positive conclusion is typically higher than expected. This is the so-called multiplicity problem of statistical inference. Here we focus on issues pertaining to the classical multiple comparisons problem arising in the comparison of several populations with respect to a single **response variable**. For other examples of multiplicity, *see* **Simultaneous Inference**.

Consider an experiment with the objective of comparing the means of I populations. Suppose a sample of n_i subjects is available for study from the ith population. Let $n_\cdot = \sum_{i=1}^{I} n_i$, and let Y_{ij} denote the response variable for the jth subject in the ith sample. We further suppose the responses are generated according to the linear model $Y_{ij} = \mu_i + E_{ij}$, where μ_i is the mean response for the ith population, $E_{ij} \sim \mathrm{N}(0, \sigma^2)$ are independently distributed, and σ^2 is a common variance parameter reflecting the extent of the sampling variability, $j = 1, \ldots, n_i$, $i = 1, \ldots, I$. Here and throughout, we make the distinction between random variables and their realized values by using upper and lower case letters, respectively. Thus, if y_{ij} denotes a realization of Y_{ij}, then $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ denotes the mean of the ith sample, and s^2 the pooled estimate of the common variance (*see* **Analysis of Variance**).

Comparisons of the population means $\mu_1, \ldots, \mu_I$ may be based on tests of hypotheses or **interval estimation**. We consider issues pertaining to hypothesis tests, and point out that directly analogous issues arise in the context of interval estimation; specific comments on this follow. In this article we emphasize applications arising in **clinical trials** and related biopharmaceutical experiments in which the I samples consist of individuals randomized to one of I groups undergoing different treatment regimens. In what follows we use related terminology and will typically make reference to treatment comparisons.

There is a variety of contexts in which one might be interested in making inferences about differences in the population means, with the specific features of the problem leading to particular analysis strategies. The structure of the problem described above is that of a one-way analysis of variance (ANOVA), suggesting that if the **null hypothesis** were the equality of all I means, a corresponding statistic following the ***F* distribution** on $(I - 1, n_\cdot - I)$ **degrees of freedom** would be a natural choice. In many contexts, however, such an approach provides inadequate insight, since rejection of the null hypothesis does not furnish information regarding the nature of the treatment differences. In general, multiple comparison procedures are directed at facilitating more detailed analyses to gain such insight, while adjusting for the multiplicity by controlling certain frequency properties of the testing procedure.

Note, however, that there are many contexts in which multiple treatment comparisons can be made without the need to adopt procedures that adjust for multiplicity. Cox [7] has pointed out that probabilities regarding the simultaneous correctness of many statements may not always be of direct relevance. Dunnett [13] points out that in situations where multiple experimental treatments are used in the same study to maximize efficiency in the use of resources rather than for the purpose of making joint inferences, it is reasonable to make treatment to reference group (control) comparisons as if the data had been collected from different studies. Examples of such scenarios include Finney [19] and Redman & Dunnett [42]. This approach is consistent with the views put forth by Cook & Farewell [6] who propose that in more general contexts when multiplicities arise, such multiplicity adjustments are often adopted unnecessarily. Cook & Farewell suggest that provided a limited number of well-defined questions are posed at the design stage, and these questions relate to different features or different treatments under study, then the case can be made for avoiding multiple

testing procedures. As can be seen by the vagueness of the above statements, it remains difficult to characterize clearly situations in which multiplicity adjustments are required and when they are not, and so it is natural that some debate in this area will continue.

In what follows we discuss and contrast strategies that are generally appropriate if there is genuine concern about the need for making adjustments for multiplicity.

Overview and Terminology

Background

For the well-known case in which $I = 2$, suppose the null and **alternative hypotheses** are $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$, respectively. Hypothesis testing procedures are typically formulated by specification of a *discrepancy measure*, a many-to-one function of the random response variables which is a stochastic measure of the "distance" between the observed responses and what would be anticipated under the null hypothesis. A discrepancy measure must have a known distribution for specified values of the parameters of interest. Upon collecting the data, one may compute a realized value for the discrepancy measure, which is typically referred to as a *test statistic*. The ***P* value** of the test is the probability, under the null hypothesis, of observing a realized value of the discrepancy measure as extreme or more extreme than that observed. Thus, small P values indicate that the data are inconsistent with the null hypothesis. The *significance level*, denoted by α, is a specified threshold value such that if the observed P-value is $\leq \alpha$, H_0 is rejected in favor of H_a (*see* **Level of a Test**). Corresponding to a given threshold significance level is a *critical value*, c, such that test statistics larger than or equal to c lead to rejection of the null hypothesis.

A type I error is said to be committed if the null hypothesis is rejected when it is in fact true (*see* **Hypothesis Testing**). The *type I error rate*, a decision-theoretic notion introduced by Neyman & Pearson [40], corresponds to the rate with which this error would be made in a hypothetically infinite population of repetitions of the trial. A test procedure is said to control the type I error rate at α if the probability of a type I error is less than or equal to α.

The type I error rate may be interpreted probabilistically and hence one may write Pr(reject $H_0|H_0$ is true) $\leq \alpha$ (the equality will typically hold when the discrepancy measure has a continuous distribution and the null hypothesis specifies a single point in the parameter space). For the case in which $I > 2$, it might be tempting to carry out multiple hypothesis tests in an effort to learn more about the nature of any potential treatment differences. The principal difficulty with this approach is that multiple hypothesis tests at a common significance level α will result in a probability of committing *one or more* type I errors that may be substantially larger than α. Structure and rigor are added to the multiple testing procedures to ensure that the type I error rate properties are known, or at least controlled.

Formulation and Terminology of Multiple Comparison Procedures

As a first step in formalizing multiple comparison procedures, Hochberg & Tamhane [26] define a *family* as a "collection of inferences for which it is meaningful to take into account some combined measure of error". Tamhane [50] further states that there should be a "contextual relatedness" for inferences grouped into a common family. Tests pertaining to this family are directed at investigating this aspect of the treatments. Note that there may be more than one family of hypotheses in a given experiment, with each family addressing a different research question. Furthermore, since these questions may be interrelated, the families might not be disjoint (i.e. they may share one or more component hypotheses). For the purposes of this discussion it is sufficient to consider a single family, and we do so for the remainder of this article.

For concreteness we consider a family as consisting of a collection of null and alternative hypotheses $\{(H_{k0}, H_{ka}), k = 1, 2, \ldots, K\}$, where K denotes the total number of hypotheses in the family. The *familywise error rate* (FWE) is defined as the probability of making one or more false positive conclusions over all hypothesis tests in a particular family. Control of the FWE is appropriate if one cannot tolerate any type I error in the family no matter how many of the K null hypotheses are true. A procedure is said to have *strong control* of the FWE if the probability of making at least one type I error over all hypothesis tests of the family is at most α, regardless of how many component null hypotheses may be true; *weak control* of the FWE at α is achieved

if this type I error rate is guaranteed to be at most α only when all null hypotheses are true. Typically multiple comparison procedures control the FWE at α, thus satisfying Pr(at least one null hypothesis is falsely rejected) $\leq \alpha$. Procedures of this sort clearly also control the type I error rates of any subset of the family, including the component tests, while guaranteeing that the FWE does not exceed a specified level.

Another term often used is the *per-comparison error rate* (PCE). Here we restrict consideration to true null hypotheses in the family, and define the PCE as the expected number of false positive conclusions divided by the number of true null hypotheses. This error rate therefore corresponds to the usual type I error rate for individual hypotheses that are tested without any adjustment for multiplicity.

Suppose that m (which is unknown) hypotheses are true and $K - m$ are false. Denote by T the number of true hypotheses that are rejected (false positives) and by F the number of false hypotheses that are rejected (true positives). T and F are random variables. Benjamini & Hochberg [3] defined the false discovery rate (FDR) as the expected value of $T/(T + F)$. By comparison, PCE is the expected value of T/m and FWE $= \Pr(T > 0)$. They proposed that a multiple testing procedure control FDR $\leq \alpha$, instead of FWE $\leq \alpha$. They showed that FDR is equivalent to FWE when $m = K$ (all hypotheses are true). Thus, it provides weak control of FWE. When several hypotheses are false, it is less conservative than controlling the FWE and may provide a useful concept for situations where strict control of the FWE is not needed.

Multiple comparison procedures may be classified as single-step or stepwise procedures. In a *single-step procedure*, multiple tests are carried out using the same critical value for each component test. Procedures that involve carrying out multiple tests in sequence, using critical values which may be unequal, are called *stepwise* multiple testing procedures. For such stepwise procedures it is convenient to arrange the test statistics in ascending order according to their significance levels, and to arrange the component hypotheses conformably. Single-step procedures are attractive in some respects since they are simpler to apply and they have direct connections to **simultaneous confidence intervals**. They generally have lower **power** than stepwise procedures, however, and so are not desirable for the purposes of hypothesis testing.

Stepwise procedures for multiple comparisons may be further classified as step-down or step-up procedures. In *step-down* procedures, formal tests are carried out in a stepwise fashion starting with the most extreme outcome (i.e. the most significant test statistic). Testing proceeds to the hypothesis corresponding to the next most extreme outcome only upon rejection of the current hypothesis. If the current null hypothesis is not rejected, then all subsequent null hypotheses (i.e. those corresponding to the test statistics with the less extreme outcomes) are not rejected. Thus, if the first test statistic does not exceed its corresponding critical value, then the testing terminates with failure to reject any null hypothesis. Critical values are derived to ensure control of the FWE. In *step-up* testing procedures, the testing begins with the statistic corresponding to the least significant outcome. Testing continues and statistics are examined of progressively more extreme outcomes until a null hypothesis is rejected. At this point, all null hypotheses corresponding to the more extreme test statistics are also rejected. As in the step-down procedures, appropriate critical values are determined to ensure control of the FWE. Examples of single-step, step-down, and step-up multiple testing procedures are provided in subsequent sections.

In many cases an overall null hypothesis may be satisfied if and only if several less restrictive hypotheses are satisfied. Let $\mathrm{H}_0 = \bigcap_{k=1}^{K} \mathrm{H}_{k0}$ be the overall null hypothesis that all H_{k0} are true, and $\mathrm{H}_\mathrm{a} = \bigcup_{k=1}^{K} \mathrm{H}_{k\mathrm{a}}$ the corresponding alternative hypothesis that at least one H_{k0} is false. Testing H_0 against H_a is known as a **union–intersection** problem, due to Roy [44]. Denoting the test statistic for H_k by t_k for $k = 1, \ldots, K$, the test statistic for H_0 is $\max(t_1, \ldots, t_K)$. The critical value c is determined so that the type I error of the test is α, which requires that c be chosen to satisfy the following K-variate probability requirement:

$$\Pr(T_1 < c, \ldots, T_K < c | \mathrm{H}_0) = 1 - \alpha.$$

The solution, $c = c_K$ say, will be larger than the α-point of the univariate statistic, the difference representing an adjustment for the multiplicity.

For the *intersection–union* problem, where $\mathrm{H}_0 = \bigcup_{k=1}^{K} \mathrm{H}_{k0}$ and $\mathrm{H}_\mathrm{a} = \bigcap_{k=1}^{K} \mathrm{H}_{k\mathrm{a}}$, the test statistic is $\min(t_1, \ldots, t_K)$. To determine the value of the critical

constant in this case, Berger [4] demonstrated that the α-point of the univariate statistic is the correct value to use so that, in effect, no multiplicity adjustment is needed.

If $\{H_{k0}\}_{k=1}^{K}$ denotes a family of null hypotheses, then the *closure* of this family is formed by considering all intersections $H_S = \bigcap_{k \in S} H_{k0}$, where $S \subseteq \{1, 2, \ldots, K\}$. A closed testing procedure operates by rejecting any H_S if and only if every H_R is rejected by an α-level test for $R \supseteq S$. Marcus et al. [37] show that this strategy controls the FWE.

Historical Remarks

Suppose the null hypothesis consists of common means ($H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$) and the alternative is that at least one mean is different. Upon application of standard ANOVA methods and rejection of the null hypothesis, it is natural to want to examine the nature of the apparent treatment differences. Fisher [20] was among the first to propose a formal multiple comparison procedure with a view to investigating potential treatment differences following a standard one-way analysis of variance. Fisher's *protected least significant difference* procedure operates as follows. If the F test from the one-way analysis of variance is carried out with a type I error rate α, and if it fails to lead to rejection of the null hypothesis of common means, the procedure terminates. If $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is rejected, then all pairwise tests are carried out with a PCE of α for each. This procedure can be shown to have only weak control of the FWE.

An alternative, suggested by Fisher, is to proceed directly to the K specific treatment comparisons of interest and carry out these tests with a PCE error rate α/K. This is referred to as a **Bonferroni** adjustment to the per-comparison error rates that maintains strong control of the FWE at α. It is well known to be conservative, particularly for highly **correlated** test statistics [41]. The Bonferroni procedure is an example of a single-step multiple test procedure since each test is carried out with the same critical value, regardless of the outcomes of any of the other tests.

Scheffé [46] developed an approach for simultaneous inference which follows naturally from the one-way analysis of variance. In particular, if one considers all **contrasts** of the form $\sum_{i=1}^{I} \ell_i \mu_i$, where $\sum_{i=1}^{I} \ell_i = 0$, Scheffé's multiple comparison procedure generates a set of tests (confidence intervals) that have a FWE (simultaneous coverage probability) less than or equal to $\alpha (\geq 1 - \alpha)$. Note that with the appropriate choice of coefficients these contrasts may correspond to pairwise comparisons. Finally, we note that if the F test leads to rejection of the null hypothesis of common means, then there exists a vector of coefficients $\ell = (\ell_1, \ldots, \ell_I)'$ such that a test of $H_0 : \sum_{i=1}^{I} \ell_i \mu_i = 0$ is rejected using Scheffé's procedure.

Several alternative strategies for multiple testing were proposed on the basis of **Studentized range** statistics. If $n_i = n, i = 1, \ldots, I$, then the Studentized range distribution is the probability distribution for the **range** $(\max\{\overline{Y}_i\} - \min\{\overline{Y}_i\})$ of the I independent sample means all from a standard normal distribution, divided by the pooled estimate of the standard error of a sample mean (i.e. the square root of a **chi-square distributed** random variable scaled by a factor $[I(n-1)]^{-1}$). This studentized range distribution is indexed by $[I, I(n-1)]$ and we let $q^{\alpha}_{I, I(n-1)}$ denote the corresponding upper $100\alpha\%$ point. The percentage points of this Studentized range distribution are provided in Harter [22] and many texts, and may be used to carry out tests or construct simultaneous confidence intervals for differences in means. Tukey [52] indicated that one could carry out $I(I-1)/2$ pairwise tests by comparing $z_{ii'} = n^{1/2}(\overline{y}_{i\cdot} - \overline{y}_{i'\cdot})/s$ to the critical value $q^{\alpha}_{I, I(n-1)}$; that is, if $|z_{ii'}| > q^{\alpha}_{I, I(n-1)}$, then the means μ_i and $\mu_{i'}$ can be inferred to be different with FWE controlled at α, $i \neq i' = 1, \ldots, I$.

So-called multiple range tests have also been developed by Newman [39], Keuls [34], and Duncan [9, 10]; the former two authors independently proposed the same testing procedure, which is often referred to as the Newman–Keuls test. Paraphrasing Miller [38], multiple range tests tend to declare two means in a set of I means significantly different provided the range of each and every subset containing the two means is significant according to an α_g level studentized range test, where g is the number of means in the subset at hand. The Neuman–Keuls test and Duncan's test differ in the way in which α_g is determined. For the Neuman–Keuls test, $\alpha_g = \alpha$ for $g = 2, 3, \ldots, I$, whereas $\alpha_g = 1 - (1-\alpha)^{g-1}$ in Duncan's test. Miller [38] provides a good illustration of the application of these two multiple range tests and argues that, while Duncan's test leads to larger α_g for larger group sizes (and hence greater power for detecting treatment effects), this is at the

expense of increasing the rate of false positive conclusions arising from multiple tests. Miller therefore favors the Neuman–Keuls approach over Duncan's. However, it is important to note that, in their original form, neither Duncan's nor the Newman–Keuls procedure controls the FWE; various modifications of these procedures have been proposed but are beyond the scope of this review (see [50] for details).

Generalizations to facilitate applications to the unbalanced one-way lay-out have been proposed by several authors [10, 11, 35, 52]. Dunnett [12] conducted a detailed simulation study designed to investigate the empirical type I error rates of various procedures for this context. He found that the preferred method for pairwise comparisons is the Tukey–Kramer procedure, which compares the statistic $(\overline{y}_{i\cdot} - \overline{y}_{i'\cdot})/[s(1/n_{i\cdot} + 1/n_{i'\cdot})^{1/2}]$, with $q^{\alpha}_{I,I(n-1)}/\sqrt{2}$. This was found to be slightly conservative, but less so than other procedures developed for this same context [21, 24, 49]. A proof of the conservativeness of the Tukey–Kramer method for the one-way model was obtained by Hayter [23].

Generalizations to the studentized augmented range distributions may be utilized if it is desired to test not only for the equality of all means, but whether all means share a specific value. Hence, any pair of means for which the simultaneous confidence intervals does not contain the null value of zero, say, may be declared to be significantly different. With a simultaneous coverage probability of $1 - \alpha$, the FWE of this test is controlled at α.

More Recent Developments in Multiple Comparisons

Single-Step Procedures Using P Values

The Bonferroni procedure which rejects any H_{k0} whose P value is $\leq \alpha/K$ is perhaps the most widely known single-step testing procedure. It is attractive in its generality, but improvements have been made to generate procedures that are less conservative. Šidák's inequality leads to the less conservative critical value corresponding to $1 - (1 - \alpha)^{1/K}$ instead of α/K [47]. The gain in power from this approach may be quite minimal, however, for cases when $K \leq 10$.

Simes [48] presents a multiple testing procedure applicable when $H_0 = \bigcap_{k=1}^{K} H_{k0}$ and $H_a = \bigcup_{k=1}^{K} H_{ka}$. Let $P_{(1)} \geq P_{(2)} \geq \cdots \geq P_{(K)}$ be P values arising from the K component tests ordered from largest to smallest. Then H_0 is rejected if $P_{(k)} \leq (K - k + 1)\alpha/K$ for some k. This test has higher power than the Bonferroni procedure for testing H_0. The FWE is shown to be controlled at α by arguments pertaining to **order statistics** of independent **uniform** (0, 1) random variables and is valid under the assumption that the component discrepancy measures are independent [48]. Correlations among the discrepancy measures can lead to serious inflation of the FWE [28]. This is an example of a union–intersection test procedure.

A number of more computationally intensive approaches have also been proposed. Brown & Fears [5] focused on binomial data and unadjusted P values arising from unconditional or conditional analyses. A permutation distribution (with fixed marginal frequencies) was then used to compute the adjusted P value corresponding to the probability (based on the permutation distribution) of realizing a P value smaller than that observed (*see* **Randomization Tests**). Westfall [53] proposed instead that one resample with replacement from the observed data set and determine whether the minimum P value of the new data set is, or is not, less than or equal to that observed in the sample. The frequency with which it is less is the adjusted P value for the comparison of interest. An advantage of this approach, suggested by Westfall & Young [54], is that it effectively addresses the correlation of the test statistics, particularly for multivariate outcomes, or when the comparisons of interest are all against a single control arm.

Stepwise Procedures Using P Values

Holm [27] presents a step-down multiple testing procedure, which he refers to as a sequentially rejective Bonferroni test, based on the ordered P values (*see* **Multiple Endpoints, *P* Level Procedures**). Again, let $P_{(1)} \geq P_{(2)} \geq \cdots \geq P_{(K)}$ be the ordered P values. Denote by $H_0^{(k)}$ the null hypothesis corresponding to $p_{(k)}$ for $1 \leq k \leq K$. The procedure compares the ordered P values with the sequence $\alpha, \alpha/2, \ldots, \alpha/K$ starting with $P_{(K)}$, then $P_{(K-1)}$, etc., continuing as long as $P_{(k)} \leq \alpha/k$, in which case we reject $H_0^{(k)}$ and go to the next ordered P value down; the first time we find $P_{(k)} > \alpha/k$, we stop testing and accept (do not reject) the remaining hypotheses.

Holm [27] points out that since the thresholds used to assess the strength of evidence against the null hypotheses are larger for all but the minimum P value, $P_{(K)}$, this approach has a higher probability of rejecting false null hypotheses than the standard Bonferroni procedure. This approach is attractive in that, as with the standard Bonferroni procedure, it is widely applicable.

Step-up procedures have also been proposed as improvements to the standard Bonferroni approach. Under Hochberg's [25] step-up procedure, all K P values are ordered as before, and testing begins by comparing the largest P value, which is $P_{(1)}$, with α, then $P_{(2)}$ with $\alpha/2$, and so on, continuing as long as the corresponding hypothesis is not rejected. The first time a rejection occurs, testing stops and all remaining hypotheses are rejected as well. In other words, one does not reject until $P_{(k)} \leq \alpha/k$, at which point $H_0^{(k)}, \ldots, H_0^{(K)}$ are rejected, $k = 1, \ldots, K$. Note that this procedure employs the same critical values as Holm's step-down procedure and any hypothesis rejected by Holm's procedure is also rejected by Hochberg's procedure; hence the latter is at least as powerful.

In Hommel's [29, 30] step-up procedure, one searches for the largest $m (1 \leq m \leq K)$ such that

$$P_{(k)} > \frac{(m-k+1)\alpha}{m}, \quad \text{for } k = 1, \ldots, m.$$

If such an m exists, then any hypothesis that has a P value $\leq \alpha/m$ is rejected. If such an m does not exist, then all hypotheses are rejected. Hommel's procedure coincides with Hochberg's for its first two steps, but after that it may reject additional hypotheses to those rejected by Hochberg's procedure.

Both Hochberg's and Hommel's procedures were developed by applying the closure principle to the procedure of Simes [48]. Hommel's procedure has slightly greater power, but is more complicated to apply. Another procedure that has slightly greater power but is also more complicated to apply, was given by Rom [43].

Stepwise Procedures Using Normal Theory

In multitreatment trials, the individual hypotheses $H_k, k = 1, \ldots, K$, are usually formulated in terms of parameters θ_k which are contrasts in the population means. Estimates $\hat{\theta}_k$ of these are determined from the data. Under the linear model assumptions stated at the beginning of this article, it follows that the $\hat{\theta}_k$ are normally distributed with $E(\hat{\theta}_k) = \theta_k$, $\text{var}(\hat{\theta}_k) = \tau_k^2\sigma^2$, and $\text{corr}(\hat{\theta}_i, \hat{\theta}_j) = \rho_{ij}$, where the τ_k^2 and ρ_{ij} are known constants which depend upon the design (e.g. on the sample sizes of the treatment groups). To test the hypothesis H_k that θ_k has a specified value 0 (say), a test statistic $t_k = \hat{\theta}_k/(\tau_k^2 s^2)^{1/2}$, where s^2 is an estimate of σ^2, is used. Under the normality and homogeneous variance assumptions, the t_k are **Student *t* statistics** and the joint distribution of the corresponding random variables is **multivariate *t***. This provides the underlying distribution theory for testing the hypotheses.

In stepwise testing the test statistics are ordered according to their P values as previously. Denote them by $t_{(1)}, t_{(2)}, \ldots, t_{(K)}$, where $t_{(1)}$ is the least significant and $t_{(K)}$ the most significant test statistic. These are compared in sequence with a set of critical constants $c_1 < \cdots < c_K$, determined so that the FWE is $\leq \alpha$.

In step-down testing, we start with $t_{(K)}$, then go to $t_{(K-1)}$, and so on. We continue to the next test in the sequence whenever we find $t_{(k)} \geq c_k$ and reject the corresponding hypothesis, stopping the first time $t_{(k)} < c_k$ and accepting (not rejecting) any remaining hypotheses.

In step-up testing we start with $t_{(1)}$, then go to $t_{(2)}$, and so on. We continue to the next test in the sequence whenever we find $t_{(k)} < c_k$ and accept (do not reject) the corresponding hypothesis, stopping the first time $t_{(k)} \geq c_k$ and rejecting any remaining hypotheses.

We illustrate the determination of the critical constants for the above step-down and step-up testing procedures for the case where one of the I treatment groups, say the Ith group, is to be compared with each of the other groups. Then the contrasts of interest are $\theta_k = \mu_k - \mu_I$, for $k = 1, \ldots, I-1$. Suppose, for simplicity, that each group has the same sample size, n, except for the Ith group which has sample size n_0. Then let $Y_{ij}(y_{ij})$ denote the random (realized) response for the jth subject and $\overline{y}_i$ the sample mean in group $i, i = 1, \ldots, I$; s^2 denotes the pooled estimator of the common variance based on $\nu = (I-1)(n-1) + n_0 - 1$ degrees of freedom (df).

Denote the random variables corresponding to the ordered t statistics $t_{(1)}, \ldots, t_{(K)}$ by $T_1, \ldots, T_K$, respectively. Then the critical constants for the step-down case with two-sided alternative hypotheses are

obtained by solving the following equations:

$$\Pr(-c_k < T_1 < c_k, \ldots, -c_k < T_k < c_k) = 1 - \alpha,$$
$$k = 1, \ldots, K = I - 1.$$

$T_1, \ldots, T_k$ have a k-variate central t distribution with ν degrees of freedom and common $\rho = n/(n + n_0)$ under the null hypothesis $H_0 = \bigcap_{i=1}^{k} H_0^{(i)}$. Note that since the critical value, c_K, is the same as the one derived for the single-step procedure, the step-down method may be thought of as a natural extension of the single-step procedure.

Tables of the critical values are readily available for the case in which the sample sizes for all but the reference group are the same ($n_i = n, i = 1, \ldots, I - 1$), and hence the discrepancy measures are equally correlated (see [2, 26]). In the case of unequal group sizes, good approximations can be obtained to the critical values by using the average correlation and interpolating from published tables. Alternatively, however, with a known correlation matrix, the critical values may be found by direct multivariate, or recursive, **numerical integration** [14].

For the step-up case, the values of the constants $c_1, c_2 \ldots, c_k$ are determined by solving

$$\Pr(-c_1 < T_{(1)} < c_1, \ldots, -c_k < T_{(k)} < c_k) = 1 - \alpha,$$
$$k = 1, 2, \ldots, K,$$

where $T_{(1)}, \ldots, T_{(k)}$ are the ordered values of the random variables $T_1, T_2, \ldots, T_k$ associated with the first k t statistics in order of significance. Note that the solutions here must be obtained recursively, starting with $k = 1$, then $k = 2$, and so on, since in order to solve for any c_k it is necessary to know the values of $c_1, \ldots, c_{k-1}$. Also, for both step-down and step-up testing, the value of c_1 is the α point of univariate Student's t. For $k > 1$, the constant c_k for step-up testing is slightly larger than the corresponding c_k for step-down testing. The first step of the step-up testing procedure corresponds to the Laska & Meisner [36] MIN test, which tests the intersection–union problem $H_0 = \bigcup_{k=1}^{K} H_0^{(k)}$ vs. $H_a = \bigcap_{k=1}^{K} H_a^{(k)}$; hence, the step-up procedure may be thought of as a natural extension of this test.

There are limited tables of the step-up constants given in Dunnett & Tamhane [15, 16] for the case of equal correlations; the case of unequal correlations is considered in Dunnett & Tamhane [17]. The step-down testing procedure is the normal theory analog of Holm's P value procedure, and the step-up testing procedure is the normal theory analog of Hochberg's P value procedure. The advantage of the normal theory procedures is that they utilize the correlation structure of the parameter estimates and have higher power when the normality and homogeneous variance assumptions hold. However, the P value procedures do not depend on such assumptions and can be used when they do not hold.

Comparisons with the Best Treatment

Now consider the case in which there is no specific reference group of interest and let $\mu_{(1)} \leq \cdots \leq \mu_{(I)}$ denote the I ordered population means, where we assume that larger values of μ_i correspond to preferred treatments. Since the means themselves are unknown, so too is the appropriate ordering given above. Nevertheless, one can conduct inference on the quantities $\mu_{(I)} - \mu_i$, the difference in the mean response for the ith treatment group from that of the unknown "best" treatment. Hsu [31] derives a method of constructing simultaneous joint one-sided upper confidence intervals for the $\mu_{(I)} - \mu_i$, $i = 1, \ldots, I$, and Hsu [32] extends these methods for two-sided intervals. For simplicity we focus on the case of common interest, namely where σ is unknown and upper bounds on $\mu_{(I)} - \mu_i$ are of main interest. Hsu [31] shows that if $U_1, \ldots, U_I$ are I independent and identically distributed standard normal random variables, $n_i = n$, $i = 1, \ldots, I$, $v = I(n - 1)$, and we let $d_{I,v}^{\alpha}$ be the constant such that

$$\Pr(U_I > U_i - d_{I,v}^{\alpha} s, i = 1, \ldots, I - 1) = 1 - \alpha,$$

then a set of $100(1 - \alpha)\%$ simultaneous confidence intervals for $\mu_{(I)} - \mu_i$ are given by $[0, D_i], i = 1, \ldots, I$, where

$$D_i = \max\left[\max_{j \neq i}(\overline{X}_j) - \overline{X}_i + d_{I,v}^{\alpha} s/\sqrt{n}, 0\right].$$

The constant $d_{I,v}^{\alpha}$ is the solution to

$$\int_0^{\infty} \int_{-\infty}^{\infty} \Phi^I(u + d_{I,v}^{\alpha} s)\, \mathrm{d}\Phi(u)\, \mathrm{d}\Psi_v(s) = 1 - \alpha,$$

where $\Phi(\cdot)$ and $\Psi_v(\cdot)$ are the distribution functions for a standard normal random variable and a $(\chi_v^2/v)^{1/2}$ random variable, respectively. Tables for the constant d are identical, except for a constant

$\sqrt{2}$, with those for the constants used in the normal theory step-down method described previously. For further information, see Hsu [33].

Medical and Biometric Applications

Comparisons Between Several Treatments and a Control

The problem is to compare K test treatments with a control treatment, which may be either a placebo or a standard treatment. Denote the unknown mean responses by $\mu_1, \ldots, \mu_{K+1}$, where $\mu_k, k = 1, \ldots, K$, denotes the mean for the kth test treatment and μ_{K+1} denotes the control mean. We formulate a multiple hypotheses testing problem where we test $H_{0k} : \mu_k = \mu_{K+1}$ vs. $H_{ak} : \mu_k \neq \mu_{K+1}$, for $k = 1, \ldots, K$. Rejection of H_{0k} in favor of H_{ak} leads us to conclude there is a difference between the kth treatment and the control.

The purpose of the trial may be to select the best candidate and to test the hypothesis pertaining to that particular candidate. Since there are K possible choices, we stipulate that the FWE be $\leq \alpha$ to ensure that the probability of declaring a false positive result is at most α. Since only one treatment is to be chosen, we may use a single-step procedure. If we reject this hypothesis, and wish to perform additional hypothesis tests to determine whether other candidates also differ significantly from the control, then we use the step-down procedure as described earlier. (Note that since we are only interested in finding a test treatment better than the control, we may prefer to formulate the hypotheses testing problem with one-sided alternatives instead of two-sided alternatives (*see* **Alternative Hypothesis**) in order to increase power. However, this decision must be made at the design stage.)

If the problem is to find all test treatments that can be shown to differ from the control, rather than selecting only one, then the problem may require a different formulation. Suppose the experiment is one of a series of similar experiments in which potential new treatments are compared with a control, and any test treatment showing promise is selected for further study. This is called *screening* (*see* **Animal Screening Systems**). In this case, interest would be in controlling the PCE rather than the FWE. Here the decision on each treatment does not depend on the decisions made with respect to the other treatments included in the experiment, which are included in the same experiment for reasons of experimental efficiency.

Comparisons Between a New Treatment and Several Standards

The test treatment may be a potential new treatment being compared with K standard treatments to determine whether it meets requirements for approval by the regulatory authority. The same hypotheses as before may be formulated, with μ_{K+1} now denoting the mean for the test treatment and μ_k the mean for the kth standard. Rejection of H_{0k} means that we conclude that the test treatment is different from the kth standard. To control the risk of any false claim, i.e. a claim that the test treatment differs from a particular standard when it does not, we would adopt a procedure that controls the FWE. Since the experimenter would like to find as many differences from the K standards as possible, one of the stepwise procedures, either step-down or step-up, should be used.

Superiority/Equivalence of a New Treatment Compared with K Standards

In the family of hypotheses used in the previous example, the alternative hypotheses, H_{ak}, were two-sided: rejection of H_{0k}, meant that we concluded there is a difference between the test treatment and the kth standard, but it could be either better or worse depending on the direction of the difference. Such a formulation is equivalent to testing a pair of hypotheses with one-sided alternatives, namely H_{0k} vs. $H^1_{ak} : \mu_{K+1} > \mu_k$ (test treatment is better than kth standard), and H_{0k} vs. $H^2_{ak} : \mu_{K+1} < \mu_k$ (test treatment is worse than kth standard).

Here we describe an alternative formulation, proposed by Dunnett & Tamhane [18]. We replace the hypothesis in each pair for determining whether the test treatment is worse with one which tests whether the test treatment is equivalent to the kth standard, namely

$$H'_{0k} : \mu_{K+1} = \mu_k - \delta \text{ vs. } H'_{ak} : \mu_{K+1} > \mu_k - \delta.$$

(*see* **Bioequivalence**; **Equivalence Trials**). Here, $\delta > 0$ is a prespecified value representing a difference that is clinically unimportant. Rejection of H'_{0k} leads

us to conclude that μ_{K+1} cannot be less than μ_k by more than an amount δ and hence, by definition, it is equivalent to the kth standard.

Dunnett & Tamhane [18] show how the stepwise multiple testing procedures described above can be adapted to this multiple testing problem. By using this superiority/equivalence testing formulation, we increase the power over the two-sided formulation, but of course this is addressing a slightly different problem. By testing simultaneously for superiority and equivalence, we also obtain more information than we would from a formulation that tests only for superiority using one-sided alternate hypotheses.

Comparisons with Both Active and Placebo Controls

D'Agostino & Heeren [8] described a trial where comparisons between a new treatment, T, two known active treatments, A_1 and A_2, and a placebo, P, are of interest. One proposal made was that the pairwise differences between treatment groups be tested with the *experimentwise error rate* controlled, which means that all comparisons are handled as a single family. Dunnett & Tamhane [16] pointed out that there were actually three families of comparisons, each answering a different question: (i) A_1 and A_2 vs. P to test the sensitivity of the experiment, defined as its ability to identify that the two known active treatments are efficacious; (ii) T vs. P to show that the new treatment is better than the placebo; and (iii) T vs. A_1 and A_2 to determine whether the new treatment can be shown to be superior to either of the known active treatments. Each family should be tested with FWE controlled at $\leq \alpha$. Failure to show the sensitivity of the experiment, or failure to find the new treatment better than placebo, would invalidate the comparisons made in (iii). In this case, controlling the FWE at α for each of the three families individually serves to control the overall error rate for the combined families at α as well, so there is no need to apply an experimentwise multiplicity adjustment. This is an example of what is known as *a priori ordered* families of hypotheses [1].

A single hypothesis test serves to cover (i), namely

$$\mathrm{H}_{01} : \mu_T - \mu_P \leq 0 \text{ vs. } \mathrm{H}_{\mathrm{a}1} : \mu_T - \mu_P > 0.$$

The following pair of hypotheses tests covers the comparisons in (ii):

$$\mathrm{H}_{02} : \mu_{A_1} - \mu_P \leq 0 \text{ vs. } \mathrm{H}_{\mathrm{a}2} : \mu_T - \mu_P > 0,$$

$$\mathrm{H}_{03} : \mu_{A_2} - \mu_P \leq 0 \text{ vs. } \mathrm{H}_{\mathrm{a}3} : \mu_{A_2} - \mu_P > 0,$$

while the following pair of hypotheses tests covers the comparisons in (iii):

$$\mathrm{H}_{04} : \mu_T - \mu_{A_1} \leq 0 \text{ vs. } \mathrm{H}_{\mathrm{a}4} : \mu_T - \mu_{A_1} > 0,$$

$$\mathrm{H}_{05} : \mu_T - \mu_{A_2} \leq 0 \text{ vs. } \mathrm{H}_{\mathrm{a}5} : \mu_{A_2} - \mu_{A_2} > 0.$$

To test H_{01}, we use an ordinary Student t test at level α, since there is only one hypothesis in the family. To test H_{02} and H_{03}, since we require both to be rejected to establish sensitivity, we use the MIN test of Laska & Meisner [36] or its extension, the step-up test described earlier in the article. To test H_{04} and H_{05}, we use the step-down test if we expect at most one of the two hypotheses to be rejected, or the step-up test if we expect both to be rejected.

Comparisons in a Dose Finding Experiment

In a dose finding experiment several dose levels of a compound along with a zero dose control are studied with respect to a specified response, usually some measure of efficacy or toxicity. The goal is to determine the lowest dose that produces a response that exceeds the control response (or, more generally, exceeds it by more than a specified amount, δ), denoted as the minimum effective dose (MED) [45] (*see* **Minimum Therapeutically Effective Dose**). Say there are K dose levels (usually, $K = 3$ or 4), and denote the mean responses for the control and the K dose levels by $\mu_0, \mu_1, \ldots, \mu_K$. Then we define

$$\mathrm{MED} = \min(k : \mu_k > \mu_0).$$

Ruberg [45] (see also Tamhane et al. [51]) formulates the problem of identifying the MED as the following multiple hypotheses testing problem:

$$\mathrm{H}_{0k} : \mu_0 = \mu_1 = \cdots = \mu_k \text{ vs.}$$

$$\mathrm{H}_{\mathrm{a}k} : \mu_0 = \mu_1 = \cdots = \mu_{k-1} < \mu_k,$$

for $k = 1, \ldots, K$. The estimated MED, or minimum detectable dose (MDD), is the lowest index k for which H_{0k} is rejected. Strong control of the FWE is needed in testing this family of hypotheses in order to

control the probability of obtaining an estimate that is less than the true MED. A class of test statistics which may be used are based on contrasts in the observed means. Various stepwise tests are given in Tamhane et al. [51] and compared in a simulation study with respect to their FWE and power under various forms of the dose response relationship.

General Remarks

The literature on multiple comparisons is voluminous and it is not possible to cover adequately all aspects and developments in an encyclopedia entry such as this. Two particular topics which are related, and warrant further mention, are selection and order restrictions. Selection problems have the general objective of identifying a favorable treatment or treatments from a collection of treatments. As might be expected, there are close links with multiple comparison problems and these connections are widely recognized (see [32] for example). Order restrictions in the hypotheses arise when there is added structure to the problem such as in dose-ranging studies when it is "known" that larger doses will be associated with nondecreasing means. Methods for the testing of order restricted hypotheses involve introducing constraints in the likelihood functions and so have links with **isotonic regression** (*see* **Isotonic Inference**).

It should be noted that most of the discussion thus far has assumed that two-sided tests are of interest. The methods described all apply for the case of one-sided tests following minimal modifications. In addition we have emphasized hypothesis testing throughout. Inferences regarding interval estimates are often equivalently possible, with the focus on the simultaneous coverage probability of collections of intervals, rather than FWE.

References

[1] Bauer, P. (1991). Multiple testing in clinical trials, *Statistics in Medicine* **10**, 871–890.

[2] Bechhofer, R.E. & Dunnett, C.W. (1988). Tables of percentage points of multivariate t distributions, *Selected Tables in Mathematical Statistics*, **11**. American Mathematical Society, Providence, pp. 1–371.

[3] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

[4] Berger, R.L. (1982). Multiparameter hypothesis testing and acceptance sampling, *Technometrics* **24**, 295–300.

[5] Brown, C.C. & Fears, R.R. (1981). Exact significance levels for multiple binomial testing with application to carcinogenicity screens, *Biometrics* **37**, 763–774.

[6] Cook, R.J. & Farewell, V.T. (1996). Multiplicity considerations in the design and analysis of clinical trials, *Journal of the Royal Statistical Society, Series A* **159**, 93–110.

[7] Cox, D.R. (1965). A remark on multiple comparison methods. *Technometrics* **7**, 223–224.

[8] D'Agostino, R.B. & Heeren, T.C. (1991). Multiple comparisons in over-the-counter drug clinical trials with both positive and placebo controls (with comments), *Statistics in Medicine* **10**, 1–31.

[9] Duncan, D.B. (1951). A significance test for differences between ranked treatments in an analysis of variance, *Virginia Journal of Science* **2**, 171–189.

[10] Duncan, D.B. (1955). Multiple range and multiple F tests, *Biometrics* **11**, 1–42.

[11] Duncan, D.B. (1957). Multiple range tests for correlated and heteroscedastic means, *Biometrics* **13**, 164–176.

[12] Dunnett, C.W. (1980). Pairwise multiple comparisons in the homogeneous variance, unequal sample size case, *Journal of the American Statistical Association* **75**, 789–795.

[13] Dunnett, C.W. (1997). Comparisons with a control, in *Encyclopedia of Statistical Sciences*, Update Vol. 1, S. Kotz, B.C. Read & D.L. Banks, eds. Wiley, New York.

[14] Dunnett, C.W. & Tamhane, A.C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts, *Statistics in Medicine* **10**, 939–947.

[15] Dunnett, C.W. & Tamhane, A.C. (1992). A step-up multiple test procedure, *Journal of the American Statistical Association* **87**, 162–170.

[16] Dunnett, C.W. & Tamhane, A.C. (1992). Comparisons between a new drug and active and placebo controls in an efficacy clinical trial, *Statistics in Medicine* **11**, 1057–1063.

[17] Dunnett, C.W. & Tamhane, A.C. (1995). Step-up multiple testing of parameters with unequally correlated estimates, *Biometrics* **51**, 217–227.

[18] Dunnett, C.W. & Tamhane, A.C. (1997). Multiple testing to establish superiority/equivalence of a new treatment compared with k standard treatments, *Statistics in Medicine* **16**, 2489–2506.

[19] Finney, D.J. (1978). Multiple assays. in *Statistical Methods in Biological Assay*, 3rd Ed. Griffin, London, Chapter 11.

[20] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[21] Genizi, A. & Hochberg, Y. (1978). On improved extensions of the T-method of multiple comparisons for unbalanced designs, *Journal of the American Statistical Association* **73**, 879–884.

[22] Harter, H.L. (1960). Tables of range and Studentized range., *Annals of Mathematical Statistics* **31**, 1122–1147.

[23] Hayter, A.J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative, *Annals of Statistics* **12**, 61–75.

[24] Hochberg, Y. (1974). Some generalizations of the T-method in simultaneous inference, *Journal of Multivariate Analysis* **4**, 224–234.

[25] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

[26] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[27] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.

[28] Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures, *Metrika* **33**, 321–336.

[29] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test, *Biometrika* **75**, 383–386.

[30] Hommel, G. (1989). A comparison of two modified Bonferroni procedures, *Biometrika* **76**, 624–625.

[31] Hsu, J.C. (1981). Simultaneous confidence intervals for all distances from the "best", *Annals of Statistics* **9**, 1026–1034.

[32] Hsu, J.C. (1984). Ranking and selection and multiple comparisons with the best, in *Design of Experiments: Ranking and Selection (Essays in Honor of Robert E. Bechhofer)*, T.J. Santner & A.C. Tamhane, eds. Marcel Dekker, New York.

[33] Hsu, J.C. (1996). *Multiple Comparisons*, Chapman & Hall, London.

[34] Keuls, M. (1952). The use of the "Studentized range" in connection with an analysis of variance, *Euphytica* **1**, 112–122.

[35] Kramer, C.Y. (1956). Extension of multiple range tests to group means with unequal number of replications, *Biometrics* **12**, 307–310.

[36] Laska, E.M. & Meisner, M.J. (1989). Testing whether an identified treatment is best, *Biometrics* **45**, 1139–1151.

[37] Marcus, R., Peritz, E. & Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analyses of variance, *Biometrika* **63**, 655–660.

[38] Miller, R.G., Jr (1981). *Simultaneous Statistical Inference*. 2nd Ed. Springer-Verlag, New York.

[39] Newman, D. (1939). The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation., *Biometrika* **31**, 20–30.

[40] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika* **20**, 175–240.

[41] Pocock, S.J., Geller, N.L. & Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics* **43**, 487–498.

[42] Redman, C.E. & Dunnett C.W. (1994). Screening compounds for clinically active drugs, in *Statistics in the Pharmaceutical Industry*, 2nd Ed. Marcel Dekker, New York, Chapter 24.

[43] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality, *Biometrika* **77**, 663–665.

[44] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[45] Ruberg, S.J. (1989). Contrasts for identifying the minimum effective dose, *Journal of the American Statistical Association* **84**, 816–822.

[46] Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika* **40**, 87–104.

[47] Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions, *Journal of the American Statistical Association* **62**, 626–633.

[48] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **73**, 751–754.

[49] Spjøtvoll, E. & Stoline, M.R. (1973). An extension of the T-method of multiple comparison to include the cases with unequal sample sizes *Journal of the American Statistical Association* **68**, 975–978.

[50] Tamhane, A.C. (1996). Multiple Comparisons. in *Handbook of Statistics*, Vol. 13, S. Ghosh & C.R. Rao eds. pp. 587–630.

[51] Tamhane, A.C., Hochberg, Y. & Dunnett, C.W. (1996). Multiple test procedures for dose finding, *Biometrics* **52**, 21–37.

[52] Tukey, J.W. (1953). The Problem of Multiple Comparisons, Mimeographed Notes, Princeton University, Reprinted in *The Collected Works of John W. Tukey, Vol. VIII – Multiple Comparisons: 1948–1983*, H.I. Braun, ed. Chapman & Hall, New York, 1994.

[53] Westfall, P. (1985). Simultaneous small-sample multivariate Bernoulli confidence intervals, *Biometrics* **41**, 1001–1013.

[54] Westfall, P.J. & Young, S.S. (1989). p value adjustments for multiple tests in multivariate binomial models, *Journal of the American Statistical Association* **84**, 780–786.

RICHARD J. COOK & C.W. DUNNETT

Multiple Endpoints in Clinical Trials

Many clinical trials and observational studies collect data on multiple endpoints. An example of this problem occurs in prevention trials where data are collected on the primary disease of interest in addition to diseases commonly observed and events that may occur as a result of the intervention. In this setting it is important to evaluate all the outcomes of interest, since an intervention that results in a decrease in the incidence of the targeted disease at the expense of an unacceptable increase in a detrimental outcome would require evaluation of the potential for future use of the intervention (*see* **Benefit/Risk Assessment in Prevention Trials**). Other examples of multiple outcomes data arise in family studies, clinical trials where treatment is applied to the left eye while the right eye serves as a control, and clinical trials where the side-effects of treatments result in outcomes that must also be examined.

The multiple endpoints problem occurs in many guises and includes recurrent observations, multiple endpoints, and correlated outcome data. These techniques can also be applied to data where there is clustering due to the study design (*see* **Cluster Randomization**). Additionally, several of these features may be present in a single data set, such as the multiple outcome setting where the outcomes of interest include recurrent events. There has been considerable work in this area over the past 20 years including a series of papers related to the workshop Statistical Methods for Multiple Events Data in Clinical Trials [9]. There are now methods available that can be readily implemented using existing software packages and potentially interesting approaches that require further development.

Several approaches exist for the analysis of recurrent event data. The earliest methods [7, 17] are based on modeling inter-event times and can be readily implemented using any standard statistical package. In their book, Andersen et al. [3] discuss methods for modeling recurrent events using general intensity models. This work arose naturally out of earlier work by Andersen & Gill [2] who proposed the modeling of recurrent event data by extending the Cox model. Wei, Lin & Weissfeld (WLW) [21] discuss the modeling of recurrent event data based on modeling the recurrences as marginal distributions. Prentice et al. [17] propose models for recurrent events based on modeling the data conditionally at each event point using the Cox regression model. They propose two different techniques based on two different definitions for the baseline hazard function: the first definition using total time and the second definition using the gap times between events. Lawless & Nadeau [10] discuss the modeling of recurrent events based on event counts and this work is extended by Cook & Lawless [5] to include a terminating event, in addition to the recurrent event. Therneau & Hamilton [19] compare the Andersen–Gill and WLW approaches to the modeling of recurrent data, in several different examples. They conclude that the Andersen–Gill method coupled with a robust estimator of the variance performs well in settings where the goal is a test of overall treatment effect. The WLW approach suffers from problems in the recurrent event setting due to its sensitivity to departures from proportionality in the margins. These models are also compared in Lin [11] who discusses approaches for assessing the fit of the proportional hazards model at each recurrence. Both the Andersen–Gill and WLW approaches have the advantage of ease of implementation using standard statistical software packages.

Other work in the area of recurrent events, based on parametric approaches for multitype events, is discussed in Abu-Libdeh et al. [1] and Fang et al. [6]. Abu-Libdeh et al. [1] use a mixed Poisson process with a random and fixed effect to model the data as a replicated multitype point process. Fang et al. [6] proposed a model based on a multiple renewal process allowing for time-dependent covariates. The estimators are obtained using standard maximum likelihood techniques.

For the analysis of multiple outcome data that include outcomes of several types, there are fewer available methods, with many methods applicable for the bivariate setting only. The method typically used in this setting is the method of Wei et al. [21] which is based on modeling each outcome as a separate margin assuming independence. The correlation between parameter estimates is obtained through a "sandwich" estimator. Lin & Wei [12] extended this method to incorporate linear regression methods for censored data. Prentice & Cai [16], developed estimators for the covariance and survival function for multivariate censored data with extensions to the

regression setting. Additional methods for estimating the correlation between two failure times have been developed via the copula model. This approach is discussed in Shih & Louis [18] and further studied in Phelps & Weissfeld [15]. More recently, Wang & Wells [20] have developed model selection and inference procedures for bivariate survival models based on the Archimedean copula. The introduction of random effects through a latent frailty variable has also been proposed for the analysis of bivariate survival data [8, 14]. The WLW method can be implemented in several standard statistical software packages, while the extension of Lin & Wei [12] and the copula approach require specialized software.

Methods

We first discuss models and notation for the multiple outcome approach. Assume that there are K types of failure and let T_{ki} denote the failure time for the ith individual and the kth type of failure where $i = 1, \ldots, n$ and $k = 1, \ldots, K$. Let C_{ki} denote the corresponding censoring time. The observed data are of the form $(Y_{ki},\ \Delta_{ki},\ \mathbf{X}_{ki})$, where $Y_{ki} = \min(T_{ki}, C_{ki})$, $\Delta_{ki} = 1$ if $T_{ki} = Y_{ki}$ and 0 otherwise, and $\mathbf{X}_{ki} = (X_{1ki}, \ldots, X_{pki})$ denote the p covariates for the kth event type on the ith individual. Two further assumptions are also made. The first is the assumption of independent censoring; that is, conditional on $\mathbf{X}_I, = (\mathbf{X}_{1I}, \ldots, \mathbf{X}_{Ki})$, the vectors T_I of the K failure times and the vectors C_I of the K censoring times are assumed to be independent. Additionally, it is assumed that $(\mathbf{T}_I, \mathbf{C}_I, \mathbf{X}_I)$, $I = 1, \ldots, n$, are independent and identically distributed random quantities.

For the kth type of failure on the ith subject, the hazard function $\lambda_{ki}(t)$ is assumed to be of the form

$$\lambda_{ki}(t) = \lambda_{k0}(t) \exp\{\beta_k' X_{ki}(t)\}, \quad t \geq 0, \qquad (1)$$

where $\lambda_{k0}(t)$ is an unspecified baseline hazard function corresponding to the kth failure type and $\beta_k = (\beta_{1k}, \ldots, \beta_{pk})'$ is the regression parameter corresponding to the kth failure type. To define the partial likelihood function corresponding to the kth type of failure, let $R_k(t) = \{l : X_{kl} \geq t\}$. Note that this is the risk set corresponding to all individuals who have not experienced the kth failure as of time t. Then the partial likelihood function corresponding to the kth failure type is given by

$$L_k(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta' X_{ki}(Z_{ki})\}}{\sum_{l \in R_k(Z_{ki})} \exp\{\beta' X_{ki}(Z_{ki})\}} \right]^{\Delta_{ki}}. \qquad (2)$$

The maximum partial likelihood estimator, $\hat{\beta}_k$, for β_k is the solution to the standard likelihood equation $\partial \log L_k(\beta)/\partial\beta = 0$. The estimator $\hat{\beta}_k$ will be consistent for β_k if the model is correctly specified. Note that in the application of the WLW approach to recurrent event data, care needs to be taken to ensure that the model for any particular recurrence is correctly specified. The marginal models are easily fit using any standard statistical package that fits a Cox regression model. WLW show that $(\hat{\beta}_1', \ldots, \hat{\beta}_K')'$ is asymptotically normal with mean $(\beta_1', \ldots, \beta_K')'$ and covariance matrix Q which is estimated via the sandwich estimator given in the Appendix to the paper. Using these estimators, simultaneous inferences on the β_ks can be carried out, and an "average effect" of the covariates can be estimated. Software for fitting this model is available through S-PLUS and an SAS macro.

A second approach to the analysis of multiple outcome data is based on the use of a copula. The Archimedean copula is defined as

$$C_\alpha(x_1, \ldots, x_K) = \phi_\alpha[\phi_\alpha^{-1}(x_1) + \cdots + \phi_\alpha^{-1}(x_K)], \quad 0 \leq x_1, \ldots, x_K \leq 1, \qquad (3)$$

where $\phi_\alpha : [0, 1], \forall [0, 4], \phi_\alpha(1) = 0, \phi_\alpha'(x) < 0$, and $\phi_\alpha''(x) > 0$. Let $T_k, k = 1, \ldots, K$, denote distinct failure times or types and let $S_k(T_k)$ denote the continuous survival function of T_k. Then the multivariate survival function can be represented as

$$\begin{aligned} C_\alpha(S_1(T_1), \ldots, S_K(T_K)) &= \phi_\alpha[\phi_\alpha^{-1}(S_1(T_1)) \\ &\quad + \cdots + \phi_\alpha^{-1}(S_K(T_K))] \\ &= S(t_1, \ldots, t_K), \\ &\quad \text{for } t_1, \ldots, t_K \geq 0. \end{aligned} \qquad (4)$$

For the purposes of modeling, several forms of $\phi_\alpha(s)$ are useful. These include the Laplace transform,

$$\phi_\alpha(s) = \exp(-s^\alpha), \qquad (5)$$

which was studied by Hougaard [8] and the Clayton–Oakes [4, 13] gamma frailty model where

$$\phi_\alpha(s) = (1 + s)^{1/(1-\alpha)}. \qquad (6)$$

While both of these models are defined via copulas, they are frailty models where α denotes the parameter of the underlying frailty. Covariates are introduced through the parameterization of the survival function and α can be estimated from the full likelihood or using a two-step approach [8, 18]. While these methods are potentially quite useful, they are still in the early stages of development with no software widely available for their implementation.

For the analysis of recurrent event data, the WLW method can be employed with k indexing the kth recurrence rather than the kth type of failure. Using this approach, all the data are used for each recurrence since an individual with two recurrences becomes a censored observation for $k > 2$. The Andersen–Gill model can be implemented by redefining the at-risk indicators for the standard partial likelihood equation. Prentice et al. [17] propose the use of a conditional model based on two different definitions of the hazard function. These definitions are based on using the time from the beginning of the study, denoted as t, or the time from the immediately preceding failure, denoted as $t - t_{n(t)}$. Thus, for this model, the hazard function is given by

$$\lambda_i(t) = \lambda_{k0}(t) \exp\{\beta_k' X_{ki}(t)\} \tag{7}$$

or

$$\lambda_i(t) = \lambda_{k0}(t - t_{n(t)}) \exp\{\beta_k' X_{ki}(Z_{ki})\}, \tag{8}$$

where the $\lambda_{k0}(\cdot)$, $k = 1, \ldots, K$, are completely arbitrary baseline hazard functions. In this case k denotes a stratification variable that may change over time for a given subject, β_k is a column vector of stratum-specific regression coefficients and $t_{n(t)}$ denotes the time of the immediately preceding recurrence. Note that these are very flexible models that can be fit using standard statistical software. The drawback of these models is that they are conditional, so that estimation of the sth recurrence is based only on the individual with $(s - 1)$ recurrences. Inferential methods are based on standard methods for the Cox proportional hazards model. All these approaches can be readily implemented using standard statistical software packages such as S-PLUS or SAS.

References

[1] Abu-Libdeh, H., Turnbull, B.W. & Clark, L.C. (1990). Analysis of multi-type recurrent events in longitudinal studies: application to a skin cancer prevention trial, *Biometrics* **46**, 1017–1034.

[2] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[3] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Methods Based on Counting Processes*. Springer-Verlag, New York.

[4] Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidences, *Biometrika* **65**, 141–151.

[5] Cook, R.J. & Lawless, J.F. (1997). Marginal analysis of recurrent events and a terminating event, *Statistics in Medicine* **16**, 911–924.

[6] Fang, J., Shi, Z., Zhang, X., Zeng, D. & Zhang, J. (1990). Parametric inference in a multiple renewal process with time dependent covariates, *Biometrics* **46**, 849–854.

[7] Gail, M.H., Santner, T.J. & Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor, *Biometrics* **36**, 255–266.

[8] Hougaard, P. (1986). A class of multivariate failure time distributions, *Journal of the American Statistical Association* **73**, 671–678.

[9] Lagakos, S. (1997). Statistical methods for multiple events data in clinical trials, *Statistics in Medicine* **16**(8).

[10] Lawless, J.F. & Nadeau, J.C. (1995). Some simple robust methods for the analysis of recurrent events, *Technometrics* **37**, 158–168.

[11] Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach, *Statistics in Medicine* **13**, 2233–2247.

[12] Lin, J.S. & Wei, L.J. (1992). Linear regression analysis for multivariate failure time observations, *Journal of the American Statistical Association* **87**, 1091–1097.

[13] Oakes, D. (1982). A model for association in bivariate survival data, *Journal of the Royal Statistical Society, Series B* **44**, 414–422.

[14] Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**, 487–493.

[15] Phelps, A.L. & Weissfeld, L.A. (1997). A comparison of dependence estimators in bivariate copula models, *Communications in Statistics-Simulation and Computation* **26**, 1583–1597.

[16] Prentice, R.L. & Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika* **79**, 495–512.

[17] Prentice, R.L., Willams, B.J. & Peterson, A.V. (1981). On the regression analysis of multivariate failure time data, *Biometrika* **68**, 373–379.

[18] Shih, J.H. & Louis, T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data, *Biometrics* **51**, 1384–1399.

[19] Therneau, T.M. & Hamilton, S.A. (1997). rhDNase as an example of recurrent event analysis, *Statistics in Medicine* **16**, 1019–2047.

[20] Wang, W. & Wells, M.T. (2000). Model selection and semiparametric inference for bivariate failure-time

data, *Journal of the American Statistical Association* **95**, 62–72.

[21] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

L.A. WEISSFELD

Multiple Endpoints, Multivariate Global Tests

The term "multiple endpoints" is often related to an investigation of a treatment effect in the setting of a **clinical trial** or biomedical research. Since there are different aspects to characterize a treatment effect, it often requires more than one **outcome measure** (or **response variable**) to characterize the efficacy of a treatment. These variables can be changes in symptoms, bodily functions, subjective assessment, or any events or phenomena that are essential to evaluate the treatment effect. We consider here the situation in which these variables are of primary importance to the process of deciding the efficacy of a treatment, and we refer here to these primary variables as *endpoints*. The following are some examples of studies that require multiple endpoints. An investigator desires to examine the effect of different dosages of an antihistamine in treating the common cold. To obtain a more complete and detailed description of the efficacy of the doses of the antihistamine, he or she measures the level of relief in various important symptoms, including runny nose, sneezing, headache, and sinus discomfort. Another example arises in a study evaluating a heartburn relief product, where a drug company compares its new product with a standard control in the relief of the heartburn episodes over a two-week period. The primary endpoints include the relief of the first episode within 30 minutes of onset of symptoms, the proportion of relief of heartburn episodes over the two weeks where relief is within 30 minutes, and the global assessment of the product at the end of the two-week trial. The process of selecting the primary endpoints is no easy task. Careful assessment, mechanism of action, clinical evidence and, sometimes, subjective judgment are required in the process. However, it is not the purpose of this article to discuss how to select the primary endpoints. Here, we assume that this step has been done and we want to discuss the different methods that can be used to analyze such multivariate data.

A variety of statistical procedures are available. In general, they can be grouped under two categories; multivariate global methods, and endpoint-specific or P level methods. Multivariate global methods are used to make an omnibus assessment of the efficacy of the treatments. They provide an overall single ***P* Value** to reflect whether there are any differences among the treatments on the several **correlated** endpoints. The endpoint specific (P level) methods are employed to evaluate how the treatments affect the individual endpoints. They involve applying individual statistical tests to each endpoint separately. In this article, we focus only on the multivariate global methods, and we leave the endpoint-specific methods to another article entitled **Multiple Endpoints, *P* Level Procedures**.

Under the category of multivariate global methods, there are two general types of test. The first type involves combining multiple endpoints into a single test statistic, so that it provides an overall (omnibus) statement to declare whether there are any significant differences among the treatments. These multivariate global tests emerged from the idea of measuring distance between multivariate populations. The notions of summing, transforming, and examining linear combinations of endpoints are fundamental to the test statistics for this first type of global test. Most of these global test methods account for the combined effect of the endpoints by incorporating their variability, intercorrelation, and often clinical relevance into the statistics. The second type of global methods arises from **Bonferroni**-type adjustments. They adjust the univariate observed P values of the individual tests to maintain the familywise type I error rate at a prescribed level, and they provide an overall probability for the trial (*see* **Level of a Test**). Familywise type I error refers to the type I error associated with rejecting at least one **null hypothesis** incorrectly over all the endpoints under consideration. Generally, any multiple testing methods can be used as a global test because, in principle, if we find rejection in any one of the multiple tests, we could reject the overall null hypothesis. For this entry, we only include P value adjustment methods which are intended specifically to be used in global testing and not as a series of tests on the individual endpoints.

Consider I treatment groups in which K correlated endpoints are measured on $n_1, n_2, \ldots, n_I$ subjects. Let Y_{ijk} represent the measurement of the kth endpoint for the jth subject in group i ($i = 1, \ldots, I$; $j = 1, \ldots, n_i$;$k = 1, \ldots, K$) and $N = n_1 + n_2 + \cdots + n_I$.

Assume that:

1. The vector $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijK})'$ has a **multivariate normal distribution** and the $\mathbf{Y}_{ij}$

are independently distributed with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}$.

2. $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iK})'$ and $\mu_{ik} = \mathrm{E}(Y_{ijk})$ is the mean of the kth endpoint for group i.
3. $\boldsymbol{\Sigma}$ is assumed to be the same for all I populations. $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)'$ consists of the diagonal elements of $\boldsymbol{\Sigma}$ and $\sigma_k^2 = \mathrm{var}(Y_{ijk})$ is the variance of the kth endpoint. The off-diagonal elements of $\boldsymbol{\Sigma}$ are denoted by $\sigma_{kk'} = \mathrm{cov}(Y_{ijk}, Y_{ijk'})$, which is the covariance of the kth and k'th endpoints.

Unless stated otherwise, all the parametric methods that are described below were developed under these assumptions.

Classical Multivariate Global Methods for Two-Sided Alternatives

One-way **multivariate analysis of variance** (MANOVA) is a multivariate analog of the one-way **analysis of variance** (ANOVA). Traditionally, we use MANOVA to compare K endpoints simultaneously among I treatment groups, where both K and I are greater than one. To test the null hypothesis $\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_I$ vs. the alternative that at least one of the equalities does not hold, several test criteria are available. Here, we briefly describe four widely used criteria which are all functions of the **eigenvalues** of $\mathbf{E}^{-1}\mathbf{H}$, where

$$\mathbf{H} = \sum_{i=1}^{I} n_i(\overline{\mathbf{Y}}_{i\cdot} - \overline{\mathbf{Y}})(\overline{\mathbf{Y}}_{i\cdot} - \overline{\mathbf{Y}})',$$

$$\mathbf{E} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_{i\cdot})(\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_{i\cdot})', \qquad (1)$$

where $\overline{\mathbf{Y}}_{i\cdot} = \sum_{j=1}^{n_i} \mathbf{Y}_{ij}/n_i$ and $\overline{\mathbf{Y}} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \mathbf{Y}_{ij}/N$. $\mathbf{H}$ is often called the model or hypothesis matrix, and $\mathbf{E}$ is the error matrix. They are generalizations of the between- and within-groups sums of squares in a one-way ANOVA.

Wilks [41] proposed the criterion Λ which is developed from the **likelihood ratio test** and is based on the determinant of $\mathbf{E}(\mathbf{E}+\mathbf{H})^{-1}$. The Wilks' Λ (*see* **Lambda Criterion, Wilks'**) is given as follows:

$$\Lambda = \prod_{w=1}^{s}\left(\frac{1}{1+\lambda_w}\right) = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}, \qquad (2)$$

where $s = \min(I-1, K)$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s$ are the eigenvalues of the characteristic equation $|\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I}_s| = 0$ ($\mathbf{I}_s$ is the identity matrix, of dimension $s \times s$).

The exact distribution of these statistics has been derived for the special cases: (i) $I = 2$ and $K \geq 1$; (ii) $I = 3$ and $K \geq 1$; (iii) $I \geq 2$ and $K = 1$; (iv) $I \geq 2$ and $K = 2$ [18]. For other cases, there are fairly good approximations available. For large sample sizes, we can use the Bartlett chi-square approximation [2] which has approximately a **chi-square distribution** with $K(I-1)$ **degrees of freedom** (df). We reject H_0 at the α level of significance if

$$\chi_b^2 = -\left[\left(\sum_{i=1}^{I} n_i - 1\right) - \tfrac{1}{2}(K+I)\right]\ln\Lambda \geq \chi_{1-\alpha}^2[K(I-1)]. \qquad (3)$$

The other modification is due to Rao [32], and is defined as

$$F_r = \frac{\nu_2}{\nu_1}\left(\frac{1-\Lambda^{1/b}}{\Lambda^{1/b}}\right) \qquad (4)$$

where $\nu_1 = K(I-1)$ and $\nu_2 = ab - [\frac{1}{2}K(I-1)] + 1$, with

$$a = \left(\sum_{i=1}^{I} n_i - 1\right) - \tfrac{1}{2}(K+I)$$

and

$$b = \left\{\frac{[K^2(I-1)^2 - 4]}{[K^2 + (I-1)^2 - 5]}\right\}^{1/2}.$$

Rao's F_r has approximately an $\boldsymbol{F}$ **distribution** with ν_1 and ν_2 df.

The second criterion for testing H_0 is the **Lawley–Hotelling trace** criterion [15, 16, 23], which is defined as the sum of the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$:

$$U = \sum_{w=1}^{s}\lambda_w = \text{trace } (\mathbf{E}^{-1}\mathbf{H}). \qquad (5)$$

Davis [5–7] gives upper percentage points of the test statistics $[(N-I)/(I-1)]U$. We reject H_0 for large values of the statistic.

The third criterion, **Pillai's trace**, was developed by Pillai [29] and Bartlett [3]. It is defined as the sum of the eigenvalues of $\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}$:

$$V = \sum_{w=1}^{s}\frac{\lambda_w}{1+\lambda_w} = \text{trace } [\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}] \qquad (6)$$

Schuurmann et al. [35] give the upper percentage points of the test statistic V, indexed by $s = \min(K, I-1)$, $m = [|K-(I-1)|-1]/2$ and $\mathbb{N} = (N-K-I-1)/2$.

All of the criteria described so far have involved all the eigenvalues. **Roy's maximum root criterion** [34] only involves the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$. His statistic is defined as

$$R = \frac{\lambda_1}{1+\lambda_1}. \tag{7}$$

The statistical significance of R can be assessed by using Pearson & Hartley's table [27] or Pillai's table [30] for $s > 5$ with the three parameters (i.e. s, m, and $\mathbb{N}$).

There are no uniformly **most powerful** MANOVA tests. Based on a study of comparative **powers**, Olson [26] concluded that Roy's largest root test has greater power than the others when the differences among groups are concentrated in one canonical dimension, while the other tests have greater power than Roy's test when the standardized measure of the distance between group means is not so heavily concentrated in a single root.

A large number of **Monte Carlo** studies have also been conducted to investigate the extent to which MANOVA tests are robust to violations of multivariate normality and equality of covariance matrices [26]. For general protection against departures from normality and from homogeneity of the covariance matrices in the **fixed-effects** model, Olson recommended the Pillai–Bartlett trace criterion (i.e. V) as the most robust of the MANOVA tests, with adequate power against a variety of alternatives (*see* **Multivariate Techniques, Robustness**).

There is a special case of MANOVA under the fixed-effects model worthy of separate discussion. When $I = 2$, the MANOVA test is equivalent to the **Hotelling's T^2 test** [15, 16]. It is the classical multivariate method to compare K endpoints simultaneously between two populations. It is the only known nontrivial **unbiased** test which is invariant with respect to affine transformations [37]. To test the hypothesis $\mathrm{H}_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ vs. $\mathrm{H}_\mathrm{a} : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0}$, we calculate T^2:

$$T^2 = \left(\frac{n_1 n_2}{n_1+n_2}\right)(\overline{\mathbf{Y}}_{1\cdot} - \overline{\mathbf{Y}}_{2\cdot})'\mathbf{S}_p^{-1}(\overline{\mathbf{Y}}_{1\cdot} - \overline{\mathbf{Y}}_{2\cdot}), \tag{8}$$

where $\overline{\mathbf{Y}}_{i\cdot}$ is a $K \times 1$ vector of sample means with $i = 1, 2$, and $\mathbf{S}_p = \left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_{i\cdot})(\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_{i\cdot})'\right] / (n_1+n_2-2)$ is a $K \times K$ pooled sample covariance matrix.

Under the null hypothesis,

$$F_{ht} = \frac{n_1+n_2-K-1}{K(n_1+n_2-2)}T^2 \tag{9}$$

has a central F distribution with K and n_1+n_2-K-1 df. If $F_{ht} > F_{1-\alpha}(K, n_1+n_2-K-1)$, then we reject the null hypothesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

The Hotelling T^2 test can also be viewed as a matrix generalization of the two-sample t test (*see* **Student's *t* Statistics**). The quantity T^2 can be interpreted as the square of the maximum possible univariate t computed on any linear combination of various endpoints. This test is appropriate when we have no specific alternative of how the various endpoints differ between two treatment groups. Roy's **union–intersection principle** shows that Hotelling T^2 test is of size α, so it has an accurate control over the type I error [34]. The robustness of Hotelling's T^2 has been examined by different authors [26]. The general conclusion is that T^2 is quite robust against nonnormality and heterogeneity of variance. However, when the sample sizes are unbalanced and the variance of the group that has fewer observations is larger than the other group, the test may not be robust. This test is best used for two-sided alternatives without prior knowledge of the relative magnitude of the treatment differences. It has poor power for alternatives which correspond to an equal (even if unknown in magnitude) beneficial treatment effect on all endpoints.

O'Brien-Type Procedures for One-Sided Alternatives

O'Brien [25] discussed the inadequacy of Hotelling's T^2 and Bonferroni method in dealing with the more common clinical setting when most or all of the efficacious measures are expected to be improved. The hypotheses are $\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_I$ vs. the alternate $\mathrm{H}_\mathrm{a} : \mu_{ik} \geq \mu_{i'k}$ for $k = 1, \ldots, K$, with strict inequality for at least one k. The ith treatment is more effective than the i'th treatment. This problem received extensive attention during the 1960s. Both Kudo [21] and Perlman [28] derived the likelihood ratio tests for this one-sided multivariate problem. The computations are difficult. Instead of tackling the exact distribution for the one-sided alternatives,

O'Brien derives his tests based on a more limited model for the alternate hypothesis. He assumes that the standardized treatment differences for the K endpoints are all of equal magnitude and in the same direction. In symbols, for a generic variable if μ_{ik} and $\mu_{i'k}$ are the means of the kth endpoint for the ith and i'th group, respectively, and σ_k is the common standard deviation of the kth endpoint, then the effect (also called the standardized effect) is

$$\frac{\mu_{ik} - \mu_{i'k}}{\sigma_k}. \tag{10}$$

O'Brien basically considered the problem of generating global tests for alternate hypotheses when the effects are positive and equal for all K endpoints. He proposed three methods to handle this one-sided hypothesis testing. One of these global methods is a simple **nonparametric** rank-sum test. For this method, we first ignore the group assignment of each subject and rank the $N(= n_1 + n_2 + \cdots + n_I)$ subjects separately for each endpoint. We then add the ranks of the K endpoints for the jth subject in the ith group to obtain

$$S_{ij} = \sum_{k=1}^{K} r_{ijk}.$$

Next we apply an appropriate univariate statistical tests to these new rank-sum data $\{S_{ij}\}$.

This method reduces the K-variate observations of each subject to a sum of K ranks. According to O'Brien, these sums are uncorrelated asymptotically, so the **central limit theorem** can ensure this nonparametric procedure to maintain the size of the test in large samples. He further showed with **simulations** that this procedure is relatively **efficient**. This nonparametric test is recommended, in particular, when the variables are not normally distributed or when the sample size is small.

The other two global methods proposed by O'Brien are based on a multivariate **general linear model**. His model assumes that the standardized treatment differences [see (10)] of the K endpoints are of the same magnitude and in the same direction. Explicitly, this assumes that the endpoints are equally important. He derived his parametric statistics by applying, respectively, the ordinary **least squares** (OLS) and the generalized least squares (GLS) methods to the standardized variables, denoted by $\{Y^*_{ijk}\}$. These are obtained by subtracting from each observation (i.e. Y_{ijk}) the overall variable mean (i.e. $\overline{Y}_{\cdot\cdot k}$) and then dividing the difference by the pooled within-group sample standard deviation (i.e. $s_{\cdot\cdot k}$).

For the OLS procedure, this is equivalent to computing the mean of the K standardized observations for each subject and performing a one-way analysis of variance on the sum variables. For the GLS procedure, O'Brien utilizes the best linear unbiased estimate provided by the GLS method and proposed the following statistic:

$$F_{\mathrm{GLS}} = \frac{\sum_{i=1}^{I} n_i [\mathbf{J}'\hat{\mathbf{R}}^{-1}(\overline{\mathbf{Y}}^*_{i\cdot} - \overline{\mathbf{Y}}^*_{\cdot\cdot})]^2}{(I-1)\mathbf{J}'\hat{\mathbf{R}}^{-1}\mathbf{J}}, \tag{11}$$

where $\mathbf{J}$ is a $K \times 1$ vector of ones, $\overline{\mathbf{Y}}^*_{i\cdot} = \sum_{j=1}^{n_i} \mathbf{Y}^*_{ij}/n_i$ is a vector of sample means of the K standardized endpoints for group i, $\overline{\mathbf{Y}}^*_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \mathbf{Y}^*_{ij}/N$ is a vector of overall sample means for the K standardized endpoints, and $\hat{\mathbf{R}}$ is a pooled sample correlation matrix. The elements of $\hat{\mathbf{R}}$ are defined by

$$\hat{R}_{uv} = \left[\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y^*_{iju} - \overline{Y}^*_{i\cdot u})(Y^*_{ijv} - \overline{Y}^*_{i\cdot v})\right] \Big/ (N-I)$$

where $u, v = 1, \ldots, K$.

This procedure incorporates, in its statistic, the information contained in the correlation matrix. An endpoint which is highly correlated to other endpoints will receive less weight, so that the individual contribution of such an endpoint to the statistic will be smaller. As reported by O'Brien, if the underlying distributions are multivariate normal, the OLS statistic follows a standard F distribution with $I-1$ and $N-I$ df and the GLS statistic approximates a standard F distribution with $I-1$ and $N-IK$ df.

As pointed out by O'Brien, the proposed procedures are designed to detect departures in which improvement was demonstrated consistently among the endpoints. Thus, it is important to anticipate a priori the same changes in the treatment effects across all K endpoints. His procedures are inappropriate to apply in situations in which nonzero treatment effects are expected to occur in only a few endpoints, or when there is no prior knowledge as to whether the treatment is going to affect the endpoints consistently. Both Pocock et al. [31] and Lehmacher et al. [24] discussed O'Brien's GLS statistic in the two-sample

case, and presented the following statistic:

$$t_{\mathrm{GLS}} = \frac{\mathbf{J}'\mathbf{R}^{-1}(\overline{\mathbf{Y}}^{*}_{1.} - \overline{\mathbf{Y}}^{*}_{2.})}{[(1/n_1 + 1/n_2)\mathbf{J}'\mathbf{R}^{-1}\mathbf{J}]^{1/2}} = \frac{\mathbf{J}'\mathbf{R}^{-1}(\overline{\mathbf{Z}}_{1.} - \overline{\mathbf{Z}}_{2.})}{(\mathbf{J}'\mathbf{R}^{-1}\mathbf{J})^{1/2}}. \quad (12)$$

If the underlying distributions are multivariate normal and $\mathbf{R}$ is known, this two-sample GLS statistic follows a standard normal distribution under H_0. If $\mathbf{R}$ is unknown and is replaced by an estimator $\hat{\mathbf{R}}$, the statistic is asymptotically normally distributed. Pocock et al. [31] adapted the application of the two-sample GLS statistic to different set of asymptotically normal test statistics which are obtained from continuous, binary and survival data. The robustness of these asymptotic normal statistics has not been established for finite sample sizes and investigations are ongoing. Another important extension that they made to the GLS statistic is to attach unequal priorities to the endpoints. O'Brien's original GLS statistic assumes that all endpoints are equally important. This assumption is necessary for O'Brien's method to achieve optimality, but it does not necessarily coincide with the clinical relevance. For this aspect, Pocock et al.proposed a simple method to attach unequal weights to the endpoints and these weights were chosen to correspond to the relative clinical importance of the endpoints. The proposed test statistic has the following form:

$$\frac{\mathbf{J}'(\mathbf{WRW})^{-1}\mathbf{W}(\overline{\mathbf{Z}}_{1.} - \overline{\mathbf{Z}}_{2.})}{[\mathbf{J}'(\mathbf{WRW})^{-1}\mathbf{J}]^{1/2}}, \quad (13)$$

where $\mathbf{W}$ is a diagonal weighting matrix assumed to be known a priori.

For analyzing data from a group sequential clinical trial, Tang et al. [39] derived a statistic assuming that the K endpoints have unequal standardized effects, which turns out to be equivalent to the modified GLS statistic proposed by Pocock et al. [31]. O'Brien's GLS statistic can be viewed as a special case of this modified statistic. Assume that patients are entered sequentially and that an interim analysis will be undertaken on the accumulated data after each accrual of $2n$ patients (*see* **Data and Safety Monitoring**). For the first j groups of data with $2n$ patients in each group (accrued between the $(m-1)$st and mth analyses, $m = 1, \ldots, j$), we have the following statistic:

$$G = \frac{(nj/2)^{1/2}\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\mathbf{d}^{(j)}}{(\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta})^{1/2}}, \quad \mathbf{d}^{(j)} = \overline{\mathbf{Y}}^{(j)}_{1.} - \overline{\mathbf{Y}}^{(j)}_{2.}, \quad (14)$$

where $\boldsymbol{\delta}$ is a $K \times 1$ vector of relative difference of interest in the K endpoints and its elements are all assumed positive, and $\mathbf{d}^{(j)}$ is the vector of differences in sample means computed from the first j groups of patients. G is normally distributed when $\mathbf{d}^{(j)}$ has a multivariate normal distribution and $\boldsymbol{\Sigma}$ is known. This statistic is intended also to be used in the power and **sample size** calculations so that a cost-effective sequential trial can be designed.

Despite the usefulness of the different extensions, there are a few problems with the O'Brien-type tests that are worth noting. First, they are based on the weighted statistics as given in (11), (12), and (14). They use $\mathbf{J}'\mathbf{R}^{-1}$ or $\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}$ as the weights. As explained and illustrated by Pocock et al. [31] and Follman [11], it is possible for these weights to have negative components for certain correlation matrices. In practice, it does not make sense to include negative weights. To make matters worse, a negative treatment difference weighted by a negative weight may mislead the investigator to conclude efficacy in favor of a treatment, even though the treatment is worse than the control. To avoid this problem, Tang et al. [38] recommended to use O'Brien's OLS test. The second issue concerns the restricted optimality of the GLS test. The GLS test is optimal only when the treatment differences for the K endpoints have the same effect size and direction as specified in the model. As mentioned previously, the GLS test is inappropriate when the nonzero treatment effects are expected to occur only in a few endpoints, or when it is not clear whether the treatment effects are of equal magnitude or direction among the endpoints. The third issue relates to the control of type I error rates. Even though O'Brien's parametric methods are more powerful than the Bonferroni method and Hotelling's T^2 when improved treatment effects are found for all endpoints, they do not always control the prescribed or nominal level of significance when the sample size is small [22]. The actual type I error may exceed the nominal and so produce a test that is liberal. In a simulation study with repeated measures setting, O'Brien also indicated that when the variances are heterogenous, the GLS procedure may

seem to enhance power, but this is at the expense of producing a liberal test.

Likelihood Ratio Test for One-Sided Alternative

Approximate Likelihood Ratio Test

The O'Brien-type statistics are derived under a restricted model of the general one-sided alternative, and they provide an optimal test for alternatives that correspond to a half-line. A half-line is simply a specified vector projection in the positive orthant. When it corresponds to $\boldsymbol{\delta}$, which is a column vector with all positive elements, it represents the specified relative size of the treatment effects for the K endpoints. For example, $\boldsymbol{\delta} = (1, 2)'$ indicates an alternate hypothesis in which the relative size of the treatment effect for the second endpoint is twice as large as that for the first endpoint. $\mathbf{J}$ of (11) is a special case of $\boldsymbol{\delta}$ which has all the elements equal to unity and assumes the relative sizes of the treatment differences among the K endpoints are all equal. The sensitivity of the O'Brien-type tests relies on the agreement of the observed relative difference with the specified $\boldsymbol{\delta}$. In other words, they require prior knowledge of $\boldsymbol{\delta}$ in order to achieve the intended power. When the choice of $\boldsymbol{\delta}$ is not clear, they are not appropriate. In this situation, a better approach is to find a test that is powerful over all possible $\boldsymbol{\delta}$ with nonnegative components. A logical candidate is to construct a test using the maximum of the GLS statistic over all feasible $\boldsymbol{\delta}$. This coincides with constructing an optimal test for alternatives which lie in the positive orthant. Tang et al. [40] provided such a test through an approximation of the likelihood ratio test and they called it the approximate likelihood ratio (ALR) test. Tang [37] give a detailed discussion on some uniformly more powerful tests which are related to the development of the ALR test.

The ALR test involves transforming the correlated vector of mean differences, $\mathbf{d}$, to a vector of independent standardized normal variables, $\mathbf{z}$, by a transformation matrix $\mathbf{A}$:

$$\mathbf{z} = \left(\frac{n}{2}\right)^{1/2} \mathbf{A}\mathbf{d}, \tag{15}$$

where $\mathbf{A}'\mathbf{A} = \boldsymbol{\Sigma}^{-1}$. Note that the matrix $\mathbf{A}$ is not unique. Tang et al. [40] give a computational **algorithm** to obtain the appropriate $\mathbf{A}$. Once $\mathbf{A}$ is selected, we can compute the test statistic

$$g(\mathbf{z}) = \sum_{k=1}^{K} (z_k \vee 0)^2 \tag{16}$$

where z_k is the kth element of $\mathbf{z}$ and $(z_k \vee 0)$ is the maximum of 0 and z_k. The P value of $g(\mathbf{z})$ under the null hypothesis is calculated by

$$\Pr[g(\mathbf{z}) \geq c] = \frac{1}{2^K} \sum_{k=1}^{K} \binom{K}{k} \Pr[\chi^2_{1-\alpha}(k) \geq c], \tag{17}$$

which is a special case of the chi-bar-square distribution [33]. It turns out that $g(\mathbf{z})$ based on $\mathbf{A}$ may not be invariant under a different ordering of the endpoints. In order to achieve permutation invariance, Tang et al. [40] suggested applying a linear ordering algorithm to the endpoints prior to the selection of $\mathbf{A}$. Tang et al. [38] provided a simplified algorithm for the ordering.

The ALR test can be powerful when the treatment improves all the endpoints, but differences in treatment effects are larger for some endpoints than for others. Theoretically, the power of the ALR test should be more stable than that of O'Brien's GLS test because O'Brien's test approximates the image space by only a half-line, while the ALR test approximates the image space by a cone. Tang et al. [40] performed some simulation studies to compare the powers of Hotelling's T^2, O'Brien's test, and the ALR test. They confirmed that the ALR test has better power than Hotelling's T^2 and O'Brien's test when the treatment effects for the various endpoints are positive but unequal. However, the ALR test can be liberal when the sample size is small (say, 40 subjects per treatment). Follmann [11] also pointed out that, for the case of two endpoints with correlation equal to 0.75 or higher, the ALR test may reject the null hypothesis when the treatment effects are negative for both enpoints.

Other Multivariate Global Methods

Follmann's X^2_+ Test for One-Sided Alternatives

In general, the likelihood ratio tests for a one-sided alternative are quite difficult to implement in a practical situation. Tang et al. [40] have produced a simpler ALR procedure, but the computations involved

are still intensive, especially when there are a large number of endpoints. Follmann [12] proposed a relatively simple multivariate test with a one-sided alternatives. For the two-sample case, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are assumed to be the endpoint differences, which are independent and identically distributed random vectors from a K-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Assume that $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_K)$ is a vector of the sample mean differences. His test rejects $\mathrm{H}_0 : \boldsymbol{\mu} = \mathbf{0}$ at level α if the statistic

$$n\overline{\mathbf{X}}' \boldsymbol{\Sigma}^{-1} \overline{\mathbf{X}} \tag{18}$$

exceeds the 2α critical value from a chi-square distribution with K df *and* the sum of the elements of the sample mean differences vector exceeds zero (i.e. $\sum_{k=1}^{K} \overline{X}_k > 0$). This test statistic is presented symbolically as the X_+^2 test. There is no formal theoretical justification for this test, but Follmann [12] provided some formal statements and proofs that the X_+^2 test is an α-level procedure for covariance known or unknown. He also gave tight bounds on the power of the test, so that an analytic approach to calculating power for one-sided multivariate studies is possible. Simulation studies show that the X_+^2 test performs better than O'Brien's GLS test when the treatment effects are all positive but not equal in magnitude.

Läuter's Standardized Sum and Principal Component Tests

Similar to O'Brien's OLS statistic, Läuter [22] used the idea of forming a weighted sum of the K endpoints and proposed three new tests to analyze the multiple endpoints data, especially when the sample size is small or when the number of endpoints is greater than the sample size.

He proposed to combine the K observations of each subject into a weighted sum, $x_{ij} = \boldsymbol{\gamma}' \mathbf{Y}_{ij}$, where $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, and $\boldsymbol{\gamma}$ is a $K \times 1$ vector of weights uniquely determined by the total covariance matrix $(\mathbf{Y} - \overline{\mathbf{Y}})(\mathbf{Y} - \overline{\mathbf{Y}})'$ where $\mathbf{Y} = (\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1}, \ldots, \mathbf{Y}_{I1}, \ldots, \mathbf{Y}_{In_I})$ is a $K \times N$ matrix of observations and $\overline{\mathbf{Y}} = (1/N)\mathbf{Y}\mathbf{J}_N\mathbf{J}_N'$ is a $K \times N$ matrix of overall means for the K endpoints. For the comparison of I populations in the sense of a one-way analysis of variance, he suggested the statistic

$$F_L = \frac{h^2}{(I-1)s^2}, \tag{19}$$

where $h^2 = \mathbf{x}'\mathbf{V}\mathbf{V}'\mathbf{x}$, $s^2 = [1/(N-I)]\mathbf{x}'[\mathbf{I}_N - (1/N)\mathbf{J}_N\mathbf{J}_N' - \mathbf{V}\mathbf{V}']\mathbf{x}$, and $\mathbf{V}$ is an $N \times (I-1)$ matrix with $\mathbf{V}'\mathbf{V} = \mathbf{I}$ and $\mathbf{J}'_N\mathbf{V} = \mathbf{0}$. Based on the theory of spherical matrix distributions developed by Dawid [8, 9] and by Fang & Zhang [10], he showed that this statistic is exactly distributed as $F_{1-\alpha}(I-1, N-I)$. No explicit form of $\mathbf{V}$ is provided in his paper.

For the case of two populations, he proposed the statistic

$$t = \frac{h}{s}, \tag{20}$$

where $h = \mathbf{x}'\boldsymbol{\kappa}$, $s^2 = [1/(N-2)]\mathbf{x}'[\mathbf{I}_N - (1/N)\mathbf{J}_N\mathbf{J}_N' - \boldsymbol{\kappa}\boldsymbol{\kappa}']\mathbf{x}$, and $\boldsymbol{\kappa}$ is an N-dimensional vector with $\boldsymbol{\kappa}'\boldsymbol{\kappa} = 1$ and $\mathbf{J}_N'\boldsymbol{\kappa} = 0$. He showed that the statistic is exactly distributed as **Student's *t*** with $N-2$ df. The explicit form of

$$\boldsymbol{\kappa} = \left[\frac{n_1 n_2}{(n_1+n_2)}\right]\begin{pmatrix}(1/n_1)\mathbf{J}_{n_1}\\ -(1/n_2)\mathbf{J}_{n_2}\end{pmatrix}$$

is given. For this special case, (20) is equivalent to applying the univariate two-sample t test to the transformed variables $\{x_{ij}\}$ with $n_1 + n_2 - 2$ df. In this paper, the author proposed three explicit forms of transformations for $\mathbf{x}$. They are the standardized sum transformation and the two **principal component** transformations.

The test based on the standardized sum transformation is useful when the effects of the K endpoints are all equal. For this test, the original observations are transformed as follows:

$$z_{ij} = \sum_{k=1}^{K} \frac{Y_{ijk}}{\left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ijk} - \overline{Y}_{\cdot\cdot k})^2\right]^{1/2}}. \tag{21}$$

Here, the original observations, Y_{ijk}, are standardized by the total variances. Recall that in O'Brien's procedure he standardized the variables by the within-sample variances.

The first principal component test is appropriate if all K variables are expected to have the same directions of treatment effects. For this test, he suggested using

$$z_{ij} = \sum_{k=1}^{K} |e_k| Y_{ijk}, \tag{22}$$

where e_k is the kth element of the first **eigenvector** obtained by solving

$$(\mathbf{Y} - \overline{\mathbf{Y}})(\mathbf{Y} - \overline{\mathbf{Y}})'\mathbf{e} = \text{diag}\,[(\mathbf{Y} - \overline{\mathbf{Y}})(\mathbf{Y} - \overline{\mathbf{Y}})']\mathbf{e}\lambda. \tag{23}$$

If we have no expectation of the direction of treatment differences, Läuter suggests using the second principal component test involving a transformation similar to (22) but with the absolute sign removed.

As stated in Läuter [22], all three proposed methods have accurate control of the type I error; but there is no discussion of the performance of these tests in terms of power and robustness.

Cureton & D'Agostino's Composite Scores

Similar to Läuter's transformed variables for the principal component test, Cureton & D'Agostino [4] presented – much earlier in time – an approach for obtaining a transformed variable through a principal component analysis. They called this transformed variable a composite score. It gives the best single measure of whatever is common to a set of at least moderately similar variables. The composite score is a weighted sum of all the variables, using the loadings on the first principal component.

The composite score of the jth subject is defined as

$$C_{ij} = a_1 Y^*_{ij1} + a_2 Y^*_{ij2} + \cdots + a_K Y^*_{ijK}, \tag{24}$$

where the $\{a_k\}$ are the loadings of the K variables on the first principal component, and the $\{Y^*_{ijk}\}$ are the standardized scores of the K variables for the jth subject in group i. The C_{ij}s are a set of deviation scores with mean 0, but they are not standard scores. To obtain the composite standard scores, we divide each C_{ij} by the standard deviation of these deviation scores, which is equal to the square root of the first eigenvalue. We can then apply the usual t statistic to these composite standard scores. It turns out that Cureton & D'Agostino's composite score is related to Läuter's principal component transformed variables, except that Cureton & D'Agostino use the standardized variables Y^*_{ijk}.

Another procedure for forming a composite score, described in Cureton & D'Agostino [4], is based on ranks. It was first presented in Kendall [19, 20]. It is a useful procedure when both the sample size and the number of endpoints are small. Ignoring the group assignment, the N subjects are first ranked on each of the K variables, using the average-rank procedure to resolve ties. The ranks of each individual are then summed, and the rank-sums are themselves ranked to obtain the composite ranks. An appropriate nonparametric rank sum test can be used on these composite ranks. This test is appropriate when the sample size is small or when normality does not hold for the data. This procedure is similar to O'Brien's rank-sum procedure.

Risk Score Test

Follmann [11] proposed a risk score statistic, which is derived on the basis of the clinical appeal rather than optimality in statistical power. Instead of using some optimal estimates as weights, he chooses weights so that the weighted sum correlates well with the occurrence of an event. For example, the risk score weights can be the estimated **logistic regression** coefficients for the endpoints from an ancillary data set. This risk score test requires the multiple endpoints to be surrogates. His derivation is based on the two treatment groups with two endpoints measured on n subjects in each group. His statistic is defined as

$$\begin{aligned} RS &= \beta_1(\overline{Y}_{1.1} - \overline{Y}_{2.1}) + \beta_2(\overline{Y}_{1.2} - \overline{Y}_{2.2}) \\ &= \beta_1 d_1 + \beta_2 d_2 \\ &= \gamma_1 d_1^* + \gamma_2 d_2^*, \end{aligned} \tag{25}$$

where $\gamma_k = \beta_k \sigma(d_k)$ and $d_k^* = d_k/\sigma(d_k)$, with $\sigma(d_k)$ the standard deviation of the mean of the group differences in the kth endpoint. Under the null hypothesis, the statistic has a normal distribution with mean zero and variance $\text{var}(\rho) = \gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2$. For a one-sided test, we reject H_0 in favor of the treatment group at 0.05 level of significance if $\boldsymbol{\gamma}(\mathbf{d^*})' > 1.645[v(\rho)]^{1/2}$. The advantage of this risk score test is having a rejection region that corresponds to the contours of constant risk, and that the correlations of the endpoints will not affect the clinically relevant rejection region. Based on his simulation study, he concluded that the risk score test is comparable to the O'Brien GLS test in terms of power when one discounts the wrongful rejections due to the negative weights arising from the GLS test.

P Value Adjustment, Global Methods

Bonferroni-Type Methods

The Bonferroni method is often used to adjust for multiple significance tests, and it is also a popular method for testing the overall hypothesis $H_0 = \bigcap\{H_k : k = 1, \ldots, K\}$, where $H_1, \ldots, H_K$ are the hypotheses for testing the K individual endpoints. Suppose that $t_1, \ldots, t_K$ is a set of K statistics with corresponding P values $P_1, \ldots, P_K$ for testing these hypotheses. Using the Bonferroni inequality, we can perform a very simple level α test of H_0:

$$\text{reject } H_0 \text{ if } P_{(1)} \leq \frac{\alpha}{K},$$

where $P_{(1)}$ is the smallest P value.

This method is easy and convenient to apply. No distributional assumptions are required. It is particularly useful in detecting nonzero treatment effects in one or some few "unknown" and distinct endpoints. Nevertheless, the method may become very conservative, especially when the endpoints are highly correlated. Also, it puts too much emphasis on the smallest P value and does not account for the collective information that the various endpoints provide. It is inappropriate in the situation in which most or all measures of efficacy are improved.

Hommel's Adjustment Methods

Hommel [13] discussed another level α test due to Rüger's inequality. It avoids the disadvantage of overemphasis on the smallest P value. The test is to

$$\text{reject } H_0 \text{ if } P_{(k)} \leq \frac{k\alpha}{K},$$

where $P_{(k)}$ is the kth smallest P value. Using this test, the investigator has to determine k before performing the K comparison tests. There are no specific guidelines on how to select k. This mainly depends on the decision of the investigator prior to the trial. Even though k is chosen in advance, there is always room for argument. To avoid choosing k in advance, Hommel proposed another level α test which combines the Bonferroni test and all $(K - 1)$ possible Rüger tests:

$$\text{reject } H_0 \text{ if } P_{(k)} \leq \frac{k\alpha}{(KC_K)} \quad \text{for at least one } k,$$

where $k = 1, \ldots, K$ and $C_K = 1 + 1/2 + \cdots + 1/K$. Hommel [14] indicated that these level α tests of H_0 are expected to be conservative in practical applications.

Simes' Adjustment Method

Simes [36] proposed a modified Bonferroni procedure based on the ordered P values for the individual tests. It is very similar to the modified Rüger test, but is less conservative because of dropping the constant C_K. The Simes test is performed as

$$\text{reject } H_0 \text{ if } P_{(k)} \leq \frac{k\alpha}{K} \quad \text{for at least one } k.$$

As Simes pointed out, his procedure does not always lead to a level α test of H_0. His simulation study showed that the level of his procedure is less than or equal to the nominal α for a large family of multivariate distributions. He also proved that the level is exactly equal to α if the test statistics are independent. This modified procedure improves on some of the major drawbacks of the Bonferroni method. It does not rely heavily on the smallest P value, and it has better power than Bonferroni's method because it has an actual significance level much closer to the nominal level. Although Simes' procedure slightly increases the computation as compared to Bonferroni, it is still very easy and convenient to perform. However, just like Bonferroni's adjustment, Simes' method does not consider the correlations between endpoints and it becomes conservative as the correlations among the K test statistics increase.

Armitage & Parmar's Empirical Method

The P value adjustment methods discussed so far do not directly account for the correlations of the endpoints. Armitage & Parmar [1] have presented a procedure which allows for correlations in the P value adjustment. Assume that the test statistics follow a multivariate normal distribution. They suggested an adjusted correction of the following form:

$$P_{\text{adj}} = 1 - (1 - P_{\min})^{K^x} \tag{26}$$

where $0 \leq x \leq 1$. The parameters $x = 1$ if the K test statistics are independent, and $x = 0$ if the K test statistics are fully correlated. In general, the empirical

formula for computing x with an arbitrary correlation structure is given as follows:

$$x = \begin{cases} \hat{x}, & \text{if } 0 < \hat{x} < 1, \\ 0, & \text{if } \hat{x} \le 0, \\ 1, & \text{if } \hat{x} \ge 1, \end{cases}$$

where

$$\hat{x} = \{1 - [\overline{\rho} + 2V(\rho) + a(p_{\min} - 0.05)]^2\}^{1/\gamma},$$

$$\overline{\rho} = \frac{1}{n_K} \sum_k \sum_l |\rho_{kl}|, \quad \text{where } k < l,$$

$$V(\rho) = \frac{1}{n_K} \sum_k \sum_l (|\rho_{kl}| - \overline{\rho})^2, \quad \text{where } k < l,$$

$$\gamma = \begin{cases} \dfrac{n_K}{n_K - 1} + \left(3 - \dfrac{n_K}{n_K - 1}\right)(\overline{\rho} + 2V(\rho)), & \\ & k \ge 3, \\ 0.4, & k = 2, \end{cases}$$

$$a = 3.25 - 2.7(100P_{\min}) + 0.4(100P_{\min})^2,$$

$$n_K = \left(\tfrac{1}{2}\right) K(K-1).$$

They found that the adjusted P values provide good approximations to the actual P values for up to five endpoints, but further studies are required to justify the adequacy of this adjustment for higher dimensions.

James's Analytic Method

James [17] presents another P values adjustment method which allows for the presence of correlations. Her method is based on an approximation derived for multinormal probabilities with equal correlation. Unlike the Bonferroni-type methods, it assumes that the test statistics follow a multivariate normal distribution, or at least an asymptotically normal one with equal correlation ρ. The adjusted P value is defined as

$$\begin{aligned} P_{\text{adj}} &= \Pr(\text{minimum } P \le P_{\min}) \\ &= 1 - \Pr(\text{all } P > P_{\min}) \\ &= 1 - \Pr\left(\bigcap_{k=1}^{K} a \le X_k \le b\right), \end{aligned}$$

where the $\{X_k\}$ are standardized multinormal with equal correlation ρ such that

$$b = \Phi^{-1}\left(1 - \frac{P_{\min}}{2}\right), \qquad a = \Phi^{-1}\left(\frac{P_{\min}}{2}\right),$$

for the two-sided case, and

$$b = \Phi^{-1}(1 - P_{\min}), \qquad a = -\infty,$$

for the one-sided case. Φ^{-1} is the inverse of the cumulative normal distribution function.

The approximation is

$$\begin{aligned} P_{\text{adj}} = 1 &- D_1(1-\rho^2) - D_2\rho^2 - D_3\rho(1-\rho) \\ &- D_4[2 - 2(1-\rho)^{1/2} - \rho - \rho^2] \end{aligned}$$

where, for a two-sided test,

$$\begin{aligned} D_1 &= (1 - P_{\min})^K, \\ D_2 &= 1 - P_{\min}, \\ D_3 &= 0, \\ D_4 &= K(K-1)\phi(b)\int_{-\infty}^{\infty} \Phi(z)^{K-2}\phi(z)^2\,\mathrm{d}z \\ &= K(K-1)\phi(b)G(K), \\ G(K) &= \int_{-\infty}^{\infty} \Phi(z)^{K-2}\phi(z)^2\,\mathrm{d}z, \\ \phi(z) &= \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2}\right), \\ \phi(z) &= \int_{-\infty}^{z} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x, \\ b &= \Phi^{-1}\left(1 - \frac{P_{\min}}{2}\right), \end{aligned}$$

and, for a one-sided test,

$$\begin{aligned} D_1 &= (1 - p_{\min})^K, \\ D_2 &= 1 - p_{\min}, \\ D_3 &= \frac{K}{2}(K-1)(1-p_{\min})^{K-2}\phi(b)^2, \\ D_4 &= \frac{K}{2}(K-1)\phi(b)\int_{-\infty}^{\infty} \Phi(z)^{K-2}\phi(z)^2\,\mathrm{d}z \\ &= \frac{K}{2}(K-1)\phi(b)G(k), \\ b &= \Phi^{-1}(1 - p_{\min}). \end{aligned}$$

With this method,

$$\text{reject } H_0 \text{ if } P_{\text{adj}} \leq \alpha$$

and the probability of rejecting the overall null hypothesis can be quoted as the minimum adjusted P value.

For the unequal correlation case, James suggested: (i) replacement of ρ with the mean correlation ρ_m defined by

$$\rho_m = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \frac{|\rho_{kl}|}{m},$$

where ρ_{kl} is the correlation between endpoint k and endpoint l, $k = 1, \ldots, K-1, l = k+1, \ldots, K$, and $m = 1/2K(K-1)$; or (ii) to replacement of ρ with ρ_{mv},

$$\rho_{mv} = \rho_m + 2 \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \frac{(|\rho_{kl}| - \rho_m)^2}{m}.$$

Estimates for the correlations, ρ_{kl}, can be obtained: (i) from a previous study or preferably a pilot trial; (ii) from the raw data if the endpoints are distributed multivariate normal; or (iii) by using the formulas provided in Pocock et al. [31] if the endpoints come from some nonnormal data.

By incorporating the correlations to the adjusted P values, James' method is definitely a solution to the problem of conservatism arising from the Bonferroni-type adjustment. She showed that the adjusted P values derived from her approximation are very close to the actual values obtained from a multivariate normal program. Furthermore, they appear to be better estimates for the true P values than the approximation proposed by Armitage & Parmar [1]. The adjustment can also be used to calculate the power of any trial with multiple testing. However, there are some trade-offs for this improvement in power. They include the increased complexity of calculations and also the requirement of the distributional assumption imposed on the test statistics. Further investigations need to be done to check the **robustness** of the method and find an error bound for the approximation. Also, it is not clear what will be the impact of various heterogeneous correlation structures on the performance of the method; particularly, when some of the correlations are negative.

Follmann's Maximum Test

Follmann [11] proposed a Bonferroni-type of method which uses the maximum of the individual test statistics. Under his model, he assumes that the statistics follow a multivariate normal distribution with equal correlation. However, instead of using a standard technique to test the maximum, he uses a critical value by the conservative Bonferroni approximation. If the maximum statistic exceeds the α/K critical value, the overall null hypothesis will be rejected at the α level of significance. This max test is equivalent to performing the Bonferroni adjustment on a one-sided P value with the assumption of normality imposed on the individual statistics. Simulations were performed to compare this max test with O'Brien's OLS test. Only two groups and two endpoints are used in his simulation. The simulation result is consistent with the known performance of the Bonferroni method, which has good power in detecting one or a few significant endpoints and performs well with a small number of endpoints. The simulations also indicate that the powers of the max test and the OLS test do not differ substantially under the alternative of equal positive treatment effects for both endpoints. However, these results may be due to the small number of endpoints used. Even though the max test is very simple, more research is needed to explore its usefulness.

Follmann [11] adapted another maximum test due to Wittes for global comparison between two groups. The method pairs the jth subjects from the two groups and reduces the data to a single paired difference for each endpoint, denoted by $d_{jk}, k = 1, \ldots, K$. $\mathbf{d}_j$ is a vector of K paired differences, which is assumed to follow a multivariate normal distribution with equal correlation, ρ. Suppose that V_j is the maximum of the K differences of the jth subjects. Follmann proposed the following statistic:

$$\overline{V}_s = \frac{\overline{V} - \mathrm{E}(V_j|\boldsymbol{\mu}, \rho)}{[\mathrm{var}(V_j|\boldsymbol{\mu}, \rho)]^{1/2}}, \tag{27}$$

where $\overline{V} = \sum_{j=1}^{n} V_j/n$ and $\mathrm{E}(V_j|\boldsymbol{\mu}, \rho)$ and $\mathrm{var}(V_j|\boldsymbol{\mu}, \rho)$ are the mean and variance of V_j, which the author suggests obtaining by **numerical integration**. Under the null hypothesis, this statistic has an asymptotic standard normal distribution by the central limit theorem. Simulations were performed to compare this maximum statistic with O'Brien's OLS test. In his simulation, two groups and two

endpoints were used. The simulation results indicate that O'Brien's method has better power than the maximum test under the alternative that both endpoints improve equally, and has poorer power than the maximum test under the alternative that only one of the two endpoints improves but that the responsive endpoint is unknown. Follmann recommended that this maximum test should be used in situations in which it is suspected that only one of the two endpoints improves and the correlation between the endpoints is large and positive. However, there are no indications as to how to justify the pairing of the subjects and how to extend this test to more than two groups.

Summary Comments

As we review each method, we can see there is no one winner for all settings. The power of each method relies on the particular setting to which it is applied. A few simulation studies [11, 12, 25, 40] have been done to compare the powers of some of the global methods. However, they basically came up with a similar conclusion, that the power of each method varies with the particular alternative assumed in the investigation. As a final note, we summarize the setting for which each global method is most appropriate.

Hotelling's T^2 and MANOVA are still the appropriate methods to be used in any trials when the investigator does not have any prior knowledge on how the treatments may affect the endpoints; for example, in a pilot study. They perform quite well even when the assumptions on multivariate normality and homogeneity of covariance are violated. However, if we have prior information about an equal improvement in the expected differences of the K endpoints, we may want to consider the one-sided global methods or the P value adjustment.

O'Brien-type parametric tests are good candidates for the situation in which we expect that most or all of the endpoints will demonstrate consistent improvement in the treatment effect. For clinical relevance, the O'Brien GLS statistic as modified by Pocock et al. [31] has the flexibility of incorporating the clinical importance of the endpoints as part of the weights, and at the same time preserving the optimality of statistical power. However, we have to be aware that the power of the GLS test may be enhanced at the expense of an increased type I error rate.

Under the expectation of equal improvement for all endpoints, the GLS test is appropriate if we can be sure that there are no negative components in the weights of the statistic [i.e. $\mathbf{J}\mathbf{R}^{-1}$ and $\boldsymbol{\delta}\boldsymbol{\Sigma}^{-1}$ in (11) and (14) respectively] and the sample sizes are large. When we are not sure about the composition of the weights, it is reasonable to use the O'Brien OLS test if we have a large sample size and the rank sum test if the sample size is small. Also, Läuter's standardized sum test seems to be appropriate here, due to its similar transformations as compared with the O'Brien OLS test. A definite advantage of using Läuter's standardized sum test is protection against an erroneously liberal type I error.

In situations in which most or all of the endpoints are expected to be improved, but it is not clear if the relative treatment differences (effects) are going to be equal, we may consider the approximate likelihood test. Again, we have to keep in mind that this gain in power may be misleading because of the inflated actual type I error rate, and also the problem with the negative weights. To solve the problem of the inflated type I error rate, some authors have suggested using the permutation distribution of the ALR test statistic instead of the chi-bar-square distribution. However, this approach is very **computer-intensive** and the level of complexity increases with the number of endpoints. If there are concerns about the complexity of computations and inflated type I error rates, we may consider using the classical tests (MANOVA tests), as long as the changes in treatment effects are not all of the same magnitude. Other candidates are the composite score tests. They can protect the type I error rates and they should perform reasonably well if the endpoints behave similarly. Of course, they will not be as powerful as the ALR test.

The next setting arises when the treatment is effective in only one or a few "unknown" and distinct endpoints. The method proposed by Simes appears to be a good candidate when the correlations among the endpoints are low or moderate (say, $\rho \leq 0.5$). This method is more powerful than the Bonferroni adjustment and the methods proposed by Hommel. The calculation is relatively easy. Also, it works quite well for multivariate distributions other than the normal. When the endpoint correlations are large, Simes' adjustment tends to be conservative. In this case, the

method proposed by James may be a better choice due to its ability to account for correlations. Unfortunately, she only provided a simple example that illustrates a power comparable to O'Brien's test when the treatment effects are almost equal and a power better than O'Brien's when the treatment effects are all positive but not equal. There are no formal power comparisons performed to support this claim in general. Further investigations on this method seem worthwhile.

By design, the composite score methods, such as the standardized sum test, the various versions of principal component composite score tests, and the risk score test, are not optimal with respect to their statistical power in detecting mean differences among groups. Therefore, under the different settings described above, there is always a better test than these composite score methods. However, as pointed by Follmann [11], statistical power should not be the only concern in evaluating treatment efficacy. Clinical considerations are also important in the evaluation process. We should not underestimate the usefulness of these methods, because they put more emphasis on a certain optimal structure of the endpoints, which may be clinically more appealing. However, more research is definitely needed in order to understand better the properties of these methods.

References

[1] Armitage, P. & Parmar, M. (1986). Some approaches to the problem of multiplicity in clinical trials, in *Proceedings of the XIIth International Biometrics Conference*. Biometric Society, Seattle.

[2] Bartlett, M.S. (1938). Further aspects of the theory of multiple regression, *Proceedings of the Cambridge Philosophical Society* **34**, 33–40.

[3] Bartlett, M.S. (1939). A note on tests of significance in multivariate analysis, *Proceedings of the Cambridge Philosophical Society* **35**, 180–185.

[4] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: an Applied Approach*. Lawrence Erlbaum, Hillsdale.

[5] Davis, A.W. (1970). Exact distributions of Hotelling's Generalized T_0^2-test, *Biometrika* **57**, 187–191.

[6] Davis, A.W. (1970). Further applications of a differential equation for Hotelling's generalized T_0^2-test, *Annals of the Institute of Statistical Mathematics* **22**, 77–87.

[7] Davis, A.W. (1980). Further tabulation of Hotelling's generalized T_0^2-Test, *Communications in Statistics – Simulation and Computation* **9**, 321–336.

[8] Dawid, A.P. (1977). Spherical matrix distributions and a multivariate model, *Journal of the Royal Statistical Society, Series B* **39**, 254–261.

[9] Dawid, A.P. (1978). Extendability of spherical matrix distributions, *Journal of Multivariate Analysis* **8**, 559–566.

[10] Fang, K.-T. & Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*. Science Press, Beijing/Springer-Verlag, Berlin.

[11] Follmann, D. (1995). Multivariate tests for multiple endpoints in clinical trials, *Statistics in Medicine* **14**, 1163–1175.

[12] Follmann, D. (1996). A simple multivariate test for one-sided alternatives, *Journal of the American Statistical Association* **91**, 854–861.

[13] Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures, *Biometrical Journal* **25**, 423–430.

[14] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test, *Biometrika* **75**, 383–386.

[15] Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.

[16] Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 23–41.

[17] James, S. (1991). Approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials, *Statistics in Medicine* **10**, 1123–1135.

[18] Johnson, R.A. & Wichern, D.W. (1988). *Applied Multivariate Statistical Analysis*, 2nd Ed. Prentice-Hall, Englewood Cliffs.

[19] Kendall, M.G. (1955). *Rank Correlation Methods*, 2nd Ed. Hafner, New York.

[20] Kendall, M.G. (1957). *A Course in Multivariate Analysis*. *Griffin Statistical Monographs and Courses* No. 2. Griffin, London.

[21] Kudo, A. (1963). A multivariate analogue of the one-sided test, *Biometrika* **15**, 403–418.

[22] Läuter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints, *Biometrics* **52**, 964–970.

[23] Lawley, D.N. (1938). A generalization of Fisher's z test, *Biometrika* **30**, 180–187. Corrections in *Biometrika* **30** (1939) 467–469.

[24] Lehmacher, W., Wassmer, G. & Reitmeir, P (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate, *Biometrics* **47**, 511–521.

[25] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1089.

[26] Olsen, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

[27] Pearson, E.S. & Hartley, H.O., eds (1972). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge University Press, Cambridge.

[28] Perlman, M.D. (1969). One-sided testing problems in multivariate analysis, *Annals of Statistics* **40**, 549–567.

[29] Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis, *Annals of Mathematical Statistics* **26**, 117–121.

[30] Pillai, K.C.S. (1960). *Statistical Tables for Tests of Multivariate Hypotheses*. Statistical Center, University of the Philippines, Manila.

[31] Pocock, S.J., Geller, N.L. & Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics* **43**, 487–498.

[32] Rao, C.R. (1951). An asymptotic expansion of the distribution of Wilks' criterion, *Bulletin of the International Statistics Institute* **33**, 177–180.

[33] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[34] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[35] Schuurmann, F.J., Krishnaiah, P.R. & Chattopadhyay, A.K. (1975). Exact percentage points of the distribution of the trace of a multivariate beta matrix, *Journal of Statistical Computation and Simulation* **3**, 331–343.

[36] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **73**, 751–754.

[37] Tang, D.-I. (1994). Uniformly more powerful tests in a one-sided multivariate problem, *Journal of the American Statistical Association* **89**, 1006–1011.

[38] Tang, D.-I., Geller, N.L. & Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints, *Biometrics* **49**, 23–30.

[39] Tang, D.-I., Gnecco, C. & Geller, N.L. (1989). Design of group sequential trials with multiple endpoints, *Journal of the American Statistical Association* **84**, 776–779.

[40] Tang, D.-I., Gnecco, C. & Geller, N.L. (1989). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials, *Biometrika* **76**, 577–583.

[41] Wilks, S.S. (1932). Certain generalizations in the analysis of variance, *Biometrika* **24**, 471–494.

(*See also* **Isotonic Inference**; **Multiple Comparisons**; **Multivariate Analysis, Overview**; **Simultaneous Inference**)

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Multiple Endpoints, *P* Level Procedures

In **clinical trials** or biomedical research, multiple endpoints are often used to investigate a treatment effect. Diseases usually affect patients in more than one way, and in order to characterize completely the efficacy of a treatment, various endpoints are required. There has been substantial debate and controversy in determining the appropriate approach to analyze these multiple endpoints. The main difficulty is how to account for the multiplicity of inferences. Sometimes this problem comes in two parts owing to the presence of multiple groups as well as multiple endpoints. The problem of multiple groups comparisons has received substantial attention, and there is an extensive literature on procedures of multiple groups comparison (*see* **Multiple Comparisons**; **Paired Comparisons**; **Simultaneous Inference**). In this article we focus on the problem of multiplicity due to the presence of multiple endpoints and describe some methods that can be used to analyze these multidimensional data.

There are researchers who advocate the simple no-adjustment approach. They apply a univariate t test (*see* **Student's *t* Statistics**) or one-way **analysis of variance** (ANOVA) to each endpoint separately and claim a significant effect for a specific endpoint if the observed ***P* value** obtained from the corresponding test is at, or less than, the prescribed level of significance, say α. This approach is legitimate if the objective of the researchers is exploratory or simply to present the comparisons examined and draw no conclusions about the treatments. However, most researchers are not satisfied with this. It is usually desired to reach conclusions about the efficacy of the treatments. The no-adjustment approach is too liberal for this, for it inflates the familywise error rate as the multiple tests are performed. The familywise error rate is the probability of making at least one type I error in the given family of inferences [9], and the type I error is the error of making a false rejection of a **null hypothesis** when it is true (*see* **Hypothesis Testing**). This inflation of the familywise error rate is undesirable, especially in clinical trials, because it may increase the chance of providing an ineffective treatment to patients. The no-adjustment approach is not recommended.

To substantiate the presence of a treatment effect, some researchers take the approach of combining the various endpoints into a single statistic. This supplies an overall probability statement about the effectiveness of a treatment. The procedures that provide a single statistic obtained by combining the multiple endpoints in some optimal way are generally called *global procedures* or *omnibus tests*. Some common omnibus tests for multiple endpoints have been reviewed (*see* **Multiple Endpoints, Multivariate Global Tests**). These global methods provide an objective and reasonable solution to a problem with multidimensional features. They consider various endpoints simultaneously and attempt to answer the question of whether the treatments in the comparison are different in an overall sense, or whether any treatments show a more significant overall improvement than the others. However, the answer to this question is usually not the end of the investigation. Rather it leads to the next question of how the treatments affect each endpoint individually. For this case we need to test the significance of each endpoint to determine treatment effects. There are two common approaches to deal with this. Some researchers employ the approach of first using a global test as a preliminary criterion. If significance is achieved, then they apply endpoint-specific analyses to the individual endpoints. This approach is used in the hope that the global test will control the familywise type I error rate, will account for the **correlations** among the endpoints, and will increase the accuracy of identifying the effect on the individual endpoint when the endpoint-specific tests are performed. However, there is disagreement on whether this joint use of global and endpoint-specific tests can actually achieve the desired protection and power. Some researchers believe that it is redundant to perform a global test because most of the endpoint-specific procedures can control the familywise error rate and they have reasonable power to detect treatment differences among various endpoints. Their approach to the analyses is simply to use only the endpoint-specific procedures. There is no clear answer which approach is better. Since both approaches need to use the endpoint-specific procedures, it is our purpose in this article to review some of the popular endpoint-specific methods as well as the recent developments in this area of research.

Endpoint-specific procedures are intended to be used to identify the significance of the individual

endpoints. These procedures are also called P level procedures because they usually involve making statistical decisions on the basis of the P levels, resulting from examining the endpoints separately. The P level is the observed level of significance. As previously stated, these P level procedures can be used in the primary analyses or as second-step analyses in conjunction with multivariate global tests. To facilitate the presentation we adopt the same grouping scheme as described by Troendle [21, 22] and classify the endpoint-specific methods into three categories: (i) Bonferroni-type tests, (ii) normal-based tests, and (iii) resampling procedures. Within each category the methods can be further classified as single-step, step-down, or step-up tests. For the single-step tests, each comparison is based on the observed values of that particular endpoint. Both step-down and step-up tests involve comparing the ordered observed P values (or test statistics) with a set of critical constants, $c_1 \leq \cdots \leq c_K$. The testing is carried out sequentially one null hypothesis at a time. If the testing is *step-down*, then it starts with the null hypothesis corresponding to the most significant (i.e. smallest) P value (or largest test statistics) and proceeds towards the least significant null hypothesis. The testing stops when an acceptance occurs the first time. Then the remaining null hypotheses are accepted also. All null hypotheses before the first accepted null hypothesis are rejected. If the testing is *step-up*, then it starts with the null hypothesis corresponding to the least significant (i.e. largest) P value (or smallest test statistics) and proceeds towards the most significant null hypothesis. The testing stops when a rejection occurs the first time and the remaining null hypotheses are rejected also. The complexity in the computations of the critical constants usually increases in the order of single-step, step-down, and step-up analyses.

Bonferroni-Type Methods

Let $P_1, P_2, \ldots, P_K$ be the observed P values for testing the hypotheses ${H_0}^1, {H_0}^2, \ldots, {H_0}^K$. The **Bonferroni inequality** [15] sets an upper bound on the overall significance level α and produces the simplest endpoint-specific procedure that controls the familywise error rate, α. Here, the familywise error rate means the probability of rejecting incorrectly at least one true null hypothesis. For testing K endpoints, the Bonferroni procedure is performed by dividing α by the number of endpoints (i.e. K). This division of α by K is often called the *Bonferroni adjustment*. Suppose the single hypothesis on the kth endpoint is denoted by

$$ {H_0}^k : \mu_{1k} = \mu_{2k} = \cdots = \mu_{Ik}, \qquad (1) $$

where I is the number of treatments in comparison and $k = 1, \ldots, K$. The test for the kth endpoint is carried out by rejecting ${H_0}^k$ at level α if $P_k \leq \alpha/K$. This is done separately for all K endpoints. Because the test for each endpoint has an α/K probability of making a type I error, the familywise type I error is at most α. The Bonferroni method is classified as a single-step test [9].

Example

Suppose we have eight endpoints and the observed P values from the univariate tests are

$$ \begin{aligned} P_1 &= 0.061, & P_2 &= 0.255, \\ P_3 &= 0.143, & P_4 &= 0.003, \\ P_5 &= 0.048, & P_6 &= 0.001, \\ P_7 &= 0.008, & P_8 &= 0.004. \end{aligned} $$

Using Bonferroni's method, we reject ${H_0}^k$ at level $\alpha = 0.05$ if $P_k \leq 0.05/8 (= 0.0063)$ for $k = 1, \ldots, 8$. So, here we reject ${H_0}^4$, ${H_0}^6$, and ${H_0}^8$ with a 0.05 level of significance. These hypotheses correspond to the fourth, sixth, and eighth endpoints.

The advantage of the classical Bonferroni test is its computational simplicity. In addition it can also be used to construct **confidence** sets. Furthermore, it has good power in detecting real treatment effects in one or a few distinct endpoints. However, it is conservative when the alternative hypothesis for most or all measures of efficacy are uniformly improved and there are no marked differences among endpoints. In 1987, Pocock et al. [17] did a **simulation** study on the performance of the Bonferroni correction in comparing K endpoints between two treatment groups, where K ranges from two to 10 endpoints. The simulated data were based on K-variate normally distributed endpoints, each with known variance, and all K endpoints equicorrelated with correlation ρ. The results showed that the Bonferroni method works well when endpoints are independent. For each number of endpoints, the Bonferroni adjustment displayed

an elevated degree of conservatism as the correlation increases. It works reasonably well provided the pairwise correlation is less than 0.5. The authors concluded that this method works well when the endpoints are at least asymptotically normally distributed with moderate to low correlations. There is no noticeable deterioration in the Bonferroni correction as the number of correlated endpoints increases from two to 10. While this study did not consider K large, it is quite clear that the Bonferroni adjustment procedure will diminish in power if K is large [8].

The classical Bonferroni method uses a factor $1/K$ to account for the possibility of the presence of K true null hypotheses and the rejection of any one of these may cause a type I error. To improve on the power of the classical Bonferroni adjustment, Holm [10] proposes a stepwise rejection Bonferroni test. This procedure is based on the fact that if one null hypothesis is rejected, then there are only $K - 1$ possible true null hypotheses that need protection from type I error. Thus, the factor $1/K$ can be changed to $1/(K - 1)$. Owing to the sequential reduction of the denominator of the factor, Holm's procedure is conventionally known as a step-down improvement of the classical Bonferroni method. In his procedure, the univariate P values are ordered such that $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(K)}$, where $H_0{}^{(1)}, H_0{}^{(2)}, \ldots, H_0{}^{(K)}$ are the corresponding null hypotheses. The procedure starts with the smallest P value and rejects $H_0{}^{(1)}$ if the smallest P value, $P_{(1)}$, is less than α/K. Given that $H_0{}^{(1)}$ is rejected, $H_0{}^{(2)}$ will be rejected if $P_{(2)} \leq \alpha/(K - 1)$. Given that $H_0{}^{(2)}$ is rejected, $H_0{}^{(3)}$ will be rejected if $P_{(3)} \leq \alpha/(K - 2)$. Given that $H_0{}^{(j-1)}$ is rejected, $H_0{}^{(j)}$ will be rejected if $P_{(j)} \leq \alpha/(K - j + 1)$, and so forth. One proceeds in this manner until the first time a null hypothesis is accepted, then the procedure stops and the remaining null hypotheses are accepted also. Of course, acceptance of a null hypothesis here means that the null hypothesis is not rejected.

To illustrate Holm's method, we use the observed P values presented in the above example and arrange them in descending order as follows:

Unordered observed P values:

$P_1 = 0.061$, $P_2 = 0.255$,
$P_3 = 0.143$, $P_4 = 0.003$,
$P_5 = 0.048$, $P_6 = 0.001$,
$P_7 = 0.008$, $P_8 = 0.004$.

Ordered observed P values:

$P_{(1)} = 0.001$, $P_{(2)} = 0.003$,
$P_{(3)} = 0.004$, $P_{(4)} = 0.008$,
$P_{(5)} = 0.048$, $P_{(6)} = 0.061$,
$P_{(7)} = 0.143$, $P_{(8)} = 0.255$.

Using Holm's method, we start with $P_{(1)}$:

$P_{(1)} = 0.001 \leq 0.05/8$
$= 0.0063 \longrightarrow$ reject $H_0{}^{(1)}$ and continue,
$P_{(2)} = 0.003 \leq 0.05/7$
$= 0.0071 \longrightarrow$ reject $H_0{}^{(2)}$ and continue,
$P_{(3)} = 0.004 \leq 0.05/6$
$= 0.0083 \longrightarrow$ reject $H_0{}^{(3)}$ and continue,
$P_{(4)} = 0.008 \not\leq 0.05/5$
$= 0.0100 \longrightarrow$ do not reject $H_0{}^{(4)}$ and the remaining hypotheses.

So, we reject $H_0{}^{(1)}$, $H_0{}^{(2)}$, and $H_0{}^{(3)}$ at level $\alpha = 0.05$. These hypotheses correspond to the fourth, sixth, and eighth endpoints.

Although Holm's test involves a slightly more complicated computations than the classical Bonferroni test, it is strictly more powerful than the Bonferroni test, except in some trivial cases. The gain in power depends on the **alternative hypotheses**. This is small if most of the null hypotheses are true. But it may become substantial when there exist many false null hypotheses. Similar to Bonferroni adjustment, Holm's test controls the familywise error rate and it can be applied to any parametric or nonparametric model and always has good power. There are no restrictions on the type of individual tests used, the only requirement being that it is possible to calculate the observed P level for each separate test.

Holm [10] also suggested an interesting extension of his test to the case of applying different weights to different hypotheses. This is useful when there is a known hierarchy of importance among the hypotheses. As before, $P_1, P_2, \ldots, P_K$ are the observed P values for testing the null hypotheses $H_0{}^1, H_0{}^2, \ldots, H_0{}^K$. Suppose $w_1, w_2, \ldots, w_K$ are positive weights indicating the importance of the hypotheses. Greater weights indicate greater importance of the hypotheses. Let $S_k = P_k/w_k$, with $S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(K)}$. $H_0{}^{(1)}, H_0{}^{(2)}, \ldots, H_0{}^{(K)}$ are the corresponding null hypotheses and $w_{(1)}, w_{(2)}, \ldots, w_{(K)}$ are the corresponding constants. This algorithm starts with the smallest $S_{(1)}$. We reject $H_0{}^{(1)}$ if $S_{(1)} \leq$

$\alpha/(w_{(1)} + w_{(2)} + \cdots + w_{(K)})$. Then, $H_0{}^{(2)}$ is rejected if $S_{(2)} \leq \alpha/(w_{(2)} + w_{(3)} + \cdots + w_{(K)})$, and so on. If we fail to reject the null hypothesis at any step, then the procedure will be stopped and the remaining null hypotheses will not be rejected. This generalized sequential rejective test increases the power for null hypotheses with high values of w_k at the cost of decreasing the power for hypotheses with small values of w_k.

Two important step-up procedures were proposed by Hommel [11] and Hochberg [7]. Both were developed by applying the closure principle [14] to the modified Bonferroni method of Simes [20]. In 1986 Simes provided a level α test under the overall null hypothesis $H_0 = \cap\{H_0{}^k : k = 1, \ldots, K\}$ with the assumption that the endpoints are independent. In the case of investigating mean difference among I treatment groups with K endpoints, H_0 can be written as $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_I$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iK})$ is a $K \times 1$ vector consisting of the means of the K endpoints for treatment group $i = 1, 2, \ldots, I$. With Simes' procedure, H_0 is rejected if $P_{(j)} \leq j\alpha/K$ for any $j = 1, 2, \ldots, K$. He then showed by simulation that under H_0, the level does not exceed α for a variety of **multivariate normal** and **gamma** test statistics even when the statistics are correlated. In 1988, Hommel [11] extended Simes' procedure to test the significance of K individual endpoints. Hommel's procedure is performed by starting in succession with $m = 1, 2, \ldots, K$ until we find the maximum m that has $P_{(K-m+j)} \geq j\alpha/m$ for $j = 1, \ldots, m$. Suppose we find the maximum m equal to t. Then we reject all $H_0{}^k, k = 1, \ldots, K$, for which $P_k \leq \alpha/t$.

The following example is a demonstration of how to perform Hommel's method.

Unordered observed P values:

$P_1 = 0.061, \quad P_2 = 0.255,$
$P_3 = 0.143, \quad P_4 = 0.003,$
$P_5 = 0.048, \quad P_6 = 0.001,$
$P_7 = 0.008, \quad P_8 = 0.004.$

Ordered observed P values:

$P_{(1)} = 0.001, \quad P_{(2)} = 0.003,$
$P_{(3)} = 0.004, \quad P_{(4)} = 0.008,$
$P_{(5)} = 0.048, \quad P_{(6)} = 0.061,$
$P_{(7)} = 0.143, \quad P_{(8)} = 0.255,$

$m = 1$:
$j = 1\ P_{(8)} = 0.255 > 0.05 \longrightarrow$ continue,

$m = 2$:
$j = 1\ P_{(8)} = 0.255 > 0.05,$
$j = 2\ P_{(7)} = 0.143 > 0.05/2$
$= 0.0250 \longrightarrow$ continue,

$m = 3$:
$j = 1\ P_{(8)} = 0.255 > 0.05,$
$j = 2\ P_{(7)} = 0.143 > 2(0.05)/3 = 0.0333,$
$j = 3\ P_{(6)} = 0.061 > 1(0.05)/3$
$= 0.0167 \longrightarrow$ continue,

$m = 4$:
$j = 1\ P_{(8)} = 0.255 > 0.05,$
$j = 2\ P_{(7)} = 0.143 > 3(0.05)/4 = 0.0375,$
$j = 3\ P_{(6)} = 0.061 > 2(0.05)/4 = 0.0250,$
$j = 4\ P_{(5)} = 0.048 > 1(0.05)/4$
$= 0.0125 \longrightarrow$ continue,

$m = 5$:
$j = 1\ P_{(8)} = 0.255 > 0.05,$
$j = 2\ P_{(7)} = 0.143 > 4(0.05)/5 = 0.040,$
$j = 3\ P_{(6)} = 0.061 > 3(0.05)/5 = 0.030,$
$j = 4\ P_{(5)} = 0.048 > 2(0.05)/5 = 0.020,$
$j = 5\ P_{(4)} = 0.008 \not> 1(0.05)/5 = 0.010 \longrightarrow$ stop.

So, the maximum $m = 4$ and the critical level is $0.05/4 (= 0.0125)$. Comparing all the P values with this critical level, we find P_4, P_6, P_7, and P_8 less than 0.0125. Therefore, we conclude that $H_0{}^4$, $H_0{}^6$, $H_0{}^7$, and $H_0{}^8$ should be rejected at the 0.05 level of significance.

Hochberg [7] also extended Simes' procedure for making inferences on individual hypotheses. His procedure is a step-up sequential rejection procedure. It uses the same adjustment factor as Holm's procedure but Hochberg starts with the largest P value and progressively reduces the factor from 1 to $1/K$. Owing to the sequential increase of the denominator of the factor, Hochberg's approach is conventionally known as a step-up procedure. This procedure rejects the hypothesis $H_0{}^{(k)}$ and all other null hypotheses corresponding to the smaller P values (i.e. $H_0{}^{(k-1)}, H_0{}^{(k-2)}, \ldots, H_0{}^{(1)}$) if $P_{(k)} \leq \alpha/(K - k + 1)$ for $k = 1, 2, \ldots, K$. To perform this procedure we start by looking at $P_{(K)}$, the largest P value; we reject $H_0{}^{(K)}, H_0{}^{(k-1)}, \ldots, H_0{}^{(1)}$ and stop further testing if $P_{(K)} \leq \alpha$. If $P_{(K)} > \alpha$, then we do not reject $H_0{}^{(K)}$ and we consider the second largest P value and reject $H_0{}^{(K-1)}, H_0{}^{(K-2)}, \ldots, H_0{}^{(1)}$ if $P_{(K-1)} \leq \alpha/2$. The procedure continues in this manner until the

first time a hypothesis $H_0^{(k)}$ is rejected. Then we stop the testing and conclude that the hypotheses $H_0^{(k-1)}, H_0^{(k-2)}, \ldots, H_0^{(1)}$ are also rejected. Recall that $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(K)}$ so we are rejecting for the $k-1$ null hypotheses with smallest P values.

To illustrate Hochberg's method, we again use the observed P values presented in the above example.

Unordered observed P values:

$P_1 = 0.061, \quad P_2 = 0.255,$
$P_3 = 0.143, \quad P_4 = 0.003,$
$P_5 = 0.048, \quad P_6 = 0.001,$
$P_7 = 0.008, \quad P_8 = 0.004.$

Ordered observed P values:

$P_{(1)} = 0.001, \quad P_{(2)} = 0.003,$
$P_{(3)} = 0.004, \quad P_{(4)} = 0.008,$
$P_{(5)} = 0.048, \quad P_{(6)} = 0.061,$
$P_{(7)} = 0.143, \quad P_{(8)} = 0.255.$

Using Hochberg's method, we start with $P_{(8)}$.

$P_{(8)} = 0.255 > 0.05$
$\longrightarrow$ continue,
$P_{(7)} = 0.143 > 0.05/2 = 0.0250$
$\longrightarrow$ continue,
$P_{(6)} = 0.061 > 0.05/3 = 0.0167$
$\longrightarrow$ continue,
$P_{(5)} = 0.048 > 0.05/4 = 0.0125$
$\longrightarrow$ continue,
$P_{(4)} = 0.008 \leq 0.05/5 = 0.0100$
$\longrightarrow$ reject $H_0^{(4)}$ and the remaining hypotheses.

So, $H_0^{(1)}, H_0^{(2)}, H_0^{(3)}$, and $H_0^{(4)}$ are rejected at the $\alpha = 0.05$ level. These hypotheses correspond to the fourth, sixth, seventh and eighth endpoints in the original unordered hypotheses.

Both of these step-up methods can be used with different types of statistics. Hochberg's procedure is easier to apply than Hommel's. In terms of power, they are both uniformly better than Holm's step-down method and Hommel's procedure is slightly more powerful than Hochberg's procedure [12].

Both Hommel's and Hochberg's methods are able to control the familywise error rate if, under the overall null hypothesis H_0, the original Simes' test can maintain the overall type I error rate close to the nominal significance level, α [12]. It is not clear if they can control the familywise error rates when the original Simes' test does not have the α-level control under H_0.

Rom [18] proposed a method to improve on Hochberg's method. He modifies the critical points of Hochberg's procedure by integrating the joint density functions of the ordered P values and finds the new critical points by solving a recursive equation through iterations. However, the power gained from performing Rom's method is small and probably not worth the increased complexity of computations.

Hochberg & Benjamini [8] took a graphic approach to improve the power of the stepwise procedures due to Holm and Hochberg. It involves plotting the complements of the individual P values (i.e. $q_{(j)} = 1 - P_{(K-j+1)}$) vs. their order (i.e. j). Let $Q_{(j)}$ be the random variable with the observed value equal to $q_{(j)}$. The authors suggested that the set of $Q_{(j)}$s will behave as an ordered sample from a **uniform distribution** over [0, 1] if all null hypotheses are true. Assume m_0 is the number of true null hypotheses. The plot of the observed value $q_{(j)}$ is approximately linear along the line with the slope $1/(m_0+1)$ passing through the origin. The P values corresponding to the false null hypotheses are smaller than the P values corresponding to the true null hypotheses, so the $q_{(j)}$s corresponding to the false null hypotheses will be located on the right-hand side of the plot. The relationship over the left-hand side of the plot should remain approximately linear with the slope $1/(m_0+1)$. On the basis of this relationship, the authors suggested estimating m_0 by fitting an ordinary **least squares** regression line through the smallest $q_{(j)}$s located on the left-hand side of the plot. See [8] for implementation of this procedure.

Normal-Based Methods

A second group of endpoint-specific procedures focus on incorporating the covariance structure of the data into the analysis. This group of procedures requires the test statistics to be multivariate normal distributed with equal variances and a known common correlation coefficient.

Assume the test statistics are K-variate normally distributed with equal correlation ρ. Armitage & Parmar [1] presented a procedure that allows for correlations in the P value adjustment. They suggested an adjusted correction for the minimum P of the

following form:

$$P_{\text{adj}} = 1 - (1 - P_{\min})^{K^x}, \tag{2}$$

where $0 \leq x \leq 1$ and $P_{\min}$ is the minimum P value. $x = 1$ if the K test statistics are independent and $x = 0$ if the K test statistics are fully correlated. The empirical formula for computing x with an arbitrary correlation structure is given in [1]. In application, we find the adjusted P value for each endpoint by replacing $P_{\min}$ with the unadjusted P value and reject $H_0{}^k$ if the adjusted P value of the kth endpoint is less than or equal to the nominal level, α.

James [13] presents another P value adjustment method that allows for the presence of correlations. Her method is based on an approximation derived for multinormal probabilities with equal correlation. Her procedure assumes that the test statistics follow a multivariate normal distribution, or at least are asymptotically normal with equal correlation ρ. The adjusted P value for the minimum per-experiment error rate is defined as

$$\begin{aligned} P_{\text{adj}} &= \Pr(\text{minimum } P \leq p_{\min}) \\ &= 1 - \Pr(\text{all } P_k > p_{\min}) \\ &= 1 - \Pr\left(\bigcap_{k=1}^{K} a \leq Z_k \leq b\right), \end{aligned} \tag{3}$$

where $p_{\min}$ is the smallest of the per-experiment error rates and $(Z_1, Z_2, \ldots, Z_K)$ are standardized multinormal random variables with equal correlation ρ, such that

$$b = \Phi^{-1}\left(1 - \frac{P_{\min}}{2}\right), \qquad a = \Phi^{-1}\left(\frac{P_{\min}}{2}\right)$$

for the two-sided case, and

$$b = \Phi^{-1}(1 - P_{\min}), \qquad a = -\infty$$

for the one-sided case.

The approximation is given in James [13]. Using her approach, we calculate the adjusted correction for each P value and declare significance to the hypothesis corresponding to the P value if the adjusted P value is less than or equal to the nominal level of significance.

James' method improved the Bonferroni-type adjustment in the sense that it can account for the correlation structure of the data. She showed that the adjusted P values derived from her approximation are very close to the actual values obtained from a multivariate normal program [19]. They appear to be better estimates for the true P values than the approximation in [1]. The adjustment can also be used to calculate the **power** of any trial with multiple testing. However, there are some tradeoffs for this improvement in power. They include the increased complexity of calculations and also the requirement of the distributional assumption imposed on the test statistics. Further investigations need to be done to find an error bound for the approximation. Also, it is not clear what is the impact of various heterogeneous correlation structures on the performance of the method, particularly when some of the correlations are negative.

In addition to the P value adjustment methods, normal theory-based hypothesis testing procedures have also been developed. They were originally derived for the purpose of comparing multiple groups, but they can also be used to test multiple endpoints when there are only two treatment groups. Let t_k be the usual t statistic for testing $H_0{}^k$, where $k = 1, \ldots, K$. If $H_0{}^1, \ldots, H_0{}^m$ are true, then the corresponding random variables $T_1, \ldots, T_m$ have Student's m-variate central t distribution with ν df and the associated common correlation ρ (*see* **Multivariate *t* Distribution**). Let $c_m{}' = t_{1-\alpha}(m, \nu, \rho)$ be the upper α point of $\max_{1 \leq k \leq m} T_k$ for $m = 1, \ldots, K$. Bechhofer & Dunnett [2] have compiled extensive tables for the distribution of the maximum of m student t variables under various correlation structures. These tables provide the critical constants $c_k{}'$. In the usual single-step test procedure, $H_0{}^k$ is rejected if $t_k \geq c_k{}'$.

Miller [15] proposed a normal-based step-down procedure in 1966 which was further studied by Naik [16] in 1975 and Marcus et al. [14] in 1976. In 1991, Dunnett & Tamhane [4] presented step-down multiple tests in the unbalanced one-way layouts. In general, one can implement the step-down procedure by ordering the statistics as $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(K)}$ and the corresponding hypotheses as $H_0{}^{(1)}, H_0{}^{(2)}, \ldots, H_0{}^{(K)}$. $H_0{}^{(k)}$ is rejected if $H_0{}^{(j)}$ is rejected for $j = k + 1, \ldots, K$ and $t_{(k)} \geq c_k{}'$. In most cases $c_k{}'$ is easy to obtain. They are extensively tabulated for the case of equal correlations [2]. Since $c_k{}' < c_K{}'$ for $k < K$, it is obvious that the step-down procedure is uniformly more powerful than the single-step procedure.

Dunnett & Tamhane [5] introduced a normal-based step-up procedure. The step-up procedure relies

on a different set of critical constants $c_1^*, \ldots, c_K^*$. It will accept $H_0^{(k)}$ if $H_0^{(j)}$ is accepted for $j = 1, \ldots, k-1$ and $t_{(k)} < c_k^*$. The step-up procedure begins by testing the smallest t statistic and working upwards, accepting one hypothesis at a time and stopping by rejecting $H_0^{(k)}, H_0^{(k+1)}, \ldots, H_0^{(K)}$ when $t_{(k)} \geq c_k^*$. The critical constants, c_k^*, are computed recursively based on the analytic approximation of the joint distribution of the ordered P values. The computational algorithm can be found in Dunnett & Tamhane [5]. With a nonnegative correlation coefficient the authors have shown that their step-up method is empirically more powerful than Hochberg's method, and in a numerical study they also achieve higher power than Hommel's method. Although this procedure is quite powerful and also it can control the familywise error rate, it requires that the parameter estimates to be normally distributed with a common variance, which is a known multiple of an unknown σ^2, and known correlations which are equal. The computations of the critical constants turn out to be rather complicated. Also, because of the equal correlation restriction it is not suitable to be used in unbalanced data situations. To resolve this problem, Dunnett & Tamhane [6] extended the step-up procedure in [5] to include unequally correlated estimates. It has been shown by simulations that the familywise error rate of applying this procedure is at most α, and the procedure enjoys a power advantage over the step-down procedure presented in [4]. However, the computations of the critical constants for this step-up method become progressively more difficult for $m > 2$.

Resampling Procedures

Westfall & Young [23] proposed determining the multiplicity adjustments through bootstrap or permutational resampling (*see* **Bootstrap Method**) for dichotomous outcomes data (*see* **Binary Data**). The adjustments through resampling offer improvements over the Bonferroni-type adjustments by incorporating the dependence structures and other distributional characteristics into the analysis. Their distributional setup assumes that there are I treatment groups with n_i subjects per group and the data are represented by $\{\mathbf{X}_{ij}\}$, where $\mathbf{X}_{ij}$ is a vector consisting of the measurements of the K endpoints ($i = 1, \ldots, I$ and $j = 1, \ldots, n_i$). The $\mathbf{X}_{ij}$ are independently and identically distributed as multivariate Bernoulli vectors $\mathrm{MVB}_k(p_i, 1, D_i)$, where D_i denotes a particular probability distribution subject to the constraint that $E(\mathbf{X}_{ij}) = p_i$. They are motivated by the belief that false significances are most likely to occur when the overall null hypothesis H_0 is true (i.e. $p_1 = \cdots = p_I = p$ and $D_1 = \cdots = D_I = D$). Therefore, they suggested computing the adjusted P values of the original tests according to this worst-case scenario.

If one wishes an unconditional analysis, then the adjusted P values will depend on the unknown population parameters p and D. These quantities and the adjusted P values may be estimated using the data and by bootstrap resampling. To implement the procedure, the authors provided the following algorithm. The first step is to compute the unadjusted P values using the test statistics of choice. Then, generate a with-replacement sample from the data. Using only the bootstrap sample, one computes the P values with the same choice of test statistics for calculating the unadjusted P values and notes whether the minimum of the P values from the bootstrap sample is less than or equal to each of the unadjusted P values. The same algorithm will be repeated for a specified number of times and the adjusted P value for the kth endpoint is defined as the proportion of samples for which the minimum of the P values is less than or equal to the unadjusted P value for the kth endpoint.

Suppose M is denoted as the number of bootstrap samples, $P_k^{(i)}$ as the P value for the kth endpoint from the ith bootstrap sample, and $P_{k,\mathrm{unadj}}$ as the unadjusted P value for the kth endpoint observed from the original data. The formula for the adjusted P value for the kth endpoint using the bootstrap approach is given by

$$P_{k,\mathrm{adj}} = \frac{1}{M} \sum_{i=1}^{M} I\left[\min\left(P_1^{(i)}, P_2^{(i)}, \ldots, P_K^{(i)}\right) \leq P_{k,\mathrm{unadj}}\right] \tag{4}$$

where $I[A] = 1$ if event A is true and $I[A] = 0$ if event A is false.

If one is interested in a conditional analysis, then one may obtain the adjusted P values exactly using the technique of Brown & Fears [3]. They suggest reporting the permutational probability of having a P value less than or equal to a given threshold, conditional on the observed marginal frequencies. Using this method, one may obtain the adjusted

P values as the proportion of permutations of the observed vectors $\{\mathbf{X}_{ij}\}$ for which the minimum of the observed P values in a permutation sample is less than or equal to the unadjusted P values. One can calculate the adjusted P values by using the same algorithm described for the bootstrap resampling with the exception that here one generates the random permutations of the observed data and the resampling requires a without-replacement sample.

These resampling techniques are improvements over the Bonferroni-type adjustments in the sense that they are able to account for the dependence structure and discreteness of the data. They are recommended if the goal of the analysis is to isolate particular comparisons from a very large set of comparisons, particularly when many true null hypotheses are expected. The permutation approach tends to be more conservative than the bootstrap approach because the bootstrap analysis approximates the nominal significance levels more closely. As expected, the bootstrap method is more powerful than the permutation method. However, the permutation method is preferred when we cannot assert that the data are multivariate binomial distributed. In this situation the permutation method is valid provided that the subjects have been properly randomized before allocation to treatment groups. In general, the resampling techniques have reasonable power when there are marked departures from the null hypothesis at only a few endpoints. But they become conservative in the case when there are many false null hypotheses.

Westfall & Young [23] show how resampling can be used to compute adjusted P values for multivariate binomial data in a single-step manner. Both Westfall & Young [24] and Troendle [21] give a step-down improvement which does not require any distributional assumptions on the data. Westfall & Young [24] consider a more general hypothesis testing setup such that the hypotheses could come from any set of hypotheses under consideration and P values are used to perform the adjustments. Troendle's step-down resampling procedure focuses more on the hypotheses for testing multiple endpoints between two treatment groups. He argues that in the case when we decided to reject $H_0^{(K)}$ which corresponds to the largest test statistic. If $H_0^{(K)}$ is true, then we have already committed a type I error. Therefore, to control the familywise type I error rate, we should assume $H_0^{(K)}$ is false and delete the component corresponding to $t_{(K)}$ before the evaluation of the next hypothesis $H_0^{(K-1)}$. $t_{(K)}$ is the largest observed test statistic. On the basis of this principle, the following stepwise resampling algorithm is derived for testing the hypotheses. The first step of the algorithm is to order the observed test statistics and hypotheses so that $t_{(1)} \leq \cdots \leq t_{(K)}$ corresponding to $H_0^{(1)}, \ldots, H_0^{(K)}$. Suppose that M resamples, each of size $2N_0$, are used to estimate α_k, which is the significance level for $H_0^{(K-k+1)}$. Using the estimated distribution of $T_{(K-k+1)}^{(K-k+1)}$ from resampling, one can estimate

$$\alpha_k = \Pr_{H_0^{(K-k+1)}} \left[T_{(K-k+1)}^{(K-k+1)} \geq t_{(K-k+1)} \right] \tag{5}$$

where $T_{(K-k+1)}^{(K-k+1)}$ is the largest of the $(K-k+1)$ test statistics corresponding to $t_{(1)}, \ldots, t_{(K-k+1)}$. However, often there are errors in estimating α_k by an estimate based on a finite number of resamples of the data. Therefore, the following estimate is proposed:

$$\alpha_j^* = \frac{1}{M} \sum_{i=1}^{M} I \left[\max_k T_{ki}^* \geq t_{(K-j+1)} \right], \tag{6}$$

where T_{ki}^* is the test statistic of the kth endpoint corresponding to the ith resampled data and $I[A]$ is the indicator function of the event A. The maximum extends over all k corresponding to $t_{(1)}, \ldots, t_{(K-j+1)}$. We generate the kth test statistic in the ith resample where $k = 1, \ldots, K$ and $i = 1, \ldots, M$ and then calculate α_j^* as defined in (6). If $\alpha_j^* \geq \alpha$, then we stop the algorithm and accept the remaining hypotheses $H_0^{(1)}, \ldots, H_0^{(K-j+1)}$. If $\alpha_j^* < \alpha$, then we reject $H_0^{(K-j+1)}$, increment j, and repeat the resampling with the component corresponding to $t_{(K-j+1)}$ deleted.

Troendle [21] has shown in general that his step-down resampling method is asymptotically conservative (i.e. the probability that any type I error is committed is asymptotically bounded above by α). In the case when the univariate tests violate the parametric assumptions, the asymptotic conservative property of this resampling method remains. The adjusted P values obtained by the resampling method are distribution-free no matter what kind of parametric tests are used to obtain the unadjusted P values or test statistics. In other words, it is not necessary to make any parametric assumption on the distributions of the data. From the simulations, the familywise type I error rate of the method has been shown to be very close to the nominal level and the precision improves

with an increasing number of resamples. It provides increased power to reject individual hypotheses when compared with the method of Hochberg [7] except when there are many false null hypotheses. It has also been shown by simulations that this step-down resampling method provides a good approximation to the step-up method of Dunnett & Tamhane [5] when the assumptions of the step-up method are satisfied. But in the case of the presence of unequal correlations, the step-down resampling method performs better than the step-up method described in [5]. It appears that Troendle's step-down resampling method is a relatively better method, especially when the distribution or correlation structure of the data is unknown.

However, the step-down resampling method has been shown to be conservative when there are many false null hypotheses. For this aspect, a step-up permutation method was proposed by Troendle [22], since step-up methods generally have higher power than step-down methods when there are many false null hypotheses. To pursue this issue, Troendle further explored a step-up alternative to handle the multiplicity of inferences for the case of comparing two treatment groups. He introduced a step-up procedure that uses permutational resampling to estimate conditional probabilities so that one can find the critical constants for which the familywise error rate is controlled asymptotically. The form of this step-up resampling procedure takes after the step-up method for normal data proposed by Dunnett & Tamhane [5]. Suppose, for some $1 \le m \le K$, that $H_0{}^{(1)}, \ldots, H_0{}^{(m)}$ are true and $H_0{}^{(m+1)}, \ldots, H_0{}^{(K)}$ are false. The method proceeds by first testing the smallest test statistic, $t_{(1)}$. It sequentially accepts the hypotheses $H_0{}^{(1)}, H_0{}^{(2)}, \ldots$ until $t_{(k)} > c_k^{**}$ for some $k = 1, \ldots, K$. Then the procedure will be stopped and the remaining hypotheses $H_0{}^{(k)}, \ldots, H_0{}^{(K)}$ will be rejected. $\{c_k^{**}\}$ is the set of critical constants that are determined recursively such that

$$\frac{1}{M}\sum_{j=1}^{M} I\left[\left(T_{(1)}^{j(m)} > q_1\right) \cup \left(T_{(2)}^{j(m)} > q_2\right) \cup \cdots \cup \left(T_{(m)}^{j(m)} > q_m\right)\right] \le \alpha, \tag{7}$$

where M is the number of distinct permutations, $m = 1, \ldots, K$, $T_{(l)}^{j(m)}$ is defined as the lth order statistics from the components of the m smallest test statistics $T_{(1)}, \ldots, T_{(m)}$ in the jth permutation, and $I[A]$ is the indicator function for event A. The exact algorithm used to determine the critical constants is quite complex and a detailed explanation is given in [22].

This step-up method enjoys all the benefits as reflected in the step-down resampling. It requires no specific distributional assumptions of the data and it is applicable to any correlation structure. It has been shown by simulations that when the correlation is equal among the endpoints, Troendle's step-up method [22] and Dunnett et al.'s step-up method [5] have a very similar familywise type I error rate as well as power. When the correlation is unequal or unknown, Troendle's step-up permutation method is shown to be slightly more powerful than the other competing procedures through simulation. As expected, the step-up permutation method provides better power than the step-down resampling method when many false null hypotheses exist. The obvious shortcoming of this procedure is the computational complexity in the determination of the critical constants as well as the adjusted P values.

Conclusion

The advantages of the Bonferroni-type methods are simplicity in computations and applicability to data of any distribution, provided the observed P values are available. In general they can control the familywise error rate. However, they do not account for the correlation structure of the endpoints. They may become overly conservative if the tests are highly correlated. The normal-based tests incorporate the correlation structure of the endpoints. Simulations have shown some of them to be slightly more powerful than the Bonferroni-type methods, but the gain in power is quite small. They are also able to control the familywise error rate.

However, despite these advantages, they require a specific distribution of the data, the calculations of the critical constants are rather cumbersome and, so far, for the purpose of testing multiple endpoints, the normal-based tests are only applicable for the case of comparing two groups. Resampling procedures use the resampling techniques to incorporate the correlation structure of the endpoints. They require no specific distributional assumptions of the data and they are applicable to any correlation structure. These methods are shown to be able to have the probability of committing any type I error bounded above by α asymptotically. Through simulation studies, the

resampling methods achieve comparable power to the normal-based methods when the data follow the specified structure as defined in the normal-based methods. But when the distribution of the data is unknown, the resampling methods are slightly more powerful than the normal-based methods. The shortcomings of these methods are the heavy dependence on the computer and that the algorithms to determine the critical constants are quite complicated. For both the normal-based and resampling methods, the step-up procedures generally appear to have more power in detecting the alternative when all or most of hypotheses are false. The step-down procedures are more sensitive to the scenario when only one or a few hypotheses are false.

There does not seem to be an all-round winner among all the methods discussed above. It appears that power of a method is often gained at the expense of increased complexity in computations. Also, the gain in power given the added complexity seems to be very small. On the basis of the consideration of a reasonable balance between computational complexity and power, we recommend Hochberg's method, which is easy to perform and, at the same time, has comparatively good power among the Bonferroni-type methods and in comparison to the normal-based and resampling methods [5, 21, 22].

References

[1] Armitage, P. & Parmar, M. (1986). Some approaches to the problem of multiplicity in clinical trials, in *Proceedings of the XIIth International Biometrics Conference*. Biometrics Society, Seattle.

[2] Bechhofer, R.E. & Dunnett, C.W. (1988). Tables of percentage points of multivariate student *t* distributions, *Selected Tables in Mathematical Statistics* **11**, 1–371.

[3] Brown, C.C. & Fears, T.R. (1981). Exact significance levels for multiple binomial testing with application to carcinogenicity screens, *Biometrics* **37**, 763–774.

[4] Dunnett, C.W. & Tamhane, A.C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts, *Statistics in Medicine* **10**, 939–947.

[5] Dunnett, C.W. & Tamhane, A.C. (1992). A step-up multiple test procedure, *Journal of the American Statistical Association* **87**, 162–170.

[6] Dunnett, C.W. & Tamhane, A.C. (1995). Step-up multiple testing of parameters with unequality correlated estimates, *Biometrics* **51**, 217–227.

[7] Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

[8] Hochberg, Y. & Benjamini, Y. (1990). More powerful procedures for multiple significance testing, *Statistics in Medicine* **9**, 811–818.

[9] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[10] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.

[11] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test, *Biometrika* **75**, 383–386.

[12] Hommel, G. (1989). A comparison of two modified Bonferroni procedures, *Biometrika* **76**, 624–625.

[13] James, S. (1991). Approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials, *Statistics in Medicine* **10**, 1123–1135.

[14] Marcus, R., Peritz, E. & Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance, *Biometrika* **63**, 655–660.

[15] Miller, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Ed. Springer-Verlag, New York.

[16] Naik, U.D. (1975). Some selection rules for comparing *p* processes with a standard, *Communications in Statistics – Theory and Methods* **4**, 519–535.

[17] Pocock, S.J., Geller, N.L. & Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics* **43**, 487–498.

[18] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality, *Biometrika* **77**, 663–665.

[19] Schervish, M.J. (1984). Algorithm AS 195. Multivariate normal probabilities with error bound, *Applied Statistics* **33**, 81–94.

[20] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **73**, 751–754.

[21] Troendle, J.F. (1995). A stepwise resampling method of multiple hypothesis testing, *Journal of the American Statistical Association* **90**, 370–378.

[22] Troendle, J.F. (1996). A permutational step-up method of testing multiple outcomes, *Biometrics* **52**, 846–859.

[23] Westfall, P.H. & Young, S.S. (1989). *P* value adjustments for multiple tests in multivariate binomial models, *Journal of the American Statistical Association* **84**, 780–786.

[24] Westfall, P.H. & Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, Vol. 1. Wiley, New York.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Multiple Imputation Methods

Missing data occur frequently in biomedical studies. For example, they occur in a survey on people's health (*see* **Surveys, Health and Morbidity**) when some people do not respond to all of the survey questions. Another example is a medical experiment (*see* **Clinical Trials, Overview**) in which some follow-up visits are skipped or some measurements are accidentally not taken.

A common technique for handling missing data is to impute, i.e. fill in, a value for each missing datum. This results in a completed data set, so that standard methods that have been developed for analyzing complete data can be applied immediately. Thus, imputing for missing values followed by using a standard complete-data method of analysis is typically easier than creating specialized techniques to analyze the incomplete data directly. Simplicity of subsequent analyses is an important practical advantage of imputation.

Imputation has other advantages in the context of the production of a data set for general use, such as a public-use file from a health survey. One such additional advantage is that the data producer can use specialized knowledge about the reasons for missing data, including confidential information that cannot be released to the public, to create the imputations. In addition, imputation by the data producer fixes the missing data problem in the same way for all users, so that consistency of analyses across users is ensured. When imputation is not carried out by the data producer, so that each user implements some method for handling missing data, the knowledge of the data producer can fail to be incorporated, analyses are not typically consistent across users, and all users expend resources addressing the missing-data problem.

Although imputation satisfies critical data-processing objectives and can incorporate knowledge from the data producer, single imputation, i.e. imputing one value for each missing datum, fails to satisfy statistical objectives concerning the validity of the resulting **inferences** based on the completed data. Specifically, for validity, the resulting estimates based on the data completed by imputation should be approximately **unbiased** for their population estimates, **confidence intervals** should attain at least their nominal coverages, and tests of **null hypotheses** should not reject true null hypotheses more frequently than their nominal levels. Because a single imputed value cannot reflect any of the uncertainty about the true underlying value, analyses that treat imputed values just like observed values systematically underestimate uncertainty. Thus, imputing a single value for each missing datum and then analyzing the completed data using standard techniques designed for complete data will result in **standard error** estimates that are too small, confidence intervals that fail to attain their nominal coverages, and ***P* values** that are too significant (*see* **Hypothesis Testing**); this is true even if the modeling for imputation is carried out carefully. For example, **large-sample** results in [24] show that for simple situations with 30% of the data missing, single imputation under the correct model followed by the standard complete-data analysis results in nominal 90% confidence intervals having actual coverages below 80%. The inaccuracy of nominal levels is even more extreme in multiparameter problems [22, Chapter 4], where nominal 5% tests can easily have rejection rates of 50% or more when the null hypothesis is true.

For particular estimates in certain situations, techniques have been developed that enable the data analyst to obtain correct estimates of variability from singly-imputed data [17, 31]. These techniques, however, require the data analyst to use nonstandard procedures that must be specially developed for each combination of imputation method and analysis, and they sometimes require the data producer to provide extra information to the data analyst. Thus, in solving one problem associated with single imputation, they lose the critical advantages inherent to imputation.

Multiple imputation [19, 22] is an approach that retains the advantages of single imputation while allowing the data analyst to obtain valid assessments of uncertainty. The basic idea is to impute two or more times for the missing data using independent draws of the missing values from a distribution that is appropriate under the posited assumptions about the data and the mechanism creating missing data. This results in two or more completed data sets, each of which is analyzed using the same standard complete-data method. The analyses are then combined in a simple way that reflects the extra uncertainty due to having imputed rather than actual

data. Multiple imputations can also be created under several different models to display sensitivity to the choice of missing-data model. Recent reviews of work on multiple imputation are given in [23] and [26].

Theoretical Motivation for Multiple Imputation

The theoretical motivation for multiple imputation is **Bayesian**, although the procedure has excellent properties from a frequentist perspective. Examples of publications containing information on the properties of multiple imputation are [5], [9], [11–13], [22, Chapter 4], [24], and [25]. More extensive references can be found in [23] and [30].

Formally, let Q be the population quantity of interest, and suppose the data can be partitioned into observed values, $\mathbf{X}_{\text{obs}}$, and missing values, $\mathbf{X}_{\text{mis}}$. If $\mathbf{X}_{\text{mis}}$ had been observed, then inferences for Q would have been based on the complete-data posterior density $p(Q|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$. Because $\mathbf{X}_{\text{mis}}$ is not observed, inferences are based on the actual posterior density $p(Q|\mathbf{X}_{\text{obs}})$, which can be expressed as

$$p(Q|\mathbf{X}_{\text{obs}}) = \int p(Q|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}) p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})\, \mathrm{d}\mathbf{X}_{\text{mis}}. \tag{1}$$

Eq. (1) shows that the actual posterior density of Q can be obtained by averaging the complete-data posterior density over the posterior predictive distribution of $\mathbf{X}_{\text{mis}}$. In principle, multiple imputations are repeated independent draws from $p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$. Thus, multiple imputation allows the data analyst to approximate (1) by separately analyzing each data set completed by imputation and then combining the results of the separate analyses.

Analyzing a Multiply Imputed Data Set

The exact computation of the posterior distribution (1) by simulation would require that an infinite number of values of $\mathbf{X}_{\text{mis}}$ be drawn from $p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$. This section summarizes simple approximations to (1) that can be used when only a small number of imputations of $\mathbf{X}_{\text{mis}}$ have been drawn. Fortunately, as indicated earlier, these approximations work very well in most practical situations.

Inferences for Scalar Q

Suppose that if the data were complete, inferences for Q would be based on a point estimate $\hat{Q}$, an associated variance estimate $\hat{U}$, and a **normal** reference distribution. When data are missing and there are M sets of imputations for the missing data, the result is M sets of complete-data statistics, say $\hat{Q}_m$ and $\hat{U}_m, m = 1, \ldots, M$.

Rubin & Schenker [24] suggested the following procedure for drawing inferences about Q from the multiply imputed data. The point estimate of Q is the average of the M completed-data estimates,

$$\overline{Q} = \sum_{m=1}^{M} \frac{\hat{Q}_m}{M},$$

and the associated **variance** estimate is

$$T = \overline{U} + (1 + M^{-1})B,$$

where $\overline{U} = \sum_{m=1}^{M} \hat{U}_m/M$ is the average within-imputation variance, and $B = \sum_{m=1}^{M} (\hat{Q}_m - \overline{Q})^2/(M-1)$ is the between-imputation variance. The approximate reference distribution for interval estimates and significance tests is a t distribution (*see* **Student's *t* Distribution**) with degrees of freedom

$$\nu = (M-1)(1 + r^{-1})^2,$$

where $r = (1 + M^{-1})B/\overline{U}$ is the estimated ratio of the between-imputation component of variance to the within-imputation component of variance.

Significance Tests for Multicomponent Q

Consider an estimand $\mathbf{Q}$ with $k > 1$ components, and suppose the goal is to obtain a significance level for a null value of $\mathbf{Q}$, say $\mathbf{Q}_0$. **Multivariate** analogs of the expressions given in the previous section for scalar Q have been derived by Rubin [22, Section 3.4] and Li et al. [13] for the situation in which the complete-data analysis that is applied to each completed data set produces both a point estimate for $\mathbf{Q}$ and an associated variance estimate. Meng & Rubin [16] developed methods for **likelihood-ratio testing** when the available information consists of point estimates and evaluations of the complete-data log likelihood ratio statistic as a function of these estimates and the completed data. Asymptotically, the procedures of Li

et al. [13] and the procedures of Meng & Rubin [16] are equally accurate.

With large data sets and large models, such as in the situation of a multiway **contingency table**, the complete-data analysis might produce only a test statistic and no estimates, unlike the situations considered by Rubin [22, Section 3.4], Li et al. [13], and Meng & Rubin [16]. With such limited information, Rubin [22, Section 3.5] provided initial methods and Li et al. [12] developed improved methods that require only the M complete-data **chi-square** statistics (or equivalently the M complete-data P values) that result from testing a null hypothesis about Q using each of the M completed data sets. These methods are less accurate than methods that use the completed-data estimates.

Creating Multiple Imputations

Ideally, multiple imputations are M independent random draws from the posterior predictive distribution of $\mathbf{X}_{\text{mis}}$ under appropriate Bayesian modeling assumptions. Such imputations are called *repeated imputations* in Rubin [22, Chapter 3]. In practice, approximations are often used and work well.

Modeling Issues

Several important issues arise in the creation of imputation models. These include the criticality of predictive models, explicit vs. implicit models, and ignorable vs. nonignorable models.

An initial modeling issue in creating imputations (see, for example, Little's discussion [14] of single and multiple imputation) is that the predictive distribution for the missing values should be conditional on all observed values. This consideration is particularly important in the context of a public-use data base. Omitting a variable from the imputation model is equivalent to assuming that the variable is not associated with the variables being imputed, at least conditionally given all other variables in the model; imputing under this assumption can result in **biases** in subsequent analyses, with estimated parameters representing conditional association pulled toward zero. Because it is not known which analyses will be carried out by subsequent users of a public-use data base, ideally it is best not to omit any variables from the imputation model. It is usually infeasible, of course, to incorporate every available variable, including **interactions**, into an imputation model, but it is desirable to condition imputations on as many variables as possible and to use subject-matter knowledge to help select those variables that are likely to be used together in subsequent complete-data analyses.

Imputation procedures can be based on explicit models or implicit models, or even combinations (see, for example, [22, Chapter 5] and [26]). An example of a procedure based on an explicit model is stochastic normal **regression** imputation, where imputations for missing values are created by adding normally distributed errors to predicted values obtained from a **least-squares** regression fit. A common type of procedure based on implicit models is hot deck imputation, which replaces the missing values for an incomplete case by the values from a matching complete case, where the matching is carried out with respect to variables that are observed for both the incomplete case and complete cases (*see* **Missing Data Estimation, "Hot Deck" and "Cold Deck"**).

Rather than attempting to match cases exactly in hot deck imputation, sometimes it is useful to define a distance function on the basis of variables that are observed for both complete and incomplete cases and then to impute values for each incomplete case from a complete case that is close with respect to this distance. In practice, the function may not be a mathematical "distance" (i.e. it can be zero without all components exactly matching). When the distance function is the absolute value of the difference between the nonrespondent's and respondent's predicted values of the variables to be imputed, the matching procedure is termed *predictive mean matching* [9, 14, 21, 33]. Hot deck imputation using predictive mean matching on the basis of an explicit prediction model (e.g. normal linear regression) is an example of an imputation procedure that combines aspects of both an implicit method and an explicit method.

The model underlying an imputation procedure, whether explicit or implicit, can be based on the assumption that the reasons for missing data are either ignorable or nonignorable [18]. The distinction between an ignorable and a nonignorable model can be illustrated by a simple example with two variables, X and Y, where X is observed for all cases, whereas Y is sometimes missing. Ignorable models assert that a case with Y missing is only

randomly different from a complete case having the same value of X. Nonignorable models assert that there are systematic differences between an incomplete case and a complete case even if they have identical X values. An important issue with nonignorable models is that because the missing values cannot be observed, there is no direct evidence in the data to address the assumption of nonignorability. It can be important, therefore, to consider several alternative models and to explore a **sensitivity analysis** of resulting inferences to the choice of model. In current practice, almost all imputation models are ignorable; limited experience suggests that in major surveys with limited amounts of missing data and careful design, ignorable models are satisfactory for most analyses (see, for example, [28]).

Incorporating Proper Variability

Multiple imputation procedures that incorporate appropriate variability across the M sets of imputations within a model are called *proper* in [22, Chapter 4], where precise conditions for a method to be proper are also given. Rubin [23] provides more intuitive statements of the conditions, and Meng [15] discusses closely related issues. Because, by definition, proper methods reflect sampling variability correctly, inferences based on the multiply-imputed data are valid from the standard repeated-sampling (i.e. design-based) frequentist perspective.

One important principle related to incorporating appropriate variability is that imputations should be random draws rather than best predictions. Imputing best predictions can lead to distorted estimates of quantities that are not linear in the data, such as measures of variability and **correlation**, and it generally results in severe underestimation of uncertainty and therefore invalid inferences.

For imputations to be proper, the variability due to estimating the model must be reflected along with the variability of data values given the estimated model. For this purpose, a two-stage procedure is often useful, as we now explain. Suppose that a Bayesian predictive distribution for $\mathbf{X}_{\text{mis}}$ has been formulated using a parameter β. Then the posterior predictive density, which appears on the right-hand side of (1), can be expressed as

$$p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}) = \int p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \beta)\, p(\beta|\mathbf{X}_{\text{obs}})\, \mathrm{d}\beta, \tag{2}$$

where $p(\beta|\mathbf{X}_{\text{obs}})$ is the posterior distribution of β, and $p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \beta)$ is derived from a parametric model for the data (e.g. normal linear regression, **loglinear**). It can be seen from (2) that a draw of a value of $\mathbf{X}_{\text{mis}}$ from its posterior predictive distribution can be obtained by first drawing a value of β from its posterior distribution and then drawing a value of $\mathbf{X}_{\text{mis}}$ conditional on the drawn value of β. Fixing β at a point estimate (e.g. the **maximum likelihood** estimate), say $\hat{\beta}$, across the M imputations and drawing $\mathbf{X}_{\text{mis}}$ from $p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \hat{\beta})$, generally leads to inferences based on the multiply imputed data that are too sharp, as shown, for example, in [22, Chapter 4], [24], and [25].

The two-stage paradigm can be followed in the context of **nonparametric methods** such as hot deck imputation as well as in the context of parametric models with formal posterior distributions for β. The simple hot deck procedure that randomly draws imputations for incomplete cases from matching-complete cases is not proper because it ignores sampling variability owing to the fact that the population distribution of complete cases is not known but rather is estimated from the complete cases in the sample. Rubin & Schenker [24, 26] discuss the use of the **bootstrap** [4] to make the hot deck procedure proper, and call the resulting procedure the *approximate Bayesian bootstrap*, since it approximates the Bayesian bootstrap [20]. The two-stage procedure first draws a bootstrap sample from the complete cases and then draws imputations randomly from the bootstrap sample. Thus, bootstrap sampling from the complete cases before drawing imputations is a nonparametric analog of drawing values of the parameters of the imputation model from their posterior distribution before imputing conditionally upon the drawn parameter values. The bootstrap has also been used in conjunction with parametric models in multiple imputation in an effort to produce imputations that reflect the variability due to estimating the parametric models [3, 9].

Choice of M

Another issue to be discussed when multiple imputations are to be created is the choice of M. This choice involves a trade-off between **simulating** more accurately the posterior distribution (2), as is possible with larger values of M, vs. using a smaller

amount of computing and storage, as occurs with smaller values of M. The effect of the value of M on accuracy depends on the fraction of information about the estimand Q that is missing, a quantity defined in [22, Chapters 3 and 4]. With ignorable missing data and just one variable, the fraction of missing information is simply the fraction of data values that are missing. When there are several variables, however, the fraction of missing information is often smaller than the fraction of cases that are incomplete because of the ability to predict missing values from observed values.

For the moderate fractions of missing information (<30%) that occur with most analyses of data from most large surveys, Rubin [22, Chapter 4] showed that a small number of imputations (say, $M = 3$ or 4) results in nearly fully **efficient estimates** of Q. In addition, Rubin & Schenker [24], Rubin [22, Chapter 4], and Li et al. [13] have shown that if proper multiple imputations are created, then the resulting inferences generally have close to their nominal coverages or significance levels, even when the number of imputations is moderate. A substantial body of work [2, 9, 24, 34], and additional publications cited in [23], supports these claims in practical cases.

Use of Iterative Simulation Techniques

Recent developments in iterative simulation, such as data augmentation [35] and Gibbs sampling [6, 7] (*see* **Markov Chain Monte Carlo**), can facilitate the creation of multiple imputations in complicated parametric models. Consider a joint model for $\mathbf{X}_{\text{obs}}$ and $\mathbf{X}_{\text{mis}}$ governed by a parameter, say θ. The data augmentation (Gibbs sampling) procedure that results in draws from the posterior distribution of θ produces multiple imputations as well. Let $\theta^{(t)}$ and $\mathbf{X}_{\text{mis}}^{(t)}$ denote the draws of θ and $\mathbf{X}_{\text{mis}}$ at iteration t in the Gibbs sample. At iteration $t+1$, a value $\theta^{(t+1)}$ is drawn from $p(\theta|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}^{(t)})$ and then a value $\mathbf{X}_{\text{mis}}^{(t+1)}$ is drawn from $p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \theta^{(t+1)})$. As t approaches infinity, $(\theta^{(t)}, \mathbf{X}_{\text{mis}}^{(t)})$ converges to a draw from $p(\theta, \mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$. Thus, for large t, $\mathbf{X}_{\text{mis}}^{(t)}$ is close to a draw from $p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$ and can be used as an imputation of $\mathbf{X}_{\text{mis}}$ in a multiple-imputation scheme. Schafer [30] developed **algorithms** that use iterative simulation techniques to multiply imputed data when there are arbitrary patterns of missing data and the missing-data mechanism is ignorable. Such techniques are being considered for use in the 2000 **census** [29] and have been used in other contexts, such as in **National Center for Health Statistics** surveys [32].

Some Recent Applications of Multiple Imputation in Health Care Research

To illustrate settings in which multiple imputation can be useful, brief descriptions of several recent applications of multiple imputation to **health services research** are now given. References to many other examples are given in [23].

Research on Health and Nutrition

Schafer et al. [32] used multiple imputation to handle incomplete observations in the National Health and Nutrition Examination Survey (NHANES) III. Each NHANES is a national survey conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of the US population and important subgroups. The data from NHANES were obtained through household interviews and through standardized physical examinations. The NHANES imputation project is significant because it demonstrates the feasibility of generating proper multiple imputations for a public-use sample survey with many observations and variables, high rates of missingness on several key variables of interest, and various patterns of **nonresponse**.

NHANES III and other NHANES were used in a simulation study conducted by Ezzati-Rice et al. [5] to evaluate the frequentist performance of model-based multiple imputations in NCHS health examination surveys. The simulations were based on a hypothetical population constructed from previous data sets to resemble populations surveyed by NHANES. The hypothetical population contained 17 variables, with both categorical and continuous variables.

From the hypothetical population, 1000 samples, each with missing values and five imputations of the missing values, were generated by a three-step process: (i) **stratified samples** were drawn to mimic some of the characteristics of NHANES sampling designs; (ii) for each generated sample, missing data

patterns were imposed on the sample using an ignorable missing data mechanism that utilized missing data patterns observed in NHANES III; (iii) for each incomplete sample, five imputations of the missing data were generated under a general location model for the complete data, which is popular when there are both categorical and continuous variables.

Ezzati-Rice et al. [5] studied the validity of multiple-imputation interval estimates for population and subdomain means and proportions. Their results indicated that the imputation model was successful in creating valid design-based repeated-sampling inferences. In other words, the interval estimates had at least nominal coverage for a wide range of estimands.

Research on AIDS

Taylor et al. [36] applied multiple imputation in a project whose goal was to estimate the distribution of times from human immunodeficiency virus (HIV) seroconversion to the onset of acquired immune deficiency syndrome (**AIDS**) from a four-year, multicentered **cohort study** of homosexual and bisexual men in the US. The subjects in the study were divided into two cohorts: (i) those men who were already infected with HIV when enrolled in the study (the "seropositives"); and (ii) those men who became infected during the follow-up period (the "seroconverters"). The seropositive cohort presents the difficulty that the times of seroconversion are unknown for its members, whereas the seroconverter cohort presents the difficulty that the dates of diagnosis with AIDS are unknown for most of its members owing to the 4-year follow-up time.

To alleviate these difficulties, Taylor et al. [36] multiply imputed the date of diagnosis with AIDS for the seroconverters who were AIDS-free during the follow-up period, using a failure time regression model (*see* **Survival Analysis, Overview**) that was estimated from the data for the seropositive cohort. The **covariates** in the regression model were chosen so that the time to AIDS diagnosis, given the covariates, could be considered nearly independent of the time since HIV infection. Each data set for the seroconverters that was completed by imputation was analyzed using **Kaplan–Meier estimation** and Greenwood's formula. The completed-data estimates were then combined to obtain an estimate of the distribution of times from HIV seroconversion to AIDS diagnosis. In this example, multiple imputation helped to "extend" the study period, so that estimates and assessments of variability were possible for longer follow-up times.

Research on Malnutrition

In studies of malnutrition, it is of interest to estimate the percentage of children who are short for their ages ("stunted") as well as the percentage of children who are light for their heights ("wasted"), where the comparison is made with a sample of normal, healthy children. Such estimation is difficult when the ages of the children in a study are reported with accuracy only to the nearest year or half-year; this inaccurate age reporting is called *age heaping*.

Heitjan & Rubin [10] considered a data set for a sample of children under 6 years of age from the Dodoma region of Tanzania. To deal with the apparently large amount of age heaping in this data set, Heitjan & Rubin [10] multiply imputed the true ages of the children. Two methods were explored for purposes of sensitivity analysis: (i) a simple procedure in which a child's age was imputed uniformly from an interval determined from the reported age; and (ii) a complex procedure in which the age-heaping process and the true age were modeled simultaneously using the sex, body measurements, and reported age of the child. The resulting estimates displayed some sensitivity to assumptions about whether the ages were rounded or truncated as well as assumptions about the width of the interval within which the ages were heaped to one reported age. The multiple-imputation inferences were, however, more similar to each other than to the inferences obtained using any method based on single imputation, which cannot reflect the proper uncertainty about the missing true ages.

Research on Hip Replacement

Dorey et al. [3] considered data on patients at UCLA Medical Center who had received a total hip arthroplasty. The routine follow-up of each such patient includes evaluations of radiographs to determine whether the prosthesis is loosening. When certain measurements from the radiographs pass a prespecified threshold, the prosthesis is considered to be at increased risk of loosening. A patient who is past this threshold at a follow-up visit frequently did not reach

this threshold by the previous visit. Thus, the time at which the threshold has been crossed is known only to be between two time points, a phenomenon known as *interval censoring*.

A standard practice for dealing with interval censoring has been to set each unknown value of the threshold-crossing time equal to the right endpoint of the known interval that contains it. Dorey et al. [3] compared this practice with approaches that multiply impute the threshold-crossing time within the interval, including a method that uses the radiographic measurements at the endpoints of the interval to predict when during the interval the threshold had been crossed. It was found that subsequent analyses were sensitive to whether imputations were created by the standard practice, or from a distribution over the interval.

Research on Drinking Behavior

Glynn et al. [8] examined data from the Normative Aging Study, a longitudinal study of community-dwelling men conducted by the US Department of Veterans affairs in Boston. Part of the study was a survey on drinking behavior. A large fraction of the men surveyed provided essentially complete information, whereas for a smaller fraction there was background information available but no information on drinking behavior. Because drinking behavior is a sensitive subject, there was concern that such nonresponse might be nonignorable.

A subsequent survey collected information on drinking behavior from about one-third of the prior nonrespondents. Glynn et al. [8] used this information to multiply impute data on drinking for the remaining nonrespondents, under the assumption that given this new information as well as the background information that had been collected previously, nonresponse was now ignorable. It was found that inferences about the effect of retirement status on drinking behavior (adjusting for age) were very sensitive to whether the multiply-imputed data were used, or, rather, only the data for the initial respondents to the survey were used.

Available Software for Multiple Imputation

There is currently only a limited amount of software for generating multiple imputations under multivariate complete-data models and for analyzing multiply imputed data sets (i.e. completing the data sets, running the complete-data analyses, and combining the complete-data outputs), but the situation appears to be improving rapidly. For generating imputations, software to implement the methodology developed in [30] has been written for the **S-PLUS** statistical package and is freely available on the **internet**. This software, which will also be expanded and incorporated into S-PLUS as a commercial add-on module to the base S-PLUS software, includes programs for multiple imputation in the contexts of incomplete **multivariate normal** data, incomplete categorical data, and incomplete data under the general location model allowing both categorical and multivariate normal variables. These programs have been used to generate imputations for NHANES III [32] and for several other multiple imputation projects. For analyzing multiply imputed data sets, a suite of programs for the Stata statistical package has been developed by J. Barnard and will be freely available on the Internet.

More software for generating multiple imputations and for analyzing multiply imputed data sets should be shortly available. Several packages, both commercial and freeware, are currently under development for generating imputations, e.g. M by J. Barnard, C. Liu, and D.B. Rubin and software within HERMES [1]. A project is also under way to develop multiple-imputation analysis software for SAS (*see* **Software, Biostatistical**).

Acknowledgment

This article is a modification and expansion of [27]. The work was supported in part by grants CA 64235 from the National Cancer Institute and DMS-970 5158 from the National Science Foundation.

References

[1] Brand, J., van Buuren, S., van Mulligen, E.M., Timmers, T. & Gelsema, E. (1994). Multiple imputation as a missing data machine, in *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care (SCAMC)*. Hanley & Beflus, Philadelphia, pp. 303–307.

[2] Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B. & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using

Bayesian logistic regression, *Journal of the American Statistical Association* **86**, 68–78.

[3] Dorey, F.J., Little, R.J.A. & Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval-censoring, *Statistics in Medicine* **12**, 1589–1603.

[4] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**, 1–26.

[5] Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B. & Schafer, J.L. (1995). A simulation study to evaluate the performance of multiple imputation in NCHS Health Examination Survey, in *Bureau of the Census Proceedings of the 1995 Annual Research Conference*. US Bureau of the Census, Washington, pp. 257–266.

[6] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 972–985.

[7] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[8] Glynn, R., Laird, N. & Rubin, D.B. (1993). The performance of mixture models for nonignorable nonresponse with follow ups, *Journal of the American Statistical Association* **88**, 984–993.

[9] Heitjan, D.F. & Little, R.J.A. (1991). Multiple imputation for the fatal accident reporting system, *Applied Statistics* **40**, 13–29.

[10] Heitjan, D.F. & Rubin, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping, *Journal of the American Statistical Association* **85**, 304–314.

[11] Herzog, T.N. & Rubin, D.B. (1983). Using multiple imputations to handle non-response in sample surveys, in *Incomplete Data in Sample Surveys*, Vol. 2: *Theory and Bibliographies*, W.G. Madow, I. Olkin & D.B. Rubin, eds. Academic Press, New York, pp. 209–245.

[12] Li, K.H., Meng, X.L., Raghunathan, T.E. & Rubin, D.B. (1991). Significance levels from repeated p values with multiply-imputed data, *Statistica Sinica* **1**, 65–92.

[13] Li, K.H., Raghunathan, T.E. & Rubin, D.B. (1991). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution, *Journal of the American Statistical Association* **86**, 1065–1073.

[14] Little, R.J.A. (1988). Missing data in large surveys (with discussion), *Journal of Business and Economic Statistics* **6**, 287–301.

[15] Meng, X.L. (1994). Multiple imputation with uncongenial sources of input (with discussion), *Statistical Science* **9**, 538–573.

[16] Meng, X.L. & Rubin, D.B. (1992). Performing likelihood ratio tests with multiply imputed data sets, *Biometrika* **79**, 103–111.

[17] Rao, J.N.K. & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika* **79**, 811–822.

[18] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

[19] Rubin, D.B. (1978). Multiple imputations in sample survey-a phenomenological Bayesian approach to nonresponse, in *American Statistical Association 1978 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 20–34.

[20] Rubin, D.B. (1981). The Bayesian bootstrap, *Annals of Statistics* **9**, 130–134.

[21] Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputation, *Journal of Business and Economic Statistics* **4**, 87–94.

[22] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[23] Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**, 473–489.

[24] Rubin, D.B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *Journal of the American Statistical Association* **81**, 366–374.

[25] Rubin, D.B. & Schenker, N. (1987). Interval estimation from multiple imputed data: a case study using agriculture industry codes, *Journal of Official Statistics* **3**, 375–387.

[26] Rubin, D.B. & Schenker, N. (1991). Multiple imputation in health-care data bases: an overview and some applications, *Statistics in Medicine* **10**, 585–598.

[27] Rubin, D.B. & Schenker, N. (1997). Imputation, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz, C. Read, & D. Banks, eds. Wiley, New York.

[28] Rubin, D.B., Stern, H.S. & Vehovar, V. (1995). Handling "don't know" survey responses: the case of the Slovenian plebiscite, *Journal of the American Statistical Association* **90**, 822–826.

[29] Schafer, J.L. (1995). Model-based imputation of census short-form items, in *Bureau of the Census Proceedings of the 1995 Annual Research Conference*. US Bureau of the Census, Washington, pp. 267–299.

[30] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

[31] Schafer, J.L. & Schenker, N. (1991). Variance estimation with imputed Means, in *American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 696–701.

[32] Schafer, J.L., Khare, M. & Ezzati-Rice, T.M. (1993). Multiple imputation of missing data in NHANES III, in *Bureau of the Census Proceedings of the 1995 Annual Research Conference*. US Bureau of the Census, Washington, pp. 459–487.

[33] Schenker, N. & Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation, *Computational Statistics & Data Analysis* **22**, 425–448.

[34] Schenker, N., Treiman, D.J. & Weidman, L. (1993). Analysis of public-use data with multiply-imputed

industry and occupation codes, *Applied Statistics* **42**, 545–556.

[35] Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association* **82**, 528–550.

[36] Taylor, J.M.G., Muñoz, A., Bass, S.M., Chmiel, J.S., Kingsley, L.A. & Saab, A.J. (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation, *Statistics in Medicine* **9**, 505–514.

JOHN BARNARD, DONALD B. RUBIN & NATHANIEL SCHENKER

Multiple Linear Regression

Multiple linear regression represents a generalization, to more than one explanatory variable, of the method of analysis known as **simple linear regression**. The term **"regression"** was first introduced by **Galton** [5, p. 246], who used it to characterize a tendency towards mediocrity (i.e. more average) observed in the offspring of parent seeds. Pearson & Lee [7] also described the relationship between the heights of father–son pairs as a regression, concluding that the estimated slope of 0.516 provided confirmation of Galton's "law of universal regression".

Today, Galton's meaning of the term regression is largely forgotten. In current usage, regression refers to the body of statistical methods used to characterize, quantitatively, the relationship between a **response** (dependent, outcome) variable, Y, which varies randomly, and one or more **explanatory** (independent, predictor) variables, $X_1, \ldots, X_k$. The adjective "multiple" indicates that, initially, at least two explanatory variables are involved in the modeling exercise. The values of the explanatory variables are assumed to be known, or under the control of an investigator. However, in many applications the observed values of $X_1, \ldots, X_k$ are also unknown, and may be observed concurrently with Y. We assume that any inherent variation in the measurement of a particular X_j, $j = 1, \ldots, k$, can be ignored. If this is not the case, we recommend the use of methods that are appropriate when there is measurement error in one or more explanatory variables (*see* **Errors in Variables**). The individual explanatory variables may be nominal (e.g. gender), categorical (e.g. different types of disease), ordered categorical (e.g. tumor grade), or interval or continuous (e.g. forced expiratory volumes) (*see* **Measurement Scale**). Multiple linear regression involves finding the best-fitting surface that relates $\mathrm{E}(Y|X_1, \ldots, X_k)$, the mean value of Y given values of $X_1, \ldots, X_k$, and $X_1, \ldots, X_k$, using an equation with a suitable functional form, such as

$$\mathrm{E}(Y|X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k. \quad (1)$$

As written, (1) characterizes the mean value, $\mathrm{E}(Y|X_1, \ldots, X_k)$, as a linear (planar) function of $X_1, \ldots, X_k$. In a sense, this may be an artifice of notation, since X_1 could represent the square root of weight in kg, i.e. $X_1 = \sqrt{W}$. Likewise, Y could be the logarithm of systolic blood pressure (SBP) in mm Hg, i.e. $Y = \log(\mathrm{SBP})$, where SBP is the outcome variable originally measured. Thus, the *linear* aspect of regression does not indicate that the model linking $\mathrm{E}(Y|X_1, \ldots, X_k)$ and $X_1, \ldots, X_k$ is necessarily a straight-line (linear) function of the explanatory variables. Rather, the average response, $\mathrm{E}(Y|X_1, \ldots, X_k)$, is a linear function of the $k + 1$ unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$. The model equation $\mathrm{E}(Y|Z) = \beta_0 + \beta_1 Z + \beta_2 Z^2$ is a special case of (1) with $X_1 = Z$ and $X_2 = Z^2$, and represents the multiple linear regression of Y on Z involving a quadratic dependence between the average response, $\mathrm{E}(Y|Z)$, and the explanatory variable, Z.

Whatever the functional form of the regression model, the objectives of linear regression are:

1. to determine whether Y and one or more of the explanatory variables are associated in some systematic way, and/or
2. to estimate or predict the value of Y, or its mean, corresponding to known values of a selected subset of $X_1, \ldots, X_k$.

The analysis is based on estimates of the $k + 1$ unknown parameters, $\beta_0, \beta_1, \ldots, \beta_k$, and their statistical properties when certain assumptions concerning the randomly varying responses, Y, are thought to be valid. The estimates of these parameters, which are known as regression coefficients, are derived from data – $n(k + 1)$-tuples $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ – using the classical method of **least squares**. However, before fitting a linear regression model to data, it is wise to examine individual scatterplots of Y vs. $X_1, \ldots, X_k$ to ensure that the functional form of the proposed relationship is sensible (*see* **Graphical Displays**).

Figure 1 shows six such scatterplots for measurements of oxygen uptake, Y, in milliliters per kilogram of body weight per minute vs. age, X_1, in years, weight, X_2, in kilograms, time, X_3, in minutes required to run 1.5 miles, resting pulse, X_4, in beats per minute, running pulse, X_5, in beats per minute, and maximum pulse, X_6, recorded while running, in beats per minute. The measurements were obtained from 31 subjects – 21 men ($X_7 = 1$) and 10 women ($X_7 = 0$) – who participated in a physical fitness workshop. For these data the notion that

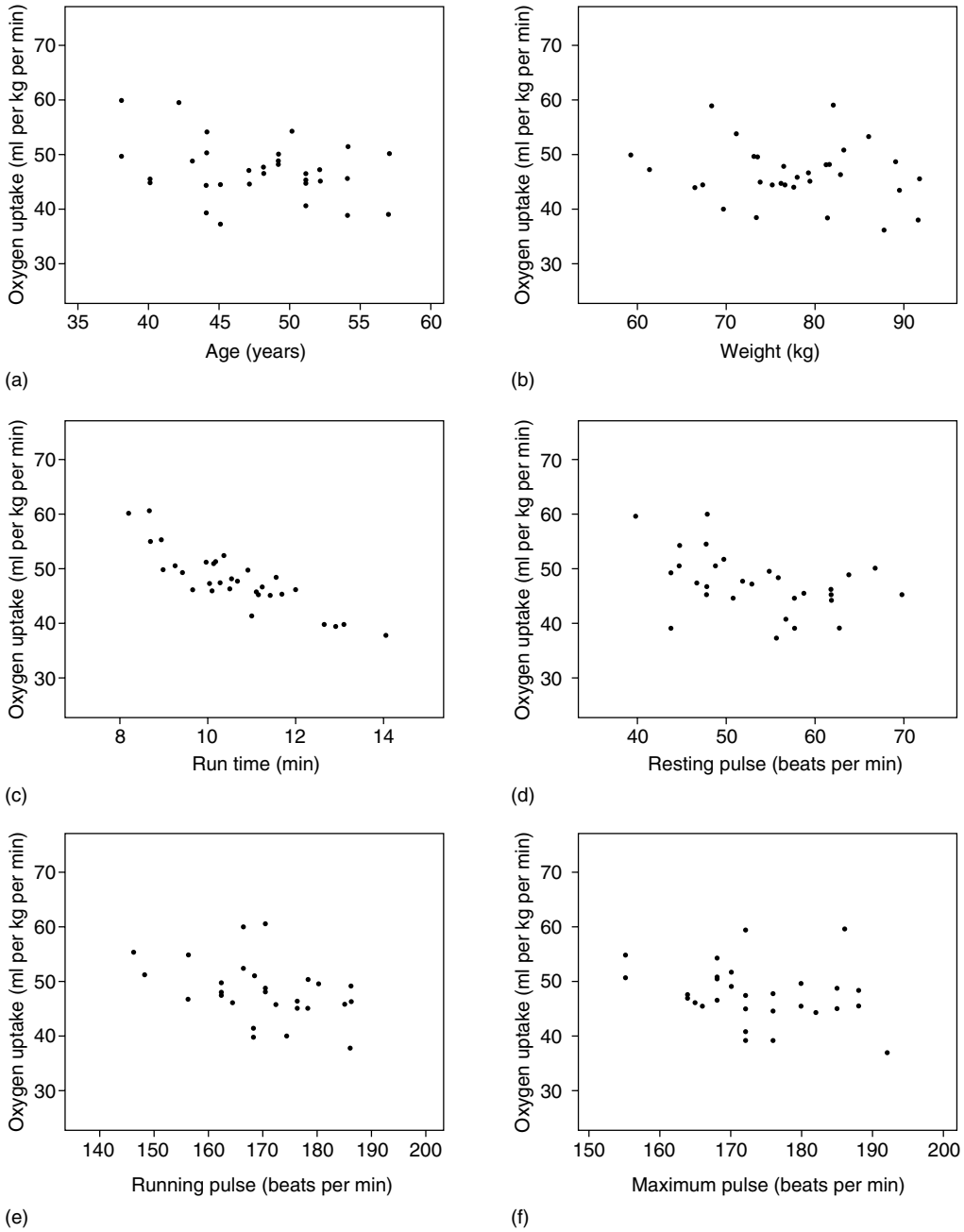

Figure 1 Scatterplots of oxygen uptake measurements, Y, vs. six explanatory measurements for a sample of 31 subjects: (a) Y vs. age; (b) Y vs. weight; (c) Y vs. running time; (d) Y vs. resting pulse; (e) Y vs. running pulse; (f) Y vs. maximum pulse

average oxygen uptake depends, systematically, in a roughly linear fashion on one or more of these potential explanatory variables seems plausible.

Least Squares Estimation

Every linear regression model consists of a systematic component – the model equation, such as that specified in (1) – and a **residual** (random, error) component, ε. Equating the sum, $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$, to Y defines the regression equation for the response. The residual, $\varepsilon = Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_k X_k$, represents the amount that an observed value of Y deviates from the predicted mean, $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$. Few $(k+1)$-tuples $(X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, in a given set of data are likely to lie on the predicted plane. The method of least squares identifies the unique values of $\beta_0, \ldots, \beta_k$ that minimize the average of the squared residuals. The details of the calculations, which are readily handled by almost any software package that includes statistical procedures, are best summarized in terms of matrix notation. Let $\mathbf{X}$ denote the data matrix, sometimes called the design matrix:

$$\begin{pmatrix} 1 & X_{11} & X_{21} & \ldots & X_{k1} \\ 1 & X_{12} & X_{22} & \ldots & X_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{kn} \end{pmatrix}$$

Each row in $\boldsymbol{X}$ represents the data obtained from a case (study unit or subject), and each column corresponds to an explanatory variable; the column of 1s is usually inserted automatically by the statistical software, and corresponds to the constant or intercept term, β_0, unless the user specifically chooses to omit it from the assumed model. If $\boldsymbol{Y}$ denotes a column vector consisting of the corresponding response measurements, $Y_1, \ldots, Y_n$, then the least squares estimator of the column vector, $\boldsymbol{\beta}$, with entries $\beta_0, \beta_1, \ldots, \beta_k$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \qquad (2)$$

This formula presupposes that the matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, i.e. that the $k+1$ simultaneous equations summarized in the single, least squares vector equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ have a unique solution in the $k+1$ unknowns, $\beta_0, \ldots, \beta_k$. When this is not the case, most software packages provide the user with a suitable warning of problems encountered in evaluating $\hat{\boldsymbol{\beta}}$. The equation of the estimated regression of Y on $X_1, \ldots, X_k$ is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k.$$

To obtain estimates of $\hat{\beta}_0, \ldots, \hat{\beta}_k$, we need to make the minimal least squares assumptions that the residuals, $\varepsilon_1, \ldots, \varepsilon_n$, are uncorrelated and have a mean value of 0 and constant variance, σ^2. We use the estimated residuals,

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_k X_{ki}, \quad i = 1, \ldots, n,$$

to estimate σ^2. The formula

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

which involves $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, the estimated residual sum of squares, emphasizes that $k+1$ parameters, $\beta_0, \ldots, \beta_k$, are estimated; the divisor, $n - (k+1) = n - k - 1$, is known as the residual **degrees of freedom** (df). Adoption of the additional assumption that the residuals are normally distributed leads to various statistical procedures that we discuss subsequently. First, however, we examine linear regression as an explanation for the observed variability in the response, Y.

Partitioning the Variability in Y

In simple linear regression, the use of a single explanatory variable to model $\mathrm{E}(Y|X)$ provides two sources for the observed variability in Y – variation due to changes in X and hence in $\mathrm{E}(Y|X)$, and residual variation in values of Y that have the same X value and hence the same mean. In multiple linear regression, each additional explanatory variable incorporated into the model for $\mathrm{E}(Y|X_1, \ldots, X_k)$ represents an additional, potential source for the observed variability in the response. Since the total variation in Y is fixed, and equal to $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the inclusion of additional explanatory variables in the model for the mean value of Y necessarily means that the magnitude of the estimated residual variation – the estimated residual sum of squares, $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ – will decrease. Consequently, the

estimated value of σ^2 tends to decrease also, as additional explanatory variables are added to the model for the mean response. For example, in the regression of oxygen uptake on the explanatory variables $X_1, \ldots, X_7$, Table 1 shows that adding age, running pulse, maximum pulse, gender, and weight to successively more complex models involving running time and all the previously used explanatory variables results in monotonically smaller estimates of σ.

However, when resting pulse is added to the previous model, the value of s increases, even though the model sum of squares increases (and hence the residual sum of squares decreases) by 0.3. The change in the model sum of squares associated with adding resting pulse to the previous model is marginal at best. Clearly, knowing a subject's age, for example, in addition to knowing the time he or she takes to run 1.5 miles, is informative in predicting that subject's oxygen uptake. However, the same subject's resting pulse does not provide important information about what his or her oxygen uptake is likely to be when the values of running time, age, running pulse, maximum pulse, gender, and weight have already been used to predict oxygen uptake.

It can be shown that the partitioning of the variability in Y is represented in the equation

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

which is alternately described by the relationship

$$\text{total sum of squares} = \text{model sum of squares} + \text{residual sum of squares.}$$

The partitioning that corresponds to any particular model equation is usually summarized in an **analysis of variance** (ANOVA) table, such as the one corresponding to the example shown in Table 2.

Table 2 ANOVA table corresponding to the regression of oxygen uptake on the explanatory variables running time, age, running pulse, maximum pulse, gender, and weight for a sample of 31 subjects

Source	SS	df	MS	F ratio
Model	729.1	6	121.52	23.8
Residual	122.3	24	5.10	
Total	851.4	30		

The ratio of the model sum of squares to the total sum of squares is called R^2, and represents the proportion of the observed variability in Y that is accounted for by modeling the mean response for Y as the assumed function of the explanatory variables in the model equation for $\mathrm{E}(Y|X_1, \ldots, X_k)$.

Interpreting the Estimated Regression Coefficients

If two values of one of the explanatory variables, say X_j, differ by one unit, then the corresponding values of the model equation differ by β_j, provided the values of all the other explanatory variables in a model equation remain the same. Therefore, $\hat{\beta}_j$ represents the estimated change in the mean response associated with a unit increase in the corresponding explanatory variable, provided the values of $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ do not change. Similar interpretations apply to each of the regression coefficients, $\beta_i, i = 1, \ldots, k$, in any postulated regression model similar to (1). Of course, each estimated coefficient and its interpretation are only applicable within the range of values of the corresponding explanatory variable that was used in fitting the linear regression model.

The value β_0 represents the mean response when all the explanatory variables in the regression model are equal to 0. In most cases, this mean response will

Table 1 Successive partitions of the observed variability in oxygen uptake measurements into the systematic (model) and residual components, as a result of incorporating additional explanatory variables in the model equation

	Subscripts of explanatory variables in the model equation for $\mathrm{E}(Y)$						
	3	1,3	1,3,5	1,3,5,6	1,3,5,6,7	1,2,3,5,6,7	1,2,3,4,5,6,7
Model	632.9	650.7	690.6	712.5	723.1	729.2	729.5
Residual	218.5	200.7	160.8	138.9	128.3	122.2	121.9
Residual df	29	28	27	26	25	24	23
s	2.74	2.68	2.44	2.31	2.27	2.26	2.30

be of no interest to the investigator, or may not belong to the range of values of $X_1, \ldots, X_k$ used in fitting the model to data. Armitage & Berry [1, pp. 313–314] show that $\hat{\beta}_0$ is equal to $\overline{y} - \hat{\beta}_1\overline{x}_1 + \cdots + \hat{\beta}_k\overline{x}_k$. This result leads to an alternative form for the fitted model, namely

$$\hat{Y} = \overline{y} + \hat{\beta}_1(X_1 - \overline{x}_1) + \cdots + \hat{\beta}_k(X_k - \overline{x}_k),$$

which reveals that the estimated mean response when $X_j = \overline{x}_j$, $j = 1, \ldots, k$, is simply the sample mean of Y. More than likely, this estimate will have a scientific meaning that an investigator can interpret sensibly.

If a regression model containing X_1, X_3, and X_5 is fitted to the oxygen uptake data discussed previously, the estimates of β_1, β_3, and β_5 are -0.26, -2.83, and -0.13, respectively. From these data we conclude that 2.83 ml per kg of body weight per min is the estimated decrease in mean oxygen uptake associated with a 1 min increase in the time that a subject takes to run 1.5 miles, provided age and running pulse are unchanged. Likewise, 0.26 ml per kg of body weight per min is the estimated decrease in mean oxygen uptake associated with a one-year increase in age, provided running time and running pulse do not change. At an age of $\overline{x}_1 = 27.7$ years, a running time for 1.5 miles of $\overline{x}_3 = 10.6$ min, and a running pulse of $\overline{x}_5 = 170$ beats per min, the estimated mean oxygen uptake among these subjects is $\overline{y} = 47.4$ ml per kg of body weight per min.

Of course, with seven potential explanatory variables, we require some systematic way of determining which of the $2^7 = 128$ possible models involving $X_1, \ldots, X_7$ represents the most satisfactory summary of the observed data. Various model or explanatory **variable selection** strategies have been developed. Some strategies depend only on the minimal, least squares assumptions, while others rely on the more powerful, and therefore more restrictive, additional premise that $Y_1, \ldots, Y_n$ are normally distributed. In the following section we describe a method of winnowing the set of all possible regression models involving the explanatory variables $X_1, \ldots, X_k$ into a shortlist of two or three, based solely on the least squares requirements of uncorrelated residuals with a common mean of 0 and constant variance represented by σ^2. Thereafter, we introduce the normal theory assumptions, and consequent methods of statistical inference, that provide a basis for differentiating among alternatives on the shortlist of candidate models for the data.

Identifying Good Candidate Models Based on all Possible Subsets

Only the advent of modern, high-speed computing, and the widespread availability of carefully written statistical software, has made the use of this approach to model selection feasible for many users. We do not intend to provide any details concerning the actual calculations involved, assuming that the interest of readers is focused elsewhere. It suffices to state that the result of the calculations is a table of all possible models, or perhaps a subset of all possible models, in which each candidate for the shortlist is identified by the list of explanatory variables it contains and the values of one or more criteria to use in comparing various candidates. Such a table is usually organized according to the number of explanatory variables used. For the oxygen uptake example, there is one model containing none of $X_1, \ldots, X_7$, seven possible models consisting of just one explanatory variable, 21 involving a pair of variables, 35 that employ three variables, a further 35 using a total of four of the k Xs, another 21 that depend on five explanatory variables, an additional seven that omit exactly one of X_1 through X_7, and one so-called full model that consists of all seven explanatory variables.

Three distinct but related numerical criteria are generally used to rank candidate models involving the same number of explanatory variables, and also to distinguish among the best models involving different numbers of explanatory variables. These are: (i) $R^2(p)$, (ii) $s^2(p)$, and (iii) **Mallow's C_p statistic** [6]. The dependence of each of these quantities on the variable p emphasizes that the value obtained changes according to p, the number of explanatory variables in the candidate model, as well as the choice of the particular subset of p explanatory variables from the full set of k Xs.

We noted previously that each time an additional explanatory variable is added to a model previously fitted, the model sum of squares increases, and hence the residual sum of squares decreases. Since R^2 is the ratio of the model sum of squares to the total

sum of squares, $R^2(p)$ will increase as p increases. In general, $s^2(p)$ decreases, although, as we have already seen in the example, the value may achieve a minimum for some value of $p < k$. Like $s^2(p)$, the C_p statistic usually decreases initially as p increases, although it is common that the value of C_p is minimized for some $p < k$. Models that are candidates for the shortlist typically have values of C_p that approach p. For the example discussed in this article, Table 3 provides a shortlist of two candidate models for each value of p, and summarizes the values of $R^2(p)$, $s^2(p)$, and C_p that were used to select these candidate models for further consideration.

To distinguish further among the various possibilities, we require statistical procedures that allow us to determine whether or not the contributions that particular explanatory variables make to a given model are important when compared with $\hat{\sigma}$, the estimated residual variability for observations.

Statistical Inference in Multiple Linear Regression

We assume that the goal in multiple linear regression modeling is to identify a fitted model that provides reasonably precise estimates of the mean response using a parsimonious set of explanatory variables. If we also assume that the residuals, $\varepsilon_1, \ldots, \varepsilon_n$, are normally distributed, the estimators of $\beta_0, \ldots, \beta_k$ have normal **sampling distributions**. Estimated **standard errors** (est se) for $\hat{\beta}_0, \ldots, \hat{\beta}_k$ are routinely produced by most computing packages. For each explanatory variable included in the model equation, the ratio of the difference, $\hat{\beta}_j - \beta_j$, to its corresponding estimated standard error follows a **Student's *t* distribution** with $n - (k + 1) = n - k - 1$ df, where k corresponds to the number of explanatory variables in the fitted model under consideration. From these results, **hypothesis tests** and/or **confidence intervals** for $\beta_0, \ldots, \beta_k$ can be evaluated. These tools will allow us to determine whether or not the association between Y and a particular explanatory variable, X_j, in a given fitted model is real ($\beta_j \neq 0$).

The same assumptions also lead to the result that the sampling distribution of $\overline{Y}$ is normal, and the corresponding estimated standard error is $s/\sqrt{n}$, where $s^2 = \hat{\sigma}^2$.

A test of the null hypothesis, $H_0 : \beta_j = 0$, is routinely used to assess the significance of the regression with respect to the explanatory variable X_j, i.e. to determine whether the data constitute statistical evidence of an association between Y and X_j. This test can be based either on the ratio $\hat{\beta}_j$/est se $(\hat{\beta}_j)$, which has a Student's t distribution with $n - k - 1$ df, or on $[\hat{\beta}_j/\text{est se } (\hat{\beta}_j)]^2$, which has an ***F* distribution** with 1 and $n - k - 1$ df.

Table 3 Shortlists of candidate models involving p explanatory variables for the regression of 31 oxygen uptake measurements on age, X_1, weight, X_2, running time, X_3, resting pulse, X_4, running pulse, X_5, maximum pulse, X_6, and gender, X_7

p	Explanatory variables included in the model equation	Corresponding values of $R^2(p)$	$s^2(p)$	C_p
1	X_3	0.743	7.53	14.1
	X_5	0.159	24.71	108.0
2	X_1, X_3	0.764	7.17	12.8
	X_3, X_5	0.761	7.25	13.3
3	X_1, X_3, X_5	0.811	5.96	7.3
	X_3, X_5, X_6	0.810	5.99	7.5
4	X_1, X_3, X_5, X_6	0.837	5.34	5.2
	X_1, X_3, X_5, X_7	0.836	5.37	5.3
5	X_1, X_3, X_5, X_6, X_7	0.849	5.13	5.2
	X_1, X_2, X_3, X_5, X_6	0.848	5.18	5.4
6	$X_1, X_2, X_3, X_5, X_6, X_7$	0.856	5.10	6.0
	$X_1, X_3, X_4, X_5, X_6, X_7$	0.856	5.35	7.2

With respect to the example, Table 3 indicates that good models involving three, four, or five explanatory variables all include running time, X_3, running pulse, X_5, age, X_1, and one or both of maximum pulse, X_6, and gender, X_7. Table 4, part (a), summarizes the results of fitting a model that contains all five explanatory variables. Neither the regression coefficient for maximum pulse nor that for gender is significantly different from 0; however, the significance levels (*see* ***P*** **Value**) corresponding to coefficients associated with the remaining three explanatory variables are all less than 0.05. These results suggest that, for the oxygen uptake data, we probably want to consider a model that involves running time, running pulse, and age, but perhaps only one of the variables gender or maximum pulse. Table 4, part (b), summarizes the results of fitting the two models that involve running time, running pulse, age, and either maximum pulse or gender. Based on these tabulated results, we conclude that the contribution of either

Table 4 Results of fitting various multiple linear regression models to the measurements on oxygen uptake (Y)

(a) Model involving age, X_1, running time, X_3, running pulse, X_5, maximum pulse, X_6, and gender, X_7

Explanatory variable	Estimated regression coefficient	Estimated standard error	Student's t statistic	Significance level
X_1	−0.200	0.094	−2.137	0.04
X_3	−2.872	0.342	−8.406	$<10^{-4}$
X_5	−0.354	0.115	−3.071	0.01
X_6	0.206	0.139	1.485	0.15
X_7	1.919	1.338	1.434	0.16

(b) Two models involving only four explanatory variables

Explanatory variable	Estimated regression coefficient	Estimated standard error	Student's t statistic	Significance level
		Model 1		
X_1	−0.198	0.096	−2.068	0.05
X_3	−2.768	0.341	−8.127	$<10^{-4}$
X_5	−0.348	0.118	−2.963	0.006
X_6	0.271	0.134	2.025	0.05
		Model 2		
X_1	−0.241	0.092	−2.629	0.01
X_3	−2.946	0.346	−8.523	$<10^{-4}$
X_5	−0.208	0.062	−3.365	0.003
X_7	2.566	1.294	1.983	0.06

(c) Final model based on X_1, X_3, and X_5

Explanatory variable	Estimated regression coefficient	Estimated standard error	Student's t statistic	Significance level
X_1	−0.256	0.096	−2.665	0.01
X_3	−2.825	0.358	−7.886	$<10^{-4}$
X_5	−0.131	0.051	−2.588	0.02

maximum pulse or gender in predicting the mean oxygen uptake is marginal, once we have used the information provided by a subject's running time, running pulse, and age. If, for the sake of model **parsimony**, we decide to select as a final model one involving just the three common explanatory variables running time, running pulse, and age, then we obtain the estimated regression coefficients and corresponding estimated standard errors given in Table 4, part (c). The values of R^2 and s^2 for this model (see Table 3) are 0.811 and 5.96, respectively, and the equation of the fitted model is

$$\hat{Y} = 111.72 - 0.256X_1 - 2.825X_3 - 0.131X_5. \quad (3)$$

The individual 95% confidence intervals for β_1, β_3, and β_5 are $(-0.45, -0.06)$, $(-3.56, -2.09)$, and $(-0.23, -0.03)$, respectively. For $\overline{Y}$, the corresponding interval estimate is (46.5, 48.3) ml of oxygen per kg of body weight per min.

Automatic Model Selection

Software packages frequently offer automatic methods of selecting variables for a final regression model from a list of candidate variables. There are three typical approaches, usually known as forward selection, backward elimination, and stepwise regression. These methods rely on significance tests known as partial F tests (*see* **Analysis of Variance**) to select an explanatory variable for inclusion in or deletion from the regression model. The forward selection approach begins with an initial model that contains only a constant term, i.e. $\mathrm{E}(Y|X_1, \ldots, X_k) = \beta_0$, and successively adds explanatory variables to the model from the set $X_1, \ldots, X_k$ until the pool of candidate variables remaining contains no variables that, if added to the current model, would contribute information that is statistically important concerning the mean value of the response. The backward elimination method begins with an initial model that contains all explanatory variables in the list, and then identifies the single variable that contributes the least information concerning the mean value of the response, i.e. results in the smallest decrease in the model sum of squares. If a partial F test identifies that this contribution is not statistically significant, then the variable is eliminated from the current model. Successive iterations of the method result in a "final" model from which no variable can be eliminated without adversely affecting, in a statistical sense, the predicted value of the mean response.

The stepwise regression method of variable/model selection combines elements of both forward selection and backward elimination. The initial model for stepwise regression is one that contains only a constant term. Subsequent cycles of the approach involve first the possible addition of an explanatory variable to the current model, followed by the possible elimination of one of the variables included in the newly augmented model. Both steps in the cycle rely on suitable partial F statistics. Succeeding cycles follow the same pattern until the set of variables in the model stabilizes, at which time a "final" model is declared.

The results produced by any of these, or most other, methods of automatic selection depend crucially on various user-selected adjustments for each procedure. For example, in forward selection the user has to specify (or use the default value of) a significance level to enter (SLE). This adjustment represents a threshold value that determines whether or not a candidate explanatory variable is eligible to be added to the current model. For backward elimination, there is a corresponding significance level to stay (SLS) – a threshold value that determines whether or not an explanatory variable is a candidate for removal from the current model. In stepwise regression, the user needs to specify both an SLE and an SLS. For example, if forward selection is used for the oxygen uptake data with a value of 0.05 for SLE, the final model involves age X_1, running time, X_3, and running pulse, X_5. When $\mathrm{SLE} = 0.10$, the final model selected by forward selection also includes maximum pulse, X_6. The use of forward selection with the default value of SLE (0.50) built into one widely used software package results in a final model consisting of all the explanatory variables except resting pulse X_4.

A separate factor that influences the results of all automatic methods of model selection in an unpredictable fashion is the underlying **correlation** structure of the data. If two explanatory variables are strongly correlated with one another, it is highly unlikely that any of the usual automatic methods of model selection will produce a final model that includes both variables. This outcome is appropriate, since curious statistical pathologies are likely to occur if two highly correlated, and therefore nearly

collinear, variables are simultaneously included in a regression model. However, the final model that automatic selection produces hides the fact that another line of modeling exists based on the second of the two highly correlated variables, and the end result of pursuing that direction might be equally satisfactory, statistically or scientifically, or perhaps even better.

In summary, automatic methods of model or variable selection provide no guarantee of identifying a "best" model in any overall scientific sense. Consequently, we recommend that investigators treat the results of an automatic approach to model selection with a healthy measure of skepticism. In particular, if there is no indication that the final model, no matter how it was arrived at, has been subjected to the diagnostic scrutiny outlined in the next section, then there are sound statistical grounds for questioning any conclusions based on the results of the model-fitting process.

Model Diagnostics

For simplicity, we assume that the result of a thoughtful, comprehensive model-fitting strategy depends on explanatory variables labeled $X_1, \ldots, X_p$. A fitted regression model and associated statistical inferences are based on various assumptions concerning the functional form of the model for $\mathrm{E}(Y|X_1, \ldots, X_p)$ and distributional properties of the residuals. Violations of these assumptions may invalidate conclusions based on the regression analysis. Therefore, it is essential to check these assumptions, using various types of diagnostic plots.

The estimated residuals,

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_p X_{pi},$$
$$i = 1, \ldots, n,$$

play an essential role in model **diagnostics**. Many computer packages offer the option of using these ordinary residuals or the corresponding standardized or studentized (*see* **Studentization**) residuals, which have a common variance. Use of either of the latter two is preferable, since the $\hat{\varepsilon}_i$s do not all have the same variance.

The following diagnostic plots furnish graphical evidence that one or more of the model assumptions may be contradicted by the data:

1. Residuals vs. the fitted values, $\hat{Y}_i$. An unsuitable functional form is usually revealed by the systematic appearance of this plot, as is nonconstant variance.
2. Residuals vs. the explanatory variables, $X_i, i = 1, \ldots, p$. Systematic patterns in these plots can indicate violations of the mean 0, constant variance assumptions, or an inappropriate model form.
3. Normal probability plot of the residuals. This plot checks the normal distribution assumption on which all the statistical inference procedures are based.
4. Residuals vs. the temporal/spatial order of data collection. Unexpected regularity in this plot suggests that the Y_is may be correlated. To prepare this diagnostic check, it is essential to record the temporal/spatial ordering when data are first collected.
5. Index plots (plot against case number) of the leverages and Cook's distance. The former are a measure of the amount of influence exerted on $\hat{Y}_i$ by the corresponding observed response, Y_i. Cook's distance is a summary measure of the influence that each case exerts on the estimated regression coefficients. These two diagnostic plots can reveal **outliers** (values of Y that are anomalous with respect to the rest of the data) or influential points (values of $(X_{1j}, \ldots, X_{pj}, Y_j)$ that strongly influence the estimated values of $\hat{\beta}_0, \ldots, \hat{\beta}_p$ and s^2).

Deviations from the expected (null) pattern in any of these plots may indicate problems that require further investigation or remedial action.

Various diagnostic plots for the oxygen uptake example are displayed in Figure 2. These reveal that two, or possibly three, of the observations are somewhat unusual relative to the rest of the data set. Omitting the two most prominent points identified in these plots from the data set results in a fitted model that is little different from (3). In particular, only the regression coefficient for run time, X_3, changes noticeably, increasing from -2.83 to -2.62; the corresponding change in the estimated standard error is a reduction from 0.358 to 0.313. The stability of the fitted model, despite the omission of possible influential points, is reassuring.

Computer packages with good facilities for multiple linear regression modeling routinely incorporate

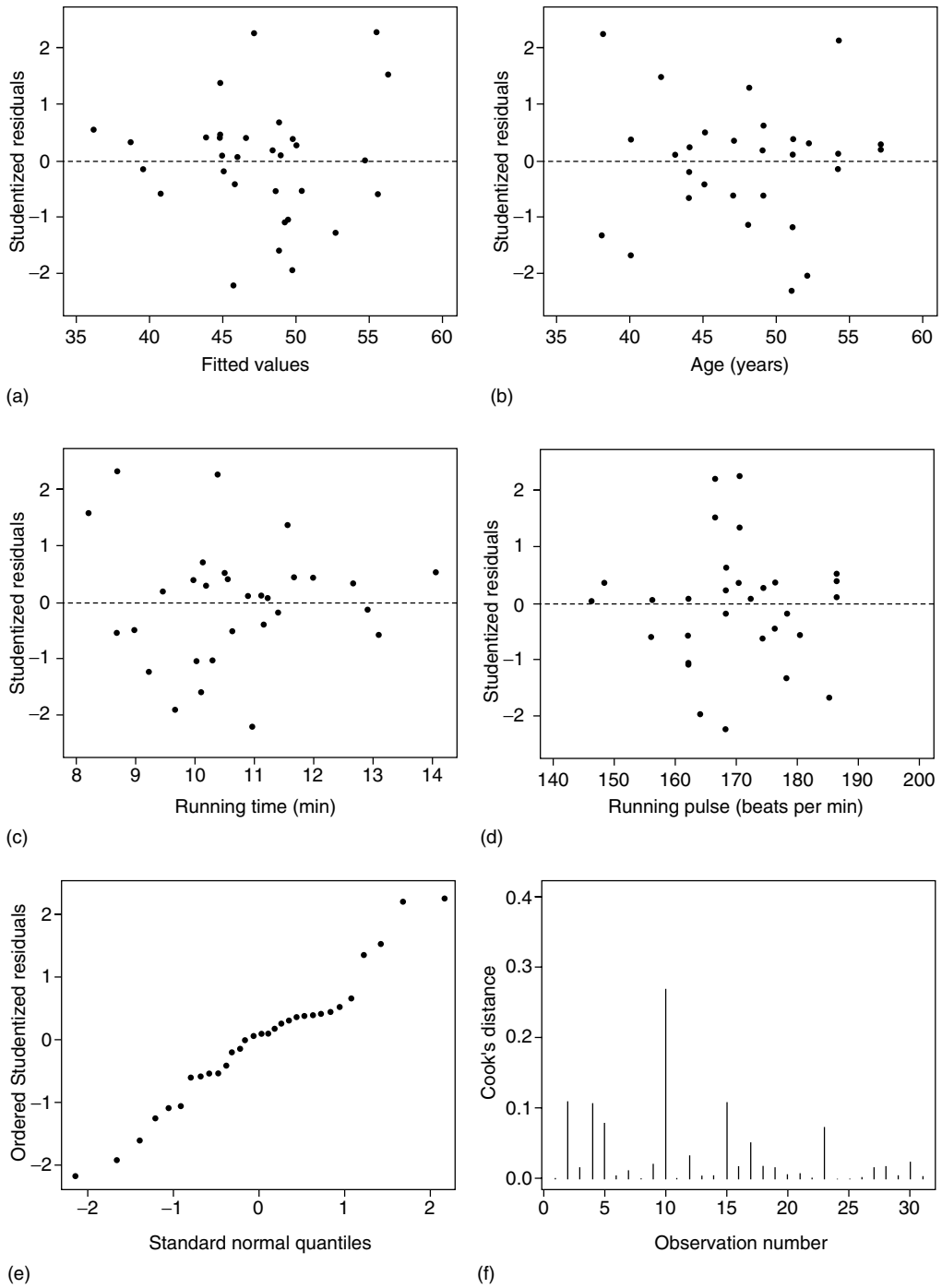

Figure 2 Diagnostic plots for the regression model $\hat{Y} = 111.72 - 0.256X_1 - 2.825X_3 - 0.131X_5$ fitted to the oxygen uptake measurements displayed in Figure 1: (a) estimated Studentized residuals ($\hat{\varepsilon}^*$) vs. $\hat{Y}$; (b) $\hat{\varepsilon}^*$ vs. age; (c) $\hat{\varepsilon}^*$ vs. running time; (d) $\hat{\varepsilon}^*$ vs. running pulse; (e) normal probability plot for $\hat{\varepsilon}^*$; (f) index plot of Cook's distance

simple methods of preparing all the diagnostic plots mentioned in the preceding list. For additional details concerning model diagnostics, see [2] or [3]. Further details concerning examination of the adequacy of a fitted regression model are found in the article, **Goodness of Fit**.

Prediction

The problem of **prediction** using a fitted regression model can be posed in two distinct ways. For example, a researcher may be interested in predicting the mean value of the response, $E(Y|X_1, \ldots, X_p)$, in the subgroup of the study population defined by the values $X_1 = x_1, \ldots, X_p = x_p$ of the explanatory variables. Under the assumption that the residuals in the model are uncorrelated and normally distributed with mean 0 and common standard deviation σ, this mean value is a parameter of the resulting normal distribution of responses in the subgroup. Hence, we can obtain both point and interval estimates of $\mu' = E(Y|X_1 = x_1, \ldots, X_p = x_p)$. The former, which is equal to

$$\hat{\mu}' = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p,$$

is the least squares (and **maximum likelihood**) estimate of the mean response, μ', when $X_1 = x_1, \ldots, X_p = x_p$, whereas the latter represents the range of plausible values for μ' that are consistent with the observed data on which the fitted model is based. Narrow interval estimates indicate that the data and fitted model permit us to estimate μ' rather precisely; wider intervals reflect a greater degree of uncertainty concerning the mean value of response in this particular subgroup. It is possible to show, mathematically, that *interval estimates* for μ' are always narrowest in the observed center of the space defined by $X_1, \ldots, X_p$, i.e. at $X_1 = \overline{x}_1 \ldots, X_p = \overline{x}_p$. At the boundaries of the explanatory space, the estimated standard error of $\hat{\mu}'$ is frequently substantially larger than the corresponding estimated standard error at $X_1 = \overline{x}_1, \ldots, X_p = \overline{x}_p$. For example, in the case of the oxygen uptake data, the median values of age, running time, and running pulse are 48, 10.47, and 170, respectively, and all three values are very close to the corresponding observed means, 47.6, 10.58, and 169.6. The estimated mean oxygen uptake for runners in the study population whose age, running time, and running pulse are equal to the median values is

$$\hat{\mu}' = 111.72 - 0.256(48) - 2.825(10.47) - 0.131(170) = 47.6$$

ml per kg per min, and the corresponding estimated standard error is 0.444. Thus, a 95% confidence interval for the mean response among such individuals is (46.7, 48.5) ml per kg per min. However, among older, slower runners who are less fit, e.g. $X_1 = 55$, $X_3 = 13.63$, $X_5 = 185$, the estimated mean oxygen uptake is only 34.9 ml per kg per min and the corresponding estimated standard error is 1.322, resulting in the much wider 95% confidence interval (32.2, 37.6). This threefold increase in the estimated standard error of $\hat{\mu}'$ reflects the increased uncertainty that is a natural consequence of trying to predict the mean oxygen uptake in a region of the space of explanatory variables that is sparsely covered by the observed data.

The second aspect of the prediction problem concerns individual response measurements in the subgroup of the study population defined by the values $X_1 = x_1, \ldots, X_p = x_p$. Some authors recommend that the predicted mean value for the subgroup is a suitable point estimate for an individual response as well; however, the estimated standard error for a predicted response is greater than the corresponding estimated standard error for the predicted mean value, for reasons that we will subsequently explain. We prefer to suggest that point estimates of individual responses, i.e. of Y for a subject with values $X_1 = x_1, \ldots, X_p = x_p$ of the explanatory variables in the fitted regression model, cannot be determined, since these would be observations from a normal distribution rather than characteristics (parameters) of the distribution. However, interval estimates for individual responses in the same subpopulation defined by the values $X_1 = x_1, \ldots, X_p = x_p$ can be evaluated. Of necessity, these interval estimates will be even wider than interval estimates for the corresponding mean response. This is because, according to the multiple linear regression model, an individual response is the sum of a particular mean response and a residual. Thus, to the uncertainty about the location of the mean response in the subgroup we need to add the uncertainty due to the residual associated with an individual response, i.e. est se $(Y) =$

$\{[\text{est se}(\hat{\mu})]^2 + [\text{est se}(\varepsilon)]^2\}^{1/2} > \text{est se}(\hat{\mu})$. To illustrate, consider runners in the example of those running pulses that have the median values 48, 10.47, and 170, respectively. A 95% prediction interval for an oxygen uptake measurement on a new runner belonging to this particular subgroup is (42.5, 52.7) – roughly five times wider than the corresponding 95% confidence interval of (46.7, 48.5) ml per kg per min for the mean response among such individuals. Also, attempting to predict individual responses in regions of the space of explanatory variables that are sparsely covered by the observed data inevitably results in interval estimates that are wider still. For example, a 95% prediction interval for oxygen uptake among young, faster runners who are quite fit, e.g. $X_1 = 40$, $X_3 = 8.62$, $X_5 = 154$, is (51.3, 62.6), whereas the corresponding interval estimate for the mean response in the same subgroup is (54.4, 59.5) ml per kg per min.

Most computer packages with well-designed routines for multiple linear regression modeling offer users the option of computing suitable confidence intervals for both aspects of the problem of prediction, as well as point estimates of the mean response for any set of values of the explanatory variables in a fitted model.

Weighted Regression

A diagnostic plot may exhibit a systematic pattern, suggesting that the variability of the estimated residuals is not constant from one observation to another. Alternatively, it may be known that some of the observations in a set of data are less reliable, i.e. more variable than the remaining observations collected.

Whatever the reason, if the residuals, $\varepsilon_1, \ldots, \varepsilon_n$, and hence the corresponding response measurements, $Y_1, \ldots, Y_n$, are correlated and/or do not all have roughly the same standard deviation, σ, the least squares estimator specified in (2) is no longer appropriate, since the minimal assumptions for least squares estimation are not satisfied.

To illustrate the general solution to such problems, we consider the simple but artificial problem of fitting a simple linear regression model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$, when the residuals, ε_i, are uncorrelated but have different, known standard deviations, $\sigma_i, i = 1, \ldots, n$. If the standard deviations were all the same, i.e. if $\sigma_i = \sigma, i = 1, \ldots, n$, then the ordinary (unweighted) least squares estimates of β_0 and β_1 in this simple linear regression context would be $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, where $\overline{x} = \sum_{i=1}^n x_i/n$, $\overline{y} = \sum_{i=1}^n y_i/n$,

$$S_{xy} = \sum_{i=1}^n x_i y_i - n^{-1} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

and

$$S_{xx} = \sum_{i=1}^n x_i^2 - n^{-1} \left(\sum_{i=1}^n x_i \right)^2 .$$

However, as we outline subsequently, when the standard deviations of the residuals are different, the correct (weighted) least squares estimates of β_1 and β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n w_i x_i y_i - \left(\sum_{i=1}^n w_i x_i \right) \left(\sum_{i=1}^n w_i y_i \right)}{\sum_{i=1}^n w_i x_i^2 - \left(\sum_{i=1}^n w_i x_i \right)^2} \quad (4)$$

and

$$\hat{\beta}_0 = \sum_{i=1}^n w_i y_i - \hat{\beta}_1 \sum_{i=1}^n w_i x_i, \quad (5)$$

where $w_i = \sigma_i^{-2} / \sum_{i=1}^n \sigma_i^{-2}$. Weighted least squares estimation associates a weight, w_i, $0 \le w_i \le 1$, $\sum_{i=1}^n w_i = 1$, with observation $i, i = 1, \ldots, n$, and these weights are inversely proportional to the variances of the corresponding residuals. Thus, the inherent reliability of each observation, which is reflected in the variance of the corresponding residual, directly determines the weight, and hence the contribution, of that observation to the estimation of the regression coefficients, β_0 and β_1, in the assumed model. Incidentally, if the standard deviations of the residuals, ε_i, are all the same, i.e. if $\sigma_i = \sigma, i = 1, \ldots, n$, then $w_i = 1/n, i = 1, \ldots, n$, and the formulas for the weighted least squares estimates of β_0 and β_1 specified in (4) and (5) simplify to the ordinary (unweighted) least squares expressions.

The preceding example shows that the solution to the problem of nonconstant residual standard deviations involves suitably weighting (transforming) the measurements associated with each observation. According to the simple linear regression

model, the residual is the difference between the observed response and the systematic component of the assumed model, i.e. $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i, i = 1, \ldots, n$. Thus, transforming the response and explanatory measurements for observation i automatically modifies the residuals in an identical manner, thereby producing transformed residuals that do satisfy the minimal assumptions for least squares estimation. The estimates of the regression coefficients that are derived using the weighted or transformed measurements are known as weighted least squares estimates.

The method of resolving the general problem represented by residuals that are correlated and/or do not all have roughly the same standard deviation, σ, involves identifying a unique, nonsingular, symmetric matrix, **T**, i.e. a matrix of suitable weights for each observed response and corresponding values of the explanatory variables, $(X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, such that $\mathbf{Z} = \mathbf{TY}$ and the effect of the same matrix, **T**, on the vector of residuals, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, produces transformed residuals, $\mathbf{T}\boldsymbol{\varepsilon}$, that are uncorrelated and do have a constant standard deviation.

Once the transformation matrix, **T**, has been identified, (2) can be used on the transformed response measurements; the vector **Y** is replaced by **Z** and the matrix, **X**, by **TX**, its corresponding equivalent in the transformed problem. Although it is possible to restate the solution for $\hat{\beta}$ in terms of the matrices **T**, **X** and the vector, **Y**, it is simpler to carry out any necessary calculations directly using **Z**, the vector of transformed response measurements, and **TX**, the matrix of transformed explanatory variable values for each observation. Once a suitable parsimonious model for $Z_1, \ldots, Z_n$ has been identified, appropriate statistical inferences concerning the estimated regression coefficients can be formulated provided that the assumption that the residuals for the estimation problem follow a normal distribution is not contradicted by the transformed data. To check all these least squares and normal theory assumptions, it is essential to examine the estimated residuals for the transformed data, i.e. $Z_i - \hat{Z}_i, i = 1, \ldots, n$, where $\hat{Z}_i$ denotes the predicted value of the transformed response for observation i and is obtained using $\hat{\boldsymbol{\beta}}$ and the vector of transformed explanatory variable values for observation i.

In the preceding discussion we did not indicate precisely how to identify the key matrix **T** by which all that previously was wrong is set right. This is because there is no universal antidote to cover all situations when one or more of the least squares assumptions fail to hold for a given set of data. Often, a simple **transformation** such as $Z_i = \sqrt{Y_i}$ or perhaps $Z_i = \log Y_i, i = 1, \ldots, n$, can put things right, particularly if the original difficulty with least squares was due to a failure of the constant variance assumption. If not, then much more effort will probably be required, as will a certain degree of data-analytic artistry which cannot be summarized in the small amount of space available here. Draper & Smith [4, pp. 112–115] discuss a numerical example in which the entries in the transformation matrix, **T**, are estimated from response measurements that are exact repeats taken at the same value of X, the only explanatory variable in the problem, or approximate repeated response measurements collected at values of X that are in close proximity to each other.

References

[1] Armitage P. & Berry, G. (1994). *Statistical Methods in Medical Research*, 3rd Ed. Blackwell Science, Oxford.

[2] Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

[3] Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London.

[4] Draper, N.R. & Smith, H. (1981). *Applied Regression Analysis*, 2nd Ed. Wiley, New York.

[5] Galton, F. (1885). Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute* **15**, 246–263.

[6] Mallows, C.L. (1973). Some comments on C_p, *Technometrics* **15**, 661–675.

[7] Pearson, K. & Lee, A. (1903). On the laws of inheritance in man. I. Inheritance of physical characters, *Biometrika* **2**, 357–462.

(*See also* **General Linear Model**)

DAVID E. MATTHEWS

Multiple Time Series

Multiple time series, also known as multivariate time series, and sometimes vector time series, refers to the analysis of observations taken simultaneously on two or more time series. Biomedical data examples include monthly recordings of death attributed to bronchitis, emphysema and asthma for males and females, and temperature, blood pressure and weight for a regularly monitored patient. Environmental examples include readings of lead concentrations at several sites at five minute intervals, air temperature readings taken at hourly intervals at a fixed height above sea level at several locations and monthly ozone levels at several recording stations. Univariate time series models are very useful for individual time series, but multiple time series models have the potential for helping to understand the *system* that gives rise to the data.

Parzen [21] gives a concise background to the theory of multiple time series and the references provide an excellent source to track its history. In this article I complement Parzen by updating many of the references and report some of the progress that has been made possible by the increased computing power that is now generally available. Most of the theoretical developments in multiple time series before and since 1985 have been driven by problems in economics, business and econometrics. Theoretical and practical contributions may be found in [2, 7, 9–11, 16, 17, 24, 26, 27, 30, 31], and [33].

Statistical Concepts

Let $\mathbf{Y_t} = (Y_{1,t}, Y_{2,t}, \ldots, Y_{m,t})'$ be a vector of m time series for $t = 1, 2, \ldots, n$. It is second-order stationary if (i) $\mathrm{E}(\mathbf{Y}_t) = \boldsymbol{\mu}$, independent of time, and (ii) the set of m series is jointly covariance stationary (*see* **Coherence Between Time Series**). Point (ii) means that the cross-covariance function $\gamma_{ij}(k) = \mathrm{cov}(Y_{i,t}, Y_{j,t-k})$ is independent of t for all i, j and k and the variance of each series, $\gamma_{ii}(0)$, is finite. The function $\gamma_{ii}(k)$ is a covariance function and is not necessarily symmetric. The function $\rho_{ij}(k) = \gamma_{ij}(k)/[\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}$ is the cross-correlation function, and the autocorrelation function is $\rho_{ii}(k) = \gamma_{ii}(k)/\gamma_{ii}(0)$. The matrix $\boldsymbol{\rho}(k)$ with the ijth entry equal to $\rho_{ij}(k)$ is the correlation matrix. The Fourier transform of $\boldsymbol{\rho}(k)$ (*see* **Fast Fourier Transform (FFT)**) is the *spectral density* matrix $\mathbf{f}(\omega) = \sum_{v=-\infty}^{\infty} \exp(-2\pi i \omega v)\boldsymbol{\rho}(k)$, $(-0.5 \le \omega \le 0.5)$, and is nonnegative definite. The complex-valued function, which is the ijth entry in this matrix, is the cross-spectral density. The real and imaginary parts of this matrix define the *cospectrum* and *quadrature spectrum*. The *amplitude*, $a_{ij}(\omega)$, and *phase*, $\phi_{ij}(\omega)$, of $\mathbf{f}(\omega)$ are obtained by considering the polar representation of the cross spectrum in the form $f_{ij}(\omega) = a_{ij}(\omega)\exp\{i\phi_{ij}(\omega)\}$. The quantity $a_{ij}(\omega)/\sqrt{[f_{ii}(\omega)f_{jj}(\omega)]}$ is called the coherency, represents the correlation between the frequency components of $Y_{i,t}$ and $Y_{j,t}$ and lies between zero and one [6, p. 212].

For m observed series the main task is to obtain sample estimates of the matrices of population quantities and use them to identify an appropriate model that will best describe $\mathbf{Y}_t$. The time domain approach involves working with estimates of $\rho_{ij}(k)$, while the frequency domain approach uses estimates of $\mathbf{f}(\omega)$. Practical progress can only be made by entertaining classes of time series models that can then be estimated. The multivariate generalization of Wold's decomposition theorem states that $\mathbf{Y}_t$ can always be represented by an infinite matrix linear combination of a vector white noise process [8, p. 158]. This leads one to consider the multivariate generalization of univariate autoregressive moving average (ARMA) processes (*see* **ARMA and ARIMA Models**) in the form $\boldsymbol{\Phi}(B)\mathbf{Y}_t = \mathbf{C} + \boldsymbol{\Theta}(B)\boldsymbol{\varepsilon}_t$, where $\boldsymbol{\Phi}(B) = \mathbf{I} - \boldsymbol{\Phi}_1 B - \ldots - \boldsymbol{\Phi}_p B^p$ and $\boldsymbol{\Theta}(B) = \mathbf{I} - \boldsymbol{\Theta}_1 B - \ldots - \boldsymbol{\Theta}_q B^q$ are matrix polynomials of orders p and q, respectively, B is the backward shift operator such that $B\mathbf{Y}_t = \mathbf{Y}_{t-1}$, all the zeros of the determinantal polynomials $|\boldsymbol{\Phi}(B)|$ and $|\boldsymbol{\Theta}(B)|$ are on or outside the unit circle and $\boldsymbol{\varepsilon}_t$ is a vector *white noise* process with $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{m,t})'$, having the properties $\mathrm{E}(\varepsilon_{i,t}) = 0$, for all i, $\mathrm{E}(\varepsilon_{i,t}\varepsilon_{j,t-k}) = 0$, for all k, $i \ne j$ and $k \ne 0$, $i \ne j$. These processes are usually termed vector ARMA (VARMA) models.

Identification

There are some major and difficult problems with VARMA modeling and many are discussed in [26], [7, pp. 244–259] and [16, p. 241]. First, if m is much larger than two or three, a rather large number of autocorrelations, cross-correlations and partial

correlations have to be interpreted. This makes the selection of the order of each polynomial somewhat difficult. In any case, these tools do not yield totally unambiguous messages about the models that should be estimated. Secondly, unique identification of a VARMA structure is not guaranteed by specifying a minimum order for the autoregressive and moving average operators. Thirdly, misspecification of multiple time series models can have more serious consequences than for univariate models. For these reasons, often the preference of applied researchers is to specify, estimate and analyze pure vector autoregressive (VAR) models. This assumes that a sufficiently long VAR model can capture the structure of $\mathbf{Y}_t$. A different approach is needed to solve the nonuniqueness problem for the general VARMA case. Lutkepohl & Poskitt [17] provide a strategy that results in a parsimonious and uniquely identifiable structure based on the echelon form of a VARMA model.

Estimation

The Gaussian likelihood of $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$ for the VARMA process is derived by Brockwell & Davis [2, p. 431]. Unlike in the univariate case, it is not possible to compute maximum likelihood estimators of $\mathbf{\Phi}(B)$ and $\mathbf{\Theta}(B)$ independently of the variance–covariance matrix of the noise series [2, p. 427] and so the maximization of the likelihood has to be done simultaneously. Efficient nonlinear optimization algorithms have to be used and it is particularly important to have good initial estimates of the parameters. This is because the likelihood function can have many local maxima that are smaller than the global maximum. Poskitt & Salau [22] show that generalized least squares estimates of VARMA models and corresponding Gaussian estimators are asymptotically convergent. In the case of estimating a pure VAR, the multivariate Durbin–Levinson algorithm for fitting autoregressions of increasing order is often used [2, p. 432]. An efficient algorithm for evaluating the exact likelihood function for VARMA models is given by Mauricio [18].

Model Checking

Univariate theory is easily extendable to yield Lagrange multiplier-type tests for vectors of estimated parameters and portmanteau-type tests for vectors of residual autocorrelations [16, pp. 298–301]. The latter procedure is, however, unlikely to be very powerful under a broad range of alternative VARMA models, since, even in the univariate case, portmanteau tests lack power [5].

Forecasting

Optimal h-step forecasts for VARMA models are straightforward to obtain [16, p. 228]. However, since a potentially large number of parameters may have been estimated in the identified multivariate model, prediction intervals for forecasts can be wider than expected. Both short- and long-term forecasting are important. For example, forecasts of wind speed and direction at several locations may be needed for 15-minute horizons, whereas environmentalists may need predictions of next year's rainfall over a wide and geographically spread region. Modeling long-range dependence by fractionally differenced VARMA models is in its infancy, but may prove to be a productive area for predicting environmental-type time series [23].

Data and Worked Examples

The World Wide Web is a useful source of information and data on biomedical and other time series (*see* **Internet**). A home page is available at the url `http://hachiman.mscs.mu.edu/research/biomedts` and there are some useful links from there. A bivariate worked example using monthly numbers of male and female deaths from bronchitis, emphysema and asthma is given by Diggle [6, p. 202]. Those data are also considered by Venables & Ripley [32, p. 373]. Both analyses concentrate on time and frequency domain statistics, and some useful plots are presented. No VARMA models are estimated. Using a state–space representation, Jones [12, p. 158] gives examples of estimating bivariate and trivariate multiple time series models for medical time series that are irregularly spaced.

Software

Sources of software (*see* **Software, Biostatistical**) for both univariate and multivariate time series analysis are provided by Aghadazeh & Romal [1] and

Rycroft [25]. The number of packages that specifically handle multiple time series is small and, to date, a comparative review of them has not been done. Ord & Lowe [20] compare five *automatic* forecasting systems for univariate analysis, only one of which provides for multiple time series analysis.

Finding innovative and useful ways to present m multiple time series, the joint behavior of all subsets of them and the statistics needed to identify VARMA models is challenging and not yet complete. Newton [19, p. 803] argues that essential features for good time series graphics are that the plots should be interactive, dynamic and linked. Few packages possess all these properties, but the statistical system in S [4] and LISP-STAT [28] have the greatest potential for the creation of useful plots for multiple time series modeling. Newton [19, p. 817] used the S system to present data on monthly ozone levels at nine recording stations with instantaneous scatter plots for all 36 pairs of series and the 45 plots of autocorrelations and cross-correlations. Other quantities, such as coherency, gain and phase need to be examined: currently, there is no readily available off-the-shelf program that presents these statistics in innovative ways.

The excellent book by Venables & Ripley [32, pp. 349–382] gives examples of graphical displays of time and frequency domain statistics for multiple time series analysis that can be produced using the S-PLUS system. VAR estimation (but not VARMA) is implemented in the system with appropriate model checking and diagnostic tools. A library of S functions for multivariate state–space and ARMA time series models is available from statlib in the S archive under the title time·series. These S functions allow for inclusion of exogenous variables and treat VAR models as a special case. They include methods for simulating, estimating and converting among different model representations. They are implemented using classes and methods so that it is easy to add new estimation methods and not difficult to add other model representations.

Scientific Computing Associates (SCA, PO Box 625, DeKalb, Illinois, USA) market a PC-based statistical system that includes a product dedicated to multivariate time series analysis and forecasting using VARMA models [15]. Box & Tiao have helped with the development of the system, and so it has a very good pedigree.

An easy-to-use and highly graphical package has been developed at the London School of Economics and is described by Koopman et al. [14] with case studies given in [13]. The software assumes that structural multivariate time series models are appropriate, with time-varying unobservable components that are themselves assumed stochastic. A full range of high quality graphical output is available to assist the modeler. The software is PC-based and a Windows version will be available shortly.

The software of Brockwell & Davis [3] allows VAR models to be estimated and forecast in a PC environment. In that software, model selection can be made simpler using the AIC criterion. The estimation routines for multivariate conditional heteroscedastic models used by Wong & Li [35] are written using the MATLAB software and may be obtained from the authors.

Extensions and Generalizations

Biomedical investigations usually involve experimental designs with deliberate replication of experiments. When these are conducted over time, relatively short and nonstationary time series arise in which trends and other features are the main interest. These kinds of data are generally known as *repeated measures* and the modeler can use time series techniques to accommodate serial dependence within the individual time series. In this way, repeated measurements can be regarded as multiple time series. A worked example using the body weight of rats is given by Diggle [6, p. 136].

State–space generalizations of multiple time series are given by Lutkepohl [16, p. 415] and Harvey [11, p. 423] (*see* **Structural Time Series Models**) and the multivariate dynamic linear model for time series is described by West & Harrison [34, p. 597]. A general methodology for the Bayesian analysis of short- and long-memory multiple integrated models is given by Ravishankar & Ray [23]. These authors illustrate their techniques by applying them to sea surface temperature series. Extensions to multivariate conditional heteroscedastic time series models, including their links to multivariate random coefficient autoregressive models, is given by Wong & Li [35]. In that paper the problems of identification, estimation and diagnostic checking are addressed. Univariate nonlinear time series modeling is itself in its infancy,

but Tong [29, p. 429] tentatively suggests the possibility of using multiple nonlinear time series models and gives an example that features threshold autoregressive series. The extra generality provided by these new developments should prove very useful for modeling biomedical- and environmental-type multivariate time series, since long-range dependence and changing conditional variance are commonly observable phenomena in biostatistical data.

References

[1] Aghadazeh, S.-M. & Romal, J.B. (1992). A directory of 66 packages for forecasting and statistical analysis, *Journal of Business Forecasting, Methods and Systems* **8**, 14–20.

[2] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Ed. Springer-Verlag, New York.

[3] Brockwell, P.J. & Davis, R.A. (1991). *ITSM: An Interactive Time Series Modelling Package for the PC*. Springer-Verlag, New York.

[4] Chambers, J.M. & Hastie, T.J. (1992). *Statistical Models in S*. Wadsworth & Brookes/Cole, Pacific Grove.

[5] Davies, N. & Newbold, P. (1979). Some power studies of a portmanteau test of time series model specification, *Biometrika* **66**, 153–155.

[6] Diggle, P. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

[7] Granger, C.W.J. & Newbold, P. (1986). *Forecasting Economic Time Series*, 2nd Ed. Academic Press, New York.

[8] Hannan, E.J. (1970). *Multiple Time Series*. Wiley, New York.

[9] Hannan, E.J. & Deistler, M. (1988). *The Statistical Theory of Linear Systems*. Wiley, New York.

[10] Hannan, E.J. & Kavalieris, L. (1984). Multivariate linear time series models, *Advances in Applied Probability* **16**, 713–723.

[11] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

[12] Jones, R.H. (1984). Fitting multivariate models to unequally spaced data, in *Time Series Analysis of Irregularly Observed Data. Springer Lecture Notes in Statistics*, Vol. 25, E. Parzen, ed. Springer-Verlag, New York, pp. 158–188.

[13] Koopman, S.J., Harvey, A.C. & Shephard, N. (1995). *Tutorials on Structural Time Series Models with STAMP 5*. Chapman & Hall, London.

[14] Koopman, S.J., Harvey, A.C., Doornik, J. & Shephard, N. (1995). *STAMP 5.0: Structural Time Series Analyzer, Modeller and Predictor*. Chapman & Hall, London.

[15] Liu, L.M. & Hudak, G.B. (1992). *Forecasting and Time Series Analysis using the SCA Statistical System*. Scientific Computing Associates, DeKalb.

[16] Lutkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*, 2nd Ed. Springer-Verlag, Berlin.

[17] Lutkepohl, H. & Poskitt, D.S. (1996). Specification of echelon-form VARMA models, *Journal of Business and Economic Statistics* **14**, 69–79.

[18] Mauricio, J.A. (1995). Exact maximum likelihood estimation of stationary vector ARMA models, *Journal of the American Statistical Association* **90**, 282–291.

[19] Newton, H.J. (1993). Graphics for time series analysis, in *Handbook of Statistics*, Vol. 9, C.R. Rao, ed. Elsevier Science, Amsterdam.

[20] Ord, K. & Lowe, S. (1996). Automatic forecasting, *American Statistician* **50**, 88–94.

[21] Parzen, E. (1985). Multiple time series, in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 719–724.

[22] Poskitt, D.S. & Salau, M.O. (1995). On the relationship between generalized least squares and Gaussian estimation of vector ARMA models, *Journal of Time Series Analysis* **16**, 617–645.

[23] Ravishankar, N.S. & Ray, B. (1998). Bayesian analysis of vector ARFIMA processes, *Australian Journal of Statistics*, to appear.

[24] Reinsel, G.C. (1993). *Elements of Multivariate Time Series Analysis*. Springer-Verlag, New York.

[25] Rycroft, R.S. (1993). Microcomputer software of interest to forecasters in comparative review: An update, *International Journal of Forecasting* **9**, 261–267.

[26] Tiao, G.C. & Tsay, R.S. (1983). Multiple time series modeling and extended sample cross-correlations, *Journal of Business and Economic Statistics* **1**, 43–56.

[27] Tiao, G.C. & Tsay, R.S. (1989). Model specification in multivariate time series (with discussion), *Journal of the Royal Statistical Society, Series B* **51**, 157–213.

[28] Tierney, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley-Interscience, New York.

[29] Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Clarendon Press, Oxford.

[30] Tsay, R.S. (1989). Parsimonious parameterization of vector autoregressive moving average models, *Journal of Business and Economic Statistics* **7**, 327–341.

[31] Tsay, R.S. (1991). Two canonical forms for vector ARMA processes, *Statistica Sinica* **1**, 247–269.

[32] Venables, W.N. & Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York.

[33] Wei, W.S. (1990). *Time Series Analysis: Univariate and Multivariate Methods*. Addison-Wesley, New York.

[34] West, M. & Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.

[35] Wong, H. & Li, W.K. (1997). On a multivariate conditional heteroscedastic model, *Biometrika* **84**, 111–123.

NEVILLE DAVIES

Multiplicative Model

In epidemiology and biostatistics, **relative risk models** are often called *multiplicative models*. In a relative risk model the effect of an exposure or other factors is described as

$$R = R_0 \times RR(z),$$

where R_0 is the background (or baseline) risk and $RR(z)$ is the **relative risk** associated with a **covariate** vector z.

The most commonly used relative risk model is the **loglinear model** $RR(z) = \exp\left(\sum_i \beta_i z_i\right)$. This is a multiplicative function since the effect of each covariate is to multiply the **risk** by a factor proportional to the covariate value. However, additive functions are also useful in describing relative risks. For example, in the assessment of **dose–response** it is often reasonable to describe the relative risk of an exposure in terms of an **additive model** for the **excess relative risk**, i.e. $ERR = RR - 1 = \beta_1 z_1$. If there is an additional exposure of interest, then it is useful to consider additive relative risk models of the form:

$$RR = 1 + \beta_1 z_1 + \beta_2 z_2$$

or

$$RR = 1 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2.$$

The second of these models is a generalization of the multiplicative excess relative risk model

$$RR = (1 + \boldsymbol{\beta}_1 z_1) \times (1 + \boldsymbol{\beta}_2 z_2).$$

Thomas [2] and Breslow & Storer [1] describe general relative risk functions that include both additive and multiplicative models. The articles on **Relative Risk Modeling** and the **Cox Regression Model** contain additional discussion of relative risk models. The articles on **Parametric Models in Survival Analysis** and **Poisson Regression in Epidemiology** present general classes of additive and multiplicative models that are useful in describing excess and relative risks. These articles also discuss methods for parameter **estimation** and **inference** with such models.

References

[1] Breslow, N.E. & Storer, B.E. (1985). General relative risk functions for case–control studies, *American Journal of Epidemiology* **122**, 149–162.

[2] Thomas, D.C. (1981). General relative risk functions for survival time and matched case–control studies, *Biometrics* **37**, 673–686.

DALE L. PRESTON

Multiplicity in Clinical Trials

The simplest randomized **clinical trial** involves the comparison of two treatments with respect to just one outcome measure. Usually, the **null hypothesis** of interest is that there is no true difference in outcome (*see* **Outcome Measures in Clinical Trials**) between the two treatment groups. In this case, the observed difference in outcome between the two treatment groups is evaluated in a statistical test (*see* **Hypothesis Testing**) to generate a ***P* value** that describes the chance of seeing the observed difference, or one more extreme, under the assumption that the null hypothesis is true. Thus, obtaining a P value of 0.01 from the test implies that the difference in outcome observed, or one more extreme, would occur by chance in one out of every 100 clinical trials involving no true difference in the effects of the two treatments. The problem of multiplicity arises whenever the clinical trial departs from this simple design and analysis of two treatments and a single outcome measure. It primarily concerns the interpretation of the multitude of hypothesis tests, often referred to as **multiple comparisons**, that might then be undertaken (*see* **Simultaneous Inference**).

The statistical problem that arises when considering multiple comparisons centers on the error rate that should be controlled. For example, if there are two outcome measures being compared between two treatments, then the conclusion that one treatment is superior to the other might be made if the P value obtained from the comparison of at least one of the outcomes is less than some level, α. The level α is the *marginal* type I error rate, and implies that the probability of concluding that one treatment is superior to the other with respect to that particular outcome measure is no more than α, irrespective of the result for the other outcome measure. In contrast, the **experiment-wise error rate** quantifies the probability of concluding that one treatment is judged superior to the other when the decision is based on data for both outcomes, despite there being no true difference between treatments with respect to either measure. In practice, the decision of superiority is often made when either or both of the marginal P values for the comparisons between the two treatments is less than α. The experiment-wise error rate is then at least α (being equal to α if the outcome measures are perfectly correlated) but may be as high as $2\alpha - \alpha^2$ if the outcome measures are independent. More generally, if K comparisons are made, each at a marginal significance level of α, then the experiment-wise error rate may be as high as $1 - (1 - \alpha)^K$, which is approximately equal to $K\alpha$ if α is small, although the exact value depends on the **correlations** between the outcome measures. The statistical problem is, therefore, aimed at controlling the experiment-wise error rate in the face of multiple comparisons. However, the appropriateness of doing this rather than controlling the marginal error rate specific to each comparison is dependent on the application being considered, and, specifically, the source of the multiplicity, and, even then, there are different methods for controlling the experiment-wise error rate. Cook & Farewell [2] give an excellent overview of the issues.

Controlling Experiment-Wise Error Rates

Global Tests

Consider the situation in which there are K comparisons of interest, each of which can be summarized by one parameter, β_k for $k = 1, \ldots, K$. Then, the global comparison involves a test of the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_k$ vs. the **alternative hypothesis** that at least one of the β_ks differs from the remaining ones. This reduces the multiple comparisons problem to a single test and so preserves the experiment-wise error rate. However, the difficulty usually encountered is that a significant test result does not identify the source of the difference. Thus, further testing of specific comparisons is usually of interest. Furthermore, as the global test seeks any departure from the null hypothesis, it lacks **power** to detect specific patterns of differences that might be of interest. For example, a global test gives equal emphasis to the situation in which β_1 and β_2 differ from the remaining β_ks, regardless of whether they do so in the same or opposite directions, whereas, in some applications, it might be anticipated that the directions of the true departures would be the same. Thus, power is lost in the comparisons. This criticism can be overcome to some extent by using tests that are sensitive to departures from the null hypothesis in the same direction, particularly if they are also of a similar magnitude [9, 14] (*see* **Multiple Endpoints, Multivariate Global Tests**).

Adjusted Marginal Tests

There are a large number of procedures that are based on the marginal tests of each β_k, but that provide criteria for determining significance so as to control the experiment-wise error rate. The most well known of these is the **Bonferroni** adjustment procedure. From each marginal test, a marginal P value is obtained and the null hypothesis is then rejected if this P value is less than α/K, where α is the desired maximum experiment-wise error rate. This procedure is conservative in that the actual experiment-wise error rate may be somewhat less than α, particularly if the quantities being compared are highly correlated. Simes [13] and Hochberg [6] have developed modified Bonferroni procedures that aim to be less conservative (*see* **Multiple Endpoints, *P* Level Procedures**). Other general procedures for marginal testing are available, some of which are described below in the context of specific multiple comparisons problems.

Summary Measures

Rather than working with K comparisons, an alternative approach involves reducing the dimension of the problem by summarizing the data involved in the K comparisons to give just one comparison. Control of the error rate for this one comparison then results directly from the reduction in the dimension of the problem.

Sources of Multiplicity

There are five major sources of multiple comparisons in clinical trials: (i) multiple treatments; (ii) multiple outcome measures; (iii) repeated measurements over time of a specific outcome measure; (iv) comparisons of outcome over subpopulations (subgroups) of subjects; and (v) interim analyses while a trial is ongoing.

Multiple Treatments

The decision to evaluate multiple treatments within a single clinical trial almost always implies some form of structure among the treatments, and so a greater interest in some comparisons than others. For example, it is very rare to design a clinical trial in which one of the treatment arms is not a standard of care, even if that one arm involves a placebo or no treatment. Thus, global tests are rarely appropriate, as they give equal weight to all comparisons and do not exploit the greater interest in particular comparisons.

In the rare circumstances in which there is no such standard, it seems natural that the main objective of the trial should be to define an ordering of treatments from worst to best in terms of some outcome. This procedure, a form of adjusted marginal test procedure, aims to identify groups of treatments such that treatments within each group are not significantly different from each other. Miller [8] gives an example in which the mean response for each of five treatments, A to E, was 16.1, 17.0, 20.7, 21.1, 26.5, respectively. Application of the Newman–Keuls procedure involves comparison of the best and worst treatments, A and E, first; this establishes a significant difference among treatments within the group of five treatments. This is followed by evaluation for differences within the two groups of four treatments, {A, B, C, D} and {B, C, D, E}, then the groups of three treatments, etc. until no further significant differences are found. The result in this example was that there was no significant difference within the group {A, B} nor within the group {B, C, D}. Thus, treatment E differed significantly from all other treatments, and treatment A differed significantly from treatments C and D. The lack of a significant difference between treatments A and B, and between treatments B and C, despite a significant difference between treatments A and C, seems somewhat confusing, but may simply reflect low **power** to detect smaller differences among treatments. This approach is appropriate when there are equal amounts of information for each treatment; the situation is more complex when this is not the case, as the power for detecting differences will vary between different pairs of treatments.

In the more general situation in which there is some structure among treatments, a sequence of tests can usually be defined in order of their importance, and, in some circumstances, marginal tests applied with no adjustment for multiple comparisons are appropriate. Some examples will illustrate the issues. A common use of clinical trials is the simultaneous evaluation of $K-1$ new treatments vs. a standard treatment, as this is considered more efficient than doing the $K-1$ separate clinical trials comparing each new treatment with the standard (though this might also depend on the practicalities of undertaking

a large trial vs. a small one). Dunnett [4] provided a procedure for adjusting each of the $(K-1)$ marginal tests to control the experiment-wise error rate. However, it is debatable whether the error rate for the comparison of treatment A vs. the standard should be affected by the fact that treatment B has also been compared with the standard. Indeed, if two separate trials had been undertaken, then each would be reported separately, with no adjustment to significance levels to reflect the existence of the other trial. Proschan & Follmann [11] discuss this issue and show that there is little difference in the proportion of marginal tests giving type I errors, whether the new treatments are compared with the standard in a single trial or in different trials. This supports the idea of not adjusting marginal test results in this situation. However, having established that some of the new treatments are superior to the standard, it might then be of interest to investigate the evidence for differences among these treatments. In this case, there might be more of a rationale for controlling the experiment-wise error rate among the comparisons in that subset, using an adjusted marginal test approach such as the Newman–Keuls procedure discussed above.

Related to this first example is a second example, in which the $K-1$ new treatments being compared to a standard treatment are different doses of the same drug. In this case, the problem can often be considered as a sequence of two questions: first, is there evidence that the drug is superior to the standard treatment (often, in this context, a placebo) and, secondly, is there a difference in effect between doses? Rather than test each dose against the placebo, it is usually better to compare a summary measure of the response obtained across all $K-1$ doses to the response in the placebo arm. Then, if this establishes an effect of the new treatment, one would test for an association between magnitude of response and dose.

In general, the latter test would also be undertaken using a summary measure, usually defined by some prior knowledge of a model that is likely to describe the association between response and dose. Although there are two hypotheses being assessed, there is little rationale for adjusting for multiple comparisons, as the hypothesis about a **dose–response** is likely to be secondary in nature, dependent on an effect of the new treatment being first established.

A third example concerns the evaluation of a combination of treatments. A trial of the combination of two antiretroviral drugs that target the human immunodeficiency virus infection, ZDV + ddI, was undertaken to compare that combination with each of ZDV and ddI separately. The doses of ZDV and ddI used in the combination arm were the same as in the two monotherapy arms, and so it was anticipated that greater toxicities would be seen in the combination arm, which was also more expensive. Thus, the combination treatment might only be recommended if it was shown to be superior to each of the two monotherapies. In this case, the error rate of interest is the probability of recommending ZDV + ddI when there is no difference between it and either ZDV or ddI alone. This probability is less than α if the marginal pairwise comparison of ZDV + ddI vs. ZDV and that of ZDV + ddI vs. ddI is undertaken using a level of α, and so no adjustment for multiple comparisons is required.

These examples help to illustrate the fact that control of the experiment-wise error rate is rarely of interest in clinical trials involving multiple treatments. Instead, more specific sequencing of hypotheses is often possible with little, if any, control of error rates necessary. Indeed, the **sample size** and power considerations of most well-designed clinical trials are dictated by some prioritization of hypotheses, which should then determine whether any error rate control is necessary.

Multiple Outcome Measures

For many diseases, there may not be a single obvious measure of outcome (*see* **Outcome Measures in Clinical Trials**). This is particularly common in trials of symptomatic diseases such as arthritis or neuropathy in which pain or other measures at different joints or in different muscles are of interest. However, it also arises in studies of diseases that have major morbidity outcome measures, such as the occurrence of both strokes and myocardial infarctions, as well as death, in cardiovascular trials. The important consideration in determining how to address the multiple comparisons problem in this context concerns how similar the various outcome measures are to one another. In the case of joint or muscular pain, the impact on a patient's **quality of life** may be similar regardless of the joint or muscle affected. In contrast, a nondebilitating stroke or myocardial infarction is clearly less important than death.

Consider a clinical trial that is designed to compare two treatments with respect to K outcome measures. Denote the true difference between treatments by β_k for $k = 1, \ldots, K$. Then the null hypothesis of interest might be $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Global tests might be considered if the outcome measures are similar in their clinical significance and the magnitude of the treatment effect is likely to be similar for each measure, so that the alternative hypothesis is $H_1 : \beta_1 = \beta_2 = \cdots = \beta_k = \beta$ for some β. O'Brien [9], Pocock et al. [10] and Wei et al. [15] present tests for this situation. These tests effectively involve a weighted average of the outcome measures, where the weights are chosen to optimize the power of the test to detect a treatment difference. If it is likely that the treatment differences vary among outcome measures, particularly if the directions of the differences might differ, then comparisons between treatments with respect to each outcome measure are likely to be more important. In this case, Follmann [5] has shown that the Bonferroni procedure applied to the K outcome measures is to be preferred, particularly when it is unknown which outcome measure is likely to have the largest treatment difference. Note that more generic global tests such as **Hotelling's T^2** test (a multivariate extension of the two-sample t test) or **chi-square tests** for $2 \times K$ **contingency tables** are rarely useful in clinical trials, as they focus on any departure from the null hypothesis, including those in which the directions of the differences (i.e. the signs of the β_ks) might differ.

Cook & Farewell [2] give a very nice critique of a colorectal cancer trial in which there were two outcome measures, tumor response and survival, of different clinical significance. In this case, the rationale for any adjustment for multiple comparisons, whether using global tests or an adjusted marginal testing approach, is weak. Specifically, global tests are of little interest because one wishes to understand how the treatments affect each of the two measures separately. Thus, marginal tests are more appropriate. However, Cook & Farewell argue against adjustment to marginal tests, on the basis that the implications for clinical practice from finding, for example, very strong evidence in favor of a tumor response and weak evidence in favor of survival, are very different from finding weak evidence in favor of a tumor response but very strong evidence in favor of survival. Thus, the relevance of an experiment-wise error rate is questionable. Indeed, in this type of situation, it might be more useful to consider a hierarchy of hypotheses that focus on the outcomes in a decreasing order defined by their clinical importance. In addition, composite outcome measures might be useful. For example, in anti-HIV trials, the first hypothesis of interest might be whether the treatments differ with respect to mortality. If they do, then other differences might be of lesser interest. If they do not, then it is natural to investigate treatment differences with respect to the composite endpoint of death or progression to AIDS, as this reflects a clinically serious outcome (AIDS) or a worse outcome (death), rather than investigate progression to AIDS separately from death.

An alternative approach for handling multiple outcome measures involves the use of summary measures. As an example, Salsburg [12] describes a clinical trial of treatment for acute painful diabetic neuropathy. In this trial, each patient was asked to score pain, numbness and weakness in each of their left and right feet, calfs, thighs, hands and arms giving 30 outcome measures in all. Instead of analyzing these 30 outcomes in a global test, or using adjusted marginal tests, two summary measures were identified for analysis. The first, the "maximum distress" score, was the maximum score across all 30 measures. The second, the "dominant symptom" score, was the score for the symptom with the worst score prior to starting study treatment. These two measures were considered clinically relevant, reflecting the most significant aspects of the disease to each patient. They also provided a more powerful analysis than that obtained by looking at all 30 measures separately, because many patients showed no distress throughout the study for certain measures.

Repeated Measurements over Time

It is very common to follow patients over time with repeated measurements of outcome (*see* **Longitudinal Data Analysis, Overview**). For example, clinical trials of antihypertensive agents often collect blood pressure measurements over time. Global tests are rarely of interest in this context. Adjusted marginal tests are also of limited value as they ignore the explicit structure of the data, and tend to be conservative because of the correlation between successive measurements. As with multiple outcome measures, this is a situation where summary measures

are often particularly valuable. For example, in antihypertensive trials, there might be interest in whether the trend in blood pressure over the duration of the study differs between treatments. In this case, the trend might be calculated for each patient using standard methods for **linear regression**, and then the average of the trends across patients compared between treatments using, for example, a two-sample t test (*see* **Student's *t* Statistics**). This is, therefore, a two-step process. Alternatively, the analysis can be done in one step, using methods for mixed effects models [7]. For example, such a model might involve a linear trend for each subject, with a further level of the model describing how these trends vary among subjects and their dependence on treatment assignment. In other applications; for example, anti-HIV trials, there might be interest in the maximal extent of viral suppression achieved and also in the durability of effect, often expressed as the change from baseline to the level at about one year after treatment started. Thus, an important advantage of the summary measure approach is that it focuses the analysis on aspects of the data that are considered most clinically relevant. However, summary measures are sometimes used in order to gain statistical power, without adequate care being given to their interpretation. For example, in anti-HIV trials, the area under the curve, formed by joining the results of successive measurements over the duration of the trial and bounded by the pretreatment level, is sometimes recommended (*see* **Bioequivalence**). However, this measure is effectively a time-weighted average and does not distinguish short-term effects from long-term effects, so that statistical significance might be obtained at the expense of clinical relevance.

An alternative approach to the analysis is to calculate a test statistic at each measurement time, and then to combine the test statistics into one value (taking into account the fact that they are correlated), as suggested by Wei & Johnson [14]. This reduces the analysis to a single test. However, this method also has the disadvantage that it does not distinguish short-term effects from long-term effects or other patterns of change that might be of clinical interest.

Subgroup Analyses

Common secondary analyses involve the investigation of whether differences between treatments vary between different subgroups of the population studied, where the subgroups are defined by the values of some covariate. From a multiple comparisons perspective, there are two problems. First, there may be many possible subpopulations defined by a particular covariate for which differences might be evaluated. Collins et al. [1] showed the hazards of this very nicely in an example in which they used the zodiac birth sign as a means of defining subpopulations and showed that the relative effect of treatment was "significant" for patients born under Scorpio, but was not for the other birth signs combined, almost undoubtedly a chance finding reflecting a type I error. This type of analysis is inappropriate. Instead, a single test of **interaction** should be undertaken to assess whether the relative effect of treatment does vary across subgroups defined by the covariate (*see* **Treatment-covariate Interaction**).

The second problem is that there may be many **covariates** used in defining different divisions of the study population. Adjusted marginal tests (applied to the tests of interaction defined by each covariate) should be considered, particularly if there is no prior reason for anticipating a difference between subpopulations. In the latter case, prespecification of the subpopulations for subgroup analyses might help as a means of identifying potential subgroups of interest. However, it is important to appreciate that prespecification is only desirable for a very limited number of subpopulations (one or two), and that prespecification of an extensive list of subpopulations does not avoid the multiple comparisons problem.

Interim Analyses

Interim analyses are sometimes undertaken during the conduct of the trial for ethical (*see* **Ethics of Randomized Trials**) and cost reasons, with the idea that a trial might be modified or terminated early if significant treatment differences are found. The successive analyses will lead to an increase in the error rate. In extreme, with continuous monitoring, the error rate can be inflated to a very high level so that a "significant" difference will be obtained with very high probability, even if there is no true difference; what Cornfield [3] termed "sampling to a foregone conclusion". Thus, control of the error rate in the face of the multiple comparisons is desirable. Methods specific to this application have been developed and are

discussed in detail elsewhere (*see* **Data and Safety Monitoring**).

Closing Remarks

The multiple comparisons problem is a major issue in the interpretation of results from clinical trials. The discussion above shows that, in many instances, the issue is better addressed by defining the specific questions of greatest clinical interest. In this way, the dimension of the multiple comparisons problem can often be reduced, and a study better designed to address these questions. When multiple comparisons are still required, then prespecification of a limited number of comparisons of primary interest is wise. In many circumstances, the analysis can proceed without further adjustment for multiple comparisons by determining that it is the marginal error rates that are of prime interest rather than the experiment-wise error rates. This requires that thought be given to the decisions that might follow from the results of the trial, and to whether these decisions are to be made on the basis of marginal hypotheses or upon the collection of hypotheses, respectively.

Although the focus has been on **hypothesis testing** and error rates, the problem of multiple comparisons is also relevant to interval estimation (*see* **Estimation, Interval**). Specifically, whenever it is considered that some control of error rates across comparisons is desirable, then similar arguments apply in the construction of **confidence intervals**.

References

[1] Collins, R., Gray, R., Godwin, J. & Peto, R. (1987). Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews, *Statistics in Medicine* **6**, 245–250.

[2] Cook, R.J. & Farewell, V.T. (1996). Multiplicity considerations in the design and analysis of clinical trials, *Journal of the Royal Statistical Society, Series A* **159**, 93–110.

[3] Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle, *American Statistician* **61**, 577–594.

[4] Dunnett, C. (1957). A multiple comparisons procedure for comparing several treatments with a control, *Journal of the American Statistical Association* **50**, 1096–1121.

[5] Follmann, D. (1995). Multivariate tests for multiple endpoints in clinical trials, *Statistics in Medicine* **14**, 1163–1175.

[6] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

[7] Laird, N. & Ware, J. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[8] Miller, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Ed. Springer-Verlag, New York.

[9] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1087.

[10] Pocock, S.J., Geller, N.L. & Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics* **40**, 487–498.

[11] Proschan, M.A. & Follmann, D.A. (1995). Multiple comparisons with control in a single experiment vs. separate experiments: why do we feel differently?, *American Statistician* **49**, 144–149.

[12] Salsburg, D.S. (1992). *The Use of Restricted Significance Tests in Clinical Trials*. Springer-Verlag, New York.

[13] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **73**, 751–754.

[14] Wei, L.J. & Johnson, W.E. (1985). Combining dependent tests with incomplete repeated measurements, *Biometrika* **72**, 359–364.

[15] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

MICHAEL D. HUGHES

Multistage Carcinogenesis Models

Cancer is a disorder of cells whereby a visible tumor is the end result of a whole series of changes which may have taken many years to develop. Cancers generally derive from the clonal expansion of a single cell (monoclonal) that is dramatically altered by the series of events.

To understand the process of carcinogenesis, the story must start at the beginning – normal cells. A normal cell has a well-defined shape and is organized within its environment of other normal cells. Growth (cell division or replication) is dictated by the stimulatory and inhibitory signals of the environment, which are normally in balance until a growth stimulus is required. In normal development and growth, growth control allows individual organs (e.g. heart, liver, lungs) to reach a specific size which is homeostatically maintained.

The process of replication brings with it the risk of mutations. Mutations may be thought of as permanent alterations in DNA (occurring within all or part of the DNA of a cell) that can impair the regulatory communication between the cell and its environment. The most generally accepted mechanism is as follows. A single mutation alters the physical nature of the cell, making it less responsive to external stimuli, resulting in frequent cell division. As genetic damage accumulates, the damaged cell becomes deaf to external stimuli. Lack of external influence eventually results in uncontrolled replication, characteristic of malignancy, and the resulting tumor (clonal mass of mutated cells) damages healthy tissue in its neighborhood or metastasizes where it may establish new colonies at distant sites. Other mechanisms exist including loss of genetic material, alterations in cellular death, alterations in cellular communication not related to mutations, and alteration in mitochondrial DNA. The net effect in all cases is general loss of homeostatic control of cellular division, growth of a tumor, and resulting damage to surrounding tissue (*see* **Cell Cycle Models**).

One hundred and forty years ago, Johannes Mueller, a German microscopist, demonstrated that cancers were made up of cells. This discovery initiated a search for the specific differences between normal and cancer cells. By 1914, the German cytologist, Theodor Boveri, concluded that malignant cells had atypical chromosomes and that any event leading to such abnormality would cause cancer. Advances in biological technology, especially in the fields of cellular and molecular biology, have identified many genes that take part in the progression from normalcy to cancer.

The process of carcinogenesis is inherently probabilistic, at least as long as it is unknown why certain individuals are afflicted with cancer under certain conditions and others are unaffected. Attention will be focused on stochastic models of carcinogenesis at the cellular level since one of the least understood aspects of tumor development is the **latent period** between cancer initiation and the appearance of tumors. Mathematical models of carcinogenesis strive to investigate the number and types of events in the progression from normalcy to malignancy and allow examination of hypothetical schemes that may be tested objectively.

There are two basic concepts that have been used in describing the events leading to carcinogenesis: hit theory and multistage theory. The biological hypothesis behind the hit theory of carcinogenesis is that a cell must be damaged a certain number of times before it loses growth control and becomes tumorigenic. The damage to the cell is thought to be caused by particles of the carcinogen hitting the nucleus of the cell. The damage incurred is dependent on the number of hits the cell receives and the dose of the carcinogenic agent. A majority of the literature on hit theory modeling comes from the area of biophysics where interest has centered on the interaction between **radiation** and target cells with respect to mutagenicity. Hit theory directly related to modeling the process of carcinogenesis does not have a pronounced history. Figure 1 displays a three-hit model of carcinogenesis.

The first mathematical model of carcinogenesis was that of Iversen & Arley [5]. Their model postulated that carcinogenic "hits" are independently and randomly distributed among all normal cells of a tissue. Each normal cell hit by the carcinogen undergoes an irreparable change which marks the onset of the cancer process. This model is frequently referred to as the "one-hit" model of carcinogenesis because only a single hit is necessary for a cell to undergo a mutation which will eventually lead to malignancy. Once the cell becomes mutated, it is assumed to lose growth control and proliferates

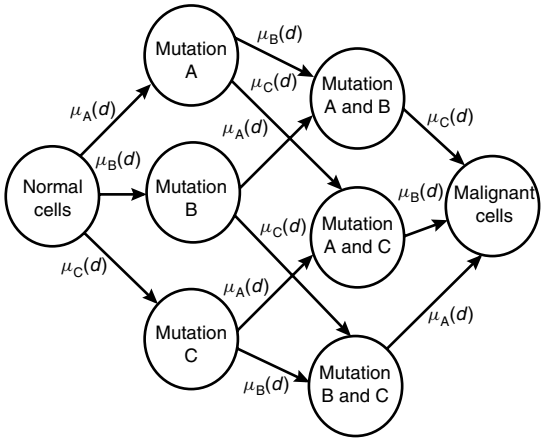

Figure 1 Three-hit model of carcinogenesis [16]. In this model, mutation rates are denoted by μ_i, ($\mu_A = \mu_B = \mu_C = \mu$). All rates are expressed as number of mutations per unit dose of carcinogen (denoted by d) for a fixed period of time

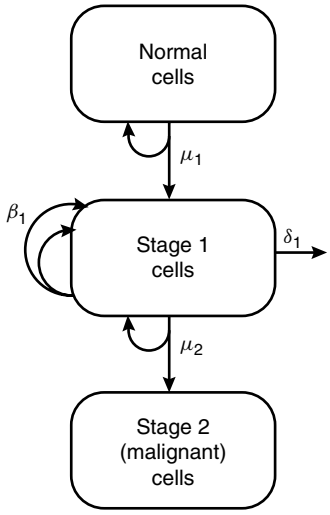

Figure 2 Two-stage model of carcinogenesis. In this model, mutation rates are denoted by μ_i, birth rates are denoted by β_i, and death/differentiation rates are denoted by δ_i. All rates are expressed as number of events per cell per unit of time

via replication. An observable tumor results when enough replicated cells have amassed to be clinically detectable.

Stochastic Models of Carcinogenesis

Many years after the development of Iversen & Arley's stochastic cancer model, Rai & Van Ryzin [16] resurrected the underlying theme of the one-hit model and adapted it to include more than a single hit, i.e. a multihit model. The biological hypothesis behind the multihit model is that a normal cell must be damaged a multiple number of times before it results in a malignant cell. The amount of damage that is incurred is dependent on the number of hits the cell receives and the dose of the carcinogenic agent. Rai & Van Ryzin further assumed that, once a cell has been subjected to at least j hits, it becomes malignant and will eventually result in a tumor. Unlike Iversen & Arley [5], they did not model the growth process of malignant cells. Mathematically, it was assumed that once a cell received j hits it instantaneously became an observable tumor.

The multihit model has been used to model the occurrence of cancer in a variety of tissues; however, it is not clear from Rai & Van Ryzin's mathematical derivation that their theory applies to entire tissues as it does to an individual cell. Disregarding this caveat, the "hit" theory of carcinogenesis was still not well received even after Rai & Van Ryzin's development of a generalized theory, and was generally abandoned at this point. This was most likely due to the perceived simplistic nature of the "hit" theory model and a lack of plausibility relative to the multistage theory of carcinogenesis.

The multistage theory of carcinogenesis also assumes several events leading to DNA damage; however, it is hypothesized that these events must occur in a particular sequence. In essence, the multistage model is an order-restricted multihit model. This theory was initially conceptualized by Muller [10] and Nordling [12] from the observation that for some carcinomas the cancer incidence rate rapidly increased with increasing age. Multistage theory continues to be a popular concept since current biological evidence suggests that genetic changes usually occur in a specific order. Figure 2 displays a two-stage model of carcinogenesis.

The two-stage model shown in Figure 2 assumes that a normal cell must pass through two unique, sequential stages before becoming malignant. This model has three types of cells: normal cells, stage-one cells, and stage-two (malignant) cells. In the small time interval $[t, t + \Delta t)$, the following events may occur:

1. A normal cell may acquire a mutation resulting in damage to a single strand of the DNA which results in one normal cell and one stage-one cell with probability $\mu_1 \Delta t + o(\Delta t)$.
2. A stage-one cell may replicate, resulting in two stage-one cells with probability $\beta_1 \Delta t + o(\Delta t)$.
3. A stage-one cell may differentiate or die, i.e. leave the system, with probability $\delta_1 \Delta t + o(\Delta t)$.
4. A stage-one cell may acquire a mutation, resulting in damage to a single strand of the DNA which results in one stage-one cell and one stage-two (malignant) cell with probability $\mu_2 \Delta t + o(\Delta t)$.

The probability of more than one event occurring in this small time interval is $o(\Delta t)$.

For the model shown in Figure 2 (and most classes of multistage models used to date) the growth of normal cells is assumed to be constant or deterministic. In the context of the model, it is assumed that the number of normal cells at any time t is constant. All intermediate cell types (in this case, stage-one cells) are assumed to undergo growth kinetics via a linear birth–death process (*see* **Stochastic Processes**). A linear birth–death process implies that the rate of growth of a cell population is proportional to the number of cells in the tissue. Further modeling assumptions are that the birth–death processes and mutation processes are stochastic and independent of one another. In addition, each cell acts independently of other cells, the transformation process is irreversible, i.e. damage to the genome is "fixed", and once a malignant cell is produced it loses growth control and will eventually result in a tumor. Mathematically, these assumptions imply that the model portrays the process of carcinogenesis as a **Markov process**. A Markov process describes the fate of any cell at time t as depending only on the present state of the cell at time t and not on the past history of that cell. More precisely, this model may be described as a continuous-time multiple **branching process** since all cell types (with the exception of malignant cells) implement growth kinetics that spawn birth–death processes from which the progeny form branching processes.

Mathematically, the main outcome studied in the context of these mathematical models of carcinogenesis is the time-to-first-entry into the malignant state, generally referred to as the **tumor incidence** rate. Let T be the associated random variable, in which case tumor incidence is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr[T \in [t, t + \Delta t) | (T \geq t)]}{\Delta t}. \qquad (1)$$

This is generally converted into a cumulative distribution function (CDF) for tumor onset by the formula

$$\Pr(T < t) = 1 - \exp\left[- \int_0^t \lambda(s)\, \mathrm{d}s \right]. \qquad (2)$$

For the simple two-stage model of carcinogenesis in Figure 2, several authors have derived a closed-form solution for the tumor incidence rate for time-constant rate parameters (Kopp–Schneider et al. [6] and Zheng [21]). The solution is given as

$$\Pr(T \leq t) = 1 - \exp[-\Lambda(t)], \qquad (3)$$

where

$$\Lambda(t) = \left(\frac{X_0 \mu_1}{\beta} \right) \left[\frac{t}{2} (\beta - \delta - \mu_2 + R) + \log \left(\frac{(\delta - \beta + \mu_2 + R) + (\beta - \delta - \mu_2 + R) \exp(-Rt)}{2R} \right) \right], \qquad (4)$$

where

$$R = [(\beta + \delta + \mu_2)^2 - 4\beta\delta]^{1/2} \qquad (5)$$

The most general formulation for the CDF is derived by Portier et al. [15]. They use the Kolmogorov backwards equations to develop a system of ordinary differential equations (ODEs) which, through a simple algebraic manipulation, can be used to derive (2) for any nonhomogeneous multistage model of carcinogenesis. The model is still required to be stochastically linear (the rate constants cannot depend upon the numbers of cells in each stage of the process). If we expand the two-stage model in Figure 2 to include a birth–death process on the normal cells (rates $\beta_0(t)$ and $\delta_0(t)$ for birth and death of normal cells, and rates $\beta_1(t)$ and $\delta_1(t)$ for stage-one cells), then the ODEs derived by Portier et al. [15] are

$$\frac{\mathrm{d}}{\mathrm{d}s} \Psi_0(s) = \beta_0(t - s)[\Psi_0(s)]^2 + \delta_0(t - s) + \mu_1(t - s)\Psi_0(s)\Psi_1(s) - [\beta_0(t - s) + \delta_0(t - s) + \mu_1(t - s)]\Psi_0(s)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}s}\Psi_1(s) = \beta_1(t-s)[\Psi_1(s)]^2 + \delta_1(t-s) - [\beta_1(t-s) + \delta_1(t-s) + \mu_2(t-s)]\Psi_1(s), \quad (6)$$

where the initial conditions are $\Psi_0(0) = 1$ and $\Psi_1(0) = 1$. The CDF for tumor incidence is calculated by solving this system from $s = 0$ to $s = T$ and plugging the solutions into the calculation

$$\Pr(T \leq t) = [\Psi_0(t)]^{m_0}[\Psi_1(t)]^{m_1}, \quad (7)$$

where m_i is the initial number of cells in stage i of the process at time $t = 0$. A detailed derivation of these ODEs would be inappropriate in this context; interested readers should refer to the manuscript by Portier et al. [15] for the details.

Even without the details, it is possible to develop systems of ODEs intuitively for more complex multistage models. Examining the form of system (7) relative to the form of the model in Figure 2, it is possible to illustrate the pattern of these equations. Starting with the end of (6) first, it is clear that in the equations pertaining to $\Psi_i(s)$, the rates of the process by which cells move out of state $i[\beta_i(t-s), \delta_i(t-s)$, and $\mu_{i+1}(t-s)]$ are summed, multiplied by $\Psi_i(s)$ and subtracted from the differential equation. The remaining terms in the differential equation for $\Psi_i(s)$ are the product of each rate for cells leaving the state i times the $\Psi.(s)$ for the eventual location of the resulting cell(s). These terms are all added to the differential equation. For example, a birth results in two cells returning to the state in which the birth occurs. For state i, the resulting product to be added to the differential equation for $\Psi_i(s)$ is $\beta_i(t-s)\Psi_i(s)\Psi_i(s)$; that is the rate for the event of birth for the proper time, $\beta_i(t-s)$, times the **generating functions** for the states of the two resulting cells, $\Psi_i(s)$ and $\Psi_i(s)$. A mutation from state i results in one cell returning to state i and the next cell going on to state $(i+1)$ so the resulting product to be added to the differential equation for $\Psi_i(s)$ is $\mu_i(t-s)\Psi_i(s)\Psi_{i+1}(s)$. Note that, for the state just prior to the malignant state (state 1 in the two-stage model), since the function for the final state [$\Psi_2(s)$ in the two-stage model] is identically zero at all times, this term drops out of the system. Finally, since death/differentiation simply removes a cell and does not place it into any state being followed by the system, the proper term to add to $\Psi_i(s)$ for a death is simply $\delta_i(t-s)$. The calculation of the CDF is a direct extension of (7) to include all stages in the more complicated model.

The most important aspect of this modification to the determination of the CDF for tumor onset is the ability to consider much more complicated and realistic models (see below) and to incorporate biochemical and pharmacological events into the determination of rate constants for the model (see Portier et al. [15]).

Towards More Realistic Models

The hit theory and multistage theory have played dominant roles in the mathematical modeling of carcinogenesis. The history of carcinogenic modeling can be described as a hierarchy of models within a respective framework, i.e. hits or stages. Generally, each newly developed model encompasses the previously developed models. Thus, mathematical models attempt to include the evolution of biological evidence in cancer biology. A natural extension in the mathematical modeling of carcinogenesis is the development of a single model which incorporates concepts from both hit theory and multistage theory. This class of models still embodies all of the mathematical models constructed under the multihit and multistage paradigms, and thus the history of carcinogenesis modeling is preserved, while simultaneously current experimental evidence is being incorporated into this class of models. In essence, a natural extension in the continuum of the mathematical modeling of carcinogenesis is being implemented. This class of models is referred to as the multipath/multistage models of carcinogenesis.

In fusing the hit theory and multistage theory of carcinogenesis, it is important to understand the notions of stages and hits in the context of the multipath/multistage model. Stages will be defined as necessary events for carcinogenesis that must occur in a specific order. Conversely, hits are defined as events that have no specific ordering and no direct bearing on carcinogenesis; however, they may augment the rate at which a stage occurs. Consequently, by definition, hits yield alternative pathways to cancer. Figure 3 displays a two-path/three-stage model of carcinogenesis. There are two possible scenarios

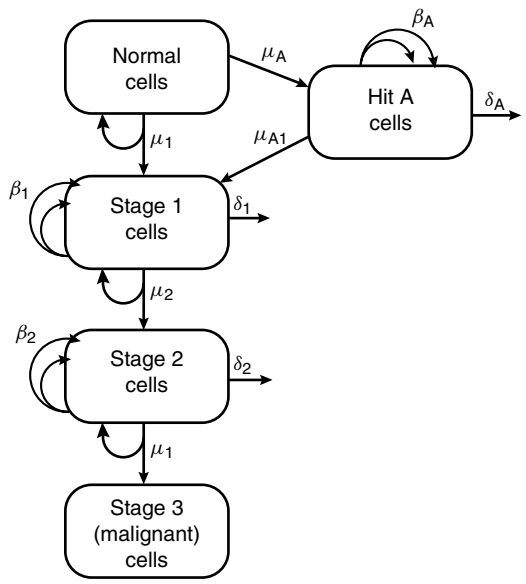

Figure 3 Two-path/three-stage model of carcinogenesis. In this model, mutation rates are denoted by μ_i, birth rates are denoted by β_i, and death/differentiation rates are denoted by δ_i. All rates are expressed as number of events per cell per unit of time

for a normal cell to be transformed into a malignant cell:

1. A normal cell may undergo three mutational events: transformation from the normal state to stage one (rate μ_1), transformation from stage one to stage two (rate μ_2), and then transformation from stage two to the malignant state (rate μ_3). This is the most direct path to carcinogenesis where three stages are traversed.
2. A normal cell may undergo four mutational events: transformation from the normal cells to hit A cells (rate μ_A), transformation to stage one (rate μ_{A1}), transformation from stage one to stage two (rate μ_2), and then transformation to the malignant state (rate μ_3).

In a modeling context, Figure 3 is a four-stage model added to a three-stage model since hits and stages are mathematically indistinguishable. However, biologically this is not simply a fourth stage added to a simple three-stage model, but a construct based on some observations regarding certain carcinogenic mechanisms. Because the hit A cells still lead to stage-one cells, this state does not really constitute a stage by the definition given. It is more closely related to a hit since passage through this stage in moving to stage one is not required, but does alter the overall mutation rate.

Experimental evidence for the multipath/multistage model is supported by current cancer research in the area of oncogenes and tumor suppressor genes. Oncogenes are thought to be genes whose activation accelerates replication. Tumor suppressor genes are thought to act in the opposite manner; they are genes whose deactivation removes some restrictions on the mechanism that regulates cell proliferation. Thus, if oncogenes are activated and tumor suppressor genes deactivated, the net result is believed to be a cell, and eventually a colony of cells, with little or no growth control (malignancy) (*see* **Tumor Growth**).

Current biological theory in the area of molecular carcinogenesis suggests that a malignant cell results from the accumulation of genetic damage to a single cell. The multipath/multistage model may possibly explain the underlying mechanisms involved in the transformation of the mechanisms by which the oncogenes and suppressor genes control replication. Three equally likely possibilities exist:

1. Oncogene activation and suppressor gene deactivation must occur in a sequential manner and induce carcinogenesis. This situation would fit the multistage theory of carcinogenesis. This theory includes models such as those by Armitage & Doll [1, 2], Neyman & Scott [11], Moolgavkar & Venzon [9], and Portier & Kopp-Schneider [14].
2. Oncogene activation and suppressor gene deactivation are not restricted to a particular order of occurrence. Thus, carcinogenesis is induced once both these events occur, regardless of order. This would directly relate to the multihit theory hypothesis. Models in this class have been proposed by Iversen & Arley [5] and Rai & Van Ryzin [16].
3. One of the events in the process, say oncogene activation, could have no direct bearing on carcinogenesis such that it is unnecessary for tumor formation. However, it may still alter the rate at which one of the other events, say suppressor gene deactivation, occurs. Thus, oncogene activation could be considered as a potential hit which augments suppressor gene deactivation.

Because suppressor gene deactivation is necessary for carcinogenesis, it is a stage in the process. Models in this class have been proposed by Portier [13] and Tan [20], and developed by Sherman & Portier [17].

Estimation Considerations

Historically, mathematical models related to the cancer process have relied on tumor response data, i.e. the presence or absence of a tumor, for parameterization. However, tumor response data are not sufficient to uniquely parameterize the simplest of mathematical cancer models. Mathematical and statistical techniques have been derived over the past several years to take advantage of some of the intermediate cancer biomarker data currently being collected [3, 4, 6, 19]. Premalignant lesion data from rodent skin papilloma studies (number of skin papillomas) and hepatocarcinogenicity studies (number and size of enzyme-altered hepatic lesions) have been used to elucidate the underlying cancer mechanisms of a variety of chemical carcinogens.

Mathematical models have also been developed to focus strictly on the growth properties of premalignant lesions. From cell labeling studies carried out over a period of time (incidence labeling data), Moolgavkar & Luebeck [8] have developed methods to estimate the birth rate of premalignant cells. Lyles [7] incorporates incidence and prevalence cell labeling data (BrdU cell labeling data and PCNA cell labeling data, respectively) to estimate the rate parameters of the cell cycle. From these methods, one may test a variety of hypotheses which may elucidate aberrant cell growth typically characterized by premalignant cell populations.

Mathematical models of carcinogenesis are not limited to using a single type of data (i.e. tumor response data alone or labeling index data alone) in the modeling process. Several pieces of information may be incorporated into a single model to more fully describe the cancer process or fill in the gaps created by previous models. Important aspects of this approach are its close ties to the underlying biology of the cancer process and the enhancement of statistical **power** in **hypothesis testing** (due to the use of additional data). Once a model and data are chosen, one may use **maximum likelihood** techniques to arrive at parameter **estimates** and **likelihood ratio tests** to examine a broad range of hypotheses.

References

[1] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

[2] Armitage, P. & Doll, R. (1957). A two-stage theory of carcinogenesis in relation to the age distribution of human cancer, *British Journal of Cancer* **11**, 161–169.

[3] Dewanji, A., Moolgavkar, S. & Luebeck, E. (1991). Two-mutation model for carcinogenesis: Joint analysis of premalignant and malignant lesions, *Mathematical Biosciences* **104**, 97–109.

[4] Dewanji, A., Venzon, D. & Moolgavkar, S. (1989). A stochastic two-stage model for cancer risk assessment. II. The number and size of premalignant clones, *Risk Analysis* **9**, 179–187.

[5] Iversen, S. & Arley, N. (1950). On the mechanism of experimental carcinogenesis, *Acta Pathologica et Microbiologica Scandinavica* **27**, 773–803.

[6] Kopp-Schneider, A., Portier, C. & Sherman, C. (1994). The exact formula for tumor incidence in the two-stage model, *Risk Analysis* **14**, 1079–1080.

[7] Lyles, C. (1996). The modeling of cell proliferation: Incorporating the cell cycle. *Unpublished doctoral dissertation*, Department of Biostatistics, University of North Carolina, Chapel Hill.

[8] Moolgavkar, S. & Luebeck, E.G. (1992). Interpretation of labeling indices in the presence of cell death, *Carcinogenesis* **13**, 1007–1010.

[9] Moolgavkar, S. & Venzon, D. (1979). Two-event models for carcinogenesis: Incidence curves for childhood and adult tumors, *Mathematical Biosciences* **47**, 55–77.

[10] Muller, H. (1951). Radiation damage to the genetic material, *Science Progress* **7**, 93–493.

[11] Neyman, J. & Scott, E. (1967). Statistical aspects of the problem of carcinogenesis, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 745–776.

[12] Nordling, C. (1953). A new theory on the cancer inducing mechanism, *British Journal of Cancer* **7**, 68–72.

[13] Portier, C. (1987). Statistical properties of a two-stage model of carcinogenesis, *Environmental Health Perspectives* **76**, 125–131.

[14] Portier, C. & Kopp-Schneider, A. (1991). A multistage model of carcinogenesis incorporating DNA damage and repair, *Risk Analysis* **11**, 535–543.

[15] Portier, C., Kopp-Schneider, A. & Sherman, C. (1996). Calculating tumor incidence rates in stochastic models of carcinogenesis, *Mathematical Biosciences* **135**, 129–146.

[16] Rai, K. & Van Ryzin, J. (1981). A generalized multihit dose-response model for low dose extrapolation, *Biometrics* **37**, 341–352.

[17] Sherman, C. & Portier, C. (1994). The multipath/multistage model of carcinogenesis, *Informatik Biometrie und Epidemiologie in Medizin und Biologie* **25**, 250–254.

[18] Sherman, C. & Portier, C. (1995). Quantitative analysis of multiple phenotype enzyme-altered foci in rat hepatocarcinogenesis experiments: The multi-path/multistage model of carcinogenesis, *Carcinogenesis* **16**, 2499–2506.

[19] Sherman, C., Portier, C. & Kopp-Schneider, A. (1994). Multistage models of carcinogenesis: An approximation method for the size and number distribution of late-stage clones, *Risk Analysis* **14**, 1039–1048.

[20] Tan, W. (1991). *Stochastic Models of Carcinogenesis*. Marcel Dekker, New York, pp. 135–212.

[21] Zheng, Q. (1994). On the exact hazard and survival functions of the MVK stochastic carcinogenesis model, *Risk Analysis* **14**, 1081–1084.

(*See also* **Extrapolation, Low Dose**; **Serial-sacrifice Experiments**; **Tumor Incidence Experiments**)

CLAIRE D. SHERMAN &
CHRISTOPHER J. PORTIER

Multistage Sampling

Multistage sampling is an extension of **cluster sampling**. The sampling units are hierarchically arranged: primary, secondary, tertiary, etc. units are established in accordance with the number of stages in the multistage sampling design. For simplicity of illustration, we consider a population that has been divided into primary sampling units (PSUs), where the PSUs are initially unstratified. A frame of M PSUs is assumed (*see* **Sampling Frames**). A first-stage sample of m PSUs is randomly selected from this frame. The population within the ith sampled PSU is itself divided into N_i secondary units, and a sample cluster of n_i units is randomly selected from this frame. In multistage sampling $n_i < N_i$ as opposed to cluster sampling where $n_i = N_i$. The sampling within each PSU is independent of the sampling in other PSUs and may be carried to any number of stages. One may use any **probability sampling** method (such as **stratified sampling** or cluster sampling) at any stage, and these methods may differ between PSUs. The enumeration or listing units on which surveys are conducted are the units selected at the last sampling stage.

Other names used for multistage sampling are *nested sampling* and *k-stage sampling* where k is the number of sampling stages in the survey design. The term *subsampling* is also used to describe the use of sample clusters as the population for the subsequent sampling stage.

Multistage sampling should not be confused with multiphase sampling. The sampling units in a multistage sample are nested. By contrast, multiphase sampling uses the same set of units at all phases included in its design. However, it is possible to use both multistage and multiphase techniques in a complex survey design (see Kish [4, Section 12.1] and Foreman [2, Section 7.4]).

The primary reason for using multistage sampling is to reduce the costs of data collection. For example, in a single-stage sample of households in a city one would have to list the households in the whole city, whereas in a multistage sample with city blocks used as PSUs, one could restrict listing activities to a sample of city blocks. One also uses multistage sampling when an exhaustive listing of the target population cannot be compiled. For instance, a list of hospital patients does not exist, but a list of hospitals does. Kish [4, Chapters 6, 9–11] and Sudman [6, Chapter 7] give detailed procedures for listing units in multistage area surveys.

The operational advantages of multistage sampling are offset by a loss in sampling efficiency. A multistage sample usually results in larger sampling error than does a **simple random sample** of the same size for the corresponding sample estimates. However, multistage sampling usually yields smaller **variances** for a unit of cost.

To facilitate discussion, we consider two-stage samples. Extensions to more sampling stages are intuitive. A sample of m PSUs is selected from the population total of M PSUs. A sample cluster of n_i secondary units is selected from the population of N_i secondary units in the ith sample PSU. The total number of listing units in the population and sample are $N = \sum_I^M N_I$ and $n = \sum_i^m n_i$, respectively.

The overall selection probability, P_{ij}, for the jth secondary unit in the ith PSU is the product $P_{ij} = P_{1i} \times P_{2ij}$, where P_{1i} and P_{2ij} are sampling fractions at the first and second sampling stages, respectively. The P_{2ij} is the **conditional probability** of selecting the jth unit provided the ith PSU is selected. One usually uses equal selection probabilities, $P_{2ij} = f_{2i}$, within PSUs, especially at the last stage. This article assumes equal selection probabilities within PSUs.

When sampling fractions within PSUs are uniform across PSUs, i.e. $f_{2i} = f_2$, N_i is frequently not a multiple of the expected sample cluster size $\mathrm{E}(n_i) = f_{2i} \times N_i$ and the expected size includes a fraction. The actual size is then variable with possible values n_i or $n_i + 1$. To minimize the effort of determining the actual sample cluster sizes in such situations, one usually uses **systematic random sampling** to select secondary units, thus allowing the random start to determine the sample cluster size from each PSU.

Estimation

In multistage sampling, **estimation** is also done in stages, starting with the last units. In two-stage samples one first estimates PSU aggregates with data from secondary units. Here, we consider samples in which simple random sampling is used at both stages without replacement. For equal probability samples of secondary units,

$$x_i' = \frac{N_i}{n_i} \sum_j^{n_i} x_{ij} \tag{1}$$

is an **unbiased** estimate of the PSU aggregate X_i. These estimates are used in place of X_i in the unbiased estimate,

$$x' = \frac{M}{m}\sum_{i}^{m} x_i', \tag{2}$$

of the aggregate X for the entire population. The variance of x' is

$$\sigma_{x'}^2 = \frac{M^2}{m}\frac{M-m}{M}S_{1X}^2 + \frac{M}{m}\sum_{I}^{M}\frac{N_I^2}{n_I}\frac{N_I - n_I}{N_I}S_{2IX}^2$$
$$= B_{x'}^2 + W_{x'}^2 \tag{3}$$

where

$$S_{1X}^2 = \frac{\sum_{I}^{M}(X_I - \overline{X})^2}{M-1}, \tag{4}$$

$$S_{2IX}^2 = \frac{\sum_{J}^{N_I}(X_{IJ} - \overline{\overline{X}}_I)^2}{N_I - 1}, \tag{5}$$

$$X_I = \sum_{J}^{N_I} X_{IJ}, \quad \overline{X} = \frac{\sum_{I}^{M} X_I}{M} = \frac{X}{M} \quad \text{and}$$

$$\overline{\overline{X}}_I = \frac{\sum_{J}^{N_I} X_{IJ}}{N_I} = \frac{X_I}{N_I} \tag{6}$$

[3, Volume I, Section 6.6, Volume II, Section 6.1]. The S_{1X}^2 is the population variance between PSU totals. The S_{2IX}^2 is the population variance between secondary units within the Ith PSU.

The variance of x' in (3) is expressed in terms of the contributions at each sampling stage. The first term is the between-PSU component of variance; it represents the contribution due to sampling PSUs. The second term is the within-PSU component of variance; it is the contribution due to sampling secondary units. The between-PSU component is the variance one gets if there were no subsampling within PSUs ($n_i = N_i$) i.e. if the sample were a one-stage cluster sample. The within-PSU component is the variance one gets if one takes all PSUs into the sample ($m = M$), i.e. if the sample were a stratified sample with M strata (*see* **Variance Components**).

In multistage sampling the estimates of means and proportions are usually ratios of estimated aggregates because the denominator, as well as the numerators, are unknown and, hence, must also be estimated. For a ratio $R = X/Y$ of aggregate variates X and Y, one may use the estimate $r = x'/y'$, where x' and y' are defined in (2). Expressions for σ_r^2 are similar in form to those for $\sigma_{x'}^2$. The variance of r for samples selected with equal probability without replacement at both stages is approximately

$$\sigma_r^2 = \frac{M^2}{m}\frac{M-m}{M}S_{1R}^2 + \frac{M}{m}\sum_{I}^{M}\frac{N_I^2}{n_I}\frac{N_I - n_I}{N_I}S_{2IR}^2$$
$$= B_r^2 + W_r^2, \tag{7}$$

where

$$S_{1R}^2 \doteq \left(\frac{1}{Y^2}\right)(S_{1X}^2 + R^2 S_{1Y}^2 - 2RS_{1XY}) \tag{8}$$

with S_{1X}^2 and S_{1Y}^2 defined in (4) and population covariance defined as

$$S_{1XY} = \frac{\sum_{I}^{M}(X_I - \overline{X})(Y_I - \overline{Y})}{M-1}, \tag{9}$$

and where

$$S_{2IR}^2 \doteq \left(\frac{1}{Y}\right)^2 (S_{2IX}^2 + R^2 S_{2IY}^2 - 2RS_{2IXY}), \tag{10}$$

with S_{2IX}^2 and S_{2IY}^2 defined in (5) and the population covariance defined as

$$S_{2IXY} = \frac{\sum_{J}^{N_I}(X_{IJ} - \overline{\overline{X}}_I)(Y_{IJ} - \overline{\overline{Y}}_I)}{N_I - 1}. \tag{11}$$

The variance approximation in (7) is good if the sample size is large enough that the coefficient of variation $V_{y'} = \sigma_{y'}/Y$ for the denominator is less than 0.05 [3, Volume I, Section 6.6], or if the sample is large enough that both $V_{x'} = \sigma_{x'}/X$ and $V_{y'}$ are less than 0.10 [1, Section 6.3] (*see* **Ratio and Regression Estimates**).

Two estimates of variance are useful in multistage estimates. When not designing samples, one may approximate the combined contribution of the first and second stages of variance simply with ultimate cluster variance estimates. An ultimate cluster is the entire sample of listing units selected from a PSU,

regardless of how many sampling stages are used. For example, in a three-stage sample where city blocks, households, and persons are selected at the first, second, and third sampling stages, persons are the listing units and blocks are the PSUs; an ultimate cluster then consists of all persons selected to the sample from one block. When two or more PSUs are selected to the sample, the ultimate cluster estimate of σ_z^2 for generic statistic z (z is the aggregate estimate x' or ratio estimate r) is

$$\hat{\sigma}_z^2 = \frac{M^2}{m} s_{cZ}^2. \tag{12}$$

If z is the aggregate estimate x' defined in (3), then

$$s_{cX}^2 = \frac{\sum_i^m (x_i' - \overline{x})^2}{(m-1)}, \tag{13}$$

where x_i' is an unbiased estimate of PSU aggregate X_i, and $\overline{x} = x'/M$. Similarly, if z is a ratio estimate r, then

$$s_{cR}^2 \doteq \left(\frac{1}{y'}\right)^2 (s_{cX}^2 + r^2 s_{cY}^2 - 2r s_{cXY}), \tag{14}$$

with

$$s_{cXY} = \frac{\sum_i^m (x_i' - \overline{x})(y_i' - \overline{y})}{m-1}. \tag{15}$$

Expressions (13) and (14) are variances between ultimate clusters.

The estimate in (12) is **consistent**. One may use it with any multistage sample, regardless of the number of stages or probability subsampling methods, provided the sampling is independent between PSUs. When the PSUs are selected with replacement, $\hat{\sigma}_{x'}^2$ for the aggregate estimate x' is unbiased. When PSUs are selected without replacement (*see* **Sampling With and Without Replacement**), $\hat{\sigma}_z^2$ overstates σ_z^2, but it may still be a useful approximation, especially when m/M is small and the between-PSU variance component is not a large portion of the total variance (see Hansen [3, Volume I, Sections 6.7 and 9.15] and Foreman [2, Section 8.5]).

When one designs samples and needs approximations for the separate variance components, one may use the following consistent estimates. An estimate of the within-PSU component in (3) or (7) is

$$w_z^2 = \frac{M^2}{m} \hat{s}_{2z}^2, \tag{16}$$

where

$$\hat{s}_{2Z}^2 = \frac{1}{m} \sum_i^m \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} s_{2iZ}^2. \tag{17}$$

When z is the aggregate estimate x',

$$s_{2iX}^2 = \frac{\sum_j^{n_i} (x_{ij} - \overline{\overline{x}}_i)^2}{n_i - 1}, \tag{18}$$

and, when z is the ratio estimate r,

$$s_{2iR}^2 \doteq \left(\frac{1}{y'}\right)^2 (s_{2iX}^2 + r^2 s_{2iY}^2 - 2r s_{2iXY}), \tag{19}$$

with

$$s_{2iXY} = \frac{\sum_j^{n_i} (x_{ij}' - \overline{\overline{x}}_i)(y_{ij}' - \overline{\overline{y}}_i)}{n_i - 1}. \tag{20}$$

A consistent estimate of the between-PSU component B^2 in (3) or (7) is

$$b_z^2 = \frac{M^2}{m} \frac{M-m}{M} (s_{cZ}^2 - \hat{s}_{2Z}^2), \tag{21}$$

where s_c^2 is defined in (13) or (14) and $\hat{s}_2^2$ is defined in (17) [3, Volume I, Section 6.7, Volume II, Section 6.4].

Sample Size and Allocation

For guidance in selecting among alternative sampling designs, it is frequently helpful to compare them with simple random samples. If one's multistage sample is self-weighting ($f_{2i} = f_2$ so $f_{ij} = f_1 \times f_2 = f$) in addition to being selected with equal probabilities at both stages, the comparison is conveniently made by approximating the variance of estimate z with

$$\sigma_z^2 \doteq \frac{\sigma_Z^2}{m\overline{n}}[1 + \delta_Z(\overline{n} - 1)] = \frac{\sigma_Z^2}{m\overline{n}} \text{ DEFF } (z), \tag{22}$$

where σ_Z^2 is the population variance between listing units for variate Z, $\overline{n} = f\overline{N}$ is the expected sample size per PSU, $\overline{N} = N/M$, and δ_Z is a measure of the homogeneity for variate Z between listing units within PSUs. The factor $\sigma_Z^2/m\overline{n}$ in (22) is the variance that would result if a simple random sample of $n = m\overline{n}$ listing units were used to estimate Z. The remaining factor is the **design effect**, DEFF, which reflects the precision lost due to use of a multistage sample instead of a simple random sample of listing

units. The DEFF and δ differ with sampling design and with the variate being estimated.

For self-weighting samples with both stages selected with equal probabilities without replacement, the appropriate value of δ is expressed by

$$\delta_Z = \frac{\frac{M-1}{M} S_{1Z}^2 - \overline{N} S_{2Z}^2}{\frac{M-1}{M} S_{1Z}^2 + \overline{N}(\overline{N}-1) S_{2Z}^2}, \tag{23}$$

where S_{1Z}^2 is defined in (4) for aggregates and in (8) for ratios, and where

$$S_{2Z}^2 = \sum_I^M \frac{N_I}{N} S_{2IZ}^2, \tag{24}$$

with S_{2IZ}^2 defined in (5) or (10). The measure of homogeneity for a multistage sample is approximately the same as that for the population. Thus, a simple estimate for the δ in (23) is

$$\delta_z' = \frac{s_{cz}^2 - \overline{n} s_{2z}^2}{s_{cz}^2 + \overline{n}(\overline{n}-1) s_{2z}^2}, \tag{25}$$

where s_{cz}^2 is defined in (13) or (14), and

$$s_{2z}^2 = \sum_i^m \frac{n_i}{n} s_{2iz}^2, \tag{26}$$

with s_{2iz}^2 defined in (18) or (19) (see Hansen et al. [3, Volume I, Section 6.8] and Foreman [2, Sections 8.2–8.3]).

Expression (22) shows that for a fixed sample size $n = m\overline{n}$, σ_z^2 varies with the factor $1 + \delta_Z(\overline{n} - 1)$ and, thus with δ_Z and $\overline{n}$. For many populations the natural clusters consist of units that are homogeneous relative to the units in the population as a whole, i.e. $\delta_Z > 0$. For example, persons residing in a city block tend to be more similar to one another than to persons in the city as a whole. Hence, as a general rule one wants to maximize the number of PSUs while minimizing the sample size within each PSU.

Increasing the number or size of sample PSUs may adversely affect survey costs. For example, the costs of listing households increases with the number of city blocks sampled when blocks are the PSUs in a multistage sample of households. To determine the most efficient sample allocation between PSUs and secondary units, one must consider both the costs and internal homogeneity of PSUs. For a two-stage sample, a simple cost function is

$$C = C_1 m + C_2 m\overline{n}, \tag{27}$$

which includes one term for each sample stage. The term for the kth sampling stage is the total cost for that stage of sampling. That term is the product of the number of units selected at the kth stage times C_k, where C_k is the average cost that is incurred when the kth-stage sample is increased by one unit. Among other costs, C_1 includes the cost for a list of all secondary units within a PSU; it also includes the costs of travel to reach PSUs if PSUs are geographically spread out. For example, in a sample of hospital discharges, C_1 includes the cost of gaining the hospital's cooperation and the cost of constructing a sampling frame of all of that hospital's eligible discharges. The second-stage cost, C_2, includes costs for collecting and processing data about the listing unit. Hansen et al. [3, Volume I, Sections 6.10–6.15] has a good discussion on constructing cost functions for multistage samples, but the costs need updating.

For two-stage samples selected with equal probabilities at both stages and the cost model in (27), the optimum sample size per PSU is

$$\text{opt. } \overline{n}_z = \left(\frac{C_1}{C_2}\frac{1-\delta_Z}{\delta_Z}\right)^{1/2} = \overline{N}\left(\frac{C_1}{C_2}\frac{S_{2Z}^2}{S_{1Z}^2 - \overline{N} S_{2Z}^2}\right)^{1/2}, \tag{28}$$

where δ is defined in (23). The optimum in (28) is independent of total survey costs. It requires only the ratio of unit costs. The optimum increases with $(1-\delta)/\delta = [(1/\delta) - 1]$; it increases as δ decreases. It also increases as the PSU costs increase relative to the secondary unit costs. The optimum number of sample PSUs depends on which of two objectives one wants satisfied by the optimized sample. When that objective is to produce estimates for a fixed value C_0 for the survey cost C defined in (27), then

$$\text{opt. } m_z = \frac{C_0}{C_1 + C_2 \text{ opt. } \overline{n}_z}, \tag{29}$$

where opt. $\overline{n}_z$ is given in (28). If, instead of a fixed survey cost, one wants to produce an estimate z with

a specified value σ_0^2 for σ_z^2 defined in (3) or (7), then

$$\text{opt. } m_z = \frac{MS_{1Z}^2 + \dfrac{N^2}{\text{opt.}\overline{n}}\dfrac{\overline{N} - \text{opt. } \overline{n}}{\overline{N}}S_{2Z}^2}{\sigma_0^2 + MS_{1Z}^2}, \quad (30)$$

where opt.$\overline{n}$ is defined in (28), the S_1^2 is defined in either (4) or (8), and S_2^2 is defined in (24) [5, Sections 10.4–10.5]. Values for C_1 and C_2, as well as estimates for S_{1Z}^2 and S_{2Z}^2, can be derived from experience in similar surveys for similar variates. For more on optimizing multistage samples, see Hansen et al. [3, Volume I, Sections 6.16–6.26, 9.7] and Foreman [2, Section 8.4].

Variable PSU Sizes

When PSUs vary in size, sampling them with equal probability is frequently not efficient. Simple random samples of varying sized PSUs will likely yield large between-PSU variance components for estimates of totals for the entire population if the PSU totals are correlated with the PSU size. Such samples will also yield sample clusters of varying sizes if the second-stage sampling fractions are uniform ($f_{2i} = f_2$). Variation in sample cluster sizes usually means increased survey costs owing to variation in work load between PSUs. Efforts to reduce variation in PSU size (by splitting large PSUs, combining small PSUs, or otherwise reconstructing PSUs) are usually too difficult or costly to be worthwhile.

There are three basic methods generally used for controlling the effects of variation in PSU sizes. Each requires some information on PSU size for every PSU. However, one usually uses approximations because the actual sizes are rarely known. One may use a stratified sample of PSUs with strata defined by PSU size. The largest PSUs or unusual PSUs can then be placed in a certainty stratum, where the estimate has no between-PSU variance component. The estimate for an aggregate X for the entire population is then the sum of estimates for stratum aggregates and the variance of that estimate is the sum of the variances for stratum estimates. If one uses simple random sampling at each stage within strata, the estimate of aggregate X_h for the hth stratum is given in (2) with variance given in (3).

Another method of controlling for varying PSU sizes is to select PSUs by **sampling with probability proportional to size** (pps). Under pps, one selects the ith PSU with probability $\pi_i = m(A_i/A) = mP_i$, provided $A_i < A/m$. The A_I is the approximation to size of the Ith PSU and $A = \sum_I^M A_I$. When $A_I > A/m$, one typically places the Ith PSU in a certainty stratum and then uses pps to select a sample of $m - 1$ PSUs from the remaining population. For sample PSUs with $A_i < A/m$, one almost always uses the probability $f_{2i} = f \times (1/p_i)$ to select the ijth secondary unit. That makes the sample self-weighting and the sample clusters approximately equal in size. Some variation in n_i is likely because of imperfect PSU size measures. When PSUs are selected with varying probabilities, an unbiased estimator of the population aggregate X is the **Horvitz–Thompson estimator**

$$x' = \sum_i^m \frac{x_i'}{\pi_i} = \frac{1}{m}\sum_i^m \frac{x_i'}{P_i}, \quad (31)$$

where x_i' is given in (1). When the P_i are uniform ($P_i = 1/M$), (31) becomes identical to (2). For a self-weighting sample with PSUs selected with pps without replacement, the variance of x' in (31) is approximately

$$\sigma_{x'}^2 = \frac{1}{m}\sum_I^M P_I \frac{1 - nP_I}{1 - P_I}\left(\frac{X_I}{P_I} - X\right)^2 + \frac{1}{m}\sum_I^M \frac{N_I^2}{P_I}\frac{(N_I - n_I)}{N_I n_I}S_{2IX}^2, \quad (32)$$

with S_{2IX}^2 defined in (5) (see Foreman [2, Section 7.5] and Hansen et al. [3, Volume I, Section 8.14, Volume II, Section 8.11]). The variance of ratio r is

$$\sigma_r^2 \doteq \left(\frac{1}{Y}\right)^2 (\sigma_{x'}^2 + R^2\sigma_{y'}^2 - 2R\sigma_{x'y'}), \quad (33)$$

where

$$\sigma_{x'y'} = \frac{1}{m}\frac{\sum_I^M P_I \dfrac{1 - nP_I}{1 - P_I}\left(\dfrac{X_I}{P_I} - X\right)\left(\dfrac{Y_I}{P_I} - Y\right)}{m - 1} + \frac{1}{m}\sum_I^M \frac{N_I^2}{P_I}\frac{(N_I - n_I)}{N_I n_I}S_{2IXY} \quad (34)$$

with S_{2IXY} defined in (11).

When two or more PSUs are selected with pps, the ultimate cluster estimate for $\sigma_{x'}^2$ is

$$\hat{\sigma}_{x'}^2 \doteq \frac{1}{m} \frac{\sum_i^m \left(\frac{x_i'}{P_i} - x' \right)^2}{m-1}, \tag{35}$$

where x_i' is an unbiased estimate of the PSU population aggregate X_i and x' is defined in (31).

Probability proportional to size and a fixed overall sampling fraction usually decrease variances compared with samples in which first-and second-stage probabilities are uniform. However, pps increases the portion of large PSUs selected to the sample. This may increase survey costs, especially if the cost of listing secondary units within PSUs is related to the number of them within the PSU. However, the work load at the second stage will be approximately the same from PSU to PSU, because the within-PSU samples are about equal. Generally, when it is economic to use pps, pps is preferred to stratification.

A third method of controlling for variations in PSU sizes depends on a ratio estimate. When one has supplementary information (from a census, administrative records, or source other than the survey) on the population total for an independent characteristic Y, one may estimate the total X by

$$x'' = \frac{x'}{y'} Y = rY, \tag{36}$$

where x' and y' are defined in (31). The variance of x'' is $\sigma_{x''} = Y^2 \sigma_r^2$, where σ_r^2 is defined in (33). When Y_I is the PSU population total N_I for listing units or some other measure of PSU size that is highly correlated with X_I, the variance of x'' may be substantially less than that of x'. The estimate x'' is biased, but the bias becomes trivial with large numbers of sample PSUs. If one uses actual PSU totals Y_i in place of estimates y_i' in (31), then y' has no sampling at the second stage. The term involving the within-PSU variance component is dropped from $\sigma_{y'}^2$ in (32).

For more on controlling for variation in PSU sizes see Hansen et al. [3, Volume I, Chapter 8] and Foreman [2, Sections 7.4–7.6].

References

[1] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[2] Foreman, E.K. (1991). *Survey Sampling Principles*. Marcel Dekker, New York.

[3] Hansen, H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample Survey Methods and Theory*. Wiley, New York.

[4] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[5] Levy, P. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

[6] Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

Bibliography

The following publications discuss methodology of actual surveys.

Bryant, E. & Shimizu, I. (1988). Sample design, sampling variance, and estimation procedures for the National Ambulatory Medical Care Survey. National Center for Health Statistics, *Vital and Health Statistics* **2**(108).

Massey, J.T., Moore, T.F., Parsons, V.L. & Tadros, W. (1989). Design and estimation for the National Health Interview Survey, 1985–94. National Center for Health Statistics. *Vital and Health Statistics* **2**(110).

National Center for Health Statistics (1963-present) Programs and Collection Procedures. National Center for Health Statistics, *Vital and Health Statistics*, Series 1.

(*See also* **Double Sampling**)

IRIS SHIMIZU

Multivariate Adaptive Splines for Analyzing Longitudinal Data

Longitudinal data represent one of the most commonly encountered data structures in health-related studies as well as in other fields including economics and finance. For every subject in a study, the outcome variable is measured repeatedly over time, and some (p) **covariates** are also collected, which may or may not vary over time. The classic method for analyzing longitudinal data is the **mixed-effects** linear model [4]. The purpose of this article is to describe MASAL [7] – **nonparametric**, method for analyzing and exploring longitudinal and **growth** curve data.

As an illustration, Figure 1 displays a subset of the data analyzed by Zhang [8]. In this data set, weights (kg) in the first one and half years were collected from 298 infants in an effort to examine the potential impact of the mother's cocaine use during pregnancy on the infant's growth after birth. In addition, information on gestational age, sex, and race is also available. The objective is to characterize the growth pattern and identify the variables that affect the growth pattern.

For the notation, suppose that data are observed for n subjects. For subject i, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})'$ be the response vector (e.g. an infant's weights) measured at T_i time points $\mathbf{t}_i = (t_{i1}, \ldots, t_{iT_i})'$ and X_i a $T_i \times p$ design matrix (e.g. a mother's cocaine use and the infant's sex). Laird and Ware [4] introduced the following mixed-effects model:

$$\mathbf{y}_i = X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i, \qquad (1)$$

where β is the $p \times 1$ fixed effect parameter vector, Z_i a random effect design matrix, $\mathbf{b}_i$ the random effects, and $\mathbf{e}_i$ the measurement error. In general, $\mathbf{e}_i$ and $\mathbf{b}_i$ are assumed independent each other and among different subjects, and follow **multivariate normal distributions** with mean 0.

Some nonparametric or **semiparametric** methods have emerged to extend the mixed-effects model. Here, the term "nonparametric" or "semiparametric" is with respect to the fixed effect. In most extensions, the time is isolated from the other covariates, and the fixed effect is decomposed as $X_i\beta + \mu(\mathbf{t}_i)$ or $X_i\beta(t)$, where X_i excludes the measurement time, but may contain other **time-dependent covariates**. In other words, the time trend is either additive or multiplicative to the covariate effects. The major advantage of such an extension is that one-dimensional smoothing (*see* **Spline Smoothing**) has been extensively studied and well implemented in the statistical literature (*see also* **Spline Function**), and the interpretation is relatively easy.

In contrast, MASAL is the unique model that is based on a multivariate nonparametric smoothing. As we know, there are usually no *a priori* reasons to believe that $X_i\beta + \mu(\mathbf{t}_i)$ or $X_i\beta(t)$ is necessarily appropriate. As we will see later, a fitted MASAL model can help us decide whether $X_i\beta + \mu(\mathbf{t}_i)$ or $X_i\beta(t)$ is appropriate.

Specifically, Zhang [7] considered a general nonparametric model

$$y_{ij} = f(x_{1,ij}, \ldots, x_{p,ij}, t_{ij}) + e_{ij}(t_{ij}), \qquad (2)$$

where f is an unknown smooth function and $e_{ij}(t_{ij})$ is an element of $\mathbf{e}_i$. Because t_{ij} can be viewed as one of the time-varying covariates, for notational convenience, we include the time as one of the covariates and write $f(X_i) = f(x_{1,ij}, \ldots, x_{p,ij}, t_{ij})$. The estimation of a MASAL model proceeds in two steps: one deals with the **covariance matrix**, and the other fits the f function for a given covariance matrix.

First, suppose that the covariance matrix, $\Sigma_i (i = 1, \ldots, n)$, is given. The f function is estimated by finding a member in the following class of functions:

$$\left\{ \sum_{k=0}^{M} \beta_k B_k(X),\, M = 0, 1, \ldots \right\},$$

where $B_k(X)$ is a basis function to be defined shortly and β_k the regression coefficient ($k = 0, 1, \ldots, M$), to minimize the weighted **least squares** (WLS).

$$\sum_{i=1}^{n} (\mathbf{y}_i - f(X_i))' \sum_{i}^{-1} (\mathbf{y}_i - f(X_i)).$$

Unlike parametric **regression**, the number of terms, M, and the individual basis, $B_k(X)$, need to be estimated from the data. Particularly, $B_k(X)$ is made of these two basis functions: $(x_l - \tau)^+$ and $x_l, l = 1, \ldots, p$, where τ is called knot and needs to be estimated, and for any number a, $a^+ = \max(a, 0)$. $B_k(X)$ is either one of the forgoing two bases, for example, $(x_1 - 2)^+$, or the product of those functions involving distinct covariates such as $(x_1 - 2)^+ x_2 (x_3 - 5)^+$.

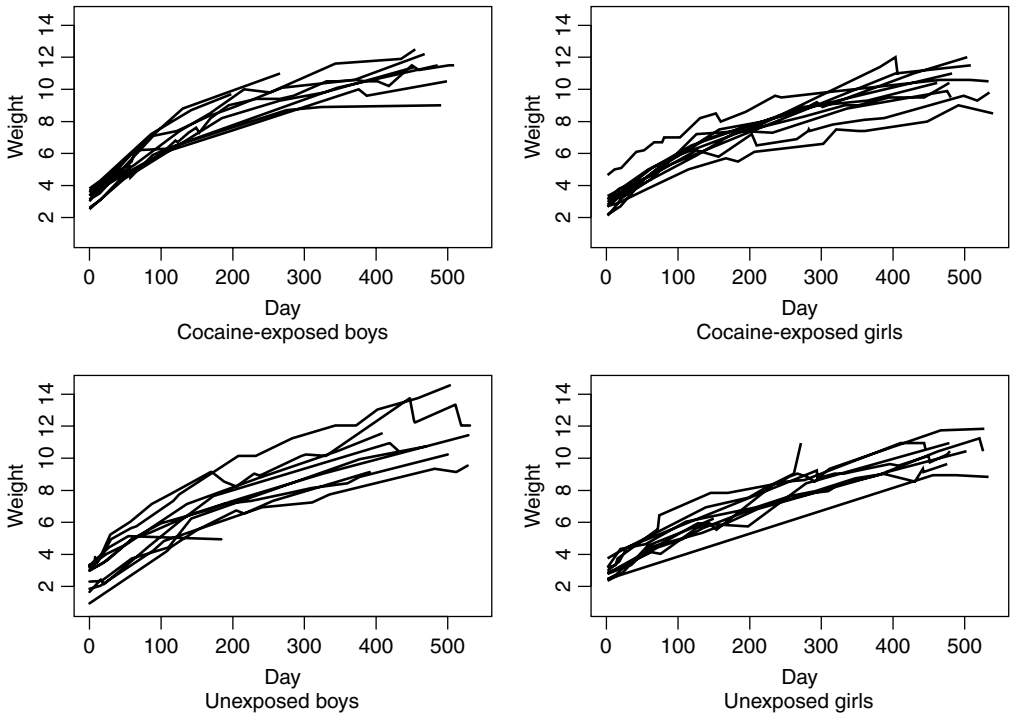

Figure 1 Growth curves of prenatally cocaine exposed and unexposed boys and girls. Ten infants are arbitrarily chosen in each group

The global minimization of the WLS is intractable. Instead, forward and backward procedures are used in practice [2, 6–8, 10] (*see* **Variable Selection**). In terms of computation, the most challenging step is to find the best τ during the forward stepwise. The fastest and exact **algorithm** is described by Zhang [7].

Next, we turn to the practical situation where Σ_i's are unknown. There are two main strategies. First, as discussed in Zhang [7], we can assume a general structure for Σ_i, for example, compound symmetry, auto-regressive correlation (*see* **ARMA and ARIMA Models**), or unstructured. Second, we decompose e_{ij} into an independent random measurement error and a systematic random variation. For example, in the analysis of growth curve data (sometimes referred to as *functional data analysis*), we may assume

$$e_{ij}(t_{ij}) = \sum_{l=1}^{L} \varphi_l(t_{ij}) b_{il} + \varepsilon_{ij}, \tag{3}$$

where similar distributions can be assumed for ε_{ij}'s and b_{il}'s to those in the linear mixed-effects model, and $\varphi_l(t)(l = 1, \ldots, L)$ is a prespecified function of t such as$\varphi(t) = t$ or $\sqrt{t}$ [5].

Once the general covariance structure, Σ_i, is chosen, the entire estimation procedure for MASAL proceeds as follows. We begin the process with initializing the parameters in Σ_i, for example, by assuming the independence of all observations. Then, the function, f, can be estimated as described above. Next, the **residuals** between $\mathbf{y}_i$ and its predicted value $\hat{f}(X_i)$ are computed and used to estimate the parameters in Σ_i. Then, the function f can be reestimated, and so on. While a theory has not been established, in nearly all applications that I have encountered, this process settles after the second iteration in the sense that there are few changes in the estimated f function and the WLS is hardly different in the subsequent iterations.

Assuming that (usually it is useful to examine a few choices)

$$e_{ij}(t_{ij}) = b_{i1} + \sqrt{t_{ij}} b_{i2} + \varepsilon_{ij}, \tag{4}$$

and applying MASAL to the infant growth data introduced earlier, the initial estimate of the f function under the independence assumption is

$$0.11 + \{0.5 - 0.28x_3 - 0.37(x_2 - 3)^+\}x_4 + (0.028 + 0.00028x_2 - 0.0008x_3)t, - \{0.016 + 0.00036(x_5 - 35)^+ - 0.0015x_1\}(t - 153)^+ \equiv \tilde{X}\beta(t) \quad (5)$$

where x_1 denotes race (white or black), x_2 the number of previous pregnancies, x_3 gender, x_4 cocaine exposure, x_5 gestational age, $\beta(t) = (0.11, 0.1, 0.01t, 0.01(t - 153)^+)'$ and $\tilde{X} = (1, \{5 - 2.8x_3 - 3.7(x_2 - 3)^+\}x_4, (28 + 0.028x_2 - 0.08x_3), 16 + 0.036(x_5 - 35)^+ - 0.15x_1)$.

Thus, this initial model requires time-varying coefficients $\beta(t)$[3] (*see* **Time-dependent Covariate**) although the design matrix is adaptively determined from the data. Using the residual estimates from this initial model, the covariance parameters are estimated, which in turn leads to the updated estimate for the f function:

$$0.12 + 0.014x_4 + 0.035t - 0.012(t - 60)^+ - 0.01(t - 150)^+ - 0.003(t - 300)^+.$$

Whereas the change from the initial model to the second one is notable, it is not so after the second iteration. As initially suggested, time-varying coefficients are not warranted in the final model. This model indicates an interesting growth pattern of the infants. First, the presence of x_4 is indicative of the importance of cocaine exposure. Second, the overall growth is increasing in time thanks to the term, $0.35t$, the growth becomes slower and slower after 2 months, 5 months, and 10 months.

Basic diagnostic procedures have been proposed [9] to assess the assumption of the covariance structure, although further investigation is clearly warranted. Owing to the highly adaptive nature of MASAL, a formal theory for statistical inference is difficult. However, **bootstrap** method [1] can be used to resample the data and consider **confidence intervals** and bands for the parameters of interest. For instance, we can test the contribution of a variable by removing it from our consideration and assess the magnitude of the degraded **goodness of fit**. Likewise, we can also test the linearity of a variable by excluding the nonlinear part of a variable. It is noteworthy, however, that these strategies need careful scrutiny.

In summary, MASAL is the only available model that is based on a high-dimensional smoothing technique and does not impose functional restrictions on time and covariates *a priori*. Not only does it accommodate time-varying covariates, but it also allows unrestricted interactions among covariates and between time and covariates. In addition, the fitted models are easy to calculate and readily interpretable. Thus, the mixed-effects multivariate adaptive splines model is handy for exploring longitudinal data. However, before the related procedures for statistical inference is thoroughly examined, MASAL should be primarily used as a way to explore the data structure in longitudinal and growth curve data, rather than a method of **hypothesis testing**. Executable MASAL programs (both stand-alone and **S-PLUS** compatible versions) are available from Heping Zhang's web site at `http://peace.med.yale.edu`.

References

[1] Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[2] Friedman, J.H. (1991). Multivariate adaptive regression splines, *The Annals of Statistics* **19**, 1–141.

[3] Hoover, D.R., Rice, J.A., Wu, C.O. & Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika* **85**, 809–822.

[4] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[5] Meredith, W. & Tisak, J. (1990). Latent curve analysis, *Psychometrika* **55**, 107–122.

[6] Zhang, H.P. (1994). Maximal correlation and adaptive splines, *Technometrics* **55**, 196–201.

[7] Zhang, H.P. (1997). Multivariate adaptive splines for the analysis of longitudinal data, *Journal of Computational and Graphical Statistics* **6**, 74–91.

[8] Zhang, H.P. (1999). Analysis of infant growth curves using multivariate adaptive splines, *Biometrics* **55**, 452–459.

[9] Zhang, H.P. (2002). Mixed effects multivariate adaptive splines model, in *Nonlinear Estimation and Classification*, *Springer Lecture Notes in Statistical Series*, Vol. 171, D. Denison, M. Hansen, C. Holmes & B. Yu, eds. pp. 293–302.

[10] Zhang, H.P. & Singer, B. (1999). *Recursive Partitioning in the Health Sciences*. Springer, New York.

(*See also* **Linear Mixed Effects Models for Longitudinal Data**; **Multilevel Models**; **Nonlinear Mixed Effects Models for Longitudinal Data**)

HEPING ZHANG

Multivariate Analysis of Variance

The multivariate analysis of variance, or MANOVA, is an extension of the univariate **analysis of variance** to multidimensional, or vector-valued, observations. The same **experimental design** or treatment layout applies to each of the observed response variables. The univariate assumption of a normal distribution is replaced by the **multivariate normal distribution** for the data vectors and the random error components in the mathematical model of the design.

The One-Way Analysis of Variance Layout

We begin with the univariate one-way layout. Random samples of observations on some variable X have been obtained under k treatments or other experimental conditions. The data are arranged as shown in Table 1. The datum x_{ij} is an observation on a normally distributed random variable with mean μ_j and a constant variance σ^2 for all $N = N_1 + \cdots + N_k$ experimental units. More generally, the observations may be represented by the linear model

$$x_{ij} = \mu + \tau_j + e_{ij},$$

where μ is an effect common to all units, τ_j is the effect of the jth treatment, and e_{ij} is a normally distributed random variable with mean $\mathrm{E}(e_{ij}) = 0$ and variance $\mathrm{var}(e_{ij}) = \sigma^2$. The purpose of the one-way analysis of variance is to test the hypothesis that the k treatment means, $\mu_1, \ldots, \mu_k$, are equal, or equivalently that the **null hypothesis** of no treatment effects,

$$\mathrm{H}_0 : \tau_j = 0, \quad j = 1, \ldots, k,$$

Table 1

	Treatment			
	1	.	.	.	k
	x_{11}	.	.	.	x_{1k}
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	$x_{N_1 1}$	.	.	.	$x_{N_k k}$
Mean	$\overline{x}_1$	.	.	.	$\overline{x}_k$
Sample size	N_1	.	.	.	N_k

is true. To test the hypothesis we compute the within-treatments sum of squares

$$E = \sum_{j=1}^{k} \sum_{i=1}^{N_j} (x_{ij} - \overline{x}_j)^2$$

and the between-treatments sum of squares

$$H = \sum_{j=1}^{k} N_j (\overline{x}_j - \overline{x})^2,$$

where $\overline{x}$ is the grand mean of all observations. The mean square $E/(N-k)$ is an **unbiased** estimate of the variance σ^2, and $H/(k-1)$ is such an unbiased estimate only if the hypothesis of equal treatment effects is true. E and H are independently distributed, so that

$$F = \frac{[H/(k-1)]}{[E/(N-k)]}$$

has the ***F* distribution** with $k-1$ and $N-k$ degrees of freedom when the null hypothesis is true. When the alternative of unequal treatment effects holds, the expected value of F will be large, and the null hypothesis should be rejected for F in excess of an appropriate critical value.

The one-way multivariate analysis of variance has the same experimental design, but the observations x_{ij} are replaced by $p \times 1$ observation vectors $\mathbf{x}_{ij}$. The p components, x_{ijh}, are measurements or other data on p response variables describing characteristics or dimensions of the experimental unit. The linear model for the vector elements is

$$x_{ijh} = \mu_h + \tau_{jh} + e_{ijh},$$

where μ_h is a general effect for the hth response variable, τ_{jh} is the effect of the jth treatment on the hth response, and e_{ijh} is a random disturbance. The vector $\mathbf{e}'_{ij} = [e_{ij1}, \ldots, e_{ijp}]$ has the multivariate normal distribution with null mean vector and **covariance matrix** $E(\mathbf{e}_{ij}\mathbf{e}'_{ij}) = \mathbf{\Sigma}$. As in the univariate model, a common covariance matrix holds for all pairs (i, j), and the $\mathbf{e}_{ij}$ of different experimental units are independently distributed. The hypothesis of equal treatment effects can be expressed as

$$\mathrm{H}_0 : \tau_{jh} = 0, \quad j = 1, \ldots, k, h = 1, \ldots, p,$$

or, in terms of the treatment mean vectors,

$$\boldsymbol{\mu}'_j = [\mu_1 + \tau_{j1}, \ldots, \mu_p + \tau_{jp}],$$

Table 2

	Treatment		
	1	. . .	k
Mean vector	$[\overline{x}_{11}, \ldots, \overline{x}_{1p}]$	. . .	$[\overline{x}_{k1}, \ldots, \overline{x}_{kp}]$
Grand mean vector	$[\overline{x}_1, \ldots, \overline{x}_p]$		

as

$$\mathrm{H}_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k.$$

The alternative is that of unequal mean vectors.

For the multivariate analysis of variance we begin by replacing the treatment means by mean vectors (Table 2). The treatment sum of squares, H, becomes the $p \times p$ matrix

$$\mathbf{H} = \begin{bmatrix} h_{11} & . & . & . & h_{1p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ h_{1p} & . & . & . & h_{pp} \end{bmatrix},$$

in which

$$h_{rs} = \sum_{j=1}^{k} N_j(\overline{x}_{jr} - \overline{x}_r)(\overline{x}_{js} - \overline{x}_s), \quad r, s = 1, \ldots, p.$$

The diagonal terms, h_{rr}, are the treatment sums of squares for the individual response variables, while the off-diagonal values, h_{rs}, are the sums of cross-products for all pairs of the variables. The within-treatments sum of squares, E, is extended to the $p \times p$ matrix

$$\mathbf{E} = \begin{bmatrix} e_{11} & . & . & . & e_{1p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ e_{1p} & . & . & . & e_{pp} \end{bmatrix},$$

with general element

$$e_{rs} = \sum_{j=1}^{k} \sum_{i=1}^{N_j} (x_{ijr} - \overline{x}_{jr})(x_{ijs} - \overline{x}_{js}).$$

The rth diagonal element of $\mathbf{E}$ is merely the one-way analysis of variance within-groups sum of squares for the rth response variable. The off-diagonal term, e_{rs}, is a corresponding sum of products of the observations on the rth and sth responses. From those univariate definitions of the h_{rs} and e_{rs} terms we can easily construct the matrices $\mathbf{H}$ and $\mathbf{E}$ for the one-way layout or, for that matter, those for any layout for which an analysis of variance is available.

The relative closeness of the elements of $\mathbf{H}$ to those of $\mathbf{E}$ is a measure of the validity of the hypothesis of equal mean vectors. We measure "closeness" by various functions of the roots of the determinantal equation $|\mathbf{H} - \lambda\mathbf{E}| = 0$ or, equivalently, the characteristic roots of the matrix $\mathbf{E}^{-1}\mathbf{H}$ (*see* **Eigenvalue**). The following are the principal test statistics for the multivariate analysis of variance:

Wilks' determinantal ratio:

$$\begin{aligned} \Lambda &= \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|} \\ &= \frac{1}{|\mathbf{E}^{-1}\mathbf{H} + \mathbf{I}|} \\ &= \frac{1}{\text{product of the characteristic roots of } \mathbf{E}^{-1}\mathbf{H}+\mathbf{I}} \end{aligned}$$

(*see* **Lambda Criterion, Wilks'**).

Roy's greatest root:

$$c_s = \text{largest characteristic root of } \mathbf{E}^{-1}\mathbf{H},$$

or

$$\theta_s = \frac{c_s}{(1 + c_s)}$$

(*see* **Roy's Maximum Root Criteria**).

Lawley–Hotelling trace:

$$\begin{aligned} T_0^2 &= \operatorname{tr} \mathbf{E}^{-1}\mathbf{H} \\ &= \text{sum of the characteristic roots of } \mathbf{E}^{-1}\mathbf{H} \end{aligned}$$

(*see* **Lawley–Hotelling Trace**).

Pillai trace:

$$V = \operatorname{tr} \mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$$

(*see* **Pillai's Trace Test**).

We give the distributional properties of the statistics in a later section.

An Example

As an illustration of the one-way multivariate analysis of variance we use $p = 3$ dimensions of the skulls of four variants of the wolf *Canis lupus L.* The data were discussed by Jolicoeur [4, 5] and given

as an example by Morrison [7]. The measurements are given in **Multivariate Analysis, Overview**. The response variables are these dimensions:

$$X_1 = \text{palatal length},$$
$$X_2 = \text{postpalatal length},$$
$$X_3 = \text{zygomatic width}.$$

The four groups consisted of male and female wolves from the Rocky Mountain and Arctic Archipelago regions of Canada. The mean vectors and covariance matrices for the four data sets are as follows:

1. *Rocky Mountain males* $(N_1 = 6)$:

$$\bar{\mathbf{x}}_1' = [126.50, 108.17, 145.17],$$
$$\mathbf{S}_1 = \begin{bmatrix} 1.5000 & 1.7000 & 2.3000 \\ 1.7000 & 6.1667 & 5.6667 \\ 2.3000 & 5.5667 & 24.9667 \end{bmatrix}.$$

2. *Rocky Mountain females* $(N_2 = 3)$:

$$\bar{\mathbf{x}}_2' = [117.33, 102.67, 128.67],$$
$$\mathbf{S}_2 = \begin{bmatrix} 5.3333 & 0.6667 & 2.6667 \\ 0.6667 & 0.3333 & -1.1667 \\ 2.6667 & -1.1667 & 10.3333 \end{bmatrix}.$$

3. *Arctic males* $(N_3 = 10)$:

$$\bar{\mathbf{x}}_3' = [115.80, 100.80, 142.40],$$
$$\mathbf{S}_3 = \begin{bmatrix} 6.1778 & 4.8444 & 6.4222 \\ 4.8444 & 9.5111 & 10.5333 \\ 6.4222 & 10.5333 & 31.8222 \end{bmatrix}.$$

4. *Arctic females* $(N_4 = 6)$:

$$\bar{\mathbf{x}}_4' = [110.83, 96.17, 137.00],$$
$$\mathbf{S}_4 = \begin{bmatrix} 5.3667 & 2.4333 & 4.2000 \\ 2.4333 & 9.7667 & 5.8000 \\ 4.2000 & 5.8000 & 25.2000 \end{bmatrix}.$$

From these we compute the between-groups sums of squares and products matrix:

$$\mathbf{H} = \begin{bmatrix} 781.16 & 585.28 & 364.29 \\ 585.28 & 445.51 & 245.66 \\ 364.29 & 245.66 & 656.74 \end{bmatrix};$$

the within-groups sums of squares and products matrix:

$$\mathbf{E} = \begin{bmatrix} 100.60 & 65.60 & 95.63 \\ 65.60 & 165.93 & 149.30 \\ 95.65 & 149.30 & 557.90 \end{bmatrix};$$

and

$$\mathbf{E}^{-1}\mathbf{H} = \begin{bmatrix} 7.88522 & 5.90989 & 3.20985 \\ 1.36789 & 1.13769 & -0.46443 \\ -1.06476 & -0.87718 & 0.75123 \end{bmatrix}.$$

The characteristic roots of $\mathbf{E}^{-1}\mathbf{H}$ are 8.5280, 1.20595, and 0.04018. The values of the four test statistics and their approximate p values are shown in Table 3. The hypothesis of equal mean vectors should be rejected at any reasonable significance level.

The Multivariate General Linear Model

The Model

We now consider the multivariate analysis of variance for the multidimensional general linear model. The model is

$$\mathbf{X} = \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\varepsilon}$$
$$= [\mathbf{A}_1\ \mathbf{A}_2]\begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{bmatrix} + \boldsymbol{\varepsilon},$$

in which

$\mathbf{X}$ = the $N \times p$ observation matrix,
$\mathbf{A}$ = the $N \times q$ design matrix of rank r for the given experimental design,
$\mathbf{A}_1$ = the $N \times r$ basis matrix for $\mathbf{A}$,
$\mathbf{A}_2$ = the $N \times (q - r)$ completion of $\mathbf{A}_1$ into $\mathbf{A}$,
$\boldsymbol{\xi}$ = the $q \times p$ parameter matrix,
$\boldsymbol{\xi}_1$ = the $r \times p$ parameter matrix corresponding to the basis of $\mathbf{A}$,

Table 3

Criterion	Statistic	p value
Roy's greatest root	8.5280	$\ll 0.01$
Wilks' Λ	0.04574	3.17×10^{-10}[a]
Lawley–Hotelling T_0^2	9.7741	1.55×10^{-47}[a]
Pillai trace V	1.4804	0.00029[a]

[a]Based on **large-sample** limiting distributions.

$\boldsymbol{\xi}_2$ = the $(q-r)\times p$ matrix of parameters for the completion $\mathbf{A}_2$ of $\mathbf{A}$, and
$\boldsymbol{\varepsilon}$ = the $N\times p$ matrix of random disturbances.

Each row of $\boldsymbol{\varepsilon}$ is independently distributed as a p-dimensional multinormal random variable with null mean vector and common covariance matrix $\boldsymbol{\Sigma}$.

The set of linear parametric functions, $\mathbf{a}'\boldsymbol{\xi} = \mathbf{a}_1'\boldsymbol{\xi}_1 + \mathbf{a}_2'\boldsymbol{\xi}_2$, is said to be *estimable* if $\mathbf{a}_1$ and $\mathbf{a}_2$ satisfy the following relation:

$$\mathbf{a}_2' = \mathbf{a}_1'(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{A}_2$$

(Roy [20]). Then the **minimum variance unbiased estimator** of $\mathbf{a}'\boldsymbol{\xi}$ is

$$\widehat{\mathbf{a}'\boldsymbol{\xi}} = \mathbf{a}_1'(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{X}.$$

General Linear Hypothesis

The multivariate general linear null hypothesis is

$$\mathrm{H}_0 : \mathbf{C}\boldsymbol{\xi}\mathbf{M} = \mathbf{0},$$

as opposed to its alternative, $\mathrm{H}_1 : \mathbf{C}\boldsymbol{\xi}\mathbf{M} \neq 0$. The $g\times q$ matrix $\mathbf{C}$ has rank $g \le r$, and is specified by the analyst. $\mathbf{C}$ is partitioned as $[\mathbf{C}_1 \quad \mathbf{C}_2]$, where its submatrices have respective dimensions $g\times r$ and $g\times(q-r)$ to conform with the partitioning of $\mathbf{A}$ and $\boldsymbol{\xi}\cdot\mathbf{M}$ is a $p\times u$ matrix that will generate linear functions of the responses or their parameters. Of necessity, $u \le p$. If the original response variables are to be used, then $\mathbf{M} = \mathbf{I}$. The null hypothesis can be written in terms of the submatrices as

$$\mathrm{H}_0 : \mathbf{C}_1\boldsymbol{\xi}_1\mathbf{M} + \mathbf{C}_2\boldsymbol{\xi}_2\mathbf{M} = \mathbf{0}.$$

The hypothesis is said to be *testable* if the rows of $\mathbf{C}_1$ and $\mathbf{C}_2$ satisfy the estimability conditions, or if

$$\mathbf{C}_2 = \mathbf{C}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{A}_2.$$

We test the null hypothesis through the $\mathbf{H}$ and $\mathbf{E}$ matrices defined by

$$\mathbf{H} = \mathbf{M}'\mathbf{X}'\mathbf{A}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{C}_1'[\mathbf{C}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{C}_1']^{-1} \times \mathbf{C}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{X}\mathbf{M},$$

$$\mathbf{E} = \mathbf{M}'\mathbf{X}'[\mathbf{I} - \mathbf{A}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1']\mathbf{X}\mathbf{M}.$$

$\mathbf{H}$ and $\mathbf{E}$ are independently distributed square symmetric matrices. $\mathbf{E}/(N-r)$ is an unbiased estimator of the population covariance matrix $\boldsymbol{\Sigma}$ as long as the linear model holds. $\mathbf{H}/g$ is only an unbiased estimator of $\boldsymbol{\Sigma}$ if H_0 is true. Examples of design and hypothesis matrices and the resulting matrices $\mathbf{H}$ and $\mathbf{E}$ have been given for some common experimental design layouts by Morrison [7].

We may compute the elements of the general $\mathbf{H}$ and $\mathbf{E}$ matrices from the univariate analysis of variance in the same manner as for the one-way MANOVA. The diagonal elements $h_{11},\ldots,h_{uu}$ of $\mathbf{H}$ are the univariate hypothesis sums of squares for the u linear functions of the multivariate response variables (or the p response variables themselves if $\mathbf{M} = \mathbf{I}$). The off-diagonal elements h_{ij} of $\mathbf{H}$ are sums of products, or bilinear forms, whose matrices are identical to the sums of squares quadratic forms on the diagonal of $\mathbf{H}$. Similarly, the diagonal elements $e_{11},\ldots,e_{uu}$ of $\mathbf{E}$ can be obtained from the corresponding univariate ANOVA. Each e_{ii} is the univariate error sum of squares for the successive u linear compounds. The off-diagonal terms e_{ij} are sums of products of all $u(u-1)/2$ pairs of the linear compounds of the observations, with the same bilinear form matrix as the quadratic forms e_{ii}. By those definitions one may avoid the matrix expressions for $\mathbf{H}$ and $\mathbf{E}$.

Test Statistics for the General Hypothesis

We now describe four common statistics for testing the general hypothesis $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\xi}\mathbf{M} = \mathbf{0}$. The first is due to Roy [19, 20], and arises from his **union–intersection method** of test construction. The statistic is the greatest characteristic root c_s of $\mathbf{E}^{-1}\mathbf{H}$, or the greatest root $\theta_s = c_s/(1+c_s)$ of $(\mathbf{H}+\mathbf{E})^{-1}\mathbf{H}$. Upper critical values of the distribution of θ_s have been computed and tabulated by Heck [1], Pillai & Bantegui [15], and Pillai [12–14]. Charts and tables of those critical values are available in current texts, e.g. Morrison [7]. The parameters of the distribution of the greatest root statistic θ_s when H_0 is true are

$$s = \min(g, u),$$
$$m = \frac{(|g-u|-1)}{2},$$
$$n = \frac{(N-r-u-1)}{2}.$$

The hypothesis $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\xi}\mathbf{M} = \mathbf{0}$ is rejected if $\theta_s > x_{\alpha;s,m,n}$. When $s = 1$ the single nonzero characteristic

root θ has the **beta distribution**, so that

$$F = \left[\frac{(n+1)}{(m+1)}\right]\frac{\theta}{(1-\theta)}$$
$$= \left[\frac{(n+1)}{(m+1)}\right]\operatorname{tr}\mathbf{E}^{-1}\mathbf{H}$$

has the F distribution with $2m+2$ and $2n+2$ degrees of freedom when H_0 is true. The Wilks' [22] determinantal ratio statistic $\Lambda = |\mathbf{E}|/|\mathbf{H}+\mathbf{E}|$ follows from the generalized **likelihood ratio test** construction. For large N,

$$\chi^2 = -\left[\frac{N-r-(u-g+1)}{2}\right]\ln\Lambda$$

has the **chi-square distribution** with gu degrees of freedom when H_0 is true, and H_0 would be rejected for large values of χ^2. Rao [17, 18] has given a transformation of Λ whose distribution can be closely approximated by that of an F variate. Exact null hypothesis distributions are also available in certain special cases. If $s = 1$, then

$$F = \left[\frac{(1-\Lambda)}{\Lambda}\right]\left[\frac{(n+1)}{(m+1)}\right]$$

has the F distribution with $2m+2$ and $2n+2$ degrees of freedom. When $s = 2$,

$$F = \left[\frac{(1-\Lambda^{1/2})}{\Lambda^{1/2}}\right]\left[\frac{(2n+2)}{(2m+3)}\right]$$

is distributed as an F variate with $4m+6$ and $4(n+1)$ degrees of freedom. The Lawley [6] and Hotelling [2, 3] statistic, $NT_0^2 = N\operatorname{tr}\mathbf{H}\mathbf{E}^{-1}$ tends to have a chi-square distribution with gu degrees of freedom when N is large and H_0 is true. Similarly, the Pillai [11] statistic, $(N-r)V = (N-r)\operatorname{tr}\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$, has a large-sample chi-square distribution with gu degrees of freedom under the null hypothesis. None of these criteria appears to be **most powerful** against all **alternative hypotheses**. Most statistical **software** systems for MANOVA give values for all of the major statistics.

Power and Robustness Properties

A number of studies have shown that the power probabilities of the four MANOVA test statistics differ only slightly for selected alternate hypotheses. Pillai & Jayachandran [16] found only small second-decimal-place differences. The Roy greatest-root test had lowest power against those alternatives, but other studies showed it surpassed the competing tests in the case of an alternative with a single large characteristic root. Olson [8–10] concluded that the Pillai trace statistic appeared to be the most robust in terms of preserving its α level (*see* **Level of a Test**) and power probabilities under nonnormality and unequal covariance matrices. The greatest-root test seemed to be affected the most by those departures from the usual model assumptions. Some further aspects of the sensitivity of MANOVA to nonnormality and heterogeneous covariance structures have been described in the article **Multivariate Techniques, Robustness**.

Simultaneous Tests and Confidence Intervals

As in the univariate case, MANOVA only indicates whether the overall hypothesis H_0 should be rejected, and not which response variables and treatment effects may have contributed to that decision. The union–intersection test based on the greatest root statistic leads directly to a **multiple comparisons** method due to Roy & Bose [21]. One may test all possible hypotheses of the sort $H_0 : \mathbf{b}'\mathbf{C}\boldsymbol{\xi}\mathbf{M}\mathbf{a} = \mathbf{0}$ for any treatment comparisons defined by $\mathbf{b}'\mathbf{C}$ and any linear functions of the response variables specified by $\mathbf{M}\mathbf{a}$ with an overall error rate not in excess of some given level α. Alternately, confidence intervals can be found for all parametric functions $\mathbf{b}'\mathbf{C}\boldsymbol{\xi}\mathbf{M}\mathbf{a}$ with a simultaneous confidence coefficient not less than $1-\alpha$. Explicit expressions for the test statistics and intervals are available in many sources on multivariate methods, e.g. Morrison [7].

References

[1] Heck, D.L. (1960). Charts of some upper percentage points of the distribution of the largest characteristic root, *Annals of Mathematical Statistics* **31**, 625–642.

[2] Hotelling, H. (1947). Multivariate quality control, illustrated by the air testing of sample bombsights, in *Selected Techniques of Statistical Analysis*, C. Eisenhart, et al., eds. McGraw-Hill, New York, pp. 111–184.

[3] Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 23–41.

[4] Jolicoeur, P. (1959). Multivariate geographical variation in the wolf *Canis lupus L., Evolution* **XIII**, 283–299.

[5] Jolicoeur, P. (1975). Sexual dimorphism and geographical distance as factors of skull variation in the wolf *Canis lupus L.*, in *The Wild Canids*, M.W. Fox, ed. Van Nostrand, New York.

[6] Lawley, D.N. (1938). A generalization of Fisher's z-test, *Biometrika* **30**, 180–187.

[7] Morrison, D.F. (1990). *Multivariate Statistical Methods*, 3rd Ed. McGraw-Hill, New York.

[8] Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

[9] Olson, C.L. (1976). On choosing a test statistic in multivariate analysis of variance, *Psychological Bulletin* **83**, 579–586.

[10] Olson, C.L. (1979). Practical considerations in choosing a MANOVA statistic: a rejoinder to Stevens, *Psychological Bulletin* **86**, 1350–1352.

[11] Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis, *Annals of Mathematical Statistics* **26**, 117–121.

[12] Pillai, K.C.S. (1964). On the distribution of the largest of seven roots of a matrix in multivariate analysis, *Biometrika* **51**, 270–275.

[13] Pillai, K.C.S. (1965). On the distribution of the largest characteristic root of a matrix in multivariate analysis, *Biometrika* **52**, 405–414.

[14] Pillai, K.C.S. (1967). Upper percentage points of the largest root of a matrix in multivariate analysis, *Biometrika* **54**, 189–194.

[15] Pillai, K.C.S. & Bantegui, C.G. (1959). On the distribution of the largest of six roots of a matrix in multivariate analysis, *Biometrika* **46**, 237–240.

[16] Pillai, K.C.S. & Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria, *Biometrika* **54**, 195–210.

[17] Rao, C.R. (1951). An asymptotic expansion of the distribution of Wilks' criterion, *Bulletin of the International Statistical Institute* **33**, 177–180.

[18] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. Wiley, New York.

[19] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[20] Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

[21] Roy, S.N. & Bose, R.C. (1953). Simultaneous confidence interval estimation, *Annals of Mathematical Statistics* **24**, 513–536.

[22] Wilks, S.S. (1932). Certain generalizations in the analysis of variance, *Biometrika* **24**, 471–494.

(*See also* **Multivariate Analysis, Overview**)

DONALD F. MORRISON

Multivariate Analysis, Bayesian

Multivariate **Bayesian** analysis is that branch of statistics that uses **Bayes' theorem** to make inferences about several, generally **correlated**, unknown quantities. The unknown quantities may index probability distributions, or they may be hypotheses or propositions, or they may be **probabilities** themselves. Such procedures have widespread biostatistical applications. Some of the basic concepts of the subject include the likelihood principle (*see* **Foundations of Probability**), multivariate **prior** and posterior distributions, **Markov chain Monte Carlo** numerical methods, and the use of Bayesian computer programs to implement the multivariate Bayesian procedures. These concepts, methods, and applications are discussed below.

Bayes' Theorem, Posterior Distributions, and Inference

In Bayesian analysis an unknown quantity is assigned a probability distribution, to represent one's degree of belief about the unknown quantity. This degree of belief is then updated via Bayes' theorem, as new information becomes available through observational data, experience, or new insights.

Multivariate Bayesian inference is based on Bayes' theorem for correlated **random variables**. The theorem asserts that the joint density of several correlated, jointly continuous, but unobservable random variables, given observations on one or more observable random variables, is proportional to the product of the **likelihood** function for the observable random variables and the probability density function of the probability distribution for the unknown variables. (If the unobservable random variables are jointly discrete, then we use the joint probability mass function instead of the joint density in Bayes' theorem; the analogous statement holds for unobservable, correlated random variables with mixed distributions.) The proportionality constant does not depend upon the unobservable quantities of interest; it is just that constant that makes the probability density function for the random variables of interest integrate to unity.

Bayes' Theorem

Symbolically, let $\boldsymbol{\Theta}$ denote a collection (vector) of k unobservable random variables, and $\mathbf{X}$ a collection (vector) of p observable random variables. Let $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ represent densities (probability mass functions) of their arguments. (Lower case letters will be used to represent observed values of the random variables designated by upper case letters.) Bayes' theorem asserts that

$$h(\boldsymbol{\theta}|\mathbf{x}) = \frac{1}{c} f(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ and $\mathbf{x}$ denote fixed values of $\boldsymbol{\Theta}$ and $\mathbf{X}$, respectively, and c denotes a constant (depending on $\mathbf{x}$, but not on $\boldsymbol{\Theta}$), which is given by

$$c = \int f(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}.$$

The integration is taken over all possible values in k-dimensional space, and the notation $f(\mathbf{x}|\boldsymbol{\theta})$ should be understood to mean the density of the conditional distribution of $\mathbf{X}$ given $\boldsymbol{\Theta} = \boldsymbol{\theta}$.

$f(\mathbf{x}|\boldsymbol{\theta})$ is the joint sampling density for $\mathbf{x}|\boldsymbol{\theta}$. When it is viewed as a function of $\boldsymbol{\theta}$, it is called the *likelihood function* (see section on "Likelihood Principle" below).

$g(\boldsymbol{\theta})$ is the *prior density* of $\boldsymbol{\Theta}$, since it is the density of $\boldsymbol{\Theta}$ prior to having observed $\mathbf{X}$ (it is a density if the variables in the $\boldsymbol{\Theta}$ array are continuous, and it is a probability mass function if they are discrete). Note that the prior density should not depend in any way on the current data set, although it certainly could and often does depend upon earlier-obtained data sets. If the prior were permitted to depend upon the current data set, then the use of Bayes' theorem in this inappropriate way would violate the laws of probability.

$h(\boldsymbol{\theta}|\mathbf{x})$ is the *posterior density* (probability mass function) of $\boldsymbol{\Theta}$, since it is the distribution of $\boldsymbol{\Theta}$ "subsequent" to having observed $\mathbf{X}$.

Bayesian inference in multivariate distributions is based on the posterior distribution of the unobservable random variables, say $\boldsymbol{\Theta}$, given the observable data (the unobservable random variable may be a vector or a matrix).

A Bayesian estimator (or posterior summary) of $\boldsymbol{\Theta}$ is generally taken to be a measure of location of the marginal posterior distribution of $\boldsymbol{\Theta}$, such as its mean, median, or mode.

For example, if there tends to be an underlying "quadratic **loss**" penalty function in an estimation problem, then the **mean** of the posterior distribution is optimal as an estimator, since it minimizes the expected loss (penalty). (For the same reason, **medians** are used with absolute error loss functions, and **modes** with binary types of decision rules.)

To obtain the marginal posterior density of $\boldsymbol{\Theta}$ given the data, it is often necessary to integrate the joint posterior density over spaces of other unobservable random variables that are jointly distributed with $\boldsymbol{\Theta}$.

For example, if the sampling distribution of $\mathbf{X}$ given $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ is $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ (*see* **Multivariate Normal Distribution**), the marginal posterior density of $\boldsymbol{\Theta}$ is obtainable by integrating the joint posterior density of $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ over all elements of $\boldsymbol{\Sigma}$ that make it positive definite.

Credibility Regions (Credible Regions)

Bayesian confidence regions (called *credibility regions or credible regions*) are obtainable for any preassigned level of credibility directly from the cumulative distribution function of the posterior distribution. We make a distinction here between "credibility" and "confidence" that is fundamental, and not just a simple choice of alternate words.

The *credibility region* is a probability region for the unknown, unobservable vector or matrix, conditional on the specific value of the observables that happened to have been observed in this instance, regardless of what values of the observables might be observed in other instances (the region is based upon $P\{\boldsymbol{\Theta}|\mathbf{X}\}$). For example, $\boldsymbol{\Omega}$ denotes a 95% credibility region for $\boldsymbol{\Theta}|\mathbf{X}$ if

$$\Pr\{\boldsymbol{\Theta} \in \Omega|\mathbf{X}\} = 95\%.$$

The *confidence region*, by contrast, is obtained from a probability statement about the observable variables, conditional on the unobservable ones, so it really represents a region based upon the distribution of where the observables are likely to be, rather than where the unobservables are likely to be (the region is based upon $P\{\mathbf{X}|\boldsymbol{\Theta}\}$). When non uniform, proper prior distributions are used, the resulting credibility and confidence regions will generally be quite different from one another.

Prediction

Predictions about a data vector(s) or matrix not yet observed are carried out by averaging the likelihood for the future observation vector(s) or matrix over the best information we have about the indexing parameters of its distribution, namely the posterior distribution of the indexing parameters given the data already observed.

Hypothesis Testing

Hypothesis testing may be carried out by comparing the posterior probabilities of all competing hypotheses, given all data observed, and selecting the hypothesis with the largest posterior probability. These notions are identical to those in univariate Bayesian analysis. In multivariate Bayesian analysis, however, in order to make posterior inferences about a given hypothesis, conditional upon the observable data, it is generally necessary to integrate out over all the components of $\boldsymbol{\Theta}$.

Likelihood Principle

The *likelihood function* is uniquely defined only up to a multiplicative constant. The likelihood function may be taken to be any constant multiple of the ordinary sampling, or frequency, function (probability mass function) of the joint distribution of all of the observable random variables given the unobservable ones.

The *likelihood principle* asserts that all relevant information about $\boldsymbol{\Theta}$ obtainable from the observable data is found in the likelihood function. The implication is that in terms of the observable data, to make inferences about $\boldsymbol{\Theta}$ we merely require the likelihood of the data, and nothing else. So if there are stopping rules (*see* **Sequential Analysis**), or other additional information about the sampling process, then such information is irrelevant for Bayesian inference. Nor should values of the observables that might have been taken, but were not, be relevant. For example, the expected value of the observables, a quantity required for invoking an **unbiasedness** principle, is not relevant for inference.

(Multivariate) Prior Distributions

The process of developing a prior distribution to express the beliefs of the analyst about the likely values of a collection of unobservables is called multivariate **subjective probability** assessment. None of the variables in a collection of unobservables, $\boldsymbol{\Theta}$, is ever known. The multivariate prior probability density function, $g(\boldsymbol{\theta})$, for continuous $\boldsymbol{\Theta}$ (or its counterpart, the prior probability mass function for discrete $\boldsymbol{\Theta}$), is used to denote the degrees of belief the analyst holds about $\boldsymbol{\Theta}$. The parameters that index the prior distribution are called *hyperparameters*.

For example, suppose $\boldsymbol{\Theta}$ is **bivariate** ($k = 2$), so that there are two unobservable, one-dimensional random variables Θ_1 and Θ_2. Suppose, furthermore (for simplicity), that Θ_1 and Θ_2 are discrete random variables, and let $g(\theta_1, \theta_2)$ denote the joint probability mass function for $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)$.

Suppose Θ_1 and Θ_2 can each assume only two values, 0 and 1, and the analyst believes the probable values to be given by those in Table 1. Thus, for example, the analyst believes that the chances that Θ_1 and Θ_2 are both 1 is 0.4, i.e.

$$\Pr\{\Theta_1 = 1, \Theta_2 = 1\} = g(1, 1) = 0.4.$$

Note that this bivariate prior distribution represents the beliefs of the analyst, and need not correspond to the beliefs of any other individual or group. Other individuals may feel quite differently about $\boldsymbol{\Theta}$.

Multivariate prior distributions are sometimes difficult to assess owing to the complexities of thinking in many dimensions simultaneously. It is easier to assess one-dimensional marginal prior distributions than it is to assess the distribution of a person's joint beliefs about several random variables simultaneously. The higher the dimension of the problem, the more this difficulty is exacerbated. We next describe several methods that have been proposed for reducing the difficulties of assessing multivariate prior distributions:

Table 1 Prior distribution: $g(\theta_1, \theta_2)$

$\theta_1 \downarrow \mid \theta_2 \rightarrow$	0	1
0	0.2	0.1
1	0.3	0.4

1. One procedure that has been proposed for assessing multivariate prior distributions involves predicting future values of observables, and then imputing backwards to an implied distribution for the unobservable, unknown indexing parameters (see Kadane et al. [22]).
2. Another proposal for assessing a multivariate prior distribution involves using the assessments of a homogeneous, informed group of experts (see Press [29; 30, Chapter 5]), and combining their assessments into a composite multivariate distribution by means of multivariate **density estimation**.
3. In another approach, Arnold et al. [2] proposed using bivariate normal and bivariate **Pareto** *natural conjugate* families for (Θ_1, Θ_2), (see discussion of natural conjugate families of prior distributions below). The natural conjugate families are conditionally specified distributions of $(\Theta_1|\Theta_2)$ and $(\Theta_2|\Theta_1)$; the appropriate family has eight hyperparameters in the case of the normal, and six hyperparameters in the case of the Pareto. It is suggested that the hyperparameters be assessed using elicitation of many values of conditional **moments** of Θ_1 fixed at various values of Θ_2, followed by elicitation of many values of conditional moments of Θ_2 fixed at various values of Θ_1. By regressing the one set of conditional assessments on the other set, we can estimate the hyperparameters of the joint prior distribution by least squares estimation. It was suggested that the same approach could be applied to more general **exponential families**.

The ability of individuals to assess correlation coefficients was studied by Gokhale & Press [19].

For an outline of numerous methods for assessing prior distributions that have been proposed, see Kass & Raftery [23] and Press [30], Chapter 5.

(Multivariate) Vague Priors

In some situations the analyst does not feel at all knowledgeable about the likely values of unknown, unobservable variables. In such cases he will probably resort to use of a "vague" (sometimes called "diffuse" or "noninformative") prior distribution.

Let $\boldsymbol{\Theta}$ denote a collection of k continuous, unknown variables, each defined on $(-\infty, +\infty)$. $g(\boldsymbol{\theta})$ is a *vague* prior density for $\boldsymbol{\Theta}$ if the elements of

$\boldsymbol{\Theta}$ are mutually independent, and if the probability mass of each variable is diffused evenly over all possible values. We write the (improper) prior density for $\boldsymbol{\Theta}$ as

$$g(\boldsymbol{\theta}) \propto \text{constant},$$

where $\propto$ denotes proportionality. Note that while this density characterization corresponds in form to the density of a **uniform distribution**, the fact that this uniform distribution must be defined over the entire real line means that the distribution is improper except in the case where the components of $\boldsymbol{\theta}$ are defined on a finite interval.

If an unobservable variable were strictly positive, such as an unknown variance, σ^2, then we could adopt a vague prior for σ^2 by considering $\log \sigma^2$ as a new variable defined on $(-\infty, +\infty)$, and taking a vague prior on the variable $\log \sigma^2$, as above. Thus

$$g(\log \sigma^2) \propto \text{constant}.$$

But by a change of variable this implies an (improper) prior for σ^2, i.e.

$$g(\sigma^2) \propto \frac{1}{\sigma^2}.$$

We next extend this idea to multidimensional variables.

The notion of "positive", one-dimensional random variables, extends, in a multivariate context, to "positive definite", when we consider an array (a matrix) of variables. Thus, if $\boldsymbol{\Sigma}$ denotes a k-dimensional square and symmetric **covariance matrix**, and if $\boldsymbol{\Sigma}$ is a positive definite matrix, a vague prior on $\boldsymbol{\Sigma}$ is given by

$$g(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(k+1)/2},$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}$.

This prior density was proposed by Jeffreys [21]. He suggested that to obtain a prior distribution for an unknown $\boldsymbol{\Theta}$, the inferences of which will be invariant under changes in the parameterization of the problem, it is necessary to adopt a prior distribution whose density is expressible as

$$p(\boldsymbol{\theta}) \propto [J(\boldsymbol{\theta})]^{1/2},$$

where $J(\boldsymbol{\theta})$ denotes the Fisher **information matrix**. (For an elaboration of such priors, see, for example, Press [28, Sections 3.6 and 3.8; 41, Sections 2.7.2–2.7.4].) For additional invariance arguments relating to these priors see Hartigan [20], Jeffreys [21], and Villegas [36]. The exponent of $|\boldsymbol{\Sigma}|$ in the Jeffreys invariant prior density presented here was first given by Geisser & Cornfield [13].

For discussions of controversial issues relating to multivariate vague prior distributions, see Press [30], Chapter 5; Stein [33], and Dawid et al. [6]

(Multivariate) Natural Conjugate Priors

It is sometimes convenient for an analyst to confine his description of his prior information about some unobservable $\boldsymbol{\Theta}$ to some preassigned family of distributions. The family most often used is called the *natural conjugate family* of prior distributions (the term and concept is attributable to Raiffa & Schlaifer [31]). The appropriate family is obtained by interchanging the roles of the observable and unobservable random variables in the likelihood function, and then "enriching" the parameters so that they have arbitrary, assignable values. A property of natural conjugate families of prior distributions is that the posterior density belongs to the same family as the prior density. So if the prior density family is **normal**, then the posterior density family will also be normal; if the prior density family is **beta**, then the posterior density family will also be beta, etc.

For example, if $L(\mathbf{x}|\boldsymbol{\theta}) = \mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, where $\mathbf{I}_p$ denotes the p-dimensional identity matrix, and $\mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, denotes the normal distribution with mean vector $\mathbf{0}$, and covariance matrix $\mathbf{I}_p$, then $L(\boldsymbol{\theta}|\boldsymbol{\varphi}, \mathbf{A}) = \mathrm{N}(\boldsymbol{\varphi}, \mathbf{A})$ is a natural conjugate prior distribution for $\boldsymbol{\Theta}$. This result is obtained by writing out the density of $(\mathbf{X}|\boldsymbol{\Theta})$ and noting that if the same density is viewed as a density of $(\boldsymbol{\Theta}|\mathbf{X})$, then the resulting density is proportional to that of a normal distribution. So we adopt a normal distribution as a prior for $\boldsymbol{\Theta}$. We then "enrich" the parameters by adopting completely general parameters for this prior, namely $(\boldsymbol{\varphi}, \mathbf{A})$ (in this way the hyperparameters do not depend upon the sample data). Next, we use our prior beliefs about $\boldsymbol{\Theta}$ to assess the hyperparameters $(\boldsymbol{\varphi},\ \mathbf{A})$.

(Multivariate) Mixture Priors

It sometimes happens that the family of prior distributions being considered for adoption to represent the analyst's prior information about a multivariate probability distribution is not sufficiently rich to capture the nuances of his information. For such situations it

has been proposed [7] that the analyst adopt a prior distribution that is a mixture (a convex combination, in particular) of natural conjugate distributions. Such a class of prior distributions is richer in hyperparameters than a single natural conjugate family, and therefore can better accommodate to the richer information the analyst has about the sampling distribution parameters.

For example, consider the very simple case in which a sample of size n has been taken from a multivariate normal distribution with unknown mean vector $\boldsymbol{\Theta}$, but known covariance matrix $\boldsymbol{\Sigma}_0$, and we have found the sample mean $\bar{\mathbf{x}}$. We can formulate the likelihood function from the fact that $(\bar{\mathbf{x}}|\boldsymbol{\theta}, \boldsymbol{\Sigma}_0) \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_0/n)$. We must next adopt a prior distribution for the unknown $\boldsymbol{\Theta}$. For this purpose we decide to use the mixture density of the normal distribution natural conjugate densities,

$$p_1(\boldsymbol{\theta}) \propto \sum \delta_i \{\mathrm{N}(\boldsymbol{\varphi}_i, \mathbf{A}_i)\}, \quad i = 1, \ldots, m,$$

with $\delta_i \geq 0, \sum \delta_i = 1$.

Note that the hyperparameters $(\delta_i, \varphi_i, A_i, i = 1, \ldots, m)$ must all be preassigned by assessment. Then, the posterior density for $\boldsymbol{\theta}$ becomes

$$p(\boldsymbol{\theta}|\bar{\mathbf{x}}, \boldsymbol{\Sigma}_0) \propto p_1(\boldsymbol{\theta}) p_2(\bar{\mathbf{x}}|\boldsymbol{\theta}, \boldsymbol{\Sigma}_0),$$

or

$$p(\boldsymbol{\theta}|\bar{\mathbf{x}}, \boldsymbol{\Sigma}_0, \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_m, \mathbf{A}_1, \ldots, \mathbf{A}_m, \delta_1, \ldots, \delta_m) \propto \sum \delta_i \left[\exp\left(\frac{-1}{2}\right)(\boldsymbol{\theta} - \boldsymbol{\varphi}_i)' \mathbf{A}_i^{-1} (\boldsymbol{\theta} - \boldsymbol{\varphi}_i)\right] \left[\exp\left(\frac{-n}{2}\right)(\bar{\mathbf{x}} - \boldsymbol{\theta})' \boldsymbol{\Sigma}_0^{-1} (\bar{\mathbf{x}} - \boldsymbol{\theta})\right],$$

or

$$p(\boldsymbol{\theta}|\bar{\mathbf{x}}, \boldsymbol{\Sigma}_0, \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_m, \mathbf{A}_1, \ldots, \mathbf{A}_m, \delta_1, \ldots, \delta_m) \propto \sum \delta_i \left\{\exp\left(\frac{-1}{2}\right)\left[(\boldsymbol{\theta} - \boldsymbol{\varphi}_i)' \mathbf{A}_i^{-1} (\boldsymbol{\theta} - \boldsymbol{\varphi}_i) + n(\bar{\mathbf{x}} - \boldsymbol{\theta})' \boldsymbol{\Sigma}_0^{-1} (\bar{\mathbf{x}} - \boldsymbol{\theta})\right]\right\}.$$

After the two quadratic forms in $\boldsymbol{\theta}$ in the last equation are combined by completing the square on θ, each term in the summation becomes a normal density in $\boldsymbol{\theta}$, but the weights must be modified. We finally obtain the posterior density as the mixture

$$p(\boldsymbol{\theta}|\bar{\mathbf{x}}, \boldsymbol{\Sigma}_0, \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_m, \mathbf{A}_1, \ldots, \mathbf{A}_m, \delta_1, \ldots, \delta_m) \propto \sum \delta_i^* \mathrm{N}(\boldsymbol{\varphi}_i^*, \mathbf{A}_i^*),$$

where

$$(\mathbf{A}_i^*)^{-1} = (\mathbf{A}_i)^{-1} + n\boldsymbol{\Sigma}_0^{-1},$$
$$\boldsymbol{\varphi}_i^* = \mathbf{A}_i^*[(\mathbf{A}_i)^{-1}\boldsymbol{\varphi}_i + n\boldsymbol{\Sigma}_0^{-1}(\bar{\mathbf{x}})]$$

and

$$\delta_i^* = \frac{\left[\delta_i |(\mathbf{A}_i^*)|^{1/2} \exp(-c_i/2)\right]}{\sum [\delta_j |(\mathbf{A}_j^*)|^{1/2} \exp(-c_j/2)]},$$

with

$$\delta_i^* \geq 0, \qquad \sum(\delta_i^*) = 1,$$

and

$$c_i = \boldsymbol{\varphi}_i' \mathbf{A}_i^{-1} \boldsymbol{\varphi}_i + n(\bar{\mathbf{x}}' \boldsymbol{\Sigma}_0^{-1} \bar{\mathbf{x}}) - (\boldsymbol{\varphi}_i^*)' \times (\mathbf{A}_i^*)^{-1} (\boldsymbol{\varphi}_i^*).$$

We note that the posterior density is in the same class as the prior, namely a mixture of normal densities. If the sampling covariance matrix were unknown as well, then the mixture would become more complicated, but could be managed completely analogously.

Exchangeability

A multivariate cumulative distribution function (cdf) that does not depend on the order in which the random variables appear is sometimes referred to as **exchangeable**. The corresponding populations are also said to be exchangeable. Suppose, for example, that $(\Theta_1 \ldots \Theta_k, \ldots)$ are one-dimensional random variables any k of which follow the joint distribution $\mathrm{N}(a\mathbf{e}, \mathbf{H})$, where $\mathbf{e}$ denotes a k-dimensional vector of ones, a denotes any scalar, and $\mathbf{H}$ denotes a covariance matrix with equal diagonal elements, and equal off-diagonal elements. If the Θ_is are permuted, the joint cdf, or joint density, does not change, or any Θ_i could be exchanged for any other, so the distribution is called exchangeable. The original concept was applied to Bernoulli sequences of trials (infinite sequences) (*see* **Binary Data**) and has now been extended to more general sequences.

In some situations in Bayesian multivariate analysis it is useful to adopt an exchangeable prior distribution to express ignorance. For instance, suppose we have observations from three multivariate normal populations with equal covariance matrices, and we wish to carry out Bayesian inference on the mean vectors to compare the closeness of the three populations (**multivariate analysis of variance**). In many situations like this it would not be unreasonable to take the prior distributions for each of the mean vectors to be the same, i.e. to assume, a priori, that the populations are exchangeable (in the absence of any information to the contrary). Thus, if $(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\eta})$ denote the mean vectors for the three normal populations, we could adopt the joint prior distribution for their mean vectors,

$$g(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\eta}) = g^*(\boldsymbol{\Theta})g^*(\boldsymbol{\Phi})g^*(\boldsymbol{\eta}),$$

where the distribution of $\boldsymbol{\Theta}$ (or of $\boldsymbol{\Phi}$, or of $\boldsymbol{\eta}$) is perhaps $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and the hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ must be assessed. Note that such an approach to multivariate prior distribution assessment not only simplifies the analysis, but it greatly reduces the number of hyperparameters that must be assessed.

Numerical Methods of Bayesian Multivariate Analysis

Numerical methods are of fundamental importance in Bayesian multivariate analysis. They are used for evaluating and approximating the normalizing constants of multivariate posterior densities, for finding marginal densities, for calculating moments and other functions of marginal densities, and for sampling from multivariate posterior densities, and for many other needs associated with multivariate Bayesian statistical inference (*see* **Numerical Integration**).

Methods for evaluating and approximating multidimensional integrals associated with multivariate posterior distributions are nicely summarized in Evans & Swartz [11]. They include discussions of the Laplace method, importance sampling and variance reduction techniques, multiple quadrature rules, and Markov chain Monte Carlo (MCMC) methods, including the Metropolis Algorithm.

MCMC, data augmentation, and related methods have been studied and explicated by Metropolis et al. [25], Geman & Geman [17], Tanner & Wong [35], Gelfand & Smith [14], Casella & George [4], Gelman & Rubin [15], Tanner [34, Chapter 6], O'Hagen [26], Chib & Greenberg [5], Gelman et al. [16, Part III], Carlin & Louis [3], and by many other authors who have made a wide variety of contributions to this rapidly expanding field.

Computer-Assisted Bayesian Multivariate Statistical Inference

It was pointed out in the previous section that it is often the case in Bayesian multivariate analysis that posterior distributions are sometimes sufficiently complicated that numerical procedures and computers are required to effect posterior inferences. Fortunately, computer programs have already been written for many of the known multivariate Bayesian inference procedures: see, for example, Spiegelhalter et al. [32], for a general computer program for Gibbs sampling (called BUGS); O'Hagen [26] for a program (based upon the APL language) which introduces Bayes' theorem graphically in one dimension (called "First Bayes"); Albert [1] for a text that introduces the use of Bayesian inference by means of MINITAB 10; Press [27; 30, Chapter 6 and its complements A & B], and Goel [18], for compilations and accompanying descriptions of many Bayesian programs. Most of such computer programs became obsolete shortly after they were written, but they are still useful (*see* **Software, Biostatistical**).

Applications

Multivariate Bayesian analysis abounds with applications in biostatistics. One important area of biostatistical application in multivariate Bayesian analysis involves Bayesian **meta-analysis** (see, for example, DuMouchel [8, 9], DuMouchel & Harris [10], and Lindley & Press [24]).

Some case studies of other applications have been collected as conference proceedings and are detailed in Gatsonis et al. [12]. There are atleast six such volumes now available under the same title (with different authors). Some biostatistical multivariate Bayesian applications presented in the first volume deal with nonignorable nonresponse (*see* **Bias from Nonresponse**); estimation of costs in a sewerage operation; the Ames salmonella/microsome assay (*see* **Mutagenicity Study**); and a cost-utility analysis of breast cancer screening (*see* **Screening Benefit, Evaluation of**).

References

[1] Albert, J. (1996). *Bayesian Computation Using MINITAB*. Duxbury Press, Belmont.

[2] Arnold, B.C., Castillo, E. & Jose Maria Sarabia, J.M. (1996). *Bayesian Analysis for Classical Distributions Using Conditionally Specified Priors*, manuscript, Department of Statistics, University of California, Riverside.

[3] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods For Data Analysis*. Chapman & Hall, London.

[4] Casella, G. & George, E.I. (1992). An introduction to Gibbs sampling, *American Statistician* **46**, 167–174.

[5] Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hasting algorithm, *American Statistician* **49**, 327–335.

[6] Dawid, A.P., Stone, M. & Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference, *Journal of the Royal Statistical Society, Series B* **35**, 189–233.

[7] Diaconis, P. & Ylvisaker, D. (1985). *Quantifying prior opinion, in Bayesian Statistics*, Vol. **2** J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds. North-Holland, Amsterdam, pp. 133–156.

[8] DuMouchel, W.H. (1989). How to perform a Bayesian meta-analysis, in *Statistical Methodology in the Pharmaceutical Sciences*, D.A. Berry, ed. Marcel Dekker, New York.

[9] DuMouchel, W.H. (1994). Predictive cross-validation for hierarchical Bayesian meta-analysis, in *Bayesian Statistics*, Vol. **5** J.M. Berger, J.M. Bernardo, D.V. Lindley & A.F.M. Smith, eds. Oxford University Press, Oxford.

[10] DuMouchel, W.H. & Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species, *Journal of the American Statistical Association* **78**, 293–315.

[11] Evans, M. & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems, *Statistical Science* **10**, 254–272.

[12] Gatsonis, C., Hodges, J.S., Kass, R.E. & Singpurwalla, N.D., eds (1993). *Case Studies in Bayesian Statistics*. Springer-Verlag, New York.

[13] Geisser, S. & Cornfield, J. (1963). Posterior distributions for multivariate normal parameters, *Journal of the Royal Statistical Society, Series B* **25**, 368–376.

[14] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

[15] Gelman, A. & Rubin, D.M. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science* **7**, 457–511.

[16] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, New York.

[17] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[18] Goel, P.K. (1988). Software for Bayesian analysis: current status and additional needs (with discussion), in *Bayesian Statistics*, Vol. **3** J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 173–188.

[19] Gokhale, D.V. & Press, S.J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution, *Journal of the Royal Statistical Society, Series A* **145**, 237–249.

[20] Hartigan, J. (1964). Invariant prior distributions, *Annals of Mathematical Statistics* **35**, 836–845.

[21] Jeffreys, H. (1961, 1966). *Theory of Probability*, 3rd Ed. Clarendon Press, Oxford.

[22] Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S. & Peters, S.C. (1980). Interactive elicitation of opinion for a normal linear model, *Journal of the American Statistical Association* **75**, 845–854.

[23] Kass, R.E. & Raftery, A.E. (1993). Bayes factors and model uncertainty, Technical Report 571. Department of Statistics, Carnegie Mellon University.

[24] Lindley, D.V. & Press, S.J. (1966). Coherent Bayesian meta-analysis, Technical Report 233. Department of Statistics, University of California, Riverside.

[25] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.

[26] O'Hagen, A. (1994). *Bayesian Inference: Volume 2B of Kendall's Advanced Theory of Statistics*. Cambridge University Press, New York/Wiley, New York.

[27] Press, S.J. (1980). Bayesian computer programs, in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner, ed. North-Holland, Amsterdam, Chapter 27.

[28] Press, S.J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd Revised Ed. Krieger, Melbourne.

[29] Press, S.J. (1985). Multivariate group assessment of probabilities of nuclear war (with discussion), in *Bayesian Statistics*, Vol. **2** J.M. Bernardo, M.H. de Groot, D.V. Lindley & A.F.M. Smith, eds. North-Holland, Amsterdam, pp. 425–462.

[30] Press, S.J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons, Inc., New York.

[31] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard University Press, Boston.

[32] Spiegelhalter, D., Thomas, A., Best, N. & Gilks, W. (1994). *BUGS: Bayesian Inference Using Gibbs Sampling*, available from MRC Biostatistics Unit, Cambridge, UK.

[33] Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution, *Proceedings of the Third Berkeley Symposium of Mathematical*

Statistics and Probability, Vol. **1** L. LeCam & J. Neyman, eds. University of California Press, Berkeley, pp. 197–206.

[34] Tanner, A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd Ed. Springer-Verlag, New York.

[35] Tanner, M.A. & Wong, W. (1987). The calculation of posterior distributions (with discussion), *Journal of the American Statistical Association* **82**, 528–550.

[36] Villegas, C. (1969). On the a priori distribution of the covariance matrix, *Annals of Mathematical Statistics* **44**, 1098–1099.

Further Reading

Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, New York.

Box, G.E.P. & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading.

de Finetti, B. (1937). La prevision: ses lois, ses sources subjectives, *Annales de l'Institut Henri Poincaré* **7** 1–68. Reprinted in English translation in *Studies in Subjective Probability*, H.E. Kyburg, Jr & H.E. Smokler, eds. Wiley, New York, 1964x.

de Finetti, B. (1974). *Theory of Probability*, Vols. I and 2, Wiley, New York.

Geisser, S. (1965). Bayesian estimation in multivariate analysis, *Annals of Mathematical Statistics* **36**, 150–159.

Leamer, E.E. (1978). *Specification Searches*. Wiley, New York.

Lindley, D.V. (1965). *Introduction to Probability and Statistics*, Vols. I and 2. Cambridge University Press, Cambridge.

Lindley, D.V. (1972). *Bayesian Statistics: A Review*. SIAM, Philadelphia.

Lindley, D.V. & Novick, M.R. (1982). The role of exchangeability in inference, *Annals of Statistics* **9**, 45–58.

O'Hagan, A. (1996). "First Bayes", A Computer Program for Windows 3.x and Windows 95, Version 1.3, available from the author at e-mail address: `aoh@maths.nott.ac.uk`.

Press, S.J. (1983). Group assessment of multivariate prior distributions, *Technological Forecasting and Social Change* **28**, 247–259.

Villegas, C. (1977). Inner statistical inference, *Journal of the American Statistical Association* **72**, 453–458.

Villegas, C. (1977). On the representation of ignorance, *Journal of the American Statistical Association* **72**, 651–654.

Villegas, C. (1981). Inner statistical inference II, *Annals of Statistics* **9**, 768–776.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

(*See also* **Multivariate Analysis, Overview**)

S. JAMES PRESS

Multivariate Analysis, Overview

Multivariate analysis is concerned with mathematical models for representing multidimensional observations, and methods for analyzing those data. A common example of multivariate data might be the lengths of a particular bone measured at six-month intervals in young children, or the set of subtest scores made by a subject on a cognitive intelligence test. The measurements or scores obtained from each person constitute a *vector-valued* observation. We assume that the vector $\mathbf{x}_i$ obtained on the ith individual, or sampling unit, is an observation on the $p \times 1$ vector random variable $\mathbf{X}$ with some p-dimensional multivariate distribution described by the density function $f(\mathbf{x})$. Traditionally, most of the multivariate methods for continuous variables assume that the population distribution is **multivariate normal**, with the density function

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \times \exp\left[-\left(\tfrac{1}{2}\right)(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

The elements of $\mathbf{x}$ each have the range $-\infty < x_i < \infty$, so the admissible values of $\mathbf{x}$ constitute p-dimensional Euclidean space. The parameters of the multivariate normal distribution are the mean vector $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$ and the **covariance matrix** $\mathrm{E}\{[\mathbf{X} - \mathrm{E}(\mathbf{X})][\mathbf{X} - \mathrm{E}(\mathbf{X})]'\} = \boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is always symmetric, and must be positive definite for the density function to exist. Otherwise, the distribution is said to be *singular*, and must be described by its cumulative distribution function. A multivariate normal random vector with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is denoted by the Wilks symbol, $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The justification for the multivariate normal distribution follows from the **central limit theorem** [3, pp. 81–82], which essentially says that sequences of properly standardized sums of independently distributed random vectors with a common well-defined distribution tend to the multivariate normal form as the number of terms in the sums increases without limit. Often the central limit theorem model of random variables as the sum of many independent underlying random components does not hold, and the multivariate distributions are not even approximately normal. In such cases appropriate **transformations** of the individual variables may give random vectors with distributions closer to that of the multivariate normal.

Some Methods for the Multivariate Normal Distribution

This review of multivariate analysis for continuous variables will be restricted to those methods that arise from the multivariate normal population model. First we extend the common univariate **hypothesis tests** and **confidence** statements on means to the mean vector $\boldsymbol{\mu}$. Next, we generalize the **analysis of variance** for one variable to the case of vector-valued observations. We then describe methods for classifying an observation vector to one of two or more unknown populations by the linear or quadratic discriminant function (*see* **Discriminant Analysis, Linear**). **Principal components analysis** and **factor analysis** will be described as means for dissecting covariance and correlation structures and their matrices.

Inferences About the Multivariate Normal Distribution

Estimation of the Mean Vector and Covariance Matrix

Assume that a random sample of N p-component observation vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ has been drawn from the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ multinormal population. The usual estimates of the mean vector and covariance matrix are the sample mean vector and the sample covariance matrix,

$$\hat{\boldsymbol{\mu}} = \left(\frac{1}{N}\right)\sum_{\mathrm{i}=1}^{N}\mathbf{x}_{\mathrm{i}} = \overline{\mathbf{x}},$$

$$\hat{\boldsymbol{\Sigma}} = \left[\frac{1}{(N-1)}\right]\sum_{\mathrm{i}=1}^{N}(\mathbf{x}_{\mathrm{i}} - \overline{\mathbf{x}})(\mathbf{x}_{\mathrm{i}} - \overline{\mathbf{x}})' = \mathbf{S}.$$

$\overline{\mathbf{x}}$ and $\mathbf{S}$ are unbiased estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. $\overline{\mathbf{x}}$ is also the **maximum likelihood** estimator, and $\mathbf{S}$ can be obtained by maximum likelihood followed by the replacement of the divisor N by $N-1$ to achieve **unbiasedness**. $\overline{\mathbf{x}}$ has the multivariate normal distribution with the obvious mean vector $\boldsymbol{\mu}$ and covariance matrix $(1/N)\boldsymbol{\Sigma}$. The distribution of $\mathbf{S}$ is more complicated. The sums of squares and products matrix

$\mathbf{A} = (N-1)\mathbf{S}$ has the **Wishart distribution** [41] with parameters degrees of freedom $n = N-1$ and covariance matrix $\boldsymbol{\Sigma}$. The density function of the Wishart distribution is

$$f(\mathbf{A}) = \begin{cases} \dfrac{|\mathbf{A}|^{1/2(n-p-1)} \exp\left(\dfrac{-1}{2}\operatorname{tr}\mathbf{A}\boldsymbol{\Sigma}^{-1}\right)}{2^{np/2}\pi^{p(p-1)/4}|\boldsymbol{\Sigma}|^{n/2}\displaystyle\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]}, \\ \qquad \mathbf{A} \text{ positive definite}, \\ 0, \qquad \text{elsewhere.} \end{cases}$$

When $p = 1$ the Wishart density is equivalent to that of a $\chi^2\sigma^2$ random variable, or to a chi-square density if $\boldsymbol{\Sigma} = 1$. The Wishart distribution also has many of the properties of the **chi-square distribution**, e.g. the sum of independent Wishart matrices with a common covariance matrix parameter is also Wishart with degrees of freedom equal to the sum of the degrees of freedom of the individual matrices. The Wishart density is useful as a starting point for deriving estimates, hypothesis tests, and distributions based on the normal distribution covariance matrix.

Another parameterization of the multivariate normal distribution is composed of the mean vector, the p standard deviations or variances, and the matrix of correlations. When the population covariance matrix is a general positive definite matrix the joint distributions of the sample sums of squares and sample correlations are rather complicated, and are not in forms that lend themselves to useful applications.

Inferences About Mean Vectors

Hypothesis tests and confidence intervals for means based on normal distribution theory can be extended to mean vectors. Since the mean vector $\bar{\mathbf{x}}$ of N independently and multinormally distributed p-component vectors is also multivariate normal with covariance matrix $(1/N)\boldsymbol{\Sigma}$, the quadratic form,

$$\chi^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}),$$

has the chi-square distribution with p degrees of freedom. If $\boldsymbol{\Sigma}$ is known, then we can test the hypothesis $\mathrm{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against the general alternative that the mean vector is not $\boldsymbol{\mu}_0$ by rejecting H_0 when χ^2 exceeds some right-hand critical value of the chi-square distribution. The test statistic is the **Mahalanobis distance** of $\bar{\mathbf{x}}$ from the hypothesized population mean vector $\boldsymbol{\mu}_0$. The statistic has an important property: it is invariant under affine transformations $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{h}$ on the sample and population mean vectors. $\mathbf{A}$ is a nonsingular (i.e. $|\mathbf{A}| \neq 0$) $p \times p$ matrix of constants, and $\mathbf{h}$ is a $p \times 1$ nonrandom vector.

If the covariance matrix is unknown and estimated by the sample matrix $\mathbf{S}$, then the hypothesis on the mean vector can be tested by the **Hotelling T^2** statistic [12]:

$$T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0).$$

When the null hypothesis is true, $F = [(N-p)/p(N-1)]T^2$ has the ***F* distribution** with degrees of freedom p and $N-p$, and $\mathrm{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ would be rejected at the α level if F exceeds the critical value $F_{\alpha;p,N-p}$. When the alternative hypothesis, $\mathrm{H}_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_1$, is true, the F statistic has the noncentral F distribution with degrees of freedom $p, N-p$, and noncentrality parameter

$$\delta^2 = N(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

We can use the noncentral F distribution to compute the **power** of the T^2 test for different alternatives, or to determine the sample size N which will provide some minimal power for a given α level.

The single-sample T^2 statistic can also be used to test the equal-means hypothesis for repeated-measures data (*see* **Analysis of Variance for Longitudinal Data**). The p repeated observations on each of the N sampling units are first transformed to $p-1$ successive differences, or $p-1$ differences from the first, last, or other repeated measure response. The null hypothesis of zero means for the $p-1$ new variables is then tested by T^2. An alternative approach to the analysis of repeated measurements or longitudinal data through **generalized estimating equations** has been given by Liang & Zeger [24] and Zeger et al. [42, 43].

The two-sample hypothesis, $\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, of equal mean vectors in two multivariate normal populations with a common covariance matrix $\boldsymbol{\Sigma}$ can also be tested by the T^2 statistic. From the independent random samples of N and M p-component observation vectors, $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$, we compute the respective mean vectors, $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, and

the within-samples covariance matrix, $\mathbf{S}$, defined by

$$\mathbf{S} = \left[\frac{1}{(N+M-2)}\right]\left[\sum_{i=1}^{N}(\mathbf{x}_{i1}-\bar{\mathbf{x}}_1)(\mathbf{x}_{i1}-\bar{\mathbf{x}}_1)' + \sum_{i=1}^{M}(\mathbf{x}_{i2}-\bar{\mathbf{x}}_2)(\mathbf{x}_{i2}-\bar{\mathbf{x}}_2)'\right].$$

($\mathbf{S}$ is an unbiased estimate of $\boldsymbol{\Sigma}$). The two-sample T^2 statistic is

$$T^2 = \left[\frac{NM}{(N+M)}\right](\mathbf{x}_1-\mathbf{x}_2)'\mathbf{S}^{-1}(\mathbf{x}_1-\mathbf{x}_2).$$

When the null hypothesis of a common mean vector is true, $F = [(N+M-p-1)/(N+M-2)p]T^2$ has the F distribution with p and $N+M-p-1$ degrees of freedom, and the null hypothesis would be rejected at the α level if $F > F_{\alpha;p,N+M-p-1}$. When the alternative hypothesis of unequal mean vectors holds, F has the noncentral F distribution with degrees of freedom $p, N+M-p-1$, and noncentrality parameter $[NM/(N+M)](\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$. As in the single-sample test, charts or tables of the noncentral F distribution can be used to find the power of the T^2 test, or to determine the sample sizes N and M that will satisfy specified power and α probabilities. Details and examples have been given by Morrison [26, Section 2.8].

Simultaneous Inferences for Mean Vectors

The previous T^2 tests only tell us whether or not hypotheses on mean vectors should be rejected. If the hypothesis has been rejected, then the test does not tell us which of the p response variables may have contributed to that decision. It is possible to test hypotheses on the individual responses or linear combinations of them with an overall, or "family", error rate α using a method due to Roy & Bose [35] (*see* **Multiple Comparisons**). The Roy–Bose simultaneous tests and confidence intervals are described under the **Hotelling T^2** entry, and in most texts on multivariate analysis, e.g. Morrison [26, Section 2.3].

Multivariate Analysis of Variance

Generalization of Univariate Analysis of Variance

The univariate **analysis of variance** (ANOVA) begins with the partition of a total sum of squares into two independent components. The error component, E, when divided by an appropriate degrees of freedom parameter, always gives an unbiased estimate of the error variance for the underlying linear model. The other component, H, only has an expected value proportional to the error variance if some hypothesis on the model's parameters is true. Under the hypothesis, H/E is proportional to an F random variable, and the hypothesis would be rejected for large values of H/E.

The **multivariate analysis of variance** (MANOVA) for p-dimensional observation vectors is a generalization of ANOVA to hypotheses on mean vectors and matrices. We assume the same mathematical model holds for each of the p response variables, and those variables are jointly distributed according to a p-dimensional multivariate normal distribution with the same covariance matrix for all N sampling units in the experimental design or other investigation. The mean vector of the distribution will depend on the hypothesis being tested, and whether or not it is true. The error sum of squares E is generalized to the $p \times p$ symmetric matrix

$$\mathbf{E} = \begin{bmatrix} e_{11} & \dots & e_{1p} \\ \vdots & \dots & \vdots \\ e_{1p} & \dots & e_{pp} \end{bmatrix},$$

in which e_{jj} = error sum of squares for the ANOVA on the jth response variable, and e_{ij} = sum of products for the ith and jth response variables, obtained by rewriting the sum of squares expression e in one variable as a sum of products in two variables. The generalization of the hypothesis sum of squares is the $p \times p$ symmetric matrix

$$\mathbf{H} = \begin{bmatrix} h_{11} & \dots & h_{1p} \\ \vdots & \dots & \vdots \\ h_{1p} & \dots & h_{pp} \end{bmatrix},$$

where h_{jj} = univariate ANOVA hypothesis sum of squares for the jth response variable, and h_{ij} = sum of products for the ith and jth response variables, obtained by rewriting the sum of squares expression in one variable for h as a sum of products in two variables.

The test statistic for the multivariate analysis of variance null hypothesis that each of the hypotheses for the individual responses is true is some function of the roots of the determinantal equation

$$|\mathbf{H} - \lambda\mathbf{E}| = 0,$$

or equivalently, the characteristic roots (*see* **Eigenvalue**) of the matrix product $\mathbf{E}^{-1}\mathbf{H}$. These include the Wilks [39] determinantal ratio $\Lambda = |\mathbf{E}|/|\mathbf{H} + \mathbf{E}|$ (*see* **Lambda Criterion, Wilks'**), the **Lawley–Hotelling** [16] statistic

$$T_0^2 = \operatorname{tr} \mathbf{H}\mathbf{E}^{-1}$$
$$= \text{the sum of the characteristic roots of } \mathbf{E}^{-1}\mathbf{H},$$

the Roy [33] greatest root statistic c_s = maximum characteristic root of $\mathbf{E}^{-1}\mathbf{H}$ (*see* **Roy's Maximum Root Criteria**), and the **Pillai** [29] **trace statistic** $V = \operatorname{tr} \mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$.

The distributions of the test statistics under the appropriate null hypotheses have been obtained for large sample sizes (*see* **Large-sample Theory**), and in some cases for small samples and special values of the experimental design parameters. The asymptotic theory of generalized **likelihood ratio test** statistics implies that the transformed Wilks Λ,

$$\chi^2 = -\left[N - r - \left(\tfrac{1}{2}\right)(p - g + 1)\right] \ln \Lambda,$$

is distributed as a χ^2 random variable with pg degrees of freedom where

N = the total number of independent observations in the experimental design,
r = the rank of the design matrix, or
$N - r$ = error degrees of freedom for the design,
g = the rank of the hypothesis matrix, $= k - 1$ for the one-way layout,
p = the number of response variables.

Similarly, NT_0^2 has the large-sample χ^2 distribution with pg degrees of freedom when the null hypothesis is true. Heck [11] computed critical values for a transformation of the Roy greatest characteristic root statistic. Charts of those percentage points and tables of additional values computed by K.C.S. Pillai may be found in Morrison [26] and other texts on multivariate analysis. The Pillai trace statistic, $(N - r)V$, has the chi-square distribution with pg degrees of freedom when N is large and the null hypothesis is true.

No statistic appears to be **uniformly most powerful** against all **alternative hypotheses**. Instead, some of the tests have highest **power** against certain types of alternatives. One advantage of the Roy criterion is that its **union–intersection** development leads directly to simultaneous tests and confidence intervals to determine which responses and treatment comparisons may have led to rejection of the overall multivariate hypothesis.

Olson [27] studied the robustness of six multivariate analysis of variance test criteria through extensive sampling studies. He concluded that the Roy greatest characteristic root statistic was least robust under nonnormality or unequal covariance matrices, while the Pillai trace measure seemed most robust.

An Example

We illustrate the one-way multivariate analysis of variance with some measurements on the skulls of the wolf *Canis lupus L.* discussed by Jolicoeur [18, 19] and given as an example by Morrison [26]. These three skull dimensions were chosen:

X_1 = palatal length,
X_2 = postpalatal length,
X_3 = zygomatic width.

The four groups consist of male and female wolves from the Rocky Mountain and Arctic Archipelago areas of northwestern Canada. The measurements, in millimeters, are reprinted in Table 1 with the kind permission of Pierre Jolicoeur.

The data form an unbalanced two-way layout, but we shall treat them as a one-way design for simplicity. The matrices for the multivariate analysis of variance are as follows.

Between-groups sums of squares and products matrix:

$$\mathbf{H} = \begin{bmatrix} 781.16 & 585.28 & 364.29 \\ 585.28 & 445.51 & 245.66 \\ 364.29 & 245.66 & 656.74 \end{bmatrix}.$$

Within-groups (error) sums of squares and products matrix:

$$\mathbf{E} = \begin{bmatrix} 100.60 & 65.60 & 95.63 \\ 65.60 & 165.93 & 149.30 \\ 95.63 & 149.30 & 557.90 \end{bmatrix},$$

$$\mathbf{E}^{-1}\mathbf{H} = \begin{bmatrix} 7.88522 & 5.90989 & 3.20985 \\ 1.36789 & 1.13769 & -0.46443 \\ -1.06476 & -0.87718 & 0.75123 \end{bmatrix}$$

Characteristic roots of $\mathbf{E}^{-1}\mathbf{H}$: 8.5280, 1.20595, 0.04018.

Values of the major multivariate analysis of variance test statistics are given in Table 2. The null

Table 1

Rocky Mountain						Arctic					
Males			Females			Males			Females		
X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
126	104	141	116	102	131	117	99	134	112	94	134
128	111	151	120	103	130	115	100	149	109	91	133
126	108	152	116	103	125	117	106	142	112	99	139
125	109	141				117	101	144	112	99	133
126	107	143				117	103	149	113	97	146
128	110	143				119	101	143	107	97	137
						115	102	146			
						117	100	144			
						114	102	141			
						110	94	132			

Group	Mean		
	X_1	X_2	X_3
Rocky Mountain males	126.50	108.17	145.17
Rocky Mountain females	117.33	102.67	128.67
Arctic males	115.80	100.80	142.40
Arctic females	110.83	96.17	137.00

Unpublished data reproduced by permission of P. Jolicoeur.

Table 2

Test criterion	Statistic	P value
Roy greatest root	8.5280	$\ll 0.01$
Wilks Λ	0.04574	3.17×10^{-10}
Lawley–Hotelling T_0^2	9.7741	1.55×10^{-47}
Pillai trace V	1.4804	0.00029

hypothesis of equal mean vectors for the three skull dimensions in each of the four region and gender groups would be rejected by each of the test statistics.

Multivariate General Linear Model and Hypothesis

The **H** and **E** matrices also follow from a general linear model for a multivariate data matrix and hypothesis tests on the model's parameters. We represent the $N \times p$ data matrix by the following linear model:

$$\mathbf{X} = \mathbf{A}\boldsymbol{\mu} + \mathbf{e} = \mathbf{A}_1\boldsymbol{\mu}_1 + \mathbf{A}_2\boldsymbol{\mu}_2 + \mathbf{e},$$

in which $\mathbf{A}$ is the $N \times q$ design matrix describing the experimental design, $\boldsymbol{\mu}$ is the $q \times p$ matrix of model parameters, $\mathbf{A}_1$ is the $N \times r$ basis of $\mathbf{A}$, $\mathbf{A}_2$ is the $N \times (q - r)$ completion of $\mathbf{A}_1$, $\boldsymbol{\mu}_1$ is the $r \times p$ submatrix of $\boldsymbol{\mu}$ corresponding to the basis of $\mathbf{A}$, and $\boldsymbol{\mu}_2$ is the $(q - r) \times p$ submatrix of the parameters corresponding to the completion of $\mathbf{A}_1$. Note that each of the p columns of $\mathbf{X}$ has the same experimental design matrix. Then the multivariate general linear hypothesis is

$$\mathrm{H}_0 : \mathbf{C}\boldsymbol{\mu}\mathbf{M} = \mathbf{C}_1\boldsymbol{\mu}_1\mathbf{M} + \mathbf{C}_2\boldsymbol{\mu}_2\mathbf{M} = \mathbf{0},$$

and its alternative is simply that $\mathbf{C}\boldsymbol{\mu}\mathbf{M} \neq \mathbf{0}$. $\mathbf{C}$ is a $g \times q$ matrix specifying the hypothesis; $\mathbf{C}_1$ and $\mathbf{C}_2$ are its $g \times r$ and $g \times (q - r)$ submatrices corresponding to the parameter matrices $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and $\mathbf{M}$ is a $p \times u$ matrix of constants that allows for hypotheses on linear compounds of the multivariate response parameters. The **H** and **E** matrices for the general model and hypothesis can be shown to be

$$\mathbf{H} = \mathbf{M}'\mathbf{X}'\mathbf{A}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{C}_1'[\mathbf{C}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{C}_1']^{-1} \times \mathbf{C}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{X}\mathbf{M},$$

$$\mathbf{E} = \mathbf{M}'\mathbf{X}'[\mathbf{I} - \mathbf{A}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1']\mathbf{X}\mathbf{M}.$$

The matrices $\mathbf{C}_2$ and $\mathbf{C}_1$ must satisfy a *testability* condition [34]:

$$\mathbf{C}_2 = \mathbf{C}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{A}_2.$$

Classification and Discrimination

The Linear Discriminant Function

Suppose two multivariate normal populations can be used to describe a p-component random variable. From respective random samples of N and M observations we have estimates $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ of their mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and the within-sample estimate $\mathbf{S}$ of their common covariance matrix $\boldsymbol{\Sigma}$. Now consider a new $(N + M + 1)$th observation vector $\mathbf{x}$ from one of the two populations. We wish to assign $\mathbf{x}$ to population 1 or 2 on the basis of the values of the linear function $\mathbf{a}'\mathbf{x}$, where the vector $\mathbf{a}$ has been chosen to maximize the squared univariate t statistic,

$$t^2(\mathbf{a}) = \frac{[\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\left\{\mathbf{a}'\mathbf{S}\mathbf{a}\left[\left(\frac{1}{N}\right) + \left(\frac{1}{M}\right)\right]\right\}},$$

subject to the condition that $\mathbf{a}'\mathbf{S}\mathbf{a} = 1$. The coefficient vector $\mathbf{a}$ maximizes the absolute distance between the means $\mathbf{a}'\bar{\mathbf{x}}_1$ and $\mathbf{a}'\bar{\mathbf{x}}_2$ given the constraint $\mathbf{a}'\mathbf{S}\mathbf{a} = 1$ on the elements of $\mathbf{a}$. The maximizing vector is given by

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

and the maximum squared t^2 $(\mathbf{a})$ is the Hotelling two-sample statistic:

$$T^2 = \left[\frac{NM}{(N+M)}\right](\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

If each population is equally likely, then we adopt the classification rule:
Assign x to population 1 if $\mathbf{a}'\mathbf{x}$ *is closer to the mean* $\mathbf{a}'\bar{\mathbf{x}}_1$, *and to population 2 if* $\mathbf{a}'\mathbf{x}$ *is closer to the other mean* $\mathbf{a}'\bar{\mathbf{x}}_2$.
Equivalently,
Assign $\mathbf{x}$ *to population 1 if* $\mathbf{a}'\mathbf{x} > \mathbf{a}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$, *and to population 2 otherwise*.
The new variable, $y = \mathbf{a}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\mathbf{x}$, is called the *sample linear discriminant function*. The midpoint, $\mathbf{a}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$, between the means of the linear discriminant function for the two samples is also subject to sampling variation, and the discriminant function is sometimes expressed as the Wald–Anderson statistic:

$$W = \frac{\mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{2}$$

(Wald [38] and Anderson [1]). Then the classification rule in terms of W is:
Assign $\mathbf{x}$ *to population 1 if* $W > 0$, *and to population 2 otherwise*.
The distribution of W is complex for small samples, and the calculation of misclassification probabilities is difficult. A large literature exists on the estimation of different kinds of misclassification rates; some references have been given by Morrison [26].

In the rather artificial case of known parameters the linear discriminant function becomes

$$y = \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

and the classification rule is:
Assign $\mathbf{x}$ *to population 1 if* $\mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 > 0$, *and to population 2 otherwise*.
The linear discriminant function y has the univariate normal distribution, and misclassification probabilities are easily calculated. The classification method can be extended to the **Bayesian** case of **prior** probabilities p and $1 - p$ for the respective populations, and misclassification costs $C(1|2)$ and $C(2|1)$ [1, 3].

For classification with k independent populations we compute the Wald–Anderson statistics,

$$W_{ij} = \frac{\mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) - (\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_j)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)}{2},$$

for all $k(k-1)$ pairs of the k sample mean vectors, and use the rule:
Assign $\mathbf{x}$ *to the ith population if* $W_{ij} > 0$ *for all* $j \neq i$.
We note that $W_{ji} = -W_{ij}$, and of course $W_{ii} = 0$. Alternatively, we can compute the sample **Mahalanobis distances**,

$$D_i^2 = (\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i), \quad i = 1, \ldots, k,$$

and assign $\mathbf{x}$ to the population with the minimum D_i^2. The two rules are algebraically identical.

When the populations have unequal covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ the likelihood ratio classification rule leads to a quadratic discriminant function the form of which, with sample mean vectors and covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$, is

$$h(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x} - \bar{\mathbf{x}}_2) - (\mathbf{x} - \bar{\mathbf{x}}_1)'\mathbf{S}_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_1) - \ln\left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\right).$$

The vector $\mathbf{x}$ is assigned to population 1 if $h(\mathbf{x}) > 0$, and to population 2 otherwise. An alternative classification rule with multivariate normal populations with unequal covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ has been proposed by Anderson & Bahadur [4]. When the parameters of the distributions are known, the vector $\mathbf{x}$ would be assigned by the linear discriminant function

$$y = \mathbf{x}'(t_1\mathbf{\Sigma}_1 + t_2\mathbf{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

to population 1 if y is greater than some constant c, and to Population 2 otherwise. A number of methods have been proposed for determining t_1, t_2, and c. When the parameters are unknown and estimated from the sample data, the linear discriminant function might be written as

$$y = \mathbf{x}'[t\mathbf{S}_1 + (1-t)\mathbf{S}_2]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

and the constant chosen as the midpoint

$$c = \frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'[t\mathbf{S}_1 + (1-t)\mathbf{S}_2]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{2}$$

between the mean values of y for the two samples. The quantity t might be chosen in some pragmatic manner, as, for example, to minimize the number of misclassified individuals.

Inferences About Covariance Matrices

Generalized likelihood ratio tests are available for hypotheses on covariance matrices of multivariate normal distributions. From a single sample we may test that a covariance matrix has a specified form or particular pattern, or that all of its correlations are zero. The generalized likelihood ratio test of zero correlations is particularly simple. We compute the determinant $|\mathbf{R}|$ of the $p \times p$ correlation matrix $\mathbf{R}$, and from it the statistic

$$\chi^2 = -\left[\frac{N - 1 - (2p+5)}{6}\right] \ln |\mathbf{R}|.$$

If the null hypothesis that all $p(p-1)/2$ population correlations are zero is true, then χ^2 has the chi-square distribution with $p(p-1)/2$ degrees of freedom for large N, and the hypothesis is rejected if $\chi^2 > \chi_{\alpha;p(p-1)/2}$.

We may test that k populations have a common covariance matrix, although the usual determinantal form of that test appears to be affected seriously by departures from normality. The hypothesis that two subsets $\mathbf{X}_1$ and $\mathbf{X}_2$ of multivariate normal random variables are independent, or that

$$\mathrm{H}_0 : \mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2') = \mathbf{\Sigma}_{12} = \mathbf{0}$$

is tenable, can be tested by a generalized likelihood ratio test due to Wilks [40], and sharpened in its chi-square approximation by Box [5]. Developments of the likelihood ratio tests can be found in Anderson [3], Morrison [26], and other sources for multivariate analysis.

Roy [33] found a union–intersection test of $\mathrm{H}_0 : \mathbf{\Sigma}_{12} = \mathbf{0}$. If the sample covariance matrix is partitioned as

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{12}' & \mathbf{S}_{22} \end{bmatrix},$$

then the Roy union–intersection test statistic is the greatest characteristic root of the matrix product $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\mathbf{S}_{12}'$ or any cyclic permutation of it, or the corresponding product in the correlation submatrices $\mathbf{R}_{ij}$. Details of the test and critical values have been given by Morrison [26]. The union–intersection test is based on the *largest squared* **canonical correlation** *coefficient*, or the greatest squared sample **correlation** between the linear compounds $U = \mathbf{a}'\mathbf{X}_1$ and $V = \mathbf{b}'\mathbf{X}_2$ of the variables in the first and second subsets. U and V were called the *canonical variates* by Hotelling [14, 15], who proposed their use for determining the nature of the correlation structure between the two subsets. Further pairs of canonical variates can be found from the second greatest, third greatest, etc. characteristic roots of the matrix product.

The Latent Structures of Covariance and Correlation Matrices

The dependence structures of multivariate normal random variables can be analyzed by various representations of the covariance or correlation matrices. The first method, principal components analysis, consists of rotating the coordinate axes of the original p variables to conform with the directions of successively smaller variation. The methodology is due to Hotelling [13], although it had been proposed much earlier by Pearson [28]. The second, factor analysis, is based on the assumption that the correlations among the p observed variables are generated by

a smaller number, m, of uncorrelated latent factor variables.

Principal Components

From the sample of N p-component observation vectors we compute the mean vector $\overline{\mathbf{x}}$ and the covariance matrix $\mathbf{S}$. The first principal component of the sample data, or of $\mathbf{S}$, is the new variable $Y_1 = \mathbf{a}_1'\mathbf{x}$, where the vector of constants $\mathbf{a}_1$ has been chosen to maximize the sample variance of Y_1, or $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$, subject to the constraint $\mathbf{a}_1'\mathbf{a}_1 = 1$. $\mathbf{a}_1$ is the characteristic vector (*see* **Eigenvector**) of $\mathbf{S}$ corresponding to the greatest characteristic root c_1, or the vector satisfying the equations $[\mathbf{S} - c_1\mathbf{I}]\mathbf{a}_1 = \mathbf{0}$, where c_1 is the largest root of the determinantal equation $|\mathbf{S} - c_1\mathbf{I}| = \mathbf{0}$. The remaining principal components are found by extracting the other $p - 1$ characteristic roots and vectors of $\mathbf{S}$: The ith principal component is the variate $Y_i = \mathbf{a}_i'\mathbf{x}$ with variance $c_i = \mathbf{a}_i'\mathbf{S}\mathbf{a}_i$. If the p characteristic roots are distinct then the characteristic vectors are mutually **orthogonal**, and the principal components are uncorrelated. The total variance of the principal components is

$$c_1 + \cdots + c_p = \operatorname{tr}\mathbf{S} = s_1^2 + \cdots + s_p^2,$$

and the proportion of the total variance due to the ith component is $c_i/tr\ \mathbf{S}$. In practice we prefer to consider the first few components that account for as much of the total variance as possible. If, for example, components 1, 2, and 3 explain 86% of the total variance, then we might consider the true dimensionality of the sample to be three rather than p.

Geometrically, $\mathbf{a}_i$ is interpretable as the vector of direction cosines of the line from the mean $\overline{x}$ through the direction of the ith greatest variation in the p-dimensional scatter plot of the data. When the c_i are distinct, the successive axes are orthogonal, or perpendicular, to one another. Each axis has an orientation that minimizes the sum of squared perpendicular distances to the data points, so that principal components analysis amounts to an "orthogonal least squares fit" of the component axes.

Certain patterned covariance matrices have distinctive principal components. If $\mathbf{S}$ is a diagonal matrix with successive diagonal elements $c_1 > c_2 > \ldots > c_p$, then the principal component axes are identical to those of the original variables. If the $p \times p$ matrix $\mathbf{S}$ has the equal-variance, equal-covariance pattern

$$\mathbf{S} = \begin{bmatrix} a & b & \ldots & b \\ b & a & \ldots & b \\ \vdots & \vdots & \cdots & \vdots \\ b & b & \ldots & a \end{bmatrix},$$

then the first principal component is $Y_1 = (x_1 + \ldots x_p)/\sqrt{p}$, and its variance is $c_1 = a + (p-1)b$. The second through pth components each have the same variance, $a - b$, and coefficient vectors that are mutually orthogonal with elements that sum to zero so that they are also orthogonal to the first component. Such a covariance matrix is called *semi-isotropic*, because it describes an ellipsoid in p-dimensional space with a single long axis (if $b > 0$), and $p - 1$ axes of equal lengths around it.

When the original variables are incommensurable or have very different variances, it may be more meaningful to extract principal components from the correlation matrix. This is equivalent to a component analysis on the standard scores $z_{ij} = (x_{ij} - \overline{x}_j)/s_j$ computed from the original observations x_{ij} obtained from the ith sampling unit's jth variable. Then the total variance for the p components will always be p.

A large-sample distribution theory for principal components has been developed by Girshick [7, 8], Anderson [2], and others. This includes tests and confidence intervals for component variances, and tests for component vectors. Asymptotic results for inferences on components extracted from correlation matrices are limited in scope, and depend on very complicated distributions. **Jackknife** and **bootstrap** methods show some promise for simpler inferences in this area.

Factor Analysis

Just as principal components were invented to explain portions of the total variance, factor analysis is designed to explain the correlation structure of multivariate data. We postulate that the observable $p \times 1$ multivariate normal random vector $\mathbf{X}$ with parameters $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{X}, \mathbf{X}') = \boldsymbol{\Sigma}$ is related to the $m \times 1$ latent vector $\mathbf{Y}$ by the linear model

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{Y} + \mathbf{e},$$

where $\mathbf{e}$ is another $p \times 1$ random vector distributed independently of $\mathbf{Y}$. The number of latent variables m is much smaller than p. $\mathbf{Y}$ has expectation $\mathrm{E}(\mathbf{Y}) = \mathbf{0}$

and covariance matrix $\text{cov}(\mathbf{Y}, \mathbf{Y}') = \mathbf{I}$. $\text{E}(\mathbf{e}) = \mathbf{0}$ and $\text{cov}(\mathbf{e}, \mathbf{e}') = \boldsymbol{\Psi}$, a diagonal matrix with ith diagonal element ψ_i. Hence, $\text{E}(\mathbf{X}) = \boldsymbol{\mu}$, and

$$\begin{aligned}\text{cov}(\mathbf{X}, \mathbf{X}') &= \boldsymbol{\Sigma} \\ &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}.\end{aligned}$$

The factor model representation of $\mathbf{X}$ has led to the decomposition $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ of the original covariance matrix. For our purposes factor analysis will consist of the estimation of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, the expression of $\boldsymbol{\Lambda}$ in an "optimal" form, and the test of the fit of the observed covariance matrix to that reproduced by the factor model.

The matrix $\boldsymbol{\Lambda}$ of **factor loading** *parameters* is not unique. If $\mathbf{T}$ is an $m \times m$ orthogonal matrix, or one such that $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$, then the new matrix $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\mathbf{T}$ is equally suited for reproducing the covariance matrix, for

$$\boldsymbol{\Sigma} = \boldsymbol{\Psi} + \boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{\Psi} + (\boldsymbol{\Lambda}\mathbf{T})(\boldsymbol{\Lambda}\mathbf{T})' = \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}'.$$

The selection of the matrix $\mathbf{T}$ is known *as factor* **rotation**, for multiplication by the orthogonal matrix $\mathbf{T}$ is equivalent to a rigid rotation of the m factor axes. In practice, $\mathbf{T}$ is often found by using such algorithms as the Kaiser varimax method [20–22] (*see* **Varimax Rotation**), or occasionally, when m is small, by successively graphically rotating pairs of the factor axes to obtain as many large and as many nearly-zero loadings as possible. Sometimes, to achieve this **oblique rotations** are employed. This makes the factors correlated.

The elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ can be estimated by maximum likelihood under the assumption that the N independent observations on $\mathbf{X}$ are from the nonsingular multivariate normal distribution. This approach is due to Lawley [23], although it has been recast by Rao [31] and others in the context of characteristic roots and vectors. Allowance has also been made by most current statistical programs for global maxima on the boundary of the parameter space. For the determination of stationary maxima the Wishart density of the sample covariance matrix $\mathbf{S}$ can be substituted for the likelihood of the original data, and then maximized with respect to $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. Lawley obtained the following nonlinear equations for the estimates:

$$\begin{aligned}\mathbf{S}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} &= \boldsymbol{\Lambda}(\mathbf{I} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}), \\ \text{diag}\,(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) &= \text{diag}\,(\mathbf{S}),\end{aligned}$$

where the second equation merely requires that the diagonal elements of the covariance matrix $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}$ reproduced by the factor model are the same as those in the original matrix $\mathbf{S}$. Unlike principal components scale transformations on the original p variables simply transform the rows of the loading matrix $\boldsymbol{\Lambda}$ by the same amounts. Hence, the factor loadings obtained from the correlation matrix of the original variables differ from those of the covariance matrix only by scale factors.

The maximum likelihood estimation process leads to a test of the fit of the m-factor model to the observed covariance structure. The generalized likelihood ratio principle leads to the following goodness-of-fit test statistic:

$$\chi^2 = \left[\frac{N - 1 - (2p + 5)}{6} - \frac{2m}{3}\right] \times \ln\left(\frac{|\hat{\boldsymbol{\Psi}} + \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'|}{|\mathbf{S}|}\right),$$

which is distributed as a chi-square variate with $[(p - m)^2 - p - m]/2$ degrees of freedom when N is large. An approximation to the test statistic is

$$\chi^2 = \left[\frac{N - 1 - (2p + 5)}{6} - \frac{2m}{3}\right] \times \sum\sum_{i<j} \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{\hat{\psi}_i \hat{\psi}_j},$$

or a measure of the closeness of the observed and reproduced covariances. In either case the hypothesis of an m-factor model is rejected for large values of the statistic.

Several other methods are available for data reduction and latent structure analysis. Multidimensional scaling was originally introduced by Richardson [32] and developed by Torgerson [36, 37] as a means of approximately representing multivariate data in lower-dimensional spaces. Gabriel [6] proposed the biplot for describing both the relationships among multivariate observations and the dependence structure of the variables (*see* **Graphical Displays**). Guttman [9, 10] proposed his simplex and circumflex models for responses ordered on a line or circle.

Nonnormal Multivariate Methods

Although in this review of multivariate methods we have largely assumed multivariate normal populations, in some cases alternative approaches not based on normality are available. **Logistic regression** may be substituted for discriminant analysis; Press & Wilson [30] have compared the two approaches with multivariate data sets, and prefer logistic regression in the case of nonnormality. Howe [17] showed that the normal-theory maximum likelihood estimation equations in factor analysis also follow from a minimum partial correlation argument independent of the multivariate normality assumption. Mooijart [25] has developed estimation methods for factor analysis models that do not assume multinormality, but do require the computation of higher-order crossproduct matrices of the variables. Some of the effects of nonnormality are described in the article **Multivariate Techniques, Robustness**.

References

[1] Anderson, T.W. (1951). Classification by multivariate analysis, *Psychometrika* **16**, 31–50.

[2] Anderson, T.W. (1963). Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics* **34**, 122–148.

[3] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[4] Anderson, T.W. & Bahadur, R.R. (1962). Classification into two multivariate normal distributions with different covariance matrices, *Annals of Mathematical Statistics* **33**, 420–431.

[5] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika* **36**, 317–346.

[6] Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika* **58**, 453–467.

[7] Girshick, M.A. (1936). Principal components, *Journal of the American Statistical Association* **31**, 519–528.

[8] Girshick, M.A. (1939). On the sampling theory of roots of determinantal equations, *Annals of Mathematical Statistics* **10**, 203–224.

[9] Guttman, L. (1954). A new approach to factor analysis: the radex, in *Mathematical Thinking in the Social Sciences*, P.F. Lazarsfeld, ed. Free Press, New York, pp. 258–348.

[10] Guttman, L. (1955). A generalized simplex for factor analysis, *Psychometrika* **20**, 173–192.

[11] Heck, D.L. (1960). Charts of some upper percentage points of the distribution of the largest characteristic root, *Annals of Mathematical Statistics* **31**, 625–642.

[12] Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.

[13] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441, 498–520.

[14] Hotelling, H. (1935). The most predictable criterion, *Journal of Educational Psychology* **26**, 139–142.

[15] Hotelling, H. (1936). Relations between two sets of variates, *Biometrika* **28**, 321–377.

[16] Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 23–41.

[17] Howe, W.G. (1955). *Some Contributions to Factor Analysis*. Oak Ridge National Laboratory, Oak Ridge.

[18] Jolicoeur, P. (1959). Multivariate geographical variation in the wolf *Canis lupus L.*, *Evolution* **XIII**, 283–299.

[19] Jolicoeur, P. (1975). Sexual dimorphism and geographical distance as factors of skull variation in the wolf *Canis lupus L.*, in *The Wild Canids*, M.W. Fox, ed. Van Nostrand Reinhold, New York.

[20] Kaiser, H.F. (1956). The Varimax Method in Factor Analysis, *Ph.D. dissertation*. University of California, Berkeley.

[21] Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

[22] Kaiser, H.F. (1959). Computer program for varimax rotation in factor analysis, *Journal of Educational and Psychological Measurement* **19**, 413–420.

[23] Lawley, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood, *Proceedings of the Royal Society of Edinburgh* **60**, 64–82.

[24] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[25] Mooijart, A. (1985). Factor analysis for non-normal variables, *Psychometrika* **50**, 323–342.

[26] Morrison, D.F. (2004). *Multivariate Statistical Methods*, 4th Ed. Duxbury Press, Pacific Grove, CA.

[27] Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

[28] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine* **2**, 559–572.

[29] Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis, *Annals of Mathematical Statistics* **26**, 117–121.

[30] Press, S.J. & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association* **73**, 699–705.

[31] Rao, C.R. (1955). Estimation and tests of significance in factor analysis, *Psychometrika* **20**, 93–111.

[32] Richardson, M.W. (1938). Multidimensional psychophysics, *Psychological Bulletin* **35**, 659–660.

[33] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[34] Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

[35] Roy, S.N. & Bose, R.C. (1953). Simultaneous confidence interval estimation, *Annals of Mathematical Statistics* **24**, 513–536.

[36] Torgerson, W.S. (1952). Multidimensional scaling. I–Theory and method, *Psychometrika* **17**, 401–419.

[37] Torgerson, W.S. (1958). *Theory and Methods of Scaling*. Wiley, New York.

[38] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups, *Annals of Mathematical Statistics* **15**, 145–162.

[39] Wilks, S.S. (1932). Certain generalizations in the analysis of variance, *Biometrika* **24**, 471–494.

[40] Wilks, S.S. (1935). On the independence of k sets of normally distributed statistical variables, *Econometrica* **3**, 309–326.

[41] Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate population, *Biometrika* **20**, 32–52.

[42] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

[43] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1989). Correction to "Models for longitudinal data: A generalized estimating equation approach", *Biometrics* **45**, 347.

DONALD F. MORRISON

Multivariate Bartlett Test

In multivariate analysis we often employ **Hotelling's T^2 test** and **multivariate analysis of variance** (MANOVA) to compare two or more mean vectors. One of the underlying assumptions for the T^2 and MANOVA tests is that the corresponding population **covariance matrices** are equal. It has shown that the T^2 and MANOVA tests are fairly robust to heterogeneity of covariance matrices as long as the sample sizes are large and equal [7]. For other cases we may use the multivariate Bartlett test to test the homogeneity of covariance matrices.

The univariate **Bartlett test** was proposed in 1937 to test the homogeneity of variances [2]. It was later extended to the multivariate case [3]. Assume independent samples of size $n_1, n_2, \ldots, n_k$ are randomly drawn from k multivariate normally distributed populations. The **null hypothesis** of equality of covariance matrices is given by

$$\mathrm{H}_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_k.$$

Suppose p is the number of variables involved, so $\mathbf{\Sigma}_i$ is of size $p \times p$. To perform the test, we calculate

$$M = \frac{|\mathbf{S}_1|^{\nu_1/2}|\mathbf{S}_2|^{\nu_2/2}\ldots|\mathbf{S}_k|^{\nu_k/2}}{|\mathbf{S}_{\mathrm{pl}}|^{\nu_{\mathrm{E}}}},$$

where $\nu_i = n_i - 1$, $\nu_{\mathrm{E}} = \sum_{i=1}^{k} \nu_i = \sum_{i=1}^{k} n_i - k$, $\mathbf{S}_i$ is the covariance matrix of the ith sample, and $\mathbf{S}_{\mathrm{pl}}$ is the pooled sample covariance matrix

$$\mathbf{S}_{\mathrm{pl}} = \frac{\sum_{i=1}^{k} \nu_i \mathbf{S}_i}{\sum_{i=1}^{k} \nu_i}.$$

A tractable expression for the exact null distribution of M exists only for the case $k = 2$ [1, 6]. For $k > 2$ populations, Box [4, 5] provided **chi-square** and ***F* distribution** approximations for the distribution of M. They are both referred to as Box's M test. For the χ^2 approximation, we use the statistic u which is given as

$$u = -2(1 - c_1) \ln M,$$

where

$$c_1 = \left[\sum_{i=1}^{k} \frac{1}{\nu_i} - \left(\frac{1}{\sum_{i=1}^{k} \nu_i} \right) \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \right].$$

For computational purposes, we may use the following form of $\ln M$:

$$\ln M = \frac{1}{2} \sum_{i=1}^{k} \nu_i \ln |\mathbf{S}_i| - \frac{1}{2} \left(\sum_{i=1}^{k} \nu_i \right) \ln |\mathbf{S}_{\mathrm{pl}}|.$$

u is approximately distributed as $\chi^2_{1-\alpha}\left[\frac{1}{2}(k-1)\ p(p+1)\right]$. We reject H_0 if $u > \chi^2_{1-\alpha}$.

For the F approximation, the statistic depends on two quantities, c_1 and c_2, where c_1 is defined as above and c_2 is defined as follows:

$$c_2 = \frac{(p-1)(p+2)}{6(k-1)} \left[\sum_{i=1}^{k} \frac{1}{\nu_i^2} - \left(\frac{1}{\left(\sum_{i=1}^{k} \nu_i \right)^2} \right) \right].$$

If $c_2 > c_1^2$, then

$$F = -2b_1 \ln M.$$

If $c_2 < c_1^2$, then

$$F = -\frac{a_2 b_2 \ln M}{a_1(1 + 2b_2 \ln M)},$$

where

$$a_1 = \tfrac{1}{2}(k-1)p(p+1), \qquad a_2 = \frac{a_1 + 2}{|c_2 - c_1^2|},$$

$$b_1 = \frac{1 - c_1 - a_1/a_2}{a_1}, \qquad b_2 = \frac{1 - c_1 - 2/a_2}{a_2}.$$

Both F_{s} are approximately distributed as $F_{1-\alpha}(a_1, a_2)$.

Olsen [7] showed that the Box M test is very sensitive to some forms of nonnormality for which the MANOVA tests are rather robust. Therefore, in some cases the M test may indicate types of covariance heterogeneity that have only inconsequential effects on the MANOVA tests. Hence, the test is not recommended as a routine diagnostic for MANOVA tests.

References

[1] Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] Bartlett, M.S. (1937). Properties of sufficiency and statistical tests, *Proceedings of the Royal Society of London, Series A* **160**, 268–282.

[3] Bartlett, M.S. (1938). Further aspects of the theory of multiple regression, *Proceedings of the Cambridge Philosophical Society* **34**, 33–40.

[4] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika* **36**, 317–346.

[5] Box, G.E.P. (1950). Problems in the analysis of growth and linear curve, *Biometrics* **6**, 362–389.

[6] Khatri, C.G. & Srivastava, M.S. (1971). On exact non-null distributions of likelihood ratio criteria for sphericity test and equality of two covariance matrices, *Sankhyā* **33**, 201–206.

[7] Olsen, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

RALPH B. D'AGOSTINO, SR & HEIDY K. RUSSELL

Multivariate Classification Rules: Calibration and Discrimination

This article emphasizes multivariate classification rules, or models, where the classification is into one of two possible states, but also discusses extensions to multistate classifications. The accuracy of fit of a model is the degree to which the predicted values, $\hat{p}_i$, coincide with the observed outcomes, y_i, when these form a sample of size N drawn from a **random variable** Y. When the outcome is binary, being either a positive ($y_i = 1$) or a negative ($y_i = 0$) outcome the predicted values are often expressed as probabilities. Because the outcome takes only two values, predicted and observed values will not match closely for each observation as they might for a continuous outcome. Instead, models can be checked for good *discrimination* (also called *resolution* or *refinement*) and *calibration* (or *reliability*). Discrimination (*see* **Discriminant Analysis, Linear**) refers to the ability of the model to correctly distinguish the two classes of outcomes with distinct predicted values. **Calibration** of a probabilistic model describes how closely the predicted probabilities agree numerically with the actual outcomes. For example, events with a predicted outcome probability of 60% should occur about six times in 10. Generally, a model is well calibrated only when predicted and observed values agree for any reasonable grouping of the observations, whether ordered by increasing predicted values or selected according to some external characteristic like a risk factor.

Although a model with good calibration will tend to have good discrimination, and vice versa, a given model may be strong in one measure and weak in another. A model that predicts all negative outcomes to occur with probability 0.49 and all positive outcomes to occur with probability 0.51 has perfect discrimination but bad calibration, whereas a model that predicts all events to occur with probability equal to the prevalence of the outcome has perfect calibration but no discrimination. However, given the choice, some have recommended that good discrimination be preferred to good calibration if prediction is the goal, because a model with good discrimination can always be recalibrated, but the rank orderings of the probabilities cannot be changed to improve discrimination [7, 21]. A model with good discrimination can distinguish when events will occur for an individual and not just on average. Diamond [4] shows that a model with predictions **uniformly distributed** over [0, 1] and perfect calibration, in which the observed event rate is the same as the predicted rate for any subset of observations, cannot have perfect discrimination (as defined by the area under the **receiver-operating characteristic (ROC) curve**).

Brier Score

Mean squared error is a standard measure of the fit of a model. For binary data, the mean squared error, $\Sigma(y_i - \hat{p}_i)^2/N$, is called the *Brier score (BS)*. Here $\hat{p}_i$ is the predicted probability of a positive outcome or event for individual i and N is the total sample size. The Brier score can be considered a weighted **loss function** in which increasing distance between observed and predicted is penalized by a quadratic measure. Other scoring rules which have been suggested are the logarithmic rule [18] and the spherical rule [20]. These three rules are called *proper scoring rules* because they cannot be improved by giving any predictions other than those consistent with the long-run frequency probabilities of the system. In other words, if the modeler knew the true event rate in each prediction category, then he could do no better than predicting that true rate.

The numerical value of the Brier score has no direct meaning, but some weak standards of comparison are available. The simplest is obtained by noting that a prediction of 0.5 for each individual results in a Brier score of 0.25. Another reference value is the Brier score calculated assuming that all individuals are given a predicted probability equal to the prevalence, $\overline{y}$. Writing this as $B_0 = \Sigma(y_i - \overline{y})^2/N = \overline{y}(1 - \overline{y})$, we see that this is simply the variance of the y_i. This suggests that we can form a statistic like the multiple **correlation** coefficient R^2 by normalizing the Brier score B_1 as $R^2 = (B_0 - B_1)/B_0$, thus allowing a comparison of Brier scores across models fit to data with different prevalences [13]. This is important because the Brier score can be changed simply by changing the prevalence. A number of decompositions of the Brier score into components of calibration and discrimination have been suggested [12, 15, 21].

Sanders' decomposition [15] is

$$BS = \frac{\sum_j n_j(\overline{y}_j - \hat{p}_j)^2}{N} + \frac{\sum_j n_j \overline{y}_j(1-\overline{y}_j)}{N},$$

where the summation is over all groups with distinct predicted probabilities $\hat{p}_j$ having an observed event proportion of $\overline{y}_j$ on a sample of size n_j. (In this article, the subscript j refers to a group of individuals while the subscript i refers to single individuals.) The first component describes calibration. It is minimized when the observed event rate is the same as the predicted rate for each unique prediction. The second component describes resolution and is minimized if each $\overline{y}_j$ equals 0 or 1. It therefore measures how well the predictions divide the outcomes into homogeneous collections. As the observed group event rates approach 0.5, the resolution decreases because the predictions are not able to differentiate between outcome categories. Geometrically, the Sanders decomposition is similar to an **analysis of variance** decomposition of the distance between observed and predicted expressed as the sum of the distance between the value of an observation and the mean of the group in which the observation is placed plus the distance between the group mean and the predicted value.

Murphy [12] noted that the Sanders resolution may be inflated by the overall prevalence of the outcome. He therefore separated the resolution component into two pieces in order to describe the resolution adjusted for prevalence. In the Murphy decomposition,

$$BS = \overline{y}(1-\overline{y}) + \frac{\sum_j n_j(\overline{y}_j - \hat{p}_j)^2}{N} - \frac{\sum_j n_j(\overline{y}_j - \overline{y})^2}{N},$$

because the first term is the baseline Brier score, B_0, the sum of the second and third terms must be the negative of the numerator of the normalized Brier score represented by R^2. This representation shows explicitly how the Brier score is affected by changing prevalence. The second term of the Murphy decomposition is just the Sanders calibration. The third term is then the part of the Sanders resolution that does not depend on the overall prevalence and can be controlled by the modeler. This corrected resolution component improves as the mean event rates within each group of observations are differentiated from each other and from the overall event prevalence. The Sanders resolution may be high simply because low prevalence makes division into homogeneous groups easy. Most groups may have zero events because few events occur. Such groups contribute little to the Murphy resolution, however, if their mean rate is nearly the same as the overall mean rate.

Yates [21] described a third decomposition of the Brier score as

$$BS = \overline{y}(1-\overline{y}) + (\hat{p}-\overline{y})^2 + \frac{\sum_j n_j(\hat{p}_j - \hat{p})^2}{N} - 2\frac{\sum_j n_j(\hat{p}_j - \hat{p})(\overline{y}_j - \overline{y})}{N}.$$

The first term is the variance of the outcomes as in Murphy's representation. The second term involves the simplest global measure of calibration, the **bias**, which is the difference between the mean predicted value and the mean event rate. Yates calls it *calibration-in-the-large* to distinguish it from the Sanders/Murphy calibration. Models that are biased may systematically over- or underpredict the true outcome rate, even when discrimination is very good. In **logistic regression**, the bias is zero. Even when a model is unbiased, however, it may still be poorly calibrated if some groups of individuals are badly overpredicted while others are underpredicted. The third term in the Yates decomposition is the variance of the predicted values and the last term is twice the covariance between the predicted and observed values. Yates calls this the covariance decomposition of the Brier score. The covariance term can also be written as $\overline{y}(1-\overline{y})(\hat{p}_1 - \hat{p}_0)$ in terms of the *mean discrimination*, the difference $\hat{p}_1 - \hat{p}_0$ between the average predicted probabilities in the positive and negative outcome groups. Yates called mean discrimination the *slope* because it is the slope estimate if the predicted probabilities are regressed against a dummy variable representing outcome status. A good model should have a large slope, with a maximum of 1 for binary variables coded as 0 or 1, and therefore the sign of the covariance term above is negative.

Yates also noted that the predicted variance in the third term of his decomposition could be made zero if the model always predicted the same value, but that this would lead to a mean discrimination of zero. Therefore the third term in the decomposition can be rewritten as

$$\frac{\sum_j n_j(\hat{p}_j - \hat{p})^2}{N} = \frac{n_0\text{var}(\hat{p}_0) + n_1\text{var}(\hat{p}_1)}{N} + \overline{y}(1-\overline{y})(\hat{p}_1 - \hat{p}_0)^2,$$

the sum of the pooled variance in the predicted values across the two outcome groups and a term involving the slope. For a given slope, the second term, $\overline{y}(1-\overline{y})(\hat{p}_1 - \hat{p}_0)^2$, is the minimum that the prediction variance can achieve. This minimum variance is reached only if within each outcome class the predicted values are the same. The pooled variance term, which Yates called the *scatter*, measures the consistency of the predictions in the two outcome groups. High values of scatter indicate predictions that vary substantially. Taken together, the Yates decomposition is then composed of terms representing prevalence, bias, scatter, and slope. For logistic regression, a model can be described by the scatter and the slope. The slope describes how well the model responds to signals that discriminate events from nonevents (i.e. positive from negative outcomes) and the scatter describes how well the model filters out noise.

One disadvantage to the Brier score and its decompositions is that because all the relevant sampling distributions are not known, **hypothesis tests** and **confidence intervals** may be unavailable [22]. A test of mean discrimination, $\hat{p}_1 - \hat{p}_0$, can be made by performing a **Student's *t* test** comparing the predicted probabilities in the two outcome groups. The pooled variance for this test is simply the scatter measure, $[n_0\text{var}(\hat{p}_0) + n_1\text{var}(\hat{p}_1)]/N$. A test of overall bias, $\hat{p} - \overline{y}$, can be made by comparing the sample bias to its standard error $\left[\sum p_i(1-p_i)/N\right]^{1/2}$ assuming that the bias follows a normal distribution based on the use of $\hat{p}$ and $\overline{y}$ [8]. Spiegelhalter [19] describes a test of calibration based on the Brier score that can be used to determine if the observed score is significantly different from the score expected under the hypothesis of perfect model calibration that the expected value of $y_i = \hat{p}_i$ for all individuals. Because the Brier score is a weighted average of independent Bernoulli random variables (*see* **Binary Data**), an asymptotic test may be constructed on the basis of a standardized normal test statistic with expectation $\sum p_i(1-p_i)/N$ and variance $\sum p_i(1-p_i)(1-2p_i)^2/N^2$. Redelmeier et al. [14] have also developed a test on the basis of a normal approximation for comparing Brier scores from two different models.

Model Discrimination

The Sanders and Murphy resolution statistics measure how well models assign different predicted probabilities to groups of observations with different event rates, but fail as true discrimination measures because they do not consider the correct ordering of the probabilities. For example, a model that assigns predicted values less than 0.5 to all observations on which an event was recorded and predicted values greater than 0.5 to all observations on which an event was not recorded would be perfectly resolved, but would mean quite the opposite of what the probabilities were intended to mean.

A simple discrimination measure that does preserve ordering is the *nonerror rate*, the proportion of observations for which the outcome falls in that category with the higher predicted value. For a binary outcome this is the probability that among events the predicted value was greater than 0.5 and among nonevents the predicted probability was less than 0.5. Predictions of 0.5 get scored as one-half. A similar type of test statistic can be formed by computing the average probability across all observations assigned to the outcome that occurs. Both these tests are discussed in Hilden et al. [8] who give appropriate standard errors.

A good measure of discrimination should also be independent of model calibration. The nonerror rate has perfect **sensitivity** and **specificity** if 0.5 completely separates the two outcome categories. Recalibrating the predictions to a uniformly lower or higher level, however, may destroy the perfect discrimination. Sensitivity and specificity defined with respect to a fixed cutpoint would also fall into this category. Mean discrimination also suffers by this criterion because its magnitude depends on the levels of the predicted probabilities in the outcome groups.

The *c*-index, or *ROC curve area*, has become a standard of model discrimination for multivariate logistic regression because it summarizes a model's pairwise discrimination and depends only on the

rank ordering of the predicted values. It is calculated as the probability that among all possible pairs of individuals with different outcomes the predicted probability for the one with positive outcome is higher than for the one with negative outcome. Any pair with equal predicted probabilities gets half credit. Hanley & McNeil [6] showed for binary outcomes that this Wilcoxon rank sum statistic (*see* **Wilcoxon–Mann–Whitney Test**) is numerically equivalent to the area under the ROC curve formed by plotting the true positive rate against the false positive rate for all possible cutpoints that divide the [0, 1] probability interval into two parts. A model with perfect discrimination has a value of 1.0 for this statistic, while a model with no discrimination in which probabilities are assigned randomly has a value of 0.5. The c-index can also be expressed as Somers' D rank correlation statistic using the linear transformation $D = 2(c - 0.5)$ [7]. Standard errors are then readily available for these statistics using, for example, formulas for the Wilcoxon statistic [11].

The ROC area statistic (as well as other discrimination and calibration measures) can be affected by specifics of the population used in the development of a predictive model, such as its prevalence and case mix. Consider a population consisting of a mixture of two groups, one with an event rate of 0.2 and the other with a rate of 0.8. The true ROC area for the population depends on the mixture proportions of the groups. A 1:1 mixture has an ROC area of 0.8, but a 4:1 mixture has an ROC area of 0.72. Thus a model developed from the first population would appear better than one from the second population even though the only difference was the composition of the population.

Another measure of discrimination that can be classified as a rank-order statistic takes the ratio of outcomes that occur among observations with predictions in a high percentile with those among observations with predictions in the corresponding low percentile. The lowest and highest quartiles, quintiles, or deciles could also be used (*see* **Quantiles**). This measure ignores the middle part of the data and so is more sensitive to discrepancies in the tails of the predicted distribution. This may be of interest in models where definitive decisions are made if the model indicates a very high or very low risk, but no decision is made otherwise.

Model Calibration

Calibration measures may be characterized as statistics that partition the data into groups and check how the average predicted risk compares with the outcome prevalence in each group. The Sanders and Murphy statistics form the groups as the sets of observations with the same predicted values. When all the predicted values are unique (as might occur in a regression model with continuous predictors), the Sanders and Murphy decompositions degenerate so that the calibration component is simply the Brier score. Another type of calibration measure defines groups of similar, but not identical, predicted values. Patients are divided either into equal-sized groups that span the probability scale with unequally sized ranges of probabilities or into unequally sized groups that split the probability scale into equally sized increments. By describing the degree of accuracy over the entire range of probabilities, calibration statistics reflect how well an instrument predicts for patients with very different likelihoods of the outcome of interest.

The most common form of calibration statistic is based on the Pearson χ^2 statistic (*see* **Chi-square Tests**) used to summarize model fit by comparing observed and expected outcomes within K groupings defined by the predictors (*see* **Goodness of Fit**). It is written

$$\sum_{j=1}^{K} \frac{(O_j - E_j)^2}{n_j \hat{p}_j (1 - \hat{p}_j)},$$

where for n_j individuals in the jth group, O_j, is the observed number of events and $E_j = n_j \hat{p}_j$ is the expected number. Hosmer & Lemeshow [9] suggest that division using equal-sized groups is preferable especially when many of the predicted probabilities are small and that the test statistic be compared with a **chi-square distribution** having eight **degrees of freedom**.

Model-Based Validation of Calibration and Discrimination

Miller et al. [10] suggest a model for checking both calibration and discrimination using an idea of Cox [3] to examine the coefficients from a logistic regression of Y_i on the logit of $\hat{p}_i$ so that $\log[\Pr(Y_i = 1)/\Pr(Y_i = 0)] = \alpha + \beta \log[\hat{p}_i/(1 - \hat{p}_i)]$. In this model the intercept, α,

is a measure of calibration and the slope, β, is a measure of discrimination. Perfect calibration and discrimination correspond to a model with $\alpha = 0$ and $\beta = 1$. The predictive probability is too low if $\alpha > 0$ and too high if $\alpha < 0$. If $\beta > 1$, then the $\hat{p}_i$ show the correct direction but do not vary enough; if $0 < \beta < 1$, then the $\hat{p}_i$ vary too much; and if $\beta < 0$, then the $\hat{p}_i$ show the wrong general direction.

Three **likelihood ratio tests** can be constructed to test discrimination and calibration: (i) the test of $H_0 : \alpha = 0$ and $\beta = 1$ is a global test for calibration and discrimination; (ii) the test of $H_0 : \alpha = 0$ given $\beta = 1$ is a test for calibration given appropriate discrimination; and (iii) the test of $H_0 : \beta = 1$ given $\alpha = 0$ is a test of discrimination given appropriate calibration. From this model, Miller et al. [10] develop regression **diagnostics** for assessing **outliers**, influence, and leverage points as they affect calibration and discrimination.

Graphical Representations of Calibration and Discrimination

Many graphs have been developed to represent calibration and discrimination (*see* **Graphical Displays**). Yates [21] describes a covariance graph for representing the components of his decomposition of the Brier score plotting the predicted values on the vertical axis against the outcome values on the horizontal axis. Because there are only two possible outcomes, this plot collapses to a vertical dotplot for each outcome group. When the number of data points is large, it is better to represent the distributions of predicted values for each outcome state as histograms. The spread of these histograms describes the scatter component. A line is then drawn connecting the means of the two distributions. This is the slope component. The bias is represented by the distance between the 45° line indicating equality of predicted and observed, and the intersection of the horizontal line drawn across from the mean predicted value and the vertical line drawn up from the mean event rate. Arkes et al. [1] describe the use of this graph in evaluating the prognosis of patients in the SUPPORT study.

The ROC curve formed by plotting the true positive rate on the vertical axis against the false positive rate on the horizontal axis gives a good depiction of the sensitivity and specificity of the diagnostic tests set up by specifying cutpoints along the probability scale.

Calibration can be represented either by a barplot giving observed and average predicted probabilities for each grouping of the observations [17] or as a continuous curve across all possible predicted values. Harrell et al. [7] have suggested using a data smoother to describe the relationship between observed and predicted in such plots. Schmid et al. [16] use a second-degree loess smoother [2]. A **moving average** could also be used. Hilden et al. [8] plot the cumulative number of events against the expected number as the predicted values increase from 0 to 1. For each of these plots the departure of the continuous curve from the 45° line of equality describes the lack of calibration evidenced by the data.

Extensions to Multistate Outcomes

Many of the calibration and discrimination measures may be easily extended to outcomes for which more than two events are possible. The Brier score is written as $(1/N)\sum(\hat{p}_i - y_i)^{\mathrm{T}}(\hat{p}_i - y_i)$, where $\hat{p}_i - y_i$ is a column vector with each element corresponding to one of the possible events that could occur on the ith individual. For the two-event case, this value is actually twice the Brier score computed previously because the complementary event state is also included in the calculation. The Brier score decompositions are also straightforward upon noting that the squared terms in the two-state case can be replaced with vector extensions. The measure of bias can be extended by forming the column vector $\mathbf{Z} = \hat{\mathbf{p}} - \overline{\mathbf{y}}$ of mean predicted and observed in each of the k states. Then $\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}$ follows a χ^2 distribution with k - 1 degrees of freedom. The calibration χ^2 statistics can be extended in the same way. For more than two states, the c-index is defined as in the two-state case, since a pair of individuals can still have only two outcomes. But the equivalence with the ROC curve area is lost. Good references for these multivariate extensions can be found in Yates [22], Winkler [20], Harrell et al. [7], and Hilden et al. [8]. Habbema et al. [5] discuss the use of a simplex to present the distribution of predicted and observed values in the three-state case.

Conclusion

Many calibration and discrimination statistics are available to measure how well binary (and multistate)

outcomes are classified by models. The *c*-index and the Hosmer–Lemeshow χ^2 statistic are the most widely used, but the careful data analyst should assess model performance in many different ways, bearing in mind that test statistics may vary from dataset to dataset. Good performance is usually easier to achieve on data with a wide variation in the covariates. For example, if individuals are either very sick or very well, it will be easier to categorize them. Performance may also suffer when a model fit to one set of data is tested on a new set, especially if the new set displays less variation in the covariates than the old one. When prediction is the goal of the modeling process, in fact, model performance should never be reported solely from the data used to develop the model. Instead, it is wisest to choose the model that displays the best calibration and discrimination on the new data [7].

References

[1] Arkes, H.R., Dawson, J.V., Speroff, T., Harrell, F.E., Alzola, C., Phillips, R., Desbiens, N., Oye, R.K., Knaus, W., Connors, A.F. & the SUPPORT Investigators (1995). The covariance decomposition of the probability score and its use in evaluating prognostic estimates, *Medical Decision Making* **15**, 120–131.

[2] Cleveland, W.S., Grosse, E. & Shyu, W.M. (1992). Local regression models, in *Statistical models in S*, J.M. Chambers & T.J. Hastie, eds. Wadsworth, Pacific Grove, pp. 309–376.

[3] Cox, D.R. (1958). Two further applications of a model for binary regression, *Biometrika* **45**, 562–565.

[4] Diamond, G.A. (1992). What price perfection? Calibration and discrimination of clinical prediction models, *Journal of Clinical Epidemiology* **45**, 85–89.

[5] Habbema, J.D.F., Hilden, J. & Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis: I. The problem, descriptive tools, and measures based on classification matrices, *Methods of Information in Medicine* **17**, 217–226.

[6] Hanley, J.A. & McNeil, B.J. (1982). The measuring and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143**, 29–36.

[7] Harrell, F.E., Lee, K.L. & Mark, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine* **15**, 361–387.

[8] Hilden, J., Habbema, J.D.F. & Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis: II. Trustworthiness of the exact values of diagnostic probabilities, *Methods of Information in Medicine* **17**, 227–237.

[9] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[10] Miller, M.E., Hui, S.L. & Tierney, W.M. (1991). Validation techniques for logistic regression models, *Statistics in Medicine* **10**, 1213–1226.

[11] Miller, R.G. (1986). *Beyond ANOVA, Basics of Applied Statistics*. Wiley, New York.

[12] Murphy, A.H. (1973). A new vector partition of the probability score, *Journal of Applied Meterology* **12**, 595–600.

[13] Poses, R.M., Bekes, C., Winkler, R.L., Scott, W.E. & Copare, F.J. (1990). Are two (inexperienced) heads better than one (experienced) head?, *Archives of Internal Medicine* **150**, 1874–1878.

[14] Redelmeier, D.A., Bloch, D.A. & Hickam, D.H. (1991). Assessing predictive accuracy: how to compare Brier scores, *Journal of Clinical Epidemiology* **44**, 1141–1146.

[15] Sanders, F. (1963). On subjective probability forecasting, *Journal of Applied Meterology* **2**, 191–201.

[16] Schmid, C.H., D'Agostino, R.B., Griffith, J.L., Beshansky, J.R. & Selker, H.P. (1997). A logistic regression model when some events precede treatment: the effect of thrombolytic therapy for acute myocardial infarction on the risk of cardiac arrest, *Journal of Clinical Epidemiology*, **50**, 1219–1239.

[17] Selker, H.P., Griffith, J.L. & D'Agostino, R.B. (1991). A tool for judging coronary care unit admission appropriateness valid for both real-time and retrospective use: a time-insensitive predictive instrument TIPI for acute cardiac ischemia: a multicenter study, *Medical Care* **29**, 610–627.

[18] Shapiro, A.R. (1977). The evaluation of clinical predictions. a method and initial application. *New England Journal of Medicine* **295**, 1509–1514.

[19] Spiegelhalter, D.J. (1986). Probabilistic prediction in patient management and clinical trials, *Statistics in Medicine* **5**, 421–433.

[20] Winkler, R.L. & Murphy, A.H. (1968). "Good" probability assessors, *Journal of Applied Meterology* **7**, 751–758.

[21] Yates, J.F. (1982). External correspondence: decompositions of the mean probability score, *Organizational Behavior and Human Performance* **30**, 132–156.

[22] Yates, J.F. (1988). Analyzing the accuracy of probability judgments for multiple events: an extension of the covariance decomposition, *Organizational Behavior and Human Decision Processes* **41**, 281–299.

CHRISTOPHER H. SCHMID &
JOHN L. GRIFFITH

Multivariate Distributions, Overview

Multivariate distributions are defined on finite-dimensional spaces. They serve as probabilistic models for dependent outcomes of random experiments. Biometric data typically comprise observations on multiple characteristics for each experimental subject, and joint distributions are central to the modeling and analyses of such data. Multivariate distributions derive from other distributions through operations including **transformations**, projections, **conditioning**, convolutions, **extreme values**, mixing, compounding, **truncating**, and **censoring**. From them derive the distributions of various sample statistics of note in statistical **inference**. In addition, multivariate distributions characterize the behavior of **stochastic processes** through properties of their finite-dimensional projections. Occasionally, multivariate distribution theory supports probabilistic proofs for mathematical theorems. In short, multivariate distributions arise throughout statistics and applied probability, and their properties are essential to an understanding of those and related fields.

Basic Concepts

Origins and Uses

To fix ideas, suppose that a pharmacologist focuses primarily on the cardiovascular system and secondarily on neurology and musculature. Let $\mathbf{X} = [X_1, X_2, X_3, X_4, X_5]'$ represent observations on the systolic X_1, and diastolic, X_2, pressures, pulse rate, X_3, and gross, X_4, and fine, X_5, motor skills. Prospects for altering the cardiovascular state of a subject through medication may be negated in part by adverse side-effects on motor skills, for example. A complete probabilistic description of the system entails the joint distribution of the variables $[X_1, X_2, X_3, X_4, X_5]$.

The origins of this topic trace to studies beginning in the early nineteenth century on **multivariate normal distributions** and their applications [1, 2, 10, 11, 18, 19, 24, 43, 50–52, 55, 57], including early biometrical investigations. Systematic studies of such distributions in two and three dimensions are credited to Bravais [2] and Schols [52], and in any finite dimensions to Edgeworth [11], including such essential concepts as **regression** and partial **correlations**.

Multivariate distributions merit scrutiny at several levels of detail. At one extreme are basics such as the stochastic **convergence** of vector sequences and multidimensional Chebychev inequalities under weak moment assumptions, which are genuinely **nonparametric**. At the other extreme are rigidly parametric models with distributions having specified functional forms. A survey of the latter follows subsequently. In between are classes of distributions exhibiting common structural features such as symmetry or unimodality, giving rise to *semiparametric* models. For many purposes it suffices to know only the structural properties of distributions rather than their explicit functional forms. For example, various concepts of multivariate unimodality are developed in Dharmadhikari & Joag-Dev [8]; some of these enable a sharpening of selected multidimensional Chebychev bounds. Further examples are cited here with reference to distributions exhibiting suitable symmetries, i.e. invariance under specified transformation groups.

Notions of stochastic ordering likewise admit several realizations in higher dimensions. These in turn give rise to useful multivariate probability inequalities. A systematic account of the latter is found in Tong [59], including many results for distributions listed here. Results pertaining to multivariate normal and related distributions are found in Tong [60]; brief surveys appear in the articles on the **Multivariate Normal Distribution** and **Multivariate *t* Distribution**. In particular, the comparative concentration of probabilities, as a stochastic ordering, is an essential concept for distributions on $\mathbb{R}^k$. Following Sherman [54], the probability measure $\mu(\cdot)$ is said to be *more peaked about* $\mathbf{0} \in \mathbb{R}^k$ than $\nu(\cdot)$ if and only if $\mu(A) \geq \nu(A)$ for every set A in the class C_k comprising the compact convex subsets of $\mathbb{R}^k$ that are symmetric under reflection about $0 \in \mathbb{R}^k$, i.e. $\mathbf{x} \in A$ implies $-\mathbf{x} \in A$. We return to this ordering subsequently.

Multivariate distributions exhibit many properties. A property is said *to characterize* a distribution if the property is unique to that distribution. A standard reference here is Kagan et al. [39], with connections to some distributions surveyed in the present article.

The choice of model is often critical in practice. Sometimes there is a clear mandate; more often the choice must be guided by experience, conjecture, and empirical validation. To aid in this choice,

numerous discrete and continuous multivariate distributions are now known, and others may be expected to emerge in the future. Our purpose here is to set forth basic concepts, to survey the principal known multivariate distributions of discrete and continuous types, and to give some insight regarding their use. It will be seen that numerous multivariate distributions arise through mixtures and compounding, in keeping with complexities intrinsic to modern experiments in many fields of inquiry. Such experiments often may be modeled conditionally in a random environment, so that unconditional distributions emerge as mixtures. Details are noted subsequently. Standard references for mixture models, their structure, analysis, and applications, are Everitt & Hand [12], Titterington et al. [58], McLachlan & Basford [46], and Lindsay [44].

Several topics encountered here are covered in greater detail elsewhere in this encyclopedia, as noted. Excellent references on the coverage of this article, including overviews for both discrete and continuous multivariate distributions and myriad technical details, are Johnson & Kotz [37, 38] and Patil & Joshi [49].

Notation

To fix notation, $\mathbb{R}^k$ designates Euclidean k-space and $\mathbb{R}^k_+$ its positive orthant; $F_{n\times k}$ is the collection of real $(n \times k)$ matrices; S_k consists of real symmetric $(k \times k)$ matrices; and S_k^0 and S_k^+ comprise the positive semidefinite and the positive definite varieties in S_k. Special arrays include the $(k \times k)$ identity $\mathbf{I}_k$, the unit vector $\mathbf{1}_k = [1, \ldots, 1]' \in \mathbb{R}^k$, the diagonal matrix diag $(a_1, \ldots, a_k)$, and the Kronecker product $\mathbf{A} \times \mathbf{B} = [a_{ij}\mathbf{B}]$. Specifically, $\mathbf{a} \in R^k$ is a column vector and $\mathbf{a}' = [a_1, \ldots, a_k]$ its transpose. The transpose, inverse, trace, and determinant of $\mathbf{A} \in F_{k\times k}$ are denoted by $\mathbf{A}', \mathbf{A}^{-1}, \mathrm{tr}\,(\mathbf{A})$, and $|\mathbf{A}|$, respectively. If $\mathbf{y} \in \mathbb{R}^k$ is random, its vector of expected values and its dispersion (or **covariance**) **matrix** are designated by $\mathrm{E}(\mathbf{y}) \in \mathbb{R}^k$ and $\mathrm{var}(\mathbf{y}) \in S_k^+$ when defined. If $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]' \in F_{n\times k}$, then conventions for the corresponding arrays are $\mathrm{E}(Y) = [\mathrm{E}(Y_{ij})] \in F_{n\times k}$ and $\mathrm{var}(\mathbf{Y}) = \mathrm{var}(\mathbf{y}) \in S_{nk}^+$ such that $\mathbf{y} = [\mathbf{y}_1', \ldots, \mathbf{y}_n']' \in \mathbb{R}^{nk}$.

The notation $\mathcal{L}(\mathbf{X})$ designates the law of distribution of $\mathbf{X} \in \mathbb{R}^k$ or $\mathbf{X} \in F_{n\times k}$, as appropriate. Abbreviations for probability density, cumulative distribution, and characteristic functions are *pdf, cdf*, and *chf*, respectively, whereas *iid* refers to a sequence of independent, identically distributed random elements (*see* **Random Variable**).

The Basic Tools

Let X_0 be a set and (Ω, B, P) a probability space with Ω as an event set, B a field of subsets of Ω, and P a probability measure. An X_0-valued random element is a measurable mapping $\mathbf{X}(\omega)$ from Ω to X_0 which, when X_0 is finite-dimensional such as the Euclidean space $\mathbb{R}^k$, is multivariate. The cumulative distribution function (cdf) of $\mathbf{X}(\omega) = [X_1(\omega), \ldots, X_k(\omega)]'$ on $\mathbb{R}^k$ is given by

$$F(x_1, \ldots, x_k) = P(\omega : X_1(\omega) \le x_1, \ldots, X_k(\omega) \le x_k), \quad (1)$$

with values in the unit interval. Corresponding to each cdf is a probability measure P_x and conversely, giving the model $(\mathbb{R}^k, B_k, P_x)$ in which B_k is the Borel field of subsets of $\mathbb{R}^k$. This provides a formal framework for multivariate distribution theory, although the details are typically suppressed in the discussion that follows.

The study of multivariate distributions draws heavily on the calculus of $\mathbb{R}^k$, on integral transforms of Fourier, Laplace, and Mellin, including **characteristic functions** on $\mathbb{R}^k$ (see [45]), and on embedding other finite-dimensional spaces in $\mathbb{R}^k$. Distributions often emerge through a change of variables and integral transforms, and their properties through inverse images. **Generating functions** for joint moments, cumulants, factorial moments, and probabilities are used routinely (*see* **Moment Generating Function**). Projection methods apply, as the distribution of $\mathbf{X}(\omega)$ on $\mathbb{R}^k$ is determined completely by the one-dimensional distributions of every linear transformation of it.

Some multivariate distribution functions admit simple expressions in closed form. More commonly, many others require expansions in multiple series fraught with problems of convergence and stability. Limit theorems (*see* **Large-sample Theory**) often suggest simple approximations. Asymptotic expansions, as well as expansions of the Cornish–Fisher and Edgeworth types, are available for many otherwise intractable multivariate distributions. Clearly, a complete catalog of probability functions would be

desirable for the distributions surveyed here. Unfortunately, this is overly ambitious owing to the aforementioned difficulties. Instead, we list functional forms where tractable, and otherwise refer the reader to excellent monographs which do supply complete details as well as further references to the archival literature.

Types of Distributions

Discrete distributions arise with counting data such as pulse rates and numbers of adult and larval insects. Continuous distributions typically associate with measurements such as systolic and diastolic pressures. A formal statement follows.

Suppose that X_0 is finite-dimensional. Each probability measure P on $[X_0, B(X_0), \cdot]$ is decomposable as a mixture

$$P = a_1 P_1 + a_2 P_2 + a_3 P_3,$$
$$a_i \geq 0,\ a_1 + a_2 + a_3 = 1, \tag{2}$$

such that P_1 assigns positive probability to the mass points of P, P_2 has absolute continuity with respect to the Lebesgue (i.e. volume) measure on $(X_0, B(X_0), \cdot)$, and P_3 is purely singular on a set in X_0 having a Lebesgue measure zero, often a linear subspace of X_0.

Corresponding to P_1, P_2, and P_3 on (R^k, B_k, P) are cdfs F_1, F_2, and F_3, respectively, as in (1). Here $F_1(x_1, \ldots, x_k)$ has a probability mass function (pmf) as given by

$$p(x_1, \ldots, x_k) = P(X_1 = x_1, \ldots, X_k = x_k), \tag{3}$$

giving the jumps of $F_1(x_1, \ldots, x_k)$ at its mass points, whereas $F_2(x_1, \ldots, x_k)$ has a corresponding probability density function (pdf) given by

$$f_2(x_1, \ldots, x_k) = \frac{\partial^k}{\partial x_1, \ldots, \partial x_k} F_2(x_1, \ldots, x_k) \tag{4}$$

for almost all $\{x_1, \ldots, x_k\}$.

Partition $\mathbf{X} \in \mathbb{R}^k$ as $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2']'$ such that $\mathbf{X}_1 \in \mathbb{R}^r$ and $\mathbf{X}_2 \in \mathbb{R}^s$, with $r + s = k$. Then the marginal cdf of X_1 is given by $F_{m1}(x_1, \ldots, x_r) = F(x_1, \ldots, x_r, \infty, \ldots, \infty)$ for either discrete, absolutely continuous, or singular distributions. The pmf for the conditional distribution $\mathcal{L}(\mathbf{X}_1|\mathbf{x}_2)$ of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is given by

$$p_1(x_1, \ldots, x_r|x_{r+1}, \ldots, x_k) = \frac{p(x_1, \ldots, x_k)}{p_2(x_{r+1}, \ldots, x_k)}, \tag{5}$$

with $p_2(x_{r+1}, \ldots, x_k)$ as the marginal function for $\mathbf{X}_2 = [X_{r+1}, \ldots, X_k]'$. A similar expression holds in the absolutely continuous case in terms of the joint pdf $f(x_1, \ldots, x_k)$ and the marginal pdf $f_2(x_{r+1}, \ldots, x_k)$.

The study of multivariate distributions is concerned mainly with functions of the discrete (3) and continuous (4) types; the principal distributions of these types are surveyed in this article under sections labeled "Discrete Distributions" and "Continuous Distributions". In practice, P_3 typically is a degenerate distribution, often concentrated on a subspace of R^k and absolutely continuous there. These pure types may be combined by mixture as in (2).

Continuous Distributions

Scope

Some finite-dimensional distributions arise from multidimensional limit theorems; many others serve as models for outcomes of random experiments; and still others originate primarily as derived distributions, often pertaining to sample statistics. Prominent among limit distributions and those from which others derive are multivariate normal distributions. Their basic features, including marginal and conditional distributions, regression functions, selected probability inequalities, central limit theorems, and Berry–Esseen bounds on convergence rates to a multivariate normal limit, are summarized under **Multivariate Normal Distribution**. Many basic distributions derived from these are known to be identical, and thus invariant, for all parent distributions in a structured class containing symmetric multivariate stable laws and numerous others. An impressive list of normal-theory statistical procedures, both univariate and multivariate, are thus genuinely semiparametric, in the sense that it suffices to identify only the structural symmetry of underlying distributions rather than their explicit functional forms. These facts bear heavily on the **robustness** and validity of normal-theory procedures for use with nonnormal data, including linear statistical models, various

data-analytic multivariate procedures, and the **multivariate Bartlett test** for equal dispersion matrices, to cite telling examples. Symmetric multivariate distributions are surveyed next; special mappings of practical note that are known to induce invariant distributions are then characterized; and derived distributions having these properties are identified subsequently. We then undertake a systematic survey of the principal distributions of random vectors and matrices of continuous types. The principal reference for these is Johnson & Kotz [38]; vector and matrix distributions having structural symmetry are found in many sources, including Dempster [6], Kariya & Sinha [40], Fang & Anderson [13], Fang et al. [15], and Fang & Zhang [14], and other references to be cited.

Symmetric Distributions

Multivariate data often are scattered more heavily away from the center of location than are multivariate normal data. Indeed, many models for contaminated errors are so. Symmetric distributions that are either more or less scattered than multivariate normal laws are described here, where by "symmetry" is meant *invariance under a specified group of transformations* acting on the space of observations. In many applications these are seen to supply useful semiparametric models, *in lieu of* overly restrictive multivariate normal models, as being germane to the outcomes of random experiments.

Distributions on $\mathbb{R}^n$. Let $S_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = [S_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \psi); \psi \in \Psi]$ be the class of *elliptical distributions* on $\mathbb{R}^n$, having location-scale parameters $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and the typical pdf

$$f(\mathbf{y}) = |\boldsymbol{\Sigma}|^{-1/2}\psi[(\mathbf{y}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})], \quad (6)$$

where $\psi(\cdot) \in \Psi$, with Ψ as the class of all such functions on $[0, \infty)$. The class $S_n(\mathbf{0}, \mathbf{I}_n)$ contains *isotropic distributions* on $\mathbb{R}^k$ for which $\mathbf{z}$ and $\mathbf{Q}\,\mathbf{z}$ have the same distribution for every real orthogonal matrix $\mathbf{Q}$ $(n \times n)$. Examples of elliptical distributions on $\mathbb{R}^n$ are given in Table 1, where **Student's *t*, Cauchy**, and stable laws with $\alpha < 2$ have tails heavier than those of multivariate normal distributions. It is known that $\mathrm{E}(\mathbf{y}) = \boldsymbol{\theta}$ and $\mathrm{var}(\mathbf{y}) = \gamma\boldsymbol{\Sigma}$, with $\gamma > 0$, whenever these moments are defined. Further properties are given in the references cited, together with review articles by Devlin et al. [7] and Chmielewski [4]. An original source is the penetrating work of Cambanis et al. [3].

Distributions on $F_{n\times k}$. Let $S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma}) = [S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma}, \psi); \psi \in \Psi]$, with parameters $(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$ such that $\boldsymbol{\Gamma} \times \boldsymbol{\Sigma} = [\gamma_{ij}\boldsymbol{\Sigma}]$, be the class of distributions on $F_{n\times k}$ having the typical pdf

$$f(\mathbf{Y}) = |\boldsymbol{\Gamma}|^{-k/2}|\boldsymbol{\Sigma}|^{-n/2}\psi \times [\mathrm{tr}\,(\mathbf{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\mathbf{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1}], \quad (7)$$

where $\mathrm{tr}\,(\cdot)$ is the trace and $\psi(\cdot)$ belongs to the class Ψ consisting of all such functions on $[0, \infty)$. This class contains symmetric stable distributions including matrix normal distributions of the type $N_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$, as given in Table 2, having $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\Theta}$ and $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Gamma} \times \boldsymbol{\Sigma}$, as well as matrix versions of other examples from Table 1, and many others. Independence of the rows of $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]'$ and multivariate normality are linked: if $L(\mathbf{Y}) \in S_{n,k}(\boldsymbol{\Theta}, \mathbf{I}_n \times \boldsymbol{\Sigma})$, then $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ are mutually independent if and only if $\mathbf{Y}$ is matrix normal on $F_{n\times k}$; see James [26].

More generally, $L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ designates the class of matrix distributions on $F_{n\times k}$ having the typical pdf

$$f(\mathbf{Y}) = |\boldsymbol{\Gamma}|^{-k/2}|\boldsymbol{\Sigma}|^{-n/2}\phi(\mathbf{D}'\boldsymbol{\Gamma}^{-1}\mathbf{D}), \quad (8)$$

Table 1 Examples of spherical distributions on $\mathbb{R}^n$ having probability density functions $f(\mathbf{x})$ or characteristic functions $\xi(\mathbf{t})$

Type	Description	Comments
Multivariate normal	$f(\mathbf{x}) = c_1 \exp(-\mathbf{x}'\mathbf{x}/2)$	
Pearson type II	$f(\mathbf{x}) = c_2(1-\mathbf{x}'\mathbf{x})^{\gamma-1}$	$\gamma > 1$
Pearson type VII	$f(\mathbf{x}) = c_3(1+\mathbf{x}'\mathbf{x})^{-\gamma}$	$\gamma > n/2$
Student's t	$f(\mathbf{x}) = c_4(1+\nu^{-1}\mathbf{x}'\mathbf{x})^{-(\nu+n)/2}$	ν a positive integer
Cauchy	$f(\mathbf{x}) = c_5(1+\mathbf{x}'\mathbf{x})^{-(n+1)/2}$	
Scale mixtures	$f(\mathbf{x}) = c_6 \int_0^\infty t^{-n/2}\exp(-\mathbf{x}'\mathbf{x}/2t)\,\mathrm{d}G(t)$	$G(t)$ a cdf
Stable laws	$\xi(\mathbf{t}) = c_7 \exp[\gamma(\mathbf{t}'\mathbf{t})^{\alpha/2}]$	$0 < \alpha < 2$ the index

Table 2 Standard pdfs for some continuous matrix distributions

Type	Description[a]	Comments
Normal $N_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$	$k_1 \exp\{-[\mathrm{tr}\,(\mathbf{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\mathbf{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1}]\}$	$\mathbf{Y} \in F_{n\times k}$
Wishart $W_k(\nu, \boldsymbol{\Sigma})$	$k_2\lvert\mathbf{W}\rvert^{(\nu-k-1)/2}\exp(-\mathrm{tr}\,\mathbf{W}\boldsymbol{\Sigma}^{-1}/2)$	$\mathbf{W} \in S_k^+$
Matrix T	$k_3\lvert\mathbf{I}_k + \nu^{-1}\mathbf{T}'\mathbf{T}\rvert^{-(\nu+r)/2}$	$\mathbf{T} \in F_{r\times k}$

[a] $k_1 = [(2\pi)^{nk/2}\lvert\boldsymbol{\Gamma}\rvert^{k/2}\lvert\boldsymbol{\Sigma}\rvert^{n/2}]^{-1}$, $k_2 = \{(2)^{\nu k/2}\pi^{k(k-1)/4}\lvert\boldsymbol{\Sigma}\rvert^{\nu/2}\prod_{i=1}^{k}\Gamma[(\nu-i+1)/2]\}^{-1}$, and $k_3 = \{(\nu\pi)^{rk/2}\prod_{i=1}^{k}\Gamma[(\nu-i+1)/2]/\Gamma[(\nu+r-i+1)/2]\}^{-1}$.

with $\mathbf{D} = (\mathbf{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1/2}$, where $\phi(\cdot)$ is a function on S_k^+ and where $\boldsymbol{\Sigma}^{-1/2}$ is a factor of $\boldsymbol{\Sigma}^{-1}$. A subclass of these is $S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$. Distributions in $L_{n,k}(\boldsymbol{\Theta}, \mathbf{I}_n, \boldsymbol{\Sigma})$ have the property that $(\mathbf{Y}-\boldsymbol{\Theta})$ and $\mathbf{Q}(\mathbf{Y}-\boldsymbol{\Theta})$ have the same distribution for every real orthogonal matrix $\mathbf{Q}(n \times n)$. For a treatment of the class $L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ and its extensions, see Dempster [6], Dawid [5], Jensen & Good [34], and selected sections of monographs on symmetric distributions as cited earlier.

Invariance properties. Basic distributions induced from the foregoing classes through mappings of selected types are invariant. It remains to identify these. To be precise, let M be a subspace of $\mathbb{R}^n$ or $F_{n\times k}$, as appropriate; let T be a mapping to a finite-dimensional space V; and consider classes of parametric families to be generated from $S_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$, and $L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ as their parameters are varied. We are concerned with distributions for V-valued random elements induced through $T(\cdot)$. The following summary is taken from Jensen & Good [34] as a primary source, where proofs are provided. Applications appear subsequently for families exhibiting suitable symmetries. These facts in turn may be used to determine, essentially by inspection, the invariance properties of numerous derived distributions beyond those to be considered here.

Property 1. If $T[c(\mathbf{y}+\mathbf{m})] = T(\mathbf{y})$ for each $c > 0$ and $\mathbf{m} \in M \subset \mathbb{R}^n$, then the distribution of $T(\mathbf{y})$ is invariant for all distributions $\mathcal{L}(\mathbf{y})$ in the class $[S_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}); \boldsymbol{\theta} \in M, \boldsymbol{\Sigma} \in S_n^+]$ consisting of elliptical families on $\mathbb{R}^n$.

Property 2. If $T[c(\mathbf{Y}+\mathbf{M})] = T(\mathbf{Y})$ for each $c > 0$ and $\mathbf{M} \in M \subset F_{n\times k}$, then $\mathcal{L}[T(\mathbf{Y})]$ is invariant for all distributions $\mathcal{L}(\mathbf{Y})$ in the class $[S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma}); \boldsymbol{\Theta} \in M, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma} \in S_{nk}^+]$ consisting of elliptical families on $F_{n\times k}$.

Property 3. If $T[(\mathbf{Y}+\mathbf{M})\mathbf{B}] = T(\mathbf{Y})$ for each $\mathbf{M} \in M \subset F_{n\times k}$ and each nonsingular matrix $\mathbf{B}(k \times k)$, then $\mathcal{L}[T(\mathbf{Y})]$ is invariant for all distributions L $(\mathbf{Y})$ in the class $[L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}); \boldsymbol{\Theta} \in M, \boldsymbol{\Gamma} \in S_n^+, \boldsymbol{\Sigma} \in S_k^+]$ consisting of left-invariant distributions on $F_{n\times k}$.

Stochastic orderings. Stochastic order relations among distributions in the classes $S_n(\mathbf{0}, \boldsymbol{\Sigma})$ and $S_{n,k}(\mathbf{0}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$ are of note. To fix ideas, consider $S_n(\mathbf{0}, \boldsymbol{\Sigma}, \psi)$ and $S_n(\mathbf{0}, \boldsymbol{\Omega}, \psi)$; let $P_{\boldsymbol{\Sigma}}(\cdot; \psi)$ and $P_{\boldsymbol{\Omega}}(\cdot; \psi)$ be their corresponding probability measures; and recall the concentration ordering of Sherman [54] for sets in the class C_n comprising the symmetric convex subsets of R^n. Then a necessary and sufficient condition that $P_{\boldsymbol{\Sigma}}(\cdot; \psi)$ should be more concentrated about $\mathbf{0}$ than $P_{\boldsymbol{\Omega}}(\cdot; \psi)$, is that $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ should be ordered so that $\boldsymbol{\Omega} - \boldsymbol{\Sigma}$ is positive semidefinite. Sufficiency is shown in Fefferman et al. [16], and necessity in Jensen [33]. Further such orderings apply when both $(\boldsymbol{\Sigma}, \psi)$ are allowed to vary in $[S_n(\mathbf{0}, \boldsymbol{\Sigma}, \psi); \boldsymbol{\Sigma} \in S_n^+, \psi \in \Psi]$; for further details see Jensen [33]. These order relations in turn extend directly, without further difficulty, to include matrix distributions in the class $[S_{n,k}(\mathbf{0}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma}); \boldsymbol{\Gamma} \in S_n^+, \boldsymbol{\Sigma} \in S_k^+, \psi \in \Psi]$.

The principal multivariate continuous distributions are surveyed next by name, although terminology is not yet completely standardized. The multivariate normal members of $S_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$ are denoted respectively by $N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $N_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma} \times \boldsymbol{\Sigma})$ as before. The notation $\chi^2(\nu, \lambda)$ designates the noncentral **chi-square distribution** having ν degrees of freedom and noncentrality parameter λ, whereas the central case is abbreviated as $\chi^2(\nu)$.

Gamma Distributions

Gamma, chi-square, and **exponential** distributions on R_+^1 are well known. Matrix and vector versions of these are considered next.

Matrix distributions. Suppose that $\mathbf{\Sigma}(k \times k)$ is positive definite, $\mathbf{W}$ is random with values in S_k^+, and $K(\cdot)$ is a constant. The pdf with $\lambda > 0, \mathbf{W} \in S_k^+$, as given by

$$f(\mathbf{W}) = K(\lambda, \mathbf{\Sigma})|\mathbf{W}|^{\lambda-1} \exp(-\mathrm{tr}\mathbf{W}\mathbf{\Sigma}^{-1}), \qquad (9)$$

$f(\mathbf{W}) = 0$ otherwise, is that of a *matric gamma distribution* [45, p. 40 ff.] Here $K(\lambda, \mathbf{\Sigma}) = \{\pi^{k(k-1)/4} \prod_{r=0}^{k-1} \Gamma[(2\lambda + r)/2]|\mathbf{\Sigma}|^{\lambda+(k-1)/2}\}^{-1}$. If $\mathbf{W} = \mathbf{Y}'\mathbf{Y}$ with $L(\mathbf{Y}) \in L_{n,k}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma})$ as in (8) such that $n \geq k$, then the pdf of $\mathbf{W}$ is

$$f(\mathbf{W}) = K(n, k, \mathbf{\Sigma})|\mathbf{W}|^{(n-k-1)/2}\phi(\mathbf{\Sigma}^{-1/2}\mathbf{W}\mathbf{\Sigma}^{-1/2}) \qquad (10)$$

for $\mathbf{W} \in S_k^+$, $f(\mathbf{W}) = 0$ otherwise, a result of Hsu [25]. Here $K(n, k, \mathbf{\Sigma}) = \pi^{nk/2-k(k-1)/4}/|\mathbf{\Sigma}|^{n/2} \prod_{i=1}^{k} \Gamma[(n - i + 1)/2]$. Moreover, if $\mathcal{L}(\mathbf{Y}) = N_{n,k}(\mathbf{M}, \mathbf{I}_n \times \mathbf{\Sigma})$ with $n \geq k$, and if $\mathbf{W} = \mathbf{Y}'\mathbf{Y}$, then $\mathbf{W}$ has a *noncentral* **Wishart distribution**, denoted by $W_k(n, \mathbf{\Sigma}, \mathbf{\Lambda})$, with noncentrality matrix $\mathbf{\Lambda} = \mathbf{M}'\mathbf{M}$. The central version is denoted by $W_k(n, \mathbf{\Sigma})$; its pdf is a special case of (9) and (10) as given in Table 3; whereas the noncentral pdf has a series expansion in special polynomials (see [38, p. 170 ff.]).

Wishart matrices arise in multivariate normal sampling, e.g. as the scaled sample dispersion matrix, and otherwise in multivariate distribution theory. Parallel remarks apply to (10) and the class $L_{n,k}(\mathbf{M}, \mathbf{I}_n, \mathbf{\Sigma})$. The noncentral Wishart distribution, although rather intractable numerically, admits approximations based on the following. As $n \to \infty$, its limit distribution is multivariate normal for standardized central and noncentral matrices, and for fixed n it is asymptotically multivariate normal as the noncentrality parameters grow in a specified manner [30].

Distributions on $\mathbb{R}_+^k$. Joint distributions for the diagonal elements of $W = [W_{ij}]$ arise in the **analysis of variance** for nonorthogonal designs, in **time series analyses**, in **multiple comparisons**, in the analysis of multidimensional **contingency tables**, in extensions of Friedman's chi-square test in two-way data based on ranks (*see* **Mantel–Haenszel Methods**), and elsewhere in statistical methodology. There is a multivariate gamma distribution on $\mathbb{R}_+^k$ for diagonal elements corresponding to expression (9), a multivariate chi-square distribution when $\mathbf{W}$ is Wishart, and a multivariate exponential distribution in the central case of the latter for the case $n = 2$. The joint distribution of $[W_{11}^{1/2}, W_{22}^{1/2}, \ldots, W_{kk}^{1/2}]$, known as a multivariate Rayleigh distribution, arises in the detection of signals from noise [47]. More general Rayleigh distributions are known [28], as are more general multivariate chi-square distributions with differing marginal degrees of freedom. The latter arise as the joint distributions of traces of diagonal blocks of a block-partitioned Wishart matrix [29].

Densities for these distributions are rather intractable, apart from special cases, typically entailing multiple series expansions in special functions from applied mathematics. Details are given in Johnson & Kotz [38, Chapter 40]. However, as $n \to \infty$, the standardized chi-square and Rayleigh distributions in the limit are multivariate normal for both central and noncentral cases, and for fixed n, the limits again are multivariate normal as the noncentrality

Table 3 Standard pdfs for some continuous distribution on $\mathbb{R}^k$

Type	Description[a]	Comments
Student's t	$k_1[1 + \nu^{-1}(\mathbf{t} - \boldsymbol{\mu})'\mathbf{R}^{-1}(\mathbf{t} - \boldsymbol{\mu})]^{-(\nu+k)/2}$	$\mathbf{t} \in \mathbb{R}^k$
Dirichlet	$k_2(1 - \sum_{j=1}^k u_j)^{\alpha_0-1} \prod_{j=1}^k u_j^{\alpha_j-1}$	$\{0 \leq u_j \leq 1, \sum_{j=1}^k u_j \leq 1\}$
Inverted Dirichlet	$k_2 \prod_{j=1}^k \nu_j^{\alpha_j-1}/[1 + \sum_{j=1}^k \nu_j]^{\alpha/2}$	$\{0 \leq \nu_j < \infty, \alpha = \sum_{j=0}^k \alpha_j\}$
Roots of $\lvert\mathbf{W} - w\mathbf{\Sigma}\rvert$	$k_3 \prod_{i=1}^k w_i^{(\nu-k-1)/2} \prod_{i<j}(w_i - w_j) \exp[-(\sum_{i=1}^k w_i)/2]$	$\{w_1 > \cdots > w_k > 0\}$
Roots of $\lvert\mathbf{S}_1 - l\mathbf{S}_0\rvert$	$k_4 \prod_{i=1}^k l_i^{(m-k-1)/2} \prod_{i=1}^k (l_i + 1)^{-(m+n)/2} \prod_{i<j}(l_i - l_j)$	$\{l_1 > \cdots > l_k > 0\}$

[a] $k_1 = \Gamma[(\nu + k)/2]/(\pi\nu)^{k/2}\Gamma(\nu/2)|\mathbf{R}|^{1/2}$, $k_2 = \Gamma(\alpha)/\prod_{j=0}^k \Gamma(\alpha_j)$, $\alpha = \alpha_0 + \alpha_1 + \cdots + \alpha_k$,

$k_3 = \pi^{k/2}/2^{\nu k/2} \prod_{i=1}^k \{\Gamma[(\nu - i + 1)/2]\Gamma[(k - i + 1)/2]\}$, and

$k_4 = \pi^{k/2} \prod_{i=1}^k \Gamma[(m + n - i + 1)/2]/\{\prod_{i=1}^k \{\Gamma[(n - i + 1)/2]\Gamma[(m - i + 1)/2]\Gamma[(k - i + 1)/2]\}\}$.

parameters grow [27]. Alternate approximations are found through multivariate normalizing transformations of the Wilson–Hilferty type; see Jensen [31] and Jensen & Solomon [35] for further details.

Student Distributions

Vector and matrix versions of **Student's *t* statistic** are considered next. Further details are given in the article on the **Multivariate *t* Distribution**, based on normal sampling models. However, multivariate normality need not be assumed here, since central versions of these distributions are seen to be invariants under symmetry of the underlying parent distributions, as noted subsequently.

Distributions on $\mathbb{R}^k$. There are two basic types. Suppose that $[X_1, \ldots, X_k]$ is multivariate normal with means $[\mu_1, \ldots, \mu_k]$, unit variances, and correlation matrix $\mathbf{R}(k \times k)$. A *type I distribution* is that of $\{T_j = X_j/S,\ j = 1, \ldots, k\}$ such that the distribution of νS^2 is $\chi^2(\nu)$ independently of $[X_1, \ldots, X_k]$. The typical pdf for this type is listed in Table 3. To consider type II distributions, suppose that $\nu\mathbf{S} = \nu[S_{ij}]$ has the central Wishart distribution $W_k(\nu, \mathbf{R})$ independently of $[X_1, \ldots, X_k]$. A *type II distribution* on $\mathbb{R}^k$ is that of $\{T_j = X_j/S_{jj},\ j = 1, \ldots, k\}$. Both types are central whenever $\mu_1 = \cdots = \mu_k = 0$ and are noncentral otherwise. These distributions arise in multiple comparisons procedures, in the construction of rectangular confidence sets for means, in the Bayesian analysis of multivariate normal data (*see* **Multivariate Analysis, Bayesian**), and in various multistage procedures. Further details are given in Johnson & Kotz [38, Chapter 37] and Tong [60, Chapter 9].

More generally, if $\mathcal{L}(X_1, \ldots, X_k, Z_1, \ldots, Z_\nu)$ is in the class $S_n(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ with $\boldsymbol{\theta}' = [\mu_1, \ldots, \mu_k, 0, \ldots, 0]$ and $\boldsymbol{\Gamma} = diag(\mathbf{R}, \mathbf{I}_\nu)$, a block-diagonal matrix, then with $\nu S^2 = Z_1^2 + \cdots + Z_\nu^2$, the central joint distribution of $\{T_j = X_j/S;\ j = 1, \ldots, k\}$ is type I multivariate t for all distributions in $S_n(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ having the required structure. This follows from invariance Property 1, so that normal-theory multiple comparisons using $\{T_1, \ldots, T_k\}$ are exact in level for linear models having spherical errors [32]. Similarly, if $L(\mathbf{Y}) \in S_{n,k}(\boldsymbol{\Theta}, \mathbf{I}_n \times \boldsymbol{\Sigma})$ with parameters $\boldsymbol{\Theta} = [\theta, \ldots, \theta]', \boldsymbol{\theta} \in R^k$; if $X_j = n^{1/2}\overline{Y}_j$ with $\{\overline{Y}_j = (Y_{1j} + \cdots + Y_{nj})/n;\ j = 1, \ldots, k\}$, and if $\mathbf{S}$ is the sample dispersion matrix, then invariance Property 2 asserts that the central distribution of $\{T_j = X_j/S_{jj}^{1/2};\ j = 1, \ldots, k\}$ is type II multivariate t for every L $(\mathbf{Y})$ in $S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \boldsymbol{\Sigma})$. Noncentral distributions generally depend on the particular distribution in $S_{n,k}(\boldsymbol{\Theta}, \mathbf{I}_n \times \boldsymbol{\Sigma})$.

Matric t distributions. Let $\mathbf{Y}$ and $\mathbf{W}$ be independent with $\mathcal{L}(\mathbf{Y}) = N_{r,k}(\mathbf{0}, \mathbf{I}_r \times \boldsymbol{\Sigma})$ and $\mathcal{L}(\mathbf{W}) = W_k(\nu, \boldsymbol{\Sigma})$ such that $\nu \geq k$, and let $\mathbf{T} = \mathbf{Y}\mathbf{W}^{-1/2}$ using any factorization $\mathbf{U}'\mathbf{U}$ of $\mathbf{W}$ with $\mathbf{W}^{1/2} = \mathbf{U}$. Then $\mathbf{T}$ has a *matric t distribution with* pdf as listed in Table 2. Alternatively, consider $\mathbf{X} = [\mathbf{Y}', \mathbf{Z}']'$ with distribution in $S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \boldsymbol{\Sigma})$ such that $n = r + \nu$ and $\nu \geq k$, and again let $\mathbf{T} = \mathbf{Y}\mathbf{W}^{-1/2}$ with $\mathbf{W} = \mathbf{Z}'\mathbf{Z}$. These variables arise from distributions in $S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \boldsymbol{\Sigma})$ in the same manner as for the multivariate normal case. From invariance Property 2, $\mathbf{T}$ has a matric t distribution for every distribution $\mathcal{L}(\mathbf{Y})$ in $S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \boldsymbol{\Sigma})$. This invariance property of $\mathcal{L}(\mathbf{T})$ transfers directly to the scaled distribution $\mathcal{L}(\mathbf{ATB})$ considered by Dickey [9] with $\mathbf{A}$ and $\mathbf{B}$ nonsingular.

Beta and F Distributions

Let X and Y be independent random gamma variates having a common scale. Then $U = X/(X + Y)$ has a **beta distribution** and $V = X/Y$ has an inverted beta distribution, with the Snedecor–Fisher ***F* distribution** as a special case of the latter. This section treats vector and matrix extensions of these distributions.

Dirichlet distributions. Let $\{Z_0, Z_1, \ldots, Z_k\}$ be independent gamma variates having a common scale and the shape parameters $\{\alpha_0, \alpha_1, \ldots, \alpha_k\}$, and let $T = Z_0 + Z_1 + \cdots + Z_k$. Then the joint distribution of $\{U_j = Z_j/T;\ j = 1, \ldots, k\}$ is the *k-dimensional Dirichlet distribution* $D(\alpha_0, \alpha_1, \ldots, \alpha_k)$ with pdf as given in Table 3. An important special case is that $\{\alpha_j = \nu_j/2;\ j = 0, 1, \ldots, k\}$ with $\{\nu_0, \nu_1, \ldots, \nu_k\}$ as positive integers and with $\{Z_0, Z_1, \ldots, Z_k\}$ as independent chi-square variates. However, in this case neither independence nor chi-square distributions are required. For if $\mathbf{y} = [\mathbf{y}_0', \mathbf{y}_1', \ldots, \mathbf{y}_k']' \in R^n$ with $\{\mathbf{y}_j \in R^{\nu_j};\ j = 0, 1, \ldots, k\}$ and $n = \nu_0 + \nu_1 + \cdots + \nu_k$ such that $L(\mathbf{y}) \in S_n(\mathbf{0}, \mathbf{I}_n)$, then invariance Property 1 ensures that $\{U_j = \mathbf{y}_j'\mathbf{y}_j/T;\ j = 1, \ldots, k\}$, but now with $T = \mathbf{y}_0'\mathbf{y}_0 + \mathbf{y}_1'\mathbf{y}_1 + \cdots + \mathbf{y}_k'\mathbf{y}_k$, has the distribution $D(\nu_0/2, \nu_1/2, \ldots, \nu_k/2)$.

A matric Dirichlet distribution is known [48] for which the random matrices $\{\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_r\}$ in S_k^+ are independent Wishart matrices as given by $\{L(\mathbf{S}_j) =$

$W_k(\nu_j, \mathbf{\Sigma}); \nu_j \geq k,\ j = 0, 1, \ldots, r\}$. If

$$\mathbf{W}_j = \left(\sum_{j=0}^{r} \mathbf{S}_j\right)^{-1/2} \mathbf{S}_j \left(\sum_{j=0}^{r} \mathbf{S}_j\right)^{-1/2} \tag{11}$$

for $j = 1, 2, \ldots, r$, then for the lower triangular square root their joint pdf is

$$f(\mathbf{W}_1, \ldots, \mathbf{W}_r) = K(\boldsymbol{\nu}) \prod_{j=1}^{r} |\mathbf{W}_j|^{(\nu_j - k - 1)/2} \times |\mathbf{I}_k - \sum_{j=1}^{r} \mathbf{W}_j|^{(\nu_0 - k - 1)/2} \tag{12}$$

for $\mathbf{W}_j$ and $(\mathbf{I}_k - \sum_{j=1}^{r} \mathbf{W}_j)$ positive definite; $f(\mathbf{W}_1, \ldots, \mathbf{W}_r) = 0$, otherwise; see [38, p. 234]. Here with $\boldsymbol{\nu} = [\nu_0, \nu_1, \ldots, \nu_k]'$ and $\nu = \nu_0 + \nu_1 + \cdots + \nu_r$, $K(\boldsymbol{\nu}) = K(\nu_0, \nu_1, \ldots, \nu_k) = \prod_{i=1}^{k} \Gamma[(\nu - i + 1)/2] / \prod_{i=0}^{r} \prod_{j=1}^{k} \Gamma[(\nu_i - j + 1)/2]$. As before, neither independence nor Wishart distributions are required. For if $\mathbf{Y} = [\mathbf{Y}_0', \mathbf{Y}_1', \ldots, \mathbf{Y}_r'] \in F_{n \times k}$ with $n = \nu_0 + \nu_1 + \cdots + \nu_r$, such that $\nu_j \geq k$ and $\mathcal{L}(\mathbf{Y}) \in S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \mathbf{\Sigma})$, then invariance Property 2 ensures that the joint pdf of $\{\mathbf{W}_1, \ldots, \mathbf{W}_r\}$ as in (11), with $\{\mathbf{S}_j = \mathbf{Y}_j' \mathbf{Y}_j;\ j = 0, 1, \ldots, r\}$, is identical to (12) for every distribution $\mathcal{L}(\mathbf{Y})$ in $S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \mathbf{\Sigma})$.

Connections among these distributions follow. When $k = 1$, (12) is Dirichlet. The ratios of quadratic forms,

$$U_j(\mathbf{a}) = \frac{\mathbf{a}' \mathbf{S}_j \mathbf{a}}{\mathbf{a}' (\mathbf{S}_0 + \mathbf{S}_1 + \cdots + \mathbf{S}_r) \mathbf{a}}, \tag{13}$$

for $j = 1, \ldots, r$ and for fixed $\mathbf{a} \in R^k$, and the ratios of traces,

$$U_j = \frac{\operatorname{tr} \mathbf{S}_j}{\operatorname{tr} (\mathbf{S}_0 + \mathbf{S}_1 + \cdots + \mathbf{S}_r)}, \tag{14}$$

for $j = 1, 2, \ldots, r$, are both Dirichlet. The special case of (12), with $r = 1$, is sometimes called a *type I multivariate beta distribution*.

Inverted Dirichlet and F distributions. The inverted Dirichlet distribution is that of $\{V_j = Z_j / Z_0;\ j = 1, \ldots, r\}$ whenever $\{Z_0, Z_1, \ldots, Z_r\}$ are independent gamma variates having a common scale and the shape parameters $\{\alpha_0, \alpha_1, \ldots, \alpha_r\}$ (see [38, p. 238]). The typical pdf is listed in Table 3. The scaled variates given by $\{V_j = \nu_0 Z_j / \nu_j Z_0;\ j = 1, \ldots, r\}$ then have a *multivariate F distribution* whenever $\{\alpha_j = \nu_j / 2;\ j = 0, 1, \ldots, r\}$ with $\{\nu_0, \nu_1, \ldots, \nu_r\}$ as positive integers. This arises in the analysis of variance in conjunction with ratios of independent mean squares to a common denominator [17]. As before, neither independence nor multivariate normality is required; take $\{V_j = \nu_0 \mathbf{y}_j' \mathbf{y}_j / \nu_j \mathbf{y}_0' \mathbf{y}_0;\ j = 1, \ldots, k\}$ with $\mathcal{L}(\mathbf{y}) \in S_n(\mathbf{0}, \mathbf{I}_n)$ as stipulated for Dirichlet distributions.

An inverted matric Dirichlet distribution due to Olkin & Rubin [48] takes $\{\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_r\}$ as before and defines $\{\mathbf{V}_j = \mathbf{S}_0^{-1/2} \mathbf{S}_j \mathbf{S}_0^{-1/2};\ 1 \leq j \leq r\}$ using the symmetric root of $\mathbf{S}_0$. The pdf $f(\mathbf{V}_1, \ldots, \mathbf{V}_r)$ is known allowing $\mathbf{S}_0$ to be noncentral. For the central case the joint pdf is given by

$$f(\mathbf{V}_1, \ldots, \mathbf{V}_r) = K(\boldsymbol{\nu}) \prod_{j=1}^{r} |\mathbf{V}_j|^{(\nu_j - k - 1)/2} | \times \left| \mathbf{I}_k + \sum_{j=1}^{r} \mathbf{V}_j \right|^{(\nu - k - 1)/2} \tag{15}$$

with $\nu = \nu_0 + \nu_1 + \cdots + \nu_r$ and $K(\boldsymbol{\nu})$ as defined following expression (12). The special case with $r = 1$ is sometimes called a *type II multivariate beta distribution*. Neither independence nor the Wishart character is required in the central case. To see this, take $\{\mathbf{S}_j = \mathbf{Y}_j' \mathbf{Y}_j;\ j = 0, 1, \ldots, r\}$ as for matric Dirichlet distributions with $\mathbf{Y} = [\mathbf{Y}_0', \mathbf{Y}_1', \ldots, \mathbf{Y}_r']'$, and conclude that $f(\mathbf{V}_1, \ldots, \mathbf{V}_r)$, as given in (15), is invariant for every $\mathcal{L}(\mathbf{Y})$ in $S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \mathbf{\Sigma})$.

Some connections among the foregoing distributions follow. When $k = 1$, $f(V_1, \ldots, V_r)$ is the pdf of the inverted Dirichlet distribution. The collections of ratios $\{V_j(\mathbf{a}) = \mathbf{a}' \mathbf{S}_j \mathbf{a} / \mathbf{a}' \mathbf{S}_0 \mathbf{a};\ 1 \leq j \leq r\}$, for fixed $\mathbf{a} \in \mathbb{R}^k$, and $\{V_j = \operatorname{tr} \mathbf{S}_j / \operatorname{tr} \mathbf{S}_0;\ 1 \leq j \leq r\}$, both have inverted Dirichlet distributions.

Other distributions of these types are known. Multivariate F distributions having correlated numerators have been found as ratios of multivariate chi-square variates to a common denominator (see [38, p. 240 ff.]), with applications in linear inference.

Distributions of latent roots. Many problems in statistics and applied probability entail the latent roots (**eigenvalues**) of random matrices. These include various topics in multivariate analysis pertaining to reduction by invariance, and in the study of energy levels of physical systems. Suppose that $\mathcal{L}(\mathbf{W}) = W_k(\nu, \mathbf{\Sigma})$, and consider the joint distribution of the

ordered roots $\{w_1 > w_2 > \cdots > w_k\}$ of the determinantal equation $|\mathbf{W} - w\boldsymbol{\Sigma}| = 0$. These entities arise in tests for hypotheses about dispersion parameters, for example. Their joint pdf is listed in Table 3. Occasionally ratios of these roots are required (see [38, p. 205]), for example in simultaneous inferences for the dispersion parameters, in which case an invariance result holds for the joint distributions of all such ratios. For if $\mathbf{W} = \mathbf{Y}'\mathbf{Y}$, such that $\mathcal{L}(\mathbf{Y}) \in S_{n,k}(\mathbf{0}, \mathbf{I}_n \times \boldsymbol{\Sigma})$, then the joint distributions of ratios of the roots of $|\mathbf{W} - w\boldsymbol{\Sigma}| = 0$ are invariant for all such matrix distributions by invariance Property 2.

To continue, suppose that $\mathbf{S}_0$ and $\mathbf{S}_1$ are independent Wishart matrices having the distributions $W_k(\nu_0, \boldsymbol{\Sigma})$ and $W_k(\nu_1, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, respectively. Then central ($\boldsymbol{\Lambda} = \mathbf{0}$) and noncentral joint distributions of the roots of

$$|\mathbf{S}_1 - l\mathbf{S}_0| = 0 \tag{16}$$

are known (see [38, pp. 181–188]), as given in Table 3 for the central case. These are the latent roots of $\mathbf{W}_1$ at (11) when $r = 1$. A further invariance property holds for the central case. For if $\mathbf{Y} = [\mathbf{Y}_0', \mathbf{Y}_1']'$ with $n = \nu_0 + \nu_1$, $\mathbf{S}_0 = \mathbf{Y}_0'\mathbf{Y}_0$, and $\mathbf{S}_1 = \mathbf{Y}_1'\mathbf{Y}_1$, then by invariance Property 3, the latent root distribution is the same for all $\mathcal{L}(\mathbf{Y})$ in $L_{n,k}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma})$.

Other Distributions

Numerous other multivariate continuous distributions are known. Multivariate versions of *Burr distributions* arise through gamma mixtures of independent **Weibull distributions** [38, pp. 288–291]. Various *multivariate exponential distributions* are known; some properties and examples are found on specializing the **multivariate Weibull distributions** treated elsewhere in this encyclopedia. Various *multivariate stable distributions* are known, as are other types of symmetric distributions mentioned earlier. *Multivariate* **extreme-value distributions** are treated in Johnson & Kotz [38, pp. 249–260], with emphasis on the bivariate case. The *Beta-Stacy distribution* (see [38, pp. 273–284]) yields a *multivariate Weibull distribution* as a special case. *Multivariate* **Pareto distributions** (see [38, pp. 285–288]) have their origins in econometrics. The *multivariate logistic distribution* (refer to [38, pp. 291–294]) is used to model binary data in the analysis of quantal responses. Kibble [41] used properties of characteristic functions to derive a bivariate distribution having normal and gamma marginals.

Discrete Distributions

Many discrete distributions have multivariate extensions. These serve as building blocks for other distributions through *compounding*, in which distributions are assigned to some or all parameters of a family. Here the principal distributions are surveyed and connections among them noted. Generic names are used including "negative" and "inverse" types, in keeping with conventional usage for univariate cases.

Few discrete multivariate probability mass functions are known in closed form, notable exceptions being the multinomial and multivariate hypergeometric functions as given in many standard textbooks. More common are multiple series expansions in special functions, or cases where joint factorial moment, or joint probability generating, functions are available. To catalog these here would require much explanation of notation and concepts, and considerable overlap with excellent sources now available. Instead we undertake a careful description of each class of distributions and relations among them. The principal references are Johnson & Kotz [37, Chapter 11] and selections from Patil & Joshi [49], to be referenced by page numbers to aid the reader. Additional citations include the inequalities of Jogdeo & Patil [36] for a number of discrete multivariate distributions, and others to be noted on occasion.

Binomial Distributions

The number of successes in n independent Bernoulli trials, each having the probability π of success, has the **binomial distribution** $B(n, \pi)$. The number of trials to k successes has a **negative binomial distribution**. Some extensions follow.

Multivariate binomial distributions. The outcome of a random experiment is classified as having or not having each of s attributes $\{A_1, \ldots, A_s\}$. If $\{X_1, \ldots, X_s\}$ are the numbers having these attributes in n independent trials, then theirs is an s-dimensional binomial distribution with parameters

$$\pi_i = \Pr(A_i) \quad i = 1, \ldots, s,$$

$$\pi_{ij} = \Pr(A_i A_j), \quad i \neq j, i, j = 1, \ldots, s, \quad (17)$$

$$\ldots$$

$$\pi_{12\ldots s} = \Pr(A_1 A_2 \ldots A_s).$$

The marginal distribution of X_i is $B(n, \pi_i)$, all having the same index n, for $i = 1, \ldots, s$. Bivariate distributions having different indices are studied in Hamdan [20] and Hamdan & Jensen [22].

For sequences of identical experiments, the limiting standardized distribution is multivariate normal as $n \to \infty$. For nonidentical sequences such that $\pi_i \to 0$ as $n \to \infty, i = 1, \ldots, s$, the limit is a multivariate Poisson distribution under conditions given later. For further developments, see [49, p. 81].

Multivariate Pascal distributions. Independent trials of the preceding type are continued until exactly k trials exhibit none of the s attributes. The joint distribution of the numbers $\{Y_1, \ldots, Y_s\}$ of occurrences of $\{A_1, \ldots, A_s\}$ during these trials is an s-dimensional Pascal distribution (see [49, p. 83]).

Multivariate negative binomial distributions. Begin with an s-variate **Poisson distribution** with parameters $\mathbf{\Lambda}$ to be defined in expression (18). Next scale each parameter using a random gamma variate with parameters (α, k). The resulting mixture is an s-variate negative binomial distribution (see [49, p. 83]), its marginals negative binomial. It reduces to the multivariate Pascal distribution when k is an integer and to the negative multinomial distribution on mixing multiple Poisson distributions to be defined.

Multinomial Distributions

Let $\{A_0, A_1, \ldots, A_s\}$ be exclusive and exhaustive outcomes having probabilities $\{\pi_0, \pi_1, \ldots, \pi_s\}$ with $0 < \pi_i < 1$ and $\pi_0 + \pi_1 + \cdots + \pi_s = 1$. The numbers $\{X_1, \ldots, X_s\}$ of occurrences of $\{A_1, \ldots, A_s\}$ in n independent trials have the **multinomial distribution** with parameters $(n, \pi_1, \ldots, \pi_s)$.

Negative multinomial distributions. If independent trials are repeated until A_0 occurs exactly k times, then the numbers of occurrences of $\{A_1, \ldots, A_s\}$ during these trials have a negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$. This distribution arises through mixtures: first as a gamma mixture of multiple Poisson distributions as noted, and secondly as a negative binomial mixture on n of multinomials. As $k \to \infty$ and $\pi_i \to 0$ such that $\{k\pi_i \to \lambda_i, 0 < \lambda_1 < \infty, i = 1, \ldots, s\}$, the negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$ converges to the multiple Poisson distribution with parameters $(\lambda_1, \ldots, \lambda_s)$. Further properties are developed in standard references (see [37, p. 292] and [49, p. 70]).

Multivariate multinomial distributions. These are the joint distributions of marginal sums in multidimensional contingency tables. Classify an outcome according to each of k criteria having the exclusive and exhaustive classes $\{A_{i0}, A_{i1}, \ldots, A_{is_i}\}$ for $i = 1, \ldots, k$. If in n independent trials $\{X_{i1}, \ldots, X_{is_i}; i = 1, \ldots, k\}$ are the numbers occurring in $\{A_{i1}, \ldots A_{is_i}; i = 1, \ldots, k\}$, then their joint distribution is called a *multivariate (also multivector) multinomial distribution*, including the k-variate binomial distribution when $s_1 = s_2 = \cdots = s_k = 1$. Further developments are given in [37, p. 312] and [49, p. 86].

Multivariate negative multinomial distributions. Continue independent trials of the preceding type until exactly t trials are classified in all of $\{A_{10}, A_{20}, \ldots, A_{k0}\}$. The numbers occurring in $\{A_{i1}, \ldots, A_{is_i}; i = 1, \ldots, k\}$ during these trials have a *multivariate negative multinomial distribution*, reducing to the negative multinomial distribution when $k = 1$, and to the multivariate Pascal distribution when $s_1 = s_2 = \cdots = s_k = 1$. For further discussion see [37, p. 314].

Hypergeometric Distributions

A collection of N items consists of $s + 1$ types: N_0 of type A_0, N_1 of type $A_1, \ldots, N_s$ of type A_s, with $N = N_0 + N_1 \cdots + N_s$. Random samples are taken from this collection.

Multivariate hypergeometric distributions. In a random sample of n items drawn without replacement, the joint distribution of the numbers of items of types $\{A_1, \ldots, A_s\}$ is an s-dimensional **hypergeometric distribution** with parameters $(n, N, N_1, \ldots, N_s)$. With replacement, their distribution is multinomial with parameters $(n, N_1/N, \ldots, N_s/N)$. As $N \to \infty$ and $N_i \to \infty$ such that $N_i/N \to \pi_i$, with $0 < \pi_i < 1$ and $\pi_1 + \cdots + \pi_s < 1$, the hypergeometric converges to the multinomial distribution with parameters $(n, \pi_1, \ldots, \pi_s)$. If instead, $N \to \infty, N_i \to \infty$, and $n \to \infty$ such that $N_i/N \to 0$ and $nN_i/N \to \lambda_i$, with $\{0 < \lambda_i < \infty, i = 1, \ldots, s\}$, then the limit distribution is multiple Poisson with parameters $(\lambda_1, \ldots, \lambda_s)$. For further properties, see [37, p. 200] and [49, p. 76].

Multivariate inverse hypergeometric distributions. If successive items are drawn without replacement until

exactly k items of type A_0 are drawn, then the numbers of types $\{A_1, \ldots, A_s\}$ thus drawn have an *s-variate inverse hypergeometric distribution* with parameters $(k, N, N_1, \ldots, N_s)$. As $N \to \infty, N_i \to \infty$, such that $N_i/N \to \pi_i$ with $0 < \pi_i < 1$ and $\pi_1 + \cdots + \pi_s < 1$, this distribution converges to the s-variate negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$.

If, instead, $N \to \infty, N_i \to \infty$, and $k \to \infty$ such that $N_i/N \to 0$ and $kN_i/N \to \lambda_i$ with $\{0 < \lambda_i < \infty, i = 1, \ldots, s\}$, then the multivariate inverse hypergeometric converges to the multiple Poisson distribution with parameters $(\lambda_1, \ldots, \lambda_s)$ (see [49, p 76].

Multivariate negative hypergeometric distributions. Sampling proceeds in two stages. First m items are drawn without replacement, giving $(x_1, \ldots, x_s)$ items of types $\{A_1, \ldots, A_s\}$. Without replacing the first sample, n additional items are drawn without replacement at the second stage, giving $(Y_1, \ldots, Y_s)$ items of types $\{A_1, \ldots, A_s\}$. The conditional distribution of $(Y_1, \ldots, Y_s)$, given that $\{X_1 = x_1, \ldots, X_s = x_s\}$, is a multivariate negative hypergeometric distribution. It arises on compounding the multinomial distribution, with parameters $(n, \pi_1, \ldots, \pi_s)$, by assigning to $(\pi_1, \ldots, \pi_s)$ an s-dimensional Dirichlet distribution and then mixing. Under alternate conditions, this distribution converges either to the multinomial or to the product of negative binomial distributions. See [49, p. 77] for further details.

Poisson Distributions

Poisson distributions on R^1 admit the following extensions.

Multiple Poisson distributions. If $\{X_1, \ldots, X_s\}$ are independent Poisson random variables with parameters $\{\lambda_1, \ldots, \lambda_s\}$, then their joint distribution is a *multiple Poisson distribution* with parameters $(\lambda_1, \ldots, \lambda_s)$.

Multivariate Poisson distributions. Let $\{X_1, \ldots, X_s\}$ have the multivariate binomial distribution with parameters as in (17), and suppose that $n \to \infty, \pi_i \to 0, i = 1, \ldots, s$, such that

$$n\left\{\pi_i - \sum_j \pi_{ij} + \sum_{j<k} \pi_{ijk} - \cdots + (-1)^{s-1}\pi_{12\ldots s}\right\} \to \lambda_i,$$

$$n\left\{\pi_{ij} - \sum_k \pi_{ijk} + \sum_{k<1} \pi_{ijkl} - \cdots + (-1)^{s-2}\pi_{12\ldots s}\right\} \to \lambda_{ij}, \qquad (18)$$

$$\ldots$$

$$n\pi_{12\ldots s} \to \lambda_{12\ldots s}.$$

Then the limiting distribution of $\{X_1, \ldots, X_s\}$ is a *multivariate Poisson distribution* with parameters given by (18). This distribution also can be derived as the joint distribution of various partial sums of 2^{s-1} independent Poisson random variables with parameters as appropriate. See Johnson & Kotz [37, p. 297] and Patil & Joshi [49, p. 82] for additional references and further details.

Multivariate Series Distributions

Further classes of discrete multivariate distributions are identified by types of their pmfs.

Multivariate logarithmic series distributions. These distributions arise through truncation and limits. If $[X_1, \ldots, X_s]$ has the s-variate negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$, then the conditional distribution of $[X_1, \ldots, X_s]$, given that $[X_1, \ldots, X_s] \neq [0, \ldots, 0]$, converges to the *s-variate logarithmic series distribution* with parameters $(\theta_1, \ldots, \theta_s)$ as $k \to 0$, where $\{\theta_i = 1 - \pi_i; i = 1, \ldots, s\}$. See [49, p. 71] for details. A modified multivariate logarithmic series distribution arises as a mixture, on n, of the multinomial distribution with parameters $(n, \pi_1, \ldots, \pi_s)$, where the mixing distribution is a logarithmic series distribution (see [49, p. 73]).

Multivariate power series distributions. A class of distributions with parameters $(\theta_1, \ldots, \theta_s) \in \Theta$, derived from convergent power series, has pmfs of the form

$$p(x_1, \ldots, x_s) = \frac{a(x_1, \ldots, x_s)\theta_1^{x_1} \ldots \theta_s^{x_s}}{f(\theta_1, \ldots, \theta_s)}, \qquad (19)$$

for $\{x_i = 0, 1, 2, \ldots; i = 1, \ldots, s\}$, $p(x_1, \ldots, x_s) = 0$, otherwise. The class of such distributions, called *multivariate power series distributions*, contains the s-variate multinomial distribution with parameters $(n, \pi_1, \ldots, \pi_s)$; the s-variate logarithmic series distribution with parameters $(\theta_1, \ldots, \theta_s)$; the s-variate

Table 4 Some discrete multivariate compound distributions

Basic distribution	Mixing parameters	Mixing distribution	References	Comments
Bivariate binomial $(n, \pi_{01}, \pi_{10}, \pi_{11})$	n	Poisson	[23]	Gives bivariate Poisson distribution
Multinomial $(n, \pi_1, \ldots, \pi_s)$	$\pi_1, \ldots, \pi_s$	Dirichlet	[37, p. 309] [49, p. 69]	Gives s-variate negative hypergeometric distribution
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Logarithmic series	[49, p. 69]	Gives s-variate modified logarithmic series distribution
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Negative binomial	[49, p. 68]	Gives s-variate negative multinomial distribution
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Poisson	[49, p. 69]	Gives multiple Poisson distribution
Multiple Poisson $(u\lambda_1, \ldots, u\lambda_s)$	u	Gamma	[49, p. 70]	Gives s-variate negative multinomial distribution
Multiple Poisson $(\lambda_1, \ldots, \lambda_s)$	$\lambda_1, \ldots, \lambda_s$	Multinormal	[56]	Gives s-variate Poisson–normal distribution
Multiple Poisson $(\lambda, \ldots, \lambda)$ $\lambda = \alpha + (\beta - \alpha)u$	u	Rectangular on (0,1)	[49, p. 80]	Gives s-variate Poisson–rectangular distribution
Multivariate Poisson $(u\lambda_1, u\lambda_{12}, \ldots, u\lambda_{12\ldots s})$	u	Gamma	[49, p. 82]	Gives s-variate negative binomial distribution
Negative multinomial $(k, \pi_1, \ldots, \pi_s)$	$\pi_1, \ldots, \pi_s$	Dirichlet	[37, p. 311] [49, p. 80]	Gives s-variate negative multinomial-Dirichlet distribution
Convolution of multinomials $(\gamma_1, \ldots, \gamma_{2^k}, \theta_1, \ldots, \theta_s)$	$\gamma_1, \ldots, \gamma_{2^k}$	Multivariate hypergeometric	[42]	Gives the distribution of numbers judged defective of k types in lot inspection

negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$; and others. See Patil & Joshi [49, p. 74] for further properties.

A nonexhaustive sampling of other discrete multivariate distributions is given next.

Bivariate Borel–Tanner distributions. A typical Borel–Tanner distribution is the distribution of the number of customers served before a queue vanishes for the first time. If service in a single-server queue begins with r customers of type I and s of type II with different arrival rates and service needs for each type, then the joint distribution of the numbers served is the *bivariate Borel–Tanner distribution* studied by Shenton & Consul [53].

Compound multivariate distributions. In many applications compound distributions arise from experiments undertaken in random environments, where the mixing distribution describes the variation of parameters of a specified model over the possible environments. Numerous bivariate and multivariate discrete distributions have been obtained through compounding, many motivated by the structure of the problem at hand. Some examples are listed in Table 4, together with references and brief comments. Series expansions, using suitable sets of orthogonal polynomials, are given in [20]– [23] for several discrete bivariate distributions.

References

An asterisk denotes a reference of historical interest.

[1] *Adrain, R. (1808). Research concerning the probabilities of the errors which happen in making observations, etc., *The Analyst; or Mathematical Museum* **1**, 93–109.

[2] *Bravais, A. (1846). Analyse mathématique sur les probabilites des erreurs de situation d'un point, *Mémoires Présentés par Divers Savants a l'Académie Royale des Sciences de l'Institut de France, Paris* **9**, 255–332.

[3] Cambanis, S., Huang, S. & Simons, G. (1981). On the theory of elliptically contoured distributions, *Journal of Multivariate Analysis* **11**, 368–385.

[4] Chmielewski, M.A. (1981). Elliptically symmetric distributions: a review and bibliography, *International Statistical Review* **49**, 67–74. (Excellent survey article on elliptical distributions).

[5] Dawid, A.P. (1977). Spherical matrix distributions and a multivariate model, *Journal of the Royal Statistical Society, Series B* **39**, 254–261. (Technical source paper on the structure of distributions).

[6] Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis*. Addison-Wesley, London. (General reference featuring a geometric approach).

[7] Devlin, S.J., Gnanadesikan, R. & Kettenring, J.R. (1976). Some multivariate applications of elliptical distributions, in *Essays in Probability and Statistics*, S. Ikeda, T. Hayakawa, H. Hudimoto, M. Okamoto, M. Siotani & S. Yamamoto, eds. Shinko Tsusho, Tokyo, pp. 365–394. (Excellent survey article on ellipsoidal distributions).

[8] Dharmadhikari, S. & Joag-Dev, K. (1988). *Unimodality, Convexity, and Applications*. Academic Press, New York.

[9] Dickey, J.M. (1967). Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution, *Annals of Mathematical Statistics* **38**, 511–518. (Source paper on matric t distributions and their applications).

[10] *Dickson, I.D.H. (1886). Appendix to "Family likeness in stature", by F. Galton, *Proceedings of the Royal Society of London* **40**, 63–73.

[11] *Edgeworth, F.Y. (1892). Correlated averages, *Philosophical Magazine, Series 5* **34**, 190–204.

[12] Everitt, B.S. & Hand, D.J. (1981). *Finite Mixture Distributions*. Chapman & Hall, New York.

[13] Fang, K.T. & Anderson, T.W., eds (1990). *Statistical Inference in Elliptically Contoured and Related Distributions*. Allerton Press, New York.

[14] Fang, K.T. & Zhang, Y.T. (1990). *Generalized Multivariate Analysis*. Springer-Verlag, New York.

[15] Fang, K.T., Kotz, S. & Ng, K.W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London.

[16] Fefferman, C., Jodeit, M. & Perlman, M.D. (1972). A spherical surface measure inequality for convex sets, *Proceedings of the American Mathematical Society* **33**, 114–119.

[17] Finney, D.J. (1941). The joint distribution of variance ratios based on a common error mean square, *Annals of Eugenics* **11**, 136–140. (Source paper on dependent F ratios in the analysis of variance).

[18] *Galton, F. (1889). *Natural Inheritance*. Macmillan, London, pp. 134–145.

[19] *Gauss, C.F. (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Muster-Schmidt, Göttingen.

[20] Hamdan, M.A. (1972). Canonical expansion of the bivariate binomial distribution with unequal marginal indices, *International Statistical Review* **40**, 277–280. (Source paper on bivariate binomial distributions).

[21] Hamdan, M.A. & Al-Bayyati, H.A. (1971). Canonical expansion of the compound correlated bivariate Poisson distribution, *Journal of the American Statistical Association* **66**, 390–393. (Source paper on a compound bivariate Poisson distribution).

[22] Hamdan, M.A. & Jensen, D.R. (1976). A bivariate binomial distribution and some applications, *Australian Journal of Statistics* **18**, 163–169. (Source paper on bivariate binomial distributions).

[23] Hamdan, M.A. & Tsokos, C.P. (1971). A model for physical and biological problems: The bivariate compound Poisson distribution. *International Statistical Review* **39**, 60–63. (Source paper on bivariate compound Poisson distributions).

[24] Helmert, F.R. (1868). Studien über rationelle Vermessungen, im Gebeite der höheren Geodäsie, *Zeitschrift für Mathematik und Physik* **13**, 73–129.

[25] Hsu, P.L. (1940). An algebraic derivation of the distribution of rectangular coordinates, *Proceedings of the Edinburgh Mathematical Society, Series 2* **6**, 185–189. (Source paper on generalizations of Wishart's distribution).

[26] James, A.T. (1954). Normal multivariate analysis and the orthogonal group, *Annals of Mathematical Statistics* **25**, 40–75.

[27] Jensen, D.R. (1969). Limit properties of noncentral multivariate Rayleigh and chi-square distributions, *SIAM Journal on Applied Mathematics* **17**, 807–814. (Source paper on limits of certain noncentral distributions).

[28] Jensen, D.R. (1970). A generalization of the multivariate Rayleigh distribution, *Sankhya, Series A* **32**, 192–208. (Source paper on generalizations of Rayleigh distributions).

[29] Jensen, D.R. (1970). The joint distribution of traces of Wishart matrices and some applications, *Annals of Mathematical Statistics* **41**, 133–145. (Source paper on multivariate chi-square and F distributions).

[30] Jensen, D.R. (1972). The limiting form of the noncentral Wishart distribution, *Australian Journal of Statistics* **14**, 10–16. (Source paper on limits of noncentral Wishart distributions).

[31] Jensen, D.R. (1976). Gaussian approximation to bivariate Rayleigh distributions, *Journal of Statistical Computation and Simulation* **4**, 259–268. (Source paper on normalizing bivariate transformations).

[32] Jensen, D.R. (1979). Linear models without moments, *Biometrika* **66**, 611–617. (Source paper on linear models under symmetric errors).

[33] Jensen, D.R. (1984). Ordering ellipsoidal measures: scale and peakedness orderings, *SIAM Journal on Applied Mathematics* **44**, 1226–1231.

[34] Jensen, D.R. & Good, I.J. (1981). Invariant distributions associated with matrix laws under structural symmetry, *Journal of the Royal Statistical Society, Series B* **43**, 327–332. (Source paper on invariance of derived distributions under symmetry).

[35] Jensen, D.R. & Solomon, H. (1994). Approximations to joint distributions of definite quadratic forms, *Journal of the American Statistical Association* **89**, 480–486.

[36] Jogdeo, K. & Patil, G.P. (1975). Probability inequalities for certain multivariate discrete distributions, *Sankhya, Series B* **37**, 158–164. (Source paper on probability inequalities for discrete multivariate distributions).

[37] Johnson, N.L. & Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Wiley, New York. (An excellent primary source with extensive bibliography).

[38] Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York. (An excellent primary source with extensive bibliography).

[39] Kagan, A.M., Linnik, Y.V. & Rao, C.R. (1973). *Characterization Problems in Mathematical Statistics*. Wiley, New York.

[40] Kariya, T. & Sinha, B.K. (1989). *Robustness of Statistical Tests*. Academic Press, New York.

[41] Kibble, W.F. (1941). A two-variate gamma type distribution, *Sankhya* **5**, 137–150. (Source paper on expansions of bivariate distributions).

[42] Kotz, S. & Johnson, N.L. (1983). Some distributions arising from faulty inspection with multitype defectives, and an application to grading, *Communications in Statistics – Theory and Methods* **12**, 2809–2821.

[43] *Laplace, P.S. (1811). Mémoire sur les integrales definies et leur application aux probabilités, *Mémoires de la classes des Sciences Mathématiques et Physiques l'Institut Impérial de France* **Année 1810**, 279–347.

[44] Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol. 5. Institute of Mathematical Statistics, Hayward.

[45] Lukacs, E. & Laha, R.G. (1964). *Applications of Characteristic Functions*. Hafner, New York. (Excellent reference with emphasis on multivariate distributions).

[46] McLachlan, G.J. & Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

[47] Miller, K.S. (1975). *Multivariate Distributions*. Huntington, New York. (An excellent reference with emphasis on problems in engineering and communications theory).

[48] Olkin, I. & Rubin, H. (1964). Multivariate beta distributions and independence properties of the Wishart distribution, *Annals of Mathematical Statistics* **35**, 261–269. Correction, *Annals of Mathematical Statistics* **37** (1966) 297. (Source paper on matric Dirichlet, beta, inverted beta, and related distributions).

[49] Patil, G.P. & Joshi, S.W. (1968). *A Dictionary and Bibliography of Discrete Distributions*. Hafner, New York. (An excellent primary source with extensive bibliography).

[50] *Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society of London, Series A* **187**, 253–318.

[51] *Plana, G.A.A. (1813). Mémoire sur divers problémes de probabilitié, *Mémoires de l'Académie Impériale de Turin* **20**, 355–408.

[52] *Schols, C.M. (1875). Over de theorie der fouten in de ruimte en in het platte vlak, *Verhandelingen der Koninklijke Akademie van Wetenschappen (Amsterdam)* **15**, 1–75.

[53] Shenton, L.R. & Consul, P.C. (1973). On bivariate Lagrange and Borel-Tanner distributions and their use in queueing theory, *Sankhya, Series A* **35**, 229–236. (Source paper on bivariate Lagrange and Borel-Tanner distributions and their applications).

[54] Sherman, S. (1955). A theorem on convex sets with applications, *Annals of Mathematical Statistics* **25**, 763–766.

[55] *Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology* **15**, 72–101.

[56] Steyn, H.S. (1976). On the multivariate Poisson normal distribution, *Journal of the American Statistical Association* **71**, 233–236. (Source paper on multivariate Poisson-normal distributions).

[57] *Student (1908). The probable error of a mean, *Biometrika* **6**, 1–25.

[58] Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

[59] Tong, Y.L. (1980). *Probability Inequalities in Multivariate Distributions*. Academic Press, New York.

[60] Tong, Y.L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

(*See also* **Multivariate Analysis, Overview**)

D.R. JENSEN

Multivariate Graphics

The domain of multivariate graphics could span about everything envisioned by mankind. Images are inherently multivariate. A satellite image can be treated as a single observation with multispectral and spatial structure. Today's computing power allows processing and manipulation of such observations. Many of today's visualization targets were unthinkable a few decades ago. For example, the human genome with over two billion base pairs is a prime visualization target. While today's popular graphics challenge is to develop visualization methods for massive data sets, the old prime objective remains: to communicate to ourselves and others.

This brief article cannot begin to cover the domain of multivariate graphics. Rather, it tours a simple trail covering a small fraction of the landscape. The visited landscape primarily concerns static graphics. The tour provides little discussion of animation and direct manipulation methods. Even for static graphics the tour calls out a few highlights and provides pointers to other highlights that, unfortunately, are left to the imagination.

The chosen trail emphasizes places that are best known by the guide, who provides a personal view about key multivariate encoding issues that extend to places not visited. A few examples are drawn for other applications, but the methods have ready application to biostatistics. Along the tour, the discussion touches on important graphic design activities with words like *focus, simplify, link*, and *enhance*. These activities are universally important.

A few pointers to the literature may help the reader explore several facets of multivariate graphics. MacEachren [57] provides a readily accessible primer on symbolization and design. The classic work covering a wide variety of visual symbols and signs is Bertin [9]. Grinstein & Levkowitz [49] cover perceptual issues in visualization. Kosslyn [55] provides a gentle introduction to application of human perception and cognition in graph design. MacEachren [58] gives an extended treatment that is a valuable resource for more advanced students.

Foley et al. [43] provide an extensive overview of computer graphics methods. The methods are most immediately relevant to low-dimensional visualization. Wegman & Carr [76] cover selected computer graphic methods and address issues in perception and connections to statistical graphics.

Gnanadesikan [47] covers many of the basics in multivariate statistics, and numerous texts have followed. The multivariate analysis literature deals with important methods such as **clustering**, **classification**, **factor analysis**, **discriminant analysis**, and dimension reduction (*see* **Battery Reduction**) that are not described here. (For a discussion of these, *see* **Classification, Overview**; **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**; **Rotation of Axes**).

Early work in multivariate statistical graphics provides a continuing source of ideas. Fienberg [41] provides an early review. Barnett [6] contains a stimulating collection of papers. The work of John Tukey (see [32]) had a profound influence on statistical graphics and is a third resource worth revisiting.

Cleveland's recent books [30, 31] capture much of his long efforts to guide scientists toward superior statistical graphics methods. Cleveland & McGill [33] provide an early survey on dynamic multivariate graphics that foreshadows the visualization revolution in computer science. Tufte [67–69] has done much to expand interest in statistical graphics and to draw attention to works of elegance and beauty that appear on the printed page (*see* **Graphical Displays**).

Many additional resources are available. The computing revolution has increased access to and usage of visualization methodology by all disciplines. Professional societies churn out videos, proceedings on CDs, and collections of papers. Wood [81] reminds us that different maps have a different agenda (*see* **Statistical Map**). Similarly, papers on graphics have different agendas. While important work appears in medical imaging and other areas (*see* **Image Analysis and Tomography**), the entertainment industry now drives much work in visualization. Establishing fruitful connections between visualization techniques and scientific applications can involve significant additional research. This tour is more about graphical methods closely connected to probabilistic inference (*see* **Inference**; **Inference, Foundations of**) than about data and model-free visualization methods.

The Goal of Multivariate Graphics: Apparently Simple Comparisons

As an overview statement at the beginning of the tour, the dominant goal of multivariate graphics is comparison. Comparisons come in three forms: (i) comparison of external images with each other; (ii) comparisons of external images with external references; and (iii) comparison of external images with the analyst's internal references. These internal references include scientific knowledge and statistical expectations or models. The multivariate graphics design goal is to facilitate meaningful comparisons. This includes converting internal references into external visual references subject to further manipulation. With external images and references available, the next step often involves transformation to simpler forms. Typically the goal is to produce simple comparison graphics that involve juxtaposition, superimposition, or the direct display of differences.

A major goal in statistical graphics design is to reduce the cognitive effort required to make comparisons. In terms of statistical graphics design, Kosslyn [55] warns "the spirit is willing, but the mind is weak". Graphic design must do everything it can to help people to understand.

One important effort reduction strategy is to re-express the standard for comparison in simpler form. For example, in a simple **regression** involving one independent variable, one can compare observed (x, y) pairs to a predicted curve $y = f(x)$. However, to assess **residuals**, humans should not have to assess differences from a changing reference line $y = f(x)$. It is better to plot the residuals directly. The reference line for residuals is a horizontal straight line, and that is as simple as a visual reference can be. As a second example the line $x = y$ is the common reference line in $Q-Q$ plots (*see* **Normal Scores**). Tukey's mean–difference plot (see [30]) transforms the reference line into a horizontal straight line. More generally, Tukey [70] says "less than fully adjusted variates should not be plotted". While the statement's context is mapping of mortality rates for exploratory analysis (*see* **Mapping Disease Patterns**), the strategy of simplifying the human cognitive tasks by removing known structure applies to all facets of multivariate graphics.

In multivariate graphics we seek apparently simple views of comparisons. The path to these apparently simple views can involve deep insights about the phenomena involved, sound statistical summaries, and careful attention to issues of human perception and cognition.

Communication Objectives

Multivariate graphics can have many different communication objectives. Communication objectives influence graph design choices. Four common objectives are to provide an overview, to tell a story, to suggest hypotheses, and to criticize a model. In providing an overview, coverage is important. Hiding details is often crucial to achieve clarity in the coverage shown. Similarly, in telling a story the predetermined message must shine through. Tufte [69] is an important resource on the topic of visual explanations. Newspapers sometimes have good graphics and these can provide valuable lessons in communication. Scientists often fail to tell simple stories because they are reluctant to suppress caveats and a host of details that qualify the basic results. Interactive network software (see [25]) alleviates the problem by showing the basic graphics and by giving access to metadata that provide the basis for appropriate interpretation.

This article emphasizes discovery objectives that include suggesting hypotheses and criticizing models. For discovery, balanced visual emphasis of the variables helps the data to speak. Once known effects have been removed, display methods that are asymmetric in visual weighting of variables tend to be poor in terms of discovery objectives. Of course a fortuitous emphasis of some variables over others occasionally leads to insight, but even then the careful analyst will move toward the symmetric position of trying all permutations of the variables.

The same method that is poor for discovery may be used to advantage in telling a story. As an example, Chernoff faces [28] provide a very asymmetric representation and are poor at conveying multivariate interpoint distances. Nonetheless, people have used them to communicate effectively by, for example, equating high salaries to smiles.

Basic Design Considerations

With Kosslyn's warning in mind, we come to multivariate graphics prepared to do battle. Our willing spirit uses the best tools in the arsenal to discover the important patterns. As Cleveland [30] says, "tools

matter". The tools include encoding variables with high perceptual accuracy of extraction, using easily discriminated symbols, reducing memory and calculation burdens, grouping information into more manageable units, and layering the information so it can be dealt with in stages. The discussion below introduces some of this in the context of univariate and bivariate graphics.

Univariate Guidelines

Cleveland & McGill [34] discuss the perceptual accuracy of extraction and indicate preferred methods for univariate comparisons. Their research had subjects judge relative magnitudes of graphically encoded univariate variables. Their results ranked the graphic encoding methods into three classes, described here as best, good, and poor.

The two best encoding methods represent variables using position along a common scale, as shown in Figure 1, and position along identical nonaligned scales. That humans do well in judging the position of a point relative to a scale should come as no surprise. Marr [59] notes the "quintessential fact of human vision – that it tells about shape and space and spatial arrangement". Locating the position of objects is a fundamental visual task. Map makers have long used position along a scale as the fundamental encoding for spatial coordinates. MacEachren's [58] review of the perception literature attests to the power and primacy of positional encoding.

Length, angle, and orientation are good encodings. Figure 2 shows that transforming line segments into a standard position converts the task of judging length into a task of judging the position of one endpoint against a scale. While this is not necessarily what people do, the example suggests that judging line length is more complicated than judging position.

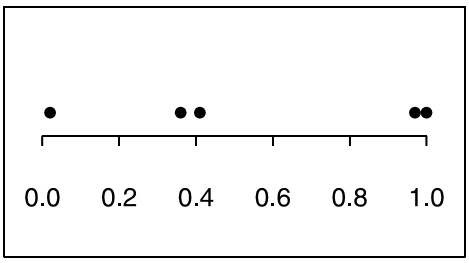

Figure 1 The best continuous univariate encoding: position along a scale

Figure 3 shows angle encoding. Rotation of the angles puts them in a position for comparison against equivalent angular scales, shown in gray. The transformation suggests that while angle comparisons work pretty well, they are more complicated than direct comparison against angle scales.

Area, volume, point density, and color saturation are poor encodings. The reader familiar with the experimental results involving Steven's Law will not be surprised by the poor results for the area and volume encodings. Steven's Law states that the perceived magnitude of a stimulus follows a power law,

$$p(x) = ax^b,$$

where x is the magnitude of the true stimulus (i.e. length, area, volume), and where the constants a and b depend on the type of stimulus. Table 1, adapted from Baird & Noma [5], provides the range of the characteristic exponents b for length, area, and volume. That is, people's perception of length tends to be directly proportional to object length. However,

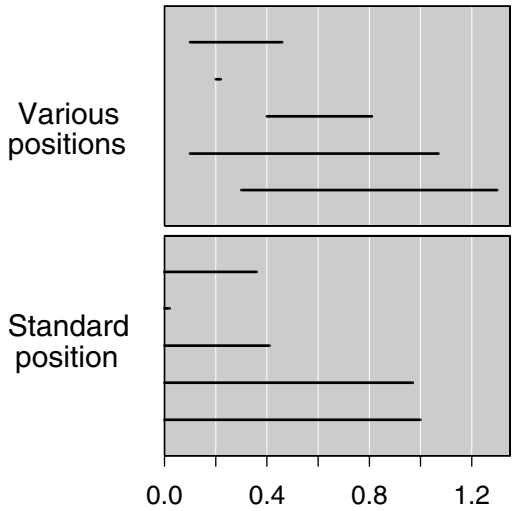

Figure 2 A good continuous univariate encoding: line length

Table 1 Exponents for Steven's Law

Encoding	Exponent range (b)
Length	(0.9, 1.1)
Area	(0.6, 0.9)
Volume	(0.5, 0.8)

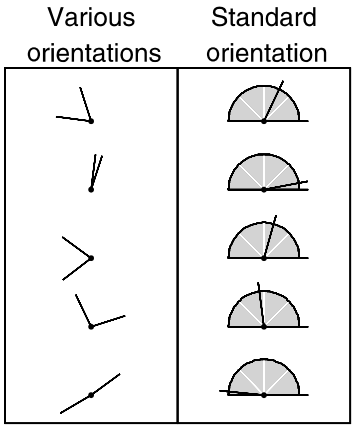

Figure 3 A good continuous univariate encoding: angle

we tend to judge area and volume nonlinearly. Consider comparing areas, one of 4 square units and the other of 1 square unit. With an exponent of 0.75, the ratio of perceived magnitudes is not 4 to 1, but 2.8 to 1. One way of describing our comparisons is that we underjudge the large areas relative to small areas. If everyone had the same exponent, then graphic encoding could adjust for systematic human bias. However, the range of values for b in Table 1 indicates substantial variability from person to person. Providing a set of reference symbols in a legend helps people calibrate to the intended interpretation, but the best strategy is to use better encodings when possible.

Weber's law is a fundamental law in human perception that has extensive ramifications concerning encoding information for accurate human decoding. A simple example gives the basic notion of the law. The probability of detecting that a 1.05 in. line is longer than a 1 in. line is about the same as the probability of detecting that a 1.05 ft. line is longer than a 1 ft. line. In absolute terms 0.05 in. is much smaller than 0.05 ft. The use of a finer resolution scale allows more accurate judgments on an absolute scale. In static graphics Cleveland [30] uses visual grid lines to provide a finer resolution scale for comparison purposes. In interactive graphics, zooming in provides a finer scale. Computer–human interface implementations that narrow focus to provide more accurate judgements include sliders (e.g. Ahlberg & Schneiderman [1] and Eick et al. [40]) and lenses (Rao & Card [63]).

Bivariate Guidelines

Tufte [67] notes that it took over 5000 years to generalize from early clay tablet maps to representing general variables using a scatterplot. Now the scatterplot is the standard for representing continuous bivariate data. The two orthogonal axes allow two coordinates to be independently encoded as a position along a common scale. In the statistical context, the bivariate visualization tasks are often to observe a functional relationship or to assess point density. These are not the most natural visual tasks. Enhancement methods reduce the amount of human visual processing required and help different humans to see the same summary.

Functional Relationships and Smoothing

When y is considered a function of x, common practice is to enhance scatterplots of (x, y) pairs by adding a smooth curve. To avoid the considerable human variability in sketching an eyeballed fit, the standard procedure is to fit the data using a computational procedure that others can replicate. Figure 4 shows a scatterplot with a smooth line generated using loess (see Cleveland et al. [35] for more details). Loess smoothes the data using weighted local regression. That is, the regression uses data local to x_0 to predict a value at x_0. Points closest to x_0 receive the greatest weight. Each smoother has many prediction points and so involves many local regressions. Each regression in the smoother shown

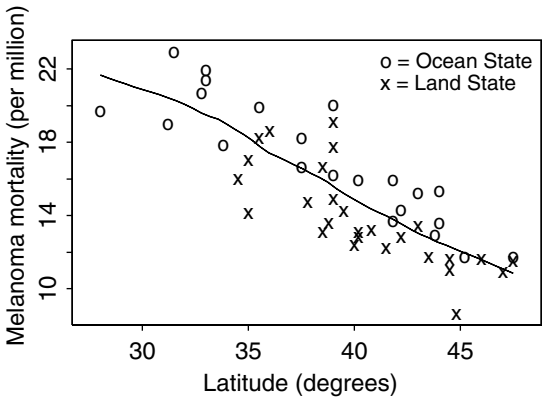

Figure 4 A smoother for scatterplots. An explicit smoother suggests the same functional relationship to different people. State mortality rates, 1950–1959

in Figure 4 used a linear model in x and included the closest 60% of the observations to the prediction point x_0. Those with the data (see Fisher & van Belle [42]) and the algorithm can reproduce the smoother. The smoother in Figure 4 draws further attention to the distinction between ocean and land states and additional modeling is appropriate. A first step might be to smooth the ocean and land states separately.

Smoothing is an extremely important enhancement technique (*see* **Nonparametric Regression**). The decomposition of data into smooth and residual parts is fundamental in statistical modeling. Hastie & Tibshirani [50] provide a good introduction to smoothing methods. Their description includes **generalized additive models** that cover the case of multiple independent variables.

Numerous smoothers are available. Historically, many researchers used cubic **splines** as smoothers. Cubic splines have a continuous second derivative, and that is sufficient to make curves appear smooth to humans. The elegant mathematical formulation behind splines *increased their popularity in segments of* the statistical community. However, there is no a priori best smoother. New methods, such as **wavelet** smoothing [11], keep appearing in statistical software. Different smoothers have different merits. Recently developed wavelets smoothers are better than many smoothers (but not necessarily all smoothers) at tracking discontinuities in the functional form.

Smoothers typically have some form of smoothing parameter that needs to be estimated or specified by the user. With computational power at hand, **cross-validation** methods have become increasingly popular as a community standard. This reduces the judgment burdens on the analyst, but of course does not guarantee a match between an empirical curve and a hypothesized true but unknown underlying curve. Hastie & Tibshirani [50] discuss cross-validation for moderate-sized applications. Golub & von Matt [48] discuss generalized cross-validation for large-scale problems.

Data Density and Density Estimates

When one looks at observations in scatterplots, one often looks for density patterns such as clusters, gaps, and outliers. However, humans are not good at visually assessing point density, so we again turn to a replicable computational method to provide an enhancement. Scott [64] provides a good introduction to **density estimation**. Like smoothing, density estimation has associated smoothing parameters and cross-validation methods to help in their selection. Scott [64] provides a discussion of cross-validation in the density estimation context.

A common task in statistics is to compare univariate distributions. Figure 5 shows a set of boxplots (*see* **Graphical Displays**). The boxplots show a caricature of the distribution. The features shown include the **median**, quartiles (*see* **Quantiles**), adjacent values, and **outliers**. Variations (see Frigge et al. [45]) may show extrema rather than adjacent values and outliners. The variation in Figure 5 uses a white line (see [15]) to provide intervals for comparing medians. If two **confidence intervals** do not overlap, then the medians are significantly different.

$Q-Q$ plots provide the preferred graphic to make detailed continuous distribution comparisons [30]. Computing a set of probability–quantile (p, q) pairs for each distribution lies behind the construction $Q-Q$ plots. For theoretical distributions, the cumulative distribution function, $F(\cdot)$, provides the correspondence between the pairs via $p = F(q)$. In simple cases the quantile function, $Q(\cdot)$, is the inverse of $F(\cdot)$ and $Q(p) = q$. Familiar p, q pairs from the standard normal distribution are (0.5, 0) and (0.975, 1.96). Order statistics approximations provide pairs for sampled data. A common approximation is $((i - 0.5)/n, x_{(i)})$, where $x_{(i)}$ is the ith order statistic and n is the sample size. Linear interpolation approximates values between the n pairs. Comparison of two distributions, denoted 1 and 2, proceeds by plotting quantile pairs $(Q_1(p), Q_2(p))$ over a range of probabilities. Figure 6 shows a $Q-Q$ plot for two batches of data. The x-axis shows quantiles from batch 1 and the y-axis shows quantiles from batch 2.

A strong merit of $Q-Q$ plots is that in simple cases they have a nice interpretation. If points fall on a straight line, then the distributions have the same shape (basically, the same moments higher than two) and the distributions differ only in the first two moments. This is the case in Figure 6, since the robust fit thin line matches the quantiles quite well. The thick line is the reference line for identical distributions. The slope of the thin line indicates the ratio of the scale estimates (for example, standard deviations). The lines are not quite parallel in Figure 6. Graphical fitting can proceed by

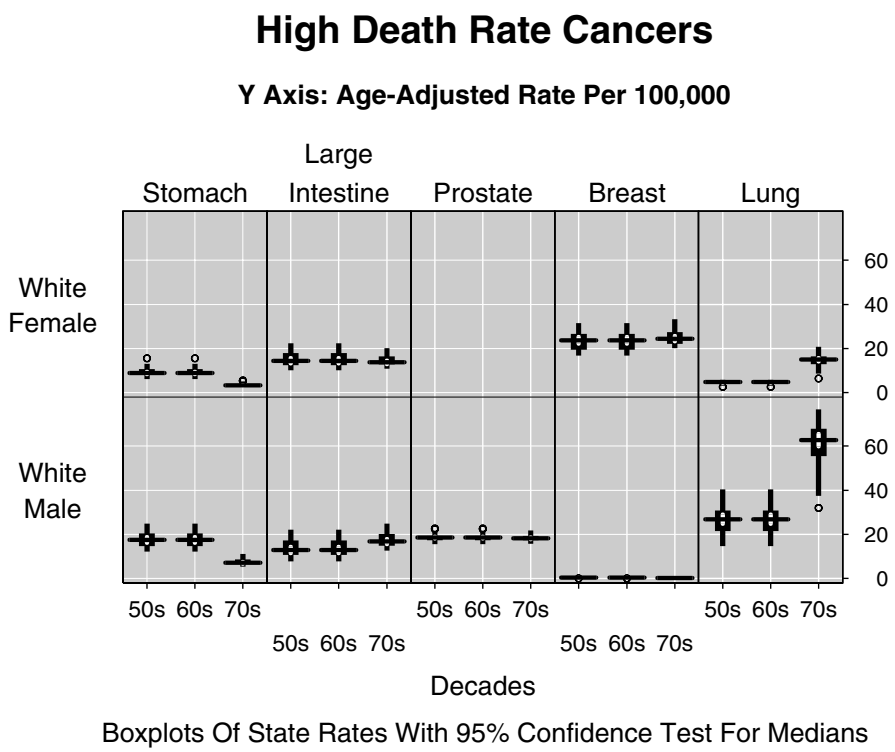

Figure 5 A variation on boxplots. The median: a long horizontal line; first and third quartiles: ends of thicker boxes; adjacent values: ends of thinner boxes; outliers: open circles; test intervals for different medians: white lines inside boxes

guessing at the ratio and multiplying this times the y-axis quantiles until the lines are parallel. When the lines are parallel, the vertical distance between the two lines gives the difference in location (or means) between the scale-adjusted distributions. In Figure 6 the lines are nearly parallel so a reasonable guess is that the distributions differ in location by about 0.5. As indicated earlier, Tukey used a mean and difference calculation to rotate the points in the plot. This simplifies the identical distribution reference line to a horizontal line with a zero intercept.

$Q-Q$ plots get around the deceptive procedure of superimposing two distribution functions or two survival curves (*see* **Survival Analysis, Overview**). As Figure 7 suggests, we are really poor at judging the distance between curves. Our visual systems naturally assess the closest differences between curves rather than the correct vertical distances (see [31]). Adding grid lines can help, but it is often better to

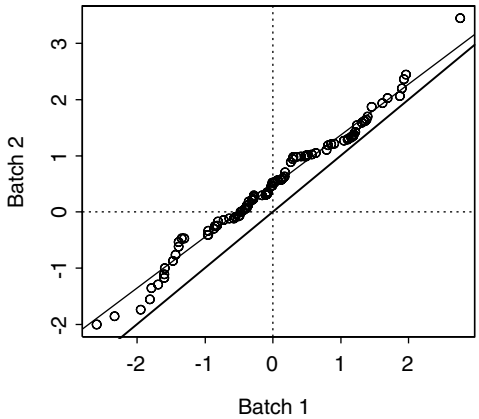

Figure 6 A two-sample $Q-Q$ plot. A good straight line fit suggests similar distributional shapes. Given similar shapes, the slope shows the ratio of scale parameters, such as standard deviations. Given a slope of one, the intercept shows the difference of location parameters, such as means. Thin line: robust fit; thick line: same distribution line

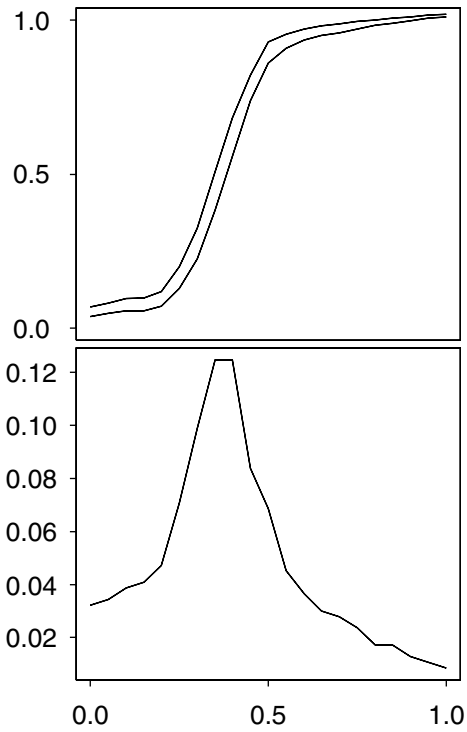

Figure 7 Explicit difference of two curves. Humans tend to see the closest differences between curves, not differences in the y direction

plot the difference explicitly or make comparisons using $Q-Q$ plots.

Jones & Cook [53] have recently generalized $Q-Q$ plots to higher dimensions and this is worth considering. Currently, analysts are more accustomed to looking at densities than cumulative distributions in higher dimensions.

Estimating point density adds another variable. For two variables, the estimated density is a third variable and can be represented as the surface $z = f(x, y)$. Many surface representations are available, such as color draped perspective wireframes and highlighted rendered polygons. For example, see Cleveland [30]. Wegman & Luo [77] note that specular reflection highlights call attention to local density anomalies. Intersecting surfaces with translucent planes help to focus attention on surface cross-sections. In fact, Cleveland et al. [35] recommend studying a sequence of two-dimensional (2D) cross-section plots to increase understanding of surface trends.

Tufte [68] notes that the pairing of contour and surface plots can aid understanding. Iso-density contour plots derive from cross-sections at fixed densities. Given a density, z_0, a contour line consists of pairs (x, y) that satisfy the equation $z_0 = f(x, y)$. A typical contour plot shows approximate contour lines for several values of z. Labeled contour lines do not have much visual impact. Several methods can provide visual impact. An easy approach is to communicate contour values by contour line thickness. Another option is to fill the regions between contour lines with color. The colors should be ordered.

A brief digression to discuss color is appropriate because color often comes into play in density representation and other facets of multivariate representation. Much literature is available on color. Good starting points are Brewer [10] and Levkowitz [56]. Humans are very sensitive to a dark-to-light scale that is referred to in the literature with terms like value, lightness, or brightness. This is an ordered scale and very important in visual interpretation. Friedhoff & Benzon [44] describe three visual processing channels, especially a high resolution dark-to-light channel. Humans get their shape information and many depth cues (linear perspective, interposition, shadow, and detail perspective) through this dark-to-light channel. Tufte [69] and others warn that when rainbow colors represent an ordered variable, lightness jumps and inconsistencies create unintended edges and patterns that can be confusing. Familiar color orderings, such as the rainbow ordering, have merit simply because they are familiar. However, the cognitive advantages of familiarity need to be balanced against the other facets of cognitive processing.

The literature also describes two other color dimensions: saturation and hue. A saturation scale goes from an achromatic color, such as medium gray, to a saturated color, such as vivid red. This scale is also ordered, but allows fewer distinctions than the dark-to-light scale. The hue dimension can be thought of as a circle that includes points between the colors of red, yellow, green, cyan, blue, magenta, and red. Hue is not an ordered scale and is good for distinguishing six or fewer categorical variables.

Returning to density representation, note that since contour plots reside in the dimension of the data, they require one fewer dimension than density surface plots. For trivariate data the estimated density

constitutes a fourth variable. Scott [64] provides density representations using sectioned, nested three-dimensional (3D) contour shells. He shows three contours distinguished by different hues. Sectioning reduces the shell to lines or ribbons and allows the viewer to see through the shell surfaces. Alternatively, surface translucence enables the viewer to see more than one surface simultaneously. In other work using four-dimensional data, Scott conditioned on the fourth variable to produce an animated sequence of 3D contour shell views. Directly showing the density surface for four-dimensional data would require a five-dimensional display.

As we push toward higher dimensions, enhancement methods remain important. Options include representing points, smoother, densities, residuals, and selected features.

Geometric Interpretation, Distance Judgments, Overplotting, and Scaling

In the multivariate context, accurate interpoint distance judgments are crucial to geometrically based visual interpretation. In dimensions above one, interpoint distance judgments are no longer equivalent to judging two values along a single scale and subtracting. For example, two points in a bivariate scatterplot are rarely parallel to an axis. The cognitive task becomes one of judging the length of an implicit line between the two points. As indicated above, humans perceive length in a plane with good perceptual accuracy of extraction. For higher-dimensional representations, assessing interpoint distances becomes increasingly difficult. The position here is that multivariate encodings for continuous variables should be ranked based on the ability of humans to judge interpoint distances.

At first, a stereo 3D scatterplot might seem ideal for representing continuous 3D data. Judging distance between points equates to judging the length of a line segment. On closer inspection, there is substantial change from two to three dimensions. Depth perception for the third coordinate derives from horizontal binocular discrepancies (parallax). The horizontal discrepancies involve only a small fraction of our full horizontal field of view. Horizontal visual acuity within this small window determines how many distinct depth planes we can resolve. This cannot match the horizontal position distinctions we make across the full field of view. Stereo 3D plots are less than ideal since humans do not judge depth as accurately as they judge vertical and horizontal position.

While stereo 3D distance judgments are not as accurate as 2D distance judgments, we live in a 3D world and have strong intuitions about 3D relationships. There are numerous cues that help us to assess depth in the real world. In the description of Friedhoff & Benzon [44], humans have three different visual processing channels: a dark-to-light high-resolution monocular shape channel described earlier; a binocular and motion parallax channel; and a low resolution color channel. Depth cues from the monocular shape channel, such as shadow, add to our depth perception. Since our depth judgment is calibrated by many depth cues and much experience, stereo 3D plots are only mildly asymmetric in the variables and this plot remains a strong candidate as the best representation for three continuous variables.

Historically, most analysts created depth views via rotation (motion parallax). This approach is very powerful. The main drawbacks are that moving points are harder to study than stationary points and that interacting with moving data is awkward. Using both stereo and rotation maintains the depth when rotations stops. The different viewing angles provided through rotation can be informative. Mostly one sees the edges of the data cloud. Different views reveal different edges. Slicing (discussed later as sectioning) helps to reveal the inside of the cloud. In the virtual reality (VR) settings one can fly through data clouds and touch points (or density features) to gain additional detail.

The first criterion for ranking multivariate encodings is the ability to convey interpoint distances. The second criterion for assessing multivariate encodings is the ability to represent many observations quickly without severe breakdown or distortion due to overplotting. Multivariate data can embody complex relationships that translate into complex geometric structure. Representing complex structure with data can require large samples just to cover the applicable domain. Seeing the structure through the noise that results from measurement error and changing measurement conditions can require much larger samples. In low dimensions a common approach uses density representations to reduce overplotting problems and to focus on the density structure (see Carr [12], Carr et al. [1], and Scott [64]).

As we move beyond five dimensions or so, our ability to judge interpoint distances deteriorates so badly that we are forced to use clustering algorithms and other computational methods to bring out patterns. Simple graphical methods do not work well except when there is a very simple geometric structure embedded in high-dimensional space. For complex multivariate structure we can lower the viewing dimensionality by sectioning (adding linear constraints – see Furnas & Buja [46]), conditioning on **categorical** variables, or by focusing attention on computed features such as local modes. This divide-and-conquer approach can reveal a plethora of patterns, but it is only the rare individual that can integrate a catalog of low-dimensional views into a coherent high-dimensional framework. Few of us claim to understand the full richness of a four-dimensional structure.

The ability to see structure depends on many factors that include scaling of data prior to graphical representation. Analysts often scale coordinates individually when they are measured in different units and jointly when measured in identical units. Transformations, such as using a **power transformation** to bring in the long tail of positive data, are common. Scaling often standardizes variables to mean zero and standard deviation one. Other options include replacing observation by ranks or normal scores. Joint scaling may spherize the data to remove **correlation** structure. The host of options includes methods for imputing values for **missing data**. Typically, graphical encoding involves scaling coordinates into the interval [0, 1] somewhere along the way. For the discussion of graphical representations below, assume that the coordinates have been scaled into [0, 1] or some other suitable interval such as $[-\pi/2, \pi/2]$.

Multivariate Representation Methods

While researchers continue to develop new multivariate representations, the representations tend to fall into a few classes. The classes include glyph plots, linked plots, nested plots, conditioned plots, geometric section plots, series plots, and composite plots. Some of the classes break into subclasses.

Glyph Plots

Glyph plots are symbol plots in which data values control the symbol parameters. For example, a circle is a glyph when one coordinate of a multivariate observation controls the circle size.

Multivariate glyph encodings typically fall into classes, namely those that use the spatial position to represent at least two of the multivariate coordinates and those that do not. A stereo scatterplot falls in the first class by representing three coordinates using the spatial position. Of the various stereo projections described in Carr [13], the infinity-enlarged (nonperspective) stereo projection has a particularly simple description as a glyph plot. The stereo projection uses two multivariate coordinates to determine the glyph position, and the glyph consists of left-eye and right-eye dots with horizontal separation (parallax) determined by the third coordinate. After appropriate routing to the eyes, the eye–brain system fuses the two dots into one dot in three dimensions.

Figure 8 is a monochrome side-by-side stereo plot adapted from the color version in Carr et al. [24]. Many can learn to fuse such images without the aid of the viewer. In a VR environment, shuttering glasses route separate full screen width images to

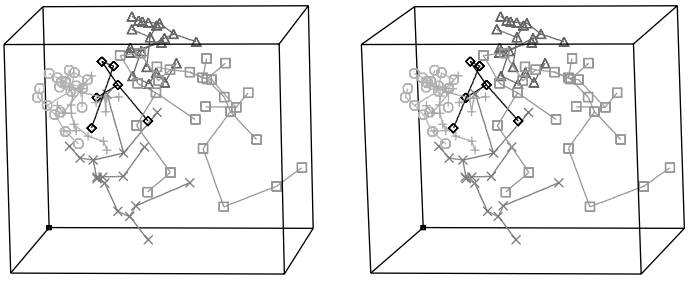

Figure 8 A side-by-side stereo pairs plot. Many can fuse the image without the aid of a viewer. The square dot should appear in the back left corner. The coordinates are principal components. Symbols represent six proposed clusters. The open squares on the right are not tightly clustered. The points connecting lines are minimal spanning tree lines

the eyes and the analyst does not have to learn the visual trick of decoupling eye convergence and lens focusing.

Figure 8 shows six clusters derived from rat spinal-chord gene-expression data, as shown later in Figure 12. The coordinates for the plot are the first three principal components derived from the nine summary measurements over time on each type of gene. In this case showing three coordinates rather than two increases the variability represented from 50% to 65%. Much variability is not captured in the plot, but looking at a flattened 2D plot is worse.

In Figure 8 the symbols distinguish the algorithmically defined clusters. A minimal spanning tree (*see* **Dendrogram**) connects the points in each cluster to provide a repeatable visual path from point to point. The visual task is to assess between-cluster separation and within-cluster tightness. While color would help to distinguish groups and overplotting is a bit of a problem, one can still see that the "octagon" and "plus" groups at mid-depth on the left are very close. The split between the two groups seems somewhat arbitrary. The "square" group on the right occupies much of the volume of the graphic. A reasonable conjecture is that this group will be divided into subgroups when a few thousand more genes are added to the analysis. A direct visual approach provides an alternative to algorithmic evaluation approaches based on questionable assumptions such as **multivariate normality**.

Carr et al. [22] suggest that glyphs using the best encoding, position along a scale, to represent coordinates will provide better judgments of interpoint distances than other glyph encodings. This motivates extending the stereo scatterplot to higher dimensions. To represent a fourth coordinate, Carr et al.selected from the class of good encodings. They chose the ray angle over line length to encode a fourth coordinate because this creates fewer ambiguities in overplotting situations.

The ray glyph has many other uses. Since people perceive ray angle more accurately than stereo depth, it is a reasonable choice for encoding a third variable, especially when the third variable is the dependent variable. For large data sets, Carr et al. [23] use hexagon binning to provide symbol congestion control. The ray angle provides a summary for a hexagon region, and with only one symbol per hexagon overplotting is not a problem. When rays represent estimated values that have confidence intervals, the authors represent the confidence intervals using arcs. Small reference wheels at the base of the ray provide an unobtrusive angular scale for comparison. Angles can be judged accurately. The drawback in the three independent variable setting is that interpoint distance assessment is not as natural as judging length in stereo plots.

For representing two dependent variables and two spatial position variables, Carr [12] uses a bivariate ray glyph. (Chambers et al. [27] describe many graphical representations, including a closely related metroglyph that represents both wind direction and speed; see also Anderson [2].) A ray pointing to the right encodes one variable (for small values the ray points down and for large values it points up) and a ray pointing to the left encodes the other. Simple ray plots provide an effective way to show four-dimensional (4D) information.

For five variables, Carr et al. [22] use ray angle and length to represent the last two coordinates. They show the ray angle in the plane of the display, for all rotations of the data to provide maximum visibility of the ray angle. The rays have to have a minimum length the keep the angle visible. Mapping the two coordinates into two spherical coordinate angles is inferior because it is difficult to decode the angle away from the display surface. The stereo parallax difference between the two ends of the ray encodes the depth angle information, and with a short ray there is almost no depth resolution. Length is a reasoned choice.

Carr et al. [22] suggest encoding a sixth coordinate using a carefully selected double-ended color scale going through gray. Color encoding, no matter how well chosen, is inferior to many other choices for representing a continuous variable. However, after five choices the options are extremely limited.

Few researchers have seriously tackled the visualization of six-dimensional data. The power of using a single glyph with well-chosen encoding remains little appreciated. A notable exception is Bayly et al. [7]. They successfully used colored stereo ellipsoids to evaluate problems in improving an electrostatic potential model. Their article includes color side-by-side stereo figures. In a long sequence of efforts Bayly (personal communication) failed to obtain insight using many of the encodings described in the entry. The breakthrough came when using

a carefully chosen stereo glyph. Understanding six-dimensional relationships is nontrivial. The selection of good encoding can be crucial.

Stereo-glyph encodings are good encodings for a difficult visualization problem. The encodings are asymmetric in terms of the coordinates. Interpoint distance assessment becomes progressively more difficult as one adds the angle, length, and color encodings. However, interpoint distance assessment using some of the methods below pales by comparison.

Glyph encodings that typically do not use coordinates to determine glyph position include Chernoff faces [28], star plots, profile plots or line-height plots, trees and castles [54], and cone plots [37]. Figure 9 shows a few examples of three encoding approaches. Chernoff faces encode variables using area of the face, shape of the face, length of the nose, location of the mouth, curve of mouth, and so on. Star plots, profile plots, and line-height plots are all variants of k-sided polygons [65]. In early work, Bertin [9] shows profile plots and Anderson [2] uses a restricted variant of star plots. Star plots have implicit [0, 1] axes at equal angles around a circle. The length of the segment along each implicit axis encodes the respective variable. Profile plots have implicit [0, 1] axes orthogonal to a horizontal base line. Segment heights indicate the magnitude of corresponding variables. Profile plots appear with minor variations. Here, the "profile" plot connects the tips of the segments and may optionally fill the polygon, hiding the construction. The version in Kleiner & Hartigan [54] draws adjacent boxes of the given height. The line-height version is a thin, detached box variation.

As one demonstration of the inferior nature of nonpositional glyphs in low dimensions, one can

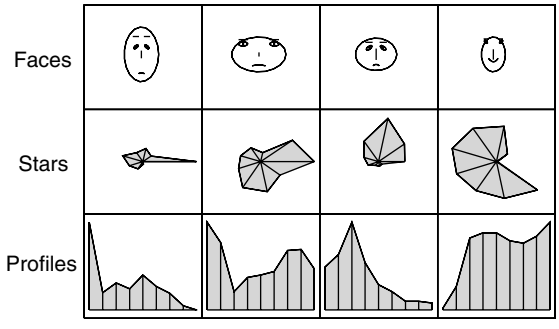

Figure 9 Three types of multivariate glyphs

generate, say, a thousand points of 3D data embedded in four dimensions. That is, select triples (u, v, w) randomly, say from a normal distribution, and then let $x_1 = f_1(u, v, w), x_2 = f_2(u, v, w), x_3 = f_3(u, v, w)$, and $x_4 = f_4(u, v, w)$, where the four functions are simple distinct polynomials. The structure in the stereo-ray glyph plot will immediately suggest the existence of a constraint and the fact that the data are not four-dimensional. Looking at 1000 Chernoff faces laid out in an two-way array using a happenstance order is not likely to give the slightest clue.

Almost any graphic can be improved. Much can be done to improve the nonpositional glyph encodings. In terms of faces, research has revealed much about our special face recognition "hardware". For example, face recognition improves if the faces are smiling and turned 15 degrees. (See Takacs [66] for a survey of the literature.) An improved face encoding might reduce the error rates cited in Chernoff & Haseeb [29]. The high similarity rankings reported in a paired comparison in an experiment by Wilkinson [80] might get even better. However, pair-comparison results do not necessarily extend to seeing patterns in a dense plot. Issues include seeing inside face outlines, judging distances when faces are not closely juxtaposed, and judging distances through interposed sequences of faces. A better layout often helps.

Like faces, the star representation is reasonably popular. The area of the star conveys a general notion of coordinate magnitude. Stars communicate when bigger is better.

In terms of layout, plotting round symbols such as faces and stars on a hexagon lattice seems reasonable. Variants of the algorithm discussed by Eick & Wills [39] can exchange icons such as stars with the objective of placing similar stars together. Taking this one step further, some of the hexagon lattice points can be strategically blank, to emphasize clusters. Basically, the notion is to use a modified stress criterion and a lattice-based variant of multidimensional scaling to strengthen the representations of interpoint distance.

Nonpositional glyphs come into their own in higher dimensions. Figure 10 is a line-height plot from Carr & Olsen [19] and shows the relative area of 159 vegetation classes for each of 26 continental US ecoregions. The area determination for each ecoregion followed after classifying 8 million pixels of a continental US AVHRR image into the

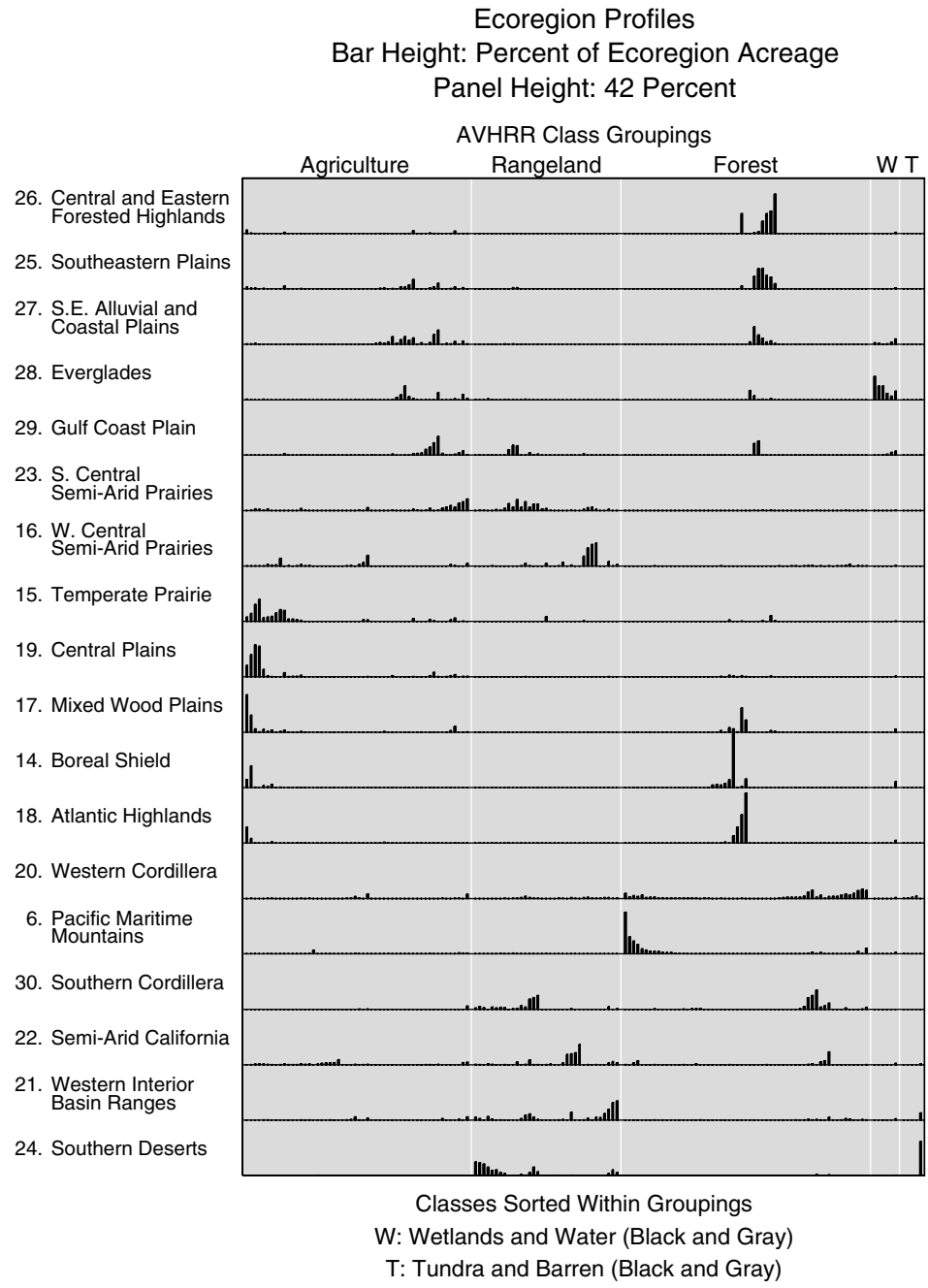

Figure 10 Line heights representing 159 variables. Separate row and column sorting creates visual clusters and simplifies appearance. Here the sorting order is the minimal spanning tree traversal order

159 vegetation classes. The theme of the article is on simplifying plot appearance by sorting rows and columns. One can infer from the example that representing 500 variables using line heights is not a problem for a small number of cases. The primary representation problem is in showing the labels. Mousing on a variable can reveal its label in an interactive setting. The real difficulty is in assessing the patterns presented.

In general, nonpositional glyphs show interpoint distances poorly. While stars and line-height glyphs are better balanced with respect to the coordinates than Chernoff faces, Kleiner & Hartigan [54] point out that most such representations remain asymmetric. That is, one can compare adjacent coordinates more easily than nonadjacent coordinates. Reordering can make a difference.

Nonpositional glyphs can represent a large number of variables. An extreme case is to encode each variable as the color of a pixel on a monitor. Thus, it is possible to represent a single case with over a million variables. The basic two problems remains: how to represent many cases, and how to interpret the graphic.

Linked Plots

Linking points across plots provides a way to connect the variables that are represented in different plots. Linking provides a weaker binding of the multivariate components than glyphs. Linking methods include linking by lines, colors, names, and pointers, and spatial linking by juxtaposition. The following discussion emphasizes line linking and color linking.

Diaconis & Friedman [38] discuss M and N plots that link points in different plots with lines. For example, they represent 4D data using two 2D scatterplots. The first plot represents the first two coordinates and the second plot represents the remaining two coordinates. A line between a bivariate point in one plot and a bivariate point in the second plot indicates that bivariate points really represent one four-coordinate point. Their general description includes linking across multiple plots of varying dimensionality. For example, a 4D representation might link a 1D plot to a 2D plot to a 1D plot.

Parallel coordinate plots are the only variation of M and N plots that have caught on. The parallel coordinate plot for p dimensions is a sequence of p univariate plots. The representation connects p coordinates with $p-1$ line segments. An early example appears in Bertin [9]. Inselberg [51] and Wegman [74] introduce the mathematical and statistical aspects of parallel coordinate plots. They and Inselberg & Dimsdale [52] describe the point-line duality and other mathematical relationships that provide a basis for extended interpretation. For example, Inselberg has used the representation to find the closest distance between two lines in four dimensions. Interpretation of some patterns requires significant background. Other patterns are easy. For example, Wegman notes that one can readily assess the correlation between adjacent variables. Many crossing segments between adjacent axes indicates a high negative correlation, and many parallel segments indicates a high positive correlation.

Carr & Olsen [18] found parallel coordinates useful for representing two variables in a map legend when space was at a premium. The parallel coordinate representation has particular merit in terms of labeling and reading selected value pairs. However, the scatterplot remains the preferred way of providing the gestalt of a functional relationship.

Parallel coordinate plots have particular merit in dynamic graphics. Carr & Nicholson [17] use unlinked parallel coordinates axes as a multivariate coordinate input device, in their case providing direct manipulation of a 4D stereo ray glyph cursor. The cumulative selection of points within a ball of the cursor and cursor movement allows higher-dimensional subset selection without the usual cross-product set restrictions. Carr & Nicholson also display marginal densities on the axes. As rotation and masking alter the 4D view, this provides information about multimodality in margin views. Parallel coordinates are very useful for dynamic subset selection.

The line-linked plot paradigm has several weaknesses. First, coordinate representation is not symmetric. That is, assessing relationships within plots is easier that assessing relationships between plots. Owing to the difficulty in following links from plot to plot, the relationships between variables in adjacent plots are easier to assess than relationships to variables in distant plots. Following lines from plot to plot is sometimes impossible owing to line overplotting. Diaconis & Friedman [38] suggest reducing the line density by connecting coordinates of only one point of each small cluster of points in the p-dimensional space. Miller & Wegman [62] take a different approach by calculating and representing

line density. The line density plots convey much more information about clusters than overplotted lines. The line density approach extends to large data sets and dynamic cluster selection, as described by Wegman & Luo [78]. However, the approach does not solve the problem of poor linking to nonadjacent plots.

In terms of understanding geometric structure, line-linked plots leave much to be desired. Humans are exceedingly poor at tomographic reconstruction. In the simplest of cases, clear structure in 2D scatterplots can be hard to fathom by looking at linked 1D margin plots. The assessment of distance between two multivariate points requires integration across all the coordinate axes, and this is complicated enough when all the coordinates are available at a glance, as in glyph displays. Beyond suggesting clusters by line overplotting density, the line-linked plots have not caught on for understanding geometric structure.

A more popular linking technique is color linking. The common application is the dynamic brushing of points in a scatterplot matrix [8]. The points brushed (selected) in one panel become highlighted in all views of their coordinates. Brushing is a focusing technique. It also serves as a conditioning technique that often lowers dimensionality.

Since preattentive vision can handle the location of distinctively colored points, color is a faster link than lines. Of course, overplotting and multiple multivariate points with the same color cause ambiguities. Color linking has several problems, some relating to color perception limitations (for example, color guidelines suggest working with six or fewer distinct hues) and some relating to poor software implementations.

In terms of implementations, the plotting order of points may not be controlled and overplotting can hide selected points. Subset memberships may not be well represented. Carr et al. [22] discuss a color scheme for representing disjoint subsets of three nonexclusive, interactively defined subsets. Their example identified a two-coordinate symmetry in a seven-dimensional particle physics data set. Sophisticated color treatment can provide subset representation and color mixing for overplotted points. Advanced graphics workstations have four color channels, red, green, blue, and alpha. The alpha channel is there specifically to provide color-blending options. Statistical graphics software has been slow to exploit the capability.

Positional linking is another option. The layout of the scatterplot matrix enables positional linking. Suppose the x coordinate of a point is 0 and well separated from the x values for other points. Then it is possible to identify the other views of the point in the scatterplot matrix by looking along the line $x = 0$. When the x coordinate is not well separated from other x coordinates, identifying the remaining coordinates of a point can be difficult unless they are highlighted by color or other means. The difficulty in finding coordinates suggests that assessing interpoint distances in a scatterplot matrix will be next to impossible, except in pathologically simple cases. While the scatterplot matrix is useful in the context of brushing, and as a multiple window display for viewing many 2D projections, the scatterplot matrix is not a good choice for perceiving higher-dimensional geometric structure. Random points on a mobius strip have obvious structure in a stereo plot, while the structure is obscure in a scatterplot matrix.

Positional linking comes into its own for small perceptual groups. Figure 11 shows positional linking of labels and dots in a dot plot. One can easily match the middle label of a group to the middle dot of a group. The grouping of labels and dots into small units of five reduces the chances of matching error. A solid list of 62 names is visually intimidating, so the groups encourage reading. More generally, Carr & Pierson [20] describe micromap designs that use position and color to link statistical summaries with small maps. The position region and shape in a small map can also link to the corresponding position and shape in a large map. Positional linking is a powerful and underutilized tool.

Nested Plots

The classic example of nesting is the casement display [72]. The basic casement display is a matrix of scatterplots, each with identical scales. The casement display partitions the data into a crossed two-way layout using two of the four coordinates. The two layout coordinates define panel membership, and the two remaining coordinates appear in the scatterplot. The casement display is not symmetric in the coordinates, sacrificing resolution for the two coordinates defining the layout.

The nesting template can be varied in many ways. Panels can show a higher-dimensional relationship. Carr [16] provides a five-dimensional nested display.

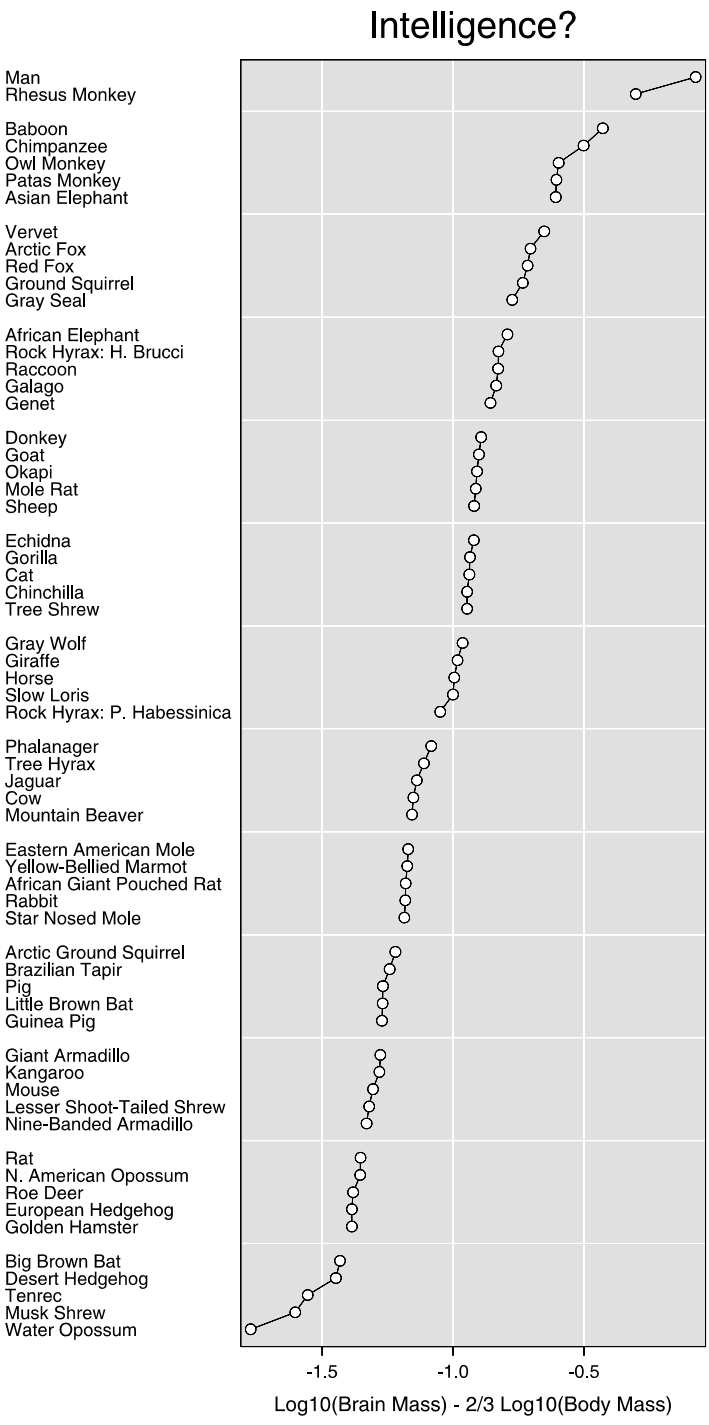

Figure 11 A dot plot with positional linking. Perceptual grouping into small units obviates the need for lines from labels to dots. The Owl Monkey is a third label, and finding the corresponding third dot is trivial even across the page. Vertical grid lines increase the perceptual accuracy of extraction

The data set concerns a dependent variable, namely protein folding energy, and four independent variables. The independent variables have seven levels and are fully crossed. The seven-by-seven panel layout presents all combinations for two of the independent variables. Each panel within the layout panel uses x and y coordinates to represent the two remaining independent variables and a ray angle to represent the folding energy. Highlighted rays distinguish local minima as computed using eight neighboring points, two for each dimension. The highlighted rays call attention to both local minima and saddlepoint troughs through space.

With only one layer of nesting in each of the horizontal and vertical directions, nested plots are equivalent to the conditioned plots described further below. The two-way layout is one of the most powerful templates available for getting two extra variables into a graph. While the resolution is typically poor for the two variables defining the layout, the resolution loss is often overshadowed by that fact that many people readily understand two-way conditioning.

Nested views can be nested. Mihalisin et al. [60, 61] describe deeper layers of nesting to handle up to 10 variables. With study, people can learn to spot certain classes of mathematical relationships in deeply nested views. Nested views are not variable-balanced views, but have a role in the arsenal of tools.

Conditioned Plots

Casement displays generalize to plots conditioned by many variables. An early exposition on conditioned plot (or coplots) appears in Cleveland et al. [35]. Conditioned plots are typically 2D plots, but they can be 3D wireframe plots or other higher-dimensional plots. People readily understand one-and two-way layouts and conditioned plots build upon this understanding.

Trellis Graphics™ automate the multiple-panel plot production process. This includes labeling of panels by factor names and graphical representation of factor levels. Trellis layout capabilities also address the issue of panel shape. Panel shape can strongly affect the perception of line slope, and banking of slopes to be close to $\pm 45^{\circ}$ is important in some problems [30]. Trellis graphics provide many sound defaults and a good framework for multivariate graphics.

Conditioned (and nested) views do not have to partition the data strictly to produce different panels. Cleveland et al. [35] introduces the notion of shingles that allow the same observations to appear in more than one panel. This is helpful when smoothing a scatterplot because it increases the number of points in the plots and addresses poor smoothing at the plot edges.

Laying out multiway panels in rows and columns across many pages is often a good start. However, general-purpose algorithms have not yet captured all the current graphical design expertise. Methods for simplifying visual appearance remain applicable. These include grouping of information, sorting and presenting the information in layers, and removal of redundant information [14, 55]. The graphics tools make it easy to apply thoughtful sorting, but the analyst still has to do the thinking.

The difficulty of seeing patterns across levels of conditioning factors and pages needs to be recognized. As indicated above, humans are not good at integrating low-dimensional relationships into higher-dimensional or overview patterns. When the information appears across pages, the limits of our short-term memories compound the difficulty. When across-panel insights occur, they are likely to be based on panels juxtaposed closely in space or time. Careful attention to the choice of layout is often the key to obtaining multivariate insights.

Geometric Section Plots

Multiple univariate sectioning is a standard technique for lowering dimensionality and seeing through objects. The cone plot [37] provides an interesting example of a nonstandard sectioned plot. One picks a vertex in p-space and a reference line through the vertex (typically defined by selecting a data point as a second point on the line through the vertex). The reference line becomes the center line for a set of cones. Connecting a different data point to the vertex creates a second line. Rotating this line around the center line defines two cones with tips that touch at the vertex. Each data point is then associated with the angle between its line and the cone center line. Each point also has a distance to the vertex. Thus, the selection of a vertex and a reference line associates an angle and a distance with each data point. Dawkins uses the natural plot for a distance and an angle, the polar coordinate plot. A

different plot results from each selection of a center line. Typically, Dawkins uses each case to define the center line and this generates a plot (or panel) for each case. Dawkins lays the panels out as a matrix and describes the geometric relationships, such as hyperplanes, observable within panels. Like nested plots and parallel-coordinate plots, a certain amount of knowledge is required before the plots are fully appreciated, and this limits the audience.

Series Plots

Andrews [3] provides a series transformation that represents a continuous multivariate observation as a curve. If X_i is the ith coordinate of an observation, then the curve for the observation is

$$y(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots$$

for $-\pi < t < \pi$. Nearly identical observations will plot similar curves. The plot can be used to assess the clustering of observations. However, the first coordinates are assigned to the lowest frequencies and the low frequencies tend to be visually dominant. Hence, the variation in the first coordinates is visually dominant. The representation is not symmetric in the variables.

The encoding can be modified to provide an animation. Slicing through the curves with a line at time t yields a set of points. Varying t then leads to the animation of points along a scale. Some have thought this might be a one-dimensional analog of the 2D grand tour mentioned below. However, Wegman & Shen [79] note that the curves are not properly defined to provide a grand tour. They provide a generalization that shows curves in three dimensions and illustrate direct manipulation of a sectioning plane. The curves intersect the plane to create a 2D scatter of points. This is advantageous in that clusters can separate better in a 2D sectioned view than they do in a 1D sectioned view.

Functions have numerous series approximations. Different series approximations can lead to different graphics. One might even try wavelet encoding.

Composite Plots

Composite plots provide a good way to look at higher dimensions. These plots often involve elements of conditioning and exploit the capacity of the analyst to understand two-way layouts. Figure 12, adapted from Carr et al. [24], provides a composite plot example. The data are rat spinal-chord gene-expression values observed at nine times. The times are gestation days 11, 13, 15, 18, 21; postbirth days 0, 7, 14, and adult. The gene-expression values range from 0 to 1 and indicate how fully the gene is functioning. The basic plot is a time-series line plot, which is equivalent to a parallel coordinate plot with axes at the time points. The overplotting of all the data in a single parallel coordinate plot is not acceptable in looking at individual genes. Figure 12 uses small multiples (see Tufte [67]) to avoid heavy overplotting. Grouping into units of four or fewer is cognitively advantageous. With four or fewer easily discriminated line textures, it is simple to match lines with the gene labels. (The original used four colors rather than four textures.) Arbitrary grouping into units of four can raise questions and miss an opportunity. Figure 12 starts the grouping process by conditioning on membership in 14 gene function groups. The gene function group name is at the left of each corresponding set of panels. Thus, Figure 12 combines elements of conditioned plots, parallel coordinates plots, and line texture linking. Showing $1008 = 112 \times 9$ values on a page is no problem.

Multiple Views, Dimensionality, Cognostics, and Projection Searches

The above review of graphics representations suggests a large number of view options. Cross these viewing options with today's data explosion and the number of possible graphics seems overwhelming. We are not going to look at all potentially insightful plots. Something needs to give.

In most fields of endeavor one feels obliged to find the obvious. Failing to find the complex is forgivable in the rare cases that someone notices. In multivariate graphics, is it important to look at low-dimensional views of the data to see if there are obvious patterns.

Frequently data relationships consist of a low-dimensional geometric structure embedded in much higher-dimensional data. For example, simple mixtures of two multivariate points generate a line segment. Mixture of three multivariate points generates a triangle. The observational process can lead

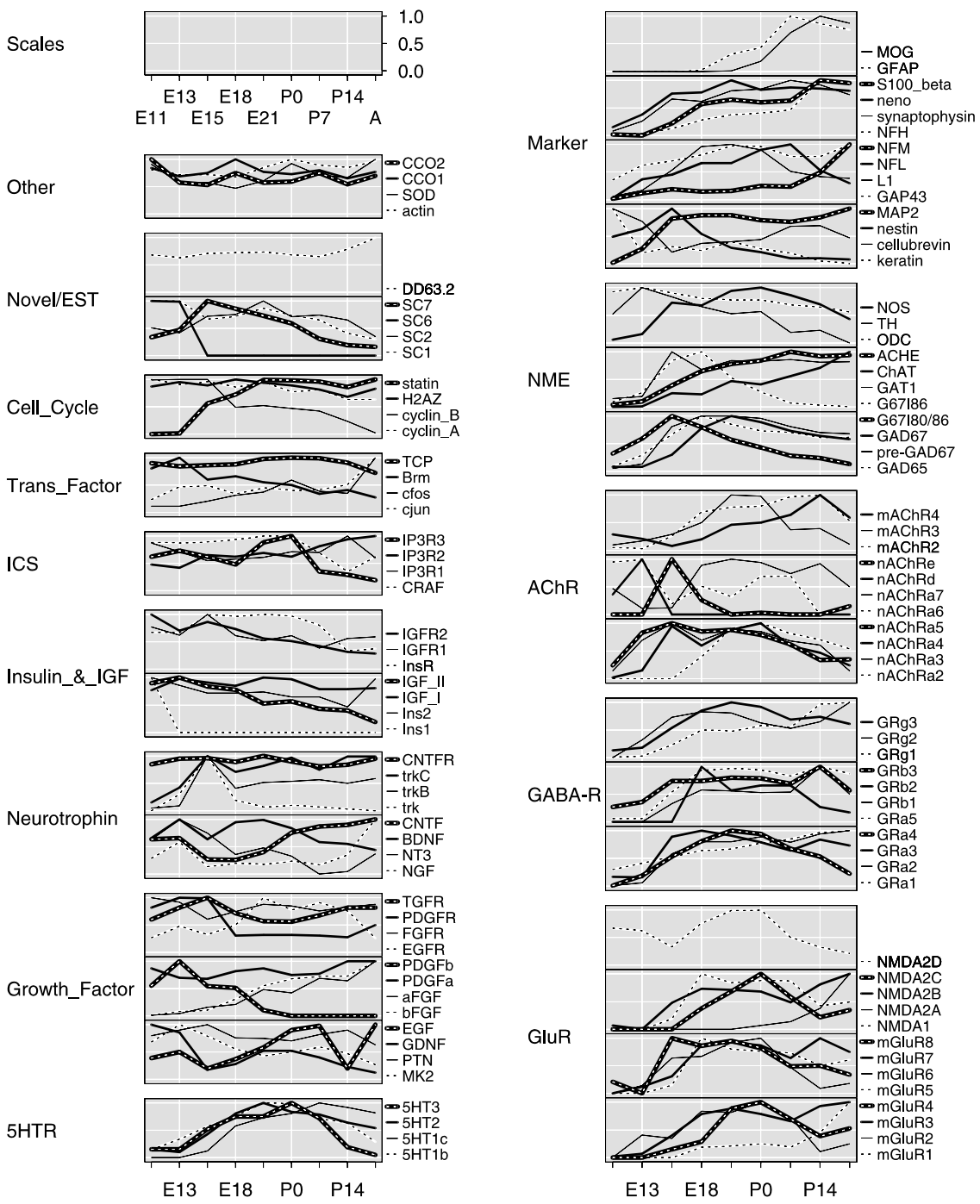

Figure 12 A multiple panel plot showing 112 time series with grouping and individual series labels

to recording a mixture value. For example, the spectral bands for pixels in a satellite image may reflect the mixture of two vegetation classes with distinct spectral signatures. Cook et al. [34] report an example of seven coordinate particle physics data that geometrically consists of a triangle with two line segments at each vertex. Since the structure was embedded in seven dimensions, the discovery process using 2D plots was nontrivial. Nonetheless, looking for the simple patterns sometimes produces results.

Carr et al. [21] advocate looking at low-dimensional margin views in part because the coordinates have an immediate interpretation. A systematic approach involves looking at 1D margin views, scatterplot matrices, stereo-scatterplot triples, and 4D plots such as the stereo-ray glyph plots. For a large number of variables, there can be layout problems. In a single screen view, panels in a scatterplot matrix can become too small to be useful. Also, increasing the dimensionality of the view causes problems. Given, say, 10 variables there are 45 2D panels and 120 3D panels. Approaches such as pan and zoom help in the 2D cases and sectioning helps in the 3D cases. However, with increasing dimensionality the notion of a quick visual sweep over the whole space is quickly lost.

In thinking about the future, Tukey [71] coined the term *cognostics* (diagnostics interpreted by a computer rather than a human). The idea was to compute features of merit and have a computer rank the plots by their potential interest to humans. Paul Tukey [73] developed some early features of merit for prioritizing scatterplots. Carr [12] developed different features of merit oriented toward large data sets. The cognostics involved binning 2D and 3D data as a starting step. One of the more nontraditional cognostics used extensions of thinning (an image processing technique) to assess the extent of skeletal structure in low-density regions. The application involved views of computational fluid dynamics (CFD) variables and the features of merit were computed in parallel, one time step behind the CFD model. The strategy was to store information and later to instantiate only high interest virtual plots. The general concept applies to looking for patterns in biological databases.

Asimov [4] proposed grand tour methods that provide an infinite sequence of 2D projected views. Wegman [75] discusses a generalized grand tour with the number of resulting coordinates being the same number as in the original data. Thus, tour views are not restricted to 2D projected views. The many views provided by grand tour sequences tend to overwhelm the mind rather than to facilitate quick insight. Similar to the cognostics idea, **projection pursuit** (see Cook et al. [36]) assists the analysts in finding interesting views.

When one is looking for structure (and constraints) in low-dimensional plots, an understanding of what happens in projection to low dimensions is very helpful. Furnas & Buja [46] shed insight into what can be seen in low-dimensional views using projection and sectioning. For example, lines in high dimensions project into lines in low dimensions.

Sectioning (also called slicing and masking in the graphics literature) adds mathematical constraints that typically lower the dimensionality. For example, adding the constraint $x = 2$ to the line $y = mx + b$ reduces the line to a point. Typically, sectioning is implemented as a logical "and" among constraints on individual coordinates. Direct manipulation provides one way of controlling the bounds $a_i < x_i < b_i$ on the ith coordinate x_i. Sectioning can apply to computed variables, including those constructed in touring sequences. Carr & Nicholson [17] describe software that enables hyperplane sectioning in the form of $a < \mathbf{c}'(\mathbf{x} - \mathbf{x}_0) < b$, where $\mathbf{c}$, $\mathbf{x}$, and $\mathbf{x}_0$ are vectors. Graphical methods, using parallel axes, define the normal vector to the hyperplane, $\mathbf{c}$, and a multivariate point, $\mathbf{x}_0$, in the hyperplane. Direct manipulation controls the scalars a and b. (An extension of the software providing stereo-ray glyph, scatterplot matrix and parallel coordinate views, alpha blending for color mixing, and other options is available; see Carr et al. [26].) Sectioning can reveal holes and other structures in the data that remain hidden in projected views.

A geometric structure is easier to identify when the viewing dimension is higher than the dimension of the structure. For example, points on a surface projected in a 2D scatterplot can saturate the plot, while the surface remains evident in a 3D scatterplot. Stereo 3D plots are the natural environment for viewing the 2D structures of such surfaces. 4D plots, such as the stereo-ray glyph plot, provide a way to look for 3D structures. Sectioning is useful to lower the dimensionality of the data until it falls below the dimensionality of the display.

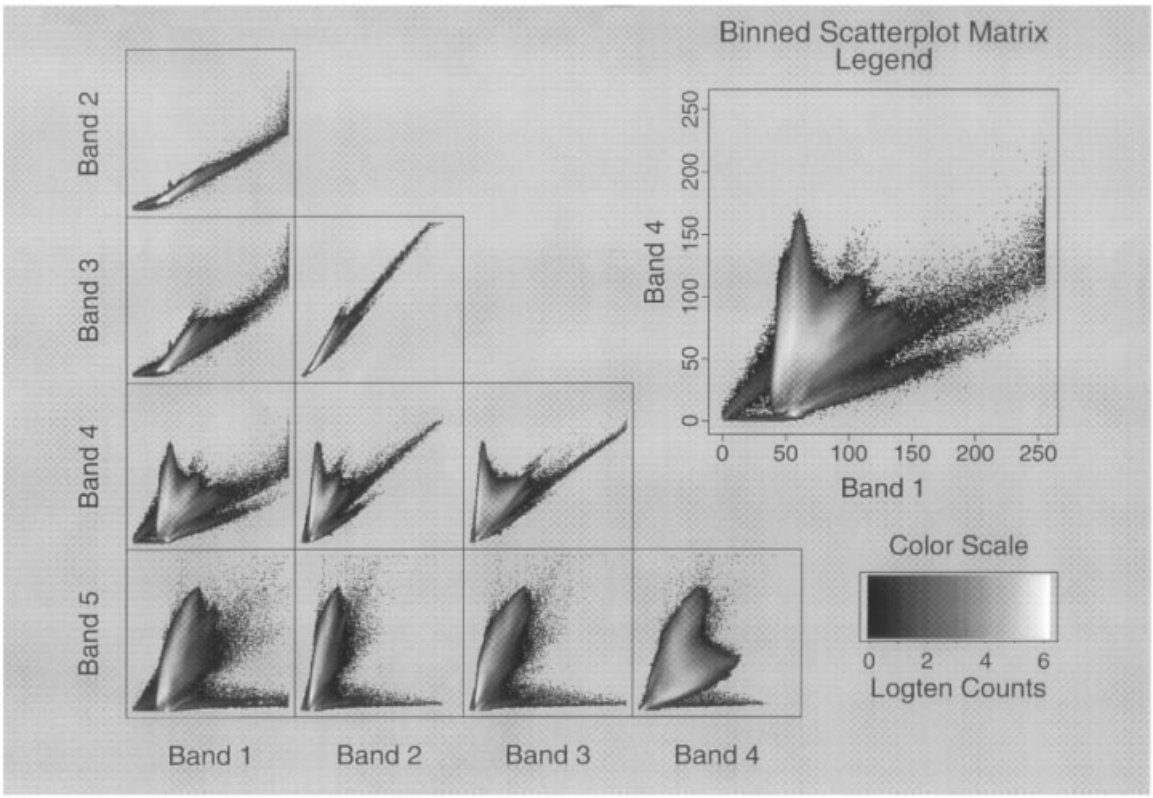

Figure 13 A binned scatterplot matrix showing over one-third of a billion nonzero point pairs. This overview crudely represents density features using gray level for counts on a log 10 scale. A pure white dot represents over a million satellite image pixels with the same bivariate spectral intensity

Massive Data Sets and Closing Remarks

Massive data sets pose new challenges. Some of the old methods work for selected problems. As one example, Figure 13 is a gray level version of a binned data scatterplot matrix presented at the 1995 American Statistical Association annual meeting. Each panel represents intensities for two spectral bands (wavelength intervals) for 54 million pixels of a multispectral satellite image. The recorded values are intensities for each spectral band, as a measure on a scale from 0 to 255. The different panels correspond to different pairings of the first five spectral bands from the seven spectral band image. (The binning discussion for six variables below omits band 6 because it has a different spatial resolution.) The approach of 2D binning and density representation scales well in terms of the sample size. Binning 54 million 2D points for each of the 15 pairings is not a problem. Visualization gets a bit more challenging since the bivariate densities (shown on a log scale) vary over five orders of magnitude. Thus, interactive methods become useful in bringing out density features.

Some tasks do not scale well to massive data sets. Challenges in this satellite data problem include six-dimensional (6D) binning without sacrificing 256 value resolutions per channel, viewing the higher-dimensional density structures, and linking density patterns in the binned panels back to the spatial coordinates of the pixels to benefit from the spatial information. With six spectral channels (variables) of interest, each with 2^8 possible intensities, there are 2^{48} potential cells. Straightforward binning that allocates space for all possible cells breaks down. Methods have to focus more on occupied cells. In this example there will be at most 54 million occupied cells. It turns out, via sorting and then binning, that there are over 12 million occupied cells. This is too many cells

for conventional methods. Making progress almost inevitably involves sacrificing intensity resolution via univariate rescaling or via use of a large number of clusters.

With resolution reduction, viewing the 6D binned density becomes feasible using a combination of conditioning and positional glyphs. For example, a two-way layout of panels can represent levels for two variables. Individual panels can show three coordinates using the 3D position. The ray angle can represent a few levels of a discrete fourth variable, and the ray length can be proportional to the log of the count. This is similar to showing small histograms in 3D space. The ray representation is advantageous owing to ambiguities resulting from overplotted histograms. Conceptually it is possible to zoom in to get higher resolution detail. A common trick in very low dimensions is to implement zooming with precomputed images. However, massive data sets require significant processing and the increased dimensionality makes it harder to anticipate where attention will be focused.

A large problem arises when it is time to link the density feature back to the 54 million spatial coordinates. Brushing in geographic space or attributing density space and viewing the result in the opposite space is no longer a trivial task. Since a typical (1280×1024) workstation screen has about 1.3 million pixels, one has to pan through many screens just to view all the pixels indicating spatial position. The next satellite, with 36 channels, possesses an even more formidable challenge. Challenges abound as we attempt to bring visualization methods to increasingly larger, higher-dimensional, higher resolution data sets.

The above description has been a whirlwind tour. We can represent a million variables. We can represent millions of cases. Finding and communicating multivariate patterns is another story. We have to work hard to find meaningful patterns. There are many barriers to the discovery of important patterns, not the least of which is the plethora of accidental patterns that will not replicate. We need statistical methods to help us find our way, and results from cognitive science so that we can more fully use the power of our minds to see.

Having found meaningful patterns does not mean that we will be able to communicate those patterns. A picture may be worth a thousand words, but in this electronic era neither a picture nor a thousand words is worth very much. Unless graphics are apparently simple, they are not likely to survive the first glance. Communicating a complex multivariate structure to a world seeking simple solutions is a major challenge. This article provides some guidance toward graphics that communicate, but much work remains ahead.

References

[1] Ahlberg, C. & Schneiderman, B. (1994). Visual information seeking: tight coupling of dynamic query filters with starfield displays, *Proceedings of the ACM Chi International Conference on Human Factors in Computing (Chi'94)*. Boston, Mass., pp. 303–317.

[2] Anderson, E. (1960). A semi-graphical method for the analysis of complex problems, *Technometrics* **2**, 287–292.

[3] Andrews, D.F. (1972). Plots of high dimensional data, *Biometrics* **28**, 125–136.

[4] Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data, *SIAM Journal of Scientific and Statistical Computing* **6**, 128–143.

[5] Baird, J.C. & Noma, E. (1978). *Fundamentals of Scaling and Psychophysics*. Wiley, New York.

[6] Barnett, V. (1981). *Interpreting Multivariate Data*. Wiley, New York.

[7] Bayly, C.I., Cieplak, P., Cornell, W.D. & Kollman, P.A. (1993). A well-behaved electostatic potential based method using charge restraints for deriving atomic charges: the RESP model, *Journal of Physical Chemistry* **97**, 10269–10280.

[8] Becker, R.A. & Cleveland, W.S. (1987). Brushing scatterplots, *Technometrics* **29**, 127–142.

[9] Bertin, J. (1967). *Semiologie Graphique*. Gauthier-Villars, Paris; *Semiology of Graphics*, translated by W.J. Berg. The University of Wisconsin Press, Madison, 1983.

[10] Brewer, C.A. (1994). Color use guidelines for mapping and visualization, in *Visualization in Modern Cartography*, A.M. MacEachren & D.R.F. Taylor, eds. Pergamon/Elsevier Science, Oxford, pp. 123–147.

[11] Bruce, A. & Goa, H. (1996). *Applied Wavelet Analysis with S-PLUS*. Springer-Verlag, New York.

[12] Carr, D.B. (1991). Looking at large data sets using binned data plots, in *Computing and Graphics in Statistics*, A. Buja & P. Tukey, eds. Springer-Verlag, New York, pp. 7–39.

[13] Carr, D.B. (1993). Production of stereoscopic displays for data analysis, *Statistical Computing and Graphics Newsletter* **4**, 2–7.

[14] Carr, D.B. (1994). Converting tables to plots, *Technical Report No. 101*. Center for Computational Statistics, George Mason University, Fairfax.

[15] Carr, D.B. (1994). A colorful variation on boxplots, *Statistical Computing and Graphics Newsletter* **5**, 19–23.

[16] Carr, D.B. (1995). Scanning a 4-D domain for local minima: a protein folding application, *Statistical Computing and Graphics Newsletter* **6**, 8–12.

[17] Carr, D.B. & Nicholson, W.L. (1988). EXPLOR4: a program for exploring four-dimensional data, in *Dynamic Graphics for Statistics*, W.S. Cleveland & M.E. McGill, eds. Wadsworth, Belmont, pp. 309–329.

[18] Carr, D.B. & Olsen, A.R. (1995). Parallel coordinate plots for representing distribution summaries in map legends, in *Proceedings 1 of the 17th International Cartography Association Conference, 10th General Assembly of the ICA*, Institute Cartogràfic de Catalunyian, Barcelona Catalunya, España, pp. 733–742.

[19] Carr, D.B. & Olsen, A.R. (1996). Simplifying visual appearance by sorting: an example using 159 AVHRR classes, *Statistical Computing and Graphics Newsletter* **7**, 10–16.

[20] Carr, D.B. & Pierson, S.M. (1996). Emphasizing statistical summaries and showing spatial context with micromaps, *Statistical Computing and Graphics Newsletter* **7**, 16–23.

[21] Carr, D.B., Littlefield, R.J., Nicholson, W.L. & Littlefield, J.S. (1987). Scatterplot matrix techniques for large N, *Journal of the American Statistical Association* **82**, 424–436.

[22] Carr, D.B., Nicholson, W.L., Littlefield, R.J. & Hall, D.L. (1986). Interactive color display methods for multivariate data, in *Statistical Image Processing and Graphics*, E.J. Wegman & D.J. DePriest, eds. Marcel Dekker, New York, pp. 215–250.

[23] Carr, D.B., Olsen, A.R. & White, D. (1992). Hexagon mosaic maps for display of univariate and bivariate geographical data, *Cartography and Geographic Information Systems* **19**, 228–236, 271.

[24] Carr, D.B., Somogyi, R. & Michaels, G.S. (1997). Simple templates for looking at data and clusters: gene expression examples, *Statistical Computing and Graphics Newsletter* **8**, No. 1, 20–29.

[25] Carr, D.B., Valliant, R. & Rope, D. (1996). Plot interpretation and information webs: a time-series example from the Bureau of Labor Statistics, *Statistical Computing and Graphics Newsletter* **7**, No. 2, 19–26.

[26] Carr, D.B., Wegman, E.J. & Luo, Q. (1997). ExplorN: design considerations past and present, Center for Computation Statistics, *Technical Report No. 137*. George Mason University, Fairfax.

[27] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Wadsworth & Brooks/Cole, Pacific Grove.

[28] Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association* **68**, 361–368.

[29] Chernoff, H. & Haseeb, R.M. (1975). Effect on classification error or random permutation of features in representing multivariate data by faces, *Journal of the American Statistical Association* **70**, 548–554.

[30] Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, Summit.

[31] Cleveland, W.S. (1994). *The Elements of Graphing Data*. Hobart Press, Summit.

[32] Cleveland, W.S., ed. (1988). *The Collect Works of John W. Tukey*. Vol. V. *Graphics 1965–1985*. Wadsworth & Brooks/Cole, Pacific Grove.

[33] Cleveland, W.S. &McGill, M.E. eds. (1988). *Dynamic Graphics for Statistics*. Chapman & Hall, New York.

[34] Cleveland, W.S. & McGill, R. (1984). Graphical perception: theory, experimentation, and application to the development of graphics methods, *Journal of the American Statistical Association* **79**, 531–554.

[35] Cleveland, W.S., Grosse, E. & Shyu, W.M. (1992). Local regression models, in *Statistical Models In S*, J.M. Chambers & T.J. Hastie, eds. Wadsworth & Brooks/Cole, Pacific Grove.

[36] Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995). Grand tour and projection pursuit, *Journal of Computational and Statistical Graphics* **4**, 155–171.

[37] Dawkins, B.P. (1995). Investigation the geometry of a p-dimensional data set, *Journal of the American Statistical Association* **90**, 350–359.

[38] Diaconis, P. & Friedman, J.H. (1980). M and N Plots. *Pub-2495*. Stanford Linear Accelerator Center, Stanford University, Stanford.

[39] Eick, S.G. & Wills, G.J. (1993). Navigating large networks with hierarchies, in *IEEE Computer Society, Proceedings of the Conference on Visualization 1993*, IEEE Computer Society, Los Alamitos, pp. 204–210.

[40] Eick, S.G., Steffen, J. & Sumner, E. (1992). Seesoft – a tool for visualization software, *IEEE Transactions on Software Engineering* **18**, 957–968.

[41] Fienberg, S.E. (1979). Graphical methods in statistics, *American Statistician* **33**, 165–178.

[42] Fisher, L.D. & van Belle, G. (1993). *Biostatistics A Methodology for The Health Sciences*. Wiley, New York.

[43] Foley, J.D., van Dam, A., Feiner, W.K. & Hughes, J.F. (1990). *Computer Graphics: Principles and Practice*. Addison-Wesley, Reading.

[44] Friedhoff, R.M. & Benzon, W. (1991). *The Second Computer Revolution, Visualization*. Freeman, New York.

[45] Frigge, M., Hoaglin, D.C. & Iglewicz, B. (1989). Some implementations of the boxplot, *American Statistician* **43**, 50–54.

[46] Furnas, G.W. & Buja, A. (1994). Prosection views: dimensional inference through sections and projections, *Journal of Computational and Graphical Statistics* **3**, 323–353.

[47] Gnanadesikan, R. (1977). *Methods of Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

[48] Golub, G.H. & von Matt, U. (1997). Generalized cross-validation for large-scale problems, *Journal of Computational and Graphical Statistics* **6**, 1–34.

[49] Grinstein, G. & Levkowitz, H., eds (1995). *Perceptual Issues in Visualization*. Springer-Verlag, New York.

[50] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, New York.

[51] Inselberg, A. (1985). The plane with parallel coordinates, *Visual Computer* **1**, 69–96.

[52] Inselberg A. & Dimsdale, B. (1994). Multidimensional lines II: proximity and applications, *SIAM Journal of Applied Mathematics* **54**, 578–596.

[53] Jones, P.G. & Cook, D. (1995). Multivariate Q-Q plots based on quantile contours, *Computing Science and Statistics* **27**, 269–278.

[54] Kleiner, B. & Hartigan, J.A. (1981). Representing points in many dimension by trees and castles, *Journal of the American Statistical Association* **76**, 260–276.

[55] Kosslyn, S.M. (1994). *Elements of Graph Design*. Freeman, New York.

[56] Levkowitz, H. (1997). *Color Theory & Modeling for Computer Graphics, Visualization and Multimedia*. Kluwer, Boston.

[57] MacEachren, A.M. (1994). *Some Truth with Maps: A Primer on Symbolization and Design*. Association of American Cartographers, Washington.

[58] MacEachren, A.M. (1995). *How Maps Work*. The Guilford Press, New York.

[59] Marr, D. (1985). Vision: the philosophy and the approach, in *Issues in Cognitive Modeling*, M. Aitkenhead & M.M. Slack, Eds. Lawrence Erlbaum, London, pp. 103–126.

[60] Mihalisin, T., Timlin, J. & Mihalisin, J. (1995). Fast visual analysis of combinatoric multidimensional data, *Computing Science and Statistics* **27**, 225–229.

[61] Mihalisin, T., Timlin, J., Schwegler, J. (1991). Visualizing multivariate function, data and distributions, *IEEE Computer Graphics and Applications* **11**, 28–35.

[62] Miller, J.J. & Wegman, E.J. (1990). Construction of line densities for parallel coordinate plots, in *Computing and Graphics in Statistics*, A. Buja & P.A. Tukey, eds. Springer-Verlag. New York, pp. 107–124.

[63] Rao, R. & Card, S.K. (1994). The table lens: merging graphical and symbolic representation in an interactive focus + context visualization for tabular information, in *Proceedings of the ACM Chi International Conference on Human Factors in Computing (Chi 1994)*, Boston, Mass., pp. 318–322.

[64] Scott, D.W. (1992). *Multivariate Density Estimation; Theory, Practice and Visualization*. Wiley, New York.

[65] Siegel, J.H., Goldwyn, R.M. & Friedman, H.P. (1971). Pattern and process of the evolution of human septic shock, *Surgery* **70**, 232–245.

[66] Takacs, B. (1996). Perception and Recognition of Human Faces, *Ph.D. Thesis*. George Mason University, Fairfax.

[67] Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire.

[68] Tufte, E.R. (1990). *Envisioning Information*. Graphics Press, Cheshire.

[69] Tufte, E.R. (1997). *Visual Explanations*. Graphics Press, Cheshire.

[70] Tukey, J.W. (1979). Statistical mapping: what should not be plotted, in *Proceedings of the 1976 Workshop on Automated Cartography*. DHEW Publication No. (PHS) 79–1254, pp. 18–26.

[71] Tukey, J.W. (1983). Another look at the future, in *Computer Science and Statistics: Proceedings of the Fourteenth Symposium on the Interface*, K.W. Heiner, R.S. Sacher & J.W. Wilkinson, eds. Springer-Verlag, New York, pp. 2–8.

[72] Tukey, J.W. & Tukey, P.A. (1983). Some graphics for studying four-dimensional data, in *Computer Science and Statistics: Proceedings of the Fourteenth Symposium on the Interface*, K.W. Heiner, R.S. Sacher & J.W. Wilkinson, eds. Springer-Verlag, New York, pp. 60–66.

[73] Tukey, P.A. (1986). Unpublished presentation at the *Computer Science and Statistics Eighteenth Symposium on the Interface*.

[74] Wegman, E.J. (1990). Hyperdimensional analysis using parallel coordinates, *Journal of the American Statistical Association* **85**, 664–675.

[75] Wegman, E.J. (1991). The grand tour in k dimensions, in *Computing Science and Statistics, Proceedings of the Twenty-second Symposium on the Interface*, E.M. Keramidas, ed. Interface Foundation of North America, Inc., Fairfax Station, pp. 127–136.

[76] Wegman, E.J. & Carr, D.B. (1993). Statistical graphics and visualization, in *Handbook of Statistics, Computational Statistics*, Vol. 9, C.R. Rao, ed. North-Holland, New York, pp. 857–958.

[77] Wegman, E.J. & Luo, Q. (1994). *Visualizing Densities*, Technical Report No. 100. Center for Computational Statistics, George Mason University, Fairfax.

[78] Wegman, E.J. & Luo, Q. (1997). High-dimensional clustering using parallel coordinates and the grand tour, *Computing Science and Statistics* **28**, 361–368.

[79] Wegman, E.J. & Shen, J. (1993). Three dimensional Andrews plots and the grand tour, *Computing Science and Statistics* **25**, 284–288.

[80] Wilkinson, L. (1982). An experimental evaluation of multivariate graphical point representations, in *Proceedings of Human Factors in Computing Systems*. Gaithersburg, MD, pp. 202–209.

[81] Wood, D.W. (1992). *The Power of Maps*. The Guilford Press, New York.

DANIEL B. CARR

Multivariate Median and Rank Sum Tests

Multivariate **nonparametric** procedures were mostly developed in the 1960s. There were some impasses from the univariate to the multivariate channels, and a basic *rank-permutation principle*, developed by Chatterjee & Sen [6], opened up the broad avenue to multivariate nonparametrics. These include the multivariate one-sample model, two- and several-sample problems, as well as **multivariate analysis of variance** (and covariance) models. In the following, we briefly review univariate nonparametric procedures, and then present more extensively their multivariate extensions using chiefly the rank-permutation principle.

Precursors of Multivariate Nonparametrics

The classical **sign test** in the univariate case, followed by the **Wilcoxon signed-rank test**, laid down the foundation of distribution-free tests for the one-sample location model, and there were subsequent developments relating to alternative tests which are characterized by *local* or *asymptotic optimality* properties against specific parametric type of alternatives; Hájek & Šidák [13] is an excellent source for these theoretical developments. In the same vein, in the two-sample case, Mood's **median** test and the **Wilcoxon–Mann–Whitney** rank sum test are, respectively, the analogs of the sign and signed rank tests; the Brown & Mood [4] and Kruskal & Wallis [15] extensions cover the multisample problems and, later on, similar tests have also been proposed for the **analysis of variance** (ANOVA) as well as the **analysis of covariance** (ANCOVA) models. A general account of multivariate nonparametrics, with some mathematical abstractions, is given in Puri & Sen [19]. Interestingly, the developments in ANCOVA nonparametrics rest heavily on the multivariate nonparametric methodology, and hence, it will be to our advantage to outline the univariate tests first, then to motivate their multivariate analogs in a coherent manner, and finally to cover **general linear models** relating to multivariate AN(C)OVA problems. We shall also include the *two-way layout* problems covering both univariate and multivariate models, where such rank tests play an important role. The median test has some nice applications in the *multivariate association* problem, and we shall touch on that too.

Sign and Signed-Rank Statistics

Consider n observations $X_1, \ldots, X_n$ from a continuous distribution F, and let θ be the median of F, i.e. θ is the unique solution of the equation $F(x) = 0.5$. We want to test for a null hypothesis $\mathrm{H}_0 : \theta = \theta_0$ (known) against an alternative H_1 that θ is $>$ (or $<$ or $\neq$) θ_0. Without loss of generality, we set $\theta_0 = 0$ (otherwise, we work with the residuals $X_i - \theta_0$). Define the sign statistic as

$$S_n = \sum_{i=1}^{n} I(X_i \leq 0), \tag{1}$$

where $I(A)$ denotes the indicator function of the set A. Note that S_n has the simple **binomial distribution** with sample size n and probability $\pi = F(0)$, where, under H_0, $F(0) = 1/2$, irrespective of the functional form of F, and F need not be symmetric about its median. This provides a simple distribution-free test for H_0 vs. H_1, based on a one-sided or two-sided critical region for binomial distributions, depending on the alternative. Moreover, note that for every α : $0 < \alpha \leq 1/2$, and $n(\geq 1)$, there exists a nonnegative $r(\leq n/2)$, such that

$$2^{-n} \sum_{i=r+1}^{n-r-1} \binom{n}{i} \leq 1 - \alpha \leq 2^{-n} \sum_{i=r}^{n-r} \binom{n}{i}. \tag{2}$$

Then, if we denote the sample order statistics by $X_{n:1} < \cdots < X_{n:n}$ (where by virtue of the assumed continuity of F, the equality signs are neglected with probability 1), we have the following *distribution-free* **confidence interval** for the population median:

$$P_F\{X_{n:r} \leq \theta \leq X_{n:n-r+1}\} \geq 1 - \alpha, \quad \text{for all } F. \tag{3}$$

As a special case, letting $r = [(n+1)/2]$, we obtain that a nonparametric point estimator of θ is given by the *sample median*:

$$\tilde{X}_n = X_{n:(n+1)/2} \text{ or } \frac{X_{n:n/2} + X_{n:n/2+1}}{2}, \quad \text{according as } n \text{ is odd or even.} \tag{4}$$

In this development neither the symmetry nor the form of F is assumed as a part of the model, and we have a distribution-free test, as well as point and interval estimators. For large values of n, we could use the **convergence** of the binomial law to a normal one, and claim that under H_0, $n^{-1/2}[S_n - n/2]$ converges in law to a standard normal one, so that the critical values can be approximated by using tables for the standard normal distribution. For n up to 50 or so, exact binomial tables can be used, although for $n \geq 20$ a normal approximation works out well. For F possibly having discontinuities, ties among the observations may occur with a positive probability, and hence some adjustments for ties are to be made. Assuming F to be symmetric about its median, adjustment for ties can be made by *sign-inversions*, and this will be described later on in a general context.

Consider next the Wilcoxon signed rank statistic. Here we assume that the cumulative distribution function (cdf) F is symmetric about its median, and write $F(x) = F_\theta(x) = F(x - \theta)$, $x \in \mathbf{R}$, where $F(x) + F(-x) = 1$, for all $x \in \mathbf{R}$. Let $R_{ni}^+ = \sum_{j=1}^n I(|X_j| \leq |X_i|)$ be the **rank** of $|X_i|$ among the n observations $|X_i|$, $j = 1, \ldots, n$, for $i = 1, \ldots, n$. Also let $S_i = \text{sign}(X_i)$ be the sign of X_i, for $i = 1, \ldots, n$. Then define

$$W_n = (n+1)^{-1} \sum_{i=1}^n S_i R_{ni}^+. \tag{5}$$

Under the null hypothesis $H_0 : \theta = 0$, the two vectors $\mathbf{S}_n = (S_1, \ldots, S_n)'$ (of signs) and $\mathbf{R}_n^+ = (R_{n1}^+, \ldots, R_{nn}^+)'$ (of absolute ranks) are distributed independently, where $\mathbf{S}_n$ takes on each of the 2^n sign-inversions with the common probability 2^{-n}, and $\mathbf{R}_n^+$ takes on each permutation of $\{1, \ldots, n\}$ with the common probability $(n!)^{-1}$. Thus, W_n is distribution-free under H_0 and, furthermore,

$$E_0\{W_n\} = 0 \quad \text{and} \quad \text{var}_0\{W_n\} = \frac{n(2n+1)}{6(n+1)}. \tag{6}$$

The exact permutation distribution of W_n can well be approximated by a normal distribution when n is large and, surprisingly, the approximation is quite good for sample size as small as 10.

If we define $X_i(a) = X_i - a$, $i \geq 1$, $a \in \mathbf{R}$, and denote the Wilcoxon signed-rank statistic based on these aligned observations by $W_n(a)$, then it is easy to see that $W_n(a)$ is a nonincreasing (step-)function of a and, hence, virtually equating $W_n(a)$ to 0, we arrive at the following rank (R-)estimator of θ:

$$\hat{\theta}_n = \text{median } \left\{\tfrac{1}{2}(X_i + X_j) : 1 \leq i \leq j \leq n\right\}. \tag{7}$$

This estimator is consistent, median unbiased and asymptotically normal for all continuous and symmetric F. This is asymptotically optimal when F is a **logistic** cdf, about 95% **efficient** when F is normal, and is usually more efficient than the sample mean when F has a heavier tail than the normal law. This R-estimator has a bounded influence function and is globally **robust**. There are other signed-rank tests based on scores $a_n(R_{ni}^+)$ instead of the R_{ni}^+, where $a_n(\cdot)$ can be skillfully chosen to have asymptotic optimality against specific F (the normal scores correspond to a normal F); but they may not have a closed expression for the derived R-estimator. Moreover, whenever the score generating function is unbounded, we may have an unbounded influence function for the derived estimators, and hence, from robustness considerations, the Wilcoxon test and estimator may dominate the scenario.

Two-Sample Location Model

Suppose that we have two independent samples: $X_1, \ldots, X_{n_1}$ are independent and identically distributed random variables with a continuous cdf F, and $Y_1, \ldots, Y_{n_2}$ are independent and identically distributed random variables with a continuous cdf G, where both F and G are unknown, but not necessarily symmetric. In a two-sample location model, we set $G(x) = F(x - \theta)$, $x \in \mathbf{R}$, and treating F as a nuisance parameter (function), we like to test for the null hypothesis H_0 that $\theta = 0$ against θ positive (or negative or nonnull); and also, we like to estimate the *shift parameter* θ in a robust, nonparametric fashion. Note that in this setup, the null hypothesis actually relates to the homogeneity of F and G.

Median Statistics

We let $N = n_1 + n_2$ and denote the combined sample order statistics by $Z_{N:1} \leq \cdots \leq Z_{N:N}$ where, by virtue of the assumed continuity of F and G, ties among the observations (and hence the $Z_{N:i}$) can be neglected with probability 1. Also let us define

$$M = \frac{N+1}{2} \text{ or } \frac{N}{2}, \quad \text{according as } N \text{ is odd or even.} \tag{8}$$

Then we consider a **two-by-two table**; where m_1 and m_2 denote the number of observations of the first and second samples having values not exceeding $Z_{N:M}$, so that $n_1 - m_1$ and $n_2 - m_2$ are the complementary entries in this table. Under the null hypothesis $F = G$, all the N observations are independent and identically distributed, and hence their joint distribution remains invariant under any (of the $N!$) permutations of the coordinates. Using this (discrete) uniform permutation probability law, Mood [17] showed that

$$P\{m_1, m_2|H_0\} = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{N}{M}},$$
$$m_1 = 0, 1, \ldots, \min(M, n_1). \tag{9}$$

Both one-sided and two-sided tests, known as the Mood median test, can be made based on this **hypergeometric** probability distribution and a version of the statistic $M_n = \{m_2 - n_2 M/N\}$. The convergence of hypergeometric to normal laws paves the way for asymptotic normality of the test statistics (under H_0). Next, we note that if we replace the Y_i by $Y_i - a$, a real, and denote the resulting median statistic by $M_n(a)$, a real, then under the shift model $G(x) = F(x - \theta), \theta$ real, (i) $M_n(a)$ is a nonincreasing (step-)function of $a, a \in \mathbf{R}$, and (ii) $M_n(\theta)$ has the same distribution as M_n under H_0. This leads to the following estimator of θ:

$$\tilde{\theta}_N = \tilde{Y}_{n_2} - \tilde{X}_{n_1}, \tag{10}$$

where $\tilde{X}_{n_1}(\tilde{Y}_{n_2})$ is the sample median of the first (second) sample observations. $\tilde{\theta}_N$ is a robust, **consistent** and asymptotically normal estimator of θ. It is quite clear that both the sample medians are least sensitive to **outliers** or gross errors in the respective samples, and hence, this median estimator scores very high with respect to robustness perspectives; the conventional difference of the sample means, although optimal for the normal shift model, is highly nonrobust to such outliers or gross errors.

Rank Sum Statistics

We denote by $R_i(S_j)$ the rank of $X_i(Y_j)$ among the N combined sample observations, for $i = 1, \ldots, n_1 (j = 1, \ldots, n_2)$. Then the Wilcoxon–Mann–Whitney rank sum statistic can be expressed in the following form:

$$W_N = n_2^{-1} \sum_{j=1}^{n_2} S_j - n_1^{-1} \sum_{i=1}^{n_1} R_i. \tag{11}$$

An equivalent representation for W_N is a generalized ***U*-statistic**

$$U_{n_1,n_2} = (n_1 n_2)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_i \le Y_j), \tag{12}$$

which automatically adjusts for ties, if there are any, while in W_N, minor adjustments for ties are needed to define the R_i and S_j properly. Note that under $H_0 : F = G$, the same permutation law as in the case of the median test prevails, and hence W_N is genuinely distribution-free under H_0. Furthermore, $E[W_N|H_0] = 0$, $V[W_N|H_0] = N^2(N+1)/\{12 n_1 n_2\}$, and under H_0, the standardized form of W_N is asymptotically normal; this convergence holds quite well even when n_1 and n_2 are as small as 9 or 10. The asymptotic distribution of W_N is normal even when the null hypothesis is not true, although its mean and variance functions would depend on the cdfs F and G. The test based on W_n is consistent against a very broad class of alternatives that $\Pr\{X < Y\}$ is not equal to 1/2, and this includes, besides the shift alternative, the *stochastic ordering* of X and Y, defined in terms of an ordering of the cdfs F and G, without necessarily being of the shift type. An important alternative in this context is known as the **Lehmann alternative**: $\overline{G}(x) = 1 - G(x) = [\overline{F}(x)]^c$, for some c (> 0), and the null hypothesis relates to $c = 1$. In this case, the alternative hypothesis distribution of W_N is also independent of F and depends on the triplet (n_1, n_2, c).

Let us now replace the Y_i by $Y_i - a$, where a is real, and denote the resulting Wilcoxon statistic by $W_N(a)$, a real. Then it is easy to check that (i) $W_N(a)$ is a nonincreasing step-function of $a \in \mathbf{R}$, and (ii) under $G(x) = F(x - \theta)$, $W_N(\theta)$ has the same distribution as W_N has under H_0. Therefore, virtually equating $W_N(a)$ to 0, we arrive at the following R-estimator of the shift parameter θ:

$$\hat{\theta}_N = \tfrac{1}{2}\{\inf[a : W_N(a) < 0] + \sup[a : W_N(a) > 0]\}$$
$$= \text{median } [(Y_i - X_j) : 1 \le i \le n_2; 1 \le j \le n_1]. \tag{13}$$

This estimator is translation-invariant, median-unbiased, consistent, robust, and asymptotically normally distributed. This is asymptotically optimal

when F is a logistic cdf and, for a normal F, it is about 95% efficient. Again, this R-estimator has a bounded influence function and is globally robust. There are other R-estimators based on **logrank** or **normal scores** statistics which are to be solved by iterative methods, and which possess asymptotic optimality properties against some specific F (namely **exponential** or normal), but in terms of robustness may not be preferable to the Wilcoxon score estimator.

Several Sample Location Models

Now consider the case of c (≥ 2) independent samples of sizes $n_1, \ldots, n_c$ respectively, drawn from populations having continuous cdfs $F_1, \ldots, F_c$, where we set

$$F_j(x) = F(x - \theta_j), \quad x \in \mathbf{R}, j = 1, \ldots, c; \quad (14)$$

the θ_j are real (location) parameters, and F need not be symmetric. In testing the homogeneity of the F_j under this shift model, we really want to test for the identity of the θ_j. On the other hand, use of specific multisample rank statistics may lead to tests for the homogeneity of the F_j against alternatives that may be broader than such shift ones. Note that under H_0, all of the samples come from a common distribution, and hence their joint distribution remains invariant under any permutation of the coordinates over the entire combined set. This generates a discrete uniform probability measure, independent of F, so that tests based on this law are distribution-free under H_0.

Brown–Mood Median Statistic

We denote the jth sample observations by X_{ji}, $i = 1, \ldots, n_j$, for $j = 1, \ldots, c$, let $N = n_1 + \cdots + n_c$, and denote the combined sample order statistics by $Z_{N:1} \leq \cdots \leq Z_{N:N}$; by virtue of the assumed continuity of the F_j, ties are neglected with probability 1. Let M be a positive integer, typically, close to $(N+1)/2$, and let m_j be the number of X_{ji} in the jth sample with values $\leq Z_{N:M}$, for $j = 1, \ldots, c$. Note that $M = m_1 + \cdots + m_c$, and we have a $2 \times c$ **contingency table** with the first-row entries $m_1, \ldots, m_c$, marginal total M, second-row entries $n_1 - m_1, \ldots, n_c - m_c$ and total $N - M$, and the bottom marginal with entries $n_1, \ldots, n_c$ and total N. The usual $2 \times c$ contingency **chi-square test** statistic (along with the **Fisher exact testing** procedure) applies to this scenario [4]. The test is consistent against possible nonhomogeneity of the c population medians.

Kruskal–Wallis Statistic

We denote the rank of X_{ji} in the combined sample by R_{ji}, for $i = 1, \ldots, n_j$, $j = 1, \ldots, c$. Let then $R_{j\cdot} = \sum_{i=1}^{n_j} R_{ji}$, $j = 1, \ldots, c$. Then the Kruskal–Wallis [15] (rank sum) statistic can be written as

$$K_N = \frac{12}{N(N+1)} \sum_{j=1}^{c} n_j^{-1} \left[R_{j\cdot} - \frac{n_j(N+1)}{2} \right]^2 \quad (15)$$

(*see* **Nonparametric Methods**). The test based on K_n is distribution-free under H_0, and when the n_j are large, it has closely central **chi-square distribution** with $c - 1$ **degrees of freedom** (df). Consistency and efficiency properties run parallel to the two-sample case.

Simple Regression Models

Another extension of the two-sample model relates to the simple **linear regression** model [12] where $X_1, \ldots, X_N$ are independent random variables with continuous dfs $F_1, \ldots, F_N$ respectively, all defined on $\mathbf{R}$, and these F_j are linearly related in their arguments; namely,

$$F_j(x) = F(x - \beta c_j), \quad j = 1, \ldots, N, \quad (16)$$

where the c_j are known (regression) constants, not all equal, and β is an unknown regression parameter. By virtue of the translation invariance of ranks, we may absorb the intercept parameter in the unknown cdf F itself, and the null hypothesis of homogeneity of all the N cdfs reduced to that of $\beta = 0$. Define R_i as the rank of X_i among the N observations, for $i = 1, \ldots, N$, and let $\bar{c}_N = N^{-1} \sum_{i=1}^{N} c_i$. Then consider a *linear rank statistic*

$$\mathcal{L}_N = \sum_{i=1}^{N} (c_i - \bar{c}_N) a_N(R_i), \quad (17)$$

where $a_N(k)$, $k = 1, \ldots, N$ are suitable **scores**, not all equal. A statistic of the median type can be defined by letting

$$a_N(k) = \text{sign}\left(k - \frac{N+1}{2}\right), \quad k = 1, \ldots, N, \quad (18)$$

while a Wilcoxon-type statistic is based on scores

$$a_N(k) = \left(k - \frac{N+1}{2}\right), \quad \text{for } k = 1, \ldots, N. \tag{19}$$

These definitions extend readily to the **multiple regression** model, in which, for some $p \geq 1$, $\boldsymbol{\beta}$ is a p-vector and so are the $\mathbf{c}_j$.

Even in the simple regression model or the several sample case, there may be certain technical problems in defining suitable R-estimators (even based median or Wilcoxon-type scores) in a closed form. There is another statistic, known as the Kendall tau statistic (*see* **Association, Measures of**), which is a mixture of median and Wilcoxon scores, and which may have some advantages. We define this as

$$T_N = \sum_{i=1}^{N} \sum_{j=1}^{N} \text{sign}(c_i - c_j)\ \text{sign}(X_i - X_j). \tag{20}$$

The double summation in the above formula can be replaced by a summation over all $\{1 \leq i < j \leq N\}$. It follows from Sen [21] that

$$\hat{\beta}_n = \text{median}\left[\frac{X_i - X_j}{c_i - c_j} : \{i, j : c_i \neq c_j\}\right] \tag{21}$$

is a robust, consistent, median-unbiased and asymptotically normal estimator of β, and that in the particular case of binary c_i (i.e. the two-sample location model), this estimator reduces to the Wilcoxon scores estimator $\hat{\theta}_N$, considered in (13). A distribution-free confidence interval for β (and hence θ) can also be obtained by using the Kendall tau statistic based on the aligned observations $X_i - bc_i$, $i \geq 1$, b real.

Two-Way Layouts

The method of *n-rankings* and *ranking after alignment* are the two popular nonparametric methods for statistical modeling and analysis of two-way layouts, covering complete **randomized block** as well as **incomplete block** designs. Both the median and rank sum procedures are popular in this context. The method of n-rankings is presented here, while the other method depending on a multivariate approach will be presented later. In that way, the impact of multivariate nonparametrics will be clearer.

Consider $n(\geq 2)$ blocks of $p(\geq 2)$ plots receiving p different treatments. Let X_{ij} be the response of the jth treatment in the ith block, for $j = 1, \ldots, p$; $i = 1, \ldots, n$. Let r_{ij} be the rank of X_{ij} within the ith block (p) observations, for $j = 1, \ldots, p$, and let $\mathbf{r}_i = (r_{i1}, \ldots, r_{ip})'$, $i = 1, \ldots, n$. The method of n-rankings is based on these n rank vectors. In fact, this does not even require that all the X_{ij} are observable; it suffices to have the realizations of the within-block ranks. This situation may arise, for example, when n judges are asked to rank, independently of each other, p objects (e.g. players), and the judgment may involve some subjective elements too. Even when the X_{ij} are observable, we do not need to assume that they come from homoscedastic normal distributions, and the treatment or block effects may not be additive. The null hypothesis H_0 of interest is the *interchangeability* of the observations within each block, so that under H_0, the $\mathbf{r}_i$ are independent and identically distributed random vectors, and each $\mathbf{r}_i$ takes on each permutation of $\{1, \ldots, p\}$ with the common probability $(p!)^{-1}$. (The adjustments for tied ranks can be made easily.) Two popular test statistics based on the method of n-ranking are the following:

1. Friedman's [10] rank sum statistic

$$\chi_r^2 = \frac{12n}{p(p+1)} \sum_{j=1}^{p} \left(\overline{r}_{j,n} - \frac{p+1}{2}\right)^2, \tag{22}$$

where $\overline{r}_{j,n} = n^{-1} \sum i = 1^n r_{ij}$, $j = 1, \ldots, p$; and

2. Brown & Mood's [4] median statistic:

$$B_r = \frac{np(p-1)}{a(p-a)} \sum_{j=1}^{p} \left(\overline{M}_{j,n} - \frac{a}{p}\right)^2, \tag{23}$$

where a is a positive integer, typically close to $(p+1)/2$, and $\overline{M}_{j,n} = n^{-1} \sum_{i=1}^{n} I(r_{ij} \leq a)$, $j = 1, \ldots, p$.

For both the statistics, null distributions are generated by the $(p!)^n$ equally likely realizations of $(\mathbf{r}_1, \ldots, \mathbf{r}_n)$, and hence they are distribution-free under H_0. Both are consistent against alternatives that are more general than the conventional ANOVA differences in the treatment effects. Moreover, for large values of n, the exact null distribution of either statistic converges to central chi-squared distribution df with $p - 1$ df. Asymptotic properties of such tests have been studied in detail in Puri & Sen [19, Chapter 7]. Extensions to incomplete block designs and more than one observation per cell have also been presented there.

Bivariate Independence Problem

This bivariate problem is essentially reducible to a quasi-univariate one, and both rank sum and median-type statistics are popular in this respect. Let $(X_i, Y_i), i = 1, \ldots, n$ be n independent and identically distributed random vectors having a continuous bivariate cdf $F(x, y)$, $(x, y) \in \mathbf{R}^2$. The hypothesis of stochastic independence of X and Y can be stated as

$$\mathrm{H}_0 : F(x, y) = F(x, \infty) \cdot F(\infty, y), \quad \text{for all} (x, y) \in \mathbf{R}^2. \tag{24}$$

Recall that independence implies uncorrelation, but the converse may not be true (without normality of F) (*see* **Correlation**). Moreover, alternatives to H_0 may be of diverse types, and naturally different test statistics may behave differently under such alternatives. One of the simple and appealing classes of alternatives is the *positive (negative) dependence* which specifies that $F(x, y) \geq (\leq) F(x, \infty) F(\infty, y)$, for all x, y, with strict inequality at least on a set of positive measure. Recall that $F_1(x) = F(x, \infty)$ and $F_2(y) = F(\infty, y)$ are the two marginal cdfs, and as a suitable measure of dependence, known as the *grade correlation* coefficient, we have the following

$$\rho_g = 12 \int \int \left[F_1(x) - \tfrac{1}{2}\right]\left[F_2(y) - \tfrac{1}{2}\right] \mathrm{d}F(x, y). \tag{25}$$

If $R_i (S_i)$ denotes the rank of $X_i (Y_i)$ among the X's (Y's), for $i = 1, \ldots, n$, then the sample counterpart of ρ_g is

$$r_{g,n} = \frac{12}{n(n^2 - 1)} \sum_{i=1}^{n} \left(R_i - \frac{n+1}{2}\right) \times \left(S_i - \frac{n+1}{2}\right), \tag{26}$$

and this is known as the **Spearman rank correlation** coefficient. A related measure of dependence is the *quadrant measure*

$$\rho_q = 4[P(X > \theta_1, Y > \theta_2) - 1], \tag{27}$$

where θ_1 and θ_2 are the population medians of X and Y. The sample counterpart of this is the following

$$Q_n = n^{-1} \sum_{i=1}^{n} \operatorname{sign}\left(R_i - \frac{n+1}{2}\right) \times \operatorname{sign}\left(S_i - \frac{n+1}{2}\right), \tag{28}$$

and is known as the quadrant test statistic. A related measure, the Kendall tau statistic, can be expressed as

$$K_n = \binom{n}{2}^{-1} \sum_{\{1 \leq i < j \leq n\}} \operatorname{sign}(R_i - R_j) \operatorname{sign}(S_i - S_j). \tag{29}$$

In each case, under H_0, the two rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ are independently distributed with a discrete uniform distribution over the permutations of $\{1, \ldots, n\}$, and hence these tests are all distribution-free under H_0. Various properties of these tests have been studied in detail in Hájek & Šidák [13] and in other contemporary nonparametric texts.

The First Spark of Multivariate Nonparametrics

The simplicity of rank-based statistical procedures in the univariate cases mentioned above stumbles into roadblocks in the bivariate or genuine multivariate cases; this impasse is primarily due to the fact that, in a bivariate or multivariate case, neither the distribution of the vector of coordinatewise signs nor the coordinatewise ranks (or absolute ranks) are generally independent of the underlying distribution, even when suitable hypotheses of invariance (similar to the univariate cases) hold. For example, suppose that we have a sample of n observations from a bivariate distribution, and we want to test the null hypothesis that both the marginal medians are equal to 0. Thus, a simple extension of the sign test would be a suitable function of the two coordinatewise sign statistics. However, the two sign-functions sign (X_i) and sign (Y_i) for an observation (X_i, Y_i) are not necessarily independent, and hence a bivariate sign test may not be genuinely distribution-free. Chatterjee [5] eliminated this drawback by an appeal to a conditional procedure based on a partitioning of the set of observations into *concordant* and *discordant* ones,

and showed that such conditional tests are conditionally distribution-free and have nice properties too. Earlier, Chatterjee & Sen [6] considered the bivariate two-sample (location/association) problem, and demonstrated that a simple *rank-permutation principle* renders conditional distribution-freeness of a large class of rank statistics, including both the median and rank sum tests. Subsequent developments in multivariate nonparametrics, reported in Puri & Sen [19], exploit this Chatterjee–Sen rank-permutation principle as extended to more complex setups, and here we shall confine ourselves to specific problems in this domain.

Multivariate Multisample Tests

Let $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(p)})'$, $j = 1, \ldots, n_i$ be n_1 independent and identically distributed random vectors having a continuous p-variate cdf $F_i(\mathbf{x})$, $\mathbf{x} \in \mathbf{R}^p$, for $i = 1, \ldots, c\,(\geq 2)$. We want to test for the null hypothesis $\mathrm{H}_0 : F_1 = \cdots = F_c = F$ (unknown), against plausible alternatives that these cdfs differ in location/scale and or association measures. We shall mainly discuss the multivariate rank sum and median procedures, studied in detail by Chatterjee & Sen [6–8] and reported in more mathematical abstraction in Puri & Sen [19]). Let $R_{ij}^{(k)}$ be the rank of $X_{ij}^{(k)}$ among the $N(= \sum_{1=1}^{c} n_i)$ observations on the kth characteristic, for $j = 1, \ldots, n_i; / i = 1, \ldots, c; k = 1, \ldots, p$. Define the *rank collection matrix* $\mathbf{R}_N$ as the $p \times N$ matrix the kth row of which contains the elements $(R_{11}^{(k)}, \ldots, R_{1n_1}^{(k)}, \ldots, R_{cn_c}^{(k)})$, for $k = 1, \ldots, p$. Thus, each row of $\mathbf{R}_N$ contains the number $\{1, \ldots, N\}$, permuted in some order, and there are $(N!)^p$ such possible realizations of $\mathbf{R}_N$. Multisample multivariate rank statistics are typically based on this rank-collection matrix, so that the distribution theory of $\mathbf{R}_N$ governs the same for such coordinatewise rank-based statistics. For $p = 1$ – that is a single characteristic – under H_0, all possible $N!$ realizations of the rank vector are equally probable. However, for $p \geq 2$, unless the p coordinates of the $\mathbf{X}_{ij}$ are independent (wherein we have only a quasi-multivariate setup), the distribution of the rank-collection matrix depends on the underlying $F_1, \ldots, F_c$, even under the null hypothesis. This characterizes the lack of distribution-freeness of multivariate multisample rank-based procedures. The Chatterjee–Sen rank permutation principle removes this impasse through a conditional approach which is easy to interpret and implement in practice. Let us permute the columns of $\mathbf{R}_N$ in such a way that the top row is in the natural order $(1, \ldots, N)$. We denote the resulting element in the kth row and rth column by $R_{Nr}^{(k)*}$, for $r = 1, \ldots, N$; $k = 2, \ldots, p$; note that $R_{Nr}^{(1)*} = r$, $1 \leq r \leq N$. We denote the derived $p \times N$ matrix by $\mathbf{R}_N^*$ and term it the *reduced rank collection* matrix. Note that the cdf of $\mathbf{R}_N^*$ depends, even under H_0, on the underlying F, but the conditional distribution of $\mathbf{R}_N$, given $\mathbf{R}_N^*$, under H_0, is independent of the underlying F, and is uniform (discrete) over the $N!$ permutations of the columns of $\mathbf{R}_N^*$. We denote this permutational (conditional) probability measure by $\mathcal{P}_N$, and advocate the use of the same in the construction of some permutationally (conditionally) distribution-free tests.

As in the univariate case, we consider the individual sample coordinatewise rank sum statistics:

$$\overline{R}_i^{(k)} = n_i^{-1} \sum_{j=1}^{n_i} R_{ij}^{(k)}, \quad k = 1, \ldots, p; \quad i = 1, \ldots, c. \tag{30}$$

Let us also define the *rank covariance* matrix $\mathbf{V}_N = ((v_{N,kq}))$ (of order $p \times p$) by

$$\mathbf{V}_N = N^{-1}(\mathbf{R}_N)(\mathbf{R}_N)' - \left(\frac{N+1}{2}\right)^2 \mathbf{1}\mathbf{1}', \tag{31}$$

and denote by $\mathbf{V}_N^-$ a generalized inverse of $\mathbf{V}_N$; under very mild regularity conditions, $\mathbf{V}_N$ is positive definite in probability, and hence, we may as well work with the actual inverse $\mathbf{V}_N^{-1} = ((v_N^{kq}))$. It is easy to verify that under $\mathcal{P}_N$, the $\overline{R}_i^{(k)}$ have all expected value $(N+1)/2$, and, furthermore,

$$\mathrm{cov}\left[\overline{R}_i^{(k)}, \overline{R}_j^{(q)} | P_N\right] = \frac{N\delta_{ij} - n_i}{n_i(N-1)} v_{N,kq}, \tag{32}$$

for $k, q = 1, \ldots, p$; $i, j = 1, \ldots, c$ where δ_{ij}, the Kronecker delta, is 1 or 0 according as $i = j$ or not. Then the multivariate generalization of the Kruskal–Wallis [15] rank sum statistic is given by

$$\mathcal{L}_N = \sum_{i=1}^{c} n_i \sum_{k=1}^{p} \sum_{q=1}^{p} v_N^{kq} \left[\overline{R}_i^{(k)} - \frac{N+1}{2}\right] \times \left[\overline{R}_i^{(q)} - \frac{N+1}{2}\right]. \tag{33}$$

The conditional (permutational) distribution of $\mathcal{L}_N$, given $\mathbf{R}_N^*$, can be obtained by enumeration of all possible $N!$ conditionally equally likely realizations of $\mathbf{R}_N$, and this provides a conditionally distribution-free test for H_0. The task becomes prohibitively laborious when N becomes large. However, it has been shown that the permutation distribution of $\mathcal{L}_N$ converges (in probability) to the central chi-square distribution with $p(c-1)$ df when N is large. The test statistic reduces to the Kruskal–Wallis statistic when $p = 1$. For $c = 2$, this corresponds to the Wilcoxon–Mann–Whitney statistics [16] and it is consistent against the same class of alternatives as in univariate case, but simultaneously for all the coordinates. The test is robust against nonnormality of F as well as for error contaminations, although unlike the parametric tests based on the characteristic roots of the between-sample sum of product matrix, normalized by the pooled within-sample sum of product matrix, it is not invariant to affine transformations on the original observations. This lack of affine-invariance is, of course, shared by most of the tests based on coordinatewise rank vectors (which are themselves not affine invariant). Also, unlike the univariate case, here tables for the exact null distribution of $\mathcal{L}_N$ are difficult to construct, as they generally depend on the reduced rank-collection matrix $\mathbf{R}_N^*$ and there are $(N!)^{p-1}$ such matrices which are themselves not necessarily equally likely, even conditionally. Nevertheless, the conditional distribution-freeness, robustness properties, simple asymptotic distribution theory, and good power properties make this multivariate rank sum test a good competitor of the classical parametric tests.

Let us present the multivariate multisample median tests side by side. As in the univariate case, we consider the coordinatewise median statistics for each sample relative to the pooled one. Let M be a positive integer, close to $(N+1)/2$, and let

$$M_i^{(k)} = \sum_{j=1}^{n_i} I(R_{ij}^{(k)} \le M), \quad k = 1, \ldots, p; \quad i = 1, \ldots, c. \tag{34}$$

We also introduce the $p \times p$ matrix $\mathbf{V}_N^0 = ((v_{N,kq}^0))$ by letting $v_{N,kk}^0 = M(N-M)/N^2, k = 1, \ldots, p$, and for $k \ne q$, defining the reduced rank collection matrix $\mathbf{R}_N^*$ as before,

$$v_{N,kq}^0 = N^{-1} \sum_{r=1}^{N} I(R_r^{(k)*} \le M, R_r^{(q)*} \le M) - \left(\frac{M}{N}\right)^2, \quad k \ne q = 1, \ldots, p. \tag{35}$$

Again under very mild regularity conditions, $\mathbf{V}_N^0$ is positive definite in probability, and we therefore assume that its inverse $(\mathbf{V}_N^0)^{-1} = ((v_N^{kq,0}))$ is well defined (otherwise, we work with a generalized inverse). Then we consider the following three types of median tests. These tests are very robust against error contaminations or outliers, but in general, for nearly multinormal distributions, they may not be asymptotically as efficient as the multivariate rank sum test, although in terms of robustness we have an opposite relative picture.

1. *Omnibus median procedure*. Recall that by letting $M_i^{(k)*} = N - M_i^{(k)}$, for $k = 1, \ldots, p, i = 1, \ldots, c$, we arrive at p sets of $2 \times c$ contingency table with the cell entries $M_i^{(k)}, M_i^{(k)*}, i = 1, \ldots, c$, for each $k = 1, \ldots, p$. Unfortunately, these tables are not necessarily stochastically independent, and hence we need to appeal to higher-dimensional categorical data models and analysis schemes to formulate appropriate test statistics; we refer to Chatterjee & Sen [6, 7] for further details. For $p \ge 2$, such a test is consistent not only against possible heterogeneity of location parameters (or medians) of the c populations, but also against possible heterogeneity of their association parameters of various orders. Note that the total df for such a complex categorical data model is much higher than $p(c-1)$, so that such an omnibus test statistic carries too many degrees of freedom. Hence, for location/shift alternatives, it may not be efficient. For this reason, we shall find it more convenient to deal more explicitly with the other two types of median procedures, which are geared toward such location alternatives and/or pairwise association alternatives only, and are generally more efficient than an omnibus procedure.

2. *Multisample multivariate median test for location*. This procedure is directed for location alternatives, and the test statistic is given by

$$\mathcal{L}_N^0 = \sum_{i=1}^{c} n_i \sum_{k=1}^{p}\sum_{q=1}^{p} v_N^{kq,0} \left(n_i^{-1} M_i^{(k)} - \frac{M}{N}\right) \times \left(n_i^{-1} M_i^{(q)} - \frac{m}{N}\right), \tag{36}$$

so that, for $p = 1$, it reduces to Brown & Mood's test statistic considered earlier. As in the case of the multivariate multisample rank sum test, for small values of $n_1, \ldots, n_c$, one can appeal to the permutational probability measure $\mathcal{P}_N$, and obtain a conditionally (permutationally) distribution-free test for the null hypothesis of homogeneity of the cdfs $F_1, \ldots, F_c$; for large sample sizes, the permutational distribution of $\mathcal{L}_N$ converges, in probability, to the central chi-square distribution with $p(c-1)$ df. Thus, for large sample sizes, the critical level of this median test can be well approximated by appropriate quantiles of this simple limiting distribution.

3. *Median tests for homogeneity of association parameters*. In addition to the p measures of marginal locations, there are $\binom{p}{2}$ measures of pairwise associations of the p variates, and similarly for higher-order associations when $p \geq 2$. In many situations, particularly when p is not so small, inclusion of all such measures results in a large-dimensional parameter space, and that in turn pulls down the level of precision that can be acquired from a given data set. For this reason, often, multivariate models are conceived wherein only first-order or pairwise association measures are included along with the location measures, so that essentially the number of parameters for each cdf is reduced to

$$p + \binom{p}{2} = \binom{p+1}{2},$$

and in the given multisample model we would therefore have

$$(c-1)\binom{p+1}{2}$$

degrees of freedom. Tests for this broader (than location) type of alternatives can be based on the collection of the entries $\{M_i^{(kq)} = \sum_{j=1}^{n_i} I(R_{ij}^{(k)} \leq M, R_{ij}^{(q)} \leq M), \quad k, q = 1, \ldots, p; i = 1, \ldots, c\}$. Moreover, such a test statistic can be decomposed into two orthogonal components, $\mathcal{L}_N^0$ described before, for location alternatives, and the complementary part for association alternatives. For further details, we refer to Chatterjee & Sen [7].

There is a basic difference in the **asymptotic relative efficiency** (ARE) picture between the univariate and multivariate cases. In both cases, for local (Pitman-type) alternatives the test statistics have appropriate noncentral chi-square distributions with comparable degrees of freedom and noncentrality parameters. However, in the univariate case, for two such test statistics under a common local alternative their noncentrality parameters are proportional to each other over the entire parameter space, so that the ratio of noncentrality parameters provides a meaningful measure of the (Pitman) ARE of one test with respect to the other (*see* **Pitman Efficiency**). In the multivariate case, the ratio of the two noncentrality parameters may depend not only on the underlying distributions through appropriate dispersion matrices that appear in their expressions, but also on the actual alternatives (namely, the direction cosines in the location case). Therefore a single measure of ARE may not be tenable for the entire parameter space under (local) alternatives, and hence one of the basic interpretations of the ARE of such tests in terms of the sample sizes needed to have equal (asymptotic) power may no longer be possible. This drawback has been minimized to a certain extent by using suitable lower and upper bounds for the ratio of noncentrality parameters, and then incorporating the univariate concepts on such bounds. However, that may not convey a definitive picture in all situations. We refer to Puri & Sen [19, Chapters 4– 5] for a comprehensive discussion of ARE of statistical tests in the multivariate case.

In the univariate case, estimators of difference of location parameters based on rank tests, particularly the median and rank sum statistics, have already been discussed earlier. These estimators extend to the multivariate case in a straightforward manner, as for each of the p coordinates, the same univariate estimation procedure can be adopted. The basic difference comes

in the interval estimation problem as extended to confidence sets in the multivariate case. The distribution-freeness of the confidence intervals in the univariate case may no longer hold in the multivariate case (as the coordinatewise rank statistics are generally not independent). However, asymptotic confidence sets may still be achieved by using the Scheffé or Tukey method along with the asymptotic chi-square distributions of related test statistics (under H_0). These asymptotic procedures are generally more robust than the classical parametric procedures, though for small sample sizes, there may not be a good resolution. We refer to Puri & Sen [19, Chapter 6] for details.

Multivariate One-Sample Rank Tests

We now present rank statistics which are the multivariate extensions of the sign and Wilcoxon signed-rank statistics in the univariate case, treated earlier. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$, $i = 1, \ldots, n$, be independent and identically distributed random vectors having a $p(\geq 1)$ variate continuous cdf, F. We denote the marginal cdfs corresponding to F by $F_{[1]}, \ldots, F_{[p]}$ respectively, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ be the vector of the marginal medians. Suppose that we want to test for the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ (specified) against alternatives that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$; without any loss of generality, we let $\boldsymbol{\theta}_0 = \mathbf{0}$. Different rank tests for this hypothesis testing problem entail different regularity conditions on the cdf. First, we consider multivariate sign tests entailing least restrictive conditions in this respect.

To motivate the scenario, we start with the bivariate sign tests, and then present the general multivariate case. Hodges [14] considered an association invariant and genuinely distribution-free sign test, although its distributions on null or alternative hypotheses are not very simple. Another association invariant bivariate sign test is due to Blumen [3], and this also shares the drawbacks of the Hodges' sign test to a certain extent. Bennett [1] used the classical likelihood ratio principle to arrive at an asymptotically distribution-free sign test. Chatterjee's [5] treatment include a conditionally distribution-free and strictly unbiased bivariate sign test, basically related to the Bennett version when the observations are iid, a condition not needed in the Chatterjee case. If we create a two-by-two table by drawing perpendicular axes through the median vector ($\mathbf{0}$ under H_0), the positive (negative) orthant $\mathbf{X} \geq (\leq)\, \mathbf{0}$ constitutes the first (second) type of *concordance* set, while the second and fourth quadrants relate to the first and second type of *discordance* sets. Although Chatterjee [5] considered independent but not necessarily identically distributed random variables, for simplicity of presentation, we consider here the independent and identically distributed case. Let $\Pr(C)$ and $\Pr(D) = 1 - \Pr(C)$ be the probabilities of concordance and discordance, respectively. Furthermore, let

$$\tau = \frac{\Pr\{\text{first type concordance}\}}{\Pr\{\text{concordance}\}} \quad \text{and}$$
$$\gamma = \frac{\Pr\{\text{first type discordance}\}}{\Pr\{\text{discordance}\}} \tag{37}$$

Then the null hypothesis of $\boldsymbol{\theta} = \mathbf{0}$ can be equivalently stated as

$$H_0^* : \tau = \gamma = \tfrac{1}{2}. \tag{38}$$

In the sample, let C_1, C_2, D_1, and D_2 be the number of observations which are concordant of the first and second types and discordant of the first and second type respectively. Note that $n = C_1 + C_2 + D_1 + D_2$. Under H_0, given $C = c$, we have

$$\Pr\{C_1 = c_1, D_1 = d_1 | C = c\} = \binom{c}{c_1} \times \binom{n-c}{d_1} 2^{-n}, 0 \leq c_1 \leq c, 0 \leq d_1 \leq n - c, \tag{39}$$

which is a product of two independent binomial distributions, and is independent of the underlying df. Therefore, a test based on this conditional law is conditionally distribution-free. Chatterjee [5] considered the test statistic

$$T_n = 4\left[C^{-1}\left(C_1 - \frac{C}{2}\right)^2 + (n-C)^{-1}\left(D_1 - \frac{n-C}{2}\right)^2\right], \tag{40}$$

and constructed a randomized test to achieve unbiasedness; he showed that, under H_0, the conditional distribution of T_n converges (in probability) to the central chi-square distribution with 2 df when n is large. With a view to presenting the general multivariate case in the same vein, we rewrite T_n in the form

$T_n = n\mathbf{S}_n'\mathbf{V}_n^{-1}\mathbf{S}_n$, where $\mathbf{S}_n' = n^{-1}\sum_{i=1}^n$ (sign (X_{i1}), sign (X_{i2})), and $\mathbf{V}_n$ has the diagonal elements equal to 1 and the off-diagonal elements equal to $n^{-1}\sum_{i=1}^n$ sign $(X_{i1}X_{i2})$. Adjustment for ties (at 0) can be made easily under this conditional setup. In this setup, we now proceed to the general p variate case, and define

$$\mathbf{S}_n = n^{-1}\sum_{i=1}^n [\text{sign}(X_{i1}), \ldots, \text{sign}(X_{ip})]',$$

$$\mathbf{V}_n = ((v_{nkq})) = \left(\left(n^{-1}\sum_{i=1}^n \text{sign}(X_{ik}X_{iq})\right)\right)_{k,q=1,\ldots,p},$$

$$\mathcal{T}_n = n\mathbf{S}_n'\mathbf{V}_n^{-1}\mathbf{S}_n, \tag{41}$$

where $\mathbf{V}_n^{-1}$ is a (generalized) inverse of $\mathbf{V}_n$. The concordance–discordance picture in the bivariate case extends to a reflection picture (over 2^p possible cells) in the multivariate case, and generates a similar conditional distribution as a product of 2^{p-1} binomial distributions. This provides the genesis of conditionally distribution-free sign tests in the multivariate case. Again, the conditional null distribution of $\mathcal{T}_n$ converge (in probability) to the central chi-square distribution with p df, so that critical levels may well be approximated by chi-square percentile points. The asymptotic power and efficiency properties of such sign tests have been presented in Puri & Sen [19, Chapter 4]. This test exploits only the reflection property of the vector of signs of the observations (under H_0). There are some other sign tests, as would be briefly introduced later on, that are based on more stringent group-invariance structures, and are appropriate only under some additional regularity assumptions on the underlying distributions that validate such invariance properties.

We consider next the multivariate signed-rank statistics [26] and exhibit their (conditional) distribution-freeness under suitable groups of transformations. As in the univariate case, we consider for each coordinate the vector of signs and absolute ranks. Also, we need to assume some sort of symmetry of the cdf F, which at least implies that all the p marginal $F_{[1]}, \ldots, F_{[p]}$ are symmetric. For the rank-permutation principle to work out, we need a more stringent symmetry-condition that the cdf is *diagonally symmetric*, implying that both $\mathbf{X} - \boldsymbol{\theta}$ and $\boldsymbol{\theta} - \mathbf{X}$ have the same distribution. In the asymptotic case, this may be relaxed to marginal symmetry.

Let R_{ij}^+ be the rank of $|X_{ij}|$ among the n observations $|X_{rj}|, r = 1, \ldots, n, i = 1, \ldots, n$, for $j = 1, \ldots, p$, and let

$$T_{nj} = n^{-1}\sum_{i=1}^n \text{sign}(X_{ij})R_{ij}^+, \quad j = 1, \ldots, p;$$

$$v_{njk} = n^{-1}\sum_{i=1}^n \text{sign}(X_{ij}X_{ik})R_{ij}^+R_{ik}^+, \quad j,k = 1, \ldots, p;$$

$$\mathbf{T}_n = (T_{n1}, \ldots, T_{np})',$$

$$\mathbf{V}_n = ((v_{njk}))_{j,k=1,\ldots,p}. \tag{42}$$

Then the multivariate signed-rank statistic is

$$\mathcal{L}_n = n\mathbf{T}_n'\mathbf{V}_n^{-1}\mathbf{T}_n. \tag{43}$$

Conditional on the rank collection matrix $\mathbf{R}_n^+$ and the collection of the n sign vectors, under H_0, the permutation (conditional) distribution of $\mathcal{L}_n$ is generated by the 2^n conditionally equally likely sign-inversions, and hence, it is a conditionally distribution-free test. For large values of n, the permutation distribution of $\mathcal{L}_n$ converges (in probability) to the central chi-square distribution with p df, so simple approximations to the critical levels can be obtained from the percentile points of this limit law. The test is consistent against the same alternative as in the univariate case, extended coordinatewise to the multivariate case. The asymptotic properties (including ARE) of this signed rank tests are essentially the same as in the case of the multisample rank sum test, and we refer to Puri & Sen [19, Chapter 4] for details. As in the multisample model, here also, estimates of the coordinatewise location parameters can be based on the coordinatewise sign or signed-rank statistics, so that there are closed expressions for them. The situation with confidence sets is somewhat more complex (as the distribution-freeness does not hold in the multivariate case); nevertheless, asymptotic solutions based on the Scheffé method work out well; we refer to Puri & Sen [19, Chapter 6] and Bickel [2] for details.

The sign and signed-rank tests are not *rotationally* or *affine-invariant*, although in the bivariate case, the Hodges [14] sign test has some invariance property. Incorporating the idea of Oja's *simplex median* [18], some rotation invariant bivariate sign tests have been developed in the recent past, and its multivariate generalizations are nicely wrapped up

by Chaudhuri & Sengupta [9], who also extended the Hodges sign test to the multivariate case, and also cited other pertinent references. While for rotationally invariant rank tests the sphericity of the underlying distribution is needed, for affine invariance elliptical symmetry, less restrictive than the sphericity, suffices, but then the resulting tests are only asymptotically distribution-free (as opposed to the conditionally distribution-freeness of the tests presented earlier). In biostatistical applications, often, the coordinate variables may relate to different types of responses or characteristics which may not be linearly compoundable, and hence the assumption of sphericity or elliptical symmetry may run contrary to the experimental setup – and hence such invariant tests having mostly mathematical appeals for large sample sizes may not be of much use. Incidentally, the permutation version of the **Hotelling T^2 test**, considered by Wald & Wolfowitz [27] more than 50 years ago, possesses the affine-invariance property, is computationally simpler, and asymptotically optimal too (when the underlying df are multinormal). On the other hand, such affine-invariant sign or signed rank tests share the nonrobustness properties of the T^2 statistics (to outliers or error contaminations as well as nonnormality). In a multivariate setup, such model departures can occur in many more ways than in an univariate model, and hence should be carefully examined before deciding on a suitable test statistic, solely on its asymptotic optimality against a specific distributional alternative.

Multivariate Paired Comparisons Tests

A natural extension of the sign test relates to **paired comparison** studies, in which $t(\geq 2)$ objects are considered in pairs $(i, j) : 1 \leq i < j \leq t$, and judged with respect to some performance characteristic(s) on a relative basis for each pair. In the case of multiple characteristics (say, p), for each pair of objects, we have a p-vector $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(p)})'$, where $X_{ij}^{(k)}$ takes on the value $+1$ or -1 according as the ith object is judged better (or not) than the jth one with respect to the kth trait, for $k = 1, \ldots, p; 1 \leq i < j \leq t$. Therefore, we are given $\binom{t}{2}$ sets of multivariate sign vectors, where for each such table we have generally multiple observations. It is possible to use a conventional (multidimensional) contingency table to analyze such paired comparisons experiments. However, the degrees of freedom for such tests would be considerably larger when t is not small. The practice is to combine these subsets in such a way that the object contrasts are highlighted with a considerable reduction of the df to enhance the power. Since we have a multivariate situation, genuinely distribution-free procedures may not be available. Sen & David [25] incorporated the Chatterjee [5] sign-inversion principle, and derived a simple paired comparison test for the homogeneity of the t objects in the bivariate case. This test statistic is conditionally (permutationally) distribution-free and has $2(t-1)$ df [compared with $t(t-1)$ in the classical contingency table approach]. An extension of this paired comparison test in the general multivariate case has been considered by Sen [24], and it contains an unifying account of the developments in this area.

Multivariate Friedman χ_r^2 Test

A direct extension of the Friedman χ_r^2 test to the multivariate case is due to Gerig [11]. He incorporated the Chatterjee–Sen rank-permutation principle to construct the permutational dispersion matrix of the individual treatment rank sums $(\sum_{i=1}^{n} r_{ij}^{(s)} = p\bar{r}_{n,j}^{(s)},\ j = 1, \ldots, p,$ for each coordinate $(s = 1, \ldots, q)$ separately, and incorporated this in the construction of a quadratic norm:

$$\mathcal{L}_n = n \sum_{s=1}^{m} \sum_{s'=1}^{m} v_n^{ss'} \sum_{j=1}^{p} \left[\bar{r}_{n,j}^{(s)} - \frac{p+1}{2} \right] \times \left[\bar{r}_{n,j}^{(s')} - \frac{p+1}{2} \right], \tag{44}$$

where the $v_{n,ss}$ are all equal to $p(p+1)/12$, while

$$v_{n,ss'} = \frac{1}{n(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[r_{ij}^{(s)} - \frac{p+1}{2} \right] \times \left[r_{ij}^{(s')} - \frac{p+1}{2} \right], \quad s \neq s' = 1, \ldots, q. \tag{45}$$

Again, this test is permutationally (conditionally) distribution-free and under the null hypothesis of homogeneity of the treatment vectors, $\mathcal{L}_n$ has asymptotically the central chi-square distribution with $q(p-1)$ df. For detailed study of the asymptotic properties of this test, we refer the reader to Puri & Sen [19, Chapter 7].

Aligned Rank Tests

An inherent problem of the method of n-ranking, be it in the univariate or multivariate case, is that it does not properly incorporate *interblock information* (as the within-block rankings are independent for different blocks). Thus, whenever the block effects are additive, it seems that if these are estimated, and *alignment* is made by subtracting these estimated block-effects from the respective block observations, then these aligned observations can be ranked across the entire set of blocks, permitting interblock comparisons to a greater extent. Rank tests based on such aligned observations are termed "aligned rank tests". Again, the process of alignment generally distorts the exact distribution-freeness of such aligned rank tests. Nevertheless, the Chatterjee–Sen rank-permutation principle can be adopted to show that such aligned rank tests are permutationally (conditionally) distribution-free. We denote the aligned rank of the jth observation in the ith block (within the entire set of $N = np$ aligned observations) by $R_{n,ij}$, and compute the averages $\overline{R}_{n,j} = n^{-1}\sum_{i=1} R_{n,ij}$, $j = 1, \ldots, p$. Also let $\overline{R}_{n,i\cdot} = p^{-1}\sum_{j=1}^{p} R_{n,ij}$, $i = 1, \ldots, n$, and define

$$V_n = [n(p-1)]^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}(\overline{R}_{n,j} - \overline{R}_{n,i\cdot})^2. \quad (46)$$

Then the aligned rank sum test for the univariate case is based on

$$\mathcal{L}_N^0 = \frac{n}{V_n}\sum_{j=1}^{p}\left(\overline{R}_{n,j} - \frac{N+1}{2}\right)^2. \quad (47)$$

A similar statistic in the multivariate case with general scores has been worked out in Sen [22], and reported in detail in Puri & Sen [19, Chapter 7]. In the particular case of aligned rank sum procedure, the test statistic involves the aligned rank average of each treatment, for each coordinate, averaged over the n blocks, and the rank correlation matrix of these aligned ranks, so the analogy with the multivariate χ_r^2 statistic is quite apparent. The asymptotic null distributions of these aligned rank tests are the same as those based on the method of n rankings, but in the nonnull case, particularly, for local alternatives, they fare better than the intrablock rank tests in terms of ARE. On the other hand, the method of n rankings does not require block-additivity, and hence in situations in which this additivity is in question, they may fare better than aligned rank tests. In passing, we may remark that median-type and rank-sum-type statistics, based either on within block rankings or aligned ones, have also been considered for **factorial designs**. These tests are also (conditionally for aligned ranks) distribution-free and possess some nice properties. For a detailed account of these procedures for replicated 2^m experiments, we refer to Sen [23].

Rank MANCOVA

One of the major advantages of the Chatterjee–Sen rank-permutation principle in the multivariate case is that it provides an access to handling the analysis of covariance in a general multivariate setup, termed MANCOVA, without any further complication. Suppose that there are p primary variates and q concomitant or **covariates**, so that we have a $(p+q)$-variate distribution of the responses. A minimal requirement for a variable to qualify as covariable or concomitant variable is that its distribution is not affected by the treatment differences that are likely to arise with the primary response variates. Therefore, granted this basic assumption, the homogeneity of the conditional distributions of the primary response variables, given the concomitant variables, across the treatment regimen, implies the homogeneity of the $(p+q)$-variate distributions over the treatment regimen. On the other hand, the concomitant variates may not contribute to any significant differences in the treatment regimen, and hence, not to lose any power, should not be treated as primary response variables. The resolution is quite simple. First, consider an appropriate rank statistic, say $\mathcal{L}_N$, for testing the homogeneity of the $(p+q)$-variate distributions for the different treatment. Next, consider the parallel test statistic just confining attention to the concomitant variates, and let it be denoted by $\mathcal{L}_N^0$. Then the covariate adjusted rank test statistic for the primary variates, denoted by $\mathcal{L}_N^*$, is given by

$$\mathcal{L}_N^* = \mathcal{L}_N - \mathcal{L}_N^0. \quad (48)$$

By the Cochran theorem on quadratic forms (*see* **Chi-square, Partition of**), $\mathcal{L}_N^*$ is easily shown to be nonnegative, and clearly, for different score functions, we would obtain different such MANCOVA test statistics. For example, if $p = 1$ and $q \geq 1$, and we use the Wilcoxon scores (for the c sample problem), we will get a version which resembles the Kruskal–Wallis

statistic, adjusted for covariates; this was considered by Quade [20], and a unified theory based on their rank-permutation principle is covered in Puri & Sen [21]. The same thing can be done for Friedman's χ_r^2 statistic, via the multivariate extension due to Gerig [11]. For aligned rank statistics too, a similar picture holds. Basically these MANCOVA rank tests are conditionally (permutationally) distribution-free), have asymptotically chi-square distributions, and are more efficient than their MANOVA counterparts where the concomitant variates are ignored (leading to possible loss of information). For details, we refer the reader to Puri & Sen [19, Chapter 5].

We conclude with the remark that for both the sign (median) and rank-sum-type procedures, the statistics can be handled as suitable functions of (generalized) U-statistics, and hence, their distribution theory (under the null as well as alternatives) can be studied under much less restrictive regularity conditions than in the case of general rank statistics. Moreover, adjustments for ties can also be made without much pain. Finally, being based on bounded scores, they have excellent robustness properties, and they are reasonably efficient for a broad class of underlying distributions, including the normal, Laplace, and the logistic cdfs. In applications in the general field of biostatistics, we therefore advocate the use of such procedures whenever the underlying distribution is suspected to be different from a normal one, as well as when we have ranked data sets.

References

[1] Bennett, B.M. (1964). A bivariate signed rank test, *Journal of the Royal Statistical Society, Series B* **26**, 457–461.

[2] Bickel, P.J. (1965). On some asymptotically nonparametric competitors of Hotelling's T^2, *Annals of Mathematical Statistics* **36**, 160–173.

[3] Blumen, I. (1958). A new bivariate sign test, *Journal of the American Statistical Association* **53**, 448–456.

[4] Brown, G.W. & Mood, A.M. (1951). On median tests for linear hypotheses, in *Proceedings of the Second Berkeley Symposium on Mathematics and Statistics Problems*, Vol. 1. University of California Press, Berkeley, pp. 159–166.

[5] Chatterjee, S.K. (1966). A bivariate sign test for location, *Annals of Mathematical Statistics* **37**, 1771–1782.

[6] Chatterjee, S.K. & Sen, P.K. (1964). Nonparametric tests for the bivariate two-sample location problem, *Calcutta Statistical Association Bulletin* **13**, 18–58.

[7] Chatterjee, S.K. & Sen, P.K. (1965). Some nonparametric tests for the bivariate two-sample association problem, *Calcutta Statistical Association Bulletin* **14**, 14–34.

[8] Chatterjee, S.K. & Sen, P.K. (1966). Nonparametric tests for the multivariate multisample location problem, in *Essays in Probability and Statistics in Memory of S.N. Roy*, R.C. Bose et al., eds. University of North Carolina Press, Chapel Hill, pp. 197–228.

[9] Chaudhuri, P. & Sengupta, D. (1993). Sign tests in multidimension: inference based on the geometry of the data cloud, *Journal of the American Statistical Association* **88**, 1363–1370.

[10] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.

[11] Gerig, T.M. (1969). A multivariate extension of Friedman's χ^2-test, *Journal of the American Statistical Association* **64**, 1595–1608.

[12] Hájek, J. (1965). Extensions of the Kolmogorov–Smirnov test to regression alternatives, in *Proceedings of the Bernoulli–Bayes–Laplace Seminar, Berkeley*, L. LeCam, ed. University of California Press, Berkeley, pp. 45–60.

[13] Hájek, J. & Šidák, Z. (1967). *Theory of Rank Tests.* Academia, Prague.

[14] Hodges, J.L., Jr (1955). A bivariate sign test, *Annals of Mathematical Statistics* **26**, 523–527.

[15] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583–621.

[16] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[17] Mood, A.M. (1950). *An Introduction to the Theory of Statistics*. McGraw-Hill, New York.

[18] Oja, H. (1983). Descriptive statistics for multivariate distributions, *Statistics & Probability Letters* **1**, 327–332.

[19] Puri, M.L. & Sen, P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[20] Quade, D. (1967). Rank analysis of covariance, *Journal of the American Statistical Association* **62**, 1187–1200.

[21] Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau, *Journal of the American Statistical Association* **63**, 1379–1389.

[22] Sen, P.K. (1969). Nonparametric tests for multivariate interchangeability, part II: the problem of MANOVA in two-way layouts, *Sankhyā, Series A* **30**, 145–156.

[23] Sen, P.K. (1970). Nonparametric inference in replicated 2^m factorial experiment, *Annals of the Institute of Statistical Mathematics* **22**, 281–294.

[24] Sen, P.K. (1995). Paired comparisons for multiple characteristics: an ANOCOVA approach, in *Statistical Theory and Applications: Papers in Honor of Herbert A. David*, H.N. Nagaraja et al., eds. Springer-Verlag, New York, pp. 247–264.

[25] Sen, P.K. & David, H.A. (1968). Paired comparisons for paired characteristics, *Annals of Mathematical Statistics* **39**, 200–208.

[26] Sen, P.K. & Puri, M.L. (1967). On the theory of rank order tests for location in the multivariate one sample problem, *Annals of Mathematical Statistics* **38**, 1216–1228.

[27] Wald, A. & Wolfowitz, J. (1944). Statistical tests based on permutations of the observations, *Annals of Mathematical Statistics* **15**, 358–372.

[28] Wilcoxon, F. (1949). *Some Rapid Approximate Statistical Procedures*. American Cyanamid Co., New York.

(*See also* **Multivariate Analysis, Overview**)

PRANAB K. SEN

Multivariate Methods for Binary Longitudinal Data

In many studies one measures a dichotomous (or **binary**) response variable and a set of **covariates** at several times for each of many individuals. When the outcome can occur only once or only the first occurrence is of interest, then methods of **survival analysis** are commonly used. However, the subject of binary longitudinal data considers the situation when the outcome can recur and the relationship of covariates with the multiple occurrences of the outcome is of interest. Examples of diseases or clinical outcomes that can recur in the same patient are common and include asthma attacks, skin cancers, urinary tract infections, myocardial infarctions, injuries, migraines, seizures in epileptics, and admissions to hospital.

The challenge to the multivariate analysis of such data arises because some individuals are more prone to recurrences than others. Thus, the repeated measures of the response variable will generally be positively correlated and regression approaches must account for these correlations (*see* **Longitudinal Data Analysis, Overview**). A variety of regression models appropriately account for these correlations and several excellent reviews have compared and discussed their alternative indications and performance [4, 9, 12, 14, 15].

Types of Models

Marginal Models

The scientific question of interest dictates the choice of model. Consider, for example, a longitudinal study in which use of antihypertensive drugs is assessed at three different times in each member of an elderly cohort. One question of interest is whether use of these drugs differs by age and sex and also changes over time because of the publication between the surveys of evidence from large-scale clinical trials demonstrating the efficacy of antihypertensive treatment in the elderly. This suggests the following model:

$$\log\left[\frac{\mu_{it}}{1-\mu_{it}}\right] = \beta_0 + \beta_1 age_{it} + \beta_2 sex_i + \beta_3 weight_{it} + \beta_4 time_t, \qquad (1)$$

where μ_{it} is the probability that the ith individual is using antihypertensive drugs at the tth survey, age_{it} and $weight_{it}$ are age and weight of the ith individual at the tth survey, sex_i is the sex of this individual, and $time_t$ is an indicator variable denoting whether the survey is before or after the demonstration of efficacy from randomized trials. Note that the model contains time-varying as well as fixed covariates. Parameters in the model are interpretable in terms of the odds of using antihypertensive drugs; for example, $\exp(\beta_4)$ is the odds of using antihypertensive drugs in the later time period relative to the earlier period, for persons of the same age, sex, and weight.

Eq. (1) is called a **marginal model** because it does not include among the independent variables the previous levels of the response variable. It commonly relates levels of covariates at each observation time to the dichotomous outcome at that time, although previous levels of covariates can also be included. The interrelationship among the repeated measures of the response variable is usually considered to be a nuisance. To complete the specification of model (1), one must also specify the relationship between the variance of the response variable and the independent variables and the association between the repeated measures of the dichotomous response variable. Commonly used and interpretable choices are that $\text{var}(Y_{it}) = \lambda\mu_{it}(1-\mu_{it})$, where λ is a constant, and that $\text{corr}(Y_{it}, Y_{is})$ depends only on the time interval between measures, $t-s$, where Y_{it} is the response variable of the ith individual at time t. Alternatively, one may parameterize the interrelationship between repeated measures of the response variable in terms of an odds ratio or allow the strength of this interrelationship to vary according to levels of the independent variables [3, 13, 18].

Mixed-Effects Models

The above marginal model is sometimes called the population-averaged approach because it assumes common regression parameters β for all individuals in the population. One may prefer to include random effects terms and the model is then called a subject-specific model (*see* **Marginal Models**). A simple model of this type is

$$\log\left[\frac{\mu_{it}}{1-\mu_{it}}\right] = \beta_{0i} + \beta_1 age_{it} + \beta_2 sex_i + \beta_3 weight_{it} + \beta_4 time_t, \qquad (2)$$

where μ_{it} is the expectation of Y_{it} conditional on a subject-specific intercept β_{0i} that is assumed to vary according to a known distribution, e.g. normal $(0, \sigma^2)$ [2, 17, 25, 26]. The repeated measures Y_{it} and Y_{is} are then assumed to be independent, conditional on the individual random effect β_{0i}.

The parameters β in models (1) and (2) have different interpretations. Consider, for example, the parameter β_3 which indexes the effect of weight in the two models. In the population-averaged approach this coefficient summarizes the average effects over the whole population of a unit difference in weight. In the subject-specific model this coefficient summarizes the effect of a unit change in weight within an individual. Neuhaus et al. [16] studied the relationship between parameters from the two models and found that, as long as the correlation between repeated response variables is positive, parameters in model (1) are always smaller in absolute value than the corresponding parameters in model (2), unless that parameter is identically zero.

Transition Models

An alternative question of scientific interest focuses on transitions between categories of the dependent variable. In the above example one may be interested in determinants of new use of antihypertensive drugs or in withdrawal or noncompliance among previous users of these drugs. In this case a model of interest may be

$$\begin{aligned}\log\text{it}[\Pr(Y_{it} = 1)] = {} & \beta_0 + \beta_1 age_{it} + \beta_2 sex_i \\ & + \beta_3 weight_{it-1} + \beta_4 time_t \\ & + \beta_5 Y_{it-1}. \end{aligned} \tag{3}$$

Or, more generally, one may assume that the current value of the response variable depends not only on the most recent values, but on all previous values of this variable [1]. Interpretation of regression coefficients for age, sex, weight, and time differs from the marginal model because in the transition model effects are conditional on the previous value of the response variable. Hence, this model is called a conditional model. Because the relationship of covariates with initiation of therapy is likely to differ from their relationship with withdrawal or noncompliance, interactions between the effects of the covariates and the previous response variables may be of particular interest.

A related conditional model is of interest when one focuses on the interrelationship among the values of the repeated responses. This model conditions on subsequent as well as previous values of the response variable:

$$\begin{aligned}\log\text{it}[\Pr(Y_{it} = 1 | Y_{-it})] & \\ = F\left(\sum_{k \neq t} Y_{ik}, \theta\right) & + \beta_1\ age_{it} + \beta_2\ sex_i \\ + \beta_2\ weight_{it} & + \beta_3\ time_t, \end{aligned} \tag{4}$$

where Y_{-it} denotes all response values for the ith person except that at the tth time and F is an arbitrary function of $\sum_{k \neq t} Y_{ik}$ and parameters θ [5, 20, 22]. This conditional expression allows for specification of the joint distribution of the repeated response variables and so this model can also be used to estimate the probability that a person uses antihypertensive drugs at all three times or at no time.

Approaches to Inference

In describing the above models we have presented only the relationship of the first and second moments of the response variable to the independent variables. Because of the interrelationships among the repeated measures, additional assumptions are often required to specify the complete likelihood. A challenge for the estimation of the marginal model (1) is the absence of joint probability distributions for multivariate binary data that yield simple expressions for marginal means [6]. Because of the intractability of the likelihood, Liang & Zeger [11] have recommended use of the **generalized estimating equations** (GEE) method, a form of **quasi-likelihood**, for estimation and inference. Estimates of the parameters β are solutions of the GEE:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{5}$$

where N is the number of subjects, $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^{\mathrm{T}}$, and $\mathbf{V}_i$ is an approximate or "working" covariance matrix of Y_i. Several alternative forms of $\mathbf{V}_i$ are commonly used, including the assumption that the correlation between any two different measures in the same person is independent of time (**exchangeable** correlation), that it depends only on the time interval

(autoregressive correlation), or that it is completely unspecified. Liang & Zeger showed that the GEE approach yields consistent estimates of the parameters $\boldsymbol{\beta}$ in (1), provided only that the specification of the expected value of the dependent variable is correct. Even if the covariance matrix is misspecified, the following robust estimate of the variance of the parameters is available:

$$\hat{V}(\hat{\boldsymbol{\beta}}) = H_1^{-1}(\hat{\boldsymbol{\beta}}) H_2(\hat{\boldsymbol{\beta}}) H_1^{-1}(\hat{\boldsymbol{\beta}}),$$

where

$$H_1(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i,$$

$$H_2(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i,$$

and $\hat{\mathbf{V}}_i$ and $\hat{\mathbf{D}}_i$ are $\mathbf{V}_i$ evaluated at $\hat{\boldsymbol{\beta}}$.

An extension of the estimating equation approach allows for a parameterization of the covariance $\mathbf{V}_i$, or alternatively of pairwise odds ratios between repeated measures, as functions of independent variables [18, 29]. Prentice & Zhao presented estimating equations for the joint estimation of both mean and covariance parameters. They also showed that these estimating equations correspond to the score equations under a quadratic exponential **likelihood** for the joint distribution of the repeated measures. Fitzmaurice & Laird [8] have also presented a likelihood-based approach that is closely related to the GEE method.

For the mixed-effects model (2), the logistic-normal likelihood is identified by the specification that the random intercept follows a normal distribution and, conditional on this intercept, the repeated measures for an individual are independent and have a logistic relationship with the independent variables. However, the likelihood does not have a closed-form expression and thus a full maximum likelihood analysis requires numerical integration [7]. One alternative is the use of Gibbs sampling techniques [27] (*see* **Markov Chain Monte Carlo**). Other approximate methods are based on penalized quasi-likelihood or marginal quasi-likelihood approaches [2, 25, 26, 28].

Transition models such as model (3) can be easily fitted with available software for logistic regression analysis if the model adequately captures the dependence on previous levels of the outcome variable. For example, Bonney has presented straightforward methods for estimation under his proposed conditional model [1]. Rosner [22] has described likelihood-based estimation under the related conditional model (4).

Two issues that frequently complicate estimation in analysis of longitudinal data are the presence of missing data and of multiple levels of clustering in the data. Laird [10] discussed general issues of missing data in longitudinal studies, focusing on likelihood-based methods. Rotnitzky & Wypij [24] presented methods to quantify the bias that can arise from missing data in both likelihood and quasi-likelihood estimation approaches. Robins et al. [21] have given modifications to the GEE approach which yield consistent estimates when missing data are missing at random (*see* **Nonignorable Dropout in Longitudinal Studies**).

Multiple levels of clustering would arise in the example presented above if siblings or spouses whose use of antihypertensive drugs might be associated were included in the longitudinal surveys. Alternatively, in hypertension studies one often takes multiple measures of blood pressure of hypertension status within a short time interval and then repeats these measures after a longer interval. In this case subjects have repeated assessments at intervals such as one year and each assessment is composed of replicate measures a short time (e.g. one week) apart. Rosner presented an extension of his model to account for such multiple levels of clustering [23]. Qaqish & Liang [19] described GEE approaches with multiple measures on individuals who are clustered within families.

References

[1] Bonney, G.E. (1987). Logistic regression for dependent binary observations, *Biometrics* **43**, 951–973.

[2] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear models, *Journal of the American Statistical Association* **88**, 9–25.

[3] Carey, V., Zeger, S.L. & Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions, *Biometrika* **80**, 517–526.

[4] Clayton, D.G. (1994). Some approaches to the analysis of recurrent event data, *Statistical Methods in Medical Research* **3**, 244–262.

[5] Connolly, M.A. & Liang, K.-Y. (1988). Conditional logistic regression models for correlated binary data, *Biometrika* **75**, 501–506.

[6] Cox, D.R. (1972). The analysis of multivariate binary data, *Applied Statistics* **21**, 113–120.

[7] Crouch, E.A.C. & Spiegelman, D. (1990). The evaluation of integrals of the form $\int f(t)\exp(-t^2)\mathrm{d}t$: applications to logistic normal models, *Journal of the American Statistical Association* **85**, 464–469.

[8] Fitzmaurice, G.M. & Laird, N.M. (1993). A likelihood-based method for analyzing longitudinal binary data, *Biometrika* **80**, 141–151.

[9] Fitzmaurice, G.M., Laird, N.M. & Rotnitzky, A.G. (1993). Regression models for discrete longitudinal responses, *Statistical Science* **8**, 284–309.

[10] Laird, N.M. (1988). Missing data in longitudinal studies, *Statistics in Medicine* **7**, 305–315.

[11] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[12] Liang, K.-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analysis for categorical data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

[13] Lipsitz, S.R., Laird, N.M. & Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, *Biometrika* **78**, 153–160.

[14] Neuhaus, J.M. (1993). Estimation efficiency and tests of covariate effects with clustered binary data, *Biometrics* **49**, 989–996.

[15] Neuhaus, J.M. (1992). Statistical methods for longitudinal and clustered designs with binary responses, *Statistical Methods in Medical Research* **1**, 249–273.

[16] Neuhaus, J.M., Kalbfleisch, J.D. & Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data, *International Statistical Review* **59**, 25–36.

[17] Pierce, D.A. & Sands, B.R. (1986). Extra-Bernoulli Variation in Regression of Binary Data. *Technical Report 46*, Dept of Statistics, Oregon State University.

[18] Prentice, R.L. & Zhao, L.P. (1991). Estimating equations for parameters in mean and covariances of multivariate discrete and continuous responses, *Biometrics* **47**, 825–839.

[19] Qaqish, B.F. & Liang, K.-Y. (1992). Marginal models for correlated binary responses with multiple classes and multiple levels of nesting, *Biometrics* **48**, 939–950.

[20] Qu, Y.S., Williams, G.W., Beck, G.J. & Goormastic, M. (1987). A generalized model of logistic regression for correlated data, *Communications in Statistics A* **16**, 3447–3476.

[21] Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**, 106–121.

[22] Rosner, B. (1984). Multivariate methods in ophthalmology with applications to other paired data situations, *Biometrics* **40**, 1025–1035.

[23] Rosner, B. (1989). Multivariate methods for clustered binary data with more than one level of nesting, *Journal of the American Statistical Association* **84**, 373–380.

[24] Rotnitzky, A. & Wypij, D. (1994). A note on the bias of estimators with missing data, *Biometrics* **50**, 1163–1170.

[25] Stiratelli, R., Laird, N.M. & Ware, J.H. (1984). Random effects models for serial observations with binary response, *Biometrics* **40**, 961–971.

[26] Waclawiw, M.A. & Liang, K.-Y. (1994). Empirical Bayes estimation and inference for the random effects model with binary response, *Statistics in Medicine* **13**, 541–551.

[27] Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association* **86**, 79–86.

[28] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

[29] Zhao, L.P. & Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika* **77**, 642–648.

(*See also* **Categorical Data Analysis**; **Distribution-free Methods for Longitudinal Data**)

ROBERT J. GLYNN & BERNARD ROSNER

Multivariate Multiple Regression

Multiple regression analysis is used whenever we wish to model the relationship between a *dependent* (**response** or *endogenous*) variable y and a set of p **explanatory** (*regressor* or *exogenous*) variables $x_1, \ldots, x_p$. Many different forms of relationship are possible, but the overwhelming emphasis in practical applications is on the *linear* relationship $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, where the β_j are parameters (since such a relationship often will hold approximately for some suitable **transformations** of the measured variables even if it does not hold for the variables themselves) (*see* **Multiple Linear Regression**).

In a typical application, values of each of these variables are observed on each of n sample individuals. Let us suppose that y_i denotes the value of the dependent variable and that $x_{i1}, \ldots, x_{ip}$ denote the corresponding values of the explanatory variables for the ith individual in the sample. The explanatory variables are usually assumed to be measured without error, but the dependent variable is subject to measurement errors. The statistical model for this situation is thus written

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

for $i = 1, 2, \ldots, n$. Here ε is a random *departure* term which represents the measurement errors; since random sampling is assumed, the ε_i are taken to be independent and identically distributed random variables each having mean zero and constant variance σ^2.

If we collect together all the y_i, β_j, and ε_i into the vectors

$$\mathbf{y} = (y_1, \ldots, y_n)',$$
$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)',$$
$$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)',$$

and write the explanatory variable values in the $n \times (p+1)$ matrix $\mathbf{X} = (x_{ij})$ for $i = 1, \ldots, n$ and $j = 0, 1, \ldots, p$, where $x_{i0} = 1$ for all i, then the above model can be written in the compact form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In this formulation $\boldsymbol{\varepsilon}$ is a random vector whose mean is $\mathbf{0}$ and whose dispersion (or **covariance matrix**) is $\sigma^2\,\mathbf{I}$.

The main interest in practical applications is the estimation of the parameters $\boldsymbol{\beta}$, **hypothesis tests** and/or **confidence intervals** for them, and perhaps **variable selection** in order to model the relationship between dependent and explanatory variables as **parsimoniously** as possible. **Least squares** estimates of the parameters are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

provided that $\mathbf{X}$ is of full rank (which is the case that we assume). For all the remaining objectives it is also necessary to assume either normality of the departures ε_i or large samples. In the case of normality, **maximum likelihood** estimates of the parameters coincide with the least squares estimates, $\hat{\boldsymbol{\beta}}$ has a **multivariate normal distribution** with mean vector $\boldsymbol{\beta}$ and dispersion matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and σ^2 is estimated from the residual mean square in the **analysis of variance** associated with the regression. For **large samples**, the least squares estimates have asymptotic normality with $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ as the dispersion matrix. These facts enable confidence regions and hypothesis tests to be constructed for elements of $\boldsymbol{\beta}$, while variable selection procedures are built out of various statistics arising in the associated analysis of variance.

Multivariate multiple regression analysis arises when we have q (> 1) dependent variables, and we wish to model the relationship between *each* of these variables and the set of explanatory variables. Attention is again focused almost exclusively on *linear* relationships. If the dependent variables are $y_1, y_2, \ldots, y_q$, the explanatory variables are $x_1, x_2, \ldots, x_p$, and all these variables are observed on each of n sample individuals, then we have to allow each y_i to have its own linear relationship with all the x_i. Thus we have to specify q different linear models, one between each y_i and the set of x_i. By analogy with the above formulation we can write these models as

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$

where

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{in})',$$
$$\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})',$$
$$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in})',$$

for $i = 1, \ldots, q$. Here y_{ij} is the value observed for the ith dependent variable on the jth sample member, ε_{ij} is the departure term corresponding to the ith dependent variable and jth sample member, $\boldsymbol{\beta}_i$ are the parameters appropriate to the ith dependent variable, and $\mathbf{X}$ is the same as before. Even more compactly, putting the q columns $\mathbf{y}_i$ side by side into the $n \times q$ matrix $\mathbf{Y}$, the q columns $\boldsymbol{\beta}_i$ similarly into the $(p+1) \times q$ matrix $\mathbf{B}$, and the q columns $\boldsymbol{\varepsilon}_i$ into the $n \times q$ matrix $\boldsymbol{\Xi}$, we can write all q linear models in the single expression

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\Xi}.$$

Once again the sample individuals are independent, but now there is association among the departure terms $\varepsilon_{1j}, \ldots, \varepsilon_{qj}$ corresponding to the same sample member. Consequently, we can treat the *rows* of $\boldsymbol{\Xi}$ as independent observations from a distribution with mean vector zero and dispersion matrix $\boldsymbol{\Sigma}$, and for most practical purposes this distribution is assumed to be multivariate normal.

Maximizing the likelihood for this model produces the estimator

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

and the same estimator results from various possible definitions of matrix least squares, also. Moreover, since $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_q)$, on picking out appropriate columns from the above equation we find that

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_i,$$

for $i = 1, \ldots, q$. That is to say, the regression coefficients have the same estimates as they would if each dependent variable was regressed *separately* on the set of explanatory variables. However, all the individual $\hat{\beta}_{ij}$ in $\hat{\mathbf{B}}$ are now intercorrelated – those within a column of $\hat{\mathbf{B}}$ because of the **correlations** among the $\mathbf{x}_i$, and those in different columns of $\hat{\mathbf{B}}$ because of the correlations among the $\mathbf{y}_i$. Hence, to control type I error rates (*see* **Level of a Test**), we cannot conduct hypothesis tests separately on each column of $\hat{\mathbf{B}}$ but instead we need *multivariate* tests for hypotheses about $\mathbf{B}$.

These tests are the analogs in **multivariate analysis of variance** (MANOVA) of the **analysis of variance** (ANOVA) ***F* tests** for multiple regression. General theory is given in the article Multivariate Analysis of Variance in the discussion on multivariate **general linear models**; we content ourselves here with specifying the relevant matrices from which the test statistics are obtained. The error matrix, by analogy with multiple regression, is given by

$$\mathbf{E} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}$$

[from which an **unbiased** estimate of $\boldsymbol{\Sigma}$ is given by $\mathbf{E}/(n-p-1)$]. The total matrix is $\mathbf{T} = \mathbf{Y}'\mathbf{Y} - n\overline{\mathbf{y}}\overline{\mathbf{y}}'$ (where $\overline{\mathbf{y}}$ is the vector of sample means of the $\mathbf{y}_i$), so the hypothesis matrix for testing the overall significance of regression is the difference of $\mathbf{T}$ and $\mathbf{E}$, namely

$$\mathbf{H} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - n\overline{\mathbf{y}}\overline{\mathbf{y}}'.$$

Any of the standard test statistics can be used to test the significance of $\mathbf{H}$, e.g. Wilks' **lambda**:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}.$$

To test the hypothesis that the $\mathbf{y}_i$ only depend on a subset of the $\mathbf{x}_i$, partition $\mathbf{B}$ into $\mathbf{B}' = (\mathbf{B}_{\mathrm{r}}'\mathbf{B}_{\mathrm{d}}')$, where $\mathbf{B}_{\mathrm{r}}$ denotes the subset of the β_{ij} that are to be retained, while $\mathbf{B}_{\mathrm{d}}$ denotes the subset of the β_{ij} that are to be deleted from the full model. If $\mathbf{X}_{\mathrm{r}}$ contains the columns of $\mathbf{X}$ corresponding to $\mathbf{B}_{\mathrm{r}}$, then the hypothesis matrix for testing $\mathrm{H}_0 : \mathbf{B}_{\mathrm{d}} = \mathbf{0}$ is

$$\mathbf{H} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - \hat{\mathbf{B}}_{\mathrm{r}}'\mathbf{X}_{\mathrm{r}}'\mathbf{Y},$$

while the error matrix is $\mathbf{E}$ as before.

Finally, the above test provides a basis for stepwise selection of a subset of the $\mathbf{x}_i$ for the most parsimonious prediction of all the $\mathbf{y}_i$, by either adding at each stage the x variable that has the most significant (partial) lambda, or by deleting the x variable that has the least significant (partial) lambda. Full details of such a scheme are given by, for example, Rencher [1].

Reference

[1] Rencher, A.C. (1995). *Methods of Multivariate Analysis*. Wiley, New York.

(*See also* **Multivariate Analysis, Overview**)

W.J. KRZANOWSKI

Multivariate Normal Distribution

Biometric data typically entail observations on multiple characteristics for each experimental subject. Multivariate normal distributions (or multinormal distributions) are often central to the modeling and analysis of such data. Reasons abound: multivariate normal distributions are tractable; they have been studied extensively; their properties are widely known; and they support a variety of known derived distributions. Indeed, many standard problems in statistical inference initially were posed in terms of multinormal distributions. Empirical evidence often points towards the normality of multivariate data. Biometric measurements, especially, may emerge as the result of many small increments due to heredity and environment, so that the approximate multivariate normality of such data rests on multidimensional **central limit** theory. In addition, it is now known that many normal-theory procedures, both univariate and multivariate, remain exact for many nonnormal multivariate distributions exhibiting suitable symmetries. Further details may be found in the article on **multivariate distributions**. Not only do multinormal distributions support an impressive list of known derived distributions, but the limiting joint distributions of numerous statistics arising in data analysis are themselves multinormal in view of central limit theory. The latter include statistics employed in **large-sample theory**, **nonparametrics**, **robust** statistics, and elsewhere. In short, a working knowledge of multivariate normal distributions and their properties is essential to the knowledgeable use and development of statistical methodologies.

Multivariate normal distributions and their applications have a rich history. Origins of these distributions, beginning with two and three dimensions, trace to the early nineteenth century [1, 5, 7, 11, 13, 21, 23, 26]. Studies in heredity, culminating in the work of Galton [10], treat bivariate **correlation** analysis in a biometric setting. Systematic developments in two and three dimensions are credited to Bravais [5] and Schols [26]. Multivariate extensions in current usage are credited to Edgeworth [8], including such essential concepts as **regression** and partial correlations under multinormality. For further details and an excellent overview, see Johnson & Kotz [18, Chapters 35 and 36] and Tong [29].

To fix notation: $\mathbb{R}^k$ designates a Euclidean k-space; $F_{n\times k}$ is the collection of real $(n \times k)$ matrices; S_k consists of real symmetric $(k \times k)$ matrices; and S_k^0 and S_k^+ comprise the positive semidefinite and the positive definite varieties in S_k, respectively. Special arrays include the $(k \times k)$ identity $\mathbf{I}_k$, the unit vector $\mathbf{1}_k = [1, \ldots, 1]' \in \mathbb{R}^k$, and the diagonal matrix diag $(a_1, \ldots, a_k)$. $\mathscr{L}(\mathbf{X})$ designates the law of distribution of $\mathbf{X} \in \mathbb{R}^k$. Abbreviations for probability density, cumulative distribution, and characteristic functions are *pdf, cdf*, and *chf*, respectively, whereas *iid* refers to a sequence of independent identically distributed random elements.

A brief survey of essential properties follows. These are listed under basic properties, probability inequalities, characterizations, and central limit theory.

Basic Properties

Suppose that $\mathbf{X} \in \mathbb{R}^k$ is random having the chf $\phi_X(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$. Then $\mathbf{X}$ is said to have the *multivariate normal distribution* on $\mathbb{R}^k$ with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^k \times S_k^0$, to be designated as $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. **Moments** of all orders are defined; the mean vector and dispersion **covariance matrix** are given by $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{X}) = \boldsymbol{\Sigma}$, respectively; and odd central moments of all orders vanish by symmetry. The class of all finite-dimensional multinormal distributions is closed under affine transformations, in the sense that, if $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on $\mathbb{R}^k$, and if $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ with $\mathbf{A} \in F_{r\times k}$ and $\mathbf{b} \in \mathbb{R}^r$ fixed, then $\mathscr{L}(\mathbf{Y}) = \mathrm{N}_r(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ on $\mathbb{R}^r$. This follows directly from elementary properties of chfs.

The distribution $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is said to be *singular of rank r*, or to be *nonsingular*, according as $\boldsymbol{\Sigma}$ has rank $\mathbf{r} < k$ or rank k, respectively. For the nonsingular case, the pdf exists and is given by

$$f(\mathbf{x}) = [(2\pi)^k |\boldsymbol{\Sigma}|]^{-1/2} \exp\left[\frac{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right]. \tag{1}$$

A reduction to **principal components** proceeds on letting $\mathbf{Y} = \mathbf{PX}$ such that $\mathscr{L}(\mathbf{Y}) = \mathrm{N}_k(\boldsymbol{\theta}, \mathbf{D}_\xi)$, where $\mathbf{P}$ is $(k \times k)$ orthogonal such that $\boldsymbol{\theta} = \mathbf{P}\boldsymbol{\mu}$, and $\mathbf{D}_\xi = \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}' = \mathrm{diag}(\xi_1, \ldots, \xi_k)$ contains the ordered **eigenvalues** $\{\xi_1 \geq \xi_2 \geq \cdots \geq \xi_k \geq 0\}$ of $\boldsymbol{\Sigma}$. In particular,

for the case that $\mathscr{L}(\mathbf{X})$ is singular of rank r, the last $k-r$ elements of $\mathbf{Y}$ are concentrated at the last $k-r$ elements of $\boldsymbol{\theta}$ with unit probability.

Joint marginal and conditional distributions of $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ emerge as follows (see the article on **multivariate distributions**). Partition $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2']'$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1', \boldsymbol{\mu}_2']'$, and $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_{ij}]$ conformably, with $\mathbf{X}_1 \in \mathbb{R}^r$ and $\mathbf{X}_2 \in \mathbb{R}^t$ such that $r + t = k$. Then the joint marginal distribution of $\mathbf{X}_1$ is itself multivariate normal on $\mathbb{R}^r$ as given by $\mathscr{L}(\mathbf{X}_1) = \mathrm{N}_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, and similarly $\mathscr{L}(\mathbf{X}_2) = \mathrm{N}_t(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ on $\mathbb{R}^t$. In like manner, the conditional distributions $\mathscr{L}(\mathbf{X}_1|\mathbf{x}_2)$ for $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, are themselves multinormal on $\mathbb{R}^r$ for every fixed $\mathbf{x}_2$. Specifically, $\mathscr{L}(\mathbf{X}_1|\mathbf{x}_2) = \mathrm{N}_r(\boldsymbol{\mu}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2})$, having the linear regression functions $\mathrm{E}(\mathbf{X}_1|\mathbf{x}_2) = \boldsymbol{\mu}_{1\cdot 2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$, and dispersion parameters $\boldsymbol{\Sigma}_{11\cdot 2} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{\Sigma}_{21})$ (Cambanis et al. [6]). Here $\boldsymbol{\Sigma}_{22}^{-}$ is any generalized inverse of $\boldsymbol{\Sigma}_{22}$. For distributions having full rank k, $\boldsymbol{\Sigma}_{22}^{-}$ becomes the unique inverse $\boldsymbol{\Sigma}_{22}^{-1}$. The matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}$ comprises the *partial regression coefficients* of $\mathbf{X}_1$ on $\mathbf{x}_2$. The squared **canonical correlations** between elements of $\mathbf{X}_1$ and $\mathbf{X}_2$ are the eigenvalues of $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$. For the case $r = 1$, the single eigenvalue is called the squared *multiple correlation coefficient* between $\mathbf{X}_1$ and $\mathbf{X}_2$.

Many further details are provided in Johnson & Kotz [18] and Tong [29], for example. These include sources for special aid tables and computational **algorithms** for evaluating various multinormal probabilities, as well as extensive reference lists.

Probability Inequalities

Basic Inequalities

Basic inequalities for multivariate normal distributions are essential. In practice, there is an excess of parameters, since $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ consist of $k(k+3)/2$ distinct parameters. Owing to limitations of special aid tables and the software now available, access to probability inequalities enables the user to employ approximate values from existing tables or algorithms, giving bounds on the required probabilities. In this spirit, some useful inequalities may be summarized as follows. Basic sources are [28, Chapter 2] and [29, Chapters 5 and 7], together with extensive reference lists provided there.

In what follows, $\boldsymbol{\Sigma} = (\sigma_{ij})$ designates any positive-definite $(k \times k)$ matrix, whereas $\mathbf{R} = (\rho_{ij})$ signifies a positive-definite correlation matrix. The special equicorrelation matrix is denoted by $\boldsymbol{\Xi}(\rho) = [(1-\rho)\mathbf{I}_k + \rho\mathbf{1}_k\mathbf{1}'_k]$ for $[-(k-1)^{-1} < \rho < 1]$. On occasion we suppose that $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k[\mathbf{0}, \mathbf{R}(\kappa)]$, its correlation matrix $\mathbf{R}(\boldsymbol{\kappa})$ having the structure $(\rho_{ij} = \kappa_i\kappa_j\omega_{ij}, i \neq j)$ such that $(|\kappa_i| \leq 1; 1 \leq i \leq k)$, where $\boldsymbol{\Omega} = (\omega_{ij})$ is a positive-definite correlation matrix. With these conventions in place, let $P_{\boldsymbol{\Sigma}}(\cdot)$ be the probability measure for $\mathscr{L}(X_1, \ldots, X_k) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; let $F_{\boldsymbol{\Sigma}}(x_1, \ldots, x_k)$ be its cdf; and let $\overline{F}_{\boldsymbol{\Sigma}}(x_1, \ldots, x_k) = P_{\boldsymbol{\Sigma}}(X_1 > x_1, \ldots, X_k > x_k)$. In addition, let $F_{\mathbf{D}}(x_1, \ldots, x_k)$ be the cdf of $\mathrm{N}_k(\boldsymbol{\mu}, \mathbf{D})$, with $\mathbf{D} = \mathrm{diag}\,(\sigma_{11}, \ldots, \sigma_{kk})$, and similarly for $\overline{F}_{\mathbf{D}}(x_1, \ldots, x_k)$. Furthermore, identify $G_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k) = P_{\boldsymbol{\Sigma}}(|X_1| \leq a_1, \ldots, |X_k| \leq a_k)$ for the case $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$, and similarly $G_{\mathbf{D}}(a_1, \ldots, a_k)$ with $\mathbf{D} = \mathrm{diag}\,(\sigma_{11}, \ldots, \sigma_{kk})$ as before. Basic probability inequalities may be summarized as follows:

Inequality 1. If $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for fixed but arbitrary $(a_1, \ldots, a_k)$, the function $F_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k)$ is increasing in each σ_{ij} for all $i \neq j$ while other values are held fixed.

Inequality 2. Suppose that $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\sigma_{ij} \geq 0$ for all $i \neq j$, then $F_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k) \geq F_{\mathbf{D}}(a_1, \ldots, a_k) \geq \prod_{i=1}^{k} F_i(a_i)$ holds for each fixed $(a_1, \ldots, a_k)$, where $F_i(\cdot)$ is the marginal cdf of X_i.

Inequality 3. Suppose that $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\sigma_{ij} \geq 0$ for all $i \neq j$, then $\overline{F}_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k) \geq \overline{F}_{\mathbf{D}}(a_1, \ldots, a_k) \geq \prod_{i=1}^{k}[1 - F_i(a_i)]$ for arbitrarily fixed $(a_1, \ldots, a_k)$.

Inequality 4. Suppose that $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$. Then $G_{\boldsymbol{\Sigma}}(c_1, \ldots, c_k) \geq G_{\mathbf{D}}(c_1, \ldots, c_k) \geq \prod_{i=1}^{k} G_i(c_i)$ for each fixed set $(c_1, \ldots, c_k)$ of positive constants, where $G_i(\cdot)$ is the marginal cdf of $|X_i|$.

Inequality 5. Suppose that $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k[\mathbf{0}, \boldsymbol{\Xi}(\rho)]$ with $P_\rho(\cdot)$ as the corresponding probability measure. Then for each fixed $a > 0$ and for all real numbers $(c_1, \ldots, c_k)$ such that $c_1 + \cdots + c_k = 0$, $P_\rho(|\sum_{i=1}^{k} c_i X_i| \leq a)$ is an increasing function of ρ.

Inequality 6. Suppose that $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k[\mathbf{0}, \mathbf{R}(\boldsymbol{\kappa})]$. Then for each fixed set of positive numbers

$(c_1, \ldots, c_k)$, the function $G_k(c_1, \ldots, c_k)$: (i) is strictly increasing in each $\kappa_i \in [0, 1]$ with other parameters held fixed; and (ii) is strictly decreasing in each $\kappa_i \in [-1, 0]$ with other parameters held fixed.

Concentration Properties

The comparative concentration of probabilities on $\mathbb{R}^k$ is an essential concept. Following Sherman [27], the probability measure $\mu(\cdot)$ is said to be *more peaked about* $\mathbf{0} \in \mathbb{R}^k$ than $\nu(\cdot)$ if and only if $\mu(\mathrm{A}) \geq \nu(\mathrm{A})$ for every set A in the class $\mathbf{C}_k$ consisting of the compact convex subsets of $\mathbb{R}^k$ that are symmetric under reflection about $\mathbf{0} \in \mathbb{R}^k$, i.e. $\mathbf{x} \in \mathrm{A}$ implies $-\mathbf{x} \in \mathrm{A}$. For two multivariate normal distributions $\mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$ and $\mathrm{N}_k(\mathbf{0}, \boldsymbol{\Omega})$ on $\mathbb{R}^k$ having ordered scale matrices, the following inequality applies. Sufficiency is shown in [2], and necessity in [17].

Inequality 7. Let $\mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$ and $\mathrm{N}_k(\mathbf{0}, \boldsymbol{\Omega})$ be multivariate normal distributions on $\mathbb{R}^k$ having ordered scale matrices such that $\boldsymbol{\Omega} - \boldsymbol{\Sigma}$ is positive semidefinite, and let $P_{\boldsymbol{\Sigma}}(\cdot)$ and $P_{\boldsymbol{\Omega}}(\cdot)$ be their corresponding probability measures. Then $P_{\boldsymbol{\Sigma}}(\cdot)$ is more concentrated about $\mathbf{0}$ than $P_{\boldsymbol{\Omega}}(\cdot)$ in the sense that $P_{\boldsymbol{\Sigma}}(A) \geq P_{\boldsymbol{\Omega}}(A)$ for every set A in the class $\mathbf{C}_k$.

Characterizations

Multivariate distributions exhibit many properties. Various characterizations of multivariate normality are concerned with properties unique to multinormal distributions on $\mathbb{R}^k$ and samples therefrom. To survey such results let $\mathbf{X} \in \mathbb{R}^k$ be random with pdf $p(\mathbf{x})$. Then its *entropy* is defined as $-\mathrm{E}[\ln p(\mathbf{X})]$. Furthermore, let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ be random vectors in $\mathbb{R}^k$; write $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)' \in F_{n\times k}$; let $(\overline{\boldsymbol{X}}, \mathbf{S})$ be the sample mean vector and the sample dispersion matrix, i.e. $\overline{\boldsymbol{X}} = n^{-1}(\mathbf{X}_1 + \cdots + \mathbf{X}_n)$ and $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{\boldsymbol{X}})(\mathbf{X}_i - \overline{\boldsymbol{X}})'/(n-1)$; and let $(\mathbf{A}_1, \ldots, \mathbf{A}_n, \mathbf{B}_1, \ldots, \mathbf{B}_n)$ be nonsingular $(k \times k)$ matrices. The following characterizations are known, a basic reference being [20].

Characteristic 1. Suppose that $\mathbf{X} \in \mathbb{R}^k$ has the pdf $p(\mathbf{x})$ with finite dispersion matrix $\mathrm{var}(\mathbf{X}) = \boldsymbol{\Sigma}$. Then $\mathscr{L}(\mathbf{X})$ has maximal entropy among all distributions on $\mathbb{R}^k$ with dispersion matrix $\boldsymbol{\Sigma}$, if and only if $\mathscr{L}(\mathbf{X})$ is multinormal, i.e. $\mathscr{L}(\mathbf{X}) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [24].

Characteristic 2. Given that $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ are iid on $\mathbb{R}^k$, then $(\overline{\boldsymbol{X}}, \mathbf{S})$ are independent if and only if each of $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ is multinormal on $\mathbb{R}^k$.

Characteristic 3. Given that $(\mathbf{X}_1, \mathbf{X}_2)$ are independent on $\mathbb{R}^k$, then $\mathbf{W} = \mathbf{X}_1 + \mathbf{X}_2$ is multivariate normal on $\mathbb{R}^k$ if and only if each of $(\mathbf{X}_1, \mathbf{X}_2)$ is multinormal on $\mathbb{R}^k$.

Characteristic 4. Let $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ be independent random vectors on $\mathbb{R}^k$. If the linear forms $\mathbf{W}_1 = \mathbf{A}_1\mathbf{X}_1 + \cdots + \mathbf{A}_n\mathbf{X}_n$ and $\mathbf{W}_2 = \mathbf{B}_1\mathbf{X}_1 + \cdots + \mathbf{B}_n\mathbf{X}_n$ are independent, then each of $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ is multivariate normal on $\mathbb{R}^k$ [12].

Characteristic 5. Suppose for $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)' \in F_{n\times k}$ that its distribution is invariant under left-orthogonal transformations, i.e. $\mathscr{L}(\mathbf{X}) = \mathscr{L}(\mathbf{PX})$ for every $(n \times n)$ orthogonal matrix $\mathbf{P}$. Then $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ are independent if and only if each of $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ is multinormal on $\mathbb{R}^k$ [14].

Characteristic 6. For $\mathbf{X} \in \mathbb{R}^k$, the distribution $\mathscr{L}(\mathbf{X})$ is multivariate normal if and only if every linear combination $L = a_1X_1 + \cdots + a_kX_k$ is normal on $\mathbb{R}^k$.

Characteristic 7. The conditional expectation $\mathrm{E}(\overline{\boldsymbol{X}} | \mathbf{X}_2 - \mathbf{X}_1, \ldots, \mathbf{X}_n - \mathbf{X}_1)$ is constant, independently of values of the conditioning variables, if and only if $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ are multinormal on $\mathbb{R}^k$ [19].

Central Limit Theory on $\mathbb{R}^k$

Central Limit Theorems

The simplest central limit theorem on $\mathbb{R}^k$ is for iid random vectors having finite means and dispersion parameters. Details follow.

Theorem 1. Let $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$ be iid random vectors on $\mathbb{R}^k$ having the finite mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$, and let $\overline{\boldsymbol{X}}_n = n^{-1}(\mathbf{X}_1 + \cdots + \mathbf{X}_n)$. Then, as $n \to \infty$, the limit distribution of $n^{1/2}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})$ is multivariate normal on $\mathbb{R}^k$, i.e. $\mathscr{L}_\infty(n^{1/2}[\overline{\mathbf{X}}_n - \boldsymbol{\mu}]) = \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$.

Limit theorems on $\mathbb{R}^k$ are also known for nonidentical summands and even for certain dependent

vector sequences. An example of the former is the following.

Theorem 2. Let $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$ be a sequence of independent random vectors in $\mathbb{R}^k$ having cdfs $[F_i(\cdot), i = 1, 2, \ldots]$ with finite means and dispersion parameters $[(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2, \ldots]$; let $\overline{\boldsymbol{X}}_n = n^{-1}(\mathbf{X}_1 + \cdots + \mathbf{X}_n)$ and $\overline{\boldsymbol{\mu}}_n = n^{-1}(\boldsymbol{\mu}_1 + \cdots + \boldsymbol{\mu}_n)$. Suppose that, as $n \to \infty$: (i) $n^{-1}\sum_{i=1}^n \boldsymbol{\Sigma}_i \to \boldsymbol{\Sigma} \neq \mathbf{0}$; and (ii) for every $\varepsilon > 0$, $n^{-1}\sum_{i=1}^n \int_{||\mathbf{x}|| > \varepsilon\sqrt{n}} ||\mathbf{x}||^2 \, \mathrm{d}F_i(\mathbf{x}) \to 0$. Then, as $n \to \infty$, the random vector $n^{1/2}(\overline{\mathbf{X}}_n - \overline{\boldsymbol{\mu}}_n) \in \mathbb{R}^k$ **converges** in law to the normal distribution $\mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$ on $\mathbb{R}^k$.

Multivariate normality is thus assured in the limit under the existence of second moments. If third moments are defined, then it is often possible to get bounds on the rate of convergence of standardized vector sums to a multinormal limit. Such bounds are called *Berry–Esseen bounds*, to be considered next.

Berry–Esseen Bounds

To proceed, let $\mathbf{C}$ be the class of all measurable convex sets in $\mathbb{R}^k$, and let $\mathbf{C}_N \subset \mathbf{C}$ be the subclass consisting of continuity sets of some probability measure $P_N(\cdot)$ to be identified. The following lemma is basic; for a proof see [16].

Lemma 1. Let $(\mathbf{X}_1, \ldots, \mathbf{X}_N)$ be an iid sequence in $\mathbb{R}^k$ whose typical member $\mathbf{X} = [X_1, \ldots, X_k]'$ has zero means, the nonsingular dispersion matrix $\boldsymbol{\Sigma}$, and finite absolute third moments $[\mathrm{E}(|X_i|^3) = \beta_{3i}; 1 \leq i \leq k]$. Let $P_N(\cdot)$ be the probability measure associated with $N^{-1/2}(\mathbf{X}_1 + \cdots + \mathbf{X}_N)$, and let $P(\cdot)$ be its multivariate normal limit having zero means and dispersion matrix $\boldsymbol{\Sigma}$. Then, for each $N = 1, 2, \ldots,$

$$\sup_{A \in C} |P_N(A) - P(A)| \leq c(k) \sum_{i=1}^k \frac{\gamma_i^{3/2}\beta_{3i}}{N^{1/2}}, \qquad (2)$$

where $\boldsymbol{\Gamma} = (\gamma_{ij}) = \boldsymbol{\Sigma}^{-1}$ and $c(k)$ is a finite positive constant depending only on k. Moreover, if $A \in C_N$, then

$$\sup_{A \in C_N} |P_N(A) - P(A)| \leq 1.595k^3 \sum_{i=1}^k \frac{\gamma_i^{3/2}\beta_{3i}}{\delta N^{1/2}}, \quad (3)$$

where δ^2 is the ratio of the smallest to the largest eigenvalue of $\boldsymbol{\Sigma}$.

Inequality (2) was given independently by Bhattacharya [4], Sazonov [25], and, in a somewhat different form, by Bergström [3]. Moreover, Bergström [3] has shown that $c(k)$ can be replaced by $c_0 k^3/\delta$, where c_0 is an absolute constant and δ^2 is as defined.

The foregoing results admit numerous applications in statistical inference in assessing the accuracy of large-sample approximate procedures. To fix ideas, consider Pearson's [22] **chi-square test** for **goodness of fit** in testing $\mathrm{H} : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ against general alternatives, where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_k]'$ are the actual and $\boldsymbol{\pi}_0 = [\pi_{10}, \ldots, \pi_{k0}]'$ the hypothetical **multinomial** probabilities. Let X_N^2 be Pearson's statistic based on a sequence of N independent trials, and let $F_N(\cdot)$ be its actual, and $\Psi_\nu(\cdot)$ its limiting cdf – central or noncentral as appropriate. Then Lemma 1 applies directly to the following effect. Suppose that the test is carried out at the nominal level α based on asymptotics, whereas its actual level is α_N, typically unknown. Then lower and upper bounds on the actual level, in terms of the nominal level α and the natural parameters of the problem, as supported by Lemma 1, are given by

$$\alpha - \frac{B(\pi_0)}{N^{1/2}} \leq \alpha_N \leq \alpha + \frac{B(\pi_0)}{N^{1/2}}, \qquad (4)$$

where

$$B(\pi) = c(\nu) \sum_{i=1}^{\nu} \left[\left(\frac{1}{\pi_i}\right) + \left(\frac{1}{\pi_k}\right)\right] \pi_i \times (1 - \pi_i)[1 - 2\pi_i(1 - \pi_i)] \qquad (5)$$

is to be evaluated at π_0, and $\nu = k - 1$. Similar bounds may be found for the actual power of the test in a sequence of N independent trials at a fixed alternative to H. Moreover, these bounds may be evaluated numerically in particular cases using expression (3). For further details see [15].

Similar developments apply in the case of Friedman's [9] test in a two-way analysis of variance based on ranks (*see* **Nonparametric Methods**). Details are given in [16].

There is by now a considerable literature pertaining to multivariate normal distributions, multidimensional limit theory with and without moments, and Berry–Esseen bounds on rates of convergence to a multinormal limit. A lengthy and detailed reference list is omitted here; access is readily available through

searching electronic databases such as the *Current Index to Statistics*, for example.

References

[1] Adrain, R. (1808). Research concerning the probabilities of the errors which happen in making observations, etc., *The Analyst; or Mathematical Museum* **1**, 93–109.

[2] Anderson, T.W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities, *Proceedings of the American Mathematical Society* **6**, 170–176.

[3] Bergström, H. (1969). On the central limit theorem in $\mathbb{R}^k$, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **14**, 113–126.

[4] Bhattacharya, R.N. (1968). Berry–Esseen bounds for the multi-dimensional central limit theorem, *Bulletin of the American Mathematical Society* **74**, 285–287.

[5] Bravais, A. (1846). Analyse mathématique sur la probabilitié des erreurs de situation d'un point, *Mémoires presentés a l'Académie Royale des Sciences, Paris* **9**, 255–332.

[6] Cambanis, S., Huang, S. & Simons, G. (1981). On the theory of elliptically contoured distributions, *Journal of Multivariate Analysis* **11**, 368–385.

[7] Dickson, I.D.H. (1886). Appendix to "Family likeness in stature", by F. Galton, *Proceedings of the Royal Society of London* **40**, 63–73.

[8] Edgeworth, F.Y. (1892). Correlated averages, *Philosophical Magazine, Series 5* **34**, 190–204.

[9] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.

[10] Galton, F. (1888). Co-relations and their measurement chiefly from anthropometric data, *Proceedings of the Royal Society of London* **45**, 134–145.

[11] Gauss, K.F. (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae. Pars Prior and Pars Posterior*. Muster-Schmidt, Göttingen.

[12] Ghurye, S.G. & Olkin, I. (1962). A characterization of the multivariate normal distribution, *Annals of Mathematical Statistics* **33**, 533–541.

[13] Helmert, F.R. (1868). Studien über rationelle Vermessungen, im Gebeite der höheren Geodäsie, *Zeitschrift für Mathematik und Physik* **13**, 73–129.

[14] James, A.T. (1954). Normal multivariate analysis and the orthogonal group, *Annals of Mathematical Statistics* **25**, 40–75.

[15] Jensen, D.R. (1973). Monotone bounds on the chi-squared approximation to the distribution of Pearson's X^2 statistics, *Australian Journal of Statistics* **15**, 65–70.

[16] Jensen, D.R. (1977). On approximating the distributions of Friedman's χ_r^2 and related statistics, *Metrika* **24**, 75–85.

[17] Jensen, D.R. (1984). Ordering ellipsoidal measures: scale and peakedness orderings, *SIAM Journal on Applied Mathematics* **44**, 1226–1231.

[18] Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

[19] Kagan, A.M., Linnik, Y.V. & Rao, C.R. (1965). On a characterization of the normal law based on a property of the sample average, *Sankhya, Series A* **27**, 405–406.

[20] Kagan, A.M., Linnik, Y.V. & Rao, C.R. (1973). *Characterization Problems in Mathematical Statistics*. Wiley, New York.

[21] Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society of London A* **187**, 253–318.

[22] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine, Series 5* **50**, 157–172.

[23] Plana, G.A.A. (1813). Mémoire sur divers problémes de probabilitié, *Mémoires de l'Académie Impériale de Turin* **20**, 355–408.

[24] Rao, C.R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.

[25] Sazonov, V.V. (1968). On the multi-dimensional central limit theorem, *Sankhya, Series A* **30**, 181–204.

[26] Schols, C.M. (1875). Over de theorie der fouten in de ruimte en in het platte vlak, *Verhandelingen der Koninklijke Akademie van Wetenschappen* **15**, 1–75.

[27] Sherman, S. (1955). A theorem on convex sets with applications, *Annals of Mathematical Statistics* **25**, 763–766.

[28] Tong, Y.L. (1980). *Probability Inequalities in Multivariate Distributions*. Academic Press, New York.

[29] Tong, Y.L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

(*See also* **Matrix Algebra**)

D.R. JENSEN

Multivariate Normality, Tests of

The **multivariate normal** density is

$$f(\mathbf{X}) = [(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}]^{-1} \times \exp[-\left(\tfrac{1}{2}\right)(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})],$$

where $\mathbf{X}$ is an m-variate random variable, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the symmetric **covariance matrix**. Most tests for multivariate normality take advantage of properties that are unique to the multivariate normal distribution. One property is the necessary, though not sufficient, condition that the marginal distributions of $\mathbf{X}$ are univariate normal. This permits the use of a univariate test of normality to determine if there is nonnormality in any of the marginals (*see* **Normality, Tests of**). However, **correlation** among the variates results in correlation between the m univariate tests, causing difficulty in determining their joint distribution. D'Agostino [6] suggests a **Bonferroni** approach, i.e. applying each test for marginal normality at the α/m level (*see* **Multiple Endpoints, *P* Level Procedures**).

A second property is that if there exists correlation between any pair of variates, then the relation is strictly linear. A third property is that any linear combination of the variates is normally distributed.

A fourth property of the multivariate normal distribution is that the quadratic form $(\mathbf{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i-\boldsymbol{\mu}) = c$ forms an ellipse of constant probability in m-space. A related property is that the angles between the marginal projection of the n observation vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ onto any of the variate planes and an arbitrary fixed vector through the mean have a **uniform distribution** over the interval $(0, 2\pi)$, denoted $\mathrm{U}(0, 2\pi)$, and these angles are independent of the vector lengths.

Based on these properties, the *scaled* **residuals** from a sample of size n,

$$\mathbf{Z}_i = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}),$$

where $\boldsymbol{\Sigma}^{1/2}$ is the symmetric square root of the covariance matrix, are m-variate standard normal, $\mathrm{N}(\mathbf{0}_m, \mathbf{I}_m)$. The squared **Mahalanobis distance** (*squared radii*)

$$R_i^2 = \mathbf{Z}_i'\mathbf{Z}_i = (\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

is the squared length of the vector from the mean $\boldsymbol{\mu}$ to the observation in m-space relative to the probability ellipses. Under multivariate normality the R_i^2 follow a **chi-square distribution** with m **degrees of freedom** (χ_m^2). More commonly, when the true parameters are unknown, the $\mathbf{Z}_i$ and R_i^2 can be estimated by

$$\mathbf{z}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}), \tag{1}$$

where $\mathbf{S}^{1/2}$ is the symmetric square root of $\mathbf{S}$, and

$$r_i^2 = \mathbf{z}_i'\mathbf{z}_i = (\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})$$

for some efficient estimators $\bar{\mathbf{x}}$ of $\boldsymbol{\mu}$ and $\mathbf{S}$ of $\boldsymbol{\Sigma}$. The $\mathrm{N}(\mathbf{0}_m, \mathbf{I}_m)$ and χ_m^2 only approximate the distributions of the $\mathbf{z}_i$ and r_i^2; the r_i^2 are more closely related to an appropriately parameterized **beta distribution**.

Multivariate analyses are sometimes distinguished by whether they are invariant under linear transformations or dependent upon the original data coordinate system. Therefore, while most of the tests described are affine-invariant, in certain cases tests which are coordinate-dependent may be more pertinent to a specific situation.

Plots

One approach to assessing multivariate normality is to use univariate probability plots to assess each of the marginal variables (*see* **Graphical Displays**). Scatter plots of all variables taken two at a time are an effective way to identify **outliers** and nonlinear relations between variables. A third approach includes ordering the marginal observations independently and plotting the ordered observations against each other taking the variates two at a time. These plots are equivalent to normal probability plots and should follow a linear pattern.

One common type of probability plot, called quantile–quantile (Q–Q) plots, is produced by plotting the **order statistics**, or sorted observations, of a sample on the horizontal axis against the expected values of the order statistics from the reference (e.g. normal) distribution on the vertical axis (*see* **Normal Scores**). Probability paper is also available for some distributions, where the probabilities are scaled to the expected values on the vertical axis. If a sample comes from the reference distribution, the plot should be approximately linear. Healy [9] suggested using probability plots of the r_i^2 relative to a χ_m^2 distribution.

However, especially for smaller samples, the r_i^2 are better approximated by a beta variable of the first kind with parameters $a = m/2$ and $b = (n - m - 1)/2$. In particular, if

$$y = \frac{nr^2}{(n-1)^2},$$

then y has the density given by

$$f(y) = \frac{y^{a-1}(1-y)}{B(a,b)},$$

where $B(a, b)$ is the beta function, given by

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

and $\Gamma(\cdot)$ is the **gamma** function.

In the bivariate case the angles θ_i made by the observation vectors with the x_1 axis are uniform over the interval $(0, 2\pi)$. Therefore, a uniform probability plot of $\theta_i^* = \theta_i/2\pi$ can provide another indication of nonnormality in the data. By defining $u_i = F(r_i^2)$, where F is the χ_2^2 distribution function, a bivariate plot of (u_i, θ_i^*) should be uniform over the unit square, although here also F may be better defined as the appropriate beta distribution. For $m > 2$, one of the $m - 1$ angles made between the projections of the data onto each of the variate planes and, say, the x_1 axis is $U(0, 2\pi)$. The remaining $m - 2$ angles have a distribution proportional to $\sin^{m-1-j}\theta_j$, $0 \le \theta_j \le \pi$, $j = 1, \ldots, m-2$ [1].

Also for the bivariate case, the distribution of the R_i^2 under normality is given by $F(R^2) = 1 - \exp(R^2/2)$, where F is the cdf of the χ_2^2 distribution. Then, the plot of $(R_{(i)}^2, Y_{(i)})$ should fall randomly about the line

$$Y_{(i)} = \log_{10}[1 - F(R_{(i)}^2)] = 0.271R_{(i)}^2 \qquad (2)$$

[11, 22], where the estimate of $Y_{(i)}$ is $\log_{10}[(n - i + 0.5)/n]$. If the r_i^2 are compared with the beta distribution, then the plot of $(R_{(i)}^2, Y_{(i)})$ should approximate the curve

$$Y_{(i)} = 0.217(n-3)\log_e\left[\frac{1 - nr_{(i)}^2}{(n-1)^2}\right], \qquad (3)$$

which reduces to (2) as n gets large. The line described by (3) gradually curves away from the straight line with slope -0.217, with more and earlier curvature for small n.

Easton & McCulloch [7] presented a multivariate Q–Q plot based on matching the observed data with a multivariate reference sample. By assigning the Euclidean distance between two points as the cost function of matching the observed $\mathbf{x}_i$ with $\mathbf{y}_j$ from the reference sample, the "best" matching is found by identifying that permutation of the data, σ^*, in the set P of all permutations which solves

$$\min_{\sigma \in P} \sum_{1}^{n} ||\mathbf{y}_i - \mathbf{x}_{\sigma(i)}||^2.$$

If the data and the reference sample have the same shape, then $\mathbf{X}$ and $\mathbf{Y}$ will have the same shape if there is an $m \times m$ matrix $\mathbf{A}$, an m-vector $\mathbf{b}$, and a permutation σ such that $\mathbf{A}\mathbf{x}_{\sigma(i)} + \mathbf{b} \approx \mathbf{y}$. Given a reference sample, the matching problem can be solved by alternating between optimizing the permutation for fixed $\mathbf{A}$ and $\mathbf{b}$, and then reoptimizing $\mathbf{A}$ and $\mathbf{b}$ for the current permutation σ. They used an assignment **algorithm** to obtain the optimal permutation at each step and a multivariate normal random sample as the reference.

If $\mathbf{x}_i{}^* = \mathbf{A}^*\mathbf{x}_{\sigma^*(i)} + \mathbf{b}^*$ is the best matching of the data to the reference sample, then the first displays to consider are probability plots of each of the components of $\mathbf{x}^*$ vs. $\mathbf{y}$. These plots will have a fuzzy appearance, but the usual deviations from linearity will appear in the presence of outliers, skewness, or heavy/light-tailedness of the data. Isolated points may stand out in the middle of these plots, and large or heterogeneous variability around the 45° line may indicate deviation from normality.

A second display from this matching procedure is a distance Q–Q plot. For this, a second reference sample $\mathbf{U}$ is drawn. Then the n Euclidean distances between $\mathbf{X}^*$ and $\mathbf{Y}$ are plotted against those obtained from matching $\mathbf{U}$ and $\mathbf{Y}$, since the distances should be similar.

Skewness and Kurtosis

Because of the popularity and good power properties of univariate moment tests for normality, it seems only natural that the first tests for assessing multivariate normality would extend the notion of **skewness** and **kurtosis** to a multivariate setting. Small [29] combined the marginal skewness and the marginal kurtosis values to obtain combined skewness and combined kurtosis tests, respectively. Let

$\mathbf{B}_1$ and $\mathbf{B}_2$ be the m-vectors of marginal skewness and kurtosis values, respectively. By applying the Johnson S_u transformations [4], component-wise to $\mathbf{B}_1$ and $\mathbf{B}_2$ as univariate transformations to normality, the transformed vectors $\mathbf{y}(\mathbf{B}_1)$ and $\mathbf{y}(\mathbf{B}_2)$ can be used to obtain

$$Q_1 = \mathbf{y}(\mathbf{B}_1)'\mathbf{U}_1^{-1}\mathbf{y}(\mathbf{B}_1)$$

and

$$Q_2 = \mathbf{y}(\mathbf{B}_2)'\mathbf{U}_2^{-1}\mathbf{y}(\mathbf{B}_2)$$

as test statistics, where $\mathbf{U}_1 = (\hat{\rho}_{ij}^3)$, $\mathbf{U}_2 = (\hat{\rho}_{ij}^4)$, and the $\hat{\rho}_{ij}$ are the sample correlations. Q_1 and Q_2 are each approximated by a χ_m^2 distribution, and an omnibus test $Q = Q_1 + Q_2$ is approximately χ_{2m}^2.

Mardia [18] presented sample estimates of multivariate skewness and kurtosis, calculated from generalized versions of the squared radii,

$$r_{ij} = \mathbf{z}_i'\mathbf{z}_j = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \overline{\mathbf{x}}).$$

Note that, using this notation, $r_{ii} = r_i^2$. Mardia's skewness and kurtosis measures are, respectively,

$$b_{1,m} = n^{-2}\sum_{i,j=1}^{n} r_{ij}^3$$

and

$$b_{2,m} = n^{-1}\sum_{1}^{n} r_{ii}^2 = n^{-1}\sum_{1}^{n}(r_i^2)^2.$$

Under multivariate normality the exact moments of the two tests are

$$\mathrm{E}(b_{1,m}) = \frac{m(m+2)[(n+1)(m+1)-6]}{(n+1)(n+3)}$$

with unknown variance, while

$$\mathrm{E}(b_{2,m}) = \frac{m(m+2)(n-1)}{(n+1)}$$

$$\mathrm{var}(b_{2,m}) = \frac{8m(m+2)(n-3)(n-m-1)(n-m+1)}{(n+1)^2(n+3)(n+5)}$$

Mardia [19]. Mardia & Zemroch [21] gave a FORTRAN subroutine for calculating $b_{1,m}$ and $b_{2,m}$.

Mardia & Foster [20] presented several omnibus tests based on combinations of $b_{1,m}$ and $b_{2,m}$. $C_w^2 = \mathbf{w}^{\mathrm{t}}\mathbf{W}^{-1}\mathbf{w}$ accounts for correlation between $b_{1,m}$ and $b_{2,m}$, where the vector $\mathbf{w}$ has elements $b_{1,m}$ and $b_{2,m}$ under transformations to normality and $\mathbf{W}$ is the correlation matrix associated with $\mathbf{w}$; entries of $\mathbf{w}$ are

$$w(b_{1,m}) = \frac{1}{6}(2f)^{1/2}6\left[\frac{4nf^2}{3}b_{1,m}\right]^{1/3} - 18f + 4$$

and

$$w(b_{2,m}) = 3\left[\frac{f_1}{2}\right]^{1/2} \times \left(\frac{(1-2)/f_1}{1 + s[2/(f_1-4)^{1/2}]^{1/3} + 2/9f_1 - 1}\right),$$

where

$$s = \frac{b_{2,m} - \mathrm{E}(b_{2,m})}{[\mathrm{var}(b_{2,m})]^{1/2}},$$

$$f = \frac{m(m+1)(m+2)}{6},$$

and

$$f_1 = 6 + [8m(m+2)(m+8)^{-2}]^{1/2} \times \sqrt{n}\{[m(m+2)/2]^{1/2}/(m+8)(n)^{1/2} + [1 + 0.5nm(m+2)/(m+8)^2]^{1/2}\}.$$

$\mathbf{W}$ is a matrix with 1s on the diagonal and $c = \mathrm{cov}[W(b_{1,m}), W(b_{2,m})]$ as the off-diagonal elements, where

$$c = (f_1/16f) - (40/9)(1 - 2/f_1) \times [1/(f_1-4)] + (n/3\sigma)(1-2/f_1)^{1/3} \times [2/(f_1-4)^{1/2}]\mathrm{cov}(b_{1,m}b_{2,m}) + \cdots$$

by Taylor expansion and $\sigma^2 = \mathrm{var}(b_{2,m})$. Alternatively, the omnibus test, $S_w^2 = W(b_{1,m}) + W(b_{2,m})$, can be used.

Two other omnibus tests are based on a normal approximation to a χ^2 variable for $b_{1,m}$,

$$\mathrm{U}(b_{1,m}) = \frac{n(b_{1,m} - 6f/n)}{(72f)^{1/2}}$$

and a transformation to normality of $b_{2,m}$,

$$\mathrm{U}(b_{2,m}) = \frac{(n)^{1/2}[b_{2,m} - m(m+2)(n-1)/(n+1)]}{[8m(m+2)]^{1/2}}.$$

The omnibus tests are

$$S_n^2 = \mathrm{U}^2(b_{1,m}) + \mathrm{U}^2(b_{2,m})$$

and

$$C_n^2 = \mathbf{b}'\mathbf{V}^{-1}\mathbf{b},$$

where

$$\mathbf{b}' = [b_{1,m} - 6f/n, \quad b_{2,m} - m(m+2)(n-1)/(n+1)]$$

and

$$\mathbf{V} = \begin{bmatrix} 72f/n^2 & 12m(8m^2 - 13m + 23)/n^2 \\ 12m(8m^2 - 13m + 23)/n^2 & 8m(m+2)/n \end{bmatrix}.$$

These four omnibus tests are approximately χ_2^2 under the null hypothesis.

Malkovich & Afifi [17] defined the distribution of a random vector $\mathbf{X}$ to have multivariate skewness if

$$\beta_1(\mathbf{C}) = \frac{\mathrm{E}\{[\mathbf{C}'\mathbf{X} - \mathbf{C}'\mathrm{E}(\mathbf{X})]^3\}^2}{\mathrm{var}(\mathbf{C}'\mathbf{X})^3} > 0$$

for some vector $\mathbf{C}$; without loss of generality, we can assume that $\mathbf{C}'\mathbf{C} = 1$. Similarly, multivariate kurtosis was defined as

$$\beta_2(\mathbf{C}) = \frac{\mathrm{E}[\mathbf{C}'\mathbf{X} - \mathbf{C}'\mathrm{E}(\mathbf{X})]^4}{\mathrm{var}(\mathbf{C}'\mathbf{X})^2} \neq 3$$

for some vector $\mathbf{C}$.

They derived ${b_1}^*$ and $({b_2}^*)^2$, based on

$$b_{1,y} = \frac{n\left[\sum_1^n (y_j - \overline{y})^3\right]^2}{\left[\sum_1^n (y_j - \overline{y})^2\right]^3}$$

and

$$b_{2,y} = \frac{n\sum_1^n (y_j - \overline{y})^4}{\left[\sum_1^n (y_j - \overline{y})^2\right]^2},$$

with $y_i = \mathbf{C}'\mathbf{x}_i$; the hypothesis of no multivariate skewness is accepted if

$${b_1}^* = \max_{\mathbf{C}} b_{1,y}(\mathbf{C}) \leq K_1.$$

The hypothesis of no multivariate kurtosis is accepted if

$$({b_2}^*)^2 = \max_{\mathbf{C}}[b_{2,y}(\mathbf{C}) - K]^2 \leq K_2,$$

where K and K_2 are appropriate constants. Since kurtosis is not symmetrically distributed, K and K_2 should be chosen to weight the minimum and maximum values (over all $\mathbf{C}$) of $b_{2,y}$, evenly, so that the probabilities of finding a significant low or high value of kurtosis when the null hypothesis is true are each $\alpha/2$; as the sample size gets large, K converges to 3. For computational purposes, let

$$\mathbf{A} = \sum_1^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \tag{4}$$

and $(\mathbf{A}^*)'\mathbf{A}\mathbf{A}^* = \mathbf{I}$, and define

$$y_j = (\mathbf{A}^*)'(\mathbf{x}_j - \overline{\mathbf{x}}), \quad j = 1, \ldots, n.$$

Then

$${b_1}^* = \max_{\mathbf{C}'\mathbf{C}=1} n\left[\sum_1^n (\mathbf{C}'\mathbf{y}_i)^3\right]^2$$

and

$$({b_2}^*)^2 = \max_{\mathbf{C}'\mathbf{C}=1} n\left[\sum_1^n (\mathbf{C}'\mathbf{y}_i)^4 - K\right]^2.$$

Computational details for the iterative method of calculating ${b_1}^*$ and $({b_2}^*)^2$ are given in Malkovich [16].

Under certain parametric restrictions, the multivariate **Pearson distribution** reduces to the multivariate normal. Bera & John [3] used Rao's score principle [25] (*see* **Likelihood**) to test those restrictions, in principle developing tests for Pearson alternatives. They defined

$$T_j = \sum_{i=1}^n \frac{z_{ij}^3}{n},$$

$$T_{jj} = \sum_{i=1}^n \frac{z_{ij}^4}{n},$$

and

$$T_{jk} = \sum_{i=1}^n \frac{z_{ij}^2 z_{ik}^2}{n},$$

where the z_{ij} are the jth component of the scaled residual $\mathbf{z}_i$, given by (1). T_j, T_{jj}, and T_{jk} are asymptotically independent and normal with means 0, 3, and

1 and variances $6/n$, $24/n$, and $4/n$, respectively. T_j is the univariate skewness test and T_{jj} is the univariate kurtosis test for the jth component of $\mathbf{z}$. Because of consistency conditions, they recommend first testing the m T_j values, using

$$C_1 = n \sum_1^m \frac{T_j^2}{6}$$

as a test for skewness, which is asymptotically χ^2_m. If C_1 is not significant, then test

$$C_2 = n \left[\frac{1}{24} \sum_1^m (T_{jj} - 3)^2 + \frac{1}{4} \sum_{j=1}^m \sum_{k=1}^{j-1} (T_{jk} - 1)^2 \right],$$

which is approximately $\chi^2_{m(m+1)/2}$. Omnibus tests are

$$C_3 = n \left[\frac{1}{6} \sum_1^m T_j^2 + \frac{1}{24} \sum_1^m (T_{jj} - 3)^2 \right]$$

and

$$C_4 = C_1 + C_2,$$

which can be tested using χ^2_{2m} and $\chi^2_{m(m+3)/2}$, respectively.

Koziol [13, 14] considered multivariate tests based on the theory of Neyman's smooth tests (*see* **Chi-square Tests**). His smooth test for skewness is algebraically equivalent to Mardia's $b_{1,m}$,

$$\hat{U}_3^2 = \frac{n b_{1,m}}{6},$$

while $\hat{U}_4^2$, the smooth test for kurtosis, contains $b_{2,m}$ as one of its components. Koziol [15] proposed

$$\tilde{b}_{2,m} = n^{-2} \sum_1^n \sum_1^n r_{ij}^4$$

and showed that

$$24n\hat{U}_4^2 = n^2 \tilde{b}_{2,m} - 6n^2 b_{2,m} + 3n^2 m(m+2).$$

Since the distribution of $\hat{U}_4^2$ is less complex than that of $\tilde{b}_{2,m}$, $\tilde{b}_{2,m}$ is only used to calculate the former statistic. $\hat{U}_3^2$ and $\hat{U}_4^2$ are asymptotically independent χ^2 variables with $\binom{p+2}{3}$ and $\binom{p+3}{4}$ df, respectively, under the null hypothesis.

Regression and Correlation Tests

The popularity and good power properties of regression and correlation tests for univariate normality suggested the extension of those methods to the multivariate case. Royston [27] proposed combinations of marginal Shapiro–Wilk W tests (*see* **Normal Scores**) for assessing multivariate normality. Using Royston's [26] **transformation** of W to normality, the test for each marginal is calculated using

$$z_i = \frac{(1 - W_i)^\lambda - \mu}{\sigma},$$

where λ, μ, and σ are functions of n. The values

$$k_i = \{\Phi^{-1}[(\tfrac{1}{2})\,\Phi(-z_i)]\}^2$$

are each approximately distributed as a χ^2_1 random variable. If the m variables were uncorrelated, then $G = \sum_1^m k_i/m \sim \chi^2_m/m$; at the other extreme, if the variates were perfectly correlated, then $G \sim \chi^2_1$. For intermediate correlations, Royston used $G \sim \chi^2_e/e$, where

$$e = \frac{m}{1 + (m-1)\overline{c}}$$

and

$$\overline{c} = \sum_1^m \sum_1^m \frac{\hat{c}_{ij}}{m^2 - m}.$$

The $\hat{c}_{ij}$ are estimates of the correlation between the k_i, which for $10 \le n \le 2000$ are calculated as

$$\hat{c}_{ij} = r_{ij}^5 \left[\frac{1 - 0.715(1 - r_{ij})^{0.715}}{0.35v} \right].$$

Malkovich & Afifi [17] presented a multivariate Shapiro–Wilk criterion, where multivariate normality is accepted if

$$\min_{\mathbf{D}} W_{\mathbf{D}} \ge K_W,$$

where $W_{\mathbf{D}}$ is the univariate Shapiro–Wilk test for the observations reduced by $z_i = \mathbf{D}'\mathbf{x}_i$. The vector $\mathbf{D}$ which gives a lower bound is given by

$$\mathbf{D}'(\mathbf{x}_1 - \overline{\mathbf{x}}) = \frac{(n-1)}{(na_1)},$$

$$\mathbf{D}'(\mathbf{x}_j - \overline{\mathbf{x}}) = -(na_1)^{-1}, \quad j > 1.$$

An approximate solution can be found by using

$$\mathbf{D} = a_1^{-1} \mathbf{A}^{-1} (\mathbf{x}_1 - \overline{\mathbf{x}}),$$

where a_1 is the first coefficient for the Shapiro–Wilk test and $\mathbf{A}$ is the matrix (4). Since any observation can be designated as x_1, the generalized statistic W^* can be obtained by identifying $\mathbf{x}_k$ as that observation for which

$$(\mathbf{x}_k - \overline{\mathbf{x}})'\mathbf{A}^{-1}(\mathbf{x}_k - \overline{\mathbf{x}}) = \max_{1 \le i \le n} (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{A}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}). \tag{5}$$

Then the order statistics $G_{(i)}$ are found, where

$$G_i = (\mathbf{x}_k - \overline{\mathbf{x}})'\mathbf{A}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$$

and the univariate Shapiro–Wilk test is applied to the $G_{(i)}$,

$$W^* = \frac{\left(\sum_1^n a_{n,j} G_{(j)}\right)^2}{(\mathbf{x}_k - \overline{\mathbf{x}})'\mathbf{A}^{-1}(\mathbf{x}_k - \overline{\mathbf{x}})}.$$

Of the set of n vectors used in (5), Fattorini [8] proposed using the vector $\mathbf{C}_M$ which minimized the Shapiro–Wilk statistic, so that $W_F{}^* \le W^*$.

Royston [27] suggested using either a χ^2 or beta distribution to transform the r_i^2 to approximate normality by

$$r_i' = \Phi^{-1}[F(r_i^2)],$$

where F is the selected cdf. Using the Shapiro–Wilk W and the normality transformation, Royston proposed the Ω test for m-normality. He further proposed examining all subsets of the m variates of size k, $k = 1, \ldots, m$. Each value of k gives $K = \binom{m}{k}$ non-independent tests $\Omega_1, \ldots, \Omega_k$. These tests may be inspected individually or further combined into a single test for each value of k:

$$\theta_k = \sum_1^k \{\Phi^{-1}[(\tfrac{1}{2})\,\Phi(-\Omega_i)]\}^2,$$

where $\theta_k \sim \chi_k^2$.

Tsai & Koziol [30] proposed a multivariate Shapiro–Francia [28] correlation test,

$$r_{m;n} = \frac{\sum_1^n (r_{(i)}^2 - \overline{r}^2)(Q_i - \overline{Q})}{\left[\sum_1^n (r_{(i)}^2 - \overline{r}^2)^2 \sum_1^n (Q_i - \overline{Q})^2\right]^{1/2}},$$

where the Q_i are the expected values of the χ_m^2 order statistics. Small values of $r_{m;n}$ indicate deviation from normality.

Other Tests

Univariate empirical distribution tests have been suggested for use with the $r_{(i)}^2$ (*see* **Goodness of Fit**; **Kolmogorov–Smirnov Test**). These include the Anderson–Darling A^2 and the Cramér–von Mises test [24], given by

$$J_n = \frac{1}{12n} + \sum_1^n \left[\frac{u_{(i)} - (i - 0.5)}{n}\right]^2,$$

where $u_{(i)} = F_m(r_{(i)})^2$, F_m being the χ_m^2 distribution function. A test of the uniformity of the angles θ_i using Rayleigh's test [12] is obtained by letting

$$\hat{\mathbf{l}}_i = r_i^{-1}\mathbf{z}_i.$$

Rayleigh's statistic is

$$\hat{\mathbf{R}} = n^{-1/2} \sum_1^n \hat{\mathbf{l}}_i.$$

$\hat{\mathbf{R}}$ is normal with mean vector $\mathbf{0}$ and covariance given by $\mathbf{V} = v\mathbf{I}$ with

$$v = m^{-1}[1 - 2/m(\{\Gamma[(m+1)/2]/\Gamma(m/2)\})^2].$$

$R_y = \hat{\mathbf{R}}'\mathbf{V}^{-1}\hat{\mathbf{R}}$ can be compared to a χ_m^2 distribution to obtain probability levels. Since they are independent, if $p_1 = \Pr(x > J_n)$ and $p_2 = \Pr(x > R_y)$, where the probabilities are obtained for the observed test statistics J_n and R_y, then $-2(\log p_1 + \log p_2) \sim \chi_4^2$.

Ward [31] proposed estimating the cumulative distribution function of $\mathbf{X}$,

$$y_i = \hat{F}(\mathbf{x}_i) = \prod_{j=1}^m \Phi(z_{ij}),$$

and using the Kolmogorov–Smirnov D or Anderson–Darling A^2 test to test the goodness of fit of the y_i to the density

$$g(y) = \frac{(-\log y)^{m-1}}{\Gamma(m)}, \quad 0 < y < 1.$$

Here, the y_i have an inherently different ordering from the r_i^2: while the r_i^2 are minimized at the point $\mathbf{x} = \overline{\mathbf{x}}$, y_i tends towards its minimum value of 0 as each of the m variates goes away from $\overline{\mathbf{x}}$.

The χ^2 test is adaptable to any null distribution, including those that are multivariate in nature. As in the univariate case, cells must be defined and the expected and observed numbers of observations found in each must be ascertained. The problems with the univariate χ^2 test, however, must also be addressed in the multivariate setting, i.e. cell size and number of cells.

For the bivariate case, Kowalski [11] used the χ_2^2 distribution to determine cell sizes. If $2c^2 = (\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ is an ellipse of constant probability based on a multivariate normal distribution, then the volume of a ring between the ellipses defined by $2c^2$ and $2(c + dc)^2$ is

$$V = \exp(-c^2) - \exp[-(c + dc)^{-2}],$$

and nV observations would be expected to occur within the ring. Comparison of expected with observed of observations within rings can then be made using the standard χ^2 test. Mason & Young [22] used the beta approximation for the r_i^2,

$$W_c = \Pr(r_i^2 < 2c^2) = 1 - \left[\frac{1 - 2c^2 n}{(n-1)^2}\right]^{(n-3)/2}.$$

To obtain rings of equal size, the approximate relationship

$$2c^2 = n\left[1 - \frac{1-i}{k}\right]^{2/(n-3)}$$

can be used for a specified number, k, of cells and $n(W_{c+dc} - W_c)$ observations will be within the ring defined by c and $c + dc$. Moore & Stubblebine [23] extended Kowalski's χ^2 test to a general dimension m.

Cox & Small [5] proposed pairwise testing for linearity between components of a multivariate distribution using Q_{ij}, the **Student's *t* statistic** of significance for the coefficient of x_j^2 when x_i is regressed on x_j and x_j^2. For the purpose of symmetry, the joint statistic $(Q_{ij}, Q_{j,i})$ is used. Then (for large samples),

$$\max(|Q_{i,j}|, |Q_{j,i}|)$$

can be referred to tables of the bivariate normal distribution, or

$$\mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q}$$

can be used as a χ_2^2 test, where $\mathbf{Q} = [Q_{i,j}\ Q_{j,i}]$,

$$\mathbf{R} = \begin{bmatrix} 1 & \tilde{\rho}_{ij}(2 - 3\tilde{\rho}_{ij}^2) \\ \tilde{\rho}_{ij}(2 - 3\tilde{\rho}_{ij}^2) & 1 \end{bmatrix},$$

and $\tilde{\rho}_{ij}$ is the observed correlation between the components. An alternative is a regression of each component x_i on all other components x_k and x_j^2. From the $m(m-1)$ regressions, the Q values may be ordered and plotted on a normal probability plot, provided the sample size is sufficiently large.

Andrews et al. [1] projected the data along directions which are chosen to be sensitive to particular types of nonnormality. Since nonnormality in the data may result in nonnormal clustering of points, the vector

$$\mathbf{d}_\alpha = \frac{\sum_1^n w_i \mathbf{z}_i}{\left\| \sum_1^n w_i \mathbf{z}_i \right\|}$$

may be used to point to these clusters, where $w_i = ||\mathbf{z}_i||^\alpha$ and α is a constant to be chosen which will determine the region of sensitivity. In particular, if $\alpha < 0$, then d_α points in the direction of nonnormal clusters near the mean, while for $\alpha > 0$, d_α points to clusters far from the mean. The observations can be projected onto the direction identified by ${d_\alpha}^* = S^{1/2} d_\alpha$, and the lengths of the projections ${d_\alpha}^*$ will, under the null hypothesis, form a univariate normal sample which can be tested using any univariate test for normality.

Andrews et al. [2] proposed the nearest distance test for ascertaining joint normality. The initial step consists of transforming the data to the unit hypercube by using the $\mathbf{z}_i$ and calculating the vector $\mathbf{y}_i$, where the entries are defined by $y_{ij} = \Phi(z_{ij})$. After calculating the distances

$$d(i, i') = \max_k \left[\min(|y_{ki} - y_{ki'}|, ||y_{ki} - y_{ki'}| - 1|)\right],$$

the nearest distance is found:

$$d_{\min} = \min_{i' \neq i} d(i, i').$$

These distances are further transformed to standard normal deviates: for each y_i let

$$w_i = \Phi \frac{1 - \exp\{-n[2d_{\min}(i)]^p\}}{1 - \exp(-1)},$$

if and only if $d_{\min}(i) < 1/2n^{1/p}$ and $d(i, i') > 1/2n^{1/p}$, $i' < i$. Under the null hypothesis, the transformed distances are independent of the coordinates from which they are measured; this independence can be measured using multiple regression. For all of the $n' \leq n$ points that follow these two conditions, fit the regression model

$$w_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \sum_{j=1}^{m} \sum_{k=1}^{m} \beta_{jk} x_{ij} x_{ik}.$$

The regression sum of squares should be compared with a $\chi^2_{(p+1)(p+2)/2}$.

Henze & Zirkler [10] presented a class of invariant consistent tests for composite multivariate normality, based on the weighted integral of the difference between the empirical characteristic function and its pointwise limit. The test is given by

$$T_\beta = n(4I_{(\mathbf{S} \text{ singular})} + D_{n,\beta} I_{(\mathbf{S} \text{ nonsingular})}),$$

where I is the indicator function and

$$D_{n,\beta} = n^{-2} \sum_{j,k=1}^{n} \exp\left[-\frac{\beta^2}{2} ||\mathbf{z}_j - \mathbf{z}_k||^2\right] + (1 + 2\beta^2)^{-m/2} - 2(1 + \beta^2)^{-m/2} n^{-1} \times \sum_{j=1}^{n} \exp\left\{\frac{-\beta^2}{[2(1 + \beta^2)]r_j^2}\right\},$$

where

$$||\mathbf{z}_j - \mathbf{z}_k||^2 = (\mathbf{x}_j - \mathbf{x}_k)' \mathbf{S}^{-1} (\mathbf{x}_j - \mathbf{x}_k)$$

and

$$\beta = \frac{1}{\sqrt{2}} \left[\frac{n(2m + 1)}{4}\right]^{1/(m+4)}.$$

T_β rejects the null hypothesis when the test value is too large. When $\mathbf{S}$ is singular, $D_{n,\beta}$ is undefined, so T_β is set to its maximum value of 4, causing rejection of the null hypothesis.

Recommendations

Since there are many types of departures from multivariate normality, a single best test may not exist. A multivariate test may dilute the effects of a subset of nonnormal components, while marginal tests may miss departures in multivariate combinations of variables. Because many of the testing procedures are difficult to program and are not included in available statistical packages, and critical values are not always available, and because of the lack of definitive power studies, it is difficult to evaluate the relative usefulness of the tests. However, the Bera & John [3] tests and Mardia's skewness, kurtosis, and omnibus tests seem to have relatively high **power** against a variety of alternatives. The T_β tests, and in particular the parameterization with $\beta = 0.5(T_{0.5})$, also seem to have good power properties over a wide variety of alternatives [10].

References

[1] Andrews, D.F., Gnanadesikan, R. & Warner, J.L. (1973). *Methods for assessing multivariate normality, in Multivariate Analysis*, 3rd Ed., P.R. Krishnaiah, ed. Academic Press, New York, pp. 95–116.

[2] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, New Jersey.

[3] Bera, A. & John, S. (1983). Tests for multivariate normality with Pearson alternatives, *Communications in Statistics – Theory and Methods* **12**, 103–117.

[4] Bowman, K.O. & Shenton, L.R. (1975). Omnibus test contours for departures from normality based on b_1 and b_2, *Biometrika* **62**, 243–250.

[5] Cox, D.R. & Small, N.J.H. (1978). Testing multivariate normality, *Biometrika* **65**, 263–272.

[6] D'Agostino, R.B. (1986). Tests for the normal distribution, in *Goodness of Fit Techniques*, R.B. D'Agostino & M.A. Stephens, eds. Marcel Dekker, New York, pp. 367–419.

[7] Easton, G.S. & McCulloch, R.E. (1990). A multivariate generalization of quantile-quantile plots, *Journal of the American Statistical Association* **85**, 376–386.

[8] Fattorini, L. (1986). Remarks on the use of the Shapiro-Wilk statistic for testing multivariate normality, *Statistica* **46**, 209–217.

[9] Healy, M.J.R. (1968). Multivariate normal plotting, *Applied Statistics* **17**, 157–161.

[10] Henze, N. & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality, *Communications in Statistics – Theory and Methods* **19**, 3595–3617.

[11] Kowalski, C.J. (1970). The performance of some rough tests for bivariate normality before and after coordinate transformations to normality, *Technometrics* **12**, 517–544.

[12] Koziol, J.A. (1983). On assessing multivariate normality, *Journal of the Royal Statistical Society, Series B* **45**, 358–361.

[13] Koziol, J.A. (1986). Assessing multivariate normality: a compendium, *Communications in Statistics – Theory and Methods* **15**, 2763–2783.

[14] Koziol, J.A. (1987). An alternative formulation of Neyman's smooth goodness of fit tests under composite alternatives, *Metrika* **34**, 17–24.

[15] Koziol, J.A. (1989). A note on multivariate kurtosis, *Biometrical Journal* **31**, 619–624.

[16] Malkovich, J.F. (1971). Tests for Multivariate Normality, PhD Dissertation. University Microfilms, Ann Arbor.

[17] Malkovich, J.F. & Afifi, A.A. (1973). On tests for multivariate normality, *Journal of the American Statistical Association* **68**, 176–179.

[18] Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications, *Biometrika* **57**, 519–530.

[19] Mardia, K.v. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies, *Sankhyā* **36**, 115–128.

[20] Mardia, K.V. & Foster, K. (1983). Omnibus tests of multinormality based on skewness and kurtosis, *Communications in Statistics – Theory and Methods* **12**, 207–221.

[21] Mardia, K.V. & Zemroch, P.J. (1975). Algorithm AS 84. Measures of multivariate skewness and kurtosis, *Applied Statistics* **24**, 262–265.

[22] Mason, R.L. & Young, J.C. (1985). Re-examining two tests for bivariate normality, *Communications in Statistics – Theory and Methods* **14**, 1531–1546.

[23] Moore, D.S. & Stubblebine, J.B. (1981). Chi-square tests for multivariate normality with application to common stock prices, *Communications in Statistics – Theory and Methods* **10**, 713–738.

[24] Paulson, A.S., Roohan, P. & Sullo, P. (1987). Some empirical distribution function tests for multivariate normality, *Journal of Statistical Computation and Simulation* **28**, 15–30.

[25] Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* **44**, 50–55.

[26] Royston, J.P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples, *Applied Statistics* **31**, 115–124.

[27] Royston, J.P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W, *Applied Statistics* **32**, 121–133.

[28] Shapiro, S.S. & Francia, R.S. (1972). Approximate analysis of variance test for normality, *Journal of the American Statistical Association* **67**, 215–216.

[29] Small, N.J.H. (1980). Marginal skewness and kurtosis in testing multivariate normality, *Applied Statistics* **29**, 85–87.

[30] Tsai, K.-T. & Koziol, J.A. (1988). A correlation procedure for assessing multivariate normality, *Communications in Statistics – Simulation and Computation* **17**, 637–651.

[31] Ward, P.J. (1988). Goodness of Fit Tests for Multivariate Normality, Ph.D. Dissertation. University of Alabama, Tuscaloosa.

HENRY C. THODE, JR

Multivariate Outliers

Outlier *rejection* based on an appropriate *test of discordancy* is just one prospect; alternatives include *accommodation* (outlier-robust) procedures, which limit the effects of possible contamination on inferring properties of the principal uncontaminated data source, or *identification* of the nature of the contamination as an interest in its own right. Such approaches need appropriate *outlier models* to reflect the possible uncontaminated or contaminated forms of the underlying population.

This same range of possible actions (rejection, accommodation, identification), of procedures and of models applies to *multivariate* also, but with a crucial difference. For univariate data outliers are extremes: they "stick out at the ends of the sample". For multivariate data, there is no natural concept of an extreme, and *outlier-detection procedures* will be needed to reveal outlying data points.

Multivariate data may contain outliers, of course. They appear as observations lying well out on the periphery of the data cloud. But what is "well out"? We need appropriate formalizations of this notion for *detecting* multivariate outliers before examining the range of models procedures for the statistical analysis of multivariate outliers (a comprehensive review is provided in Chapter 7 of [7]).

Interest in **outliers** (or "spurious observations" or "mavericks") goes back to the origins of statistical enquiry. As observations in a univariate sample, they are not only extreme but are "extremely extreme" – they trigger concern for whether they are truly part of the population under investigation or reflect "contamination" from some other low-incidence source. Early preoccupation with rejecting such observations "to restore the integrity of the sample", or of steadfastly retaining them, so as not to "distort the message of the data" has been replaced in modern times by more sophisticated considerations.

Thus, whilst *rejection* based on an appropriate statistical procedure (a *test of discordancy*) is still a prospect, alternatives include *accommodation* (outlier-robust) procedures which limit the effects of possible contamination on inferring properties of the principal uncontaminated data source, or *identification* of the nature of the contamination as an interest in its own right. Such approaches need to be based on appropriate *outlier models* to reflect the possible uncontaminated or contaminated forms of the underlying population and its distributional characteristics. Such attitudes and approaches to outlier study provide a sound basis for handling outliers (see, for example, Barnett [3]) and for the vast array of specific statistical inference procedures which have been developed for studying outliers in recent times. Barnett & Lewis [7] provide a comprehensive review of the field, not only for basic univariate data, but for more structured relational models such as linear models (**regression** and designed experiments), **time series**, directional data (*see* **Circular Data Models**), **categorical data**, and general multivariate data (*see* **Diagnostics**; **Multivariate Analysis, Overview**).

In all these cases the ranges of possible actions (rejection, accommodation, identification), of procedures, and of models are essentially similar to those for univariate data. A crucial difference, however, is in the indication provided by the data of the presence of outlying observations. For univariate data the outliers are extreme values: they "stick out at the ends of the sample". In more structured or higher-dimensional data the stimulus is less obvious. Outliers in regression data may be detected by extreme residuals as distinctly breaking the pattern of relationship. For general multivariate data there is no natural concept of an extreme, and *outlier detection procedures* will be needed to reveal outlying data points.

The fact that multivariate data may contain outliers is not in dispute. They are often clearly signaled by observations lying well out on the periphery of the data cloud. See, for example, the observations marked A and B on the two-dimensional scatter plot shown in Figure 1. But what do we mean by "lying well out"? We need to consider appropriate formalizations of this notion for *detecting* multivariate outliers, and then proceed to examine the range of procedures for the statistical analysis of multivariate outliers (a comprehensive review is provided in [7, Chapter 7].

Principles for Multivariate Outlier Detection

We need some way of essentially "ordering" the multivariate data. No natural unambiguous ordering principle is possible in more than one dimension; but progress can be made using more modest *subordering principles*. Barnett [2] categorizes these in four

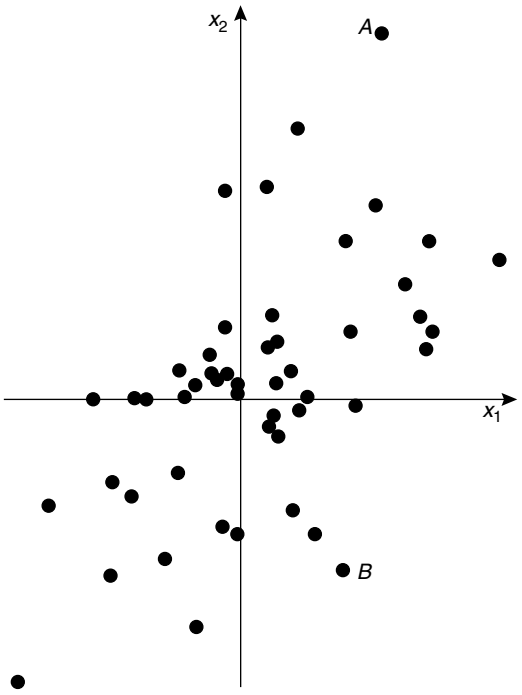

Figure 1 A bivariate sample

types: *marginal, reduced* (or *aggregate*), *partial*, and *conditional*. For outlier study, *reduced subordering* is almost the only principle that has been employed.

With reduced subordering we transform any multivariate observation $\mathbf{x}$, of dimension p, to a scalar quantity $R(\mathbf{x})$. We can then order a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in terms of the values $R_j = R(\mathbf{x}_j)$, $j = 1, 2, \ldots, n$. That observation $\mathbf{x}_i$ which yields the maximum value $R_{(n)}$ is then a candidate for declaration as an outlier – provided its extremeness is surprising relative to the basic model F. Specifically, an outlier $\mathbf{x}_i$ will be adjudged *discordant* if $R_{(n)}$ is unreasonably (statistically) large in relation to the distribution of $R_{(n)}$ under F. Thus the principle of a *test of discordance* is the same as it was for a univariate outlier.

However, many problems now arise which are specific to the multivariate case. Clearly, we may lose useful information on multivariate structure by employing reduced (or any other form of) subordering. So how are we to choose the reduction measure $R(\mathbf{x})$? Subjective choice is fraught with danger; multivariate data do not reveal their structure readily (or reliably) to casual observation.

Barnett [4] has considered general principles for the detection of multivariate outliers. He proposes two possibilities as follows.

Principle A

The most extreme observation is the one, $\mathbf{x}_i$, whose omission from the sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ yields the largest incremental increase in the maximized **likelihood** under F for the remaining data. If this increase is surprisingly large, declare $\mathbf{x}_i$ to be an outlier.

This principle requires only the basic model F to be specified. If we are prepared to adopt an alternative (contamination) model $\overline{F}$, e.g. of *slippage type* with one contaminant, we can set up a more sophisticated principle, as follows.

Principle B

The most extreme observation is the one, $\mathbf{x}_i$, whose assignment as the contaminant in the sense of $\overline{F}$ maximizes the difference between the log likelihoods of the sample under F and $\overline{F}$. If this difference is surprisingly large, declare $\mathbf{x}_i$ to be an outlier.

Such principles have been applied: Barnett & Lewis [7, Section 7.3] discuss applications to **multivariate normal, exponential,** and **Pareto** models. Often, however, the reduction metric $R(\mathbf{x})$ is chosen in an ad hoc (if intuitively supported) manner. For example, it is common to represent a multivariate observation $\mathbf{x}$ by means of a *distance measure*,

$$R(\mathbf{x}; \mathbf{x}_0, \boldsymbol{\Gamma}) = (\mathbf{x} - \mathbf{x}_0)' \boldsymbol{\Gamma}^{-1} (\mathbf{x} - \mathbf{x}_0),$$

where $\mathbf{x}_0$ reflects the location of the underlying distribution and $\boldsymbol{\Gamma}^{-1}$ applies a differential weighting to the components of the multivariate observation inversely related to their scatter or to the population variability. For example, $\mathbf{x}_0$ might be the zero vector, the true mean $\boldsymbol{\mu}$, or the sample mean $\bar{\mathbf{x}}$, and $\boldsymbol{\Gamma}$ might be the **covariance matrix** $\mathbf{V}$ or its sample equivalent $\mathbf{S}$, depending on the state of our knowledge about $\boldsymbol{\mu}$ and $\mathbf{V}$.

If the basic model F were multivariate normal, $\mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, the corresponding form,

$$R(\mathbf{x}; \boldsymbol{\mu}, \mathbf{V}) = (\mathbf{x} - \boldsymbol{\mu})' \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu}),$$

has substantial practical appeal in terms of probability density ellipsoids and turns out to have much broader statistical support, including accord with *Principle A*. In fact, the sample shown in Figure 1 is from a

bivariate normal distribution, and the appropriate elliptic density contours in Figure 2 highlight any intuitive concern we had for the observations A and B. For other distributions, $R(\mathbf{x}; \mathbf{x}_0, \mathbf{\Gamma})$ may or may not be appropriate, as we shall find later, but it is nonetheless widely used.

A **Bayesian** approach to outlier detection ("detection of spuriosity") arises from the work of Guttman [23]. It is interesting to note that the implicit concept of extremeness used to detect the outlier is again expressible in terms of the distance metric $R(\mathbf{x}; \boldsymbol{\mu}, \mathbf{V})$.

We now proceed to consider *accommodation procedures* and then *tests* of *discordancy* for multivariate outliers. With the greater complexity of the multivariate case, however, there also exist a wide range of *informal proposals* for outlier detection and processing – we shall consider some of these.

Accommodation

For multivariate data we often need statistical methods that are specifically robust against (i.e. which *accommodate*) outliers as manifestations of contamination. Such accommodation procedures exist for estimating parameters (often with specific regard to the multinormal distribution) and for various multivariate procedures (such as **principal components** and **discriminant analysis**).

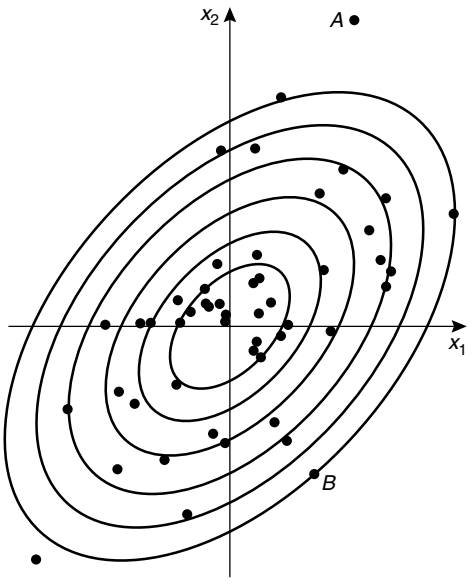

Figure 2 The sample with probability density ellipses

Suppose that, under a basic model F, $\mathbf{X}$ has mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. Outlier robust estimation of $\boldsymbol{\mu}$ and of $\mathbf{V}$ (also the correlation matrix, $\mathbf{R}$) has been widely examined, both in terms of individual components, and of the overall forms, of the mean vector and covariance matrix. For $\boldsymbol{\mu}$, the starting point is in an obvious generalization of the work of Anscombe [1]. For $\mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, we order the sample, and if $R_{(n)}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{V})$ (or $R_{(n)}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{S})$, depending on our knowledge about $\mathbf{V}$, is sufficiently large, then we omit the observation $\mathbf{x}$, yielding $R_{(n)}$, before estimating $\boldsymbol{\mu}$ from the residual sample; if $R_{(n)}$ is not sufficiently large, then we use the overall sample mean, $\bar{x}$. Such *adaptive trimming* is revised by Golub et al. [21], who employ a similar approach but based on *Winsorization* or "*semi-Winsorization*" (*see* **Trimming and Winsorization**). This approach can be extended to sequential trimming or Winsorization of several sufficiently extreme values. Guttman [23] considers the posterior distribution of $\mathbf{a}$ for a basic model $\mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$ and mean-slippage alternative $\mathrm{N}(\boldsymbol{\mu} + \mathbf{a}, \mathbf{V})$ for at most one of the observations.

Some qualitative effects of outliers on estimation of $\mathbf{V}$, and corresponding attitudes to **robust** estimation, are considered by Devlin et al. [15] and by Campbell [11], who claims that outliers "tend to deflate correlations and possibly inflate variance", although this may be rather too simplistic a prescription. A tangible form for outlier-robust M-estimators of $\boldsymbol{\mu}$ and $\mathbf{V}$, relevant to an elliptically symmetric basic model, is considered by Maronna [28] and by Campbell [11].

An early proposal for *direct* robust estimation of the matrix $\mathbf{V}$ in positive-definite form was made by Gnanadesikan & Kettenring [20]. It involves selective iterative *trimming* of the sample based on values of some measure $R(\mathbf{x}; \mathbf{x}^*, \mathbf{I})$, where $\mathbf{x}^*$ is a robust estimator of $\boldsymbol{\mu}$. The procedure is intuitively appealing, but only limited empirical investigation is reported.

Rousseeuw & van Zomeren [30] are also concerned with outlier robust estimation of $\boldsymbol{\mu}$ and $\mathbf{V}$. Rejecting the M-estimators of Campbell [11] in view of the low breakdown point, they suggest alternative robust estimators with a higher breakdown point.

Gnanadesikan [19, Section 5.2.3] includes detailed proposals for constructing robust estimators of the *individual elements* of $\boldsymbol{\mu}$ and $\mathbf{V}$ as well as for directly "multivariate" estimators of $\boldsymbol{\mu}$ and $\mathbf{V}$. One difficulty

with such an approach is that, if we form estimators of **V** or of **R** from separate robust estimators of their elements, the resulting **V** and **R** may not be positive-definite. Devlin et al. [14] propose a remedy involving "shrinking" **R** until it is positive-definite, and rescaling it if an estimate of **V** is required. (They also review various ad hoc estimators of the correlation coefficient, for a bivariate normal distribution, based on partitioning the sample space, on transformations of Kendall's τ (*see* **Association, Measures of**) or on **normal scores**; see also [26]).

Another approach to outlier-robust estimation of correlation uses the ideas of convex hull "peeling" or ellipsoidal peeling (see [8, 32] for details) following the suggestion by Barnett [2] that the most extreme group of observations in a multivariate sample are those lying on the convex hull (with those on the convex hull of the remaining sample, the second most extreme group, etc). Figure 3 shows the successive convex hulls for the bivariate sample presented in Figure 2.

Any form of multivariate analysis is, of course, likely to be susceptible to outliers as the manifestation of contamination. Proposals for modified forms of multivariate analysis which give protection against outliers include those for *principal component analysis* using M-estimators [11], for *canonical variate analysis* (*see* **Canonical Correlation**) [12], and for *discriminant analysis* [10]. Critchley & Vitiello [13] further examine Campbell's approach to the influence of outliers in linear discriminant analysis with particular regard to estimates of misclassification probabilities. Outlier robust (accommodation) methods have also been advanced for **analysis of covariance** [9], **correspondence analysis** [16], and **multidimensional scaling** [31].

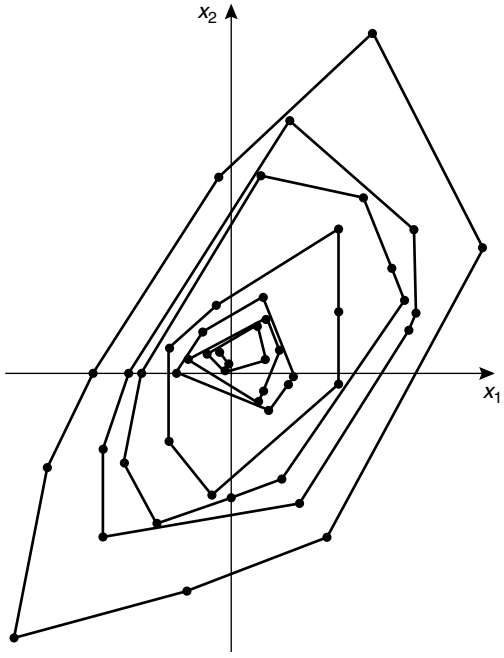

Figure 3 Convex hulls for the sample

Discordancy Tests

The notion of a test of discordancy is as relevant to multivariate data as it is to univariate samples, although conceptual and manipulative difficulties have limited the number of formal and specific proposals. Most work centres on the normal distribution, which proves amenable to the construction of tests of discordancy with desirable statistical properties and a useful degree of unity of form.

Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is a sample of n observations from a p-dimensional normal distribution, $N(\boldsymbol{\mu}, \mathbf{V})$. A possible alternative model which would account for a single contaminant is the slippage alternative, obtained as a multivariate adaptation of the univariate *models A* (slippage of the mean) and *B* (slippage of the variance) discussed by Ferguson [18]. A test of discordancy can be based on the *two-stage maximum likelihood ratio* principle (i.e. Principle B, above). Models A and B have been studied extensively with various assumptions about what parameter values are known [7, Section 7.3].

As an example, consider model A with **V** known. Here we are led to declare as the outlier $\mathbf{x}_{(n)}$ the observation $\mathbf{x}_i$ for which $R_i(\overline{\mathbf{x}}, \mathbf{V}) = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$ is a maximum, so that implicitly the observations have been ordered in terms of the reduced form of subordering based on the distance measure $R(\mathbf{x}; \overline{\mathbf{x}}, \mathbf{V})$. Furthermore, we will declare $\mathbf{x}_{(n)}$ a *discordant* outlier if

$$R_{(n)}(\overline{\mathbf{x}}, \mathbf{V}) = (\mathbf{x}_{(n)} - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_{(n)} - \overline{\mathbf{x}}) = \max R_j(\overline{\mathbf{x}}, \mathbf{V})$$

is significantly large.

The null distribution of $R_{(n)}(\overline{\mathbf{x}}, \mathbf{V})$ is neither readily determined in exact form nor very tractable, but

it has been widely studied. Critical values for discordancy tests for models A and B under various assumptions about parameter values being known or unknown are given in [7, Tables XXX–XXXIV].

For model A with $\mathbf{V}$ unknown, using a similar likelihood approach, it seems at first sight that quite a different principle is advanced for the declaration of an outlier $\mathbf{x}_i$ and for the assessment of its discordancy. Here we lead to implicitly ordering the multivariate observations in terms of reduced subordering based on the values of $|\mathbf{A}^{(j)}|$, which are $\sum(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$, where the sum is taken over all observations *except* $\mathbf{x}_j$. The $|\mathbf{A}^{(j)}|$ are ordered, and the observation corresponding to the smallest value of $|\mathbf{A}^{(j)}|$ is declared an outlier.

Thus the outlier is the observation whose removal from the sample effects the greatest reduction in the "internal scatter" of the data set, and it is adjudged discordant if this reduction is sufficiently large. This approach was first advanced by Wilks [33], but the distinction of principle for declaring an outlier in the case of unknown $\mathbf{V}$, compared with the case where $\mathbf{V}$ is known, turns out to be less profound than might appear at first sight since it is possible to reexpress the internal scatter in terms of the distance measure $R(\overline{\mathbf{x}}, \mathbf{S})$.

Thus the outlier is *again* that observation whose "distance" from the body of the data set is a maximum, provided we replace $\boldsymbol{\mu}$ and $\mathbf{V}$ by $\overline{\mathbf{x}}$ and $\mathbf{S}$.

Frequently we encounter multivariate data for which the normal distribution is quite unsuitable, in view of manifest **skewness**. Thus we might need to consider models expressing skewness. Two such prospects are provided by a *multivariate exponential* model and a *multivariate Pareto* model.

Many forms of multivariate exponential distribution have been proposed. One of these, due to Gumbel [22], has for the bivariate case a probability density function

$$f(x_1, x_2) = [(1 + \theta x_1)(1 + \theta x_2) - \theta] \times \exp(-x_1 x_2 - \theta x_1 x_2).$$

Applying a directional form of Principle A, an appropriate reduction measure,

$$R(\mathbf{X}) = X_1 + X_2 + \theta X_1 X_2,$$

is obtained. Thus an upper outlier is detected as the observation (x_{1i}, x_{2i}) which yields the largest value of $R(\mathbf{x})$ over the sample of n observations. It is judged discordant if the corresponding $R_{(n)} = \max R(\mathbf{x}_i)$ is sufficiently large. The distribution of $R_{(n)}$ is tractable. See Barnett [4], who also considers a discordancy test for an "upper" outlier in another skew bivariate distribution: namely, the (one of the two) Pareto distribution(s) considered by Mardia [27] which has probability density function

$$f(x_1, x_2) = a(a + 1)(\theta_1 \theta_2)^{a+1} \times (\theta_2 x_1 + \theta_1 x_2 - \theta_1 \theta_2)^{-(a+2)},$$

for $x_1 \geq \theta_1 \geq 0$, $x_2 \geq \theta_2 \geq 0$, and $a > 0$. The correlation coefficient is $\rho = a^{-1} (a > 2)$.

This time the appropriate restricted form of Principle A (assuming θ_1, θ_2, and a are known) yields a reduction measure

$$R(\mathbf{X}) = \left(\frac{x_1}{\theta_1}\right) + \left(\frac{x_2}{\theta_2}\right) - 1.$$

This is again tractable and yields the critical value γ_α for a level-α discordancy test for an upper outlier, satisfying

$$\delta \gamma_\alpha^{(a+1)} - (a + 1)\gamma_\alpha + a = 0,$$

with $\delta = 1 - (1 - \alpha)^{1/n}$.

Informal Methods for Multivariate Outliers

A host of informal proposals have been made for detecting outliers in multivariate data by quantitative or graphical methods (*see* **Graphical Displays**). These cannot be regarded as tests of discordancy; they may be based on derived reduction measures (but with no supporting distribution theory) or, more commonly, they are presented simply as aids to intuition in picking out multivariate observations which are suspiciously aberrant from the bulk of the sample.

Various forms of initial processing of the data, involving transformation, study of individual marginal components of the observations, judicious reduction of the multivariate observations to scalar quantities in the forms of reduction measures or linear combinations of components, changes in the coordinate bases of the observations, and appropriate methods of graphical representation, can all help to identify or highlight suspicious observations. If

several such procedures are applied simultaneously (or separately) to a set of data they can help to overcome the difficulty caused by the absence of a natural overall ordering of the sample members. An observation which clearly stands out on one, or preferably more, processed re-representations of the sample becomes a firm candidate for identification as an outlier.

An early example of an informal graphical procedure is described by Healy [25], who proposes plotting the *ordered* $R_j(\bar{\mathbf{x}}, \mathbf{S})$ against the expected values of the **order statistics** of a sample of size n from χ_p^2.

We should not underestimate the importance of the *marginal samples* (that is, the univariate samples of each component value in the multivariate data) in the occurrence of outliers. It is perfectly plausible for contamination to occur in one of the marginal variables alone – for example, by misreading or an error of recording. It could even happen that a single marginal variable is intrinsically more liable to contamination. We must be careful, however, not to adopt too simplistic an approach in examining this prospect. This is illustrated in some work by Barnett [5], who considers a sample from a bivariate normal distribution where contamination may have occurred by slippage of the mean of the *first component only* for one observation. The detection of such an outlier is by no means simple.

Graphical and pictorial methods are often advanced in relation to multivariate outliers. Rohlf [29] remarks as follows:

> Despite the apparent complexity of the problem, one can still characterize outliers by the fact that they are somewhat isolated from the main cloud of points. They may not "stick out on the end" of the distributions as univariate outliers must, but they must "stick out" somewhere.

With this emphasis it is natural to consider different ways in which we can merely *look at the data* to see if they seem to contain outliers. A variety of methods employing different forms of pictorial or graphical representation have been proposed with varying degrees of sophistication. Review of such methods of "informal inference" applied to general problems of analysis of multivariate data including the detection of outliers are presented by Gnanadesikan [19] and by Barnett [6]. (See [7, Section 7.4], for further details on published contributions in this spirit.)

It can be useful to perform a preliminary principal components analysis on the data, and to look at sample values of the projection of the observations on to the principal components of different order. Gnanadesikan & Kettenring [20] discuss this in some detail, remarking how the first few principal components are sensitive to outliers inflating variances or covariances (or correlations, if the principal components analysis has been conducted in terms of the sample correlation matrix, rather than the sample covariance matrix), whilst the last few are sensitive to outliers adding spurious dimensions to the data or obscuring singularities. Some modifications of approach to outlier detection by principal components analysis are suggested by Hawkins [24] and by Fellegi [17].

Another way in which informal quantitative and graphical procedures may be used to exhibit outliers is to construct reduced univariate measures. Gnanadesikan & Kettenring [20] consider various possible classes of such measures which are all similar in principle to the "distance" measure discussed above. Particularly extreme values of such statistics, possibly demonstrated by graphical display, may reveal outliers of different types. For graphical display of outliers, the "gamma-type probability plots" of ordered values, with appropriately estimated shape parameters, are also a useful approximate procedure and have been widely considered.

We have already noted the way in which outliers may affect, and be revealed by, the correlation structure in the data. Some proposals for identifying multivariate outliers specifically consider this matter. Gnanadesikan & Kettenring [20] suggest that we examine the product–moment correlation coefficients $r_{-j}(s, t)$ relating to the sth and tth marginal samples after the omission of the single observation $\mathbf{x}_j$. As we vary j we can examine, for any choice of s and t, the way in which the correlation changes – substantial variations reflecting possible outliers. Devlin et al. [14] make use of the *influence function* to investigate how outliers affect correlation estimates in bivariate data ($p = 2$). Influence functions of other statistics (apart from the correlation coefficient) have also been proposed as a basis for detecting outliers.

We noted earlier the characterization of multivariate outliers suggested by Rohlf [29] – that they are separated from other observations "by distinct gaps". Rohlf has used this idea to develop a *gap test* for multivariate outliers based on minimum spanning trees (*see* **Graphical Displays**). He argues that a single

isolated point will be connected to only one other point in the minimum spanning tree by a relatively large distance, and that *at least one* edge connection from a cluster of outliers must also be relatively large. Accordingly, an informal gap test for outliers is proposed based on this principle.

This article can only briefly review the very wide range of concepts and methods for multivariate outliers. Barnett & Lewis [7, Chapter 7] provide the entrée to more detailed study.

References

[1] Anscombe, F.J. (1960). Rejection of outliers, *Technometrics* **2**, 12–147.

[2] Barnett, V. (1976). The ordering of multivariate data (with discussion), *Journal of the Royal Statistical Society, Series A* **139**, 318–354.

[3] Barnett V. (1978). The study of outliers: purpose and model, *Applied Statistics* **27**, 242–250.

[4] Barnett, V. (1979). Some outlier tests for multivariate samples, *South African Statistical Journal* **13**, 29–52.

[5] Barnett, V., ed. (1981). *Interpreting Multivariate Data*. Wiley, Chichester.

[6] Barnett, V. (1983). Marginal outliers in the bivariate normal distribution, *Bulletin of the International Statistical Institute* **50**, 579–583.

[7] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Ed. Wiley, Chichester.

[8] Bebbington, A.C. (1978). A method of bivariate trimming for robust estimation of the correlation coefficient, *Applied Statistics* **27**, 221–226.

[9] Birch, J.B. & Myers, R.H. (1982). Robust analysis of covariance, *Biometrics* **38**, 699–713.

[10] Campbell, N.A. (1978). The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics* **27**, 251–258.

[11] Campbell, N.A. (1980). Robust procedures in multivariate analysis, I: robust covariance estimation, *Applied Statistics* **29**, 231–237.

[12] Campbell, N.A. (1982). Robust procedures in multivariate analysis, II: robust canonical variate analysis, *Applied Statistics* **31**, 1–8.

[13] Critchley, F. & Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis, *Biometrika* **78**, 677–690.

[14] Devlin, S.J., Gnanadesikan, R. & Kettenring, J.R. (1975). Robust estimation and outlier detection with correlation coefficients, *Biometrika* **62**, 531–545.

[15] Devlin, S.J., Gnanadesikan, R. & Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components, *Journal of the American Statistical Association* **76**, 354–362.

[16] Escoffier, B. & Le Roux, B. (1976). Factor's stability in correspondence analysis. How to control the influence of outlying data (in French), *Cahiers de l'Analyse des Données* **1**, 297–318.

[17] Fellegi, I.P. (1975). Automatic editing and imputation of quantitative data, *Bulletin of the International Statistical Institute* **46**, 249–253.

[18] Ferguson, T.S. (1961). On the rejection of outliers, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, J. Neyman, ed. University of California Press, Berkeley and Los Angeles, pp. 253–287.

[19] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

[20] Gnanadesikan, R. & Kettenring, J.R. (1972). Robust estimates, residual and outlier detection with multiresponse data, *Biometrics* **28**, 81–124.

[21] Golub, G.H., Guttman, I. & Dutter, R. (1973). Examination of pseudo-residuals of outliers for detection spuriosity in the general univariate linear model, in *Multivariate Statistical Inference*, D.G. Kabe & P.R. Gupta, eds. North-Holland, Amsterdam.

[22] Gumbel, E.J. (1960). Bivariate exponential distributions, *Journal of the American Statistical Association* **55**, 698–707.

[23] Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity – a Bayesian approach, *Technometrics* **15**, 723–738.

[24] Hawkins, D.M. (1974). The detection of errors in multivariate data using principal components, *Journal of the American Statistical Association* **69**, 340–344.

[25] Healy, M.J.R. (1968). Multivariate normal plotting, *Applied Statistics* **17**, 157–161.

[26] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

[27] Mardia, K.V. (1962). Multivariate Pareto distributions, *Annals of Statistics* **33**, 1008–1015.

[28] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter, *Annals of Statistics* **4**, 51–67.

[29] Rohlf, F.J. (1975). Generalisation of the gap test for the detection of multivariate outliers, *Biometrics* **31**, 93–101.

[30] Rousseeuw, P.J. & van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association* **85**, 633–639.

[31] Spence, I. & Lewandowski, S. (1989). Robust multidimensional scaling, *Psychometrika* **54**, 501–513.

[32] Titterington, D.M. (1978). Estimation of correlation coefficients by ellipsoidal trimming, *Applied Statistics* **27**, 227–234.

[33] Wilks, S.S. (1963). Multivariate statistical outliers, *Sankhyā, Series A* **25**, 407–426.

(*See also* **Multivariate Distributions, Overview**)

VIC BARNETT

Multivariate Survival Analysis

Multivariate survival analysis deals with methods designed for the study of correlated failure time observations taken on a single individual or a group of individuals. Examples of applications include epidemiologic studies on the familial tendency in chronic disease incidence, follow-up studies of recurrent diseases (*see* **Repeated Events**), litter-matched carcinogenicity experiments on animals (*see* **Tumor Incidence Experiments**), or **clinical trials** on paired human organs.

Extensions of both non and semiparametric methods of univariate survival analysis to the multivariate setting have turned out to be quite difficult and resulted in many different approaches. In general, the analysis of multivariate failure time data requires a rigorous specification of both censoring mechanisms and timescales on which failure time observations are recorded. Multivariate censoring arises when failure time observations are registered on several different time clocks. This is common to familial studies where life lengths or age at the onset of a disease are measured on a different timescale for each member of the family. Moreover, each family member may be subject to his/her own censoring mechanism. So-called univariate censoring may arise when serial observations are taken on a single individual. These may be, for example, times of successive episodes of asthmatic attacks or times of the onset of specific stages of a nonrecurrent disease. The univariate censoring model applies also to many twin studies and matched-pair experiments. Finally, the occurrence of a single event may be recorded on several timescales. Examples are provided by **staggered entry** models where of interest are both the calendar time of entry into a clinical trial and the duration of the subsequent time period on trial.

From a more mathematical point of view, both univariate censoring models and models involving multiple time measurements of a single event can be formulated usually within the classical marked **point process** framework of survival analysis [4]. However, the martingale-based methods for the derivation of estimates and their large-sample properties are, in general, insufficient in this setting and are replaced by methods related to empirical processes techniques. These methods also apply to models involving multiple clock experiments. Such experiments can only seldom be defined using marked point processes in real time because concepts such as "past–present–future" do not have a clear-cut interpretation in the case of multiparameter processes.

Nonparametric Estimation

Survival function estimation is the center-point of nonparametric approaches towards the analysis of multivariate failure time data. From a practical point of view, multivariate survival functions are seldom of interest on their own; however, they play an important role as an auxiliary tool in other inferential problems. Examples of applications include **density estimation** in **graphical displays** of data, regression analyses with uncensored or censored covariates, estimation of dependence parameters, **goodness-of-fit** tests, and so forth.

In the special univariate case, estimation of the unknown survival function S of a possibly censored failure time T is usually based on the Kaplan–Meier estimate. The choice of this estimator can be justified by both **nonparametric maximum likelihood** and self-consistency principles [20, 30]. In addition, the estimate is the sample analog of the **product-integral** representation of survival functions in terms of cumulative hazards [23]. The latter concept does not have a unique extension to multivariate distributions, whereas both maximum likelihood and self-consistency principles do not lead, in general, to unique and consistent estimates of the multivariate survival functions in the presence of censoring [47, 53]. As a result, there are many different approaches to this estimation problem.

Here, we assume first that the multivariate vector of interest is of the form $(T, \mathbf{Z})$ where T is a univariate failure time subject to censoring and $\mathbf{Z}$ is a vector of uncensored covariates. In this case, nonparametric estimation methods usually aim to recover the parameters of the conditional survival function,

$$S(t|z) = P(T > t|\mathbf{Z} = z) = \pi_{[0,t]}(1 - A(\mathrm{d}u|z)), \tag{1}$$

where $A(t|z)$ is the cumulative hazard function of the conditional distribution of the failure time T given the covariate $\mathbf{Z} = z$. Under identifiability

assumptions such as the conditional independence of the failure time T and censoring time $\tilde{T}$ given the covariate $\mathbf{Z}$, estimation methods are often based on Beran's [5] conditional **Nelson–Aalen** and **Kaplan–Meier** estimates. The conditional Nelson–Aalen estimate assumes the form

$$\hat{A}(t|z) = \int_{[0,t]} \frac{\sum_{i=1}^{n} W_i(z) N_i(\mathrm{d}u)}{\sum_{i=1}^{n} W_i(z) Y_i(u)},$$

where $(N_i, Y_i), i = 1, \ldots, n$, are counting and risk processes $N_i(t) = I(T_i \leq t \wedge \tilde{T}_i)$ and $Y_i(t) = I(T_i \wedge \tilde{T}_i \geq t)$ associated with a sample of n individuals under study, and $W_i(z), i = 1, \ldots, n$, are weights dependent only on their covariates $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. The conditional Kaplan–Meier estimate is obtained by substituting this estimate into the right-hand side of (1). A general counting process formulation of **nonparametric regression** allows one also to accommodate intensity models with time-dependent covariates or recurrent failure time events. In both settings, different choices of the weights $W_i(z)$ lead to different types of nonparametric regression estimates. Examples include regressogram, kernel, and nearest neighbor estimates. Under regularity conditions, these estimates are uniformly consistent and asymptotically Gaussian at a rate dependent on the choice of the smoother [13, 35].

Under suitable identifiability conditions, the conditional Kaplan–Meier estimate remains consistent when covariates are subject to censoring and provides a convenient tool in the estimation of the joint distribution function of failure time vectors $\mathbf{T} = (T_1, \ldots, T_m)$ subject to multivariate censoring. In what follows we consider the simplest bivariate censoring model and assume that the observable data are of the form $(X_1, X_2, \delta_1, \delta_2)$ where $X_l = T_l \wedge \tilde{T}_l$, $\delta_l = I(T_l \leq \tilde{T}_l)$ and $T = (T_1, T_2)$, $\tilde{T} = (\tilde{T}_1, \tilde{T}_2)$ are independent failure and censoring variables.

As an ad hoc modification of Efron's self-consistency algorithm, Pruitt [47] proposed to consider the sample analog of the identity

$$F(t_1, t_2) = \sum_{i,j=0,1} F_{ij}(t_1, t_2),$$

where $F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2)$ and $F_{ij}(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2, \delta_1 = i, \delta_2 = j)$. Independence of the failure and censoring times implies

$$\begin{aligned} F(t_1, t_2) = &\int_{[0,t_1]\times[0,t_2]} P(T_1 \leq t_1, \\ &T_2 \leq t_2 | T_1 > u, T_2 > v) \\ &\times \mathrm{E}N_{00}(\mathrm{d}u, \mathrm{d}v) + \mathrm{E}N_{11}(t_1, t_2) \\ &+ \int_{[0,t_1]\times[0,t_2]} \frac{P(v < T_2 \leq t_2 | T_1 = u)}{P(T_2 > v | T_1 = u)} \\ &\times \mathrm{E}N_{10}(\mathrm{d}u, \mathrm{d}v) \\ &+ \int_{[0,t_1]\times[0,t_2]} \frac{P(u < T_1 \leq t_1 | T_2 = v)}{P(T_1 > u | T_2 = v)} \\ &\times \mathrm{E}N_{01}(\mathrm{d}u, \mathrm{d}v), \end{aligned} \tag{2}$$

where $N_{ij}(t_1, t_2) = I(X_1 \leq t_1, X_2 \leq t_2, \delta_1 = i, \delta_2 = j)$. Pruitt's estimate of the distribution function F is the solution to the sample counterpart of this equation obtained by replacing the expected processes $\mathrm{E}N_{ij}(t_1, t_2)$ by their empirical counterparts and by approximating the subdistribution functions $P(v < T_j \leq t_j | T_{3-j} = u)$ using the conditional Kaplan–Meier estimate. In special cases, such as the censoring of only one component of the vector $\mathbf{T}$ or univariate censoring of ordered failure times $T_1 < T_2$, the solution of (2) has an explicit form and the corresponding estimates assume the form of averaged conditional Kaplan–Meier estimates. In the first of these two cases the estimate is also fully efficient [1]. The consistency and asymptotic normality of Pruitt's estimate in the presence of bivariate censoring is shown by van der Laan [55] who also provides a fully efficient survival function estimate designed for data subject to bivariate censoring [56].

There are several alternative estimators based on ad hoc representations of the survival function of the failure times (T_1, T_2) in terms of the subdistribution functions of the observable data. Such estimates are useful in that they are very simple to implement in practice. In particular, they do not rely on smoothing techniques and therefore can be applied towards analysis of data sets of small or moderate sample sizes. One possible choice corresponds to the sample analog of the identity $S(t_1, t_2) = S_1(t_1) S_2(t_2) M(t_1, t_2)$, where S_l are the marginal survival functions and

$$M(t_1, t_2) = \exp \int_{[0,t_1]\times[0,t_2]} \mathrm{d} \log F. \tag{3}$$

The estimate of the bivariate survival function is obtained by replacing the unknown marginals by

their respective Kaplan–Meier estimators, whereas the function M is approximated by

$$\widehat{M}(t_1, t_2) = \prod_{u_j \le t_j} \prod_{\alpha} \left\{ \sum_{i=1}^{n} \prod_{l \in \alpha} [Y_{il}(u_l) - N_{il}(\Delta u_l)] \prod_{l \notin \alpha} Y_{il}(u_l) \right\}^{(-1)^{|\alpha|}},$$

where $|\alpha|$ is the cardinality of a set $\alpha \subseteq \{1, 2\}$ and $(N_{il}, Y_{il}), l = 1, 2, i = 1, \ldots, n$, are the marginal counting and risk processes associated with an independent identically distributed (iid) sample of censored failure times [14]. Alternatively, (3) is the unique solution of the Volterra integral equation:

$$M(t_1, t_2) = 1 + \int_{[0,t_1]\times[0,t_2]} M(u_1-, u_2-) B(\mathrm{d}u_1, \mathrm{d}u_2), \quad (4)$$

where B is a standardized version of the covariance of the marginal martingales, $M_{il} = N_{il} - \int_0 Y_{il}\, \mathrm{d}A_l$, $A_l(\mathrm{d}t) = -S_l(\mathrm{d}t)/S_l(t-), l = 1, 2$. An estimate of this covariance function is given by Prentice & Cai [46] who further propose to recover the function M by the solution to the sample analog of (4). In practice, the two estimates are in close numerical agreement and are both consistent and asymptotically Gaussian in the presence of bivariate censoring. References [8], [15], [18], [22], [24], [58], and [60] discuss the asymptotic properties of these estimates, extensions to multivariate data and also provide applications to regression analyses and testing dependencies. Other examples of survival function estimates designed for multivariate censored data are given in [9], [10], [54], [7], and [53]. Pruitt [49] provides a useful overview of the small sample properties of some of these estimates along with software available through statlib at the Carnegie Mellon University. Some care has to be taken in the use of these estimates, however, as they may define signed measures and may fail to be consistent in the presence of more complex censoring mechanisms [19, 45, 48, 53].

Nonparametric Testing

Testing independence, k sample homogeneity, and symmetry are examples of common testing problems involving multivariate data. In the case of bivariate data, nonparametric tests for independence are often based on **rank correlation** statistics such as logrank and Spearman rank correlation or Kendall's τ. Examples of censored data analogs of these tests are given in [6], [11], [12], and [39]. In particular, given an iid sample $(X_{i1}, X_{i2}, \delta_{i1}, \delta_{i2}), i = 1, \ldots, n$, of the observed withdrawal times and censoring indicators, Oakes' [39] analog of Kendall's τ test rests on a comparison of the number of pairs of observations known to correspond to concordant and discordant failure times:

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i<j} a_1(i, j) a_2(i, j),$$

where the score $a_l(i, j), l = 1, 2$, assumes values $\delta_{il}, 0$ and $-\delta_{jl}$ if sign $(X_{il} - X_{jl}) = 1, 0$ and -1, respectively. The logrank (or Savage scores) correlation test is given by

$$R_n = \frac{1}{n} \sum_{i=1}^{n} [\hat{A}_1(X_{i1}) - \delta_{i1}][\hat{A}_2(X_{i2}) - \delta_{i2}],$$

where $\hat{A}_l, l = 1, 2$, are Nelson–Aalen estimates of the marginal cumulative hazard functions. Standardized versions of these statistics are asymptotically mean zero normal and provide consistent tests against alternatives of signed dependence. In the case of uncensored data, under alternatives remote to independence, rank correlation statistics and Kendall's τ statistic also provide consistent estimates of common nonparametric association measures. However, this does not carry over to censored data. In the case of both uncensored and censored data, multivariate extensions of Kendall's test and rank correlation tests are not uniquely defined. Some examples are given in [2] and [18] (*see* **Multivariate Median and Rank Sum Tests**).

Omnibus Kolmogorov–Smirnov type tests for independence are due to Pons & de Turckheim [44]. In the bivariate case, such tests can be based on statistics

$$U_n = \sup_{t_1 \le \tau_1, t_2 \le \tau_2} \sqrt{n} |\hat{S}(t_1, t_2) - \hat{S}_1(t_1) \hat{S}_2(t_2)|,$$

where $\hat{S}_i, i = 1, 2$, are the marginal Kaplan–Meier estimates, and $\hat{S}$ is an estimate of the joint survival function of the underlying failure times. Alternatively, such tests can also be based on the supremum norm statistic comparing estimates of bivariate and

marginal cumulative hazards. Critical values of the tests are obtained based the bootstrap approximation to the null distribution of U_n. Bootstrap methods apply also to other testing problems involving multivariate data such as k sample homogeneity or symmetry tests [50].

Rank tests designed for these problems are discussed in [3], [16], [17], [41], [58], and [61], among others. In particular, symmetry tests arise in matched-pair experiments and are often based on paired or signed rank tests. In the case of uncensored data, paired rank tests (or conditional rank tests) are scores tests derived from the conditional likelihood of ranks given the observed paired ranks [52]. Evaluation of the scores of such tests and derivation of both finite sample and asymptotic properties are, in general, quite difficult; however, special choices such as the paired Wilcoxon and logrank tests gained some popularity due to their good performance as compared with other symmetry tests. Wilcoxon signed-rank tests derive their form from the marginal likelihood of signed ranks in the log-linear model $\log T_{2i} = \log T_{1i} + \theta + \varepsilon_i$, where θ is an unknown shift parameter and ε_i is an error term with a known symmetric distribution. Special cases include the sign, and signed Wilcoxon and **normal scores** tests. Their censored data analogs are limited to models involving univariate censoring. The paper [17] provides a comparison of the **asymptotic relative efficiency** of the two classes of symmetry tests.

Regression Models

Regression analysis of multivariate data is often based on **random effects** (or **frailty**) models. The idea of the use of frailty models is due to Vaupel et al. [57] who introduced this concept to model heterogeneity in univariate survival models. In the multivariate setting, frailty models are used to induce dependence among failure times by way of a random effect accounting for possible genetic, environmental and other factors linking the marginal failure processes.

The simplest random effects models for multivariate data are derived under the assumption that m individuals, such as family members, share a common unobserved factor W. Here, W is a nonnegative random variable such that, conditionally on W and the vector of observable covariates $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$, the failure times $(\mathbf{T}_1, \ldots, \mathbf{T}_m)$ of the m individuals are independent random variables with hazard functions

$$\alpha_i(t|\mathbf{Z}, W) = W\alpha_{0i}(t)\exp(\beta_i^{\mathrm{T}}\mathbf{Z}_i) \quad i = 1, \ldots, m. \tag{5}$$

The conditional survival function of the failure times $\mathbf{T} = (T_1, \ldots, T_m)$ given the covariates $\mathbf{Z}$ is

$$S(t_1, \ldots, t_m|\mathbf{Z}) = \int \exp\left[-w\sum_{i=1}^{m}\exp(\beta_i^{\mathrm{T}}\mathbf{Z}_i)\int_0^{t_i}\alpha_{0i}(u)\,\mathrm{d}u\right] \times F_W(\mathrm{d}w),$$

where F_W is the distribution of the frailty variable W. Typically, this distribution belongs to some known parametric family such as the family of gamma, positive stable, **inverse Gaussian distribution** and **lognormal distribution** [40, 26, 34]. Apart from the conditional independence assumption, the model (5) stipulates that, conditionally on W, components of the vector $\mathbf{T}$ follow the proportional hazard model. However, with the exception of Hougaard's positive stable frailty model, averaging over W leads to models with marginals that do not satisfy the proportional hazard model assumption. Different choices of the frailty distribution give rise also to different types of dependence among the marginal failure times. More flexible frailty models allowing for several types of random effects and different degrees of association among the components of the vector $\mathbf{T} = (T_1, \ldots, T_m)$, can be obtained by assuming that W is a vector or matrix of correlated variables. Examples of such models were provided by Yashin et al. [62] and McGilchrist [34].

Inference methods in these models are quite difficult. A frequent approach rests on **nonparametric maximum likelihood** and **EM algorithm** [38]; however, only partial results covering the gamma and correlated frailty models are available at the present time [36, 37, 42, 43] and much attention is paid to alternate approaches to regression analysis of multivariate data. In particular, in the case of litter-matched experiments. Holt & Prentice [28] propose the model (5) with systematic rather than random effects. The paper [25] provides a detailed derivation of the properties of the Cox and maximum likelihood estimates in this model. Another common choice corresponds to the so-called "marginal

approach" (*see* **Marginal Models for Multivariate Survival Data**). Instead of modeling the conditional distribution of T_j, $j = 1, \ldots, m$, given a frailty variable, it imposes semiparametric assumptions on the marginal distribution of T_js and leaves the joint dependence structure among the components of the vector $(T_1, \ldots, T_m)$ unspecified. Under suitable identifiability assumptions, the unknown parameters of marginals can be estimated by modifying estimation procedures developed for univariate regression analyses. Wei et al. [59] and Lin & Wei [32] provide the joint asymptotic structure of the estimates in multivariate models with marginals satisfying the Cox proportional hazard and accelerated failure time model assumptions. The marginal approach can further be strengthened by assuming that the failure time data follow a multivariate **copula** model. In the absence of covariates, these are multivariate models in which the joint survival function S of the vector $\mathbf{T} = (T_1, \ldots, T_m)$ is of the form $S = C_\theta(S_1, \ldots, S_m)$, where S_l are the marginal survival functions and $\{C_\theta : \theta \in \Theta\}$ is a parametric family of distribution functions on the unit cube with uniform marginals [21, 27]. The parameter θ accounts for the joint dependence among the marginal failure times. In these models, the dependence parameter and the parameters of the marginals can be based on a two-stage estimation process: the parameters of the marginals are first estimated as in univariate analyses, the resulting estimates are next used to construct a **pseudo-likelihood** for the unknown dependence parameter θ [27, 31, 51]. Although the resulting estimates are inefficient, this estimation approach is relatively easy to implement and has a good practical performance. Other examples of semiparametric models that can be used in the analysis of multivariate failure time data are surveyed in [26] and [33].

Acknowledgment

Research supported by the NSF Grant DMS 9504507 and NIH Grant R01 CA 65595-02.

References

[1] Akritas, M.G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring, *Annals of Statistics* **22**, 1299–1327.

[2] Akritas, M.G. & Siebert, M. (1996). A test for partial correlation with censored astronomical data, *Monthly Notices of the Royal Astronomical Society* **278**, 919–924.

[3] Albers, W. (1988). Combined rank tests for randomly censored paired data, *Journal of the American Statistical Association* **83**, 1159–1162.

[4] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[5] Beran, R. (1981). Nonparametric regression with randomly censored survival data, *Technical Report*, University of California, Berkeley.

[6] Brown, W.B., Hollander, M. & Korwar, R.M. (1974). Nonparametric tests of independence for censored data with applications to heart transplant studies, in *Reliability and Biometry: Statistical Analysis of Lifelengths*, F. Proschan & R.G. Serfling, eds. SIAM, Philadelphia.

[7] Burke, M.D. (1988). Estimation of a bivariate survival function under random censorship, *Biometrika* **75**, 379–382.

[8] Cai, J. & Prentice, R.L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data, *Biometrika* **82**, 151–164.

[9] Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data, *Biometrika* **68**, 417–422.

[10] Campbell, G. & Földes, A. (1982). Large sample properties of nonparametric bivariate estimators with censored data, in *Nonparametric Statistical Inference*, B.V. Gnedenko, M.L. Puri & I. Vincze, eds. North-Holland, Amsterdam.

[11] Cuzick, J. (1982). Rank tests for association with right censored data, *Biometrika* **89**, 351–364.

[12] Dabrowska, D.M. (1986). Rank tests for independence for bivariate censored data, *Annals of Statistics* **14**, 250–264.

[13] Dabrowska, D.M. (1987). Nonparametric regression with censored survival time data, *Scandinavian Journal of Statistics* **14**, 181–197.

[14] Dabrowska, D.M. (1988). Kaplan–Meier estimate on the plane, *Annals of Statistics* **16**, 1475–1489.

[15] Dabrowska, D.M. (1989). Kaplan–Meier estimate on the plane: Weak convergence, LIL and the bootstrap, *Journal of Multivariate Analysis* **29**, 308–325.

[16] Dabrowska, D.M. (1989). Rank tests for matched pair experiments with censored data, *Journal of Multivariate Analysis* **28**, 88–114.

[17] Dabrowska, D.M. (1990). Signed rank tests for censored matched pairs, *Journal of the American Statistical Association* **85**, 478–485.

[18] Dabrowska, D.M. (1998). Weak convergence of a product integral dependence measure, *Scandinavian Journal of Statistics*, to appear.

[19] Dabrowska, D.M. & Lee, W. (1996). Nonparametric estimation of transition probabilities in a two-stage duration model, *Journal of Nonparametric Statistics* **7**, 75–103.

[20] Efron, B. (1967). The two-sample problem with censored data, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, 831–853.

[21] Genest, C. & MacKay, N. (1986). Copules Archimèdiennes et familles de lois bidimensionnelles dont les marges sont données, *Canadian Journal of Statistics* **14**, 145–159.

[22] Gill, R.D. (1992). Multivariate survival analysis. Proceedings of the Second World Congress of the Bernoulli Society, Uppsala, Sweden, *Theory of Probability and Its Applications* **37**, 19–36, 307–328.

[23] Gill, R.D. & Johansen, S. (1990). A survey of product integration with a view towards application in survival analysis, *Annals of Statistics* **18**, 1501–1555.

[24] Gill, R.D., van der Laan, M. & Wellner, J.A. (1995). Inefficient estimators for three multivariate models, *Annales de l'Institut Henri Poincaré* **31**, 545–597.

[25] Gross, S.T. & Huber, C. (1987). Matched pair experiments: Cox and maximum likelihood estimation, *Scandinavian Journal of Statistics* **14**, 27–42.

[26] Hougaard, P. (1986). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678.

[27] Hougaard, P. (1987). Modeling multivariate survival, *Scandinavian Journal of Statistics* **14**, 291–304.

[28] Holt, J.D. & Prentice, R.L. (1974). Survival analysis in twin studies and matched pair experiments, *Biometrika* **61**, 17–30.

[29] Joe, H. (1993). Parametric families of multivariate distributions with given margins, *Journal of Multivariate Analysis* **46**, 262–282.

[30] Johansen, S. (1978). The product limit estimator as maximum likelihood estimator, *Scandinavian Journal of Statistics* **5**, 195–199.

[31] Liang, K.Y., Self, S.G., Banden-Roche, K.J. & Scott, L.Z. (1995). Some recent developments for regression analysis of multivariate failure time data, *Lifetime Data Analysis* **1**, 403–416.

[32] Lin, J.S. & Wei, L.J. (1992). Linear regression analysis for multivariate failure time observations, *Journal of the American Statistical Association* **87**, 1091–1097.

[33] Marshall, A.W. & Olkin, I. (1988). Families of multivariate distributions, *Journal of the American Statistical Association* **83**, 824–841.

[34] McGilchrist, C. (1993). REML estimation for survival models with frailty, *Biometrics* **49**, 221–225.

[35] McKeague, I.W. & Utikal, K.J. (1990). Inference for a nonlinear counting process regression model, *Annals of Statistics* **18**, 1172–1187.

[36] Murphy, S.A. (1994). Consistency in a proportional hazard model incorporating random effects, *Annals of Statistics* **22**, 712–731.

[37] Murphy, S.A. (1995). Asymptotic theory for the frailty model, *Annals of Statistics* **23**, 182–198.

[38] Nielsen, G.G., Gill, R.D., Andersen, P.K. & Sorensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics* **19**, 25–43.

[39] Oakes, D. (1982). A concordance test for independence in the presence of censoring, *Biometrics* **38**, 451–455.

[40] Oakes, D. (1989). Bivariate survival models introduced by frailties, *Journal of the American Statistical Association* **84**, 487–493.

[41] O'Brien, P.C. & Fleming, T.R. (1987). A paired Prentice-Wilcoxon test for censored paired data, *Biometrics* **43**, 169–180.

[42] Parner, E. (1996). Consistency in the correlated frailty model, *Technical Report*, University of Aarhus.

[43] Parner, E. (1996). Asymptotic normality in the correlated frailty model, *Technical Report*, University of Aarhus.

[44] Pons, O. & de Turckheim, E. (1991). Tests for independence for bivariate censored data based on the empirical joint hazard function, *Scandinavian Journal of Statistics* **18**, 21–39.

[45] Pons, O., Kaddour, A. & de Turckheim, E. (1992). A nonparametric approach to dependence for bivariate censored data, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel, eds. Kluwer, Boston, pp. 381–392.

[46] Prentice, R.L. & Cai, J. (1992). Covariance survivor function estimation using censored multivariate failure time data, *Biometrika* **79**, 495–512.

[47] Pruitt, R.C. (1991). Bivariate survival curve estimation using nonparametric smoothing techniques, *Technical Report*, University of Minnesota.

[48] Pruitt, R.C. (1993). Identifiability of bivariate survival curves from censored data, *Journal of the American Statistical Association* **88**, 573–579.

[49] Pruitt, R.C. (1993). Small sample comparison of six bivariate survival curve estimators, *Journal of Statistical Computation and Simulation* **45**, 147–167.

[50] Romano, J.P. (1988). A bootstrap revival of some nonparametric distance tests, *Journal of the American Statistical Association* **83**, 698–709.

[51] Shih, J.H. & Louis, T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data, *Biometrics* **51**, 1384–1399.

[52] Snijders, T. (1981). Rank tests for bivariate symmetry, *Annals of Statistics* **9**, 1087–1095.

[53] Stute, W. (1993). Consistent estimation under random censorship when covariables are present, *Journal of Multivariate Analysis* **45**, 89–103.

[54] Tsai, W.-Y., Leurgans, S. & Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring, *Annals of Statistics* **14**, 1351–1365.

[55] van der Laan, M.J. (1994). Modified EM-equations-estimator of the bivariate survival function, *Mathematical Methods in Statistics* **3**, 213–243.

[56] van der Laan, M.J. (1996). Efficient estimator of the bivariate survival function and repairing NPMLE, *Annals of Statistics* **24**, 596–627.

[57] Vaupel, J.W., Manton, K.G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **16**, 439–454.

[58] Wei, L.J. & Lachin, J.M. (1985). Two sample asymptotically distribution free tests for incomplete multivariate

observations, *Journal of the American Statistical Association* **79**, 653–661.

[59] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

[60] Wellner, J. (1994). Covariance formulas via marginal martingales, *Statistica Neerlandica* **48**,.

[61] Woolson, R.F. & Lachenbruch, P.A. (1980). Rank tests for censored matched pairs, *Biometrika* **67**, 597–606.

[62] Yashin, A.I., Vaupel, J.W. & Iachine, I.A. (1995). Correlated individual frailty: An advantageous approach to survival analysis of bivariate data, *Mathematical Population Studies* **5**, 1–10.

DOROTA M. DABROWSKA

Multivariate *t* Distribution

Student's *t* distributions on $\mathbb{R}^1$ arise through the use of Studentized statistics in normal-theory sampling models. Multivariate versions on $\mathbb{R}^k$ typically derive through Studentization from **multivariate normal** models through an impressive array of methods employed in the analysis of biomedical and other data. Properties of these procedures in turn rest on those of the corresponding joint distributions, and their implementation requires the availability of special aid tables or their software equivalents. On occasion, multivariate t distributions themselves serve to model the errors of a random experiment, offering greater flexibility and heavier tails than multivariate normal models.

There are two basic types of t distributions on $\mathbb{R}^k$. Type I distributions emerge on scaling each component of a random normal vector by a single random scalar. Type II distributions entail separate scalings by elements of a further random vector, itself having a joint distribution on $\mathbb{R}^k$. Details are supplied subsequently. Excellent references are [5, Chapter 27] and [8, Chapter 9].

There is by now a considerable literature pertaining to multivariate t distributions on $\mathbb{R}^k$, mostly of type I. This literature may be categorized roughly as follows for brevity. Broad topics include: (i) basic multidimensional t distributions and their properties; (ii) computations for and approximations to these distributions; and (iii) special aid tables to support their many applications. Their methodological origins encompass: (iv) **multiple comparisons** for means, including Dunnett's [2] pairwise comparison of treatments with a control; (v) **simultaneous confidence** bounds for location parameters; (vi) ranking and selection problems; (vii) **Bayesian** analyses; (viii) the analysis of repeated measurements (*see* **Longitudinal Data Analysis, Overview**); (ix) **prediction** intervals in regression; (x) **diagnostics** for **outliers**; (xi) topics in **estimation**; and (xii) as primary models for error distributions arising in multilinear models as noted. A lengthy and detailed reference list is omitted here; access is readily available through searching electronic databases such as the *Current Index to Statistics*.

In what follows $\mathbb{R}^k$ denotes Euclidean k-dimensional space; abbreviations include *pdf* and *cdf* for probability density and cumulative distribution functions, respectively; and $\mathcal{L}(\mathbf{X})$ designates the law of distribution of $\mathbf{X} = (X_1, \ldots, X_k)' \in \mathbb{R}^k$. In particular, $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution on $\mathbb{R}^k$ having the mean $\boldsymbol{\mu} \in \mathbb{R}^k$ and the positive-definite $(k \times k)$ dispersion **covariance matrix** $\mathrm{V}(\mathbf{X}) = \boldsymbol{\Sigma} = (\sigma_{ij})$; and $\chi^2(\nu)$ identifies the central **chi-square distribution** having ν **degrees of freedom**.

Type I Distributions on $\mathbb{R}^k$

Basic Properties

Suppose that the distribution of $\mathbf{X} \in \mathbb{R}^k$ is given by $\mathcal{L}(\mathbf{X}) = \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{\Sigma})$ and that $\mathcal{L}(\nu S^2/\sigma^2) = \chi^2(\nu)$ independently of $\mathbf{X}$. In practice, S^2 is typically an error mean square from an **analysis of variance** based on ν degrees of freedom. Then, with $\{T_i = X_i/S; 1 \le i \le k\}$, their joint distribution is a type I t distribution on $\mathbb{R}^k$, to be designated as $\mathcal{L}(T_1, \ldots, T_k) = t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Its pdf takes the form

$$f(\mathbf{t}) = C(k, \nu)\left[\frac{1 + (\mathbf{t} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{t} - \boldsymbol{\mu})}{\nu}\right]^{-(\nu+k)/2}, \tag{1}$$

where $C(k, \nu) = \Gamma[(\nu + k)/2]/(\pi\nu)^{k/2}\Gamma(\nu/2)|\boldsymbol{\Sigma}|^{1/2}$. Here $\boldsymbol{\mu}$ is the center of symmetry, $\boldsymbol{\Sigma}$ is the matrix of scale parameters, and ν is the number of degrees of freedom. The distribution is said to be *central* whenever $\boldsymbol{\mu} = \mathbf{0}$, and to be *noncentral* otherwise. The first two moments when defined are $\mathrm{E}(\mathbf{T}) = \boldsymbol{\mu}$ for $\nu > 1$, and $\mathrm{var}(\mathbf{T}) = [\nu/(\nu - 2)]\boldsymbol{\Sigma}$ for $\nu > 2$. For arbitrary $\boldsymbol{\Sigma}$, the one-dimensional marginal distributions are scaled t distributions on $\mathbb{R}^1$. However, if $\boldsymbol{\Sigma}$ is replaced by a positive-definite correlation matrix $\mathbf{R} = [\rho_{ij}]$, then each marginal is a Student's t distribution on $\mathbb{R}^1$, central or noncentral as appropriate, having ν degrees of freedom. These k-dimensional distributions are elliptically contoured from expression (1), and the case $\nu = 1$ yields elliptical **Cauchy** distributions on $\mathbb{R}^k$.

Joint marginal and conditional distributions of $t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ emerge as follows. Partition $\mathbf{T} = [\mathbf{T}_1', \mathbf{T}_2']'$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1', \boldsymbol{\mu}_2']'$, and $\boldsymbol{\Sigma} = [\Sigma_{ij}]$ conformably, with $\mathbf{T}_1 \in \mathbb{R}^r$ and $\mathbf{T}_2 \in \mathbb{R}^t$ such that $r +$

$t = k$. Since $\mathcal{L}(\mathbf{T})$ is elliptical, it follows from the theory of elliptical distributions [1] that the joint marginal distribution of $\mathbf{T}_1$ is $\mathcal{L}(\mathbf{T}_1) = t_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \nu)$, and similarly for $\mathbf{T}_2$. In like manner, the conditional distributions $\mathcal{L}(\mathbf{T}_1|\mathbf{t}_2)$ are elliptical on $\mathbb{R}^r$, having the linear **regression** functions $\boldsymbol{\mu}_{11.2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{t}_2 - \boldsymbol{\mu}_2)$ and scale parameters $\kappa(\mathbf{t}_2)\boldsymbol{\Sigma}_{11.2}$ with $\boldsymbol{\Sigma}_{11.2} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$, where $\kappa(\mathbf{t}_2)$ depends on the conditioning value $\mathbf{t}_2$. These properties hold even without moments, where now $\boldsymbol{\mu}_{11.2}$ represents the center of symmetry and $\boldsymbol{\Sigma}_{11.2}$ the matrix of scale parameters.

Probability Inequalities

Basic inequalities for these distributions are essential. In practice, there is an excess of parameters for $t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, even in the central case with $\boldsymbol{\mu} = \mathbf{0}$, since $(\boldsymbol{\Sigma}, \nu)$ consist of $[k(k+1)+2]/2$ distinct parameters. Owing to limitations of available tables, access to probability inequalities enables the user to employ approximate values from existing tables giving bounds on the required probabilities. In this spirit, we summarize some useful inequalities as follows. Basic references are [7, Chapter 3] and [8, Chapter 9], together with extensive reference lists.

In what follows, $\boldsymbol{\Sigma} = [\sigma_{ij}]$ designates any positive-definite $(k \times k)$ matrix, whereas $\mathbf{R} = [\rho_{ij}]$ denotes a positive-definite correlation matrix. An equicorrelation matrix is denoted by $\boldsymbol{\Xi}(\rho) = [(1-\rho)\mathbf{I}_k + \rho\mathbf{1}_k\mathbf{1}_k']$ for $\{-(k-1)^{-1} < \rho < 1\}$. With these conventions in place, let $P_{\boldsymbol{\Sigma}}(\cdot; \nu)$ be the probability measure for $\mathcal{L}(T_1, \ldots, T_k) = t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$; let $F_{\boldsymbol{\Sigma}}(t_1, \ldots, t_k; \nu)$ be its cdf; and let $\overline{F}_{\boldsymbol{\Sigma}}(t_1, \ldots, t_k) = P_{\boldsymbol{\Sigma}}(T_1 > t_1, \ldots, T_k > t_k; \nu)$. Furthermore, let $F_{\mathbf{D}}(t_1, \ldots, t_k)$ be the cdf of $t_k(\boldsymbol{\mu}, \mathbf{D}, \nu)$, with $\mathbf{D} = \text{diag}(\sigma_{11}, \ldots, \sigma_{kk})$, and similarly for $\overline{F}_D(t_1, \ldots, t_k)$. Finally, identify $G_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k; \nu) = P_{\boldsymbol{\Sigma}}(|T_1| \le a_1, \ldots, |T_k| \le a_k; \nu)$ for the case $\mathcal{L}(\mathbf{T}) = t_k(\mathbf{0}, \boldsymbol{\Sigma}, \nu)$, and similarly $G_D(a_1, \ldots, a_k; \nu)$, with $\mathbf{D} = \text{diag}(\sigma_{11}, \ldots, \sigma_{kk})$ as before. Basic probability inequalities may be summarized as follows.

Property 1. If $\mathcal{L}(\mathbf{T}) = t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, then for fixed but arbitrary $\{a_1, \ldots, a_k\}$ and ν, the function $F_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k; \nu)$ is increasing in each σ_{ij} for all $i \neq j$, while other values are held fixed.

Property 2. Suppose that $\mathcal{L}(\mathbf{T}) = t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. If $\sigma_{ij} \ge 0$ for all $i \neq j$, then $F_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k) \ge F_{\mathbf{D}}(a_1, \ldots, a_k) \ge \prod_{i=1}^k F_i(a_i)$ holds for each fixed $\{a_1, \ldots, a_k\}$, where $F_i(\cdot)$ is the marginal cdf of T_i.

Property 3. Suppose that $\mathcal{L}(\mathbf{T}) = t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. If $\sigma_{ij} \ge 0$ for all $i \neq j$, then $\overline{F}_{\boldsymbol{\Sigma}}(a_1, \ldots, a_k) \ge \overline{F}_{\mathbf{D}}(a_1, \ldots, a_k) \ge \prod_{i=1}^k [1 - F_i(a_i)]$ for arbitrarily fixed $\{a_1, \ldots, a_k\}$.

Property 4. Suppose that $\mathcal{L}(\mathbf{T}) = t_k(\mathbf{0}, \boldsymbol{\Sigma}, \nu)$. Then $G_{\boldsymbol{\Sigma}}(c_1, \ldots, c_k) \ge G_{\mathbf{D}}(c_1, \ldots, c_k) \ge \prod_{i=1}^k G_i(c_i)$ for each fixed set $\{c_1, \ldots, c_k\}$ of positive constants, where $G_i(\cdot)$ is the marginal cdf of $|T_i|$.

Property 5. Suppose that $\mathcal{L}(\mathbf{T}) = t_k[\mathbf{0}, \boldsymbol{\Xi}(\rho), \nu]$. Then for each fixed $a > 0$ and for all real numbers $\{c_1, \ldots, c_k\}$ such that $c_1 + \cdots + c_k = 0$, $P_\rho(|\sum_{i=1}^k c_i T_i| \le a)$ is an increasing function of ρ.

Type II Distributions on $\mathbb{R}^k$

Basic Structure

Type II t distributions on $\mathbb{R}^k$ have origins essentially as follows. Suppose that $\mathcal{L}(\mathbf{X}) = \text{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Independently of $\mathbf{X}$, let $\nu\mathbf{S} = \nu[S_{ij}]$, of order $(k \times k)$, have a central **Wishart distribution** $W_k(\nu, \boldsymbol{\Xi})$ having ν degrees of freedom and the matrix $\boldsymbol{\Xi}$ of scale parameters, such that diagonal elements of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Xi}$ are the same. If we let $\{T_i = X_i/S_i; 1 \le i \le k\}$, with $\{S_i^2 = S_{ii}; 1 \le i \le k\}$ as the diagonal elements of $\mathbf{S}$, then $\mathcal{L}(T_1, \ldots, T_k)$ is said to have a type II t distribution on $\mathbb{R}^k$.

Expressions for pdfs of such distributions are not available in closed form. Nonetheless, probability inequalities for such distributions are known under special structure for dependencies among the elements of $\mathbf{X}$ and of $\mathbf{S}$. An example follows.

Probability Inequalities

Suppose that $\mathcal{L}(\mathbf{X}) = \text{N}_k(\boldsymbol{\mu}, \mathbf{R})$, its correlation matrix $\mathbf{R} = [\rho_{ij}]$ having the structure $\{\rho_{ij} = \kappa_i\kappa_j\omega_{ij}; i \neq j\}$, such that $\{|\kappa_i| \le 1; 1 \le i \le k\}$, where $\boldsymbol{\Omega} = [\omega_{ij}]$ is a positive-definite correlation matrix. Furthermore, let $\mathcal{L}(\nu\mathbf{S}) = W_k(\nu, \boldsymbol{\Lambda})$ independently of $\mathbf{X}$, such that $\boldsymbol{\Lambda} = [\lambda_{ij}]$ is a correlation matrix with structure $\{\lambda_{ij} = \lambda_i\lambda_j; i \neq j\}$, for some $\{\lambda_1, \ldots, \lambda_k\}$.

Let $\{T_i = X_i/S_i; 1 \le i \le k\}$, with $\{S_i^2 = S_{ii}; 1 \le i \le k\}$, and denote the cdf $G_k(c_1, \ldots, c_k) = \Pr(|T_1| \le c_1, \ldots, |T_k| \le c_k)$. With these conventions, the following monotone properties of $G_k(c_1, \ldots, c_k)$ apply, as shown in [7, Theorem 3.1.2].

Property 6. For each fixed set of positive numbers $\{c_1, \ldots, c_k\}$, the function $G_k(c_1, \ldots, c_k)$ is: (i) strictly increasing in each $\kappa_i \in [0, 1]$ with other parameters held fixed; (ii) strictly decreasing in each $\kappa_i \in [-1, 0]$ with other parameters held fixed; and (iii) strictly increasing in each $|\lambda_i|$ with other parameters held fixed.

Peakedness Ordering

The comparative concentration of probabilities is an essential concept for distributions on $\mathbb{R}^k$. Following Sherman [6], the probability measure $\mu(\cdot)$ is said to be *more peaked about* $\mathbf{0} \in \mathbb{R}^k$ than $\nu(\cdot)$ if and only if $\mu(A) \ge \nu(A)$ for every set A in the class $\mathbf{C}_k$ comprising the compact convex subsets of $\mathbb{R}^k$ that are symmetric under reflection about $\mathbf{0} \in \mathbb{R}^k$, i.e. $\mathbf{x} \in A$ implies $-\mathbf{x} \in A$. For two type I t distributions $t_k(\mathbf{0}, \boldsymbol{\Sigma}, \nu)$ and $t_k(\mathbf{0}, \boldsymbol{\Omega}, \nu)$ on $\mathbb{R}^k$ having ordered scale matrices, the following inequality applies. Sufficiency is shown in [3], and necessity in [4].

Property 7. Let $t_k(\mathbf{0}, \boldsymbol{\Sigma}, \nu)$ and $t_k(\mathbf{0}, \boldsymbol{\Omega}, \nu)$ be type I multivariate t distributions on $\mathbb{R}^k$ having ordered scale matrices such that $\boldsymbol{\Omega} - \boldsymbol{\Sigma}$ is positive-semidefinite, and let $P_{\boldsymbol{\Sigma}}(\cdot; \nu)$ and $P_{\boldsymbol{\Omega}}(\cdot; \nu)$ be their corresponding probability measures. Then $P_{\boldsymbol{\Sigma}}(\cdot; \nu)$ is more concentrated about $\mathbf{0}$ than $P_{\boldsymbol{\Omega}}(\cdot; \nu)$ in the sense that $P_{\boldsymbol{\Sigma}}(A; \nu) \ge P_{\boldsymbol{\Omega}}(A; \nu)$ for every set A in the class $\mathbf{C}_k$.

References

[1] Cambanis, S., Huang, S. & Simons, G. (1981). On the theory of elliptically contoured distributions, *Journal of Multivariate Analysis* **11**, 368–385.

[2] Dunnett, C.W. (1955). A multiple comparisons procedure for comparing several treatments with a control, *Journal of the American Statistical Association* **50**, 1096–1121.

[3] Fefferman, C., Jodeit, M. & Perlman, M.D. (1972). A spherical surface measure inequality for convex sets, *Proceedings of the American Mathematical Society* **33**, 114–119.

[4] Jensen, D.R. (1984). Ordering ellipsoidal measures: scale and peakedness orderings, *SIAM Journal on Applied Mathematics* **44**, 1226–1231.

[5] Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

[6] Sherman, S. (1955). A theorem on convex sets with applications, *Annals of Mathematical Statistics* **25**, 763–766.

[7] Tong, Y.L. (1980). *Probability Inequalities in Multivariate Distributions*. Academic Press, New York.

[8] Tong, Y.L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

(*See also* **Multivariate Distributions, Overview**)

D.R. JENSEN

Multivariate Techniques, Robustness

Many statistical techniques are based on assumptions about either the form or the structure of the data to which they are to be applied. This is as much the case in **multivariate analysis** as it is in other branches of statistics. For example, a **hypothesis test** about a population mean vector may assume normality of the sampled data (an assumption about the *form* of the data), while a comparison between the mean vectors of two populations may additionally require equality of the population dispersion matrices (an assumption about the data *structure*). In practical applications it is accepted that such assumptions will at best only be approximations to the truth, but the hope is that the validity and outcome of the technique will be largely unaffected within the sort of range of departures from these assumptions that might reasonably be encountered for real data.

In the *Dictionary of Statistical Terms* [17], a statistical procedure is described in broad terms as being *robust* if it is "not very sensitive to departure from the assumptions on which it depends". The **robustness** of a technique is thus, loosely, the extent to which it is unaffected by such departures, so that a study of the robustness of standard techniques is important from the point of view of practical statistics. In the univariate situation such study is generally restricted to inferential techniques arising from either hypothesis testing or **estimation** (whether point or interval). A theoretical distinction that is sometimes made in these studies is one between *criterion* robustness and *inference* robustness: the former is when the behavior of any criterion on which the inference depends (e.g. a particular test statistic or estimator) is largely unaffected by departures from assumptions, so that by implication the resulting inferences are also unchanged; in the latter case, the behavior of the criterion may be subject to appreciable changes under departures from assumptions but the resulting inferences are nevertheless unaffected. However, in practical terms this distinction is rather academic – the practitioner merely wants to know whether a technique is robust or not, and if not then what are the main areas of sensitivity.

In the multivariate case interest has again mainly focused on inferential techniques, particularly on hypothesis testing. However, the issue is now wider, as descriptive techniques such as canonical variate analysis (*see* **Canonical Correlation**) and modeling techniques such as **factor analysis** also invoke assumptions about the data, and thus have to be examined for robustness. We therefore consider three broad areas: inference, ordination and classification, and latent variable models, and give an overview of the robustness of the main techniques encountered under each of them. We should be clear at the outset about our frame of reference. We are here concerned only with the robustness of *standard* multivariate techniques to departures from those assumptions on which they are founded, and our aim is to give a general overview rather than to discuss the many detailed points that have been raised in the literature. What we are *not* concerned with is consideration of any specifically robust procedures, by which is meant procedures that remain unaffected by **outliers** or other data contaminants. This is a different aspect of robustness to the one given above, and such procedures are considered elsewhere in these volumes.

Inference

Most standard multivariate inferential procedures are concerned either with statements about the mean vector and/or the dispersion matrix of a single **multivariate normal distribution**, or with the comparison of these parameters across several multivariate normal populations. Moreover, there are now many special estimators of these parameters that are designed to be robust with respect either to outlying observations or to non-normality, so we will not explicitly consider estimation here. However, most extant tests of hypotheses about these parameters are based on the assumption of normality and (for multiple-population comparisons) equality of dispersion matrices, so we focus in this section on the robustness of these tests. It is convenient to subdivide the discussion into single-population and multiple-population tests, and within this subdivision to consider tests of means and tests of dispersion matrices separately.

For single-population tests of the mean vector, **Hotelling's T^2** is the most common normal-based test statistic. There have been various **simulation** studies of the performance of this test statistic (e.g. [4, 7, 16], and [24]). The consensus view is that the test is

more sensitive to **skewness** than to **kurtosis** of the underlying distribution: it is reasonably robust with regard to the nominal significance level (*see* **Level of a Test**) when the underlying distribution is symmetrical, but the actual significance level is greater than the nominal level for skew distributions, and the discrepancy increases both with increasing skewness and with number of variables p.

Single-population tests of dispersion matrices, by contrast, are more sensitive to kurtosis than to skewness in the parent population [19]. There are various standard hypotheses about dispersion matrices, but probably the most common is for independence (either mutual or blockwise). Reassuringly, the study by Pillai & Hsu [21] shows that none of the usual test statistics is seriously affected if departures from normality are only slight.

Turning to multiple-population tests, it is convenient to consider tests of equality of dispersion matrices first (as tests of equality of means usually make this prior assumption as well). Here, unfortunately, the situation is *not* very promising. The standard test for equality of a set of dispersion matrices is the **likelihood ratio test** and its associated approximations [22, p. 449], and available studies (e.g. [10], [15], and [20]) demonstrate its extreme nonrobustness to non-normality. In particular, it is very sensitive to kurtosis, and a significant value of the test statistic could be due equally to kurtosis as to departures from the null hypothesis. Indeed, Layard [15] goes so far as to say that its usefulness is suspect.

Turning finally to tests of equality of a set of g population means, we encounter the most complicated setup of those considered in this Section. Essentially we are in the realms of one-way **multivariate analysis of variance** (MANOVA), in which we assume that individual observations come from multivariate normal distributions having possibly different means but a common dispersion matrix. The complicating feature is that there are now several "standard" test statistics for the null hypothesis (each one sensitive to a different form of departure from the null hypothesis). The likelihood ratio statistic is commonly known as **Wilks' lambda** (λ), the **union–intersection principle** yields **Roy's maximum root**, $\phi_{\max}$, and the two other most common statistics are the **Lawley–Hotelling trace** T_g^2, and the **Pillai trace** V. (All these statistics reduce to the same quantity, Hotelling's generalized T^2, in the case of $g = 2$ populations.)

One of the earliest studies of these statistics was that by Ito & Schull [11], who demonstrated that when the sample sizes are the same from each population then moderate heterogeneity of dispersion matrices does not affect T_g^2 seriously, but when there is inequality in sample sizes then both the significance level and power of T_g^2 can be affected seriously. Analytical investigations of the effect of non-normality on λ and $\phi_{\max}$ were conducted by Davis [5, 6]. Broadly speaking, both statistics are reasonably robust to non-normality providing that the number of residual **degrees of freedom** is moderate to large, but if the underlying populations have large skewness or kurtosis then the effects on the true significance level can be large when there are fewer than about 30 residual degrees of freedom. Generally, effects on significance level for excess skewness are opposite to those for excess kurtosis, while each effect is itself reversed as sample sizes become more unbalanced. The widest comparison of all four statistics was in the 1974 **Monte Carlo** study by Olson [20], although he only considered the case of equal sample sizes from all populations. His conclusions were that kurtosis affects significance levels only mildly, the powers of all tests suffer under kurtosis, and inequality of dispersion matrices is the most serious violation. The worst statistic of the four in these respects is $\phi_{\max}$ (not surprisingly, as it is based on just a single extreme eigenvalue), while the best is Pillai's V. Moreover, robustness generally improves as either the number of groups g or the number of variables p is *reduced* (so, including variables "for the sake of it" is not to be recommended!).

Ordination and Classification

Ordination is the geometrical representation of multivariate samples as points in a low-dimensional space, while classification is the process of allocating individuals in multivariate samples to distinct populations. Often, prior ordination of multivariate data will facilitate subsequent classification. Most ordination techniques are purely descriptive and do not rely on any assumptions about the data, so there is no need to study their robustness. However, a genuine concern is the effect that outliers, contaminations of various types, or perturbations of the data might have on the results of the analysis, and some techniques have been

subjected to scrutiny with this in mind. Sibson [23] has quantified the effect that perturbations of the input dissimilarities might have on a metric scaling representation of the data. This quantification is given in terms of the **Procrustes** statistic between the original and perturbed configurations of points. Langron & Collins [14] have provided a similar quantification for the results of generalized Procrustes analysis, while Jackson [12, pp. 365–369] surveys the methods that have been developed to combat the effects of outliers on **principal components analysis**.

The one ordination technique which *does* make an explicit assumption about the data is canonical variate analysis, in which multivariate observations taken from g separate populations are represented geometrically by points in space in such a way that the ratio of between-population to within-population scatter is maximized (thereby highlighting population differences). This technique assumes a common dispersion structure across populations, and so is closely related to one-way multivariate analysis of variance. Robustness considerations associated with hypothesis testing in the latter situation have already been discussed above, but also of relevance is the robustness or otherwise of the ordination of the populations in the space of the canonical variates. Unfortunately, very little appears to have been done on this aspect. Campbell [3] offers some heuristic approaches for examining the effect of within-population heterogeneity in specific practical cases, but all that can be said in general is that standard canonical variate analysis will give increasingly misleading results as the population dispersion matrices become more disparate [18, p. 194].

By contrast to ordination, the area of discrimination and **classification** has received considerable attention with regard to robustness. The primary focus has been on the case of two groups, and specifically on the linear discriminant function (*see* **Discriminant Analysis, Linear**). This function was first derived in 1936 by Fisher [8] from the point of view of discriminating between two populations. Its derivation in this context was simply as the linear combination of variables that maximizes the ratio of squared difference of sample means to pooled within-sample variance, without recourse to any assumptions about the form or structure of the data. In this (discrimination) usage, therefore, the function is reasonably robust to a variety of data types. However, the function also arises as the optimal function for allocating future individuals to one of two multivariate normal populations that have equal dispersion matrices, where optimality is judged in terms of the expected cost incurred by misclassification of future individuals [25]. If the restriction to equal dispersion matrices is relaxed, but the assumption of normality is retained, then the optimal function becomes a quadratic discriminant function. These two functions are used very often in practice and so have been subjected to many robustness studies, in which the effects of departures from normality and from equality of population dispersion matrices have been examined. The criterion on which this examination is based is usually the error rate incurred in the classification of future individuals (*see* **Misclassification Error**). There are many complicated technical issues involved in some of these studies, and space limitations preclude any detailed discussion of them here; for an excellent summary survey see [18, pp. 152–161]. However, some very broad conclusions that may be drawn are that: both functions will perform very poorly if the data are continuous but non-normal with heavy tails and large skewness, while the quadratic function may still perform reasonably well if the distributions are symmetric even through heavy-tailed; the linear function will perform reasonable well on **binary data** (scored 0 and 1) providing that the true log **likelihood ratio** increases as the number of subjects who score 1 increases, but not otherwise; and the allocatory performance of the linear function with mixtures of discrete and continuous variables depends upon the similarity of the correlation patterns among the continuous variables in the two groups, being good when the similarity is high and poor when the similarity is low.

Latent Variable Models

A popular technique for the analysis of covariance structure, particularly in the social sciences, is factor analysis. This technique had been used since the early part of the twentieth century, but computational and theoretical advances in the 1960s and early 1970s by Jöreskog and co-workers not only greatly popularized its use but also led to developments in more general covariance structure analysis. The whole area now comprises a series of techniques that can be categorized under the heading of **structural equation models** [2]. All these techniques rest on the assumption of underlying sets of latent variables for the

explanation of observed covariances among a set of manifest variables. Models of varying levels of complexity can be built and analyzed with the aid of such software packages as **LISREL** (an acronym for "linear structural relationships", developed by Jöreskog & Sörbom [13]) or EQS (denoting "structural equations", developed by Bentler [1]). The fitting of these models is achieved by optimizing one of a number of available "fit functions" which measure the discrepancy between data and model. While some of these functions are fairly general ones, others have been derived from distributional assumptions about the data. In particular, the "default" choice of many researchers is the function resulting from **maximum likelihood** assuming normality of manifest variables, so again a study of robustness is an important consideration.

As in discrimination and classification, effects of non-normality on the results of the maximum-likelihood technique can be divided up according as the manifest variables are continuous but non-normal, categorical (and hence non-normal), or a mixture of the two types. For continuous non-normal variables, Bollen [2, p. 418] summarizes the available studies by saying that violation of normality does not generally affect the **bias** or **consistency** of estimates, but excessive kurtosis usually eliminates asymptotic **efficiency** and makes the estimated asymptotic covariance matrix and the chi-square **goodness-of-fit** test potentially inaccurate. In the case of discrete or categorical variables, the evidence from available studies [2, p. 435] suggests that not much distortion will occur on using the maximum likelihood procedure if the skewnesses and kurtoses of the variables are close to normal distribution values, but the chi-square goodness-of-fit statistic tends to be too big if the variables are highly skewed. However, as yet there has been little investigation in this area (and virtually nothing in the case of mixed continuous/categorical data), so these results can only be considered to be preliminary ones.

Finally, Henly [9] reports a large-scale Monte Carlo investigation of the robustness of a range of estimators to misspecification of the fit function. As well as concluding that robustness of the maximum likelihood and the normal theory generalized least squares estimators cannot be taken for granted, she provides a comprehensive survey and reference list of previous work in this area.

In summary, therefore, across all the different areas the broad concerns are the effects that departures from normality and/or equal dispersion matrices have on the standard techniques, in particular on type I error (*see* **Level of a Test**) and **power**. As a single overall conclusion it is probably fair to say that departures from normality are relatively unimportant providing that the true distribution is approximately symmetric with kurtosis near the normal value, but lack of equality of dispersion matrices can have much more serious consequences for those techniques that are aimed at analyzing grouped data.

References

[1] Bentler, P.M. (1985). *Theory and Implementation of EQS: A Structural Equations Program*. BMDP Statistical Software, Los Angeles.

[2] Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

[3] Campbell, N.A. (1984). Canonical variate analysis with unequal covariance matrices: generalizations of the usual solution, *Mathematical Geology* **16**, 109–124.

[4] Chase, G.R. & Bulgren, W.G. (1971). A Monte Carlo investigation of the robustness of T^2, *Journal of the American Statistical Association* **66**, 499–502.

[5] Davis, A.W. (1980). On the effects of moderate multivariate nonnormality on Wilks' likelihood ratio criterion, *Biometrika* **67**, 419–427.

[6] Davis, A.W. (1982). On the effects of moderate multivariate nonnormality on Roy's largest root test, *Journal of the American Statistical Association* **77**, 896–900.

[7] Everitt, B.S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one and two-sample T^2 statistic, *Journal of the American Statistical Association* **74**, 48–51.

[8] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–184.

[9] Henly, S.J. (1993). Robustness of some estimators for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **46**, 313–338.

[10] Ito, K. (1969). On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures, in *Multivariate Analysis*, Vol. II, P.R. Krishnaiah, ed. Academic Press, New York, pp. 87–120.

[11] Ito, K. & Schull, W.J. (1964). On the robustness of the T_0^2 test in multivariate analysis of variance when variance-covariance matrices are not equal, *Biometrika* **51**, 71–82.

[12] Jackson, J.E. (1991). *A User's Guide to Principal Components*. Wiley, New York.

[13] Jöreskog, K.G. & Sörbom, D. (1986). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods*. Scientific Software Inc., Mooresville.

[14] Langron, S.P. & Collins, A.J. (1985). Perturbation theory for generalized Procrustes analysis, *Journal of the Royal Statistical Society, Series B* **47**, 277–284.

[15] Layard, M.W.J. (1974). A Monte Carlo comparison of tests for equality of covariance matrices, *Biometrika* **61**, 461–465.

[16] Mardia, K.V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test, *Applied Statistics* **24**, 163–171.

[17] Marriott, F.H.C. (1990). *A Dictionary of Statistical Terms*, 5th Ed. Longman, Singapore.

[18] McLachlan, G.J. (1992). *Discriminant Analysis and Pattern Recognition*. Wiley, New York.

[19] Muirhead, R.J. & Waternaux, C.M. (1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for non-normal populations, *Biometrika* **67**, 31–43.

[20] Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

[21] Pillai, K.C.S. & Hsu, Y.S. (1979). Exact robustness studies of the test of independence based on four multivariate criteria and their distribution problems under violations, *Annals of the Institute of Statistical Mathematics* **31**, 85–101.

[22] Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

[23] Sibson, R. (1979). Studies in the robustness of multidimensional scaling: perturbation analysis of classical scaling, *Journal of the Royal Statistical Society, Series B* **41**, 217–229.

[24] Srivastava, M.S. & Awan, H.M. (1982). On the robustness of Hotelling's T^2-test and distribution of linear and quadratic forms in sampling from a mixture of two multivariate normal populations, *Communications in Statistics - Theory and Methods* **11**, 81–107.

[25] Welch, B.L. (1939). Note on discriminant functions, *Biometrika* **31**, 218–220.

W.J. KRZANOWSKI

Multivariate Weibull Distribution

Weibull distributions on $\mathbb{R}^1$ have wide usage in modeling cumulative damage, fatigue, life lengths, and as tolerance distributions for quantal responses. These distributions figure prominently in **survival analysis** as models for survival distributions (*see* **Parametric Models in Survival Analysis**). Applications in the biomedical sciences include survival analysis, cumulative damage to the liver and other organs, systems reliability for pacemakers and prosthetic devices in biomedical engineering, and life lengths of surgical repairs and organ duration including transplants. The typical cumulative distribution function (*cdf*) is $F(x) = 1 - \exp(-\lambda x^{\alpha})$, with λ and α as the scale and shape parameters, respectively. The *reliability function* $\overline{F}(x)$, also called the *survival function* $S(x)$, is given by $\overline{F}(x) = 1 - F(x) = \Pr(X > x) = S(x)$. The *failure rate function*, also known as the *intensity function*, the **hazard rate** *function*, and the force of mortality *function*, is given by $r(x) = f(x)/\overline{F}(x)$, with $f(x)$ as the probability density function (*pdf*) of the distribution (*see* **Survival Distributions and Their Characteristics**). In what follows $\mathcal{L}(X) = W(\lambda, \alpha)$ designates the Weibull model on $\mathbb{R}^1$, with $\mathcal{L}(X)$ as its law of distribution, and *iid* refers to independent and identically distributed **random variables**. Weibull distributions on $\mathbb{R}^n$ encompass joint distributions exhibiting essential Weibull characteristics. Since the distribution of $X^{1/\alpha}$ is Weibull on $\mathbb{R}^1$ whenever X has an exponential distribution, multivariate Weibull distributions associate in a natural way with multivariate exponential distributions of various types. Details follow.

A characteristic property of the one-dimensional Weibull family is its closure under the operation of taking minima. Specifically, if $\{X_1, \ldots, X_n\}$ are iid random variables on $[0, \infty)$, and if $Y = \min\{X_1, \ldots, X_n\}$, then $\mathcal{L}(Y)$ is Weibull on $\mathbb{R}^1$ if and only if each $\mathcal{L}(X_i)$ is Weibull on $\mathbb{R}^1$ for $1 \le i \le n$. See Dubey [6] for further details. For additional characterizations let $\mathcal{L}(X)$ be **exponential** with unit mean, and let Z have a stable distribution on $\mathbb{R}^1$ with index α, independently of X. Then the Weibull distribution, $\mathcal{L}(T) = W(1, \alpha)$, can be characterized in terms of X as $\mathcal{L}(T) = \mathcal{L}(X^{1/\alpha})$ as noted, and in terms of (X, Z) as $\mathcal{L}(T) = \mathcal{L}(XZ^{-1})$; see [21] and [22].

On $\mathbb{R}^n$ let $\overline{F}(t_1, \ldots, t_n) = \Pr(T_1 > t_1, \ldots, T_n > t_n)$ be the joint survival function of the positive random variables $[T_1, \ldots, T_n]$. Basic properties of $\overline{F}(s, t)$ for $n = 2$, with obvious extensions, are as follows. The marginal survival function for S is given by $\overline{F}(s, 0)$, for $U = \min\{S, T\}$ is $\overline{F}(u, u)$, and for $V = \min\{aS, bT\}$ is $\overline{F}(av, bv)$ under arbitrary scaling. Let $R(\mathbf{t}) = -\log \overline{F}(\mathbf{t})$. Then the *hazard gradient* is the vector $\mathbf{r}(\mathbf{t}) = [\partial R(\mathbf{t})/\partial t_1, \ldots, \partial R(\mathbf{t})/\partial t_n] = [r_1(\mathbf{t}), \ldots, r_n(\mathbf{t})]$ when defined. These have been studied in Johnson & Kotz [11] and in Shaked [20], where bounds are provided, and elsewhere.

The foregoing facts are central to the study of multivariate distributions exhibiting Weibull characteristics. Many multivariate distributions have Weibull marginals; five classes may be identified here as follows:

C1. $[T_1, \ldots, T_n]$ are independent such that $\{\mathcal{L}(T_i) = W(\lambda_i, \alpha); 1 \le i \le n\}$.

C2. $[T_1, \ldots, T_n]$ are generated as $\{T_i = \min(X_J; i \in J); 1 \le i \le n\}$. Here, $\mathcal{J}$ is a class of nonempty subsets of $\{1, 2, \ldots, n\}$ such that for each i, $i \in J$ for some $J \in \mathcal{J}$, whereas $\{X_J; J \in \mathcal{J}\}$, are independent random variables having distributions $\{W(\lambda_J; \alpha), J \in \mathcal{J}\}$.

C3. For arbitrary positive constants $\{a_1, \ldots, a_n\}$, $U = \min\{a_1 X_1, \ldots, a_n X_n\}$ has a Weibull distribution $\boldsymbol{W(\lambda(\mathbf{a}), \alpha)}$ on $\mathbb{R}^1$ for some $\lambda(\mathbf{a}) = \lambda(a_1, \ldots, a_n) > 0$ and $\alpha > 0$.

C4. $[T_1, \ldots, T_n]$ have a joint distribution with Weibull minima such that $U_s = \min\{T_i; i \in S\}$ has a Weibull distribution $W(\lambda_s, \alpha)$ on $\mathbb{R}^1$ for some $\lambda_s > 0$, for every nonempty subset S of $\{1, 2, \ldots, n\}$.

C5. $[T_1, \ldots, T_n]$ have a joint distribution with arbitrary Weibull marginals $\{\mathcal{L}(T_i) = W(\lambda_i, \alpha_i); 1 \le i \le n\}$, on $\mathbb{R}^1$.

Class C2 comprises joint distributions of minima of overlapping subsets of independent Weibull variates, as in [5] and [14]. Related work by Shaked [20] considers minima over sets of random size.

Examples of **bivariate distributions** in these classes are known. Consider the typical function $\overline{F}(t_1, t_2) = \exp\{-[\lambda_1 c_1^{\alpha} t_1^{\alpha} + \lambda_2 c_2^{\alpha} t_2^{\alpha} + \lambda_{12} \max(c_1^{\alpha} t_1^{\alpha}, c_2^{\alpha} t_2^{\alpha})]\}$ on the positive quadrant, such that $\alpha > 0, c_1 > 0, c_2 > 0, \lambda_1 > 0, \lambda_2 > 0,$ and $\lambda_{12} \ge 0$.

If $c_1 \neq c_2$ and $\lambda_{12} > 0$, then this distribution belongs to class C3. Furthermore, consider functions of the type $\overline{F}(t_1, t_2) = \exp\{-[\lambda_1 t_1^{\alpha_1} + \lambda_2 t_2^{\alpha_2} + \lambda_{12} \max(t_1^{\alpha_1}, t_2^{\alpha_2})]\}$ such that $\alpha_1 > 0, \alpha_2 > 0, \lambda_1 > 0, \lambda_2 > 0$, and $\lambda_{12} \geq 0$. If $\alpha_1 \neq \alpha_2$, then this distribution belongs to class C5. If $\alpha_1 = \alpha_2 = \alpha$ and $\lambda_{12} > 0$, then this distribution belongs to class C2. The typical marginal distribution in the class C4 takes the form $\Pr[\min_{i \in S}(T_i) > t] = \exp(-\lambda_S t^{\alpha})$ for some $\lambda_S > 0$. For further reference see Lee [13].

A subclass of C5, due to Marshall & Olkin [17], is generated as $[T_1, \ldots, T_n] = [X_1^{1/\alpha_1}, \ldots, X_n^{1/\alpha_n}]$, where $[X_1, \ldots, X_n]$ follow the multivariate exponential distribution of those authors (*see* **Multivariate Outliers**). The typical survival function takes the form

$$\overline{F}(\mathbf{t}) = \exp\left[\sum_J \lambda_J \max_{i \in J}(\mathbf{t}_i^{\alpha})\right]$$

with $\alpha > 0$ and $\lambda_J > 0$ for $J \in \mathcal{J}$, where sets J are elements of the class $\mathcal{J}$ consisting of nonempty subsets of $\{1, \ldots, n\}$, such that, for each i, it is true that $i \in J$ for some $J \in \mathcal{J}$. For extensions see Arnold [1]. Weibull distributions of these types are studied extensively in [19].

Lee [13] considered distributions of the foregoing types and proposed the classification scheme adopted here. He further established that the inclusion relations C1 $\subset$ C2 $\subset$ C3 $\subset$ C4 $\subset$ C5 are strict.

Distributions on $\mathbb{R}^n$ are said to have the *increasing failure rate average (IFRA)* property whenever $\mathrm{E}[h(T_1, \ldots, T_n)] \leq \{\mathrm{E}[h^{\gamma}(T_1/\gamma, \ldots, T_n/\gamma)]\}^{1/\gamma}$ for every nonnegative continuous nondecreasing function $h(\cdot)$ and for all $0 < \gamma \leq 1$. Block & Savits [2] show that the Weibull distributions on $\mathbb{R}^n$ of Marshall & Olkin [17], and distributions in the class C2, are all IFRA for the cases $\alpha_i \geq 1$, $1 \leq i \leq n$, and $\alpha \geq 1$, respectively. Related developments pertaining to multivariate hazard rates and their applications are undertaken by Johnson & Kotz [11].

Krishnaiah [12] generalized from Weibull distributions for the case $n = 2$ on taking $[T_1, T_2] = [X_1^{1/\alpha_1}, X_2^{1/\alpha_2}]$ such that $[X_1, X_2]$ follow a bivariate **gamma distribution**. Maxim et al.. [18] considered the choice of optimal designs under multivariate Weibull sensitivity models. Connections are known between particular multivariate Weibull distributions and the multivariate **extreme-value** distributions of Gumbel [7]; see also Johnson & Kotz [10, p. 249 ff].

More recent studies are germane. Repeated failure time measurements are modeled in [3] using conditionally independent Weibull failure times $W(\lambda, \alpha)$, whereas α is then assigned a gamma distribution over subjects. The unconditional mixture on $\mathbb{R}^n$ is a multivariate Burr distribution (see [10, p. 288 ff.]). If, instead, α is given a stable distribution as in [4], then the resulting mixture is genuinely multivariate Weibull having Weibull marginals. See also [8, 9, 15], and [16], where related matters and further mixtures are treated. The univariate Weibull distribution is often employed in discussions of **accelerated life** testing and failure models. Parallel developments are reported in [15] for the class of multivariate Weibull distributions as treated in [8].

References

[1] Arnold, B.C. (1967). A note on multivariate distributions with specified marginals, *Journal of the American Statistical Association* **62**, 1460–1461.

[2] Block, H.W. & Savits, T.H. (1980). Multivariate increasing failure rate average distributions, *Annals of Probability* **8**, 793–801.

[3] Crowder, M. (1985). A distributional model for repeated failure time measurements, *Journal of the Royal Statistical Society, Series B* **47**, 447–452.

[4] Crowder, M. (1989). A multivariate distribution with Weibull connections, *Journal of the Royal Statistical Society, Series B* **51**, 93–107.

[5] David, H.A. (1974). Parametric approaches to the theory of competing risks, in *Reliability and Biometry, Statistical Analysis of Life Length*, F. Proschan & R.J. Serfling, eds. SIAM, Philadelphia, pp. 275–290.

[6] Dubey, S.D. (1966). Characterization theorems for several distributions and their applications, *Journal of Industrial Mathematics* **16**, 1–22.

[7] Gumbel, E.J. (1958). *Statistics of Extremes*, 2nd Ed. Columbia University Press, New York.

[8] Hougaard, P. (1986). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678.

[9] Jaisingh, L.R., Dey, D.K. & Griffith, W.S. (1993). Properties of a multivariate survival distribution generated by a Weibull and inverse-Gaussian mixture, *IEEE Transactions on Reliability* **42**, 618–622.

[10] Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

[11] Johnson, N.L. & Kotz, S. (1975). A vector multivariate hazard rate, *Journal of Multivariate Analysis* **5**, 53–66.

[12] Krishnaiah, P.R. (1977). On generalized multivariate gamma type distributions and their applications in reliability, in *The Theory and Applications of Reliability*, Vol. 1, C.P. Tsokos & I.N. Shimi, eds. Academic Press, New York, pp. 475–494.

[13] Lee, L. (1979). Multivariate distributions having Weibull properties, *Journal of Multivariate Analysis* **9**, 267–277.

[14] Lee, L. & Thompson, W.A. (1974). Results on failure time and pattern for the series system, in *Reliability and Biometry, Statistical Analysis of Life Length*, F. Proschan & R.J. Serfling, eds. SIAM, Philadelphia, pp. 291–302.

[15] Lu, J.C. (1990). Least squares estimation for the multivariate Weibull model of Hougaard based on accelerated life test of system and component, *Communications in Statistics – Theory and Methods* **19**, 3725–3739.

[16] Lu, J.C. (1992). Effects of dependence on modeling system reliability and mean life via a multivariate Weibull distribution, *Journal of the Indian Association for Productivity, Quality, and Reliability* **17**, 1–22.

[17] Marshall, A.W. & Olkin, I. (1967). A multivariate exponential distribution, *Journal of the American Statistical Association* **62**, 30–44.

[18] Maxim, L.D., Hendrickson, A.D. & Cullen, D.E. (1977). Experimental design for sensitivity testing: the Weibull model, *Technometrics* **19**, 405–412.

[19] Moeschberger, M.L. (1974). Life tests under dependent competing causes of failure, *Technometrics* **16**, 39–47.

[20] Shaked, M. (1977). Bounds for the distributions and hazard gradients of multivariate random minimums, in *The Theory and Applications of Reliability*, Vol. 1, C.P. Tsokos & I.N. Shimi, eds. Academic Press, New York, pp. 227–242.

[21] Shanbhag, D.N. & Sreehari, M. (1977). On certain self-decomposable distributions, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **38**, 217–222.

[22] Williams, E.J. (1977). Some representations of stable random variables as products, *Biometrika* **64**, 167–169.

D.R. JENSEN

Mutagenicity Study

In many laboratory experiments, data are analyzed via various statistical approaches, each depending on the nature of the endpoint and aspects of the particular assay under study [39]. Among major endpoints of interest, damage to genetic components such as DNA is of increasing interest. Mutation induction, or *mutagenicity*, and other forms of genotoxicity represent fruitful areas of study, particularly as continuing biotechnological advances allow for more precise genetic study.

Mutagenicity is perhaps the most varied of any toxicological endpoint, since the potential mechanisms of genetic damage are so numerous. Indeed, there exist well over 100 different genotoxicological assays, employing all sorts of animal and microbial systems [18, 50]. Statistical themes associated with mutagenicity experiments include proper identification of the **sampling distribution** [24, 31, 40], construction of appropriate biomathematical models to characterize **dose response** and other features of genetic damage [2, 3, 17, 22, 30], **sample size determination** [7, 30, 35, 42], and the selection of statistical approaches associated with these models and methods [28]. A basic introduction to these issues is given in the compilation by Kirkland [21]. Here, attention will be directed at some of the basic issues of dose–response analysis encountered in mutagenesis experiments.

Microbial Systems in Environmental Mutagenesis: The *Salmonella* Assay

Perhaps the most well known of all modern mutagenicity assays is the Ames/*Salmonella* microsome assay [33], employing the bacteria *Salmonella typhimurium* to identify damage to DNA after exposure to toxic agents. The assay is based on strains of the bacterium unable to synthesize histidine, an amino acid required for growth. This production deficiency can be reversed into a production capability via point mutations at selected sites on the bacterial genome. DNA damage is indicated by mutation of the bacteria from histidine-dependency to a self-sustaining state. Thus, the mutated cells can grow in a microenvironment (such as a Petri plate) containing only minimal amounts of histidine; greater mutant yield at higher exposures to the environmental chemical suggests that mutagenesis increases with increasing dose. Observational accuracy of the assay is enhanced by use of a selective medium for the growth environment, so that only mutant colonies grow after exposure to the environmental toxin.

The observations in such a microbial assay are the mutated colony counts, Y_{ij}, for the jth replicate plate at the ith dose ($i = 1, \ldots, T; j = 1, \ldots, R_i$). Commonly, the number of replicate plates is taken to be constant, e.g. $R_i = 3$, for all i. (Table 1 presents an example of data from a *Salmonella* assay, using the chemical agent 1,3-butadiene.) A natural candidate for the sampling distribution of Y_{ij} is the **Poisson distribution** for count variables. Indeed, many processes that drive the Poisson sampling assumption are active with the colony count data seen here. These include:

1. The microbes' responses are independent.
2. Each locally organized group of microbes experiences the same environment (ignoring controlled variables such as dose).
3. The plated number of microbes is large, say 10^8.
4. The probability that a plated microbe gives rise to a visible mutant colony is small, between 10^{-5} and 10^{-9} [30].

A fifth assumption crucial to adoption of the Poisson distribution, however, is that both the environment and the number of microbes plated should remain relatively stable across replicate plates (within a dose group). If this is not the case, extra-Poisson variability results, and one is forced to model or otherwise account for the consequent **overdispersion** in Y_{ij}.

Table 1 *Salmonella* mutagenicity results for 1,3-butadiene in strain TA1535

Dose (ppm)	Mutants per plate, Y_{ij}			Average plate count, $\overline{Y}_i$
0.000	20	31	27	26.0
0.002	100	92	89	93.7
0.007	147	123	178	149.3
0.014	216	170	181	189.0
0.020	176	154	183	171.0
0.030	154	153	149	152.0

Testing for Extra-Poisson Variability

Testing for extra-Poisson variability is critical when analyzing count data from mutagenicity experiments. Standard trend tests for Poisson-distributed counts (see Trend Tests for Counts and Proportions) become too sensitive in the presence of overdispersion, increasing type I error rates above the nominally specified α level [38]. Thus, the data analyst must be aware of the nature and level of any overdispersion in the data before selecting a test for dose response.

Extra-Poisson variability in a set of ostensibly homogeneous observations is most easily assessed by a variance-to-mean comparison. The test employs Fisher's dispersion statistic [14, 15]:

$$X_i^2 = \sum_{i=1}^{T} \frac{(Y_{ij} - \overline{Y}_i)^2}{\overline{Y}_i}, \qquad (1)$$

where $\overline{Y}_i$ is the ith dose sample mean ($i = 1, \ldots, T$) (*see* **Poisson Distribution**). Asymptotically, X_i^2 is distributed as $\chi^2(R_i - 1)$, (i.e. with $R_i - 1$ degrees of freedom), hence $X_i^2 > \chi^2_{1-\alpha}(R_i - 1)$ implies significant extra-Poisson variability. This test is $C(\alpha)$-optimal against the **negative binomial distribution** [35], a standard alternative to the Poisson model. To extend (1) to multiple test groups, simply aggregate the individual X_i^2 statistics into a sum of per-dose contributions [28]: $X^2 = \sum_{i=1}^{T} X_i^2$. Asymptotically, X^2 is distributed as $\chi^2\left(\sum_{i=1}^{T} R_i - T\right)$. (A slightly more powerful test of extra-Poisson variability may be constructed using the similar statistic $C^2 = \sum_{i=1}^{T} \sum_{j=1}^{R_i} (Y_{ij} - \overline{Y}_i)^2 / \overline{Y}$, where $\overline{Y} = \sum_{i=1}^{T} R_i Y_i / \sum_{i=1}^{T} R_i$. The null reference distribution of C^2 is complex, however, and hence somewhat difficult to apply in practice [11].) Dean [13] discusses additional extensions to more complex regression settings; see also [20].

A common application of these tests is to large, controlled databases where similar controlled trials and replicate data are available. Margolin et al. [30] reported on such data, demonstrating that the Poisson model generally is inadequate to describe statistical variability when analyzing *Salmonella* mutagenicity experiments. Further research has suggested that *Salmonella* data often exhibit extra-Poisson variability, and that the parent distribution of the data is better described by a negative binomial distribution [11, 28, 30]. Other publications indicate, however, that the Poisson model can be acceptable in select instances [8], particularly for test data obtained by imposing strict protocol adherence and by harvesting all the data on the same day [16, 26]. Given these conflicting considerations, testing for extra-Poisson variability via X^2 is necessary in practice.

Trend Tests for Dose Response

For mutagenicity experiments where no extra-Poisson variability is evidenced, the well-known Cochran–Armitage trend test for dose response may be employed. Under Poisson sampling, the test is known to be optimal against any monotone dose–response function [32, 48]. Given the observed counts, Y_{ij} ($j = 1, \ldots, R_i; i = 1, \ldots, T$), and denoting x_i as some increasing score (dose, log dose, etc.) associated with treatment level i, the test statistic is

$$Z_{\mathrm{CA}} = \frac{\sum_{i=1}^{T} x_i \left[\left(\sum_{j=1}^{R_i} Y_{ij} \right) - R_i \overline{Y} \right]}{(\overline{Y} S_x^2)^{1/2}}, \qquad (2)$$

in which $S_x^2 = \sum_{i=1}^{T} R_i (x_i - \overline{x})^2$ and $\overline{x} = \sum_{i=1}^{T} R_i x_i / \sum_{i=1}^{T} R_i$. Asymptotically, Z_{CA} is distributed as standard normal, and significant dose response is suggested when Z_{CA} is larger than an appropriate upper-α quantile, $z_{1-\alpha}$.

If the response is overdispersed and the Poisson model is invalid, a form of Cochran–Armitage trend test may still be constructed. Under a negative binomial parent model, one simply replaces $(\overline{Y} S_x^2)^{1/2}$ in (2) with an updated estimator based on estimating the negative binomial variance [28]. The result is

$$Z_{\mathrm{NB}} = \frac{\sum_{i=1}^{T} x_i \left[\left(\sum_{j=1}^{R_i} Y_{ij} \right) - R_i \overline{Y} \right]}{[\overline{Y}(1 + \hat{\delta}\overline{Y}) S_x^2]^{1/2}}. \qquad (3)$$

The dispersion estimator, $\hat{\delta}$, is calculated by the **method of moments**, **maximum likelihood** (ML), or maximum quasi-likelihood [10, 36, 49]. For ML estimation, differentiating the log likelihood function

with respect to δ achieves a nonlinear equation,

$$\psi\left(\frac{1}{\delta}\right)\sum_{i=1}^{T} R_i + \sum_{i=1}^{T}\sum_{j=1}^{R_i}\left[\ln(1+\delta\overline{Y}_i) - \psi\left(Y_{ij}+\frac{1}{\delta}\right)\right] = 0, \qquad (4)$$

the solution of which provides $\hat{\delta}_{\mathrm{ML}}$. In (4), $\psi(\cdot)$ is the digamma function, which is available from tables [1] or computer algorithms [5]. Once calculated, $\hat{\delta}_{\mathrm{ML}}$ is substituted for $\hat{\delta}$ in (3). As above, significant dose response is suggested when Z_{NB} is larger than a standard normal quantile, $z_{1-\alpha}$.

When the parent distribution of the data is not known or indeterminate, it is possible to construct a generalized score statistic (*see* **Likelihood**) for testing trend producing a statistic recommended by Astuti and Yanagawa [4]

$$\mathrm{Z_{GCA}} = \frac{\sum_{i=1}^{T} R_i(x_i-\overline{x})\overline{Y}_i}{\sqrt{\sum_{i=1}^{T}(x_i-\overline{x})^2\sum_{j=1}^{R_i}(Y_{ij}-\overline{Y})^2}}, \qquad (5)$$

where $\overline{x}$, $\overline{Y}_i$, and $\overline{Y}$, are defined as above. In large samples, $\mathrm{Z_{GCA}}$, is distributed approximately as standard normal; thus we reject $\mathrm{H_0}$ in favor of an increasing (decreasing) trend in the mean response when $\mathrm{Z_{GCA}}$ is greater (less) than or equal to $z_{1-\alpha}(-z_{1-\alpha}$. (Two-sided testing is also possible.) This is similar to the basic statistics in (2) or (3); the major difference is that the terms estimating the Poisson or negative binomial variances are replaced by the more robust empirical variance estimate $\sum_{j=1}^{R_i}(Y_{ij}-\overline{Y})^2$.

Trend Test Under Nonmonotone Dose Response

An important, additional feature illustrated by the data in Table 1 is a curious downturn in the response at higher doses. This form of nonmonotone dose response is common in the *Salmonella* assay, and is observed with other mutagenesis assays, including data for chromosomal damage in yeast (*Saccharomyces cerevisiae*) or mold (*Aspergillus nidulans*) [37, 42], *in vitro* cell transformation data, or *in vitro* chromosome aberration data with human lymphocytes [43].

In *Salmonella*, the downturn phenomenon is driven by a number of possible mechanisms, the most common of which is a consequence of the experimental scheme employed to identify the mutational events. Since mutagenesis is identified by growth on a selective medium, any other mechanism that hinders growth will compete with the desired outcome. Thus, for example, high exposures of the toxic agent can lead to cell death or perhaps chemical/threshold induced increases in DNA repair. These factors conspire to *reduce* the yield of mutated cells by killing or neutralizing the microbes before mutations can be observed. The result is a downturn at the higher doses, i.e. an *umbrella response* [25]. As might be expected, simple tests for monotone trend such as Z_{CA}, Z_{NB}, or Z_{GCA} can exhibit large losses in **power** to detect any increase in dose response in the presence of a downturn [12], and suitable alternatives must be employed.

Nonlinear biomathematical and empirical models for this downturn phenomenon are available, which, when coupled with the negative binomial sampling model on the Y_{ij}s, involve fairly advanced statistical methodology [9, 23, 30]. A simple alternative is to identify statistically the point of maximal departure from background mutation levels, and then test for increasing trend up to that point. A number of different procedures have been suggested for such an approach [6, 46], the majority of which analyze the **ranks** of the observations by nonparametric techniques [37].

For example, a nonparametric approach developed by Simpson & Margolin [46] estimates recursively the point of maximal response (the *umbrella point*), and then tests for a significant increase in dose response up to that point. One begins by calculating a series of two-sample **Mann–Whitney** statistics [27], W_i, comparing the ith dose group with – collectively – all preceding dose groups, starting at $i = T$ and working backwards. That is, W_T tests whether the response information at $i = T$ departs significantly above the pooled response information at $i = 1, \ldots, T-1$. If so, an estimate, $\mathbf{h}$, of the umbrella index is $\mathbf{h} = T$. If not, one discards the information at $i = T$, and repeats the process at $i = T-1$. In this way, the test employs the values of $W_i (i = T, T-1, \ldots, 2)$ to estimate the umbrella index via recursive pretesting. In effect, it selects

this estimate as the largest dose index, $\mathbf{h}$, such that $W_{\mathbf{h}}$ is larger than a specified critical value $c_{\mathbf{h}}$; i.e. $\mathbf{h} = \max\{i \in \{2, \ldots, T\} : W_i > c_i\}$, where $c_2 = 0$ and $c_i = c_i(\mathbf{q})(i = 3, \ldots, T)$ is the $\mathbf{q}$-quantile of the distribution of W_i under the null hypothesis of no dose response. The tuning parameter $\mathbf{q} \in (0, 1)$ must be prespecified, and is usually set at, or near, $\mathbf{q} = 0.5$ [37, 47].

Once the umbrella index, $\mathbf{h}$, is estimated, the recursive procedure calculates the conditional Jonckheere–Terpstra trend statistic $U_{\mathbf{h}} = W_1 + \cdots + W_{\mathbf{h}}$ [19], and an increasing dose response is suggested when $U_{\mathbf{h}}$ is larger than the $(1 - \pi)$-quantile of the distribution of U_i under the null hypothesis of no differences among the doses $(i = 2, \ldots, T)$, for $0 < \pi < 1$. For fixed $\mathbf{q}$ and prespecified significance level α, a conservative choice for π is given as $\pi = \alpha(1 - \mathbf{q})/(1 - \mathbf{q}^{T-1})$ [46]. A computing algorithm for these calculations is described in [45], and extensions and other research issues involved with rank-based testing for *Salmonella* data are discussed in [44].

Example: 1,3-Butadiene

Consider the example of the airborne toxin 1,3-butadiene, the data for which appear in Table 1. These values represent a *Salmonella* mutagenic response after gaseous exposure to the chemical, where an increase in the number of observed colonies suggests a mutagenic effect.

Begin by testing whether the data exhibit any overdispersion relative to the Poisson sampling model. Employing the aggregated dispersion statistic using (1) to the data in Table 1 gives $X^2 = 22.134$ on 12 df ($P = 0.004$). Strong departure from Poisson variability is evidenced. If the negative binomial sampling model were considered as an adequate replacement to the Poisson for these data, then a trend test for increasing dose response based on (3) gives $Z_{\mathrm{NB}} = 1.285$, with one-sided $P = 0.099$. Marginal suggestion of an increasing trend is given.

For these data, however, a downturn in the dose response is evident. This could affect the trend test's ability to identify properly a positive dose response. Application of the recursive test for an umbrella response is therefore appropriate: setting the tuning parameter to $\mathbf{q} = 0.5$, the umbrella index is estimated as $\mathbf{h} = 6$, with a corresponding rank-based trend statistic given by $U_6 = 105.5$. Rejection occurs at significance levels as low as $\alpha = 0.005$, providing strong evidence of a positive dose response for these mutagenicity data.

References

[1] Abramowitz, M. & Stegun, I.A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th Ed. Wiley-Interscience, New York.

[2] Ager, D.D. & Haynes, R.H. (1987). Mathematical description of the interactions between cellular inactivating agents, *Radiation Research* **110**, 129–141.

[3] Alvord, W.G., Driver, J.H., Claxton, J. & Creason, J.P. (1990). Methods for comparing Salmonella mutagenicity data sets using nonlinear models, *Mutation Research* **240**, 177–194.

[4] Astuti, E.T. & Yanagawa, T. (2002). Trend test for count data with extra-Poisson variability, *Biometrics* **58**, 398–402.

[5] Bernardo, J.M. (1976). Algorithm AS103. The digamma function, *Applied Statistics* **25**, 315–317.

[6] Bernstein, L., Kaldor, J., McCann, J. & Pike, M.C. (1982). An empirical approach to the statistical analysis of mutagenesis data from the Salmonella test, *Mutation Research* **97**, 267–281.

[7] Bishop, J.B. & Kodell, R.L. (1980). The heritable translocation assay: Its relationship to assessment of genetic risk for future generations, *Teratogenesis, Mutagenesis, and Carcinogenesis* **1**, 305–322.

[8] Bogen, K.T. (1994). Applicability of alternative models of revertant variance to Ames-test data for 121 mutagenic carcinogens, *Mutation Research* **322**, 265–273.

[9] Breslow, N. (1984). Extra-Poisson variability in log-linear models, *Applied Statistics* **33**, 38–44.

[10] Clark, S.J. & Perry, J.N. (1989). Estimation of the negative binomial parameter k by maximum quasi-likelihood, *Biometrics* **45**, 309–316.

[11] Collings, B.J. & Margolin, B.H. (1985). Testing goodness of fit for the Poisson assumption when observations are not identically distributed, *Journal of the American Statistical Association* **80**, 411–418.

[12] Collings, B.J., Margolin, B.H. & Oehlert, G.W. (1981). Analyses for binomial data, with applications to the fluctuations test for mutagenicity, *Biometrics* **37**, 775–794.

[13] Dean, C.B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* **87**, 451–457.

[14] Fisher, R.A. (1950). The significance of deviations from expectation in a Poisson series, *Biometrics* **6**, 17–24.

[15] Fisher, R.A., Thornton, H.G. & MacKenzie, W.A. (1922). The accuracy of the plating method of estimating the density of bacterial populations, *Journal of Applied Biology* **9**, 325–359.

[16] Hamada, C., Wada, T. & Sakamoto, Y. (1994). Statistical characterization of negative control data in the Ames

Salmonella/microsome test, *Environmental Health Perspectives* **102**, Supplement 1, 115–119.

[17] Haynes, R.H. (1989). Mutagenesis and mathematics: the allure of numbers, *Environmental and Molecular Mutagenesis* **14**, 200–205.

[18] Hollstein, M., McCann, J., Angelosanto, F. & Nichols, W. (1979). Short-term tests for carcinogens and mutagens, *Mutation Research* **65**, 133–226.

[19] Jonckheere, A.R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika* **41**, 133–145.

[20] Kim, B.S. & Park, C. (1992). Some remarks on testing goodness of fit for the Poisson assumption, *Communications in Statistics - Theory and Methods* **21**, 979–995.

[21] Kirkland, D.J. (1989). *Statistical Evaluation of Mutagenicity Test Data*. Cambridge University Press, Cambridge.

[22] Krewski, D., Leroux, B.G., Bleuer, S.R. & Broekhoven, L.H. (1993). Modeling the Ames Salmonella/micro some assay, *Biometrics* **49**, 499–510.

[23] Leroux, B.G. & Krewski, D. (1993). AMESFIT: A microcomputer program for fitting linear-exponential dose-response models in the Ames Salmonella assay, *Environmental and Molecular Mutagenesis* **22**, 78–84.

[24] Lockhart, A.-M., Piegorsch, W.W. & Bishop, J.B. (1992). Assessing overdispersion and dose response in the male dominant lethal assay, *Mutation Research* **272**, 35–58.

[25] Mack, G.A. & Wolfe, D.A. (1981). K-sample rank tests for umbrella alternatives, *Journal of the American Statistical Association* **76**, 175–181.

[26] Mahon, G.A.T., Middleton, B., Robinson, W.D., Green, M.H.L., Mitchell, I. & Tweats, D.J. (1989). Analysis of data from microbial count assays, in *Statistical Evaluation of Mutagenicity Test Data*, D.J. Kirkland, ed. Cambridge University Press, Cambridge, pp. 26–65.

[27] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[28] Margolin, B.H. (1985). Statistical studies in genetic toxicology: a perspective from the U.S. National Toxicology Program, *Environmental Health Perspectives* **63**, 187–194.

[29] Margolin, B.H., Collings, B.J. & Mason, J.J. (1983). Statistical analysis and sample-size determinations for mutagenicity experiments with binomial response, *Environmental Mutagenesis* **5**, 705–716.

[30] Margolin, B.H., Kaplan, N. & Zeiger, E. (1981). Statistical analysis of the Ames Salmonella/microsome test, *Proceedings of the National Academy of Sciences* **76**, 3779–3783.

[31] Margolin, B.H., Kim, B.S. & Risko, K.J. (1989). The Ames Salmonella/microsome mutagenicity assay: issues of inference and validation, *Journal of the American Statistical Association* **84**, 651–661.

[32] Margolin, B.H., Resnick, M.A., Rimpo, J.Y., Archer, P., Galloway, S.M., Bloom, A.D. & Zeiger, E. (1986). Statistical analyses for in vitro cytogenetic assays using Chinese hamster ovary cells, *Environmental Mutagenesis* **8**, 183–204.

[33] Maron, D.M. & Ames, B.N. (1983). Revised methods for the salmonella mutagenicity test, *Mutation Research* **113**, 173–215.

[34] Murphy, S.A., Tice, R.R., Smith, M.G. & Margolin, B.H. (1992). Contributions to the design and analysis of in vivo SCE experiments, *Mutation Research* **271**, 39–48.

[35] Neyman, J. & Scott, E.L. (1966). On the use of $C(\alpha)$ optimal tests of composite hypotheses, *Bulletin de l'Institut International de Statistique (Calcutta)* **41**, 477–497.

[36] Piegorsch, W.W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter, *Biometrics* **46**, 863–867.

[37] Piegorsch, W.W. (1992). Nonparametric methods to assess non-monotone dose response: applications to genetic toxicology, in *Order Statistics and Nonparametrics: Theory and Applications*, P.K. Sen & I.A. Salama, eds. North-Holland, Amsterdam, pp. 419–430.

[38] Piegorsch, W.W. (1993). Biometrical methods for testing dose effects of environmental stimuli in laboratory studies, *Environmetrics* **4**, 483–505.

[39] Piegorsch, W.W. (1994). Environmental biometry: Assessing impacts of environmental stimuli via animal and microbial laboratory studies, in *Handbook of Statistics*, Vol. 12: Environmental Statistics, G.P. Patil & C.R. Rao, eds. North-Holland/Elsevier, Amsterdam, pp. 535–559.

[40] Piegorsch, W.W., Lockhart, A.-M.C., Margolin, B.H., Tindall, K.R., Gorelick, N.J., Short, J.M., Carr, G.J., Thompson, E.D. & Shelby, M.D. (1994). Sources of variability in data from a lacI transgenic mouse mutation assay, *Environmental and Molecular Mutagenesis* **23**, 17–31.

[41] Piegorsch, W.W., Margolin, B.H., Shelby, M.D., Johnson, A., French, J.E., Tennant, R.W. & Tindall, K.R. (1995). Study design and sample sizes for a lacI transgenic mouse mutation assay, *Environmental and Molecular Mutagenesis* **25**, 231–245.

[42] Piegorsch, W.W., Zimmermann, F.K., Fogel, S., Whittaker, S.G. & Resnick, M.A. (1989). Quantitative approaches for assessing chromosome loss in Saccharomyces cerevisiae: general methods for analyzing downturns in dose response, *Mutation Research* **224**, 11–29.

[43] Richardson, C., Williams, D.A., Amphlett, G., Phillips, B., Allen, J.A. & Chanter, D.O. (1989). Analysis of data from in vitro cytogenetic assays, in *Statistical Evaluation of Mutagenicity Test Data*, D.J. Kirkland, ed. Cambridge University Press, Cambridge, pp. 141–154.

[44] Schumacher, M. & Schmoor, C. (1991). Statistical analysis of the Ames assay, in *Statistics in Toxicology*, L. Hothorn, ed. Springer-Verlag, Heidelberg, pp. 5–19.

[45] Simpson, D.G. & Dallal, G.E. (1989). BUMP: a FORTRAN program for identifying dose-response curves

subject to downturns, *Computers and Biomedical Research* **22**, 36–43.

[46] Simpson, D.G. & Margolin, B.H. (1986). Recursive nonparametric testing for dose-response relationships subject to downturns at high doses, *Biometrika* **73**, 589–596.

[47] Simpson, D.G. & Margolin, B.H. (1990). Nonparametric testing for dose-response curves subject to downturns: asymptotic power considerations, *Annals of Statistics* **18**, 373–390.

[48] Tarone, R.E. (1982). The use of historical control information in testing for a trend in Poisson means, *Biometrics* **38**, 457–462.

[49] van de Ven, R. (1993). Estimating the shape parameter for the negative binomial distribution, *Journal of Statistical Computation and Simulation* **46**, 111–123.

[50] Waters, M.D., Stack, H.F., Brady, A.L., Lohman, P.H.M., Haroun, L. & Vainio, H. (1987). Activity profiles for genetic and related tests, Appendix I, in *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Genetic and Related Effects: An Updating of Selected IARC Monographs from Volumes 1 to 42*, IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, ed. International Agency for Research on Cancer, Lyon, France, pp. 687–696.

(*See also* **Biological Assay, Overview**)

WALTER W. PIEGORSCH

Mutation

Most of the time when chromosomes duplicate, the deoxyribonucleic acid (**DNA**) is faithfully copied. Rarely, however, mutation occurs, i.e. there is a change due to an error in the copying process, leading to a mutant allele (*see* **Gene**). There are different kinds of mutations, depending on which base-pair in a DNA sequence is miscopied, such as missense mutations that change an amino acid in a protein and nonsense mutations that result in the formation of an incomplete protein. Whereas nonsense mutations usually lead to lack of biologic activity, and hence have a large effect, missense mutations can vary in effect from none to serious disease. The word mutation is now often used to mean a mutant allele, especially if the allele is rare and has a deleterious effect, in contrast to a **polymorphism**. But all polymorphisms can be assumed to have arisen by mutation, making evolution possible.

ROBERT C. ELSTON

National Center for Health Statistics (NCHS)

The National Center for Health Statistics (NCHS) is the principal health statistics agency of the US, with responsibility for designing and maintaining a variety of general-purpose descriptive health surveys on a continuous basis and disseminating these data for widespread use (*see* **Surveys, Health and Morbidity**). The NCHS came into being in 1960 through the combining of the National Health Survey and the National Office of Vital Statistics. The principal health surveys conducted by the NCHS include **population-based** surveys (the National Health Interview Survey and the National Health and Nutrition Examination Survey); record-based surveys (the National Health Care Surveys); facility-based surveys (the Master Facility Inventory and the National Nursing Home Survey); and the Cooperative National Vital Statistics System covering births, deaths, fetal deaths, marriages, and divorces (*see* **Vital Statistics, Overview**).

History

The National Vital Statistics System of the US achieved essentially complete coverage for births and deaths in 1933. The collection and archiving of vital records are the responsibilities of each of the states and territories. The federal system is a cooperative program between each of the registration areas and the NCHS (and its predecessor agencies). In coordination with the Association for Vital Records and Health Statistics, which represents the registrars of each of the states, NCHS periodically revises the model certificates for registration of vital events, identifies the data items to be reported to NCHS for national tabulation, and specifies quality criteria for the accuracy, completeness, and timeliness of reporting vital events.

Drawing on the experience of such surveys as the nationwide health survey of 1935–1936 done under the auspices of the Work Projects Administration (WPA) and local health surveys conducted in Hagerstown and Baltimore, Maryland, which showed that interview-based health surveys were feasible, and encouraged by recent advances in population sampling and statistical methods developed by the US Bureau of the Census, the newly created US National Committee of Vital and Health Statistics issued a report calling for the development of a National Morbidity Survey. Largely in response to this report, Congress passed the National Health Survey Act in 1956. The first component of the National Health Survey to be implemented was the National Health Interview Survey which went into the field in July 1957, and has been conducted continuously since then. The second component, the National Health Examination Survey (subsequently to become the National Health and Nutrition Examination Survey when a major nutrition component was added in 1971) was initiated in 1960 and has been conducted periodically seven times with plans for initiating it as a continuously ongoing survey in 1998. The third component was the Health Records Survey which began in 1963 with the creation of a National Master Facility Inventory (NMFI) designed as a comprehensive file of inpatient facilities in the US. Drawing a sample of hospitals from the NMFI, the National Hospital Discharge Survey began in 1965 and has been conducted continuously since. Additional subcomponents of the renamed National Health Care Survey were added over the years including the National Ambulatory Medical Care Survey, the National Survey of Ambulatory Surgery, the National Hospital Ambulatory Medical Care Survey, and the National Home and Hospice Care Survey. The National Nursing Home Survey has been conducted intermittently since 1973 and is based more on characteristics of facilities than on individual patient records.

The NCHS has been based in a number of agencies since its inception, but since 1987 it has been a component of the **Centers for Disease Control (CDC)** of the Department of Health and Human Services with most of its staff of about 500 persons located in Hyattsville, Maryland.

Survey Programs of the National Center for Health Statistics

The National Health Interview Survey

The National Health Interview Survey (NHIS) is a continuous **cross-sectional** survey of the noninstitutionalized, civilian population of the US designed to produce national data on the incidence of acute illnesses and injuries, prevalence of chronic

conditions and impairments, extent of disabilities, utilization of health care services (*see* **Health Care Utilization Data**), and, on a periodic basis, information on other health-related topics such as health insurance coverage, knowledge and attitudes about HIV/**AIDS**, use of medical devices, immunization status of children, and indicators of progress toward achieving the objectives set in *Healthy People 2000: National Health Promotion and Disease Prevention Objectives*. The sampling plan follows a **multistage sampling** design such that the sample scheduled each week is representative of the **target population** and the weekly samples can be cumulated over time. For the years 1985–1994 a typical NHIS sample consisted of 198 primary sampling units (counties or metropolitan statistical areas) with approximately 7500 segments (defined to contain an expected eight households each) yielding about 59 000 assigned households, of which about 15% were vacant or out of scope. The expected final sample of about 49 000 households comprised a **probability sample** of about 125 000 persons. The annual response rates for the core survey have been between 94% and 98% over the years, with somewhat lower response rates for the special topic areas.

The data are collected through personal household interviews conducted by interviewers employed and trained by the Bureau of the Census according to procedures specified by the NCHS. The questionnaires are developed in conjunction with statisticians and experts in cognitive psychology. Extensive pretesting and field trials are conducted before the survey goes into each annual cycle. All members of the household aged 17 and older are invited to participate and to respond for themselves. For children and for persons not at home, information is provided by a proxy (an adult resident of the household). For some topics, a random subsample of household members is selected and special techniques have been developed for eliciting confidential information on sensitive topics in the household setting. The NHIS often serves as a **sampling frame** for follow-on studies and special callbacks will be made to household members selected for such studies (*see* **Call-backs and Mail-backs in Sample Surveys**). These targeted population studies have included the Longitudinal Study on Aging, the Teenage Attitudes and Practices Survey, the Access to Care Followup Study, and the Disability Followup Survey.

The National Health and Nutrition Examination Survey

Through the National Health and Nutrition Examination Surveys, the NCHS provides estimates of the **prevalence** of selected diseases and conditions, normative distributions for a variety of physiologic, anthropomorphic, and nutritional measures, and assessments of exposures to environmental hazards such as lead and pesticides. Seven health examination surveys were conducted by the NCHS between 1960 and 1995. With the addition of the specific mandate to monitor the nutritional status of the US in 1970, the three Health Examination Surveys conducted in the 1960s were followed by the first National Health and Nutrition Examination Survey (NHANES I) in 1970–1973, NHANES II in 1976–1980, and NHANES III in 1988–1994. The surveys were designed to obtain nationally representative information on the health and nutritional status of the population using interviews, physical examinations, and standardized clinical and laboratory tests. This information is of two kinds: prevalence data of selected diseases and health conditions; and normative population data on the distribution of such measurements as height, weight, blood pressure, visual and auditory acuity, and a variety of blood chemistries such as cholesterol levels, vitamin levels, and metabolites of pesticides and other environmental exposures. The several surveys have targeted different age segments of the population, with the latest survey covering the entire noninstitutionalized US population aged two months or older. A special Hispanic Health and Nutrition Examination Survey (HHANES) was conducted in 1983–1984 covering three areas with high concentrations of Hispanic Americans: Mexican Americans in the Southwest; Cubans in Miami (Dade County), Florida; and Puerto Ricans in the New York City area. Oversampling has been used in other surveys to obtain more precise estimates for selected subgroups of the population. In NHANES III, black and Mexican Americans, as well as children and older persons were oversampled.

The sample design of NHANES III employed a **stratified** multistage probability sample of counties, blocks, and persons randomly selected from households. The periods 1988–1991 and 1991–1994 each constituted national samples of the US population. Eighty-one counties were selected from 26 states from which approximately 40 000 persons of all races

were selected and about 30 000 agreed to participate in the medical examination. The examinations are conducted in specially designed mobile examination centers that provide a standardized environment for obtaining the measurements and biologic samples. Some of the 30 topics investigated in the NHANES III were: high blood pressure, high blood cholesterol, obesity, passive smoking, lung disease, osteoporosis, HIV, hepatitis, *Helicobacter pylori*, immunization status, diabetes, allergies, growth and development, blood lead, anemia, food sufficiency, dietary intake including fats, antioxidants, and nutritional blood measures.

NHANES I Epidemiologic Followup Study

The NHANES I Epidemiologic Followup Study (NHEFS) is a national longitudinal study designed to investigate the relationships between clinical, nutritional, and behavioral factors assessed at baseline NHANES I, and subsequent morbidity, mortality, and institutionalization. The NHEFS population includes the 14 407 participants who were 25–74 years of age when first examined in NHANES I (1971–1975). NHEFS provides data on mortality, morbidity, and hospital utilization as well as changes in risk factors, functional limitation, and institutionalization between NHANES I and the follow-up recontacts. The first wave (1982–1984) of data collection was conducted for all members of the NHEFS cohort. Continued follow-ups of the NHEFS population were conducted in 1986, 1987, and 1992.

National Survey of Family Growth

The National Survey of Family Growth monitors childbearing practices and reproductive health through periodic household interview surveys of a national sample of women aged 15–44. The survey provides data on contraception, infertility, use of family planning and infertility services, sexual activity, family formation, family size and related aspects of maternal and child health. Conducted periodically since 1973, the survey was conducted most recently in 1995.

The National Health Care Survey

The National Health Care Survey (NHCS), originally consisting of four discrete record-based surveys, is now an integrated survey of a wide variety of health care providers. The NHCS was built upon the following four continuing surveys: the National Hospital Discharge Survey, the National Ambulatory Medical Care Survey, the National Nursing Home Survey, and the National Health Provider Inventory (formerly the National Master Facility Inventory). The new surveys include the National Survey of Ambulatory Surgery, the National Hospital Ambulatory Medical Care Survey, and the National Home and Hospice Care Survey.

National Hospital Discharge Survey

The National Hospital Discharge Survey (NHDS) is the principal source of information on inpatient utilization of nonfederal short-stay hospitals. It includes data on diagnoses, procedures, length of stay, expected source of payment, and patterns of use of care in hospitals, and on the size, location, and ownership of hospitals, but does not include individual identifiers for patients. Conducted annually since 1965, the NHDS currently is based on data abstracted from a sample of approximately 274 000 patient records from a sample of 525 hospitals from a universe of about 8000 short stay hospitals. Only hospitals with six or more beds and an average length of stay for all patients of less than 30 days are included in the sample.

National Ambulatory Medical Care Survey

The National Ambulatory Medical Care Survey (NAMCS) provides statistics on the characteristics of patients and services provided by office-based physicians. The sample consists of 40 000 visits from approximately 3000 physicians drawn from a sampling frame of licensed physicians in office-based, patient care practices compiled from files maintained by the American Medical Association and the American Osteopathic Association. Data collection is carried out by the participating physicians using a form that takes only one or two minutes to complete. Up to 10 patient visits per physician are reported each day during about a one-week period. The sampling rate and duration depend on the number of patients the physician expects to see. The data collected for each patient visit include information on the patient's symptoms, diagnostic procedures, physician's diagnoses, and medications ordered or provided, as well

as patient management and planned future treatment. The NAMCS has been conducted annually from 1974 to 1981, in 1985, and annually since 1989.

National Survey of Ambulatory Surgery

The National Survey of Ambulatory Surgery (NSAS) provides information on the use of free-standing and hospital-based ambulatory surgery centers in the US. Although most surgery is still performed on an inpatient basis, advances in medical technology have enabled a wide variety of surgical and diagnostic treatments now to be performed in an ambulatory setting. The NSAS, which began in April 1994, provides detailed data on this expanding area of health care. Information collected includes patient characteristics, diagnoses, surgical and diagnostic procedures, and administrative information such as patient disposition and expected sources of payment.

National Hospital Ambulatory Medical Care Survey

The National Hospital Ambulatory Medical Care Survey produces statistics that are representative of the experience of the US population receiving health care in hospital emergency departments and outpatient departments. This hospital-based survey, which is based on abstracts from medical records, provides information similar to that collected in the office-based ambulatory survey: demographic characteristics of patients, patients' complaints, physicians' diagnoses, diagnostic/screening services, procedures, medications, disposition, types of health care professional seen, expected sources of payment, and causes of injury where applicable. Data collection began in 1992 with an annual sample of 70 000 visits to 440 hospitals.

National Nursing Home Survey

The National Nursing Home Survey is based on self-administered questionnaires and interviews with administrators and staff in a sample of about 1500 long-term care facilities. Information is obtained on both the providers of services and on the nursing home patients. Data about the facilities include characteristics such as size, ownership, Medicare/Medicaid certification, occupancy, days of care provided, and expenses. For patients, data are obtained on demographic characteristics, health status, and services received. The survey has been conducted periodically since 1963, most recently in 1995.

National Nursing Home Survey Followup

The National Nursing Home Survey Followup is a longitudinal study that follows the cohort of surviving current residents and discharged residents sampled from the 1985 National Nursing Home Survey. Its primary purpose is to provide data on the flow of persons in and out of long-term care facilities and hospitals. The National Nursing Home Survey Followup provides data on the subjects' vital status, living arrangements, nursing home stays, hospital stays, and sources of payment for stays. The study population consisted of approximately 6600 subjects with follow-up conducted at two-year intervals using **computer-assisted telephone interviews**.

National Home and Hospice Care Survey

The National Home and Hospice Care Survey provides data on home health agencies and hospices and their current patients and discharges. This survey was instituted in 1992, in response to the rapid growth in the number of these agencies throughout the US. The annual survey is based on personal interviews with administrators and staff of approximately 1500 sample agencies. Information is obtained on diagnoses, types and length of services provided, number of visits, patient charges, health status, and reason for discharge.

National Health Provider Inventory

The National Health Provider Inventory (NHPI) is a comprehensive national listing of nursing homes, residential care facilities, hospices, and home health agencies. The NHPI serves as a sampling frame for several sample surveys. In addition, it is an important source of national statistics on the number, type, and geographical distribution of health providers in the US. Conducted periodically since 1963 under different survey titles – National Master Facility Inventory, 1971–1976; Inventory of Long-Term Care Places, 1986 – it was conducted in its

present form in 1991. The NHPI provides names and addresses of almost 56 000 facilities, including more than 7800 home health agencies and hospices, more than 15 500 nursing homes, and more than 31 000 board and care homes. Information about such items as type of facility, ownership, size, location, and resident characteristics is collected from questionnaires sent directly to agencies and facilities.

National Vital Statistics System

The National Vital Statistics System is responsible for US official vital statistics. The registration of vital events – births, deaths, marriages, divorces, fetal deaths, and induced terminations of pregnancy – is the responsibility of each of the states and is carried out under the civil registration laws of each state. However, standard forms for the collection of the data, model procedures for the uniform registration of the events, and standards for quality and timeliness of the statistical reports of events are developed and recommended for state use through cooperative activities of the states and the NCHS. The NCHS shares the costs incurred by the states in providing vital statistics data for national use. It produces annual data for the US and for states, counties, and other local areas, and monthly provisional data for the US and each state. Typically, the record of each event is filed by the responsible party (physician, hospital, or funeral director) with the local registrar of the town, city, or county where the event occurs. The local registrar inspects the report for completeness and accuracy, retains a local copy, and sends the original to the state health department. The state vital records office maintains permanent archives of the records, processes the statistical information, and provides summaries for state and local use. A uniform data set from the individual records is transmitted electronically to the NCHS for compilation into national statistical files. Two principal sets of reports are prepared from these data: monthly provisional estimates assembled on a "current flow" basis and final annual national vital statistics volumes. Before 1997, the provisional mortality data had been based on a 10% sample of deaths, but since then all records received have been incorporated into the provisional reports. The contents and quality standards for the uniform data set are reviewed approximately every 10 years. Rigorous security controls have been instituted to ensure the confidentiality of the records and prevent inadvertent identification of individuals. The annual volumes provide detailed tabulations of a wide variety of characteristics of the approximately 4 000 000 births, 2 100 000 deaths, and 70 000 fetal deaths that occur in the US and of the 2 300 000 marriages and 1 200 000 divorces.

The NCHS linked files of live birth and infant death records are research files for exploring the complex relationships between infant death and risk factors present at birth. The linked files include information from the birth certificate such as birth weight, mother's age, and prenatal care, linked to information from the death certificate for the same infant, such as cause of death and age at death (*see* **Record Linkage**). The files are birth cohort linked files. They are based on deaths under one year of age to all infants born in a calendar year. Each file contains approximately 40 000 linked records. The first annual national linked file was for the 1983 cohort under a pilot project. Beginning with the birth cohort of 1987, linked files are part of the National Vital Statistics System.

National Death Index

Working with state offices, NCHS established the National Death Index (NDI), a central computerized index of death record information, in 1979 as a resource to aid epidemiologists and other health and medical investigators in determining whether persons in their studies have died and, if so, to provide the names of the states in which those deaths occurred, the dates of death, and the corresponding death certificate numbers. The NDI is available to investigators solely for statistical purposes in biomedical research and is not accessible for legal, administrative, or genealogic purposes.

National Maternal and Infant Health Survey

The National Maternal and Infant Health Survey (NMIHS) collects data to study factors related to poor pregnancy outcomes, including low birth weight, stillbirth, infant illness, and infant death. The NMIHS is a followback survey of informants named on vital records. The 1988 survey was based

on 10 000 live births, 4000 fetal deaths, and 6000 infant deaths and was the first national survey that included data on those three pregnancy outcomes simultaneously. National Natality Surveys had been conducted in 1963, 1964–1966, 1968–1969, 1972, and 1980. A National Fetal Mortality Survey was done in 1980, and a National Infant Mortality Survey was conducted in 1964–1966. A 1991 longitudinal follow-up to the NMIHS was conducted to obtain additional information about respondents from the 1988 survey. The NMIHS provides data on socioeconomic and demographic characteristics of mothers, prenatal care, pregnancy history, occupational background, health status of mother and infant, and types and sources of medical care received.

National Mortality Followback Survey

Data on characteristics of deceased persons are provided in the National Mortality Followback Survey. The survey is based on questionnaires sent to informants listed on the death certificates to obtain additional data on socioeconomic characteristics of deceased persons, use of and payment for hospitals and institutional care during the last year of life, and factors related to health status. The 1986 survey was a national sample of approximately 1% of US resident deaths of persons 25 years of age and over. The survey was conducted annually from 1961 to 1968 and in 1986. The most recent survey was initiated in 1993 and is the first survey to collect information from medical examiners and coroners for external causes of death.

Data Dissemination

The dissemination of its vital and health statistics, of summary reports, and of research findings is an essential part of the mission of NCHS. Historically, published reports have been the principal modes of dissemination, but with advancing technologies, electronic products are becoming more common.

Principal Publications

Vital Statistics of the United States has been published annually since 1937. It contains the final summaries of mortality and natality data in extensive demographic and geographic detail. Marriage and divorce data have also been available since 1946.

Monthly Vital Statistics Reports are based on monthly and cumulative data on vital events.

The *Vital and Health Statistics* series (the "Rainbow Series") contains detailed reports of the background, methodology, analytical studies, and tabulations from the various NCHS data collection programs. Each program has its own series with a distinctively colored cover.

Advance Data from Vital and Health Statistics are summary reports that provide the first release of data from the various surveys.

Health, United States is a comprehensive annual report from the Secretary of Health and Human Services to the President and Congress describing the nation's health.

Electronic Products

Data are disseminated in great detail through a variety of electronic products. As technology progresses, the specific types of product have varied. Public-use data tapes had been a dominant form for sharing micro-level files with researchers. These data are now released in cartridge format and on CD-ROMs. Diskettes are available with detailed tables from a variety of surveys. Most recently, the NCHS Home Page (`http://www.cdc.gov/nchswww/nchshome.htm`) has been established to provide instantaneous access to a range of statistical information about health status (*see* **Quality of Life and Health Status**) and use of health services in the US.

References

[1] Kovar, M.G. (1989). Data systems of the National Center for Health Statistics, *Vital and Health Statistics* **1**. NCHS, Hyattsville.

[2] Massey, J.T., Moore, T.F., Parsons, V.L. & Tadros, W. (1991). Design and estimation for the National Health Interview Survey, 1985–94, *Vital and Health Statistics* **2**. NCHS, Hyattsville.

[3] National Center for Health Statistics (1964). Health survey procedure: concepts, questionnaire development, and definitions in the Health Interview Survey, *Vital and Health Statistics* **1**. NCHS, Hyattsville.

[4] National Center for Health Statistics (1965). Origin, program, and operation of the U.S. National Health Survey, *Vital and Health Statistics* **1**. NCHS, Hyattsville.

[5] National Center for Health Statistics (1965). Plan and initial program of the Health Examination Survey, *Vital and Health Statistics* **1**. NCHS, Hyattsville.

[6] National Center for Health Statistics (1994). Plan and operation of the third National Health and Nutrition Examination Survey, 1988–94, *Vital and Health Statistics* **1**. NCHS, Hyattsville.

[7] US Government Printing Office (1950). History and organization of the Vital Statistics System, in *Vital Statistics of the United States*, Vol. 1. US GPO, Washington, Chapter 1, pp. 2–19.

MANNING FEINLEIB

National Institutes of Health (NIH)

The National Institutes of Health (NIH), which began as a one-room Laboratory of Hygiene in 1887, today is one of the world's foremost biomedical research centers, and the federal focal point for biomedical research in the US. The goal of NIH research is to acquire new knowledge to help prevent, detect, diagnose, and treat disease and disability, from the rarest genetic disorder to the common cold. NIH works toward achieving its mission by: conducting research in its own laboratories (Intramural Research Programs); supporting the research of nonfederal scientists in universities, medical schools, hospitals, and research institutions throughout the country and abroad (Extramural Research Programs); helping in the training of research investigators; and fostering communication of biomedical information.

The NIH is one of eleven health agencies of the United States Department of Health and Human Services and consists of 27 separate institutes and centers (see Table 1). From a budget of about $300 in 1887, the NIH budget has grown to over $27 billion in 2003, (according to "The National Institutes of Health," a public information brochure published by the NIH.) Biostatistics and biostatisticians have played, and continue to play, an important role in the Intramural and Extramural Research Programs at the NIH.

Biostatistics first appeared as a recognized discipline at the National Institutes of Health in the years 1946–1948. The Division of Statistical Methods in the US Public Health Service was established with **Harold Dorn** as its first Head to support the research of the then new NIH.

The degree of formal statistical training of his first recruits (**Jerry Cornfield**, **Sam Greenhouse**, Jack Lieberman, **Nathan Mantel**, and **Marvin Schneiderman**) varied, but their experience in the applications of statistics to problems of biology and medicine was minimal [12, 13]. Within a few years, **Sid Cutler, Max Halperin**, Bill Haenszel, Harold Kahn, Sam

Table 1 The Institutes, Centers, and Divisions of the US National Institutes of Health

National Cancer Institute
National Eye Institute
National Heart, Lung, and Blood Institute
National Human Genome Research Institute
National Institute on Aging
National Institute on Alcohol Abuse and Alcoholism
National Institute of Allergy and Infectious Diseases
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institute of Biomedical Imaging and Bioengineering
National Institute of Child Health and Human Development
National Institute on Deafness and Other Communication Disorders
National Institute of Dental and Craniofacial Research
National Institute of Diabetes and Digestive and Kidney Diseases
National Institute on Drug Abuse
National Institute of Environmental Health Sciences
National Institute of General Medical Sciences
National Institute of Mental Health
National Institute of Neurological Disorders and Stroke
National Institute of Nursing Research
National Library of Medicine
National Center for Research Resources
National Center for Complementary and Alternative Medicine
National Center for Minority Health and Health Disparities
Center for Information Technology
Center for Scientific Review
Clinical Center
John E. Fogarty International Center

Marcus, Felix Moore, and others rounded out the cadre of statisticians. Morton Kramer headed a statistics group at the separate National Institute of Mental Health.

Harold Dorn (NIH tenure 1946–1963) was the initial force in the recruitment and the building of a biostatistical presence at NIH (Figure 1). Jerry Cornfield (NIH tenure 1947–1967), hired by Dorn, is characterized by many as a leader who created the theoretical foundation for extensive methodological research in epidemiology and **clinical trials** (Figure 2), and whose forceful influence with physicians and epidemiologists enhanced the prestige of biostatistics at NIH. Unfortunately, both suffered relatively early and untimely deaths.

Many biostatisticians who stayed at NIH for much of their careers tended to move around to different institutes. The initial group of Harold Dorn, Marvin Schneiderman, Jerome Cornfield, Jacob Leiberman, Nathan Mantel, and Samuel Greenhouse arrived in about 1946 (Figure 3). The first "splitting off" of the individuals came in about 1948, when Max Halperin also arrived (Figure 4). A more detailed listing of the comings and goings of the early arrivals is given in Table 2.

Figure 1 Harold Dorn (*ca.* 1950)

Figure 2 Jerry Cornfield (1974)

In 1948 both the statistical group and the NIH were quite small. In the next decade the NIH grew tremendously, as did the numbers and international reputation of its statisticians. Several papers explore the environment conducive to scientific collaboration at NIH in the early decades [1–4, 6–11, 14], indicating that NIH stands as a model emulated throughout the US and the world for the interface of statistics and medicine. At one point in the 1960s there was concern about the absence of intermediate level statisticians to take over in the future and the impact on recruiting younger statisticians that this absence would have. This concern arose after statisticians who would have been very successful in serving as leaders in the future, individuals such as Seymour Geisser and Marvin Zelen, left the NIH for universities.

The medical leaders at NIH have been strong supporters of biostatisticians over the years, and biostatistics as a discipline has flourished. Indeed, for more than 50 years NIH has been home to many of the most influential biostatisticians and the most

Table 2 Chronologic overview of (≥5 year) tenures of early (entry before 1965) NIH Biostatisticians

Statistician	Years, 1st institute		Years, 2nd institute		Years, 3rd institute	
Lieberman, Jacob E.	1947–56	NCI	1957–1962	DRS	1963–70	NHI
Dorn, Harold	1947–56	NCI	1957–59	DRS	1960–62	NHI
Cornfield, Jerome	1947–56	NCI	1957–58	DRS	1960–67	NHLBI
Moore, Felix	1947–57	NHI				
Mantel, Nathan	1947–74	NCI				
Marcus, Samuel C.	1948–60	NCI				
Greenhouse, Samuel W.	1948–53	NCI	1954–66	NIMH	1967–74	NICHD
Cutler, Sidney J.	1948–75	NCI				
Schneiderman, Marvin A.	1948–80	NCI				
Sadowsky, Doris A.	1949–53	NCI	1954–79	NINDS		
Kramer, Morton	1949–75	NIMH				
Kahn, Harold	1950–51; 1960–70	NHI	1957–60	OD	1971–75	NEI
Halperin, Max	1951–55	NHI	1955–58	DBS	1966–77	NHLBI
Kroll, Bernard H.	1951–58	NIMH	1959–82	NINDS		
Loveland, Donald	1951–59	NCI	1970–74	NICHD		
Haenszel, William	1952–76	NCI				
Gordon, Tavia	1954–58	NHI	1958–60	NCI	1966–77	NHLBI
Pollack, Earl S.	1954–77	NIMH	1977–85	NCI		
Geisser, Seymour	1955–61	NIMH	1962–65	NIAMD		
Bailar, John C. III	1955–80	NCI				
Morrison, Donald F.	1956–63	NIMH				
Ederer, Fred	1957–64	NCI	1964–71	NHLBI	1971–86	NEI
Chiazze, Leonard Jr	1957–66	NCI				
Goldberg, Irving D.	1957–66	NINDS				
Crittenden, Margaret	1958–61	NCI				
Gurian, Joan M.	1958–64	NCI	1965–71	NHLBI		
Gehan, Edmund A.	1958–67	NCI				
Rosen, Beatrice M.	1958–81	NIMH				
Deutchberger, Jerome	1959–68	NINDS				
Myers, Max H.	1960–86	NCI				
Markush, Robert E.	1961–66	NHI	1967–69	NINDS	1970–74	NIMH
Jackson, Esther C.	1961–77	NINDS				
Pettigrew, Karen	1961-present	NIMH				
Schachter, Joseph	1962–65	NHI	1965–67	DRS	1971–74	NIAID
Hawkins, C. Morton	1962–66	NINDS				
Weiss, William	1962–84	NINDS				
Seigel, Daniel	1963–67	NHI	1967–76	NICHD	1977–91	NEI
Zelen, Marvin	1963–67	NCI				
Pettigrew, Hugh	1963–89	NCI				
Gart, John J.	1965–91	NCI				

Abbreviations: OD, Office of the Director, NIH; NCI, National Cancer Institute; NHI and NHLBI, National Heart and Heart, Lung, and Blood Institute; NEI, National Eye Institute; NIAMD, National Institute for Arthritis and Musculoskeletal Diseases; NIMH, National Institute of Mental Health; NICHD, National Institute of Child Health and Human Development; NINDS, National Institute of Neurological Disorders and Stroke; DRS, Division of Research Services; DBS, Division of Biologics Standards.

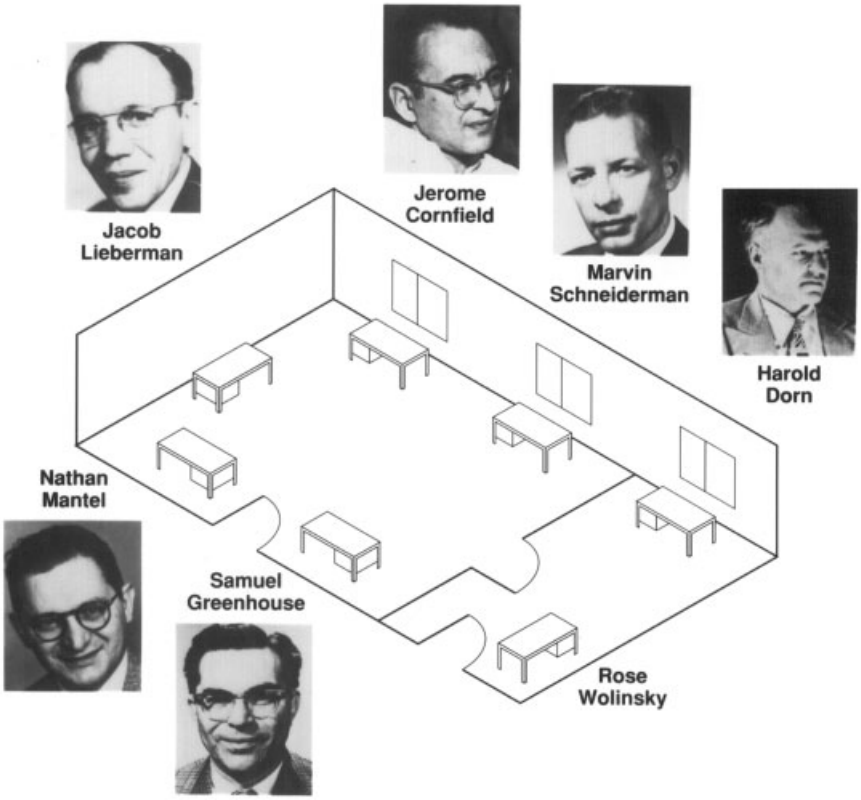

Figure 3 US Public Health Service, Division of Statistical Methods, 1947

important developments in the design and analysis of biomedical experiments. The statistical foundations for epidemiologic **case–control studies**, the use of **regression** models for identification of high-risk individuals, and key methodology for the conduct of modern clinical trials all originated with NIH statisticians.

NIH has also played a major role in the training of biostatisticians. In 2004, 16 US universities offer doctoral training programs (27 separate programs) with funding from 10 NIH institutes.

In celebration of 50 years of biometry at the NIH, a conference was held in January 1993 to acknowledge and review contributions of those pioneers who laid the foundation in the 1940s and continued, through their contributions, persuasiveness, and perseverance, to foster the strong presence of biostatistics at NIH that exists today [5].

Biostatistics, representing the science of the design of biomedical experiments (*see* **Experimental Design**) and the analysis of quantitative data, is more important today than ever. The goal of biostatisticians at the NIH today is to continue to provide statistical insight and rigor to NIH investigations. At the 1993 conference celebrating 50 years of biostatistics at NIH, the diversity and depth of the ongoing biostatistical collaboration was demonstrated by the range of topics presented. Some examples are: **time series** for modeling counts from a relapsing–remitting disease (P.S. Albert); a comparison of **likelihood**-based and marginal **generalized estimating equation** methods for analyzing repeated **ordered categorical** responses with *missing data* (S.D. Mark, M.H. Gail); the utility of large-simple trials in the evaluation of **AIDS** treatment strategies (S. Ellenberg); mixed effects regression models for studying the natural history of prostate disease (J.D. Pearson); partial **questionnaire design** for case–control studies (S. Wacholder); stochastic curtailment (*see* **Data**

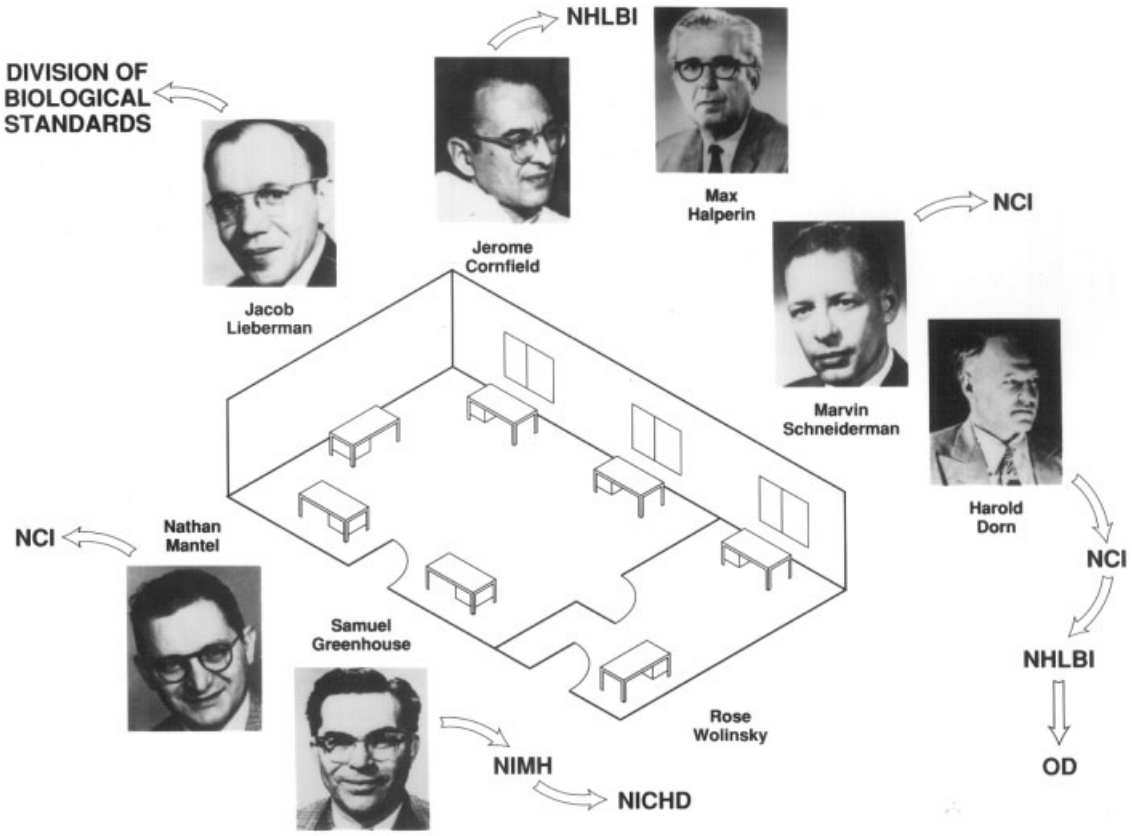

Figure 4 The methodologic "big bang"

and Safety Monitoring) and conditional **power** in matched case–control studies (S. Hunsberger); and a comparison of tumor incidence analyses applicable in single-sacrifice animal experiments [5].

The ongoing collaborations among those interested in developing and applying new statistical methods and those interested in solving biomedical problems account for the continued vitality of biostatistics at NIH.

References

[1] Byar, D.P. (1990). Discussion of papers on "Historical and methodological developments in clinical trials at the National Institutes of Health", *Statistics in Medicine* **9**, 903–906.

[2] Colton, T., Greenhouse, S.W., Zelen, M., Gehan, E.A., Friedewald, W., DeMets, D., Ware, J.H., Gordon, T., Lachin, J.M. & Wittes, J. (1990). Remembrances of Max Halperin, *Statistics in Medicine* **9**, 863–870.

[3] Ederer, F. (1982). Jerome Cornfield's contributions to the conduct of clinical trials, *Biometrics* **38**, 25–32.

[4] Ellenberg, J.H. (1995). Some perspectives on the career of Samuel W. Greenhouse: the first 75 years, *Statistics in Medicine* **14**, 1615–1619.

[5] Ellenberg, J.H., Gail, M.H. & Simon, R.M. (1994). National Institutes of Health Conference on Current Topics in Biostatistics, *Statistics in Medicine* **13**, 399–794.

[6] Gehan, E.A. & Schneiderman, M.A. (1990). Historical and methodological developments in clinical trials at the National Cancer Institute, *Statistics in Medicine* **9**, 871–880.

[7] Greenhouse, S.W. (1982). A tribute to Jerome Cornfield, *Biometrics* **38**, 3–6.

[8] Greenhouse, S.W. (1982). Jerome Cornfield's contributions to epidemiology, *Biometrics* **38**, 33–45.

[9] Greenhouse, S.W. (1990). Some historical and methodological developments in early clinical trials at the National Institutes of Health, *Statistics in Medicine* **9**, 893–901.

[10] Halperin, M., DeMets, D.L. & Ware, J.H. (1990). Early methodological developments for clinical trials at the National Heart, Lung and Blood Institute, *Statistics in Medicine* **9**, 881–892.

[11] Kramer, M. (1975). Some perspectives on the role of biostatistics and epidemiology in the prevention and control of mental disorders, *Milbank Memorial Fund Quarterly*, **Summer**.

[12] Mantel, N. (1976). A personal perspective on statistical techniques for quasi experiments, in *On the History of Statistics and Probability*, D.B. Owen, ed. Marcel Dekker, New York, pp. 103–129.

[13] Mantel, N. (1982). Jerome Cornfield and statistical applications to laboratory research: a personal reminiscence, *Biometrics* **38**, 17–23.

[14] Schneiderman, M.A. (1977). The numerate sciences – epidemiology and biometry, *Journal of the National Cancer Institute* **59**, 633–644.

JONAS H. ELLENBERG & DENNIS O. DIXON

National Surgical Adjuvant Breast and Bowel Project

During the second half of the twentieth century, the scientific method, a process by which hypotheses generated from laboratory and clinical investigation are tested in **randomized** clinical trials, began to be more frequently used. This seminal advance, which marked the transition from anecdotalism and inductivism to science, has accounted for most of the progress that has been made in the management of breast cancer for the past 50 years.

During that time I was fortunate to have been associated with the National Surgical Adjuvant Breast and Bowel Project (NSABP), a cooperative group that I viewed as an extension of my laboratory and that conducted what is now known as translational research. The findings from large randomized clinical trials carried out by my NSABP colleagues and I have, in large part, been responsible for several paradigm shifts that have occurred in the management of breast cancer during the last half century. This overview provides an account of the origin and early years of the NSABP, a description of the major findings obtained from NSABP breast cancer trials involving more than 50 000 women, and a commentary about the accomplishments of the group.

The Origin and Early Trials of the NSABP (1957–1970)

The NSABP grew out of the Surgical Adjuvant Chemotherapy Projects, a program sponsored by the National Institutes of Health (NIH) Cancer Chemotherapy National Service Center (CCNSC) to test the effectiveness of various anticancer drugs used with cancer surgery. The rationale for the project was based on clinical and laboratory findings obtained during the 1950s which suggested that systemic agents administered during and shortly after "curative" operations could improve the outcome of cancer patients. There was evidence that tumor cells were dislodged into the blood during surgery, thus making the procedure less effective; that the growth of cancer cells injected into the blood of animals could be impaired by chemotherapy; and that thiotepa (TSPA) and several other agents might be effective in destroying such cells in humans.

In the spring of 1957, 23 surgeons were invited by Dr I.S. Ravdin, chairman of the CCNSC Clinical Studies Panel, to attend a meeting at Stone House on the NIH campus to discuss the creation of the Surgical Adjuvant Chemotherapy Breast Project (later known as the NSABP), which had as its goal the conduct of clinical trials. Each of the surgeons who participated in that project agreed to abide by specific criteria for the inclusion of patients that had been outlined in a predefined **protocol** and to adhere to strict randomization procedures that divided the patients into treatment and control groups. Randomization was planned to prevent bias in selecting patients for a particular treatment. There were also plans for centralized data collection, evaluation, and review of pathologic material, as well as a program for long-range follow-up. The willingness among this group of surgeons to follow a predefined protocol represented a radical departure from conventional practice and set the stage for the more sophisticated protocols that would subsequently be designed by the NSABP.

Early in 1964 an executive committee was formed to coordinate and direct the study and to provide more effective liaison among project participants. I was co-chairman of the group and was subsequently appointed chairman and principal investigator of the Surgical Adjuvant Chemotherapy Breast Project on 9 May 1967. For the next three years the operations and statistical centers of the group remained at Roswell Park Cancer Institute in Buffalo, New York, while I interacted with them from Pittsburgh.

The first Surgical Adjuvant Chemotherapy Breast Project trial, called Phase I (a term that should not be confused with the current definition of a Phase I trial), compared the outcome of patients treated by radical mastectomy with or without the administration of TSPA. That study accrued 826 eligible patients between April 1958 and October 1961, a remarkable achievement at the time. The results of the Phase I study, which subsequently became known as NSABP B-01, provided the first evidence that the use of chemotherapy could alter the natural history of some patients and demonstrated, for the first time, that the outcome of patients with one to three positive axillary nodes was different from that of patients with four or more positive nodes [10, 12]. However, because so few patients seemed to benefit from

the chemotherapy and because it resulted in toxicity, surgeons were reluctant to accept the use of that treatment modality. When patient entry into the Phase I program was completed, a new study, Phase II, was initiated. Its objective was to evaluate the worth of 5-fluorouracil (5-FU), as compared with TSPA, and the value of postoperative radiotherapy and prophylactic oophorectomy. Findings from the Phase II trial demonstrated no advantage from the use of 5-FU over TSPA and showed that the toxicity resulting from the 5-FU regimen was actually of greater magnitude.

In 1961, as part of the Phase II study, a randomized trial (NSABP B-02) was begun to resolve the uncertainties about the worth of administering postoperative radiation therapy as an adjunct to radical mastectomy. Through five years of follow-up [11], the data failed to confirm conclusions derived from anecdotal information that had been obtained from the use of similar radiation techniques that had indicated an improvement in survival. Our findings resulted in great controversy. It is interesting to note that, 40 years later, uncertainty still exists with regard to the worth of using postoperative radiation therapy to improve survival outcome.

Because uncertainty also existed at that time with regard to the use of prophylactic oophorectomy as an adjunct to radical mastectomy, in 1961 the NSABP initiated the B-03 study, a randomized clinical trial that was designed to evaluate that treatment regimen in premenopausal breast cancer patients. Preliminary findings from B-03 demonstrated no difference in either recurrence or survival data among patients who had been treated by either oophorectomy, TSPA, or placebo [34]. Accruing patients to the study was difficult because there was a lack of appreciation of the urgency for resolving that question – a situation that was to prevail for the next 30 years. B-03, as well as the trial to evaluate postoperative radiation therapy, was never updated because, after the NSABP relocated to Pittsburgh, the data that had been stored at Roswell Park were never made available.

The Phase I and Phase II studies related patient outcome to the location of breast tumors. At that time, it was widely believed that patients with tumors in the inner quadrants of the breast had a poorer prognosis than did those with lesions in the outer quadrants. An evaluation of more than 1000 patients in the Phase I and Phase II studies demonstrated that the location of a tumor was unrelated to prognosis and, thus, led us to conclude that there was no justification for selecting specific surgical or radiation therapy approaches for treatment based upon tumor location. Other information from the patients entered into those trials demonstrated that the larger the tumor the more likely that axillary nodes would be positive, that more nodes would be involved, and that patient outcome would be poorer. This led to the conclusion that, in itself, size was not necessarily related to "earliness" or "lateness" of a tumor and was not as consequential as other tumor and/or host factors that determine the development of metastases. We also correlated recurrence and survival rates with number of lymph nodes examined in surgical specimens. Results indicated that examination of a greater number of nodes in a specimen was no more meaningful in determining prognosis than examination of only a few. Those findings are still relevant to current arguments about the management of axillary nodes in breast cancer patients.

The interval between 1957 and 1969 marked a learning period in the conduct of clinical trials, especially with regard to experimental design. In retrospect, the early trials of the Surgical Adjuvant Chemotherapy Breast Project were too complicated and represented a desire on the part of investigators to answer too many questions at once. This circumstance led to my view that clinical trials should be kept simple and that only a few questions should be answered in any single study. Although the overall results of those trials were disappointing, they were the first to demonstrate that cooperative studies using adjuvant therapy could be effectively conducted among large groups of investigators nationwide.

In addition to marking the rise in the use of the clinical trials mechanism for evaluating adjuvant therapy in the treatment of operable breast cancer, the interval between 1957 and 1969 was noteworthy because of the amount of information that was made available for augmenting existing biologic hypotheses and for generating new ones. The period was particularly significant because it marked my introduction to the clinical trials process, which, in turn, stimulated my interest in tumor metastases. During the late 1950s and 1960s, my laboratory associates and I published our findings in more than 50 scientific papers. The information from our studies led me to formulate an alternative hypothesis that became the basis for a new generation of NSABP clinical trials.

Paradigm Shifts

A New Surgical Paradigm

If any scientific basis existed for the disagreement that persisted for several decades about the operative management of breast cancer, then it related to differences in perception about the biology of the disease, particularly in terms of tumor spread. Two divergent hypotheses were at the center of the variance in opinion. One, the concept formulated by William S. Halsted at the end of the nineteenth century, gave rise to the paradigm that governed the surgical management of breast cancer for most of the twentieth century. The Halstedian paradigm was based on an anatomic and mechanistic perception of tumor spread that was in keeping with the understanding of the biology of tumor metastases at the time [5]. The tenets of this hypothesis gave rise to an anatomic basis for cancer surgery in which the Halsted radical mastectomy, characterized by *en bloc* dissection, became the hallmark of a surgical approach that emphasized that curability could more effectively be achieved as a result of more expansive, meticulously performed surgical procedures. The use of radiation therapy after surgery was governed by the same principles.

The alternative hypothesis, which I synthesized nearly 25 years ago, contended that cancer is a systemic disease that involves a complex spectrum of host–tumor interrelations and that variations in local–regional therapy are unlikely to substantially affect survival [6]. That premise was formulated from a series of laboratory and clinical investigations that my associates and I carried out from 1958 to 1970 to obtain a better understanding of the biology of metastases. All our findings had the same characteristic, i.e. they did not conform to the concepts that served as the basis for the principles of the Halstedian hypothesis but, rather, provided a matrix for the formulation of an alternative thesis. That hypothesis, which we developed in 1968, is biologic rather than anatomic and mechanistic in concept, and its components are completely antithetical to those of Halsted's thesis. Consequently, we hypothesized that variations in the local–regional treatment of breast cancer, i.e. different surgical regimens, were unlikely to affect patient outcome. Thus, in August 1971 we implemented the first of two clinical trials designed to test the validity of the principles upon which our alternative hypothesis was based. In that study, approximately 1700 women [14] without clinical evidence of axillary node involvement were randomized among three treatment regimens: (1) radical mastectomy; (2) total (simple) mastectomy with local–regional irradiation, but no axillary dissection; or (3) total mastectomy and removal of axillary nodes only if they later became clinically positive.

The significant aspect of the B-04 study was that 40% of the patients with clinically negative nodes treated by radical mastectomy were found to have histologically positive nodes. Thus, about 40% of the patients in the groups treated by total mastectomy alone had positive axillary nodes that were left unremoved. Despite this therapeutic nonconformity, no significant difference in overall treatment failure, distant metastasis, or overall survival has been noted among the three groups through more than 20 years of follow-up [33]. The negation of the primacy of radical mastectomy and the principles upon which it was based eliminated most of the biologic considerations that might have contraindicated the performance of breast-conserving operations. Consequently, in 1976 we instituted a second trial, NSABP B-06, to re-evaluate the principles of our hypothesis and, at the same time, to appraise the worth of lumpectomy plus axillary dissection for the surgical management of breast cancer.

In the B-06 study, women with breast tumors of less than or equal to 4 cm in size were randomly assigned to one of three treatment groups: (1) total mastectomy, (2) lumpectomy, or (3) lumpectomy followed by breast irradiation. The lumpectomy removed enough tissue to ensure that the margins of resected specimens were free from tumor. Women in all treatment groups had an axillary dissection, and those with positive nodes received chemotherapy. In almost 2000 women through at least 12 years of follow-up, results (first reported in 1985 [15]) demonstrated the efficacy of lumpectomy and radiation therapy for the treatment of breast cancer [22]. There continues to be no significant difference in either distant distant disease-free survival (DDFS) or overall survival among the three treatment groups. Of most importance is the observation that no difference in DDFS or overall survival has yet occurred despite the fact that the total mastectomy group (by virtue of breast removal) had no breast tumor recurrence, that lumpectomy patients who received radiation therapy had an ipsilateral tumor recurrence rate of 10%, and that patients treated with lumpectomy alone demonstrated a tumor

recurrence rate of 40%. Aside from indicating the efficacy of breast conservation, the findings from B-06, just as did those from B-04, repudiated Halstedian principles of breast cancer management and provided support for our alternative hypothesis. Moreover, they indicated that a large proportion of women with breast cancer could be treated with breast conservation.

In the years between the initiation of the B-04 study and the report of findings from B-06, i.e. from 1971 to 1985, a radical shift occurred in the treatment of primary breast cancer. Most significantly, the events described led to emancipation from conventional thinking about breast cancer and its treatment and set the stage for new scenarios that were to occur in rapid succession. In a sequence that represented an orderly scientific process, one paradigm governing breast cancer management was replaced by another. As a result, Halstedian principles of cancer surgery must now be viewed as nothing more than historical "milestones" against which cancer progress can be measured, and the Halstedian paradigm must now be permitted to assume its proper place in the annals of surgical history.

The Systemic Therapy Paradigm

After a hiatus of almost a decade, the NSABP launched a new trial to evaluate adjuvant therapy. Not until the early 1970s, after Skipper and his associates defined the concept of a growth fraction in a tumor cell population and provided an array of tumor growth kinetic principles that were instrumental in formulating a hypothesis that postulated the value of adjuvant chemotherapy, did such a trial become supportable [35, 36]. The NSABP initiated the B-05 trial to evaluate adjuvant therapy and to test that hypothesis in 1972. In that study, L-phenylalanine mustard (L-PAM) was administered after radical mastectomy to patients with positive axillary nodes. The results, reported in 1975, were the first (subsequent to our 1958 trial) to demonstrate that systemic adjuvant therapy could alter the natural history of certain cohorts of patients with primary breast cancer [13]. Within a short time, that conclusion was confirmed by findings from a study by other investigators in which cyclophosphamide, methotrexate, and 5-fluorouracil (CMF) was used [1].

In general, postoperative chemotherapy for the treatment of node-positive, premenopausal women (less than 50 years of age) has been widely accepted. Of particular importance in the treatment of such women was the report of findings from NSABP B-15, which demonstrated that patients treated with doxorubicin (Adriamycin) and cyclophosphamide (AC) over 63 days had the same outcome as those who received six months of conventional CMF administered on each of 84 days [18]. In 1984 the NSABP implemented B-16, a trial designed to determine whether tamoxifen plus chemotherapy (AC) was more effective than tamoxifen alone in improving the Disease-free Survival (DFS), DDFS, and overall survival of node-positive, tamoxifen-responsive patients aged 50 years and older. Results from almost 1200 eligible patients provided the first definitive information to demonstrate a greater benefit from tamoxifen plus AC than from tamoxifen alone [19]. Moreover, our studies had failed to demonstrate an unfavorable interaction between chemotherapy and tamoxifen when that regimen was administered simultaneously to tamoxifen-responsive patients. As a result of the findings from the B-15 and B-16 studies, we concluded that AC therapy given over 63 days produced results comparable with those from CMF therapy given over 154 days. We also recommended that node-positive patients aged 50 years and older should receive tamoxifen in addition to chemotherapy.

In the late 1980s it was hypothesized that failure to achieve a greater therapeutic effect from the use of systemic therapy was due to inadequate drug administration. It was considered that a higher dose intensity, i.e. amount of drug administered per unit of time, would result in a better outcome. On the basis of this consideration, we implemented NSABP B-22, a study in node-positive patients that involved manipulating the total dose and intensity of cyclophosphamide in an AC combination. We reported [25] no significant difference in DFS among the treatment groups. Results from a companion study (NSABP B-25), in which the intensity and total dose of cyclophosphamide were increased to levels beyond that in B-22, showed that an even greater drug intensity and/or cumulative dose than was administered in B-22 failed to demonstrate a benefit beyond that obtained using the "standard" dose and intensity of the drug. Failure to note a difference in outcome among the groups was unrelated to either differences in amount and intensity of cyclophosphamide or to dose delays and intervals between courses of therapy. We concluded that, until

there was more information about the worth of more intensive high-dose therapy, increasing the total dose of cyclophosphamide was inappropriate in the treatment of women with primary breast cancer.

At an NIH Consensus Conference held in 1985, it was decided that a lack of information precluded recommending therapy other than surgery for breast cancer patients with negative axillary nodes [3]. Findings from four NSABP randomized clinical trials involving more than 8000 patients subsequently indicated the propriety of using systemic therapy to treat such women. Two of the studies, NSABP B-13 and B-19, were conducted to evaluate the worth of adjuvant chemotherapy in patients with estrogen-receptor (ER)-negative tumors.

In B-13, 760 women were randomly assigned to either methotrexate (M) and sequentially administered fluorouracil (F) (M→F) followed by leucovorin, or to surgery and no chemotherapy [16, 23]. In B-19, a total of 1095 women with the same eligibility requirements were randomly assigned to receive either M→F or cyclophosphamide (C) together with MF (CMF) as conventionally used [23]. The aim of the B-19 trial was to determine if the alkylating agent cyclophosphamide contributed an additional benefit when used in a chemotherapeutic regimen. Data from both studies led us to conclude that M→F and CMF were effective for women with ER-negative tumors and negative axillary nodes. In the younger age group, treatment with CMF clearly resulted in a better DFS and survival; in the older age group, a benefit from both regimens was apparent, although it was less clear which regimen was most effective. Because severe toxicity was less frequent after M→F therapy, that regimen was recommended for older women with associated medical problems that might preclude the use of more toxic agents.

Two additional trials, B-14 and B-20, were conducted by the NSABP in patients with ER-positive tumors. The aim of B-14, a randomized, double-blind, placebo-controlled trial initiated in 1982, was to determine the effectiveness of adjuvant therapy with tamoxifen in patients with negative axillary nodes [17]. That study, which involved more than 2800 randomized and 1200 registered tamoxifen-treated patients, has, arguably, provided some of the most compelling information that has been gathered during the past decade. Through 10 years of follow-up, a significant advantage was observed in DFS and survival among tamoxifen-treated women 49 years of age or younger and 50 years old or older [24]. Tamoxifen therapy was also associated with a significant reduction in the incidence of contralateral breast cancer, a finding that led us subsequently to consider the use of that drug for the prevention of breast cancer. No additional benefit, however, was observed from tamoxifen administration beyond five years.

Before the B-14 findings became available, however, we concluded that the degree of benefit achieved with tamoxifen in this patient population was unlikely to be sufficiently great to eliminate the need for other trials to test potentially more effective regimens. As a result of that consideration, we implemented NSABP B-20, a study that involved more than 2300 women and was aimed at testing the hypothesis that the addition of either M→F or CMF to tamoxifen (MFT, CMFT) would result in a greater benefit than that which could be achieved with tamoxifen alone [26]. In that trial, chemotherapy plus tamoxifen resulted in significantly better DFS and survival than that observed with tamoxifen alone [26]. When compared with tamoxifen alone, MFT and CMFT reduced both the rate of ipsilateral breast tumor recurrence (IBTR) after lumpectomy and the rate of recurrence at other local, regional, and distant sites. Of particular significance was the observation that the rate of treatment failure was reduced after the administration of both types of chemotherapy, regardless of the size of a patient's tumor, the degree of tumor ER positivity, progesterone-receptor (PgR) level, or age. In addition, we failed to identify any subgroup of patients who did not benefit from chemotherapy.

Because the B-14 findings showed that tamoxifen significantly reduced the rate of treatment failure at local and distant sites, the rate of tumors in the contralateral breast, and the incidence of IBTR, and because all subgroups of patients benefited and the benefits were attained with a relatively low incidence of undesirable side-effects, we concluded that tamoxifen was justified for women who met the eligibility criteria of the study participants. Also, when analyses of the B-20 study failed to identify a subgroup of women with ER-positive tumors and negative nodes who failed to benefit from either MFT or CMFT, we concluded that women similar to those in the trial might be considered candidates for chemotherapy plus tamoxifen. However, because patients with small (1 cm or less), mammographically identified lesions were rarely enrolled in the B-14

and B-20 studies, no information was obtained to indicate whether or not such women "should receive tamoxifen or tamoxifen and chemotherapy".

The findings from the four studies conducted in axillary node-negative patients raised several issues. They demonstrated that, after surgery alone, the overall prognosis of such patients was sufficiently poor to warrant the use of systemic therapy. In fact, the prognosis of some of those women was worse than that of women with positive nodes. That situation prevails in women with ER-negative as well as in women with ER-positive tumors. Although some patients in the four studies did not need systemic therapy, many women derived substantial benefit from it. Findings from appropriate statistical analyses that we conducted in search of inconsistencies in treatment effect among node-negative patients in our studies failed to identify subgroups of women with either ER-negative or ER-positive tumors who did not achieve some benefit from systemic adjuvant therapy, i.e. chemotherapy in the former and tamoxifen plus chemotherapy in the latter. However, as has previously been mentioned, definitive information about the propriety of using tamoxifen and/or chemotherapy for the management of patients with tumors of 1.0 cm or less in size is not yet available.

Consequently, when asked the question, Should *all* patients with negative nodes and ER-positive tumors be treated with tamoxifen plus chemotherapy?, we are unable to provide as precise an answer as we would like, despite the extent and credibility of our findings. Certainly, there are patients who either will not benefit from or will not need such therapy. Because other investigators have also been unable to identify these women with precision, we have taken the position that, until markers are found that will be able to do so with greater certainty than now exists, patients who have been judged to be candidates for such therapy should not be denied the chance to receive it so that they may experience the benefit that has been demonstrated.

In a recent attempt to identify subpopulations of node-negative women with ER-positive tumors who might require more aggressive forms of therapy and to distinguish them from women for whom such therapy is unnecessary, we assessed the outcomes, through 10 years of follow-up, of more than 4000 node-negative patients in B-14 who received either placebo or tamoxifen, taking into account their age at surgery, ER status, PgR status, tumor size, tumor S-phase fraction, and tumor nuclear grade [2]. Tumor size and S-phase were viewed as continuous variables [2]. Perhaps the most significant of the findings from that study was the observation of an extreme heterogeneity of outcomes among a population that has, until recently, been considered to have a favorable prognosis. For example, the 10-year DFS for 35-year-old women in the B-14 trial varied from about 80% to less than 40%, depending upon the interaction of the various prognostic variables being assessed. Thus, a group of node-negative patients who received tamoxifen or placebo consisted of women who displayed myriad heterogeneous outcomes. These observations indicate that more aggressive therapy is warranted for some, but not all, patients in the node-negative, ER-positive population. Before a specific therapeutic regimen is prescribed, however, each patient's prognosis must first be determined as accurately as is currently possible.

The remarkably low incidence of IBTR observed in all of the node-negative and node-positive studies following lumpectomy, breast irradiation, and systemic therapy has further justified the use of breast-conserving surgery for most women. Those findings support the author's long-held contention that the effect of local–regional therapy, i.e. surgery and radiation therapy, should no longer be considered independent of the effect of systemic therapy. Thus, the two separate and independent paradigms that governed the management of breast cancer, one that related to the management of local and regional disease by surgery and radiation therapy and the other that involved governing the treatment of micrometastatic disease, have merged into a single, interdependent paradigm.

The Preoperative Chemotherapy Paradigm

Hypotheses formulated from biologic and clinical information obtained during the 1980s led us to initiate NSABP B-18, the first randomized clinical trial (involving more than 1500 women) that was designed to evaluate the role of preoperative chemotherapy in the treatment of primary operable breast cancer [27, 28]. Although the use of such therapy failed to improve the overall benefit beyond that of patients who were randomly assigned to receive the same therapy postoperatively, the findings demonstrated that preoperative chemotherapy could be used

without fear of decreasing the DFS or survival of patients who received it. The most compelling findings were those that demonstrated that women whose tumors displayed a clinical and pathological complete response as the result of preoperative therapy had a more favorable outcome than did women whose tumors displayed either a clinical complete response or a clinical partial response. Thus, we concluded that the response of a breast tumor to preoperative chemotherapy could serve as a surrogate or intermediate end point for determining the response of micrometastases to systemic therapy. Because breast tumor response could be determined within weeks after preoperative chemotherapy was administered, it became possible to predict a patient's outcome and then to provide her with information so that she and her physician could consider other treatment strategies without having to postpone therapy until a treatment failure occurred.

Another finding of particular importance demonstrated that the downstaging of large tumors after the use of preoperative chemotherapy permitted more patients to be treated with lumpectomies. As a consequence, I have proposed that women with tumors judged by surgeons to be too large for lumpectomies, or women whose surgeons are ambivalent about performing that procedure, should initially have the option of receiving preoperative chemotherapy to determine whether the primary tumor sufficiently decreases in size so that lumpectomy and radiation therapy, rather than mastectomy, can be carried out in an attempt to enhance quality of life without increasing the risk for distant disease. Finally, the finding that preoperative chemotherapy downstages axillary lymph node status, i.e. converts nodal status from positive to negative, must be taken into account before a decision with regard to the management of axillary nodes can be made.

Whether or not preoperative chemotherapy is sufficiently important to replace postoperative systemic therapy remains to be seen. At least at this time, there is justification to suggest its use in certain circumstances.

The Breast Cancer Prevention Paradigm

The concept that tamoxifen could be used to prevent breast cancer had its origins in the late 1970s and 1980s, when the drug was shown to be of value in a variety of laboratory and clinical settings. Particularly germane to the concept of breast cancer prevention was tamoxifen's demonstrated ability to reduce the incidence of contralateral breast cancer [9]. To test that thesis, in 1992 the NSABP implemented the P-1 trial [8, 29]. In that study, women at increased risk for invasive breast cancer were randomly assigned to receive either placebo or tamoxifen for five years. The study findings demonstrated that tamoxifen decreased the overall risk of invasive and noninvasive breast cancer by 50%, a reduction that occurred in various age groups and categories of risk. The incidence of ER-positive tumors (but not of ER-negative tumors) was also reduced. The findings obtained in women who had a history of lobular carcinoma *in situ* (LCIS) or atypical hyperplasia – pathological entities considered to increase the risk of invasive breast cancer – were of particular importance, as they not only provided the only quantitative information available from a clinical trial to indicate the magnitude of their risk, but also demonstrated that the risk could be substantially reduced by tamoxifen administration. These findings are particularly relevant to those that were recently obtained from our studies in which strategies for the treatment of ductal carcinoma *in situ* (DCIS) were evaluated.

In 1985, because of uncertainty regarding the management of DCIS, we initiated B-17, the first randomized clinical trial to test the hypothesis that the treatment of localized DCIS by lumpectomy with tumor-free specimen margins followed by radiation therapy was more effective than was lumpectomy alone in preventing the subsequent occurrence of invasive tumor in the ipsilateral breast [20]. The 1993 report of our findings supported that hypothesis and demonstrated that postoperative breast irradiation markedly reduced the subsequent occurrence of invasive ipsilateral breast tumors. When the outcome of patients was examined relative to a wide array of pathologic and mammographic characteristics, we failed to find a discriminant that identified any group of DCIS patients who did not benefit from postoperative radiation therapy.

Because our previous trials and the studies of other investigators had demonstrated a benefit from tamoxifen administration in a variety of settings, it was considered that the drug might interfere with either the development of a primary invasive cancer from its start, or with the progression of residual DCIS to invasive cancer in women with a history of DCIS.

Consequently, in 1991 we initiated a second randomized clinical trial, NSABP B-24, to test the hypothesis that treatment with postoperative radiation therapy and tamoxifen would be more effective in patients who had had DCIS removed either with or without tumor-free specimen margins than would radiation therapy alone in preventing invasive and noninvasive cancers in the ipsilateral and contralateral breast [30]. The results of that study demonstrated that the risk of ipsilateral breast cancer was lower in women treated with tamoxifen, regardless of whether specimen margins were tumor free and regardless of whether DCIS was associated with comedonecrosis. Because the benefit from tamoxifen was due to a decrease in the rate of ipsilateral, contralateral, and metastatic invasive breast cancers, it seemed reasonable to conclude that focusing only on the frequency with which ipsilateral breast tumors occurred was too limited and that an assessment of the effect of treatment on all the sites combined seemed more appropriate. Finally, because women in the P-1 trial who had a history of LCIS or atypical hyperplasia were thought to be at sufficiently high risk of developing an invasive cancer to warrant being considered candidates for tamoxifen administration, it seemed reasonable to recommend that women with DCIS should also be considered candidates for tamoxifen, since they are at an even greater risk for developing invasive disease, even after they have been treated with radiation therapy.

There has often been more of an emphasis on the adverse effects of tamoxifen than on the benefits resulting from its use. Findings from the NSABP P-1 and B-24 trials, as well as the results of other NSABP studies that have evaluated tamoxifen, have failed to justify concerns about quality-of-life issues, liver damage, hepatoma, retinal toxicity, and cancers at other sites [4, 31, 32]. The excess risk of endometrial cancer [7, 21] and of vascular-related events such as stroke, deep-vein thrombosis, and pulmonary embolism that were observed in the tamoxifen group, as compared with those in the placebo group in these studies, has caused the most concern. In the P-1 study, less than 1 woman per 100 (0.7%) in the tamoxifen group developed endometrial cancer over a five-year period. All such cancers were International Federation of Gynecology and Obstetrics (FIGO) stage 1, and no deaths from endometrial cancer have, to date, been reported. The undesirable vascular events in the tamoxifen group in excess of those in the placebo group over a five-year period were few: 0.2%–0.3% of women experienced a stroke, approximately 0.2% had a pulmonary embolism, and between 0.2% and 0.3% exhibited deep-vein thrombosis. Those events occurred less frequently in women 49 years of age or younger and were slightly more frequent in women 50 years of age or older, being approximately 1% for endometrial cancer and less than 1% for each of the vascular-related events over five years. In view of the relatively few side-effects that resulted from tamoxifen administration, we concluded that its use as a breast cancer preventive agent may be appropriate in many women at increased risk for the disease.

Summary and Comments

This article has provided an overview of some of the more important accomplishments achieved by the NSABP during the past four decades. Subsequent to each of the advances that I have noted, the same paradoxical situation occurred that has relevance for the conduct of future breast cancer research. Because the extent of the unknown is often recognized only as knowledge expands, it is not surprising that, after each demonstration of a therapeutic advance in breast cancer management, issues arise that cannot immediately be resolved. That consequence of accomplishment has, all too often, resulted in confusion and in pessimism about the meaning of the results obtained. Thus, uncertainty arises with regard to the clinical application of the findings. As a result, success may now create more havoc than does failure. This circumstance is, indeed, unfortunate because those putative uncertainties do not detract from either the credibility or the importance of the findings that gave rise to them. It is, indeed, rare, if not impossible, for a single study to provide enough information to eliminate all uncertainties associated with positive achievement. In essence, every answer generates a whole set of new questions.

With the demonstration of the worth of systemic adjuvant therapy for the treatment of invasive breast cancer, postoperative breast irradiation after lumpectomy for the treatment of DCIS, and tamoxifen for reducing the incidence of invasive and noninvasive breast cancer in women at increased risk for the disease, the same questions have arisen. They relate to who will benefit from treatment and who will not;

who will not need the therapy because they will never demonstrate a treatment failure; how much of a benefit is worthwhile; and whether or not the toxicity and mortality encountered justify its administration. Despite these uncertainties, the use of adjuvant therapy is considered to be a major advance in the treatment of breast cancer. Similarly, the use of breast irradiation following lumpectomy for the treatment of both invasive and noninvasive breast cancer has been a major advance in the local–regional treatment of that disease despite the questions that have arisen. Also, the use of tamoxifen to obtund, and perhaps prevent, the development of a phenotypically expressed cancer before its diagnosis denotes a similar advance. Nevertheless, research directed toward identifying cohorts of patients who either do or do not benefit from those therapies that demonstrate an overall advantage must be vigorously pursued because such information will permit more precisely identifying patients according to their need for and response to therapy.

In conclusion, I believe that the continued use of the scientific process is imperative if progress is to continue in breast cancer research and treatment. Although critics of the clinical trials mechanism are numerous and their objections variable, e.g. that such trials take too long, are too cumbersome, are too costly, and are in need of replacement by other mechanisms, until other alternatives become available, such studies continue to provide the most appropriate way of obtaining the kind of information necessary for verifying hypotheses and for evaluating therapies. It is unfortunate that many critics of clinical trials do not participate in them, do not understand the complexities and diligence necessary in their conduct to obtain credible data, and would prefer to continue to believe in the worth of retrospective information for therapeutic decision-making. On the other hand, there is little disagreement that there is a need for some clinical trials to be made simpler, that they be subject to fewer, less rigid rules and regulations, and that the media look upon them more favorably so as to eliminate the fear created by negative publicity, which inhibits patients from participating in them. Finally, the contributions made by the NSABP during the past 40 years to advance the treatment of breast cancer could not have occurred without the cooperation of more than a hundred thousand physicians, biostatisticians, nurses, administrative and support staff, technicians, and, above all, the patients, who participated in approximately 30 major NSABP clinical trials. Their passion in carrying out this effort against a single disease is, in itself, a unique undertaking in medical history.

References

[1] Bonadonna, G., Brusamolino, E., Valagussa, P., Rossi, A., Brugnatelli, L., Brambilla, C., DeLena, M., Tancini, G., Bajetta, E., Musumeci, R. & Veronesi, U. (1976). Combination chemotherapy as an adjuvant treatment in operable breast cancer, *New England Journal of Medicine* **294**, 405–410.

[2] Bryant, J., Fisher, B., Gunduz, N., Costantino, J.P. & Emir, B. (1998). S-phase fraction combined with other patient and tumor characteristics for the prognosis of node-negative, estrogen-receptor-positive breast cancer, *Breast Cancer Research and Treatment* **51**, 239–253.

[3] Consensus Conference. Adjuvant Chemotherapy for Breast Cancer (1985). *Journal of the American Medical Association* **254**, 3461–3463.

[4] Day, R., Ganz, P.A., Costantino, J.P., Cronin, W.M., Wickerham, D.L. & Fisher, B. (1999). Health-related quality of life and tamoxifen in breast cancer prevention: a report from the National Surgical Adjuvant Breast and Bowel Project P-1 study, *Journal of Clinical Oncology* **17**, 2659–2669.

[5] Fisher, B. (1970). The surgical dilemma in the primary therapy of invasive breast cancer: a critical appraisal, *Current Problems in Surgery* **October**, 1–53.

[6] Fisher, B. (1980). Laboratory and clinical research in breast cancer: a personal adventure: the David A. Karnofsky memorial lecture, *Cancer Research* **40**, 3863–3874.

[7] Fisher, B. (1996). A commentary on endometrial cancer deaths in tamoxifen-treated breast cancer patients, *Journal of Clinical Oncology* **14**, 1027–1039.

[8] Fisher, B. (1999). National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial: a reflective commentary, *Journal of Clinical Oncology* **17**, 1632–1639.

[9] Fisher, B. & Redmond, C. (1991). New perspective on cancer of the contralateral breast: a marker for assessing tamoxifen as a preventive agent, *Journal of the National Cancer Institute* **83**, 1278–1280.

[10] Fisher, B., Ravdin, R.G., Ausman, R.K., Slack, N.H., Moore, G.E. & Noer, R.J. (1968). Surgical adjuvant chemotherapy in cancer of the breast: results of a decade of cooperative investigation, *Annals of Surgery* **168**, 337–356.

[11] Fisher, B., Slack, N.H., Cavanaugh, P.J., Gardner, B. & Ravdin, R.G. (1970). Postoperative radiotherapy in the treatment of breast cancer: results of the NSABP clinical trial, *Annals of Surgery* **172**, 711–732.

[12] Fisher, B., Slack, N., Katrych, D. & Wolmark, N. (1975). Ten-year follow-up results of patients with carcinoma

of the breast in a cooperative clinical trial evaluating surgical adjuvant chemotherapy, *Surgery, Gynecology and Obstetrics* **140**, 528–534.

[13] Fisher, B., Carbone, P., Economou, S.G., Frelick, R., Glass, A., Lerner, H., Redmond, C., Zelen, M., Band, P., Katrych, D.L., Wolmark, N. & Fisher, E.R. (1975). L-phenylalanine mustard (L-PAM) in the management of primary breast cancer: a report of early findings, *New England Journal of Medicine* **292**, 117–122.

[14] Fisher, B., Montague, E., Redmond, C., Barton, B., Borland, D., Fisher, E.R., Deutsch, M., Schwarz, G., Margolese, R., Donegan, W., Volk, H., Konvolinka, C., Gardner, B., Cohn, I., Lesnick, G., Cruz, A.B., Lawrence, W., Nealon, T., Butcher, H. & Lawton, R. (1977). Comparison of radical mastectomy with alternative treatments for primary breast cancer: a first report of results from a prospective randomized clinical trial, *Cancer* **39**, Supplement 6, 2827–2839.

[15] Fisher, B., Bauer, M., Margolese, R., Poisson, R., Pilch, Y., Redmond, C., Fisher, E., Wolmark, N., Deutsch, M., Montague, E., Saffer, E., Wickerham, L., Lerner, H., Glass, A., Shibata, H., Deckers, P., Ketcham, A., Oishi, R. & Russell, I. (1985). Five-year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer, *New England Journal of Medicine* **312**, 665–673.

[16] Fisher, B., Redmond, C., Dimitrov, N.V., Bowman, D., Legault-Poisson, S., Wickerham, D.L., Wolmark, N., Fisher, E.R., Margolese, R., Sutherland, C., Glass, A., Foster, R. & Caplan, R. (1989). A randomized clinical trial evaluating sequential methotrexate and fluorouracil in the treatment of patients with node-negative breast cancer who have estrogen-receptor-negative tumors, *New England Journal of Medicine* **320**, 473–478.

[17] Fisher, B., Costantino, J., Redmond, C., Poisson, R., Bowman, D., Couture, J., Dimitrov, N.V., Wolmark, N., Wickerham, D.L., Fisher, E.R., Margolese, R., Robidoux, A., Shibata, H., Terz, J., Paterson, A.H.G., Feldman, M.I., Farrar, W., Evans, J., Lickley, H.L. & Ketner, M. (1989). A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors, *New England Journal of Medicine* **320**, 479–484.

[18] Fisher, B., Brown, A.M., Dimitrov, N.V., Poisson, R., Redmond, C., Margolese, R.G., Bowman, D., Wolmark, N., Wickerham, D.L., Kardinal, C.G., Shibata, H., Paterson, A.H.G., Sutherland, C.M., Robert, N.J., Ager, P.J., Levy, L., Wolter, J., Wozniak, T., Fisher, E.R. & Deutsch, M. (1990). Two months of doxorubicin-cyclophosphamide with and without interval reinduction therapy compared with six months of cyclophosphamide, methotrexate, and fluorouracil in positive-node breast cancer patients with tamoxifen-nonresponsive tumors: results from NSABP B-15, *Journal of Clinical Oncology* **8**, 1483–1496.

[19] Fisher, B., Redmond, C., Legault-Poisson, S., Dimitrov, N.V., Brown, A.M., Wickerham, D.L., Wolmark, N., Margolese, R.G., Bowman, D., Glass, A.G., Kardinal, C.G., Robidoux, A., Jochimsen, P., Cronin, W., Deutsch, M., Fisher, E.R., Myers, D.B. & Hoehn, J.L. (1990). Postoperative chemotherapy and tamoxifen compared with tamoxifen alone in the treatment of positive-node breast cancer patients aged 50 years and older with tumors responsive to tamoxifen: results from the National Surgical Adjuvant Breast and Bowel Project B-16, *Journal of Clinical Oncology* **8**, 1005–1018.

[20] Fisher, B., Costantino, J., Redmond, C., Fisher, E.R., Margolese, R., Dimitrov, N., Wickerham, D.L., Wolmark, N., Deutsch, M., Ore, L., Mamounas, E. & Kavanah, M. (1993). Lumpectomy compared with lumpectomy and radiation therapy for the treatment of intraductal breast cancer, *New England Journal of Medicine* **328**, 1581–1586.

[21] Fisher, B., Costantino, J.P., Redmond, C., Fisher, E.R., Wickerham, D.L. & Cronin, W. (1994). Endometrial cancer in tamoxifen-treated breast cancer patients: findings from NSABP B-14, *Journal of the National Cancer Institute* **86**, 527–537.

[22] Fisher, B., Anderson, S., Redmond, C., Wolmark, N., Wickerham, L. & Cronin, W. (1995). Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer, *New England Journal of Medicine* **333**, 1456–1461.

[23] Fisher, B., Dignam, J., Mamounas, E.P., Costantino, J.P., Wickerham, D.L., Redmond, C., Wolmark, N., Dimitrov, N.V., Bowman, D.M., Glass, A.G., Atkins, J.N., Abramson, N., Sutherland, C.M., Aron, B.S. & Margolese, R.G. (1996). Sequential methotrexate and fluorouracil for the treatment of node-negative breast cancer patients with estrogen receptor-negative tumors: eight-year results from National Surgical Adjuvant Breast and Bowel Project (NSABP) B-13 and first report of findings from NSABP B-19 comparing methotrexate and fluorouracil with conventional cyclophosphamide, methotrexate, and fluorouracil, *Journal of Clinical Oncology* **14**, 1982–1992.

[24] Fisher, B., Dignam, J., Bryant, J., DeCillis, A., Wickerham, D.L., Wolmark, N., Costantino, J., Redmond, C., Fisher, E.R., Bowman, D.M., Deschenes, L., Dimitrov, N.V., Margolese, R.G., Robidoux, A., Shibata, H., Terz, J., Paterson, A.H.G., Feldman, M.I., Farrar, W., Evans, J. & Lickley, H.L. (1996). Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors, *Journal of the National Cancer Institute* **88**, 1529–1542.

[25] Fisher, B., Anderson, S., Wickerham, D.L., DeCillis, A., Dimitrov, N., Mamounas, E., Wolmark, N., Pugh, R., Atkins, J.N., Meyers, F.J., Abramson, N., Wolter, J., Bornstein, R.S., Levy, L., Romond, E.H., Caggiano, V., Grimaldi, M., Jochimsen, P. & Deckers, P. (1997).

Increased intensification and total dose of cyclophosphamide in a doxorubicin-cyclophosphamide regimen for the treatment of primary breast cancer: findings from National Surgical Adjuvant Breast and Bowel Project B-22, *Journal of Clinical Oncology* **15**, 1858–1869.

[26] Fisher, B., Dignam, J., Wolmark, N., DeCillis, A., Emir, B., Wickerham, D.L., Bryant, J., Dimitrov, N.V., Abramson, N., Atkins, J.N., Shibata, H., Deschenes, L. & Margolese, R.G. (1997). Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer, *Journal of the National Cancer Institute* **89**, 1673–1682.

[27] Fisher, B., Brown, A., Mamounas, E., Wieand, S., Robidoux, A., Margolese, G., Cruz, A.B., Fisher, E.R., Wickerham, D.L., Wolmark, N., DeCillis, A., Hoehn, J.L., Lees, A.W. & Dimitrov, N.V. (1997). Effect of preoperative chemotherapy on local-regional disease in women with operable breast cancer: findings from National Surgical Adjuvant Breast and Bowel Project B-18, *Journal of Clinical Oncology* **15**, 2483–2493.

[28] Fisher, B., Bryant, J., Wolmark, N., Mamounas, E., Brown, A., Fisher, E.R., Wickerham, D.L., Begovic, M., DeCillis, A., Robidoux, A., Margolese, R.G., Cruz, A.B. Jr, Hoehn, J.L., Lees, A.W., Dimitrov, N.V. & Bear, H.D. (1998). Effect of preoperative chemotherapy on the outcome of women with operable breast cancer, *Journal of Clinical Oncology* **16**, 2672–2685.

[29] Fisher, B., Costantino, J.P., Wickerham, D.L., Redmond, C., Kavanah, M., Cronin, W.M., Vogel, V., Robidoux, A., Dimitrov, N.V., Atkins, J., Daly, M., Wieand, S., Tan-Chiu, E., Ford, L., Wolmark, N. & other National Surgical Adjuvant Breast and Bowel Project Investigators (1998). Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 study, *Journal of the National Cancer Institute* **90**, 1371–1388.

[30] Fisher, B., Dignam, J., Wolmark, N., Wickerham, D.L., Fisher, E.R., Mamounas, E., Smith, R., Begovic, M., Dimitrov, N.V., Margolese, R.G., Kardinal, C.G., Kavanah, M.T., Fehrenbacher, L. & Oishi, R.H. (1999). Tamoxifen in treatment of intraductal breast cancer: National Surgical Adjuvant Breast and Bowel Project B-24 randomised controlled trial, *Lancet* **353**, 1993–2000.

[31] Ganz, P.A., Day, R., Ware, J.E. Jr, Redmond, C. & Fisher, B. (1995). Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project breast cancer prevention trial, *Journal of the National Cancer Institute* **87**, 1372–1382.

[32] Gorin, M.B., Day, R., Costantino, J.P., Fisher, B., Redmond, C.K., Wickerham, L., Gomolin, J.E., Margolese, R.G., Mathen, M.K., Bowman, D.M., Kaufman, D.I., Dimitrov, N.V., Singerman, L.J., Bornstein, R., Wolmark, N. & Kaufman, D. (1998). Long-term tamoxifen citrate use and potential ocular toxicity, *American Journal of Ophthalmology* **125**, 493–501.

[33] Progress Report of the National Surgical Adjuvant Breast and Bowel Project (1998). 17–24.

[34] Ravdin, R.G., Lewison, E.F., Slack, N.H., Dao, T.L., Gardner, B., State, D. & Fisher, B. (1970). Results of a clinical trial concerning the worth of prophylactic oophorectomy for breast carcinoma, *Surgery, Gynecology and Obstetrics* **131**, 1055–1064.

[35] Skipper, H.E. (1971). Kinetics of mammary tumor cell growth and implications for therapy, *Cancer* **28**, 1479–1499.

[36] Skipper, H.E. & Schabel, F.M., Jr (1973). Quantitative and cytokinetic studies in experimental tumor models, in *Cancer Medicine*, J.F. Holland & E. Frei III, eds. Lea & Febiger, Philadelphia, pp. 629–650.

BERNARD FISHER

Natural History Study of Prognosis

Prognosis is defined as a forecast as to the probable outcome of an attack of disease; the prospect as to recovery from a disease as indicated by the nature and symptoms of the case [10]. In the context of research, it is a prediction of the future course of a disease following its onset. The natural history of a disease is the evolution of a disease in the absence of medical intervention [14]. Today, however, this definition may be too narrow for many diseases in which there are widely accepted treatments that are offered universally. For example, if a patient is diagnosed with severe diabetes, insulin may be prescribed. The description of disease course would include the influence of the insulin, since few patients would continue without medical intervention. Accordingly, we use the term *clinical course* to describe the natural history of a disease that has been affected by medical intervention. In this discussion the use of the term *natural history* may also, include clinical course.

The term prognosis implies an outcome. Natural history studies of prognosis are studies in which the course of a disease is observed over time as outcomes such as death, relapse of a tumor, or acquisition of AIDS following HIV-1 infection. **Rates,** such as percent of patients recurring within five years or a 10 year survival rate, are used to summarize these outcomes [16]. While a rate is a simple description of the outcome in a natural history study of prognosis, it can sometimes mask different processes. For example, in a study of cutaneous malignant melanoma, approximately one-quarter of the Stage I patients had a prophylactic lymph node dissection. The authors found that while this procedure was associated with an early survival advantage, this advantage was not sustained over a longer follow-up period. Thus, a conclusion based on five years of follow-up, as opposed to 10 years, would have been erroneous [33].

Sometimes we express outcomes as technical results, e.g. percent change in tumor size, or whether coronary dilatations exceed the diameter of normal adjacent segments by 1.5 times [2], or specific values of laboratory markers like CD4 T-cell counts [31]. While a laboratory marker is a more easily measured outcome than an outcome such as "pain", and one that lends itself more readily to rigorous statistical analysis, either type is relevant only if it is truly prognostic and related to the clinical outcomes.

It is important to distinguish between the concepts of **risk** and prognosis. Risk is related to the likelihood of getting a disease, while prognosis is about the progress and outcome of the disease, once it has been acquired. Similarly, risk factors are those characteristics that increase the probability of getting the disease, while **prognostic factors** are those variables associated with different possible outcomes of the disease, or time to an outcome, once the disease is present. Of course, some variables may be both a risk factor and a prognostic factor for the same disease. Variables that have both risk and prognostic roles may work in opposite directions. For example, men are more likely than women to develop coronary disease in middle age (risk), and also are more likely to die from it if they get it (prognosis) [29]. Conversely, younger women are less at risk of getting breast cancer, but once diagnosed with breast cancer, young age is associated with worse prognosis [24].

This article discusses design considerations for studies of prognosis, as well as analytical approaches.

Design Considerations

Design considerations in natural history studies of prognosis include the definition and selection of the population and samples to be studied, the definition of time zero, the definitions of outcomes, and potential sources of **bias**. We described definitions of outcomes. The following describes these other design issues.

Cohort Studies

The usual design for a natural history study of prognosis would be a **cohort study** [25]. In a cohort study we follow a subject population over time, during which periodic assessments or observations are scheduled and the occurrence of predefined outcomes in the cohort is recorded. The cohort can be prospectively or retrospectively sampled. A retrospective cohort might include all the patients presenting at a hospital within a period of time with a specific diagnosis or set of symptoms (*see* **Cohort Study, Historical**). We assemble the cohort via a medical record review. For example, in a retrospective study

of the natural history of ductal adenocarcinoma of the pancreas in patients under 40 years old, all patients seen at the Mayo Clinic from 1970 to 1985 were reviewed [21].

A prospective cohort study utilizes an inception cohort, i.e. a group of subjects assembled near the onset ("inception") of disease. This cohort typically cannot be gathered at a single point in time because the screening and recruitment process requires some time. Studies of this type typically are designed to have a recruitment or accrual period when the cohort is identified, and a specified follow-up period after the last subject has been entered into the cohort. Thus, the length of follow-up may vary by patient, depending on their early vs. late recruitment. In a study of the long-term prognosis of lupus nephritis, an inception cohort of 87 patients at Montreal's General Hospital and Children's Hospital seen between 1967 and 1983 was studied. The inception point was the day of first renal biopsy [17]. While prospective studies clearly are less susceptible to **selection bias, retrospective studies** provide preliminary answers more quickly and at a relatively low cost. Researchers use retrospective cohort study results to plan a prospective study more effectively.

Bias

There are many types of biases that can occur in natural history studies of prognosis and that effect the conclusions. The most frequent are sampling or selection bias, **length bias,** and dropout bias.

Sampling or selection bias results, also known as assembly bias, results from the method by which patients are sampled or selected to be in the study [25]. This type of bias may occur, for example, when the selected cases are not representative of the diseased population, or when the groups of patients differ in terms of the extent of disease under study. For example, in a study of infective endocarditis, it was shown that the clinical spectrum of the disease could be distorted by referral patterns when a community cohort in Minnesota was compared with a Mayo Clinic cohort [39]. In general, natural history studies of prognosis based only on patients in academic institutions (or tertiary care centers) tend to suggest poorer prognosis, unless the disease is treated only at academic institutions. This is because subjects may be referred to these centers only after treatment efforts at the community level have failed. In cancer studies conducted at tertiary care cancer research centers, which tend to receive higher proportions of more advanced cases, the natural history prognostic estimates appear worse than those derived from a representative sample of patients at all stages of the disease. This is analogous to the susceptibility bias that is present in cohort studies performed to estimate the risk of acquiring disease based on exposure [14].

Length bias occurs in several ways. An example is the natural history estimate of the **latent period** from HIV infection to the onset of AIDS. Early studies of HIV-infected cohorts estimated that this interval was at least four years, on the basis of **surveillance** data from the **Centers for Disease Control** [7]. Researchers continuously updated these estimates to a current estimate of eight years or more on the basis of cohorts from the US, Canada, Australia, and the Netherlands [40]. Patients with shorter latency presented earlier with the disease and affected the estimate. Another form of length bias combined with selection bias occurs when a cohort is not assembled at the inception of disease, but is selected on the basis of availability of patients (e.g. accepting patients at different stages of the disease), and information is collected after the inception of each subject's disease. In this case the cohort includes only those subjects who are available for study, and thus may not include patients with the same time disease-inception in whom a failure event (death, relapse) has already occurred. This introduces length bias into the estimates of time to progression within the natural history of disease if the analysis does not take this into account [6].

In dropout bias the patients who drop out of a study or are lost to follow-up may be different from the patients remaining in the study. Since the study conclusions are based only on available data, if the dropout pattern is not random, then this introduces a form of selection bias. In addition to aggressive retention and follow-up of subjects, a carefully designed observation schedule is helpful in assessing the magnitude of the problem due to dropout (*see* **Bias, Overview**).

Definition of Time Zero

For a cohort study to be meaningful, a well-defined time zero should be defined for the disease of interest. We define time zero as the time of onset of the first or specified symptoms or the date of diagnosis. For

diseases in which **screening** is routinely performed, such as certain screening tests for newborns, this definition is easy. In other cases, such as breast cancer, this determination is less clear; the point of inception might be based on a routine mammogram, on the discovery of a lump during self-examination test, or the onset of symptoms. The importance of an unambiguous definition for time zero in a natural history study of prognosis cannot be overemphasized. The goal of describing the time course until outcome events occur will be thwarted in the absence of an unambiguous starting point.

In summary, the design of natural history studies of prognosis must include a carefully defined cohort, appropriate selection and assembly procedures, and clearly defined starting points (time zero), follow-up procedures, and unambiguous definitions of outcome.

Analysis Approaches

Dichotomous Outcomes and Logistic Regression

The choice of appropriate analysis tools depends on the type of outcome that the study is describing. In the first step of analysis in a natural history study of prognosis we estimate the rate or percentage of those who had a given outcome, e.g. the percentage who died. At the next step we relate the percentage outcome to prognostic factors. In the simplest case there is a single, dichotomous prognostic factor. In a study of severe cardiac events such as death and myocardial infarction, researchers assessed the prognostic value of dipyridamole first-pass radionuclide ventriculography using a **Fisher's exact test** [3]. A **chi-square** is also typically used.

When there are several prognostic variables, epidemiologists typically use logistic regression to analyze further the natural history of prognosis. In **logistic regression** we define a dichotomous outcome variable Y, such as alive at the end of three years ($Y = 0$) or died within the first three years since time zero ($Y = 1$). We model the outcome variable as a function of the prognostic variables or **explanatory variables** [19]. The prognostic variables can be dichotomous, such as sex, continuous, such as age, or categorical, such as race. In a cohort study examining the prognostic value of ultrasound findings at birth in predicting disabling cerebral palsy, both dichotomous variables, such as presence of ultrasound findings, and continuous variables, such as gestation age at birth, were explored [38]. In this study the logistic regression modeling showed that certain ultrasound findings (parenchymal echodensities/lucencies) were strongly prognostic (**odds ratio** of 15) of disabling cerebral palsy, even after accounting for well-known prognostic factors such as gestation age and birth weight.

The relationship between variables can also be studied using a regression tree (*see* **Tree-structured Statistical Methods**). In a study on the impact of disease activity on long-term prognosis of lupus nephritis, the researchers used a regression tree technique which gave easily interpretable and useful results [17].

Survival Analysis

By far the most popular analysis tool in natural history studies of prognosis is **survival analysis** [9]. In this approach we assemble a cohort at a particular time zero in the course of disease, and follow the subjects until either the outcome occurs (called "failure"), or a prespecified point for the end of the study is reached. This prespecified time point can be a total follow-up time or a total number of observed failures. A description of the natural history of the group is given by a **Kaplan–Meier** curve or a product–limit estimator (*see* **Life Table**) [22]. This curve gives the estimated survival experience of the study subjects as a function of time. Sometimes, patients exit the study before the failure event occurred, owing to loss to follow-up, patient withdrawal from the study, or other reasons. The observations on these patients are defined as **censored**, and the calculation adjusts for these. The estimation of the survival probabilities at each point in time provides a description of the overall prognosis of the disease in question. The outcome variable is, therefore, the time until an event occurs.

In the next step, specific prognostic factors for the disease are examined. For example, in a study of risk **stratification** and prognosis of patients with recent onset of angina, the researchers examined whether a positive thallium stress test and the number of clinical risk characteristics as well as the number of involved arteries were prognostic predictors of future medical events [4]. The analytical tools used to examine the effect of a single prognostic factor on survival are the Mantel–Cox test, **logrank test**, the Wilcoxon–Gehan test or the Peto test [28]. These tests examine whether a dichotomous or categorical

factor is significantly associated with better or worse prognosis of the disease, depending on the value of this factor. The tests differ primarily in the weight that they place on observations at the early vs. later part of the Kaplan–Meier curve.

When there is interest in the joint effect of several factors on survival or when a factor is continuous and there is no desire to categorize it (and hence lose some information), the most popular analytical tool is the **Cox regression model**, which is **semiparametric**. The Cox model is a regression model proposed for survival distributions in which the time to failure is related to the prognostic factors and it allows the introduction of the censored observations into the model [8].

The typical use of the Cox model is in the context of a **proportional hazards** model. A proportional hazards model possesses the property that different individuals have **hazards** that are proportional to one another over time, depending on the values of the fixed **covariates**. A typical example of the use of a Cox model is in the study of unknown primary carcinomas [1]. The fixed covariates used were pathology subtype (adenocarcinoma, neuroendocrine carcinoma), number of sites (1, 2, 3+), gender, and involved organ site.

Other covariates included in a Cox model may be time-varying covariates (*see* **Time-dependent Covariate**). When the covariates are time-varying, the hazards are no longer proportional with respect to the time-varying covariate, since the value of the covariate may change over time. Thus, there is no meaning to the concept of proportional hazards. Time-varying covariates may be very important in understanding the natural history of prognosis. CD4 T-cell counts are typical time-varying covariates for understanding the natural history of HIV infection [15].

In addition to the Cox nonparametric approach to survival analysis, there are **parametric survival models**, including, in particular, **accelerated survival models** that serve well for natural history studies of prognosis. An example is a study based on the **Framingham** Heart Study data which used a **Weibull distribution** based accelerated failure time model to model the time to angina pectoris when there is **interval censoring** [36].

The development of martingale theory and **counting processes** gave researchers many more opportunities to expand and allow for more complex survival models [13]. These models allow for **multivariate survival** outcomes and flexibility in modeling, resulting in better understanding of the natural history of complex diseases such as **AIDS** [5]. In this study both seroconversion time and progression to AIDS were studied. The counting processes models also allow for approaches such as modeling time-varying effects using **spline functions** to examine, for example, whether tumor necrosis is a prognostic factor for early recurrence and death in lymph-node-positive breast cancer [16]. The issue here was that the effect of tumor necrosis appeared to be changing with time, and the spline modeling enabled the researchers to provide a better description of the prognosis of these patients than more traditional survival analysis approaches.

Longitudinal Data Analysis

Another way to study the natural history of prognosis in terms of data analysis is to use models for the analysis of **longitudinal data**. This approach is useful especially when the interest is not necessarily in the presence/absence of an outcome, but rather in the changes in a parameter over time. The parameter may be the lung function of cystic fibrosis patients, or a molecular biomarker called prostate specific antigen (PSA) in the study of the natural history of prostate disease [37]. The longitudinal model in this example estimates the natural history of the biomarker both as a function of time (or age) and as a function of other prognostic factors, such as age at diagnosis, or presence of cancer (as a time-varying parameter). In a study of this type the subjects are measured repeatedly over time. In natural history studies of prognosis, which are based on a cohort of subjects, there are frequently unequal intervals between follow-up visits of different patients and unequal numbers of measurements for different patients. In this situation an appropriate approach is a mixed effects model or **random effects** model [26, 27]. This approach models the change in a parameter across time, using the repeated measurements and modeling the **correlation** between the measurements of an individual. For example, in a study of neurodevelopment in children who were perinatally infected with HIV, the number of measurements ranged from two to nine and the timing was different between children [35].

The mixed effects model assumes a linear relationship between the outcome variable and

the explanatory variables. Moreover, it assumes a **normal distribution** for the error of the so-called individual random effects. The **generalized estimating equations** approach (GEE) relaxes these assumptions and generalizes the model to dichotomous and other outcome variables [30]. Both of these methods have received considerable attention in the last decade, with much research devoted to expanding these methods. As the methods are developed and become more accessible, they may be better suited for application in natural history studies. In a study relating the outcome of poor visual acuity to the genetic type of retinas pigmentia, the GEE approach was used to combine correlated data of two eyes per individual [23] (*see* **Correlated Binary Data**).

Combining Data from Several Studies

There are several indirect approaches to natural history studies of prognosis that are based on existing data and previous studies. One approach takes the raw data from several studies with similarly defined populations, and combines the data for an overall natural history study. In a study addressing the prognostic variables for survival in hepatocellular carcinoma, data on patients from three consecutive clinical trials conducted by the Eastern Cooperative Oncology Group were combined and a Cox proportional hazards model fit to the combined data set [11]. The study is a natural history study because none of the therapeutic approaches worked, and ignoring the treatment or allowing the treatment to be a covariate did not change results. The definition of the patient population was identical in all studies, and thus the homogeneity of the combined study population was established and generalizability was possible (*see* **Validity and Generalizability in Epidemiologic Studies**). Combining the studies increased substantially the **power** to detect significance of prognostic factors for survival.

Another approach to combining data from several studies is **meta-analysis** [20]. In meta-analysis one combines the results from several studies, either by summarizing all of the studies that meet certain criteria, such as in a study of inconsistent prognoses of post-acute myocardial infarction [32], or by actual data synthesis. An example of the latter is a study of prognosis and outcomes of patients with community-acquired pneumonia where **odds ratios**, rate differences, and **confidence intervals** were calculated as well as overall mortality and the relationship with prognostic factors [12].

Multiple Outcomes

Sometimes a natural history study can have multiple outcomes. This can be true for all types of analysis approaches. It can occur in the context of logistic regression where there are two or more dichotomous outcomes that can be observed, or multivariate survival variables that are observed. One example is time to metastases of breast cancer at various sites as multiple outcomes. This problem can be dealt with both as a **competing risks** problem in the context of a Cox model, or in a generalization of joint outcomes [41].

Conclusion

The choice of analytical approach depends on many factors related to the design of the study and the definition of outcomes. Sometimes more than one approach is applied to the same data and conclusions are compared [34]. For example, consistency may be a problem when there are **missing data** [18]. The use of more than one analytical approach helps to understand better the natural history of prognosis and prevents erroneous conclusions.

References

[1] Abbruzzese, J.L., Abbruzzese, M.C., Hess, K.R., Raber, M.N., Lenzi, R. & Frost, P. (1994). Unknown primary carcinoma: natural history and prognostic factors in 657 consecutive patients, *Journal of Clinical Oncology* **12**, 1272–1280.

[2] Bal, E.T., Plokker, H.W.T., van dem Berg, E.M.J., Ernst, S.M.P.G., Mast, E.G., Gin, R.M.T.J.G. & Ascoop, C.A.P.L. (1991). Predictability and prognosis of PTCA-induced coronary artery aneurysms, *Catheterization and Cardiovascular Diagnosis* **22**, 85–88.

[3] Bassevich, R., Zafrir, N., Sulkes, J. & Lubin, E. (1994). Dipyridamole first-pass radionuclide ventriculography: prediction of future cardiac events, *American Journal of Cardiology* **74**, 1229–1232.

[4] Castaner, A., Roig, E., Serra, A., De Flores, T., Magrina, J., Azqueta, M., Sanz, G. & Betriu, A. (1990). Risk stratification and prognosis of patients with recent onset angina, *European Heart Journal* **11**, 868–875.

[5] Chiarotti, F., Palombi, M., Schinaia, N., Ghirardini, A. & Bellocco, R. (1994). Median time from seroconversion to AIDS in Italian HIV-positive haemophiliacs: different parametric estimates, *Statistics in Medicine* **13**, 163–175.

[6] Cnaan, A. & Ryan, L.M. (1989). Survival analysis in natural history studies of disease, *Statistics in Medicine* **8**, 1255–1268.

[7] *Confronting AIDS, Directions for Public Health, Health Care, and Research (1986). National Academy of Sciences, Public Health Reports, 101, National Academy Press, Washington, pp. 326–327.*

[8] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[9] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[10] *Dorland's Illustrated Medical Dictionary* (1994). Saunders, Philadelphia.

[11] Falkson, G., Cnaan, A., Schutt, A., Ryan, L.M. & Falkson, H.C. (1988). Prognostic factors in hepatocellular carcinoma, *Cancer Research* **48**, 7314–7318.

[12] Fine, M.J., Smith, M.A., Carson, C.A., Mutha, S.S., Sankey, S.S., Weissfeld, L.A. & Kapoor, W.N. (1996). Prognosis and outcomes of patients with community-acquired pneumonia, a meta-analysis, *Journal of the American Medical Association* **275**, 134–141.

[13] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[14] Fletcher, R.H., Fletcher, S.W. & Wagner, E.H. (1988). *Clinical Epidemiology: The Essentials*. Williams & Wilkins, Baltimore.

[15] Fore, A.J., Kramer, A., Grund, B. & Hannan, P. (1994). Segmented Cox models can distinguish short and long term AIDS progression markers, *International Conference on AIDS*, Abstract No. PCO256. University of Minnesota, Minneapolis, MN.

[16] Gilchrist, K.W., Gray, R., Fowble, B., Tormey, D.C. & Taylor IV, S.G. (1993). Tumor necrosis is a prognostic predictor for early recurrence and death in lymph node-positive breast cancer: a 10-year follow-up study of 728 Eastern Cooperative Oncology Group patients, *Journal of Clinical Oncology* **11**, 1929–1935.

[17] Goulet, J.-R., MacKenzie, T., Levinton, C., Hayslett, J.P., Ciampi, A. & Esdaile, J.M. (1993). The longterm prognosis of lupus nephritis: the impact of disease activity, *Journal of Rheumatology* **20**, 59–65.

[18] Greenland, S. & Finkle, W.D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analysis, *American Journal of Epidemiology* **142**, 1255–1264.

[19] Hosmer, D.W., Jr & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[20] Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis, Correcting Error and Bias in Research Findings*. Sage, Newbury Park.

[21] Ivy, E.J. Sarr, M.G. & Reiman, H.M. (1990). Nonendocrine cancer of the pancreas in patients under age forty years, *Surgery* **108**, 481–487.

[22] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation for incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[23] Katz, J., Zeger, S. & Liang, K.-Y. (1994). Appropriate statistical methods to account for similarities in binary outcomes between fellow eyes, *Investigative Ophthalmology and Visual Science* **35**, 2461–2465.

[24] Kelsey, J.L. & Berkowitz, G.S. (1988). Breast cancer epidemiology, *Cancer Research* **48**, 5615–5623.

[25] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research, Principles and Quantitative Methods*. Van Nostrand Reinhold, New York.

[26] Laird, N.M. & Ware, J.H. (1982). Random effects models for longitudinal data: an overview of recent results, *Biometrics* **38**, 963–974.

[27] Laird, N.M., Donnelly, C. & Ware, J.H. (1992). Longitudinal studies with continuous responses, *Statistical Methods in Medical Research* **1**, 225–247.

[28] Lee, E.T. (1992). *Statistical Methods for Survival Data Analysis*. Wiley, New York.

[29] Lerner, D.J. & Kannel, W.B. (1986). Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population, *American Heart Journal* **111**, 383–390.

[30] Liang, K.Y. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[31] Lin, D.Y., Fischl, M.A. & Schoenfeld, D.A. (1993). Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials, *Statistics in Medicine* **12**, 835–842.

[32] Marx, B.E. & Feinstein, A.R. (1995). Methodologic sources of inconsistent prognoses for post-acute myocardial infarction, *American Journal of Medicine* **98**, 537–550.

[33] Meyskens, F.L., Jr, Berdeaux, D.H., Parks, B., Tong, T., Loescher, L. & Moon, T.E. (1988). Cutaneous malignant melanoma (Arizona Cancer Center Experience), *Cancer* **62**, 1207–1214.

[34] Moriguchi, S., Hayashi, Nose, Y., Maehara, Y., Korenaga, D. & Sugimachi, K. (1993). A comparison of the logistic regression and the Cox proportional hazard models in retrospective studies on the prognosis of patients with gastric cancer, *Journal of Surgical Oncology* **52**, 9–13.

[35] Nozyce, M., Hittelman, J., Muenz, L., Durako, S.J., Fischer, M.L. & Willoughby, A. (1994). Effect of perinatally acquired human immunodeficiency virus infection on neurodevelopment in children during the first two years of life, *Pediatrics* **94**, 883–891.

[36] Odell, P.M., Anderson, K.M. & D'Agostino, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model, *Biometrics* **48**, 951–959.

[37] Pearson, J.D., Morrell, C.H., Landis, P.K., Carter, H.B. & Brant, L.J. (1994). Mixed-effects regression models for studying the natural history of prostate disease, *Statistics in Medicine* **13**, 587–601.

[38] Pinto-Martin, J., Riolo, S., Cnaan, A., Holzman, C., Susser, M.W. & Paneth, N. (1995). Cranial ultrasound prediction of disabling and nondisabling cerebral palsy at age two in a low birth weight population, *Pediatrics* **95**, 249–254.

[39] Steckelberg, J.M., Melton, L.J., Ilstrup, D.M., Rouse, M.S. & Wilson, W.R. (1990). Influence of referral bias on the apparent clinical spectrum of infective endocarditis, *American Journal of Medicine* **88**, 582–588.

[40] Veugelers, P.J., Page, K.A., Tindall, B., Schechter, M.T., Moss, A.R., Winkelstein, W.W., Cooper, D.A., Craib, J.P., Charlebois, E., Coutinho, R.A. & van Griensven, K.J.P. (1994). Determinants of HIV disease progression among homosexual men registered in the tricontinental seroconversion study, *American Journal of Epidemiology* **140**, 747–758.

[41] Zedeler, K., Keiding, N. & Kamby, C. (1992). Differential influence of prognostic factors on the occurrence of metastases at various anatomical sites in human breast cancer, *Statistics in Medicine* **11**, 281–294.

(*See also* **Predictive Modeling of Prognosis**; **Time-varying Treatment Effect**).

AVITAL CNAAN

Negative Binomial Distribution

The negative binomial distribution has been used in biostatistical literature since Student [31]. He proposed that the error in counting red blood cells using the hemocytometric camera follows that distribution. The method consists of stretching a few drops of blood on a special slide that has a camera with a grid superimposed. After some chemical treatment to preserve the material and to color the blood cells, the specimen is read on a microscope. The reader counts the number of red blood cells in a fixed number of quadrats (say 100). From that, given a proportionality constant, the total number of red blood cells of the donor is estimated.

This first application is emblematic, because the negative binomial was proposed as an alternative to the **Poisson distribution** for modeling counts when data show a certain degree of extra-Poisson variation (*see* **Overdispersion**). In this application, this extra variation is due to the fact that the intensity rate is not homogeneous, since the blood is not uniformly spread on the slide.

The Negative Binomial Distribution

Definition

The distribution is obtained from the expansion of $(Q - P)^{-R}$, where $Q - P = 1$ and $P > 0$, similarly to a derivation for the **binomial distribution**. The probability mass function equals

$$\Pr(y) = \frac{(R + y - 1)!}{y!(R - 1)!} \left(\frac{P}{Q}\right)^y \left(1 - \frac{P}{Q}\right)^R \quad (1)$$

and is defined for nonnegative integer values y. The value R need not be an integer, but when it is, this is known as the *Pascal distribution*. The mean is RP and the variance is RPQ. Therefore, the variance exceeds the mean, and the distribution shows greater tail probabilities than the Poisson. The probabilities can be calculated from the binomial distribution, observing that [16]

$$\frac{(R + y - 1)!}{(y)!(R - 1)!} \left(\frac{P}{Q}\right)^y \left(1 - \frac{P}{Q}\right)^R = \frac{R}{R + y} \frac{(R + y)!}{(y)!(R)!} \left(\frac{P}{Q}\right)^y \left(1 - \frac{P}{Q}\right)^R. \quad (2)$$

Example The observed number of accidents at work during a five-week period are reported in Table 1; the fitted negative binomial and Poisson expected counts [17] are also shown. The negative binomial distribution has higher probability for the zero count and has a longer right-hand tail than the Poisson distribution with the same expected value (*see* **Accident Proneness**).

Derivation A

The negative binomial distribution can be derived in two ways.

Suppose that we are interested in the number of trials (say N) until R successes have occurred, when the probability of success is p for each trial. Then the probability distribution of N is

$$\Pr(N = y + R) = \frac{(R + y - 1)!}{y!(R - 1)!} p^R (1 - p)^y \quad (3)$$

for the reparameterization $P = (1 - p)/p$. This distribution is also called the *binomial waiting-time distribution* and, in the special case $R = 1$, it is the **geometric distribution**, the discrete analogue of the exponential distribution (*see* **Parametric Models in Survival Analysis**).

Table 1 Negative binomial and Poisson expected frequencies for observed counts of work accidents during a five-week period

y_i	f_i	Negative binomial	Poisson
0	447	442	406
1	132	140	189
2	42	45	45
3	21	14	7
4	3	5	1
≥5	2	2	0.1

Reproduced from [17] by permission of the Royal Statistical Society

Example In the *surveillance* of rare health events, such as congenital birth defects, the probability that a birth has a malformed baby can be modeled as a function of the number of nonmalformed babies born since the previous malformation occurred (negative binomial with $R = 1$). This approach is used in the SETS scheme [7].

Derivation B

Suppose that a heterogeneous Poisson process generates the observed counts. This could be modeled as a mixture of Poisson distributions, the expected value μ of which is the realization of a **gamma distribution** with density

$$\Pr(\mu) = [\nu^\kappa \Gamma(\kappa)]^{-1} \mu^{\kappa-1} \exp\left(-\frac{\mu}{\nu}\right), \qquad (4)$$

with $\kappa > 0$ and $\nu > 0$. Then

$$\begin{aligned}\Pr(y) &= \int_0^\infty \left[\frac{\mu^y \exp^{-\mu}}{y!}\right] [\nu^\kappa \Gamma(\kappa)]^{-1} \mu^{\kappa-1} \\ &\quad \times \exp\left(-\frac{\mu}{\nu}\right) \mathrm{d}\mu \\ &= [\nu^\kappa \Gamma(\kappa) y!]^{-1} \int_0^\infty \{\mu^{y+\kappa-1} \\ &\quad \times \exp[-\mu(\nu^{-1}+1)] \,\mathrm{d}\mu\} \\ &= \frac{(y+\kappa-1)!}{y!(\kappa-1)!} \left(\frac{\nu}{\nu+1}\right)^y \left(\frac{1}{\nu+1}\right)^\kappa, \qquad (5)\end{aligned}$$

which is negative binomial with parameters $\{\kappa, \nu\}$ (where $R = \kappa$ and $P = \nu$). This mixture could arise (factorizing $\mu = \theta t$) either by varying the *rate* θ of occurrence of the events for constant time span (t) or the population at risk for each observation, or by varying t. We cannot distinguish between these two mechanisms from the data, and usually one of the two sources of heterogeneity is controlled by the study design.

Example: Varying rates. The observed regional variation of the rate of surgical interventions can be split into a component reflecting the underlying rate variability among areas, service availability, physicians practice styles, etc. which can be modeled by the gamma density, and a random Poisson component [13].

Example: Varying population-time. In counting the number of larvae caught in a trap, the total catch is the sum of the larvae that emerged from distinct egg masses, the number M of which is unknown. The negative binomial distribution arises as a Poisson stopped sum of a logarithmic series variable [30].

Estimation of Parameters

Method of Moments

The **method of moments** [19] is the simplest way to obtain estimates of the parameters of the negative binomial. The method equates the population parameters to their sample counterparts (the mean to the sample mean and the variance to the sample variance) and then solves for the parameter estimates

$$\hat{P} = \frac{s^2}{\overline{y} - 1}, \qquad (6)$$

$$\hat{R} = \frac{\overline{y}^2}{(s^2 - \overline{y})}, \qquad (7)$$

where $\overline{y}$ and s^2 are the sample mean and variance of the observed counts. This method could give negative values if $s^2 < \overline{y}$. A different approach uses the observed proportion of zero counts, p_0 [3]:

$$\hat{R}\hat{P} = \overline{y}, \qquad (8)$$

$$p_0 = (1 + \hat{P})^{-\hat{R}}. \qquad (9)$$

This method can be used if $\overline{y} > -\log p_0$.

Iterative methods of moments were used by Scheaffer & Leavenworth [28] and Clayton & Kaldor [9]. The first used a normal approximation for a transformation introduced by Anscombe [3] based on $\sinh^{-1}$ to develop a new formula based on the variance of the transformed variable. The second, given initial estimates, is based on estimates of the parameters of the Poisson distributions. From their mean and variance, one updates the estimates of the negative binomial parameters, and repeats until convergence. A further refinement is to equate the Pearson χ^2 to its expected value, instead of using the variance (see also [23]). In most circumstances the method of moments works well. However, a warning has been posed when the mean is small and the sample size does not exceed 20 [8].

Maximum Likelihood

Maximum likelihood (ML) estimation was discussed by Fisher [15], who gave formulas for the case of moderately small sample sizes. The ML estimator is indeterminate when $s^2 < \overline{y}$, and Anscombe [4] comments that the ML estimator does not have a distribution, since there is a nonzero probability of observing a data set with sample variance less than the sample mean. A useful reparameterization is $\alpha = 1/\kappa$, which yields the Poisson distribution as $\alpha \to 0$; the parameter α is called the dispersion parameter, because the negative binomial variance is $V(y) = \mu + \alpha\mu^2$ [27]. To facilitate the specification of a linear model for the mean value, the other parameter is assumed to be the mean $\mu = \nu/\alpha$. The negative binomial density becomes

$$P(y) = \frac{(y + \alpha^{-1} + 1)!}{y!(\alpha^{-1})!} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^y \left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}. \tag{10}$$

When μ is assumed as the parameter of the Poisson likelihood the gamma prior has only one parameter. Maximum likelihood estimates can be obtained using the Newton–Raphson method [21] (*see* **Optimization and Nonlinear Equations**). This method produces biased results in small sample sizes, and the distribution of $\hat{\alpha}$ is discrete in those cases. It is recommended, therefore, to utilize other methods, such as those listed below.

Conditional Maximum Likelihood

The conditional maximum likelihood estimator for the dispersion parameter of the negative binomial distribution was first proposed by Kalbfleisch & Sprott [20] and evaluated by Anraku & Yanagimoto [2]. The negative binomial likelihood can be factored into two terms: the first term allows the estimation of μ, and the second term allows the estimation of the dispersion parameter given μ. The relationship between the conditional (CL) and the unconditional ML estimator (UL) is

$$\frac{1}{2}\frac{\overline{y}}{1+\alpha\overline{y}} < CL - UL < \frac{\overline{y}}{1+\alpha\overline{y}}. \tag{11}$$

The UL estimator is smaller than the CL, for any sample. Simulation studies seems to support CL against ML. It should be noticed that both estimators are defined only for $s^2 > \overline{y}$.

Maximum Extended Quasi-likelihood

Extended quasi-likelihood was proposed by Nelder & Pregibon [26] in the context of the **generalized linear model**. Clark & Perry [8] considered it for the estimation of the dispersion parameter of the negative binomial distribution. One of the merits of this approach (see [24]) is the ability to handle overdispersed as well as *underdispersed* data (*see* **Overdispersion**). The extended quasi-likelihood for the negative binomial distribution is

$$\begin{aligned} l(\mu_i, \alpha) = {} & y_i \ln\left(\frac{\mu_i}{y_i}\right) - \frac{1+\alpha y_i}{\alpha}\ln\left(\frac{1+\alpha\mu_i}{1+\alpha y_i}\right) \\ & - \frac{1}{2}\ln(2\pi) - \ln(1+\alpha y_i) - \frac{1}{2}\ln\left(y_i + \frac{1}{6}\right) \\ & - \frac{1}{2}\ln\left(1+\frac{\alpha}{6}\right) + \frac{1}{2}\ln\left(\alpha y_i + 1 + \frac{\alpha}{6}\right), \end{aligned} \tag{12}$$

and α is estimated by maximizing this. A simulation study of Piegorsch [27] ended by concluding that "it is more prudent to recommend general use of the maximum *quasi*-likelihood approach as long as the sample size is adequate (above 20), and α is not very small".

Regression Models with Random Effects

Random Intercept Model

Suppose that y_i follows the Poisson distribution with expected value $\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta} + u_i)t_i$. The term u_i is a random intercept that can be modeled by assuming that $\exp(u_i)$ follows a gamma density with mean 1 and variance α ([21]; see also [5]). This is consistent with the interpretation of u_i as a random noise with expected value zero.

Example In a paper by Zeger & Edelstein [32], y_i denoted the number of deaths in village i with t_i person-years at risk, $\mathbf{x}_i$ was a vector of predictor variables, and u_i was a random effect for the effects of baseline health status on mortality. The model allowed the underlying mortality rate to vary from village to village.

Extensions

First, an important extension allows the dispersion parameter to vary as a function of covariates, such as by letting $\log \alpha_i = \mathbf{z}_i \boldsymbol{\gamma}$. Using this approach, Manton et al. [22] analyzed the geographic distribution of age-specific lung cancer rates among North Carolina counties using a model with scale parameter varying by age group and shape parameter by birth cohort.

Secondly, in longitudinal studies a series of measurements $\mathbf{y}_i$ is recorded for the ith subject. When the responses are counts, the negative binomial model allows subject-specific covariates and random terms for each subject. In most cases, however, we have to cope with time-varying covariates. Morton [25] adapts the negative binomial model by proposing that the conditional variance take the form $V(y_{ij}|u_i) = \phi\mu_{ij}$, obtaining the estimates of β, α, and ϕ using a quasi-likelihood approach.

The reader familiar with generalized linear models (GLMs) will be aware that taking $\log \mu_i$ as a linear function of the covariates and incorporating random terms results in a noncanonical link, so that standard errors are only asymptotically equivalent to those obtained from a GLM extended quasi-likelihood model with canonical link $\log[\mu/(\mu + \alpha)]$. Models with random effects on the same scale as the fixed effects can be more attractive [6], while other methods avoid the specification of the form of the mixing distribution (see [18] and [1] for an extension to the general **exponential family**).

Empirical Bayes Estimates

Suppose that we want to estimate the rate of a certain disease by areas within a country. Given $\{y_i\}$, the observed event counts, and $\{E_i\}$, the expected counts for the areas given some standard reference rates, the maximum likelihood estimator for θ_i is $r_i = y_i/E_i$, while the empirical Bayesian estimator for θ_i is the mean of the posterior density $f(\theta_i|y_i)$ [14],

$$f(\theta_i|y_i) = \frac{f(y_i|E_i\theta_i)\, f(\theta_i|\kappa, \nu)\, \mathrm{d}\theta_i}{\int f(y_i|E_i\theta_i)\, f(\theta_i|\kappa, \nu)\, \mathrm{d}\theta_i}, \qquad (13)$$

which is gamma $[y_i + \kappa, \nu/(\nu E_i + 1)]$ with mean $\hat{\theta}_i = (y_i + \kappa)/(E_i + \nu^{-1})$. The parameters κ and ν can be estimated from the marginal distribution $\int f(y_i|\theta_i)\, f(\theta_i)\, \mathrm{d}\theta_i$, which is the negative binomial distribution.

Example The empirical Bayes estimator (ebmr) of the rate ratios for lip cancer of 56 Scottish counties is reported by Clayton & Kaldor [9]. The ebmr values are shrunk toward the mean, and nonzero estimates are produced even for areas with observed zero counts. The degree of shrinkage is proportional to the variance of the gamma density and, for a given area, to the amount of population at risk (i.e. the E_i) (*see* **Shrinkage Estimation**).

Test of Overdispersion for Count Data

A *score* test with one degree of freedom (*see* **Likelihood**), based on the negative binomial variance, uses the statistic

$$T_1^2 = \frac{\left[\sum_i^n (y_i - \hat{\mu}_i)^2 - (1 - h_i)\hat{\mu}_i\right]^2}{2\sum_i^n \hat{\mu}_i^2}, \qquad (14)$$

where $\hat{\mu}_i$ is the expected value and h_i is the diagonal element of the hat matrix derived from fitting a regression model to the observed counts [11, 12]. Under a quasi-likelihood approach, the statistic is

$$T_2^2 = \frac{1}{2n}\left[\sum_i^n \frac{(y_i - \hat{\mu}_i)^2 - (1 - h_i)\hat{\mu}_i}{\hat{\mu}_i}\right]^2. \qquad (15)$$

The first test is an extension of the Fisher dispersion test (*see* **Poisson Distribution**) to the general case of regression models (see also [10]).

Conclusions

The interest in the negative binomial distribution arises from the frequency of occurrence in fieldwork of count data with over- or underdispersion with respect to the Poisson variance. Regression models can be built and maximum likelihood or extended maximum quasi-likelihood estimates of the mean and dispersion parameters can be obtained. Currently, some software provides commands for negative binomial regression analyses (see `glm`, `nbreg`, and `gnbreg` in *Stata* [29]).

Applications can be found in **geographical analysis**, in experimental designs, in epidemiology for the analysis of cohort studies, and in **longitudinal data analysis**. In most of these cases, the negative binomial distribution is considered as an alternative to the Poisson. The assumption of a specific form for the density of random terms, using a scale different from that of the fixed effects, limits the popularity of such modeling.

References

[1] Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models, *Statistics and Computing* **6**, 251–262.

[2] Anraku, K. & Yanagimoto, T. (1990). Estimation for the negative binomial distribution based on the conditional likelihood, *Communications in Statistics – Simulation and Computation* **19**, 771–786.

[3] Anscombe, F.J. (1949). The statistical analysis of insect counts based on the negative binomial distribution, *Biometrics*, **5**, 165–173.

[4] Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions, *Biometrika* **36**, 358–382.

[5] Breslow, N.E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics* **33**, 38–44.

[6] Breslow, N.E. & Clayton, D. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**, 9–25.

[7] Chen, R. (1978). A surveillance system for congenital malformations, *Journal of the American Statistical Association* **73**, 323–327.

[8] Clark, S.J. & Perry, J.N. (1989). Estimation of the negative binomial parameter k by maximum quasi-likelihood, *Biometrics* **45**, 309–316.

[9] Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671–681.

[10] Collings, B.J. & Margolin, B.H. (1985). Testing goodness-of-fit for the Poisson assumptions when the observations are not identically distributed, *Journal of the American Statistical Association* **80**, 411–418.

[11] Dean, C. (1992). Testing for overdispersion in Poisson and binomial regression models, *Journal of the American Statistical Association* **87**, 451–457.

[12] Dean, C. & Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association* **84**, 467–472.

[13] Diehr, P. (1984). Small area statistics: large statistical problems. Editorial, *American Journal of Public Health* **74**, 4.

[14] Efron, B. & Morris, C. (1975). Data analysis using Stein's estimation and its generalization, *Journal of the American Statistical Association* **70**, 311–319.

[15] Fisher, R.A. (1953). A note on the efficient fitting of the negative binomial, *Biometrics* **9**, 197–200.

[16] Freund, J.E. & Walpole, R.E. (1980). *Mathematical Statistics*. Prentice-Hall, London.

[17] Greenwood, M. & Yule, G.U. (1920). An inquiry into the nature of frequency distributions of multiple happenings, *Journal of the Royal Statistical Society* **83**, 255.

[18] Hinde, J.P. (1982). Compound Poisson regression models, in *GLIM82*, R. Gilchrist, ed. Springer-Verlag, New York.

[19] Johnson, N.L. & Kotz, S. (1969). *Discrete Distribution*. Houghton Mifflin, Boston.

[20] Kalbfleisch, J.D. & Sprott, D.A. (1973). Marginal and conditional likelihoods, *Sankyā, A* **35**, 311–328.

[21] Lawless, J.F. (1987). Negative binomial and mixed Poisson regression, *Canadian Journal of Statistics* **15**, 209–225.

[22] Manton, K.G., Woodbury, M.A. & Stallard, E. (1981). A variance component approach to categorical data models with heterogeneous cell population: analysis of spatial gradients in lung cancer rates in North Carolina counties, *Biometrics* **37**, 259–269.

[23] Marshall, R.J. (1991). Mapping disease and mortality rates using empirical Bayes estimators, *Applied Statistics* **40**, 283–294.

[24] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[25] Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation, *Biometrika* **74**, 247–257.

[26] Nelder, J.A. & Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika* **74**, 221.

[27] Piegorsch, W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter, *Biometrics* **46**, 863–867.

[28] Scheaffer, R.L. & Leavenworth, S. (1976). The negative binomial model for counts in units of varying size, *Journal of Quality Technology* **3**, 158–163.

[29] StataCorp. (1997). *Stata Statistical Software Release 5.0*, Stata Corporation, College Station.

[30] Stuart, A. & Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 1. Wiley, New York, pp. 180–181.

[31] Student (W.S. Gosset) (1907). On error of counting with an haemocytometer, *Biometrika* **5**, 351–360.

[32] Zeger, S.L. & Edelstein, S.L. (1989). Poisson regression with a surrogate X; an analysis of vitamin A and Indonesian children's mortality, *Applied Statistics* **38**, 309–318.

(*See also* **Contagious Distributions**; **Generalized Linear Model**; **Poisson Regression**)

ANNIBALE BIGGERI

Nelson–Aalen Estimator

The Nelson–Aalen estimator is a nonparametric estimator which may be used to estimate the cumulative hazard rate function from censored survival data (*see* **Survival Distributions and Their Characteristics**). Since no distributional assumptions are needed, one important use of the estimator is to check graphically the fit of parametric models, and this is the reason why it was originally introduced by Nelson [10, 11]. Independently of Nelson, Altshuler [2] derived the same estimator in the context of **competing risks** animal experiments. Later, by adopting a counting process formulation, Aalen [1] extended its use beyond the survival data and competing risks setups, and studied its small and large sample properties using martingale methods. The estimator is nowadays denoted the Nelson–Aalen estimator, although other names (the Nelson estimator, the Altshuler estimator, the Aalen–Nelson estimator, the empirical cumulative hazard estimator) are sometimes used as well. Below we present a number of situations where the Nelson–Aalen estimator may be applied and exemplify its use in one particular case. Furthermore, we indicate how counting processes provide a framework which allows for a unified treatment of all these diverse situations, and we summarize the most important properties of the Nelson–Aalen estimator. A detailed account is given in [3, Section IV.1].

Survival Data

Consider first the survival data situation, where we want to study the time to death (or some other event) for a homogeneous population with hazard rate function $\alpha(t)$ and cumulative hazard rate function $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. Assume that we have a sample of n individuals from this population. Our observation of the survival times for these individuals will typically be subject to right censoring, meaning that for some individuals we only know that their true survival times exceed certain censoring times. The censoring is assumed to be independent in the sense that the additional knowledge of censorings before any time t does not alter the risk of failure at t (*see* **Censored Data**). We denote by $t_1 < t_2 < \cdots$ the times when deaths are observed and let d_j be the number of individuals who die at t_j.

The Nelson–Aalen estimator for the cumulative hazard rate function then takes the form

$$\widehat{A}(t) = \sum_{t_j \le t} \frac{d_j}{r_j}, \tag{1}$$

where r_j is the number of individuals at risk (i.e. alive and not censored) just prior to time t_j. Thus the Nelson–Aalen estimator is an increasing right-continuous step function with increments d_j/r_j at the observed failure times. The variance of the Nelson–Aalen estimator may be estimated by

$$\widehat{\sigma}^2(t) = \sum_{t_j \le t} \frac{(r_j - d_j)d_j}{(r_j - 1)r_j^2}. \tag{2}$$

It may be shown (see below) that the Nelson–Aalen estimator (1) as well as the variance estimator (2) are almost unbiased. In large samples the Nelson–Aalen estimator, evaluated at a given time t, is approximately normally distributed, so a standard $100(1-\alpha)\%$ confidence interval for $A(t)$ takes the form

$$\widehat{A}(t) \pm z_{1-\alpha/2}\widehat{\sigma}(t), \tag{3}$$

with $z_{1-\alpha/2}$ the $1-\alpha/2$ fractile of the standard normal distribution. The approximation to the normal distribution is improved by using a log transform giving the confidence interval

$$\widehat{A}(t) \exp\left[\pm z_{1-\alpha/2}\frac{\widehat{\sigma}(t)}{\widehat{A}(t)}\right]. \tag{4}$$

This interval is satisfactory for quite small sample sizes [5].

Right censoring is not the only kind of data incompleteness in survival analysis. Often, e.g. in epidemiological applications, individuals are not followed from time zero (in the relevant time scale, typically age), but only from a later entry time (conditional on survival until this entry time). Thus, in addition to right censoring, the survival data are subject to left truncation. For such data we may still use the Nelson–Aalen estimator (1) and estimate its variance by (2). The number at risk, r_j, now is the number of individuals who have entered the study before time t_j and are still in the study just prior to t_j. For left-truncated data the numbers at risk, r_j, may be low for small values of t_j. This will result in estimates $\widehat{A}(t)$ which have large sampling errors. But because the increments of the Nelson–Aalen estimator are

uncorrelated (see below), the uncertainty induced for small time values has no influence on the increment $\widehat{A}(t) - \widehat{A}(s)$ of the Nelson–Aalen estimator over a later time interval $(s, t]$. An estimator for the variance of this increment is $\widehat{\sigma}^2(t) - \widehat{\sigma}^2(s)$.

Quite often we want to estimate the survival distribution function $S(t) = \exp[-A(t)]$, representing the probability that an individual will be alive at time t. This may be done from right-censored and/or left-truncated survival data by the **Kaplan–Meier estimator**. The relation $A(t) = -\ln S(t)$ suggests that the cumulative hazard rate function alternatively may be estimated as minus the logarithm of the Kaplan–Meier estimator. Even though this estimator numerically will be close to the Nelson–Aalen estimator, the latter is the canonical one from a theoretical point of view. Furthermore, the Nelson–Aalen estimator may be used in a number of different situations (see below) while the alternative estimator applies only to the survival data situation.

An Illustration

To give an illustration of the Nelson–Aalen estimator we use data from a randomized clinical trial for patients with histologically verified liver cirrhosis. Patients were recruited from several hospitals in Copenhagen between 1962 and 1969 and were followed until death, lost to follow-up or until the closing date of the study, October 1, 1974. The time variable of interest is time since entry into the study. Patients are right censored if alive on October 1, 1974, or if lost to follow-up before that date.

We consider only the 138 placebo-treated male patients. Their median age at entry was 57 years, while the lower and upper quartiles were 51 and 66 years, respectively. Of the 138 patients, 88 died during the study. The Nelson–Aalen estimate for these patients is shown in Figure 1 with 95% confidence intervals computed according to (4). Even though the cumulative hazard rate function provides a useful summary measure (e.g. [6, Section 2.3]), it is usually the hazard rate function itself which is the entity of real interest. So when interpreting the estimate in Figure 1, we mainly focus on the "slope" of the curve. The estimate of the cumulative hazard rate function is steeper for the first 9–10 months after randomization than at later times. Therefore we have evidence that the risk of dying for these patients is highest just after randomization. (This may, at least in part, be due to heterogeneity which is not accounted for in our simple analysis.) The hazard rate function is approximately 0.3 per year for the first 9–10 months and slightly below 0.2 per year thereafter when estimated as the average slope of the curve over the relevant time periods. More formal procedures for smoothing the Nelson–Aalen estimate in order to obtain an estimate for the hazard rate function itself are available but will not be considered here (*see* **Smoothing Hazard Rates**). A further discussion and analysis of the cirrhosis data is given in [12]. The data were also used for illustrative purposes in [3].

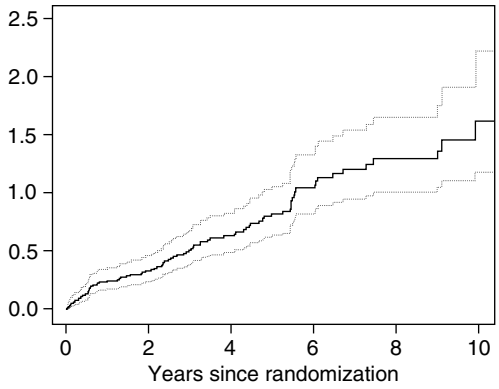

Figure 1 Nelson–Aalen estimate of the cumulative hazard rate function for death for 138 placebo-treated male patients with liver cirrhosis, with 95% log-transformed confidence intervals

Multi-state Models and Recurrent Events

The survival analysis setup considered above may be generalized in two directions. More than one type of event may be considered for each individual under study, and/or the event in question may happen more than once for each individual. Examples of the first type are competing risks with two or more causes of death and the Markov illness–death model with the states "healthy", "diseased", and "dead" (*see* **Counting Process Methods in Survival Analysis**). More generally, we may consider any **Markov process** with a finite number of states which may be used to model the life history of an individual. An example of the second type is an inhomogeneous **Poisson process** with intensity $\alpha(t)$ modeling the occurrence of some recurrent event like episodes of

angina pectoris in patients with coronary heart disease or infections in AIDS patients. For both of these two types of situations we observe the times when events occur for a number of individuals (modeled as iid copies of the relevant process) who need not all be observed over the same interval of time. The Nelson–Aalen estimator may then be applied to estimate cumulative intensities.

To be specific, consider a finite-state Markov process with transition intensities $\alpha_{gh}(t)$ for $g \neq h$. Focusing on fixed g and h in what follows, we drop the subscripts and write just $\alpha(t)$ for the $g \to h$ transition intensity. Furthermore, we denote by $t_1 < t_2 < \cdots$ the times when transitions from g to h are observed. Let d_j be the number of individuals who experience a $g \to h$ transition at t_j, and write r_j for the number of individuals in state g (i.e. at risk for a $g \to h$ transition) just prior to time t_j. Then the cumulative $g \to h$ transition intensity $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ may be estimated by (1) and its variance by (2). Similarly, the integrated intensity of an inhomogeneous Poisson process may be estimated with the t_js denoting the times of observed events, and the d_js and r_js being the corresponding numbers of events and numbers at risk, respectively. An illustration of the use of the Nelson–Aalen estimator to estimate integrated Markov transition intensities is given by Keiding & Andersen [9].

Two Other Applications

For the situations considered so far, (1) and (2) apply with r_j the number at risk at t_j for the event in question. The use of the Nelson–Aalen estimator is, however, not restricted to such situations. We mention here two other applications and return to a general discussion below.

Relative Mortality

Our first example considers right-censored and/or left-truncated survival data, but they no longer come from a homogeneous population. Rather, we assume that the hazard rate function of the ith individual may be written as the product $\alpha(t)\mu_i(t)$, where $\alpha(t)$ is a relative mortality common to all individuals and $\mu_i(t)$ is the hazard rate function at time t for a person from an external standard population corresponding to the ith individual (e.g. of the same sex and age as individual i). Typically the $\mu_i(t)$ will be known from published life tables for the general population. In this situation the Nelson–Aalen estimator may be used to estimate the cumulative relative mortality $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. All that is required is that r_j in (1) be taken to denote the sum of the external rates $\mu_i(t_j)$ for all individuals at risk just prior to t_j. An illustration of this use of the Nelson-Aalen estimator is provided by Breslow & Day [7, Chapter 5].

An Epidemic Model

A simple model for the spread of an infectious disease in a community is the following (*see* **Epidemic Models, Stochastic**). At the start of the epidemic, i.e. at time $t = 0$, some individuals make contact with individuals from elsewhere and are thereby infected with the disease. There are no further infections from outside the community during the course of the epidemic. Let $S(t)$ and $I(t)$ denote the number of susceptibles and infectives, respectively, just prior to time t. Assuming random mixing, the infection intensity in the community at time t becomes $\alpha(t)S(t)I(t)$, where $\alpha(t)$ is the infection rate per possible contact. We denote by $0 < t_1 < t_2 < \cdots$ the times when individuals are infected and let d_j denote the number infected at t_j. Then the cumulative infection rate, $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$, may be estimated by the Nelson–Aalen estimator (1) where now $r_j = S(t_j)I(t_j)$; see Becker [4, Section 7.6] for an illustration.

Counting Process Formulation and Small Sample Properties

In general we consider the occurrences of some events of interest (e.g. deaths, occurrences of a disease, infections), and denote by $0 < t_1 < t_2 < \cdots$ the times when an event is observed. We assume that two or more events cannot occur at the same time, so that there are no tied observations. (The handling of ties is discussed briefly below.) Then the process $N(t)$ counting the number of observed events in the time interval $[0, t]$ is a (univariate) counting process. The behavior of $N(t)$ is governed by its intensity process $\lambda(t)$ given heuristically by

$$\lambda(t)\,\mathrm{d}t = \Pr(\text{event occurs in } [t, t + \mathrm{d}t) | \mathcal{F}_{t-}).$$

Here $\mathcal{F}_{t-}$ represents all the information available to the researcher just before time t. The counting process satisfies Aalen's multiplicative intensity model if we may write its intensity process as

$$\lambda(t) = \alpha(t)Y(t), \tag{5}$$

for some unknown function $\alpha(t)$ and some observable process $Y(t)$ whose value at time t is known just prior to t. All the situations considered above give counting processes which fulfill (5). Survival data from a homogeneous population, finite-state Markov processes, and the inhomogeneous Poisson process, all give a $Y(t)$ process which is the number at risk just prior to time t. For the model for relative mortality, $Y(t)$ is the sum of the $\mu_i(t)$ for those at risk just before t, while for the epidemic model, $Y(t) = S(t)I(t)$. The common structure of all these models when formulated as counting processes is the reason why the Nelson–Aalen estimator may be applied to all these diverse problems.

In fact, the counting process formulation provides a framework which makes it simple to study the statistical properties of the Nelson–Aalen estimator. We briefly indicate a few main steps and refer to [3, Section IV.1.1] for a thorough treatment. First, we note that, with $r_j = Y(t_j)$, we may write the Nelson–Aalen estimator (1) as

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y(s)}\,\mathrm{d}N(s), \tag{6}$$

where $J(s) = I(Y(s) > 0)$ and 0/0 is interpreted as 0. Then using (5), (6), and the decomposition $N(t) = \int_0^t \lambda(s)\,\mathrm{d}s + M(t)$ of a counting process into a sum of its integrated intensity process and a local square integrable martingale $M(t)$, we obtain

$$\widehat{A}(t) = A^*(t) + M^*(t). \tag{7}$$

Here $A^*(t) = \int_0^t J(s)\alpha(s)\,\mathrm{d}s$ is almost the same as $A(t)$ when there is only a small probability that $Y(s) = 0$ for some $s \le t$, while $M^*(t) = \int_0^t [J(s)/Y(s)]\,\mathrm{d}M(s)$ is a stochastic integral and as such is a local square integrable martingale. Relation (7) is the key to studying the statistical properties of the Nelson–Aalen estimator. Since $M^*(t)$ has expected value zero for any given t, we have $\mathrm{E}\widehat{A}(t) = \mathrm{E}A^*(t)$, so the Nelson–Aalen estimator is almost unbiased. Furthermore, an unbiased estimator for the variance of $M^*(t)$ is its optional variation process $\int_0^t [J(s)/Y(s)^2]\,\mathrm{d}N(s)$. Thus the variance estimator (2) is almost unbiased when there are no ties. Finally, a martingale has uncorrelated increments, and by (7) this is (almost) the case for the Nelson–Aalen estimator as well.

In the presence of ties, i.e. when the number of events d_j at t_j exceeds one, the process $N(t)$ counting occurrences of events in $[0, t]$ may have jumps of size two or larger and is therefore no longer a counting process. Often, however, we may write $N(t) = \sum_{i=1}^n N_i(t)$, where $N_i(t)$ is a counting process registering the events for individual i. If we consider a homogeneous population where the rates of occurrence of the events are the same for all individuals, we may adopt the discrete extension of the model described in [3, pp. 180–181]. For this extended model, the arguments of [8, pp. 94–96], apply, to show that the variance estimator (2) is almost unbiased also in the presence of ties. This justifies the use of the tie-corrected estimator (2) for all situations considered above, except for the model with relative mortality and the epidemic model. Within the framework of the extended model the Nelson–Aalen estimator is a **nonparametric maximum likelihood** estimator; see [3, Section IV.1.5] for details and further discussion.

Weak Convergence and Confidence Bands

By (7) the martingale central limit theorem may be used to prove that, considered as a stochastic process, the Nelson–Aalen estimator (properly normalized) converges weakly to a mean zero Gaussian martingale. In particular, for a fixed t it is asymptotically normally distributed, a fact that was used in connection with the confidence intervals (3) and (4). The weak convergence result also makes it possible to derive confidence bands for A, i.e. limits which contain $A(t)$ for all t in an interval $[\tau_1, \tau_2]$ with a prespecified probability.

One important class of such confidence bands are the equal precision bands. The standard and log-transformed equal precision bands are obtained by replacing $z_{1-\alpha/2}$ in (3) and (4) by $d_{1-\alpha}$, the $1-\alpha$ fractile in the distribution of the supremum of the absolute value of a standardized Brownian bridge (over a certain time interval). This fractile may be found (approximately) by solving (with respect to d)

the nonlinear equation

$$\frac{4\phi(d)}{d} + 2\phi(d)\left(d - \frac{1}{d}\right)\ln\left[\frac{\widehat{\sigma}(\tau_2)}{\widehat{\sigma}(\tau_1)}\right] = \alpha,$$

where $\phi(d)$ is the standard normal density function. The equal precision bands require $\widehat{\sigma}(\tau_1) > 0$, so they cannot be extended all the way down to $t = 0$. Typically, one will also omit the largest values of t. The standard equal precision band has poor small sample properties, so even with sample sizes in the hundreds the use of the log transformed confidence band is recommended [5]. As an illustration we use once more the liver cirrhosis example. Considering the interval from 4 months (1/3 year) to 8 years, we have $\widehat{\sigma}(1/3) = 0.027$ and $\widehat{\sigma}(8) = 0.163$, so that $d_{0.95} = 2.99$. Therefore the 95% log transformed equal precision band for the cumulative hazard rate function between 4 months and 8 years may be obtained from (4) by using the fractile 2.99 instead of the value 1.96 used for the pointwise confidence intervals in Figure 1. A detailed study of the weak convergence of the Nelson–Aalen estimator and the derivation of confidence bands are provided by [3, Section IV.1.2-3]. Here another class of confidence bands, the Hall–Wellner bands, is also discussed.

We finally note that **semi-Markov processes** (or Markov **renewal processes**), where the transition intensities (only) depend on the sojourn times in the states, do not give rise to counting processes which fulfill the multiplicative intensity model (5). Thus the results outlined above do not immediately extend to such models. However, it turns out that enough of the above structure is preserved to be able to define Nelson–Aalen estimators also for such semi-Markov processes and to derive identical asymptotic results for these as for the case of Markov processes; see [3, Section X.1] for a discussion and further references.

References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[2] Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments, *Mathematical Biosciences* **6**, 1–11.

[3] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Becker, N.G. (1993). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

[5] Bie, O., Borgan, O. & Liestøl, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties, *Scandinavian Journal of Statistics* **14**, 221–233.

[6] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. **1**: The Analysis of Case-Control Studies, IARC Scientific Publications, Vol. **32**. International Agency for Research on Cancer, Lyon.

[7] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*. Vol. **2**: The Design and Analysis of Cohort Studies, IARC Scientific Publications, Vol. **82**. International Agency for Research on Cancer, Lyon.

[8] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[9] Keiding, N. & Andersen, P.K. (1989). Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process, *Applied Statistics* **38**, 319–329.

[10] Nelson, W. (1969). Hazard plotting for incomplete failure data, *Journal of Quality Technology* **1**, 27–52.

[11] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**, 945–965.

[12] Schlichting, P., Christensen, E., Andersen, P.K., Fauerholdt, L., Juhl, E., Poulsen, H. & Tygstrup, N., for The Copenhagen Study Group for Liver Diseases (1983). Prognostic factors in cirrhosis identified by Cox's regression model, *Hepatology* **3**, 889–895.

ØRNULF BORGAN

Nephrology

Nephrology is defined as the scientific study and treatment of the kidney and its diseases. Acute renal failure usually occurs in previously normal kidneys, often after a major injury or surgery, and typically the patient's renal function returns to normal after a short period. Chronic renal failure, which is usually associated with kidney disease, results in the irreversible deterioration of kidney function. When the condition becomes terminal the patient is said to have end-stage renal failure (ESRF). Two forms of renal replacement therapy are available for patients with ESRF; renal dialysis and **transplantation**.

Dialysis partially cleans the blood of toxic waste products normally excreted by the kidney. Hemodialysis and continuous ambulatory peritoneal dialysis (CAPD) are the most common treatment methods. Other kidney functions are supplemented by drugs. Transplantation is the preferred treatment for most patients as it replaces all functions of the kidney, but the main bar to expanding this is the worldwide shortage of donor organs.

The clinical challenge in the management of dialysis patients is the dialysis prescription. Inadequate dialysis can lead to toxin build-up, anemia, infection, nutritional problems and an increased risk of mortality. For the transplant recipient, rejection and infection are the primary complications.

Renal Disease Registers

Several regional, national, and international registries collecting data on ESRF patients have been established worldwide (*see* **Disease Registers**). The range of data collected extends to patient demography and treatment modality, acceptance rates (**incidence**) and patient stock levels (**prevalence**), patient outcome and co-morbidity factors, dialysis duration and adequacy and treatment center parameters (structure, staffing, etc.). The reports prepared by the Registries [1] indicate the areas of key interest and the statistical techniques applied.

Excepting for the registries in the US, Holland, Finland, and Germany (under development), participation is voluntary. Return rates for the first established European Dialysis and Transplant Association–European Renal Association (EDTA–ERA) registry have been particularly poor in recent years (less than 50% for some countries) but other registries claim high compliance rates of over 90% [1]. Similar registries exist in the transplant arena in the US, UK, Europe, and elsewhere [14]. These databases tend to be less ambitious, having implemented the concept of a core database of key readily available data items. For specific additional studies, more detailed information is then collected from co-opted centers for a limited time.

In addition to these registries of observational data, there have been a number of **clinical trials** in nephrology, of which the National Cooperative Dialysis Study (see, for example, Laird et al. [6]) is probably the most comprehensive.

Nephrology Journals

A number of journals publish articles on the study and treatment of renal disease in the clinical and laboratory setting. Many papers include reference to, and use of, statistical methods, although often the content is quite low. Use of the t, Wilcoxon, and χ^2 tests and analysis of variance are widely reported, as are Kaplan–Meier and Cox survival analyses. Most authors quote the package used but few comment on the analysis method chosen or its limitations, and even fewer acknowledge statistical advice. Generally, the statistical content is higher in the transplantation journals and in *Kidney 1International*.

Patient Survival Under Alternative Treatment Modalities

Choosing an Appropriate Start Point and Study Cohort

For many patients, the onset of ESRF can only be defined retrospectively as the long-term prognosis may be uncertain when regular dialysis first begins, so the start point is often set at 90 or 120 days after the start of dialysis. Also, several patients will change dialysis modality in the first months of treatment. Similarly, when comparing patients' survival prospects under dialysis and transplantation, setting the start point as the date of listing for transplant excludes the potential bias of including patients unsuitable for transplantation.

Modeling Treatment Changes

Several authors have avoided the issue of patients changing treatment by censoring (*see* **Censored Data**) when a change occurs. To minimize the effect of possible nonindependent censoring, the survival time is censored at the point of change, but a death within the following two months is considered a death on the previous modality (unless the death is unrelated to treatment) (see, for example, Nelson et al. [12]). Similarly, transplantees are censored at the date of grafting, as the chance of a patient receiving a kidney is governed by donor availability and not by dialysis mode or duration.

This is inadequate when the total survival experience under alternative treatments is of interest (*see* **Survival Analysis, Overview**). For a transplantee, for example, it is necessary to account for the time the patient survived on dialysis prior to receiving a graft. The **Cox regression model** with **time-dependent covariates** can be used for inference about **relative risks** in this situation.

In its simplest form (with a single treatment change), the model would be

$$h_i(t) = h_0(t) \times \exp[\mathbf{z}_i'\beta + I_i(t)\gamma], \qquad (1)$$

where $h_0(t)$ is the baseline **hazard**, $\mathbf{z}_i$ is the $p \times 1$ vector of fixed covariates, and $I_i(t) = 0$ if patient i has not changed treatment at time t, 1 otherwise, with $\beta_1, \ldots, \beta_p$, γ the regression coefficients to be estimated.

For a patient changing between dialysis treatments, this model may not be unreasonable, but for a change from dialysis to transplantation, model (1) is not entirely appropriate as it assumes the transplantation hazard remains constant. It is well recognized that the hazard after transplantation is not constant over time, and that the initial postoperative period has the greatest hazard. An extension to (1) that would allow for transient effects would be to include multiple indicators to reflect the clinically relevant cutpoints, t_1, t_2, and t_3, say, namely

$$\begin{aligned} h_i(t) = h_0(t) \times \exp[\mathbf{z}_i'\beta &+ I_i(0 \le t - t_i^* < t_1)\gamma_1 \\ &+ I_i(t_1 \le t - t_i^* < t_2)\gamma_2 \\ &+ I_i(t_2 \le t - t_i^* < t_3)\gamma_3], \end{aligned} \qquad (2)$$

where t_i^* is the time from entry to the study to transplant.

The effects would then be modeled by way of the step function. An alternative approach for exploring transient hazard effects is to assume the hazard after transplant has an exponential decay (*see* **Parametric Models in Survival Analysis**), namely

$$\begin{aligned} h_i(t) = h_0(t) &\times I_i(t) \\ &\times \exp\{\beta_0 + \beta_1 \exp[-\gamma(t - t_i^*)]\}. \end{aligned} \qquad (3)$$

Other covariates can be controlled for by **stratification**. Mauger et al. [10] have developed this further to estimate the times at which the hazard and survival curves for dialysis patients awaiting transplantation cross the transplant hazard and survival curves.

Inherent in models (1) and (2) is the assumption that the relative risks, $\beta_1, \ldots, \beta_p$, do not change over time. Various studies have shown that the risks attributable to human leukocyte antigen (HLA) mismatching for transplant survival are transient. For further exploration of issues surrounding the analysis of transplant outcome per se, *see* **Transplantation**.

Occurrence of Repeated Events

Repeated events, usually of an adverse nature, arise in many clinical settings, including nephrology. For example, infections in the peritoneal cavity and at the point of insertion of the catheter are common complications that hamper significantly the success of CAPD as a treatment. Statistical models have been developed to understand better the etiology and incidence of infection in these patients. Infection occurrence can be recorded in one of two ways:

1. Number of infections, x_i, say, over follow-up time T_i.

In this situation, a **Poisson process** for counts is appropriate, and both **fixed effects** and **random effects** models have been explored. Vonesh [15] fitted a mixed effect **gamma** Poisson **multiplicative model** for modeling individual peritonitis infection rates, namely

$$\lambda(\mathbf{z}_i) = \lambda_i \exp(\mathbf{z}_i'\beta), \qquad (4)$$

with

$$\lambda_i \sim \Gamma(\alpha, \gamma), \qquad (5)$$

and compared the results with a fixed effects model with $\lambda_i = \lambda_0$ for all i. The conclusions were similar,

but the mixed model, by giving less weight to the few high-risk cases, provided a better fit and more realistic **standard errors**.

2. Distinct recurrence times, t_{ij}, $j = 1, \ldots, n_i$, where n_i is the number of infections occurring in follow-up time T_i.

When distinct times to infection are available, survival models with covariates can be applied. Common approaches when repeated event times are recorded are either to reduce the data to counts and fit **Poisson regression** models as described above, or, if the first event is of greatest clinical significance and the number of repeated events small, to consider time to the first occurrence and discard second and subsequent events.

Hierarchical models with **frailty** allow full use of the data and have been used to evaluate the relative risks for repeat infections at the catheter insertion with censoring. McGilchrist & Aisbett [11] assumed the recurrence times were independent and suggested fitting a multiplicative Cox model of the form:

$$h_i(t) = h_0(t) \times f_i \exp(\mathbf{z}_i' \beta), \tag{6}$$

where f_i is the frailty term and

$$\ln(f_i) \sim \mathrm{N}(0, \sigma^2). \tag{7}$$

If independence between the repeated infections is not assumed, then more complex models are required and Lindsey [7] has explored how this can be achieved through **counting processes** and the use of **loglinear models**.

Repeated Measurements

In many clinical trials and studies in nephrology, outcome takes the form of repeated measures of a (continuous) biochemical response recorded over the treatment period. For example, in a trial to compare the biocompatibility and functionality of alternative dialysis membranes, white blood cell (WBC) counts were recorded before dialysis and at 15, 60, 120, and 210 minutes after the start of dialysis. Many authors have considered the issues associated with the analysis of serial measurements in the context of clinical research (e.g. Matthews et al. [9]). Some of the simplest approaches are to choose an appropriate summary measure such as the area under the curve, value at a fixed time or when the response reaches a peak, rate of increase/decrease (slope) for each patient, and use **analysis of variance** to compare treatment effects. If the response relative to a pretreatment baseline is appropriate, **analysis of covariance** can be used.

However, if interest lies in evolution of the response, and patient **covariates** are also appropriate, this simple approach will be inadequate and more complex models that allow for within individual **correlation** are required. Linear models with fixed and random effects and **time series** models for characterizing the covariance structure are alternate approaches that have received considerable attention in the literature. Rochon [13] uses data from a trial to evaluate the efficacy of erythropoietin for treating anemia in patients with ESRF to illustrate the use of **ARMA** covariance models with time dependence in the covariance matrix. In the trial, patients were allocated at random (*see* **Randomization**) to receive one of three drug doses and hemoglobin measurements were measured weekly for 26 weeks to ensure target levels were reached. Observations were relatively stable at the beginning of the study, but became more variable towards the end, so that the usual ARMA assumption of constant **variances** and covariances over time was not entirely appropriate and a model allowing for heteroscedasticity (*see* **Scedasticity**) was fitted.

Biomedical Time Series – Patient Monitoring

Biomedical time series also arise in the routine ongoing monitoring of both dialysis and transplant patients. The function of a transplanted kidney is monitored regularly for signs of rejection, while patients on long-term dialysis are routinely tested for high levels of toxic waste and low levels of other blood constituents, such as WBC.

Physiological variation gives rise to typically noisy series that are subject to different types of abrupt change. For example, the WBC count will rise temporarily following the administration of steroids, and to assess accurately the underlying level, these values need to be filtered out. Similarly, a functioning kidney may exhibit alternating periods of improvement and deterioration, some of which may be self-correcting, others which may require medical intervention. Thus, the simple cumulative sum

(CUSUM) procedure is not suitable as there is a need to detect both the occurrence and form of each change.

The multiprocess Kalman filter methodology was developed by Gordon & Smith [5] specifically to analyze these renal monitoring series. An appropriate form of model for the series is selected [an AR(1) for the WBC data], which can be expressed in the general form:

$$y_t = \mathbf{H}_t\theta_t + \delta y_t^{(j)}, \tag{8}$$

$$\theta_t = \mathbf{G}_t\theta_{t-1} + \delta\theta_t^{(j)}, \tag{9}$$

where y_t is the reading at time t, θ_t is a vector of parameters, $\mathbf{H}_t$ and $\mathbf{G}_t$ are known regression and transition matrices, and j represents one of four possible states – stable state, impulse (temporary perturbation), change of level, and transient (temporary perturbation). The components $\delta y_t^{(j)}$ and $\delta\theta_t^{(j)}$ are assumed **normally distributed**:

$$\delta y_t^{(j)} \sim \mathrm{N}(0, \lambda^{-1}R_y^{(j)}) \quad \text{and} \quad \delta\theta_t^{(j)} \sim \mathrm{N}(0, \lambda^{-1}\mathbf{R}_\theta^{(j)}), \tag{10}$$

where the elements of R_y and $\mathbf{R}_\theta$ are chosen to reflect the different states. The quantities of interest and of greatest clinical importance are the associated probabilities that a change of level is obtained at time t given $y_1, \ldots, y_t$, and that a change of level occurred at time t given $y_1, \ldots, y_{t+1}$.

Institutional Comparisons

Comparison of institutional success rates has received considerable attention in nephrology, particularly in the transplantation arena. Of particular interest is the work of Gilks [4]. He proposed a hierarchical Bayes model for distinguishing between systematic and random variation in center success rates and showed for a study of UK transplant center success rates from 1978 to 1984 that the phenomenon of persistent center variability in raw success rates to suggest institutional variation was an illusion.

Missing Data

Random missing values occur (*see* **Missing Data**), especially in multicenter registries, because data may be unavailable at some units and from patient migration to other cities and missed appointments. Informative missing values arise particularly in studies that involve monitoring renal function. Typically, patients who show a more rapid decline in function are those more likely to terminate the study early and this tends to occur particularly among patients enrolling with poor function.

Clinical Trials

There are numerous examples in the literature of **clinical trials** in nephrology. The use of randomized controlled trials and **factorial designs** and **crossover designs** have proved particularly popular. These designs are appropriate as the majority of trial "treatments" are not expected to cure or permanently alter the state of the patient.

Examples of recent trials that have presented particular statistical challenges, although not unique to nephrology, include the treatment of side effects of renal dialysis (multivariate repeated measurements evaluated using **Bayesian methods**) [3], and complications arising from long-term dialysis (**paired comparisons** with a "no preference" option in the analysis of pain) [8].

Future Developments

The examples we have considered here were chosen to illustrate the range of statistical methods that have found application in nephrology in recent years. While not unique to nephrology, many of the more complex models described are examples of the application of recent statistical advances and methodological developments. The computational aspects of fitting these models are necessarily complex and further evaluation of the performance of the **algorithms** used for estimating the model parameters, and of the properties of the estimates obtained, are likely to be an area of continuing research.

Extensions to the Cox model by way of assessing the model through the use of various diagnostics such as martingale residuals (*see* **Counting Process Methods in Survival Analysis**), estimated **explained variation**, and assessing a model's predictive ability though **cross-validation** are just some developments that could have application in nephrology studies. A

recently proposed **nonparametric method** for estimating relationships between covariates and hazard rates that relaxes the assumption that covariate effects are linear and additive on the log hazard [2] may also be relevant.

While the survival of the patient is the primary outcome, there is increasing interest in the development of **co-morbid** conditions among renal patients, such as heart disease and malignancy (in transplant recipients). Further developments are needed to study these **competing risks** better and to assess the influence of a patient's treatment history, and factors specific to different "treatment" episodes, on their future status.

With the increasing computer power and the continuing development of renal disease databases there is potential for greater use of **simulation**, not only in resource planning and allocation but also in assessing the validity and **robustness** of model estimates. The problem of missing values, particularly in large multicenter databases, is likely to be a continuing one and awareness of the effect and handling of all types of missing values is an area for further study.

References

[1] D'Amico, G. & Striker, G.E. (1995). Proceedings from the Symposium on Renal Replacement Therapy Throughout the World: The Registries, *American Journal of Kidney Diseases* **25**, 113–205.

[2] Escolono, S., Golmard, J.L. & Mallet, A. (1995). A fully non-parametric approach to survival models with explanatory variables, *Communications in Statistics – Theory and Methods* **24**, 3027–3054.

[3] Farrow, M. & Goldstein, M. (1993). Bayes linear methods for grouped multivariate repeated measurement studies with application to crossover trials, *Biometrika* **80**, 39–59.

[4] Gilks, W.R. (1987). Some applications of hierarchical models in kidney transplantation, *Statistician* **36**, 127–136.

[5] Gordon, K. & Smith, A.F.M. (1990). Modeling and monitoring biomedical time series, *Journal of the American Statistical Association* **85**, 328–337.

[6] Laird, N.M., Berkey, C.S. & Lowrie, E.G. (1983). Modeling success or failure on dialysis therapy. The National Cooperative Dialysis Study, *Kidney International* **23**, S101–S106.

[7] Lindsey, J.K. (1995). Fitting parametric counting processes using log-linear models, *Applied Statistics* **44**, 201–212.

[8] Matthews, J.N.S. & Morris, K.P. (1995). An application of Bradley–Terry type models to the measurement of pain, *Applied Statistics* **44**, 243–255.

[9] Matthews, J.N.S., Altman, D.G., Campbell, M.J. & Royston, P. (1990). Analysis of serial measurements in medical research, *British Medical Journal* **300**, 230–235.

[10] Mauger, E.A., Wolfe, R.A. & Port, F.K. (1995). Transient effects in a Cox proportional hazards regression model, *Statistics in Medicine* **14**, 1553–1565.

[11] McGilchrist, C.A. & Aisbett, C.W. (1991). Regression with frailty in survival analysis, *Biometrics* **47**, 461–466.

[12] Nelson, C.B., Port, F.K., Wolfe, R.A. & Guire, K.E. (1992). Comparison of continuous ambulatory peritoneal dialysis and hemodialysis patient survival with evaluation of trends during the 1980's, *Journal of the American Society for Nephrology* **3**, 1147–1155.

[13] Rochon, J. (1992). ARMA covariance structures with time heteroscedasticity for repeated measures experiments, *Journal of the American Statistical Association* **87**, 777–784.

[14] Terasaki, P.I. & Cecka, J.M., eds. (1994). *Clinical Transplants 1994*. UCLA Tissue Typing Laboratory, Los Angeles.

[15] Vonesh, E.F. (1990). Modeling peritonitis rates and associated risk factors for individuals in continuous ambulatory peritoneal dialysis, *Statistics in Medicine* **9**, 263–271.

C.A. ROGERS

Network Sampling

Conventional sampling and network sampling differ with regard to the number of different selection units at which the same population element is countable in the survey. Conventional sampling postulates that every population element is uniquely linked to one and only one selection unit at which it is enumerable in the survey. Network sampling is not subject to this restriction. The network sampling paradigm indicates that every population element belongs to a network of selection units at which it is countable and the network sizes may vary, including possibly null networks without any selection units. Conventional sampling may be viewed as a special case of network sampling in which every population element is linked to a network, and each network contains one and only one selection unit.

Flexibility with respect to network sizes provides network sampling with design alternatives that may be superior to those based on conventional sampling. Network sampling is sometimes preferable when multiple selection units seem to be inextricably linked to the same population elements. However, network sampling may be fostered as a design strategy to improve survey efficiency whether or not multiple selection units are inextricably linked to the same population elements.

Historical Perspective

Network sampling is a relatively new kind of sample design that emerged in the early 1960s in response to an **estimation** problem involving a sample survey of medical providers [7] designed to estimate the **prevalence** of cystic fibrosis, a relatively rare and often lethal genetic disease of children. In the survey, medical providers reported each individual they had treated for the disease. The estimation problem arose because in designing the survey it had been implicitly assumed that each patient had been treated by only one medical source. The survey designers did not realize that it was common practice for the same patient to be treated by multiple medical sources during the course of the disease. The mistake became apparent when multiple medical providers in the survey reported the same patients. If not adjusted, the conventional estimation procedure would have counted the same patients as many times as reported by different medical sources and the estimate would have been **biased**. The conventional procedure would have been biased even if duplicate reports of the same patients had not been counted.

Birnbaum & Sirken [1] derived three **unbiased** estimators for network sampling which addressed the effects of multiple reporting on the sample selection probabilities. Their estimators differ from one another with respect to kinds of information required about network sizes of population elements that are counted in the sample survey.

Initially, network sampling was applied only in surveys for which multiple selection units appeared to be inextricably linked to the same population elements. Many were establishment surveys which, like the cystic fibrosis survey, involved estimating population **prevalence** rates based on counts of individuals having transactions with establishments whose constituents overlap. For example, Laska et al. [9], estimated the number of different individuals receiving mental health clinic care in a sample survey of the patients of several mental health clinics with overlapping clients (see also [18]). Establishment surveys are not the only venues in which network sampling is applied when multiple selection units are inextricably linked to the same population elements. Levy & Sirken [13] applied network sampling in estimating the number of defective statistical statements in texts of technical publications on the basis of number of defective statements that straddle sampled lines of text. Faulkenberry & Garoui [6] and Hendricks et al. [8] applied network sampling in agricultural surveys to estimate the number of farms that overlapped sampled area land segments.

It was not until the 1970s that network sampling was applied as a deliberate strategy to foster design efficiency. This development occurred after Sirken [19] demonstrated that fostering network sampling could substantially increase survey yields and decrease sampling errors, particularly in population surveys of relatively rare events. He proposed fostering network sampling in household surveys by linking individuals to households of relatives and others with whom they had well-defined relationships and who could serve as good proxy respondents. Network sampling based on kinship relationships was applied

in several health surveys, including diabetes prevalence surveys [30], cancer prevalence surveys [4, 5], surveys of births and marriages [15], surveys of recent decedents [17], and surveys of the Jewish population [26].

Subsequently, it became apparent that network sampling had potential for reducing **measurement errors** as well as sampling errors. For example, network sampling using kinship counting rules was applied in a post-enumeration population survey to improve estimates of **census** population undercoverage [14, 30], and network sampling based on friendship counting rules was applied in drug use surveys to improve the quality of response on sensitive questions [16]. More recently, Sirken et al. [31] demonstrated the utility of fostering network sampling in **population-based** establishment surveys of disease prevalence which link individuals having multiple transactions with the same establishments. Population-based establishment surveys are particularly applicable when free-standing establishment frames with good measures of establishment size are not available.

Since Birnbaum & Sirken [1] first proposed several network sampling estimators, the theory of network sampling has been extended in several directions. Sirken & Nathan [28] developed several "hybrid" network estimators based on combinations of counting rules. Network sampling theory has been extended to ratio estimation [27] (*see* **Ratio and Regression Estimates**), and to complex types of sampling, including **stratified sampling** [10, 20], **cluster sampling** [11], and **multistage sampling** [31]. Also, relationships between network sampling and other sample designs involving multiple linkages of selection units to the same population elements have been investigated. For example, Casady & Sirken [2] compared stratified network sampling and multiple frame sampling estimators; Sirken [23, 24] and Casady et al. [3] derived several unbiased network estimators for dual system sampling, and Sudman et al. [32] discuss conventional sampling, network sampling, and **capture–recapture** sampling as alternative design strategies for sampling rare and elusive populations. Sampling textbooks by Levy & Lemeshow [12] and by Thompson [33] give special attention to network sampling and other sampling strategies involving multiple linkages for sampling rare and elusive populations.

Sample Design Strategies

Surveys are complex measurement instruments involving multiple interdependent design features, each feature usually having several options, and each option having its respective cost, error, and other survey effects. Designing surveys involves selecting an options set that contains one option for each design feature, and the goal is to select the optimum option set which has the greatest overall utility in meeting the survey objectives.

In addition to the sample selection procedure (i.e. stratified, cluster, multistage sampling, etc.) three survey design features are key to network sampling designs and are particularly important in understanding the survey circumstances that favor network sampling [22].

1. *Counting rules* link population elements to selection units at which they are eligible to be enumerated in the survey.
2. *Estimators* are algebraic **algorithms** for counting and weighting the population elements enumerated in the sample survey to estimate population parameters.
3. *Respondent rules* specify the sources that are eligible to provide information about population elements that are enumerated in the survey.

Each of these design features and some of their respective options are discussed in the next section, and their design effects in terms of sampling errors, measurement biases, and survey costs are discussed in the final section.

Survey Design Features

Counting Rules

Counting rules specify conditions for linking population elements to selection units at which they are eligible to be counted in the survey [22]. Groupings of selection units that are linked to the same population elements are called *networks*, and groupings of population elements that are linked to the same selection units are called *clusters*.

A requirement of conventional sampling is that the conditions specified by counting rules have the property of linking each population element to a network that contains one and only one selection

unit at which it is countable in the survey. Counting rules satisfying this condition are called *conventional counting rules*. *De facto* and *de jure* residence rules in household surveys are examples of conventional counting rules. The *de facto* rule links individuals to the households at which they happen to be located. The *de jure* rule links individuals to their legal places of residence.

Network sampling permits multiple linkages of selection units to the same population element. Counting rules of this type are called *network counting rules*. Network counting rules in household surveys typically link individuals to households of kinfolk or others (such as friends or neighbors) with whom they have close social relationships. Frequently, the rules also link individuals to their own households so that individuals without any relationships to others will not be missed in the survey.

In the following example, we compare effects of a conventional and two network counting rules on the formation of networks and clusters:

1. *Conventional rule*. Individuals are uniquely linked to their own households.
2. *Sibling rule*. Individuals are linked to households of their siblings.
3. *Conventional/sibling rule*. Individuals are linked to their own and their siblings' households.

Assume a census of diabetes prevalence is conducted on a fictional population of seven individuals residing in four households. Table 1 shows the within-household and between-household relationships of these individuals. Note that the three sons of the head of household A are siblings to one another.

Table 2 compares the effects of the counting rules on the formation of clusters. It lists the individuals that are countable at every household by each of the three rules. At household A, for example, three individuals are countable (A_1, A_2, and A_3) by the conventional rule, two individuals are countable (B_1 and C_1) by the sibling rule, and five individuals are countable (A_1, A_2, A_3, B_1 and C_1) by the conventional/sibling rule.

Table 3 compares the counting rules' effects on the formation of networks. It lists the households at which the seven individuals are countable by each counting rule. For example, B_1 is countable at one household (B) by the conventional counting rule, at two households (A and C) by the sibling counting rule, and at three households (A, B, and C) by the conventional/sibling counting rule. On the other hand, A_1 is countable once (household A) by the conventional and by the conventional/sibling rule but is missed by the sibling rule.

Continuing with this example, to estimate diabetes prevalence requires knowing which of the seven

Table 1 Fictional population of households and individuals

Households	Individuals	Within-household relationship	Relationship to head of household A
A	A_1	Head	Self
	A_2	Wife	Wife
	A_3	Son	Son
B	B_1	Head	Son
C	C_1	Head	Son
	C_2	Wife	Daughter in law
D	D_1	Head	Unrelated

Table 2 Effects of three counting rules on the formation of clusters

Household	Conventional rule	Sibling rule	Sibling/conventional rule
A	A_1, A_2, A_3	B_1, C_1	A_1, A_2, A_3, B_1, C_1
B	B_1	A_3, C_1	A_3, B_1, C_1
C	C_1, C_2	A_3, B_1	A_3, B_1, C_1, C_2
D	D_1	–	D_1

Table 3 Effects of three counting rules on the formation of networks

Individuals	Conventional rule	Sibling rule	Conventional sibling rule
A_1	A	–	A
A_2	A	–	A
A_3	A	B, C	A, B, C
B_1	B	A, C	A, B, C
C_1	C	A, B	A, B, C
C_2	C	–	C
D_1	D	–	D

individuals were diabetic. This information would be obtained by enumerating all individuals that are countable at every household (Table 2). Summing the unweighted numbers of individuals that are diabetics is an unbiased estimation procedure for the conventional counting rule because that rule counts each individual once and only once. However, this would be a biased estimation procedure for either of the network counting rules. The sibling rule is biased because it counts three individuals (A_3, B_1, and C_1) twice and fails to count any of the other four individuals. It is a biased estimation procedure for the conventional/sibling rule because three individuals (A_1, B_1, and C_1) are each counted three times.

In this example the conventional estimation procedure would be unbiased for the conventional/sibling counting rule if duplicate enumerations were eliminated for individuals counted multiple times. In some sample surveys, however, eliminating duplicate enumerations is not a sufficient condition for unbiasedness.

Survey Estimators

Two network estimators (multiplicity A and B) and a conventional estimator are compared, assuming **simple random sampling** of selection units to estimate N, the population prevalence of a **binomial** variable. Both network estimators are weighted sums of population elements that are countable at sample selection units in compliance with the counting rule adopted in the survey. The estimators differ, however, with respect to the ways they count the population elements, and the network information they use to determine the network weights.

The first multiplicity estimator (N_a) counts the same population element every time it is countable at the same or different sample selection units, and weights the element by the ratio of the number of times it is countable at the sample selection unit to the number of times it is countable at all selection units. The second multiplicity estimator (N_b) counts the same population element at most once at any sample selection unit and weights the countable elements by the inverse of the number of selection units in its network. For example, in a household survey using a conventional/sibling counting rule in which two siblings that were enumerated in a sample household have three other siblings living in another household, each of the two siblings at the sample household would get a weight of two-fifths based on the multiplicity A estimator, and a weight of one-half based on the multiplicity B estimator. Both network estimators are unbiased if every population element is linked to at least one selection unit by the counting rule. The multiplicity A estimator is one of the three unbiased network estimators originally proposed by Birnbaum & Sirken [1].

The conventional estimator (N_c) is an unweighted sum of the population units that are countable at sample selection units in compliance with the conventional counting rule.

Assume a population of N elements $I = \{I_1, \ldots, I_\alpha, \ldots, I_N\}$ with a specified attribute, and a **sampling frame**, $H = \{H_1, \ldots, H_i, \ldots, H_L\}$, containing L selection units households. Denote the links between elements and units by the indicator variable:

$$\delta_{\alpha,i} = \begin{cases} 1, & \text{if } I_\alpha \text{ is linked to } H_i, \\ & \alpha = 1, \ldots, N; i = 1, \ldots, L, \\ 0, & \text{otherwise} \end{cases}$$

The general form of the network estimator based on a sample of l selection units is

$$\hat{N} = \frac{L}{l} \sum_i^l \lambda_i,$$

where $\lambda_i = \sum_\alpha W_{\alpha i} \delta_{\alpha i}$ is the weighted sum of the population elements countable at $H_i, i = 1, \ldots, L$, and $W_{\alpha i}$ is the network weight assigned to I_α, $i = 1, \ldots, N$, when I_α is counted at $H_i, i = 1, \ldots, L$. The network estimator is unbiased if and only if

$$\sum_i W_{\alpha i} \delta_{\alpha i} = 1, \quad \alpha = 1, \ldots, N.$$

The multiplicity estimator $\hat{N}_a$ assigns the network weights

$$W_{\alpha i} = \frac{S_{\alpha i}}{S_\alpha}, \quad \alpha = 1, \ldots, N, i = 1, \ldots, L,$$

where $S_{\alpha i}$ is the number of times I_α is linked to H_i, $i = 1, \ldots, L$, and $S_\alpha = \sum_i^L S_{\alpha i}$ is the number of times I_α is linked to all L selection units.

The multiplicity estimator $\hat{N}_b$ assigns the network weights

$$W_{\alpha i} = \frac{1}{S_\alpha^*}, \quad \alpha = 1, \ldots, N; \quad i = 1, \ldots, L,$$

where $S_\alpha^* = \sum_i^L \delta_{\alpha i}$ is the number of different H_i, $i = 1, \ldots L$, to which I_α is linked.

The conventional estimator is a special case of the network estimator in which $S_\alpha = 1$, $\alpha = 1, \ldots, N$, and thus $W_{\alpha i} = 1$, $\alpha = 1, \ldots, N$; $i = 1, \ldots, L$.

It is noteworthy that the multiplicity estimators $\hat{N}_a$ and $\hat{N}_b$ require the network weights for the I_α that are enumerated at sample selection units and not at any others. Hence, it is often feasible and cost-effective to collect the information needed to calculate the weights from the selection units at which the population elements are enumerable or possibly from other information sources that are identified by the sample selection units.

Respondent Rules

Essentially, three kinds of information are collected for individuals enumerated in population surveys that are based on network sampling:

1. *Eligibility information* identifies the individuals that are countable at sample addresses in compliance with network counting rules.
2. *Topic information* calibrates countable individuals on survey topic related variables.
3. *Network information* is used to determine network weights.

For example, a diabetes prevalence survey based on the conventional/sibling rule and the multiplicity A estimator or the multiplicity B estimator would collect the following kinds of information at sample households:

1. Eligibility information – listing of resident individuals and their nonresident siblings that are countable at the household.
2. Topic information – identifies the countable individuals that have diabetes.
3. Network information – for each countable diabetic, determines either the number of his or her siblings for multiplicity estimator B (in this instance self-evident from the eligibility information) or the number of different households in which the diabetic and siblings reside for multiplicity estimator A.

Most population surveys based on network sampling collect all three types of information from the sample households at which individuals are counted. (The individual himself and/or other sample household residents may be specified as eligible within-household respondents.) Otherwise, sample households may identify other sources, such as *de jure* residences of individuals who are not residents of sample households; or the information could be obtained from multiple sources with eligibility information being ascertained at sample households, and topic and network information being obtained from *de jure* residences.

Nonhouseholds are also potential information sources. Population-based establishment surveys are notable examples of network sample surveys in which establishments are the principal sources of information about individuals enumerated in household sample surveys. Population-based establishment surveys use network counting rules such as the following: "individuals with diabetes are countable at every household whose residents were treated by the same medical providers". Suppose, for example, there are two individuals living in different households that are being treated by the same medical provider and one of the households is selected in the sample and the other is not. The population survey determines that a resident at the sample household has diabetes, and then obtains from that household the name and address of his or her medical provider. A follow-up survey is conducted with that provider who provides information about all diabetics in his practice, which in this example would be two diabetics. The network weight assigned by network estimators A and B to each diabetic would be equal to one-half.

Survey Design Effects

Sampling Errors

Assuming simple random sampling of l of L selection units with replacement, the sampling **variance** of the network estimator of N is

$$\text{var}(\hat{N}) = \frac{1}{l}\text{var}(N),$$

where

$$\text{var}(N) = \frac{1}{L}\sum_{i}^{L}(\lambda_i - \overline{\lambda})^2$$

and

$$\overline{\lambda} = \frac{1}{L}\sum_{i}^{L}\lambda_i = \frac{1}{L}\sum_{i}^{L}\sum_{\alpha}^{N} W_{\alpha i}\delta_{\alpha i} = \frac{N}{L}.$$

Clearly, the ideal network counting rules and weights would have $\lambda_i = \overline{\lambda}$, $i = 1, \ldots, L$, and $\text{var}(\lambda) = 0$. This would occur if, for example, the network counting rule linked every population element to all L selection units, that is, $S_\alpha = N$, $\alpha = 1, 2, \ldots, N$, an arrangement quite unlikely to be practicable.

Sirken [21] decomposed the variance of the multiplicity B estimator into components having interpretations that are useful for guiding the selections of counting rules and network weights:

$$\text{var}(\hat{N}_\text{b}) = \frac{1}{l}\left\{\sum_{\alpha=1}^{N}\frac{1}{L}\sum_{\beta\neq\alpha}^{N}\frac{\delta_{\alpha i}\delta_{\beta i}}{S_\alpha S_\beta} + \hat{\lambda}\frac{\text{var}(\gamma)}{\text{E}(\gamma)} + \overline{\lambda}[\text{E}(\gamma) - 1] + \overline{\lambda}(1 - \overline{\lambda})\right\}$$

where

$$\gamma_{\alpha i} = \frac{\delta_{\alpha i}}{S_\alpha}, \quad \alpha = 1, \ldots, N; i = 1, \ldots, L,$$

$$\text{E}(\gamma) = \frac{N}{R},$$

$$\text{var}(\gamma) = \frac{1}{R}\sum_{\alpha}^{N}\frac{1}{S_\alpha} - \left(\frac{N}{R}\right)^2,$$

and

$$R = \sum_{\alpha}^{N}\sum_{i}^{L}\delta_{\alpha i} = \sum_{\alpha=1}^{N} S_\alpha$$

is the total number of links between selection units and population elements.

The fourth term of $\text{var}(\hat{N}_\text{b})$ is independent of network counting rules and weights, depending only on the population parameter $\overline{\lambda}$. Ignoring the first three terms and $\text{var}(\hat{N}_\text{b})$ becomes $\text{var}(\hat{N}_\text{c})$.

Network counting rules and weights affect the first term and network counting rules affect the second and third terms.

The first term is nonnegative, and measures heaping of population elements within selection units. It is equal to zero if none of the population elements is linked to more than one selection unit. The second term is also nonnegative. It is a measure of network size variability, and it equals zero if and only if $S_\alpha = S$, $\alpha = 1, \ldots, N$. The third term is nonpositive, and is a measure of clustering of population elements within networks that equals zero if and only if $s = 1$, $\alpha = 1, \ldots, N$.

In summary, it is desirable from the viewpoint of sample error effects to select network counting rules and weights that minimize heaping of population elements within selection units, maximize **clustering** of selection units within networks, and minimize network size variability.

Though network sampling typically yields more population elements, it is not necessarily more design-effective than conventional sampling. Thus, the sampling error of neither $\hat{N}_\text{a}$ nor $\hat{N}_\text{b}$ is necessarily less than that of $(\hat{N}_\text{c})$. Nor for that matter are the sampling errors of one network estimator more design-effective than the other. However, assuming that none of the elements is linked to more than one selection unit, which is a very strong assumption most likely to be approximated when $\overline{\lambda} = N/L$ is small,

$$\text{var}(\hat{N}_\text{b}) \leq \text{var}(\hat{N}_\text{a}) \leq \text{var}(\hat{N}_\text{c}),$$

where

$$\text{var}(\hat{N}_\text{c}) = \frac{1}{l}\overline{\lambda}(1 - \overline{\lambda}),$$

$$\text{var}(\hat{N}_\text{a}) = \frac{1}{l}\overline{\lambda}\left[\frac{1}{l}\sum_{i}^{L}\sum_{\alpha}^{N}\left(\frac{S_{\alpha i}}{S_\alpha}\right)^2 - \overline{\lambda}\right],$$

and

$$\text{var}(\hat{N}_\text{b}) = \frac{1}{l}\overline{\lambda}\left[\frac{1}{L}\sum_{\alpha}^{N}\frac{1}{S_\alpha^*} - \overline{\lambda}\right].$$

It is apparent that

$$\text{var}(\hat{N}_\text{a}) = \text{var}(\hat{N}_\text{b})$$

if and only if

$$S_{\alpha i} = \frac{S_\alpha}{S_\alpha^*}, \quad \alpha = 1, \ldots, N,$$

which implies that I_α is linked the same number of times to each of the S_α^* selection units in its network. Also

$$\text{var}(\hat{N}_\text{c}) = \text{var}(\hat{N}_\text{b})$$

if and only if

$$S_\alpha^* = 1(\alpha = 1, \ldots, N).$$

Under the conditions stated above, the **design effect** (DE) of network sampling compared to conventional sampling is

$$DE = \frac{\text{var}(\hat{N}_\text{b})}{\text{var}(\hat{N}_\text{c})} = \frac{h - \overline{\lambda}}{1 - \overline{\lambda}},$$

where

$$h = \frac{1}{N} \sum_\alpha^N \frac{1}{S_\alpha^*}$$

is the inverse of the harmonic mean of S_α. If $\overline{\lambda}$ is small, then the design effect is approximately

$$DE = h.$$

Measurement Biases

The major sources of survey measurement bias are coverage, **nonresponse**, and response errors. Coverage errors occur if population elements are erroneously counted when they are unlinked to selection units, or missed when linked population elements are not counted. Nonresponse errors occur when information is not obtained for linked elements, and response errors occur when invalid information is obtained for linked elements. It is noteworthy that network estimators are subject to response errors in network information as well as response errors in survey information.

Network sampling can be an effective design strategy for handling measurement biases especially when conventional sampling is predisposed to large measurement biases, such as when conventional counting rules fail to link population elements to any selection units (e.g. incomplete sampling frames) and/or link elements to units that fail to report them in the survey.

For example, survey experiments demonstrated the efficacy of network sampling in overcoming measurement biases associated with estimating the number of deaths in conventional household surveys of population change. These surveys adopt counting rules that link living persons to their *de jure* residences, and deceased persons to their terminal *de jure* residences. Royston et al. [17] demonstrated that population change survey estimates of the numbers of deaths are subject to measurements biases due to lack of coverage of deaths occurring in nursing homes and other institutional establishments, and due to underreporting of noninstitutional deaths.

Household sample survey experiments conducted in North Carolina [29] compared the effectiveness of network and conventional sampling to estimate N, the number of North Carolinians that died during a calendar period. The experiment tested three counting rules:

Rule 1. Decedents are linked to their terminal households of residence.

Rule 2. Decedents are linked to residences of surviving spouses, siblings and children residing in counties that were decedents' terminal residences.

Rule 3. Decedents are linked by rules 1 and 2.

Rule 1 implies conventional sampling and rules 2 and 3 imply network sampling. The experiment compared the coverage and response biases and the total undercounts of deaths associated with the three counting rules.

The findings of the experiment are summarized in Table 4 for decedents in the age range 65–84. Undercoverage bias represents the fraction of N deaths unlinked to any households, and underreporting bias represents the fraction of linked deaths that are unreported in the survey. The total undercount of decedents is

$$\text{total undercount} = (1 - g) + g(1 - f),$$

where g is undercount bias and f is underreporting bias.

Total percentage undercounts are substantially less for network sampling than for conventional

Table 4 Comparison of three counting rules

Counting rule	Total undercount	Undercoverage bias	Underreporting bias
1	0.29	0.22	0.09
2	0.22	0.16	0.07
3	0.15	0.06	0.07

sampling. The undercounts are about a half and a third smaller for rules 2 and 3 respectively than for rule 1. Most of the undercount differences are accounted for by differences in the undercoverage rate, but underreporting bias was also less for network sampling.

Rule 1 failed to cover any of the institutional deaths representing about 29% of North Carolina deaths in this age group. On the other hand, rule 3 covered about two-thirds of the institutional deaths because they were survived by relatives residing in the decedents' county of terminal residences. Rule 2 missed about 16% of all deaths that were not survived by relatives residing in the decedents' counties of terminal residence.

About 7% of all deaths were unreported by households eligible to report them by rules 2 and 3, and 9% by rule 1. Nearly all the unreported deaths occurred at terminal decedent residences without any surviving relatives when the survey was conducted. This is not a particularly surprising finding after one realizes that the event of death itself often precipitates household dissolution and dislocation.

In this example, sampling error effects as well as the measurement biases favor network sampling. The sample design effect of rule 3 compared with rule 1 is substantially less than one, implying that network sampling would attain equivalent precision comparable with conventional sampling with a substantially smaller sample size.

Survey Costs

Since the yields are greater and the interviews are longer, data collection costs are greater for surveys based on network sampling than for those based on conventional sampling. Consequently, to be cost-effective, network sampling must be more efficient than conventional sampling, a necessity most likely to be realized when conventional sampling is subject to large sampling and/or measurement errors.

Assume a sampling frame of L households containing N individuals with a specified attribute. A network sample survey is conducted to estimate N based on a simple random sample of l households. A simplified version of the expected field costs model proposed by Sirken [25] is

$$C = lc,$$

where the unit cost is

$$c = c_1 + \overline{\lambda}\overline{s}c_2$$

in which c_1 is the expected cost of contacting a household, c_2 the expected cost of enumerating an individual linked to a household, $\overline{\lambda} = N|L$, and $\overline{s} = (1/N)\sum_i^N S_\alpha$.

What gains in data quality would be required to overcome the lower field costs of conventional sampling? Assuming measurement biases are a standoff, and that none of the elements is linked to more than one selection unit, network sampling is a cost-effective design alternative to traditional sampling when the following inequality is satisfied:

$$\overline{s} < 1 - (1 - \overline{\lambda})(1 - \theta),$$

where θ is ratio of conventional sampling and network sampling unit field costs. If $\overline{\lambda}/\theta$ is small, network sampling enhances design efficiency whenever $\overline{s} < \theta$.

Values of the parameters $\overline{s}$ and θ vary depending on the particular set of network sampling options that is selected for the counting rule, the estimator, and the respondent rule. For example, the parametric values would be different for an option set comprising the sibling counting rule, the multiplicity A estimator, and a household respondent rule allowing proxy respondents than they would be for an option set comprising the conventional/sibling rule, the multiplicity B estimator, and a self-respondent rule disallowing any proxy respondents. The optimum network sampling design option set represents that particular option set from among all feasible network sampling option sets for which the $\overline{s}$ is smallest relative to its θ. Unless $\overline{s} > \theta$ for the optimum network sampling option set, network sampling would be more cost-effective than conventional sampling.

References

[1] Birnbaum, Z.W. & Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates, *Vital and Health Statistics*, PHS Publication No. 1, Series 2, No. 11. US Government Printing Office, Washington.

[2] Casady, R.J. & Sirken, M.G. (1980). A multiplicity estimator for multiple frame sampling, *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 601–605.

[3] Casady, R.J., Nathan, G. & Sirken, M.G. (1985). Alternative dual system network estimators, *International Statistical Review* **53**, 183–197.

[4] Czaja, R., Warnecke, R.B. Eastman, E., Royston, P., Sirken, M. & Tuteur, D. (1984). Locating patients with rare diseases using network sampling: frequency and quality of reporting, in *Proceedings of the Fourth Conference on Health Survey Research Methods*. Public Health Publication No. 84–3346, National Center for Health Service Research, US Department of Health and Human Services, pp. 311–324.

[5] Czaja, R.F., Snowden, C.B. & Casady, R.J. (1986). Reporting bias and sampling errors in a survey of a rare population using multiplicity counting rules, *Journal of the American Statistical Association* **81**, 411–419.

[6] Faulkenberry, D. & Garoui, A. (1991). Estimating a population total using an area frame, *Journal of the American Statistical Association* **86**, 445–449.

[7] Kramm, E.R., Crane, M.M., Sirken, M.G. & Brown, M.L. (1962). A cystic fibrosis pilot study in three New England states. *American Journal of Public Health* **52**, 2041–2057.

[8] Hendricks, W.A., Searles, D.T. & Horvitz, D.G. (1965). A comparison of three rules for associating farms and farmland with sample area segments in agriculture surveys, in *Estimation of Areas in Agricultural Statistics, Food and Agriculture*. Organization of the United Nations, Rome, pp. 191–198.

[9] Laska, E.M., Meisner, M., Wanderling, J.A. & Kushner, H.B. (1995). Estimating population sizes when duplicates are present, *Statistics in Medicine* **15**, 1635–1646.

[10] Levy, P.S. (1977). Optimum allocation in stratified random network sampling for estimating prevalence of attributes of rare populations, *Journal of the American Statistical Association* **72**, 758–763.

[11] Levy, P.S. (1977). Estimation of rare events by simple cluster sampling with multiplicity, in *American Statistical Association 1977 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 963–966.

[12] Levy, P.S. & Lemeshow, S.A. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

[13] Levy, P.S. & Sirken, M.G. (1972). Quality control of statistical reports. in *American Statistical Association 1972 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 356–359.

[14] Marks, E. & Ockay, C. (1978). A model for network (multiplicity): estimation of census under coverage, in *American Statistical Association 1978 Proceedings of the Section on Survey Research Method*. American Statistical Association, Alexandria.

[15] Nathan, G., Schmeltz, O. & Kenvin, J. (1977). Multiplicity Study of Marriages and Births in Israel, *Vital and Health Statistics*, Series 2, No. 78. DHEW Publication No. (PHS) 79–1352. US Government Printing Office, Washington.

[16] Rittenhouse, J.D. & Sirken, M.G. (1981). A note on networks, nominations, and multiplicity, as contributory to heroin estimation. *Administrative Report*, National Institute of Drug Abuse, Department Health and Human Services, Washington.

[17] Royston, P.N., Sirken, M.G. & Bergsten, J. (1978). Bias and sampling errors and mortality counts based on network sampling. *American Statistical Association 1978 Proceedings of the Section on Social Statistics, American Statistical Association*, Alexandria, pp. 471–475.

[18] Shepard, D.S. & Neutra, M.E. (1977). A pitfall in sampling medical visits, *American Journal of Public Health* **67**, 743–750.

[19] Sirken, M.G. (1970). Household surveys with multiplicity, *Journal of the American Statistical Association* **65**, 257–266.

[20] Sirken, M.G. (1972). Stratified sample surveys with multiplicity, *Journal of the American Statistical Association* **67**, 224–227.

[21] Sirken, M.G. (1972). Variance components of multiplicity estimators, *Biometrics* **28**, 869–873.

[22] Sirken, M.G. (1975). The counting rule strategy in sample surveys, in *American Statistical Association 1975 Proceedings of the Section on Social Statistics, American Statistical Association*, Alexandria, pp. 119–123.

[23] Sirken, M.G. (1978). Dual system estimators based on multiplicity estimators, in *Developments in Dual System Estimation of Population Size and Growth*. University Alberta Press, Edmonton, Alberta, pp. 81–91.

[24] Sirken, M.G. (1979). A dual system network estimator, in *American Statistical Association 1979 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 340–342.

[25] Sirken, M.G. (1983). Handling missing data by network sampling, in *Incomplete Data in Sample Surveys, Part III*, Vol. 2. Academic Press, New York, pp. 81–90.

[26] Sirken, M.G. & Goldstein, S. (1973). Use of multiplicity rules in surveys of Jewish populations, in *Proceedings of Sixth World Congress of Jewish Studies*. Hebrew University, Jerusalem, Israel, pp. 47–57.

[27] Sirken, M.G. & Levy (1973). Multiplicity estimation of proportions based on ratios of random variables, *Journal of the American Statistical Association* **64**, 65–73.

[28] Sirken, M.G. & Nathan, G. (1988). Hybrid network estimators, in *American Statistical Association 1988 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 459–461.

[29] Sirken, M.G. Royston, P.N. & Bridges, M.P. (1977). Counting rule bias in household surveys of decedents, in *American Statistical Association 1977 Proceedings of the Section on Social Statistics*. American Statistical Association, Alexandria, pp. 347–351.

[30] Sirken, M.G., Graubard, B.L. & McDaniel, M.J. (1978). National network surveys of diabetes, in *American Statistical Association 1978 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 631–635.

[31] Sirken, M.G., Shimizu, I. & Judkins, D. (1995). Population based establishment surveys. in *American Statistical Association 1995 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, pp. 470–473.

[32] Sudman, S., Sirken, M.G. & Cowan, C.C. (1988). Sampling rare and elusive populations, *Science* **240**, 991–996.

[33] Thompson, S.K. (1992). *Sampling*. Wiley, New York.

MONROE SIRKEN

Neural Network

Modern regression and classification techniques (*see* **Classification, Overview**) have seen a rapid development over the last 15 years. We hear of names such as smoothers (*see* **Nonparametric Regression**), **projection pursuit**, **additive model**, CART (*see* **Tree-structured Statistical Methods**), MARS, belief networks, and many more, each designed to accommodate some of the deficiencies of our more traditional linear-model-based regression techniques. *Neural networks* appeared in the early 1980s and their coming established a new and popular branch of applied statistical modeling, practiced mainly in the computer science and engineering community. Neural network models tend to be far more ambitious than traditional statistical models, and more successful on large-scale problems. The initial response from the statistics community was either rejection or heavy skepticism. At the time of this writing neural networks have flourished for 13 years, and are used successfully in a large variety of applications including face recognition, handwriting and speech recognition, stock market prediction, medical diagnosis (*see* **Computer-aided Diagnosis**), system control, and genetic modeling.

While neural networks probably get more attention than they deserve in the scientific community at large, they in turn get less attention than they deserve from statisticians. This article gives an overview of the technology, and attempts to place it in the broad context of flexible regression and classification. Along with these models comes a lot of new, redefined, and often attractive terminology, which we emphasize in the text in *italics*.

Figure 1 depicts a neural network with three predictors or *inputs*, a single hidden layer of four hidden units, and an output layer of two responses or *output* units.

If we denote the vector of p inputs by $\mathbf{x}$, and the vector of K outputs by $\mathbf{y}$, then this model can be written more traditionally as

$$z_j = \sigma(\alpha_{j0} + \alpha_j^{\mathrm{T}}\mathbf{x}), \quad j = 1, \ldots, m = 4,$$
$$\hat{y}_k = f_k(\beta_{k0} + \beta_k^{\mathrm{T}}\mathbf{z}), \quad k = 1, \ldots, K = 2 \quad (1)$$

1. The *activation function* σ is used to introduce a nonlinearity at the hidden layer, and is often taken to be the sigmoid $\sigma(z) = 1/(1 + \mathrm{e}^{-z})$.

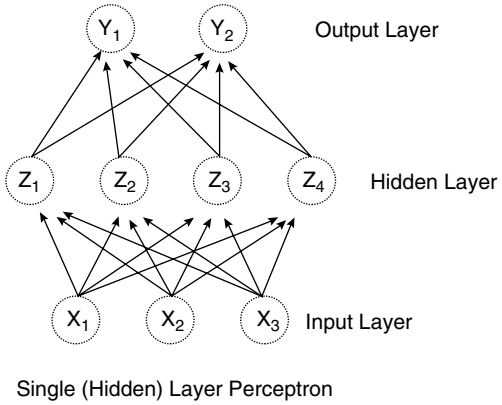

Figure 1 A network diagram represents a neural network model with three inputs (predictors), four hidden units, and two outputs (responses). This configuration is often referred to as a single layer perceptron

2. The parameters α_{jl} and β_{kj} are known as *weights*, and define linear combinations of the input vector $\mathbf{x}$ and hidden unit output vector $\mathbf{z}$, respectively.
3. The intercepts α_{j0} and β_{k0} are known as *biases*.
4. The function f_k permits a final transformation of the output, and the typical choices are:
   (i) $f_k(v) = v$: identity, suitable for regression with quantitative responses;
   (ii) $f_k(v) = 1/[1 + \exp(-v)]$: inverse logit, suitable when responses should lie in [0, 1] (as in two-class nonparametric **logistic regression** problems);
   (iii) $f_k(\mathbf{v}) = \mathrm{e}^{v_k} / \sum_{l=1}^{K} \mathrm{e}^{v_l}$: inverse multiple logit, used for k-class classification. Note here that each f_k requires the entire vector of outputs $\mathbf{v}$.

Neural networks can have more than one hidden layer of units. A *multilayer perceptron* (MLP) simply repeats the hidden layer several times, creating even more complex models. In what follows we focus on the *single layer perceptron* (SLP).

Each hidden unit can be thought of as a nonlinear basis function which creates a new derived variable z_j from a linear combination of the inputs. The responses are then regressed on these transformed data z_j either linearly or via logistic regression. When the model is *learned* or fit to the data, an optimal set of basis functions is learned at the same time.

Early *perceptron* models used *hard or binary thresholding* functions $\sigma(z) = 1(z > 0)$ at the hidden units, with strong neurophysiological implications–the hidden unit (neuron) either fires or does not in response to its inputs. This biological interpretation is not taken too seriously today, and soft threshold functions such as the sigmoid permit differentiation and smoother learning algorithms.

Versions of neural networks look like familiar statistical models:

1. linear regression, when $\sigma(z) = z$, as long as $m \geq p$;
2. logistic regression with two or more classes; again, $\sigma(z) = z$ and the appropriate versions of f_k are used;
3. series and basis expansion models, where we regress the response on a few appropriately placed basis functions.

In addition, there are interesting comparisons with the *projection pursuit* model [3]:

$$\hat{y}_k = \sum_{j=1}^{m} f_{kj}(\alpha_{kj}^{\mathrm{T}}\mathbf{x}),$$

where the functions f_{kj} and the directions α_{kj} are fit simultaneously. The projection pursuit model typically invests many parameters on each function f_{kj} for a given direction α_{kj}, while the neural network can be seen to use exactly two.

Learning Neural Network Models

The computational paradigm central to most neural network systems is *backpropagation*, which is essentially gradient descent using the chain rule plus a few bells and whistles.

Suppose we use least squares on a sample of *training data* to learn the parameters or *weights* in (1):

$$R(\alpha, \beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}(y_k^i - \hat{y}_k^i)^2,$$

a criterion nonlinear in the parameters. Since we anticipate using gradient descent algorithms, we require derivatives. All the derivatives will be sums over the N observations, and we display only the ith component (denoted by superscript i):

$$\frac{\partial R^i}{\partial \beta_{kj}} = -2(y_k^i - \hat{y}_k^i) f_k'(\beta_k^{\mathrm{T}}\mathbf{z}^i) z_j^i,$$

$$\frac{\partial R^i}{\partial \alpha_{jl}} = -\sum_{k=1}^{K} 2(y_k^i - \hat{y}_k^i) f_k'(\beta_k^{\mathrm{T}}\mathbf{z}^i)\beta_{kj}\sigma'(\alpha_j^{\mathrm{T}}\mathbf{x}^i) x_l^i.$$

Given these derivatives, a gradient update at the $(r+1)$th iteration has the form

$$\beta_{kj}^{(r+1)} \leftarrow \beta_{kj}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R^i}{\partial \beta_{kj}^{(r)}},$$

$$\alpha_{jl}^{(r+1)} \leftarrow \alpha_{jl}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R^i}{\partial \alpha_{jl}^{(r)}},$$

where γ_r is the *learning rate* which can change with iteration number r. The standard initialization is to use *random* (Gaussian) starting values for the parameters.

On-line learning refers to a similar gradient descent algorithm, but where we take a separate gradient step in response to each observation one at a time (as opposed to the *batch* mode as presented). The algorithm then has two distinct phases of operation in response to each new training pair $(\mathbf{x}^i, \mathbf{y}^i)$:

1. the *feed-forward* phase, in which $\mathbf{x}^i$ is filtered up through the network to produce a prediction $\hat{\mathbf{y}}^i$:
2. the *backpropagation* phase in which the error $(\mathbf{y}^i - \hat{\mathbf{y}}^i)$ is filtered back and apportioned to each of the coefficients, which are modified in turn in a small way to reduce such a future error.

There are many different variants of this basic approach – some examples are:

1. Batch algorithms typically use second derivative information as well as gradients.
2. On-line algorithms are useful for very large datasets, since changes are made continuously without having to cycle through the entire training data set. The name *on-line* refers to their use in dynamical systems which respond continuously to a changing environment.

3. When the responses are categorical, a different criterion such as likelihood can be used for fitting, but the basic algorithm is the same.

Neural network models are typically overparameterized, often with more parameters than observations. Furthermore, the parameters are intrinsically aliased – in fact, the model is perfectly symmetrical with respect to the hidden units. The symmetry is resolved by using random starting values, similar in flavor to the fitting of mixture models. Although the overparameterization can be controlled somewhat by restricting the number of hidden units, other strategies for regularization are popular, as follows.

Early Stopping. Gradient descent will converge slowly to a local minimum, which in a saturated model will fit the training data perfectly. If an independent validation set is available, then one can monitor the performance of the model on this validation set after each one or set of updates. The usual bias-variance tradeoffs will cause this validation error to eventually increase as the training error approaches zero, and will provide an optimal place to stop.

Weight Decay. Similar to **ridge regression**, various proposals exist for shrinking the weights towards zero (*see* **Shrinkage**). As a simple example, consider adding a penalty of the form $\lambda(\sum \alpha_{kj}^2 + \sum \beta_{jl}^2)$ to the criterion R. This introduces a hyperparameter λ, which shrinks between the unrestricted model ($\lambda = 0$) and the constant model ($\lambda = \infty$). This approach is similar in flavor to cubic smoothing splines (*see* **Spline Function**) [4]. A large basis of cubic splines is created, with a knot at each data point, but then the large number of coefficients is fit subject to the rather stringent smoothness constraints of the derived function.

Model Averaging. Even when the model is not overfit, the criterion R has multiple minima, and the random starting weights will lead to one. One can restart the algorithm many times, and then combine the solutions in some way. Simple averaging works well; [5] proposes an approximate form of Bayesian posterior averaging.

Discussion

The *field* of neural networks is large, and there are many large and important conferences each year, which at their inception in the mid-1980s were devoted to various aspects of the model described in this chapter. The annual NIPS Neural Information Processing Conference in Denver, Colorado is one such venue. The mid-1990s versions of these conferences would be more aptly described as applied statistics than neural networks. Neural network models tend to be currently viewed as one of a number of flexible regression models.

There have been many ingenious modifications and restrictions to the neural network model to broaden its range of applications. Two examples are *bottleneck* networks for nonlinear principal components (the inputs are also the outputs), and networks with duplicated weights to mimic autoregressive models.

While this field is strongly application driven, some interesting theory has been produced [1]. For example, neural networks are *universal approximators* – given enough parameters they can approximate smooth functions arbitrarily well.

This author has seen many examples of inappropriate use of neural network models. For example, their use is unlikely to be appropriate in small biostatistical binary regression problems, where the primary goal is to understand the effect of the inputs on the event represented by the binary response. The converse problem, of course, abounds as well. How often do we struggle to fit large traditional parametric models to huge datasets with our inappropriate software. When prediction is a goal, the *black box* neural network model will often deliver close to the best fit with very few tears.

There are many commercial software packages available for fitting neural networks, and many free packages as well. An SAS (*see* **Software, Biostatistical**) module exists, and functions for S-PLUS (*see* **S-PLUS and S**) contributed by B.D. Ripley are available from the Statlib archive `http://lib.stat.cmu/S`. For statisticians interested in reading more about neural networks, I strongly recommend the books by Bishop [2] and Ripley [5]. Both of these synthesize the vast neural network literature, contain key references, and use a style and language familiar to statisticians.

References

[1] Barron, A.R. (1991). Universal Approximation Bounds for Superpositions of a Sigmoid Function, *Technical Report*. University of Illinois, Urbana-Champaign.

[2] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

[3] Friedman, J.H. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.

[4] Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[5] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

(*See also* **Computer-intensive Methods**; **Nonlinear Regression**; **Prediction**)

TREVOR HASTIE

Neurology

Neurology is the specialty of medicine which deals with the structure and function of the nervous system, and with its diseases. Although the word "neurology" was introduced as long ago as 1681, this specialty of medicine was closely allied with **psychiatry** until the early decades of the twentieth century – both being concerned with "nerve" disorders. Their proximity is still apparent today within conjoint areas such as neuropsychiatry. The main neurological disorders (approximate point **prevalence** per 100 000) are migraine (20 000), **stroke** (500), **epilepsy** (500), Parkinson's disease (200), cerebral palsy (60), multiple sclerosis (50), and primary brain tumours (50); others, such as syringomyelia (chronic progressive disease of the spinal cord), Huntington's chorea (a hereditary, chronic muscular twitching), polymyositis (simultaneous inflammation of many muscles), muscular dystrophy (progressive wasting and atrophy of muscles), and motor neurone disease (degeneration of motor neurones) are much rarer (each <10 per 100 000); the most interesting biostatistically are stroke and epilepsy.

ANTHONY L. JOHNSON

Neuropathology

Neuropathology is the study of the nature, causes, structure, and function of nervous system diseases. This broad area of medicine includes pathology of cerebrovascular disease (**stroke**, hypoxia, ischemia, aneurysms, hemorrhage), neurodegenerative diseases and dementia, craniocerebral trauma, neurometabolic and demyelinating disorders, as well as many other related topics covered in a standard reference text on the subject such as *Greenfield's Neuropathology* [3]; see also [4]. Complementary aspects of neuropathology have from its beginning been the topographical, anatomical emphasis of the French school and the cellular emphasis on pathogenesis of the German school. Modern technology is helping to integrate these two aspects of the field. Biostatistical methods and applications in neuropathology are as broad as the field itself, ranging from **simple linear regression** models to detailed histological **image analyses**. The most important elements in the nervous system are neurons, their bodies (somata), branches (dendrites) and firing channels (axons), and the central questions of neuropathology are concerned with how these most complex cells in our bodies grow, degenerate, and die by natural aging or disease processes. From a biostatistical perspective, image analysis perhaps serves best to orient one's thinking about a field whose main aim is to help clinical and basic scientists recognize the cellular features of central nervous system diseases and whose central tool remains the light microscope. By special staining techniques one is able in microscopic neuropathologic image analysis to observe complex alterations of neurons, glia and astrocytes, degenerative-reductive changes such as axonal demyelination, and productive-accumulative changes such as neurofibrillary tangles and senile plaques seen in Alzheimer's dementia.

For instance, it is well known that neurons may aggregate in functionally significant ways to form discrete layers within the cerebral cortex of macrocolumnar arrays of neurons that occur vertically across the individual layers [8], and hence it is conceivable that a disorder like schizophrenia, involving disturbances in several key corticolimbic brain areas [1], might involve unusual arrangements of neurons [2] arising from specific neurodevelopmental disturbances [11, 12, 15]. Although schizophrenia is also known currently to be a neurochemical disorder involving dysfunction of glutamatergic circuitry [9], biostatistical image analysis has helped demonstrate significant *post mortem* differences in the arrangements of pyramidal neurons in layer IV of the cingulate cortex for normal control, schizoaffective and schizophrenic subjects [5]. Employing a second-moment estimator for stationary **point processes** [10] and a modified **bootstrap** procedure for statistical **inference**, a **stereological** analysis of these spatial point patterns showed not only that the psychotic subjects had fewer pyramidal neurons per cubic micrometer in this brain region but also that these cells tended to be more regularly spaced (nonoverlapping) for these subjects. This finding suggested that there may be increased inhibitory distances among these neurons, and, therefore, that there may be a relative expansion of the neuropil surrounding these cells. Since the neuropil is the site where most neural connections are found, this finding implied further that the pyramidal neurons of the cingulate cortex in schizophrenics may be engaging in more synaptic connections than those of normal individuals.

Neuropathology also uses macroscopic imaging modalities such as computed tomography and magnetic resonance imaging for studies of brain structure, and functional magnetic resonance imaging and single photon emission computed tomography for studies of brain function, to provide data for detailed yet lower resolution analyses. Volumetric studies by computed tomography and magnetic resonance imaging help quantify both **cross-sectional** and **longitudinal** changes in sizes and shapes of various brain structures over time for the same individual or cohorts of individuals. The growing literature on differences between normal aging, Alzheimer's dementia, and vascular dementia provide a further example of the biostatistics of neuropathology. For instance, a lack of neuroimaging data may have contributed to an underrecognition of mixed Alzheimer's and vascular dementia cases in community-based studies [6, 14]. An analysis of computed tomography brain scans from histopathologically confirmed Alzheimer's dementia cases [7] reported a 10-fold increase in the yearly rate of medial temporal lobe atrophy relative to losses experienced by **control** subjects, from 1.5% per year to 15.1% per year. This excessive atrophy was evidence for a neuropathological cascade process and not a simple acceleration of the normal aging process, and was

confirmed through use of stereological techniques that compared regional patterns of neuronal cell loss in the hippocampus related to normal aging to that associated with Alzheimer's dementia [13], providing additional information on neurodegenerative processes involved.

These two brief examples of current neuropathology research have been included here, and many more could have been cited, to demonstrate to the reader that new technologies such as those involved in microscopic cellular imaging and macroscopic structural and functional brain imaging enable neuroscientists to ask and to answer entirely new quantitative questions regarding the neuropathology of the developing brain, the mature brain, and the aging brain. It is anticipated that, by further evolution, the predominantly qualitative findings of the past will continue to be enlarged, discarded and refined through the kinds of detailed quantitative result provided by biostatistical analyses.

References

[1] Beneš, F.M. (1988). Post-mortem structural analyses of schizophrenic brain: study designs and the interpretation of data, *Psychiatric Developments* **3**, 213–226.

[2] Beneš, F.M. & Bird, E.D. (1987). An analysis of the arrangement of neurons in the cingulate cortex of schizophrenic patients, *Archives of General Psychiatry* **44**, 608–616.

[3] Blackwood, W. & Corsellis, J.A.N., eds (1992). *Greenfield's Neuropathology*. Edward Arnold, Chicago.

[4] Brumback, R.A. & Leech, R.W. (1995). *Neuropathology and Basic Neuroscience*. Springer-Verlag, New York.

[5] Diggle, P.J. Lange, N. & Beneš, F.M. (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy, *Journal of the American Statistical Association* **86**, 618–625.

[6] Hebert, R. & Brayne, C. (1995). Epidemiology of vascular dementia, *Neuroepidemiology* **14**, 240–257.

[7] Jobst, K.A., Smith, A.D., Szatmari, M., Esiri, M.M., Jaskowski, A., Hindley, N., McDonald, B. & Molyneux, A.J. (1994). Rapidly progressing atrophy of medical temporal lobe in Alzheimer's disease, *Lancet* **343**, 829–830.

[8] Mountcastle, V.B. (1979). An organizing principle for cerebral function: the unit module and the distributed system, in *The Neurosciences Fourth Study Program*, F.O. Schmitt & F.G. Worden, eds. MIT Press, Cambridge, Mass., pp. 21–42.

[9] Olney, J.W., Sesma, M.A. & Wozniak, D.F. (1993). Glutamergic, cholinergic, and GABAergic systems in posterior cingulate cortex: interactions and possible mechanisms of limbic disease, in *Neurobiology of Cingulate Cortex and Limbic Thalamus*, B.A. Bogt & E.M. Gabriel, eds. Birkhäuser, Boston, pp. 557–580.

[10] Ripley, B.D. (1976). The second order analysis of stationary point processes. *Journal of Applied Probability* **13**, 255–266.

[11] Weinberger, D.R. (1987). Implications of normal brain development for the pathogenesis of schizophrenia, *Archives of General Psychiatry* **44**, 660–669.

[12] Weinberger, D.R. (1995). From neuropathology to neurodevelopment, *Lancet* **346**, 552–557.

[13] West, M.J., Coleman, P.D., Flood, D.G. & Troncoso, J.C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease, *Lancet* **344**, 769–772.

[14] White, L. Petrovitch, H. Ross, G.W., Masaki, K., Abbott, R.D., Teng, E.L., Rodriguez, B.L., Blanchette, P.L., Havlik, R.J., Wergowske, G., Chiu, F., Foley, D.J., Murdaugh, C. & Curb, J.D. (1996). Prevalence of dementia in older Japanese-American men in Hawaii: the Honolulu-Asia Aging Study, *Journal of the American Medical Association* **276**, 955–960.

[15] Wolf, S.S. & Weinberger D.R. (1996). Schizophrenia: a new frontier in developmental neurobiology, *Israel Journal of Medical Science* **32**, 51–55.

NICHOLAS LANGE

Neyman, Jerzy

Born: April 16, 1894, in Bendery, Russia.
Died: August 5, 1981, in Berkeley, California, USA.

Reproduced by permission of the Royal Statistical Society

Jerzy Neyman, one of the principal architects of modern statistics, was Director of the Statistical Laboratory, University of California, Berkeley. He was born into a Polish family in Bendery, and died in Berkeley at the age of 87. With Neyman's passing, history has closed a chapter on the early development of this important scientific field.

At the time of his birth, there was no Poland as a nation. The "Poland proper" had been divided among Germany, Austria, and Russia. Neyman's father was a lawyer. When Neyman was 12 years old, his father died of a heart attack. His caring mother moved her family to Kharkov, where he attended school and college. Although he was born a Pole, Neyman spoke Russian almost as early as he spoke Polish. At an early age, he could also speak Ukrainian, German, French, and Latin fluently. Upon his graduation from high school, through his mother's arrangement, he joined a student group making a journey to see Europe outside Russia. Before entering the college in Kharkov, he decided to study mathematics instead of pursuing his father's profession. He received his mother's support and encouragement. "She had respect for intellectual activity", Neyman fondly recalled to Constance Reid in the late 1970s. (Reid published her book entitled *Neyman – From Life* in 1982 [33].) In 1921, after a Polish–Soviet peace treaty, Neyman was sent to Poland in a repatriation of prisoners of war program between the two countries. Thus Neyman saw his fatherland Poland for the first time when he was 27 years old!

Neyman's interest in mathematics was reinforced when he studied with the Russian probabilist S.N. Bernstein at the University of Kharkov. When he read Henri Lebesgue's *Leçons sur l'intégration et la recherche des functions primitives*, Neyman was fascinated by sets, measure, and integration. During his college days he had proved five theorems on the Lebesgue integral on his own. His article entitled "Sur une théorème metrique concernant les ensembles fermés", published in 1923 [9], was one of his early research papers in pure mathematics. His candidate thesis at the University of Kharkov (1916) was on the integral of Lebesgue. In 1917, Neyman returned to the university for a postgraduate study. In the following year he was a *docent* at the Institute of Technology, Kharkov. At the University of Warsaw, Neyman studied mathematics with Waclaw Sierpinski. He earned the Doctor of Philosophy degree from the University of Warsaw in 1924. The oral examination consisted of *Rigorosum Major* in mathematics and *Rigorosum Minor* in philosophy. No one knew more statistics than Neyman to examine him on the subject.

In the little spare time that he had during his student days, Neyman was heavily involved in teaching to earn a living. He also gave supplementary lectures for professors at the university, and taught mathematics and statistics to college students.

Neyman first heard of **Karl Pearson** from his reading of Pearson's book. *The Grammar of Science* [32]. Apparently, he was influenced by Pearson's philosophical views expressed in the book.

Neyman's contact with statistics occurred early in his academic career. It appears that he had studied applications of mathematical statistics with Bernstein at the University of Kharkov. But he learned most statistics through his work on his own, especially in agricultural experimentation. He had held a position of "senior statistical assistant" at the National Agricultural Institute in Bydgoszcz, Poland, in 1921, and he was a special lecturer at the Central College of Agriculture in Warsaw in 1922.

In the fall of 1925, Sierpinski and Kazimierz Bassalik, the director of the National Agricultural Institute, were awarded a Polish Government Fellowship for Neyman to study mathematical statistics with Karl Pearson in London. Neyman was well prepared in mathematics and in statistics. While in London, Neyman and a young man about his own age, Pearson's son, **Egon S. Pearson**, became good friends.

During the academic year 1926–27, Neyman was on a Rockefeller fellowship to study pure mathematics in Paris. He attended lectures given by Emile Borel at the University of Paris and also lectures by Lebesgue and Jacques Hadamard at the Collège de France. In addition, he had some of his own notes read at these institutes. Quite possibly, the year of studying mathematics in Paris had prepared him well for his joint endeavor with Egon Pearson in the development of statistical theory in the years to come.

Neyman and Pearson's joint work formally started in the spring of 1927, when Pearson visited Neyman in Paris. While there are no records of what transpired during the ten days during which they worked together, they must have laid out plans for their future joint project. At the end of the 1926–27 academic year, Neyman went back to Poland, and in 1928 he became head of the Biometric Laboratory at the Nencki Institute in Warsaw. He carried out his joint work with Pearson mostly through correspondence between Warsaw and London. Between 1928 and 1934, they published seven of their 10 most important papers on the theory of testing statistical hypotheses [20–26] (*see* **Hypothesis Testing**).

In developing their theory, Neyman and Pearson recognized the need to include **alternative hypotheses**; and they perceived the errors in testing hypotheses concerning unknown population values based on sample observations which are subject to variation. They called the error of rejecting a true hypothesis the first kind of error, and the error of accepting a false hypothesis the second kind of error. They placed the importance on the probability of rejecting a hypothesis when it is false. They called this probability the **power** of a test. They proposed a term "**critical region**" to denote a set of sample statistic values leading to the rejection of the hypothesis being tested. The "size" of a critical region is the probability of making the first kind of error, which they called the level of significance.

They called an hypothesis which completely specifies a probability distribution a simple hypothesis. An hypothesis which is not a simple hypothesis is a composite hypothesis. An hypothesis concerning the **mean** of a **normal distribution** with a known **standard deviation**, for example, is a simple hypothesis. The hypothesis is a composite hypothesis if the standard deviation is unknown.

It is now difficult for us to imagine how one could perform a statistical test without these concepts. But the Neyman–Pearson theory was a considerable departure from traditional hypothesis testing at the time. They were severely criticized for their new theory by the leading authorities of the field, especially by **R.A. Fisher**.

Neyman and Pearson used conceptual mathematics and logical reasoning to develop the theory of hypothesis testing. They emphasized "the importance of placing in a logical sequence the stages of reasoning in the solution of ... inference". In their initial papers [20, 21], it seems that they were leading the reader, step by step, in their development of the theory. They relied on the concept of the **likelihood ratio** in testing hypotheses concerning parameters in known probability distributions; and they elucidated their ideas further with specific examples and numerical computations.

After they had laid a solid mathematical foundation for their theory, they applied it to the problem of two samples [22], and to the problem of k samples [24]. In one of their joint papers [26], they used the likelihood ratio to establish an objective criterion for determining the best (in the sense of power of test) critical regions for testing a simple hypothesis and a composite hypothesis. That was a high point of their accomplishments. The landscape of statistical hypothesis testing would no longer be the same.

In 1934, Neyman joined the faculty of E.S. Pearson's Department of Applied Statistics at University College London. Between 1934 and 1938 they published only three more joint papers on testing hypotheses [27, 28], possibly because of Pearson's involvement in administrative responsibilities. Neyman, however, was still very productive during that period. From time to time, Neyman published many papers on hypothesis testing on his own; but most of the fundamental work was contained in his joint publications with Pearson.

When he was still in Poland, Neyman had developed the idea of **confidence interval** estimation. He

even gave his lectures on confidence interval estimation rather than hypothesis testing in his class at University College London in 1934. He published his work in 1937 [10]. At that time, many statisticians confused the confidence interval with the **fiducial** interval, a concept developed by Fisher. That confusion was soon dispelled by Fisher himself [3]. Neyman clarified the difference between the two in his *Lectures and Conferences* [11].

To put it in very simple terms, the difference between confidence interval and fiducial interval lies in the assumption regarding the population value being estimated. Consider a sample of size n and mean $\overline{X}$ from a normal distribution with an unknown mean u and unit **variance**. The basic quantity for estimating u is the product $n^{1/2}(\overline{X} - u)$. According to the confidence interval theory, the product $n^{1/2}(\overline{X} - u)$ is subject to variation, because the sample mean $\overline{X}$ is a random variable. For a given probability, say 0.95, $\Pr\{\overline{X} - 1.96/\sqrt{n} < u < \overline{X} + 1.96/\sqrt{n}\} = 0.95$. After a sample is taken and the sample mean (say $\overline{X}_0$) is determined, the product becomes $n^{1/2}(\overline{X}_0 - u)$, which is not subject to variation. The interval $\{\overline{X}_0 - 1.96/\sqrt{n},\ \overline{X}_0 + 1.96/\sqrt{n}\}$ becomes a confidence interval, and 0.95 becomes the corresponding confidence coefficient. In the fiducial interval argument, there is a range of values of u, each of which could have generated the sample mean $\overline{X}_0$. For a given probability 0.95, one can find two values of u depending on $\overline{X}_0$, say $u' < u''$, such that fiducial $\Pr\{u' < u < u''|\overline{X}_0\} = 0.95$. The interval (u', u'') is the fiducial interval and 0.95 is the fiducial probability.

In addition to the theory of statistical inference, Neyman made contributions to many other branches of statistics, such as the design of agricultural experimentation (in 1923, 1925, and 1935), the theory of sampling (in 1925, 1938 and 1939), a class of **"contagious" distributions** (1939), and others. He even used the "storks bring babies" example to show how to avoid reaching a wrong conclusion by misusing a **correlation** between variables, the so-called spurious correlation [11].

Neyman's work on applications of statistical methods in practical problems was very extensive. He considered practical problems as a source of inspiration for theoretical statisticians. His publications related to **biostatistics** and health include: virulent bacteria and disease (with R. Iwaskiewicz [5]); recovery and relapse of cancer patients (with E. Fix [4]); accident proneness (with G.E. Bates [1]); a **stochastic** model of an epidemic (with E.L. Scott [30]) (*see* **Epidemic Models, Stochastic**); multiphasic **screening** and diagnosis (with M.F. Collen et al. [2]); health-pollution (in 1972); a view of biometry [15]; energy crisis, pollution, and health (in 1975); environmental pollution and public health [17] (*see* **Environmental Epidemiology**); some problems of biometry deserving particular attention (with R. Bartoszynski et al. [18]); **radiation**-related public health studies (in 1980); probability models in medicine and biology [19]; and on understanding the mechanism of radiation effects (with P.S. Puri [29]).

There was an interesting feature in Neyman's approach to practical problems. He had the ability to visualize the phenomena behind the data and a model of the mechanism that creates the phenomena. He would express the model in mathematical terms to produce new probability distributions, or new stochastic models. Only then would he find appropriate statistical methods with which to analyze the data on hand [7].

Neyman devoted a considerable amount of time and effort to three major projects of research. The first was his joint studies of galaxies with E.L. Scott and C.D. Shane. Over a span of 25 years, Neyman published 24 papers on the subject. He reported their work on the spatial distribution of galaxies (in 1953, with Scott and Shane), on the problem of expansion of clusters of galaxies (in 1954, with Scott), on the statistical approach to the problems of cosmology (in 1958, with Scott), and on the relation of galaxies in clusters in the presence of instability and absorption (in 1961, with Scott).

In a separate effort, Neyman organized and edited the volume *The Heritage of Copernicus* [16] for the National Academy of Sciences, to commemorate the 500th anniversary of the birth of the great Polish astronomer.

The second major project was on cloud seeding and weather modification, jointly with Scott. This project covered a period of over 20 years from the early 1950s to 1980, and was reported in 26 publications. In this project, Neyman had the opportunity to witness several cloud seeding experiments, and to evaluate the designs and the outcomes. Their conclusion was as follows: "Methods of evaluation of the effects of cloud seeding proposed by commercial

operators could not be considered scientific. After a few years and a few experiments nothing remarkable could be asserted" [6].

His third project was on cancer and carcinogenesis, but on a much smaller scale than the other two. This was also a joint effort with Scott. There were two theories regarding the formation of cancerous clones available to them: the one-stage mutation theory and the two-stage mutation theory. According to the one-stage mutation theory, the growth of a clone of abnormal cells is from a single mutant cell; while according to the two-stage theory, a second mutation is necessary in order to produce a cancer clone. Their main objective was to construct stochastic models that would best describe the process of formation of cancerous clones [31] (*see* **Dose–Response Models in Risk Analysis**).

In the spring of 1937, Neyman delivered a series of lectures on mathematical statistics and probability at the Graduate School in the US Department of Agriculture in Washington, DC. That was the first time that the American statistical public had the opportunity to hear statistical theory from Neyman in person. The lecture notes were subsequently published in 1937 [11], and revised and expanded in 1952 [14], under the title *Lectures and Conferences on Mathematical Statistics and Probability*. Among the reviews of the 1937 book, there was one written by William Feller, published in *Zentralblatt*, which reads in part as follows:

> The point of departure for the author is always actual practical problem, and he never loses sight of the applications. At the same time his goal is always a truly rigorous mathematical theory. He appears to insist on absolute conceptual clarity and rigor, not only as a sound foundation, but also because it is really useful and necessary, particularly where the practical problem goes beyond the mathematical aspect. . ..

Feller's words should apply equally well to Neyman's other publications.

In 1938, Neyman accepted a mathematics professorship from the University of California at Berkeley, and he established the Statistical Laboratory, with himself as the director. That was the beginning of one of the preeminent statistical centers in the world. In 1955, Neyman established the Department of Statistics. He retained the title Director of the Statistical Laboratory.

Neyman was a very dynamic person, full of ideas and energy. Soon after the Statistical Laboratory was established and the teaching program was in good order, he began to plan a symposium of mathematical statistics and probability "to mark the end of the war and to stimulate the return to theoretical research". The symposium had the participation of leading authorities in theoretical probability and in mathematical statistics, as well as those in applied fields. The *Proceedings* of the symposium, edited by Neyman [13], were published in 1949 to "stimulate research and foster cooperation between the experimenter and the statistician".

The success of the symposium prompted Neyman to plan a series of symposia, once every five years. The number of participants and the coverage grew from one symposium to the next. The Sixth Berkeley Symposium, held in three different periods in 1970 and 1971, was attended by 240 leading authors in 33 subject areas in the theory of probability, in mathematical statistics, and in scientific fields using applications of statistics. The *Proceedings*, edited by LeCam et al. [8], were published in 1972, in six volumes and 3397 pages – a gigantic undertaking!

These symposia supplemented the teaching programs and research activities normally carried out in universities and other academic institutions. They also had a great deal of influence on the attitude of the theoretical statisticians and research scientists, making them recognize the need for and the advantage of applications of statistics.

During the 40 years during which he was in Berkeley, Neyman had students coming from all over the world to attend his lectures and to learn the proper way of conducting research. Neyman was a generous man. He helped students financially in any way he could. He recommended students for the university scholarships and he secured Federal grants for the support of both students and the faculty. At times, when he could not obtain the funds that he needed to support students from any other sources, Neyman took the money out of his own pocket!

Neyman was a member of the National Academy of Sciences, the American Academy of Arts and Sciences, the Royal Swedish Academy, and the National Academy of Poland. He was a fellow of the Royal Society of London, and he was honorary president of the **International Statistical Institute**.

Neyman used to say: "Statistics is the servant to all sciences". In many ways, Neyman had expanded the domain and improved the quality of the service.

References

[1] Bates, G.E. & Neyman, J. (1952). Contribution to the theory of accident proneness. I. An optimistic model of correlation between light and severe accidents, in *University of California Publications in Statistics*, Vol. I, pp. 215–254.

[2] Collen, M.F., Rubin, L., Neyman, J., Dantzig, G.B., Baer, R.M. & Siegelaub, A.B. (1964). Automated multiphasic screening and diagnosis, *American Journal of Public Health* **54**, 741–750.

[3] Fisher, R.A. (1936). The fiducial argument in statistical inference, *Annals of Eugenics* **6**, 391–398.

[4] Fix, E. & Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients, *Human Biology* **23**, 205–241.

[5] Iwaszkiewicz, R. & Neyman, J. (1931). Counting virulent bacteria and particles of virus, *Acta Biologiae Experimentalis* **6**, 101–142.

[6] LeCam, L. (1979). Neyman, Jerzy in *International Encyclopedia of Social Sciences*, *Biographical Supplement* **18**, 587–590.

[7] LeCam, L. (1995). Neyman and stochastic models, *Probability and Mathematical Statistics* **15**, 37–45.

[8] LeCam, L. Neyman, J. & Scott, E.L., eds (1972). *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vols. I-VI. University of California Press, Berkeley pp. 3397.

[9] Neyman, J. (1923). Sur une théorème metrique concernant les ensembles fermés, *Fundamenta Mathematicae* **5**, 328–330.

[10] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London, Series A* **236**, 333–380.

[11] Neyman, J. (1938). *Lectures and Conferences on Mathematical Statistics*. Graduate School, US Department of Agriculture, Washington.

[12] Neyman, J., (1949). Contribution to the theory of the chi-square test, in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 239–273.

[13] Neyman, J. ed. (1949). *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley. 501 pp.

[14] Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd Ed. Graduate School, US Department of Agriculture, Washington, ix + 350 pp.

[15] Neyman, J. (1974). A view of biometry: an interdisciplinary domain concerned with chance mechanisms operating in living organisms; illustration: urethan carcinogenesis, in *Reliability and Biometry, Statistical Analysis of Lifelength*, F. Prochan & R.J. Serfling, eds. SIAM, Philadelphia, pp. 183–201.

[16] Neyman, J. ed. (1974). *The Heritage of Copernicus: Theories "Pleasing to the Mind"*. MIT Press, Cambridge, Mass.

[17] Neyman, J. (1975). Assessing the chain: energy crisis, pollution and health, *International Statistical Review* **43**, 253–267.

[18] Neyman, J. (1977). Some problems in biometry deserving particular attention, in *Proceedings of the Symposium to Honor Jerzey Neyman*, R. Bartoszynski, E. Fidelis & W. Klonecky, eds. Polish Scientific Publishers, Warszawa, pp. 257–264.

[19] Neyman, J. (1979). Probability models in medicine and biology: avenues for their validation for humans in real life, *Management Science* **25**, 931–938.

[20] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, part I, *Biometrika*, **20-A**, 175–240.

[21] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, part II, *Biometrika* **20-A**, 263–294.

[22] Neyman, J. & Pearson, E.S. (1930). On the problem of two samples, *Bulletin of the Polish Academy, Series A*, 73–96.

[23] Neyman, J. & Pearson, E.S. (1931). Further notes on chi-square distribution, *Biometrika* **22**, 298–305.

[24] Neyman, J. & Pearson, E.S. (1931). On the problem of k samples, *Bulletin of the Polish Academy, Series A*, 460–481.

[25] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London, Series A* **231**, 289–337.

[26] Neyman, J. & Pearson, E.S. (1933). The testing of statistical hypotheses in relation of probabilities *a priori*, *Proceedings of the Cambridge Philosophical Society* **29**, 492–510.

[27] Neyman, J. & Pearson, E.S. (1936). Contribution to the theory of statistical hypotheses testing and critical region of Type A and Type A_1, *Statistical Research Memoirs* **1**, 1–37.

[28] Neyman, J. & Pearson, E.S. (1938). Contribution to the theory of testing statistical hypotheses, part II, *Statistical Research Memoirs* **2**, 25–57.

[29] Neyman, J. & Puri, P.S. (1981). A hypothetical stochastic mechanism of radiation effects in single cells, *Proceedings of the Royal Society of London, Series B* **213**, 139–160.

[30] Neyman, J. & Scott, E.L. (1964). A stochastic model of epidemics, in *Stochastic Models in Medicine and Biology*, J. Gurland, ed. University of Wisconsin Press, Madison, pp. 45–83.

[31] Neyman, J. & Scott, E.L. (1967). Statistical aspect of the problem of carcinogenesis, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics*

and Probability, Vol. 4, L. LeCam & J. Neyman, eds. University of California Press, Berkeley, pp. 745–776.

[32] Pearson, K. (1937). *The Grammar of Science*, 3rd Revised Enlarged Ed. Dutton, New York (a paperback edition was published in 1957 by Meridian).

[33] Reid, C. (1982). *Neyman – From Life*. Springer-Verlag, New York, 298 pp.

CHIN LONG CHIANG

Neyman–Pearson Lemma

In the mid-1920s, **Jerzy Neyman** and **Egon S. Pearson** set out to elaborate a theory of testing statistical hypotheses (*see* **Hypothesis Testing**). This resulted in the publication of two papers [3] in 1928. They proposed as a *principle* the use of **likelihood ratio tests**. Suppose that the observable random vector possesses, with respect to some measure μ, a density $f(x, \theta)$ that depends on a parameter $\theta \in \Theta$. One hypothesis specifies that $\theta \in \mathrm{H}_1$, while the alternative specifies that $\theta \in \mathrm{H}_2$, where the H_i are subsets of Θ. Neyman & Pearson proposed to reject H_1 if the ratio $r(x) = \sup[f(x, \theta); \theta \in \mathrm{H}_2] / \sup[f(x, \theta); \in \mathrm{H}_1]$ exceeds a specified value c determined to insure that the probabilities of rejection of H_1 if true do not exceed a specified limit.

They showed that many of the tests then in use could be derived from that principle. They also introduced the concept of **power** of a test, the probability for $\theta \in \mathrm{H}_2$ that H_1 be rejected.

This, however, was a *principle*. It needed a mathematical justification. To describe the justification with a simple notation, we shall use test functions instead of "critical regions". A test function ϕ is a measurable function defined on the sample and such that $0 \leq \phi \leq 1$. If the measures are noted $P_\theta, \theta \in \Theta$, the power of ϕ at θ is $\int \phi \, \mathrm{d}P_\theta$. Critical regions correspond to functions ϕ that take only values zero or one, being one in the critical region.

The Neyman–Pearson lemma [4] initially covered only the case in which each H_i is reduced to a single point θ_i. It then says the following.

Lemma. Let $r(x) = f(x, \theta_2)/f(x, \theta_1)$ and let ϕ be such that $\phi(x) = 0$ if $r(x) < c$ and $\phi(x) = 1$ if $r(x) > c$. Then, for any other test function ψ such that $\int \psi \, \mathrm{d}P_{\theta_1} \leq \int \phi \, \mathrm{d}P_{\theta_1}$, one has $\int \psi \, \mathrm{d}P_{\theta_2} \leq \int \phi \, \mathrm{d}P_{\theta_2}$.

This lemma allowed Neyman & Pearson [5, 6] to solve the problem of selection of critical regions and justify their likelihood ratio principle in many problems where there exist *uniformly most powerful tests*. That includes the case in which the parameter θ is real and the densities form a family with monotone likelihood ratios (e.g. exponential families) if the hypotheses are of the form $\{\theta \leq a_1\}$ against $\{\theta > a_2\}$. The case of two-sided hypotheses of the kind $\{\theta = 0\}$ against $\{\theta; |\theta| > a_2\}$ was not covered. Neyman & Pearson introduced for such cases a concept of **unbiasedness**; requiring, for instance, that the derivative of the power function vanish at $\theta = 0$.

This led Neyman & Pearson to consider problems of the following type. Given a finite set of functions f_i, $i = 1, 2, \ldots, m+1$, and constants c_i, find test functions ϕ such that $\int \phi f_i \, \mathrm{d}\mu = c_i, i = 1, \ldots, m$, and such that $\int \phi f_{m+1} \, \mathrm{d}\mu$ be maximized. The method of Lagrange multipliers suggests the introduction of positive constants $k_i, i = 1, \ldots, m$, and the sum $s(x) = \sum_{i=1}^{m} k_i f_i$. One then takes $\phi = 0$ where $f_{m+1} < s$ and $\phi = 1$ where $f_{m+1} > s$. For a precise statement, see, for instance, [2, p. 96].

With the advent of Wald's Theory of statistical decision functions [7] (*see* **Decision Theory**), such problems were essentially subsumed under the following general theorem.
If the family of measures $\{P_\theta\}$ is dominated, then all admissible tests are either Bayes solutions or limits of them.

Another more recent form of a Neyman–Pearson lemma is that given by Huber & Strassen [1]. These authors consider sets $\mathrm{H}_i = \{P; P \leq v_i\}$ where the v_i are Choquet capacities alternating of order two, that is such that $v(A \cup B) + v(A \cap B) \leq v(A) + v(B)$. They show that in this case there exist least informative pairs (P_1, P_2) with $P_i \in \mathrm{H}_i$ and that the likelihood ratios of these pairs provide the optimal tests between the H_i.

References

[1] Huber, P. & Strassen, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities, *Annals of Statistics* **1**, 251–263.

[2] Lehmann, E. (1986). *Testing Hypotheses*, 2nd Ed. Wiley, New York.

[3] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I, *Biometrika* **20A**, 175–240; Part II, *Biometrika* **20A**, 263–294.

[4] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London, Series A* **231**, 289–337.

[5] Neyman, J. & Pearson, E.S. (1933). The testing of statistical hypotheses in relation to probabilities *a priori*, *Proceedings of the Cambridge Philosophical Society* **24**, 492–510.

[6] Neyman, J. & Pearson, E.S. (1938). Contributions to the theory of testing statistical hypotheses, *Statistical Research Memoirs* **1**, 1–37; **2**, 25–57.
[7] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

(*See also* **Inference**)

L.M. LE CAM

Nightingale, Florence

Born: May 12, 1820, in Florence, Italy.
Died: August 13, 1910, in London, UK.

Florence Nightingale was the second daughter of Fanny and William Nightingale. The Nightingales were a wealthy family and had two large country houses, one in Derbyshire and the other near Romsey in Hampshire, as well as rooms in Mayfair, London. Florence and her elder sister, Parthenope, were educated at home, initially by a governess and later also by their father who was a graduate of Trinity College, Cambridge. Their education was classical and included only basic mathematics.

Florence was fascinated by numbers and in 1840 announced her wish for further tuition in mathematics. This request met with family disapproval, particularly from her mother who thought this unnecessary study for a girl, who should by now be forming social ambitions. Florence won the support of her Aunt Mai, who managed to persuade Florence's father to allow the tuition, and for a few weeks Florence was tutored by James Sylvester [2]. It is not clear when she was introduced to statistics, but she used to study the blue books and statistics on public health and hospitals in the early hours of the morning [1]. Thus she was familiar with the problems of hospital administration well before her thoughts turned to **nursing**.

Florence believed that God had called her to do something, but did not know what. She was discontented with life and with herself. She seemed to gain most satisfaction from helping the poor and sick living near her family's Derbyshire home. Sometime in 1844 the realization came to her that her vocation lay in hospitals among the sick, but, anticipating family disapproval, it was not until December 1845 that she proposed her plan to go to Salisbury Infirmary to work as a nurse. The scheme was totally opposed by her parents. Florence was not to be deterred, and in 1851 persuaded her family to allow her to undertake a course of training at Kaiserswerth in Germany. She finally managed to convince her family that she should be allowed to practice nursing and in 1853 took a position as superintendent of an Institution for Sick Gentlewomen in Distressed Circumstances in Harley Street, London.

Once the Institution was running smoothly, Florence started visiting other hospitals to collect facts to establish the case for reforming conditions for hospital nurses. As her reputation grew, doctors contacted her, asking her to recommend nurses. Florence recognized the need for a training school capable of producing a supply of respectable, reliable and qualified nurses. She established such a school in 1860.

In 1853 the Crimean War began; the initial battles were horrific in terms of casualties. There were few facilities for the sick and wounded, and the problem was compounded by a cholera epidemic. Sidney Herbert was the Secretary of State for War responsible for the treatment of the sick and wounded. He was also a friend of Florence and knew of her work in the Harley Street Institution. He wrote to her inviting her to lead, with the government's sanction and at the government's expense, a party of nurses to Scutari to work in the British Army hospital.

The party of 38 nurses arrived at Scutari on November 4, 1854. The first hurdle to overcome was to be accepted by the Army medical department, which had declared itself against the introduction of women as nurses for soldiers. By the end of November there were 8000 men in the hospitals and the conditions were terrible, and so the Army medical authorities grudgingly had to accept all offers of help. All the administrative systems of the hospital had collapsed, and slowly Florence got her way. Sanitary improvements, cleaner surroundings and an improved diet were rewarded with a massive decrease in the

death rate. Woodham-Smith [7] provides a detailed account of Florence's time in Scutari.

Florence returned to England in July 1856 a national hero, and the position of the nursing profession had now been established. Florence was determined to reform the British Army. In 1857, largely through her efforts, a Royal Commission was set up to investigate the disasters of the Crimean War. Since women were not allowed to serve as members of the Commission or to testify, Florence wrote and compiled facts about the war and sent them to the Commission. She worked with **William Farr** on compiling the data she had collected in Scutari and devised some novel **graphical displays**, most notably her coxcomb [2, 6], to help convey her message. Her work was published in 1858 [3].

On December 21, 1858, Florence was elected a Fellow of the Statistical Society of London (now the **Royal Statistical Society**); she was one of the first female fellows of the society. She became an honorary member of the **American Statistical Association** in 1874.

Soon after her return from Scutari, Florence became reclusive and only saw people by appointment, usually at her home in South Street, Mayfair. She was a prolific writer, and in 1859 published her famous work *Notes on Nursing – What It Is And What It Is Not*. The book was very popular and was expanded and republished in 1860 [4]. It has sold millions of copies all over the world.

In 1859 she began her campaign for uniform hospital statistics. She devised a disease classification system and some model forms for the collection of the data. Again she worked with William Farr on this project. In 1860 this work was presented at the International Statistical Congress, held in London [5]. For a short time London hospitals made returns on these model forms, the data being published in the *Journal of the Statistical Society*. **Adolphe Quetelet** attended the congress, where Florence and he first met. She was fascinated by his statistical work, which had a great influence on her thinking [1].

In 1860 Florence established the Nightingale Training School for nurses at St Thomas's Hospital, London. This was financed by the Nightingale Fund established from donations made in recognition of her work in Scutari. It became the model for schools of nursing everywhere.

Florence's work in the mid-1860s was dominated by the contributions she made to sanitary improvements in India, both for the Army in India and the people of India. In 1907 she became the first woman to be awarded the Order of Merit. Three years later she died peacefully at the age of 90.

References

[1] Diamond, M. & Stone, M. (1981). Nightingale on Quetelet, *Journal of the Royal Statistical Society, Series A* **144**, 66–79.

[2] Grier, B. & Grier, M. (1978). Contributions of the passionate statistician, *Research in Nursing and Health* **1**, 103–109.

[3] Nightingale, F. (1858). *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army, Founded Chiefly on the Experience of the Late War*. Harrison & Sons, London.

[4] Nightingale, F. (1860). *Notes on Nursing – What It Is, And What It Is Not*. Harrison & Sons, London.

[5] Nightingale, F. (1860). Hospital statistics, in *Programme of the Fourth Session of the International Statistical Congress*. Eyre & Spottiswoode for HMSO, London, pp. 63–71.

[6] Wainer, H. (1995). A rose by another name, *Chance* **8**, 46–51.

[7] Woodham-Smith, C. (1950). *Florence Nightingale 1820–1910*. Constable, London.

NICOLA J. CRICHTON

Noise and White Noise

Noise is familiar to us as an *unmusical* sound or, technically speaking, *interference* in a communication channel. The effect of *snow* on a television set is a form of visual noise. In communication engineering, noise arises from all the uncontrolled sources of fluctuations of voltage, current, thermal agitation of electrons in resistors, etc. Shannon [6] depicted a communication system schematically as in Figure 1. Any difference between the message at source and at destination could be attributable to inherent disturbances, or noise, in any, or all, of the activities indicated. Engineers spend much time and effort to control and reduce the noise in a system.

In statistical methods such as regression analysis, noise is the random sampling component and the "message or signal" the underlying model. Most time series are subject to noise, usually represented as a stochastic element in an appropriate model description. A typical problem is to estimate $X(t + \Delta)$ for positive, negative or zero Δ, from an observed process $\{Y(t)$; where t is a real value$\}$ with assumed model $Y(t) = X(t) + Z(t)$, where $Z(t)$ is the *noise* component. Barahona & Poon [1] give details of methods that can be used to distinguish deterministic chaos from random noise in short time series.

To allow the investigator to identify the underlying model or signal from the collection of noisy observations, he or she needs to determine the nature and structure of the noise. There is a range of tools available to analyze noise in time series data; see, for example, [5]. The **fast Fourier transform (FFT)**, introduced by Cooley & Tukey [4], can be used to estimate the power spectra of noise (*see* **ARMA and ARIMA Models**). Computer software is also available to fit autoregressive moving-average models to the noise process, using the analysis of autocorrelation and partial autocorrelation functions [3, 5].

White Noise

Biostatistical time series data often occur as sequences of observations equally spaced in time, and can therefore be analyzed using discrete-time models [2, 3, 5]. If $\{Z_t\}$ is a set of random

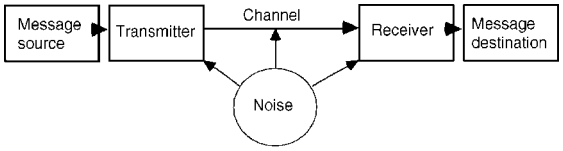

Figure 1 Main components of a communication system, indicating where noise can affect the message sent

variables with $\mathrm{E}(Z_t) = \mu_t$, and autocovariance function (*see* **Autocorrelation Function**) $\gamma_{t,s} = \mathrm{cov}\{Z_t, Z_s\} = \mathrm{E}[\{Z_t - \mu_t\}\{Z_s - \mu_s\}]$, then the stochastic process is said to be **stationary** if $\mu_t = \mu$ and $\gamma_{t,s} = \gamma(|t - s|)$. In many applications the noise process is assumed stationary and ergodic, i.e. its parameters can be estimated from a single realization of the process.

White noise, in discrete time, is simply the name given to a stochastic process $\{Z_t\}$ with $\mathrm{E}(Z_t) = 0$, and

$$\gamma_{t,s} = \begin{cases} \sigma^2, & t = s, \\ 0, & t \neq s. \end{cases}$$

Such a process has a constant spectral density. Autoregressive moving-average processes are derived from discrete-time white noise. The step from discrete to continuous time is not easy and involves an understanding of the Wiener process [2].

References

[1] Barahona, M. & Poon, C.F. (1996). Detection of nonlinear dynamics in short, noisy time-series, *Nature* **381**, 215–217.

[2] Bartlett, M.S. (1960). *An Introduction to Stochastic Processes*. Cambridge University Press, Cambridge.

[3] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, California.

[4] Cooley, J.W. & Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series, *Mathematics of Computation* **19**, 297–301.

[5] Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

[6] Shannon, C.E. (1948). A mathematical theory of communication, *Bell Systems Technical Journal* **27**, 379.

(*See also* **Spectral Analysis**)

CLIVE J. LAWRENCE

Nominal Data

Nominal data arise from a specific type of *measurement*. As defined by the psychologist S.S. Stevens [9], measurement is the assignment of numbers to entities according to some rule. Close examination of such rules allows us to classify the various types of measurements into a small set of possibilities. Measurements are, of course, taken for many different reasons to answer many different questions, usually (but not always) on experimental units or subjects or on observational units. Measurements produce data which take on values along some **measurement scale**. The measurement type(s) of data to be analyzed dictates the analyses that yield justifiable, and meaningful interpretations.

Stevens classified measurement scales into four types:

1. nominal;
2. ordinal;
3. interval;
4. ratio.

Nominal measurements give rise to a measurement scale which classifies entities by labeled categories. The categories are unordered. All we can state when comparing measurements on different entities is that they are equal or unequal. There is no particular mathematical significance given to the labeled categories themselves. The scale produces labels to be attached to the entities so that the entities can be classified into (usually) a discrete number of categories (hence, statisticians often view nominal measurements simply as producing unordered categorical data).

Ordinal and interval measurements produce scales on which measurements can be ranked (ordinal), or scale values on which arithmetic can be defined (interval). Letter grades given to students on an examination, or preoperative condition scores of patients (poor, fair, good, and so forth) follow ordinal scales; temperature measured in Fahrenheit or centigrade falls on an interval scale (the absence of an absolute zero-point is noteworthy for interval scales). Lastly, ratio measurements have all the features of interval measurements, with the additional property that levels on the scale may be expressed as ratios. Zero-points are well-defined (as in the case of the Kelvin temperature scale). A very good, and lengthy, discussion of measurement theory can be found in Krantz et al. [5], Suppes et al. [10], and, especially, Luce et al. [6] and a nice summary is given in Wallsten [11].

Ordinal measurements are usually categorical in nature, and can be studied with **ordered categorical data** techniques. Statisticians often assume that measurements on interval and ratio scales can be modeled as continuous random variables. One can of course use techniques for nominal data to analyze ordinal measurements, but not ordinal techniques for nominal data. It clearly is best to match the level of measurement with the analytic techniques used.

A good example of a nominal measurement scale is gender, measured on people – two categories, which clearly cannot be ordered in a meaningful way. A nominal scale can have any number of levels – assume that the number of levels is I. Other examples include race, religion, and, of course, the outcome of a possibly fatal disease (alive or dead). Other examples of nominal data, as well as a little more on the measurement aspects of nominal scales, can be found in [7]. Nominal data are common in all areas of science, and the techniques for their analyses have expanded greatly since the 1960s. There is a large variety of such techniques – any of the tools of **categorical data analysis** are appropriate, as long as categories are not assumed to be ordered. Here, we briefly mention these techniques.

We assume that interest is on a set of experimental units, or some observational subjects (people or otherwise). Let this set be of size N. For simplicity, we also assume that this set of units arises from a simple random sample, and that we are taking measurements on $S \geq 1$ nominal scales. Responses given by the units are unknown, and are governed by some statistical mechanism (more on this below), so that the nominal scales may be termed *nominal variables*. Analytic techniques, whether they be data analytic or statistical, depend on the number of variables under study, and hence, the discussion here is divided into the three possibilities: univariate, bivariate, and multivariate.

Univariate Techniques

For the analysis of a single nominal variable, we record the frequencies of units falling into the I

categories. These frequencies can be displayed graphically, in a variety of different ways. When $I = 2$, these **binary data** are often assumed to arise from Bernoulli random variables, so that statistical techniques based on the **binomial distribution** are usually appropriate when studying the frequencies. We usually use 0 and 1 to code the two categories. If interest is not on the frequencies, but on how many subjects must be sampled before the frequencies attain certain values, then the **negative binomial distribution** may be appropriate. For these probability models, we let p be the probability that a unit takes on the value 1, or falls into the "second" category. Frequently, interest centers on the odds of a measurement of 1 vs. a measurement of 0, which equals $p/(1-p)$. The logarithm of this **odds ratio** is often termed a *logit*, and is a useful transformation of binary data.

For $I > 2$ categories, or **polytomous data**, we define p_i as the probability of falling into the ith category. The I frequencies can be displayed using many graphical techniques (*see* **Exploratory Data Analysis**). Relevant probability distributions include the polytomous generalization of the binomial, the **multinomial distribution**, and occasionally, the **Poisson distribution** (which arises naturally from the stochastic **Poisson process** by counting how many events occur in some fixed interval of time, for example).

Statistical inference centers on learning about the unknown parameters of these distributions. Often, one might hypothesize models for such parameters; for example, one could test for equal probabilities ($p_1 = p_2 = \cdots = p_I = 1/I$). A variety of logits could be calculated and studied. These hypotheses are often modeled loglinearly by postulating that the logarithms of the probabilities (another useful theoretic transformation) are linear functions of various parameters (*see* **Loglinear Model**). More sophisticated forms of **generalized linear model** arise by relaxing distributional and scale assumptions concerning the response variable. The standard, albeit asymptotic for large N, approach to such testing is via *chi-square tests*. Other approaches are possible, including the use of other distributions (such as the **beta-binomial distribution**).

If some of the $\{p_i\}$ are zero, some of the categories must have zero frequencies regardless of the magnitude of N. Such categories are called structural zeros, and inferential procedures must take them into account.

Bivariate Techniques

Assume that one has a pair of nominal variables, one with I_1 categories, and the second with I_2 categories. We record the frequencies of the units falling into the $I_1 I_2$ categories of the cross classification of the two variables. These frequencies are arrayed into a two-dimensional **contingency table**, of size $I_1 \times I_2$. We usually assume that a multinomial distribution with probabilities $\{p_{ij}\}$, or a set of multinomial distributions (one for each row of the table, for example), generated these data.

Hypotheses about the probabilities are usually translated into loglinear models. Maximum likelihood estimates of model parameters often have closed form solutions, although iterative algorithms (such as **iterative proportional fitting**) can almost always be used.

Without question, the most common hypothesis is independence of the two variables; that is, a test of whether p_{ij} can be factored into the product $p_{i\cdot}p_{\cdot j}$, for all i, j. The study of independence has led to the construction of a wide range of measures of association between two nominal variables; two of the oldest and still most widely used are the **Goodman–Kruskal** indices. Odds ratios can also be used to quantify association.

This very standard hypothesis of independence can be tested in a variety of ways, although the most common is still via Pearson's χ^2 statistic, which for large N, is **chi-square distributed**. Note, however, that **exact tests for categorical data** do indeed exist (which are not asymptotic **likelihood ratio tests**); the most widely used is **Fisher's exact test**.

In the case of square, $I \times I$ tables, a range of special hypotheses arise, including tests of symmetry and **quasi-symmetry** of the table (for example, does $p_{ij} = p_{ji}$?), and the **McNemar test**, which looks at equality of the upper and lower triangles. Such a **square contingency table** is common when the two variables have exactly the same categories (perhaps they are measurements on the same units, separated in time; such as success or failure of some drug at time 1 and time 2). If both variables are binary, we obtain a **two-by-two table**.

Yet another possibility, although one that is used more often when $S > 2$ variables, is a regression approach, using the logit transformation(s) of one of the variables as the response variable. Techniques for such a **logistic regression** are well understood, and

increasingly used. Explanatory variables can be of many types, and allow for a wide range of interactions.

Multivariate Techniques

Lastly, suppose that $S > 2$, so that the contingency table is multiway. Data analytic and statistical techniques are designed to highlight the many interactions that can exist among the variables, and to test for different types of independence. We refer those interested to the article on **Categorical Data Analysis**, or to one of the many textbooks on the subject, especially Agresti [1] and Bishop et al. [3] for the more advanced student, or to the primers by Agresti [2], Wickens [12], and Fienberg [4]. The chapter by Sobel [8] presents a nice overview of multivariate techniques.

Acknowledgments

This work was supported by a grant from the National Science Foundation. I thank Alan Agresti, Carolyn Anderson, David Budescu, and Laura Koehly for their helpful comments.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

[3] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[4] Fienberg, S.E. (1980). *The Analysis of Cross-Classified, Categorical Data*, 2nd Ed. MIT Press, Cambridge, Mass.

[5] Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A. (1971). *Foundations of Measurement*, Vol. I, Academic Press, New York.

[6] Luce, R.D., Krantz, D.H., Suppes, P. & Tversky, A. (1990). *Foundations of Measurement*. Vol. III: Representation, Axiomatization, and Invariance. Academic Press, New York.

[7] Reynolds, H.T. (1985). Nominal data, in *Encyclopedia of Statistical Science*, Vol. 6, S. Kotz, & N.L. Johnson, eds. Wiley, New York, pp. 256–261.

[8] Sobel, M.E. (1995). The analysis of contingency tables, in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg & M.E. Sobel, eds. Plenum, New York, pp. 251–310.

[9] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 46–52.

[10] Suppes, P., Krantz, D.H., Luce, R.D. & Tversky, A. (1990). *Foundations of Measurement*. Vol. II: Geometrical, Threshold, and Probabilistic Representations. Academic Press, New York.

[11] Wallsten, T.S. (1985). Measurement theory, in *Encyclopedia of Statistical Science*, Vol. **5** S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 387–389.

[12] Wickens, T.D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*. Erlbaum, Hillsdale.

(*See also* **Chi-square Tests**; **Polytomous Data**)

STANLEY WASSERMAN

Noncentral *t* Distribution

The noncentral t distribution is a two-parameter family with a **unimodal** density having parameters ν (**degrees of freedom**) and δ (noncentrality parameter). Degrees of freedom are usually positive integers, although any positive value is theoretically possible. The noncentrality parameter can take on any real value, although when it equals zero the distribution reduces to a special case; namely, the well known (central) **Student's *t* distribution**.

Suppose that X and Y are stochastically independent **random variables** such that X is **normally distributed** $\mathrm{N}(\delta, 1)$ and Y is central χ^2 with ν degrees of freedom (df). Then the ratio

$$R = \frac{X}{\sqrt{(Y/\nu)}}$$

follows the noncentral t distribution. (If Y is noncentral χ^2, then R has the doubly noncentral t distribution.) The most important application relates to the testing of the **null hypothesis** H_0, that the mean of a normal population with unknown variance equals μ_0. If H_0 is true, the statistic $T = \sqrt{n}(\overline{x} - \mu_0)/s$ follows the central t distribution with $n - 1$ df, but if the mean instead equals μ_1, then T follows the noncentral t distribution with noncentrality parameter $\delta = \sqrt{n}(\mu_1 - \mu_0)/\sigma$. Similarly, for testing the equality of means for two normal populations having common unknown variance, one may construct the statistic $T = \sqrt{[n_1 n_2/(n_1 + n_2)]}(\overline{x}_1 - \overline{x}_2)/s_\mathrm{p}$, where s_p is the pooled standard deviation. Under the null hypothesis, T follows the central t distribution with $n_1 + n_2 - 2$ df, but when the two population means differ by an amount $\Delta = \mu_1 - \mu_2$ the relevant noncentrality parameter is $\delta = \sqrt{[n_1 n_2/(n_1 + n_2)]}\ \Delta/\sigma$. The **power** of either test is expressible with the cumulative distribution function (cdf) of the relevant noncentral t, evaluated at the appropriate critical point(s) of the corresponding central t distribution.

Some moments of the noncentral t are

$$\mathrm{E}(R) = \delta\sqrt{\frac{\nu}{2}}\left[\Gamma\left(\frac{\nu-1}{2}\right)\Big/\Gamma\left(\frac{\nu}{2}\right)\right],$$

$$\mathrm{var}(R) = \frac{\nu}{\nu-2}(1+\delta^2) - [\mathrm{E}(R)]^2,$$

$$\mu_3(R) = \frac{\nu}{\nu-3}\mathrm{E}(R)\left[\frac{3(1+\delta^2)}{(\nu-2)} - 2\delta^2\right] + 2[\mathrm{E}(R)]^3.$$

The noncentral t density can be written as

$$f(\nu, \delta; t) = \frac{\exp(-\delta^2/2)}{\sqrt{(\pi\nu)}}\left[\Gamma\left(\frac{\nu+1}{2}\right)\Big/\Gamma\left(\frac{\nu}{2}\right)\right] \times \left[\frac{\nu}{\nu+t^2}\right]^{(\nu+1)/2} S(\nu, \delta; t),$$

where

$$S(\nu, \delta; t) = \sum_{j=0}^{\infty}\left[\Gamma\left(\frac{\nu+j+1}{2}\right)\Big/\Gamma\left(\frac{\nu+1}{2}\right)j!\right] \times \left[\frac{t\delta\sqrt{2}}{\sqrt{(\nu+t^2)}}\right]^j.$$

Although this series converges rather slowly for large arguments, it satisfies the equation

$$\frac{\mathrm{d}^2 S(t)}{\mathrm{d}t^2} = \left[\frac{\delta^2\nu t}{(\nu+t^2)^2} - \frac{3t}{\nu+t^2}\right]\frac{\mathrm{d}S(t)}{\mathrm{d}t} + \frac{\delta^2\nu^2(\nu+1)}{(\nu+t^2)^3}S(t).$$

This relationship is useful for recasting noncentral t distributions in terms of S-system differential equations, which are used in solving their densities, cdfs, and **quantiles**, as well as **moments** of integer and fractional orders [4].

Alternative expressions for the noncentral t are available; for example,

$$f(\nu, \delta; t) = K\exp\left[-\frac{\nu\delta^2/2}{\nu+t^2}\right]\left(\frac{\nu}{\nu+t^2}\right)^{(\nu+1)/2} \times Hh_\nu\left[-\frac{\delta t}{\sqrt{(\nu+t^2)}}\right],$$

where

$$K = \nu!\left[2^{(\nu-1)/2}\sqrt{(\pi\nu)}\Gamma\left(\frac{\nu}{2}\right)\right]^{-1}$$

and

$$Hh_\nu(x) = \frac{1}{\nu!}\int_0^\infty u^\nu \exp\left[\frac{-(u+x)^2}{2}\right]\mathrm{d}u.$$

For a random sample of size n from an $\mathrm{N}(\mu, \sigma^2)$ population, with $\mu > 0$, the sample coefficient of

variation $s/\overline{x}$ has a distribution related to noncentral t, although the tails of the former map to small values of the latter. More specifically, if R has $\nu = n - 1$ and $\delta = \mu\sqrt{n}/\sigma$, then $\Pr(s/\overline{x} > c) = \Pr(0 \leq R \leq \sqrt{n}/c)$.

Other situations in which noncentral t distributions are relevant include one-sided *tolerance* limits for a normal distribution (*see* **Tolerance Interval**), lot acceptance sampling plans, **confidence** limits on one-sided normal **quantiles** and **binomial** proportions, and one-sided tolerance limits in **linear regression**.

Noncentral t tables have occasionally been published; a fresh and extensive compilation [1] has appeared recently. See also [2] and [3].

References

[1] Bagui, S.C. (1993). *CRC Handbook of Percentiles of Noncentral t-Distributions*. CRC Press, Boca Raton.

[2] Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics: Continuous-Univariate Distributions*, Vol. 2. Houghton Mifflin, Boston.

[3] Owen, D.B. (1968). A survey of properties and applications of the noncentral t distribution, *Technometrics* **10**, 445–478.

[4] Voit, E.O. & Rust, P.F. (1990). Evaluation of the noncentral t distribution with S-systems, *Biometrical Journal* **32**, 681–695.

PHILIP F. RUST

Noncompliance, Adjustment for

The randomized **clinical trial** (RCT) is arguably the most important contribution of statistical science to human health. It provides a scientific basis for the **unbiased** evaluation of preventive and therapeutic treatments, not by controlling sources of variation as done in the laboratory, but by balancing them across treatment groups (*see* **Randomized Treatment Assignment**). In the simple RCT design (*see* **Clinical Trials, Overview**), each participant is randomly assigned to one treatment (e.g. placebo; *see* **Blinding or Masking**) or another (active agent) as indicated by the value of $R_i = 0$ or 1. Given his or her random assignment, a patient actually takes either the placebo or active treatment as indicated by $D_i(R_i) = 0$ or 1, and then experiences a health outcome $Y_i[D_i(R_i)] = 0$ or 1, $i = 1, \ldots, n$.

In the ideal trial, 100% of participants comply with the treatment to which they were randomly assigned so that $D_i(R_i) = R_i$. In practice, however, compliance is less than perfect as illustrated in Table 1 with data from the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) [8, 10, 16]. Of the 337 subjects reported in the table, 123 (36%) fail to reach the (arbitrary) 60% compliance level used for illustration below. Note that the noncompliance (<60%) rates in the placebo ($R = 0$) and cholestyramine ($R = 1$) arms are quite different – $46/172 = 27\%$ and $77/165 = 47\%$, respectively.

Such a RCT is designed to address several questions including:

1. Is there sufficient evidence to reject the **null hypothesis** that the treatment and placebo have the same effect on the health outcome (*see* **Outcome Measures in Clinical Trials**)?
2. What is the average difference in outcome caused by being randomized to the treatment, rather than placebo, group?
3. Among persons who comply with their treatment regimen, what is their average improvement as a result of receiving treatment rather than placebo?

Questions 1 and 2 are typically addressed with an **intention-to-treat (ITT) analysis**, in which the average response is compared across randomization groups ($R = 0$ vs. $R = 1$) without regard to the treatment that was actually received. The target of estimation implicit in the ITT analysis has been called the *programmatic effectiveness* [14], which combines the average therapeutic effect of the treatment and the rate of patient compliance to the treatment regimen. Question 3 focuses upon the therapeutic benefit alone. The average benefit among compliers has been referred to as the *biologic efficacy* [14]. While efficacy is often of scientific interest, the ITT analysis alone is inadequate to estimate it. Compliance information must also obviously be used.

Even when addressing questions 1 and 2, compliance information is relevant. As compliance decreases, effectiveness and the **power** to detect a treatment difference also decrease [5]. Hence, sample sizes must be increased to maintain a desired level of power (*see* **Sample Size Determination for Clinical Trials**). A second design strategy to contend with imperfect compliance is inclusion of a *run-in period* [13] after which persons with poor compliance are dropped from the study. **Randomization** to treatment group occurs at the end

Table 1 Two-by-two-by-two table displaying presence (1) or absence (0) of improvement in cholesterol level by 20 units (Y); treatment group assignment; $R = 0$ for placebo and $R = 1$ for cholestyramine; and compliance level: <60%; ≥60%

		Placebo ($R = 0$)			Treatment ($R = 1$)		
		Compliance			Compliance		
		<60%	≥60%		<60%	≥60%	
Y	0	$m_{00} = 42$	$m_{01} = 98$	$m_{0\cdot} = 140$	$n_{00} = 50$	$n_{01} = 16$	$n_{0\cdot} = 66$
	1	$m_{10} = 4$	$m_{11} = 28$	$m_{1\cdot} = 32$	$n_{10} = 27$	$n_{11} = 72$	$n_{1\cdot} = 99$
		$m_{\cdot 0} = 46$	$m_{\cdot 1} = 126$	$m_{\cdot\cdot} = 172$	$n_{\cdot 0} = 77$	$n_{\cdot 1} = 88$	$n_{\cdot\cdot} = 165$

of the run-in period (*see* **Compliance Assessment in Clinical Trials**).

Compliance is easy to discuss in the abstract, but can be difficult to measure in practice. In many trials, patients are asked to make diary entries of pills taken. Or, in some studies, investigators repeatedly collect patient sera samples from which estimates of drug concentrations are made. In these cases, a longitudinal data set is generated, from which compliance can be inferred using methods such as those discussed by Lim [15] and Kim & Lagakos [12].

Until recently, RCT practitioners were reticent to estimate efficacy using compliance information collected after randomization because substantial **selection biases** can occur. Recently, the method of *instrumental variables*, common in econometrics, has been applied to the RCT to contend with selection bias inherent in estimating efficacy. Because the method leads to the ITT test of the null hypothesis, and to unbiased estimates of efficacy in many situations, they are attractive to many, although certainly not all, statistical scientists. Below, the main ideas of this approach are illustrated using the LRC-CPPT data. Extensions to different kinds of outcomes and compliance measures are then summarized. Connections to instrumental variables methods in econometrics and to Rubin's causal model are important and are discussed in detail by Angrist et al. [2]. One of the more complex models for compliance is presented in Efron & Feldman [8] and in a paper critical of their approach by Albert & DeMets [1].

Adjusting for Compliance to Estimate Efficacy

The ITT estimate of **relative risk** (RR) representing programmatic effectiveness for the LRC-CPPT data in Table 1 is $RR_{\mathrm{ITT}} = (66/165)/(140/172) = 0.49$ with an approximate 95% **confidence interval** (0.40, 0.60). These data provide evidence that assignment to receive cholestyramine reduces the risk of failing to improve the cholesterol level in comparison with a group assigned to receive a placebo. However, only 88 of the 165 participants (53%) in the treatment group actually took more than 60% of their medication. Hence, the relative risk estimate, RR_{ITT}, probably underestimates the biologic impact of treatment, since only about half of the treatment group actually received a substantial fraction of the drug assigned.

The question remains, what is the treatment effect among the subgroup of persons who complied with their treatment assignment? The direct, but naive, way to estimate efficacy from Table 1 is to compare the compliers in the treatment and placebo groups. From Table 1, we have $RR_{\mathrm{E_I}} = (16/88)/(98/126) = 0.23$. Use of this simple estimate raises serious concerns, because it is subject to selection bias. The compliers in the treatment and placebo groups may not be comparable in ways other than their treatment status. There is evidence to this effect – the fraction of compliers in the placebo group is 73% as compared with 53% in the treatment group. Ingesting cholestyramine has been described as similar to eating sand; perhaps the placebo was more tolerable; hence, the compliant subgroups in the two arms are not necessarily comparable. This potential for differential subgroup selection bias is one basis for serious concerns about such efficacy estimates among many clinical trialists [6].

There is, however, another estimate of efficacy that can be useful in some circumstances because it does not suffer from this selection bias. Consider an imaginary, perfect placebo that is identical to the treatment in every possible way except that it has no biologic activity. Further assume that compliance is conditionally independent of the health outcome given treatment assignment; that is, the effect of the treatment (or lack thereof) on the health outcome is not the cause of the differential compliance. Then, in a clinical trial in which participants receive the drug only if they are in the treatment group and comply $[D(1) = 1; D(0) = 0]$, randomization guarantees two conditions:

1. The expected fraction of compliers would be the same in the two treatment groups. The observed fraction in the cholestyramine group (53%) is an unbiased estimate of this common compliance rate.
2. Among the noncompliers, no drug is received (not quite true in the LRC-CPPT example where the compliance level is actually a continuum) so that their expected rate of health outcome would be the same across the two treatment groups.

Under the assumptions above, these two facts allow us to infer the expected entries in the placebo side of Table 1 for an imaginary, perfect placebo. As shown

by Sommer & Zeger [23], m_{10} can be estimated by $m_{..}n_{10}/n_{..}$. That is, the expected rate of successful outcomes in a perfect-placebo, noncompliant subgroup can be estimated by the observed rate in the treatment, noncompliant group. The expected number of successful outcomes in the placebo, compliant subgroup can then be obtained by subtracting the imputed values for the noncompliant group from the marginal totals $m_{0\cdot}$ and $m_{1\cdot}$. This leads to the alternative relative risk estimate [23]:

$$RR_{E_2} = \frac{n_{01}/n_{\cdot 1}}{\hat{m}_{01}/(\hat{m}_{01} + \hat{m}_{11})},$$

where $\hat{m}_{11} = m_{1\cdot} - (m_{..}/n_{..})n_{10}$ and $\hat{m}_{01} = m_{0\cdot} - (m_{..}/n_{..})n_{00}$. For the cholestyramine example, the new estimate of treatment efficacy is $RR_{E_2} = (16/88)/(87.9/91.7) = 0.190$. It is different from the naive estimate RR_{E_1} because it is corrected for the differential compliance rates in the two treatment arms.

RR_{E_2} is a simple example of an *instrumental variables* estimator, commonly used in econometric research [4]. Recent applications of this approach to clinical trials are by Permutt & Hebel [19], Robins [20] and Baker & Lindeman [3]. A detailed discussion of the use of instrumental variables for causal inference (*see* **Causation**) within Rubin's framework is given by Angrist et al. [2].

More Realistic Applications

In an analysis of data from the Multicenter Diltiazem Post-infarction Trial (MDPIT), Oakes et al. [18] include a measure of compliance as a **covariate** in Cox's **proportional hazard's** model [7] to estimate a **relative hazard** parameter for each of three comparisons; treatment vs. placebo for compliers; treatment vs. placebo for noncompliers; and compliers vs. noncompliers among persons allocated to the placebo group. Their treatment vs. placebo comparison among compliers is a **survival analysis** analog of RR_{E_1}. But their analysis also summarizes the evidence about selection bias in the comparison of the noncompliant subgroups and hence is a sensible first step. White & Pocock [24] extend this approach by considering **time-dependent** indicators of compliance.

Robins [20], Robins & Tsiatis [21] and Mark & Robins [17] develop and apply an **accelerated failure-time model** [7] to estimate the causal relative hazard of a treatment taking account of compliance information. Following Rubin's framework for causal inference [11, 22], they assume that each individual can be thought of as having a latent survival time associated with each treatment. The causal effect is defined as the difference in the latent survival times for the treatments being compared. Under the accelerated failure-time model, the effect of treatment is to rescale (make faster or slower) time so that the parameter of interest is the fractional increase in survival for an individual on vs. off treatment. Their use of compliance information is a survival analog of what is done in the simple **binary** case above. The strength of the Robins & Tsiatis method is that, as in the binary case, the test of the null hypothesis of no treatment effect is identical to the ITT test of the same hypothesis. See also Goetghebeur & Lapp [9].

Goetghebeur & Molenberghs [9] have developed this approach for the case of a **binary** response, but with an ordinal measure of compliance, as occurs in the cholestyramine example above. To do so, they make the additional *monotone treatment effect* assumption that, if a failure occurred at a given level of treatment, then it would also have occurred at all lower levels of treatment. This rules out the possibility of drugs being toxic and therefore less effective at higher doses, for example.

Efron & Feldman [8] take account of a continuous measure of compliance with a continuous outcome reduction in serum cholesterol level. Because they are interested in a possible interaction whereby the treatment effect will differ depending on a person's tendency to comply, they must use the observed compliance information in both the treated and control groups. Hence, they do not protect against unequal selection from an imperfect placebo as was done in the binary example above. Instead, they assume that an individual's inherent tendency to comply is measured by their quantile of compliance within their own treatment group. With this assumption, they use both the conditional means and variances of cholesterol change, given the level of compliance, to assess the causal effect of treatment as a function of compliance. This approach is distinct in its assumptions from the others described above. A critique of the Efron & Feldman efficacy estimate is provided by Albert & DeMets [1].

In summary, *programmatic effectiveness* and *biologic efficacy* are both important targets for estimation in clinical trials. The ITT analysis estimates

effectiveness. This article reviews the use of instrumental variable methods as one approach to estimating the efficacy of the treatment among persons who comply. In the case of a binary treatment indicator and binary outcome, this approach leads to an estimator that is not biased by differences between treatment groups in compliance when the differences are due to factors that are conditionally independent of the outcome given the treatment assignment. Hence, they will be useful when the decision to comply does not depend on the outcome being studied. Several extensions to more realistic situations are also summarized.

References

[1] Albert, J.M. & DeMets, D.L. (1994). On a model-based approach to estimating efficacy in clinical trials: analgesia during labor, *Statistics in Medicine* **13**, 2323–2335.

[2] Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996). Identification of causal effects using instrumental variables, *Journal of the American Statistical Association* **91**, 444–472.

[3] Baker, S.G. & Lindeman, K.S. (1994). The paired availability design: a proposal for evaluating epidural analgesia during labor, *Statistics in Medicine* **13**, 2269–2278.

[4] Bowden, R.J. & Turkington, D.A. (1984). *Instrumental Variables*. Cambridge University Press, Cambridge.

[5] Brown, B.W. (1984). Problems related to protocol nonadherence, in *Cancer Clinical Trials: Methods and Practice*, M.E. Buyse, M.J. Staquet & R.J. Sylvester, eds. Oxford University Press, Oxford.

[6] Byar, D.P., Simon, R.M., Friedewald, W.T., Schlesselman, J.J., Demets, D.L., Ellenberg, J.H., Gail, M.H. & Ware, J.H. (1976). Randomized clinical trials: perspectives on some recent ideas, *New England Journal of Medicine* **295**, 74–80.

[7] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[8] Efron, B. & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials, *Journal of the American Statistical Association* **86**, 9–26.

[9] Goetghebeur, E. & Lapp, K. (1997). The effect of treatment compliance in a placebo-controlled trial: regression with unpaired data, *Applied Statistics* **46**, 351–364.

[10] Goetghebeur, E. & Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance, *Journal of the American Statistical Association* **91**, 928–934.

[11] Holland, P.W. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **81**, 945–968.

[12] Kim, H.M. & Lagakos, S.W. (1994). Assessing drug compliance using longitudinal marker data, with application to AIDS, *Statistics in Medicine* **13**, 2141–2153.

[13] Lang, J.M. (1990). The use of run-in to enhance compliance, *Statistics in Medicine* **9**, 87–95.

[14] Last, J.M. (1988). *A Dictionary of Epidemiology*, 2nd Ed. Oxford University Press, Oxford.

[15] Lim, L-Y. (1992). Estimating compliance to study medication from serum drug levels: application to an AIDS clinical trial of zidovudine, *Biometrics* **48**, 619–630.

[16] Lipid Research Clinic Program (1984). The Lipid Research Clinics Coronary Primary Prevential Trial results, parts I and II, *Journal of the American Medical Association* **251**, 351–374.

[17] Mark, S.D. & Robins, J.M. (1993). A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial, *Controlled Clinical Trials* **14**, 79–97.

[18] Oakes, D., Moss, A.J., Fleiss, J.L., Bigger, J.T., Jr, Therneau, T., Eberly, S.W., McDermott, M.P., Manatunga, A., Carteen, E., Benhorn, J. & The Multicenter Diltiazem Post-Infarction Trial Research Group (1993). Use of compliance measures in an analysis of the effect of diltiazem on mortality and reinfarction after myocardial infarction, *Journal of the American Statistical Association* **88**, 44–49.

[19] Permutt, T. & Hebel, J.R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight, *Biometrics* **45**, 619–622.

[20] Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in *Health Services Research Methodology: A Focus on AIDS*, L. Sechrest, H. Freeman & A. Bailey, eds. National Center for Health Services Research, US Public Health Service, Washington.

[21] Robins, J.M. & Tsiatis, A. (1991). Correcting for noncompliance in randomized trials using rank preserving structural equation failure time model, *Communications in Statistics-Theory and Methods* **20**, 2609–2631.

[22] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies, *Journal of Educational Psychology* **66**, 688–701.

[23] Sommer, A. & Zeger, S.L. (1991). On estimating efficacy from clinical trials, *Statistics in Medicine* **10**, 45–52.

[24] White, I.R. & Pocock, S.J. (1996). Statistical reporting of clinical trials with individual changes from allocated treatment, *Statistics in Medicine* **15**, 249–262.

SCOTT L. ZEGER

Nondifferential Error

Suppose a response variable Y has a conditional distribution $F(y|x)$ given a true exposure measurement, $X = x$. Suppose that instead of measuring X, one measures an error-prone version of X, say Z. Then the error process is said to be nondifferential if $F(y|x, z) = F(y|x)$, namely if Y and Z are conditionally independent given X. In usual cases, though not in all cases, the effect of analyzing the model $F(y|x)$ by substituting Z for X will be to **bias** estimates of **exposure effect** toward the **null hypothesis** (*see* **Bias Toward the Null**) when the exposure error is nondifferential. However, if exposure measurements are differential, the bias can be in any direction.

The term nondifferential error can also be applied to errors in the outcome measure, Y. Suppose that one measures the error-prone version W of Y, rather than Y itself. Then the error process is nondifferential if W is conditionally independent of X given Y.

(*See also* **Bias in Observational Studies**; **Differential Error**; **Measurement Error in Epidemiologic Studies**; **Misclassification Error**; **Validity and Generalizability in Epidemiologic Studies**)

MITCHELL H. GAIL

Non-Fourier Waveforms

While the discrete Fourier transform (DFT) is indispensable for analyzing **stationary** data, whose frequency behavior is fixed over time, we often have data that exhibits periods of high-frequency behavior, which do not extend over the whole time period. The DFT is not necessarily the "best" representation of such data, and we wish to find sets of basis function, which are able to isolate such behavior. Quite recently, there has been much interest in a class of **orthogonal** transforms that give information about the data at different times and scales. The functions that are used to represent the data are no longer the well-known sines and cosines of Fourier theory but are functions that are known as **wavelets**. These wavelets, as the name suggests, are "small" waves, meaning that they are only nonzero in a finite time interval. This compact support is critical for the success of these representations and allows us to study different scale behavior at different times.

A wavelet function, $\psi(\cdot)$, must satisfy two basic properties, namely,

1. The integral of $\psi(\cdot)$ is zero:

$$\int_{-\infty}^{\infty} \psi(t)\,\mathrm{d}t = 0. \tag{1}$$

2. The square of $\psi(\cdot)$ integrates to unity:

$$\int_{-\infty}^{\infty} \psi^2(t)\,\mathrm{d}t = 1. \tag{2}$$

The orthogonal transform associated with wavelets as basis functions is known as the discrete wavelet transform (DWT) and it results in a set of coefficients, which relate to different times and scales. In the same way, that the **fast Fourier transform** (FFT) has revolutionized the implementation of **spectral** methods, there exists a fast method of calculating the coefficients of the DWT, which is in fact computationally even faster than the FFT.

Arguably the oldest wavelet function is named after Haar [2]:

$$\psi(t) \equiv \begin{cases} \dfrac{-1}{\sqrt{2}}, & -1 < t \leq 0; \\ \dfrac{1}{\sqrt{2}}, & 0 < t \leq 1; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The Haar wavelet at different scales is shown in Figure 1.

More recently, many other families of wavelets have appeared that have more desirable properties, for example, in terms of their "smoothness", than the Haar wavelet. In particular, a class of wavelets

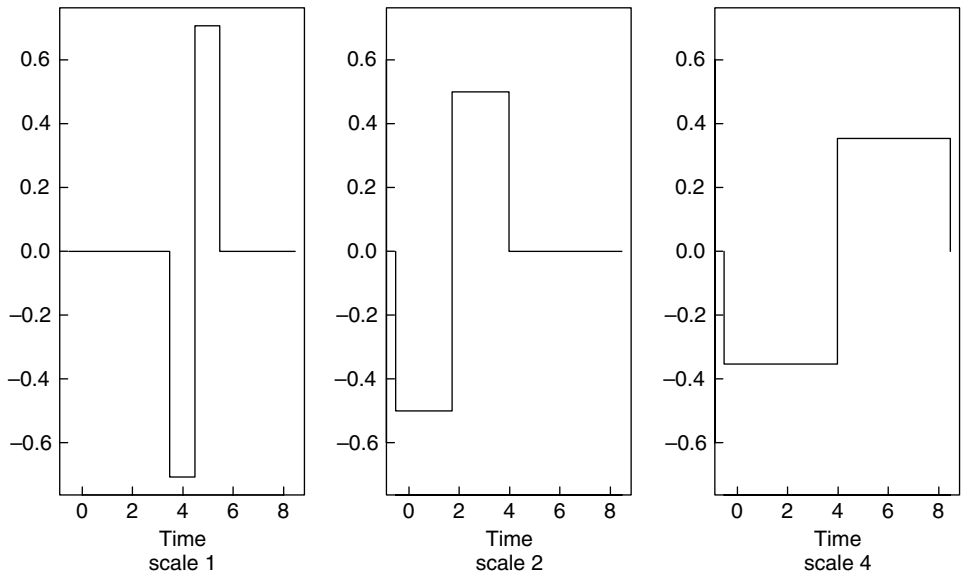

Figure 1 The Haar Wavelet at different scales

introduced by Daubechies [1] have proved popular and gained widespread use.

Given a time domain sequence $\boldsymbol{g} = (g_0, \ldots, g_{N-1})^\top$, the calculation of the N DWT coefficients, $\boldsymbol{w} = (w_0, \ldots, w_{N-1})^\top$ can be written in matrix form as

$$\boldsymbol{w} = \boldsymbol{W}\,\boldsymbol{g}, \tag{4}$$

where $\boldsymbol{W}$ is the $N \times N$ orthogonal wavelet matrix.

For illustrative purposes, when $N = 8$ and using the Haar wavelet, we have

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \end{bmatrix} = \begin{bmatrix} (g_1 - g_0)/\sqrt{2} \\ (g_3 - g_2)/\sqrt{2} \\ (g_5 - g_4)/\sqrt{2} \\ (g_7 - g_6)/\sqrt{2} \\ (g_3 + g_2 - g_1 - g_0)/2 \\ (g_7 + g_6 - g_5 - g_4)/2 \\ (g_7 + \cdots + g_4 - g_3 - \cdots - g_0)/\sqrt{8} \\ (g_7 + \cdots + g_0)/\sqrt{8} \end{bmatrix}. \tag{5}$$

The first four rows of this matrix correspond to unit scale changes (*high-frequency behavior*), the next two rows represent changes on a scale of two, the seventh row represents changes on a scale of four, while the final row represents the average at scale 8 (*low-frequency behavior*). In addition, each of the rows corresponds to different, nonoverlapping, time intervals in the data. For a good introduction to the DWT with direct comparisons to the DFT see [4].

Non-Fourier waveforms find a wealth of applications, particularly in areas where the underlying process has time-varying properties, for example, the study of arrhythmia in electrocardiogram (ECG) data, or denoising medical images (*see* **Clinical Signals**). For some statistical applications of wavelets, see [3].

References

[1] Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

[2] Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme, *Mathematische Annalen* **69**, 331–371.

[3] Ogden, R.T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston.

[4] Strang, G. (1993). Wavelet transforms versus Fourier transforms, *Bulletin of the American Mathematical Society* **28**, 288–305.

(*See also* **Time Series**)

E.J. McCoy

Nonignorable Dropout in Longitudinal Studies

Most longitudinal studies (*see* **Longitudinal Data Analysis, Overview**) are designed to collect data on every individual in the sample in a planned sequence of observation times. However, longitudinal studies habitually suffer from the problem of attrition; that is, some individuals "drop out" of the study prematurely. For example, suppose for each individual we plan to make a sequence of T observations on outcome variables $Y_1, \ldots, Y_T$. In addition, for each individual we may have a set of fixed covariates, X, and these are assumed to be fully observed. In that case, the term *dropout* refers to the special case where if Y_k is missing, then $Y_{k+1}, \ldots, Y_T$ are also missing. This gives rise to a monotone data pattern (see Figure 1) in contrast to the nonmonotone patterns that arise when data are missing intermittently. Note that intermittent missing data give rise to a considerably larger number of potential missing data patterns but, apart from that, do not raise any further technical considerations. When there is dropout in a longitudinal study, the key issue is whether those who drop out and those who remain in the study differ in any further relevant way. If they do not, then analyses restricted to those remaining in the study yield valid (albeit inefficient) inferences. If they do differ, then such analyses are potentially biased.

In the statistical literature, three different types of dropout have been distinguished [2, 4, 5]: *completely random, random*, and *nonignorable* dropout (see [3] for a more refined classification of dropout). Often the term "informative" dropout is used to refer to nonignorable dropout, e.g [1] (*see* **Diggle–Kenward Model for Dropouts**). To clarify the distinction between these different types of dropout, it is helpful to introduce a dropout indicator variable, D, for each individual. Let $D = k$ if an individual drops out between the $(k-1)$th and kth observation time, and $D = T + 1$ if there is no dropout. That is, when $D = k$ we only observe $Y_1, \ldots, Y_{k-1}$, and the remaining $Y_k, \ldots, Y_T$ are missing. Note, however, that D is recorded for all individuals. With completely random dropout, individuals leave or drop out of the study in a process that is independent of any other observed variables. That is, $\Pr(D = k|X, Y_1, \ldots, Y_T) = \Pr(D = k)$, and the probability of dropout does not depend on an individual's outcomes $Y_1, \ldots, Y_T$. Little [3] distinguishes completely random dropout from *covariate-dependent* dropout. In the latter, $\Pr(D = k|X, Y_1, \ldots, Y_T) = \Pr(D = k|X)$, and the probability of dropout depends on values of the fixed covariates X, but, given X, it is conditionally independent of an individual's outcomes $Y_1, \ldots, Y_T$. Note, however, that if dropout depends on covariates that have not been fully observed or on covariates that have been omitted from the model for the longitudinal outcomes, then dropout is no longer said to be *covariate-dependent*. With random dropout, the process can depend on the outcomes that have been observed in the past, but, given this information, it is conditionally independent of all future (unrecorded) values of the outcome variable following dropout. That is, $\Pr(D = k|X, Y_1, \ldots, Y_T) = \Pr(D = k|X, Y_1, \ldots, Y_{k-1})$, and the probability of dropout depends only on outcomes that have been observed or recorded. Finally, in the case of nonignorable or informative dropout, the dropout process, $\Pr(D = k|X, Y_1, \ldots, Y_T)$, depends on unobserved values of the outcome variable. That is, dropout is said to be nonignorable when the probability of dropout depends on the unrecorded values of the outcome variable that would have been observed had the individual remained in the study.

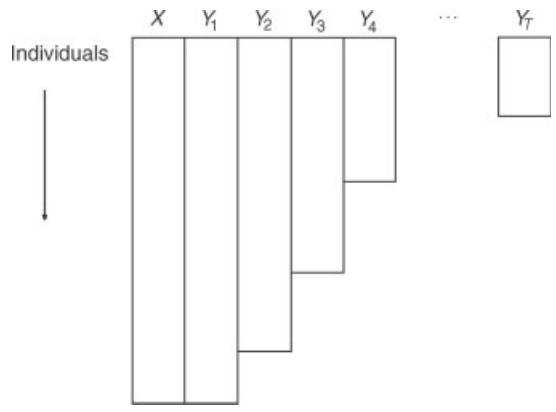

Figure 1 Schematic representation of a monotone data pattern (adapted from Little [3])

Completely random, covariate-dependent, and random dropout are often referred to as being *ignorable* (provided that the parameters of the dropout process are distinct from the parameters of the model for the longitudinal outcomes) [5]. We caution, however, that the use of the term ignorable does not imply that

the individuals with missing data can simply be disregarded. Rather, the term ignorable is used to indicate that it is not necessary to specify an explicit model for dropout in likelihood-based or **Bayesian inference** concerning the parameters in any model for longitudinal outcomes. Ignorable dropout can often be handled using standard statistical software (e.g. BMDP5V or SAS PROC MIXED), where those who drop out are included in the likelihood-based analysis (*see* **Software, Biostatistical**). However, with nonignorable dropout, the dropout mechanism cannot be ignored in likelihood-based or Bayesian inference [4, 5]. With nonignorable dropout, inference is only possible once assumptions are made about the dropout process. Recently, quite a number of methods have been proposed for handling nonignorable dropout, and all of these methods make particular assumptions about the dropout process. A general overview of the statistical literature on methods for modeling dropout can be found in [3]. However, it is worth stressing that, short of tracking down the individuals who have left the study, any assumptions made about the dropout process are not verifiable. Therefore, it is important to assess carefully the sensitivity of inferences to a variety of plausible assumptions concerning the dropout process.

Finally, there are some subtle issues concerning the identifiability of models for nonignorable dropout. That is, for any given set of data, some parameters may not be estimable from the information in the data since the likelihood is exactly the same for a whole range of parameter values. In general, nonignorable dropout models are nonidentifiable unless some arbitrary constraints are imposed on the model.

References

[1] Diggle, P. & Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion), *Applied Statistics* **43**, 49–94.

[2] Laird, N.M. (1988). Missing data in longitudinal studies, *Statistics in Medicine* **7**, 305–315.

[3] Little, R.J.A. (1995). Modelling the dropout mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**, 1112–1121.

[4] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[5] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

GARRETT FITZMAURICE

Nonlinear Growth Curve

Traditional growth curve analysis is associated with the monitoring of development of individuals over time. A classic case involves recordings made on a group of children, say, of y (height or weight) at times x (age). For each child the points (x, y) can be plotted to give individual growth curves. Traditionally, again, low-degree polynomials would be fitted to the curves and the resulting parameter estimates used for inferences such as comparisons between different groups of children.

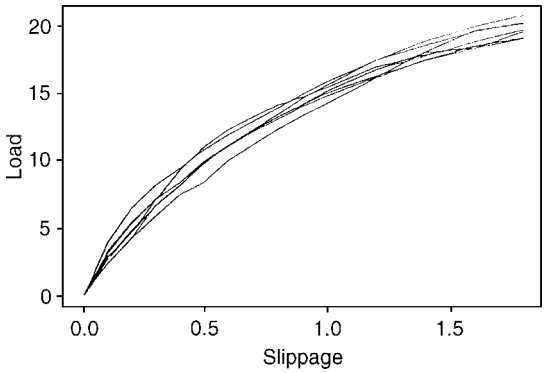

Figure 1 Timber slip under loading: load vs. slippage

Linear Growth Curves

Suppose that the observed value for a particular individual at time x_j is Y_j, $j = 1, \ldots, p$, and that the curve is to be represented as a quadratic (for $p > 3$). Then the statistical model is

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \varepsilon_j,$$

where $\beta = (\beta_0, \beta_1, \beta_2)'$ is the 3×1 vector of regression coefficients and ε_j is the "error" term. As an example, consider the data given in Table 1, plotted in Figure 1. The x values here are the loads required to produce slippage x of a timber specimen in a clamp. There are eight specimens, each producing 15 points on the curve. The eight individual curves in Figure 1 are quite close, and, by eye, it would not be unreasonable to entertain a quadratic fit.

Table 1 Timber slip data

	Timber specimen							
Slip	1	2	3	4	5	6	7	8
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.10	2.38	2.69	2.85	2.46	2.97	3.96	3.17	3.36
0.20	4.34	4.75	4.89	4.28	4.68	6.46	5.33	5.45
0.30	6.64	7.04	6.61	5.88	6.66	8.14	7.14	7.08
0.40	8.05	9.20	8.09	7.43	8.11	9.35	8.29	8.32
0.50	9.78	10.94	9.72	8.32	9.64	10.72	9.86	9.91
0.60	10.97	12.23	11.03	9.92	11.06	11.84	11.07	11.06
0.70	12.05	13.19	12.14	11.10	12.25	12.85	12.13	12.21
0.80	12.98	14.08	13.18	12.23	13.35	13.83	13.15	13.16
0.90	13.94	14.66	14.12	13.24	14.54	14.85	14.09	14.05
1.00	14.74	15.37	15.09	14.19	15.53	15.79	15.11	14.96
1.20	16.13	16.89	16.68	16.07	17.38	17.39	16.69	16.24
1.40	17.98	17.78	17.94	17.43	18.76	18.44	17.69	17.34
1.60	19.52	18.41	18.22	18.36	19.81	19.46	18.71	18.23
1.80	19.97	18.97	19.40	18.93	20.62	20.05	19.54	18.87

The matrix version of the quadratic model equation is $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{Y}$ is the $p \times 1$ vector with jth component Y_j, ε is $p \times 1$ with components ε_j, and $\mathbf{X}$ is $p \times 3$ with jth row $(1, x_j, x_j^2)$. The simplest assumption for the ε_js, $j = 1, \ldots, p$, is that they all have mean 0, all have the same variance σ^2, and are uncorrelated. In this case, least squares estimates and their standard errors can be obtained for the βs. More formal inferences can be made under the standard additional assumption that the ε_js are normally distributed. More generally, a *multivariate normal distribution* N $(\mathbf{0}, \mathbf{E})$, in which the ε_js may be correlated, will be adopted for ε. Often, a structured form will be adopted for $\mathbf{E}$ depending on a parameter vector τ.

The description above applies to the measurements made on one individual (e.g. timber specimen) only. To extend this to a group of individuals, denote the set of data points for the ith by $\{(x_{ij}, y_{ij}) : j = 1, \ldots, p_i\}$. Here, the individual subscript i is applied to x as well as to y to allow the possibility of different x values between individuals, and to p to allow the possibility of different numbers of measurements between individuals. In addition, the βs will also be enhanced with the subscript i: individuals tend to have different growth parameters. The linear model equation for the ith individual is then

$$Y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \beta_{2i} {x_{ij}}^2 + \varepsilon_{ij}, \qquad (1)$$

or, in vector notation, $\mathbf{Y}_i = \mathbf{X}_i \beta_i + \varepsilon_i$.

To perform an analysis on the group as a whole, rather than on individuals separately, the individual

model equations are linked via the parameters β_i. In the simplest case, where the β_is are supposed to differ from one another only by random variation between individuals in a homogeneous population, they can be represented as $\beta + \mathbf{b}_i$, where the $\mathbf{b}_i$s are independent with mean $\mathbf{0}$ and covariance matrix $\mathbf{B}$. The model for $\mathbf{Y}_i$ becomes

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{X}_i\mathbf{b}_i + \varepsilon_i.$$

In the more general formulation of Laird & Ware [2], the design matrices multiplying β and $\mathbf{b}_i$ are allowed to differ: thus, the model becomes

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i. \tag{2}$$

Nonlinear Models

The elements of the rows of $\mathbf{X}_i$ determine the shape of the fitted curve for individual i. In the quadratic example, row j of $\mathbf{X}_i$ is $(1, x_{ij}, x_{ij}^2)$, the presence of x_{ij}^2 making the curve nonlinear. However, this is not the nonlinearity referred to in the title of this subsection. The general model so far considered is linear in the regression parameter β, whatever the shape of curve determined by the elements of $\mathbf{X}_i$. It is only when parameters occur nonlinearly in the equation that the model is said to be nonlinear.

As an example, consider an exponential regression curve $y = \beta_0 + \beta_1 \exp(-\gamma x)$. Here, the parameters β_0 and β_1 occur linearly, but the parameter γ does not. We assume here that the value of γ is unknown and to be estimated from the data, like β_0 and β_1. Otherwise, γ would just be a known constant and the model would be linear. In the case of timber slippage, there is a well-established empirical law, $y = \alpha[1 - \exp(-\gamma x)]$, of exponential form with the constraint $\beta_1 = -\beta_0 = \alpha$. The exponential regression model for individual i is

$$Y_{ij} = \beta_{0i} + \beta_{1i} \exp(-\gamma_i x_{ij}) + \varepsilon_{ij}, \tag{3}$$

ε_{ij} being the usual zero-mean error term. The (3) model is similar to that of (1), the β parameters appearing linearly, but now with the additional γ parameter appearing nonlinearly. In vector notation this is $\mathbf{Y}_i = \mathbf{X}_i(\gamma_i)\beta_i + \varepsilon_i$, where $\mathbf{X}_i(\gamma_i)$ is $p \times 2$ with jth row $[1, \exp(-\gamma_i x_{ij}]$: the design matrix $\mathbf{X}_i$ now depends on the γ parameter. The more general Laird–Ware form, corresponding to (2), has

$$\mathbf{Y}_i = \mathbf{X}_i(\gamma_i)\beta + \mathbf{Z}_i(\gamma_i)\mathbf{b}_i + \varepsilon_i; \tag{4}$$

here, $\mathbf{Y}_i$ is of length p_i, $i = 1, \ldots, n$, and the error vectors ε_i are taken to be independent with mean $\mathbf{0}$ and covariance matrices $\mathbf{E}_i(\tau)$.

Not all nonlinear models present themselves in the standard form given above. For instance, the Bleasdale–Nelder form $y = (\alpha_1 + \alpha_2 x^{\phi})^{-\kappa}$ ([1], Section 8.1) has no linear parameters. However, it can be reformulated as $\beta(1 + \alpha x^{\phi})^{-\kappa}$, where $\beta = \alpha_1^{-\kappa}$ and $\alpha = \alpha_2/\alpha_1$, which does have a linear parameter.

Nonlinear Parameter Homogeneous Over Individuals

The statistical analysis is much simpler when there is no variation in γ_i over individuals and we consider this case first. Such homogeneity of γ, and of τ in the covariance structure, is tenable when these parameters are somehow more fundamental than the $\mathbf{b}_i$s, i.e. more like constants of nature than the individually varying $\mathbf{b}_i$-characteristics. For instance, in the exponential curve, β_{0i} and β_{1i} are location and scale parameters, respectively, for the measurements on individual i, whereas γ determines the intrinsic shape of the regression curve. The general model, (4), can be written as

$$\mathbf{Y}_i = \mathbf{X}_i(\gamma)\beta + \mathbf{Z}_i(\gamma)\mathbf{b}_i + \varepsilon_i, \tag{5}$$

where now γ has lost its subscript i.

Suppose that the random coefficients $\mathbf{b}_i$ have mean $\mathbf{0}$ and covariance matrix $\mathbf{B}(\tau)$, and that the ε_i and $\mathbf{b}_i$ are all independent. Then the $\mathbf{Y}_i$ are independent $p_i \times 1$ vectors with means and covariance matrices given by

$$\begin{aligned} \mu_i = \mathrm{E}(\mathbf{Y}_i) &= \mathbf{X}_i(\gamma)\beta, \\ \Sigma_i = \mathrm{cov}(\mathbf{Y}_i) &= \mathbf{Z}_i(\gamma)\mathbf{B}(\tau)\mathbf{Z}_i(\gamma)' + \mathbf{E}_i(\tau). \end{aligned} \tag{6}$$

Nonlinear Parameter Randomly Varying Over Individuals

In many circumstances it is necessary to allow for random variation in all the parameters, e.g. when modeling biological data, in view of the large natural variation among living things. Unfortunately, however, allowing the γ_i to vary randomly over individuals introduces a higher level of difficulty.

Referring again to the general model, (4), to calculate μ_i and Σ_i the operations $\mathrm{E}(\cdot)$ and $\mathrm{var}(\cdot)$ must be taken over the joint distribution of $\mathbf{b}_i$ and γ_i now, not just that of $\mathbf{b}_i$ in (6). To evaluate these, some distributional specifications will have to be made for γ_i. This will depend on the context, i.e. on what kind of parameter γ is; for instance, if γ were a rate parameter in an exponential decay model, a distribution on $(0, \infty)$ would be appropriate. In most practical cases **numerical integration** will be needed to evaluate these expectations and covariances.

Normal Likelihood Theory

For the between-individuals model, which describes the distribution of parameters over the population of individuals, suppose that $(\mathbf{b}_i, \gamma_i)$ has a joint probability density $f_{b\gamma}(\mathbf{b}_i, \gamma_i)$. Then, denoting the conditional density of $\mathbf{Y}_i$ given the individual parameters by $f(\mathbf{y}_i|\mathbf{b}_i, \gamma_i)$, the likelihood is $\mathbf{\Pi} f(\mathbf{y}_i)$, where $f(\mathbf{y}_i)$ is the unconditional density of $\mathbf{Y}_i$ obtained by integration over $(\mathbf{b}_i, \gamma_i)$:

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i|\mathbf{b}_i, \gamma_i) f_{b\gamma}(\mathbf{b}_i, \gamma_i)\, \mathrm{d}\mathbf{b}_i\, \mathrm{d}\gamma_i.$$

The integral is usually intractable for realistic models.

A certain amount of simplification is possible in some cases. If the conditional distribution of $\mathbf{b}_i|\gamma_i$ is taken as normal, so is that of $\mathbf{Y}_i|\gamma_i$. Then the expression for $f(\mathbf{y}_i)$ can be reduced by integrating out $\mathbf{b}_i$:

$$\begin{aligned} f(\mathbf{y}_i) &= \int f(\mathbf{y}_i|\mathbf{b}_i, \gamma_i) f_{b|\gamma}(\mathbf{b}_i|\gamma_i) f_\gamma(\gamma_i)\, \mathrm{d}\mathbf{b}_i\, \mathrm{d}\gamma_i \\ &= \int f(\mathbf{y}_i|\gamma_i) f_\gamma(\gamma_i)\, \mathrm{d}\gamma_i; \end{aligned}$$

here $f(\mathbf{y}_i|\gamma_i)$ is the multivariate normal density with mean $\mathrm{E}(\mathbf{Y}_i|\gamma_i)$ and covariance matrix $\mathrm{cov}(\mathbf{Y}_i|\gamma_i)$, and $f_\gamma(\gamma_i)$ is the marginal density of γ_i in the $(\mathbf{b}_i, \gamma_i)$ distribution. This integral, over γ_i only, is of smaller dimension than the preceding one. If there were no between-individuals variation in γ_i, $f_\gamma(\gamma_i)$ would be concentrated at a single point, say γ, and then the integral would reduce to $f(\mathbf{y}_i|\gamma)$, the density of $\mathrm{N}(\mu_i, \Sigma_i)$ with μ_i and Σ_i given in (6). The assumption of conditional normality of $\mathbf{b}_i$ given γ_i is just a natural extension of the usual assumption of unconditional normality of $\mathbf{b}_i$ given a fixed γ.

In general, $\mathbf{Y}_i$ will not have a normal distribution, even though it is conditionally normal given γ_i. Once $f_\gamma(\gamma_i)$ has been specified, $f(\mathbf{y}_i)$ can be computed by numerical integration, and a likelihood function thus obtained for inference.

Mean Curves

Consider again the exponential model (3), and suppose that $(\beta_{0i}, \beta_{1i}, \gamma_i)$ has mean $(\beta_0, \beta_1, \gamma)$ over the population of individuals. The "mean curve", obtained by inserting these mean parameter values, has the same exponential form. However, the "curve of means", i.e. the curve resulting from taking the expectation of (3) over the joint distribution of $(\beta_{0i}, \beta_{1i}, \gamma_i)$, is not generally of this form. This is because γ_i occurs nonlinearly in the equation: the form of the curve, $\mathrm{E}(Y_{ij})$ vs. x_{ij}, is determined by the form of $\mathrm{E}[\beta_{1i} \exp(-\gamma_i x_{ij})]$.

The curve of means defines average y responses at given x values, whereas the mean curve exhibits the characteristics of individual trajectories such as turning points and asymptotes. In linear models, and more generally in cases where γ_i is homogeneous over individuals, the mean curve and the curve of means coincide.

Further Reading

The literature on nonlinear growth curves is wide and scattered, much of it concerned with numerical approximations to the awkward integrations mentioned above. Some representative works are listed, and further details and examples (including timber slippage) are given, in Hand & Crowder [1, Section 8.4].

References

[1] Hand, D.J. & Crowder, M.J. (1996). *Practical Longitudinal Data Analysis*. Chapman & Hall, London.

[2] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

(*See also* **Random Coefficient Repeated Measures Model**)

M.J. CROWDER

Nonlinear Mixed Effects Models for Longitudinal Data

Experimental and observational studies in the medical, biological, and social sciences often result in data collected on specific *subjects* (or, in a more general sense, specific *experimental units*) that can be regarded as a sample drawn from the population of interest. When several observations are collected on each subject, the data are described as *repeated measures*. If the multiple measurements are indexed by the time of the observation, we say they are **longitudinal data**.

In the analysis of such data, we often must account for the variation between the subjects but, in doing so, we are not interested in the change in the response associated with specific subjects as much as we are interested in estimating the overall variation in the response within the population that can be attributed to subject-to-subject differences. *Mixed-effects models* are a flexible and powerful class of statistical models for use with longitudinal data and other types of repeated measures. These models incorporate both **fixed effects**, which we will write as $\boldsymbol{\beta}$, a vector of length p representing parameters associated with the entire population or with certain repeatable levels of experimental factors, and **random effects**, which we will write as $\mathbf{b}_i, i = 1, \ldots, m$, a collection of m vectors of length q that model the variation associated due to subject i in the observed collection of m subjects.

A fundamental difference between the fixed effects, $\boldsymbol{\beta}$, and the random effects, $\mathbf{b}_i$, is that the components of $\boldsymbol{\beta}$ are parameters in the statistical model but the components of $\mathbf{b}_i$ are not. The statistical model incorporates parameters that determine the *distribution* of the random effects. Sometimes these parameters can be expressed as **variance components** in the overall model.

Linear Mixed-effects Models

Laird and Ware [6] formulated a linear mixed-effects model for univariate repeated measures data as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$
$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \quad i = 1, \ldots, m,$$
$$\boldsymbol{\varepsilon}_i \perp \boldsymbol{\varepsilon}_j, \quad \mathbf{b}_i \perp \mathbf{b}_j, \quad i \neq j \quad \boldsymbol{\varepsilon}_i \perp \mathbf{b}_j, \quad \forall\, i, j, \tag{1}$$

where $\mathbf{y}_i$ is the vector of length n_i of responses for subject i; $\mathbf{X}_i$ is the $n_i \times p$ model **matrix** for subject i with respect to $\boldsymbol{\beta}$; and $\mathbf{Z}_i$ is the $n_i \times q$ model matrix for subject i and the random effects $\mathbf{b}_i$. The symbol $\perp$ indicates independence of random variables. The columns of the model matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ are based on any **covariates** that are observed along with the response for subject i.

The data shown in Figure 1 and described in [9, Appendix A.19]. are an example of **growth** *curve* data for which a **linear mixed-effects model** could be appropriate. Here the response is the height and the only covariate measured is the (scaled and centered) age of the boy at each observation. Each boy's height was measured nine times over a fixed range of age but the specific ages of the measurements vary from boy to boy.

In applying model (1) to these data, we may choose just an intercept and an age term for the fixed effects, the model matrix $\mathbf{X}_i$, in which case $\boldsymbol{\beta}$ would be of length 2, or we could use a quadratic model with an intercept, a column of age and a column of age^2 for which $\boldsymbol{\beta}$ would have length 3. The random effects model matrix could be a single column for the intercept, representing a random additive shift in height for each subject, or could have the intercept column and an age column, representing a random shift in the intercept and a random shift in the growth rate.

Nonlinear Mixed-effects Models

Although the linear mixed-effects model (1) is a versatile model, it does not encompass all of the mixed-effects models that are used in biostatistics. In some fields, the experimenters will have externally determined models of the mechanism determining the response as a function of the covariates. For example, it is common in **pharmacokinetics** to use **compartment models** [1, Chapter 5] to predict the subject's serum concentration of a drug as a function of the time since administration of the drug.

The data shown in Figure 2 and described in [9, Appendix A.29] are typical longitudinal data from

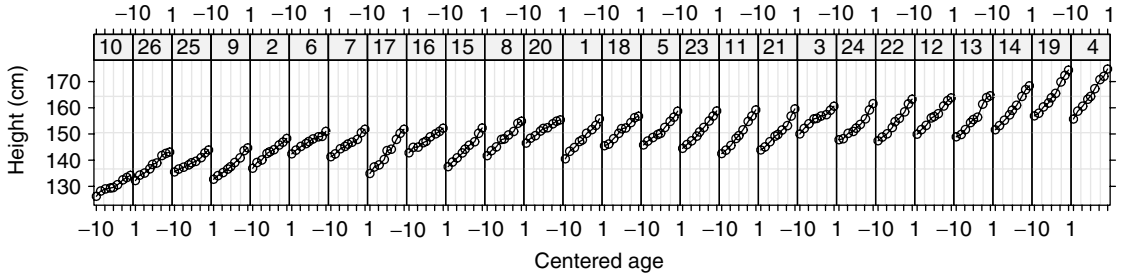

Figure 1 Heights (cm.) of a sample of 26 boys from Oxford, England versus their scaled, centered age

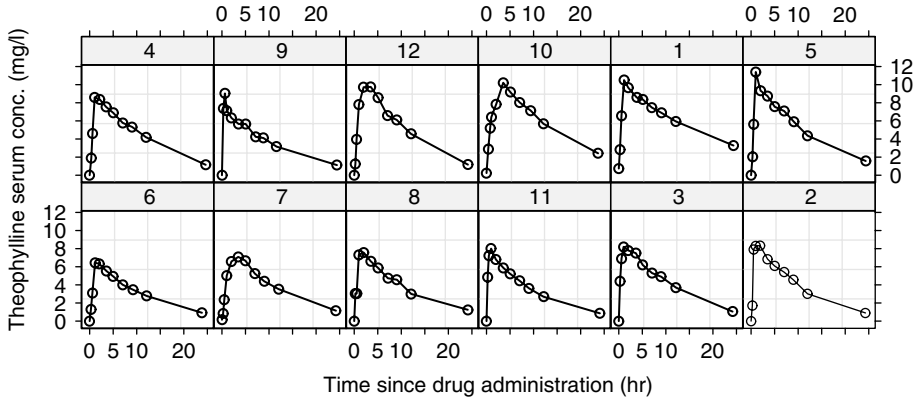

Figure 2 Serum concentration of theophylline (mg/l) versus time since administration of an oral dose of the drug. Each panel shows the concentration profile for one subject

a clinical pharmacokinetic study. The concentration profile for a single subject could be modeled as

$$c_t = \frac{Dk_e k_a}{Cl\,(k_a - k_e)}\left[\exp(-k_e t) - \exp(-k_a t)\right]. \quad (2)$$

where c_t is theophylline concentration in the serum at time t after an initial oral dose of D. The parameters in the model are the *elimination* rate constant k_e, the *absorption* rate constant k_a, and the *clearance* Cl.

The concentration profile (2) reflects many of the properties that we would expect in such data, such as $c_0 = 0$, a predicted concentration of zero at time zero, and $\lim_{t\to\infty} c_t = 0$, complete elimination of the drug after a long period of time.

All three of the parameters k_e, k_a, and Cl occur nonlinearly in (2). The desire to incorporate fixed effects and random effects in a population model based on nonlinear models for pharmacokinetic data motivated early formulations and development of nonlinear mixed-effects models [10]. One form [7] of the statistical model for the jth observation on the ith subject is

$$y_{ij} = f(\boldsymbol{\phi}_{ij}, \mathbf{x}_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i, \quad (3)$$

where the underlying model function f depends on a subject-specific parameter, $\boldsymbol{\phi}_{ij}$ and the values of covariates $\mathbf{x}_{ij}$ and is nonlinear in at least one component of $\boldsymbol{\phi}_{ij}$. The random effects and the fixed effects determine the subject-specific parameter through model matrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ of suitable dimension as

$$\boldsymbol{\phi}_{ij} = \mathbf{A}_{ij}\boldsymbol{\beta} + \mathbf{B}_{ij}\mathbf{b}_i, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i. \quad (4)$$

Using the vector function $\mathbf{f}_i$ with components $\{\mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)\}_j = f(\boldsymbol{\phi}_{ij}, \mathbf{x}_{ij})$, we can write model as

$$\begin{aligned} &\mathbf{y}_i = \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i) + \boldsymbol{\varepsilon}_i, \quad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \\ &\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \quad i = 1, \ldots, m \\ &\boldsymbol{\varepsilon}_i \perp \boldsymbol{\varepsilon}_j, \quad \mathbf{b}_i \perp \mathbf{b}_j, \quad i \neq j \quad \boldsymbol{\varepsilon}_i \perp \mathbf{b}_j, \quad \forall\, i, j. \end{aligned} \tag{5}$$

As in the linear mixed-effects model (1), the fixed effects $\boldsymbol{\beta}$ are common to the entire population and the random effects $\mathbf{b}_i$ are specific to the subject. In this model, the parameter vector for the underlying nonlinear model, $\boldsymbol{\phi}_{ij}$, can depend on both the subject and the observation because the model matrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ can depend on the observation j. This allows the model to incorporate **time-dependent covariates**.

The use of the matrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ may be confusing. To make this more concrete, let us consider a specific model [9, §8.2] for the theophylline data. The parameters Cl, k_e, and k_a, in (2) must be positive for the model to be physically meaningful so we reexpress (2) in terms of the logarithms of these parameters.

$$\begin{aligned} c_t = \frac{D \exp(lKe + lKa - lCl)}{\exp(lKa) - \exp(lKe)} &\left\{\exp\left[-\exp(lKe)\, t\right]\right. \\ &\left. - \exp\left[-\exp(lKa)\, t\right]\right\}, \end{aligned} \tag{6}$$

where $lCl = \log(Cl)$, $lKe = \log(k_e)$ and $lKa = \log(k_a)$. The subject-specific parameter is $\boldsymbol{\phi}_{ij} = \boldsymbol{\phi}_i = (lCl, lKe, lKa)'$ and the covariate vector for the jth observation on the ith subject is $\mathbf{x}_{ij} = \left(D_i, t_{ij}\right)'$, representing the dose given to subject i and the time of the jth concentration measurement for subject i.

As shown in [9, Section 8.2], a reasonable nonlinear mixed-effects model for these data can be formulated with $\boldsymbol{\beta}$ having the same components as $\boldsymbol{\phi}_i$ and with random effects only for lCl and lKe. That is, $\boldsymbol{\phi}_i$ has length 3 as does $\boldsymbol{\beta}$ while $\mathbf{b}_i$ has length 2. The matrices $\mathbf{A}_{ij} = \mathbf{I}_3$, the 3×3 identity matrix, and $\mathbf{B}_{ij}$, which is the first two columns of $\mathbf{I}_3$, do not change with i or j.

It is possible to use a more general form of the dependence of $\boldsymbol{\phi}_{ij}$ on $\boldsymbol{\beta}$ and $\mathbf{b}_i$ [4, Section 4.2] but the ability to reparameterize the nonlinear model $f(\boldsymbol{\phi}_{ij}, \mathbf{x}_{ij})$ in another set of parameters $\mathbf{g}(\boldsymbol{\phi})$ allows other forms of dependence on $\boldsymbol{\beta}$ and $\mathbf{b}_i$ to be rewritten as (4).

Estimation of Parameters

The parameters in the nonlinear mixed-effects model (5) are $\boldsymbol{\beta}$, the fixed effects, σ^2, the variance of the within-subject random variation, and $\boldsymbol{\theta}$, some set of parameters that determine $\boldsymbol{\Sigma}$, the variance–covariance matrix of the random effects or, equivalently, the set of parameters that determine the relative variance–covariance Σ/σ^2. Although many estimation criteria for these parameters have been proposed [4], the most common is **maximum likelihood**.

To evaluate the **likelihood** function, we must determine the **marginal** density of the $n = \sum_{i=1}^{m} n_i$-dimensional response $\mathbf{y}$ from the model (5)

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\mathbf{y}) &= \prod_{i=1}^{M} p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \\ &= \prod_{i=1}^{M} \int p(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \sigma^2)\, p(\mathbf{b}_i|\boldsymbol{\theta}, \sigma^2)\, \mathrm{d}\mathbf{b}_i. \end{aligned} \tag{7}$$

In theory, it is straightforward to define the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$, $\widehat{\sigma^2}$, and $\hat{\boldsymbol{\theta}}$ as the values that optimize (7). In practice, evaluation of the integral in (7) and **optimization** of the resulting likelihood or log-likelihood function is difficult. Software for estimating the parameters in a nonlinear mixed-effects model uses approximations to the integral in (7) or to the likelihood function itself. The nlme package for **S-PLUS** and for **R** [5] uses an approximation to the log-likelihood based on conditional estimates $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and conditional modes $\hat{\mathbf{b}}_i(\boldsymbol{\theta})$ obtained from a penalized nonlinear **least squares** problem [9, 7] (*see* **Penalized Maximum Likelihood**). PROC NMIXED in SAS uses adaptive Gaussian quadrature [8], which also is based on the conditional modes $\hat{\mathbf{b}}_i(\boldsymbol{\theta})$ of the random effects. The NONMEM program [2, 3] which is widely used in pharmacokinetics, can use different approximations to the log-likelihood including a first-order Taylor expansion of the model function around $\mathbf{0}$, the expected value of the random effects vector $\mathbf{b}$.

This is but a brief introduction to the formulation of nonlinear mixed-effects models and to methods of estimating the parameters in such models. There are, naturally, many other aspects of these models that we

have not discussed but are covered in references such as [4, 9].

References

[1] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

[2] Beal, S. & Sheiner, L. (1980). The NONMEM System, *American Statistician* **34**, 118–119.

[3] Beal, S. & Sheiner, L. (1984). NONMEM Users' Guide. University of California, San Francisco.

[4] Davidian, M. & Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London.

[5] Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**, 299–314.

[6] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[7] Lindstrom, M.J. & Bates, D.M. (1990). Nonlinear mixed-effects models for repeated measures data, *Biometrics* **46**, 673–687.

[8] Pinheiro, J.C. & Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics* **4**, 12–35.

[9] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-effects Models in S and S-PLUS*. Statistics and Computing, Springer.

[10] Sheiner, L.B. & Beal, S.L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis–Menten model: routine clinical pharmacokinetic data, *Journal of Pharmacokinetics and Biopharmaceutics* **8**, 553–571.

(*See also* **Random Coefficient Repeated Measures Model**; **Nonlinear Growth Curve**)

DOUGLAS M. BATES

Nonlinear Regression

One of the most common situations in statistical analysis is that of data which consist of observed responses Y_i thought to be related to corresponding k-dimensional inputs $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_k})'$. Nonlinear regression is used when this situation may be represented by the regression equation

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where the form of the **expectation** function f is entirely known except for the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ and f is a nonlinear function of $\boldsymbol{\theta}$. In this model, the random variable Y_i, often called the response or dependent variable, represents the response for case i and the variable $\mathbf{x}_i$, usually called the explanatory (*see* **Explanatory Variables**) or independent variable, may represent an experimental setting or predetermined conditions associated with the ith response. A wide variety of assumptions concerning the error terms $\{\varepsilon_i\}$ are possible, but the most frequent one is that they are independent and identically distributed (iid) normal random variables.

Linearity or nonlinearity of a model depends on how the parameters occur in the expectation function, but not on how the explanatory variables do, and a nonlinear regression model is one in which at least one of its parameters appears nonlinearly. For example,

$$f(\mathbf{x}, \boldsymbol{\theta}) = \theta_1 + \theta_2 x_1 + \theta_3 x_2^2 + \theta_4 x_1 x_2$$

is a linear model as the expression is linear in the parameters, whereas

$$f(\mathbf{x}, \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \tag{2}$$

is a nonlinear model, being nonlinear in θ_2. More formally, *nonlinear* means that at least one of the derivatives of the expectation function f with respect to the parameters is a nontrivial function of at least one of those parameters.

Some nonlinear models can be transformed to a linear one by taking a suitable transformation (*see* **Transformations**) of the data. For example, in the model in (2) we can take a log transformation of the response variable, $Y_i^* = \log Y_i$, and setting $\theta_1^* = \log \theta_1$, $\theta_2^* = \theta_2$, we obtain a new expectation function

$$f^*(\mathbf{x}, \boldsymbol{\theta}^*) = \theta_1^* + \theta_2^* x.$$

This model is now linear in its unknown parameters θ_1^* and θ_2^* and therefore we could estimate these parameters using linear regression (*see* **Linear Regression, Simple**) of the logarithm of the data on the explanatory variable $\mathbf{x}_i$. However, assuming that the error in Y_i is additive, the model for Y_i^* is

$$\begin{aligned} Y_i^* &= \theta_1^* + \theta_2^* x_i + \log\left[1 + \frac{\varepsilon_i}{f(\mathbf{x}_i, \boldsymbol{\theta})}\right] \\ &= \theta_1^* + \theta_2^* x_i + \varepsilon_i^*, \end{aligned}$$

and now the variance of the error becomes dependent on x_i through the expectation function $f(\mathbf{x}_i, \boldsymbol{\theta})$. Thus, transforming the data results in a transformation of both the expectation function and the disturbance term, and so the usual assumption of constant variance and normality required for simple linear regression may no longer be valid. This may lead to serious deficiencies in the estimates of $\boldsymbol{\theta}^*$ and hence of $\boldsymbol{\theta}$. The decision whether or not to make a linearizing transformation depends very much on the nature of the errors, and linearization should only be used when the transformed data are adequately described by a model with an additive normal error.

Another type of transformation of a model is a parameter transformation where the parameters of the new model are related to the parameters of the old one by an expression which involves parameters only and not the explanatory variables $\mathbf{x}_i$. For example, consider the following models:

$$Y_i = \theta_1 \frac{x_i}{x_i + \theta_2} + \varepsilon_i$$

and

$$Y_i = \frac{x_i}{\theta_1^* x_i + \theta_2^*} + \varepsilon_i.$$

If we define $\theta_1^* = 1/\theta_1$ and $\theta_2^* = \theta_2/\theta_1$, then the two models (parameterizations) produce identical values of Y_i for the same value of x_i. However, the statistical properties of estimators in one of these models may be much better than in the other, in the sense that the former may have properties closely approaching those of estimators in linear regression models. This is an important feature of nonlinear regression and we give more details regarding this problem in the final section.

If the joint distribution of $\{\varepsilon_i\}$ in the model in (1) is assumed known, then the parameter $\boldsymbol{\theta}$ can be estimated through the use of the **maximum likelihood** method. Under the assumption that the $\{\varepsilon_i\}$

are iid normal random variables with a constant but unknown variance σ^2, the maximum likelihood estimates of $\boldsymbol{\theta}$ are the **least squares** values, which, by definition, minimize the sum of squares

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} [Y_i - f(\mathbf{x}_i, \boldsymbol{\theta})]^2$$

over all possible values of $\boldsymbol{\theta}$. Other methods of estimation include a weighted least squares method which allows for variance heterogeneity in the model, robust estimation (*see* **Robust Regression**), which may be considered as an alternative to the least squares method if **outliers** are present in data, and *quasi-likelihood* estimation which does not require the distribution of ε_i to be known explicitly, except for its first two moments. These methods and others, including Bayes estimation (*see* **Bayesian Methods**), are discussed by Bates & Watts [4] and Seber & Wild [18]. In what follows we discuss mainly the least squares method assuming the validity of the assumption of iid normal error.

In contrast to linear models, in nonlinear regression it is not possible to write down an explicit expression for the least squares estimators of the parameters. In addition, the least squares estimators in this case have essentially unknown properties for finite sample sizes. In particular, they are usually neither unbiased, normally distributed, nor *minimum variance estimators* as are those in linear models. The estimators achieve these properties only asymptotically as the sample sizes approach infinity (*see* [11] and **Large-Sample Theory** for a detailed development of asymptotic theory). It is not possible, in general, to present any guidelines as to how large the sample size must be for these asymptotic properties to be closely approximated.

This difference between estimators in linear and nonlinear models should be remembered, since in practice inference regions for nonlinear models are often approximated using linear models. A number of measures and procedures have been developed for studying the behavior of the estimators in nonlinear cases and they provide information on the adequacy of linear approximation inference regions and indicate situations where by changing the parameters a nonlinear model behaves more like a linear model. A brief description of some measures of nonlinearity is given in the final section.

Examples of Applications

An important step in nonlinear regression is the specification of the model, including both the expectation function and the characteristics of the disturbance (*see* **Model, Choice of**). Frequently the expectation function is tentatively suggested by theoretical investigations and the analyst's job is then to find the simplest form of the model and the parameter estimates which provide an adequate fit of the model to the data, subject to the assumptions about the disturbance. However, in many situations, particularly in the biological sciences, the underlying processes are complex and no physically meaningful models for the expectation functions may be advanced. In such cases, the statistician may suggest a model which has the same sort of behavior as the data. In this search, particularly helpful may be the literature, such as the handbook by Ratkowsky [16], which examines commonly used nonlinear regression models.

Nonlinear regression has been applied to a wide range of situations and we now give a selection of examples.

Example 1

Processes producing sigmoidal growth curves are widespread in many fields of study. In medicine, for example, we may be interested in the normal growth of infants (*see* **Growth and Development**) or the growth of a tumor, and the effect of treatments upon such growth. It is usually assumed that the theoretical growth curve belongs to a known parametric family of curves which start at some fixed point and increase their growth rate monotonically to reach an inflection point; after this the growth rate decreases to approach asymptotically some final value. A number of functions have been proposed for modeling such curves (e.g. [18, Chapter 7]), many of which are claimed to have some underlying theoretical basis. In most cases, however, the approach to the analysis of growth data is purely empirical and involves fitting a parametric family of curves to the data. Two examples of growth curves are the Gompertz function

$$f(x, \boldsymbol{\theta}) = \theta_1 \exp[-\exp(\theta_2 - \theta_3 x)]$$

and the logistic function

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1}{1 + \exp(\theta_2 - \theta_3)x}.$$

Usually, the variable x represents time, but similar models occur also in situations when the explanatory variable is an increasing intensity of some other factor, such as the amount of nutrients in a diet and its effect on weight gain.

Example 2

Biological assay (bioassay) is a method for estimating the potency of a drug or material by the study of the reaction caused by its use on experimental subjects. In an indirect assay specified doses are each given to a set of experimental units and the resulting responses are recorded. For quantitative responses the assay can then be analyzed by fitting a nonlinear dose–response regression model to the data. Various functions having different characteristics have been proposed to fit dose–response curves of a generally sigmoidal shape. For example, in radioligand assays (*see* **Radioimmunoassay**), in which potency estimation involves the relation between counts of radioactivity and dose, many researchers have found the regression function to be satisfactorily represented by the four-parameter logistic model (e.g. [10]), which can be written as

$$f(x, \boldsymbol{\theta}) = \theta_1 + \frac{\theta_2}{1 + \exp(\theta_3 - \theta_4 x)},$$

where x is the logarithm of the dose. This representation is only one of many possible parameterizations and, although different parameterizations give the same predicted value of the response for a specified dose, from a statistical perspective they are not all equivalent [17].

Example 3

The forced expiratory volume Y, a measure of how much air a subject can breathe, is widely used epidemiologically for screening against chronic respiratory disease. Cole [7] discusses a nonlinear model relating Y to height ($x_{i,1}$) and age ($x_{i,2}$) of the form

$$Y_i = x_{i,1}^{\theta_1}(\theta_2 + \theta_3 x_{i,2}) + \varepsilon_i.$$

This model appears to be equally suitable for predicting maximum oxygen uptake which is the best-known measure of an individual's capacity to deliver oxygen to, and to use oxygen in, exercising muscle [14]. In this case $x_{i,1}$ denotes the body weight.

Example 4

It is not necessary for the expectation function to be an explicit function of the parameters and the explanatory variables. For example, in an important class of models, known as compartment models, the expected response is given by the solution to a set of linear differential equations. These models are commonly used in **pharmacokinetics**, where the exchange of materials in biological systems is studied. To estimate the parameters in such models several methods can be used. In some cases it is possible to obtain an analytic solution to the system of differential equations and then use the expectation function, corresponding to the compartment for which data are available, in a standard nonlinear estimation program. However, special techniques have been developed which allow one to avoid solving explicitly for the expectation function and its derivatives [5, 13].

In other situations, the expectation function may be the solution to a nonlinear differential equation or a partial differential equation which has no analytic solution. Then the values of the expectation functions must be determined numerically for any given parameter values. In such situations numerical derivatives or derivative-free *optimization* procedures often have to be used to calculate the least squares estimates. Dalgaard & Larsen [8], for example, proposed an **algorithm** for the analysis of data obtained by vitreous fluorophotometry, a method in clinical eye research. The model involves the *diffusion* equation, and the parameters are the diffusion coefficient in the vitreous body of the eye and the permeability of the blood–retinal barrier.

Calculating Parameter Estimates

Most of the methods of estimation in nonlinear regression require an estimator obtained by maximizing or minimizing some objective function of $\boldsymbol{\theta}$, like $S(\boldsymbol{\theta})$ for the least squares estimator. Since these optimization problems can seldom be solved analytically, the optimal value of $\boldsymbol{\theta}$ must be in most cases located by iterative techniques using a computer. One option is to use standard nonlinear optimization algorithms such as modifications of the Newton method or conjugate gradient method, which are widely available in a number of numerical libraries (see, for example, [12] and [18, Chapter 15]). However, when choosing an

algorithm, some understanding of how it works and its limitations may be necessary. Another option is to use specialized software which exploits the particular structure of a nonlinear problem. For least squares, most such methods are based on the Gauss–Newton algorithm, which we now briefly describe.

For a given data set, the values of $\mathbf{x}_i$ are fixed, and as we vary $\boldsymbol{\theta}$ through all possible values, the expectation vectors

$$\begin{aligned}\mathbf{f}(\boldsymbol{\theta}) &:= (f_1(\boldsymbol{\theta}), \ldots, f_n(\boldsymbol{\theta}))' \\ &:= (f(\mathbf{x}_1, \boldsymbol{\theta}), \ldots, f(\mathbf{x}_n, \boldsymbol{\theta}))'\end{aligned}$$

generate a surface in R^n, called the *expectation surface*. The Gauss–Newton algorithm uses a linear approximation to this surface as follows:

1. Near an initial point $f(\boldsymbol{\theta}^{(0)})$ the expectation surface is approximated by its tangent plane.
2. The observation vector $\mathbf{y}$ is projected onto the tangent plane by linear regression to obtain a new parameter vector $\boldsymbol{\theta}^{(1)}$.
3. The tangent plane is calculated at $f(\boldsymbol{\theta}^{(1)})$ and the procedure is continued until either convergence or abnormal termination.

Formally the process can be justified using a Taylor series expansion. Suppose that $\boldsymbol{\theta}^{(0)}$ is an approximation to the least squares estimate $\hat{\boldsymbol{\theta}}$. For $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}^{(0)}$ we have the linear Taylor expansion:

$$f_i(\boldsymbol{\theta}) \approx f_i(\boldsymbol{\theta}^{(0)}) + \sum_{s=1}^{p} \frac{\partial f_i}{\partial \theta_s}(\boldsymbol{\theta}^{(0)})(\theta_s - \theta_s^{(0)}),$$
$$i = 1, \ldots, n,$$

or

$$\mathbf{f}(\boldsymbol{\theta}) \approx \mathbf{f}(\boldsymbol{\theta}^{(0)}) + \mathbf{V}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}), \tag{3}$$

where $\mathbf{V}_0 = \mathbf{V}(\boldsymbol{\theta}^{(0)}) = [\partial f_i(\boldsymbol{\theta}^{(0)})/\partial \theta_s]$ is the derivative matrix evaluated at $\boldsymbol{\theta}^{(0)}$. With approximation (3) the original model can be rewritten as follows:

$$\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^{(0)}) \approx \mathbf{V}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. Since this approximation is linear in the parameter $\boldsymbol{\beta}$, we can use linear regression to obtain the least squares estimate of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^{(0)} = (\mathbf{V}_0'\mathbf{V}_0)^{-1}\mathbf{V}_0'\mathbf{r}^{(0)},$$

where $\mathbf{r}^{(0)} = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(0)})$ is the residual vector at $\boldsymbol{\theta}^{(0)}$. This suggests that for a given current approximation $\boldsymbol{\theta}^{(0)}$, the next approximation should be

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \hat{\boldsymbol{\beta}}^{(0)}. \tag{4}$$

This provides an iterative scheme for obtaining $\hat{\boldsymbol{\theta}}$ and usually is referred to as the Gauss–Newton method. If the starting value $\boldsymbol{\theta}^{(0)}$ is sufficiently close to a local minimum $\hat{\boldsymbol{\theta}}$, the algorithm will converge. The unmodified algorithm, however, is rarely used in practice. To deal primarily with ill-conditioning of the derivative matrix $\mathbf{V}$, which may cause, for example, the sum of squared residuals at $\boldsymbol{\theta}^{(i+1)}$ to be greater than $\boldsymbol{\theta}^{(i)}$ on the ith iteration, and to avoid having to code and specify the derivatives, modifications to this method, as well as alternative methods, have been suggested. The most common of these modifications are the Hartley and Levenberg–Marquardt algorithms. Another modification, described in [4], forms the basis of the algorithm implemented in S-PLUS (*see* **S-PLUS and S**). When the residuals for the fitted model are large, the above algorithms may converge very slowly or even fail. In such situations two alternative methods can be used: the NL2SOL algorithm and the algorithm due to Gill and Murray (see [18, Chapter 14] for details about these and other methods).

Optimization programs require specification of convergence criteria and initial parameter estimates $\boldsymbol{\theta}^{(0)}$. Obtaining good starting values is essential to guarantee convergence, and many methods for determining them have been proposed (e.g. [15] and [4]). Convergence criteria are often based on the relative change in $S(\boldsymbol{\theta})$ from one iteration to the next or on the relative change in the components of $\boldsymbol{\theta}$. Although they are usually reliable, they are not unambiguous and criteria using different rules have also been proposed [2]. In practice, the concurrent use of several criteria is often recommended.

Confidence Regions and Effects of Nonlinearity

Several methods for determining confidence regions for the parameter $\boldsymbol{\theta}$ (*see* **Estimation**) have been proposed. These include an exact method, usually referred to as the lack-of-fit method, the **likelihood** confidence regions, and linearization confidence

regions (e.g. [11] and [9]). Only the first method gives exact confidence regions, but it is computationally the most difficult. Likelihood regions are generally easier to calculate but are still computationally tedious compared with the linearization method, which is the most commonly used. In this method confidence regions and confidence intervals are obtained using the method of linear regression with the estimated variance–covariance matrix approximated by $(\hat{\mathbf{V}}'\hat{\mathbf{V}})^{-1}s^2$, where $\hat{\mathbf{V}}$ is the derivative matrix calculated at $\hat{\boldsymbol{\theta}}$ and s^2 is the estimate of the residual variance. Unlike the exact and likelihood regions, the linearization region is always ellipsoidal. However, the extent to which such a region approximates the exact region depends on the extent of nonlinearity of the model, and in some cases the discrepancies may be very large.

To provide information on the adequacy of the linear approximation, Beale [6] and Bates & Watts [1, 3, 4] have developed a number of measures of nonlinearity using a geometric approach. For nonlinear models the expectation surface is curved, and this is in marked contrast to the linear models where the expectation surface is always planar. This nonplanarity has been called the *intrinsic nonlinearity*. The term *intrinsic* is suitable, as this nonlinearity becomes determined the moment a model/data set combination has been defined, and it cannot be altered by reparameterization.

The other way in which linear and nonlinear models differ involves the position on the expectation surface of the values having equal increments in $\boldsymbol{\theta}$. For a linear model, points $f(\boldsymbol{\theta})$ representing constant values of $\boldsymbol{\theta}$ are straight, parallel and equally spaced for equal increments of $\boldsymbol{\theta}$, which in general is not true for nonlinear regression models. This nonuniformity of the spacing defines the second component of nonlinearity, termed *parametric effects nonlinearity*. This component depends on how parameters occur in the expectation function and it can be changed by reparameterization.

To quantify the extent of the two components of nonlinearity, Bates & Watts derived curvature measures based on second-order derivatives of the expectation function. As found by these authors, and in subsequent studies, the intrinsic nonlinearity is typically small for most models of practical interest. In contrast, the parametric curvature is usually large enough to affect linear approximation inference regions. However, a large parametric curvature can often be substantially reduced by finding an appropriate reparameterization. At present little guidance is available as to the choice of a reparameterization, although occasionally certain transformations are recommended from practical experience or are suggested by the model itself (e.g. [15–17] for the four-parametric logistic model).

For detailed information about measures of curvature and nonlinearity and their applications in assessing the appropriateness of linear approximations, see [4] and [18].

References

[1] Bates, D.M. & Watts, D.G. (1980). Relative curvature measures of nonlinearity (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 1–25.

[2] Bates, D.M. & Watts, D.G. (1981). A relative offset orthogonality convergence criterion for nonlinear least squares, *Technometrics* **123**, 179–183.

[3] Bates, D.M. & Watts, D.G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares, *Annals of Statistics* **9**, 1152–1167.

[4] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

[5] Bates, D.M., Wolf, D.A. & Watts, D.G. (1986). Nonlinear least squares and first order kinetics, in *Computer Science and Statistics: The Interface*, D.M. Allen, ed. Elsevier, New York, pp. 71–81.

[6] Beale, E.M.L. (1960). Confidence regions in non-linear estimation (with discussion), *Journal of the Royal Statistical Society, Series B* **22**, 41–88.

[7] Cole, T.J. (1975). Linear and proportional regression models in the prediction of ventilatory function, *Journal of the Royal Statistical Society*, Series A **138**, 297–337.

[8] Dalgaard, P. & Larsen, M. (1990). Fitting numerical solutions of differential equations to experimental data: a case study and some general remarks, *Biometrics* **46**, 1097–1109.

[9] Donaldson, J.R. & Schnabel, R.B. (1987). Computational experience with confidence regions and confidence intervals for nonlinear least squares, *Technometrics* **29**, 67–82.

[10] Finney, D.J. (1976). Radioligand assay, *Biometrics* **32**, 721–740.

[11] Gallant, A.R. (1987). *Nonlinear Statistical Models*. Wiley, New York.

[12] Gill, P.E., Murray, W. & Wright, M.H. (1981). *Practical Optimization*. Academic Press, London.

[13] Jennrich, R.I. & Bright, P.B. (1976). Fitting systems of linear differential equations using computer generated exact derivatives, *Technometrics* **18**, 385–399.

[14] Nevill, A.M. & Holder, R.L. (1994). Modelling maximum oxygen uptake – a case-study in non-linear regression model formulation and comparison, *Applied Statistics* **43**, 653–666.
[15] Ratkowsky, D.A. (1983). *Nonlinear Regression Modeling*. Marcel Dekker, New York.
[16] Ratkowsky, D.A. (1990). *Handbook of Nonlinear Regression Models*. Marcel Dekker, New York.
[17] Ratkowsky, D.A. & Reedy, T.J. (1986). Choosing near-linear parameters in the four-parameter logistic model for radioligand and related assays, *Biometrics* **42**, 575–582.
[18] Seber, G.A.F. & Wild, C.J. (1989). *Nonlinear Regression*. Wiley, New York.

(*See also* **Optimization and Nonlinear Equations**)

ADAM W. KOLKIEWICZ

Nonlinear Time Series Analysis

Nonlinear time series analysis is currently a most active area of research in **time series** and dynamical systems literature. This is partly because of the enormous interest in nonlinear dynamical systems, including **chaos**, since the early 1970s. A number of recent textbooks on time series analysis have included some coverage of this important area; see, for example [3] and [5]. For a wider and more in-depth coverage, we refer to [9, 16] and [19]. These cover roughly the work up to the early 1990s. For more recent surveys, we refer to [18] and [20].

Nowadays, the term "nonlinear time series" is used in both the statistical and the dynamical systems literature. Although there is much common ground, the emphases are different, with the former focusing on the randomness generated by a stochastic system/model and the latter on the randomness generated by a purely deterministic system/model. In this article, we shall concentrate on the former, although passing references will be made to the latter where appropriate.

To date, biostatistical applications using the statistical analysis of nonlinear time series include epidemiology (e.g. [21] and [22]) and physiology (e.g. [19]) and others. For similar applications, but using the dynamical systems approach, see, for example [6, 10, 12, 13] and [21].

Why Nonlinearity?

At the simplest level, linearity may be exemplified by the doubling of the deflection of the indicator of the balance if we place two copies of this article instead of one on the balance. Departure from this doubling is often associated with nonlinearity. In fact, it is obvious that we do not expect the balance to keep doubling the deflection of its indicator every time we double the number of copies on it. This simple example illustrates a fundamental fact: in reality, *linearity cannot hold in the large (or globally) although it may hold in the small (or locally)*. For a deeper discussion of the notion of nonlinearity in the context of time series analysis, we refer to, for example, [1, 16], and [19].

In time series analysis, linear methodology culminates in the classic family of autoregressive moving average models (*see* **ARMA and ARIMA Models**) including the vector cases, e.g. [2] and [11]. However, there are many observable phenomena that cannot be explained properly if we restrict ourselves to linear systems/models. The above, rather simple, example of a balance is, in fact, an example of the so-called saturation effect, in that the indicator has only a finite range of deflection. The saturation effect is common to many real systems, since almost all real systems have finite capacity/energy only.

Cycles are another often-observed phenomenon in many time series and these may be due to the bifurcation of a single global linear dynamics into multiple local dynamics. For example, Watier & Richardson [22] have studied the cycles observed in the monthly notifications of *Salmonella typhimurium* in France from January 1978 to December 1988. They have suggested that the increase in the prevalence of some phage types from the cold period to the warm period could alter the epidemiological system from one approximately linear dynamics to another and they have modeled the switching by reference to the notification of seven months previously. They have used the fitted nonlinear time series model (specifically a self-exciting threshold autoregressive model, whose form will be given in the next section) to explain the cycles by showing that, in the absence of the background noise that is modeled as a sequence of independent identically distributed random variables, the nonlinear dynamics of the fitted model generate cycles that bear some resemblance to the observed cycles. As a digression, the Watier–Richardson example also shows a possible intimate connection between nonlinearity (e.g. different dynamics corresponding to different states of the phage types) and nonstationarity (e.g. different dynamics corresponding to different seasons), in some cases (*see* **Stationarity**). We can sometimes put both notions, at least formally, on the same footing by treating time as a covariate/state variable as well.

Returning to our discussion, there are other observable phenomena that can only be explained properly by reference to nonlinear dynamics. These include chaos, synchronization (i.e. how the

frequency of oscillations of a system can get locked into that of an external oscillator), jump phenomenon (i.e. how the amplitude of oscillations of the output signal of a system is affected when we alter the frequency of the input signal), and others; e.g. [19].

Various Approaches to Nonlinear Time Series Modeling

Parametric

There are essentially two major approaches to nonlinear time series modeling in statistical literature; namely, the parametric approach and the nonparametric approach. In the former, several classes were proposed in the late 1970s and the early 1980s. These included the threshold autoregressive models (TAR), the autoregressive models with conditional heteroscedasticity (ARCH), bilinear models (BL), the nonlinear moving average models (NLMA), the polynomial autoregressive models, random coefficient autoregressive models (RAC), and others. Tong [19] has given a fairly comprehensive list of these various parametric models, together with some of their probabilistic structure and statistical estimation. Granger & Terasvïrta [9] and Priestley [16] are also relevant whilst Engle & McFadden [7], Nicholls & Quinn [15] and Rao & Gabr [17] give more details on specific models.

Given the vast number of parametric families of models, it would be impossible to survey them properly in this article. However, in order to give the readers some general flavor, we refer to Watier–Richardson's model mentioned in the last two sections. Let $\{X_t\}$ denote the monthly notifications of cases of *Salmonella typhimurium* in France over the period January 1978 to December 1988. Their fitted model takes the form

$$X_t = 74.84 + 0.378X_{t-1} - 0.367X_{t-4} + 0.341X_{t-12} + \varepsilon_t^{(1)}, \quad (1)$$

if $X_{t-7} \leq 81$, and

$$X_t = -0.392 + 0.658X_{t-1} + 0.314X_{t-11} + \varepsilon_t^{(2)}, \quad (2)$$

if $X_{t-7} > 81$. Here, $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$ denote two independent sequences of zero-mean, independent and identically distributed random variables (to be referred to as iid), where $\varepsilon_t^{(1)}$ has an estimated standard deviation 32.53 and $\varepsilon_t^{(2)}$ 14.66. Essentially, the nonlinearity is modeled by dividing the state space (i.e. the space of values of X) into two regimes, one corresponding to $X_{t-7} \leq 81$ and the other to $X_{t-7} > 81$, and inside each regime a linear autoregressive model is used. This model is an example of the class of self-exciting threshold autoregressive models; the adjective "self-exciting" refers to the fact that the division is effected by reference to a covariate that happens to be some past value of the response. We have already mentioned some of the properties of this fitted model in the last section.

The threshold models are currently one of the more popular nonlinear time series models in the literature. It is clear that conditional on the past X values up to and including $t-1$, the expectation of X_t is a piecewise linear function of the past X values. The primary, although by no means exclusive, objective of the above model is clearly the general drift of the time series. In econometric literature, often the interest lies not so much in the drift but rather in the diffusion about this drift (e.g. the volatility of the stock market). Translated into statistical language, the focus is then primarily on the conditional variance term. Perhaps for this reason, the class of ARCH models is particularly popular in the econometric literature. In its simplest form, it may be written as

$$X_t = \varepsilon_t \sqrt{V_t}, \quad (3)$$

where $\{\varepsilon_t\}$ are iid and

$$V_t = \gamma + \phi_1 X_{t-1}^2 + \ldots + \phi_q X_{t-q}^2, \quad (4)$$

with $\gamma > 0$, $\phi_i \geq 0$ for all i. Note that $\mathrm{E}[X_t|X_{t-1}, X_{t-2}, \ldots] = 0$ and $\mathrm{var}[X_t|X_{t-1}, X_{t-2}, \ldots] = V_t$.

Once the parametric form of the model is defined, we could then proceed with general statistical inference, including model identification, estimation of parameters and testing of hypotheses (e.g. testing for linearity), and so on. The method of conditional least squares seems to play a particularly important role in many of the studies. By this we mean the minimization, with respect to the unknown parameters, of $\sum\{X_t - \mathrm{E}[X_t|X_{t-1}, \ldots, X_{t-k}]\}^2$, where the summation is over the data set and k is an appropriately chosen order. (See, for example, [19] for details.) We return to the issue of an appropriately chosen order later. All

in all, although a significant number of asymptotic results are now available, there are still many outstanding problems, as we can see from the above references.

Nonparametric

Putting the drift and the diffusion together, we may write down a more general model in the form

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + \varepsilon_t g(X_{t-1}, \ldots, X_{t-q}), \tag{5}$$

where $\{\varepsilon_t\}$ is the usual iid. This is also called a nonlinear autoregressive model. If we assume that $g(\cdot) \equiv 1$, we have the so-called nonlinear autoregressive model of order p with homogeneous noise. If we assume that f is piecewise linear and g is piecewise constant, then we recover the self-exciting threshold autoregressive model. If we assume that f is identically zero and g is $\sqrt{V_t}$ as in (4), then we recover an ARCH model.

In the nonparametric approach, we do not assume any particular parametric form for f or g. Instead, we typically make only minimal assumptions, such as differentiability. The general idea is that "we let the data tell us what f and g are".

There are numerous ways to estimate f and g and we refer to [18] and [20] for some references. To give some general flavor, we describe briefly the locally polynomial approach, which is one of the currently favored approaches. (For details, see, for example, [8].) Let $p(\cdot)$ denote a probability density function, called the kernel. Consider the following local sum of squares:

$$S(\alpha) = \sum (X_t - \alpha)^2 p\left(\frac{X_{t-1} - x_1}{h}\right), \tag{6}$$

where the summation is taken over $t = 2, \ldots, n$, n being the sample size of the X data and h is a positive real constant, usually called the bandwidth of the kernel $p(\cdot)$. Here, α is treated as an unknown parameter as in the classical regression setup. The major significance of the above sum is its local characteristic: observations far from x_1 are given less weight. Minimizing $S(\alpha)$ with respect to α yields the estimate

$$\hat{\alpha} = \frac{\sum X_t p\left(\frac{X_{t-1} - x_1}{h}\right)}{\sum p\left(\frac{X_{t-1} - x_1}{h}\right)}, \tag{7}$$

which coincides with the Nadaraya–Watson estimate of $\mathrm{E}[X_t | X_{t-1} = x_1]$, i.e. $f(x_1)$. Here, we call it the locally constant estimate of $f(x_1)$. We can generalize it to yield a locally polynomial estimate in an obvious way by replacing α in $S(\alpha)$ by $\sum_{j=0}^{q} \alpha_j (X_{t-1} - x_1)^j$, where q denotes the degree of the local polynomial. We can further generalize to obtain similar estimates of $\mathrm{E}[X_t | X_{t-1} = x_1, \ldots, X_{t-k} = x_k]$, i.e. $f(x_1, \ldots, x_k)$, for $k > 1$. For example, the locally constant estimate of $f(x_1, \ldots, x_k)$ may take the form

$$\hat{f}(x_1, \ldots, x_k) = \frac{\sum_{t=k+1}^{n} X_t p\left(\frac{X_{t-1} - x_1}{h}\right) \ldots p\left(\frac{X_{t-k} - x_k}{h}\right)}{\sum_{t=k+1}^{n} p\left(\frac{X_{t-1} - x_1}{h}\right) \ldots p\left(\frac{X_{t-k} - x_k}{h}\right)}. \tag{8}$$

Here, we have effectively taken the associated k-dimensional kernel $p(u_1, \ldots, u_k)$ as a product of k one-dimensional kernels $p(u_j)$, $j = 1, \ldots, k$. (We have abused the notation of $p(\cdot)$.)

With the above nonparametric tools, many important issues can be addressed. For example, in practice, we often have to determine an appropriate order for a nonlinear autoregressive model even though we do not wish to commit ourselves to any particular parametric form of the autoregression. Let us denote the true order by k_0. To describe one nonparametric solution, let us omit the entry $t = j$ in (8). We have the so-called leave-one-out estimate of $f(x_1, \ldots, x_k)$, say $f_j(x_1, \ldots, x_k)$. Let L denote a fixed positive integer, which is much smaller than n. Minimizing

$$\sum_{j=L+1}^{n} [X_j - \hat{f}_j(X_{j-1}, \ldots, X_{j-k})]^2 \tag{9}$$

with respect to $k = 1, \ldots, L$ gives us an estimate $\hat{k}_{\mathrm{CV}}$ of k_0. It turns out that $\hat{k}_{\mathrm{CV}}$ is a consistent estimate of k_0. Moreover, the sample size requirement

for a "good" performance of $\hat{k}_{\mathrm{CV}}$ is not excessive; see [20].)

Other important issues that have been addressed in the context of nonparametrics include tests for independence, linearity, determinism, initial-value sensitivity, nonlinear forecasting, and others.

Stationarity and Prediction

The problem of obtaining sufficient and necessary conditions for strict stationarity for a nonlinear time series model is much deeper than its linear counterpart. The latter problem has a complete solution. This is far from being the case for the former.

If the model can be recast in the form of a **Markov chain** on a Euclidean space (of suitable dimension), then we can appeal to the extensive literature concerning the ergodicity of such Markov chains. (See, for example, [14].) Chan & Tong [4] were probably the first to point out that there is an intimate connection between strict stationarity and the notion of stability in deterministic difference equations. Specifically, let us consider the following nonlinear difference equation

$$z_t = f(z_{t-1}, \ldots, z_{t-k}). \tag{10}$$

We may think of this as the skeleton of the nonlinear autoregressive model; namely

$$X_t = f(X_{t-1}, \ldots, X_{t-k}) + \varepsilon_t. \tag{11}$$

Now, if $|z_t| \to \infty$ as $t \to \infty$, then we say that the system (10) is unstable. Otherwise, we say that it is stable. In fact, under suitable conditions on f, z_t may converge to a finite limit, say z^*, regardless of the initial value z_0. Chan & Tong [4] have proved that, under mild conditions on the distribution of ε, if the z system has the above properties, then the corresponding nonlinear autoregressive model (11) defines an ergodic Markov chain and hence a strictly stationary time series $\{X_0, X_1, \ldots\}$ if X_0 is endowed with the unique invariant distribution of the ergodic Markov chain.

For a strictly stationary time series $\{X_t\}$ with finite variance, the conditional expectation of X_{t+m}, given observations up to and including t, is the m-step-ahead least square predictor of X_{t+m}. Let us denote this by $\hat{X}_t(m)$. It need not be linear in $X_t, X_{t-1}, \ldots$. If $\{X_t\}$ is, in fact, given by a nonlinear autoregressive model of order k, then $\hat{X}_t(m)$ is a nonlinear function of $X_t, X_{t-1}, \ldots, X_{t-k+1}$ only. Unlike the case with linear Gaussian time series models, (i) $\mathrm{var}(X_{t+m}|X_t, \ldots, X_{t-k+1})$ is a function of $X_t, \ldots, X_{t-k+1}$ and need not be monotonically increasing with respect to m; (ii) if it exists, the conditional probability density function of X_{t+m}, given $X_t, \ldots, X_{t-k+1}$, need not be unimodal.

References

[1] Bickel, P.J. & Bühlmann, P. (1996). *What is a Linear Process?* Technical Report, University of California, Berkeley.

[2] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis*. Holden Day, San Francisco.

[3] Brockwell, P.J. & Davis, R.A. (1993). *Time Series: Theory and Methods*. Springer-Verlag, Heidelberg.

[4] Chan, K.S. & Tong, H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability* **17**, 666–678.

[5] Chatfield, C. (1996). *The Analysis of Time Series: An Introduction*. Chapman & Hall, London.

[6] Cutler, C.D. & Kaplan, D.T. (1997). *Nonlinear Dynamics and Time Series*. Fields Institute Communications, Toronto.

[7] Engle, R.F. & McFadden, A. (1993). *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam.

[8] Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

[9] Granger, C.W.J. & Terasvïrta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford.

[10] Grenfell, B.T., Kleczkowski, A., Ellner, S. & Bolker, B.M. (1995). Nonlinear forecasting and chaos in ecology and epidemiology: measles as a case study, in *Chaos and Forecasting*, H. Tong, ed. World Scientific, Singapore, pp. 321–345.

[11] Hannan, E.J. (1970). *Multiple Time Series*. Wiley, New York.

[12] Hao, B-L. (1990). *Chaos II*. World Scientific, Singapore.

[13] Holden, A.V. (1986). *Chaos*. Manchester University Press, Manchester.

[14] Meyn, S.P. & Tweedie, R.L. (1994). *Markov Chains and Stochastic Stability*. Springer–Verlag, Heidelberg.

[15] Nicholls, D.F. & Quinn, B.G. (1982). *Random Coefficient Autoregressive Models: An Introduction*. Springer–Verlag, Heidelberg.

[16] Priestley, M.B. (1988). *Non-linear and Non-stationary Time Series Analysis*. Academic Press, London.

[17] Rao, T.S. & Gabr, M.M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Springer–Verlag, Heidelberg.

[18] Tjøstheim, D. (1994). Non-linear time series: a selective review, *Scandinavian Journal of Statistics* **21**, 97–130.

[19] Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.

[20] Tong, H. (1995). A personal overview of non-linear time series from a chaos perspective (with discussion and comments), *Scandinavian Journal of Statistics* **22**, 399–445.

[21] Tong, H. (1995). *Chaos and Forecasting*. World Scientific, Singapore.

[22] Watier, L. & Richardson, S. (1995). Modelling of an epidemiological time series by a threshold autoregressive model, *Statistician* **44**, 353–364.

HOWELL TONG

Nonparametric Maximum Likelihood

Consider estimation of the survival function of a univariate or multivariate time variable T of interest, based on observing n subjects from a particular population (*see* **Survival Distributions and Their Characteristics**). In an early phase of understanding the survival function of T, or equivalently the **hazard** of T, one typically has only limited knowledge about the shape of these quantities. For example, it might be reasonable to assume that the hazard of T is increasing, but all other assumptions are hard to justify. In such situations there is a need for a nonparametric or semiparametric estimation method of the distribution of T (*see* **Semiparametric Regression**). Nonparametric maximum likelihood estimation (NPMLE) has received a lot of attention over the last few decades.

For simplicity, we will assume that the n observations on the subjects are independent and identically distributed (iid). Let $\mathbf{x} = x_1, \ldots, x_n$ represent n iid observations on a random variable X with probability distribution P_{θ_0} indexed by a parameter θ_0 which is known to be an element of a set Θ. The set of possible probability distributions $\{P_\theta : \theta \in \Theta\}$ is called a *model* for the distribution of X. If $\Theta \subset \mathbb{R}^k$, then the model is called *parametric*, and if Θ is infinite dimensional, then the model is often referred to as being *semiparametric*. For semiparametric models the structure $\theta \to P_\theta$ and/or the structure of the set Θ usually imply that the model is still strictly smaller than the set of all possible probability distributions, which explains the name *semiparametric model*. In statistics, an important goal is to find an estimator of θ_0 based on the data $\mathbf{x}$.

A celebrated estimation method is to define a **likelihood** of the data $\mathbf{x}$ as a function of θ and maximize this likelihood over all θ in Θ (*see* **Maximum Likelihood**). This method makes sense in the classical parametric models where there is only one sensible candidate for the likelihood, but in semiparametric models containing both continuous and discrete probability distributions one typically needs a generalization of maximum likelihood estimation. In this article, we will define maximum likelihood estimation, the generalization, provide illustrations, and elaborate on potential problems of maximum likelihood estimation in semiparametric models.

To start with we will assume that the model is dominated by a single measure μ so that standard maximum likelihood estimation (MLE) can be defined. This allows one to identify each probability distribution P_θ with a density p_θ with respect to (wrt) μ. This density p_θ is defined by the property that the probability that $\mathbf{X}$ falls in a set A (under P_θ) is computed by integrating p_θ over the set A wrt μ. If the P_θs are continuous probability distributions on $\mathbb{R}^k$, μ being the Lebesgue measure, then one can choose p_θ to be the classical derivative of the cumulative distribution function corresponding with P_θ (e.g. the normal density). And if the P_θs are discrete probability distributions on a finite set of outcomes, μ being the counting measure on this set, then $p_\theta(w)$ equals the probability that $\mathbf{X} = w$ (e.g. the **binomial distribution**).

The *likelihood* at θ of a given data vector $\mathbf{x}$ is defined as the density of $\mathbf{X}$ evaluated at $\mathbf{x}$ and is thus given by:

$$L_\mu(\theta|\mathbf{x}) = \prod_{i=1}^{n} p_\theta(x_i),$$

where we indexed the likelihood by μ to stress that the likelihood depends on the choice of dominating measure μ. Assume now that for this given data vector $\mathbf{x}$ there exists a $\hat{\theta}$ for which the likelihood $L(\hat{\theta}|\mathbf{x})$ is larger than or equal to the likelihood $L(\theta|\mathbf{x})$ for any $\theta \in \Theta$. Then $\hat{\theta}$ is called the *maximum likelihood estimator* of θ_0 for the model $\{P_\theta : \theta \in \Theta\}$.

Under "regularity" conditions, a maximum likelihood estimator $\hat{\theta}$ of θ is root-n **consistent**, asymptotically normally distributed and **efficient** (see [1]). We consider a classical textbook example of a parametric maximum likelihood estimator.

Example 1

Let $T_1, \ldots, T_n$ be n iid observations of a survival time T of interest which is known to be **exponentially** distributed with parameter $\lambda > 0$; the density of T is given by $f_\lambda(t) = \lambda \exp(-\lambda t)$ and its hazard equals λ and is thus constant. Thus, the likelihood of an observed data vector $\mathbf{t}$ of $\mathbf{T} = (T_1, \ldots, T_n)$ is given by

$$L(\lambda|\mathbf{t}) = \prod_{i=1}^{n} f_\lambda(t_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} t_i\right).$$

In iid models it is more convenient to maximize the log likelihood, which is here given by

$$\log(L(\lambda|\mathbf{t})) = n\log(\lambda) - \lambda\sum_{i=1}^{n} t_i.$$

Since $\lambda = 0$ is not a maximizer, the maximizer falls in the interior of the parameter space. Thus, the maximum likelihood estimator λ_n solves the score equation:

$$0 = \frac{\mathrm{d}}{\mathrm{d}\lambda}\log[L(\lambda|\mathbf{t})]\bigg|_{\lambda_n} = \frac{n}{\lambda_n} - \sum_{i=1}^{n} t_i.$$

Thus, $\lambda_n = 1/\bar{t}$, $\bar{t}$ being the sample mean of the t_is.

Note that in the definition of a maximum likelihood estimator it is essential that the likelihoods at different elements of the model are comparable. A semiparametric model typically contains both discrete and continuous probability distributions. In this case, the model is often not dominated by a single probability distribution so that we need a generalization of the definition given above. For every pair θ_1, θ_2 one can define the density (likelihood) of P_{θ_1} and of P_{θ_2} wrt $\mu_{\theta_1,\theta_2} \equiv P_{\theta_1} + P_{\theta_2}$, even when P_{θ_1} is discrete and P_{θ_2} is continuous. Now, we will call $\hat{\theta}$ a maximum likelihood estimator of θ_0 if

$$L_{\mu_{\hat{\theta},\theta}}(\hat{\theta}|\mathbf{x}) \geq L_{\mu_{\hat{\theta},\theta}}(\theta|\mathbf{x}), \quad \text{for all } \theta \in \Theta.$$

This definition is due to Kiefer & Wolfowitz [9]. It is easily verified that if the model is dominated by a single measure μ, then the two definitions agree with each other. In the case that the model is semiparametric, the maximum likelihood estimator is called the *nonparametric maximum likelihood estimator*. This name can be misleading since it does *not* mean that no assumptions on the distribution of X have been imposed.

It is typically possible to identify the nonparametric maximum likelihood estimator with a finite dimensional vector so that it can be defined as the maximizer of a real valued functional on a euclidean space. This often allows one to consider the MLE as a vector solution of a set of estimating (score) equations (*see* **Likelihood**). In some semiparametric models these estimating equations can be explicitly solved. In general, numerical iterative subroutines for determining such a maximum or solution of a multivariate estimating equation are available. In some applications specific algorithms might arise naturally.

There exist several examples in the literature where one can explicitly solve for the nonparametric maximum likelihood estimator by deriving a set of score equations. The following example shows how one can obtain score equations via differentiation along one-dimensional submodels (see [5]).

Example 2

Let $T_1, \ldots, T_n$ be n iid copies of a univariate survival time T with cumulative distribution function F, where F is completely unspecified. Let F_n be the empirical cumulative distribution which jumps $1/n$ at each observation. By definition, we have that $\hat{F}$ is a nonparametric maximum likelihood estimator of F if for any distribution F_1:

$$\int \log\left[\frac{\mathrm{d}\hat{F}}{\mathrm{d}(\hat{F}+F_1)}\right]\mathrm{d}F_n \geq \int \log\left[\frac{\mathrm{d}F_1}{\mathrm{d}(\hat{F}+F_1)}\right]\mathrm{d}F_n. \tag{1}$$

Suppose $\hat{F}$ exists. Let μ_n be a dominating measure of $\hat{F}$. Then $\hat{f} \equiv \mathrm{d}\hat{F}/\mathrm{d}\mu_n$ maximizes

$$f \to \int \log(f)\,\mathrm{d}F_n \text{ over all } f = \mathrm{d}F/\mathrm{d}\mu_n, \quad \text{with } F \text{ dominated by } \mu_n. \tag{2}$$

In parametric models we could derive empirical score equations for the maximum likelihood estimator by differentiating the loglikelihood wrt the parameters that indexed the density of T. Differentiation wrt such a one-dimensional parameter means that one moves from p_{θ_n} to some local alternative in the particular direction implied by this parameter. Therefore, in order to imitate this procedure of obtaining score equations in our nonparametric case it is natural to define one-dimensional submodels $\varepsilon \to [1 + \varepsilon h(t)]\hat{f}(t)$ which go through $\hat{f}(t)$ at $\varepsilon = 0$ and where h represents a direction in which we move from $\hat{f}$ to an alternative. This one-dimensional submodel should consist of densities and therefore we need to enforce $\hat{F}h \equiv \int h(t)\,\mathrm{d}\hat{F}(t) = 0$. Thus, we can represent h with $h'(t) - \hat{F}h'$ for any uniformly bounded function h'. We will delete the prime ($'$) and just work with directions $h - \hat{F}h$, h being any uniformly bounded function.

Because $\hat{f}$ maximizes (2) we have that

$$\phi_h(\varepsilon) \equiv \int \log\{\hat{f}_{\varepsilon,h}(t)\}\,\mathrm{d}F_n(t)$$

is maximized at $\varepsilon = 0$, where $\hat{f}_{\varepsilon,h}(t) \equiv \{1 + \varepsilon[h(t) - \hat{F}h]\}\hat{f}(t)$. Consequently, the derivative of the one-dimensional log likelihood $\phi_h(\varepsilon)$ at ε equals zero. Thus

$$\int [h(t) - \hat{F}h]\,\mathrm{d}F_n(t) = 0 \text{ for all uniformly bounded } h,$$

which implies that $\hat{F}h = \int h(t)\,\mathrm{d}F_n(t)$ for all uniformly bounded h. This proves that $\mathrm{d}\hat{F} = \mathrm{d}F_n$. In other words, the maximum likelihood estimator in the completely nonparametric model is given by the empirical cumulative distribution function F_n.

Alternatively, one can view the likelihood as a function of a finite-dimensional vector of parameters and obtain score equations by differentiating wrt these parameters. In the following success story of nonparametric maximum likelihood estimation we present both approaches for obtaining score equations, each of them leading to different insights in the maximum likelihood estimator.

Example 3: Univariate Right-censoring Model

In survival applications T is often either observed or right-**censored** by a censoring variable C. Let the distributions F and G of T and C be completely unspecified and we will assume that C is independent of T. A right-censored observation on T can be represented as

$$Y = [\tilde{T} = T \wedge C, \Delta = I(T \leq C)].$$

In other words, we observe the minimum of failure and censoring and the failure indicator. The density of Y can be represented as

$$p_{F,G}(\tilde{t}, \delta) = \{\mathrm{d}F(\tilde{t})\overline{G}(\tilde{t})\}^{\delta} + \{S(\tilde{t})\,\mathrm{d}G(\tilde{t})\}^{1-\delta}.$$

So the log likelihood is given by:

$$\sum_{i=1}^{n} \log[\mathrm{d}F(\tilde{T}_i)]\Delta_i + \log[S(\tilde{T}_i)](1 - \Delta_i) + \log[\overline{G}(\tilde{T}_i)]\Delta_i + \log[\mathrm{d}G(\tilde{T}_i)](1 - \Delta_i).$$

We are concerned with estimating F. For maximum likelihood estimation we concentrate on maximizing the relevant part of the likelihood wrt F given by

$$\sum_{i=1}^{n} \log[\mathrm{d}F(\tilde{T}_i)]I(\Delta_i = 1) + \log[S(\tilde{T}_i)]I(\Delta_i = 0).$$

One can now obtain a set of score equations by differentiating the log likelihood wrt a class of one-dimensional submodels $\mathrm{d}F_\varepsilon(x) = (1 + \varepsilon h(x))\,\mathrm{d}F_n(x)$, $\int h\,\mathrm{d}F_n = 0$, through the nonparametric maximum likelihood estimator F_n, as in the preceding example. This shows that F_n solves the self-consistency equations (see [4] and [5]):

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} E_{F_n}[I(T_i \leq t)|\tilde{T}_i, \Delta_i],$$

which provides an important heuristic interpretation to the estimator, explained in detail in the next example.

The following alternative way of obtaining score equations provides a closed-form solution F_n. First, note that F_n is discrete on the uncensored observations $T_1, \ldots, T_m$. Then we can reparameterize the log likelihood in terms of the jumps $\mathrm{d}\Lambda$ of the hazard at the uncensored observations by substitution of $\mathrm{d}F(\tilde{T}_i) = \mathrm{d}\Lambda(\tilde{T}_i)S(\tilde{T}_i-)$ and $S(\tilde{T}_i) = \prod T_j \leq \tilde{T}_i[1 - \mathrm{d}\Lambda(T_j)]$, where for a $\tilde{T}_i$ with $\Delta_i = 0$ we have $S(\tilde{T}_i) = S(\tilde{T}_i-)$. This shows that the log likelihood can be expressed as

$$\sum_{i=1}^{n} \left\{ \Delta_i \log[\mathrm{d}\Lambda(T_i)] + \sum_{T_j < \tilde{T}_i} \log[1 - \mathrm{d}\Lambda(T_j)] \right\}.$$

Consider this as a function of $(\lambda_1, \ldots, \lambda_m) \equiv [\mathrm{d}\Lambda(T_1), \ldots, \mathrm{d}\Lambda(T_m)]$. Let $\boldsymbol{\lambda}_n = (\lambda_{1n}, \ldots, \lambda_{mn})$ be the maximizer of the log likelihood. Let k_j be the number of uncensored observations at T_j, $j = 1, \ldots, m$, so with continuous data $k_j = 1$. Differentiation of the log likelihood wrt λ_1 at $\boldsymbol{\lambda}_n$ yields

$$0 = \frac{k_1}{\lambda_{1n}} - \frac{1}{1 - \lambda_{1n}}\sum_{i=1}^{n} I(\tilde{T}_i > T_1).$$

Simple algebra yields:

$$\lambda_{1n} = \frac{k_1}{k_1 + \sum_{i=1}^{n} I(\tilde{T}_i > T_1)}.$$

Similarly, differentiation wrt λ_j yields

$$\lambda_{jn} = \frac{k_j}{k_j + \sum_{i=1}^{n} I(\tilde{T}_i > T_j)}, \quad j = 1, \ldots, m.$$

In other words, the jump of Λ_n at T_j is given by the number of uncensored observations at T_j divided by the number of subjects at risk just before T_j. The corresponding estimator F_n of F is the well-known **Kaplan–Meier estimator**.

If the supremum of the likelihood over the model is attained by an element at the boundary of the model, where the boundary is not part of the model, then the maximum likelihood estimator will not exist. For example, if the model for a real-valued random variable X consists of all continuous densities, then the supremum over the likelihood of n iid observations is attained by the empirical probability distribution, which is not an element of the model (see, for example, [12]). However, if we restrict in this iid setting the model to all monotone decreasing (increasing) densities, then the maximum likelihood estimator appears to be a histogram density estimator with variable bandwidth [3, 7]. The latter estimator can be computed as the left derivative of the convex minorant of the empirical cumulative distribution function, which yields an extremely fast algorithm for computing this beautiful maximum likelihood estimator. It should be noted that in these problems where the support of the maximum likelihood estimator is unknown as well, score equations as derived in the preceding example do not uniquely characterize the maximum likelihood estimator (see [8]).

The monotone density model and right-censored data model provide examples for which the maximum likelihood estimator is very sensible. Roughly speaking, it can be shown that if the nonparametric maximum likelihood estimator of a parameter of P_θ is root-n consistent, then it will also be asymptotically efficient (see, for example, [6]). In other words, if it works, then it results in asymptotically optimal estimators of root-n estimable parameters.

The fact that the nonparametric maximum likelihood estimator tries to be efficient at every element in the model is one of the main reasons why it performs poorly in truly high-dimensional models (e.g. involving several covariates). In [10] and [11] it is reasoned that for semiparametric estimation with high-dimensional data structures one should sacrifice the global efficiency and instead search for estimators which are consistent at every element of the model *and* are efficient at a specified subset of the model. By now, their presented method has been used and applied in several high-dimensional censored data models.

This article concludes with an example where the NPMLE is inconsistent, but can be repaired successfully by slightly transforming the data.

Example 4: Bivariate Right censoring

Suppose that we are concerned with estimation of the bivariate lifetime distribution S_0 of a particular population of twins (*see* **Twin Analysis**). Let $T = (T_1, T_2)$ be the corresponding bivariate survival time (*see* **Multivariate Survival Analysis**) of a randomly drawn twin from this population and assume that each twin is subject to right-random censoring by an irrelevant censoring vector $C = (C_1, C_2)$. The iid observations on n twins are now

$$Y_i \equiv (\tilde{T}_i, \Delta_i) \equiv [T_i \wedge C_i, \quad I(T_i \leq C_i)],$$

with components given by:

$$\tilde{T}_{ij} = \min\{T_{ij}, C_{ij}\},$$
$$\Delta_{ij} = I(T_{ij} \leq C_{ij}), \quad j = 1, 2.$$

In other words, for twin 1 we observe the minimum of censoring and survival and we observe if this minimum is the actual survival time of interest, and similarly for twin 2. Each bivariate randomly right-censored observation Y_i tells us that $T_i = (T_{1i}, T_{2i})$ has fallen in a region $B(Y_i)$ in the plane where this region is a dot if both T_{1i} and T_{2i} are observed (uncensored), it is a half-line if only one of the survival times is right censored (singly censored), and it is a right-upper quadrant if both T_{1i} and T_{2i} are right censored (doubly censored). Therefore, the data can be nicely presented as in Figure 1 (disregard the strips, at this stage). If we had observed all $T_i, i = 1, \ldots, n$, then the NPMLE $S_n(t_1, t_2)$ of $S(t_1, t_2)$ would equal the fraction of the T_i which is larger than (t_1, t_2). In other words, we would give each observation T_i weight $1/n$ and sum up the weights of the T_is with $T_i > (t_1, t_2)$.

In general, we can still give all uncensored observations weight $1/n$. An observation Y_i only tells us that $T_i \in B(Y_i)$; so we want to give the mass $1/n$ to $B(Y_i)$ in an appropriate way. Assume that by using the observations in $B(Y_i)$ we are able to obtain a good estimator $P_{F_n^0}(T \in \cdot | T \in B(Y_i))$ of the conditional distribution $P(T \in \cdot | T \in B(Y_i))$ of T_i given that $T_i \in B(Y_i)$. Then a natural thing to do is to redistribute the mass $1/n$ corresponding with the censored

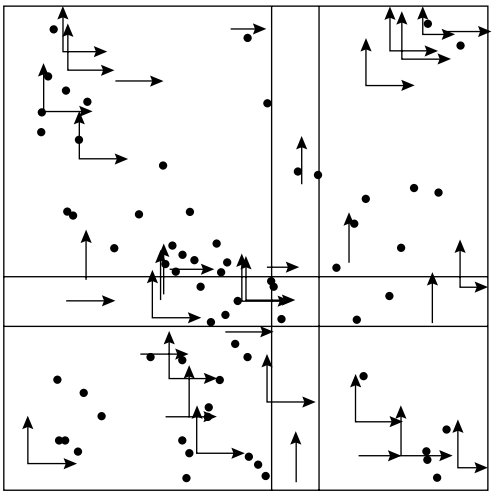

Figure 1 Right-censored bivariate data; • = uncensored, →= censored.

observation Y_i over $B(Y_i)$ as follows (assume for convenience that the estimate is discrete): a point $s > t$ gets the following fraction of the mass $1/n : P_{F_n^0}(T = s|T \in B(Y_i))$. If we do the redistribution for all censored observations, then we obtain a new estimator F_n^1, which might be an improvement.

This suggests the following **algorithm**:

1. Let $\{s_1, \ldots, s_k\}$ be a set of points in the plane which contains all uncensored T_i and it is such that each $B(Y_i)$ (lines and quadrants) contains at least one of these s_is.
2. Give each s_i a weight $f_n^0(s_i) > 0$. $\sum_{i=1}^k f_n^0(s_i) = 1$. Set the count $M = 0$.
3. Compute a new estimator f_n^{M+1}, as follows:

$$f_n^{M+1}(s_i) = \sum_{j=1}^{n} P_{f_n^M}(T = s_i|T \in B(Y_j))\frac{1}{n}, \quad i = 1, \ldots, k. \tag{3}$$

In other words, a point s_i gets from each observation Y_j mass $1/nP_{f_n^M}(T = s_i|T \in B(Y_j))$, which is zero if $s_i \notin B(Y_j)$ and it is 1 if s_i equals the observed T_j.

4. Replace M by $M + 1$ and go to step 3.

This is the **EM algorithm** [2, 14]. f_n^M can be shown to converge to a solution f_n of (3) with $f^M = f^{M+1} = f_n$. Eq. (3) in f_n is the well-known self-consistency equation of Efron [4], which is solved by NPMLE. The latter can be shown as in the univariate right-censoring example by differentiating the log likelihood along one-dimensional submodels.

If an uncensored observation receives mass from a censored observation at step M of the EM algorithm, then this influences the conditional probabilities $P_{f_n^{M+1}}[T = s|T \in B(Y_j)]$ of each region $B(Y_j)$ which contains this uncensored observation. If the underlying F is continuous, then the half-lines $B(Y_j)$ corresponding with singly censored observations will with probability one not contain any uncensored observations. Then the conditional probabilities over lines do not change at a step of the EM algorithm by mass given to the uncensored observations; i.e. a singly censored observation with half-line $B(Y_i)$ does not listen to information given by uncensored observations, but its redistribution over the half-line might change by mass given by doubly censored and other singly censored observations interacting with $B(Y_i)$. Since uncensored observations around a half-line provide a lot of information about the distribution over the half-line, there is no reason to expect good performance of S_n. Indeed, S_n is not consistent for continuous data [13].

Suppose now that the uncensored components of the singly censored observations Y_i are **interval censored** by a lattice partition $A_{k,l} = (u_k, u_{k+1}] \times (v_l, v_{l+1}]$, in the sense that (we do as if) it is only known to lie in $(u_k, u_{k+1}]$ (if T_1 is uncensored) or in $(v_l, v_{l+1}]$ (if T_2 is uncensored) (see Figure 1). In this way we have reduced the original data Y_i by putting an additional slight transformation on top of it, $i = 1, \ldots, n$. The interval-censored singly censored observations tell us that T_i has fallen in a strip around the original singly censored observation, which will contain other uncensored observations, and hence we expect a better result from the EM algorithm. In [15] it is shown that if one reduces the data in this manner, then the NPMLE based on the reduced data is efficient for the reduced data and efficient for the original data if the reduction converges to zero slowly enough when n converges to infinity.

References

[1] Bickel, P.J., Klaassen, A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[2] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[3] Devroye, L. (1987). *A Course in Density Estimation*. Birkhauser, Basel.

[4] Efron, B. (1967). The two sample problem with censored data, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 831–853.

[5] Gill, R.D. (1989). Non- and semiparametric maximum likelihood estimators and the von Mises method (Part I), *Scandinavian Journal of Statistics* **16**, 97–128.

[6] Gill, R.D. & van der Vaart, A.W. (1993). Non- and semiparametric maximum likelihood estimators and the von Mises method, II, *Scandinavian Journal of Statistics* **20**, 271–288.

[7] Grenander, U. (1956). On the theory of mortality measurement. Part II *Skandinavisk Aktuarietidskrift* **39**, 125–153.

[8] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser, Basel.

[9] Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics* **27**, 887–906.

[10] Robins, J.M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, in *American Statistical Association 1993 Proceedings of the Section on Biopharmaceuticals*. American Statistical Association, Alexandria, pp. 24–33.

[11] Robins, J.M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, in *Aids Epidemiology, Methodological Issues*. Birkhauser, Basel.

[12] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[13] Tsai, W.-Y., Leurgans, S. & Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring, *Annals of Statistics* **14**, 1351–1365.

[14] Turnbull, B.W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 290–305.

[15] van der Laan, M.J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE, *Annals of Statistics* **24**, 596–627.

MARK VAN DER LAAN

Nonparametric Methods

Many of the earliest statistical procedures proposed and studied rely on the underlying assumption of distributional normality. How well these procedures operate outside the confines of this normality constraint varies from setting to setting. Although there were a few isolated attempts to create statistical procedures that were valid under less restrictive sets of assumptions that did not include normality, such as the early introduction of the essence of the **sign test** procedure by Arbuthnott [2] in 1710, and the **rank correlation** procedure considered by Spearman [51] in 1904, it is generally agreed that the systematic development of the field of nonparametric statistical **inference** traces its roots to the fundamental papers of Friedman [18], Kendall [31], Kendall & Babington Smith [33], Mann & Whitney [38], and Wilcoxon [58].

The earliest work in nonparametric statistics concentrated heavily on the development of **hypothesis testing** that would be valid over large classes of probability distributions – usually the entire class of continuous distributions, but sometimes with the additional assumption of distributional symmetry. Most of this early work was intuitive by nature and based on the principle of ranking (*see* **Ranks**) to de-emphasize the effect of any possible **outliers** on the conclusions. Point and interval **estimation** expanded out of this hypothesis testing framework as a direct result of centering of the test statistics and test inversion, respectively (*see* **Estimation, Interval**).

Most distribution-free test procedures (and associated confidence intervals) are based on one or more of the following three fundamental properties.

Result 1. Let $Z_1, \ldots, Z_n$ be a **random sample** from some probability distribution and let A be a subset of the common domain for the Zs. If $I(t)$ represents the indicator function for this subset A, then the random variable $V = \sum_{i=1}^{n} I(Z_i)$ has a **binomial distribution** with parameters n and $p = \Pr(Z_i \in A)$.

Result 2. Let $Z_1, \ldots, Z_n$ be a random sample from a continuous distribution with cumulative distribution function (cdf) $F(\cdot)$, and let R_i denote the rank (from least to greatest) of Z_i among the n Zs, for $i = 1, \ldots, n$. Then the vector of ranks $\mathbf{R} = (R_1, \ldots, R_n)$ has a joint distribution that is **uniform** over the set of all permutations of the integers $(1, \ldots, n)$.

Result 3. Let Z be a **random variable** with a probability distribution that is symmetric about the point θ. Define the indicator function $\Psi(\cdot)$ by

$$\Psi(t) = 1, \quad \text{if } t > 0,$$
$$= 0, \quad \text{if } t \leq 0.$$

Then the random variables $|Z - \theta|$ and $\Psi(Z - \theta)$ are independent.

Statistics based solely on Result 1 are referred to as counting statistics, those based solely on Result 2 are commonly known as ranking statistics, and those based on an appropriate combination of all three results are called **signed-rank statistics**. Over the years of development in the field, distribution-free procedures have certainly become more sophisticated, both in the problems they address and in their complexity. However, the underlying premise behind almost all such hypothesis tests continues to rest with these three basic results or with modifications thereof.

Much of the early work in distribution-free hypothesis tests followed the general approach of mimicking a standard normal theory procedure for a statistical problem by replacing the sample values with some combination of rank or counting statistics. The first nonparametric test statistics looked quite similar in form to their classical normal theory counterparts. However, more recent advances in nonparametric statistics have been less tied to previously developed normal theory structure and, in fact, there have been a number of settings where nonparametric procedures were the first to be developed, and classical procedures followed a few years later.

It is the intent of this article to provide some brief overview of nonparametric statistics. However, the field has grown over the years to such a size that one must rely on standard textbooks in the area for a truly complete picture. The very first such textbooks in nonparametric statistics were the pioneering works of Siegel [49] and Fraser [17], both arriving on the scene in the infancy of the field. Walsh [55–57] published a three-volume handbook covering those nonparametric procedures available at the time. Other texts and reference books have added to the literature

of nonparametric statistics over the years, including the applications-oriented books by Bradley [3], Conover [5], Daniel [10], Gibbons [21], Hollander & Wolfe [27], and Marascuilo & McSweeney [39]. The text by Lehmann [36] occupies an intermediate place in the literature. It has a general application orientation, but a considerable amount of the basic underlying theory of some of the procedures is also presented in a substantial appendix. Textbooks dealing primarily with the theory of rank tests and associated point estimators and confidence intervals have been published by Gibbons [20], Hájek [22], Hájek & Šidák [23], Hettmansperger [25], Noether [44], Pratt & Gibbons [46], and Randles & Wolfe [47]. The monograph by Kendall [32] covers the specialized topic of rank correlation methods. These resources vary on the extensiveness of their bibliographies, but it is safe to say that the vast majority of published literature in the field of nonparametric statistics is cited in at least one of these volumes.

One of the necessities in the application of distribution-free test procedures and confidence intervals is the availability of the exact null distributions of the associated test statistics (*see* **Null Hypothesis**). Extensive tables of many of these null distributions are available in some of the applications-oriented texts mentioned previously. In addition, recent software developments have made it a good deal easier both to compute the appropriate test statistics and to obtain the associated ***P* values** for many of these test procedures (*see* **Software, Biostatistical**). Of particular note in this regard are the Minitab and **StatXact** software packages, for both their rather complete coverage of the basic nonparametric procedures and their ability to circumvent the need for exact null distribution tables by providing the associated exact or approximate P values for many of the test procedures (*see* **Exact Inference for Categorical Data**). StatXact also has the option of actually generating the required exact null distributions for some of the better known test statistics, including the appropriate modifications necessary in the case of tied observations.

We first turn our attention to brief descriptions of the most commonly used nonparametric procedures in standard statistical settings involving one, two, or more samples, including one- and two-way **analysis of variance** and **correlation**. In each case, the emphasis will be on the description of the problem and a particular standard approach to its solution, rather than on attempting to cover the myriad of different nonparametric procedures that are commonly available for the problem.

Finally, we will discuss briefly a few nonstandard topics where the development of nonparametric methods has been particularly motivated by the need to analyze medical and health sciences data. Included in these topics will be **censored data** and **survival analysis**, as well as **proportional hazards models**, **counting processes**, and **bootstrap methods**.

One-sample Location Problem

Continuity Assumption Only

Let $Z_1, \ldots, Z_n$ be a random sample arising from an underlying probability distribution that is continuous with cdf $F(\cdot)$ and median θ. Here the primary interest is in inference about θ.

Test Procedure. For this setting, we are interested in testing the null hypothesis that $\theta = \theta_0$, where θ_0 is some preset value appropriate for the problem. If no additional assumptions are reasonable about the form of the underlying F, the most commonly used inference procedures are those associated with the sign statistic

$$B = [\text{number of sample } Z\text{s that exceed } \theta_0]$$

(*see* **Sign Tests**). The properties of B follow from the basic counting Result 1 with the set $A = (\theta_0, \infty)$. In particular, B has a binomial distribution with number of trials n and success probability $p = \Pr(Z_1 > \theta_0)$. When the null hypothesis is true, we have $p = 1/2$ (since θ_0 is then the **median** of the underlying distribution) and the null distribution of B does not depend on the form of F. The associated level α sign procedure for testing H_0, vs. the alternative $H_1 : \theta > \theta_0$, is to reject H_0 if the observed value of B exceeds b_α, the upper αth percentile for the null distribution of B, namely, the binomial distribution with parameters n and $p = 1/2$ (*see* **Level of a Test**). The appropriate tests for the other directional alternatives $\theta < \theta_0$ and $\theta \neq \theta_0$ rely on the fact that the binomial distribution with n trials and $p = 1/2$ is symmetric about its mean $n/2$.

Point Estimation and Confidence Intervals/ Bounds. Natural nonparametric **confidence intervals** and confidence bounds for θ are associated with these sign test procedures through the common process of inverting the appropriate hypothesis tests. These intervals and bounds are based on the ordered sample observations $Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(n)}$. The $100(1-\alpha)\%$ confidence interval for θ associated in this manner with the level α two-sided sign test is given by $(Z_{(n+1-b_{\alpha/2})}, Z_{(b_{\alpha/2})})$, where $b_{\alpha/2}$ is again the upper $(\alpha/2)$th percentile for the binomial distribution with parameters n and $p = 1/2$. The corresponding $100(1-\alpha)\%$ lower and upper confidence bounds for θ (obtained by inverting the appropriate one-sided sign tests) are given by $Z_{(n+1-b_\alpha)}$ and $Z_{(b_\alpha)}$, respectively. The Hodges–Lehmann [26] point estimator of θ associated with the sign test is $\tilde{\theta}$ = median $\{Z_1, \ldots, Z_n\}$.

Continuity and Symmetry Assumption

Let $Z_1, \ldots, Z_n$ be a random sample from an underlying probability distribution that is continuous and symmetric about its median θ. Once again the primary interest is in inference about θ.

Test Procedure. We remain interested in testing the null hypothesis that $\theta = \theta_0$. However, the additional symmetry assumption now enables us to provide generally more powerful test procedures. For this setting, the most commonly used inference procedures are those associated with the **Wilcoxon signed-rank test** statistic [58],

$$T^+ = \sum_{i=1}^{n} R_i \Psi_i,$$

where $\Psi_i = 1, 0$ as $Z_i >, < \theta_0$, and R_i is the rank of $|Z_i - \theta_0|$ among $|Z_1 - \theta_0|, \ldots, |Z_n - \theta_0|$. Thus, the Wilcoxon signed-rank statistic corresponds to the sum of the $|Z - \theta_0|$ ranks for those Zs that exceed the hypothesized median value θ_0. [Since we have a continuous underlying distribution, the probability is zero that there are ties among the absolute values of the $(Z_i - \theta_0)$s. Likewise, the probability is zero that any of the Z_is actually equals θ_0. However, these events may occur in actual data sets. In such an event, it is standard practice to discard the Z_is that equal θ_0 and reduce n accordingly. Ties among the absolute values of the $(Z_i - \theta_0)$s are generally broken by assigning average ranks to each of the absolute differences within a tied group.]

Properties of T^+ under $H_0: \theta = \theta_0$ derive directly from Result 3, which yields the independence of the ranks of the $|Z_i - \theta_0|$s and the Ψ_is, and Result 2, which implies that the ranks of the $|Z_i - \theta_0|$s are uniformly distributed over the set of permutations of the integers $(1, \ldots, n)$ under H_0. The associated null distribution of T^+ does not depend on the form of the underlying $F(\cdot)$ and has been extensively tabled (see, for example, [27] and [59]). The associated level α signed-rank procedure for testing H_0 vs. the alternative $H_1: \theta > \theta_0$ is to reject H_0 if the observed value of T^+ exceeds t_α, the upper αth percentile for the null distribution of T^+. The appropriate tests for the other directional alternatives $\theta < \theta_0$ and $\theta \ne \theta_0$ rely on the fact that the null distribution of T^+ is symmetric about its mean $n(n+1)/4$.

Point Estimation and Confidence Intervals/ Bounds. Once again, natural confidence intervals and confidence bounds for θ are associated with these signed-rank procedures through inversion of the appropriate hypothesis tests. These intervals and bounds are based on the ordered values of the $M = n(n+1)/2$ Walsh averages of the form $W_{ij} = (Z_i + Z_j)/2$, for $1 \le i \le j \le n$. Letting $W_{(1)} \le \cdots \le W_{(M)}$ denote these ordered Walsh averages, the $100(1-\alpha)\%$ confidence interval for θ associated with the level α two-sided signed-rank test is given by $(W_{(M+1-t_{\alpha/2})}, W_{(t_{\alpha/2})})$, where once again $t_{\alpha/2}$ is the upper $(\alpha/2)$th percentile for the null distribution of T^+. The corresponding $100(1-\alpha)\%$ lower and upper confidence bounds for θ (obtained by inverting the appropriate one-sided signed-rank tests) are given by $W_{(M+1-t_\alpha)}$ and $W_{(t_\alpha)}$, respectively. The Hodges–Lehmann [26] point estimator of θ associated with the signed-rank test is $\hat{\theta}$ = median $\{W_{ij}, 1 \le i \le j \le n\}$.

We note that both the sign and signed-rank inference procedures can be applied to paired replicates data (X_i, Y_i), where X_i represents a pretreatment measurement on a subject and Y_i represents a posttreatment measurement on the same subject, and we collect such paired data from $i = 1, \ldots, n$ independent subjects. The appropriate sign or signed-rank procedures are then applied to the post-minus-pre differences $Z_i = Y_i - X_i, i = 1, \ldots, n$.

Two-sample Location Problem

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from the continuous probability distributions with cdfs $F(\cdot)$ and $G(\cdot)$, respectively. We consider here the case where $G(y) = F(y - \Delta)$, with $-\infty < \Delta < \infty$; that is, the X and Y distributions differ only by a possible location shift Δ, and we are interested in inference about Δ.

Test Procedure. For this setting, the appropriate null hypothesis is that $\Delta = \Delta_0$, where Δ_0 is some preset value (often zero) of interest for the shift. The most commonly used nonparametric inference procedures for this setting are those associated with the rank sum version of the **Wilcoxon–Mann–Whitney** [38, 58],

$$W = \sum_{j=1}^{n} R_j,$$

where R_j is the rank of Y_j among the combined sample of $N = (m + n)$ observations $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$. (Once again, ties among the Xs and/or Ys are broken by assigning average ranks to each of the observations within a tied group.)

Properties of W under H_0: $\Delta = 0$ (corresponding to no differences between the X and Y probability distributions) follow directly from the basic ranking Result 2, which implies that the joint ranks of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are uniformly distributed over the set of permutations of the integers $(1, \ldots, N)$ under H_0. The associated null distribution of W does not depend on the form of the common (under H_0) underlying distribution $F(\cdot)$ and has been extensively tabled (see, for example, [27] and [59]). The associated level α rank sum procedure for testing H_0 vs. the alternative H_1: $\Delta > 0$ is to reject H_0 if the observed value of W exceeds w_α, the upper αth percentile for the null distribution of W. The appropriate tests for the other directional alternatives $\Delta < 0$ and $\Delta \neq 0$ rely on the fact that the null distribution of W is symmetric about its mean $n(m + n + 1)/2$.

Point Estimation and Confidence Intervals/ Bounds. As in the one-sample setting, natural confidence intervals and bounds for Δ are associated with these rank sum procedures through inversion of the appropriate hypothesis tests. These intervals and bounds are based on the ordered values of the mn differences $U_{ij} = Y_j - X_i$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Letting $U_{(1)} \leq \cdots \leq U_{(mn)}$ denote these ordered differences, the $100(1 - \alpha)\%$ confidence interval for Δ associated with the level α two-sided rank sum test is given by $U_{(\{[n(2m+n+1)+2]/2\}-w_{\alpha/2})}$, $U_{(w_{\alpha/2}-[n(n+1)/2])}$, where once again $w_{\alpha/2}$ is the upper $(\alpha/2)$th percentile for the null distribution of W. The corresponding $100(1 - \alpha)\%$ lower and upper confidence bounds for Δ (obtained by inverting the appropriate one-sided rank sum tests) are given by $U_{(\{[n(2m+n+1)+2]/2\}-w_\alpha)}$ and $U_{(w_\alpha-[n(n+1)/2])}$, respectively. The Hodges–Lehmann [26] point estimator of Δ associated with the rank sum test is $\hat{\Delta} =$ median $\{U_{ij}, i = 1, \ldots, m, j = 1, \ldots, n\}$.

Other Two-sample Problems

The possibility of differences in location between the X and Y distributions is certainly the most common problem of interest in the two-sample setting. However, there are circumstances where differences in scale are of primary concern, as well as situations where it is important to detect differences of *any* kind between the X and Y distributions. For discussion on nonparametric two-sample procedures designed for scale differences, see **Wilcoxon-type scale tests**. The development of nonparametric procedures designed to be effective against *any* differences between the X and Y distributions was initiated by the pioneering work of Kolmogorov [34] and Smirnov [50]. These papers have inspired a substantial body of research on such omnibus two-sample procedures (*see* **Kolmogorov–Smirnov Test**).

One-way Analysis of Variance: $k \geqslant 3$ Populations

This is a direct extension of the two-sample location problem. The data now represent k mutually independent random samples of observations from continuous probability distributions with cdfs $F_1(x) = F(x - \tau_1)$, $F_2(x) = F(x - \tau_2), \ldots, F_k(x) = F(x - \tau_k)$, where $F(\cdot)$ is the cdf for a continuous population with median θ and $\tau_1, \ldots, \tau_k$ represent the additive effects corresponding to belonging to population $1, \ldots, k$, respectively. Here, our interest is in possible differences in the population effects $\tau_1, \ldots, \tau_k$.

Test Procedures. For the one-way analysis of variance setting, we are interested in testing the null hypothesis H_0: $[\tau_1 = \cdots = \tau_k]$, corresponding to no differences in the medians of the k populations. For this setting, the most commonly used test procedures correspond to appropriate extensions of the Mann–Whitney–Wilcoxon joint ranking scheme as specifically directed toward the particular alternative of interest. For testing the null H_0 vs. the standard class of general alternatives H_1: (not all τ_is equal), the Kruskal–Wallis [35] test is the most popular procedure. For one-sided ordered alternatives of the form H_2: $(\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$, with at least one strict inequality), the appropriate extension is that proposed independently by Jonckheere [28] and Terpstra [54]. Finally, for umbrella alternatives H_3: $(\tau_1 \leq \tau_2 \leq \cdots \leq \tau_{q-1} \leq \tau_q \geq \tau_{q+1} \geq \cdots \geq \tau_k$, with at least one strict inequality), with either the peak of the umbrella, q, known a priori or estimated from the data, the standard test procedures are those proposed by Mack & Wolfe [37].

Multiple Comparisons and Contrast Estimation. After rejection of H_0: $(\tau_1 = \cdots = \tau_k)$ with an appropriate test procedure, one is most often interested in deciding *which* of the populations are different and then in estimating the magnitudes of these differences. This leads to the use of **multiple comparison** procedures, based either on pairwise or joint rankings of the observations. With pairwise rankings, where two-sample ranks are used to compare separately the sample data for each of the $\binom{k}{2}$ pairs of populations, the most commonly used multiple comparison procedures are those considered by Dwass [12], Steel [53], and Critchlow & Fligner [7] for two-sided all-treatment differences, and by Hayter & Stone [24] for one-sided all-treatment differences. The corresponding two-sided all-treatment multiple comparison procedure based on joint rankings, where the sample data from all k populations are ranked jointly, has been studied by Nemenyi [43] and Damico & Wolfe [8], while the joint rankings multiple comparison procedure for one-sided treatments vs. control decisions can be found in [43] and [9]. Point estimation of any **contrasts** in the τs (that is, any linear combination $\beta = \sum_{i=1}^{k} a_i \tau_i$, with $\sum_{i=1}^{k} a_i = 0$) is discussed in Spjøtvoll [52]. Simultaneous two-sided confidence intervals for *all* simple contrasts of the form $\tau_j - \tau_i$ have been developed by Critchlow & Fligner [7], while the corresponding simultaneous one-sided confidence bounds (*see* **Simultaneous Inference**) were studied by Hayter & Stone [24].

Two-way Analysis of Variance

We consider here the standard two-way layout setting, where the data consist of one observation on each combination of k treatments and n blocks (*see* **Randomized Complete Block Designs**). The observation in the ith block and jth treatment combination, denoted by X_{ij}, arises from a continuous probability distribution with cdf $F(x - \beta_i - \tau_j)$, where $F(\cdot)$ is the cdf for a continuous distribution with median θ, for $i = 1, \ldots, n$; $j = 1, \ldots, k$. Moreover, the nk Xs are assumed to be mutually independent random variables. (This is known as the additive two-way layout model.) Here, our interest is in possible differences among the treatment effects $\tau_1, \ldots, \tau_k$.

Test Procedures. For the two-way layout with one observation per cell, we are interested in testing the null hypothesis H_0: $(\tau_1 = \cdots = \tau_k)$, corresponding to no differences in the k treatment effects. For this setting, the most commonly used procedures correspond to appropriate extensions of the sign test procedure for paired replicates data as specifically directed toward a particular alternative of interest. For testing the null H_0 vs. the standard class of general alternatives H_1: (not all τ_is equal), the Friedman [18] test procedure is based on within-blocks ranks of the observations across treatment levels. For ordered alternatives of the form H_2: $(\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$, with at least one strict inequality), the appropriate test based on within-blocks ranks is that given by Page [45].

Multiple Comparisons and Contrast Estimation. After rejection of H_0: $(\tau_1 = \cdots = \tau_k)$ with an appropriate test procedure, one can use either the multiple comparison procedure studied by Nemenyi [43] and McDonald & Thompson [40] to reach the $k(k-1)/2$ all-treatments two-sided decisions of the form $\tau_i = \tau_j$ vs. $\tau_i \neq \tau_j$, or the corresponding treatments vs. control multiple comparison procedure due to Nemenyi [43], Wilcoxon & Wilcox [60], and Miller [41] to reach the $k-1$ treatments vs. control one-sided decisions of the form $\tau_j > \tau_{\text{control}}$. A method for point

estimation of a contrast in the τs can be found in Doksum [11].

Independence

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a continuous bivariate probability distribution (*see* **Bivariate Distributions**). The most common distribution-free tests for the independence of the X and Y variables are those considered by Kendall [31] and Spearman [51] (*see* **Rank Correlation**). The null distribution properties of both of these test procedures are based on the basic Result 2 and the fact that the ranks of the Xs and the separate ranks of the Ys are themselves independent under the independence of X and Y. Approximate $100(1-\alpha)\%$ confidence intervals and bounds for the Kendall correlation coefficient $\gamma = \{2\Pr[(Y_2 - Y_1)(X_2 - X_1) > 0] - 1\}$ have been provided by Noether [44], Fligner & Rust [16], and Samara & Randles [48].

Censored Data

One of the areas where nonparametric methods have played a major role in the analysis of medical and health sciences data in particular has been that of survival analysis of censored lifetime data. We discuss the basic concepts involved in dealing with censored data in the one-sample setting and then provide brief descriptions of the most important nonparametric methods available for other selected settings.

There are times in the collection of data that we are prevented from actually observing the values of all of the observations. Such censoring leading to only partial information about the random variables of interest can be a direct result of the statistical design governing our data collection or it can be purely a consequence of additional random mechanisms affecting our data collection process.

Considerable attention in the literature has been devoted to three particular types of censoring, which we now describe. The first of these, known as type I censoring, corresponds to a fixed (preset) censoring time, t_c, at which the study is to come to an end. In this setting, instead of observing the random variables $Z_1, \ldots, Z_n$ of interest, we are only able to observe the truncated variables $W_i = \min(Z_i, t_c), i = 1, \ldots, n$. Type I censoring corresponds to medical and health sciences studies conducted for a fixed period of time after initiation and no entry to the study once begun.

A second type of censoring, known as type II censoring, corresponds to collecting survival (lifetime) data until a fixed number, say $r < n$, of the subjects have failed. Once this has occurred, the study is terminated. In this setting, we only observe the r smallest lifetimes (i.e. the first r **order statistics**) among $Z_1, \ldots, Z_n$. All we know about the remaining $n - r$ unobserved lifetimes is that they are at least as long as the final observed failure.

A third type of censoring, called random censoring, is probably the most common and the most complicated type of censoring associated with medical and health sciences data. In this setting, not only are the lifetimes random but the censoring times are also random. In clinical trials, for example, such random censoring could correspond to a study where not all subjects enter the study at the same time, but the study ends at one time, or to subjects leaving a study because they moved from the area or because of serious side-effects leading to discontinuation of the treatment.

Probably the earliest nonparametric approach to dealing directly with censored lifetime data was provided by **Kaplan & Meier** [30] in their development of the product limit estimator for the survival function $S(t) = 1 - G(t), -\infty < t < \infty$ (*see* **Survival Distributions and Their Characteristics**). The first two-sample rank procedure designed specifically to test hypotheses with censored data was provided by Gehan [19]. He proposed a direct extension of the Mann–Whitney form of the Mann–Whitney–Wilcoxon test statistic that provided a natural way to handle censored values occurring in either the X and/or Y sample data. A generalization of the Gehan two-sample test to the k-sample ($k \geq 3$) setting has been provided by Breslow [4]. For additional discussion of such rank-based procedures for censored data, the reader is referred to [42].

Other Important Nonparametric Approaches

Brief mention must also be made here of three other major initiatives in the development of nonparametric approaches to the analysis of medical and health

sciences data. Paramount among such developments is that of the proportional hazards model initially proposed by Cox [6] (*see* **Cox Regression Model**). Seldom has any single paper had such an impact on further research in the field. Kalbfleisch & Prentice [29] provide a nice discussion of the analysis of survival data by the use of the Cox proportional hazards model and extensions thereof. A second important thrust of more recent vintage has been the application of **counting process methods in survival analysis**. For a good discourse on this important methodology, the reader is referred to [1]. Finally, we need to mention the advent of the **bootstrap** as an important tool in the analysis of medical data. The survey articles [14] and [15] serve very well as introductions to this important topic, and its application to the analysis of censored data is discussed in [13].

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes, *Philosophical Transaction of the Royal Society of London* **27**, 186–190.

[3] Bradley, J.V. (1968). *Distribution-Free Statistical Tests*. Prentice-Hall, Englewood Cliffs.

[4] Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship, *Biometrika* **57**, 579–594.

[5] Conover, W.J. (1980). *Practical Nonparametric Statistics*, 2nd Ed. Wiley, New York.

[6] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[7] Critchlow, D.E. & Fligner, M.A. (1991). On distribution-free multiple comparisons in the one-way analysis of variance, *Communications in Statistics – Theory and Methods* **20**, 127–139.

[8] Damico, J.A. & Wolfe, D.A. (1987). Extended tables of the exact distribution of a rank statistic for all treatments: multiple comparisons in one-way layout designs, *Communications in Statistics – Theory and Methods* **16**, 2343–2360.

[9] Damico, J.A. & Wolfe, D.A. (1989). Extended tables of the exact distribution of a rank statistic for treatments versus control multiple comparisons in one-way layout designs, *Communications in Statistics – Theory and Methods* **18**, 3327–3353.

[10] Daniel, W.W. (1978). *Applied Nonparametric Statistics*. Houghton-Mifflin, Boston.

[11] Doksum, K. (1967). Robust procedures for some linear models with one observation per cell, *Annals of Mathematical Statistics* **38**, 878–883.

[12] Dwass, M. (1960). Some k-sample rank-order tests, in *Contributions to Probability and Statistics*, I. Olkin, S.G. Ghurye, H. Hoeffding, W.G. Madow & H.B. Mann, eds. Stanford University Press, Stanford, pp. 198–202.

[13] Efron, B. (1981). Censored data and the bootstrap, *Journal of the American Statistical Association* **76**, 312–319.

[14] Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society of Industrial Applications in Mathematics, CBMS-National Science Foundation Monograph, Vol. 38.

[15] Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science* **1**, 54–77.

[16] Fligner, M.A. & Rust, S.W. (1983). On the independence problem and Kendall's tau, *Communications in Statistics – Theory and Methods* **12**, 1597–1607.

[17] Fraser, D.A.S. (1957). *Nonparametric Methods in Statistics*. Wiley, New York.

[18] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.

[19] Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples, *Biometrika* **52**, 203–223.

[20] Gibbons, J.D. (1971). *Nonparametric Statistical Inference*. McGraw-Hill, New York.

[21] Gibbons, J.D. (1976). *Nonparametric Methods for Quantitative Analysis*. Holt, Rinehart, and Winston, New York.

[22] Hájek, J. (1969). *Nonparametric Statistics*. Holden Day, San Francisco.

[23] Hájek, J. & Sidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

[24] Hayter, A.J. & Stone, G. (1991). Distribution free multiple comparisons for monotonically ordered treatment effects, *Australian Journal of Statistics* **33**, 335–346.

[25] Hettmansperger, T.P. (1984). *Statistical Inferenc Based on Ranks*. Wiley, New York.

[26] Hodges, J.L., Jr & Lehmann, E.L. (1963). Estimates of location based on rank tests, *Annals of Mathematical Statistics* **34**, 598–611.

[27] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*. 2nd Ed. Wiley, New York.

[28] Jonckheere, A.R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika* **41**, 133–145.

[29] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[30] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[31] Kendall, M.G. (1938). A new measure of rank correlation, *Biometrika* **30**, 81–93.

[32] Kendall, M.G. (1962). *Rank Correlation Methods*, 3rd Ed. Griffin, London.

[33] Kendall, M.G. & Babington Smith, B. (1939). The problem of m rankings, *Annals of Mathematical Statistics* **10**, 275–287.

[34] Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione, *Giornale dell'Istituto Italiano degli Attuari* **4**, 83–91.

[35] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583–621.

[36] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

[37] Mack, G.A. & Wolfe, D.A. (1981). K-sample rank tests for umbrella alternatives, *Journal of the American Statistical Association* **76**, 175–181.

[38] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[39] Marascuilo, L.A. & McSweeney, M. (1977). *Nonparametric and Distribution-free Methods for the Social Sciences*. Wadsworth, Belmont.

[40] McDonald, B.J. & Thompson, W.A., Jr (1967). Rank sum multiple comparisons in one- and two-way classifications, *Biometrika* **54**, 487–497.

[41] Miller, R.G., Jr (1966). *Simultaneous Statistical Inference*. McGraw-Hill, New York.

[42] Miller, R.G., Jr, Gong, G. & Muñoz, A. (1981). *Survival Analysis*. Wiley, New York.

[43] Nemenyi, P. (1963). Distribution-free multiple comparisons, PhD Thesis. Princeton University.

[44] Noether, G.E. (1967). *Elements of Nonparametric Statistics*. Wiley, New York.

[45] Page, E.B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks, *Journal of the American Statistical Association* **58**, 216–230.

[46] Pratt, J.W. & Gibbons, J.D. (1981). *Concepts of Nonparametric Theory*. Springer-Verlag, New York.

[47] Randles, R.H. & Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

[48] Samara, B. & Randles, R.H. (1988). A test for correlation based on Kendall's tau, *Communications in Statistics – Theory and Methods* **17**, 3191–3205.

[49] Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

[50] Smirnov, N.V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples, *Bulletin of Moscow University* **2**, 3–16 (in Russian).

[51] Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology* **15**, 72–101.

[52] Spjøtvoll, E. (1968). A note on robust estimation in analysis of variance, *Annals of Mathematical Statistics* **39**, 1486–1492.

[53] Steel, R.G.D. (1960). A rank sum test for comparing all pairs of treatments, *Technometrics* **2**, 197–207.

[54] Terpstra, T.J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking, *Indagationes Mathematicae* **14**, 327–333.

[55] Walsh, J.E. (1962). *Handbook of Nonparametric Statistics*. Van Nostrand, Princeton.

[56] Walsh, J.E. (1965). *Handbook of Nonparametric Statistics*, Vol. II. Van Nostrand, Princeton.

[57] Walsh, J.E. (1968). *Handbook of Nonparametric Statistics*, Vol. III. Van Nostrand, Princeton.

[58] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

[59] Wilcoxon, F., Katti, S.K. & Wilcox, R.A. (1973). Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test, in *Selected Tables in Mathematical Statistics*, Vol. 1, H.L. Harter & D.B. Owen, eds. American Mathematical Society, pp. 171–259.

[60] Wilcoxon, F. & Wilcox, R.A. (1964). *Some Rapid Approximate Statistical Procedures*, 2nd Ed. American Cyanamid Co., Lederle Laboratories, Pearl River.

DOUGLAS A. WOLFE

Nonparametric Regression Analysis of Longitudinal Data

Introduction

Longitudinal data involve repeated measurements that are recorded over a period of time on the same subject. The number of measurements for each subject may be different and is denoted by n_i for the ith subject when there are a total of n subjects in the study. We use $N = \sum_{i=1}^{n} n_i$ to denote the total number of observed measurements on all subjects. The time points at which those measurements were taken are also often different and are denoted by $t_{i1}, \ldots, t_{in_i}$. We use $Y_{ij} = Y(t_{ij})$ to denote a measurement for the ith subject at the jth time point, and $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{in_i})$ to denote the observed vector for the ith subject. This leads to a correlation structure between the repeated measurements within the same subject. Longitudinal data arise commonly in health sciences and engineering research, but different terms have been applied to describe them. They are usually referred to as "longitudinal data" in biomedical applications, where a small number of repeated measurements, n_i, over time per subject is common, and as "functional data" in engineering and biological applications, where n_i is often large. Statistical approaches to analyze such data have also been intrinsically different for longitudinal and functional data. Longitudinal data are treated as vectors, $\mathbf{Y_i}$, with subject-specific dimension n_i for the ith subject, while functional data are regarded as realizations of random processes with smooth paths, $Y(t)$, that are observed at discrete time points. Parametric GEE-based marginal models and parametric random effects models are the predominant approaches for longitudinal data, and non- or semiparametric approaches are the standard practice to analyze functional data. Recent challenges in the biomedical and biological fields prompted the development of more complex and flexible approaches to model longitudinal data. Nonparametric regression, well known to be more data adaptive and less restrictive than parametric approaches, thus emerged as a promising alternative to handle longitudinal data. For readers searching for such nonparametric approaches in the literature, a keyword to include is "functional data" in addition to "longitudinal data". The two books [20] and [21] on functional data analysis provide an excellent introduction to this topic.

In this article, we focus on situations in which the responses for the experimental subjects are longitudinal data. The covariates can be a baseline vector (X), a time-varying covariate vector $(X(t_{ij}))$, which is longitudinal data itself, or a combination of both. Key issues in nonparametric regression for such data include inference for the overall mean and nonparametric fixed effects, and modeling of the within-subject covariance structure through nonparametric random effects. We will use the fecundity data set described in the next section to illustrate these issues.

We begin with nonparametric mean function estimation, treating the overall mean as a function on a time interval, $[0, T]$, over which data were recorded for the subjects. This overall mean function is assumed to be smoothed and is often referred by researchers as the mean curve.

Estimating the Overall Mean as a Function of Time

We assume that the overall mean function (or mean curve), $\mu(t) = E(Y(t))$, is an unknown but smooth function on $[0, T]$. Hence, $E\ (Y_{ij}) = \mu(t_{ij})$. If we ignore for the moment the within-subject correlations, then the mean function can be regarded as a nonparametric regression function with the regressor being the time variable. A scatter-plot smoother can then be applied to all N observed data points (t_{ij}, Y_{ij}) to estimate the mean function $\mu(\cdot)$. Specifically, the estimate at a particular time t is

$$\hat{\mu}(t) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_{ij} Y_{ij}. \tag{1}$$

This is simply a weighted average of all the N measurements, where the weights w_{ij} depend on the design points t_{ij} and the particular smoother. The choice of the smoother can be subjective and common choices include the kernel method [8], local polynomials [6], [14] and [32] and splines [22], [23] and [24].

Standard software such as S-plus can be employed easily to obtain a mean function estimate. The only difference with respect to the standard nonparametric regression setting is that repeated measurements

are available for each subject. For this reason, the estimated mean function in (1) can be expected to be consistent if n tends to infinity, provided the time points t_{ij} are spread out over the design interval $[0, T]$.

However, standard nonparametric smoothing methods may need to be adjusted for longitudinal data. An example for such an adjustment is the choice of smoothing (or tuning) parameter that is required by all smoothing methods. For the popular cross-validation method, it was shown in [22] that a leave-one-subject-out scheme should be employed for longitudinal data rather than the standard leave-one-observation-out scheme. All smoothing procedures require the proper choice of a tuning parameter. This problem is less understood and studied in longitudinal settings, and further research is needed.

Example of Reproductive Fecundity Data: We illustrate here the mean function estimate through a data set collected on 1000 female Mediterranean fruit flies (medflies) that was analyzed in [6]. Daily egg production, in terms of the number of eggs laid, were recorded individually until death for each of the 1000 female medflies. This results in a sample of 1000 longitudinally recorded fecundity curve data, with $Y_{ij} =$ number of eggs laid on day j by fly i. The goal is to explore the reproductive behavior of medflies through the pattern and modes of variation of these fecundity curves, $Y(t)$. Such information is important because reproduction is considered by evolutionary biologists as the single most important life history trait besides lifetime itself. See [6] and the references therein for details and information on biological features of the experiment.

Figure 1 provides the mean function estimates at various bandwidths based on a local linear scatter-plot smoother. Details of the procedure, including the leave-one-subject-out cross-validated bandwidth choice, are available in [6]. Because of the large number of subjects (there are 1000 flies) in the study and the dense recording per subject (repeated measurements were available daily), cross-validation selected a very small bandwidth at 2.5 days. The mean curve based on the larger bandwidth 15 is smoother but has larger bias than the mean curve based on the smaller bandwidth 2.5, as is expected for any nonparametric smoothing procedure.

The smoothing weights for the scatter-plot smoothing procedure in Figure 1 were determined by the choice of the local linear smoother and

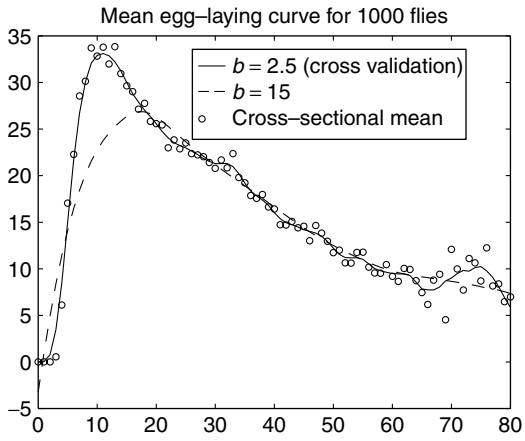

Figure 1 Mean egg-laying curves of 1000 female Mediterranean fruit flies. (a) Daily cross-sectional means of flies alive at the beginning of the day (○); (b) smoothed mean curve with fixed bandwidth $b = 15$ (- - - - - -); (c) smoothed mean curve with cross-validated bandwidth choice $b = 2.5$ (———)

the corresponding bandwidths following the standard practice in nonparametric regression. The within-subject correlation structure was not incorporated. An intriguing question for longitudinal data is how to effectively adjust the weights w_{ij} in (1) in the smoothing step to reflect the within-subject correlation structure of $\mathbf{Y_i}$. This was cleverly demonstrated for the case of smoothing splines in [29], and for local polynomial smoothers in [27] and [28]. It was shown that the asymptotic variance of the mean function estimators can be minimized if the weights are selected properly. However, these optimal weights require the use of the true within-subject correlation structure and do not necessarily minimize the asymptotic bias. The bias issue is more elusive and has not been resolved.

Nonparametric Fixed-effects Covariates

The procedures and discussion in the previous section apply directly to covariates other than time. To estimate the regression function $E(\mathbf{Y_i}|\mathbf{X_i})$, corresponding to fixed effects of covariates $\mathbf{X_i}$, simply replace in the scatter-plot smoother the t_{ij} by X_{ij} for time-varying covariates, and by $\mathbf{X_i}$ for vector covariates. Another framework mimicking the marginal approach of (Generalized Estimating Equations) can be found

in [14], [27], [28] and [29], where $E(Y_{ij}|X_{ij}) = \mu_{ij} = h(g(X_{ij}))$, with h a known and differentiable link function, and g an unknown smooth function. Here, the covariance structure of the response $\mathbf{Y}_i$ is also assumed to be a function of the means μ_{ij}, as suggested in the generalized linear model setting. The asymptotic variance of the minimum variance estimate, $\hat{g}(\cdot)$, was derived in [29] for smoothing splines and in [28] for local polynomial estimators. Additional results on semiparametric marginal models are also available in these two articles and [15] and [18].

Nonparametric Random effects

The overall mean function in Section 2 represents the population average, but individual trajectories may vary owing to subject effects, which also contribute to correlated repeated measurements within the same subject. Subject-specific random effects can be added for example by assuming as follows:

$$Y_i(t) = \mu(t) + v_i(t) + e_i(t), \quad i = 1, \ldots, n, \quad (2)$$

where μ is the unknown overall mean function, v_i are unknown subject-specific random effects reflecting the individual variation from the overall mean function, and e_i are measurement errors independent of v_i. The random effects are often regarded as realizations of a mean zero random process with smooth paths. It is thus expected that both the smoothed mean and random effects functions in (2) can be approximated using some basis functions. A B-spline basis, such as cubic splines, is a common choice and was proposed independently in [23] and [24], resulting in the following mixed-effects model:

$$Y_i(t) = \sum_{k=1}^{K} \beta_k B_k(t) + \sum_{k=1}^{K} b_k B_k(t) + e_i(t). \quad (3)$$

Here, β_k are coefficients, b_k are random variables with mean zero, $B_k(\cdot)$ is a basis of spline functions on $[0, T]$, and $e(t)$ is the measurement error. Consequently, the first summand yields the population mean function and corresponds to a fixed effect, while the second summand represents the random effects attributed to subject variations and describes the within-subject correlation structure. It is possible to use separate bases for random and fixed effects in equation (3). If we further assume normality for b_k and e_i, then model (3) becomes a linear mixed-effects model and thus can be fitted using either S-PLUS LME or SAS PROC MIXED. This computational advantage is an attractive feature of the B-spline approach. However, it requires the choice of the spline basis and the number of basis functions K, which in turn involves fairly complex choices of the number and location of knots. Cross-validation procedures or information-based criteria such AIC and BIC are among the suggestions in [23] for the choice of knots, but the issue is elusive and remains unsettled.

The B-spline basis approach may have difficulties for a data set that requires a large number of basis functions. This is because the degrees of freedom involved in (3) may be too small or even negative for sparse data. One solution is to use instead a local basis such as local polynomials in model (3) as only a low-degree polynomial is needed locally to fit the data, and often a linear polynomial suffices. This is explored in [32]. The trade-off is computation time as smoothing is done locally at each point while the B-spline approach is a global smoothing procedure.

Another remedy for the B-spline procedure is proposed in [10] and is based on the reduced rank procedure that involves the use of principal components. This approach aims at reducing the actual number of parameters needed in the nonparametric mixed-effects model (3) so that one can increase the degrees of freedom. However, it often involves a compromise in terms of computational feasibility and model flexibility. An alternative is the principal component analysis approach, a commonly used method for functional data that can be adapted to longitudinal data.

Principal Component Analysis Approach

Simply put, this approach just replaces the prespecified basis B_k in (3) by the eigenfunctions of the covariance operator of the response. This is the essence of principal component analysis and results in a data-adaptive basis that can effectively reduce the number of basis functions needed to model the random effects. They are effective dimension reduction tools for longitudinal data. The concept is similar to the problem to find the best K-dimensional linear model for stochastic processes, and has been extended in [4] and [22] to the case of functional or longitudinal data and termed functional principal components analysis.

Here, the response functions $Y_i(t)$ are considered realizations of smooth L^2 process with mean $\mu(t)$ and covariance function $\text{cov}(Y(s), Y(t)) = \gamma(s,t)$. The covariance function γ allows a spectral decomposition into orthonormal eigenfunctions $\rho_k(\cdot)$.

$$\gamma(s,t) = \sum_{k=1}^{\infty} \lambda_k \rho_k(s)\rho_k(t), \tag{4}$$

with ordered nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

Let $\langle \cdot, \cdot \rangle$ denote the inner product in L^2 space. The Karhunen-Loève representation for a randomly selected curve is

$$Y(t) = \mu(t) + \sum_{k=1}^{\infty} A_k \rho_k(t), \tag{5}$$

where the random variables a_k correspond to the principal component scores, and are given by

$$A_k = \langle \rho_k, Y - \mu \rangle. \tag{6}$$

The principal components A_k in (6) are uncorrelated random variables with

$$E(A_k) = 0, \quad \text{var}(A_k) = \lambda_k, \quad \sum_{k=1}^{\infty} \lambda_k < \infty,$$

that is, the kth eigenvalue in (4) corresponds to the variance of the kth principal component as in the multivariate case. The principal components A_k and basis functions ρ_k in (5) can be interpreted as defining the variation of the stochastic process about its mean function, and $A_1\rho_1$ explains the maximum amount of variation in Y among all functions that involve a single real-valued random variable. Similarly, the function $A_2\rho_2$ explains the maximum additional amount of process variation that is unexplained by $A_1\rho_1$, and so forth, for $k = 3, 4, \ldots$.

Methods to estimate the eigenfunctions and principal components are described in [22] on the basis of smoothing splines. The leave-one-subject-out cross-validation method to select the number of eigen-basis is also first proposed there. Theoretical properties of the functional principal components estimates can be found in [19] under the hypothetical assumption that the entire process $Y(t)$ is observable. Similar theoretical results for another functional principal components approach based on kernel smoothers are provided in [2].

Compared to the nonparametric mixed-effects model in (3), functional principal component analysis is more data adaptive and therefore typically requires fewer basis functions. It also has the advantage to allow direct interpretations in terms of modes of variation of the underlying process and is favored by biologists to explore the covariance structure of the data. Although functional principal component analysis is not yet available on standard statistical packages, it is not difficult to write code in either MATLAB or S-PLUS to perform this analysis.

Example of Fecundity data: To incorporate subject-specific random effects of the reproduction process of medflies, we perform principal component analysis on the fecundity data. As in [6], we restricted the analysis to the first 50 days of lifetime owing to high variability of the fecundity curves beyond day 50. This circumvents the problem of the eigen-analysis being dominated by erratic tail behavior of the data and provides a more sensible analysis. See [6] for more details. The rest of the principal component analysis presented below is based on the first 50 days of daily egg counts for the 167 flies that lived beyond day 50. The egg-laying curves $Y_i(t)$, $i = 1, \ldots, 167$, are considered as realizations of a stochastic process on the interval $T = [0, 50]$.

We applied the eigen-analysis based on local linear smoothing as described in [6]. The optimal number of principal components based on AIC is nine, but there is little gain after four components as these four components explain 95.88% of the total variation of the fecundity data. The eigenfunctions corresponding to the largest four eigenvalues of the covariance function, $\gamma(s,t)$, are shown in Figure 2. This example demonstrates how principal component analysis effectively reduces the dimension of the data from an infinite-dimensional curve to a few components. In fact, selecting two components may suffice as they explain already 82% of the total variation based on Figure 2. To see whether one can reasonably predict the shape of the fecundity curve using just two principal components, we proceed to fit the fecundity curves of four randomly selected flies using the Karhunen-Loève representation (5) with two components.

Figure 3 exhibits the observed and predicted egg-laying profiles of individual flies, as well as the overall mean curve. The overall mean curve would be the predicted curve when no random effects are included. As can be seen form Figure 3, the functional PCA

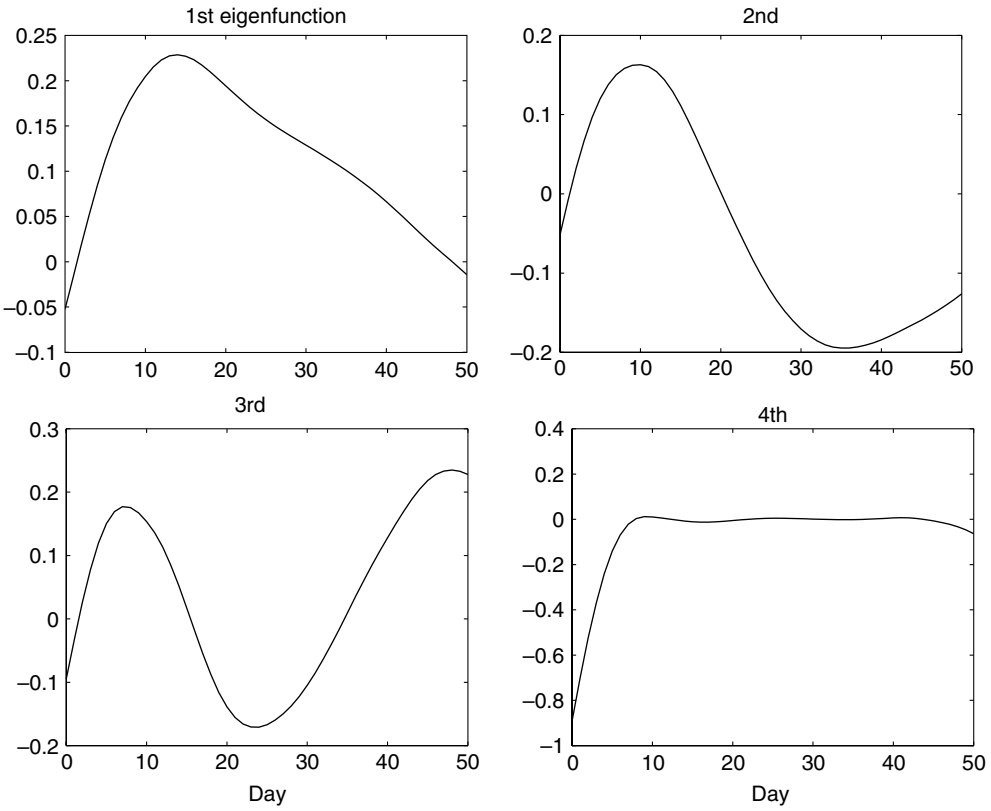

Figure 2 First four eigenfunctions for egg-laying data. The fraction of variation explained by each of these components are 0.6183, 0.2090, 0.0779 and 0.0536 respectively

approach seems to fit the curves reasonably well and has much less bias than the overall mean curve. We have thus demonstrated the effectiveness of the nonparametric principal components analysis approach for longitudinal data.

Although the fecundity data illustrated here were sampled at the same time points (daily in this case), the procedure can as well be applied to longitudinal data sampled at irregular time points.

Nonparametric Mixed-effects Models with Covariates

The procedures in the previous section allow to handle the time effects for a sample of individuals from the same population. When there are other covariates $\mathbf{X}$ that affect the longitudinal response data, one needs to incorporate these covariate effects in addition to the time effects. This has been explored for vector covariates in [6] by taking the conditional expectation with respect to $\mathbf{X}$ on both sides of (5). As a result, we have the following model:

$$E(Y_i(t)|\mathbf{X_i}) = \mu(t) + \sum_{k=1}^{\infty} E(A_{ik}|\mathbf{X_i})\rho_k(t) \qquad (7)$$

Procedures to estimate all the components including the mean function, the eigenfunctions and the conditional principal components $E(A_{ik}|\mathbf{X_i})$ can be found in [6], which also includes a semiparametric index model to tackle situations when the vector $\mathbf{X}$ is high dimensional. It is interesting to note here that the fixed effect of a covariate in (7) is derived from the unconditional principal components A_{ik} and eigenfunctions $\rho_k(t)$. This is because although A_{ik} has overall mean zero, the conditional mean $E(A_{ik}|\mathbf{X_i})$ in (7) is not zero and thus contributes to the fixed effects. Additional random effects can then be added to the model through $b_{ik} = A_{ik} - E(A_{ik}|\mathbf{X_i})$ to reach

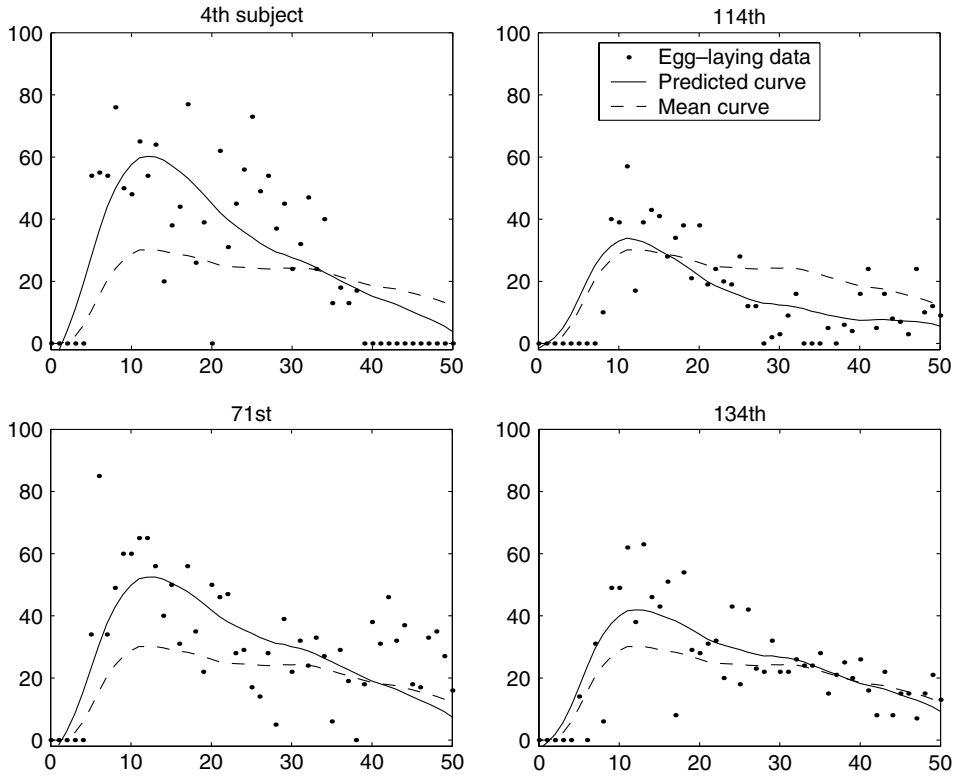

Figure 3 Observed data, predicted egg-laying curves and mean egg-laying curves for four randomly selected subjects

the following nonparametric mixed-effects model:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{\infty} E(A_{ik}|\mathbf{X_i})\rho_k(t) + \sum_{k=1}^{\infty} b_{ik}\rho_k(t). \tag{8}$$

Other types of mixed-effects functional PCA regression models include [3] and [25], and more parsimonious semiparametric mixed-effects models include [26] and [33].

Other Non- and Semiparametric Regression Approaches

So far, we have discussed briefly a few nonparametric approaches for longitudinal/functional data. There are many other non- and semiparametric alternatives. One of them is the Generalized Additive Model (GAM) of the form:

$$E(Y_i(t)) = \beta_0 + \sum_{k=1}^{P} g_k(X_{itk}) + e_i(t), \tag{9}$$

where $(Y_i(t), X_{it1}, \ldots, X_{itP})$ is observed at time t for the ith subject with P-dimensional covariates denoted by X_{itj} at time t_j. The functions g_k in (8) are unknown smooth functions. Backfitting algorithms were proposed in [1] and [30], and inference procedures studied in [16].

When the data exhibit a common shape or structure, a smooth curve g can be used to model this common shape with individual responses adjusted by some parametric transformation of the common curve. This is referred to as self-modelling regression (SEMOR) in [13] and studied in [12] and [17] among others. A more general semiparametric model that includes SEMOR is recently proposed in [11].

Another approach that has emerged recently to model longitudinal data is the varying-coefficients model of the form:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \beta(t)X_i(t) + e_i(t). \tag{10}$$

This was first applied to longitudinal data in [9], and subsequently studied in [5], [7] among others. See the review of this model in [31] for details.

References

[1] Berhane, K. & Tibshirani, R.J. (1998). Generalized additive models for longitudinal data, *Canadian Journal of Statistics* **26**, 517–535.

[2] Boente, G. & Fraiman, R. (2000). Kernel-based functional principal components, *Statistics and Probability Letters* **48**, 335–345.

[3] Capra, W.B. & Müller, H.G. (1997). An accelerated time model for response curves, *Journal of the American Statistical Association* **92**, 72–83.

[4] Castro, P.E., Lawton, W.H. & Sylvestre, E.A. (1986). Principal modes of variation for processes with continuous sample curves, *Technometrics* **28**, 329–337.

[5] Chiang, C.T., Rice, J.A. & Wu, C.O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables, *Journal of the American Statistical Association* **96**, 605–619.

[6] Chiou, J.M., Müller, H.G. & Wang, J.L. (2002). Functional quasi-likelihood regression models with smooth random effects, *Journal of Royal Statistical Society, Series B* **65**, 405–423.

[7] Fan, J. & Zhang, J.T. (2000). Statistical estimation in varying-coefficient models, *Annals of Statistics* **27**, 1491–1518.

[8] Hart, J.P. & Wehrly, T.E. (1986). Kernel regression estimation using repeated measurements data, *Journal of the American Statistical Association* **81**, 1080–1088.

[9] Hoover, D., Rice, J., Wu, C. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika* **85**, 809–822.

[10] James, G.M., Hastie, T.J. & Sugar, C.A. (2000). Principal component models for sparse functional data, *Biometrika* **87**, 587–602.

[11] Ke, C. & Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications, *Journal of the American Statistical Association* **96**, 1272–1298.

[12] Kneip, A. & Gasser, T. (1988). Convergence and consistency results for self-modeling nonlinear regression, *Annals of Statistics* **16**, 82–112.

[13] Lawton, W.H., Sylvestre, E.A. & Maggio, M.S. (1972). Self-modeling nonlinear regression, *Technometrics* **14**, 513–532.

[14] Lin, X. & Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error, *Journal of the American Statistical Association* **95**, 520–534.

[15] Lin, X. & Carroll, R.J. (2001). Semiparametric regression for clustered data using generalized estimating equations, *Journal of the American Statistical Association* **96**, 1045–1056.

[16] Lin, X. & Zhang, D.W. (1999). Inference in generalized additive mixed models by using smoothing splines, *Journal of Royal Statistical Society, Series B* **61**, 381–400.

[17] Lindstrom, M.J. (1995). Self-modeling with random shift and sacle parameters and a free-knots spline shape function, *Statistics in Medicine* **14**, 2009–2021.

[18] Moyeed, R.A. & Diggle, P.J. (1994). Rates of convergence in semi-parametric modelling of longitudinal data, *Australian Journal of Statistics* **36**, 75–93.

[19] Pezzulli, S. & Silverman, B.W. (1993). Some properties of smoothed principal components analysis for functional data, *Computational Statistics* **8**, 1–16.

[20] Ramsay, J.O. & Silverman, B.W. (1997). *Functional Data Analysis*. Springer, New York.

[21] Ramsay, J.O. & Silverman, B.W. (2002). *Applied Functional Data Analysis*. Springer, New York.

[22] Rice, J.A. & Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of Royal Statistical Society, Series B* **53**, 233–243.

[23] Rice, J.A. & Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves, *Biometrics* **57**, 253–259.

[24] Shi, M., Weiss, R.E., & Taylor, J. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves, *Applied Statistics* **45**, 151–163.

[25] Staniswalis, J.G. & Lee, J.J. (1998). Nonparametric regression analysis of longitudinal data, *Journal of the American Statistical Association* **93**, 1403–1418.

[26] Wang, Y. (1998). Mixed-effects smoothing spline analysis of variance, *Journal of Royal Statistical Society, Series B* **60**, 159–174.

[27] Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation, *Biometrika* **90**, 43–52.

[28] Wang, N., Carroll, R.J. & Lin, X. (2005). Ecient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association* **100**, 147–157.

[29] Welch, A., Lin, X. & Carroll, R.J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods, *Journal of the American Statistical Association* **97**, 482–493.

[30] Wild, C.J. & Yee, T.W. (1996). Additive extension to generalized estimating equation methods, *Journal of Royal Statistical Society, Ser. B* **58**, 711–725.

[31] Wu, C. & Yu, K.F. (2002). Nonparametric varying-coefficient models for the analysis of longitudinal data, *International Statistical Institute Review* **70**, 373–393.

[32] Wu, H. & Zhang, J.T. (2002). Local polynomial mixed-effects models for longitudinal data, *Journal of the American Statistical Association* **97**, 883–897.

[33] Zhang, D., Lin, X., Raz, J. & Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data, *Journal of the American Statistical Association* **93**, 710–719.

J.-L. WANG

Nonparametric Regression

Nonparametric methods of regression seek to describe the relationship between an explanatory variable, x and a **response variable**, Y, whilst making only the most general type of assumptions about the functional form of this relationship. The idea is that when a scatterplot (*see* **Graphical Displays**) does not indicate any simple pattern of dependency of Y on x, the data are allowed to "speak for themselves" in determining which function fits them best. With data such as these, nonparametric regression provides a useful tool for exploratory analysis. The resulting fitted curve may suggest a simple parametric model, and provides a method of prediction when this is not the case.

If n data pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$ have been observed then the relationship may be modeled by

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where ε_i is an error term with zero mean and the function $m(x)$ is the conditional expectation, $\mathrm{E}(Y|x)$. In linear regression, for example, this function is defined by $m(x) = \alpha + \beta x$, where α and β are unknown parameters. Nonparametric regression methods, on the other hand, make only the assumption that $m(x)$ is a smooth function whose form is unknown. The problem is then to find an automatic method for constructing an estimate of $m(x)$ from the data. There are a number of nonparametric techniques for doing this, the most popular being kernel regression [7] and spline smoothing [2]. Hardle [3] describes both these and other nonparametric regression methods.

In the context of kernel regression the simplest method is due to Nadaraya [5] and Watson [8]. Their method works by estimating $\mathrm{E}(Y|x)$, i.e. $m(x)$, by a weighted average of the responses. The weights are chosen so that Y_i provides a large contribution to the average if x_i is close to x, and a small contribution if x_i is distant. Specifically, the Nadaraya–Watson estimator is defined by

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K(x - x_i) Y_i}{\sum_{i=1}^{n} K(x - x_i)}.$$

In this equation K is a kernel function, typically positive, symmetric about zero and integrating to one. The parabolic shaped kernel given by

$$K(u) = \begin{cases} \frac{3}{4} h^{-1} \left[1 - \left(\frac{u}{h} \right)^2 \right], & -h < u < h, \\ 0, & \text{otherwise}, \end{cases}$$

is a typical choice. This kernel is defined in terms of a scaling parameter, h, called the *bandwidth* (or *window width* by some authors). In practice, the value of h should be chosen so that the range of the kernel function is appropriate for the scale of the data. The manner in which this should be done has been the subject of much research. See Chapter 5 of Simonoff's monograph [6], for example. With the use of this kernel, $K(x - x_i)$ will take smaller values as the relative distance $|x - x_i|/h$ increases.

Kernel estimators fit into the class of local **polynomial regression** models. In this approach $\mathrm{E}(Y|x)$ is estimated by fitting a weighted polynomial regression, of degree r, say, to data whose x_i value is local to x. The kernel function defines the weights. The Nadaraya–Watson estimator corresponds to the case $r = 0$.

Spline smoothing is a somewhat different approach to nonparametric regression; its motivating philosophy is as follows. In fitting a curve to a scatterplot there are two (conflicting) interests. In the first place the regression function should be close to the data points. Thus, if $\tilde{m}(x)$ is a function estimating $m(x)$, then we wish the residual sum of squares, $\sum_{i=1}^{n} [Y_i - \tilde{m}(x_i)]^2$, to be small. However, there is usually good reason to believe that the function $m(x)$ is relatively smooth, so our estimator $\tilde{m}(x)$ should reflect this. The smoothness of $\tilde{m}(x)$ (assumed twice differentiable) can be measured by $\int \tilde{m}''(x)^2 \, \mathrm{d}x$, with larger values indicating greater roughness. Hence, the expression

$$\sum_{i=1}^{n} [Y_i - \tilde{m}(x_i)]^2 + \lambda \int \tilde{m}''(x)^2 \, \mathrm{d}x$$

is a penalty criterion, taking large values if $\tilde{m}(x)$ fits the data poorly or is unacceptably rough. It can be shown that the particular estimator, $\hat{m}(x)$, which minimizes this criterion is a cubic spline – that is, a piecewise cubic polynomial. The balance between goodness of fit to the data and smoothness of function is controlled by the parameter λ, with larger values

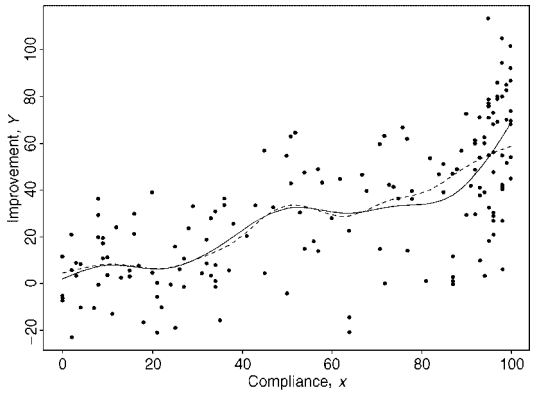

Figure 1 Nonparametric regressions for cholostyramine data [1]. Men were supposed to take six packets of cholostyramine (a cholesterol-lowering drug) per day. Compliance, x, is the percentage of the intended dose actually taken, whilst improvement, Y, is the decrease in total plasma cholesterol over the period of the treatment. The broken line is fitted using kernel regression, the unbroken one using spline smoothing

of λ leading to smoother $\hat{m}(x)$. The choice of an optimal λ, resulting in an estimator $\hat{m}(x)$ which picks up true trend but ignores haphazard variation in the data, is of considerable importance in the practical implementation of spline smoothing.

Both kernel and spline methods are illustrated in Figure 1 using data on the drug cholostyramine, given in [1]. The fitted curves bring out structure in the data, such as the rapid rise in the response for compliance of 90% or more, which would have been lost had a linear regression been fitted.

Nonparametric regression techniques can be generalized to cope with multiple explanatory variables, although the resulting methods tend to be computer-intensive. If there are p explanatory variables, with x_{ij} denoting the value of the jth one for the ith observation, then a natural extension of (1) is

$$Y_i = \sum_{j=1}^{p} m_j(x_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n.$$

This type of model is referred to as a **generalized additive model**, and is discussed in detail in [4].

References

[1] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
[2] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.
[3] Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.
[4] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, London.
[5] Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and its Applications* **10**, 186–190.
[6] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer, New York.
[7] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*. Chapman & Hall, London.
[8] Watson, G.S. (1964). Smooth regression analysis, *Sankhyā A* **26**, 101–116.

MARTIN L. HAZELTON

Nonrandomized Trials

A comparative **clinical trial** is a planned experiment in human subjects involving two or more treatments, where the primary purpose is to evaluate the relative effectiveness of the treatments. Often, the comparison is between two treatments, a proposed new treatment for the disease and standard therapy. Usually, the standard therapy has been utilized in other clinical trials, often those at the same institution(s) conducting the comparative trial, so patient characteristics related to **prognosis** may be known. In many clinical trials, patients entering are randomized to the available treatments (*see* **Randomized Treatment Assignment**), but there may be circumstances for conducting a nonrandomized trial in which the **control** group might be from: a historical control series of patients; concurrent controls, either at the same or other institutions; control patients from a computerized database (*see* **Administrative Databases**); or controls from articles reported in the literature. Byar et al. [4] have summarized the major arguments for conducting randomized clinical trials (RCTs), and Pocock [13] gives reasons why nonrandomized studies are likely to yield misleading results.

Articles giving arguments in favor of conducting nonrandomized trials are given by Gehan and Freireich [7–9]. The general arguments for conducting a nonrandomized trial involving a historical control treatment are that all knowledge is historical and that modifications are made as evidence accumulates. In a nonrandomized historical control trial (HCT), results of the proposed new treatment administered to consecutive patients are compared with those from a historical control group. This approach is consistent with the acquisition of knowledge by application of the principles of the scientific method. Results for patients on the new treatment can be compared with predictions of outcome for the standard treatment based on the premise that the past is the best guide to knowledge about the future. Even advocates of RCTs must accept some historical data; namely, the results of their own studies, to make predictions about future results for treatment. Confirmation of results of experimental therapies in multiple studies makes their acceptance more likely.

HCTs require much smaller numbers of patients and shorter time periods for their conduct than RCTs designed to meet equivalent objectives. Also, it is likely that a larger number of patients will be available for an HCT, since some patients will not accept randomization to treatment, whereas they would accept assignment to a new treatment that has promise of being better than a standard therapy. If an investigator is studying a proposed new (A) vs. standard (B) therapy in a nonrandomized trial and sufficient prior data are available so that the response rate (say p) for the standard therapy may be assumed known, then the number of patients required to compare A with B is only one-quarter of that required for an RCT with an equivalent statistical significance level (*see* **Level of a Test**) and **power** [9]. When the response rate is not assumed known, but is estimated from a historical control series of moderate size, Makuch & Simon [11] give tables for the number of patients needed. For example, if there are 100 patients in the historical control series and the response rate of the standard therapy is 40% and it is desired to detect a 20% improvement in response rate for the proposed new therapy at a 5% significance level (one-sided test) with a statistical power of 80%, then 52 patients are needed on the new therapy rather than 76 patients in an RCT. When the historical control series is moderate in size (about 50 patients or more), the number of patients required on the new treatment is generally less than that for the new treatment group alone in an RCT (*see* **Sample Size Determination for Clinical Trials**).

A clinical investigator conducting a nonrandomized trial with a historical control group has no ethical dilemma when advising his patients about entry into the study (*see* **Ethics of Randomized Trials**). If it is accepted that all clinical investigators should seek results that are better than those observed in the past, then it follows that no study should be started unless there is some evidence suggesting that the new therapy is at least as good or possibly better than the standard therapy. In such a circumstance, a clinical investigator doing a nonrandomized study would be entering all patients on the new treatment that was predicted to be better. In contrast, the ethical basis of the RCT depends upon the absence of convincing evidence about the relative merits of the two treatments.

Further, ethical concerns arise in an RCT, but not in an HCT, when interim results suggest that the new treatment is better than the standard therapy at some level of statistical significance, say $P = 0.25$. Such results could arise in a clinical trial designed to accrue a fixed number of patients or in

a sequential trial (*see* **Data and Safety Monitoring**) when a boundary point for deciding in favor of one of the treatments had not quite been reached. In such a circumstance, could a clinical investigator honestly seek the informed consent (*see* **Ethics of Randomized Trials**) of a potential patient by stating that the evidence favoring each of the treatments was equivalent? Of course, the same ethical concerns could also arise in an HCT, if interim results suggested that the standard therapy was better.

For randomized comparative trials sponsored by the **National Institutes of Health**, **data and safety monitoring boards** (DSMBs) have been set up to evaluate data accrued during the trials at regular intervals and decide about their continuation, without informing the clinical investigators participating. In such circumstances, the participating clinical investigators could ask for the informed consent of patients to enter the trial without the burden of knowing the interim results.

The outstanding criticism of nonrandomized studies involving historical control groups is that, consciously or unconsciously, patients may be selected to receive the new treatment that have a more favorable prognosis than patients receiving the standard therapy (*see* **Bias from Historical Controls**). Thus, the clinical trial of the new treatment may yield a positive result because the prognosis of the group of patients and not the treatment was more favorable. When a large body of data is available on the standard treatment, techniques for determining prognostic factors are well known [3], and knowledge of these factors may be used to stratify (*see* **Stratification**) patients or to adjust the comparison of the new vs. standard treatment by the use of **regression** models.

If a **multivariate multiple regression** model involving prognostic factors (e.g. a **logistic** or **proportional hazards** model) is available relating the outcome of treatment to prognostic factors and the model explains a substantial amount of the total variation in outcome (*see* **Explained Variation Measures in Survival Analysis**), the model may be used to test for treatment effects after adjustment for the prognostic features of the patients. An example of a proportional hazards model comparing disease-free survival and survival between treatment groups in a breast cancer HCT study is given by Buzdar et al. [5]. Simon [14] describes some statistical regression models, and their usefulness in **prediction** and **inference**, and gives methods for examining adequacy of fit (*see* **Goodness of Fit**) and some uses of regression models in clinical oncology. A crucial assumption needed to utilize the regression modeling approach is that the relationship of patient characteristics to outcome is essentially the same in the historical control and current study period.

Planners of RCTs rely primarily on randomization and secondarily on stratification and regression models as techniques for ensuring the validity of comparisons between treatments, whereas those preferring nonrandomized studies with historical control groups can use only the latter two procedures. Arguing that a historical control group might not be comparable with a new treatment group involves asserting either that there was an unknown prognostic feature of major importance (in addition to those known) or that factors related to the difference in time periods were responsible for the observed treatment difference. If such unfavorable events did occur in a single study, then the investigator who did not randomize would have to discover in a subsequent confirmation study that the new treatment was not as beneficial as expected, whereas the investigator who randomized would discover this within the trial.

Nonrandomized studies, possibly involving concurrent treatment groups, might be used to resolve controversial questions, such as those when preliminary data suggest that one treatment is substantially better than the other or when radically different types of therapy are being compared. In the latter circumstance, proponents of the differing forms of therapy (e.g. radiotherapy vs. radical surgery) might enter consecutive patients on a single therapy and use stratification and/or regression models involving prognostic factors to test for differences in effectiveness between treatments. Gehan [8] gives some examples in clinical oncology of studies in osteosarcoma, brain metastases, and localized stomach cancer in which RCTs were planned, but had substantial difficulties in their conduct because the studies involved controversial questions. Patients were unwilling to give informed consent to be randomized to the differing forms of therapy.

Baker & Lindeman [1] have proposed a nonrandomized paired availability design for evaluating epidural analgesia during labor. In their application, the design consists of independent pairs of experimental (epidural analgesia available) and control (epidural analgesia not available) groups, each group being patients treated at a hospital before

(control group) or after (experimental group) epidural analgesia became available. The fundamental characteristics of the design are: the intervention is availability of treatment; the **target population** from which subjects arise is well defined with little migration in or out; and the study involves many pairs of control and experimental groups. Baker & Lindeman developed a test of the **null hypothesis** that the receipt of intervention will increase response (measured by the percentage of patients having Caesarean sections) by some specified nonzero amount, and applied their results to a study of epidural analgesia.

Since clinical research in recent years has produced many efficacious new treatments, it is reasonable to ask which of the important advances in an area of clinical medicine in the last 25 years can be attributed to HCTs and which to RCTs. At least in cancer research there is strong evidence that new treatment regimens for acute leukemia, choriocarcinoma, lymphoma, lung cancer, osteosarcoma, breast cancer, testicular cancer, and sarcoma have come from nonrandomized studies. It should also be stated that RCTs have debunked false claims made for some new treatments [6]. Grage & Zelen [10] pointed out that historical controls tend to exaggerate the value of a new treatment, using as an example the treatment of metastatic colorectal carcinoma to the liver.

Although Pocock [13] argues strongly in favor of the RCT, he gives the requirements for a nonrandomized study involving a valid historical control group [12]. These are: the control group has received a precisely defined treatment in a recent previous study; the criteria for **eligibility**, work-up, and evaluation of treatment must be the same; important prognostic features should be known and be the same for both treatment groups; and there should be no unexplained indications leading one to expect different results. A further proviso might be that if there are some modest differences between treatment groups with respect to these features, then it should be established that these were not sufficient to explain any observed differences in outcome between treatment groups.

The use of computerized data banks with information on previous patients in a given institution has been advocated as a substitute for the RCT by some enthusiasts. For example, Starmer et al. [15] give an example of the utilization of a data bank in the management of chronic illness. However, Byar [2] gives strong arguments why databases should not be a substitute for RCTs. Literature controls might be considered for a nonrandomized study in which the control group is made up of patients previously reported in the literature. Of course, it is possible that there are substantial differences in patient selection and the experimental environment, such that a meaningful comparison of the new with the standard therapy would not be possible. Gehan & Freireich [9] give some examples in which it seemed reasonable to compare a new with the standard therapy involving control patients from the literature.

In clinical oncology, **phase I trials**, designed to find a maximum tolerated dose of therapy, and **phase II trials**, designed to determine whether or not the therapy is worth pursuing further, are generally nonrandomized, but such trials do not include a standard therapy comparison group and are conducted prior to the comparative (phase III) trial. **Outcomes research** studies are also nonrandomized, and a critical component of such research is inferring the relationship between the therapeutic process and outcomes – the major difficulty being the evaluation of processes when these are not randomized.

Without question, the "gold standard" in clinical research is the RCT. However, a nonrandomized trial with a historical control group might be considered in some circumstances. For planners of such studies, the difficulty to be overcome is to demonstrate that the groups of patients receiving the new vs. the standard therapy had comparable probabilities of favorable outcomes.

References

[1] Baker, S.G. & Lindeman, K.S. (1997). The paired availability design: a proposal for evaluating epidural analgesia during labor, *Statistics in Medicine* **13**, 2269–2278.

[2] Byar, D. (1980). Why databases should not replace randomized clinical trials, *Biometrics* **36**, 337–342.

[3] Byar, D. (1984). Identification of prognostic factors, in *Cancer Clinical Trials: Methods and Practice*, M. Buyse, M. Staquet & R. Sylvester, eds. Oxford University Press, Oxford, pp. 423–443.

[4] Byar, D.P., Simon, R.M. et al. (1976). Randomized clinical trials: perspectives on some recent ideas, *New England Journal of Medicine* **295**, 74–80.

[5] Buzdar, A.U., Gutterman, J.U., et al. (1978). Intensive post-operative chemoimmunotherapy for patients with Stage II and Stage III breast cancer, *Cancer* **41**, 1064–1075.

[6] Chalmers, T.C. & Block, J.B. (1972). Controlled studies in clinical cancer research, *New England Journal of Medicine* **285**, 75–78.

[7] Freireich, E.J. & Gehan, E.A. (1979). The limitations of the randomized clinical trial, in *Methods in Cancer Research*, Vol. 17, H. Busch & V. DeVita, eds. Academic Press, New York, Chapter 8, pp. 277–310.

[8] Gehan, E.A. (1984). The evaluation of therapies: historical control studies, *Statistics in Medicine* **3**, 315–324.

[9] Gehan, E.A. & Freireich, E.J. (1974). Non-randomized controls in cancer clinical trials, *New England Journal of Medicine* **290**, 198–203.

[10] Grage, T.B. & Zelen, M. (1982). The Controlled Randomized Trial and the Evaluation of Cancer Treatment – The Dilemma and Alternative Designs, *UICC Technical Report Series* **70**, 23–47.

[11] Makuch, R. & Simon, R. (1972). Sample size considerations for non-randomized comparative studies, *Journal of Chronic Diseases* **33**, 175–181.

[12] Pocock, S.J. (1976). The combination of randomized and historical controls in clinical trials, *Journal of Chronic Diseases* **29**, 175–188.

[13] Pocock, S.J. (1985). *Clinical Trials: A Practical Approach*. Wiley, New York, pp. 50–62.

[14] Simon, R. (1984). Use of regression models: statistical aspects, in *Cancer Clinical Trials: Methods and Practice*, M. Buyse, M. Staquet & R.J. Sylvester, eds. Oxford University Press, Oxford, pp. 444–466.

[15] Starmer, F. et al. (1974). Data bank use in management of chronic disease, *Computers and Biomedical Research* **7**, 111–116.

(*See also* **Adaptive and Dynamic Methods of Treatment Assignment**)

EDMUND A. GEHAN

Nonresponse

It would be preferred in virtually any **sample survey** to obtain an answer to every questionnaire item from every sample member who is eligible to participate. Unfortunately, almost all surveys fail to achieve that level of performance. Nonresponse may occur at the level of the sample unit (*unit nonresponse*) or in an individual questionnaire item *(item nonresponse)*. It is a potential source of error in survey estimates because it may cause some segments of a **target population** to be underrepresented. Also, nonresponse may reduce statistical **power** by resulting in a measured sample that is smaller than the desired sample size. Concern about nonresponse has increased over the past two decades as survey researchers have observed a general decline in response rates [6, 7, 14, 19–21]. This article reviews the causes of survey nonresponse, the nature of nonresponse error and its potential impact on survey estimates, and techniques for measuring and reporting nonresponse.

Causes of Nonresponse

Total Nonresponse

Some researchers categorize nonresponse as **nonsampling error**, considering it as a function only of the data collection process. However, as will be described below, at the unit level, nonresponse may be a function of the sampling process as well. Therefore, overall it is most appropriate to consider survey error attributable to nonresponse as *nonobservation error* [14]. This broader concept appropriately casts survey nonresponse as the failure of the survey process to obtain full participation from all eligible members of a sample.

The *total nonresponse* for any particular measured variable is the sum of the two levels of nonresponse: unit nonresponse and item nonresponse (both expressed as percentages). For example, if a survey fails to obtain participation from 20% of the eligible sample, and if among the 80% who participate no response is obtained from 10% for the variable of interest, total nonresponse for that variable is $0.20 + 0.10 \times 0.80 = 28\%$. Therefore, although in most surveys unit nonresponse accounts for the largest proportion of total nonresponse, it is important for researchers to account for both levels of nonresponse, in combination as well as individually.

The above example applies to the typical **cross-sectional** sample survey. Total nonresponse is both more complicated and usually more serious in the case of sample attrition in a longitudinal **panel study**, a design in which repeated measures are to be obtained from the same respondents over two or more waves (observation points). Total nonresponse for any particular variable in a panel study is cumulative over the nonresponse at each wave. Therefore, it is important to account for overall nonresponse for the duration of a panel study as well as wave-specific nonresponse.

Unit Nonresponse

Unit nonresponse occurs when members of a sample who are eligible to participate in a survey either do not participate at all or participate only partially, such that sufficient data are not obtained to include them in the analysis. Depending on the survey design, a unit may be an individual, a household, an institution or organization (e.g. a school), or other group (*see* **Unit of Analysis**). Unit nonresponse is a function of two components. The first is failure to contact sample members (noncontacts), which appears to account for most cases of unit nonresponse [20]. Examples of this type of problem include cases where, for the entire survey period, persons are not at home (e.g. due to business or vacation travel) or constantly use a telephone answering machine or voice-mail system. The second component of unit nonresponse is failure to obtain participation. This includes two types of cases. One is where a sample member refuses to complete an interview/questionnaire. The other is where a sample member who otherwise would complete an interview/questionnaire is unable to do so, for example because of an illness or injury, or being too busy with other activities.

Although hard data are lacking, researchers have attributed recent increases in unit nonresponse to secular trends in the US, especially in urbanized areas [14, 21], whereby the population has become more mobile, resulting in people being available at their usual place of residence for shorter periods. In particular, the Council of American Survey Research Organizations (CASRO) [7] noted that is has become more difficult to contact women because of their increased participation in the labor force.

Kessler et al. [16] observed that there is an increased tendency for entire households to be away from home when interviewers call because of growing proportions of single-member households and dual-earning couples, increased commuting time, and an increase in evening activities outside the home. Moreover, people appear to be more protective of their more limited personal time at home, making them less receptive to requests to participate in a survey [6].

Two particular aspects of survey methodology may effect unit nonresponse. The first is the mode of initiating contact and collecting data. In general, unit nonresponse is lowest for face-to-face interviews, slightly larger for telephone interviews, and largest for mail (postal) surveys [1, 10]. The second is the burden participation places on respondents. For example, nonresponse tends to be larger for longer interviews/questionnaires, less salient survey topics, and sensitive or threatening survey topics. Additionally, nonresponse tends to vary among population subgroups. In particular, nonresponse tends to be greater among young adults, the elderly, the poor, persons with little education, and persons with certain disabilities such as impaired hearing [1, 14]. Other factors such as characteristics of the survey sponsor and the time of year also may effect unit nonresponse [11, 15]. Finally, the sampling process may contribute to unit nonresponse in instances when problems with the **sampling frame** prevent the researcher from contacting a sample member, such as if the frame contains erroneous or out-of-date address and/or telephone information.

A special case of unit nonresponse is when a respondent begins to participate in a survey but fails to complete an interview/questionnaire. These *partial completes* occur more often in face-to-face and telephone interview surveys than in mail surveys, when the respondent "breaks off" from an interview before it is completed (for example, because of a lack of time or a refusal to answer any more questions). Further complicating the matter is that some so-called *complete* questionnaires may include unanswered questions (item nonresponse). Thus, although partial completes usually are categorized as a component of unit nonresponse, they are strongly related to item nonresponse. The larger the item nonresponse for a case, the more likely it is to be considered a partial complete. Because there is no standard definition of a partial complete, the researcher's decision about whether to include a case in the analysis or to count it as a unit nonresponse usually is guided by two factors: the proportion of the questions answered by the respondent; and whether responses are obtained for questionnaire items measuring key variables for the analysis.

Item Nonresponse

Item nonresponse occurs when an eligible sample member participates in a survey but does not provide a usable response to one or more of the survey questionnaire items (questions). For a survey where the questionnaire contains some items that do not apply to all respondents, item nonresponse refers only to questions that apply to a particular respondent. Thus, the data record for some respondents regarded as "completes", because their overall participation in a survey was acceptable, may include **missing values** for some variables.

Item nonresponse is a function of two components. The first is failure to obtain an answer to a question, which may occur for several reasons. The most obvious one is when a question is presented but the respondent refuses to answer or is unable to provide the requested information. However, item "nonresponse" also includes cases where a question is not presented because a respondent or interviewer does not follow instructions correctly (e.g. does not understand that a question applies to the respondent, or records only one response to a multiple response question). Also, an interviewer may fail to present a question because the interviewer feels uncomfortable with the subject matter, or the interviewer may fail to encourage a respondent properly to answer a difficult question. Finally, item "nonresponse" also includes cases where a respondent "answers" a question but the response is not recorded (by the respondent or interviewer).

The second component of item nonresponse is failure to obtain a usable response to a question. In a mail survey or other type of self-administered questionnaire study a respondent's handwriting may be illegible, or a respondent may record two response choices where only one response is requested and/or logical (e.g. a respondent may record both "yes" and "no" for a dichotomous question). For an interview survey an interviewer may fail to probe properly to obtain a clear and complete response. This is particularly a problem for open-ended questions or "other – specify" type responses (*see* **Interviewing**

Techniques). Similar problems of lack of clarity and completeness are even more likely to occur in a mail survey or other type of self-administered questionnaire study.

A special issue for item nonresponse is when a respondent answers "don't know" to a question, which may or may not be a case of missing data. In some cases, a respondent may use this response as a convenient and polite way to refuse to answer. In others, it may indicate that a respondent is unable to answer because of inability to retrieve the necessary information (e.g. from memory or records), the respondent has no opinion about the subject of a question, or because none of the response choices is appropriate. In still other cases, "don't know" may indicate that the respondent has no knowledge about the subject. For example, "I don't know" may be a valid response to a question such as "Whom would you call if a member of your household needs emergency medical treatment?". Before data collection begins, the researcher should anticipate the possibility that a respondent may answer "don't know" to virtually any question and decide whether to treat it as a nonresponse or as a valid response to be included in the analysis. This decision, even for identical questions, may vary from one survey to another depending on the purposes of the study and how the data will be interpreted. In general, the issue is whether it is reasonable to expect respondents to have adequate knowledge (e.g. about their age) or hold an opinion (e.g. about their health status) so as to be able to answer a question, as opposed to a situation where "don't know" may be a relevant substantive response (e.g. indicating a lack of knowledge about a health service).

Nonresponse Error and its Potential Impact

Nonresponse error is a function of two components: the magnitude of nonresponse (i.e. nonresponse rate); and the extent to which nonrespondents systematically differ from respondents. Concern about survey nonresponse has been driven mainly by the increase in unit nonresponse rates, probably because unit nonresponse results in fewer cases in the analysis, reducing statistical power and the precision of survey estimates. All things being equal, a higher rather than lower survey response rate is preferred because a higher rate indicates that unit nonresponse is lower. Moreover, when unit nonresponse is low, total nonresponse probably is low (for most variables) because item nonresponse is fairly low in most well-conducted surveys [8]. In general, the lower the total nonresponse rate the less concern about whether the participation of nonrespondents would have changed the survey estimates, simply because there are relatively few nonrespondents.

However, a high response rate does not mean that nonresponse error necessarily is trivial. For example, nonrespondents may differ substantially from respondents in terms of relatively rare characteristics [16] or if most nonrespondents are concentrated within one or two sample strata or population subgroups [23]. Also, nonresponse error may be substantial even when nonresponse is low if the factors causing nonresponse are associated strongly with important variables in the study. Ironically, as Kessler et al. [16] have observed, sometimes the techniques used to increase response rates and reduce nonresponse (e.g. special types of contact strategies or monetary incentives) can *increase* nonresponse error if they are more effective among some population subgroups than others, and if an underrepresented subgroup differs strongly from the others in terms of key study variables.

Nonresponse often is correlated with important demographic or other background characteristics (e.g. education in mail surveys, or health status in interview surveys) [6, 8]. But even if nonrespondents are similar to respondents on those characteristics, they may differ in terms of other important variables, such as attitudes and behaviors [23]. For example, persons who engage in risky behaviors may be less willing than others to participate in a survey about the epidemiology of the human immunodeficiency virus [22]. Therefore, the key nonresponse issue is *nonobservation* **bias**, whereby the absence of nonrespondents from the analysis causes one or more survey estimates to be consistently lower or higher than their population parameter (true value). This may substantially change the univariate and/or **multivariate distributions** of the survey data and result in erroneous interpretations of a study's findings.

The methods for taking account of sampling error in survey estimates (e.g. **confidence intervals**) assume that nonresponse error (as well as other nonsampling error) is zero. As nonresponse error increases, the model on which the computation and

interpretation of **inferential** statistics are founded becomes less appropriate. Therefore, the investment in carefully designing and selecting a large, random sample of a target population to minimize sampling error may be subverted by a substantial nonresponse bias. In most cases when the potential for nonresponse bias is relatively large, a researcher should consider using data collection strategies that have been shown to obtain high response rates. Also, the researcher may consider employing various strategies for obtaining information about the nonrespondents that can be used to take nonresponse error into account in the analysis. These decisions about survey design must balance available resources with a study's objectives. For example, an exploratory study may be able to tolerate larger amounts of both nonsampling and sampling errors than a study that is intended to provide a rigorous test of an important hypothesis. In most cases a researcher will invest more resources in the latter type of study.

Measuring and Reporting Nonresponse

Unfortunately, no model exists for taking nonresponse error into account in a way similar to that for assessing statistical inferences based on **probability theory**. Moreover, in most cases it is very difficult or impossible to estimate nonresponse error because reliable knowledge, independent of the survey, about the population regarding the variables measured in the survey usually is not available. Tests for nonresponse bias usually are limited to comparisons of nonrespondents with respondents, or respondents with the target population, in terms of aggregate characteristics such as socioeconomic status and other demographic variables based on data from existing sources (sometimes with questionable reliability) such as the US Census, institutional records (e.g. from schools or clinics), information that may be available from the sampling frame (e.g. residential location), or interviewer observation (e.g. type of dwelling unit) [9, 14]. Such comparisons may be useful if the criterion variables are strongly associated with the main variables of interest in the survey. Although it rarely is possible to make comparisons directly regarding the main variables of interest, researchers sometimes try to approximate this by comparing early respondents with late respondents [18] or by conducting brief follow-up interviews with subsamples of nonrespondents [8, 14, 16].

In addition to using techniques that tend to reduce nonresponse, researchers also sometimes apply various *post hoc* adjustments to improve the representativeness of survey estimates. These include techniques such as **poststratification** and a variety of strategies to impute values for missing questionnaire items (*see* **Missing Data**; **Multiple Imputation Methods**).

Survey researchers usually report a *response rate* (or a similar rate that may go by another name, such as cooperation rate or completion rate) rather than a *nonresponse rate*. It seems reasonable to regard a survey nonresponse rate as the complement of its response rate, obtained by subtracting the percentage response rate from 100% [14]. Unfortunately, despite attempts to encourage survey researchers to adopt a standard definition of response rate, there is no universally accepted definition for either a survey response rate or nonresponse rate [5, 7, 14, 17, 20, 24]. Thus, there is considerable confusion in comparing the quality of survey data across studies (meta-analysis), time periods, and survey methods.

The prevailing concept appears to be that a survey response rate should reflect the degree to which a survey succeeds in obtaining the cooperation of all potential respondents in the sample. Accordingly, the response rate may be calculated as the proportion of sample members known or estimated to be eligible for participation in the survey, from whom a complete/usable set of data is obtained. While there appears to be general agreement about the numerator for the response rate calculation (i.e. complete/usable cases), there is substantial variation in specifying the denominator, especially regarding the definition and estimation of eligible sample members [19, 20]. Moreover, the factors that effect eligibility vary with the sampling design and data collection procedures. In particular, this issue becomes quite complex in surveys using methods such as **random-digit dialing** telephone interviews, in which it is difficult and sometimes impossible to determine the eligibility of a substantial proportion of the initial sample.

The best guidance for computing response rates is available from the American Association for Public Opinion Research (AAPOR) both in the form of a report [2] and a Response Rate Calculator spreadsheet that may be downloaded free at the AAPOR web site [3]. However, until standard definitions of response and nonresponse rates are adopted, it is recommended that survey reports state how the response rate (or

similar term) is calculated, including the definition of eligible sample members [8, 17, 24].

Response rates rarely are reported for individual items. Item nonresponse usually is indicated by reporting the number of cases and/or **degrees of freedom** when presenting results in the text and/or tables of a survey report. However, when appropriate, it is recommended that an item nonresponse rate should be calculated as the proportion of respondents from whom a usable response was not obtained to a questionnaire item, from among the number of respondents who were eligible to answer that item.

Finally, because nonresponse error is not necessarily a direct function of the response rate, and because the definition and calculation of response rate are not consistent, it is not possible unequivocally to specify acceptable levels of survey response/nonresponse. However, some gross guidelines are that a survey with a response rate lower than 50% is very likely to contain a substantial nonresponse error [4, 10, 12]. A response rate greater than 75% generally may be regarded as good to excellent [10, 13]. However, it is strongly recommended that virtually any response rate should be compared with the response rate for other surveys addressing similar topics, dealing with similar populations, and using similar methods. Also, the study's goals should be considered: the more at stake in terms of the study's findings, the less the tolerance for nonresponse error.

References

[1] Aday, L.A. (1989). *Designing and Conducting Health Surveys*. Jossey-Bass, San Francisco.

[2] American Association for Public Opinion Research (AAPOR) (1998). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for RDD Telephone Surveys. Available at `http://www.aapor.org/ethics/stddef.html`

[3] American Association for Public Opinion Research (AAPOR) Response Rate Calculator. Available at `http://www.aapor.org/ethics/outcome=calculator.html`

[4] Babbie, E.R. (1973). *Survey Research Methods*. Wadsworth, Belmont.

[5] Bailar, B. & Lanphier, C.M. (1978). *Development of Survey Methods to Assess Survey Practices*. American Statistical Association, Washington.

[6] Bradburn, N.M. (1992). Presidential address: a response to the nonresponse problem, *Public Opinion Quarterly* **56**, 391–397.

[7] Council of American Survey Research Organizations (1982). On the Definition of Response Rates. A Special Report of the CASRO Task Force on Completion Rates. CASRO, Port Jefferson, New York, `http://www.casro.org/resprates.cfm`

[8] Czaja, R. & Blair, J. (1996). Designing Surveys: A Guide to Decisions and Procedures. Pine Forge Press, Thousand Oaks.

[9] de Vaus, D.A. (1986). *Surveys in Social Research*. George Allen & Unwin, London.

[10] Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. Wiley, New York.

[11] Edwards, P., et al.(2002). Increasing response rates to postal questionnaries: systematic review, *British Medical Journal* **324**((),18) 1183–1191.

[12] Erdös, P.L. (1983). *Professional Mail Surveys*. Robert E. Krieger, Malabar.

[13] Fowler, F.J. Jr. (1988). *Survey Research Methods*. Sage, Beverly Hills.

[14] Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

[15] Heberlein, T.A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature, *American Sociological Review* **43**, 447–462.

[16] Kessler, R.C., Little, R.J.A. & Groves, R.M. (1995). Advances in strategies for minimizing and adjusting for survey nonresponse, *Epidemiologic Reviews* **17**, 192–204.

[17] Kviz, F.J. (1977). Toward a standard definition of response rate, *Public Opinion Quarterly* **41**, 265–267.

[18] Lin, I.-F. & Schaeffer, N.C. (1995). Using survey participants to estimate the impact of nonparticipation, *Public Opinion Quarterly* **59**, 236–258.

[19] Slattery, M.L., Edwards, S.L., Caan, B.J., Kerber, R.A. & Potter, J.D. (1995). Response rates among control subjects in case-control studies, *Annals of Epidemiology* **5**, 245–249.

[20] Spaeth, M.A. (1992). Response rates at academic survey research organizations, *Survey Research* **23**, 18–20.

[21] Steeh, C.G. (1981). Trends in nonresponse rates, 1952–1979, *Public Opinion Quarterly* **45**, 40–57.

[22] Tourangeau, R. and Smith, T.W. (1996). Asking sensitive questions: the impact of data collection mode, question format, and question content, *Public Opinion Quarterly* **60**, 275–304.

[23] Warwick, D.P. & Lininger, C.A. (1975). *The Sample Survey: Theory and Practice*. McGraw-Hill, New York.

[24] Wiseman, F. & McDonald, P. (1980). *Toward the Development of Industry Standards for Response and Nonresponse Rates*. Marketing Science Institute, Cambridge, Mass.

F.J. KVIZ

Nonsampling Errors

It has become conventional to partition the total survey error into components representing sampling and nonsampling errors. Sampling error arises from the sampling process itself, i.e. from the fact that we are making **inferences** from observations on a randomly chosen subset of units, rather than observing the whole population. Nonsampling errors include all the errors not attributable to this incomplete enumeration. Every step in the survey process is a potential source of nonsampling error, from imperfections in the initial specification and listing of the **target population**, through failure to obtain complete information from all units drawn in the sample (*see* **Nonresponse**) or to obtain correct information from the units that we do contact, to errors in recording and managing the data after the survey has been completed (*see* **Data Management and Coordination**).

Sampling error is relatively easy to deal with, at least in principle. We can reduce its effect by increasing the sample size or by clever choice of design and estimator (*see* **Estimation**). Moreover, we can estimate its size internally from the sample measurements themselves. In contrast, nonsampling errors often increase as we increase the sample size or the complexity of the sampling procedure and, although special surveys can be designed to get information on some components, it is difficult to measure the size of most components without external information of some sort. Unfortunately, the nonsampling component of the total error is likely to be at least as large as the sampling component in a well-designed survey. Since the impact of this component of total survey error is not captured by conventional formulas for the **standard error**, published estimates of survey error almost always underestimate the true state of affairs.

In the following sections we look at some specific sources of nonsampling error, with special reference to health surveys (*see* **Surveys, Health and Morbidity**), under three general headings: coverage errors (frame errors and nonresponse); **measurement errors** (question and format effects, respondent errors, interviewer effects); and processing errors. The choice of survey mode can have a substantial impact on all these components. In health research, this usually involves a choice among personal interviews, telephone interviews or mailed questionnaires. Some useful advice on the relationship between the survey mode and data quality for a variety of health outcomes is given by Van der Zouwen et al. [32], Siemiatydi [30], and Sibbald et al. [28]. Once the mode has been chosen, most methods aimed at reducing the nonsampling errors involve more resources being spent on preparation, pre-testing and piloting, training and supervision, and processing. These methods tend to be expensive and, with a limited budget, mean that the sample size will need to be reduced.

Coverage Errors

Coverage errors arise when the population from which the sample is really drawn differs from the target population. Two major sources of such errors are deficiencies in the **sampling frame** or listing from which the sample of units is drawn, and a failure to elicit responses from every unit that is drawn in the sample.

Frame Errors

A key requirement in the early stages of planning for any survey is the development of a frame, i.e. a list of units from which the sample will be drawn. In a telephone survey the frame will consist of a list of phone numbers, in a mail survey it will be a list of addresses, while in a personal interview survey it might be a list of households, area sampling units, hospitals or physicians' practices. Except in the very simplest situations, the population defined by the frame is likely to differ from the target population whose characteristics we really want to measure. For example, in a telephone survey any member of the target population who is not accessible by telephone will be excluded [7]. Similarly, in any survey based on a register or list of patients held by a health facility, such as a practice, a certain proportion will have either moved address or left the facility entirely [25].

Frame error can take the form of overcoverage. This can occur when the frame contains units that do not belong to the target population, or when there are multiple or duplicate listings such that single population units are identified with more than one frame element. However, the most common type of frame error takes the form of undercoverage (or incomplete coverage), with some units in the target population omitted from the frame. The effect of

undercoverage depends both on the proportion of missing units and on the magnitude of the difference between the values of the missing units and those listed in the frame. For example, consider a simple **mean** or proportion, θ. If we use subscripts T, F, and NC to denote values for the target population, the frame, and the units not covered by the frame respectively, then we have

$$\theta_T = \theta_F + \pi_{NC}(\theta_{NC} - \theta_F),$$

where π_{NC} denotes the proportion of the target population units that are not covered by the frame. We see that the **bias** (i.e the difference between the target and frame population values) is the product of the proportion of undercoverage and the difference between the means of the units in the frame and the omitted units. Most health surveys are concerned with more complex quantities than means and proportions (**relative risks**, **regression** parameters, etc.). Here the effect cannot be expressed quite so simply, but the basic idea still applies; there is little bias from undercoverage if the proportion of units not covered by the frame is small or if parameter values for the omitted units are very similar to those of the frame units.

Unfortunately there is no way to detect the presence of undercoverage either from the frame or from the sample itself, and no simple way to overcome the problem completely. Some ways to help alleviate this and other frame problems are discussed by Lessler & Kalsbeek [19, pp. 80–102] and Groves [13, pp. 81–128].

Nonresponse

Even if we have a reasonably complete frame from which to draw our sample, we may not be able to elicit responses from every unit. People may not return mailed questionnaires, or they may be out when the interviewer calls. Some people may be unwilling or unable to respond even if they are contacted. Some units may not provide any information at all (*unit* nonresponse) while others may provide responses for some items but not others (*item* nonresponse). Nonresponse can be regarded as another aspect of undercoverage, and the effect is very similar to that for incomplete frames; the degree of bias depends both on the response rate and on the extent to which nonresponders differ from responders. The response bias will be small if the proportion of nonrespondents is small (i.e. a high response rate) or if there is little difference between responders and nonresponders. Differences between responders and nonresponders can be substantial in many health surveys. For example, readiness to respond may be influenced by recently experienced health events, which may engender greater interest in participating in a health survey [3], or by health status, with those having the symptoms under investigation more likely to respond [20, 31]. One difference between nonresponse and frame undercoverage is that we do at least know the proportion of nonrespondents so that the possibility of a problem is clearly signalled, even if we have no idea of its size. Perhaps for this reason, there has been more attention paid to nonresponse than to any other source of nonsampling error.

As with most nonsampling errors, prevention is usually the best form of cure. It is hard to do much to control differences between responders and nonresponders, but it may be possible to increase the response rate. The choice of survey mode can have a big impact on response rates. In general, response rates for postal surveys tend to be lower, but older people who feel threatened by face-to-face interviews with a stranger may respond well to a mail survey [15]. The design of questions (*see* **Questionnaire Design**) and the quality of interviewers (*see* **Interviewing Techniques**) can also affect the response rate [4]. Extra training and extensive piloting can improve things here. Once the survey is in progress, we can make vigorous attempts to contact initial nonrespondents. For example, we might get interviewers to call back several times if a person is not at home, or send several reminder letters with a mailed questionnaire (*see* **Call-backs and Mail-backs in Sample Surveys**). Providing incentives such as paying people to take part may also improve response rates in some circumstances (although this can accentuate differences between responders and nonresponders if, for example, low-income people are more likely to be attracted by the offer [27]). Most of these measures are costly and implementation will usually have to be at the expense of sample size. This tradeoff will be worthwhile if the extra responses are sufficiently different to alter the survey estimate.

Getting an indication of the size of the difference between responders and nonresponders is difficult. Direct subsampling of nonrespondents, although expensive, may be worthwhile in some circumstances.

If we have auxiliary information on all units listed in the frame, then we can calculate differences between the means of the auxiliary variables for respondents and nonrespondents. If these variables are **correlated** with the study variables, this will give some indication of the potential problems. For example, Andersen et al. [1] compared respondent reports of care received with medical record data and derived adjustments. Frequently, however, such auxiliary information is not available. Another approach is possible if we are prepared to assume that willingness to respond lies on a continuum of cooperation. Then we may get some idea of the likely magnitude of problems by looking at differences between estimates for early and late responders, or those who respond only after additional prompting [15]. A number of studies have shown that nonresponders are more like late responders than early responders [9, 29].

One way of getting direct information on nonrespondents which is particularly useful in health surveys is through the use of proxy (or surrogate) respondents. Questionnaires constructed using concrete items which require less interpretation by the proxy and a shorter range of possible responses are more likely to yield responses congruent with subject response. In health surveys, studies have shown that proxies are able to report accurately on areas of health and functioning, although they tend to rate patients as slightly more impaired than patients rate themselves [11, 22]. Agreement between subject and proxy tends to be lower for conditions that are not observable, relatively private and not likely to be discussed, such as mental conditions and general aches and pains [22]. The best agreement is achieved in subject–proxy pairs where the respondents live together; correlation is reduced as contact between subjects is reduced.

Finally, a whole range of statistical procedures have been proposed to mitigate the effects of unit nonresponse. These include **post-stratification** and weighting adjustments based on estimates of the probability of response. These estimates might be based on auxiliary information, for example, or on extra information collected from respondents. A common procedure uses data on how often each respondent has been available for interview in the past week. A good review of these procedures is given in Lessler & Kalsbeek [19, pp. 161–233]. The most common procedure of all, particularly for item nonresponse, is to impute the missing values from respondent data. Many different imputation procedures have been proposed and a good overview can be found in Kalton & Kasprzyk [16]. There can be problems making inferences, and particularly with estimating precision, if too many values are imputed. Rao [26] and the ensuing discussion give some idea of the problems (*see* **Missing Data**; **Missing Data Estimation, "Hot Deck" and "Cold Deck"**; **Multiple Imputation Methods**).

Measurement Errors

Measurement errors arise from complex interactions among the survey mode, the instrument (i.e. the questionnaire in most health surveys), the particular question, the respondent, and, in personal interview and telephone surveys, the interviewer. For convenience, we group common problems under three general headings, but most problems involve all of these components to some extent.

Question and Format Effects

It is obvious that asking the right questions is critical if we are to obtain good information about the quantities in which we are interested. Common sense tells us that questions should be clear and unambiguous and expressed in language that the respondent can understand. Unfortunately, the situation is much more complex than this. The survey mode (personal interview, telephone interview, mailed questionnaire) can have a big impact. For example, telephone respondents tend to indicate a more favorable health status than mail respondents [23]. Differences can be large; in a study reported in Moore [24], 44% of people interviewed personally answered "Yes" when asked if they favored contraceptives being made freely available to unmarried women, in contrast to 75% of those questioned by telephone or mail. Even for a given choice of mode, very subtle changes in the wording, context, format, and layout of a questionnaire can have a measurable effect on the survey response. The order in which questions are asked affects the way that people respond in all modes, but even seemingly inconsequential factors such as the placement of instructions and the color of print has been shown to affect the responses. **Questionnaire design** is a specialist subject with a huge literature of its own. A

good introduction can be found in Kalton & Schuman [17] which is essential, if chastening, reading for anyone planning a survey for the first time.

Respondent Errors

The respondent is the ultimate source of information. Even if the question is understood clearly, he or she must have access to the information that is sought and must be able (and willing) to access this information accurately. Accuracy of recall is related to respondent motivation, the degree of detail required, the significance of the event and the time elapsed since it occurred, and also to the nature of the topic. For example, illness in healthy subjects may be underreported because it is not of current concern to the respondent (*see* **Recall Bias**).

Many surveys ask about events that occur in a specific time-period, such as the number of visits to a doctor over the past year. This requires respondents to place events in time, and a common distortion is "telescoping", where an event is remembered as having happened more recently than was actually the case. Fortunately, this has the opposite effect to loss of recall and the two errors may partially offset each other. Surveys asking for sensitive or personal information (e.g. about diet, sexual activity or alcohol consumption) may engender a "social desirability bias" resulting from the wish of a respondent to convey a positive image in keeping with social norms and to avoid criticism. This can distort the measurement of the variable of interest significantly [14, 33].

Methods to reduce respondent measurement errors require some understanding of their causes. The literature on this topic is wide-ranging, and includes work in cognitive psychology on memory and judgment as well as work in social psychology, survey methodology, and other disciplines. A good introduction is given by Groves [13, Chapter 9]. The Survey Research Center at Michigan has conducted a long-term study aimed at improving the quality of reporting of health events. Some of the results are summarized by Cannell et al. [5]. Successful techniques tried by Cannell and his colleagues at Michigan include the use of instructions to respondents asking them to think carefully about their responses and emphasizing that accurate and complete answers are important, the use of feedback, and securing a formal agreement of respondent commitment. Some of their findings are surprising. For example, they found that longer questions sometimes gave an increase in the number of health events reported, suggesting that the common advice to "keep the questions short" might be better phrased as "keep the questions simple" (see [17]).

Interviewer Effects

In personal interview and telephone surveys, the interviewer introduces a further source of measurement error. Some interviewers may simply not adhere to the survey protocol. Most problems stem from the interaction between the interviewer and respondent. The effect is likely to vary according to the type of question, with attitude questions, questions requiring probing, fixed-alternative and forced-choice items, together with poorly worded and ambiguous questions, being particularly susceptible to interviewer variability. When questions are unclear and consistently require additional interviewer input, there is a greater likelihood that results may be influenced by the interviewer; different interviewers may interpret questions differently, or may rephrase questions in a directive manner [10]. "Acquiescence bias" arising from the disposition to answer "yes" (or, less commonly, "no") regardless of the question asked, may be more severe in interviews with respondents who are of low socioeconomic status, or belong to minority cultures, when the interviewer is perceived to be of higher status.

The impact of interviewer variability depends on several things. For simple means and proportions, the **variance** of the sample estimate in **simple random sampling** is inflated by a factor $1 + (n - 1)\rho_{\text{int}}$, where n is a weighted average of the interviewers' case-loads and ρ_{int} is the intra-interviewer correlation as defined by Kish [18]. This correlation is a scale-free measure of the size of the variability among interviewers. The effect is similar with more complex survey designs. The impact depends both on the interviewer variability, as measured by ρ_{int}, and on the size of the case-load. Even very small values of ρ_{int} can have a big impact on the precision of the estimate if the average case-load is large. Most medical and health surveys are interested in more complex issues such as making comparisons between subgroups, comparing **relative risks**, estimating **regression** coefficients, and so on. There is a general belief that the impact of interviewer variability is much less severe for more complex parameters.

The special case of comparisons between subgroups is examined by Davis & Scott [8]. They show that the impact depends on the distribution of the case-loads between the subgroups and on the interaction between the interviewers and the members of the subgroups. The effect will usually be smaller than for a single mean but can be almost as large if the case-loads are very unbalanced and the interviewer effect differs between subgroups.

Most suggestions for reducing interviewer effects involve putting effort into the initial selection of interviewers, and into their training and supervision. A quality control protocol for checking interviewing consistency using audio tapes of randomly selected interviews has been shown to reduce interviewer variability [10]. The number of interviewers involved in a survey is an important factor since small differences between interviewers may give rise to appreciable reductions in the precision of sample estimates if each interviewer has a large case-load. With a constant ρ_{int}, the impact of interviewer variance can be reduced by increasing the number of interviewers and so reducing the number of individuals responding to each interviewer. However, this will usually result in a more heterogeneous pool of interviewers, particularly in heath surveys, where the interviewer often needs special expertise, and in less intensive training and supervision, all of which will tend to increase ρ_{int}. Data quality can sometimes be improved by careful deployment of the interviewers. We have seen above that making sure that interviewers see respondents from all subgroups can improve the precision of subgroup comparisons. Matching interviewers to respondents, such as using an interviewer of the same gender for examining sexual behavior [6], can also sometimes be effective. For example, older white male interviewers gained more reports of substance abuse in a study by Johnson & Parsons [15].

Processing Errors

Once the respondents have answered the questions, the responses have to be coded, edited and entered in a machine-readable form. Supplemental editing will usually be needed to clean the data. Finally, the raw data will be manipulated into a form suitable for analysis. Missing values may be imputed at this stage. The processing stage is the least glamorous but often the most important step in the whole survey process. Errors can creep in at every step; we may find coding errors, transcription errors, and errors introduced by the editing. However, we have an opportunity to remedy some of the nonsampling errors introduced earlier in the survey operation. All large survey organizations have their own specialized editing procedures for detecting inconsistencies and unlikely responses. In some cases we may have to check back with the original respondent to get clarification of responses that do not pass the editing checks.

Coding and transcription errors can be minimized by cutting down the human component of the process as far as possible with the use of computer-assisted data entry techniques. **Computer assisted telephone interviewing** (CATI) is one of these techniques. Providing interviewers with laptop computers so that data entry and editing can be carried out at the time of the original interview as with CATI is another. Most of the effective methods to control human errors involve careful selection, training, and supervision of personnel, just as with other sources of nonsampling errors. A good survey of modern methods for process control in surveys is given by Lyberg et al. [23].

Further Reading

Good general surveys of the broad field of nonsampling errors can be found in the books by Groves [13] and Lessler & Kalsbeek [19] and in the collections edited by Biemer et al. [2] and Lyberg et al. [21].

References

[1] Andersen, R., Kasper, J. & Frankel, M.R. (1970). *Total Survey Error. Applications to Improve Health Surveys*. Jossey-Bass, San Francisco.

[2] Biemer, P.R., Groves, R.M., Groves, K.E.K., Lyberg, L.E., Mathiowetz, N.A. & Sudman, S. eds. (1991). *Measurement Errors in Surveys*. Wiley, New York.

[3] Brambilla, D.J. & McKinlay, S.M. (1987). The comparison of responses to mailed questionnaires and telephone interviews in a mixed mode health survey, *American Journal of Epidemiology* **126**, 962–971.

[4] Cannell, C.F., Miller, P.V. & Okesenberg, L. (1981). Research on interviewing techniques, *Sociological Methodology* **12**, 389–437.

[5] Cannell, C.F., Fowler, F.J., Kalton, G., Okesenberg, L. & Bischoping, K. (1989). New quantitaive techniques for

presenting survey questions, *Bulletin of the International Statistical Institute* **53**, 481–495.

[6] Catania, J.A., Binson, D., Canchola, J., Pollack, L.M., Hauck, W. & Coates, T.J. (1996). Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behaviour, *Public Opinion Quarterly* **60**, 345–375.

[7] Davis, P.B., Lay Yee, R., Chetwynd, J. & McMillan, N. (1993). The New Zealand Partner Relations Survey: methodological results of a national telephone survey, *Journal of Acquired Immune Deficiency Syndome* **7**, 1509–1516.

[8] Davis, P. & Scott, A. (1995). The effect of interviewer variance on domain comparisons, *Survey Methodology* **21**, 99–106.

[9] de Marco, R., Verlato, G., Zanolin, E., Bugiani, M. & Drane, J.W. (1994). Nonresponse bias in an EC respiratory health survey in Italy, *European Respiratory Journal* **7**, 2139–2145.

[10] Edwards, S., Slattery, M.L., Mori, M., Berry, T.D., Caan, B.J., Palmer, P. & Potter, J.D. (1994). Objective system for interviewer performance evaluation for use in epidemiological studies, *American Journal of Epidemiology* **140**, 1020–1028.

[11] Epstein, A.M., Hall, J.A., Tognetti, J., Son, L.H. & Conant, L. (1989). Using proxies to evaluate quality of life, *Medical Care* **27**, 291–298.

[12] Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

[13] Hebert, J.R., Clemow, L., Pbert, L., Ockene, I.S. & Ockene, J.K. (1995). Social desirability bias in diet self-report may compromise the validity of dietary intake measures, *International Journal of Epidemiology* **2**, 389–398.

[14] Hebert, J.R., Bravo, G., Korner-Bitensky, N. & Voyer, L. (1996). Refusal and information bias associated with postal questionnaires and face-to-face interviews in very elderly subjects, *Journal of Clinical Epidemiology* **49**, 373–381.

[15] Johnson, T.P. & Parsons, J.A. (1994). Interviewer effects on self-reported substance use among homeless persons, *Addictive Behaviours* **19**, 83–93.

[16] Kalton, G. & Kasprzyk, D. (1986). The treatment of missing survey data, *Survey Methodology* **12**, 1–16.

[17] Kalton, G. & Schuman, H. (1982). The effect of the question on survey response: a review, *Journal of the Royal Statistical Society, Series A* **145**, 42–73.

[18] Kish, L. (1962). Studies of interviewer variance for attitudinal items, *Journal of the American Statistical Association* **57**, 92–115.

[19] Lessler, J.T. & Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. Wiley, New York.

[20] Locker, D. & Grusher, M. (1988). Response trends and non-response bias in a mail survey of oral and facial pain, *Journal of Public Health* **48**, 20–25.

[21] Lyberg, L.E., Biemer, P.P., Collins, M., de Leeuw, E., Dippo, C., Scwarz, N. & Trewin, D. eds. (1997). *Survey Measurement and Process Quality*. Wiley, New York.

[22] Magaziner, J., Bassett, S.S., Hebel, J.R. & Gruber-Baldini, A. (1996). Use of proxies to measure health and functional status in epidemiological studies of community-dwelling women aged 65 years and older, *American Journal of Epidemiology* **143**, 283–292.

[23] McHorney, C.A., Kosinski, M. & Ware, J.E. (1994). Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey, *Medical Care* **32**, 551–567.

[24] Moore, D.S. (1985). *Statistics: Concepts and Controversies*. Freeman, San Francisco.

[25] Pope, D. & Croft, P. (1996). Surveys using general practice registers: who are the non-responders?, *Journal of Public Health Medicine* **18**, 6–12.

[26] Rao, J.N.K. (1996). On variance estimation with imputed data, *Journal of the American Statistical Association* **91**, 499–506.

[27] Schweitzer, M. & Asch, D.A. (1995). Timing payments to subjects of mail surveys: cost effectiveness and bias, *Journal of Clinical Epidemiology* **48**, 1325–1329.

[28] Sibbald, B., Addington-Hall, J., Brenneman, D. & Freeling, P. (1994). Telephone versus postal surveys of general practitioners: methodological considerations, *British Journal of General Practice* **44**, 297–300.

[29] Siemiatydi, J. & Campbell, S. (1984). Nonresponse bias and early versus all responders in mail and telephone surveys, *American Journal of Epidemiology* **120**, 291–301.

[30] Siemiatydi, J. (1979). A comparison of mail, telephone and home interview strategies for household health surveys, *American Journal of Public Health* **69**, 238–245.

[31] Tennant, A. & Badley, E.M. (1991). Investigating non-response bias in a survey of disablement in the community: implications for survey methodology, *Journal of Epidemiology and Community Health* **45**, 247–250.

[32] Van der Zouwen, J. & De Leeuw, E. (1990). The relationship between mode of administration and quality of data in survey research, *International Sociological Association Paper* 90S23659.

[33] Welte, J.W. & Russell, M. (1993). Influence of socially desirable responding in a study of stress and substance abuse, *Alcoholism, Clinical and Experimental Research* **17**, 758–761.

A. SCOTT & P. DAVIS

Normal Clinical Values, Design of a Study

Ideally, a study to determine normal ranges for a cross-section of a specified population should start by selecting a random (i.e. probability) sample of healthy individuals from that population (*see* **Probability Sampling**). Since participation in such a study requires informed consent, which many randomly selected individuals may decline, an attempt should be made to test whether those agreeing to participate are representative of the entire population. Unfortunately, these conditions have only rarely been met in normal range studies. There have been exceptions, however, and we will mention these later.

The key stumbling block is the word "healthy". If one were interested in a cross-section of the entire population (as defined, for example, by location, age, or gender), a reasonably straightforward plan would be to select individuals according to a probability sample from existing **census** records and then invite them to participate in a study to measure the variation in some physiological function or blood test. For example, the National Health and Nutrition Examination Surveys (NHANES) conducted by the **National Center for Health Statistics** were based on samples of the civilian noninstitutionalized US population following a highly stratified (*see* **Stratification**), **multistage sampling** design. More familiar perhaps to epidemiologists is the **Framingham Heart Study**. Begun in 1948, this was a prospective study of risk factors in myocardial heart disease. A list of town residents was stratified by family size and precinct of residence. A sample of two families was selected at random from every three successive families in each stratum. All persons aged 30–59 in each selected family were included in the study.

The problem becomes more complicated when one wants to restrict the selected sample of reference subjects to healthy persons. At some point in the selection process a screen must be applied to detect those who do not qualify as healthy. This requires the development of a medical history questionnaire (*see* **Questionnaire Design**) (which many people might refuse to fill out) and evaluation of each individual return. Nevertheless, probability sampling may be applied to obtain the initial sample of individuals. A good example is a study undertaken in 1974 by Munan et al. [9] of the University of Sherbrooke in Quebec province, Canada, to establish population-based normal values for a variety of blood constituents. A total of 900 households in the Eastern Townships of Quebec (just north of the Vermont–New Hampshire border) were chosen randomly from 75 census enumeration districts. Within these families, approximately 2400 persons agreed to participate and were interviewed through a standard questionnaire. This questionnaire covered certain exclusion criteria listed in the next section. On the basis of the responses to these criteria, slightly over 50% of these individuals were selected as reference subjects.

During the 1970s, many studies were undertaken to develop normal ranges from large samples of individuals. The guiding principle in most cases was to obtain data from individuals presumed a priori to be healthy, so that every measurement could be used in the calculations. Often, blood donors were assumed to fit this description, although we recognize now that paying blood donors can attract many who are not really healthy. In addition, their ages are younger, in general, than the population as a whole. Other sources of subjects included attendees at a well-person screening clinic, or an outpatient clinic for some disability that was not expected to influence the variables measured. For example, an orthopedic clinic might meet this specification for many blood constituents. In some studies, data were obtained from instruments set up in a booth at an amusement park or a fair that gathered large numbers of presumed healthy persons. Under these circumstances, only the most superficial criteria for judging health could be applied.

If the variable for which normal values are desired is an indicator of a specific disease, the general health status of an individual in the population is not of concern; rather, the question is whether or not a prospective reference subject is suffering from that particular disease. It is then more feasible to obtain a statistically **random sample**, provided one has available some highly sensitive diagnostic tests, apart from the variable of interest, to detect the presence of the disease. The study of biochemical tumor markers often fits this description. An example is prostatic specific antigen (PSA), now well known as an indicator of prostate cancer. In a recent study by Oesterling et al. [11], a sample of almost 4000

white males aged 40–79 years, without a history of prostate cancer, was selected at random from a population of over 100 000 in Olmsted County, Minnesota. Only slightly over 55% agreed to participate in a prostatic evaluation, so the potential **bias** of self-selection loomed (*see* **Selection Bias**). Comparison of the medical records of those who agreed to participate, and those who declined, did not show "significant differences". Of this slightly over 2000-man group, about 25% were selected at random for the evaluation, which included not only the PSA measurement but also two additional diagnostic tests for prostate cancer. From this sample, the authors determined normal ranges for PSA as a function of age.

It remains that the vast majority of published normal ranges for blood constituents are not derived from random samples but from "samples of convenience"; for example, hospital or laboratory employees, blood donors, medical students, or other groups assumed to be "healthy" (or, at least, functioning normally in their daily occupations) and convenient to the laboratory desiring to obtain these ranges. In some sample groups, the restriction to a certain age group will be obvious. The situation is worse for many physiological functions where an "average normal" value is repeated in many medical textbooks. A prime example is cardiac output, for which the figure of 5–6 l/min is often cited as representing the average 73 kg male.

Exclusion Criteria

In 1975, Alström et al. [1] of the Scandinavian Committee on Reference Values (SCRV) published an extensive, rather rigid, set of exclusion criteria, including upper limits for blood pressure, hematocrit, serum cholesterol and triglycerides, urine albumin, and glucose. In addition, the use of therapeutic drugs or drugs of abuse was barred, as well as a long list of diagnosed diseases. Immediately on publication, these criteria were applied to a randomly drawn sample of potential reference subjects in Kristianstad, Sweden. Results [2] showed that use of these criteria excluded all but 11% of the men and 24% of the women from further testing. Clearly, some loosening of the criteria was needed. The criteria employed by Munan et al. were simpler and worked better. Even here, however, about 50% of the randomly selected persons who agreed to participate were finally excluded from the study, and the differences, if any, between selected and original population were not explored. The exclusion criteria of Munan et al. were as follows:

1. No chronic disease reported at the time of interview.
2. No disease leading to confinement to bed during the 15 days preceding the interview.
3. No medication intake in the 48 hours preceding the interview, except for vitamins.
4. No alcohol intake in the 48 hours preceding the interview.
5. Did not smoke more than 50 cigarettes during the 48 hours preceding the interview.

During the 1980s, the Expert Panel on the Theory of Reference Values (EPTRV) of the International Federation of Clinical Chemistry (IFCC) published a series of recommended practices with respect to the determination of normal values and ranges. One of these recommendations [6] dealt with the selection of reference subjects. The EPTRV did not formally adopt the rigid criteria of the SCRV, although they suggested the 1975 list of criteria as a guide. The EPTRV specifically recommended the exclusion of individuals suffering from systemic disease and such pathophysiological disorders as renal failure, congestive heart disease, chronic respiratory disease, liver disease, malabsorption syndrome and nutritional anemia. Moreover, individuals on either therapeutic drugs or drugs of abuse were barred, as well as persons using oral contraceptives, alcohol, and tobacco. Possible further criteria for exclusion include pregnancy and recent surgery.

It appears that the problem of developing a set of exclusion criteria that will bar those in a state of ill health from contributing to a normal range study, while not excluding the great majority of potential reference subjects who agree to participate, has not yet been resolved satisfactorily. It would seem, however, that every randomly sampled individual who agrees to participate should be allowed to do so; that is, should complete a medical questionnaire and should be measured for the variables of interest. Then, blind to these data, a reasonable set of predetermined exclusion criteria should be applied to the information in the questionnaires. This would allow comparison of the distributions of measured values in the group of subjects acceptable under the criteria and in those excluded. From such comparisons, information may be obtained not only for determining normal

ranges but also for judging the effects of the exclusion criteria on the distribution of each variable.

Partitioning Criteria

The planning for a normal range study should include consideration of separate ranges for different subgroups of the population. A smaller number of reference subjects within a given subgroup will provide an **unbiased** and more precise range for future application to patients of that subgroup than would a larger number of subjects from many subgroups with clinically important differences in the variable measured. However, as discussed elsewhere, the fact that mean values of different subgroups are significantly different by the usual statistical tests is an insufficient reason for calculating separate normal ranges. The differences must have a known physiological basis or, at least, be large enough to be recognized as clinically important even if the medical world is not sure why they exist. Depending on the sample size in each subgroup, this will demand mean differences whose probability levels under the **null hypothesis** are far smaller than 5% or 1%.

Most commonly, demographic groups (gender, age, race) are the partitioning criteria that come to mind first. Since these separate subgroups are not equally represented in the population sampled (except, in most cases, for gender), different sampling rates must be applied to each subgroup to assure an adequate number of reference subjects for calculating a normal range of acceptable precision (see next section). For example, the criterion of age usually does not imply equally spaced age intervals across the life span. Rather, it refers to approximate age intervals that are known to be associated with important physiological changes. Thus, separate normal ranges may be desired for selected variables in women before and after menopause. This, then, becomes the partitioning criterion, not a specified age boundary. However, information on menstruation will not become available until the medical questionnaires are received from those women willing to participate in the study. It may be, therefore, that an additional random sample of older women will be needed to bolster the sample size in the postmenopausal group.

Other partitioning criteria that might be clinically significant for certain variables include blood group, ethnic background, geographic location, stage of pregnancy, and tobacco use. If one or another of these is thought to be important, the relevant questions must be included in the questionnaire. A recent document published by the National Committee for Clinical Laboratory Standards (NCCLS) [10] contains a sample questionnaire as a planning guide.

Partitioning, if warranted clinically, may improve the homogeneity of within-group data. But whether the measurements are eventually to be subdivided into separate groups or not, certain rules followed during the execution of the study will help greatly to achieve a homogeneous distribution of the results. Among these are conditions on the premeasurement behavior of the reference subjects. For example, they should not engage in strenuous exercise during 24 hours preceding the drawing of a blood specimen. Specimens should be obtained in a fasting state and should be drawn during the morning hours to eliminate variations due to **circadian** rhythms. It is obviously important that the same measurement process (e.g. analytical method) be used throughout the study, and that stable environmental conditions be maintained in the laboratory during this time.

Sample Size

Curiously, the EPTRV of the IFCC did not address the question of how many reference subjects should be sampled to provide the data for calculating a normal range (*see* **Sample Size Determination**). Until recently, this issue has emerged only sporadically in the clinical literature.

Sample size calculations assume that the data represent a random sample from the given population. As indicated above, we believe that a statistically random sample of healthy individuals can, and should, be drawn from the population, although this practice has not been followed in the great majority of studies from which published normal ranges have been derived. Beyond the implicit assumption of random sampling, recommended sample sizes depend on the statistical method used to estimate the normal range. Various parametric and **nonparametric methods** have been described elsewhere.

For a given sample size, the most precise estimating procedure becomes available when the reference values are **normally distributed** or can be transformed to a scale on which they are normal (*see* **Transformations**). In this case, the percentile

limits (*see* **Quantiles**) of the range are estimated from the sample **mean** $\overline{x}$ and **standard deviation** s. Assuming large samples (say, $n \geq 100$), the **variance** of each limit on the normal scale is given approximately by $(3/n)\sigma^2$. For comparison, if these limits had been estimated by the corresponding sample quantiles, their large-sample variances would be approximated by the formula $\text{var}Q_p = p(1-p)/nf_p^2$, where f_p is the normal ordinate at the pth percentile. For $p = 0.025$ or 0.975, $\text{var}Q_p = (7.13/n)\sigma^2$. In other words, on the normal scale, the variance of the quantile estimator of a 95% normal limit would be expected to be 2.4 times larger than that of the normal estimator. The weighted quantile estimate (Harrell–Davis formula), described elsewhere (*see* **Normal Values of Biological Characteristics**), is more precise than the simple quantile, but the reduction in variance is not great, averaging about 20%.

Let us assume that when a normalizing transform can be found, the normal estimators will be used to derive normal limits. Then, a general criterion is needed for obtaining a recommended sample size from which to estimate normal limits for a variable that is normally distributed on either the original or transformed scale. Until recently, this problem was not addressed. Instead, in the few papers considering the question of sample size for normal ranges, it was assumed that the great majority of distributions of clinically important variables were either not normal and could not be normalized satisfactorily through the usual transforming functions, or that the process of trying to find a successful transform and testing for normality was generally too difficult and expensive for most clinical investigators to undertake.

For example, Reed et al. [12] include a table of exact nonparametric 90% **confidence intervals** for upper and lower normal limits for sample sizes from 120 to 369. They recommend a minimum sample size of 120, since this is the smallest number for which an exact nonparametric 90% confidence interval can be calculated (i.e. for any smaller sample size, the lower 90% confidence bound for the lower normal limit is less than the smallest observation, while the upper 90% confidence bound for the upper normal limit is greater than the largest observation). More recently, Miller et al. [8] arrived at a rule for the minimum number of reference subjects based on **bootstrap**-like resampling from a highly skewed empirical distribution: creatine kinase in women. They found that the standard deviation of the quantile $Q_{0.975}$ declined approximately as the square root of sample size up to size 400, but thereafter the decline was less than expected. These authors concluded that samples of less than 200 subjects were clearly inadequate to define 95% normal limits for highly skewed distributions using the sample quantile estimator, but that 400 would probably be sufficient.

It was Linnet [7], however, who proposed a general criterion for sample size that can be applied conveniently to normally distributed variables. This criterion and examples of its application are given in Harris & Boyd [5]. Briefly, Linnet suggested that the width of the 90% or 95% confidence interval for the true normal limit (e.g. the 97.5th percentile of the sampled population) be set equal to a fixed proportion of the normal range; say, 0.1, 0.2, or 0.3. To illustrate the application of this criterion, suppose that we are concerned with a variable whose measurements form a skewed distribution that can be normalized by a square root transformation; that is, the Box–Cox (3) transform parameter $\lambda = 1/2$. Then the width of the 95% normal range, after backtransforming the normal estimators to the original scale, is given by

$$W_{\text{nr}} = (\overline{y} + 2.0s_y)^2 - (\overline{y} - 2.0s_y)^2 = 8\overline{y}s_y, \quad (1)$$

where $\overline{y}$ is the mean and s_y the standard deviation of the square roots.

Similarly, the width of the 90% confidence interval for the true 97.5th percentile will be the difference between the square of the upper limit of the confidence interval computed from the normal estimator minus the square of the lower limit. We noted earlier that the **large-sample** variance of the normal estimator is $(3/n)\sigma^2$. Substituting s_y for σ, the width of the confidence interval on the original scale may be written $W_{\text{ci}} = 11.40\overline{y}s_y(1 + 2.0C_y)/n^{1/2}$, where C_y is the coefficient of variation of the distribution of square roots. Then, the ratio, say R, of W_{ci} to W_{nr} becomes $R = (1/n^{1/2})(1.25 + 2.85C_y)$. Setting $R = 0.2$, say, we obtain an expression for the minimal required sample size n as a function of C_y. This is graphed in Figure 1 ($\lambda = 1/2$).

For $C_y = 0.2$, the sample size must exceed 100 to satisfy the condition on R. This number rises to 400 when $C_y = 1.0$. Also included in Figure 1 are similar graphs for $R = 0.2$ when the power transformation parameter λ is 1/3 or 1/4. Harris & Boyd [5, p. 72] present a graph for the **lognormal distribution**, but

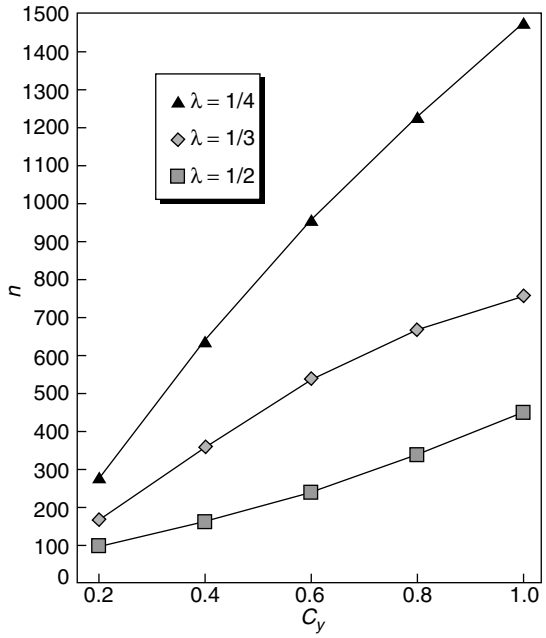

Figure 1 Sample size needed to estimate 2.5% or 97.5% normal limit with confidence interval 20% as wide as 95% normal range, using square root, cube root, or fourth root transformation to normalize the distribution; $C_y = CV$ on transformed scale

here n is a function of the coefficient of variation (*see* **Standard Deviation**) of the measured values on the original scale. They recommend that a pilot sample of 150–200 reference subjects be measured to determine a suitable power transform, if necessary, to achieve normality, and the appropriate coefficient of variation. Then, the foregoing results may be used to estimate the final number of subjects needed to attain the desired value of R. In circumstances where such numbers cannot be obtained, the investigators must settle for wider confidence limits relative to the overall normal range or for a smaller confidence level.

Even if investigators finally decide that no transform path will produce an acceptable normal distribution of their data, it is useful to estimate the Box–Cox power parameter because in most cases a power transform will greatly reduce **skewness** and **kurtosis**. Then, a baseline sample size can be estimated following Linnet's ratio. If a nonparametric estimation procedure is selected, general formulas for Linnet's criterion are not possible because confidence intervals and normal ranges are expressed only in terms of the rankings of the order statistics. In this case, we return to the definition of the normal range as a **tolerance interval** to contain at least a proportion p of the population with confidence $(1 - \alpha)$. To avoid an overly wide interval, Table A.30 in Hahn & Meeker [4] gives the nonparametric sample sizes needed to satisfy these conditions while guaranteeing a specified low probability of including a given proportion greater than p. For example, setting $p = 0.95, 1 - \alpha = 0.90$ and specifying that the probability of including a proportion greater than 0.98 should be no more than 0.1, the minimal sample size is 258.

Given the data from a study of this size, one may calculate Linnet's ratio on an ad hoc basis using the table of Reed et al. [12] to obtain 90% confidence intervals for the upper and lower normal limits. Then one may judge whether additional observations are needed to achieve a satisfactory R value. For sample sizes greater than 369, large-sample approximate confidence limits are given by the **order statistics** $x_{(r)}$ and $x_{(s)}$, where r is the largest integer less than or equal to $np + 1/2 - z_{\alpha/2}[np(1-p)]^{1/2}$, s is the smallest integer greater than or equal to $np + 1/2 + z\alpha/2[np(1-p)]^{1/2}$, and $z_{\alpha/2}$ is the **standard normal deviate** cutting off the upper 100 $(\alpha/2)$% of the normal curve.

References

[1] Alström, T., Gräsbeck, R., Hjelm, M. & Skandsen, S. (1975). Recommendations concerning the collection of reference values in clinical chemistry and activity report by the Committee on Reference Values of the Scandinavian Society for Clinical Chemistry and Clinical Physiology, *Scandinavian Journal of Clinical Laboratory Investigations* **35**, Supplement 144, 1–44.

[2] Berg, B., Nilsson, J.E., Solberg, H.E. & Tryding, N. (1981). Practical experience in the selection and preparation of reference individuals: empirical testing of the provisional Scandinavian recommendations in *Reference Values in Laboratory Medicine*, R. Gräsbeck & T. Alström, eds. Wiley, New York, pp. 55–64.

[3] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformation, *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

[4] Hahn, G.J. & Meeker, W.Q. (1991). *Statistical Intervals*. Wiley, New York.

[5] Harris, E.K. & Boyd, J.C. (1995). *Statistical Bases of Reference Values in Laboratory Medicine*. Marcel Dekker, New York.

[6] International Federation of Clinical Chemistry, Expert Panel on Theory of Reference Values (1984). Part 2.

Selection of individuals for the production of reference values, *Journal of Clinical Chemistry and Clinical Biochemistry* **22**, 203–208.

[7] Linnet, K. (1987). Two-stage transformation systems for normalization of reference distributions evaluated, *Clinical Chemistry* **33**, 381–386.

[8] Miller, W., Chinchilli, V.M., Greumer, H.-D. & Nance, W.E. (1984). Sampling from a skewed population distribution as exemplified by estimation of the creatine kinase upper reference limit, *Clinical Chemistry* **30**, 18–23.

[9] Munan, L., Kelly, A., PetitClerc, C. & Billon, B. (1978). *Atlas of Blood Data*. University of Sherbrooke, Quebec.

[10] National Committee for Clinical Laboratory Standards (NCCLS) (1995). *How to Define and Determine Reference Values in the Clinical Laboratory; Approved Guideline*. NCCLS Document C28-A, Villanova.

[11] Oesterling, J.E., Jacobsen, S.J., Chute, C.G. et al. (1993). Serum prostate-specific antigen in a community-based population of healthy men, *Journal of the American Medical Association* **270**, 860–864.

[12] Reed, A.H., Henry, R.J. & Mason, W.B. (1971). Influence of statistical method used on the resulting estimate of normal range, *Clinical Chemistry* **17**, 275–284.

(*See also* **Normal Clinical Values, Reference Intervals for**)

EUGENE K. HARRIS

Normal Clinical Values, Reference Intervals for

The idea of trying to detect or diagnose a disorder in an individual by comparing a relevant measurement from that individual with a range of values expected in a healthy or "normal" population is both ancient and intuitive. However, it was made systematic and scientific only in the 1950s when Wootton et al. [33] proposed the use of specific probability distributions to describe variation in biochemical measurements in blood samples from 100 active, working volunteers. Since then an enormous number of papers on "reference values", "normal ranges", "reference ranges" or "reference intervals" (all may be regarded as synonymous) have appeared in the medical, laboratory-based, and statistical literatures. A key controversy in the 1960s was whether hospital patients, a convenient resource for hospital biochemistry laboratories, were a suitable population on which to base reference ranges for application to the general public. This was proposed by Pryce [19], but was effectively dismissed by the work of Elveback [8] who showed major differences in the distributions of important serum variables between healthy working individuals and patients at the Mayo Clinic.

Leaving aside the vexed question of how to choose appropriate reference populations, the notion that an "abnormal" value (i.e. outside defined reference limits) is reliably indicative of disease is open to several criticisms. Since reference ranges are intended to include a fixed proportion of the population (see the section "Definitions" below), inevitably a proportion of "normal" individuals will lie outside the limits and will be falsely classified as "abnormal". Moreover, concentrations of biochemical analytes form a continuous scale, so useful diagnostic information may be lost by dichotomizing at fixed values. Since, nowadays, clinicians tend to order the batteries of tests that are routinely available from multichannel analyzers, there is an obvious problem of multiplicity (*see* **Multiple Comparisons**) when interpreting the results. More fundamentally, for effective discrimination between normal and diseased states the distributions of test values in diseased populations are required. Such distributions are rarely available, doubtless due to the high cost and difficulty of obtaining enough appropriate data.

Despite these and other criticisms, reference ranges remain a continuing feature of everyday clinical practice. In most cases they are used more as informal indicators that "something is wrong" in a particular physiological system when forming a clinical picture of a patient, rather than as formal decision rules in screening (*see* **Decision Analysis in Diagnosis and Treatment Choice**; **Screening, Overview**).

Harris & Boyd [10] is a comprehensive recent text on statistical aspects of reference values in laboratory medicine. It includes useful contextual information and several example data sets.

Definitions

A univariate $p\%$ reference interval is a pair of numbers (the *reference limits*) that enclose the central $p\%$ of a sample of observations (the set of *reference values*) obtained from a specified group of individuals (the *reference subjects*). Thus $(100 - p)/2\%$ of the values lie below the lower limit and the same proportion above the upper limit. The reference subjects are (supposedly) representative of some larger population, more or less well defined. A typical value of $p\%$ in clinical settings is 95%.

When extended to k-dimensional multivariate distributions, the reference interval becomes a reference region and is bounded by a suitable $(k - 1)$-dimensional surface. If the distribution of the reference values is **multivariate normal**, the surface is an ellipsoid.

Note that, by default, reference intervals are **"cross-sectional"**; that is, derived from samples with one observation per subject, though they may depend on factors such as age. Construction of intervals for changes, as in studies of human growth (*see* **Growth and Development**), or in individual patient monitoring, requires longitudinal data and generally more complex methods of statistical analysis (*see* **Longitudinal Data Analysis, Overview**; **Multilevel Models**), and is not considered here.

Design and Sample Size

The choice of reference subjects should be appropriate to the clinical use to which the resulting intervals will be put. In some cases, this may involve

measurements on special groups. For example, children with cerebral palsy are known to grow less rapidly than unaffected individuals, so age-specific reference intervals for, say, height should be constructed from a suitable population of children with the condition, not those from the general population. The selection of individuals is considered by the International Federation of Clinical Chemistry (IFCC) [12] in the clinical chemistry context and by Altman & Chitty [3] for studies of fetal growth.

Choice of sample size, n, is not necessarily straightforward, since no comparison of an outcome variable between groups (as in a **clinical trial**) is to be made. Since adequate information about the tails of the reference distribution is required, methods tend to focus on the precision of estimated reference limits. Stated simply, the width of a **confidence interval** for a centile (*see* **Quantiles**) depends on $\sqrt{n}$, so n may be determined by specifying this width, either in absolute (measurement) units or relative to the width of the reference interval. For example, a possible criterion (illustrated by Harris & Boyd [10, pp. 68–69]) is to take n large enough to ensure that the width of a 90% confidence interval for a reference limit is no more than 20% of the width of a 95% reference interval. One may use **nonparametric methods** or the properties of the **normal distribution** to calculate n. In the latter case, one must allow for estimation of a shape parameter if **transformation** towards normality is needed (see below). It is less clear how to use the approach in the age-specific case, since the width of the confidence interval will vary with age. In addition, the question of a suitable choice of ages arises (*see* **Sample Size Determination**).

Estimation

Univariate Case: Homogeneous Samples

One may determine reference intervals from a set of reference values either by nonparametric methods of quantile estimation or by fitting parametric densities and calculating intervals from the resulting parameters. Nonparametric estimation is distribution-free and therefore **"robust"**, but is usually inefficient in comparison with parametric modeling when a suitable distribution has been found. For example, if the underlying distribution is normal, the **variance** of the simplest nonparametric estimator of the 97.5th centile is 2.44 times that of the normal-based estimator, representing a relative efficiency of only 41%. The situation is less unfavorable with skewed distributions.

One may use the simple quantile estimator obtained by ordering the values and choosing the appropriate **order statistics** (with interpolation if needed). Harrell & Davis [9] proposed a method that gives efficiency gains of about 20% at the cost of additional computation. It is equivalent to applying **bootstrap** resampling to the simple estimator. Alternatively, kernel **density estimators** [27, 28] may be used.

As regards parametric estimation, the most popular method involves the normal distribution or a functional transformation towards normality. Transformation is often needed because the distribution of reference values is positively **skewed**. If we denote the measurement variable by Y, the most commonly used functions are the identity (Y is normal), log Y (Y is **lognormal**), $(Y^{\lambda} - 1)/\lambda$ (Box–Cox transformation) [5], $[\exp(\gamma Y) - 1]/\gamma$ (scaled exponential transformation, closely related to Box–Cox) [17], and $\log(Y + C)$ (origin-shifted logarithmic transformation, e.g. [23]). If we choose the value of the shape parameter (λ, γ, or C) correctly, the transformed measurements will have zero skewness. However, there is still no guarantee that they will be normal. Residual **kurtosis** (nonnormal tails) or other peculiarities may remain. Kurtosis may be removed by further transformation, such as the modulus function [15]. The IFCC [13] recommend use of the exponential transformation, followed by the John–Draper modulus transformation [15], if needed, to remove kurtosis. **Maximum likelihood** is the preferred method of estimating the parameters of these distributions. In most cases, the densities for Y corresponding to the power, origin-shifted logarithmic and exponential transformations may be shown by Taylor expansion to be very similar. They lead to reference intervals that are for practical purposes indistinguishable. (See the example below for an illustration of this assertion.)

Univariate Case: Subgroups and Covariates

The distributions of clinical variables such as blood pressure or serum biochemicals are usually age- or sex-specific. To retain their proper meaning, reference intervals should be constructed to take such **covariates** into account. Categoric variables present

little difficulty provided sample sizes are adequate, because the methods described above may be applied to subgroups. Modeling continuous variables such as age is more challenging and has stimulated statisticians to propose a variety of approaches to estimation. Wright & Royston [34] summarize the main techniques and compare them using several real data sets. Nonparametric techniques have included a window-based quantile estimator followed by polynomial smoothing, both over age and over normal equivalent deviates of selected quantiles [11, 18], and bivariate kernel density estimation [21]. Most current parametric approaches stem in essence from the methods of Cole [6] and Cole & Green [7]. Cole [6] applied the Box–Cox power transformation to remove skewness of Y in each of several contiguous age groups. In a second stage, he smoothed each of the parameter estimates across age by fitting polynomials (*see* **Polynomial Regression**). He called the resulting functions the L, M, and S curves; they represent the age-specific **power**, **median**, and coefficient of variation (*see* **Standard Deviation**) for Y. In a later refinement, which may be described as a **semiparametric** method, Cole & Green [7] used natural cubic **splines** to model the LMS curves by **penalized maximum likelihood**. This method is extremely flexible. Fully parametric methods are proposed by Royston [22], Altman [2], and Royston & Wright [25]. The last is based on the Manly [17] exponential transformation and uses fractional polynomials [24] to represent the age-specific parameter curves. No age grouping is needed.

Example

In a study [14] to compute age-specific reference intervals, the serum concentration of immunoglobulin-G (IgG) was measured in 298 children aged between six months and six years. A scatter plot of IgG concentrations against age is shown in Figure 1.

The continuous lines in Figure 1 are estimated 2.5th and 97.5th age-specific centile curves for IgG. In fact, the curves show the fitted values from three separate parametric models, based on the Box–Cox, shifted logarithmic, and scaled exponential transformations, respectively. (Further details of the modeling, which is based on fractional polynomials and maximum likelihood estimation, are given by Royston & Wright [25].) The results from the three models are essentially identical, and this is typical of such comparisons. The dashed lines show the fit from Cole & Green's natural cubic spline algorithm [7], using respectively 1, 5, and 3 equivalent **degrees of freedom** for the L, M, and S parameter curves. Although there are differences between the parametric and semiparametric curves, they are minor.

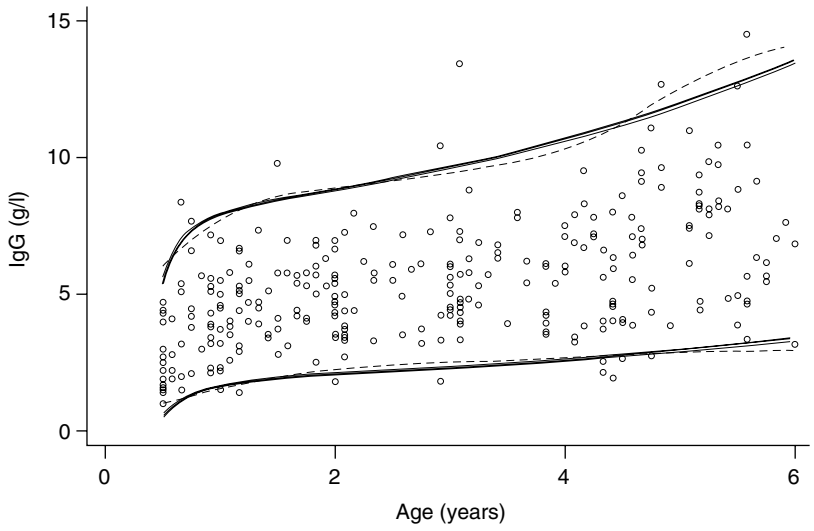

Figure 1 Serum IgG concentrations with estimated 95% age-specific reference intervals according to four models. Continuous lines: three parametric models based on data transformation; broken lines, semiparametric model based on natural cubic splines. See text for further details

Parametric or Nonparametric?

Ultimately, parametric methods are more rewarding because, *provided the model is approximately correct*, they usually offer greater precision and (in the age-specific case) smoother centile curves than do nonparametric methods. A second advantage is in dissemination; parametric models have concise equations that are easily published and, for example, incorporated into computer **software**. A disadvantage is that more initial effort may be required to find a suitable model, if one even exists. Since the validity of parametric models depends strongly on the correctness of their assumptions, such models must be subjected to rigorous **goodness-of-fit** checks (see below).

Multivariate Case

Methods of constructing reference regions for several (usually related) laboratory variables are described in detail by Albert & Harris [1]. Their approach is parametric and is based on the multivariate normal distribution, with preliminary univariate transformation of nonnormal components. Nonparametric **multivariate** density estimation does not seem to have been proposed in the context of clinical reference regions, nor has the age-specific case been discussed. In both cases, construction of reference regions is not straightforward. It is likely that since clinicians will wish to inspect the individual test results anyway, multivariate reference regions are little used in practice, despite the problem of multiplicity mentioned above.

Goodness of Fit and Outliers

Essentially, two methods of assessing goodness of fit of a $p\%$ reference interval have been proposed: direct (e.g. [11]) and model-based (e.g. [6]). The first involves counting the number of points that lie between the reference limits. Apart from **binomial** sampling variation, this number should be approximately $np/100$. The method is applicable in all cases except when the simple quantile estimator has been used to estimate the interval. Unless the sample size is enormous, however, it lacks **power**. The second method relies on the assumption that the data, possibly after transformation, have a known distribution, usually normal. Data transformed to standard normal (i.e. with mean 0 and variance 1) are known as SD scores or Z scores (*see* **Normal Scores**). The normality assumption may be checked by well-known graphical methods such as a histogram or a normal quantile–quantile (Q–Q) plot and by hypothesis tests such as the Shapiro–Francia [26] or Anderson–Darling [31] statistics.

Two major difficulties arise with the model-based method. First, if a shape parameter has been estimated, the distribution of a normality test statistic is affected, invariably in the direction of conservatism (not rejecting the **null hypothesis** often enough). Linnet [16] offered a corrected Anderson–Darling test when the data have been power transformed. Secondly, it is unclear how best to assess the fit of models for age-specific reference intervals. Departures from the model may be age dependent, and overall plots and tests of Z scores may be insensitive to them. Further research is needed in this area.

The detection of **outliers** is mainly relevant to the parametric approaches, as nonparametric methods tend to be robust. There is a huge statistical literature on outliers (a well-known text is [4]) and little may need adding in the present context. A particular problem that arises with power transformation of Y (namely, whether the transformation depends on just a few values) is considered by Tango [32].

Computation

When the reference values are homogeneous and apparently normally distributed, their sample mean and SD are all that are required to calculate any desired reference interval. For example, a $p\%$ interval is given by mean $\pm \mathrm{SD} \cdot \Phi^{-1}[(1 - p/100)/2]$. Essentially the same formula is used when Y has been transformed; the limits are calculated on the transformed scale and finally back-transformed. For the normal age-specific case, Altman [2] proposed a simple approximate solution using absolute **residuals** that does not require iteration. Some difficulty arises in finding maximum likelihood estimates of the parameters when Y has been transformed. **Estimation** is iterative as no closed-form solution is available, and, in general, it is necessary to use special purpose software (either stand-alone programs or routines written for use with particular statistical packages).

Software Sources

Solberg [29] has implemented the exponential/modulus transformation method recommended by the IFCC [13] for homogeneous samples in a stand-alone program (RefVal). It is available from Dr H. E. Solberg, Department of Clinical Chemistry, Rikshospitalet, N-0027 Oslo, Norway. Under the auspices of the **World Health Organization**, an MS-DOS-based package GROSTAT [20] has been developed that implements the nonparametric methods of Healy and colleagues [11, 18]. It is available from Dr H. Pan, Institute of Education, Bedford Way, London WC1H 0AL, UK. A FORTRAN program that implements the semiparametric method of Cole & Green [7] is available from Dr T. J. Cole, MRC Dunn Nutrition Unit, Downham's Way, Cambridge, UK. Software that implements the parametric methods of Royston & Wright [25] for use with the package Stata [30] may be obtained from Dr P. Royston, Department of Medical Statistics and Evaluation, Imperial College School of Medicine, Ducane Road, London W12 0NN, UK.

References

[1] Albert, A. & Harris, E.K. (1987). *Multivariate Interpretation of Clinical Laboratory Data*. Marcel Dekker, New York.

[2] Altman, D.G. (1993). Construction of age-related reference centiles with absolute residuals, *Statistics in Medicine* **12**, 917–924.

[3] Altman, D.G. & Chitty, L.S. (1994). Charts of fetal size: 1. methodology, *British Journal of Obstetrics and Gynaecology* **101**, 29–34.

[4] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Ed. Wiley, Chichester.

[5] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

[6] Cole, T.J. (1988). Fitting smoothed centile curves to reference data (with discussion), *Journal of the Royal Statistical Society, Series A* **151**, 385–418.

[7] Cole, T.J. & Green, P.J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood, *Statistics in Medicine* **11**, 1305–1319.

[8] Elveback, L.R. (1970). How high is high? A proposed alternative to the normal range, *Proceedings of the Mayo Clinic* **47**, 93–97.

[9] Harrell, F.E. & Davis, C.S. (1982). A new distribution-free quantile estimator, *Biometrika* **69**, 635–640.

[10] Harris, E.K. & Boyd, J.C. (1995). *Statistical Bases of Reference Values in Laboratory Medicine*. Marcel Dekker, New York.

[11] Healy, M.J.R., Rasbash, J. & Yang, M. (1988). Distribution-free estimation of age-related centiles, *Annals of Human Biology*, **15**, 17–22.

[12] International Federation of Clinical Chemistry (IFCC) (1984). Panel on Theory of Reference Values. The theory of reference values. Part 2. Selection of individuals for the production of reference values, *Journal of Clinical Chemistry and Clinical Biochemistry* **22**, 203–208.

[13] International Federation of Clinical Chemistry (IFCC) (1987). Panel on Theory of Reference Values. The theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits, *Journal of Clinical Chemistry and Clinical Biochemistry* **25**, 645–656.

[14] Isaacs, D., Altman, D.G., Tidmarsh, C.E., Valman, H.B. & Webster, A.D.B. (1983). Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges for IgG, IgA, IgM, *Journal of Clinical Pathology* **36**, 1193–1196.

[15] John, J.A. & Draper, N.R. (1980). An alternative family of transformations, *Applied Statistics* **29**, 190–197.

[16] Linnet, K. (1988). Testing normality of transformed data, *Applied Statistics* **37**, 180–186.

[17] Manly, B.F.J. (1976). Exponential data transformations, *Statistician* **25**, 37–42.

[18] Pan, H.Q., Goldstein, H. & Yang, Q. (1990). Nonparametric estimation of age-related centiles over wide age ranges, *Annals of Human Biology* **17**, 475–481.

[19] Pryce, J.D. (1960). The normal range, *Journal of the American Medical Association* **212**, 883–884.

[20] Rasbash, J., Pan, H. & Goldstein, H. (1991). *GROSTAT-A Program for Estimating Age Related Centiles using Piecewise Polynomials*. Institute of Education, London.

[21] Rossiter, J.E. (1991). Calculating centile curves using kernel density estimation methods with application to infant kidney lengths, *Statistics in Medicine* **10**, 1693–1701.

[22] Royston, P. (1991). Constructing time-specific reference ranges, *Statistics in Medicine* **10**, 675–690.

[23] Royston, P. (1992). Estimation, reference ranges and goodness of fit for the three-parameter lognormal distribution, *Statistics in Medicine* **11**, 897–912.

[24] Royston, P. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[25] Royston, P. & Wright, E.M. (1998). A method for estimating age-specific reference intervals ('normal ranges'), *Journal of the Royal Statistical Society, Series A* **161**, 79–101.

[26] Shapiro, S.S. & Francia, R.S. (1972). An approximate analysis of variance test for normality, *Journal of the American Statistical Association* **67**, 215–216.

[27] Sheather, S.J. & Marron, J.S. (1990). Kernel quantile estimators, *Journal of the American Statistical Association* **85**, 410–416.

[28] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, New York.
[29] Solberg, H.E. (1995). RefVal: a program implementing the recommendations of the International Federation of Clinical Chemistry on the statistical treatment of reference values, *Computer Methods and Programs in Biomedicine* **48**, 247–256.
[30] StataCorp (1996). *Stata Reference Manual, Version 5.0*. Stata Press, College Station.
[31] Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association* **69**, 730–737.
[32] Tango, T. (1986). Estimation of normal ranges in clinical laboratory data, *Statistics in Medicine* **5**, 335–346.
[33] Wootton, I.D.P., King, E.J. & Smith, J.M. (1951). The quantitative approach to hospital biochemistry, *British Medical Bulletin* **7**, 307–311.
[34] Wright, E.M. & Royston, P. (1997). A comparison of statistical methods for age-related reference intervals, *Journal of the Royal Statistical Society, Series A* **160**, 47–69.

(*See also* **Normal Clinical Values, Design of a Study**; **Normal Values of Biological Characteristics**)

PATRICK ROYSTON

Normal Distribution

This article describes the univariate normal distribution, with brief references also to the **bivariate normal distribution** and the **multivariate normal distribution**. For each of these, there are brief historical remarks, and discussions of distributional properties, **sampling distributions**, and related applications.

The Univariate Normal Distribution

Historical Remarks

The univariate normal distribution is probably *the* most important distribution in classical statistical theory and methods. This distribution has a "bell-shaped" continuous probability density function. In the history of statistics, it was first discovered by the German mathematician **Carl F. Gauss** in the early nineteenth century while studying certain problems in physics and astronomy. As a result, this distribution is also known as the *Gaussian distribution*.

Distributional Properties

A continuous univariate **random variable** X is said to follow a normal distribution with parameters μ and σ^2 if its probability density function $f(x)$ is of the form

$$f(x) = \left[\frac{1}{(2\pi)^{1/2}\sigma}\right] \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right],$$
$$-\infty < x, \mu < \infty, \text{ and } \sigma > 0;$$

in symbols, $X \sim \mathrm{N}(\mu, \sigma^2)$. It can be shown by elementary calculus that the mean and the variance of this distribution are, respectively, μ and σ^2.

A random variable Z is said to follow a *standard* normal distribution if $Z \sim \mathrm{N}(0, 1)$ (*see* **Standard Normal Deviate**). The distribution function of Z, $\Phi(z)$, is then given by

$$\Phi(z) = \int_{-\infty}^{z} \left[\frac{1}{(2\pi)^{1/2}}\right] \exp\left(\frac{-u^2}{2}\right) \, \mathrm{d}u.$$

Since the function $\exp(-u^2/2)$ is symmetric about 0, $\Phi(z)$ satisfies $\Phi(z) = 1 - \Phi(-z)$ for all z. The $(1-\alpha)$th **quantile** [or $100(1-\alpha)$th percentile] of a standard normal distribution, often denoted as z_α in most textbooks, is the quantity that satisfies $\Phi(z_\alpha) = 1 - \alpha, \alpha \in (0, 1)$. Since the function $\Phi(z)$ cannot be expressed in a closed form, tables for the numerical values of $\Phi(z)$ and z_α are needed. Such tables can be found in most statistics books.

The following theorem, which can be proved by calculus, shows how a normal distribution with mean μ and variance σ^2 is related to the standard normal distribution.

Theorem 1. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then the random variable $Z = (X - \mu)/\sigma$ is distributed according to the standard normal distribution.

As a simple application of Theorem 1, it follows that:

Fact 2. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then:

1. $\Pr[a < X \le b] = \Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)$ for all $a < b$;
2. The $(1-\alpha)$th quantile of the distribution of X is $\mu + z_\alpha \sigma$.

Thus, tables for the standard normal distribution can be used for any univariate normal distribution.

More detailed distribution properties of the univariate normal distribution can be found in Patel & Read [4] and other related sources.

Sampling Distributions

The **chi-square distribution, Student's *t* distribution**, and the ***F* distribution**, generally considered to be the cornerstones of classical statistical analysis, are closely related to the normal distribution. Specifically, let $X_1, X_2, \ldots, X_N$ be a random sample of size N from an $\mathrm{N}(\mu, \sigma^2)$ distribution; let

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and} \quad \mathbf{S}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

denote the sample mean and the sample variance, respectively. Then:

1. $\overline{X}$ has an $\mathrm{N}(\mu, \sigma^2/N)$ distribution and $(N-1)S^2/\sigma^2$ has a chi-square distribution with $N-1$ **degrees of freedom**. Furthermore, $\overline{X}$ and S^2 are independent.
2. $t = N^{1/2}(\overline{X} - \mu)/S$ has a Student's t distribution with $N-1$ degrees of freedom, where $S = \sqrt{S^2}$ is the sample standard deviation.

3. t^2 has an F distribution with degrees of freedom $(1, N-1)$.

Another important sampling distribution result related to the normal distribution is a fundamental theorem in probability theory, called the **central limit theorem**:

Theorem 3. Let $X_1, X_2, \ldots, X_N$ be a random sample of size N from any population with mean μ and finite variance σ^2. Then, for every fixed z,

$$\lim_{N\to\infty} \Pr[\{N^{1/2}(\overline{X}-\mu)/\sigma\} \le z] = \Phi(z).$$

This theorem provides an approximation for the distribution of $\overline{X}$ when N is large. In most applications,

$$\Pr[a < \overline{X} \le b] \doteq \Phi(N^{1/2}(b-\mu)/\sigma) - \Phi(N^{1/2}(a-\mu)/\sigma)$$

when $N \ge 30$.

Related Applications

In various applications of statistical **inference** problems – including, **estimation** of the parameters and **hypothesis testing** – the above sampling distribution results are applied. These include the following:

1. For inference on μ when σ^2 is known, the results for the distribution of $\overline{X}$ and Theorem 3 may be applied.
2. For inference on μ when σ^2 is unknown under the assumption of normality, Student's t distribution may be used.
3. When making statistical inference on σ^2 under the assumption of normality, the chi-square distribution may be used.
4. The normal distribution may be applied for inference on the difference of two normal means when their variances are known. Similarly, the Student's t distribution may be applied for the same purpose when the variances are unknown but equal; in the hypotheses-testing problem, this is known as the two-sample t test (*see* **Student's *t* Statistics**).
5. The F distribution may be applied for testing the equality of $k \ge 2$ normal means when the variances are assumed to be equal but unknown. This method is known as the one-way **analysis of variance** (ANOVA) method. When $k = 2$, it reduces to the two-sample t test as a special case.
6. Other results related to Theorem 3 are useful in **large-sample** inference problems. For example, it is known that under regularity conditions the **maximum likelihood** estimator has an asymptotically normal distribution, and that the asymptotic null distribution of $-2\ln\lambda$ is chi-square, where λ is the likelihood ratio function in a **likelihood ratio test**.

The Bivariate Normal Distribution

Historical Remarks

Studies of the bivariate normal distribution seem to begin in the middle of the nineteenth century, and moved forward dramatically when **Galton** published his work [3] on the applications of correlation analysis in genetics. As **Karl Pearson** noted in his 1920 *Biometrika* paper [6], "In 1885 Galton had completed the theory of bivariate normal correlation" but, because he "was very modest and throughout his life underrated his own mathematical powers, he did not at once write down the equation" of the bivariate normal density function. Consequently, it was Pearson himself who gave a definitive mathematical formulation of the bivariate normal distribution in his 1896 paper [5] on **regression** and heredity.

Distributional Properties

A two-dimensional random vector (X_1, X_2) is said to have a bivariate normal distribution if their joint density function $f(x_1, x_2)$ is of the form

$$f(x_1, x_2) = [2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}]^{-1} \exp\left[-\tfrac{1}{2}Q_2(x_1, x_2; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right], \quad -\infty < x_1, x_2 < \infty;$$

where

$$Q_2(x_1, x_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (x_1-\mu_1, x_2-\mu_2)\boldsymbol{\Sigma}^{-1} \times \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}$$

defines an ellipse centered at $(\mu_1, \mu_2) \equiv \boldsymbol{\mu}$ (which is the mean vector), the 2×2 matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

is the **covariance matrix**, and $\rho \in (-1, 1)$ is the **correlation** coefficient.

The marginal and conditional distributions of a bivariate normal random vector are univariate normal. For details, see the article on **bivariate normal distribution** and Tong [7, Section 2.1].

Sampling Distributions and Related Applications

The sampling distribution results involve the distributions of the sample mean vector $\overline{\mathbf{X}}$, the sample covariance matrix $\mathbf{S}$, the independence property of $\overline{\mathbf{X}}$ and $\mathbf{S}$, and the distribution of the sample correlation coefficient. Those results may be applied for estimation and hypotheses testing purposes. For details, see the article on **bivariate normal distribution**, Anderson [1, Section 2.3], and Tong [7, Sections 2.1 and 2.2].

The Multivariate Normal Distribution

Historical Remarks

The development of the multivariate normal distribution theory, which originated mainly from the studies of regression analysis and multiple and partial correlation analysis (*see* **Multiple Linear Regression**), was treated comprehensively for the first time by **Edgeworth** in his 1892 paper [2]. The development of the sampling distribution theory under the assumption of normality (such as Fisher's work on the distributions of sample correlation coefficients and **Hotelling's T^2 distribution**) then followed.

Distributional Properties

An n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is said to follow a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and covariance matrix $\boldsymbol{\Sigma}_{n\times n} = (\sigma_{ij})$, in symbols $\mathbf{X} \sim \mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its joint probability density function is of the form

$$f(\mathbf{x}) = [(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}]^{-1} \exp\left[-\tfrac{1}{2}Q_n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right], \quad \mathbf{x} \in \mathbb{R}^n,$$

where

$$Q_n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})'.$$

The marginal and conditional distributions of a multivariate normal random vector are also normal. For details, see the article on **Multivariate Normal Distribution**, Anderson [1, Sections 2.3, 2.4 and 2.5], and Tong [7, Section 3.3].

Sampling Distributions and Related Applications

The sampling distribution results also involve $\overline{\mathbf{X}}$, $\mathbf{S}$, and their independence property; in particular, the results are related to Hotelling's T^2 distribution and the **Wishart distribution** (generalizations of the Student's t distribution and chi-square distribution, respectively). There also exist distributional results on the sample regression equations and various types of sample correlation coefficients. Such results have been found useful for the purposes of prediction and correlation analysis. For details on sampling distributions, see the article on **Multivariate Normal Distribution**, Anderson [1, Chapters 4, 5, and 7], Tong [7, Sections 3.4 and 3.5], and other related sources. For related applications in inference, a classical reference is Anderson [1].

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Edgeworth, F.Y. (1892). Correlated averages, *Philosophical Magazine, Series 5* **34**, 190–204.

[3] Galton, F. (1888). Co-relations and their measurements, chiefly from anthropometric data, *Proceedings of the Royal Society of London* **45**, 135–145.

[4] Patel, J.K. & Read, C.B. (1982). *Handbook of the Normal Distribution*. Marcel Dekker, New York. Revised Ed. 1996.

[5] Pearson, K. (1896). Mathematical contributions to the theory of evolution – III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society of London, Series A* **187**, 253–318.

[6] Pearson, K. (1920). Notes on the history of correlation, *Biometrika* **13**, 25–45.

[7] Tong, Y.L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

(*See also* **Normal Scores**; **Normality, Tests of**)

Y.L. TONG

Normal Scores

Normal scores are the **expectations** of the **order statistics** of a sample from the standard **normal distribution**. They are widely used in plots (*see* **Normality, Tests of**) and tests to assess the **goodness of fit** of the normal distribution to data, and also in other procedures (*see* **Normality, Tests of**).

The key idea in assessing fit is that a sample from a continuous distribution will tend to have a shape characteristic of that distribution. This shape is given by the *order statistics* of the sample – that is, its ordered values – comparison of which with their expected values will be informative about distributional fit. Their most common use is in normal scores plots, as outlined below, but the same basic idea can be extended to quantitative tests of fit.

Another use of normal scores is in **transformations** to normality, when it is desired to apply a technique suitable for normal data to data that are visibly nonnormal.

Normal Scores Plots

A widely used procedure for assessing normality and screening data for **outliers** is a normal scores plot, in which the sample order statistics are plotted against normal scores. This is a *Q–Q plot* of the data and the normal scores, and is sometimes also called a *rankit plot*. It is also used to assess the fit of **linear regression** *models*, where a common assumption is that the **residuals** have approximate normal distributions.

The upper two rows of Figure 1 show normal scores plots for six simulated normal samples. In each case, the intercept and slope provide rough estimates of the mean and standard deviation of the sample. For example, the top left panel shows data from the N(0, 1) distribution, while the data in the top right panel have mean and standard deviation approximately −4 and 3. The three panels in the bottom row of the figure show (from left) data from a distribution skewed to the right, to the left, and a long-tailed distribution. In each case there is a systematic departure from the straight-line pattern seen in the top six panels, indicating various types of nonnormality. Outliers would show up as observations lying well away from the upper or lower tails of the data, as in the central panel in the middle row and the left and right panels in the bottom row – although in fact none of those observations are outliers.

Tests of Fit

As shown in Figure 1, purely graphical assessment of normality can be inconclusive, especially when sample sizes are small. Consequently, tests using normal scores have been proposed for use when it is critical to know whether data are normal. For example, *Shapiro–Wilk tests* are based on the correlation between the data and normal scores, with low values of the **correlation coefficient** indicating nonnormality. These and other tests of normality should be used in conjunction with a normal scores plot. For references and a fuller account, see D'Agostino [1].

Transformation

When data clearly do not have a normal distribution, but it is desired to use a procedure based on the assumption of normality, one possibility is to replace each original data value with its corresponding normal score and to perform the procedure on the transformed data.

As an example, suppose that it is intended to perform a test to compare the locations of two sets of data the shapes of which are similar, but that it is clear (perhaps from normal scores plots) that the two samples are not normal. More precisely, we have two random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, where the distributions of X and $Y - \theta$ are the same for some θ, and we wish to test the null hypothesis that $\theta = 0$. A test statistic based on normal scores is $T = \sum_{r=1}^{N} e(r, N) I_r$, where $N = n + m$, the normal score $e(r, N)$ is defined below, and

$$I_r = \begin{cases} 1, & \text{if the } r\text{th largest of the combined} \\ & \quad \text{sample } X_1, \ldots, X_n, \\ & \quad Y_1, \ldots, Y_m \text{ is an } X, \\ 0, & \text{otherwise.} \end{cases}$$

This is like a rank test (*see* **Nonparametric Methods**) but with the ranks replaced by normal scores.

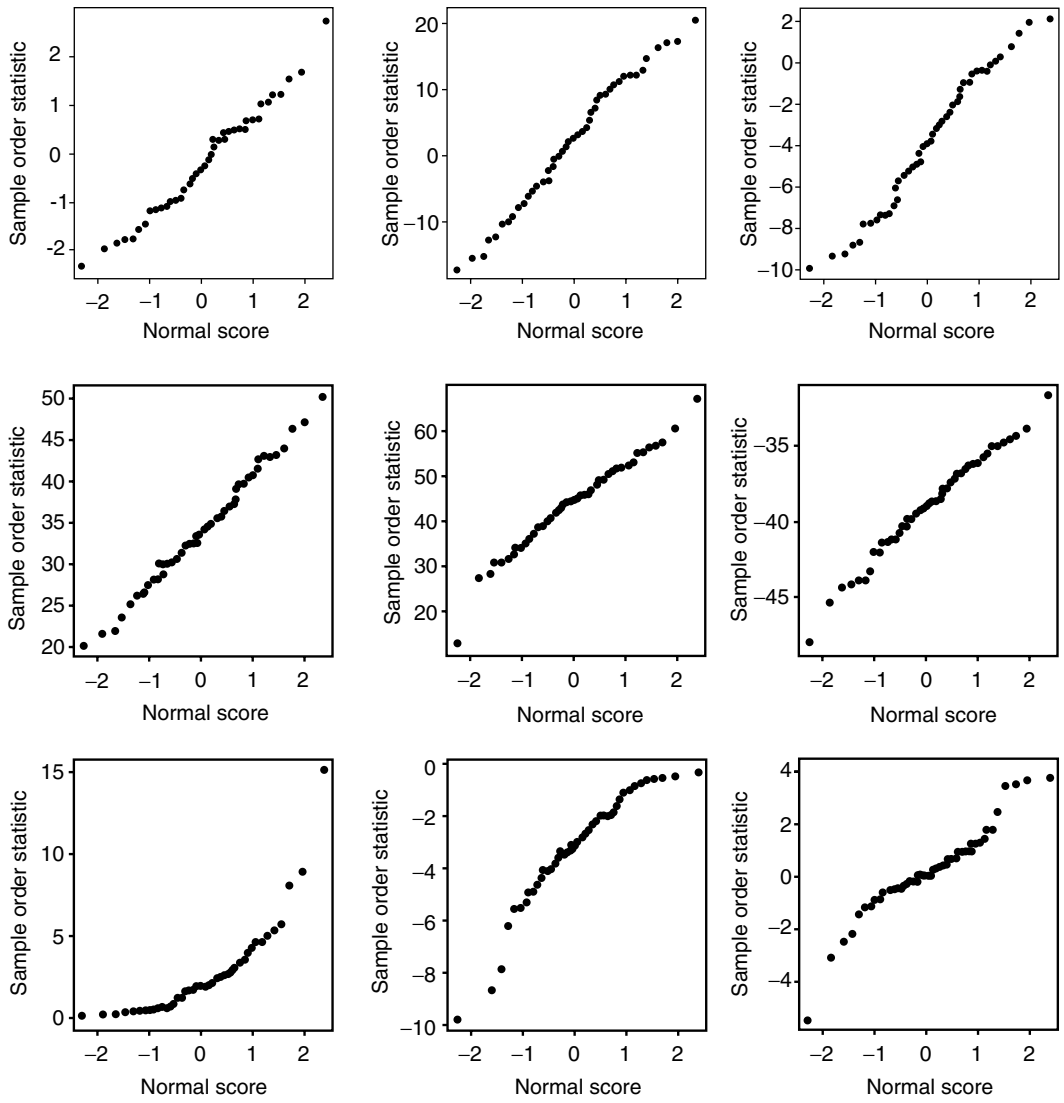

Figure 1 Normal scores plots for simulated samples of size 49. The top two rows show data from normal distributions. The bottom row shows data from a distribution skewed to the right (left), a distribution skewed to the left (center), and a long-tailed distribution (right)

Its advantage over rank tests is its high power when the distribution of X is close to normal. In practice, it may be more convenient to replace the exact normal scores by their approximation give in (1) below. For more details and related tests, including references to tabulated significance points for T, see Gibbons [2].

Computation

Suppose that $Z_1, \ldots, Z_n$ is a sample from the standard normal distribution, and let $Z_{(1)} \leq \cdots \leq Z_{(n)}$ denote the corresponding order statistics. Let $\Phi(z)$ and $\phi(z)$, respectively, denote the cumulative distribution function (cdf) and probability density

function pdf of the standard normal distribution. Then the exact value of the rth normal score is

$$\mathrm{E}(Z_{(r)}) = e(r,n) = \frac{n!}{(r-1)!(n-r)!} \times \int_{-\infty}^{\infty} z\Phi(z)^{r-1}[1-\Phi(z)]^{n-r}\phi(z)\,\mathrm{d}z.$$

This is awkward to work with, and an approximation that is adequate for most purposes is

$$e(r,n) = \Phi^{-1}\left(\frac{r-\frac{3}{8}}{n+\frac{1}{4}}\right), \qquad (1)$$

where $\Phi^{-1}(\cdot)$ is the inverse of the normal cdf. The approximation given in (1) is readily calculated in standard statistical **software** packages.

Royston [5] gives FORTRAN algorithms for $e(r,n)$ (available in machine-readable form from URL `http://lib.stat.cmu.edu/apstat/177`), while Harter [3] has tabulated their values for $n = 2(1)100(25)250(40)400$; see also Pearson & Hartley [4].

References

[1] D'Agostino, R.B. (1982). Departures from normality, tests for, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, Chichester, pp. 315–324.

[2] Gibbons, J.D. (1985). Normal scores tests, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson, eds. Wiley, Chichester, pp. 362–367.

[3] Harter, H.L. (1961). Expected values of normal order statistics, *Biometrika* **48**, 151–165.

[4] Pearson, E.S. & Hartley, H.O. (1976). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge University Press, Cambridge.

[5] Royston, J.P. (1982). Algorithm AS177: expected normal order statistics (exact and approximate), *Applied Statistics* **31**, 161–165.

A.C. DAVISON

Normal Values of Biological Characteristics

This article focuses on clinical biochemistry (e.g. blood glucose or serum cholesterol) and hematology measurements (e.g. hemoglobin or hematocrit) because this is the area where most progress has been made in the collection and statistical treatment of normal values. There are, of course, many nonchemical variables used to assess physiological status. A familiar example is brachial arterial pressure which is routinely measured before a clinical examination. Other examples will be noted in passing. The statistical methods described below should be applicable to all clinical (or, more generally, biological) variables that are measured on a continuous quantitative scale.

The "normal range" is commonly understood to be that interval which contains the measured values of a specified clinical variable in 95% of a population of healthy individuals. Since this range is estimated from a sample of individuals, we should rather define the estimate as a **tolerance interval** that has probability $100(1-\alpha)\%$ of including *at least* 95% of the population. In routine clinical practice the normal range is used as a **prediction** interval for the value in a healthy patient. However, when used for this purpose over and over again, as is always the case, it no longer satisfies the limited conditions of a statistical prediction interval (i.e. to predict the next observation or the **mean** or all of a fixed number of future observations). Therefore, we fall back on the definition of the normal range as a statistical tolerance interval. The interval is usually not estimated directly but only as the range of values between the estimated 2.5th and 97.5th percentiles of the population distribution. In many clinical situations, only one percentile (usually the upper) is important for diagnosis (*see* **Quantiles**).

The word "normal" is somewhat ambiguous, usually meaning "typical" but sometimes representing the "ideal". It also refers to a particular statistical distribution. Moreover, an analogous range (perhaps overlapping) could be defined for persons suffering from a particular disease. For these reasons, the term "reference range" has supplanted the normal range, at least in the clinical laboratory. However, since the discussion here will be confined to reference ranges in healthy individuals, and since the term "normal range" is still commonly used outside the clinical laboratory, we will continue this usage here.

A study to determine normal values requires, first of all, a set of criteria for judging whether a potential reference subject is really healthy. This and other aspects of the design of a study to establish normal clinical values are considered elsewhere (*see* **Normal Clinical Values, Design of a Study**). We assume here that a collection of measured values on a given variable has been obtained from a group of healthy subjects (one value per person). We describe first various statistical procedures (including tests of **outliers**) for deriving a normal range from this collection. Next, we examine the question of whether two or more collections of normal values obtained, say, from different sexes or races or geographic locations are sufficiently distinct to warrant separate normal ranges. We then review some recently published methods for estimating time-dependent normal ranges. We close with a brief discussion of multivariate normal indexes or regions.

Methods of Estimation

Parametric

Both parametric and **nonparametric** methods are routinely used to estimate the percentile limits of the normal range. Since most clinical variables are not **normally distributed**, the parametric method applies some mathematical **transformation** to the original values in the hope that they will conform to a normal distribution on the transformed scale. The log transform has often been used when the observed distribution shows positive skewness (*see* **Lognormal Distribution**). Occasionally, the square root or $\log(x + C)$ transforms are tried, estimating C by trial and error using the coefficients of **skewness** and **kurtosis** as guides. Unfortunately, the more general Box–Cox **power transform** [3] has only rarely been applied to normal values.

The Expert Panel on the Theory of Reference Values (EPTRV) of the International Federation of Clinical Chemistry (IFCC) has recommended [8] either the Box–Cox function or Manly's exponential transform [12] to remove skewness, followed by a second transform (e.g. the modulus function of

John & Draper [10]), if necessary, to remove any residual kurtosis. The Anderson–Darling **goodness-of-fit** test is recommended to test final agreement with a normal distribution. If the transform(s) have been successful, the normal range would be estimated as $\overline{x} \pm 1.96s$ (or $2.0s$) where $\overline{x}$ and s are the mean and **standard deviation** of the measurements on the transformed scale. The estimated 2.5th and 97.5th percentiles would be backtransformed to the original measurement scale for practical use.

Strictly speaking, the multiple of s should be taken from a table of normal tolerance factors for the given sample size, n, and level of confidence, $1 - \alpha$. For example, for $n = 100$ and $\alpha = 0.05$, the proper multiplier for at least 95% coverage would be 2.23. The factor 2.0 offers only about 50% confidence. However, assuming this value for simplicity, the large-sample **standard error** of the estimated upper or lower 95% normal limit on the normalized scale is given by $s(3/n)^{1/2}$, and **confidence limits** for the 2.5th and 97.5th percentiles of the population may be calculated as usual.

Nonparametric

When the parametric method achieves a normal distribution, it produces more precise estimates of the true normal limits than would a nonparametric method applied to the same data on the normalized scale (*see* **Standard Normal Deviate**). Clearly, however, some sophistication in statistics and computing is needed to pursue this approach. Moreover, there is no guarantee that a transform will be found to normalize the distribution. Therefore, since the collection of normal values has been carried out by clinicians or clinical chemists, often without statistical advice, the normal limits are more likely to be estimated by the simplest nonparametric method, namely ranking the data by order of magnitude and determining the 2.5th and 97.5th sample quantiles, interpolating between adjacent data points if necessary. Computing exact nonparametric confidence limits for the population percentiles is more complicated. Let $x_{(r)}$ and $x_{(s)}$ be the rth and sth **order statistics,** the sample values whose ranks are r and s, respectively. Then, $x_{(r)}$ and $x_{(s)}$ will be the bounds of a $100(1 - \alpha)\%$ confidence interval for the true percentile ζ_p if

$$\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \geq 1 - \alpha. \qquad (1)$$

The summation may be written as the difference between two **binomial distributions** (the first summing from r to n and the second from s to n) and may be calculated from the incomplete **beta distribution**. Reed et al. [14] have listed the ranks of the ordered observations that provide 90% confidence intervals for $\zeta_{0.025}$ and $\zeta_{0.975}$ from samples in the size range 120–369.

The drawback of the simple nonparametric method is that it places great weight on two or three sample quantiles in the tails of the observed distribution. A more recent nonparametric method for estimating the percentiles that define the normal range is the weighted quantile procedure proposed by Harrell & Davis [5]. These authors estimate ζ_p by a weighted average of all the observed order statistics, expressed as

$$\hat{\zeta}_p = \sum_{i=0}^{n} W_{n,i} x_{(i)}, \qquad (2)$$

where the weight function is given by the difference between two incomplete beta functions, and the sum of the weights equals unity. The Harrell–Davis estimate has been shown to be equivalent to a **bootstrapped** estimate, i.e. the average of a large number of resamples of the reference values. Bootstrapping is probably more familiar than incomplete beta functions to nonstatisticians; more importantly, it provides a standard error of the estimate for use in calculating confidence limits. With modern computing equipment, the Harrell–Davis formula may be speedily applied to many percentiles of the population. Results could then be smoothed by eye to provide a nonparametric guideline that would allow the clinician to estimate the exact percentile corresponding to any measured value in a new patient. Discussion and examples of the Harrell–Davis estimate are given in Harris & Boyd [6, Chapter 2].

The Harrell–Davis estimate produces a more precise result than the sample quantile, but the variance does not appear to decline by more than 20%. Therefore, it remains considerably less efficient than the parametric estimator, provided the normalizing transform(s) succeed. However, Linnet [11] has demonstrated that sampling variation in the estimated transformation parameter may widen the confidence interval around a normally estimated limit by 25%, increasing the variance of the normal estimator by 1.25^2, or a factor of 1.56. This implies that, in practice, the efficiency of the simple quantile on the

normal scale is not 41% (1/2.4) but 64% (1.56/2.4), while that of the weighted quantile estimate is actually about 77%. This conclusion has been supported by empirical results with distributions of blood chemistries [17]. Given the ease with which bootstrapping programs may be implemented today, the Harrell–Davis estimate should be more widely used, especially where the statistical expertise and the programs necessary to apply the parametric method may not be available.

Treatment of Outliers

Observations whose values differ substantially from the bulk of the observed distribution should be investigated further to discover, if possible, why they occurred. One or two outliers in a large sample of reference values (say, more than 100) may indicate sick persons whose conditions affected the variable being measured. If this is confirmed by further investigation, then these observations should be deleted before calculating normal ranges. A cluster of outlying observations is more likely to indicate a temporary dysfunction in the analytical system producing the measured values, or possibly the inadvertent use of a different analytical system than that used to obtain the rest of the values. In such cases, one might expect the outlying observations to have been obtained within a relatively short time period, although this would not necessarily be the case if the analytical problem were highly transient and unpredictable. Again, the circumstances surrounding these observations should be investigated, and if clear discrepancies with expected conditions are found, these observations, too, should be deleted from the reference sample.

If no specific reason(s) can be found to explain extreme results, and especially if they occur in the long tail of a skewed distribution, they should be retained. As Barnett & Lewis [2] and others have pointed out, such apparently aberrant values may be the result of chance selection from a skewed distribution representing all the observations. In this case, a normalizing transform may be tried, at least to separate the bulk of the distribution from the outliers. If this succeeds, then one may either apply an outlier test based on the assumption that the population distribution is normal (e.g. one of the Dixon ratio tests discussed by Barnett & Lewis), or estimate the mean and standard deviation of the (assumed normal) population by a **robust** method. Healy [7] has described a symmetric **trimming** procedure for this purpose. If normalizing transforms do not seem to work, then extreme observations should still be retained (given no reason for doubting them), and a nonparametric method applied to estimate the normal limits. If the total number of observations is small (50–100), the bootstrapped estimates are preferred over simple quantiles to minimize the effects of outliers and provide estimates of the standard errors of the estimated limits.

Separate Normal Ranges for Population Subgroups

During the 1960s and 1970s, many studies of blood constituents showed statistically significant differences in mean values of men and women and often significant trends with age. Most of these studies were based on large samples of presumed healthy individuals (e.g. blood donors or attenders at a well-person screening center). An example of a highly significant trend with age was shown in serum albumin by Wilding et al. [19]. Results are listed in Table 1.

The weighted **least-squares** slope is −0.054/decade, with standard error 0.000445, a highly significant decline with age. However, results like these rarely affect routine clinical practice. Laboratory reports to clinicians list normal ranges by sex for very few blood constituents and by age for none. This is partly due to the expense of obtaining separate normal ranges for the same variable, but more importantly, either no physiological basis has been found to explain the subgroup differences or the statistical difference is considered of no clinical significance. Small differences of no clinical

Table 1 Mean and standard deviation (g/100 ml) of serum albumin in men, by age [19]. Reproduced from [6] by permission of Marcel Dekker Inc.

Age decade (i)	Number (n_i)	Mean ($\overline{x}_i$)	Standard deviation (s_i)
20–29	96	4.41	0.20
30–39	721	4.35	0.21
40–49	1268	4.29	0.21
50–59	1112	4.24	0.22
60–69	415	4.19	0.22
70–79	105	4.13	0.30
Combined	3717	4.27	0.225

importance between subgroup means will inevitably become statistically significant if the sample size is large enough. In addition, it may be shown [6] that separate normal ranges are very little narrower than the combined range unless the difference between the means of subgroups is far greater than that required to achieve statistical significance. A clear example of this is seen in Table 1 by comparing the weighted average standard deviation in each decade with the standard deviation over all decades combined.

Many nonchemical variables show much stronger effects of aging. Examples are forced expiratory volume (FEV) and pulmonary diffusing capacity, both measures of lung function that decline sharply with increasing age after 20–29 years. In addition, the FEV is generally higher in men than in women.

Where separate normal ranges for men and women are recognized in practice, the physical or physiological basis for the difference is well known and accepted. With respect to blood tests, separate ranges for men and women are routinely reported for serum calcium, creatine and uric acid as well as hemoglobin and hematocrit. Cholesterol in various forms and triglycerides are known to increase significantly in healthy women after menopause; however, the normal range in cholesterol has been replaced by two decision points: 200 and 240 mg/dl. Wong et al. [20] have found substantial racial and gender differences in creatine kinase, while Sinton et al. [18] have presented evidence favoring separate reference ranges for alkaline phosphatase in pre and post-menopausal women.

Is there a statistical criterion that might help to answer the question of whether statistically significant subgroup differences should be recognized clinically, assuming that a physiological basis for such differences is known? Sinton et al.suggested that separate normal ranges not be considered unless the difference between subgroup means is at least 25% as large as the 95% normal range for the combined group. This is a rather stringent guideline. For example [6, p.79], with 400 subjects in each of two subgroups, the usual z-statistic for comparing the two means would have to exceed 16.3 before separate ranges would be recommended under this criterion. Separate ranges for calcium or high-density lipoproteins would not meet this criterion, nor would racial and gender differences in creatine kinase, although the data of Wong et al.make it clear that these separate categories for creatine kinase are justified.

Harris & Boyd [6] argue that real differences between two population subgroups imply that normal limits based on the combined group would cut off proportions much less than or much greater than the nominal 2.5% for either one subgroup or the other. From **simulation** studies of two normal subgroup distributions with 120 subjects in each, they find that a z-statistic of 5–8 would imply that in the subgroup with the smaller mean, less than 0.5% of the distribution would exceed the upper 95% normal limit based on the combined group, while only 0.5% of the distribution with the larger mean would be less than the lower normal limit of the combined group. These results assume the larger standard deviation to be within 30% of the smaller. (In most cases, subgroup standard deviations are little different despite statistically significant differences in mean values.) These authors recommend $z = 5$ as a minimal criterion for recognizing two separate subgroup normal ranges when 120 subjects have been sampled in each group. Under this criterion, the ratio of the difference between subgroups means to the width of the combined 95% normal range is 15.3%, considerably less than the ratio of 25% proposed by Sinton et al. [18] However, the critical value of z depends on the sample size according to the formula $z^* = 5(n/120)^{1/2}$. Thus, for $n = 400$, say, z^* would rise to 10.2.

When three or more subgroups are involved, it is not immediately clear how to apply this criterion. One possibility is to carry out **analysis of variance** followed by simultaneous comparison of paired means. Any pair of means whose difference is statistically significant should then be reassessed using the higher z^* values (*see* **Multiple Comparisons**).

Time-dependent Normal Values

In the preceding section we developed the idea that justifying separate normal ranges for demographic or age subgroups requires a more stringent statistical criterion than simply the result of a standard significance test of subgroup differences. However, clinicians and clinical chemists have long recognized that certain clinically important variables will change with the aging of a healthy individual. For example, as noted above, serum cholesterol in healthy women rises after menopause. Changes in selected biochemistries during infancy and childhood have also been reported, as have changes during pregnancy. Recently, Oesterling

et al. [13] have found that prostate-specific antigen in healthy men shows a clear increase with age beyond 50 years. All of these studies have been **cross-sectional** (single-sample) from healthy subjects at different ages or duration of pregnancy.

Age-specific percentile values for height and weight in normal children have long been familiar to parents and physicians (*see* **Anthropometry**; **Growth and Development**). Within the past decade, however, British statisticians have published a variety of new statistical methods for analyzing cross-sectional age- or other **covariate**-dependent measurements. The aim of these methods has been to produce mathematically smooth curves across the entire age span, in contrast to the earlier habit of arbitrarily defining age groups and joining with a straight line the estimated percentiles in adjacent age groups. Some of these techniques are too complicated to include here but have been described in Harris & Boyd [6].

Simpler methods proposed by Isaacs et al. [9], Royston [16], and Altman [1] should meet the objectives of smooth percentile curves without predefined age groups. Both Isaacs et al.and Royston seek a single transform function to convert the conditional distribution of the measurements at any age to normal form, assuming that they are not normally distributed on the original scale. Isaacs et al., working with immunoglobulin data, chose the logarithmic transform for IgA and IgM and the square root for IgG. Royston tried both log and $\log(x + C)$ functions on increasing values of serum cholesterol with age in women and on declining fetal triglycerides with gestational age.

The success of the transform function was assessed in the following way. Before and after applying the transform, a **polynomial** of appropriate degree was fitted to the scatter diagram of values vs. age (or other covariate). Then, the **residuals** from this curve were tested for normality. Royston [15] suggests the Shapiro–Wilk test for large samples. The polynomial estimates the mean value of the variable at any given age within the span of the data. Therefore, a normal distribution of the residuals implies that all conditional distributions by age are normal.

On the other hand, failure to achieve normality may arise from a nonconstant standard deviation of the residuals on the transformed scale across the age range. This question is critical to the establishment of other percentile curves (e.g. the 90th or 95th). Both Isaacs et al. [9] and Royston [15] suggested subdividing the residuals into a number of age brackets, testing the homogeneity of standard deviations across age, and, if needed, fitting a straight line or quadratic to represent the change of standard deviation with age. The drawback with this procedure is that it requires arbitrary division of the data into age groups. Altman [1] resolved this problem by noting, first, that if the standard deviation is a function of age, then age-standardized residuals (residuals divided by their standard deviations) should be used to test normality. But how are these to be determined without age-grouping? Assuming normality to begin with, Altman showed that if the absolute values of the unstandardized residuals were plotted against age, and a linear or quadratic curve, as suggested by the plot, were fitted, then multiplying the coefficients by $(\pi/2)^{1/2}$ produces an equation for the standard deviation of the original (signed) residuals as a function of age. Using this equation to derive age-standardized residuals, their conformity to a normal distribution may be tested. An example of the use of this technique is given in Harris & Boyd [6, p. 170].

From the estimated mean at any age, the age-dependent equation for the standard deviation, and evidence that the age-standardized residuals are normally distributed, upper or lower percentile curves can easily be calculated on the transformed scale and backtransformed for clinical use. The final test of validity is to calculate the proportion of observations falling outside each estimated percentile. These proportions should agree closely with expected values.

Multivariate Normal Indexes or Regions

Many diagnostic tests are grouped around specific organ systems, such as the liver (whose status is judged by measuring different enzymes) or the lung (various tests of breathing capacity) or the kidneys (serum creatinine, uric acid, urea nitrogen). One might expect that measurements on a cluster of organ-based variables would be **correlated**. Then, a single multivariate index capturing these correlations along with the normal means and variances would seem to be an attractive diagnostic tool. In fact, although multivariate normal indexes and regions have been discussed in the clinical literature for over 20 years, they are seldom, if ever, used in regular clinical practice.

A multivariate statistical index would usually be limited to a linear function of the measured variables,

in contrast to the common use of **nonlinear** models applied to measurements of physiological functions in clinical medicine. For example, in determining the condition of the cardiovascular system, several measured variables are often combined into a nonlinear physico-mathematical model. A simple example is the Fick principle for calculating cardiac output. This is defined as systemic flow (l/min) obtained by dividing the body's oxygen consumption (ml/min) by the difference between arterial and mixed venous (i.e. pulmonary arterial) oxygen contents. Textbooks in clinical medicine will frequently quote a single "average" normal value for each of these measured variables.

In one of the rare studies of the potential use of **multivariate normal** regions or indices, Durbridge [4] presented to clinicians over a 4-month period a "distance" index reflecting six hepatobiliary tests. Results for the individual tests were also presented in the laboratory report along with the obligatory normal ranges. Durbridge reported that about one-third of the physicians found that the index failed to shed further light on organ status, beyond that given by the individual tests. However, the majority of clinicians either thought the index was useful to them or said they needed more experience to reach a conclusion. Given the innate conservatism of medical practice, this result should not discourage further studies of the use of multivariate normal regions. Of course, values of individual variables and associated normal ranges will continue to be included in reports to the clinician. Moreover, anomalies between the multivariate index and the univariate normal ranges must be investigated and explained to the clinician. These include cases where the multivariate index is "significantly" high but none of the variables included in the index falls outside its normal range, or where one or more of the univariate results lie outside their normal ranges while the multivariate index is within bounds. Without clear interpretation of such results, the physician will find the conflicting information confusing and will probably reject any further use of the multivariate index.

The conventional multivariate index for k variables in the ith patient may be written as $D_k^2(\mathbf{x}_i) = (\mathbf{x}_i - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$, where x_i represents the patient's vector of results, $\overline{\mathbf{x}}$ is the mean vector over the (presumed nondiseased) reference group, and $\mathbf{S}$ is the sample variance-**covariance matrix** from this group. Then $D_k^2(\mathbf{x}_i)$ represents the "distance" between this patient's profile of test results and the mean profile for the reference group (*see* **Mahalanobis Distance**). The general form $D_k^2(\mathbf{x})$, with data coming solely from the reference group, may be used to define the boundary of a multivariate normal region, assuming that the reference profiles form a sample from a k-variate normal distribution. However, the multivariate index is more suitable for clinical practice because it is a single number for the ith patient just like an individual test result. In addition, the multivariate index computed for each member of the reference group can be used nonparametrically. That is, the 95th percentile (say) of its true distribution may be estimated by the nonparametric methods described earlier (e.g. the Harrell–Davis procedure). This estimate can then serve as a guideline for making a decision about a particular patient's index.

References

[1] Altman, D.G. (1993). Construction of age-related reference centiles using absolute residuals, *Statistics in Medicine* **12**, 917–924.

[2] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Ed. Wiley, New York.

[3] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

[4] Durbridge, T.C. (1983). Clinical acceptance of a multitest reference region for biochemical-panel results, *Clinical Chemistry* **29**, 1724–1726.

[5] Harrell, F.E. & Davis, C.E. (1982). A new distribution-free quantile estimator, *Biometrika* **69**, 635–640.

[6] Harris, E.K. & Boyd, J.C. (1995). *Statistical Bases of Reference Values in Laboratory Medicine*. Marcel Dekker, New York.

[7] Healy, M.J.R. (1979). Outliers in clinical chemistry quality-control schemes, *Clinical Chemistry* **25**, 675–677.

[8] International Federation of Clinical Chemistry, Expert Panel on Theory of Reference values (1987). Part 5. Statistical treatment of collected reference values. Determination of reference limits, *Journal of Clinical Chemistry and Clinical Biochemistry* **25**, 645–656.

[9] Isaacs, D., Altman, D.G., Tidmarsh, C.E., Valman, H.B. & Webster, A.D. (1983). Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges for IgG, IgA, IgM, *Journal of Clinical Pathology* **36**, 1193–1196.

[10] John, J.A. & Draper, N.R. (1980). An alternate family of transformations, *Applied Statistics* **29**, 190–197.

[11] Linnet, K. (1987). Two-stage transformation systems for normalization of reference distributions evaluated, *Clinical Chemistry* **33**, 381–386.

[12] Manly, B.F.J. (1976). Exponential data transformations, *Statistician* **25**, 37–42.

[13] Oesterling, J.E., Jacobsen, S.J., Chute, C.G., Guess, H.A., Girman, C.J., Panser, L.A. & Lieber, M.M. (1993). Serum prostate-specific antigen in a community-based population of healthy men, *Journal of the American Medical Association* **270**, 860–864.

[14] Reed, A.H., Henry, R.J. & Mason, W.B. (1971). Influence of statistical method used on the resulting estimate of normal range, *Clinical Chemistry* **17**, 275–284.

[15] Royston, J.P. (1982). An extension of Shapiro and Wilk's test for normality to large samples, *Applied Statistics* **31**, 115–124.

[16] Royston, P. (1991). Estimation, reference ranges and goodness of fit for the three-parameter log-normal distribution, *Statistics in Medicine* **10**, 675–690.

[17] Shultz, E.K., Willard, K.E., Rich, S.S., Connelly, D.P. & Critchfield, G.C. (1985). Improved reference-interval estimation, *Clinical Chemistry* **31**, 1974–1978.

[18] Sinton, T.J., Cowley, D.M. & Bryant, S.J. (1986). Reference intervals for calcium, phosphate, and alkaline phosphatase as derived on the basis of multichannel-analyzer profiles, *Clinical Chemistry* **32**, 76–79.

[19] Wilding, P., Rollason, J.G. & Robinson, D. (1972). Patterns of change for various biochemical constituents detected in well population screening, *Clinica Chimica Acta* **41**, 375–387.

[20] Wong, E.T., Cobb, C., Umehara, M.K., Wolff, G.A., Haywood, L.J., Greenberg, T. & Shaw, S.T., Jr (1983). Heterogeneity of serum creatine kinase activity among racial and gender groups of the population, *American Journal of Clinical Pathology* **79**, 582–586.

(*See also* **Normal Clinical Values, Reference Intervals for**)

EUGENE K. HARRIS

Normality, Tests of

The **normal distribution** has played a major role in statistical analysis ever since the work on the theory of errors by **Gauss** and **Laplace** in the early decades of the 1800s. It is used as a model for populations and random processes. It is the major distribution in asymptotic situations (*see* **Large-sample Theory**) because of the **central limit theorem**. Many statistical procedures are based on the assumption of normality. In response to this widespread use, numerous techniques for judging or testing for normality (or for departures from normality) have been developed.

Moment Tests: $\sqrt{b_1}$, and b_2

The field of tests for normality was initiated by **Karl Pearson** [18], who realized that deviations from normality could be described by the standardized third and fourth **moments** of a distribution, defined as

$$\sqrt{\beta}_1 = \frac{\mu_3}{\sigma^3} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\sigma^4}. \tag{1}$$

Here μ_i is the ith central moment for $i = 3, 4$ and σ^2 is the variance,

$$\mu_i = \mathrm{E}(X - \mu)^i, \quad \sigma^2 = \mathrm{E}(X - \mu)^2, \quad \text{and}$$
$$\mu = \mathrm{E}(X). \tag{2}$$

If a distribution is symmetric about its mean, then $\sqrt{\beta}_1 = 0$. Values different from zero indicate **skewness** and so nonnormality. β_2 characterizes **kurtosis** (or peakedness and tail thickness) of a distribution. For the normal distribution, $\beta_2 = 3$; other values indicate nonnormality (see Figure 1 for illustrations of varying $\sqrt{\beta}_1$ and β_2).

Tests of normality following from this are based on the sample third and fourth standardized moments, respectively, given as

$$\sqrt{b}_1 = \frac{m_3}{m_2^{3/2}} \quad \text{and} \quad b_2 = \frac{m_4}{m_2^2}, \tag{3}$$

where

$$m_k = \sum \frac{(X - \overline{X})^k}{n} \quad \text{and} \quad \overline{X} = \sum \frac{X}{n}. \tag{4}$$

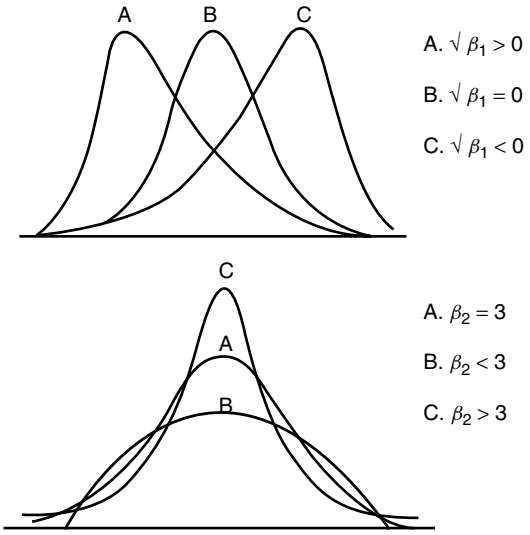

Figure 1 Distributions with varying $\sqrt{\beta}_1$ and β_2

Here, n is the sample size. Extensive tables of critical points and approximations for the sampling distributions of $\sqrt{b}_1$ and b_2 are readily available [2; 9, Chapter 9; 10; 11; 16; 17].

These moment statistics can be applied separately to tests of nonnormality due specifically to skewness or kurtosis. They can also be applied jointly for an omnibus test of nonnormality by employing various suggestions given by D'Agostino & Pearson [8]. One particular statistic for an omnibus test is given by

$$K^2 = X^2(\sqrt{b}_1) + X^2(b_2), \tag{5}$$

where $X(\sqrt{b}_1)$ and $X(b_2)$ are the **standard normal deviates** equivalent (in probability) to observing $\sqrt{b}_1$ and b_2 [8; 9, Chapter 9]. D'Agostino et al. [12] supply a simple computer macro program. Bowman & Shenton [4; 9, Chapter 7] present graphs that make possible the performance of the test for samples of size $n < 1000$. This test is available for basically all practical applications.

Normal Probability Plots

The moment tests are formal procedures for statistical inference (*see* **Hypothesis Testing**). A **graphical display** called normal probability plotting or probit plotting was developed as an informal technique for judging deviations from normality. The objective is

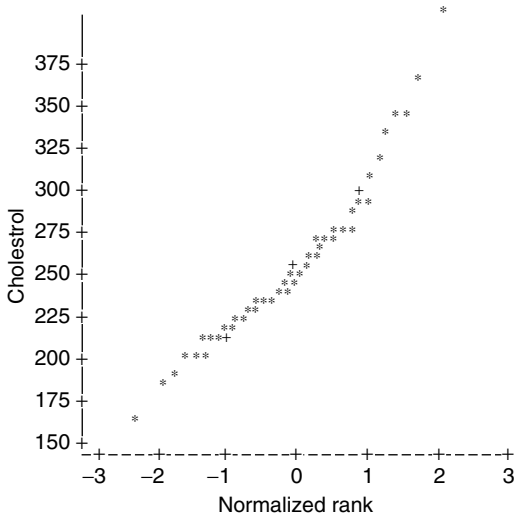

Figure 2 Normal probability plot of cholesterol data

to graph the data in such a way that, if the underlying population is normally distributed, then the graph will be a straight line. The deviation from linearity indicates the degree and type of nonnormality. Figure 2 shows such a plot for cholesterol data from the Framingham Heart Study. The ordered observations $X_{(1)} \leq \ldots \leq X_{(n)}$ are plotted on the vertical axis. The horizontal axis contains the inverse of the cumulative of the standard normal distribution

$$Z_{(i)} = \Phi^{-1}_{(q_i)}, \quad i = 1, \ldots, n. \tag{6}$$

There has been much discussion on the appropriate value of q_i (*see* **Normal Scores**). One standard choice is $q_i = (i - 0.5)/n$ for $i = 1, \ldots, n$. More details are given in D'Agostino & Stephens [9, Chapter 2] and by Brown & Hettmansperger [5].

Figure 3 contains a number of idealized normal probability plots with possible explanations for the deviation from a straight line. D'Agostino et al. [12] have emphasized the usefulness of the joint use of the moment tests with the normal probability plots in data analysis.

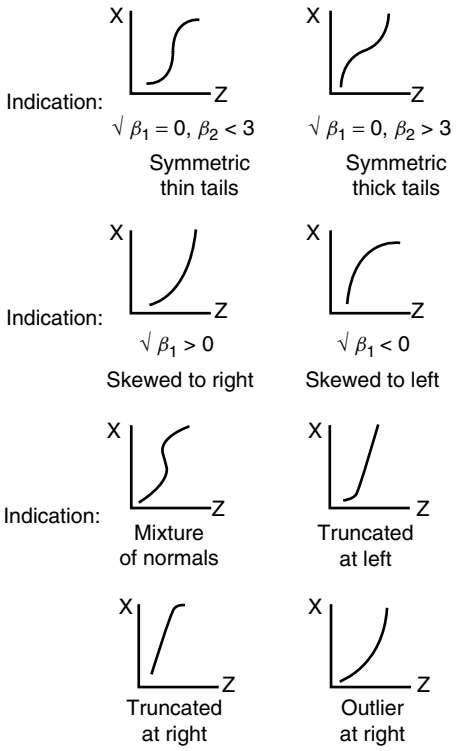

Figure 3 Normal probability plots with diagnosis nonnormality

Chi-square Test

The **chi-square test** developed by Karl Pearson in 1900 [19] can also be used for testing for normality. For this test the data are categorized into k categories. Each category has O_i observed values for $i = 1, \ldots, k$ with $n = \sum O_i$. Under the null hypothesis of normality, expected values e_i are computed. The chi-square statistic is then computed as

$$X^2 = \sum \frac{(O_i - e_i)^2}{e_i}. \tag{7}$$

For large samples this statistic has an approximate **chi-square distribution**. The appropriate **degrees of freedom** depend upon how the expected values are obtained and the choice and number of categories k. A full discussion of this test is given by David Moore in D'Agostino & Stephens [9, Chapter 3]. A nice feature of this test is that it can be employed for **censored** samples. The moment tests need complete samples.

Empirical Distribution Function (EDF) Tests

Another general procedure applicable for testing normality is the class of tests called the edf test. For these

tests the theoretical cumulative distribution function of the normal distribution, $F(x : \mu, \sigma)$, is contrasted with the empirical distribution function (edf) of the data, defined as

$$F_n(x) = \frac{\#(X \leq x)}{n}. \tag{8}$$

A famous test in this class is the Kolmogorov test [14] (*see* **Kolmogorov–Smirnov Test**), defined by the test statistic

$$D = \sup_x |F_n(x) - F(x, \mu, \sigma)|. \tag{9}$$

Large values of D indicate nonnormality. If μ and σ are known, then the original Kolmogorov test can be used. When they are not known they can be replaced by sample estimates. Stephens [23], employing **simulations**, developed adjusted critical values for D of (9) for this situation.

There are a large number of edf tests which involve various weighting of the deviations $F_n(x) - F(x, \mu, \sigma)$, with or without μ and σ known. Stephens [23] simulated critical values for four of these: the Cramér–von Mises W^2 test [6], the Kuiper V test [15], the Watson U^2 [24], and the Anderson–Darling A test [1], when μ and σ not known. Many of these tests are given in detail in D'Agostino & Stephens [9, Chapter 4]. Stephens has also extended these tests to censored samples [9, Chapter 4].

Transformation Tests

A variation of the edf tests involves first transforming the observations into independent observations free of unknown parameters and then applying an edf test. Quesenberry [9, Chapter 6] presents a general theory of these. These **transformation** procedures unfortunately require **randomization** of the data – a feature many find unattractive.

Regression and Correlation Techniques

Probably the most interesting innovative test of normality after the moment tests is the Shapiro & Wilk W test [22] and the extension of it by Shapiro & Francia [21]. The original W statistic can be viewed in a number of different ways. It is the ratio of the best linear estimator of σ to the sample standard deviation. It can also be viewed as the R^2 (square of the **correlation** coefficient) obtained from the normal probability plots. In this latter framework it arises in a **regression** and correlation context. Computationally, W is

$$W = \frac{\left(\sum w_{(i)} X_{(i)}\right)^2}{(n-1)S^2}, \tag{10}$$

where $w_{(i)}$ are the optimal weights for the **least squares** estimate of σ, and S^2 is the sample variance. Originally, the test required extensive computer work to obtain the $w_{(i)}$. D'Agostino [7] developed a simple approximation to W, called the D test. However, Shapiro & Francia produced more direct approximations of the $w_{(i)}$ which produced a true extension of W for large sample sizes [21]. Stephens [9, Chapter 5], and Royston [20] also generated transformations of W that provide good approximations to the null distribution.

Other Tests

There are a number of other tests for normality based on the sample **range**, spacings of observations, the properties of the **characteristic function, *U*-statistics**, etc. Some of these are outlined in D'Agostino & Stephens [9, Chapter 9].

Power Studies, Recommendations

Given the large number of tests for normality, the choice of which to use in practice is not easy. Fortunately, a large number of **power** studies have been performed. Many of these are summarized in D'Agostino & Stephens [9, Chapter 9]. Other, more modest ones, are recent [3, 13]. While it is not possible to give definitive answers, some general recommendations can be made:

1. Popular textbook tests such as the chi-square and Kolmogorov test have poor power in comparison to other tests and should not be used to test for normality.
2. $\sqrt{b_1}$ and b_2 have excellent power over a range of alternative distributions which deviate from normality with respect to skewness and kurtosis, respectively. These appear to be especially powerful as one-sided tests (*see* **Alternative Hypothesis**).

3. K^2 of (5) and other such tests are sensitive over a wide range of alternatives. They are *omnibus* tests.
4. The most powerful of the edf tests is the Anderson–Darling test as modified by Stephens. Its power is comparable to that of K^2.
5. The Shapiro–Wilk W test and its extensions are very sensitive omnibus tests. For many skewed distributions they are the most powerful.
6 While the D'Agostino D test is an omnibus test, it has the best power for the distributions with $\beta_2 > 3$.

Lastly, the usefulness of an investigation to judge the normality or nonnormality of data usually comes not from deciding to accept or reject normality, but rather from understanding the nature of the nonnormality and then performing appropriate statistical analyses given this knowledge. To achieve this knowledge one needs to look deeply at the data. The use of formal tests in addition to informal analyses, such as come from normal probability plots, is needed to achieve this understanding.

References

[1] Anderson, T.W. & Darling, D.A. (1954). A test of goodness-of-fit, *Journal of the American Statistical Association* **49**, 765–769.

[2] Anscombe, F.J. & Glynn, W.J. (1983). Distribution of the kurtosis statistic b_2 for normal statistics, *Biometrika* **70**, 227–234.

[3] Bera, A.N. & McKenzie, C.R. (1986). Tests for normality with stable alternatives, *Journal of Statistical Simulation and Computation* **19**, 37–52.

[4] Bowman, K.O. & Shenton, B.R. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2, *Biometrika* **62**, 243–250.

[5] Brown, B.M. & Hettmansperger, T.P. (1996). Normal scores, normal plots and tests for normality, *Journal of the American Statistical Association* **91**, 1668–1675.

[6] Cramér, H. (1928). On the composition of elementary errors. Second paper: statistical applications, *Skandinavisk Aktuarietidskrift* **11**, 141–180.

[7] D'Agostino, R.B. (1971). An omnibus test of normality for moderate and large size samples, *Biometrika* **58**, 341–348.

[8] D'Agostino, R.B. & Pearson, E.S. (1973). Testing for departures from normality. I. Fuller empirical results for the distribution of b_2 and b_1, *Biometrika* **60**, 613–622.

[9] D'Agostino, R.B. & Stephens, M.A. (1986). *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

[10] D'Agostino, R.B. & Tietjen, G.L. (1971). Simulation probability points for b_2 for small samples, *Biometrika* **58**, 669–672.

[11] D'Agostino, R.B. & Tietjen, G.L. (1973). Approaches to the null distribution of $\sqrt{b_1}$, *Biometrika* **60**, 169–173.

[12] D'Agostino, R.B., Belanger, A.J. & D'Agostino, R.B., Jr (1990). A suggestion for using powerful and informative tests of normality, *American Statistician* **44**, 316–321.

[13] Gan, F.F. & Koehler, K.J. (1990). Goodness-of-fit test based on P-P probability plots, *Technometrics* **32**, 289–303.

[14] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione, *Giornale dell' Institute Italiano degli Attuari* **4**, 83–91.

[15] Kuiper, N.H. (1960). Tests concerning random points on a circle, *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen, Series A* **63**, 38–47.

[16] Pearson, E.S. & Hartley, H.O. (1966). *Biometrika Tables for Statisticians*, Vol. 1. Cambridge University Press, Cambridge.

[17] Pearson, E.S. & Hartley, H.O. (1972). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge University Press, Cambridge.

[18] Pearson, K. (1895). Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society of London* **91**, 343.

[19] Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in a random sampling, *Philosophical Magazine, 5th Series* **50**, 157–175.

[20] Royston, J.P. (1982). An extension of the Shapiro and Wilk's W test for normality to large samples, *Applied Statistics* **31**, 161–165.

[21] Shapiro, S.S. & Francia, R.S. (1972). Approximate analysis of variance test for normality, *Journal of the American Statistical Association* **67**, 215–216.

[22] Shapiro, S.S. & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples), *Biometrika* **61**, 644–646.

[23] Stephens, M.A. (1974). EDF statistics for goodness-of-fit and some comparisons, *Journal of the American Statistical Association* **65**, 1597–1600.

[24] Watson, G.S. (1961). Goodness-of-fit tests on a circle, *Biometrika* **48**, 109–114.

(*See also* **Multivariate Normality, Tests of**; **Multivariate Techniques, Robustness**; **Robustness**)

RALPH B. D'AGOSTINO, SR

Nuisance Parameter

Consider a probabilistic model involving a set of parameters θ divided into two subsets θ_1 and θ_2 for which **inference** is required only about θ_1. The set θ_2 are then called *nuisance parameters*. Because the likelihood $f(x|\theta)$ involves θ_2, estimation for θ_1 must also involve the nuisance parameters.

Handling nuisance parameters is conceptually straightforward in a **Bayesian** context. The marginal posterior distribution for θ_1 may be found by integrating out the nuisance parameter as

$$f(\theta_1|x) = \int f(\theta_1, \theta_2|x)\, \mathrm{d}\theta_2.$$

If θ_2 is high-dimensional, this computation may be numerically intensive, but is always theoretically possible by **simulation**.

In non-Bayesian inference, a standard way to handle nuisance parameters is to find a set of statistics T_2 that are **sufficient** for the nuisance parameters θ_2 given θ_1 and to make inferences for θ_1 based on the **conditional** sampling distribution of $T_1|T_2, \theta_1$, where $\{T_1, T_2\}$ are the complete set of sufficient statistics for x. Because T_2 is sufficient for θ_2, the distribution of $T_1|T_2$ cannot involve θ_2 and therefore provides information only about θ_1 [1, pp. 27–28]. It can be shown [3, p. 145] that this is the only way to make the inference independent of the nuisance parameters.

This conditional approach simplifies the problem because it assumes that the probabilistic properties of the sampling procedure are completely independent of the nuisance parameters, but this simplification can result in a loss of **information** and suboptimal inference. Consider the full **likelihood** written as $f(x|\theta) = f(T_1, T_2|\theta_1, \theta_2)$. Because T_2 is sufficient for θ_2, we may write

$$f(x|\theta) = f(T_1|T_2, \theta_1) f(T_2|\theta_1, \theta_2).$$

Use of $f(T_1|T_2, \theta_1)$ to make inferences about θ_1 is then optimal only if the second term $f(T_2|\theta_1, \theta_2) = f(T_2|\theta_2)$; that is, only if T_2 is **ancillary** for θ_1 given θ_2.

An example involves the comparison of two **binomial** probabilities, p_1 and p_2, drawn from samples with size n_1 and n_2. Let x_1 be the number of successes in the first group and let x_2 be the number of successes in the second group. Then, reparameterizing in terms of the log **odds ratio**

$$\delta = \log\left[\frac{p_2/(1-p_2)}{p_1/(1-p_1)}\right]$$

and the logit of the success probability in the first group (*see* **Logistic Regression**),

$$\alpha = \log\left[\frac{p_1}{(1-p_1)}\right],$$

the likelihood may be written

$$L = \frac{\exp(\alpha x_1)}{(1+\exp\alpha)^{n_1}} \times \frac{\exp\left[(\alpha+\delta)x_2\right]}{\left[1+\exp(\alpha+\delta)\right]^{n_2}}.$$

It can be shown that sufficient statistics for α and δ are $x_\cdot = x_1 + x_2$ and x_2. For fixed δ, inference about α may be made from the total number of successes, $x_\cdot$ which are then sufficient. Thus, the distribution of x_2 given $x_\cdot$ is independent of α and, accordingly, can be used to make inferences about δ. When $\delta = 0$, this distribution is the **hypergeometric** and leads to the well-known **exact test** of significance due to Fisher [1, pp. 43–48] (*see* **Fisher's Exact Test**). This "exact" inference is not optimal, however, because $x_\cdot$ is not ancillary for δ given α.

Another common technique for handling nuisance parameters for which sufficient statistics exist is to make inferences for θ_1 from a sampling distribution conditional on θ_2 with the sufficient statistics substituted for θ_2. For example, the mean $\overline{x}$ of a sample of size n from a $\mathrm{N}(\mu, \sigma^2)$ distribution follows a $\mathrm{N}(\mu, \sigma^2/n)$ distribution. Often, if we do not know σ^2, we may replace it by its sample estimate s^2 in making inferences about μ. This procedure is exact if $\sigma^2 = s^2$ and will be good enough for large samples. Of course, in this problem, the exact solution for the unknown nuisance parameter is based on the pivotal quantity $t = n(\overline{x} - \mu)/s$ that follows a **Student's *t* distribution**, but in many problems such pivotal quantities may not exist. In general, assuming nuisance parameters to be known will place too much precision on the estimates of the other parameters because of the failure to consider the added variation from the unknown parameters. It can be shown that the asymptotic variance of an **efficient estimator**, when some parameters are unknown, is always at least as large as its value when they are all known [2, pp. 437–438].

References

[1] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Ed. Chapman & Hall, New York.

[2] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[3] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd Ed. Wiley, New York.

(*See also* **Estimation**; **Two-by-Two Table**)

CHRISTOPHER H. SCHMID

Null Hypothesis

A *null hypothesis* (usually symbolically represented as H_0:) is a statement about a population or set of populations which is tested through completion of an experiment. If only one population is being studied, then the null hypothesis is generally stated in the form that a parameter, θ (a measure of the population), is equal to a hypothetical value, θ_0; thus, $H_0{:}\theta = \theta_0$. To illustrate, if the parameter of interest is the mean, μ, then a statement of a null hypothesis might be $H_0{:}\mu = \mu_0$, where μ_0 is specified by the experimenter. For example, it might be hypothesized that the mean length of hospital stay for patients undergoing coronary artery bypass surgery (CABG) is 7 days (or $H_0{:}\mu = 7$ days). If $k \geq 2$ populations are being investigated, then the null hypothesis is generally presented as a statement specifying how the value of a parameter, θ_i for the ith population, is related to that parameter from each of the other $k - 1$ populations. Most often this hypothetical statement takes the form that these values, θ_j, $j = 1, 2, \ldots, k$, are all equal, i.e. $H_0{:}\theta_1 = \theta_2 = \cdots = \theta_k$. To illustrate, if the parameter of interest is the mean, μ, then the statement of a null hypothesis might be $H_0{:}\mu_1 = \mu_2 = \cdots = \mu_k$. For example, if μ_1 is the mean length of hospitalization for a population of males post-CABG, and μ_2 is the comparable parameter for a population of females, then $H_0{:}\mu_1 = \mu_2$. The usual presentation of a null hypothesis that values of the parameter across the k populations are equal suggests the derivation of the term *null* (a nullity or no difference) in the testing terminology. The word *null* is not always used, and one may call this underlying hypothetical statement merely the *hypothesis*, or equivalently the *statistical hypothesis* or *tested hypothesis*.

A null hypothesis most often takes the form that a parameter, θ, for a single population is equal to a specified value, θ_0, or that values of a parameter for a set of populations are all equal. However, a null hypothesis can be directional, that is, specifying a bound on a parameter, say, $H_0{:}\theta < \theta_0$, or a relationship among values of a parameter for several populations, for example, $H_0{:}\theta_1 < \theta_2$ for $k = 2$ populations.

A researcher should note the distinction between the null hypothesis (which is a statement about the *statistical* character of a parameter for one or several populations) and the *clinical* or *subject matter* hypothesis. The latter hypothesis frequently expresses a parameter relationship desired for the area of study and thus is not stated in the null (no effect) form. Often this clinical hypothesis indicates the magnitude of an anticipated effect of importance in the subject area. A finding from an experiment which results in the rejection of the null hypothesis is said to be *statistically significant*. This conclusion must be interpreted in light of what constitutes an *important or relevant* finding in the subject area (clinically) (*see* **Clinical Significance Versus Statistical Significance**; **Hypothesis Testing**).

(*See also* **Alternative Hypothesis**)

M.A. SCHORK

Number Needed to Treat (NNT)

Introduction

The number needed to treat (NNT) was originally proposed as a way of presenting the results of randomized **clinical trials** with **binary** outcome [16, 32, 33, 47, 48, 50]. Defined as the inverse of the **absolute risk** reduction (ARR), the number needed to treat is the average number of patients needed to be treated to prevent an adverse outcome in one additional patient compared to a control or standard treatment group. For example, in the Diabetes Control and Complications Trial (DCCT) the five-year **risk** of neuropathy in type 1 diabetic patients was 16.9% in the standard treatment group compared to 6.7% in the intensive insulin treatment group [18]. The absolute effect of the treatment can be described by $\text{ARR} = 16.9\% - 6.7\% = 10.2\%$. This translates to $\text{NNT} = 1/0.102 = 9.8 \approx 10$, that is, on average 10 patients are needed to be treated with intensive diabetes therapy to prevent one additional case of neuropathy compared to the standard therapy. For an adequate interpretation of NNTs, the characteristics of patients being treated, the outcome being measured, and the type and duration of interventions being compared have to be known.

NNT as well as ARR represent absolute measures of the treatment effect. Relative effect measures such as the **odds ratio** (OR), the **relative risk** (RR), or the relative risk reduction (RRR) frequently result in impressive numbers, even though the absolute effect of the treatment might be low. For example, if the two risks are $\pi_0 = 0.6$ and $\pi_1 = 0.1$, then $\text{RRR} = 83\%$, $\text{ARR} = 0.5$ and $\text{NNT} = 2$; if the two risks are $\pi_0 = 0.006$ and $\pi_1 = 0.001$, then $\text{RRR} = 83\%$ remains the same, but $\text{ARR} = 0.005$ and $\text{NNT} = 200$. Owing to the low baseline risk, the absolute effect of the treatment is also low, which is described by ARR and NNT. The information given by ARR and NNT is mathematically identical. However, the statement "200 patients are needed to be treated in order to avoid one event" is potentially more informative and comprehensible than "the treatment reduces the risk of an event by 0.005". Several studies demonstrated that assessment of health-care intervention effects by consumers is affected by the way in which study results are presented. The inclination of physicians to prescribe drugs and to treat patients is stronger when study results are presented by means of relative effect measures than when the same study is described by using absolute effect measures [12, 23, 40]. Health authority members are more willing to support health programs when results are expressed as RRRs compared with absolute effect measures [22]. Likewise, more patients assent to receive a therapy when potential benefits are reported in terms of RRR rather than ARR or NNT [29].

NNT has become the standard for presenting results of randomized clinical trials in the journal *Evidence-Based Medicine* [47] and the *ACP Journal Club* [1] and use of NNT to express study results is suggested in the **CONSORT** explanation and elaboration document [6]. However, the widespread application and extension of NNT in different settings is not without difficulties and care is required to use and interpret NNT appropriately. Recent developments regarding NNT are given by the development of methods to express benefit as well as harm, the calculation, presentation, and interpretation of confidence intervals, the application in screening studies, public health research, **epidemiology** (case–control and cohort studies), crossover studies, studies measuring continuous and time-to-event data, risk-benefit analyses, and systematic reviews. In the following, the characteristics and application areas of NNT are summarized.

General Characteristics

Relation to Other Effect Measures

A large number of effect measures exist to express the magnitude of difference between two groups concerning the risk of an adverse event. Let π_0 be the risk in the control group and π_1 be the risk in the treatment group. In the case of a beneficial treatment ($\pi_0 > \pi_1$) the most frequently used effect measures derived from a simple **2 x 2 table** are

$$\begin{aligned}
&\text{Absolute risk reduction:} && ARR = \pi_0 - \pi_1 \\
&\text{Relative risk:} && RR = \frac{\pi_1}{\pi_0} \\
&\text{Relative risk reduction:} && RRR = \frac{\pi_0 - \pi_1}{\pi_0} \\
& && \quad = 1 - RR
\end{aligned}$$

Odds ratio: $$OR = \frac{\pi_1 \times (1-\pi_0)}{\pi_0 \times (1-\pi_1)}$$

Number needed to treat: $$NNT = \frac{1}{\pi_0 - \pi_1} = \frac{1}{ARR}$$

The relation between NNT and ARR is obvious. It is helpful in practice to also express NNT as a function of RR, RRR, OR, and the control event rate π_0. The respective formulae are given by

$$NNT = \frac{1}{(1-RR) \times \pi_0} = \frac{1}{RRR \times \pi_0} \quad (1)$$

$$NNT = \frac{1-(1-OR) \times \pi_0}{(1-OR) \times \pi_0 \times (1-\pi_0)} = \frac{1}{(1-OR) \times \pi_0} + \frac{OR}{(1-OR) \times (1-\pi_0)} \quad (2)$$

Similar formulas are published for the case of harmful treatments [7, 9], for considering desirable instead of adverse outcomes [39], and for the inverse definitions of RR and OR [28].

Quantifying Benefit and Harm

NNT represents the inverse of the difference of two risks. On principle, the difference of two risks can be positive, zero, or negative. The concept of NNT was originally developed for the situation of a beneficial treatment, so that the risk of an adverse event in the treatment group is lower than in the control group [33]. Thus, calculating the risk difference as control minus treatment leads to a positive ARR value. Considering only beneficial treatments, the term "number needed to treat" was proposed to describe the inverse of ARR. In the case of a harmful treatment, this calculation leads to a negative risk difference and a negative NNT. To avoid negative numbers, the risk difference is calculated as treatment minus control if the risk of the treatment group is higher than that of the control group leading to a positive value called absolute risk increase (ARI). To describe the inverse of ARI, the unfavorable term "number needed to harm" (NNH) was used [39]. Recognizing that NNT and NNH are not good abbreviations, Altman suggested the terminology "number of patients needed to be treated for one additional patient to benefit" (NNTB) or "be harmed" (NNTH) [2]. This terminology should be used when it is necessary to indicate the direction of the effect. In the case of desirable outcomes, such as healing or improvement of **quality of life**, the order of the two probabilities in the calculation of NNT is reversed. Here, NNTB represents the average number of patients needed to be treated to gain one additional beneficial outcome compared to a control or standard treatment group [55].

Confidence Intervals

As with other estimated effect measures, it is important to document the uncertainty of the estimation by means of an appropriate **confidence interval**. In principle, confidence intervals for NNTs can be obtained by inverting and exchanging the confidence limits of the corresponding risk difference [17]. Nevertheless, calculating, presenting, and interpreting confidence intervals for NNTs is not straightforward. Owing to the reciprocal **transformation**, the NNT has undesirable statistical properties [34]. To obtain meaningful confidence intervals for NNT two issues have to be considered. Firstly, the unusual scale of NNT has to be taken into account, and secondly, an appropriate method to calculate confidence intervals for the risk difference is required.

The key to understand the confidence interval for NNT is that the domain of NNT is the union of 1 to ∞ (in the NNTB region) and $-\infty$ to -1 (in the NNTH region). The best value of NNT indicating the largest possible beneficial treatment effect is 1, the NNT value indicating no treatment effect (ARR = 0) is $\pm\infty$, and the worst NNT value indicating the largest possible harmful effect is -1. Values between -1 and 1 are impossible for NNT. Owing to estimation uncertainty, the estimated NNT may be negative even when the true NNT is positive and vice versa. Even when the sign of the estimated and true NNT are identical, the estimation uncertainty can be so large that neither a harmful nor a beneficial effect can be excluded. In this case, the confidence interval covers both the NNTB and the NNTH region. Thus, the result NNT = 10 with confidence limits 4 and -20 means that the two regions 4 to ∞ and -20 to $-\infty$ form the confidence interval. To make this clear, a confidence interval for an NNT estimate that is not statistically significant should be presented as NNTB = 10 (NNTB 4 to ∞ to NNTH 20) [2]. This presentation indicates that a beneficial treatment effect of NNTB = 10 is estimated, but the uncertainty

of this estimation is so large that a more beneficial effect up to NNTB = 4 and a less beneficial effect up to NNTB = ∞ (no effect at all) as well as a harmful effect up to NNTH = 20 is compatible with the observed data.

For large sample sizes and risks not close to 0 or 1, the usual Wald method can be used to calculate confidence intervals for risk differences (*see* **Estimation, Interval**). However, Wald confidence intervals have poor coverage probabilities and a propensity to aberrations in many practical situations. Thus, Newcombe proposed to calculate confidence intervals for risk differences based upon Wilson scores [42]. This method was also recommended for NNT [8].

Let n_0 and n_1 be the number of patients in the control and the treatment group, respectively, and let e_0 and e_1 be the number of patients having an event in the control and the treatment group, respectively. The risks of an event in the two groups can then be estimated by the proportions $p_0 = e_0/n_0$ and $p_1 = e_1/n_1$. The effect measures can be estimated by ARR $= p_0 - p_1$ and NNT $= 1/(p_0 - p_1)$. Using this notation, the $100 \times (1 - \alpha)\%$ confidence interval for ARR based upon Wilson scores is given by:

$$LL(ARR) = p_0 - p_1 - \delta \text{ and}$$
$$UL(ARR) = p_0 - p_1 + \varepsilon, \qquad (3)$$

where

$$\delta = \sqrt{(p_0 - l_0)^2 + (u_1 - p_1)^2},$$
$$\varepsilon = \sqrt{(u_0 - p_0)^2 + (p_1 - l_1)^2},$$
$$\mathrm{l}_i = \varphi_i - \sqrt{\varphi_i^2 - \psi_i}, \mathrm{u}_i = \varphi_i + \sqrt{\varphi_i^2 - \psi_i}, \mathrm{i} = 0, 1,$$
$$\varphi_i = \frac{2e_i + z_{1-\alpha/2}^2}{2(n_i + z_{1-\alpha/2}^2)}, \psi_i = \frac{e_i^2}{n_i^2 + n_i z_{1-\alpha/2}^2}, \mathrm{i} = 0, 1,$$

and $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the **standard normal** distribution.

The corresponding confidence limits for NNT can then be calculated by LL(NNT)=1/UL(ARR) and UL(NNT)=1/LL(ARR) in consideration of the NNT scale ranging from 1 through ∞ to -1 (see above). An SAS program can be used for calculations [8].

Confidence intervals for NNT based upon Wilson scores seem to be adequate for most practical applications. For very small sample sizes or applications, which require that the true confidence level under no circumstances remains under the nominal level, exact [14] or quasi-exact methods [15] should be used (*see* **Exact Inference for Categorical Data**).

Extensions and Applications

The principle of NNT has been extended and suggested for use in a wide variety of circumstances. The most important ones are summarized below.

Screening

Rembold extended the NNT concept to compare strategies for disease **screening** [44]. The analogous statistic termed "number needed to screen" (NNS) describes the number of people that need to be screened to prevent one death or adverse event. In clinical trials that directly investigate the benefit of a screening strategy, the point and interval estimation of NNS is identical to that of NNT. However, the intervention under study is a screening strategy applied to a population, rather than a treatment applied to patients. If no study exists that evaluates directly the benefit of a screening strategy, NNS estimation can be performed by combining the knowledge of clinical trials investigating the benefit of treating **risk factors** and the **prevalence** of persons with inadequately treated risk factors in the community. Under the assumption that screened individuals with positive results will show full compliance with subsequent treatment, NNS can be calculated by dividing the corresponding NNT by the prevalence of unaware or untreated disease.

Expressing the absolute effect of screening strategies as NNS values has the same advantages as the presentation of treatment effects by means of NNTs. However, the NNS approach has some limitations. Firstly, the division of NNTs by an estimated prevalence of untreated disease is subject to propagation of errors. A method to calculate confidence intervals for NNS taking the uncertainty of both the NNT and the prevalence estimation into account is required. Secondly, NNS values calculated from clinical trials investigating the benefit of a screening strategy directly (*see* **Screening Trials**) may not be comparable to NNS values calculated from NNTs divided by the prevalence of unaware or untreated disease. The former may be more affected by participation and selection effects than the latter. Hence, Richardson suggested to multiply the directly estimated NNS

by the participation rate adjusted for **selection** to obtain an NNS value free of participation and selection effects [45]. However, this method is even more exposed to propagation of errors. Moreover, the benefit of a screening strategy should be described including participation and selection effects. Analogous to the **intention to treat** analysis of clinical trials, the gold standard is the unadjusted NNS estimated from trials directly investigating the benefit of screening strategies.

Public Health Research

The NNT statistic relates to those patients actually treated and gives no information how many people of all patients with the disease or of the total population will benefit from the treatment. Heller & Dobson proposed two new statistics offering a public health perspective [27]. The idea is similar to that of NNS calculated by NNT divided by the prevalence of unaware or untreated disease. The "disease impact number" (DIN) takes into account the number of people in the population with the disease, not just those eligible for treatment according to the entry criteria of the considered clinical trial. DIN is calculated by dividing NNT by the proportion of patients with the disease who are eligible for treatment. The "population impact number" (PIN) takes into account the total size of the population from which the patients with the disease are drawn. PIN is calculated by dividing DIN by the prevalence of disease in the population. DINs and PINs suffer from limitations similar to those of indirectly estimated NNS values. Owing to the division of NNTs by estimated proportions they are subject to greater **random error** than NNT. However, they may play a role as communication tool for treatment effects from a population perspective [52].

Case–control Studies

Bjerre & LeLorier proposed to use the NNTH statistic to express the magnitude of harmful exposures effects in **case–control studies** [11]. As information about the absolute risk is not directly available from case–control studies, they calculated NNTH by using the odds ratio provided by the case–control study and the unexposed event rate obtained from external sources. Although not mentioned by the authors, an additional advantage of this approach is that adjusted NNTs can be calculated by using adjusted ORs to estimate the corresponding NNT values (see next section). Confidence intervals for NNTH are calculated by transforming the confidence limits of OR. Unfortunately, to calculate NNTH as function of OR, formula (1) was used, which actually represents the relation between NNT and RR. Thus, NNTH is systematically underestimated, that is, the exposure effect is overestimated. The magnitude of this error is negligible if OR and RR are approximately equal. Thus, in case-control studies, in which usually rare diseases are investigated, the error is unimportant. However, in situations where OR and RR are quite different, either formula (1) with RR or formula (2) with OR must be applied to obtain correct results. Let $NNTH_{1,\mathrm{OR}}$ be the *NNTH* value calculated by formula (1) with OR and let $NNTH_{\mathrm{true}}$ be the true *NNTH*. It can be shown that $(NNTH_{\mathrm{true}} - NNTH_{1,\mathrm{OR}})/NNTH_{\mathrm{true}} = \pi_1$, that is, the relative error of $NNTH_{1,\mathrm{OR}}$ equals the exposed event rate [7]. Even, if the correct formula is used, a limitation of this approach is that the confidence interval for NNTH takes into account the uncertainty of the OR estimation but not that of the unexposed event rate. A possible solution is given by the methods developed by King & Zeng for point and interval estimation of risk differences in case–control studies based upon **Bayesian methods** or a range of possible values for the unexposed event rate [30].

Cohort Studies

The NNT concept has been applied to compare exposed and unexposed persons in **cohort studies** [9]. For this application, the term "number needed to be exposed" (NNE) was suggested. When it is necessary to distinguish between harmful and beneficial **exposures**, the abbreviations NNEH and NNEB should be used. In the case of a harmful exposure, NNEH represents the average number of persons needed to be exposed for one additional case of disease or death compared to the unexposed persons. NNEs are calculated as a function of the odds ratio and the unexposed event rate by means of formula (2). This approach allows the calculation of adjusted NNEs by using adjusted odds ratios, estimated, for example, by multiple **logistic regression**. Within the framework of logistic regression, the adjusted odds ratio is constant over the distribution of the considered **confounders**. However, the event rates and their differences are dependent on the confounder

values. Thus, NNE also varies with the values of the confounding variables, which has to be taken into account when adjusted NNEs are estimated. Two methods were proposed to calculate adjusted NNEs. In the first approach, the mean risk of the unexposed persons is used and NNE is calculated for the corresponding confounder profile. In the second approach, NNE is calculated for some fixed confounder profiles, which gives an impression about different absolute effects of the exposure in cohorts with varying confounder values. A similar principle is applied to calculate pooled NNTs in meta-analysis (see below).

Confidence intervals for adjusted NNEs can be calculated indirectly via confidence intervals for the corresponding risk difference. Within the framework of logistic regression analysis applied to prospective cohort data, risk differences between the exposed and unexposed persons can be expressed as functions of the logistic regression coefficients. Thus, approximate **standard errors** and confidence intervals for risk differences can be calculated by means of the multivariate **delta method** [9]. In contrast with the calculation of NNTs in case–control studies, this method takes the estimation uncertainties of both the odds ratio and the unexposed event rate into account. The adequacy of the approximate confidence intervals was investigated via **simulations** demonstrating sufficient quality for most epidemiological applications [10].

Continuous Data

NNT represents a summary statistic for the comparison of two groups concerning a binary outcome. Nevertheless, in some applications, investigators want to express their study results in terms of NNT although the outcome variable is measured in a continuous scale (*see* **Random Variable**). One obvious method to calculate NNTs for continuous outcomes is to dichotomize the response in both groups and to apply the usual methods. Alternatively, one can dichotomize the difference of the responses between the two groups. Walter examined the probability that the difference of the responses between the two groups is larger than the minimally important difference (*see* **Sample Size Determination for Clinical Trials**) [55]. Without loss of generality, we assume that higher response values correspond to adverse outcomes (such as hypertension). Let X_0 and X_1 be the control and treatment responses of a given subject and c be the minimally important difference. The probability described above is given by $\theta = P(X_0 - X_1 > c)$. The continuous data version of NNT is then calculated by $NNT = 1/\theta$. Under the assumption of **bivariate normality** of (X_0,X_1), θ is given by

$$\theta = \Phi\left(\frac{\mu_0 - \mu_1 - c}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1}}\right), \qquad (4)$$

where Φ denotes the distribution function of the standard normal distribution, μ_0 and μ_1 and σ_0 and σ_1 are the means and standard deviations of X_0 and X_1, respectively, and ρ is correlation of X_0 and X_1 [55]. Estimation of θ and NNT is performed by substituting the usual estimates of μ_0, μ_1, σ_0, σ_1 and ρ into (4). Formulas for the standard error of the estimated probability θ can be derived by means of the delta method both for paired and unpaired data [55].

It should be noted that formula (4) is first of all only useful in studies, which provide an estimate of ρ (such as **crossover** studies, see below). In all designs considered so far (randomized clinical trials with parallel group design, cohort studies, and case–control studies with two independent groups) the within-subject correlation is not estimable. In this case, Walter proposed to use a variety of different assumed values of ρ and investigate the **sensitivity** of θ to the unknown correlation value [55]. Alternatively, in studies observing independent groups, the first mentioned approach of dichotomizing the response in both groups could be used.

In practice, continuous outcomes are frequently subject to random **measurement error**. Even in the case of **nondifferential** measurement error, dichotomization of continuous variables leads to a **bias** in the estimated proportions and estimated NNTs. Walter & Irwig investigated the effect of measurement error in continuous outcomes on NNT estimation [56], and methods to reduce the bias by adjusting for measurement error are in development [38]. In general, even in the case of no measurement error, one should be aware of the potential loss of information due to **categorizing of continuous variables**. Hence, calculation of NNTs from continuous data can only serve as supplement to the analysis of data in the original continuous scale by using means and differences of means.

Crossover Studies

Originally, the NNT statistic was developed for use in studies investigating two independent groups. Walter systematically examined NNT estimators and their variances for both crossover and parallel group designs [55]. Owing to the undesirable statistical properties of NNT, it is preferable to calculate the standard errors of the corresponding risk differences instead of the NNTs themselves. The NNT estimators are identical in both designs, whereas standard errors are different. Approximate confidence intervals for risk differences can be calculated in both designs by using the Wald method [55]. As described before, confidence intervals for NNTs can be obtained by inverting and exchanging confidence limits of the corresponding risk difference [17]. For the parallel group design, it was shown that the Wald method is unreliable in many practical situations. The same holds for crossover studies, in which it is preferable to calculate confidence intervals for the difference between paired proportions based upon the Wilson score method [41].

As crossover studies provide an estimate of the within-subject correlation, the continuous data version of NNT based upon the minimally important difference (see above) can be estimated directly. Under the assumption of normality of the continuous response, NNT can be estimated by using equation (4). Without making distributional assumptions, NNT is given by the inverse proportion of subjects for which the difference between the responses is larger than the minimally important difference [55].

Survival Data

The concept of NNT was originally developed for binary outcomes measured at a specific fixed time point. Nevertheless, NNTs are also calculated and presented for studies where the outcome is the time to an event (*see* **Survival Analysis, Overview**). Unfortunately, unclear and questionable methods have been used for point and interval estimation of NNT in studies in which follow-up times are not equal for all patients. Owing to the application of questionable *ad hoc* methods, different and confusing results have been published for the same data [8].

First, it should be noticed that in studies with varying follow-up times, NNT would also vary according to the length of follow-up. In such studies, no single NNT value exists. NNT can be calculated at any time point after the start of the treatment. Frequently used methods to analyze survival times are given by **Kaplan–Meier** survival curves providing estimates of the survival probabilities $S_0(t)$ and $S_1(t)$ of the control and treatment group, respectively, and the **Cox regression model**, providing an estimate of the **hazard ratio** (HR), possibly adjusted for other **prognostic** variables. Altman & Andersen proposed to estimate NNT by means of

$$NNT(t) = \frac{1}{S_1(t) - S_0(t)} \tag{5}$$

if the survival probabilities $S_0(t)$ and $S_1(t)$ are given, or by

$$NNT(t) = \frac{1}{(S_0(t))^{\mathrm{HR}} - S_0(t)} \tag{6}$$

if the assumption of **proportional hazards** is fulfilled and $S_0(t)$ and the HR for the comparison of the control and treatment group are given [3]. If one fixed time point is specified, one NNT value is obtained. Otherwise, (5) and (6) will lead to a NNT curve as a function of time.

To get an NNT statistic independent of time, Lubsen et al. proposed to calculate NNTs by the reciprocal of the difference of two **hazards** [36]. However, this approach requires the assumption of constant hazards. Moreover, the difference of hazards is not the same as the difference of risks. Thus, this approach leads to a statistic with a different meaning than that of the usual NNT. It should be noted that in the presence of confounders survival probabilities are dependent on the confounder values even if we can assume a constant HR. Thus, NNT not only depends on time but also on confounders. Altman & Andersen proposed to calculate NNT curves for different subsets of patients with varying **prognosis** [3]. However, more work is required to develop methods for estimation of adjusted NNTs from survival times.

Combining and Pooling

Risk-benefit Analysis

The decision about the use of a treatment should not be based upon its effect on the target event alone. Adverse side effects attributable to treatment as well as costs of therapy and costs avoided by preventing target events should also be considered (*see* **Decision Analysis in Diagnosis and Treatment Choice**). The

"threshold NNT" (NNT_T) was defined as the NNT value at which the therapeutic benefit equals the therapeutic risks [25, 26, 51]. If the estimated NNT is below the threshold NNT, then treatment should be administered. If the estimated NNT is above the threshold NNT, the patients should not be treated because the risks and costs of treatment are larger than the expected benefit. The threshold NNT is given by

$$NNT_T = \frac{TEC + TEV}{DC + AER \times (AEC + AEV)}, \quad (7)$$

where TEC represents the costs of treating one target event, TEV the value of one target event avoided (given in the same economic units as costs), DC the direct costs of therapy, AEC the costs of treating one adverse side effect, AEV the value of the side effect and AER the event rate of the side effect [51]. Similar formulas for considering multiple side effects and omitting costs can be found elsewhere [51].

While the concept of the threshold NNT seems to be appealing, the practical application is challenging. For an adequate decision making, the estimation uncertainties should be taken into account. The specification of the data (costs and values) required for the calculation of the threshold NNT is not easy and the quantification of these data uncertainties is much more difficult. Especially, the values one is willing to pay for one target or one side effect avoided are highly subjective. Thus, it is quite important to disclose all data and assumptions used for calculating a threshold NNT.

Combined NNT Measures for Different Outcomes

Several approaches have been published to combine the NNTB of the target event and the NNTH of a side effect into one measure incorporating benefit as well as harm. Let π_0, π_1 be risks of the target event and ν_0, ν_1 the risks of the side effect in the control and the treatment group, respectively. We consider the case of an adverse target event, an adverse side effect and a treatment that is beneficial concerning the target event ($\pi_0 > \pi_1$) but harmful concerning the side effect ($\nu_1 > \nu_0$). For other situations, appropriate modifications of the following measures are required. Riegelman & Schroth proposed the combined measure

$$NNT_{\text{comb}} = \frac{1}{(\pi_0 - \pi_1) - (\nu_1 - \nu_0)}, \quad (8)$$

that is, the reciprocal of the difference between ARR of the target event and ARI of the side effect [46]. The authors proceeded by adjusting this measure for the qualities and timings of the considered outcomes [46]. This procedure was criticized because a decision analysis has to be carried out before the quality-adjusted NNT can be calculated [19]. Thus, the intuitive meaning, which is one advantage of the NNT statistic, is lost. It is only possible to interpret a quality-adjusted NNT if the underlying decision analysis is understood. The statistical properties of the quality-adjusted NNT statistic have not been investigated and no methods to calculate confidence intervals have been developed.

A second approach of an NNT measure incorporating benefit and harm was proposed by Schulzer & Mancini [49]. They tried to calculate the number of patients needed to treat to produce one "unqualified success" (US), that is, the situation in which one adverse target event is avoided while simultaneously no treatment-induced side effect occurred. The NNT for one unqualified success is given by

$$NNT_{\text{US}} = \frac{1}{(\pi_0 - \pi_1) \times [1 - (\nu_1 - \nu_0)]}. \quad (9)$$

Formula (9) is based upon the assumption that the target event and the adverse side effect are independent in both the untreated and the treated population. This assumption will rarely be true in practical applications. Although a procedure was proposed to handle situations in which an association between the prevention of a target event and the induction of a side effect is expected [37], this approach suffers from the lack of an appropriate method to estimate the association from the data. Moreover, no adequate method to calculate confidence intervals for NNT_{US} has been developed.

Willan et al. proposed the benefit–risk ratio

$$R = \frac{NNTH\,(\text{side effect})}{NNTB\,(\text{target effect})} = \frac{\pi_0 - \pi_1}{\nu_1 - \nu_0}$$
$$= (\pi_0 - \pi_1) \times NNTH\,(\text{side effect}), \quad (10)$$

which can be interpreted as increase in the expected number of prevented target events achieved for each additional adverse side effect induced by treatment [57]. For large sample sizes, Willan et al. developed a statistical procedure to construct confidence intervals for the benefit–risk ratio based upon **Fieller's theorem** [57].

The development of a combined NNT statistic incorporating benefit and harm of multiple events is not straightforward. Before one of the proposed combined NNT measures considering multiple events can be routinely applied in practice, more work is required concerning the practical utility of these measure as well as their statistical properties, especially in small samples.

Meta-analysis

Since NNT has been advocated as a useful effect measure for systematic reviews [39], a number of authors have pointed out that particular caution is needed in deriving pooled NNTs in **meta-analyses** [5, 13, 20, 21, 53]. A single pooled NNT value over all studies in a meta-analysis may be misleading, especially if there is a variation in the baseline risk, different lengths of follow-up, differences in the outcomes considered, or different clinical settings. The naive approach of simply adding the raw totals of all considered trials as if the data came from one trial should be avoided. The calculation of a pooled NNT should be based upon a pooled effect measure, which should be independent of the baseline risk. Using empirical data, Furukawa et al. showed that the relative effect measures OR and RR calculated by means of an appropriate **fixed** or **random effects** regression model often appear to be reasonably constant across different baseline risks [24]. Meaningful NNTs can be obtained by inserting the pooled RR or OR from meta-analyses in formula (1) or (2). If there is variation in the baseline risk, different NNTs relevant to specific patient subgroups should be calculated [20, 24, 53]. If there is evidence that even the relative effect measures vary substantially between subgroups in a meta-analysis, no meaningful pooled NNT can be calculated.

Conclusion

The use of NNT as effect measure for the comparison of risks between two groups has been advocated in medical journals for several years [16, 33, 39, 43, 47, 50] but was recently criticized [53, 58] or even rejected [28]. There seems to be a gap in the assessment of the practical usefulness of NNTs between some statisticians and clinicians [4, 28, 35]. Some mathematical arguments against the use of NNTs, such as undesirable distributional properties, are surely justified. However, strict mathematical arguments lose their importance when NNT is considered as a way of presenting results, not as a tool for statistical computations [4, 35]. A clear distinction should be made between data analysis and subsequent risk communication [54]. In the light of the effects on consumers of the scale in which benefits and risks are reported, it is frequently advisable to choose a statistical model and a corresponding appropriate summary measure for the task of data analysis, but alternative effect measures to report the most important results. For the translation of research findings to consumers, the number needed to treat may represent a useful tool, because it gives an intuitive impression of the absolute effect of a therapy or an intervention. NNTs contain the same information as risk differences, but in the unit of patient numbers instead of probabilities, which is easier to understand.

The attempt to extend and apply the simple NNT concept developed for randomized clinical trials with two independent groups and a binary outcome for a variety of other settings led to the development of more sophisticated approaches and procedures for NNT calculation. Some useful approaches have been developed, but situations remain for which further work is needed to calculate meaningful NNTs, for example, survival time studies or the combination of NNTs for multiple outcomes. These extension and adjustment procedures can alleviate problems with NNTs. However, the extended and adjusted NNTs can no more be considered as "one simple single yardstick" [31]. Particular caution is required to apply and interpret NNTs adequately in practice, especially in meta-analyses and in the presence of confounders. Nevertheless, if handled appropriately, NNTs represent a useful communication tool to express the absolute effects of interventions and exposures.

References

[1] ACP Journal Club (1997). Glossary, *ACP Journal Club* **126**, 28.

[2] Altman, D.G. (1998). Confidence intervals for the number needed to treat, *British Medical Journal* **317**, 1309–1312.

[3] Altman, D.G. & Andersen, P.K. (1999). Calculating the number needed to treat where the outcome is time to an event, *British Medical Journal* **319**, 1492–1495.

[4] Altman, D.G. & Deeks, J.J. (2000). Comments on the paper by Hutton, *Journal of the Royal Statistical Society – A* **163**, 415–416.

[5] Altman, D.G. & Deeks, J.J. (2002). Meta-analysis, Simpson's paradox, and the number needed to treat, *BMC Medical Research Methodology* **2**, 3.

[6] Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D.R., Gøtzsche, P.C., & Lang, T. for the CONSORT Group. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration, *Annals of Internal Medicine* **134**, 663–694.

[7] Bender, R. (2000). Expressing the number needed to treat as a function of the odds ratio and the unexposed event rate (Rapid Electronic Letter), *British Medical Journal* (eBMJ: `http://www.bmj.com/cgi/eletters/320/7233/503#EL7`).

[8] Bender, R. (2001). Calculating confidence intervals for the number needed to treat, *Controlled Clinical Trials* **22**, 102–110.

[9] Bender, R. & Blettner, M. (2002). Calculating the "number needed to be exposed" with adjustment for confounding variables in epidemiological studies, *Journal of Clinical Epidemiology* **55**, 525–530.

[10] Bender, R. & Kuss, O. (2003). Confidence intervals for adjusted NNEs: a simulation study (Letter), *Journal of Clinical Epidemiology* **56**, 205–206.

[11] Bjerre, L.M. & LeLorier, L. (2000). Expressing the magnitude of adverse effects in case-control studies: "The number of patients needed to be treated for one additional patient to be harmed", *British Medical Journal* **320**, 503–506.

[12] Bobbio, M., Demichelis, B. & Giustetto, G. (1994). Completeness of reporting trial results: Effect on physicians' willingness to prescribe, *Lancet* **343**, 1209–1211.

[13] Cates, C. (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis, *BMC Medical Research Methodology* **2**, 1.

[14] Chan, I.S.F. & Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions, *Biometrics* **55**, 1202–1209.

[15] Chen, X. (2002). A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small samples, *Statistics in Medicine* **21**, 943–956.

[16] Cook, R.J. & Sackett, D.L. (1995). The number needed to treat: a clinically useful measure of treatment effect, *British Medical Journal* **310**, 452–454.

[17] Daly, L.E. (1998). Confidence limits made easy: Interval estimation using a substitution method, *American Journal of Epidemiology* **147**, 783–790.

[18] DCCT Research Group (1995). The effect of intensive diabetes therapy on the development and progression of neuropathy, *Annals of Internal Medicine* **122**, 561–568.

[19] Dowie, J. (1998). The 'number needed to treat' and the 'adjusted NNT' in health care decision-making, *Journal of Health Services Research and Policy* **3**, 44–49.

[20] Ebrahim, S. (2001). The use of numbers needed to treat derived from systematic reviews and meta-analysis. Caveats and pitfalls, *Evaluation & the Health Professions* **24**, 152–164.

[21] Ebrahim, S. & Smith, G.D. (1999). The 'number need to treat': Does it help clinical decision making? *Journal of Human Hypertension* **13**, 721–724.

[22] Fahey, T., Griffiths, S. & Peters, T.J. (1995). Evidence based purchasing: understanding results of clinical trials and systematic reviews, *British Medical Journal* **311**, 1056–1095.

[23] Forrow, L., Taylor, W.C. & Arnold, R.M. (1992). Absolutely relative: how research results are summarized can affect treatment decisions, *American Journal of Medicine* **92**, 121–124.

[24] Furukawa, T.A., Guyatt, G.H. & Griffith, L.E. (2002). Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses, *International Journal of Epidemiology* **31**, 72–76.

[25] Guyatt, G.H., Sackett, D.L., Sinclair, J.C., Hayward, R., Cook, D.J. & Cook, R.J. for the Evidence-Based Medicine Working Group. (1995). Users' guides to the medical literature. IX. A method for grading health care recommendations, *Journal of the American Medical Association* **274**, 1800–1804.

[26] Guyatt, G.H., Sinclair, J.C., Cook, D.J. & Glasziou, P. for the Evidence-Based Medicine Working Group & for the Cochrane Applicability Methods Working Group. (1999). Users' guides to the medical literature. XVI. How to use a treatment recommendation, *Journal of the American Medical Association* **281**, 1836–1843.

[27] Heller, R.F. & Dobson, A. (2000). Disease impact number and population impact number: population perspectives to measures of risk and benefit, *British Medical Journal* **321**, 950–952.

[28] Hutton, J.L. (2000). Number needed to treat: properties and problems, *Journal of the Royal Statistical Society – A* **163**, 403–415.

[29] Hux, J.E. & Naylor, C.D. (1995). Communicating the benefits of chronic preventive therapy: does the format of efficacy data determine patients' acceptance of treatment? *Medical Decision Making* **15**, 152–157.

[30] King, G. & Zeng, L. (2002). Estimating risk and rate levels, ratios and differences in case-control studies, *Statistics in Medicine* **21**, 1409–1427.

[31] Kristiansen, I.S., Gyrd-Hansen, D., Nexøe, J. & Nielsen, J.B. (2002). Number needed to treat: easily understood and intuitively meaningful? Theoretical considerations and a randomized trial, *Journal of Clinical Epidemiology* **55**, 888–892.

[32] Laupacis, A. & Sackett, D.L. (1998). Number needed to treat (NNT), in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds. Wiley, Chichester, pp. 3081–3088.

[33] Laupacis, A., Sackett, D.L. & Roberts, R.S. (1988). An assessment of clinically useful measures of the consequences of treatment, *New England Journal of Medicine* **318**, 1728–1733.

[34] Lesaffre, E. & Pledger, G. (1999). A note on the number needed to treat, *Controlled Clinical Trials* **20**, 439–447.

[35] Lesaffre, E. & Pledger, G. (2000). Comments on the paper by Hutton, *Journal of the Royal Statistical Society – A* **163**, 417.

[36] Lubsen, J., Hoes, A. & Grobbee, D. (2000). Implications of trial results: the potentially misleading notations of number needed to treat and average duration life gained, *Lancet* **356**, 1757–1759.

[37] Mancini, G.B. & Schulzer, M. (1999). Reporting risks and benefits of therapy by use of the concepts of unqualified success and unmitigated failure: applications to highly cited trials in cardiovascular medicine, *Circulation* **99**, 377–383.

[38] Marschner, I.C., Emberson, J., Irwig, L. & Walter, S.D. (2003). Adjustment for bias in the number needed to treat (NNT) when outcome is based on a continuous quantity measured with error, *Journal of Clinical Epidemiology* (submitted for publication).

[39] McQuay, H.J. & Moore, A. (1997). Using numerical results from systematic reviews in clinical practice, *Annals of Internal Medicine* **126**, 712–720.

[40] Naylor, C.D., Chen, E. & Strauss, B. (1992). Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Annals of Internal Medicine* **117**, 916–921.

[41] Newcombe, R.G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data, *Statistics in Medicine* **17**, 2635–2650.

[42] Newcombe, R.G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods, *Statistics in Medicine* **17**, 873–890.

[43] Nuovo, J., Melnikow, J. & Chang, D. (2002). Reporting number needed to treat and absolute risk reduction in randomized controlled trials, *Journal of the American Medical Association* **287**, 2813–2814.

[44] Rembold, C.M. (1998). Number needed to screen: development of a statistic for disease screening, *British Medical Journal* **317**, 307–312.

[45] Richardson, A. (2001). Screening and the number needed to treat, *Journal of Medical Screening* **8**, 125–127.

[46] Riegelman, R. & Schroth, W.S. (1993). Adjusting the number needed to treat: Incorporating adjustments for the utility and timing of benefits and harms, *Medical Decision Making* **13**, 247–252.

[47] Sackett, D.L. (1996). On some clinically useful measures of the effects of treatment, *Evidence-Based Medicine* **1**, 37–38.

[48] Sackett, D.L. & Cook, R.J. (1994). Understanding clinical trials. What measures of efficacy should journals provide busy clinicians, *British Medical Journal* **309**, 755–756.

[49] Schulzer, M. & Mancini, G.B. (1996). 'Unqualified success' and 'unmitigated failure': number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events, *International Journal of Epidemiology* **25**, 704–712.

[50] Sinclair, J.C. & Bracken, M.B. (1994). Clinically useful measures of effect in binary analyses of randomized trials, *Journal of Clinical Epidemiology* **47**, 881–889.

[51] Sinclair, J.C., Cook, R.J., Guyatt, G.H., Pauker, S.G. & Cook, D.J. (2001). When should an effective treatment be used? derivation of the threshold number needed to treat and the minimum event rate for treatment, *Journal of Clinical Epidemiology* **54**, 253–262.

[52] Smeeth, L. & Ebrahim, S. (2000). Commentary: DINs, PINs, and things – clinical and population perspectives on treatment effects, *British Medical Journal* **321**, 952–953.

[53] Smeeth, L., Haines, A. & Ebrahim, S. (1999). Numbers needed to treat derived from meta-analyses – sometimes informative, usually misleading, *British Medical Journal* **318**, 1548–1551.

[54] Walter, S.D. (2000). Choice of effect measure for epidemiological data, *Journal of Clinical Epidemiology* **53**, 931–939.

[55] Walter, S.D. (2001). Number needed to treat (NNT): estimation of a measure of clinical benefit, *Statistics in Medicine* **20**, 3947–3962.

[56] Walter, S.D. & Irwig, L. (2001). Estimating the number needed to treat (NNT) index when the data are subject to error, *Statistics in Medicine* **20**, 893–906.

[57] Willan, A.R., O'Brien, B.J. & Cook, D.J. (1997). Benefit-risk ratios in the assessment of the clinical evidence of a new therapy, *Controlled Clinical Trials* **18**, 121–130.

[58] Wu, L.A. & Kottke, T.E. (2001). Number needed to treat: Caveat emptor, *Journal of Clinical Epidemiology* **54**, 111–116.

(*See also* **Categorical Data Analysis**; **Evidence-based Medicine**; **Risk Assessment in Clinical Decision Making**)

RALF BENDER

Numerical Analysis

Numerical analysis is concerned with the accurate and efficient evaluation of mathematical expressions, especially on computers with **floating point arithmetic**. While scientists have always been concerned to some extent with numerical computation, the modern discipline of numerical analysis is almost entirely a product of the period since 1950, during which there has been an explosion in the availability of electronic computers. There are three main issues:

1. to organize computations so that there is minimum accumulation of error in floating point arithmetic
2. to organize computations efficiently, so that they consume the least possible resources
3. to obtain accurate numerical approximations to quantities which may not have explicit mathematical expressions.

In other words, do it accurately, do it quickly, and do it cheaply.

It might be thought that numerical evaluation consists largely of translating textbook formulas into a **computer programming language** – a mere coding exercise – but this is very far from the case. Direct translations of expressions from mathematical theory are seldom optimal, and very often are found to fail in circumstances when a less obvious numerical process would have succeeded.

The focus on numerical results means that one is not limited to direct expressions, but can evaluate functions which are defined only indirectly, for example through integrals, differential equations, series, or as solutions to equations. An important example is the **maximum likelihood** estimator for a nonlinear statistical model (*see* **Nonlinear Regression**). Indeed, an indirect method is often preferred for numerical evaluation even when a direct expression exists.

While efficiency and accuracy are both aims, it is accuracy which takes precedence, since a slightly slower accurate program is invariably preferred to a faster one with unreliable accuracy. Errors arise from three sources: (i) errors in the input data; (ii) computation errors due to finite precision arithmetic; and (iii) approximation error. The first of these is not under the control of the calculation; in fact it might be considered to be the special concern of the statistician. Computation error appears because of the difference between exact arithmetic and the finite-length arithmetic available on digital computers and hand calculators. Approximation error occurs when the computed expression is not exactly equal to the theoretical quantity even in exact arithmetic. An integral is replaced with a sum, for example, or an infinite series is evaluated only to a finite number of terms.

Efficiency is usually measured by counting basic floating point operations (flops), such as additions, subtractions, multiplications, and divisions. Another consideration is to minimize the use of computer memory and other space requirements, especially for large jobs. More recent concerns which arise from modern **computer architecture** include parallel computing (designing algorithms so that they can be evaluated in parallel streams on fast computers with multiple processors) and local referencing (minimizing unnecessary paging of virtual memory).

It is also desirable to keep programs simple and understandable, thus making the programs easy to maintain and to modify. Users must often choose between using compact programs which can be tinkered with for their own use, and using sophisticated high-performance software from public libraries, which cannot be modified and must be taken somewhat on trust.

Most biostatisticians can benefit from familiarity with numerical analysis. An understanding of the numerical methods being used and an idea of when they will perform well or poorly is necessary even for users of standard statistical package programs (*see* **Software, Biostatistical**). You must still understand the program's purpose and limitations to know whether it applies to your particular situation or not. More importantly, many problems cannot be solved by simple application of a standard program. If you develop your own software, a knowledge of numerical analysis can help avoid numerical pitfalls that can occur easily in a number of problems.

A justifiably popular text on scientific computing is Press et al. [9], which contains a lot of advice on routines to use. Other good general texts on numerical analysis are by Atkinson [3] and Stoer & Bulirsch [12]. An introduction with some statistical orientation is by Thisted [13]. An elegant and elementary introduction to the fundamental ideas of numerical analysis is given by Stewart [11].

The remainder of this article discusses the basic ideas of accuracy and describes briefly key topics in

numerical analysis which are treated at more depth in separate articles. Pointers to available software are given at the end of the article.

Conditioning

The concept of conditioning refers to the intrinsic difficulty of a numerical problem. A problem is ill-conditioned if it is sensitive to perturbations in the data, and well-conditioned if it is not. Conditioning is often quantified by a *condition number* which refers to the amplification of relative errors. Suppose that x is the exact argument to a function f but unfortunately only an approximation $\tilde{x}$ is available. The *condition number* κ of f at x is defined, with respect to a given norm $(||\cdot||)$, by the relation

$$\frac{||f(\tilde{x}) - f(x)||}{||f(x)||} \approx \kappa \frac{||\tilde{x} - x||}{||x||}$$

for $\tilde{x}$ near x. If κ is large, then errors in x are magnified in the evaluation of $f(x)$, while the opposite is true if κ is small. If $\kappa = 10^k$, then k is roughly the number of significant figures of accuracy we can expect to lose in the computation.

For univariate, differentiable functions, the condition number is essentially $\kappa = |xf'(x)/f(x)|$. For example, $f(x) = (x-1)^6$ is an ill-conditioned function near $x = 1$, while $f(x) = x^{1/2}$ is well-conditioned for any $x > 0$.

For general multivariate functions, the specific definition of condition number depends on the problem. For example, the computation of regression coefficients from a **multiple regression** is ill-conditioned when the design matrix **X** displays **collinearity**. Conditioning also depends on the quantity of interest. A **least squares** regression may be ill-conditioned from the point of view of the regression coefficients but well-conditioned from the point of view of the fitted values.

Stability

A stable **algorithm** is one which evaluates a function to the accuracy allowed by the function's condition number. A stable algorithm therefore will evaluate a well-conditioned function accurately, and will do as well as can be expected on an ill-conditioned problem. For example, consider the problem of computing the sample **variance** of the three numbers:

$$62, 63, 64$$

using four-digit decimal arithmetic. A commonly taught formula for the variance is

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right),$$

where n is the sample size and the x_i are the observations. Since the data are given to two significant figures, it might be thought that carrying four significant figures through the calculation will leave a more than adequate safety margin. In this case $\sum x_i^2 = 3844 + 3969 + 4096$, which is 1191×10^1 in 4-digit arithmetic. Similarly, $n\overline{x}^2 = 3(63^2)$ is 1191×10^1 to 4 digits. Therefore, s^2 is computed to be 0, a 100% error compared with the true value of 1. Alternative algorithms are available: for example, $s^2 = \sum(x_i - \overline{x})^2/(n-1)$, which evaluates to $[(-1)^2 + 0^2 + 1^2]/2 = 1$ – the correct answer in this case. The first formula is unstable, while the second formula is stable. There are many other algorithms for computing the sample variance, some of which are of great interest to manufacturers of hand calculators; see Chan et al. [4] for a discussion.

The error in the first formula above arises in the rounding errors of $\sum x_i^2$ and $n\overline{x}^2$, and the error is revealed when the difference is taken of the two large and nearly equal quantities. This is often called *subtractive cancellation*, although rounding error occurred not in the subtraction but in the previous summation. It is a general principle that one cannot add a large value to a floating point number and later subtract it without losing accuracy. One concern, therefore, of numerical analysis is to limit the growth in size of intermediate quantities in calculations. For example, a summation is generally stable if the summands are all of one sign. In this case, the partial sums cannot be greater in absolute value than the final sum.

There is often a close relationship between stability in numerical analysis and in statistics. Frequently, parameters which are statistically interpretable because they measure some invariant characteristic of a problem appear also in a stable algorithm, because of the need to compute quantities which do not grow without bound. Even the small example above gives

an example of this, as the $x_i - \overline{x}$ are the well-known residuals, while in the textbook formula $\sum x_i^2$ and $n\overline{x}^2$ are merely intermediate quantities, not statistically useful quantities in their own right.

Let $\tilde{f}(x)$ be the approximation to $f(x)$ which arises from an algorithm. The algorithm is called *backwardly stable* if $\tilde{f}(x)$ can be shown to be equal to the exact evaluation of f at $\tilde{x}$, where $\tilde{x}$ is close to x. In this way, a (backwardly) stable algorithm will compute a well-conditioned function accurately, and will compute an ill-conditioned function as accurately as is allowed by its conditioning.

Although proving error bounds is an important part of modern numerical analysis, the specific bounds obtained are usually pessimistic and are seldom used in practice. In general, rounding-error analyses are less valued for their final bounds than for the insight they provide about a numerical algorithm. A thorough treatment of rounding-error analyses can be found in Higham [6].

Floating Point Arithmetic

There are an infinite number of real numbers, but only a finite number can be represented on a computer. Therein lies the fundamental difference between exact and computer arithmetic, alluded to above. Numbers are represented on computers in floating point form, i.e. $f \times \beta^e$ in terms of a base β, fraction f, and exponent e. For example,

$$2.597 \times 10^{-3}$$

is a base-10 floating point number with four figures of accuracy. Most computers use base 2, and the resulting arithmetic is called binary arithmetic.

Finite computer arithmetic produces three types of errors. When an arithmetic operation produces a number with an exponent that is too large, the result is said to have *overflowed*. Similarly, an arithmetic operation that produces an exponent that is too small is said to have *underflowed*. Even within the limits of the exponent, most numbers cannot be represented exactly on floating point arithmetic of a fixed word length. The resulting inaccuracy is called *rounding error*. It is a central concern of numerical analysis that rounding errors do not accumulate during a long computation (*see* **Floating Point Arithmetic**).

Linear Equations and Matrix Computations

The theory and practice of solving a linear system

$$\mathbf{Ax} = \mathbf{b}$$

for $\mathbf{x}$, and, more generally, the whole subject of computations involving matrices, is now very well developed (*see* **Matrix Computations**). Here, we outline two applications of interest to biostatisticians.

In least squares regression of a response vector $\mathbf{y}$ on a design matrix $\mathbf{X}$, numerical analysts have influenced statisticians to move away from the normal equations for the regression coefficients in favor of methods based on the decomposition

$$\mathbf{X} = \mathbf{QR},$$

where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{R}$ is upper triangular. This is because the $\mathbf{QR}$ approach is backwardly stable, while the normal equations are not.

Conditioning for the least squares problem is determined by that of $\mathbf{X}$, which can be analyzed through the singular value decomposition

$$\mathbf{X} = \mathbf{UDV}^{\mathrm{T}},$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{D}$ is diagonal containing the singular values. The condition number of $\mathbf{X}$ is usually defined to be the ratio of the largest to the smallest singular value. If the columns of $\mathbf{X}$ are standardized; say, by dividing by the sample standard deviation of the column, then the singular values entirely capture the idea of ill-conditioning and collinearity for the least squares problem. The singular value decomposition therefore gives statisticians the means to quantify collinearity, and there are those who propose its routine use in regression computations for that reason [9, Section 15.4].

Numerical linear algebra is dealt with in more detail in the article on **Matrix Computations**.

Optimization and Nonlinear Equations

Optimization means to find that value of $\mathbf{x}$ which maximizes or minimizes a given function $f(\mathbf{x})$. This is a central concern in statistics, because statistical **estimation** principles such as least squares, maximum likelihood, posterior mode (*see* **Bayesian Methods**) and M-estimation (*see* **Robustness**) are defined in

terms of optimizing an appropriate objective function. Numerical optimization strategies come into play when the statistical model is nonlinear and analytic estimators of the parameters are not available.

A closely related problem is that of solving nonlinear equations. Many algorithms for optimizing $f(\mathbf{x})$ are, in fact, derived from algorithms for solving $\partial f/\partial \mathbf{x} = 0$, where $\partial f/\partial \mathbf{x}$ is the derivative vector of f with respect to $\mathbf{x}$.

Details are given in the article on **Optimization and Nonlinear Equations**.

Interpolation and Approximation

The purpose here is accurately to approximate complex functions with ones which are easy to evaluate. For example, rational function approximations to the standard normal distribution function and its inverse allow it to be computed rapidly within statistical programs. Typical methods include series expansions, rational functions, and polynomials (see, for example, Press et al. [9, Chapter 5] and **Polynomial Approximation**). A great many approximation formulas are given in Abramowitz & Stegun [1].

Numerical Integration

After matrix computations, numerical integration is one of the largest areas of numerical analysis. A large number of sophisticated and reliable methods are available for numerical integration in one dimension. Unfortunately, for statisticians wanting to evaluate mixture models or Bayesian marginal posteriors, the picture is less clear in high dimensions. Statisticians have made a substantial contribution to high-dimensional integration through the development of efficient **Monte Carlo** methods. A survey of integration methods is given in the article on **Numerical Integration**.

Available Software

The final goal of numerical analysis is to make numerical methods generally available through high-quality portable software. Numerical analysts were also early users of the **internet**, and a wide range of software is available online. Netlib is the most extensive collection of numerical programs. Its URL is `http://www.netlib.org`.

Worthy of special mention are the LINPACK library [5] for linear algebra and the EISPACK library [10] for **eigenvalue** computations, both from the Argonne National Laboratory. These are published, documented and freely available, and have gained wide acceptance by statisticians and other scientists. The two libraries have now been combined and updated as LAPACK [2]. Other libraries of note include the QUADPACK library [8] for numerical integration, and the SLATEC library – an enormous library of FORTRAN programs.

The Guide to Available Mathematical Software (GAMS) at `http://gams.nist.gov` provides a virtual database of documented and supported programs, searchable by program and problem type. The journal *ACM Transactions of Mathematical Software* is a source of refereed software, also searchable by GAMS.

Commercial subroutine libraries include the NAG Library (Numerical Algorithms Group) and the IMSL Mathematics and Statistics Libraries. LINPACK, EISPACK, and other routines have also been incorporated into the interactive matrix programming language, MATLAB [7].

Another popular commercial source is Numerical Recipes [9], accessible through `www.nr.com`. Numerical Recipes supplies smaller, understandable programs, which may be modified by users for specific applications. Netlib, GAMS, NAG, and IMSL provide more sophisticated routines designed for high performance on large problems. Considerable effort has been expended to make the high-performance routines efficient, memory-compact, and capable of trapping most errors.

Programs developed by statisticians, dealing specifically with statistical problems, can be found at Statlib, the statistical database maintained at Carnegie-Mellon University. The URL is `http://lib.stat.cmu.edu`.

References

[1] Abramowitz, M. & Stegun, I.A. (1962). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington. Reprinted by Dover, New York, 1965.

[2] Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., DuCroz, J., Greenbaum, A., Hammerling, S., McKenney, A., Ostrouchov, S. & Sorensen, D. (1995). *LAPACK Users' Guide*, Release 2.0, 2nd Ed. SIAM Publications, Philadelphia.

[3] Atkinson, K.E. (1989). *An Introduction to Numerical Analysis*, 2nd Ed. Wiley, New York.

[4] Chan, T., Golub, G. & Leveque, R. (1983). Algorithms for computing the sample variance: analysis and recommendations, *American Statistician* **37**, 242–247.

[5] Dongarra, J.J. et al.(1979). *LINPACK Users' Guide*. SIAM, Philadelphia.

[6] Higham, N.J. (1996). *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia.

[7] Moler, C., Little, J. & Bangert, S. (1987). *Pro-Matlab User's Guide*. The Math Works, Sherborn.

[8] Piessens, R., De Doncker-Kapenga, E., Überhuber, C.W. & Kahaner, D.K. (1983). *Quadpack, a Subroutine Package for Automatic Integration*. Springer-Verlag, Berlin.

[9] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in Fortran*. Cambridge University Press, Cambridge.

[10] Smith, B.T., Boyle, J.M., Ikebe, Y., Klema, V.C. & Moler, C.B. (1970). Matrix Eigensystem Routines: EISPACK Guide, 2nd Ed., in *Lecture Notes in Computer Science*, Vol. 6. Springer-Verlag, New York.

[11] Stewart, G.W. (1996). *Afternotes on Numerical Analysis*. Society for Industrial and Applied Mathematics, Philadelphia.

[12] Stoer, J. & Bulirsch, R. (1993). *Introduction to Numerical Analysis*, 2nd Ed. Springer-Verlag, New York.

[13] Thisted, R.A. (1988). *Elements of Statistical Computing. Numeric Computation*. Chapman & Hall, New York.

(*See also* **Computer Algebra**)

GORDON K. SMYTH

Numerical Integration

Numerical integration is the study of how the numerical value of an integral can be found. Also called *quadrature*, which refers to finding a square whose area is the same as the area under a curve, it is one of the classical topics of **numerical analysis**. Of central interest is the process of approximating a definite integral from values of the integrand when exact mathematical integration is not available. The corresponding problem for multiple dimensional integration is known as multiple integration or *cubature*.

Numerical integration has always been useful in biostatistics to evaluate distribution functions and other quantities. Emphasis in recent years on **Bayesian** and **empirical Bayesian** methods and on mixture models has greatly increased the importance of numerical integration for computing **likelihoods** and posterior distributions and associated **moments** and derivatives. Many recent statistical methods are dependent especially on multiple integration, possibly in very high dimensions.

Although there exist many high-quality automatic integration programs, no program can be expected to integrate all functions, even in one dimension. It is therefore useful for the user to know something about the limitations of the commonly used methods.

This article describes classical quadrature methods and, more briefly, some of the more advanced methods for which software is widely available. The description of the elementary methods in this article borrows from introductory notes by Stewart [31]. An excellent general reference on numerical integration is [5]. More recent material can be found in [7] and [29]. Recent surveys of numerical integration with emphasis on statistical methods and applications are [9] and [8].

Trapezoidal Rule

The simplest quadrature rule in wide use is the *trapezoidal rule*. Like many other methods, it has both a geometric and an analytic derivation. The idea of the geometric derivation is to approximate the area under the curve $y = f(x)$ from $x = a$ to $x = b$ by the area of the trapezoid bounded by the points $(a, 0)$, $(b, 0)$, $[a, f(a)]$, and $[b, f(b)]$. This gives

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}[f(a) + f(b)].$$

The analytic derivation is to interpolate $f(x)$ at a and b by a linear polynomial.

The trapezoidal rule cannot be expected to give accurate results over a larger interval. However, by summing the results of many applications of the trapezoidal rule over smaller intervals, we can obtain an accurate approximation to the integral over any interval. We begin by dividing $[a, b]$ into n equal intervals by the points $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$. Specifically, if $h = (b-a)/n$ is the common length of the intervals, then $x_i = a + ih, i = 0, 1, \ldots, n$. Applying the trapezoidal rule to each interval $[x_{i-1}, x_i]$ gives the *composite trapezoidal rule*

$$\int_a^b f(x)\,\mathrm{d}x \approx h\left\{\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{n-1}) + \frac{f(x_n)}{2}\right\}.$$

An error formula for the composite trapezoidal rule can be obtained from polynomial approximation theory. If f is twice continuously differentiable on (a, b), then the error of integration decreases as h^2, so that doubling the number of points reduces the error by a factor of four.

Simpson's Rule

More sophisticated quadrature rules can produce higher-order error terms. Even more popular than the trapezoidal rule is *Simpson's rule*:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

Simpson's rule can be derived by interpolating $f(x)$ by a quadratic polynomial at a, $(a+b)/2$, and b.

As with the trapezoidal rule, Simpson's rule is usually applied to many short intervals. Letting the x_i be as above for n even, and writing $f_i = f(x_i)$ the *composite Simpson rule* is

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{n-2} + 4f_{n-1} + f_n).$$

If $\mathrm{CS}_h(f)$ denotes the result of applying the composite Simpson rule to f over the interval $[a, b]$, and if f has a continuous fourth derivative on (a, b), then

$$\int_a^b f(x)\,\mathrm{d}x - \mathrm{CS}_h(f) = -\frac{(b-a)f^{(4)}(\xi)}{180}h^4$$

for some $\xi \in [a, b]$. Although Simpson's rule was derived to integrate quadratic polynomials exactly on each interval, the presence of the fourth derivative in the error term signals that it in fact integrates cubics exactly as well. This property follows from the fact that Simpson's rule is a special case of Gaussian quadrature, treated below.

Newton–Cotes Formulas

The trapezoidal rule integrates any linear polynomial exactly. In general, we might look for an $(n+1)$-point rule which integrates exactly any polynomial of degree n. Such a quadrature rule is the *Newton–Cotes formula*.

Let $x_0, x_1, \ldots, x_n$ be distinct points in the interval $[a, b]$. We wish to determine constants $A_0, A_1, \ldots, A_n$ such that

$$\int_a^b f(x)\,\mathrm{d}x = A_0 f(x_0) + \cdots + A_n f(x_n)$$

for any polynomial f of degree $\leq n$. Strictly speaking, as Newton–Cotes usually refers to formulas with equally spaced abscissas, this is a slight generalization.

Although there is an elegant analytic expression for the A_i in terms of *Lagrange polynomials* [31], they are difficult to evaluate stably. For rules of low degree, one can substitute in $f(x) = 1$, $f(x) = x$, $f(x) = x^2$, etc. to obtain a system of linear equations which can be solved for the A_i.

Clenshaw–Curtis Integration

Newton–Cotes formulas with equally spaced abscissas are of practical use only for small point numbers, say $n \leq 8$. For n as low as nine, the coefficients A_i vary in sign. As n increases, the coefficients become large in absolute value, leading to unstable evaluation of the integral. This problem can be avoided by choosing the abscissas in a more sophisticated way. One choice for which the coefficients are not only positive but have stable analytic expressions is the Chebyshev points on $[a, b]$,

$$x_i = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{i\pi}{n}\right), \quad i = 0, 1, \ldots, n.$$

Define the modified Fourier coefficients,

$$a_j = \frac{2}{n}{\sum_{i=0}^{n}}'' f(x_i)\cos\left(\frac{ij\pi}{n}\right),$$

where $''$ indicates that the first and last terms in the sum are to be halved. If n is even, then the *Clenshaw–Curtis formula* can be written

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\left[a_0 - \frac{2a_2}{(1)(3)} - \frac{2a_4}{(3)(5)} - \cdots - \frac{2a_{n-2}}{(n-3)(n-1)} - \frac{a_n}{(n-1)(n+1)}\right].$$

Like other formulas of the Newton–Cotes type, Clenshaw–Curtis will integrate exactly polynomials of order n or less. In practice, it does rather better than other rules of the same order, because of the bounded variation properties of Chebyshev polynomials. The error of Clenshaw–Curtis integration can be estimated from the rate of decrease of the coefficients a_j. O'Hara & Smith [22] suggest the use of bounds such as $\max(2|a_{n-4}|, 2|a_{n-2}|, |a_n|)$ for the approximation error.

Treatment of Singularities

Provided that the integrand f is sufficiently smooth, the Newton–Cotes formulas converge as $n \to \infty$. It sometimes happens, however, that one has to integrate a function with a singularity. Suppose, for example, that, for x near zero,

$$f(x) \approx \frac{c}{x^d}$$

for some constant c and $0 < d < 1$. Then $\int_0^1 f(x)\,\mathrm{d}x$ exists, but the Newton–Cotes formulas will not obtain good results because f is not at all polynomial on $[0, 1]$. A better approach is to incorporate the singularity into the quadrature rule itself.

First define

$$g(x) = x^d f(x),$$

and look for a rule that evaluates the integral

$$\int_0^1 g(x)x^{-d}\,\mathrm{d}x,$$

where g is a well-behaved function on [0, 1]. The function x^{-d} is called a *weight function*. Given any modest number of points $x_0, \ldots, x_n$ in the interval (0,1], the method of undetermined coefficients can easily determine an integration rule of the form

$$\int_0^1 g(x)x^{-d}\,\mathrm{d}x = A_0 f(x_0) + \cdots + A_n f(x_n)$$

by substituting in $g(x) = 1, g(x) = x, g(x) = x^2$, etc.

The appearance of derivatives in the error terms for Newton–Cotes rules (and for the Gaussian rules below) shows that the method is troubled not only by singularities in the integrand, but by singularities in its derivatives as well. A weight function may therefore need to remove singularities in the derivatives as well as in the function itself.

Gaussian Quadrature

A polynomial of degree n is determined by its $n + 1$ coefficients. We have seen that the $n + 1$ coefficients $A_0, \ldots, A_n$ in the $(n + 1)$-point Newton–Cotes formula can be chosen to make the rule exact for polynomials of degree n or less. The idea behind Gaussian quadrature is that the abscissas $x_0, \ldots, x_n$ represent another $n + 1$ degrees of freedom, which may be used to extend the exactness of the rule to polynomials of degree $2n + 1$.

Gauss quadrature formulas have the form

$$\int_a^b f(x)w(x)\,\mathrm{d}x \approx A_0 f(x_0) + \cdots + A_n f(x_n),$$

where $w(x)$ is a weight function which is greater than zero on the interval $[a, b]$. The correct choice for $x_0, \ldots, x_n$ turns out to be the zeros of an orthogonal polynomial $P_{\{n+1\}}$ of order $n + 1$. An important point is that the coefficients A_i are positive. Moreover, $A_0 + A_1 + \cdots + A_n = \int_a^b w(x)\,\mathrm{d}x$, so no coefficient can be larger than $\int_a^b w(x)\,\mathrm{d}x$. Consequently, we cannot have a situation in which large coefficients create large intermediate results that suffer cancellation when they are added.

Gaussian quadrature has error formulas similar to those for Newton–Cotes formulas. Specifically, if f is $2n + 2$ times continuously differentiable on (a, b), and $G_n f$ is the quadrature approximation, then

$$\int_a^b f(x)w(x)\,\mathrm{d}x - G_n f = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b p_{n+1}^2(x)w(x)\,\mathrm{d}x,$$

where $\xi \in [a, b]$. If f does not satisfy the smoothness property, then the accuracy of Gaussian quadrature is generally reduced by at least an order of magnitude. However, it is a consequence of the positivity of the coefficients A_i that Gaussian quadrature converges for any continuous function as $n \to \infty$.

Particular Gauss formulas arise from particular choices of the interval $[a, b]$ and the weight function $w(x)$. The workhorse is *Gauss–Legendre* quadrature in which $[a, b] = [-1, 1]$ and $w(x) = 1$, so that the formula approximates the integral

$$\int_{-1}^1 f(x)\,\mathrm{d}x.$$

The corresponding orthogonal polynomials are called Legendre polynomials.

If we take $[a, b] = [0, \infty)$ and $w(x) = \mathrm{e}^{-x}$, we get a formula to approximate

$$\int_0^\infty f(x)\mathrm{e}^{-x}\,\mathrm{d}x.$$

This is called *Gauss–Laguerre* quadrature.

If we take $[a, b] = [-\infty, \infty]$ and $w(x) = \mathrm{e}^{-x^2}$, we get a formula to approximate

$$\int_{-\infty}^\infty f(x)\mathrm{e}^{-x^2}\,\mathrm{d}x$$

This is *Gauss–Hermite* quadrature.

Computing the abscissas and coefficients for these and other Gauss rules in a stable and efficient manner is a challenging nonlinear problem. Two successful **algorithms** are those of Golub & Welsch [12] and Sack & Donovan [28]. A FORTRAN program implementing the Golub–Welsch method can be obtaining by searching the NETLIB* database for GAUSSQ. The expense of computing the abscissas and coefficients is sufficiently great that they are usually stored and reused rather than generated afresh for each problem.

Simpson's rule is actually a variant of the Gauss–Legendre three-point rule in which x_0 and x_n are constrained to be the end points. Rules with such constraints are called *Gauss–Radau* or *Gauss–Lobatto* quadrature [5].

Progressive Formula

Despite their optimal properties, the Gaussian formulas are not universally used in practice. The main reason for this is the difficulty of determining in advance the required number of points to achieve a given level of accuracy. In some cases, mathematical analysis of the function to be integrated makes it possible to use the analytic error bounds of the quadrature rules. It is more common, however, to estimate the error empirically by applying the same quadrature rule twice with different point numbers. Often the point number is doubled until the successive values of the integral agree to the required number of figures.

A succession of integration formulas with increasing point numbers is said to be *nested* or *progressive* if each formula reuses the abscissas of the earlier formulas. The composite Simpson and Clenshaw–Curtis rules with n doubling at each step are important examples of progressive formulas. Gaussian formulas are generally not progressive, as the abscissas at any point number are different from those for any other point number. The relative advantage of the Gauss formulas is therefore lost in the expense of computing addition abscissas and function evaluations.

One possibility is to construct progressive formulas starting or finishing with a Gaussian formula. Kronrod [18] gave a method for adding points to a Gauss–Legendre formula in an optimal way. The Kronrod rule adds $n+1$ points to a n-point Gauss–Legendre formula, resulting in a rule which integrates exactly polynomials of order $3n+1$ (n even) or $3n+2$ (n odd). The desirable properties of Gaussian quadrature are preserved in that the abscissas remain in the integration interval and the coefficients A_i remain positive. When the n-point Gauss rule is combined with its Kronrod optimal extension, a very economical pair of formulas result for the simultaneous calculation of an approximation for an integral and the respective error estimate. The problem of extending arbitrary quadrature formulas in a progressive fashion was studied by Patterson [17, 23, 24], who also gave a stable computation for the Kronrod rules. Together, the Kronrod and Patterson methods provide a nested sequence of quadrature rules based on an initial Gauss rule, and are the basis of some of the most widely used integration programs.

Adaptive Methods

A quadrature rule is *adaptive* if it compensates for a difficult subrange of an integrand by automatically increasing the number of quadrature points in the awkward region. Adaptive strategies divide the integration interval into subintervals and, typically, employ a progressive formula in each subinterval with some fixed upper limit on the number of points allowed. If the required accuracy is not achieved by the progressive formula, then the subinterval is bisected and a similar procedure carried out on each half. This subdivision process is carried out recursively until convergence is achieved in each of the terminating subintervals. Most general purpose integration programs are adaptive, since such a strategy can be successful over a very wide range of integrands.

Multiple Integration: Product Rules

Multiple integration is concerned with the numerical approximation of integrals of two or more variables. It is not a simple extension of one-dimensional integration. The diversity of possible integration regions and singularities for d-dimensional functions is daunting. As a general rule, it is not possible to obtain the same accuracy with higher-dimensional integrals as with one-dimensional integrals for reasonable computing times.

The problem addressed by multiple integration is to evaluate integrals of the form

$$\int f = \int_{a_d}^{b_d} \int_{a_{d-1}(x_d)}^{b_{d-1}(x_d)} \cdots \int_{a_1(x_1,\ldots,x_d)}^{b_1(x_1,\ldots,x_d)} f(x_1, x_2, \ldots, x_d) \times \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_d.$$

The most obvious approach is to treat the multiple integral as a nested sequence of one-dimensional integrals, and to use one-dimensional quadrature with respect to each argument in turn. The resulting multiple integration formula is a *product rule*.

Suppose that the integration region is a hyperrectangle, so that the integration interval $[a_j, b_j]$

for x_j in the above integral is independent of $x_{j+1}, \ldots, x_d$. If Gauss quadrature is used to integrate f with respect to x_j, with abscissas $x_{j0}, x_{j1}, \ldots, x_{jn}$ and coefficients $A_{j0}, A_{j1}, \ldots, A_{jn}$, then the product rule is

$$\int f \approx \sum_{i_0, i_1, \ldots, i_d = 0}^{n} A_{0i_0} A_{1i_1} \ldots A_{di_d} \times f(x_{0i_0}, x_{1i_1}, \ldots, x_{di_d}).$$

This rule integrates exactly any sum of monomials $x_1^{\alpha} x_2^{\beta} \ldots x_n^{\gamma}$, where each $\alpha, \beta, \ldots, \gamma$ is an integer between zero and $2n + 1$, a result which derives directly from the corresponding result for the one-dimensional Gauss rule.

The number of evaluations of f in the product rule is $(n + 1)^d$, which grows exponentially with d. Rapid growth in the number of function evaluations usually limits the practical use of product rules to around five or six dimensions. One special case common in statistical applications is that in which $x_1, \ldots, x_d$ are exchangeable. This arises when $x_1, \ldots, x_d$ is an independent sample from some distribution and f is a function of the probability density. In that case only one evaluation of f is needed for all points which are permutations of one another. The total number of evaluations required is then $\binom{n+d}{d}$, which is considerably smaller than $(n + 1)^d$, so that calculations for sample sizes up about 10 are manageable.

Despite the above limitations, Gauss product rules have been the basis of at least one general approach to implementing Bayesian analysis methods, discussed in [20] and [30].

Rules of Polynomial Degree

As with quadrature, most cubature rules are designed to integrate a certain class of polynomials exactly. A rule is said to be of polynomial degree r if it integrates exactly any sum of monomials $x_1^{k_1} \ldots x_d^{k_d}$ with $k_1 + \cdots + k_d \leq r$. Although Gauss product rules integrate certain monomials of higher order, they do not integrate x_j^{2n+2} exactly and are therefore of polynomial degree $2n + 1$.

By allowing rules that are not product rules, it is usually possible to find rules which are more efficient than the Gauss product rules in the sense of having polynomial degree $\geq 2n + 1$ yet requiring fewer than $(n + 1)^d$ points. Methods for constructing rules of prescribed polynomial degree are surveyed in [3]. For a compilation of such rules see [32] and [4].

Polynomial rules of degrees five and seven on the hyper-rectangle serve as basic integrating rules for the popular multiple integration program ADAPT [11], which is described further below.

Globally Adaptive Algorithms

One-dimensional adaptive programs usually consider each subinterval in turn, subdividing each until a specified accuracy is obtained. This straightforward strategy is called *locally adaptive* because the behavior of the algorithm in each local subinterval depends only on the error estimates in that interval. However, for multiple integrals it is often unknown at the beginning of the calculation whether the given accuracy can be obtained in a reasonable amount of time. A popular adaptive strategy, originally proposed by van Dooren & de Ridder [33], always subdivides the integration subregion with the largest error. Such a strategy is known as *globally adaptive* because it makes subdivision decisions using information about all the current subregions. Although globally adaptive algorithms require more memory space to maintain the current subregion list and take more time to select subregions for subdivision, at each stage in the calculation the global estimate for $\int f$ is in some sense the best one available using the computation that has been done so far.

The globally adaptive program ADAPT [11] and its successor DCUHRE [1, 2] build on the work of van Dooren & de Ridder [33]. ADAPT uses the difference between nested pairs of polynomial rules, of degrees seven and five, respectively, to estimate the error in each subregion. Some of the degree seven integrand values are also used to compute fourth differences in directions parallel to each of the coordinate axes. When a subregion is selected for subdivision, it is divided in half in the direction of largest absolute fourth difference. This clever strategy for halving in only one direction, using fourth differences to measure integrand irregularity, is probably one of the main reasons for the practical effectiveness of the algorithm. The later program DCUHRE gives the user a choice of integration rules, uses a more sophisticated error estimate, and is organized to facilitate parallel integration of a vector of related integrands.

Lattice Methods

Lattice rules were originally called "number theoretic" or "quasi-random" methods [32]. The integration region is translated to the unit cube, and the integral approximated by a multiple sum of the form

$$Qf = \frac{1}{n_1 n_2 \ldots n_t} \times \sum_{j_t=0}^{n_t-1} \cdots \sum_{j_1=0}^{n_1-1} f\left(\frac{j_1}{n_1}\mathbf{z}_1 + \cdots + \frac{j_t}{n_t}\mathbf{z}_t\right),$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_t$ are carefully selected integer vectors. This is the simple unweighted mean of the integrand evaluated over a regular lattice of abscissas in the unit cube. For the method to work well, the integrand must be transformed to be periodic in the cube so that its Fourier coefficients go to zero rapidly.

A lattice method originated by Korobov [16] and extended by Patterson & Cranley [25] is implemented in the NAG library* routines D01GCF and D01GDF. Lattice methods are not yet in widespread use, but there is some evidence [10, 29] that they can outperform other available methods when the number of dimensions is between about 10 and 20.

Monte Carlo Methods

The idea of estimating an integral by random sampling is a natural one in a statistical context. In the classical **Monte Carlo method** [13, 19], points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are chosen randomly in the integration region and the integral is estimated by

$$\overline{f} = \frac{V}{n}\sum_{i=1}^{n} f(\mathbf{x}_i),$$

where V is the volume of the integration region. Convergence is guaranteed almost surely by the central limit theorem under very weak conditions on f. Moreover, the rate of convergence is independent of the dimensionality. The error $\overline{f} - \int f$ is approximately normal with mean zero and standard deviation

$$\frac{\sigma(f)}{\sqrt{n}}V,$$

where

$$\sigma^2(f) = \frac{1}{V}\int f^2 - \left(\frac{1}{V}\int f\right)^2$$

is the variance of f. Finally, and most importantly, a free estimate of the error is available as $\sigma^2(f)$ may be estimated by the sample variance of $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$.

The slow $n^{-1/2}$ rate of convergence means that Monte Carlo methods are usually limited to low accuracy; say, three significant figures. However, this accuracy can be achieved with comparable work for any number of dimensions and for a very wide range of integration regions. In many statistical applications higher accuracy is not required; computational error need only be small relative to the inherent statistical uncertainty that enters the process of drawing inferences from data.

Practical use of the Monte Carlo method depends on techniques for reducing the $\sigma^2(f)$ variance term in the error. Central amongst these is *importance sampling*, in which $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are sampled from a distribution which is as much like f in shape as possible. This has the effect of sampling most densely in those parts of the integration region where the integrand is greatest. Specifically, write the integral as

$$\int f(\mathbf{x})g(\mathbf{x})\,\mathrm{d}x,$$

where g is a density function on the integration region, and f is as close to a constant function as possible. Points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are sampled from g, and the integral is approximated as before by $\overline{f}$. The standard deviation of $\overline{f}$ is now

$$\frac{\sigma_g(f)}{\sqrt{n}},$$

where

$$\sigma_g^2(f) = \int f^2 g - \left(\int fg\right)^2$$

is the variance of f with respect to the density g.

Other variance reduction techniques include **stratified sampling** and **antithetic** acceleration. Antithetic acceleration involves generating pairs of identically distributed but negatively correlated points $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$. This tends to produce negatively correlated terms in the sum; the more negative the correlation, the lower the variance of the sum. See [5], [13], [15], or [8] for references to variance reduction methods. The use of Monte Carlo integration to solve Bayesian problems is treated more fully in the article, **Markov chain Monte Carlo**.

The NAG library subroutine D01GBF uses an adaptive Monte Carlo algorithm to integrate over a hyper-rectangle. The number of subregions is doubled at each iteration, and in each the integral and variance are estimated by Monte Carlo sampling. Algorithms also exist which are adaptive in terms of the importance sampling density. Such algorithms refine the importance sampling density adaptively so as to minimize σ^2 during the Monte Carlo process [5, 21].

Conclusions

If a large number of well-behaved one-dimensional integrands are to be integrated, and the user is willing to do some analytic analysis to obtain efficiency, then it is hard to go past the classical Gauss quadrature methods. More usually, though, users will choose to use an automatic integration program of some kind, using computer time to save their own time and to gain reliability.

Reliable and well-documented software for numerical integration can be found by searching the NIST GAMS online catalogue at `http://gams.nist.gov` under class "h2". See [14] for brief reviews of much of this software. It is also worth searching the STATLIB database for statistical functions based on these routines. Simple integration programs, suitable for modification by users, can be found in [27]. Most major statistical and mathematical programming languages include numerical integration programs, often based on the programs found in GAMS.

In one and two dimensions there is a wealth of reliable and effective programs. The leading one-dimensional package currently is QUADPACK by Piessen et al. [26]. This is available from the NETLIB database and is cross-classified by GAMS. It has also been incorporated into the NAG, IMSL, and SLATEC subroutine libraries. QUADPACK provides a suite of programs designed for different types of difficulties, such as singularities and oscillatory integrands, and includes a decision tree to guide the user in choosing the appropriate routine. The program QAGS is a particularly robust general purpose integration program, as is the non-QUADPACK program CADRE [6] which is included in the IMSL library. In statistical applications, however, the integrands are often smooth with a single dominant peak, so the more efficient programs QNG and QAG, which use higher-order Gauss, Gauss–Kronrod and Patterson rules, may suffice.

So far there is no reliable suite of programs for multiple integration. Up to 10 or perhaps 15 dimensions, globally adaptive routines such as ADAPT and DCUHRE can be recommended. When the number of dimensions exceeds about 20, Monte Carlo methods are the only ones possible. Mark 20 of the NAG library includes 10 multiple integration programs, including one which implements a Monte Carlo method.

References

[1] Bernsten, J., Espelid, T.O. & Genz, A. (1991). Algorithm 698: DCUHRE: an adaptive multidimensional integration routine for a vector of integrals, *ACM Transactions on Mathematical Software* **17**, 452–456.

[2] Bernsten, J., Espelid, T.O. & Genz, A. (1991). An adaptive algorithm for the approximate calculation of multiple integrals, *ACM Transactions on Mathematical Software* **17**, 437–451.

[3] Cools, R. (1992). A survey of methods for constructing cubature formulae, in *Numerical Integration: Recent Developments, Software and Applications*, T.O. Espelid & A.C. Genz, eds. Kluwer, Dordrecht, pp. 1–24.

[4] Cools, R. & Rabinowitz, P. (1993). Monomial cubature rules since "Stroud": a compilation, *Journal of Computational and Applied Mathematics* **48**, 309–326.

[5] Davis, P.J. & Rabinowitz, P. (1984). *Methods of Numerical Integration*. Academic Press, New York.

[6] de Boor, C. (1971). CADRE: An algorithm for numerical quadrature, in *Mathematical Software*, J.R. Rice, ed. Academic Press, New York, pp. 417–449.

[7] Evans, G. (1993). *Practical Numerical Integration*. Wiley, New York.

[8] Evans, M. & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems, *Statistical Science* **10**, 254–272.

[9] Flournoy, N. & Tsutakawa, R.K., eds. (1991). *Statistical Multiple Integration, Contemporary Mathematics 115*. American Mathematical Society, Providence.

[10] Genz, A. (1984). Testing multiple integration software, in *Tools, Methods and Languages for Scientific and Engineering Computation*, B. Ford, J.-C. Rault & F. Thomasset, eds. North-Holland, New York, pp. 208–217.

[11] Genz, A. & Malik, A.A. (1980). An adaptive algorithm for numerical integration over an N-dimensional rectangular region, *Journal of Computational and Applied Mathematics* **6**, 295–302.

[12] Golub, G.H. & Welsch, J.H. (1969). Calculation of Gaussian quadrature rules, *Mathematics of Computation* **23**, 221–230.

[13] Hammersley, J.M. & Handscomb, D.C. (1964). *Monte Carlo Methods*. Methuen, London.

[14] Kahaner, D.K. (1991). A survey of existing multidimensional quadrature routines, in *Statistical Multiple Integration, Contemporary Mathematics 115*, N. Flournoy & R.K. Tsutakawa, eds. American Mathematical Society, Providence, pp. 9–22.

[15] Kalos, M.H. & Whitlock, P.A. (1986). *Monte Carlo Methods*, Vol. 1. Basics. Wiley, New York.

[16] Korobov, N.M. (1959). The approximate calculation of multiple integrals, *Doklady Akademii Nauk SSSR* **124**, 1207–1210 (in Russian).

[17] Krogh, F.T. & van Snyder, W. (1991). Algorithm 699: a new representation of Patterson's quadrature formulae, *ACM Transactions on Mathematical Software* **17**, 457–461.

[18] Kronrod, A.S. (1965). *Nodes and Weights of Quadrature Formulas*. Consultants Bureau, New York (authorized translation of the Russian).

[19] Metropolis, N.C. & Ulam, S.M. (1949). The Monte Carlo method, *Journal of the American Statistical Association* **44**, 335–341.

[20] Naylor, J.C. & Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions, *Applied Statistics* **31**, 214–225.

[21] Oh, M.-S. (1991). Monte Carlo integration via importance sampling: dimensionality effect and an adaptive algorithm, in *Statistical Multiple Integration, Contemporary Mathematics 115*, N. Flournoy & R.K. Tsutakawa, eds. American Mathematical Society, Providence, pp. 165–187.

[22] O'Hara, H. & Smith, F.H. (1968). Error estimation in the Clenshaw-Curtis quadrature formulae, *Computing Journal* **11**, 213–219.

[23] Patterson, T.N.L. (1968). The optimum addition of points to quadrature formulae, *Mathematical Computation* **22**, 847–856.

[24] Patterson, T.N.L. (1989). An algorithm for generating interpolating quadrature rules of the highest degree of precision with preassigned nodes for general weight functions, *ACM Transactions on Mathematical Software* **15**, 123–136.

[25] Patterson, T.N.L. & Cranley, R. (1976). Randomization of number theoretic methods for multiple integration, *SIAM Journal Numerical Analysis* **13**, 904–914.

[26] Piessens, R., De Doncker-Kapenga, E., Überhuber, C.W. & Kahaner, D.K. (1983). *Quadpack, a Subroutine Package for Automatic Integration*. Springer-Verlag, Berlin.

[27] Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1986). *Numerical Recipes*. Cambridge University Press, Cambridge.

[28] Sack, R.A. & Donovan, A.F. (1972). An algorithm for Gaussian quadrature given modified moments, *Numerical Mathematics* **18**, 465–478.

[29] Sloan, I.H. & Joe, S. (1994). *Lattice Methods for Multiple Integration*. Clarendon Press, Oxford.

[30] Smith, A.F.M., Skene, A.M., Shaw, J.E.H. & Dransfield, M. (1985). The implementation of the Bayesian paradigm, *Communication in Statistics – Theory and Methods* **14**, 1079–1102.

[31] Stewart, G.W. (1996). *Afternotes on Numerical Analysis*. Society for Industrial and Applied Mathematics, Philadelphia.

[32] Stroud, A.H. (1971). *Approximate Calculation of Multiple Integrals*. Prentice-Hall, Englewood Cliffs.

[33] van Dooren, P. & de Ridder, L. (1976). An adaptive algorithm for numerical integration over an N-dimensional cube, *Journal of Computational Applied Mathematics* **2**, 207–217.

(*See also* **Polynomial Approximation**)

GORDON K. SMYTH

Nursing

Most people will have some idea of what nursing is, perhaps believing it is the general care of sick people, as opposed to medical attention to their disease. In fact, the definition of nursing has been and continues to be a topic of much debate in the nursing profession. **Florence Nightingale** [6], in her work *Notes on Nursing – What It Is And What It Is Not*, said, in essence, that what nursing has to do is "...put the patient in the best condition for nature to act upon him". A more recent and widely accepted definition of nursing is that of Virginia Henderson [4]: "The unique function of the nurse is to assist the individual, sick or well, in performance of those activities contributing to health or its recovery (or to a peaceful death) that he would perform unaided if he had the necessary strength, will or knowledge. And to help him gain independence as rapidly as possible."

History of the Nursing Profession

Nursing was not suddenly invented during the Crimean War. In the middle ages several religious orders provided and staffed hospitals caring for the sick and needy. The world's first school of nursing was established by Theodore Fliedner at Kaiserswerth in 1833, and Florence Nightingale undertook a course there in 1851.

In 1854, Florence Nightingale led, at the government's expense, a party of nurses to Scutari to work in the British Army hospital treating the sick and wounded of the Crimean War. From the inception of this mission, Florence Nightingale was to be an administrator. Woodham-Smith [7] suggests it was not as an angel of mercy that she was asked to go to Scutari; the consideration of overwhelming importance was the opportunity to advance the cause of nursing. Florence Nightingale returned to England in 1856 as a national hero. The position of nursing as a profession was now established.

In 1860, Florence Nightingale established, at St. Thomas's Hospital London, a school of nursing that became a model for schools of nursing everywhere. Since that time the nursing profession and nurse training have continued to change and develop as health care advances. Nurses work as advanced specialists in all clinical areas and have an increasingly important role in the community, in general practice, and in health education.

Nursing Research

Florence Nightingale is often regarded not only as the founder of the nursing profession but also as the first nurse researcher. Her research influenced health care in general, and nursing specifically. Nightingale's *Notes on Nursing* [6] describe her initial research activities, which focus on the importance of a healthy environment in promoting the patient's physical and mental well-being. She changed the attitudes of the military and society towards the care of the sick.

Following Nightingale's work there was very little further nursing research carried out until the 1950s, when a research trend started, particularly in the US. Through the 1950s and 1960s, a number of nursing journals started to appear; for example, *Nursing Research* was first published in 1952. Through the 1960s, clinical research started to expand as specialty groups such as pediatric, obstetric, and community nursing developed standards of care. By the 1970s, nursing research was a growing activity in the UK, and the *Journal of Advanced Nursing* began publication in 1976. The teaching of research methods was introduced to the nursing curriculum in the 1970s, providing nurses with a basic understanding of research methods.

Nursing research has always struggled for funding. A major political victory for nursing research was the creation of the National Center for Nursing Research, in the US, in 1985.

The Briggs report [1] suggested the need for nursing practice to be based on research. Subsequent reports and nursing authorities, such as the Royal College of Nursing, have continued to stress this issue; however, there is still little evidence of wide-scale adoption of this idea.

Types of Study Used in Nursing Research

Cross-sectional surveys have been commonly used in nursing research: however, randomized controlled trials (*see* **Clinical Trials, Overview**) have not been used much in nursing research studies. The dominant approaches to nursing research through the 1980s have been the so-called qualitative approaches, phenomenology, ethnography, and grounded theory,

approaches borrowed from sociology, anthropology, and psychology (*see* **Social Sciences**). They are thought by many nurse researchers to be a better way of gaining an understanding of the rather subjective and personal phenomena with which nursing care is concerned. Such approaches have no place for statistics.

The preference of these qualitative techniques shown by nursing in the past two decades during a time when medicine has been dominated by the randomized controlled trial has enhanced the divide between nursing and medicine. Each side often chooses to disregard the knowledge of the other in the belief that the philosophical approach of the other side is irrelevant.

Statistical Development

Florence Nightingale is often referred to as "The Passionate Statistician". She strongly believed in the importance of figures, and that the collection of reliable data was essential to any worthwhile determination of policy. While in Scutari, Nightingale organized the record-keeping practices, resulting in a systematic method for data collection. From her data it was clear that preventable or, as she called them, "Zymotic", diseases were responsible for many more deaths than were wounds. She used these data to support her case for sanitary reforms, after which mortality declined rapidly, from 42.7% in February 1855 to 2.2% by June 1855.

Nightingale believed in the value of publicizing statistical findings through diagrams. In 1857, she reported to Sidney Herbert "I have written to Dr Farr for the diagram which is to affect thro' the Eyes what we may fail to convey to the brains of the public through their word-proof ears" (quoted in Diamond & Stone [2] from Florence Nightingale's letter to Sidney Herbert of August 7, 1857). Nightingale drew on the help and statistical advice of **William Farr**. The statistical tables in her reports owed much to Farr and the diagrams were prepared under Farr's guidance, although the inspiration for them lay with Florence Nightingale. She herself claimed to have invented her "coxcombs", which were polar area charts, a type of pie chart. An example, shown in Figure 1, is a reproduction of a diagram in Nightingale [5]. It clearly shows the dramatic reduction in mortality following the commencement of sanitary improvements.

Nightingale was elected a Fellow of the **Royal Statistical Society** in 1858, being proposed by William Farr and seconded by **William Guy**, she was among the first (if not the first) female fellows of the society. In 1859, she began a campaign for uniform hospital and surgical statistics to enable one to ascertain the relative mortality in different hospitals. She worked on this idea with Farr and he got the "model forms" that Florence Nightingale had drawn up discussed at an International Statistical Congress in 1860. The "model forms" duly went into hospitals, and uniform statistics for hospitals were published in the ***Journal of the Royal Statistical Society*** from 1862 to 1866, by which time the forms were considered to be out of date and based on a classification of disease that was too rudimentary.

In addition to working with Farr, Florence Nightingale held the work of **Quetelet** in high regard. The influence of Quetelet is discussed by Diamond & Stone [2]. Florence Nightingale wanted to endow a chair of Applied Statistics at Oxford University (it would have been the first chair of statistics), but a dispute with **Francis Galton** over the purpose of the endowment led Nightingale to revoke the legacy [2].

Use of Statistical Methods

It is perhaps surprising, given Nightingale's strong and passionate belief in the need for statistics, that there has been no further development of statistics in nursing research; indeed, the 1980s and 1990s have seen relatively little use of any statistics in nursing research. There have been occasional series of articles about statistical methods in journals such as *Nursing Research* in the Methodology Corner. There have also been a few articles, similar to those appearing in the medical literature, discussing **clinical significance versus statistical significance** and giving examples of misuse of statistical methods in the nursing literature.

The majority of published nursing research has used little more than elementary univariate methods. However, the relatively extensive use of **Cronbach's alpha** for measuring reliability and of **factor analysis** illustrate that nursing tends to use approaches from psychology rather than medicine.

Nursing journals do not currently make extensive use of statistics. When statistics are used,

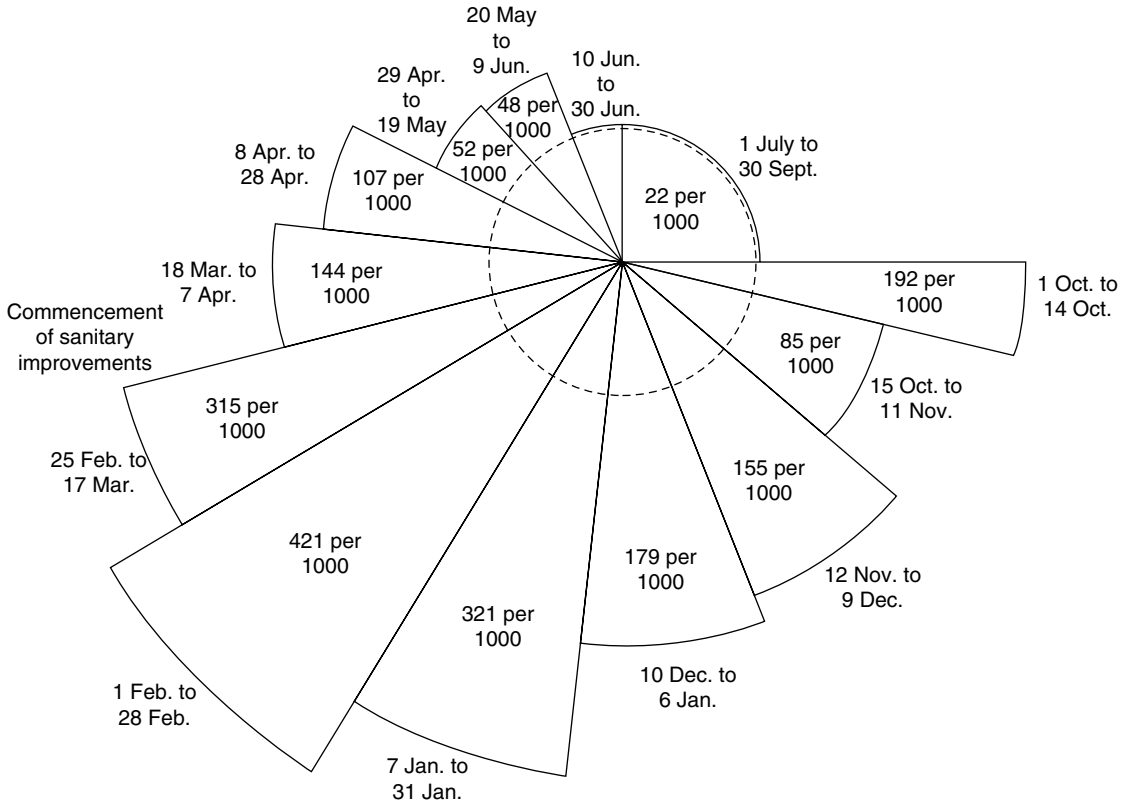

Figure 1 A diagram representing mortality in the hospitals at Scutari and Kulali, from October 1, 1854, to September 30, 1855. The area within the dashed circumference represents the average annual mortality in the military hospitals in and near London, 20.9 per 1000

they are often used and presented poorly. **Confidence intervals** are rarely reported in nursing journals.

The Future

Funk et al. [3] found that a great or moderate barrier to the use of research for 68% of the nurses in their sample was that the statistical analyses were not understandable. The amount of statistical education that nurses receive is negligible, and this is an issue that will need to be addressed if the research-based practice initiative is to become a reality (*see* **Teaching Statistics to Medical Students**).

The emphasis on systematic reviews and **meta-analysis** has started to have an impact on nurses. Good quality systematic reviews of the literature on nursing problems would be welcomed, but will not be easy. The current criteria used to assess study quality will need expansion and development if they are to work well for nursing literature, in which randomized controlled trials are rare. Bayesian approaches to meta-analysis (*see* **Bayesian Methods**) look a promising approach, but they will need to be developed to summarize qualitative as well as quantitative aspects of research.

References

[1] Briggs, A. (1972). *Report of the Committee on Nursing*. HMSO, London.

[2] Diamond, M. & Stone, M. (1981). Nightingale on Quetelet: part I, *Journal of the Royal Statistical Society, Series A* **144**, 66–79.

[3] Funk, S.G., Champagne, M.T., Wiese, R.A. & Tornquist, E.M. (1991). Barriers to using research findings

in practice: the clinician's perspective, *Applied Nursing Research* **4**, 90–95.

[4] Henderson, V. (1966). *The Nature of Nursing*. Macmillan, New York.

[5] Nightingale, F. (1859). *A Contribution to the Sanitary History of the British Army During the Late War with Russia*. Harrison & Sons, London.

[6] Nightingale, F. (1860). *Notes on Nursing - What It Is, And What It Is Not*. Harrison & Sons, London.

[7] Woodham-Smith, C. (1950). *Florence Nightingale 1820–1910*. Constable, London.

NICOLA J. CRICHTON

Nutritional Epidemiology

Epidemiology is the study of the etiology of illness and related phenomena in human populations [38]. Nutritional epidemiology, a branch of epidemiology, seeks to unfold the causal relationship between aspects of the diet and occurrence of human illness (*see* **Causation**). Historically, nutritional epidemiology was concerned mainly with nutritional deficiency diseases where a gross deficiency in a particular food or nutrient caused an untoward condition to occur. An early example, which took place in 1753, was the observation that consumption of lemons and oranges prevented the occurrence of scurvy among sailors on British ships, and this has led to the discovery of vitamin C deficiency as a cause of scurvy.

In recent years, the focus of nutritional epidemiology has been shifted from nutritional deficiency syndromes to the dietary determinants of chronic diseases such as heart disease and cancer. The underlying premise in contemporary nutritional epidemiology is that a person's long-term habitual diet has an impact on the occurrence of chronic disease. However, because the etiology of chronic diseases is a great deal more complex than that of deficiency syndromes, this shift in focus has indeed presented immense challenges. Whereas the occurrence of a deficiency syndrome typically has a single cause (deficiency in a food or nutrient item), the risk of a chronic disease not only can be attributed to numerous causal factors, including genetic, environmental, personal lifestyle (e.g. smoking, drinking, physical exercise) as well as dietary, but also the factors exert varying effects with complex **interactions** on disease occurrence. Moreover, a person's diet is made up of a myriad of dietary components, all of which tend to be correlated with each other, and some of which may increase the risk of disease while others may have a protective effect. Whereas a deficiency syndrome has a short **latent period** of exposure to a single cause (the time interval between onset of deficiency and onset of disease), many chronic diseases have latent periods of exposure that are protracted and ill-defined. Because humans are exposed to most dietary factors for their entire lives, there is no clear standard for comparison. Unlike a deficiency syndrome, where the exposure variable can be categorized as "not deficient" or "deficient", the degree of risk for a chronic disease attributed to most dietary factors varies on a continuum. Also, choosing the most relevant time (person's age) at which to begin measuring diet relative to disease onset (the reference period) is difficult and subjective because the reference period is seldom known and may vary not only from one disease to another but also among persons. Additionally, diet can at any time affect the disease process, and its effects may vary over time. The greatest challenge of all which confronts nutritional epidemiology of chronic diseases is how to measure accurately and precisely a person's long-term diet. Clearly these factors make it difficult to attribute the occurrence of a chronic disease to any single food or nutrient item, and consequently any observed relationship between a food or nutrient item and chronic disease must be interpreted with care and replicated in multiple studies. Notwithstanding, nutritional epidemiology has made important contributions to our knowledge regarding the influence of diet on the etiology of human diseases. The intent of this article is to present a brief overview of nutritional epidemiology. A comprehensive and lucid treatise of the subject is given by Willett [44]. Other general references on nutritional epidemiology include [20] and [25].

Types of Nutritional Epidemiologic Study

Different methods and procedures can be used to carry out a nutritional epidemiologic study on the dietary etiology of human disease occurrence, and comprehensive accounts on epidemiologic study designs are available in the literature [25, 38]. Most of the nutritional epidemiology studies conducted to date are **observational** in nature, in that the allocation of persons to dietary exposure group is not under the control of the investigator. Instead, disease frequency is observed and compared between groups of subjects with different dietary exposures. In this section, we select a sampling of research findings from the different types of nutritional epidemiologic studies.

Group-based correlational studies, which correlate the aggregate disease rate with the average dietary intake for different groups of people, provided the earliest clues that a person's diet may affect the risk of chronic disease. A *geographic correlational study* compares the disease rate and average food intake of groups of people living in diverse geographic areas

(*see* **Ecologic Study**). As an example, Armstrong & Doll in 1975 [3] correlated the cancer **incidence** and mortality rates with the per capita consumption of foods and nutrients from various countries. The **correlations** ranged from 0.7 to 0.8 for meat and animal fat consumption with colon cancer incidence and mortality in men and women; for fat intake with breast cancer incidence and mortality in women; and for fat intake with mortality from cancer of the corpus uteri. These remarkably high correlations stimulated further research on intake of animal products and cancer risk.

Other group-based correlational studies that make use of experiments ongoing in nature are studies of migrants, time trend, and special populations. **Migrant studies** compare disease rates defined by migration status. The disease rates of the first and second generation migrants are compared with rates of people in the country of origin as well as people in the host country. Haenszel & Kurihara in 1968 [15] compared the mortality from cancer and other diseases of the first generation Japanese migrants (Issei), US born second generation Japanese (Nisei), the Japanese in Japan, and US whites. Table 1 summarizes the results for selected causes of death for men. The standardized mortality ratio (SMR) is the ratio of the rate of each index group to the rate of the standard population (Japanese in Japan) statistically adjusted to the age distribution of the standard population. The rate of the standard population is reexpressed as 100 [38] (*see* **Standardization Methods**).

It can be seen that stomach cancer and CVA mortality rates show steady progressions from those in the parent country, where rates are high, to those in the host country, where rates are low. Similarly for colon cancer and heart disease, the mortality experience of the Issei and Nisei increased dramatically towards that of the US whites. The mortality rates for all four sites are higher in the Issei than in the Nisei, and part of this difference can be attributed to the considerable age difference between the two groups with the Issei being older than the Nisei. (The SMR corrects for age differences between the index groups and the standard population, but not between index groups.)

Table 1 SMR comparing male mortality rates with those of Japanese men in Japan, adapted from Haenszel et al. [15]

Cause of death	Japan	Issei	Nisei	US whites
Stomach cancer	100	74	38	17
Colon cancer	100	374	288	489
Intracranial lesions of vascular origin (CVA)	100	32	24	37
Arteriosclerotic heart disease	100	226	165	481

Migrant studies have provided strong evidence for the existence of environmental causes for chronic disease by finding that disease rates of the migrant populations diverged from those of the people in their country of origin and approached the rates of the people in their host country. Since the migrant populations and the people in their country of origin share the same genetic background, the change in disease rates must be attributed to environmental and lifestyle factors.

Studies of time trends can also be helpful in determining the role of the environment in disease etiology. For instance, mortality from stomach cancer in the US declined by more than 30% between 1950 and 1960, and this decline is coincident with a dramatic increase in the per capita consumption of fresh fruits and vegetables (*see* **Morbidity and Mortality, Changing Patterns in the Twentieth Century**). *Studies of special populations* whose diet is restricted also provide a unique opportunity to evaluate the role of the environment and lifestyle factors on disease etiology. For example, disease rates may be compared between Mormons, who abstain from caffeine intake, and a similar group of non-Mormon individuals, to assess the role of caffeine on disease development.

Although useful for generating diet–disease hypotheses, group-based correlational studies have fundamental weaknesses. The most crucial drawback of group-based data is **confounding**. Exposure data are collected at the group rather than at the individual subject level, making confounding a virtual certainty because the high correlations between exposure variables at the individual subject level cannot be disentangled. For example, in many populations where meat intake is high, vegetable intake tends to be low, and consequently any apparent association between average meat consumption and disease rate would be confounded by vegetable intake. When comparing populations of different races (and genetic backgrounds), as in a geographic correlation study, confounding by different genetic predispositions to diseases also renders the findings equivocal.

Table 2 Relative risks for hip fracture by quintiles of calcium intake (mg/day), adapted from Lau et al. [23] and Cooper et al. [11]

	Hong Kong			Britain		
	Calcium intake	Relative risk		Calcium intake	Relative risk	
Quintile	Range	Women	Men	Range	Women	Men
Q1 (low)	<75	1.9	2.1	<433	1.2	6.2
Q2	75–82	1.9	1.4	433–566	1.4	5.8
Q3	83–128	1.1	1.7	567–683	1.1	3.3
Q4	129–243	1.2	1.5	684–837	1.2	6.2
Q5 (high)	≥244	1.0	1.0	≥838	1.0	1.0

Table 3 Odds ratios for prostate cancer by quantiles of dietary fat intake, for cases diagnosed after age 69 and their matched controls, adapted from Kolonel et al. [21]

Quantile[a]	Total	Caucasians	Japanese	Filipinos	Hawaiians	Chinese
Q1	1.0	1.0	1.0	1.0	1.0	1.0
Q2	1.1	2.0	0.6	4.0	1.2	1.1
Q3	1.5	2.3	0.8	5.8	1.3	1.6
Q4	1.7	2.6	1.2	2.8		

[a]Quartiles for total, Caucasians, Japanese, and Filipinos, and tertiles for Hawaiians and Chinese.

The quality of the **cause of death** data on the death certificates can also vary considerably among countries (*see* **Death Certification**), and this may render the disease rates not comparable.

The types of study described below record dietary exposure and disease status from individual subjects, thus avoiding many of the drawbacks inherent in group-based data. The most widely used study design in nutritional epidemiology is the **case–control study**. With this study design, people with the disease (cases) and comparable individuals without the disease in question (**controls**) are asked about their dietary and nondietary exposures. Nutritional factors associated with disease occurrence are determined by comparing the past diet of the cases with that of the controls.

Two case–control studies were undertaken, one in Hong Kong [23] and one in the UK [11], to assess the effect of dietary calcium on hip fracture in men and women. Dietary calcium intake was estimated based on the frequency of consumption of nine food items in the Hong Kong study and six food items in the British study. A total of 400 radiologically confirmed fracture cases, 400 hospital controls, and 400 community controls were recruited in Hong Kong, and 300 cases were compared with 600 community controls in the UK. A clear protective effect of calcium intake on hip fracture was found in all groups excepting the British women (Table 2).

Case–control studies have largely been consistent in demonstrating the adverse effect of dietary fat for prostate cancer. One of these studies, conducted in Hawaii, included 452 histologically confirmed prostate cancer cases diagnosed between 1977 and 1983, and 899 age-matched (*see* **Matching**) population controls [21]. The participants were administered a diet history questionnaire with over 100 food items. Table 3 shows a monotonic effect of saturated fat on the risk of prostate cancer in older cases (those diagnosed after age 69), and the finding is consistent in most of the ethnic groups.

The case–control study offers many strengths in the investigation of the dietary etiology of chronic diseases. It is relatively inexpensive and efficient, typically requiring several hundred study participants and 2–4 years to complete the study; individual diet is measured and related to the risk of disease; and confounding can be minimized through appropriate selection of controls, by using appropriate statistical analysis techniques, or both. The case–control study is often the only feasible choice of study design for rare diseases. The primary disadvantage in all case–control studies is that the measurement

of exposure data relies on recall. This is particularly a problem in nutritional studies where cases must remember their past diets before the onset of disease. The diet may also have changed as a result of the disease process, and the current diet has been found to influence the recall of past diet. The results from a case–control study will be invalid if the cases remember their past diets differently than the controls do (**recall bias** of exposure), and if the controls are not comparable with the cases (**selection bias**). Healthy control bias can be especially problematic in studies of nutritional epidemiology if only health-conscious volunteers serve as controls.

Cohort studies avoid the inherent problems which afflict the case–control study, namely recall bias and selection bias. In this study design, diet is measured on a large number of disease-free individuals who are then monitored for disease occurrence. To date, cohort studies in nutritional epidemiology have been few in number, primarily because of their considerable high cost in terms of resources and length of follow-up. More than a decade of follow-up on thousands or even hundreds of thousands of individuals may be required in a cohort study of diet and chronic disease. Because a person's diet is likely to change with age, it is important to record dietary intake on several occasions from each study subject throughout the follow-up period. Repeated measurements will provide more accurate information about the average long-term exposure (*see* **Longitudinal Data Analysis, Overview**). As noted earlier, nutritional epidemiology of chronic disease is based on the premise that long-term dietary exposure affects disease risk.

In one large cohort study, 43 757 male health professionals in the US were recruited to investigate the etiology of the intake of dietary fiber on the occurrence of myocardial infarction. The cohort members completed a mailed dietary questionnaire with over 100 food items; they were then followed for 6 years for the occurrence of heart disease. Intake of dietary fiber, particularly from cereal, was found to be statistically associated with a decreased risk for myocardial infarction [32]. Table 4 shows the **relative risks** adjusted for relevant **covariates**.

Another cohort study example is the Iowa Women's Health Study, used to investigate the postulation that aspects of the diet influence the occurrence of endometrial cancer [47]. A cohort of over 23 000 women was recruited and a questionnaire with 127 food items was administered to each participant. After 7 years of follow-up, dietary intake was correlated with the incidence of endometrial cancer. Although the findings are on the whole equivocal, caloric intake from animal sources and intake of processed meat appear to be associated with a slight increase in the risk of endometrial cancer, especially during the early years of follow-up (Table 5).

In an *intervention study* or *controlled trial* (*see* **Clinical Trials, Overview**), study participants are randomly allocated to the different dietary regimens (*see* **Randomization**). If the subjects have

Table 4 Relative risks for myocardial infarction by quintile of energy-adjusted dietary fiber, adapted from Rimm et al. [32]

Type of fiber	Q1 (low)	Q2	Q3	Q4	Q5
Total	1.00	1.01	0.96	0.92	0.64
Fruit	1.00	0.93	0.83	0.84	0.82
Vegetable	1.00	1.06	0.98	1.00	0.84
Cereal	1.00	0.98	0.90	0.88	0.73

Table 5 Relative risks for endometrial cancer by tertile of intake for selected dietary factors, adapted from Zheng et al. [47]

	≤4 years after cohort entry			≥5 years after cohort entry		
	T1 (low)	T2	T3	T1 (low)	T2	T3
Caloric intake from animal foods	1.0	1.3	1.2	1.0	0.9	0.9
Total meat	1.0	1.0	1.3	1.0	1.0	0.9
Red meat	1.0	0.9	1.2	1.0	0.9	0.9
Seafood	1.0	1.4	1.0	1.0	1.4	2.0
Processed meat	1.0	1.4	1.6	1.0	1.0	1.3
Dairy products	1.0	1.2	1.2	1.0	0.8	1.0
Eggs	1.0	1.2	1.4	1.0	1.4	1.3

the disease, the dietary component is tested as a therapeutic agent, and if the subjects are disease-free, the dietary component is tested as a chemopreventive agent. The controlled trial has one crucial advantage over the cohort study described above, and that is randomization. Because random allocation of subjects to different exposure groups tends to reduce confounding, the controlled trial is able to establish with greater confidence whether a dietary component is a causal factor. However, practical limitations concerning **compliance**, dosage, latency (*see* **Latent Period**), and cost seriously diminish the usefulness of intervention studies in nutritional epidemiology. Most free-living individuals will not strictly follow a dietary regimen, and the persons in the control group may adopt a diet similar to the test diet, particularly if it is perceived to be beneficial (a phenomenon known as control drift). These compliance issues will obscure the differences between the treatment groups, making it more difficult to ascertain a true effect of the diet on disease risk. An example of "control drift" was found in the Multiple Risk Factor Intervention Trial (MRFIT) which randomly allocated 12 866 men at high risk for coronary heart disease to either a special intervention program or usual care [27]. The intervention program was a three-pronged intervention aimed at smoking cessation, blood pressure, and serum cholesterol reduction through lifestyle and dietary modification. Although serum cholesterol dropped from 254 to 236 mg/dl over 72 months in the intervention group, a similar reduction occurred in the nonintervention group: from 254 to 240 mg/dl.

A randomized, double blind, placebo controlled trial of 29 133 male smokers, the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study, was undertaken to determine whether supplementation of alpha-tocopherol and beta-carotene would prevent lung cancer [7]. The study failed to confirm results from observational studies that had found a protective effect for these dietary components. Lack of precise dosage and latency information have been postulated as factors in the results.

Community intervention trials are experimental studies carried out at the population level; they are generally concerned with the effectiveness of an education or incentive program on behavior. For example, a number of participating communities may be randomly allocated either to receive information about the benefits of low-fat diets (test) or to receive no such information (control). **Cross-sectional** surveys, based typically on a sample of subjects from each community, are conducted before and after the information campaign to determine if there is a greater change in diet in the test communities than in the control communities. "Contamination" between groups, where the control communities also receive the test information, can be a problem in community trials.

Variation in Dietary Intake

As noted previously, the underlying premise of nutritional epidemiology is that a person's true "average" diet affects the occurrence of chronic diseases. It would be an easy task to ascertain a person's true diet provided that a person eats the same foods and the same quantity of each food day in and day out (no day-to-day variation), and that a person's diet and its nutrient content can be measured perfectly (no measurement error). In reality, a person's diet varies not only from day to day, but food preference and quantity consumed may be altered as one ages and as circumstances change. In contrast to simple exposures such as smoking, a person's diet is a composite exposure consisting of many food and beverage items consumed in varying amounts, and each food and beverage item contains numerous macro and micronutrients. It is unlikely that an instrument ever will ever be developed which will ascertain dietary intake exactly. Indeed, accurate assessment of a person's long-term average diet is a major concern of nutritional epidemiologic research today, and methods and procedures for estimating the long-term diet are still evolving. In this section we delineate the concepts and definitions of variation in dietary intakes. Measurement error will be discussed in the next section.

The day-to-day variation in a person's food and nutrient intake is assumed to be random, that is, a person's daily diet fluctuates randomly about his or her "average" diet. This assumption implies that the average dietary intake over a number of randomly selected days will approach a person's true habitual intake, and that any deviation of intake on a given day from the true average diet is a reflection of imprecision (sampling error) rather than **bias**.

Some dietary components have been found to be more variable than others. Macronutrients, such as

fat which is present in most foods, tend to show less day-to-day variation than substances that are present in only a few foods or whose amount is very high in a particular food. For instance, eating a mango can drastically increase the intake of beta-carotene for that day. The day-to-day dietary intake variation for a given person is referred to as *within-person variance*. And the variation in dietary intake among different persons is called *between-person variance*. To estimate within-person variance, more than one measurement per person is required.

It might be useful to depict the within-person and between-person components of dietary variation by a statistical model. The **random effects analysis of variance** model, also called the **variance components** model, is given as:

$$Y_{ij} = \mu + \alpha_i X_i + \varepsilon_{ij},$$

where Y_{ij} is the nutrient or food of interest for the ith person and the jth day of measurement, μ is the mean intake across persons and days, X_i is a **dummy variable** that identifies person i, α_i represents the random effect due to person i, and ε_{ij} represents the random day-to-day variability, or the within-person variability.

The variance components model assumes that these conditions hold: the variance of ε_{ij}, designated by σ_ε^2, is constant across days and persons, and the covariance between ε_{ij} and $\varepsilon_{i'j}$ is zero when $i \neq i'$. The variance of α_i, designated by σ_α^2, is constant across persons, and the covariance between α_i and $\alpha_{i'}$ is zero when $i \neq i'$. Models of dietary components often violate the first variance homogeneity assumption because persons with higher intake tend to have a larger within-person variance (i.e. the variance of ε_{ij} is proportional to Y_{ij}). The assumption is usually upheld after a suitable **transformation** (usually log) is applied to the nutrient data Y_{ij}.

The total variance (σ_y^2) is the sum of the between-person variance (σ_α^2) and the within-person variance (σ_ε^2). (With m replicated observations per person, total variability is $\sigma_\alpha^2 + \sigma_\varepsilon^2/m$.) The ratio of these quantities ($\sigma_\varepsilon^2/\sigma_\alpha^2$) gives an indication of the relative importance of the within-person to the between-person variance components. A ratio close to zero indicates that most of the dietary intake variability occurs between persons, in which case the intake level for an individual would be fairly constant from day-to-day. A ratio around one indicates an equal split between the two components, and ratios greater than one indicate that the within-person variance exceeds between-person variance. Another useful statistic based on the variance components is the "intraclass correlation coefficient" (*see* **Correlation**), which will be described later under **validation studies**.

Numerous investigators have estimated the within-person and between-person variance components for the daily intake of common nutrients based on repeated food records [5, 16, 24, 28, 39]. Table 6 presents a summary of the results in the form of ratios. Although between-person variation is substantial, within-person variability is the larger variance

Table 6 Ratio of within-person to between-person variance components, adapted from [5, 16, 24, 28], and [39]

Nutrient	Number of studies	Median ratio	Range of ratios
Energy (kcal)	12	1.4	0.8-2.2
Protein	10	1.4	1.2–3.9
Carbohydrate	9	1.2	0.8–2.0
Fat	9	1.3	0.9–2.8
Percent of calories from fat	8	2.4	1.3–4.8
Saturated fat	8	1.5	1.0–2.8
Cholesterol	11	4.4	1.8–6.8
Vitamin C	9	2.3	1.6–4.0
Vitamin A	7	4.6	1.6–>100
Iron	8	2.4	1.5–3.6
Calcium	10	1.6	1.0–2.6
Zinc	6	2.4	1.7–11.7
Dietary fiber	3	1.7	1.1–2.2

component, accounting for 55% (carbohydrate) to 82% (vitamin A) of the total variation.

Measurement Error

Error in measuring the exposure variable is a common concern in all etiologic research. Measurement error may be minimal in simple exposures such as smoking history, but it may be quite consequential in complex exposures such as long-term dietary cholesterol intake. **Measurement error** occurs when the true exposure value in the ith person, X_i, is not directly observable, but instead a surrogate value, Z_i, is measured. Measurement error, defined as $(Z_i - X_i)$, may be *random* (unbiased) or *systematic* (biased). The error is random if the expected (long-range average) value of Z_i is X_i, and is systematic if the expected value of Z_i is not X_i. Random measurement error will occur if a person is just as likely to overreport as underreport, by the same amount, the consumption of a food item. **Systematic error** will occur if a person is more likely to underreport than overreport, or vice versa, the consumption of an item.

Systematic measurement error of the exposure variable may or may not bias the exposure–disease association. If measurement error is constant for all persons (e.g. all persons underreport the use of cooking oil by the same amount), the exposure–disease association will not be biased, and the **power** will not be diminished. However, systematic measurement error is seldom, if ever, the same for all individuals.

Although systematic measurement error may not be constant for all persons, the overall average error in persons with disease (cases) may be the same as that in persons without disease (controls). This is called **nondifferential measurement error**. Nondifferential error will not bias the exposure–disease association, but it will reduce the statistical power. **Differential measurement error** occurs when the overall average error in the cases is not the same as that in the controls (e.g. the cases on average underestimated the use of cooking oil more so than the controls). Systematic errors which are differential between cases and controls will lead to invalid estimates of the exposure–disease association, and the extent and direction of the bias are difficult to predict.

The consequence of **random measurement error** in the exposure variable will depend on the specific situation. If the exposure variable is a single continuous or dichotomous variable, random measurement error will attenuate the exposure–disease association, that is, **bias towards the null** the **correlation** coefficient, **regression** coefficient or relative risk. Random measurement error also tends to inflate the **standard deviation** for the association, thereby reducing power of statistical tests. If the exposure variable is **polytomous** (with more than two exposure levels), the **odds ratio** or relative risk for the most extreme exposure level will be biased toward the null value, while those for the intermediate levels can be biased away from the null [6, 13, 26]. When confounding variables are measured with random error, the effect of the exposure variable on disease risk may be biased away from the null, even if the exposure variable were measured without error [22]. Also random measurement error in the exposure variable can lead to incomplete adjustment of confounding, resulting in residual confounding in the adjusted exposure–disease **association** estimates [2, 14].

It is a useful practice to estimate the extent of measurement error in the exposure variables so that more reliable exposure–disease associations can be ascertained by taking into account these errors. Correction of measurement error may help to clarify whether an observed null exposure–disease association is real or attenuated. A *reproducibility study* with repeated measurements of the exposure variables can be deployed to estimate random exposure measurement error (*see* **Reliability Study**). To evaluate systematic measurement error, a **validation study** of the dietary instrument against a **"gold standard"**, or at least a more superior instrument, is required.

The simplest approach to reduce random measurement error in the exposure variable is to use the average value obtained from repeated measurements of the exposure, as an average based on replicated values has less random error than a single measured value. For example, study subjects are asked to keep food records on several occasions. This approach may be feasible in a small etiologic study but is likely to be prohibitive in a large study.

An alternative strategy to minimize the effect of exposure measurement error on the exposure–disease association is to estimate the association correcting statistically, for the measurement error in the exposure variable. The general approach is to quantitate the statistical relationship between the measured value, Z_i, and the true value, X_i, of the exposure

variable obtained from a **"calibration"** study and to use this relationship in the correction of measurement error (*see* **Measurement Error in Epidemiologic Studies**). The "calibration" study is either a reproducibility or validation substudy based on a sample of subjects taken from the main etiologic study. Statistical methods pertaining to the correction of exposure measurement error abound in the literature, and they are still evolving. Because of the broad nature of these methods, which encompass different statistical models and assumptions and different types of measurement error, the account given below is intended only to provide a brief and incomplete sketch of the topic, with a sampling of references for the reader to turn to for more information.

The basic concept underlying the measurement error correction methods can be depicted by a simplified model. The expected value of X, given the observed Z, is substituted for every study participant in the main etiologic study, based on the calibration substudy information. $E(X|Z)$ is then substituted for Z in the disease etiology model in the main study to obtain an estimate of the true exposure–disease association. A critical assumption of this model is that the relationship between X and Z in the main study is the same as that in the calibration substudy sample. Wacholder et al. [41] warn that the value from the "gold standard" in the calibration study is almost always measured with error, albeit with less error than the measurements used in the main study. They show that correction to such an "alloyed gold standard" only partially eliminates the bias when the measurement errors between the two methods are moderately to strongly positively correlated. But when the measurement errors are either inversely correlated, uncorrelated, or weakly positively correlated, the corrected exposure–disease association estimate will tend to be overcorrected (anticonservative). It is clear that care needs to be exercised in the application of these statistical correction techniques.

The two examples shown below illustrate the effect of exposure measurement errors on the exposure–disease association and the role of the statistical correction method. Both the examples assume that results from a calibration substudy are available, that the relationship between X and Z in the calibration substudy can be extrapolated to the main study, and that the exposure measurement error is nondifferential, that is, it does not depend on disease status.

Example 1

This example illustrates a statistical technique for correcting systematic and random measurement errors in a dichotomous exposure variable [12]. Results from a validation study are required. Table 7 shows the frequencies for the association between a "true" exposure (X) and disease status.

The relationship between the true exposure (X) and the observed exposure (Z) from the calibration substudy is given in Table 8. Even with a moderately high **sensitivity** of 0.60 and **specificity** of 0.70, the exposure–disease associations are severely attenuated. As shown in Table 9, the true odds ratio of 5.0 becomes 1.6 and the true relative risk of 2.3 becomes 1.2.

The odds ratio or relative risk estimates can be corrected for measurement error when sensitivity (ξ) and specificity (ψ) estimates are available from a validation substudy, using the identities in Table 10:

In our example,

$$a_{\text{true}} = \frac{400(0.70) - 190}{0.6 + 0.7 - 1} = 300,$$

$$b_{\text{true}} = \frac{400(0.6) - 210}{0.6 + 0.7 - 1} = 100,$$

Table 7 Association between true exposure (X) and disease status

	True exposure		
Disease status	Yes	No	Total
Yes	300	100	400
No	150	250	400
Total	450	350	800

Case–control study: odds ratio $= (300 \times 250)/(150 \times 100) = 5.00$
Cohort study: relative risk $= (300/450)/(100/350) = 2.33$

Table 8 Relationship between true exposure (X) and observed exposure (Z)

	True exposure		
Observed exposure	Yes	No	Total
Yes	60	30	90
No	40	70	110
Total	100	100	200

Sensitivity $= \xi = 60/(60 + 40) = 0.60$.
Specificity $= \psi = 70/(30 + 70) = 0.70$.

Table 9 Association between observed exposure (Z) and disease status

	Observed exposure		
Disease status	Yes	No	Total
Yes	210	190	400
No	165	235	400
Total	375	425	800

Case–control study: odds ratio $= (210 \times 235)/(165 \times 190) = 1.57$
Cohort study: relative risk $= (210/375)/(190/425) = 1.25$

Table 10 Correction identities

	Exposure		
Disease	Yes	No	
Yes	a	b	n_{D}
No	c	d	n_{ND}

$$a_{\mathrm{true}} = \frac{n_{\mathrm{D}}\,\psi - b_{\mathrm{obs}}}{\xi + \psi - 1}, \qquad b_{\mathrm{true}} = \frac{n_{\mathrm{D}}\,\xi - a_{\mathrm{obs}}}{\xi + \psi - 1}$$
$$c_{\mathrm{true}} = \frac{n_{\mathrm{ND}}\,\psi - d_{\mathrm{obs}}}{\xi + \psi - 1}, \qquad d_{\mathrm{true}} = \frac{n_{\mathrm{ND}}\,\xi - c_{\mathrm{obs}}}{\xi + \psi - 1}$$

$$c_{\mathrm{true}} = \frac{400(0.70) - 235}{0.6 + 0.7 - 1} = 150,$$

$$d_{\mathrm{true}} = \frac{400(0.6) - 165}{0.6 + 0.7 - 1} = 250.$$

This technique can be generalized to correct for differential measurement error if sensitivity and specificity estimates are available separately for cases and controls. Note the **standard error** for the corrected association will be larger than that for the uncorrected value to account for the sampling error in the estimation of sensitivity and specificity. A corrected standard error is not available in the literature. Rarely are epidemiologic studies focused on one dichotomous exposure variable without other covariates.

Example 2

This example illustrates that random measurement error in two dietary exposures will tend to bias their correlation towards zero. If information is available from a reproducibility study for the two measurements W and U, and no systematic error is present, then the following variance components model hold,

$$W_{ij} = \mu_W + \alpha_i X_i + \varepsilon_{ij},$$
$$U_{ij} = \mu_U + v_i X_i + \zeta_{ij},$$

so that the overall variances are $\mathrm{var}W = s_{\mathrm{b}}^2 + s_{\mathrm{w}}^2/n$ and $\mathrm{var}U = v_{\mathrm{b}}^2 + v_{\mathrm{w}}^2/m$. All errors are assumed to be uncorrelated. The estimate of the "true" correlation can be computed as

$$r_{\mathrm{true}} = r_{\mathrm{obs}} \left[\left(1 + \frac{s_{\mathrm{w}}^2}{(s_{\mathrm{b}}^2 n)}\right)\left(1 + \frac{v_{\mathrm{w}}^2}{(v_{\mathrm{b}}^2 m)}\right)\right]^{1/2}$$

where r_{obs} is the observed correlation between W_{obs} and U_{obs}, s_{w}^2 is the within-person variability of W from a reproducibility study, s_{b}^2 is the between-person variability of W from a reproducibility study, n is the number of replicated values of W from a reproducibility study, v_{w}^2 is the within-person variability of U from a reproducibility study, v_{b}^2 is the between-person variability of U from a reproducibility study, and m is the number of replicated values of U from a reproducibility study.

The proof is given below:

$$\begin{aligned}
r_{\mathrm{obs}} &= \frac{\mathrm{cov}(W_{\mathrm{obs}}, U_{\mathrm{obs}})}{[\mathrm{var}(W)\mathrm{var}(U)]^{1/2}} \\
&= \frac{\mathrm{cov}(\mu_W + a_i X_i + \varepsilon_{ij}, \mu_U + v_i X_i + \zeta_i j)}{[(s_{\mathrm{b}}^2 + s_{\mathrm{w}}^2/n)(v_{\mathrm{b}}^2 + v_{\mathrm{w}}^2/m)]^{1/2}} \\
&= \frac{\mathrm{cov}(\mu_W, \mu_U)}{[s_{\mathrm{b}}^2 v_{\mathrm{b}}^2 (1 + s_{\mathrm{w}}^2/(n s_{\mathrm{b}}^2))(1 + v_{\mathrm{w}}^2/(m v_{\mathrm{b}}^2))]^{1/2}} \\
&= \frac{\mathrm{cov}(\mu_W, \mu_U)}{(s_{\mathrm{b}}^2 v_{\mathrm{b}}^2)^{1/2}} \\
&\quad \times \frac{1}{[(1 + s_{\mathrm{w}}^2/(n s_{\mathrm{b}}^2))(1 + v_w^2/(m v_{\mathrm{b}}^2))]^{1/2}} \\
&= r_{\mathrm{true}} \frac{1}{[(1 + s_{\mathrm{w}}^2/(n s_{\mathrm{b}}^2))(1 + v_{\mathrm{w}}^2/(m v_{\mathrm{b}}^2))]^{1/2}}.
\end{aligned}$$

When only one variable U is measured with error, the formula relating the true and observed correlations becomes

$$r_{\mathrm{true}} = r_{\mathrm{obs}} \left(1 + \frac{v_{\mathrm{w}}^2}{(v_{\mathrm{b}}^2 m)}\right)^{1/2}.$$

Similarly, if only random error is present in the exposure variable, the slope in a **linear regression**

can be corrected as

$$b_{\text{true}} = b_{\text{obs}} \left(1 + \frac{v_{\text{w}}^2}{(m v_{\text{b}}^2)}\right)^{1/2}.$$

As an example suppose a reproducibility study includes 14 food records from which fat and vitamin E intake were computed. Assume the within to between person variance ratios of 1.5 for fat and 4.0 for vitamin E. In a large epidemiologic study, the correlation between these two nutrients was found to be 0.55. An estimate of the correlation corrected for measurement error is

$$0.55 \times \left[\left(1 + \frac{1.5}{14}\right)\left(1 + \frac{4.0}{14}\right)\right]^{1/2} = 0.66.$$

Again, the standard deviation for the corrected correlation coefficient or regression coefficient must account for the variability in the estimation of s_{b}^2, s_{w}^2, v_{b}^2, and v_{w}^2. Rosner et al. [34] derived a standard deviation for the corrected correlation based on the **delta method**. Readers are referred to Beaton et al. [5] for a more general formula where the measurement errors cannot be assumed uncorrelated, and to Kupper [22] for the effect on partial correlations when a confounder is measured with error.

Statistical principles and procedures for the correction of exposure measurement errors in relative risk and odds ratio estimates under more complex models and assumptions are expounded in the following references: [1, 2, 4, 31, 33, 35–37], and [43]. These methods address multiple continuous or categorical exposure variables, confounding variables, both systematic and random measurement errors, and other situations often encountered in nutritional epidemiologic research.

Dietary Assessment Methods

As noted earlier, the day-to-day variation and the measurement errors inherent in ascertaining a person's long-term diet constitute the major challenge in modern day nutritional epidemiology. Continuing research efforts are still being devoted to developing better methods for measuring a person's "average" diet. All the instruments used for assessing dietary intake rely on information supplied directly by the study participants, usually in the form of a questionnaire. Selection of a dietary assessment instrument for a given nutritional epidemiologic study is motivated by the intended use of the dietary data, and considerations include whether the short-term or long-term "average" diet is relevant, what dietary components are most germane, and whether absolute intake or relative intake (i.e. the ranking of individuals by intake) is desired (*see* **Nutritional Exposure Measures**).

The most commonly used tools for dietary assessment are food records, 24-hour recalls, and food frequency questionnaires. Food records and 24-hour recalls, methods that assess recent diet, are generally not feasible for use in large-scale epidemiology studies. Both methods provide information on the total diet (daily calories) and can give information concerning patterns of food consumption.

Food records are arguably the most accurate method for assessing dietary intake. Participants are required to record in a diary all foods at the time they are eaten, as well as ingredients to all recipes. Weighing foods before eating and any leftovers afterward is a common method for evaluating the amount of food consumed. The record-keeping typically covers 3–7 days. Clearly, this technique puts a heavy demand on the participants and is suitable only for literate and motivated volunteers. Another disadvantage with food records is that the very act of recording of foods consumed can change dietary behavior, either by avoiding foods that are considered undesirable or by simplifying the diet to facilitate the transcription. However, food diaries are often the assessment method of choice when high accuracy in the measurement of diet is needed, as for example in validation studies (described later).

In the *24-hour recall* method, a trained interviewer elicits information about the foods consumed and their amounts, during the past day. This technique is rapid, typically taking 10–20 minutes to complete, and is not very burdensome to the study participants. The quality of the 24-hour recall is directly related to the skill of the interviewer, who uses structured probes to facilitate an individual's memory, but who must be careful not to influence the responses. Interviews should be unannounced to avoid having the persons change their diets for ease of recall. Telephone administration of 24-hour recalls is feasible, although the estimation of amounts is more difficult (*see* **Interviewing Techniques**).

The crucial drawback of food records and 24-hour recalls is that the large within-person variability in most dietary components causes such short-term

dietary information to be highly imprecise, deviating substantially from a person's usual or average diet. We illustrate this with the variance components model. Person k has true mean intake of $\mu + \alpha_k$, but with a single record or recall the measured value would be $\mu + \alpha_k + \varepsilon_{kj}$ with variance σ_ε^2. Averaging data across multiple records improves the estimation of the person's true average diet; with m replicated days of collection, person k's measured value would be $\mu + \alpha_k + \sum_j(\varepsilon_{kj})/m$ with variance σ_ε^2/m. It can be seen that the within-person variation becomes negligible as m increases. The number of replicated days required to characterize a person's "true" long-term average diet with high precision has been estimated from numerous dietary variability studies. Briefly, the estimates range from 4 days to 15 days for energy intake, 6 to 14 for fat intake, 6 to 23 days for vitamin C intake, and 47 to 105 days for vitamin A intake. Adjustment for calories (described later) tends to decrease the number of days slightly. Because many different methods were used to estimate the required number of replicated measurements, interested readers are directed to the following selected references: [5, 16, 24, 28], and [39].

The dietary assessment method best suited for use in large-scale nutritional epidemiologic studies is the *food frequency questionnaire* (FFQ), also referred to as the *diet history method*. With the FFQ technique, the frequency (and sometimes amount) of consumption of a list of commonly eaten foods is obtained. The list may consist of only a few highly selected food items, or as many as 100–200 food items, depending on the etiologic hypothesis being tested. The questions may be open ended, where the respondent gives the frequency of consumption for each food item as times per day, week, month or year, or they may be close ended where several frequency categories are listed. Seasonal food items are incorporated by asking for their intake during the season they are available. Many of these questionnaires have "write-in" options where more detail is obtained about specific foods or where participants can add food items important to their diets that are not covered by the list. If the amount consumed is also estimated, typically by incorporating usual portion size or serving size of each food item, the questionnaire is then referred to as "quantitative". The serving sizes of foods with natural units such as eggs are relatively easy to assess. For other foods, aids such as household measures, food models, or photographs are often used to facilitate estimation.

The FFQ method is flexible in that it can be used for short-term or long-term dietary recall, and for estimating a partial or comprehensive diet; it can be administered by an interviewer or self-administered. When the long-term average diet is desired, as is the case with most nutritional epidemiologic research, it is important to inquire about a person's diet covering the relevant time frame or reference period. In a case–control study, the case is asked about his or her "usual" diet before the onset of symptoms, and the control is typically asked about last year's diet, provided there has not been a recent change in the diet. In a cohort study, it is important to administer the FFQ more than once to increase the precision of the estimate of a person's average diet.

A comprehensive diet (or total caloric intake) is sometimes required as overnutrition or undernutrition may have a direct effect on disease risk. Also, measurement of dietary components relative to the total diet may be of interest (see caloric adjustment below).

A drawback of the FFQ is that its list format makes it "population sensitive". A questionnaire that is well suited to one population (e.g. Caucasians) may not include the necessary items to cover adequately the diet in another group (e.g. Japanese). It is crucial to ensure that a FFQ covers all the commonly consumed food items in the population. The FFQ interview can be lengthy, typically requiring between 1 and 2 hours to complete, and this may adversely affect the response rate. Because the FFQ inquires only about consumption of selected foods, it measures relative rather than absolute dietary intake. Administration of different FFQs to the same person is likely to produce different absolute values on intakes, such as grams of fat consumed per day. However, the dietary intakes of a group of persons as assessed by the two FFQs should show comparable rankings. Measurement of relative intake is generally adequate for etiologic research because comparison between cases and controls is of central interest.

Dietary Components

Diet consists of many substances, such as nutrients, additives, contaminants, and other unknown compounds. Information collected on a limited set of relevant foods may be adequate when the dietary

component of interest is concentrated in those foods. In most studies on nutritional etiology of disease, however, a wide variety of dietary components are of interest, such as intake of total calories; macronutrients, including fat, protein, and carbohydrate; micronutrients, including vitamin A, vitamin E, and iron; and intake of particular foods or groups of food, such as red meat or fruits.

In many of the early studies, information on a brief list of foods was collected, and analysis was focused on food intake. More recently, questionnaires on comprehensive diet were introduced, and the emphasis was on nutrient rather than food intake. There is now some realization that both nutrients and foods are important. A nutrient may react in the same way biologically regardless of its food source, in which case the nutrient intake is relevant. Also, nutrient analyses allow comparison across studies from populations with different food intakes. An example where food intake may be the relevant exposure is the recent interest in the influence of soy product consumption on the risk of certain cancers. Analyses often incorporate both nutrients and food sources. For instance, fat from meat sources and fat from dairy products can be studied as separate exposures. Because food choices are correlated, it is desirable to study food patterns and disease risk. Analysis of food patterns with techniques such as **factor analysis** has been attempted but is very preliminary.

Use of dietary supplements can substantially alter a person's dietary profile, and information on supplement use should be collected. However, the nutrient data should be analyzed separately from food alone and from foods and supplements, as it is not known if a nutrient from a food and from a supplement behave the same way biologically.

Computation of Nutrients

Computation of nutrients requires information on the nutrient composition of the food. Information generally comes from published nutrient composition tables, but may be supplemented by food analysis. National nutrient composition data are available from many countries; the US Department of Agriculture (USDA) publishes information about the nutritional content of foods in the US. Investigators need to use food composition data from food sources similar to those under study whenever possible. Nutrient content can vary dramatically by locale. For example, the carotene content of plants is affected by soil type, and the iodine content of fish is affected by seawater content. The nutrient composition data in national tables usually represents an average value from analyses of that food from different sources and locales. The tables generally have information on hundreds of nutrients, given as units, such as grams and milligrams, per 100 grams of food. The nutrient composition data have varying degrees of accuracy. Macronutrients such as protein and fat content tend to vary less between food samples than micronutrients. Some nutrients, like selenium, are so variable between samples that usefulness of the nutrient composition data is questionable.

The **algorithm** for nutrient computation is to compute, for each food in the recall or FFQ, the daily grams of consumption. If portion size is asked, each serving size needs to be assigned a weight in grams. Otherwise, weight in grams should be assigned to a standard serving size. Daily grams are computed for each food as frequency of intake per day times the gram weight of that food. These quantities are compared against a food composition table, and daily nutrients from each food are computed as daily grams of consumption times the nutrient content of that food per 100 grams, multiplied by 100. These quantities are summed across foods for each person to obtain nutrients per day. Note that each food on the dietary assessment instrument must be associated with a food in the composition table. This assignment is very labor intensive for food records and 24-hour recalls and requires a person knowledgeable about nutrition, such as a dietitian. A questionnaire specific food composition table is required for FFQs where several foods are grouped into a single question. For example, an item for beef may include steak and roasts. The corresponding item in the food composition table will be a weighted average of nutrient composition data for each beef item. Complex questionnaires generally require initial preparation before comparison with food composition data, such as adjustment for oils added during cooking and fats eaten on meats.

Reproducibility and Validity

With food frequency questionnaires (FFQ), volunteers are asked to estimate their own usual diets. In the parlance of the variance components model,

a participant is being asked to recall the true average intake $\mu + \alpha_k$, thereby "eliminating" the within-person day-to-day variability. It is of course essential to gauge how well the FFQ measures true average diet. Desirable qualities for a FFQ are reproducibility and validity.

A *reproducible*, or reliable, instrument will give consistent answers on repeated administrations. To study reliability, a FFQ is given to the same person at two or more points in time and their responses are compared. The time period between administrations cannot be too short so that participants remember what they reported earlier. It also cannot be too long as an intervening dietary change can affect responses. Therefore, a period of several months is typical. Correlations between multiple administrations of questionnaires have ranged from 0.5 to 0.7; these correlations compared favorably with the reproducibility for many biological measures, such as blood pressure, over similar time intervals.

A *valid* instrument measures what it is intended to measure. A valid instrument is reproducible, although the reverse is not necessarily true. To verify validity of an instrument, measurements are compared against accurate measurements from a "gold standard". Because no perfect gold standard exists for diet, FFQs are typically compared against superior dietary measurements, mostly commonly against repeated food records, Burke's dietary history, or repeated 24-hour recalls. Biological markers are also used. Generally, a **random sample** of subjects from the population of interest is asked to give information on their diets via multiple records or recalls. These records or recalls need to span different days of the week and different seasons for the average across measurements to match closely the person's true average diet. Timing of the administration of the FFQ is problematic. Administration prior to the recalls and records prevents the more detailed record keeping from altering the questionnaire responses, that is, a learning effect. However, the reference periods will be different, in that the questionnaire will ask about the year prior to the period of detailed dietary information. Administration several months after the last record or recall has the benefit that the reference periods will be similar. As the participants cannot be told at the beginning that a subsequent questionnaire will be requested, drop out can be a problem; the records or recalls of persons unwilling to do the subsequent questionnaire will be unusable. Also, fatigue may detrimentally affect the quality of responses to the FFQ. It is crucial that the same food composition table be used to compute the nutrient intakes from both dietary methods so that the difference between them cannot be attributed to differences in the nutrient computation methods.

Ideally, correlation between the test method (FFQ) and the gold standard should only reflect the extent to which the FFQ measures diet accurately. However, the food record, the 24-hour recall, and the FFQ methods all require the participants to record their diets, which may induce spurious correlations: a person who is a poor recorder of diet will tend to report low intake values on all instruments, regardless of true intake. In this regard, use of biomarkers as the gold standard in validation studies seems appealing, as the measurement procedure is completely unrelated to recording of the diet. However, there are serious limitations to their use, as few biomarkers exist and those that do, such as doubly-labeled water and 24-hour urine nitrogen, only measure current nutritional status, not "usual" status. In addition, biomarkers are not only inordinately expensive, but they can only validate one nutrient at a time, and are dependent on a person's metabolism as well as dietary intake.

Analysis of Validation Studies

Measurements of validity need to be adjusted or stratified (*see* **Stratification**) for variables that will be controlled for in the final epidemiologic analysis. For instance, validation should be performed separately for different gender, age, ethnic, and education groups if the questionnaire is intended for use in these groups. Inclusion of groups with diverse diets in the validation study will increase between-person variation which will inflate the correlation. The between-person variance will be reduced through stratification or adjustment in the epidemiologic analysis, and therefore the adjustment should be performed in the validation study also. For instance, without adjustment for sex, a FFQ that poorly assesses calories could yield a high correlation with a "gold standard", simply because it is able to distinguish the substantial difference in caloric intake between men and women.

Several statistical methods can be used to compare the performance of a dietary questionnaire with that of a "gold standard". The choice of statistical method will depend on the intended use of the

validation study. Means and standard deviations can be presented for both methods, and a **paired *t* test** used to compare the means. For most etiologic studies, it may be sufficient for a questionnaire simply to rank people correctly, in which case a systematic overestimation or underestimation in nutrient values will not bias the exposure–disease association. A single number comparing the two dietary methodologies for each nutrient is desirable as a moderate number of nutrients are generally compared in a validation study. The most frequently used agreement measure is Pearson's correlation coefficient r on log transformed nutrient data, which measures the linear relation between the two measurements. A disadvantage to this statistic is that it depends not only on the agreement between the methods, but also on between-person variability in the population. Use of r, however, allows for easy comparison with past studies. The intraclass correlation coefficient, defined as $r_I = (\sigma_\alpha^2 - \sigma_\varepsilon^2)/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$ from the variance components model, measures **agreement**, rather than correlation, because it accounts for the between-person variability. The r_I can be thought of as the proportion of the total variation accounted for by between-person variability. Often in a validation study, an investigator would like to know if the questionnaire performs equally well for different groups. The correlation coefficients can be statistically compared between subgroups by converting the correlations to Fisher z statistics (*see* **Correlation**), which approximately have a **normal distribution**, and comparing the z' values by a **chi-square statistic** [40].

The **kappa** statistic is a measure of agreement for **nominal** variables that adjusts for chance agreement. To use kappa in validation studies of dietary questionnaires, the nutrients from each of the two measurements must be categorized. A disadvantage to this statistic is that the agreement depends on the categorization, and it must be decided whether to use different cutpoints for the food records and the questionnaire, in which case the kappa measures correlation, or whether to use the same cutpoints, in which case agreement is measured. A weighted kappa is a generalization where cells other than those representing complete agreement are counted as partial agreements; with specific weights, the weighted kappa is related to the intraclass correlation coefficient. Regression coefficients cannot be used to measure the strength of the relationship, since the slope is not scale-free but depends on the standard deviations of the measurements.

Numerous studies have been conducted to validate the FFQ against repeated food records or 24-hour recalls, or Burke's diet history [25]. A summary of the results is presented in Table 11. It can be seen that the correlation varies substantially between studies, attributed to differences in the period between assessments, the number of repeated food records or recalls, and the populations studied. Most correlations appear to be in the range of 0.35 to 0.60. In one of the most detailed validation studies, women completed a FFQ at the beginning of the study, collected four 1-week food records at 3-month intervals, and then completed another FFQ [44]. In this study, the correlations ranged from 0.28 for iron to 0.61 for carbohydrate. Adjustment for caloric intake tended to improve the correlations. Questionnaires can be validated for food intake as well, although the correlations tend to be low because of the high within-person variability for foods.

Table 11 Pearson's correlation coefficients between FFQs and a superior dietary assessment method, adapted from [25]

Nutrient	Number of investigations	Median	Range	Interquartile range
Energy (kcal)	12	0.48	0.29–0.74	0.35–0.58
Protein	10	0.42	0.18–0.80	0.41–0.58
Carbohydrate	7	0.48	0.27–0.60	0.42–0.58
Fat	12	0.52	0.08–0.94	0.36–0.59
Cholesterol	6	0.50	0.42–0.67	0.46–0.60
Vitamin C	9	0.46	0.33–0.64	0.38–0.58
Vitamin A	6	0.40	0.21–0.63	0.33–0.51
Vitamin E	4	0.49	0.39–0.64	0.44–0.56
Calcium	3	0.63	0.61–0.66	–

Statistical Methods

As noted previously, nutritional epidemiology investigates the influence of the diet (exposure variable) on disease occurrence or death (outcome variable). Because nutritional epidemiology is a branch of chronic disease epidemiology, all of the statistical principles and methods for chronic disease epidemiology [8, 9, 38] are applicable to nutritional epidemiology. Thus, the exposure–disease association in nutritional epidemiology is quantified by the odds ratio, **risk** ratio, or **hazard ratio**, depending on the study design, and the statistical models used to estimate this association include the **logistic regression**, **Cox regression**, and **Poisson regression** models. What is unique about nutritional epidemiology is that the primary exposure variable (a person's long-term diet) is imprecisely measured, due to considerable within-person variation and measurement errors. As highlighted in the previous sections, many of the statistical methods were developed to address this problem. In this section, several other statistical topics that are especially germane to nutritional epidemiology will be highlighted.

Grouping of the Exposure Variable

In the statistical analysis of data from nutritional epidemiology, the exposure variable (food or nutrient intake) is often initially categorized into approximately evenly sized groups, such as tertiles, quartiles or quintiles (*see* **Quantiles**), and then entered into the appropriate statistical model as indicator or dummy variables. The cutpoints are generally based on the distribution of the controls, although the joint distribution of cases and controls can also be used and provides the best dispersion of counts by exposure category. This grouping of continuous exposure variables serves several purposes: the distribution of a nutrient tends to be skewed to the right and categorization dampens the effect of the extreme values; the effect of measurement error is reduced in that the dietary questionnaire needs only to categorize people into broad categories of intake; and the relationship between exposure and risk of disease occurrence can be assessed for **dose–response** trend. If the exposure is to be used as a continuous variable, then it is important to check that the relationship is indeed monotonic prior to analysis. Statistical evaluation of trend can be performed in a number of ways, such as using a continuous predictor or assigning scores to categories [38].

Multiple Comparisons

Some dietary questionnaires have more than 100 food items, which are then converted to between 20 and 30 nutrients. Statistical analysis of exposure–disease associations for all of these exposure variables will give rise to many statistical tests of "significance". Moreover, statistical analysis if often repeated in subgroups defined by such factors as gender, race, age, or subcategory of disease (e.g. stage of disease). It is not uncommon to perform an overwhelmingly large number of statistical tests from a single study, giving a high likelihood of finding many "statistically significant" test results purely by chance (*see* **Multiple Comparisons**). It is crucial that the investigator is aware of this problem. It is equally important to make clear in the research report which hypotheses constitute a priori or a posteriori tests. A stricter criterion for reporting "statistical significance" should be deployed for a posteriori tests.

Multicollinearity

Certain types of eating patterns are generally seen together in individual diets, such as high fat and low fiber diets. These eating patterns sometimes create very high correlations, or multicollinearity, between dietary exposures, leading to problems in model estimation. If two exposure variables, such as fat and dietary fiber, are highly inversely correlated, then the regression coefficient (and hence the effect measure) for fat will vary depending on whether dietary fiber is in the model. Therefore, with multicollinearity, the regression coefficient does not reflect any underlying effect of the variable on disease, but rather a marginal effect that depends on what other variables are included in the model. Additionally, standard errors of regression coefficients are inflated when the independent variables in the model are highly correlated with each other, and the correlated variables may not individually be statistically significant even if there is a strong relationship between the set of predictor variables and the outcome variable. In nutritional epidemiologic studies, correlations between foods and nutrients need to be investiaged prior to building a model with multiple nutritional predictors.

Remedial measures for multicollinearity, such as ridge regression, are available. However, effects of nutrients with near perfect correlation cannot be estimated separately.

Energy Adjustment Methods

Almost all of a person's total energy (caloric) intake is contributed by three macronutrients: intake of fat, protein, and carbohydrate. (Alcohol intake may also contribute substantially to some peoples' energy intake.) The current thinking of some nutritional epidemiologists is that it is not enough to estimate the effect of a food or nutrient on the risk of disease without giving due consideration to total energy intake [44]. For example, if high fat and energy intake were found to elevate the risk for colon cancer, then it is important to distinguish whether the apparent effect of fat intake on colon cancer actually acts through its contribution to the energy content (higher fat intakes results in higher energy content) or whether there is a specific effect of fat, independent of energy intake, on colon cancer. Statistically, what is needed is to estimate the association between fat intake and colon cancer adjusting for energy intake. Another justification for energy adjustment is that the same amount of a nutrient consumed will have less potency on a large person than on a smaller person (here, energy intake can be thought of as a surrogate for body size). Energy adjustment has also been advocated for micronutrients, such as vitamins and minerals, even though they have no appreciable energy content.

Not all nutritional epidemiologists agree with the need for energy adjustment in the estimation of nutrient-disease associations, and even though a variety of statistical methods have been proposed for energy adjustment, none is generally accepted as a standard. Indeed, energy adjustment is currently highly contentious [10, 17–19, 29, 30, 42, 44–46]. This section presents a brief sketch of the four proposed energy adjustment models. In these models, D denotes disease status, N denotes calories from the nutrient of interest, and T denotes total caloric intake. The exact specification of the regression model $M(\cdot)$ is not given, but the logistic and Cox **proportional hazards** models are the common choices.

Standard Model. $M(D) = \beta_{0S} + \beta_{1S}N + \beta_{2S}T + \varepsilon$, where the variables N and T are entered in the model simultaneously.

Residual Model. $M(D) = \beta_{0R} + \beta_{1R}R + \beta_{2R}T + \varepsilon$, where R is the **residual** from the linear regression model of N on T : $N = \alpha_0 + \alpha_1 T$.

Partition Model. $M(D) = \beta_{0P} + \beta_{1P}N + \beta_{2P}(T - N) + \varepsilon$, where the caloric intake is partitioned into calories from the nutrient of interest (N) and those from other sources ($T - N$).

Nutrient Density Model. $M(D) = \beta_{0N} + \beta_{1N}(N/T) + \beta_{2N}T + \varepsilon$, where the calories from the nutrient are divided by the total calories (N/T) to give the proportion of calories from the nutrient of interest.

The following identities show that the first three models are in fact equivalent when N and T are continuous. However, the models are not equivalent when either N or T or both are categorized and represented by indicator variables. Brown et al. [10] noted that the residual model is the most powerful and **robust** of the three models with categorized variables.

Relationship between standard and residual models:

$$\beta_{0S} = \beta_{0R} - \alpha_0\beta_{1R},$$
$$\beta_{1S} = \beta_{1R},$$
$$\beta_{2S} = \beta_{2R} - \alpha_1\beta_{1R}.$$

Relationship between standard and partition models:

$$\beta_{0S} = \beta_{0P},$$
$$\beta_{1S} = \beta_{1P} - \beta_{2P},$$
$$\beta_{2S} = \beta_{2P}.$$

Relationship between residual and partition models:

$$\beta_{0P} = \beta_{0R} - \alpha_0\beta_{1R},$$
$$\beta_{1P} - \beta_{2P} = \beta_{1R},$$
$$\beta_{2P} = \beta_{2R} - \alpha_1\beta_{1R}.$$

No one model appears clearly superior to another. The choice of model with continuous variables may be guided by the meaning of the parameters contained in each model. In the standard model, β_{1S} measures the effect on D of increasing N by 1 unit while keeping total calories constant, that is, the effect of substituting calories from sources other than N (denote by N') with calories from N. β_{2S} represents the effect on D of increasing T by 1 unit while

keeping N constant, that is, the effect of increasing calories from N'.

In the residual model, the above identities show that β_{1R} also measures the effect of substituting calories from N' with calories from N. The parameter β_{2R} represents the effect on D of increasing T by 1 unit while holding R constant. Recall that $R = N - \alpha_0 - \alpha_1 T$. Substituting $T' = (N + N') + 1$ into the equation, R is constant only when an increase of 1 unit in N' is matched with a concomitant increase in N of $\alpha_1/(1 - \alpha_1)$ units.

The partition model parameter β_{1P} measures the effect of increasing N by 1 unit while holding N' constant, that is, the effect on D of adding 1 calorie from N. By the identity with β_{2S}, β_{2P} represents the effect of increasing calories from N' by 1 unit.

The nutrient density model has a more complex structure that involves the reciprocal of caloric intake, making its coefficients rather difficult to interpret.

The partition model appears to have the parameters with the most straightforward interpretation, although it cannot be used to energy-adjust food or micronutrient intake. Pike et al. [29, 30] point out that the interpretation of β_1 in all four models is complicated by the fact that N' is itself made up of many dietary components. Suppose the nutrient of interest is fat. A β_{1S} of 0 would indicate that a substitution of 1 nonfat calorie with 1 fat calorie has no effect, that is, that calories from different sources have the same influence on the risk of D. This conclusion would be incorrect if calories from carbohydrate have a protective effect and calories from protein and alcohol have a direct effect on risk. The effect of calories from fat is being compared against the average effect of calories from the other components in these models.

Wacholder et al. [42] argue that it is not possible to distinguish the generic effect on the delivery of energy from the nutrient of interest and any specific effect of that nutrient. They point out that the true model of interest is

$$M(D) = \beta_0 + \beta_N N + \beta_{N'} N' + \beta_T T + \varepsilon,$$

which is a non**identifiable** problem because $T = N + N'$. All four models given above are special cases of this model, where one of the parameters is excluded. Therefore, the parameters in the energy adjustment models are confounded by the missing parameter and cannot clearly distinguish between a generic caloric effect and a specific nutrient effect. It is clear that when energy and nutrient intakes are too highly correlated, the variables measure nearly the same function (highly **collinear**), and their effects on D cannot be segregated. In this case, the residuals from the regression of N on T will have limited variability, as most of it will have been explained by calories, and should not be used to represent an independent exposure variable.

Energy adjustment methods can be used as a tool to investigate the joint effects of energy and individual nutrients on disease risk, but it is clear that much care must be taken in its application and interpretation.

Acknowledgment

This project was supported in part by NIH grants P01-CA-33619 and P30-CA-71789 from the National Cancer Institute.

References

[1] Armstrong, B.G. (1985). Measurement error in the generalised linear model, *Communications in Statistics – Simulation and Computation* **14**, 529–544.

[2] Armstrong, B.G. (1990). The effects of measurement errors on relative risk regressions, *American Journal of Epidemiology* **132**, 1176–1184.

[3] Armstrong, B. & Doll, R. (1975). Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices, *International Journal of Cancer* **15**, 617–631.

[4] Armstrong, B.G., Whittemore, A.S. & Howe, G.R. (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer, *Statistics in Medicine* **8**, 1151–1163.

[5] Beaton, G.H., Milner, B.A., Corey, P., McGuire, V., Cousins, M., Stewart, E., de Ramos, M., Hewitt, D., Grambsch, P.V., Kassim, K. & Little, J.A. (1979). Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation, *American Journal of Clinical Nutrition* **32**, 2456–2259.

[6] Birkett, N.J. (1992). Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure, *American Journal of Epidemiology* **136**, 356–362.

[7] Blumberg, J. & Block, G. (1994). The alpha-tocopherol, beta-carotene cancer prevention study in Finland, *Nutrition Reviews* **54**, 242–250.

[8] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. I: *The Analysis of Case-Control Studies*. World Health Organization, International Agency for Research on Cancer, Lyon.

[9] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of*

Cohort Studies. World Health Organization, International Agency for Research on Cancer, Lyon.

[10] Brown, C.C., Kipnis, V., Freedman, L.S., Hartman, A.M., Schatzkin, A. & Wacholder, S. (1994). Energy adjustment methods for nutritional epidemiology: the effects of categorization, *American Journal of Epidemiology* **139**, 323–338.

[11] Cooper, C., Barker, D.J.P. & Wickham, C. (1988). Physical activity, muscle strength, and calcium intake in fracture of the proximal femur in Britain, *British Medical Journal* **297**, 1443–1446.

[12] Copeland, K.T., Checkoway, H., McMichael, A.J. & Holbrook, R.H. (1977). Bias due to misclassification in the estimation of relative risk, *American Journal of Epidemiology* **105**, 488–495.

[13] Dosemeci, M., Wacholder, S. & Lubin, J. (1990). Does misclassification of exposure always bias a true effect toward the null value?, *American Journal of Epidemiology* **132**, 746–748.

[14] Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564–569.

[15] Haenszel, W. & Kurihara, M. (1968). Studies of Japanese migrants. I. Mortality from cancer and other diseases among Japanese in the United States, *Journal of the National Cancer Institute* **40**, 43–68.

[16] Hartman, A.M., Brown, C.C., Palmgren, J., Pietinen, P., Verkasalo, M., Myer, D. & Virtamo, J. (1990). Variability in nutrient and food intakes among older middle-aged men. Implications for design of epidemiologic and validation studies using food recording, *American Journal of Epidemiology* **132**, 999–1012.

[17] Howe, G.R. (1989). The first author replies. Re: Total energy intake: implications for epidemiologic analyses, *American Journal of Epidemiology* **129**, 1314–1315.

[18] Howe, G.R., Miller, A.M. & Jain, M. (1986). Re: Total energy intake: implications for epidemiologic analyses, *American Journal of Epidemiology* **124**, 157–159.

[19] Kipnis, V., Freedman, L.S., Brown, C.C., Hartman, A., Schatzkin, A. & Wacholder, S. (1993). Interpretation of energy adjustment models for nutritional epidemiology, *American Journal of Epidemiology* **137**, 1376–1380.

[20] Kohlmeier, L. & Helsing, E., eds (1989). *Epidemiology, Nutrition and Health: Proceedings of the First Berlin Meeting on Nutritional Epidemiology, Berlin, 1988*. Smith-Gordon, London; Nishimura Company Niigata-Shi, Japan, pp. 1–109.

[21] Kolonel, L.N., Yoshizawa, C.N. & Hankin, J.H. (1988). Diet and prostatic cancer: a case-control study in Hawaii, *American Journal of Epidemiology* **127**, 999–1012.

[22] Kupper, L.L. (1984). Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies, *American Journal of Epidemiology* **120**, 643–648.

[23] Lau, E., Donnan, S., Barker, D.J.P. & Cooper, C. (1988). Physical activity and calcium intake in fracture of the proximal femur in Hong Kong, *British Medical Journal* **297**, 1441–1443.

[24] Liu, K., Stamler, J., Dyer, A., McKeever, J. & McKeever, P. (1978). Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol, *Journal of Chronic Diseases* **31**, 399–418.

[25] Margetts, B.M. & Nelson, M., eds. (1991). *Design Concepts in Nutritional Epidemiology*. Oxford University Press, New York.

[26] Marshall, J.R., Priore, R., Graham, S. & Brasure, J. (1981). On the distortion of risk estimates in multiple exposure level case-control studies, *American Journal of Epidemiology* **113**, 464–473.

[27] Multiple Risk Factor Invention Trial Research Group (1982). Multiple risk factors intervention trial: risk factor changes and mortality results, *Journal of the American Medical Association* **248**, 1465–1477.

[28] Nelson, M., Black, A.E., Morris, J. & Cole, T.J. (1989). Between and within subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision, *American Journal of Clinical Nutrition* **50**, 155–167.

[29] Pike, M.C., Bernstein, L. & Peters, R.K. (1989). Re: Total energy intake: implications for epidemiologic analyses, *American Journal of Epidemiology* **129**, 1312–1313.

[30] Pike, M.C., Peters, R.K. & Bernstein, L. (1993). Re: Total energy intake: implications for epidemiologic analyses, *American Journal of Epidemiology* **137**, 811–812.

[31] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**, 331–342.

[32] Rimm, E.B., Ascherio, A., Giovannucci, E., Spiegelman, D., Stampfer, M.J. & Willett, W.C. (1996). Vegetable, fruit, and cereal fiber intake and risk of coronary heart disease among men, *Journal of the American Medical Association* **275**, 447–451.

[33] Rosner, B.A. (1996). Measurement error models for ordinal exposure variables measured with error, *Statistics in Medicine* **15**, 293–303.

[34] Rosner, B. & Willett, W.C. (1988). Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing, *American Journal of Epidemiology* **127**, 377–386.

[35] Rosner, B., Spiegelman, D. & Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error, *American Journal of Epidemiology* **132**, 734–745.

[36] Rosner, B., Spiegelman, D. & Willett, W.C. (1992). Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error, *American Journal of Epidemiology* **136**, 1400–1413.

[37] Rosner, B., Willett, W.C. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error, *Statistics in Medicine* **8**, 1051–1069.

[38] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company, Boston.

[39] Sempos, C.T., Johnson, N.E., Smith, E.L. & Gilligan, C. (1985). Effects of intraindividual and interindividual variation in repeated dietary records, *American Journal of Epidemiology* **121**, 120–130.

[40] Snedecor, G.W. & Cochran, W.G. (1989). *Statistical Methods*, 8th Ed. Iowa State University Press, Ames.

[41] Wacholder, S., Armstrong, B. & Hartge, P. (1993). Validation studies using an alloyed gold standard, *American Journal of Epidemiology* **137**, 1251–1258.

[42] Wacholder, S., Schatzkin, A., Freedman, L.S., Kipnis, V., Hartman, A. & Brown, C.C. (1994). Can energy adjustment separate the effects of energy from those of specific macronutrients?, *American Journal of Epidemiology* **140**, 848–855.

[43] Willett, W. (1989). An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies, *Statistics in Medicine* **8**, 1031–1040.

[44] Willett, W. (1990). *Nutritional Epidemiology*. Oxford University Press, New York.

[45] Willett, W. & Stampfer, M.J. (1986). Total energy intake: implications for epidemiologic analyses, *American Journal of Epidemiology* **124**, 17–27.

[46] Willett, W.C. & Stampfer, M.J. (1993). Re: Total energy intake: implications for epidemiologic analyses, *American Journal of Epidemiology* **137**, 812–813.

[47] Zheng, W., Kushi, L.H., Potter, J.D., Seller, T.A., Doyle, T.J., Bostick, R.M. & Folsom, A.R. (1995). Dietary intake of energy and animal foods and endometrial cancer incidence, *American Journal of Epidemiology* **142**, 388–394.

LYNNE R. WILKENS & JAMES LEE

Nutritional Exposure Measures

In all methods of measurement of nutritional exposure, some estimate of the weight of food consumed is required, and for the determination of nutrient or other food component intake, either an appropriate description for use with food tables is needed or a portion must be available for chemical analysis. Different methods are available to assess the weight of food, and they vary in accuracy, complexity, and cost [1, 5, 6]:

1. Food frequency questionnaires are designed to assess long-term habits, over months or years, and comprise a list of foods most informative about the nutrients or foods of interest [5]. The length of this list generally does not exceed 150 items. Various methods to assess portion sizes may be used, for example fitting average portion weights derived from other data to the respondents' chosen food and frequency selections. To assess the frequency of food consumption, accompanying the food list is a multiple response grid in which respondents attempt to estimate how often selected foods are eaten. Up to ten categories ranging from never, or once a month or less, to six times per day is a usual format. Because responses are standardized, food frequency questionnaires can be analyzed quickly and easily so that large numbers of individuals can be investigated relatively inexpensively. This method has been used particularly in **cohort studies**.
2. Diet history is usually conducted by trained interviewers, who obtain more detailed information on usual foods consumed, portion sizes, recipes, and frequency of food consumption over the recent past. The diet history is less commonly used in epidemiology but is frequently used in clinical dietetics.
3. Twenty-four hour recalls are based on interviews on written information about the previous day's intake, and the actual foods consumed are described, together with information on portion weights. This method is also more costly due to the variety of foods possibly consumed (at least 5000 different food items are available in most Westernized food suppliers), all of which require estimation of portion size and individual computer coding. This method is used in **surveillance** and **cross-sectional studies**, and potentially for nested case–control studies and validation studies.
4. Daily written records of the description and amount of food are kept at the time of consumption. In some studies, the consumer is asked to weigh food as it is served. The method requires substantial resources for data entry, but if records are kept for a sufficient length of time, this method has been generally used in research to assess the accuracy of other methods such as food frequency questionnaires, and for **nested case–control studies**. However, regression dilution may be considerably underestimated because of error correlation between the reference and test method [3, 4].
5. Checklists are precoded lists of foods for rapid data entry to be completed every day for several days [2]. The consumer is given one checklist per day, and asked to check off which foods are eaten from the list. Portion sizes and the list of foods are as for food frequency questionnaires, but errors in the estimation of frequency of consumption are avoided by this method.

Food frequency questionnaires or diet history methods are used in **case–control studies** to assess diet retrospectively prior to onset of symptoms. However, estimates of past dietary consumption are closely related to present consumption, and the discrepancy between actual and recalled past diet is greater the longer the period of recall attempted [5] (*see* **Recall Bias**).

Measurement errors can arise from food table databases, assessment of portion size, daily variation, inaccurate frequency categorization in food frequency questionnaires, and underreporting. All methods have different types of error structure, so that the magnitude of the error varies according to the method and may not be predictable. Relative **validation studies**, using biomarkers of intake as the reference method, suggest that regression dilution from written record methods is considerably less than with food frequency questionnaires [3]. All methods may

be subject to systematic **bias** from under- or overreporting.

References

[1] Bingham, S. (1987). The dietary assessment of individuals: methods, accuracy, new techniques and recommendations, *Nutrition Abstracts and Reviews* **57**, 705–742.

[2] Bingham, S., Gill, C., Welch, A., Cassidy, A., Runswick, S., Sneyd, M., Thurnham, D., Key, TJ.A., Roe, L., Khaw, K.T. & Day, N.E. (1997). Validation of dietary assessment methods in the UK arm of EPIC *International Journal of Epidemiology* **26**, S137–151.

[3] Day, McKeown, N., Wong, M.Y., Welch, A. & Bingham, S. (2001). Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium *Int. J. Epidemiol.* **30**, 309–317.

[4] Kipnis, V., Midthune, D., Freedman, L.S., Bingham, S.A., Schatzkin, A., Carroll. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications *American Journal of Epidemiology* **153**, 394–403.

[5] Margetts, B.M. and Nelson, M., eds. (1997). *Design Concepts in Nutritional Epidemiology*, 2nd Ed. Oxford University Press, London.

[6] Willett, W. (1998). *Nutritional Epidemiology*. 2nd edition Oxford University Press, New York.

(*See also* **Nutritional Epidemiology**)

S. BINGHAM

Nyquist Frequency

While a **time series** $X(t)$ can often be thought of as having values for all real values of t for computation we must make recordings at discrete time points. Suppose we choose to digitize the record by taking values at time intervals Δt apart, giving $X(\Delta t), X(2\Delta t), \ldots, X(N\Delta t)$ to make inferences about the original $X(t)$.

The choice of time intervals is important since the digitizing, which samples the series, has two important consequences:

1. We have no information about phenomena which have frequencies above the *Nyquist frequency* or *folding frequency* of $1/2\Delta t$ cycles per unit time.
2. These missing effects may distort our perception of those cyclic phenomena which have frequencies below the Nyquist frequency. This effect is called *aliasing*.

Some definitions are given as follows: A function $g(t)$ is periodic if

$$g(t) = g(t \pm s) = g(t \pm 2s) = \cdots = g(t \pm ks) = \cdots,$$

and the smallest (nonzero) s value is called the period of the function. The frequency f is the number of periods per unit time; that is, $f = 1/s$ cycles per unit time. Thus, $\cos(2\pi t/s)$ has period s, while $\cos(2\pi ft)$ has frequency f. Our Nyquist frequency of $1/2\Delta t$ cycles per unit time will correspond to a minimum period of $2\Delta t$. Note that mathematicians like to work in angular frequencies $\omega = 2\pi f$ radians per unit time.

Sampling limits the frequency range because of the nature of periodic functions. Suppose $\Delta t = 1$ and the signal contains a function which is periodic; say, $g(t)$ with frequency 3/2, i.e. period 2/3. Then we observe in sequence

$$g(0), g\left(\frac{1}{3}\right), g\left(\frac{2}{3}\right) = g(0), g\left(\frac{1}{3}\right), g\left(\frac{2}{3}\right), \ldots,$$

a signal which repeats in steps of 3, giving an apparent frequency of 1/3. If we regard our periodic component $g(t)$, with period $f_b = k/2\Delta t + f_0$, $f_0 < f_N$, as being expressed in terms of complex exponentials $\exp(i2\pi f_b t)$, then, since $t = m\Delta t$, they become $\exp(i2\pi f_b m\Delta t) = \exp(i2\pi f_0)$. For an example see Figure 1.

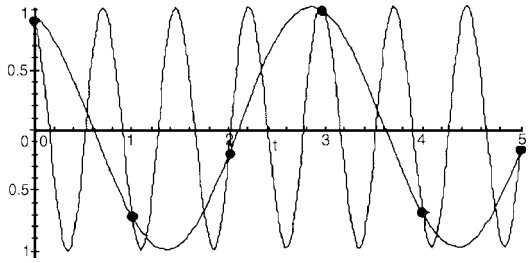

Figure 1 Alias of a signal after sampling

If the power spectrum of the original series (*see* **Spectral Analysis**) is $h(f)$, then we can show that the power spectrum of the observed, digitized, series $h_{\mathrm{d}}(f)$ is

$$h_{\mathrm{d}}(f) = \sum_{k=-\infty}^{\infty} h\left(f + \frac{k}{\Delta t}\right), \quad -\frac{1}{2\Delta t} < f \leq \frac{1}{2\Delta t}.$$

Our observed spectrum is thus the result of folding the original over the Nyquist range.

This means that the observed value of the power spectrum at f_0 is made up not only of $h(f_0)$ but also the values of the original spectrum at the aliases to f_0; that is, $f_0 \pm 1/\Delta t$, $f_0 \pm 1/2\Delta t$, $f_0 \pm 1/3\Delta t, \ldots$. So when you digitize a sequence it is vital that there are no components with appreciable power whose frequency lies outside the Nyquist range. Indeed, we can prove that if the power outside this range is exactly zero, then the original series can be reconstructed exactly from the digitized one.

Comprehensive accounts are given in [1] and [2].

References

[1] Koopmans, L.H. (1974). *The Spectral Analysis of Time Series*. Academic Press, New York.

[2] Priestley, M.B. (1985). *Spectral Analysis and Time Series*. Academic Press, London.

G.J. JANACEK

Oblimin Rotation

Oblimin rotation is a general form of performing an **oblique rotation** of vectors comprising the matrix $\mathbf{V}$ of dimension $(p \times k)$ associated with **principal components** or factors in order to transform these quantities into new variables by the relationship $\mathbf{B} = \mathbf{V\Theta}$ [$\mathbf{B}$ is a matrix of dimension $(p \times k)$ and $\mathbf{\Theta}$ is a matrix of dimension $(k \times k)$] such that $\mathbf{B}$ will approximate **simple structure** (*see* **Rotation of Axes**). Oblimin rotation is similar in nature to the *Orthomax* **orthogonal rotation** procedure. Like Orthomax, the Oblimin rotations are also quartic solutions and are a general solution of the following expression:

$$Q = \sum_{g<j=1}^{k(k-1)/2} \left[p \sum_{i=1}^{p} b_{ij}^2 b_{ig}^2 - c \left(\sum_{j=1}^{p} b_{ij}^2 \right) \times \left(\sum_{i=1}^{p} b_{ig}^2 \right) \right],$$

where p is the number of original variables, k is the number of retained components or factors, b_{ij} are the coefficients of the vectors defining the rotation, and c is an arbitrary constant. Unlike the Orthomax procedure, Oblimin rotations require Q to be *minimized*.

The Oblimin expression includes a number of procedures that have been derived independently. *Covarimin* rotation [1, 7], which is obtained by setting $c = 1$, has been criticized by some as being too close to an orthogonal rotation. Similarly, setting $c = 0$ produces the *Quartimin* rotation [1, 6] which has been felt to be too oblique. As a compromise, Carroll [2] derived the *Biquartimin* rotation by setting $c = 1/2$. Standard errors for the vector coefficients produced by Oblimin rotation were given by Jennrich [4] and Clarkson [3].

There are several versions of these procedures. The expression above is referred to as *raw* Oblimin. In a manner similar to **Varimax rotation**, *normal* Oblimin may be obtained by dividing each b_{ij}^2 or b_{ig}^2 by the corresponding diagonal term of $\mathbf{VV}'$. The original Oblimin procedures involved both the primary and reference vectors. A later procedure, called *Direct Oblimin* [5, 6], simplified this by doing away with the reference vectors.

The direct Oblimin rotation is available in the SPSS software package (*see* **Software, Biostatistical**) [8]. Table 1 provides an example of the direct Oblimin solution with the original principal component characteristic vectors. This example deals with depression data collected in the Framingham Study. The principal components analysis is performed on a sample of 1660 subjects (*see* **Principal Components Analysis** for data description).

Table 1 Framingham depression data: characteristic and direct Oblimin rotated vectors

	Characteristic vectors			Direct Oblimin rotation		
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$
Effort	0.60	0.15	0.41	0.12	0.08	0.67
Restless	0.39	0.07	0.55	−0.06	−0.10	0.70
Depress	0.77	−0.13	−0.10	0.69	0.08	0.16
Happy	0.70	−0.23	−0.06	0.68	−0.05	0.15
Lonely	0.64	−0.23	−0.21	0.71	0.01	−0.02
Unfriend	0.35	0.67	−0.33	0.04	0.83	−0.06
Enjoylife	0.52	−0.27	−0.27	0.69	−0.03	−0.13
Feltsad	0.72	−0.23	−0.20	0.76	0.02	0.02
Disliked	0.34	0.72	−0.22	−0.06	0.83	0.06
Getgoing	0.58	0.20	0.47	0.04	0.10	0.73

In recent years, these solutions seemed to have lost favor to some two-stage procedures such as **Orthoblique** or Harris–Kaiser rotation, **Promax rotation**, and **Optres rotation**.

References

[1] Carroll, J.B. (1953). An analytical solution for approximating simple structure in factor analysis, *Psychometrika* **18**, 23–38.

[2] Carroll, J.B. (1957). Biquartimin criterion for rotating to oblique simple structure in factor analysis, *Science* **126**, 1114–1115.

[3] Clarkson, D.B. (1979). Estimating the standard errors of rotated factor loadings by jackknifing, *Psychometrika* **44**, 297–314.

[4] Jennrich, R.I. (1973). Standard errors for obliquely rotated factor loadings, *Psychometrika* **38**, 593–604.

[5] Jennrich, R.I. (1979). Admissible values of γ in direct oblimin rotation, *Psychometrika* **44**, 173–177.

[6] Jennrich, R.I. & Sampson, P.F. (1966). Rotation for simple loadings, *Psychometrika* **31**, 313–323.

[7] Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

[8] *SPSS-X™ Users' Guide*, 3rd Ed. SPSS Inc., Chicago.

(*See also* **Axes in Multivariate Analysis**; **Factor Analysis, Overview**)

J. EDWARD JACKSON

Oblique Rotation

Given a matrix $\mathbf{V}$ of dimension $(p \times k)$ consisting of a set of k vectors, usually defining a set of **principal components** or factors, a new set of transformed variables may be obtained by a rotation of $\mathbf{V}$, namely $\mathbf{B} = \mathbf{V}\boldsymbol{\Theta}$. $\mathbf{V}$ is often the **factor loading matrix** or factor matrix from the initial step in a principal components analysis or a **factor analysis**. Such a rotation is said to be **oblique** if the resultant rotated axes in the vector space are not at right angles to each other. Most rotation procedures are designed to approximate a **simple structure**. The matrix $\boldsymbol{\Theta}$ of dimension $(k \times k)$ defines the angles of rotation and the matrix $\mathbf{B}$ of dimension $(p \times k)$ defines the vectors determining the new variables obtained by rotation.

The purpose of using oblique rotation is to obtain a matrix $\mathbf{B}$ that exhibits a better pattern of simple structure than would be obtained by using an orthogonal rotation. This improvement towards a simple structure comes at the cost of loss of **orthogonality**. If the main purpose of rotation is to cluster groups of the original variables, then this may present no problem. As in the case of **orthogonal rotation**, there are many methods of obtaining oblique rotations and, again, there is a general quartic formula, the **Oblimin rotation**, which produces a number of these procedures. Unlike orthogonal rotation, the more popular oblique procedures are not members of this family and are generally two-stage procedures. Among these procedures are **Orthoblique** or Harris–Kaiser, **Promax**, and **Optres rotations**.

(*See also* **Axes in Multivariate Analysis**)

2 Oblique Rotation

J. EDWARD JACKSON

Observational Study

An observational study is a study in which conditions are not under the control of the investigator, unlike an **experimental study**. In particular, the exposures or treatments of interest are not assigned at random to experimental units by the investigator (*see* **Randomization**). Thus, **associations** between exposure and health outcome, say, may result from **confounding** by factors associated both with exposure and outcome.

Epidemiologic studies of disease etiology in humans are almost always observational because it is unethical to allocate people to receive potentially harmful exposures (*see* **Ethics of Randomized Trials**). Although the investigator does not control the allocation of exposure in observational studies, it is possible to mimic experimental designs in many respects and, by proper collection of observational data, to examine critically the hypothesis that the exposure has a causal impact (*see* **Causation**) on health outcome. The process of causal induction from observational data was brilliantly described by Hill [1, 2] (*see* **Hill's Criteria for Causality**).

Observational data are particularly subject to confounding in studies of therapeutic effects because factors that cause a doctor or patient to select a particular treatment are also often strongly related to health outcome. Such confounding has been called 'confounding by indication' [3]. Whenever possible, an **experimental design**, the controlled **clinical trial**, should be used to evaluate such treatments.

References

[1] Hill, A.B. (1953). Observation and experiment, *New England Journal of Medicine* **248**, 995–1001.

[2] Hill, A.B. (1965). The environment and disease: association or causation, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[3] Miettinen, O.S. (1983). The need for randomization in the study of intended effects, *Statistics in Medicine* **2**, 267–271.

MITCHELL H. GAIL

Observer Reliability and Agreement

Many measurements in medical practice and research are based on observations made by clinicians. As these measurements are prone to error, observer reliability and agreement are important issues in medicine. The terms "observer reliability" and "**agreement**" are often used interchangeably, but in theory they are different concepts. Reliability coefficients express the ability to differentiate among subjects. They are ratios of **variances**: in general, the variance attributed to the difference among subjects divided by the total variance [11, 12]. Agreement refers to conformity. Agreement parameters determine whether the same value is achieved if a measurement is performed twice, either by the same observer or by different observers [4]. In homogeneous populations one can imagine that reliability might be low while agreement is high; in a heterogeneous population, reliability and agreement measures will correspond well [13].

The parameters for assessment of observer reliability and agreement differ according to the scale of measurement. The possible scales of measurement are: categorical data on a **nominal** scale (e.g. the judgment of presence or absence of a sign or symptom); categorical data on an ordinal scale (e.g. judgment of the degree of severity of a lesion); and data on a continuous scale (e.g. the measurement of blood pressure). For each of these scales, reliability and agreement parameters will be presented and discussed. Table 1 summarizes several characteristics of the parameters.

Categorical Data

With a **binomial** outcome, the results of two observers rating N subjects or one observer rating N subjects twice, may be presented in a **two-by-two table** as shown in Table 2 [3]. The level of agreement is 50% (both positive scores) plus 20% (both negative scores). However, this measure does not discriminate between actual agreement and agreement which arises due to chance. A measure which attempts to correct for chance agreement is the **kappa** (κ) coefficient [4]. κ represents the extra amount of agreement observed above chance ($p_o - p_e$), divided by the amount of agreement which could maximally occur above chance ($1 - p_e$). In Table 2, the observed proportion of agreement (p_o) is the proportion of X-ray films agreed on by the two rheumatologists as being positive (50/100), plus the proportion agreed on as negative (20/100), i.e. 0.70. The expected proportion of chance agreement is calculated assuming independence of the observers: $65/100 \times 65/100 = 0.42$ of the X-rays would be scored positive and $35/100 \times 35/100 = 0.12$ would be scored negative by both observers: $p_e = 0.54.\kappa = (p_o - p_e)/(1 - p_e) = (0.70 - 0.54)/(1 - 0.54) = 0.35$. Usually, κ-values lie between 0 and +1, where 0 indicates only chance agreement and 1 indicates perfect agreement. However, κ-values can be negative, when there is less agreement than expected by chance (*see* **Kappa and its Dependence on Marginal Rates**).

Table 1 Characteristics of the presented parameters of agreement

Characteristics	κ	κ_w	ICC	Generalizability study	Limits of agreement
Scale	Categorical nominal	Categorical ordinal	Continuous	Continuous	Continuous
Measure of reliability or agreement	Agreement	Agreement	Reliability	Reliability	Agreement
Distinguish between random and nonrandom errors	No	No	No	Yes	Yes
Applicable to more than two observers	Yes	Yes	Yes	Yes	No
Expressed in metric unit of measurement	No	No	No	Yes	Yes

κ = kappa
κ_w = weighted kappa
ICC = intraclass correlation coefficient

Table 2 Agreement between two rheumatologists rating hand radiographs of 100 patients according to presence or absence of evidence of erosions [3]

	Rheumatologist B		
Rheumatologist A	Present	Absent	Total
Present	50	15	65
Absent	15	20	35
Total	65	35	100

In Table 3, an example is presented where both **random errors** and nonrandom errors in the measurements occur [3]. Two rheumatologists again show 70% agreement, but now rheumatologist A is more likely to score X-rays as positive (75%) than rheumatologist B (55%). The κ-value is 0.37, indicating a level of agreement which is slightly higher than for the example in Table 2.

The κ-coefficient can be extended for observations with more than two nominal categories [2]. If the number of categories increases, the opportunities for disagreement will increase, and consequently κ will tend to be lower.

In the situation of a scale with more than two classes with a logical sequence (ordinal scale), a weighted κ-coefficient (κ_w) has been proposed [5]. This reflects the fact that disagreements between adjacent categories are less serious than disagreements over more categories. κ_w adjusts for the seriousness of disagreement by assigning weights (between 0 and 1) to partial agreement cells, where 0 means total disagreement and 1 total agreement. Using quadratic weights in this calculation, κ_w becomes equivalent to the intraclass correlation coefficient [10] (*see* **Correlation**). This shows the similarity of reliability and agreement under specific conditions.

κ is the most widely accepted measure of agreement when considering categorical data. κ includes both random errors and nonrandom errors. However, the interpretation of κ is difficult. The practice of calculating a **confidence interval** for κ and assessing whether it differs statistically significantly from zero is of no use: the question is not whether an association is present or not, but how close to perfect the observer agreement is. Although proposals for interpretation of different κ-values have been made, the interpretation is not as straightforward as suggested [1, 9]. Several factors should be taken into account. First, κ is dependent on the number of classes. With more classes it is more difficult to classify the subjects correctly and lower κ-values are usually found. Secondly, κ is dependent on the **prevalence** of the attribute being measured. High underlying prevalences result in a high level of expected agreement, leaving less room for actual agreement. Supposing that the rheumatologists both scored 80% of the X-rays as positive, agreeing in 65% of the X-rays on positive scores and 5% on negative scores. In that case, the agreement would again be 70%, but the κ-value would only be 0.06. Thirdly, the κ-value also depends on differences in the marginals of the two observers, that is, in the presence of **bias**. In the case of bias, the κ-value can, paradoxically, become higher [7, 8]. A single κ-value does not differentiate between random and nonrandom errors. Therefore, presentation of the complete table together with the κ-value is very important. The table shows the prevalences of the scores, whether there is bias as well as random errors and, in the case of more classes, which classes are most difficult to distinguish. This information is indispensable for a proper interpretation of the κ-value and forms the basis for improvement of the observer agreement.

Table 3 Agreement between two rheumatologists rating hand radiographs of 100 patients for presence of erosions with bias in their evaluation [3]

	Rheumatologist B		
Rheumatologist A	Present	Absent	Total
Present	50	25	75
Absent	5	20	25
Total	55	45	100

Continuous Data

Intraclass Correlation Coefficient

The intraclass correlation coefficient (ICC) is a measure observer reliability designed for continuous variables, although it can also be used for ordinal data. The ICC is defined as the ratio of the **variance** of interest (often the variance between subjects) to the total variance [11, 12]. These variances are derived from **analyses of variance** (ANOVA). The structure of the ANOVA model depends, among other things, on whether the observers are drawn at random from a large population of observers (**random effects**) or whether they are the only observers of interest

(**fixed effects**), and on whether each observer rates each subject or not [11, 12]. These factors determine the appropriate ICC formula. The ICC includes both random errors and systematic differences. The ICC ranges from 0 with no agreement to +1 with perfect agreement.

The ICC avoids the problem of the Pearson correlation coefficient that a linear relationship is mistaken for agreement, but is, like other correlation coefficients, dependent on the range of the variables measured. With larger ranges, that is, in a more heterogeneous population, the value of ICC is higher. This reflects the fact that in heterogeneous populations subjects are easier to distinguish than in homogeneous populations. Although the ICC is designed to measure how well patients can be distinguished from each other despite measurement errors, the ICC is also used as a measure of agreement. In this case, caution should also be exercised by comparing ICCs between populations and by extrapolating results on ICCs to populations which differ with respect to heterogeneity [13]. The ICC is a ratio of variances and, therefore, difficult to interpret clinically. Presentation of its **variance components**, whose square roots are expressed in the metric units of measurement (*see* **Unit of Analysis**), would be more informative clinically.

Generalizability Studies

A more extensive analysis using ANOVA is proposed under the term "generalizability" studies [6]. In any measurement situation there are multiple sources of error variance. Besides intra-observer and inter-observer disagreements, variability among subjects may arise on different days, after different diets or in stressful situations. The error variance can be calculated for each source of error. By taking the square root of the error variances the **standard errors** of measurement (SEMs) are obtained, expressed in the dimension of the original measurement. An important goal of generalizability studies is to identify and measure variance components which contribute errors to a measurement. They provide a lot of information on observer reliability. They identify sources of error (e.g. intra-observer or inter-observer) and determine the relative importance of each component. This provides useful information for strategies to prevent or minimize errors. Moreover, the SEMs resulting from the analyses have direct use in clinical practice and research. Therefore, generalizability studies are powerful tools in assessing intra-observer and inter-observer reliability (*see* **Validity and Generalizability in Epidemiologic Studies**).

Example. In a three way **factorial experiment** (random effects model), 15 patients are rated by five different observers on three different days. Assuming that the patient characteristics being judged are stable over these days, the measurements at the three different moments can be used to assess the intra-observer reliability. The results of the ANOVA analysis are presented in Table 4. From an investigation of the variance components (σ^2), observations on different days (W: within observers) or in **interaction** of days with the other two main effects (P × W, B × W, P × B × W) appear to contribute little to the total variation. A greater proportion of variance is contributed by the different observers (B: between observers), and the largest by the inter-individual differences among the patients (P) and the patient−observer interaction (P × B). The variance components corresponding to the various facets of the measurement design are the major results of a generalizability study. The square root of these variance components equals the SEM. Hence, the generalizability studies identify the sources of error, assess the relative contribution to the measurement and provide direct usable clinical information.

Limits of Agreement

A simple method which measures agreement and distinguishes between random and nonrandom errors was proposed by Bland & Altman [2]. In order to assess whether there are systematic differences between two rheumatologists A and B, the scores of the two observers are subtracted (d) and plotted against the mean of the measurements (Figure 1) [3]. The confidence interval around $d(d \pm t_{n-1}SE)$ is calculated to assess whether statistically significant bias exists between the two rheumatologists. Furthermore, limits of agreement can be calculated, based on the mean difference between the rheumatologists (d) and the **standard deviation** of these differences. Approximately 95% of the differences will lie between $d - 2$ and $d + 2$ standard deviations, which are called the limits of agreement. These formulas hold if the differences are not dependent on the

Table 4 Fictitious example of ANOVA analysis of generalizability study

Facet	SS	df	MS	σ^2 : SEM^2	% σ^2
Patients (P)	14 000	14	1000	$\sigma^2_P = (MS_P - MS_{PW} - MS_{PB} + MS_{PWB})/n_B n_W = 58.67$	53.9
Inter-observer (B)	2400	4	600	$\sigma^2_B = (MS_B - MS_{BP} - MS_{BW} + MS_{PBW})/n_P n_W = 11.33$	10.4
Intra-observer (W)	200	2	100	$\sigma^2_W = (MS_W - MS_{WP} - MS_{WB} + MS_{PBW})/n_P n_B = 0.47$	0.4
P × B	4200	56	75	$\sigma^2_{PB} = (MS_{PB} - MS_{PBW})/n_W = 23.33$	21.4
P × W	1400	28	50	$\sigma^2_{PW} = (MS_{PW} - MS_{PBW})/n_B = 9$	8.3
B × W	160	8	20	$\sigma^2_{BW} = (MS_{BW} - MS_{PBW})/n_P = 1$	0.9
P × B × W	560	112	5	$\sigma^2_{PBW} = 5 = MS_{error}$	4.6

P = patients
B = between observers
W = within observers
SS = sum of squares
df = degrees of freedom
MS = mean square
σ^2 = estimate of variance
SEM = standard error of measurement

value of the mean (e.g. larger differences with higher means). If this is not the case, **transformations** are required to make the differences independent of the mean.

The method of Bland & Altman [2] clearly visualizes systematic differences and random errors. Moreover, errors are expressed in terms of the scale of measurement (*see* **Measurement Scale**), which enables a direct clinical interpretation of the results. The minimum acceptable level of agreement depends on the clinical use and situation. Deciding whether errors are acceptable is always a question of clinical, not statistical, judgment.

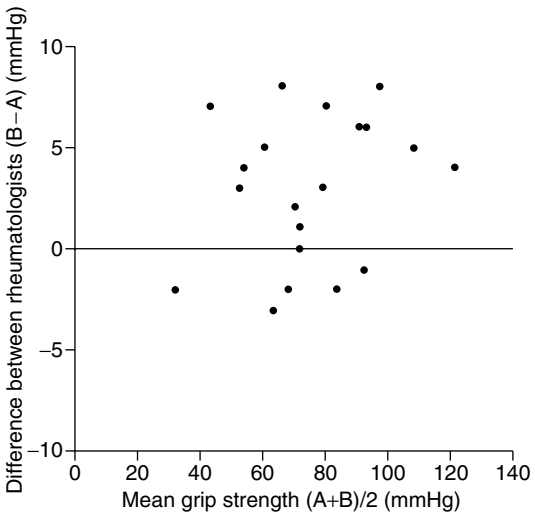

Figure 1 Difference between rheumatologist A and rheumatologist B's readings of grip strength measurements plotted against mean measurements for 20 patients [3]

Clinical Relevance

Assessing observer reliability and agreement is essential for interpretation of clinical observations both in research and in medical practice. Even more important than being aware of a suboptimal observer reliability or agreement is coping with or anticipating it. Tracing the sources and types (bias or random error) of the disagreements is the beginning of wisdom. For that purpose, presenting one single coefficient is insufficient and a visual presentation of the data is advisable. Generalizability studies, which aim to determine the origin of the variation and their relative contribution to measurement errors, are most valuable in this respect. Such studies are able to measure, among other things, the contribution of intra-observer and inter-observer variation to the total of measurement errors. The solutions for intra-observer and inter-observer disagreements have to be sought in standardization of the measurements and consensus meetings about clinical observations. Knowledge about the origins of the errors helps in this process. If the agreement cannot be improved by these

strategies, multiple measurements may be a solution. Depending on the major source of disagreement, these multiple measurements should be performed either by different observers or by the same observer. In medical research, increasing the sample size is also an option for coping with random errors. However, one should note that increasing sample size does not prevent bias. In general, improvement of observer reliability or agreement of clinical observations may have much impact on the **quality of health care**.

References

[1] Altman, D.G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall, London.

[2] Bland, J.M. & Altman, D.G. (1996). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **i**, 307–310.

[3] Brennan, P. & Silman, A. (1992). Statistical methods for assessing observer variability in clinical measures, *British Medical Journal* **304**, 1491–1494.

[4] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[5] Cohen J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213–220.

[6] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York.

[7] Feinstein, A.R. & Cicchetti, D. (1990). High agreement but low kappa: I. The problems of two paradoxes, *Journal of Clinical Epidemiology* **43**, 543–549.

[8] Feinstein A.R., Cicchetti D. (1990). High agreement but low kappa: II. Resolving the paradoxes, *Journal of Clinical Epidemiology* **43**, 551–558.

[9] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[10] Fleiss, J.L. & Cohen, J. (1973). The equivalence of weighted kappa and the intra class correlation coefficient as measures of reliability, *Educational and Psychological Measurement* **33**, 613–19.

[11] McGraw, K.O. & Wong, S.P. (1996). Forming inferences about intraclass correlation coefficients, *Psychologists* **1**, 30–46.

[12] Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability, *Psychological Bulletin* **86**, 420–428.

[13] Stratford, P. (1989). Reliability: consistency or differentiating among subjects?, *Physical Therapy* **69**, 299–300.

(*See also* **Agreement, Modeling of Categorical**)

H. DE VET

Occupational Epidemiology

Especially since the late 1970s, numerous epidemiologic studies have revealed elevated **risks** of cancer, cardiovascular disease, and neurologic and other disorders among various occupational groups [106]. Strong evidence that many disorders are work-related was unobtainable from the clinical experiences or **case reports** that used to be the basis for identifying occupational risks. The introduction of modern, rigorous principles for epidemiologic research and the integration of epidemiologic courses in **occupational health** training have played an important role in this development. Many textbooks on epidemiologic methods have appeared since the early 1980s, and some of these have specifically focused on occupational epidemiology [41, 75, 115]. The availability of computers and statistical packages also has facilitated the progress.

It is hardly possible to predict the directions in occupational epidemiology that will lead to the most important future achievements. Only the more general aspects and principles of occupational epidemiology can be illustrated here by examples drawn from the several different subject matter areas. The challenge in occupational epidemiology has been, and will be, to identify adverse agent(s) or processes rather than to associate health risks with occupational groups or titles, because successful prevention can only be based on the elimination or reduction of specific exposures. The difficulties in this respect are often considerable, however, as occupational exposures tend to be mixed, and lifestyle factors may interfere.

The refinements of methods in epidemiologic research [63] came long after the first few epidemiologic studies of occupational disorders. In 1843, **W.A. Guy** studied "pulmonary consumption" in letter press printers and identified a higher risk among compositors than among pressmen [99]. The observation in 1879 of an increased occurrence of lung cancer among Schneeberg miners [70], and the excess of bladder cancer among German aniline workers reported around the turn of the century [135], are other examples of early occupational epidemiology.

A more recent example of occupational epidemiology, from 1948, demonstrated a high **proportional mortality** of lung cancer among British workers exposed to inorganic arsenic [78]. A few years later an increased risk of lung cancer was demonstrated among gas workers [49] and also bladder cancer in rubber workers [39]. Important studies of the risk of lung cancer in asbestos workers [142] and in underground miners were published in the 1960s [164, 165]. Studies on chemically induced cardiovascular disease also appeared relatively early – for example, among workers exposed to carbon disulfide [76, 155].

There is now, in the twenty-first century, an increasing interest in the effects of work stress and psychosocial determinants of the risk of cardiovascular disease. Other recent studies concern neurologic disorders and their relationships to occupational exposures. Ergonomic risk factors and musculoskeletal disorders as well as reproductive hazards from occupational exposures are other aspects that have attracted interest since the 1980s. The health effects of electromagnetic fields have been among the most intriguing questions in occupational and **environmental epidemiology** during the 1990s [89, 141, 154], although the initiating study in this respect concerned cancer in children [166].

For the future, as for today, a central issue in occupational epidemiology will be the assessment of the effect of single as well as combined exposures. The recently developed tools of **molecular epidemiology** may increase the **power** of epidemiologic studies to detect risks at lower exposure levels and in smaller worker groups. These new tools are already being used in occupational epidemiology to define biomarkers of exposure or early effects [75, 161], to identify susceptible individuals and to specify cancers by their mutational patterns [13].

Defining Research Questions for Occupational Epidemiology

New ideas for occupational studies have come from a variety of sources. A clinical observation has often suggested a connection between a disease and an exposure; sometimes toxicologic data from animal experiments have indicated a possible health hazard. Suggestions for a study may also have originated from observations or suspicions among workers about an adverse health effect. Still other leads for study derive from an examination of death records

in various occupational groups (*see* **Occupational Mortality**) or from studies that link **census** data on job titles with cancer registry data or other **disease registries** or **causes of death**. Clues for new studies may also come from **case–control studies** that routinely tend to assess a number of exposures; should some unexpected **associations** appear, further studies are usually warranted.

The Role of Clinical Observations for Epidemiologic Research

The concerns about asbestos exposure as a cause of lung cancer and mesothelioma have probably generated more studies in occupational epidemiology than any other job-related health risk. The first suspicion of a lung cancer risk was raised by two case reports in 1935 [60, 102], and other such reports followed before the association between exposure and disease was clearly assessed in 1955, both in England [50] and California [32]. The risk of mesothelioma was not noticed until 1960; again the first suspicions arose from clinical observations [163].

The perception of possibly different risk patterns for the different types of asbestos may in part explain the numerous studies that have been conducted worldwide on asbestos exposure, but little gain in specific knowledge has been achieved in this respect. Indeed, between 1977, when the IARC (**International Agency for Research Against Cancer**) Monograph on asbestos exposure and cancer risk was published, and 1986, when this material was updated in the Monograph Supplement 7 [85], there was little new information regarding the effect of specific types of asbestos in spite of more than a doubling of the number of available studies. Several of these studies on asbestos in many countries have probably been motivated by the need to convince both the medical community and the authorities in each country with local studies on the health risk.

An even more clear-cut example of how a clinical observation initiated a large number of epidemiologic studies arose from the discovery of a cluster (*see* **Clustering**) of paranasal cancer cases among furniture workers in High Wycombe, UK [2, 104]. The many subsequent studies from various parts of the world have been convincingly consistent [85]. Similarly, a cluster of nasal cancers was traced to boot and shoe manufacturing [1], and again, this link was confirmed in many studies from several countries [85]. The rarity of this type of cancer certainly facilitated the recognition of a causal relationship (*see* **Causation**) to occupational exposures.

The report in 1974 of liver angiosarcomas among workers exposed to vinyl chloride provides still another example of how the observation of a cluster of cases of a rare tumor [46] gave rise to a number of further studies. Some of these included also other cancers and cardiovascular disease [51]. The first report on human liver angiosarcoma [46] referred to animal experiments that indicated the possibility of an oncogenic effect of vinyl chloride, but this knowledge seemed not to have reduced workplace exposures. A note added to the report on the human angiosarcomas described unpublished data showing liver angiosarcoma and other tumors in animals exposed to vinyl chloride. The consistent findings in humans and animals convincingly established the risk of liver angiosarcoma from exposure to vinyl chloride.

Studies of the association of phenoxy herbicides with soft tissue sarcomas and lymphomas represent another theme in occupational epidemiology that arose from clinical observations and a case series report [67]. In contrast to the vinyl chloride case, there has been no convincing experimental evidence of a cancer risk from phenoxy herbicides. However, in particular, the 2,4,5-trichlorophenoxyacetic acid was known to contain varying amounts of 2,3,7,8-tetrachlorodibenzodioxin and other dioxins, for which there was growing evidence of cancer risks from animal data from 1977 onwards [83, 159]. Although there has been some inconsistency among the ensuing studies, it seems that soft tissue sarcomas are mainly related to dioxin exposure, whereas the lymphomas, especially the non-Hodgkin lymphomas, might be caused by phenoxy herbicides themselves [68].

Clinical observations also have stimulated epidemiologic studies in other areas than cancer. Painters with severe neurasthenic or psychoorganic syndromes in the 1970s raised the question of a role for long-term solvent exposure. Following some initial case–control and **cohort studies** in Sweden and Denmark indicating an effect [17, 112, 119], further epidemiologic research has been conducted and essentially confirmed both acute effects and the

syndromes that may appear after long-term solvent exposure [30, 43, 160].

Animal Data Initiating Occupational Epidemiology

Sometimes the initial clue that precipitates epidemiologic research arises from animal studies. An example of strong animal evidence of a cancer risk preceding epidemiologic research concerns lung cancer among workers with exposure to chloromethyl methyl ether [54, 84]. The animal data existing in this case have apparently been so convincingly corroborating the epidemiologic findings that preventive actions were taken in many countries and few further human studies have followed.

Animal carcinogenesis studies of trichloroethylene also triggered epidemiologic investigations in the late 1970s [20, 156]. The early results were less convincing of a cancer risk, and only by aggregating the results from the later three most informative studies [5, 19, 150], and comparing the observed to expected numbers of liver and biliary tract cancers as well as non-Hodgkin's lymphomas, could an IARC Working Group conclude that there was limited evidence for a carcinogenic effect from trichloroethylene in humans [88].

Many epidemiologic studies also followed animal studies showing that formaldehyde caused cancer in the nasal cavity. An excess of nasal and nasopharyngeal cancers appeared as a fairly consistent finding in several of these ensuing studies [85]. Still, the IARC Working Group that evaluated this agent considered the available studies to provide only limited evidence for a carcinogenic effect in humans. This and the previous example indicate the problems and the latitude involved in trying to assess finally a cancer risk.

Animal data suggest carcinogenic risks from lead, cadmium, and beryllium, but epidemiologic findings have been relatively weak, although finally convincing enough for cadmium and beryllium to permit a conclusion about sufficient evidence for a carcinogenic effect also in humans [86]. A more recent meta-analysis suggests also a cancer risk for workers exposed to lead [57].

These examples notwithstanding, it is perhaps surprising that relatively few epidemiologic studies have been initiated in response to animal studies, especially in view of the large number of chemicals tested. For diseases other than cancer, there are even fewer examples of animal studies leading to epidemiologic investigations, but there is also a relative lack of animal studies about other effects than cancer. Furthermore, the principles of occupational epidemiology have been less developed for such other diseases.

Record Linkage Studies

Epidemiologists have linked mortality or cancer registry data with census or **death certificate** information on occupation (*see* **Record Linkage**) in efforts to discover new occupational health hazards. Even when there is an increased risk of some disorder in an occupational group, the imprecise measure of exposure in such linkage studies attenuates the effect, however. This dilution problem may explain why the associations found have usually been weak and have contributed relatively little new knowledge. **Census** data reflect the occupational status at a point in time (e.g. during a particular week), and are therefore inherently poor measures of the occupational exposure that may, or may not, have occurred over many years.

A source of potential **confounding** (see discussion later) in registry linkage studies is the geographic variation in disease incidence, which may be real or may reflect local preferences in diagnostic practice (*see* **Geographic Patterns of Disease**). A common job in an area with a high incidence of some disease may therefore be associated with an artifactually increased occupational risk of the disease. For example, the linkage of registry data in Sweden indicated an excess of brain cancer in glass workers, but further evaluation showed that there was also a locally increased risk for others living in the relatively small area where the glassworks were located [169].

These limitations do not imply that registry studies are futile, however. For example, a Nordic registry linkage study has contributed essential information regarding the risk of lung cancer in connection with silica exposure [103]. A **proportional mortality study** linking causes of death to job titles on the death certificates also suggested an occupational risk for leukemia from exposure to electric and magnetic fields [113]. Similarly, record linkage studies gave an early indication that multiple sclerosis was associated with solvent exposure [114] – a connection that now seems rather likely in the light of a recent meta-analysis [98].

Concerns of Workers and Others

Sometimes workers perceive adverse health effects and attract the interest of epidemiologists. For example, a group of men exposed to bromochloropropane, a pesticide for nematodes, noted that none of them had fathered any children. This risk was later confirmed in epidemiologic studies [167, 168]. In contrast, suspicions that work with video display units could cause spontaneous abortions have created considerable concern and many studies, but no consistent effect has been demonstrated [105]. There are probably many small-size negative or nonpositive studies relating to workers' anxiety about a health hazard that are unpublished and unknown but that nevertheless reassure the workers involved.

Sometimes media reports have initiated a study, as for example reports of a cluster of childhood leukemia and non-Hodgkin's lymphoma in the vicinity of the Sellafield nuclear plant in England (*see* **Leukemia Clusters**). A subsequent case–control study suggested that paternal exposure could have been the cause [59]; much controversy and other studies have followed, and the risks remain unclear. In general, however, there are surprisingly many observations that paternal exposures may play an important role for hazards affecting the next generation, especially perhaps for birth defects [45].

Options in Study Design

Cohort Studies

Cohort studies are often regarded as the most valid and informative type of epidemiologic study. Cohorts are defined by a common event for its members. This event is usually of a somewhat complex nature, involving employment during a defined period at a particular industry; exposure to a specific agent may preferably also be required. Employment records or trade union registers are almost always the starting point for defining an occupational cohort. **Cross-sectional studies** of specific exposures in the past or data gathered as a consequence of biological monitoring programs may also define suitable cohorts for follow-up. For example, cohorts can be defined based on surveillance programs for lead in blood or some solvent metabolite, such as trichloroacetic acid or mandelic acid in urine (reflecting exposure to trichloroethylene and styrene, respectively (*see* **Surveillance of Diseases**)).

The analysis of occupational cohorts may be based on either **cumulative incidence** or **incidence density**, and these rates can be compared between the exposed and unexposed in terms of a rate ratio (**relative risk**) or, more rarely, a rate difference. In countries with sound mortality statistics and cancer registries, the observed numbers of specific causes of death or cancer types in a cohort are usually compared to expected numbers as based on the general population rates. These expected numbers are calculated by the "**person-years** method" from the national (or regional) rates, and the relative risk in the exposed cohort, compared to the general (or regional) population, is expressed as the standardized mortality ratio, SMR [33, 58] (*see* **Standardization Methods**). In cancer incidence studies the corresponding measure of effect is usually referred to as SIR, i.e. the standardized incidence ratio.

Occupational cohorts are usually historical or retrospective in character (*see* **Cohort Study, Historical**) but may also include some prospective follow-up. Still, the accuracy of the exposure assessment is usually limited by that in the retrospective phase of the study. A purely prospective cohort study would permit more accurate exposure assessment in principle, but this design is rarely used because it may take decades to complete. Many cohort studies fail to address adequately the changes in the pattern of exposure over time. Inaccuracies in exposure assessment can therefore be severe for those individuals who change jobs and who acquire new exposures, which in combination with the earlier exposures may enhance multistage development of diseases such as cancer.

It may be difficult to trace individuals of an occupational cohort in countries without registries of the living population and of deaths, or because of restrictions in the use of identifying information. In many countries tracing may therefore rely on driving license registries, telephone directories, and writing and calling people with similar family names living in the vicinity of a factory at issue.

A successful follow-up includes 95% or more of the cohort, as is possible in countries with good registries. Emigrants are difficult but not impossible to trace in contrast to "guest workers", who often come from less developed countries without population statistics. If the follow-up requires health

examinations for assessing the health outcome, then a reference cohort usually needs to be established for sound comparisons; the participation rate may drop to 80% or less.

To account for a **latent period** between exposure and disease, especially in cancer studies, new cases and cumulative person-years may be ignored in the analysis for a certain period of time after the start of exposure. Alternatively, the cases and the person-years, along with the observed and expected number of cases, might be analyzed separately according to the time period since first exposure. Further aspects of the analysis of occupational cohort data are discussed below in connection with the healthy worker effect.

Cross-Sectional Studies

A traditional approach in occupational epidemiology has been the cross-sectional study, i.e. to examine an exposed and a nonexposed group at a particular point in time and to compare the **prevalence** of some disease or symptoms in the two groups according to degree of exposure. Cohort or case–control studies of incident disease are preferable, however, because they are not distorted by factors that influence survival or persistence of disease following disease onset. Nevertheless, many medically less serious health problems may be studied by a cross-sectional approach, especially as there is no other realistic possibility regarding, for example, lung or renal dysfunctions, neurobehavioral or neurophysiologic disturbances, or musculoskeletal and other nonlethal disorders. A problem with the cross-sectional design is, however, that the more severely affected workers might have left their jobs, resulting in an underestimate of the true health effects of a particular exposure.

Studies of pregnancy outcome, such as the occurrence of malformations or low **birthweight** may be regarded as cross-sectional. Also, the prevalence of pregnancies that terminate in spontaneous abortions may be compared between women with or without an exposure of interest. Similar to other cross-sectional studies in occupational epidemiology, the exposure information gathered in studies on reproduction may pertain to an entire period, e.g. pregnancy, or even before.

The prevalence **odds ratio** is sometimes used to measure risk in cross-sectional studies, but when the prevalence rate is large, as for abortions or musculoskeletal and other common disorders, the prevalence odds ratio poorly approximates the more intelligible **prevalence ratio**. Hence, when the prevalence rate is 10% in the unexposed and 40% in the exposed, the odds ratio is 6.0, whereas the prevalence ratio is only 4.0. Furthermore, a potentially confounding factor has different effects on the prevalence ratio than on the odds ratio. Thus, the use of **logistic regression** to adjust the odds ratio for confounding is of little utility in cross-sectional studies of common symptoms or disorders. Further details in this regard may be found elsewhere [16, 116, 170].

Cross-sectional studies in occupational health are usually applied also to data involving molecular markers such as DNA or protein adducts to indicate an early effect of an exposure or a sort of subclinical disorder. There are both shorter overviews and extensive conference proceedings on adduct studies with a variety of examples in occupational and environmental health [22, 73, 161]. Some further aspects of the use of molecular biological data are raised in the section "Use of Molecular Epidemiology in Occupational Health" below.

Case–Control (Case–Referent) Studies

Etiologic factors for rare diseases are usually best studied by case–control designs, unless exposures are very unusual (or extremely common). Except when the case–control study is nested in a cohort, the study population is open or dynamic in occupational case–control studies. Together with the time period involved, the **study population** forms the base for a study. An open base can be predetermined by defining the study population in geographic or administrative terms, but, alternatively, the boundaries may be secondarily laid down by the way the cases are recruited. That is, the study base can be either primary or secondary [110].

In a study with a primary base, all cases of the disease (or a representative sample of these cases) in an area are ascertained from cancer registries (or other disease registries when existent) or hospital files. The cases are compared in terms of various exposures with a sample of subjects from the study base, i.e. the **controls**. This approach implies that the general population is the reference for estimating the odds ratio. If the study base is secondary, then one would have to recruit the controls similarly to

the cases, such as by taking patients with other diseases in a hospital to serve as controls for the cases with the disease of interest from that hospital [109] (*see* **Case–Control Study, Hospital-based**). Results obtained from a secondary base tend to be less reliable than those derived from a primary base.

Especially since the 1980s, the case–control study has become widely used in occupational health for studying the effects of exposures that are not confined to any particular industry, as, for example, in studies of cancer risks from pesticide use in farming and forestry. The risk of exposure to a particular industrial process or agent can also be studied by locating the case–control study to a restricted population living in the area where a particular factory is located. Examples include studies of lung cancer risk from exposure to arsenic in copper smelter workers as well as in the general population [18, 127]. As an alternative to a cohort design, a **case–cohort study** may be useful by allowing multiple case–control comparisons against a common control group.

When other disease entities are used as controls, there is the possibility that these may be associated with exposure. In this case, the exposure frequency of the diseased controls does not reflect the exposure frequency in the base population and the risk ratio is **biased** (*see* **Bias in Case–Control Studies**). If a mix of other disorders are intended to be used as the controls, then some disease entities may be associated with the exposure and should therefore be excluded. Should unrelated disorders be misjudged and also excluded, no bias in the estimated rate ratio (odds ratio) would result, as the relation of exposed to nonexposed among the remaining, properly selected, controls is not affected. Appropriate exclusions may easily be misunderstood and lead to skeptical comments, however, unless clear arguments are given for leaving out some disorders from the control series. Since more than one occupational exposure might be of interest as influencing the occurrence of a disease, a refinement might be necessary in the selection of control disorders because some conditions might be related to some but not all exposures under consideration.

Nested Case–Control (Case–Referent) Studies

A nested case–control study is obtained if the cases as well as the sample of controls are drawn from a closed population, that is, within a cohort. The nested case–control study is usually applied to gather information on exposures and confounders not assessable for all cohort members in the main study. For example, the combined effect of an industrial exposure and smoking (or other exposure) might be of interest. Then, if the distribution of smoking is not known for the cohort members, smoking status need be determined only for the cases and for a sample of the base population, that is, a sample of the cohort members; see, for example, the nested case–control study of lymphohematopoietic cancer in a cohort of workers manufacturing styrene-butadiene rubber [140].

Proportional Mortality Studies and Mortality Odds Ratio Studies

As already mentioned, the proportional mortality study has been applied in occupational epidemiology for many years [78] and may be seen as a kind of cross-sectional study at the time of death, even though the deaths considered are not simultaneous. The principle is to calculate the proportion of deaths from a particular disease out of all deaths and calculate the ratio of the proportions of cause-specific deaths for exposed and nonexposed individuals, the proportional mortality ratio (PMR). **Stratifications** and standardizations for age and other factors may be applied. Another possibility is to use national or regional proportions of specific causes of death for comparisons. The proportional mortality study tends to be somewhat insensitive because any excess mortality would not only affect the numerator but also increase the denominator.

Proportional mortality data may also be analyzed by a case–control approach – sometimes more specifically referred to as a mortality odds ratio study. Analogously to a hospital-based case–control study, other deaths than those from the disease of interest are used as controls [8, 111]. Thus, the ratio of odds of the cause of death of interest to the other deaths for the exposed and nonexposed, respectively, namely the mortality odds ratio, can be estimated as the exposure odds ratio for the cases and for the other deaths. Control diseases should be excluded if they are suspected to be related to the exposure as in case–control studies with hospital controls. The aforementioned copper smelter study may illustrate this approach as comparing various types of cancer and cardiovascular deaths against a common control group of deaths unlikely to be related to the exposure [18].

Correlational Studies

With a continuous census over time in an open study base, a comparison could even be made with regard to incidence rates in regions with a more or less concentrated representation of the type of industry and exposure under consideration. Although studies of this type have been presented in the field of occupational epidemiology, this design cannot be recommended because of the lack of information on exposures and confounders at the individual level and because of the dilution with nonexposed individuals. The design, also called an **ecologic study**, is perhaps more useful in environmental epidemiology, for example in studying health effects of air or water pollution.

Character of Exposed Populations and the Healthy Worker Effect

Being able to work usually requires good health, which means that there is some selection regarding who will enter a particular job as well as who is expelled from it. More skilled jobs tend to recruit workers with different lifestyles from workers in less skilled jobs, and health-related departures from the labor force may be concentrated among low socioeconomic groups [47]. A particular group of workers is therefore likely to be healthier than the general population and also tends to differ from other workers. This health-related selection process, called the "healthy worker effect" [107], makes it difficult to find proper comparison groups and explains why various worker groups often enjoy better health outcomes and have smaller risks than expected [6, 44].

One can distinguish between a healthy worker effect in the period shortly after hire and a healthy worker survivor effect operating on a long-term basis. The latter may cause cumulative exposure to become associated with good health among the long-term employees and have a tendency to depress the upper end of an exposure–response curve.

Cohort Studies

In cohort studies, the healthy worker effect is usually evidenced by a total mortality of about 90% or less than expected. A decrease in cardiovascular deaths tends to contribute most to the healthy worker effect, but other causes of death may also be below expected levels. Sometimes the observed number of deaths is as low as only about 50%–60% of the expected, as, for example, in some studies for cardiovascular disease [121], other noncancer deaths [24], as well as cancer [157].

When the healthy worker effect is strong, the comparison with expected numbers based on national or regional rates is questionable, but often there is no alternative reference population. One should be cautious, however, in concluding that there is no risk when national or regional rates are taken as the reference. Even with an appropriate reference, studies showing no effect should be looked upon as essentially uninformative or "nonpositive" rather than "negative" unless there is a large number of cases [4, 74]. If a cohort is large enough, then internal comparisons regarding exposure–response relationships might offer the better comparability, but often there is no unexposed reference group in such studies. Regarding the early period of follow-up, preemployment measures such as health exams create strong selection for a healthy worker effect; a further concern is that cases might have been selectively lost to follow-up – for example, due to sorting out of deceased individuals from company registries.

The healthy worker effect has been relatively weak in many cohort studies from the Nordic countries. Possibly the low unemployment rate that prevailed for a long period of time made it necessary for employers to recruit even people with a marginal health prognosis. In contrast, a more pronounced healthy worker effect may occur in countries and time periods characterized by a high unemployment rate. Usually the healthy worker effect is greater in the younger age groups in a cohort and in the early phase of follow-up [125].

In many studies, the higher risk ratios have appeared among workers with short-time employment rather than among those who have been employed for a long time, and risk may decrease with increasing duration of employment and exposure. The reason for the poor health outcome in short-time employees is usually sought in certain lifestyle characteristics. Part of the explanation may also be that only those workers remaining healthy stay on the job long enough to achieve a higher degree of exposure measured as years of employment or as a product of exposure concentration and time. Allowing for a time

lag following initial employment tends to reduce this healthy worker survivor effect.

Adjustment for length of follow-up and employment status (if associated with the disease, independently of the exposure) may also reduce bias from the healthy survivor effect [55, 123, 151]. Computer programs are available that can provide appropriate person-time data for such adjustments [124]. Arrighi & Hertz-Picciotto [6] compared methods to deal with the healthy worker survivor effect in a study of exposure to arsenic on lung cancer risk. The so-called G method of Robins et al. [136] was thought to be most appropriate, but a lagged analysis worked relatively well except for diseases with a short induction-latency time.

Cross-Sectional Studies

A recent study of symptoms of the respiratory tract, lung function and airway responsiveness in relation to occupational and smoking histories in underground bituminous coal miners and nonmining controls illustrates selection problems in cross-sectional studies [131]. Miners with the longest duration of work at the coal face responded less often to methacholine than miners who had never worked at the coal face, and miners who responded to methacholine were less likely to have worked in dusty jobs than miners not responding. It was concluded that these findings probably resulted from health-related job selection. Similarly, a cross-sectional study of animal feed workers revealed a decreasing prevalence of most chronic respiratory symptoms with increasing years of exposure to dust and endotoxin [144]. Thus, the healthy worker effect may lead to underestimation of risk in cross-sectional studies and even obscure a risk altogether.

Case–Control Studies

The healthy worker effect can also influence the results of case–control studies if exposed individuals tend to be healthier than other members of the population constituting the study base. A more subtle, reversed, and less obvious healthy worker effect can occur in hospital-based case–control studies. If the working population with the exposure of interest is healthier than others in the study base, then the controls would less often be exposed, as many of them come from the unexposed part of the study population with less good health. The result of the healthy worker effect in such case–control studies using hospital (or deceased) controls would therefore be an exaggerated rate ratio (odds ratio). The same reasoning applies to proportional mortality studies. Park et al. [122] have presented parallel analyses showing that mortality odds ratios (MORs) and proportional mortality ratios (PMRs) were higher than standardized mortality ratios (SMRs) for some causes of death; it seems likely that "the truth" might be somewhere between the different estimates obtained. In case–control studies this reverse healthy worker effect is avoided when population controls are enrolled.

A related phenomenon in case–control studies may arise when controls are recruited by **random digit dialing**. Subjects answering the telephone are less likely to be working or to have an exposed job, especially if exposure is associated with a job that demands long working hours. Thus, the exposure frequency in the study base might be underestimated, resulting in an exaggerated estimate of the effect of the exposure.

Assessment of Exposure

Conferences held in the early 1990s reflect the efforts made to improve exposure assessment in occupational epidemiology [15, 72, 77]. Specific knowledge is required for preventive measures to reduce or eliminate hazardous agents or processes from the work environment. The proper assessment of exposure is therefore a key issue in any study of work-related adverse health effects. There are conceptual difficulties in defining exposure and dose, and further problems in accurately measuring or classifying the exposure. Errors in this regard can also affect adjustments for confounding.

Records showing the specific job tasks of the workers are available in many companies and usually form the basis for cohort studies, but can also be used for exposure assessment in case–control and proportional mortality studies. Case–control studies often rely on questionnaire information or interviews, however, and may therefore be subject to **recall bias** or **interviewer bias** (observer bias).

Measures of Exposure

Measures of exposure are usually either exposure intensity, exposure duration, or cumulative exposure. For acute diseases, peak exposure intensity is often

particularly relevant. Cumulative exposure measured as duration or as time-integrated intensity are often used in studies of chronic disease. With ionizing **radiation** as the paradigm, cumulative exposure is commonly taken as a proper determinant of risk for genotoxic and carcinogenic agents, and there are specific arguments in support of such a measure [134]. These arguments are based on an assumption of linear kinetics in the metabolism. However, a literature survey on cancer studies has shown that intensity measures of exposure often yielded larger relative risks than duration of exposure, and intensity measures also often yield monotonically increasing exposure–response curves [26].

Furthermore, in a **pharmacokinetic** study relating cumulative exposure to tissue dose for insoluble, respirable dust particles and toxic metabolites of a nonpolar organic solvent, Smith [146] found no linear relationship between cumulative exposure and tissue dose. It was suggested that this observation could explain why a disproportionally high risk of pulmonary effects is commonly seen for workers with relatively short but intense dust exposures. Specific measures of exposure that result in large apparent risks and clear **dose–response** relationships have been suggested for particular diseases, such as silicosis [40].

Job–Exposure Matrices

Occupations or job tasks are easier to recall and report correctly than exposures to specific agents like metals, solvents, or pesticides. Hoar [79] proposed a **job–exposure matrix** to translate job task histories into estimates of exposure to specific agents. A job–exposure matrix consists of jobs on one axis and specific exposures to substances or other agents on the other, with the matrix elements describing the likelihood of an individual's exposure to a specific substance in a given job, either in **binary** or **polytomous** categories. A matrix may also dichotomize exposure on a probability basis [29].

However, it is necessary to adapt the job–exposure matrix to the country or region and the type of industry where it is to be used. A population-specific job–exposure matrix may therefore be preferable to general job–exposure matrices developed elsewhere. Such a matrix can be constructed from the results of in-depth interviews of a job-stratified sample of cohort members [97].

Ronneberg [138] used a job–exposure matrix in a study of Norwegian aluminum smelter workers. Jobs held by cohort members were identified from personnel records; work tasks and their locations were determined for all jobs, and information was gathered about changes in exposure conditions over time. Then the jobs were combined into categories thought to represent similar exposure conditions, and time-weighted average exposures were estimated on a relative scale.

A specific job–exposure matrix for chlorinated solvents assigned semiquantified estimates of the probability and intensity of exposure to each four-digit job category of the Standard Industrial Classification and Standard Occupational Classification codes in the US [61]. The matrix was also designed to account for the changing patterns of use of these solvents by decade from the 1920s to the 1980s. An algorithm was applied to assign each study subject a unique lifetime probability of exposure and an estimated score of cumulative exposure for each of the solvents. An important goal of the matrix was to reduce the number of **false positive** exposure assessments.

The latter principle is corroborated by a study of astrocytomas and exposure to methylene chloride showing that the odds ratios increased with increasing **specificity** of the exposure assessment [53]. The risk estimate more than tripled compared with the risk estimate obtained without taking probability of exposure and exposure by decades into account and coding for industries and occupations.

There have been many comparisons and evaluations of the validity of the various approaches to assess exposure. Structured questionnaire information is commonly used, but underreporting of exposure remains a problem [90, 128]. On the basis of a large-scale study from Canada [143], Dewar et al. [48] found that the assessment of exposure by an expert team was more efficient than the use of a job–exposure matrix. This may explain why interviews resulted in several increased odds ratios in a study of mental retardation and parental occupation, whereas the use of a job–exposure matrix did not [137].

Men and women with the same job title may have different exposure patterns, indicating a need for gender-specific job–exposure matrices [108]. A comparison of information on exposure to dusts, gases,

and fumes from a job–exposure matrix with questionnaire data indicated a better agreement in men than in women and suggested that men had a more accurate recollection of exposure – especially well-educated men [80]. Smoking habits had no effect on the perception of exposure. For women, the perception of exposure did not vary significantly according to respiratory symptoms. In men, however, subjects without chronic cough or chronic bronchitis even had a significantly higher perception of exposure than the others, but no difference was shown for wheezing, dyspnoea, or asthma.

Stengel et al. [152] compared the performance of experts vs. job–exposure matrices in studies of glomerulonephritis and bladder cancer. Categories of exposure as obtained from both experts and job–exposure matrices were dichotomized, using different cutoff points for exposure and nonexposure. **Sensitivity** of the job–exposure matrices vis-à-vis the experts was low (23%–63%), whereas **specificity** was rather high (87%–98%). Assuming an odds ratio of 3 and an exposure prevalence of 10%, and taking the experts' classification of exposure to be completely correct, the use of a job–exposure matrix led to attenuation of the odds ratio by a factor of 1.5–2.1, and to a loss of power equivalent to a reduction in the number of subjects by a factor of 5–10. On the other hand, the job–exposure matrix performed better than self-reported exposure in discriminating high-risk subgroups in a study of lung cancer and asbestos exposure among construction workers [56].

Job–exposure matrices have also been applied to assess physical exposures such as electric and magnetic fields [21, 91] as well as to study aspects of work organization such as work control, social support, and psychological and physical job demands [92, 65]. Also, a kind of job-exposure matrices have recently been elaborated for assessing exposure to carcinogens in some European countries [93, 94].

Some Other Aspects on Exposure Assessment

Data from biological monitoring programs may be helpful for exposure assessment in cohort studies. For example, when animal studies indicated a cancer risk from trichloroethylene in the late 1970s, existing data from routine monitoring of the metabolite trichloroacetic acid in urine could be used for defining cohorts for follow-up with regard to cancer [20, 156]. Although trichloroacetic acid in urine clearly indicated exposure, the proper measure of exposure (e.g. peak values or simple averages) was not evident.

Hygienists are needed not only for judgments about whether a particular exposure is likely to have occurred, but also to evaluate documents on previously measured exposures. Measurement strategies and methods of sampling and analysis have varied over the years. It is especially important to consider the sampling strategies when hygienists' measurements are used for epidemiologic purposes [158]. The reason is that such measurements usually have been made for control of the work environment after changes in an industrial process for hygienic or technical reasons and therefore tend to underestimate the average daily exposures.

Considerable differences in exposure may occur between workers from the same factory and with the same job titles [96]. Only one-quarter of some worker groups had individual **mean** exposures within a two-fold range for 95% of the individuals. Furthermore, about one-third of the worker groups had a greater than 10-fold range for 95% of the individuals. There were also large day-to-day variations, especially for outdoor workers and when the process was intermittent. Indoor work in a continuous process led to more homogenous exposures. Others have reported similar observations regarding exposure variation [36, 126] and a suggested approach to deal with the problem of retrospective exposure assessment by Bayesian methods has lead to some discussion [37, 133].

Uncertainties are likely to affect any exposure assessment, causing some individuals to be taken as more exposed and others as less exposed than they really are. In principle, one should emphasize the need for a positive **predictive value** rather than sensitivity of a job-exposure matrix or a questionnaire for assessing exposure to agents of interest [132]; the reverse is true for confounding factors, however.

Sometimes the presence of recall bias or observer bias in case–control data can be revealed by comparing the odds ratio for those with and without reported exposure but within job categories with potential exposure. If the latter, who report no exposure, show a decreased risk in comparison to those in clearly unexposed jobs, it is likely that an increased odds ratio for those reported exposed reflects some bias in the assessment of exposure [9].

If there are no systematic influences on the exposure assessment, then the result is **nondifferential error**, which usually leads to risk estimates that are

biased towards the null value (*see* **Misclassification Error**). Nondifferential misclassification makes it difficult to discover adverse health effects by attenuating the risk estimates. However, even in the presence of nondifferential error, chance sometimes may lead to exaggerated risk estimates. Little attention has been paid to this possibility in reporting of study results, but it has been well illustrated by computer **simulations** [148]. Dosemeci et al. [52] showed that discretizing continuous exposure data can lead to biases away from the null if nondifferential error acts on the continuous exposure measurement.

Assessment of exposure in limited time windows may yield particular insights. A time-window approach that is sensitive to recent exposures may enable one to detect a late stage effect for cancer or other diseases. For example, exposure to radon and radon progeny in mines in the 5–15 years before lung cancer seems to have had an important effect on risk [25].

Confounding in Occupational Epidemiology

Determinants of risk that are associated with the exposure under consideration can spuriously increase the apparent risk from this exposure. Such determinants are called "**confounders**". Confounders can also obscure an effect, either when the confounding risk factor tends to be more common in absence of the exposure or when it is protective. In principle, confounding may explain all or part of an association of a disease with an exposure, either because the control of a known confounding factor is incomplete, or because the confounder has not been identified. Mismeasurement or nondifferential misclassification of a confounder can lead to poor control for confounding [3, 132]. However, as long as the exposure under study has a quite strong effect, incomplete control of confounding is not too deleterious for risk estimation [14, 62].

Often the concern about confounding in occupational epidemiology has been focused on lifestyle factors such as smoking and alcohol use or socioeconomic class. Even for a strong risk factor like smoking for lung cancer, the confounding influence is quite modest because smoking tends to be nearly equally prevalent among the occupationally exposed and the unexposed [7, 14].

The most important confounders to consider in occupational epidemiology are other work-related exposures and factors [42]. For example, it is difficult to investigate the role of welding fumes on lung cancer risk because asbestos has often been used in protective equipment in the welding process, and there is also considerable exposure to magnetic fields from electric arc welding, which may or may not be a risk factor. Likewise, in the artistic glass industry, there has been exposure to many different and potentially carcinogenic metals or metallic compounds, but again, asbestos has also been present to protect from the warm glass [86].

Sometimes various exposures are inextricably linked. For example, some phenoxy herbicides have contained impurities of chlorinated dibenzodioxins as a result of the manufacturing process. The association between exposures of this kind is so tight that there is no way to control properly for confounding of one compound to find out the effect of another. Instead, one has to consider the effect of these exposures en bloc [7]. Similarly, occupational job titles might sometimes have to be viewed as blocs of exposures. The best possibilities for prevention occur, however, when specific exposures or processes can be identified as hazardous.

Since a worker may be exposed to a complex array of occupational and other agents of physical or psychosocial character, there is considerable potential for some mutually confounding effects in occupational studies. For this and other reasons, it has become increasingly common to consider many exposures, especially in case–control studies. For example, Blair et al. [27] considered some 150 occupations and about as many industrial categories in a study of lymphoma. When a great number of exposures are analyzed, **false positive** findings may result from the play of chance in the many comparisons. There is also the possibility that confounding from one or more of the exposures associated with increased risk may explain some other positive associations as well. More interest should probably be devoted to this possibility than to the consequences of **multiple comparisons** because exposures for consideration in a study are not randomly selected but are usually included on the basis of some evidence or suspicion of an adverse effect.

Interaction of Exposure Effects

When multiple exposures occur in occupational settings it would be useful to know whether synergistic

or antagonistic **interactions** are present (*see* **Synergy of Exposure Effects**). Not many examples in this respect were found in a literature survey [10], but, for example, combined exposure to vinyl chloride and arsenic increased the risk of respiratory cancer, and the combined exposure to phenoxy herbicides and solvents seemed to increase the risk of lymphoma. This latter interaction has also been recently confirmed by other data in which the combination of phenoxy herbicides and solvent exposure gave an odds ratio of 8.6 vs. 2.6 and 1.4, respectively, for the two exposures alone [130]. Exposure to plastic and rubber chemicals resulted in an odds ratio of 2.2, which increased to 4.8 in the presence of solvents.

As in these examples, the sample sizes available for assessing such occupational interactions are usually small, resulting in considerable uncertainty in the estimates. A background of differently combined exposures may explain inconsistent findings obtained in different studies on the same agent, as can interactions with factors outside the work environment. In the latter respect, the strong synergistic effect of smoking and asbestos exposure on the risk of lung cancer is a classical example [66]. Arsenic and smoking also act synergistically to increase the risk of lung cancer [129]. A synergistic effect of smoking and exposure to radon progeny seems likely as well, although the results differ to some extent between studies [25].

Use of Molecular Epidemiology in Occupational Health

The great achievements in molecular biology over the past decade have also influenced occupational epidemiology [71, 73, 147, 161]. Chemical adducts to deoxyribonucleic acid (DNA) or various proteins like hemoglobin and albumin have been used as either markers of exposure or taken as early adverse health effects. A somewhat later development has involved attempts to evaluate exposure effects in relation to metabolic **polymorphism** and to detect mutations, for example in the p53 **gene**. Specific mutations in the p53 gene in squamous skin cancers have been associated with UV-light exposure [31] and in liver tumors with widespread exposure to aflatoxin B_1 and hepatitis B virus [34, 81], and, by analogy, it is likely that characteristic mutations may result also from occupational exposures to carcinogens.

Studies involving adducts are usually of a cross-sectional design and tend to reflect rather recent exposures due to the turnover of cells and proteins in blood. There is a similarity in this respect to investigations based on chromosomal aberrations, sister chromatid exchanges or micronuclei. These latter type of studies became common in the late 1970s and early 1980s [149]. It has been unclear to what extent chromosomal damage implies any serious effect, but a cohort follow-up indicates that chromosomal aberrations might be predictive of cancer development [64]. The relation to known exposures to carcinogens remains uncertain, however [28] In an evaluation of the cancer risk to humans from styrene exposure, an IARC Working Group took into special account the many studies indicating chromosomal damage [87]. Even so, there now seems to be a decreasing enthusiasm to use chromosomal damage as an outcome measure.

The case–control design is well suited to detecting exposures that can lead to mutations in oncogenes or tumor suppressor genes [147]. The cases are divided into subentities defined by some mutational characteristic, and each such subentity of cases is compared with controls regarding exposure. For example, Taylor et al. [153] compared 62 cases of acute myeloid leukemia with 630 controls. The 10 leukemia patients who were positive for *ras*-mutation were found to have worked more often in high risk occupations. Odds ratios between 1.9 and 7.2 were obtained for the various exposure categorizations made. In contrast, the odds ratios for the *ras*-negative cases ranged from 0.6 to 0.9.

The case–control design is also useful for studying the impact of the genetically determined polymorphism of enzymatic activity and metabolic capacity that determines the susceptibility to risk from an occupational exposure. For example, individuals with one form of the polymorphic CYP1A1 gene appeared to be more susceptible to risks from smoking and occupational exposures such as asbestos than those with other alleles [38, 100]. Although these results have not been confirmed, the applied epidemiologic design used in these studies defines a valid approach for studying metabolic activity and occupational exposures as well as smoking. Efforts to identify individuals at increased occupational risk for bladder cancer because of a glutathione-S-transferase M1 deficiency can serve as another example of this type of study [35].

Glutathione-S-transferase polymorphism may also influence the susceptibility to nonmalignant asbestos-related disease [145].

Occupational exposures may even increase the risk for a clearly inherited disorder. Few individuals with the genetic trait for familial amyloid polyneuropathy develop the disease, suggesting that some other factors might have been involved in the clinically overt cases. In a case–control study of this disorder, solvent exposure appeared as a fairly strong risk factor, with an 11-fold risk for the more heavily exposed [69].

More time is needed to evaluate the role of **molecular epidemiology** for identifying health risks in the workplace. The complexity and the costs involved in these studies will remain a major hindrance for future development, even though some interesting and important results are likely to appear. Identification of genetically determined susceptibility to occupational exposures raises ethical concerns because persons without elevated susceptibility may be selected for employment. Instead, the proper goal should be to create a safe work environment even for those individuals who are more susceptible.

The Etiologic Contribution of Occupational Exposures

The proportion of disease burden attributable to specific exposures or jobs is rather substantial. In Germany, for example, about 250 asbestos-associated lung cancers and 400 mesotheliomas have been recognized and compensated for each year [23]. Lung cancer claims among the underground uranium mine workers in Thuringia and Saxony ranked second to asbestos. Lung cancers related to silicotic scar tissue and to chromium (VI) and arsenic compounds and other chemicals were also subject to compensation.

Estimates of the etiologic contribution of occupational exposures to morbidity or mortality may be obtained by calculating the so-called population **attributable risk** or etiologic fraction. Any particular occupational exposure is quite rare in the general population, however, and can therefore cause only a limited proportion of disease. The overall burden of diseases related to various occupational exposures may nevertheless be considerable. A study of lung cancer in Norway indicated a population attributable risk of 22%–35% for occupations with definitely hazardous exposures [95]. The estimate rose to 37%–47% when jobs with "possibly exposed" categories were also included. Asbestos exposure was the main single risk factor. Attributable risks may add up to more than 100% due to interaction between risk factors. Not surprisingly, therefore, the contribution from smoking could still be estimated to be 82%. Quite similar estimates have more recently been reported also from Sweden and Finland [12, 117].

The quantitative impact of working conditions on cardiovascular diseases in Denmark has been suggested to account for 16% of the premature cardiovascular mortality in men and 22% in women [118]. Including sedentary work as an occupational risk factor, the etiologic fractions rose to 51% and 55% for men and women, respectively. Monotonous high-paced work and shift work were considered the most important single factors, whereas the impact of rather rare chemical exposures to carbon disulfide, nitroglycol, lead, arsenic, carbon monoxide, and other agents was marginal.

Estimates of attributable risk for musculoskeletal and neurologic disorders can be calculated as well. Olsen et al. [120] estimated the population attributable risk for coxarthrosis, a degenerative condition of the hip joint, as 40% for physical workload on the job, 55% for sports, and 15% for excess weight. Overall these three risk factors could account for about 80% of the "idiopathic" coxarthrosis. Landtblom et al. [98] reviewed 10 studies of multiple sclerosis and found relative risks near two for exposure to solvents. Assuming the frequency of relevant solvent exposure to be in the range of 10%–20% in an industrialized country, one would estimate a population attributable risk of about 10% or more. These few examples indicate that the contribution of occupational exposures to cancer as well as other disorders is not negligible.

Concluding Remarks

Identification of risks from occupational exposures and quantification of the associated burden of diseases should lead to prevention efforts. Mandated and voluntary changes in the work environment and proper supervision to ensure compliance with regulations may be as beneficial to health as attempts to change personal habits and lifestyle. Continuous

epidemiologic surveillance is important to obtain information about the long-term impact of preventive measures.

When an epidemiologic study indicates a health risk associated with an industrial process, required changes in the production process may be costly. Workers and the management often hold different views of the balance between the costs and health benefits of preventive measures. It is not unusual, however, that an improvement in the work environment also improves productivity; even so there is usually considerable resistance from an industry to accept epidemiologic evidence of a risk and to improve the work environment.

The method of presenting epidemiologic results is critical to a successful prevention strategy when serious health effects are indicated, such as excess cancer deaths or malformations. When the study pertains to a particular plant or company, it is advisable first to inform the management as well as the employees or their representatives. Mass media may be interested as well, but untimely information through the media can create controversy and hostility towards occupational health research. Press conferences in the presence of management and worker representatives can be useful for limiting negative publicity for a company, because, eventually, the mass media will get access to the information when a scientific report is published. It is therefore advisable to provide information in a more controlled manner [11].

In view of the technical difficulties of conducting epidemiologic studies and a natural reluctance to accept that an industrial process may be harmful, it is not surprising that interpretations of data may differ and serious controversies arise. For example, in 1966 Hueper [82] reconsidered his early warnings, in view of the European experiences, of lung cancer risk from exposure to radon progeny, and accused government officials of having impeded studies of this health hazard among uranium miners in the US. By 1966, Hueper's suspicions had been confirmed by the first report on an excess of lung cancer among these miners [164]. As late as 1971, B. MacMahon wrote in the preface to a comprehensive report on lung cancer in uranium miners [101]:

> The epidemic now in progress among American uranium miners could readily have been – and indeed was – predicted on the basis of past experience in other parts of the world. Less predictable was the extent of the scientific, legal and political controversy that the American experience would engender. Although ... few medical experiences have been so carefully documented, diametrically opposite opinions are still held and expressed not only regarding the interpretation of the facts that have emerged but as to the nature of the facts themselves.

When epidemiologic study results are weak or inconsistent it is indeed difficult to come up with a tenable judgment on the health risk involved. Some subjectivity is unavoidable in such situations, but decision makers may get some guidance from ethical considerations. Hence, it seems reasonable to give the benefit of the doubt to those suffering the risk [4], and in balancing benefits against risk one has to be clear about who takes the risk and who has the benefit. In occupational health, the situation is more complicated than in medical treatment, where the risk of adverse side-effects might be weighed against benefits for the same individual [162]. A comprehensive discussion of the ethical guidelines in occupational health can be found elsewhere [139].

References

[1] Acheson, E.D., Cowdell, R.H. & Jolles, B. (1970). Nasal cancer in the Northamptonshire boot and shoe industry, *British Medical Journal* **1**, 385–393.

[2] Acheson, E.D., Hadfield, E.H. & Macbeth, R.G. (1967). Carcinoma of the nasal cavity and accessory sinuses in woodworkers, *Lancet* **i**, 311–312.

[3] Ahlbom, A. & Steineck, G. (1992). Aspects of misclassification of confounding factors, *American Journal of Industrial Medicine* **21**, 107–112.

[4] Ahlbom, A., Axelson, O., Stöttrup Hansen, E., Hogstedt, C., Jensen, U. & Olsen J. (1990). Interpretation of "negative" studies in occupational epidemiology, *Scandinavian Journal of Work, Environment and Health* **16**, 153–157.

[5] Anttila, A., Pukkala, E., Sallmen, M., Hernberg, S. & Hemminki, K. (1995). Cancer incidence among Finnish workers exposed to halogenated hydrocarbons, *Journal of Occupational Medicine* **37**, 797–806.

[6] Arrighi, H.M. & Hertz-Picciotto, I. (1993). The evolving concept of the healthy worker survivor effect, *Epidemiology* **5**, 189–196.

[7] Axelson, O. (1978). Aspects on confounding in occupational health epidemiology, *Scandinavian Journal of Work, Environment and Health* **4**, 85–89.

[8] Axelson, O. (1979). The case-referent (case control) study in occupational health epidemiology, *Scandinavian Journal of Work, Environment and Health* **5**, 91–99.

[9] Axelson, O. (1980). A note on observational bias in case-referent studies in occupational health epidemiology, *Scandinavian Journal of Work, Environment and Health* **6**, 80–82.

[10] Axelson, O. (1991). Cancer and combined exposures to occupational and environmental factors, *Recent Results in Cancer Research* **122**, 60–70.

[11] Axelson, O. (1994). Dynamics of management and labor in dealing with occupational risks, in *The Identification and Control of Environmental and Occupational Disease*, M.A. Mehlman, ed. Princeton Scientific Publishing, Princeton.

[12] Axelson, O. (2002). Alternative for estimating the burden of lung cancer from occupational exposures-some calculations based on data from Swedish men, *Scand J Work Environ Health* **28**, 58–63.

[13] Axelson, O. & Söderkvist, P. (1991). Characteristics of disease and some exposure considerations, *Applied Occupational and Environmental Hygiene* **6**, 428–435.

[14] Axelson, O. & Steenland, K. (1988). Indirect methods of assessing the effect of tobacco use in occupational studies, *American Journal of Industrial Medicine* **13**, 105–118.

[15] Axelson, O. & Westberg, H. (1992). Introductory note to the concepts of exposure and dose in occupational epidemiology, *American Journal of Industrial Medicine* (Special issue) **21**, 3–4.

[16] Axelson, O., Fredriksson, M. & Ekberg, K. (1994). Use of prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies, *Occupational and Environmental Medicine* **51**, 574.

[17] Axelson, O., Hane, M. & Hogstedt, C. (1976). A case-referent study on neuropsychiatric disorders among workers exposed to solvents, *Scandinavian Journal of Work, Environment and Health* **2**, 14–20.

[18] Axelson, O., Dahlgren, E., Jansson, C.-D. & Rehnlund, S.O. (1978). Arsenic exposure and mortality, a case referent study from a Swedish copper smelter, *British Journal of Industrial Medicine* **35**, 8–15.

[19] Axelson, O., Seldén, A., Andersson, K. & Hogstedt, C. (1994). Updated and expanded Swedish cohort study on trichloroethylene and cancer risk, *Journal of Occupational Medicine* **36**, 556–562.

[20] Axelson, O., Andersson, K., Hogstedt, C., Holmberg, B., Molina, G. & de Verdier, A. (1978). A cohort study on trichloroethylene exposure and cancer mortality, *Journal of Occupational Medicine* **20**, 194–196.

[21] Baris, D., Armstrong, B.G., Deadman, J. & Theriault, G. (1996). A case cohort study of suicide in relation to exposure to electric and magnetic fields among electrical utility workers, *Occupational and Environmental Medicine* **53**, 17–24.

[22] Bartsch, H., Kadlubar, F. & O'Neill, I., eds. (1993). Biomarkers in human cancer – Part II. Exposure monitoring and molecular dosimetry, *Environmental Health Perspectives* **99**, 2–309.

[23] Baur, X., Marczynski, B., Rozynek, P. & Voss, B. (1994). Bronchopulmonale Prakanzerosen und Tumoren – Risikogruppen aus arbeitsmedizinischer Sicht (Bronchopulmonary precancerous conditions and tumors – risk groups from the occupational medicine viewpoint), *Pneumologie* **48**, 825–834.

[24] Beall, C., Delzell, E. & Macaluso, M. (1995). Mortality patterns among women in the motor vehicle manufacturing industry, *American Journal of Industrial Medicine* **28**, 325–337.

[25] BEIR IV. Committee of Biological Effects of Ionizing Radiations, US National Research Council. (1988). *Health Risk of Radon and other Internally Deposited Alpha-Emitters*. National Academy Press, Washington.

[26] Blair, A. & Stewart, P.A. (1992). Do quantitative exposure assessments improve risk estimates in occupational studies of cancer?, *American Journal of Industrial Medicine* **21**, 53–63.

[27] Blair, A., Linos, A., Stewart, P.A., Burmeister, L.F., Gibson, R., Everett, G., Schuman, L. & Cantor, K.P. (1993). Evaluation of risks for non-Hodgkin's lymphoma by occupation and industry exposures from a case-control study, *American Journal of Industrial Medicine* **23**, 301–312.

[28] Bonassi, S., Hagmar, L., Stromberg U., Montagud, A.H., Tinnerberg, H., Forni, A., Heikkila, P., Wanders, S., Wilhardt, P., Hansteen, I.L., Knudsen, L.E. & Norppa, H. (2000). Chromosomal aberrations in lymphocytes predict human cancer independently of exposure to carcinogens. European Study Group on Cytogenetic Biomarkers and Health *Cancer Research* **60**, 1619–1625.

[29] Bouyer, J. & Hemon, D. (1993). Comparison of three methods of estimating odds ratios from a job exposure matrix in occupational case-control studies, *American Journal of Epidemiology* **137**, 472–481.

[30] Brackbill, R.M., Maizlish, N. & Fischbach, T. (1990). Risk of neuropsychiatric disability among painters in the United States, *Scandinavian Journal of Work, Environment and Health* **16**, 182–188.

[31] Brash, D.E., Rudolph, J.A., Simon, J.A., Lin, A., McKenna, G.J., Baden, H.P., Halperin, A.J. & Ponten, J. (1991). A role of sunlight in skin cancer, UV-induced p53 mutations in squamous cell carcinoma, *Proceedings of the National Academy of Sciences* **88**, 10124–10128.

[32] Breslow, L. (1955). Industrial aspects of bronchogenic neoplasms, *Diseases of the Chest* **28**, 421–430.

[33] Breslow, N.E. (1984). Elementary methods of cohort analysis, *International Journal of Epidemiology* **13**, 112–115.

[34] Bressac, B., Kew, M., Wands, J. & Ozturk, M. (1991). Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa, *Nature* **350**, 429–431.

[35] Brockmoller, J., Kerb, R., Drakoulis, N., Staffeldt, B. & Roots, I. (1994). Glutathione S-transferase M1 and its variants A and B as host factors of bladder cancer

susceptibility, a case-control study, *Cancer Research* **54**, 4103–4111.

[36] Burstyn, I. & Kromhout H. (2000). Are the members of a paving crew uniformly exposed to bitumen fume, organic vapor, and benzo(a)pyrene? *Risk Analysis* **20**, 653–663.

[37] Burstyn, I. & Kromhout, H. (2002). Critique of bayesian methods for retrospective exposure assessment, *The Annals of Occupational Hygiene* **46**, 429–431.

[38] Caporaso, N., Hayes, R.B., Dosemeci, M., Hoover, R., Ayesh, R., Hetzel, M. & Idle, J. (1989). Lung cancer risk, occupational exposure, and the debrisoquine metabolic phenotype, *Cancer Research* **49**, 3675–3679.

[39] Case, R.A.M. & Hosker, M.E. (1954). Tumour on the urinary bladder as an occupational disease in the rubber industry in England and Wales, *British Journal of Preventive and Social Medicine* **8**, 39–50.

[40] Checkoway, H. & Rice, C.H. (1992). Time-weighted averages, peaks, and other indices of exposure in occupational epidemiology, *American Journal of Industrial Medicine* **21**, 25–33.

[41] Checkoway, H.A., Pearce, N.E. & Crawford-Brown, D.J. (1989). *Research Methods in Occupational Epidemiology*. Oxford University Press, New York.

[42] Checkoway, H., Savitz, D.A. & Heyer, N.J. (1991). Assessing the effects of nondifferential missclassification of exposures in occupational studies, *Applied Occupational and Environmental Hygiene* **6**, 528–533.

[43] Cherry, N.M., Labréche, F.P. & McDonald, J.C. (1992). Organic brain damage and occupational solvent exposure, *British Journal of Industrial Medicine* **49**, 776–781.

[44] Chia, S.E. & Shi L.M. (2002). Review of recent epidemiological studies on paternal occupations and birth defects, *Occupational and Environmental Medicine* **59**, 149–155.

[45] Choi, B.C. (1992). Definition, sources, magnitude, effect modifiers, and strategies of reduction of the healthy worker effect, *Journal of Occupational Medicine* **34**, 979–988.

[46] Chreech, J.L. & Johnson, M.N. (1974). Angiosarcoma of liver in the manufacture of polyvinyl chloride, *Journal of Occupational Medicine* **16**, 150–151.

[47] Dahl, E. (1993). Social inequality in health – the role of the healthy worker effect, *Social Science and Medicine* **36**, 1077–1086.

[48] Dewar, R., Siemiatycki, J. & Gerin, M. (1991). Loss of statistical power associated with the use of a job-exposure matrix in occupational case-control studies, *Applied Occupational and Environmental Hygiene* **6**, 508–515.

[49] Doll, R. (1952). The causes of death among gasworkers with special reference to cancer of the lung, *British Journal of Industrial Medicine* **9**, 180–185.

[50] Doll, R. (1955). Mortality from lung cancer in asbestos workers, *British Journal of Industrial Medicine* **12**, 81–86.

[51] Doll, R. (1988). Effects of exposure to vinyl chloride. An assessment of the evidence, *Scandinavian Journal of Work, Environment and Health* **14**, 61–78.

[52] Dosemeci, M., Wacholder, S. & Lubin, J. (1990). Does non-differential missclassification of exposure always bias a true effect toward the null value?, *American Journal of Epidemiology* **132**, 746–748.

[53] Dosemeci, M., Cocco, P., Gomez, M., Stewart, P.A. & Heineman, E.F. (1994). Effects of three features of a job-exposure matrix on risk estimates, *Epidemiology* **5**, 124–127.

[54] Figueroa, W.G., Raszkowski, R. & Weiss, W. (1973). Lung cancer in chloromethyl ether workers, *New England Journal of Medicine* **288**, 1096–1097.

[55] Flanders, W.D., Cardenas, V.M. & Austin, H. (1993). Confounding by time since hire in internal comparisons of cumulative exposure in occupational cohort studies, *Epidemiology* **4**, 336–341.

[56] Fletcher, A.C., Engholm, G. & Englund, A. (1993). The risk of lung cancer from asbestos among Swedish construction workers, self-reported exposure and a job exposure matrix compared, *International Journal of Epidemiology* **22**, Supplement 2, S29–S35.

[57] Fu, H. & Bofetta, P. (1995). Cancer and occupational exposure to inorganic lead compounds, a meta-analysis of published data, *Occupational and Environmental Medicine* **52**, 73–81.

[58] Gardner, M.J. (1986). Considerations in the choice of expected numbers for appropriate comparisons in occupational cohort studies, *Medicina del Lavoro* **77**, 23–47.

[59] Gardner, M.J., Snee, M.P., Hall, A.J., Powell, C.A., Downes, S. & Terrell, J.D. (1990). Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria, *British Medical Journal* **300**, 423–429.

[60] Gloyne, S.R. (1935). Two cases of squamous carcinoma of the lung occurring in asbestosis, *Tubercle* **17**, 5–10.

[61] Gomez, M.R., Cocco, P., Dosemeci, M. & Stewart, P.A. (1994). Occupational exposure to chlorinated aliphatic hydrocarbons, job exposure matrix, *American Journal of Industrial Medicine* **26**, 171–183.

[62] Goodman, M., Kelsh, M., Ebi, K., Iannuzzi, J. & Langholz, B. (2002). Evaluation of potential confounders in planning a study of occupational magnetic field exposure and female breast cancer, *Epidemiology* **13**, 50–58.

[63] Greenland, S., ed. (1987). *Evolution of Epidemiologic Ideas. Annotated Readings on Concepts and Methods*. Epidemiology Resources Inc., Chestnut Hill.

[64] Hagmar, L., Brögger, A., Hansteen, I.L., Heim, S., Högstedt, B., Knudsen, L., Lambert, B., Linnainmaa, K., Mitelman, F. & Nordenson, I. (1994). Cancer risk in humans predicted by increased levels of chromosomal aberrations in lymphocytes. *Nordic study group on the health risk of chromosome damage, Cancer Research* **54**, 2919–2922.

[65] Hammar, N., Alfredsson, L., Johnson, JV. (1998). Job strain, social support at work, and incidence of myocardial infarction. *Occup Environ Med* **55**, 548–553.

[66] Hammond, E.C., Selikoff, I.J. & Seidman, H. (1979). Asbestos exposure, cigarette smoking and death rates, *Annals of the New York Academy of Sciences* **330**, 473–490.

[67] Hardell, L. (1977). Soft-tissue sarcomas and exposure to phenoxyacetic acids – a clinical observation, *Läkartidningen* **74**, 2753–2754 (in Swedish).

[68] Hardell, L., Eriksson, M., Axelson, O. & Hoar Zahm, S. (1994). Cancer epidemiology, in *Dioxins and Health*, A. Schecter, ed., Plenum Press, New York, Chapter 16 pp. 525–547.

[69] Hardell, L., Holmgren, G., Steen, L., Fredrikson, M. & Axelson, O. (1995). Occupational and other risk factors for clinically overt familial amyloid polyneuropathy, *Epidemiology* **6**, 598–601.

[70] Härting, F.H. & Hesse, W. (1879). Der Lungenkrebs, die Bergkrankheit in den Schneeberger Gruben, *Vierteljahrsschrift für Gerichtliche Medicin und Öffentliches Gesundheitswesen* **30**, 296–307; **31**, 102–132; **31**, 313–337.

[71] Hayes, R.B. (1985). Genetic susceptibility and occupational cancer, *Medicina del Lavoro* **86**, 206–213.

[72] Heederik, D., Boleij, J.S.M., Kromhout, H. & Smid, T. (1991). Use and analysis of exposure monitoring data in occupational epidemiology; an example of an epidemiological study in the Dutch animal food industry, *Applied Occupational and Environmental Hygiene* **6**, 458–464.

[73] Hemminki, K. (1992). Use of molecular biology techniques in cancer epidemiology, *Scandinavian Journal of Work, Environment and Health* **18**, Supplement 1, 38–45.

[74] Hernberg, S. (1981). "Negative" results in cohort studies. How to recognize fallacies, *Scandinavian Journal of Work, Environment and Health* **7**, Supplement 4, 121–126.

[75] Hernberg, S. (1992). *Introduction to Occupational Epidemiology*. Lewis, Chelsea.

[76] Hernberg, S., Partanen, T., Nordman, C.H. & Sumari, P. (1970). Coronary heart disease among workers exposed to carbon disulphide, *British Journal of Industrial Medicine* **27**, 313–325.

[77] Herrick, R.F. & Stewart, P.A. (1991). International workshop on retrospective exposure assessment for occupational epidemiologic studies. Preface, *Applied Occupational and Environmental Hygiene* **6**, 417–420.

[78] Hill, A.B. & Faning, E.L. (1948). Studies in the incidence of cancer in a factory handling inorganic compounds of arsenic. I. Mortality experience in the factory, *British Journal of Industrial Medicine* **5**, 1–6.

[79] Hoar, S.K. (1982). Job-exposure matrices in occupational epidemiology, *Journal of the National Cancer Institute* **69**, 1419–1420.

[80] Hsairi, M., Kauffmann, F., Chavance, M. & Brochard, P. (1992). Personal factors related to the perception of occupational exposure, an application of a job exposure matrix, *International Journal of Epidemiology* **21**, 972–980.

[81] Hsu, I.C., Metcalf, R.A., Sun, T., Welsh, J.A., Wang, N.J. & Harris, C.C. (1991). Mutational hotspot in the p53 gene in human hepatocellular carcinomas, *Nature* **350**, 427–428.

[82] Hueper, W.C. (1966). *Occupational and Environmental Cancers of the Respiratory System*. Springer-Verlag, Berlin.

[83] Huff, J.E., Salmon, A.G., Hooper, N.K. & Zeise, L. (1991). Long-term carcinogenesis studies on 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin and hexachlorodibenzo-p-dioxins, *Cell Biology and Toxicology* **7**, 67–94.

[84] IARC (1974). *Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man. Some Aromatic Amines, Hydrazine and Related Substances, N-Nitroso Compounds and Miscellaneous Alkylating Agents*, Vol. 4. International Agency for Research on Cancer, Lyon.

[85] IARC (1987). *Monographs on the Evaluation of Carcinogenic Risk to Humans. Overall Evaluations of Carcinogenicity. An Updating of IARC Monographs*, Vols 1–42, Suppl. 7. International Agency for Research on Cancer, Lyon.

[86] IARC (1993). *Monographs on the Evaluation of Carcinogenic Risks to Humans*. Vol. 58, Beryllium, Cadmium, Mercury, and Exposures in the Glass Manufacturing Industry. International Agency for Research on Cancer, Lyon.

[87] IARC (1994). *Monographs on the Evaluation of Carcinogenic Risks to Humans*. Vol. 60, Some Industrial Chemicals. International Agency for Research on Cancer, Lyon.

[88] IARC (1995). *Monographs on the Evaluation of Carcinogenic Risks to Humans*. Vol. 63, Dry Cleaning, Some Chlorinated Solvents and Other Industrial Chemicals. International Agency for Research on Cancer, Lyon.

[89] Jauchem, J.R. & Merritt, J.H. (1991). The epidemiology of exposure to electromagnetic fields; an overview of the recent literature, *Journal of Clinical Epidemiology* **44**, 895–906.

[90] Joffe M. (1992). Validity of exposure data derived from a structured questionnaire, *American Journal of Epidemiology* **135**, 564–570.

[91] Johansen, C., Raaschou-Nielsen, O., Skotte, J., Thomsen, B.L., Olsen, J.H. (2002). Validation of a job exposure matrix for assessment of utility worker exposure to magnetic fields, *Applied Occupational and Environmental Hygiene* **17**, 304–310.

[92] Johnson, J.V. & Stewart, W.F. (1993). Measuring work organization exposure over the life course with a job-exposure matrix, *Scandinavian Journal of Work, Environment and Health* **19**, 21–28.

[93] Kauppinen, T., Pajarskiene, B., Pondniece, Z., Rjazanov, V., Smerhovsky, Z., Veidebaum, T. & Leino, T.

(2001). Occupational exposure to carcinogens in Estonia, Latvia, Lithuania and the Czech Republic in 1997, *Scand J Work Environ Health* **27**, 343–345.

[94] Kauppinen, T., Toikkanen, J., Pedersen, D., Young, R., Arhens, W., Boffetta, P., Hansen, J., Kromhout, H., Maqueda Blasco, J., Mirabelli, D., de la Orden-Rivera, V., Pannett, B., Plato, N., Savela, A., Vincent, R. & Kogevinas, M. (2000). Occupational exposure to carcinogens in the European Union, *Occup Environ Med* **57**, 10–18.

[95] Kjuus, H., Langård, S. & Skjaerven, R. (1986). A case-referent study of lung cancer, occupational exposures and smoking. III. Etiologic fraction of occupational exposures, *Scandinavian Journal of Work, Environment and Health* **12**, 210–215.

[96] Kromhout, H., Symanski, E. & Rappaport, S.M. (1993). A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents, *Annals of Occupational Hygiene* **37**, 253–270.

[97] Kromhout, H., Heederik, D., Dalderup, L.M. & Kromhout, D. (1992). Performance of two general job-exposure matrices in a study of lung cancer morbidity in the Zutphen cohort, *American Journal of Epidemiology* **136**, 698–711.

[98] Landtblom, A.-M., Flodin, U., Söderfeldt, B., Wolfson, C. & Axelson, O. (1996). Organic solvents and multiple sclerosis. A synthesis of the current evidence, *Epidemiology* **7**, 429–433.

[99] Lilienfeld, A.M. & Lilienfeld, D.E. (1979). A century of case-control studies, progress?, *Journal of Chronic Diseases* **32**, 5–13.

[100] London, S.J., Daly, A.K., Fairbrother, K.S., Holmes, C., Carpenter, C.L., Navidi, W.C. & Idle, J.R. (1995). Lung cancer risk in African-Americans in relation to a race-specific CYP1A1 polymorphism, *Cancer Research* **55**, 6035–6037.

[101] Lundin, F.E., Wagoner, J.K. & Archer, V.E. (1971). *Radon Daughter Exposure and Respiratory Cancer; Quantitative and Temporal Aspects*. NIOSH and NIEHS Joint Monograph No. 1. Department of Health, Education and Welfare, Public Health Service, Washington.

[102] Lynch, K.M. & Smith, W.A. (1935). Pulmonary asbestosis III. Carcinoma of the lung in asbestos-silicosis, *American Journal of Cancer* **24**, 56–64.

[103] Lynge, E., Kurppa, K., Kristofersen, L., Malker, H. & Sauli, H. (1986). Silica dust and lung cancer, results from the Nordic occupational mortality and cancer incidence registers, *Journal of the National Cancer Institute* **77**, 883–889.

[104] Macbeth, R. (1965). Malignant disease of the paranasal sinuses, *Journal of Laryngology* **79**, 592–612.

[105] McDonald, A. (1995). Work and pregnancy, in *Epidemiology of Work Related Diseases*, J.C. McDonald, ed. British Medical Journal Publishing Group, London.

[106] McDonald, J.C. ed. (1995). *Epidemiology of Work Related Diseases*. British Medical Journal Publishing Group, London.

[107] McMichael, A.J. (1976). Standardized mortality ratios and the "healthy worker effect", scratching beneath the surface, *Journal of Occupational Medicine* **18**, 165–168.

[108] Messing, K., Dumais, L., Courville, J., Seifert, A.M. & Boucher, M. (1994). Evaluation of exposure data from men and women with the same job title, *Journal of Occupational Medicine* **36**, 913–917.

[109] Miettinen, O.S. (1985). The "case-control" study: valid selection of subjects (with dissents, comment and response), *Journal of Chronic Diseases* **38**, 543–558.

[110] Miettinen, O.S. (1985). *Theoretical Epidemiology, Principles of Occurrence Research in Medicine*. Wiley, New York.

[111] Miettinen, O.S. & Wang, J.-D. (1981). An alternative to the proportionate mortality ratio, *American Journal of Epidemiology* **114**, 144–148.

[112] Mikkelsen, S. (1980). A cohort study of disability pension and death among painters with special regard to disabling presenile dementia as an occupational disease, *Scandinavian Journal of Social Medicine* **16**, Supplement, 34–43.

[113] Milham, S. (1982). Mortality from leukaemia in workers exposed to electrical and magnetic fields, *New England Journal of Medicine* **307**, 249.

[114] Milham, S., Jr. (1983). *Occupational Mortality in Washington State 1950–1979*. DHHS Pub. No. (NIOSH) 83–116. Centers for Disease Control and Prevention, Cincinnati, pp. 1663–1664.

[115] Monson, R.R. (1990). *Occupational Epidemiology*. CRC Press, Boca Raton.

[116] Nurminen, M. (1995). To use or not to use the odds ratio in epidemiologic analyses?, *European Journal of Epidemiology* **11**, 365–371.

[117] Nurminen, M., Karjalainen, A. (2001). Epidemiologic estimate of the proportion of fatalities related to occupational factors in Finland, *Scand J Work Environ Health* **27**, 161–213.

[118] Olsen, O. & Kristensen, T.S. (1991). Impact of work environment on cardiovascular diseases in Denmark, *Journal of Epidemiology and Community Health* **45**, 4–10.

[119] Olsen, J. & Sabroe, S. (1980). A case-reference study of neuropsychiatric disorders among workers exposed to solvents in the Danish wood and furniture industry, *Scandinavian Journal of Social Medicine* **16**, Supplement, 44–49.

[120] Olsen, O., Vingård, E., Köster, M. & Alfredsson, L. (1994). Etiologic fractions for physical work load, sports and overweight in the occurrence of coxarthrosis, *Scandinavian Journal of Work, Environment and Health* **20**, 184–188.

[121] Ott, M.G., Skory, L.K., Holder, B.B., Bronson, J.M. & Williams, P.R. (1983). Health evaluation of employees

occupationally exposed to methylene chloride. Mortality, *Scandinavian Journal of Work, Environment and Health* **9**, Supplement 1, 8–16.

[122] Park, R., Krebs, J. & Mirer, F. (1994). Mortality at an automotive stamping and assembly complex, *American Journal of Industrial Medicine* **26**, 449–463.

[123] Pearce, N. (1992). Methodological problems in time-related variables in occupational cohort studies, *Revue d'Épidemiologie et de Santé Publique* **40**, S43–S54.

[124] Pearce, N. & Checkoway, H. (1987). Epidemiologic programs for computers and calculators. A simple computer program for generating person-time data in cohort studies involving time-related factors, *American Journal of Epidemiology* **125**, 1085–1091.

[125] Pearce, N.E., Checkoway, H. & Shy, C.M. (1986). Time-related factors as potential confounders and effect modifiers in studies on an occupational cohort, *Scandinavian Journal of Work, Environment and Health* **112**, 97–107.

[126] Peretz, C., Goren, A., Smid, T. & Kromhout, H. (2002). Application of mixed-effects models for exposure assessment, *The Annals of Occupational Hygiene* **46**, 69–77.

[127] Pershagen, G. (1985). Lung cancer mortality among men living near an arsenic-emitting smelter, *American Journal of Epidemiology* **122**, 684–694.

[128] Pershagen, G. & Axelson, O. (1982). A validation of questionnaire information on occupational exposure and smoking, *Scandinavian Journal of Work, Environment and Health* **8**, 24–28.

[129] Pershagen, G., Wall, S., Taube, A. & Linnman, L. (1981). On the interaction between occupational arsenic exposure and smoking and its relationship to lung cancer, *Scandinavian Journal of Work, Environment and Health* **7**, 302–309.

[130] Persson, B. & Fredrikson, M. (1995). A pooled analysis of non-Hodgkin's lymphoma and the role of rare occupational exposures and interaction of risk factors, in *Occupational Exposures and Malignant Lymphoma*, B. Persson, ed. Linköping University Medical Dissertations No. 475. Faculty of Health Sciences, Linköping University, Linköping.

[131] Petsonk, E.L., Daniloff, E.M., Mannino, D.M., Wang, M.L., Short, S.R. & Wagner, G.R. (1995). Airway responsiveness and job selection; a study in coal miners and non-mining controls, *Occupational and Environmental Medicine* **52**, 745–749.

[132] Plato, N. & Steineck, G. (1993). Methodology and utility of a job-exposure matrix, *American Journal of Industrial Medicine* **23**, 491–502.

[133] Ramachandran, G. (2001). Retrospective exposure assessment using Bayesian methods, *The Annals of Occupational Hygiene* **45**, 651–667.

[134] Rappaport, S.M. (1993). Biological considerations in assessing exposures to genotoxic and carcinogenic agents, *International Archives of Occupational and Environmental Health* **65**, S29–S35.

[135] Rehn, L. (1906). Blasenkrankungn bei Anilinarbeitern, *Verhandlungen der Deutschen Geselllschaft für Chirurgie* **35**, 313–318.

[136] Robins, J.M., Blevins, D., Ritter, G. & Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients, *Epidemiology* **3**, 319–336.

[137] Roeleveld, N., Zielhuis, G.A. & Gabreëls, F. (1993). Mental retardation and parental occupation, a study on the applicability of job exposure matrices, *British Journal of Industrial Medicine* **50**, 945–954.

[138] Ronneberg, A. (1995). Mortality and cancer morbidity in workers from an aluminium smelter with prebaked carbon anodes. Part I, Exposure assessment, *Occupational and Environmental Medicine* **52**, 242–249.

[139] Samuels, S.W., ed. (1986). The environment of the work place and human values, *American Journal of Industrial Medicine* **9**, 1–113.

[140] Santos-Burgoa, C., Matanoski, G.M., Seger, S. & Schwartz, L. (1992). Lymphohematopoietic cancer in styrene-butadiene polymerization workers, *American Journal of Epidemiology* **136**, 843–854.

[141] Savitz, D.A., Pearce, N.E. & Poole, C. (1989). Methodological issues in the epidemiology of electromagnetic fields and cancer, *Epidemiological Reviews* **11**, 59–78.

[142] Selikoff, I.J., Churg, J. & Hammond, E.C. (1964). Asbestos exposure and neoplasia, *Journal of the American Medical Association* **188**, 22–26.

[143] Siemiatycki, J., ed. (1991). *Risk Factors for Cancer in the Workplace*. CRC Press, Boca Raton.

[144] Smid, T., Heederik, D., Houba, R. & Quanjer, P.H. (1992). Dust- and endotoxin-related respiratory effects in the animal feed industry, *American Review of Respiratory Disease* **146**, 1474–1479.

[145] Smith, C.M., Kelsey, K.T., Wiencke, J.K., Leyden, K., Levin, S. & Christiani, D.C. (1994). Inherited glutathione-S-transferase deficiency is a risk factor for pulmonary asbestosis, *Cancer Epidemiology, Biomarkers and Prevention* **3**, 471–477.

[146] Smith, T.J. (1992). Occupational exposure and dose over time, limitations of cumulative exposure, *American Journal of Industrial Medicine* **21**, 35–51.

[147] Söderkvist, P. & Axelson, O. (1995). On the use of molecular biology data in occupational and environmental epidemiology, *Journal of Occupational and Environmental Medicine* **37**, 84–90.

[148] Sohrahan, T. & Gilthorpe, M.S. (1994). Non-differential missclassification of exposure always leads to an underestimate of risk, an incorrect conclusion, *Occupational and Environmental Medicine* **51**, 839–840.

[149] Sorsa, M. (1980). Cytogenetic methods in the detection of chemical carcinogens, *Journal of Toxicology and Environmental Health* **6**, 1077–1080.

[150] Spirtas, R., Stewart, P.A., Lee, J.S., Marano, D.E., Forbes, C.D., Grauman, D.J., Pettigrew, H.M., Blair, A., Hoover, R.N. & Cohen, J.L. (1991). Retrospective cohort mortality study of workers at an aircraft

maintenance facility. I. Epidemiological results, *British Journal of Industrial Medicine* **48**, 515–530.

[151] Steenland, K. & Stayner, L. (1991). The importance of employment status in occupational cohort mortality studies, *Epidemiology* **2**, 418–423.

[152] Stengel, B., Pisani, P., Limasset, J.C., Bouyer, J., Berrino, F. & Hemon, D. (1993). Retrospective evaluation of occupational exposure to organic solvents, questionnaire and job exposure matrix, *International Journal of Epidemiology* **22**, Supplement 2, S72–S82.

[153] Taylor, J.A., Sandler, D.P., Bloomfield, C.D., Shore, D.L., Ball, E.D., Neubauer, A., McIntyre, O.R. & Liu, E. (1992). ras Oncogene activation and occupational exposures in acute myeloid leukemia, *Journal of the National Cancer Institute* **84**, 1626–1632.

[154] Theriault, G. (1992). Electromagnetic fields and cancer risks, *Revue d'Épidemiologie et de Santé Publique* **40**, S55–S62.

[155] Tiller, J.R., Schilling, R.S.F. & Morris, J.N. (1968). Occupational toxic factor in mortality from coronary heart disease, *British Medical Journal* **4**, 407–411.

[156] Tola, S., Vilhunen, R., Järvinen, E. & Korkola, M.-L. (1980). A cohort study of workers exposed to trichloroethylene, *Journal of Occupational Medicine* **22**, 737–740.

[157] Torchio, P., Lepore, A.R., Corrao, G., Comba, P., Settimi, L., Belli, S., Magnani, C. & di Orio, F. (1994). Mortality study on a cohort of Italian licensed pesticide users, *Science of the Total Environment* **149**, 183–191.

[158] Ulfvarson, U. (1992). Validation of exposure information in occupational epidemiology, *American Journal of Industrial Medicine* **21**, 125–132.

[159] van Miller, J.P., Lalich, J.J. & Allen, J.R. (1977). Increased incidence of neoplasms in rats exposed to low levels of tetrachlorodibenzo-p-dioxin, *Chemosphere* **6**, 537–544.

[160] van Vliet, C., Swaen, G.M., Volovics, A., Tweehuysen, M., Meijers, J.M., de Boorder, T. & Sturmans, F. (1990). Neuropsychiatric disorders among solvent-exposed workers. First results from a Dutch case-control study, *International Archives of Occupational and Environmental Health* **62**, 127–132.

[161] Vineis, P. (1992). Uses of biochemical and biological markers in occupational epidemiology, *Revue d'Épidemiologie et de Santé Publique* **40**, S63–S69.

[162] Vineis, P. & Soskolne, C.L. (1993). Cancer risk assessment and management. An ethical perspective, *Journal of Occupational Medicine* **35**, 902–908.

[163] Wagner, J.C., Sleggs, C.A. & Marchand, P. (1960). Diffuse pleural mesothelioma and asbestos exposure in North Western Cape Province, *British Journal of Industrial Medicine* **17**, 260–271.

[164] Wagoner, J.K., Archer, V.E., Carrol, B.E., Holaday, D.A. & Lawrence, P.A. (1964). Cancer mortality patterns among U.S. uranium miners and millers, 1950 through 1962, *Journal of the National Cancer Institute* **32**, 787–801.

[165] Wagoner, J.K., Miller, R.W., Lundin, Jr, F.E., Fraumeni, J.F. & Haij, N.E. (1963). Unusual mortality among a group of underground metal miners, *New England Journal of Medicine* **269**, 281–289.

[166] Wertheimer, N. & Leeper, E. (1979). Electrical wiring configurations and childhood cancer, *American Journal of Epidemiology* **109**, 273–284.

[167] Whorton, M.D. (1980). Recovery of testicular function among DBCP workers, *Journal of Occupational Medicine* **22**, 177–179.

[168] Whorton, M.D., Krauss, R.M., Marshall, S. & Milby, T.H. (1977). Infertility in male pesticide workers, *Lancet* **ii**, 1259–1261.

[169] Wingren, G. & Axelson, O. (1992). Cluster of brain cancer spuriously suggesting occupational risk among glassworkers, *Scandinavian Journal of Work, Environment and Health* **18**, 85–89.

[170] Zocchetti, C., Consonni, D. & Bertazzi, P.A. versus Lee, J. (1995). Letters to the Editor. Estimation of prevalence rate ratios from cross-sectional data, *International Journal of Epidemiology* **5**, 1064–1067.

OLAV AXELSON

Occupational Health and Medicine

Occupational health is an activity organized to protect the health of employees from harmful consequences arising out of their work. It includes industrial medicine and occupational medicine which also provides medical **surveillance** services. The aims are to reduce the frequency and severity of occupational diseases, i.e. diseases caused or exacerbated by the occupational environment, and hence to reduce premature death and disability. The prevention or reduction of occupational disease is emphasized, and this involves changes in the occupational environment that may be achieved by those practicing occupational hygiene. The importance given to occupational health is dependent on the social attitudes of the population in which it is based, and many of the occupational health questions now considered would have seemed of trivial concern a few decades ago. Occupational health includes safety, sometimes emphasized by the use of the term "health and safety", which is concerned with the reduction of accidents in the workplace.

Important elements are first the identification of adverse health effects of occupational exposure to a pollutant (*see* **Occupational Epidemiology**) and, secondly, the implementation of measures to reduce exposure and hence the frequency or severity of adverse health effects. This second step may involve standard setting, i.e. the setting of exposure limits predicted to lead to minimal adverse health effects. The setting of such limits may utilize **dose–response** data but also takes into account practicality and the milieu in which employers and workers interact.

Historical Development

The Italian physician, Bernardino Ramazzini (1633–1714), has been referred to as the "father of occupational medicine" [13]. He was professor of medicine at the Universities of Modena and Padua. He stressed the importance of direct examination of workers and introduced the concept of the taking of an occupational history. He described diseases associated with a wide range of occupations and their causes [24], including diseases caused by the inhalation of dusts and gases and those caused by poor ergonomic practices. Ramazzini's publications were the main source on illnesses caused by work for over a century [18]. Popper [22] drew the distinction between "occupational diseases" and "workers diseases", the former restricted to diseases caused directly by some intrinsic feature of the occupation, such as exposure to a chemical, while the latter also includes diseases occurring for socioeconomic reasons associated with the occupation. Ramazzini had noted that breast cancer was more prevalent in nuns than other women, and attributed this to celibacy.

In the eighteenth century there was a growing concern on the effects of industrialization. **Guy** [10] analyzed the proportion of deaths due to pulmonary consumption in broad occupational groupings and attributed an excess of such deaths to poor conditions, such as inadequate drainage and ventilation, an inadequate water supply, and overcrowding, in both dwellings and workshops. **William Farr** (1807–1883), the first compiler of Abstracts in the General Register Office of England and Wales, introduced a classification of occupations in 1851 that was used for the analysis of **occupational mortality** from official **vital statistics.** Later, **Hill** [12] used national insurance statistics to examine sickness absences of printers, cotton weavers, and spinners. From about this time, and particularly after the end of World War II, there was increasing attention to research into the extent and causes of occupational diseases, and this research necessitated the application of statistical concepts and methods into the design and analysis of studies.

Types of Study

The relationship between the occupational environment and health has been studied in a variety of ways. Insofar as occupational diseases may be caused by industrial pollutants, then basic studies of the interaction between pollutants and biological systems have been carried out in toxicology, by *in vitro* experiments, and by animal experimentation.

Studies of humans have involved experimentation, looking at acute effects, but such studies are infeasible, and unethical, for chronic effects. Such effects have been studied by epidemiologic investigations

(*see* **Observational Study**). The outcome variables include mortality, the occurrence or presence of a disease, and the value of a variable that measures some function of health (*see* **Health Status Instruments, Measurement Properties of**). For example, measures of lung function, such as the forced expiratory volume, are indicators of health but do not indicate disease unless grossly abnormal. Disease outcomes include diseases specific to occupational exposure, e.g. the pneumoconioses are caused by the inhalation of dust or fibers. They also include nonspecific diseases, the incidence of which may be increased by occupational exposure. For example, the frequency of lung cancer is increased by exposure to asbestos and a number of chemicals, but it occurs also in the absence of such exposures.

The main study designs employed have been:

1. **Cross-sectional study** – a study examining the relationship between disease, or a measure of health, and occupational conditions at a given time.
2. **Cohort study** – a study in which subjects are selected and followed-up over time to observe their mortality, morbidity, or changes in some functional measure of health, and to relate these to the exposure within the occupation. Cohort studies have been used particularly in the study of cancer, and frequently are *historical cohort studies* in which the cohort is defined in terms of existing historical data, such as records of employees. Follow-up is then from some time in the past – often many years or decades – to the present (*see* **Cohort Study, Historical**).
3. **Case–control (case-referent) study** – A study in which cases of disease are identified and noncases are chosen as **controls** or referents. The previous occupational exposure history of cases and controls are then ascertained and compared to give estimates of the **association** between the occupational environment and the disease. Case–control studies may be nested within a cohort study so that the cohort study is used to identify cases of disease in an occupational population and the case–control study is used to obtain more detailed information on exposure on a smaller number of cases and controls than would be practical for all members of a large cohort (*see* **Case–Control Study, Nested**) [16, 14].

Landmark Studies

One of the earliest controlled **clinical trials** took place in the area of occupational health. J. Lind, a naval surgeon on the *Salisbury*, divided a group of 12 seamen suffering from scurvy into six groups of 2. One of the treatments was two oranges and a lemon a day, and the "most sudden and visible good effects" were observed in the two seamen receiving this treatment. Regrettably it took another 50 years before lemon juice was supplied as a dietary supplement on British naval vessels [15; 21, pp. 14, 15].

The association between exposure to occupational environments and cancer has been a constant theme. The first malignant disease to be associated with a particular occupation was cancer of the scrotum in chimney sweeps, described by Percivall Pott [23]; see [20] and [27]. Härting & Hesse [11] reported an excess of respiratory cancers in underground metal miners, later shown to be due to radon daughters. This was the first association between an external agent and an internal cancer. The studies of Doll [6] on lung cancer amongst gas-workers, and of Case et al. [2] on bladder cancer and exposure to chemicals, were important studies, not only with respect to the identification of occupational carcinogens, but also in the early use of historical prospective cohort studies and the use of the **person-years at risk** method of mortality analysis. The risk of exposure to asbestos in producing lung cancer has been identified and confirmed using historical prospective studies [7, 25, 19], and the strong link between asbestos exposure and mesothelioma has been evaluated using similar studies following the identification of the link by Wagner et al. [26].

Particular Statistical Concepts, Problems, and Techniques

Most of the statistical methods used in occupational health are not unique to that area. Problems of potential **bias** due to **confounding** are often present in epidemiologic studies since the workers in groups to be compared may differ systematically in other characteristics. The confounding effect of smoking in studies on occupational respiratory diseases is a particular problem because of the large effect on health of the smoking habit (*see* **Smoking and Health**). A particular type of bias, which may be

considered as due to confounding, but is more usefully considered separately, is the "healthy worker effect", which may arise in any epidemiologic study in which a workforce is compared with the general population or different workforces are compared with one another. A second problem is that due to the *latency* of occupational cancers, i.e. an excess of cancer due to occupational exposure does not occur until many years after the exposure (*see* **Latent Period**). Related to this, during a follow-up study of current workers the extent of exposure is constantly changing during the period of follow-up when adverse health effects are being noted. This makes the linking of the extent of exposure to the health effect difficult. A full account of many of the statistical methods used in occupational health is given by Checkoway et al. [3].

Healthy Worker Effect

Studies of occupational health are carried out in groups of workers who select themselves, and are selected by employers, into particular occupations. These selection processes lead to the "healthy worker effect", which occurs because they are likely to eliminate the most unhealthy from entering the workforce and also may mean that those developing ill health are less likely to remain in the job. As noted by McMichael [17], these selection effects lead to lower mortality rates than would be expected. Fox & Collier [9] noted that the healthy worker effect consisted of three components, which they attributed to selection (*see* **Selection Bias**), survival, and length of follow-up. They found that the effect was particularly marked in the first 10 years after the start of employment in which there was exposure to vinyl chloride. The consequence of the healthy worker effect is that comparisons of mortality between employed groups and the general population may be biased unless workers are followed-up after they have left the employed group and unless comparisons take account of time since the start of employment. Methods of analysis using internal comparisons may take account of the initial selection criteria provided that these were applied similarly to the groups under comparison.

The Analysis of Mortality – External Comparisons

One method of analysis of follow-up studies where the outcome is mortality consists of the comparison of the observed number of deaths due to all or specific causes with the number that would be expected taking into account the age distribution of the cohort studied, the period during which follow-up occurred, and the varying lengths of follow-up of the different subjects in the study. The indirect method of standardization is the basis of the method, referred to as the *person-years* method (*see* **Standardization Methods**). This method was first used by Doll [6] and has been commonly applied in occupational mortality studies. The total follow-up period is divided into periods, usually of 5 years, and within each of these periods the ages of those persons followed within that period of time are similarly subdivided, again usually into 5-year groups. The expected number of deaths is calculated by multiplying the person-years at risk in each of the age–period intervals by the age- and period-specific death rates of a standard population. The standard population is a national or regional population for which death rates are available. The usual measure of effect is the ratio of observed to expected deaths which, in analogy with indirect standardization, is often referred to as an SMR.

The method has usually been applied to single groups or descriptively to a few subgroups, but may be extended to take account of other variables recorded for the persons followed-up using **Poisson regression** [1].

The Analysis of Mortality – Internal Comparisons

A disadvantage of the person-years method is the implicit assumption that the death rates in the occupational group would be the same as in the standard population except for factors within the occupational environment. While this assumption is not strictly necessary, the healthy worker effect has to be taken into account in the interpretation of the results. The method does assume that the death rates in the occupational group and the standard population are in a fixed proportion in all the age–period groups, i.e. that the standard death rates apply to the occupational group at least proportionally. The approach may be modified by working internally within the occupational group and avoiding the use of a standard population for comparison. This leads to internally calculated SMRs, and comparison of more than two of these depends on **proportional hazards** of the subgroups across the age-period strata. This problem may be overcome by the use of directly

standardized rates, which leads to the *standardized rate ratio* (SRR).

All these methods involve **stratification** and become imprecise when there are several variables to take into account. Stratification can be avoided by using the most general method, i.e. Poisson modeling (*see* **Poisson Regression**).

As it is often required to assess the mortality of an occupational group in the context of the population of which the workers are a part and also to assess dose–response relationships within the occupational group, or to compare subgroups within the whole occupational group, a combination of external and internal methods is appropriate. For example, Checkoway et al. [4], in a study of workers employed in the mining and processing of diatomaceous earth, compared the mortality due to lung cancer in the whole group. The methods of analysis included comparison with US national death rates using the SMR and internal analyses using Poisson regression to examine the trend with duration of employment. The combination of methods led to the conclusion that there was an excess of deaths due to lung cancer in the workforce, compared with the US population, and that this excess was associated with duration of employment in dust-exposed jobs.

Time-Related Exposure and Covariates

A measure of exposure to an occupational agent that may be used is cumulative exposure, i.e. the intensity of exposure accumulated to give a time-weighted cumulative measure. Clearly this measure will continue to increase as long as exposure is occurring. Methods of analysis relating health outcomes to exposure have to take this into account, not only with respect to the appropriate exposure to link with the outcome event, but also, equally importantly, to allocate the earlier years, when the event did not take place, to the lower cumulative exposure; failure to do this results in biased results [8]. Other variables that may be included in the analysis as **covariates** (confounders) may also be time-dependent (*see* **Time-dependent Covariate**) – for example, smoking and age. One way of dealing with this problem is through a proportional hazards model [5] with time-related variables. Another way when there are no covariates is to use the person-years method, or corresponding internal methods, with each individual transferring from one cumulative exposure category to the next when the cumulative exposure reaches the appropriate values.

Latency

Many diseases do not occur until some time after the exposure that has caused the disease. In particular, occupational cancers usually do not occur, i.e. they cannot be diagnosed, until at least 10 years after exposure. This feature should be incorporated into the analysis since, otherwise, very recent exposure, that cannot be relevant to the disease, is included as if that exposure were relevant. One method of dealing with this problem is to begin the follow-up after an interval, of say 10 years, since the start of exposure, and to ignore deaths and person-years at risk within this interval. This method does not allow regression-type methods including cumulative exposure but may be extended for this situation by defining a lagged cumulative exposure; for example, the lagged cumulative exposure relevant to the disease risk of a 45-year-old worker would be the cumulative exposure at age 35.

Allowing for latency also helps to reduce the influence of the healthy worker effect.

Future Developments

The perceived importance of occupational diseases is dependent on societal attitudes. The major occupational effects of early industrialization have been eliminated, and the large excess death rates due to exposure to asbestos and other occupational pollutants have been identified and preventive measures taken, although the **excess mortality** continues to occur because of the long latency effects. Concern has moved to the possibility or suspicion of smaller effects. Methods will need to be developed to identify small effects of occupational pollutants. As the estimation of small effects may not be achievable by epidemiologic studies, there is likely to be an increasing emphasis on biological methods and on the translation of biological findings into meaningful **risk assessments** for exposed populations.

The concept of individual susceptibility to disease has been a long-standing concept, but there has been insufficient knowledge for most occupational diseases on how to identify those who would be most susceptible to occupational exposure. This means

that it has not been practical to screen out of the exposed workforce those most likely to develop disease due to exposure to a pollutant. Advances in **molecular epidemiology** may contribute to this area by leading to the possible identification of susceptible individuals, who could be advised to avoid employment in particular industries.

References

[1] Berry, G. (1983). The analysis of mortality by the subject-years method, *Biometrics* **39**, 173–184.

[2] Case, R.A.M., Hosker, M.E., McDonald, D.B. & Pearson, J.T. (1954). Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. Part I, *British Journal of Industrial Medicine* **11**, 75–104.

[3] Checkoway, H., Pearce, N. & Crawford-Brown, D.J. (1989). *Research Methods in Occupational Epidemiology*. Oxford University Press, New York.

[4] Checkoway, H., Heyer, N.J., Demers, P.A. & Breslow, N.E. (1993). Mortality among workers in the diatomaceous earth industry, *British Journal of Industrial Medicine* **50**, 586–597.

[5] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[6] Doll, R. (1952). The causes of death among gas-workers with special reference to cancer of the lung, *British Journal of Industrial Medicine* **9**, 180–185.

[7] Doll, R. (1955). Mortality from lung cancer in asbestos workers, *British Journal of Industrial Medicine* **12**, 81–86.

[8] Enterline, P.E. (1976). Pitfalls in epidemiologic research: an examination of the asbestos literature, *Journal of Occupational Medicine* **18**, 150–156.

[9] Fox, A.J. & Collier, P.F. (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry, *British Journal of Preventive and Social Medicine* **30**, 225–230.

[10] Guy, W.A. (1844). A third contribution to a knowledge of the influence of employments upon health, *Journal of the Statistical Society of London* **7**, 232–243.

[11] Härting, F.H. & Hesse, W. (1879). Der Lungenkrebs, die Bergkrankheit in den Schneeberger Kobaltgruben, *Vjschr Gericht Med Offentl Gesundheitswesen* **31**, 102–132, 313–337.

[12] Hill, A.B. (1929). An investigation of sickness in various industrial occupations, *Journal of the Royal Statistical Society* **92**, 183–238.

[13] Koelsch, F., ed. (1912). *Bernardo Ramazzini, der Vater der Gewerbehygiene*. Stuttgart.

[14] Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977). Methods of cohort analysis: appraisal by application to asbestos mining (with discussion), *Journal of the Royal Statistical Society, Series A* **140**, 469–491.

[15] Lind, J. (1753). *A Treatise of the Scurvy*. Sands, Murray & Cochran, Edinburgh.

[16] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics* **29**, 479–486.

[17] McMichael, A.J. (1976). Standardized mortality ratios and the "healthy worker effect": scratching below the surface, *Journal of Occupational Medicine* **18**, 165–168.

[18] Milles, D. (1985). From workers' diseases to occupational diseases: the impact of experts' concepts on workers' attitudes, in *The Social History of Occupational Health*, P. Weindling, ed. Croom Helm, London, pp. 55–77.

[19] Newhouse, M.L. (1969). A study of the mortality of workers in an asbestos factory, *British Journal of Industrial Medicine* **26**, 294–301.

[20] Ogle, W. (1885). Letter to the Registrar-General on the mortality in the registration districts of England and Wales during the ten years 1871–80, in *Supplement to the 45th Annual Report of the Registrar General of Births. Deaths, and Marriages, in England*, p. xxiii.

[21] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.

[22] Popper, M. (1882). *Lehrbuch der Arbeiterkrankheiten und Gewerbehygiene*. Zwanzig Vorlesungen, Stuttgart.

[23] Pott, P. (1775). *Chirurgical Observations*, Vol. 3. L. Hawes, W. Clark, and R. Collins, London, pp. 177–183.

[24] Ramazzini, B. (1700). *De Morbis Artificum* (translated by W.C. Wright as *Diseases of Workers*. Hafner, New York, 1964).

[25] Selikoff, I.J., Churg, J. & Hammond, E.C. (1964). Asbestos exposure and neoplasia, *Journal of the American Medical Association* **188**, 22–26.

[26] Wagner, J.C., Sleggs, C.A. & Marchand, P. (1960). Diffuse pleural mesothelioma and asbestos exposure in the North Western Cape Province, *British Journal of Industrial Medicine* **17**, 260–271.

[27] Waldron, H.A. (1983). A brief history of scrotal cancer, *British Journal of Industrial Medicine* **40**, 390–401.

G. BERRY

Occupational Mortality

The study of occupational mortality involves the systematic tabulation of mortality by occupational groups or by socioeconomic groups (*see* **Social Classifications**) when these are defined by occupation. Three main methods are used to conduct these studies.

The first method, **cross-sectional studies**, utilizes the number of deaths occurring to persons in a given occupation during a given time period divided by the number of persons in that occupation in the middle of the period. The source for the numerator is usually **death certificates**; the denominator is usually based on the **census**. As the age distribution varies considerably between the occupational groups, an age **standardization** is needed in order to compare the mortality of different occupational groups. In the cross-sectional studies, the comparative mortality figure (CMF, direct standardization) or the standardized mortality ratio (SMR, indirect standardization) are used as summary measures of an occupational group's relative mortality.

The second method, death certificate studies, involves the distribution of deaths by cause for a given occupational group compared with the distribution for a total population without regard to occupation. Such studies are often sex-specific or limited to the male population. Here **proportional mortality ratios (PMRs)** are used as summary measures for each occupational group's relative mortality from a given cause of death.

The third method, follow-up studies (*see* **Cohort Study**), is based on individually matched records and typically on census data. A census population is followed up for deaths and emigrations, and maybe also for new census data, which allows separate analyses of persons who stayed in an occupation from one census to the next. In these studies various methods are used for the matching of individual records (*see* **Record Linkage**). In the UK, for example, the study population is flagged in the National Health Service Central Register; whereas in the Nordic countries the matching is based on the personal identification numbers used in both censuses and death and emigration registrations. In the follow-up studies, the CMFs and SMRs are often used as summary measures for an occupational group's mortality. However, each individual in these studies has a record containing the census characteristics, the number of **person-years at risk**, and the eventual **cause of death**. It is therefore possible to use these data sets also for internal comparisons of the mortality between occupational groups; for example, controlled for sex, age, marital status, and region.

Cross-Sectional Studies

The study of occupational mortality is closely linked to procedures developed in England and Wales, where the first cross-sectional study was published in 1855 [37]. Since then, occupational mortality studies have been published every ten years; no other country has a similar record. The potentials and limitations of cross-sectional studies are therefore best illustrated by this series of data.

It was realized in England in the late 1840s that

> if the age of the various classes of society ... are abstracted from the census returns ... and if the deaths are abstracted in the same classes ... the relative mortality ... can be satisfactorily deduced ... and much light will be thrown upon the causes which really influence the health and well being of the working, middle and higher classes [36].

Farr used this method to tabulate the mortality for 1851 by occupation, and commented that

> the professions and occupations of men open a new field of inquiry, on which we are now prepared to enter, not unconscious, however, of the peculiar difficulties that beset all inquiries into the mortality of limited, fluctuating, and sometimes ill-defined sections of the population [37].

The methodological problems entailed in occupational mortality studies were thus realized from the very beginning. In 1851, miners, bakers, butchers, and inn and beershop keepers experienced the heaviest rates of mortality.

CMFs were first calculated for the 1880–1882 data. The 1900–1902 data showed a variation in the CMF from 600 or below for clergyman, priest, minister; gardener, nurseryman, seeds-man; gamekeeper; and farmer, grazier, farmer's son, etc.; to 1800 or above for inn-, hotel-servant; costermonger, hawker; tin miner; and general laborer [38].

In the 1910–1912 decennial supplement, Stevenson "included for the first time an attempted grading of the male working population into eight social

classes as determined by occupation", where "I represented as far as possible the middle and upper classes ... III skilled labour, and V unskilled labour" [46]. The social classes II and IV were intermediate classes, and textile workers, miners, and agricultural workers formed separate classes. **Infant mortality** showed a steep gradient from 76 deaths per 1000 births in social class I to 153 in social class V. The CMFs for men aged 25–65 for the classes I–V were 88, 94, 96, 93, and 142. Commercial clerks with an CMF of 108 formed 28% of class I and thus inflated the overall CMF for class I. This gave rise to a discussion about criteria for grouping of occupations into social classes, a discussion which has persisted ever since.

The changing composition of the labor force made it necessary to shift from an industry based classification to one in which it was "no longer necessary to assign the head of a tinplate etc. works to the same social class as his labourers" [39]. In 1921–1923, the relative mortality for occupations and social classes were presented both by the CMFs and by summary measures similar to SMRs. The CMFs for social classes I–V were 821, 942, 951, 1007, and 1258; and the SMRs were 82, 94, 95, 101, and 125. The two methods of calculation gave similar results, illustrating a subsequently often observed **robustness** of occupational mortality data.

From a previous holistic view of occupation as encompassing both occupational risks and living conditions, an interest in separating the two aspects emerged in the 1930s. To address this question, the 1930–1932 decennial supplement added tables on the mortality of married women by the social class of their husbands. Similar gradients of SMRs from social classes I–V were found; for men 90, 94, 97, 102, and 111; and for women 81, 89, 99, 103, and 113; and it was concluded "that the contribution made by actual work done to men's social mortality gradient from all causes must be small compared with the contribution made by the accompanying environmental, economic and selective factors" [47].

The 1951 classification included 600 occupations. However, the more detailed classification increased the risk for discordance between numerators and denominators. There was a tendency for people to be given more prestigious occupational titles on death certificates than they had in the census: in 1949–1953, for example, 1443 deaths were registered among company directors compared with only 98 expected deaths based on the number of company directors registered at the census and the mortality rates for all men. As a supplement to the SMRs for deaths in persons aged 20–64, PMRs were therefore presented for deaths in persons aged 65–74 to avoid the problem of discordance between the numerators and denominators. The post-war concern about equity was reflected in systematic tabulations of trends in social class differences in the 1949–1953 decennial supplement. The SMRs for men from social classes I–V were 98, 88, 101, 104, and 118 [40], showing that the English society still had a disadvantaged social class V. The relatively high SMR for social class I was partly due to **misclassification** [24]. A new observation was that the social class gradient of some diseases changed as they became more frequent. The mortality from lung cancer almost tripled in men under 65 years from 1921–1923 to 1930–1932, and the social class gradient changed from an excess risk in social class I in 1921–1923, to equal risks across social classes in 1931–1932, and to a clear excess risk in social class V in 1949–1953. Coronary heart disease increased rapidly in men after the war. In 1930–1932 it was a disease of "the better classes" [11], with an SMR gradient from 237 in social class I to 67 in social class V, but the gradient in 1949–1953 was from 150 in social class I to 89 in social class V.

"The most disturbing feature of the [1959–1963 decennial supplement] when compared with earlier analyses [was] the apparent deterioration in social class V" [41]. The SMRs for the five social classes were 76, 81, 100, 103, and 143. A new and shorter occupational classification was used in 1960, but the results for social class V remained "even when the rates were adjusted to the 1950 classification" [41]. The social class gradient in lung cancer had become even steeper, with SMRs of 53 in social class I and 148 in social class V. The change in the social distribution of coronary heart disease had continued, and the SMRs were now 98 for social class I and 112 for social class V.

When it came to results for specific occupations in 1959–1963, the concern about discordance between numerators and denominators clearly influenced the interpretation. Following a review of occupational cancers, Adelstein wrote that "although the exercise known as occupational mortality is not a useful tool as an early warning system, it remains a valuable analysis of mortality of groupings of occupations

as a back-up and reference system" [1]; a modest aim compared with the expectations a hundred years earlier.

The social class III was in 1970–1972 divided into nonmanual and manual workers. The SMRs for men over the six social classes were 77, 81, 99, 106, 114, and 137. Similar patterns were seen for women and for stillbirths, and mortality of infants and children. Gradients across social classes showed, for example, that social class V had an almost fourfold risk of respiratory diseases but the same risk of pancreatic cancer as social class I. Some occupations stood out with high risks for specific diseases; for example, butchers with cancer of the lung and maxillary sinus, and electro- and dip platers with cancer of the lung [33].

The 1972 smoking rates for men by occupational order showed a high positive **correlation** with the lung cancer mortality. Under these circumstances the previous concern about the influence of living conditions on occupational mortality became a concern about "life-styles of persons" [33]. In a paper on "work or way of life", Fox & Adelstein found that only 18% of the variation in mortality between occupational orders remained when the mortality was standardized for social class [12].

In 1979–1983, the social gradient in mortality for men went from an SMR of 66 in social class I to 165 in social class V; the official comment being that "these data are subject to serious bias and do not represent usable estimates of mortality by social class" [34]. However, aggregated into nonmanual workers (social classes I, II, and IIIN) and manual workers (social classes IIIM, IV, and V), where a serious misclassification would not be expected, the data showed "an overall fall in all-cause mortality from 1970–1972 to 1979–1983 for both manual and nonmanual occupational classes, but the rate of decline [had] been greater in nonmanual groups. Thus the social gap [had] widened" [26].

As a consequence of concern about biases, the 150 year old practice of combining census and death certificate data was abolished with the next decennial supplement, that for 1990.

Cross-sectional studies of occupational mortality were undertaken in several other countries – for example, in France in 1907–1908 [19], and in the US in 1930 [50] and 1950 [14–16, 20] – but in no other country did these studies have the same importance for the social debate as in England and Wales.

Death Certificate Studies

At the time of the 1990 decennial supplement for England and Wales, advancement in technology had rendered the cross-sectional method obsolete, as the overall mortality of an occupational group could now be estimated from the individually matched records of a follow-up study known as the Longitudinal Study (see below). However, the death certificate data were used for search of specific associations between detailed occupations and causes of deaths using only PMRs. This analysis included 1.8 million deaths from 1979–1980 and 1982–1990. New observations were, for example, an excess risk of leukemia, lymphoma, aplastic anemia, and agranulocytosis in teachers, "suggesting a possible hazard from exposure to childhood infections" [8].

In the US, large-scale death certificate studies started in the 1970s. The first study was from Washington State and covered deaths from 1950 to 1971 [27]. An update included deaths from 1950 to 1979. The systematic tabulation of PMRs for detailed occupation and cause of death revealed, for example, an increased risk of leukemia in workers exposed to electric and magnetic fields and a deficit of multiple sclerosis among outdoor workers [28].

Similar studies were undertaken in California [35], Massachusetts [10], Utah [3], and Rhode Island [21]. A detailed analysis of the large number of solvent exposed jewelry workers from Rhode Island revealed an excess mortality from mental disorders, kidney diseases, liver, and kidney cancer [9].

A standardized reporting, coding, and registration scheme for occupation on death certificates from 12 states started in the US in 1984. A PMR analysis for broad groups has been published for the 270 000 deaths occurring in 1984 for persons above the age of 20 [42]. Data from death certificates from 1985 to 1991 are available on public-use data tapes [31].

Follow-Up Studies

That the numerator/denominator bias in cross-sectional studies could be overcome in follow-up studies of census populations was realized as early as the 1920s in the Nordic countries, among others [7]. It was, however, 50 years later that such studies started to emerge. The earliest studies were a 4 month follow-up of the 1960 census population from the

US [22], a 5 year follow-up of the Norwegian 1960 census population [48], a 17 year follow-up of a sample of the French 1954 census population [5], and a 10 year follow-up of the Swedish 1960 census population [45].

The 1970 censuses were used for follow-up studies of the national populations In Denmark [2, 4], Finland [25, 44], and Norway [23], and a joint analysis was made of the occupational mortality in Norway, Sweden, Finland, and Denmark for the 10 year period 1971–1980 [32]. The Longitudinal Study from the UK is a follow-up study of a 1% sample of the 1971 census population, where the sample is, in addition, continuously supplemented with 1% of births and immigrations [13]. A 6 month follow-up of the Italian 1980 census population was published in 1995 [43]. In addition to sex, age, and occupation, these studies often also include other census variables such as marital status, housing conditions, region, family composition, etc.

Some important and fairly consistent observations have been made from the follow-up studies of census populations, such as:

1. All marginal groups of the labor market have an excess overall mortality compared with the working population. In the Nordic countries in 1971–1980, the SMR for economically inactive men was 233, when the economically active men were used as the standard. The SMR for economically inactive women, the majority being housewives, was 151 [32]. In England and Wales in 1971–1975, the SMRs for unemployed men or men with an inadequately described occupation were 306 and 185, respectively, when all men were used as the standard [13].
2. There is a social class gradient in the overall mortality. The SMRs for men in England and Wales in 1971–1975 varied from 80 in social class I to 115 in social class V [13]. The short follow-up of the US 1960 census showed for white men aged 25–64 a mortality ratio of 0.92 for white-collar workers and of 1.07 for blue-collar workers, when the mortality of all men was used as the standard [22].
3. Farmers have a low overall mortality in many countries. The mortality ratio for white male agricultural workers was 0.76 in the US study [22], the SMR for farmers in the Nordic countries was 87 [32], and farm workers in Italy had a 20% deficit in overall mortality compared with all economically active men [43].
4. When studied, the social class gradient in mortality seems to have widened. In Finland in 1971–1983, the overall mortality for men aged 35–49 declined for all occupational classes, but in 1984–1990 it increased for workers and farmers while declining further for white-collar employees [49]. In France, the overall mortality for men aged 35–60 decreased by 28%–30% for professionals, foremen, and salaried employees, but by only 7%–12% for skilled, unskilled, and farm workers from 1955–1959 to 1975–1980 [6].

The follow-up studies of census data are often criticized for lack of information on personal habits, especially tobacco smoking. Such data are available in a follow-up study of 300 000 US veterans who were interviewed about their smoking habits in 1954 and 1957. A follow-up study of this cohort provides smoking adjusted **relative risks** [17, 18], but the number of deaths for a given combination of occupation and cause of death is often small.

However, data from smaller cohorts with a broad range of recorded variables may often provide useful supplementary information to the census studies. Data from the Longitudinal Study have shown, for example, that men unemployed at the time of the census have an excess mortality in the subsequent years [30]. Very useful supplementary data came from the British Regional Heart Study, where unemployed men had an excess mortality even when controlled for age, town, social class, smoking, alcohol intake, and pre-existing diseases [29].

However, at present follow-up studies of census populations provide the most comprehensive data on overall mortality by occupational groups in unselected populations.

References

[1] Adelstein, A.M. (1972). Occupational mortality: cancer, *Annals of Occupational Hygiene* **15**, 53–57.

[2] Andersen, O. (1985). *Mortality and Occupation 1970-80.* Statistical Studies No. 41. Danmarks Statistik, København (in Danish).

[3] Bangerter, N.H., Dandoy, S., Elison, G. & Brockert, J.E. (1985). Utah's Occupational Health Surveillance System, 1980–82. Collection of Technical Reports. Utah Department of Health, Salt Lake City.

[4] Danmarks Statistik (1979). *Mortality and Occupation*. Statistical Studies No. 37. Danmarks Statistik, København (in Danish).

[5] Desplanques, G. (1976). *Adult Mortality by Social Environment* 1955–1971. No. 195 des Collections de I'INSEE, Série D, No. 44 (in French).

[6] Desplanques, G. (1985). *Adult Mortality. Results of Two Longitudinal Studies (Period 1955–1980)*. No. 479 des Collections de I'INSEE, Série D, No. 102 (in French).

[7] Det Statistiske Departement (1921). Meeting of Nordic Statisticians, København, 29–31 August 1921. SM 4: 64: 3. Det Statistiske Departement, København (in Danish).

[8] Drever, F., ed. (1995). *Occupational Health. Decennial Supplement.* Office of Population Censuses and Surveys. HMSO, London.

[9] Dubrow, R. & Gute, D.M. (1987). Cause-specific mortality among Rhode Island jewelry workers, *American Journal of Industrial Medicine* **12**, 579–593.

[10] Dubrow, R. & Wegman, D.H. (1982). *Occupational Characteristics of White Male Cancer Victims in Massachusetts* 1971–73. National Institute for Occupational Health, US Department of Health and Human Services, Cincinnati.

[11] Editorial (1959). Health and social class, *Lancet* **i**, 303–305.

[12] Fox, A.J. & Adelstein, A.M. (1978). Occupational mortality: work or way of life?, *Journal of Epidemiology and Community Medicine* **32**, 73–78.

[13] Goldblatt, P., ed. (1990). *Longitudinal Study*. Mortality and Social Organization. Office of Population Censuses and Surveys. HMSO, London.

[14] Guralnick, L. (1962). *Mortality by Occupation and Industry among Men 20 to 64 Years of Age: United States*, 1950 Vital Statistics, Special Reports, Vol. 53 No. 2. National Center for Health Statistics, Washington.

[15] Guralnick, L. (1963). *Mortality by Occupation and Industry among Men 20 to 64 Years of Age: United States*, 1950 Vital Statistics, Special Reports, Vol. 53 No. 4. National Center for Health Statistics, Washington.

[16] Guralnick, L. (1963). *Mortality by Occupation and Industry among Men 20 to 64 Years of Age: United States*, 1950 Vital Statistics, Special Reports, Vol. 53 No. 5. National Center for Health Statistics, Washington.

[17] Hrubec, Z., Blair A.E. & Vaught, J. (1995). *Mortality Risks by Industry among U.S. Veterans of Known Smoking Status*. 1954–1980. Vol. 2. Department of Health and Human Services, Public Health Service, National Cancer Institute, NIH Publication No. 95–2747.

[18] Hrubec, Z. Blair, A.E., Rogot, E. & Vaught, J. (1992). *Mortality Risks by Occupation among U.S. Veterans of Known Smoking Status.* 1954–1980. Vol. 1. Department of Health and Human Services, Public Health Service, National Cancer Institutes, NIH Publication No. 92–3407.

[19] Huber, M. (1912). Occupational mortality, *Bulletine de la Statistique Générale de la France*, **1**, 402. Librairie Félix Alcan, Paris (in French).

[20] Kaplan, D.L., Parkhurst, E. & Whelpton, P.K. (1961). Comparability of Reports on Occupation from Vital Records and the 1950 Census. Vital Statistics, Special Reports, Vol. 53 No. 1. National Center for Health Statistics, Washington.

[21] Kelley, B.C. & Gute, D.M. (1986). *Surveillance Cooperative Agreement between NIOSH and States (Scans) Program. Rhode Island* 1980–82. National Institute for Occupational Health, US Department of Health and Human Services, Cincinnati.

[22] Kitagawa, E.M. & Hauser, P.M. (1973). *Differential Mortality in the United States: a Study in Socioeconomic Epidemiology*. Harvard University Press, Cambridge, Mass.

[23] Kristofersen, L. (1979). *Occupational Mortality.* Reports from Statistical Central Bureau 79/19. Statistical Central Bureau, Oslo (in Norwegian).

[24] Logan, W.P.D. (1959). Occupational mortality, *Proceedings of the Royal Society of Medicine* **52**, 463–468.

[25] Marin, R. (1986). *Occupational Mortality* 1971–80. Series no. 129, Central Statistical Office of Finland, Helsinki.

[26] Marmot, M.G. & McDowall, M.E. (1986). Mortality decline and widening social inequalities, *Lancet* **ii**, 274–276.

[27] Milham, S. (1976). *Occupational Mortality in Washington State* 1950–71. Vols. I–III. National Institute for Occupational Safety and Health, US Department of Health and Human Services, Cincinnati.

[28] Milham, S. (1983). *Occupational Mortality in Washington State* 1950–79. National Institute for Occupational Safety and Health, US Department of Health and Human Services, Cincinnati.

[29] Morris, J.K., Cook, D.G. & Shaper, A.G. (1994). Loss of employment and mortality, *British Medical Journal* **308**, 1135–1139.

[30] Moser, K.A., Goldblatt, P.O., Fox, A.J. & Jones, D.R. (1987). Unemployment and mortality: comparison of the 1971 and 1981 longitudinal study census samples, *British Medical Journal* **294**, 86–90.

[31] National Center for Health Statistics (1993). Mortality by occupation, industry, and cause of death: 12 reporting states, in *Final Data from the Centers for Disease Control and Prevention/National Center for Health Statistics.* Public-use Data Tapes 1985–91.

[32] Nordic Statistical Secretariat (1988). *Occupational Mortality in the Nordic Countries* 1971–1980. Statistical Reports of the Nordic Countries, No. 49, Nordic Statistical Secretariat, Copenhagen.

[33] Office of Population Censuses and Surveys (1978). *Occupational Mortality 1970–72. Decennial Supplement.* England and Wales. HMSO, London.

[34] Office of Population Censuses and Surveys (1986). *Occupational Mortality 1979–80, 1982–83. Decennial Supplement. Great Britain. Part I. Commentary.* HMSO, London.

[35] Petersen, G.R. & Milham, S. (1980). *Occupational Mortality in the State of California 1959–61*. National

Institute for Occupational Health, US Department of Health and Human Services, Cincinnati.

[36] Registrar-General (1846). *Seventh Annual Report on Births, Deaths, and Marriages in England.* HMSO, London.

[37] Registrar-General (1855). *Fourteenth Annual Report on Births, Deaths, and Marriages in England.* HMSO, London.

[38] Registrar-General (1908). *Supplement to the Sixty-fifth Annual Report on Births, Deaths, and Marriages in England. Part II.* HMSO, London.

[39] Registrar-General (1927). *Report on Occupational Mortality during 1921–23. Decennial Supplement. Part II*. HMSO, London.

[40] Registrar-General (1954). *Occupational Mortality. Decennial Supplement. England and Wales. Part I.* HMSO, London.

[41] Registrar-General (1971). *Occupational Mortality Tables. Decennial Supplement. England and Wales.* HMSO, London.

[42] Rosenberg, H.M., Burnett, C., Maurer, J. & Spirtas, R. (1993). Mortality by occupation, industry, and cause of death: 12 reporting states, 1984, in *National Center for Health Statistics. Monthly Vital Statistics Report 42*, no. 4, supplement, pp. 1–63.

[43] Roseo, G., ed. (1995). *Occupational Mortality in Italy in the 1980s*. Istituto Superiore per la Prevenzione e la Sicurezza del Lavoro, Roma (in Italian).

[44] Sauli, H. (1979). *Mortality, in Occupational Mortality in 1971–75*. Studies No. 54, Central Statistical Office of Finland, Helsinki.

[45] Statistiska Centralbyrån (1981). *Mortality Register 1961–1970.* Promemorior från SCB 81:5, Örebro (in Swedish).

[46] Stevenson, T.H.C. (1923). The social distribution of mortality from different causes in England and Wales, 1910–12, *Biometrika* **15**, 382–400.

[47] Stocks, P. (1938). The effects of occupation and of its accompanying environment on mortality, *Journal of the Royal Statistical Society* **101**, 669–708.

[48] Tønnesen, B.L. (1974). *Selected Aspects of the Mortality in Norway 1960–64 Compared with Other Countries*, Statistical Central Bureau, Oslo (in Norwegian).

[49] Valkonen, T., Martelin, T., Rimpalä, A., Notkala, V. & Savela, S. (1993). *Socio-economic Mortality Differences in Finland 1981–90. Population 1993:1*. Statistics Finland, Helsinki.

[50] Whitney, J.S. (1934). *Deaths Rates by Occupation, Based on Data of the U.S. Census Bureau, 1930*. National Tuberculosis Association, New York.

(*See also* **Occupational Epidemiology**; **Occupational Health and Medicine**; **Vital Statistics, Overview**)

ELSEBETH LYNGE

Odds Ratio

If two events, E_1 and E_2, have respective probabilities $\Pr(E_1)$ and $\Pr(E_2)$, the odds ratio comparing E_1 with E_2 is $[\Pr(E_1)/\{1-\Pr(E_1)\}]/[\Pr(E_2)/\{1-\Pr(E_2)\}]$, namely the ratio of the **odds** of E_1 to the odds of E_2.

MITCHELL H. GAIL

Odds

If an event E has probability $\Pr(E)$, the odds of the event is defined as $\Pr(E)/\{1 - \Pr(E)\}$.

MITCHELL H. GAIL

Office for National Statistics (ONS) (formerly OPCS)

The Office for National Statistics (ONS) lies at the centre of the UK Government Statistical Service, which aims to "provide Parliament, Government and the wider community with the statistical information, analysis and advice needed to improve decision making, stimulate research and inform debate". The Director is also the Head of the Government Statistical Service and Registrar General for England and Wales. He has a leading international role as a member of the UN Statistical Commission and of the Board of EUROSTAT, the statistical office of the European Union. The ONS was formed in 1996 as a result of the merger of the two major statistical agencies in the UK, the Office of Population Censuses and Surveys and the Central Statistical Office. These two government departments had long and distinguished histories [1, 2]. The new organization employs approximately 3000 people, spread between its offices in London, Newport, Southport, and Titchfield. The merger brought together the main collection, processing, analysis, and publication of social and economic statistics into a single organization. The Registrar General also has responsibility for the administration of the vital registration of births, marriages, and deaths and for the running of the National Health Service Central Register for England and Wales (*see* **Vital Statistics, Overview**).

In partnership with other government departments ONS aims to provide, for the UK, "the main statistical advisory service for policy formulation, resource allocation, planning, and research on the number, characteristics and health of the population. It collects and analyses data from a wide variety of sources, and publishes and interprets the statistics. Major customers are government, business, researchers and the general public." Its main outputs, increasingly available on the web [3], are:

1. regular publications, including the quarterly *Population Trend*s and *Health Statistics Quarterly*;
2. large-scale databases covering mortality, fertility, cancer incidence and survival, congenital anomalies, morbidity in general practice, and longitudinal data from the ONS Longitudinal Study; and
3. advice and interpretation of recent patterns and trends.

It conducts a significant fraction of its work in the health field on a commissioned basis, with the Department of Health for England and the National Health Service (NHS) as principal customers. ONS also plays a key role in providing a service to government covering the main **survey** collections. In recent years, this has included surveys on children, psychiatric morbidity, health and nutrition, dental health and disability, carried out for the Department of Health, as well as the regular General Household Survey.

The UK Government Statistical Service is distributed among the constituent countries of the UK. The Registrars General for Scotland and Northern Ireland have similar responsibilities for **censuses**, the administration of vital registration, and associated statistical activities to those held by the Registrar General for England and Wales. The Government Statistical Service also supports the Chief Medical Officers for England, Scotland, Wales, and Northern Ireland, supports the NHS in each of these countries, and provides policy advice to ministers. Statisticians in different policy departments in government take the lead on infectious diseases (Public Health Laboratory Service; *see* **Communicable Diseases**), **occupational health** (Health and Safety Executive), transport, and home accidents (Departments of Transport and Trade and Industry, respectively), health in prisons (Home Office), and the health of the armed forces (Ministry of Defence).

References

[1] Nissel, M. (1987). *People Count: a History of the General Register Office*. HMSO, London.

[2] Ward, R. & Doggett, T. (1991). *The First Fifty Years of the Central Statistical Office*. HMSO, London.

[3] `www.statistics.gov.uk`

JOHN FOX, KAREN DUNNELL & MICHAEL COLEMAN

Oncology

The subject of statistical methods applied to oncology covers a wide range of applications. Accordingly, the subject is considered here in three separate areas:

1. biostatistics in cancer epidemiology;
2. methods for cancer **clinical trials**; and
3. tumor modeling.

The three areas are interrelated, the most ubiquitous connecting factor being the seminal work of Mantel & Haenszel [30] on **relative risk** estimation in **case–control studies**. From this single paper, an enormous body of methodologic work has followed, and innumerable studies in the field have made use of the **Mantel–Haenszel** technique or of adaptations thereof.

Biostatistics in Cancer Epidemiology

Definitions

Although epidemiology of a kind which entails at least a basic application of scientific method has been intermittently apparent for centuries, biostatistics has only made a formal contribution since the twentieth century. The role of biostatistics has included advances in the methodology of study design as well as in analysis of data, and it is timely to define the three major design categories used in cancer epidemiology.

First, there is the prospective **cohort study**, in which a cohort of healthy individuals have certain characteristics recorded and are subsequently monitored for incidence of or mortality from a particular cancer. The aim of the study is to assess whether the characteristics recorded at recruitment are predictive (*see* **Prediction**) of subsequent disease. A characteristic which is related to incidence of disease is usually referred to as a **risk** factor. Related to this is the **prevention trial** in which healthy individuals are randomized (*see* **Randomization**) to receive or not to receive a particular intervention (for example, vitamin supplementation), and are subsequently followed up for the cancer in question to assess whether the intervention has a preventive effect.

Secondly, and most commonly in cancer epidemiology, there is the **retrospective** case–control design. In a study of this design, a series of cancer patients (cases) is compared with a series of subjects who do not have the cancer in question (controls), with respect to retrospectively assessed potential risk factors. This type of study has the advantage of being relatively quick to perform, but is more prone to **biases**, for example from differential recall between cases and controls (*see* **Recall Bias**), than the prospective cohort study [40].

Finally, there is **descriptive epidemiology**. This is not, strictly speaking, a design at all, being the practice and methodology of analysis of routinely collected national or regional figures on cancer incidence or mortality, to determine differences in morbidity or mortality by time or by geographic area.

To understand the landmark developments in statistical methods in epidemiology, it is necessary to define the fundamental measure of epidemiology, the **relative risk** (RR). Let D denote the event of disease occurring, $RF+$ denote the presence of a certain risk factor, and $RF-$ its absence. Let Pr denote probability. The relative risk is defined as

$$RR = \frac{\Pr(D|RF+)}{\Pr(D|RF-)}.$$

Historical Development

The first study of cancer epidemiology with a substantial biostatistical input was the Lane–Claypon case–control study of breast cancer [26]. This innovative study was the first rigorous case–control study addressing multiple risk factors and using a large cases base. Its findings included relationships of high parity with lowered breast cancer risk, late marriage with increased risk, miscarriage with increased risk and a family history of cancer with increased risk. This very much set the scene for later research, indicating a concentration on factors related to fertility and parity.

In a retrospective case–control study, the relative risk is not directly estimable but it may be approximated by the **odds ratio** (OR). This quantity is invariant to the design of the study. Suppose that the number of subjects with and without the disease and the risk factor are as in Table 1.

The odds ratio is defined as

$$OR = \frac{ad}{bc}.$$

Table 1 Symbolic tabulation of disease status by risk factor status

Risk factor	Disease present	Disease absent
Present	a	b
Absent	c	d

This is sufficient for the case in which there is only one risk factor of interest and it is assumed unrelated to other risk factors. In the case in which other factors are related both to disease risk and to the risk factor under study (such factors are known as **confounders**), it is clear that some amendment or complication of the formula is necessary. For this purpose, one of the most important developments has been the proposed estimate of **Mantel & Haenszel** [30]. A simple example of confounding arises in the assessment of the effect of alcohol drinking on lung cancer risk. It is known that smoking has a strong predisposing effect, and that heavy drinkers are more likely to be smokers than the rest of the population. How do we adjust our estimate of the effect of excessive drinking on lung cancer for the common association of disease and risk factor with smoking?

Let i denote smoking status, $i = 0$ representing nonsmokers and $i = 1$ representing smokers and suppose that we have a case–control study with data as in Table 2.

Thus there are two **two x two tables**, one for nonsmokers and one for smokers. A solution which intuitively recommends itself is a weighted average of the odds ratios calculated in each table separately. Mantel & Haenszel suggested weights $b_i c_i / N_i$, where $N_i = a_i + b_i + c_i + d_i$. This approximates weighting by inverses of the variances and gives the Mantel–Haenszel estimate familiar to epidemiologists

$$OR_{\mathrm{MH}} = \frac{\sum((a_i d_i)/N_i)}{\sum((b_i c_i)/N_i)}.$$

Table 2 Symbolic tabulation of lung cancer by heavy drinking and smoking

Heavy drinker	Lung cancer cases	Controls
Present	a_i	b_i
Absent	c_i	d_i

An accompanying **chi-square test** similarly adjusted for stratum was also developed. The original development of this estimate had no particularly strong theoretic backing, but the estimate has subsequently proved very **robust** and useful, and has been shown to possess various desirable properties, in particular its equivalence to the OR estimate from **conditional logistic regression** in the case of a matched study (see below). The disadvantage of the estimate is the lack of a universally applicable **variance** estimate. Different variance formulas apply depending on the sparseness of the data [18].

A copious amount of research has built on the Mantel–Haenszel estimate. For a review, see Gail [20]. The range of applications includes adaptations of the estimate for use in randomized trials, prospective **cohort studies**, **meta-analyses**, and studies in which risk factors are measured with error (*see* **Measurement Error in Epidemiologic Studies**).

Other landmarks in biostatistics as applied to cancer epidemiology include the development of **generalized linear models**. This revolutionized statistical analysis in many fields but was of particular benefit in epidemiology, enabling estimation of relative risks and odds ratios in a **regression** framework, notably **logistic regression** and **Poisson regression** [6, 7]. This made it considerably easier to adjust for confounding variables in analysis of **observational data**.

Current Approaches and Ongoing Problems

Other major developments include the evolution of methods to deal with mismeasurement of risk factors, beginning with Bross's [8] definition of the problem, with approaches including simple **likelihood**-based methods [15], imputation (*see* **Multiple Imputation Methods**), and regression methods [39] and, more recently, stochastic estimation [38] and **structural equations models** [25].

The development of computer software and hardware since the 1960s has gone hand in hand with the development of statistical models, in particular software for fitting generalized linear models. There are now several computer packages available which perform logistic regression for unmatched case–control studies, Poisson regression for cohort studies and conditional logistic regression for matched case–control studies (*see* **Software, Biostatistical**).

The advent of generalized linear models has also been influential in changing the approach to descriptive epidemiology [18]. Poisson regression has proved useful in providing simple estimates of changes in disease rates over time (see, for example, Lee et al. [29]). Recent methodologic work has included **Bayesian methods** for smoothing changes over time or area [3], the latter having a specific application in cancer mapping (*see* **Geographic Epidemiology**) [10], techniques for identifying clusters (*see* **Clustering**) [4] and methods for prediction of future incidence or mortality [24].

Likely Future Developments

Problems of measurement error are of particular interest because of the perceived need to ascertain the effect of dietary factors (*see* **Nutritional Exposure Measures**) on cancer risk. This is especially relevant to the ongoing series of cohort studies of diet and cancer [37]. In cancer epidemiology there is also interest in other variables which are often impossible to measure definitively, such as history of exposure to electromagnetic fields. In this case, approaches involving measurement by several different methods are likely to be productive [25].

The growth of **genetic epidemiology** [1] raises particular methodologic problems. These are likely to stimulate considerable statistical research in the future.

Another area of particular interest at the moment is in descriptive epidemiology, in terms of analysis and prediction of incidence and mortality rates. Because of recent changes in therapy or the introduction of preventive or **screening** policies, public health departments are particularly interested in modeling changes in rates, contemporaneously with, or following upon, changes in policy. Techniques of interest include linear modeling [24], the use of **excess mortality** models [27], and Bayesian prediction [10].

Methods for Cancer Clinical Trials

Definitions and Historical Development

Whether for cancer or any other complaint, the basic principle of a clinical trial is the same: a comparative study in which one group (the **control** group) receives a traditional regime or a placebo and another group receives a new regime or treatment (study group). There are, however, methodologic features which tend to dominate in cancer clinical trials. First, time to a given event – for example, death or recurrence – is usually the primary outcome variable of a trial in cancer treatment, rather than a dichotomous "cure" result or a posttreatment **cross-sectional** measure of a continuous attribute. Secondly, the nature of the disease, in particular the tendency of cancers at certain anatomical sites to recur after a period of remission, renders long-term success of therapy relatively difficult to achieve. Thus, full evaluation of the effect of the new therapy may involve prolonged follow-up. Another corollary of this difficulty in achieving long-term success is that fairly small differences in prognosis may be considered worth pursuing.

The seminal papers in the field are therefore methodologic papers on **survival analysis** [11, 35] and discursive works on the design and analysis of survival studies [34, 35]. Both have proved very useful in clinical trials both within and outside the discipline of oncology. The two survival analysis methods which have gone into widespread use are the **logrank test** [33], which is another adaptation of the Mantel–Haenszel method, and **Cox regression** [11]. The latter is a method which assumes a constant proportional effect (of treatment, prognostic feature, and so on) on the instantaneous hazard of death or recurrence, but assumes no distributional form for the survival probability (*see* **Proportional Hazards, Overview**).

Following from the work on the methods of survival analysis, considerable effort has gone into the calculation of **sample sizes required for clinical trials** in general and cancer trials in particular; see, for example, Freedman [19].

Current Approaches and Ongoing Problems

Due to the necessity of establishing treatment effects which may be small in absolute terms, the necessity for large clinical trials has long been appreciated in oncology [34]. For this reason, **multicenter trials** and overviews of separate clinical trials are a regular feature of cancer treatment research (*see* **Meta-analysis of Clinical Trials**) [14, 32].

Another factor in cancer therapeutic research is the need for timely results. Chemotherapies often have toxic effects, and patients should not be subjected to potentially harmful experimental treatments for any

longer than is absolutely necessary. On the other hand, the wider patient population is anxious for effective new therapies to be introduced as soon as possible. Consequently, a considerable amount of research on stopping rules has been conducted (*see* **Sequential Analysis**) [44].

In conjunction with developments in cancer prevention, the methodology of disease prevention trials is an ongoing issue. In a trial of a preventive measure, the ultimate aim is to assess the effect of the measure on future disease incidence or mortality in currently healthy individuals. Thus a small number of disease/death events will occur in a large number of subjects. For this reason, research has focused on methods of reducing the numbers of patients required, shortening the observation period and simplifying the organizational aspects. Strategies include use of **surrogate endpoints** for reducing the size and period of the trial [36], and cluster randomization (*see* **Group-randomization Designs**) to reduce complexity [21, 43].

Likely Future Developments

Bayesian methods are becoming more common in trial design and analysis [41], with particular progress in terms of monitoring (*see* **Data and Safety Monitoring**) and early stopping. This trend is likely to continue.

Pressure on resources combined with the need to assess preventive measures give an impetus to research on efficient designs of preventive measures. In particular, surrogate outcomes are likely to be more frequently adopted in trials of cancer screening regimens [13]. Another potentially useful strategy is the **factorial design** in which more than one intervention is assessed in the same trial [5].

Tumor Modeling

Definitions and Historical Development

Traditionally, there has been a need to devise mathematical models of tumor initiation and development, first to understand better epidemiologic observations of exposure and subsequent disease and, secondly, to model realistically the results of animal experiments in carcinogenesis and tumor promotion (*see* **Tumor Growth**; **Tumor Incidence Experiments**). The seminal works on the subject are the multistate models of Armitage & Doll [2] and Day & Brown [12] (*see* **Multistage Carcinogenesis Models**). Mathematical forms for tumor development models include **exponential, Weibull**, and Gompertz (*see* **Parametric Models in Survival Analysis**) models. For a review, see Gart et al.[22, Chapter 6].

Current Approaches and Ongoing Problems

In recent years, a major aim of tumor modeling has been to describe and estimate the parameters of tumor development in humans and its arrest by early detection and treatment. This is necessary to understand how programs of screening for cancers have worked in the past and for ongoing monitoring of such programs in the future. Although the ultimate aim of screening is to reduce mortality from the disease in question, it is desirable to have early indicators of whether this is likely to be achieved or whether changes are necessary to the screening regime before results on mortality are available.

Although the methods developed are applicable to screening for a variety of diseases, their primary application has been in screening for breast cancer and cervical cancer [42], cancers for which the effect of screening has been demonstrated, researched and documented over a long period of time. With recent interest in screening for colorectal cancer, modeling of the cancer process at this anatomical site is becoming a target for research [28].

Probably the most commonly used model is the **Markov chain** with discrete state space [42]. Such a model divides the development of the tumor into stages – for example, no disease, lymph node negative, and lymph node positive – with exponentially distributed times spent within the stages. These have been used to predict the effect of different screening regimes for breast cancer [31].

One reason for use of the Markov chain model is the mathematical tractability of the exponential distribution of time spent in each state which is implicit in the Markov model. It is possible, however, to use other distributions for time spent in states. Other models used in the past include the Weibull [17] and **nonparametric** models [9].

Likely Future Developments

Computer programming and estimation from complex biologic models are likely to become easier with the development of theory, practice, and

software in Gibbs sampling (*see* **Markov Chain Monte Carlo**) [23]. Gibbs sampling has already been used to estimate screening **sensitivity** and the average time spent in the preclinical screen-detectable period for colorectal cancer (*see* **Screening, Sojourn Time**) [28].

With the possibility of chemoprevention of various cancers [5], a new application for mathematical models is likely to be the modeling of the biologic processes underlying the chemoprevention strategies. Another possible new application is in modeling of early detection in circumstances in which a randomized trial of screening is not considered appropriate, such as in screening of **gene** carriers for breast cancer.

References

[1] Andrieu, N. & Demenais, F. (1994). Role of genetic and reproductive factors in breast cancer, *Genetic Epidemiology* **11**, 285.

[2] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

[3] Bernardinelli, L. & Montomoli, C. (1992). Empirical Bayes versus fully Bayes analysis of geographical variation in disease risk, *Statistics in Medicine* **11**, 983–1007.

[4] Besag, J. & Newell, J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A* **154**, 143–155.

[5] Bonanni, B., Guerrieri-Gonzaga, A., Rotmensz, N., Torrisi, R., Pigatto, F., Cazzaniga, M., Mora, S., Diani, S., Robertson, C. & Decensi, A. (2000). Hormonal therapy and chemoprevention. *Breast Journal* **6**, 317–323.

[6] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. I: *The Analysis of Case–Control Studies*. International Agency for Research on Cancer, Lyon.

[7] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. I: *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[8] Bross, I. (1954). Misclassification in 2×2 tables, *Biometrics* **10**, 478–486.

[9] Chen, J.S. & Prorok, P.C. (1983). Lead time estimation in a controlled screening program, *American Journal of Epidemiology* **118**, 740–751.

[10] Clayton, D. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, in *Geographical and Environmental Epidemiology: Methods for Small Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. Oxford University Press, Oxford.

[11] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **30**, 187–220.

[12] Day, N.E. & Brown, C.C. (1980). Multistage models and primary prevention of cancer, *Journal of the National Cancer Institute* **64**, 977–989.

[13] Day, N.E. & Duffy, S.W. (1996). Trial design based on surrogate end points – application to comparison of different breast screening frequencies, *Journal of the Royal Statistical Society, Series A* **159**, 49–60.

[14] Early Breast Cancer Trialists' Co-operative Group (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy, *Lancet* **339**, 1–15.

[15] Elton, R.A. & Duffy, S.W. (1983). Correcting for the effect of misclassification bias in a case–control study using data from two different questionnaires, *Biometrics* **39**, 659–665.

[16] Esteve, J., Benhamou, E. & Raymond, L. (1994). *Statistical Methods in Cancer Research*, Vol. IV: *Descriptive Epidemiology*. International Agency for Research on Cancer, Lyon.

[17] Esteve, J., Parker, L., Roy, P., Herrmann, F., Duffy, S., Frappaz, D., Lasset, C., Hill, C., Sancho-Garnier, H. Michaelis, J. & Philip, T. (1995). Is neuroblastoma screening evaluation needed and feasible?, *British Journal of Cancer* **71**, 1125–1131.

[18] Flanders, W.D. (1985). A new variance estimator for the Mantel–Haenszel odds ratio, *Biometrics* **41**, 637–642.

[19] Freedman, L.S. (1982). Tables of the number of patients required in clinical trials using the logrank test, *Statistics in Medicine* **1**, 121–129.

[20] Gail, M.H. (1991). A bibliography and comments on the use of statistical models in epidemiology in the 1980s, *Statistics in Medicine* **10**, 1819–1885.

[21] Gambia Hepatitis Study Group (1987). The Gambia Hepatitis Intervention Study, *Cancer Research* **47**, 5782–5787.

[22] Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research*, Vol. III: *The Design and Analysis of Long-Term Animal Experiments*. International Agency for Research on Cancer, Lyon.

[23] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[24] Hakulinen, T. & Dyba, T. (1994). Precision of incidence predictions based on Poisson distributed observations, *Statistics in Medicine* **13**, 1513–1523.

[25] Kaaks, R., Riboli, E., Esteve, J., van Kappel, A. & van Staveren, W. (1994). Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equations models, *Statistics in Medicine* **13**, 127–142.

[26] Lane-Claypon, J.E. (1926). *A Further Report on Cancer of the Breast, with Special Reference to its Associated Antecedent Conditions*. Ministry of Health, London.

[27] Larsson, L.G., Nyström, L., Wall, S., Rutqvist, L.E., Andersson, I., Bjurstam, N., Fagerberg, G., Frisell, J. & Tabar, L. (1996). The Swedish randomised mammography trials: analysis of their effect on the breast cancer related excess mortality, *Journal of Medical Screening* **3**, 129–132.

[28] Launoy, G., Smith, T.C., Duffy, S.W. & Bouvier, V. (1997). Colorectal cancer mass-screening: estimation of faecal occult blood test sensitivity taking into account cancer mean sojourn time, *International Journal of Cancer*, **73**, 220–224.

[29] Lee, H.P., Day, N.E. & Shanmugaratnam, K. (1988). *Trends in Cancer Incidence in Singapore 1968–82*. International Agency for Research on Cancer, Lyon.

[30] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[31] Organizing Committee and Collaborators, Falun Meeting (1996). Breast-cancer screening with mammography in women aged 40–49 years, *International Journal of Cancer* **68**, 693–699.

[32] Parmar, M.K.B., Spiegelhalter, D.J. & Freedman, L.S. (1994). The CHART trials: Bayesian design and monitoring in practice, *Statistics in Medicine* **13**, 1297–1312.

[33] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society, Series A* **135**, 185–206.

[34] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient II. Analysis and examples, *British Journal of Cancer* **35**, 1–39.

[35] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient I. Introduction and design, *British Journal of Cancer* **34**, 585–612.

[36] Prentice, R.L. (1989). Surrogate end points in clinical trials: definition and operating criteria, *Statistics in Medicine* **8**, 431–440.

[37] Riboli, E. (1992). Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC), *Annals of Oncology* **3**, 783–791.

[38] Richardson, S. & Gilks, W.R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models, *American Journal of Epidemiology* **138**, 430–442.

[39] Rosner, B., Willett, W. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error, *Statistics in Medicine* **8**, 1051–1069.

[40] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[41] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society, Series A* **157**, 357–387.

[42] Stevenson, C.E. (1995). Statistical models for cancer screening, *Statistical Methods in Medical Research* **4**, 18–32.

[43] Tabar, L., Fagerberg, G., Duffy, S.W., Day, N.E., Gad, A. & Gröntoft, O. (1992). Update of the Swedish two-county program of mammographic screening for breast cancer, *Radiologic Clinics of North America* **30**, 187–210.

[44] Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*. Ellis Horwood, Chichester.

STEPHEN W. DUFFY

Operations Research, Simulation

Introduction

Simulation is a versatile problem-solving methodology that involves the abstraction of a real-life system into a symbolic model format and provides an alternative to a purely mathematical analytical solution. A **Monte Carlo** simulation generates random number values (in a process similar to spinning a wheel of fortune), which can be interpreted as actual real-life circumstances or discrete events (*see* **Pseudo-random Number Generator**). Using a computer, a large number of possible combinations of circumstances can be generated to replicate the uncertainties in a real-life scenario, and their impact on selected outcome variables can be evaluated. One can manipulate and experiment with such a model to produce information about the behavior of the real-life scenario over time.

Simulation models can provide estimates of a system's performance measures, for example, the length of a queue in a system with customers waiting for service, the time spent by those customers in the queue, as well as the total time spent in the system. Simulation techniques can also be used to provide financial estimates in the area of investment appraisal, for example, the overall net present value (NPV) for a capital-budgeting investment decision; the likelihood of competitors entering a market and the effect of those competitors on market share. In addition to these performance estimates, simulation models can provide a vehicle to evaluate the effect of changes to a system's input parameters or changes to various operating strategies in a system, for example, changes to the number of staff serving customers, changes in the arrival rates or service rates of customers, changes in future interest rates or discount rates included in an investment decision analysis, and changes in future revenues and costs.

A simulation model must be developed in a way that accurately captures the real-life variation that exists in the system. It must also be systematically verified and validated before it is implemented to provide valuable insight for a decision maker. Such a model becomes a powerful tool for decision makers, often helping to reduce the time needed to evaluate decisions and saving significant costs when good alternatives or solutions are identified. Experimenting with such a model can be less expensive and less disruptive than experimenting with the real-world system. The decision maker can evaluate alternatives in the "safe" environment of a symbolic model.

The Simulation Process

The following stages are part of any simulation study:

(1) Problem Statement
(2) Information Gathering and Data Preparation
(3) Model Conceptualization
(4) Model Implementation
(5) Analysis and Interpretation
(6) Verification and Validation
(7) Implementation and Documentation.

Problem Statement:

Any real-life system tends to be "messy" in terms of its structure and the risk and uncertainty of the system parameters. At the beginning of any simulation study, one must first understand the problems or opportunities that exist in the system. What are the user's stated and unstated goals? Is there a gap between the current situation and the desired one? What information do we need and what criteria will be used to evaluate the decisions?

This first stage is accomplished when we have a complete description of the problem scenario including a problem statement, objective criteria for evaluation, and an understanding of the data that will be needed to represent the real-life system in a simulation model format.

Information Gathering

In this stage of a simulation study, information about the real-life system definition and operation must be gathered. Accurate data are needed to build the simulation model. What are the relevant data needed to abstract the real-life system? What are the uncertainties that exist in these data? Can we assess probability distributions for these uncertainties and how will this happen? What measures of effectiveness are appropriate to analyze these data?

The output from this stage is an understanding of all relevant data and the key factors needed to build a simulation model that accurately represents the real-life scenario.

Model Conceptualization

The art or craft of model building is the abstraction of the real-life scenario into a symbolic model of variables and relationships. This art requires a simplification of the real-life situation that both captures sufficient detail to be acceptable to decision makers yet does not become prohibitively complex. A balance must be struck between keeping the model understandable and representing a reasonable abstraction of the real-life complexity. One big advantage of simulation modeling over analytical models is the number and scope of complexities that can be included. Mathematical models used in optimization are often very limited by their assumptions and therefore applicable to only very simple circumstances. Simulation models can capture much greater complexity in most situations (*see* **Model, Choice of**).

For example, if the simulation model is built to capture a process flow, for instance, a hospital emergency room or an obstetrics–gynecology clinic, one can capture the real-life routing logic that represents various types of arrivals to the system as well as various service patterns and interventions that might occur in the system. If the model is built to capture a capital-budgeting decision, for example, a 10-year cash flow projection for a new ambulatory care facility, one can capture the relevant resource costs, utilization factors, and revenues over a defined window of time. In either case, this stage requires the conceptualization of performance measures that will be used to evaluate effectiveness as well as identification of the key relationships among the inputs and the performance measures.

Model Implementation

In this stage of a simulation study, the conceptual model is translated into the requirements of a chosen computer software package. Although general-purpose languages can be used to program simulation logic, special purpose simulation packages are available that are flexible, inexpensive, and user friendly. The advantages of these packages include the ability to represent **random variables** with probability distributions, the ability to control iterations of the model, and the automatic generation of output statistics, risk profiles, and **sensitivity** and scenario results. Simulation packages are available in several categories. For the simulation of discrete event processes, software such as GPSS/H, Simscript, ARENA, and Process Model are widely used [2, 4–6]. For static simulation models such as the simulation of financial spreadsheet projections, @RISK and Crystal Ball products are available [3, 7].

Analysis and Interpretation

The execution of a simulation model using a computer is designed to replicate all of the combinations of variation that exist in the input parameters of a real-life scenario. The simulation model will produce statistical estimates or probability distributions for all of the outcome variables, based on the variation in the multiple input factors (*see* **Estimation**). This is accomplished by repetitively executing the model generating different sets of inputs and saving the output results. For each execution or iteration of the model, each random input variable is sampled from its defined probability distribution. With a set of sampled input values, the model is evaluated in terms of the outcome variables and these results are saved.

For a discrete event process flow model such as a hospital emergency room, the input variables might represent types of patients arriving with different types of demands for care, arrival time intervals, types of available resources, and service times. The outcome variables might be an estimate of the time an arrival occurred, the time that service started, and the time that service is completed. From these data, estimates of the wait time, server idle time, and queue length can be determined.

For a cash flow financial projection model such as a capital-budgeting investment analysis, the input variables might represent the types of costs, the types of revenues, the growth rates or projections for growth rates, the presence of competitors, and market share. The outcome variables might be yearly after-tax cash flows and net present value for the time frame analyzed.

During each model iteration, a new set of random variables is sampled and a new set of outcome variables are saved. Upon a predetermined number of iterations or a predetermined termination point, all of the outcome variables can be summarized in terms of statistics (e.g. **means**, **standard deviations**, minimum and maximum), risk profiles, sensitivity results, and scenarios.

Once the model is producing outcome measures, it is important to determine which input variables have

the greatest impact on these measures. One needs to pay attention to this type of information as the most sensitive inputs should be modeled with the greatest care. Sensitivity analysis allows the analyst to explore how changes to the input variables impact the output measures. Some typical types of sensitivity analyses would explore the choice of a probability distribution for a particular input or the choice of a particular parameter value. If the number of input factors is small, sensitivity may be explored changing one variable at a time or the **interactions** of changes to two or more variables simultaneously.

Verification and Validation

The output from a simulation model must be checked for accuracy (verification) and reasonableness (validity). Completing these requirements can be time consuming but are critical, requiring the analyst to carefully check all of the inputs and relationships to ensure that the process of developing the conceptual model as well as the process of translating the conceptual model into a computer model was done accurately. Initially, one should take steps to ensure that the conceptual model makes sense to a group of participants (conceptual model validity). Once the conceptual model is agreed to, its translation into the computer model should be verified for accuracy. Finally, in order to validate the computer model, the analyst should ask whether the model results resemble the real-life output or expected output (face validity)? Do the results make sense? Are the results consistent with how the participants perceive the system should behave? (*See* **Model Checking**.)

It is essential that a user have confidence that a simulation model is credible before decisions can be made on the basis of the model. One should design the process of verification and validation to ensure that confidence.

Implementation and Documentation

Once the model has been developed, verified, and validated, simulation experiments may be executed, analyzed, and documented. A simulation study is considered successful when its results have been understood, accepted, and acted upon. Documenting all assumptions, conceptual logic, programmed model logic, and data results in a way that is readable and understandable by both an analyst and the user will enable future modification to the model as well as provide a foundation for future simulation modeling.

Conclusion

One creates simulation models in order to provide valuable insight to decision makers by replicating real-life decision scenarios and providing more complete information about the risks and uncertainties in those environments. Once a model has been verified and validated as accurately representing the real-world environment, many "what if" questions about the system can be posed.

Simulation models have been an indispensable tool for decision makers in many different public and private environments. Several excellent models in the area of health care are illustrative. Rossi describes the use of a simulation model to forecast short- and medium-term projections of HIV/AIDS epidemic indicators for evaluating prevention campaigns, alternate health care strategies for AIDS patients, and drug supply needs [11]. This model also estimates the number of intravenous drug users and the number of AIDS cases that are not appropriately tracked by surveillance and monitoring systems. On the website, `www.informs-cs.org`, an annual conference of simulation practitioners publishes proceedings with complete texts of articles relating to a wide variety of applied simulation models. In the Health Care Track for the Winter 2003 Conference, papers on emergency management of critical incident response, mobile examination centers, a geriatric center, and patient care processes in an ambulatory care center, to name a few, present applied simulation models that enhance decision making in health care settings [1, 8–10]. Past conference papers are also available demonstrating the widespread application of simulation methods to the health care environment.

References

[1] Brady, T.F. (1997–2003). Emergency management: capability analysis of critical incident response, in *Winter 2003 Simulation Conference Proceedings*, `www.informs-cs.org`

[2] For information about ARENA simulation software, see `www.arenasimulation.com`, 2004.

[3] For information about Crystal Ball simulation software, see `www.crystalball.com`, 2004.

[4] For information about GPSS/H simulation software, see www.wolverinesoftware.com/products.htm, 2004.

[5] For information about Process Model simulation software, see www.processmodel.com, 2004.

[6] For information about Simscript simulation software, see www.caciasl.com, 2004.

[7] For information about @RISK simulation software, see www.Palisade.com, 2004.

[8] Martin, E., Gronhaug, R. & Haugene, K. (2003). Proposals to reduce over-crowding, lengthy stays and improving patient care: study of the geriatric department in Norway's largest hospital, in *Winter 2003 Simulation Conference Proceedings*, www.informs-cs.org

[9] Morrison, B.P. & Bird, B.C. (2003). A methodology for modeling front office and patient care processes in ambulatory health care, in *Winter 2003 Simulation Conference Proceedings*, www.informs-cs.org

[10] Osidach, V.Z. & Fu, M.C. (2003). Computer simulation of a mobile examination center, in *Winter 2003 Simulation Conference Proceedings*, www.informs-cs.org

[11] Rossi, C. & Schinaia, G. (1998). The Mover-Stayer model for the HIV/AIDS epidemic in action, *INTERFACES* **28**(3), 127–143.

PHOEBE D. SHARKEY

Operations Research

Operations research methods have been applied widely to improve efficiency and effectiveness in health service delivery, as well as to assist managers and policy-makers in planning and implementation. The conceptual framework used in operations research to assist managers in solving their decision problems has several key attributes:

1. There is an identified decision-maker who is expected to use the results from an operations research analysis.
2. Essential elements of the manager's decision or problem can be mathematically modeled, including constraints on alternatives or options.
3. Measurable indicators of the decision-maker's goals or preferences can be applied to identify optimal solutions.

The focus on decision-making, measurable objectives, constrained decision-making, and opportunity costs are common to most operations research applications. With the growth in high-speed personal computing, plus user-friendly software, many operations research models can now be used directly by managers, clinicians, and their staffs.

Operations research methods have been applied in several areas of health care:

1. The efficient allocation of resources (infrastructure) to meet health care needs of patients or a defined population.
2. Structuring the care processes to ensure that the flow of patients and resources is efficient.
3. Information systems support and patient classification systems to assist managers in understanding the relationship of patient characteristics (e.g. diagnosis, service types, care setting) to resource requirements and costs.

These tools have applications in the design and management of large managed care organizations and integrated delivery systems, as well as in the management of physician group practices and community health center clinics (*see* **Health Services Organization in the US**). A significant barrier to the application of operations research methods has been the timely availability of relevant data at reasonable cost. With advances in health management information systems and their application in hospitals and physician offices, this is becoming less of a barrier than ever before. Advances in medical informatics are making it possible to integrate complex patient care and management data (*see* **Administrative Databases**) in support of clinical and management decision-making.

Operations research models to support decision-making can be categorized in several ways. Models may be either deterministic or stochastic, depending on the importance of random variability in the problem and its solution (*see* **Model, Choice of**). Making a decision regarding the numbers of hospital beds needed in a community may use a deterministic model to predict demand and market characteristics, while an appropriate model to schedule hospital admissions and operating rooms optimally would need to incorporate the effects of random arrivals of emergency admissions and variations in operating time from patient to patient.

Another classification of operations research models concerns whether the model is designed to provide optimal solutions (*see* **Optimization and Nonlinear Equations**) or has a more limited capability to predict and compare the outcomes from decision alternatives or trends (e.g. a **simulation** model). One may want a decision model to determine the optimal number of operating rooms on the basis of a known future demand for surgery. However, a model would be needed to estimate future demand for inpatient vs. outpatient surgery on the basis of trends in population and market characteristics, and in surgical technologies. Since there will be uncertainty about future demand, the predictive model might be used to estimate high, average, and low future scenarios. These estimates would be used in the optimization model to determine what impact, if any, uncertainty of future demand would have on the optimal number of operating rooms. Analysis of the sensitivity (*see* **Sensitivity Analysis**) of optimal solutions due to uncertainty is a standard part of the application of operations research methods. Alternatively, one might not seek an optimal decision but employ simulation modeling to estimate the numbers of operating rooms that would be needed under alternative demand projections. Simulation modeling is frequently easier to learn and to apply than many optimization models, and may be appropriate when the assumptions underlying the optimization model cannot be satisfied.

Operations research models may operate over different time horizons in different decision-making environments. Real-time decision-making is needed when scheduling elective hospital admissions or deciding when specific pharmaceuticals should be reordered to avoid being out of stock when a prescription needs to be filled. These models are generally embedded in computer information systems so they can be updated continually (*see* **Database Systems**). This contrasts with models used in strategic and capital investment planning, which may be updated and used annually or less frequently.

What distinguishes operations research from related fields of biostatistics and economics is the focus on a decision-maker, constraints on alternative choices, the emphasis on tools that can provide the decision-maker with optimal solutions, and the explicit consideration of opportunity cost. However, the methods used by operations researchers are drawn from statistics, economics, computer science, and other areas of applied mathematics and behavioral sciences. Models used to find optimal solutions include **linear programming**, **decision theory**, and a range of nonlinear optimization methods (e.g. dynamic programming, quadratic programming). Optimization models may incorporate stochastic models or stochastic models may be used independently of optimization to predict system performance under alternative assumptions. Some frequently used stochastic models are queuing and Markov models (*see* **Markov Processes**), inventory control, and quality control models.

Health workforce models are used to determine the efficient allocation of health care personnel for meeting the health care needs of patients or a defined population. Such models can be used to predict future physician requirements, or to determine optimal staffing in individual health care facilities.

Decision-making occurs at all levels of the US health care system, ranging from Congress and corporate boards to physician offices and pharmacies. Operations research provides methods that can support decision-making at all these levels. The **Analytic Hierarchy Process** is a method for strategic planning that uses hierarchical methods to elicit utilities and preferences. These methods have been used by corporate boards of directors in strategic decision-making. Physicians and other health care providers need to make decisions regarding the best care for individual patients. Decision analysis models allow the provider to compare alternative treatments in terms of expected outcomes and costs (e.g. medical management vs. surgery). Bayesian decision analysis models can help to quantify the level of uncertainty and its relationship to individual patient risk factors (*see* **Bayesian Methods**; **Bayesian Decision Models in Health Care**).

One of the most widely used operations research optimization technique is linear programming. This technique has been applied extensively in industrial management, but less so in health care management. As health care is becoming a more highly organized and vertically integrated service industry, applications of linear programming are increasing.

Operations research has contributed to the development of patient classification methodologies that can be used to predict patient care resource requirements and costs. **Diagnosis Related Groups (DRG)** is a methodology using diagnoses, surgical procedures, and patient characteristics for estimating the cost of inpatient care. DRGs are widely used as the basis for paying for inpatient hospital care, and also as a measurement of a hospital's "output" that can be used with operations research techniques.

All of the applications and techniques of operations research share a common focus on the needs of the decision-maker for making better decisions. The complexity inherent in the provision of health care services has limited the easy translation of operations research applications from other services industries. Where successful health care applications have been developed, their impact has been substantial. DRGs, physician requirements modeling, and applications of decision analysis are probably the most notable examples of widely used and highly successful operations research applications.

DONALD STEINWACHS

Ophthalmology

Ophthalmology is a branch of medical science concerned with the structure, functions, and diseases of the eye; and *optometry* is an art or occupation consisting of the examination of the eye for defects and faults of refraction, and prescription of correctional lenses and exercises which does not include the use of drugs or surgery. The eye has a wide variety of structures in which both physical and functional defects can have serious consequences regarding the ability to perform in everyday life, and in particular in occupations and professions.

The history of ophthalmology dates from ancient times and documents many issues pertinent to modern ophthalmology. While many of the technical details involved with today's surgical procedures differ from those of the past, many basic strategies for combating eye disease (such as "couching cataracts" which dislodges an opacified lens, or using a magnet to remove foreign ocular bodies) have their roots in antiquity. Indeed, some of the diseases described thousands of years ago, such as trichiasis (inverted eyelashes rubbing against the eye), diplopia (double vision), and night blindness, still constitute significant ophthalmologic problems today.

Johannes Kepler (1571–1631), pioneered the physics of vision and refraction, and shifted the perception of the "organ of vision" from the lens to the retina. During the seventeenth and eighteenth centuries, cataracts were known to occur in the lens, Daviel's extraction of lenses was used in place of the older "couching of cataracts", and spectacles, used since the thirteenth century, were prescribed for patients following lens removal. The nineteenth century brought new advances in anatomy and physiology, including early attempts at corneal transplants, a description of "diabetic retinopathy" (a condition which currently is one of the leading causes of blindness in the Western world), and the invention of the ophthalmoscope, enabling observation of the inner eye.

The twentieth and twenty-first centuries have brought further advances in technology (e.g. tonometers to measure intraocular pressure, lasers to treat a variety of retinal abnormalities, and echography or ultrasound to aid in diagnosis and management), and the application of modern scientific knowledge: the agent of trachoma, a conjunctival infection, was discovered in China; strabismus surgery to correct misaligned eyes increasingly addressed functional, rather than cosmetic, issues; and **genes** associated with certain eye disorders have been identified. Despite these advances, eye disorders are still of major concern.

Types of Data Collected

There are many types of data from ophthalmic research: some are direct measures of visual function and others are more descriptive of certain characteristics associated with visual conditions. Data related to visual function can be related to *central vision* and/or *peripheral vision*. Central vision is the sharp, clear vision that most people use when looking directly at an item or object, such as reading or driving, while peripheral vision is the progressively less clear vision found at further distances from the point of fixation. Tests of central vision include tests to measure visual acuity, contrast sensitivity, reading speed, and others.

Peripheral vision is often quantified and described using perimetry tests such as with visual field analyzers. Tests of either central or peripheral visual function depend on the ability of the person being tested to focus the test object on to the retina and for that image to be correctly interpreted by the brain.

Refractive error is a measure of the amount of optical correction needed to focus an image precisely on the retina. The measure is a combination of three components (sphere, cylinder, and axis) that are often summarized into a spherical equivalent measurement: sphere $+\frac{1}{2}$ cylinder. Subjective refraction is performed by presenting various lenses to an individual and asking which makes the vision clearer. Retinoscopy is an objective measurement of the refractive error that uses a streak of light to determine the needed refractive correction. When performed by a skilled examiner, it is extremely accurate, but the measurement of refraction can often be more generally measured subjectively. Combinations of objective and subjective measurements are also possible. Refractive error measurements are usually performed before further visual function testing and corrective lenses matching the refractive error can be worn by the person during the later tests of visual function to determine, for example, the "best corrected visual acuity" or the "best corrected reading

speed". The change in refractive error over time is an important issue in any surgery involving the cornea, but there has been little development in the analysis of refractive error measurements.

Visual acuity (VA) is a common measure of visual function and generally consists of a "Snellen fraction" of two numbers based on a measurement taken at 6 m (20 ft) from the testing object. The first number indicates the distance separating the test object from the test subject; the second indicates the distance at which the object subtends an angle of 5 minutes. The largest symbol used in the US subtends an angle of 5 minutes at a distance of 200 ft (60 m). An individual being tested is asked to read the smallest symbol on the chart that they can clearly see. Therefore, if a person is 6 m from the visual acuity chart and they can clearly read the line on the chart that subtends the desired angle at 6 m, then the visual acuity of that person would be recorded as 6/6. If they are unable to read the letters on that line but are able to read the largest symbol, their vision would be recorded as 6/60. If the individual is unable to recognize the largest test symbol, the distance at which he recognizes it is recorded (i.e. if it was seen at 3 m the vision would be recorded as 3/60). Whenever an individual is unable to recognize any symbol, then VA is recorded as one of the following measurements (in order of increasing poor visual acuity): distance at which he can count fingers (CF), distance at which hand motion can be detected (HM), light projection ability (being able to project the direction from which light is entering the eye from a small penlight), light perception ability (LP), or no light perception ability (NLP). The **World Health Organization** defines blindness as a VA less than 3/60, and in the US blindness is legally defined as a VA less than 20/200.

A variety of methods have been suggested for computing the **mean** of a set of VA measurements, all of which are expressed in fractional (Snellen) notation: (i) treat the VA as a fraction, convert to decimals, compute the mean, then express the result as a fraction with a numerator of 6 m (or 20 ft); (ii) if all VAs are all measurable at the same distance (or convertible to $6/X$ or $20/X$), then compute the mean of denominators and report the average VA as a fraction with 6 (or 20) in the numerator; and (iii) convert the VA fraction to decimals, average the logarithms of the decimals, take the antilogarithm of the resulting mean, and report the average VA as a fraction with 6 (or 20) in the numerator. This latter method of calculating the mean VA is most consistent with the geometric progression of VA used in the newer standardized VA charts recommended for visual research. The logarithm of the Snellen fraction is approximately the logarithm of the minimal angle of resolution (logMAR) and the new charts provide a step of 0.1 logMAR units for each higher line on the chart. An additional feature of the newer charts is the uniformity in the number of letters on each line of the chart. Earlier visual acuity charts had more smaller letters on lower lines of the charts and fewer, larger letters on higher lines, while the newer charts have a standard five letters per line. Hence, the precision is constant over all levels of measured acuity. Scores are computed from 0.1 logMAR units for each complete line correctly read plus 0.02 units for each additional letter.

Calculating a mean of a set of VA values with both chart-measurable (Snellen fraction) and nonchart measurable (CF, HM, LP, and NLP) visions is more difficult and less standardized because there is no universally agreed upon numerical equivalent for the nonchart measurable (coded) visions. Arbitrary Snellen fractions can be assigned for these coded values; however, it is usually preferable to report the **median** VA or describe the data categorically. To compute the change of vision among VA measurements with a mixed continuous and categorical scale, one is forced either to assign scores or describe the change among category percentages.

Contrast threshold, another measure of visual function, reflects the degree of contrast (darkness) needed to distinguish an item from its underlying background. Some disorders such as optic neuritis may not affect VA (which is usually tested under high-contrast conditions) but may affect the amount of contrast needed to perform tasks such as reading. There are varying tests to measure contrast threshold, with the most common being the Pelli–Robson chart that consists of triplets of letters at decreasing contrast levels. The "log contrast sensitivity" is determined by the faintest triplet of letters for which two of the three letters were correctly identified, although newer methods of scoring have been proposed. The data are reported either as log contrast sensitivity which ranges from 0.00 (the darkest segment) to 2.25 (the lightest segment) or as contrast threshold, which is equal to the reciprocal of the antilog of the contrast sensitivity.

A whole spectrum of other measures of visual function exist. For example, special cards of various sizes test reading rates; and measurements of ability to distinguish different colors can be done using the Farnsworth–Munsell 100-hue test, the Ishihara test, or Hardy–Rand–Rittler plates – the latter two contain dots of primary and secondary colors which test whether an individual can distinguish a particular pattern among the colors. There are tests to determine if the individual is able to fuse correctly the images from both eyes into a single image. Goldmann and Humphrey visual field tests determine the location and extent of peripheral vision or the lack of peripheral vision, which is one of the complications of glaucoma. Instruments for measurement of vision-related **quality of life** have recently become popular with the development of instruments such as the Activities of Daily Vision Scale (ADVS), the VF14, the National Eye Institute Visual Function Questionnaire (NEI-VFQ), and the Refractive Status and Vision Profile (RSVP). These instruments measure the impact of eye disease on the functioning in activities of daily life.

Other commonly occurring ocular measurements are derived from external physical examination of the cornea, iris, lens, retina, and other parts of the eye either during clinical examinations or from photographs, or fluorescein angiography. Retinal photographs and angiography aid in documenting the pathology of retinal vasculature, retinal pigment epithelium, and some aspects of choroidal vasculature. These methods require a trained observer to quantify the features of the disease. A common method of evaluating photographs or angiograms for the presence of particular characteristics is to use numerical scales or grading systems which assign scores that reflect the relative severity of a characteristic. Often, the grades are determined by comparing the characteristics with a set of "standard" photographs or angiograms chosen to represent the severity of the characteristic with a numerical score. Two common examples of these methods include: (i) grading of the degree of nuclear opacity of the lens for classifying cataracts using the four-step Lens Opacity Classification System II (LOCS II) or the decimal LOCS III scales via photograph or slit lamp examination; and (ii) grading of diabetic retinopathy from stereoscopic color fundus photographs using the modified Airlie House classification schemes or from fluorescein angiograms as developed for the Early Treatment Diabetic Retinopathy Study.

Unfortunately, time and space limit the description of the types of ophthalmologic data, and although many have been listed or described, many more have remained unmentioned. But more important than the type(s) of data is the fact that the data from the two eyes of an individual may differ. Some diseases or conditions are *bilateral* (affecting both eyes) and other conditions are *unilateral* (affecting only one eye). Furthermore, bilateral conditions may occur at different times or to different extents. Some treatments may be eye-specific, while others are systemic or affect the person and both eyes. All of these features make analyzing ophthalmologic data more challenging to statisticians.

Common Statistical Issues

The field of ophthalmology and vision research provides a rich source for both the application and the development of statistical theory and methodology. It is safe to say that one could find an application of nearly every major statistical method to the wide array of ophthalmologic data discussed above.

Because of the natural pairing of eyes within individuals, however, ophthalmology has provided particular inspiration for the development of methods for paired, correlated data (*see* **Correlated Binary Data**). Interest in providing designs and analyses that "correctly" estimate and/or account for the associations inherent in paired data have both contributed to and benefited from the increased research in methods for analyzing correlated data over the past decade. Ederer's oft-quoted 1973 paper [2] with the title "Shall we count numbers of eyes or number of subjects?" lays the foundation for interest in this area. In this article, Ederer notes that paired measurements are usually correlated, and explains the consequence of this correlation: positively correlated observations provide *less* statistical information than uncorrelated data when data analysis concerns the *average* of paired measurements, but they provide *more* information than uncorrelated data when the analysis concerns the *difference* of the paired measurements. Thus, using an experimental example, if one is investigating two treatments, where one of the two is applied to each patient (and thus both eyes simultaneously), then the **variance** of the mean

treatment difference using data from both eyes of N patients may not be as small as if $2N$ independent patients provided only one eye observation each. Conversely, if each person receives one treatment in their right eye and the other treatment in their left eye, then the variance of the mean treatment difference using both eyes may be smaller than if using one eye of $2N$ patients. Ederer's article does not go into additional detail, but it provides the basis for subsequent work that treats the eye as the fundamental **unit of analysis**. Thus, the correlation between eye measurements becomes a component in the analysis. This correlation may be a **nuisance parameter** that is required to obtain the "correct" **standard errors**, or it may be the parameter of scientific interest.

It is important to note that it is often useful to consider the patient as the fundamental unit of analysis. Obviously, there are a number of ophthalmologic-related outcomes that can only be obtained on the "person level" instead of the "eye level"; for example, quality of life. For the evaluation of some systemic or bilateral treatments for ocular disease (such as gangcyclovir for CMV retinitis), summarization to a person-level outcome is necessary. For some studies, it is impractical or unethical to observe outcomes on both eyes of patients. For example, with some highly experimental surgical procedures, it may be desirable not to treat one eye of a patient in the event that complications occur or the procedure fails, thereby not risking the loss of vision in at least one eye (although the risk of vision loss from surgery may be lower than the risk associated with no treatment). Eye-specific selection criteria for entry in some studies may also result in patients providing only one eye meeting **eligibility criteria**.

Even when observations specific to each eye are collected, it may be easier to conduct analyses on either a subset of the data or to transform the eye-specific responses to person-specific responses and then apply statistical methods that are appropriate for independent data. Obvious subsets include analyses based on only left eyes, right eyes, or one eye chosen at random. These analyses can suffer from lower efficiency and can possibly provide discordant results [4]. Transformations to person-level responses include using only the worse or better eye of patients, the average of left- and right-eye responses, or a "composite response" approach, which is similar to using the worse eye of each patient but creates additional ordinal steps by treating patients with identical values in both eyes as having a severity score one step greater than patients having worse vision in one eye [3]. Thus, for example, a patient with retinopathy scores of "30" in each eye is assigned a composite response of "30/30", while a patient with retinopathy scores of 30 and 10 is classified as "30/<30", one step lower on the composite response scale. Therefore, if each eye were measured on a scale of (10, 20, 30, 40), the corresponding ordinal composite response index would be (10/10, 20/<20, 20/20, 30/<30, 30/30, 40/<40, 40/40).

The primary advantage of these transformation methods is that they allow one to use standard statistical techniques for analysis. However, there are several problems with these approaches. From a **regression** viewpoint, only models relating person-specific **covariates** can be appropriately constructed for person-level outcomes. To relate eye-specific covariates, some **transformation** of the covariates must be obtained (e.g. the average covariate value). This is a clear example of the **error-in-variables** or exposure **misclassification** problem that will result in estimated effects which are biased relative to the true relationship. In addition, the direct clinical interpretation of an estimated relationship can be lost when using these techniques. It has been shown that the relationship of a person-level covariate to a composite response score can be biased by up to 33% relative to the relationship to the original eye score, with the bias a function of the **correlation** between outcomes [3]. The interpretation of a relationship becomes less clear when based on the average of two eye-specific covariates [4], depending on the goal of the analysis.

In cases in which eye-specific outcomes are collected, it is generally better to construct a statistical model or test which incorporates the relationship between eye responses. When data can be assumed to be from a **multivariate normal distribution**, regression models and testing methods date back to the earliest interest in **multivariate analysis**. Two influential papers by Rosner [8, 9] detail regression models for **bivariate normal data**, with specific application to problems in ophthalmology. These models are flexible enough to allow the response of each eye to be modeled as a function of covariates that are specific to the individual (e.g. gender) and covariates that are specific to the eye (e.g. intraocular pressure). Standard **maximum likelihood** methods are used for estimating and testing parameters.

The single correlation parameter in the standard bivariate normal distribution that, in our context, represents the correlation between left and right eye responses, is often referred to as the *intraclass correlation*. While the models of Rosner detail maximum likelihood estimates of this parameter in conjunction with regression modeling of the bivariate expectations, several other methods have been proposed to estimate this parameter.

When the data cannot be assumed to have a multivariate normal distribution, such as with **binary** and categorical data, different strategies need to be used because of the dearth of an overall multivariate distribution for categorical responses with interpretable parameters. If the scientific questions concern only testing an hypothesis, **nonparametric methods** have been developed for correlated data. For constructing the association between covariates (risk factors and treatment assignments) and ocular outcomes, there has been a tremendous interest in bivariate and more general multivariate models for correlated categorical data that are directly applicable and often inspired by ophthalmologic data. In particular, the papers by Rosner [8, 9] also detail models for **binomially distributed** data. In the context of strategies for correlated binary data, these methods can be classified as *conditional* models, since they model the probability of an outcome in an eye conditional on the outcome of the fellow eye in addition to the covariates of interest. The interpretation of the regression parameters from these models can then be viewed as the effect of a covariate after controling for the outcome status in the fellow eye, and can be viewed as extensions of **beta-binomial** methods. Building upon these models, Rosner & Milton [10] consider significance testing (*see* **Hypothesis Testing**) while Donner [1] and others have proposed alternate approaches based upon adjusting standard **chi-square** statistics.

By contrast to the conditional approaches, many other strategies and models for correlated binary data have been suggested. In particular, methods which model the marginal outcome probability, such as the **generalized estimating equations** methods, have been extensively developed and applied to ophthalmologic problems [7]. The interpretation of the regression parameters from these models is different from the conditional approach in that the parameter represents the effect of a covariate *not* adjusting for the outcome in the fellow eye. Because of the variety of methods that have been proposed, several papers [4, 6] have made comparisons among the approaches with specific application to ophthalmology problems. The key results from these papers have highlighted the inadequacy of ignoring the correlation between ocular outcomes, and have emphasized the difference between conditional and marginal approaches.

The above methods are appropriate when investigating factors associated with the prevalence or incidence of disease outcomes from two eyes. Other scientific questions might concern the factors associated with the time to a particular event, such as the time from enrollment in a clinical trial to a visual acuity worse than 20/200. A variety of methods have been developed for this type of "survival" or "failure time" outcome that occurs from paired eyes, including those that extend the popular **proportional hazards** (**Cox regression**) models.

Ophthalmologic data have also provided inspiration for the investigation of a wide range of biostatistical problems beyond regression models for correlated data. A particularly interesting area is the measurement of visual field function throughout the central retina (the "visual field"). Automated perimetry has gradually become more accepted as the **gold standard** for determining field loss. Individuals are asked to fixate on a central point while flashes of light of varying brightness are presented to the individual at varying locations. The patient then presses a button when the flash is seen. Computerized analysis packages and pre-programmed test patterns have made the standardization of visual field tests much more effective, and produce much more reliable and reproducible results. The result of the test is to produce a grid that maps the ability to detect light across the visual field. Several papers [5] have described methods for analyzing this type of data for the purposes of diagnosis. However, detecting changes over time has proven a formidable task because of the high degree of variability within and between tests, particularly in field areas of damage.

Other ophthalmologic problems have motivated biostatistical research. These include, for example, models and methods for time to visual loss, determining numerical scales and schemes for clinical outcome grading, studying the spatial pattern of retinal structures, estimating **agreement** in binocular data, and interim analyses of clinical trials (*see* **Data and Safety Monitoring**) with long response times or correlated outcomes. It is certain that the field will

continue to provide interesting problems and inspire further methodologic developments in biostatistics.

References

[1] Donner, A. (1989). Statistical methods in ophthalmology: an adjusted chi-square approach, *Biometrics* **45**, 605–611.

[2] Ederer, F. (1973). Shall we count numbers of eyes or numbers of subjects?, *Archives of Ophthalmology* **89**, 1–2.

[3] Gange, S.J., Linton, K., Scott, A., DeMets, D. & Klein, R. (1996). A comparison of methods for correlated ordinal measures with ophthalmic applications, *Statistics in Medicine* **14**, 1961–1974.

[4] Glynn, R.J. & Rosner, B. (1992). Accounting for the correlation between fellow eyes in regression analysis, *Archives of Ophthalmology* **110**, 381–387.

[5] Hilton, S., Katz, J. & Zeger, S. (1996). Classifying visual field data, *Statistics in Medicine* **15**, 1349–1364.

[6] Katz, J., Zeger, S. & Liang, K.-Y. (1994). Appropriate statistical methods to account for similarities in binary outcomes between fellow eyes, *Investigative Ophthalmology and Visual Science* **35**, 2461–2465.

[7] Podger, M.J., Hiller, R. & the Framingham Eye Studies Group (1996). Associations of types of lens opacities between and within eyes of individuals: an application of second-order generalized estimating equations, *Statistics in Medicine* **15**, 145–156.

[8] Rosner, B. (1982). Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes, *Biometrics* **38**, 105–114.

[9] Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations, *Biometrics* **40**, 1025–1035.

[10] Rosner, B. & Milton, R.C. (1988). Significance testing for correlated binary outcome data, *Biometrics* **44**, 505–512.

MARTA J. MARSH & STEPHEN J. GANGE

Optimal Design

Informally, an *optimal design* is one which makes best use of the experimental resources available. We need to define what is meant by "best", to see which designs are best and to see if different definitions of best give markedly different designs (*see* **Experimental Design**).

Models and Parameter Estimates

The aim of any experiment is to estimate the parameters in some model as accurately as possible. Thus, it is natural to define best in terms of the **variance** of the parameter estimates for a **regression** model, and in terms of the variance of estimable **contrasts** of treatment effects in an experimental design model. For either situation we need to specify the model for which parameter estimates are required and the *design region*, the region in which it is possible to experiment.

Let the vector of observed responses be denoted by $\mathbf{y}$ and let the model be given by $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, where we assume that the errors in the observations are independent and have a constant variance denoted by σ^2.

In **simple linear regression**, for example, we have $\boldsymbol{\beta}' = (\beta_0, \beta_1)$ and all the entries in the first column of $\mathbf{X}$ would be 1 and the entries in the second column of $\mathbf{X}$ would be the x values at which the responses were observed. Thus, for each observed y we would have $\mathrm{E}(y) = \beta_0 + \beta_1 x$.

In a design model, $\mathbf{X}$ is the usual design matrix. For a completely randomized design to compare t treatments, for example, we have $\mathrm{E}(y_{ij}) = \mu + \tau_i$, $\boldsymbol{\beta}' = (\mu, \tau_1, \tau_2, \ldots, \tau_t)$ and $\mathbf{X}$ has $t + 1$ columns. If the first n_1 responses are those associated with treatment 1, the next n_2 with treatment 2 and so on, then the first column of $\mathbf{X}$ has all entries equal to 1, the second column has the first n_1 entries equal to 1 and the remainder equal to 0, the third column has the first n_1 entries equal to 0, the next n_2 equal to 1 and the remainder equal to 0, and so on. A **simulation** study comparing 15 designs for 12 models is given in [14].

In regression models it is usually the case that the parameter estimates are unique and we can show that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Thus, $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ and the matrix $X'X$ is called the **information matrix**. If we are interested in predicting the response at some point $\mathbf{x}$ in the design region we would have $\hat{y}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}$, with $\mathrm{var}(\hat{y}(\mathbf{x})) = \sigma^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$. Then we say that a design is *optimal* if it minimizes $\mathrm{var}(\hat{y}(\mathbf{x}))$ or some other function of $\mathrm{var}(\hat{\boldsymbol{\beta}})$. A case study comparing the estimates obtained using a 10-point data set with those obtained from smaller, optimally chosen, subsets appears in [12].

Optimality Criteria

The four most frequently used optimality criteria are A, D, E, and G optimality. In all cases the expressions are given in terms of the **eigenvalues**, $\lambda_1, \lambda_2, \ldots, \lambda_p$ of the information matrix $\mathbf{A} = \mathbf{X}'\mathbf{X}$.

A design is said to be *A-optimal* if the sum of the variance of the parameter estimates is minimized. Thus, we seek to minimize $\mathrm{tr}(\mathbf{A}^{-1}) = \sum_i \lambda_i^{-1}$.

A design is said to be *D-optimal* if the generalized variance of the parameter estimates is minimized. Thus, we seek to minimize $\det(\mathbf{A}^{-1}) = \prod_i \lambda_i^{-1}$.

A design is said to be *E-optimal* if it minimizes the variance of the least well-estimated normalized contrast. Thus, we seek to minimize the largest eigenvalue of $\mathbf{A}^{-1}$; that is, we aim to minimize the largest value of λ_i^{-1}.

A design is said to be *G-optimal* if it minimizes the maximum variance of a predicted value over the design region (*see* **Minimax Theory**).

Other optimality criteria may be preferred; see, for instance, [7] and [10].

It is important to realize that the optimum design is dependent on the postulated model. For example, in a simple linear regression $\mathrm{var}(\hat{\beta}_1) = \sigma^2 / \sum_i (x_i - \overline{x})^2$, where $\overline{x} = \sum_i x_i / N$. Then, with a, perhaps coded, design region $-1 \le x \le 1$, the variance of $\hat{\beta}_1$ is minimized by having half the responses at $x = -1$ and half at $x = 1$. If, however, the initial assumption is incorrect and the response is really quadratic, say, then it will be impossible to detect it with this optimum design. This is why various authors recommend that modifications of optimal designs be used in practice (see, for example, [1], [9], and [16]).

Optimality and Block Designs

In design models the estimates of the elements of $\boldsymbol{\beta}$ are not, in general, unique, but some functions

of them, such as the difference of two treatment estimates, are uniquely estimated. Such functions are called *estimable functions* and can be represented by $\mathbf{c}'\boldsymbol{\beta}$. As we are interested in the design setting, in the treatment parameter estimates and not in the block parameter estimates, we need to use a modified information matrix.

We are only interested in *connected* designs; that is, designs in which every pairwise difference of treatment effects is estimable. Assume there are t treatments and b blocks. Let $Y_{ij} = \tau_i + \beta_j + E_{ij}$, $1 \leq i \leq t$, $1 \leq j \leq b$, although some treatments may not appear in some blocks so the corresponding Y_{ij} do not exist. (*see* **Randomized Complete Block Designs**) It should be noted here that we have assumed that the block and treatment effects do not interact. This is the usual assumption in the literature on blocks designs, and if this is not the case then the conclusions that are drawn from the experiment may be invalid. It can be helpful to test for nonadditivity; Read [13] gives a test and a discussion. Let $\mathbf{N}$ be a $t \times b$ incidence matrix in which the (i,j)th position is the number of times that treatment i is in block j. Let $\mathbf{R}$ be a diagonal matrix with the treatment replication numbers on the diagonal and let $\mathbf{K}$ be a matrix with the block sizes on the diagonal. Then the *information matrix* $\mathbf{A}$ is defined by $\mathbf{A} = \mathbf{R} - \mathbf{N}\mathbf{K}^{-1}\mathbf{N}'$. Let $\mathbf{T}$ be the vector of treatment totals and let $\mathbf{B}$ be the vector of block totals. Then the *reduced normal equations* are given by $\mathbf{A}\hat{\boldsymbol{\tau}} = \mathbf{q}$, where $\mathbf{q} = \mathbf{T} - \mathbf{N}\mathbf{K}^{-1}\mathbf{B}$. We now need an expression for $\hat{\boldsymbol{\tau}}$. In any block design, $\mathbf{A}$ is not of full rank and so no inverse exists. However, it can be shown that in a connected design the rank of $\mathbf{A}$ is $v - 1$. Hence, to get an expression for $\hat{\boldsymbol{\tau}}$ we need to calculate a generalized inverse for $\mathbf{A}$, $\boldsymbol{\Omega}$ say. A *generalized inverse* of $\mathbf{A}$ is a matrix $\boldsymbol{\Omega}$ such that $\mathbf{A}\boldsymbol{\Omega}\mathbf{A} = \mathbf{A}$. One way to get $\boldsymbol{\Omega}$ is to calculate the eigenvalues and corresponding **eigenvectors** of $\mathbf{A}$. Suppose that λ_i is an eigenvalue of $\mathbf{A}$ with corresponding eigenvector $\mathbf{z}_i$. These eigenvectors can be normalized so that $\mathbf{z}_i'\mathbf{z}_m = 0$, unless $i = m$, when it equals 1. Then we can write $\mathbf{A}$ in canonical form as $\mathbf{A} = \sum_i \lambda_i \mathbf{z}_i \mathbf{z}_i'$, and a generalized inverse of $\mathbf{A}$ is $\boldsymbol{\Omega} = \sum_i \lambda_i^{-1} \mathbf{z}_i \mathbf{z}_i'$, where this summation is over the nonzero eigenvalues only. Then $\hat{\boldsymbol{\tau}} = \boldsymbol{\Omega}_q$ and the **covariance matrix** of $\hat{\boldsymbol{\tau}}$ is $\sigma^2\boldsymbol{\Omega}$. Note that the variance of estimable functions of the treatment parameters is independent of the particular generalized inverse used. The previous definitions of optimality can now be used with $\boldsymbol{\Omega}$ replacing $\mathbf{A}^{-1}$ throughout.

Efficiency

The final topic that we will discuss is that of *efficiency*. This is the ratio of the optimality value of the optimal design to that of the proposed design. Clearly, the efficiency of the optimal design is 1, and so all other designs have a lower efficiency. The higher the efficiency, the better the design.

There are other efficiency measures that are unique to designed experiments. For blocked experiments the variance of the pairwise differences of treatments (*see* **Paired Comparisons**) is compared with the variance that would be obtained in a completely randomized design with the same replication as the blocked experiment for each of the treatments. In a completely randomized design with equal replication r, say, for each of the treatments, we see that $\text{var}(\hat{\tau}_i - \hat{\tau}_m) = 2\sigma^2/r$. For a block design in which each treatment is replicated r times, the *pairwise efficiency factor* is the ratio of $2\sigma^2/r$ to $\text{var}(\hat{\tau}_i - \hat{\tau}_m)$. The ratio of $2\sigma^2/r$ to the average variance of the treatment differences is the *average efficiency factor*.

For **balanced incomplete block designs**, we get $\boldsymbol{\Omega} = k/\lambda v(\mathbf{I} - \mathbf{J}/v)$ and so each pairwise efficiency factor is $\lambda v/rk$ and hence so is the average efficiency factor. For **partially balanced incomplete block designs** with two associate classes (PBIBD (2)) a similar result can be obtained. In this case, there are two pairwise efficiency factors, one for comparing first associates and one for comparing second associates. The values are given for the designs tabulated by Clatworthy [4], as well as the general expressions for any PBIBD (2).

Bounds on the efficiency factors of pairwise comparisons may be found in [8]. **Algorithms** for constructing designs such as GENDEX [6] and the algorithm given by Paterson et al. [11] look for designs which are close to the upper bound for efficiency.

Other Topics

Optimal designs for bivariate **logistic regression** appear in [7]. Optimal design of LD50 bioassay (*see* **Biological Assay, Overview**) is discussed in [10] and references cited therein. Optimal design of **validation studies** is considered in [15]. Optimal design of **Phase II clinical trials** is discussed in [5]. An introduction to the issues involved in the design of

optimal **crossover trials** is given in [3]. A package to construct optimal designs for a batch reaction catalyzed by a soluble enzyme where the activity of the enzyme decays with time is described in [2].

References

[1] Atkinson, A.C.& Donev, A.N. (1992). *Optimum Experimental Designs*. Oxford University Press, Oxford.

[2] Balcao, V.M.& Malcata, F.X. (1993). STADEERS: a software package for the statistical design of experiments pertaining to the estimation of parameters in rate expressions that describe enzyme-catalyzed processes, *Computer Applications in the Biosciences* **9**, 629–637.

[3] Carriere, K.C. (1994). Crossover designs for clinical trials, *Statistics in Medicine* **13**, 1063–1069.

[4] Clatworthy, W.H. (1973). *Tables of Two-Associate-Class Partially Balanced Designs*, National Bureau of Standards (US) Applied Mathematics Series No. 63. Government Printing Office, Washington.

[5] Ensign, L.G., Gehan, E.A., Kamen, D.S.& Thall, P.F. (1994). An optimal three-stage design for Phase II clinical trials, *Statistics in Medicine* **13**, 1727–1736.

[6] GENDEX (1993). GENDEX: An algorithmic toolkit for designers of experiments, in *Proceedings of Statistics, 73*. University of Wollongong, Australia.

[7] Heise, M.A.& Myers, R.H. (1996). Optimal designs for bivariate logistic regression, *Biometrics* **52**, 613–624.

[8] Jarrett, R. (1989). A review of bounds for the efficiency factor of block designs, *Australian Journal of Statistics* **31**, 118–129.

[9] Krewski, D.& Goddard, M. (1990). Principles of bioassay design, *Drug Information Journal* **24**, 381–394.

[10] Markus, R.A., Frank, J., Groshen, S.& Azen, S.P. (1995). An alternative approach to the optimal design of an LD50 bioassay, *Statistics in Medicine* **14**, 841–852.

[11] Paterson, L.J., Wild, P.& Williams, E.R. (1988). An algorithm to generate designs for variety trials, *Journal of Agricultural Science* **111**, 133–136.

[12] Preston, S.L.& Drusano, G.L. (1996). Nonparametric expectation maximization population modeling of ganciclovir, *Journal of Clinical Pharmacology* **36**, 301–310.

[13] Read, C.B. (1988). Tukey's test for nonadditivity, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz& N.L. Johnson, eds. Wiley, New York, pp. 364–366.

[14] Scharfstein, D.O.& Williams, P.L. (1994). Design of developmental toxicity studies for assessing joint effects of dose and duration, *Risk Analysis* **14**, 1057–1071.

[15] Stram, D.O., Longnecker, M.P., Shames, L., Kolonel, L.N., Wilkens, L.R., Pike, M.C.& Henderson, B.E. (1996). Cost-efficient design of a diet validation study, *American Journal of Epidemiology* **142**, 353–362.

[16] Wong, W.K.& Lachenbruch, P.A. (1996). Designing studies for dose response, *Statistics in Medicine* **15**, 343–359.

D.J. STREET

Optimization and Nonlinear Equations

Optimization means finding that value of $\mathbf{x}$ which maximizes or minimizes a given function $f(\mathbf{x})$. The idea of optimization goes to the heart of statistical methodology, as it is involved in solving statistical problems based on **least squares**, **maximum likelihood**, posterior mode (*see* **Bayesian Methods**), and so on.

A closely related problem is that of solving a nonlinear equation,

$$\mathbf{g}(\mathbf{x}) = \mathbf{0}$$

for $\mathbf{x}$, where $\mathbf{g}$ is a possibly multivariate function. Many algorithms for minimizing $f(\mathbf{x})$ are in fact derived from algorithms for solving $\mathbf{g} = \partial f/\partial \mathbf{x} = \mathbf{0}$, where $\partial f/\partial \mathbf{x}$ is the vector of derivatives of f with respect to the components of $\mathbf{x}$.

Except in linear cases, optimization and equation solving invariably proceed by iteration. Starting from an approximate trial solution, a useful algorithm will gradually refine the working estimate until a predetermined level of precision has been reached. If the functions are smooth, a good algorithm can be expected to converge to a solution when given a sufficiently good starting value.

A good starting value is one of the keys to success. In general, finding a starting value requires heuristics and an analysis of the problem. One strategy for fitting complex statistical models, by maximum likelihood or otherwise, is to progress from the simple to the complex in stages. Fit a series of models of increasing complexity, using the simpler model as a starting value for the more complicated model in each case. Maximum likelihood iterations can often be initialized by using a less efficient moment estimator (*see* **Method of Moments**). In some special cases, such as **generalized linear models**, it is possible to use the data themselves as starting values for the fitted values.

An extremum (maximum or minimum) of f can be either *global* (truly the extreme value of f over its range) or *local* (the extreme value of f in a neighborhood containing the value); see Figure 1. Generally it is the global extremum that we want. (A maximum likelihood estimator, for example, is by definition the global maximum of the likelihood.)

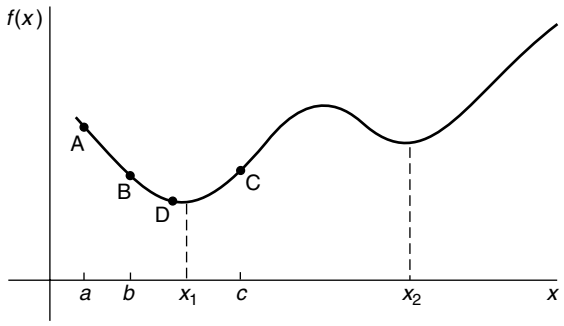

Figure 1 The function $f(x)$ has a local minimum at x_2 and a global minimum at x_1. The points $\mathrm{A} = [a, f(a)]$, $\mathrm{B} = [b, f(b)]$, and $\mathrm{C} = [c, f(c)]$ bracket the global minimum. The next point tried by a golden section search would be D

Unfortunately, distinguishing local extrema from the global extremum is not an easy task. One heuristic is to start the iteration from several widely varying starting points, and to take the most extreme (if they are not equal). If necessary, a large number of starting values can be randomly generated. Another heuristic is to perturb a local extremum slightly to check that the **algorithm** returns to it. Two relatively recent types of algorithms, simulated annealing and genetic algorithms, are often used successfully on problems where there are a large number of closely competing local extrema. Simulated annealing handles multiple minima by introducing a stochastic element into the interaction, which allows it to escape from local extrema by temporarily increasing the objective function. Genetic algorithms handle multiple minima by remembering at each iteration a set of candidate parameter estimates instead of just one.

This article discusses *unconstrained optimization.* Sometimes, however, $\mathbf{x}$ must satisfy one or more constraints. An example is some of the components of $\mathbf{x}$ being known a priori to be positive. In some cases the constraints may be removed by a suitable transformation ($x_i = \mathrm{e}^{z_i}$, for example), or by use of Lagrange multipliers.

One must choose between algorithms which use derivatives and those which do not. In general, methods which use derivatives are more powerful. However, the increase in speed does not always outweigh the extra overheads in computing the derivatives, and it can be a great convenience for the user not to have to program them.

Algorithms are also distinguished by the amount of memory they consume. Storage requirements are typically order N or order N^2, where N is the dimension of $\mathbf{x}$. In most biostatistical applications, N is not usually so large that storage becomes an issue.

If you can calculate first and second derivatives of f, then the well-known *Newton's method* is simple and works well. It is crucially important, though, to check the function value $f(\mathbf{x})$ at each iteration, and to implement some sort of backtracking strategy, to prevent the Newton iteration from diverging to distant parts of the parameter space from a poor starting value. If second derivatives are not available, then *quasi-Newton* methods, of which Fisher's method of scoring is one, can be recommended. General-purpose quasi-Newton algorithms build up a working approximation to the second-derivative matrix from successive values of the first derivative. If computer memory is very critical, then *conjugate gradient* methods make the same assumptions as quasi-Newton methods but require only order N storage [8, Section 10.6]. If even first derivatives are not available, the Nelder–Mead *downhill simplex* algorithm is compact and reasonably robust. However, the slightly more complex *direction-set* methods or Newton methods with finite difference approximations to the derivatives should minimize most smooth differentiable functions, with fewer function evaluations. Whilst all the above comments apply generally, the one-dimensional problem is something of a special case. In one dimension, once one can provide an interval which contains the solution, there exist efficient "low-tech" algorithms robust enough to take on all problems.

A practical introduction to root finding and optimization is given in Chapters 9, 10, and 15 (Sections 15.5 and 15.7) of *Numerical Recipes* [8]. More specialist texts are Dennis & Schnabel [2], Fletcher [3], and Gill et al.[4]. A classic but technical text on solving nonlinear equations is Ortega & Rheinboldt [6]. A survey of available software is given by Moré & Wright [5].

One Dimension

The case where x is one-dimensional is qualitatively simpler than the multidimensional case. This is because a solution can be trapped between bracketing values, which are gradually brought together. A root of $g(x)$ is bracketed in the interval (a, b) if $g(a)$ and $g(b)$ have opposite signs. A minimum of $f(x)$ is bracketed by a triplet of values, $a < b < c$, if $f(b)$ is less than both $f(a)$ and $f(c)$.

The simplest and most robust method for finding a root in a bracketing interval is *bisection*. That is, we evaluate the function g at the midpoint of (a, b) and examine its sign. The midpoint then replaces whichever end point has the same sign. After k iterations, the root is known to lie in an interval of length $(b - a)/2^k$.

The equivalent method for function minimization is the *golden section search*. Given a bracketing triplet of points, the next point to be tried is that which is a fraction 0.38197 of the way from the middle point of the triplet to the farther end point (Figure 1). One then drops whichever of the end points is farthest from the new minimum. The strange choice of step size ensures that at each iteration the middle point is always the same fraction of the way from one end point to the other (the so-called golden ratio). After k iterations, the minimum is bracketed in an interval of length $(c - a)0.61803^k$.

Bisection and golden section search are linear methods, in that the amount of work required increases linearly with the number of significant figures required for x. There are a number of other methods, such as the *secant method*, the *method of false position, Muller's method*, and *Ridder's method*, which are capable of *superlinear convergence*, wherein the number of significant figures liberated by a given amount of computation increases as the algorithm converges. The basic idea is that g should be roughly linear in the vicinity of a root. These methods interpolate a line or a quadratic polynomial through two or three previous points, and use the root of the polynomial as the next iterate. They therefore converge more rapidly than bisection or golden search when the function g is smooth, but can converge slowly when g is not well approximated by a low-order polynomial. They also require modification if they are not to risk throwing the iteration outside the bracketing interval known to contain the root.

It is an advantage to use one of the higher-order interpolating methods when the function g is nearly linear, but to fall back on the bisection or golden search methods when necessary. In that way a rate of convergence at least equal to that of the bisection or golden section methods can be guaranteed, but higher-order convergence can be enjoyed when it is

possible. Brent [1, 8] has published methods which do the necessary bookkeeping to achieve this, and which can be generally recommended for root finding or minimizing in one dimension. Brent's algorithms do not require the derivatives of f or g to be supplied. However, the method for minimizing a function can be easily modified to make use of the derivative when it is available [8].

Newton's Method

The most celebrated of all methods for solving a nonlinear equation is *Newton's method*, also called *Newton–Raphson*. Newton's method is based on the idea of approximating $\mathbf{g}$ with its linear Taylor series expansion about a working value $\mathbf{x}_k$. Let $\mathbf{G}(\mathbf{x})$ be the matrix of partial derivatives of $\mathbf{g}(\mathbf{x})$ with respect to $\mathbf{x}$. Using the root of the linear expansion as the new approximation gives

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{G}(\mathbf{x}_k)^{-1}\mathbf{g}(\mathbf{x}_k)$$

(see Figure 2).

The same algorithm arises for minimizing $f(\mathbf{x})$ by approximating f with its quadratic Taylor series expansion about $\mathbf{x}_k$. In the minimization case, $\mathbf{g}(\mathbf{x})$ is the derivative vector (gradient) of $f(\mathbf{x})$ with respect to $\mathbf{x}$ and the second derivative matrix $\mathbf{G}(\mathbf{x})$ is symmetric. Beware, though, that Newton's method as it stands will converge to a maximum just as easily as to a minimum.

If f is a log **likelihood** function, then $\mathbf{g}$ is the score vector and $-\mathbf{G}$ is the observed **information matrix**. Newton's method for maximizing the likelihood is based on the same quadratic expansion which underlies asymptotic maximum likelihood theory.

Newton's method is powerful and simple to implement. It will converge to a fixed point from any sufficiently close starting value. Moreover, once it starts to home in on a root, the convergence is quadratic. This means that, if the error is ε, the error after one more iteration is of order ε^2. In other words, the number of significant places eventually doubles with each iteration. However, its global convergence properties are poor. If $\mathbf{x}_k$ is unlucky enough to occur near a turning point of $\mathbf{g}$, then the method can easily explode, sending the next estimate far out into the parameter space (Figure 3). In fact, the set of values for which Newton's method does and does not converge can produce a fractal pattern [8].

The problems with Newton's method are: (i) an inability to distinguish maxima from minima; and (ii) poor global convergence properties. Both problems can be solved effectively through a *restricted step* suboptimization [3]. Suppose we want to minimize $f(\mathbf{x})$. A condition for a minimum is that $\mathbf{G}(\mathbf{x})$ be positive definite. We therefore add a diagonal matrix to $\mathbf{G}$ to ensure that it is positive definite:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [\mathbf{G}(\mathbf{x}_k) + \lambda_k \mathbf{I}]^{-1}\mathbf{g}(\mathbf{x}_k).$$

It is always possible to choose λ_k sufficiently large so that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. A simple algorithm then is to choose λ_k just large enough to ensure a descent step. As the iteration converges to a minimum, λ_k can be decreased towards zero so that the algorithm enjoys superlinear convergence. This is the algorithm of choice when derivatives of f are available.

Solving $\mathbf{g}(\mathbf{x}) = 0$, when $\mathbf{g}$ is not the gradient of some objective function f, is slightly more difficult.

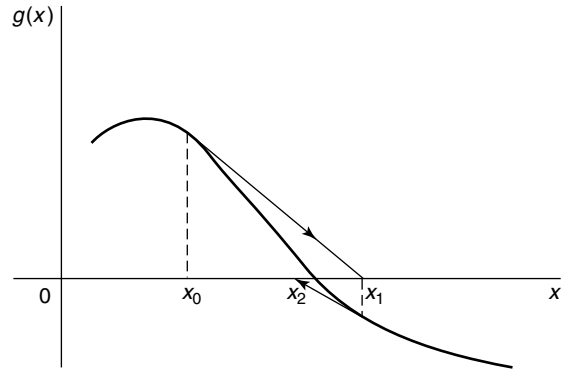

Figure 2 Newton's method converges quadratically from the starting value x_0

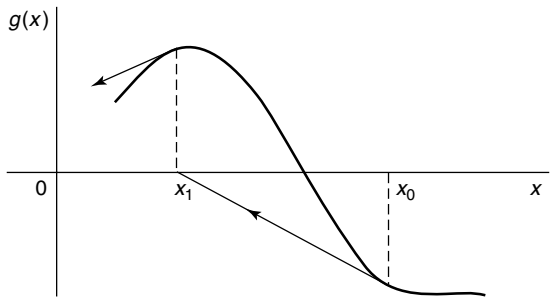

Figure 3 Newton's method diverges from the starting value x_0

One can manufacture a stand-in objective function by defining

$$f(\mathbf{x}) = \mathbf{g}(\mathbf{x})^{\mathrm{T}}\mathbf{g}(\mathbf{x}).$$

Then the root of $\mathbf{g}$ occurs at a minimum of f. Note, however, that $\mathbf{g}$ is not the derivative of f, so that the above restricted step strategy is not available. In this case we can replace the Newton step with the *line search* strategy,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{G}(\mathbf{x}_k)^{-1}\mathbf{g}(\mathbf{x}_k),$$

where $0 < \alpha_k < 1$. It is always possible to choose α_k sufficiently small that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. The line search idea is to implement a one-dimensional suboptimization at each step, minimizing $f(\mathbf{x}_{k+1})$ approximately with respect to α_k.

Both the restricted step and the line search algorithms have global convergence properties. They can be guaranteed to find a local minimum of f and a root of $\mathbf{g}$ if such exist subject only to some standard regularity conditions such as differentiability.

Quasi-Newton Methods

One of the drawbacks of Newton's method is that it requires the analytic derivative $\mathbf{G}$ at each iteration. This is a problem if the derivative is very expensive or difficult to compute. In such cases it may be convenient to iterate according to

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{A}_k^{-1}\mathbf{g}(\mathbf{x}_k),$$

where $\mathbf{A}_k$ is an easily computed approximation to $\mathbf{G}(\mathbf{x}_k)$. For example, in one dimension, the *secant method* approximates the derivative with the difference quotient

$$a_k = \frac{g(x_k) - g(x_{k-1})}{x_k - x_{k-1}}.$$

Such an iteration is called a *quasi-Newton* method. If $\mathbf{A}_k$ is positive definite, as it usually is, an alternative name is *variable metric* method.

One positive advantage to using an approximation in place of $\mathbf{G}$ is that $\mathbf{A}_k$ can be chosen to be positive definite, ensuring that the step will not be attracted to a maximum of f when one wants a minimum. Another advantage is that $\mathbf{A}_k^{-1}\mathbf{g}(\mathbf{x}_k)$ is a descent direction from $\mathbf{x}_k$, allowing the use of line searches.

The best known quasi-Newton method in statistical contexts is Fisher's method of scoring, which is treated in more detail below. Among general purpose quasilikelihood algorithms, the best is probably the *Broydon–Fletcher–Goldfarb–Shanno* (BFGS) algorithm. BFGS builds upon the earlier and similar *Davidon–Fletcher–Powell* algorithm. BFGS starts with a positive-definite matrix approximation to $\mathbf{G}(\mathbf{x}_0)$, usually the identity matrix. At each iteration it makes a minimalist (rank two) modification to $\mathbf{A}_k^{-1}$ to gradually approximate $\mathbf{G}(\mathbf{x}_k)^{-1}$. Both DFP and BFGS are robust algorithms showing superlinear convergence.

Statisticians might fall into the trap of thinking that the final approximation $\mathbf{A}_k^{-1}$ is a good approximation to $\mathbf{G}^{-1}(\mathbf{x}_k)$ at the final estimate. Since $\mathbf{A}_k$ is chosen to approximate $\mathbf{G}(\mathbf{x}_k)$ only in the directions needed for the Newton step, it is useless for the purpose of providing standard errors for the final estimates.

Fisher's Method of Scoring

Of frequent interest to statisticians is the case where $f(\mathbf{x})$ is a log likelihood function and $\mathbf{x}$ is the vector of unknown parameters. Then $\mathbf{g}$ is the score vector and $-\mathbf{G}$ is the observed information matrix. For many models (curved **exponential families** are the major class), the Fisher **information**, $\mathcal{I}(\mathbf{x}) = \mathrm{E}[-\mathbf{G}(\mathbf{x})]$, is much simpler in form than $-\mathbf{G}(\mathbf{x})$ itself. Furthermore, since $\mathcal{I}(\mathbf{x}) = \mathrm{var}[\mathbf{g}(\mathbf{x})]$, $\mathcal{I}(\mathbf{x})$ is positive definite for any parameter value $\mathbf{x}$ for which the statistical model is not degenerate. The quasi-Newton method which replaces $-\mathbf{G}(\mathbf{x})$ with $\mathcal{I}(\mathbf{x})$ is known as *Fisher's method of scoring* [7, Section 5g]. Fisher scoring is linearly convergent, at a rate which depends on the relative difference between observed and expected information [10].

Consider the special case of nonlinear least squares, in which context Fisher scoring has a very long history and is known as the *Gauss–Newton* algorithm. The objective function is

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - \mu(\mathbf{t}_i, \boldsymbol{\beta})]^2,$$

where the y_i are observations and μ is a general function of covariate vectors $\mathbf{t}_i$ and the vector of unknown parameters $\boldsymbol{\beta}$. Write $\mathbf{y}$ for the vector of y_i, $\boldsymbol{\mu}$ for the vector of $\mu(\mathbf{t}_i, \boldsymbol{\beta})$, and $\dot{\boldsymbol{\mu}}$ for the derivative

matrix of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\beta}$. The Fisher scoring iteration becomes

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + (\dot{\boldsymbol{\mu}}^{\mathrm{T}}\dot{\boldsymbol{\mu}})^{-1}\dot{\boldsymbol{\mu}}^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\mu}),$$

where all terms on the right-hand size are evaluated at $\boldsymbol{\beta}_k$. The updated estimate is obtained by adding to $\boldsymbol{\beta}_k$ the coefficients from the multiple regression of the residuals $\mathbf{y} - \boldsymbol{\mu}$ on the derivative matrix $\dot{\boldsymbol{\mu}}$. Gauss–Newton therefore solves the nonlinear least squares problem by way of a series of linear regressions.

The Gauss–Newton algorithm can be speeded-up considerably in the special case that some of the $\boldsymbol{\beta}$ appear linearly in μ. For example, if

$$\mu(t_i; \boldsymbol{\beta}) = \beta_1 \exp(-\beta_3 t_i) + \beta_2 \exp(-\beta_4 t_i),$$

then β_1 and β_2 are linear parameters. In such cases, the Gauss–Newton iteration can be restricted to the nonlinear parameters, β_3 and β_4. This idea is known as *separable* least squares; see, for example, Seber & Wild [9, Section 14.7]. Smyth [10] discusses the same principle in the context of maximum likelihood estimation.

Perhaps the most important application of Fisher scoring is to generalized linear models (GLMs). GLMs extend the idea of nonlinear regression to models with nonnormal error distributions, including **logistic regression** and **loglinear** models as special cases. GLMs assume that y_i is distributed according to a probability density or mass function of the form

$$p(y; \theta_i, \sigma^2) = a(y, \sigma^2) \exp\left\{\frac{1}{\sigma^2}[y\,\theta_i - b(\theta_i)]\right\}$$

for some functions b and a (a curved exponential family). We find that $\mathrm{E}(y_i) = \mu_i = b'(\theta_i)$ and $\mathrm{var}(y_i) = \sigma^2 v(\mu_i)$, where $v(\mu_i) = b''(\theta_i)$. If the mean μ_i of y_i is as given above for the nonlinear least squares, then the Fisher scoring iteration for $\boldsymbol{\beta}$ is a slight modification of the Gauss–Newton iteration:

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + (\dot{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{V}^{-1}\dot{\boldsymbol{\mu}})^{-1}\dot{\boldsymbol{\mu}}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\mathbf{V}$ is the diagonal matrix of the $v(\mu_i)$. The update for $\boldsymbol{\beta}$ iteration is still obtained from a linear regression of the residuals on $\dot{\boldsymbol{\mu}}$, but now the observations are weighted inversely according to their variances.

Classical GLMs assume a link-linear model of the form

$$h(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$$

for some link function h. In that case the Fisher scoring update can be reorganized as

$$\boldsymbol{\beta}_{k+1} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z},$$

where $\mathbf{z}$ is a working vector with components $z_i = h'(\mu_i)(y_i - \mu_i) + h(\mu_i)$ and $\mathbf{W}$ is a diagonal matrix of working weights $1/[h''(\mu_i)^2 v(\mu_i)]$. The updated $\boldsymbol{\beta}$ is obtained from weighted linear regression of the working vector $\mathbf{z}$ on $\mathbf{X}$. Since $\mathbf{X}$ remains the same throughout the iteration, but the working weights change, this iteration is known as *iteratively reweighted least squares* (IRLS).

When the observations y_i follow an exponential family distribution, observed and expected information coincide, so that Fisher scoring is the same as Newton's method. For GLMs this is so if h is the *canonical link*. We can conclude from this that IRLS is quadratically convergent for logistic regression and loglinear models, but linearly convergent for binomial regression with a probit link, for example (*see* **Quantal Response Models**). In practice, rapid linear regression is difficult to distinguish from quadratic convergence.

Nonderivative Methods

The Nelder–Mead *downhill simplex algorithm* is a popular derivative-free optimization method. It is based on the idea of function comparisons amongst a simplex of $N + 1$ points. Depending on the function values, the simplex is reflected or shrunk away from the maximum point. Although there are no theoretical results on the convergence of the algorithm, it works very well on a range of practical problems. It is a good choice if a once-off solution is required with minimum programming effort or if the function to be minimized is not differentiable.

If you are prepared to use a more complex program and if the function to be optimized is smoothly differentiable, the best performing methods for optimization without derivatives are quasi-Newton methods with difference approximations for the gradient vector. These programs require only the objective function as input, and compute difference approximations for the derivatives internally. Note that this is different from computing numerical derivatives and inputting them as derivatives to a program designed to accept analytic derivatives. Such a strategy is unlikely to be successful, as the numerical derivatives

are unlikely to show the assumed analytic behavior.

Close competitors to the finite-difference methods are *direction set methods*. These methods perform one-dimensional line searches in a series of directions which are chosen to be approximately *conjugate*, i.e. orthogonal with respect to the second derivative matrix. The best current implementation is given by Brent [1].

EM Algorithm

The **EM algorithm** is not an optimization method in its own right, but rather a statistical method of making optimization easier. The idea is the possibility that the log likelihood $\ell(\mathbf{y};\boldsymbol{\theta})$ might be easier to maximize if there were additional observations or information. Let $\mathbf{z}$ be the completed data, and let $\ell(\mathbf{z};\boldsymbol{\theta})$ be the log likelihood for $\mathbf{z}$. Maximizing the *incomplete* likelihood $\ell(\mathbf{y};\boldsymbol{\theta})$ is equivalent to maximizing the conditional expectation of the *complete* likelihood given $\mathbf{y}$, $\mathrm{E}[\ell(\mathbf{z};\boldsymbol{\theta})|\mathbf{y})$. In most cases, the EM is applied when the complete likelihood can be maximized in one step. However, the conditional expectation changes with the updated estimate of $\boldsymbol{\theta}$. So the optimization proceeds by alternate steps of expectation and maximization – hence the name "EM".

The EM algorithm is linearly convergent, at a rate which depends on the proportion of observed to unobserved Fisher information. Let ρ be the fraction of the Fisher information for a particular parameter in the complete log likelihood $\ell(\mathbf{z};\boldsymbol{\beta})$ which is not actually observed. Then the error in the estimate for that parameter after each iteration is $\varepsilon_{k+1} \approx \rho\varepsilon_k$. The proportion ρ can in applications be very close, or even equal, to one for some parameters, so that convergence can be very slow. On the other hand, the EM algorithm normally converges even from poor starting values. The iteration can often be speeded up by *Aitkin acceleration*, which attempts to convert linear convergence into quadratic convergence [8, p. 92].

Software

Optimization software is included in the commercial subroutine libraries IMSL and NAG, and in many statistical programs such as SAS, **S-PLUS, R**, MATLAB, and Gauss (*see* **Software, Biostatistical**). Publicly available software can be obtained by searching the NETLIB online library at `http://www.netlib.org/`. The guides and software provided by the Optimization Technology Center at the Argonne National Laboratory at the URL, `http://www.ece.northwestern.edu/OTC`, are also worth considering. Less elaborate programs suitable for user modification can be found in *Numerical Recipes* [8].

References

[1] Brent, R.P. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs.

[2] Dennis, J.E. & Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs.

[3] Fletcher, R. (1987). *Practical Methods of Optimization*, 2nd Ed. Wiley, New York.

[4] Gill, P.E., Murray, W. & Wright, M.H. (1981). *Practical Optimization*. Academic Press, New York.

[5] Moré, J.J. & Wright, W.J. (1993). *Optimization Software Guide*. Society for Industrial and Applied Mathematics, Philadelphia.

[6] Ortega, J.M. & Rheinboldt, W.C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York.

[7] Rao, C.R. (1973). *Linear Statistical Inference*, 2nd Ed. Wiley, New York.

[8] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in Fortran*. Cambridge University Press, Cambridge.

[9] Seber, G.A.F. & Wild, C.J. (1989). *Nonlinear Regression*. Wiley, New York.

[10] Smyth, G.K. (1996). Partitioned algorithms for maximum likelihood and other nonlinear estimation, *Statistics and Computing* **6**, 201–216.

(*See also* **Numerical Analysis**)

GORDON K. SMYTH

Optres Rotation

Optres rotation [2, 3] is a nonquartic method of performing an oblique rotation of a matrix $\mathbf{V}$ of dimension $(p \times k)$ made up of vectors associated with **principal components analysis** or **factor analysis** in order to transform these quantities into new variables by the relationship $\mathbf{B} = \mathbf{V}\boldsymbol{\Theta}$ such that $\mathbf{B}$ will approximate a **simple structure**. The matrix $\mathbf{B}$ is of dimension $(p \times k)$ and the matrix $\boldsymbol{\Theta}$ is of dimension $(k \times k)$ (*see* **Rotation of Axes**). Optres rotation is essentially an enhancement of a conventional **oblique rotation**. The principle feature of Optres rotation is that it requires, for each vector, a specification of the variables (called *salient* variables) whose rotated coefficients should be maximized and the nonsalient variables whose coefficients should be minimized. This specification is done by means of an algorithm applied to the results of another oblique rotation such as **Promax**. The Optres rotation then produces a new rotation subject to the constraints associated with the identified variables. Since Promax, itself, requires an orthogonal rotation, such as **Varimax**, for its starting approximation, Optres rotation is a three-step procedure.

The Optres procedure is based on the following criterion:

$$Q_j = M({v_{j(\mathrm{s})}}^2) - cM({v_{j(\mathrm{ns})}}^2) = \max,$$

where $M({v_{j(\mathrm{s})}}^2)$ is the mean of the squared loadings of the salient variables of component or factor j (*see* **Factor Loading Matrix**), $M({v_{j(\mathrm{ns})}}^2)$ is the mean of the squared loadings of the nonsalient variables of component or factor j, and c is a constant designed to give equal importance to these two quantities. Hakstian recommended $c = 50$, but Cureton [1] felt that $c = 100$ was more appropriate.

For the Decathlon example given in the article **Rotation of Axes**, the Optres solutions for $c = 50$ and $c = 100$ are given in Table 1 along with the original principal component characteristic vectors (*see* **Eigenvector**).

References

[1] Cureton, E.E. (1976). Studies of the PROMAX and Optres rotations, *Multivariate Behavioral Research* **4**, 449–460.

[2] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[3] Hakstian, A.R. (1972). Optimizing the resolution between salient and non-salient factor pattern coefficients, *British Journal of Mathematical and Statistical Psychology* **25**, 229–245.

(*See also* **Axes in Multivariate Analysis**)

J. EDWARD JACKSON

Table 1 Decathlon: Characteristic and Optres rotated vectors

	Characteristic vectors				Optres $c = 50$				Optres $c = 100$			
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
100 m run	0.69	0.22	−0.52	−0.21	0.79	0.01	0.01	−0.00	0.79	0.01	−0.00	−0.02
Long jump	0.79	0.18	−0.19	0.09	0.45	0.04	0.37	0.02	0.44	0.03	0.36	0.01
Shotput	0.70	−0.53	0.05	−0.18	0.05	0.73	0.05	−0.07	0.05	0.73	0.05	−0.07
High jump	0.67	0.13	0.14	0.40	0.02	−0.01	0.66	0.00	0.01	−0.02	0.66	0.00
400 m run	0.62	0.55	−0.08	−0.42	0.71	0.00	−0.08	0.55	0.71	0.00	−0.09	0.54
110 m hurdle	0.69	0.04	−0.16	0.35	0.21	−0.02	0.55	−0.20	0.20	−0.02	0.55	−0.20
Discus	0.62	−0.52	0.11	−0.23	0.01	0.74	−0.01	−0.01	0.00	0.74	−0.01	−0.00
Pole vault	0.54	0.09	0.41	0.44	−0.25	0.04	0.69	0.11	−0.26	0.04	0.69	0.11
Javelin	0.43	−0.44	0.37	−0.24	−0.19	0.71	−0.03	0.18	−0.20	0.71	−0.03	0.20
1500 m run	0.15	0.60	0.66	−0.28	0.01	−0.02	0.02	0.90	0.00	−0.01	0.01	0.90

Order Statistics

The order statistics of a collection of data or **random variables** are their ordered values. More precisely, if $X_1, \ldots, X_n$ is a collection of random variables with ordered values

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

then their rth order statistic is the rth smallest among them, $X_{(r)}$. Thus $X_{(1)}$ and $X_{(n)}$ are, respectively, the sample minimum and maximum. Clearly the order statistics are not independent. In principle the order statistics are certain to be unequal when the underlying data are continuous, but in practice this depends on the degree of rounding to which they are subject.

The order statistics are widely used as the basis of estimators and assessment of fit, but they have many other uses. Simple statistics based on them include:

1. The sample **median**,

$$\text{median}_j X_j = \begin{cases} X_{((n+1)/2)}, & n \text{ odd}, \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}, & n \text{ even}. \end{cases}$$

2. The α trimmed mean,

$$\frac{1}{n-2r} \sum_{j=r+1}^{n-r} X_{(j)}, \tag{1}$$

where r is the integer part of αn, which gives the sample average when $\alpha = 0$ and is interpreted as the median when $\alpha = 0.5$ (*see* **Trimming and Winsorization**).

3. The median absolute deviation (MAD),

$$\text{median}_i |X_i - \text{median}_j X_j|.$$

4. The interquartile range (IQR), $X_{(m_2)} - X_{(m_1)}$, where m_2 and m_1 are the integer parts of $\frac{3}{4}n$ and $\frac{1}{4}n$.
5. The **range**, $R = X_{(n)} - X_{(1)}$.

The first two of these are estimates of the location of the sample, while the last three can be used to estimate its dispersion. The median, trimmed mean, interquartile range, and MAD are little affected by **outliers**, while the range is badly affected by them because it depends on the most extreme observations in the sample. The median, trimmed mean, interquartile range, and range are linear combinations of order statistics, a class of quantities called L statistics.

Linear interpolation between adjacent order statistics is often used in small samples; an example is provided by the median above.

For illustration, consider the following data, which show the differences between the numbers of attacks of angina pectoris that were suffered by 12 patients, each of whom was given two different treatments [4, p. 16]:

$$-25 \quad 2 \quad 2 \quad 3 \quad 5 \quad 7 \quad 7 \quad 7 \quad 9 \quad 14 \quad 19 \quad 42.$$

Their median is 7, their 25% trimmed mean 6.33, their interquartile range 7, their MAD 4.5, and their range 67. If the lowest difference of -25 is replaced by -125, than the values of the first four are unchanged, but the range becomes 167. The average and sample standard deviation for the original data are 7.67 and 15.11; for the modified data they are -0.67 and 40.69. Statistics such as the average, sample standard deviation, and range are sensitive to the values of extreme observations, and for this reason *robust estimates* of location and scale are often based on combinations of nonextreme order statistics (*see* **Robustness**). In this example the rounding is severe enough to give a number of ties.

Exact Distributions

Suppose that $X_1, \ldots, X_n$ are independent identically distributed continuous random variables with cumulative distribution function (cdf) F and probability density function (pdf) f. Then the rth order statistic, $X_{(r)}$, has cdf

$$\Pr(X_{(r)} \leq x) = \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i} \tag{2}$$

and pdf

$$\frac{n!}{(n-r)!(r-1)!} F^{r-1}(x)[1 - F(x)]^{n-r} f(x). \tag{3}$$

Eq. (2) follows because the event $X_{(r)} \leq x$ occurs if and only if at least r of X_j are less than or equal to x, and these events are independent and

each has probability $F(x)$. Expression (3) is obtained on differentiating (2). When the distribution F is discrete, the possibility of ties among the $X_1, \ldots, X_n$ makes the corresponding formulas messy.

If F is known, then expressions for the mean, variance, and other moments of $X_{(r)}$ can be established from (3). For example, if the X_j have the **exponential** pdf $f(x) = \lambda \exp(-\lambda x), x > 0$, then it is fairly straightforward to show that for $r = 1, \ldots, n-1$ the differences of order statistics $X_{(r+1)} - X_{(r)}$ are independent exponential variables with means $\lambda^{-1}/(n-r)$. Consequently

$$\lambda \mathrm{E}(X_{(r)}) = e(r, n) = \sum_{j=1}^{r} (n + 1 - j)^{-1}.$$

The quantities $e(r, n)$, sometimes called exponential scores, are useful in assessing the fit of the exponential distribution. In a plot of the order statistics against the $e(r, n)$, a straight line indicates perfect fit of the distribution to the data, with the slope giving an estimate of λ^{-1}, while departures from exponentiality will show as different types of curvature. Probability plots such as this are widely used for assessing fit. The best known of them is based on **normal scores**, and is used to check normality of a sample of data. In such plots it often suffices to replace the rth expected order statistic by $F^{-1}[r/(n+1)]$, which is an approximation to the r/n **quantile** of F, though in particular cases better approximations are sometimes available.

The joint pdf of any subset of order statistics can be written down in a form that generalizes (3), and expressions for their covariances and other properties obtained. For example, the expected value of the range of $n = 12$ independent normal variables with variance σ^2 is 3.26σ, so for the sample above an estimate of σ based on the range is $R/3.26 = 20.55$. Estimates such as this are easy to calculate and hence are widely used in quality control, where the sample size may be tiny, but their sensitivity to outliers makes them too unreliable for general use.

The exact joint distribution of order statistics is also useful in testing for outliers. For example, if $X_1, \ldots, X_n$ is thought to be a sample from the exponential distribution, except that the largest observation may be an outlier, then a standard test rejects the null hypothesis of no outlier for large values of $T = X_{(n)} / \sum X_j$. The statistic T has a **beta distribution** when there is no outlier, so exact calculation of the significance level is straightforward. Barnett & Lewis [3] is a compendium of such tests, many of which rely on order statistics.

The cdf (2) is the basis for a simple nonparametric **confidence interval** for the p quantile of F, i.e. the value x_p such that $F(x_p) = p$; we suppose that $f(x_p) > 0$, so x_p is unique. Suppose that we choose s and r so that as nearly as possible,

$$\Pr(x_p \le X_{(s)}) = \alpha, \qquad \Pr(X_{(r)} \le x_p) = \alpha;$$

$(X_{(s)}, X_{(r)})$ is then an approximately equi-tailed $1 - 2\alpha$ confidence interval for x_p. Since $x_p = F^{-1}(p)$, we should take the values of r and s so that

$$\sum_{j=0}^{s-1} \binom{n}{j} p^j (1-p)^{n-j} \doteq \alpha;$$

$$\sum_{j=r}^{n} \binom{n}{j} p^j (1-p)^{n-j} \doteq \alpha.$$

For illustration, suppose that a confidence interval is required for the median $x_{0.5}$ of the population of differences underlying the sample given above. In this case $n = 12$, and with $p = 0.5$ we find that $s = 3$ and $r = 10$ gives endpoints (2,14) for a 96% confidence interval for $x_{0.5}$.

Approximate Distributions

The exact formulas (2) and (3) for the pdf and cdf of the rth order statistic can be difficult to work with. Fortunately, there is a simple approximation for the pdf of a central order statistic. Suppose that r is the integer part of pn, where $0 < p < 1$, and let x_p denote the p quantile of F. As before, we suppose that $f(x_p) > 0$. Then in large samples the approximate distribution of $X_{(r)}$ is

$$X_{(r)} \dot{\sim} \mathrm{N}\left[x_p, \frac{p(1-p)}{n f^2(x_p)}\right]. \tag{4}$$

Thus the sample median can be used to estimate the median of the population from which the sample is drawn, and other order statistics can be used to estimate their corresponding quantiles. The result (4) shows that the limiting distribution of the median of a sample of size n from the $\mathrm{N}(\mu, \sigma^2)$ distribution is normal with mean μ and variance $\pi\sigma^2/(2n)$,

and that for such a sample the IQR and MAD have expected values 1.35σ and 0.675σ; thus IQR/1.35 and MAD/0.675 are the corresponding robust estimates of σ.

One consequence of (4) is that although nonparametric estimation of a population quantile is straightforward, it is much harder to give an accurate statement of its uncertainty in small samples. For we see from (4) that the problem of estimating the variance of $X_{(r)}$ is tantamount to estimating the density $f(\cdot)$ at x_p, and accurate **density estimation** needs large samples because nonparametric density estimators converge only slowly.

Under mild conditions on F, (4) extends to any fixed number of central order statistics. Consider $X_{(r_1)} \leq \cdots \leq X_{(r_k)}$, where r_j is the integer part of np_j, and $0 < p_1 < \cdots < p_k < 1$, and for $j = 1, \ldots, k$ let $x_{p_j} = F^{-1}(p_j)$ and suppose that $f(x_{p_j})$ is positive and finite. Then the joint limiting distribution of $X_{(r_1)}, \ldots, X_{(r_k)}$ is **multivariate normal** with means $x_{p_1}, \ldots, x_{p_k}$ and covariances given by $p_i(1 - p_j)/[nf(x_{p_i})f(x_{p_j})]$, where $j \geq i$. This implies, for example, that the IQR has a limiting normal distribution, because it is a difference of the two approximately bivariate normal random variables obtained by taking $p_1 = 0.25$ and $p_2 = 0.75$.

These limiting distributions do not apply to an order statistic $X_{(r)}$ for any fixed r, since in that case $r/n \to 0$ as $n \to \infty$, corresponding to $p = 0$. In this situation the approximate distributions that apply are those for extreme values, and it is possible to show that if there are sequences a_n and b_n such that $Y_{(n+1-r)} = a_n(X_{(n+1-r)} - b_n)$ has a nondegenerate limiting distribution as $n \to \infty$, then the pdf of $Y_{(n+1-r)}$ must have the form

$$\frac{1}{\tau(r-1)!}\left[1 + \kappa\left(\frac{y-\eta}{\tau}\right)\right]^{-r/\kappa - 1} \times \exp\left\{-\left[1 + \kappa\left(\frac{y-\eta}{\tau}\right)\right]^{-1/\kappa}\right\},$$
$$-\infty < \eta, \kappa < \infty, \tau > 0,$$

where the range of y is such that $1 + \kappa(y - \eta)/\tau > 0$. Such distributions are useful in a wide range of applications, more details of which are given in the article on **extreme values**, and in [8] and [10].

L statistics such as (1) are widely used, but appropriate distributional approximations are not immediate from the discussion above because they depend neither on extreme values nor on a fixed number of order statistics. Such a linear combination can be written as

$$T = n^{-1}\sum_{r=1}^{n} w\left(\frac{r}{n}\right) X_{(r)},$$

where $w(u)$ is a function of u, for $0 \leq u \leq 1$, which ascribes weights to the order statistics. For the α trimmed mean, for example,

$$w(u) = \begin{cases} 1, & \alpha \leq u \leq 1 - \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Then under mild conditions on $w(\cdot)$ and F, a general result [12] is that T has an approximate normal distribution with mean and variance

$$\int w[F(x)]x \, \mathrm{d}F(x),$$
$$\frac{2}{n}\int_{-\infty}^{\infty} \mathrm{d}x \int_{x}^{\infty} \mathrm{d}y \, w[F(x)]w[F(y)]F(x)[1 - F(y)]. \tag{5}$$

This result is theoretically useful because the conditions are mild and easy to verify for particular choices of $w(\cdot)$, and because estimates of the variance in (5) can form the basis of uncertainty statements for the estimator T.

References

[1] Arnold, B.C. & Balakrishnan, N. (1989). *Relations, Bounds, and Approximations for Order Statistics. Lecture Notes in Statistics*, Vol. 53. Springer-Verlag, New York.

[2] Arnold, B.C., Balakrishnan, N. & Nagaraja, H.N. (1992). *A First Course in Order Statistics*. Wiley, New York.

[3] Barnett, V.D. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Ed. Wiley, New York.

[4] Bland, M. (1995). *Medical Statistics*, 2nd Ed. Clarendon Press, Oxford.

[5] David, H.A. (1981). *Order Statistics*, 2nd Ed. Wiley, New York.

[6] David, H.A. (1985). Order statistics, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson. eds. Wiley, New York.

[7] Galambos, J. (1985). Order statistics, in *Handbook of Statistics*, Vol. 4, P.R. Krishnaiah & P.K. Sen, eds. Elsevier, Amsterdam, Chapter 17, pp. 359–382.

[8] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd Ed. Krieger, Malabar.

[9] Johnson, N.L., Kotz, S. & Balakrishnan, N. *Continuous Univariate Distributions*, Vol. I, 2nd Ed. Wiley, New York.

[10] Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics*. Springer-Verlag, New York.
[11] Resnick, S.I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer-Verlag, New York.
[12] Stigler, S.M. (1974). Linear functions of order statistics with smooth weight functions, *Annals of Statistics* **2**, 676–693.

Bibliography

The huge literature on order statistics to 1981 is surveyed by David [5], with a more recent book-length treatment by Arnold et al. [2]. David [6] gives a brief survey, while Galambos [7] is a somewhat longer fairly mathematical account. Reiss [10] focuses particularly on approximate distributions useful in nonparametric statistics, while Resnick [11] discusses extreme-value approximations. Arnold & Balakrishnan [1] give a compendium of order statistic approximations. References to order statistics of the more common continuous distributions are given by Johnson et al. [9].

A.C. DAVISON & F. DORSAZ

Ordered Alternatives

In experiments to determine whether a particular chemical causes cancer in laboratory animals, the dose response is an important indication of possible carcinogenicity. If the test chemical is a carcinogen, then in general the higher the dose the more animals develop cancer. The dose response in toxicity assays is an important example of ordered alternatives.

Do children tend to lie more as they grow older? Here we would like to known if the lying tendency is related monotonically to age. Does drinking during pregnancy increase the risk of congenital malformation? Are younger cancer patients more prone to multiple malignancies? These are some illustrative examples of ordered alternatives in biostatistics.

Let $d_1 < d_2 < \cdots < d_{k-1} < d_k$ denote ordered values associated with k groups. There may, for instance, be k levels of chemical dose in toxicity assays. They may represent k age groups. The values are not necessarily quantitative but can be ordinal values such as $1, 2, \ldots, k$, denoting unquantifiable k values. We assume that the increasing d value also increases the outcome value under the **alternative hypothesis**, generating ordered alternatives.

Ordered Alternative in Binomial Probabilities

Suppose that n_i animals are assigned to the d_i dose group, and that $n_i\mu_i$ animals are expected to develop a particular type of tumor, where μ_i is an unknown probability of developing the tumor during the study period. We are interested in testing the **null hypothesis** that $\mu_1 = \mu_2 = \cdots = \mu_k$ against the ordered alternative that $\mu_1 \le \mu_2 \le \cdots \le \mu_k$ with at least one strict inequality. Let Y_i denote the number of animals developing the tumor of interest in the d_i dose group. Y_i is a **binomial** random variable. We would like to determine whether the higher dose increases the risk of the tumor. For such an analysis, the usual **chi-square test** with $k-1$ **degrees of freedom** is inappropriate, because it is directed toward finding if μ_is are different rather than if μ_is increase with increasing dose. The χ^2 statistic is given by $\sum n_i(\hat{\mu}_i - \hat{\mu})^2/\hat{\mu}(1-\hat{\mu})$, where $\hat{\mu}_i = y_i/n_i$ is the observed proportion of animals developing the tumor of interest and $\hat{\mu} = \sum y_i / \sum n_i$ is the overall observed proportion of animals developing the tumor.

An appropriate test statistic for the ordered alternative is

$$Z_{\mathrm{L}} = \frac{\sum d_i Y_i - \hat{\mu} \sum n_i d_i}{\left[\hat{\mu}(1-\hat{\mu}) \sum n_i (d_i - \overline{d})^2\right]^{1/2}},$$

where $\overline{d}$ is $\sum n_i d_i / \sum n_i$ (*see* **Trend Test for Counts and Proportions**). If $Z_{\mathrm{L}} \ge z_{1-\alpha}$, then the null hypothesis is rejected and the chemical is considered to cause the tumor. Here, $z_{1-\alpha}$ is the upper α cutoff point of the standard **normal distribution**. In applying the Z_{L}, it is assumed that μ_i is related to d_i linearly, i.e. $\mu_i = a + bd_i$. When such an assumption is improper, we can apply

$$Z_{\mathrm{M}} = \frac{\sum i Y_i - \hat{\mu} \sum i n_i}{\left[\hat{\mu}(1-\hat{\mu}) \sum n_i (i - \overline{i})^2\right]^{1/2}},$$

where $\overline{i} = \sum i n_i / \sum n_i$. If $Z_{\mathrm{M}} \ge z_{1-\alpha}$, then the null hypothesis is rejected in favor of the ordered alternative. Both Z_{L} and Z_{M} statistics are considered **nonparametric** [1], while the Z_{M} statistic is applicable to broader ordered alternatives than the Z_{L} statistic.

Ordered Alternative in Poisson Means

The test statistic based on $\sum d_i Y_i$ or $\sum i Y_i$ is applicable to ordered alternatives in means of k **Poisson** distributions. Let Y_i be a Poisson random variable with mean $w_i\mu_i$ for $i = 1, \ldots, k$, where w_i is a known weight variable and μ_i is an unknown parameter which is ordered under the alternative hypothesis, namely $\mu_1 \le \mu_2 \le \cdots \le \mu_k$ with at least one strict inequality. For example, suppose that Y_i is a binomial random variable with sample size n_i and probability $\mu_i = p(d_i)$, and that $p(d_i)$ is an unknown function of given dose $d_i, d_1 < d_2 < \cdots < d_k$. Furthermore, assume that μ_i is very small. Then Y_i is considered a Poisson random variable with mean of $n_i\mu_i$.

We know that $(Y_1, \ldots, Y_k)$ is a **multinomial** vector with $(N, w_1\mu_1/\sum w_i\mu_i, \ldots, w_k\mu_k/\sum w_i\mu_i)$, conditional on $N = \sum Y_i$. Lee [6] showed that both the $\sum d_i Y_i$ and $\sum i Y_i$ statistics are applicable to testing the null hypothesis against the ordered alternative, although $\sum i Y_i$ is more broadly applicable than

$\sum d_i Y_i$. For given N and w_is, we can obtain exact distributions of $\sum d_i Y_i$ and $\sum i Y_i$ statistics under the null hypothesis that μ_is are constant for all i. For notational convenience assume that $\sum w_i = 1$ and that $\sum w_i \mu_i = 1$. Under the null hypothesis, $T_{\mathrm{M}} = (\sum i Y_i - n \sum i w_i)/\{N[\sum i^2 w_i - (\sum i w_i)^2]\}^{1/2}$ is an asymptotic standard normal random variable, and thus if N is large and $T_{\mathrm{M}} \geq z_{1-\alpha}$, then the null hypothesis is rejected. We can apply $T_{\mathrm{L}} = (\sum d_i Y_i - n \sum d_i w_i)/\{N[\sum d_i^2 w_i - (\sum d_i w_i)^2]\}^{1/2}$ to testing somewhat narrower ordered alternatives [6].

In testing ordered alternatives in Poisson means, we can apply the best asymptotic normal **(BAN)** estimation method [4]. For a small to moderate sample size, the BAN estimation method can fail [7]. Furthermore, test statistics $\sum i Y_i$ and $\sum d_i Y_i$ are asymptotically **efficient** [10].

Ordered Alternative in Multinomial Probabilities

Let $(Y_1, Y_2, \ldots, Y_k)$ be a k-variate multinomial vector with sample size N and probabilities $(w_1\mu_1, w_2\mu_2, \ldots, w_k\mu_k)$, where $\sum w_i \mu_i = 1$ and $\sum w_i = 1$. Under the null hypothesis, $\mu_i = \mu_j$ for $i \neq j$, and under the alternative, $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$, with at least one strict inequality. We can apply $\sum i Y_i$ to testing the ordered alternative. Examples of ordered alternatives in multinomial probabilities can be found in [5] and [6].

Nonparametric Tests for Ordered Alternatives in Means of Continuous Distributions

Suppose that Y_{ij} is an independent random variable with distribution function $F(y_{ij} - \mu_i)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. This is a one-way **analysis of variance** situation where the form of the underlying distribution is unknown. Under the null hypothesis, $\mu_i = \mu_{i'}$ for $i \neq i'$, and under the alternative, $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$ with at least one strict inequality. If $k = 2$, then we can apply the **Wilcoxon–Mann–Whitney** two-sample test statistic. For $k > 2$, we compute the Mann–Whitney version of the Wilcoxon–Mann–Whitney statistic for every pair of $1 \leq i < i' \leq k$, and sum $k(k-1)/2$ Mann–Whitney statistics. This statistic was proposed for testing the ordered alternative by Jonckheere [3].

Let $I(a < b) = 1$ if $a < b$, 0 if $a > b$, and $\frac{1}{2}$ if $a = b$. For $i < i'$, let $U_{ii'} = \sum_{u=1}^{n_i} \sum_{v=1}^{n_{i'}} I(y_{iu} < y_{i'v})$. The Jonckheere statistic is given by $J = \sum_{i=1}^{k-1} \sum_{i'=i+1}^{k} U_{ii'}$. The exact null distribution of J for $k = 3, 4, 5, 6$ and small n_is is available [2]. Under the null hypothesis, $\mathrm{E}_0(J) = (N^2 - \sum_{i=1}^{k} n_i^2)/4$ and $\mathrm{var}_0(J) = [N^2(2N+3) - \sum_{i=1}^{k} n_i^2(2n_i+3)]/72$, where $N = \sum n_i$. For a large $\min(n_i, i = 1, \ldots, k)$, if $[J - \mathrm{E}_0(J)]/[\mathrm{var}_0(J)]^{1/2} \geq z_{1-\alpha}$, then the null hypothesis is rejected in favor of the ordered alternative at significance level α. When some blocks are incompletes because of missing observations, see the methods presented in [9].

When observations are collected in n blocks of size k, we cannot apply the Jonckheere test because of the possible block effect. k is the number of treatment groups. In this case, the Page statistic [8] is a relatively simple and easy to use method to test the ordered alternative. Suppose that Y_{ij} is an independent random variable with distribution function $F(y_{ij} - \mu_i - \theta_j)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n, \theta$ is a block effect, and the parameter of interest is μ, the treatment effect. Under the null hypothesis, $\mu_i = \mu_{i'}$ for $i \neq i'$, and under the alternative, $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$ with at least one strict inequality.

Let r_{ij} be the **rank** of y_{ij} among $(y_{1j}, y_{2j}, \ldots, y_{(i-1)j}, y_{ij}, \ldots, y_{kj})$, k observations of the jth block. Let $R_i = \sum_{j=1}^{n} r_{ij}$. The Page statistic for testing the ordered alternative is given by $L = \sum_{i=1}^{k} i R_i$. Exact cutoff points for given significance levels of 0.001, 0.01, and 0.05 are given for $k = 3, \ldots, 8$ for small to moderate n [2]. Under the null hypothesis, $\mathrm{E}_0(L) = nk(k+1)^2/4$ and $\mathrm{var}_0(L) = n(k^3 - k)^2/144(k-1)$. For a large n, if $[L - \mathrm{E}_0(L)]/[\mathrm{var}_0(L)]^{1/2} \geq z_{1-\alpha}$, then the null hypothesis is rejected in favor of the alternative hypothesis at the given significance level α. When some blocks are incomplete because of missing observations, see the methods presented in [9].

References

[1] Cox, D.R. (1970). *The Analysis of Binary Data*. Chapman & Hall, London.

[2] Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods*. Wiley, New York.

[3] Jonckheere, A.R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika* **41**, 133–145.

[4] Jorgenson, D.W. (1960). Multiple regression analysis of a Poisson process, *Journal of the American Statistical Association* **56**, 235–245.

[5] Lee, Y.J. (1977). Maximin tests of randomness against ordered alternatives: the multinomial distribution case, *Journal of the American Statistical Association* **72**, 673–675.

[6] Lee, Y.J. (1980). Test of trend in count data: multinomial distribution case, *Journal of the American Statistical Association* **75**, 1010–1014.

[7] Lee, Y.J. (1985). Tests of monotone trend in K Poisson means, *Journal of Quality Technology* **17**, 44–49.

[8] Page, E.B. (1963). Ordered hypothesis for multiple treatments: a significance test for linear ranks, *Journal of the American Statistical Association* **58**, 216–230.

[9] Park, E. & Lee, Y.J. (2000). Non-parametric test of ordered alternatives in incompletes blocks, *Statistics in Medicine* **19**, 1329–1337.

[10] Tarone, R.E. & Gart, J.J. (1980). On the robustness of combined tests for trends in proportions, *Journal of the American Statistical Association* **75**, 110–116.

(*See also* **Isotonic Inference**; **Isotonic Regression**)

Y.J. LEE

Ordered Categorical Data

The methodology considered herein complements general methodology for categorical data analysis (*see*, for example, **Categorical Data Analysis**; **Contingency Table**; **Loglinear Model**), but it differs from the general methodology in that a natural ordering is assumed for one or more of the categorical variables of interest. Ordered categorical data are ubiquitous in the medical and health sciences, and research into methodology for the analysis of such data has been vigorous over the past 20 or so years.

An observed variable that is ordered categorical may arise as the result of **categorizing a continuous variable**, or it may be an inherently categorical measurement. It is routine medical practice to classify patients as being at various degrees of risk for developing a disease or condition according to specified ranges on risk factors (*see* **Prognostic Factors for Survival**). In cardiovascular disease, for example, important risk factors are serum cholesterol and blood pressure levels. These variables are intrinsically continuous, but for some purposes, such as classifying patients into groups or levels of risk, it is useful to analyze them as ordered categorical variables; see, for example, [49]. However, in classifying levels of pain relief attained with a treatment it is often only possible to arrive at subjective categorizations, such as "no relief", "some relief", "considerable relief", or "complete relief". Likewise, clinical assessments of mental health status are sometimes based on a numerical scoring system for one or more functional items or tasks (*see* **Scores**), and at other times according to a subjective clinical evaluation. Even when the observed data arise from the subjective assignment of an observed response into a category on an ordered scale, it might be assumed that there does indeed exist an underlying continuous random variable for which the discrete classification is an imperfect measure. The issue of whether the observed variable or the underlying continuous variable is of primary interest sparked a contentious debate between **Karl Pearson** [50] and **G. Udny Yule** [58, 59] in the early part of this century. Those issues will not be revisited here, but rather I cover methodology developed from both points of view.

It should be noted that methodology developed for categorical data in general can be applied to the analysis of ordered categorical data. There are, however, important advantages to using models and methodology developed to explicitly take account of the ordinal structure of the categories. In particular, models for ordered categorical data tend to be more **parsimonious** than their more general counterparts, thus resulting in more **efficient** inferences and facilitating interpretation of parameters and economical presentation of results. The ability to focus hypothesis tests on restricted alternatives specific to the ordinal structure of the data often results in tests that are substantially more **powerful** (for the alternatives of interest) than those that cover omnibus alternatives; see, for example [3, pp. 269, 282–283] (*see* **Isotonic Inference**).

The models and methodology outlined herein focus on two general areas of statistical analysis: (i) **association** and (ii) **regression**. The study of **correlation** and **linear regression** are central to any development of models and methodology for the analysis of continuous variables, and the situation is no different for categorical variables. Many of the models, measures, and parameters that I consider here have analogous counterparts in **general linear models** for continuous data, e.g. in ordinary linear regression and in the **analysis of variance**. There are, of course, important issues particular to the analysis of ordered categorical data, and those are the primary focus of this article.

Models for the analysis of association are the first topic I consider. The **odds ratio** serves as the focus of much of the modern analysis of **binary data**, and the models I consider for the analysis of association build on that work by focusing also on the use of odds ratios in summarizing associations. Regression models for the analysis of an ordered categorical response are covered subsequently. The approach taken is to focus on models that generalize (extend) **logistic regression** modeling of a binary response. I conclude with a brief discussion of recent developments and ongoing lines of research on the analysis of ordered categorical data.

It is impossible to cover all approaches to the analysis of ordered categorical data here. It is inevitable that some topics are omitted, intentionally or otherwise, or not covered as fully as some others. For example, ordinal probit models (*see* **Quantal Response Models**) are mentioned only briefly. In such cases, I have attempted to provide citations to papers and books that will assist the interested reader in exploring those areas.

Models for Association

The correlation coefficient is the standard measure for the assessment of (linear) association between two continuous variables, and the parameters in a classical linear regression are readily related to it. The odds ratio plays the parallel role in the modeling and analysis of categorical data. Here, I work within the context of two-way contingency tables; citations to literature on multivariate generalizations are given subsequently. Continuous variables are introduced in the subsequent section on regression-type models. There are immediate connections between models for odds ratios and regression models for categorical data, just as there are immediate connections between partial correlations and regression coefficients in ordinary linear regression.

Odds Ratios for $I \times J$ Contingency Tables

Let π_{ij} denote the probability associated with the cell in row i and column j of the $I \times J$ table $(i, = 1, \ldots, I; \quad j = 1, \ldots, J)$, and let n_{ij} denote the corresponding observed count. The letters R and C are used to denote the row variable and the column variable, respectively. A loglinear representation of the saturated model (i.e. the only constraints placed on the cell probabilities follow from the axioms of probability) is

$$\ln \pi_{ij} = \lambda + \lambda_i^{\mathrm{R}} + \lambda_j^{\mathrm{C}} + \lambda_{ij}^{\mathrm{RC}},$$

where the λ terms are parameters to be estimated from the data. The model of statistical independence is the reduced model where $\lambda_{ij}^{\mathrm{RC}} = 0$ for all pairs (i, j). The association between R and C is captured by the $\lambda_{ij}^{\mathrm{RC}}$ terms, and contrasts of the $\lambda_{ij}^{\mathrm{RC}}$ terms are readily interpreted in terms of log odds ratios for 2×2 subtables of the complete table, i.e.

$$\ln\left(\frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}}\right) = \lambda_{ij}^{\mathrm{RC}} - \lambda_{ij'}^{\mathrm{RC}} - \lambda_{i'j}^{\mathrm{RC}} + \lambda_{i'j'}^{\mathrm{RC}}.$$

All possible local log odds ratios may be completely characterized by a set of $(I-1)(J-1)$ local log odds ratios, and the model of independence for the complete table is equivalent to independence in all $(I-1)(J-1)$ subtables. When at least one of the variables is ordered, there are parsimonious models between statistical independence and the saturated model that have been found to be extremely useful in empirical work for characterizing the pattern of association.

Loglinear Models of Association

Goodman [29] and Haberman [35] considered simple loglinear models for the association in contingency tables having ordered categories. Goodman specified models in terms of the basic set of local odds ratios $\theta_{ij} = (\pi_{ij}\pi_{i+1,j+1})/(\pi_{i,j+1}\pi_{i+1,j})$, and in particular the models of *uniform association, row effects association, column effects association*, and row + column *effects association* are:

$$\begin{aligned}
\text{uniform} &: \theta_{ij} = \theta, \\
\text{row effects} &: \theta_{ij} = \theta_i, \\
\text{column effects} &: \theta_{ij} = \theta_j, \\
\text{row + column effects} &: \theta_{ij} = \theta_{1i}\theta_{2j},
\end{aligned}$$

where $i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$. The uniform association model takes the local association, as characterized by local odds ratios, to be constant throughout the table. The row effects association model summarizes the local association in terms of $I-1$ odds ratios θ_i, and, likewise, the column effects association model summarizes the local association in terms of $J-1$ local odds ratios θ_j. The row + column effects model allows for additive variation of the local log odds ratios. Corresponding loglinear models for the cell probabilities restrict the $\lambda_{ij}^{\mathrm{RC}}$ parameters by assigning scores to the categories of one or both of the variables cross-classified. For example, in the uniform association model $\lambda_{ij}^{\mathrm{RC}} = \beta u_i v_j$, where the u_i are equally spaced and the v_j are equally spaced. Note that $\ln\theta = \beta[(u_{i+1} - u_i)(v_{j+1} - v_j)]$, and that there is no loss in generality in assuming the scores $u_i = i$ and $v_j = j$. This model is also known as the model of linear-by-linear association [3, 35] when the scores, or the distances between them, are fixed but not assumed to be equally spaced. Inferences are necessarily sensitive to the choice of scores [34], but these simple loglinear models provide a useful starting point for exploring associations between ordered categorical variables.

Example 1: Association Between Mental Health Status and Socioeconomic Status

In a now classic study of mental health in Manhattan, New York, Srole et al. [53] explore the relationship,

among others, between mental impairment and parents' socioeconomic status. Table 1, from that study, has been used extensively (see, for example [3, 23, 29, 32], and [33]) to illustrate the utility and application of models for ordered categorical data.

Residual degrees of freedom and deviance statistics for loglinear association models applied to this table are reported in Table 2. Note that under the assumption of multinomial sampling for the n_{ij} a deviance statistic equals minus two times the log likelihood ratio statistic, with an asymptotic **chi-square distribution** under the assumed model, for testing model **goodness of fit** (*see* **Likelihood Ratio Tests**). The hypothesis of statistical independence between mental health status and parents' socioeconomic status is rejected, both by the omnibus test of goodness of fit (deviance = 47.42, on 15 df) and by the more focused test of independence versus uniform association (deviance = 47.42 − 9.90 = 37.52, on 1 df). The uniform association model provides an excellent fit to these data (deviance = 9.90, on 14 df), and likelihood ratio comparisons of uniform association to the row effects, column effects, and row + column effects association models do not yield evidence of departures from uniform association in favor of these more general models. The **maximum likelihood** estimate $\hat{\beta}$ of β under uniform association is 0.09, with an estimated asymptotic **standard error** of 0.015. Hence, there is strong evidence that mental health status and parents' socioeconomic status are positively associated (i.e. higher socioeconomic status is associated with better mental health), but the strength of that association is quite small, i.e. an approximate 95% **confidence interval** for the uniform local odds ratio is (1.06, 1.13), as compared with the value 1.00 for independence.

Table 1 The Midtown Manhattan Study: mental health and parents' socioeconomic status

Parents' socioeconomic status	Mental health status			
	Well	Mild symptom formation	Moderate symptom formation	Impaired
A (high)	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F (low)	21	71	54	71

Table 2 Models summary for Table 1

Model	Degrees of freedom	Deviance
Independence	$(I-1)(J-1)=15$	47.42
Uniform association	$(I-1)(J-1)-1=14$	9.90
Row effects	$(I-1)(J-2)=10$	6.83
Column effects	$(I-2)(J-1)=12$	6.28
Row + column effects	$(I-2)(J-2)=8$	3.05

All of the above models are appropriate when both variables are ordered, but assuming fixed scores does have limitations. Assigning fixed scores to a variable implies a specified ordering and spacing to the categories. That may not be desirable or realistic for some ordered variables, and it is inappropriate for a nominal variable. The row effects (column effects) model does not assume fixed scores for the row (column) categories, and hence it can be used to estimate row (column) scores, and to analyze nominal–ordinal (ordinal–nominal) cross-classifications of counts; see, for example, [1], and [23]. The row + column effects model includes both fixed and estimated scores for both variables.

RC and Quasi-Symmetric RC Models

The preceding loglinear models are immensely useful, but they cannot be used to describe all data sets encountered in practice. An intrinsically nonlinear model that is a simple generalization of the linear-by-linear association model is

$$\ln(\pi_{ij}) = \lambda + \lambda_i^{\mathrm{R}} + \lambda_j^{\mathrm{C}} + \beta\mu_i\nu_j,$$

where the μ_i and ν_j are scores to be estimated from the data [11, 29]. Hereafter, we refer to this model as the *RC association* model. Setting $\Delta_i^{\mathrm{R}} = \mu_{i+1} - \mu_i$ and $\Delta_j^{\mathrm{C}} = \nu_{j+1} - \nu_j$, the local log-odds ratios are of the form

$$\ln(\theta_{ij}) = \beta\Delta_i^{\mathrm{R}}\Delta_j^{\mathrm{C}}.$$

The row effects and column effects models are special cases of the RC model where one of the two sets of scores are constrained to be equally spaced.

There are many instances in biostatistical practice where there is a one-to-one correspondence between the categories of two, or more, ordered categorical variables. It is frequently the case in

this situation that questions arise regarding issues of symmetry, be they in terms of the joint distribution of the variables, the marginal distributions of the variables, or in the association as summarized by local odds ratios. Models that exhibit symmetric marginal distributions are said to satisfy marginal homogeneity, and models of symmetric association are said to exhibit **quasi-symmetry**. Quasi-symmetry corresponds to symmetry in the association, but asymmetry in the joint distribution. The special case of the RC association model where the row and column scores are constrained to be equal (i.e. $\mu_i = \nu_i$, $i = 1, \ldots, I$) is a restricted form of quasi-symmetry. Models of this type have been considered by Agresti [2], Becker [17], and Goodman [29, 32], among others. Note that the uniform association model is the special case of the quasi-symmetric RC model where the scores are assumed to be equally spaced. Introducing the constraint $\lambda_i^{\mathrm{R}} = \lambda_i^{\mathrm{C}} (i = 1, \ldots, I)$ along with symmetric association results implies a symmetric joint distribution, and hence marginal homogeneity. In general, symmetry in the joint distribution is equivalent to marginal homogeneity and symmetric association holding simultaneously.

The RC association model is not a loglinear model or **generalized linear model**, and as such it is not readily estimated with standard computer programs for fitting generalized linear (or loglinear) models (*see* **Software, Biostatistical**). **Algorithms** that can be used to fit the model, as well as some generalizations of it, are described in Goodman [29], Becker [18], and Haberman [36]. It is also possible to program in some statistical packages, such as GLIM and SAS, to fit the RC association model; see, for example [1, Appendix D].

Example 2: Reliability of Self-Reported Passive Smoking Histories

Exposure to passive **smoking** has been studied as a possible risk factor for lung cancer, and hence reliable measures of exposure to passive smoking are required. The data I consider here are from a study of the reliability of passive smoking histories [51]. A sample of 177 controls from a case–control study of lung cancer were interviewed on two different occasions with respect to exposure to occupational and residential passive smoke. Subjects were to indicate the number of regular smokers with whom they had resided. The results for the residential histories are given in Table 3.

Table 3 Exposure to passive smoking in the residence: number of smokers resided with

First interview	Second interview			
	0	1	2	3+
0	19	7	1	2
1	2	15	18	10
2	1	5	5	6
3+	1	4	3	18

Table 4 Models summary for Table 3

Model	Degrees of freedom	Deviance
Independence	9	69.02
Uniform association	8	24.41
Row effects	6	13.47
Column effects	6	17.02
Row + column effects	4	10.05
RC	$(I-2)(J-2) = 4$	10.36
Quasi-symmetric RC	$(I-2)(J-1) = 6$	10.59

Residual degrees of freedom and deviance statistics for both loglinear association models and RC-type models are presented in Table 4. Again, there is strong evidence of a departure from statistical independence (deviance = 69.02, on 9 df), but in this case models more general than uniform association are better suited for summarizing the degree to which the first and second interview responses are associated. The conditional test of quasi-symmetric RC association vs. RC association provides a direct means for testing the assumption of homogeneous scores; there is no reason to reject the null hypothesis that the scores are homogeneous (deviance = $10.59 - 10.36 = 0.23$, on 2 df). In addition, the quasi-symmetric RC model provides a reasonable summary of these data (deviance = 10.59, on 6 df). The $\hat{\Delta}_i$, $i = 1, 2, 3$, scaled for direct comparison to the uniform distances equal 1, are (1.67, 0.36, 0.97), and $\hat{\beta} = 0.88$. The interested reader is referred to Becker [14] for discussion of how the estimated scores from the fit of the quasi-symmetric RC model may be used to measure and characterize the reliability of the residential passive smoking histories within the framework developed by Darroch & McCloud [26] for analyzing observer agreement (*see* **Agreement, Measurement of**). Other relevant

references on this subject are [2, 4, 7, 14, 19, 47], and [56].

Comments

Models for local association provide a useful starting point for the analysis of two-way cross-classifications where at least one of the variables cross-classified is ordered categorical. These models are not predicated on the assumption of an underlying continuous distribution for the ordered categorical variable(s), but there are close connections between some of the models and the **bivariate normal distribution** for continuous data [15, 30, 39]. Generalizations of the RC model to accommodate more general association structures have been developed (see, for example, [17, 20, 32], and [33]), as have generalizations for higher-way contingency tables (see, for example, [6, 16, 20, 22, 28, 29], and [33]).

In the case of the doubly ordered contingency table, global cross-ratios [24] provide an alternative odds-ratio measure of association. Models for global cross-ratios have been considered by Dale [25], Heagerty & Zeger [37], Molenberghs & Lesaffre [48], and Williamson et al. [57], among others. All of these authors also consider the joint modeling of the marginal distributions (*see* **Marginal Models**) and the association structure, which can be extremely useful for some purposes. Global cross-ratio models also do not require the assumption of an underlying continuous bivariate distribution, but certain models can be motivated in terms of a multivariate Plackett distribution [25, 48]. The joint modeling of marginal distributions and local association structure is considered by Lang & Agresti [43]. Multivariate probit models [10, 13, 41, 44, 46] (*see* **Quantal Response Models** for a definition of the univariate probit) are an approach to the analysis of ordered categorical that do assume an underlying continuous distribution conditional on covariates, the **multivariate normal**. In all of these papers the authors also specify regression models for the (univariate marginal) response distributions, and so now we turn to the modeling of ordered categorical responses.

Regression Models for an Ordered Categorical Response

Logistic regression is a regression model for a binary response variable. The link between the response, "success" or "failure", to a set of predictor variables or covariates is through a linear regression model for the log-odds for success versus failure. That is, let $\pi(\mathbf{x}_i)$ denote the probability of success for subject i, or for a group of n_i subjects, with covariate vector $\mathbf{x}_i$. Let logit $(\mathbf{x}_i)] = \ln\{\pi(\mathbf{x}_i)/[1 - \pi(\mathbf{x}_i)]\}$ denote the so-called logit for observation i. Then the standard logistic regression model is

$$\text{logit}[\pi(\mathbf{x}_i)] = \beta_0 + \beta'\mathbf{x}_i,$$

where β_0 is the intercept term.

In moving from the case of a binary response to an ordered categorical response there are multiple approaches to defining logits for the response. The three logits that figure most prominently in the biostatistics literature are adjacent-categories logits, continuation-ratio logits, and cumulative logits. For the purposes of exposition we take the number of categories for the response to be J, and hence for each set of logits it is necessary to construct $J - 1$ logits. Let $\pi_j(\mathbf{x}_i)$ denote the probability of response category j, given the covariate vector $\mathbf{x}_i$, and let $\gamma_j(\mathbf{x}_i) = \sum_{k=1}^{j} \pi_k(\mathbf{x}_i)$. Then the respective logits are defined as follows:

Adjacent-categories logits:

$$L_j(\mathbf{x}_i) = \ln\left[\frac{\pi_j(\mathbf{x}_i)}{\pi_{j+1}(\mathbf{x}_i)}\right], \quad j = 1, \ldots, J-1.$$

Continuation-ratio logits:

$$L_j(\mathbf{x}_i) = \ln\left\{\frac{\pi_j(\mathbf{x}_i)}{[1 - \gamma_j(\mathbf{x}_i)]}\right\}, \quad j = 1, \ldots, J-1.$$

Cumulative logits:

$$L_j(\mathbf{x}_i) = \ln\left\{\frac{\gamma_j(\mathbf{x}_i)}{[1 - \gamma_j(\mathbf{x}_i)]}\right\}, \quad j = 1, \ldots, J-1.$$

The definitions of the adjacent-categories logits and the cumulative logits are invariant with respect to reversing the ordering of the response categories, but the high-to-low ordering of the categories gives a different set of continuation ratio logits than are obtained with the low-to-high ordering. The decision of which set of logits to use, and which ordering in the case of continuation-ratio logits, should be based on the substantive questions of interest that are to be addressed in the statistical analysis.

Given a set of logits $L_j(\mathbf{x}_i)$, a general logit linear regression model is of the form

$$L_j(\mathbf{x}_i) = \beta_{0j} + \beta_j'\mathbf{x}_i.$$

Note that there is a separate intercept β_{0j} for each logit. The interpretation of the regression coefficients is specific to the form of the logits modeled. Models where the vectors of regression coefficients β_j are constrained to be equal for all $j = 1, \ldots, J-1$ are referred to as **proportional-odds models**. It is most common to see the proportional-odds assumption used with cumulative logits, and McCullagh [45] has been highly influential in promoting the use of proportional-odds cumulative logit models.

In general, adjacent-categories linear logit models can be estimated using software for fitting loglinear or generalized linear models, and continuation-ratio linear logit models can be estimated using software for fitting logistic regression models. Fitting cumulative linear logit models under the proportional-odds assumption requires specialized software, such as that documented in Stokes et al. [54]. The interested reader is referred to Agresti [3] for discussion of how standard statistical software can be used to fit certain ordered categorical linear logit models.

Example 3: Evaluation of Treatment Effects for Relief of Arthritis Pain, by Gender

It is common in clinical trials of treatments for **pain** relief to record an ordered categorical response variable with categories ranging from no relief to substantial or complete relief. The data used in Table 5 to illustrate and compare the results obtained with the three types of logits are from a clinical study comparing an active treatment to a placebo treatment for arthritis pain [42]. There are three response categories for the degree of pain relief.

Table 5 Response to therapy for arthritis pain

		Response to treatment		
Gender	Treatment	Marked	Some	None
Female	Active	16	5	6
Female	Placebo	6	7	19
Male	Active	5	2	7
Male	Placebo	1	0	10

Table 6 reports the deviance statistics and estimated regression coefficients for each of the three types of logits under the proportional-odds assumption. The models were fit using the lower levels of pain relief as the numerators in odds, and the continuation-ratio logits were formed going from "none" to "marked". That is, the continuation-ratio logits were constructed for the comparisons "none" vs. "some + marked", and "some" vs. "marked". The adjacent-categories logits compare "some" vs. "marked" and "none" vs. "some", and the cumulative logits make the comparisons "none" vs. "some + marked" and "none + some" vs. "marked". The gender variable was coded 0 for "female" and 1 for "male", and the treatment variable was coded 0 for "active" and 1 for "placebo".

The data are congruent with proportional odds for all three types of logits in this case. That is, all three deviance statistics are consistent with good fits to the data, and test statistics for testing the proportional-odds assumption (not reported here) fail to reject it. In general, it is possible that proportional odds will be supported for a subset of the adjacent-categories logits, continuation-ratio logits, and cumulative logits, but not all three. The estimates of $\hat{\beta}_{\text{gender}}$ are all significantly greater than zero (e.g. evaluate $\hat{\beta}_{\text{gender}}/se(\hat{\beta}_{\text{gender}})$ relative to the distribution of a N(0,1) random variable), suggesting that females tended to achieve a greater degree of pain relief. The estimates of $\hat{\beta}_{\text{treatment}}$ are all significantly greater than zero, suggesting that subjects on the active treatment achieved higher levels of pain relief.

Comments

There are close connections between adjacent-categories logit models and the association models of the section on models for association. The connections between the two approaches to modeling ordered categorical data parallel those for loglinear models and logit models for a binary response [21]. Goodman [31] focuses on regression formulations

Table 6 Models summary for Table 5

Logits	Deviance	$\hat{\beta}_{\text{gender}}$ (se)	$\hat{\beta}_{\text{treatment}}$ (se)
Adjacent-categories	3.36	0.741 (0.325)	1.076 (0.293)
Continuation-ratio	4.42	0.997 (0.460)	1.551 (0.409)
Adjacent-categories	2.71	1.319 (0.529)	1.797 (0.473)

of loglinear association models, and Anderson [12] considers a family of models that can be viewed as generalizations of the regression formulation of an RC association model. Anderson's stereotype ordered regression model, is equivalent to

$$\ln\left[\frac{\pi_j(\mathbf{x}_i)}{\pi_{j+1}(\mathbf{x}_i)}\right] = \beta_{0j} + \phi_j \beta^t x_i,$$

where the ϕ_j are ordered and $j = 1, \ldots, J-1$. Note that the stereotype model follows from the general logit linear model when the regression coefficient vectors are assumed to be parallel, i.e. $\beta_j = \phi_j \beta$. In addition, just as RC models can be motivated in terms of multivariate normality, Anderson provides a motivation for stereotype models in terms of ordered multivariate normal distributions. It is, however, important to recognize that neither log multiplicative association models nor stereotype regression models are reliant on normality assumptions. Rather, the point is that the models are also appropriate in those cases where the assumption of latent normality is assumed or reasonable.

The discussed papers by Anderson [12] and McCullagh [45] provide good background on issues that arise in the analysis of an ordered categorical response, and the theoretical perspectives that can be used to motivate the different logit formulations. Among the issues discussed in both papers are: (i) the utility of the models in data analysis; (ii) theoretical results for statistical inference; (iii) the computation of **maximum likelihood** estimates; and (iv) the stochastic ordering of response distributions.

Concluding Remarks

The development and application of models to the analysis of ordered categorical responses continues to be an area in which important contributions are being made. It was noted in concluding the section on models for association that there have been recent developments along the lines of the simultaneous modeling of ordered categorical responses and the association between them. An overlapping area in which there has been much recent activity is the analysis of repeated measures data (*see* **Longitudinal Data Analysis, Overview**) in the form of an ordered categorical response. Such data arise, for example, in longitudinal studies, **crossover** experiments, and studies of families or sibships (*see* **Generalized Linear Models for Longitudinal Data**). General approaches to the analysis of repeated measures data have been applied to such problems, including methodology based on **generalized estimating equations** [37, 40, 55] and methodology based on incorporating **random effects** into the models [8, 27, 38, 40]. The vast majority of this work builds on the models and literature summarized herein.

This entry has focused on models for the analysis of ordered categorical data, emphasizing the study of association and the formulation of regression relationships. There are many important issues that can not be covered here due to space limitations. A few examples include methods for **exact inference** (see, for example, [5]), measures (rather than models) of association (*see*, for example, [1] and [23] and **Association, Measures of**), and inference for association model scores under order restrictions [9, 52] (*see* **Isotonic Regression**).

References

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

[2] Agresti, A. (1988). A model for agreement between ratings on an ordinal scale, *Biometrics* **44**, 539–548.

[3] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[4] Agresti, A. (1992). Modeling patterns of agreement and disagreement, *Statistical Methods in Medical Research* **8**, 201–218.

[5] Agresti, A. (1992). A survey of exact inference for contingency tables, with discussion, *Statistical Science* **7**, 131–153.

[6] Agresti, A. & Kezouh, A. (1983). Association models for multidimensional cross-classifications of ordinal variables, *Communications in Statistics – Theory and Methods* **12**, 1261–1276.

[7] Agresti, A. & Lang, J.B. (1993). Quasi-symmetric latent class models, with application to rater agreement, *Biometrics* **49**, 131–140.

[8] Agresti, A. & Lang, J.B. (1993). A proportional odds model with subject-specific effects for repeated ordered categorical responses, *Biometrika* **80**, 527–534.

[9] Agresti, A., Chuang, C. & Kezouh, A. (1987). Order-restricted score parameters in association models for contingency tables, *Journal of the American Statistical Association* **82**, 619–623.

[10] Aitchison, J. & Silvey, S. (1957). The generalization of probit analysis to the case of multiple responses, *Biometrika* **44**, 131–140.

[11] Andersen, E.B. (1980). *Discrete Statistical Models with Social Science Applications*. North-Holland, Amsterdam.

[12] Anderson, J.A. (1984). Regression and ordered categorical variables, with discussion, *Journal of the Royal Statistical Society, Series B* **46**, 1–30.

[13] Ashford, J.R. & Sowden, R.R. (1957). Multivariate probit analysis, *Biometrics* **26**, 535–546.

[14] Becker, M.P. (1989). Using association models to analyze agreement data: two examples, *Statistics in Medicine* **8**, 1199–1207.

[15] Becker, M.P. (1989). On the bivariate normal distribution and association models for ordinal categorical data, *Statistics and Probability Letters* **8**, 435–440.

[16] Becker, M.P. (1989). Models for the analysis of association in multivariate contingency tables, *Journal of the American Statistical Association* **84**, 1014–1019.

[17] Becker, M.P. (1990). Quasisymmetric models for the analysis of square contingency tables, *Journal of the Royal Statistical Society, Series B* **52**, 369–378.

[18] Becker, M.P. (1990). aximum likelihood estimation of the RC(M) association model, *Applied Statistics* **39**, 152–166.

[19] Becker, M.P. & Agresti, A. (1992). Loglinear modeling of pairwise interobserver agreement on a categorical scale, *Statistics in Medicine* **11**, 101–114.

[20] Becker, M.P. & Clogg, C.C. (1989). Analysis of sets of two-way contingency tables using association models, *Journal of the American Statistical Association* **84**, 142–151.

[21] Bishop, Y.M.M. (1969). Full contingency tables, logits, and split contingency tables, *Biometrics* **25**, 119–128.

[22] Clogg, C.C. (1982). Some models for the analysis of association in multiway contingency tables having ordered categories, *Journal of the American Statistical Association* **77**, 803–815.

[23] Clogg, C.C. & Shihadeh, E.S. (1994). *Statistical Models for Ordinal Variables*. Sage, Thousand Oaks.

[24] Dale, J.R. (1984). Local versus global association for bivariate ordered responses, *Biometrika* **71**, 507–514.

[25] Dale, J.R. (1986). Global cross-ratios for bivariate, discrete, ordered responses, *Biometrics* **42**, 909–917.

[26] Darroch, J.N. & McCloud, P.I. (1986). Category distinguishability and observer agreement, *Australian Journal of Statistics* **28**, 420–428.

[27] Ezzet, F. & Whitehead, J. (1991). A random effects model for ordinal responses from a crossover trial, *Statistics in Medicine* **10**, 901–906. (Comment and Reply: **12** (1993) 2147–2151.

[28] Gilula, Z. & Haberman, S.J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models, *Journal of the American Statistical Association* **83**, 760–771.

[29] Goodman, L.A. (1979). Simple models for the analysis of cross-classifications having ordered categories, *Journal of the American Statistical Association* **74**, 537–552.

[30] Goodman, L.A. (1981). Association models and the bivariate Normal for contingency tables with ordered categories, *Biometrika* **68**, 347–355.

[31] Goodman, L.A. (1983). The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds, *Biometrics* **39**, 149–160.

[32] Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *Annals of Statistics* **13**, 10–69.

[33] Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables, *International Statistical Review* **54**, 243–309.

[34] Graubard, B.I. & Korn, E.L. (1987). Choice of column scores for testing independence in ordered 2× K contingency tables, *Biometrics* **43**, 471–476.

[35] Haberman, S.J. (1974). Loglinear models for frequency tables with ordered classifications, *Biometrics* **29**, 205–220.

[36] Haberman, S.J. (1996). Computation of maximum likelihood estimates in association models, *Journal of the American Statistical Association* **90**, 1438–1446.

[37] Heagerty, P. & Zeger, S.L. (1996). Marginal regression models for clustered ordinal measurements, *Journal of the American Statistical Association* **91**, 1024–1036.

[38] Hedeker, D. & Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis, *Biometrics* **50**, 933–944.

[39] Holland, P.W. & Wang, Y.J. (1987). Dependence functions for continuous bivariate densities, *Communications in Statistics – Theory and Methods* **16**, 863–876.

[40] Kenward, M.G., Lesaffre, E. & Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random, *Biometrics* **50**, 945–953.

[41] Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data, *Statistics in Medicine* **14**, 1341–1352.

[42] Koch, G.G. & Edwards, S. (1988). Clinical efficacy trials with categorical data, in *Biopharmaceutical Statistics for Drug Development*, K.E. Peace, ed. Marcel Dekker, New York, pp. 403–451.

[43] Lang, J.B. & Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical data, *Journal of the American Statistical Association* **89**, 625–632.

[44] Lesaffre, E. & Molenberghs, G. (1991). Multivariate probit analysis: A neglected procedure in medical statistics, *Statistics in Medicine* **10**, 1391–1403.

[45] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[46] Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

[47] Melia, M.B. & Diener-West, M. (1994). Modeling interrater agreement for pathologic features of choroidal melanoma, in *Case Studies in Biometry*, N. Lange,

L. Ryan, L. Billard, D. Brillinger, L. Conquest & J. Greenhouse, eds. Wiley, New York, pp. 323–338.

[48] Molenberghs, G. & Lesaffre, E. (1994). Marginal modeling or correlated ordinal data using a multivariate Plackett distribution, *Journal of the American Statistical Association* **89**, 633–644.

[49] National Cholesterol Education Program (1993). Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults: Summary of the second report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel II), *Journal of the American Medical Association* **269**, 3015–3023.

[50] Pearson, K. & Heron, D. (1913). On theories of association, *Biometrika* **9**, 159–315.

[51] Pron, G.E., Burch, J.D., Howe, G.R. & Miller, A.B. (1988). The reliability of passive smoking histories in a case–control study of lung cancer, *American Journal of Epidemiology* **127**, 267–273.

[52] Ritov, Y. & Gilula, Z. (1991). The order-restricted RC model for ordered contingency tables: estimation and testing for fit, *Annals of Statistics* **19**, 2090–2101.

[53] Srole, L., Langner, T.S., Michael, S.T., Kirkpatrick, P., Opler, M.K. & Rennie, T.A.C. (1978). *Mental Health in the Metropolis: The Midtown Manhattan Study*, Revised Ed. New York University Press, New York.

[54] Stokes, M.E., Davis, C.E. & Koch, G.G. (1995). *Categorical Data Analysis Using the SAS System*. SAS Institute, Cary.

[55] Stram, D.O., Wei, L.J. & Ware, J.H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent data, *Journal of the American Statistical Association* **83**, 631–637.

[56] Tanner, M.A. & Young, M.A. (1985). Modeling agreement among raters, *Journal of the American Statistical Association* **80**, 175–180.

[57] Williamson, J.M., Kim, K. & Lipsitz, S.R. (1995). Analyzing bivariate ordinal data using a global odds ratio, *Journal of the American Statistical Association* **90**, 1432–1437.

[58] Yule, G.U. (1900). On the association of attributes in statistics, *Philosophical Transactions of the Royal Society of London, Series A* **194**, 257–319.

[59] Yule, G.U. (1912). On the methods of measuring association between two attributes (with discussion), *Journal of the Royal Statistical Society* **75**, 579–642.

(*See also* **Polytomous Data**; **Trend Test for Counts and Proportions**)

MARK P. BECKER

Orders of Magnitude

It is often useful to compare the limiting behavior of a function $f(x)$ with some known simple function $g(x)$ as x tends to L:

1. If $f(x)/g(x)$ remains bounded as x tends to L, then we say that $f(x)$ *is at most of the order of g(x)* and we write $f(x) = O[g(x)]$ as $x \to L$.
2. If $f(x)/g(x)$ tends to zero, then we say that *f(x) is of a smaller order than g(x)*, and we write $f(x) = o[g(x)]$ as $x \to L$.
3. If $f(x)/g(x)$ tends to 1 as x tends to L, then we say that $f(x)$ *is asymptotically equal to g(x)*, and we write $f(x) \sim g(x)$ as $x \to L$.

Note that $O(1)$ stands for any bounded function, $o(1)$ stands for any function tending to zero, and $O(x)$ stands for any function which is at most of order x, as x tends to ∞. The above definition can be applied to any sequence a_n by considering $a_n = f(n)$, for $n \in N$.

Listed below are some simple rules for asymptotic calculations:

1. If $f_1(x) = O[g_1(x)]$, and $f_2(x) = O[g_2(x)]$, then $f_1(x) + f_2(x) = O[g_1(x) + g_2(x)]$.
2. $f(x) = o[g(x)]$ implies that $f(x) = O[g(x)]$.
3. If $f_1(x) = O[g_1(x)]$ and $f_2(x) = o[g_2(x)]$, then $f_1(x) f_2(x) = o[g_1(x) g_2(x)]$.
4. The order of magnitude of a sum of a finite number of terms is the largest order of magnitude of the summands.

Example: $o(1) + O(n^{-1/2}) + O(n^{-1}) = o(1)$.
Example: Let $f(x) = \log(1 + x)$. By the Taylor expansion of f at $x = 0$, we have

$$f(x) = x + o(x) = O(|x|) \quad \text{as } x \to 0.$$

Example: $n \log(1 + n^{-1}) = n[n^{-1} + o(n^{-1})] = 1 + o(1)$.

For further examples, see [1]

Reference

[1] Cramér, H. (1946). *Mathematical Methods of Statistics* Princeton University Press, Princeton.

MEI-LING TING LEE

Ornstein–Uhlenbeck Process

The Ornstein–Uhlenbeck process (OUP) is a diffusion process (*see* **Brownian Motion and Diffusion Processes**) – a continuous-time **stochastic process** which satisfies the strong Markov property and has continuous sample paths – which has a drift, towards some mean taken to be zero without loss of generality, which is proportional to its displacement from that mean, i.e. the process has instantaneous mean $-\beta X$ when at X, where $\beta > 0$, and constant instantaneous variance σ^2.

If the process is denoted by $\{X(t), t \in \mathbb{R}\}$, then the conditional distribution is of the form

$$X(s+t)|X(s) = x \sim \mathrm{N}\left\{\exp(-\beta t)x, [1-\exp(-2\beta t)]\frac{\sigma^2}{2\beta}\right\}.$$

The limiting distribution of $X(t)$ is $\mathrm{N}(0, \sigma^2/2\beta)$, and if

$$X(t) \sim \mathrm{N}\left(0, \frac{\sigma^2}{2\beta}\right) \tag{1}$$

for some t, then the process is **stationary** and (1) holds for all t.

The OUP is the only diffusion process that is both stationary and Gaussian.

The stationary process has autocovariance function

$$\gamma(t) = \exp(-\beta t)\frac{\sigma^2}{2\beta}$$

and hence **autocorrelation function**

$$\rho(t) = \exp(-\beta t).$$

The OUP can be obtained as the solution to the following stochastic differential equation:

$$\mathrm{d}X = -\beta X\,\mathrm{d}t + \sigma\,\mathrm{d}W,$$

where $W(t)$ is a Wiener process.

The process generalizes to two or more dimensions; then the conditional distribution is of the form

$$\mathbf{X}_{(s+t)}|\mathbf{X}(s) = \mathbf{x} \sim \mathrm{N}(\exp(\mathbf{B}t)\mathbf{x}, \mathbf{\Lambda} - \exp(\mathbf{B}t)\mathbf{\Lambda}\exp(\mathbf{B}'t)),$$

where $\mathbf{\Lambda}$ and $\mathbf{B}$ are matrix constants, $\mathbf{B}$ is stable [i.e. $\exp(\mathbf{B}t) \to 0$ as $t \to \infty$, or equivalently the **eigenvalues** of $\mathbf{B}$ all have negative real parts], and the limiting distribution is

$$\mathbf{X}_{(t)} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Lambda}).$$

Genesis

The OUP arose originally [27] as a model for the velocity of a particle suspended in a fluid and undergoing Brownian motion. The particle's velocity is assumed to be affected by collisions with the molecules of the fluid, resulting in both random fluctuations and an overall frictional effect. The process representing the position of a particle which has velocity given by an OUP will be referred to as the integrated OUP. The integrated OUP improves on the Wiener process as a model for Brownian motion, in that the Wiener process has nowhere-differentiable sample paths, and hence undefined velocity.

The OUP also arises as the limiting diffusion approximation to certain urn models. In the Ehrenfest urn model, there are $2N$ balls divided between two urns, A and B. At each time step of the process, one of the balls is selected (completely at random) and moved to the other urn. Thus if at time t there are i balls in A, and hence $2N - i$ in B, then at time $t+1$ there are $i-1$ balls in A with probability $i/2N$ and $i+1$ balls in A with probability $1 - i/2N$. As $N \to \infty$ and the interval between steps becomes small, the process converges to an OUP. Details are given in, for example, [9]; Jacobsen [7] gives some alternative urn models leading to the OUP, as part of a historical account of the origins of the OUP.

Applications

Neural Models

An important application of the OUP is in modeling the activity of individual nerve cells. Stein's [23] model assumes that membrane potential in a nerve cell changes discontinuously at random times due to excitatory (positive) and inhibitory (negative) inputs, with exponential decay of the potential between inputs; when the potential exceeds some threshold, the neuron fires. If the individual inputs are small but frequent, then the behavior of the membrane

potential in the Stein model can be well approximated by an OUP. This diffusion approximation has become a basic model in the study of nerve-cell activity (e.g. [14]); Tuckwell [26] gives a self-contained introduction to stochastic models in this area, and describes results for the OUP model. The importance of the interspike interval, the time taken for the potential to reach the level that triggers activity, has led to particular interest in first-passage time problems for the OUP.

Velocity Models

The original use of the OUP, as a model for the velocity of a particle, continues in a number of fields. Stokes et al. [24], for example, consider cell motility effects in physiological processes. They present a model in which an individual microvessel endothelial cell follows an integrated OUP, with parameters dependent on the biochemistry of the medium in which it is moving.

Heubach & Watkins [6] model the rotational velocity of a white blood cell. They show that, in the limit as the number of receptors on the cell becomes large, and in a uniform concentration of chemoattractant, the rotational velocity follows an OUP.

Movement Models

The OUP can also be used for the direct modeling of movement (rather than velocity). Dunn & Gipson [3], for example, use a bivariate OUP to represent the location over time of an animal exhibiting home range behavior as a model with which to interpret data from radio tracking. They also use a higher-dimensional form to model simultaneously the movements of multiple animals. Worton [28] reviews some applications and refinements.

Population Models

The OUP arises as a limiting diffusion approximation in many population processes. Andersson & Djehiche [1] consider a stochastic spatial epidemic model (*see* **Epidemic Models, Spatial**). They show that the differences between their model and its deterministic approximation converge, as the number of subpopulations increases, to an OUP.

Kämmerle [8] and Möhle [18] consider bisexual versions of the Moran and Wright–Fisher models, respectively, and show that in each case the backward process – the number of ancestor-pairs of a given generation – converges when suitably normed to an OUP (*see* **Population Genetics**).

Other Biological Models

Martins [17] uses an OUP to represent the change in (mean) phenotype within a species, under stabilizing selection. Properties of the OUP can thus be used to describe the behavior over time of between-species variability in the phenotype.

Taylor et al. [25] analyze serial CD4 T-cell measurements from the Multicenter **AIDS** Cohort Study. They use the integrated OUP as part of their model, to incorporate **correlation** between measurements in a way that allows the testing of immunologic theories about the dynamics of CD4 cell numbers.

Newell et al. [19] describe an experiment in which individual human subjects, of various age groups, stood still on a platform equipped to measure the downward pressure exerted. The authors found that the dynamics of the center of pressure could be modeled adequately by an OUP, with parameters dependent upon the age group of the subject.

Statistical Applications

In the applications mentioned so far, the OUP is used directly as a model for some process of interest, or emerges as a limiting case or approximation to some such model. Alternatively, the OUP may be a limiting approximation to a statistical quantity.

For example, in **genetic mapping** of quantitative trait loci (QTL), Lander & Botstein [13] consider the LOD score or **likelihood ratio** $Z(t)$ at all locations t in the genome simultaneously. In the limit as the **genetic markers** become dense and the number of individuals tested becomes large, under the **null hypothesis** of no QTLs, the random process $Z(t)$ converges to the square of an OUP. Thus knowledge of the **extreme-value** behavior of the OUP enables testing of that hypothesis. Lander & Botstein assume a normal distribution for the phenotype of interest; Kruglyak & Lander [10] give a nonparametric version of the test, based on the Wilcoxon rank-sum statistic, in which the OUP plays essentially the same role.

Inference

Parameter estimation for an OUP is typically by **maximum likelihood**; we assume for this section that the mean parameter, denoted by μ, is unknown.

If the process is observed continuously over some interval $[0, s]$, say, then the variance parameter σ can be calculated exactly as

$$\sigma^2 = s^{-1} \lim_{n\to\infty} \sum_{j=1}^{2^n} \{X(js2^{-n}) - X[(j-1)s2^{-n}]\}^2.$$

The likelihood for the drift parameter, β, and for the mean, μ, can then be written down – see, for example, Tuckwell [26] or Guttorp [5] – and we can thus obtain maximum likelihood estimates:

$$\hat{\mu} = \left\{ [X(s) - X(0)] \int_0^s X^2(t)\,\mathrm{d}t - \frac{1}{2}[X^2(s) - X^2(0) - \sigma^2 s] \int_0^s X(t)\,\mathrm{d}t \right\} \times \left\{ s \int_0^s X^2(t)\,\mathrm{d}t - \left[\int_0^s X(t)\,\mathrm{d}t \right]^2 \right\}^{-1}$$

and

$$\hat{\beta} = \left\{ [X(s) - X(0)] \int_0^s X(t)\,\mathrm{d}t - \left(\frac{s}{2}\right)[X^2(s) - X^2(0) - \sigma^2 s] \right\} \times \left\{ s \int_0^s X^2(t)\,\mathrm{d}t - \left[\int_0^s X(t)\,\mathrm{d}t \right]^2 \right\}^{-1}.$$

Alternatively, given discrete observations $X(t_1), \ldots, X(t_n)$ on the process, the conditional density for each observation is

$$X(t_i)|X(t_{i-1}) = x \sim \mathrm{N}(\nu_i + \exp[-\beta(t_i - t_{i-1})]x, \phi_i), \quad i = 2, \ldots, n,$$

where

$$\nu_i = \{1 - \exp[-\beta(t_i - t_{i-1})]\}\mu$$

and

$$\phi_i = \{1 - \exp[-2\beta(t_i - t_{i-1})]\}\frac{\sigma^2}{2\beta},$$

and so the log likelihood is of the form

$$\sum_{i=2}^{n} \left(-\frac{1}{2}\ln(\phi_i) - \frac{1}{2}\{x_i - \nu_i - \exp[-\beta(t_i - t_{i-1})]x_{i-1}\}^2/\phi_i \right).$$

If, rather than conditioning on $X(t_1)$, it is assumed that the process is in equilibrium at time t_1, then there is an additional term in the log likelihood

$$-\frac{1}{2}\ln\left(\frac{\sigma^2}{2\beta}\right) - (x_1 - \mu)^2\frac{\beta}{\sigma^2},$$

and the information contained in $X(t_1)$ may be much greater than in other observations. The likelihood can be maximized numerically; if *regular* discrete observations are available, then the relationship with the AR(1) process (see below) may be exploited, and estimation techniques from the **time series** literature are available.

Dunn & Gipson [3], Dunn & Brisbin [2], and Worton [28] consider inference based on the likelihood for discrete observations in higher-dimensional cases, the latter investigating the performance of estimation based on the OUP likelihood in the case where the true model is not an OUP, and using **bootstrapping** to obtain **standard errors**.

Polson & Roberts [20] describe **Bayesian** model selection techniques for diffusions, which allow comparison between an OUP, say, and an alternative diffusion model.

As an alternative to likelihood-based approaches, Kutoyants [11] and Kutoyants & Pilibossian [12] use minimum distance estimation, based on the L_2 and L_1 norms, respectively.

Further Properties

First Passage Times

The first passage time of a stochastic process from x_0 to $a > x_0$, say, is

$$T = \inf_{t\geq 0}\{t : X(t) \geq a | X(0) = x_0\},$$

with the obvious change if $a < x_0$. A time-varying threshold $a(t)$ can also be considered.

The first passage time of the OUP to a fixed threshold is of particular interest since it represents the interspike interval in neural models (see above). Ricciardi & Sato [22] consider the first passage time for the OUP, and Giorno et al. [4] consider the asymptotic case for the OUP and related diffusions. Lefebvre considers the two-dimensional case [16] and the integrated OUP [15].

Related Processes

An OUP $X(t)$ is related to the Wiener process $W(t)$ by

$$X(t) = \exp(-\beta t)W[\exp(2\beta t)],$$
$$W(t) = t^{1/2}X\left(\frac{\ln t}{2\beta}\right).$$

An OUP observed at regular discrete time intervals (of length 1, without loss of generality) is of the form

$$X_{t+1} = \exp(-\beta t)X_t + \varepsilon_t,$$

where the ε_ts are independent $\mathrm{N}(0, [1\exp(-2\beta)]\ \sigma^2/2\beta)$ random variables, and so the process is just a Gaussian first-order autoregression or AR(1) process (*see* **ARMA and ARIMA Models**).

A spatial discretization of the (integrated) OUP, in which velocity, and hence position, is restricted to integer values, is developed by Renshaw [21].

References

[1] Andersson, H. & Djehiche, B. (1995). Limit theorems for multitype epidemics, *Stochastic Processes and Their Applications* **56**, 57–75.

[2] Dunn, J.E. & Brisbin, I.L., Jr (1985). Characterizations of the multivariate Ornstein Uhlenbeck process in the context of home range analysis, in *Statistical Theory and Data Analysis*, K. Matusita, ed. Elsevier, Amsterdam, pp. 181–205.

[3] Dunn, J.E. & Gipson, P. (1977). Analysis of radiotelemetry data in studies of home range, *Biometrics* **33**, 85–101.

[4] Giorno, V., Nobile, A.G. & Ricciardi, L.M. (1990). On the asymptotic-behavior of first-passage-time densities for one-dimensional diffusion-processes and varying boundaries, *Advances in Applied Probability* **22**, 883–914.

[5] Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman & Hall, London.

[6] Heubach, S. & Watkins, J. (1995). A stochastic model for the movement of a white blood cell, *Advances in Applied Probability* **27**, 443–475.

[7] Jacobsen, M. (1996). Laplace and the origin of the Ornstein–Uhlenbeck process, *Bernoulli* **2**, 271–286.

[8] Kämmerle, K. (1989). Looking forwards and backwards in a bisexual Moran model, *Journal of Applied Probability* **27**, 880–885.

[9] Karlin, S. & Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

[10] Kruglyak, L. & Lander, E.S. (1995). A nonparametric approach for mapping quantitative trait loci, *Genetics* **139**, 1421–1428.

[11] Kutoyants, Y. (1991). Minimum distance parameter-estimation for diffusion type observations, *Comptes Rendus de l'Académie des Sciences, Série I (Mathématique)* **312**, 637–642.

[12] Kutoyants, Y. & Pilibossian, P. (1994). On minimum L_1-norm estimation of the parameter of the Ornstein–Uhlenbeck process, *Statistics and Probability Letters* **20**, 117–123.

[13] Lander, E.S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* **121**, 185–199.

[14] Lánský, P. (1984). On approximations of Stein's neuronal model, *Journal of Theoretical Biology* **107**, 631–647.

[15] Lefebvre, M. (1989). Moment generating function of a first hitting place for the integrated Ornstein–Uhlenbeck process, *Stochastic Processes and Their Applications* **32**, 281–287.

[16] Lefebvre, M. (1996). On the first passage time probability problem for bidimensional Ornstein–Uhlenbeck processes, *Sankhyā, Series A* **58**, 179–185.

[17] Martins, E.P. (1994). Estimating the rate of phenotypic evolution from comparative data, *American Naturalist* **144**, 193–209.

[18] Möhle, M. (1994). Forward and backward processes in bisexual models with fixed population sizes, *Journal of Applied Probability* **31**, 309–332.

[19] Newell, K.M., Slobounov, S.M., Slobounova, E.S. & Molenaar, P.C.M. (1997). Stochastic processes in postural center-of-pressure profiles, *Experimental Brain Research* **113**, 158–164.

[20] Polson, N.G. & Roberts, G.O. (1994). Bayes factors for discrete observations from diffusion processes, *Biometrika* **81**, 11–26.

[21] Renshaw, E. (1987). The discrete Uhlenbeck–Ornstein process, *Journal of Applied Probability* **24**, 908–917.

[22] Ricciardi, L.M. & Sato, S. (1988). First-passage-time density and moments of the Ornstein–Uhlenbeck process, *Journal of Applied Probability* **25**, 43–57.

[23] Stein, R.B. (1965). A theoretical analysis of neuronal variability, *Biophysical Journal* **5**, 173–194.

[24] Stokes, C.L., Lauffenburger, D.A. & Williams, S.K. (1991). Migration of individual microvessel endothelial cells: stochastic model and parameter measurement, *Journal of Cell Science* **99**, 419–430.

[25] Taylor, J.M.G., Cumberland, W.G. & Sy, J.P. (1994). A stochastic model for analysis of longitudinal AIDS data, *Journal of the American Statistical Association* **89**, 727–736.

[26] Tuckwell, H.C. (1988). *Introduction to Theoretical Neurobiology*, Vol. 2. *Nonlinear and Stochastic Theories*. Cambridge University Press, Cambridge.

[27] Uhlenbeck, G.E. & Ornstein, L.S. (1930). On the theory of Brownian motion, *Physical Review* **36**, 823–841.

[28] Worton, B.J. (1995). Modelling radio-tracking data, *Environmental and Ecological Statistics* **2**, 15–23.

PAUL G. BLACKWELL

Orthoblique Rotation

Orthoblique or *Harris-Kaiser Rotation* [1, 2] is a nonquartic method of performing an **oblique rotation** of a matrix $\mathbf{V}$ of dimension $(p \times k)$ made up of vectors associated with **principal components analysis** or **factor analysis** in order to transform these quantities into new variables by the relationship $\mathbf{B} = \mathbf{V\Theta}$ such that $\mathbf{B}$ will approximate a **simple structure**. The matrix $\mathbf{B}$ is of dimension $(p \times k)$ and the matrix $\mathbf{\Theta}$ is of dimension $(k \times k)$ (*see* **Rotation of Axes**). The principal feature of Orthoblique rotation is that the solution is obtained from the product of a number of orthogonal matrices which are either diagonal or orthonormal. One of these matrices is the solution of an **orthogonal rotation** such as **Varimax**, making this a two-step rotation. Another is a diagonal matrix made up of the characteristic roots (*see* **Eigenvalue**) associated with the retained characteristic vectors (*see* **Eigenvector**). This latter matrix is raised to a power which essentially affects the normalization of the original vectors. A power of zero implies a model where the clusters are independent and will be a definite oblique rotation. Higher powers will approach an orthogonal rotation. At various times, **Quartimax**, Equimax, and Varimax have each been suggested for the initial rotation, but Hakstian & Abell concluded that the effect of this choice is far outweighed by the choice of power.

For the Decathlon example given in the article, **Rotation of Axes**, the Orthoblique solutions for powers of 0.35 and 0.75 are given in Table 1 along with the original principal component characteristic vectors.

References

[1] Hakstian. A.R. & Abell, R.A. (1974). A further comparison of oblique factor transformation methods, *Psychometrika* **39**, 429–444.

[2] Harris, C.W. & Kaiser, H.F. (1964). Oblique factor analytic solutions by orthogonal transformations, *Psychometrika* **29**, 347–362.

(*See also* **Axes in Multivariate Analysis**)

J. EDWARD JACKSON

Table 1 Decathlon Data: characteristic and orthoblique-rotated vectors

	Characteristic vectors				Orthoblique power = 0.35				Orthoblique power = 0.75			
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
100 m run	0.69	0.22	−0.52	−0.21	0.89	0.05	0.02	−0.16	0.88	0.11	0.09	−0.14
Long jump	0.79	0.18	−0.19	0.09	0.56	0.09	0.43	−0.05	0.60	0.16	0.46	−0.02
Shotput	0.70	−0.53	0.05	−0.18	0.14	0.81	0.10	−0.13	0.20	0.82	0.16	−0.13
High jump	0.67	0.13	0.14	0.40	0.11	0.04	0.74	0.04	0.20	0.12	0.73	0.07
400 m run	0.62	0.55	−0.08	−0.42	0.81	0.02	−0.05	0.43	0.80	0.05	0.02	0.45
110 m hurdle	0.69	0.04	−0.16	0.35	0.30	0.04	0.61	−0.21	0.37	0.12	0.61	−0.18
Discus	0.62	−0.52	0.11	−0.23	0.10	0.81	0.03	−0.06	0.14	0.81	0.09	−0.06
Pole vault	0.54	0.09	0.41	0.44	−0.19	0.09	0.79	0.20	−0.08	0.15	0.75	0.22
Javelin	0.43	−0.44	0.37	−0.24	−0.14	0.76	0.02	0.17	−0.09	0.74	0.07	0.16
1500 m run	0.15	0.60	0.66	−0.28	0.02	−0.03	0.07	0.93	0.04	−0.05	0.08	0.93

Orthogonal Designs

In the design of experiments, the term *orthogonal* is used widely and in different contexts. Orthogonal *designs* is a term generally used by statisticians in the context of experiments in which a number of treatments (with or without a **factorial** structure) are to be compared and in which it is desirable to eliminate the variability due to nuisance factors such as blocks, rows, columns, etc. This term, in a very different context, has been used, for example, by Geramita & Wallis [1].

Orthogonal Designs in Comparative Experiments

In comparative experiments two factors (say, treatments and blocks) are said to be *orthogonal* if and only if the condition of *proportional frequency* is satisfied. This condition can be explained as follows. Suppose the two factors involved are A and B, where A has a levels and B has b levels. Furthermore, let n_{ij} be the number of times level i of A appears with level j of B, let $n_{i\cdot}$ be the number of times level i of A appears in the whole design, and let $n_{\cdot j}$ be the number of times level j of B appears in the whole design: $i = 1, \ldots, a$; $j = 1, \ldots, b$. Then A and B are orthogonal if and only if the condition $n_{ij} \propto n_{i\cdot} n_{\cdot j}$ holds for all values of i, j. In particular, if each level of A appears equally often with each level of B, the condition trivially holds. The simplest examples of orthogonal design are **randomized complete blocks designs**, where each of the v treatments under comparison appears precisely once in each block.

An advantage of this type of orthogonality is that if A and B are orthogonal, then under a standard **additive model**, the *best linear* **unbiased** *estimator* of any **contrast** among the levels of A is uncorrelated with the best linear unbiased estimator of any contrast among the levels of B. As a consequence, the sums of squares due to A and B in the **analysis of variance** can be partitioned orthogonally.

Similar properties hold for higher order layouts, involving more than two factors. For example, in the case of **Latin square designs** with s treatments, each treatment appears in each row and each column precisely once. The three factors – treatments, rows, and columns – are mutually orthogonal; that is, treatments are orthogonal to each of rows and columns, and the rows are orthogonal to columns.

Latin squares can be generalized to **Graeco–Latin squares**. Two $s \times s$ Latin squares are said to be orthogonal if, when one of the squares is superimposed on the other, each of the s^2 ordered pairs of symbols from the two separate squares appears once in the superimposed arrangement, called a Graeco-Latin square. In a Graeco-Latin square, any pair of the factors, rows, columns, symbols of the first square, and symbols of the second square are orthogonal.

A set of Latin squares of the same order is said to form a set of mutually orthogonal Latin squares if each pair in the set is orthogonal. If a set of mutually orthogonal Latin squares contains three or more squares, then by superimposing these one over the others, a generalization of Graeco-Latin square is obtained, which may be called a hyper-Graeco-Latin square. A further generalization is provided by orthogonal arrays. An orthogonal array of size N, s symbols, k constraints, and index t is an $k \times N$ array having s symbols with the property that in any $k \times t$ subarray, every s^t ordered t-plets occurs equally often (say, λ times each) as a column. The integer λ is called the *index* of the array. It is easily seen that a Latin square of order s is equivalent to an orthogonal array of size s^2, s symbols, three constraints, strength two and index unity. In general, one can convert an orthogonal array to an orthogonal multifactor design by identifying the rows of the orthogonal array with the factors of the design.

Orthogonality in Response Surface Designs

In **response surface** experiments, it is often assumed that the expected response to the quantitative input variables is a smooth function; say, a polynomial. In particular, suppose that the expected response to the quantitative variables $x_1, x_2, \ldots, x_p$ is a linear function of p unknown parameters. A design for fitting this function is said to be orthogonal if the columns of the matrix of input variables are mutually orthogonal. In an orthogonal response surface design, the **least squares** estimators of the parameters of the surface are mutually uncorrelated. Furthermore, the least squares estimator of any parameter depends only on the values in that column of the matrix of **explanatory variables** and the data.

Orthogonal Factorial Structure

In experiments where the treatments have a factorial structure (*see* **Factorial Experiments**), it is often desirable to have designs in **incomplete blocks** such that the usual least squares estimators of factorial effects belonging to different main effects and interactions are mutually uncorrelated. Such designs are called designs with orthogonal factorial structure. Details on these are available in Gupta & Mukerjee [2].

Other Orthogonal Designs

Geramita & Wallis [1] define an orthogonal design of order n and type $(s_1, \ldots, s_t)$, where the s_is are positive integers, as an $n \times n$ matrix A with entries from the set $\{0, \pm x_1, \ldots, \pm x_t\}$, the x_is being commuting indeterminates, such that

$$AA' = \left(\sum_{i=1}^{t} s_i x_i^2\right) I_n,$$

where A' is the transpose of A and I_n is an identity matrix of order n. These orthogonal designs can be regarded as generalizations of Hadamard matrices. For a review on Hadamard matrices, see Hedayat & Wallis [3].

References

[1] Geramita, A.V. & Wallis, J.S. (1979). *Lecture Notes in Pure and Applied Mathematics*, Vol. 45: *Orthogonal Designs: Quadratic Forms and Hadamard Matrices*. Marcel Dekker, New York.

[2] Gupta, S. & Mukerjee, R. (1989). *Lecture Notes in Statistics*, Vol. 59: *A Calculus for Factorial Arrangements*. Springer-Verlag, Berlin.

[3] Hedayat, A. & Wallis, W.D. (1978). Hadamard matrices and their applications, *Annals of Statistics* **6**, 1184–1238.

(*See also* **Orthogonality**)

ALOKE DEY

Orthogonal Rotation

Given a matrix $\mathbf{V}$ of dimension $(p \times k)$ often consisting of a set of k vectors defining a set of principal components or factors, a new set of transformed variables may be obtained by a rotation of $\mathbf{V}$; namely, $\mathbf{B} = \mathbf{V}\boldsymbol{\Theta}$. $\mathbf{V}$ is often the **factor loading matrix** or factor matrix from the initial step in a **principal components analysis** or a **factor analysis**. Such a rotation is said to be *orthogonal* if the resultant rotated axes in the vector space are at right angles to each other (*see* **Orthogonality**). Most rotation procedures are designed to approximate **simple structure** (*see* **Rotation of Axes**). $\boldsymbol{\Theta}$, a matrix of dimension $(k \times k)$, defines the angles of rotation and the resulting matrix, $\mathbf{B}$, of dimension $(p \times k)$, contains the new rotated vectors defining the transformed variables.

A great many of the orthogonal rotation schemes belong to a class called *Orthomax* rotation. These start with the general expression:

$$Q = \sum_{j=1}^{k} \left[\sum_{i=1}^{p} b_{ij}^4 - \frac{c}{p} \left(\sum_{i=1}^{p} b_{ij}^2 \right)^2 \right],$$

where p is the number of original variables, k is the number of retained components or factors, b_{ij} are the coefficients of the vectors defining the rotation, and c is an arbitrary constant. As the concept of a simple structure requires the b_{ij} to be as large or small as possible, these rotation methods are designed to determine the b_{ij} such that Q is maximized. These are sometimes referred to as *quartic* solutions since they involve fourth powers of the coefficients.

If $c = 1$, then the resulting solution is called **Varimax rotation** in which the sums of squares of $\mathbf{B}$ are maximized *columnwise*. **Quartimax rotation** is obtained by setting $c = 0$ and maximizes the sums of squares of $\mathbf{B}$ *rowwise*. A compromise, *Equimax*, which maximizes the sums of squares across both rows and columns is obtained by setting $c = k/2$. Standard errors for the vector coefficients produced by Orthomax rotations were given by Archer & Jennrich [1].

A method not related to Orthomax is the Minimum Entropy Solution [3], which minimizes:

$$H = -\sum_{i=1}^{p} \sum_{j=1}^{k} b_{ij}^2 \ln(b_{ij}^2).$$

A method of obtaining a single rotation to simultaneously obtain a simple structure for two sets of vectors was obtained by Hakstian [2].

References

[1] Archer, C.O. & Jennrich, R.I. (1973). Standard errors for rotated factor loadings, *Psychometrika* **38**, 581–592.

[2] Hakstian, A.R. (1976). Two-matrix orthogonal rotation procedures, *Psychometrika* **41**, 267–272.

[3] McCammon, R.B. (1966). Principal component analysis and its application in large-scale correlation studies, *Journal of Geology* **74**, 721–733.

(*See also* **Axes in Multivariate Analysis**)

J. EDWARD JACKSON

Orthogonality

In the study of a system in which several factors contribute to a net "effect" it is convenient, though not always possible, to consider the contribution of each factor independently of the others. Coordinate geometry is a trivial example – any point on a plane can be specified in terms of its x and y coordinates, and if we move a point around by varying its x coordinate, its y coordinate is unaffected. This is because movement in the x direction is *orthogonal* to that in the y direction. The orthogonality (Greek *orthos* = straight; erect) of two entities implies that they are, in some sense to be described below, perpendicular to each other. In applied statistics, orthogonality may loosely be regarded as a descriptive term for the ability to disentangle individual effects.

We start with a brief overview of the relevant formal mathematical notions; this provides a basis for referral in later sections. The reader wishing to skip the mathematical detail in the next few paragraphs may find that the subsequent section "Orthogonal Vectors and Matrices" helps to fix ideas in a geometrical context. The section "Orthogonality in Experimental Design" is undoubtedly the most important one from a biostatistical viewpoint; other sections outline more peripheral material which may be encountered in the literature.

Mathematical Definition

Orthogonality arises most naturally in the context of inner product spaces. An inner product is a suitably-behaved scalar (and possibly complex) valued, linear operation (usually denoted by $\langle \cdot, \cdot \rangle$) between pairs of elements in a vector space $\mathcal{V}$. Elements x and y are said to be *orthogonal*, with respect to the space $\mathcal{V}$ and inner product $\langle \cdot, \cdot \rangle$, if $\langle x, y \rangle = 0$.

It is natural to think of elements of $\mathcal{V}$ as having a "size" or "length". This is achieved by defining a *norm* on the space, which has length-like properties. The norm of x is usually denoted by $||x||$: we are solely interested in that defined by $||x|| = \langle x, x \rangle^{1/2}$. If x and y are orthogonal and both have unit length, they are called *orthonormal*.

A subset $\mathcal{S}$ of elements of $\mathcal{V}$ is called a *subspace* of $\mathcal{V}$ if $\mathcal{S}$ is itself an inner product space with the same addition and scalar multiplication defined as in $\mathcal{V}$. The set of elements of $\mathcal{V}$ which are orthogonal to every element of $\mathcal{S}$ is denoted by $\mathcal{S}^{\perp}$. A fundamental result (the *projection theorem*) says that if x is in $\mathcal{V}$ and $\mathcal{S}$ is a subspace of $\mathcal{V}$, then x can be uniquely written as $x = x_1 + x_2$, where $x_1 \in \mathcal{S}$ and $x_2 \in \mathcal{S}^{\perp}$. x_1 is the *orthogonal projection* of x onto $\mathcal{S}$, and satisfies $||x - x_1|| = \inf[||x - y|| : y \in \mathcal{S}]$. Furthermore, the mappings $x \mapsto x_1$ and $x \mapsto x_2$ are linear.

An *orthonormal basis* for $\mathcal{V}$ is a (possibly infinite and uncountable) collection $\{e_i\}$ of elements of $\mathcal{V}$ which are mutually orthonormal and which *span* $\mathcal{V}$ (in other words, every element of $\mathcal{V}$ can be expressed as a linear combination of the $\{e_i\}$). If $\{e_i\}$ is such a basis, define $\hat{a}_i(x) = \langle x, e_i \rangle$ for all $x \in \mathcal{V}$ [so that $x = \sum_i \hat{a}_i(x) e_i$], and denote by $\overline{\hat{a}_i(x)}$ the complex conjugate of $\hat{a}_i(x)$. Then the *Parseval relation* states that, for $x, y \in \mathcal{V}$,

$$\langle x, y \rangle = \sum_i \hat{a}_i(x) \overline{\hat{a}_i(y)}. \tag{1}$$

In particular,

$$||x||^2 = \sum |\hat{a}_i(x)|^2. \tag{2}$$

The latter result is especially important because it offers a means of assessing the relative contributions of each e_i to the magnitude of x; alternatively, it provides a decomposition of the magnitude of x into individual effects arising from each of its components.

Our discussion of inner product spaces ends here; it is necessarily brief and details have been omitted in order to present the most relevant results clearly and concisely. Thorough treatments may be found in the many available texts on real and complex analysis, e.g. [12].

Orthogonal Vectors And Matrices

The theory of inner product spaces is most easily visualized in terms of standard k-dimensional Euclidean space. Let $\mathbf{u} = (u_1, \ldots, u_k)'$ and $\mathbf{v} = (v_1, \ldots, v_k)'$ be vectors in $\mathbb{R}^k$, and define their inner product to be the standard vector product

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}'\mathbf{v} = \sum_{i=1}^{k} u_i v_i. \tag{3}$$

The norm is thus defined by

$$||\mathbf{u}|| = \left(\sum_{i=1}^{k} u_i^2 \right)^{1/2}, \qquad (4)$$

which justifies our thinking of it as a measure of "length". With this inner product, two vectors are orthogonal if they are perpendicular to each other. The most obvious orthonormal basis for $\mathbb{R}^k$ is the Cartesian one; the Parseval relation, (2), is then Pythagoras's theorem in k dimensions.

An *orthogonal matrix* is a square matrix whose columns, considered as vectors in Euclidean space, are mutually orthonormal. Orthogonal matrices can be regarded as rotations and/or reflections of axes in Euclidean space, as they map any one set of mutually perpendicular vectors into another. Orthogonal matrices play an important role in multivariate analysis (*see* **Multivariate Analysis, Overview**), because any set of n measurements on each of k ordinal variates may be represented as an $n \times k$ data matrix $\mathbf{X}$, regarded as a set of n points in k-dimensional space. Often interest centers on finding a subspace of dimension $p < k$ which contains most of the information in $\mathbf{X}$; this amounts to finding an appropriate set of axes on which to plot the data, or alternatively to finding an orthogonal matrix, of dimension $k \times k$, which will map the data points to their positions within the coordinate system defined by these axes. Perhaps the most important tool for such purposes is the *singular value decomposition*, which allows any matrix to be factorized into orthogonal components. Statistical applications of this decomposition are discussed in [5].

In experimental design, a design matrix $\mathbf{X}$ (*see* **Analysis of Variance**) which is not square is usually described as being orthogonal if its columns are mutually orthogonal; alternatively, $\mathbf{X}$ is orthogonal if $\mathbf{X}'\mathbf{X}$ is either a diagonal matrix or block diagonal with blocks corresponding to submodels of interest.

Random Variables

Here we consider a **random variables** with finite second moments. An inner product may be defined as

$$\langle X, Y \rangle = \mathrm{E}(XY)$$

under which the norm of a random variable is its root mean square. X and Y are orthogonal if $\mathrm{E}(XY) = 0$. If $\mathrm{E}(X)$ or $\mathrm{E}(Y)$ is zero, then orthogonality equates to lack of **correlation**; hence, if X and Y are uncorrelated, then the random variables $X^* = X - \mathrm{E}(X)$ and $Y^* = Y - \mathrm{E}(Y)$ are orthogonal. Some authors describe uncorrelated random variables as being orthogonal regardless of whether or not either of them has zero mean; although this is technically incorrect according to the definition given above, the practice is sufficiently widespread that the point is worth mentioning to avoid confusion.

Regression and Conditional Expectation

Suppose we wish to approximate one random variable, Y, by a linear function of k others, say $X_1, \ldots, X_k$. Denote by $\mathcal{S}$ the space of all linear functions of the $\{X_i\}$, i.e. $\mathcal{S} = (Z : Z = \sum_{i=1}^{k} a_i X_i, a_1, \ldots, a_k \in \mathbb{R})$. Suppose $\tilde{Y}$ is the orthogonal projection of Y onto $\mathcal{S}$; then, by the projection theorem, $\tilde{Y}$ minimizes $||Y - \tilde{Y}||$ (alternatively, $||Y - \tilde{Y}||^2$), so it is the linear predictor with the *minimum* **mean square error** (MMSE). If one of the predictors is a constant ($X_1 \equiv 1$, say), then, because $Y - \tilde{Y}$ is orthogonal to all the predictors, we must have $\mathrm{E}[1 \times (Y - \tilde{Y})] = 0$, so the prediction error is a zero-mean random variable.

A similar interpretation can be given to the idea of conditional expectation. Here the aim is to approximate Y by *any* function of the $\{X_i\}$, not just linear functions; so now we are seeking our predictor in the subspace $\mathcal{T} = [Z : Z = h(X_1, \ldots, X_k), h : \mathbb{R}^k \mapsto \mathbb{R}]$. In this case, the function yielding the MMSE predictor is the conditional expectation of Y given $X_1, \ldots, X_k$.

A useful introduction to these ideas is given in [7].

Orthogonal Processes

In **time series analysis**, reference is frequently made to *orthogonal processes* (or *processes with orthogonal increments*). Orthogonal processes are **stochastic processes** in continuous time, $W(t)$, say, such that when I_1 and I_2 are disjoint intervals on the real line,

$$\mathrm{cov}\left[\int_{I_1} \mathrm{d}W(t), \int_{I_2} \mathrm{d}W(t) \right] = 0. \qquad (5)$$

Note that this definition uses the "uncorrelated" interpretation of orthogonality of random variables

rather than its strict definition. The Wiener process (*see* **Brownian Motion and Diffusion Processes**) is an example of such a process. Orthogonal processes arise in the specification of continuous-parameter time series models and in the Spectral Representation Theorem (*see* **Spectral Analysis**). The ideas generalize straightforwardly to spatial processes.

Orthogonality in Experimental Design

For the applied statistician, the most important role orthogonality has to play is in isolating effects due to different factors (*see* **Experimental Design**). To explore this, we start by examining linear regression, as this is perhaps the easiest scenario to visualize. Some points related to the analysis of variance and to **factorial experiments** are then discussed, and finally the ideas are extended to general likelihood-based estimation.

Linear Regression

Suppose we have a dependent variate y, which we wish to regress on p covariates $x_1, \ldots, x_p$ (*see* **Multiple Linear Regression**). A sample of n observations on each variate is taken; the x-data are assembled into an $n \times p$ data matrix $\mathbf{X}$ (the *design matrix*), and the y-data are put into a $n \times 1$ vector $\mathbf{y}$. The regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a vector of uncorrelated errors with equal variance. If $\tilde{\mathbf{y}}$ denotes the fitted value of $\mathbf{y}$, then the standard technique for fitting the model (*see* **Least Squares**) is to minimize the residual sum of squares $\sum(y - \tilde{y})^2 = (\mathbf{y} - \tilde{\mathbf{y}})'(\mathbf{y} - \tilde{\mathbf{y}}) = ||\mathbf{y} - \tilde{\mathbf{y}}||^2$, if the vectors are considered as points in $\mathbb{R}^n$ and the norm and inner product are the standard Euclidean ones defined in (3) and (4). Columns of $\mathbf{X}$ also represent points in $\mathbb{R}^n$, and the span of all linear combinations of these columns is a subspace, $\mathcal{S}$ say, which is a hyperplane of dimension equal to the rank of $\mathbf{X}$. Each value of $\boldsymbol{\beta}$ defines a point in this subspace.

Now consider the Normal equations (*see* **Multiple Linear Regression**) for the estimation of the parameter vector $\boldsymbol{\beta}$:

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \tag{6}$$

If the matrix $(\mathbf{X}'\mathbf{X})$ is diagonal, then (6) has the trivial solution

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ij} y_i}{\sum_{i=1}^{n} x_{ij}^2}, \quad j = 1, \ldots, p. \tag{7}$$

In this case no element of $\hat{\boldsymbol{\beta}}$ depends on any other; hence, if we decide to hold one or more β values fixed, then estimates of the others are not affected. Furthermore, the information matrix for $\boldsymbol{\beta}$ is diagonal (*see* **Information Matrix**), so that individual parameter estimates are uncorrelated, and tests of the significance of individual parameters are independent of the values taken by other parameters. See [13, Section 3.5] for more detail on the advantages that accrue from an orthogonal design matrix in linear regression.

It is instructive to link this with concepts introduced earlier. In a geometrical context, $(\mathbf{X}'\mathbf{X})^{-1}$ will be a diagonal matrix if and only if its columns, regarded as n-dimensional vectors, are mutually orthogonal. The Parseval relation then gives us

$$||\mathbf{y}||^2 = \sum_{j=1}^{p} \hat{\beta}_j^2 ||\mathbf{x}_j||^2 + ||\mathbf{e}||^2, \tag{8}$$

where $\mathbf{e}$ is the vector of residuals from the fitted model, and $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. This is just a decomposition of the sum of squares into individual model components and residuals.

There is also a close parallel with the ideas discussed earlier in the section "Random Variables", for we have

$$\mathbf{x}_j'\mathbf{x}_k = \sum_{i=1}^{n} x_{ij} x_{ik} = n\hat{\mathrm{E}}(X_j X_k),$$

the notation $\hat{\mathrm{E}}(\cdot)$ being used to denote a sample estimate of the expectation. The space considered here is thus effectively the sample version of that considered previously. In particular, if a constant term is included as a linear predictor (giving rise to a column of ones in $\mathbf{X}$) and the linear predictors are mutually orthogonal, then all nonconstant predictors must have zero mean and the sample correlation matrix must be diagonal.

It is unusual to find an example of linear regression where the design matrix is orthogonal. It may be possible to find a set of orthogonal linear combinations of the original variables (using **principal components analysis**, for example), but this is generally only useful in modeling if the principal components themselves have some physical interpretation. One exception to this general rule is in **polynomial regression**, where terms of successively higher degree in a predictor variable X may be parameterized in such a way as to ensure that columns of the design matrix corresponding to each power of X are mutually orthogonal. We seek a sequence $[f_p(\cdot) : p = 0, 1, 2, \ldots]$ such that f_p is a polynomial of degree p and

$$\sum_{i=1}^{n} f_p(x_i) = 0, \quad p = 1, 2, \ldots,$$

$$\sum_{i=1}^{n} f_p(x_i) f_q(x_i) = 0, \quad p \neq q.$$

Setting the coefficient of x^p in $f_p(\cdot)$ to be unity, we obtain a set of simultaneous equations for the coefficients in each $f_p(\cdot)$. For example, we have

$$\begin{aligned} f_1(x) &= x - \overline{x}, \\ f_2(x) &= x^2 + a_{21}x + a_{20}, \end{aligned} \tag{9}$$

where

$$a_{21} = \frac{\sum x_i^2(\overline{x} - x_i)}{\sum x_i^2 - n\overline{x}^2},$$

$$a_{20} = -\left(a_{21}\overline{x} + n^{-1}\sum x_i^2\right).$$

Further discussion of orthogonal polynomials in this context may be found in [13, Section 8.2].

We now present an example to clarify the ideas presented so far. Figure 1 is a scatterplot showing a clear linear relationship between energy requirement and body weight for a sample of 64 grazing sheep in Australia. Let us fit constant, linear and quadratic models to the data, using both a "naive" and an orthogonal parameterization. Denoting by y_i the energy requirement for the ith individual, and by x_i its body weight ($i = 1, \ldots, 64$), the two parameterizations are:

$$y_i = \beta_0^{(1)} + \beta_1^{(1)} x_i + \beta_2^{(1)} x_i^2 + \varepsilon_i, \tag{10}$$

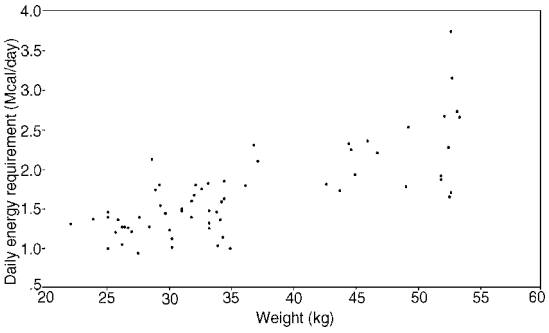

Figure 1 Daily energy requirement versus body weight for a sample of 64 sheep. This is data set no. 241 from [6]. The mean energy requirement is 1.69 Mcal/day, and the mean weight is 35.94 kg

$$\begin{aligned} y_i = \beta_0^{(2)} &+ \beta_1^{(2)}(x_i - 35.9359) \\ &+ \beta_2^{(2)}(x_i^2 - 78.2152x + 1430.4947) + \varepsilon_i. \end{aligned} \tag{11}$$

(the coefficients in the second parameterization being obtained using Eq. (9)). The fitted constant, linear and quadratic models under each parameterization are shown in Table 1, together with P values for each parameter. Notice the following:

1. The parameter estimates, and assessment of the significance of model components, differ under the first parameterization, depending on which model is fitted. In particular, for the quadratic model no components are significant, so the misspecification of the model has serious consequences. Under the second parameterization there is no such ambiguity, and individual parameter estimates are unchanged by the inclusion of other terms in the model. This is convenient for computation, as we only need to estimate each parameter once.
2. The equation of the fitted model is the same in both cases. This is because of the uniqueness of the least squares predictor, as guaranteed by the projection theorem.

The full benefits of orthogonality here result from the appearance of the quantity $\mathbf{X}'\mathbf{X}$ in the normal equations (6), which itself is a consequence of the least squares approach to the estimation. It is clear that these same properties of an orthogonal design will continue to hold as long as estimation is carried

Table 1 Parameter estimates and significance levels for models fitted to sheep data

	Parameterization 1		Parameterization 2	
Model	Estimate	P value	Estimate	P value
Constant	$\hat{\beta}_0^{(1)} = 1.693$	$<5 \times 10^{-5}$	$\hat{\beta}_0^{(2)} = 1.693$	$<5 \times 10^{-5}$
Linear	$\hat{\beta}_0^{(1)} = 0.133$	0.4640	$\hat{\beta}_0^{(2)} = 1.693$	$<5 \times 10^{-5}$
	$\hat{\beta}_1^{(1)} = 0.0434$	$<5 \times 10^{-5}$	$\hat{\beta}_1^{(2)} = 0.0434$	$<5 \times 10^{-5}$
Quadratic	$\hat{\beta}_0^{(1)} = 0.941$	0.3257	$\hat{\beta}_0^{(2)} = 1.693$	$<5 \times 10^{-5}$
	$\hat{\beta}_1^{(1)} = -7.779 \times 10^{-4}$	0.9879	$\hat{\beta}_1^{(2)} = 0.0434$	$<5 \times 10^{-5}$
	$\hat{\beta}_2^{(1)} = 5.650 \times 10^{-4}$	0.3895	$\hat{\beta}_2^{(2)} = 5.650 \times 10^{-4}$	0.3895

out by least squares or weighted least squares (so that the quantity $\mathbf{X}'\mathbf{X}$ is replaced by $\mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is a diagonal matrix). However, most inferential procedures (such as hypothesis testing) rely on **likelihood**-based estimation rather than least squares methods. For simple linear regression, the two methods coincide and an orthogonal design matrix gives rise to a diagonal information matrix; it is the information matrix that plays a central role in likelihood estimation. We shall return to this point after discussing the role played by orthogonality in the analysis of variance.

Analysis of Variance

We have just seen that in linear regression, an orthogonal design matrix enables us to estimate the effect of each covariate individually, regardless of which other covariates are included in the model. The same is true of the analysis of variance, for ANOVA model parameters are estimated by least squares, and the same geometrical argument applies – see, for example, [10, Chapter I.4]. However, for ANOVA models we have the additional advantage that, for certain types of experiment, it is possible to obtain an orthogonal design matrix in a relatively straightforward manner. A simple example illustrating this is given by [4, Appendix C].

Unfortunately, much of the literature on analysis of variance and experimental design is extremely unclear as to exactly what is meant by orthogonality in this context; the term is often used in a descriptive rather than a mathematical sense, and the justifications for its use are varied. The various strands may be connected as follows: suppose we fit an ANOVA model to data, by least-squares estimation of a parameter vector $\boldsymbol{\theta}$ (which is typically a vector of treatment effects), and that $\boldsymbol{\theta}$ can be partitioned into two subvectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Let us also fit a model containing only those parameters in $\boldsymbol{\theta}_1$ – we refer to this as the *reduced model*. Then the design of the experiment is orthogonal for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ if either of the following equivalent conditions are satisfied:

1. The sums of squares attributable to parameters in $\boldsymbol{\theta}_1$ are the same in the ANOVA table for the reduced model as in that for the full model.
2. The least-squares estimate of $\boldsymbol{\theta}_1$ is the same in the reduced model as in the full model.

These conditions result from an orthogonal design matrix, as outlined previously; however, either one of them is often cited as a definition of orthogonality, even in situations where there is no explicit mention of a design matrix. Perhaps the most helpful interpretation of such a "descriptive" view of orthogonal design (*see* **Orthogonal Designs**) is that an underlying design matrix exists, whose columns corresponding to $\boldsymbol{\theta}_1$, regarded as vectors, are orthogonal to those corresponding to $\boldsymbol{\theta}_2$. Note that a design can be orthogonal only for certain parameter combinations.

In the analysis of variance, frequent use is made of *orthogonal* **contrasts** for making treatment comparisons simultaneously. Suppose there are a total of t treatments in an experiment; any contrast is a linear combination of at least two of them whose coefficients sum to zero, and represents a comparison of treatments. The contrasts themselves may be regarded as vectors in $\mathbb{R}^{(t-1)}$, and are defined to be orthogonal if these vectors are orthogonal. If the contrasts are orthogonal and each treatment is applied to the same number of observational units, then it can be shown (see, for example, [10, p. 85]) that the design matrix corresponding to the contrasts (rather than to

the original treatments) is orthogonal, and hence their effects can be assessed independently.

Schemes giving rise to orthogonal designs for the analysis of variance include randomized block designs (*see* **Graeco–Latin Square Designs**; **Latin Square Designs**; **Randomized Complete Block Designs**). These may all be regarded as special cases of *orthogonal arrays*, which provide the combinatorial framework within which a particular orthogonal design may be generated. Orthogonal arrays are used extensively in Taguchi design; a useful introduction and extensive reference list appears in [8].

Orthogonality in Likelihood Theory

A brief account is now given of the role of orthogonality in more general parameter estimation settings. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ be a model parameter vector to be estimated, and let $l(\boldsymbol{\theta})$ be its log likelihood. Then parameters θ_i and θ_j are said to be orthogonal if $\mathrm{E}(\partial^2 l/\partial\theta_i \partial\theta_j) = 0$. The implication is that, in the large-sample normal distribution theory for the **maximum likelihood** estimator $\hat{\boldsymbol{\theta}}$, the estimators $\hat{\theta}_i$ and $\hat{\theta}_j$ are asymptotically independent. Furthermore, the standard error of $\hat{\theta}_j$ remains the same whether θ_i is regarded as known or unknown. If all k parameters are mutually orthogonal, then the information matrix is diagonal and we have a clear parallel with the linear regression case explored previously.

Now suppose that θ_i and θ_j are two parameters, and that θ_i is given. It has been shown [2] that, as θ_i is varied, the maximum likelihood estimate of θ_j changes much more slowly when the two parameters are orthogonal than otherwise. The most important consequence of this is that, for large samples, fixing θ_i has a minimal effect upon inference regarding θ_j, providing it is fixed close to its maximum likelihood estimate.

Ideally, of course, we would like the estimate of θ_j to be completely unaffected by changes in θ_i. Examples of distributions for which this is possible are given by [1, Chapter 9] – an important class is the **exponential family** with θ_i as part of the canonical parameter and θ_j as the complementary part of the expectation parameter.

These ideas have particular relevance in a generalized linear modeling framework, where the parameters represent effects of different predictor variables. The information matrix for a **generalized linear model** takes the form $\mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{X}$ is the design matrix and $\mathbf{W}$ is a diagonal matrix. Thus, if $\mathbf{X}$ is orthogonal then the information matrix is diagonal and the discussion in the preceding paragraphs applies. When the maximum likelihood estimate can be obtained directly by weighted least squares, we can assess the effect of each predictor individually as though it were the only one in the experiment; otherwise, fixing one parameter at a "sensible" value (i.e. close to its maximum likelihood estimate) will have an effect on other parameter estimates, but the effect will be smaller than if the design matrix were not orthogonal.

Orthogonal Functions

Inner products may be defined on $L^2(\mathbb{R})$, the collection of complex-valued square-integrable functions on the real line: for two functions f and g in this class, the most commonly-used inner product is given by

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)\overline{g(x)}\,\mathrm{d}x,$$

where $\overline{g(\cdot)}$ denotes the complex conjugate of $g(\cdot)$. The function space is infinite-dimensional, and any orthonormal basis for the space must have an infinite number of elements. One such basis is the collection of complex exponential functions $[e_\omega : e_\omega(x) = (2\pi)^{-1}\exp(-i\omega x), \omega \in \mathbb{R}^+]$ used in Fourier theory. For any function f which is in $L^2(\mathbb{R})$ (and satisfies certain other regularity conditions), the quantity $\langle f, e_\omega\rangle = G(\omega)$, say, is the *Fourier coefficient* at frequency ω. The Parseval relation, (2), becomes

$$\int_{-\infty}^{\infty} |f(x)|^2\,\mathrm{d}x = \int_{-\infty}^{\infty} |G(\omega)|^2\,\mathrm{d}\omega,$$

so that the basis $\{e_\omega\}$ effects an orthogonal decomposition of the variation in f into cyclical components at each frequency $\omega \in \mathbb{R}^+$.

Applications in Distribution Theory

Orthogonal functions have an important role to play in distribution theory, and give rise to methods for characterizing and approximating distributions, establishing their properties, and also for finding distributions which satisfy certain requirements. For

example, the **characteristic function** of a continuous random variable is just its expansion in terms of the Fourier basis above. Other choices of basis underlie methods such as the Edgeworth expansion. A good account of the application of such methods in distribution theory is given by [11]; see also [9].

Applications in Time Series Analysis

Any realization of a time series (discrete or continuous parameter) may be regarded as a function defined on some interval and may thus, providing it is suitably well-behaved, be expressed as a linear combination of functions which are orthogonal over that interval in exactly the same way as if it were deterministic (the only difference is that the coefficients in the linear representation are random variables rather than constants). This is the underlying idea behind the **spectral analysis** of time series. Wavelet techniques (see [3], for example) are another example using different basis functions. In some situations, there may be a physical reason for wishing to decompose a time series into orthogonal components (spectral analysis was originally developed as a means of detecting periodicities in noisy data); in others, statistical properties (such as approximate uncorrelatedness) of the coefficients in the orthogonal representation may be useful in their own right.

Concluding Remarks

Orthogonality has applications throughout mathematics and statistics; the topics covered here are merely those which have some relevance in biostatistics. Recurring themes include the Parseval relation, which simultaneously forms the basis for an extension of Pythagoras's theorem and for variance decomposition, and the projection theorem, asserting the existence of a unique least squares linear predictor, for example, which is orthogonal to the prediction error (this property is called the *orthogonality principle* by some authors). In closing, we emphasize that orthogonality is best exploited in *linear* systems (a consequence of the linearity of a vector product in the underlying mathematical framework).

Given the enormous range of subjects within which the notion of orthogonality may be usefully employed, the references below have been chosen to provide a broad (and, where possible, recent) overview of the relevant ideas. Most contain extensive bibliographies for more detailed reading.

References

[1] Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

[2] Cox, D.R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion), *Journal of the Royal Statistical Society, Series B* **49**, 1–39.

[3] Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia.

[4] Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*. Chapman & Hall, London.

[5] Good, I.J. (1969). Some applications of the singular value decomposition of a matrix, *Technometrics* **11**, 823–831.

[6] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski E. (1994). *A Handbook of Small Datasets*. Chapman & Hall, London.

[7] Karr, A.F. (1992). *Probability*. Springer-Verlag, New York.

[8] Logothetis, N. & Wynn, H.P. (1989). *Quality Through Design: Off-line Quality Control, and Taguchi's Contributions*, Vol. 7. *Oxford Series on Advanced Manufacturing*. Oxford University Press, Oxford.

[9] Lukacs, E. (1970). *Characteristic Functions*. Griffin, London.

[10] Mead, R. (1988). *The Design of Experiments: Statistical Principles for Practical Application*. Cambridge University Press, Cambridge.

[11] Ord, J.K. (1972). *Families of Frequency Distributions*. Griffin, London.

[12] Rudin, W. (1987). *Real and Complex Analysis*, 3rd Ed. McGraw-Hill, New York.

[13] Seber, G.A.F. (1977). *Linear Regression Analysis*. Wiley, New York.

(*See also* **Matrix Algebra**)

R.E. CHANDLER

Otorhinolaryngology

Otorhinolaryngology (ORL) is a speciality in medicine dealing with diseases of the ears, nose, and throat (ENT). The practice of ORL continues to change and expand, as does research in this field. It covers areas such as audiology, immunology, **oncology**, microbiology, facio-plastic surgery, implantation of advanced medical technologies, **genetics**, and communicative development [8].

Modern otorhinolaryngology has expanded so much over the past few decades that subspecialities have developed, including pediatric ORL, head and neck surgery, and otology. ENT surgeons care for with patients of all age groups with a focus on children (upper airway infections, middle ear diseases) and elderly people (deafness).

As most of the diseases dealt with by otolaryngologists interfere with communication skills (hearing, speech), the consequences for individual quality of life and for society are tremendous. This is especially true because many of the diseases in ORL are highly **prevalent**, for example:

1. Seventy percent of children experience one or more episodes of acute otitis media before the age of 3 years [17]. Middle ear effusion that is associated with hearing loss is present in 20–40% of all ears of preschool children, dependent on the season [20].
2. ENT operations (adeno-tonsillectomy, tympanostomy tubes) are by far the most common surgical procedure in Western societies [5].
3. Around 10% of all people older than 65 years possess a hearing aid [4].

In addition to these highly prevalent diseases, the otorhinolaryngologist can be faced with several relatively rare but life-threatening diseases, such as acute airway obstruction, congenital syndromes with craniofacial malformations, and malignant tumors of head and neck.

In many respects, research in ORL follows the same lines as in other specialities, including the biostatistical aspects of these studies. Some landmark studies on selected areas are:

1. A randomized **clinical trial** of the Veterans Affairs Study Group comparing combinations of surgery, chemotherapy, and radiotherapy in patients with laryngeal cancer with respect to survival rate and other outcomes. Statistical analyses are straightforward comparing **Kaplan–Meier** curves of patients' survival and disease-free interval by **logrank tests** [18] (*see* **Survival Analysis, Overview**).
2. A newly developed PCR assay (polymerase chain reaction, a gene amplification method) for *Streptococcus pneumoniae* is compared with classic bacterial culture of the middle ear fluid. The assay is to be used for the etiological diagnosis of acute otitis media [19]. **Sensitivity** and **specificity** of the PCR method were assessed in 180 middle ear fluid samples of 125 children with acute otitis media. Standard statistical procedures (**chi-square test** and **McNemar test**) were applied to the data.
3. Triggered by the high incidence of permanent sensorineural hearing loss among US soldiers returning from World War II, many studies (human and animal experiments) have been published on noise-induced hearing loss [6]. Typical of these kinds of studies is the graphical presentation of **mean** and **standard deviation** threshold shifts (TS) for exposed and unexposed groups. Such permanent or temporary TSs [expressed in decibel (dB) loss] are measured by pure tone audiometry at frequencies ranging from 500 Hz to 8000 Hz. For statistical testing many authors compare the mean of three frequencies (1000 Hz, 2000 Hz, and 4000 Hz, important for discrimination of speech) for each exposure group.
4. Typical for pediatric infectious diseases in ORL is the favorable natural course. As a result, the effect of treatment in clinical practice seems large. In randomized clinical trials, however, only modest effects are demonstrated, underscoring the importance of this type of study. An example of this phenomenon is a three-way trial in 177 children with recurrent episodes of throat infection that met the US standards for tonsillectomy or adino-tonsillectomy [12]. Children were randomized to either a tonsillectomy group, an adino-tonsillectomy group, or a control group and followed for three years. The authors assumed a Poisson distribution for comparing mean rates of occurrence of episodes of throat infection using a generalized linear model.

As ORL diseases are so diverse and ORL research touches so many disciplines, including laboratory, clinical, and epidemiologic research (*see* **Observational Study**), all study designs and statistical procedures are applied in this field. Some specific characteristics of ORL bear consequences for biostatistics.

Specific Statistical Issues in ORL Research

Dependency

In principle each patient in an ORL study provides the investigator with one nose, one throat, and two ears. The ears do not act independently, but neither do they perform in perfect concert. Analyses of data on hearing, middle ear function, and morphology of the ear have to cope with this partial dependency. This feature of paired, dependent organs is not confined to ORL. Other medical disciplines such as **ophthalmology**, orthopedics, and **nephrology** share this feature. With a view to the **power** of the study, it is tempting to analyze the data as if they are independent. This will lead to **confidence intervals** (CI) that are too narrow. Until recently, conservative solutions for this problem were found in selecting just one ear for each patient, but this led to considerable loss of information. The construction of a 95% CI for the prevalence of middle ear disease with the ear as the unit of analysis, accounting for the dependency in pairs of ears, is given below.

Let θ denote the true probability for an ear to have middle ear disease and let γ denote the true **conditional probability** that one ear is affected when the other ear is also affected. Suppose a sample of n pairs of ears ($2n$ single ears) is drawn. Let n_{l} denote the number of left ears and n_{r} the number of right ears that are affected. Denote by n_{lr} the number of pairs where both ears are affected. The **maximum likelihood** estimates for the prevalence of θ and the conditional probability γ are given by

$$\hat{\theta} = \frac{n_{\mathrm{l}} + n_{\mathrm{r}}}{2n}, \qquad \hat{\gamma} = \frac{2n_{\mathrm{lr}}}{n_{\mathrm{l}} + n_{\mathrm{r}}}.$$

An approximate 95% CI for θ is then given by

$$\hat{\theta} \pm 2\left(\frac{\hat{\theta}(1-\hat{\theta}) + (\hat{\theta}\hat{\gamma} - \hat{\theta}^2)}{2n}\right)^{1/2}.$$

When both ears are independent ($\hat{\theta} \approx \hat{\gamma}$) the 95% CI reduces to the conventional confidence interval for θ from $2n$ Bernoulli trials with success probability θ. If there is positive dependency, $\hat{\gamma} > \hat{\theta}$, then the width of the correct CI is on average larger than the conventional CI.

More advanced methods for the analyses of correlated ears with the possibility to include **covariates** are described by Rosner [15], Dallal [3], and Le & Lindgren [10]. The outcome variables in the models given by Rosner are **normally distributed** (**multiple linear regression**) or **binomially** distributed (**logistic regression**). Dallal modifies the assumptions of the logistic regression model of Rosner. Le & Lindgren give another logistic model for the analysis of correlated ears (*see* **Correlated Binary Data**).

Random effect logistic models for indistinguishable data are another possibility to model the dependency between ears and to include covariates. In these models, the probability θ that an ear is affected does not only depend on the covariate values but also on extra random variation. Let θ_i denote the probability for each ear in the ith pair to be affected. The θ_i is modeled by

$$\mathrm{logit}(\theta_i) = \ln\left[\frac{\theta_i}{(1-\theta_i)}\right] = \mathbf{z}_i'\boldsymbol{\beta} + \sigma\mathbf{u}_i.$$

In this expression $\mathbf{z}_i$ is the covariate vector for the ith pair, $\boldsymbol{\beta}$ the vector of the regression parameters, $\mathbf{u}_i$ a realization of a standard normal distribution, and $\sigma > 0$.

The issue of dependent ears also provides elegant potential for within-person trials. Maw [11] became famous for his trials on surgical treatment for persistent middle ear disease in children. The design of the trial allowed the child to act as his or her own control, in that only one of the bilaterally affected ears was treated with a ventilation tube. The unoperated ear in the group of children not receiving surgery either to the adenoids or tonsils allowed documentation and examination of the natural history of the untreated condition over a long period (*see* **Crossover Designs**; **Unit of Analysis**).

Fluctuating Natural Course

Many infectious diseases in ORL have a very fluctuating, and usually favorable natural course, that is,

episodes of disease in different degrees of severity are followed by episodes without the disease and vice versa in an apparently random order. A typical example is the occurrence of otitis media with effusion (OME) in children. There are four states for characterizing a middle ear with respect to OME based on tympanometry measurement of the compliance of the tympanic membrane at different levels of external air pressure:

A. Normal (aerated) ear: no effusion present in the middle ear cavity. A middle ear pressure corresponding to ambient air pressure. Maximal compliance at ambient air pressure.
B. OME ear: middle ear cavity filled with fluid. No compliance of the tympanic membrane.
C1. Intermediate: slight negative pressure in the middle ear. Fluid in some instances. Maximal compliance at pressures below ambient air pressure.
C2. Intermediate: moderate negative pressure in the middle ear, fluid in approximately half of the cases. Maximal compliance at pressures below ambient air pressure.

A typical pattern for the development of the disease is from type A to type B via type C1 and type C2, followed by the same sequence in reverse order. It takes at least several weeks to develop OME and to recover from it. Short episodes (1 or 2 months) are common, in many cases not even reaching the type B state. Some children's ears, however, have persistent middle ear effusion (type C2 and B) that may have long-term consequences for hearing and language development.

A description of the natural course of OME is very important as a first step in identifying those children who will have persistent disease and therefore need treatment. As there is no simple model that fits the empirical data on the natural history of OME, most statistical techniques fail in this description. For instance, **Markov** models were explored but gave no valid descriptions. Instead we used several graphical techniques [22]. A useful technique describing dichotomized data (type B vs. the other types) on ears measured by tympanometry nine times in a period of 2 years is shown in Figure 1. It gives some insight into **cumulative incidence**, rate of improvement, and rate of recurrence. Unfortunately, these graphical techniques are not suitable for studying determinants of natural course, nor do they provide precision parameters.

Simplified models and survival analytic techniques are more suitable when the focus is on duration of OME episodes. Let X denote the duration of an OME episode. Suppose X has an **exponential distribution** with density $a\exp(-ax)$, $a > 0$, and X is measured in months. If after 3 months 50% of the patients have recovered from OME, the monthly rate of recovery is given by the solution of the equation $\exp(-3a) = 0.5$. The proportion of patients still diseased after x months is then given by $\exp(-ax)$. Such a simple model does help to describe natural course (with an appealing effect on public awareness) but might fail proper fit [21].

An extension of the **life table** to repeating and changing events described by Hoover [7] can be used to model the cumulative episodes of middle ear disease. Tests for changes in the number of episodes within age intervals and differences between individuals are possible.

Dynamics of Disease Over Time

Some typical ORL diseases now very prevalent, such as OME, seemed rare in the past. Other diseases, such as acute mastoiditis, have become less frequent over time. These patterns are subject to **bias** as concepts and classification of disease also change over time [2]. Until the nineteen nineties, OME was considered a sterile inflammation of the middle ear, distinct from but correlated with acute otitis media and upper airway infections. Nowadays, the general belief is that OME is just one of several expressions of infections of the upper respiratory system. This idea is supported by **principal components analysis** of data on signs and symptoms of a large number of patients [16].

In addition, demographic trends such as aging of the population and improving socioeconomic conditions and changes in disease management lead to changes in frequency of diseases or its sequelae [23]. Societal views on **quality of life** are also major issues in ORL research [14]. For instance, utilities for technology-supported hearing programs (cochlear implants) or radical head and neck surgery (vs. radiotherapy in case of advanced laryngeal carcinoma) may vary over time and between groups of patients. One should recognize these differences in the analysis and interpretation of descriptive statistics on ORL diseases and/or frequency of treatment options.

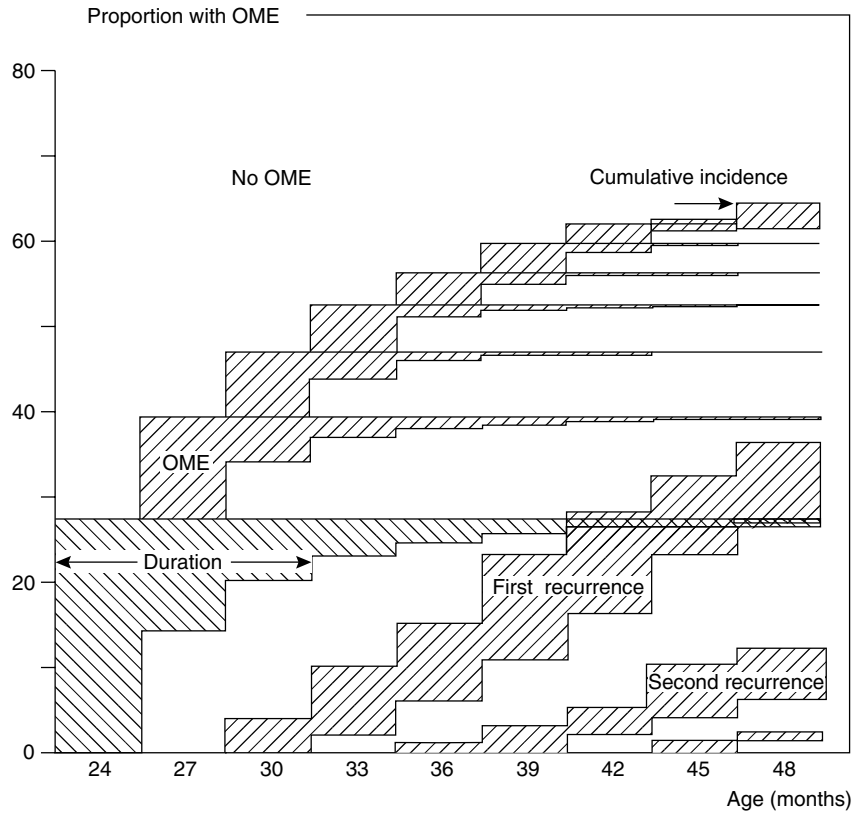

Figure 1 Natural history of OME. Children with tubes excluded; $n = 1217$ ears. Reproduced from Zielhuis et al., *Lancet*, vol. i, pp. 311–314. © The Lancet Ltd 1989

Dynamics of Management Over Time

ORL is a medical discipline that has experienced large changes in the popularity of treatment options over time. These were partly the result of the dynamics of the disease itself, partly a result of newly developed techniques or drugs (note that antibiotics have only been available since World War II). Some changes, however, are due to new scientific insights into the effectiveness of specific treatments. A typical example of this in the field of ORL is the practice of tonsillectomy, alone or in combination with adenoidectomy [2, 5, 11]. In the 1950s and 1960s it was almost routine to remove the tonsils and/or adenoids. During the 1970s and 1980s the belief in the effectiveness of these surgical procedures disappeared, leading to a one-third reduction of their frequency [5]. Since the late 1980s there has been some revival of the tonsillectomy and/or adenoidectomy for specific indications [12]. In the same period the rate of surgery with ventilation tubes has risen quite rapidly, with stabilization in recent years.

The practice of ORL in general is considered sensitive to the results of epidemiologic research, bearing consequences for medical practice. Despite a high standard of international scientific communication in journals and at conferences, large international differences do occur. For instance, European countries still prefer surgical treatment of middle ear disease, while US physicians (most of them pediatricians) prefer medical treatment (e.g. antibiotics or steroids) [23]. To a certain extent these international differences provide opportunities for **ecologic studies**. Medical audit and formal assessment of cost–effectiveness or cost–benefit ratios of medical procedures and technologies is also expanding into the field of otolaryngology [1, 9] (*see* **Health Economics**). Alongside standard statistical procedures, techniques like **factor analysis**, uncommon in this

medical field, are being introduced. An example is the Glasgow Benefit Inventory as a measure of patient benefit from ORL interventions [13]. With a principal components analysis of patient responses to five such interventions, subscales for different types of benefit are developed. The use of these and other evaluation instruments will assist evaluation research as well.

Concluding Remarks

Biostatisticians entering the field of ORL are confronted with large numbers of patients and a diversity of diseases and treatment options. Most of the statistical procedures used are similar to those used in other branches of medicine. Special features are the potential of each patient having two ears, the fluctuating course of many of the common ORL diseases, and the diversity of management options.

References

[1] Berman, S., Roark, R. & Luckey, D. (1994). Theoretical cost effectiveness of management options for children with persisting middle ear effusions, *Pediatrics* **93**, 353–363.

[2] Black, N.A. (1984). Is glue ear a modern phenomenon? A historical review of the medical literature, *Clinical Otolaryngology* **9**, 155–163.

[3] Dallal, G.E. (1988). Paired Bernoulli trials, *Biometrics* **44**, 253–258.

[4] Davis, A.C. (1996). Epidemiology, in *Scott Brown's Otolaryngology*, 6th Ed., A.G. Kerr, ed. Butterworth Heinemann, Oxford, pp. 23–24.

[5] Derkay, C.S. (1993). Pediatric otolaryngology procedures in the United States: 1977 1987, *International Journal of Pediatric Otorhinolaryngology* **25**, 1–12.

[6] Henderson, D., Subramaniam, M. & Boettcher, F.A. (1993). Individual susceptibility to noise-induced hearing loss: an old topic revisited, *Ear and Hearing* **14**, 152–168.

[7] Hoover, R. (1996). Extension of the life table to repeating and changing events, *American Journal of Epidemiology* **143**, 1266–1276.

[8] Kerr, A.G., ed. (1996). *Scott Brown's Otolaryngology*, 6th Ed. Butterworth Heinemann, Oxford.

[9] Kleinman, L.C., Kosecoff, J., Dubois, R.W. & Brook, R.H. (1994). The medical appropriateness of tympanostomy tubes proposed for children younger than 16 years in the United States, *Journal of the American Medical Association* **271**, 1250–1255.

[10] Le, C.T. & Lindgren, B.R. (1990). Statistical methods for determining risk factors of chronic otitis media with effusion, *Statistics in Medicine* **9**, 1495–1500.

[11] Maw, A.R. (1995). *Glue Ear in Childhood. A Prospective Study of Otitis Media with Effusion*. MacKeith Press, London.

[12] Paradise, J.L., Bluestone, C.D., Colborn, K., Bernand, B.S., Rockette, H.E. & Kurs-Laysky, M. (2002). Tonsillesctomy and adinotonsillesctomy for recurrent throat infection in moderately affected children, *Pediatrics* **110**, 7–15.

[13] Robinson, K., Gatehouse, S. & Browning, G.G. (1996). Measuring patient benefit from otorhinolaryngological surgery and therapy, *Annals of Otology, Rhinology, & Laryngology* **105**, 415–422.

[14] Rosenfeld, R.M., Goldsmith, A.J., Tetlus, L. & Balzaro, A. (1997). Quality of life for children with otitis media, *Archives of Otolaryngology–Head & Neck Surgery* **123**, 1049–1054.

[15] Rosner, B. (1984). Multivariate methods in ophthalmology with applications to other paired data situations, *Biometrics* **40**, 1025–1035.

[16] Schilder, A.G.M., Zielhuis, G.A., Straatman, H. & van den Broek, P. (1992). An epidemiological approach to the etiology of middle ear disease in The Netherlands, *European Archives of Otorhinolaryngology* **249**, 370–373.

[17] Teele, D.W., Klein, J.O. & Rosner, B.A. (1980). Epidemiology of otitis media in children, *Annals of Otorhinolaryngology* **89**, Supplement 68, 5–6.

[18] Veterans Affairs Laryngeal Cancer Study Group (1991). Induction chemotherapy plus radiation compared with surgery plus radiation in patients with advanced laryngeal cancer, *New England Journal of Medicine* **324**, 1685–1690.

[19] Virolainen, A., Salo, P., Jero, J., Karma, P., Eskola, J. & Leinonen, M. (1994). Comparison of PCR assay with bacterial culture for detecting streptococcus pneumoniae in middle ear fluid of children with acute otitis media, *Journal of Clinical Microbiology* **32**, 2667–2670.

[20] Zielhuis, G.A., Rach, G.H. & van den Broek, P. (1989). Screening for otitis media with effusion in preschool children, *Lancet* **i**, 311–314.

[21] Zielhuis, G.A., Rach, G.H. & van den Broek, P. (1990). The natural course of otitis media with effusion in preschool children, *European Archives of Otorhinolaryngology* **247**, 215–221.

[22] Zielhuis, G.A., Straatman, H., Rach, G.H. & van den Broek, P. (1990). Analysis and presentation of data on the natural course of otitis media with effusion in children, *International Journal of Epidemiology* **19**, 1037–1044.

[23] Zuijlen, D.A., van Schilder, A.G.M., Balen, F.A.M. & van Hoes, A.W. (2001). National differences in incidence of acute mastoiditis:relationship to prescribing patterns for acute otitis media? *The Pediatric infectious disease journal* **20**, 140–144.

GERHARD A. ZIELHUIS, HUUB STRAATMAN &
ANNE G.M. SCHILDER

Outcome Measures in Clinical Trials

Clinical trials are an important source of information used by health care decision makers to establish health policy, and to make judgments about prescribing treatments for specific patients. In recent years, the clinical evaluation of therapeutic interventions has moved beyond simple measures of safety and efficacy to include measures of symptom improvement, **quality of life**, and cost-effectiveness (*see* **Clinical Epidemiology**; **Health Economics**). This has resulted in more complex experimental designs using multiple response variables to address multiple objectives (*see* **Multiplicity in Clinical Trials**). The purpose of this article is to address the statistical considerations used to select the outcome measures in clinical trials. Issues relating to clinical endpoints, quality of life assessments, and economic evaluations are discussed.

Clinical Endpoints

Primary Outcomes

In regulated environments evaluating pharmaceutical interventions, it is recommended that clinical trials have a single primary objective. In confirmatory trials, this typically includes **hypothesis testing** about the comparative efficacy of the therapies under study. The hypothesis may be one of superiority of one treatment compared to a placebo or other control, or it may be the **equivalence** of two treatments. The primary outcome variable or endpoint is the measure capable of providing the best evidence directly related to the primary objective of the trial.

It is readily agreed that the most clinically relevant outcomes are those directly relating treatment to the patient's health status. In studies of serious or life-threatening diseases, this may be an important clinical event (e.g. mortality or myocardial infarction), measured either as a **binary** outcome or by the time from **randomization** to the event. In some trials, health status may be measured according to disease rating scales, usually with ordered categories of severity (*see* **Ordered Categorical Data**) (e.g. the New York Heart Association classification of congestive heart failure). When study variables are continuous (e.g. blood pressure or serum cholesterol) the effect of treatment is usually based on a comparison of change or percent change from baseline. Response data are often collected on a repeated basis at several time points during the trial. Also, patients may experience several primary events during the period of observation. These data can be analyzed using models for recurrent events or using the time to the first event.

It is important that the primary outcome variable be measured without **bias** in a reliable manner using validated instruments with adequate **sensitivity** to detect real change in a patient's health status. These assessments should be made prior to initiation of the trial, using experience gained from previous trials or from the published literature. The **power** and **sample size for the trial** should be based on the primary outcome variable.

Surrogate Endpoints

There are situations in which short-term measures of response to treatment may provide reliable indicators of long-term patient outcome. The presence of serum antibody at protective levels following vaccine administration is one example. In the area of cardiovascular disease, it has been shown using **meta-analysis** [20], and with long-term clinical endpoint trials, that lipid-lowering by diet or drugs does confer the benefits of a reduced risk of coronary heart disease [36, 37] and increased survival [36]. This suggests that serum cholesterol levels may be a viable **surrogate endpoint** marker.

The use of surrogate endpoints also has its drawbacks [14]. Pocock [33] notes that their use has been a "deceptively attractive pursuit" in many areas of clinical trials research. His view was formed in part by HIV research, with a reliance on CD4 counts as a potential surrogate that could hopefully shortcut the need for long-term trials. However, the Concorde trial [5] of immediate and deferred use of zidovudine failed to show a difference in clinical outcomes, despite a significant change in CD4 cell counts.

The statistical methodology needed to explore surrogate endpoints has undergone considerable development in recent years. Prentice [35] proposed a definition and operational criteria for surrogate endpoints. Fleming et al. [15] discussed data analysis methods.

Responder or Threshold Variables

Investigators may want to re-express continuous or ordinal variables into categorical or dichotomous response variables (*see* **Categorical Data Analysis**). These are often called responder or threshold analyses since they are usually specifications of a positive clinical result. A typical example may be a reduction in serum LDL-cholesterol to a level below the recommended levels established by the National Cholesterol Education Program [31]. When using these types of variables, it is important to recognize that the loss of information can result in a loss of power for the study, and **regression to the mean** effects can introduce bias when the underlying variables are subject to **measurement error** [4].

Composite Variables

Constructed variables which are composites of multiple measurements or endpoints are sometimes used to provide a single summary response. At the individual level, this involves combining the univariate responses in some clinically sensible manner. This technique is often used in quality of life assessments. Composite endpoints can also be formed from multiple clinical event data. For example, clinical trials of antiretroviral therapy for patients infected with HIV may use the combined clinical endpoint of any new or recurrent AIDS defining event or death. Composite endpoints of heart disease or death have also been used for cardiovascular clinical trials. Composite variables may offer advantages in terms of increased statistical power due to the reduction in dimensionality of the multiple endpoints, or to the increased incidence of the composite event when the incidences of the individual events are low. However, this advantage is offset if the treatment does not affect each endpoint consistently. Also, the clinical interpretation of the constructed variables can be difficult.

Secondary Outcomes

Secondary outcome variables are either supportive measures to help interpret the primary results, or response variables, related to secondary objectives or hypotheses. Trials may also include **explanatory variables** to be used for generating hypotheses to be tested in future studies.

Compliance

Although difficult to measure, it is well known that patient noncompliance can dramatically impact estimates of treatment effect in trials (*see* **Compliance Assessment in Clinical Trials**). In a study comparing two treatment programs to reduce levels of serum LDL-cholesterol, Oster et al. [31] showed that patient compliance differed between programs and, not surprisingly, noncompliance affected the likelihood of patients achieving their target levels. Efron & Feldman [13] proposed a compliance–response **regression** to measure the relationship between the treatment effect and compliance using data from a primary **prevention trial** of a cholesterol lowering drug (*see* **Noncompliance, Adjustment for**).

Multiplicity Considerations

Inclusion of **multiple endpoints** in a trial raises concern about an increased probability of drawing a **false positive** conclusion. This is the principal reason that studies subject to regulatory review use only a single primary endpoint, or use appropriate multiplicity adjustments in those situations in which multiple hypotheses are being tested. Cook & Farewell [6] take a less restrictive viewpoint in allowing studies to consider a few well-chosen, important clinical outcomes and computing a marginal ***P* value** for each. This position recognizes the P value as providing a measure of the strength of the evidence against the **null hypothesis**, rather than as a decision making criterion wherein the type I error rate α (*see* **Hypothesis Testing**) is interpreted as a rejection rate.

Understanding the distinction between multiple endpoints and multiple hypotheses is particularly important in circumstances in which the underlying disease may be measured across a number of meaningful dimensions. One such example is described by Gormley et al. [19], who studied the effect of finasteride in men with benign prostate hyperplasia (BPH). They recognized BPH as a multifaceted disease and evaluated the effect of treatment on a biochemical variable, dihydrotestosterone, on physiologic measures, prostate size, and urinary flow rate, and on a subjective assessment of BPH symptoms using a validated questionnaire. If evidence of a treatment effect depends on any single variable being impacted by treatment, then attention focuses on the most extreme P value, and a **Bonferroni**-type adjustment

(see, for example, [22]) is appropriate (*see* **Multiple Endpoints, *P* Level Procedures**). If evidence of a treatment effect requires that each of the variables is affected by treatment, then no adjustment to the individual P values is necessary. Tests of a global hypothesis proposed by O'Brien [30] can also be performed using the individual analyses to help interpret the results (*see* **Multiplicity in Clinical Trials**).

Quality of Life

Patient perceptions of health-related **quality of life** have recently received a higher priority in clinical trials. This is due in part to the broadened definition of health proposed in 1947 by the **World Health Organization** as "a state of complete physical, mental, and social well being, and not merely the absence of disease and infirmatory" [40]. One of the first clinical applications of quality of life was by Karnofsky & Burchenal [27], who outlined the basic criteria necessary for evaluating chemotherapeutic agents. These included measures of survival, but also subjective variables, such as performance status, symptom status, mood, and well-being. There are several important statistical considerations in the design, analysis, and interpretation of quality of life studies. Fletcher et al. [16] and Cox et al. [7] offer recommendations on how to keep quality of life evaluations simple.

Measurement Characteristics

The most important statistical consideration in using quality of life questionnaire instruments in clinical evaluations is an assessment of its measurement characteristics (*see* **Health Status Instruments, Measurement Properties of**). Quality of life (QoL) is by definition a multidimensional construct. A typical instrument consists of components or dimensions which usually include a number of individual questions or items. The scoring of an instrument begins with a response scale for each item. Scores are usually provided for each dimension and an overall score is sometimes provided. Although controversial, the scoring scheme may sometimes use weights based on individual patient preferences [38], or data analytic techniques based on **factor analysis** or **discriminant analysis**.

Quality of life scales used in clinical trials should have demonstrated reliability and validity, and be responsive to changes in health status. Reliability is assessed through examination of the internal consistency at a single administration of the instrument, using **Cronbach's alpha**. The test–retest properties of the scales can be evaluated using the concordance **correlation** coefficient, or the intraclass correlation. The correlation coefficient is not an appropriate measure of agreement in assessing reliability [2].

Face validity is a subjective judgment of whether the instrument appears to cover its intended topics clearly and unambiguously. It can be maximized by including individuals of diverse backgrounds, including both patients and healthcare providers, among the developers. Construct validity is a formal measure of the association between the QoL scores and other objective and subjective measures of health status.

Responsiveness checks the ability of a scale to detect clinically meaningful changes in health status. It is the most difficult property to assess prior to its use in a trial. Responsiveness is usually assessed through measuring the longitudinal association of changes in QoL scores with changes in other measures of health status. Jaeschke et al. [24] defines a minimal important difference as the change from baseline in QoL score associated with a patient perceived change in health status based on a global rating.

Selection of Instruments

The choice of dimensions to use in a trial depends on the nature of the disease and the candidate treatments. For example, symptomatic conditions such as asthma or rheumatoid arthritis are likely to have a negative impact on a patient's quality of life that may potentially be offset with therapy. Alternately, for patients with high blood pressure or high levels of serum cholesterol, interest focuses on the adverse experiences associated with therapy, or on the reduction in both quality and length of life associated with coronary heart disease.

Most clinical trials use standard questionnaire instruments. These are generally categorized as generic or disease-specific. Generic instruments cover a broad range of quality of life dimensions and are particularly useful when there is uncertainty regarding the appropriate dimensions, or when there is not much known about the therapeutic effects. They are also useful for making comparisons among

different diseases, which is sometimes necessary when establishing health policy. The disadvantage of generic instruments is that their responsiveness to changes in health status is reduced. Health indices are meant to provide an overall measure of a patient's quality of life by using a continuum from 0 (death) to 1 (perfect health). Health states with diminished quality are scored less than 1. The clinical usefulness of health indices is limited unless individual component scores are also made available.

Well-known examples of generic instruments include the MOS 36-item short form health survey (SF-36) [39], the Sickness Impact Profile [1], and the Nottingham Health Profile [23]. The Quality of Well-Being (QWB) index [26] is a health index that is an additive combination of three function scales (mobility, physical activity, and social activity), and a scale for symptoms and health problems. Fryback et al. [17] provides a quantitative link between the eight subscales of the SF-36 and the QWB as a means of providing a single health utility summary score representative of the SF-36 profile.

The quality adjusted life year (QALY) method of analysis [38] accumulates the 0 to 1 quality scores over the lifetime of individuals to combine the concepts of quality and length of survival. The number of QALYs gained is often used as a measure of health benefit in cost–effectiveness analysis (*see* **Health Economics**). A related method measures TWIST (Time WIthout Symptoms of disease and Toxicity of treatment) and Q-TWIST provides a quality adjustment to TWIST [18] (*see* **Quality of Life and Survival Analysis**). While conceptually attractive, both the QALY and Q-TWIST have limitations in interpretation, since separate evaluations of quality and survival are usually of interest.

Disease-specific instruments focus only on relevant dimensions of quality of life that impact patients with the underlying target illness. As such, they have greater responsiveness and a higher level of patient acceptance. Disease-specific quality of life is most applicable with symptomatic disease, where patients are able to detect changes in their health status due to therapy. For example, Juniper et al. [25] evaluated the measurement properties of an asthma QoL instrument developed for use in clinical trials. The 32-item instrument included dimensions for activities, symptoms, emotions, and exposure to environmental stimuli, and was developed through an item-generation and item-reduction process involving adult asthmatics.

Disease-specific QoL evaluations may have limited usefulness in guiding health policy decisions involving multiple diseases and therapeutic interventions. Fletcher et al. [16] recommends using both a generic and a disease-specific instrument in trials. This approach assures the focus needed to identify the QoL impact of changes in health status, and yet allows the possibility of detecting unexpected effects, and may allow a comparison with other diseases.

Analysis

Special considerations in the analysis of QoL data in a clinical trial are related to the multidimensional aspect of both the concept and the instrument. Cox et al. [7] recommend simplicity in the scoring and weighting schemes used. The analysis should also evaluate the sensitivity of the results to alternate scoring and weighting schemes. Fletcher et al. [16] recommend that scores can be expressed as the percentage out of the maximum achievable, to facilitate the interpretation of aggregated dimension scores.

The separate dimension mean scores, dimension-by-treatment mean scores, and the overall treatment mean scores are the key values for interpretation. The possibility of a dimension-by-treatment qualitative interaction (*see* **Treatment-covariate Interaction**) should be explored, since a drug may have a positive impact on some dimensions and a negative impact on others. A complete interpretation of the trial results needs to consider the multivariate character of the data. This raises issues of **multiplicity**, which can be addressed by a protocol specification of a limited number of dimensions thought to be impacted by therapy. An overall summary score can be derived, but limits the clinical utility of the analysis. Bonferroni and modified Bonferroni-type procedures [22] based on the ordered P values can be used if the researcher is to treat the multiple dimensions as separate multiple hypotheses. Tests of significance of departures from a global null hypothesis of treatment equivalence can be undertaken using the methods of O'Brien [30]. Global tests and overall summary scores are more applicable for evaluating disease-specific QoL, since there is less possibility of an interaction.

Effect Size

Two forms of effect size are available to help interpret the magnitude of changes observed in a trial. Deyo et al. [8] consider the difference in **means** relative to the **standard deviation** at baseline. This approach transforms the change score into a standardized unit of measurement that can be compared to other instruments. The index of responsiveness [21] considers the pre-minus-post mean changes relative to the within-patient standard deviation during a stable period. This allows differences to be measured against changes that individuals experience in normal day-to-day variability. Fletcher et al. [16] note that, for an individual patient, a treatment effect of one to two units of responsiveness is probably important. When considering average treatment effects for a randomized group, it is likely that effects that are larger than one-third of the between-patient standard deviation are noticed by individual patients. The availability of estimated effect sizes for established therapies is useful in providing a benchmark for interpreting the results of trials involving experimental therapies.

Economic Analysis

Economic endpoints are becoming an increasingly important part of clinical trials. Data on individual patients' use of health care resources can be collected and combined with cost data, relevant to the trial perspective, to perform cost comparisons or cost–effectiveness analysis (*see* **Health Economics**).

There has been considerable debate on the appropriateness of randomized trials to support valid conclusions about economic endpoints [34]. Drummond & Davis [11] discuss some of the methodologic issues. There is no consensus on the most relevant choice of outcome measure, clinically important effect sizes [12], or the analytic method to be used in comparing treatments on the basis of cost–effectiveness [3, 29]. Dudley et al. [10] note the complexity of analyzing cost data that are subject to **censoring**.

Advances in the methodology are being made, as a number of trials have recently been completed that have included economic evaluations. Pedersen et al. [32] collected data on hospitalization and coronary revascularization procedures, as part of the Scandinavian Simvastatin Survival Study. They concluded that simvastatin therapy, to reduce levels of serum cholesterol, has the potential for significant cost offsets, in addition to reducing mortality and morbidity. Mark et al. [28] supplemented data from the one-year Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) trial with projected **life expectancy** data and hospital cost data, to evaluate the cost–effectiveness of thrombolytic therapy with tissue plasminogen activator (t-PA) compared to streptokinase for the treatment of acute myocardial infarction. Using cost–effectiveness analysis, they concluded that the added costs of t-PA therapy were consistent with other well-accepted medical technologies, when evaluated relative to the benefits of therapy. Similarly, the Diabetes Control and Complications Trial (DCCT) Research Group [9] examined the cost–effectiveness of aggressive approaches to the management of insulin-dependent diabetes mellitus. They used a **Monte Carlo simulation** model, based on the clinical outcome results of the DCCT, supplemented with other clinical, epidemiologic, and cost studies.

These three examples illustrate innovative ways of utilizing clinical and health care utilization data, collected as part of randomized trials, to perform cost–effectiveness analysis. In each case, the researchers were able to overcome certain limitations in the trial with supplemental data and modeling techniques.

References

[1] Bergner, M., Bobbitt, R., Carter, W. & Gibson, B. (1981). The sickness impact profile: development and final revision of a health status measure, *Medical Care* **19**, 787–805.

[2] Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **i**, 307–310.

[3] Chaudhary, M.A. & Stearns, S.C. (1996). Estimating confidence intervals for cost effectiveness ratios: an example from a randomized trial, *Statistics in Medicine* **15**, 1447–1458.

[4] Cochran, W.G. (1968). The errors of measurement in statistics, *Technometrics* **10**, 637–666.

[5] Concorde Coordinating Committee (1994). Concorde: MRC/ANRS randomized double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection, *Lancet* **343**, 871–881.

[6] Cook, R.J. & Farewell, V.T. (1996). Multiplicity considerations in the design and analysis of clinical trials,

Journal of the Royal Statistical Society, Series A **159**, 93–100.

[7] Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Spiegelhalter, D.J. & Jones, D.R. (1992). Quality of life assessment: can we keep it simple?, *Journal of the Royal Statistical Society, Series A* **155**, 353–393.

[8] Deyo, R.A., Diehr, P. & Patrick, D.L. (1991). Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation, *Controlled Clinical Trials* **12**, 142S–158S.

[9] Diabetes Control and Complications Trial Research Group (1996). Lifetime benefits and costs of intensive therapy as practiced in the diabetes control and complications trial, *Journal of the American Medical Association* **276**, 1409–1415.

[10] Dudley, R.A., Harrell, F.E., Smith, L.R., Mark, D.B., Califf, R.M., Pryor, D.B., Glower, D., Lipscomb, J. & Hlatky, M. (1993). Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery, *Journal of Clinical Epidemiology* **46**, 261–271.

[11] Drummond, M.F. & Davies, L.M. (1991). Economic analysis alongside clinical trials: revisiting the methodological principles, *International Journal of Technology Assessment in Health Care* **7**, 561–573.

[12] Drummond, M.F. & O'Brien, B.J. (1993). Clinical importance, statistical significance and the assessment of economic and quality of life outcomes, *Health Economics* **2**, 205–212.

[13] Efron, B. & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials, *Journal of the American Statistical Association* **86**, 9–17.

[14] Fleming, T.R. & DeMets, D.L. (1996). Surrogate endpoints in clinical trials: are we being misled?, *Annals of Internal Medicine* **125**, 605–613.

[15] Fleming, T.R., Prentice, R.L., Pipe, M.S. & Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research, *Statistics in Medicine* **13**, 955–968.

[16] Fletcher, A.E., Gore, S.M., Jones, D.R., Fitzpatrick, R., Spiegelhalter, D.J. & Cox, D.R. (1992). Quality of life measures in health care. II: design, analysis, and interpretation, *British Medical Journal* **305**, 1145–1148.

[17] Fryback, D.G., Lawrence, W.F., Martin, P.A., Klein, R. & Klein, B.E.K. (1997). Predicting quality of well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study, *Medical Decision Making* **17**, 1–9.

[18] Gelber, R.D., Cole, B.F., Gelber, S. & Goldhirsch, A. (1995). Comparing treatments using quality adjusted survival: the Q-TWIST method, *American Statistician* **49**, 161–169.

[19] Gormley, G.J., Stoner, E., Bruskewitz, R.C., Imperato-McGinley, J., Walsh, P.C., McConnell, J.D., Andriole, G.L., Geller, J., Bracken, B.R., Tenover, J.S., Vaughan, E.D., Pappas, F., Taylor, A., Binkowitz, B. & Ng, J., for the Finasteride Study Group (1992). The effect of finasteride in men with benign prostatic hyperplasia, *New England Journal of Medicine* **327**, 1185–1191.

[20] Gould, A.L., Rossouw, J.E., Santanello, N.C., Heyse, J.F. & Furberg, C.D. (1995). Cholesterol reduction yields clinical benefit. A new look at old data, *Circulation* **91**, 2274–2282.

[21] Guyatt, G.H., Walter, S. & Norman, G. Measuring change over time: assessing the usefulness of evaluative instrument, *Journal of Chronic Disease* **40**, 171–178.

[22] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

[23] Hunt, S.M., McEwen, J. & McKenna, S.P. (1986). *Measuring Health Status*. Croom-Helm, London.

[24] Jaeschke, R., Singer, J. & Guyatt, G.H. (1989). Ascertaining the minimal clinically important difference, *Controlled Clinical Trials* **10**, 407–415.

[25] Juniper, E.F., Guyatt, G.H., Epstein, R.S., Ferrie, F.J., Jaeschke, R. & Hillier, T.K. (1992). Evaluation of impairment of health-related quality of life in asthma: development of a questionnaire for use in clinical trials, *Thorax* **47**, 76–83.

[26] Kaplan, R.M. & Anderson, J.P. (1988). A general health policy model: update and application, *Health Services Research* **23**, 203–235.

[27] Karnofsky, D.A. & Burchenal, J.H. (1949). Clinical evaluation of chemotherapeutic agents in cancer, in *Evaluation of Chemotherapeutic Agents*. C.M. Macleod, Ed. Columbia University Press, New York.

[28] Mark, D.B., Hlatky, M.A., Califf, R.M., Naylor, C.D., Lee, K.L., Armstrong, P.W., Barbash, G., White, H., Simoons, M.L., Nelson, C.L., Clapp-Channing, N., Knight, D., Harrell, F.E., Simes, J. & Topol, E.J. (1995). Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction, *New England Journal of Medicine* **332**, 1418–1424.

[29] O'Brien, B.J., Drummond, M.F., LaBelle, R.J. & Willian, A. (1994). In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care, *Medical Care* **32**, 150–163.

[30] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1087.

[31] Oster, G., Borok, G.M., Menzin, J., Heyse, J.F., Epstein, R.S., Quinn, V., Benson, V., Dudl, R.J. & Epstein, A. (1996). Cholesterol-reduction intervention study (CRIS): a randomized trial to assess effectiveness and costs in clinical practice, *Archives of Internal Medicine* **156**, 731–739.

[32] Pedersen, T.R., Kjekshus, J., Berg, K., Olsson, A.G., Wilhelmsen, L., Wedel, H., Pyorala, K., Miettinen, T., Haghfelt, T., Faergeman, O., Thorgeirsson, G., Jonsson, B. & Schwartz, J.S., for the Scandinavian Simvastatin Survival Study Group (1995). Cholesterol lowering and the use of healthcare resources, results of the Scandinavian Simvastatin Survival Study, *Circulation* **93**, 1796–1802.

[33] Pocock, S.J. (1996). Clinical trials: a statistician's perspective, in *Advances in Biometry*, P. Armitage & H.A. David, eds. Wiley, New York.

[34] Powe, N.R. & Griffiths, R.I. (1995). The clinical-economic trial: promise, problems, and challenges, *Controlled Clinical Trials* **16**, 377–394.

[35] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.

[36] Scandinavian Simvastatin Survival Study Group (1994). Randomized trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S), *Lancet* **344**, 1383–1389.

[37] Shepherd, J., Cobbe, S.M., Ford, I., Isles, C.G., Lorimer, A.R., MacFarlane, P.W., McKillop, J.H. & Packard, C.J. (1995). Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia, *New England Journal of Medicine* **20**, 1301–1307.

[38] Torrance, G.W. & Feeny, D. (1989). Utilities and quality adjusted life years, *International Journal of Technology Assessment in Health Care* **5**, 559–575.

[39] Ware, J. & Sherbourne, C.D. (1992). The MOS 36-item short form health survey (SF-36). I. Conceptual framework and item selection, *Medical Care* **30**, 473–483.

[40] World Health Organization (1947). The constitution of the World Health Organization, *WHO Chronicle* **1**, 29.

JOSEPH F. HEYSE

Outcomes Research

The multidisciplinary field of study that seeks to understand and improve the end results of particular health care practices and interventions is now broadly and commonly known as outcomes research [9]. The US Agency for Healthcare Research and Quality defines such end results to include not only survival and other biomedical endpoints, but also such patient-reported outcomes as symptom status, functional status, and experiences with the health care system [1]. At the US National Cancer Institute, outcomes research "describes, interprets, and predicts the impact of various influences, especially (but not exclusively) interventions on 'final' endpoints that matter to decision makers: patients, providers, private payers, government agencies, accrediting organizations, and society at large" [20, p. III-4]. In cancer, these final endpoints (outcomes) include not only survival and disease-free survival (*see* **Survival Analysis, Overview**), but health-related **quality of life**, patient perceptions of and satisfaction with health care, and economic burden. Final outcomes are to be distinguished from a variety of "intermediate" outcomes, for example, smoking quit rates, colorectal cancer screening rates, tumor shrinkage rates. While these latter are central to evaluating the proximate success of a particular prevention, screening, or treatment intervention, the pivotal question in almost every instance is whether improvement in the intermediate outcome increases the likelihood of improvement in one or more final outcomes.

Outcomes research, so defined, and **quality-of-care** research are not synonymous, but are closely linked. Indeed, much of the logic and vocabulary of outcomes research can be traced to Donabedian's seminal work on assessing and assuring the quality of health care [12, 13]. His paradigm usefully distinguishes three components of health care: "structure" (the nature and quantity of the physical, human, and financial resources available for providing care); "process" (the health care services and products delivered to individuals and populations), and "outcome" (the resulting impact on health and well-being). The paradigm posits a fundamental functional relationship among the components (see Figure 1). The structure of care feeds into and supports the production of the processes of care, and structure and process together influence outcomes. While, as Donabedian notes, this tripartite typology is a simplifying abstraction of the complex reality of healthcare delivery, it offers what has become an enduring framework for defining and analyzing the quality of care. Specifically, quality may be measured by the appropriateness of structure (e.g. was the physician specialty certified?), process (e.g. was the service or procedure performed consistent with evidence-based practice guidelines?), or outcome (e.g. did the patient have improved length of life or quality of life?).

While the valued end product of the healthcare system is good outcomes, it is not always easy or even feasible to judge quality by outcomes alone. For example, the 40-year old hypertensive who begins pharmacological and behavioral therapy to reduce blood pressure can expect to see major "final outcome" benefits only years later, when (all else being equal) she will be at lower risk than otherwise for heart disease, stroke, and microvascular problems. Moreover, all else might not be equal. If this individual later develops adult onset diabetes, she may yet suffer similar target organ damage even if her blood pressure is controlled. In cases like this (and they are numerous in the chronic disease domain), when the outcomes of interest cannot be immediately observed or else might be influenced by myriad systematic and random factors, the focus of quality-of-care assessment shifts to process or structure variables. In particular, quality is frequently measured by the extent to which the processes of care *that are expected to produce valued outcomes* are performed. In the example above, reducing this individual's blood pressure may be judged good quality care because substantial evidence exists that, on average, it will eventually lead to improvements in final outcomes that matter to her.

For outcomes research to achieve its potential to improve health care quality, progress must continue on three broad fronts, leading specifically to: (1) valid, reliable, appropriately responsive, and feasible outcome measures, (2) sound evidence about the impact of interventions on the outcomes of interest (*see* **Evidence-based Medicine**), (3) the capacity and commitment by research investigators and sponsoring organizations to translate findings into information useful to patients, providers, and other decision makers [18].

1. There has been significant progress in recent decades to develop, test, and use measures of patient-reported health outcomes, largely in research studies but sometimes also in clinical practice. These include

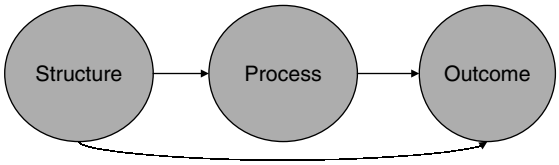

Figure 1 Components of health care

measures of health-related quality of life, at several levels: generic (non-disease-specific) indexes of functional status (*see* **Quality of Life and Health Status**), such as the SF-36 [32] or the Sickness Impact Profile [5]; disease-specific measures, such as the FACT G [8] in cancer or the VF-14 [30] for vision assessment; and even disease-subtype-specific measures such as the FACT-B [28], which combines the FACT G items with additional questions targeted for breast cancer. There are scales measuring patient report of symptom frequency and bother [10]. Measures of global well-being may be preference-based [14, 17] or psychometrically derived (*see* **Psychometrics, Overview**). For evaluating patient perceptions of and satisfaction with healthcare, the CAHPS (Consumer Assessment of Health Plans survey) instrument is being increasingly applied [16]. While there are few standardized instruments for measuring the economic burden of disease, the guiding principle in most applications is the same: capture the explicit or implied monetary value of the resources consumed by disease and its treatment.

Descriptions and analyses of a wide variety of quality-of-life measures are available in [6, 25, 29]. A comprehensive evaluation of the state of the science in cancer outcomes assessment has been completed by the Cancer Outcomes Measurement Working Group, a 35-member task force appointed by the National Cancer Institute [19]. Key findings from the working group underscore the importance of modern psychometric approaches, such as item-response theory modeling (*see* **Rasch Models**), for improving the technical quality and feasibility of patient-reported outcomes assessment. Throughout, this working group was guided by the evaluation criteria recommended by the Medical Outcomes Trust [21].

2. The central inferential problem in outcomes research can be summarized symbolically as: $O = f[P, S, X, E]$, which says that outcome (O) is a function of process (P) and structural variables (S), additional factors (X) such as patient characteristics and **risk factors**, and **random error** (E); it is understood further, that S may influence P, but at a point in time, both may potentially influence O (Figure 1). The approaches available for understanding the impact of P and S on O, while adjusting for X and taking account of E, are of course the standard strategies for scientific inference. These include randomized control trials (*see* **Clinical Trials, Overview**) [7] and a variety of **observational study** designs, ranging from retrospective [4] and prospective **cohort studies** [26], to **case–control studies** [27], to systematic expert judgment and synthesis (e.g. through a Delphi-type process) [22]. Observational studies are increasingly employing techniques such as **instrumental variables** [24] and **propensity scores** [11] in an effort to compensate or correct for **selection biases** arising because patients are not randomized to interventions.

An important motivating concern in outcomes research is whether there is an unexplained or unwarranted variation across population groups in valued health outcomes. In terms of the variables defined above, does O vary systematically with X? If so, does the impact of P and S on O also vary with X (that is, should there be **interaction** terms allowing for the possibility that the observed effectiveness of interventions may vary across population groups)?

While most such analyses are still conducted by individual investigators or small teams, there is a growing trend towards larger collaborative projects conducted by multidisciplinary groups comprising clinicians, statisticians, psychometricians, economists, behavioral scientists, and others [3, 23].

3. The challenges in translating the findings of outcomes research into knowledge useful to decision makers have been underscored by a recent US government report examining "The Outcomes of Outcomes Research" [31]. The report framed its analysis around a four-tiered pyramid (see Figure 2) depicting the levels of impact that outcomes research might have. In ascending the pyramid, one moves from research that largely adds to the knowledge base (level 1), to research that affects medical practice policy (level 2), to research that affects medical care delivery (level 3), to research that changes health outcomes (level 4). The report concludes that most of the federally funded outcomes research it reviewed represented level 1 studies that contributed to understanding the **epidemiology** of disease, the impact of specific interventions, or methodological issues (such

Figure 2 Outcomes research levels of impact [31]

as **meta-analysis** or other inferential techniques). There were only a few examples where outcomes research had been incorporated into policy practice or clinical decision making, or where interventions had then been shown to improve patient outcomes [31]. In response, there has been a heightened focus in the first half of this decade on initiatives to enhance the dissemination, diffusion, and adoption of evidence-based interventions (e.g. [2, 15]). At the same time, much (level 1) work remains, to understand better the behavioral, social, economic, and environmental factors that jointly determine whether, and how rapidly, any given evidence-based intervention scales the "outcomes pyramid."

In sum, the central tasks of outcomes research are measurement, inference, and translation into practice. The central aim is to improve outcomes important to decision makers, most typically through efforts to enhance the quality of care.

References

[1] Agency for Healthcare Research and Quality (2003). *Outcomes Research Fact Sheet: What is Outcomes Research*. Available at `http://www.ahrq.gov/clinic/outfact.htm`. Last accessed on December 28, 2003.

[2] Agency for Healthcare Research and Quality (2004). *Translating Research into Practice II: RFA HS-00-008*. Available at `http://grants.nih.gov/grants/guide/rfa-files/RFA-HS-00-008.html`. Last accessed on January 18, 2004.

[3] AHCPR (1990). *AHCPR Program Note: Medical Treatment Effectiveness Research*. Agency for Health Care Policy and Research, Rockville.

[4] Bach, P.B., Cramer, L.D., Warren, J.L. & Begg, C.B. (1999). Racial differences in the treatment of early-stage lung cancer, *New England Journal of Medicine* **341**(16), 1198–1205.

[5] Bergner, M., Bobbit, R.A., Carter, W.B. & Gilson, B.S. (1981). The sickness impact profile: development and final revision of a health status measure, *Medical Care* **19**, 787–805.

[6] Bowling, A. (1994). *Measuring Disease: a Review of Disease-Specific Quality of Life Measurement Scales*. Open University Press, Buckingham.

[7] Bypass Angioplasty Revascularization Investigation (BARI) Investigators (1997). Five year clinical and functional outcome comparing bypass surgery and angioplasty in patients with multivessel coronary disease: a multicenter randomized trial, *Journal of the American Medical Association* **277**, 715–721.

[8] Cella, D.F., Tulsky, D.S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., Silberman, M., Yellen, S.B., Winicour, P. & Brannon, J. (1993). The functional assessment of cancer therapy scale: development and validation of a general measure, *Journal of Clinical Oncology* **11**, 570–579.

[9] Clancy, C.M. & Eisenberg, J.M. (1998). Outcomes research: measure the end results of health care, *Science* **282**, 245–246.

[10] Cleeland, C.S. (1991). Pain assessment in cancer, in *Effect of Cancer on Quality of Life*, D. Osoba, ed. CRC Press, Boca Raton, pp. 293–304.

[11] D'Agostino, R.B.J. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group, *Statistics in Medicine* **17**, 2265–2281.

[12] Donabedian, A. (1985). *Explorations in Quality Assessment and Monitoring*, Vol. 1. *The Definition of Quality and Approaches to Its Assessment*. Health Administration Press, Ann Arbor.

[13] Donabedian, A. (1988). The quality of care: how can it be assessed? *Journal of the American Medical Association* **260**, 1743–1748.

[14] Feeny, D., Furlong, W., Torrance, G.W., Goldsmith, C.H., Zhu, Z., DePauw, S., Denton, M. & Boyle, M. (2002). Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system, *Medical Care* **40**, 113–128.

[15] Feussner, J.R., Kizer, K.W. & Demakis, J.G. (2000). The quality enhancement research initiative (QUERI): from evidence to action, *Medical Care* **38**(6 Suppl.), I1–I6.

[16] Hayes, R.D., Shaul, J.A. & Williams, V.S.L. (1997). Psychometric properties of the consumer assessment of health plans survey measures, *Medical Care* **37**(Suppl.), MS22–MS31.

[17] Kaplan, R.M., Anderson, J.P. & Ganiats, T.G. (1993). The quality of well-being scale: rationale for a single quality of life index, in *Quality of Life Assessment: Key*

Issues in the 1990s, S.R. Walker & R.M. Rossev, eds. Kluwer Academic, London, pp. 65–94.

[18] Lipscomb, J. & Donaldson, M.S. (2003). Outcomes research at the National Cancer Institute: measuring, understanding, and improving the outcomes of cancer care, *Clinical Therapeutics* **25**(2), 699–712.

[19] Lipscomb, J., Gotay, C.C. & Snyder, C.F. (2004). *Outcomes Assessment in Cancer*. Cambridge University Press, Cambridge, forthcoming 2005.

[20] Lipscomb, J. & Snyder, C.F. (2002). The outcomes of cancer outcomes research: focusing on the National Cancer Institute's quality-of-care initiative, *Medical Care* **40**(Suppl.), III-3–III-10.

[21] Lohr, K.N. & (for the Scientific Advisory Committee of the Medical Outcomes Trust) (2002). Assessing health status and quality-of-life instruments: attributes and review criteria, *Quality of Life Research* **11**, 193–205.

[22] McGlynn, E.A., Asch, S.M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A. & Kerr, E.A. (2003). The quality of health care delivered to adults in the United States, *New England Journal of Medicine* **348**, 2635–2645.

[23] National Cancer Institute (2004). *Cancer Care Outcomes Research and Surveillance Consortium (CanCORS)*. Available at `http://healthservices.cancer.gov/cancors/`. Last accessed January 17, 2004.

[24] Newhouse, J.P. & McClellan, M.D. (1998). Econometrics in outcomes research: the use of instrumental variables, *Annual Review of Public Health* **19**, 17–34.

[25] Patrick, D.L. & Erickson, P. (1993). *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. Oxford University Press, New York.

[26] Potosky, A.L., Harlan, L.C., Sanford, J.L., Gilliland, F.D., Hamilton, A.S., Albertsen, P.C., Eley, J.W., Liff, J.M., Deapen, D., Stephenson, R.A., Legler, J., Ferrans, C.E., Talcott, J.A. & Litwin, M.S. (1999). Prostate cancer practice patterns and quality of life: the prostate cancer outcomes study, *Journal of the National Cancer Institute* **91**(20), 1719–1724.

[27] Ridker, P.M., Hennekens, C.H. & Miletich, J.P. (1999). GP20210A mutation in prothrombin gene and risk of myocardial infarction, stroke, and venous thrombosis in a large cohort of U.S. men, *Circulation* **99**, 999–1004.

[28] Ritz, L.J., Nissen, M.J., Swenson, K.K., Farrell, J.B., Sperduto, P.W., Sladek, M.L., Lally, R.M. & Schroeder, L.M. (2000). Effects of advanced nursing care on quality of life and cost outcomes of women diagnosed with breast cancer, *Oncology Nursing Forum* **27**, 923–932.

[29] Spilker, B. (1996). *Quality of Life and Pharmacoeconomics in Clinical Trials*. Lippincott-Raven, Philadelphia.

[30] Steinberg, E.P., Tielsch, J.M., Schein, O.D., Javitt, J.C., Sharkey, P., Cassard, S.D., Legro, M.W., Diener-West, M., Bass, E.B. & Damiano, A.M. (1994). The VF-14: an index of functional impairment in patients with cataract, *Archives of Ophthalmology* **112**, 630–638.

[31] Tunis, S. & Stryer, D. (2004). *The Outcome of Outcomes Research at AHCPR*. Publication No. AHCPR 99-R044. Available at `http://www.ahrq.gov`. Last accessed January 16, 2004.

[32] Ware, J.E. & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36). 1. Conceptual framework and item selection, *Medical Care* **30**, 473–483.

JOSEPH LIPSCOMB

Outliers

Outliers are an important issue in data analysis. Since a single undetected outlier can destroy an entire analysis, analysts should worry about the origins, relevance, detection, and correct handling of outliers.

Intuitively, an outlier is an observation so discordant from the majority of the data as to raise suspicion that it could not plausibly have come from the same statistical mechanism as the rest of the data. Apparent discordancy is what distinguishes an *outlier* from a *contaminant* – an observation that did not come from the same mechanism as the rest of the data but was generated in some other way. A contaminant may appear ordinary (and not outlying), while an outlier could be, but is not necessarily, a contaminant.

Outliers can arise in several ways.

- They may be contaminants generated by some other statistical mechanism. For example, if the seeds used in a plant growth experiment contain some foreign seed, then the plants produced from the foreign seed will be contaminants and may be outliers.
- They may result from procedural errors in data gathering. For example, misreading an instrument will produce a contaminant that may be outlying. It is generally accepted that several percent of even high-quality data are wrong, and some of these errors may be outlying.
- The analyst may have a misconception of the true model (*see* **Misspecification**). For example, if an instrument is thought to produce **normally distributed** readings, but actually produces **Cauchy-distributed** readings, then some valid correct readings will be severe outliers *relative to the wrongly assumed normal model*.
- In structured data such as **multiple regression** or **analysis of variance** data, they may be a symptom of some other model failure. Heteroscedasticity (*see* **Scedasticity**) and nonlinearity, as well as nonnormality, may cause what appear to be isolated outliers.

Different possible origins of outliers call for different resolutions. Where the outliers are caused by imperfect modeling, the model should be refined so that they are accommodated. Where the outliers result from errors of execution, the primary concern is to minimize the damage they do to the analysis. Where the outliers result from mixtures of distributions or other types of contamination, there may be interest in identifying the outliers and (in the mixture case) estimating the characteristics of the mixture components. It is not always clear in applications what caused the outliers, and so "one size fits all" recipes for dealing with outliers are inappropriate.

Here, we concentrate on methods of *identifying* outliers – that is, of deciding whether observations really are implausibly discordant. This has two parts – deciding which of the observations are most discordant from the majority; and measuring their discordancy.

The first task – locating the most discordant observations – seems trivial at first glance, but is so only in the simplest case of univariate **random samples**. Here, it is only the largest and the smallest of the data that could be potentially discordant. Locating the most discordant observations, however, can be very difficult in "structured" data sets such as **time series**, analysis of variance, multiple regression, and **multivariate** data. Here, simple **diagnostics** like regression **residuals** cannot be relied upon to locate discordant observations.

Estimating the parameters of a model without risk of serious damage from outliers is addressed by the techniques of *robust estimation* (*see* **Robustness**), the most familiar example of which may be the box and whisker plot with its outlier identification rules (*see* **Graphical Displays**). The objectives of robust estimation and outlier identification are logically connected – if the potential outliers are located, then removing them from the sample and fitting the model to the remaining data will neutralize them and so provide robust estimates. Similarly, sturdy estimates of unknown parameters will provide a good picture of the model fitting the majority of the data, thereby helping locate potential outliers. The details of both approaches are less simple than they seem though, and the connection is not enough for either technology to supersede the other.

Likelihood Models for Outliers

Outliers may be flagged and investigated in varying degrees of formality. For example, making a normal probability plot (*see* **Normal Scores**) of residuals from a multiple regression and labeling as outliers any cases that seem visually too far from the line

is a method of flagging questionable cases. Being informal though, it suffers from dependence on perceptions of how large a deviation is too large and may be criticized for having no obvious theoretical basis. More formal statistical models for outliers remove the subjective element and are valuable even if used only as benchmarks to assess other less formal models. A general framework that permits the modeling of data sets that might contain outliers is the "parameter shift" model. Each "good" (scalar or vector) observation $\boldsymbol{X}_i$ in the sample is modeled as following a specified statistical model with unknown parameter vector $\boldsymbol{\theta}$

$$\boldsymbol{X}_i \sim g_i(., \boldsymbol{\theta}).$$

There may also be one or more contaminating observations $\boldsymbol{X}_i$ with distribution

$$\boldsymbol{X}_i \sim g_i(., \boldsymbol{\theta}, \alpha_i),$$

where the parameter α_i drives the contaminant's displacement, and is most conveniently parameterized so that a contaminant with $\alpha_i = 0$ has the same distribution as a "good" value. Commonly, different contaminants are modeled as having different α_i, but in some settings, a common α for all is sensible. Contaminants with sufficiently extreme α_i values will be outlying and so potentially identifiable; those with less extreme α_i values will be indistinguishable from the good observations. This model is less restrictive than it might seem. An outlier that resulted from a recording error of transposing digits would not plausibly be explained by this parameter shift model, but since a good choice of α_i could be fit to the data, this conceptual failing is arguably unimportant. For example, a transposition error of writing 84 for X_i instead of 48 is exactly the same as adding $\alpha_i = 84 - 48$ to the value actually generated by the model.

This distributional approach to describing outliers is attractive for those who like to work with formal models, as it allows much of the processing of outliers to be formalized and codified. Using say **maximum likelihood** to estimate the parameters $\boldsymbol{\theta}$ and α_i automatically gives a robust estimate of $\boldsymbol{\theta}$, and the **likelihood** gives a formal outlier test through a test of the **null hypothesis** $\alpha_i = 0$.

The likelihood model describes contaminants, not outliers. Contaminants that are not outlying are undetectable; however, since they are undetectable precisely because they behave like "good"observations, their invisibility means that failing to detect them usually does not have serious bad consequences.

Computational Issues

Fitting the model is less trivial than it might seem as it requires looking at all possible partitionings of the data into the "good" and the "potentially suspect" subsets. In the simplest cases, this can be done by inspection. For a single sample from a normal distribution with mean-shift contamination, it is easy to show that the highest likelihood results when the extreme **order statistics** of the sample are labeled as "potentially suspect". This means that no other types of partitioning need be studied.

In multiple regression, it would be equally intuitive that the cases most likely to be outliers would be those with the largest residuals, or the largest **studentized** residuals. Here though intuition is misleading: a pair of outliers can so distort the regression line that both have small residuals. This is called "masking". They may also make the residual of a "good" case large – this is termed "swamping". Thus, the most discordant observations need not necessarily have large (or even nonzero) residuals. Reliable identification of potentially suspect cases in multiple regression requires the substantial computational exercise of looking at all possible partitioning of the cases into "good" and "potentially suspect" subsets, selecting the partition with the largest likelihood.

"High-breakdown" methodologies do indeed guarantee locating even large numbers of outliers, however badly they may be placed, provided enough computation is done. These methodologies are inherently **computer-intensive** as the guarantee of locating the outliers requires potentially investigating all partitions of the data into an "inner half" and an "outer half". After this exhaustive search, all outliers along with some inliers can be expected to be in the outer half where using some cutoff rule should distinguish the outliers from the more extreme inliers.

A much lighter computation is required for "sequential identification" methods. In these, the single observation from the whole sample whose deletion would lead to the greatest increase in likelihood is flagged as potentially suspect, and temporarily removed from the sample. The single observation in the remaining sample whose deletion would lead to the greatest improvement in model

fit is then identified as another potentially suspect observation, and is also temporarily removed from the sample. This stripping of observations continues until some stopping rule is reached.

Sequential identification methods are of two types. In **"forward selection"**, a discordancy measure is calculated as each new potentially suspect observation is identified and the process stops when it first fails to find an observation sufficiently discordant to be called an outlier. In "backward elimination" a preset number of potentially suspect observations is identified in one pass, and then in a second pass, each of them is tested sequentially to see whether it really is sufficiently discordant to be called an outlier. A case that is not discordant is then reincluded with the "good" observations and the testing of the remaining potentially suspect observations repeated.

Both theoretical reasoning and practical experience show that backward elimination is much better than forward selection, which can miss arbitrarily severe outliers. This is because the masking effect may cause the outlier diagnostics of all the outliers to appear modest so that forward selection stops before all outliers have been located.

Intuitive though sequential identification is, an even more intuitive approach is one-pass identification of all cases whose outlier diagnostics are large. Because of masking and swamping though, this approach is much worse than even forward selection and should not be used at all. This leaves backward elimination as the most reliable approach with low computational requirements.

Another "backward elimination" method starts with some small subset of the data that is outlier-free, and then sequentially adds observations that appear not to be discordant, reestimating parameters as each new apparently concordant observation is added. Success with this approach depends on finding a starting subset of cases that is not only outlier-free but also informative enough to correctly distinguish the inliers from the outliers in the not-yet-classified group. Starting with a high-breakdown estimate gives reliable results, but at high computational cost. Methods using full-sample estimates – for example, the cases with the smallest absolute residuals from a full-sample **least squares** regression fit – may succeed, but come with no guarantees.

Even high breakdown methods are not completely bulletproof: see the example in Hawkins and Olive [5] in which even high breakdown regression methods saw nothing surprising about men less than an inch tall, but whose head circumference was some five feet.

The second phase, of deciding whether to label a suspect case outlying, may be based on a **likelihood ratio** or a case diagnostic such as a studentized residual. Two different philosophies on outlier labeling are to use fixed cutoff values; and to use **multiple comparison** tests. An example of the first approach is to flag as outlying any cases whose regression residual exceeds 2.5 standard deviations regardless of sample size. This rule will delete a fixed percentage of the *cases* in data sets consisting of only good data. A sound multiple comparison method is the **Bonferroni** approach of testing the externally studentized residual ("outlier t", or RSTUDENT) of a regression against the α/n **quantile** of a **Student's *t* distribution**, where n is the sample size and α a significance **level**. In this approach, a fixed proportion of good *data sets* will have one or more observations wrongly labeled outliers.

Particular Cases

The easiest situation is mean-shift outliers from a univariate normal distribution, with inliers distributed as $N(\mu, \sigma^2)$, both parameters unknown, contaminated by one or more $N(\mu + \alpha_i, \sigma^2)$ outliers. The scaled deviations from the mean, $(X_i - \overline{X})/s$, where $\overline{X}$ is the sample mean and s the standard deviation, are effective for both identifying and testing a single outlier. Multiple outliers can be found using sequential identification either by starting with the full sample and stripping one suspect observation at a time, or by starting with the easy-to-find "inner half" of cases and adding concordants.

Outliers in the univariate normal can also be modeled by scale contamination, where the outliers are $N(\mu, \sigma^2(1 + \alpha))$. This (with a common α) can be thought of as mixing the mean-contamination model's displacements over a $N(0, \alpha\sigma^2)$ distribution, and is effective for outliers occurring symmetrically to the left and the right of the inliers.

Flagging and testing for outliers from a **chi-square distribution** use the scaled quantities $X_i/\overline{X}$. Cochran's test for the largest such ratio is classical, but outlier tests can also be performed on the smallest such ratios.

In homoscedastic normal linear modeling (both linear regression and analysis of variance), a general-purpose approach to outlier identification and testing is by variance analysis. Let S_0 with ν degrees of freedom denote the residual sum of squares from a fitted model using some set of cases. Write S_1 for the residual sum of squares obtained after deleting a suspect observation. Then, the pseudo-F ratio

$$F = \frac{(\nu - 1)(S_0 - S_1)}{S_1} \tag{1}$$

may be compared with the fractiles (Bonferroni-corrected or fixed) of an ***F* distribution** with 1 and $\nu - 1$ **degrees of freedom** to decide whether the suspect case is or is not within plausible limits. This F ratio is the square of the "outlier t" statistic that software often produces as a case diagnostic.

In analysis of variance with replicates, fitting the ANOVA model on the full sample and after removal of individual readings leads to tests for individual outliers. A different model is the "slippage" model in which all the replicate readings in some cell are displaced by the same amount. Reducing the original data to cell means and taking these means as the basic data to which to apply outlier screening methods addresses the slippage model.

Multivariate outliers are commonly modeled by the **multivariate normal distribution** $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown mean vector and **covariance matrix**. The mean contamination model holds that the outliers are $N(\boldsymbol{\mu} + \boldsymbol{\alpha}_i, \boldsymbol{\Sigma})$, and mixing the outlier displacements over a zero-mean normal distribution leads to the scale-contamination model in which the outliers are $N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Gamma})$. With this baseline model, if the sample mean vector and covariance matrix are written as $\overline{X}$ and S respectively, then the discrepancy of a single case X_i can be measured by its squared **Mahalanobis distance** from the mean $D_i = (X_i - \overline{X})' S^{-1} (X_i - \overline{X})$. The traditional multivariate outlier statistic, Wilk's lambda, is a monotonic function of D_i. Sequential deletion successively trims the case with the largest D_i from the current sample mean vector using the current sample covariance matrix.

The likelihood outlier model can handle **generalized linear models**. If the deviance of any fitted model is S_0 and the deviance obtained by deleting some suspect case and refitting is S_1, then the deviance explained by deleting the case is $S_0 - S_1$, and this can be referred to its asymptotic chi-squared distribution to get an outlier test. This framework covers **Poisson** and **logistic regression** and **loglinear modeling**, among others.

The generalized linear model framework is also useful for outliers in **contingency tables** – for example individual cells whose frequencies violate independence, which holds in the rest of the table. Deleting a particular cell, refitting the model, and computing the change in deviance between the two fits gives an outlier test statistic for that cell.

Time series can have two types of outlier – an "additive outlier" displaces a single reading from where it should have been, but does not affect the subsequent observations. "Innovative outliers" displace the whole time series from their point of occurrence onward, and, apart from their occurrence at an unexpected time, can be modeled by the intervention analysis likelihood.

Example

Rousseeuw and Leroy [7] discuss a data set relating the body weight and brain weight of 25 animals alive today, along with three dinosaurs. Least-squares regression of $Y = \log_e$ (brain weight) on $X = \log_e$ (body weight) gives (standard errors in parentheses)

$$\begin{aligned} Y = 2.555 \; &+ \; 0.496 \quad X \\ &(0.413) \quad (0.078). \end{aligned} \tag{2}$$

Species	Body weight	Brain weight	Outlier t after deletion number				
			0	1	2	3	4
Diplodocus	11 700	50.0	−2.507	−2.507	−3.646	−6.649	
Brachiosaurus	87 000	154.5	−2.505	−3.644	−3.644	−6.804	
Triceratops	9400	70.0	−2.094	−2.835	−6.045	−6.045	
Human	62	1320.0	1.789	1.861	2.098	3.232	
Residual SS			60.988	48.733	31.372	12.117	8.217

With 28 data points, a maximum of four deletions (15% of the data) seems reasonable. Starting with the full data set, and sequentially deleting the case with the largest absolute externally studentized residual for four iterations gives the following set of externally studentized residual of each of the four suspect cases and residual sum of squares of the fitted regression.

The pseudo-F ratios for the successive deletions are 6.29, 13.28, 36.55, and 10.44 with 1 numerator degree of freedom and denominator degrees of freedom 25, 24, 23, and 22. Start with the last of these, 10.44, whose right tail area is 0.004. Multiplying this figure by the remaining sample size, 25, gives a Bonferroni ***P* value** of 0.1, which argues against making the fourth deletion. Going to the next F ratio of 36.55 gives a P value of 3.6×10^{-6}, indicating that the deletion of the third dinosaur, and by implication, its two even more extreme companions, is clearly called for.

With an initial sample size of 28, the Bonferroni 5% and 1% points for the outlier t statistic would be the two-sided $5/28 = 0.179$ and $1/28 = 0.036\%$ points of a t distribution with 25 degrees of freedom, which are 3.50 and 4.13, respectively. None of these four cases is close to significance, illustrating the masking effect. Successively deleting the three dinosaurs makes humans the most outlying species. At this point, humans' t of 3.232 corresponds to a right tail area of 0.0038, which after Bonferroni adjustment, makes us unremarkable. The regression found after deleting the dinosaurs is

$$\begin{array}{cccc} Y = 2.150 & + & 0.752 & X \\ & & (0.201) \quad (0.046). & \end{array} \tag{3}$$

Evidence of the damaging effect of the dinosaurs is that the slope fitted after their removal is 3.3 standard errors away from the full-sample estimate.

The high breakdown regression fit using the least **trimmed** squares criterion gives the model $X = 1.88 + 0.776W$, close to what we get with ordinary regression after deleting the dinosaurs.

Further Reading

The text by Barnett and Lewis [2] provides an exhaustive coverage of the standard outlier situations and models. Additional useful theoretical material on the robust estimation aspects of outliers may be found in Hampel et al. [3]. Rousseeuw and Leroy [7] is a valuable exposition of the high breakdown approach, though readers should be aware that many of the specifics (notably computational procedures and outlier flagging rules) are now obsolete. Better **algorithms** for high breakdown estimation in the difficult multivariate location/scatter case are given in [6, 8], and for the regression case, in [4].

Atkinson and Riani [1] provide discussion centered on sequential identification starting from an outlier-free subset of cases.

Hawkins and Olive [5] provide theoretical and empirical support for routine use of a variety of multiple regression methods, including sequential outlier identification, in analysis of real data. They also argue that in games against nature, in which a malicious opponent tries to place outliers where you cannot find them in a tolerable amount of time, the opponent will always win, given large enough data sets. In other words, it is impossible to guarantee finding even huge outliers in a large data set if an opponent is allowed to hide them. This argument generalizes to other outlier settings – for example, multivariate location/scatter.

References

[1] Atkinson, A. & Riani, R. (2000). *Robust Diagnostic Regression Analysis*. Springer, New York.

[2] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*. Wiley, New York.

[3] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley, New York.

[4] Hawkins, D.M. & Olive, D.J. (1999). Improved feasible solution algorithms for high breakdown estimation, *Computational Statistics and Data Analysis* **30**, 1–11.

[5] Hawkins, D.M. & Olive, D.J. (2002). Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm. (with discussion), *Journal of the American Statistical Association* **97**, 136–148, rejoinder 156–159.

[6] Rocke, D.M. & Woodruff, D.L. (2001). Robust cluster analysis and outlier identification,in 2000 Proceedings, American Statistical Association, Arlington, VA, Biometrics Section.

[7] Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

[8] Rousseeuw, P.J. & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics* **41**, 212–223.

DOUGLAS M. HAWKINS

Overdispersion

The phenomenon which has come to be termed *overdispersion* arises when the empirical variance in the data exceeds the nominal variance under some presumed model. Overdispersion is often observed in the analysis of discrete data, for example count data analyzed under a Poisson assumption, or data in the form of proportions analyzed under a binomial assumption. Support for overdispersion is most likely first obtained when the "full" model is fitted (*see* **Generalized Linear Model**), and the Pearson or deviance **residuals** are predominantly too large [4]; the corresponding Pearson and deviance **goodness-of-fit** statistics indicate a poor fit (*see* **Chi-square Tests**).

The **Poisson** and **binomial** distributions are derived from simple, but fairly strict assumptions, and it is not surprising that these do not apply generally in practice. Fitting either of these distributions assumes a special mean–variance relationship, since both distributions are fully characterized by a single parameter. This can be contrasted with the analysis of continuous data using a normal assumption. Consider a simple example where a set of univariate observations is modeled by a common $N(\mu, \sigma^2)$ distribution. In estimating the fitted distribution, the sample average of the data points defines the location of the normal distribution on the number line (μ), while the sample variance determines the spread of the fitted bell-curve (σ^2). The normal distribution is characterized by two parameters, while the Poisson and binomial distributions are completely specified when only the mean, or the probability of success, respectively, is determined; the variance is fixed by the mean.

The existence of overdispersion is not a recent observation. Student [27] comments upon this problem, and Fisher [13] discusses a goodness-of-fit statistic for testing the adequacy of the Poisson distribution in the single sample problem, i.e. when the counts $Y_i, i = 1, \ldots, n$, are assumed to be independent variates from a Poisson distribution with common mean. The statistic is $\sum_{i=1}^{n}[(Y_i - \overline{Y})^2/\overline{Y}]$, where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$, and is called the sample index of dispersion (*see* **Poisson Distribution**).

An important question to ask when a model is suspect is: Will the lack-of-fit affect inference and lead to incorrect conclusions? If the effect is negligible, and if the efforts involved in fitting a more "exact" model are substantial, then the approximate inference obtained under the simpler model may well suffice. In what follows the effect of overdispersion is shown to be nonignorable, and can be quite drastic, so inference under a Poisson or binomial model when overdispersion is present may be very misleading. We focus on the analysis of count and categorical responses, because these are the two areas where overdispersion most commonly arises, with binary responses being an important special case of the latter.

Effect of Overdispersion on Standard Poisson or Binomial Analyses

The effect of overdispersion is determined primarily by how it arises and its degree of incidence. One common way that overdispersion arises in the analysis of proportions is through a failure of the binomial independence assumption. In animal litter studies, for example, responses of animals in a litter are often positively correlated; so, too, in dental studies for responses on individual teeth for a single individual. Let Y denote a binomial response, $Y = \sum_{j=1}^{m} Y_j$, where Y_j are independent binary variates taking values 0 or 1 with probabilities $(1 - p)$ and p, respectively. Then, $E(Y) = mp$, and $\text{var}(Y) = mp(1 - p)$. If the independence assumption does not hold, and the correlation $(Y_j, Y_k) = \tau > 0$, then $E(Y) = mp$, and $\text{var}(Y) = \text{var}(\sum_{j=1}^{m} Y_j) = mp(1 - p)[1 + \tau(m - 1)]$, leading to overdispersion with respect to the binomial model. Note that $\tau < 0$ leads to underdispersion, which is rare in practice, but might correspond to competition among the binary variates for a positive response.

Another way overdispersion may arise is through a failure of the binomial assumption of constant probability of success from trial to trial. This might occur if the population can be subdivided into naturally occurring subunits, for example colonies, where the probability of a positive response varies over these subunits.

Similarly, for count data, failure of the assumptions underlying the use of the Poisson distribution generally leads to overdispersion. In particular, the probability of an event may vary over individuals or over time. For a simple example, suppose the

response is the number of days absent due to illness over a period of time, in a situation where the number of episodes of illness, Y, are Poisson (μ) distributed but will likely lead to consecutive days of absence. The distribution of the number of days of absence due to illness will then be overdispersed with respect to the Poisson model. If A represents the number of days absence during a single episode, A and Y being independent, then the total number of days of absence is $\sum_{i=1}^{Y} A_i$, which has mean and variance, $\mathrm{E}(\sum_{i=1}^{Y} A_i) = \mu\mathrm{E}(A)$, and $\mathrm{var}(\sum_{i=1}^{Y} A_i) = \mu\mathrm{E}(A)\{\mathrm{E}(A)+[\mathrm{var}(A)/\mathrm{E}(A)]\} > \mu\mathrm{E}(A), \text{if} \mathrm{E}(A) > 1$.

In some contexts the Poisson or binomial variation is only a minute part of the overall variability. This is typical, for example, in cancer **mapping** studies, where the distribution of rates over a region is to be determined (*see* **Clustering**). In many cases it is the spatial variation of the cancer mortality rates which is the main component of the dispersion.

Overdispersion cannot be ignored. In fact, many statistical packages routinely incorporate overdispersion (*see* **Software, Biostatistical**). The magnitudes and signs of the estimated covariate effects in a loglinear or logistic analysis can be quite similar whether or not overdispersion is properly accounted for, so a researcher may gain no hint of an inappropriate analysis by there being strikingly strange estimated effects. However, the standard errors of the estimated regression parameters will be underestimated; these will reflect only the Poisson or binomial variation. The precision of the resulting estimates will be too high and P values for testing the significance of the included covariates will be correspondingly too low. This will very likely lead to incorrect inference, unless the overdispersion is almost negligible.

Testing for Overdispersion

In many situations the presence of overdispersion is clearly indicated by the presence of overly large values of the Pearson or deviance goodness-of-fit statistics, even when the full model is fitted. Formal tests for overdispersion have also been discussed in the literature. Score tests (*see* **Likelihood**) for overdispersion [3, 9, 10] compare the sample variance with what is expected under the model. For the testing of extra-Poisson variation, the adjusted score test statistic, for testing the null hypothesis H_1: $\tau = 0$ in the model with overdispersed variance function $\mu_i + \tau\mu_i^2$, is

$$T_{P1} = \frac{\sum_{i=1}^{n}\{(y_i - \hat{\mu}_i)^2 - (1 - \hat{h}_i)\hat{\mu}_i\}}{\left(2\sum_{i=1}^{n}\hat{\mu}_i^2\right)^{1/2}}. \qquad (1)$$

In (1) h_i is the ith diagonal element of the hat matrix, $\mathbf{H}$, for **Poisson regression**; $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$, where $\mathbf{W} = \mathrm{diag}(\mu_1, \ldots, \mu_n)$, and $\mathbf{X}$ is $n \times p$ with ijth entry $\mu_i^{-1}(\partial\mu_i/\partial\beta_j)$, $\mu_i = \mu_i(\mathbf{x}_i; \boldsymbol{\beta})$. For loglinear regression, $\log\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$, and $\mathbf{X}$ is the usual matrix of covariates. Estimates $\hat{\mu}_i$ and $\hat{h}_i$ are obtained by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, its maximum likelihood estimate under the Poisson assumption. The statistic T_{P1} converges in distribution, as $n \to \infty$, to a standard normal under H_1. For testing the null hypothesis H_2: $\tau = 0$, in the model with variance function $(1 + \tau)\mu_i$, the score test statistic is

$$T_{P2} = \frac{1}{\sqrt{(2n)}}\sum_{i=1}^{n}\left\{\frac{(y_i - \hat{\mu}_i)^2 - (1 - \hat{h}_i)\hat{\mu}_i}{\hat{\mu}_i}\right\}, \qquad (2)$$

which is also asymptotically ($n \to \infty$) distributed as standard normal, under H_2.

It is interesting to note that the test statistic for testing H_2 when considering $\mu_i \to \infty$ asymptotics, for fixed n, is equivalently

$$T'_{P2} = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \qquad (3)$$

which has a limiting distribution of $\chi^2(n - p)$. This is just the Pearson statistic, which is traditionally used to assess correct specification of the mean. The derivation of the tests for extra-Poisson variation assumes that the regression specification is correct. Hence, the Pearson statistic arises as a test of either of two alternative hypotheses: namely, $\mu_i = \mu_i(\mathbf{x}_i, \boldsymbol{\beta})$ is incorrectly specified, or the distribution of the counts has variance form $\phi\mu_i$. We would not, however, interpret a significantly large Pearson statistic as indicating overdispersion in a **generalized linear model**, unless the mean had been reasonably well modeled. Otherwise, the apparent overdispersion could reflect missing covariates, e.g. interaction terms, implying systematic lack of fit, or, the functional form of the mean may be inappropriate. Pregibon [22] develops

a test for checking the form of the mean in generalized linear models; for multinomial models, O'Hara Hines et al. [21] develop diagnostic tools for this purpose.

Smith & Heitjan [25] develop a test for overdispersion which results when the vector of coefficients in the mean is considered to be random. The score tests T_{P1} and T_{P2} arise from a random intercept model, and can therefore be considered a special case of Smith & Heitjan's test. Dean [8] discusses testing for overdispersion with **longitudinal** count data.

A score test for the adequacy of the binomial model against an alternative with variance form $m_i p_i(1-p_i)[1+\tau(m_i-1)]$ is

$$T_B = \frac{\left(\sum_{i=1}^{n} (\hat{p}_i(1-\hat{p}_i))^{-1}[(y_i - m_i\hat{p}_i)^2 + \hat{p}_i(y_i - m_i\hat{p}_i) - y_i(1-\hat{p}_i)] \right)}{\left[2\sum_{i=1}^{n} m_i(m_i-1) \right]^{1/2}}, \quad (4)$$

which has a standard normal distribution as $n \to \infty$. This variance form is obtained from the correlated binomial model, discussed in the previous section. Prentice [23] derives this statistic as a score test statistic against **beta-binomial** model alternatives; Tarone [28] derives it by considering correlated binomial alternatives.

As an example in the application of the tests, consider the data, given in [29], from a clinical trial of 59 patients with epilepsy, 31 of whom were randomized to receive the anti-epilepsy drug Progabide and 28 of whom received a placebo. The total seizure count over four follow-up periods is taken here as the response variable. Table 1 shows the results of a Poisson regression analysis of the effect of Progabide on seizure rate which includes two covariates and their interactions, plus terms for a main and interaction treatment effect. The data and this fitted model have been discussed at length in Breslow [4] and his results are given here. The Pearson and deviance goodness-of-fit statistics clearly indicate lack-of-fit of the Poisson model. The score test statistics T_{P1} and T_{P2} have observed values 36.51 and 37.56, respectively. The overdispersion here is very substantial.

Table 1 Loglinear Poisson regression fit to the epilepsy data; case no. 207, with high leverage, omitted. Reproduced from *Statistica Applicata*, Vol. 8, pp. 23–41, by permission of Rocco Curto Editore

Coefficient	Value	Std. error	t Statistic
Intercept	3.079	0.451	6.833
ln(base count/4)	−0.074	0.201	−0.366
Age/10	−0.511	0.153	−3.332
ln(base count/4): age/10	0.351	0.068	5.164
Progabide	−0.610	0.191	−3.197
Progabide: ln(base count/4)	0.204	0.088	2.325

Deviance = 408.4; Pearson $\chi^2 = 456.52$; df = 52.

Accounting for Overdispersion

With overdispersion present, the use of the Poisson or binomial **maximum likelihood** equations for estimating the regression parameters in the mean is still valid. These are the usual generalized linear model (GLM) or **quasi-likelihood** (QL) estimating equations, and they are unbiased estimating equations regardless of any misspecification of the variance structure. However, the estimated variances of the parameter estimates will be in error, and possibly severely so.

If there are alternative "overdispersed" models which are postulated, then certainly one could proceed by maximum likelihood estimation. This will be discussed further below. There are, however, some robust, simple methods for adjusting standard errors to account for overdispersion and these will be considered first.

McCullagh & Nelder [20] suggest a simple adjustment, which is to multiply $\text{var}(\hat{\boldsymbol{\beta}})$, obtained from the Poisson or binomial model, by an estimate of the GLM scale factor, ϕ. This estimate is usually the Pearson or deviance statistic divided by its degrees of freedom. This is appropriate if the overdispersion gives rise to a variance model which is a constant times the nominal variance, e.g. $\phi\mu$ for counts and $\phi mp(1-p)$ for proportions. This variance form may also well approximate other, possibly more complicated, variance structures in certain situations; for example, for count data, when $\text{var}(y_i) = \mu_i + \tau\mu_i^2$ and $\tau\mu_i$ do not vary greatly with i.

If the sample is large, then an empirical variance estimate can be computed. This is called the "sandwich" variance estimate, cf. Liang & Zeger [17].

For loglinear models, for example, the sandwich estimator is

$$\text{var}(\hat{\beta}) = \left[\sum_{i=1}^{n} \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \left[\sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)^2 \mathbf{x}_i \mathbf{x}_i'\right] \times \left[\sum_{i=1}^{n} \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1}. \quad (5)$$

Unless very large samples are available, the sandwich estimator tends to underestimate the true variance.

Resampling techniques, although typically **computer-intensive**, have become popular for providing estimates of the variance of regression parameters. **Bootstrap** and **jackknife** estimates are discussed in [12], and there are methods of approximating these which require less computing effort, e.g. the one-step jackknife estimate. The bootstrap estimates are considered to be quite accurate. Table 2, from [4], compares these estimators in the analysis of the data from [29], mentioned earlier. Notice how much larger these estimates of the standard errors are compared with those obtained from the Poisson model. The treatment effect is no longer significant when overdispersion is taken into account.

Model-based methods for incorporating overdispersion lead to mixture models or **random effects models**. A simple Poisson random effects model can be derived by considering a model with individual-specific random effects $\nu_i > 0$ where, conditional on $(\nu_i, \mathbf{x}_i)$, the distribution of Y_i is Poisson with a mean of $\nu_i \mu_i(\mathbf{x}_i; \boldsymbol{\beta})$, and the ν_i are continuous, independent variates with probability density function $p(\nu; \tau)$ depending on a parameter τ. If $\mu_i(\mathbf{x}_i; \boldsymbol{\beta})$ takes the common form $\exp(\mathbf{x}_i' \boldsymbol{\beta})$, then the fixed and random effects are added on the logarithmic scale and the random effects can be construed as representing covariates which are unavailable. The probability function of Y in the mixed model is

$$\int_0^\infty \frac{(\mu\nu)^y \exp(-\mu\nu)}{y!} p(\nu)\, \mathrm{d}\nu \quad (6)$$

and the score function for estimating the regression parameters has the intuitively appealing form

$$\sum_{i=1}^{n} [y_i - \mu_i \mathrm{E}(\nu_i | y_i)] \frac{1}{\mu_i} \left(\frac{\partial \mu_i}{\partial \beta_r}\right) = 0. \quad (7)$$

This equation, with $\mathrm{E}(\nu_i | y_i)$ omitted, is the maximum likelihood equation for Poisson regression [Eq. (8), with $\sigma_i^2 = \mu_i$]. If the distribution of ν is specified, then full maximum likelihood estimation can be performed; if ν is assumed to be **gamma**, then this leads to a **negative binomial distribution** for the counts. However, it is more common to adopt the more robust approach of specifying only the first two moments for Y, i.e. μ_i and σ_i^2, respectively; the parameters in the mean are then estimated using the quasi-likelihood estimating equation,

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\sigma_i^2} \left(\frac{\partial \mu_i}{\partial \beta_r}\right) = 0, \quad (8)$$

for count data, together with another estimating equation for the additional parameter τ in σ_i^2.

There are many important reasons for the widespread use of the quasi-likelihood approach. For generalized linear models with a full likelihood, these are the maximum likelihood equations. From the viewpoint of estimating equations (*see* **Estimating Functions**), Godambe & Thompson [14] (see also Nelder's discussion of that paper) derive important optimality properties of the estimators. When $\sigma^2 = \mu\tau$, estimation of the regression coefficients is not affected by the value of τ. Estimation is easy with standard software. An important point is that

Table 2 Overdispersion adjusted standard errors for Table 1 coefficients. Reproduced from *Statistica Applicata*, Vol. 8, pp. 23–41, by permission of Rocco Curto Editore

Coefficient	Scale method	Sandwich	One-step jackknife	True jackknife	Bootstrap (nb = 5000)
Intercept	1.263	0.711	0.792	0.792	0.870
ln(base count/4)	0.564	0.326	0.368	0.369	0.424
Age/10	0.430	0.237	0.264	0.263	0.291
ln(base count/4): age/10	0.190	0.104	0.117	0.117	0.137
Progabide	0.535	0.403	0.440	0.448	0.466
Progabide: ln(base count/4)	0.246	0.188	0.210	0.214	0.226

the asymptotic variance of $\tilde{\boldsymbol{\beta}}$, the estimate of $\boldsymbol{\beta}$, is independent of the choice of the estimating function for τ, and depends only on the first two moments of the distribution. This is also, asymptotically, a very **efficient estimator** for a wide range of models. Simulation studies have been conducted to investigate the performance of $\tilde{\boldsymbol{\beta}}$ in small samples; they support the **unbiasedness** and efficiency of this estimator.

A popular method for estimating τ is **pseudo-likelihood**. Davidian & Carroll [7] derived the pseudo-likelihood estimating equation as the maximum likelihood equation when residuals are normally distributed. An alternative simple choice is equating the Pearson statistic to its degrees of freedom.

Nonparametric methods of modeling random effects for handling overdispersion have been shown to be useful. Lindsay [18] is a comprehensive source on the topic. He discusses the geometry and theory of mixture models and describes a plethora of applications where mixture models are used, including overdispersion, measurement errors, and latent variable models for **cluster analysis**. Practical issues for estimation of nonparametric mixing distributions, such as algorithms and computational problems, are discussed at length in [2] and [16].

The preceding overdispersion models incorporated a single random effect which was independent over subjects. More general models might include multiplicative random effects [25]; large numbers of random effects are common in animal breeding experiments, where it is the prediction of the random effects, representing sire effects, which is the main focus of the study. In studies of the **geographic** distribution of cancer mortality rates, or disease incidence, the random effects may represent area-specific effects and there may be good reason to suspect that such area effects are not independent, and in some circumstances may be quite similar within small neighborhoods.

A general body of theory which synthesizes the incorporation of several random effects, which are not necessarily independent, falls under the heading of generalized linear mixed models. It permits a simple incorporation of overdispersion, as discussed previously, and can also model dependences in outcome variables or random effects, as are typical in repeated measures design (*see* **Nonlinear Mixed Effects Models for Longitudinal Data**). A generalized linear mixed logistic model specifies that

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}, \tag{9}$$

where p_i and $\mathbf{x}_i$ are the probability of a positive response and the vector of covariates, corresponding to the ith proportion, respectively, $\mathbf{z}_i$ is a vector of covariates, and $\boldsymbol{\gamma}$ is distributed with a mean of zero, and finite variance matrix. Conditional on $\boldsymbol{\gamma}$, the responses are supposed binomially distributed. As an aside, note that the representation above elucidates that apparent overdispersion can be induced by missing covariates, or by **outliers**. Residual **diagnostics** are important for identifying the latter.

In generalized linear mixed models the random effects are usually assumed to be Gaussian, and maximum likelihood estimation involves q-dimensional integration; here q is the dimension of $\boldsymbol{\gamma}$. Alternative simpler approaches for inference have been proposed, using generalizations of moment methods or penalized quasi-likelihood [5]. The penalized quasi-likelihood is a Laplace approximation to the integrated likelihood, with some seemingly harmless other approximations added. Breslow & Clayton [5] provide simple algorithms for estimation using an iterative fitting procedure which updates both the parameter values and a modified response variable at each step. They also evaluate the performance of their estimators. It seems that bias corrections are required for small samples, and these are given in [6].

Lee & Nelder [15] discuss hierarchical generalized linear models, where the distribution of the random components is not restricted to be normal, and where, like penalized quasi-likelihood, estimation avoids numerical integration. Maximizing what they call the h-likelihood, a posterior density, gives fixed effects estimators which are asymptotically equivalent to those obtained using the corresponding **marginal likelihood**. Here, "asymptotically" refers to cluster sizes tending to infinity, and the random effects are cluster-specific. This is important to note because many applications with random effects involve several small-sized clusters. In general, their asymptotic arguments require that the total number of random effects remains fixed, as the overall sample size becomes large. However, they also derive properties of their estimators on a model-by-model basis, and some models require less strict assumptions. When there are many random effects to be estimated, none of the procedures described here will be simple,

as can be expected when dealing with complicated mechanisms for incorporating overdispersion.

Technical Reference Texts and Software Notes

The general topic of overdispersion is discussed in McCullagh & Nelder [20, Sections 4.5, 5.5, 6.2]. Chapters 9 and 10 of that text are also relevant and discuss quasi-likelihood and joint modeling of mean and dispersion. Diggle et al. [11] discuss random effects models for longitudinal data, and Chapter 5 of Lindsey [19] is devoted to the topic of overdispersion in models for categorical data.

Software for incorporating overdispersion includes S-PLUS [26], function glm, SAS ([24], procedures LOGISTIC, MIXED GENMOD, and CATMOD), and GLIM [1].

References

[1] Baker, R.J.& Nelder, J.A. (1987). *GLIM, 3.77*. Numerical Algorithms Group, Oxford, England.

[2] Böhning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models, *Journal of Statistical Planning and Inference* **47**, 5–28.

[3] Breslow, N.E. (1989). Score tests in overdispersed GLMs, in *Workshop on Statistical Modeling*, A. Decarli, B.J. Francis, R. Gilchrist& G.U.H. Seeber, eds. Springer-Verlag, New York, pp. 64–74.

[4] Breslow, N.E. (1996). Generalized linear models: checking assumptions and strengthening conclusions, *Statistica Applicata* **8**, 23–41.

[5] Breslow, N.E.& Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**, 9–25.

[6] Breslow, N.E.& Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**, 81–91.

[7] Davidian, M.& Carroll, R.J. (1987). Variance function estimation. *Journal of the American Statistical Association* **82**, 1079–1081.

[8] Dean, C.B. (1991). Estimating equations for mixed Poisson models, in *Estimating Functions*, V.P. Godambe, ed. Oxford University Press, Oxford, pp. 35–46.

[9] Dean, C.B. (1992). Testing for overdispersion in Poisson and binomial regression models, *Journal of the American Statistical Association* **87**, 451–457.

[10] Dean, C.& Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association* **84**, 467–472.

[11] Diggle, P., Liang, K.Y.& Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

[12] Efron, B.& Tibshirani R.J. (1993). An Introduction to the Bootstrap. Chapman& Hall, London.

[13] Fisher, R.A. (1950). The significance of deviations from expectation in a Poisson series, *Biometrics* **6**, 17–24.

[14] Godambe, V.P.& Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with discussion), *Journal of Statistical Planning and Inference* **22**, 137–152.

[15] Lee, Y.& Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series B* **58**, 619–678.

[16] Lesperance, M.L.& Kalbfleisch, J.D. (1992). An algorithm for computing the non-parametric MLE of a mixing distribution, *Journal of the American Statistical Association* **87**, 120–126.

[17] Liang, K.Y.& Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[18] Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Institute of Mathematical Statistics, Hayward.

[19] Lindsey, J.K. (1993). *Models for Repeated Measurements*. Oxford University Press, New York.

[20] McCullagh, P.& Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman& Hall, London.

[21] O'Hara Hines, R.J., Lawless, J.F.& Carter, E.M. (1992). Diagnostics for a cumulative multinomial generalized linear model, with applications to grouped toxicological mortality data, *Journal of the American Statistical Association* **87**, 1059–1069.

[22] Pregibon, D. (1980). Goodness-of-link tests for generalized linear models, *Applied Statistics* **29**, 15–24.

[23] Prentice, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, *Journal of the American Statistical Association* **81**, 321–327.

[24] SAS Institute Inc. (1990). *SAS Release 6*.03 edition. SAS Institute Inc., Cary, NC.

[25] Smith, P.J.& Heitjan, D.F. (1993). Testing and adjusting for departures from nominal dispersion in generalized linear models, *Applied Statistics* **42**, 31–41.

[26] Statistical Sciences (1995). *S-PLUS*, Version 3.3, StatSci, a division of Math-Soft, Inc., Seattle.

[27] "Student" (1919). An explanation of deviations from Poisson's law in practice, *Biometrika* **12**, 211–215.

[28] Tarone, R.E. (1979). Testing the goodness-of-fit of the binomial distribution, *Biometrika* **66**, 585–590.

[29] Thall, P.F.& Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**, 657–671.

(*See also* **Binary Data**; **Correlated Binary Data**; **Logistic Regression**; **Loglinear Model**)

CHARMAINE B. DEAN

Overmatching

Overmatching refers to the unnecessary or inappropriate use of **matching** in a **cohort** or **case–control study**. Matching on intermediate factors on the causal pathway (*see* **Causation**) can inappropriately attenuate estimates of exposure effect, and matching on factors that are not **confounders** can needlessly reduce the **power** of the study [1, pp. 104–106]. Overmatching is also sometimes used to describe elaborate matching schemes that make it difficult to find suitable **controls** satisfying all matching criteria (*see* **Matched Analysis**).

Reference

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1. The Analysis of Case-Control Studies. International Agency for Research on Cancer, Lyon.

MITCHELL H. GAIL

P Value

The *P* value is probably the most ubiquitous statistical index found in the applied sciences literature and is particularly widely used in biomedical research. It is also fair to state that the misunderstanding and misuse of this index is equally widespread. A complete discussion of all issues relevant to the *P* value could touch on many of the subtlest and most difficult areas in statistical **inference**. In this article we will focus on those aspects most relevant to the interpretation of results arising from biomedical research.

The *P* value, as it is used and defined today, was first proposed as part of a quasi-formal method of inference by **R.A. Fisher**, and popularized in his highly influential 1925 book, *Statistical Methods for Research Workers* [13]. It is defined as the probability, under a given simple hypothesis H_0 (the **null hypothesis**), of a statistic of the observed data, plus the probability of more extreme values of the statistic (*see* **Hypothesis Testing**). Other names that have been used for the *P* value include "tail-area probability," "associated probability," and "significance probability" [16]. It can be written as

$$P \text{ value} = \Pr(\mathrm{t}(\mathbf{X}) \geq \mathrm{t}(\mathbf{x})|\mathrm{H}_0),$$

where $\mathbf{X}$ is a random vector, $\mathbf{x}$ is a vector of observed results, and $\mathrm{t}(\mathbf{X})$ is a function of the data, known as a *test statistic* (e.g. the sample average).

This seemingly simple mathematical definition belies the complexity of its interpretation, and even the occasional difficulty of its calculation in real-world settings. The seeds of this difficulty are embodied in its definition, which is partly conditional (depending on observed data for calculation), and partly unconditional (calculated over a set of unobserved outcomes defined by the study design) [26]. It is therefore an index that seems to measure simultaneously an "error rate" (pre-experiment, unconditional perspective) and the strength of inferential evidence (post-experiment, conditional perspective). Fisher's original purpose for the *P* value was in the latter category, as an index that denoted the statistical compatibility between observed results and a hypothetical distribution. He meant it to be used informally as a measure of statistical evidence. The larger the statistical distance (and the smaller the *P* value), the greater was the evidence against the null hypothesis.

While the basic idea had undeniable appeal, this definition posed several problems, some logical, some practical, which Fisher never fully resolved. They included the following:

1. How a measure of distance from a single hypothesis could be interpreted as a measure of evidence without consideration of other hypotheses [6].
2. How the probability of unobserved "more extreme" results were relevant to the evidential meaning of the observed result [19, 21].
3. How to calculate a *P* value when the experimental design (e.g. **sequential**) or execution (unanticipated events) rendered the sample space uncertain [8, 11].
4. How to choose an appropriate test statistic [9, 10, 20].
5. How the numerical magnitude of the *P* value should be interpreted operationally [2, 3, 12].

These questions were difficult to address because the *P* value was not part of any formal system of inference. This contrasted with the **likelihood ratio**, which was a part of **Bayes' theorem**. But while Fisher also developed the idea of mathematical **likelihood**, he eschewed Bayes' theorem as a useful tool in scientific inference. In its stead, he developed a panoply of methods which were meant to be tools in what he regarded as the fluid and nonquantifiable process of scientific reasoning. He offered various "rules of thumb" for the use of these tools. Among such rules were the suggestion that if the *P* value were less than 0.05, one might start doubting the null hypothesis. He was clear that the response to such a finding would generally be to conduct another experiment, or gather more data.

Had this been the full extent of the *P* value's theoretical foundation, it is doubtful that it would occupy as central a role as it does today. It is ironic that it became popularized, and indeed reified, because of the development of another method that did not include it, and which explicitly denied that conditional inference could be part of a "scientific" method [18].

The *P* Value as an "Observed Error Rate"

In 1933, **J. Neyman** and **E.S. Pearson** (N–P) proposed the "hypothesis test", which involved

"rejecting" or "accepting" null or **alternative hypotheses** with predetermined frequencies when they were true [24]. The probabilities of making the wrong decision were called error rates, and designated into two classes: α, (or type I) error (the probability of rejecting the null hypothesis when it was true) and β (type II) error (the probability of accepting the null hypothesis when the alternative was true) (*see* **Level of a Test**). A critical value for a test statistic is determined (via a likelihood ratio), and the null hypothesis is rejected if the statistic exceeds that critical value, and is accepted if not (*see* **Critical Region**). This methodology borrowed several of Fisher's ideas, most notably that of mathematical likelihood, and, with its introduction of the concept of an alternative hypothesis and associated **power**, appeared to address some of the logical problems posed by Fisher's less formal system. But N–P explicitly rejected P values, because for the "error rates" to have meaning (and for a method to be "scientific"), an hypothesis had to be rejected or accepted; a result could not reflect back on underlying hypotheses in a graded way, which use of the P value implied.

The N–P method provided the formal framework of statistical inference (or at least decision-making) that the P value lacked, but it was a framework that encouraged the misinterpretation of P values. The juxtaposition of the two approaches has been the source of considerable confusion ever since. Since both P values and type I error rates were tail-area probabilities under the null hypothesis, the P value came to be interpreted as an "observed type I error", and defined in most applied textbooks that way. While it is mathematically true that the P value is smallest alpha level at which one could still justify the rejection of the null hypothesis, this fact does make not make the P value an error rate. (It should be stressed that the confusion we are discussing here is not the more common lay misperception that P value represents the probability of the null hypothesis.) The outcome cannot fall within the region over which the P value is calculated; by definition, the observed outcome is always on the border of that region, and is usually the most probable. In an applied setting, this confusion is manifested in the inability to quantitatively distinguish between the very different inferential implications of a result reported as "$P \leq 0.05$" vs. "$P = 0.05$". It is fascinating how few users of P values recognize the profound difference in inferential weight introduced by that subtle change in inequality sign, whereas many agonize over the far smaller difference between $P = 0.04$ and $P = 0.07$.

The identification of the P value as a form of *post hoc* type I error rate, or as an "improved" estimate of that error, created a powerful illusion that is undoubtedly the source of its appeal; that a deductively derived error rate and inductive measure of inferential strength were identical, and that the "objective" qualities of the first could be bestowed upon the latter.

Fisher and Neyman fought vigorously in print over which approach was preferable, and Fisher in particular expressed profound dismay at seeing his "significance probability" absorbed into hypothesis testing ("acceptance procedures", in his words) [14, 23]. But textbook and **software** writers, either not wanting to confuse readers, or perhaps themselves not being clear about the issues, obscured the distinction, and technology triumphed over philosophy [17].

The *P* Value as a Measure of Evidence

We have seen above how the conditionality of the P value makes it inappropriate to view as a post-test α level. We will see that its unconditional characteristics render it problematic as an evidential measure as well. Fisher's proposal that the P value be used as a measure of evidence appeared based on the intuitively appealing idea that the more unlikely an event was under the null hypothesis, the more "evidence" that event provided against the hypothesis. The tail area appeared to be a convenient way in which to index the statistical distance between the null hypothesis and the data. Fisher himself did not seem wedded to that measure, simultaneously endorsing the use of the mathematical likelihood as an alternative.

If the P value were used informally as Fisher originally intended, it could be viewed as equivalent to any other functional transformations of the distance between the observed statistic and its null value, like a Z-score, or standardized likelihood. But its use in any formal way poses several difficulties. Because the P value is calculated only with respect to one hypothesis, and has no information, by itself, of the magnitude of the observed effect (or equivalently, of power), it implicitly excludes the magnitude of effect from the definition of "evidence". Small deviations from the null hypothesis in large experiments can have the same P value as large deviations in small experiments. The likelihood functions for these two

results are quite different, as are any data-independent summaries of the likelihoods. This difference is also reflected in the perspective of most scientists, who would typically draw different conclusions from such a pair of results.

The corrective in the biomedical literature has been to urge the reporting of P values together with estimates of effect size, and of the precision of the measured effect, usually reported as a **confidence interval** [1, 15, 25]. This does not completely solve the problem of representing the evidence with the same number when the data appear quite different, but it at least gives scientists additional information upon which to base conclusions.

The dependence on only one hypothesis means that data with different inferential meaning can have the same P value [4]. A converse problem occurs when the same data can be represented by two different P values. This problem is created by the inclusion of "unobserved" outcomes in the tail area calculation of the P value. Experiments of different design can have different "unobserved" outcomes even if the observed data are identical. The classic example involves the contrast between a fixed sample size experiment and one in which a data-dependent stopping-rule is used. Suppose that two treatments, A and B, are applied to each subject in a clinical trial, and the preferred treatment for that subject recorded. The sequence of preferred treatments are five A s followed by one B. If this was planned as an experiment with $n = 6$, then the one-sided P value based on the **binomial distribution** is

$$\binom{6}{1}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^1 + \frac{1}{2}^6 = 0.11.$$

If the experiment had been designed to stop when the first B success was observed, then the more extreme results would consist of longer sequences of As, and the P value based on the **negative binomial** would be

$$\frac{1}{2}\left(\frac{1}{2}\right)^5 + \frac{1}{2}\left(\frac{1}{2}\right)^6 + \cdots = 0.031.$$

While this is an idealized situation, this issue is manifest in the real-world applications by the discussions of how to measure appropriately the evidence provided by biomedical experiments that have stopped unexpectedly or because of large observed differences. In one trial of medical therapy known by its acronym of "ECMO", a variety of P values could be derived from the data presented [22].

Perhaps the most illuminating examinations of the inferential meaning of P values have used **Bayesian** or likelihood approaches. Bayesian analyses show that the P value is approximately the Bayesian posterior probability, under a diffuse **prior**, of an effect being in the direction opposite to the one observed, relative to the null hypothesis. More generally, the one-sided P value is the lower bound on that probability for all unimodal symmetric priors centered on the null [7].

Both likelihood and Bayesian arguments show that the P value substantially overstates the evidence against a simple, "sharp" null hypothesis, particularly for P values above 0.01. In the **normal** (Gaussian) case, the minimum likelihood ratio for the null, which is the minimum Bayes factor as well, is substantially higher than the associated P value, as shown in Table 1 [5, 12].

Since the P value is usually calculated relative to a sharp null hypothesis, most users interpret its magnitude as reflecting on the null hypothesis. The standardized likelihood $[\exp(-Z^2/2)]$ is the smallest Bayes factor that can multiply the prior odds of the null hypothesis to calculate the posterior odds. We see that the odds are not changed nearly as much as the P value's magnitude would suggest; nor is the probability. In the range of very small P values ($P < 0.001$) the quantitative differences are not likely to be important in practice. But in the range which includes many P values reported in biomedical research – that is, $0.01 < P < 0.10$ – the differences between most users' impression of the P value's meaning and its

Table 1 The relationship between the observed Z-score, (**standard normal deviate**), the fixed-sample size two-sided P value, the Gaussian standardized likelihood $\exp(-Z^2/2)$, and the smallest possible Bayesian posterior probability when the prior probability on the simple null hypothesis is 0.5. The latter is calculated using Bayes' theorem, and equals stand.lik./(1 + stand.lik.)

Z-score	P value (two-sided)	Gaussian standardized likelihood	Min. $\Pr(H_0\|Z)$ when $\Pr(H_0) = 0.5$
1.64	0.10	0.26	0.21
1.96	0.05	0.15	0.13
2.17	0.03	0.09	0.08
2.58	0.01	0.036	0.035
3.29	0.001	0.0045	0.0045

maximum inferential weight is striking. When one-sided P values are used, the contrast is even more marked.

One-sided vs. Two-sided *P* Values

Much has been written about how one could choose whether to cite a one- or two-sided P value. This is somewhat academic, since the *de facto* standard in the biomedical literature is for two-sided tests. Some writers have stressed that this is little more than a semantic distinction – that is, about what label should be attached to a $Z = 1.96$ – and that if the result is completely reported, a reader could make the appropriate adjustment if they judge it appropriate. It is interesting to note that this distinction becomes more than academic if one dichotomizes results into "significant" and "not significant". Then, whether the test was one- or two-sided is an important factor in assessing the meaning of the verdict. However, this assumes that not even the direction of the result is reported, which is unlikely.

This issue is similar to that confronting the experimenter who stops a trial before its planned end. In both situations, an experimenter's intentions or thoughts are taken as relevant to the strength of evidence, such that two persons with the same data could quote different P values. There is no clear resolution to this, since those who focus on the unconditional aspects of the P value will see such considerations as important, and those who regard it primarily as a conditional evidential measure will insist that such considerations should be irrelevant. In general, the custom that most or all P values be reported as two-sided seems a good one, with the condition that if one-sided P values are used, this be indicated clearly enough so their value could be doubled by a reader. If the P values are in a range in which doubling makes a substantive difference, the evidence is equivocal enough so there will be controversy regardless of how the P value is reported.

Conclusions

The P value represented an attempt by R.A. Fisher to provide a measure of evidence against the null hypothesis. Its peculiar combination of conditionality and unconditionality, combined with its absorption into the hypothesis testing procedure of Neyman and Pearson, have led to its misinterpretation, widespread use, and seeming imperviousness to numerous criticisms. Since P values are not likely to soon disappear from the pages of medical journals or from the toolbox of statisticians, the challenge remains how to use them and still properly convey the strength of evidence provided by research data, and how to decide whether issues of design or analysis should be incorporated into their calculation.

References

[1] Altman, D.G. & Gardner, M. (1992). Confidence intervals for research findings, *British Journal for Obstetrics and Gynaecology* **99**, 90–91.

[2] Barnard, G. (1966). The use of the likelihood function in statistical practice, in *Proceedings of the Fifth Berkeley Symposium*, Vol. 1. University of California Press, Berkeley, pp. 27–40.

[3] Berger, J. (1986). Are P-values reasonable measures of accuracy?, in *Pacific Statistics Congress*, I. Francis, B. Manly & F. Lam eds. North-Holland, Amsterdam.

[4] Berger, J.O. & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity, *American Scientist* **76**, 159–165.

[5] Berger, J. & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P-values and evidence, *Journal of the American Statistical Association* **82**, 112–139.

[6] Berkson, J. (1942). Tests of significance considered as evidence, *Journal of the American Statistical Association* **37**, 325–335.

[7] Casella, G. & Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem, *Journal of the American Statistical Association* **82**, 106–111.

[8] Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle, *American Statistician* **20**, 18–23.

[9] Cox, D. (1977). The role of significance tests, *Scandinavian Journal of Statistics* **4**, 49–70.

[10] Cox, D. & Hinckley, D. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[11] Dupont, W. (1983). Sequential stopping rules and sequentially adjusted P values: does one require the other (with discussion)?, *Controlled Clinical Trials* **4**, 3–10.

[12] Edwards, W., Lindman, H. & Savage, L. (1963). Bayesian statistical inference for psychological research, *Psychological Review* **70**, 193–242.

[13] Fisher, R. (1973). *Statistical Methods for Research Workers*, 14th Ed. Hafner, New York. Reprinted by Oxford University Press, Oxford, 1990.

[14] Fisher, R. (1973). *Statistical Methods and Scientific Inference*, 3rd Ed. Macmillan, New York.

[15] Gardner, M. & Altman, D. (1986). Confidence intervals rather than *P* values: estimation rather than hypothesis testing, *Statistics in Medicine* **292**, 746–750.
[16] Gibbons, J. & Pratt, J. (1975). *P*-values: interpretation and methodology, *American Statistician* **29**, 20–25.
[17] Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Kruger, L. (1989). *The Empire of Chance*. Cambridge University Press, Cambridge.
[18] Goodman, S.N. (1993). *P*-values, hypothesis tests and likelihood: implications for epidemiology of a neglected historical debate (with commentary and response), *American Journal of Epidemiology* **137**, 485–496.
[19] Hacking, I. (1965). *The Logic of Statistical Inference*. Cambridge University Press, Cambridge.
[20] Howson, C. & Urbach, P. (1989). *Scientific Reasoning: the Bayesian Approach*. Open Court, La Salle.
[21] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Oxford University Press, Oxford.
[22] Lin, D.Y. & Wei, L.J. (1989). Comments on "Investigating therapies of potentially great benefit: ECMO", *Statistical Science* **4**, 324–325.
[23] Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd Ed. The Graduate School, US. Department of Agriculture, Washington.
[24] Neyman, J. & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypothesis, *Philosophical Transactions of the Royal Society Series A* **231**, 289–337.
[25] Rothman, K. (1978). A show of confidence, *New England Journal of Medicine* **299**, 12–13.
[26] Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference*. Reidel, Dordrecht.

STEVEN N. GOODMAN

Pain

There are many situations in clinical medicine when it is necessary to assess how much pain a patient is feeling. This may be either to ascertain how painful a clinical procedure is, or to compare the pain relief, or pain inflicted, by two or more modes of treatment. In the medical literature, there is a fairly clear distinction between acute pain and chronic pain.

Types of Pain

Chronic Pain

There are many ailments where patients suffering from long-term illnesses are assessed for how much pain they are experiencing. In particular, patients with rheumatoid arthritis, back pain, and cancer often fall into this category. Owing to the wide range of illnesses, such studies appear in many specialist areas of the medical literature. A typical example is a study of chronic low back pain by Marchand et al. [10].

Acute Pain

There are many examples where patients undergoing a medical procedure are asked postoperatively about the level of pain they are suffering. Such studies include patients having hip or knee replacements, hysterectomies, and wisdom tooth removal. While they cover many different types of operation, these studies are particularly common in the anesthesia literature. A typical example is a study by Hommeril et al. [6] exploring pain relief after hip and knee arthroplasty.

Experimental Pain

A further category is that of experimental pain. An example is where volunteers immerse their hands in ice or are subjected to small electric shocks. These experiments usually yield straightforward data that can be analyzed using conventional statistical methods. A good example of such studies is that by Bjorkman & Elam [1].

Historical Development

To study pain it is first necessary to measure it reliably. Unfortunately, it is not possible (yet) to attach an instrument to a patient to obtain a direct objective measure of how much their pain hurts. There are two common methods of getting round this problem.

The 10 cm Visual Analog Scale (VAS)

This is the most common method for measuring pain in the medical literature. It is used in all three types of pain study described earlier. A standard 10 cm scale is illustrated in Figure 1. The patient is asked to mark on the line how much his or her pain hurts. This is a very important component of the literature and is addressed in detail in later sections.

Questionnaires

There are several pain questionnaires. The one which seems to be the most used and respected is the McGill Pain Questionnaire published by Melzack [12] in 1975. However, more recently, Thomas [16] published the Glasgow Pain Questionnaire. Neither of these is disease specific. This is an important point because there is an expanding literature of pain scores developed to assess the pain and disability associated with particular conditions. Two good examples are the Aberdeen Back Pain [14] score and Psoriasis Disability Index [5].

Measurement of **quality of life** is an expanding area of the medical literature. The most commonly used questionnaires are the Nottingham Health Profile [7] and the SF-36 [3]. Both have particular domains dedicated to pain. Methodology for calculating sample sizes for the SF-36 was described by Julious [8]. In order to study a particular disease it is important to note that if either of these is to be used, then a disease-specific score such as those mentioned above should also be used. This is because they are both general health questionnaires.

Two other approaches are worth mentioning. Rather than a VAS score, the patient is asked to fill in a four- or five-point categorical scale ranging from no/little pain to very bad/unbearable.

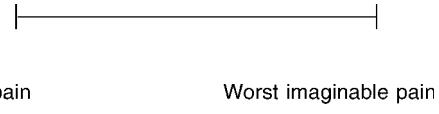

Figure 1 Standard 10 cm scale

This is most useful when the aim is to obtain a single pain assessment and the subsequent analysis is straightforward. The other method is to use the amount of analgesia consumed by a patient as a surrogate for pain. An example of this sort is discussed later.

Different Types of Study

Chronic Pain

The usual aims for studying chronic pain are (i) to ascertain the "average" amount of a pain a patient suffers in daily life and (ii) to intervene with a treatment that will hopefully reduce the suffering. Any of the pain measures discussed above can be used, although the questionnaires are particularly popular. Usually, data are collected at baseline and at a fixed point in time after treatment. These studies do not present with any particular statistical problems and can be analyzed using conventional techniques.

Acute Pain

The questionnaire-based measures of pain that are appropriate to chronic pain are clearly not appropriate for measuring acute pain. There are two reasons for this. First, in acute pain studies the patients are often in great distress. Secondly, some of the questionnaires take several minutes to complete, which is too onerous for the patient and not suitable if the experimenter wishes to assess pain levels at regular intervals. The most common form of measurement of acute pain is the 10 cm VAS. The properties of the VAS have been explored by many authors. A good source of references can be found in [13].

A common study design is to allocate randomly patients recovering from an operation under general anesthetic to receive pain relief from different analgesic agents (*see* **Randomization**). Patients are then asked to fill in a VAS at regular intervals. These intervals are not always of equal length, and such studies vary from 2 h to 48 h in duration. The most common form of analysis in the medical literature is to use either *t* tests (*see* **Student's *t* Statistics**) or **Wilcoxon–Mann–Whitney tests** to compare **mean** pain levels at each time point between groups.

A typical example by Doyle et al. [4] is worth describing. Forty children undergoing appendicectomy were allocated randomly to receive one of two bolus doses of morphine upon waking from the operation. Each group then used the same patient-controlled analgesia (PCA) regime to control their own pain. An observer filled in a VAS every 4 h for 48 h for each child at rest and during movement. The data were then analyzed using a Wilcoxon–Mann–Whitney test at each time point. The slight difference with this study is that the patients did not fill in the VAS but an observer did.

A study by Hommeril et al. [6] is also worth illustrating. In a double-blind (*see* **Blinding or Masking**) randomized study 32 subjects were allocated to receive either ketoprofen or morphine after undergoing hip and knee arthroplasty. The ketoprofen regime was a continuous infusion over 13 h, whereas the morphine regime was a single large bolus dose. The patients filled in a VAS upon recovery and 1, 3, 5, 7, 9, and 13 h after recovery. Patients suffering from extreme pain were allowed extra help in the form of an intravenous paracetamol injection. In total, nine of these injections were administered. The data were analyzed using a Mann–Whitney test at each time point. No attention was given to the fact that the study had in effect changed due to the extra "treatment" received by some of the patients.

Experimental Pain

In these studies the pain is usually very short lived and can be reproduced by an identical stimulus. Hence **crossover designs** are quite popular. The use of crossover designs in acute pain studies is not possible. There are examples of crossover designs for chronic pain studies in the medical literature, although the washout period must be such that the effect of the drug administered in the first experimental period has time to wear off.

Landmark Studies

Anybody wishing to learn about pain should consult the volumes edited by Melzack & Ward [13] which cover the entire subject from a medical viewpoint. One paper well worth a mention is that by Matthews et al. [11]. Here, they demonstrate in a major medical journal the flaws of repeated significance testing at successive time points (*see* **Sequential Analysis**). Kelly reviewed 12 journals from the anesthesia literature from 1991 and found that 59 out of 71 papers used this flawed method as the basis of

their analysis. The logic of comparing bolus dosages of analgesia is criticized in a *British Journal of Anaesthesia* editorial [9]. However, despite this, the practice is still very common.

Particular Statistical Concepts, Problems, and Techniques

The flaws in making significant tests at repeated time points are described by Matthews et al. [11] as:

1. A single curve joining the mean values at each point may hide important variation between patients within the same group.
2. The analysis does not take account of the fact that measurements at different time points are from the same patients.
3. Successive observations on a given patient are likely to be **correlated**. Note that this is particularly true for pain data.
4. Dividing the results into "significant" or "non-significant" introduces an artificial dichotomy into serial data (*see* **Hypothesis Testing**).

Statistically valid alternatives used in the medical literature are either to use repeated measures **analysis of variance** or to use summary measures as suggested by Matthews et al. [11]. The latter route is certainly the easier to interpret and in my view is more appropriate for most medical applications. Problems with both approaches include **missing data**. A patient may be in a deep sleep or decide to drop out of the study at any point. In most pain studies it is necessary for ethical reasons to allow the patient to take additional analgesia if they request it. This use of so-called escape analgesia is often ignored when it comes to data analysis. However, such data clearly **confound** the intended comparison.

Solutions

The key to using a summary measure approach is to ensure that the chosen measure is clinically meaningful. For this reason, summaries such as time to taking first escape analgesia, rate of pain reduction over a fixed time period, minimum pain, or the number of unacceptable pain observations in a fixed time period are all appealing in various circumstances. Here are three examples that illustrate the use of summary measures.

Bray et al. [2] compared the efficacy of morphine infusion against patient-controlled analgesia in a double-blind randomized study on 30 children undergoing major abdominal surgery. The time period of interest was the first 4 h after the child recovered from the general anesthesia from the operation. Prior to commencing the study, the authors decided to ask the children to fill in a 10 cm VAS every half hour for 4 h once awake after the operation. The outcome of interest to the clinician was how long each child's pain is at an unacceptable level. Hence, a VAS score of 50 mm or more was defined as unacceptable. The authors simply counted how many of the pain scores were unacceptable and compared them between groups using a Mann–Whitney test. This simple summary encapsulated precisely the aim of the clinicians in performing the study and provided very clear and easily interpretable results which they readily accepted.

Seymour et al. [15] explored the efficacy of ketoprofen and paracetamol for pain relief in patients recovering from the extraction under general anesthesia of their third wisdom tooth. A total of 206 patients entered the study into one of five groups, a placebo and two dosages of each drug. Patients filled in a 10 cm VAS at 15, 30, 45, 60, 90, 120, 180, 240, 300, and 360 min after administration of the analgesia. The investigators were interested in: (i) How quickly does each treatment act? (ii) For how long does the treatment have an effect? (iii) What is the maximum pain relief provided from the treatment? For each of these a summary measure was used. The speed with which the drug acted was measured by simply looking at the change in pain score over the first hour. The effective length of the treatment was assessed by observing the times at which patients made requests for escape analgesia, and the resulting information was analyzed using the **logrank test**. The maximum pain relief was calculated for each patient as the difference between their baseline pain and their lowest VAS score. A global summary in the form of the area under the curve (*see* **Bioequivalence**) of the pain profile of each patient was also calculated. The interpretation of such a quantity is not clear and is complicated by patients taking escape analgesia. This is a popular measure with pharmaceutical companies but is not in my view as useful as carefully thought out summaries such as those illustrated which clearly provide useful information for the researchers' questions.

The use of the cumulative amount of analgesia consumed by a patient as a measure of pain was mentioned earlier. This is appealing; however, a warning to be wary of such studies is necessary. Welchew & Breen [17] compared the pain relief received between patient-controlled on-demand fentanyl delivered epidurally or intravenously. Here, it is the method of delivery that is being tested rather than the analgesia. The subjects were 20 patients undergoing upper abdominal surgery. They used two methods of analysis. Each patient filled in a 10 cm VAS every hour for 24 hours after recovering from the operation. The Mann–Whitney test was used to compare the means at each time point. They also plotted the mean cumulative doses of drug consumed by patients in each group and compared these at each time point with an unpaired t test. Apart from the obvious statistical error, they forgot that, because of the safety lockouts on the patient-controlled analgesia equipment, patients in each group would automatically receive differing amounts of drug.

Anticipated Developments and Unresolved Problems

There is great potential for statistics to make a major and direct impact upon research into pain and pain relief. The use of summary measures should be explained and illustrated with the appropriate medical literature. There seems to be a characteristic shape to a pain curve which could be modeled. The parameters of such a curve could then be compared between different treatment groups, giving a comparison of the pain profiles rather than many individual time points. Methods for dealing with missing data and for patients taking additional analgesia should be developed.

Increasingly, chronic pain is been assessed by qualitative methods. While these do not readily lend themselves to statistical analysis, and indeed are sometimes directly opposed to classical statistical concepts of randomization and **outcome variables**, there is some merit in their aim to provide better quality information on pain.

References

[1] Bjorkman, R. & Elam, M. (1993). Diclofenac evaluation in a human experimental model of central pain, *Pain* **54**, 197–202.

[2] Bray, R.J., Woodhams, A.M. & Vallis, C.J. (1996). A double-blind comparison of morphine infusion and patient-controlled analgesia in children, *Paediatric Anaesthesia* **6**, 121–127.

[3] Brazier, J.E., Harper, R. & Jones, N.M.B. (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care, *British Medical Journal* **305**, 160–164.

[4] Doyle, E., Mottart, K.J. & Marshall, C. (1994). Comparison of different bolus doses of morphine for patient-controlled analgesia in children, *British Journal of Anaesthesia* **72**, 160–163.

[5] Finlay, A.Y. & Kelly, S.E. (1987). Psoriasis – an index of disability, *Clinical Experimental Dermatology* **12**, 8–11.

[6] Hommeril, J.L., Bernard, J.M. & Gouin, F. (1994). Ketoprofen for pain after hip and knee arthroplasty, *British Journal of Anaesthesia* **72**, 383–387.

[7] Hunt, S., McEwan, P. & McKenna, S.P. (1985). Measuring health status: a new tool for clinicians and epidemiologists, *Journal of the Royal College of General Practitioners* **35**, 185–188.

[8] Julious, S.A., George, S. & Campbell, M.J. (1995). Sample size for studies using the short form 36 (SF-36), *Journal of Epidemiology and Community Health* **49**, 642–644.

[9] Kehlet, H. (1994). Postoperative pain relief – what is the issue? *British Journal of Anaesthesia* **72**, 375–378.

[10] Marchand, S., Charest, J. & Li, J. (1993). Is TENS purely a placebo effect? A controlled study on chronic low back pain, *Pain* **54**, 99–106.

[11] Matthews, J.N.S., Altman, D.G. & Campbell, M.J. (1990). Analysis of serial measurements in medical research, *British Medical Journal* **300**, 230–235.

[12] Melzack, R. (1975). The McGill pain questionnaire: major properties and scoring methods, *Pain* **1**, 277–299.

[13] Melzack, R. & Ward, P.D. (1994). *Textbook of Pain*, 3rd Ed. Churchill Livingstone, Edinburgh.

[14] Ruta, D.A., Garratt, A.M. & Wardlaw, D. (1994). Developing a valid and reliable measure of health outcome for patients with low-back-pain, *Spine* **19**, 1887–1896.

[15] Seymour, R.A., Kelly, P.J. & Hawkesford, J.E. (1996). The efficacy of ketoprofen and paracetamol (acetaminophen) in postoperative pain after 3rd molar surgery, *British Journal of Clinical Pharmacology* **41**, 581–585.

[16] Thomas, R.J., McEwan, P. & Asbury, A.J. (1996). The Glasgow pain questionnaire: a new generic measure of pain; development and testing, *International Journal of Epidemiology* **25**, 1060–1067.

PETER J. KELLY

Paired Comparisons

The term *paired comparisons* is used in two rather different senses. This article is not concerned with the comparison of two treatments by the use of **matched pairs** (*see* **Paired *t* Test**), but focuses on the *method of paired comparisons*. In this method, t "treatments" are compared pairwise in situations in which full measurement of the responses is not feasible or is inappropriate [7].

A good example is provided by preference testing where, on the basis of the $\frac{1}{2}t(t-1)$ preferences expressed by each of n respondents, one wishes to test whether there are significant differences among the treatments (*see* **Hypothesis Testing**), and, if so, to rank the treatments or, better still, place them on some preference scale. In each paired comparison involved, the preferred treatment receives a score of 1 and the other a score of 0. Subsequent analysis is based on the (total) *scores*, $a_1, \ldots, a_t$, attained by the treatments at the end of the experiment. This requires in all $N = \frac{1}{2}nt(t-1)$ paired comparisons, which are assumed to be stochastically independent.

It will be seen that the method is not confined to preference testing, the essential feature being the allocation of 1s and 0s for each paired comparison. "Treatments" is a covering term that can also denote "objects", "individuals", and so on.

An interesting application is given by Jin et al. [5] in their study of whether there is segregation distortion in the human leukocyte antigen complex. For a parent of genotype ij, $i \neq j$, at a particular locus ($i = 1, \ldots, t$; $j = 1, \ldots, t$), a score of 1 is given to whichever allele, i or j, is transmitted to an offspring. Here, block size two (pairing) is forced on the experimenter, as it is for matching body parts or in round robin tournaments.

Sometimes it is possible for a respondent to rank all t treatments directly. If this is easily done, it may be an appropriate procedure. However, the method of paired comparisons allows inconsistencies to show up. Thus, if $A_1 \rightarrow A_2$ signifies that treatment A_1 scores 1, it may happen that $A_1 \rightarrow A_2$ and $A_2 \rightarrow A_3$ but $A_3 \rightarrow A_1$. Such *circular triads* [6] are an indication that the respondent finds it difficult to be consistent. This may be due to small perceived differences among treatments, or to the fact that the *merit* or *worth* of the treatments cannot be represented on a linear scale. If such a representation is possible, we say that a *linear* (paired comparison) *model* holds.

It can be shown [4, 6] that small values of the total number of circular triads are equivalent to large values of the sum of squares of the scores. Clearly, either of these would lead us to suspect the null hypothesis H_0 of no treatment differences. A simple approximate test may be made by rejecting H_0 if

$$\frac{4}{nt}\sum_{i=1}^{t}(a_i - \overline{a})^2 > \chi^2_{1-\alpha}(t-1), \qquad (1)$$

where $\overline{a} = \frac{1}{2}n(t-1)$ and $\chi^2_{1-\alpha}(t-1)$ is the upper α significance point of the **chi-square distribution** with $t-1$ **degrees of freedom**. The approximation is good except when N is small, in which case exact tables of upper 5% and 1% points are available in [4]. **Multiple comparison** methods, based on the a_i, may be used to rank the treatments (see [4]).

Of course, it is not always possible to make all pairings equally often. The analysis of such unbalanced experiments, where A_i and A_j are compared n_{ij} times ($i = 1, \ldots, t$; $j = 1, \ldots, t$; $n_{ij} \geq 0$) is treated in [1].

The foregoing procedures are applicable generally, but the interpretation of the results is simpler when a linear model can be assumed. An important example is the **Bradley–Terry model** [3], for which

$$\Pr(A_i \rightarrow A_j) = \frac{\pi_i}{(\pi_i + \pi_j)}, \qquad i = 1, \ldots, t, \quad j = 1, \ldots, t, \qquad (2)$$

where $\pi_i (> 0)$ represents the merit of $A_i (i = 1, \ldots, t)$. To fix the scale of the π_i, it is usual to impose the constraint $\sum \pi_i = 1$. General theory may be used to test H_0: $\pi_i = 1/t (i = 1, \ldots, t)$ and to estimate the π_i. It can be shown that the resulting estimates are ranked exactly as the a_i. Moreover, the method of **maximum likelihood** also allows the immediate estimation of the π_i for unbalanced experiments.

Ties in individual paired comparisons produce a complication. The simplest way of handling them is to allot a score of $\frac{1}{2}$ to each of the two tied treatments, especially when the number of ties is relatively small. For a full discussion, see [4].

A review of the subject, with emphasis on the Bradley–Terry model, is given in [2]. The general account [4] also deals with issues such as (i) designs

for paired comparison experiments when not all possible pairings can be made, (ii) multivariate paired comparisons, and (iii) selection and ranking of treatments.

References

[1] Andrews, D.M. & David, H.A. (1990). Nonparametric analysis of unbalanced paired-comparison or ranked data, *Journal of the American Statistical Association* **85**, 1140–1146.

[2] Bradley, R.A. (1984). Paired comparisons: some basic procedures and examples, in *Handbook of Statistics*, Vol. 4, P.R. Krishnaiah & P.K. Sen, eds. North-Holland, Amsterdam, pp. 299–326.

[3] Bradley, R.A. & Terry, M.E. (1952). The rank analysis of incomplete block designs, I: the method of paired comparisons, *Biometrika* **39**, 324–345.

[4] David, H.A. (1988). *The Method of Paired Comparisons*, 2nd Ed. Oxford University Press, New York.

[5] Jin, K., Speed, T.P., Klitz, W. & Thomson, G. (1994). Testing for segregation distortion in the HLA complex, *Biometrics* **50**, 1189–1198.

[6] Kendall, M.G. & Babington Smith, B. (1940). On the method of paired comparisons, *Biometrika* **31**, 324–345.

[7] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 273–286.

HERBERT A. DAVID

Paired *t* Test

A basic concept in study design is that extraneous sources of variation should be controlled, so that the comparison is made among groups which are alike except for the intervention (*see* **Experimental Design**). Techniques which may be used include **stratification** and **analysis of covariance**. Data which are naturally paired form useful strata and often facilitate a more precise comparison than one in a set of unrelated subjects. This leads to the paired t test. The cost of pairing lies in the effort needed to determine the values of the **matching** variables (which is low when there are natural pairs such as litter mates or pre- and post-intervention measures), and the reduction of **degrees of freedom** (df) from $2(n-1)$ to $n-1$ (*see* **Student's *t* Statistics**). The loss of degrees of freedom is almost never a problem when n is greater than 15. The statistic is based on the difference of the members of the pair, so the variance of the difference is $2\sigma^2(1-\rho)$, where ρ is the **correlation** coefficient, and σ^2 is the variance of a single observation. If $\rho = 0$, then the variance is the same as that of the difference between unpaired observations ($2\sigma^2$). In this case, the loss is that of half of the degrees of freedom. When the pairing is effective, $\rho > 0$, and the difference will have a smaller variance than if the observations were unpaired. The tradeoff is almost always in favor of pairing.

The paired t test is used to compare mean differences when the observations have been obtained in pairs, and are thus dependent. Examples include observations made on weight before and after an intervention in a subject, serum cholesterol of two members of a family, and blood concentration of a toxin in litter mates. Subjects are often matched on characteristics such as age, gender, race, and study site (in multicenter studies). In each of these examples, the two observations are made on the same response and will be correlated. Accounting for this is necessary to analyze the study properly. For paired data, it is natural to form the differences, and perform the analysis on the differences. An interesting recent account is by Senn & Richardson [5] who discuss the study which **W.S. Gosset** (Student) used in his paper. Moses [3] gives a summary of the theoretical aspects of this test.

We assume that there are n pairs of observations, and that each pair is independent of the other pairs. Denote the paired observations as x_i and y_i, and their difference as d_i for $i = 1, \ldots, n$. The analysis assumes that the observations are normally distributed with means μ_x and μ_y. The random variable $D = X - Y$ is then normally distributed with mean $\mu_d = \mu_x - \mu_y$ and variance σ_d^2. The **null hypothesis** is $\mathrm{H}_0 : \mu_d = 0$. A Student t statistic can be computed as

$$t = \frac{\overline{d}}{(s_d/\sqrt{n})}.$$

This statistic is then compared to **Student's *t* distribution** with $n-1$ df. This analysis is equivalent to a **randomized blocks** analysis of variance in which the pairs correspond to blocks.

If the variances in X and Y are equal, the variance of D is $2\sigma^2(1-\rho)$, as noted above. If $\rho < 0$, the variance is increased. This would not usually be the case when the matching process is effective. If the variances are unequal, the variance of D is $\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$. The relative efficacy of a paired t-test is $1/(1-\rho)$ as compared with a two-independent-sample t test in a parallel comparison. For example, if $\rho = 1/3$, the relative efficacy is 1.5. This means that 100 paired observations would have the same power to detect differences as an unpaired study of 150 observations per group.

Example

We give here, in Table 1, data used by Student (Gosset) as cited in Senn & Richardson [5]. The data were on the amount of sleep gained under two soporic drugs.

The means given here agree with those of Senn & Richardson, but the standard deviations are slightly larger. The values they report were those given by Student, who used the divisor n in computing the variance. Using the unpaired t-test we obtain a $t = 1.86$ with 18 df and a ***P* value** of 0.0792. By pairing the observations, as we should, the result is $t = 1.58/(0.39/\sqrt{10}) = 4.06$ with 9 df and a P value of 0.0028. It is better to report the mean of the differences and its standard deviation rather than showing only the t statistic or the P value (or worse, NS for "not significant", or some number of asterisks!).

Table 1

Patient	Dextro	Laevo
1	0.7	1.9
2	−1.6	0.8
3	−0.2	1.1
4	−1.2	0.1
5	−0.1	−0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4
Mean	0.75	2.33
Standard deviation	1.79	2.00

Robustness

The **robustness** properties correspond to those of the one-sample t test. The effect of nonnormality is fairly small if n is at least 30, since the distribution of d will be close to normal in that case. If one difference (or a few) appear to be quite large (i.e. **outliers**) the results can be affected. Outliers can be considered a form of nonnormality. They affect the variance of the observations, and can also affect the **skewness** of the distributions. The P values reported from an analysis are often given as $P < 0.05$ or $P < 0.01$. Since the P value depends on the behavior of the distribution in its tails, nonnormality generally means that statements such as $P < 0.001$ are rarely accurate (the quoted P value for the example thus should be regarded as $P < 0.01$). Lack of independence among the pairs (which might arise if multiple members of a litter or a family were included in a study, i.e. clustering) can seriously affect the level of the test. If the correlation between any pair of differences is γ, the variance of the differences is $\sigma^2(1 + 2\gamma)$. Thus, the estimated variance is biased. If $\gamma > 0$, the denominator of the t statistic is too small, and the significance levels are incorrect. If the correlation holds only among certain pairs (e.g. independent clusters would lead to a block diagonal **covariance matrix**), the analysis is more complex, but the estimated variance is still biased. Lack of common variance in X and Y does not formally affect the analysis. However, unless the primary interest is in the difference between the observations, the lack of common variance indicates that X and Y do not have the same distribution, although they might have the same mean. If the variance differs over the pairs, heteroscedasticity concerns arise (*see* **Scedasticity**). Rosner [4] has suggested a **random effects** model which accounts for this. Missing values can create problems. Usually, only one member of the pair is missing. If the missing value is related to the mean value within the pair, the missingness is not random, and the t test is affected (*see* **Missing Data**). For further discussion of these points, see Miller [2] or Madansky [1].

Several alternatives exist to the paired t test. These are useful if the distribution is not normal and there is concern that this may affect the performance of the test. The **sign test** uses the number of positive (or negative) signs as a **binomial** variable with probability parameter 1/2 under H_0, and, for large samples, computes the **standard normal deviate**, z, to test H_0. The **asymptotic relative efficiency** (ARE) of this test is 0.637 when the differences are normal. The **signed-rank** test ranks the absolute values of the differences and sums the ranks corresponding to the positive (or negative) signs. The ARE of this test is 0.955. The **normal scores** test replaces the ranks of the differences by their expected values under normality and computes a t test on these. Its ARE is 1.0. For observations from nonnormal distributions, the efficiencies of the **nonparametric** procedures may be higher than indicated here and the t test can be very inefficient.

References

[1] Madansky, A. (1988). *Prescriptions for Working Statisticians*. Springer-Verlag, New York.

[2] Miller, R.G. (1985). *Beyond Anova*. Wiley, New York.

[3] Moses, L. (1985). Matched pairs t-tests, in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 289–203.

[4] Rosner, B. (1982). A generalization of the paired t-test, *Applied Statistics* **31**, 9–13.

[5] Senn, S. & Richardson, W. (1994). The first t-test, *Statistics in Medicine* **13**, 785–803.

HENRY HSU & PETER A. LACHENBRUCH

Pairwise Independence

Suppose that two variables are considered as potentially **explanatory** for a further variable, called the **response**, and that the dependence of the response on each of the variables taken alone and on both acting jointly is of main interest. For an appropriately chosen scale and measure of dependence, suppose furthermore that the effects of both variables turn out to be (essentially) additive (*see* **Additive Model**). This means that the effect of one of them on the response is (nearly) the same no matter at which level the other explanatory variable is fixed. Often, this is described as the absence of an **interaction** but the presence of two main effects.

An important role of pairwise **independence** of explanatory variables is then as follows: it is certain that no dependence reversal can occur for the (nearly) additive effect of one of the explanatory variables in comparison with the effect of this variable taken alone. To put it differently, if the explanatory variables are nearly independent, and have essentially additive effects on the response, then the overall effect of just one of them coincides at least qualitatively with the corresponding effects considered conditional given the other variable. A strong reversal of treatment success as related to variable B occurs instead in the 2^3 contingency table displayed in Table 1, since the explanatory variable pair B, C is highly dependent.

For both discrete and for continuous responses, further discussions of dependence reversal in spite of essentially additive main effects are to be found, for instance, in Snedecor & Cochran [5, p. 472], Good & Mittal [1], Wermuth [6,7], and Guo & Geng [2]. In a **contingency table** context early insights are due to Yule [8] and Simpson [4], (*see* **Simpson's Paradox**).

Mutual Dependence in Spite of Pairwise Independences

In general, no mutual independence results even if several variables are all pairwise independent. Instead, more complicated types of dependencies may still exist, which are often called higher-order interactions. An important implication is that methods of analysis relying completely on pairwise associations, **correlation**-based techniques or **correspondence analysis**, will overlook the existing dependencies in such situations and are therefore likely to lead to misleading interpretations.

An empirical example with four **binary** variables is due to Lienert [3]. He reported on symptoms after LSD intake. The 2^3 contingency table shown in Table 2 is an adaptation of his results. The three transient symptoms, recorded to be present (level 1) or absent (level 2), are distortions in affective behavior (A), distortions in thinking (B), and dimming of consciousness (C). There is a strong three-way interaction, as reflected for instance in the quite distinct **odds ratios** at the two levels of C; at the same time, the frequencies in the three marginal tables show all three symptom pairs as being close to independence.

With completely randomized designs (*see* **Randomization**) it will typically occur that – at the time a study starts – not only observed variables but also unobserved variables will essentially be both pairwise, and mutually, independent. Note, however, that even with this technique it is not possible to avoid dependencies with unobserved intermediate variables,

Table 1 Dependence reversal because of strongly associated explanatory variables

	C, site				Overall; that is, summed over sites	
	$k = 1$		$k = 2$			
A,	B, treatment		B, treatment		B, treatment	
outcome	$j = 1$	$j = 2$	$j = 1$	$j = 2$	$j = 1$	$j = 2$
$i = 1$ (success)	96 (96%)	600 (60%)	400 (40%)	4 (4%)	496 (45%)	604 (55%)
$i = 2$	4	400	600	96	604	496
Sum	100	1000	1000	100	1100	1100
Odds ratio	16		16		0.67	

Table 2 Symptoms after LSD intake: mutual dependence and pairwise independence

	C, dimming of consciousness			
	$k = 1$ (yes)		$k = 2$	
A, distorted affective behavior	B, distorted thinking		B, distorted thinking	
	$j = 1$, (yes)	$j = 2$	$j = 1$, (yes)	$j = 2$
$i = 1$ (yes)	21	5	4	16
$i = 2$	2	13	11	1
Odds ratio	27.30		0.023	

i.e. with unrecorded variables related to both treatment and outcome, but occurring unnoticed before observing outcome. Typical examples are noncompliance of some patients (*see* **Compliance Assessment in Clinical Trials**) or, more generally, unrecorded treatment effects or changes in measurement devices before treatment outcome is established.

References

[1] Good, I.J. & Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables, *Annals of Statistics* **15**, 694–711.

[2] Guo, J. & Geng, Z. (1995). Collapsibility of logistic regression coefficients, *Journal of the Royal Statistical Society, Series B* **57**, 263–267.

[3] Lienert, G.A. (1970). Konfigurationsfrequenzanalyse einiger Lysergsäurediäthylamid-Wirkungen, *Arzneimittelforschung* **20**, 912–913.

[4] Simpson, E.H. (1951). The interpretation of interactions in contingency tables, *Journal of the Royal Statistical Society, Series B* **13**, 238–241.

[5] Snedecor, G.W. & Cochran, W.G. (1967). *Statistical Methods*, 6th Ed. Iowa State University Press, Ames.

[6] Wermuth, N. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable, *Journal of the Royal Statistical Society, Series B* **49**, 353–364.

[7] Wermuth, N. (1993). Association structures with few variables: characteristics and examples, in *Population Health Research*, K. Dean, ed. Sage, London, pp. 181–203.

[8] Yule, G.U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* **2**, 121–134.

(*See also* **Statistical Dependence and Independence**)

NANNY WERMUTH

Panel Study

In a panel study a number of individuals are followed for a given period of time. At each of a predetermined set of time points several measurements on each individual are taken. Data obtained from a panel study are called *panel data*. A panel study designed to have observations at k time points is called a k-wave panel design. Under this definition the term *panel study* could be used to refer to a large range of studies in biostatistics, particularly in epidemiology and **clinical trials**, although in many cases the term panel study is not used. The main advantage of a panel study is that individual changes over time can be modeled and the unobserved heterogeneity across individuals and over time can be taken into account.

In many panel studies, especially those lasting for a long period, attrition or loss to follow-up is an important issue. To keep the study population at a proper size during the study, two variations of the simple panel study can be used [3]. One is the *rotating panel study*, which replaces a part of the previous panel by a new panel at some time points and each individual only stays in the study for a certain period. Another is the *split panel design*, which recruits a new panel at some time points and keeps following all the panels until the end of the study.

Two important design issues are the calculation of sample size and the choice of time points to take the observations. For studies with continuous or categorical outcomes, standard sample size calculation procedures for repeated measurement models can be used. For panel studies that measure time to event (*see* **Survival Analysis, Overview**) the exact time of an event may not be obtained. In these cases the panel data are interval **censored**. With a predetermined number of waves, optimal time points can be obtained by using **optimal design** procedures [2]. These choose designs according to a criterion based on the Fisher **information matrix**, e.g. the D-optimality criterion. Similar procedures can be used for estimating duration of diseases or other recurrent events [1] (*see* **Repeated Events**). For the variable k there is a tradeoff between increasing costs and the gains in efficiency.

Several models can be used for the analysis of panel data. A simple model for continuous outcomes is a linear model with normally distributed variance components. Suppose a panel consists of n individuals and k observations are taken on each individual. Letting y_{ij} be observation j from subject i, this model can be written as [7]

$$y_{ij} = u_{ij} + \mathbf{x}_{ij}\boldsymbol{\beta}, \tag{1}$$

where $\mathbf{x}_{ij}$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of coefficients. Here u_{ij} is a term which can be either random or fixed, representing the variation among individuals and over time. For random u_{ij}, y_{ij}, $j = 1, \ldots, k$, may be correlated and several correlation patterns can be modeled by imposing properties on the u_{ij}. A simple model assumes that $u_{ij} = u_i + e_{ij}$, where u_i and e_{ij} are independent random errors. This model results in compound symmetry correlation in the linear models. Another commonly used model for u_{ij} in panel studies is the **time series** model or **serial correlation** model. An example is the AR(1) model (*see* **ARMA and ARIMA Models**) in which $u_{ij} = au_{ij-1} + e_{ij}$, for some constant a. A more complicated model may be a combination of these two. In (1) $\boldsymbol{\beta}$ (or a part of $\boldsymbol{\beta}$) can also be random. Models with random β are called **random coefficient** models [11]. For example, in a linear growth model, the growth rate for each subject may be different and may be considered as a random variable. Other kinds of models for panel studies include **lagged dependent variables** and lagged independent variables as **covariates**. These models are also known as the *dynamic* models.

To include discrete outcomes we may use (1) as a model for a latent variable, then discretize y_{ij} [6] or, more generally, assume that conditional on u_{ij}, y_{ij} belongs to the **exponential family**. This model is known as the generalized linear mixed model and can be used to model several types of data such as nonnormal continuous data and **binary**/count data (*see* **Generalized Linear Models for Longitudinal Data**). However, there are other models which are not included in this family [5], such as the **loglinear** models that are widely used in panel studies. A large amount of work has been done in the context of modeling repeated measurement data.

The main problem in analyzing panel data is the correlation between repeated measures. Several procedures have been developed for this purpose. For linear models with normally distributed errors, **marginal models** can easily be found and either **least squares** or generalized least squares can be used to estimate $\boldsymbol{\beta}$. The parameters for the variance components can be estimated using either the **maximum likelihood** estimator (MLE) or the

residual (or **restricted**) **maximum likelihood** estimator (REML), or by the **method of moments**.

The analysis of discrete data is much more difficult. For generalized linear mixed models the marginal likelihood function normally does not have a closed form and **numerical integration** has to be used to obtain the MLE. An alternative is to use approximations either when the **variance components** are small or the number of observations from each subject is large.

When $u_{ij} = u_i$ and u_i is fixed, the number of parameters for the subject effects increases with increasing total sample size so that the asymptotic properties of the MLE do not apply here. One approach to deal with this problem is to use conditioning (*see* **Conditionality Principle**). For example, for binary data the total responses are **sufficient statistics** for the subject effects. Conditioning on them can eliminate the subject effects and yield a consistent estimate of β. Conditioning is easy to perform when k is small but complicated even for moderate k. Conditioning can also be used for models with random u_{ij}.

In the last 10 years the **generalized estimating equation** (GEE) procedure [9] has been widely used. To use GEE only a marginal model and a **covariance matrix** for the outcomes are needed. This procedure gives consistent estimates as long as the model for the mean is correct, and in many cases it is effective compared with the MLE if the "working covariance matrix" is close to the true one.

Some special issues in panel data analysis are worth consideration. Often, in panel data u_{ij} contains both random subject effects and serial correlation. For the analysis of continuous outcome models, the Prais–Winston transformation can be used and the correlation coefficient can be tested by the generalized **Durbin–Watson test** [8]. For generalized linear mixed models the MLE is difficult to obtain and some approximation methods may be used.

Attrition also causes problem in the analysis of panel data, especially when loss to follow-up is outcome related. In panel studies loss to follow-up often depends on the previous observations. This is classified in missing data analysis as missing at random (MAR) [10]. In MAR data the missing mechanism depends on the observed data (here the previous observations) but not the missing data (here the outcome after the loss to follow-up). Several procedures have been proposed for analyzing MAR data and even more complicated missing mechanisms [4] (*see* **Nonignorable Dropout in Longitudinal Studies**).

References

[1] Albert, P.S. & Brown, C.H. (1991). The design of a panel study under an alternating Poisson-process assumption, *Biometrics* **47**, 921–932.

[2] Chappell, L. (1991). Sampling design of multiwave studies with an application to the Massachusetts Healthcare Panel Study, *Statistics in Medicine* **10**, 1945–1958.

[3] Curtin, L. & Feinleib, M. (1991). Considerations in the design of longitudinal surveys of health, in *Statistical Models for Longitudinal Studies of Health*, J.H. Dwyer, M. Feinleib, P. Lippert & H. Hoffmeister, eds. Oxford University Press, New York.

[4] Fitzmaurice, G.M., Heath, A.F. & Clifford, P. (1996). Logistic regression models for binary panel data with attrition, *Journal of the Royal Statistical Society, Series B* **159**, 249–264.

[5] Fitzmaurice, G.M., Laird, N.M. & Rotnitzky, A.G. (1993). Regression models for discrete longitudinal responses (with discussion), *Statistical Science* **8**, 284–309.

[6] Hamerle, A. & Ronning, G. (1995). Panel analysis for qualitative variables, in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg & M.E. Sobel, eds. Plenum Press, New York.

[7] Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge University Press, Cambridge.

[8] Hsiao, C. (1995). Panel analysis for metric data, in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg & M.E. Sobel, eds. Plenum Press, New York.

[9] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[10] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[11] Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, Oxford.

B. JONES & J. WANG

Parallel-line Assay

Parallel-line assays are the most common type of indirect analytical dilution assays (*see* **Biological Assay, Overview**). The response, which is measured on subjects at each of several fixed dose levels of the standard and test preparations, may be either quantitative or an "all-or-none" (quantal) variable, such as "dead" or "surviving" for each subject. In the parallel-line assay the measured response or an appropriate transformation of response, referred to as the response metameter, is a linear function of the logarithm of the dose administered. The condition of similarity, which is a prerequisite of all analytical dilution assays, specifies that the test preparation behaves like a dilution of the standard preparation. Furthermore, the expected response (y_S) of the standard is generally assumed to be linear in some known power of dose; that is,

$$\mathrm{E}[y_S|d_S] = \alpha_S + \beta x_S,$$

where

$$x_S = d_S^{\lambda}.$$

When $\lambda \to 0$, then $x_S = \log d_S$. If we next also assume that a dose of the standard (d_S) is equivalent to the product of the relative potency parameter ρ, times the dose of the test, d_T for all dose levels, then

$$\mathrm{E}[y_T|\log d_T] = \mathrm{E}[y_S|\log(\rho d_T)]$$

and

$$x_S = \log d_S = \log(\rho d_T) = \log\rho + x_T.$$

Therefore (see Figure 1),

$$\mathrm{E}[y_T|x_T] = \alpha_S + \beta\log\rho + \beta x_T = \alpha_T + \beta x_T.$$

When the responses are equivalent for the standard and test, i.e.

$$\alpha_S + \beta x_S = \alpha_S + \log\rho + \beta x_T.$$

Then

$$\log\rho = \frac{\alpha_T - \alpha_S}{\beta}.$$

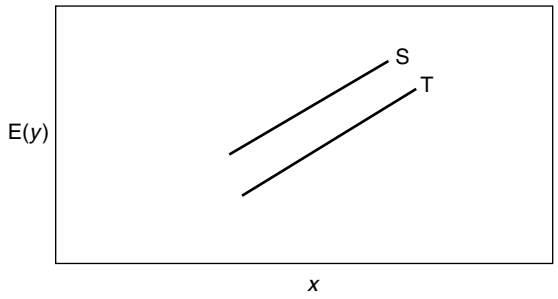

Figure 1 Expected response (E(y)) vs. dose metameter (x) for standard (S) and test (T) preparations in a parallel-line assay

Relative Potency Estimation and Validity Tests

Quantitative Responses

Estimates of the parameters are readily calculated from standard **least squares** regression programs. The common slope and intercepts for test and standard preparations are estimated as follows (see Table 1 for notation):

$$\hat{\beta} = \frac{\sum_{S,T} S_{xy}}{\sum_{S,T} S_{xx}},$$

$$\hat{\alpha}_T = \overline{y}_T - \hat{\beta}\overline{x}_T,$$

and

$$\hat{\alpha}_S = \overline{y}_S - \hat{\beta}\overline{x}_S.$$

Accordingly, the log relative potency estimate is

$$\log\hat{\rho} = \frac{\hat{\alpha}_T - \hat{\alpha}_S}{\hat{\beta}} = (\overline{x}_S - \overline{x}_T) - \frac{(\overline{y}_S - \overline{y}_T)}{\hat{\beta}}.$$

Before proceeding to calculate the **confidence intervals** for ρ, it is useful to verify that the assumptions of the model are not seriously violated. Table 1 summarizes in general form the **analysis of variance** (ANOVA) used to validate the assumptions. The method of least squares used to estimate the regression parameters, and to determine the confidence interval for ρ, further assumes that the y_i are

independent normally distributed variables and that

Table 1 Analysis of variance in quantitative parallel-line assay

Source	df	Sums of squares	Mean squares	F
Total	$N-1$	$\mathrm{SS}_y = \sum_{\mathrm{S,T}} \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} \left(y_{pij} - \overline{y}\right)^2$	$\mathrm{MS}_y = \dfrac{\mathrm{SS}_y}{N-1}$	
Among doses	$K-1$	$\mathrm{SS}_{\mathrm{D}} = \sum_{\mathrm{S,T}} \sum_{i=1}^{K_p} n_{pi} \left(\overline{y}_{pi} - \overline{y}\right)^2$	$\mathrm{MS}_{\mathrm{D}} = \dfrac{\mathrm{SS}_{\mathrm{D}}}{K-1}$	$\dfrac{\mathrm{MS}_{\mathrm{D}}}{\mathrm{MS}_{\mathrm{E}}}$
Preparations	1	$\mathrm{SS}_{\mathrm{P}} = \sum_{\mathrm{S,T}} N_p \left(\overline{y}_p - \overline{y}\right)^2$	$\mathrm{MS}_{\mathrm{P}} = \mathrm{SS}_{\mathrm{P}}$	$\dfrac{\mathrm{MS}_{\mathrm{P}}}{\mathrm{MS}_{\mathrm{E}}}$
Common slope	1	$\mathrm{SS}_{\mathrm{R}} = \dfrac{\left(\sum_{\mathrm{S,T}} S_{xy}\right)^2}{\sum_{\mathrm{S,T}} S_{xx}}$	$\mathrm{MS}_{\mathrm{R}} = \mathrm{SS}_{\mathrm{R}}$	$\dfrac{\mathrm{MS}_{\mathrm{R}}}{\mathrm{MS}_{\mathrm{E}}}$
Parallelism	1	$\mathrm{SS}_{\mathrm{PL}} = \sum_{\mathrm{S,T}} \left(\dfrac{S_{xy}^2}{S_{xx}}\right) - \mathrm{SS}_{\mathrm{R}}$	$\mathrm{MS}_{\mathrm{PL}} = \mathrm{SS}_{\mathrm{PL}}$	$\dfrac{\mathrm{MS}_{\mathrm{PL}}}{\mathrm{MS}_{\mathrm{E}}}$
Nonlinearity	$K-4$	$\mathrm{SS}_{\mathrm{NL}} = \mathrm{SS}_{\mathrm{D}} - \mathrm{SS}_{\mathrm{P}} - \mathrm{SS}_{\mathrm{R}} - \mathrm{SS}_{\mathrm{PL}}$	$\mathrm{MS}_{\mathrm{NL}} = \mathrm{SS}_{\mathrm{NL}}/(K-4)$	$\dfrac{\mathrm{MS}_{\mathrm{NL}}}{\mathrm{MS}_{\mathrm{E}}}$
Within doses (error)	$N-K$	$\mathrm{SS}_{\mathrm{E}} = \mathrm{SS}_y - \mathrm{SS}_{\mathrm{D}}$	$\mathrm{MS}_{\mathrm{E}} = \dfrac{\mathrm{SS}_{\mathrm{E}}}{N-K} = \hat{\sigma}^2$	

Notation (adapted from [8]):

$$n_{pi} = \text{number of observations at dose } i \text{ of preparation } p,$$

$$y_{pij} = \text{response for subject } j \text{ to dose } i \text{ of preparation } p,$$

$$x_{pij} = \text{dose metameter for subject } j \text{ to dose } i \text{ of preparation } p,$$

where

$$i = 1, 2, \ldots, K_p, \qquad p = \text{S(Standard) or T(Test);}$$

and

$$K_p = \text{number of dose levels of preparation } p,$$

$$N_p = \sum_{i=1}^{K_p} n_{pi},\ N = \sum_{\mathrm{S,T}} N_p,\ K = \sum_{\mathrm{S,T}} K_p,$$

$$S_{yy} = \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (Y_{pij} - \overline{y}_p)^2,$$

where

$$\overline{y}_p = \frac{\sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} y_{pij}}{N_p}, \qquad S_{xx} = \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (x_{pij} - \overline{x}_p)^2,$$

in which

$$\overline{x}_p = \frac{\sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} x_{pij}}{N_p},$$

and

$$S_{xy} = \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (y_{pij} - \overline{y}_p)(x_{pij} - \overline{x}_p).$$

$\text{var}(y_i) = \sigma^2$ for all dose levels d_i. Absence of heteroscedasticity should be verified, whenever possible, by an appropriate statistical test.

The five **hypothesis tests** of interest for parallel-line assays, as shown in Table 1, are:

1. $H_1 : \alpha_T = \alpha_S$ and $\beta_S = \beta_T = 0$. Reject if

$$F = \frac{MS_D}{MS_E} > F_{(K-1,N-K)},$$

the critical level of the ***F* distribution** on $(K - 1, N - K)$ **degrees of freedom** (df). Failure to reject indicates the likelihood of poor selection of dose levels, e.g. that they were outside a dose range for observing changes in response for either of the two preparations. (Calculated variance ratios, F, in the analysis of variance are often conventionally compared to tabulated $F_{(df_1,df_2)}$ values corresponding to a 5% significance level (*see* **Level of a Test**).)

2. H_2: Reject linearity if

$$F = \frac{MS_{NL}}{MS_E} > F_{(K-4,N-K)}.$$

If this hypothesis is rejected, indicating statistical invalidity in the linearity assumption, then a nonlinear regression function, such as a quadratic, may be applicable. More usually, a different part of the dose range might be used, or, alternatively, a **transformation** of the response variable might achieve linearity.

3. $H_3 : \beta_S = \beta_T = \beta$. If

$$F = \frac{MS_{PL}}{MS_E} > F_{(1,N-K)},$$

then there is evidence of fundamental invalidity of the assay and the assumption of a constant relative potency is suspect.

4. H_4: Reject equivalence of mean responses for the test and standard preparations if

$$F = \frac{MS_P}{MS_E} > F_{(1,N-K)}.$$

If the dose-ranges for the two preparations are not equivalent, then the parallel lines may be far apart even when the mean responses are the same. Although in bioassay the differences between treatment means are not of primary interest, a significant mean square for H_4 indicates a design problem associated with the doses chosen for the experiment. On the other hand, a nonsignificant test cannot necessarily be interpreted as assurance that the dose ranges were appropriate [8].

5. $H_5 : \beta = 0$; reject if

$$F = \frac{MS_R}{MS_E} > F_{(1,N-K)}.$$

It is expected that a well-designed assay will have a highly significant common regression coefficient.

If the analysis of variance does not indicate any validity concerns and there was no evidence of heteroscedasticity, **Fieller's theorem** [6] for ratio estimators can be applied to obtain the confidence interval for ρ. In a parallel-line assay the 95% confidence intervals for $\log \rho$ are calculated as

$$\log \hat{\rho}_L, \log \hat{\rho}_U = (\bar{x}_S - \bar{x}_T) + \left[(\log \hat{\rho} - \bar{x}_S + \bar{x}_T) \pm \frac{\hat{\sigma} t}{\hat{\beta}} \left\{ (1 - g) \left(\frac{1}{N_S} + \frac{1}{N_T} \right) + \frac{(\log \hat{\rho} - \bar{x}_S + \bar{x}_T)^2}{\sum S_{xx}} \right\}^{1/2} \right] \Big/ (1 - g),$$

where

$$g = \frac{t^2 \hat{\sigma}^2}{\hat{\beta}^2 \sum_{S,T} S_{xx}}$$

and t has $N - K$ degrees of freedom.

Quantal Responses

Toxicological assays frequently involve **quantal responses**. In quantal assays the response is the proportion of subjects responding at a fixed dose level, d_i. If the proportion responding is plotted against $\log d_i$, the curve is generally sigmoid in shape. This leads to consideration of various probability density functions for the tolerance distribution (*see* **Quantal**

Response Models). The normal and the **logistic** are the two most frequently used tolerance distributions, but others such as the **Cauchy** or the angular have been proposed for some applications. To illustrate, when the normal probability density is specified, the expected proportion responding at dose, d_i is

$$P_i = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{y_i=(x_i-\mu)/\sigma} \exp\left(-\frac{1}{2}u^2\right) \mathrm{d}u.$$

Hence the linearizing transformation is the normal equivalent deviate (ned)

$$y_i = -\frac{\mu}{\sigma} + \frac{1}{\sigma}x_i = \alpha + \beta x_i,$$

where y_i is usually referred to as the probit of P_i [2, 3]. (In order to avoid negative values, most references add 5 to each ned. [8].)

When the logistic transformation is used

$$P_i = \frac{1}{1 + \exp[-(\alpha + \beta x_i)]}$$

and

$$y_i = \text{logit}\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta x_i.$$

Often, the estimates of ρ will be quite similar regardless of which linearizing transformation is employed, since the curves will overlap over a wide range of responses, with only the extremes differing. There are varying viewpoints about how to select an appropriate theoretical unknown tolerance distribution. Furthermore, validity tests seldom discriminate between the normal and logistic functions in practice, so that for estimation purposes there is limited basis for preferring one over the other.

If the probability of responding at a particular dose is a **binomial**, such that

$$\Pr(r_i \text{ responding}) = \binom{n_i}{r_i} P_i^{r_i} Q_i^{n_i - r_i}$$

and P is defined by the ned above, then the parameters α and β can be estimated iteratively by usual **maximum likelihood** procedures (*see* **Optimization and Nonlinear Equations**). Estimation involves weighting the observed proportions at specified doses since the binomial responses will have unequal variances. (Weights are symmetric about the middle of the tolerance distribution, with lesser weights at the extremes.) For the probit transformation, the weight at each fixed log dose d_i is $n_i w_i$, where n_i is the number of subjects at $\log d_i$ and

$$w_i = \frac{\Phi_i}{P_i Q_i}.$$

Φ_i is the ordinate of the standard normal density curve corresponding to y_i, the expected probit of P_i. While the calculations are somewhat laborious, they are readily programmed and available in standard packages such as SAS [11] or GLIM [9] (*see* **Software, Biostatistical**). Following the notation adopted above for quantitative response curves, the weighted sums of squares and cross products obtained by maximum likelihood estimation are used to calculate the estimated log relative potency,

$$\log \hat{\rho} = \frac{\hat{\alpha}_\mathrm{T} - \alpha_\mathrm{S}}{\hat{\beta}} = \overline{x}_\mathrm{S} - \overline{x}_\mathrm{T} - \frac{\overline{y}_\mathrm{S} - \overline{y}_\mathrm{T}}{\hat{\beta}}$$

and

$$\hat{\beta} = \frac{\sum_{\mathrm{S,T}} S_{xy}}{\sum_{\mathrm{S,T}} S_{xx}} = \frac{\sum_{\mathrm{S,T}} Snw(x - \overline{x})(y - \overline{y})}{\sum_{\mathrm{S,T}} Snw(x - \overline{x})^2},$$

where $\overline{x}_\mathrm{S}$, $\overline{x}_\mathrm{T}$, $\overline{y}_\mathrm{S}$, and $\overline{y}_\mathrm{T}$ are weighted means, e.g. $\overline{x}_\mathrm{S} = Snwx/Snw$, etc.

As before, the confidence interval for $\log \rho$ is calculated by application of Fieller's theorem. Since the standard normal density is assumed for the tolerance distribution, the within-subject error (σ^2) is assumed to be 1 and normal distribution values are used for specifying confidence intervals. The 95% confidence intervals are, therefore,

$$\log \hat{\rho}_\mathrm{L}, \log \hat{\rho}_\mathrm{U}$$

$$= (\overline{x}_\mathrm{S} - \overline{x}_\mathrm{T}) + \Bigg[(\log \hat{\rho} - \overline{x}_\mathrm{S} + \overline{x}_\mathrm{T})$$

$$\pm \frac{1.96}{\hat{\beta}} \Bigg\{ (1 - g) \sum_{\mathrm{S,T}} \left(\frac{1}{Snw}\right)$$

$$+ \frac{(\log \hat{\rho} - \overline{x}_\mathrm{S} + \overline{x}_\mathrm{T})^2}{\sum_{\mathrm{S,T}} S_{xx}} \Bigg\}^{1/2} \Bigg] \Big/ (1 - g),$$

where

$$g = \frac{1.96^2}{\hat{\beta}^2 \sum_{\mathrm{S,T}} S_{xx}}.$$

Validation of the model assumed for quantal assays is analogous to the ANOVA performed for quantitative parallel line assays. In quantal assays the statistical significance for hypothesis tests is compared to the **chi-square distribution** with appropriate degrees of freedom. Hypothesis tests of interest are:

1. H_1: reject if

$$X^2 = \sum_{\mathrm{S,T}} S_{yy} - \sum_{\mathrm{S,T}} \left\{ \frac{S_{xy}^2}{S_{xx}} \right\} > \chi^2_{K_\mathrm{S}+K_\mathrm{T}-4}.$$

   H_1 is a test for nonlinearity of regression. When the calculated X^2 is not significant, the observed data are not inconsistent with the assumption of a normal log tolerance distribution. Where H_1 is rejected, consideration should be given as to whether there is a systematic deviation from normality in the dose–response relationship or whether there is nonsystematic heterogeneity. In the latter case, it is customary to employ a heterogeneity factor for the variance in calculating the confidence intervals; the variances are multiplied by the X^2 calculated for nonlinearity divided by its degrees of freedom (*see* **Overdispersion**). The estimated variance is, therefore,

$$\hat{\sigma}^2 = \frac{X^2}{K_\mathrm{S} + K_\mathrm{T} - 4}.$$

   All variances will thus have only $K_\mathrm{S} + K_\mathrm{T} - 4$ degrees of freedom when this adjustment is made.
2. H_2: reject if

$$X^2 = \sum_{\mathrm{S,T}} \left\{ \frac{S_{xy}^2}{S_{xx}} \right\} - \frac{\left(\sum_{\mathrm{S,T}} S_{xy} \right)^2}{\sum_{\mathrm{S,T}} S_{xx}} > \chi_1^2.$$

   H_2 is a test for parallelism of the transformed regression lines, and therefore a nonsignificant X^2 verifies the fundamental validity of the assay.

A common extension employed in quantal assays introduces an additional parameter, C, in the estimation to incorporate a proportion responding (dying) at zero dose level, which provides for a natural event (death) rate in the absence of exposure. This leads to an adjusted proportion responding at each dose level:

$$P_i' = C + (1 - C) P_i.$$

Therefore,

$$P_i = \frac{P_i' - C}{1 - C}.$$

Estimation of parameters, including C, can proceed as previously using standard iterative approaches to obtain the maximum likelihood estimates. The expression for P_i above is generally referred to as Abbot's formula [1].

Detailed illustrations of the design and analysis of quantal assays are provided in [7, 8], and [10].

Additional Remarks

The formulas presented in Table 1 are general in form and do not depend on having a symmetrical, balanced design for the assay or equal spacing between doses. For both design and efficiency reasons, a symmetric design with equal numbers of subjects at each dose level is preferable. Modern statistical analysis software such as GLIM [9] or SAS [11] can readily accommodate either unsymmetric or symmetric designs. At least three dose levels of the test and standard preparations are required to test all validity assumptions in the ANOVA. With regard to quantal assays, analyses using normal or logistic tolerance distributions can also be done in most general statistical software packages, such as SAS, GLIM, or **S-PLUS** (for examples, see references to statistical software at the end of this section or [12]).

The methodology is easily extended to designs in which multiple test preparations are simultaneously compared to the same standard. In parallel-line assays with more than one test preparation, optimal allocation among r test preparations given a total number of subjects would be

$$N_\mathrm{s} = r^{1/2} N_\mathrm{T}$$

assuming that the variance of ρ is approximately proportional to $N_\mathrm{s}^{-1} + N_\mathrm{T}^{-1}$.

Finney [8] demonstrates that the analyses described above may provide very similar estimates of the relative potency and its confidence intervals even when the normality assumption is not fulfilled, but other model assumptions are not seriously violated. Information derived from evaluating validity across a related class of independent assays is preferred to justify an appropriate choice for the response metameter.

A useful indicator of the optimality of assay design is the value of g from Fieller's theorem. In quantitative parallel-line assays, g can be rewritten as

$$g = \frac{F_{\text{tabulated}}}{F_{\text{calculated}}},$$

where the numerator is the variance ratio for regression. Well-designed and well-executed assays will generally have values of $0 < g < 0.05$, whereas high values of g indicate some lack of efficient design. A comparative measure proposed by Bliss & Cattrell [4] to evaluate the sensitivity among assays is $\hat{\sigma}/\hat{\beta}$. Smaller values of $\hat{\sigma}/\hat{\beta}$ are indicative of "better sensitivity", whereas larger values indicate lesser precision, less steep slopes or a combination of both.

Finney [7, 8] and Brown [5] give detailed discussions of considerations important for optimizing design of quantitative or quantal parallel-line assays.

References

[1] Abbott, W.S. (1925). A method of computing the effectiveness of an insecticide, *Journal of Economic Entomology* **18**, 265–267.

[2] Bliss, C.I. (1934). The method of probits, *Science* **79**, 38–39.

[3] Bliss, C.I. (1934). The method of probits – a correction, *Science* **79**, 409–410.

[4] Bliss, C.I. & Cattrell, M. (1943). Biological assay, *Annual Review of Physiology* **5**, 479–539.

[5] Brown, B.W. Jr. (1966). Planning a quantal assay of potency, *Biometrics* **22**, 322–329.

[6] Fieller, E.C. (1940). The biological standardization of insulin, *Journal of the Royal Statistical Society, Supplement* **7**, 1–64.

[7] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, Cambridge.

[8] Finney, D.J. (1978). *Statistical Method in Biological Assay*, 3rd Ed. Griffin, London, pp. 148–178, 297–315.

[9] Francis, B., Green, M. & Payne, C., eds (1994). *The GLIM System. Release 4 Manual*, Clarendon Press, Oxford, pp. 429–451.

[10] Hubert, J.J. (1984). *Bioassay*, 2nd Ed. Kendall/Hunt, Dubuque, pp. 26–39.

[11] SAS Institute (1994). *SAS/STAT® User's Guide*, Version 4, 4th Ed. SAS Institute, Cary, Chapters 24, 35, and 36.

[12] Venables, W.N. & Ripley B.D. (1994). *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, pp. 189–195.

(*See also* **Radioimmunoassay**; **Slope–Ratio Assay**)

CAROL K. REDMOND

Parametric Models in Survival Analysis

Survival analysis in biostatistical applications involves the analysis of times to events (*see* **Survival Analysis, Overview**); for conciseness we refer to these times as lifetimes. Sometimes the objective is to model or describe the distribution of lifetimes in a single homogeneous population of individuals. More generally, we wish to compare distributions or to assess the relationship of **explanatory variables** to lifetimes. In some instances different types of events may occur to an individual, and the joint distribution of several lifetimes may be of interest.

Parametric survival models are ones in which the distribution of lifetimes is specified up to a parameter $\boldsymbol{\theta}$ that is of finite, and usually rather small, dimension. A major advantage of parametric models is the availability of straightforward methods of **estimation** and **inference** based on the **likelihood** function. In addition, parametric representations facilitate the accumulation of scientific evidence across different but similar studies. However, fully parametric models involve stronger assumptions than semi- or nonparametric models. The choice of a parametric model is more often based on a combination of tractability and ability to fit the data than on any deep physical motivation, and it is therefore important to check the adequacy of models and to consider the sensitivity of inferences and conclusions to plausible variations in them. The next two sections describe the main families of models used for parametric survival analysis, and how such analysis is carried out.

Parametric Models

Following standard terminology, let $T \geq 0$ be a **random variable** representing lifetime and assume that T has a continuous distribution with cumulative distribution function (cdf) $F(t) = \Pr(T \leq t)$, probability density function (pdf) $f(t) = F'(t)$, survivor function $S(t) = \Pr(T \geq t)$, **hazard** function $h(t) = f(t)/S(t)$, and cumulative hazard function $H(t) = \int_0^t h(u)\,\mathrm{d}u$. Any one of these five functions specifies the distribution of T (*see* **Survival Distributions and Their Characteristics**). It is useful to note that $S(t) = \exp[-H(t)]$.

Parametric models are specified in terms of a parameter $\boldsymbol{\theta}$, and in this case we write $F(t;\boldsymbol{\theta}), f(t;\boldsymbol{\theta})$, and so on. A wide variety of models has been used for univariate lifetime distributions. The merits of specific families of distributions are often discussed both in terms of their ability to fit existing data and the shapes of their hazard functions, which specify the instantaneous probability of death or failure at time t, given survival up to t. In the following subsection we describe some important parametric families. Two additional subsections discuss models involving covariates and models that are multivariate or involve random effects.

Some Important Parametric Families

Below we summarize the most widely used parametric lifetime distribution models. Unless specified otherwise, the range of the lifetime variable t is $t \geq 0$.

Exponential distribution. The **exponential distribution** has survivor, density, and hazard functions of the form

$$S(t) = \exp(-\lambda t), \quad f(t) = \lambda \exp(-\lambda t) \quad h(t) = \lambda,$$

where $\lambda > 0$ is a parameter; it is easily seen that $\mathrm{E}(T) = \lambda^{-1}$. The exponential family has constant hazard functions and the associated "lack of memory" property: $\Pr(T \geq t + x | T \geq t) = \Pr(T \geq x)$. These characteristics are obviously restrictive, and limit the applicability of the model. Approximate sample-size calculations for clinical trials (*see* **Sample Size Determination in Survival Analysis**) are sometimes based on it, but require careful application [22]. It is a useful fact that if T is an arbitrary random variable with cumulative hazard function $H(t)$, then the transformed variable $H(T)$ has a standard ($\lambda = 1$) exponential distribution.

Weibull distribution. The **Weibull distribution** with inverse scale parameter $\lambda > 0$ and shape parameter $\delta > 0$ has survivor, density, and hazard functions

$$S(t) = \exp[-(\lambda t)^{\delta}],$$

$$f(t) = \delta\lambda(\lambda t)^{\delta-1} \exp[-(\lambda t)^{\delta}], \quad h(t) = \delta\lambda(\lambda t)^{\delta-1}.$$

The two parameters allow the Weibull density to take a variety of shapes, and the hazard function is either monotone increasing, decreasing, or constant according to whether $\delta > 1$, $\delta < 1$, or $\delta = 1$. The case $\delta = 1$ gives the exponential distribution, and, more generally, T^δ has an exponential distribution with hazard function λ^δ. The Weibull distribution often fits biostatistical survival data well, and in some applications there is physical justification for a Weibull model through weakest link or multistage waiting time arguments (e.g. Armitage & Doll [3]) (*see* **Multistage Carcinogenesis Models**).

Lognormal distribution. The **lognormal** distribution has the property that log lifetime, $Y = \log T$, is normally distributed with mean μ and variance σ^2. This gives a two-parameter family with survivor and density functions

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right),$$

$$f(t) = \frac{1}{(2\pi)^{1/2}\sigma t}\exp\left[-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right],$$

where $\Phi(x)$ is the standard normal cdf. The hazard functions are nonmonotonic: they rise from 0 at $t = 0$ to a maximum, and then decrease to 0 as $t \to \infty$. This shape is implausible for some, but by no means all, survival analysis applications. Moreover, the proportion of the population with lifetimes greater than the hazard function mode can vary widely according to the values of (μ, σ). The lognormal model has the advantage that when there is no **censoring** of lifetimes we can apply simple normal distribution inference methods to the log lifetimes (see the section "Model Fitting and Inference" below). However, censoring is more the rule than the exception in survival data, and so this is a relatively minor advantage.

Log-logistic distribution. The log-logistic family has survivor, density, and hazard functions of the form

$$S(t) = \frac{1}{1 + (\lambda t)^\delta}, \quad f(t) = \frac{\delta\lambda(\lambda t)^{\delta-1}}{[1 + (\lambda t)^\delta]^2},$$

$$h(t) = \frac{\delta\lambda(\lambda t)^{\delta-1}}{1 + (\lambda t)^\delta},$$

where $\lambda > 0$ and $\delta > 0$ are inverse scale and shape parameters. It derives its name from the fact that $Y = \log T$ has a **logistic distribution**. This family is similar to the lognormal family, and its hazard functions also increase from 0 at $t = 0$ to a maximum, and then decrease to 0 as t becomes large. It is slightly more convenient than the lognormal family for survival analysis, since its survivor and hazard functions have simple closed-form expressions.

Location-scale models for $Y = \log T$. The Weibull, lognormal, and log-logistic distributions share a property: log lifetime Y has a **location-scale** distribution with survivor and density functions of the form

$$S_Y(y) = G\left(\frac{y - \mu}{\sigma}\right), \qquad f_Y(y) = \frac{1}{\sigma}g\left(\frac{y - \mu}{\sigma}\right), \tag{1}$$

where $\mu(-\infty < \mu < \infty)$ is a location parameter, $\sigma > 0$ is a scale parameter, and $G(z)$ and $g(z) = -G'(z)$ are the survivor function and pdf for the standardized variable $Z = (Y - \mu)/\sigma$. When T is Weibull the distribution of $Y = \log T$ is called the **extreme-value** distribution. When T is lognormal, Y is normal, and when T is log-logistic, Y is logistic. The three respective survivor functions for Z are:

1. extreme-value $\quad G(z) = \exp(-e^z)$,
2. normal $\quad G(z) = 1 - \Phi(z)$,
3. logistic $\quad G(z) = (1 + e^z)^{-1}$,

where $-\infty < z < \infty$ in each case.

Other lifetime models can be formed by choosing other distributions for Z or Y. Sometimes the distribution of Z is allowed to depend upon one or two additional parameters, thus giving very flexible three- or four-parameter lifetime distributions. Two such families are the generalized log **gamma** [16; 23, Section 5.3] and the Burr–**Pareto** [6, 25]. The latter, for example, gives a three-parameter distribution with survivor function for T of the form

$$S(t) = [1 + \alpha^{-1}(\lambda t)^\delta]^{-\alpha}, \tag{2}$$

where $\alpha > 0, \lambda > 0$, and $\delta > 0$. This is obtained from (1) by taking $G(z) = (1 + \alpha^{-1}e^z)^{-\alpha}$ and defining $\lambda = e^{-\mu}$ and $\delta = \sigma^{-1}$. It may also be obtained from a Weibull **frailty** model with gamma-distributed

random effects (e.g. [12]). The special case $\alpha = 1$ gives the log-logistic distribution, and as $\alpha \to \infty$ the Weibull distribution is obtained, so the Burr–Pareto model may be used to discriminate between these two models.

The generalized log gamma model is similar, and includes the Weibull and lognormal distributions as special cases. An even more comprehensive four-parameter family is obtained by choosing Z to have a standardized log ***F* distribution** [20, pp. 28, 63; 7]. This includes as special cases all of the distributions considered thus far.

Some other models. A number of other parametric models are occasionally used as lifetime distributions; two of the more common models are the gamma and the **inverse Gaussian**. In some circumstances it is convenient to formulate a model so as to give certain shapes for the hazard function. For example, the Gompertz distribution has loglinear hazard, $\log h(t) = \alpha_1 + \alpha_2 t$. Another example arises in contexts where the hazard function is thought to exhibit the so-called "bathtub" shape, in which it decreases from a local maximum at $t = 0$ to a local minimum, and then increases thereafter. None of the models discussed so far allows hazard functions of this shape. Models that do have been proposed by various authors; Hjorth [18], for example, considers a three-parameter family with $h(t) = \lambda t + \alpha/(1 + \beta t)$, where $\lambda > 0, \alpha > 0$, and $\beta > 0$. If $\lambda \leq \alpha\beta$, then the hazard function has the bathtub shape.

Finally, lifetimes are by convention usually taken to be nonnegative, and the models discussed so far all assume $T \geq 0$. In some applications it is argued that there exists an unknown minimum or "threshold" time $\gamma > 0$ before which the event in question cannot occur. In that case we can model the lifetime distribution by replacing t by $t - \gamma$ in previous expressions for survivor functions, hazard functions, and so on. For example, in the case of the exponential distribution this yields a two-parameter model with pdf $f(t) = \lambda \exp[-\lambda(t - \gamma)], t \geq \gamma$. A practical difficulty when γ is unknown is that threshold parameters are difficult to estimate precisely unless there is a considerable amount of data.

Regression Models

In most survival analysis applications there are groups of individuals to be compared, or explanatory variables, such as individual characteristics or environmental conditions, whose relationship to lifetime is to be examined. A small number of homogeneous groups may be compared by fitting separate lifetime distributions for each group, but more generally, regression models that employ covariates to model the effects of explanatory variables are used. For example, in studying the survival time from diagnosis of patients with multiple myeloma, Krall et al. [21] considered 16 covariates representing factors such as the white blood-cell count at diagnosis, the presence or absence of infection at diagnosis, and the sex and age of the subject.

Survival analysis may also involve covariates that vary over time. **Time-varying covariates** are usually handled through hazard-based **semiparametric** models, especially Cox's proportional or multiplicative hazards model [1, Chapter VI; 20, Chapters 4 and 5] (*see* **Cox Regression Model**). This is to a large extent because **"partial" likelihood** methods of analysis for such models [9] tend to be simpler than methods based on fully parametric models. Most of the discussion below deals with cases where the covariates are fixed, i.e. constant over time, since that is the main domain of application of parametric models. We return to time-varying covariates later in the article.

Let T denote a lifetime and $\mathbf{x}$ a $p \times 1$ vector of covariates associated with each individual. Fully parametric regression models are in principle obtained by taking any parametric lifetime distribution and allowing its parameters to depend upon $\mathbf{x}$ in some specified way. For example, Weibull regression models are obtained by taking

$$S(t|\mathbf{x}) = \exp\{-[\lambda(\mathbf{x})t]^{\delta(\mathbf{x})}\}, \tag{3}$$

where we write $S(t|\mathbf{x})$ to denote the survivor function of T given $\mathbf{x}$. To complete the model we must specify how $\lambda(\mathbf{x})$ and $\delta(\mathbf{x})$ depend on $\mathbf{x}$. In many applications it happens that $\delta(\mathbf{x})$ does not vary much with $\mathbf{x}$, and so $\delta(\mathbf{x}) = \delta$ is assumed fixed. Models with $\lambda(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, are often used; this form is flexible and automatically constrains $\lambda(\mathbf{x})$ to be nonnegative.

Although a very wide range of parametric regression models is possible, two types of models dominate. These are termed **accelerated failure time** (AFT) and **proportional hazards** (PH) models; each

has both fully parametric and semiparametric versions. They are described in turn.

Accelerated failure time models. It was noted previously that for several common parametric survival models the distribution of log lifetime Y is of location-scale form (1). An important class of regression models is obtained by allowing the location parameter μ in (1) to depend on $\mathbf{x}$, so that the survivor function of Y, given $\mathbf{x}$, is

$$S_Y(y|\mathbf{x}) = G\left[\frac{y-\mu(\mathbf{x})}{\sigma}\right], \tag{4}$$

where $G(z)$ is a specified survivor function on $-\infty < z < \infty$. We may also write (4) as

$$Y = \mu(\mathbf{x}) + \sigma\varepsilon, \tag{5}$$

where the "error" ε has a distribution with survivor function $G(z)$. In the case where $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, we have a linear model, although the errors are not necessarily normal. The name "accelerated failure time" derives from the corresponding model for T given $\mathbf{x}$, which has survivor function

$$S(t|\mathbf{x}) = G\left[\frac{\log t - \mu(\mathbf{x})}{\sigma}\right] = S_0\left[\frac{t}{\alpha(\mathbf{x})};\delta\right], \tag{6}$$

where $\alpha(\mathbf{x}) = \exp[\mu(\mathbf{x})]$, $\delta = \sigma^{-1}$, and $S_0(t;\delta) = G(\delta \log t)$. The effect of the covariates is to alter the time scale for t multiplicatively, i.e. either to accelerate or decelerate time. Other models that alter the time scale are also possible; in particular, nonlinear **transformations** of t may be used. In this sense, (6) is actually a special type of "accelerated failure time" model.

The most widely used parametric AFT models are those for which ε in (5) has either a standard extreme value, logistic, or normal distribution, corresponding to T being Weibull, log-logistic and lognormal, respectively. However, other distributions may be used, as described in the preceding subsection. It may be noted that the Weibull model, (3), is an AFT model only if $\delta(\mathbf{x}) = \delta$. Semiparametric versions of the AFT family are also based on (4), but do not assume any specific form for $G(z)$ (*see* **Semiparametric Regression**).

Proportional hazards models. A proportional hazards (PH) family of regression models is one for which the hazard function of T given $\mathbf{x}$ is of the form

$$h(t|\mathbf{x}) = h_0(t)r(\mathbf{x}), \tag{7}$$

where $r(\mathbf{x})$ is a positive-valued function and $h_0(t)$ is a "baseline" hazard function. The hazard function for any individual is proportional to $h_0(t)$, hence the name of the family. Fully parametric PH models specify parametric forms $h_0(t;\boldsymbol{\alpha})$ and $r(\mathbf{x};\boldsymbol{\beta})$ for the two components of (7); in its semiparametric version [9], $h_0(t)$ is left unspecified. The specification $r(\mathbf{x};\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$ is often used.

From (7) and the relationship $S(t) = \exp[-H(t)]$ it follows that the survivor function of T given $\mathbf{x}$ is of the form

$$S(t|\mathbf{x}) = S_0(t)^{r(\mathbf{x})}, \tag{8}$$

where $S_0(t) = \exp[-H_0(t)]$ is the baseline survivor function. A feature of PH models is that if $S_0(t;\boldsymbol{\alpha})$ is in a family of parametric models, then $S(t|\mathbf{x})$ is not always in the same family. It is, however, if $h_0(t)$ is of the form $\alpha_1 h_1(t;\alpha_2)$; this includes the Weibull family, (3), with $\delta(\mathbf{x}) = \delta$. It is also easily checked that the family of Weibull regression models, (3), with $\delta(\mathbf{x}) = \delta$, is both a PH and an AFT model, and that it is the only set of distributions with this property. In other words, the class of PH and the class of AFT models are distinct aside from models (3) with $\delta(\mathbf{x}) = \delta$.

The AFT and PH models make fairly strong assumptions about the relationship between T and $\mathbf{x}$. Among other things, they imply that the survivor functions for individuals with different covariate vectors never cross. Two extensions that relax these assumptions are often useful. The first retains a location-scale model, (4), for $\log T$, but allows the scale parameter σ to depend on $\mathbf{x}$; the parametric form $\sigma(\mathbf{x}) = \exp(\boldsymbol{\gamma}'\mathbf{x})$ is convenient (e.g. [28]). The second is a relaxation of the PH assumption that allows the regression coefficients in $r(\mathbf{x})$ of (7) to change over time, giving

$$h(t|\mathbf{x}) = h_0(t)r[\mathbf{x};\boldsymbol{\beta}(t)].$$

A simple example is the two-step model in which there is a specified value τ such that $\boldsymbol{\beta}(t) = \boldsymbol{\beta}_1$ for $0 \le t \le \tau$ and $\boldsymbol{\beta}(t) = \boldsymbol{\beta}_2$ for $t > \tau$.

Covariates may be linked to lifetimes in a variety of other ways. For example, **additive hazards models** with

$$h(t|\mathbf{x}) = h_0(t;\alpha) + r(\mathbf{x};\boldsymbol{\beta}) \tag{9}$$

are in some circumstances more plausible than PH models. In another direction, so-called **proportional-odds regression** models [5, 25], in which

$$\log\left[\frac{S(t|\mathbf{x})}{1-S(t|\mathbf{x})}\right] = \log\left[\frac{S_0(t;\alpha)}{1-S_0(t;\alpha)}\right] + \alpha(\mathbf{x};\boldsymbol{\beta}),$$

are sometimes useful; there is an obvious connection with **logistic regression** models for binary responses, and such models are especially useful in situations where T is a discrete variable.

Let us now consider briefly time-varying covariates, in which case we write $\mathbf{x}(t)$ instead of $\mathbf{x}$. Situations in which time-varying covariates can arise include (i) clinical studies where the treatment assigned to an individual may change during the course of the study, (ii) observational studies in which time-dependent environmental variables or exposures affect individuals, and (iii) studies where covariates internal to an individual are prognostic with respect to survival; for example, CD4 lymphocyte counts for subjects infected with the human immunodeficiency virus (HIV) are prognostic for time to death (*see* **AIDS and HIV**). In addition, synthetic time-dependent covariates may be used to test a PH assumption [9].

A general treatment of time-varying covariates is delicate if we allow "internal" covariates; see [20, Chapter 5] or [1, Section III.5]. We restrict the discussion here to "external" covariates whose values are determined independently of individuals who are under study, and assume that an individual's hazard function at time t depends only on covariates $\mathbf{x}(t)$ whose value can be determined at time t.

Such covariates are readily incorporated into models via the hazard function. For PH models, (7), or additive hazards models, (8), for example, we merely replace $\mathbf{x}$ with $\mathbf{x}(t)$. The same thing can be done for AFT models by writing down the hazard function $h(t|\mathbf{x})$ that corresponds to (4), but the resulting models are less appealing than in the case of multiplicative or additive hazards.

Even when the hazard function depends in a simple way on time-varying covariates, calculation of survival probabilities is more involved than for fixed covariates, as is the comparison of survival distributions for individuals with different covariate values. In particular, if we condition on the external covariate history $\mathbf{x}^* = [\mathbf{x}(s), s \geq 0]$, then

$$\Pr(T \geq t|\mathbf{x}^*) = \exp\left\{-\int_0^t h[s|\mathbf{x}(s)]\,\mathrm{d}s\right\}. \quad (10)$$

This requires knowledge of all covariate values over $(0, t)$.

Other Models

Survival models that involve more elaborate structure are sometimes needed. For example, there may be several modes of death or failure, such as in carcinogenicity studies in which animals are determined at autopsy to have died from one of several causes (*see* **Tumor Incidence Experiments**). This is often referred to as a competing modes of death or **competing risks** problem; the observable data consist of a lifetime $T \geq 0$ and mode of death C, which is in some set $(1, \ldots, k)$. Models with covariates are conveniently expressed in terms of mode-specific hazard functions [23, Chapter 9].

$$h_j(t|\mathbf{x}) = \lim_{\Delta t \to 0} \frac{\Pr(T < t + \Delta t, C = j | T \geq t, \mathbf{x})}{\Delta t}, \quad j = 1, \ldots, k.$$

Parametric representations $h_j(t|\mathbf{x};\boldsymbol{\theta})$ similar to those used for the hazard functions of parametric lifetime distributions, discussed above, can be employed here.

Multivariate lifetime models in which m lifetimes $(T_{i1}, \ldots, T_{im})$ are associated with an individual i are sometimes needed; for example, T_{i1} and T_{i2} might represent the ages at death of identical twins [19]. It is also occasionally useful to associate individual latent failure times $T_{i1}, \ldots, T_{ik}$ with multiple modes of failure, as described in the preceding paragraph (*see* **Competing Risks**). Various parametric families have been proposed for **multivariate survival analysis**.

Discrete or continuous mixture models are also employed frequently. For example, discrete mixtures have been used in situations where a fraction $1 - p$ of individuals have an effectively infinite lifetime whereas the remaining fraction p have lifetimes that follow a distribution with survivor function $S_1(t;\boldsymbol{\theta})$. A main area of application is in connection with the survival times of cancer patients, when a fraction $1 - p$ of patients are cured (*see* **Cure Models**). The long-term survivors cannot be distinguished a

priori, so a randomly selected individual has survivor function

$$S(t;\boldsymbol{\theta}, p) = 1 - p + pS_1(t;\boldsymbol{\theta}).$$

Models in which both p and $S_1(t)$ depend upon covariates may be considered (e.g. [15]).

More general mixture models are obtained when individuals or groups ("clusters") of individuals have unobservable random effects associated with them. "Frailty" models, for example, assume that there is a positive-valued random effect α associated with an individual and that, conditional on α, the individual's lifetime distribution has cumulative hazard function $\alpha H(t)$. The random effect is assumed to have a cdf G, in which case the unconditional survivor function for T is

$$S(t) = \int_0^\infty \exp[-\alpha H(t)]\, \mathrm{d}G(\alpha), \qquad (11)$$

which is a Laplace transform. Multivariate lifetime models may similarly be obtained by assuming that there is a common random effect α associated with a group of lifetimes $T_1, \ldots, T_m$, which, given α, are independent (e.g. [11]). Different parametric choices for $H(t)$ and $G(\alpha)$ in (11) provide a wide variety of models.

Model Fitting and Inference

Model fitting involves estimation of the unknown parameters in the models on the basis of observed data. More generally, we may wish to estimate, or test hypotheses about, certain features of the model. Maximum likelihood methods are favored for most applications, and software is widely available for the most common lifetime models. Standard procedures are described below, followed by a discussion of model checking and an example.

Likelihood-based Estimation and Inference

Suppose that individuals $i = 1, \ldots, n$ have independent lifetimes T_i with density and survivor functions $f_i(t;\boldsymbol{\theta})$, and $S_i(t;\boldsymbol{\theta})$, respectively. The index i is used with f and S to indicate that they may depend upon covariates associated with individual i. Data on lifetimes are frequently right-censored, so we will assume that for some individuals ($i \in D$) the exact lifetime t_i is known, and for others ($i \in C$) only the fact that t_i exceeds the observed censoring time c_i is known. The likelihood function for $\boldsymbol{\theta}$ is based on the joint pdf of the observed data. Under the assumption that the censoring of an individual at a point in time cannot be related to future events, the likelihood function is proportional to

$$L(\boldsymbol{\theta}) = \prod_{i\in D} f_i(t_i;\boldsymbol{\theta}) \prod_{i\in C} S_i(c_i;\boldsymbol{\theta}). \qquad (12)$$

The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is obtained by maximizing $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, or equivalently, the log likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i\in D} \log f_i(t_i;\boldsymbol{\theta}) + \sum_{i\in C} \log S_i(c_i;\boldsymbol{\theta}). \qquad (13)$$

With the models considered here, $\ell(\boldsymbol{\theta})$ can typically be maximized by solving the maximum likelihood or "score" equations $\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$.

Maximum likelihood **large-sample theory** provides several ways of constructing tests or confidence intervals for model parameters [23, Appendix C]. The two which are most readily available in software are termed the Wald and **likelihood ratio test** procedures, and are described briefly.

Suppose the $p \times 1$ parameter vector $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta}' = (\boldsymbol{\phi}', \boldsymbol{\lambda}')$, where the $r \times 1 (r \le p)$ parameter $\boldsymbol{\phi}$ is the parameter of interest and $\boldsymbol{\lambda}$ is a **nuisance parameter**. The $p \times p$ observed **information matrix** $\mathbf{I}(\boldsymbol{\theta})$ and its inverse,

$$\mathbf{I}(\boldsymbol{\theta}) = -\frac{\partial^2\ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{pmatrix} \mathbf{I}_{\phi\phi} & \mathbf{I}_{\phi\lambda} \\ \mathbf{I}_{\lambda\phi} & \mathbf{I}_{\lambda\lambda} \end{pmatrix},$$

$$\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{pmatrix} \mathbf{V}_{\phi\phi} & \mathbf{V}_{\phi\lambda} \\ \mathbf{V}_{\lambda\phi} & \mathbf{V}_{\lambda\lambda} \end{pmatrix},$$

play a key role in the Wald method. Under the hypothesis that $\boldsymbol{\phi} = \boldsymbol{\phi}_0$, the quantity $\hat{\mathbf{V}}_{\phi\phi}^{-1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)$ is approximately standard r-variate normal in large samples, where $\hat{\mathbf{V}}$ stands for $\mathbf{V}(\hat{\boldsymbol{\theta}})$. Equivalently, the Wald statistic (*see* **Likelihood**),

$$W_1(\boldsymbol{\phi}_0) = (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)^{\mathrm{T}} \hat{\mathbf{V}}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1} (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0), \qquad (14)$$

is approximately distributed as the **chi-square distribution** with r **degrees of freedom**, $\chi^2(r)$. The statistic (14) can be used to test H: $\boldsymbol{\phi} = \boldsymbol{\phi}_0$; large values of $W_1(\boldsymbol{\phi}_0)$ provide evidence against the hypothesis. **Confidence** regions with approximate confidence

coefficient q are obtained as the set of parameter values $\boldsymbol{\phi}$ that satisfy $W_1(\boldsymbol{\phi}) \leq \chi_q^2(r)$, where $\chi_q^2(r)$ is the qth **quantile** of $\chi^2(r)$.

The likelihood ratio method utilizes the statistic

$$W_2(\boldsymbol{\phi}_0) = 2\ell(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) - 2\ell[\boldsymbol{\phi}_0, \hat{\boldsymbol{\lambda}}(\boldsymbol{\phi}_0)], \qquad (15)$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\phi}}', \hat{\boldsymbol{\lambda}}')'$; and $\hat{\boldsymbol{\lambda}}(\boldsymbol{\phi}_0)$ is the value of $\boldsymbol{\lambda}$ that maximizes $\ell(\boldsymbol{\phi}_0, \boldsymbol{\lambda})$. Under H: $\boldsymbol{\phi} = \boldsymbol{\phi}_0$, $W_2(\boldsymbol{\phi}_0)$ is approximately $\chi^2(r)$ in large samples, and tests and confidence regions are obtained in the same way as with the Wald statistic.

The methods based on (14) and (15) produce two-sided confidence regions consisting in the case $r = 1$ of parameter values on either side of $\hat{\phi}$. If one-sided intervals are wanted, they may be based on sign $(\hat{\phi} - \phi_0)W_1(\phi_0)^{1/2}$ or sign $(\hat{\phi} - \phi_0)W_2(\phi_0)^{1/2}$, assumed to be approximately standard normal.

The asymptotic theory upon which these methods rely requires that the models satisfy some mild regularity conditions. Of the models described earlier, the only cases which are problematic concern threshold parameters. For censored data a "large" sample is essentially one for which n is large and not too high a proportion of lifetimes is censored. For practical purposes what we desire is that significance levels (*see* ***P*** **Value**) or confidence coefficients calculated from the asymptotic χ^2 or normal distributions for W_1 or W_2 are sufficiently close to their true values. If there is doubt about the adequacy of the approximations, they may be checked by **simulation**. A warning about accuracy is signaled by any large differences in confidence intervals or significance levels based on $W_1(\boldsymbol{\phi}_0)$ and $W_2(\boldsymbol{\phi}_0)$; in such cases intervals based on $W_2(\boldsymbol{\phi}_0)$ are likely to be more accurate. Analytic adjustments to improve accuracy have been developed (e.g. [4]), but are not yet widely available in software. Parametric **bootstrap** procedures [13] provide another possibility for improving accuracy.

Most survival analysis software relies solely on Wald or likelihood ratio procedures. Major packages with parametric survival analysis capabilities include SAS (see PROC LIFEREG and PROC RELIABILITY), SAS JMP, **S-PLUS** (see especially the function Censor Reg), SYSTAT (see the Survival module), and LIMDEP. All handle estimation for accelerated failure time regression models based on the Weibull, lognormal, and log-logistic distributions. In addition to right censoring, they handle interval censoring: in this case the ith lifetime t_i is known to lie in an interval (a_i, b_i), yielding the likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [S_i(a_i; \boldsymbol{\theta}) - S_i(b_i; \boldsymbol{\theta})].$$

The likelihood (12) is a special case of this.

More complicated parametric models involving multivariate lifetimes, competing failure modes, or random effects may also be handled by maximum likelihood. Although software is not widely available for specific models, general purpose **optimization** software allows one to deal quite easily with most situations. Problems that cause difficulty tend to be ones in which the observed data are uninformative about certain aspects of the model, thus leading to flat regions in the likelihood function (see, for example, [15]).

Model Checking

Parametric survival models assume a specific form for lifetime distributions and, in addition, the dependence upon any covariates must be specified. At preliminary stages of analysis it is useful to group individuals so that within groups they have similar values of important covariates. Nonparametric (**Kaplan–Meier**) estimates $\hat{S}(t)$ of the survivor function for each group may be compared graphically to detect prominent features of the data, to assess the shape of the lifetime distributions, and to see whether a fairly simple regression model will suffice. A particularly useful procedure is to plot $\log[-\log \hat{S}(t)]$ (vertical axis) vs. $\log t$ (horizontal axis) for each group. If a PH model is reasonable, then the curves should be roughly parallel in the vertical direction [see (8)]; if the curves are roughly parallel in the horizontal direction, an AFT model is suggested. If Weibull distributions are reasonable, then the curves should be roughly linear [see (3)]; differences in slope indicate a shape parameter that depends upon x. This is often termed a Weibull probability plot and is an example of probability or hazard plotting, which can also be used to explore other models.

Graphical tools such as scatter plots or box plots are also valuable (*see* **Graphical Displays**), but must be modified to deal with censored observations. For box plots the empirical quantiles for a group of individuals may be determined from the Kaplan–Meier estimate, provided they are not beyond the largest observation. For plots that involve the lifetimes

directly it is important to plot lifetimes and censoring times with different symbols.

Probability plots of empirical survivor functions can provide checks on parametric models, as suggested above. We may also compare fitted parametric models $S_0(t;\hat{\boldsymbol{\theta}})$ for baseline distributions, or models without covariates, with nonparametric estimates. Formal **goodness-of-fit** tests based on this idea exist for certain parametric models when there are strict limitations on censoring and the presence of covariates (e.g. [23, Chapter 10]), and in some cases for more complex situations [1, Sections VI.3 and VII.6]. However, most regression model checking is based on model expansion or on graphical assessment of **residuals**.

Model expansion involves fitting models with additional parameters that represent specific types of departures from the current model; the need for the extra parameters may be assessed via likelihood ratio or Wald tests. Score tests for which only the current model has to be fitted are also useful on occasion (e.g. [23, Section 10.2.2]). Some examples of model expansion are: (i) adding covariates representing **interactions** or other effects, as a check on a specified regression model; (ii) allowing σ in a location-scale model (4) to depend on $\mathbf{x}$, as a check on the AFT assumption; and (iii) using the Burr–Pareto distribution (2) in order to test the assumption of a baseline Weibull distribution ($\alpha = \infty$).

Location-scale models (4) are the most widely used parametric regression models, and for them residuals are naturally defined as

$$\hat{z}_i = \frac{y_i^* - \mu(\mathbf{x}_i;\hat{\boldsymbol{\beta}})}{\hat{\sigma}}, \tag{16}$$

where $y_i^* = \min(y_i, \log c_i)$ is either a log lifetime or log censoring time, depending on what was observed. If the model is appropriate, the $\hat{z}_i$s should look roughly like a censored random sample from the distribution with survivor function G. Probability or hazard plots of the $\hat{z}_i$s may be used to assess the baseline distribution G, and plots of the $\hat{z}_i$s vs. covariates or other factors can be used to check on the constancy of σ or to look for systematic departures from the assumed specification $\mu(\mathbf{x};\boldsymbol{\beta})$. Departures from the location-scale form itself are harder to detect and are best examined by model expansion or the graphical methods mentioned at the beginning of this subsection. Plots should designate censored and uncensored residuals with different symbols. Sometimes it may be useful to adjust censored $\hat{z}_i$s upwards to give imputed uncensored residuals; a way of doing this is described below.

In general, residuals for an arbitrary parametric survival model may be defined as

$$\hat{r}_i = H(t_i^*|\mathbf{x}_i;\hat{\boldsymbol{\theta}}), \tag{17}$$

where $t_i^* = \min(t_i, c_i)$ and $H(t|\mathbf{x}) = -\log S(t|\mathbf{x})$ is the cumulative hazard function. The motivation for (17) is that the variables $H(T_i|\mathbf{x}_i;\boldsymbol{\theta}_0)$ have a standard exponential distribution if the model is correct and $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. If the model is appropriate, the $\hat{r}_i s, i = 1, \ldots, n$, should look roughly like a censored sample of exponential variates. These residuals can be plotted in ways described above to check on the model. It will be noted that the AFT residuals, (16), are related to the residuals (17) by $\hat{r}_i = -\log G(\hat{z}_i)$. Residuals based on score or log likelihood contributions by each individual are also used.

Sometimes censored residuals $\hat{r}_i$ are adjusted upwards to give imputed values for corresponding uncensored residuals. This is usually done by adding either 1 or log 2 to r_is for $i \in C$. The motivation for doing this is that if r_i is a standard exponential random variable then $\mathrm{E}(r_i|r_i > r) = r + 1$ and $\Pr(r_i > r + \log 2|r_i > r) = 0.5$. If such imputed values are shown on plots, then it is also advisable to show the censored residuals from which these were calculated.

Conclusions about explanatory variables are generally not affected much by mild misspecification of the baseline distribution in an AFT or PH model, or by mild departures from the AFT or PH frameworks themselves. Moreover, unless there is a considerable amount of data it is difficult to discriminate among parametric distributions with rather similar shapes, such as the Weibull, log-logistic, and log-normal. One should, however, check carefully for significant departures from assumed models. Observations that are unusual or highly influential should also be scrutinized; in some cases it is a good idea to refit models with certain observations omitted. The fitting of expanded models is encouraged both as a **model checking** device and as a way to assess the sensitivity of inferences to variations in the model that are plausible on the basis of the observed data.

An Example

As a brief example we consider data on survival times for leukemia patients given by Feigl & Zelen [17] in an early paper on regression analysis of lifetime data. Survival times are in weeks from diagnosis, and there are two covariates: white blood-cell count (WBC) at diagnosis and a binary covariate AG that indicates a positive or negative (positive $= 1$, negative $= 0$) test related to white blood-cell characteristics. The data are given in Table 1, where *wbc* denotes white blood-cell count divided by 1000. The original data had no censored lifetimes, but for illustrative purposes three of the lifetimes have here been replaced with censoring times.

Exploratory analysis suggests an AFT regression model in which covariates $x_1 = AG$ and $x_2 = \log(wbc)$ are included. Figure 1 shows a plot of log (survival time) vs. log (wbc), with AG positive- and negative-valued observations denoted by P and N, respectively. Small letter ps denote the three censoring times. There are 17 AG-positive and 16 AG-negative subjects; two of the AG-positive subjects have $wbc = 100$ and $t = 1$, and their symbols are overlaid in the figure. The plot suggests that lifetimes tend to be shorter for subjects with higher WBC and longer for AG-positive subjects.

Accelerated failure time models (5) with extreme value and logistic error distributions were fitted, corresponding to Weibull and log-logistic lifetime distributions. Figure 1 suggests possibly different effects of WBC for subjects with $AG = 1$ and $AG = 0$, so models with and without a log $(wbc) - AG$ interaction term were fitted in each case. The two distributions gave very similar results and in neither case was the interaction term significant. The maximum log-likelihood values $\ell(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma})$ for the models without the interaction terms (i.e. with $\mu(x)$ in (5) given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2$) were -52.9 (Weibull) and -53.0 (log-logistic), and so provide no evidence to favor one distribution over the other. Model checks described below also showed the two models to be comparable. For convenience we consider only the Weibull model in the remainder of the discussion.

Table 1 Leukemia survival data

Time	*AG*	*wbc*	Time	*AG*	*wbc*
65	1	2.3	56	0	4.4
140[a]	1	0.75	65	0	3.0
100	1	4.3	17	0	4.0
134	1	2.6	7	0	1.5
16	1	6.0	16	0	9.0
106[a]	1	10.5	22	0	5.3
121	1	10.0	3	0	10.0
4	1	17.0	4	0	19.0
39	1	5.4	2	0	27.0
121[a]	1	7.0	3	0	28.0
56	1	9.4	8	0	31.0
26	1	32.0	4	0	26.0
22	1	35.0	3	0	21.0
1	1	100.0	30	0	79.0
1	1	100.0	4	0	100.0
5	1	52.0	43	0	100.0
65	1	100.0			

[a] Denotes a censoring time; *wbc* = WBC ÷ 1000.

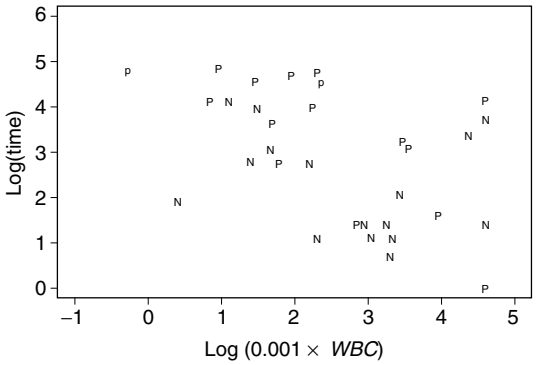

Figure 1 Leukemia data: log lifetime vs. log *wbc*

For the Weibull model the estimates (with standard errors obtained from the observed information matrix given in parentheses) are

$$\hat{\beta}_0 = 3.841(0.534), \qquad \hat{\beta}_1 = 1.177(0.427),$$

$$\hat{\beta}_2 = -0.366(0.150), \quad \hat{\sigma} = 1.119(0.164).$$

There is clearly no evidence against the hypothesis that $\sigma = 1$, suggesting that the exponential model could be used. The effects of WBC and AG are both significant, and are in the directions suggested by Figure 1.

Model checks may be based on residuals $\hat{z}_i$ of the form (16). We use the extreme value residuals with $\hat{\sigma} = 1.119$, but residuals based on the exponential model give essentially the same picture. Plots of the $\hat{z}_i$s against x_{1i}, x_{2i}, or fitted values $\mu(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ do not suggest any major problems with the model. Figure 2 shows a **diagnostic** probability plot designed to check on the assumed baseline extreme

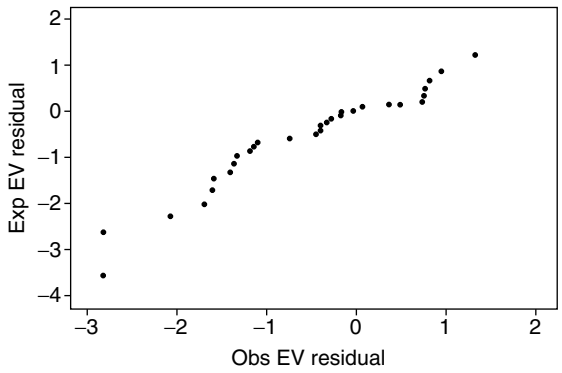

Figure 2 Leukemia data: probability plot of residuals

value distribution. This is obtained by treating the $\hat{z}_i$s as a censored sample of 33 log lifetimes, and computing the Kaplan–Meier estimate $\hat{S}(z)$ of the survivor function of Z based on them. Figure 2 is an extreme value probability plot of the points $\hat{z}_i$, $w_i = 0.5\hat{S}(\hat{z}_i - 0) + 0.5\hat{S}(\hat{z}_i + 0)$, based on plotting $\hat{z}_i$ vs. $\log(-\log w_i)$ for the uncensored residuals. The plot is roughly linear, and provides no evidence against the extreme value assumption.

One feature of the data is worth noting: for the AG-positive group there is an individual with a high WBC along with a reasonably large lifetime. For the AG-negative group there are two such individuals. It is clear from Figure 1 that these observations are very influential. If they were omitted, then the effects of WBC and AG would be increased substantially. We have no reason to isolate these observations and so they are retained, but their influence is noted.

Bibliographic Notes

Parametric models for lifetime data have been in existence for a long time, but received increasing attention from about 1950. The books by Lawless [23] and Cox & Oakes [10] reference many models. Early work on parametric methods for the regression analysis of survival data emphasized exponential, Weibull, and lognormal models, and may especially be found in the biostatistics and reliability literature. Papers with biostatistics applications include those by Sampford & Taylor [27], Feigl & Zelen [17], Pike [26], and Zippin & Armitage [29]. The focus for parametric analysis quickly became the accelerated failure time and location-scale models. The books by Kalbfleisch & Prentice [20, Chapter 3] and Lawless [23, Chapter 6] provide detailed treatments. Other good sources include Cox & Oakes [10], Andersen et al. [1], and Collett [8]. Some papers that study specific families of models include Farewell & Prentice [16], Bennett [5], Ciampi et al. [7], and Anderson [2]. Escobar & Meeker [14] consider diagnostics and influence analysis. Lawless [24] discusses general classes of lifetime regression models.

The proportional or multiplicative hazards model gained popularity quickly following the landmark paper by Cox [9]. The emphasis has been very much on semiparametric methods, however, and relatively little fully parametric inference for these models is found in the literature.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Anderson, K.M. (1991). A nonproportional hazards Weibull accelerated failure time regression model, *Biometrics* **47**, 281–288.

[3] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

[4] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[5] Bennett, S. (1983). Log-logistic regression models for survival data, *Applied Statistics* **32**, 165–171.

[6] Burr, I.W. (1942). Cumulative frequency distributions, *Annals of Mathematical Statistics* **13**, 215–232.

[7] Ciampi, A., Hogg, S.A. & Kates, L. (1986). Regression analysis of censored survival data with the generalized F family – an alternative to the proportional hazards model, *Statistics in Medicine* **5**, 85–96.

[8] Collett, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, London.

[9] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[10] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[11] Crowder, M.C. (1989). A multivariate distribution with Weibull connections, *Journal of the Royal Statistical Society, Series B* **47**, 447–452.

[12] Dubey, S.D. (1968). A compound Weibull distribution, *Naval Research Logistics Quarterly* **15**, 179–188.

[13] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

[14] Escobar, L. & Meeker, W.Q. (1992). Assessing influence in regression analysis with censored data, *Biometrics* **48**, 507–528.

[15] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* **38**, 1041–1046.

[16] Farewell, V.T. & Prentice, R.L. (1977). A study of distributional shape in life testing, *Technometrics* **19**, 69–76.

[17] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[18] Hjorth, U. (1980). A reliability distribution with increasing, decreasing, constant, and bathtub-shaped failure rates, *Technometrics* **22**, 99–108.

[19] Hougaard, P., Harvald, B. & Holm, N.V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881–1930, *Journal of the American Statistical Association* **87**, 17–24.

[20] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. Wiley, New York.

[21] Krall, J., Uthoff, V. & Harley, J. (1975). A step-up procedure for selecting variables associated with survival, *Biometrics* **31**, 49–57.

[22] Lachin, J.M. & Foulkes, M.A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification, *Biometrics* **42**, 507–519.

[23] Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data*, 2nd Ed.Wiley, New York.

[24] Lawless, J.F. (1986). A note on lifetime regression models, *Biometrika* **73**, 509–512.

[25] Pettitt, A.N. (1984). Proportional odds models for survival data and estimates using ranks, *Applied Statistics* **33**, 109–175.

[26] Pike, M.C. (1966). A method of analysis of a certain class of experiments in carcinogenesis, *Biometrics* **22**, 142–161.

[27] Sampford, M.R. & Taylor, J. (1959). Censored observations in randomized block experiments, *Journal of the Royal Statistical Society, Series B* **21**, 214–237.

[28] Smyth, G.K. (1989). Generalized linear models with varying dispersion, *Journal of the Royal Statistical Society, Series B* **51**, 47–60.

[29] Zippin, C. & Armitage, P. (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter, *Biometrics* **22**, 665–672.

J.F. LAWLESS

Parental Effects

A parental effect refers to a situation where, conditional on the individual's own genotype, the phenotype of an individual depends upon the mother's or father's phenotype or **genotype**. For example, a factor complicating **gene** discovery in asthma is the possibility, based on epidemiologic studies, that maternal phenotype influences the inheritance of asthma and atopy [10, 12]. The presence of asthma and associated phenotypes such as atopy in children has been consistently associated with an increased prevalence of asthma or atopy in mothers [10]. The differential risk of transmission between parents may be fourfold. The mechanism, or mechanisms, for these parental effects are unknown, but possibilities include genomic imprinting (see below) or maternal modification of the developing infant's immune system by transmission of immune factors across the placenta or through breast milk. The latter is likely to be affected by a complex **interaction** between maternal and fetal genetic and environmental factors. Similar parental effects have been noted in other immunologic disorders, most notably type I diabetes [21], rheumatoid arthritis [9], inflammatory bowel disease [1], and selective IgA deficiency [20], suggesting that parental effects on the developing infant's immune system may be an important common process modifying genetic diseases that are immunologic in origin. The remainder of this article describes the phenomena of genomic imprinting and maternal effects and their implications for genetic analysis. While there are clear theoretic differences between these two mechanisms, in practice they may be difficult to distinguish in **complex diseases**.

Genomic Imprinting

Imprinting refers to the situation where the relationship between a genotype and a phenotype (or disease-status) in an offspring depends upon which parent passed on the disease susceptibility or phenotype-modifying gene. When imprinting exists, the **penetrance** of the disease susceptibility allele will be different for maternally derived vs. paternally derived alleles. The observation that certain genes are expressed differently depending on whether they are inherited from the father or mother implies that genetic alteration of a gene or its expression has taken place. For example, a chromosomal deletion of a certain part of human chromosome 15 in a father results in an offspring with Prader–Willi syndrome. However, when the same part of chromosome 15 is missing in a mother, the offspring has Angelman syndrome [2] (*see* **Genetic Counseling**).

Concrete examples of genomic imprinting derive largely from studies of transgenic mice [16, 17]. However, imprinting has been suggested to play a role in several complex human diseases in addition to asthma, including bipolar affective disorder [4] and type 2 diabetes mellitus [8]. The mechanisms causing imprinting are poorly understood, but are thought to involve DNA methylation. The effect of imprinting can range from total inactivation of a gene and its expression (*see* **Gene Expression Analysis**) to the reduced expression in specific tissues. Interestingly, the imprinting effect can appear to be **heritable** only in a single generation. That is, the effect is unmasked if it passes through the nonimprinting sex. For instance, a gene inactivated by maternal imprinting that is inherited by a son will be reactivated in the next generation, i.e. the offspring of the son inheriting the gene. (Note, however, that should the son then have a daughter, her children will not express the phenotype.) Conversely, the same gene inherited by a daughter will remain inactivated in the next generation.

For a quantitative phenotype assessed in a nuclear family, the possible existence of imprinting upon offspring phenotype may be crudely assessed by estimating a basic variance component model (*see* **Variance Component Analysis**, Equation (2)) or its extensions. The basic variance component model is

$$Y_i = \mu_i + G_i + C_i + E_i, \qquad (1)$$

where Y_i is a continuous trait measured on individual i, μ is the conditional trait mean, and G_i, C_i and E_i are independent random variables with zero means and represent genetic factors, factors common to relatives (i.e. familial environmental factors), and factors specific to an individual (including measurement error, assumed to arise from nongenetic environmental factors), respectively (*see* **Familial Correlations**).

The result of imprinting at a locus on the expression of a quantitative phenotype will be to reduce the expected phenotypic covariance between parents and offspring relative to that between sibs (*see* **Genetic**

Correlations and Covariances). Gametic imprinting would be suggested if the difference between the parent–offspring covariance and twice the covariance between paternal half sibs derived from (1) were significantly less than zero. Imprinting can be explicitly assessed by extending (1):

$$Y_i = \mu_i + G_{if} + G_{im} + C_i + E_i, \qquad (2)$$

where the subscript f denotes components of genetic variance derived from the father, and the subscript m denotes components of genetic variance derived from the mother. Note that in an imprinting model, covariances between pairs of relatives are also estimated separately for male–male, female–female, and male–female relationships [3].

The phenomenon of imprinting has potentially important implications for genetic analysis. The reduction of the expected phenotypic covariance between parents and offspring relative to that between sibs caused by imprinting can lead to greatly reduced power to detect **linkage** for both quantitative and qualitative traits when such imprinting effects are not considered in the analysis. Similarly, **association** analyses are compromised when the differential risk for outcome among individuals with the same genotype (but alleles derived from different parent genders) is not considered. In an imprinting situation, genotypes of high risk are considered in the same "exposure" category as genotypes conferring no increased risk due to imprinting, biasing the **association** towards the null. For these reasons, inclusion of terms reflecting imprinting effects in models of quantitative and qualitative traits is important for **segregation analysis**, linkage analysis (*see* **Linkage Analysis, Model-free**; **Software for Genetic Epidemiology**) and association analysis (*see* **Disease-marker Association**; **Family-based Case–Control Studies**).

Maternal Effects

Maternal effects arise when, for reasons that may be environmental (e.g. *in utero* environmental effects or the effects of breast feeding), genetic, or a combination of the two, the phenotype of an offspring depends more upon maternal phenotype than on paternal phenotype. Such an observed effect may arise due to maternal genotype (any genetic effects on the mother's *in utero* environment or other nontransmitted maternal genotype effect) or maternal phenotype (any characteristic in the mother – possibly nongenetic – that influences the child's phenotype). For example, children of mothers with the genetic disorder phenylketonuria may develop mental retardation and small head size regardless of the child's genotype unless dietary intervention during pregnancy is pursued [5]. This disorder in the child is due to the maternal genotype rather than transmitted genes carried by the child.

The basic variance component model given in (1) (*see* **Genetic Correlations and Covariances**; **Variance Component Analysis**) can be extended to allow for maternal effects [25] by splitting each of the original components of variance into two components, representing the direct expression of each individual's genotype and environmental components as well as indirect maternal components of variance:

$$Y_i = \mu_i + (G_{oi} + C_{oi} + E_{oi}) + (G_{mi} + C_{mi} + E_{mi}), \qquad (3)$$

where the subscript o denotes components of variance reflecting a direct effect of an individual's genotype and environmental exposures, and the subscript m denotes components of variance reflecting an indirect effect of the maternal phenotype. Note that this formulation explicitly ignores epistatic sources of maternal genetic variation.

As with imprinting effects, inclusion of terms reflecting maternal effects in variance components models of quantitative traits can be readily extended to segregation analysis and variance-component-based linkage analysis.

Failure to account for maternal effects (when present) in variance component and segregation analysis can result in misleading inferences regarding mode of phenotypic inheritance. This is also true for linkage and association analysis of qualitative traits, where failure to account for potential maternal or paternal effects can **bias** tests of the **null hypothesis** regarding linkage or association to a particular genetic variant possessed by affected individuals [22]. For example, in the **transmission-disequilibrium test** (TDT) setting, maternal effects that increase risk for a disease among children (regardless of the child's genotype) will result in an excess of affected child–mother pairs, compared with

affected child–father pairs. This may lead to a false conclusion of a maternal imprinting genetic effect, when no such genetic effect (in the children) exists.

Paternal effects, while far less common than maternal effects, can be dealt with analytically in the same way as maternal effects.

Linkage Analysis

Methods to test and account for imprinting when evaluating linkage have been proposed for model-based linkage analysis, model-free allele-sharing methods and variance component methods. Model-based linkage analysis can be performed by specifying male and female recombination fractions separately [15], or by fixing the recombination fraction of the assumed imprinting gender at 0.5 and then estimating the recombination fraction of the other gender [7]. Alternatively, a four-penetrance model can be created in which the disease locus **heterozygotes** have different penetrances depending on the parental origin of the particular alleles [18]. Because imprinting can occur for dominant and recessive modes of inheritance, maximized lod score (mod score) analysis can also be pursued under this four-penetrance model [14, 18]. Tests of linkage can be carried out via **likelihood ratio testing** of the four-penetrance model under linkage vs. the same model under no linkage. Explicit tests of imprinting can be carried out by comparing the four-penetrance model under linkage with a standard three-penetrance model, which assumes the two types of heterozygotes have equal penetrance. The GENEHUNTER-IMPRINTING software can perform parametric lod score linkage analysis under the four-penetrance model (*see* **Software for Genetic Epidemiology**).

Imprinting can be detected in model-free linkage as differential results when stratifying family sets according to paternal and maternal meioses [13]. For quantitative traits, **marker** allele sharing can be estimated for maternally derived and paternally derived alleles separately, and variance components or Haseman–Elston regression used to assess linkage as departure from the expected 25% sharing of a particular parental allele [6]. A similar method for qualitative traits has also been described [11]. Under the variance component framework, linkage in the presence of imprinting can be assessed by comparing the likelihood of the data using the **maximum likelihood** estimators (MLEs) of the parent-specific major-gene variance components with the likelihood obtained when constraining these to 0. Imprinting can be tested in a manner similar to that described for the model-based methods described above by comparing a likelihood where both parent-specific major gene variance components are estimated with a model where they are constrained to be equal (but not necessarily 0). A similar strategy can be employed for Haseman–Elston regression by estimating regression coefficients for paternal allele sharing and maternal allele sharing separately, and testing whether they are equal to 0 (test of linkage) or whether they are equal to each other (test of imprinting). These analyses can be carried out using available identity-by-descent (*see* **Identity Coefficients**) estimation software and standard statistical packages for variance components and linear regression. Testing for maternal or paternal effects not due to imprinting can be allowed for in linkage analyses by the inclusion of parental phenotype through conditioning, covariate adjustment or, in the case of a quantitative phenotype, inclusion of random effects representing the indirect parental components of individual phenotypic variance (as in (3) above).

Association Analysis

Association analyses that incorporate parental effects must include parental information. For parental genotype effects and imprinting effects, family-based association methods are needed. The most commonly used method of family-based association, the TDT, has been extended to allow for parent-of-origin effects [19, 22–24]. As mentioned above, failure to account for imprinting or parental effects can bias TDT results because decreased expression of phenotypes among children of the imprinting gender may dilute transmission effects. Simple comparisons of transmission estimates for father–child pairs vs. mother–child pairs can also lead to erroneous conclusions about imprinting and genetic effects, when nontransmitted parental effects influence the phenotype [22]. For these reasons, recent methods have proposed the use of conditional logistic [22] (*see* **Logistic Regression, Conditional**) or log-linear models [23], expressed in terms of direct genetic effects, imprinting effects, and additional (nontransmitted) parental effects. For

both approaches, trios can be characterized by the joint mother, father, and child genotypes, resulting in 15 distinct trio-types whose expected frequencies can be calculated theoretically according to the (child) genotype [genetic association], imprinting, and parental genotype (or environment) effect sizes. Logistic regression conditional on trio-type or log-linear regression of expected trio-type cell counts can be used to estimate these effect parameters and test for association using **likelihood ratio** methods. Such methods can be carried out in standard statistical software packages for conditional logistic regression and Poisson regression.

Conclusion

While the theory is well developed for parental effects, little empirical work in human genetics has focused on these issues. The identification of imprinted genes, methodologic development to assess and incorporate parental effects into genetic analysis, and further understanding of the imprinting mechanism all represent important challenges for **genetic epidemiology**.

References

[1] Akolkar, P.N., Gulwani-Akolkar, B., Heresbach, D., Lin, X.Y., Fisher, S., Katz, S. & Silver, J. (1997). Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease, *American Journal of Gastroenterology* **92**, 2241–2244.

[2] Cassidy, S.B., Dykens, E. & Williams, C.A. (2000). Prader-Willi and Angelman syndromes: sister imprinted disorders, *American Journal of Medical Genetics* **97**, 136–146.

[3] Eisen, E.J. & Legates, J.E. (1966). Genotype-sex interaction and the genetic correlation between the sexes for body weight in Mus musculus, *Genetics* **54**, 611–623.

[4] Grigoroiu-Serbanescu, M., Nothen, M., Propping, P., Poustka, F., Magureanu, S., Vasilescu, R., Marinescu, E. & Ardelean, V. (1995). Clinical evidence for genomic imprinting in bipolar I disorder, *Acta Psychiatrica Scandinavica* **92**, 365–370.

[5] Guttler, F. & Guldberg, P. (2000). Mutation analysis anticipates dietary requirements in phenylketonuria, *European Journal of Pediatrics* **159**, Supplement 2, S150–S153.

[6] Hanson, R.L., Kobes, S., Lindsay, R.S. & Knowler, W.C. (2001). Assessment of parent-of-origin effects in linkage analysis of quantitative traits, *American Journal of Human Genetics* **68**, 951–962.

[7] Heutink, P., van der Mey, A.G., Sandkuijl, L.A., van Gils, A.P., Bardoel, A., Breedveld, G.J. et al. (1992). A gene subject to genomic imprinting and responsible for hereditary paragangliomas maps to chromosome 11q23-qter, *Human Molecular Genetics* **1**, 7–10.

[8] Huxtable, S.J., Saker, P.J., Haddad, L., Walker, M., Frayling, T.M., Levy, J.C. et al. (2000). Analysis of parent-offspring trios provides evidence for linkage and association between the insulin gene and type 2 diabetes mediated exclusively through paternally transmitted class III variable number tandem repeat alleles, *Diabetes* **49**, 126–130.

[9] Koumantaki, Y., Giziaki, E., Linos, A., Kontomerkos, A., Kaklamanis, P., Vaiopoulos, G., Mandas, J. & Kaklamani, E. (1997). Family history as a risk factor for rheumatoid arthritis: a case–control study, *Journal of Rheumatology* **24**, 1522–1526.

[10] Moffatt, M.F. & Cookson, W.O. (1998). The genetics of asthma. Maternal effects in atopic disease, *Clinical and Experimental Allergy* **28**, Supplement 1, 56–61; discussion 65–66.

[11] Olson, J.M. & Elston, R.C. (1998). Using family history information to distinguish true and false positive model-free linkage results, *Genetic Epidemiology* **15**, 183–192.

[12] Palmer, L.J. & Cookson, W.O.C.M. (2000). Genomic approaches to understanding asthma, *Genome Research* **10**, 1280–1287.

[13] Paterson, A.D., Naimark, D.M. & Petronis, A. (1999). The analysis of parental origin of alleles may detect susceptibility loci for complex disorders, *Human Heredity* **49**, 197–204.

[14] Risch, N. (1984). Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes, *American Journal of Human Genetics* **36**, 363–386.

[15] Smalley, S.L. (1993). Sex-specific recombination frequencies: a consequence of imprinting? *American Journal of Human Genetics* **52**, 210–212.

[16] Solter, D. (1988). Differential imprinting and expression of maternal and paternal genomes, *Annual Review of Genetics* **22**, 127–146.

[17] Solter, D. (1998). Imprinting, *International Journal of Developmental Biology* **42**, 951–954.

[18] Strauch, K., Fimmers, R., Kurz, T., Deichmann, K.A., Wienker, T.F. & Baur, M.P. (2000). Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization, *American Journal of Human Genetics* **66**, 1945–1957.

[19] Umbach, D.M. & Weinberg, C.R. (2000). The use of case–parent triads to study joint effects of genotype and exposure, *American Journal of Human Genetics* **66**, 251–261.

[20] Vorechovsky, I., Webster, A.D., Plebani, A. & Hammarstrom, L. (1999). Genetic linkage of IgA deficiency to the major histocompatibility complex: evidence for allele segregation distortion, parent-of-origin penetrance differences, and the role of anti-IgA antibodies in disease

predisposition, *American Journal of Human Genetics* **64**, 1096–1109.

[21] Warram, J.H., Krolewski, A.S., Gottlieb, M.S. & Kahn, C.R. (1984). Differences in risk of insulin-dependent diabetes in offspring of diabetic mothers and diabetic fathers, *New England Journal of Medicine* **311**, 149–152.

[22] Weinberg, C.R. (1999). Methods for detection of parent-of-origin effects in genetic studies of case–parents triads, *American Journal of Human Genetics* **65**, 229–235.

[23] Weinberg, C.R., Wilcox, A.J. & Lie, R.T. (1998). A log-linear approach to case–parent triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting, *American Journal of Human Genetics* **62**, 969–978.

[24] Wilcox, A.J., Weinberg, C.R. & Lie, R.T. (1998). Distinguishing the effects of maternal and offspring genes through studies of "case–parent triads", *American Journal of Epidemiology* **148**, 893–901.

[25] Willham, R.L. (1972). The role of maternal effects in animal breeding. 3. Biometrical aspects of maternal effects in animals, *Journal of Animal Science* **35**, 1288–1293.

DANI FALLIN & LYLE J. PALMER

Pareto Distribution

The Pareto distribution comes in several different forms. Its original form, with cumulative distribution function

$$F_X(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\nu}, \quad \sigma, \nu > 0,\ x \geq \sigma, \tag{1}$$

was introduced by Vilfredo Pareto in [8] as a model for the distribution of income, and now commonly goes by the name of the classical Pareto distribution or *Pareto distribution of the first kind*. The parameter σ is the scale parameter, and ν is related to the expected value of X by the equation $\mu = \nu\sigma/(\nu - 1)$. Two other forms of this distribution were also proposed by Pareto. One of them, which is now known as the *Pareto distribution of the second kind*, takes the following form,

$$F_X(x) = 1 - \left(1 + \frac{x - \mu}{\sigma}\right)^{-\nu}, \quad x > \mu,\ \sigma, \nu > 0, \tag{2}$$

and is obtained from the classical Pareto distribution by a shift in location, as well as the addition of the location parameter μ. The other one, the *Pareto distribution of the third kind*, is given by

$$F_X(x) = 1 - \frac{C\exp(-bx)}{(x + C)^{\nu}}, \quad x > 0,\ C, \nu, b > 0. \tag{3}$$

The fourth form of the Pareto distribution

$$F_X(x) = 1 - \left[1 + \left(\frac{x - \mu}{\sigma}\right)^{1/\gamma}\right]^{-\nu}, \quad x > \mu,\ \alpha, \gamma, \sigma > 0, \tag{4}$$

is a direct generalization of (2), with the parameter γ providing additional flexibility, and has been defined in Arnold [1], which is a general reference book for this distribution; see also Arnold [2], and Johnson et al. [6].

For the classical Pareto distribution, the **maximum likelihood** estimators (MLEs) of σ and ν are given by

$$\hat{\sigma} = X_{1:n}, \qquad \hat{\nu} = n\left[\sum_{j=1}^{n} \ln\left(\frac{X_j}{\hat{\sigma}}\right)\right], \tag{5}$$

where $X_{1:n}$ is the smallest **order statistic**, from which one may obtain the following **unbiased** estimators of σ and ν,

$$\sigma^* = \left[1 - \frac{1}{(n-1)\hat{\nu}}\right]\hat{\sigma}, \qquad \nu^* = \left(1 - \frac{2}{n}\right)\hat{\nu}. \tag{6}$$

If ν is known, then the MLE of σ is $\hat{\sigma} = X_{1:n}$, while the best linear unbiased estimator (BLUE) is given by

$$\sigma^* = \frac{\nu n - 1}{\nu n\sigma}\hat{\sigma}. \tag{7}$$

If σ is known then the MLE of ν is given by $\hat{\nu}$ above with $\hat{\sigma}$ replaced by σ.

Parameter **estimation** for the Pareto distribution in (2) is discussed in [7], where exact explicit expressions for the BLUEs of μ and σ are given for the case in which ν is known.

For the situation in which a location parameter μ (*see* **Location–Scale Family**) is added to the classical Pareto distribution, exact explicit expressions for the BLUEs of μ and σ have recently been derived in [3], for complete as well as right-**censored** samples. The **robustness** of these BLUEs is also examined in [3], where recommendations are made as to how to protect against the situation in which the strict distributional assumptions are not completely satisfied.

Many of the applications of the distribution in (2) (with $\mu = 0$) stem from the fact that it is the mixture (or compound) of **exponential** distributions, $f_{X|\Theta}(x|\theta) = \theta\exp(-\theta x)$, where Θ has a **gamma** distribution, $f_{\Theta}(\theta) = [\sigma/\Sigma(\nu)](\sigma\theta)^{\nu-1}\exp(-\sigma\theta)$. This is used, for example, in the analysis of heart transplant data in [9], where X represents the survival time of a patient entering the study if no heart transplant were received, and the hazard rate θ varies from patient to patient and is assumed to follow a gamma distribution. It is also used in the study of remission rates of psychiatric patients in [4], where X represents the time to rehospitalization for psychiatric reasons, and the differing hazard rates among patients are reflected by different values of θ, which are assumed to follow a gamma distribution.

The multivariate form of (2), with survival function (*see* **Survival Distributions and Their Characteristics**),

$$S_X(\mathbf{x}) = \Pr\left[\bigcap_{i=1}^{k}(X_i \geq x_i)\right] = \left[1 + \sum_{i=1}^{k}\left(\frac{x_i - \mu_i}{\sigma_i}\right)\right]^{-\nu}, \qquad (8)$$

has been used (with $k = 2$) in [5] to model the injuries to two drivers in a road accident, where X_i is a measure of the severity of the injuries to driver $i, i = 1, 2$. In the same paper, (8) has been used to model the severity of a lesion in a patient as assessed by two physicians, and is discussed as a model for the liability of two people who are related to each other of acquiring a certain disease (*see* **Frailty**). In the former situation, X_i is a measure of the severity of the lesion as assessed by doctor $i, i = 1, 2$, and in the latter case, X_i is a measure of how likely each person is to acquire the disease.

References

[1] Arnold, B.C. (1983). *Pareto Distributions*. International Cooperative Publishing House, Fairland.

[2] Arnold, B.C. (1985). Pareto distribution, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, New York, pp. 568–574.

[3] Childs, A. & Balakrishnan, N. (1998). Generalized recurrence relations for moments of order statistics from non-identical Pareto and truncated Pareto random variables with applications to robustness, in *Handbook of Statistics – Order Statistics and Their Applications*, C.R. Rao & N. Balakrishnan, eds. Elsevier Science, North-Holland, to appear.

[4] Fox, P.D. & Kraemer, H.C. (1971). A probability model for the remission rate of discharged psychiatric patients, *Management Science* **17B**, 694–699.

[5] Hutchinson, T.P. (1979). Four applications of a bivariate Pareto distribution, *Biometrical Journal* **21**, 553–563.

[6] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd Ed. Wiley, New York.

[7] Kulldorff, G. & Vännman, K., (1973). Estimation of the location and scale parameters of a Pareto distribution by linear functions of order statistics, *Journal of the American Statistical Association* **68**, 218–227.

[8] Pareto, V. (1897). *Cours d'Economie Politique*. Rouge et Cie, Paris.

[9] Turnbull, B.W., Brown, B.W., Jr & Hu, M. (1974). Survivorship analysis of heart transplant data, *Journal of the American Statistical Association* **69**, 74–80.

(*See also* **Bivariate Distributions**; **Minimum Variance Unbiased (MVU) Estimator**; **Multivariate Distributions, Overview**; **Outliers**; **Random Variable**)

A. CHILDS & N. BALAKRISHNAN

Parsimony

In the development of statistical models, parsimony is often viewed as desirable or, even, as embodying a logical principle. In application, this requires that models be based on as few assumptions as possible and include as few parameters as possible. For example, Box & Jenkins [1] advocated the choice of **time series** models with the smallest number of parameters which gives adequate representation, and **Mallow's C_p** statistic [2], which is used to choose between alternative regression models, incorporates a penalty for additional parameters.

Such thinking is linked to the more general philosophical principle of Occam's Razor, "A plurality (of reasons) should not be posited without necessity", which was named after William of Occam (1280–1349).

References

[1] Box, G.E.P. & Jenkins, G.M. (1970). *Times Series Analysis Forecasting and Control*. Holden Day, San Francisco.

[2] Mallows, C. (1973). Some comments on C_p, *Technometrics* **15**, 661–675.

VERN T. FAREWELL

Partial Likelihood

Somewhat surprisingly, the **robustness** of the **likelihood** function extends to various approximate likelihoods used in models where likelihoods may be difficult to calculate, or that may not be characterized by a complete parametric specification. The partial likelihood is one such approximate likelihood.

Partial likelihoods seem to have been used formally for the first time in D.R. Cox's introduction of the semiparametric **proportional hazards** model (*see* **Cox Regression Model**) [1], and were discussed as a separate topic in a later paper [2]. The partial likelihood is based on a simple **conditional probability** argument. Suppose that $(A_1, B_1), (A_2, B_2), \ldots, (A_K, B_K)$ is a collection of pairs of observations. If the probability of these outcomes depends on a parameter θ, then the likelihood for θ based on the $2K$ observations can be written:

$$\begin{aligned}
&\Pr(A_K B_K A_{K-1} B_{K-1} \ldots, A_1 B_1) \\
&\quad = \left[\prod_{k=2}^{K} \Pr(A_k B_k | A_{k-1} B_{k-1} \ldots A_1 B_1)\right] \Pr(A_1 B_1) \\
&\quad = \left[\prod_{k=2}^{K} \Pr(A_k | B_k A_{k-1} B_{k-1} \ldots A_1 B_1)\right] \Pr(A_1 | B_1) \\
&\qquad \times \left[\prod_{k=2}^{K} \Pr(B_k | A_{k-1} B_{k-1} \ldots A_1 B_1)\right] \Pr(B_1).
\end{aligned}$$

In general, all four terms in this product may depend on θ; the first two terms comprise the partial likelihood for θ based on the As in the sequence (A_k, B_k).

The loss of information about θ caused from dropping the last two terms will depend on how strongly the conditional distribution of B_ks, given the previous events in the sequence, depends on θ. Balanced against the loss of **efficiency** is the possibility that the partial likelihood may be simpler to compute or may not depend on highly dimensional **nuisance parameters**. The partial likelihood for the proportional hazards regression model does not include an infinite dimensional nuisance parameter that is part of the full likelihood.

Consider a regression problem with right-**censored data** with no tied values in the set of observation times. Set $X_j = \min(T_j, U_j)$, $\delta_j = I_{X_j = T_j}$, and denote the observed data by (X_j, δ_j, Z_j), $j = 1, \ldots, n$, where $\mathbf{Z}_j$ denotes a vector of covariates for the jth case. Let $T_1^{\mathrm{o}} < \cdots < T_L^{\mathrm{o}}$ denote the L ordered times of observed failures, and set $T_0^{\mathrm{o}} \equiv 0$ and $T_{L+1}^{\mathrm{o}} \equiv \infty$. Let (k) be the case label for the patient failing at T_k^{o}. Let B_k be the event describing the observed times of censoring in the interval $[T_{k-1}^{\mathrm{o}}, T_k^{\mathrm{o}}]$, the case labels associated with the censored times, and the fact that a failure has been observed at T_k^{o}. Finally, let A_k specify the label, (k), of the case failing at T_k^{o}. The likelihood of the observed data is $\Pr(B_1 A_1 \ldots B_L A_L B_{L+1})$.

In the proportional hazards regression model the conditional **hazard** for T_j given $\mathbf{Z}_j$ is $\lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_j)$. If the conditional distributions B_k contain little information about the regression parameter $\boldsymbol{\beta}$, then a reasonable partial likelihood for $\boldsymbol{\beta}$ would be

$$\left[\prod_{k=2}^{L} \Pr(A_k | B_k A_{k-1} B_{k-1} \ldots A_1 B_1)\right] \Pr(A_1 | B_1).$$

For fixed k, since there are no ties in the observation times, it can be established that

$$\begin{aligned}
\Pr(A_k | B_k A_{k-1} B_{k-1} \ldots A_1 B_1) &= \frac{\lambda(T_k^{\mathrm{o}} | \mathbf{Z}_{(k)})}{\sum_{j \in R_k} \lambda(T_k^{\mathrm{o}} | \mathbf{Z}_j)} \\
&= \frac{\exp(\boldsymbol{\beta}' \mathbf{Z}_{(k)})}{\sum_{j \in R_k} \exp(\boldsymbol{\beta}' \mathbf{Z}_j)},
\end{aligned}$$

where R_k is the set of cases at risk at time T_k^{o}, i.e. $\{j : X_j \geq T_k^{\mathrm{o}}\}$ (*see* **Risk Set**). The partial likelihood for $\boldsymbol{\beta}$ will thus be

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{L} \frac{\exp(\boldsymbol{\beta}' \mathbf{Z}_{(k)})}{\sum_{j \in R_k} \exp(\boldsymbol{\beta}' \mathbf{Z}_j)}.$$

The infinite dimensional nuisance parameter λ_0 does not appear in the partial likelihood.

The derivation of the partial likelihood for the proportional hazards model requires more care when there are **tied survival times** and **time-dependent covariates**; the details for these cases can be found in [1] and [4].

The general construction of a partial likelihood is very similar. Suppose a data vector $\mathbf{X}$ has

density $f_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector parameter $(\boldsymbol{\phi}, \boldsymbol{\beta})$, and that the primary inference question is about $\boldsymbol{\beta}$ in the presence of the nuisance parameter $\boldsymbol{\phi}$. If $\mathbf{X}$ can be transformed into a sequence $(V_1, W_1, V_2, W_2, \ldots, V_N, W_N)$, then the likelihood for $\boldsymbol{\theta}$ can be written as

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) &= f_{V_1,W_1,\ldots,V_N,W_N}(v_1, w_1, \ldots, v_N, w_N;\boldsymbol{\theta}) \\ &= \prod_{n=1}^{N} f_{W_n|V_1,W_1,\ldots,V_n}(w_n|v_1, w_1, \ldots, \\ &\quad v_{n-1}, w_{n-1}, v_n;\boldsymbol{\theta}) f_{V_n|V_1,W_1,\ldots,V_{n-1},W_{n-1}}(v_n|v_1,w_1,\ldots, \\ &\quad v_{n-1}, w_{n-1};\boldsymbol{\theta}) \\ &= \left[\prod_{n=1}^{N} f_{W_n|Q_n}(w_n|q_n;\boldsymbol{\theta})\right] \\ &\quad \times \left[\prod_{n=1}^{N} f_{V_n|P_n}(v_n|p_n;\boldsymbol{\theta})\right], \end{aligned} \tag{1}$$

where $P_1 = \{\emptyset\}$, $Q_1 = V_1$, and, for $n = 2, \ldots, N$,

$$P_n = (V_1, W_1, \ldots, V_{n-1}, W_{n-1})$$

and

$$Q_n = (V_1, W_1, \ldots, W_{n-1}, V_n).$$

When the first product on the right-hand side of (1) depends only on $\boldsymbol{\beta}$, Cox called this term the partial likelihood for $\boldsymbol{\beta}$ based on W in the sequence $(V_1, W_1, \ldots, V_N, W_N)$. Because of the way the conditioning events are used in the construction, a partial likelihood is not the likelihood of either the observable or derived data, so the usual properties of likelihood-based estimates do not hold automatically.

There are a number of results that are true for partial likelihoods generally. Cox [2] shows that asymptotic sampling distributions for maximum partial likelihood estimates follow from standard arguments about score statistics (*see* **Likelihood**) applied to the partial likelihood score. Wong [6] gives a more formal account, and shows how to calculate the **asymptotic relative efficiency** of a maximum partial likelihood estimate compared to an estimate from the full likelihood. Oakes [5] and Efron [3] illustrate asymptotic efficiency calculations for maximum partial likelihood estimators in the proportional hazards model, finding conditions under which the partial likelihood estimators are fully efficient.

Cox [2] cites a number of difficulties that the concept of partial likelihood presents, including the possibility that since the conditioning argument used to construct the partial likelihood is not necessarily unique, different partial likelihood estimates may be obtained from the same observations. The principal difficulty from the practitioner's perspective may be that there is no direct, routine method for constructing a partial likelihood in the presence of difficult likelihoods or highly dimensional nuisance parameters.

References

[1] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[2] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[3] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association* **72**, 557–575.

[4] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[5] Oakes, D. (1977). The asymptotic information in censored survival data, *Biometrika* **64**, 441–448.

[6] Wong, W.H. (1986). Theory of partial likelihood, *Annals of Statistics* **14**, 88–123.

(*See also* **Marginal Likelihood**; **Survival Analysis, Overview**)

DAVID HARRINGTON

Partially Balanced Incomplete Block Design

Partially balanced incomplete block designs (PBIBDs) are used in the testing of drugs [2] and in the design of **factorial experiments**.

As with other **incomplete block designs**, PBIBDs are defined on a set of v treatments. The set of v treatments is structured: for each treatment the set of the remaining $v-1$ treatments is partitioned into a set of subsets called *associate classes*. This partitioning is usually based on inherent attributes of the treatments. For example, consider a combination drug study (see [4]). Treatments having the same level of factor 1 are first associates, treatments having the same level of factor 2 are second associates and treatments which have neither factor 1 nor factor 2 at the same level are third associates.

We will define the concept of an association scheme and of a PBIBD and then describe some association schemes that have been used in the biosciences.

If a set of v treatments has an *association scheme with m associate classes* defined on it, then:

1. any two distinct treatments are ith associates for exactly one value of i, $1 \leq i \leq m$;
2. each treatment has exactly n_i ith associates, $1 \leq i \leq m$;
3. for any pair of ith associates, x and y, say, there are a fixed number of treatments which are both ith associates of x and hth associates of y, and this number is independent of the particular pair of ith associates chosen.

The *parameters of the association scheme* are v and $n_i, 1 \leq i \leq m$. Since each treatment is an associate of every other, $\sum_i n_i = v - 1$.

A design based on a set of v treatments with an m-associate class association scheme defined on them is a *partially balanced incomplete block design with m associate classes* (PBIBD(m)). There are b blocks of size k and each treatment is replicated r times. If treatments x and y are ith associates then there are λ_i blocks containing both x and y. The PBIBD determines the association scheme, but not conversely (see [8]).

The parameters of a PBIBD(m) are not independent of each other. Counting plots, we see that there are v treatments each replicated r times and there are b blocks each with k plots. Thus $vr = bk$. Counting pairs, we have that $\sum_i n_i \lambda_i = r(k-1)$.

For example, let the treatments be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and let the pairs of first associates be $\{1, 6\}$, $\{2, 7\}$, $\{3, 8\}$, $\{4, 9\}$ and $\{5, 10\}$. Any other pair of treatments is a pair of second associates. So $v = 10$, $n_1 = 1$, $n_2 = 8$. If the blocks of the designs are $\{1, 2, 3, 4, 5\}$, $\{1, 2, 3, 9, 10\}$, $\{1, 7, 8, 4, 5\}$, $\{1, 7, 8, 9, 10\}$, $\{6, 2, 8, 4, 10\}$, $\{6, 2, 8, 9, 5\}$, $\{6, 7, 3, 4, 10\}$, and $\{6, 7, 3, 9, 5\}$, then $b = 8$, $k = 5$, $r = 4$, $\lambda_1 = 0$, $\lambda_2 = 2$.

The most commonly used association schemes are the *factorial association schemes*. These include the group-divisible, with two classes, and the rectangular, with three. Both of these schemes have been generalized to have more classes.

For the *group-divisible association scheme* there is a set of $v = mn$ treatments which may be viewed as being divided into m subsets, (or *groups*, in the nonalgebraic sense) each of n treatments. Treatments in the same group are first associates; treatments in different groups are second associates. We see that $n_1 = n - 1$ and $n_2 = n(m-1)$. This association scheme arises naturally when there are two factors, the first with m levels and the second with n levels, and the second factor is nested in the first. For example, the first factor might be one of three types of support provided to new mothers and the second factor hospitals. Each hospital treats all new mothers in the same way and so hospitals are nested within type of support. The hospitals can be compared within type of support and the types of support can be compared across hospitals.

If instead the two factors are not nested but crossed, then the corresponding association scheme is the *rectangular association scheme*. This scheme has three associate classes and is most easily visualized by writing the treatments in an $m \times n$ array. Treatments in the same row are first associates (having the same level of factor 1), treatments in the same column are second associates (having the same level of factor 2) and treatments which have neither factor 1 nor factor 2 at the same level are third associates. Combination drug studies provide one example of this structure (see [4]).

Other schemes can be built up by repeated nesting and crossing or a mix of these.

For complete diallel cross experiments the *Latin-square-type (L_i-type) association scheme* is

appropriate. In this scheme there are n^2 treatments which are laid out in an $n \times n$ array. In an L_2-type scheme, a pair of treatments in the same row or the same column are first associates and a pair of treatments not in the same row or column are second associates. In an L_i-type scheme, we also need $i-2$ mutually **orthogonal Latin squares** of order n. Then a pair of treatments are first associates if they occur in the same row, or in the same column, or in a cell with the same symbol when any of the Latin squares is superimposed on the square array. We see that $n_1 = i(n-1)$ and $n_2 = (n-1)(n-i+1)$ for $i \leq n-1$. If $i = n$, then the scheme is a group-divisible one with n subsets and with the names of the first and second associate classes reversed. If $i = n+1$, then all treatments are first associates of each other and the PBIBD will be a **balanced incomplete block design**.

Lattice designs with $n-1$ replicates are examples of L_2-type designs. Lattice designs were introduced by Yates [9] for crop cultivar trials and have been extended by various authors, most recently Patterson & Williams [6] to α-designs. These designs have been shown to give slightly more precision than **randomized complete block designs** for immunosorbent assays [1].

For complete half-diallel cross experiments the triangular association scheme is appropriate. In the *triangular association scheme* there are $v = n(n-1)/2, n \geq 5$, treatments. Arrange these treatments in a symmetrical $n \times n$ array with the diagonal cells empty. First associates are those in the same row or column. Treatments not in the same row or column are second associates.

For a partial diallel cross the scheme proposed by Kempthorne & Curnow [5], and attributed to G.W. Brown, is equivalent to a cyclic association scheme. In the *cyclic association scheme* use the integers modulo v, usually written as $Z_v = \{0, 1, 2, \ldots, v-1\}$, as the treatment names. Partition the $v-1$ nonzero entries in Z_v by $\{1, 2, \ldots, v-1\} = D \cup E$, where $D = \{d_1, d_2, \ldots, d_{n_1}\}$. For each $d_i \in D$ we must have that $v - d_i \in D$. There are $n_1(n_1-1)$ differences of distinct elements of D. These differences must contain each element of D equally often and each element of E equally often. The first associates of treatment i are the treatments in $i + D = \{i + d_1, i + d_2, \ldots, i + d_{n_1}\}$, where the sum (modulo v) always lies between 0 and $v-1$.

Further details about PBIBDs may be found in [7] and [8]. Clatworthy [3] provides an extensive tabulation of the two-associate-class PBIBDs mentioned here.

References

[1] Bauske, E.M., Hewings, A.D., Kolb, F.L.& Carmer, S.G. (1994). Variability in enzyme-linked immunosorbent assays and control of experimental error by use of experimental designs, *Plant Diseases* **78**, 1206–1210.

[2] Bourne, D.W.A. (1997). A First Course in Pharmacokinetics and Biopharmaceutics, located at `http://gaps.cpb.uokhsc.edu/gaps/pkbio/pkbiof.html`.

[3] Clatworthy, W.H. (1973). *Tables of Two-Associate-Class Partially Balanced Designs*, National Bureau of Standards (US) Applied Mathematics Series No. 63. National Bureau of Standards, Washington.

[4] Hung, H.M.J. (1996). Global tests for combination drug studies in factorial trials, *Statistics in Medicine* **15**, 233–247.

[5] Kempthorne, O.& Curnow, R.N. (1961). The partial diallel cross, *Biometrics* **17**, 229–250.

[6] Patterson, H.D.& Williams, E.R. (1976). A new class of resolvable incomplete block designs, *Biometrika* **63**, 83–92.

[7] Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. Wiley, New York.

[8] Street, A.P.& Street, D.J. (1987). *Combinatorics of Experimental Design*. Clarendon Press, Oxford.

[9] Yates, F. (1940). Lattice squares, *Journal of Agricultural Science* **30**, 672–687.

D.J. STREET & A.P. STREET

Partner Study

Partner studies are epidemiologic investigations of transmission of infectious diseases in susceptible partners of known infected individuals. Widely used for studying transmission of sexually transmitted diseases (STD) such as Human Immunodeficiency Virus (HIV) (*see* **AIDS and HIV**), these studies are applicable in situations in which "partnerships" that link known susceptible and infectious individuals can be identified. Eligibility typically requires that exposure to infection in susceptible individuals is limited to a previously infected partner. This allows potential bias associated with uncertainty about the source of infection to be reduced or eliminated. Partner studies focus primarily on the influence of measured **covariates** on the probability of transmission, but can also be used to gain knowledge of key parameters such as the *infectivity* (transmission risk per unit of exposure). Statistical analyses of data from these studies provide a number of unique challenges, most of which arise from the incomplete nature of data describing exposure and infection (*see* **Infectious Disease Models**).

Study Designs and Data

Study designs for investigations of infectious disease transmission depend on linking infection outcomes in susceptible individuals to exposure to infectious individuals. Thus, factors that characterize exposure as well as those related to the nature of infectiousness and susceptibility are important considerations in design and analysis. Because of the difficulty in directly observing such factors, most studies are carried out in restricted settings where exposure, susceptibility, and infectiousness can be more clearly identified. Classical examples include observation of disease outbreaks in households and island communities, where exposure can be assumed to be limited to a group. The primary focus of statistical analyses for such studies has been on describing epidemic spread at the group level [2]. For sexually transmitted diseases (STD), transmission must occur via sexual contacts within partnerships. This suggests that sexual partnerships are an ideal experimental unit on which to study transmission, and motivates the notion of a partner study. Partner studies have recently become a popular tool for investigating HIV transmission, although similar designs have undoubtedly been used to study other STDs. Most studies focus on a particular mode of sexual contact in a select type of partnership (e.g. heterosexual transmission in monogamous couples), but other modes of contact (e.g. needle sharing among intravenous drug users) can be investigated as well. Because the majority of statistical work has focused on monogamous couples, most of the development here will be restricted to this setting.

A partner study consists of a sample of partnerships, each composed of an infected individual called the *index case*, and a susceptible individual referred to here as the *partner*. For a single such couple, complete data on transmission include the total number of contacts and time of occurrence of each, information about the degree of susceptibility of the partner and infectiousness of the index case at the time each contact is made, and indicators of whether or not each contact induced infection in the susceptible partner. In actual studies, the timing and outcome of individual contacts is unknown, and exposure histories and infection indicators can only be measured following an interval of contact. As a result, direct biological measures of susceptibility and infectiousness at the time transmission occurred are almost always unavailable. The vast majority of studies can be described as either **cohort**, with a sample of partnerships consisting of a known infected and susceptible individual followed over time and infections observed and exposure measured at intervals; **case–control**, in which infected and uninfected susceptible individuals and their index case partners are recruited and exposure histories and covariates are ascertained retrospectively; or **cross-sectional**, where only infection status and risk behaviours are collected, but no direct measures of exposure are available. Kim & Lagakos [8] and Jewell & Shiboski [4] provide more details about study designs.

The remainder of this entry will focus primarily on retrospective studies of sexually transmitted diseases which exhibit minimal latency periods (i.e. infected individuals become infectious almost immediately after getting infected) (*see* **Latent Period**) and for which no immunity or recovery is possible. HIV provides a good example [1]. This simplifies the presentation and best reflects the existing statistical work in this area. Such studies are generally characterized by incomplete observations of relevant exposure and infection data. In Figures 1 and 2 are

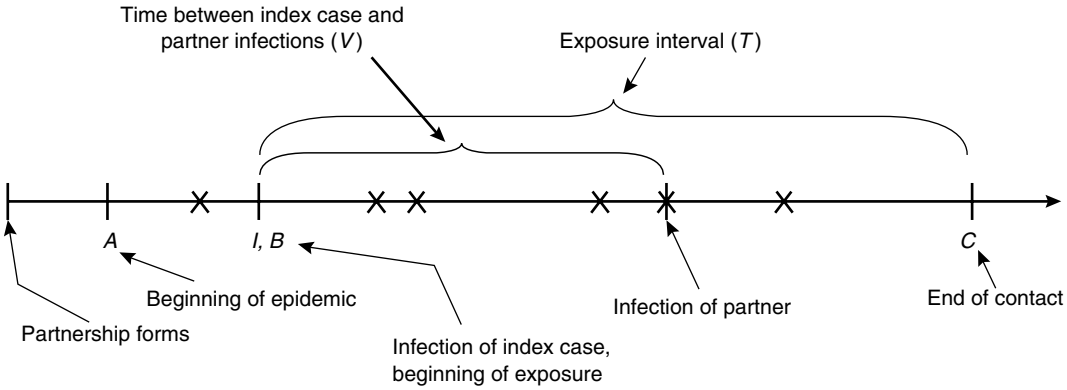

Figure 1 Exposure history from a long-term monogamous partnership formed prior to the earliest possible time of infection of the index case, A. The index case is infected at chronological time I, which also marks the beginning of exposure B. Exposure ends at C, where the infection status Y of the partner is first determined. The duration of exposure T and the time between the infections of the two partners V are also indicated. Contacts are indicated by the symbol "**X**"

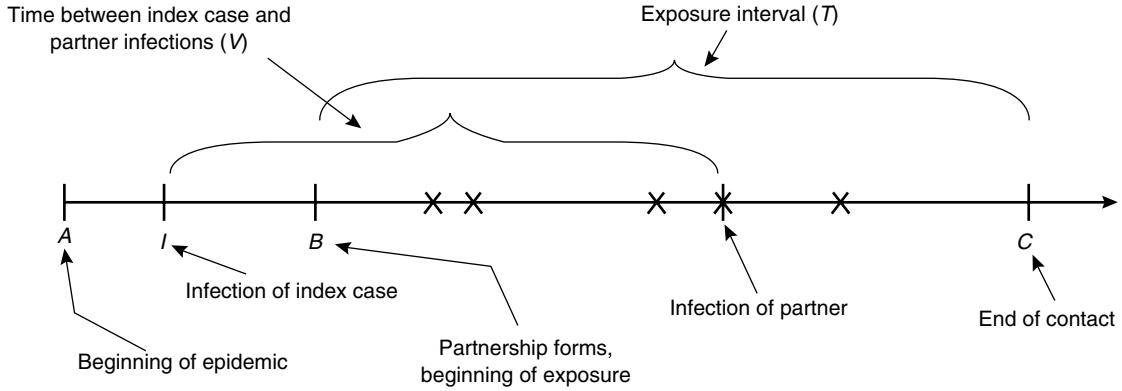

Figure 2 Exposure history from a short-term monogamous partnership, where I, the time of infection of the index case, is bounded below by A (see Figure 1), and contact (and hence exposure) begins at a subsequent time B. Contact ends at C, where the infection status Y of the partner is first determined. Contacts are indicated by the symbol "**X**"

provided schematic representations of exposure data commonly observed in retrospective studies. Both represent exposure histories for single partnerships. Let A denote a chronological time which provides a lower bound on the infection time of the index case, denoted I. This could represent the beginning of an epidemic, date of sexual debut, or other relevant information, and is useful in cases when the actual data of infection is unknown. Let B denote the chronological time when actual exposure to infection in the partner begins, C the time exposure ends or the partnership is recruited, and T the duration of exposure, $C - B$. In "long-term" partnerships which are formed prior to I (Figure 1), $B \equiv I$; for "short-term" partnerships which commence after the infection of the index case (Figure 2), $B > I$. In both situations, the time V between index case infection and partner infection along with the actual degree of exposure during the period $[B, C]$ are the quantities of primary interest, but neither can be observed directly. Rather, the infection status of the partner, denoted by the **binary** random variable Y, is ascertained at the time the partnership is recruited along with retrospective information on exposure.

In some cases (e.g. couples where the index case is infected with HIV via blood transfusion), the time

of infection of the index case I and the exposure interval $[B, C]$ can be identified along with an estimate of the number or rate of contacts during this period. Here, observations of V are **censored** on the left for infected partners ($Y = 1$) and on the right for uninfected partners ($Y = 0$). This form of interval censoring is also known as *current status data* on V. In cases in which the time of index case infection is unknown, it is impossible to identify correctly the beginning of the exposure interval. In this case, observations are limited to the knowledge that infections for both partners occurred in the interval $[A, C]$, and that the infection of the index case preceded the partner's. For a long-term partnership (Figure 1), only A, C, Y, and contact information are observed. For a short-term partnership (Figure 2), the initiation of contact B is observed as well. Both of these types of observations are examples of *doubly censored current status data* on the time to partner infection V. In practice, retrospective studies may contain a mixture of observations of both types, in addition to more complex exposure patterns.

Statistical Models

Statistical interpretation of partner study data is usually based on simplified models for the *transmission probability* of infection in the susceptible partner, defined as the probability of infection occurring, conditional on a specified exposure history. Consider the case of a single couple from a retrospective study in the case that the time of the infection of the index case is known. Let $K(x)$ represent a **counting process** with intensity $\mu(x)$ which measures contacts in the interval beginning at the initiation of contact B. Let $\lambda(x)$ represent the probability that infection is transmitted via a contact occurring at time x following the index case's infection at I. This represents the transmission risk associated with a single contact and is often referred to as the *infectivity* of the disease. Necessarily, $\lambda(x)$ depends on both susceptibility of the partner and infectiousness of the index case, and the relative influence of these cannot be investigated in the absence of additional data. If complete data in the sense described above are available, the conditional probability of the partner being infected following an exposure of length $T = C - B$, given the observed sample path of $K(x)$ can be written:

$$P_{t,k} = \Pr[Y = 1 | K(x) : B - I \le x \le C - I; \; K(C - I) = k] = 1 - \prod_{j=1}^{k} [1 - \lambda(x_j)], \quad (1)$$

where $(x_1, \ldots, x_k)$ are the times of contacts. This probability affords two interpretations in a survival analysis context: (i) the probability that the partner's true infection time V does not exceed $C - I$; and (ii) the probability that the number of contacts $K(V)$ required for infection to occur does not exceed $K(C - I) = k$. If $\lambda(x)$ is assumed to be a constant, then this reduces to the *constant infectivity model* $P_k = 1 - (1 - \lambda)^k$, which is frequently used as a null model against which more complex models are compared. Since the times $(x_1, \ldots, x_k)$ of contacts are unlikely to be available in any of the study designs mentioned above, most analyses are based on the marginal transmission probability based on the time V between infections of the two partners, or on the discrete scale provided by the number of contacts $K(V)$ required for infection to occur. The marginal probability for becoming infected following an exposure of duration t beginning at B is computed from (1) by taking the expectation with respect to the sample paths of the contact process $K(x)$ [16]. When $K(x)$ is a nonhomogeneous **Poisson process** with intensity $\mu(x)$,

$$P_t = 1 - \exp\left[- \int_{B-I}^{C-I} \lambda(x)\mu(x) \, \mathrm{d}x \right] = \frac{G(C - I) - G(B - I)}{1 - G(B - I)}, \quad (2)$$

where G denotes the distribution function for the infection time V, with **hazard** function $\lambda(x)\mu(x)$. For long-term partnerships $B \equiv I$, and the right hand side equals $G(C - I)$. When $\lambda(x)$ and $\mu(x)$ are both constant this is the **exponential** model, the time-based analog of the constant infectivity model mentioned above. The infectivity $\lambda(x)$ is clearly not estimable without external information on $\mu(x)$. If this is available, the model affords a **proportional hazards** interpretation, with $\lambda(x)$ representing the "baseline" hazard of infection associated with a unit contact rate. The effects of a vector Z of covariates can also be considered, as discussed below. Because

of these close ties with conventional **survival analysis** methods, models based on P_t are in general preferred to those based on the marginal transmission probability P_k for becoming infected following an exposure of k contacts. In the constant infectivity case, these models are almost indistinguishable.

When the time of infection of the index case is unknown, the exposure interval is also impossible to specify. If, as discussed above, a lower bound A on the time I of index case infection is known, and if additional information on the distribution F of this time in the interval $[A, C]$ is available, the transmission probability P can be computed as follows [6]:

$$P = \frac{1}{F(C)}\left[\int_A^B \frac{G(C-s)-G(B-s)}{1-G(B-s)}\,\mathrm{d}F(s) + \int_B^C G(C-s)\,\mathrm{d}F(s)\right] \tag{3}$$

The first integral on the right-hand side of (3) is necessary for short-term partnerships, where it is unknown whether or not the index case infection preceded the initiation of contact B (Figure 2). In long-term partnerships formed prior to A this term vanishes, consistent with the assumption that the index case was infected subsequent to this time (Figure 1).

Estimation and Inference

Estimation and **inference** for parameters in transmission models are most commonly based on **maximum likelihood** methods. For a sample of n partnerships from a retrospective or cross-sectional study, the log **likelihood** function (conditional on the observed exposure information) is

$$\sum_{i=1}^{n} y_i \log P_i + (1-y_i)\log(1-P_i), \tag{4}$$

where P_i represents the transmission probability as described above. In the case in which index case infection times are known (i.e. current status information is available on partners' infection times), **nonparametric maximum likelihood** estimates of the distributions of V and $K(V)$ can be obtained using **isotonic regression** [3, 7]. If index case infection times are unknown (i.e. doubly censored current status information is available on partners' infection times), nonparametric estimation of G from (3) is also possible in the case of long-term partnerships if the distribution F of index case infection times is assumed to be uniform [6].

In many cases it is desirable to estimate the infectivity $\lambda(x)$ directly, and to investigate covariate effects in a regression setting. Using the proportional hazards structure of (2), which is appropriate for current status information on partner infection in a retrospective design, the complementary log–log transformation (*see* **Binary Data**) yields the following regression model:

$$\log[-\log(1-P_i)] = \log\left[\int_{B_i-I_i}^{C_i-I_i} \lambda(x)\,\mathrm{d}x\right] + \log\mu_i + \beta z_i. \tag{5}$$

Here, the contact rate μ_i for the ith partnership has been assumed to be constant and $\log\mu_i$ represents an offset term. For parametric choices of $\lambda(x)$ the techniques of **generalized linear models** can be used to estimate parameters in this model. For example, when this parameter is assumed to be constant in (2) (or (1)), (4) is a standard generalized linear model for the binary outcome Y and can be treated using standard techniques and **software** [3]. When no parametric assumptions are made about $\lambda(x)$ the resulting models are **semiparametric**. For (5), the constraint that the integrated infectivity is nondecreasing in the duration of exposure must also be taken into account in estimation. Estimation techniques include estimating $\lambda(x)$ and β jointly using **penalized maximum likelihood** and the **EM algorithm** [16]; **profile likelihood** [12]; or modified **generalized additive models** [15, 16, 18]. The regression coefficients β can also be estimated separately, treating the infectivity as a **nuisance parameter** using an approximate **partial likelihood** approach [13]. Estimation methods for doubly censored current status observations are poorly developed in the semiparametric setting. Recent results indicate that such data generally provide inconsistent estimates (*see* **Consistent Estimator**) of the distribution G unless additional assumptions (e.g. smoothness) are imposed [5], and that, for estimation of β, methods for current status data can be applied with little loss of **efficiency** [10]. Fully parametric models for such data can be also handled using standard maximum likelihood techniques.

Inference about the parameters introduced in the models above is typically focused on two areas: (i) the effects of covariates on the probability of disease transmission; and (ii) investigation of the properties of the infectivity $\lambda(x)$. Techniques for parameters in fully parametric transmission models are straightforward, and in many cases follow standard practice for generalized linear models. By contrast, asymptotic theory for parameter estimates in semiparametric models is not well developed. In addition, there are a number of complicating issues in inference which arise from the lack of information most studies provide about exposure and the infection process. These include the possibilities of measurement error in exposure data related to retrospective **ascertainment**, and the potential for uncontrolled heterogeneity of infectivity between partnerships reflecting unmeasured factors influencing infectiousness and susceptibility. Ignoring these factors may lead to biased estimates of λ and β [3, 14, 16]. Unfortunately, data from partner studies is usually extremely limited, making it very difficult or impossible to correct for these problems using existing methods. For example, data are rarely sufficient to allow mixture models incorporating heterogeneity into λ to be fit, except in cases in which the latter is assumed to be constant [20]. Similarly, **validation** data for use in modeling measurement error are rarely if ever available. Finally, even in well conducted studies with valid exposure information it is rare that data are extensive enough to yield much detailed information about the form of the infectivity. Despite these limitations, partner studies remain valuable tools in understanding infectious disease transmission, and are worthy of further statistical research.

Many of the above techniques can be generalized to apply in other settings, including studies in which identification of the index case is uncertain [9], and in which multiple potential index cases may be associated with each partner [11, 17, 19]. Obtaining reliable estimates of the per-contact infectivity is extremely difficult in these situations, but the effects of covariates on the probability of transmission can still be investigated.

References

[1] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans*. Oxford University Press, New York.

[2] Becker, N.G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, New York.

[3] Jewell, N.P. & Shiboski, S.C. (1990). Statistical analysis of HIV infectivity based on partner studies, *Biometrics* **46**, 1133–1150.

[4] Jewell, N.P. & Shiboski, S.C. (1993). The design and analysis of partner studies of HIV transmission, in *Methodological Issues in AIDS Behavioral Research*, D.G. Ostrow & R. Kessler, eds. Plenum, New York.

[5] Jewell, N.P. & van der Laan, M. (1994). Generalizations of current status data, *Lifetime Data Analysis* **1**, 101–109.

[6] Jewell, N.P., Malani, H. & Vittinghoff, N.P. (1994). Nonparametric estimation for a form of doubly censored data with application to two problems in AIDS, *Journal of the American Statistical Association* **89**, 7–18.

[7] Kaplan, E.H. (1990). Modeling HIV infectivity: must sex acts be counted?, *Journal of Acquired Immune Deficiency Syndrome* **3**, 55–61.

[8] Kim, M.Y. & Lagakos, S.W. (1990). Estimating the infectivity of HIV from partner studies, *Annals of Epidemiology* **1**, 117–128.

[9] Magder, L. & Brookmeyer, R. (1993). Analysis of infectious disease data from partner studies with unknown source of infection, *Biometrics* **49**, 1110–1116.

[10] Rabinowitz, D. & Jewell, N.P. (1996). Regression with doubly censored current status data, *Journal of the Royal Statistical Society, Series B* **58**, 541–550.

[11] Rhodes, P., Halloran, M.E. & Longini, I.M. (1996). Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility, *Journal of the Royal Statistical Society, Series B* **58**, 751–762.

[12] Rossini, A. & Tsiatis, A.A. (1996). A semiparametric proportional odds regression model for the analysis of current status data, *Journal of the American Statistical Association* **91**, 713–721.

[13] Satten, G. (1996). Rank-based inference in the proportional hazards model for interval censored data, *Biometrika* **83**, 355–370.

[14] Shiboski, S. (1994). Statistical interpretation of data from partner studies of heterosexual HIV transmission, in *Modeling the AIDS Epidemic*, E. Kaplan & M. Brandeau, eds. Raven Press, New York.

[15] Shiboski, S. (1998). Generalized additive models for current status data, *Lifetime Data Analysis*, to appear.

[16] Shiboski, S.C. & Jewell, N.P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data, *Journal of the American Statistical Association* **87**, 360–372.

[17] Shiboski, S.C. & Padian, N. (1996). Population and individual based approaches to the design and analysis of epidemiological studies of STD transmission, *Journal of Infectious Diseases* **174**, Supplement 2, 188–200.

[18] van der Laan, M. (1995). Locally efficient estimation with current status data and high dimensional covariates, *Group in Biostatistics Technical Report*, no. 55. University of California, Berkeley.

[19] Wick, D. & Self, S.G. (1997). Estimating disease attack rates in heterogeneous interacting populations with applications to HIV vaccine trials, *Annals of Statistics*, **25**, 642–661.

[20] Wiley, J.A., Herschkorn, S.J. & Padian, N.S. (1989). Heterogeneity in the probability of HIV transmission per sexual contact: the case of male-to-female transmission in penile–vaginal intercourse, *Statistics in Medicine* **8**, 93–102.

S.C. SHIBOSKI

Pascal, Blaise

Born: June 19, 1623, in Clermont, France.
Died: August 19, 1662, in Paris, France.

Blaise Pascal was one of the most influential figures in French life and literature. Here, only his contributions to mathematics are considered.

In 1631, Pascal's father Etienne moved from Clermont to Paris in order to secure his son a better education, and in 1635 he was one of the founders of Marin Mersenne's "Academy", to which he introduced his son at the age of 14. The younger Pascal immediately put this new source of knowledge to good use, producing (at the age of 16) his famous *Essay pour les coniques*.

In the succeeding years, Pascal *fils* designed and had built the first mechanical adding machine (there is now a computer language called "Pascal") and conducted experiments into the nature of a vacuum (the "Pascal" is the SI unit of pressure), but his chief mathematical contribution was to lay the foundations of the **theory of probability**.

Before his time, probability calculations amounted to no more than the enumeration of equally probable outcomes in games of chance, but Pascal introduced the important idea of **expectation** and used recursively the fact that if expectations of gain X and Y are equally probable, the expectation is $\frac{1}{2}(X+Y)$. He also introduced the **binomial distribution** for equal chances and, with its help, and that of mathematical induction applied to expectations, solved the Problem of Points for two players.

This problem was the topic of correspondence between Pascal and Pierre de Fermat in 1654 which, together with Pascal's contemporary *Traité du triangle arithmétique*, includes three methods of solution. Two players stake equal money on being the first to win n points in a game in which the winner of each point is determined by the toss of a coin. If such a game is interrupted when one player still lacks a points and the other b, how should the stakes be divided between them?

Fermat and Pascal independently concluded that the problem could be solved by noting that at most $a+b-1$ more tosses will settle the game, and that if this number of tosses is imagined to have been made, the resulting 2^{a+b-1} possible games (each equally probable) may be classified according to the winner in each case, the stakes then being divided accordingly. Thus, the real game, of indeterminate length, is embedded in an imaginary game of fixed length. Apart from this novel idea, however, such a solution by enumeration was straightforward, but Pascal offered both an independent method based on expectations which is valid for any number of players, and, in the *Traité du triangle arithmétique*, the solution for two players in terms of the binomial distribution, proved by induction. He did not give the binomial distribution algebraically, but by reference to the "arithmetical triangle" of binomial coefficients, the properties of which he elaborated in his *Traité* (whence the name "Pascal's Triangle").

Pascal's seminal work laid the foundations of probability theory, influencing Christian Huygens (*De Ratiociniis in Ludo Aleae*, 1657) and thence James **Bernoulli** (*Ars Conjectandi*, 1713). Already in 1710, we find the binomial distribution being applied to a biological problem (John Arbuthnot, *An Argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes*) and giving rise to the first test of significance (*see* **Hypothesis Testing**).

References

The literature on Pascal is vast, but his contributions to probability are fully covered in:

Edwards, A.W.F. (1987). *Pascal's Arithmetical Triangle*. Griffin, London/Oxford University Press, New York.

Edwards, A.W.F. (2002). *Pascal's Arithmetical Triangle*. Johns Hopkins University Press, Baltimore.

A.W.F. EDWARDS

Paternity Testing

Paternity testing on the basis of **blood group** typing has been known since 1920, but at that time was only used to prove nonpaternity by way of exclusion. For example, in the case of a child with blood group AB, and mother with blood group B, a man with blood group O could be excluded from paternity, because the true father must have the component A in his phenotype. In 1938, Essen-Möller [2] was the first author to propose a statistical approach in nonexclusion situations to quantify the available genetic information. For the standard trio case (child, mother, putative father) he defined the two alternative hypotheses "paternity" vs. "unrelated" and, on the basis of estimated population frequencies, evaluated the likelihoods:

$$X = L(\text{observed phenotypes}|\text{"paternity"})$$

and

$$Y = L(\text{observed phenotypes}|\text{"unrelated"}).$$

Essen-Möller then used X and Y in a **Bayesian methods** approach with equal prior probabilities to calculate the probability of paternity as

$$W = \frac{X}{X+Y},$$

thus starting a never-ending discussion regarding the appropriateness of introducing prior probabilities into the statistical argument. It has to be pointed out, though, that already at this stage it was evident that the likelihood ratio X/Y (now known as the paternity index ([1] and [5]) contained the complete genetic evidence for a given case. Because of differing legal situations in different countries, development since the 1950s has not always been parallel, but by now has led to more or less general scientific agreement and standardization regarding the following aspects:

1. Use of high-quality **genetic marker** systems with adequately estimated gametic population frequencies.
2. Exact formulation of hypotheses for standard trio cases, as well as complex relationships in deficiency cases.
3. Standard **likelihood** calculation on the basis of well-defined hypotheses and adequate parameter estimates.
4. **Likelihood ratio** as the basis of decision making.
5. Additional statistics for the general and case-specific probability of exclusion.

Marker Systems

The basis for all statements are the genetic marker systems used for typing and the knowledge of the genetic parameters governing these systems. In the case of fully **penetrant** marker systems, these parameters are the mutation rates, the recombination fractions and the gametic population frequencies of the respective allelic expressions (*see* **Gene Frequency Estimation**). Quality aspects regarding the selection of markers make it preferable to have systems with a negligible amount of mutation, free recombination between marker loci and no **linkage disequilibrium**, which is equivalent to intragametic independence, thus allowing easy calculation of overall likelihoods by multiplication of marginal frequencies [4]. In addition, the quality of such a marker system is measured by its power to exclude paternity for an unrelated man (general power of exclusion, *GPE*), which is a direct function of the number of alleles and their population frequencies. The historical development has led from blood groups, protein systems, enzyme systems (*see* **Polymorphism**) to the highly polymorphic **HLA system** with all its complexity of closely linked loci and linkage disequilibrium, to the now available (over)abundance of **DNA sequence** marker systems. Within the group of DNA markers the multilocus probes (MLPs) are being used less and less due to the impossibility of determining the **genotypes** corresponding to a given set of phenotypic signals. The standard systems now being used almost worldwide are the two groups of single locus probes (SLPs) with a continuous distribution of allelic "fragment length" and the short tandem repeat (STR) systems with discrete alleles [3]. Although SLP systems have extreme discrimination powers (high **heterozygosity**, high *GPE*), the problem of measurement error for the continuous fragment lengths in these systems leads to complications in their statistical evaluation. It can be expected that STRs, with their simplicity of allele and genotype definition, their ever-increasing number, their good discrimination power and their increasing ease of typing, will become the standard in all kinds of relationship testing.

Hypotheses

The basis of all statistical argument in cases of nonexclusion is well-defined hypotheses. In the standard trio case, the following are usually the two alternatives:

H_1: The putative father is the true biological father of the child.

H_2: The putative father is not the biological father of the child: he is an unrelated random male from a given population.

It must be pointed out that the second hypothesis is not simply the negation of the first hypothesis and has to be clearly distinguished from the situation where the putative father could be related to the true father (for example, he is his brother). Furthermore, the ethnic background(s) of the putative father, the implicitly postulated true father and, to a lesser extent, that of the mother, are part of the hypotheses. For more complex situations in deficiency cases (such as when the putative father is not available for testing, but information on blood relatives is available) two (or more) hypotheses must be well defined and can be represented by the corresponding pedigree(s) that include all individuals necessary to represent the postulated relationships, irrespective of the availability for marker typing. Figure 1 represents the situation of a case with a deceased putative father whose parents are both available for typing.

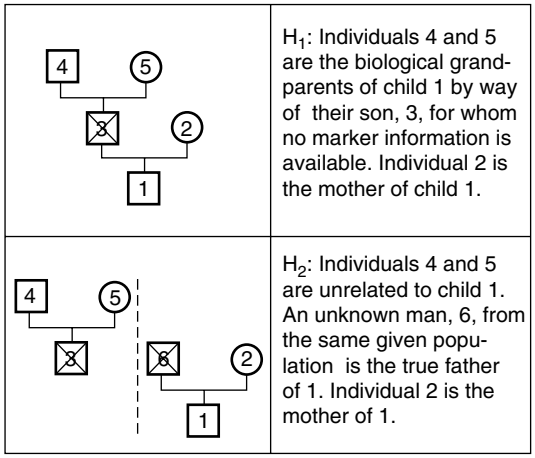

Figure 1 Hypothesis formulation and representation by way of corresponding pedigrees

Likelihood Calculation

Every conceivable well-defined hypothesis can be represented as a pedigree. Consequently, the calculation of the likelihood of the given hypothesis for the observed phenotypes is identical to the probability of the observed phenotypes given the corresponding pedigree. Let $g_{i,j}$ be the ith possible genotype for the jth person, with population frequency $f_{i,j}$, ph_j be the genotype of the jth person, and $\Pr(g_{i,j}|g_{i,p}, g_{i,m})$ be the probability of $g_{i,j}$ conditional on the genotypes of j's parents. Then the general formulation of such a likelihood is

$$L(\text{phenotypes}|\text{pedigree}) = \sum_{g_{i,1}} \sum_{g_{i,2}} \cdots \sum_{g_{i,k}} \prod_{k} P(ph_k|g_{i,k}) \times \prod_{\text{founders}} f_{i,k} \prod_{\text{descendants}} P(g_{i,k}|g_{i,p}, g_{i,m})$$

(*see* **Elston–Stewart Algorithm**).

For the case of high-quality marker systems, this general formulation is considerably simplified:

1. The nested summation for all individuals reduces to a single term for one-to-one correspondence of genotype and phenotype, which is the case for single locus DNA markers with no silent alleles.
2. Marker systems having complete penetrance reduce the number of terms.
3. In the case of intragametic independence, population frequencies are given by the product of the allele frequencies for each marker.
4. Transmission probabilities are 1 for a homozygous parent and 1/2 for a **heterozygous** parent.

For example, consider the simple case of a trio with the following typing results:

child (a/c), mother (c/d), putative father (a/b).

Letting f_i be the population frequency of allele i, the likelihoods of the hypotheses evaluate to

$$X = L(\text{genotypes}|H_1) = f_a f_b f_c f_d,$$
$$Y = L(\text{genotypes}|H_2) = 2 f_a f_b f_a f_c f_d.$$

The general likelihood formulation can handle dominance or untyped individuals by summation over

all compatible genotypes, closely linked loci with linkage disequilibrium by using **haplotype** frequencies on the population level and recombination fractions in the transmission probabilities, and even possible mutations can be modeled into this component (if reliable estimates are available). Ihm & Hummel [6] were the first to present an algorithm to evaluate the likelihood for a complex hypothesis in relationship testing and today many software packages for pedigree analysis (*see* **Software for Genetic Epidemiology**) can perform this calculation, given the parameter estimates.

Likelihood Ratio, *PI*, and Posterior Probability, *W*

In the case of two well-defined hypotheses with corresponding likelihoods X and Y, the likelihood ratio statistic X/Y has been termed the *paternity index*, *PI*. The *PI* contains the complete genetic information and has the general property that with increasing *PI* the probability of H_1 in comparison with H_2 also increases. Consequently, the decision process for the acceptation or rejection of H_1 ("Paternity") is the process of setting a decision threshold for the *PI* statistic. In most countries, a likelihood ratio of $PI = 1000 : 1$ has been chosen for the verbal conclusion "Paternity practically proven". Some countries are asking for $PI = 10\,000 : 1$. The power of the marker systems now available for typing is strong enough to exceed this threshold in almost all situations when there is paternity, and, in nonpaternity situations, one or several exclusions will most likely be detected. The evaluation of type I and type II errors (*see* **Hypothesis Testing**) would be theoretically possible on the basis of the two distributions of all possible genotype constellations conditional on the two hypotheses. Because of the fact that new marker systems are constantly being developed and introduced into paternity testing, these conditional distributions would have to be re-evaluated for any new set of marker systems.

The alternative approach, which dates back to the original paper of Essen-Möller [2] is the calculation of posterior probabilities using **Bayes' theorem**. Given the conditional likelihoods, X_j, for two or more mutually exclusive hypotheses H_j with prior probabilities p_j, the posterior probability of H_i is given by

$$\Pr(H_i|\text{phenotypes}) = \frac{p_i X_i}{\sum_j p_j X_j}.$$

This expression can only be evaluated if, in addition to the conditional likelihoods, X_j, the prior probabilities of the hypotheses H_j are given. For the simple situation of two alternative hypotheses, the above expression reduces to

$$W = \Pr(\text{paternity}|\text{phenotypes}) = \frac{pX}{pX + (1-p)Y},$$

and, if one is willing to set both prior probabilities to $p = 1/2$,

$$W = \frac{X}{X+Y},$$

the formula given by Essen-Möller [2]. It has to be pointed out that the **information** contained in W (with equal priors) is identical to that of *PI* since

$$W = \frac{PI}{PI+1}.$$

A threshold of $PI = 1000$ is equivalent to $W = 0.9990$.

The argument about the use of the paternity index only, the use of the probability of paternity with equal priors, or the estimation of prior probabilities from "similar" previous cases and thereby the introduction of nongenetic evidence, has governed the scientific debate for decades. However, it has recently ended because the strength of the genetic evidence is so extreme that, no matter what decision procedure is used, the amount of error in standard cases is negligible. In complex deficiency cases with two alternative hypotheses, the likelihood ratio is again the statistic presenting the complete genetic information for the postulated relationship. Since, in many situations, direct exclusions are structurally not possible – for example, child, mother and one putative grandparent – the likelihood ratio symmetrically provides information against the postulated relationship. The complex situation with more than two mutually exclusive hypotheses is not directly accessible. Pairwise comparisons with likelihood ratios is feasible but the comparison of one conditional likelihood, X_i, simultaneously with all others is only possible by way of the general Bayesian formulation using

assigned priors. The approach using equal priors for all hypotheses allows easy calculation, but provides nothing more than a representation of the conditional likelihoods, X_i, normed. Nevertheless, in this way, preponderance of one of the hypotheses becomes easily visible.

General and Case-specific Probability of Exclusion

In addition to the above defined likelihood statements, there are two further statistics used in paternity testing, one characterizing the general power of a marker system to detect an exclusion (*GPE*), the other characterizing its power to exclude an unrelated man in the case of a specific mother–child phenotype (*SPE*).

Let M be a marker system with $i = 1, \ldots, n$ different phenotypes, ph_i. For a given mother–child constellation $(ph_\mathrm{m} : ph_\mathrm{c})$ the set of all phenotypes $\{ph_i\}$ contains a uniquely defined subset of phenotypes genetically compatible for a father of the child. The sum of the frequencies of these phenotypes is the probability that a random man is not excluded (*RMNE*):

$$RMNE(ph_\mathrm{m} : ph_\mathrm{c}) = \sum_{\text{compatible}} f(ph_i),$$

and the specific probability of exclusion consequently is defined as

$$\begin{aligned} SPE(ph_\mathrm{m} : ph_\mathrm{c}) &= 1 - RMNE(ph_\mathrm{m} : ph_\mathrm{c}) \\ &= 1 - \sum_{\text{compatible}} f(ph_i). \end{aligned}$$

Obviously, *SPE* depends on the phenotypes of the given mother–child constellation. For any marker system M the general power to exclude a random man, *GPE*, is given as follows. Let $i = 1, \ldots, l$ be the number of all possible mother–child constellations $(ph_\mathrm{m} : ph_\mathrm{c})_i$. Let f_i be the corresponding frequencies of these constellations (easily calculated from the general likelihood formulation for pedigrees; see above) and SPE_i the corresponding specific probability of exclusion defined above. Then, the general exclusion probability of this system M is given by

$$GPE_M = \sum_{i=1}^{l} f_i SPE_i.$$

For a set S of k independent marker systems M_i, the overall exclusion probability is simply

$$GPE_S = 1 - \prod_{i=1}^{k} (1 - GPE_{M_i}).$$

Given the new DNA polymorphisms available for testing, it is no problem to reach overall exclusion probabilities $\geq$99.9% with sets of 8–10 independent single locus systems. Cost–benefit analysis has shown that for a given level of GPE_S the overall cost, C_S, is minimized if markers, M_i, with corresponding costs, C_i, are selected according to the magnitude of the benefit–cost relationship, R_i, defined by

$$R_i = \frac{-\log_{10}(1 - GPE_i)}{C_i}.$$

Note that in deficiency cases the evaluation of *SPE* may be very complex, or for certain hypotheses structurally equal to zero and thus meaningless.

Conclusion

Paternity testing on the basis of marker typing has become a highly efficient and standardized procedure due to the large amount of information now available. Marker systems with high discrimination power are available, statistical procedures have been developed to describe the general quality of the marker system (*GPE*) as well as statistics for the given case (*PI*, *W*, *SPE*) – see, for example, [7]. Complex relationship testing as an extension of paternity testing can be performed on the basis of exact formulation of hypotheses and corresponding likelihood statements. Quality assessment requires continuous quality control of the laboratory work and a good data base to determine relevant population parameters of the marker systems used. The selection of marker systems with a high benefit–cost ratio leads to improved discrimination power at reduced cost.

References

[1] Baur, M.P., Elston, R.C., Gürtler, H., Henningsen, K., Hummel, K., Matsumoto, H., Mayr, W., Morris, J.W., Niejenhuis, L., Polesky, H., Salmon, D., Valentin, J. & Walker, R.H. (1986). No fallacies in the formulation of the paternity index, *American Journal of Human Genetics* **39**, 528–536.

[2] Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis – Theoretische Grundlagen, *Mitteilungen der Anthropologischen Gesellschaft in Wien* **68**, 9–53.

[3] Evett, I.W. & Buckleton, J.S. (1995). Statistical analysis of STR data, in *Advances in Forensic Haemogenetics*, Vol. **6** A. Carracedo & B. Brinkman, eds. Springer-Verlag, Berlin, pp. 79–86.

[4] Gjertson, D.W. & Morris, J.W. (1995). Assessing probability of paternity and the product rule in DNA systems, *Genetica* **96**, 89–98.

[5] Gürtler, H. (1956). Principles of blood group statistical evaluation of paternity cases at the University Institute of Forensic Medicine Copenhagen, *Acta Medicinae Legalis et Socialis Liege* **9**, 83–93.

[6] Ihm, P. & Hummel, K. (1975). Ein Verfahren zur Ermittlung der Vaterschaftswahrscheinlichkeit aus Blutgruppenbefunden unter beliebiger Einbeziehung von Verwandten, *Zeitschrift für Immunitätsforschung* **149**, 406–416.

[7] Walker, R.H., ed. (1983). *Inclusion Probabilities in Parentage Testing*. American Association of Blood Banks, Arlington.

MAX P. BAUR

Path Analysis in Genetics

The method of **path analysis**, introduced by the pioneer population geneticist Sewall Wright [40, 41] over 70 years ago, is a form of structural linear regression analysis of standardized variables whose purpose is two-fold: to explain the interrelationships among a given set of variables, and to evaluate the relative importance of varying causes influencing a variable of interest. Algebraic equations that specify the structural interrelationships among the variables are known as **structural equations**. Although path analysis can be pursued strictly through structural equations, path diagrams that specify the proposed structural relationships schematically are often more helpful for conceptualizing a complex path model quickly, and for determining the internal consistency and limitations of a given model. Algebraic manipulation of structural equations or analysis of path diagrams using a standard set of tracing rules [19] both yield identical results.

Path analysis was originally developed by Sewall Wright for a systematic study of **inbreeding** and systems of mating. Despite its application in a variety of contexts and despite its inherent appeal, it remained a secluded method for a long time. In later years this method has been used extensively in sociological and econometric modeling in the guise of structural equation models (e.g. [9]). The method was largely neglected by the statistical community perhaps because, around the same time that path analysis was being developed, statisticians were overwhelmed by the beauty and logic of the formal statistical methods that were being rapidly developed. A relative lack of mathematical formalization appears to have rendered path analysis less attractive [15]. Starting in the mid 1970s, however, path analysis was widely applied in human genetics and **genetic epidemiology**, perhaps partly aided by the development of formal tests of hypotheses [33].

Path analysis is especially useful for studying the genetic epidemiology of complex phenotypes. In terms of specifying the interrelationships among variables in a path model, there is a theoretical basis at least for some of the relationships in families. For example, **genes** cause phenotypes (*see* **Genotype**) and not the other way around, and the genes of children are determined (caused) by those of their parents. When the direction of "causation" is less clear, ambiguous correlational and other associative relationships, such as *copaths* [3] and *delta paths* [38], are used. Formal tests of hypotheses are now routine in such situations, and path analysis enables testing whether the theoretical model is consistent with the observed interrelationships among variables (i.e. assessment of **goodness of fit**). A rich battery of path models have been developed over the years for dealing with a variety of situations in genetic epidemiology and behavioral genetics; see, for example, [25].

For the purposes of evaluating familial resemblance, path models are based on *multifactorial inheritance*, which means that multiple genetic and/or familial environmental effects contribute to the resemblance within families where none of the individual effects is large. In turn, the genetic effects are assumed to be polygenic (*see* **Polygenic Inheritance**) (i.e. many genes each with a small and equal effect). While some models accommodate interactions at the polygenic level, polygenic effects are often assumed to be additive, and this is the case in the path analysis models used in genetic epidemiology.

Under a given path model, correlations among the variables (or, more specifically, among phenotypes of family members) can be derived either by taking the mathematical expectations of products of (standardized) variables, or directly from the path diagram following a simple set of tracing rules (see [19]). The gist of the method of path analysis consists of a comparison of these model-based correlations, which are functions of unknown parameters, with the actual data. Fitting path models to family data needs appropriate statistical methods for obtaining consistent and efficient estimates of the parameters of the model, as well as for testing specific null hypotheses. Path analysis usually considers a series of hypotheses and rejects some as being inconsistent with the data, thus leading to what appears to be a plausible model for the observed data. For recent reviews of path analysis in genetic epidemiology see [25] and [16].

The basic path model for familial resemblance described in what follows can be extended to suit other needs. For example, it is easily expanded to include temporal and developmental variations across time, to assess heterogeneity among multiple studies, and to assess multiple correlated traits simultaneously (i.e. **multivariate analysis**). The complexity of the models depends on the particular need as well as

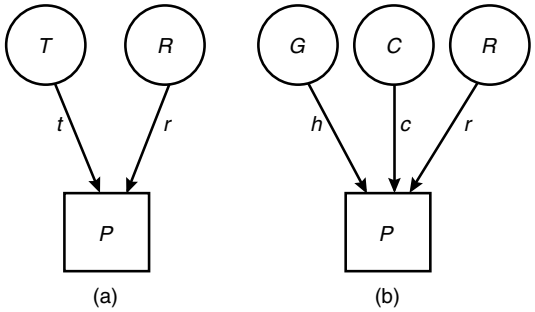

Figure 1 Path diagram for the basic model. Part (a) depicts the generic model involving a combined familial component (T) and a residual (R). Part (b) depicts the specific model with latent genetic (G), familial environmental (C), and residual (R) sources of phenotypic (P) variation in an individual. Path coefficients are defined in the text

the type of data collected and the type of family structures represented in the families.

The Basic Model

At an individual level, one may distinguish between two variations of the basic model (Figure 1). A *generic* model (Figure 1(a)) postulates that the measured variable (the phenotype P) is derived from the additive effects of two sources: the total familial effect, where both genetic and familial environmental effects are lumped together (T), and a residual that is not familial (R). A *specific* model (Figure 1(b)) further splits the total familial effect (T) into a separate genetic effect (G) which is assumed here to be polygenic with additive gene action, and a familial environmental effect (C). The observed (measured) variables are shown in squares and latent (unobserved) variables are shown in circles.

The Generic Model

In terms of the corresponding unstandardized variables (P^*, T^*, and R^*, each with zero mean), the generic basic model may be expressed as

$$P^* = T^* + R^*, \tag{1}$$

where the familial and residual effects are assumed to be uncorrelated. The total phenotypic variance is given by

$$\sigma_p^2 = \sigma_t^2 + \sigma_r^2, \tag{2}$$

where σ_x^2 denotes the variance of X. The components (σ_t^2 and σ_r^2) are called variance components. The basic structural equation underlying this model is obtained by dividing both sides of (1) by σ_p:

$$P = tT + rR, \tag{3}$$

where P, T, and R denote standardized variables, and the standardized partial regression coefficients t and r, called path coefficients, are given by

$$t = \frac{\sigma_t}{\sigma_p} \quad \text{and} \quad r = \frac{\sigma_r}{\sigma_p}. \tag{4}$$

The variance of P in (3) defines what is called the equation for complete determination:

$$t^2 + r^2 = 1, \tag{5}$$

the components of which provide the variance components relative to the total phenotypic variance. Therefore, the components of (2) may be referred to as absolute variance components, while those of (5) may be called relative variance components.

The Specific Model

Likewise, the specific model may be expressed as

$$P^* = G^* + C^* + R^*. \tag{6}$$

Although exceptions exist in the literature (e.g. see [25]), here all three causes are assumed to be uncorrelated with each other and to act additively to produce the phenotype. The corresponding structural equations, absolute and relative variance components, and path coefficients are given by

$$\begin{gathered} P = hG + cC + rR, \\ \sigma_p^2 = \sigma_g^2 + \sigma_c^2 + \sigma_r^2, \\ h^2 + c^2 + r^2 = 1, \\ h = \frac{\sigma_g}{\sigma_p}, c = \frac{\sigma_c}{\sigma_p}, r = \frac{\sigma_r}{\sigma_p}. \end{gathered} \tag{7}$$

Historically, the classic quantitative genetic model postulated that the phenotype was merely derived from the additive effects of genes and (residual) environments, where the environment was totally nonfamilial. This is equivalent to the specific model

without the familial environment (C). The equation for complete determination would reduce to $h^2 + r^2 = 1$. Because all familial effects are then only due to genes, h^2, the proportion of phenotypic variance explained by the genetic component (G), was simply called **heritability**. With the introduction of familial nongenetic effects, vis-à-vis the familial environment (C), some prefer to call h^2 the genetic heritability. Likewise, one may define cultural heritability (c^2) as the proportion of the phenotypic variance explained by the familial environment. In the specific basic model, h^2 and c^2 are the two unknown parameters, since r^2 is readily obtained from (7). If it were possible to measure the underlying genes and environments (G and C), then this model would be identified and we could estimate the heritabilities directly from individual data. However, lacking such measures, we have to resort to indirect methods of estimating them and testing hypotheses about them from particular types of family data.

Path Models for Familial Resemblance

Extending the basic model for an individual (Figure 1) to groups of related individuals (families) involves additional assumptions. These concern specification of marital resemblance and **assortative mating**, transmission from parents to offspring and other effects that are unique to subjects and their relatives. One of the simplest extensions is to nuclear families consisting of parents and offspring, which demonstrate most of the concepts.

Nuclear families are the most frequently used study design for assessing family resemblance, and have the advantage of providing the most representative sample of the population to which the results are to be generalized. However, phenotypic data alone from nuclear families do not provide enough information to differentiate between genetic and family environmental (cultural) sources of family resemblance. Therefore, such data can be used only to assess the combined effect of both genetic and cultural effects. To resolve the genetic and cultural sources of variance, several strategies have been developed which use either additional types of family members such as twins, adoptees, and extended families, or use measured variables to index the latent familial environment.

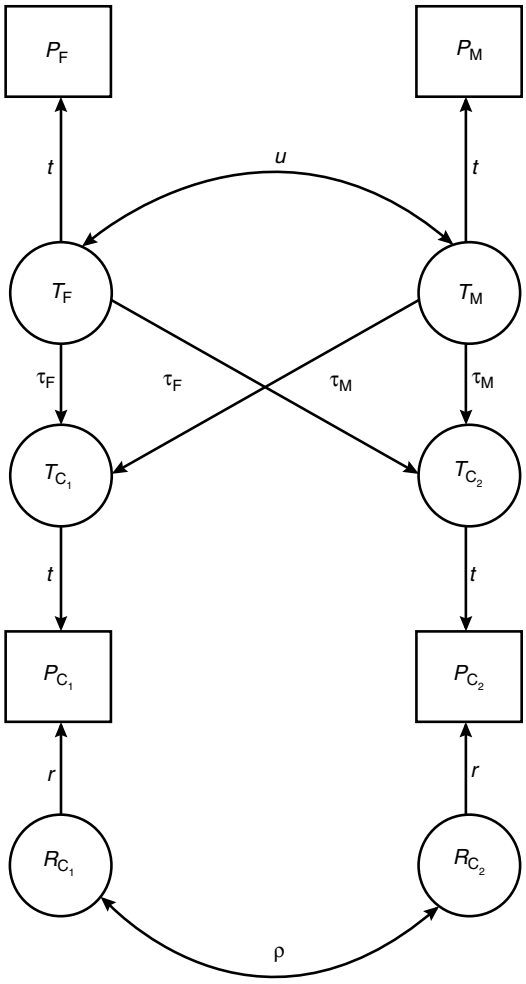

Figure 2 TAU path model of resemblance among relatives in nuclear families. P denotes a phenotype, T indicates transmissible influences, and the subscripts F, M, C_1, and C_2 denote father, mother, and two children, respectively. R denotes nontransmitted residual environment. Path coefficients are defined in the text

Simple Familial Model for Nuclear Families (TAU Model)

In the absence of additional information, the familial components h^2 and c^2 are not resolvable in nuclear families. However, the basic generic model (Figure 1(a)) is useful in this context. This model of familial resemblance, shown in Figure 2, has been termed the TAU model by Rice et al. [37]. The TAU model for nuclear families in Figure 2 depicts a father

Table 1 Expected correlations for nuclear families predicted under the TAU (Figure 2) and BETA (Figure 3) models

Correlation between	TAU model Expected correlations	BETA model Expected correlations
(P_F, P_M)	t^2u	uc^2
(P_F, P_C)	$t^2(\tau_F + u\tau_M)$	$(1/2)h^2 + c^2(f_F + uf_M)$
(P_M, P_C)	$t^2(\tau_M + u\tau_F)$	$(1/2)h^2 + c^2(f_M + uf_F)$
(P_{C_1}, P_{C_2})	$t^2(\tau_M^2 + \tau_F^2 + 2u\tau_M\tau_F) + r^2\rho$	$(1/2)h^2 + c^2\psi$

Note: $\psi = b^2 + f_F^2 + f_M^2 + 2uf_Ff_M$.

(F), mother (M), and two children (C_1 and C_2) yielding four correlations (r_{FM}, r_{FC}, r_{MC}, and $r_{C_1C_2}$). An additional latent residual variable (R) reflecting additional environments shared only by sibs is included in order to account for the possibility that sibling correlations can be higher than parent–offspring correlations. The parameters of the model include: t (the square root of the proportion of phenotypic variance explained by familial factors); u (the correlation between spousal transmissible effects); ρ (the residual sibling correlation); and τ_F and τ_M (the effects of the transmissible factor of fathers and mothers, respectively, on that of a child). Using structural equations or path analysis tracing rules [19], the four correlations can be expressed in terms of the path coefficients, as shown in Table 1. **Maximum likelihood** techniques are then used to estimate the path coefficients and test hypotheses about the model, as discussed later. However, because there is insufficient information in nuclear family data to estimate all five path coefficients, we need to make some additional assumptions. In one case, τ_F may be fixed at the value of 1/2 (expected under polygenic inheritance), and τ_M can differ from 1/2 to allow for maternal influences. Alternatively, if the sibling phenotypic correlation is less than both parent–child correlations, then ρ may be fixed at 0 and both τ_M and τ_F estimated in addition to t and u. Another alternative is the pseudo-polygenic model, where both τ_F and τ_M are fixed to 1/2 and only t, u, and ρ are estimated. It is called pseudo-polygenic because the τs may equal 1/2 due to nongenetic causes as well. Extensions of the TAU model include sex differences in the parameters and alternative sources of spouse resemblance (i.e. social vs. phenotypic homogamy).

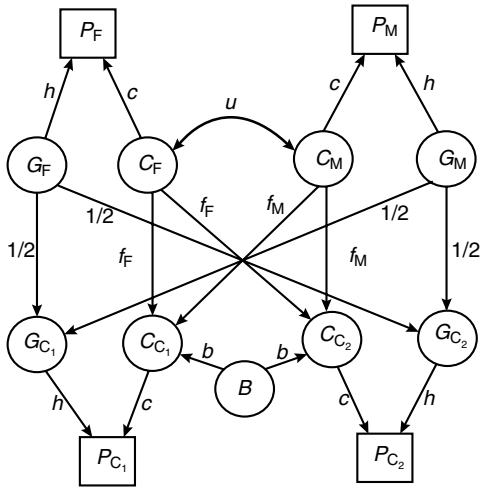

Figure 3 BETA path model of resemblance among relatives in nuclear families. P, G, and C denote phenotype, genotype, and familial environment, respectively. Subscripts F, M, C_1, and C_2 denote father, mother, and two children, respectively. B is the nontransmitted common sibship environment. Path coefficients are defined in the text

Genetic and Familial Environmental Effects (BETA Model)

A more informative model which permits differentiation between the genetic and cultural heritabilities as noted in (7) above is depicted in Figure 3 for nuclear families. This model was termed the BETA model by Rice et al. [37]. As in the specific basic model, each phenotype is assumed to be caused by genes (G), family environments (C), and a residual (R). Also, there is another latent variable (B) reflecting additional shared sibling resemblance. The parameters in the BETA model include: h (square

root of genetic heritability); c (square root of cultural heritability); u (correlation between familial environments of mates); b (additional correlation between sibling familial environments); and f_F and f_M (paternal and maternal, respectively, cultural transmission). Genetic transmission is fixed at 1/2 since each parent transmits half his/her genes to his/her offspring. The correlations expressed in terms of the path coefficients under the BETA model are also given in Table 1 (last column). A more detailed discussion of the BETA model and its theoretical development can be found elsewhere [3, 4, 31–33, 41]. Extensions of the BETA model include intergenerational differences in the genetic and environmental heritabilities (e.g. h^2 in children vs. h^2z^2 in adults), and alternative sources of spouse resemblance.

When nuclear families consisting of fathers, mothers, and children are used, the BETA model is not identified, since there are more unknown parameters (six) than observed correlations (r_{FM}, r_{FC}, r_{MC}, and $r_{C_1C_2}$). There are two basic approaches which have been taken to resolve the BETA model. The first method stays within the basic nuclear family approach and involves introducing additional imputed variables into the model, which generates additional equations and affords resolution of the model. For example, one approach was taken by incorporating the concept of an environmental index measured for each individual in the family, which was assumed to be an imperfect estimate of C [33], thus generating two "observed" variables on each member (the phenotype P and an index I). This provides enormous power for resolution of genetic and familial environmental effects. Unfortunately, unless the indices are based on extensive information, they are known to underestimate the familial environmental effect (c^2) and overestimate the genetic heritability (h^2) [35].

Alternately, extended families or pedigrees provide additional information on other relationships such as uncle–niece, half sibs, and first cousins. While this information helps to fit contemporary path models (such as the BETA model) to analyze the data, it also poses two challenges. First, it is difficult to model cultural transmission adequately for more remote relatives. Secondly, as the degree of relatedness decreases, the information about genetic heritability relies on increasingly larger multiples of small differences in correlations. For these reasons, arbitrarily large pedigrees are not ideal for path analysis. Therefore, when extending nuclear family designs, the benefits associated with increased identifiability should be weighed against the potential error associated with the design. However, nuclear families with twin offspring provide a more appealing design. Several investigators have taken advantage of an even more complex study design involving twins and their spouses and offspring [6, 25].

Simple Twin Model of Family Resemblance

One of the simplest designs capable of resolving genetic and cultural heritabilities is the twin study (*see* **Twin Analysis**). Monozygotic (MZ) twins share all of their genetic material, while dizygotic (DZ) twins are genetically only as similar as full siblings and share half of the **genotype** (*see* **Zygosity Determination**). Assuming that the correlation due to familial environmental effects is identical for both types of twins, the most commonly used estimates of genetic (h^2) and cultural (c^2) heritabilities are: $h^2 = 2(r_{MZ} - r_{DZ})$, and $c^2 = 2r_{DZ} - r_{MZ}$, where r_{MZ} and r_{DZ} are the observed twin correlations estimated from the data. A path diagram of the basic twin model is shown in Figure 4. Each member of the twin pair has his/her their own phenotype (P), genotype (G), and unique or residual environment (R), but the pair shares a common familial environment (C). The additive genetic correlation (path between G_1 and G_2) is 1.0 for MZ and 1/2 for DZ twins. The standardized path coefficients h and c are the square roots of the genetic and cultural heritabilities, and r [derived such that $r^2 = 1 - (h^2 + c^2)$] represents the effect of the unique environment on the phenotype. A powerful (yet rare) design for directly estimating genetic heritability involves the study of MZ twins reared apart (e.g. [22]). Since these twin pairs are assumed to share exactly the same genes and no common environments, the simple MZ twin correlation reflects the genetic heritability.

Adoption Model of Family Resemblance

Adoption studies permit a direct assessment of genetic and cultural heritabilities because, under ideal

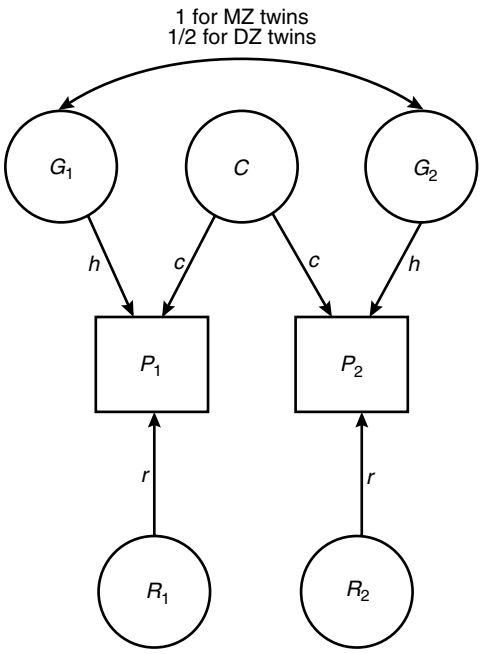

Figure 4 Path model of twin resemblance. P_1 and P_2 are the observed quantitative phenotypes for two members of a twin pair; G_1 and G_2 denote latent additive genetic effects which are correlated (1 for MZ twins and 1/2 for DZ twins); C is the latent familial environment shared by both members of a twin pair; R_1 and R_2 are latent residual environmental influences not shared by the twins. Path coefficients are defined in the text

circumstances, any resemblance between an adopted child and his/her biological parents will be due strictly to genetic effects, and resemblance between an adopted child and adoptive parents (and between adopted siblings) should be attributable strictly to familial environmental effects. In principal, adoption studies are one of the most powerful designs for separating genetic and cultural effects, but in practice full adoption studies are rare, especially those with measurements on the biological father. A noteworthy exception is the Colorado Adoption Project (e.g. [8]). The more common partial adoption studies include those where measures are available only on the adoptee and one set of relatives, usually the adoptive family. While this design allows direct estimation of the cultural effect, critical assumptions regarding selective placement are not testable. An additional crucial assumption is that adoptions take place at (or shortly after) birth so as to preclude any shared environmental resemblance with the biological parents.

Other Developments

Combining multiple family designs into a single approach by using the most informative relationships (e.g. [12]) is one way to overcome the weaknesses of individual approaches. One of the most appealing designs combines nuclear family data with the basic twin design which permits: explicit modeling of complex effects (i.e. separation of genetic and cultural heritabilities); testing critical assumptions (e.g. parent–offspring, sibling, and marital resemblance); and some cross-checks for consistency of the model fit to the data (e.g. see [25]). In one of the more powerful twin-family designs, twins (as well as nontwins) are included as offspring in the nuclear family (e.g. [12]). Another combined approach incorporates data on spouses and offspring of adult twins [5]. Here, the cousins (offspring of MZ twins) are biological half siblings. While these combined designs provide a wealth of information regarding marital resemblance, maternal vs. paternal effects, and sex-linkage, they present difficulties in modeling cultural transmission across the resulting nonnuclear relationships.

Spouse resemblance is a critical aspect in path models (*see* **Assortative Mating**), and if not accounted for can lead to bias in the genetic and cultural heritabilities [24]. Alternate models of spouse resemblance include correlations between the transmissible factors and/or genotypes of mates (called social homogamy; [33]), or between the phenotypes of mates (called phenotypic homogamy), usually modeled as a copath [3]. Models for phenotypic homogamy in the presence of cultural inheritance were developed and studied extensively by Wright [41], Cloninger et al. [4], and Jencks [14]. A path model for generalized assortative mating, called mixed homogamy, was presented by Rao et al. [32], in which both phenotypic homogamy and social homogamy were incorporated. However, distinguishing between these sources of marital resemblance is difficult using only nuclear family data, and additional information such as twins [6] or multiple measurements in the spouses are needed in order to identify the model. Heath & Eaves [11] presented a model involving MZ and DZ twin pairs, their spouses, and offspring which allows one to

estimate all possible sources of resemblance between mates and to test alternative hypotheses.

Temporal and developmental trends have also been explored (e.g. [13, 27, 28, 7], and [30]). Models with temporal changes in genetic and cultural heritability that give rise to changes in family resemblance have been developed for family data (e.g. [13, 27], and [28]) twin data (e.g. [7]) and adoption data (e.g. [26]), and some of these will be discussed later.

It must be clear from the preceding discussion of familial models that we cannot fit any given model to any type of data set. A given model requires certain type(s) of data. For example, even the simple TAU model cannot be fitted to nuclear family data without making additional assumptions. On the other hand, data on MZ and DZ twins alone can resolve genetic (h^2) and familial environmental (c^2) components under certain assumptions. The type of data at hand determines what kinds of questions can be answered by a study.

Statistical Analysis

When dealing with family data of one type or another, the method of choice is maximum likelihood under the assumption that all the observed variables within a family jointly follow a **multivariate normal distribution**. To illustrate the method, let us consider a random sample of nuclear families with variable sibship sizes. Let $\mathbf{X}' = (X_{11}, X_{21}, X_{31}, X_{32}, \ldots, X_{3s})$ denote the row vector of phenotypes for the father (X_{11}), mother (X_{21}), and s children (X_{3k}, $k = 1, \ldots, s$). Assuming that the means and variances are equal for all children, the vector of means ($\boldsymbol{\mu}'$) of $\mathbf{X}'$ is given by (μ_1, μ_2, μ_3) for the father, mother, and offspring, and the corresponding variances are σ_1^2, σ_2^2, and σ_3^2. A correlation matrix ($\mathbf{R}$) is defined as the one that consists of the intercorrelations among all variables in $\mathbf{X}'$. Depending on the model, many of the elements of the correlation matrix can be further equated. For example, in the nuclear family TAU model, there are four independent correlations:

$$\begin{array}{lll} \text{spouse} & \rho(X_{11}, X_{21}), & \\ \text{father–child} & \rho(X_{11}, X_{3k}), & k = 1, \ldots, s, \\ \text{mother–child} & \rho(X_{21}, X_{3k}), & k = 1, \ldots, s, \\ \text{sibling} & \rho(X_{3k}, X_{3l}), & k \neq l = 1, \ldots, s. \end{array} \tag{8}$$

In path analysis, these correlations are expressed as functions of the path coefficients of the model (see Table 1). The covariance matrix ($\boldsymbol{\Sigma}$) is defined in terms of the correlations (or path coefficients) and variances ($\boldsymbol{\Sigma} = \boldsymbol{\sigma}'\mathbf{R}\boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is a diagonal matrix of standard deviations).

Assuming that the data vector $\mathbf{X}$ for a given family follows a multivariate normal distribution, the log-likelihood function for the family is given by

$$\ln L = -\left(\frac{1}{2}\right)[\ln|\boldsymbol{\Sigma}| + (\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})] + \text{constant}, \tag{9}$$

where $|\boldsymbol{\Sigma}|$ is the determinant and $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix $\boldsymbol{\Sigma}$. The total log-likelihood function for a sample is derived by summing across the families. Because the likelihood function is computed separately for each family and the model is fitted directly to the data, a fixed family structure is not required. Therefore, missing data are allowed, as are variable sibship sizes. Parameters of the model (path coefficients, means, and variances) are estimated simultaneously by maximizing $\ln L$. Maximization routines commonly used in genetic epidemiology include GEMINI and ALMINI [17]. Tests of hypotheses are conducted by comparing the log likelihoods across different models. The **likelihood ratio test**, which is the difference between $-2\ln L$ with $k + w$ parameters estimated and $-2\ln L$ when only k of the parameters are estimated, is asymptotically distributed as a $\boldsymbol{\chi}^{\mathbf{2}}$ with w degrees of freedom. The only requirements for this asymptotic distribution are that the reduced model (with only k parameters) must be nested within the more general model (with $k + w$ parameters) and the reduced model must not be on the boundary of the more general model (*see* **Generalized Linear Model**).

The familial sources of variation may also be estimated as variance components (*see* **Genetic Correlations and Covariances**). The maximum likelihood methods described here are formally equivalent for both variance components analysis and path analysis. The advantage of path analysis is that the model is schematically represented, and all assumptions are obvious from examining the diagram. Several computer programs for estimating familial effects using family data are currently available (*see* **Software for Genetic Epidemiology**).

Modeling Assumptions

Implicit in the path models are several assumptions. Linearity and additivity (*see* **Additive Model**) of effects represent two fundamental assumptions. Choice of scale is important for these assumptions to hold, and such a scale may not exist for some variables. Secondary assumptions include the absence of gene–gene interactions (dominance and epistasis (*see* **Genotype**), and **gene–environment interactions**). It is also assumed that the genes and familial environment (partly) determine the phenotype, and not vice versa, and that parental genes and environments determine those of the offspring. While these assumptions may be reasonable for many phenotypes, in some situations they may not be. For example, the phenotype may temporally influence the environment, such as when high blood pressure may modify a person's subsequent lifestyle. Additional assumptions concern the resemblance between parents (*see* **Assortative Mating**). We generally assume that spouses resemble each another only due to their correlated environment (u). However, if spouse selection is based in part on the phenotype of interest, and if there is a partial genetic determination for that trait, then the correlations between parents may also include a genetic component (mixed homogamy). Violations of any of these modeling assumptions can give rise to errors in inference [24].

Maximum likelihood methods require certain distributional assumptions. For example, the assumption of multivariate normality is critical, especially for hypothesis testing which is sensitive to moderate departures from normality [23]. Deviations from normality can be controlled using suitable data **transformations**. Although transformations do not guarantee multivariate normality, they tend to minimize the impact [31].

Nonrandom Samples

The maximum likelihood techniques described for random samples can be used to analyze data that have been collected under an **ascertainment** scheme that provides a selected sample, if adjustment for ascertainment is incorporated into the likelihood function. One method of adjusting for single ascertainment through a proband is by conditioning the likelihood function for a particular family on the phenotypic value of the proband [1]. The conditional likelihood for a family given the proband's phenotype value is then:

$$L(\mathbf{X}|X_{\mathrm{p}}) = \frac{f_k(\mathbf{X})}{f_1(X_{\mathrm{p}})}, \tag{10}$$

where $\mathbf{X}$ is the $(k+1)$-dimensional vector of data for a family, X_{p} is the proband's phenotypic value, and f_k is the k-variate normal density function. This conditional likelihood approach is generally used for any sample of families each of which is ascertained through a single proband, and for that reason is termed the *generic* method [36]. If a particular variable (such as the father's phenotype) is the variable on which selection occurs in every family, then there is no information for estimating the mean and variance of that variable and these parameters must be estimated from other sources.

Additional information regarding ascertainment can be used to increase the efficiency of **estimation** and **power** of hypothesis tests [36]. For example, there are different methods to correct for ascertainment when probands are selected by truncation on the distribution of a phenotype or a correlated trait, such that families of probands whose selection variable lies above a defined threshold are included in the sample (otherwise they are excluded). Provided the region of truncation is known and adherence to the sampling scheme is maintained, then *specific* methods to correct for ascertainment are available and provide more efficient parameter estimates and more powerful hypothesis tests. If the region of truncation is not known and must be estimated, then the generic method and the specific method are approximately equal in efficiency and power. The generic method is computationally simpler and easier to implement and is relatively robust (*see* **Robustness**) to departures from adherence to the truncation sampling scheme.

Another method for adjusting for nonrandom ascertainment was developed by Hanis & Chakraborty [10]. This method yields accurate estimates of parameters for most types of nonrandom sampling, and is robust against departures from multivariate normality. However, it is computationally intensive and provides only approximate tests of hypotheses, and while it can be theoretically applied to multiple ascertainment schemes, the sample sizes required for implementation become impracticably large [34].

Temporal/Developmental Trends

Three kinds of temporal trends may be distinguished for quantitative traits. First, the mean of the population can change systematically over time. These mean effects are relatively easy to detect, and the clinical relevance of such trends quantify how the average individual in the population will change over time.

The second kind of temporal trend occurs when the variance changes over time/age. Such heteroscedasticity with time, a second-order effect, is more difficult to diagnose and is often viewed as a nuisance that must be dealt with rather than an effect of primary interest. Temporal changes in the mean and variance can be dealt with outside of familial models simply by analyzing the residual phenotype after adjusting for age in both the mean and variance.

The third kind of temporal trend manifests in the correlations (or covariances) between related individuals. These trends are more complex because two ages are involved (one for each individual), and are both interesting and important to investigate from a genetic epidemiological point of view. It is natural to ask whether the familial components within the path model framework also vary as a function of time. The presence of secular trends (e.g. increased obesity in the population over time) could be manifesting as intergenerational differences in genetic or cultural heritabilities. Developmental trends, as for example growth spurts characteristic in adolescence before attaining more stable adult values, also may be a function of intergenerational differences in the genetic or cultural heritabilities.

Several models take the approach of assuming a particular temporal structure of the genotype (and environment) that gives rise to the temporal changes in family resemblance [2, 7, 13, 18]. For example, Eaves et al. [7] presented a developmental model that includes one constant pool of genes acting with constant effect throughout all time, and at each time point a specific set of genes that is active at that time and no other. Such a model predicts exponential trends in overall heritability. Boomsma & Molenaar [2] model the genotype (and environment) as a first-order autoregressive process (*see* **ARMA and ARIMA Models**), with each pool of genes depending only on the previous time's pool. This model gives rise to a simplex temporal structure. Each of these models make assumptions about the underlying genetic mechanisms and therefore anticipate a particular kind of temporal change. Other models [27, 28, 30] make no assumptions about the gene action. There is a myriad of alternative genetic mechanisms that can cause temporal variability. For example, there can be a variable lag time between gene action and observed product, or more than one common set of genes acting independently over time, or pools of genes acting with one magnitude of effect at one time and a different magnitude at another time, or environmental triggers of gene action, etc.

A simple extension of path models that incorporates time-dependent effects using cross-sectional family data may be considered for analyzing phenotypes that warrant temporal effects. For example, the basic structural equation in (7) may be extended so that each parameter is made a function of the individual's age (A):

$$P(A) = h(A)G + c(A)C + r(A)R. \qquad (11)$$

It is possible to describe the temporal variability in $h(A)$ and $c(A)$ as continuous functions of age and the two basic parameters at birth (h and c). An important feature of this extension is that it enables testing null hypotheses of alternate forms of temporal trends as well as no temporal variation at all. The simple TAU model has been expanded to incorporate temporal trends in the familial correlations using cross-sectional data [27]. The path coefficients are functions of an individual's age [i.e. $t(A)$ and $r(A)$], and the residual sibling correlation is a function of the absolute age difference between the sibs. The effects of parental transmissible factors on those of their children (τ_F, τ_M) are the same for each child, regardless of age, and the spouse correlation is also assumed to be a constant. The functional forms available for modeling the age trends are chosen such that virtually any shape of trend can be modeled. For example, it is possible to choose a function that permits the **heritability** to increase or decrease monotonically, or to increase (decrease) up to a certain age and then decrease (increase) again to a different value. A detailed discussion of the available functional forms and the justification for their parameterization can be found elsewhere [27].

A restriction in this model based on cross-sectional family data is that the genes, habits, etc. that parents give to their children are constant in their effects over time, although all individuals express these factors in the phenotype to varying degrees as a function of their age at the time [$t(A)$]. However, temporal

models based on longitudinal or repeated measures family data (e.g. [30]) allow for temporal changes in several components, including the mean and variance functions, the transmissibility, the marital residual correlation (which is a function of the cohabitation time), and the sibling residual correlation. In the repeated measures temporal model, the familial and nonfamilial components are modeled as a correlation between the residuals of a pair of repeated measures on the same individual. The later component exists even if the phenotype is nonfamilial. Another feature is that with repeated measures the model is completely resolvable in nuclear families, whereas the classic TAU model and the cross-sectional trend extension are not fully identified.

Multivariate Models

A simple approach for addressing multivariate questions using only univariate path models involves preadjusting one variable for the effects of a second variable using **multiple linear regression**. The residual score from the regression represents the independent effects of the first variable after those from the second have been removed. Comparison of the univariate results prior to and after the adjustment suggests whether the same or different familial factors underlie each of the traits. While this approach is relatively simple to perform, a more informative approach is to analyze correlated variables simultaneously with multivariate methods. Although increased computation is associated with **multivariate analysis**, there are several advantages, including greater power to detect effects, and the ability to resolve bivariate correlations into genetic and familial environmental effects.

Formal multivariate path analysis decomposes the phenotypic covariance of two or more traits into additive genetic and familial environmental correlations. Figure 5 illustrates the path diagram for two correlated traits P_1 and P_2 for the simple case of uncorrelated G and C in an individual. The expected correlation between P_1 and P_2 is

$$r_{P_1P_2} = h_1 r_G h_2 + c_1 r_C c_2, \tag{12}$$

where $h_1 r_G h_2$ is the standardized genetic covariance and $c_1 r_C c_2$ is the standardized environmental covariance. This expression reduces to $h^2 + c^2$ when $P_1 = P_2$.

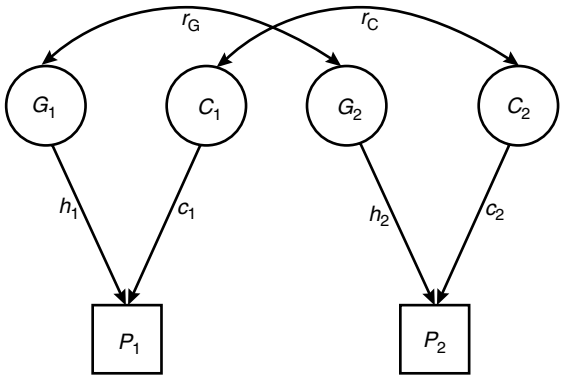

Figure 5 Path model of genetic (G) and familial environmental (C) influences on two correlated phenotypes (P_1 and P_2) measured on the same individual

In the context of path analysis of multivariate family data, the basic structural model is invoked repeatedly for each phenotype, so that the system of linear structural equations defining a multivariate system is given by ($P_i = h_i G_i + c_i C_i, i = 1, \ldots, k$) which can be represented in general matrix formulation simply as

$$\mathbf{P} = \mathbf{hG} + \mathbf{cC}, \tag{13}$$

where $\mathbf{P}$, $\mathbf{G}$, and $\mathbf{C}$ represent ($k \times 1$) column vectors of phenotypic, genetic, and environmental standardized deviations from the mean, respectively; $\boldsymbol{h}$ represents a ($k \times k$) diagonal matrix containing the square roots of the genetic heritabilities of the k traits; and $\mathbf{c}$ is a ($k \times k$) diagonal matrix containing the square roots of the proportions of phenotypic variance that are attributable to the environment.

The structural equation is used to derive the matrix of expected covariances within individuals, or the phenotypic correlation matrix $\mathbf{R}_P$:

$$\mathrm{E}(\mathbf{R}_P) = \mathbf{hR}_G\mathbf{h}' + \mathbf{cR}_C\mathbf{c}' + \mathbf{hsc}' + \mathbf{cs}'\mathbf{h}', \tag{14}$$

where $\mathbf{R}_G$ is the matrix of correlations among the genotypes for each variable within an individual, $\mathbf{R}_C$ is the environmental correlation matrix, and $\mathbf{s}$ is a full, nonsymmetric matrix consisting of the genotype–environment correlations among all measures.

In the context of a family design, the algebraic derivations for expected correlations among relatives are essentially obtained by using the univariate

model. However, each variable in the diagram represents not a single measure, but a $(k \times 1)$ column vector of measures, and path coefficients represent $(k \times k)$ diagonal matrices of path coefficients. Tracing rules for multivariate diagrams is straightforward, consisting of the conventional path analysis rules [19] with extensions developed by Vogler [39].

There is little difference between the analysis of multiple phenotypes measured simultaneously and the analysis of the same character in an individual at different points in time. When multiple measures are obtained at discrete points that are identical for all individuals, the multivariate methods presented here can be applied directly to developmental or longitudinal data.

Combined Path Analysis

Many traits are multifactorial and complex in the sense that there are a number of genetic and/or nongenetic determinants. For each of these traits it is highly unlikely that anything close to "the gene" will ever be discovered that explains most of the variation for that phenotype and solves the whole puzzle. Instead, the genetic component of many of these traits is expected to be in the oligogenic (few genes) to polygenic (many genes) range. What we know so far leads us to expect an elaborate interplay between many factors in the development of the phenotypes, including gene–gene and gene–environment interactions. While much analytic progress can be made using the standard genetic tools of path analysis, **segregation analysis**, and **linkage analysis**, a combined approach is potentially more powerful and holds great promise for these traits. Indeed, a combined approach may be the only way to disentangle the interplay of the multiple underlying processes. The utility of this approach has been demonstrated through combined segregation and path analysis.

The general model [29] includes both a segregation component (g) and a multifactorial path model component (m), as well as fixed effects (f), so that the phenotype (P), univariate or multivariate, is determined by

$$P = g + m + f + r, \qquad (15)$$

where r is the residual. In this formulation g can denote any number of major genes, with whatever pleiotropic effects they may have on the (multivariate) phenotype(s), and m denotes an arbitrary path model so that m may be a linear expression with many terms containing multiple heritable components. Likewise, f represents any number of fixed covariate effects (e.g. sex, age, measured genotypes). Eq. (15) defines a set of regression equations for each person in a pedigree that are then linked via both m and g. **Interactions** among effects (e.g. $g \times f$) can be included in the model. No distributional assumptions need be made about observable fixed effects such as sex. The segregation portion of the model is an independent additive term in the regression equations, above and beyond any specified by the path analysis part of the combined model. For models with a segregation component, the log likelihood of each pedigree is partitioned as a sum of conditional probabilities over every possible latent genotypic vector. With large numbers of subjects per pedigree, however, an exact likelihood formulation can be impracticable to compute. Finally, considerable progress has been made in combining path and regressive models by extending the regressive models to incorporate the simple TAU model [21] as well as the BETA [20] model.

Conclusion

Going from genes and their products to the expression of normal variation involves a leap of faith, and complex statistical procedures are often necessary to demonstrate a relationship between the two. Here we have discussed one such procedure (path analysis) that attempts to relate the observations to the underlying determinants.

Important considerations include alternate family designs (e.g. families, twins, adoptions) as well as alternate models of family resemblance (e.g. univariate, multivariate, TAU or BETA, marital resemblance, temporal trends). The increasing availability of computer programs that empower the genetic epidemiologist to create appropriate models to fit their data, instead of forcing their data to fit existing models, is particularly helpful in testing alternate modes of genetic and cultural transmission. The relatively new approach of combining models (e.g. path and segregation analysis) should yield new insights into determining how the multiplicity of causes underlying a phenotype are interrelated. Figure 6 illustrates a possible scenario for complex phenotypes

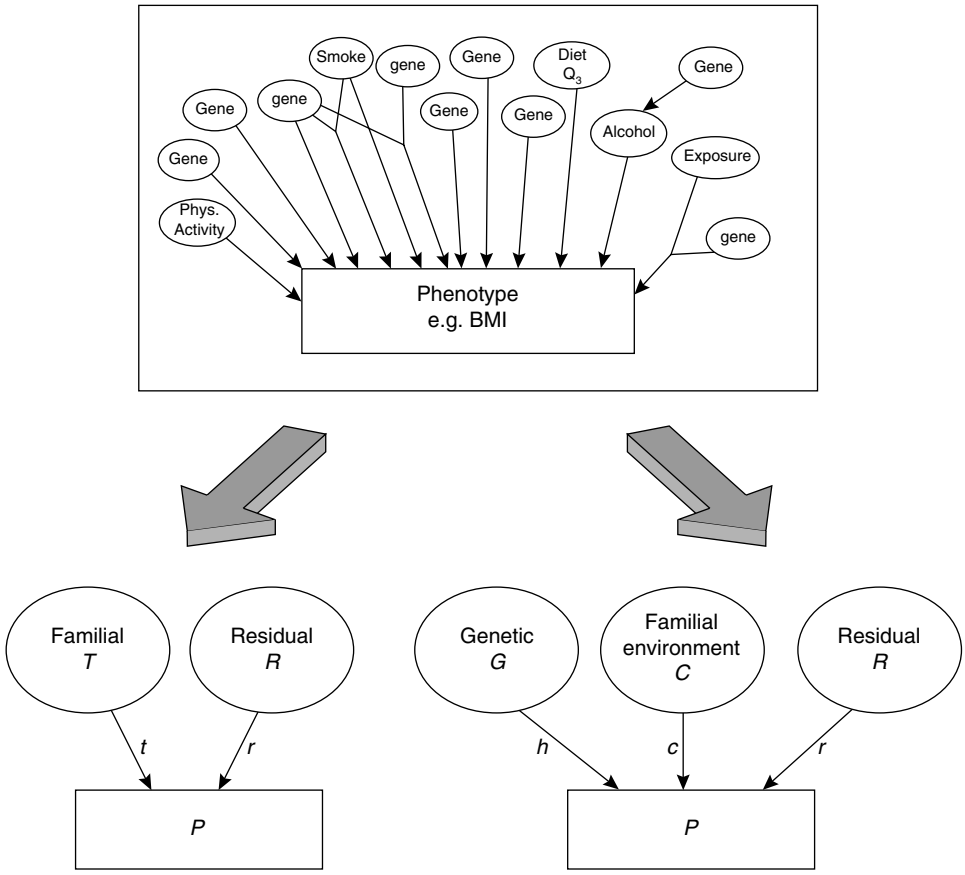

Figure 6 Hypothetical model of the underlying genetic and environmental factors giving rise to a complex phenotype. Modeling approximations are shown at the bottom

using body mass index (BMI) as an example. The figure illustrates how complex phenotypes are determined by multiple genes, multiple environments, and interactions among them. The complicated reality is often approximated by simple and feasible models, as shown at the bottom of Figure 6. Path analysis offers a promising method for disentangling these underlying causes by formulating simple refutable hypotheses.

References

[1] Boehnke, M. & Lange, K. (1984). Ascertainment and goodness of fit of variance components for pedigree data, in *Genetic Epidemiology of Coronary Heart Disease: Past, Present and Future*, D.C. Rao., R.C. Elston, L.H. Kuller, M. Feinleib, C. Carter & R. Havlik, eds. Liss, New York, pp. 173–192.

[2] Boomsma, D.I. & Molenaar, P.C.M. (1987). The genetic analysis of repeated measures. I. Simplex models, *Behavior Genetics* **17**, 111–123.

[3] Cloninger, C.R. (1980). Interpretation of intrinsic and extrinsic structural relations by path analysis: theory and applications to assortative mating, *Genetical Research* **36**, 133–145.

[4] Cloninger, C.R., Rice, J. & Reich, T. (1979). Multifactorial inheritance with cultural transmission and assortative mating. II. A general model of combined polygenic and cultural inheritance, *American Journal of Human Genetics* **31**, 176–198.

[5] Corey, L.A. & Nance, W.E. (1978). The monozygotic half-sib model: a tool for genetic epidemiological research, *Progress in Clinical and Biological Research* **24A**, 201–209.

[6] Eaves, L.J., Fulker, D.W. & Heath, A.C. (1989). The effects of social homogamy and cultural inheritance on the covariances of twins and their parents: a LISREL model, *Behavior Genetics* **19**, 113–122.

[7] Eaves, L.J., Long, J. & Heath, A.C. (1986). A theory of developmental change in quantitative phenotypes applied to cognitive development, *Behavior Genetics* **16**, 143–162.

[8] Fulker, D.W. & DeFries, J.C. (1983). Genetic and environmental transmission in the Colorado Adoption Project: path analysis, *British Journal of Mathematical and Statistical Psychology* **36**, 175–188.

[9] Goldberger, A.S. & Duncan, O.D. (1973). *Structural Equation Models in the Social Sciences*. Seminar Press, New York.

[10] Hanis, C.L. & Chakraborty, R. (1984). Nonrandom sampling in human genetics: familial correlations, *IMA Journal of Mathematics Applied in Medicine and Biology* **1**, 193–213.

[11] Heath, A.C. & Eaves, L.J. (1985). Resolving the effects of phenotype and social background on mate selection, *Behavior Genetics* **15**, 15–30.

[12] Heath, A.C., Kendler, K.S. Eaves, L.J. & Markell, D. (1985). The resolution of cultural and biological inheritance: informativeness of different relationships, *Behavior Genetics* **15**, 439–465.

[13] Hopper, J.L. & Mathews, J.D. (1982). Extensions to multivariate normal models for pedigree analysis, *Annals of Human Genetics* **46**, 373–383.

[14] Jencks, C. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books, New York.

[15] Kang, K. & Seneta, E. (1980). Path analysis: an exposition, in *Developments in Statistics*, Vol. 3, P.R. Krishnaia, ed. Academic Press, New York, pp. 217–246.

[16] Khoury, M.J., Beaty, T.H. & Cohen, B.H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press, New York, pp. 220–232.

[17] Lalouel, J.M. (1979). GEMINI – a computer program for optimization of general nonlinear functions. Department of Medical Biophysics and Computing, *Technical Report 14*, Salt Lake City.

[18] Lange, K. (1986). Cohabitation, convergence and environmental covariances, *American Journal of Medical Genetics* **24**, 483–491.

[19] Li, C.C. (1975). *Path Analysis: A Primer*. Boxwood, Pacific Grove.

[20] Li, Z., Bonney, G.E. & Rao, D.C. (1994). Genetic analysis combining path analysis with regressive models: the BETA path model of polygenic and familial environmental transmission, *Genetic Epidemiology* **11**, 431–442.

[21] Li, Z., Bonney, G.E. Lathrop, G.M. & Rao, D.C. (1994). Genetic analysis combining path analysis with regressive models: the TAU model of multifactorial transmission, *Human Heredity* **44**, 305–311.

[22] Loehlin, J.C. (1978). Identical twins reared apart and other routes to the same destination, *Progress in Clinical and Biological Research* **24A**, 69–77.

[23] McGue, M., Wette, R. & Rao, D.C. (1987). A Monte Carlo evaluation of three statistical methods used in path analysis, *Genetic Epidemiology* **4**, 129–155.

[24] McGue, M., Wette, R. & Rao, D.C. (1989). Path analysis under generalized marital resemblance: evaluation of the assumptions underlying the mixed homogamy model by the Monte Carlo method, *Genetic Epidemiology* **6**, 373–388.

[25] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer, Boston.

[26] Phillips, K. & Fulker, D.W. (1989). Quantitative genetic analysis of longitudinal trends in adoption designs with application to IQ in the Colorado Adoption Project, *Behavior Genetics* **19**, 621–658.

[27] Province, M.A. & Rao, D.C. (1985). Path analysis of family resemblance with temporal trends: applications to height, weight, and Quetelet Index in Northeastern Brazil, *American Journal of Human Genetics* **37**, 178–192.

[28] Province, M.A. & Rao, D.C. (1985). A new model for the resolution of cultural and biological inheritance in the presence of temporal trends: application to systolic blood pressure, *Genetic Epidemiology* **2**, 363–374.

[29] Province, M.A. & Rao, D.C. (1995). General purpose model and a computer program for combined segregation and path analysis (SEGPATH): automatically creating computer programs from symbolic language model specifications, *Genetic Epidemiology* **12**, 203–219.

[30] Province, M.A., Tishler, P. & Rao, D.C. (1989). Repeated-measures model for the investigation of temporal trends using longitudinal family studies: application to systolic blood pressure, *Genetic Epidemiology* **6**, 333–347.

[31] Rao, D.C., McGue, M., Wette, R. & Glueck, C.J. (1984). Path analysis in genetic epidemiology, in *Human Population Genetics: The Pittsburgh Symposium*, A. Chakravarti, ed. Van Nostrand Reinhold, New York, pp. 35–81.

[32] Rao, D.C., Morton, N.E. & Cloninger, C.R. (1979). Path analysis under generalized assortative mating. I. Theory, *Genetical Research* **33**, 175–188.

[33] Rao, D.C., Morton, N.E. & Yee, S. (1974). Analysis of family resemblance. II. A linear model for familial correlation, *American Journal of Human Genetics* **26**, 331–359.

[34] Rao, D.C. & Wette, R. (1989). Nonrandom sampling in genetic epidemiology: an implementation of the Hanis-Chakraborty Method for multifactorial analysis, *Genetic Epidemiology* **6**, 461–470.

[35] Rao, D.C. & Wette, R. (1990). Environmental index in genetic epidemiology: an investigation of its role, adequacy, and limitations, *American Journal of Human Genetics* **46**, 168–178.

[36] Rao, D.C., Wette, R. & Ewens, W.J. (1988). Multifactorial analysis of family data ascertained through truncation: a comparative evaluation of two methods of statistical inference, *American Journal of Human Genetics* **42**, 506–515.

[37] Rice, J., Cloninger, C.R. & Reich, T. (1978). Multifactorial inheritance with cultural transmission and assortative mating. I. Description and basic properties of the

unitary models, *American Journal of Human Genetics* **30**, 618–643.

[38] Van Eerdewegh, P. (1982). Statistical selection in multivariate systems with applications in quantitative genetics, *Unpublished doctoral dissertation*. Washington University, St Louis.

[39] Vogler, G.P. (1985). Multivariate path analysis of family resemblance, *Genetic Epidemiology* **2**, 35–53.

[40] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.

[41] Wright, S. (1978). *Evolution and the Genetics of Populations:* Vol. 4. *Variability Within and Among Natural Populations*. University of Chicago Press, Chicago.

(*See also* **Familial Correlations**)

D.C. RAO & TREVA RICE

Path Analysis

Path analysis is a statistical methodology that begins with a model. Each model consists of a system of equations and assumptions about the relationships between a set of variables. The variables can be *observed, latent* (unobserved), or *disturbances* (errors). In some applications the observed variables are nonstochastic, but in nearly all models the latent and disturbance variables are **random variables**. The hypothesized model originates with the researcher. Path analysis provides an **algorithm** for understanding the direct, indirect, and total effect of one variable on another in this hypothesized structure. It also allows a test of whether the hypothesized model is consistent with the covariances **(correlations)** of observed variables. A lack of fit (*see* **Goodness of Fit**) is an indicator of an error in model specification (*see* **Misspecification**). A "good" fit lends plausibility to the model, but it does not rule out the possibility of other models with as good, or superior, fits.

Path analysis was invented by Sewall Wright in graduate school. His first publication using the method was in 1918 [12]. One of the most common misunderstandings about path analysis is the belief that it is a tool to "discover" causal relationships (*see* **Causation**). This critique accompanied the introduction of path analysis and periodically appears in the contemporary literature. Wright explicitly denies this in several places: ". . . the method of path coefficients is not intended to accomplish the impossible task of deducing causal relations from the values of the correlation coefficients" [15, p. 193]. Instead, Wright describes path analysis as a method that

> depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them [13, p. 557].

Thus, path analysis is best seen as a method to test the implications of a given model structure.

The primary components of path analysis are: (i) the path diagram, (ii) the estimation of path coefficients, and (iii) the decomposition of effects. The next sections provide details on each.

Path Diagrams

Path diagrams are *pictorial or graphical representations of a system of equations and assumptions* (*see* **Graphical Displays**). The diagrams are most useful for the analysis of two or more equations that link observed or latent variables and disturbances or errors to each other in a system. The path diagram is easy to interpret and it readily reveals relationships that might be missed if viewing only the equations of a model. Furthermore, the tracing of effects across variables also is made easier with the diagrams.

Figure 1 contains standard and some nonstandard notational conventions in path analysis. A basic distinction between the variables is whether they are latent (unobserved), observed, or disturbance (error) variables. Each observed variable is enclosed in a box. Latent variables appear in ellipses or ovals. Disturbances are not enclosed in either, although some researchers place disturbances in ovals to signify that they are latent.

Other symbols are best illustrated by way of examples. Figure 2(a) is the path diagram for a multiple **regression** equation with four **explanatory variables** and one dependent **response variable**. The straight line with multiple arrow heads coming down from it and pointing to the X_1 to X_4 variables indicates that each pair of variables is associated. The reason for their association is not explained by the model. In the language of econometrics, these are correlated exogenous variables.

Each straight single headed arrow signifies the influence of an X on Y_1. The coefficients corresponding to these paths are the *path coefficients*. Originally, Wright used path coefficients that were equivalent to what are commonly called standardized regression coefficients. The standardized coefficient is the product of the unstandardized coefficient times the ratio of the standard deviation of the explanatory variable to the standard deviation of the "dependent" variable. Now it also is common to use unstandardized regression coefficients in path diagrams. The final symbol in Figure 2(a) is the disturbance, ζ_1, that summarizes the other influences on Y_1 besides those of the Xs. Exogenous variables or disturbances (errors) that are not connected by two-headed arrows are treated as having no linear association. Thus, the lack of two-headed arrows linking ζ_1 and the Xs represents the usual multiple regression assumption that the disturbance is uncorrelated with the explanatory variables.

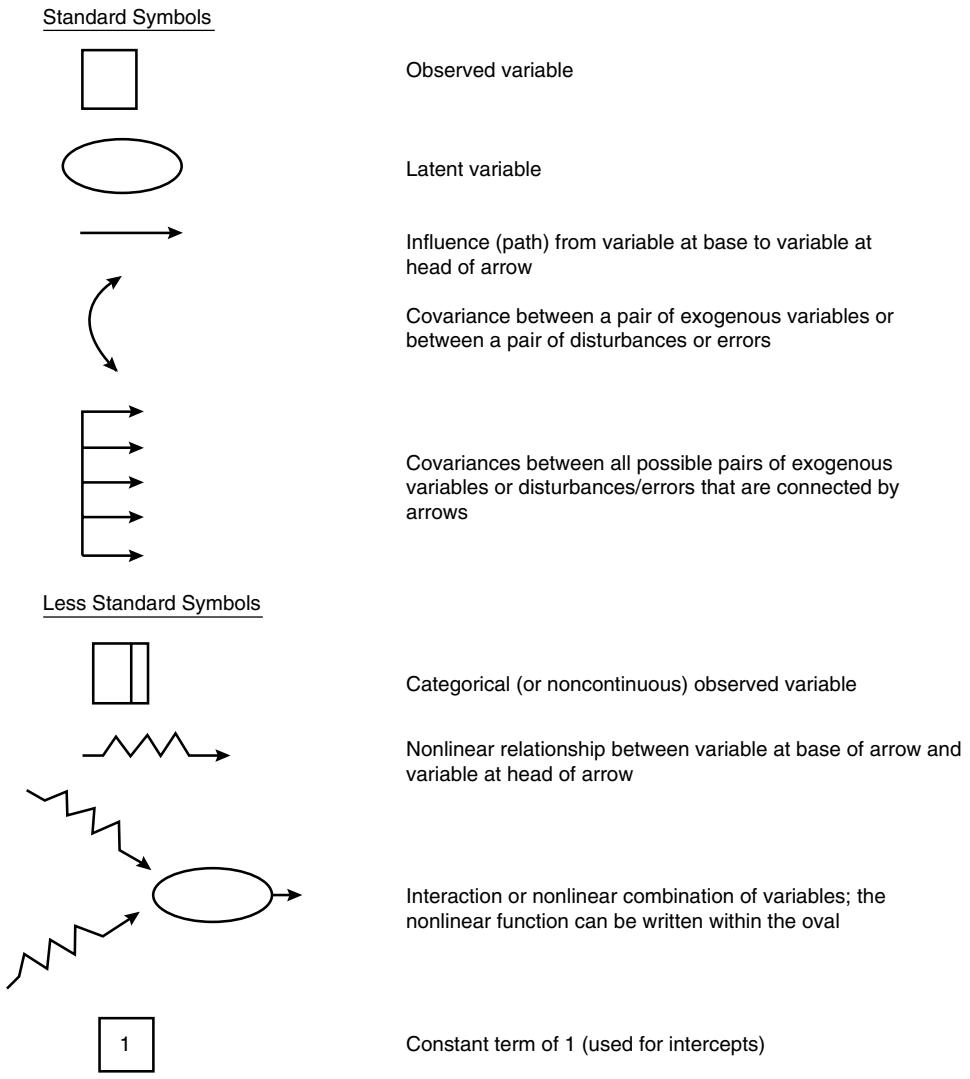

Figure 1 Symbols for path analysis

Figure 2(b) contains an example of a simultaneous equations model. The equations that correspond to this path diagram are:

$$Y_1 = \alpha_1 + \gamma_{11} X_1 + \zeta_1,$$
$$Y_2 = \alpha_2 + \beta_{21} Y_1 + \gamma_{22} X_2 + \zeta_2,$$
$$Y_3 = \alpha_3 + \beta_{31} Y_1 + \beta_{32} Y_2 + \zeta_3,$$
$$\text{cov}(x_k, \zeta_i), \qquad \text{cov}(\zeta_i, \zeta_j) \neq 0,$$
$$\text{E}(\zeta_i) = 0, \qquad i, j = 1, 2, 3 \text{ for } i \neq j, k = 1, 2.$$

The β and γ coefficients correspond to the straight, single-headed arrows. Sometimes these coefficients are placed on the arrows in the diagram. The αs are equation intercepts. They often do not appear in a path diagram to avoid cluttering the figure. The curved arrows indicate correlations among the exogenous variables. These are not explained in the model. Although econometricians rarely use path diagrams, they have considerable experience with simultaneous equation models such as this.

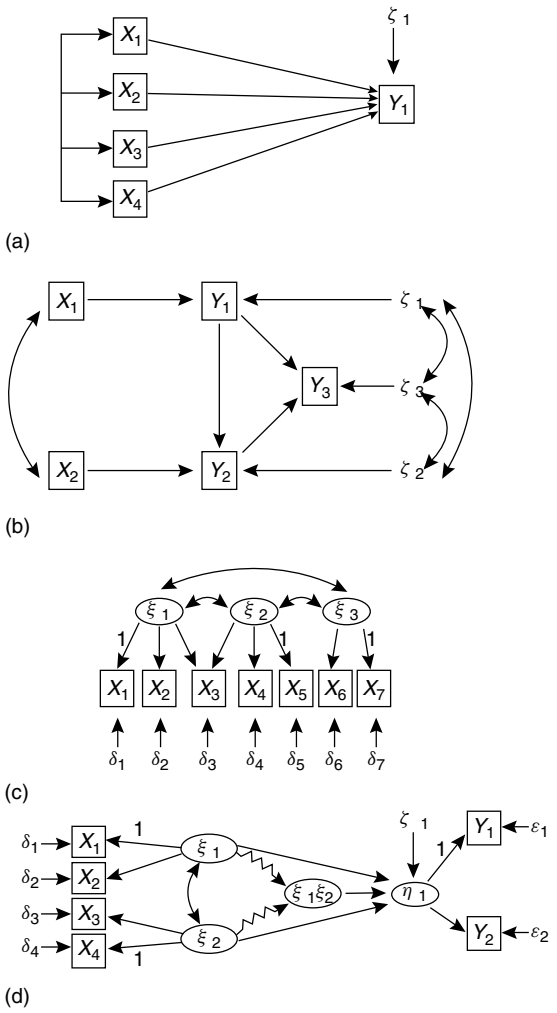

Figure 2 Examples of path diagrams. (a) Multiple regression; (b) simultaneous equation model; (c) confirmatory factor analysis; (d) general model with product interaction of latent variables

A factor analysis example is given in Figure 2(c). The three latent variables (the ξs) are in ovals. The indicators of the latent variables are the X variables that are in boxes. Errors of measurement (δs) are unenclosed. The lack of two-headed arrows connecting them to each other or to the latent variables signifies the assumption that these variables are uncorrelated.

The final illustration of a path diagram is a more complicated general model [Figure 2(d)]. This model has both a latent variable structure that allows the latent variables to influence each other and it has a measurement model linking the measures to the latent variables that influence them. In addition, this model includes a product interaction of latent variables. The nonlinear function of the latent variables that forms the interaction term is inside the middle oval in the diagram. The nonlinear linkage of variables is diagrammed with a saw-toothed, single-headed arrow from its components to the product latent variable.

Researchers can construct a wide variety of other path diagrams from the basic symbols given in Figure 1. In all cases the diagram is just an alternative to a system of equations and assumptions that show the relationship between variables in a model.

Estimation of Model Parameters

Wright originally used the path diagrams as an aid to writing the variances and covariances of the observed variables in terms of the parameters of the model. The parameters included the coefficients in the equations as well as the variances and covariances of the exogenous variables and disturbances (errors). He then would solve for the model parameters in terms of the variances and covariances of the observed variables. Contemporary estimation procedures are more sophisticated and use **maximum likelihood** and **least squares** methods (*see* **Structural Equation Models** for more details).

Decomposition of Effects

Path analysis makes a distinction between the direct, indirect, and total effects of one variable on another. The direct effect is the influence of one variable on another that is not mediated by any other variable that is part of the model. It corresponds to the coefficient of a variable in the original structural form of an equation. The indirect effect is the effect of one variable on another that is mediated by, or passes through, at least one other variable in the system. The total effect is the sum of the direct and indirect effects. In the special case of simultaneous equation models, the total effect equals the reduced-form coefficients that econometricians use.

Figure 2(b) is useful for illustrating these definitions. The variable X_1 has a direct effect on Y_1 of γ_{12} and X_1 has zero direct effect on Y_2 and Y_3. The indirect effect of X_1 on Y_2 is $\gamma_{11}\beta_{21}$, the product of the

paths that connect these variables. The X_1 variable has two indirect paths to Y_3: $\gamma_{11}\beta_{21}\beta_{32}$ and $\gamma_{11}\beta_{31}$. The total effect is simply the sum of the direct effect and all the indirect effects. Since the X_1 variable has zero direct effect on Y_3, its total effect on Y_3 is the sum of its indirect effects, i.e. $\gamma_{11}\beta_{21}\beta_{32} + \gamma_{11}\beta_{31}$.

The decompositions of effects in the multiple regression in Figure 2(a) or the confirmatory **factor analysis** of Figure 2(c) are less interesting since all the effects are direct effects. Figure 2(d) with the interaction of latent variables complicates the decomposition since in such models the influences of the component variables in the interaction depend on the value that the other components of the interaction take. See Stolzenberg [11] for guidance on interpreting the decomposition of effects in models with interaction terms.

Several issues surround the discussion of the decomposition of effects. First, general formulas exist for the decomposition of effects [7] (*see* **Structural Equation Models**) or for finding the effects that operate only through specific variables in the system [3]. Secondly, the decomposition of effects is always with respect to a specific path model. If a researcher adds intervening variables between two variables that previously had a direct relation, then in the new model this new link would be indirect. Thirdly, definitions of the decomposition of effects differ among authors when unanalyzed associations between exogenous variables or disturbances are in the model (e.g. [5]). Two-headed arrows connecting variables signify that the association between these variables is not explained by the model structure. This creates ambiguity in understanding the effects that might occur with a change in an exogenous variable. For instance, in Figure 2(b) the X_1 and X_2 exogenous variables have a linear association, but we do not show the reason for that association in the path diagram. If, in reality, X_1 affects X_2, a shift in X_1 will lead to real influences beyond those estimated with the decomposition of effects given above. In the decomposition of the covariance (correlation) between two variables into its structural components (see the section "Implied Covariance Matrix" in **Structural Equation Models**), terms that involve the covariance (correlation) of exogenous variables or disturbances are sometimes called *spurious effects*. A final issue to consider is that there is some work that has redefined the decomposition of effects to conform to specific definitions of causality that involve the manipulation of variables in the model. These definitions can lead to different decomposition formulations than those described here [9].

Sewall Wright's Contributions with Path Analysis

Wright's applications of path analysis led to many innovations. His first application in 1918 modeled size as a latent variable that underlied the dimensions of rabbit bones [12]. His latent variable analysis was developed without knowledge of Spearman's [10] article that independently proposed factor analysis. In 1925, Wright [14] employed path analysis to estimate a longitudinal **panel** analysis with latent variables in a study of corn and hog associations. This was decades before similar models would become more commonplace in the social sciences. Goldberger [6] suggests Wright's use of path analysis in supply and demand models was a forerunner of the simultaneous equation approach later developed by econometricians. The influence of Wright's path analysis on sociometrics is clearly seen in the many sociological publications on the technique that began to appear in the 1960s and 1970s (e.g. [1, 2], and [4]). Provine's [8] biography documents Wright's many contributions to evolutionary biology.

Since Wright's early work, the path analysis technique has evolved in many directions. Its current form is best represented in **structural equation models**. These models and path analysis arc widely known in the social sciences and are receiving renewed interest in the biological sciences.

References

[1] Blalock, H.M. (1964). *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill.

[2] Blalock, H.M., ed. (1971). *Causal Models in the Social Sciences*. Aldine-Atherton, Chicago.

[3] Bollen, K.A. (1987). Total, direct, and indirect effects in structural equation models, in *Sociological Methodology 1987*, C.C. Clogg, ed. American Sociological Association, Washington, pp. 37–69.

[4] Duncan, O.D. (1966). Path analysis: sociological examples, *American Journal of Sociology* **72**, 1–16.

[5] Duncan, O.D. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.

[6] Goldberger, A.S. (1972). Structural equation methods in the social sciences, *Econometrica* **40**, 979–1001.

[7] Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8*. Scientific Software Inc., Mooresville.
[8] Provine, W.B. (1986). *Sewall Wright and Evolutionary Biology*. University of Chicago Press, Chicago.
[9] Sobel, M.E. (1990). Effect analysis and causation in linear structural equation models, *Psychometrika* **55**, 495–515.
[10] Spearman, C. (1904). General intelligence, objectively determined and measured, *American Journal of Psychology* **15**, 201–293.
[11] Stolzenberg, R. (1979). The measurement and decomposition of causal effects in nonlinear and nonadditive models, in *Sociological Methodology 1980*, K.F. Schuessler, ed. Jossey-Bass, San Francisco, pp. 459–488.
[12] Wright, S. (1918). On the nature of size factors, *Genetics* **3**, 367–374.
[13] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.
[14] Wright, S. (1925). Corn and hog correlations, *Bulletin No. 1300*. US Department of Agriculture, Washington.
[15] Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics* **5**, 161–215.

KENNETH A. BOLLEN

Pattern Recognition

The term *pattern recognition* refers to a technology that recognizes and analyzes patterns automatically by machine. Human beings are of course subjective experts in both the discernment and imposition of patterns in nature; pattern recognition by machines commonly excludes gestalt theory, although figure/ground phenomena and other aspects of sensory perception are common to both. The present brief treatment of various topics in pattern recognition favors breadth over depth, application over theory. Pattern recognition has been employed successfully in many areas of application including optical character recognition, speech recognition, face recognition, remote sensing (sonar, radar, and satellite), mining, and medical image processing (*see* **Image Analysis and Tomography**). Specific examples include automatic reading of addresses and postal codes, both machine- and handwritten; training a programmable computer chip to recognize its owner's voice in executing a limited command vocabulary; employee identification through video camera security systems; classification of picture elements (pixels) as containing water, land, or ice in a satellite image of the earth's surface; tracking cracks and fissures in geological core samples; breast cancer detection in digital mammography (*see* **Screening, Overview**); karyotyping for identification of chromosomal abnormalities; detection of ventricular fibrillation; and semiautomatic classification of volume elements (voxels) in human brain images as representing white matter, gray, matter or cerebrospinal fluid. Data in all of these cases are multivariate objects consisting of one-dimensional temporal waveforms, two- or three-dimensional images in space, and four-dimensional **time series** of image volumes. Pattern recognition involving automatic scene analysis is also known as image understanding and computer vision [7]. Relevant temporal/spatial patterns (signals) are often obscured by irrelevant detail (noise). In addition to analyzing differences in signal intensity and contrast, pattern recognition also employs color, shape, and texture gradients, as in counting and categorizing different types of white blood cells for instance. Statistical models for pattern recognition problems help to decide what is and is not relevant and take advantage of some form of averaging and aggregation across replications when possible and advisable to reduce extraneous sources of variance and thus increase signal-to-noise ratio. While pattern recognition techniques are widely used, many are patented and only available commercially. There do not yet exist definitive solutions to any one of the aforementioned general problems.

Relatives of Pattern Recognition in Other Fields

In biostatistics, pattern recognition is not as well established as in engineering and industrial applications, yet its status is changing as biostatisticians participate in continued cross-fertilization with these and other disciplines. The reasons for the relative paucity of pattern recognition applications in biostatistics to date may involve preferences in engineering and industry for certainty over uncertainty, determinism over randomness, in applications that do not seem to warrant formal statistical **inference**. While the availability of massive amounts of data may indeed favor deterministic **algorithms** and reduce the need for classical distributional assumptions in some cases, statistical models of uncertainty and inference continue to play a central role in pattern recognition. Some authors use the term *statistical pattern recognition* to describe the line of research that probes structure and pattern in very large data sets without recourse to classical assumptions that may be too inflexible for practical use. Yet there do not appear to be major differences between statistical pattern recognition and pattern recognition *per se*. What differences, then, exist between pattern recognition and its relatives: **artificial intelligence**, expert systems, artificial **neural networks**, and machine learning?

Artificial intelligence is "the science of making machines do the sorts of things done by human minds" [3]. Expert systems "attempt to organize the knowledge of human experts in [a] particular field. . . [and] contain domain specific knowledge" [6]. An artificial neural network is "a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use" [1]. Machine learning is "generally taken to encompass automatic learning procedures based on logical or binary operations that learn a task from a series of examples" [15]. A distinguishing characteristic of pattern recognition is perhaps that no direct analogy is made in its methodology to underlying biological processes, animal or human brains and

minds, real or imagined. Pattern recognition appears most closely related to machine learning, **decision theory**, and **information** theory, all of which rely heavily on statistical theory and methods. In fact, the related fields cited here have their origins in multidisciplinary interactions between engineers building pattern recognition machines, biologists, mathematicians, psychologists, and statisticians. Statistics is one of the oldest fields to study pattern recognition problems and has perhaps seen the greatest duplication of its approaches in other fields [6].

Classification

"Recognition" often means **"classification"**. In optical character recognition, for instance, automatic algorithms are used to classify intensity patterns in a scanned digital image of a printed page as members of an alphabet of allowable characters. Figure 1 is a portion of text, from [21], that has been scanned, blurred, and degraded with different types of noise. A commercial optical character recognition algorithm, set at default values of tuning constants, recognized italics as italics and correctly classified 98.7% (221/224) of the characters in the original, sharper image. Instances of errors were "machrnes" and "recogmtion". The process took 2 seconds (112 characters per second) on an IBM-compatible PC with a Pentium® 133 MHz chip. While a 1.3% error rate may be unacceptable in certain cases, as in automatic processing of lottery tickets for example, it is perhaps surprising that an off-the-shelf computer program could do so well with no tuning or training whatsoever. An automatic spell-checker, another pattern-recognizer, could have made good guesses in replacing these nonwords to achieve a perfect match.

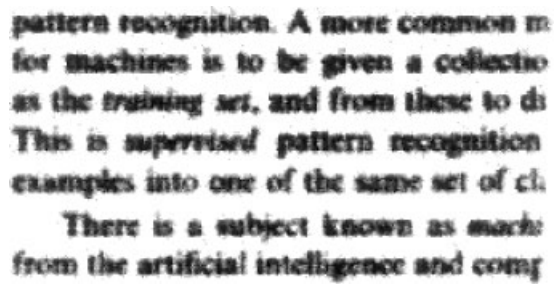
pattern recognition. A more common m
for machines is to be given a collectio
as the *training set*, and from these to di
This is *supervised* pattern recognition
examples into one of the same set of cl
There is a subject known as machi
from the artificial intelligence and comp

Figure 1 Optical character recognition is one example of pattern recognition; text source, [21]

Adding a small amount of Gaussian **noise** to the original image and repeating the optical character recognition task, without retaining any information on this particular set of characters from the previous run, resulted in slight degradation of results (adding "superrised" and "mtelligence" as errors). Blurring the image slightly yielded similar results; adding simulated dust and scratches doubled the processing time and decreased accuracy to 90.6% (203/224). With such good performance, it may also be surprising that none of the characters in the degraded image in Figure 1, which includes Gaussian noise, blur, dust, and scratches, was able to be classified correctly by the machine. A human reader can easily obtain a perfect, meaningful classification. Although a more finely tuned test could surely have done better, this small experiment demonstrates that, in pattern recognition, subjective (prior) knowledge plays an important role and that adding noise to a signal is not necessarily an invertible process. Another example of the current necessity of subjective human judgments is in the classification of pixels in human brain images as white matter, gray matter, or cerebrospinal fluid, the application mentioned in the introduction. While sophisticated and highly successful pattern recognition algorithms exist to get the job done, this task remains at best semiautomatic, requiring a technician to clean up and make sense of automatic results that are not realistic despite all efforts to the contrary by neuroscientists building the tools. Nonetheless, a job that would take hours if done manually takes minutes if done semiautomatically. A major attractive feature of pattern recognition is its ability to yield objective, automatic, and rapid results. Tedium and human error in repetitive, routine tasks can often be removed automatically by designing and implementing clever algorithms. Errors can be further reduced through screening of results by human operators in a multistage process of quality control.

Feature Detection, Extraction, and Classification

Classical pattern recognition proceeds in two stages, feature detection and extraction followed by classification. Pattern features, such as the number of closed loops in hand-written characters, are detected and extracted at a first stage. **Principal component analysis** and singular value decomposition (*see* **Matrix**

Computations) are examples of dimension reduction techniques that are often used in pattern recognition as preprocessing strategies for feature extraction (*see* **Reduced Rank Regression**). At a second stage, vectors of features, the dimensions of which are usually much lower than those of the original multivariate objects, are classified and a decision is made (e.g. "This is a valid alphanumeric string"). Modern pattern recognition involves bi-directional feedback and iteration between these two stages. The classifier adapts itself to subtleties of detected features, creating new features to be extracted, classified, refined, and so forth. Instead of detecting a fixed list of features according to fixed rules, new classes are spawned and eliminated as dynamic algorithms adapt to local data characteristics. These dynamic algorithms may be either entirely deterministic or include a random component. Features and their classifications can also be arranged in dynamic hierarchies and dependency relationships, such as characters, words, sentences, paragraphs, etc. These hierarchies are traversed and altered, either bottom-up or top-down or both, according to some **goodness-of-fit** criterion, to converge at some final pattern classification and ultimate decision.

Invariants and Deformable Templates

One important property of feature detectors and classifiers for pattern recognition is their invariance with respect to minor pattern distortions or deformations. In the optical character recognition example, placements of characters on the page should not matter much, for an "r" is generally an "r" regardless of its orientation. Identification of invariants constitutes important prior knowledge of the problem at hand. Such invariance (in this case, translation, rotation, and scale), as well as other forms of invariance (such as reflection), suggest a class of deformable template models. In deformable template approaches to pattern recognition individual instances of a particular class of objects of interest are represented as deformations of an archetypal, template shape, or form, or texture, etc. from which they all derive according to some specified mechanism, usually a random one. All machine-written or hand-written instances of "r" are seen as distorted instances of some common template for "r", all triangles share common properties that qualify them as members of the geometric class "triangle", all normal electrocardiograms are modeled as perturbations of the same basic waveform of a normal template electrocardiogram, and so forth (*see* **Clinical Signals**). Identification of invariants to be preserved from one instance to the next is a key component in the design of modern pattern recognition systems. Such identification suggests how multivariate objects are best represented as input to the algorithm for the first-stage feature detection and extraction mentioned previously. Until the late 1980s, for instance, most text recognition consisted of matrix matching methods in which a data matrix of 0s (for white) and 1s (for black) containing a particular character was matched with the best fitting matrix from a fixed library of font-specific character shapes. One approach to optical character recognition in the 1990s consists of passing the data matrix containing a character (getting to this point is a major problem in itself) through a series of machine "experts", one for each possible character, that seek to detect the presence of features that are invariant to irrelevant transformations, such as font changes, minor pixel shifts, etc. within the group defining their character class. As another example, in the analysis of human growth, triangles of landmarks for developing structures can be located on two-dimensional images collected over time for each subject and the changing patterns of shapes recognized and analyzed statistically. Shape changes may in some cases be invariant with respect to changes in size. A pattern recognition algorithm for classification could then represent these triangles, with size removed, by two of their interior angles. In general, the best results from pattern recognition systems appear to be those built by human subject-matter experts who have worked out important features and their invariants in painstaking detail and have embodied this knowledge in clever algorithms. Yet this is should come as no surprise, since in biostatistical science detailed substantive models, when available, generally outperform rote, empirical ones.

Learning from Examples: Supervised and Unsupervised Learning

Pattern recognition is an exercise in learning from examples. Under supervised learning, the machine is given a labeled collection of examples. In optical character recognition, for instance, examples of scanned alphanumeric data along their correct character assignments are used to train the algorithm to

recognize pattern features in data that are indicative of particular characters and hence to assign character labels to new test data correctly. This training usually involves some form of **cross-validation**. Unsupervised learning, which occurs less frequently in practice, refers to a pattern recognition exercise that does not involve preassigned labels in either training or test phases of algorithm development. The emphasis in unsupervised learning is on the discovery of new groupings and common characteristics amongst previously unclassified ensembles of examples.

Linear Discriminant and Nearest Neighbor Rules: The Curse of Dimensionality

Pattern recognition systems built according to well-established criteria from statistical decision and information theory are known to outperform systems built using *ad hoc* methods. A **loss function** is defined through consideration of losses incurred by making different sorts of classification errors, such as the **false positives** and **false negatives** of medical diagnoses. A classification rule is then chosen which minimizes total risk, i.e. total expected loss, when assigning class labels. When classification probabilities are assumed to be known, the best that a pattern recognition system can do is determined by Bayes' rule, minimizing Bayes' risk (*see* **Bayesian Methods**). Criteria for label assignment can be parametric or nonparametric. When feature vectors have been generated by **multivariate normal** (Gaussian) **distributions**, with different means yet a common variance–**covariance matrix**, then the optimal classification (Bayes) rule that minimizes total **Mahalanobis distance** is the well-known linear discriminant function (*see* **Discriminant Analysis, Linear**). (Note the distinction between this pattern recognizer and Fisher's linear discriminant, which instead maximizes the ratio of between-class to within-class variance.) Many other parametric, **likelihood**-based classifiers exist. Nonparametric, nearest-neighbor pattern recognition assigns labels to new examples by measuring distances to their nearest neighbors under some particular metric. Metrics for determining nearness may be either fixed or adaptive, Euclidean or non-Euclidean. Class labels are determined by assigning a new example to that class that minimizes the combined distance from the example to a fixed number k (much lower than the total number of examples) of previous examples of that same class over all possible class assignments. When $k = 1$, and for feature vectors of dimension $p \geq 2$, nearest-neighbor pattern recognition is equivalent to Dirichlet tessellation or tiling of the feature space. When $k = 1$ and $p = 2$, nearest-neighbor rules produce a feature space tiling by what are also known as Voronoi or Theissen polygons [21]. Linear discriminant, nearest-neighbor, and other pattern recognition rules, whether parametric or nonparametric, can be extended to allow for reject options or doubt probabilities for uncertain classifications. When p is large, pattern recognition is subject to the so-called curse of dimensionality, i.e. the inability in high-dimensional spaces to obtain enough examples to pack the feature space densely. All examples are close to an edge rather than uniformly spread throughout that space, and thus geometric intuition gained from low-dimensional problems is often inapplicable [8]. The curse of dimensionality is particularly vexing because it may be the extremely high dimension of a multivariate problem that recommends it for treatment by pattern recognition techniques to begin with. There appears to be no substitute in the applied pattern recognition problems of biostatistics for careful selection and representation of the most important features of interest, for incorporation of prior knowledge whenever possible, and for careful checking of results from automatic algorithms by end users.

Prediction and Generalization: Bias/Variance Tradeoff

The construction of pattern recognition systems and the evaluation of their performance usually involves some form of prediction. One way to assess the predictive performance of a pattern recognizer is to estimate its generalization ability, that is, how well it performs on an entirely new set of examples (test set) after having recognized patterns in previous, similar examples (training set). Some form of bias/variance tradeoff in algorithm design is required to avoid too much variance, or overfitting, on the one hand, and too much **bias**, or underfitting, on the other. This is a type of model selection. For a fixed number of examples, bias/variance tradeoffs within nested families of models (*see* **Hierarchical Models**) relate to algorithmic complexity: if a pattern recognizer is too simple, then it is biased and makes systematic prediction

errors, whereas if it is too complex, then it will make unreliable predictions. An additional and somewhat complementary way of assessing predictive performance is to fix algorithmic complexity and to graph generalization error as a function of training set size. This produces a learning curve which, in general, is a monotonically decreasing function. Repeating this exercise for a pattern recognizer of different complexity enables a comparison: when learning curves for two algorithms cross, one outperforms the other for small vs. large numbers of examples.

Layered Feed-forward Networks and Their Equivalent Statistical Models

Artificial neural networks embody many key pattern recognition ideas and principles. The artificial neural network school gained a large following in the 1960s and 1970s, dwindled somewhat in the late 1970s and early 1980s, and is experiencing a resurgence in the 1990s. Many of the ideas and principles of artificial neural networks are common to biostatistical sciences, although sometimes under different names. Figure 2 depicts perhaps the simplest artificial neural network, the McCullough–Pitts neuron, or single-unit perceptron [5, 13], whose algebraic formula is $y = \text{sign}\left(w_0 + \sum_{i=1}^{p} w_i x_i\right)$, where sign$(\cdot)$ equals $+1$ if its argument is nonnegative and -1 otherwise. This artificial neural network is a type of multiple linear regression of y on $x_1, \ldots, x_p$ with unknown weights (parameters) to be estimated and without any formal error term (although many can of course be specified at this point). This pattern recognizer is not ordinary **least squares** regression but rather some form of discriminant analysis with an indicator variable,

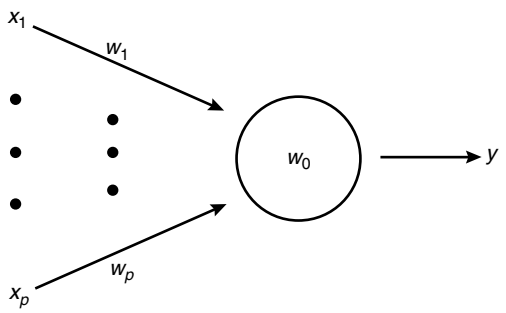

Figure 2 This pattern recognizer is a simple form of artificial neural net [5]

y, as the output of the "neuron" modeled as a linear function of its inputs, the xs, which are themselves indicator variables. The neuron "fires" or does not fire depending on whether or not the summation is positive. Networks of artificial neurons such as these are constructed by forming interconnected banks of such elements with xs feeding into every such node, such as the circle in Figure 2, and the outputs of these nodes feeding into similar nodes at another level, and so forth. Such an arrangement is a layered feed-forward neural network or, equivalently, a multilayer perceptron [10]. This toy example shows that, in one sense, artificial neural networks "can be regarded as a graphical notation for specific regression and classification techniques" [18]. Layers in between input and output are called *hidden layers* since they are not observable directly. Nodes at the same level are not connected to one another directly in a feed-forward net; otherwise, the construction is a recurrent net, Hopfield net, Boltzmann machine, or associative memory the stable patterns of which can be recognized using **Markov chain Monte Carlo** ("Gibbs sampler") methodology [5, 21].

The layered feed-forward neural network is an important pattern recognizer and a powerful tool for general problems of classification and function approximation [8]. Many forms of biostatistical models can be represented as artificial neural nets, and vice versa [20]. For instance, when output y is an indicator variable and is modeled as a function of inputs x that is loglinear in its parameters, we have a form of multiple **logistic regression** or logistic discrimination (*see* **Discriminant Analysis, Linear**) that classifies new examples according to maximum posterior probability. This is equivalent to the "softmax" approach of artificial neural networks, and indeed predated it by decades. Extensions to more than two classes lead to **canonical correlation** analysis [12] and flexible discriminant analysis by optimal scoring [4, 9]. When regression functions are not necessarily linear and are allowed to be members of a general feature space spanned by smooth functions of x, artificial neural networks have been shown to be equivalent to various forms of **generalized additive models**, **projection pursuit** regression, and multivariate adaptive regression **splines**. Expansions by Fourier and **wavelets** bases are also possible. For more detail on the relationships between layered feed-forward nets and these models, and well as radial basis functions, **nonparametric**

regression, and high dimensional **density estimation**, see [5], [8], and [20] and discussions and references therein. Demonstrations of close connections between statistical methods and artificial neural networks such as these provide some idea of the depth and breadth of pattern recognition – a truly multidisciplinary field.

Optical Character Recognition of Handwritten Zip Codes

One prized accomplishment of the neural network school is in automatic recognition of handwritten zip codes, which are US postal codes [5, 11]. This artificial neural network received multivariate input of dimension $p = 256$ consisting of a 16×16 matrix of 0s and 1s representing a single numeric character. Each input was processed by three hidden layers consisting of a total of 1256 nodes connected by a total of 63 660 links and 9760 parameters, to yield an output classification as one of the digits 0 through 9. The network was trained to recognize patterns in a training set of 7291 handwritten zip code digits, adapting to data features by back-propagation, which is a form of gradient descent. The resulting discrimination rule misclassified about 1% of the training set cases, rejecting 12% of them as undecidable, and misclassified 5.1% (102/2007) new test cases. The network was subsequently refined by pruning to reduce problems of overfitting through reductions in the number of free parameters, to retain a 1% misclassification rate with only a 9% case rejection rate.

Some Benchmark Data Sets for Biostatistical Pattern Recognition Research

The zip code data set has become a benchmark for research in artificial neural networks and in other approaches to statistical pattern recognition. There does not yet appear to be a well-defined collection of benchmark data sets for strictly biostatistical pattern recognition problems. However, some biostatistical pattern recognition applications are mentioned in [20] and [21] with accompanying data sets available through Statlib. These include data on R.A. Fisher's irises, hemophilia, Cushing's syndrome, diabetes (in North American Indians, and for analysis of an oral glucose tolerance test), renal disease and hypertension, and differential diagnosis of liver disease, tobacco viruses, and rock crabs.

The pattern recognition literature is enormous, spanning many disciplines. In addition to the works cited herein, the articles that appear after the Reference list are useful sources for further study.

References

[1] Aleksander, I. & Morton, H. (1990). *An Introduction to Neural Computing*. Chapman & Hall, London.

[2] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

[3] Boden, M.A. (1987). Pattern recognition, in *The Oxford Companion to the Mind*, R.L. Gregory, ed. Oxford University Press, Oxford pp. 48–50.

[4] Brieman, L. & Ihaka, R. (1984). Nonlinear discriminant analysis via ACE and scaling, *Technical Report 40*. Department of Statistics, University of California, Berkeley.

[5] Cheng, B. & Titterington, D.M. (1994). Neural networks: a review from a statistical perspective (with comments), *Statistical Science* **9**, 2–54.

[6] Cherkassky, V., Friedman, J.H. & Weschler, H., eds. (1994). *From Statistics to Neural Networks, Theory and Pattern Recognition Applications*. Series F: Computer and Systems Sciences, Vol. 136. NATO-PCO Database. Springer-Verlag, Berlin.

[7] Duda, R.O., Hart, P.E. & Stork, D.G. (1997). *Pattern Classification and Scene Analysis*, 2nd Ed. Wiley, New York.

[8] Friedman, J.H. (1994). An overview of predictive learning and function approximation, in *From Statistics to Neural Networks, Theory and Pattern Recognition Applications*, V. Cherkassky, J.H. Friedman & H. Weschler, eds. Series F: Computer and Systems Sciences, Vol. 136. NATO-PCO Database. Springer-Verlag, Berlin.

[9] Hastie, T., Tibshirani, R. & Buja, A. (1994). Flexible discriminant analysis by optimal scoring, *Journal of the American Statistical Association* **89**, 1255–1270.

[10] Hertz, J., Krogh, A. & Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Santa Fe Institute, Santa Fe.

[11] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. & Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition, *Neural Computation* **1**, 541–551.

[12] Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, London.

[13] McCulloch, W.S. & Pitts, W. (1948). The statistical organization of nervous activity, *Biometrics* **4**, 91–99.

[14] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

[15] Michie, D., Spiegelhalter, D.J. & Taylor, C.C., eds. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.

[16] Minsky, M. (1963). Steps toward artificial intelligence, in *Computers and Thought*, E.A. Fiegenbaum & J. Feldman, eds. McGraw-Hill, New York, pp. 406–450.

[17] Moody, J.E. (1991). Note on generalization, regularization and architecture selection in nonlinear learning systems, in *First IEEE-SP Workshop on Neural Networks in Signal Processing*. IEEE Computer Society Press, Los Alamitos, pp. 1–10.

[18] Poggio, T. & Girosi, F. (1993). Learning algorithms and network architectures, in *Exploring Brain Functions: Models in Neuroscience*, Proceedings of the Dahlem Conference, T. Poggio & F. Girosi, eds. Wiley, New York. [This paper also appeared in *Brain Theory*, A. Aertsen, ed. Elsevier, Amsterdam, 1993].

[19] Ripley, B.D. (1986). Statistics, images and pattern recognition, *Canadian Journal of Statistics* **14**, 83–111.

[20] Ripley, B.D. (1995). Neural networks and related methods for classification, *Journal of the Royal Statistical Society, Series B* **56**, 409–456.

[21] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

(*See also* **Factor Analysis, Overview**; **Multivariate Analysis, Overview**; **Multivariate Graphics**; **Tree-structured Statistical Methods**)

NICHOLAS LANGE

Pearl, Raymond

Born: June 3, 1879, in Farmington, New Hampshire.
Died: November 17, 1940, in Hershey, Pennsylvania.

Photograph supplied by The Alan Mason Chesney Medical Archives of the Johns Hopkins Medical Institutions

Raymond Pearl was a member of the original faculty of the Johns Hopkins University School of Hygiene and Public Health. In a very real sense he established biostatistics as a central discipline for the schools of public health that were to follow.

Pearl's primary training was in biology, with A.B. and Ph.D. degrees in that subject from Dartmouth College in 1899 and the University of Michigan in 1902, respectively. He remained at Michigan as an instructor in zoology until 1905. By then, Pearl recognized a need for sound statistical methods for application in biology, because in the study of human biology it would not be enough to know the individual organism in as much detail as possible; one needed to know the characteristics of the group or groups to which the individual belonged. The most informative characteristics were usually quantitative, and mathematical models and statistical methods were required to study them.

Accordingly, he arranged to spend a year studying biometrics with **Karl Pearson** at University College, London. The impression Pearl made on Pearson was apparently favorable, as he served as a ***Biometrika*** associate editor from 1906 to 1910.

After a year as a junior faculty member in biology at the University of Pennsylvania, Pearl joined the Department of Biology of the Maine Agricultural Experiment Station in 1907, where he rose to department chairman and remained until 1918.

By 1917, Pearl was nationally known and respected, particularly for his application of statistical methods to population studies. In that year, he accepted an appointment as wartime chief of the statistical division of the US Food Administration (later the **Food and Drug Administration (FDA)**).

Coincidentally, the Rockefeller Foundation had decided to make funds available to establish an institute of hygiene, a new type of institution for increasing scientific knowledge, not only about how to interrupt the spread of contagious disease, but also about how to promote health and prevent disease generally. William Henry Welch, a Johns Hopkins University School of Medicine pathology professor, had been selected to organize the new institution as a component of Johns Hopkins. Welch saw in Pearl just the individual to "provide a general theoretical and philosophical approach to quantitative methods as applied to human health and disease" [1, p. 63].

Pearl thus became the first professor and chairman of the Department of Biometry and Vital Statistics of the Johns Hopkins University School of Hygiene and Public Health, the first institution of its kind anywhere. His series of appointments at Johns Hopkins were Professor of Biometry and Vital Statistics in the School of Hygiene and Public Health, 1918–1925; Professor of Biology in the School of Medicine, 1923–1940, and in the School of Hygiene and Public Health, 1930–1940; and Research Professor and Director of the Institute of Biological Research, 1925–1930.

Pearl published more than 700 books, articles, and reviews. In his own listing they fell into the categories: animal behavior; biology – general; biometric theory; biometry – general; duration of life (*see* **Life Expectancy**) and biology of death; effects of physical agents on organisms; **eugenics**; evolution; food and economics; general physiology; genetic technique (*see* **Genetic Epidemiology**); heredity and breeding (*see* **Human Genetics, Overview**); pathology; physiology of **reproduction**; population; poultry husbandry; **public health** and hygiene; sex;

teratology; tuberculosis; variation and **correlation**; **vital statistics**; zoological and physiological technique; and miscellaneous.

One major line of research involved the mathematical modeling of population sizes (*see* **Population Growth Models**), especially using the **logistic** function. With his Johns Hopkins colleague Lowell Reed, Pearl showed that US total population sizes as reported from decennial **census** data from 1790 to 1910 matched very closely a certain logistic curve [2]. The significance of this, to Pearl, was that the parameters of the model were easy to explain in terms of such concepts as the natural limit of the population size, which, in turn, lent strong credibility to **predictions** from the fitted curve. The publication appeared in 1920; the predicted 1940 population size differed by only 3.7% from that reported from the actual 1940 census count.

Naturally, predictions of national population size attracted the attention of the general public, as did quite a few of Pearl's other research endeavors. He concluded from the analysis of data from a large group of his fellow Baltimore citizens that those using alcohol in moderation lived longer on the average than those abstaining and than those using it to excess. He also produced analyses indicating the negative impact of cigarette use on life span (*see* **Smoking and Health**).

Pearl was elected to the US National Academy of Sciences at the relatively early age of 37, the American Philosophical Society, and the American Academy of Arts and Sciences. He served as **American Statistical Association** president in 1939. The University of Maine, Dartmouth College, and St John's College (Annapolis) conferred honorary doctorates. He founded, and was first editor of, two journals, *Quarterly Review of Biology* in 1926 and *Human Biology* in 1929.

When Pearl died suddenly in 1940 at age 61, his obituary appeared in many publications, including the *American Journal of Public Health, Science, The Scientific Monthly, The New York Times*, and *Newsweek*. The obituary in Baltimore's *Evening Sun* was written by his close friend H.L. Mencken.

References

[1] Fee, E. (1987). *Disease & Discovery: A History of the Johns Hopkins School of Hygiene and Public Health 1916–1939*. The Johns Hopkins University Press, Baltimore.

[2] Pearl, R. & Reed, L. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation, *Proceedings of the National Academy of Sciences* **6**, 275–288.

DENNIS O. DIXON

Pearson Distributions

This system of **probability** distributions (often known as *Pearson curves*) was developed by **Karl Pearson** in 1894 and 1895 to provide flexible descriptions of the nonnormal distributions encountered in his biometric research. The original papers are reproduced in [5].

Apart from the fitting of models for observed frequency distributions, the Pearson distributions have also been used to provide approximations to other theoretical distributions, and to study the effect of nonnormality on **sampling distributions**. An early book by W.P. Elderton, now available under joint authorship [1], provided a popular guide to the Pearson curves. For a fuller description, see [3].

Pearson noticed a difference equation satisfied by adjacent probabilities in the **hypergeometric distribution**, and produced an analogous differential equation for continuous distributions with probability density function (pdf) $f(x)$:

$$f'(x) = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2}. \qquad (1)$$

Variations in the four parameters of (1) produce a wide range of unimodal distributions, including **U-** and **J-shaped distributions**. Pearson identified 12 types, the **normal distribution** being a limiting case of several types. The types can conveniently be distinguished by the indices of **skewness** and **kurtosis**, $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$, respectively (see Figure 1).

Several well-known distributions are special cases: the **beta distribution** (I), the **chi-square** or **gamma** (III), the ***F* distribution** (VI), **Student's *t*** (VII), the **exponential** (X), and the **Pareto** (XI). Pearson advocated the fitting of the distributions by the **method of moments**, equating the observed and theoretical values of the first four **moments**. **Efficient** methods such as **maximum likelihood** are now regarded as preferable.

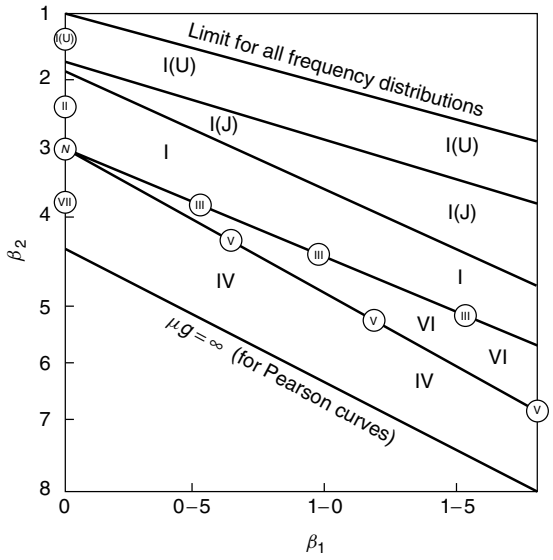

Figure 1 A chart relating types of Pearson distributions to measures of skewness and kurtosis (based, with permission, on Table 43 in [4])

Another general system of continuous distributions, by Johnson [2], is based on **transformations** of a normally distributed variable.

References

[1] Elderton, W.P. & Johnson, N.L. (1969). *Systems of Frequency Curves*. Cambridge University Press, Cambridge.

[2] Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation, *Biometrika* **36**, 149–176.

[3] Ord, J.K. (1985). Pearson system of distributions, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, New York, pp. 655–659.

[4] Pearson, E.S. & Hartley, H.O. (1966). *Biometrika Tables for Statisticians*, 3rd Ed. Cambridge University Press, Cambridge.

[5] Pearson, K. (1948). *Early Statistical Papers*. Cambridge University Press, Cambridge.

PETER ARMITAGE

Pearson, Egon Sharpe

Born: August 11, 1895, in Hampstead, UK.
Died: June 12, 1980, in Midhurst, UK.

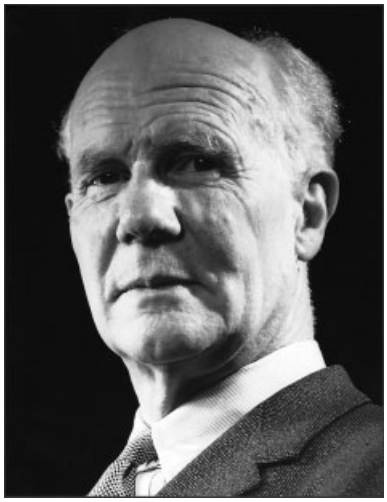

Reproduced by permission of the Royal Statistical Society

Egon Pearson was the only son among the three children of **Karl Pearson**. After education in Oxford and Winchester, in 1914 he went to Trinity College, Cambridge, to read mathematics. War service at the Admiralty and Ministry of Shipping delayed his graduation until 1920.

After leaving Cambridge, Pearson became a statistics lecturer in his father's department at University College, London. Here he helped his father to edit ***Biometrika***. In 1924 he was appointed assistant editor, and after his father's death in 1936 he became managing editor, a position he held until 1965. After Karl Pearson had retired in 1933 his department was split into two, with Egon Pearson becoming head of the new statistics department and **R.A. Fisher** becoming head of the **eugenics** department. In 1935 Pearson was made professor.

From the 1920s Pearson began to develop his personal philosophy of statistics, with two main strands. Most famously, he embarked on an important collaboration with **Jerzy Neyman** which led to the philosophy of statistical **inference**, now known as "**Neyman–Pearson** theory". In this work, Neyman and Pearson introduced ideas of the **alternative hypothesis** and the **power** of a test [2]. Later papers included ideas of suspended judgment in addition to formal acceptance or rejection of a hypothesis [3, 4] (*see* **Hypothesis Testing**).

Pearson's other main field of activity was the promotion of use of statistical methods in industry, with an emphasis on model building. His interest in this area arose from meeting W.H. Shewhart of Bell Telephone Laboratories when Pearson was visiting North America. An influential paper [6] led to the formation of the Industrial and Agricultural Section of the **Royal Statistical Society** (RSS) in 1933 and the appearance of the *Supplement* to the ***Journal of the Royal Statistical Society***, starting in 1934. In 1936 Pearson broke new ground with the publication of a handbook (BS600) on statistical methods in standardization.

During World War II Pearson worked for the Ordnance Board, after which he returned to University College, where he remained until, and indeed after, his retirement in 1960. Among his other publications, the two volumes of statistical tables (with H.O. Hartley) [8] were especially valuable, their title disguising their rich content.

Pearson was made CBE in 1946. He received the Guy medal in gold from the RSS in 1955, and was president in 1955–1956. He was elected FRS in 1966.

Further details of Pearson's life are given by Moore [1] in an eightieth birthday tribute. Many of his important papers are republished in [5] and [7].

References

[1] Moore, P.G. (1975). A tribute to Egon Sharpe Pearson, *Journal of the Royal Statistical Society, Series A* **138**, 129–130.

[2] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, Parts I and II, *Biometrika* **20A**, 175–240, 263–294.

[3] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypothesis, *Philosophical Transactions of the Royal Society of London, Series A* **231**, 289–337.

[4] Neyman, J. & Pearson, E.S. (1933). The testing of statistical hypotheses in relation to probabilities *a priori*, *Proceedings of the Cambridge Philosophical Society* **24**, 492–510.

[5] Neyman, J. & Pearson, E.S. (1967). *Joint Statistical Papers of J. Neyman and E.S. Pearson*. Cambridge University Press, Cambridge, and University of California Press, Berkeley.

[6] Pearson, E.S. (1933). Statistical method in the control and standardization of the quality of manufactured product, *Journal of the Royal Statistical Society, Series A* **96**, 21–60.

[7] Pearson, E.S. (1966). *The Selected Papers of E.S. Pearson*. Cambridge University Press, Cambridge.

[8] Pearson, E.S. & Hartley, H.O. (1954 and 1972). *Biometrika Tables*, Vol. 1 (1954) and Vol. 2 (1972) Cambridge University Press, Cambridge.

Pearson, Karl

Born: March 27, 1857, in London, UK.
Died: April 27, 1936, Coldharbour under Dorking, Surrey, UK.

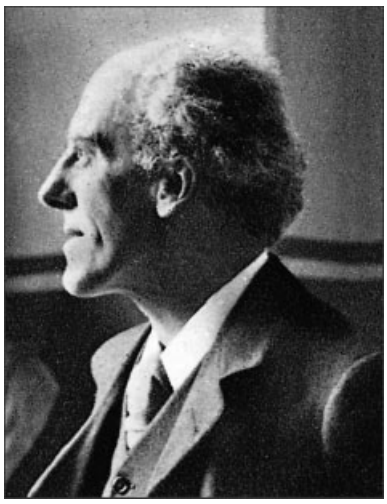

Reproduced by permission of the Royal Statistical Society

The founder of biometrics, Karl Pearson was one of the principal architects of the modern theory of mathematical statistics. He was a polymath whose interests ranged from astronomy, mechanics, meteorology, and physics to the biological sciences in particular (including anthropology, **eugenics**, evolutionary biology, heredity, and medicine). In addition to these scientific pursuits, he undertook the study of German folklore and literature, and of the history of the Reformation and German humanists (especially Martin Luther). He also contributed hymns to the *Socialist Song Book*. Pearson's writings were prodigious: he published more than 650 papers in his lifetime, of which 400 are statistical. Over a period of 28 years, he founded and edited six journals and was a cofounder of the journal ***Biometrika***. University College London (UCL) houses the main collection of Pearson's papers, consisting of 235 boxes containing family papers, scientific manuscripts, and 16 000 letters.

Largely because of his interests in evolutionary biology, Pearson created, almost single-handedly, the modern theory of statistics in his Biometric School at UCL from 1892 to 1903 (which was practiced in the Drapers' Biometric Laboratory from 1903 to 1933). These developments were underpinned by Charles Darwin's ideas of biological variation and "statistical" populations of species – arising from the impetus of statistical and experimental work of his colleague and closest friend, the Darwinian zoologist, W.F.R. Weldon (1860–1906). Additional developments emerged from **Francis Galton's** law of ancestral heredity. Pearson also devised a separate methodology for problems of eugenics in the Galton Eugenics Laboratory from 1907 to 1933.

In his creation of biometrics, out of which the discipline of mathematical statistics had developed by the end of the nineteenth century, Pearson introduced a new vernacular for statistics (including such terms, for example, as the **standard deviation, mode**, homoscedasticity, heteroscedasticity (*see* **Scedasticity**), **kurtosis**, and the product-moment **correlation** coefficient). Like a number of scientists at the end of the nineteenth century, Pearson was interested in the developing etymology in various disciplines, especially biology. Though he attempted to coin a number of biological words, the only word that survived him is "siblings", which he used "to cover a group of brothers and sisters regardless of sex".

Family and Education

Karl was the second of three children born to William Pearson and Fanny Smith. His mother came from a family of seamen and mariners, and his father was a barrister. The Pearsons were a Yorkshire family of dissenters and of Quaker stock. By the time he was in his twenties, Pearson had rejected Christianity and had become a Freethinker, which involved the "rejection of all myths as explanation and the frank acceptance of all ascertained truths to the relation of the finite to the infinite". Though he did not regard himself as an atheist, "he vigorously denied the possibility of a god ... because the idea of one and all of them by contradicting some law of thought involves an absurdity". To Pearson "religion was the relation of the finite to the infinite". Politically, he was a socialist whose outlook was similar to the Fabians, but he never joined the Fabian Society (despite requests from Sidney and Beatrice Webb). Socialism was a form of morality for Pearson; the

moral was social and the immoral was antisocial in conduct.

There were a number of lawyers in the Pearson family, including William's brother Robert and Robert's son, Hugh, as well as William's eldest child, Arthur, all of whom read law at the Inner Temple in London. William was a very hard-working and taciturn man who was never home before 7 p.m. he continued to work until about midnight and was usually up at 4 a.m. reviewing his briefs. To both of his sons, William regularly emphasized the importance of hard work, especially so once they were at Cambridge University. The children only really spent any time with their father during vacations. In a letter to Karl, his elder brother Arthur described the experience of being home with their father as "simply purgatory ... the governor never spoke a word". In this desolate atmosphere, with her husband working incessantly and never talking to anyone when he was home, Fanny was deeply unhappy in her marriage. Thus she transferred her love to her two sons, and she was deeply affectionate to Karl, who was, without doubt, her favorite child.

For a short time in 1866, both boys received tuition from a Mr William Penn, who had started a small school at Harrow, near London. As a child, Karl was rather frail, delicate, often ill, and prone to depression. On a number of occasions he received tuition at home because he was too unwell to go to school. After the Pearsons moved to 40 Mecklenburgh Square, in Holborn, London, in June 1866 (where they stayed until 1875), Karl and Arthur began attending University College London School.

When they went up to Cambridge, at least one of the Pearson boys was expected to read mathematics. The Cambridge Mathematics Tripos was, at that time, the most prestigious degree in any British university. Although his father urged him to read mathematics, Arthur settled on Classics. Thus when Karl was 15 years old, his father was already looking for someone to prepare him for the Mathematics Tripos. Karl first received tuition from the Reverend Louis Hensley at Hitchin (50 miles from Cambridge). Subsequently he was tutored in mathematics at Merton Hall in Cambridge, by John Edward Rendall Harris, John P. Taylor, and Edward John Routh.

By the spring of 1875, Pearson was ready to take the entrance examinations at various Cambridge colleges. His first choice was Trinity College, where he failed the entrance exam; his second choice was King's College, from which he received an Open Scholarship on April 15, 1875. Pearson found that the highly competitive and demanding system leading up to the Mathematical Tripos was the tonic he needed. Though he had been a rather delicate and sickly child with a nervous disposition, he came to life in this environment and his health improved. In addition to the highly competitive and intellectually demanding system, students of the Mathematics Tripos were expected to take regular exercise as a means of preserving a robust constitution and regulating the working day. Pearson carefully balanced hard mathematical study against such physical activities as walking, skating, ice-hockey, and lawn tennis.

As a diversion from studying mathematics, Pearson read works from such Romantics as Goethe and Shelley in his second year. He also read Rousseau and Dante, and wrote a couple of articles on Spinoza for the *Cambridge Review*. Pearson's time at King's College left its legacy through his revolt over the compulsory divinity examination. Near the beginning of his third year in 1877, he decided that he no longer wished to be compelled to attend church services. Pearson also refused to retake one of his divinity papers, as it would have interfered with his study for the Mathematics Tripos. The events that transpired led eventually to King's College abolishing the whole system of compulsory divinity examinations in March 1878.

Pearson spent the rest of 1878 in preparation for the Mathematics Tripos examination, which he took in January 1879. He graduated with honors; subsequently, he received a Fellowship from King's College which he held for seven years. He was made an Honorary Fellow of King's in 1903.

A couple of weeks after Pearson had taken his degree, he began to work in Professor James Stuart's engineering workshop and read philosophy during the Lent Term in preparation for his trip to Germany. After making arrangements with Kuno Fischer, Pearson left for Heidelberg in April 1879. Philosophically and professionally, his time in Germany was a period of self-discovery. The romanticist and idealist discovered positivism: Pearson thus adopted and coalesced two different philosophical traditions to fulfill two different needs. Around this time, he began to write the *New Werther*, a literary work on idealism and materialism, written in the form of letters to his fiancée from a young man wandering in Germany. For Pearson the book was "about conflict

between the ideal and the real, spirit and matter". The book was published in 1880 under the pseudonym of Loki (a mischievous Norse god).

Germany and University College London

During his time in Heidelberg Pearson read Berkeley, Fichte, Locke, Kant, and Spinoza, but he subsequently abandoned philosophy because "it made him miserable". He studied physics under Quincke and metaphysics under Kuno Fischer. He then considered becoming a mathematical physicist, but decided not to pursue this since he "was not a born genius". In November 1879, he went to Berlin to hear Kirchhoff and Helmholtz and began to study Roman law under Bruns, Baron, and Dernburg. A year later, he took rooms back in London at the Inner Temple and read law at Lincoln's Inn. He became a barrister at the end of 1881 and practiced law for a very short time only. Still searching for some direction when he returned to London, Pearson lectured on socialism, Marx and Lassalle at the working men's clubs and on Martin Luther at Hampstead, near London, from 1880 to 1881.

By 1882 Pearson had decided that he did not want to pursue the law because it depressed him, and he decided instead to "devote his time to the religious producing of German literature before 1300". Later that year his work on *The Trinity, A Nineteenth Century Passion-Play, The Son; or Victory of Love* was published. From 1882 to 1884, he lectured on German society from the medieval period up to the sixteenth century. He became so competent in German that by the late spring of 1884 he was offered a post in German at Cambridge. In his pursuit of German history, Pearson consulted his friend, the Cambridge University librarian, Henry Bradshaw, who taught him the meaning of thoroughness and patience in research. With Bradshaw's help, Pearson finished in 1887 *Die Fronica: Ein Beitrag zur Geschichte des Christusbildes im Mittelalter* (which involved a collection of the so-called Veronica portraits of Christ).

Nevertheless, Pearson found all these pursuits deeply dissatisfying, and he "longed to be working with symbols rather than words". He then began to write some papers on the theory of elastic solids and fluids as well as some mathematical physics papers on optics and ether squirts. He deputized mathematics at King's College, London and for Professor Rowe at UCL in 1883. Between 1879 and 1884 he applied for more than six mathematical posts and he received the Chair of "Mechanism and Applied Mathematics" at UCL in June 1884.

During Pearson's first six years at UCL, he taught mathematical physics, hydrodynamics, magnetism, electricity, and his specialty, elasticity, to engineering students. Nearly all of his teaching on dynamics, general mechanics, and statics was based on geometrical methods. He finished editing the incomplete manuscript of William Kingdom Clifford's *The Common Sense of Exact Science* in 1885 and a year later he finished Todhunter's *History of the Theory of Elasticity*.

The Gresham Lectures on Geometry and Curve Fitting

Pearson was a founding member of the Men's and Women's Club established in London in 1885 "for the free and unreserved discussion of all matters in any way connected with the mutual position and relation of men and women". Among the various members was Marie Sharpe, whom he married in June 1890. They had three children, Sigrid, Helga, and Egon. Six months after his marriage, he took up another teaching post as Gresham Professor of Geometry, which he held for three years concurrently with his post at UCL. As Gresham Professor, he was responsible for giving 12 lectures a year, delivered on four consecutive days, from Tuesdays to Fridays, during the Michaelmas, Easter, and Hilary terms. The hour-long lectures, which were free to the public, were held at Gresham Collage, in London, and began at 6 p.m. Between February 1891 and November 1893, Pearson delivered 38 lectures. His first eight lectures formed the basis of his book, *The Grammar of Science*, which was published in several languages.

Pearson's earliest teaching of statistics can, in fact, be found in his lecture of November 18, 1891 when he discussed **graphical** statistics and the mathematical **theory of probability**, with a particular interest in **actuarial methods**. Two days later he introduced the histogram – a term he coined to designate a "time-diagram" to be used for historical purposes. He introduced the standard deviation in his Gresham lecture of January 31, 1893. Pearson's early Gresham lectures on statistics were influenced by the work of **Edgeworth**, Jevons, and Venn. Until November 1893, these lectures covered fairly conventional statistical and probability methods. While the material in

these lectures was not original in content, Pearson's approach to teaching was highly innovative. In one of his lectures, he scattered 10 000 pennies over the lecture room floor and asked his students to count the number of heads or tails: "the result was very nearly half heads and half tails, thus proving the law of average and probability".

Pearson's last 12 Gresham lectures signified a turning-point in his career owing, in particular, to his relationship with Weldon – who was the first biologist Pearson met who was interested in using a statistical approach for problems of Darwinian evolution. Their emphasis on Darwinian population of species not only implied the necessity of systematically measuring variation, but also prompted the reconceptualization of statistical populations. Moreover, it was this mathematization of Darwin which led to a paradigmatic shift for Pearson from the Aristotelian essentialism underpinning the earlier use and development of social and **vital statistics**. Weldon's questions not only provided the impetus for Pearson's seminal statistical work, but also led eventually to the creation of the Biometric School at UCL.

In Pearson's first published statistical paper of October 26, 1893, he introduced the **method of moments** as a means of curve-fitting asymmetrical distributions. One of his aims in developing the method of moments was to provide a general method for determining the values of the parameters of a **frequency distribution** (i.e. central tendency, variation, **skewness**, and kurtosis). In 1895 Pearson developed a general formula to use for subsets of various types of frequency curve and defined the following curves: type I (asymmetric **beta** density curve), type II (symmetric beta curve), type III (**gamma** curve), type IV (family of asymmetric curves), and type V (**normal** curve). In his first supplement to this family of curves in 1901 he defined types VI and VII (type VII is now known as **"Student's" distribution**), and then in his second supplement in 1916 types VIII and IX. Many of his curves were J-shaped, U-shaped, and skewed. Pearson derived all of his curves from a differential equation whose parameters were found from the moments of the distribution. As Churchill Eisenhart remarked in 1974, "Pearson's family of curves did much to dispel the almost religious acceptance of the normal distribution as the mathematical model of variation of biological, physical and social phenomena" (*see* **Pearson Distributions**). Though the method of moments is not widely used by biostatisticians today, it still remains a very powerful tool in econometrics.

The Biometric School

Following the success of his Gresham lectures, Pearson began to teach statistics to students at UCL in October 1894. By 1895, four years after Pearson first started to teach statistics, he had worked out the mathematical properties of the product-moment correlation coefficient (which measures the relationship between two continuous variables) and **simple regression** (used for the linear prediction between two continuous variables). By then, Francis Galton had determined graphically the idea of correlation and regression for the normal distribution only. Because Galton's procedure for measuring correlation involved measuring the slope of the regression line (which was a measure of regression instead), Pearson kept Galton's r to symbolize correlation. Pearson later used the letter b (from the equation for a straight line) to symbolize regression. After Weldon had seen a copy of Pearson's paper on correlation, he suggested to Pearson that he should extend the range for correlation from 0 to $+1$ (as used by Galton) so that it would include all values from -1 to $+1$.

In this seminal paper on "Regression, heredity and panmixia" in 1896, Pearson introduced **matrix algebra** into statistical theory (Arthur Cayley, who taught at Cambridge when Pearson was a student, created matrix algebra by his discovery of the theory of invariants during the middle of the nineteenth century). In the same paper, Pearson also introduced the following statistical methods: η as a measure for a curvilinear relationship, the **standard error** of an estimate, **multiple regression**, and multiple, part and partial correlation, and he also devised the coefficient of variation as a measure of the ratio of a standard deviation to the corresponding mean, expressed as a percentage.

From 1896 to 1911, Pearson devised more than 18 methods of correlation. By the end of the nineteenth century he began to consider the relationship between two discrete variables. In 1900, he devised the tetrachoric correlation and the phi-coefficient for dichotomous variables (*see* **Categorical Data Analysis**). The tetrachoric correlation requires that both X and Y represent continuous, normally distributed and linearly related variables, whereas the phi-coefficient

was designed for so-called point distributions, which implies that the two classes have two point values or merely represent some qualitative attribute. Nine years later, he devised the **biserial correlation** when one variable is continuous and the other is discontinuous. With his son Egon, he devised the polychoric correlation in 1922 (which is very similar to **canonical correlation** today). Though not all of Pearson's correlational methods have survived him, a number of these methods are still the principal tools used by psychometricians for test construction (*see* **Psychometrics, Overview**). Following the publication of his first three statistical papers in *Philosophical Transactions of the Royal Society*, Pearson was elected a Fellow of the Royal Society in 1896. He was awarded the Darwin Medal from the Royal Society in 1898.

Pearson's Chi-square Tests

At the turn of the century, Pearson reached a fundamental breakthrough in his development of a modern theory of statistics when he found the exact **chi-square distribution** from the family of Gamma distributions and devised the chi-square **goodness-of-fit** test. The test was constructed to compare observed frequencies in an empirical distribution with expected frequencies in a theoretical distribution to determine "whether a reasonable graduation had been achieved" (i.e. one with an acceptable probability). This landmark achievement was the outcome of the previous eight years of curve fitting for asymmetrical distributions and, in particular, of Pearson's attempts to find an empirical measure of a goodness-of-fit test for asymmetric curves.

Four years later, he extended this to the analysis of manifold **contingency tables** and introduced the "mean square contingency coefficient" which he also termed the **chi-square test** of independence (which **R.A. Fisher** termed the chi-square statistic in 1923). While Pearson used $n - 1$ for his **degrees of freedom** for the chi-square goodness-of-fit test, R.A. Fisher claimed in 1924 that Pearson also used the same degrees of freedom for his chi-square test of independence. However, in 1913, Pearson introduced what he termed a "correction" (rather than degrees of freedom) for his chi-square test of independence of 1904. Thus, he wrote, if x is the number of rows and λ the number of columns, then on average the correction for the number of cells is $(x - 1)(\lambda - 1)/N$. (As may be seen, Fisher's degrees of freedom for the chi-square statistic as $(r - 1)(c - 1)$ is very similar to that used by Pearson in 1913.)

Pearson's conception of contingency led at once to the generalization of the notion of the association of two attributes developed by his former student, **G. Udny Yule**. Individuals could now be classed into more than two alternate groups or into many groups with exclusive attributes. The contingency coefficient and the chi-square test of independence could then be used to determine the extent to which two such systems were contingent or noncontingent. This was accomplished by using a generalized theory of association along with the mathematical theory of independent probability.

Pearson's Four Laboratories

Pearson established and ran four laboratories. He set up the Drapers' Biometric Laboratory in 1903 following a grant from the Worshipful Drapers' Company (which continued to fund Pearson's work in this laboratory until his retirement in 1933). The methodology incorporated in the Drapers' Biometric Laboratory was twofold: the first was mathematical, and included the use of Pearson's statistical methods, matrix algebra, and analytical solid geometry. The second involved the use of such instruments as integrators, analyzers, curve plotters, the cranial coordinatograph, silhouettes, and cameras. The problems investigated by the biometricians included natural selection, Mendelian genetics (*see* **Mendel's Laws**) and Galton's law of ancestral inheritance, craniometry, physical anthropology, and theoretical aspects of mathematical statistics. By 1915, Pearson established the first degree course in mathematical statistics in Britain.

Though Pearson did not accept the generality of Mendelism, he did not reject it completely, as is commonly believed. When William Bateson published his fiercely polemical attack on Weldon in 1902, Bateson saw Mendelism as a tool for discontinuous variation only. As a biometrician, most of the variables that Pearson and his co-workers analyzed were continuous, and only occasionally did they examine discontinuous variables. While Pearson and Weldon used Galton's law of ancestral inheritance for continuous variables, they used Mendelism for discontinuous variables. Indeed, Pearson argued that his

chi-square test of independence was the most appropriate statistical tool for the analysis of Mendel's discrete data for dominant and recessive alleles (such as color of eyes where brown is dominant and blue is recessive). Even today, Pearson's chi-square tests remain the most widely used technique for analyzing Mendelian data.

A year after Pearson had established the Biometric Laboratory, the Worshipful Drapers' Company gave him a grant so that he could establish an Astronomical Laboratory equipped with a transit circle and a 4-inch (10 cm) equatorial refractor. Hence, he also referred to his two observatories as the Transit House and the Equatorial House. Pearson was interested in determining the correlations of stellar rotations and the variability in stellar parallax. He was also instrumental in setting up a degree course in astronomy in 1914 at UCL.

In 1907, Francis Galton (who was then 85 years old) wanted to step down as director from the Eugenics Record Office which he had set up three years earlier, and he asked Pearson if he would take it on. Though Pearson had "great hesitation in taking any initiative at all ... because he did not want Galton to think that [he] was carrying all things into the biometric vortex!", he took on the directorship reluctantly and renamed the office the Galton Eugenics Laboratory. Pearson made very little use of his biometric methods in this laboratory; instead he developed a completely different methodology for problems relating to eugenics. This methodology was underpinned by the use of actuarial death rates (*see* **Actuarial Methods**) and by a very highly specialized use of family pedigrees assembled in an attempt to discover the inheritance of various diseases (which included, for example, such conditions as alcoholism, cancer, diabetes, epilepsy, paralysis, and pulmonary tuberculosis; *see* **Population Genetics**).

These family pedigrees became the vehicle through which Pearson could communicate statistical ideas to the medical community by stressing the importance of using quantitative methods for medical research. This tool enabled doctors to move away from concentrating on individual pathological cases or "types" and to see, instead, a wide range of pathological variation of the disease (or condition) of the doctors' specialty. Such work attracted the interest of **Major Greenwood**, who was the first medically qualified person to take an interest in Pearson's statistics in 1902, and who subsequently became Reader of Medical Statistics in the University of London in 1922 (the first position to be held at a university in Britain). The statistical work of Pearson and Greenwood was further promulgated by their student, **Austin Bradford Hill**, who had the greatest impact on the successful adoption of mathematical statistics in the medical community. In 1924, Pearson set up the Anthropometric Laboratory, this was made possible by a gift from one of his students, Ethel Elderton. The laboratory was open to the public and used to collect and display statistics related to problems of heredity.

In the spring of 1909, Galton was discussing the future of the Eugenics Laboratory with Pearson. While Galton thought that Pearson would have been "the most suitable man for the first Galton Professor", Pearson let Galton know that he was "wholly unwilling to give up superintendence of the Biometric Laboratory [he] had founded and confine [his] work to Eugenics Research". A month later, Galton added a codicil to his will stating that he desired that the first professor of the post should be offered to Pearson on such condition that Pearson could continue to run his Biometric Laboratory. Six months after Galton's death in January 1911, Pearson first learned about Galton's codicil to his will. He then relinquished the Goldsmid Chair of Applied Mathematics after 27 years of tenure to take up the Galton Chair. The Drapers' Biometric and the Galton Eugenics laboratories, which continued to receive separate funding, then became incorporated into the Department of Applied Statistics. The essential aim in combining both laboratories was to enable Pearson to give up his undergraduate teaching of applied mathematics and to devote himself "solely to what had been for many years the main element of [his] research: the advancement of the modern theory of statistics".

Statistical Charts and Gunnery Computations

Pearson then proceeded to raise funding for a new building for his Department of Applied Statistics. Adequate funding had been raised by 1914 and contracts for the fittings had been made. In the early summer of 1914, the new laboratory was complete and preparations were under way for the occupation and fitting up of the public museum and the Anthropometric Laboratory. It was hoped that the building

would be occupied by October 1915. These developments and further biometric work were shattered by the onset of World War I. The new laboratory building was taken over by the government to be used as a military hospital. Pearson and his co-workers took on special war duties. They produced statistical charts for the Board of Trade's Labor Department as well as for its Census production. Pearson was also involved in elaborate calculations concerned with antiaircraft guns and bomb trajectories "both through air and air and water". By June 1919, Pearson was in possession of his building and plans were under way for the opening in October 1920. It was not until December 4, 1922, when the work had been completed, that the building was occupied.

His wife, Marie Sharpe, died in 1928, and in 1929 he married Margaret Victoria Child, a co-worker in the Biometric Laboratory. Pearson was made Emeritus Professor in 1933, and was given a room in the Zoology Department at UCL which he used as the office of *Biometrika*. From his retirement until his death in 1936, he published 34 articles and notes, and continued to edit *Biometrika*. Pearson was twice offered honors by King George V, but declined on both occasions. He also declined the **Royal Statistical Society** Guy Medal in the society's centenary year in 1934. Pearson believed that "all medals and honours should be given to young men, they encourage them when they begin to doubt whether their work was of value". Pearson accepted an honorary Doctorate of Science (D.Sc.) from the University of London in 1934 because if he had refused he "would have hurt the executive of the university where he had worked" for nearly half a century.

Pearson's statistical achievement not only provided continuity from the mathematical and statistical work that preceded him (including that of Francis Ysidro Edgeworth, Francis Galton, **Adolphe Quetelet**, and John Venn), but also engendered the modern theory of mathematical statistics in the twentieth century which, in turn, provided the foundation for such statisticians as R.A. Fisher who went on to make further advancements for a modern theory of statistics.

Bibliography

Archival Sources

Papers and correspondence of Karl Pearson, held in the Manuscript Room at University College London.

Secondary Sources

Eisenhart, C. (1974). Karl Pearson, in *Dictionary of Scientific Biography*, Vol. 10, C.C. Gillispie, ed. Charles Scribner's Sons, New York, pp. 447–473.

Hilts, V. (1981). *Statist and Statistician*. Arno Press, New York.

Mackenzie, D. (1981). *Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge*. Edinburgh University Press, Edinburgh.

Magnello, M.E. (1993). Karl Pearson: evolutionary biology and the emergence of a modern theory of statistics. *D. Phil. thesis*, University of Oxford.

Magnello, M.E. (1996). Karl Pearson's Gresham lectures: W.F.R. Weldon, speciation and the origins of Pearsonian statistics, *British Journal for the History of Science* **29**, 43–64.

Magnello, M.E. (1997). Karl Pearson's mathematisation of inheritance. From Galton's ancestral heredity to Mendelian Genetics (1895–1909), *Annals of Science* **55**, 35–94.

Norton, B. (1978). Karl Pearson and the Galtonian tradition: studies in the rise of quantitative social biology. *Ph.D. thesis*, University of London.

Norton, B. (1978). Karl Pearson and statistics: the social origin of scientific innovation, *Social Studies of Science* **8**, 3–34.

Pearson, E. (1936). Karl Pearson: an appreciation of some aspects of his life and work Part 1, 1857–1905, *Biometrika*, **28**, 193–257.

Pearson, E. (1937). Karl Pearson: an appreciation of some aspects of his life and work. Part 2, 1906–1936, *Biometrika* **29**, 161–248.

Pearson, K. (1914–1930). *The Life, Letters and Labours of Francis Galton*, 3 vols. Cambridge University Press, Cambridge.

Porter, T.M. (1986). *The Rise of Statistical Thinking, 1820–1900*. Princeton University Press, Princeton.

Semmel, B. (1958). Karl Pearson: socialist and Darwinist, *British Journal of Sociology* **9**, 111–125.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, Mass.

Major Publications of Karl Pearson

(1888). *The Ethic of Freethought*. E.W. Allen, London. 2nd Ed., 1901.

(1892). *The Grammar of Science*. Adam & Charles Black, London. 2nd Ed., 1900; 3rd Ed., 1911.

(1893). Asymmetrical frequency curves, *Nature* **48**, 615–616.

(1894). Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society, Series A* **185**, 71–110. Abstract in *Proceedings of the Royal Society of London* **54**, (1893) 329–333.

(1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material, *Philosophical Transactions of the Royal Society, Series A* **186**,

343–414. Abstract in *Proceedings of the Royal Society of London* **57**, 257–260.

(1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society, Series A* **187**, 253–318. Abstract in *Proceedings of the Royal Society of London*, **59**(3), (1895) 69–71.

With Filon, L.N.G. (1898). Mathematical contributions to the theory of evolution. IV. On the probable error of frequency constants and on the influence of random selection on variation and correlation, *Philosophical Transactions of the Royal Society, Series A* **191**, 229–311. Abstract in *Proceedings of the Royal Society of London*, **62**, (1897) 173–176.

(1898). *Chances of Death and other Studies in Evolution*. Edward Arnold, London.

(1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable, *Philosophical Transactions of the Royal Society, Series A* **195**, 1–47.

(1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Fifth Series* **50**, 157–75.

(1904). Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation, *Drapers' Company Research Memoirs. Biometric Series* **I**, 1–37.

(1905). Mathematical contributions to the theory of evolution. XIII. On the general theory of skew correlation and non-linear regression, *Drapers' Company Research Memoirs. Biometric Series* **II**, 54–92.

(1909). On a new method of determining correlation between a measured character A and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A, *Biometrika* **6**, 96–105.

(1909). The theory of ancestral contributions of a Mendelian population mating at random, *Proceedings of the Royal Society* **81**, 225–229.

M.E. MAGNELLO

Pedigrees, Sequential Sampling

The great majority of genetic epidemiologic study designs involve data not just on individuals, but rather on *families*. Generally speaking, we can envision collection of such data as involving two separate operations: first, *individuals* are selected from the target population, or ascertained (*see* **Ascertainment**); secondly, *intrafamilial sampling* among the relatives of these ascertained individuals is conducted. The resulting pedigrees are thus composed of ascertained individuals and (subsets of) their relatives. (In many applications, only affected individuals are eligible for the initial ascertainment, which results in ascertainment bias.)

One technique sometimes used for obtaining families is sequential sampling, in which intrafamilial sampling proceeds in stages, with the decision to continue (or stop) the sampling process made subsequent to completion of each stage. For example, a sequential sampling protocol might require observation of all first-degree relatives of any previously observed affected individual. In this case, all first-degree relatives of initially ascertained individuals would be evaluated, and, if any of them were affected, then all of their (previously unobserved) first-degree relatives would be sampled, and so forth, with sampling continuing or stopping at each stage, depending on whether any newly observed individuals were affected. Such schemes can obviously result in highly variable observed pedigree structures, depending on the ascertainment scheme, the underlying genetic etiology of the disease under study, and the particular intrafamilial sampling rule employed.

Cannings & Thompson [2] defined a broad class of sequential sampling procedures, by allowing for any intrafamilial sampling scheme meeting the following restrictions: (i) the decision as to which individuals to sample next depends only on the phenotypes of individuals who have already been observed, and possibly on other, auxiliary considerations (see below); and (ii) all observed relatives are included in the final sample. These restrictions preclude, for instance, sampling schemes in which the investigator "looks ahead" for affected individuals and then preferentially includes them in the sample, or in which individuals are dropped from the sample after determining that they are unaffected. (When these restrictions are violated, unbiased parameter estimation can become impossible.) It is interesting to note that these criteria permit sampling (and stopping) decisions to be based on auxiliary factors such as cost-effectiveness, or the current **likelihood ratio**, which introduces a sequential element in the classical statistical sense [7]. Also, many other sampling schemes that do not entail sequential decision-making *per se*, nevertheless belong to the class of schemes defined by these restrictions. For instance, sampling of all and only full-siblings of any ascertained individual(s) satisfies (i) and (ii) above.

Sampling family members in a sequential manner can, in some circumstances, greatly increase the amount of genetic information in the sample [4]. For example, Boehnke et al. [1] simulated a quantitative trait under the mixed model (*see* **Segregation Analysis, Mixed Models**), and compared the efficiency of nonsequential and sequential sampling schemes. They found that sampling sequentially could increase **power** to detect segregation at a dominant major locus by over 60%. (However, the amount of increased efficiency actually achieved depends upon the underlying genetic etiology of the disease under study.) Thus, sequential schemes for intrafamilial sampling constitute an extremely flexible and efficient class of procedures for sampling of pedigrees in human genetics.

It is worth noting, however, that sequential sampling introduces certain technical complications into the calculation of **likelihood**, particularly when conducted in conjunction with ascertainment through affected individuals. Strictly speaking, observed pedigree structure is a random variable, which may depend upon genetic parameters as well as the sampling procedures, and should then be included in the formal (likelihood) model. Cannings & Thompson [2] showed that, when intrafamilial sampling is sequential, conditioning the likelihood on the observed pedigree structure does not produce bias in genetic parameter estimates, provided that ascertainment is random, i.e. not a function of phenotype (see also [5]), or single (see also [3] and [5]), and provided that the Cannings & Thompson criteria given above are not violated. However, when ascertainment is neither random nor single, conditioning likelihoods on the observed pedigree structure will introduce (asymptotic) bias into parameter estimation when families are sequentially sampled [5]. In

fact, this result extends to nonsequentially sampled families whenever the observed pedigree structure depends on which among a set of relatives happen to have been the initially ascertained individuals [5]. This result applies not only to segregation analysis, but to linkage analysis as well [6]. It is also important to note that while **maximum likelihood** parameter estimates may not be (asymptotically) affected by the use of sequential sampling procedures, standard errors of estimators cannot be calculated in the usual way when intrafamilial sampling is sequential. As a result, it may be inappropriate to assume that test statistics (e.g. $-2 \times$ the natural logarithm of the likelihood ratio) follow their canonical distributions (e.g. a **chi-square distribution**) (*see* **Likelihood Ratio Tests**).

References

[1] Boehnke, M., Young, M.R. & Moll, P.P. (1988). Comparison of sequential and fixed-structure sampling of pedigrees in complex segregation analysis of a quantitative trait, *American Journal of Human Genetics* **43**, 336–343.

[2] Cannings, C. & Thompson, E.A. (1977). Ascertainment in the sequential sampling of pedigrees, *Clinical Genetics* **12**, 208–212.

[3] Hodge, S.E. & Vieland, V.J. (1996). The essence of single ascertainment, *Genetics* **144**, 1215–1223.

[4] Thompson, E.A. (1981). Optimal sampling for pedigree analysis: sequential sampling schemes for sibships, *Biometrics* **37**, 313–325.

[5] Vieland, V.J. & Hodge, S.E. (1995). Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework, *American Journal of Human Genetics* **56**, 33–43.

[6] Vieland, V.J. & Hodge, S.E. (1996). The problem of ascertainment for linkage analysis, *American Journal of Human Genetics* **58**, 1072–1084.

[7] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

VERONICA VIELAND

Peirce, Charles Sanders

Born: September 10, 1839, in Cambridge, Massachusetts.
Died: April 19, 1914, in Milford, Pennsylvania.

C.S. Peirce is regarded by many as the greatest philosopher the US has ever produced and is considered to be the creator of pragmatism. He wrote and published articles upon a wide range of subjects including logic, mathematics, metaphysics, psychology, optics, astronomy, chemistry, religion, and other subjects. Many of his ideas are of significance to statisticians. These ideas include his thoughts about statistical **inference**, **random samples**, **randomization** of treatments, the definition of probability, formulation of hypotheses (*see* **Hypothesis Testing**), and errors of measurement (*see* **Measurement Error in Epidemiologic Studies**).

Charles Sanders Peirce (pronounced Purse) was the second of five children born to Benjamin Peirce, the leading American mathematician of that time, and Sarah Hunt (Mills) Peirce.

Charles entered Harvard in 1855, graduated in 1859, and ranked seventy-first in a class of 91 graduates. In July of 1861, he joined the US Coast Survey as a computing aide to his father, who had been employed with the Survey since 1852 and who served as superintendent of the Survey, from 1867 to 1874. Charles married Harriet Melusina (Zina) Fay on October 16, 1861. He received an M.A. degree from Harvard in 1862 and an S.B. degree in chemistry, summa cum laude, from the Lawrence Scientific School at Harvard in 1863.

He prepared 11 lectures on the logic of science to be given at Harvard in the spring of 1865. It appears that the lectures were not given at Harvard but were given in revised form at Lowell Institute in 1866. In 1867, he was elected to membership as resident fellow in the American Academy of Arts and Sciences. He worked as an assistant at Harvard Observatory from 1869 to 1872 and lectured on philosophy in 1869–1870. In 1870, he was sent by his father to find suitable observation sites in Europe along the path of totality to observe the solar eclipse of December 22, 1870. He was a member of an observation party near Catania, Sicily.

In 1871, he was in charge temporarily of the Coast Survey. He included astronomical observations made at Harvard Observatory from 1872 to 1875 in *Photometric Researches* (1878). In 1872, he became assistant at the Coast Survey and held this position until 1884. In 1877, he was elected as a fellow of the National Academy of Sciences. He was appointed lecturer in logic at Johns Hopkins University in 1879 and in 1880 he was elected as a member of the London Mathematical Society.

In 1876, he was separated from his wife, Melusina. His father died in 1880. Charles divorced Melusina in 1883 and married Juliette Froissy. He lost his position at Johns Hopkins University in 1884 for unspecified reasons and never held another academic position. In 1887, he and Juliette moved to a country home, Arisbe, near Milford, Pennsylvania. Possibly because of incompleted work, he was forcibly retired from the Coast Survey in 1891.

He continued to write articles for periodicals, *The Monist* and *Nation*, for example [1]. He became an editor and contributor for *Century Dictionary*. He continued to give lectures, which included lectures on the history of science at the Lowell Institute in 1892, a lecture on number at Bryn Mawr College in 1896, lectures on logic at the Lowell Institute in 1903, and lectures on the scientific method before the Philosophy Club at Harvard in 1907. Because of the strong friendship and support of William James, he informally added Santiago (Saint James) to his name.

In Peirce's study of inductive science, the need for random samples is explicitly stated. Peirce ("Lessons from the history of science", *c*. 1896 [2]) says, "A sample is a *random* one, provided it is drawn by such machinery, artificial or physiological, that in the long run any one individual of the whole lot would get taken as often as any other."

Peirce thought of the need for experimental randomization much before **R.A. Fisher**. Gustav Fechner had performed a multifactor experiment in 1860 to study sensory perception. In 1885, Peirce and Joseph Jastrow (see [4]) designed a similar experiment that used randomization of treatments. They presented slightly different weights in random order to subjects who were asked to state which of the two orders they had received. Two well-shuffled decks of cards were used to make the randomization.

Peirce was interested in the meaning and interpretation of probability and tried to give an operational definition of probability. In a letter to William James in 1909, he gave a long-range definition of probability

and he continued to work on a definition of probability in terms of endlessly diminishing oscillations.

On the subject of data analysis, Peirce [2] emphasized the importance of formulating hypotheses before examining the sample. He said, "We must first decide for what character we propose to examine the sample and only after that decision examine the sample." Peirce gave as an example the examination of a random sample of names from a biographical reference. If one allowed the data to suggest hypotheses, one might reach some very unlikely conclusions. For example, he listed several eminent men whose death dates met the following criterion "... All eminent men die in years whose date doubled and increased by one gives a number whose last figure is the same as that in the ten's place of the date itself."

Peirce [3] wrote a paper in 1870 entitled, "On the theory of errors of observation". The paper included a lengthy theoretical development that resulted in the normal law of errors (*see* **Normal Distribution**) and the analysis of a large data set. He studied the reaction times of an untrained 18-year-old boy to signals received. He recorded 500 measurements for each of 24 days and fit normal curves to the data sets.

He died at Milford, Pennsylvania, on April 19, 1914, at the age of 74. His unpublished papers and his library were left to the Harvard department of philosophy.

References

[1] Brent, J. (1993). *Charles Sanders Peirce*. Indiana University Press, Bloomington.

[2] Peirce, C.S. (1931). *Collected Papers of Charles Sanders Peirce*, C. Hartshorne & P. Weiss, eds. Harvard University Press, Cambridge, Mass.

[3] Peirce, C.S. (1976). *The New Elements of Mathematics*, Vol. III/1 *Mathematical Miscellanea*, C. Eisele, ed. Mouton, Paris.

[4] Stigler, S. (1978). Mathematical statistics in the early states, *Annals of Statistics* **6**, 239–265.

J. LEROY FOLKS

Penalized Maximum Likelihood

The standard approach to analyzing a set of observations $x_i, i = 1, \ldots, n$, thought to be a random sample from some population with density $f(\cdot)$ is to assume a parametric form for f and then estimate the parameters of the density using, for example, **maximum likelihood**. This parametric approach has excellent properties if the assumed density is the correct one, but can lead to grossly incorrect inferences if the assumed density is inappropriate.

Nonparametric **density estimation** methods avoid this problem by weakening the assumption on the true density to be that f is unspecified, except that it is assumed to be smooth (that is, at least a specified number of derivatives of f are square integrable, where that number is usually two or four). The nonparametric maximum likelihood estimator (MLE) of f is not a reasonable density estimator, since the maximizer of the nonparametric log likelihood is a set of (Dirac) spikes at the observations $\{x_i\}$. That is, it is not consistent with a smooth density because of its great roughness. Maximum penalized likelihood estimation modifies the log likelihood to discourage the roughness of the nonparametric MLE by penalizing the estimator if it becomes too rough. The maximum penalized likelihood estimator (MPLE) [3] is the maximizer of

$$L(f) = n^{-1} \sum_{i=1}^{n} \log f(x_i) - \Phi(f)$$

subject to $\int f = 1$, where $\Phi(f) \geq 0$ is a roughness penalty that decreases as f becomes smoother. The resultant estimator provides a tradeoff between fidelity to the data (from the log likelihood) and smoothness (from the roughness penalty). The estimator also has a **Bayesian** interpretation, with the **prior** for the density having the form $\exp[-\Phi(f)]$ and the posterior mode being the final estimate.

Different choices of Φ yield different estimators. A common strategy is to take $g = \log f$ and $\Phi(g) = \alpha \int [g^{(m)}(u)]^2 \, du, \alpha \geq 0$, where $g^{(m)}$ is the mth derivative of g [14, 15]. Then, $\hat{f} = \exp(\hat{g})$ (note that since the MPLE is defined through exponentiating an estimate of the log-density, it is nonnegative).

The smoothing parameter α controls the amount of smoothing, with larger α resulting in a smoother estimate. As $\alpha \to 0$, the MPLE approaches the nonparametric MLE of Dirac spikes, while as $\alpha \to \infty$, the MPLE becomes the MLE within a parametric family that depends on Φ. For example, if f is defined on the nonnegative numbers, and $\Phi(g) = \alpha \int [g''(u)]^2 \, du$, the limiting family is **exponential**, while if $\Phi(g) = \alpha \int [g'''(u)]^2 \, du$, the limiting family is Gaussian. If f is bounded away from zero, and $\int [f^{(2m)}(u)]^2 \, du$ is finite, then a roughness penalty based on mth order derivatives gives an estimator the **mean square error** (MSE) of which converges to zero at the rate $O(n^{-4m/(4m+1)})$. Note that the estimator does not achieve the usual parametric rate of $\text{MSE} = O(n^{-1})$. This deficiency can be viewed as being the price that one pays for weakening the parametric assumption merely to smoothness.

Penalized likelihood estimators are often called **spline** estimators, since many such estimators take the form of polynomial splines with knots at the order statistics. Asymptotically, the MPLE is approximately a local-bandwidth kernel estimator. If $\Phi(g) = \alpha \int [g''(u)]^2 \, du$, for example, then

$$\hat{f}(x) \approx \frac{f(x)^{1/4}}{n\alpha^{1/4}} \sum_{i=1}^{n} K\left[\frac{(x - x_i) f(x)^{1/4}}{\alpha^{1/4}}\right],$$

with kernel function

$$K(u) = \frac{1}{2} \exp\left(\frac{-|u|}{\sqrt{2}}\right) \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right),$$

away from the boundary. Further theoretical analysis of the MPLE can be found in [1], [2], [7], and [8].

The MPLE is difficult to compute, and for that reason various adaptations of it have been proposed. Scott et al. [13] converted the penalized likelihood to one on discrete data by binning the observations, calling this the discrete maximum penalized likelihood estimator (DMPLE), and gave conditions where the DMPLE converges to the MPLE as the bins narrow. O'Sullivan [10] and Gu [6] described other spline-based density estimators.

Penalized likelihood methods also can be generalized to multivariate data, although with increasing

Table 1

	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
Staff	0	0	0	0	0	0	1	2	3	1	1	3	3	4	3
Trainee	1	1	0	0	1	0	1	3	5	4	0	2	5	0	0

computational difficulties. Natural roughness penalties take the form of sums of squared derivative terms (one for each dimension), such as (for bivariate data)

$$\Phi = \iint \left[\left(\frac{\partial f}{\partial u} \right)^2 + \left(\frac{\partial f}{\partial v} \right)^2 \right] \mathrm{d}u \, \mathrm{d}v$$

(Scott et al. [12], who used a discrete version of this penalty), or

$$\Phi = \iint \left[\left(\frac{\partial^2 \gamma}{\partial u^2} \right)^2 + \left(\frac{\partial^2 \gamma}{\partial v^2} \right)^2 \right] \mathrm{d}u \, \mathrm{d}v,$$

where $\gamma = \sqrt{f}$ [4]. Granville & Rasson [5] proposed and investigated a penalty based on $g = \log f$.

Penalized likelihood estimation has also proven valuable in the estimation of cell probabilities in large sparse **contingency tables**. Consider a K-cell **multinomial** vector $n_i, i = 1, \ldots, K$, with $\sum n_i = n$ and underlying cell probabilities $p_i, i = 1, \ldots, K$. Under the usual asymptotic model of a fixed number of cells K with $n \to \infty$, the frequency estimator $\overline{p}_i = n_i/n, i = 1, \ldots, K$, is **consistent** and fully **efficient**.

Tables in which the sample size is not large compared with the number of cells are called sparse tables, and the usual asymptotics (*see* **Large-sample Theory**) do not provide a reasonable model for them. Sparse asymptotics, where K and n both become infinite at the same rate, provide a theoretical framework for the analysis of large sparse tables. In this situation the frequency estimator is no longer useful, since its good properties require the number of observations in each cell to become infinite. The frequency estimator is not sparse asymptotic consistent, in the sense that

$$\sup_{1 \le i \le K} \left| \frac{\overline{p_i}}{p_i} - 1 \right| \neq o_p(1).$$

In contrast, maximum penalized likelihood estimation leads to a sparse asymptotically consistent estimator. The estimator is the maximizer of

$$\sum_{i=1}^{K} n_i \log p_i - \alpha \sum_{i=1}^{K-1} (\log p_i - \log p_{i+1})^2, \alpha > 0,$$

where α is the smoothing parameter. Assuming appropriate smoothness and boundary conditions on the underlying probability vector $\mathbf{p}$, the MPLE $\hat{\mathbf{p}}$ is sparse asymptotic consistent, with convergence rate [16]

$$\sup_{1 \le i \le K} \left| \frac{\hat{p}_i}{p_i} - 1 \right| = O_p[K^{-2/5} (\log K)^{2/5}].$$

An application of maximum penalized likelihood estimation to sparse categorical data is illustrated in Table 1. The data, originally from [9], and further analyzed in [11] and [17], represent the performance of staff and trainees in correctly interpreting diagnostic tests given to psychologically disturbed patients, as measured by counts of people at the different performance levels. The question of interest here is whether the performance of staff and trainees differs.

In Figure 1 are given the observed relative frequencies [part (a)], and the (smoothed) penalized likelihood estimates [part (b)] (staff probabilities are represented by solid lines connecting $\times$'s, while trainee probabilities are represented by dotted lines connecting $\circ$'s). Because of the roughness in the estimates, it is difficult to see from the unsmoothed relative frequencies what the relative shape of the distributions is. In contrast, the smoothed estimates clearly show multimodality in both group's scores, with lower scores (less than 72) more probable for trainees, and higher scores (greater than 72) more probable for staff members.

Notes

1. IMSL provides a Fortran subroutine to calculate a discrete maximum penalized likelihood density estimate.
2. RKPACK-II, a collection of Ratfor routines for penalized likelihood density estimation, is available using a World Wide Web browser at the URL, at `http://www.stat.purdue.edu/~chong/rkpk2.shar.gz`.

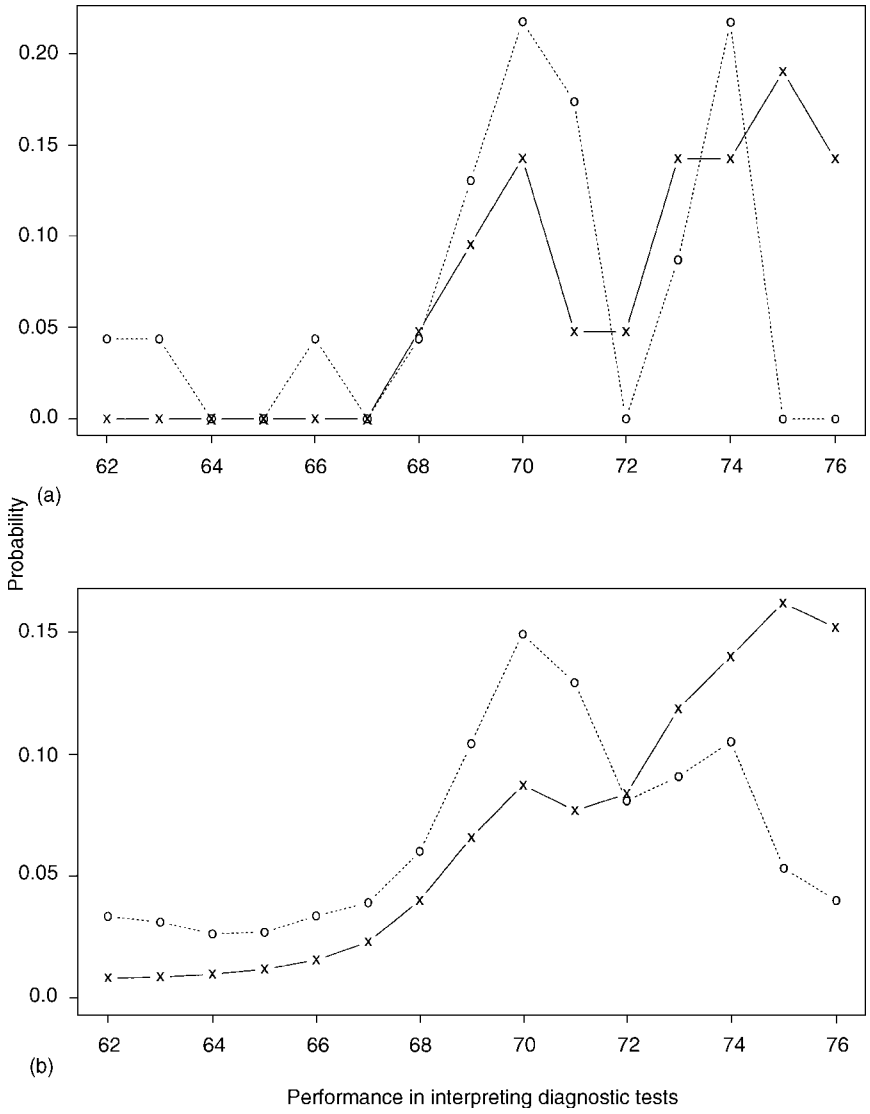

Figure 1 (a) Observed relative frequencies; (b) smoothed likelihood estimates

References

[1] Cox, D.D.& O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators, *Annals of Statistics* **18**, 1676–1695.

[2] de Montricher, G.M., Tapia, R.A.& Thompson, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods, *Annals of Statistics* **3**, 1329–1348.

[3] Good, I.J.& Gaskins, R.A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika* **58**, 255–277.

[4] Good, I.J., Holtzman, G.I., Deaton, M.L.& Bernstein, L.H. (1989). Diagnosis of heart attack from two enzyme measurements by means of bivariate probability density estimation: statistical details, *Journal of Statistical Computation and Simulation* **32**, 68–76.

[5] Granville, V.& Rasson, J.P. (1995). Multivariate discriminant analysis and maximum penalized likelihood density estimation, *Journal of the Royal Statistical Society, Series B* **57**, 501–518.

[6] Gu, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm, *Journal of the American Statistical Association* **88**, 495–504.

[7] Klonias, V.K. (1982). Consistency of two nonparametric penalized likelihood estimators of the probability density function, *Annals of Statistics* **10**, 811–824.

[8] Klonias, V.K. (1984). On a class of nonparametric density and regression estimators, *Annals of Statistics* **12**, 1263–1284.

[9] Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical production, *Psychological Monographs* **76**, No. 28.

[10] O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators, *SIAM Journal on Scientific and Statistical Computation* **9**, 363–379.

[11] Pettitt, A.N. (1984). Tied, grouped continuous and ordered categorical data: a comparison of two models, *Biometrika* **71**, 35–42.

[12] Scott, D.W., Tapia, R.A.& Thompson, J.R. (1978). Multivariate density estimation by discrete penalized likelihood methods, in *Graphical Representation of Multivariate Data*, P.C.C. Wang, ed. Academic Press, New York, pp. 169–181.

[13] Scott, D.W., Tapia, R.A.& Thompson, J.R. (1980). Nonparametric probability density estimation by discrete penalized likelihood criteria, *Annals of Statistics* **8**, 820–832.

[14] Silverman, B.W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method, *Annals of Statistics* **10**, 795–810.

[15] Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method, *Annals of Statistics* **12**, 898–916.

[16] Simonoff, J.S. (1983). A penalty function approach to smoothing large sparse contingency tables, *Annals of Statistics* **11**, 208–218.

[17] Simonoff, J.S., Hochberg, Y.& Reiser, B. (1986). Alternative estimation procedures for $\Pr(X < Y)$ in categorized data, *Biometrics* **42**, 895–907.

(*See also* **Density Estimation**)

JEFFREY S. SIMONOFF

Penetrance

There is a relationship between the underlying **genotype** and the observed phenotype of an individual that must be defined in describing the genetic basis of a trait. The function that describes this relationship is called the *penetrance function*, and is defined in a general sense as the conditional probability that an individual with genotype g expresses phenotype y : $\Pr(y|g)$ for a discrete trait, or $\phi(y|g)$ for a continuous trait, where ϕ is the probability density function. Specification of such a penetrance function, along with associated parameter estimates, is a necessary component in computations that are based on **maximum likelihood** methods and requires relationships between genotypes and phenotypes (*see* **Genetic Counseling**; **Linkage Analysis, Model-based**; **Segregation Analysis, Complex**).

When the relationship between the genotype and the phenotype is simple, as is the case for most **blood groups**, this conditional probability is either 0 or 1, depending on the particular combination of genotype and phenotype. For many traits, however, the relationship between specified underlying genotypes and possible resulting phenotypes is not so clear. For these traits, the penetrance function can be thought of as a quantitative model that summarizes our lack of understanding about the underlying mechanisms relating genotype to phenotype. In defining the penetrance function in these cases, we make the implicit assumption that $\Pr(y|g)$ is an average over all possible individuals with genotype g, and that these individuals have been randomized for all factors that have an impact on this probability.

There are a number of commonly used penetrance functions. For some phenotype–genotype combinations, $\Pr(y|g) = k$, where $0 < k < 1$ but k is a constant. This model is called a *fixed-penetrance* model, and has been well described in experimental organisms. For example, in the fruit fly *Drosophila melanogaster* there is a **gene** for which 90% of homozygotes for allele i have an interrupted wing vein [5]. This *reduced penetrance* probability remains 90% in offspring of matings between obligate homozygotes for the i allele, regardless of whether the i/i parents have normal or interrupted wing veins. In humans, there are many genetic disorders that behave in a similar manner to this reduced penetrance example in flies. For example, in the autosomal dominant disorder split-hand deformity, $\Pr(\text{disease}|\text{genotype} = Dd) \cong 0.7$, where D is the disease allele and d is the normal allele [7]. For phenotypes that are easily dichotomized into normal vs. abnormal phenotypes, this narrower definition of penetrance is often used: that of the *penetrance* of a genotype, where it is specifically the penetrance $\Pr(y = \text{disease}|g)$ that is of interest, possibly for different underlying genotypes, g. This narrower definition of penetrance is used only in the context of inherited diseases (in humans) or unusual phenotypes (in nonhumans).

Frequently, the penetrance differs among individuals as a function of some identifiable **covariate**, which can then be incorporated into the model. Covariates might include such intrinsic characteristics of an individual such as age or sex, or might include extrinsic characteristics such as diet or exposure to some environmental factor. The penetrance function may then be defined as $\Pr(y|g, c)$, where c may be a vector of covariates. It is quite common, for example, to assume that males and females have different penetrance functions for certain phenotypes, such as the probability of baldness.

Age is a particularly common covariate in penetrance functions for human diseases. Examples include breast cancer, Alzheimer's disease, and Huntington's disease. For each, the probability that disease has developed increases with the age of the individual. Typically, either a cumulative **normal distribution**, or a straight-line penetrance function (with a minimum and maximum age of penetrance and a linear increase between them) is used. Other functions are also possible, including use of liability classes for age intervals, with fixed but increasing penetrance for each interval [6]. The parameters for the functions will vary across genotypes, including the possibility of 0% penetrance at all ages for some genotypes (e.g. the nonsusceptible genotype for Huntington's disease), or the possibility of sporadic forms of the disease with later mean onset than the "genetic" forms (e.g. early-onset Alzheimer's disease [3]). In model-based analyses of diseases with sporadic cases, computation of **likelihood** for linkage analysis (*see* **Linkage Analysis, Model-based**) or **genetic counseling** requires using the density function and age at onset for affected individuals, and the cumulative distribution function and age at last examination or at death for unaffected individuals [4] since ultimately what is desired is the reverse probability, $\Pr(g|y)$.

A disease with age-dependent penetrance can also be affected by other covariates. For example, $\varepsilon 4/\varepsilon 4$ homozygotes at the ApoE locus have a lower mean onset age of Alzheimer's disease than do individuals with other ApoE genotypes [2]. Separate age-dependent penetrance functions for different ApoE genotypes increased the likelihood in a linkage analysis [3], as is expected if this model fits the data better than one which did not account for this discrete covariate.

The penetrance function may, in principle, also take into account genotypes at multiple loci, e.g. $\Pr(y|g_1, g_2)$, where g_1 and g_2 are genotypes at two loci. In humans, relatively few such multilocus systems have been characterized well enough to define precise multilocus penetrance functions. Until the specific interacting loci have been identified, the penetrance function will generally be parameterized in terms of only one locus, which should be adequate for describing penetrance in unrelated individuals unless there is **linkage disequilibrium** at the population level between this locus and additional contributory loci. However, for computations on related individuals, the increased sharing of alleles identical-by-descent (*see* **Identity Coefficients**) in closely, vs. more distantly, related individuals means that a single-locus penetrance function will only approximate the correct penetrance probabilities in computations on pedigrees. While this has little effect on initial linkage analyses [1], the resulting model-misspecification may reduce the ease by which such a disease locus can be localized to a very small genomic region by **multipoint linkage analysis**.

References

[1] Greenberg, D.A. & Hodge, S.E. (1989). Linkage analysis under "random" and "genetic" reduced penetrance, *Genetic Epidemiology* **6**, 259–264.

[2] Jarvik, G.P., Larson, E.B., Goddard, K., Schellenberg, G.D. & Wijsman, E.M. (1996). Influence of apolipoprotein E genotype on the transmission of Alzheimer disease in a community-based sample, *American Journal of Human Genetics* **58**, 191–200.

[3] Levy-Lahad, E., Wijsman, E.M., Nemens, E., Anderson, L., Goddard, K.A.B., Weber, J.L., Bird, T.D. & Schellenberg, G.D. (1995). A familial Alzheimer's disease locus on chromosome 1, *Science* **269**, 970–973.

[4] Margaritte, P., Bonaiti-Pellie, C., King, M.C. & Clerget-Darpoux, F. (1992). Linkage of familial breast cancer to chromosome 17q21 may not be restricted to early-onset disease, *American Journal of Human Genetics* **50**, 1231–1234.

[5] Rothwell, N.V. (1983). *Understanding Genetics*, 3rd Ed. Oxford University Press, Oxford, pp. 82–86.

[6] Terwilliger, J.D. & Ott, J. (1994). *Handbook of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore, Chapter 9.

[7] Thompson, M.W., McInnes, R.R. & Willard, H.F. (1991). *Genetics in Medicine*, 5th Ed. Saunders, London, p. 83.

(*See also* **Gene-environment Interaction**; **Genetic Liability Model**)

ELLEN M. WIJSMAN

Person-years at Risk

In follow-up studies of subjects subsequent to various treatments or exposures and in the study of chronic disease where the **incubation period** or length of illness may be months or years, consideration must be given to the time the subjects were under observation or to the time intervening between the initial exposure and the eventual outcome, e.g. recovery, onset of disease, or death. If the probability of a given outcome is related to time, outcome measures are affected by the length of the observational period [1].

Person-years at risk are units of measurement which combine persons and time by summing individual units of time (years and fractions of years) during which subjects in a study population have been exposed to the risk of the outcome under study. A person-year is defined as the equivalent of the experience of one individual for one year [2]. Each subject contributes only as many years as he or she has been actually observed (or exposed); a subject under observation for one year contributes one person-year; six months would contribute one-half person-year, etc. Person-years at risk frequently comprise the denominator of calculations of **incidence rates** measured over extended and variable time periods or of measures of morbidity and mortality resulting from chronic exposure to environmental hazards such as industrial toxic waste materials or cigarette smoke.

References

[1] Kahn, H.A. & Sempos, C.T. (1989). *Statistical Methods in Epidemiology*. Oxford University Press, New York.

[2] Last, J.M., ed. (1983). *A Dictionary of Epidemiology*. Oxford University Press, New York.

ROBERT A. ISRAEL

Person-years of Life Lost

For public health planning purposes it is important to have measures of the impact of particular diseases and of the potential impact of preventive measures, and one of the most useful and intuitive of such measures of impact is *person-years of life lost*, sometimes termed, *potential years of life lost*, and generally abbreviated as *PYLL*.

In principle, the measure is straightforward. In a stationary or **life table** population, the person-years of life lost due to a particular cause of death represents the difference between the expectation of life of the population and the expectation of life with the cause in question eliminated (*see* **Life Expectancy**). Standard methods are available for the calculation of cause-deleted life tables [2]. In an actual population, the PYLL are obtained by taking the deaths at each age due to the cause being examined, and calculating the years those people would have lived according to the cause-deleted life table [1]. Like life expectancy, the PYLL may be calculated from birth or any other age, or over any given age range. Also like life expectancy, it gives greatest weight to the deaths at the youngest ages.

In practice, there are a number of complications, arising from two main sources: judgments about the value of life at different ages; and the problem of **competing risks**.

Person-years of life lost is often used as a measure in cost-of-illness studies, and in this context it is common to restrict its calculation to the years of economically active life (e.g. between 15 and 65). In other contexts, it is generally calculated up to a particular upper age limit, e.g. 65, 70, etc. The use of such cutoffs is virtually universal, despite the obvious theoretical weakness in arbitrarily weighting or valuing all years lived (or lost) inside the limits at unity and all those lived outside the limits at zero.

One reason for excluding deaths at older ages is the problem of **competing risks** and the difficulty of determining a single cause of death in older people with multiple potentially lethal conditions. Competing risks are, in fact, an issue at all ages, and the various approximate methods which have been used to calculate the index differ principally in the extent to which they attempt to come to terms with competing causes of death. It is still common, for instance, to base the calculation on the assumption that, if the cause were eliminated, the individuals "saved" would all survive to the upper age limit being used (e.g. 65 years). That clearly overestimates the years lost, because in that period such individuals would remain at risk of dying from other causes. A second approximation involves applying the total mortality to the lives saved; this underestimates the PYLL, because it includes the cause that is meant to have been eliminated.

Apart from these technical issues, the principal shortcoming of the method is that basing it on cause-deleted tables incorporates an inadequate model of disease causation and prevention. Public health activities are principally aimed at elimination or reduction of risk factors rather than causes of death, and as in the case of smoking, one risk factor may result in deaths from a variety of causes. Where appropriate, the use of **attributable fractions** to convert PYLL due to particular causes to PYLL due to particular risk factors is a more realistic approach.

References

[1] Hakulinen, T. & Teppo, L. (1976). The increase in working years due to elimination of cancer as a cause of death, *International Journal of Cancer* **17**, 429–435.

[2] Keyfitz, N. (1977). *Applied Mathematical Demography*. Wiley, New York.

L. SMITH

Petty, William

Born: May 26, 1623, in Romsey, Hampshire, UK.
Died: December 16, 1687, in London, UK.

Petty was a precocious youth. His father was a clothier, and the young Petty learnt many of the skills of the business while he was learning Latin and Greek at school. At the age of 14 he became a cabin boy and attracted attention in France, during a period of recovery from a broken leg, as "le petit matelot anglais qui parle latin et grec". He spent a few years abroad, teaching English and navigation, and studying at the Jesuit College in Caen. He returned to England to join the Royal Navy, but in 1643, after the outbreak of the Civil War, he returned to the continent, studying medicine and mathematics in Utrecht, Amsterdam, and Leyden. After a period in Paris, he returned to England in 1646, joining his father's business. During this period, he sketched the idea of a national scientific society, later to be embodied in the Royal Society, of which he would become a founder member.

After a while, he moved to Oxford to continue his medical studies, and in 1649 he was made a Doctor of Physic and Fellow of Brasenose College. By 1651 he had become Professor of Anatomy and Vice-Master of Brasenose.

In about 1652 he was appointed Professor of Music at Gresham College, apparently with the influential help of his friend **John Graunt**. But almost immediately he left for Ireland, where the Commonwealth government was engaged in resettlement and in the forfeiting of estates of Irish landowners who were in debt to the government. For this purpose a survey of these estates was required, and Petty was entrusted with the task, which eventually expanded to the complete mapping of Ireland. He became involved in Commonwealth politics, but after the Stuart restoration in 1660 he found himself on good terms with the monarchists, and was knighted by Charles II on the granting of the Charter to the Royal Society in 1662. Petty had by now become a key figure in London intellectual circles.

In the meantime, Graunt, hitherto a prosperous trader, had become bankrupt, and Petty was able to repay Graunt's earlier generosity by giving him material support until Graunt's death in 1674.

Petty shared Graunt's interest in demographic matters (*see* **Demography**), and published many essays on **vital statistics**, inventing the term "political arithmetic". He was not a pioneer of Graunt's calibre, and earlier rumours that he had written Graunt's *Observations on the Bills of Mortality* appear to be unfounded. His economic writings include a *Political Anatomy of Ireland*, and *A Treatise of Taxes and Contributions*. For a modern statistician perhaps his greatest contribution was the advocacy of a national statistical office, to enumerate population (*see* **Censuses**), evaluate property, organize tax collection and improve the public health. His assessment of the importance of official statistics can perhaps be judged from this delightful quotation: "God sent me the use of things, and notions, whose foundations are secure and the superstructures mathematical reasoning; for want of which props so many Governments doe reel and stagger."

Greenwood's epitaph [2] is appropriate: "Careless, happy-go-lucky, tendentious; yes, all of that. But anyone who has felt the exhilaration ... in the doing of sums concerning biological problems, feels his heart warmed by the arithmetic knight errant who had so many statistical adventures" (see also [1] and [3]).

References

[1] Fitzmaurice, E. (1896). Petty, Sir William, *Dictionary of National Biography* **45**, 113–119.

[2] Greenwood, M. (1948). *Medical Statistics from Graunt to Farr*. Cambridge University Press, Cambridge.

[3] Strauss, E. (1954). *Sir William Petty, Portrait of a Genius*. Bodley Head, London.

PETER ARMITAGE

Pharmaceutical Industry, Statistics in

Pharmaceutical statistics is an emerging discipline that is going through the classical stages of growth of a developing science: first, as a loose collection of topics; next, as a recognized subspecialty; and finally as a largely separate discipline. In this respect, it is following closely on the heels of medical statistics as it emerges from the science of statistics, which in its turn has followed the emergence of statistics from mathematics. These developments can be traced through the founding of journals devoted to these respective subjects. **Karl Pearson** founded ***Biometrika*** in 1900 because he found that his papers were too mathematical for the biologists and too biological for the mathematicians. ***Biometrics*** was started after World War II, amongst other reasons, because *Biometrika* was perceived as being too abstract for the life scientist. ***Statistics in Medicine*** has had a phenomenal growth since its beginning in the 1980s and this is surely at least partly due to the increasing numbers of statisticians working exclusively in medicine and with no interest (say) in agriculture. More recently, a new journal has appeared, ***The Journal of Biopharmaceutical Statistics***, and although we can expect that many papers on pharmaceutical statistics will continue to appear in *Statistics in Medicine* in the same way that papers on medical statistics continue to appear in *Biometrics*, no doubt we can expect to see a growth in this more specialist outlet for pharmaceutical statistics. Perhaps eventually we will even see the need for a specialist *Encyclopedia of Pharmaceutical Statistics*!

Obviously, there is still a considerable overlap between pharmaceutical statistics and medical statistics. There are, however, many topics which are given a different emphasis even if attention is restricted to a field of common interest: the design and analysis of **clinical trials**. For example, two of the most important topics of research in the statistics of clinical trials in the last quarter of the twentieth century have been **survival analysis** and group sequential (*see* **Data and Safety Monitoring**) methods. Although both of these are of some importance to pharmaceutical statistics, in fact it is only a small minority of trials in which survival is the main outcome, and even fewer which are run sequentially. On the other hand, **crossover trials** are much more important than in medical statistics generally and there are various other topics such as **pharmacokinetic** modeling that really apply only in connection with drug development [18].

Drug Development

Before describing the role of the statistician in drug development, it is appropriate to say something about drug development itself. It is the process of finding and producing therapeutically useful pharmaceuticals and of turning them into high-quality formulations of usable, effective, and safe medicines. It is also the process, however, of delivering valuable, reliable and trustworthy information about appropriate doses and dosing intervals and about likely effects and side effects of these treatments. Drug development is carried out by *sponsors* (mainly pharmaceutical companies), and its acceptability is judged by *regulators* such as the **Food and Drug Administration (FDA)** of the US or the **Medicines Control Agency (MCA)** in the UK. It is an extremely complex business and the risks are high, but the potential rewards are also considerable.

The phrase *drug development* is used in two different senses within the pharmaceutical industry. In its more general sense, it covers all the activities leading to the eventual marketing of a pharmaceutical. In a more restricted sense, a distinction is made between drug research, which covers the study of basic mechanisms of action and the identification of candidate substances, and drug development proper, which covers the business of producing suitable formulations and studying their effects in man.

It takes many years for a project to reach drug development proper. First, basic research must be undertaken to validate concepts and mechanisms. The choice of therapeutic areas to investigate will depend on commercial potential and this in turn requires a thorough understanding of disease areas and current therapies in order to establish unmet medical need. Intelligence reports regarding these matters will continue to be produced throughout the life of a project. Next, a lead compound must be identified for a particular indication. This will then be subject to a battery of screening tests to

assess its potential in terms of therapeutic activity. Back-up compounds will also be investigated. If a compound looks promising it will also be evaluated from both safety and practical points of view. Will it be easy to formulate? How many steps are involved in the synthesis? How difficult will it be to manufacture in large quantities? Before a treatment can go into development, not only must satisfactory answers have been obtained to all these questions, but a viable pharmaceutical formulation permitting further study must be available. This can be an extremely delicate matter, involving work to develop suitable solutions, pills, patches, or aerosols, as the case may be (*see* **Preclinical Treatment Evaluation**).

If and when a molecule is accepted into drug development proper, animal studies will be undertaken in order to check safety and to establish a dose at which studies in man may be undertaken. Once basic toxicological work has been undertaken, **phase I** may begin and the first such studies may start. These will be single-dose studies in which lower doses are tried first and cautiously increased until a maximum tolerated dose may be established. In many indications such studies are carried out on healthy volunteers but where the treatment is highly aggressive (and hence intended for serious diseases) patients will be used instead. In the meantime longer-scale toxicological studies with animals will have been completed. Pharmacokinetic studies in man will be undertaken in which the concentration–time profile of the drug in blood will be measured at frequent intervals in order to establish the rate at which the drug is absorbed and eliminated (*see* **Bioavailability and Bioequivalence**). These studies together, if successful, will permit multiple dose studies to be undertaken.

Once maximum tolerated doses have been established, **phase II** begins and dose-finding studies in patients are started. This is usually an extremely difficult phase of development but, if the drug proves acceptable, the object is that preliminary indications of efficacy should be available and that a firm recommendation for doses and dose schedules should emerge. Once these studies have been completed, the pivotal phase III studies can begin. These have the object of proving efficacy to a skeptical regulator and also of obtaining information on the safety and tolerability of the treatment.

A successfully completed development program results in a dossier: an enormous collection of clinical trial and other reports, as well as expert summaries covering not only the clinical studies as regards efficacy and safety but also preclinical studies and other technical reports as well as details of the manufacturing process. The purpose of this dossier is to reassure the regulator as to quality, safety, and efficacy of the pharmaceutical. If successful, the package leads to registration, but even during the review process, phase IV studies may have been initiated in order to discover more about the effect of the treatment in specialist subpopulations, or perhaps with the object of providing data to cover price negotiations with purchasing authorities. These *reimbursers* may include national health services but also agents acting for private health care plans.

Once a drug has been launched on the market the process of monitoring and "pharmacovigilance" begins in earnest, since the drug will now be used by far more people than was ever the case in the clinical trials in phases I to III, and rare side effects, which could not be detected earlier, may now appear [18]. Some further phase IV **postmarketing surveillance** studies may be initiated and further work extending indications or preparing new formulations may be undertaken.

It is important to stress that all this work is carried out to an extremely high standard [1, 5, 9, 18]. A sponsor will have standard operating procedures to cover all aspects of the trial, from informed consent for patients (*see* **Ethics of Randomized Trials**), to report writing, passing such diverse matters as the handling of human specimens, monitoring, drug disposal and accounting, source data verification, database management (*see* **Data Management and Coordination**), and, of course, statistical analysis en route. These, in turn, will reflect regulatory guidelines [2]. The results will be closely examined by regulators and this examination will, in some cases, involve reanalysis of the data and even study-site visits [4, 5, 9] (*see* **Drug Approval and Regulation**). Of course, in-house quality control and assurance also receives a great deal of attention and the typical sponsor will have a number of auditors and monitors, who, although not necessarily statisticians, will make important contributions to data quality. The net result is that the average standard that applies to clinical trials within the pharmaceutical industry is far higher than that which applies outside.

Statistics in Drug Development: The Current Position and Recent Past

A top pharmaceutical company will employ between 50 and 150 statisticians in a number of sites worldwide dealing with various aspects of drug development and research. Activities in which statisticians can become involved include [18]:

1. Optimization in chemical, pharmaceutical, or manufacturing development
2. Animal toxicology experiments
3. Bioassay (*see* **Biological Assay, Overview**)
4. Project prioritization and portfolio management
5. Quality control
6. Pharmacoeconomics (*see* **Pharmacoepidemiology, Overview**)
7. Pharmacokinetics
8. Epidemiology and drug monitoring
9. Design and analysis of clinical trials.

The last of these has always been the most important area of activity for the pharmaceutical statistician but others, in particular epidemiology and pharmacoeconomics, are growing in importance. The potential for the statistician to contribute to portfolio management is also considerable [17, 18].

The randomized clinical trial (RCT) provides the *fiducial framework* which permits the regulator to mandate the sponsor to carry out drug development. The allocation of patients to treatment usually involves an element of **randomization**, thus reducing the sponsor's ability to manipulate results. Each trial has an extensive protocol covering all aspects of its conduct. It will also include a detailed description of the data which will be collected and the intended analysis and these must be presented or accounted for in the final report (*see* **Clinical Trials Protocols**). A concern, sometimes amounting to an obsession, with prespecified analyses is, in fact, one of the characteristic features of pharmaceutical statistics.

It is the norm, of course, for trials to be designed jointly by at least one physician and one statistician. Indeed, the Good Clinical Practice Guidelines of the European Union state, "Access to biostatistical expertise is necessary before and throughout the entire procedure, commencing with the design of the protocol and ending with completion of the Final Report." There has been a remarkable growth in the number of statisticians employed by the pharmaceutical industry since the 1970s. The first phase of this growth was in the primary industry, but latterly there has been an explosion in the secondary industry: the Contract Research Organizations (CROs) (*see* **Proprietary Biostatistical Firms**). This is illustrated by the change in membership of **Statisticians in the Pharmaceutical Industry (PSI)**, a UK-based body, since it was founded in the late 1970s. In the late 1980s, through the creation of the category of Associate Member, it was opened to those not working in the primary industry and they now exceed ordinary members in number.

This growth in pharmaceutical statistics has taken place both in North America and in Europe and, as regards the statistical expertise available to sponsors, there is no disparity between the two regions. (Japan, although actively involved in the harmonization of drug regulations, has lagged behind in this respect.) As regards regulators, however, until recently the discrepancy was considerable. The FDA employs about 120 statisticians in human health but until the 1990s there were none in the European Union. The **Royal Statistical Society** (RSS) was so concerned by this state of affairs that it issued a report in an attempt to encourage statistical representation in drug regulation [12]. The position has now improved considerably, with statisticians, for example, in Germany, Sweden and the UK. This is a welcome development, since employing physicians to consider statistical arguments is clearly an inefficient and ineffective use of resources.

Statistical topics which have been of particular interest to pharmaceutical statisticians, whether working for regulators or sponsors, are:

1. Design and analysis of crossover trials [16]
2. **Equivalence** testing [18]
3. Pharmacokinetic and pharmacodynamic models [10, 14, 21]
4. Repeated measures
5. Dose finding [22]
6. **Bayesian methods**
7. **Nonlinear random effect** modeling [10, 21]
8. Professional and practical statistical matters [1, 5].

Two Examples

Statisticians employed in the industry can be involved in a wide variety of tasks. Two examples are given here to illustrate the breadth.

Formoterol

The first concerns the development of a new formulation of a drug, formoterol (eformoterol in the UK) [19]. Formoterol is a beta-agonist used to treat asthma and a series of trials carried out over a number of years had succeeded in demonstrating that it had a rapid onset and 12 hours duration of action [11]. This was a considerable improvement over the drugs then available, which had a 4–5 hour duration.

A single-dose dry powder formulation had been developed and it was desired to show that a dry powder multidose formulation would be equivalent. Normally, one would attempt to do this by measuring concentration–time profiles of the active substance in the blood, but this was impossible because (a) at 12 μg for a standard dose (80 000 doses weigh less than a gram), formoterol is not detectable by bioassay, and (b) even if it were, 90% of it is swallowed and although some of this is eliminated in "the first pass", the portion that remains might be a large (but irrelevant for efficacy) part of what you would detect. Hence, it is necessary to study the effects in patients to compare the formulations.

This raises an issue of sensitivity, however, and to show that equivalence is not falsely concluded simply because the top of the **dose–response** curve has been reached for each formulation, it is necessary to compare various doses of each formulation: a parallel groups design. It is also necessary to include a placebo (*see* **Blinding or Masking**). As a result, it was decided to compare three doses of each formulation with a placebo. There were thus seven treatments. **Sample size calculations** showed that for a parallel group trial 200 patients per group would be needed for the target precision. For these sorts of numbers, however, one might as well start a development from scratch. On the other hand, a seven period crossover was out of the question: for reasons of patient compliance, five was the most periods that could be contemplated.

In the end an **incomplete blocks design** was chosen in 21 sequences replicated six times and five periods chosen in such a way that each patient received five different treatments. Each treatment appeared equally often in each period, so that each pair of treatments was equally represented amongst the 126 patients. The design proved to be a considerable success but the new formulation was not. The trials showed unequivocally that the new formulation was less potent: at least 24 μg of the multidose dry powder were necessary to show the same effect as 6 μg of the single-dose formulation.

Project Prioritization

The second example concerns work on project prioritization: how does one choose which drug development projects to pursue [17]? Everyone agrees that probability of success, cost to develop, time to develop, and potential sales are the important considerations, but there is no general consensus on how to combine them.

A simple insight is sufficient to show, however, that it is necessary to dig deeper into projects to establish their worth. If two drugs have identical overall probabilities of success, identical overall costs to develop, identical rewards and so forth, any index based on weighting these factors would have to score them identically. However, if one project were such that it would fail early if it failed it would be much more valuable than another which would fail late if it failed. Hence, what is needed is an index that goes into the cost and probability architecture of the projects. Such indices can be based on **decision analysis**, a subject developed originally by statisticians, and it is thus just as important to the pharmaceutical industry for statisticians as for marketers to be involved in this problem.

Statistics, Drug Development, and the Future

The pharmaceutical industry is currently facing a strong economic challenge as a consequence of rapidly increasing health care costs and price curtailment policies. This has various implications for the work of the pharmaceutical statistician. First, price regulation is forcing sponsors to consider pharmacoeconomics and to provide reimbursers with evidence regarding "value for money". There has been a rapid growth in the number of pharmacoeconomic studies. Many of these have been carried out through the medium of the RCT. The RCT is not, however, ideally suited to this and it will be necessary in the future to combine information from a number of sources using complex models, whereby current knowledge is used to predict the probable health effects and cash flow resulting from introduction of

a pharmaceutical. This will be a great collaborative opportunity for the statistician. Since the fiducial framework of the RCT does not cover this, it is doubtful that reimbursers will be able to mandate sponsors to carry out these studies and as a consequence we shall see the rise of third party analysis. Of course, the industry will also have to carry out these analyses, if only to predict what conclusions others will come to.

Secondly, specialization is the order of the day, and more and more work of all kinds, including statistical work, is being subcontracted by sponsors. The CROs seem to provide sponsors with an attractive option to disinvest: something which is only done with difficulty with your own staff. However, business is value added and if the CROs are adding the value, this is an area of business lost to the primary industry. Where will it all end? Will CROs eventually become the commissioning partners and trawl the primary industry for molecules to develop? Will the major pharmaceutical companies gradually turn themselves into nothing more than the providers of venture capital?

The third economic concern of the industry is speed. All sponsors are trying to shorten development times. One consequence for statistics is that the time available to complete the final analysis has been reduced from weeks to days. This means, of course, that all computer programs have to be developed before the trial is over and validated and run on blinded data before finally being applied to the "cleaned" and decoded database. The irony is that this practice, a reflection not only of economic but also of regulatory pressure for prespecified analyses, means that at a time when **exploratory data analysis** has never been easier, it has been rendered virtually obsolete. This development is not all bad: far from it. Exploratory data analysis has always carried with it the danger of overinterpretation. Far more attention will have to be given to prior specification, whether the framework is frequentist or **Bayesian** (*see* **Inference**). Nevertheless, the trend is worrying and many companies would do well to consider the possibility of creating secondary databases for nonregulatory analyses: the sort that could be carried out at leisure and whose results could inform the design of further trials. More serious, however, is that some companies are reducing development times by moving towards doing activities in parallel. This is extremely costly when a project fails and what is needed is to carry out the sort of statistical decision analysis described above to see if it is justified.

Finally, the new concern with economics means marketing will have to change. The golden era of the "detail man" is over and marketing will be carried out by highly qualified sellers to a few specialist buyers. Evidence (which in this context is just another word for data and statistical analysis) will become more important in selling the product and the need for statistical support for marketing will increase.

Subgroup analysis (*see* **Treatment-covariate Interaction**) will become more important. Recently, the **National Institutes of Health (NIH)** and FDA have insisted that women and minorities be adequately represented in clinical trials [7, 8]. This is not as simple as it might seem. If a highly effective treatment for lung cancer is found it will have been found on the basis of studies where most patients were male. (To insist that the same numbers of females should be represented would simply delay recruitment and the eventual proof of efficacy). Can it then only be prescribed for men? Must further trials be run until the same number of women have been studied? Bayesian methods or frequentist approaches which consider **mean square error** will be needed to deal with this problem [18]. The issue can only become more important since the Human Genome Project will deliver ever more ways (and perhaps more relevant ways) of classifying patients into subgroups [15]. (It may, of course eventually deliver a host of novel therapies, although there are reasons for not hoping for too much in the near future from this quarter [3].)

An important movement affecting clinical practice is **evidence-based medicine** [13], an idea which originated at McMaster with Sackett and colleagues. This requires the individual physician to integrate global information on the efficacy of treatment into his or her daily medical practice by formulating precise questions concerning the care of the individual, searching medical databases (*see* **Administrative Databases**) for the answer, critically appraising the evidence, deciding on a treatment, and carrying out a follow-up evaluation. A president of the RSS has even suggested that it should be extended to public policy making generally [20]. If it does it will only increase the pressure and need for more information. The quality of information available in dossiers submitted to regulatory authorities far exceeds that in

published reports. This raises an interesting possibility: will there soon be a requirement that regulatory dossiers, together with standard summaries, be made available on-line? If so, will there be commercial organizations whose sole purpose is to "surf the Web" and produce analyses which they will then sell to health care purchasers (*see* **Internet**)? Will the industry statistician be constantly looking over his or her shoulder, knowing that all analyses will eventually be repeated and modified by others who come after?

So far potential developments in pharmaceutical statistics have been looked at from the point of view of the requirements of the user. What of statistics itself? Here, various trends can be discerned. The first is the rapid rise in the use of Bayesian methods. The next few decades will show whether these are so successful that frequentist methods completely disappear by the time the DeFinetti–Lindley limit of 2020 has been reached [6]. However, whether or not we all become Bayesians, there are two Bayesian lessons that all pharmaceutical statisticians are having to learn. The first is the importance of **random effects** models; for example, models that permit appropriately combining information obtained from repeated measures on a given patient in a particular center with further information on the response of all patients in that center, and even with information on all patients in the trial, in order to produce better descriptions of the effect of the treatment for him or her. The second lesson is the importance of using prior knowledge, in particular biological knowledge, when choosing models for analysis (*see* **Model, Choice of**).

One of the advantages brought by progress in computing is that more difficult calculations can be undertaken. Currently there is a lack of satisfactory methods for dealing with **missing data**, patients who take rescue medication, noncompliance, and so forth. Some of the difficulties are inherent, but in some cases the pharmaceutical statistician is held back simply because computation is too difficult. These are topics that will receive more attention in the future simply because the ability to deal with them increases. Also included under this heading are nonlinear models, as used for example in repeated measure designs for pharmacodynamic dose–response.

Thus, a forecast of the duties of the pharmaceutical statistician and the nature of pharmaceutical statistics during the next decade might suggest:

1. That (s)he will have to know more about Bayesian methods.
2. Random effect models will be extremely important in his or her work.
3. In general, he or she will more often have to carry out complex and nonlinear modeling than is currently the case.
4. The incorporation of biological and pharmacological insights in planning and analysis will be more important.
5. The RCT will continue to be important but the proportion of statisticians not working on planning and analyzing RCTs but instead in health and economic modeling, epidemiology, and so forth will rise.
6. The statistician will become more involved in producing the clinical expert report (the overall summary of evidence for an application) [5].
7. There will be a much greater chance that the statistician will not be working in the primary industry.
8. Communication will be of the essence.

References

[1] Chuang-Stein, C. (1996). On the job training of pharmaceutical statisticians, *Drug Information Journal* **30**, 351–357.

[2] CPMP Working Party on Efficacy of Medicinal Products (1995). Biostatistical methodology in clinical trials in applications for marketing authorizations for medical purposes, *Statistics in Medicine* **14**, 1659–1682.

[3] Horrobin, D.F. (1996). The philosophical basis of drug target selection, *Pharmaceutical Medicine* **10**, 29–42.

[4] Lewis, J.A. (1983). Clinical trials: statistical developments of practical benefit to the pharmaceutical industry, *Journal of the Royal Statistical Society, Series A* **146**, 362–393.

[5] Lewis, J.A. (1996). Statistics and statisticians in the regulation of medicines, *Journal of the Royal Statistical Society, Series A* **159**, 359–362.

[6] Lindley, D. (1974). In preface to DeFinetti (translated Machi & Smith), *Theory of Probability*. Wiley, Chichester.

[7] Merkatz, R.B., Temple, R., Subel, S., Feiden, K. & Kessler, D.A. (1993). Women in clinical trials of new drugs. A change in Food and Drug Administration policy, *New England Journal of Medicine* **329**, 292–296.

[8] National Institutes of Health (1994). NIH Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research, *NIH Guide, Vol. 23, no. 10*, March 11, 1994.

[9] O'Neill, R.T. (1991). The biometrical role in U.S. drug regulations, in *Biometrie in der Chemisch-Pharmazeutischen Industrie 4*, J. Vollmar, ed. Fischer-Verlag, Stuttgart.

[10] Racine, A. & Dubois, J.P. (1989). Predicting the range of carbamazapine concentrations in patients with epilepsy, *Statistics in Medicine* **8**, 1327–1338.

[11] Richardson, W. & Bablok, B. (1992). Clinical experience with formoterol, in *Formoterol: Fast and Long-Lasting Bronchodilation*, S.T. Holgate, ed. Royal Society of Medicine Services, London.

[12] Royal Statistical Society Working Party on Statistics in Drug Regulation (1991). Statistics and statisticians in drug regulation in the United Kingdom, *Journal of the Royal Statistical Society, Series A* **154**, 413–419.

[13] Sackett, D.L., Richardson, W., Rosenberg, W. & Haynes, R.B. (1997). *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, Edinburgh.

[14] Sanathanan, L.P. & Peck, C. (1991). The randomized concentration-controlled trial: an evaluation of its sample size efficiency, *Controlled Clinical Trials* **12**, 780–794.

[15] Savill, J. (1997). Prospecting for gold in the human genome, *British Medical Journal* **314**, 43–45.

[16] Senn, S.J. (1993). *Cross-over Trials in Clinical Research*. Wiley, Chichester.

[17] Senn, S.J. (1996). Some statistical issues in project prioritization in the pharmaceutical industry, *Statistics in Medicine* **15**, 2689–2702.

[18] Senn, S.J. (1997). *Statistical Issues in Drug Development*. Wiley, Chichester.

[19] Senn, S.J., Lillienthal, J., Patalano, F. & Till. D. (1997). An incomplete blocks cross-over in asthma: a case study in collaboration, in *Cross-over Trials*, Hothorn, ed. Fischer-Verlag, Stuttgart.

[20] Smith, A.F.M. (1996). Mad cows and ecstasy: chance and choice in an evidence-based society, *Journal of the Royal Statistical Society, Series A* **159**, 367–383.

[21] Steimer, J.-L., Vozeh, S., Racine, A., Holford, N. & O'Neill, R. (1994). The population approach: rationale, methods and applications in clinical pharmacology and drug development, in *Pharmacokinetics of Drugs*, Welling & Balant, eds. Springer-Verlag, New York.

[22] Wong, W.K. & Lachenbruch, P.A. (1996). Designing studies for dose response, *Statistics in Medicine* **15**, 343–359.

STEPHEN SENN

Pharmacoepidemiology, Adverse and Beneficial Effects

Pharmacoepidemiology has been defined as "the study of the distribution and determinants of drug-related events in populations and the application of this study to efficacious drug treatment" [26]. Similar definitions have been given by several authors [37, 53]. The term "drug" in the definition is generally understood to include biologics, such as vaccines, and the populations are understood to be human. The emphasis is on studies of the safety and effectiveness of drugs used for medical purposes. Both randomized (*see* **Clinical Trials, Overview**) and nonrandomized (**observational**) designs are used, with the latter being more common, especially for the study of adverse effects. Pharmacoepidemiology may be regarded as a subdiscipline of both **clinical epidemiology** and clinical pharmacology [53]. However, clinical pharmacologists typically use small, carefully controlled studies to examine drug **pharmacokinetics** (absorption, distribution, metabolism, and excretion) and pharmacodynamics (the relationship between the drug level and drug effects), while pharmacoepidemiologists typically examine drug effects in larger populations under conditions more representative of clinical practice. Pharmacoepidemiology is an essential component of risk management of pharmaceutical products. Risk Management "encompasses processes for identifying and assessing the risks of specific health hazards, implementing activities to eliminate or minimize those risks, communicating risk information, and monitoring and evaluating the results of the interventions and communications" [56].

Current US federal regulations require evidence of both safety and effectiveness of drugs prior to approval for marketing (*see* **Drug Approval and Regulation**). However, such evidence is limited by the extent, duration, and patient characteristics of preapproval clinical trials. In addition, unexpected potentially beneficial effects are sometimes found after marketed use and questions may arise about the effectiveness of various drugs under conditions of use and in patient populations not included in premarketing clinical trials. A well-known international guideline on the extent of patient exposure to assess the clinical safety of drugs intended for chronic use in the treatment of non-life-threatening conditions summarized limitations of preapproval information on safety by noting, first, that it is expected that short-term adverse events with a cumulative 3-month incidence of about 1% or more should be well characterized prior to approval; secondly, that events where the rate of occurrence changes over a longer period of time may need to be characterized depending on their severity and importance to the risk–benefit assessment of the drug; and thirdly, that adverse events occurring in less than one in 1000 patients treated are not expected to be characterized prior to market approval [22]. Thus, it is often necessary to conduct pharmacoepidemiologic studies of risks and benefits of drugs and vaccines under conditions of marketed use (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**).

To design and interpret such studies it is essential to understand the clinical pharmacology of the drug and the pathophysiology and natural history of the diseases which the drug is used to treat or prevent. It is also essential to understand basic principles of epidemiologic study design (*see* **Pharmacoepidemiology, Study Designs**) and to identify and avoid potential sources of **bias**.

Some Common Sources of Bias in Pharmacoepidemiologic Studies

In the epidemiologic literature *bias* refers to an error which causes an estimate of a parameter to differ in a systematic way from the true value [26] in the source population, also known as the *study base*, whose person-time experience (*see* **Person-years at Risk**) the study is designed to sample [31, 57]. Numerous authors have provided methodologic approaches by which sources of bias in epidemiologic studies may be categorized [41, 43, 51] (*see* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**; **Bias, Overview**). We will briefly discuss some of the more common sources of bias in epidemiologic studies of drug effects.

Selection bias refers to errors arising because the estimated exposure effect among subjects included in the study differs from that which would have been obtained from including the entire study base [41]. For example, selection bias may occur when the

cases included in a study represent a nonrandomly selected subset of all of those arising from the study base [45]. Selection bias may also occur in **hospital-based case–control studies** if the drug exposure is related either positively or negatively to the diagnoses used to select **controls** or the drug exposure in the catchment population of the diagnoses used to identify cases differs from that in the catchment population of the diagnoses used to identify controls [57]. Differences in catchment populations are a particular problem with hospital-based studies because in many teaching hospitals patients may be drawn from hundreds of miles away for treatment of certain illnesses requiring particular skill in management and from only the immediate surrounding few miles for treatment of common disorders that are often used to select controls (*see* **Hospital Market Area**). Selection bias can produce serious distortions in estimates of disease natural history or treatment outcomes of patients drawn from referral centers [1, 28].

Confounding in epidemiologic studies occurs when exposure groups differ with respect to an extraneous factor related to the outcome. Estimates of exposure effects that fail to account appropriately for the imbalance are subject to bias. A full discussion of the assessment and control of confounding is beyond the scope of the present article and may be found in standard textbooks and in the current literature [41, 58, 60]. However, it is useful to mention *confounding by indication*, a particularly problematic form of confounding in studies of medical interventions when an indication for the intervention is itself a risk factor for the outcome under study [44, 59]. Studies in which confounding by indication has been an important consideration include mortality among asthma patients using long-acting inhaled beta agonists [7], myocardial infarction among hypertensive patients prescribed calcium channel blockers [38], and renal cell carcinoma in association with the use of diuretics [19]. A common way to avoid some obvious confounding by indication is to compare adverse outcomes of two drugs used for the same indication [7, 38]. However, even when the nominal indication is the same for two drugs, there may be subtle differences in patient characteristics and clinical judgments which lead to the choice of one drug over the other, are not documented in medical records, and yet which may be risk factors for the outcome.

A form of bias, which is closely related to but conceptually distinct from confounding by indication is *protopathic bias* [12, 44]. This occurs when early symptoms of a disease which is present but not yet recognized lead a patient to take a drug, which then appears to be the cause of the disease when it is eventually diagnosed. A classic example of this form of bias was seen in early studies of the antiulcer drug cimetidine, where a higher than expected incidence of gastric carcinoma was found among users than among nonusers. It is likely that many of the cancers were present but undiagnosed at the time the cimetidine was started. Subsequent studies with this class of drug have shown that elevations in gastric cancer risk diminish with duration of follow-up, returning to baseline with long-term use [23]. Not only protopathic bias but also confounding by indication was likely present in the association between cimetidine and gastric carcinoma. Peptic ulcer is both an indication for cimetidine and a risk factor for gastric carcinoma, with Helicobacter pylori being causally related to both peptic ulcer disease and gastric carcinoma.

Information bias arises from inaccuracies in the information collected on subjects in the study, resulting in **misclassification** of exposure, outcomes, or **covariates**. For example, patient recall of previous drug exposures has been shown to be subject to error, with the extent of inaccuracy differing by medication type, duration of therapy, recall interval and patient age [25, 61] (*see* **Recall Bias**). The misclassification of outcome is said to be *differential (nondifferential)* with respect to exposure if the misclassification probability differs (does not differ) depending on exposure. **Differential** and **nondifferential** misclassification of exposure with respect to outcome are defined similarly (*see* **Bias, Nondifferential**). In a simple cross-classification of exposure and outcome, nondifferential misclassification creates a **bias toward the null** [41]. However, even slight deviations from completely nondifferential misclassification can produce large biases away from the null [3].

When both exposure and outcome are misclassified and the misclassifications are **correlated**, the bias may be in either direction even when the misclassifications are nondifferential [4]. With more than two exposure levels, nondifferential misclassification will bias the most extreme category to the null but can bias intermediate levels of exposure in either direction [2]. Bias due to misclassification of **confounders** results in loss of ability to control confounding and

cannot be adequately dealt with by methods used to control confounding [13]. One practical conclusion from all of these findings is that it is *not* correct to conclude that risks of adverse drug effects estimated from inaccurate information are likely to represent underestimates simply because the misclassification may be presumed to be nondifferential.

Misclassification is a particular problem in epidemiologic studies using hospital discharge diagnosis codes to define outcomes and confounders, because the codes often do not correctly reflect discharge diagnoses recorded in the medical records [20]. This may occur through miscoding, use of nonspecific codes, omissions of codes in complicated patients with many different diagnoses, or failure to modify a code for an admission to "rule out" a condition when the condition was ruled out. For example, in a sample of about 1000 hospitalizations with the discharge diagnosis of acute myocardial infarction (AMI), medical record review found that 26% did not meet clinical criteria for AMI. Most were hospitalizations to rule out AMI in which the code remained even though AMI had been ruled out [20]. One approach which avoids some of problems with information bias is to use computer-based discharge diagnosis codes to identify potential cases and to confirm these by medical record review [40].

Adverse Drug Effects

Pharmacologic Classification

To help guide evaluation of adverse drug effects, clinical pharmacologists have classified them into two types, designated A and B, depending on their relationship to known pharmacological properties of the drug [39]. Type A ("augmented") effects are caused by exaggerated pharmacological actions of a drug. Such effects are also sometimes called "mechanism-based" adverse effects. They are somewhat predictable on the basis of the pharmacology of the drug and are typically dose-dependent. Examples include hypotension with anti-hypertensive drugs and gastrointestinal hemorrhage with nonsteroidal anti-inflammatory drugs [14]. Most type A effects are likely to have been at least identified before market approval. However, the predisposing factors, **dose–response** relationships, warning signs, spectrum of severity and long-term consequences may not have been adequately characterized at the time of initial marketing. Subsequent studies may reveal an increased risk in some patients with impaired metabolic clearance, concomitant use of drugs with competing metabolism, or increased target-organ sensitivity. Pharmacoepidemiologic studies of type A effects should be aimed not only at quantifying risks but also at finding ways to anticipate and reduce the risk through identification of predisposing factors and improved dosing guidelines [8].

Type B ("bizarre") effects are those that are not expected from the known pharmacologic properties of a drug given in usual doses to patients who metabolize the drug in a normal way [39]. Such effects include idiosyncratic, immunologic, allergic, pseudo-allergic, teratogenic, or carcinogenic reactions for which mechanisms are often unknown. Type B effects are typically rare, serious, unpredictable, not dose-dependent, and unlikely to have been adequately characterized or even recognized before market approval. The liver, blood, and skin are among the most common sites of type B reactions to drugs, while some vaccines have been associated with type B reactions of the nervous system [16, 21, 24, 36, 42, 55, 64]. Both drugs and biologics have been associated with rare allergic and pseudo-allergic type B reactions [9]. Pharmacologic studies of type B effects are typically constrained by the rarity of the events, but should attempt to identify patient subgroups at increased risk whenever this can be done.

Perhaps the most comprehensive example of using epidemiologic information to identify risks and benefits in different patient subgroups and providing this information to patients and physicians is given by the US prescribing information for oral contraceptives [34]. A more limited example is given by studies of agranulocytosis in association with the angiotensin-converting enzyme inhibitor captopril; it was found that the risk was extremely low except in well-defined subgroups in whom use of the drug could generally be avoided [5].

Timing of Adverse Effects in Relation to Duration of Therapy

One of the most important aspects to consider in both the clinical and epidemiologic evaluation of adverse drug effects is the timing in relation to duration of therapy [15, 17, 52, 54, 62]. Some effects, such as angioedema with angiotensin-converting enzyme inhibitors, are more common early in

therapy [50]. Others, such as tardive dyskinesia with phenothiazines, are typically seen only after prolonged exposure to the drug. For some effects there may be a time window of highest risk. For example, onset of Guillain–Barré syndrome following the so-called "swine flu" vaccine was highest 17 days after vaccination and declined thereafter [46]. Serum-sickness-like reactions to drugs typically occur from 7 to 21 days after starting therapy [9]. Depletion of susceptibles may also affect the hazard function for adverse events in relation to duration of therapy [62]. Accounting for timing of adverse effects in relation to duration of therapy requires collecting information on timing of the event not only in relation to last exposure to the drug but also in relation to duration of therapy. Failure to account properly for timing may result in over- or underestimation of risks and can create artifactual treatment-by-subgroup **interactions** when comparing patient groups with different temporal patterns of usage [15].

Beneficial Effects

Intended Beneficial Effects: Efficacy, Effectiveness and Outcomes Research

Efficacy refers to the benefits of an intervention as measured under ideal circumstances in a randomized, controlled clinical trial conducted in a homogeneous set of patients with careful attention to the protocol. Clinical trials conducted to provide demonstrations of drug efficacy needed for drug approval are typically conducted under conditions which maximize internal validity of the trial itself at the possible expense of external validity – generalizability to usual clinical practice [48] (*see* **Validity and Generalizability in Epidemiologic Studies**). *Effectiveness* refers to the benefits of the intervention as measured under conditions intended to resemble closely the settings and patient populations where the intervention will be used in clinical practice. Effectiveness depends not only on efficacy but also on ease of administration, acceptability to patients and prescribers, compliance, and impact on use of health care resources. Consequently, effectiveness of an intervention depends not only on the intervention, but also on the setting in which it is delivered.

Recently there has been an increased emphasis on judging the results of health interventions in terms of their ability to improve *health outcomes*, i.e. changes in health status noticeable by patients, rather than exclusively in terms of their ability to improve laboratory tests or physiological parameters [10, 11, 48]. The field of **outcomes research** seeks to evaluate the overall effects of different interventions on health outcomes in clinical practice [10, 11, 27].

Because of the difficulty and expense of conducting randomized clinical trials in a clinical practice setting, observational studies are often used for the comparison of treatment outcomes [11, 27]. Observational studies of drug effectiveness are subject to selection bias, which may be impossible to control because the factors which lead to the choice of one therapy over another may not be fully reflected in any data source [6, 12, 28–30, 32, 59]. Selection bias in the study of intended effects of drugs may be more difficult to overcome than in the study of unintended effects [30]. Confounding by the indication for therapy must be considered in all pharmacoepidemiologic studies, and is particularly difficult to control in observational studies of intended drug effects [30, 59, 63]. Because most differences in effectiveness between active agents are likely to be moderate, observational studies are especially prone to distortion caused by bias and confounding [30, 32, 63]. This cautionary note also applies to observational studies of vaccine effectiveness, though to a lesser extent, because the effect sizes are typically much larger than for effectiveness studies of drugs [33, 47] (*see* **Vaccine Studies**). As an alternative to observational studies of drug effectiveness in clinical practice, randomized effectiveness trials have been conducted [35, 49]. Such studies have the potential for producing more valid estimates of effectiveness than can be obtained from observational studies.

Unintended Beneficial Effects

Some current indications for drug treatment began with the serendipitous finding of an unexpected **association** between drug exposure and a beneficial effect. The initial hypothesis of a beneficial drug effect usually arises from **case series** or laboratory observations, followed by formal epidemiologic studies. As useful as such studies have been in providing quantitative estimates of benefit, randomized clinical trials are essential for **hypothesis testing**. For example, data from randomized clinical trials of beta-carotene

in the prevention of lung cancer have not confirmed findings of earlier observational studies which had suggested a protective effect [18].

References

[1] Ballard, D.J., Bryant, S.C., O'Brien, P.C., Smith, D.W., Pine, M. & Cortese, D.A. (1994). Referral selection bias in the Medicare hospital mortality prediction model: are centers of referral for Medicare beneficiaries necessarily centers of excellence?, *Health Services Research* **8**, 771–784.

[2] Birkett, N.J. (1992). Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure, *American Journal of Epidemiology* **136**, 356–362.

[3] Brenner, H. (1993). Inferences on the potential effects of presumed nondifferential exposure misclassification, *Annals of Epidemiology* **3**, 289–294.

[4] Brenner, H., Savitz, D.A. & Gefeller, O. (1993). The effects of joint misclassification of exposure and disease on epidemiologic measures of association, *Journal of Clinical Epidemiology* **46**, 1195–1202.

[5] Bristol-Myers Squibb Company (1997). Physicians' prescribing information for Capoten™. Princeton.

[6] The Coronary Drug Project Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project, *New England Journal of Medicine* **303**, 1038–1041.

[7] Crane, J., Pearce, N., Burgess, C. & Beasley, R. (1995). Asthma and the beta agonist debate, *Thorax* **50**, Supplement 1, S5–S10.

[8] De Groen, P.C., Lubbe, D.F., Hirsch, L.J., Daifotis, A., Stephenson, W., Freedholm, D., Pryor-Tillotson, S., Seleznick, M.J., Pinkas, H. & Wang, K.K. (1996). Esophagitis associated with the use of alendronate, *New England Journal of Medicine* **335**, 1–21.

[9] deShazo, R.D. & Kemp, S.F. (1997). Allergic reactions to drugs and biologic agents, *JAMA* **278**, 1895–1906.

[10] Ellwood, P.M. (1988). Shattuck Lecture – Outcomes management: a technology of patient experience, *New England Journal of Medicine* **318**, 1549–1556.

[11] Epstein, R.S. & Sherwood, L.M. (1996). From outcomes research to disease management: a guide for the perplexed, *Annals of Internal Medicine* **124**, 832–837.

[12] Feinstein, A.R. (1985). *Clinical Epidemiology: The Architecture of Clinical Research*. W.B. Saunders, Philadelphia, p 301.

[13] Greenland, S. & Robins, J.M. (1985). Confounding and misclassification, *American Journal of Epidemiology* **122**, 495–506.

[14] Griffin, M.R., Ray, W.A. & Schaffner, W. (1988). Nonsteroidal anti-inflammatory drug use and death from peptic ulcer in elderly persons. *Annals of Internal Medicine* **109**, 359–363.

[15] Guess, H.A. (1989). Behaviour of the exposure odds ratio in a case-control study when the hazard function is not constant over time. *Journal of Clinical Epidemiology* **42**, 1179–1184.

[16] Guess, H.A. (1993). How should acute hepatic drug effects be studied epidemiologically?, *Epidemiology* **4**, 487–489.

[17] Hemmelgarn, B., Suissa, S., Huang, A., Boivin, J.F. & Pinard, G. (1997). Benzodiazepine use and the risk of motor vehicle crash in the elderly, *JAMA* **278**, 27–31.

[18] Hennekens, C.H., Buring, J.E., Manson, J.E., Stampfer, M., Rosner, B., Cook, N.R., Belanger, C., LaMotte, F., Gaziano, J.M., Ridker, P.M., Willett, W. & Peto, R. (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease, *New England Journal of Medicine* **334**, 1145–1149.

[19] Hiatt, R.A. Tolan, K. & Quesenberry, C.P. Jr (1994). Renal cell carcinoma and thiazide use: a historical, case-control study, *Cancer Causes and Control* **5**, 319–325.

[20] Iezzoni, L.I., Burnside, S., Sickles, L., Moskowitz, M.A., Sawitz, E. & Levine, P. (1988). Coding of acute myocardial infarction. Clinical and policy implications, *Annals of Internal Medicine* **109**, 745–751.

[21] Institute of Medicine (1994). *Adverse Events Associated With Childhood Vaccines: Evidence Bearing on Causality*. National Academy Press, Washington.

[22] International Conference on Harmonization (1995). *The Extent of Population Exposure to Assess Clinical Safety: For Drugs Intended for Long-Term Treatment of Non-Life-Threatening Conditions*, ICH-E1A, Federal Register 60 FR 11270. US Government Printing Office, Washington.

[23] Johnson, A.G., Jick, S.S., Perera, D.R. & Jick, H. (1996). Histamine-2 receptor antagonists and gastric cancer, *Epidemiology* **7**, 434–436.

[24] Kaufman, D.W., Kelly, J.P., Levy, M. & Shapiro, S. (1991). *The Drug Etiology of Agranulocytosis and Aplastic Anemia*. Oxford University Press, New York.

[25] Kelly, J.P., Rosenberg, L., Kaufman, D.W. & Shapiro, S. (1990). Reliability of personal interview data in a hospital based case-control study, *American Journal of Epidemiology* **131**, 79–90.

[26] Last, J.M. (1988). *A Dictionary of Epidemiology*, 2nd Ed. Oxford University Press, New York, pp. 13, 98.

[27] Maklan, C.W., Greene, R. & Cummings, M.A. (1994). Methodological challenges and innovations in patient outcomes research, *Medical Care* **32**, Supplement 7, JS13–JS21.

[28] Melton III, L.J. (1985). Selection bias in the referral of patients and the natural history of surgical conditions, *Mayo Clinic Proceedings* **60**, 880–889.

[29] Michels, K.B. & Braunwald, E. (2002). Estimating treatment effects from observational data, *JAMA* **287**, 3130–3132.

[30] Miettinen, O.S. (1983). The need for randomization in the study of intended effects, *Statistics in Medicine* **2**, 267–271.

[31] Miettinen, O.S. (1985). *Theoretical Epidemiology – Principles of Occurrence Research in Medicine*. Wiley, New York, Chapter 3, pp. 46–68.

[32] Moses, L.E. (1985). Statistics in practice: statistical concepts fundamental to investigations, *New England Journal of Medicine* **312**, 890–897.

[33] Orenstein, W.A., Bernier, R.H. & Hinman, A.R. (1988). Assessing vaccine efficacy in the field. Further observations, *Epidemiologic Reviews* **10**, 212–241.

[34] Ortho Pharmaceutical Corporation (1997). Physicians' prescribing information for Ortho-Cept™. Raritan.

[35] Oster, G., Borok, G.M., Menzin, J., Heyse, J.F., Epstein, R.S., Quinn, V., Benson, V.V., Dudl, R.J. & Epstein, A. (1995). A randomized trial to assess effectiveness and cost in clinical practice: rationale and design of the Cholesterol Reduction Intervention Study (CRIS), *Controlled Clinical Trials* **16**, 3–16.

[36] Park, B.K., Pirmohamed, M. & Kitteringham, N.R. (1992). Idiosyncratic drug reactions: a mechanistic evaluation of risk factors, *British Journal of Pharmacology* **34**, 377–395.

[37] Porta, M.S. & Hartzema, A.G. (1991). The contribution of epidemiology to the study of drugs, in *Pharmacoepidemiology: An Introduction*, 2nd Ed., A.G. Hartzema, M.S. Porta & H.H. Tilson, eds. Harvey Whitney, Cincinnati. pp. 2–17.

[38] Psaty, B.M., Heckbert, S.R., Koepsell, T.D., Siscovick, D.S., Raghunathan, T.E., Weiss, N.S., Rosendaal, F.R., Lemaitre, R.N., Smith, N.L., Wahl, P.W., Wagner, E.H. & Furberg, C.D. (1995). The risk of myocardial infarction associated with antihypertensive drug therapies, *Journal of the American Medical Association* **274**, 620–625.

[39] Rawlins, M.D. & Thompson, J.W. (1991). Mechanisms of adverse drug reactions, in *Textbook of Adverse Drug Reactions*, 4th Ed., D.M. Davies, ed. Oxford University Press, Oxford, pp. 18–45.

[40] Ray, W.A. & Griffin, M.R. (1989). Use of Medicaid data for pharmacoepidemiology, *American Journal of Epidemiology* **129**, 837–849.

[41] Rothman, K.J. & Greenland, S. (1998). Precision and validity in epidemiology studies, in: *Modern Epidemiology*, 2nd Edition, Chapter 8, K.J. Rothman & S. Greenland, eds, Lippincott, Williams & Wilkins, Philadelphia, pp. 115–134.

[42] Roujeau, J.C. & Stern, R.S. (1994). Medical progress: severe adverse cutaneous reactions to drugs, *New England Journal of Medicine* **331**, 1272–1285.

[43] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[44] Salas, M., Hofman, A. & Stricker, B.H. (1999). Confounding by indication: an example of variation in the use of epidemiologic terminology, *American Journal of Epidemiology* **149**(11), 981–983.

[45] Savitz, D.A. & Pearce, N. (1988). Control selection with incomplete case ascertainment, *American Journal of Epidemiology* **127**, 1109–1117.

[46] Schonberger, L.B., Bregman, D.J., Sullivan-Bolyai, J.Z., Keenlyside, R.A., Ziegler, D.W., Retailliau, H.F., Eddins, D.L. & Bryan, J.A. (1979). Guillain-Barré syndrome following vaccination in the National Influenza Immunization Program, United States, 1976–1977, *American Journal of Epidemiology* **110**, 105–123.

[47] Shapiro, E.D., Berg, A.T., Austrian, R., Schroeder, D., Parcells, V., Margolis, A., Adair, R.K. & Clemens, J.D. (1991). The protective efficacy of polyvalent pneumococcal polysaccharide vaccine, *New England Journal of Medicine* **325**, 1453–1460.

[48] Simon, G., Wagner, E. & VonKorff, M. (1995). Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs, *Journal of Clinical Epidemiology* **48**, 363–373.

[49] Simon, G.E., VonKorff, M., Heiigenstein, J.H., Revicki, D.A., Grothaus, L., Katon, W. & Wagner, E.H. (1996). Antidepressant choice in primary care-effectiveness and cost of fluoxtine vs tricyclic antidepressants, *Journal of the American Medical Association* **275**, 1897–1902.

[50] Slater, E.E., Merrill, D.D., Guess, H.A., Roylance, P.J., Cooper, W.D., Inman, W.H. & Ewan, P.W. (1988). Clinical profile of angioedema associated with angiotensin converting-enzyme inhibition, *Journal of the American Medical Association* **260**, 967–970.

[51] Steineck, G. & Ahlbom, A. (1992). A definition of bias founded on the concept of the study base, *Epidemiology* **3**, 477–482.

[52] Stephens, M.D.B. (1984). Assessment of causality in an industrial setting, *Drug Information Journal* **18**, 307–313.

[53] Strom, B.L. (1994). What is pharmacoepidemiology?, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, New York, Chapter 1, pp. 3–13.

[54] Suissa, S., Blais, L., Spitzer, W.O., Cusson, J., Lewis, M. & Heinemann, L. (1997). First-time use of newer oral contraceptives and the risk of venous thromboembolism. Contraception, **56**, 141–146.

[55] US Food and Drug Administration, Drug Induced Liver Toxicity, Available at: `http://www.fda.gov/cder/livertox/default.htm`. Accessed June 26, 2002.

[56] US Food and Drug Administration Task Force on Risk Management: Managing the risks from medical product use – creating a risk management framework, May 1999. Available at: `http://www.fda.gov/oc/tfrm/riskmanagement.pdf`. Accessed June 27, 2002, Part 4, p. 73.

[57] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case control studies. I. Principles, *American Journal of Epidemiology* **135**, 1019–1028.

[58] Walker, A.M. (1991). *Confounding, in Observation and Inference – An Introduction to the Methods of Epidemiology*. Epidemiology Resources, Inc., Newton Lower Falls, pp. 119–128.

[59] Walker, A.M. (1996). Confounding by indication, *Epidemiology* **7**, 335–336.

[60] Weinberg, C.R. (1993). Toward a clearer definition of confounding, *American Journal of Epidemiology* **137**, 1–8.

[61] West, S.L., Savitz, D.A., Koch, G., Strom, B.L., Guess, H.A. & Hartzema, A. (1995). Recall accuracy for prescription medications: self-report compared with database information, *American Journal of Epidemiology* **142**, 1103–1112.

[62] Yola, M. & Lucien, A. (1994). Evidence of the depletion of susceptibles effect in nonexperimental pharmacoepidemiologic research, *Journal of Clinical Epidemiology* **47**, 731–737.

[63] Yusuf, S., Collins, R. & Peto, R. (1984). Why do we need some large, simple randomized trials?, *Statistics in Medicine* **3**, 409–420.

[64] Zimmerman, H.J. (1978). *Hepatotoxicity: The Adverse Effects of Drugs and Other Chemicals on the Liver*. Appleton-Century Crofts, New York.

(*See also* **Drug Utilization Patterns**)

H.A. GUESS

Pharmacoepidemiology, Overview

Pharmacoepidemiology, the study of patterns of medication use in the population and their effects on disease, is a new field. The need for this area of research became evident in 1961, during the thalidomide catastrophe, when it was realized that drugs prescribed for therapeutic purposes could produce unexpected risks. The entry of thalidomide, a new hypnotic drug, on the market was accompanied by a sudden sharp increase in the frequency of rare birth defects, characterized by the partial or complete absence of limbs [29, 31]. Consequently, several countries either instituted agencies to regulate drugs or expanded the mandate of existing agencies [51]. These agencies were previously interested only in the demonstration of a drug's efficacy, but now required proof of a drug's safety before it was tested in humans let alone before it was marketed for use by the general population (*see* **Drug Approval and Regulation**). These proofs of safety, based on toxicologic and pharmacologic studies, were necessary before randomized controlled trials (RCTs) (*see* **Clinical Trials, Overview**) could be conducted on human subjects, primarily to demonstrate the efficacy of a drug.

The use of the epidemiologic approach to characterize population patterns of medication use and to assess their effects developed as a complement to RCTs for several reasons. First, RCTs were designed to assess the efficacy and effectiveness of a drug, providing as well some data on its safety with respect to commonly arising side-effects. However, rare side-effects typically cannot be identified in clinical trials because of their small size. For example, to detect a **relative risk** of 2, for a side-effect having an **incidence** of 1 per 100, we would require a two-arm trial with over 3000 subjects per arm ($\alpha = 0.05$, $\beta = 0.1$). If the incidence of the side-effect is 1 per 10 000, the sample size per arm would need to be over 300 000. Clearly, these sample sizes are rarely if ever used in RCTs, yet the number of people who will be using these drugs will be in the millions.

Secondly, RCTs usually restrict the study subjects to people without coexisting disease, who are therefore not taking other medications that could interact with the study medication. They are also restricted with respect to age, rarely including children and elderly subjects. Yet, the elderly will be major consumers of most of these medications, along with other medications they are using for coexisting diseases (*see* **Co-morbidity**; **Drug Interactions**).

Thirdly, RCTs will usually be based on a short follow-up that typically assesses medication use for a period of 3–12 months. Yet, again, subjects may be using these medications for years, so that the effect of the prolonged use of these medications remains clearly unknown from the RCT data. Finally, there are situations where the RCT is either unethical or inapplicable. For example, it would be ethically unacceptable today in North America or Europe to assess the long-term effects of a new anti-hypertensive agent against placebo in an RCT, although this has been done in China [12] (*see* **Ethics of Randomized Trials**). Yet, a large number of hypertensive patients from the general population are either untreated or do not comply with their treatment (*see* **Compliance Assessment in Clinical Trials**). They could be used as the reference group for a **nonrandomized** study based on a **cohort study** design.

Although pharmacoepidemiology can be simply regarded as an application of epidemiologic principles and methods to the field of medications, it is now developing as a discipline of its own because of the special nature of drugs. Indeed, the ways by which drugs are prescribed, employed, marketed and regulated impose certain constraints on epidemiologic research into their use and effects. This field poses challenges that often require special solutions not found in other domains of application, such as cancer, cardiovascular, occupational or infectious disease epidemiology; medications are marketed rapidly, practice patterns of prescribing by physicians are variable and profiles of drug use by patients are complicated by varying compliance patterns. This complex and dynamic context in which pharmacoepidemiology is situated, as well as the available sources of data, have given rise to unique statistical challenges. The fact that the lifetime of a drug on the market is relatively short and can suddenly be shortened still further by a regulatory or corporate withdrawal, often imposes major constraints on studies of its effects. These studies must be conducted rapidly and use existing data in an efficient way without compromising validity (*see* **Validity and Generalizability in Epidemiologic Studies**).

In this article we describe several areas where biostatistical input has served to advance pharmacoepidemiology. Although the last two decades have witnessed an explosion of methodologic advances put forward by biostatistics in the design and analysis of epidemiologic studies, most of these have been fundamental to the field of epidemiology in general. We do not discuss these areas here since they are dealt with extensively elsewhere (*see* **Analytic Epidemiology**; **Descriptive Epidemiology**). Instead, we focus on the biostatistical aspects that produced unique methodologic advances, specific to problems posed by pharmacoepidemiologic research.

The Case–Crossover Design

When conducting a **case–control study**, the selection of **controls** is usually the most challenging task. The fundamental principle is that selected controls should be representative of the source population which gave rise to the cases [33] a principle often difficult to implement in practice, especially when dealing with acute adverse events and transient exposures.

For example, we may wish to study the **risk** of ventricular tachycardia in association with the use of inhaled beta agonists in asthma. This possible effect has been hypothesized on the basis of clinical study observations of hypokalemia and prolonged Q-T intervals as measured on the electrocardiogram in patients after beta agonist exposure [1]. These unusual cardiac deviations were observed only in the 4-hour period following drug absorption. Thus, a case–control study of this issue would first select cases with this adverse event and investigate whether the drug was taken during the 4-hour span preceding the event. For **controls**, on the other hand, the investigator must define a time point of reference for which to ask the question about use of this drug in the "past 4 hours". However, if, for example, the drug is more likely to be required during the day, but controls can only be reached at home in the evening, the **relative risk** estimate will be **biased** by the differential timing of responses for cases and controls.

Consequently, when dealing with the study of transient drug effects on the risk of acute adverse events, Maclure [30] proposed the *case–crossover design*, which uses the cases as their own controls. The case–crossover design is simply a **crossover** study in the cases only. The subjects alternate at varying frequencies between exposure and nonexposure to the drug of interest until the adverse event occurs, which does for all subjects in the study since all are cases by definition. Each case is investigated to determine whether exposure occurred within the predetermined effect period, namely within the 4 hours previous to the adverse event in our example. This occurrence is then classified as having arisen either under drug exposure or nonexposure on the basis of the effect period. Thus, each case is either exposed or unexposed. For the reference information, data on the average drug use pattern are necessary to determine the probability of exposure in the time window of effect. This is done by obtaining data for a sufficiently long period of time to derive a stable estimate. In our example, we might determine the average number of times a day each case has been using beta agonists (two inhalations of 100 μg each) in the past year. This will allow us to estimate the proportion of time that each asthmatic is usually spending as "exposed" in the 4-hour effect period. This proportion is then used to obtain the number of cases **expected** on the basis of time spent in these "at-risk" periods, for comparison with the number of cases observed during such periods. This is done by forming a **two-by-two table** for each case, with the corresponding control data as defined above, and combining the tables using the **Mantel–Haenszel** technique as described in detail by Maclure [30]. The resulting **odds ratio** is then given by $OR = \sum a_i N_{0i} / \sum (1 - a_i) N_{1i}$, where a_i is 1 if case i is exposed, 0 if not, N_{0i} is the expected number of unexposed periods and N_{1i} the expected number of exposed periods during the reference time span.

Table 1 displays hypothetical data from a case–crossover study of 10 asthmatics who experienced ventricular tachycardia. These were all queried regarding their use of two puffs of inhaled beta agonist in the last 4 hours and on average over the past year. The fact of drug use within the effect period is defined by a_i with three cases having used beta agonists in the 4-hour period prior to the adverse event. The usual frequency of drug use per year is converted to a ratio of the number of exposed periods to the number of unexposed periods, the total number of 4-hour periods being 2190 in one year. Using the Mantel–Haenszel formula to combine the 10 two-by-two tables, the estimate of the odds ratio is 3.0, and the 95% **confidence interval** (1.2, 7.6).

Table 1 Hypothetical data for a case–crossover study of beta agonist exposure in last 4 hours and the risk of ventricular tachycardia in asthma

Case no.	Beta agonist use[a] in last 4 hours (ai)	Usual beta agonist use in last year	Periods of exposure (N_{1i})	Periods of non-exposure (N_{0i})
1	0	1/day	365	1825
2	1	6/year	3	2184
3	0	2/day	730	1460
4	1	1/month	12	2178
5	0	4/week	208	1982
6	0	1/week	52	2138
7	0	1/month	12	2178
8	1	2/month	24	2166
9	0	2/day	730	1460
10	0	2/week	104	2086

[a]Inhalations of 200 μg: 1 = yes, 0 = no

The case–crossover design depends on several assumptions to produce **unbiased** estimates of the odds ratio. Greenland [14] presented examples where the odds ratio estimates from this approach can be biased. For example, the probability of exposure cannot vary over the time. Similarly, **confounding** factors must be constant over time. Finally, there cannot be **interaction** between unmeasured subject characteristics and the exposure. Nevertheless, this approach is being used successfully in several studies [38, 43]. It has also been adapted for application to the **risk assessment** of vaccines [11].

Confounding Bias

Because of the lack of **randomization**, the most important limitation of **observational studies** in pharmacoepidemiology is whether an important confounding factor is biasing the reported relative risk estimate. A factor is considered a confounder if it is associated, at each level of drug exposure, with the adverse event, and with exposure to the drug itself. Two approaches exist to address the problem of confounding variables, which we describe in the context of the case–control design, although they apply equally well to a cohort design. The first is to select controls that are matched to the cases with respect to all confounding factors and to use the appropriate corresponding techniques of analysis for matched data, usually **conditional logistic regression** (*see* **Matched Analysis**; **Matching**). This approach, although often appropriate, has been shown to be susceptible to bias from residual confounding due to coarse matching [5]. The second solution is to select controls unmatched with respect to these confounding factors, but to measure these confounders for all study subjects and use statistical techniques based on either **stratification** or **multiple regression**, permitting removal of their effect on the risk from the effect of the drug under study. This approach can also lead to residual confounding if the confounder data are not analyzed properly. For example, the risk of venous thromboembolism has been found to be higher among users of newer oral contraceptive drugs than users of older formulations [22, 41, 52], after controlling for the effect of age. A recent study, however, showed that when confounding by the woman's age is analyzed using finer age bands, the relative risk is substantially reduced [10].

Beyond such difficulties generic to epidemiology, the context of pharmacoepidemiology has produced several situations where confounding requires particular statistical treatment. Some are described in this section.

Missing Confounder Data

It is at times impossible to obtain data on certain important confounding variables. A frequent situation encountered in pharmacoepidemiology is that of complete data for the cases and incomplete data for the controls of a case–control study. This is often encountered in "computerized database" studies based on **administrative databases** where cases have likely been hospitalized and thus have an extensive medical dossier. For these cases, the investigator will thus have access to ample information on potential confounding variables. However, if the controls are population-based (*see* **Case–Control Study, Population-based**), it is unlikely they were hospitalized and will not provide comparable data on confounders in the absence of medical charts. Consequently, confounder data will only be available in the cases, and not in the controls.

We can assess whether a factor is a confounder on the basis of data available solely for the cases, so that if the factor is deemed not to be a confounder, then the final analysis of the risk of the drug under study will not need to be adjusted. The approach is described by Ray & Griffin [37] and was

used in the context of a study of nonsteroidal anti-inflammatory drug (NSAID) use and the risk of fatal peptic ulcer disease [15]. The strategy is based on the definition [5] of a confounder C (C+ and C− denote presence and absence) in the assessment of the **association** between a drug exposure E (E+ and E− denote exposure or not to the drug) and an adverse condition D (D+ and D− denote cases and controls, respectively). Confounding is present if both following conditions are satisfied:

1. C and E are associated in the control group (in D−).
2. C and D are associated in E+ and in E−.

Assuming, in the absence of **effect modification**, a common odds ratio between E and D (OR_{ED}) in C+ and in C−, condition 1 becomes equivalent to: C and E are associated in the case group (D+). Thus, if in the cases we find no association between the potential confounder and drug exposure, confounding by this factor can be excluded outright, without having to verify condition 2. In this instance, the analysis involving drug exposure in cases and controls can be performed directly without any concern for the confounding variable. This strategy for assessing confounding is extremely valuable for several case–control studies in pharmacoepidemiology, since if confounding is excluded by this technique, crude methods of analysis can be used to obtain a valid estimate of the odds ratio. However, this is not often the case.

As an example, we use data from a case–control study conducted using the Saskatchewan computerized databases to assess whether theophylline, a drug used to treat asthma, increases the risk of acute cardiac death [46]. In this study, the 30 cases provided data on theophylline use, as well as on smoking, possibly an important confounder. However, the 4080 controls only had data available on theophylline use and not on smoking. Table 2 displays the data from this study. The crude odds ratio between theophylline use and cardiac death is 4.3((17/13)/(956/3124)). Because of the **missing data** on smoking, it is only possible to partition the cases, but not the controls, according to smoking. The odds ratio between theophylline use and smoking among the cases is estimable and found to be 7.5((14/5)/(3/8)), thus indicating that smoking is indeed a strong confounder.

Table 2 Data from a case–control study of theophylline use and cardiac death in asthma, with the smoking confounder data missing for controls

	Cases		Controls	
	E	$\overline{E}$	E	$\overline{E}$
Notation:				
Combined	a	c	b	d
Stratified by smoking:				
Smokers	a_0	c_0	[a]	[a]
Nonsmokers	a_1	c_1	[a]	[a]
Data:				
Combined	17	13	956	3124
Stratified by smoking:				
Smokers	14	5	[a]	[a]
Nonsmokers	3	8	[a]	[a]

[a]These frequencies are missing for controls.

An approach was recently developed to permit the estimation of the *adjusted* odds ratio of the theophylline by cardiac death association, in the absence of confounder data among the controls [45]. The adjusted odds ratio is given by

$$OR_{\text{adj}} = \frac{P_0(w-y)}{(1-P_0)y}, \tag{1}$$

where $y = \{v - [v^2 - 4(r-1)rwx]^{1/2}\}/[2(r-1)]$, $v = 1 + (r-1)(w+x)$ when $r \neq 1$ (and $y = wx$ when $r = 1$), r is the odds ratio between exposure and confounder among the cases, x is the probability of exposure among the controls, P_0 is estimated by $a_0(a_0 + c_0)$, and w is the **prevalence** of the confounder among the controls [45]; w is the only unknown and must be estimated from external sources. An estimate of the variance of OR_{adj} in (1) exists in closed form. For the illustrative data, an external estimate of smoking prevalence among asthmatics, obtained from a Canadian general population health survey, is 24%. Using this estimate, the adjusted odds ratio is 2.4, much lower than the crude estimate of 4.3, with 95% confidence interval (1.0, 5.8).

This statistical approach was developed specifically to address the frequent problem of missing confounder data in pharmacoepidemiology. When using computerized databases, these data are more often missing only in the control series of a case–control study. This technique, based on statistical reasoning, allows us to derive adjusted estimates of relative risk with few assumptions. Extensions of this type

of approach to a **regression** context will expand its usefulness.

The Case–Time–Control Design

Case–control studies in pharmacoepidemiology that assess the intended effects of drugs are often limited by their inability to obtain a precise measure of the indication of drug exposure. Adjustment for this crucial confounding factor becomes impossible and an unbiased estimate of the drug effect is unattainable [49]. This bias, arising from confounding by indication, is a major source of limitation in pharmacoepidemiology [32]. Here again, a within-subject approach, similar to the case–crossover design, has been developed. By using cases and controls of a conventional case–control study as their own referents, the *case–time–control design* eliminates the biasing effect of unmeasured confounding factors such as drug indication [44]. This approach is applicable only in situations where exposure varies over time, which is typically often the case for medications.

The correct application of the case–time–control design is based on a specific model for the data, a model that entails inherent assumptions and imposes certain conditions for the approach to be valid. The model, based on a case–control sampling design, is presented for a dichotomous exposure that varies over time and that is measured only for two consecutive time periods, the current period and the reference period. The logit of exposure $L_{ijkl} = \text{logit}\{\Pr(E_{ijkl} = 1)\}$, is given by

$$L_{ijkl} = \mu + S_{il} + \pi_j + \Theta_k \qquad (2)$$

where E_{ijkl} represents the binary exposure for group i, period j, outcome k and subject l within group i, μ represents the overall exposure logit, S_{il} is the effect of study subject l in group i, π_j is the effect of period j and Θ_k is the effect of event occurrence k. More specifically, $i = 0, 1$ denotes the case–control group (1 = case subjects, 0 = control subjects), $j = 0, 1$ denotes the period (1 = current period, 0 = reference period), $k = 0, 1$ denotes the event occurrence (1 = event, 0 = no event) and $l = 1, \ldots, n_i$ designates the study subject within group i, with n_1 case subjects and n_0 control subjects. The confounding effect of unmeasured severity or indication is inherently accounted for by S_{il}.

The period effect, measured by the log of the odds ratio, is given by $\delta_\pi = \pi_1 - \pi_0$ and estimated from the control subjects. The net effect of exposure on event occurrence is given by $\delta_\Theta = \Theta_1 - \Theta_0$. The case subjects permit one to estimate the sum $\delta_\Theta + \delta_\pi$ so that the effect of exposure on event occurrence δ_Θ, is estimable by subtraction. The estimation of the odds ratio is based on any appropriate technique for matched data, such as conditional logistic regression.

Three basic assumptions are inherently made by this logit model. The first is the absence of effect modification of the exposure–outcome association by the unmeasured confounder, i.e. the exclusion of the $S_{il}\Theta_k$ interaction term in model (2). The second is the absence of effect modification of the exposure–outcome association by period, i.e. a null value for the $\pi\Theta_k$ interaction term. The third is the lack of effect modification of the exposure–period association by the confounder, represented by the absence of an $S_{il}\pi_j$ interaction term in model (2). Greenland [14] presented examples of the bias that can occur with this approach when the model contains the latter interaction.

The approach is illustrated with data from the Saskatchewan Asthma Epidemiologic Project [40], a study conducted to investigate the risks associated with the use of inhaled beta agonists in the treatment of asthma. Using databases from Saskatchewan, Canada, a cohort of 12 301 asthmatics was followed during 1980–87. All 129 cases of fatal or near-fatal asthma and 655 controls were selected. The amount of beta agonist used in the year prior to the index date was found to be associated with the adverse event. In comparing low (12 or less canisters per year) with high (more than 12) use of beta agonists. The crude odds ratio for high beta-agonist use is 4.4, with 95% confidence interval (2.9, 6.7). Adjustment for all available markers of severity, such as oral corticosteroids and prior asthma hospitalizations as confounding factors, lowers the odds ratio to 3.1, with 95% confidence interval (1.8, 5.4), the "best" estimate one can derive from these case–control data using conventional tools.

The use of inhaled beta agonists, however, is known to increase with asthma severity which also increases the risk of fatal or near-fatal asthma. It is therefore not possible to separate the risk effect of the drug from that of disease severity. To apply

the case–time–control design, exposure to beta agonists was obtained for the 1-year current period and the 1-year reference period. Among the 129 cases, 29 were currently high users of beta agonists and were low users in the reference period, while 9 cases were currently low users of beta agonists and were high users previously. Among the 655 controls, 65 were currently high users of beta agonists and were low users in the reference period, while 25 were currently low users of beta agonists and were high users previously. The case–time–control odds ratio, using these discordant pairs frequencies for a matched pairs analysis, is given by $(29/9)/(65/25) = 1.2$, with 95% confidence interval (0.5, 3.0). This estimate, which excludes the effect of unmeasured confounding by disease severity, indicates a minimal risk for these drugs.

The case–time–control approach provides an unbiased estimate of the odds ratio in the presence of confounding by indication, a common obstacle in pharmacoepidemiology. This is possible despite the fact that the indication for drug use (in our example, disease severity) is not measured, because of the within-subject analysis. Nevertheless, as mentioned above, its validity is subject to several strict assumptions. This approach must therefore be used with caution.

Risk Functions Over Time

Most epidemiologic studies assessing a risk over time routinely assume that the hazards are constant or proportional. Rate ratios are then estimated by **Poisson regression** models or Cox's **proportional hazards** model [6]. Often, deviations from these simplifying assumptions are addressed at the design stage, by restricting the study to a specific follow-up period where the assumptions are satisfied. For instance, to study the risk of cancer associated with an agent considered to be an initiator of the disease, the first few years of follow-up after the initiation of exposure will not be accounted for in the analysis, to allow for a reasonable **latency period**. On the other hand, if the agent is suspected to be a cancer promoter, these same first years will be used in the analysis. Since such considerations are mostly dealt with at the design stage, little attention has been paid to the analytic considerations of this issue.

In pharmacoepidemiology, the risk of an adverse event often varies strongly with the duration of use of a drug. Figure 1 shows three different risk profiles, typical of drug exposure. Figure 1(a) displays the usual profile of risk associated with an acute effect. The drug will affect susceptible subjects early, reflected by the early sharp rise in the curve, and once these subjects are eliminated from the cohort, the remaining subjects will return to some lower constant baseline risk. The peak can occur almost immediately, as with allergic reactions to antibiotics, or may take a certain time to affect the organ, as with gastrointestinal hemorrhage subsequent to NSAID use. This profile of risk was used to explain variations in the risk of agranulocytosis associated with the

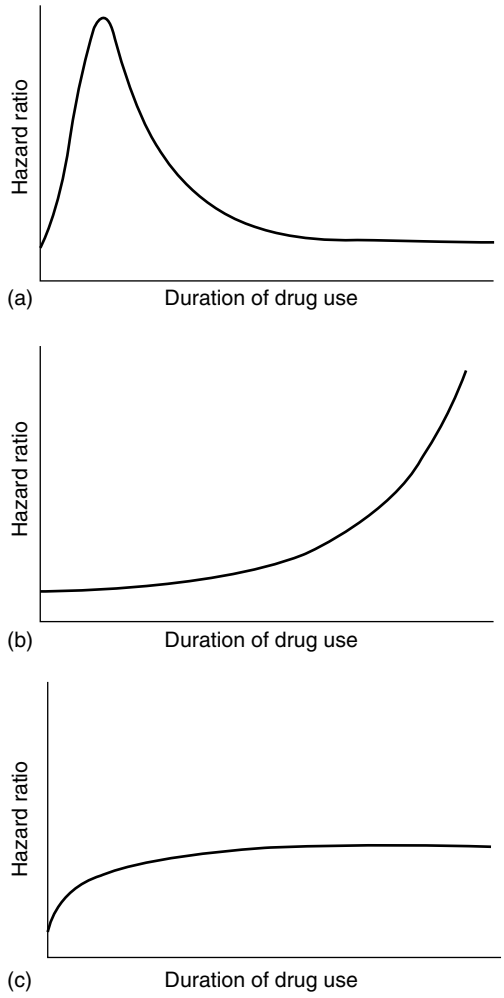

Figure 1 Different risk profiles by duration of drug use: (a) acute effect; (b) increasing risk; (c) constant risk

use of the analgesic dipyrone [16]. It has recently been used to assess the risk profile of oral contraceptives [47]. Figure 1(b) shows a gradually increasing model of risk, associated with diseases of longer latency such as cancer. Figure 1(c) displays the constant hazard model, after a rapid rise in the risk level.

Unfortunately, such graphs are not part of the analysis plan of most pharmacoepidemiologic studies at this point, despite the existence of appropriate techniques [9, 16]. The next few years should see an increasing use of **spline functions** and other similar tools to model the risk of drugs by their duration of use. The wider access to newer statistical software such as **S-PLUS** [42] among researchers in pharmacoepidemiology, as well as the publication of papers that simplify the understanding of these sophisticated approaches [13, 16], will encourage their wider use in a field where they are clearly pertinent.

Probabilistic Approach for Causality Assessment

The traditional epidemiologic approach to assess whether a drug causes an adverse reaction is based on the Hill criteria [18], that require the association to be biologically plausible, strong, specific, consistent and temporally valid (*see* **Hill's Criteria for Causality**). These criteria are applied to the results of pharmacoepidemiologic studies and, depending on the number of criteria satisfied, provide a level of confidence regarding causality of the drug (*see* **Causation**). The result of this exercise will be, if a drug is judged to cause an adverse reaction, that all exposed cases, or at least some etiologic proportion of these cases [39], are due to the drug. This approach is valuable for **inferences** to the population, but does not allow cases to be assessed individually, does not incorporate the specifics of the case, and does not entirely address the unique features of drugs as an exposure entity in epidemiology.

The study of individual cases of adverse reactions has been the mainstay of national pharmacovigilance centers throughout the world for several decades [3, 51] (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**). When a case report of a suspected drug-associated adverse event is received, the natural question is whether the drug actually caused the event. Several qualitative approaches have been proposed to answer this question [20, 25]. Recently, however, a formal quantitative approach using biostatistical foundations has been put forward [26–28]. It is based on **Bayes' theorem**, which can be used in the following way:

$$\Pr(D \rightarrow E|B, C) = \frac{\Pr(D \rightarrow E|B)\Pr(C|D \rightarrow E, B)}{\Pr(C)},$$

where $D \rightarrow E$ denotes that the drug D causes the adverse event E, B represents the background characteristics of the case that are known to affect the risk, while C represents the case information.

This Bayes' theorem approach allows us to estimate the posterior probability that an adverse event was caused by a drug by separating the problem into two components. The first component is the **prior probability** of the event given the baseline characteristics of the patients. This is estimated from existing data obtained from clinical trials or epidemiologic studies. The second component is the probability of case information given that the drug caused the event.

The primary limitation of this approach is the scarcity of data available to estimate the two components. In many instances, it is difficult to find the clinical and epidemiologic data necessary to estimate the prior probability, especially for rare clinical conditions that have not been the object of extensive population-based research. The same limitation applies to the second probability component because of the problems of finding cases relevant to proven drug causation. **Case series** that apply directly to the case being assessed are often difficult to find. These limitations are real but not limited to the Bayesian approach – they are a general problem in the assessment of individual cases. The authors of these methods suggest that the primary purpose of the Bayesian approach is to provide a framework in which subjective judgments relevant to assessment of an individual case are coherently combined.

To facilitate the use of the Bayesian method, the equation is usually expressed in terms of **odds** rather than probabilities. This formulation simplifies somewhat the need for data, since epidemiologic studies more frequently report odds ratios than absolute probabilities. The relative **likelihood** may also be easier

to estimate subjectively. This equation formulated in terms of odds is given as

$$\begin{aligned}\text{posterior odds} &= \Pr(D \rightarrow E|B, C)/\Pr(D \not\rightarrow E|B, C)\\ &= [\Pr(D \rightarrow E|B)/\Pr(D \not\rightarrow E|B)]\\ &\qquad \text{prior odds}\\ &\times[\Pr(C|D \rightarrow E, B)/\Pr(C|D \not\rightarrow E, B)]\\ &\qquad \text{likelihood ratio.}\end{aligned}$$

This approach was applied on several occasions [19, 23, 24, 34] and was recently made user-friendly by computerizing [21]. The increasing amount of new epidemiologic data on disease distribution and risk factors combined with new clinical and pharmacologic insights on drug effects will make this probabilistic approach more effective in future uses.

Methods Based on Prescription Data

One of the distinguishing features of pharmacoepidemiology is the use of computerized administrative health databases to answer research questions reliably and with sufficient rapidity. The usual urgency of concerns related to drug safety makes these databases essential to perform such risk assessment studies. In particular, databases containing only information on prescriptions dispensed to patients, and no outcome information on disease diagnoses, hospitalizations or vital status, have been the object of interesting statistical developments. These standalone prescription drug databases, that do not require to be linked to outcomes databases, are more numerous and usually more easily accessible than the fully linked databases. They provide a source of data that allows the investigation of patterns of drug use that can yield some insight into the validity of risk assessment studies as well as generate and test hypotheses about these risks. In this section we briefly review some of the resourceful uses of these drug prescription databases in pharmacoepidemiology.

A technique that was developed specifically for the context of drug databases is prescription sequence analysis [36]. Prescription sequence analysis is based on the situation when a certain drug A is suspected of causing an adverse event that itself is treated by a drug B. To apply this technique, the computerized drug database is searched for all patients using drug A. For these subjects, all patients prescribed drug B in the course of using drug A are identified and counted. Under the **null hypothesis** that drug A does not cause the adverse event treated by drug B, this number of subjects should be proportional to the duration of use of drug A relative to the total period of observation. This extremely rapid method of assessing the association between drug A and drug B is assessed for its **random error** with a **Monte Carlo** simulation analysis. This technique was applied to assess whether patients using the antivertigo or antimigraine drug flunarizine (drug A) causes mental depression, as measured by the use of antidepressant drugs (drug B). The authors found that the number of patients starting on antidepressant drugs during flunarizine use was in fact lower than expected [36]. They thus concluded, using this rapid approach based solely on drug prescription data, that this drug probably does not cause mental depression. An extension of prescription sequence analysis, called prescription sequence symmetry analysis, was recently proposed [17]. Using a population of new users of either drug A or B, this approach compares the number of subjects who used drug A before drug B to that using B before A. Under the null hypothesis, this distribution should be symmetrical and the numbers should be equal.

Another function of these databases is to use the prescriptions as **covariate** information to explain possible confounding patterns. The concept of *channeling* of drugs was put forward as an explanation of unusual risk findings [48]. For example, a case–control study conducted in New Zealand found that fenoterol, a beta agonist bronchodilator used to treat asthma attacks, was associated with an increased risk of death from asthma [8]. Using a prescription drugs database, it was found that severe asthmatics, as deemed from their use of other asthma medications prescribed for severe forms of the disease, were in fact channeled to fenoterol, probably because fenoterol was felt by prescribers to be a more potent bronchodilator than other beta agonists [35]. This phenomenon of channeling can be assessed rapidly in such databases, provided medications can be used as proxies for disease severity. This approach can be subject to bias, however, as it has been used with **cross-sectional** designs that cannot differentiate the directionality of the association. An application of channeling using a **longitudinal** design was recently presented [4]. It indicated that channeling can vary according to the timing of exposure, i.e. that disease severity was not associated with first-time use of a

drug, but subsequently severe patients were more likely to be switched to that drug. This type of research into patterns of drug prescribing and drug use can be very useful in understanding the results of case–control studies with limited data on drug exposures and subject to confounding by indication (*see* **Pharmacoepidemiology, Adverse and Beneficial Effects**).

These prescription drug databases have also been used to study patterns of interchange in the dispensing of NSAIDs. Such research is important because switching patterns permits the identification of brands that may not be well tolerated and result in the prescription of another agent. By using a **stochastic** approach, Walker et al. [50] estimated the **transition** probabilities from one NSAID to another. For a set of k different brands of NSAID, they derived the expected marginal distributions of the transition matrix that corresponds to a global equilibrium state by solving a system of $k + 1$ equations with $k + 1$ unknowns. By comparing these expected values with the observed marginals of the transition matrix, it was possible to assess whether this population had reached this stable state and for which drugs. Such models can be used rapidly to assess patterns of interchange and identify potentially harmful agents.

Finally, these prescription drugs databases may in certain situations provide all the necessary data for a conventional cohort or case–control study. For instance, the use of beta blockers to treat hypertension and other cardiac diseases has been hypothesized to cause depression. A prescription for an antidepressant drug can be used as a proxy for the outcome of depression. In this way, a standalone prescription drug database can provide data on exposure to beta blockers, on the outcome of depression, as well as on covariate information from other medications [2, 7].

Acknowledgments

Samy Suissa is a senior research scholar supported by the Fonds de la Recherche en Santé du Quebec (FRSQ). The McGill Pharmacoepidemiology Research Unit is funded by the FRSQ and by grants from the National Health Research and Development Programme (NHRDP) and the Medical Research Council (MRC) of Canada.

References

[1] Aelony, Y., Laks, M.M. & Beall, G. (1975). An electrocardiographic pattern of acute myocardial infarction associated with excessive use of aerosolized isoproterenol, *Chest* **68**, 107–110.

[2] Avorn, J., Everitt, D.E. & Weiss, S. (1986). Increased antidepressant use in patients prescribed beta-blockers. *Journal of the American Medical Association* **255**, 357–360.

[3] Baum, C., Kweder, S.L. & Anello, C. (1994). The spontaneous reporting system in the United States, in *Pharmacoepidemiology*, 2nd Ed, B.L. Strom, ed. Wiley, New York, pp. 125–138.

[4] Blais, L., Ernst, P. & Suissa, S. (1996). Confounding by indication and channeling over time: the risks of beta-agonists, *American Journal of Epidemiology* **144**, 1161–1169.

[5] Breslow, N. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. I: *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

[6] Breslow, N. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*, 2nd Ed. International Agency for Research on Cancer, Lyon.

[7] Bright, R.A. & Everitt, D.E. (1992). Beta-blockers and depression: evidence against an association, *Journal of the American Medical Association* **267**, 1783–1787.

[8] Crane, J., Pearce, N., Flatt, A., Burgess, C., Jackson, R., Kwong, T., Ball, M. & Beasley, R. (1989). Prescribed fenoterol and death from asthma in New Zealand 1981–1983: case-control study, *Lancet* **1**, 917–922.

[9] Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve, *Journal of the American Statistical Association* **83**, 414–425.

[10] Farmer, R.D.T., Lawrenson, R.A., Thompson, C.R., Kennedy, J.G. & Hambleton, I.R. (1997). Population-based study of risk of venous thromboembolism associated with various oral contraceptives, *Lancet* **349**, 83–88.

[11] Farrington, C.P., Nash, J. & Miller, E. (1996). Case series analysis of adverse reactions to vaccines: a comparative evaluation, *American Journal of Epidemiology* **143**, 1165–1173.

[12] Gong, L., Zhang, W., Zhu, Y., Zhu, J., Kong, D., Page, V., Ghadirian, P., Lehorier, J. & Hamet, P. (1996). Shanghai trial of Nifedipine in the elderly (STONE), *Journal of Hypertension* **19**, 1–9.

[13] Greenland, S. (1995). Dose-response and trend analysis in epidemiology: alternatives to categorical analysis, *Epidemiology* **6**, 356–365.

[14] Greenland, S. (1996). Confounding and exposure trends in case-crossover and case-time-control design, *Epidemiology* **7**, 231–239.

[15] Griffin, M.R., Ray, W.A. & Schaffner, W. (1988). Nonsteroidal anti-inflammatory drug use and death from peptic ulcer in elderly persons, *Annals of Internal Medicine* **109**, 359–363.

[16] Guess, H.A. (1989). Behavior of the exposure odds ratio in a case-control study when the hazard function is not

constant over time, *Journal of Clinical Epidemiology* **42**, 1179–1184.

[17] Hallas, J. (1996). Evidence of depression provoked by cardiovascular medication – a prescription sequence symmetry analysis, *Epidemiology* **7**, 478–484.

[18] Hill, A.B. (1965). The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[19] Hutchinson, T.A. (1986). A Bayesian approach to assessment of adverse drug reactions: evaluation of a case of acute renal failure, *Drug Information Journal* **20**, 475–482.

[20] Hutchinson, T.A., Leventhal, J.M., Kramer, M.S., Karch, F.E., Lipman, A.G. & Feinstein, A.R. (1979). An algorithm for the operational assessment of adverse drug reactions. II. Demonstration of reproducibility and validity, *Journal of the American Medical Association* **242**, 633–638.

[21] Hutchinson, T.A., Dawid, A.P., Spiegelhalter, D.J., Cowell, R.G. & Roden, S. (1991). Computer aids for probabilistic assessment of drug safety I: A spread-sheet program, *Drug Information Journal* **25**, 29–39.

[22] Jick, H., Jick, S.S., Gurewich, V., Myers, M.W. & Vasilakis, C. (1995). Risk of idiopathic cardiovascular death and nonfatal venous thromboembolism in women using oral contraceptives with differing progestogen components, *Lancet* **346**, 1589–1593.

[23] Jones, J.K. (1986). Evaluation of a case of Stevens-Johnson syndrome, *Drug Information Journal* **20**, 487–502.

[24] Kramer, M.S. (1986). A Bayesian approach to assessment of adverse drug reactions: evaluation of a case of fatal anaphylaxis, *Drug Information Journal* **20**, 505–518.

[25] Kramer, M.S. Leventhal, J.M., Hutchinson, T.A. & Feinstein, A.R. (1979). An algorithm for the operational assessment of adverse drug reactions. I. Background, description, and instructions for use, *Journal of the American Medical Association* **242**, 623–632.

[26] Lane, D.A. (1984). A probabilist's view of causality assessment, *Drug Information Journal* **18**, 323–330.

[27] Lane, D.A. (1986). The Bayesian approach to causality assessment, *Drug Information Journal* **20**, 455–461.

[28] Lane, D.A., Kramer, M.S., Hutchinson, T.A., Jones, J.K. & Naranjo, C.A. (1987). The causality assessment of adverse drug reactions using a Bayesian approach, *Journal of Pharmaceutical Medicine* **2**, 265–283.

[29] Lenz, W. (1966). Malformations caused by drugs in pregnancy, *American Journal of Diseases of Children* **112**, 99–106.

[30] Maclure, M. (1991). The case-crossover design: a method for studying transient effects on the risk of acute events, *American Journal of Epidemiology* **133**, 144–153.

[31] McBride, W.G. (1961). Thalidomide and congenital abnormalities, *Lancet* **ii**, 1358.

[32] Miettinen, O.S. (1983). The need for randomization in the study of intended effects, *Statistics in Medicine* **2**, 267–271.

[33] Miettinen, O.S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. Wiley, New York.

[34] Naranjo, C.A., Lanctot, K.L. & Lane, D.A. (1990). The Bayesian differential diagnosis of neutropenia associated with antiarrhythmic agents, *Journal of Clinical Pharmacology* **30**, 1120–1127.

[35] Petri, H. & Urquhart, J. (1991). Channeling bias in the interpretation of drug effects, *Statistics in Medicine* **10**, 577–581.

[36] Petri, H., De Vet, H.C.W., Naus, J. & Urquhart, J. (1988). Prescription sequence analysis: a new and fast method for assessing certain adverse reactions of prescription drugs in large populations, *Statistics in Medicine* **7**, 1171–1175.

[37] Ray, W.A. & Griffin, M.R. (1989). Use of Medicaid data for pharmacoepidemiology, *American Journal of Epidemiology* **129**, 837–849.

[38] Ray, W.A., Fought, R.L. & Decker, M.D. (1992). Psychoactive drugs and the risk of injurious motor vehicle crashes in elderly drivers, *American Journal of Epidemiology* **136**, 873–883.

[39] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company, Boston.

[40] Spitzer, W.O., Suissa, S., Ernst, P., Horwitz, R.I., Habbick, B., Cockcroft, D., Boivin, J.F., McNutt, M., Buist, A.S. & Rebuck, A.S. The use of beta-agonists and the risk of death and near death from asthma, *New England Journal of Medicine* **326**, 501–506.

[41] Spitzer, W.O., Lewis, M.A., Heinemann, L.A., Thorogood, M. & MacRae, K.D. (1996). Third generation oral contraceptives and risk of venous thromboembolic disorders: an international case-control study. Transactional Research Group on Oral Contraceptives and the Health of Young Women, *British Medical Journal* **312**, 83–88.

[42] StatSci (1995). *S-PLUS Version 3.3*. StatSci, a division of MathSoft, Inc., Seattle.

[43] Sturkenboom, M.C., Middelbeek, A., de Jong, L.T., van den Berg, P.B., Stricker, B.H. & Wesseling, H. (1995). Vulvo-vaginal candidiasis associated with acitretin, *Journal of Clinical Epidemiology* **48**, 991–997.

[44] Suissa, S. (1995). The case-time-control design, *Epidemiology* **6**, 248–253.

[45] Suissa, S. & Edwardes, M. (1997). Adjusted odds ratios for case-control studies with missing confounder data in controls. *Epidemiology* **8**, 275–280.

[46] Suissa, S., Hemmelgarn, B., Blais, L. & Ernst, P. (1996). Bronchodilators and acute cardiac death, *Journal of Respiratory and Critical Care Medicine* **154**, 1598–1602.

[47] Suissa, S., Blais, L., Spitzer, W.O., Cusson, J., Lewis, M. & Heinemann, L. (1997). First-time use of newer oral contraceptives and the risk of venous thromboembolism, *Contraception* **56**, 141–146.

[48] Urquhart, J. (1989). ADR crisis management, *Scrip* **1388**, 19–21.

[49] Walker, A.M. (1996). Confounding by indication, *Epidemiology* **7**, 335–336.
[50] Walker, A.M., Chan, K.W.A. & Yood, R.A. (1992). Patterns of interchange in the dispensing of non-steroidal anti-inflammatory drugs, *American Journal of Epidemiology* **45**, 187–195.
[51] Wiholm, B.E., Olsson, S., Moore, N. & Wood, S. (1994). Spontaneous reporting systems outside the United States, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, New York, pp. 139–156.
[52] World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception, (1995). Venous thromboembolic disease and combined oral contraceptives: results of international multicentre case-control study. World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception, *Lancet* **346**, 1575–1582.

(*See also* **Drug Approval and Regulation**; **Pharmacoepidemiology, Study Designs**)

S. SUISSA

Pharmacoepidemiology, Study Designs

The field of **pharmacoepidemiology** includes the study of the use of and effects of pharmaceuticals in populations [10]. In general, the study designs used in pharmacoepidemiology are the same as those used in other areas of **clinical epidemiology**. There are three key differences, however.

First, because pharmacoepidemiology studies are usually performed after drug marketing, and because 500–3000 patients are generally studied prior to drug marketing, pharmacoepidemiology studies usually must include substantially larger numbers of patients in a **cohort study** or, alternatively, tap an equivalently sized population for a **case–control study**, in order to contribute new useful information.

Secondly, because at least one randomized **clinical trial** was already performed prior to drug marketing, pharmacoepidemiology studies are less likely to use randomized clinical trial study designs; many of the same limitations which the premarketing randomized clinical trials were subject to would apply as well to any postmarketing randomized clinical trial, and so they would not be able to contribute new useful information. For example, because of the need for huge sample sizes, randomized clinical trials are not an efficient means of studying uncommon adverse effects, or the effects of drugs in types of patients commonly excluded from such trials.

Thirdly, because pharmacoepidemiology questions often arise as regulatory, commercial and public health crises, answers must often be obtained very quickly.

This need for rapidly performed studies of massive sample size has led to a series of special approaches which have characterized the field of pharmacoepidemiology, and will be the primary focus of this article. Other analytic issues which are special to pharmacoepidemiology include the need for special attention to the drug regimen. In many other areas of clinical epidemiology, "exposure" is frequently treated as a dichotomous variable. Even when studying drugs, most randomized clinical trials specify a single fixed dose of the drug of interest. In contrast, once a drug is on the market, questions of how its effects vary with different doses, different durations of therapy, and different regimens (e.g. intermittent vs. continuous administration) are often the questions which are of greatest clinical importance. For reasons of space, these have not been discussed here. The reader is referred elsewhere for such considerations [2].

Data Resources Used in Pharmacoepidemiology

Historically, the primary data resource used in pharmacoepidemiology was the spontaneous reporting system [1, 18]. This is a nonsystematic collection of case reports (*see* **Case Series, Case Reports**) of adverse events following use of drugs, considered by the treating physician as possibly due to the drug. These are reports to the medical literature but, more voluminously, to regulatory bodies (*see* **Drug Approval and Regulation**). As case reports, they are useful primarily for generating rather than testing hypotheses. For example, case reports of acute flank pain following the use of suprofen generated formal studies which tested and confirmed the resulting hypothesis [14], while analogous spontaneous reports of anaphylactic reactions to tolmetin were not confirmed when these hypotheses were formally tested [13].

Another approach to pharmacoepidemiology studies uses **vital statistics** data and **drug utilization** data to perform analyses of secular trends, searching for whether trends in drug exposure over time or across geographic areas correlate with trends in disease occurrence [9]. For example, with the marketing of oral contraceptives, mortality rates from pulmonary embolism increased, but only in women of reproductive age [5]. While this type of study is easy to perform, it obviously is prone to many difficulties, including the limitations inherent in vital statistics data as well as identifying which of the possible **correlations** truly reflect cause vs. simply coincidence (*see* **Hill's Criteria for Causality**). For example, in a study using data from 18 different cancer registries around the world to investigate the relationship between sales of methyldopa and the development of biliary carcinoma [12], the results from one of the cancer registries showed an artifactual association, caused by changes in coding practices in that registry.

Over the past two decades, pharmacoepidemiology has been to the fore in scientific fields using automated databases of claims information (*see* **Administrative Databases**) for its research [7, 11].

In clinical epidemiology studies, the largest expense of the study is generally that of data collection, and this is particularly problematic in the very large studies of pharmacoepidemiology. However, using these automated databases (*see* **Database Systems**), the substantial cost of data collection for these very large studies is borne by the underlying insurance system, rather than the study. A large proportion of pharmacoepidemiology studies are now using such systems. However, given the relatively small number of exposed diseased individuals relative to unexposed diseased individuals, in using an automated database to implement such a study one needs to obtain a huge number of medical records of diseased individuals, relatively few of whom are exposed. Inasmuch as most of the cost of this type of study is the cost of obtaining these medical records, this is inefficient, and new methods are needed to enable one to sample from unexposed diseased individuals. The major weakness of these systems is the uncertain validity of the data, especially the diagnosis data [11]. Furthermore, these systems can lack information about key potential **confounding** variables, if they do not come to medical attention. Thus, there is increasing interest in the exploration of medical record databases, as opposed to claims databases, for such research [3].

Historically, there have been a few ongoing systems of *ad hoc* data collection, tailored for pharmacoepidemiology research. One of these has been a hospital-based system of collecting drug exposures and outcomes in hospitals, pioneered by the Boston Collaborative Drug Surveillance Program [4]. This system has the advantage of known **incidence rates** and high-quality data collected on site in the hospital, but an inability to study either the many important drugs used in outpatients, or uncommon adverse reactions, even if serious. New data collection in this system was abandoned many years ago, although a few hospitals have mounted analogous systems elsewhere.

Another such system has been the system of case–control surveillance developed by the Drug Epidemiology Unit, now the Slone Epidemiology Unit at Boston University [8]. This also focuses on hospitals, but collects information on prior drug exposures as possible causes of the hospitalization, performing **hospital-based case–control studies**. This system suffers from the uncertain validity of the drug exposure data obtained from patients and from restriction to hospital-based case–control studies, with the inherent problems of **selection bias**. In recent years, data collection for this system has been curtailed.

Finally, many pharmacoepidemiology studies are still designed as *ad hoc* clinical epidemiology studies, whether cohort or case–control. The choices among these designs are discussed elsewhere.

Special Methodological Approaches Used to Apply These Resources

In applying the special resources described above, unique challenges are confronted by pharmacoepidemiologists, in part because of the nature of the information being collected, and in part because of the large sample sizes required. These are each discussed in turn below.

Validity of Exposure and Outcome Data

In order to perform a valid epidemiologic study, one obviously must have valid information on both exposure and outcome. This can be very problematic in pharmacoepidemiologic studies. The best measures of disease occurrence are medical records, as patients often do not understand the details of the diseases which they have. Case–control studies have a major advantage here, as patients can be identified from their health care providers. However, medical records are very poor sources of information about prior drug use, as drugs tend to be recorded very incompletely [15]. Obtaining drug histories directly from patients can be problematic, however, as most patients cannot identify the drugs they are on now, even less the drugs they took in the past. Special techniques have been developed by pharmacoepidemiologists in order to maximize the validity of the drug data collected from patients as part of case–control studies, e.g. the use of indicator prompts and pictorial handouts. The details of these approaches are beyond the scope of this article, but the reader is referred elsewhere for them [17]. However, suffice it to say that much work remains to be done on these issues.

In contrast, data on drug exposure from claims databases are extremely valid, as they represent documentation of the exact drugs dispensed to patients. Reimbursement by insurance carriers varies according to the identity and amount of the drug [11]. While

this does not assure compliance with the dispensed drug, it is a level better than prescribing information, as many prescribed drugs are never dispensed. In contrast, however, the diagnosis information in these databases is of uncertain validity, as reimbursement generally does not depend on diagnosis and, especially, on correct and precise diagnoses [11]. As such, in studies using these databases, considerable attention needs to be paid to obtaining validation of these diagnoses.

Special Study Designs Used to Increase Efficiency

Because of the large numbers of individuals included in many pharmacoepidemiology studies, even when using claims databases, pharmacoepidemiologists seek special study designs to enable the data processing and, especially, any medical record review to be more efficient. In general, pharmacoepidemiologists are studying diseases of low incidence. Furthermore, any given drug is used by a small proportion of the population. As such, one is investigating a low **prevalence** of exposure and a low **incidence** of disease. To the degree one includes general population samples, therefore, one collects information on a large number of people who do not contribute much additional statistical information to the investigation. Case–control studies can be useful toward this end, when the prevalence of exposure is high. However, for many drugs this is not applicable.

Another approach which is used is the **nested case–control study**. In this design, an investigator first creates a cohort of exposed individuals and then, within that cohort, identifies cases and a **random sample** of noncases for the study. This design is efficient and allows one to use conventional statistical methods for the analysis of case–control studies, which is a major advantage. However, it can result in logistical problems in identifying the sample of noncases, as they must be at risk of developing the disease at the same time as the case, and identifying these risk sets can be difficult [16].

Another design beginning to be used by pharmacoepidemiologists is the **case–cohort study**, an approach pioneered in **occupational epidemiology** [6]. In this situation, one identifies a cohort of exposed subjects in advance, and the subset of people who are cases, as with the nested case–control study. However, instead of sampling randomly from the known noncases those at risk of developing the disease at the same time as the case, one randomly samples from the entire cohort, creating a subcohort, which also can include some of the cases. Then, the distribution of exposures and **confounders** in the case group can be compared for analytic purposes to the distribution in the subcohort. This approach has the advantage of achieving much smaller sample sizes with a simple sampling scheme. Also, one can study multiple outcomes within the same study. It has the disadvantage of requiring analyses different from those of normal case–control studies. Instead, a modified Cox **proportional hazards** approach is used and, until recently, software was not available to implement this.

Conclusions

In conclusion, pharmacoepidemiologists use the same methods of study design as do other clinical epidemiologists. However, because of the special characteristics of the field, there are special issues of study design and analysis which arise. These have been discussed briefly in this article.

References

[1] Baum, C., Kweder, S.L. & Anello, C. (1994). The spontaneous reporting system in the United States, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, Chichester, pp. 125–138.

[2] Guess, H.A. (1989). Behavior of the exposure odds ratio in a case–control study when the hazard function is not constant over time, *Journal of Clinical Epidemiology* **42**, 1179–1184.

[3] Hall, G. (1992). Pharmacoepidemiology using a UK database of primary care records, *Pharmacoepidemiology and Drug Safety* **1**, 33–37.

[4] Lawson, D.H. & Beard, K. (1994). Intensive hospital-based cohort studies, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, Chichester, pp. 157–170.

[5] Markush, R.E. & Seigel, D.G. (1969). Oral contraceptives and mortality from thromboembolism in the United States, *American Journal of Public Health* **59**, 418–434.

[6] Prentice, R.L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[7] Ray, W.A. & Griffin, M.R. (1989). The use of Medicaid data for pharmacoepidemiology, *American Journal of Epidemiology*, **129**, 837–849.

[8] Shapiro, S. (1994). Case–control surveillance, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, Chichester, pp. 301–322.

[9] Stolley, P.D. (1982). The use of vital and morbidity statistics for the detection of adverse drug reactions and for monitoring of drug safety, *Journal of Clinical Pharmacology* **22**, 499–504.

[10] Strom, B.L. (1994). What is pharmacoepidemiology?, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, Chichester, pp. 3–14.

[11] Strom, B.L. & Carson, J.L. (1990). Use of automated databases for pharmacoepidemiology research, *Epidemiologic Reviews* **12**, 87–107.

[12] Strom, B.L., Hibberd, P.L. & Stolley, P.D. (1985). No evidence of association-between methyldopa and biliary carcinoma, *International Journal of Epidemiology* **14**, 86–90.

[13] Strom, B.L., Carson, J.L., Schinnar, R., Sim, E. & Morse, M.L. (1988). The effect of indication on the risk of hypersensitivity reactions associated with tolmetin sodium vs. other nonsteroidal antiinflammatory drugs, *Journal of Rheumatology* **15**, 695–699.

[14] Strom, B.L., West, S.L., Sim, E. & Carson, J.L. (1989). The epidemiology of the acute flank pain syndrome from suprofen, *Clinical Pharmacologic Therapy* **46**, 693–699.

[15] Strom, B.L., Carson, J.L., Halpern, A.C., Schinnar, R., Snyder, E.S., Stolley, P.D., Shaw, M., Tilson, H.H., Joseph, M., Dai, W.S., Chen, D., Stern, R.S., Bergman, U. & Lundin, F. (1991). Using a claims database to investigate drug-induced Stevens-Johnson syndrome, *Statistics in Medicine* **10**, 565–576.

[16] Suissa, S. (1994). Novel approaches to pharmacoepidemiology study design and statistical analysis, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, Chichester, pp. 629–646.

[17] West, S.L. & Strom, B.L. (1994). Validity of pharmacoepidemiology drug and diagnosis data, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, Chichester, pp. 549–580.

[18] Wiholm, B.-E., Olsson, S., Moore, N. & Wood, S. (1994). Spontaneous reporting systems outside the United States, in *Pharmacoepidemiology*, 2nd Ed., B.L. Strom, ed. Wiley, Chichester, pp. 139–156.

(*See also* **Dose-response in Pharmacoepidemiology**; **Drug Approval and Regulation**; **Drug Interactions**; **Drug Utilization Patterns**; **Pharmacoepidemiology, Adverse and Beneficial Effects**; **Postmarketing Surveillance of New Drugs and Assessment of Risk**)

B.L. STROM

Pharmacogenetics

Most **complex diseases** are syndromes rather than distinct diseases, and probably have multiple environmental and genetic determinants [3, 13, 29]. A component of this complexity in diseases that can be treated with pharmacotherapy is often a highly variable response to pharmacological therapy among individual patients [21]. Pharmacogenetics is the study of the role of genetic determinants in the variable response to therapy. Ideally, we would be able to stratify a population needing treatment into those likely, or unlikely, to respond to treatment as well as those likely, or unlikely, to experience adverse side-effects [24].

Variability in individual drug treatment response may be due to many factors, including the severity and type of disease, treatment compliance, the presence of other illness, the use of other drugs (drug–drug **interaction**), environmental exposures, sex and age. However, family-based studies have suggested that genetic factors underlie a significant proportion of the observed treatment **variance** in many diseases [11].

Although many pharmacogenetic mechanisms are possible, specific **DNA** sequence variants may alter response to drugs in three main ways (Table 1) [12]:

1. Variation in the metabolism of a drug among individuals, especially in enzymes involved in the catabolism or excretion of a drug. An important example is the highly genetically diverse cytochrome P450 system, known to have many pharmacogenetic effects [10, 27].
2. Variation among population members with respect to drug adverse effects that are not based on the drug's action.
3. Variation in the drug treatment target or target pathways. In this category, a population is conceptually divided into responders and nonresponders, and analysis of genetic variants (*see* **Genotype**) is used in an attempt to distinguish these groups [24].

Table 1 Possible pharmacogenetic mechanisms[a]

Genetic variants associated with:	Type of mechanism
Altered uptake, distribution or metabolism of the drug administered	Pharmacokinetic
An unintended action of a drug outside of its therapeutic indication	Idiosyncratic
Alterations in the drug target or a component of the drug pathway leading to altered drug efficacy; or	Pharmacodynamic
Differences in the expression of a physiological phenotype (*see* **Genotype**) such that a given target may not be disease-associated in a given patient	

[a]Reproduced by permission of Churchill Livingstone.

The current trend in genetic analysis of complex human diseases is away from family-based strategies using microsatellite **markers** towards single nucleotide polymorphisms (SNPs) [SNPs are discussed in the article, Markers] genotyping and different analytical strategies based on **association** and **haplotype analysis** [14, 18, 20]. Since response to drug treatment generally varies with age, and the number and type of medications for a given disease is changing rapidly, it is unlikely that family-based treatment data will be available in the foreseeable future for most complex diseases. In the absence of these data, **case–control** association studies are the approach of choice (*see* **Disease-marker Association**). Case–control association analyses are now recognized as being well suited for localizing susceptibility loci [23], and they are intrinsically more powerful than **linkage analyses** in detecting weak genetic effects [3].

Pharmacogenetic studies have generally investigated associations between drug response phenotypes and genotyped DNA sequence variants, most commonly "SNPs". The last decade has seen dramatic increases in molecular genetic technologies that can potentially be used to understand the biologic basis of pharmacogenomics [22]. Because of their potential biologic importance, the common SNPs in the human genome increasingly have been the subject of large-scale cataloguing projects funded by both government and industry groups [2, 5, 9] (*see* **Human Genome Project**). Limitations related to cost and the current incomplete status of SNP databases has meant that the association analysis of SNPs in pharmacogenomics has so far been limited to polymorphisms within biologically plausible candidate loci (*see* **Gene**).

Although genetic information has only recently begun to be integrated into the clinical trial setting, there is now a growing list of candidate genes being investigated for association with treatment response in many different diseases [30]. Most pharmacogenomic studies published to date have essentially been *post hoc* genetic studies undertaken using DNA and phenotypic data from subjects who had been enrolled in a conventional clinical trial.

Statistical Issues

The testing of large numbers of SNPs for association with one or more traits raises important statistical issues regarding the appropriate false-positive rate of the tests and the level of statistical significance to be adopted given the multiple testing (*see* **Multiple Comparisons**) involved [19]. Many relevant areas of genetic statistics are under active methodological development [18, 25]. Two major potential statistical issues in pharmacogenetics relate to population **stratification** and statistical power.

Genetic Heterogeneity and Population Stratification

In addition to variation in allele frequencies, there is also a high degree of variation in the strength of **linkage disequilibrium** in a given chromosomal region among populations of different origins [31] and also between different genomic regions [6, 26]. Such genetic heterogeneity is a major challenge to the discovery of genes that modulate pharmacogenetics pathways. An important limitation of case–control association studies related to heterogeneity is the potential that undetected population stratification will produce misleading evidence of association.

Population stratification may cause spurious associations in a case–control study when allelic frequencies vary across subpopulations in a study cohort. For example, if there is an imbalance in ethnic group representation between the case and control cohorts, one could detect a spurious association [4]. Such population stratification may result from recent **admixture** or from poorly matched cases and controls. Genotyping of unlinked panels of SNPs, chosen without regard to the phenotype of interest, can be used to ensure that case and control populations are genetically homogeneous. Methods have recently been developed to assess population stratification and, if necessary, to test correctly for association in the presence of such stratification [15–17]. However, neither systematic testing for population stratification nor application of these new statistical methods has yet been incorporated into the great majority of pharmacogenetic studies of complex human diseases.

Statistical Power

Growing experience with complex disease genetics has made clear the need to minimize type I error in genetic studies [7, 18]. Table 2 shows some simple

Table 2 Sample size requirements for case–control analyses of SNPs (one control per case; detectable difference of OR $\geq$ 1.5; power = 80%)

	Dominant model[c]			Recessive model[d]		
		Sample size required[e]			Sample size required[e]	
Allele frequency[a]	Exposure[b]	$\alpha = 0.05$	$\alpha = 0.005$	Exposure[b]	$\alpha = 0.05$	$\alpha = 0.005$
10%	19%	1162	1934	1%	16 730	27 822
20%	36%	834	1388	4%	4370	7366
30%	51%	818	1360	9%	2094	3484
40%	64%	936	1556	16%	1316	2188
50%	75%	1200	1994	25%	980	1630
60%	84%	1732	2882	36%	834	1388

[a]Frequency of risk-increasing allele in controls.
[b]Exposure (=prevalence) in controls assuming a diallelic locus with a dominant or recessive allele at the **Hardy–Weinberg equilibrium**.
[c]OR of 1.5 between cases and controls for possession of at least one copy of disease-associated SNP by case.
[d]OR of 1.5 between cases and controls for possession of two copies of disease-associated SNP by case.
[e]Required sample size = cases plus controls.

estimation of required sample sizes needed to detect a true **odds ratio** (OR) of 1.5 with 80% power and type I error probability (α) of either 0.05 or 0.005. Both mode of inheritance (dominant, recessive) (*see* **Genotype**) and allele frequency can have dramatic effects on required sample sizes (Table 2). Even for the "best case scenario" – a common SNP acting in a dominant fashion – a relatively large sample size of more than 800 subjects is required at an α of 0.05 (Table 1).

Multiple testing issues are likely in many genetic association studies of candidate loci where either multiple SNPs in one gene, multiple SNPs in several loci, or both [28] are tested, suggesting that a lower α such as 0.005 is probably more realistic than an α of 0.05. Using an α of 0.005 or assuming an uncommon SNP (allele frequency $\leq$0.10) that acts in a recessive fashion points to the need for very large sample sizes, i.e. more than 10 000 cases. Finally, Table 2 assumes an effect size (OR = 1.5) which, in the context of common, multifactorial diseases, may be quite large. Assuming a smaller effect may be more realistic for many genes, and would lead to concomitantly higher required sample sizes. **Simulation** studies have also suggested that genes of small effect are not likely to be detectable by association studies in sample sizes of less than 500 [8].

While these power calculations are simple and fairly conservative, they suggest that the sample sizes used in many of the small case–control pharmacogenetic association studies conducted to date had insufficient power to detect even a large effect associated with a SNP. This suggests that larger-scale studies than most of those currently being performed using data derived from standard clinical trials will be needed. As other researchers have suggested [1], the integration of genetic information into clinical trials will likely require a paradigm shift in the conduct and design of clinical trials. A central problem has been that the parameters of the **mutation**(s) affecting drug response (mode of inheritance, allele frequency, effect size) are not generally known at the start of a clinical trial. Study design remains one of the areas most in need of attention in pharmacogenetics.

Future Directions and Issues

The ultimate goal of pharmacogenetics is to understand the role that sequence variation among individuals and populations plays in the variability of responses to pharmaceuticals. The frequency and **penetrance** of a sequence variant affecting responsiveness to a particular drug and potential interactions with other genetic and environmental factors must ultimately be assessed in multiple population-based samples. A SNP must be relatively common and have a significant impact upon phenotype to be important at the population level in determining treatment response. These criteria become particularly important when extrapolating from specific clinical trials to general clinical use in the highly heterogeneous populations that are the current major markets for drug therapeutics. It is clear that large well-characterized cohort studies of population-based and ethnically diverse samples will be critical to the future success of any diagnostic SNP-based pharmacogenetic tests and for cost-effectiveness studies.

Thus far, pharmacogenetics studies have been limited to the **candidate gene** model. A new direction that is technically feasible at present, but as yet remains unexplored in pharmacogenomics, are SNP-based whole genome screens for variants associated with variation in drug response [31]. Other future directions include the use of pharmacogenomic data for the study of **gene–environment interactions** in determining response to pharmacologic therapy and for homogeneity testing and improving study design [12].

References

[1] Cardon, L.R., Idury, R.M., Harris, T.J., Witte, J.S. & Elston, R.C. (2000). Testing drug response in the presence of genetic information: sampling issues for clinical trials, *Pharmacogenetics* **10**, 503–510.

[2] Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. & Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998–2003, *Science* **282**, 682–689.

[3] Elston, R. (1995). The genetic dissection of multifactorial traits, *Clinical and Experimental Allergy* **2**, 103–106.

[4] Ewens, W. & Spielman, R. (1995). The transmission/disequilibrium test: history, subdivision, and admixture, *American Journal of Human Genetics* **57**, 455–464.

[5] Gray, I.C., Campbell, D.A. & Spurr, N.K. (2000). Single nucleotide polymorphisms as tools in human genetics, *Human Molecular Genetics* **9**, 2403–2408.

[6] Jorde, L.B., Watkins, W.S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A. & Leppert, M. (1994). Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region, *American Journal of Human Genetics* **54**, 884–898.

[7] Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**, 241–247.

[8] Long, A.D. & Langley, C.H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits, *Genome Research* **9**, 720–731.

[9] Masood, E. (1999). As consortium plans free SNP map of human genome (news), *Nature* **398**, 545–546.

[10] Meyer, U.A. (1994). Pharmacogenetics: the slow, the rapid, and the ultrarapid, *Proceedings of the National Academy of Sciences* **91**, 1983–1984.

[11] Nebert, D.W. (1999). Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist?, *Clinical Genetics* **56**, 247–258.

[12] Nebert, D.W. & Weber, W.W. (1990). armacogenetics, in *Principles of Drug Action: The Basis of Pharmacology*, W.B. Pratt & P. Taylor, eds. Churchill Livingston, New York.

[13] Olson, J.M., Witte, J.S. & Elston, R.C. (1999). Genetic mapping of complex traits, *Statistics in Medicine* **18**, 2961–2981.

[14] Palmer, L.J. & Cookson, W.O.C.M. (2001). Using single nucleotide polymorphisms (SNPs) as a means to understanding the pathophysiology of asthma, *Respiratory Research* **2**, 102–112.

[15] Pritchard, J.K. & Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies, *American Journal of Human Genetics* **65**, 220–228.

[16] Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. (2000). Association mapping in structured populations, *American Journal of Human Genetics* **67**, 170–181.

[17] Reich, D.E. & Goldstein, D.B. (2001). Detecting association in a case–control study while correcting for population stratification, *Genetic Epidemiology* **20**, 4–16.

[18] Risch, N.J. (2000). Searching for genetic determinants in the new millennium, *Nature* **405**, 847–856.

[19] Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science* **273**, 1516–1517.

[20] Schork, N.J., Fallin, D. & Lanchbury, J.S. (2000). Single nucleotide polymorphisms and the future of genetic epidemiology, *Clinical Genetics* **58**, 250–264.

[21] Schork, N.J., Fallin, D., Tiwari, H.K. & Schork, M.A. (2001). armacogenetics, in *Handbook of Statistical Genetics*, D.J. Balding, ed. Wiley, Chichester, pp. 741–764.

[22] Shi, M.M. (2001). Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies, *Clinical Chemistry* **47**, 164–172.

[23] Silverman, E.K. & Palmer, L.J. (2000). Case–control association studies for the genetics of complex respiratory diseases, *American Journal of Respiratory Cell and Molecular Biology* **22**, 645–648.

[24] Stephens, J.C. (1999). Single-nucleotide polymorphisms, haplotypes, and their relevance to pharmacogenetics, *Molecular Diagnostics* **4**, 309–317.

[25] Terwilliger, J.D. & Goring, H.H. (2000). Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design, *Human Biology* **72**, 63–132.

[26] Watkins, W.S., Zenger, R., O'Brien, E., Nyman, D., Eriksson, A.W., Renlund, M. & Jorde, L.B. (1994). Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region, *American Journal of Human Genetics* **55**, 348–355.

[27] Weber, W.W. (1984). Acetylation pharmacogenetics: experimental models for human toxicity, *Federation Proceedings* **43**, 2332–2337.

[28] Witte, J.S., Elston, R.C. & Cardon, L.R. (2000). On the relative sample size required for multiple comparisons, *Statistics in Medicine* **19**, 369–372.

[29] Witte, J.S., Elston, R.C. & Schork, N.J. (1996). Genetic dissection of complex traits (with discussion), *Nature Genetics* **12**, 355–358.

[30] Wolf, C.R. & Smith, G. (1999). Pharmacogenetics, *British Medical Bulletin* **55**, 366–386.

[31] Zavattari, P., Deidda, E., Whalen, M., Lampis, R., Mulargia, A., Loddo, M., Eaves, I., Mastio, G., Todd, J.A. & Cucca, F. (2000). Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection, *Human Molecular Genetics* **9**, 2947–2957.

LYLE J. PALMER

Pharmacokinetics and Pharmacodynamics

The final stage of establishing the safety and efficacy of a new drug, prior to submitting an application for marketing approval, typically involves one or more randomized **clinical trials** comparing the drug's effect to that in an appropriate control group. Sample size requirements in these pivotal (Phase III) trials may dictate that the number of treatment arms be kept to a minimum. As a result, it is not uncommon to study just one dose level of the new drug in pivotal trials. In this situation it is essential that adequate investigation of possible dosing strategies be carried out in earlier stages of clinical development to allow determination of the "best" dose for inclusion in Phase III testing.

Pharmacokinetic (PK) and pharmacodynamic (PD) modeling are key tools in proper dose selection. *Pharmacokinetics* attempts to characterize the fate of the drug in the body following dosing, primarily by sampling its concentration–time profile in the circulation. *Pharmacodynamics* investigates the relationship between the response induced by the drug and its circulating concentration. A common mnemonic is that pharmacokinetics deals with "what the body does to the drug", while pharmacodynamics focuses on "what the drug does to the body". Whereas the response of interest in PK modeling is always concentration of the drug and/or its metabolites, there is often more flexibility in the choice of response in PD models. Typically, it is some relevant biochemical or physiological marker of drug activity, that can be measured easily during the early stages of clinical development. In some cases this may be the intended final clinical endpoint (e.g. reduction in blood pressure). More commonly, the PD response, while providing some measure of the drug's biological activity, serves only as a **surrogate** for the clinical outcome ultimately of interest.

PK/PD modeling plays a role not only in human studies, but also in animal experiments conducted during drug development. Although animal studies involve certain unique aspects, the main issues in model derivation and fitting parallel those encountered with human data. Accordingly, the exposition here considers modeling of data from human subjects only. Understanding the PK and PD characteristics of a new drug is important, not only to identify promising dosing regimens to be tested in pivotal efficacy trials, but also to obtain information needed for labeling, to allow the prescribing physician to use the drug sensibly in clinical practice.

Pharmacokinetic Modeling

Objectives

Underlying most dosage regimens is the idea of a "therapeutic window", i.e. a range within which drug concentrations should be maintained to achieve clinical benefit. Concentrations that are too low may not achieve efficacy, whereas higher levels may result in undesirable side-effects. For instance, most antibiotics require a certain minimum inhibitory concentration (MIC) to be sustained to maintain efficacy against a particular target organism, but concentrations too far in excess of the MIC may be toxic. An effective dosing regimen should aim to reach concentrations within the therapeutic window as quickly as possible, and to stay within the desired range by suitable choices of maintenance dose and dosing interval. On the basis of the drug's PK characteristics in a "typical" subject, a dosing strategy can be designed which should maintain concentrations in the desired range (*see* **Minimum Therapeutically Effective Dose**).

For some classes of agents, assuming a common therapeutic window for all patients may be an oversimplification. Subject-specific genetic, physiologic, or demographic factors may alter the relationship between drug concentration and effect to such a degree that the desirable concentration range may differ substantially across patients, particularly if the therapeutic window is narrow. Even if the target concentration range is the same for all subjects, if differences in their PK characteristics are sufficiently marked it may be necessary to use different dosing strategies for some subjects to maintain concentrations in that range. For PK/PD modeling to provide useful guidance for dosing patients, it is thus a matter of practical importance that the models capture not only "typical" PK/PD behavior; they also need to explain variation in this behavior as fully as possible.

Following the target concentration paradigm described above, the primary clinical application of PK information is the development of a dosing regimen that maintains circulating drug concentrations

within the therapeutic window appropriate for a given patient. The relevant concentration range may be derived from an understanding of the PD relationship, if available. In the earliest stages of clinical testing, some guidance on the appropriate range may be obtained from preclinical work (e.g. from efficacy studies in a relevant animal model, or from theoretical considerations involving binding affinity and targeted receptor occupancy).

To use the drug in clinical practice, three fundamental questions must be answered: (i) *How* should the drug be administered? (ii) *How much* should be given? (iii) *How often* should it be given? In addition to these three basic questions, two auxiliary issues often need to be investigated prior to approval of a drug. These are: characterization of potential interactions with other drugs likely to be used in the target population, and identification of patient characteristics such as sex, weight, ethnic group, or renal function, which exert an effect on the kinetics of the drug substantial enough to warrant dose adjustment.

The main goal of PK modeling is to address these issues, using information obtained in clinical studies. Developing a *model* to describe the kinetic properties of a drug allows one to go beyond purely empirical conclusions, based on dosing regimens that have actually been studied, to predict outcomes for regimens that have not been tested. Evaluating the effects of subject characteristics, changes in physiology or disease state, or other drugs, on kinetic behavior is also naturally accomplished within a suitable model framework.

Design Aspects of Pharmacokinetic Studies

The fundamental data in any PK study are concentration measurements at timepoints following single or multiple doses of the drug. A prerequisite, therefore, is the availability of an accurate and reproducible assay to determine drug levels in any biological matrix of interest. This requirement is not trivial, but lies beyond the scope of this review, and we shall assume that a reliable assay is available. Conceptually, one might be interested in drug concentrations over time in any one of several target tissues. However, ethical and practical considerations limit the ability to obtain concentration measurements in human subjects, so that data are generally available only for blood (serum or plasma concentrations), and possibly urine. Even if biopsies can be obtained for other tissues, reliable determination of concentration is difficult as homogeneous distribution of the drug is unlikely for many tissues. To simplify the exposition, in what follows we consider drug or metabolite concentrations obtained in blood, as this covers most PK studies. We refer to "plasma concentrations", although in some cases drug levels may be assayed in serum, or in whole blood. When available, urine concentration data can provide further insight into elimination characteristics.

It is helpful to identify two broad categories of studies where PK data may be generated. In the first of these, characterization of the kinetic behavior is a primary goal. Such "classical" PK studies involve intensive sampling of the concentration–time profile in a relatively small number of subjects, often healthy volunteers. In contrast, most clinical studies do not have PK as the main focus, but some concentration measurements may be obtained during their conduct. In this situation, relatively few concentration measurements are available for a given subject, but the number of subjects is often considerably larger than that in a classical PK study, and information is obtained in patients, the target population of interest. Occasionally in clinical studies, more intensive sampling may also be carried out in a subgroup of subjects; this is quite common in Phase II trials, but not usual in Phase III.

Some Fundamental Pharmacokinetic Concepts

Studies carried out primarily to address pharmacokinetic issues are often referred to as ADME studies, since they are designed to characterize four fundamental aspects of a drug's kinetics: absorption, distribution, metabolism, and excretion.

Absorption refers broadly to the process by which the drug proceeds from the site of administration to the site of measurement within the body, typically the bloodstream. An understanding of absorption characteristics is important for drugs intended for extravascular routes of administration, e.g. oral agents, or drugs given by subcutaneous injection. Both the rate and extent of absorption are of practical interest. The rate may have implications for time to symptom relief, the extent provides information on the efficiency of the proposed route of administration.

Distribution deals with the question: Where does the drug go? Because of practical constraints in measuring drug levels in tissues, detailed information on

distribution is difficult to obtain directly in humans. Assessment of distribution is typically carried out most extensively in preclinical animal studies, using a radioactively labeled drug. The relevance of such preclinical results to humans is subject to the usual uncertainties surrounding interspecies extrapolation.

Metabolism refers to the conversion of one chemical species to another. It typically plays an important role in the elimination of drugs from the body, as very few drugs are excreted unchanged. The liver is a major site of drug metabolism, through microsomal enzymes which act upon the drug; metabolism in the gastrointestinal system is also common. In some cases it is of interest to characterize the kinetics of major metabolites as well as those of the original drug, particularly if these metabolites exhibit significant pharmacologic or toxic activity.

Excretion refers to the loss of unchanged drug from the body. The term *elimination*, used to describe any loss of drug, encompasses both metabolism and excretion. Elimination occurs primarily, though not exclusively, through the kidneys and the liver. Physiological or demographic factors, as well as the presence of other medications, can frequently affect elimination. It is important to understand which factors exert a substantial effect on elimination patterns to facilitate any dose adjustments that might be necessary as a result.

The key data obtained in any PK study are samples taken post-dosing from the concentration–time curve. Appropriate summaries calculated from this profile can provide information on specific ADME characteristics of the drug. Table 1 summarizes key PK quantities pertaining to absorption, distribution, and elimination that can be calculated from the concentration profile (*see* **Bioavailability and Bioequivalence**). Rates (e.g. of absorption or elimination) are usually defined in terms of *first-order kinetics*. This corresponds to an assumption that the rate of change of drug concentration at a particular time is

Table 1 Common pharmacokinetic parameters

Parameter	Explanation	Remarks
Area below the curve, AUC	Area below the concentration–time curve	Often taken as a measure of total exposure to the drug
Clearance, Cl	Volume of plasma cleared of drug per unit time	$Cl = \text{dose/AUC}$
Volume of distribution, V	The volume that would be occupied if the total amount of drug in the body were at the same concentration as that in the plasma	Gives an idea of the extent to which drug is distributed to tissues. If $V \gg$ plasma volume, then this indicates extensive distribution outside the plasma
Peak concentration, $C_{\max}$	Maximum achieved plasma concentration	Values of $C_{\max}$ that are "too high" may have implications for toxicity
$T_{\max}$	Time at which peak concentration occurs	Mainly relevant for extravascular administration; gives some information on the rate of absorption
Absorption rate, k_a	Fractional rate of drug absorption if first order kinetics are assumed	Relevant for extravascular routes of administration.
Elimination rate, k_{el}	Fractional rate of drug elimination from the body during the terminal phase	$k_{el} = \text{Cl}/V$
Terminal half-life, $t_{1/2}$	Time taken for the plasma concentration to fall by one half during the elimination phase	$t_{1/2} = \log 2/k_{el}$
Bioavailability, F	Exposure relative to that for the same dose given IV. Exposure is measured by the total AUC	$F = \text{AUC}_{\text{other route}}/\text{AUC}_{\text{IV}}$

proportional to the drug concentration at that time, which is approximately true in many cases. The concept of an elimination half-life that is independent of concentration also presupposes first-order elimination. *Zero-order kinetics*, reflecting a constant rate of change of drug concentration, sometimes obtain, for example, intravenous infusion at a constant rate, or in the absorption of certain drugs from the gastrointestinal tract.

The relationships between the quantities shown in the table are model-independent, unless stated otherwise. That is, their validity does not rest on any particular parametric model to describe the underlying concentration profile. Estimation of these quantities from experimental data is discussed below.

Compartmental Modeling

If intensive sampling of individual PK profiles has been conducted, significant progress can be made in estimating the PK parameters described in Table 1, without assuming a specific parametric functional form for a given subject's response profile. If the goal is to go beyond simple summarization of the data at hand, for instance to project PK behavior for other dosage regimens, then a more sophisticated model framework is usually invoked. Similarly, stronger assumptions about the form of the model describing individual subject profiles are generally needed to make progress on inference in the sparse data case. A suitable basis to support derivation of relevant parametric models is provided by the theory of compartmental modeling. Data analyses are classified as noncompartmental or compartmental, according to the underlying assumptions.

The basic idea in compartmental modeling is to represent the body as a system of compartments that communicate reversibly with each other. One should think of a "compartment" in this context not so much as a particular anatomic or physiologic region, but rather as a tissue or group of tissues with similar blood flow and drug affinity. For example, the liver and kidneys, being highly perfused organs, are often considered as being in the same compartment as the circulation. Compartments are assumed to be "well-mixed", assuring uniform distribution of drug throughout. On the basis of such a representation, a mathematical model can be derived to describe drug disposition over time in different compartments. Such models usually assume that drug transfer in and out of compartments follows *linear kinetics*, with transfer rates that are linear in concentration (i.e. zero or first order). This leads to a system of linear differential equations describing the transfer of drug between compartments. The associated PK rate constants are parameters that will generally need to be estimated from experimental data.

An illustration is the case of a two-compartment model to describe drug kinetics following a single intravenous (IV) bolus injection. This model is illustrated in the central part of Figure 1. Following essentially instantaneous absorption of the drug into the circulation, it is assumed to distribute into two compartments. The central compartment represents the blood, extracellular fluid and highly perfused organs and tissues. The second (peripheral) compartment may be thought of as other, poorly perfused, tissues. For the case shown in Figure 1, elimination is assumed to occur from the central (plasma) compartment only. The rate constants, k_{10}, k_{12}, and k_{21}, referred to as *microconstants* in the PK literature, govern the assumed first-order kinetic transfers into and out of the relevant compartments.

For this model, writing the differential equations to describe the system, and solving, shows that the function describing the concentration of drug in the plasma at time t after dosing may be written in

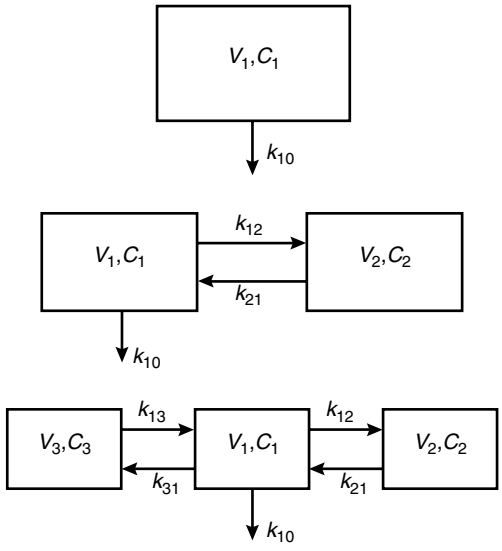

Figure 1 One-, two-, and three-compartment models to describe drug kinetics following a single intravenous bolus injection

biexponential form:

$$C(t) = A_1 \exp(-\lambda_1 t) + A_2 \exp(-\lambda_2 t). \quad (1)$$

Here the parameters λ_1 and λ_2 satisfy the relationships $\lambda_1 + \lambda_2 = k_{12} + k_{21} + k_{10}$ and $\lambda_1 \lambda_2 = k_{21} k_{10}$. The secondary parameters, A_1 and A_2, satisfy $A_1 = [D(\lambda_1 - k_{21})]/[V_P(\lambda_1 - \lambda_2)]$ and $A_2 = [D(k_{21} - \lambda_2)]/[V_P(\lambda_1 - \lambda_2)]$, where D and V_P represent the amount of drug administered and the volume of the plasma compartment, respectively. Solving the linear differential equations that arise in compartmental models, for which transfers follow linear kinetics, is most naturally accomplished using Laplace transforms. Details for the most common cases are given in Gibaldi & Perrier [6].

If the model above is a reasonable description of the system, then we would expect the concentration–time profile for a subject to decline in two distinct exponential phases. During the *distribution phase* there may be a relatively rapid disappearance of the drug from the plasma compartment as it distributes to the tissues. After equilibrium is reached between the plasma and tissue compartments, elimination exerts the dominant effect on plasma concentration, resulting in a shallower decline during the *(terminal) elimination phase*. This type of behavior is evident in Figure 2; these data, taken from a larger study, represent concentrations of indomethacin measured in four subjects following a single IV bolus dose. The constraint $\lambda_1 > \lambda_2$ is sometimes implemented to ensure **identifiability** of the model; nonnegativity of all four parameters in the model is required to yield sensible interpretations. The model is nonlinear in λ_1 and λ_2; nonlinearity in certain parameters is the norm for compartmental models. Following intravenous dosing, assuming linear kinetics, a general representation of the function describing plasma concentration is as a sum of exponential terms, the number of exponentials corresponding to the number of compartments.

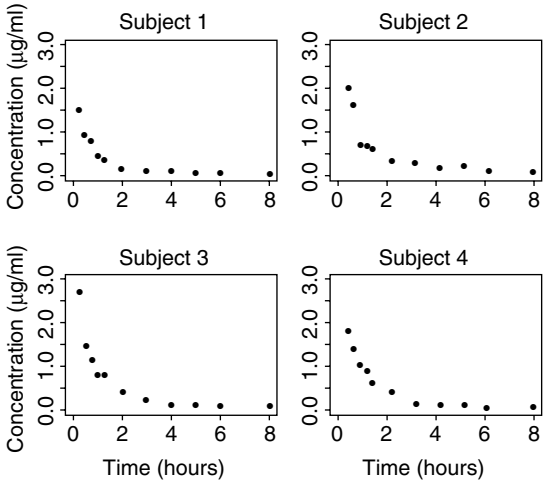

Figure 2 Concentrations of indomethacin in four subjects, following a single intravenous bolus injection

Use of the two-compartment model described above is particularly common. Modification to incorporate elimination from the tissue compartment is straightforward. If distribution to tissues is minimal, a single (plasma) compartment model may be adequate to describe the concentration profile. If distribution to tissues occurs very rapidly, and few sampling times fall in the distribution phase, then it is possible to miss the fact that two compartments are appropriate, as the observed concentration profile will likely appear to follow a monoexponential decline. Where the data indicate three distinct exponential decay phases, the usual interpretation is that tissues fall into two groups, one for which the movement of drug from the plasma to tissue occurs at a moderate rate, the second ("deep tissue compartment") consisting of very poorly perfused tissues such as bone and fat. The distribution of drug from the plasma compartment to these tissues is markedly slower.

Another important extension of the two-compartment model given above is to accommodate extravascular routes of administration such as oral dosing, or subcutaneous or intramuscular injection. In such cases the model is adjusted to include an absorption phase. If absorption occurs from the gastrointestinal tract, a lag time is sometimes included as well. A general method for adjusting the expected plasma concentration function following bolus IV dosing (instantaneous input) to account for a different input process is described by Gibaldi & Perrier[[6], Appendix B]. One multiplies the Laplace transform of the concentration function for IV dosing by that of the input function, and takes the inverse Laplace transform of the product. Application of this method to the case of a two-compartment model with first-order absorption into, and first order elimination from, the plasma compartment yields a function for the plasma concentration at time t that may be written

as the sum of three exponential terms:

$$C(t) = A_1 \exp(-\lambda_1 t) + A_2 \exp(-\lambda_2 t) + A_3 \exp(-k_a t), \quad (2)$$

where the six parameters in the model are known functions of the microconstants and other PK parameters. Inspection of this function suggests the possibility of identifiability problems, depending on the relative magnitudes of the absorption rate constant, k_a, and those associated with the distribution and elimination phases (λ_1 and λ_2, respectively). For most drugs, one expects that $k_a > \lambda_2$. For certain drugs with very rapid elimination, fitting techniques may sometimes interchange the two, a phenomenon referred to as "flip-flop" behavior. In general, k_a may be larger or smaller than λ_1, so that distinguishing between them cannot be done on any theoretical basis. A further difficulty arises if $k_a \approx \lambda_1$. In that case it is likely that the data support fitting only two exponential terms. This would correspond to a one-compartment model with first-order absorption, even though experience from other routes may suggest two compartments are needed. Thus, it is not always possible to determine k_a unambiguously from concentration data following oral administration only. Infrequent sampling during the absorption phase often compounds the problem. In theory, any ambiguity can be resolved by also characterizing kinetics following IV administration in the same subject; this may not always be practicable.

Nonlinear Kinetics

In many cases, linear kinetics obtain in only part of the concentration range. Differences in kinetic behavior may occur at higher concentrations, often due to some type of saturation phenomenon, e.g. of plasma protein or tissue binding, or of some capacity-limited metabolic process. Elimination kinetics are frequently affected. For example, suppose that clearance of the drug occurs primarily through metabolism in the liver, and that the activity of a particular enzyme involved acts as a rate-limiting step. Then the rate of metabolism may be linear over a portion of the concentration range, but reach a plateau at higher concentrations. Borrowing from standard modeling approaches in enzyme kinetics, this type of behavior is usually described by a rectangular hyperbola, with parametric form given by the **Michaelis–Menten equation:**

$$\frac{dC(t)}{dt} = \text{rate} = \frac{V_m C}{K_m + C}, \quad (3)$$

where V_m represents the maximal rate and K_m the concentration at which half the maximal rate is attained. Capacity-limited metabolism can also give rise to nonlinear absorption kinetics, e.g., in the case of drugs that undergo a significant degree of metabolism in the liver before being absorbed into the circulation (the so-called *first pass effect*). If one or more transfer rates in a compartmental model obey Michaelis–Menten kinetics, then the associated system of differential equations will be nonlinear. In this situation, an analytic solution for concentration functions is not possible, except in certain special cases. Saturable, or otherwise nonlinear, kinetics are to be expected at very high concentrations of a drug. Nonlinear kinetic behavior within the therapeutic concentration range is less frequent and its existence can complicate dosing considerably.

Multiple Dosing

Most drugs require multiple administration to maintain effect. If fixed doses are given, separated by a constant dosing interval, then peak plasma levels following later doses will usually be higher than that occurring after the first dose, a phenomenon known as *accumulation*. The degree of accumulation following successive doses attenuates until an eventual steady state is reached for the plasma profile. At the steady state, average input and output over the dosing interval are the same. It is straightforward to show that the average concentration at steady state is proportional to the half-life divided by the length of the dosing interval. The time to reach steady state is a function solely of the drug half-life; to attain 90% of steady-state concentration requires about 3.3 half-lives. Only in the case of intravenous infusion at a constant rate is the expected concentration at steady state actually constant; for an input function corresponding to discrete doses at regular intervals, a "sawtooth" pattern about some overall average steady-state level is expected. If fixed doses are given, at constant intervals, then the size and frequency of doses must be balanced to preclude excessive fluctuations in level between doses. More frequent dosing reduces the

degree of fluctuation between peak and trough levels, but must be balanced against practical aspects such as patient compliance.

If one makes the assumption that the kinetic behavior following a single dose is unchanged, regardless of how many other doses may already have been given, then it is straightforward in principle to predict plasma concentration following multiple doses. One applies the *principle of superposition*, which simply adds the expected contribution of the new dose to the concentration at any timepoint to the expected concentration at that time from all preceding doses. Proceeding recursively, the expected concentration is then simply a sum of terms, one per administered dose, of the concentration function evaluated at the time post-dosing for each of the individual doses. If doses are given at times t_k, and f is the function describing plasma concentration after a single dose, then

$$C(t) = \sum_{k:t_k<t} f(t - t_k). \tag{4}$$

It is important to recognize that the principle of superposition represents an *assumption*, which may or may not be true in practice. Quite often it applies until concentrations exceed the range where linearity of kinetics is preserved. Measurement of concentrations following multiple dosing is needed to provide an empirical check on the validity of models derived from the principle of superposition.

Further Reading

Several texts cover various aspects of PK in more depth. Gibaldi & Perrier [6] provide mathematical details of derivations for the most commonly used compartmental models. A more applied perspective is taken by Rowland & Tozer [13]. Other approaches that have been suggested for modeling PK data include the use of physiologically based models [1], and of stochastic compartmental models (e.g. [11]). A partial list of current and future issues in applying PK to clinical therapeutics includes: development of controlled release formulations, differential PK behavior of stereoisomers, use of stable isotopes in determining bioavailability, appropriate dosing of therapeutic proteins manufactured by recombinant DNA technology, and the challenges in delivery and dosing that accompany the promise of gene therapy.

Statistical Aspects

The development so far has focused on derivation of theoretically based models to describe the expected plasma concentration profile. In practice, concentration values sampled from a subject's profile will be subject to measurement error. Estimation of underlying model parameters is necessarily based on *observed* concentration data and involves the explicit or implicit assumption of a *statistical model*, characterizing the variability in the data. To be useful in practice, statistical modeling must also reflect the design and the analysis objectives accurately. These are generally sufficiently different for the "intensive" and sparse sampling cases that distinct modeling issues arise in the two situations. We now consider each case separately.

Individual PK Modeling and Inference

Objectives

For early **Phase I** studies, with intensive sampling of kinetic profiles, the primary objectives in analysis are usually: (i) to establish the correct functional form of the relevant PK model (e.g. one or two compartments) and (ii) to obtain a preliminary idea of typical values of the model parameters. A subsidiary objective may be characterization of the variability in concentration measurements about the postulated mean function. Intersubject variation in PK characteristics is also of some interest, but estimating this variability on the basis of a relatively small number of subjects who may not be representative of the target patient population clearly has the potential to mislead.

Noncompartmental Analysis

For data from "classical" studies, where relatively frequent samples are taken from the concentration profile, considerable progress can be made in estimating subject-specific PK parameters using noncompartmental methods. For instance, the area under the curve for a given subject can be estimated by application of the trapezoidal rule to the subject's observed profile. For curves exhibiting two distinct phases of exponential decay, a plot of log(concentration) against time should reveal two separate phases of approximately linear decline. Fitting a straight line

to later concentration measurements on this plot yields an estimate of the terminal rate of exponential decline. The terminal half-life may then be estimated by dividing log 2 by this slope estimate. Other parameters may be estimated by exploiting relationships such as those shown in Table 1. Quantifying the uncertainty associated with such noncompartmentally derived parameter estimates, e.g. by calculation of standard errors, is not straightforward, however. The absence of replication precludes calculation of any type of "pure error" estimate of within-subject variation to be used in calculating standard errors. Furthermore, although such methods do provide empirical summaries of the observed concentration profiles, they do not facilitate prediction of concentrations for other dosing regimens, without the assumption of a specific model form.

Compartmental Analysis

In estimating subject-specific parameters within a compartmental modeling framework, the key statistical techniques are those used in heteroscedastic nonlinear regression problems. Denote the jth concentration measurement for the ith subject, taken at time t_{ij}, by y_{ij}, $j = 1 \ldots m_i$, $i = 1 \ldots n$. For convenience, use the vector $\mathbf{x}_{ij}$ to summarize covariate information pertinent to the response y_{ij}; typically, this involves the measurement time, t_{ij}, and information on dosing history for the ith subject up to that time. Assume that

$$\mathrm{E}(y_{ij}|\boldsymbol{\beta}_i) = f(\mathbf{x}_{ij}, \boldsymbol{\beta}_i); \quad \mathrm{cov}(\mathbf{y}_i|\boldsymbol{\beta}_i) = \mathbf{R}_i. \qquad (5)$$

Here, f is a function whose form is known from the relevant compartmental model and which depends nonlinearly on $\boldsymbol{\beta}_i$, a $p \times 1$ vector of unknown PK parameters, specific to the ith subject. The within-subject **covariance matrix** $\mathbf{R}_i$, is assumed to have the same form for all subjects. We restrict attention here to the case where $\mathbf{R}_i$ is diagonal; **serial correlation** in the within-subject concentration measurements is certainly possible, but methods commonly used in PK modeling do not account for it. In contrast, heterogeneity of variance in concentration usually is accommodated when specifying the form of $\mathbf{R}_i$. Typically, a pattern of increasing variance is seen as the mean concentration increases, suggesting that variance might be modeled as a smooth function of the mean. The most common implementation of this idea in the PK literature assumes that the variance is proportional to some power of the mean, corresponding to diagonal elements of $\mathbf{R}_i$ of the form $\mathrm{var}y = \sigma^2\mu^{2\theta}$. The value of θ appearing in the exponent may be known, or may need to be estimated from the data.

Estimating the subject-specific parameters $\boldsymbol{\beta}_i$ from the ith subject's data is thus a **nonlinear regression** problem. Ordinary **least squares** fitting is one possibility, but the heterogeneity of variance in concentration measurements suggests adopting a weighted approach instead, with weights w_{ij} chosen to be inversely proportional to the response variance: $w_{ij} = \sigma^2/\mathrm{var}y_{ij}$.

If the weights are considered to be known constants, then the weighted least squares (WLS) estimate of $\boldsymbol{\beta}_i$ is the value that minimizes $\sum_{j=1}^{n_i} w_{ij}\{y_{ij} - f(\mathbf{x}_{ij}, \boldsymbol{\beta}_i)\}^2$. In practice, the correct choice of weights, w_{ij}, will not be known. However, a simple extension of the WLS approach is to use *estimated* weights in a scheme such as the following (assuming θ known). First obtain a preliminary unweighted estimator $\boldsymbol{\beta}_i^{\mathrm{p}}$, e.g. by ordinary least squares. On the basis of this estimate, form estimated weights $\hat{w}_{ij} = 1/(\hat{\mu}_{ij}^{2\theta})$, where $\hat{\mu}_{ij} = f(\mathbf{x}_{ij}, \boldsymbol{\beta}_i^{\mathrm{p}})$. Use these weights to reestimate $\boldsymbol{\beta}_i$ by weighted least squares. Update the weights on the basis of the new estimate of $\boldsymbol{\beta}_i$ and iterate until convergence. Denote the final estimate by $\hat{\boldsymbol{\beta}}_i^{\mathrm{GLS}}$. This method is known as generalized least squares (GLS), or sometimes as iteratively reweighted least squares (IRLS) (*see* **Generalized Linear Model**).

If θ is unknown, then it may be estimated from the data. An overview of methods for estimating variance parameters in the context of GLS regression may be found in Carroll & Ruppert [2]. An alternative method, commonly used by pharmacokineticists in fitting individual profile data, is to maximize the normal theory **likelihood** for the ith subject's data jointly in $\boldsymbol{\beta}_i$, θ, and σ; this method is referred to in the PK literature as *extended least squares* [12].

In practice, it is important to implement *some* type of weighting scheme in estimating the regression parameters, as failure to do so may give insufficient weight to lower concentration values, resulting in poor estimation of parameters pertaining to the terminal phase. Weights based on predicted rather than observed concentrations are preferable to avoid misbehavior due to one or two "bad" observed y values, and some iteration is desirable to wash out the effect

of poor initial weights. Small deviations from the "optimal" weighting scheme are unlikely to affect inference very much, so exact characterization of the value of θ is not critical. With this in mind, a less formal procedure is simply to examine plots of weighted Studentized residuals following GLS fits for a small grid of possible values of θ (e.g. 0, 0.5, and 1.0), and choose the value which most nearly corrects for the heterogeneity of response variance. Investigating the pattern of within-subject heteroscedasticity on the basis of **residuals** from a preliminary fit has the advantage that it is straightforward to pool the within-subject residuals across all subjects to obtain a single estimate of θ. This presupposes that a common value of θ is appropriate for all subjects, which is likely to be reasonable for PK data, provided assay procedures for determining concentrations are comparable.

In compartmental modeling, the parameters $\boldsymbol{\beta}_i$ appearing in the polyexponential functions usually employed in fitting plasma concentration data are not always the quantities of primary interest. Pharmacokinetic parameters that may be of more practical interest are generally expressible as some known (possibly nonlinear) function, h say, of the elements of $\boldsymbol{\beta}_i$. For example, the terminal half-life in the case of a two-compartment model following intravenous bolus dosing may be calculated as $\log 2/\beta_{4i}$, where β_{4i} is the exponential decay rate for the elimination phase. Estimation of such derived PK parameters is straightforward: one simply substitutes the estimate of $\boldsymbol{\beta}_i$ obtained by GLS, or other method of choice, into the appropriate expression. That is, an estimate of $h(\boldsymbol{\beta}_i)$ is given by $h(\hat{\boldsymbol{\beta}}_i)$.

Approximate expressions for estimating standard errors for the $\hat{\boldsymbol{\beta}}_i$, or functions of $\hat{\boldsymbol{\beta}}_i$, may be obtained by suitable application of standard asymptotic theory for weighted nonlinear regression. Details may be found in most nonlinear regression texts (e.g. [14]). In the context of parameter estimation based on data from only a single subject, where one may be attempting to estimate five or six parameters from only a dozen or so data points, the quality of such approximations may be quite poor.

Estimates of subject-specific PK parameters may be used for subsequent analyses in a number of ways. They may form the basis of prediction of achieved concentrations for different dosing regimens contemplated in future studies. If population modeling is of interest, then subject-specific estimates may be used as raw data for certain types of population analyses.

Software

Nonlinear regression routines with the flexibility to allow iterative reweighting are available in most major statistical packages (*see* **Software, Biostatistical**). These routines generally require explicit specification of the regression function, and possibly its first derivatives with respect to the parameters. More specialized packages, such as NONLIN or ADAPT, widely used by pharmacokineticists, allow the user to choose from a library of preexisting models. This obviates the need for explicit specification of the regression function and derivatives, which can be particularly helpful in situations where complex multiple dosing schemes have been used, or where explicit solution of the differential equations is not possible. Generally, these programs also allow iterative reweighting, with a few different options for the form of the dependence of the variance on the mean.

Practical questions which may arise, no matter what fitting routine is used, are the determination of suitable starting values, conventions for handling concentration values below the assay detection limit, possible model identifiability problems, and suitable choice of parameterization. In addition, convergence to sensible values, and the reliability of standard error estimates based on asymptotic approximations, always tricky issues in nonlinear regression, can be especially troublesome in fitting compartmental models.

Population PK Modeling and Inference

Objectives

For sparse concentration data obtained in the clinical setting, different statistical approaches are usually needed. In part, this follows from the design; obtaining individual estimates from each subject's data is generally not feasible. Analysis objectives also differ, with greater emphasis on characterizing differences in kinetic behavior among subjects. Population analysis thus has two primary goals: (i) estimation of subject-specific PK parameters, and (ii) understanding inter-subject variation in PK characteristics.

Population Modeling

Noncompartmental techniques are generally less useful in the population setting; the extra structure that

follows from a compartmental modeling approach is typically necessary to make much progress. Accommodation of two levels of random variation, within and between subjects, is also necessary. The relevant statistical techniques are those associated with mixed-effect models. A **two-stage** model framework proves useful; we first present this model in fairly general form, and then explain the interpretation of its different components. Let $\mathbf{x}_{ij}$, y_{ij}, $\mathbf{y}_i$, $\boldsymbol{\beta}_i$, $\mathbf{R}_i$, and f have the same interpretations as before, and consider the following model:
Stage 1:

$$\mathbf{y}_i = f(\mathbf{x}_i, \boldsymbol{\beta}_i) + \mathbf{e}_i,$$
$$\mathrm{E}(\mathbf{e}_i|\boldsymbol{\beta}_i) = \mathbf{0}; \qquad \mathrm{cov}(\mathbf{e}_i|\boldsymbol{\beta}_i) = \mathbf{R}_i. \tag{6}$$

Stage 2:

$$\boldsymbol{\beta}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{b}_i,$$
$$\mathrm{E}(\mathbf{b}_i) = \mathbf{0}; \qquad \mathrm{cov}(\mathbf{b}_i) = \mathbf{D}. \tag{7}$$

Interpretation of stage 1, which models data for the ith subject, $i = 1, \ldots, m$, is as before. The second stage models the intersubject variation in $\boldsymbol{\beta}_i$. In the first term, $\mathbf{A}_i\boldsymbol{\beta}$, the systematic component of the variation, is modeled in terms of a $k \times 1$ vector of fixed effects, $\boldsymbol{\beta}$, and a $p \times k$ matrix $\mathbf{A}_i$ incorporating subject-specific covariate information. The second term, $\mathbf{b}_i$, is a $p \times 1$ vector of random effects, reflecting residual (unexplained) variation in the PK parameters $\boldsymbol{\beta}_i$. In the simplest possible case, $\mathbf{A}_i = \mathbf{I}$, so that $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{b}_i$, corresponding to the situation where all of the variation in $\boldsymbol{\beta}_i$ is considered random, about an overall population average $\boldsymbol{\beta}$, with covariance matrix $\mathbf{D}$. If possible, one would like to identify subject-specific covariates that account for some of the variation in $\boldsymbol{\beta}_i$. This can be done by a suitable choice of the "design matrix" $\mathbf{A}_i$.

To fix ideas, suppose that the PK parameters $\boldsymbol{\beta}_i$ correspond to the bivariate vector of clearance and distribution volume for the ith subject, (Cl_i, V_i). We might wish to investigate the belief that clearance for a given subject depends on the subject's age. One possibility would be to formulate a model that postulates a linear dependence of clearance on age. Suppose none of the measured subject-level covariates plausibly affects the volume of distribution, so that all between-subject variation in that parameter is considered unexplained. Then, writing β_{1i} and β_{2i} for clearance and volume of the ith subject, respectively, this situation would be described by the following model:

$$\beta_{1i} = \beta_1 + \beta_3\mathrm{Age}_i + b_{1i},$$
$$\beta_{2i} = \beta_2 + b_{2i}. \tag{8}$$

Thus, at the second stage, one attempts to build a parametric regression model, where the subject-specific PK parameter is the response and the model attempts to separate the variation in this response into a systematic component, predictable by measured covariates, and a residual (random) component. As in any regression model, the residual variation will reflect both the true random variation in the population, as well as any variation attributable to covariates that have not been measured. In the case above the regression model is linear in the fixed-effect parameters, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$. For the example described above, $\mathbf{A}_i$ is the following 2×3 matrix:

$$\begin{bmatrix} 1 & 0 & \mathrm{Age}_i \\ 0 & 1 & 0 \end{bmatrix}.$$

Dependence on other covariates would be incorporated by adding columns to $\mathbf{A}_i$ and corresponding elements to the vector of fixed-effect parameters, $\boldsymbol{\beta}$, as necessary.

The need to accommodate both within-subject and intersubject variation correctly in population analysis, and the potential utility of mixed-effects modeling as a way of doing this, was first recognized by Sheiner et al. [15].

Inference

The two-stage model above is quite similar to the two-stage mixed-effects model, described by Laird & Ware [8], used to derive inferential methods for *linear* repeated measurement data. This suggests adapting inferential methods for the linear case for use in fitting population PK models. This approach forms the basis of most commonly used methods for population PK modeling. In the linear case, normality of $\mathbf{e}_i$ and $\mathbf{b}_i$, and linearity of the model in $\mathbf{b}_i$, allow the marginal distribution of $\mathbf{y}_i$ to be derived explicitly. **Multivariate normality** of this marginal distribution allows application of standard **maximum likelihood** or **restricted maximum likelihood** methods.

In contrast, nonlinear dependence of f on $\boldsymbol{\beta}_i$, and consequently on $\mathbf{b}_i$, makes exact derivation of the

marginal impossible, as the necessary integration over the random-effects distribution cannot be performed analytically. Dependence of the covariance matrix $\mathbf{R}_i$ on $\mathbf{b}_i$ complicates matters further. Most inferential methods for PK modeling of sparse data within the framework above rely on approximations to the marginal likelihood. Specifically, a Taylor series expansion in $\mathbf{b}_i$ is used to provide an approximation that is linear in $\mathbf{b}_i$, yielding an approximation to the marginal likelihood. One can then apply methods for linear mixed-effects models to carry out approximate inference on the parameters $\boldsymbol{\beta}$ and $\mathbf{b}_i$, and the independently varying components of the matrices $\mathbf{R}_i$ and $\mathbf{D}$.

Most of the commonly used methods are based on some variation of one of two approximations. The first of these involves a Taylor series expansion in $\mathbf{b}_i$ about zero, the mean of the random effects distribution, giving a first-order approximation to the likelihood. This approximation was first suggested by Sheiner et al. [15], and is referred to as the *first-order* (FO) method in the PK literature. Lindstrom & Bates [9] suggested a refinement, wherein accuracy of the approximation may be improved by expanding about the current estimate of the subject-specific random effect in an iterative fashion. A variation of their method has been adopted by pharmacokineticists; the resulting inferential techniques being referred to as *first-order conditional estimation* (FOCE). At the time of writing, these methods are implemented in two software packages: NONMEM, a package widely used by pharmacokineticists, developed in the early 1980s, and a more recent suite of contributed functions to **S-PLUS**, `nlme`, which implement the method of Lindstrom & Bates.

Application of these methods allows inference on the fixed-effects and the independently varying covariance components. In addition, empirical Bayes estimates of the individual random effects $\mathbf{b}_i$ (and thus of the subject-specific PK parameters $\boldsymbol{\beta}_i$) are obtained. By analogy with the linear mixed-effects model, the random-effect estimates may be viewed as subject-level residuals, and thus prove useful in screening covariates for inclusion in the second-stage model to account for intersubject variation. Formal testing of the need to include particular covariates may be implemented using **likelihood ratio test** techniques.

Other inferential methods have been proposed in the recent statistical literature. Each assumes some type of **hierarchical model** framework; most attention has focused on relaxing assumptions pertaining to the second-stage (random-effects) distribution. Among the proposed approaches are completely nonparametric estimation of the random-effects distribution [10], and a semiparametric alternative, developed by Davidian & Gallant [3]. Recent advances in using **Markov chain Monte Carlo** (MCMC) techniques to deal with computational complexities have allowed practical implementation of Bayesian analysis in the context of hierarchical regression models. Wakefield [17] illustrates application of these methods in the context of population PK modeling. Two recent texts that review modeling and inference in nonlinear hierarchical regression are Davidian & Giltinan [4], and Vonesh & Chinchilli [16].

PD Modeling

Objectives

The primary goal of PD modeling is to characterize the concentration–effect relationship for some appropriate measure of response (*see* **Dose-response in Pharmacoepidemiology**). This can be used to deduce the appropriate target concentration range to guide dosing. Just as subjects may exhibit different PK behavior resulting from genetic or physiological characteristics, their response to circulating concentrations may also differ. Substantial differences might result in a need for subject-specific target concentration ranges – particularly if the separation between toxicity and efficacy response curves is narrow. In such cases an understanding of the factors giving rise to differential sensitivity to the drug may be important in order to allow accurate prediction of the correct target concentration range for a new patient.

Design Aspects

Data suitable for PD modeling are most often obtained in a setting where PK (concentration) data are also being collected. The goal is to characterize the relationship between response and concentration, so that concentration plays the role of "independent variable" in PD modeling. Sometimes, concentration values are obtained concurrently with the measurement of the PD response, but sampling schedules

for PK and PD measurements may differ. Conceptually, this can be accommodated by incorporating an appropriate lag term in the model, or by using predicted concentration values, from subject-specific PK parameter estimates, at the PD measurement timepoints. Some studies employ a dose–titration design, in which each subject receives multiple dose levels of the drug, separated by a washout period. More commonly, each subject is dosed at only one of the possible levels, although multiple dosing at the particular level assigned may occur. In this case, concentration determinations may be made at several timepoints post-dosing, ensuring some variation in the achieved concentrations, but any one subject is unlikely to achieve concentration levels that span the entire range of interest. Thus, unlike PK modeling, it is rarely possible to characterize the full concentration–effect profile for a particular subject on the basis of that subject's data only. As a result, population-based fitting methods are usual in PD modeling.

Model Derivation

In comparison to PK modeling, both theoretical and practical aspects of PD modeling are less well developed. This is not surprising; given the potential variety in the choice of PD endpoints, development of a unified theory for modeling may be too much to expect. Potential issues in PD modeling include (i) choice of pharmacologic endpoint – factors to be considered in making this choice include clinical relevance, ease and variability of measurement, and sensitivity to changes in drug concentration; (ii) possible lag between changes in plasma concentration and PD effect, a phenomenon known as *hysteresis*; (iii) error in concentration measurements; (iv) the possibility of diurnal or other temporal variation in response; (v) feedback loops resulting in up- or downregulation of response; and (vi) possible formation of metabolites affecting response. This list is not exhaustive.

The most commonly used model relating PD response to concentration is the so-called sigmoid $E_{\max}$ model, proposed by Holford & Sheiner [7]. There are several possible parameterizations; the following is common:

$$y = E_0 + \frac{E_{\max} - E_0}{1 + (\mathrm{EC}_{50}/C)^{\gamma}}. \tag{9}$$

This model is also referred to as the four-parameter logistic function. The parameters have the following interpretation. E_0 and $E_{\max}$ are the response at zero concentration and the maximal response, respectively. EC_{50} is the concentration eliciting a response halfway between E_0 and $E_{\max}$, and γ (the *Hill coefficient*) is a parameter governing the steepness of the concentration–effect relationship.

In some situations the biochemistry or physiology of response is such that a definite lag occurs between a change in the circulating concentration of the drug and a change in response. One possible view is that it is the drug concentration at the site where it elicits its effect (e.g. by binding to receptors in the target tissue) that is important, more than plasma concentrations. Measurement of drug levels at the effect site is generally infeasible, of course, so plasma concentrations act as a surrogate. The lag may be interpreted as the time it takes for changes in plasma concentrations to be propagated to the effect site. Adopting this view, one approach to modeling the lag in response is to incorporate a hypothetical "effect-site compartment", with a term that describes the absorption process required for the drug to reach this compartment (see, for example, [7]). Recent work by Jusko and colleagues [5] develops a more general approach, resulting in a family of *indirect pharmacodynamic models*. The flexibility of the indirect modeling approach is appealing; given its relatively recent introduction, there have been few reported applications in the literature to date.

Inference

The data structure in PD modeling problems usually parallels that in population PK analysis: relatively infrequent repeated measurements on each of a number of subjects, according to an unbalanced design. A similar two-stage model is appropriate, with the first stage modeling concentration–response data for each subject, and the second stage describing intersubject variation in parameters of the first-stage PD model. Since this model is nonlinear in certain parameters, inferential methods for nonlinear mixed effects models are appropriate for population PD modeling, as well as for population PK modeling. Either sequential or simultaneous fitting of population PK and PD models is possible. Details are beyond the scope of this article; a more extensive discussion may be found in Davidian & Giltinan [4].

In contrast to PK modeling, there is no requirement that the response variable in a PD model be continuous. For discrete or binary responses, existing

methods for population PK analysis will generally not be appropriate. The natural model framework in these cases is that of generalized mixed models; see, for example, [16]. This is an area of considerable interest in the current statistical literature. Given the interest in characterizing intersubject variation in PD parameters, subject-specific modeling is more suitable than methods that focus on marginal inference. **Bayesian** hierarchic modeling, implemented by appropriate **Markov chain Monte Carlo** (MCMC) methods, appears particularly promising in this context.

Conclusion

Pharmacokinetic and pharmacodynamic modeling address an important practical need during drug development, namely that of learning how to dose a drug sensibly to achieve therapeutic benefit. Several factors can make this a challenging problem: a steep concentration–effect relationship, poor separation of the dose–response curves for efficacy and toxicity, and high intersubject variability in PK or PD characteristics. Drugs that exhibit good separation of the efficacy and toxicity curves are generally easy to use in practice – even in the presence of high intersubject variation it may be possible to find a single dosing regimen that is suitable for all patients. If the therapeutic window is not so wide, then some individualization of dosing may be required, depending on the degree of intersubject variation. Ideally, simple adjustments to dosing based on easily measured subject characteristics, such as age, sex, or weight, are adequate. For drugs where simple adjustments prove inadequate, a possible recourse is to implement some degree of therapeutic concentration monitoring initially. With this approach a patient might be started out on a dosage regimen based on "typical" kinetic behavior. One or two blood draws are taken and assayed for drug concentration, thereby allowing an assessment of how different the subject's concentration profile is from the norm. Beliefs about the values of the subject's kinetic parameters may be updated in a Bayesian fashion, and possible dose corrections may be implemented. Since therapeutic concentration monitoring is somewhat resource-intensive, its use is generally limited to drugs with a narrow therapeutic window, and a high degree of unexplained intersubject variability in kinetics, such as theophylline, digoxin, or phenytoin. At the most challenging end of the dosing spectrum are drugs with poor separation of efficacy and toxicity, a high degree of intersubject PK and PD variability, and with few covariates of predictive value. If no measure of a patient's sensitivity is available to guide a sensible choice of target concentration range, it is sometimes possible to titrate dose to a specific effect level. This requires the availability of an easily measured PD endpoint upon which dose titration can be based. In many situations it may not be practicable to dose a drug in this manner, though in some cases this approach can be successful. The anticoagulating agent coumadin, frequently titrated for an individual subject according to the coagulation parameter prothrombin time, represents an example of a drug dosed in this fashion. The distilled grain extract Stolichnaya is another.

References

[1] Bischoff, K.B. & Dedrick, R.L. (1968). Thiopental pharmacokinetics, *Journal of Pharmaceutical Sciences* **57**, 1347–1357.

[2] Carroll, R.J. & Ruppert, D. (1988). *Transformations and Weighting in Regression*. Chapman & Hall, New York.

[3] Davidian, M. & Gallant, A.R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine, *Journal of Pharmacokinetics and Biopharmaceutics* **20**, 529–556.

[4] Davidian, M. & Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London.

[5] Dayneka, N.L., Garg, V. & Jusko, W.J. (1993). Comparison of four basic models of indirect pharmacodynamic responses, *Journal of Pharmacokinetics and Biopharmaceutics* **21**, 457–478.

[6] Gibaldi, M. & Perrier, D. (1982). *Pharmacokinetics*. Marcel Dekker, New York.

[7] Holford, N.H.G. & Sheiner, L.B. (1981). Understanding the dose-effect relationship: clinical application of pharmacokinetic-pharmacodynamic models, *Clinical Pharmacokinetics* **6**, 429–453.

[8] Laird, N.M. & Ware, J.H. (1982). Random effects models for longitudinal data, *Biometrics* **38**, 963–974.

[9] Lindstrom, M.J. & Bates, D.M. (1990). Nonlinear mixed effects models for repeated measurement data, *Biometrics* **46**, 673–687.

[10] Mallet, A. (1986). A maximum likelihood estimation method for random coefficient regression models, *Biometrika* **73**, 645–656.

[11] Matis, J.H. & Wehrly, T.E. (1979). Stochastic models of compartmental systems, *Biometrics* **35**, 199–220.

[12] Peck, C.C., Beal, S.L., Sheiner, L.B., & Nichols, A.I. (1984). Extended least squares regression: a possible

solution to the choice of weights problem in analysis of individual pharmacokinetic data, *Journal of Pharmacokinetics and Biopharmaceutics* **12**, 545–558.

[13] Rowland, M. & Tozer, T.N. (1995). *Clinical Pharmacokinetics: Concepts and Applications*. Williams & Wilkins, Baltimore.

[14] Seber, G.A.F. & Wild, C.J. (1989). *Nonlinear Regression*. Wiley, New York.

[15] Sheiner, L.B., Rosenberg, B. & Marathe, V.V. (1977). Estimation of population characteristics of pharmacokinetic parameters from routine clinical data, *Journal of Pharmacokinetics and Biopharmaceutics* **5**, 635–651.

[16] Vonesh, E. & Chinchilli, V. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker, New York.

[17] Wakefield, J. (1996). The Bayesian analysis of population pharmacokinetic models, *Journal of the American Statistical Association* **91**, 62–75.

(*See also* **Multilevel Models**)

DAVID GILTINAN

Phase I Trials

The initial phase of human experimentation in the development of chemotherapeutic drugs involves finding a dose that produces an acceptable level of toxicity. What is acceptable obviously depends on the disease in question: in diseases with significant morbidity and mortality, particularly cancer, the acceptable level of toxicity is quite high. Indeed, the toxicity of the drug may be to some extent a measure of its potential efficacy. That any toxicity is acceptable is an acknowledgment that, in the absence of any information regarding the optimal therapeutic dose of the drug, the maximum benefit is presumed to occur at the highest possible dose.

The Phase I trial has, as a general objective, the determination of the maximum dose of a drug, either alone or as part of a combination, that will, when administered by a specific schedule and route, produce an acceptable level of toxicity. This dose is usually referred to as the maximum tolerable dose (MTD), although in animal studies the same acronym may refer to the largest dose that will not kill any animals. A typical definition of acceptable toxicity (for an anticancer agent) might be, for example, "toxicity of grade 3 or worse in not more than one out of three patients", where grade 3 toxicity is further defined according to criteria that are as objective as possible. An occurrence of toxicity that is unacceptable is said to be a dose-limiting toxicity (DLT). The MTD is most frequently defined in terms of the frequency of occurrence of the **binary** outcome associated with the presence/absence of DLT. Mathematically, if Y denotes the binary response with $Y = 1$ denoting the occurrence of DLT, then one seeks a dose x_{MTD} where $\Pr(Y = 1|x_{\mathrm{MTD}}) = q_0$, where q_0 is usually in the range $0.1-0.4$ (although this is often not explicitly defined). However, in cases where toxicity is almost always restricted to a single quantitative parameter; for example, white blood count (WBC), it would be possible to define the MTD in terms of the WBC itself; for example, the highest dose where the median nadir WBC is at least 2000 (see the section on continuous outcomes below). One must also define the time period over which the evaluation of toxicity is to be performed. The MTD is typically defined only with respect to relatively acute toxicities, such as those occurring within three to four weeks of drug administration.

Once an MTD is established, one presumes that this dose will be used in further evaluations of efficacy in **Phase II trials**; however, this logical progression is complicated by the fact that patient populations in Phase I and II trials are likely to be dissimilar.

Traditional Methods

When defined in terms of the presence or absence of DLT, the MTD can be defined as some **quantile** of a tolerance distribution or **dose–response** curve. For a given sample size, the most effective way of estimating this quantile would be to determine, for each patient in the sample, the exact dose of the agent in question at which DLT first appears. Such data are nearly impossible to gather, however, as it is impractical to give each patient more than a small number of discrete dosages. Furthermore, the data obtained from sequential administration of different doses to the same patient would almost surely be **biased**, as one could never distinguish the cumulative effects of the different doses from the acute effects of the current dose level. Extended washout periods (*see* **Crossover Designs**) between doses are not a solution, since the condition of the patient, and hence the response to the drug, is likely to change rapidly for the typical patient in a Phase I trial. For this reason, almost all Phase I trials involve the administration of only a single dose level to each patient and the observation of the frequency of occurrence of DLT in all patients treated at the same dose level.

There are two significant constraints on Phase I trial design. The first is the ethical requirement to approach the MTD from below, so that one must start at a dose level believed almost certainly to be below the MTD, and gradually escalate upward. The second is the fact that the number of patients typically available for a Phase I trial is relatively small, say $15-30$, and is not driven traditionally by rigorous statistical considerations requiring a specified degree of precision in the estimate of MTD. The pressure to use only small numbers of patients is large – literally hundreds of drugs per year need to be tested, and each combination with other drugs, each schedule, and each route of administration

requires a separate trial. Furthermore, the number of patients for whom it is considered ethically justified to participate in a trial with little evidence of efficacy is limited (*see* **Ethics of Randomized Trials**).

As a consequence of the above considerations, the traditional Phase I trial design utilizes a set of fixed dose levels that have been specified in advance; that is, $x \in \{a_1, a_2, \ldots, a_K\}$. The choice of the initial dose level a_1, and the dose spacing, are discussed in more detail below. Beginning at the first dose level, small numbers of patients are entered, typically three to six, and the decision to escalate or not depends on a prespecified **algorithm** related to the occurrence of DLT. When a dose level is reached with unacceptable toxicity, then the trial is stopped.

Initial Dose Level and Dose Spacing

The initial dose level is generally derived either from animal experiments if the agent in question is completely novel, or by conservative consideration of previous human experience if the agent in question has been used before but with a different schedule, route of administration, or with other concomitant drugs. A common starting point based on the former is from 1/10 to 1/3 of the mouse LD_{10}, the dose that kills 10% of mice, adjusted for the size of the animal on a per kilogram basis or by some other method.

Subsequent dose levels are determined by increasing the preceding dose level by decreasing fractions, a typical sequence being $\{a_1, a_2 = 2a_1, a_3 = 1.67a_2, a_4 = 1.5a_3, a_5 = 1.4a_4$, and thereafter $a_{k+1} = 1.33a_k, \ldots\}$ Such sequences are often referred to as "modified Fibonacci". (In a "true" Fibonacci sequence the increments are approximately 2, 1.5, 1.67, 1.60, 1.63, and then 1.62 thereafter, converging on the golden ratio.) With some agents, particularly biological agents, the dose levels may be determined by log (i.e. $\{a_1, a_2 = 10a_1, a_3 = 100a_1, a_4 = 1000a_1, \ldots\}$) or half-log (i.e. $\{a_1, a_2 = 3a_1, a_3 = 10a_1, a_4 = 30a_1, a_5 = 100a_1, \ldots\}$) spacing.

Escalation Algorithms

A wide variety of dose escalation rules may be used. For the purposes of illustration, we describe the following:

1. Evaluate three patients at a_k:
   - (i) if zero of three has DLT, then increase dose to a_{k+1} and go to step 1;
   - (ii) if one of three has DLT, then go to step 2;
   - (iii) if more than one of three have DLT, then go to step 3.
2. Evaluate an additional three patients at a_k:
   - (i) if one of six has DLT, then increase dose to a_{k+1} and go to step 1;
   - (ii) if more than one of six have DLT, then go to step 3.
3. Discontinue dose escalation.

If the trial is stopped, then the dose level below that at which excessive DLT was observed is the MTD. Some protocols may specify that if only three patients were evaluated at that dose level, then an additional three should be entered, for a total of six, and that process should proceed downward, if necessary, so that the MTD becomes the highest dose level where no more than one toxicity is observed in six patients. The actual q_0 that is desired is generally not defined when such algorithms are used, but clearly $0.17 \leq q_0 \leq 0.33$, so we could take $q_0 \approx 0.25$.

While such an algorithm makes common sense, only brief reflection is needed to see that the determination of MTD on rigorous statistical grounds is extremely tenuous. Consider a trial where the frequency of DLT for three consecutive dose levels following the algorithm above is, respectively, zero of three, one of six, and two of six. Ignoring at this point the sequential nature of the escalation procedure, the pointwise 80% **confidence intervals** for the rate of DLT at the three dose levels are, respectively, 0 – 0.54, 0.02 – 0.51, 0.09 – 0.67. Although in such an outcome the middle dose would be taken as the estimated MTD, there is not even reasonably precise evidence that the toxicity rate for *any* of the three doses is either above or below the implied target rate of approximately 0.25.

A more rigorous statistical analysis of this problem would produce simultaneous confidence intervals for the three dose levels, account for the sequential sampling algorithm (*see* **Sequential Analysis**), and also account for the fact that the toxicity rates are presumably nondecreasing with increasing dose. Such an analysis is quite complex for even such a simple example, and would not alter the basic point regarding the lack of precision about toxicity rates. Furthermore, even the latter analysis does not

quantify the imprecision in the estimate of MTD itself.

Crude comparisons among different dose escalation algorithms can be made by examining the level-wise operating characteristics of the design, i.e. the probability of escalating to the next dose level given an assumption regarding the underlying probability of DLT at the current dose level. Usually, this calculation is a function of simple **binomial** success probabilities. For example, in the algorithm described above, the probability of escalation is $B(0, 3; p_x) + B(1, 3; p_x)B(0, 3; p_x)$, where $B(r, n; p_x)$ is the probability of r successes (toxicities) out of n trials (patients) with underlying success probability at the current dose level p_x. When the probability of escalation is plotted over a range of p_x, one can characterize algorithms as relatively "aggressive" or "conservative".

More useful comparisons need to involve consideration of the entire dose–response curve, which, of course, is unknown. Many features of traditional dose escalation algorithms can be studied by formulating the design as a discrete **Markov chain** [11]. The states in the chain refer to treatment of a patient or group of patients at a dose level, with an absorbing state corresponding to the stopping of the trial. If the design can be formulated such that one has constant transition probabilities for various assumptions about the true dose–response curve, then much information, such as the number of patients treated at each dose level, can be calculated exactly by carefully computing the appropriate quantities summarized from successive powers of the transition probability matrix **P**. Usually, however, **simulation** studies are a more practical tool for this purpose. As with exact computations, one needs to specify a range of possible dose–response scenarios, and then simulate the outcome of a large number of trials under each scenario. Many different aspects of the algorithm can be compared, such as the mean and variability of x_{MTD} (or, more usefully, the probability of DLT at x_{MTD}), the average number of patients treated, and the percentage of patients treated at doses where p_x is either undesirably small or large. In particular, one can study the sensitivity of the design to features of the dose–response curve that will be unknown in practice, such as the number of dose levels between the starting dose and the true MTD, and the steepness of the dose–response curve near the MTD.

Alternative Approaches

A Bayesian Approach: The Continual Reassessment Method

The small sample size and low information content in the data derived from traditional methods have suggested to some the usefulness of Bayesian approaches (*see* **Bayesian Methods**) to estimate the MTD. In principle, this approach allows one to combine any prior information available regarding the value of the MTD with subsequent data collected in the Phase I trial to obtain an updated estimate reflecting both.

The most clearly developed Bayesian approach to Phase I design is the continual reassessment method (CRM) proposed by O'Quigley and colleagues [7, 9] From among a small set of possible dose levels, say $\{a_1, \ldots, a_6\}$, experimentation begins at the dose level that the investigators believe, based on all available information, is the most likely to have an associated probability of DLT equal to the desired q_0. It is assumed that there is a simple family of monotone dose–response functions ψ such that for any "dose" x and desired probability of toxicity q there exists a unique θ where $\psi(x, \theta) = q$ and, in particular, $\psi(x_{\mathrm{MTD}}, \theta_0) = q_0$. An example of such a function is $\psi(x, \theta) = [(\tanh x + 1)/2]^{\theta}$. Note that ψ is not assumed to be necessarily a dose–response function relating a characteristic of the dose levels to the probability of toxicity. That is, x does not need to correspond literally to the dose of the drug. The uniqueness constraint implies, in general, the use of single-parameter models, and explicitly eliminates popular two-parameter dose–response models like the logistic (*see* **Quantal Response Models**).

A **prior distribution** $g(\theta)$ is assumed for the parameter θ such that for the initial dose level; for example, a_3, either $\int_0^{\infty} \psi(a_3, \theta) g(\theta)\, \mathrm{d}\theta = q_0$ or, alternatively, $\psi(a_3, \mu_a) = q_0$, where $\mu_a = \int_0^{\infty} \theta g(\theta)\, \mathrm{d}\theta$. The particular prior used should also reflect the degree of uncertainty present regarding the probability of toxicity at the starting dose level; in general, this will be quite vague.

After each patient is treated and the presence or absence of toxicity observed, the current distribution $g(\theta)$ is updated along with the estimated probabilities of toxicity at each dose level, calculated by either of the methods above. The next patient is then treated at the dose level minimizing some measure of the distance between the current estimate of the

probability of toxicity and q_0. After a fixed number, n, of patients have been entered sequentially in this fashion, the dose level selected as the MTD is the one that would be chosen for a hypothetical $n + 1$st patient. Confidence intervals for the probability of toxicity at the selected dose level are available [6], though not for x_{MTD} itself, although some consistency results are available [10].

Although the prior distribution $g(\theta)$ can be chosen to be extremely vague, some practitioners object philosophically to the Bayesian approach, and it is clear in the Phase I setting that the choice of prior can have a measurable effect on the estimate of MTD [4]. However, the basic framework of CRM can be adapted easily to a non-Bayesian setting and can conform in practice more closely to traditional methods [10]. For example, there is nothing in the approach that prohibits one from starting at the same low initial dose as would be common in traditional trials, or from updating after groups of three patients rather than single patients. Allowing for some ad hoc deterministic rules to start the trial off, the Bayesian prior can be abandoned entirely and the updating after each patient can be fully **likelihood** based [8].

Storer's Two-stage Design

Storer [11, 12] has explored a combination of more traditional methods implemented in such a way as to minimize the numbers of patients treated at low dose levels and to focus sampling around the MTD; these methods also utilize an explicit dose–response framework to estimate the MTD.

The design has two stages and uses a combination of simple dose-escalation algorithms. The first stage assigns single patients at each dose level, and escalates upward until the first toxicity is reached; then, the dose is decreased one level and the second stage begun. The second stage incorporates a fixed number of cohorts of three patients; successive cohorts are entered at the next lower, same, or next higher dose level, respectively, according to whether the current cohort experiences two or more, one, or zero dose-limiting toxicities. After completion of the second stage a dose–response model (*see* **Dose-response in Pharmacoepidemiology**; **Dose–Response Models in Risk Analysis**) is fit to the data and the MTD estimated by **maximum likelihood**. For example, one could use a **logistic** model where logit $\Pr[Y = 1|x] = \alpha + \beta x$, whence the estimated MTD is $x_{\mathrm{MTD}} = (\mathrm{logit} q_0 - \hat{\alpha})/\hat{\beta}$.

The particular second-stage algorithm described above is designed with a target percentile of $q_0 = 0.33$ in mind. Although other percentiles could be estimated from the same estimated dose–response curve, a target q_0 different from 0.33 would probably lead one to use a modified second-stage algorithm.

Extensive simulation experiments using this trial design in comparison with a more traditional design demonstrated the possibility of reducing the variability of point estimates of the MTD, and reducing the proportion of patients treated at very low dose levels, without markedly increasing the proportion of patients treated at dose levels where the probability of DLT is excessive; say, 0.5. Storer [12] also evaluated different methods of providing confidence intervals for the MTD. Standard likelihood-based methods that ignore the sequential sampling scheme (a **delta method**, a method based on **Fieller's theorem**, and a **likelihood ratio** method) are often markedly anticonservative. More accurate confidence sets can be constructed by simulating the distribution of any of those test statistics at trial values of the MTD; however, the resulting confidence intervals are often extremely wide. Furthermore, the methodology is purely frequentist, and may be unable to account for minor variations in the implementation of the design when a trial is conducted.

With some practical modifications, the two-stage design described above has been implemented in a real Phase I trial [2]. The major modifications included: (i) a provision to add additional cohorts of three patients, if necessary, until the estimate of β in the fitted logistic model becomes positive and finite; (ii) a provision that the recommended Phase II dose is not higher than the highest dose level at which patients have actually been treated, in the event that the estimate of x_{MTD} is above that dose; and (iii) a provision to add additional intermediate dose levels if, in the judgment of the protocol chair, the nature or frequency of toxicity at a dose level precludes further patient accrual at that dose level.

Continuous Outcomes

Although not common in practice, it is useful to consider the case where the major outcome defining toxicity is a continuous measurement; for example, the nadir WBC. This may or may not involve

a fundamentally different definition of the MTD in terms of the occurrence of DLT. For example, suppose that DLT is determined by $Y < c$, where c is a constant, and we have $Y \sim \mathrm{N}(\alpha + \beta x, \sigma^2)$. Then $x_{\mathrm{MTD}} = [c - \alpha - \Phi^{-1}(q_0)\sigma]/\beta$ has the traditional definition that the probability of DLT is q_0. The use of such a model in studies with small sample size makes some distributional assumption imperative. Some sequential design strategies (*see* **Sequential Analysis**) in this context have been described by Eichhorn & Zacks [3].

Alternatively, the MTD can be defined in terms of the mean response, i.e. the dose where $\mathrm{E}(Y) = c$. For the same simple linear model above, we then have that $x_{\mathrm{MTD}} = (c - \alpha)/\beta$. Fewer distributional assumptions are needed to estimate x_{MTD}, and **stochastic approximation** techniques might be applied in the design of trials with such an endpoint [1]. Nevertheless, the use of a mean response to define MTD is not generalizable across drugs with different or multiple toxicities, and, consequently, has received little attention in practice.

A recent proposal for a design incorporating a continuous outcome is that of Mick & Ratain [5]. This is also a two-stage study, which for a hypothetical study of etoposide assumes a simple **linear regression** model relating dose to the WBC nadir $\ln(\mathrm{WBC}) = \alpha + \beta_1 \ln(\mathrm{WBC_{pre}}) + \beta_2 x$, where $\mathrm{WBC_{pre}}$ is the pretreatment WBC. The first phase uses cohorts of two patients. Ad hoc rules for dose escalation are determined by the toxicity experience in the current cohort; however, the model is fit each time and cohorts of two are added until at least eight patients have been treated and β_2 is significantly different from 0 at the 0.05 level of significance. In the second stage of the study, the dose for the next cohort of two patients is determined by fitting the regression model to the accumulated data and estimating the dose that leads to a mean nadir WBC of 2.5; that is, $a_{k+1} = [\ln(2.5) - \hat{\alpha} - \hat{\beta}_1 - \ln(\mathrm{WBC_{pre}})]/\hat{\beta}_1$. This continues until at least eight patients have been treated and β_2 is significantly different from 0 at the 0.001 level of significance.

Simulation studies of this design using a **pharmacokinetic model** and historic database demonstrated a clear increase in precision in the MTD estimated from the model-based dose escalation method, as compared with the MTD estimated from a more traditional design. The average sample size was also measurably smaller. Though such results are promising, the method applies only to situations where the dose limiting toxicity is a single continuous outcome. Furthermore, the simulation studies that are needed to establish the usefulness of the method in specific situations often require the use of human pharmacokinetic data that may not yet be available.

References

[1] Anbar, D. (1984). Stochastic approximation methods and their use in bioassay and Phase I clinical trials, *Communications in Statistics – Theory and Methods* **13**, 2451–2467.

[2] Berlin, J., Stewart, J.A., Storer, B., Tutsch, K.D., Arzoomanian, R.Z., Alberti, D., Feierabend, C., Simon, K. & Wilding, G. (1997). Phase I clinical and pharmacokinetic trial of penclomedine utilizing a novel, two-stage trial design, to appear.

[3] Eichhorn, B.H. & Zacks, S. (1973). Sequential search of an optimal dosage, *Journal of the American Statistical Association* **68**, 594–598.

[4] Gatsonis, C. & Greenhouse, J.B. (1992). Bayesian methods for Phase I clinical trials, *Statistics in Medicine* **11**, 1377–1389.

[5] Mick, R. & Ratain, M.J. (1993). Model-guided determination of maximum tolerated dose in Phase I clinical trials: evidence for increased precision, *Journal of the National Cancer Institute* **85**, 217–223.

[6] O'Quigley, J. (1992). Estimating the probability of toxicity at the recommended dose following a Phase I clinical trial in cancer, *Biometrics* **48**, 853–862.

[7] O'Quigley, J. & Chevret, S. (1991). Methods for dose finding studies in cancer clinical trials: a review and results of a Monte Carlo study, *Statistics in Medicine* **10**, 1647–1664.

[8] O'Quigley, J. & Shen, L.Z. (1996). Continual reassessment method: a likelihood approach, *Biometrics* **52**, 673–684.

[9] O'Quigley, J., Pepe, M. & Fisher, L. (1990). Continual reassessment method: a practical design for Phase I clinical studies in cancer, *Biometrics* **46**, 33–48.

[10] Shen, L.Z. & O'Quigley, J. (1996). Consistency of continual reassessment method under model misspecification, *Biometrika* **83**, 395–405.

[11] Storer, B. (1989). Design and analysis of Phase I clinical trials, *Biometrics* **45**, 925–937.

[12] Storer, B. (1993). Small-sample confidence sets for the MTD in a Phase I clinical trial, *Biometrics* **49**, 1117–1125.

BARRY E. STORER

Phase II Trials

Clinical trials of new medical treatments may be classified into three successive phases. **Phase I trials** typically are small studies to determine the maximum safe dose of a drug, biological agent or combination regimen [26]. Once a dose and schedule of a new experimental regimen E have been determined, its therapeutic activity is evaluated in a Phase II trial. Phase II trials in cancer are usually single-treatment-group studies whose primary goal is to determine whether E has a level of anti-disease activity sufficiently promising to warrant further development. Phase II results also frequently serve as the basis for additional single-treatment-group studies involving E in combination with other drugs or in other dosage schedules. Phase II trials are important because they are the primary means of selecting treatments for more rigorous evaluation in Phase III trials.

Phase II trials can be categorized broadly with regard to objectives. The objective of most single-agent Phase II trials in oncology is to determine whether the agent has any anti-disease activity. This is a Phase IIA trial. Gehan [15] proposed the first Phase IIA design, a two-stage design in which n_1 patients are treated in stage 1, the trial is stopped if there are no successes in the first stage, and otherwise an additional n_2 patients are treated in stage 2. The stage 1 sample size is chosen to control type II error probability β (*see* **Hypothesis Testing**), specifically $n_1 \geq \log(\beta)/\log(1-p_1)$ (see Phase IIA trials below). The stage 2 sample size is chosen to obtain an estimate of the success probability, having **standard error** no larger than a given magnitude. The size of n_2 depends on the number of successes in the first stage, since that stage provides the estimate of success probability on which to base computation of the **binomial** standard error.

Often, the goal of a Phase II trial of a combination regimen E is to determine whether the new treatment is promising, compared with a standard treatment S. This is a Phase IIB trial. In this case, E is already known to be active. An important consideration in IIB trials is that it is clinically undesirable to continue a trial of an experimental treatment that proves not to be promising compared with S, to make way for potentially more promising new treatments. It is also important to recognize the comparative aspect of Phase IIB trials, which may lead to formal use of historical data on S in the evaluation of E, and possibly to a randomized trial [31].

Short-term response is usually used as the measure of treatment effect in Phase II trials. In oncology, partial tumor response often is not a validated measure of patient benefit. In general, the comparison of survival between responders and nonresponders is not valid for demonstrating that treatment has extended survival for responders [1]. Because response is often viewed as a necessary but not sufficient condition for extending survival, response may be used in Phase II trials for screening promising treatments. To evaluate the effectiveness of a regimen in prolonging survival, however, a Phase III trial of survival is required.

Phase IIA Trials

Single-stage Designs

The simplest Phase II design is a single-arm (i.e. treatment group) single-stage trial in which patients are treated with E. Treatment success typically is defined as a **binary** variable such as greater than 50% tumor shrinkage. The data consist of the **random variable** Y_n, the number of successes, and n, the number of patients evaluated. Y_n is **binomial** in n and the unknown success probability p. The sample size is determined so that, given a fixed standard response rate p_0 and a target response rate $p_1 = p_0 + a$, which represents a medically important improvement over p_0, a test of $H_0 : p \leq p_0$ vs. $H_1 : p \geq p_1$ has type I error probability (significance level) $\leq \alpha$ and type II error probability $\leq \beta$. The test is determined by a cutoff, r, with H_0 rejected if $Y \geq r$ and H_1 rejected if $Y < r$. The consequences of a type I error are that an uninteresting or even inferior treatment is likely to become the basis for a Phase III trial. The consequence of a type II error is that a promising treatment has been lost or its detection delayed. The required sample size n and test cutoff r are determined by specifying α, β, p_0 and δ. Reasonable values for δ are usually 0.15 to 0.20, since $\delta < 0.15$ usually leads to unrealistically large n, while $\delta > 0.20$ leads to a trial yielding very little information about E.

An alternative method of designing a single-stage trial is to choose n to obtain a **confidence interval** of given width and level (coverage probability) to estimate p. Ghosh [16] provides a good approximate

confidence interval, or the exact binomial confidence interval of Clopper & Pearson [6] may be used.

Multi-stage Designs

The most serious limitation of the single-stage design is that it has no provision for early termination if the interim observed response rate is unacceptably low. The first multistage design introduced was Gehan's two-stage design described above. Schultz et al. [23] and Fleming [13] provide a general multistage framework for Phase II trials in which n_j patients are accrued at the jth stage and a decision is made to stop or continue the trial at the end of each stage (*see* **Sequential Analysis**). At stage j, H_1 is rejected and the trial is terminated if S_j, the cumulative number of successes up to that point, $\leq a_j$. H_0 is rejected and the trial is terminated if $S_j \geq r_j$. The trial continues to the next stage if $a_j < S_j < r_j$. If the trial continues to the Kth (final) stage, then one of the two hypotheses must be rejected; hence, $a_K = r_K - 1$. The sample sizes in each stage and test cutoffs must be chosen to provide overall test error rates α and β. The actual sample size is thus random. Fleming [13] provides an explicit method for determining the test cutoffs to satisfy the error probability constraints, although the number of stages and division of patients among the stages are somewhat arbitrary. Bellisant et al. [3] present a simulation study evaluating several multistage Phase II designs.

Therneau et al. [39] provide an efficient enumeration **algorithm** that provides optimal $\{a_j, r_j\}$ boundary values for given $K, n_1, \ldots, n_k, \alpha, \beta, p_0$ and p_1.

Simon [24] derives two-stage designs that either (i) minimize the expected sample size (the optimal design) under the **null hypothesis** or (ii) minimize the maximum sample size (the **minimax** design) for given α, β, p_0, and p_1. One need not specify n_1 or n_2 because these are viewed as design parameters to be optimized. An important distinction between the two-stage version of the Fleming design and Simon's designs is that the latter allow only rejection of H_1 or continuation, but not rejection, of H_0 at the interim test. Simon [24] tabulates design parameters and operating characteristics for a broad range of parameter values, and a computer program to obtain these values is available on Statlib (OTSDEXEC.ZIP)

Garnsey-Ensign et al. [14] provide an optimal three-stage design that is essentially a combination of the Gehan [15] and optimal Simon [24] designs. At stage 1, the design stops with rejection of H_1 if there are no successes at that point; otherwise, it continues to stage 2 and (possibly) stage 3. Rejection of H_1 is possible at any stage, but H_0 may be rejected only at the final test. The design is optimal in that the expected sample size under the null hypothesis is minimized for given α, β, p_0 and p_1 subject to the requirement that $n_1 \geq 5$.

When computing a confidence interval for the response probability, one should account for interim decision rules (*see* **Data and Safety Monitoring**). Methods for adjusting one-sample confidence intervals for a binomial parameter computed after trials with interim stopping rules (*see* **Sequential Analysis**) have been discussed by a number of authors, including Jennison & Turnbull [18], Atkinson & Brown [2], and Duffy & Santner [10].

Bayesian Designs

Sylvester [29] proposes **decision-theoretic Bayesian methods** for Phase II clinical trials. He optimizes the sample size and decision cutoff of a single-stage design to determine whether a new drug is active by minimizing the Bayes risk. The approach assumes that $\Pr[\Theta_E = p_0] = 1 - \Pr[\Theta_E = p_1]$, with $p_1 > p_0$, where Θ_E is the response rate of regimen E, and p_1 and p_0 are response rates at which E would and would not be considered promising, respectively, i.e. they assume that Θ_E may take on two possible values. A more general approach is given by Brunier & Whitehead [4], who use a **beta prior distribution** for Θ_E and derive optimal Bayesian designs based on a **utility** function that accounts for both cost and the number of patients treated in a future Phase III trial.

Herson [17] proposes the use of predictive probability (PP) as a criterion for early termination of Phase II trials to minimize the number of patients exposed to an ineffective therapy. The PP of an event, such as concluding that E is or is not promising according to some decision rule, is the conditional probability of that event given the current data, computed by averaging over the posterior distribution of the parameter, which is Θ_E in the present context. Mehta & Cain [21] provide charts of early stopping rules based on the posterior probability of $\Theta_E > p_1$, where p_0 is a fixed level at which E would be considered active.

Phase IIB Clinical Trials

Most Phase IIB trials evaluate one or more new treatments relative to a standard therapy S; hence, they are inherently comparative even though a standard treatment arm usually is not included. In the designs described above, it is common to assume that p_0 is a known constant and to determine n to obtain a test of $p = p_0$ vs. $p = p_0 + \delta$ for given type I and type II error rates, α and β. For Phase IIB trials, where p_0 represents the activity level of available regimens, the numerical value of p_0 used in this computation is often a statistical estimate $\hat{p}_0$ based on historical data, rather than a known constant. The empirical difference $\hat{p} - \hat{p}_0$, which is the basis for the test, is thus the difference between two statistics and has variance larger than the assumed $p(1-p)/n$. Consequently, the sample size computed under a model ignoring the fact that $\hat{p}_0$ is a statistic is incorrect.

Makuch & Simon [19] derived single-stage Phase II designs for binary endpoints that take into consideration the number of patients in a single historical control series (*see* **Nonrandomized Trials**). Similar results were presented by Dixon & Simon [9] for time-to-event endpoints. Thall & Simon [31] derived optimal single-stage Phase II designs that incorporate historical data from one or more trials of S and account for the variability inherent in $\hat{p}_0$. They considered both binary and **normally distributed** responses. Because the variability between historical studies sometimes exceeds what is predicted by a binomial model for binary responses, they used a **beta-binomial** model to account for possible extra binomial variation. Their results indicate that it is sometimes best to randomize (*see* **Randomization**) a proportion of patients to S, and they derived the total sample size and optimal proportions for allocation to E and S that minimize $\text{var}(\hat{p} - \hat{p}_0)$. Their results indicate that an unbalanced randomization may be superior to a single-arm trial of E alone. When the uncertainty in historical control outcomes is great, either because of interstudy variability or lack of historical controls, the traditional single-arm Phase IIB trial used in oncology is not reliable. They also showed that ignoring the variability in $\hat{p}_0$ may lead to trials with actual values of α and β much higher than their nominal values.

The above method for dealing with the variability of an estimate of p_0 may be regarded as a particular approach to a more general problem. Given that in a Phase II trial the success rate of E ultimately must be compared to that of S, and that uncertainty regarding the response rate of S will always exist, the general problem is to account for this uncertainty when planning the trial and interpreting its results. A different statistical approach is based on the Bayesian framework, in which the success probabilities of E and S are regarded as random rather than fixed parameters. To underscore this distinction, we denote the random response probabilities by Θ_E and Θ_S.

Thall & Simon [32, 33] present a Bayesian approach to Phase II clinical trials in which patient response is binary and the accumulating data are monitored continuously. Their designs require an informative beta prior for Θ_S, a flat or weakly informative beta prior for Θ_E, a targeted improvement for Θ_E over Θ_S, and lower and upper bounds m and M on the allowable sample size. The maximum sample size M is chosen to obtain a given level of reliability in the posterior distribution of Θ_E. Depending upon the specific objectives, the posterior distribution of Θ_E is updated when each patient response is observed. The trial may be terminated if E is shown with high posterior probability to be either promising or not promising compared with S, or if the predictive probability of either conclusion is small. Otherwise, the trial continues. Although the framework for determining early termination bounds and M is Bayesian, the operating characteristics of the design are evaluated using frequentist criteria, and the design parameters are determined on that basis. An alternative design stops early only if E is not promising compared with S, and does not stop early if E is promising. This design would be preferred when it is desirable to continue the trial if the new treatment is promising, rather than terminate it early. A menu-driven computer program "Multcomp" to implement this design is available via anonymous FTP from `odin.mdacc.tmc.edu` in directory/pub/source.

Randomized Phase II Trials

The response rates obtained in different Phase II trials of the same treatment often vary widely. Simon et al. [25] cite a number of factors as the sources of this variability, including patient selection, definition of response, inter-observer variability in response

evaluation (*see* **Observer Reliability and Agreement**), drug dosage and schedule, reporting procedures, and sample size. To deal with these problems, they propose randomizing patients among several experimental treatments in Phase IIA trials, with **ranking** and selection methods used to interpret results. They do not require that a standard treatment arm be included. Specifically, they propose that sample size be computed to ensure that, if one group of treatments has response rate $p_0 + \delta$ and the rest have rate p_0, then a "select-the-best" strategy will choose one of the superior treatments with a desired probability. For example, 44 patients in each of three arms will ensure a 90% chance of choosing a treatment with response rate 0.35 when the other two treatments have response rate 0.20.

Randomized strategies for Phase IIB evaluation of new treatments have been considered by Whitehead [40, 41], Strauss & Simon [28], and Thall & Estey [30]. Whitehead assumes that the success rates of the experimental treatments are random and may be considered as independent draws from a beta prior distribution. Given the total number of patients, N, he derives the number of treatments, k, and number of patients per treatment, n, which maximize the expected success probability of the treatment selected based on the largest observed success rate. The constraint is that $nk = N$.

Strauss & Simon [28] study properties of a sequence of randomized Phase II trials. At each of k stages, $2n$ patients are randomized between a new experimental treatment and the better of the two treatments from the previous stage, starting with a known standard S at stage 1. The better of the two treatments at each stage, the "winner", thus becomes the new standard, and is then compared with the next experimental treatment. The goal is to select a single treatment of Phase III evaluation. Similar to Whitehead [40], Strauss & Simon assume that the success probabilities of the experimental treatments are independent draws from a beta prior distribution, either with fixed mean equal to that of S or with distribution adapted to the data in that its mean equals that of the latest winner.

Multiple Outcomes

The designs discussed in the preceding sections are based on a single binary outcome. Patient response in clinical trials is an inherently multidimensional phenomenon, however, with the possibility of both adverse events and efficacy outcomes (*see* **Multiplicity in Clinical Trials**). In addition to evaluating treatment efficacy, a Phase II trial must determine whether an experimental treatment is sufficiently safe to allow its evaluation in a large randomized trial.

Thall et al. [34, 37] present a general Bayesian strategy for monitoring multiple outcomes in single-arm clinical trials. Each patient's response is characterized as a **multinomial** variable that records the specific combination of events occurring for the patient in the course of the trial. This includes both adverse events and efficacy outcomes, possibly occurring at different study times. They use a Dirichlet–multinomial model to accommodate general discrete multivariate responses (*see* **Multivariate Distributions, Overview**), and they provide Bayesian decision criteria for early termination of studies with unacceptably high rates of adverse outcomes or with low rates of efficacy outcomes. Each stopping rule is constructed either to control the rate of an adverse event or to achieve a specified level of improvement of an efficacy event rate for the experimental treatment, compared with that of standard therapy. They avoid explicit specification of costs and a loss function, and evaluate the joint behavior of the multiple decision rules using frequentist criteria. Their approach accommodates a broad range of clinical situations, including settings in which observation of certain endpoints is conditional on the occurrence of earlier events. They illustrate the approach with a variety of single-arm cancer trials.

Etzione & Pepe [12] propose a Bayesian criterion for monitoring two adverse outcomes in a pilot toxicity study, in the case where the occurrence of one event precludes the occurrence of the other.

Bryant & Day [5] extend Simon's [24] minimax design to the setting with an efficacy endpoint and a toxicity endpoint. They determine two-stage designs that minimize the maximum expected number of patients entered when the treatment is unacceptable either in terms of clinical response or toxicity. Conaway & Petroni [7, 8] also propose Phase II designs accounting for both efficacy and toxicity, and formulate hypotheses in terms of tradeoffs between these outcomes.

Discussion

In oncology, nearly any clinical trial that is not a dose-finding study and that does not contain a randomized control group is called a Phase II trial. Consequently, the Phase II category is quite heterogeneous with regard to objectives and characteristics. Unfortunately, these differences are not always recognized, and statistical designs developed for one type of Phase II trial are sometimes inappropriately applied to another type.

Until recently, most statistical designs developed for Phase II clinical trials were applicable primarily to the objectives of Phase IIA trials. These include the designs of Gehan [15], Schultz et al. [23], Fleming [13], Simon [24], and Therneau et al. [39]. These and other Phase II designs are reviewed by Mariani & Marubini [20]. Phase IIB trials have the objective of determining whether a new regimen has a level of anti-disease activity that is promising relative to the best available regimens. In dealing with combination regimens, sometimes involving a complex sequential treatment program for the patient, it is not relevant to show that the regimen is active. Rather, the focus often is on determining whether the combination regimen under test is sufficiently active, compared with the activity level of the best available standard therapy, to warrant a Phase III trial. Hence, Phase IIB trials are inherently comparative. Although Phase IIA designs can be used for Phase IIB trials if the response probability for the best available regimen is known accurately, this is seldom the case. Usually, the comparative aspects of Phase IIB trials are either suppressed or not addressed directly. This can have two undesirable effects. The first is that the results with the experimental regimen may appear so promising that a Phase III trial is difficult to conduct, since randomization to a control arm appears unethical. The second is that the results are inappropriately interpreted as promising and a Phase III trial is conducted when it is not warranted. In general, we believe that the comparative aspects of Phase IIB trials should be addressed directly, that specific control groups should be identified, and that uncertainties arising from the use of nonrandomized control groups of finite size should be quantified. The designs of Thall & Simon [31–33] and Thall et al. [34, 37] address this. These designs are, however, quite different from those developed for the simple Phase IIA trials.

An alternative to conducting a Phase IIB trial is to use a Phase III randomized design, allowing one or several experimental regimens, with early termination of a treatment arm if early results with that regimen are sufficiently discouraging. The designs described by Ellenberg & Eisenberger [11], Thall et al. [38], Thall et al. [35, 36], Wieand & Therneau [42], Schaid et al. [22], and Storer [27] are of this type. It is often difficult to organize a Phase III trial of an experimental regimen, however, without some earlier Phase II experience with that regimen.

References

[1] Anderson, J.R., Cain, K.C. & Gelber, R.D. (1983). Analysis of survival by tumor response, *Journal of Clinical Oncology* **1**, 710–719.

[2] Atkinson, E.N. & Brown, B.W. (1985). Confidence limits for probability of response in multistage clinical trials, *Biometrics* **41**, 741–744.

[3] Bellisant, E., Benichou, J. & Chastang, C. (1990). Application of the triangular test to phase II cancer clinical trials, *Statistics in Medicine* **9**, 907–917.

[4] Brunier, H.C. & Whitehead, J. (1994). Sample sized for phase II clinical trials derived from Bayesian decision theory, *Statistics in Medicine* **13**, 2493–2502.

[5] Bryant, J. & Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials, *Biometrics* **51**, 1372–1383.

[6] Clopper, C.J. & Pearson, E.S. (1934). The use of confidence of fiducial limits illustrated in the case of the binomial, *Biometrika* **26**, 404–413.

[7] Conaway, M. & Petroni, G. (1995). Bivariate sequential designs for phase II trials, *Biometrics* **51**, 656–664.

[8] Conaway, M. & Petroni, G. (1996). Designs for phase II trials allowing for a tradeoff between response and toxicity, *Biometrics* **52**, 1375–1386.

[9] Dixon, D.O. & Simon, R. (1988). Sample size considerations for studies comparing survival curves using historical controls, *Journal of Clinical Epidemiology* **41**, 1209–1214.

[10] Duffy, D.E. & Santner, T.J. (1987). Confidence intervals for a binomial parameter based on multistage tests, *Biometrics* **43**, 81–93.

[11] Ellenberg, S.S. & Eisenberger, M.A. (1985). An efficient design for phase III studies of combination chemotherapies, *Cancer Treatment Reports* **69**, 1147–1154.

[12] Etzioni, R. & Pepe, M.S. (1994). Monitoring of a pilot toxicity study with two adverse outcomes, *Statistics in Medicine* **13**, 2311–2321.

[13] Fleming, T.R. (1982). One sample multiple testing procedure for phase II clinical trials, *Biometrics* **38**, 143–151.

[14] Garnsey-Ensign, L., Gehan, E.A., Kamen, D. & Thall, P.F. (1994). An optimal three-stage design for phase II clinical trials, *Statistics in Medicine* **13**, 1727–1736.

[15] Gehan, E.A. (1961). The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent, *Journal of Chronic Diseases* **13**, 346–353.

[16] Ghosh, B.K. (1979). A comparison of some approximate confidence intervals for the binomial parameter, *Journal of the American Statistical Association* **74**, 894–900.

[17] Herson, J. (1979). Predictive probability early termination plans for phase II clinical trials, *Biometrics* **35**, 775–783.

[18] Jennison, C. & Turnbull, B.W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials, *Technometrics* **25**, 49–58.

[19] Makuch, R.W. & Simon, R.M. (1980). Sample size considerations for non-randomized comparative studies, *Journal of Chronic Diseases* **33**, 175–181.

[20] Mariani, L. & Marubini, E. (1996). Design and analysis of phase II cancer clinical trials: a review of statistical methods and guidelines for medical researchers, *International Statistical Review* **64**, 61–88.

[21] Mehta, C.R. & Cain, K.C. (1984). Charts for the early stopping of pilot studies, *Journal of Clinical Oncology* **2**, 676–682.

[22] Schaid, D.J., Wieand, S. & Therneau, T.M. (1990). Optimal two-stage screening designs for survival comparisons, *Biometrika* **77**, 507–513.

[23] Schultz, J.R., Nichol, F.R., Elfring, G.L. & Weed, S.D. (1973). Multiple stage procedures for drug screening, *Biometrics* **29**, 293–300.

[24] Simon, R. (1989). Optimal two-stage designs for phase II clinical trials, *Controlled Clinical Trials* **10**, 1–10.

[25] Simon, R., Wittes, R.E. & Ellenberg, S.S. (1985). Randomized phase II clinical trials, *Cancer Treatment Reports* **69**, 1375–1381.

[26] Storer, B.E. (1989). Design and analysis of phase I clinical trials, *Biometrics* **45**, 925–937.

[27] Storer, B.E. (1990). A sequential phase II/III trial for binary outcomes, *Statistics in Medicine* **9**, 229–235.

[28] Strauss, N. & Simon, R. (1995). Investigating a sequence of randomized phase II trials to discover promising treatments, *Statistics in Medicine* **14**, 1479–1489.

[29] Sylvester, R.J. (1988). A Bayesian approach to the design of phase II clinical trials, *Biometrics* **44**, 823–836.

[30] Thall, P.F. & Estey, E.H. (1993). A Bayesian strategy for screening cancer treatments prior to phase II clinical evaluation, *Statistics in Medicine* **12**, 1197–1211.

[31] Thall, P.F. & Simon, R. (1990). Incorporating historical control data in planning phase II clinical trials, *Statistics in Medicine* **9**, 215–228.

[32] Thall, P.F. & Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials, *Biometrics* **50**, 337–349.

[33] Thall, P.F. & Simon, R. (1994). A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials, *Controlled Clinical Trials* **15**, 463–481.

[34] Thall, P.F. & Simon, R. (1996). New statistical strategy for monitoring safety and efficacy in single-arm clinical trials, *Journal of Clinical Oncology* **14**, 296–303.

[35] Thall, P.F., Simon, R. & Ellenberg, S.S. (1989). Two stage selection and testing designs for comparative clinical trials, *Biometrika* **75**, 303–310.

[36] Thall, P.F., Simon, R. & Ellenberg, S.S. (1989). A two-stage design for choosing among several experimental treatments and a control in clinical trials, *Biometrics* **45**, 537–547.

[37] Thall, P.F., Simon, R. & Estey, E.H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes, *Statistics in Medicine* **14**, 357–379.

[38] Thall, P.F., Simon, R., Ellenberg, S.S. & Shrager, R. (1988). Optimal two-stage designs for clinical trials with binary response, *Statistics in Medicine* **7**, 571–579.

[39] Therneau, T.M., Wieand, H.S. & Chang, S.M. (1990). Optimal designs for a grouped sequential binomial test, *Biometrics* **46**, 771–781.

[40] Whitehead, J. (1985). Designing phase II studies in the context of a program of clinical research, *Biometrics* **41**, 373–383.

[41] Whitehead, J. (1986). Samples sizes for phase II and III clinical trials: an integrated approach, *Statistics in Medicine* **5**, 459–464.

[42] Wieand, H.S. & Therneau, T.M. (1987). A two-stage design for randomized trials with binary outcomes, *Controlled Clinical Trials* **8**, 20–28.

RICHARD SIMON & PETER F. THALL

Phase-type Distributions in Survival Analysis

When studying time to failure of some sort, it is often natural to imagine that the process goes through a number of stages, or phases. In cancer epidemiology, this is an old recognition; the **multistage model** of Armitage & Doll [2] postulated that development of cancer went through a number of stages corresponding to a simple homogeneous Markov model (*see* **Markov Processes**). This assumption fits very well with the **Weibull** hazard function that is often observed for cancer incidence as a function of age. The Armitage–Doll model has been an inspiration in the understanding of carcinogenesis, although it has now been overtaken by more complex models.

The Armitage–Doll model is a special case of a phase-type distribution. In general, consider a time-continuous **Markov chain** with a finite number of states and constant transition intensities. One state is supposed to be absorbing, the remaining states being transient. Assume that the Markov process starts somewhere in the transient space, and consider the time until absorption. A transition time defined in this way is said to have a phase-type distribution.

Considerable attention has been paid to this kind of distribution in probability theory, especially in **queuing** theory where they are a means for developing computationally manageable models. Some basic references are [3, 5, 9], and [10]. Estimation problems have been considered by Asmussen et al. [4] and Olsson [11]. Phase-type distributions may also be of use in **survival analysis**, both for the insights they give and as an estimation tool; see [1]. In particular, phase-type models can deepen our understanding of **hazard** functions. An important modern use of a simple phase-type model has been given by Longini et al. [8].

The Armitage–Doll and Longini distributions are so-called series models (see Figure 1, where the αs are constant transition intensities). Such simple series models may help in answering the question: Why do hazard functions often increase with time? This is observed very often, like the increasing incidence of cancer and heart disease and of general mortality as a function of age. Phase-type theory tells us that the hazard function for the time to move from the initial to the final state in a series model is always increasing with time, and reaching a maximum level equal to the smallest transition intensity. This increase is in spite of the constant transition intensities, so no aging in the sense of increasing intensities is needed to see an increasing hazard. This point has been made repeatedly in cancer epidemiology [6], but may have a more general validity.

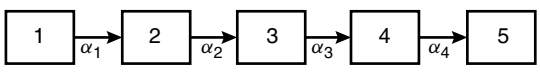

Figure 1 Series model with five states

However, hazard functions do not always increase inexorably; sometimes they decrease, or they have an initial increase followed by a decline. In the phase-type framework such phenomena can also be explained. To do this, one must introduce the concept of *quasi-stationary distribution* on the transient state space. If the Markov chain starts out with this initial distribution, then the hazard function of the time until absorption is constant. A quasi-stationary distribution often exists [1, 7], and can be considered as a suitable normalized limiting distribution on the transient state space.

Under certain assumptions (like irreducibility of the transient state space) the hazard function of any phase-type distribution will approach the constant value consistent with the quasi-stationary initial distribution. For some typical processes, allowing movements back and forth between states, the following will be observed [1]: if the process starts out in a state close enough to the absorbing one, then the hazard function will be monotonically decreasing. If the process starts out in an intermediate distance from the absorbing state, then one will observe a hazard function that first increases and then declines. If the starting state is far enough out, then the hazard function will be monotonically increasing.

In general, there are two forces influencing the Markov process: one is the "pull" of the absorbing state, and the second is the diffusion on the transient state. It is the balance of these two forces that determines the shape of the hazard function.

References

[1] Aalen, O.O. (1995). Phase type distributions in survival analysis, *Scandinavian Journal of Statistics* **22**, 447–463.

[2] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–15.

[3] Asmussen, S. (1987). *Applied Probability and Queues*. Wiley, New York.

[4] Asmussen, S., Nerman, O. & Olsson, M. (1996). Fitting phase type distributions via the EM algorithm, *Scandinavian Journal of Statistics* **23**, 419–441.

[5] Bobbio, A. & Cumani, A. (1992). ML estimation of the parameters of a PH distribution in triangular canonical form, in *Computer Performance Evaluation*, G. Balbo & G. Serazzi, eds. Elsevier, Amsterdam, pp. 33–46.

[6] Day, N.E. & Brown, C.C. (1980). Multistage models and primary prevention of cancer, *Journal of the National Cancer Institute* **64**, 977–989.

[7] Keilson, J. (1979). *Markov Chain Models – Rarity and Exponentiality*. Springer-Verlag, New York.

[8] Longini, I.M., Clark, W.S., Byers, R.H., Ward, J.W., Darrell, W.W., Lamp, G.I. & Hethcote, H.W. (1989). Statistical analysis of the stages of HIV-infections using a Markov model, *Statistics in Medicine* **8**, 831–843.

[9] Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore.

[10] O'Cinneide, C.A. (1990). Characterization of phase type distributions, *Communications in Statistics – Stochastic Models* **6**, 1–57.

ODD O. AALEN

Physicians' Health Study

The Physicians' Health Study was a randomized, double-blind, placebo-controlled, 2×2 factorial trial of aspirin (325 mg every other day) and β-carotene (50 mg every other day) among 22 071 apparently healthy US male physicians. The aspirin arm was terminated early, on 25 January 1988 (median treatment and follow-up, five years) due primarily to the emergence of a statistically extreme ($P < 0.00001$) 44% reduction in risk of a first myocardial infarction. The β-carotene arm ended as scheduled on 31 December 1995 (median treatment and follow-up, 12 years), and showed no overall benefit or harm of β-carotene supplementation on cancer or cardiovascular disease. The Physicians' Health Study cohort continues to provide important information regarding the causes and prevention of cardiovascular disease, cancer, and other chronic diseases.

Background and Rationale

A critical element that must be considered when designing a trial is the balance between the evidence supporting the hypothesis being tested and the gap in knowledge that may be filled by the results. Achieving this balance is a particularly delicate matter in randomized clinical trials. For both ethical and practical reasons, there must be sufficient belief in the potential of the agent to be tested to justify exposing half of the subjects to it, just as there must be sufficient doubt about the agent to allow withholding treatment from the other half [15].

In the late 1970s, we viewed these conditions to be applicable to the use of aspirin in the primary prevention of myocardial infarction and the use of β-carotene in the primary prevention of cancer. For aspirin, a plausible Nobel prize-winning mechanism [38] had been advanced for the possible use of aspirin in the treatment and prevention of cardiovascular disease [33, 34]. The epidemiologic evidence, however, was not consistent. A large case series from California [10], two case–control studies [5, 18], and a prospective cohort study had suggested a possible small beneficial effect of aspirin in men and women [12], while another case–control study had found no association between aspirin use and fatal coronary disease [17]. With respect to β-carotene, more than 20 epidemiologic studies conducted in various parts of the world had generally found reduced risks of cancers at various sites associated with higher intakes of dark green and yellow vegetables, which are abundant in β-carotene [26]. In addition, the use of β-carotene-containing multivitamins or supplements was increasing rapidly, with as many as one-third to one-half of US adults reporting daily vitamin consumption [39], based largely on the belief, unsupported by randomized trials, that β-carotene could prevent cancer.

Thus, randomized trials testing whether aspirin and β-carotene conferred benefits were both important and timely. The Physicians' Health Study was designed to accomplish these and other objectives.

Methods

Detailed descriptions of the Physicians' Health Study have been published elsewhere [16, 28, 29]. In brief, from 1981 to 1984, invitation letters, consent forms, and enrollment questionnaires were sent to 261 248 male physicians between 40 and 84 years of age residing in the US who were registered with the American Medical Association in 1981. Of these, 59 285 were willing to participate in the trial, 26 062 of whom were excluded because they indicated on the baseline questionnaire a history of myocardial infarction, stroke, or transient ischemic attack; cancer (except nonmelanoma skin cancer); current renal or liver disease; peptic ulcer; gout; or contraindication to or current use of either aspirin or β-carotene. The remaining 33 223 willing and eligible physicians were enrolled in an 18-week run-in phase to assess willingness to participate and compliance. Of these, 22 071 physicians were randomized using a 2×2 **factorial design** to one of four treatment groups: aspirin, β-carotene, both, and neither. A total of 11 037 physicians were randomized to aspirin and 11 034 to aspirin placebo; 11 034 physicians were randomized to β-carotene and 11 037 to β-carotene placebo. Active aspirin consisted of one 325 mg tablet (as Bufferin, supplied by Bristol-Myers Products) to be taken every other day. Active β-carotene consisted of one 50 mg capsule (as Lurotin, supplied by BASF AG) to be taken every other day.

Follow-up

All follow-up was conducted by mailed questionnaire. The initial follow-ups were at 6 and 12 months, then yearly thereafter.

Endpoints

Endpoints (*see* **Outcome Measures in Clinical Trials**) were considered confirmed or refuted only after an endpoints committee made up of two internists, a cardiologist, and a neurologist, all of whom were blinded to the treatment assignments, had examined all available information, including medical records.

Blood Samples

Kits containing Vacutainer tubes containing ethylenediamine tetraacetic acid, complete instructions for blood draws, polypropylene cryopreservation vials, and coldpacks were mailed to all 33 223 physicians participating in the run-in phase. They were asked to have their blood drawn into the Vacutainer tubes, to fractionate the blood by centrifugation, and to return the samples in the coldpack by prepaid overnight courier. Upon receipt in the laboratory, each sample was divided into aliquots and stored at -82°C. Specimens were received from 14 916 (68%) of the randomized physicians. None of the samples collected between 1982 and 1984 has inadvertently thawed during storage.

Results

Aspirin

The aspirin component of the trial was terminated early in 1988, after an average follow-up of 60.2 months, due primarily to the emergence of a statistically extreme benefit on risk of a first myocardial infarction. Aspirin takers had a 44% reduction in risk of a first myocardial infarction, with significant benefits on fatal and nonfatal events [28, 29]. As a consequence of early termination, there were insufficient numbers of strokes upon which to draw firm conclusions. However, the available data did not suggest any reduction in stroke, and there was, in fact, a possible but not statistically significant 19% increase in nonfatal stroke. Because of aspirin's effect on platelet aggregation, a particular concern with its use is a possible increase in hemorrhagic stroke. For this subgroup of strokes, there was the suggestion of a possible increased risk in the aspirin group (23 events vs. 12 events), although the numbers were small and did not achieve conventional statistical significance ($P = 0.06$). There were also too few cardiovascular deaths upon which to draw firm inferences.

β-carotene

Virtually no early or late differences in the overall incidence of malignant neoplasms or cardiovascular disease, or in overall mortality, were observed between the β-carotene and placebo groups. In the β-carotene group, 1273 men had a malignant neoplasm (except nonmelanoma skin cancer), as compared with 1293 in the placebo group (relative risk, 0.98; 95%, 0.91 to 1.06). There were also no significant differences in the number of cases of lung cancer (82 in the β-carotene group vs. 88 in the placebo group); the number of deaths from cancer (386 vs. 380), deaths from any cause (979 vs. 968), or deaths from cardiovascular disease (338 vs. 313); the number of men with myocardial infarction (468 vs. 489); the number with stroke (367 vs. 382); or the number with any one of the previous three endpoints (967 vs. 972). Among current and former smokers, there were also no significant early or late differences in any of these endpoints. There were, however, suggestions that among the prospective subgroup of men with the lowest blood levels of β-carotene at randomization (*see* **Treatment-covariate Interaction**), β-carotene reduced the risk of total and prostate cancer, and also reduced the risk of subsequent vascular events among the subgroup of men with angina at baseline.

Discussion of Results

The Physicians' Health Study was the first randomized trial to demonstrate a benefit of aspirin in reducing the risk of a first myocardial infarction. This finding was not supported by the smaller British Doctors Trial [27], but has been confirmed in two more recently reported primary prevention trials, the Thrombosis Prevention Trial [37] and the Hypertension Optimal Treatment trial [13]. An overview (*see* **Meta-analysis of Clinical Trials**) of all four available trials shows a highly significant ($P < 0.00001$)

33% reduction in risk of a first myocardial infarction attributed to aspirin [14], while there are still too few strokes and deaths upon which to draw firm conclusions.

With regard to β-carotene, the Physicians' Health study was by far the longest randomized trial of this agent – at 12 years of treatment and follow-up, it was two to three times longer than any other randomized trial of β-carotene. Since the Physicians' Health Study in 1982, three other trials of shorter duration have been designed and completed. The Chinese Cancer Prevention Study, conducted among a nutritionally deficient rural population, found a statistically significant 9% decrease in total mortality, a 13% decrease in total cancer mortality, and a 21% decrease in gastric cancer deaths, and a non-significant 10% decrease in cerebrovascular mortality among those assigned to a combined daily treatment of β-carotene (15 mg), vitamin E (30 mg), and selenium (50 μg) [4]. In the six-year Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC), conducted among male Finnish smokers, the risk of cancer was not lower among the men on active treatment (50 mg of alpha-tocopherol, or 20 mg of β-carotene, or both) compared with those in the placebo arm. In fact, β-carotene use was associated with statistically significant increases in lung cancer (18%), total mortality (8%), and mortality from ischemic heart disease (12%) [36]. The Beta-Carotene and Retinol Efficacy Trial (CARET), conducted among men and women at high risk of lung cancer, was stopped prematurely, with an average of just four years of treatment and follow-up, because preliminary results showed an increased risk of lung cancer, death from lung cancer, and death from cardiovascular disease in the active treatment group [25]. It must be noted that the possible increased risk did not reach the prespecified O'Brien–Fleming boundary ($P < 0.007$) for early termination (*see* **Data and Safety Monitoring**). Thus, the null results from the Physicians' Health Study are the most accurate, least biased, and least confounded results of β-carotene in primary prevention to date.

Other Benefits of the Physicians' Health Study

The Physicians' Health Study was the first large-scale study ever to collect by mail and store a large number of blood samples. Approximately 15 000 randomized physicians (68%) contributed blood samples at baseline, which were stored as aliquots in liquid nitrogen. This bank of blood samples has offered the opportunity to conduct large-scale blood-based epidemiologic studies using a prospective, nested, case–control design. Completed studies include investigations on the influence of a variety of potential markers on risk of coronary heart disease, including homocysteine [35] and inflammatory markers such as C-reactive protein [30], fibrinogen [22], and soluble intercellular adhesion molecule-1 [31]. The stored blood samples have also allowed for the rapid evaluation of hypotheses generated from the identification of possible genetic markers. For example, a 1992 case–control study [7] suggested an association between a polymorphism of the angiotensin-converting-enzyme (ACE) gene and the occurrence of myocardial infarction. Prospective blood-based data from the Physicians' Health Study reported shortly thereafter, however, showed that the presence of the D allele of the ACE gene conferred no appreciable increase in the risk of ischemic heart disease or myocardial infarction [21].

In addition to blood-based epidemiologic investigations, the Physicians' Health Study has been an excellent cohort for testing a variety of timely and important hypotheses from observational data. These range from a positive association between smoking and risk of cataract [9] to an inverse association between exercise and risk of type 2 diabetes [23] and stroke [20], as well as the investigation of a number of dietary factors (fish consumption [2, 3, 24], vitamins B_6 and folate [8], and moderate alcohol consumption [1, 6, 11]) on risk of various diseases.

The Physicians' Health Study has also provided an opportunity to explore the decreases in mortality from cardiovascular disease that began in the mid 1960s. In 1982, when the study began, 28% of American men and 19% of US physicians smoked, though only 11% of Physicians' Health Study participants reported being current smokers. That difference could be one reason for the low rate of incidence of cardiovascular events observed in the Physicians' Health Study. With respect to case fatality rates for myocardial infarction, among participants in the Physicians' Health Study, the average time from onset of symptoms of myocardial infarction to emergency medical care was 1.8 hours, while the population

average was 4.8 hours [32]. Shifting the population average toward that observed among Physicians' Health Study participants could significantly decrease mortality and morbidity from myocardial infarction.

Finally, a number of methodological advances have been tested in the Physicians' Health Study. The use of a prerandomization run-in, for example, enhanced the validity and made the trial more efficient – **compliance** with pill taking increased up to 40% and duration of follow-up decreased 7%. In addition, enrolling 11 152 fewer subjects in the trial resulting from the use of the run-in led to substantial cost savings [19] (*see* **Cost-effectiveness in Clinical Trials**). The Physicians' Health Study also demonstrated the advantages of the 2×2 factorial trial, a research design allowing for the independent and simultaneous evaluation of two hypotheses. Finally, the trial showed the feasibility of conducting a large-scale randomized trial entirely by mail at a fraction of the usual cost for randomized trials.

Conclusion

The Physicians' Health Study has made and should continue to make important substantive and methodologic contributions to the investigation of the causes and prevention of cardiovascular disease, cancer, and other chronic diseases.

References

[1] Ajani, U.A., Hennekens, C.H., Spelsberg, A. & Manson, J.E. (2000). Alcohol consumption and risk of type 2 diabetes mellitus among US male physicians, *Archives of Internal Medicine* **160**, 1025–1030.

[2] Albert, C.M., Hennekens, C.H., O'Donnell C.J. et al. (1998). Fish consumption and risk of sudden cardiac death, *Journal of the American Medical Association* **279**, 23–28.

[3] Albert, C.M., Manson, J.E., Hennekens, C.H. & Ruskin, J.N. (1997). Fish consumption and the risk of myocardial infarction, *New England Journal of Medicine* **337**, 497–498; discussion 498–499.

[4] Blot, W.J., Li, J.Y., Taylor, P.R. et al. (1993). Nutrition intervention trials in Linxian, China: supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population, *Journal of the National Cancer Institute* **85**, 1483–1492.

[5] Boston Collaborative Drug Surveillance Group (1974). Regular aspirin intake and acute myocardial infarction, *British Medical Journal* **1**, 440–443.

[6] Camargo, C.A. Jr, Hennekens, C.H., Gaziano, J.M., Glynn, R.J., Manson, J.E. & Stampfer, M.J. (1997). Prospective study of moderate alcohol consumption and mortality in US male physicians, *Archives of Internal Medicine* **157**, 79–85.

[7] Cambien, F, Poirier, O, Lecerf, L. et al. (1992). Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction, *Nature* **359**, 641–644.

[8] Chasan-Taber, L., Selhub, J., Rosenberg, I.H. et al. (1996). A prospective study of folate and vitamin B6 and risk of myocardial infarction in US physicians, *Journal of the American College of Nutrition* **15**, 136–143.

[9] Christen, W.G., Manson, J.E., Seddon, J.M. et al. (1992). A prospective study of cigarette smoking and risk of cataract in men, *Journal of the American Medical Association* **268**, 989–993.

[10] Craven, L.L. (1956). Prevention of coronary thrombosis and cerebral thrombosis, *Mississippi Valley Medical Journal* **78**, 213–215.

[11] Gaziano, J.M., Gaziano, T.A., Glynn, R.J. et al. (2000). Light-to-moderate alcohol consumption and mortality in the Physicians' Health Study enrollment cohort, *Journal of the American College of Cardiology* **35**, 96–105.

[12] Hammond, E.C. & Garfinkel, L. (1975). Aspirin and coronary heart disease: findings of a prospective study, *British Medical Journal* **2**, 269–271.

[13] Hansson, L., Zanchetti, A., Carruthers, S.G. et al. (1998). Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomized trial. HOT Study Group, *Lancet* **351**, 1755–1762.

[14] Hebert, P.R. & Hennekens, C.H. (2000). Overview of aspirin in the primary prevention of cardiovascular disease, *Archives of Internal Medicine* **160**, 3123–3127.

[15] Hennekens, C.H. & Buring J.E. (1987). *Epidemiology in Medicine*. Little, Brown & Company, Boston.

[16] Hennekens, C.H., Buring, J.E., Manson, J.E. et al. (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease, *New England Journal of Medicine* **334**, 1145–1149.

[17] Hennekens, C.H., Karlson L.K. & Rosner, B. (1978). A case-control study of regular aspirin use and coronary deaths, *Circulation* **58**, 35–38.

[18] Jick, H. & Miettinen, O.S. (1976). Regular aspirin use and myocardial infarction, *British Medical Journal* **1**, 1057.

[19] Lang, J.M., Buring, J.E., Rosner, B., Cook, N. & Hennekens, C.H. (1991). Estimating the effect of the run-in on the power of the Physicians' Health Study, *Statistics in Medicine* **10**, 1585–1593.

[20] Lee, I.M., Hennekens, C.H., Berger, K., Buring, J.E. & Manson, J.E. (1999). Exercise and risk of stroke in male physicians, *Stroke* **30**, 1–6.

[21] Lindpaintner, K., Pfeffer, M.A., Kreutz, R. et al. (1995). A prospective evaluation of an angiotensin-converting-enzyme gene polymorphism and the risk of ischemic heart disease, *New England Journal of Medicine* **332**, 706–711.

[22] Ma, J., Hennekens, C.H., Ridker, P.M. & Stampfer, M.J. (1999). A prospective study of fibrinogen and risk of myocardial infarction in the Physicians' Health Study, *Journal of the American College of Cardiology* **33**, 1347–1352.

[23] Manson, J.E., Nathan, D.M., Krolewski, A.S., Stampfer, M.J., Willett, W.C. & Hennekens, C.H. (1992). A prospective study of exercise and incidence of diabetes among US male physicians, *Journal of the American Medical Association* **268**, 63–67.

[24] Morris, M.C., Manson, J.E., Rosner, B., Buring, J.E., Willett, W.C. & Hennekens, C.H. (1995). Fish consumption and cardiovascular disease in the physicians' health study: a prospective study, *American Journal of Epidemiology* **142**, 166–175.

[25] Omenn, G.S., Goodman, G.E., Thornquist, M.D. et al. (1996). Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease, *New England Journal of Medicine* **334**, 1150–1155.

[26] Peto, R., Doll, R., Buckley, J.D. & Sporn, M.B. (1981). Can dietary beta-carotene materially reduce human cancer rates?, *Nature* **290**, 201–208.

[27] Peto, R., Gray, R., Collins, R. et al. (1988). Randomized trial of prophylactic daily aspirin in British male doctors, *British Medical Journal* **296**, 313–316.

[28] Physicians' Health Study (1988). Preliminary report: findings from the aspirin component of the ongoing Physicians' Health Study, *New England Journal of Medicine* **318**, 262–264.

[29] Physicians' Health Study Research Group (1989). Final report on the aspirin component of the ongoing Physicians' Health Study. Steering Committee of the Physicians' Health Study Research Group, *New England Journal of Medicine* **321**, 129–135.

[30] Ridker, P.M., Cushman, M., Stampfer, M.J., Tracy, R.P. & Hennekens, C.H. (1997). Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men, *New England Journal of Medicine* **336**, 973–979.

[31] Ridker, P.M., Hennekens, C.H., Roitman-Johnson, B., Stampfer, M.J. & Allen, J. (1998). Plasma concentration of soluble intercellular adhesion molecule 1 and risks of future myocardial infarction in apparently healthy men, *Lancet* **351**, 88–92.

[32] Ridker, P.M., Manson, J.E., Goldhaber, S.Z., Hennekens, C.H. & Buring, J.E. (1992). Comparison of delay times to hospital presentation for physicians and nonphysicians with acute myocardial infarction, *American Journal of Cardiology* **70**, 10–13.

[33] Roth, G.J. & Majerus, P.W. (1975). The mechanism of the effect of aspirin on human platelets. I. Acetylation of a particulate fraction protein, *Journal of Clinical Investigation* **56**, 624–632.

[34] Smith, J.B. & Willis, A.L. (1971). Aspirin selectively inhibits prostaglandin production in human platelets, *Nature New Biology* **231**, 235–237.

[35] Stampfer, M.J., Malinow, M.R., Willett, W.C. et al. (1992). A prospective study of plasma homocyst(e)ine and risk of myocardial infarction in US physicians, *Journal of the American Medical Association* **268**, 877–881.

[36] The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers, *New England Journal of Medicine* **330**, 1029–1035.

[37] The Medical Research Council's General Practice Research Framework (1998). Thrombosis prevention trial: randomized trial of low-intensity oral anticoagulation with warfarin and low-dose aspirin in the primary prevention of ischaemic heart disease in men at increased risk, *Lancet* **351**, 233–241.

[38] Vane, J.R. (1971). Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs, *Nature New Biology* **231**, 232–235.

[39] Willett, W., Sampson, L., Bain, C., et al. (1981). Vitamin supplement use among registered nurses, *American Journal of Clinical Nutrition* **34**, 1121–1125.

P.J. SKERRETT & CHARLES H. HENNEKENS

Pillai's Trace Test

Pillai in 1955 [24] proposed the trace test for the following three testing problems: (i) equality of mean vectors of lp-variate normal distributions with the common but unknown **covariance matrix** (*see* **Multivariate Normal Distribution**); (ii) independence between two sets of variates distributed jointly as a normal distribution with unknown mean vector; and (iii) equality of covariance matrices of two p-variate normal distributions with unknown mean vectors.

In all these problems the test statistic can be expressed as $V^{(s)} = \text{trace}(\mathbf{B})$, where $\mathbf{B} = \mathbf{S}_1(\mathbf{S}_1 + \mathbf{S}_2)^{-1}$, $\mathbf{S}_1$ and $\mathbf{S}_2$ being $p \times p$ matrices, s being the number of nonzero characteristic roots of B (*see* **Eigenvalue**). The problem (i) can be seen to be equivalent to the **multivariate analysis of variance** (MANOVA) problem, and, in that case, $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the sums of squares and cross-products matrices "due to hypothesis" and "due to error", respectively. For problem (ii), $\mathbf{S}_1 = \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$, $\mathbf{S}_2 = \mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$, where $\mathbf{S}_{ij}$, $i, j = 1, 2$, is the partitioned sums of squares and cross-products matrix corresponding to the two sets of variates, based on a random sample from the distribution. For problem (iii), $\mathbf{S}_1$ and $\mathbf{S}_2$ are the sums of squares and cross-products matrices based on **random samples** from the two distributions with **degrees of freedom** (df) ν_1 and ν_2, respectively.

The trace test has also been considered by Bartlett [3] for (ii), and Nanda [20] has considered the distribution of this statistic when $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent central **Wishart** matrices.

We assume $s = p$; if $s < p$ for (i), then the parameters can be modified appropriately. The null distribution of $V^{(s)}$ takes the same form in each of these three problems. Nanda [20] has derived the distribution of $V^{(s)}$ when s = 2,3, and $|\nu_1 - p| = 1$. Pillai [24] has suggested approximation of the null distribution of $V^{(s)}/s$ by $B[s(2m + s + 1), s(2n + s + 1)]$, where $m = (|\nu_1 - p| - 1)/2$, $n = (\nu_2 - p - 1)/2$ (*see* **Beta Distribution**), and recommended this approximation when $m + n > 30$; this amounts to approximating the cut-off points of the test statistic based on the upper percentage points of the ***F* distribution**. Pillai & Mijares [32], and Mijares [17] have tabulated upper 5% and 1% points of $V^{(s)}$ based on this approximation for small s and various m and n; see also Pillai [25, 26]. Pillai & Jayachandran [30] have obtained the **moment generating function** of $V^{(s)}$ and derived explicit expression for its cumulative distribution function for $s = 3$, $m = 1, 2, 3$ and $s = 4$, $m = 0, 1$, along with the corresponding values of the exact upper 5% and 1% points for several n. Mikhail [18] has derived the exact null distribution of $V^{(2)}$. Davis [4, 5] has also obtained the null distribution for $m = 2, 3$, and studied the accuracy of Pillai's beta approximation. The exact null distribution of $V^{(s)}$ is obtained by Krishnaiah & Chang [13] in the general case. Exact percentage points of $V^{(s)}$ are tabulated by Schuurmann et al. [35]; see also Anderson [1].

The limiting distribution of $\nu_2 V^{(s)}$, as $\nu_2 \to \infty$, is the **chi-square distribution** with $\nu_1 p$ df. The asymptotic expansion for the null distribution of $V^{(s)}$ is derived by Muirhead [19] and Fujikoshi [6].

Khatri & Pillai [11] derived the first two **moments** of $V^{(s)}$ in the noncentral case for MANOVA (linear alternative).

Pillai & Jayachandran [31] derived the noncentral distribution of $V^{(s)}$ for (i) and (ii) for the 2-root case, and tabulated the **power** functions for $p = 2$. Pillai [27] has obtained the moment generating function in the noncentral case for (i) and (ii), and later Khatri & Pillai [12] derived the noncentral distribution in a series form for (i). Pillai & Al-Ani [28] derived the noncentral distribution in the 2-root and 3-root cases for (iii). See also Pillai & Sudjana [33] for some numerical studies on the nonnull distribution.

Fujikoshi [6] has obtained asymptotic expansion of the nonnull distribution of $V^{(s)}$ for MANOVA and the test of independence (with local alternatives) when ν_2 is large; later, Fujikoshi [7] obtained a similar expansion in terms of noncentral chi-square distributions when ν_1 is large. For similar asymptotic expansions for the nonnull distribution, see Lee [14, 15].

Kiefer & Schwarz [10] have shown that Pillai's trace test is **Bayes** and admissible (*see* **Decision Theory**) for the MANOVA problem. It follows from the results of Olkin & Perlman [21] that this trace test is **unbiased** for problems (i) and (ii). Perlman [23] has shown that the power function of this test enjoys the monotonicity property in the respective noncentral parameters in both cases, if the cutoff point is not too large; for details, see Perlman [23]. It follows from the result of Anderson & DasGupta [2]

that the one-sided trace test for (iii) enjoys the monotonicity property of its power function when the **alternative** is one-sided; Giri [8] has shown that the one-sided trace test is locally best invariant (*see* **Hypothesis Testing**) for (iii). The locally best property of the trace test among all invariant tests is shown by John [9] for (i), (ii), and (iii). Schwarz [36] has studied the trace test and shown that it has the local **minimax** property for (i) and (ii).

Mikhail [18] found that the trace test has better power than that of any other standard test for the MANOVA problem when $p = 2$. Pillai & Jayachandran [30] compared the trace test with other standard tests for the MANOVA problem, and found that the trace test has the maximum power for small deviations from the null hypothesis when $p = 2$.

Schatzoff [34] calculated the expectation of the observed significance value of the trace test under different alternatives for the MANOVA problem, and found that the trace test is quite sensitive. On the basis of permutation distributions (*see* **Randomization Tests**), Mardia [16] has shown that the trace test is moderately robust for problem (i), but highly sensitive for (iii). On the basis of a Monte Carlo study, Olson [22] found that the trace test is most robust among the standard tests for the MANOVA problem with respect to deviations from normality and homogeneity of covariance matrices. Pillai & Sudjana [33] carried out some robustness study of the trace test for (i) and (iii) in the 2-root case; a **robustness** study for (ii) has been carried out by Pillai & Hsu [29].

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Anderson, T.W. & DasGupta, S. (1964). A monotonicity property of the power functions of some tests of the equality of two covariance matrices, *Annals of Mathematical Statistics* **35**, 1059–1063.

[3] Bartlett, M.S. (1939). A note on tests of significance in multivariate analysis, *Proceedings of the Cambridge Philosophical Society* **35**, 180–185.

[4] Davis, A.W. (1970). On the null distributions of sum of the roots of a multivariate beta distribution, *Annals of Mathematical Statistics* **41**, 1557–1562.

[5] Davis, A.W. (1972). On the distributions of the latent roots and traces of certain random matrix, *Journal of Multivariate Analysis* **2**, 189–200.

[6] Fujikoshi, Y. (1970). Asymptotic expansions of the distributions of test statistics in multivariate analysis, *Journal of Science of Hiroshima University Series A-1* **34**, 73–141.

[7] Fujikoshi, Y. (1972). Asymptotic formulas for the distributions of the determinant and trace of a noncentral matrix, *Journal of Multivariate Analysis* **2**, 208–218.

[8] Giri, N.C. (1968). On tests of equality of two covariance matrices, *Annals of Mathematical Statistics* **39**, 275–277. [Correction: **39** (1968) 1764].

[9] John, S. (1971). Some optimal multivariate tests, *Biometrika* **58**, 123–127.

[10] Kiefer, J. & Schwarz, R. (1965). Admissible Bayes character of T^2-, R^2-, and other fully invariant tests for classical multivariate normal problems, *Annals of Mathematical Statistics* **36**, 747–770.

[11] Khatri, C.G. & Pillai, K.C.S. (1967). On the moments of traces of two matrices in multivariate analysis, *Annals of the Institute of Statistical Mathematics* **19**, 143–156.

[12] Khatri, C.G. & Pillai, K.C.S. (1968). On the noncentral distributions of two test criteria in multivariate analysis of variance, *Annals of Mathematical Statistics* **39**, 215–226.

[13] Krishnaiah, P.R. & Chang, T.C. (1972). On the exact distributions of the traces of $S_1(S_1 + S_2)^{-1}$ and $S_1 S_2^{-1}$, *Sankhyā, A* **34**, 153–160.

[14] Lee, Y.S. (1971). Distribution of the canonical correlations and asymptotic expansion for distributions of certain independence statistics, *Annals of Mathematical Statistics* **42**, 526–537.

[15] Lee, Y.S. (1971). Asymptotic formulae for the distributions of a multivariate test statistic: power comparison of some multivariate tests, *Biometrika* **58**, 647–651.

[16] Mardia, K.V. (1971). The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model, *Biometrika* **58**, 105–127.

[17] Mijares, T.A. (1964). *Percentage Points of the Sum* $V_I^{(s)}$ *of s Roots* ($s = 1 - 50$). The Statistical Center, University of Philippines, Manila.

[18] Mikhail, M.N. (1965). A computation of tests of the Wilks-Lawley hypothesis in multivariate analysis, *Biometrika* **52**, 149–156.

[19] Muirhead, R.J. (1970). Asymptotic distributions of some multivariate tests, *Annals of Mathematical Statistics* **41**, 1002–1010.

[20] Nanda, D.N. (1950). Distribution of the sum of roots of a determinantal equation, *Annals of Mathematical Statistics* **21**, 432–439.

[21] Olkin, I. & Perlman, M.D. (1980). Unbiasedness of invariant tests for MANOVA and other multivariate problems, *Annals of Statistics* **8**, 1326–1341.

[22] Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

[23] Perlman, M.D. (1974). On the monotonicity of the power functions of tests based on traces of multivariate beta matrix, *Journal of Multivariate Analysis* **4**, 27–30.

[24] Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis, *Annals of Statistics* **8**, 1326–1341.

[25] Pillai, K.C.S. (1957). *Concise Tables for Statisticians*. The Statistical Center, University of Philippines, Manila.

[26] Pillai, K.C.S. (1960). *Statistical Tables for Tests of Multivariate Hypothesis*. The Statistical Center, University of Philippines, Manila.

[27] Pillai, K.C.S. (1968). On the moment generating function of Pillai's $V^{(s)}$ criterion, *Annals of Mathematical Statistics* **39**, 877–880.

[28] Pillai, K.C.S. & Al-Ani, S. (1970). Power comparisons of tests of equality of two covariance matrices based in individual characteristic roots, *Journal of the American Statistical Association* **65**, 438–446.

[29] Pillai, K.C.S. & Hsu, Y.S. (1979). Exact robustness studies of the test of independence based on four multivariate criteria and their distribution problems, *Annals of the Institute of Statistical Mathematics* **31**, 85–101.

[30] Pillai, K.C.S. & Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypothesis based on four criteria, *Biometrika* **54**, 195–210.

[31] Pillai, K.C.S. & Jayachandran, K. (1970). On the exact distribution of Pillai's $V^{(s)}$ criterion, *Journal of the American Statistical Association* **65**, 447–454.

[32] Pillai, K.C.S. & Mijares, T.A. (1959). On the moments of the trace of a matrix and approximation to its distribution, *Annals of Mathematical Statistics* **30**, 1135–1140.

[33] Pillai, K.C.S. & Sudjana (1975). Exact robustness studies of tests of two multivariate hypotheses based on four criteria and their distribution problems under violations, *Annals of Statistics* **3**, 617–638.

[34] Schatzoff, M. (1966). Sensitivity comparisons among tests of the general linear hypothesis, *Journal of the American Statistical Association* **61**, 415–435.

[35] Schuurmann, F.J., Krishnaiah, P.R. & Chattopadhyay, A.K. (1975). Exact percentage points of the distribution of the trace of a multivariate beta matrix, *Journal of Statistical Computation and Simulation* **3**, 331–395.

[36] Schwarz, R.E. (1967). Locally minimax tests, *Annals of Mathematical Statistics* **38**, 340–359.

(*See also* **Lambda Criterion, Wilks'**; **Multivariate Analysis, Overview**; **Multivariate Techniques, Robustness**).

SOMESH DASGUPTA

Pinel, Philippe

Born: April 20, 1745, in Jonquières, France.
Died: October 25, 1826, in Paris, France.

Phillipe Pinel received his training in medicine from the schools at Toulouse and Montpellier. In 1778 he went to Paris where he eventually became associated with a group of thinkers known as the Idéologues who took their inspiration, in part, from the work of the mathematician, the Marquis de Condorcet. Condorcet believed that the "calculus of probabilities" was the key to extending a scientific understanding to problems dealing with human society – what eighteenth-century thinkers referred to as the "science of man".

With the changing political climate following the French Revolution, Pinel was able to find positions in the Parisian hospitals that treated the mentally ill. In 1793 he was appointed to Bicêtre and in 1795 he was appointed to an analogous position at Salpêtrière, then a repository for the inveterately insane. From this institutional base, Pinel was able to test the success of "moral treatment" of the insane, which involved removing their physical restraints and treating them in a more humane manner. Also, Pinel was able to use the hospital setting to observe large numbers of patients, thereby facilitating statistical comparison. In a speech read before the Institut de France in 1807, Pinel [2] reported on a statistical "experiment" that he conducted to demonstrate that "moral treatment" was superior to the traditional treatment (which involved bleeding and other more "physical" interventions by the physician). As discussed in Goldstein [1], Pinel used no control group of patients who had been treated by the traditional means; however, he did establish a 93% rate of cure for maniacs and melancholics treated by "moral" methods. Two years later Pinel [3] presented a more detailed analysis of his statistical findings in *Traitè Medico-Philosophique sur l'Alienation Mentale*, in which he declared that the "fundamental principle of the calculus of probabilities" entails acquiring knowledge of the number of cases "favorable and contrary".

Thus, Pinel is important not only for the history of **psychiatry**, but also for the **history of biostatistics**. By placing emphasis on more humanitarian treatment of the insane, his work marks a major shift toward more "enlightened" treatment of the mentally ill. By placing emphasis on statistical comparison to prove efficacy, he foreshadowed a methodological approach that would become increasingly important throughout the course of the nineteenth and twentieth centuries.

References

[1] Goldstein, J. (1987). *Console and Classify: The French Psychiatric Profession in the Nineteenth Century*. Cambridge University Press, Cambridge.

[2] Pinel, P. (1807). Résultats d'observation et construction des tables pour servir à déterminer le degré de probabilité de la guérison des aliénés, *Institut de France, Mémoires de la Classe des Sciences Mathèmatique et Physique* **8**, 169–205.

[3] Pinel, P. (1809). *Traité Medico-Philosophique sur l'Alienation Mentale*, 2nd Ed. Paris.

J. ROSSER MATTHEWS

Pitman Efficiency

The Pitman efficiency is an index for comparing test procedures or estimators. It is of special importance for comparing procedures in large samples. Given two procedures, $\mathcal{P}_1$ and $\mathcal{P}_2$, with a particular definition of "goodness" of the procedures, if $\mathcal{P}_1$ requires n_1 observations and $\mathcal{P}_2$ requires n_2 observations to achieve the same level of "goodness", the **efficiency** of $\mathcal{P}_1$ relative to $\mathcal{P}_2$ is rel. eff. $(\mathcal{P}_1, \mathcal{P}_2) = n_2/n_1$. This is the basic idea in the Pitman relative efficiency index. We illustrate this first with two examples.

Example 1

Suppose that $X_1, \ldots, X_n$ are independent identically distributed (iid) **random variables** (a **random sample**) from a **lognormal distribution** $\ln(\mu, \sigma^2)$ (i.e. $\log X_1 \sim \mathrm{N}(\mu, \sigma^2)$). The expected value of X_1 is $\xi = \exp(\mu + \sigma^2/2)$. Assume that μ is unknown and σ^2 is known. An unbiased estimator of ξ is the sample mean $\overline{X}_n$, having a variance $\mathrm{var}\{\overline{X}_n\} = [w_1(\mu, \sigma^2)]/n$, where $w_1(\mu, \sigma^2) = \mathrm{var}\{X_1\} = \xi^2[\exp(\sigma^2) - 1]$. The **maximum likelihood** estimator (MLE) is $\hat{\xi}_n = \exp(\overline{Y}_n + \sigma^2/2)$, where $Y_i = \log X_i (i = 1, \ldots, n)$, $\overline{Y}_n = (1/n)\sum_{i=1}^{n} Y_i$. The MLE, $\hat{\xi}_n$, is a **biased** estimator of ξ, having the **mean square error** (MSE)

$$\begin{aligned}\mathrm{MSE}\{\hat{\xi}_n\} &= \xi^2\left[\exp\left(\frac{2\sigma^2}{n}\right) - 2\exp\left(\frac{\sigma^2}{2n}\right) + 1\right] \\ &= \frac{\sigma^2}{n}\xi^2\left(1 + \frac{7}{4}\frac{\sigma^4}{n} + \cdots\right) \\ &= \frac{\sigma^2}{n}\xi^2 + O\left(\frac{1}{n^2}\right).\end{aligned}$$

Let n_1 be the value of n for which $\mathrm{MSE}\{\hat{\xi}_n\} = c$ and let n_2 be the value of n for which $\mathrm{var}\{\overline{X}_n\} = c$. For small values of c we need large values of n_1 and n_2, and the index of relative efficiency of the MLE vs. $\overline{X}_n$ is approximately

$$\frac{n_2}{n_1} \approx \frac{\exp(\sigma^2) - 1}{\sigma^2} = 1 + \frac{\sigma^2}{2} + \frac{\sigma^4}{6} + \cdots.$$

This shows that $\overline{X}_n$ is very inefficient, compared with the MLE, if σ^2 is large. This result is intuitively clear, since $\overline{X}_n$ is not a **sufficient statistic** for this family of distributions.

Example 2

Consider a random sample $X_1, \ldots, X_n$, from a **normal distribution**, $\mathrm{N}(\mu, \sigma^2)$, with known variance σ^2 and unknown mean μ. In quality control problems we are often interested in testing whether $p = \Pr\{X_1 \geq \xi_0\}$ is greater than some value p_0. Both p_0 and ξ_0 are specified constants. This problem is equivalent to that of testing the **null hypothesis** $\mathrm{H}_0 : \mu \geq \mu_0$ against the **alternative** $\mathrm{H}_1 : \mu < \mu_0$, where $\mu_0 = \xi_0 - Z_{1-p_0}\sigma$ and where Z_{1-p_0} is the $(1 - p_0)$th **quantile** of $\mathrm{N}(0, 1)$. It is well known that the uniformly **most powerful** (UMP) test of H_0 vs. H_1, at level of significance α, is given by the test function

$$\phi_1(\overline{X}_n) = \begin{cases} 1, & \text{if } \overline{X}_n \leq \mu_0 - Z_{1-\alpha}\sigma/\sqrt{n}. \\ 0, & \text{otherwise.} \end{cases}$$

The **power** of this test is given by the function,

$$\psi_n^{(1)}(\mu_1) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma}\sqrt{n} - Z_{1-\alpha}\right), \quad \text{for } \mu_1 \leq \mu_0,$$

where $\Phi(\cdot)$ is the standard normal distribution function. Clearly, for each $\mu_1 < \mu_0$, $\lim_{n\to\infty} \psi_n(\mu_1) = 1$. Moreover, $\psi_n^{(1)}(\mu_1)$ is a strictly increasing function of n, and the value of n for which $\psi_n^{(1)}(\mu_1) \geq 1 - \beta$ is

$$n_1 = \min\left\{n : n \geq \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{(Z_{1-p_1} - Z_{1-p_0})^2}\right\}.$$

Note that in terms of the original testing problem, μ_1 corresponds to $p_1 = P_{\mu_1}\{X \geq \xi_0\}$, or $\mu_1 = \xi_0 - \sigma Z_{1-p_1}$. Obviously, if $p_1 < p_0$, then $\mu_1 < \mu_0$.

An alternative test procedure is to count the number, K_n, of items in the sample having values greater than ξ_0 (number of conforming items). K_n has a **binomial distribution** $B(n, p)$. We consider the test of the null hypothesis $\mathrm{H}_0 : p \geq p_0$, against the alternative $\mathrm{H}_1 : p < p_0$. The UMP test, based on K_n, of level α, is

$$\phi_2(K_n) = \begin{cases} 1, & \text{if } K_n < C_\alpha, \\ \gamma_\alpha, & \text{if } K_n = C_\alpha, \\ 0, & \text{if } K_n > C_\alpha, \end{cases}$$

in which $C_\alpha = B^{-1}(\alpha; n, p_0)$, and $\gamma_\alpha = [\alpha - B(C_\alpha - 1; n, p_0)]/[b(C_\alpha; n, p_0)]$. Here $B^{-1}(\alpha; n, p_0)$ denotes the α-quantile of $B(n, p_0)$; $B(C_\alpha; n, p_0)$ and

$b(C_\alpha; n, p_0)$ are the cumulative distribution function (cdf) and probability density function (pdf) of $B(n, p_0)$, respectively, at C_α. The power function of ϕ_2 is

$$\psi_n^{(2)}(p_1) = P_{p_1}\{K_2 < C_\alpha\} + \gamma_\alpha P_{p_1}\{K_2 = C_\alpha\}$$
$$= B(C_\alpha - 1; n, p_1) + \gamma_\alpha b(C_\alpha; n, p_1).$$

In large samples,

$$C_\alpha \doteq np_0 - Z_{1-\alpha}[np_0(1-p_0)]^{1/2}$$

and

$$\psi_n^{(2)}(p_1) \cong \Phi\left(\sqrt{n}\frac{p_0 - p_1}{[p_1(1-p_1)]^{1/2}} - Z_{1-\alpha}\left[\frac{p_1(1-p_1)}{p_0(1-p_0)}\right]^{1/2}\right).$$

Let n_2 denote the smallest n needed so that $\psi_n^{(2)}(p_1) \geq 1 - \beta$. We have

$$n_2 \doteq \frac{(Z_{1-\alpha}[p_0(1-p_0)]^{1/2} + Z_{1-\beta}[p_1(1-p_1)]^{1/2})^2}{(p_0 - p_1)^2}.$$

The ratio of n_2 to n_1 gives the relative efficiency of $\phi_1(\overline{X}_n)$ vs. $\phi_2(K_n)$. This is given by

$$\frac{n_2}{n_1} \simeq \frac{(Z_{1-\alpha}[p_0(1-p_0)]^{1/2} + Z_{1-\beta}[p_1(1-p_1)]^{1/2})^2}{(Z_{1-\alpha} + Z_{1-\beta})^2} \times \frac{(Z_{1-p_1} - Z_{1-p_0})^2}{(p_0 - p_1)^2}.$$

For $\alpha = \beta = 0.05$ and $p_0 = 0.95$, we obtain the relative efficiency values shown in Table 1. These results are not surprising, since the binomial testing is a **nonparametric** (distribution-free) method, while the test ϕ_1 is based on the minimal sufficient statistic, $\overline{X}_n$ for the normal model.

Pitman [5] introduced an index of **asymptotic relative efficiency** for tests and for estimators. An exposition of the theory can be found in Lehmann's book on estimation [4] and in Pitman's book [6]. A more general approach was given by Hoeffding & Rosenblatt [3]. We now present the essential results.

Table 1

p_1	n_2/n_1
0.945	4.33
0.94	4.21
0.935	4.10
0.93	4.00
0.925	3.91
0.92	3.82

Asymptotic Relative Efficiency of Test Procedures

Let $X_1, X_2, \ldots$ be a sequence of iid random variables having a common cdf $F_\theta(x)$, which depends on a real parameter θ. Let $T_n = t(X_1, \ldots, X_n)$ be a real-valued statistic, such that

$$Z_n = \frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \xrightarrow[n \to \infty]{\text{d}} \text{N}(0, 1)$$

(*see* **Convergence in Distribution and in Probability**). It is often the case that $\sigma_n(\theta) = c(\theta)w_n$, where $c(\theta) > 0, w_n > 0$ and $w_n \downarrow 0$ and $n \to \infty$. Typically, $w_n = n^{-\alpha}$ or $w_n = (n \log n)^{-\alpha}$ for $\alpha > 0$. Moreover, $\mu_n(\theta) \to \mu(\theta)$ as $n \to \infty$. The problem is to test $\text{H}_0 : \theta \leq \theta_0$ against $\text{H}_1 : \theta > \theta_0$ at level of significance α (*see* **Level of a Test**). Consider a sequence of test procedures, which reject H_0 whenever $(T_n - \mu(\theta_0))/\sigma_n(\theta_0) \geq k_n$, where $k_n \to Z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. The power functions corresponding to this sequence of tests are, for some $\theta > \theta_0$,

$$\psi_n(\theta; T_n) = P_\theta\{T_n \geq \mu(\theta_0) + k_n\sigma_n(\theta_0)\}$$
$$\cong \Phi\left(\frac{\mu(\theta) - \mu(\theta_0)}{c(\theta_0)w_n}\frac{c(\theta_0)}{c(\theta)} - Z_{1-\alpha}\frac{c(\theta_0)}{c(\theta)}\right).$$

We assume the following:

1. $\mu(\theta)$ has a continuous derivative, $\mu'(\theta)$, in a neighborhood of θ_0 and $\mu'(\theta) > 0$;
2. $c(\theta)$ is continuous in a neighborhood of θ_0, and $c(\theta_0) > 0$.

Notice that under these assumptions, for $w_n \downarrow 0$ as $n \to \infty$, for a fixed θ in the required neighborhood of θ_0, $\lim_{n\to\infty} \psi_n(\theta; T_n) = 1$. Thus, the above large sample test is **consistent**. We consider its limiting power, for a sequence of alternatives $\theta_n \to \theta_0$. More specifically, let $\theta_n = \theta_0 + \delta w_n$ with $\delta > 0$. Then, for

large values of n, $\mu(\theta_n) \approx \mu(\theta_0) + \delta w_n \mu'(\theta_0)$ and $c(\theta_0)/c(\theta_n) \to 1$. The asymptotic power is

$$\lim_{n\to\infty} \psi_n(\theta_n; T_n) = \Phi\left(\frac{\delta\mu'(\theta_0)}{c(\theta_0)} - Z_{1-\alpha}\right) = \psi^*,$$

where $0 < \psi^* < 1$.

The function

$$J(\theta; T) = \frac{[\mu'(\theta)]^2}{[c(\theta)]^2}$$

is called the *asymptotic efficacy* of T_n (see [6, p. 351]). Let $\{V_n\}$ be an alternative sequence of test statistics, having asymptotic efficacy $J(\theta; V) = [\eta'(\theta)]^2/[d(\theta)]^2$. Consider the case in which $w_n = n^{-1/2}$ for *both* $\{T_n\}$ and $\{V_n\}$. In this case, the sample size n' required for attaining the same limiting power ψ^* is such that

$$\lim_{n\to\infty} \frac{n}{n'(n)} = \frac{J(\theta_0; V)}{J(\theta_0; T)} = \left(\frac{\eta'(\theta_0)}{\mu'(\theta_0)}\right)^2 \frac{c^2(\theta_0)}{d^2(\theta_0)}.$$

This limit is called the Pitman asymptotic relative efficiency (ARE) of $\{V_n\}$ compared with $\{T_n\}$, and denoted by $\text{eff}(\theta_0; V_n, T_n)$. More generally, if $w_n = n^{-\alpha}$, $\alpha > 0$, then the Pitman ARE is

$$\text{eff}(\theta_0; V_n, T_n) = \left(\frac{J(\theta_0, V)}{J(\theta_0, T)}\right)^{1/2\alpha}.$$

Note that if Z_n and $W_n = (V_n - \eta_n(\theta))/v_n(\theta)w_n$ do not have the same asymptotic distribution then the ARE index is not defined.

Example 3

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $G(x;\theta) = F(x - \theta)$, where θ is the **median** of $G(x;\theta)$; that is, $\theta = G^{-1}(\frac{1}{2};\theta)$. We assume that $F(x)$ is symmetric around zero; that is, $F^{-1}(\frac{1}{2}) = 0$. We further assume that $F \in \Omega_s$, where Ω_s is the family of symmetric distributions, which are absolutely continuous, with pdf $f(x)$, having a continuous derivative $f'(x)$. Let $R_j (j = 1, \ldots, n)$ be the **ranks** of $Y_j = |X_j|$ and let $s(x) = I\{x > 0\}$ be the sign of x. A **score** statistic for testing the hypothesis $\text{H}_0 : \theta \le 0$ vs. $\text{H}_1 : \theta > 0$ is

$$\overline{V}_n = \frac{1}{n}\sum_{j=1}^{n} \phi\left(\frac{R_j}{n+1}\right) s(X_j),$$

where $\phi(\cdot)$ is a score function; that is, $\phi(u)$ is nondecreasing on (0, 1) $\phi(u) \ge 0$ and $\int_0^1 \phi(u)\,\text{d}u < \infty$. A large class of nonparametric test statistics can be expressed as such score functions. As shown by Hettmansperger [2, p. 105], the efficacy of V_n, for $\theta_0 = 0$, is

$$J(0, V_n) = \frac{\left[\int_0^1 \phi(x) \dfrac{f'\{F^{-1}[(x+1)/2]\}}{f\{F^{-1}[(x+1)/2]\}}\,\text{d}x\right]^2}{\int_0^1 \phi^2(x)\,\text{d}x}.$$

Using this function, one can obtain the Pitman ARE of two different score statistics, when $f(x)$ is a specified density, symmetric around zero.

Asymptotic Relative Efficiency of Estimators

Let $\mathcal{F}$ be a family of distributions depending on a real parameter θ. Consider the problem of estimating a real parameter $w(\theta)$. Given a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$, let $\hat{w}_1(\mathbf{X}_n)$ and $\hat{w}_2(\mathbf{X}_n)$ be two estimators of $w(\theta)$. We assume that:

1. $E_\theta\{\hat{w}_i(\mathbf{X}_n)\} = w_{i,n}(\theta), i = 1, 2$; and
2. $V_\theta\{\hat{w}_i(\mathbf{X}_n)\} = \sigma^2_{i,n}(\theta), i = 1, 2$.

The efficacy of $\hat{w}_i(\mathbf{X}_n)$ is defined as

$$J_n(\theta; \hat{w}_i) = \frac{[w'_{i,n}(\theta)]^2}{\sigma^2_{i,n}(\theta)}.$$

The relative efficiency of $\hat{w}_2$ compared with $\hat{w}_1$ is defined as (see [1])

$$\text{RE}_n(\theta; \hat{w}_2, \hat{w}_1) = \frac{J_n(\theta; \hat{w}_2)}{J_n(\theta; \hat{w}_1)}.$$

The asymptotic relative efficiency is

$$\text{ARE}(\theta; \hat{w}_2, \hat{w}_1) = \lim_{n\to\infty}\left[\frac{J_n(\theta; \hat{w}_2)}{J_n(\theta; \hat{w}_1)}\right].$$

If both estimators are asymptotically unbiased, in other words, $w_{i,n}(\theta) \to w(\theta)$ as $n \to \infty, i = 1, 2$ and if $\sigma^2_{i,n}(\theta) = [\sigma^2_i(\theta)]/n + o(1/n)$ as $n \to \infty$, then

$$\text{ARE}(\theta; \hat{w}_2, \hat{w}_1) = \frac{\sigma^2_1(\theta)}{\sigma^2_2(\theta)}.$$

If the family $\mathcal{F}$ satisfies the Cramér–Rao regularity conditions (see [7]), the mean square error of an estimator $\hat{w}(\theta)$ is often compared to the **Cramér–Rao** lower bound for variances of unbiased estimators of $w(\theta)$, and the relative efficiency is defined as

$$\mathrm{RE}_n(\theta; \hat{w}) = \frac{w'(\theta))^2}{nI(\theta)\mathrm{MSE}_\theta\{\hat{w}\}},$$

where $I(\theta)$ is the Fisher **information** function. Thus, if $\mathrm{MSE}_\theta\{\hat{w}\} = \sigma^2(\theta)/n + o(1/n)$ as $n \to \infty$, then $\lim_{n\to\infty} \mathrm{RE}_n(\theta; \hat{w}) = [w'(\theta)]^2/I(\theta)\sigma^2(\theta)$.

Example 4

Consider the problem of Example 1, for estimating the mean ξ of a lognormal distribution $\ln(\mu, \sigma^2)$, with σ^2 known. Here, $w(\mu) = \exp(\mu + \sigma^2/2)$, and the Fisher information is $I(\mu) = \sigma^2$. Thus, the relative efficiency of the unbiased estimator $\overline{X}_n$ is

$$\mathrm{RE}_n(\mu; \overline{X}_n) = \frac{[\exp(\mu + \sigma^2/2)]^2\sigma^2}{nV\{\overline{X}_n\}} = \frac{\sigma^2}{\exp(\sigma^2) - 1}.$$

In this case $\mathrm{RE}_n(\mu, \overline{X}_n) < 1$ for all σ^2, and $\mathrm{RE}_n(\mu, \overline{X}_n) \to 0$ as $\sigma^2 \to \infty$.

If the distribution of X depends on k parameters, $\theta_1, \ldots, \theta_k, k > 1$, the definition of relative efficiency of estimators of $w(\theta_1, \ldots, \theta_k)$ is more complicated. The reader is referred to Zacks [8].

References

[1] DeGroot, M.H.& Raghavachari, M. (1970). Relations between the Pitman efficiency and Fisher information, *Sankhyā* **32**, 314–324.

[2] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

[3] Hoeffding, W.& Rosenblatt, J.R. (1955). The efficiency of tests, *Annals of Mathematical Statistics* **26**, 52–63.

[4] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[5] Pitman, E.J.G. (1948). *Notes on Nonparametric Statistical Inference*. Institute of Statistics, University of North Carolina, Chapel Hill, unpublished.

[6] Pitman, E.J.G. (1979). *Some Basic Theory of Statistical Inference*. Chapman& Hall, London.

[7] Zacks, S. (1980). *Parametric Statistical Inference: Basic Theory and Modern Approaches*. Pergamon Press, Oxford.

[8] Zacks, S. (1985). Pitman efficiency, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz& N.L. Johnson, eds. Wiley, New York, pp. 731–735.

S. ZACKS

Placebos

History of Placebo

The placebo, which originates from Latin meaning "I shall please", was used for centuries as therapy for patients whom physicians were unsure of how to treat or for whom no useful treatment was available [26]. Attitudes towards therapies in Western cultures have shifted dramatically over the last four decades. Principles of safety, efficacy, and informed consent of participants are now well entrenched (*see* **Ethics of Randomized Trials**); the randomized controlled trial (*see* **Clinical Trials, Overview**) dominates as the method for evaluating therapy and patients are increasingly involved in decisions about their care [38]. The role of the placebo has evolved in response to all of these factors. In research, the notion of placebo control arose to account for those beneficial or harmful effects not directly attributable to the therapy of interest. Despite some dissent, placebos and placebo-controlled trials remain the benchmark by which all new drugs are evaluated and regulated [50]. In clinical care, any part of an encounter can have some therapeutic value and influence the patient's response, including taking a history, stating a diagnosis, or making repeated measurements (e.g. blood pressure) or assurances about prognosis. Knowledgeable clinicians are interested in determining which parts of the placebo effect they should implement to optimize their patients' health.

Definitions

Many, similar definitions for "placebo", "placebo effect", and "placebo response" exist. Here, we use the word *placebo* to mean an inert substance or sham procedure designed to appear identical to the active substance or procedure but without known therapeutic effect. *Placebo effect* or *placebo response* is the psychophysiologic effect associated with placebos, thought to be primarily operative through the expectations or symbolic meaning of the administered therapy [5, 36, 39]. This implies both positive and negative effects, although the latter are often referred to as the *nocebo effect* [19]. More recently, the placebo effect has been further characterized as a *true* placebo effect versus a *perceived* placebo effect. The perceived placebo effect, which is the effect commonly quoted and discussed in the literature, includes the true placebo effect plus other nonspecific effects, including the natural course of disease, **regression to the mean**, unidentified co-intervention effects, and other time-dependent effects [15, 39]. The natural course of any disease or symptom will change in severity over time and may resolve on its own (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**). Regression to the mean, where a follow-up measurement of a patient's disease or symptoms gives a more normal reading, is a well-recognized phenomenon. Patients enrolled in a trial may influence clinical outcomes by implementing unidentified parallel interventions (co-intervention), for example, by starting a weight-loss program while in a trial of a diabetes drug. Time-dependent effects might include the increasing skill of the investigator in measuring study endpoints, or the decreasing "white coat hypertension" effect as patients become used to having their blood pressure measured. Since these nonspecific effects are inherent in any therapy (i.e. active intervention, placebo, or no-treatment groups), it has been argued that we should be focusing on the true placebo effect as a measurement target [15, 23].

Magnitude of the Placebo Effect

For decades, the placebo response has been assumed to be similar across disease categories at approximately 35% improvement from baseline [3]. This figure has been applied equally to the proportion of patients improving and to the degree of improvement per patient per outcome. More recently, doubt has been cast on the constant placebo response as evidence accumulates that both the true and perceived placebo effects are variable [15, 36, 55]. A review of 75 trials of antidepressant medications revealed that the placebo response rate varied from 12.5 to 51.8% and had increased over time [55]. A recent systematic review of randomized trials with placebo and no-treatment arms (*see* **Meta-analysis of Clinical Trials**), concluded that there was evidence of a mild true placebo effect for some subjective outcomes (e.g. pain and anxiety) but not for more objective outcomes (blood pressure, weight loss, asthma outcomes) [23]. For example, placebo was associated with a reduction in **pain** by a mean of 0.65 cm

on a 10 cm visual analog scale, as compared with no treatment. The authors measured the difference in outcomes between the two arms at the end of the treatment period rather than the change in outcomes from baseline, thus attempting to control for the nonspecific effects and determine the true placebo effect [22]. Their concluding caution against the use of the placebo and its effects for therapeutic purposes outside of a controlled clinical trial raised a storm of criticism. Although the validity of the findings were questioned, in terms of widely varying populations and diseases, heterogeneity and low statistical **power** [1, 46], wrong choice of placebo [27], and contamination of the "no-treatment" arms [14], in our opinion, it is really the generalizability of the findings that is in question (*see* **Validity and Generalizability in Epidemiologic Studies**). Within randomized controlled trials, where patients acknowledge by their consent that they know they may not receive an active treatment, the perceived (total) placebo effect may be more muted than in clinical practice. Furthermore, even within randomized trials, unblinding can occur and **bias** clinician and patient – a situation bound to happen with a no-treatment arm. Finally, and most important, a finding that placebo therapies and no-treatment arms do not differ does not negate the possibility that a true placebo effect occurred (and occurred equally) in both groups.

It is important to keep in mind that placebos are not risk free. Placebos may be harmful if they delay access to effective therapy for the disease under investigation, if their nonspecific symptomatic effects mask a condition that has effective treatment or via direct nocebo effect [1]. Where placebos are used without patient consent, any revelation of the deception may seriously undermine patient–physician relationship, which is itself a powerful source of "placebo effect".

Influences on the Placebo Effect

The attitude and behavior of the clinician toward the treatment and the patient, the attitude of the patient toward her own health and the treatment, as well as external, cultural, family, and media influences, all influence the placebo effect [5, 13, 22, 28, 35, 36]. Treatment variables including appearance, invasiveness, impressiveness, perceived plausibility, past experience, and cost all appear to play a role.

Provider factors can produce major placebo effects and, although not well studied, are considered integral to the "art of medicine". The white coat of the clinician, the stethoscope, the interest, empathy, authority, and compassion displayed in the interview with the patient and the motivation and skillfulness with which a diagnosis or therapeutic path is pursued, each may influence the patient [5, 36, 39]. In a study of 200 British patients presenting to a physician with abnormal symptoms for which no firm diagnosis could be made, the patients were randomly assigned to a negative or positive consultation, and placebo treatment or no treatment [7]. The positive consultation consisted of a firm diagnosis and reassurance that the patient would be better in a few days while in the negative consultation, the physician confessed uncertainly. Two weeks later, 64% of those who received a positive consultation reported improvement compared with 39% of those who received a negative consultation. The percentage of the placebo treated group who improved was not significantly different than the percentage in the untreated group.

Patient expectations, prior experiences with similar therapies, severity of current complaints, and suggestibility likely play a role in the placebo effect but have not been rigorously studied [2, 17, 29, 34, 36, 41, 42]. In a 10-week study of exercise, 48 healthy young adults were randomized to a control aerobics training program or an exercise program with constant reminders of the aim to improve both aerobic capacity and psychological well-being [13]. After 10 weeks, significant increases in fitness levels were seen in both groups, however, self-esteem and psychological well-being were only significantly improved in the experimental group, not the control group. Cultural differences in placebo responsiveness have also been explored. A review of 117 studies of ulcer treatment, and 37 studies of treatment for anxiety, showed Germany had the highest placebo healing rates for ulcers, but experienced only moderate placebo rates for treatments of anxiety. In comparison, Italy had the lowest placebo effects for anxiety, while Brazil had the lowest rates for ulcers [35]. Similarly, a retrospective analysis of deaths attributed to lymphatic cancer found that Chinese-Americans who were born in "Earth years", and therefore deemed by Chinese medical history to be especially susceptible to diseases involving lumps, nodules, or tumors, had a lower mean age of death than those born in other years. No

such relationship could be seen in Caucasian controls who died of similar causes [35, 36].

Patients who enter into trials compared with those who do not and patients who adhere to their treatment, whether placebo or not, compared with those who do not (*see* **Compliance Assessment in Clinical Trials**), may have different outcomes. In a large study of beta-blockers to prevent a recurrence of a myocardial infarction, it was found that those who took more than 80% of their medications had a lower mortality rate compared to poor adherers whether they received beta-blockers or placebo [35, 39]. A **Cochrane** methods review pilot looked at patients enrolled in phase III randomized controlled trials (*see* **Clinical Trials, Overview**) versus patients who were similar but not enrolled. In 23 out of 25 reports, better outcomes were documented for patients within the trials compared to those who were not. Overall, there were lower mortality and lower rates for complications of therapy [21]. An analysis of the CAMIAT (Canadian Amiodarone Myocardial Infarction Arrhythmia Trial) study [24] showed that adherence to placebo or amiodarone, both predicted mortality. Patients who received placebo and were considered noncompliant had over a two-fold increased risk of sudden cardiac death, cardiac mortality, and all-cause mortality compared to those placebo patients who were compliant.

Drug and treatment characteristics, apart from the pharmacology of the active ingredient(s), can have powerful placebo effects. Studies suggest that subjects react to the "meanings" or suggestion of the color and quantity of drugs. Red suggests "hot" or "danger", blue suggests "down" or "quiet", and a quantity of two means "more than one" [6, 12, 35]. A group of medical students were told that they were participating in a single-blinded study (*see* **Blinding or Masking**) on the psychological and physiologic effects of two drugs, a sedative and a stimulant, and received one or two placebo capsules of either blue or pink without attribution of any effect [6]. The consent form they were asked to sign gave a brief description of the stimulant or sedative side effects they may expect to experience. Study results showed that the students tended to experience stimulant reactions to the pink capsules, while the blue capsules produced depressant effects. Two capsules tended to produce a greater effect than one for both psychological changes (e.g. drowsiness) and physiologic changes (e.g. pulse) [37]. In other words, analgesia associated with an injection of saline solution was reversed with an opiate antagonist such as naloxone and enhanced with an opiate agonist such as proglumide, which suggests that these patients were experiencing a physiologic response such as the release of endogenous opioids [4, 35]. Furthermore, the placebo response can be shown to produce a typical **pharmacokinetic** profile of activity [56].

Surgery and other mechanical or invasive procedures have been thought to produce exaggerated placebo responses, ever since the original studies of sham internal mammary artery ligation for the treatment of angina, where 80% of patients in the sham treatment arm reported substantial improvement [9]. More recently, the value of an extremely common procedure, arthroscopic lavage and debridement in patients with osteoarthritis of the knee, was questioned when it was found no better than sham surgery for outcomes of pain and physical function [37]. Reviews of trials examining placebo or comparing different placebo routes indicate that injections are more powerful than pills, and devices or procedures, such as sham ultrasound or sham acupuncture seem to be associated with stronger placebo effects than oral placebos [15, 28]. However, the extent to which the response is influenced by the procedure administrator versus the procedure itself remains unclear.

Ethics of Employing Placebo in Research

Many questions have been raised about the ethics of using placebos in human research. Strong critics of the placebo argue that it is not ethical to assign subjects to any intervention that has even the potential of being less efficacious than current therapy (*see* **Ethics of Randomized Trials**). The opposing view asserts that very few "standard" treatments have been proven effective by today's research standards, that placebo-controlled trials are a necessary first step with the smallest sample size to establish whether further research with a drug or treatment is warranted and that patients tend to improve within these trials regardless of allocation because of the close attention and follow-up. Certain situations, such as life-threatening conditions where a proven effective treatment exists, are agreed to be inappropriate for placebo-only allocation [50]. One of the areas of medicine where the use of placebo has been debated widely is in depression. Critics worried that randomization to placebo might lead

to serious harm including suicide. When this was reviewed, not only were completed and attempted suicide rates similar for placebo, standard antidepressants (imipramine, amitriptyline, trazodone), and the investigational antidepressants (fluoxetine, sertraline, paroxetine), but also the clinical responses were not very different among the three groups either [30]. Another large review noted that the response to placebo in depressed patients was highly variable (10–50%) between trials, and has increased over the years with increasing trial length [55]. All of these factors suggest that the use of active treatment controls instead of placebo would make accurate evaluation of interventions difficult.

Although it has been well argued that research subjects, by way of informed consent, indicate their willingness to freely participate in a placebo-controlled trial, there is some evidence that patients may not be fully informed about risks and benefits of the treatment options [25, 43]. Studies of consent documents have shown that they sometimes overstate the benefits and understate the risks of research protocols and are frequently written beyond the reading level of patients [11, 33, 49]. Even when consent forms are accurate and appropriate, patients may confuse treatment in a clinical trial with that of individualized medical care, overestimate the benefits of participating in a trial, and underestimate the risks they may be involved in [20]. A recent survey of patients participating in cancer trials showed that while they were satisfied with the consent form and considered themselves well informed, a large percentage of these patients did not recognize that the treatment being described was not a standard therapy, that there was potential risk to themselves, and that the study drug was unproven [25]. This research suggests that some of the placebo effect is already evident at this early stage before any treatment is actually given – the trust in trial clinicians, the hope for therapeutic benefit, and so on may color the patient's memory of consent.

Guidelines for the Use of Placebos in Research

Several prominent guidelines provide some direction for the use of placebos in research. All agree on basic principles such as noncoercion and fully informed consent for participants. In other matters, the guidelines differ somewhat. The Declaration of Helsinki has the strictest guidelines and is often cited by critics of the placebo. First ratified by the World Medical Association (WMA) in 1964, it cites "The benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic and therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where no proven prophylactic, diagnostic, or therapeutic methods exists" [57]. However, its stand has been criticized as too strict as "best ... methods" may not be the same as "best available" or "most cost-effective" comparators [40] and many standard therapies have not been rigorously compared to determine which is the best. The International Conference on Harmonization (ICH), a committee with representatives from the drug industry and regulatory authorities internationally, has published ICH E10 – Guideline for Choice of Control Group and Related Issues in Clinical Trials [53]. These **guidelines**, citing the unique scientific usefulness and general safety of placebos, allow the use of placebo controls even when effective treatment exists unless subjects would be exposed to an unacceptable risk of death or permanent injury or if the toxicity of standard therapy is so severe that "many patients have refused to receive it". ICH E10 advocates the use of modified study designs whenever possible, such as add-on studies, **factorial designs**, or "early escape" from ineffective therapy [53].

Various national regulatory and research bodies have formulated their own guidelines. The Tri-Council Policy Statement (TCPS) from Canada's three main research granting agencies, requires that grant recipients abide by the Declaration of Helsinki, prohibiting the use of placebo-controlled trials if there is an existing effective therapy available. However, it allows for broad exemptions, for example, for exceptional circumstances where effective treatment is not available to patients due to cost constraints or short supply, use in refractory patients, for testing add-on treatment to standard therapy, where patients have provided an informed refusal of standard therapy for a minor condition for which patients commonly refuse treatment, or when withholding such therapy will not lead to undue suffering or the possibility of irreversible harm [51]. The US **Food and Drug Administration's** Code of Federal Regulations (CFR), revised in April 2002, accepts placebo-controlled trials and does not stipulate any restrictions but strongly advocates the need for informed

consent [10, 18]. The primacy of informed patient consent in placebo-controlled trials has spurred a number of critiques [25, 32, 48], challenges [52], and recommendations to improve processes of obtaining consent [11, 31, 44, 47, 54].

Patients themselves have become involved. The National Depressive and Manic-depressive Association, a large patient-directed illness-specific organization in the United States, recently developed a consensus development panel to discuss the controversy regarding the use of placebo [8]. In their guidelines, they acknowledge that mood disorders are episodic, chronic conditions that are associated with considerable morbidity and have no curative or fully preventative treatments. Despite the Declaration of Helsinki, it was agreed that mood disorder research was not at the point where noninferiority trials (trials designed only to rule out that the new therapy is worse than the control) (*see* **Equivalence Trials**) involving active controls could be considered scientifically valid designs. Therefore, the guidelines cite "placebo is justified when testing a new antidepressant with a novel mechanism of action that has a substantial probability of efficacy with an acceptable adverse effect risk. However, placebo is also ethical in studies of new drugs in a class because the newer members may offer important advantages over the original drug."

Despite the guidelines that currently aid researchers regarding the use of placebos in research, there are ongoing issues that need to be addressed. One could debate the "proof" of efficacy compared to placebo for many drugs in current use, such that the counterargument has been made that it might be unethical to insist on the use of these drugs in control groups. The Declaration of Helsinki's view on placebo was established in an attempt to deter pharmaceutical companies and research organizations from exploiting people in poorer populations, who may not have access to proven treatments. This concern is not relevant in many countries. Advocating the use of placebos only when no standard treatment exists leaves the efficacy of drugs in certain patient groups largely unknown. Many clinicians need to know whether a therapy is "better than nothing" and could therefore be an alternative for patients who do not respond to the conventional treatment or cannot tolerate it [45]. In chronic diseases such as hypertension, diabetes, and vascular disease, combinations of therapies are increasingly prevalent. It is likely a more efficient assessment of a new drug to test it against placebo initially before embarking on multiple, more complicated, larger sample size, drug add-on trials.

Innovations to Improve Research Involving Placebo

Apart from improving the process of informed consent as described above, several themes of innovative trial design are developing. All attempt to minimize patient exposure to placebo only or increase the efficiency of the design to lower sample size. The first design, now commonly used, is the add-on trial where one group gets both the standard therapy and the new therapy while the other group gets the standard therapy only. The difference between the mean response of the combination standard/new therapy group and the mean of the standard therapy group alone is a reasonable estimate of effect of the new therapy provided the two therapies do not interact with each other.

The second design was developed for antidepressant drug trials to address the high placebo response rates [55]. This two-phase randomized **crossover** method initially randomizes more patients to placebo than active therapy and, in the second phase, crosses over only placebo nonresponders [16]. It is intended to minimize the overall placebo response rate and thus the sample size required to show a clinically important difference (*see* **Sample Size Determination for Clinical Trials**).

The third design avoids the bias of patient selection according to placebo response. A 2^2 factorial design involving standard and new therapies has four groups: double placebo, new therapy, standard therapy, and combination standard therapy/new therapy. With a $1:1:1:1$ allocation ratio, there are two estimates of the efficacy of the new therapy–the difference between the new therapy group mean and the double placebo group mean, and the difference between the combination standard therapy/new therapy group mean and the standard therapy group mean. These two estimates are pooled together to measure the overall efficacy of the new therapy by computing the mean of the two estimates. These two estimates should be similar to each other provided the new therapy does not **interact** with the standard therapy. However, one half of the patients are given either the double placebo or the new therapy of yet unproven efficacy. This may present an ethical problem if the

standard therapy has excellent efficacy and clinicians are reluctant to deny patients access to it. If the investigator were to drop these two groups from the third design, the result is the first, add-on design, and the measure of interaction between the two therapies has to rely on theory, not data.

A fourth design is possible where there are two known-to-be effective therapies available and where some clinicians might use the first standard, some might use the second standard, and some might use both standards together. Here, a 2^3 factorial design should be considered: triple placebo, new therapy, first standard therapy, second standard therapy, double combination of new therapy and the first standard therapy, double combination of the new therapy with the second standard therapy, double combination of the first standard and second standard therapies, triple combination of the new therapy, the first standard therapy and the second standard therapy with 8 groups and an equal allocation ratio. The design would provide four estimates of the efficacy of new therapy, namely: (1) the difference in the mean of new therapy group and the mean of the triple placebo group, (2) the difference in the mean of the double combination of new therapy and first standard group and the mean of the first standard group, (3) the difference in the mean of the double combination of the new therapy and second standard therapy and the mean of the second standard group and (4) the difference between the mean of the triple combination of new therapy, first standard therapy and second standard therapy and the mean of the double combination of the first and second standard therapies. These four estimates should be similar to each other provided the new therapy does not interact with either of the first or second standard therapies, an assumption that can be checked.

Finally, if the two ethically challenged groups, placebo and new therapy, are dropped from the fourth design, then the remaining fractional factorial design has six groups that permit three estimates of efficacy of the new therapy, namely, (2), (3), and (4). Again these three estimates should be similar to each other provided the two standard therapies do not interact with new therapy, and these three estimates should be similar to the one estimate of the efficacy of the new therapy that it is not possible to obtain from the six groups, namely (1). This six-group fractional factorial design still permits the efficacy of the new therapy to be estimated by pooling together the three estimates with the mean of (2), (3), and (4). Provided the interaction assumption is reasonable, this three-term mean should provide adequate evidence of the efficacy of the new therapy. This fifth design should be ethically acceptable since no patient is being denied the benefit of a known therapy.

Summary

The placebo's role in medicine has been and continues to be in transition. For decades, the dispute regarding the placebo revolved around its ability (or not) to induce psychological or physiological effects. More recently, the debate has focused on the utilization of the placebo-controlled trial, its usefulness, and whether, despite informed consent, it impinges on patient autonomy and the practice of beneficence. There are currently several international guidelines that direct the researchers on the use of placebos; however, there are prominent differences between them. The future is likely to bring developments in trial designs where a new therapy's effect can be compared to placebo yet no individual subject is exposed only to placebo, designs that lower the placebo response rate so that treatment effect may be ascertained more efficiently, and designs that allow further exploration and exploitation of factors influencing the placebo effect.

Acknowledgments

Dr. Holbrook is the recipient of a Canadian Institutes for Health Research Career Investigator award.

References

[1] Bailar, J.C. III. (2001). The powerful placebo and the Wizard of Oz, *The New England Journal of Medicine* **344**, 1630–1632.

[2] Barsky, A.J., Saintfort, R., Rogers, M.P. & Borus, J.F. (2002). Nonspecific medication side effects and the nocebo phenomenon, *JAMA* **287**, 622–627.

[3] Beecher, H.K. (1955). The powerful placebo, *JAMA* **159**, 1602–1606.

[4] Benedetti, F. (1996). The opposite effects of the opiate antagonist naloxone and the cholecystokinin antagonist proglumide on placebo analgesia, *Pain* **64**, 535–543.

[5] Benson, H. & Friedman, R. (1996). Harnessing the power of the placebo effect and renaming it "remembered wellness", *Annual Review of Medicine* **47**, 193–199.

[6] Blackwell, B., Bloomfield, S.S. & Buncher, C.R. (1972). Demonstration to medical students of placebo responses and non-drug factors, *Lancet* **1**, 1279–1282.

[7] Brody, H. & Waters, D.B. (1980). Diagnosis is treatment, *The Journal of Family Practice* **10**, 445–449.

[8] Charney, D.S., Nemeroff, C.B., Lewis, L., Laden, S.K., Gorman, J.M., Laska, E.M., Borenstein, M., Bowden, C.L., Caplan, A., Emslie, G.J., Evans, D.L., Geller, B., Grabowski, L.E., Herson, J., Kalin, N.H., Keck, P.E., Jr., Kirsch, I., Krishnan, K.R., Kupfer, D.J., Makuch, R.W., Miller, F.G., Pardes, H., Post, R., Reynolds, M.M., Roberts, L., Rosenbaum, J.F., Rosenstein, D.L., Rubinow, D.R., Rush, A.J., Ryan, N.D., Sachs, G.S., Schatzberg, A.F., Solomon, S. (2002). National depressive and manic-depressive association consensus statement on the use of placebo in clinical trials of mood disorders, *Archives of General Psychiatry* **59**, 262–270.

[9] Cobb, L., Thomas, G.I., Dillard, D.H., Merendino, K.A. & Bruce, R.A. (1959). An evaluation of internal-mammary artery ligation by a double blind technic, *The New England Journal of Medicine* **260**, 1118.

[10] Code of Federal Regulations. Title 21 Volume 5. Revised April 1, 2002 Accessed on November 18, 2002 at `http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?FR=314.126`. 2002.

[11] Coyne, C.A., Xu, R., Raich, P., Plomer, K., Dignan, M., Wenzel, L.B., et al. (2003). Randomized, controlled trial of an easy-to-read informed consent statement for clinical trial participation: a study of the eastern cooperative oncology group, *Journal of Clinical Oncology* **21**, 836–842.

[12] de Craen, A.J., Roos, P.J., Leonard d, V. & Kleijnen, J. (1996). Effect of colour of drugs: systematic review of perceived effect of drugs and of their effectiveness, *BMJ* **313**, 1624–1626.

[13] Desharnais, R., Jobin, J., Cote, C., Levesque, L. & Godin, G. (1993). Aerobic exercise and the placebo effect: a controlled study, *Psychosomatic Medicine* **55**, 149–154.

[14] Einarson, T.E., Hemels, M. & Stolk, P. (2001). Is the placebo powerless? *The New England Journal of Medicine* **345**, 1277–1279.

[15] Ernst, E. & Resch, K.L. (1995). Concept of true and perceived placebo effects, *BMJ* **311**, 551–553.

[16] Fava, M., Evins, A.E., Dorer, D.J. & Schoenfeld, D.A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach, *Psychotherapy and Psychosomatics* **72**, 115–127.

[17] Greene, P.J., Wayne, P.M., Kerr, C.E., Weiger, W.A., Jacobson, E., Goldman, P., Kaptchuk, T. (2001). The powerful placebo: doubting the doubters. *Advance in Mind-Body Medicine* **17**, 298–307.

[18] Guidance for institutional review boards and clinical investigators. 21 CFR Part 50. 1998 Update. Accessed on November 18, 2002 at `http://www.fda.gov/oc/ohrt/irbs/appendixb.html`. 1998.

[19] Hahn, R.A. (1997). The nocebo phenomenon: concept, evidence, and implications for public health, *Preventive Medicine* **26**, 607–611.

[20] Horng, S. & Miller, F.G. Is placebo surgery unethical? *The New England Journal of Medicine* **347**, 137–139.

[21] How do the outcomes of patients treated within randomized control trials compare with those of similar patients treated outside these trials? `http://hiru.mcmaster.ca/ebm/trout/`, accessed 2002-11-17. 2001.

[22] Hrobjartsson, A. (2002). What are the main methodological problems in the estimation of placebo effects? *Journal of Clinical Epidemiology* **55**, 430–435.

[23] Hrobjartsson, A. & Gotzsche, P.C. (2001). Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment, *The New England Journal of Medicine* **344**, 1594–1602.

[24] Irvine, J., Baker, B., Smith, J., Jandciu, S., Paquette, M., Cairns, J., et al. (1999). Poor adherence to placebo or amiodarone therapy predicts mortality: results from the CAMIAT study. Canadian Amiodarone Myocardial Infarction Arrhythmia Trial, *Psychosomatic Medicine* **61**, 566–575.

[25] Joffe, S., Cook, E.F., Cleary, P.D., Clark, J.W. & Weeks, J.C. (2001). Quality of informed consent in cancer clinical trials: a cross-sectional survey, *Lancet* **358**, 1772–1777.

[26] Kaptchuk, T.J. (1998). Powerful placebo: the dark side of the randomised controlled trial, *Lancet* **351**, 1722–1725.

[27] Kaptchuk, T.J. (2001). Is the placebo powerless? *The New England Journal of Medicine* **345**, 1277–1279.

[28] Kaptchuk, T.J., Goldman, P., Stone, D.A. & Stason, W.B. (2000). Do medical devices have enhanced placebo effects? *Journal of Clinical Epidemiology* **53**, 786–792.

[29] Khan, A., Leventhal, R.M., Khan, S.R. & Brown, W.A. (2002). Severity of depression and response to antidepressants and placebo: an analysis of the food and drug administration database, *Journal of Clinical Psychopharmacology* **22**, 40–45.

[30] Khan, A., Warner, H.A. & Brown, W.A. (2000). Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: an analysis of the Food and Drug Administration database, *Archives of General Psychiatry* **57**, 311–317.

[31] Lavori, P.W., Sugarman, J., Hays, M.T. & Feussner, J.R. (1999). Improving informed consent in clinical trials: a duty to experiment, *Controlled Clinical Trials* **20**, 187–193.

[32] Lilford, R.J. (2003). Ethics of clinical trials from a Bayesian and decision analytic perspective: whose equipoise is it anyway? *BMJ* **326**, 980–981.

[33] Macklin, R. (1999). The ethical problems with sham surgery in clinical research, *The New England Journal of Medicine* **341**, 992–996.

[34] Mataix-Cols, D., Rauch, S.L., Manzo, P.A., Jenike, M.A. & Baer, L. (1999). Use of factor-analyzed symptom dimensions to predict outcome with serotonin reuptake inhibitors and placebo in the treatment of obsessive-compulsive disorder, *American Journal of Psychiatry* **156**, 1409–1416.

[35] Moerman, D.E. (2000). Cultural variations in the placebo effect: ulcers, anxiety, and blood pressure, *Medical Anthropology Quarterly* **14**, 51–72.

[36] Moerman, D.E. & Jonas, W.B. (2002). Deconstructing the placebo effect and finding the meaning response, *Annals of Internal Medicine* **136**, 471–476.

[37] Moseley, J.B., O'Malley, K., Petersen, N.J., Menke T.J., Brody, B.A., Kuykendall, D.H., Hollingsworth, J.C., Ashton, C.M., Wray, N.P. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *The New England Journal of Medicine* **347**, 81–88.

[38] O'Connor, A.M., Rostom, A., Fiset, V., Tetroe, J., Entwistle, V., Llewellyn-Thomas, H., et al. (1999). Decision aids for patients facing health treatment or screening decisions: systematic review, *BMJ* **319**, 731–734.

[39] Papakostas, Y.G. & Daras, M.D. (2001). Placebos, placebo effect, and the response to the healing situation: the evolution of a concept, *Epilepsia* **42**, 1614–1625.

[40] Riis, P. (2000). Perspectives on the fifth revision of the declaration of Helsinki, *JAMA* **284**, 3045–3046.

[41] Rochon, P.A., Binns, M.A., Litner, J.A., Litner, G.M., Fischbach, M.S., Eisenberg, D., Kaptchuk, T.J., Stason, W.B., Chalmers, T.C. (1999). Are randomized control trial outcomes influenced by the inclusion of a placebo group?: a systematic review of nonsteroidal anti-inflammatory drug trials for arthritis treatment, *Journal of Clinical Epidemiology* **52**, 113–122.

[42] Rosenberg, N.K., Mellergard, M., Rosenberg, R., Beck, P. & Ottosson, J.O. (1991). Characteristics of panic disorder patients responding to placebo, *Acta Psychiatrica Scandinavica, Supplementum* **365**, 33–38.

[43] Rothman, K.J. & Michels, K.B. (1994). The continuing unethical use of placebo controls, *The New England Journal of Medicine* **331**, 394–398.

[44] Slater, E.E. (2002). IRB reform, *The New England Journal of Medicine* **346**, 1402–1404.

[45] Solomon, D.A. (1995). The use of placebo controls, *The New England Journal of Medicine* **332**, 62.

[46] Spiegel, D., Kraemer, H. & Carlson, R.W. (2001). Is the placebo powerless? *The New England Journal of Medicine* **345**, 1276–1279.

[47] St. Joseph's Health care/McMaster University/Hamilton Health Sciences. Informed consent checklist. `http://www.fhs.mcmaster.ca/csd/forms/ic-chklist.pdf`. 7-30-2003.

[48] Steinbrook, R. (2002). Protecting research subjects–the crisis at Johns Hopkins, *The New England Journal of Medicine* **346**, 716–720.

[49] Tattersall, M.H. (2001). Examining informed consent to cancer clinical trials, *Lancet* **358**, 1742–1743.

[50] Temple, R. & Ellenberg, S.S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues, *Annals of Internal Medicine* **133**, 455–463.

[51] Tri-Council Policy statement. Ethical conduct for research involving humans. Updated November 21, 2000. Accessed on November 20, 2002 at. `http://www.nserc.ca/programs/ethics/english/policy.htm`. 2000.

[52] Truog, R.D., Robinson, W., Randolph, A. & Morris, A. (1999). Is informed consent always necessary for randomized, controlled trials? *The New England Journal of Medicine* **340**, 804–807.

[53] U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry: E 10 Choice of Control group and Related Issues in Clinical Trials. Available at: `http://www.fda.gov/cder/guidance/4155fnl.pdf.`, accessed July 24, 2003. 2001.

[54] Verheggen, F.W., Jonkers, R. & Kok, G. Patients' perceptions on informed consent and the quality of information disclosure in clinical trials, *Patient Education and Counseling* 96 A.D.**29**, 137–153.

[55] Walsh, B.T., Seidman, S.N., Sysko, R. & Gould, M. (2002). Placebo response in studies of major depression: variable, substantial, and growing, *JAMA* **287**, 1840–1847.

[56] Weiner, M. & Weiner, G.J. (1996). The kinetics and dynamics of responses to placebo, *Clinical Pharmacology and Therapeutics* **60**, 247–254.

[57] World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects, *JAMA* (2000). **284**, 3043–3045.

ANNE HOLBROOK, CHARLES H. GOLDSMITH & MOVA LEUNG

Point Processes

Point processes provide the appropriate mathematical models for describing a countable set of points randomly located in some space; renewal and Poisson processes on a line are the simplest and best-known examples. There is a rich variety of applicable point process models of relevance in many fields. As discussed in [17], the origins of point process theory lie in the areas of **life tables** and renewal theory, counting problems starting from work of **S.D. Poisson** and leading to applications in particle physics and population processes (*see* **Population Growth Models**), and communication engineering.

Illustrative examples of point processes include: times of emission of pulses in a nerve fiber; times of arrival of patients at an intensive care unit; survival times of patients following onset of a disease; locations of trees in a forest, ants' nests in a region, particles in space, or a specified type of cell in a section (or volume) of tissue; locations and instantaneous directions of movement of insects, spermatozoa, etc. in a region; times of occurrence, locations and magnitudes of earthquakes in a given region.

It may be helpful to read the present article in conjunction with those on **renewal processes** and **Poisson processes**. Especially in the latter, many key point process ideas are introduced, often in a simple intuitive way using appealing examples. Here, we sketch in a more formal way the foundational ideas, in the hope that the reader may be assisted and motivated to read the literature. Even those parts of the point process literature that look forbiddingly mathematical are often based on very simple ideas which a nonmathematician may well be able to grasp without needing to understand all the detailed mathematics. Further aims, especially in later sections, are to show some of the rich variety of point process and **stochastic process** models that can be built from more basic point processes, and to discuss some aspects of statistical inference for point process data.

Representing Realizations

The intuitive paradigm of a point process realization should mostly be, as above, a (finite or) countable set $\mathbf{x} = \{x_i\}$ of (distinct) points randomly located in some *state space* S. This space would commonly be one-dimensional, often representing time, or two- or three-dimensional. In principle, dealing with earthquakes would require a time dimension, three spatial dimensions for the location of the epicenter, and a further dimension for the magnitudes. Thus, for many applications, the state space S will be a Euclidean space, and essentially, this is assumed in the present article. More general spaces can be used, and indeed are required to deal with aspects of point process theory and some applications, for example in stochastic geometry and **stereology** (see [5] and [9]); some of these applications are discussed in a later section.

Two other points of view concerning realizations are possible. The first, applicable only when the space is one-dimensional and ordered, like the real line, involves representing the realization as the sequence of "intervals" between successive points and telling where these points should be located in relation to the origin. Renewal processes are usually defined in this way. The second involves viewing the realization as a counting measure $N(\cdot)$. (A *measure* is a function defined for sets, taking nonnegative real values or the value $+\infty$, and additive, even countably so, over disjoint sets. In respect of the latter property a measure behaves like **probabilities**. A measure is called a *counting measure* if its values are nonnegative integers or $+\infty$ (*see* **Counting Process Methods in Survival Analysis**).) The connection between this and the initial view is that $N(B) = \#\{x_i \in B\}$, the number of points from the set $\{x_i\}$ which lie in B, considered for suitable subsets B of the space. Such subsets should include at least all possible bounded sets which are intervals (on a line), rectangles or disks (in a plane), or boxes or balls (in space), as well as unions and intersections of these. In addition, it is usual to assume that $N(B)$ is finite whenever the set B is bounded. Such an assumption, often called *local finiteness*, is likely to be satisfied in most applications.

The counting (measure) view of realizations, although maybe not the most intuitive, is not without practical relevance, in that suitable counts will often provide an appropriate description of the data. Moreover, this view is necessary for developing a nice mathematical theory of point processes in general spaces, in particular because it facilitates the treatment of possible "multiple" points. (In the set view of realizations the points x_i must, strictly speaking, be distinct.)

A disadvantage of adopting the counting viewpoint is that the resultant theory depends heavily on the theory of measure and integration, which is often regarded as difficult, and not common knowledge among those interested in applications of point processes. Nevertheless, many of the ideas of point process theory can be explained with minimal dependence on measure and integration theory, sometimes by resorting to the more heuristic "set" view of realizations. This is the approach adopted in the present article. For a more mathematical approach consult the references listed at the end of this article, especially [24].

Describing Point Processes

To describe a point process we need, in principle, all probabilities such as $\Pr(N(B) = n), n = 0, 1, \ldots$, for suitable sets B, and also all joint probabilities

$$\Pr(N(B_1) = n_1, N(B_2) = n_2), \quad n_1, n_2 = 0, 1, \ldots,$$

for suitable B_1 and B_2. For fixed B, the collection of probabilities

$$\{\Pr(N(B) = n) : n = 0, 1, \ldots\}$$

is one of the *one-dimensional distributions* of the process. Similarly, for fixed B_1 and B_2 the collection

$$\{\Pr(N(B_1) = n_1, N(B_2) = n_2) : n_1, n_2 = 0, 1, \ldots\}$$

of joint probabilities is one of the *two-dimensional distributions* of the point process.

In general, for a fixed positive integer k and (Borel) subsets $B_1, \ldots, B_k$ of the state space we need

$$\{\Pr(N(B_1) = n_1, \ldots, N(B_k) = n_k) : n_1, \ldots, n_k = 0, 1, \ldots\}.$$

This is simply the joint distribution of $N(B_1), \ldots, N(B_k)$, the numbers of points in the sets $B_1, \ldots, B_k$. These joint distributions, for all possible k and subsets $B_1, \ldots, B_k$, are called the *finite-dimensional distributions* of the point process. For a given point process, they must be mutually consistent; for instance, the marginals of any two-dimensional distribution must coincide with the separate one-dimensional distributions for the specified sets B_1 and B_2. Furthermore, for example, since for disjoint B_1 and B_2 the counts must be additive in the sense that $N(B_1 \cup B_2) = N(B_1) + N(B_2)$ (with probability one), the family of one-dimensional distributions of a point process must be an "additive" family of probability distributions. Such additivity and consistency conditions, though in a sense simple, are highly restrictive: they put severe limitations on the structure of the possible finite-dimensional distributions for a point process, and in general make it a nontrivial task to specify (a new) point process model.

If a suitable consistent, additive family of potential finite-dimensional distributions can be found, then we have specified a point process. (Some would regard this conclusion as resting on a variant of a stochastic process existence theorem usually attributed to Kolmogorov.) Thus the properties of a point process can be regarded as being determined by the collection of all its finite-dimensional distributions.

At least when the state space is Euclidean, i.e. $S = \mathbb{R}^d$ for some positive integer d, a point process is called *(strictly) stationary* if its probabilistic properties are invariant under translation: specifically, if for every u in the state space S, every positive integer k, every collection of subsets $B_1, \ldots, B_k$ of the state space, and every collection $n_1, \ldots, n_k$ of nonnegative integers, we have

$$\Pr(N(B_1) = n_1, \ldots, N(B_k) = n_k) = \Pr(N(B_1 + u) = n_1, \ldots, N(B_k + u) = n_k),$$

where $B + u = \{x + u : x \in B\}$ denotes the set of all points obtained by translating B by u. Thus for a stationary point process the distribution, and hence expected value, of the number of points falling in any translate of a given set B is the same as that for the set B. One consequence is that $\mathrm{E}N(B) = \mu|B|$, where $|B|$ denotes the Lebesgue measure (length, area, volume, etc.) of B and μ, called the *intensity*, denotes the expected number of points falling in any set U having $|U| = 1$ (that is, any set having unit length, area or volume, as appropriate). Also, for any stationary point process, it can be shown that the limit

$$\rho = \lim_{h \to 0+} \frac{\Pr\big(N(B_h) > 0\big)}{h^d},$$

where $B_h = (0, h]^d$, exists and $0 \le \rho \le \infty$. The quantity ρ is often called the *rate* of the point process, though this terminology is not standardized in the literature. Note that $\rho \le \mu$ since $\Pr\big(N(B_h) >$

$0) \leq \mathbb{E}N(B_h)$ for all h. For spatial point processes ($d \geq 2$) one may also be interested in *isotropy*, a concept analogous to stationarity, defined by all the finite-dimensional distributions being invariant under rotations.

A point process is called *simple* if, with probability one, its realizations have no multiple points, i.e. if $\Pr(N(\{x\}) = 0 \text{ or } 1 \text{ for every } x \in S) = 1$. From any point process N it is possible to derive an associated simple point process N^* by "forgetting" all multiplicities in the original process N. The process N^* is stationary whenever N is stationary, and therefore has an intensity μ^*. Clearly, $\mu^* \leq \mu$. Furthermore, for a point process N that is both simple and stationary $\rho = \mu\ (= \mu^*)$, a result often known as Korolyuk's theorem.

A point process is termed *mixing* if events defined in terms of the numbers of points in widely separated sets are close to independent. (For a formal definition see, for example, [17].) Some such property is needed to ensure consistent estimation of the intensity of a stationary point process; many use the related notion of ergodicity (cf. [25] or [43, p. 194]).

It is important to be able to look at the distribution of a point process when we view that process from an arbitrary point in the process. The *Palm distribution* of a point process is, more formally, the conditional distribution of the process given that, for example, there is a point of the process at x. For a stationary point process this conditioning can be reduced to the demand that there be a point at the origin. Such conditioning requires care because the conditioning event is clearly one of probability zero, and a rigorous introduction is beyond the scope of the present article. For a stationary point process on the real line it is by means of the Palm distribution that we must approach, for example, the distribution of the times between successive points ("intervals") when starting from a description of the process in terms of its finite-dimensional distributions. There is also an inversion formula which allows the latter distributions to be expressed in terms of the Palm distribution. The simplest consequence of this is that the intensity of a stationary simple point process is equal to the reciprocal of the mean time between successive points. For a development of these ideas see [41, Chapter 4] or [45]. For a stationary isotropic point process the Palm distribution is needed at least in order to define formally the nearest-neighbor distribution function and the K-function which play an important role in statistical inference for such processes (see the section "Statistical Inference" later in this article).

Some Basic Point Processes

A *homogeneous Poisson process* in a Euclidean space $S = \mathbb{R}^d$ can be defined by the following two requirements:

1. $N(B) \sim \text{Poi}(\lambda|B|)$, where $|\cdot|$ denotes Lebesgue measure (length, area or volume) on $\mathbb{R}^d$ and B denotes any bounded subset of the state space S.
2. $N(B_1), \ldots, N(B_k)$ are mutually independent random variables whenever $B_1, \ldots, B_k$ are pairwise disjoint bounded subsets of S.

A point process satisfying the latter condition is often called *completely random*. Observe that requirements 1 and 2 together specify the form of all finite-dimensional distributions and ensure the process is (strictly) stationary. In addition, requirement 1 ensures that λ is the intensity of the process.

Any homogeneous Poisson process has the fundamental property that, given the number of points in a bounded subset of the state space, these points are distributed independently and uniformly over the subset. This "conditional" property is an important tool in proving other results about homogeneous Poisson processes and about processes derived from them.

The point process that results from such conditioning of is a particular type of *Bernoulli process* (also called sample process or binomial process).

In general, such a process is defined on a compact (i.e. closed and bounded) set W by requiring that all its realizations have the same fixed total number of points and that these points are distributed independently and identically over W according to some specified probability distribution. Such processes are straightforward to simulate.

A Poisson process with *intensity function* $\lambda(u)$, where $\lambda(u)$ is nonnegative for all u in S and $\int_B \lambda(u)\,\mathrm{d}u$ is finite for all bounded sets B, is defined as for a homogeneous Poisson process but with requirement 1 replaced by

1. $N(B) \sim \text{Poi}(\int_B \lambda_u\,\mathrm{d}u)$, for any bounded subset B of the state space S.

Such a process is also called an *inhomogeneous* or *nonhomogeneous Poisson process*, a terminology which is not meant to exclude homogeneous Poisson processes as the special cases for which the intensity function is in fact constant. The class of inhomogeneous Poisson processes includes, for example, processes for which $\lambda(u)$ – or perhaps preferably, $\ln \lambda(u)$ – exhibits some trend with u, is a periodic function, or is dependent on the values of some associated explanatory variables. For processes on the real line, viewed as a time axis, the intensity function is commonly referred to as the instantaneous intensity function.

In practice, many point processes will be defined most simply by means other than direct specification of their finite-dimensional distributions. When $S = [0, \infty)$ we can define an *(ordinary) renewal process* by the set $\{L_1, L_1 + L_2, L_1 + L_2 + L_3, \ldots\}$ of random points, where $L_1, L_2, L_3, \ldots$ are independent and identically distributed (lifetime) random variables. If L_1 is allowed a distribution different from that of the other random variables the process is called a *modified renewal process*. Counting properties of a renewal process are less simple to describe, except when the common distribution of $L_1, L_2, L_3, \ldots$ is an **exponential distribution**, in which case the process reduces to a homogeneous Poisson process on $[0, \infty)$.

A wide variety of types of point process can be defined in terms of simpler point processes, such as Poisson or renewal processes. Their finite-dimensional distributions could, in principle, then be derived, although this may be difficult or tedious in practice, and unnecessary in full detail. Sometimes it is possible to work with probability **generating functions** or **moment generating functions** of relevant finite-dimensional distributions, and at other times with certain summary measures or **moments**.

Three broad classes of point processes can be constructed from Poisson processes by introducing further randomness. One of these is the class of *compound Poisson processes*. Such a process is obtained from a Poisson process, homogeneous or inhomogeneous, by replacing each point, independently of the other points, by a random number of new points all of which are placed at the associated point of the original process. In general a process of this type would have points with multiplicity greater than one. Another class is that of *mixed Poisson processes* (cf. [22]), which are defined by allowing the parameter λ of a homogeneous Poisson process to have a specified distribution (*see* **Contagious Distributions**). Such processes provide one of the simplest classes of processes which are not mixing; essentially the lack of mixing is a consequence of the dependence of the number of points in even widely separated sets on the common value of λ. The third class is that of *doubly stochastic Poisson processes* or *Cox processes*. Such a process is obtained from an inhomogeneous Poisson process by allowing its intensity function to be a realization of some other stochastic process, which might be thought of as representing an underlying (usually unobservable) environmental heterogeneity. One type of Cox process on $S = \mathbb{R}$ is obtained taking the stochastic process governing the intensity function to be a continuous-time **Markov chain** with finitely many states. Such processes are also called *Markov modulated Poisson processes* [40]. In the simplest case, the Markov chain has just two states, where these correspond to a high level and a low (or even zero) level for the intensity function. Both the mixed Poisson and Cox process models allow for **overdispersion** (relative to a Poisson process), in that the counts may have a variance greater than their mean.

Product Densities

For point process properties, an approach which many find both intuitively appealing and useful is to consider the so-called *product densities* of the process. These are defined, under further conditions which will be mentioned in the next section, for a point process which is simple. They can be described in terms of differentials as follows: the first-order product density is given, for any u in the state space S, by $m_1(u)\,\mathrm{d}u = \Pr(N(\mathrm{d}u) = 1)$, where the right-hand side can be thought of as the probability of the event "a point at u". In the case of a Poisson process with intensity function $\lambda(u)$ we have $m_1(u) = \lambda(u)$. For a general point process $m_1(u)$ itself is called the *intensity function* of the process.

The second-order product density is given, for any distinct u_1, u_2 in the state space S, by

$$m_2(u_1, u_2)\,\mathrm{d}u_1\,\mathrm{d}u_2 = \Pr(N(\mathrm{d}u_1) = 1, N(\mathrm{d}u_2) = 1),$$

where the right-hand side is the probability of the event "a point at each of u_1 and u_2". In general, for any positive integer k and any distinct $u_1, \ldots, u_k$

in the state space, the kth-order product density is given by

$$m_k(u_1, \ldots, u_k)\, \mathrm{d}u_1 \ldots \mathrm{d}u_k = \Pr(N(\mathrm{d}u_1) = 1, \ldots, N(\mathrm{d}u_k) = 1).$$

For a stationary point process we must have $m_1(u) = \mu$, the intensity of the point process, and $m_2(u_1, u_2)$ must be a function just of $u_2 - u_1$, or even just of the (Euclidean) distance $\mathrm{d}(u_1, u_2)$ between u_1 and u_2 when the process is also isotropic.

In the case of a homogeneous Poisson process with intensity λ we have $m_k(u_1, \ldots, u_k) = \lambda^k$ for all k and distinct $u_1, \ldots, u_k$, and the covariance density is identically zero. For a renewal process, on $S = [0, \infty)$, it can be shown that

$$m_k(u_1, \ldots, u_k) = h_1(u_1)h(u_2 - u_1) \ldots h(u_k - u_{k-1})$$

for all k and distinct $u_1, \ldots, u_k$, where h is the (ordinary) renewal density and h_1 the modified renewal density, which takes account of the fact that the origin need not be a renewal point (*see* **Renewal Processes**).

Since $m_1(u)\, \mathrm{d}u = \Pr(N(\mathrm{d}u) = 1) = \mathrm{E}N(\mathrm{d}u)$, it follows that $\mathrm{E}N(B) = \int_B m_1(u)\, \mathrm{d}u$ gives the expected number of points in B. Similarly,

$$\mathrm{E}N(B)[N(B) - 1] = \int_B \int_B m_2(u_1, u_2)\, \mathrm{d}u_1\, \mathrm{d}u_2$$

yields the second factorial moment of $N(B)$. These ideas lead to consideration of moment measures.

Moment Measures

For any point process N, a set function $M_1(\cdot)$ can be defined for (Borel) subsets B of S by $M_1(B) = \mathrm{E}N(B)$. This inherits the nonnegativity and additivity properties of N and so is itself a measure, variously called the *mean measure*, the *intensity measure* or the *first moment measure* of N. This and higher-order equivalents provide for a point process the analogs of the ordinary moments of a random variable.

The simplest aspects of the dependence structure of a point process are embodied in its *second moment measure* $M_2(\cdot)$. This is defined for subsets B_1 and B_2 of the state space starting from its value $M_2(B_1 \times B_2) = \mathrm{E}N(B_1)N(B_2)$ for the "rectangle" $B_1 \times B_2$ in S^2, the Cartesian product of the state space with itself, and extending in a standard manner to a measure $M_2(\cdot)$ on S^2. Since $M_2(B_1 \times B_2) = M_2(B_2 \times B_1)$, $M_2(\cdot)$ is a symmetric measure. In a similar way, starting from $C(B_1 \times B_2) = M_2(B_1 \times B_2) - M_1(B_1)M_1(B_2)$ we can define the *covariance measure* of the point process N. (Observe that C, although being additive for disjoint sets, may take negative values.) Putting $B_1 = B_2 = B$ gives $M_2(B \times B) = \mathrm{E}\{N(B)^2\}$, and also the *variance function*, a set function, defined by $\mathrm{var}N(B) = C(B \times B)$. When the state space is the real line and $B = (0, t]$, it is the function $V(t) = \mathrm{var}N((0, t])$ which is usually called the variance function.

A disadvantage of the moment measures is that the second and all higher-order moment measures of any point process have "diagonal concentrations" [17, Section 5.4]. At least for simple point processes, this disadvantage can avoided by introducing *factorial moment measures*, which are the point process analogs of factorial moments for a random variable. The first factorial moment measure coincides with the mean measure, while the second factorial moment measure $M_{[2]}(\cdot)$ can be defined by

$$M_{[2]}(B_1 \times B_2) = M_2(B_1 \times B_2) - M_1(B_1 \cap B_2)$$

for subsets B_1 and B_2 of S. Observe that when B_1 and B_2 are disjoint the right-hand side reduces to $M_2(B_1 \times B_2)$, but that when $B_1 = B_2 = B$ (say) it reduces to the second factorial moment of $N(B)$. In some circumstances when $S = \mathbb{R}^d$ for some d, the factorial moment measures can be defined by densities with respect to Lebesgue measure (length, area, volume, etc.) in the appropriate dimensional Euclidean space, these being the product densities introduced heuristically in the previous section.

A useful result for moment measures is *Campbell's theorem* (see [17, Section 6.4] or [27, Section 3.2]), the simplest version of this being that for a wide class of functions g and subsets B of the state space

$$\mathrm{E}\left\{\int_B g(u)N(\mathrm{d}x)\right\} = \int_B g(u)M_1(\mathrm{d}u).$$

An alternative, possibly more immediately meaningful expression of this is

$$\mathrm{E}\left\{\sum_{i:X_i \in B} g(X_i)\right\} = \int_B g(u)m_1(u)\, \mathrm{d}u,$$

where $\{X_i\}$ is the set of random points corresponding to the point process N and $m_1(u)$ its intensity

function. The latter integral simplifies to $\mu \int_B g(u)\,\mathrm{d}u$ whenever the point process is stationary with intensity μ. Campbell's theorem can be extended to results for higher-order moment measures, and also to results involving the Palm distribution; see, for example, [45, Chapter 4].

Operations on Point Processes and Associated Limit Results

There are a number of operations by which new point processes can be generated from a process or processes already defined. Useful tools for dealing with these operations – generating functionals – will be discussed in a later section.

Superposition

For two point processes N_1 and N_2, not necessarily independent, a new process, N, their *superposition*, can be defined by $N = N_1 + N_2$, where this is to be interpreted as $N(B) = N_1(B) + N_2(B)$ for subsets B of the state space (or in intuitive terms as the pooling of the sets of points for the two processes). When the processes are independent, the distribution of the superposition can be viewed as a convolution of the distributions of the summand processes. In particular, the superposition of two independent homogeneous Poisson processes with respective intensities λ_1 and λ_2 is another homogeneous Poisson process, with intensity $\lambda = \lambda_1 + \lambda_2$. By iteration, the superposition of several point processes can be considered and, under appropriate conditions, the superposition of a countable number of point processes (the latter being needed, for example, in "Cluster processes"; see below).

Random Deletion

Given a point process N, consider the operation of *random deletion* or *random thinning* (sometimes called Bernoulli deletion) whereby, given a realization of N, each point in that realization is deleted with probability $1 - p$ and retained with probability p, independently of all other points in the realization. The intensity function for the process of retained points is then clearly $pm_1(u)$, where $m_1(u)$ is the corresponding intensity for the original process N. Thus, if N is stationary with intensity μ, then the process of retained points is stationary, with intensity $p\mu$. If N is a homogeneous Poisson process, then so is the new process; furthermore, so is the process of deleted points, and these two processes are independent.

Random Translation

For any point process N the operation of *random translation* can be defined as follows: given a realization of N, each point in that realization is shifted, independently of all other points in the realization, the shift having some specified distribution function on S, where this distribution is the same for each point in the original realization. (The shifts are thus assumed independent and identically distributed.) If the original point process is stationary, then so also is the resultant randomly translated process; furthermore, if N is homogeneous Poisson, then so is the resultant process and their intensities are the same.

Cluster Processes

Here each point of an "input" point process is replaced by the points of some subsidiary point process or cluster, and the superposition of all these clusters is then the "output" process or *cluster process*. In the simplest situation the input might be a homogeneous Poisson process with specified intensity and the clusters independent and identically distributed point processes, each with its origin translated to the associated point of the input process (often called the cluster center). Then, when $S = \mathbb{R}$, two particular types of cluster structure are: (i) a finite renewal process with the number of points either fixed or following a specified distribution, and the interval (lifetime) distribution also specified; or (ii) a process in which the number of points is either fixed or follows a specified distribution, and these points are placed independently and identically according to a specified distribution on S. The resultant cluster processes are respectively termed *Bartlett–Lewis* and *Neyman–Scott (cluster) processes*. Processes of the latter type can also be considered for Euclidean state spaces $S = \mathbb{R}^d$. Compound Poisson processes are special cases in which all the points of a given cluster are placed at the cluster center.

State Space Transformation

In this case the operation is not defined directly on a process or processes, but initially as a mapping

which transforms points of the state space S into points of a new state space S^*. Such a mapping then transforms the set of points of any given point process realization in S into a corresponding realization in S^*. Under this type of operation an inhomogeneous Poisson process on S is transformed into another such process on S^* [27, Section 2.3]. However, if the initial process were a homogeneous Poisson process the homogeneity would not in general be preserved.

Limit Results

Associated with the operations of superposition, random deletion and random translation are certain limit results concerning point processes. The simplest is that the superposition of a large number of independent and suitably sparse point processes is approximately a Poisson process.

For point processes in Euclidean state spaces, the simplest limit theorem for deletions is as follows: suppose that points of an initial point process are subject to Bernoulli deletions, with retention probability p for any individual point, and that the scale is contracted so as to balance the deletions and preserve the intensity. From suitable initial point processes, it is possible to prove convergence as $p \to 0$ to a homogeneous Poisson process.

Among limit results for random translations, arguably the simplest allows the points of a suitable initial point process to move with independent and identically distributed random velocities, and yields convergence to a homogeneous Poisson process after a long time.

Substantial generalizations of these basic results can be considered, and once again generating functional methods can be used. Here, we have not attempted even a definition of what is meant by convergence of point processes: for this and other details see, for example, [17, Chapter 9].

Generating Functionals

Various **generating functions** (probability generating functions, **moment generating functions** or Laplace transforms, and **characteristic functions**) are useful tools in the study of random variables; probability generating functions are especially helpful with nonnegative integer-valued random variables. All the above generating functions have "functional" relatives that are useful in the study of point processes, as a means of compactly summarizing information about point processes and enabling that information to be manipulated. We mention here only one type of functional. (A functional is a function whose argument is itself a function, rather than a real number or vector of real numbers.)

The *probability generating functional* G of a point process N can be defined, for functions h mapping the state space S to values in [0, 1] and equal to one outside some bounded set, by $G[h] = \mathrm{E}\left(\exp\left\{\int_s \ln h(u) N(\mathrm{d}u)\right\}\right)$. This can be considered heuristically as

$$G[h] = \mathrm{E}\left(\prod_{u \in S} h(u)^{N(\mathrm{d}u)}\right) = \mathrm{E}\left(\prod_{u : N(\{u\}) > 0} h(u)^{N(\{u\})}\right).$$

The middle expression allows us to think of G as the joint probability generating function of all the random variables $N(\mathrm{d}u), u \in S$. Working heuristically with this form, property (ii) of a homogeneous Poisson process with intensity λ, and the form of the probability generating function of a Poisson-distributed random variable, we find that the probability generating functional of a homogeneous Poisson process with intensity λ is

$$\begin{aligned} G[h] &= \prod_{u \in S} \mathrm{E}(\exp\{[h(u) - 1]\lambda \, \mathrm{d}u\}) \\ &= \exp\left\{\lambda \int_s [h(u) - 1] \, \mathrm{d}u\right\}. \end{aligned}$$

Consider now the superposition N of two independent point processes N_1 and N_2 having respective probability generating functionals G_1 and G_2. Then the probability generating functional of the superposition N is given by $G[h] = G_1[h]G_2[h]$. The probability generating functional of the process obtained from N by random deletions, as described earlier is $G[1 - p + ph]$, where G is the probability generating functional of N. Using these ideas it is possible to give proofs of the assertions about Poisson processes made in the previous section; see, for example, [17, Section 8.2].

Markov Point Processes and Some Related Processes

A wide class of point processes on some bounded subset S of a Euclidean space can be defined by

specifying their densities, i.e. likelihood ratios, with respect to a homogeneous Poisson process of unit intensity defined on S. Among these processes are some, known as *Markov point processes*, which allow the introduction of a form of spatial dependence that is local or Markov (see [14, Section 8.5.5], [20, Section 4.9] or [45, Section 5.5]). For the simplest such processes, the densities (with respect to the unit Poisson process) have a representation as a canonical **exponential family** (see, for example, [31, Chapter 2] and references cited there), with a small number of parameters. Since **maximum likelihood** estimation is in principle straightforward for canonical exponential families, inference for such Markov point processes should also be so. The main difficulty is that the **likelihood function** for such a Markov point process family would usually involve the (parameter-dependent) normalizing constant, for which an explicit closed form is generally impossible. In such a case, the conventional approach to maximum likelihood estimation, based on solving the likelihood equation(s), is not feasible. However, recent advances in computing power and statistical technology have made it possible to bypass calculation of the normalizing constant and conventional maximum likelihood estimation by using the approach known as **Markov chain Monte Carlo**, which involves large-scale **simulation**. For an introduction to these ideas, see [21].

The simplest nontrivial example of such a Markov point process family is the family of *Strauss processes*. These can be defined by densities (with respect to the unit Poisson process) which are proportional to $\beta^{n(\mathbf{x})}\gamma^{t(\mathbf{x})}$, where $\mathbf{x}$ here denotes a particular realization (i.e. a set of finitely many points from S), $n(\mathbf{x})$ is the number of points in $\mathbf{x}$ and $t(\mathbf{x})$ counts the number of r-neighbors of $\mathbf{x}$, that is the number of pairs of points in $\mathbf{x}$ that are within a prespecified distance r of each other. The parameter β, which must be nonnegative, relates to the intensity of the process, while γ is an interaction parameter and must satisfy $0 \leq \gamma \leq 1$. When $\gamma = 1$ the process reduces to a homogeneous Poisson process with intensity β. The case $\gamma = 0$ corresponds to a simple inhibition process, often called a *hard core process*, in which Poisson realizations are conditioned to have no pairs that are r-neighbors. Cases with $0 < \gamma < 1$ yield processes exhibiting less strict inhibition. Another such family, consisting of so-called *triplets processes* [21], can be generated in a similar way by bringing in a further statistic $w(\mathbf{x})$ which counts the number of triples of points that are mutual r-neighbors. The resultant exponential family has a three-dimensional canonical statistic $(n(\mathbf{x}), t(\mathbf{x}), w(\mathbf{x}))$. Such processes can, like the Strauss process, be fitted by Markov chain Monte Carlo methods [21]. Similar comments apply to the *area interaction* processes introduced in [6].

Another class of point processes including Strauss processes is the class of *pairwise interaction processes*, defined by densities (with respect to the unit Poisson process) which are proportional to $\beta^{n(\mathbf{x})} \prod_{i \neq j} h(d(x_i, x_j))$ where $d(x_i, x_j)$ denotes the usual Euclidean distance between x_i and x_j, and h an interaction function, which must be a suitable bounded function. Pairwise interaction processes are a special case of the much wider class of *Gibbs processes* [14, 45], which have their origins in statistical mechanics.

Important in the simulation of the processes discussed in this section, and in Markov chain Monte Carlo methods of inference for such processes, is a class of spatio-temporal stochastic processes known as *spatial birth and death processes* [7, 45]. These are continuous-time pure jump **Markov processes** whose state space is the set of all possible realizations of point processes on S (that is, all finite subsets of S which is assumed, as above, to be a bounded subset of a Euclidean space), and whose only possible transitions are either the "birth" of a new point, or the "death" of a point in the preceding point process realization. (Note that a spatial birth and death process is Markov as regards time). The essence of the connection with simulation is that, under certain conditions, the limiting distribution of a spatial birth and death process is a Markov point process as introduced above. A consequence of this is that realizations of such a Markov point process can be generated as observations on the relevant spatial birth and death process after it has been running for a long time [21, 36].

Statistical Inference

The remarks that follow do scant justice to a difficult area which has been the subject of much recent growth, but may serve as an introduction.

Any statistical analysis of point process data should be backed by suitable **graphical displays**.

Where feasible these should include a plot of the point process realization; in itself, this may show some form of interaction between points. For example, there may be a tendency to **clustering** (in a biological application, perhaps as a result of local reproduction), or to *inhibition* (possibly arising from competition for space or nutrients); *regularity* is likely to be observed when there is inhibition with a minimum permissible distance between points and a sufficiently high intensity. Other plots of data summaries are possible. Some of these may play a purely descriptive or summary role; others may be relevant in fitting particular point process models, or assessing the **goodness of fit** of such models. Whatever the approach, it should be driven primarily by the needs of the person who collected the data.

Assume that the data are a partial realization of a stationary point process: for example, only those points within a bounded window W may be observed. The prime interest may lie in estimation of the intensity. In a forestry application, this may give valuable information, for example on the overall quantity of wood in the forest. However, in such an application it may be necessary to use a more complex model: one that introduces supplementary information on the sizes of individual trees, represented as a mark (see the next section) attached to each point in the point process, may lead to better information on the overall quantity of wood.

If stationarity does not seem a reasonable assumption, then it may be of interest to estimate the intensity function of the process. Here nonparametric *kernel density estimation* techniques could be used (*see* **Density Estimation**); or, based on an inhomogeneous Poisson process model, a specific parametric form could be fitted for the intensity function. These approaches are discussed, for example, in [14, Sections 8.2.4, 8.5.1] and [44, Section 13.3]. Graphical displays in the form of a plot or contour plot of the estimated intensity function could be provided, respectively, for real line or planar data.

In the case of data which are a partial realization of a stationary isotropic point process, three other functions are often considered (see [14, Sections 8.2.6, 8.4] or [20, Chapter 2]). One is $F(r) = \Pr(d(u, \mathbf{x}) \leq r), r > 0$, the distribution function of the distance $d(u, \mathbf{x})$ from an arbitrary point u in S to the nearest point of the process; this is often called the *empty space function*. Another is $G(r) = \Pr(d(x, \mathbf{x}\backslash\{x\}) \leq r), r > 0$, the distribution function of the distance $d(x, \mathbf{x}\backslash\{x\})$ from an arbitrary point x of $\mathbf{x}$ to the nearest other point of the process $\mathbf{x}$. This is the *nearest-neighbor distribution function* of the process. Finally, there is the so-called (Ripley) *K-function* [38] or *reduced second moment function* which can be defined for $r > 0$ by

$$K(r) = \mu^{-1}\mathrm{E}(\text{number of further points of } \mathbf{x} \text{ within distance } r \text{ of an arbitrary point of } \mathbf{x}),$$

where μ is the intensity of the process. To define the latter two functions formally requires consideration of the Palm distribution of the process; the expectation defining the K-function is in fact an expectation with respect to the Palm distribution of the process.

For a homogeneous planar Poisson process with intensity λ it follows that $F(r) = G(r) = 1 - \exp\{-\lambda\pi r^2\}$ and $K(r) = \pi r^2$, $r > 0$. For a clustered point process, $F(r)$ for small r will be less than the corresponding value for a homogeneous Poisson process, while $G(r)$ and $K(r)$ for values of r close to the range of clustering will each be greater than the corresponding Poisson value. For a point process showing inhibition, $F(r)$ for values of r larger than the range of inhibition will exceed the homogeneous Poisson equivalent, while $G(r)$ and $K(r)$ for values of r close to the range of inhibition will each be less than the corresponding Poisson value.

Since $F = G$ for a homogeneous Poisson process, various proposals for assessing Poissonness of a given point process, or as some would say *complete spatial randomness* (often abbreviated to CSR, see [20] or [14]) of that process, are based on comparing F with G. For example, Diggle [19] considered the statistic $\sup_r |F(r) - G(r)|$. Van Lieshout and Baddeley [30] suggested the function $J(r) = [1 - G(r)]/[1 - F(r)]$, defined for r such that $F(r) < 1$, as a useful summary measure to indicate the strength and range of interpoint interactions in a point process. A homogeneous planar Poisson process with intensity λ satisfies $J(r) \equiv 1$. Furthermore, $J(r) < 1$ indicates clustering and $J(r) > 1$ inhibition or regularity, while for many point processes $J(r)$ is constant for r beyond the range of spatial interaction.

The immediately preceding remarks refer to the "true" functions being considered, whereas in practice these functions would usually be estimated from some point process realization. Estimation of the functions F, G, and K on the basis of the points in a bounded window raises special problems of *edge*

effects, which have been discussed by Baddeley [4]. Such effects are of two main types: sampling bias that is size-dependent and related to the well-known problem of **length-biased** sampling (for example, widely separated nearest-neighbor pairs are less likely to be represented in a fixed bounded sampling window), and **censoring** effects (which arise, for example, because the nearest point to a given point inside the window may be outside the window and therefore unobserved). Ways of dealing with these effects are discussed by Ripley [39] and Baddeley [4]. For example, extensions of Campbell's theorem play a key role in assessing bias in the estimation of F, G, and K.

Estimates $\hat{F}$, $\hat{G}$, and $\hat{K}$ can be plotted separately, along with their respective Poisson equivalents based on the estimated intensity. These plots can be used to assess the fit of a homogeneous Poisson process model. Such assessment may be assisted by use of **Monte Carlo** tests (cf. [20]). Here, for example based on F, one would simulate 99 independent realizations from the homogeneous Poisson model with the estimated intensity, and then construct the upper and lower envelopes for F,

$$U_F(r) = \max_i \hat{F}_i(r), \qquad L_F(r) = \min_i \hat{F}_i(r),$$

where the maximum and minimum are taken over the estimates $\hat{F}_i$ of F from each of the 99 simulations. The functions U_F and L_F are then plotted with $\hat{F}$ and its Poisson equivalent. To the extent that $\hat{F}$ lies between U_F and L_F the Poisson model is regarded as acceptable. (Note that, while $(L_F(r), U_F(r))$ gives a 98% **confidence interval** for $F(r)$ for any specified value of r, it cannot be asserted that the same confidence coefficient applies for all values of r in some interval – this is a problem of **simultaneous confidence intervals**.) A similar approach based on G, K, or J could be used; each function embodies somewhat different information from the others, so the plots should be complementary.

Variations on the above plots are possible. For example, one could use a probability plot of P–P type (*see* **Graphical Displays**) where $\hat{F}(r)$ is plotted against the corresponding Poisson equivalent $\hat{F}_{\mathrm{Poi}}(r)$ for each r, and also plot the pairs $(L_F(r), \hat{F}_{\mathrm{Poi}}(r))$ and $(U_F(r), \hat{F}_{\mathrm{Poi}}(r))$ to give corresponding envelope functions. In the case of the K-function it has been found useful to plot either $\hat{K}(r) - \pi r^2$ against r, or $[\hat{K}(r)/\pi]^{1/2} - r$ against r; (see, for example, [14] or [20]).

Indications of deviations from Poissonness shown in plots like those discussed above provide clues as to what type of non-Poisson model may be appropriate. A more detailed description of any observed clustering or inhibition could be attempted by the formulation and fitting of a more complex model (cf. [14, Section 8.5]): for example, this might be some simple Poisson cluster process or a Strauss process. The choice of a suitable model and its fitting may not be entirely simple matters and it is likely that, at least at this stage, the investigator would need to consult with a statistician knowledgeable about point processes. The parameters of a reasonably fitting model provide a summary of the original data: for a homogeneous Poisson process this summary would involve just the intensity; for a Strauss process, as described in the previous section, the parameter β is related to the intensity of the process, while γ describes interactions between neighboring points. Since the case $\gamma = 1$ reduces to a Poisson process, it is in principle possible to assess Poissonness parametrically within the family of Strauss processes by testing the hypothesis $\gamma = 1$.

One of the problems that has impeded development of inference for point process models is the difficulty, and in most cases impossibility, of writing down an expression for the likelihood function. A notable exception is that the likelihood can be written down explicitly for any (inhomogeneous) Poisson process on a line which has been observed over a fixed time interval $(0, T)$; (see [28] or [42]). As indicated in the previous section, there is much current interest (see, for example, [21] and [36]) in the use of Markov chain Monte Carlo methods. These enable likelihood-based inference to be implemented for parametric point process models even when a likelihood function cannot be written down explicitly. There is also the possibility of using pseudo-likelihood methods, in which the likelihood function is replaced by another closely related function that is then used as if it were the likelihood; see [5, 8], and the references therein.

Marked Point Processes and Other Related Processes

As we have indicated above, when modeling the location of earthquakes in a region, or trees in a forest, it may be appropriate to introduce a further random quantity, often called a *mark*, associated

with each point in an underlying point process. The resultant process is known as a *marked point process*, and can be considered as a point process on the product space $S \times M$ consisting of all pairs (x, y) where x is from the original state space S and y from a space M consisting of all possible marks. For the two examples indicated above, the mark space would be taken as $M = (0, \infty)$, though other choices for M are possible. For example, a compound Poisson process can be viewed as a marked (Poisson) point process with M the set of nonnegative integers. Although the general theory of point processes does not strictly need extension to cover marked point processes, the special structure of the new state space leads to related special structure in, for example, moment measures and product densities; see, for example, [45, Chapter 4]. There is, in particular, a version of Campbell's theorem for marked point processes (see also [44, Section 14.2]).

The case of a point process where there are two types of point, for example two types of tree in the forest, can be covered by using a two-point mark space, e.g. $M = \{1, 2\}$. In addition, this setting allows joint consideration of the point process of "retained" points and the point process of "deleted" points in the random deletion context discussed earlier. For Bernoulli deletions it is even easy to write down an expression for the joint probability generating functional of the two processes (probability generating functional of the marked point process) in terms of the probability generating functional of the original process. If the mark space is a finite set, without loss of generality $M = \{1, 2 \ldots, s\}$, then the marked point process is often called a *multitype* or *multivariate point process* (see [12, Chapter 5] or [20, Chapter 6]).

The theory of marked point processes is often helpful in providing language and a unifying framework within which other processes can be considered. **Markov and semi-Markov processes** with finitely many states can be viewed as multitype point processes [12, Section 3.2], although most properties of such processes can be conveniently derived without using this connection. Alternating renewal processes are a special case of semi-Markov processes in which there are two types of mark and two types of lifetime, alternating over time. Shot noise processes (cf. [12, Section 5.6] or [42, Chapter 4]) can also be viewed as marked point processes, where in this case the mark attached to each point is a possibly random multiple (independent and identically distributed for each point) of a fixed function, e.g. a negative exponential function with fixed decay parameter, which represents a 'blip' of electric current associated with that point. The actual shot noise process, which is not a point process, is the superposition (sum) of all the "blips" of current.

Marked point processes provide a framework for treating many probability models of interest in stochastic geometry and stereology. For example, consider a stochastic process whose realizations are a (finite or) countable number of line segments in the plane, with each segment specified by a random location, orientation and length. Such a process can be viewed as a marked point process in which the points of the underlying point process give the locations of the midpoints of the line segments, while two marks attached to each such point record respectively the orientation and length of the line segment to be associated with that point. (An example involving positions and orientations of flies on a leaf is quoted in [44, p. 265].) The simplest such model involving line segments assumes that the underlying point process is a homogeneous Poisson process in the plane. Another type of model can be built from an underlying homogeneous Poisson process in the plane by supposing that each point is independently marked with a positive number drawn from some specified distribution, the same for each point. Then each point of the underlying process is replaced by a disk centered at that point and having radius determined by the mark associated with that point. The union of all such disks then constitutes a realization of the desired process which, of course, is not itself a point process; it is an example of a random set process. Both these types of process are examples of *Poisson grain models* or *Boolean models*; (see, for example, [44, Appendix F] or [45, Chapter 3]). One application of such models is to modeling the distribution of cells over a region.

Rather than consider a process of line segments, it is sometimes of interest to consider a process of lines (infinite in length); see [17, Section 10.6] or [27, Chapter 7] for some discussion of such processes.

Martingale Theory of Point Processes

A different approach to point processes is needed for dealing with processes, such as arise in the study of queueing or communication systems or in **survival**

analysis, which evolve dynamically over time and in a manner that may depend on the past history of the process. For a point process N_t, where t represents time and $N_t = N((0, t])$ in our previous notation, the *stochastic intensity function* $\lambda(t|\mathcal{H}_{t-})$ can be defined heuristically by

$$\lambda(t|\mathcal{H}_{t-})\,\mathrm{d}t = \Pr(N(\mathrm{d}t) = 1|\mathcal{H}_{t-}),$$

where $\mathcal{H}_{t-}$ is a history of N_t up to but not including time t. Using the mathematically well-developed theory of martingales (*see* **Counting Process Methods in Survival Analysis**), the *martingale theory of point processes* provides an approach to formalizing the notion of a stochastic intensity and to solving a wide variety of problems by means of the stochastic calculus that results. It is beyond the scope of the present article to enter into details of this extensive and technically rather difficult theory; see [2, 11] and [42], the latter comprehensive work being focused on applications to survival analysis.

Guide to the Literature

The annotated references that follow may be of some assistance. An expanded version of the present article with similar aims but introducing some more advanced topics can be found in Milne [35]. For those interested primarily in spatial data, especially in biological applications, [20] is highly recommended and [32] and [33] may prove useful. The examples in parts of [14, 38, 39] and [44] are also good, and in all these books there is some discussion of theory; [23, Chapter 5] offers a succinct overview of the theory of point processes and is well motivated by examples; [27] is a masterly survey of the many beautiful properties and applications of Poisson processes and a good introduction to many aspects of general point process theory. More detailed exposition of various aspects of the theory can be found in [10, 16], Grandell [12, 22, 37] and [43]. A much more comprehensive, yet readable, presentation of the mathematical theory is given in [17] and Daley & Vere-Jones [18]. A systematic and careful development of the mathematical foundations of point process theory is in [26] and [34], though these works are usually considered difficult, even by probabilists.

Point process theory, with a view to applications in stochastic geometry, is dealt with in [1, 44], and [45]. A good introduction to stochastic geometry, including its connections with point process theory, is provided in [3].

For those interested especially in statistical inference from point process data [13, 14, 20, 38, 44] and [45] contain examples as well as an introduction to relevant theory; in particular, [14, Section 8.6] and [20, Chapters 6, 7] deal with aspects of inference for multitype point process data.

Pre-1973 references are well covered in [15]. There are many references and an excellent coverage of both theory and applications of point processes, as at 1971, in [29].

References

[1] Ambartzumian, R.V. (1990). *Factorization Calculus and Geometric Probability*, Encyclopedia of Mathematics and its Applications, Vol. 33. Cambridge University Press, Cambridge. (Chapters 7 and 8 contain an interesting summary of point process theory, used in Chapters 9 and 10 for applications to random geometry are made.)

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York. (A comprehensive development of the martingale theory of point processes with applications to survival analysis.)

[3] Baddeley, A.J. (1999a). A crash course in stochastic geometry, in *Stochastic Geometry: Likelihood and Computation*, O.E. Barndorff-Nielsen, W.S. Kendall & M.N.M. van Lieshout, eds. Chapman & Hall/CRC, London, pp. 1–35.

[4] Baddeley, A.J. (1999b). Spatial sampling and censoring, in *Stochastic Geometry: Likelihood and Computation*, O.E. Barndorff-Nielsen, W.S. Kendall & M.N.M. van Lieshout, eds. Chapman & Hall/CRC, London, pp. 37–78.

[5] Baddeley, A.J. (2001). Likelihoods and pseudolikelihoods for Markov spatial processes, in *State of the Art in Probability and Statistics: Festschrift for Willem R. van Zwet*, Lecture Notes-Monograph Series, Vol. 36, M.C.M. de Gunst, C.A.J. Klaassen & A.W. van der Vaart, eds. Institute of Mathematical Statistics, Hayward CA, pp. 21–49.

[6] Baddeley, A.J. & Lieshout, M.N.Mvan (1995). Area-interaction point processes, *Annals of the Institute of Statistical Mathematics* **47**, 601–619.

[7] Baddeley, A.J. & Møller, J. (1989). Nearest-neighbor Markov point processes and random sets, *International Statistical Review* **57**, 89–121.

[8] Baddeley, A.J. & Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (With discussion), *Australian & New Zealand Journal of Statistics* **42**, 283–322.

[9] Baddeley, A.J. & Vedel Jensen, E.B. (2004). *Stereology for Statisticians*. Chapman and Hall/CRC, Boca Raton.

[10] Brandt, A., Franken, P. & Lisek, B. (1990). *Stationary Stochastic Models*. Wiley, Chichester. (Adopts a very general approach to the existence and uniqueness of stationary distributions for stochastic systems. Relevant point process theory is summarized in Chapters 2 and 3. Applications, particularly to queueing systems, are explored in later chapters.)

[11] Brémaud, P. (1981). *Point Processes and Queues. Martingale Dynamics*. Springer-Verlag, New York. (A readable account of the martingale approach to point processes. It was the first book to deal with applications of this approach in queueing theory.)

[12] Cox, D.R. & Isham, V. (1980). *Point Processes*. Chapman & Hall, London. (A good introduction to the theory of point processes. Mathematical detail is subservient to presentation of the ideas.)

[13] Cox, D.R. & Lewis, P.A.W. (1966). *Statistical Analysis of Series of Events*. Methuen, London. (A good introduction, now somewhat dated, to aspects of statistical analysis of point processes in which the events occur in a one-dimensional continuum; see also Lewis (1972).)

[14] Cressie, N.A. (1991). *Statistics for Spatial Data*. Wiley, New York. (A comprehensive work covering relevant theory together with applications to geostatistical data, lattice data and spatial patterns. The latter topic includes a useful discussion of point process theory.)

[15] Daley, D.J. & Milne, R.K. (1973). The theory of point processes: a bibliography, *International Statistical Review* **41**, 183–201. (Comprehensive annotated bibliography of pre-1973 references.)

[16] Daley, D.J. & Vere-Jones, D. (1972). A summary of the theory of point processes, in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P.A.W. Lewis, ed. Wiley-Interscience, New York, pp. 299–383. (An excellent survey paper which was the seed for Daley & Vere-Jones (1988).)

[17] Daley, D.J. & Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York. (A fine comprehensive introduction, written so as to be accessible to a wide range of readers. Some history and a simple development of important ideas precedes treatment of general theory.)

[18] Daley, D.J. & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd Ed., Elementary Theory and Methods, Vol. 1. Springer-Verlag, New York. (First volume of an enhanced edition of Daley & Vere-Jones (1988); second in preparation.)

[19] Diggle, P.J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns, *Biometrics* **35**, 87–101.

[20] Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London. (Written especially for biologists, with an emphasis on spatial point processes and appropriate methods for their statistical analysis.)

[21] Geyer, C. (1999). likelihood inference for spatial point processes, in *Stochastic Geometry: Likelihood and Computation*, O.E. Barndorff-Nielsen, W.S. Kendall & M.N.M. van Lieshout, eds. Chapman & Hall/CRC, London, pp. 79–140.

[22] Grandell, J. (1997). *Mixed Poisson Processes*, Monographs on Statistics and Applied Probability, Vol. 77, Chapman & Hall, London.

[23] Guttorp, P. (1995). *Stochastic Modelling of Scientific Data*. Chapman & Hall, London. (A fine presentation of a variety of stochastic process models well motivated by discussion of real applications. Chapter 5 provides an excellent introduction to point processes.)

[24] Karr, A.F. (1988). stochastic processes, point, in *Encyclopedia of Statistical Sciences*, Vol. 8, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 852–859.

[25] Karr, A.F. (1991). *Point Processes and their Statistical Inference*, 2nd Ed. Marcel Dekker, New York. (Primarily a study of inference for point processes. Heavily dependent on martingale methods, but attempts to avoid the analytic detail needed for a strict development of continuous-time martingales.)

[26] Kerstan, J., Matthes, K. & Mecke, J. (1974). *Unbegrenzt teilbare Punktprozesse*. Akademie-Verlag, Berlin. (Revised and expanded as Matthes et al. (1978).)

[27] Kingman, J.F.C. (1993). *Poisson Processes*. Clarendon Press, Oxford. (A fascinating introduction to theory and applications of Poisson processes.)

[28] Kutoyants, Yu. (1998). *Statistical Inference for Spatial Poisson Processes*. Lecture Notes in Statistics, Vol. 134, Springer-Verlag, New York. (Develops estimation and testing for spatial Poisson processes.)

[29] Lewis, P.A.W. (1972). *Stochastic Point Processes: Statistical Analysis, Theory and Applications*. Wiley-Interscience, New York. (A large and useful volume containing the proceedings of a conference held at the IBM Research Center, Yorktown Heights, New York, in August 1971.)

[30] Lieshout, M.N.M. van & Baddeley, A.J. (1996). A nonparametric measure of spatial interaction in point patterns, *Statistica Neerlandica* **50**, 344–361.

[31] Lindsey, J.K. (1996). *Parametric Statistical Inference*. Oxford University Press, Oxford. (A modern presentation of parametric statistical inference emphasizing the central role of the likelihood function. Mostly nontechnical, with a focus on statistical ideas. Contains a summary of exponential family theory.)

[32] Matérn, B. (1960). *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*. *Meddelanden fran Statens Skogsforskningsinstitut* **49**(5), pp. 1–144. (The author's doctoral thesis, it contains an interesting mix of theory and applications, very much ahead of its time.)

[33] Matérn, B. (1986). *Spatial Variation*, 2nd Ed., Lecture Notes in Statistics, Vol. 36. Springer-Verlag, Berlin. (Published version of Matérn (1960).)

[34] Matthes, K., Kerstan, J. & Mecke, J. (1978). *Infinitely Divisible Point Processes*. Wiley, Chichester. (One of the most detailed and comprehensive treatments of general point process theory, though not for beginners. Intuition

is left largely to the reader. Kerstan et al. (1974) was like a first edition.)

[35] Milne, R.K. (2001). point processes and some related processes, in *Stochastic Processes: Theory and Methods*, Handbook of Statistics, Vol. 19, D.N. Shanbhag & C.R. Rao, eds. Elsevier, Amsterdam, pp. 599–641.

[36] Møller, J. (1999). Markov chain Monte Carlo and spatial point processes, in *Stochastic Geometry: Likelihood and Computation*, O.E. Barndorff-Nielsen, W.S. Kendall & M.N.M. van Lieshout, eds. Chapman & Hall/CRC, London, pp. 141–172.

[37] Reiss, R.-D. (1993). *A Course on Point Processes*. Springer-Verlag, New York. (A well-structured graduate-level introduction to basic theory. Applications considered include nonparametric curve estimation, sampling from finite populations, models for exceedances, spatial data and image analysis.)

[38] Ripley, B.D. (1981). *Spatial Statistics*. Wiley, New York. (A good account of the theory and statistical analysis of spatial processes, including spatial point processes.)

[39] Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge. (Published version of a prize-winning essay which goes beyond the author's 1981 book to discuss the special challenges of statistical inference in this context.)

[40] Rydén, T. (1995). Consistent and asymptotically normal parameter estimates for Markov modulated Poisson processes, *Scandinavian Journal of Statistics* **22**, 295–303.

[41] Sigman, K. (1994). *Stationary Marked Point Processes: An Intuitive Approach*. Chapman & Hall, New York. (A detailed study of stationarity and its applications in queueing theory. The development is mathematical but with a strong intuitive flavor.)

[42] Snyder, D.L. & Miller, M.I. (1991). *Random Point Processes in Time and Space*, 2nd Ed. Springer-Verlag, New York. (A good account of the theory and applications of point processes from an engineering viewpoint. The first edition (1975, Snyder only) was written before the martingale theory was well developed.)

[43] Srinivasan, S.K. (1974). *Stochastic Point Processes*. Griffin, London. (Presents a useful heuristic account of the theory of point processes, outlining some applications in statistical physics, queueing and reliability, and biology.)

[44] Stoyan, D. & Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics*. Wiley, Chichester. (The last part of this wide-ranging three-part work contains an exposition of the theory of point fields (point processes), including discussion of many interesting examples. Intended as an introduction for nonmathematicians.)

[45] Stoyan, D., Kendall, W.S. & Mecke, J. (1995). *Stochastic Geometry and Its Applications*, 2nd Ed. Wiley, Chichester. (General theory of point processes and random measures is developed to support study in the remaining chapters of aspects of stochastic geometry.)

ROBIN K. MILNE

Poisson Distribution

The Poisson distribution arises as a limiting form of **binomial distribution**. It is named after the French mathematician, **Siméon Denis Poisson** (1781–1840), but was introduced earlier in 1718, by **A. De Moivre**, who presented the limiting result in approximate form.

The Poisson distribution is also important in its own right; for instance, when rare events of some sort occur randomly in time or space. The variable of interest is the number of events observed in a continuous interval. The interval may be an interval of time, but it may also refer to some other measure of length, or it may be an area or volume segment.

The Poisson distribution is used in many areas of medical research and, in particular, in toxicology, bacteriology and epidemiology. Examples include the number of abnormal cells in a fixed area of a histological slide, the count of bacteria surviving treatment in a fixed volume of bacterial suspension, the number of white blood cells in a drop of blood, the number of new breast cancer cases registered per month by the National Cancer Registry or the number of live births in Greater London during the month of January.

In this article we use nontechnical and heuristic arguments. More detailed information can be found in [3] and [4].

Assumptions

For count data to follow a Poisson distribution three conditions need to be met:

1. In any very small interval (smaller, say, than a millisecond or a cubic nanometer) the probability of an event occurring is proportional to the size of the interval.
2. The probability that the interval contains two or more events gets smaller as the interval gets smaller and can, therefore, for all practical purposes, be ignored.
3. What happens in any small interval is independent of what happens in any other small interval that does not overlap the first.

These conditions imply that events occur over time or space at a constant rate on average, each event occurring independently and at random.

Probability Distribution

We give a heuristic derivation of the Poisson probability distribution using the context of white blood cell counts.

Suppose that Ms Smith has on average 6000 white blood cells in a cubic millimeter of blood. It is reasonable to assume that the probability of finding a white cell in a small drop of blood is proportional to the size of the drop and that, for a sufficiently small drop, the probability of finding two or more white cells is negligible. Since cells move independently, the presence of a white cell in one small drop of blood is not expected to affect the presence or absence of a white cell in any other nonoverlapping drop of the same size. A drop of 0.0005 cubic millimeter ($V = 0.5 \times 10^{-3}\text{mm}^3$) of blood from Ms Smith is examined. If several such drops were examined, some of them would, by random variation alone, contain no white blood cells, others would contain one, or two, and some would contain as many as, say, eight white blood cells. We would expect the average number of white blood cells to be three when the average is taken over a large number of drops of size 0.0005 cubic millimeters ($0.5 \times 10^{-3}\text{mm}^3 \times 6000\text{cells/mm}^3 = 3$ cells). We write $\mu = 3$.

We make a hypothetical split of the 0.0005 cubic millimeter drop of blood into yet smaller subdrops, say, $n = 60$ or 100 or 300 subdrops of equal size. The size of each subdrop ($V/n\text{mm}^3$) gets smaller when the number n gets larger. When the subdrops are sufficiently small, most of them will contain no white blood cells, some will contain one and it is very unlikely that any subdrop will contain more than one white blood cell. The small subdrops are approximating a sequence of n binomial trials, in each of which there is a probability μ/n of finding a white blood cell and a probability $1 - (\mu/n)$ of not finding one. The probability that in the whole series of n subdrops there are exactly x white blood cells is given by the binomial probability

$$\frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x}. \quad (1)$$

The assumptions underlying this binomial probability get more accurate as n increases. What happens when n increases indefinitely? We can replace $n(n-1)(n-2)\cdots(n-x+1)$ by n^x, since x will be negligible in comparison with n. We can also replace $[1-(\mu/n)]^{n-x}$ with $[1-(\mu/n)]^n$ since $[1-(\mu/n)]^x$ will

approach one as n increases. It is a standard mathematical result that $[1 - (\mu/n)]^n$ approaches $e^{-\mu}$ when n increases indefinitely, e being the base of the natural (or Napierian) logarithms ($e = 2.718\ldots$). Thus, when n increases indefinitely, the binomial probability in (1) takes the form

$$\frac{n^x}{x!}\left(\frac{\mu}{n}\right)^x e^{-\mu},$$

which equals

$$P_x(\mu) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2\ldots. \qquad (2)$$

and defines the Poisson probability distribution.

Interestingly, the limiting approximation derived by De Moivre did not involve the Napierian e, since this was not introduced until the days of Euler (1707–1783).

Table 1 illustrates how the binomial distribution approaches the Poisson when n gets larger and the binomial probability $\pi = \mu/n$ gets smaller, while the mean $\mu = n\pi$ is kept constant.

Matsunawa [5] gives accuracy bounds for the Poisson–binomial approximation. A rule of thumb is that the Poisson approximation to the binomial is good if $n \geq 20$ and $\pi \leq 0.05$ and very good if $n \geq 100$ and $\pi \leq 10$.

The Poisson probabilities in (2) can also be derived from a **Poisson process** by solving a set of differential equations.

Table 1 Probabilities for three binomial distributions with $n\pi = 3$ and for the Poisson distribution with mean 3

Number of events x	$\pi = 0.05$ $n = 60$	$\pi = 0.03$ $n = 100$	$\pi = 0.01$ $n = 300$	Poisson mean $\mu = 3$
0	0.046	0.048	0.049	0.050
1	0.145	0.147	0.149	0.149
2	0.226	0.225	0.224	0.224
3	0.230	0.227	0.225	0.224
4	0.172	0.171	0.169	0.168
5	0.102	0.101	0.101	0.101
6	0.049	0.050	0.050	0.050
7	0.020	0.021	0.021	0.022
8	0.007	0.007	0.008	0.008
9	0.002	0.002	0.003	0.003
10	0.001	0.001	0.001	0.001
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Properties

A Poisson random variable X takes the values $x = 0, 1, 2, \ldots$ with probabilities defined by (2). Thus, $P_0 = e^{-\mu}$, $P_1 = \mu e^{-\mu}$, $P_2 = \frac{1}{2}\mu^2 e^{-\mu}$, etc. with $P_1 + P_2 + \cdots = 1$. The whole distribution is characterized entirely by the one parameter μ. The **moment generating function** is of the form $E(e^{tx}) = \exp[\mu(e^t - 1)]$, with the first two moments $E(X) = \mu$ and $E(X^2) = \mu^2 + \mu$. Both the mean $E(X)$ and the variance $\text{var}(X) = E(X^2) - [E(X)]^2$ are thus equal to μ.

The shape of the distribution for $\mu = 1$, 3, 7 and 15 is shown in Figure 1. For small values of μ the distribution is skewed, and it gets more symmetric as μ increases. For $\mu \geq 10$, the distribution is close to symmetric.

Normal Limit

When μ increases indefinitely the Poisson distribution approaches the **normal distribution** with mean μ and variance μ. De Moivre derived the normal limit to the binomial distribution, from which the result for the Poisson distribution follows. Using a continuity correction, the probability that a Poisson random variable takes the integer value x is approximated by the probability that a normally distributed random variable takes values between $x - 0.5$ and $x + 0.5$. The probability that a Poisson variable with mean μ takes values greater than, or equal to, x is correspondingly approximated by the tail area of a normal distribution beyond the standardized normal deviate $z = (|x - \mu| - 0.5)/\sqrt{\mu}$.

Variance-stabilizing Transform

The square root **transformation** $Y = \sqrt{X}$ of a Poisson random variable X stabilizes the variance, with $\text{var}(Y) = \frac{1}{4}$ for all values of μ (*see* **Delta Method**; **Power Transformations**). This allows checking the Poisson assumption, and for Poisson random variables with large enough means one can get approximate results using methods appropriate for normally distributed random variables with constant variance.

Sum of Poisson Variables

The Poisson distribution has a reproductive property: let $X_1, X_2, \ldots, X_k$ be k independent Poisson

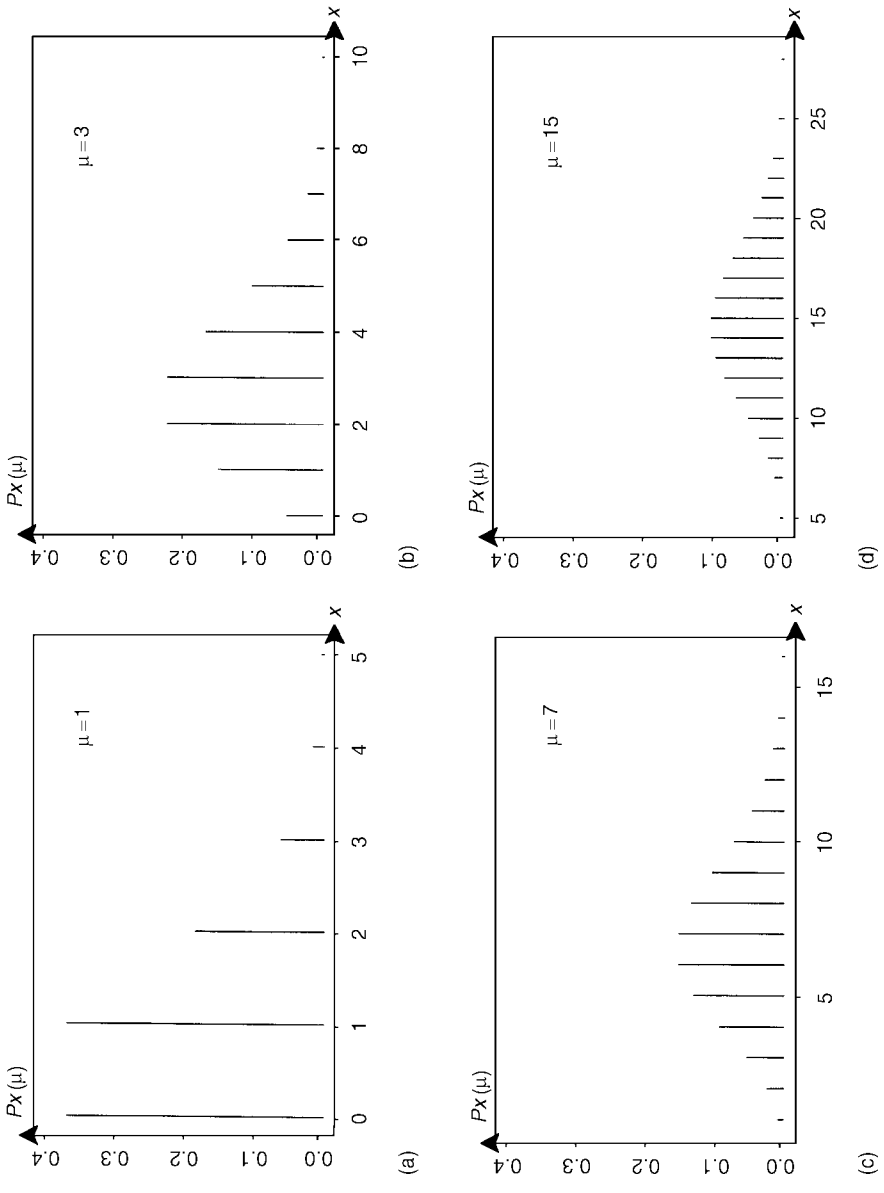

Figure 1 Poisson distribution for various values of the mean μ

random variables with parameters $\mu_1, \mu_2, \ldots, \mu_k$, respectively. Then the sum $X_1 + X_2 + \cdots + X_k$ is also a Poisson random variable with parameter $\mu_1 + \mu_2 + \cdots + \mu_k$.

Binomial and Multinomial

Let X_1 and X_2 be two independent Poisson random variables with parameters μ_1 and μ_2, respectively. The conditional distribution of X_1, given $X_1 + X_2 = n$, is binomial with index n and probability $\mu_1/(\mu_1 + \mu_2)$. Correspondingly, for k independent Poisson variables $X_1, X_2, \ldots, X_k$ with parameters $\mu_1, \mu_2, \ldots, \mu_k$, the conditional distribution, given $X_1 + X_2 + \cdots + X_k = n$, is multinomial with index n and probabilities $\mu_s/(\mu_1 + \mu_2 + \cdots + \mu_k)$, $s = 1, 2, \ldots, k$. This relation between sampling distributions is central to the theory of a **loglinear model** for **categorical data analysis**.

Hypergeometric

Let X_{11}, X_{12}, X_{21}, and X_{22} be four independent Poisson random variables with parameters μ_{11}, μ_{12}, μ_{21} and μ_{22}, respectively. They may be viewed as cell counts in a **2 × 2 contingency table**. Conditional on both margins $X_{i1} + X_{i2} = m_i$ and $X_{1j} + X_{2j} = n_j$, $i, j = 1, 2$, the random variable X_{11} follows a noncentral **hypergeometric distribution** with noncentrality parameter $\theta = \mu_{11}\mu_{22}/\mu_{12}\mu_{21}$. This conditional distribution is used in matched case control studies and in conditional logistic regression. For $\theta = 1$, i.e. when the odds ratio in the 2×2 frequency table is unity, the distribution reduces to the standard hypergeometric distribution.

Heterogeneity

In empirical studies, observed counts often exhibit larger variance than would be expected from the Poisson assumption. One mechanism generating this larger spread is heterogeneity in the average event rate over the population under study. Such **overdispersion** relative to the Poisson was noted already in 1920 by Greenwood & Yule [2], who suggested a model where the mean μ is not constant, but a random variable with a **gamma distribution** (*see* **Accident Proneness**). This leads to a two-parameter **negative binomial distribution** for the count. Mixing distributions other than the gamma distribution have been discussed, as well as distribution-free approaches and quasi-likelihood methods for handling overdispersion. If sources of heterogeneity in the Poisson means are known and measured, then **Poisson regression** methods are appropriate.

Test for Heterogeneity

For a set of k observed counts $X_1, X_2, \ldots, X_k$, it may be of interest to test the hypothesis that they are drawn at random from a single Poisson distribution, as opposed to being drawn from several Poisson distributions with different means. A reasonable test statistic for detecting heterogeneity would compare the spread of the observed counts relative to their average. Let $\overline{X} = \sum_{i=1}^{k} X_i/k$ denote the sample mean. Under the null hypothesis of common mean μ the Poisson *heterogeneity* or *dispersion* test statistic:

$$T = \frac{\sum_{i=1}^{k}(X_i - \overline{X})^2}{\overline{X}} \tag{3}$$

is approximately distributed as a **chi-square distribution** with $k - 1$ degrees of freedom (*see* **Chi-square Tests**). The heterogeneity test in (3) was first introduced by "Student" in 1907 (*see* **Gosset, William Sealy**). Armitage & Berry [1] justify it from two different points of view: for normally distributed observations with constant variance the sum of squares divided by the variance follows a χ^2 distribution. Here $\hat{\mu} = \overline{X}$ is an estimator for the variance under the null hypothesis. One may also view (3) as the usual χ^2 test statistic $T = \sum_{i=1}^{k}(X_i - \hat{\mu}_i)^2/\hat{\mu}_i$, with X_i the observed count and $\hat{\mu}_i = \overline{X}$ the corresponding expected count under the null hypothesis.

In practice, it may be easy to show that a given sample of counts does not originate from a Poisson distribution. The result is useful, since it reveals the presence of some kind of nonrandomness. If the Poisson distribution is rejected, then a distribution involving two (or more) parameters is needed to describe the data. It may, however, take a very large sample to distinguish between, for example, the negative binomial and some other two-parameter distributional form. In general, one should view determination of a specific mechanism for nonrandomness as a biological, rather than a statistical, problem (*see* **Contagious Distributions**).

References

[1] Armitage, P.& Berry, G. (1994). *Statistical Methods in Medical Research*, 3rd Ed. Wiley, New York.

[2] Greenwood, M.& Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society* **83**, 255–279.

[3] Haight, F.A. (1967). *Handbook of the Poisson Distribution*. Wiley, New York.

[4] Johnson, N.L.& Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Wiley, New York.

[5] Matsunawa, T. (1982). *Some strong ε-equivalence of random variables*, *Annals of the Institute of Statistical Mathematics* **34**, 209–224.

(*See also* **Generating Functions**; **Poisson Regression**)

JUNI PALMGREN

Poisson Processes

The **Poisson distribution** has often been rediscovered, but goes back to the work of the French mathematician S.-D. Poisson (1837). It arises as a limiting form of the **binomial distribution**, when a rare event is observed occasionally in a large number of repetitions. Thus, suppose that there are a number n of repetitions of some random experiment, the repetitions being statistically independent. Suppose that there is some outcome of the experiment which we describe as a "success", and which has a constant probability p. The probability that there are exactly r successes in the n trials is then given by the binomial probability

$$b(n, p; r) = \binom{n}{r} p^r q^{n-r}, \quad r = 0, 1, 2, \ldots, n. \tag{1}$$

If n is large and p is small, in such a way that the mean number of successes $\mu = np$ is moderate, then $b(n, p; r)$ is well approximated by the Poisson probability

$$\pi_r(\mu) = \frac{\mu^r \mathrm{e}^{-\mu}}{r!}. \tag{2}$$

In fact, it is simple to prove that, as n tends to infinity for fixed μ,

$$b\left(n, \frac{\mu}{n}; r\right) \longrightarrow \pi_r(\mu). \tag{3}$$

A famous example of the occurrence of the Poisson distribution (2), is the stream of particles emitted from a piece of radioactive material and detected in a Geiger counter. One can think of each atom as being an independent trial, with a very small probability of emission in a given interval of time. There are many atoms, and the total number of particles emitted in a time interval will thus have the distribution of (2), where μ is the mean number of emissions in that interval and will normally be proportional to the length of the interval.

This is an example of a Poisson process occurring in one (time) dimension. If we mark out along a time axis the instants at which particles are emitted, then we have a random set of points with the property that, if $N(a, b)$ denotes the number of points between a and b, then for any $a < b$,

$$\Pr[N(a, b) = r] = \pi_r(\mu), \quad r = 0, 1, 2, \ldots, \tag{4}$$

where

$$\mu = \lambda(b - a). \tag{5}$$

Moreover (and this is an essential part of the definition) the random variables $N(a, b)$ for disjoint intervals (a, b) are independent (where disjoint means that no two overlap).

Many other random series of points, such as mutations at a chromosome locus (*see* **Gene**), or arrivals at a queue, are found to be, more or less exactly, Poisson processes in this sense. In many cases, however, the rate λ is not constant, and then (5) must be generalized to

$$\mu = \int_a^b \lambda(t)\, \mathrm{d}t, \tag{6}$$

where $\lambda(t)$ represents the instantaneous rate of occurrence at time t. The best way to visualize this is that $\lambda(t)h$ is, for small h, the probability that at least one of the random points occurs between t and $t + h$.

The one dimension need not be time, but could be spatial, and there is an obvious generalization to two or more dimensions. For instance, the stars visible in the sky, or the (centers of the) spots of a skin rash, might be modeled as random sets of points in two dimensions. The general structure is simple, although there are technical mathematical niceties which are here ignored. We have a space S, which might be ordinary space of some dimension but could be more complicated. The object of interest is a random set of points in S, sufficiently dispersed that for typical subsets A of S the number of these points falling in A is a finite random variable $N(A)$. (Think of S as the plane, and the sets A as being bounded sets like circles, triangles, and rectangles.) The random set of points is then called a Poisson process if

1. the random variable $N(A)$ has a Poisson distribution (2), where $\mu = \mu(A)$ depends on A
2. for disjoint sets $A_1, A_2, \ldots, A_n$, the random variables $N(A_j)(j = 1, 2, \ldots, n)$ are independent.

In most cases the mean

$$\mu(A) = \mathrm{E}[N(A)] \tag{7}$$

is given in terms of a rate function $\lambda(x)$ on S by

$$\mu(A) = \int_A \lambda(x)\, \mathrm{d}x (A \subseteq S), \tag{8}$$

but it can be a general (nonatomic) measure on S.

The simple structure of the Poisson process leads to a number of properties which are extremely powerful in modeling random sets of points. Suppose, for instance, that we have two independent Poisson processes on the same S, and that the numbers of their points falling in a set A are, respectively, $N(A)$ and $N'(A)$. Then superposing these two sets gives a new random set, which also turns out to be a Poisson process. The number of points of the new process falling in A is $N(A) + N'(A)$, which has expectation $\mu(A) + \mu'(A)$.

Again, a mapping from one space to another preserves the Poisson property. If we have a Poisson process on a space S, and a mapping f from S to another space S^*, then f maps the random set of points into (usually) a Poisson process on S^*, and the expectation measure maps in a simple way. The only thing that can go wrong is that the induced measure on S^* may have atoms, in which case f may pile points on top of one another.

Often the random points of the Poisson process come with numerical values attached to them; the stars have their magnitudes, the arrivals at a queue their service requirements. This leads to the concept of a marked Poisson process, where a mark, which may be numerical or more complicated, is associated with each point. If the marks of different points are independent random variables (with distributions depending on the positions of the points), there is a powerful product space representation from which detailed calculations can be made. Thus, a point of the process at x, with a mark m, can be plotted as a random point (x, m) in the product space $S \times M$, where M is the space of possible marks. Then it can be shown that this random set of points is a Poisson process on the product space, and there is an explicit formula for the expected number of points in a given subset.

Suppose that a swarm of bees attacks an unfortunate victim. It might be that the instants of successive stings form a (one-dimensional) Poisson process, and that the amount of venom in each sting is a random quantity independent from bee to bee whose distribution might depend on the time of the sting. Then we have a marked Poisson process, and the product space representation is two-dimensional, each bee being represented by a point whose coordinates are the time at which it stings and the amount of venom. The victim might be interested in the total amount of venom he receives. This is an example of a Poisson sum of the form

$$\sum g(x), \tag{9}$$

where the sum is taken over all the points x of a Poisson process on S, and g is a real-valued function on S. A result known as Campbell's theorem, after the work of N.R. Campbell (1909), gives an explicit formula for the **moment generating function** of the sum (9). This and details and proofs of the properties already cited may be found in [2].

All the results so far described apply whatever the dimension of the space S. For some problems S can be quite a complicated geometrical object. For example, some fibrous substances (such as paper) when viewed microscopically resemble a random array of lines in three-space, and this can be represented as a random set of points in a four-dimensional manifold. However, there are properties which are peculiar to Poisson processes on the line. A one-dimensional Poisson process can always be mapped, by an increasing function, into a Poisson process which is homogeneous in the sense that it satisfies (4) and (5) for some λ. A homogeneous Poisson process is a renewal process in the sense that the intervals between successive points are independent and identically distributed, their common probability density being

$$\lambda \exp(-\lambda x), \quad x > 0. \tag{10}$$

This property is characteristic of the one-dimensional homogeneous Poisson process, and has no analogue in higher dimensions.

One useful aspect of this special case is that, in one dimension, it is easy to construct an alternative model for a random set of points (or **point process** as it is then usually called) which is not Poisson. It is only necessary to change the interval distribution from (10), or to introduce some dependence between successive intervals. It is much more difficult, in more than one dimension, to devise usable models for non-Poisson point processes. There are, however, variants of the Poisson process which can sometimes provide useful models.

One of these is the Poisson cluster process, in which each of the points of a Poisson process gives rise to a random number of daughters, the daughters being dispersed about the parent point in a random way (think of seedlings around trees). Another is the Cox process, which is a Poisson process whose rate function is itself a random process. These are, in principle, very general classes of point processes,

but the generality is paid for in great complexity of calculation. At the same time, they have special properties which limit their usefulness; for instance, a Cox process is always overdispersed in the sense that the variance of the number of points in any set always exceeds its mean (there is equality for a Poisson process), so that no Cox process could be a realistic model for a random set which was underdispersed.

The other significant aspect of the theory of Poisson processes, which is not covered in [2], is that of **limit theorems** and approximations. The Poisson distribution started from a limit result (3), and there are many situations where modifying a non-Poisson process leads approximately or asymptotically to a Poisson process. The best starting point of this aspect is [1].

References

[1] Barbour, A.D., Holst, L. & Janson, S. (1992). *Poisson Approximation*. Clarendon Press, Oxford.
[2] Kingman, J.F.C. (1993). *Poisson Processes*. Clarendon Press, Oxford.

(*See also* **Stochastic Processes**)

J.F.C. KINGMAN

Poisson Regression in Epidemiology

Various authors [3, 9, 11, 12] have noted that **Poisson regression** can be used to analyze cohort survival data (*see* **Cohort Study**). This formulation also leads to a unification of **risk** estimation based on internal comparison of rates among members of a cohort with various exposure levels and classical epidemiologic methods based on external rates that yield standardized mortality ratios or standardized incidence ratios [2, 5] (*see* **Standardization Methods**).

Poisson regression is an important alternative to **partial-likelihood**-based analysis of the **proportional hazards** model (*see* **Cox Regression Model**) and to parametric analyses of such models (*see* **Survival Analysis, Overview**) for two main reasons. First, it provides an efficient and intuitive method for dealing with cumulative exposures and other **time-dependent covariates** and for allowing risk to depend on multiple time scales (e.g. attained age, time since exposure, or calendar time). Secondly, it facilitates the consideration of a broad range of risk models including those that allow for the direct parametric description of baseline rates, absolute excess rates, and **relative risks**.

Breslow & Day [4] offer a general discussion of the use of Poisson regression in the analysis of cohort survival data. Some of the most extensive applications of these methods have involved studies of radiation effects on mortality and cancer incidence in the atomic bomb survivors [14].

Poisson Regression of Survival Data

The data from cohort survival studies typically consist of information on whether or not the event of interest occurred, the event or censoring time, t, and a vector of possibly time-dependent covariates, $\mathbf{z}$, for each cohort member. Since interest centers on **hazard rates** it is natural and useful for the purposes of analysis or summarization to reorganize such data into an event–time table defined by a cross-classification over a set of time intervals and covariate categories. The data for each cell in such a table include the total number of events, c_{is}, the total time (person-years) at risk, R_{is}, and representative values of the covariates, z_{is} for time period i and category s. For each cell the ratio of the number of events to the time at risk is a crude hazard rate. The analysis involves **regression** methods to smooth these rates as a function of time and other **covariates**.

When such tables are produced as simple summaries of a data set, it is common to limit the number of time periods and other factors used to define the table. However, for modeling rates it is appropriate to use detailed tables with many cells based on a relatively fine **stratification** over time and other factors. For example, a rate table to be used in an analysis of an occupational cohort study (*see* **Occupational Epidemiology**) might be defined in terms of age, year, age at first exposure, sex, and cumulative exposure with hundreds or even thousands of cells. An event–time table for a **clinical trial** might involve follow-up time, age at entry, sex, and treatment. Although not usually necessary in practice, the methods can be applied to a table based on individual subjects where the only grouping is on time. This suggests the close connection between the use of Poisson regression methods for the analysis of rates and the Andersen–Gill **counting process** method [1] for analysis of hazard functions.

If it is assumed that the hazard, λ_{is} is constant within each cell, then the expected number of events in the cell is given by

$$\mathrm{E}(c_{is}) = R_{is} \times \lambda_{is}.$$

In terms of a parametric function, $\lambda(t_i, z_{is}, \theta)$ for the rates, the log **likelihood** for the survival data under the piecewise constant hazard assumption is

$$\sum_{l,s} c_{is} \ln(\lambda(t_l, z_{is}, \theta)) - R_{is} \times \lambda(t_i, z_{is}, \theta),$$

which is equivalent to the log likelihood that would arise if the event counts in the table were independent **Poisson** random variables. Thus, Poisson regression can be used to estimate the parameters in this model.

With this approach, modeling rates in terms of time is straightforward since, in contrast to Cox regression, there is no distinction between time-dependent and time-independent covariates. This is because the time-dependent computations are carried out when the event – time table is constructed and are not repeated each time a model is fitted.

Using External Rates or Expected Cases

In some situations one has external data on the expected rates λ_q^e stratified by time and other factors (e.g. age, calendar time period, and sex but not exposure or treatment related factors). In this case, it is possible to compute the expected number of cases for each cell in the table as $C_{is}^e = R_{is} \times \lambda_q^e$, where $q = q(is)$ denotes the external rate strata corresponding to cell is. In this case, Poisson regression can be used to model the relative hazard, ρ_{is}, since

$$\mathrm{E}(c_{is}) = C_{is}^e \times \rho_{is} = R_{is} \times \lambda_q^e \times \rho_{is}.$$

When **person-years** are replaced by expected numbers of cases, this type of analysis is known as the subject-years method or standardized mortality ratio (SMR) regression [4].

Models for Rate Regression

Following the pioneering work of Cox [8], the most commonly used hazard function model is the **log-linear** proportional hazards model

$$\lambda(t, z, \theta) = \lambda_0(t, \alpha) \times \exp(\beta z). \tag{1}$$

Here λ_0 is a baseline hazard for an individual with covariate $z = 0$.

Other models are also important, however. For example, in **dose–response** studies it is often useful to consider models in which the **excess relative risk** is a linear function of dose d; that is

$$\lambda(t, z, \theta) = \lambda_0(t, \alpha) \times (1 + \beta d).$$

Preston [16] has described a flexible general class of parametric **additive hazard models** of the basic form

$$\lambda_0(t, \alpha, z_0) + \lambda_{\mathrm{EAR}}(t, \beta, z_1) \tag{2}$$

and

$$\lambda_0(t, \alpha, z_0)[1 + \lambda_{\mathrm{ERR}}(t, \beta, z_1)], \tag{3}$$

in which λ_0 represents the baseline or background rates and λ_{EAR} and λ_{ERR} describe the excess absolute or excess relative risks. In these models baseline rates are usually assumed to be loglinear functions of the covariates while the excess risks are modeled as linear or products of linear and loglinear functions of the covariates.

One reason for the popularity of the Cox regression model is that it allows one to focus (perhaps too much) on the **relative risk** while treating the baseline hazard as completely unspecified. A similar simplification is possible in the analysis of **relative risk models** for rates using Poisson regression. This is accomplished by the inclusion of a **multiplicative** parameter for each time interval leading to models such as

$$\tau_i \exp(\beta z) \quad \text{or} \quad \tau_i(1 + \beta d). \tag{4}$$

This approach can also be extended to allow stratification over additional factors, in which case the model is similar to the stratified Cox regression model. Preston et al. [17] describe an efficient **algorithm** for models with large numbers of stratum parameters.

Parameter Estimation and Inference

Parameter estimates for Poisson regression models are computed using **maximum likelihood** methods. Models in which the rates depend on the parameters through a linear function βz, are **Generalized Linear Models (GLM)**. Parameter estimates for GLMs can be computed using iteratively reweighted **least squares** with person-years (or cases for subject-years analysis or standardized mortality/incidence ratio regression) as an "offset". These methods are available in all of the major statistical packages including GLIM, SAS, and S-PLUS (*see* **Software, Biostatistical**). However, the more general rate function models such as (3) and (4) are not GLMs. In this case, it is necessary to make use of special software to define the likelihood and possibly its derivatives. The Epicure package [17] is designed to work with models in the general class described by (1)–(4) above.

Inference about parameters of interest can be carried out using the standard asymptotic methods, including Wald, score, and **likelihood ratio tests**. However, because of the nonlinear nature of the models and, in many applications, the limited information on **excess risks**, asymptotic **standard errors** and hence **hypothesis tests** and **confidence intervals** based on Wald tests can be misleading. Score or likelihood ratio tests and **profile-likelihood**-based confidence intervals should be emphasized when working with additive hazard models.

An important issue concerns the assessment of **goodness of fit** for Poisson regression models derived from detailed event–time tables. Because rate modeling often involves relatively rare events and event–time tables with many cells, the rates or the number of events in each cell of the table can be quite small. In this case, neither the global deviance nor the Pearson **chi-square statistic** provides reasonable guidance as to goodness of fit. The total deviance is often much smaller than the putative **degrees of freedom** (the number of cells in the table minus the number of free parameters in the model). Pregibon [15] developed generalized regression **diagnostics** that can be used for regression models in exponential families. While such diagnostics may be useful in looking for lack of fit and other problems with fitted models [10], they should be interpreted with caution since the underlying data are not independent Poisson counts. In view of these issues, the most effective general method for the assessment of goodness of fit when using Poisson regression to analyze rates is to make use of likelihood ratio tests designed to detect specific departures from models of interest, such as time dependence or nonlinearity, or to make use of **Akaike's criterion** or related statistics to compare alternative (possibly nonnested) models.

Creating Event–Time Tables

The creation of an adequate event–time table is often the most difficult aspect of carrying out analyses of rates using Poisson regression. Among other features, an ideal program for the construction of event–time tables would:

1. allow for categorization on multiple time scales (age, year, length of follow-up, etc.), as well as multiple time-independent and time-dependent factors with variable length intervals in each of these scales;
2. allow for late entry, disjoint follow-up intervals, and multiple events;
3. include procedures for the computation of and categorization on time-dependent quantities;
4. allow computation and storage of counts for multiple event types along with representative values (often time-at-risk weighted means) for covariates of interest for each cell in the table;
5. have efficient procedures for handling the large, sparse tables that can arise when one stratifies on multiple time scales;
6. be able to deal with the data structures that can arise in describing complex exposure histories; and
7. facilitate the incorporation of external rates.

Several computer programs are currently available for the creation of event–time tables. However, many of these programs, e.g. OCMAP [6] and O/E [13], are designed for specific applications and are of limited use in more general problems. Procedures for the creation of such tables in the major statistical programs are extremely limited or nonexistent. The DATAB module in Epicure [17] and "Person-years" [7] are probably the most flexible general-purpose programs for event–time tabulation available at this time. Hopefully, there will be major improvements in this area over the next few years.

Summary

Poisson regression is a powerful tool for the analysis of rates from cohort survival studies that facilitates simple, straightforward analyses of temporal patterns, baseline risks, excess relative or **absolute risks**, and other aspects of hazard functions that may be difficult to assess with other methods. The application of Poisson regression requires that data on individual subjects be organized into event–time tables stratified on time and other factors of interest and, for the most interesting models, specialized software capable of dealing with nonlinear Poisson regression models is also required. The tools needed to conduct these analyses are available today but it is likely that they will be more fully developed in the years to come.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Berry, G. (1983). The analysis of mortality by the subjects years method, *Biometrics* **39**, 173–184.

[3] Breslow, N.E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–99.

[4] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*. IARC Scientific Publications 82, International Agency for Research on Cancer, Lyon.

[5] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[6] Caplan, R.J., Marsh, G.M. & Enterline, P.E. (1983). A generalized effective exposure modeling program for assessing dose-response in epidemiologic investigations, *Computers in Biomedical Research* **16**, 587–596.

[7] Coleman, M. (1986). Cohort study analysis with a FORTRAN computer program, *International Journal of Epidemiology* **15**, 134–137.

[8] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[9] Frome, E. (1983). The analysis of rates using Poisson regression models, *Biometrics* **39**, 665–675.

[10] Frome, E. & Morris, M.D. (1989). Evaluating goodness of fit of Poisson regression models in cohort studies, *American Statistician* **43**, 144–147.

[11] Holford, T.R. (1980). The analysis of rates and survivorship using log-linear models, *Biometrics* **36**, 299–305.

[12] Laird, N. & Oliver, D. (1981). Covariance analysis of censored survival data using log-linear models, *Journal of the American Statistical Association* **76**, 231–240.

[13] Monson, R.R. (1974). Analysis of relative survival and proportional mortality, *Computers in Biomedical Research* **7**, 325–332.

[14] Pierce, D.A., Shimizu, Y., Preston, D.L., Vaeth, M. & Mabuchi, K. (1996). Studies of the mortality of atomic bomb survivors. Report 12, Part I. Cancer Mortality 1950–1990, *Radiation Research* **146**, 1–27.

[15] Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics* **9**, 705–724.

[16] Preston, D.L. (1990). Modeling radiation effects on disease incidence, *Radiation Research* **124**, 343–344.

[17] Preston, D.L., Lubin, J., Pierce, D.A. & McConney, M.E. (1993). *Epicure, Users' Guide*. Hirosoft International, Seattle.

DALE L. PRESTON

Poisson Regression

For response variables that have counts or frequencies as outcomes it is often reasonable to assume an underlying **Poisson** distribution and describe the impact of **explanatory variables** on their means by some **regression** function. Poisson regression models, as a widely applicable class of models particularly useful in biostatistics, emerged in the late 1970s; see, for example, [6, 11, 21–25, 28, 29, 31], and [32].

As an example consider the data given in Table 1, taken from [27]. Randomly chosen household members from a **probability sample** of Oakland, CA, were asked to note which stressful events had occurred within the last 18 months and to report the month of occurrence of these events. A scattergram of the data indicates a decline of recalls as events lie farther in the past, possibly due to the fallibility of human memory (see Figure 1). To define a Poisson regression model, assume that (i) the number of recalls is a random variable Y distributed as Poisson with mean μ, and (ii) μ is some function of X, the number of months before interview. Plotting logarithms of frequencies against months suggests a linear relationship

$$\log \mu = \alpha + \beta x.$$

For this **loglinear model**, the mean satisfies the exponential relationship,

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}.$$

A one-unit increase in X has a multiplicative effect of e^{β} on μ, i.e. the mean of Y at $x + 1$ equals the mean of Y at x multiplied by e^{β}.

Most of the widely available software packages are capable of fitting **generalized linear models**, and can be used to obtain **maximum likelihood** estimates for the parameters of Poisson regression models as well. For these data one finds $\hat{\alpha} = 2.803$ and $\hat{\beta} = -0.0838$; hence

$$\hat{\mu} = 16.5 \times 0.920^{x},$$

indicating a negative trend in time.

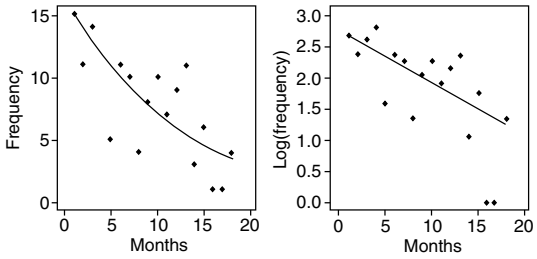

Figure 1 Scattergram of observed frequencies and their logarithms against months before interview. Solid lines represent fitted means and respective values for the linear predictor for the Poisson regression model mentioned in the text

Using the relationship between the **multinomial** and conditional Poisson distributions, this is shown to be equivalent to an exponential decay model for the probability of remembering an event. For a more detailed discussion, see [27] or [33].

Definition

To define the basic version of a Poisson regression model, suppose that we have observations $y_1, \ldots, y_n$ for the response variable $Y_1, \ldots, Y_n$, assumed to be independently distributed Poisson variates with means $\mu_1, \ldots, \mu_n$, i.e.

$$f(y_i|\mu_i) = \frac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i). \qquad (1)$$

The systematic component of the model is specified by some regression function η, depending on regression parameters $\beta_1, \ldots, \beta_k$, with each component relating values $x_{i1}, \ldots, x_{ik}$ of explanatory variables to respective means, i.e.

$$\mu_i = \eta_i(\beta) = \eta_i(x_{i1}, \ldots, x_{ik}; \beta_1, \ldots, \beta_k). \qquad (2)$$

Table 1 Distribution by months prior to interview of stressful events reported from subjects: 147 subjects reporting exactly one stressful event in the period from 1 to 18 months prior to interview. Reprinted from [27, p. 3] by permission of Academic Press, Inc.

Months	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Number	15	11	14	17	5	11	10	4	8	10	7	9	11	3	6	1	1	4

Often, this relationship is such that some monotone transformation g of the means is connected to a *linear predictor* of explanatory variables,

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j.$$

In this situation g is called the *link function* and the model defined in this manner is an instance of a **generalized linear model** (see [35] and [36] or, for an introductory text, [18]). For $\eta_i(\beta) = \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)$ we have the familiar loglinear model,

$$\log \mu_i = \sum_{j=1}^{k} x_{ij}\beta_j.$$

For the model specified by the stochastic component (1) and regression function (2), the log **likelihood** function is written as

$$\ell_y(\beta) = \sum_{i=1}^{n} \{y_i \log[\eta_i(\beta)] - \eta_i(\beta) - \log(y_i!)\}. \quad (3)$$

It may be worthwhile noting that this reduces to a k-parameter **exponential family** log likelihood,

$$\ell_y(\beta) = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} x_{ij} y_i\right) \beta_j - \sum_{i=1}^{n} \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right) - \sum_{i=1}^{n} \log(y_i!), \quad (4)$$

with jointly **sufficient statistics** $\sum_{i=1}^{n} x_{ij} y_i$, $j = 1, \ldots, k$, if the model is log linear.

Some Special Cases

Loglinear Models for Contingency Tables

Suppose, in obvious notation, y_{ij} with indices $i = 1, \ldots, I$ and $j = 1, \ldots, J$ form a two-dimensional **contingency table**, according to some classifying factors A and B having I and J categories, respectively. A common method for analyzing data of this kind is to assume that cell frequencies Y_{ij} are independently distributed as Poisson and to use loglinear models, where in an **analysis-of-variance**-like fashion logarithms of expected cell frequencies μ_{ij} are assumed to be sums of several effects, e.g. for the **multiplicative model**,

$$\log(\mu_{ij}) = \beta_o + \beta_i^A + \beta_j^B, \quad (5)$$

subject to some constraints on the βs. Sums of independent Poisson variates are again distributed as Poisson with means equal to the sum of respective means. Row totals Y_{i+}, column totals Y_{+j}, and grand total Y_{++} are, therefore, Poisson variates with means $\mu_{i+} = \mu_{i1} + \cdots + \mu_{iJ}$, $\mu_{+j} = \mu_{1j} + \cdots + \mu_{Ij}$, and $\mu_{++} = \sum_{i,j} \mu_{ij}$, respectively. Under the assumption of the multiplicative model these quantities are related by

$$\mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{\mu_{++}},$$

showing that the joint distribution of the contingency table is, in a multiplicative manner, completely determined by the marginal distributions.

The Poisson model assumption implies that marginals are random. If, instead, the total is fixed by the sampling design, it may be more appropriate to assume a multinomial distribution for the table. Formally, the multinomial model can be inferred from the Poisson model by conditioning on the total y_{++} (*see* **Conditional Probability**). For the probability π_{ij} of an observation falling into row i and column j, we then have $\pi_{ij} = \mu_{ij}/\mu_{++}$, and from assuming the multiplicative model (5), it follows that

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \quad (6)$$

where π_{i+} and π_{+j} are the marginal probabilities of an observation falling into row i and column j, respectively. Hence, row and column variables A and B are independently distributed.

Likewise, if row totals are fixed, then each row may be assumed to be multinomially distributed. Again, this can be inferred from the Poisson model by conditioning on the row totals, and the multiplicative model (5) implies identical distributions for the rows – a condition usually called *homogeneity*. It may be worthwhile noting that maximum-likelihood estimates for the parameters in the Poisson models are identical to those obtained for some other sampling designs, such as the multinomial designs just mentioned, making this class of model particularly interesting and useful.

Loglinear models for two- and higher-dimensional contingency tables, used to describe the association and interaction structure connecting the variables, are

Table 2 Number of recurrences of superficial bladder cancer for 31 male patients with grade 2, stage T_1, solitary primary tumors and respective times under observation (in months) by size of primary tumor. Subset of data analyzed in [38]

Size	Recurrences	Time under observation
$\leq$3 cm	1	2, 3, 6, 8, 9, 10, 11, 13, 14, 16, 21, 22, 24, 26, 27
	2	7, 13, 15, 18, 23
	3	20
	4	24
>3 cm	1	1, 5, 17, 18, 25
	2	18, 25
	3	4
	4	19

discussed in the article on **Loglinear Model**. Usually, the goal is to find a **parsimonious** model that fits the data well and allows meaningful substantive interpretation. Most commonly, this search is restricted to **hierarchical models**.

Multiplicative Rate Models

If occurrences of some kind of event are counted over time, then often interest lies in the rate at which events occur. The rate describes the instantaneous risk for an event to happen at a given point in time. To be more specific, the probability of observing exactly one event in the interval ranging from t to $t+h$, divided by its length h, is assumed to tend to some value $\lambda(t)$, as h tends to 0. $\lambda(t)$, as a function of time t, is called the *rate* or *intensity function*.

An important special case, termed the **Poisson process**, assumes that waiting times between successive events are independent and **exponentially distributed** with common mean $1/\lambda$. Here, the rate function is constant over time, $\lambda(t) \equiv \lambda$. Furthermore, the number $Y(t)$ of events that occur up to time t is distributed as Poisson with mean $\mu = \lambda t$. Note that the mean of $Y(t)/t$ equals the rate λ. This suggests a Poisson regression approach

$$\log \lambda = \log\left(\frac{\mu}{t}\right) = \alpha + \beta x$$

for modeling the dependence of the rate function on an explanatory variable X. This can be rewritten as

$$\log \mu = \alpha + \beta x + \log t,$$

with $\log(t)$ as an *offset*, i.e. a variable in the linear predictor, the corresponding regression parameter of which is set equal to 1. Observe that this defines a multiplicative model for the rate function,

$$\lambda = \mathrm{e}^{\alpha}(\mathrm{e}^{\beta})^{x}, \tag{7}$$

with a *baseline rate* $\lambda_0 = \exp(\alpha)$ and proportionality factor $\exp(\beta x)$.

For illustrative purposes a subset of the data analyzed in [38] is reprinted in Table 2. For 31 male patients, who have been treated for superficial bladder cancer, the number of recurrent tumors has been recorded for some time after removal of the primary tumor. Defining X to be 1 for larger primary tumors (>3 cm) and 0 otherwise, and assuming a Poisson process with rate (7), yields parameter estimates $\hat{\alpha} = -1.95$ and $\hat{\beta} = 0.385$. The (baseline) rate for smaller tumors is (estimated as) 0.142, the rate for larger tumors being 1.47 times larger. In terms of waiting times between recurrences, means are estimated as 7.06 and 4.80 months, respectively.

Now suppose that we have recorded, for n individuals, time under observation, t_i, and the number y_i of events occurred. Observation times are assumed to be nonrandom and counts to be mutually independent. We also have a set of explanatory variables $x_{i1}, \ldots, x_{ik}$ available for each subject. Under the assumption of *proportional rates*, $\lambda_i = \lambda_0 \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)$, we have

$$\mu_i = \lambda_i \times t_i = \lambda_0 \exp\left[\log(t_i) + \sum_{j=1}^{k} x_{ij}\beta_j\right]$$
$$= \lambda_0 t_i \prod_{j=1}^{k} \exp(x_{ij}\beta_j), \tag{8}$$

i.e. a loglinear model for the mean of the Poisson process, involving the logarithm of observation times as an "explanatory" variable, with the associated regression parameter fixed at a value of 1.

If the process is such that it can be characterized by a time-varying rate function $\lambda(t)$, it is called a *nonhomogeneous Poisson process*. Writing

$$\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u$$

for the *integrated rate* or *intensity function*, the number of occurrences of the event in period until time point t is again distributed as Poisson, but with mean equal to $\Lambda(t)$. Note that events in nonoverlapping time intervals are independent, but waiting times between successive events are, contrary to the homogeneous process with constant rate, neither identically distributed nor independent. In this situation model (8) can be modified, using a *baseline rate* function $\lambda_0(t|\alpha)$, possibly depending on some additional parameter α, to give

$$\mu_i = \exp\left\{\log[\Lambda_0(t_i|\alpha)] + \sum_{j=1}^{k} x_{ij}\beta_j\right\}.$$

Choosing $\Lambda_0(t|\alpha)$ to be t, t^α, or $\exp(\alpha t)$ corresponds to an exponential, a **Weibull**, or an **extreme value** intensity function, respectively, and results in a loglinear model for the Y_is. Disregarding constant terms, the likelihood function for this model is

$$L(\alpha, \beta) = \prod_{i=1}^{n}\left[\Lambda_0(t_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right]^{y_i} \times \exp\left[-\Lambda_0(t_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right]. \quad (9)$$

If times for occurrences of each event were known, a multiplicative term, depending on the parameter α, would be added to (9); see [32].

There is a close connection to **relative risk models**, which are very frequently used in epidemiology. This class of models assumes that risk factors interact in a multiplicative way. See [9] and [12], and, for a critical review, [26].

Proportional Hazard Models for Censored Survival Times

Now suppose that individuals are under observation until either a single event of interest occurs or the period of observations ends for some other reason. For each subject the data are of the form (y_i, c_i), where y_i is the time under observation, and c_i is an indicator variable for **censoring**, taking the value 1 if the event has occurred at time y_i, and the value 0 if the event has not occurred until time y_i. This is a similar situation to the one in the previous example, but with one terminal event that stops the process; interest, however, lies in the analysis of the *survival times* y_i (*see* **Survival Analysis, Overview**).

The distribution of the survival time can be uniquely described by the rate function, in the context of survival analysis usually called **hazard rate** or *force of mortality*. As before, a common approach assumes **proportional hazard** rates,

$$\lambda(y_i|\alpha, \beta) = \lambda_0(y_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right),$$

with a *baseline hazard* $\lambda_0(y_i|\alpha)$.

Assuming a noninformative censoring mechanism (and continuous survival times), the kernel of the likelihood function is $\prod_{i=1}^{n} f(y_i)^{c_i} \times S(y_i)^{1-c_i}$, where $f(y_i)$ denotes the density for the ith survival time, and $S(y_i) = 1 - F(y_i)$, the *survival function*, i.e. the probability for the ith survival time to exceed y_i. The ratio $f(y_i)/S(y_i)$ is identical to the hazard function. For proportional hazard rates, the likelihood function can therefore be expressed as

$$L_{y,c}(\alpha, \beta) = \prod_{i=1}^{n}\left[\Lambda_0(y_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right]^{c_i} \times \exp\left[-\Lambda_0(y_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right],$$

where $\Lambda_0(y_i|\alpha) = \int_0^{y_i} \lambda_0(u|\alpha)\,\mathrm{d}u$ denotes the cumulative baseline hazard rate. Writing, as we did before, $\mu_i = \exp\left\{\log[\Lambda_0(y_i|\alpha)] + \sum_{j=1}^{k} x_{ij}\beta_j\right\}$, $L(\alpha, \beta)$ is the likelihood function for n independent Poisson variates C_i with means μ_i. Aitkin & Clayton [2] used

this fact to bring survival analysis into the framework of generalized linear models (see also [7, 28], and [31]).

If no assumptions on the functional form of the baseline hazard function are made, then this is Cox's proportional hazards model [13, 14] that can be fitted by maximizing a "**partial likelihood**" (*see* **Cox Regression Model**). Another **semiparametric model**, due to Breslow [7], assumes a piecewise exponential distribution for the survival times, the baseline hazard function in this case is constant over prespecified intervals of time (*see* **Grouped Survival Times**). To be more specific, suppose that the time axis is split into intervals $(a_{p-1}, a_p]$, $p = 1, \ldots, P$, with $0 = a_0 < a_1 < \cdots < a_p < a_{p+1} = \infty$. The baseline hazard can now be written as

$$\lambda_0(y|\alpha) = \exp(\alpha_p), \quad \text{if } a_{p-1} < y \leq a_p.$$

To simplify notation, for individual i and interval $(a_{p-1}, a_p]$ the proportional hazard assumption can be expressed in terms of a constant λ_{ip}, where

$$\lambda_{ip} = \exp\left(\alpha_p + \sum_{j=1}^{k} x_{ij}\beta_j\right). \tag{10}$$

Define P_i to be such that y_i is contained in interval $(a_{P_{i-1}}, a_{P_i}]$ and e_{ip} to be the exposure time of individual i in the pth interval, i.e.

$$e_{ip} = \begin{cases} a_p - a_{p-1}, & \text{if } p = 1, \ldots, P_i - 1, \\ y_i - a_{P_{i-1}}, & \text{if } p = P_i \end{cases}.$$

Also, introduce an extended censoring indicator variable to be

$$c_{ip} = \begin{cases} 1, & \text{if } p = 1, \ldots, P_i - 1, \\ c_i, & \text{if } p = P_i. \end{cases}$$

Disregarding constant terms, the likelihood function is then

$$L_{y,c}(\alpha, \beta) = \prod_{i=1}^{n} \prod_{p=1}^{P_i} (\lambda_{ip} e_{ip})^{c_{ip}} \exp(-\lambda_{ip} e_{ip}),$$

where λ_{ip} is defined by (10). Since this is a Poisson likelihood for the "counts" c_{ip}, the *piecewise exponential model* reduces to a loglinear model. If intervals are chosen such that their endpoints correspond to observed times of death, i.e. t_is with $c_i = 1$, then maximum likelihood estimates for the regression parameters β are found to be close to those obtained from the Cox model; see [3] and [39].

For an example consider the data printed in Table 3. For 33 patients treated for papillary thyroid carcinoma, survival time, censoring indicator, age, and gender are reported. This is a small subset of the data analyzed in [30]. For cutpoints $a_1 = 0.5$, $a_2 = 1$,

Table 3 Survival times: *time* in years, censoring indicator *cens* (= 0 for censored), *gender* (1 for male), and *age* for 33 patients treated for papillary thyroid carcinoma. Subset of data analyzed in [30]

Time	Cens	Gender	Age	Time	Cens	Gender	Age
27.42	0	1	21	2.33	0	1	76
8.50	1	2	31	1.33	0	1	46
0.13	1	1	62	0.08	1	2	84
0.83	1	1	53	2.83	0	2	69
5.92	0	2	52	2.25	1	2	90
1.92	0	2	67	0.25	1	2	52
0.92	1	1	73	3.42	0	2	71
11.67	0	2	56	1.92	1	2	75
0.17	1	1	57	3.00	1	1	69
5.00	1	2	71	1.00	1	1	75
0.08	1	1	53	8.50	1	2	73
0.08	1	2	53	4.17	0	2	36
0.92	1	2	48	3.50	1	1	38
5.08	1	2	65	1.25	1	2	69
5.42	1	2	49	0.33	1	2	77
0.25	0	1	61	0.67	1	1	87
0.17	1	1	71				

$a_3 = 2$, and $a_4 = 3$, the piecewise-constant baseline hazard function is, up to a constant, estimated as

$$\lambda_0(y) = \begin{cases} 0.45, & \text{if } y \leq 0.5, \\ 0.36, & \text{if } 0.5 < y \leq 1, \\ 0.11, & \text{if } 1 < y \leq 2, \\ 0.16, & \text{if } 2 < y \leq 3, \\ 0.20, & \text{if } 3 < y. \end{cases}$$

Regression parameters, estimated for gender and age, are -0.70 and 0.04, respectively.

Log-nonlinear Models

While loglinear models do have some desirable properties, it may not always be possible to find a parameterization such that the regression function is linear on the log scale. An example of this is given in [20], using a log-logistic regression function. The data come from a **radioimmunoassay**, a widely used technique to measure the quantity of a given biological substance by identifying the amount of a radioactive labeled antibody from a reagent by subsamples of increasing concentration. The response variable is the amount of radioactive material remaining measured in counts per minute. If these are very large, a normal distribution for the counts may be assumed, but if this is not the case, an underlying Poisson distribution seems to be more appropriate. For counts $y_1, \ldots, y_n$ and concentrations $x_1, \ldots, x_n$ a regression function of the form

$$\eta_i(\beta_1, \ldots, \beta_4) = \beta_1 + \frac{\beta_2}{1 + \exp\{-[\beta_3 + \beta_4 \log(x_i)]\}} \quad (11)$$

can be used to describe the relationship between mean counts and concentrations. Note that this model cannot be transformed into a loglinear one.

Other examples of log-nonlinear models arise frequently in the analysis of contingency tables, when specific structure in the data suggests inclusion of nonlinear **interaction** effects into the regression function. See, for instance, [1, pp. 287–293].

Likelihood Inference

When adopting a modeling approach it seems to be natural to estimate the parameters of a model by maximizing the likelihood function, or, equivalently, its logarithm. The likelihood function contains all the relevant information about the mechanism that generated the data as well as the data actually observed. The larger its value the stronger the support given, by the data, to the corresponding value of the parameters. When dealing with Poisson regression models, maximum-likelihood estimation is, by far, the most often used method to obtain estimates for the unknown parameters.

Poisson regression models as defined above are instances of curved exponential family models; even if the model is loglinear, it is an exponential family model. So a much more general theory applies to this class of models. Here, only the special case will be considered. Readers interested in a general and detailed treatment are referred to, for example, Barndorff-Nielsen & Cox [5].

To maximize the log likelihood function one usually calculates partial derivatives with respect to all the parameters, sets them equal to 0, and solves this system of equations for the unknowns. For a Poisson regression model with log likelihood (3), the *estimating equations*

$$u_j(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \eta_i(\beta) \frac{1}{\eta_i(\beta)} [y_i - \eta_i(\beta)] = 0$$

need to be solved. If the model is loglinear, then this simplifies to

$$u_j(\beta) = \sum_{i=1}^{n} x_{ij} \left[y_i - \exp\left(\sum_{h=1}^{k} x_{ih} \beta_h \right) \right] = 0.$$

A generally applicable method for obtaining estimates numerically is provided by the *Fisher scoring algorithm* (*see* **Optimization and Nonlinear Equations**). In the present case, this is seen to be an *iteratively reweighted least squares procedure*, where, in each step of the iterative algorithm, a weighted **least squares** problem is to be solved. As a particular consequence to this fact, methods developed for diagnosing linear regression models can be modified for generalized linear models. To define *leverage* and *influence* one only needs to refer to respective quantities calculated from the last iteration step (*see* **Diagnostics**). Formulas needed to do so are lengthy to write down, but most of the widely used software packages provide, at least as an option, the figures. For more on diagnostics for generalized linear models

see [15]. Software packages found useful for fitting Poisson regression models include GLIM [20] and **S-PLUS** [10].

Not much is known about existence and uniqueness of maximum likelihood estimators in the general case. For loglinear models, however, if all observed sufficient statistics involved are larger than 0, then maximum likelihood estimates for the means, i.e. $\hat{\mu}_i = \eta_i(\hat{\beta})$, do exist and are unique, which is also true for $\hat{\beta}$, if the design matrix is of full rank. For a more detailed discussion, see [1] and the references therein.

A statistic capable of measuring the amount of support given by the data to a particular value of the parameter compared to its maximum likelihood estimate is the *deviance*, defined as minus two times the logarithm of the *normed likelihood*:

$$\begin{aligned} D_y(\beta) &= -2\log\left(\frac{L_y(\beta)}{L_y(\hat{\beta})}\right) \\ &= -2[\ell_y(\beta) - \ell_y(\hat{\beta})] \\ &= -2\sum_{i=1}^{n}\left\{y_i \log\left[\frac{\eta_i(\beta)}{\eta_i(\hat{\beta})}\right] \right. \\ &\qquad \left. - [\eta_i(\beta) - \eta_i(\hat{\beta})]\right\}. \end{aligned}$$

The deviance cannot be negative. It provides a measure of distance between the model described by β and the model characterized by the most likely parameter $\hat{\beta}$ and can, thus, be used to construct likelihood regions. Assuming β to be the "true" parameter, the deviance has an asymptotic χ^2 distribution with k degrees of freedom, where k is the dimension of the parameter β. This admits an interpretation of likelihood regions as confidence sets.

To obtain a measure of **goodness of fit** similar to the residual sum of squares in normal linear regression, the likelihood for the *maximal model* that perfectly fits the data can be compared to the likelihood of the model under consideration. This statistic is usually written as

$$\mathrm{dev}_y = 2\sum_{i=1}^{n}\left\{y_i \log\left[\frac{y_i}{\eta_i(\hat{\beta})}\right] - [y_i - \eta_i(\hat{\beta})]\right\}, \tag{12}$$

and termed *deviance* as well. Assuming the null model to be correct, the expected value for the latter statistic is approximately equal to the number of residual **degrees of freedom**, i.e. the number of observations minus the number of parameters in the model.

The deviance is a very important tool in searching for a "good", i.e. a parsimonious and well fitting, model, as it can be used to compare nested **hierarchical** models. Suppose we have a model with parameter β and a smaller one with a parameter γ, which can be obtained from β by setting r components to 0. Then, assuming the smaller model to be the correct one, the difference of deviances (12) is asymptotically χ^2 distributed with r degrees of freedom. Note that this is a **likelihood ratio test** for the smaller model with the null hypothesis against the larger model as the alternative.

The deviance is a useful measure of discrepancy, frequently supposed to have an approximate χ^2 distribution. However, this is to be taken with care, as χ^2 is not, in general, guaranteed to be a large sample distribution of (12). The deviance itself can be approximated by

$$X^2 = \sum_{i=1}^{n}\frac{[y_i - \eta_i(\hat{\beta})]^2}{\eta_i(\hat{\beta})}, \tag{13}$$

which is known as the *Pearson goodness-of-fit statistic* (*see* **Chi-square Tests**).

Another way of performing significance tests of hypotheses about single parameters is by applying a *Wald test* (*see* **Likelihood**). This uses the approximate normality of the maximum likelihood estimates and computes, as the test statistic, the ratio of the estimate of the parameter of interest and its asymptotic standard error. The formula is complex, but, again, many statistical packages provide the figures for the Wald test, sometimes under the heading *t-test*, as well as observed significance values. For more detailed accounts on likelihood inference for a generalized linear model with some emphasis on the Poisson regression model see [1, 19, 34] and [35] and the references therein.

An obvious way of defining **residual** quantities is to use square roots of contributions to the sums in (12) or (13), and attach the appropriate signs. Denoting raw residuals by $r_i = y_i - \eta_i(\hat{\beta})$, we have

$$\begin{aligned} r_i^{\mathrm{D}} = \mathrm{sgn}(r_i)(-2\{y_i \log\left[\frac{y_i}{\eta_i(\hat{\beta})}\right] \\ - [y_i - \eta_i(\hat{\beta})]\})^{1/2} \end{aligned} \tag{14}$$

for the *deviance residuals* and

$$r_i^{\mathrm{P}} = \frac{y_i - \eta_i(\hat{\beta})}{[\eta_i(\hat{\beta})]^{1/2}} \tag{15}$$

for the *Pearson residuals*. In any case, large residuals indicate large contributions to the respective goodness-of-fit statistics. Both deviance and Pearson residuals can (and should) be standardized. This requires computation of leverages for all observations. See [37] and [15] for more on residuals in generalized linear models.

For the time trend model fitted to the Stress Recall Data one calculates a deviance of 24.57 with 16 degrees of freedom. The deviance is 1.5 times larger than its approximate expected value, indicating a moderate amount of *overdispersion* (see [8, 16], and [17] for more on the phenomenon of overdispersion in Poisson regression models). Compared to a model with only the constant term included, we see a difference of deviances of 26.67. Referring to its approximate **chi-square distribution** (with 1 degree of freedom) clearly confirms the time trend. The regression parameter for the explanatory variable "months before interview" has been estimated as -0.0837, with an asymptotic standard error of 0.017, resulting in a t-value of -4.99, which is definitely large enough to reject the hypothesis of no time trend. The smallest deviance residual is -1.99, the largest 2.04, and there is no obvious pattern suggesting any specific inadequacies in the model.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Aitkin, M. & Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM, *Applied Statistics* **29**, 156–163.

[3] Aitkin, M., Laird, N. Francis, B. (1983). A reanalysis of the Stanford heart transplant data (with comments and rejoinder), *Journal of the American Statistical Association* **78**, 264–292.

[4] Aitkin, M., Anderson, D., Francis, B. & Hinde. J. (1989). *Statistical Modeling in GLIM*. Clarendon Press, Oxford.

[5] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[6] Bishop, Y.M.M., Fienberg, S.E. & Holland, P. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

[7] Breslow, N. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–100.

[8] Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models, *Journal of the American Statistical Association* **85**, 565–571.

[9] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1. *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

[10] Chambers, J.M. & Hastie, T.J. (1992). *Statistical Models in S*. Wadsworth & Brooks, Pacific Gove.

[11] Charnes, A., Frome, E.L. & Yu, P.L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family, *Journal of the American Statistical Association* **71**, 169–172.

[12] Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

[13] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[14] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[15] Davison, A.C. & Snell, E.J. (1991). Residuals and diagnostics, in *Statistical Theory and Modelling. In Honour of Sir David Cox*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, pp. 83–106.

[16] Dean, C.B. (1992). Testing for overdispersion in Poisson and binomial regression models, *Journal of the American Statistical Association* **87**, 451–457.

[17] Dean, C. & Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association* **84**, 467–472.

[18] Dobson, A. (1990). *An Introduction to Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[19] Firth, D. (1991). Generalized linear models, in *Statistical Theory and Modelling. In Honour of Sir David Cox*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, pp. 55–82.

[20] Francis, B., Green, M. & Payne, C., eds (1993). *The GLIM System. Release 4 Manual*. Clarendon Press, Oxford.

[21] Frome, E.L. (1981). Poisson regression analysis, *American Statistician* **35**, 262–263.

[22] Frome, E.L. (1983). The analysis of rates using Poisson regression models, *Biometrics* **39**, 665–674.

[23] Frome, E.L. & DuFrain, R.J. (1986). Maximum likelihood estimation for cytogenic dose-response curves, *Biometrics* **42**, 73–84.

[24] Frome, E.L. & Morris, M.D. (1989). Evaluating goodness of fit of Poisson regression models in cohort studies, *American Statistician* **43**, 144–147.

[25] Frome, E.L., Kutner, M.H. & Beauchamp, J.J. (1973). Regression analysis of Poisson-distributed data, *Journal of the American Statistical Association* **68**, 935–940.

[26] Greenland, S. & Maldonado, G. (1994). The interpretation of multiplicative-model parameters as standardized parameters, *Statistics in Medicine* **13**, 989–999.

[27] Haberman, S. (1978). *Analysis of Qualitative Data*. Vol. 1. *Introductory Topics*. Academic Press, New York.

[28] Holford, T.R. (1980). The analysis of rates and survivorship using log-linear models, *Biometrics* **36**, 299–306.

[29] Koch, G.G., Atkinson, S.S. & Stokes, M.E. (1986). Poisson regression, in *Encyclopedia of Statistics*, Vol. 7, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 32–41.

[30] Ladurner, D. & Seeber, G. (1984). Das papilläre Schilddrüsenkarzinom-Prognose und prognostische Faktoren, *Langenbeck's Archiv für Chirurgie* **363**, 43–55.

[31] Laird, N. & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques, *Journal of the American Statistical Association* **76**, 231–240.

[32] Lawless, J.F. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association* **82**, 808–815.

[33] Lindsey, J.K. (1995). *Modelling Frequency and Count Data*. Clarendon Press, Oxford.

[34] Lindsey, J.K. (1996). *Parametric Statistical Inference*. Clarendon Press, Oxford.

[35] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[36] Nelder, J.A. & Wedderburn, R.W. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

[37] Pierce, D.A. & Schafer, D.W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association* **81**, 977–986.

[38] Seeber, G.U.H. (1989). On the regression analysis of tumour recurrence rates, *Statistics in Medicine* **8**, 1363–1369.

[39] Selmer, R. (1990). A comparison of Poisson regression models fitted to multiway summary tables and Cox's survival model using data from a blood pressure screening in the city of Bergen, Norway, *Statistics in Medicine* **9**, 1157–1165.

(*See also* **Categorical Data Analysis**)

GILG U.H. SEEBER

Poisson, Siméon–Denis

Born: June 21, 1781, in Pithiviers, Loiret, France.
Died: April 25, 1840, in Paris, France.

Poisson's formative years were spent in the provinces where he came of modest family. Encouraged by a dedicated teacher at the École Centrale at Fontainebleau, he was admitted to the École Polytechnique in Paris in 1798, where, through the backing of **Laplace**, he was appointed to the academic staff in 1800. He stayed at the École Polytechnique, replacing Fourier as professor in 1806, while gaining other posts, notably a professorship at the Faculty of Sciences at the Sorbonne in 1816. He had been elected to the Paris Academy of Sciences in 1812. Through his life, characterized by hard work and dedication to scientific research and the teaching of science, Poisson accommodated himself to the various changes of political regime. In his research he treated a very broad range of subjects in a great number of publications, of which those on **probability** and its applications are relatively few [4], the most important being the book, published close to the end of his life: *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile* (1837).

In the *Recherches*, the **Poisson distribution** is derived as the limit of the distribution function of the (Pascal) distribution with mass function:

$$\binom{m+t-1}{m-1} p^m q^t, \quad t = 0, 1, 2, \ldots,$$

as $m \to \infty$, $q \to 0$ in such a way that $qm \to \omega =$ const. > 0 [5].

Indeed, the Poisson mass function $e^{-\omega}\omega^t/t!$, $t = 0, 1, 2, \ldots$, occurs significantly earlier than in Poisson's work [1–3]. However, he considered the **"Cauchy" distribution** with density $f(x) = 1/\pi (1 + x^2)$, $-\infty < x < \infty$, some 20 years before Cauchy. The use of the notion of a **random variable**, of the cumulative distribution function, and the definition of the density as its derivative, may be original with Poisson [4].

Apart from his many scientific articles and memoirs, Poisson was heavily involved in administration and pedagogical activity. He had the habit of saying "Life is good for only two things: to study mathematics and to teach it."

References

[1] David, F.N. (1962). *Games, Gods and Gambling: The Origins and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era*. Griffin, London. (De Moivre on the Poisson approximation.)

[2] Kendall, M.G. (1968). Thomas Young on coincidences, *Biometrika* **55**, 249–250. (The Poisson distribution in the game of rencontre in 1819.)

[3] Seneta, E. (1983). Modern probabilistic concepts in the work of E. Abbe and A. De Moivre, *Mathematical Scientist* **8**, 75–80. (Includes a discussion of the origins of the Poisson approximation to the binomial.)

[4] Sheynin, O.B. (1977–1978). S.D. Poisson's work in probability, *Archive for History of Exact Sciences* **18**, 245–300. (The most extensive account of Poisson's work in probability and statistics.)

[5] Stigler, S.M. (1982). Poisson on the Poisson distribution, *Statistics and Probability Letters* **1**, 33–35.

E. SENETA

Polya Process

The Polya process is a nonstationary birth process (*see* **Stochastic Processes**). Let $X(t)$ be the number of individuals present at time t, with $X(0) = 0$ at $t = 0$. The transition (conditional) probabilities of $X(t)$,

$$P_{0k}(0, t) = \Pr[X(t) = k | X(0) = 0],$$

satisfy the following differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{00}(0, t) = -\lambda_0(t) P_{00}(0, t) \tag{1a}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{0k}(0, t) = -\lambda_k(t) P_{0k}(0, t) + \lambda_{k-1}(t) P_{0,k-1}(0, t), \tag{1b}$$

for $k = 1, 2, \ldots$, where

$$\lambda_k(t) = \frac{1 + \lambda k}{1 + \lambda t}, \quad \text{for } k = 0, 1, \ldots, \tag{2}$$

is a function of both the number k and time t. The probability **generating function** of $X(t)$, as derived in [2], is

$$G_X(s, t) = [1 + \lambda t - \lambda t s]^{-1/\lambda},$$

which leads to the formula for $P_{0k}(0, t)$:

$$P_{0k}(0, t) = \binom{1/\lambda + k - 1}{k} \left(\frac{1}{1 + \lambda t}\right)^{1/\lambda} \times \left(\frac{\lambda t}{1 + \lambda t}\right)^{k}, \tag{3}$$

for $k = 0, 1, \ldots$. Formula (3) is a **negative binomial distribution** with parameters $1/\lambda$ and $1/(1 + \lambda t)$. The expectation and the variance of $X(t)$ are, respectively,

$$\mathrm{E}[X(t)] = t \quad \text{and}$$
$$\mathrm{var}[X(t)] = t[1 + \lambda t].$$

The Polya process is closely related to the Polya distribution generated by the **Polya urn** scheme [3–5, 7]. Suppose that an urn contains $(a + b)$ balls, of which a balls are red and b balls are black. A ball is drawn at random from the urn, its color noted. The ball is replaced and c balls of the same color are added. After the first draw, there are $(a + b + c)$ balls in the urn. The procedure is repeated.

Among n draws, the number of times resulting in red balls, $X(n)$, has the following probability distribution:

$$P_{\mathrm{k,n}} = \binom{n}{k} \frac{\left\{\begin{array}{c} a(a + c)(a + 2c) \ldots [a + (k - 1)c] \\ \times b(b + c)(b + 2c) \ldots [b + (n - k - 1)c] \end{array}\right\}}{\left\{\begin{array}{c} (a + b)(a + b + c)(a + b + 2c) \\ \times \ldots [a + b + (n - 1)c] \end{array}\right\}}, \tag{4}$$

for $k = 0, 1, \ldots, n$. The expectation and the variance of $X(n)$ are, respectively,

$$\mathrm{E}[X(n)] = np \quad \text{and}$$
$$\mathrm{var}[X(t)] = \frac{npq[1 + nr]}{[1 + r]},$$

where $p = a/(a+b)$, $q = b/(a+b)$, and $r = c/(a+b)$. If $c = 0$, then (4) is a **binomial distribution**; if $c = -1$, then (4) is a **hypergeometric distribution**.

The constant c is a "contagion" factor of the distribution (*see* **Contagious Distributions**). If red balls stand for infection, then the occurrence of an infection (drawing a red ball) increases the amount of infection in the population, and increases the probability of having an infected case (a red ball) in the future. Drawing a black ball will have the opposite effect. The number c can be either positive or negative. If c is positive, then the contagion is positive; if c is negative, then the contagion is negative (such as immunity). The Polya process and Polya distribution have applications in studying contagious diseases, in the theory of cosmic radiation [1], and in the analysis of accident, insurance, and sickness statistics [6].

Polya [3, 7] has shown that, within a time interval $(0,t)$, if $n \to \infty$, $p \to 0$, and $r \to 0$, so that $np \to t$ and $nr \to \lambda t$, then formula (4) in the Polya distribution approaches formula (3) in the Polya process as a limit.

References

[1] Arley, N. (1948). *On the Theory of Stochastic Processes and Their Applications to the theory of Cosmic Radiation*. Wiley, New York.

[2] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Krieger, New York.

[3] Eggenberger, F. & Polya, G. (1923). Über die statistik verketteter Vorgänge, *Zeitschrift für Angewandte Mathematik and Mechanik* **3**, 279–289.

[4] Feller, W. (1950). *An Introduction to Probability Theory and Its Applications*, 3rd Ed. Wiley, New York.

[5] Fisz, M. (1953). *Probability Theory and Mathematical Statistics*, 3rd Ed. Wiley, New York.

[6] Lundberg, O. (1940). *On Random Processes and Their Applications in Sickness and Accident Statistics*. Uppsala, Sweden.

[7] Polya, G. (1930). Sur quelques points de la théorie des probabilités, *Annales de L'Institut Henri Poincare* **1**, 117–161.

(*See also* **Accident Proneness**).

CHIN LONG CHIANG

Polya's Urn Model

The Polya urn model was introduced by Eggenberger & Polya [2]. The urn initially contains balls of two colors (W white balls and R red balls). A ball is drawn at random. After each drawing, the chosen ball is returned together with s balls of the same color. The process is repeated n times. Let X be the number of times a red ball is drawn. The exact distribution of X can be found in Johnson & Kotz [3, p. 177] and can be approximated by the well-known Polya distribution [4, p. 64]. It is interesting to note that when $s = 0$, X is a simple **binomial** random variable, and when $s = -1$, X is a **hypergeometric** random variable.

This simple urn model can be generalized. For example, after each draw, t balls of color opposite to that chosen are also added. Furthermore, s and t may be negative, and s and t may be random variables. For these general cases, the properties of X have been studied through the theory of **branching processes** [1] (*see* **Polya Process**).

This model applies to many practical situations in medical studies. The case in which s and t are random variables has an interesting and important application to **sequential methods in clinical trials** (*see* **Adaptive and Dynamic Methods of Treatment Assignment**). Specifically, in comparing two treatments A and B, suppose that eligible patients occur singly and must be treated when they arrive. For each patient's treatment assignment, a ball is selected at random from the urn with replacement. If it is white, assign this patient to Group A; otherwise, assign this to Group B. When the response of a previous patient to treatment A is a success, we add s white balls and t red balls to the urn. However, if the response is a failure, we add s' whites and t' reds to the urn. In practice, $\{s, t, s', t'\}$ are chosen so that this type of randomized Polya urn scheme tends to put more patients on the better treatment, but also provides reliable data to evaluate the treatment difference due to its random nature. The urn scheme can also be modified to deal with the case in which more than two treatments are involved in the study [5], using an urn and balls with more than two colors.

References

[1] Athreya, K.B.& Karlin, S. (1968). Embedding of urn schemes into continuous time Markov branching process and related limit theorems, *Annals of Mathematical Statistics* **39**, 1801–1817.

[2] Eggenberger, F.& Polya, G. (1923). Über die statistik verketteter Vorgänge, *Zeitschrift für Angewandte Mathematik und Mechanik* **1**, 279–289.

[3] Johnson, N.L.& Kotz, S. (1977). *Urn Models and Their Application*. Wiley-Interscience, New York.

[4] Johnson, N.L. and Kotz, S. (1986). *Encyclopedia of Statistical Sciences*, Vol. 7. Wiley-Interscience, New York.

[5] Wei, L.J. (1979). The generalized Polya's urn design for sequential medical trials, *Annals of Statistics* **7**, 291–296.

(*See also* **Contagious Distributions**).

L.J. WEI

Polygenic Inheritance

Many quantitative, metric, or continuous traits do not show simple Mendelian transmission (*see* **Mendel's Laws**) and their expression is believed to be controlled by polygenes. Strictly speaking, polygenic inheritance refers to the mode of inheritance of characters or traits whose genetic component is determined by polygenes (i.e. many **genes**) with individually small effects, as opposed to monogenic (i.e. single gene) or oligogenic (i.e. a few genes) inheritance.

Plate [26] appears to be the first to use the term "polygenic". Fisher [8] and Mather [21] greatly enriched the meaning of "polygenic inheritance". The main properties of polygenic inheritance were summarized elegantly by Lerner [20], as follows: (i) most metric and meristic (polychotomous, or categorical) traits are affected by a number of genetic loci; (ii) the effects of allelic substitution at each of the segregating loci are usually relatively small and interchangeable, in the sense that identical phenotypes may be displayed by a great variety of **genotypes**; (iii) the phenotypic expression of polygenic traits is subject to considerable modification by environmental influences; (iv) most populations have great reserves of genetic variability, often carried in the gene pool in balanced systems; (v) there is nothing biochemically exceptional about genes controlling polygenic inheritance; (vi) polygenic traits show a continuous rather than a discontinuous distribution.

However, it should be noted that, for dichotomous or binary traits, which are either present or absent in any one individual, a combination of polygenes and environmental factors can determine an underlying or latent risk or liability toward the trait (*see* **Genetic Liability Model**). The values of the liability may or may not be directly observable, and individuals having the trait are those whose liability values exceed a threshold that is sometimes unknown.

Historical Background

Around the turn of this century with the rediscovery of Mendel's work, the field of human heredity was deeply divided and witnessed bitter confrontation between the biometricians, represented by **Karl Pearson** and W.F.R. Weldon, and the Mendelians, led by the Cambridge geneticist William Bateson. The biometricians held that the observed correlations between relatives in large populations were incompatible with Mendelian inheritance [25], and the Mendelians seem to have considered discontinuous genetic variation as incompatible with anything but obviously discontinuous phenotypic variation. A series of experimental and theoretical discoveries made during the 1910s and 1920s eventually led to the settlement of the arguments between the two schools.

Shull [29] demonstrated the rapid fixation of one or another random combination of quantitative traits in self-fertilized lines of maize and interpreted this as due to automatic fixation of Mendelian units. Johannsen [16] showed that heritable and nonheritable agencies were jointly responsible for the variation in seed weight, and that their effects were of the same order of magnitude. The effects of discontinuity of the genotype could be smoothed out and continuous variation realized in the phenotype by the action of the environment. Nilsson–Ehle [23] found that similar factors with smaller individual effects could account for continuous quantitative variation if enough of them were segregating. Each factor would be inherited in the Mendelian way, and its change would be discontinuous. Yet with a number of such factors, having similar and cumulative action, many different dosages would be possible. Continuity would be completed by the blurring effect of noninheritable agencies. These findings were also independently documented by East around the same period [4, 5]. Nilsson–Ehle and East both recognized that their findings were consistent with the cumulative effects of many genes, each with a small effect for a continuous phenotypic variable, and thus laid the one cornerstone of the "polygenic" theory of continuous variations. Sax [28] further established the Mendelian nature of quantitative traits by demonstrating linkage of minor quantitative differentials with marker genes. Yule [37] showed mathematically that Mendelian variation could be retained indefinitely in a randomly breeding population. It was on the basis of these works that **Ronald A. Fisher** was able to synthesize the idea of continuous phenotypic variation of the biometricians with the genetic discontinuity of the Mendelian mechanism of inheritance. He showed [8] how **Galton's** own data could be reconciled with Mendelian inheritance. There were three main aspects to Fisher's argument. First, following

the experimental results of East and Nilsson–Ehle, Fisher adopted the polygenic model in which a large number of genes (infinitely many in Fisher's analysis) of individually small effect were responsible for continuous variation and family resemblance in humans. Secondly, Fisher questioned the generalization drawn from Mendel's experiments that all hybrids for a given character would resemble one or the other parent (i.e. he relaxed the assumption of complete dominance). Thirdly, he allowed in various ways for the effects of **assortative mating**. His model predicts very well the pattern in the correlations between different kinds of relatives.

Polygenic Models

Polygenic models, also called **variance component models** or random effects models, are widely used in the field of animal and plant breeding (see, for example, [13, 14]). In human genetics, polygenic models can be used to identify, among other traits associated with the definition of disease, the subset that have a significant genetic component in their etiology [15, 30]. Moreover, a good, well-tested polygenic model can be utilized in **genetic counseling** [19]. Polygenic models also have the merit that the parameterization is straightforward: various cultural, environmental and complex genetic effects can be easily incorporated into the models. Another attraction of the models is their mathematical tractability when **multivariate normality** is assumed.

Under the polygenic model, the phenotype is the result of the joint action of infinitely many genes, each with a small contribution to the phenotype, and of the environment. In mathematical terms, the polygenic models assume that the phenotypic value y is the additive combination of genotypic value v and independent environmental effect e, namely

$$y = v + e.$$

In general, y can be an arbitrary function of v, but such a model would be too complex to be useful in reality.

The polygenic value v can be further decomposed into various components. For example, v can be decomposed into the mean value μ in the reference population, the additive component a, which is the combination of main effects of all loci, and the dominance effect d, which is the combination of the interactions between homologous genes at each loci (see [1]), i.e.

$$v = \mu + a + d,$$

where a and d are usually assumed to be normal. In practice, phenotypes often have to be transformed to approximate normality prior to analysis.

Let y_i and y_j be random variables defining the trait values of individuals i and j, respectively. Also, let σ_a^2 be the additive genetic variance, σ_d^2 the variance due to dominance, and σ_e^2 the environment variance. Denote ϕ_{ij} as the kinship coefficient for i and j (*see* **Inbreeding**). Then, under the formulation of Fisher, the covariance between the phenotypes for any two individuals, i and j, in a given pedigree, is

$$\mathrm{cov}(y_i, y_j) = 2\phi_{ij}\sigma_a^2 + \Delta_{ij}\sigma_d^2 + \delta_{ij}\sigma_e^2.$$

A multivariate normal distribution with this covariance structure is usually assumed. Under some fairly stringent conditions, Lange [17] proved a **central limit theory** for pedigrees.

In vector notation, and denoting $\mathbf{\Phi} = (\phi_{ij})$ and $\mathbf{\Delta} = (\Delta_{ij})$, the above model can be rewritten as

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{a} + \mathbf{d} + \mathbf{e}, \tag{1}$$

where $\mathbf{1}$ is an $n \times 1$ vector of 1s, $\mathbf{a} \sim \mathrm{N}(\mathbf{0}, 2\sigma_a^2\mathbf{\Phi})$, $\mathbf{d} \sim \mathrm{N}(\mathbf{0}, \sigma_d^2\mathbf{\Delta})$, and $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$.

Components $\mathbf{a}$, $\mathbf{d}$, and $\mathbf{e}$ are taken to be independent so that the variance of the phenotype $\mathbf{y}$ is simply the sum of the variances of the components.

A number of other assumptions are usually made to limit the complexity of the model. These include **Hardy–Weinberg equilibrium** and linkage equilibrium (*see* **Linkage Disequilibrium**) for all loci, no epistasis (*see* **Genotype**), and absence of **gene–environment interaction**, of **assortative mating**, and of correlated environments among relatives. Note that some of the assumptions can be relaxed without substantially complicating the mathematics. For example, the assumption of no correlated environments can be lifted by adding a common environment component $\mathbf{c}$ to the model with $\mathbf{c} \sim \mathrm{N}_n(\mathbf{0}, \sigma_d^2\mathbf{C})$, where the element c_{ij} of $\mathbf{C}$ is 1 if i and j share the same environment, and 0 otherwise.

Thus, a general polygenic model is

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{z}_1 + \mathbf{z}_2 + \cdots + \mathbf{z}_k + \mathbf{e}, \tag{2}$$

where the $\mathbf{z}_i$s are k random components, with $\mathbf{z}_i \sim \mathrm{N}_n(\mathbf{0}, \sigma_i^2\mathbf{\Sigma}_i)$; $\mathbf{e}$ is the environmental effect, with $\mathbf{e} \sim$

$N_n(\mathbf{0}, \sigma_e^2\mathbf{I})$. $\mathbf{z}_1, \ldots, \mathbf{z}_k$ and $\mathbf{e}$ are assumed mutually independent. The $\mathbf{\Sigma}_i$s are known and are at least semi-positive-definite. Without loss of generality, we can assume the $\mathbf{\Sigma}_i$s are invertible since we can always reparameterize the model so that, say, $\tilde{\mathbf{\Sigma}}_i = \mathbf{\Sigma}_i + (\varepsilon/\sigma_i^2)\mathbf{I}$, which is invertible, and $\sigma_e'^2 = \sigma_e^2 - \varepsilon$, where ε is small enough so that $\sigma_e^2 - \varepsilon > 0$. Note that here we have assumed a constant mean μ for all individuals. This is for convenience only. In general, the model can include some **fixed effects** such as sex and generation effects, or some **covariates** like age. Incorporation of fixed effects and covariates is straightforward.

Parameter Estimation

Maximum likelihood is the method of choice in parameter estimation for polygenic models because of its theoretical soundness. There are two approaches to maximum likelihood estimation: Fisher's method of scoring and the **EM algorithm**. Most recent theoretical developments in this area are aimed at reducing the computational burdens of likelihood estimation for variance components. In animal breeding genetics, the maximum likelihood method and its related method, **restricted maximum likelihood**, has been applied to estimation in polygenic models, often taking advantage of relatively simple structure in animal breeding data [9, 13, 14, 22]. In human genetics (*see* **Human Genetics, Overview**), the method of scoring is often the method of choice when pedigrees are not too large or complex. Otherwise, the EM method is often used.

The Method of Scoring

Lange et al. [19] provide an excellent exposition on the use of scoring method in parameter estimation for polygenic models. Consider the model in (1). Let $\mathbf{\Omega} = 2\sigma_a^2\mathbf{\Phi} + \sigma_d^2\mathbf{\Delta} + \sigma_e^2\mathbf{I}$, and $\boldsymbol{\gamma} = (\mu, \sigma_a^2, \sigma_d^2, \sigma_e^2)'$. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be the observed trait values. Individuals with missing values are deleted prior to analysis.

The log likelihood for the pedigree is then

$$L = -\tfrac{1}{2}\log|\mathbf{\Omega}| - \tfrac{1}{2}(\mathbf{y} - \mu\mathbf{1})'\mathbf{\Omega}^{-1}(\mathbf{y} - \mu\mathbf{1}),$$

since

$$\frac{\partial \mathbf{\Omega}^{-1}}{\partial\theta} = -\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\theta}\mathbf{\Omega}^{-1},$$

$$\frac{\partial L}{\partial\mu} = \mathbf{1}\mathbf{\Omega}^{-1}(\mathbf{y} - \mu\mathbf{1}),$$

and

$$\frac{\partial L}{\partial\sigma_k^2} = -\frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_k^2}\right) + \frac{1}{2}\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_k^2}\mathbf{\Omega}^{-1}(\mathbf{y} - \mu\mathbf{1}),$$

where

$$\frac{\partial\mathbf{\Omega}}{\partial\sigma_a^2} = 2\mathbf{\Phi}, \quad \frac{\partial\mathbf{\Omega}}{\partial\sigma_d^2} = \mathbf{\Delta}, \quad \text{and } \frac{\partial\mathit{\Omega}}{\partial\sigma_e^2} = \mathbf{I}.$$

The second partial derivatives of L are [19, 28]:

$$\frac{\partial^2 L}{\partial\mu^2} = -\mathbf{1}\mathbf{\Omega}^{-1}\mathbf{1}',$$

$$\frac{\partial^2 L}{\partial\mu\partial\sigma_k^2} = \frac{\partial^2 L}{\partial\sigma_k^2\partial\mu} = -\mathbf{1}\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_k^2}\mathbf{\Omega}^{-1}(\mathbf{y} - \mu\mathbf{1}),$$

$$\frac{\partial^2 L}{\partial\sigma_k^2\partial\sigma_l^2} = \frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_k^2}\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_l^2}\right) - \frac{1}{2}(\mathbf{y} - \mu\mathbf{1})'\mathbf{\Omega}^{-1} \times \left(\frac{\partial\mathbf{\Omega}}{\partial\sigma_k^2}\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_l^2} + \frac{\partial\mathbf{\Omega}}{\partial\sigma_l^2}\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_k^2}\right) \times \mathbf{\Omega}^{-1}(\mathbf{y} - \mu\mathbf{1}).$$

Since $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$, $\mathrm{E}[\mathbf{A}(\mathbf{Y} - \mu\mathbf{1})] = \mathbf{A}\mathrm{E}(\mathbf{Y} - \mu\mathbf{1}) = 0$, $\mathrm{E}[(\mathbf{Y} - \mu\mathbf{1})'\mathbf{A}(\mathbf{Y} - \mu\mathbf{1})] = \mathrm{tr}(\mathbf{A\Omega})$, where $\mathbf{A}$ and $\mathbf{B}$ are constant matrices and $\mathbf{Y}$ is a random vector with mean $\mu\mathbf{1}$ and variance $\mathbf{\Omega}$, the **information matrix** $\mathbf{I}(\gamma)$ can be obtained by taking expectations of the above second partial derivatives as

$$\mathbf{I}(\gamma) = \begin{pmatrix} \mathbf{1}\mathbf{\Omega}^{-1}\mathbf{1}' & 0 & 0 & 0 \\ 0 & \rho_{11} & \rho_{12} & \rho_{13} \\ 0 & \rho_{21} & \rho_{22} & \rho_{23} \\ 0 & \rho_{31} & \rho_{32} & \rho_{33} \end{pmatrix},$$

where

$$\rho_{kl} = \rho_{lk} = \frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_k^2}\mathbf{\Omega}^{-1}\frac{\partial\mathbf{\Omega}}{\partial\sigma_l^2}\right).$$

If there are m pedigrees, then the overall score vector and information matrix are the summation of score vectors and information matrices over all m pedigrees.

Now let $\gamma^{(i)}$ be the value of the parameters at the ith iteration, and let the score vector $\mathbf{S}[\gamma^{(i)}]$ be

the vector of partial derivatives of L with respect to individual parameters evaluated at $\gamma^{(i)}$. Then, the scoring algorithm takes the new value to be

$$\gamma^{(i+1)} = \gamma^{(i)} + \mathbf{I}^{-1}[\gamma^{(i)}]\mathbf{S}[\gamma^{(i)}].$$

If the iteration converges to a unique value $\hat{\gamma}$, then standard large sample theory shows that $\hat{\gamma}$ is asymptotically multivariate normal with mean γ_0 and covariance matrix $\mathbf{I}^{-1}(\hat{\gamma}_0)$, where γ_0 is the true parameter value.

The EM Algorithm

The EM equations for estimating γ are:

$$\mu^* = \frac{1}{n}\mathrm{E}_\gamma[\mathbf{1}'(\mathbf{y} - \mathbf{a} - \mathbf{d} - \mathbf{e})|\mathbf{y}],$$

$$(\sigma_a^*)^2 = \frac{1}{n}\mathrm{E}_\gamma[\mathbf{a}'(2\mathbf{\Phi})^{-1}\mathbf{a}|\mathbf{y}],$$

$$(\sigma_d^*)^2 = \frac{1}{n}\mathrm{E}_\gamma(\mathbf{d}'\mathbf{\Delta}^{-1}\mathbf{d}|\mathbf{y}),$$

and

$$(\sigma_e^*)^2 = \frac{1}{n}\mathrm{E}_\gamma(\mathbf{e}'\mathbf{e}|y),$$

where

$$\mathrm{E}_\gamma[\mathbf{a}'(2\mathbf{\Phi})^{-1}\mathbf{a}|\mathbf{y}] = \mathbf{p}'(2\mathbf{\Phi})^{-1}\mathbf{p} + \mathrm{tr}\left[(2\mathbf{\Phi})^{-1}\left(\frac{\mathbf{I}}{\boldsymbol{\sigma}_\mathbf{e}^2} + \frac{(2\mathbf{\Phi})^{-1}}{\sigma_a^2}\right)^{-1}\right],$$

$$\mathrm{E}_\gamma(\mathbf{d}'\mathbf{\Delta}^{-1}\mathbf{d}|\mathbf{y}) = \mathbf{q}'\mathbf{\Delta}^{-1}\mathbf{q} + \mathrm{tr}\left[\mathbf{\Delta}^{-1}\left(\frac{\mathbf{I}}{\boldsymbol{\sigma}_\mathbf{e}^2} + \frac{\mathbf{\Delta}^{-1}}{\sigma_d^2}\right)^{-1}\right],$$

and

$$\mathrm{E}_\gamma(\mathbf{e}'\mathbf{e}|\mathbf{y}) = \mathbf{h}'\mathbf{h} + \mathrm{tr}\left[\left(\frac{4\mathbf{I}}{\sigma_e^2} + \frac{(2\mathbf{\Phi})^{-1}}{\sigma_a^2} + \frac{\mathbf{\Delta}^{-1}}{\sigma_d^2}\right)^{-1}\right],$$

where $\mathbf{p} = \mathrm{E}_\gamma(\mathbf{a}|\mathbf{y})$, $\mathbf{q} = \mathrm{E}_\gamma(\mathbf{d}|\mathbf{y})$, and $\mathbf{h} = \mathrm{E}_\gamma(\mathbf{e}|\mathbf{y})$.

Other Methods

Both the scoring and the EM methods require repeated inversion of certain matrices of an order equivalent to the pedigree size. For small or moderate-size pedigrees, this is not a problem. However, for large pedigrees, the inversion of large matrices may not be practically feasible. Thompson & Shaw [33] proposed a method for parameter estimation using the EM algorithm that avoids repeated inversion of large matrices. Their method can be extended to multivariate traits [34]. Guo & Thompson [10, 11] proposed a Monte Carlo EM algorithm (*see* **Monte Carlo Methods**) to solve this problem.

Likelihood Computation

An EM algorithm that provides maximum likelihood estimates may be of limited use if it does not also provide a computable form of the likelihood because likelihood values are needed for statistical inferences and in monitoring the convergence process. Although direct evaluation of the likelihood for polygenic models is theoretically possible, as done by Lange et al. [19], it is practically infeasible for large pedigrees because of its demand for repeated inversion of large matrices.

Elston & Stewart [6] proposed an algorithm, later known as the **Elston–Stewart algorithm** or the "peeling algorithm", to compute the **likelihood** on pedigrees for simple genetic models. They showed that the algorithm can compute likelihoods not only for major gene models but also for simple polygenic models (i.e. with additive component only). The algorithm was later generalized to more complex pedigrees and more complex genetic models [2, 18, 24]. Thompson & Shaw [33] proposed a peeling algorithm for computing the likelihood function on large complex pedigrees for a simple polygenic model. The algorithm has been extended to complex polygenic models with multiple random effects [32] and to multiple traits [34].

Approximations to Polygenic Models

The polygenic model assumes an infinite number of loci, each contributing a small amount to the variation of the trait of interest. Another class of genetic model, called the mixed model, assumes that the expression of the trait is determined by a major gene plus polygenic background (*see* **Segregation Analysis, Mixed Models**). This model has proved to be a useful alternative to classical Mendelian models in the analysis of pedigree data. The exact calculation of the likelihood for the mixed model is virtually

impossible except for very small pedigrees [24]. To circumvent the problem, Hasstedt [12] proposed an approximation to the likelihood on large pedigrees for the mixed model. Guo & Thompson proposed a method based on a Monte Carlo EM algorithm [10].

Fernando et al. [7] and Stricker et al. [31] took a different approach, postulating a finite polygenic model that was suggested earlier by Cannings et al. [2]. The model approximates polygenic inheritance by postulating that trait values are determined by a small number of biallelic loci having equal and additive effects. This approach is computationally fast, and is attractive as an alternative to the traditional mixed model in linkage analysis.

Concluding Remarks

Polygenic inheritance is an important, and perhaps common, mode of inheritance in genetics. Sewall Wright [35, 36] was an avid advocate of the notion that each phenotype is affected by multiple genes, and that the same genes can affect more than one phenotype (pleiotropic effects). At least in plant genetics, the increasing numbers of Mendelian marker loci have helped confirm Robertson's [27] prediction that "the distribution of gene effects will probably be of an exponential kind (so that the smaller the range of the effect specified, the greater the number of loci concerned)"; see [3].

As an approximation to polygenic inheritance, polygenic models have been useful in establishing a genetic basis for many traits. Under certain assumptions, they flexibly include various random and fixed effects and are easy to parameterize. However, this flexibility in no way means that it would be easy to disentangle the joint action of a large number of loci, and their mathematical tractability is not a licence to apply the models indiscriminately to any data at hand. Many assumptions underlying the polygenic model are critical to ease the mathematical tractability, and are often difficult to verify in practice. For example, the assumption of no gene–environment interaction is difficult to verify in reality, especially so when environmental factors affecting the trait of interest have not been completely identified. Indeed, the mathematical tractability always comes with a price: restricted utility and the burden of verifying the underlying assumptions. Nonetheless, polygenic models, if used judiciously, can provide better insight into the genetic mechanism underlying traits of interest.

References

[1] Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, New York.

[2] Cannings, C., Thompson, E.A. & Skolnick, M.H. (1978). Probability functions on complex pedigrees, *Advances in Applied Probability* **10**, 26–61.

[3] Cox, T.S. (1995). Simultaneous selection for major and minor resistance genes, *Crop Science* **35**, 1337–1346.

[4] East, E.M. (1910). A Mendelian interpretation of variation that is approximately continuous, *American Naturalist* **44**, 65–82.

[5] East, E.M. (1916). Studies on size inheritance in *Nicotiana*, *Genetics* **1**, 164–176.

[6] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[7] Fernando, R.L., Stricker, C. & Elston, R.C. (1994). The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance, *Theoretical and Applied Genetics* **88**, 573–580.

[8] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[9] Gianola, D. (1986). On selection criteria and estimation of parameters when the variance is heterogeneous, *Theoretical and Applied Genetics* **72**, 671–677.

[10] Guo, S.W. & Thompson, E.A. (1991). Monte Carlo estimation of variance component models, *IMA Journal of Mathematical Applications in Medicine and Biology* **8**, 171–189.

[11] Guo, S.W. & Thompson, E.A. (1992). A Monte Carlo method for combined segregation and linkage analysis, *American Journal of Human Genetics* **51**, 1111–1126.

[12] Hasstedt, S.J. (1982). A mixed-model likelihood approximation on large pedigrees, *Computers and Biomedical Research* **15**, 295–307.

[13] Henderson, C.R. (1986). Recent developments in variance and covariance estimation, *Journal of Animal Science* **63**, 208–216.

[14] Henderson, C.R. (1986). Estimation of variances in animal model and reduced animal model for single traits and single records, *Journal of Dairy Science* **69**, 1394–1402.

[15] Hopper, J.L. & Mathews, J.D. (1982). Extensions to multivariate normal models for pedigree analysis, *Annals of Human Genetics* **46**, 373–383.

[16] Johannsen, W. (1909). *Elemente der Exakten Erblichkeitslehre*. Fischer, Jena.

[17] Lange, K. (1978). Central limit theorems for pedigrees, *Journal of Mathematical Biology* **6**, 59–66.

[18] Lange, K. & Elston, R.C. (1975). Extensions to pedigree analysis I. Likelihood calculations for simple and complex pedigrees, *Human Heredity* **25**, 95–105.

[19] Lange, K., Westlake, J. & Spence, M.A. (1976). Extensions to pedigree analysis. III. Variance components

by the scoring method, *Annals of Human Genetics* **39**, 485–491.

[20] Lerner, I.M. (1968). *Heredity, Evolution and Society*. Freeman, San Francisco, pp. 137–138.

[21] Mather, K. (1941). Variation and selection of polygenic characters, *Journal of Genetics* **41**, 159–193.

[22] Meyer, K. (1987). Restricted maximum likelihood to estimate variance components for mixed models with two random factors, *Genetique, Selection, Evolution* **19**, 49–68.

[23] Nilsson-Ehle, H. (1909). *Kreuzunguntersuchungen an Häfer und Weizen*, Vol. 5. Lunds University Aarskr. N.F., pp. 1–22.

[24] Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human genetics, *American Journal of Human Genetics* **31**, 161–175.

[25] Pearson, K. (1904). On the generalized theory of alternative inheritance with special reference to Mendel's law, *Philosophical Transactions of the Royal Society, Series A* **203**, 53–86.

[26] Plate, L. (1913). *Vererbungslehre; mit besonderer berucksichtigung der abstammungslehre und des menschen*. Engelmann, Leipzig.

[27] Robertson, A. (1967). The nature of quantitative genetic variation, in *Heritage from Mendel*, R.A. Brink, ed. University of Wisconsin Press, Madison, pp. 265–280.

[28] Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*, *Genetics* **8**, 552–560.

[29] Shull, G.H. (1908). The composition of a field of maize, *Report of the American Breeders Association* **4**, 298–301.

[30] Sing, C.F., Boerwinkle, E., Moll, P.P. & Templeton, A.R. (1988). Characterization of genes affecting quantitative traits in humans, in *Proceedings of the Second International Conference on Quantitative Genetics*, B.S. Weir, E.J. Eisen, M.M. Goodman & G. Namkoong, eds. Sinauer, Sunderland.

[31] Stricker, C., Fernando, R.L. & Elston, R.C. (1995). Linkage analysis with an alternative formulation for the mixed model of inheritance: the finite polygenic mixed model, *Genetics* **141**, 1651–1656.

[32] Thompson, E.A. & Guo, S.W. (1991). Evaluation of likelihood ratios for complex genetic models, *IMA Journal of Mathematical Applications in Medicine and Biology* **8**, 149–169.

[33] Thompson, E.A. & Shaw, R.G. (1990). Pedigree analysis for quantitative traits: variance components without matrix inversion, *Biometrics* **46**, 399–413.

[34] Thompson, E.A. & Shaw, R.G. (1992). Estimating polygenic models for multivariate data on large pedigrees, *Genetics* **131**, 971–978.

[35] Wright, S. (1968). *Evolution and the Genetics of Populations*, Vol. I. University of Chicago, Chicago.

[36] Wright, S. (1980). Genic and organismic selection, *Evolution* **34**, 825–843.

[37] Yule, G.U. (1906). On the theory of inheritance of quantitative compound characters and the basis of Mendel's law: a preliminary note, in *Proceedings of the Third International Congress on Genetics,* pp. 140–142.

(*See also* **Genetic Correlations and Covariances**; **Heritability**; **Identity Coefficients**)

SUN-WEI GUO

Polymorphism Information Content

The polymorphism information content (PIC) value is a measure of polymorphism introduced by Botstein et al. [1] to describe a genetic marker's usefulness for **linkage analysis** when attempting to localize on a chromosome the **gene** locus involved in a rare dominant disease. The **genotype** of a person with the disease is assumed to be **heterozygous**, and the PIC value is then defined as the probability that the marker genotype of such a person's offspring would allow one to deduce which marker allele was derived from the affected parent. If there are n alleles at the marker locus, the relative frequency of the ith being p_i, then, assuming **Hardy–Weinberg equilibrium**, they showed that in this situation

$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - 2 \sum_{i=1}^{n-1} \sum_{j=1}^{n} p_i^2 p_j^2. \qquad (1)$$

Estimating PIC and its standard error are discussed in [4].

As the number of alleles increases, each with decreasing population frequency, the PIC value approaches its maximum value of unity.

Most authors who quote PIC values base them on (1). However, Chakravarti & Buetow [3] derived analogous formulae for mapping rare traits with other modes of inheritance, and Chakravarti [2] derived expressions for the joint PIC value of two markers for a family structure consisting of four grandparents, two parents, and their offspring.

References

[1] Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms, *American Journal of Human Genetics* **32**, 314–331.

[2] Chakravarti, A. (1991). Information content of the Centre d'Etude du Polymorphisme Humain (CEPH) family structures for linkage studies, *Human Genetics* **87**, 721–724.

[3] Chakravarti, A. & Buetow, K.H. (1985). A strategy for using multiple linked markers for genetic counseling, *American Journal of Human Genetics* **37**, 984–997.

[4] Shete, S., Tiwari, H. & Elston, R.C. (2000). On estimating the heterozygosity and polymorphism information content value, *Theoretical Population Biology* **57**, 265–271.

(*See also* **Linkage Information Content**)

ROBERT C. ELSTON

Polymorphism

In biology, polymorphism refers to the many different forms that an organism can have. In genetics, it refers to the many different **genotypes** and phenotypes or gene products associated with a particular **gene** locus. A locus was originally defined to be polymorphic if the frequency of its least common allele is too large to be due simply to recurrent mutation, operationally taken to be at least 1% [1]. However, a locus at which there are 1000 distinct alleles each with a frequency of 0.1% would be considered extremely polymorphic, and so a more accurate definition of a polymorphic locus would put an upper bound, such as 99%, on the frequency of its most common allele. DNA polymorphisms abound. Many authors now use the terms "polymorphism" or "gene polymorphism" to denote a common allele at a locus that does not cause disease (*see* **Mutation**).

Reference

[1] Ford, E.B. (1940). Polymorphism and taxonomy, in *The New Systematics*, J. Huxley, ed. Clarendon Press, Oxford, pp. 493–513.

ROBERT C. ELSTON

Polynomial Approximation

A polynomial is a function that can be written in the form

$$p(x) = c_0 + c_1 x + \cdots + c_n x^n,$$

for some coefficients $c_0, \ldots, c_n$. If $c_n \neq 0$, then the polynomial is said to be of order n. A first-order (linear) polynomial is just the equation of a straight line, while a second-order (quadratic) polynomial describes a parabola.

Polynomials are just about the simplest mathematical functions that exist, requiring only multiplications and additions for their evaluation. Yet they also have the flexibility to represent very general nonlinear relationships. Approximation of more complicated functions by polynomials is a basic building block for a great many numerical techniques.

There are two distinct purposes to which polynomial approximation is put in statistics. The first is to model a nonlinear relationship between a **response variable** and an **explanatory variable** (*see* **Nonlinear Regression**; **Response Variable**). The response is usually measured with error, and the interest is on the shape of the fitted curved and perhaps also on the fitted polynomial coefficients. The demands of **parsimony** and interpretability ensure that one will seldom be interested in polynomial curves of more than third or fourth order in this context.

The second purpose is to approximate a difficult to evaluate function, such as a density or a distribution function, with the aim of fast evaluation on a computer. Here, there is no interest in the polynomial curve itself. Rather, the interest is on how closely the polynomial can follow the special function, and especially on how small the maximum error can be made. Very high order polynomials may be used here if they provide accurate approximations. Very often, a function is not approximated directly, but is first transformed or standardized so as to make it more amenable to polynomial approximation.

On either type of problem, substantial benefit can be had from orthogonal polynomials (*see* **Orthogonality**). Orthogonal polynomials can be used to make the polynomial coefficients uncorrelated, to minimize the error of approximation, and to minimize the sensitivity of calculations to roundoff error.

Suppose that the function to be approximated, $f(x)$, is observed at a series of values $x_1, \ldots, x_N$. In general, we will observe $y_i = f(x_i) + \varepsilon_i$, where the ε_i are unknown errors. The task is to estimate $f(x)$ for new values of x. If the new x is within the range of the observed abscissae, then the problem is *interpolation*. If it is outside, then the problem is **extrapolation**. Polynomials are useful for interpolation, but notoriously poor at extrapolation.

Polynomial approximation is relatively straightforward and good enough for many purposes. There are, however, many other ways to approximate functions. Many functions, for example, can be more economically approximated by rational functions, which are quotients of polynomials. A survey of approximation methods is given by Press et al. [4, Chapter 4].

Most numerical analysis texts include a treatment of polynomial approximation. Atkinson [2, Chapter 4] gives a nice treatment of minimax approximation using Chebyshev polynomials. Many specific polynomial approximation formulae to functions used by statisticians are given by Abramowitz & Stegun [1]. Many statistical texts mention polynomial regression. Kleinbaum et al. [3, Chapter 13] give a very accessible treatment, while that of Seber [5, Chapter 8] is more detailed and mathematical.

Taylor's Theorem

Use of polynomials can be motivated by Taylor's theorem. A well-behaved function f can be approximated about a point x by

$$f(x + \delta) \approx f(x) + f'(x)\delta + f''(x)\frac{\delta^2}{2!} + \cdots.$$

The right-hand side, which is a polynomial in δ, is an accurate approximation provided that δ is small.

The trouble with Taylor's theorem is that the error of approximation is not evenly distributed. The formula is very accurate for δ near zero, but becomes increasingly inaccurate as δ increases. Consider the cubic Taylor series expansion for e^x about zero on the interval $[-0, 3]$ (Figure 1). The approximation is accurate for x near zero, but becomes poor for large values of x. Meanwhile, there are other cubic polynomials which follow e^x with good accuracy over the entire interval. The Holy Grail of polynomial approximation is to find the polynomial that minimizes the maximum deviation of the polynomial from

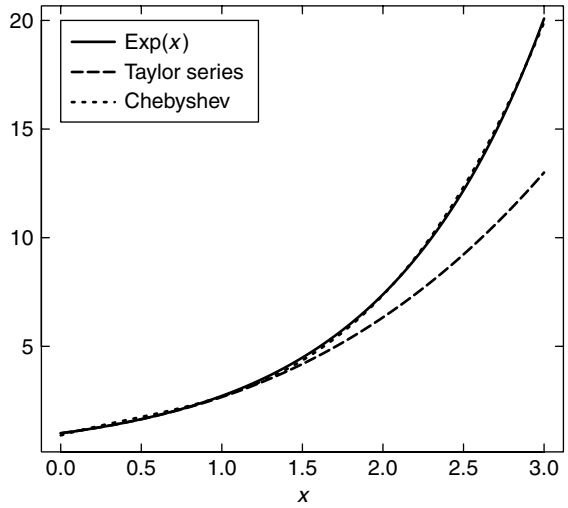

Figure 1 The cubic Taylor series approximation to $\exp(x)$ is accurate only near zero. The cubic Chebyshev polynomial approximation is indistinguishable from the function itself

the function over the entire interval, the so-called *minimax* polynomial.

Orthogonal Polynomials

The general polynomial $p(x)$ above was written in terms of the *monomials* x^j. This is known as the *natural form* of the polynomial. The trouble with the natural form is that the monomials all look very similar when plotted on [0, 1]; that is, they are very highly correlated. This means that small changes in $p(x)$ may arise from relatively large changes in the coefficients $c_0, \ldots, c_n$. The coefficients are not well determined when there is measurement or roundoff error.

The general polynomial can just as well be written in terms of any sequence of basic polynomials of increasing degree,

$$p(x) = a_0 p_0(x) + a_1 p_1(x) + \cdots + a_n p_n(x),$$

where the degree of $p_j(x)$ is j, for $j = 0, \ldots, n$. There is a linear relationship between the original coefficients c_j and the new coefficients a_j to make the resulting polynomial the same in each case.

The idea behind orthogonal polynomials is to select the basic polynomials $p_j(x)$ to be as different from each other as possible. Two polynomials p_i and p_j are said to be *orthogonal* if $p_i(X)$ and $p_j(X)$ are uncorrelated as X varies over some distribution. *Legendre polynomials* are uncorrelated when X is uniform on $(-1, 1)$. *Chebyshev polynomials* are uncorrelated when X is Beta(1/2,1/2) on $(-1, 1)$. *Laguerre polynomials* are uncorrelated when X is gamma on $(0,\infty)$. *Hermite polynomials* are uncorrelated when X is standard normal on $(-\infty, \infty)$.

Any sequence of orthogonal polynomials can be calculated recursively using a three-term recurrence formula. For example, the Chebyshev polynomials satisfy

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= x, \\ p_2(x) &= 2x^2 - 1, \\ &\cdots \\ p_{n+1}(x) &= 2x p_n(x) - p_{n-1}(x), \quad n \geq 1. \end{aligned}$$

Another important property of orthogonal polynomials is that $p_n(x)$ changes sign (and has a zero) n times in the interval of interest. The zeros of the nth order Chebyshev polynomial occur at

$$x_k = \cos\left(\pi \frac{k - 0.5}{n}\right), \quad k = 1, \ldots, n.$$

The Chebyshev polynomials also have the property of bounded variation. The local maxima and minima of Chebyshev polynomials on $[-1, 1]$ are exactly equal to 1 and -1, respectively, regardless of the order of the polynomial. It is this property which makes them valuable for minimax approximation. In fact, an excellent approximation to the nth order minimax polynomial on an interval can be obtained by finding the polynomial that satisfies $p_n(x) = f(x)$ at the zeros of the $(n + 1)$th order Chebyshev polynomial. Figure 1 shows the use of a third- order Chebyshev polynomial to approximate the function e^x on the interval [0, 3]. The error is less than 0.18 over the whole interval.

As another example, consider the problem of approximating the tail probability of the normal probability distribution function, $1 - \Phi(x)$, for $x > 0$. Since the tail probability decreases rapidly as x increases, we consider the ratio of the tail probability to the normal density function $[1 - \Phi(x)]/\phi(x)$. Finally, we transform x to $y = (x - 1)/(x + 1)$, which takes values on $[-1, 1]$. The resulting function $f(y) = \{1 - \Phi[x(y)]\}/\phi[x(y)]$ looks nearly linear

and can be well approximated by a polynomial. The tenth-order polynomial which interpolates $f(y)$ on the Chebyshev points on $[-1, 1]$ approximates $f(y)$ to nine or more significant figures, and hence gives an approximation to $\Phi(x)$ that remains accurate to ten significant figures for very large values of x.

Polynomial Regression

Now we turn to the case in which the nonlinear function is observed with error. Suppose that we observe $(x_i, y_i), i = 1, \ldots, N$, where

$$y_i = f(x_i) + \varepsilon_i,$$

where f is some nonlinear function and the ε_i are uncorrelated with mean zero and constant variance.

Consider height as a function of age for 318 girls who were seen in a disease study [6] in East Boston in 1980 (Figure 2). Height might be described roughly by a straight line over a short range of ages – say, ages 5–10 – but over wider age ranges a more general function is needed. We initially fit a sixth-order polynomial,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 + \beta_6 x_i^6 + \varepsilon_i,$$

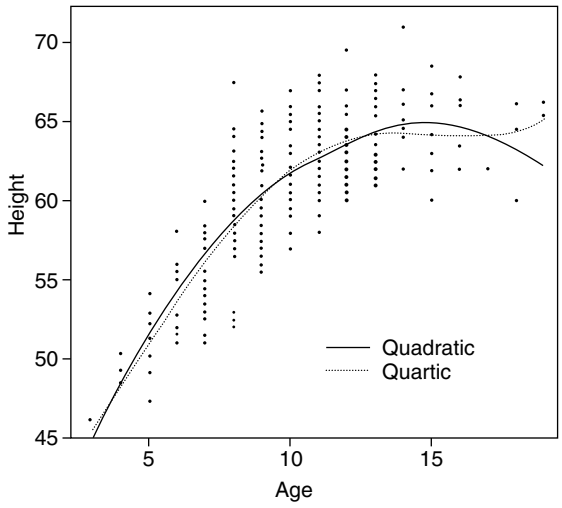

Figure 2 Height (in inches) vs. age (in years) for 318 girls who were seen in 1980 in the Childhood Respiratory Disease Study in East Boston, Massachusetts

with the intention of decreasing the order later if a simpler polynomial is found to fit just as well. This leads to a **multiple linear regression** problem for the coefficients $\beta_0, \ldots, \beta_6$, in which the design matrix is

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^6 \\ 1 & x_2 & x_2^2 & \ldots & x_2^6 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{318} & x_{318}^2 & \ldots & x_{318}^6 \end{pmatrix}.$$

The columns of this matrix are very nearly **collinear**, which will make the least squares problem ill-conditioned. Nevertheless, we can obtain results from a statistical package: the regression overall is highly significant, with an F statistic of 135.7 on 6 and 311 df. However, the table of coefficients and standard errors offers little guidance as to what order of polynomial is required (Table 1). None of the regression coefficients are significantly different from zero, a reflection of the high correlations between the coefficients (Table 2). We could determine the smallest adequate order for the polynomial by fitting, in turn, a fifth-order polynomial, a fourth-order, a third-order, and so on. At each step we could test for the neglected monomial term using an adjusted F statistic. A more satisfactory solution, however, is to use orthogonal polynomials.

Table 1 Coefficients and standard errors for polynomial regression of height on age for the respiratory disease study

Coefficient	Value	se	t value	P value
β_0	80.2384	32.9342	2.4363	0.0154
β_1	−26.9075	23.0292	−1.1684	0.2435
β_2	7.8563	6.3456	1.2381	0.2166
β_3	−1.0296	0.8856	−1.1627	0.2459
β_4	0.0712	0.0663	1.0737	0.2838
β_5	−0.0025	0.0025	−1.0020	0.3171
β_6	0.0000	0.0000	0.9503	0.3427

Table 2 Correlation matrix for the polynomial regression coefficients

	β_0	β_1	β_2	β_3	β_4	β_5
β_1	−0.9935					
β_2	0.9774	−0.9950				
β_3	−0.9558	0.9824	−0.9960			
β_4	0.9313	−0.9650	0.9860	−0.9969		
β_5	−0.9058	0.9451	−0.9721	0.9888	−0.9975	
β_6	0.8805	−0.9241	0.9559	−0.9776	0.9910	−0.9980

Many statistical programs allow one to compute a sequence of polynomials which are orthogonal with respect to the observed values of x; that is, which satisfy

$$\sum_{k=1}^{N} p_i(x_k)p_j(x_k) = 0, \quad i \neq j.$$

(The function ORPOL is part of PROC MATRIX or PROC IML in SAS. In **S-PLUS** or R, the function is `poly` (*see* **Software, Biostatistical**).) It is also convenient to choose the polynomials so that

$$\sum_{k=1}^{N} p_i(x_k)^2 = 1, \quad i = 0, 1, \ldots, N-1.$$

In terms of these polynomials, the multiple regression model becomes

$$y_i = \alpha_0 p_0(x_i) + \alpha_1 p_1(x_i) + \cdots + \alpha_6 p_6(x_i) + \varepsilon_i,$$

where again there is a linear relationship between the coefficients α_j of the orthogonal polynomials and the original β_j. This model has the same fitted values, sums of squares, and F ratio as the original model. However, because of orthogonality, the least squares estimates of the α_j are uncorrelated and have identical standard errors, which greatly simplifies interpretation. In fact, each estimated coefficient $\hat{\alpha_j}$ is unchanged by the actual order of the polynomial which has been fitted.

The estimated coefficients and standard errors for the orthogonal polynomial regression are given in Table 3. In this case, the **Student's *t* statistics** and ***P* values** for the coefficients directly relate to the significance of cubic, quartic, and so on, components of the regression. We can see that the fifth- and sixth-order terms are not required, but that the third- and fourth-order terms are approaching significance. A plot of the quadratic and quartic fitted values against age (Figure 2) shows that the quartic fit might be preferred, because the quadratic is not monotonic in the observed range.

Table 3 Coefficients and standard errors for orthogonal polynomial regression of height on age for the respiratory disease study

Coefficient	Value	se	t value	P value
α_0	60.2119	0.1426	422.1543	0.0000
α_1	65.0285	2.5435	25.5669	0.0000
α_2	−31.3549	2.5435	−12.3276	0.0000
α_3	4.4838	2.5435	1.7629	0.0789
α_4	4.9562	2.5435	1.9486	0.0522
α_5	−2.1465	2.5435	−0.8439	0.3994
α_6	2.4170	2.5435	0.9503	0.3427

References

[1] Abramowitz, M. & Stegun, I.A. (1962). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington. Reprinted by Dover, New York, 1965.

[2] Atkinson, K.E. (1989). *An Introduction to Numerical Analysis*, 2nd Ed. Wiley, New York.

[3] Kleinbaum, D.G., Kupper, L.L. & Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*. PWS-Kent, Boston.

[4] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in Fortran*. Cambridge University Press, Cambridge. (Also available for C, Basic, and Pascal.).

[5] Seber, G.A.F. (1977). *Linear Regression Analysis*. Wiley, New York.

[6] Tager, I.B., Weiss, S.T., Rosner, B. & Speizer, F.E. (1979). Effect of parental cigarette smoking on pulmonary function in children, *American Journal of Epidemiology* **110**, 15–26.

(*See also* **Numerical Analysis**; **Numerical Integration**)

GORDON K. SMYTH

Polynomial Regression

It has been known for over a century [12] that a polynomial may be found which will approximate an arbitrary continuous function on a finite interval as closely as may be desired. This, one presumes, is the mathematical motivation for the introduction of polynomials into data analysis, which appears to have occurred in the early twentieth century (e.g. [3]).

At least in the initial stages of analysis, statistical models are often defined to be linear both in the parameters and in the covariates. In the simple situation of a single continuous covariate, X, the predicted value of the outcome variable in such a model is $\beta_0 + \beta_1 X$. The parameters β_0 and β_1 are estimated from the data, usually by **maximum likelihood**. If a technique such as analysis of **residuals** suggests that a straight line is a poor fit, one may wish to extend the model to accommodate nonlinearity in the relation between the outcome and X. One way is to fit a polynomial $\beta_0 + \beta_1 X + \beta_2 X^2 + \ldots$. This polynomial regression model, though no longer linear in X, is still linear in the parameters and therefore may be estimated by standard **multiple linear regression** techniques. Such models can be used for normal errors regression, but also for the **generalized linear model** (such as **logistic regression**) and Cox proportional hazards regression (*see* **Cox Regression Model**).

Polynomial regression has two main uses. The first is as a method of representing curved regression relationships in observational data and in designed experiments. The approach has been described by many writers, for example [11]. The second is essentially as a diagnostic for curvature (*see* **Diagnostics**). This is popular in the analysis of epidemiological data, where one may fit multiple regression models which involve a risk factor, X, of central interest together with several **confounding** variables. A hypothesis test for including a quadratic term (i.e. in X^2) in the model serves as a crude overall test of curvature.

Although polynomials in more than one variable (see Definitions, below) are known, they appear to be scarcely, if at all, used in biostatistical applications.

Definitions

A univariate polynomial of degree k in a scalar variable X is defined as

$$\sum_{i=0}^{k} \beta_i X^i,$$

where $X^0 \equiv 1$. A model may, of course, contain univariate polynomials in several covariates. A bivariate polynomial of degree k in a vector $\mathbf{X} = (X_1, X_2)'$ is defined as

$$\sum_{0 \le i+j \le k} \beta_{ij} X_1^i X_2^j,$$

with extension in similar fashion to higher dimensions (numbers of components in $\mathbf{X}$). The βs are parameters that need to be estimated from the data.

Advantages of Polynomials

As model functions, polynomials have several advantages. They are simple and familiar functions that are easily described and reported in nonstatistical journals. Fitting is easy using standard regression software, available in all statistical packages (*see* **Software, Biostatistical**). The resulting curves are independent of the origin and scale chosen for X, and at least for low-order polynomials ($k \le 2$) are very smooth. Derivatives of any order of the fitted curves with respect to X are easily calculated and are also smooth. Since the models are fully parametric, the standard array of tools for statistical inference – hypothesis tests of parameters, confidence intervals, and so on – is available.

Drawbacks

Unfortunately, polynomials also suffer from severe disadvantages. The low-order models are quite inflexible in the range of curve shapes they offer, whereas the higher-order ones often display artefacts such as "end effects" and "wiggliness". The Runge phenomenon [9] is an extreme manifestation of the latter (see Examples). Polynomials cannot have asymptotes as X tends to infinity, so they are unsuitable models for such relationships. Certain types of nonlinear model (*see* **Nonlinear Regression**) are almost invariably preferable. To avoid inaccuracies in parameter

estimates due to the high correlations between the powers of X in the model, special techniques such as centering or orthogonalization are needed (*see* **Orthogonality**). It is sometimes difficult to know whether a feature of the fitted polynomial (such as a peak or a nadir) is real or not. This matters in certain applications where the shape of the unknown true function is of central importance. An example is the so-called J-shaped risk relationship in epidemiology [4]. Polynomials are nonrobust in that a small number of values can considerably affect the parameter estimates and hence the entire curve shape.

Related Techniques

Royston & Altman [8] have described "fractional polynomials" which include noninteger and fractional powers of X, and have illustrated their uses with several biostatistical examples. Fractional polynomials are much more flexible than conventional polynomials and share most of their advantages. Other approaches are variants of **nonparametric regression**, and include the scatterplot smoother [1], the **spline function** ([5, 10]) and the **generalized additive model** (GAM) [7]. Although the statistical literature on nonparametric regression modeling is large and growing, alternatives to polynomials such as GAMs and fractional polynomials do not seem to be much used by authors of medical and epidemiologic papers. The situation may change as software becomes more widely available.

Examples

We shall illustrate some features of polynomial regression using two data sets. The first comprises observations on 478 patients rescued by the helicopter ambulance service of the Royal London Hospital [2]. Each patient received an injury severity score (X) between 1 and 76 at the rescue location (1 being least severe), and either died ($Y = 0$) or recovered ($Y = 1$) following subsequent hospital treatment. We attempt to predict the probability of recovery as a function of the severity score. Figure 1 shows the results of fitting several logistic regression models to the data.

The continuous line, a quadratic polynomial, shows a biologically implausible rise in the probability of recovery when $X > 60$. A cubic model (short dashes) is a significantly better fit ($P = 0.02$) but is showing some "waviness" near $X = 55$. A fractional polynomial of degree 1 (long dashes) has

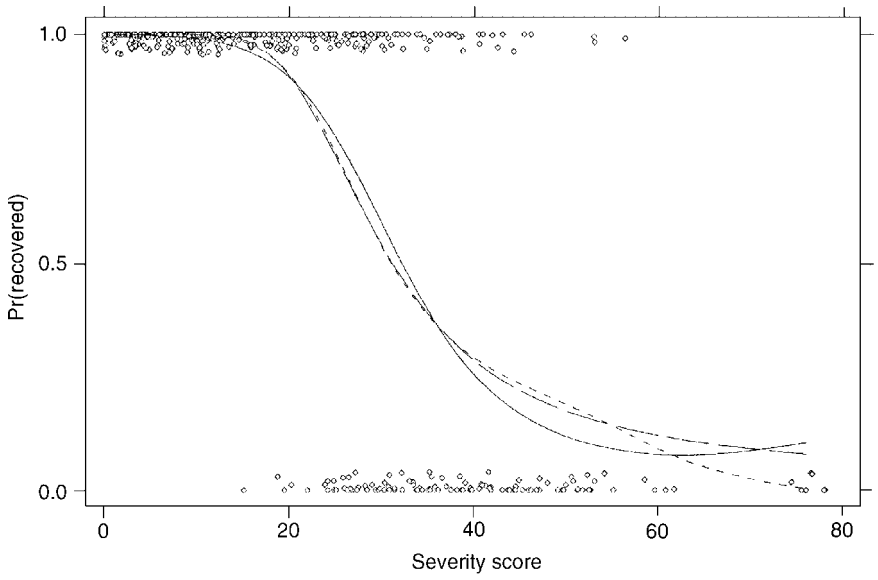

Figure 1 Fitted curves for three logistic regression models for recovery following rescue by helicopter ($n = 478$): quadratic (———); cubic ( ); fractional polynomial of degree 1 (– – –). The original binary observations (◊) are "jittered" randomly to separate the points

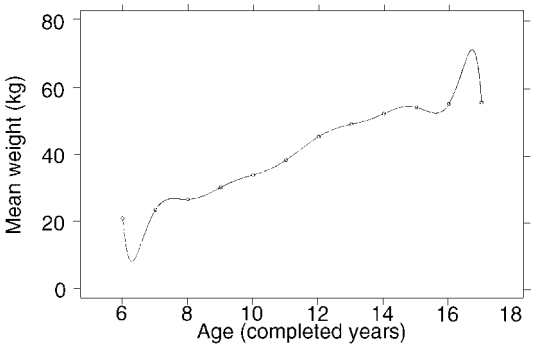

Figure 2 Mean girls' weight plotted against age group ($n = 12$), together with fitted eleventh order polynomial exhibiting the Runge phenomenon

a log **likelihood** only 0.6 lower than the cubic, but has a smoother and more convincing curve shape than either of the polynomial models. The fractional polynomial function found here is $\beta_0 + \beta_1 X^{-1/2}$, which has an asymptote of β_0 as X tends to infinity.

We use a second data set to illustrate the Runge phenomenon. Figure 2 is a plot of the mean bodyweights of several thousand US girls according to completed years of age between 6 and 17 years [6].

There are 12 pairs of observations. The continuous line shows the behavior of an eleventh order polynomial fitted to the data. The fitted line passes exactly through every observation, but between observations it oscillates wildly (the Runge phenomenon), particularly near the extreme ages. It is therefore quite useless for estimating the mean weight at intermediate ages.

The example is artificial in that an interpolating polynomial has been used, something one would be unlikely to do in practice since the "model" is clearly overfitted. However, the same thing can happen in realistic situations and its presence may escape the notice of the analyst. For example, connecting the fitted values with straight line segments between observations, as is often done automatically by statistical software, may conceal the oscillations in the true fitted curve.

References

[1] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

[2] Evans, S.J.W. (1996). Personal communication.

[3] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, London.

[4] Goetghebeur, E.J.T. & Pocock, S.J. (1995). Detection and estimation of J-shaped risk-response relationships, *Journal of the Royal Statistical Society, Series A* **158**, 107–121.

[5] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, New York.

[6] Hamill, P.V.V., Drizd, T.A., Johnson, C.L., Reed, R.B. & Roche, A.F. (1977). *NCHS Growth Curves for Children, Birth–18 Years, United States*. US Department of Health, Education and Welfare, Washington.

[7] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, New York.

[8] Royston, P. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[9] Runge, C. (1901). Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten (On empirical functions and the interpolation between equidistant ordinates), *Zeitschrift Mathematische Physik*, **46**, 224–243.

[10] Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion), *Journal of the Royal Statistical Society, Series B* **47**, 1–52.

[11] Snedecor, G.W. & Cochran, W.G. (1989). *Statistical Methods*, 8th Ed. Iowa State University Press, Ames.

[12] Weierstrass, K. (1885). *Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen reeller Argumente (On the Analytic Representability of Arbitrary Functions of a Real Argument)*. Sitzungsberichte der Akademie, Berlin, pp. 633–639, 789–805.

PATRICK ROYSTON

Polytomous Data

In many studies the response of interest is restricted to a fixed set of possible values, the so-called response categories. Response variables of this type are called *polytomous*, or, less frequently, *polychotomous*. Examples are side effects of medical treatment with the possible categories headache, insomnia, or nausea, or several types of infection that may follow an operation. Most rating scales have fixed response categories that measure, for example, the medical condition after some treatment in categories such as good, fair, or poor, or the severity of symptoms in categories such as none, mild, moderate, or marked.

These examples show immediately that there are at least two cases to be distinguished, namely the case in which response categories are mere labels which have no inherent ordering and the case in which categories are ordered. In the first case the response Y is measured on a nominal scale (*see* **Nominal Data**). Instead of using the numbers $1, \ldots, k$ for the response categories, any set of k numbers would do. In the latter case the response is measured on an ordinal scale, on which the ordering of the categories and the corresponding numbers may be interpreted, but not the distance or spacing between categories (*see* **Ordered Categorical Data**). Examples of nominal and ordinal response categories are given in Figures 1 and 2. In the nominal case the response categories are scattered over the plane in order to show that categories have no systematic ordering. In the ordinal case the response categories are given on a straight line, thus illustrating the ordering of the categories.

Another type of response category that contains more structure than the nominal case but is not

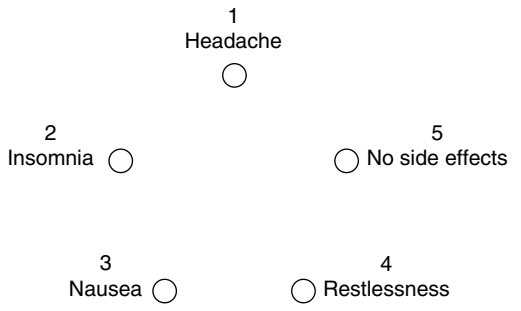

Figure 1 Side effects as example for nominal response categories where the numbers $1, \ldots, k$ are mere labels

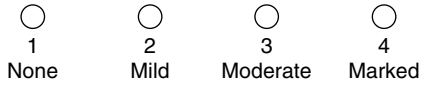

Figure 2 Severity of symptoms as ordered categories

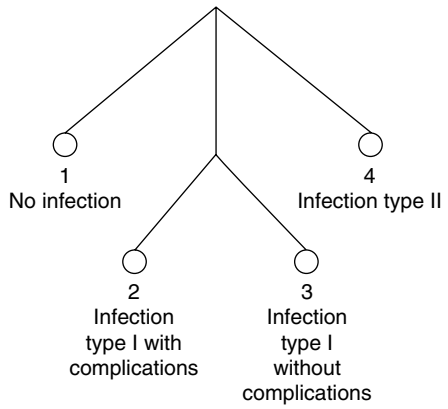

Figure 3 Type of infection as nested structure

captured by simple ordering is given by the nested or hierarchical response category. An example in which the basic response is in the categories "no infection", "infection type I," and "infection type II" is shown in Figure 3. However, for infection type I two cases have to be distinguished; namely, infection with and without additional complications. Thus, one has splits on two levels; first the split into basic categories and then the conditional split for the outcome "infection type I".

Nominal Response: The Multinomial Logit Model

Let $1, \ldots, k$ represent the response categories, i.e. $Y \in \{1, \ldots, k\}$, and let $\mathbf{x}' = (x_1, \ldots, x_p)$ be a vector of **explanatory variables** containing an intercept. The objective is to investigate the effect of the explanatory variables upon the probability of response categories $\pi_r(\mathbf{x}) = \Pr(Y = r|\mathbf{x}), r = 1, \ldots, k$. In the case of nominal response categories the most widespread model is the *multinomial logit model*, which assumes

$$\pi_r(\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_r)}{\sum_{s=1}^{k} \exp(\mathbf{x}'\boldsymbol{\beta}_s)}, \quad r = 1, \ldots, k, \qquad (1)$$

with the parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ being specific for the response categories. The simple structure of the model becomes obvious if one considers two response categories r and s, for which one obtains the **odds**

$$\frac{\pi_r(\mathbf{x})}{\pi_s(\mathbf{x})} = \exp[\mathbf{x}'(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s)] \tag{2}$$

or, equivalently, the log odds

$$\log \frac{\pi_r(\mathbf{x})}{\pi_s(\mathbf{x})} = \mathbf{x}'(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s). \tag{3}$$

Thus, by (2), the odds that the response is in category r instead of category s are specified by an exponential term depending on the difference of parameters $\boldsymbol{\beta}_r - \boldsymbol{\beta}_s$, whereas the log odds or logits depend linearly on the explanatory vector $\mathbf{x}$, with weights determined by the difference $\boldsymbol{\beta}_r - \boldsymbol{\beta}_s$.

Parameter interpretation may be illustrated by the example from Figure 1 with dichotomous covariates such as treatment ($\mathbf{x}_\mathrm{T}$; one, new; zero, conventional) and a continuous covariate such as age centered at 30 years ($\mathbf{x}_\mathrm{A}$). Then one has, for example, for the response categories 2 (insomnia) and 5 (no side effect)

$$\begin{aligned}\frac{\Pr(\text{insomnia})}{\Pr(\text{no side effects})} &= \frac{\pi_2(\mathbf{x})}{\pi_5(\mathbf{x})} \\ &= \exp(\beta_{02} - \beta_{05}) \\ &\quad \times \exp(\beta_{\mathrm{T}2} - \beta_{\mathrm{T}5})^{x_\mathrm{T}} \\ &\quad \times \exp(\beta_{\mathrm{A}2} - \beta_{\mathrm{A}5})^{x_\mathrm{A}}.\end{aligned}$$

That means that $\exp(\beta_{02} - \beta_{05})$ gives the baseline odds that insomnia occurs instead of no side effects, $\exp(\beta_{\mathrm{T}2} - \beta_{\mathrm{T}5})$ is the factor by which these odds increase or decrease for the new treatment, and $\exp(\beta_{\mathrm{A}2} - \beta_{\mathrm{A}5})$ gives the multiplicative increase or decrease per year of age.

Since in the model given in (1) it is only possible to investigate the effect of $\mathbf{x}$ upon the preference of response categories over the alternatives, not all of the parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ are identifiable. It is common to choose a reference category, for example k, and take $\boldsymbol{\beta}_k = \mathbf{0}$. Then the model reduces to

$$\pi_r(\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{k-1} \exp(\mathbf{x}'\boldsymbol{\beta}_s)}, \quad r = 1, \ldots, k-1,$$

and

$$\pi_k(\mathbf{x}) = \frac{1}{1 + \sum_{s=1}^{k-1} \exp(\mathbf{x}'\boldsymbol{\beta}_s)}.$$

Although (1) seems to imply strong assumptions about the relation between explanatory variables and the response variable, this is not generally true. In particular, for categorical explanatory variables the model may just represent a reparameterization of the conditional response categories; that is, it is a saturated model.

For simplicity, let us consider the case of two dichotomous explanatory variables, such as gender and treatment. With x_G being a dummy variable for gender (1, male; 0, female) and x_T coding treatment (1, new treatment; 0, conventional), consider the logit model

$$\begin{aligned}l_r(\mathbf{x}) &= \log \frac{\pi_r(\mathbf{x})}{\pi_k(\mathbf{x})} \\ &= \beta_{0r} + x_\mathrm{G}\beta_{\mathrm{G},r} + x_\mathrm{T}\beta_{\mathrm{T},r} + x_\mathrm{G}x_\mathrm{T}\beta_{\mathrm{G}\times\mathrm{T},r}, \\ r &= 1, \ldots, k-1.\end{aligned} \tag{4}$$

The model includes the explanatory variables x_G and x_T and their product $x_\mathrm{G}x_\mathrm{T}$, the latter representing the **interaction** between gender and treatment. For each category $r = 1, \ldots, k-1$, there are four parameters $\beta_{0r}, \beta_{\mathrm{G},r}, \beta_{\mathrm{T},r}$, and $\beta_{\mathrm{G}\times\mathrm{T},r}$ yielding a total of $4(k-1)$ parameters. However, for each of the four combinations of gender and treatment there are $k-1$ probabilities $\pi_r(x_\mathrm{G}, x_\mathrm{T}) = \Pr(Y = r|x_\mathrm{G}, x_\mathrm{T}), r = 1, \ldots, k-1$. The last one is determined by $\sum_r \Pr(Y = r|\mathbf{x}) = 1$. Consequently, the model contains as many parameters as probabilities, and any set of probabilities can be represented in the form given in (4). This may be explicitly shown by solving the system of equations resulting for different combinations of gender and treatment.

Generally, for categorical covariates any set of response probabilities $\pi_1(x), \ldots, \pi_k(x)$ may be represented by the multinomial logit model if interaction terms are included into the linear term. For three dichotomous variables for example, one has to include x_1, x_2, and x_3 and all of the interactions x_1x_2, x_1x_3, x_2x_3, and $x_1x_2x_3$. Parameter interpretation may again be based on the form given in (2), or directly on the logit form given in (3) or (4). The intercept $\beta_{0r} = \log[\pi_r(0, 0)/\pi_k(0, 0)]$ gives the odds of category r in comparison with the reference category k

if $x_G = 0$ and $x_T = 0$. The effect of x_G, given by

$$\beta_{G,r} = l_r(1,0) - l_r(0,0) = \log\left\{\frac{[\pi_r(1,0)/\pi_k(1,0)]}{[\pi_r(0,0)/\pi_k(0,0)]}\right\},$$

represents the difference between the logit for $x_G = 1$ and the logit for $x_G = 0$ (holding $x_T = 0$ constant). It may also be seen as the logarithm of the **odds ratio** between the corresponding odds. The interaction effect, given by

$$\beta_{G\times T,r} = l_r(1,1) - l_r(-,1) - l_r(1,0) + l_r(0,0) = \frac{\left[\dfrac{\pi_r(1,1)}{\pi_k(1,1)} \Big/ \dfrac{\pi_r(0,1)}{\pi_k(0,1)}\right]}{\left[\dfrac{\pi_r(1,0)}{\pi_k(1,0)} \Big/ \dfrac{\pi_r(0,0)}{\pi_k(0,0)}\right]},$$

represents the ratio between two odds ratios, reflecting the transition from $x_G = 0$ to $x_G = 1$ for $x_T = 1$ in the nominator and for $x_T = 0$ in the denominator.

If the interaction vanishes – that is, $\beta_{G\times T} = 0$ – the two explanatory variables determine the response separately. A model of this type is strongly related to **loglinear models**, or to graphical models that investigate forms of conditional independence (*see* **Antidependence Models**). Suppose that Y, x_G, and x_T are any three categorical random variables. If x_G and x_T are conditionally independent given Y (which corresponds to a special loglinear model), then the interaction terms between x_G and x_T may be omitted in the multinomial logit model, meaning that the response is determined by regressors x_G and x_T separately. For more than two explanatory variables, there is a larger variety of interaction terms that may be omitted. The resulting models correspond to loglinear models of varying complexity, and therefore various structures of conditional independence are possible. Loglinear as well as graphical models are considered extensively in [7]; for graphical models, see also [26].

In the linear term $\mathbf{x}'\boldsymbol{\beta}$, various types of variables may be included. If a covariate is dichotomous it may be used directly in (0–1)-coding, whereas if it is polytomous one has to use several **dummy variables**. Continuous scaled covariates may enter the linear term by using x, but also in polynomial form including x^2, x^3, etc. Moreover, interactions such as in the simple example above may be used. The essential restriction is the linear form $\mathbf{x}'\boldsymbol{\beta}$, which does not mean linear in the covariates but only linear in the parameters. Of course, if continuous variables are included, or only part of the interaction terms are used, model adequacy should be checked (see later section).

Ordinal Response

The multinomial logit model is based on the logits $l_r(\mathbf{x}) = \log[\pi_r(\mathbf{x})/\pi_k(\mathbf{x})]$, where k is the reference category. By assuming $l_r(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_r$, each logit $l_r(\mathbf{x})$ has its own category specific parameter $\boldsymbol{\beta}_r$. Consequently, a permutation of the response categories does not change the validity of the model. This is also easily seen from the form given in (1), where it is obvious that no ordering of response categories is used.

Cumulative Logits

If the categories are ordered, simpler models that assume more structure will be more appropriate. By using the ordering of categories, one may consider the cumulative logits

$$l_r^c(x) = \log[\Pr(Y \le r|\mathbf{x})/\Pr(Y > r|\mathbf{x})] = \log\left\{\frac{[\pi_1(\mathbf{x}) + \cdots + \pi_r(\mathbf{x})]}{[1 - \pi_1(\mathbf{x}) - \cdots - \pi_r(\mathbf{x})]}\right\}.$$

These cumulative logits represent the logarithm of the odds that the response is below or in category r instead of being above category r. By cumulating over responses $1, \ldots, r$, the ordering is explicitly used. A simple model that is based on cumulative logits is the proportional-odds or cumulative logit model

$$l_r^c(x) = \log\left[\frac{\Pr(Y \le r|\mathbf{x})}{\Pr(Y > r|\mathbf{x})}\right] = \beta_{0r} + \mathbf{x}'\boldsymbol{\beta}, \quad r = 1, \ldots, k-1, \tag{5}$$

where the intercept is now separated; in other words, in $\mathbf{x}' = (x_1, \ldots, x_p)$, only the covariates or explanatory variables are collected. The advantage of the model given in (5) is that by using the ordering the number of parameters is strongly reduced. One has intercepts $\beta_{01}, \ldots, \beta_{0,k-1}$ and only one weight parameter $\boldsymbol{\beta}$ instead of $k-1$ weight parameters

$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}$ as in the multinomial logit model. Thus, the model given in (5) is simpler with respect to parameter economy. The additional structure assumed in (5) becomes obvious by consideration of the cumulative logits for two covariate values $\mathbf{x}_1$ and $\mathbf{x}_2$. One obtains

$$l_r^c(\mathbf{x}_1) - l_r^c(\mathbf{x}_2) = \log \frac{\Pr(Y \le r|\mathbf{x}_1)/\Pr(Y > r|\mathbf{x}_1)}{\Pr(Y \le r|\mathbf{x}_2)/\Pr(Y > r|\mathbf{x}_2)} = (\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\beta},$$

meaning that the ratios of cumulative odds with different covariate values do not depend on the category but only on the difference $\mathbf{x}_1 - \mathbf{x}_2$. This property describes a form of strict stochastic ordering [16]. One has either $l_r^c(\mathbf{x}_1) > l_r^c(\mathbf{x}_2)$ for all r or $l_r^c(\mathbf{x}_1) < l_r^c(\mathbf{x}_2)$ for all r. That means that the total response is shifted toward lower or higher response categories by the transition from subpopulation x_1 to subpopulation x_2.

This shifting is also supported if the model is motivated by an underlying continuous response. In some cases the categorical variable Y may be considered as a coarser version of a latent variable $\tilde{Y}$. For example, the severeness of a headache may subjectively have a continuous scale but a response in categories such as none, mild, or moderate, which are determined by the investigator. One assumes that the observable variable Y and the latent variable $\tilde{Y}$ are linked by

$$Y = r_3, \quad \text{if } \theta_{r-1} \le \tilde{Y} < \theta_r, \tag{6}$$

with threshold values $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = \infty$. Thus the observable response is in category r if the continuous underlying response is in interval $[\theta_{r-1}, \theta_r)$. If the continuous latent variable is given by $\tilde{Y} = -\mathbf{x}'\boldsymbol{\beta} + \varepsilon$, with an error variable that follows the logistic distribution function $F(x) = \exp(x)/[1 + \exp(x)]$, one immediately obtains the cumulative logit model given in (5), with $\beta_{0r} = \theta_r$, $r = 1, \ldots, k-1$. The response mechanism in (6), together with $\tilde{Y} = -\mathbf{x}'\boldsymbol{\beta} + \varepsilon$, shows that transition from population $\mathbf{x}_1$ to population $\mathbf{x}_2$ means a changing of the expectation from $\mathrm{E}\tilde{Y} = -\mathbf{x}_1\boldsymbol{\beta}$ to $\mathrm{E}\tilde{Y} = -\mathbf{x}_2\boldsymbol{\beta}$, and therefore a shifting on the latent scale resulting either in lower or higher response categories.

By assuming an alternative but fixed distribution function F for the noise variable x, one obtains from (6) the more general model

$$\Pr(Y \le r|\mathbf{x}) = F(\beta_{0r} + \mathbf{x}'\boldsymbol{\beta}).$$

Models of this type reduce the number of parameters by exploiting the ordinal response. This advantage is often crucial, since the response in categories contains less information than a continuous response, as is usually assumed in classic linear *regression* models. For further information and applications of cumulative models, see [5, 6, 14, 16], and [22].

Continuation-ratio or Sequential Logits

In many applications the ordering of the response categories is due to a sequential mechanism. If the response is, for example, tonsil size with categories "not enlarged", "enlarged", and "greatly enlarged" [12], then starting from the normal state "not enlarged" (category 1) one may reach "enlarged" (category 3) only if the intermediate state has previously been reached, whatever the duration in this state is. Similar responses are found in discrete duration data. If the time until recovery is measured in months, it will take at least 10 months only if the recovery process has lasted at least through 9 months previously.

Appropriate modeling of categories that are reached only successively may be based on conditional models that model explicitly the process of transition. Let us consider the continuation-ratio or sequential logits

$$l_r^s(\mathbf{x}) = \log[\Pr(Y = r|\mathbf{x})/\Pr(Y \ge r|\mathbf{x})],$$

where the underlying odds determine that the response is in category r instead of a category above or in r. These odds may be rewritten in the form

$$\frac{\Pr(Y = r|\mathbf{x})}{\Pr(Y \ge r|\mathbf{x})} = \Pr(Y = r|Y \ge r, \mathbf{x}),$$

therefore specifying the conditional probability of a response in category r given category r is reached.

The corresponding *continuation-ratio* logit model or *sequential logit* model has the form

$$l_r^s(x) = \log\left[\frac{\Pr(Y = r|\mathbf{x})}{\Pr(Y \ge r|\mathbf{x})}\right] = \mathbf{x}'\boldsymbol{\beta}_r. \tag{7}$$

The model may be seen as consecutive dichotomous steps. Starting in category 1, the first step, namely the potential transition to category 2, is determined by a dichotomous logit model with parameter $\boldsymbol{\beta}_1$. Given that category 2 is reached, the next step –

potential transition to category 3 – is determined by a dichotomous logit model with parameter $\boldsymbol{\beta}_2$. More generally, the rth step is determined by a dichotomous logit model with parameter β_r. Consequently, the parameters have to be interpreted in the same way as for dichotomous logit models, but with reference to the specific step modeled. For the tonsil size example, the model given in (7) means that one is explicitly investigating the effects of covariates upon the transition from enlarged tonsils to greatly enlarged tonsils, given one has at least enlarged tonsils. Actually, any dichotomous model may be used in the consecutive steps. The general sequential model is given by

$$\Pr(Y = r|Y \geq r, \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}_r),$$

with a fixed response function F and the parameter vector possibly depending on r.

An alternative way of interpreting the consecutive steps is within the context of survival analysis. If the response categories correspond to failure in discrete time (in weeks or months), the conditional odds

$$\frac{\Pr(Y = r|x)}{\Pr(Y \geq r|x)}$$

may also be interpreted as a *discrete hazard function* (*see* **Discrete Survival-time Models**), the conditional odds representing the hazard of failure at discrete time point t (corresponding to a time interval). For discrete duration models and the connection to sequential models, see [8, 10, 21], and [23]. Sequential models for ordinal data have been considered in [9, 17], and [25].

Nested or Hierarchical Responses

Let us consider the infection example given in Figure 3. The essential point is that, for response data of the nested type, some hierarchy has to be taken into account. At the first level one has the response into categories "no infection" (category 1), "infection type I" (categories 2 and 3), and "infection type II" (category 4). At the second level, infection type I is split into categories 2 and 3. Of course, one may fit the multinomial logit model for categories 1–4. Then, however, the information that categories 2 and 3 are more similar than the other categories is not used. This becomes even more important if the split into subcategories with and without complications occurs later in time than the primary first level split.

In this case an appropriate hierarchical model is a two-step model. The first level may be modeled by a trichotomous multinomial model with reference category 1, given by

$$\log \frac{\Pr(Y \in \{2, 3\}|\mathbf{x})}{\Pr(Y = 1|\mathbf{x})} = \mathbf{x}'\boldsymbol{\beta}_1^{(1)},$$
$$\log \frac{\Pr(Y = 4|\mathbf{x})}{\Pr(Y = 1|\mathbf{x})} = \mathbf{x}'\boldsymbol{\beta}_2^{(1)}.$$

The conditional split on the second level may be modeled by dichotomous logits

$$\log \frac{\Pr(Y = 3|Y \in \{2, 3\}, \mathbf{x})}{\Pr(Y = 2|Y \in \{2, 3\}, \mathbf{x})} = \mathbf{x}'\boldsymbol{\beta}^{(2)}.$$

The principle is the conditional modeling of consecutive splits. This principle allows hierarchies with more than two levels. In a straightforward way, any category – for example, "infection type I with complications" – can be split up further. The additional structure could be modeled conditionally given the response group "infection type I with complications". Not only dichotomous splits are possible. If infection type I is split into five categories of different forms of complications on the second level, one would use a multinomial model. Consequently, on each level one may use a nominal or an ordinal model, corresponding to the type of response categories on the level.

McCullagh & Nelder [17] illustrate the approach of consecutive dichotomous splits by a study of mortality and fertility of lactating cows with consecutive trials of insemination leading or not leading to pregnancy. Further examples have been given in [3, 15, 18], and [24].

Further approaches to the modeling of ordinal data include "adjacent categories logits", where the logits between adjacent categories of the ordinal scale are considered [1, Chapter 6], and the stereotype model [4], where the ordering results from the modeling.

Estimation, Testing, and Goodness of Fit

The data usually are given as pairs $(Y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, of the response variable Y_i and the covariate vectors $\mathbf{x}_i$. The estimation of the models is most often based on the **maximum likelihood** principle.

The kernel of the multinomial log likelihood has the form

$$l = \sum_{i=1}^{n} \sum_{r=1}^{k-1} y_{ir} \log[\pi_r(\mathbf{x}_i)] + (1 - y_{i1} - \cdots - y_{i,k-1}) \times \log[1 - \pi_1(\mathbf{x}_i) - \cdots - \pi_{k-1}(\mathbf{x}_i)], \quad (8)$$

where the probabilities are given by $\pi_r(\mathbf{x}_i) = \Pr(Y_i = r|\mathbf{x}_i)$ and y_{ir}, $r = 1, \ldots, k-1$, are dummy variables coding the response by $y_{ir} = 1$ if $Y_i = r$ and $y_{ir} = 0$ otherwise. When maximizing (8) the response probabilities have to be replaced by their parametric form. The models considered here are special cases of a (multivariate) **generalized linear model** for multinomial response. That means that some (known) *transformation* of the response probabilities depends linearly on the covariates, as in the models given in (3), (4), and (5). Generalized linear models provide a unified framework in which parameters can be estimated, and the influence of specific variables can be tested. For details, see [2, 11], and [17].

Of particular interest is the **goodness of fit** of models, which may be assessed by considering the difference between the data and the fitted values. If the explanatory variables are categorical, only a limited number of patterns (say g) for the vector $\mathbf{x}$ is possible. Then one may consider grouped data $(\mathbf{p}_i, \mathbf{x}_i)$, $i = 1, \ldots, g$, where $\mathbf{p}_i' = (p_{ir}, \ldots, p_{ik})$ is the vector of relative frequencies computed from the n_i observations taken at pattern $\mathbf{x}_i$. The goodness of fit of the models may be checked by the Pearson statistic

$$\chi^2 = \sum_{i=1}^{g} n_i \sum_{j=1}^{k} \frac{(p_{ir} - \hat{\pi}_{ir})}{\hat{\pi}_{ir}}$$

(*see* **Chi-square Tests**), or the deviance

$$D = 2 \sum_{i=1}^{g} n_i \sum_{r=1}^{k} p_{ir} \log\left(\frac{p_{ir}}{\hat{\pi}_{ir}}\right),$$

where $\hat{\pi}_{ir}$ denotes the maximum likelihood estimate of $\pi_{ir} = \Pr(Y_i = r|x_i)$ (*see* **Likelihood Ratio Tests**). If $n_i \to \infty$ for fixed g (n_i/n converges to $\lambda_i \in (0, 1)$), then both statistics are asymptotically χ^2 distributed with $g(k-1) - p$ degrees of freedom, where p is the length of the parameter vector $\boldsymbol{\beta}$. When applying χ^2 or D, one has to keep in mind that this fixed cell type of asymptotics assumes that the local number of observations n_i must not be too low (*see* **Likelihood Ratio Tests**). The distribution of χ^2 and D is quite different for sparse data ($n \to \infty, g \to \infty$) (see [19] and [20]). If the covariates are continuous then the number of patterns is the same as the number of groups. In this case χ^2 and D will fail as goodness-of-fit measures. Hosmer & Lemeshow [13] give a modification for this case, based on grouping within intervals.

References

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

[2] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. Wiley, New York.

[3] Amemiya, T. (1975). Qualitative response models, *Annals of Economic and Social Measurement* **4**, 363–372.

[4] Anderson, J.A. (1984). Regression and ordered categorical variables, *Journal of the Royal Statistical Society, Series B* **46**, 1–30.

[5] Anderson, J.A. & Phillips, R.R. (1981). Regression, discrimination and measurement models for ordered categorical variables, *Applied Statistics* **30**, 22–31.

[6] Aranda-Ordaz, F.J. (1983). An extension of the proportional-hazard model for grouped data, *Biometrics* **39**, 110–118.

[7] Christensen, R. (1990). *Log-Linear Models*. Springer-Verlag, New York.

[8] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[9] Cox, C. (1988). Multinomial regression models based on continuation ratios, *Statistics in Medicine* **7**, 435–442.

[10] Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve, *Journal of the American Statistical Association* **83**, 414–425.

[11] Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd Ed. Springer-Verlag, New York.

[12] Holmes, M.C. & Williams, R. (1954). The distribution of carriers of Streptococcus pyogenes among 2413 healthy children, *Journal of Hygiene* **52**, 165–179.

[13] Hosmer, D. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[14] Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present, *Applied Statistics* **39**, 74–85.

[15] Kahn, L.M. & Morimune, K. (1979). Unions and employment stability: a sequential logit approach, *International Economic Review* **20**, 217–236.

[16] McCullagh, P. (1980). Regression model for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–127.

[17] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, New York.

[18] Morawitz, B. & Tutz, G. (1990). Parameterizations for business survey data, *ZOR – Methods and Models of Operations Research* **34**, 143–156.

[19] Osius, G. & Rojek, D. (1992). Normal goodness-of-fit tests for parametric multinomial models with large degrees of freedom, *Journal of the American Statistical Association* **87**, 1145–1152.

[20] Read, I. & Cressie, N. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.

[21] Ryu, K. (1994). Group duration analysis of the proportional hazard model: minimum chi-squared estimators and specification tests, *Journal of the American Statistical Association* **89**, 1386–1397.

[22] Terza, J.V. (1985). Ordinal probit: a generalization, *Communications in Statistics – Theory and Methods* **14**, 1–11.

[23] Thompson, W.A. (1977). On the treatment of grouped observations in life studies, *Biometrics* **33**, 463–470.

[24] Tutz, G. (1989). Compound regression models for categorical ordinal data, *Biometrical Journal* **31**, 259–272.

[25] Tutz, G. (1991). Sequential models in ordinal regression, *Computational Statistics and Data Analysis* **11**, 275–295.

[26] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

(*See also* **Logistic Regression**; **Multinomial Distribution**; **Proportional-odds Model**)

GERHARD TUTZ

Popper, Karl R.

Born: July 28, 1902, in Vienna, Austria.
Died: September 17, 1994, in London, UK.

In his *Autobiography* (see [13] or *Unended Quest* [7]), Popper recalls being upset, as a child, by the sight of hunger and poverty in the streets of Vienna. His father was a lawyer and a "radical liberal", his mother was a musician. At the University of Vienna after World War I he studied a variety of subjects from physics, mathematics, and history of music to education, psychology, and philosophy, while earning his living mostly from schoolteaching. The encounter with Marxism, and to a lesser extent with psychoanalysis, made him conscious that he valued free critical thinking over dogmatic theories. After listening with awe to a lecture by Einstein he was reinforced in the conviction that what makes conjectures scientific is that they are open to refutation by experience.

Although not a member of the *Wiener Kreis*, he interacted with members of the Circle. He resisted their idea that philosophical questions are reducible to questions of language. His first manuscript, critical of the doctrines of the Circle, boldly attempted to solve "the two fundamental problems in the theory of knowledge" (the problems of induction and of demarcation) [1]. The manuscript was read and discussed by Feigl, Carnap, Schlick, and Frank, among others. His next work, entitled *Logik der Forschung* (1934), first published in a series edited by Frank & Schlick, would (so Popper thought) defeat the theses of logical positivism, through showing that there is no such thing as "inductive logic" ("induction is a myth"), and that the criterion of demarcation between science and pseudo-science is that scientific propositions are refutable (falsifiable), while pseudo-scientific ones are irrefutable (unfalsifiable) [2]. Valid reasoning is deductive, science is hypothetico-deductive, refutation is a case of *modus tollens*. Scientific theories are neither verifiable nor probable; they are hypothetical. The book was a great success. Popper was invited to lecture abroad. In Prague he met Tarski, whose theory of truth was to influence decisively his own philosophical realism. In England he met with Ayer and Russell, as well as with German-speaking refugees such as Schrödinger and Hayek. On account of the rise of Nazism in Austria, he then accepted a teaching position in the University of New Zealand at Canterbury. And that is how, in 1937, he became a philosophy professor.

"The *Poverty* and the *Open Society* were my war effort": they were an attack on totalitarianism, and a plea for a free society in which people "send their theories to die in their stead"; that is, expose their ideas to criticism and exchange arguments rather than blows [3, 4]. Both texts were published after World War II, thanks to Hayek, who also gave their author the opportunity to return to Europe and teach in London. Popper became a professor in logic and scientific method at the London School of Economics and stayed there for the rest of his career (1945–1969). In spite of a somewhat eccentric marginality he became a celebrity on the British philosophical scene, and was knighted in 1965.

Conjectures and Refutations (1963) expands on the thesis, already present in *Logik der Forschung*, that truth is not manifest, searching for truth takes work, cognitive delusions abound, and scientific theories are mere conjectures, the truth of which can never be conclusively established [5]. Scientific theories, however, have a degree of testability, and as a theory passes more severe tests, it gets more highly "corroborated" (which does not mean confirmed, but so far not disconfirmed). *Objective Knowledge* (1972), a collection of essays, marks a turn towards an evolutionary epistemology [6]. Essay 3 on "epistemology without a knowing subject" presents the reader with the stimulating (and realist) view that scientific problems, theories, and arguments develop within a "third world", largely autonomous with respect to the first (world of physical objects) and second (world of mental states); and that the objective traits of the growth of knowledge (world 3) are of far greater interest to epistemology than the subjective thoughts of scientists (world 2). Essay 6 "on clouds and clocks" generalizes Popper's epistemic theory (of learning through a trial and error-elimination process), via a neo-Darwinist scheme, to a metaphysical theory (of evolution through random variation and natural selection within an "open", i.e. indeterministic, universe). This is rehearsed in a clear and simple style by *A World of Propensities* [12]. *The Self and its Brain* (1977) is a fascinating dialogue between a philosopher (P) and a brain scientist (E) on the mind–body interaction [8].

While tackling a great variety of new subjects, Popper never ceased to work on his initial work,

translated into English in 1959 under the title *The Logic of Scientific Discovery* [2], revised many times, and completed by a huge *Postscript* from which three new volumes have been extracted (see [9, 10], and [11]). The reader of the *Encyclopedia of Biostatistics* will be most interested by *Realism and the Aim of Science* (1982), in which Popper re-examines in detail the problems of truth, corroboration, and **probability**. He maintains that inductive reasoning, that is [9], concluding validly from empirical observations to general regularities, is an illusion. There are neither "inductive sciences", as Mill or Whewell thought, nor a "logic" of induction, as Carnap tried to build, nor a "statistical" or "probabilistic" induction, as statisticians like **R. A. Fisher** tried to sketch. According to Popper, scientific hypotheses do not arise from experience, but are an expression of our freedom to conjecture; they do not have to be plausible, they have to be testable. He is an adept of a selective or Darwinian theory of learning (as opposed to an instructive or Lamarckian theory). We do not learn from the facts, we learn "from our mistakes" (from conjecturing and being refuted).

Popper's views on inductive **inference** gave rise to hot controversies. Inductivists objected that while Popper emphasizes the negative aspect of scientific methods (how to detect and reject errors), he does not take account of the positive aspect (how to make judicious hypotheses). The choice of a theory is guided by experience. Rationally favoring some hypothesis implies careful sampling of data, and estimation of the degree of inductive support the hypothesis gets from known data. Thus, statistical or probabilistic induction does exist and is in use. Deductivists maintained that empirical facts cannot model our ideas, and they produced more "proofs" of the impossibility of induction. Even if mutual concessions and rephrasing have now brought inductivists and deductivists somewhat closer to each other, inductivism still emerges as the winner of the contest; inductivists outnumber deductivists. But, as inductive methods remain hard to "justify", Popper's critical examination of the topic of induction remains an interesting challenge.

References

[1] Popper, K.R. (1932). *Die beiden Grundprobleme der Erkenntnistheorie I: Das Induktions-problem; II: Das Abgrenzungsproblem*. Wien, dittoed; Mohr (Siebeck), Tobingen, 1979.

[2] Popper, K.R. (1934). *Logik der Forschung*. Springer-Verlag, Wien. English translation by Popper & Freed, *The Logic of Scientific Discovery*. Hutchinson, London, and Basic Books, New York, 1959; 9th Ed. Revised, 1977.

[3] Popper, K.R. (1944–1945). The Poverty of Historicism I, II, III, *Economics, New Series* **11**(42), 86–103; **11**(43), 119–137; **12**(46), 69–89;Revised Ed.: *The Poverty of Historicism*. Routledge & Kegan Paul, London, 1957.

[4] Popper, K.R. (1945). *The Open Society and Its Enemies*, Vol. 1, Plato, & Vol. 2, Hegel and Marx. Routledge & Kegan Paul, London, 5th Ed. Revised, 1966; Princeton Paperback, Princeton, 1971.

[5] Popper, K.R. (1963). *Conjectures and Refutations. The Growth of Scientific Knowledge*. Routledge & Kegan Paul, London; 4th Ed. Revised, 1972.

[6] Popper, K.R. (1972). *Objective Knowledge: An Evolutionary Approach*. Oxford University Press, London; Revised Ed., 1979.

[7] Popper, K.R. (1976). *Unended Quest*. Open Court, La Salle (autobiography).

[8] Popper, K.R. (1977). *The Self and Its Brain. An Argument for Interactionism*, with John C. Eccles. Springer-Verlag, New York.

[9] Popper, K.R. (1982). *Realism and the Aim of Science. From The Postscript to the Logic of Scientific Discovery*, Vol. I. W.W. Bartley III, ed. Hutchinson, London.

[10] Popper, K.R. (1982). *The Open Universe. An Argument for Indeterminism. From The Postscript to the Logic of Scientific Discovery*, Vol. II. Hutchinson, London.

[11] Popper, K.R. (1982). *Quantum Theory and the Schism in Physics. From the Postscript to the Logic of Scientific Discovery*, Vol. III. Hutchinson, London.

[12] Popper, K.R. (1990). *A World of Propensities*. Thoemmes, Bristol.

[13] Schilpp, A., ed. (1974). *The Philosophy of Karl Popper*. Open Court, La Salle; 2 vols (1323 pp.) (Popper's Autobiography, 33 Critical Essays, Popper's Replies to His Critics, Popper's Bibliography).

A. FAGOT-LARGEAULT

Population Genetics

Population genetics, as the name implies, is concerned with the analysis of factors affecting the genetic composition of a population. It is thus centrally concerned with evolutionary questions through the change in the genetic composition of a population over time as directed by natural selection, mutation, migration, and other factors, with questions associated with genetic diseases in human populations, and finally with plant and animal breeding programs, in which an attempt is made to change the genetic constitution of a population by artificial methods. To the extent that this genetic composition is in part random and can be described in quantitative terms, population genetics is a subject of prime interest to biometricians. Here we focus on biometrical aspects of evolutionary population genetics, although we will note the importance of several aspects of the evolutionary theory on human genetics and on plant and animal breeding programs.

Early Population Genetics Theory

It is a well-known curiosity that Darwin published his theory of evolution by natural selection many years before the (Mendelian) hereditary mechanism (*see* **Mendel's Laws**) was firmly established. One matter in particular troubled him: under the blending theory of heredity current during his time, which assumed that the characteristics of any offspring are a blend of the corresponding characteristics of the two parents, the variation upon which his theory of evolution by natural selection depended would soon be dissipated unless some strong variation-creating agency could be established causing offspring not to resemble their parents. Such an agency, however, would make inoperative another central component of the Darwinian theory, namely the resemblance between parents and offspring.

It was one of the early triumphs of population genetics theory, stressed repeatedly by **Fisher**, that the Mendelian hereditary system resolved this difficulty at a stroke. The main early result responsible for reaching this important breakthrough was the well-known **Hardy–Weinberg equilibrium** law. This law implies, for two alleles (or **gene** types) A_1 and A_2 at some gene locus "A", with no selection, mutation, or any other disturbing force, and if random changes in frequencies arising from **stochastic processes** in finite populations can be ignored, that if the **genotype** frequencies in any generation take the arbitrary values X (for A_1A_1), $2Y$ (for A_1A_2), and Z (for A_2A_2), then following one generation of random mating these frequencies will be of the form x^2, $2x(1-x)$ and $(1-x)^2$, where $x = X + Y$, and that under random mating these frequencies will be maintained in all future generations. This result follows essentially from the "quantal" nature of the gene.

Unfortunately this law is too often taught to students for the wrong reason: the **binomial** form of genotype frequencies which are derived under this law is certainly interesting and convenient, but what is crucial is not this but the unchanging nature of genotype frequencies, as indicated by the law, if no disturbing forces such as selection and mutation exist. The only significant area where the law does not come to grips with an important question concerns the fact that it does not handle the stochastic changes in gene frequencies implied by the finite nature of every population. This stochastic behavior is a matter we return to again later.

The quantal nature of the Mendelian scheme ensures that the natural tendency of a Mendelian population is to maintain genetic variation. Of course this variation will tend to be lost by the very process of replacing an "inferior" allele by a "superior" one, but this is an entirely different matter from a natural tendency of a hereditary system itself to destroy variation. It is thus a crucial concern to quantify the concept of variation in a Mendelian population, not only among genotypic frequencies but, more important to biometricians, among those measurable characteristics whose variation is preserved through the preservation of genetic variation.

The Correlation Between Relatives

In the early years of the century, biometricians noted various regular forms for the **correlation** of certain metrical traits such as height between various types of relatives. It was one of the early triumphs of population genetics theory, attributable to Fisher [3], to show that the form of these correlations could be accounted for by Mendelian considerations. Suppose, in the simplest case, that all individuals of genotypes A_1A_1 have measurement m_{11} for this character, all

individuals of genotype A_1A_2 have measurement m_{12}, and all individuals of genotype A_2A_2 have measurement m_{22}. Then it is easy to compute the mean $\overline{m}$ and the variance σ^2 of the measurement as

$$\overline{m} = m_{11}x^2 + 2m_{12}x(1-x) + m_{22}(1-x)^2$$

and

$$\sigma^2 = m_{11}^2x^2 + 2m_{12}^2x(1-x) + m_{22}^2(1-x)^2 - \overline{m}^2.$$

Further, since there is a one-to-one correspondence between measurement value and genotype, it is straightforward to write down all possible combinations for this measurement between any pair of relatives (mother–daughter, brother–brother, etc.), together with their probabilities, using as a key part of the argument the Mendelian rules for the transmission of genes from parent to offspring. From this it is straightforward to find the covariance, and thus also the correlation, between the two relatives for the measurement in question. When this is done, a remarkable series of formulas arises [3]. We find that all such correlations are of the form

$$\text{corr} = \frac{(\alpha\sigma_A^2 + \beta\sigma_D^2)}{\sigma^2},$$

where α and β are simple constants such as 0, 1/2, 1/4 (and more generally of the form $(1/2)^k$, for some integer k), and σ_A^2 and σ_D^2 are defined by

$$\sigma_A^2 = 2x(1-x)[xm_{11} + (1-2x)m_{12} - (1-x)m_{22}]^2$$

and

$$\sigma_D^2 = x^2(1-x)^2(m_{11} - 2m_{12} + m_{22})^2.$$

It is a matter of simple algebra to show that $\sigma^2 = \sigma_A^2 + \sigma_D^2$, so that the total variance in the measurement has been subdivided into two components which enter differently into the correlation between various forms of relatives. We see here the beginnings of the concept of the **analysis of variance**, to flower so greatly in Fisher's hands in the 1920s, and first appearing in the biometrical context of the correlation between relatives (*see* **Familial Correlations**).

The Additive and the Dominance Variances

The two components σ_A^2 and σ_D^2 have interpretations well beyond being simply components of the total variance in the measurement. Our later main focus is on σ_A^2, but we first briefly discuss σ_D^2. Clearly this component is always zero if $m_{12} = 1/2(m_{11} + m_{22})$, that is if the measurement for the heterozygote A_1A_2 is the average for that of the two homozygotes A_1A_1 and A_2A_2. When this occurs we say that there is no dominance in the measurement, and the component σ_D^2 of the total variance is nonzero only if dominance does exist. Thus σ_D^2 is called the dominance variance, and this explains the suffix in the notation.

We have just noted the occurrence of the so-called "additive genetic variance" σ_A^2 in the formulas for the correlation between various relatives. This variance is, however, of even greater value, and its significance is more clearly demonstrated, when we consider evolutionary aspects of a genetic population. To see this, suppose that the three genotypes A_1A_1, A_1A_2, and A_2A_2 have (viability) fitnesses w_{11}, w_{12}, and w_{22}, respectively, and that no other form of fitness differential (for example, fertility differentials) exists. Then the frequency x of the allele A_1 will, in general, change from generation to generation, and it is straightforward to find the formula for the change Δx in the frequency of A_1 from one generation to the next. If we define the mean fitness $\overline{w}$ as a special case of a mean measurement by the formula

$$\overline{w} = w_{11}x^2 + 2w_{12}x(1-x) + w_{22}(1-x)^2,$$

then it is also straightforward to find the change $\Delta\overline{w}$ in $\overline{w}$ from one generation to the next, and from this we find that, to a close approximation,

$$\Delta\overline{w} = \sigma_A^2.$$

The statement embodied by this formula, that the increase in mean fitness is approximately equal to the additive genetic variance, has been called Fisher's [4] "Fundamental Theorem of Natural Selection". More precisely, this statement is the "conventional wisdom" version of the Fundamental Theorem. It is, however, uncertain that this result was what Fisher claimed to have achieved, and for an alternative interpretation of Fisher's theorem, see [2].

Whatever the correct interpretation of the theorem might be, it quantifies in genetical terms the main tenet of the Darwinian theory, that for evolution by natural selection to occur, leading in some sense to an "improvement" in the population, measured here by an increase in mean fitness, there must be variation

in the population, here variation in fitnesses of the three genotypes.

However, the theorem shows that not all the variation in fitness is relevant to evolution, since only the additive part of the variance is involved in the above formula. To understand why this is so, we must first investigate more closely the interpretation of the additive genetic variance σ_A^2. The key points to note are, first, that at any given locus an individual passes on a gene to an offspring, not the two genes comprising his/her genotype at this locus, and second that evolution concerns the changes in gene frequency. It is thus necessary to define the concept of a "gene fitness" to understand the differential transmission of genes from parent to offspring as brought about by natural selection and the consequent change in mean fitness. (More correctly we should define an "allele fitness", and use "allele frequency" rather than "gene frequency" above, but we follow here the standard, albeit incorrect, usage.) We define "gene fitnesses" by noting that an A_1 gene combines with another A_1 gene, with probability x, to form an individual of fitness w_{11}, and with an A_2 gene, with probability $1-x$, to form an individual of fitness w_{12}. We thus ascribe a fitness $w_{11}x + w_{12}(1-x)$ to the gene A_1, and similarly we ascribe a fitness $w_{12}x + w_{22}(1-x)$ to the gene A_2. Since a randomly chosen gene is A_1 with probability x and A_2 with probability $1-x$, we can define a mean and a variance of gene fitness by using the probability distribution shown in Table 1.

The mean gene fitness is easily found, from this distribution, to be identical with the mean genotype fitness $\overline{w}$ given above, but the variance in gene fitness is found to be different from the variance in genotype fitness, being $x(1-x)[xw_{11} + (1-2x)w_{22}]^2$. Recalling that at any locus each individual has two genes in the genotype, we define the variance in the fitness of any individual due to the genes in his/her genotype to be twice this value, or (with the general measurement replaced by fitness), by the additive genetic variance in fitness.

If the additive genetic variance comprises all the variance in fitness, then the dominance variance is zero. This occurs, as noted above, if and only if the fitness of the heterozygote A_1A_2 is the average of the fitnesses of the two homozygotes. This in turn occurs if and only if genotype fitnesses are completely determined, in an additive way, by genes within genotypes. In this case it is not surprising that all the variance in genotype fitness, being determined solely by genes, is available for evolution.

Table 1

Gene (allele)	A_1	A_2
Gene fitness	$w_{11}x + w_{12}(1-x)$	$w_{12}x + w_{22}(1-x)$
Probability	x	$1-x$

If the additive genetic variance is zero, then either one or the other gene is absent from the population (a case we ignore from now on as trivial), or the two genes are equally "fit" (as determined by the definition of gene fitness just described). As a result there will be no change in gene frequencies through natural selection and the population is at an (internal) equilibrium point. This observation explains the occurrence of constant genetic variation in a population: if all the variance in fitness is dominance variance, then the population remains static in its genetic composition.

The above considerations also have considerable significance for artificial selection. The (narrow) **heritability** for any character is defined as the ratio of the additive genetic variance in a character to the total variance, and if this is zero, the two genes A_1 and A_2 have equal values for this character and no increase in the mean value of the character is possible through a change in gene frequencies.

All of the above shows that the concepts developed in early and elementary population genetics theory form the basis and origin of many general procedures in biostatistics. In particular, the concepts of subdividing a total variance into meaningful components, of seeking the relevance of each component, and of showing that various important quantities can be written in terms of these components, all find their origins in the simple arguments developed above.

Further Developments of Population Genetics Theory

The population genetics theory sketched above needs to be extended in many directions. For example, there may well be more than two alleles at the locus in question. The genes at different loci often interact (epistasis; *see* **Genotype**) in defining some character. Ecological considerations, such as geographical dispersal and the interactions of two or more species and their consequent mutually interactive genetic evolution, need to be examined. The stochastic effects which are inevitable in small populations are of

potential importance. While these extensions to the theory are the subject of much present-day population genetics, here we focus the discussion on only two of these extensions, namely multilocus theory and the (single locus) stochastic theory.

If (as is indeed the case) most genetic characteristics, and in particular those undergoing selection, are determined by the genes at many loci, then a proper analysis of genetic evolution must involve all such genes. The appropriate vehicle for doing this turns out to be the gamete, that is (in effect) a set of single chromosomes carrying the genes at various loci on them. Evolution then concerns changes in gametic frequencies. A description of gametic evolution is extremely complex and here we focus on the particular case where only two loci, A and B, are of interest, and where also each of these loci allows only two allelic types, namely A_1 and A_2 at the A locus and B_1 and B_2 at the B locus. We thus consider the evolution of the frequencies of the four gametes A_1B_1, A_1B_2, A_2B_1, and A_2B_2, which we denote by y_1, y_2, y_3, and y_4, respectively.

This evolution is particularly simple if knowledge of the gene at the A locus on any chromosome confers no information about the gene present at the B locus on the chromosome. It is not difficult to see that this occurs if and only if the equation $y_1y_4 = y_2y_3$ holds. However, this will not necessarily be the case and, when selection exists at the two loci, the so-called "coefficient of **linkage disequilibrium** D (more appropriately called the coefficient of association), defined by

$$D = y_1y_4 - y_2y_3,$$

will often not be zero.

This fact has several implications for biostatistics, of which we mention two here and one later when discussing human genetics. First, if all two-locus coefficients of association are zero, the overall additive genetic variance of the entire genome is simply the sum of the constituent single-locus additive genetic variance values. If these coefficients are not zero, then there is no simple relation between the overall additive genetic variance and the single-locus values. In the 1950s, when the concept of the analysis of variance was used systematically in population genetics, an attempt was made to subdivide the overall genome-wide additive genetic variance for some character into meaningful components, in particular components associated with single loci. Since the total variance does not usually divide into the sum of single locus components, this venture essentially failed. Secondly, when some character depends on the genes at many loci, the correlations between various relatives for that character are far more complex that those given by the single-locus formula above. Some simplification and indeed some explicit expressions for these correlations are possible when all possible two-locus coefficients of association are zero but in general, when these coefficients are nonzero, the correlations are hopelessly complicated, containing upwards of 100 terms even for a character determined by two loci. Thus the hope of a biostatistical investigation of realistic correlations, at least when coefficients of association are nonzero, seems to be doomed.

A classical calculation concerning D is the following. If there is no selection at either locus, no mutation or indeed any other directed force, and (most important) if the population consists of one large randomly mating group, then the value of D decreases geometrically each generation. More specifically, if D is the value of the coefficient of association in one generation and D' the value in the following generation, and if the recombination fraction between the two loci is θ, then

$$D' = (1 - \theta)D.$$

Thus, after t generations, $D^{(t)} = (1 - \theta)^t D^{(0)}$, so that, unless the two loci are closely linked, the value of D should rapidly decrease to zero. The inference that the occurrence of association between two loci implies that they are closely linked is made often in population and in **human genetics**. We discuss this inference further below.

It was very early recognized in population genetics theory that stochastic effects need to be considered for a complete picture of evolution to be attempted, since in a finite population one cannot avoid the chance events that lead to unequal transmissions of genes from parent to offspring. This inequality occurs for two reasons: first, a **heterozygous** parent A_1A_2 might, by chance alone, transmit the A_1 gene more (or less) frequently to his/her offspring than the A_2 gene, thus causing a (random, nonselected) change in the frequencies of A_1 and A_2 in the daughter generation. Secondly, by chance alone, individuals of the same genotype might transmit different numbers of genes to the following generation: as an extreme example, if one of a pair of identical twins is accidentally killed

in early life while the other survives to reproduce, their contributions to the next generation differ even though the two twins have identical genotype and hence identical fitness.

The fact that stochastic changes in gene frequency will arise was recognized very early on, and Fisher [4] and Wright [11] devised stochastic models allowing for these changes. The model at the center of their (similar but independent) work is now called the Wright–Fisher model. This is a Markov chain and represents perhaps the first major use of **Markov chains** in science (although it appears that neither Fisher nor Wright ever used the term "Markov chain"). The model is as follows. The **random variable** of interest is the number of A_1 genes in the population. The possible values of this number are $0, 1, \ldots, 2N$, where N is the population size. It is assumed in the model that if, in any generation, there are i such genes, then the number in the following generation has a binomial distribution with parameter $i/2N$ and index $2N$.

Clearly the states $i = 0$ and $i = 2N$ are absorbing, so interest focuses on the probability that a given absorbing state is entered and, from the point of view of the preservation of genetic variation, on the mean time until one or other absorbing state is entered. This then concerns a qualitatively different outcome from that in infinitely large populations, where the amount of genetic variation remains fixed. The rate at which this variation is lost can be assessed in two ways. First, one can find the leading nonunit **eigenvalue** of the matrix of transition probabilities of the Markov chain. For the Wright–Fisher model, this is easily seen to be $1 - 1/(2N)$. This suggests that, in populations of large size, genetic variation tends to be lost, on average, very slowly. This was the approach of Fisher and Wright. Secondly, a more direct assessment can be made by calculating the mean time until the population consists entirely of A_1, or entirely of A_2, genes, given some initial frequencies of A_1 and A_2. Unfortunately, this mean time cannot in practice be calculated exactly in the Wright–Fisher model. However, both Fisher and Wright were able to approximate the behavior of a population following the Wright–Fisher model by a diffusion process, for which these mean times are easily calculated. These show that the mean time until one or other gene is lost from the population is proportional to the population size. This is thus large in a large population, and this agrees with the conclusion reached as a result of the eigenvalue calculation. It is also possible to show that these mean times differ only trivially from the (unknown) values in the Wright–Fisher model. It is interesting that both Fisher and Wright used (again without knowing it) forward Kolmogorov equation methods to establish the results they needed, and never became aware of the far simpler and more appropriate methods available through use of the Kolmogorov backward equation.

Many extensions and variations to the Markov chain model have appeared in the population genetics literature over the past 40 years. The number of alleles can be arbitrary, selection can exist, and more realism has been introduced by noting that the population includes both males and females, has a geographical structure, can change in size over time, and so on. Similarly, very extensive generalizations to the diffusion process used by Fisher and Wright have been made, again incorporating more realistic features such as those described for the Markov chain model (*see* **Brownian Motion and Diffusion Processes**).

Fisher, Wright, and Haldane [5] also developed theory for, and used, **branching processes** to discuss the probability of survival of a new mutant gene. Once again, the introduction and first significant applications of what would now be viewed as standard equipment in any biometrician's toolkit was carried out in the early research in population genetics theory.

Retrospective Theory

The analyses of Markov chains, diffusion processes and branching processes by Fisher, Wright, and Haldane, and the generations of population geneticists following them (in particular Kimura [7]), are part of the *prospective* theory of population genetics: given certain selective values, mutation rates, population sizes, and so on, statements were made about the likely evolution forward in time of a population under the Mendelian hereditary mechanism. This sort of analysis was indeed needed in order to show that, contrary to the views held even in many scientific circles early in the century, the Darwinian paradigm provided a valid and workable theory for evolution. Two factors have, however, changed the broad direction of population genetics theory in recent years. First, for all practical purposes and for all reasonable

people, the Darwinian theory is indeed validated, so that there is no longer any need for further validating arguments. Secondly, the nature of the genetic material is now known. Early population genetics theory in effect conceptualized different alleles (i.e. different types of genes) as differently colored billiard balls, with the A_1 allele corresponding, say, to a red billiard ball, an A_2 allele to a blue ball, and so on. No information was available about the interior constitution of the billiard balls, and the coloring allocation (i.e. the labeling as A_1, A_2, and so on) was entirely arbitrary. All that was important was the discrete "quantal" nature of the color of the billiard ball (red, blue, etc., but never a mixture), and, apart from rare mutations, the faithful transmission by a parent of the color, unmixed with any other color, to the genetic make-up of his/her offspring.

The knowledge of the structure of genes as **DNA sequences** has completely altered population genetics, since now the description of a gene is no longer simply an arbitrary label (such as A_1 or A_2) but the actual DNA material of which the gene is made. In other words, real rather than arbitrary descriptions for genes are now available. This has led, among other things, to a blossoming of the *retrospective* theory of population genetics, in which a sample of genes is taken, their DNA examined, and the questions asked relate to the way in which, through evolution, the population arrived at its presently observed state. The broad form of population genetics theory needed to answer such questions is clearly closely allied to statistics, in that data deriving from both a stochastic process and also a sampling procedure are available and an inference deriving from those data is wanted.

Retrospective questions have entered the popular imagination: When did "Eve" live? Where did she live? What do we even mean by "Eve"? What inferences can we draw, given DNA information from a number of present-day species, about the tree of evolution leading to these species? These questions can be answered only by assuming various stochastic process population genetic models, none of which can be claimed to describe with close fidelity the actual path that evolution happens to have taken. Unfortunately, the estimates and inferences drawn are often sensitive to the assumptions made in those models, and this has led to much acrimony on the answers to the kind of question raised above.

To population geneticists, perhaps the most interesting and controversial question in retrospective population genetics concerns the so-called neutral theory [8]. This theory claims that a very large proportion of the DNA variation that we see within populations, and also much of the DNA difference that we observe between species, was not directed by natural selection but arose through purely random stochastic effects. The analysis of this theory thus requires as a starting point information concerning the properties of the Wright–Fisher and other stochastic prospective evolutionary models when no selection acts, followed by a further analysis of what properties we might expect to see in samples of genes if the theory is true. Substantial work in population genetics theory, which is in effect no more than a statistical analysis of present-day data testing certain evolutionary hypotheses, has been carried out to assess the acceptability of this theory.

An important vehicle for essentially all retrospective intrapopulation genetic inferences is the *coalescent process* of Kingman [9]. Given a sample of n genes, the coalescent traces their ancestry back in time. At some point two genes will have a common ancestor and a coalescence may be said to have occurred. Eventually all n genes will have a common ancestor and the final coalescence will have occurred. The properties of this coalescent process often provide, by moving forward in time, the simplest way of deriving the probability distributions needed for inferences from those data. An extended discussion of this process, focusing on biological questions, is given in [6], and extended mathematical and statistical aspects are discussed in [1] and [10].

Human Genetics

Several aspects of human genetics research, which focuses largely on the elucidation of the genetic basis of diseases, have parallels with, or even derive directly from, evolutionary population genetics theory.

First, there is a clear parallel with parts of the retrospective theory of population genetics and human genetics. In both cases we start with a sample of genes and attempt to make some inference from them. This is particularly true of questions concerning the population genetics concept of allelic association: if we see an association between a disease

and a marker locus (*see* **Disease-marker Association**), then one is tempted to infer that the disease and the marker loci are closely linked. This argument is supported by the fact, noted above, that in a randomly mating population the degree of allelic association decreases geometrically fast over succeeding generations. However, the assumption that the population from which the sample of genes is drawn is mating at random is crucial for this conclusion, and in practice, for human populations, this assumption is far from the truth. Simple geographical considerations imply nonrandom mating, as do the existence of racial groups. Thus if a certain disease is prevalent in some country and a certain unlinked marker gene is also prevalent there and if a sample of individuals is taken in the USA in which a significant proportion of individuals has recent ancestry from that country, then an association between the disease and the marker gene will be observed. However, this association has nothing to do with linkage between disease and marker loci. Approaches to this problem, using both population genetics theory focusing on geographical features and human genetics theory in which this form of association is separated from association due to *linkage*, have both been undertaken. This is but one example of the interwining of human genetics, one of the central areas of biometrical research, and evolutionary population genetics during the next decade will lead to an increasing linkage between evolutionary and human population genetics, and hence to an increasing use of biometrical methods in both areas.

References

[1] Donnelly, P.J. & Tavaré, S. (1995). Coalescents and the genealogical structure under neutrality, *Annual Review of Genetics* **29**, 542–551.

[2] Ewens, W.J. (1989). An interpretation and proof of the fundamental theorem of natural selection, *Theoretical Population Biology* **36**, 167–180.

[3] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[4] Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

[5] Haldane, J.B.S. (1930). *The Causes of Evolution*. Longmans Green, London.

[6] Hudson, R.R. (1991). Gene genealogies and the coalescent process, in *Oxford Surveys in Evolutionary Theory*, Vol. 7, D. Futuyma & J. Antonovics, eds. Oxford University Press, Oxford.

[7] Kimura, M. (1957). Some problems of stochastic processes in genetics, *Annals of Mathematical Statistics* **28**, 882–901.

[8] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

[9] Kingman, J.F.C. (1982). The coalescent, *Stochastic Processes and Their Applications* **13**, 235–248.

[10] Tavaré, S. (1992). Calibrating the clock: using stochastic processes to measure the rate of evolution, in *Calculating the Secret of Life*, E.S. Lander & M.S. Waterman, eds. National Academy Press, Washington.

[11] Wright, S. (1931). Evolution in Mendelian populations, *Genetics* **16**, 97–159.

(*See also* **Genetic Correlations and Covariances**)

W.J. EWENS

Population Growth Models

Forecasting the growth of human populations has, for at least two centuries, been a problem that has taxed many scholars. The methods proposed have included mathematical models, sometimes including the search for an elusive "law" of human population growth; statistical forecasts; and component projections. These methods have addressed geographical areas ranging in size from local areas with very small populations to the entire world. This article summarizes the major features of each of the three types of method and highlights the purposes to which they are put.

The first person to propose a *mathematical growth model* was probably **Thomas Malthus**. In his 1798 essay he observed that "Population, when unchecked increases in a geometrical ratio" (Malthus, [16]). This can be expressed by geometric or exponential growth models:

$$P_t = P_0(1 + r)^t.$$

or

$$P_t = P_0 \mathrm{e}^{r\ t},$$

where P_0 is population at time 0, P_t is population at time t, r is the rate of growth, and t is time.

These models are still widely used and often are reasonably accurate for relatively small t. However, in the long term it is clear that populations cannot grow exponentially. This was observed first by **Quetelet** [24] who, inspired by Newton's law of viscosity, outlined a theory by which, as populations grow, they are restricted in size by increasing density. This idea was developed by Verhulst [27], who concluded that growth was a function of size and that an S-shaped curve was appropriate. He termed this the "logistic" curve (*see* **Logistic Distribution**). Much later, **Pearl** & Reed [18] made the logistic curve popular for the long-term prediction of population growth, Pearl attempting to provide empirical evidence for the mathematics through the famous fruit fly experiments. This model can be expressed as

$$P_t = \frac{a}{1 + \exp(-abt)},$$

where a and b are estimated coefficients and it can be shown that a represents the upper asymptote, t is time, and P_t is the population at time t.

The logistic model was very popular in the 1930s when slowing rates of population growth were common in most countries where reliable population data were collected. After World War II the model was not used a great deal because it was shown to be very inaccurate in a number of empirical situations. For example, Yule [28] predicted a population of England and Wales in 1971 of 58–59 million compared with the actual figure of 49 million. Brass [5] notes that the major problem was that "resource based" constraints on population growth had not continued to slow in the 1950s. However, improved techniques to estimate the model and changing demographic patterns may mean that, for short-term forecasts, the logistic model may be a possibility in the early part of the twenty-first century.

The mathematics of these simple models was developed by Sharpe & Lotka [25] who described a simple one-sex deterministic population model (one-sex because Sharpe & Lotka [25] calculate the population of men and simply assume that there are enough women to maintain population growth). Pollard [20] provides an excellent review of this model and sensibly develops it in terms of women. Using this model, Keyfitz [9] developed a formula to calculate population momentum. Although the model has not been widely used in recent years, the observations by Bongaarts & Amin [3] that population momentum will be an increasingly important factor in the growth of populations in less developed countries mean that it may gain a new prominence.

In the immediate post World War II period the development of **stochastic processes** for applications such as particle physics or telephone exchange problems led to the possibility of developing simple stochastic models for the growth of human populations. The development of these models was reviewed by Kendall [8] and more recently by Alho [1].

They typically start with a simple set of equations to determine both births and deaths and base these on a **Poisson process**. Kendall [8] was able, in addition, to include migration. These models have not been widely used despite a number of attempts to generalize them [7, 19] to take account of the dynamics between the sexes [7, 20]. The lack of success of these latter attempts stems largely from the

necessity to assume that one sex is dominant. This means that if the sex ratio diverges greatly from unity, then one sex will soon become extinct. In addition, to be mathematically tractable, the two-sex model has been dependent on the linear model. It is possible that improved numerical methods will permit a nonlinear exposition, but it seems unlikely that they will ever be particularly popular for scholars working on human populations. Pollard [20] reviews the early work, and recently Pollard & Höhn [23] revisited the problem and provided a number of extensions. A comprehensive review is provided by Pollard [22].

The development of *statistical models* for **time series** and their popularity, particularly in econometrics, led to their use to forecast population growth. To forecast total population size, Lee [11, 12] was at the forefront of adapting econometric models, particularly those of Box & Jenkins [4] (*see* **ARMA and ARIMA Models**) to predict population growth. Thorough reviews are provided by Land [10], Lee & Tuljapurkar [13] and Alho [2]. The advantage of these models is, of course, that they were developed partly for **prediction** and so interval estimates of future population size are possible. However, their use has not been widespread probably because they suffer, as do most of the mathematical models, from the problem that population growth depends on the dynamics between births, deaths, and **migration**. Therefore, most commentators would agree that it is important that population growth models take this into account.

This need to control explicitly for fertility, mortality and migration has led to *component methods* of population projection being, without doubt, the most commonly used method to estimate future population size. Component methods are based on the following simple balancing equation:

$$P_t = P_0 + {}_tB_0 - {}_tD_0 + {}_tI_0 - {}_tE_0,$$

where P_t is population at time t, and P_0 is population at time 0, ${}_tB_0$ is births between times 0 and t, ${}_tD_0$ is deaths between times 0 and t, ${}_tI_0$ is immigrants between time 0 and t, and ${}_tE_0$ is emigrants between time 0 and t.

Starting with a base population, P_0, these models account separately for trends in each of the components of population growth to make a projection of future population size. In themselves, projections are just internally consistent forecasts dependent on a set of assumptions. Methodologists always caution that a range of projections should be used but, in practice, planners and policy makers usually believe one. The mathematics of component methods were developed by Leslie [14] and the *"Leslie" matrix* is the basis for most computational models of component projections. There are many computer packages to make component projections.

The key elements in any component projections are the models used to predict future mortality, fertility, and migration. To describe these fully is outside the scope of this article but it should be noted that there are a myriad of strategies ranging from the simple (and almost always false) assumption that the three components will remain constant to very sophisticated mathematical and statistical models of future trends. Mathematical models have been proposed by a number of authors, most notably Brass [5] who developed a relational model to predict future trends in fertility and mortality. Among many authors, Murphy [17] and Pollard [21] have developed models for mortality, and there are many time series models to predict fertility, although their success has varied. Migration has not been modeled well in most countries particularly at a subnational level.

A particularly notable set of forecasts for the world was provided by Lutz [15], and Alho [2] develops an excellent strategy to estimate the uncertainty in these forecasts. In general, component models of population growth are reasonable at a national level in the short term, but difficulties, particularly in modeling future fertility, mean there are many examples of extremely inaccurate forecasts in the medium to long term.

For small areas, typically populations below 20 000, component models of population growth are used sometimes. More common, though, are simple models based on expected numbers of persons per household or on **census** data updated by ancillary information from, say, an electoral roll. In addition, **regression** models have been used a little; for example, by Erickson [6]. These models can be reasonably accurate for estimating the total population size where there is little population change. However, Simpson et al. [26] show that when age-specific estimates are required such models provide inaccurate forecasts and there is little alternative to a local census.

In summary, there have been many attempts to model population growth. If the aim is to forecast future population, then in the short term there have been a number of successful models, but in the longer term population growth models have been

less successful. However, they have been very useful in understanding the dynamics of change in human populations.

References

[1] Alho, J. (1990). Stochastic models in population forecasting, *International Journal of Forecasting* **6**, 521–530.

[2] Alho, J. (1997). Scenarios, uncertainty and conditional forecasts of the world population, *Journal of the Royal Statistical Society, Series A* **160**, 71–86.

[3] Bongaarts, J. & Amin, S. (1996). *The Prospects for Future Population Growth in South Asia*. IUSSP Seminar on Fertility in South Asia, Islamabad, December 16–19, 1996.

[4] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis*, Revised Ed. Holden-Day, San Francisco.

[5] Brass, W. (1974). Perspectives in population prediction, *Journal of the Royal Statistical Society, Series A* **137**, 532–583.

[6] Erickson, E.P. (1974). A regression model for estimates of population changes of local areas, *Journal of the American Statistical Association* **79**, 867–875.

[7] Goodman, L.A. (1953). Population growth of the sexes, *Biometrics* **9**, 212–225.

[8] Kendall, D.G. (1949). Stochastic processes and population growth, *Journal of the Royal Statistical Society, Series B* **11**, 203–264.

[9] Keyfitz, N. (1971). On the momentum of population growth, *Demography* **8**, 71–80.

[10] Land, K. (1986). Methods for national population forecasts: a review, *Journal of the American Statistical Association* **81**, 888–901.

[11] Lee, R.D. (1974). National fertility, population cycles and the spectral analysis of births and marriages, *Journal of the American Statistical Association* **69**, 607–617.

[12] Lee, R.D. (1979). Time series models of population growth, in *Prospect of Population: Methodology and Assumptions*. United Nations, New York.

[13] Lee, R.D. & Tuljapurkar, S. (1994). Stochastic population forecasts for the US: beyond the high, medium and low, *Journal of the American Statistical Association* **89**, 1175–1189.

[14] Leslie, P.H. (1945). On the use of matrices in certain population mathematics, *Biometrika* **33**, 183–212.

[15] Lutz, W. ed. (1994). *The Future Population of the World*. Earthscan, London.

[16] Malthus, T.R. (1798). *An Essay on the Principle of Population*. Printed for J. Johnson in St Paul's Churchyard, London, Chapter 2.

[17] Murphy, M. (1995). The Prospect of Mortality: England and Wales and the United States of America, 1962–1989, *British Actuarial Journal* **1**, 331–350.

[18] Pearl, R. & Reed, L.J. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation, *Proceedings of the National Academy of Sciences* **6**, 275–288.

[19] Pollard, A.H. (1948). The measurement of reproductivity, *Journal of the Institute of Actuaries* **74**, 288–318.

[20] Pollard, J.H. (1973). *Mathematical Models for the Growth of Human Populations*. Cambridge University Press, Cambridge.

[21] Pollard, J.H. (1996). On the changing shape of the Australian mortality curve, *Health Transition Review, Supplement* **6**, 283–300.

[22] Pollard, J.H. (1997). Modelling the Interaction between the sexes, *Mathematical and Computer Modelling*, to appear.

[23] Pollard, J.H. & Höhn, C. (1993). The interaction between the sexes, *Zeitschrift für Berolkerungswissenschaft* **19**, 203–228.

[24] Quetelet, A. (1835). *Essai de Physique Sociale*, Vol. 1. Paris.

[25] Sharpe, F.R. & Lotka, A.J. (1911). A problem in age distribution, *Philosophical Magazine* **21**, 435–438.

[26] Simpson, S., Diamond, I., Tonkin, P. & Tye, R. (1996). Updating small area estimates in England and Wales, *Journal of the Royal Statistical Society, Series A* **139**, 235–247.

[27] Verhulst, P.-J. (1845). Recherches mathèmatiques sur la Loi de Accroisement de la Population, *Nouveaux Mémoires de l'Academie Royale des Sciences et Belles Lettres (Brussels)* **18**, 3–41.

[28] Yule, G.U. (1924). A mathematical theory of evolution based on the conclusions of Dr J.C. Willis, F.R.S., *Philosophical Transactions of the Royal Society, Series B* **213**, 21–87.

(*See also* **Demography**)

IAN DIAMOND

Population-based Study

A population-based study is a study of properties of a well-defined population, such as individuals residing in a defined geographic region in a given time period. The size of such a population can be estimated, and, if all cases of a disease arising from such a population are identified, **rates** of disease can be calculated. Valid **sampling frames** can be constructed for estimating the **prevalence** of risk factors and other characteristics of such a population. **Population-based case–control studies** yield not only estimates of **relative risk** for given exposures but also estimates of exposure-specific **absolute risk**. The latter are obtained by combining information on the overall **risk** of disease in the population with information on the prevalences and relative risks of various exposure levels.

MITCHELL H. GAIL

Postmarketing Surveillance of New Drugs and Assessment of Risk

The assessment of the safety of newly marketed drugs (either prescription or over-the-counter products) is recognized as a fundamental public health responsibility in every developed region of the world. Efforts are under way to harmonize international standards for surveillance and reporting of adverse drug events to regulatory bodies of the US, the European Union (EU), and Japan [3]. The goal of this article is to review the role, rationale, objectives, and design of quantitative methods of surveillance strategies developed for drug safety assessment over the past 30 years. We emphasize, with several examples, the quantitative methods that have been developed in response to regulatory requirements to monitor adverse drug event data for the purpose of assessing potential risk (*see* **Drug Approval and Regulation**; **Pharmacoepidemiology, Adverse and Beneficial Effects**).

The Role of a Monitoring System

Finney [6], in a seminal paper which focused on monitoring for adverse reactions to drugs used in medical practice and on the methods for detecting drug associated reactions, stated that "the primary duty of a drug monitoring system is less to demonstrate danger or to estimate incidence than to initiate suspicions". The role of a monitoring system is to initiate a "suspicion" and not to establish cause and effect relationships or to estimate risk. When "suspicion builds", it is the task of more formal experimental and epidemiologic studies to confirm the suspicion. Generally, quantitative surveillance methods will have a hypothesis generating goal.

Types of Monitoring Systems

Finney [7] described the major components of a "monitoring system" as consisting of the reporters, the patients, the drugs, and the events, stating that there were three ways by which drug–events were reported to a system: by patient, by drug, and by event. In a related paper, Finney [6] expanded the discussion of these three methods of ascertainment by showing how these different methods impacted the statistical methodologies used either to signal or to evaluate causality of a drug–event association (*see* **Causation**) or to assess the **risk** of a drug-related event.

Finney stressed the need to identify a "reference population" as the source of the data coming into the monitoring system and to understand the underlying assumptions related to the data being reported (e.g. that events are independent and are representative of some underlying population).

The pioneering study by Inman et al. [12], in their investigation of thromboembolic disease and the steroidal content of oral contraceptives, illustrates the application of these principles to a monitoring system. An example of event ascertainment by patient (i.e. population based drug monitoring) was initiated at the Kaiser-Permanente, Department of Medical Research, in the early 1970s [8]. A unique feature of this **population-based** approach was its ability to screen over 75 000 patients for potential drug–event **associations**. Any combination of a drug and an event recorded within the system could be investigated. Once a signal occurred, the evaluation of the association for causality could be explored using traditional epidemiologic approaches (e.g. strength of association, level of statistical significance to rule out chance, consistency of the finding, specificity of the event, time relationships, biologic plausibility, and gradients, etc. [9]; *see* **Hill's Criteria for Causality**). More recent examples of population-based monitoring include the Boston Collaborative Drug Surveillance Program, Group Health Cooperative of Puget Sound, various Medicaid databases, and the Saskatchewan database [20].

Perhaps the largest and most systematic drug–event monitoring systems are those set up by governmental regulatory bodies to monitor safety of new drugs once marketed (e.g. in the US and UK [3, 23]). In the US, postmarketing surveillance systems require that manufacturers collect and send to a central point reports for which there is a suspicion that a drug and an event are associated. What is not required in the US is that the suspicion be established as a causality [13].

The particular quantitative surveillance methods for comparing drug–event reporting rates depend on a variety of issues like the method of event

ascertainment, the choice of **control** groups, the specific design of the database, and the various sources of **bias** (underreporting, patient characteristics associated with differential drug prescribing, treatment patterns, and concomitant drug treatment). Therefore, as we intend to illustrate several surveillance methods developed to signal and to alert within the reporting system available in the US, we describe the system in the US in some detail.

The Drug–Event Monitoring System at the Food and Drug Administration (FDA)

The **Food and Drug Administration (FDA)** Center for Drug Evaluation and Research has had a drug–event-based monitoring system since the late 1960s. In the US, the manufacturer and holders of new drug applications (NDA)s are required to report those adverse drug experiences (ADEs) associated with their products which come to their attention. The specific reporting requirements are described in federal regulations [25]. It is worth noting that some specific requirements for reporting may be modified pending publication of final rules and the implementations of certain recommendations of the International Conference on Harmonization (ICH).

In general, the reporting requirements for manufacturers creates a **stratification** for the timing of when some reports are sent to and received by FDA. These requirements involve distinguishing between "serious and nonserious" adverse events and between labeled and unlabeled adverse drug events (ADEs), each term being defined in regulations. Serious unlabeled ADEs must be submitted by the reporter within 15 days of learning of the event, whereas nonserious ADEs are submitted periodically in less frequent intervals of time, like quarterly for the first 3 years post approval and then every 6 months thereafter. The underlying philosophy is to obtain reports from health practitioners as soon as possible to identify new adverse events associated with exposure to new drugs. Where it is possible, an additional goal is to characterize the unique patient features which make patients susceptible to the adverse event and then to change the drug label accordingly to inform patients and prescribers about how to minimize the occurrence of these events.

The basic instrument for collecting the information on possible drug–event associations is called the Medwatch form, which includes demographic information about the patient, a description of the reaction(s) that occurred, an outcome reflecting what happened to the patient, an identification of the suspected drug and concomitant drugs and other history (see Figure 1).

Figure 2 shows the number of reports of labeled and unlabeled and serious and nonserious ADEs received for all marketed drugs in the United States entered into a computer database at FDA over the past 25 years. From Figure 2, it can be seen that there has been a continuous increase in the number of drug-associated adverse events reported to FDA. The histogram gives the number of reports entered into FDAs computer file. Starting in 1984, FDA distinguished between a manufacturer's periodic (labeled) and 15-day (serious and unlabeled) reports, and direct voluntary reports to FDA from health care providers.

Some Applications of Quantitative Methods

In this Section we describe four different situations for which specific statistical methods have been applied to postmarketing surveillance of adverse drug reaction reports. In the first three applications, we assume a centralized report registry like that of the FDA and that reporters follow similar instructions on what to report, use a standardized report form, and report either voluntarily or in accordance with some national program of reporting.

We distinguish between three situations: (i) monitoring a change in the reporting pattern of a specific drug–event; (ii) monitoring the comparative reporting of a drug–event for several drugs in the same class; and (iii) monitoring the comparative reporting of drug–events reported from multiple sources. The focus in each situation is on identifying an increased frequency of ADEs.

Monitoring a Change in a Specific Drug–Event Reporting Pattern

FDAs guideline [3] describes an "arithmetic" and a "statistical" approach to signal generation. The arithmetic approach is deterministic and is not discussed further. The statistical approach is based on the concept of a reporting rate where the numerator consists

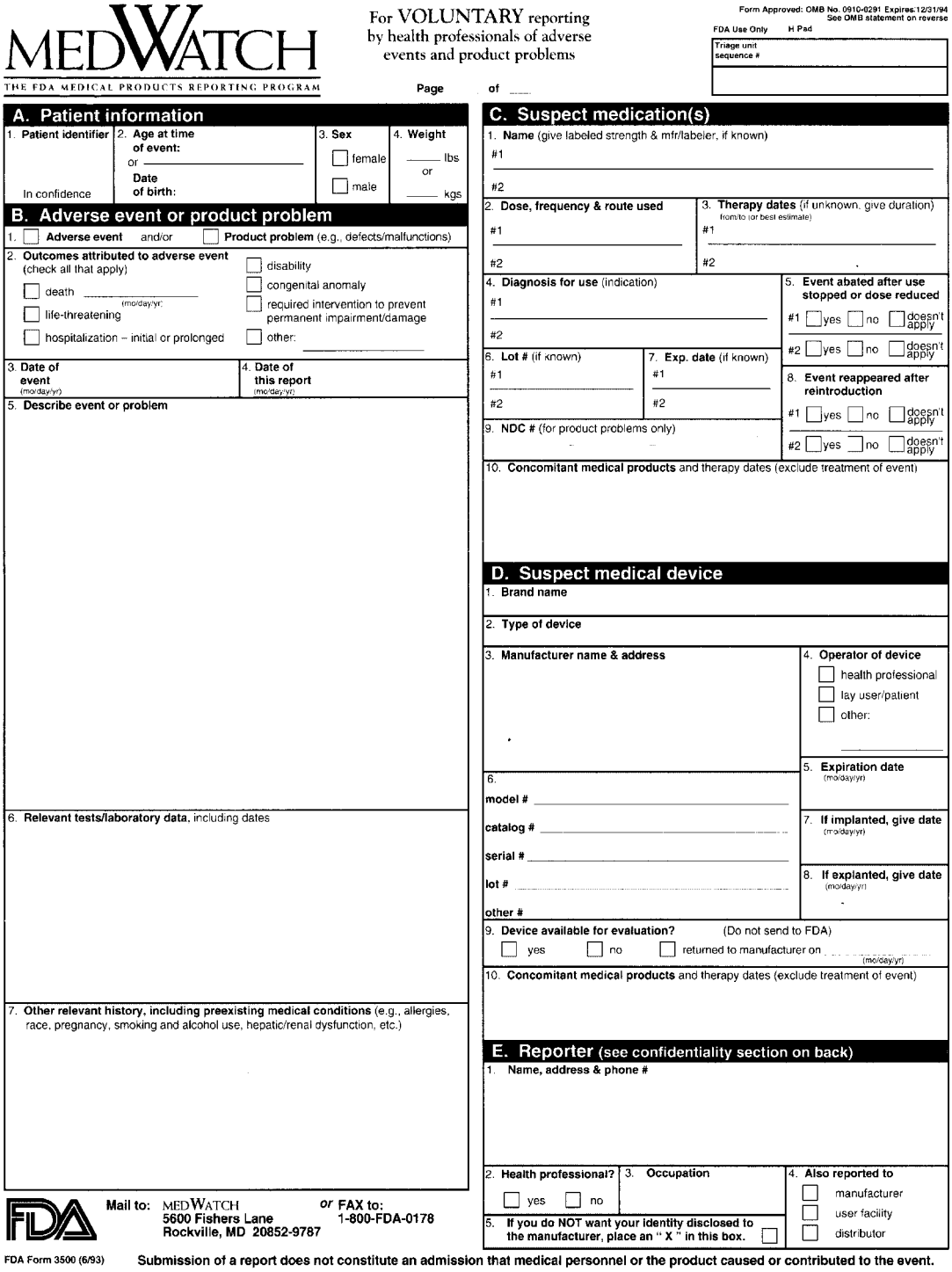

MEDWATCH
THE FDA MEDICAL PRODUCTS REPORTING PROGRAM

For VOLUNTARY reporting by health professionals of adverse events and product problems

Form Approved: OMB No. 0910-0291 Expires:12/31/94
See OMB statement on reverse

FDA Use Only H Pad

Triage unit sequence #

Page of

A. Patient information

1. Patient identifier In confidence
2. Age at time of event: or Date of birth:
3. Sex ☐ female ☐ male
4. Weight ____ lbs or ____ kgs

B. Adverse event or product problem

1. ☐ Adverse event and/or ☐ Product problem (e.g., defects/malfunctions)
2. Outcomes attributed to adverse event (check all that apply)
☐ death ____ (mo/day/yr) ☐ life-threatening ☐ hospitalization – initial or prolonged ☐ disability ☐ congenital anomaly ☐ required intervention to prevent permanent impairment/damage ☐ other: ____
3. Date of event (mo/day/yr)
4. Date of this report (mo/day/yr)
5. Describe event or problem
6. Relevant tests/laboratory data, including dates
7. Other relevant history, including preexisting medical conditions (e.g., allergies, race, pregnancy, smoking and alcohol use, hepatic/renal dysfunction, etc.)

C. Suspect medication(s)

1. Name (give labeled strength & mfr/labeler, if known) #1 #2
2. Dose, frequency & route used #1 #2
3. Therapy dates (if unknown, give duration) from/to (or best estimate) #1 #2
4. Diagnosis for use (indication) #1 #2
5. Event abated after use stopped or dose reduced #1 ☐ yes ☐ no ☐ doesn't apply #2 ☐ yes ☐ no ☐ doesn't apply
6. Lot # (if known) #1 #2
7. Exp. date (if known) #1 #2
8. Event reappeared after reintroduction #1 ☐ yes ☐ no ☐ doesn't apply #2 ☐ yes ☐ no ☐ doesn't apply
9. NDC # (for product problems only)
10. Concomitant medical products and therapy dates (exclude treatment of event)

D. Suspect medical device

1. Brand name
2. Type of device
3. Manufacturer name & address
4. Operator of device ☐ health professional ☐ lay user/patient ☐ other: ____
5. Expiration date (mo/day/yr)
6. model # catalog # serial # lot # other #
7. If implanted, give date (mo/day/yr)
8. If explanted, give date (mo/day/yr)
9. Device available for evaluation? (Do not send to FDA) ☐ yes ☐ no ☐ returned to manufacturer on ____ (mo/day/yr)
10. Concomitant medical products and therapy dates (exclude treatment of event)

E. Reporter (see confidentiality section on back)

1. Name, address & phone #
2. Health professional? ☐ yes ☐ no
3. Occupation
4. Also reported to ☐ manufacturer ☐ user facility ☐ distributor
5. If you do NOT want your identity disclosed to the manufacturer, place an " X " in this box. ☐

FDA

Mail to: MEDWATCH 5600 Fishers Lane Rockville, MD 20852-9787 or FAX to: 1-800-FDA-0178

FDA Form 3500 (6/93)

Submission of a report does not constitute an admission that medical personnel or the product caused or contributed to the event.

Figure 1 Medwatch form

ADVICE ABOUT VOLUNTARY REPORTING

Report experiences with:

- medications (drugs or biologics)
- medical devices (including in-vitro diagnostics)
- special nutritional products (dietary supplements, medical foods, infant formulas)
- other products regulated by FDA

Report SERIOUS adverse events. An event is serious when the patient outcome is:

- death
- life-threatening (real risk of dying)
- hospitalization (initial or prolonged)
- disability (significant, persistent or permanent)
- congenital anomaly
- required intervention to prevent permanent impairment or damage

Report even if:

- you're not certain the product caused the event
- you don't have all the details

Report product problems – quality, performance or safety concerns such as:

- suspected contamination
- questionable stability
- defective components
- poor packaging or labeling

How to report:

- just fill in the sections that apply to your report
- use section C for all products except medical devices
- attach additional blank pages if needed
- use a separate form for each patient
- report either to FDA or the manufacturer (or both)

Important numbers:

- 1-800-FDA-0178 to FAX report
- 1-800-FDA-7737 to report by modem
- 1-800-FDA-1088 for more information or to report quality problems
- 1-800-822-7967 for a VAERS form for vaccines

If your report involves a serious adverse event with a device and it occurred in a facility outside a doctor's office, that facility may be legally required to report to FDA and/or the manufacturer. Please notify the person in that facility who would handle such reporting.

Confidentiality: The patient's identity is held in strict confidence by FDA and protected to the fullest extent of the law. The reporter's identity may be shared with the manufacturer unless requested otherwise. However, FDA will not disclose the reporter's identity in response to a request from the public, pursuant to the Freedom of Information Act.

The public reporting burden for this collection of information has been estimated to average 30 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send your comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to:

Reports Clearance Officer, PHS
Hubert H. Humphrey Building,
Room 721-B
200 Independence Avenue, S.W.
Washington, DC 20201
ATTN: PRA

and to:
Office of Management and Budget
Paperwork Reduction Project
(0910-0230)
Washington, DC 20503

Please do NOT return this form to either of these addresses.

FDA Form 3500-back **Please Use Address Provided Below – Just Fold In Thirds, Tape and Mail**

Department of Health and Human Services
Public Health Service
Food and Drug Administration
Rockville, MD 20857

Official Business
Penalty for Private Use $300

NO POSTAGE NECESSARY IF MAILED IN THE UNITED STATES OR APO/FPO

BUSINESS REPLY MAIL
FIRST CLASS MAIL PERMIT NO. 946 ROCKVILLE, MD

POSTAGE WILL BE PAID BY FOOD AND DRUG ADMINISTRATION

MEDWATCH
The FDA Medical Products Reporting Program
Food and Drug Administration
5600 Fishers Lane
Rockville, MD 20852-9787

Figure 1 (*Continued*)

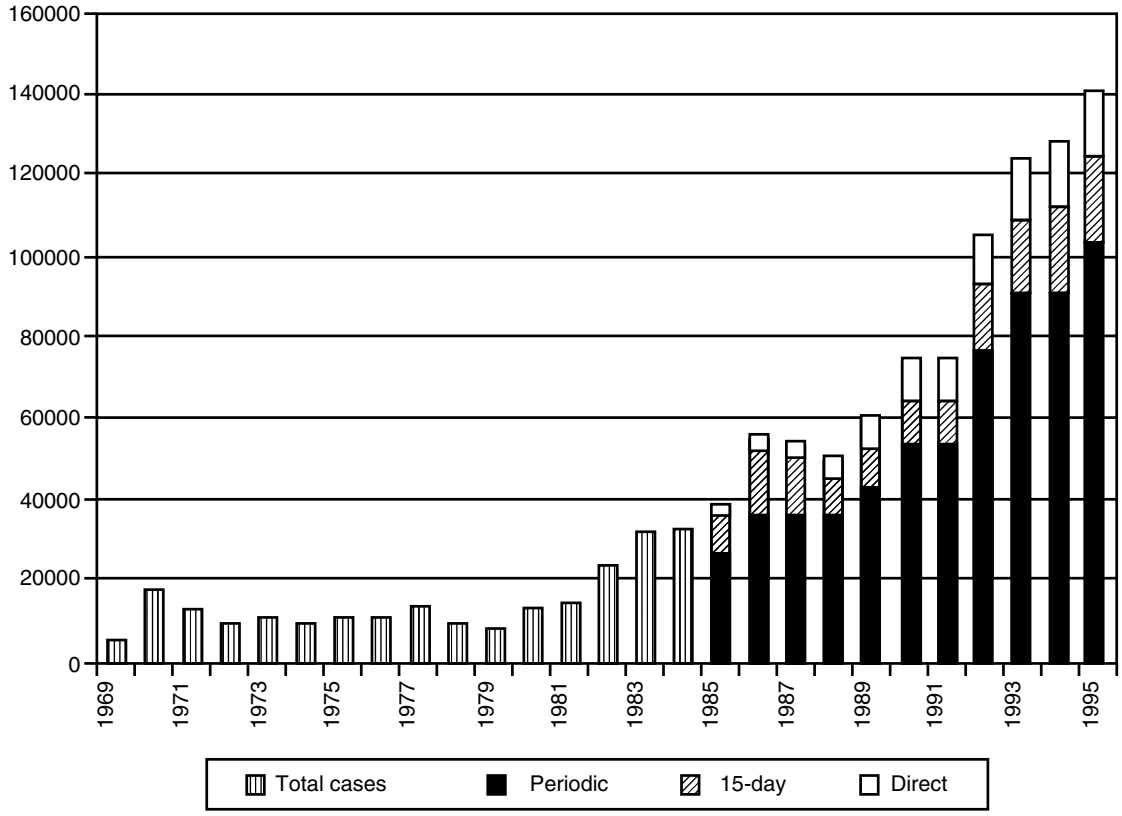

Figure 2 Count of all domestic and foreign reported adverse events by type of report by year, 1969–1995

of ADE reports and the denominator consists of some measure of exposure [22].

In proposing this approach it was recognized that the numerator, consisting of ADE reports, was subject to reporting biases, and that the denominator, consisting of estimated drug exposure based on sales data or prescription surveys, is a crude estimate of exposure. The approach attempts to detect changes from one period of reporting relative to an earlier reporting period.

Adopting the notation of Prause [17], the FDA approach assumes that ADEs are rare, and the number of reported ADEs (x_i) for a given drug–event (in the ith interval) follows a **Poisson distribution** with parameter $c_i p_i$, where i denotes period and x_i denotes the number of reported adverse reactions of a specific type of event for a given drug, c_i denotes the corresponding sales or any other estimate of drug usage for this period, and p_i denotes the rate of a reported adverse reaction per sales unit in period i. Let c_{i-1} be the sales in the historical (previous) period.

We define R_i as the proportion of the ith interval sales out of the sum of the sales for the i and $i-1$ intervals, i.e.

$$R_i = \frac{c_i}{(c_{i-1} + c_i)},$$

and denote the observed number of ADEs for the ith and $(i-1)$th intervals as x_i and x_{i-1}, respectively. X_i and X_{i-1} can be modeled as two independent Poisson random variables. Thus, the following **hypothesis testing** criteria could be used as a basis for an alert system:

$$\mathrm{H}_0 : p_i \leq p_{i-1} \text{ vs. } \mathrm{H}_1 : p_i > p_{i-1}. \qquad (1)$$

FDA proposed an asymptotic procedure to test this hypothesis based on the Poisson assumption. For the comparison of p_i and p_{i-1}, a normal approximation (*see* **Normal Distribution**) gives the test

statistic,

$$Z_{\mathrm{FDA}} = \frac{x_i - \left(\frac{R}{1-R}\right) x_{i-1}}{\left[x_i + \left(\frac{R}{1-R}\right)^2 x_{i-1}\right]^{1/2}}. \qquad (2)$$

A signal occurs when $Z_{\mathrm{FDA}} > Z_{1-\alpha}$, where Z is a standardized normal variate (*see* **Standard Normal Deviate**) and α is the tail probability. This approach signals an increased frequency of reports and assumes: (i) an **unbiased** comparison of the number of reports of a specific drug–event reaction during a recent appropriate time period with the number of reports of the same drug–event reaction during a previously observed comparative time period; and (ii) that differences in drug exposure between the intervals compared are taken into account. Norwood & Sampson [16] proposed a **binomial** approach based on the conditional distribution of x_i given $x_i + x_{i-1}$ to test the hypothesis (1).

Both the FDA Poisson-based method and the modification proposed by Norwood & Sampson test the **null hypothesis** that there is no change between the historical and the current period in the number of ADE rates adjusted for sales against the **alternative hypothesis** that there is an increase in the sales-adjusted ADE rates.

Using the same notation as above, Norwood & Sampson define

$$\Pr\left(X_i = x_i | x_i + x_{i-1}\right) = \binom{y}{x_i} R^{x_i} (1-R)^{y-x_i}, \qquad (3)$$

where $y = x_i + x_{i-1}$.

The normal approximation to the above conditional binomial distribution leads to the following asymptotic version of the exact test:

$$Z_{\mathrm{NS}} = \frac{x_i - (x_i + x_{i-1})\, R}{\left[x_i + x_{i-1} R_i\,(1-R_i)\right]^{1/2}}. \qquad (4)$$

Norwood & Sampson presented several examples of how to apply this approach to ADEs associated with a particular product line involving various dosage forms (capsules, tablets, liquid, and powder). In one of their examples, they observed four ADEs in the historical period with 18 000 sales, giving a historical reporting rate of 0.22 per thousand sales, compared to two ADEs in 5 650 sales, or a reporting rate of 0.35 per thousand in the current period. For these data the sales ratio $R = 5.65/(5.65 + 18) = 0.24$. The ***P* value** corresponding to the binomial test was 0.44 and thus the observed reporting change could be explained by chance.

CUSUM Approach

Praus et al. [17] propose the use of a cumulative sum (CUSUM) chart to address this question. They assume that a background incidence level k_0 of reported ADEs per unit of sales volume is known, together with its **standard deviation**. This information may, for example, be estimated by a run-in-series of earlier periods. Let $k_1 > k_0$ be an increase in the level that one considers important to detect and which serves as the rejection criterion.

The cusum score S_i for the period i is then defined in relation to its value in the previous period $i - 1$ by

$$S_i = \max\left(0,\, S_{i-1} + \frac{x_i}{c_i} - k_r\right), \quad i > 0, \qquad (5)$$

where $S_0 = 0$.

Here, k_r, where $k_0 < k_r \leq k_1$, is called the reference level and is usually taken as the mean of k_0 and k_1. x_i and c_i are defined as the number of ADRs and the sales volume in period i, respectively, as indicated earlier.

Whenever S_i exceeds a detection boundary h, an increased incidence level is suspected (alert case). The average run length ARL_0 under the background incidence k_0, the standard deviation of k_0, and the difference $k_1 - k_r$ will determine h, and the average run length ARL_1 under an assumed increased incidence level k_1 is determined similarly. The equations defining h and ARL_i cannot be solved explicitly, but tables and nomograms for practical use are available.

Praus et al. [17] applied the cusum procedure to the detection of serious ADEs for diphtheria vaccine. The main advantage of the cusum method in this diphtheria example is that it allows one to detect trends over time. One limitation of a cusum method is the need to have data on a drug that has been on the market for a long time to obtain a stable background rate.

Monitoring the Comparative Reporting of a Specific Drug–Event for Several Drugs in the Same Class

Tsong [21] extended the work of Rossi et al. [18] and Hsu [11] to compare the reports of ADEs associated with several nonsteroidal anti-inflammatory drugs (NSAIDs). The comparison of one marketed drug to another using data from an ADE registry introduces additional problems such as ensuring that the indication is the same for each drug compared, that the patients are comparable, that the patterns of concomitant medications are similar, and that the reporting patterns are the same if each drug is first marketed in different years. The **power** of this comparative approach has been investigated by Tsong [22].

Comparing Two Drugs First Marketed in the Same Year

Let X_{1i} and X_{2i} denote the number of reports of drug A and drug B, respectively, in year i; and C_{1i} and C_{2i} denote the number of prescriptions of drug A and drug B, respectively, in year i. Let P_{1i} and P_{2i} denote the ADE reporting rates of drug A and drug B, respectively, in year i; P_{1i} and P_{2i} are estimated by $\hat{p}_{1i} = x_{1i}/c_{1i}$ and $\hat{p}_{2i} = x_{2i}/c_{2ii}$, and $R_i = p_{1i}/p_{2i}$ denotes the ratio of reporting rates in year i.

Let $x_i = x_{1i} + x_{2i}$ be the combined number of ADE reports of both drugs, and $p_i = c_{1i}/(c_{1i} + c_{2i})$ be the proportion of prescriptions for drug A out of the total prescriptions combining drug A and drug B.

Assuming that x_{li} and p_i are given, under the hypothesis of equal reporting rates of drug A and drug B, x_{1i}, the **random variable** for the number of reports of drug A (and x_{1i} is the realization of X_{1i}) can be assumed to be distributed as BIN (x_i, P_i). For testing $H_0 : R_i = 1$ vs. $\mathrm{H_a} : R_i \neq 1$, the binomial test is the conditionally most powerful test, namely

$$Z = \frac{x_{1i} - (x_i c_i)}{x_i c_i (1 - p_i)^{1/2}} \tag{6}$$

When x_i or x_{1i} is small, the exact binomial p-values can be determined. Otherwise the normal approximation of the binomial test is the Z-test comparing against $Z_{\alpha/2}$, the $(1 - \alpha/2)$th percentile of the standard normal distribution; typically, $\alpha = 0.05$. R_i is estimated by $\hat{R}_i = (x_{1i}/c_{1i})/(x_{2i}/c_{2i})$, and its **confidence interval** can be calculated using Cornfield's method.

Monitoring Drug–Events Reported from Multiple Sources

Moussa [15] proposed a probabilistic approach to monitoring ADEs in situations where ADE reports are being collected from multiple registries (*see* **Disease Registers**) and collected at one central location (e.g. a **World Health Organization (WHO)** registry comprised of several National registries). In this situation, additional heterogeneities and irregularities associated with each reporting center or region are to be expected. In Moussa's approach each subpopulation is considered a cluster (which is defined as a homogeneous group of reports on a specific drug–event from one subpopulation in a specified time unit). The reports in a cluster are assumed to conform to a Bernoulli distribution, with the parameter varying between clusters of the same subpopulation according to a two-parameter **beta distribution**. An additional assumption is made that the cluster size follows a **negative binomial distribution**. The four parameters of the compound model are estimated by **maximum likelihood**.

Moussa uses these assumptions to construct a one-sided cumulative sum test to signal an alert. His model is applied to reports of intestinal hemorrhage associated with a specific drug submitted over 16 months to the UK Committee on Safety of Medicine, where the clusters are administrative subpopulations of the UK.

Other Computer-assisted Surveillance Methods

We now turn to several other surveillance methods that were not originally developed for drug–event monitoring but which can easily be adapted for that purpose. Recall that each report to FAD's system already has a drug–event linkage, so that aspect of the event–exposure relationship is known. Other earlier methods only assumed a reported event but assumed no exposure linked to the event. Computer-assisted surveillance was used to alert health agencies of unexpected increases in congenital malformations, not necessarily linking these events to any particular exposure. Hill et al. [10] compared the local incidence of congenital malformations to the expected national experience in various administrative areas in the UK. They used the cumulative sum

techniques of Ewan & Kemp [5] and Woodward & Goldsmith [24]. Bjerkedal & Bakketeig [1] used a system which incorporated all births in Norway on a monthly basis, and selected **International Classification of Diseases (ICD)** codes associated with a description of the event. They established control limits which would produce one false alarm every 230 months. Ericson et al. [4] compared two surveillance systems using two registries run in parallel in Sweden. One was a specific report card and the second was a computerized analysis of births, and the comparative **specificity** and timeliness of these systems was evaluated. The authors concluded that the specific report system was the only one suitable for monitoring malformations.

An alternative method to the cusum technique was introduced by Chen [2]. Chen describes a system that would be suitable for surveillance of congenital malformations in a single hospital or in several hospitals based on the number of consecutive births occurring between the birth of an infant with a specific monitored malformation and the birth of a second infant with the same malformation. The groups of such consecutive births are called sets, and the set size is a random variable with an assumed **geometric distribution**. Chen compared the relative efficiencies of the sets and the cusum techniques in monitoring the occurrence of rare events. Chen found the sets and cusum techniques to be comparable. The former was somewhat less efficient when monitoring a single hospital, although it was computationally simpler.

Levin & Kline [14] used a modification of Page's cumulative sum procedure to investigate unusual fluctuations in the proportion of spontaneous abortions. The monitoring of Nosocomial infections by the National Nosocomial Infections Study (NNIS) for the **Centers for Disease Control (CDC)** motivated Shore & Quade [19] to propose a surveillance system to detect an increase in the mean of the Poisson distribution of cases of a disease. Like Chen, they suggest an alert or signal to be based on the run-length distribution based on the tail probability of a geometric distribution.

Thus we see that several methods of assessment of risk related to newly and currently marketed drugs have been proposed. For the most part, this area of research is limited by the completeness of the numerator data and the biases associated with the reported ADEs and the difficulties in obtaining an adequate denominator as a surrogate for exposure. The most useful areas of work have focused on the search for increased frequencies of ADEs for a specific drug–event combination.

References

[1] Bjerkedal, T. & Bakketeig, L.S. (1975). Surveillance of congenital malformations and other conditions of the newborn, *International Journal of Epidemiology* **4**, 31–36.

[2] Chen, R. (1978). A surveillance system for congenital malformations, *Journal of the American Statistical Association* **73**, 323–327.

[3] Department of Health and Human Services, Public Health Service, Food and Drug Administration, United States (1992). *Guideline for Postmarketing Reporting of Adverse Drug Experiences*, Docket No. 85D-0249.

[4] Ericson, A., Kallen, B. & Winberg, J. (1977). Surveillance of malformations at birth: A comparison of two record systems run in parallel, *International Journal of Epidemiology* **6**, 35–41.

[5] Ewan, W.D. & Kemp, K.W. (1960). Sampling inspection of continuous processes with no autocorrelation between successive results, *Biometrika* **47**, 363–380.

[6] Finney, D.J. (1971). Statistical aspects of monitoring for dangers in drug therapy, *Methods of Information in Medicine* **10**, 1–8.

[7] Finney, D.J. (1971). Statistical logic in the monitoring of reactions to therapeutic drugs, *Statistical Logic in Drug Monitoring* **10**, 237–245.

[8] Friedman, G.D. (1972). Screening criteria for drug monitoring, *Journal of Chronic Disease* **25**, 11–20.

[9] Hill, A.B. (1971). The computer surveillance of congenital malformations, in *Principles of Medical Statistics*, 9th Ed. R. & R. Clark, Edinburgh.

[10] Hill, G.B., Spicer, C.C. & Weatherall, J.A.C. (1968). The computer surveillance of congenital malformations, *British Medical Bulletin* **24**, 215–218.

[11] Hsu, J.P. (1985). Refinement of the Methods of Analysis for the NSAID Study, *CDER Division of Biometrics Memorandum*. FDA, Rockville.

[12] Inman, W.H.W., Vessey, M.P. & Westerholm, B. (1970). Thromboembolic disease and the steroidal content of oral contraceptives: A report to the committee on safety of drugs, *British Medical Journal* **2**, 203–209.

[13] Johnson, J.M. (1992). Reasonable possibility: Causality and postmarketing surveillance, *Drug Information Journal* **26**, 553–558.

[14] Levin, B. & Kline, J. (1985). The cusum test of homogeneity with an application in spontaneous abortion epidemiology, *Statistics in Medicine* **4**, 469–488.

[15] Moussa, M.A.A. (1978). Statistical problems in monitoring adverse drug reactions, *Methods of Information in Medicine* **17**, 106–112.

[16] Norwood, P.K. & Sampson, A.R. (1988). A statistical methodology for postmarketing surveillance of adverse drug reaction reports, *Statistics in Medicine* **7**, 1023–1030.

[17] Praus, M., Schindel, F., Fescharek, R. & Schwarz, S. (1993). Alert systems for post-marketing surveillance of adverse drug reactions, *Statistics in Medicine* **12**, 2383–2393.

[18] Rossi, A.C., Hsu, J.P. & Faich, G.A. (1987). Ulcerogenicity of piroxicam: Analysis of spontaneous report data, *British Medical Journal* **294**, 147–150.

[19] Shore, D.L. & Quade, D. (1989). A surveillance system based on a short memory scheme, *Statistics in Medicine* **8**, 311–322.

[20] Strom, B.L. (1994). *Pharmacoepidemiology*, 2nd Ed., Part III. *Systems Available for Pharmacoepidemiology Studies*. Wiley, Chichester.

[21] Tsong, Y. (1992). False alarm rates of statistical methods used in determining increased frequency of reports on adverse drug reaction, *Journal of Biopharmaceutical Statistics* **2**, 9–30.

[22] Tsong, Y. (1995). Comparing reporting rates of adverse events between drugs with adjustment for year of marketing and secular trends in total reporting, *Journal of Biopharmaceutical Statistics* **5**, 95–114.

[23] Wood, S.M. & Coulson, R. (1993). Adverse drug reactions on-line information tracking (ADROIT), *Pharmaceutical Medicine* **7**, 203–213.

[24] Woodward, R.H. & Goldsmith, P.L. (1964). *Cumulative Sum Techniques*. Oliver & Boyd, Edinburgh.

[25] 21 CFR Code of Federal Regulations, Part 20,310.305 and 314.80, Department of Health and Human Services, Food and Drug Administration (Thursday October 27, 1994) 34046–34064.

(*See also* **Drug Utilization Patterns**; **Scan Statistics for Disease Surveillance**)

C. ANELLO & R. O'NEILL

Poststratification in Survey Sampling

To stratify is to partition a set into disjoint subsets, called strata. Poststratification is partitioning a sample, i.e. after the sampling has been performed. It is an important method in statistical survey practice, e.g. in opinion polls. In textbooks it is less prominent.

Poststratification can be combined with any sampling method, but a typical situation is that the sample is viewed as obtained by **simple random sampling**. Suppose that we are investigating some gender-related property and note that there are more men than women in the sample. It is then natural to estimate within each sex group and then combine the two estimates into one with the help of the correct sex ratio in the **target population**. This is precisely the idea of poststratification.

Since poststrata have random sizes, the method results in larger estimator **variance** than stratification before sampling (*see* **Stratified Sampling**). It is still often used for practical reasons, cost or simplicity. In multipurpose studies, different poststratifications can be employed on the same sample for different estimating purposes. Sometimes stratification cannot be done in advance because it is not known to which stratum an individual belongs. In political opinion polls it is thus frequent to ask not only about voting intentions but also how you voted at the last election, and then to poststratify according to the latter variable.

Poststratification in not too many groups, within which the study variable varies less than in the whole population, can diminish **variance** substantially. Since the number of strata grows at a multiplicative rate with the introduction of new stratifiers, there is a risk of overstratification, which increases variance again. If the sampling scheme is somehow unbalanced, e.g. so that there are discrepancies between **sampling frame** and target population, or there is **nonresponse**, poststratification can also reduce **bias**.

Mathematically, the situation can be expressed as follows. Consider a study variable Y and a stratification variable X. The purpose is to estimate the population average of Y. If the sampling procedure is balanced, then this will be the same as the **expectation** of Y, $\sum yp_{xy}$, where the sum is over all possible values of the two variables, and p_{xy} denotes the probability that the ith observation (usually the ith individual) of the sample has the values x and y. The point is that the p_{xy} are unknown, whereas the marginal distribution of the stratification variable,

$$p_x := \sum_y p_{xy},$$

is known, or we have a good estimate of it from some other (large) study.

Let s denote the sample, x_i, y_i the values of the two variables for the individual i, n_x the number of sampled individuals with $x_i = x$, and

$$\overline{y}_x := \sum_{i \in s, x_i = x} \frac{y_i}{n_x}$$

if $n_x > 0$ and zero otherwise. Then the estimator of the population Y-average poststratified with respect to X is defined as

$$\hat{y} := \sum_x p_x \overline{y}_x.$$

The choice $\overline{y}_x = 0$ in empty poststrata is for mathematical convenience. It could well be argued that in such cases one should choose the full sample average or a combination of $\overline{y}_{x'}$, where the x' are stratifier values deemed to yield y-values close to those corresponding to x.

If, given strictly positive n_x, $\overline{y}_x$ is an **unbiased** estimator of the conditional expectation of Y, given $X = x$, then the poststratified estimator will have no bias, provided all $n_x > 0$. The overall bias will be $\sum_x p_x \Pr(n_x = 0)$ which is $\sim q^n$ in the usual sampling schemes, for sample size n and $q = \max_x(1 - p_x)$.

Provided the different y_i are uncorrelated, the conditional variance will be

$$\mathrm{var}(\hat{y} | n_x, \text{ all } x) = \sum_{x; n_x > 0} \frac{p_x^2 \sigma^2(Y|x)}{n_x},$$

where $\sigma^2(Y|x)$ is the conditional variance of Y, given $X = x$:

$$\sigma^2(Y|x) := \sum_y y^2 \frac{p_{xy}}{p_x} - \mu^2(X|x),$$

$$\mu(Y|x) := \sum_y y \frac{p_{xy}}{p_x}.$$

If sampling is simple and random without replacement (*see* **Sampling With and Without Replacement**) and the model is the classical one of fixed given y-values in the population, then this is still approximately true, the risk of sampling the same individual more than once being disregarded.

The overall variance is then obtained as

$$\sum_x p_x^2 \sigma^2(Y|x) \mathrm{E}\left(\frac{1}{n_x} | n_x > 0\right) \Pr(n_x > 0) + \mu^2(Y|x) \Pr(n_x = 0)(1 - \Pr(n_x = 0)).$$

Here, the latter term can often be disregarded and the expectation of $1/n_x$ approximated by

$$\frac{1}{np_x} + \frac{(1 - p_x)}{(np_x)^2}.$$

In practice, the second term is usually also neglected, leading to the impression that poststratification "always pays". As pointed out, however, this can be misleading: if within-strata variances are not considerably smaller than the overall variance of the study variable, and the number of poststrata, i.e. the number of different x-values, is of the same order as n (a situation that can be met with in practice) the second term is not negligible.

From a theoretical viewpoint, poststratification has certain desirable properties: it is **maximum likelihood** in a general model, and it has a conditional **minimal variance** property among a natural class of estimators (*see* **Estimation**). The question whether **inference** statements should be made conditionally upon poststratification or not is discussed in two articles by Jagers et al. and by Smith, the latter of which also gives an overview. The method is looked at from a **Bayesian** angle by Little.

Bibliography

Jagers, P., Odén, A. & Trulsson, L. (1985). Post-stratification and ratio estimation: usages of auxiliary information in survey sampling and opinion polls. *International Statistical Review* **53**, 221–238.

Little, R.J.A. (1993). Post-stratification: a modeler's perspective, *Journal of the American Statistical Association* **88**, 1001–1012.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Smith, T.M.F. (1991). Post-stratification, *Statistician* **40**, 315–323.

PETER JAGERS

Power Divergence Methods

Numerous asymptotic test statistics have been proposed to assess the fit of a collection of counts with a **multinomial** or product-multinomial probability model. Minimizing lack of fit, as measured by such a statistic, is a plausible approach to estimating unknown model parameters. Power divergence provides a unifying conceptual and computational framework for these statistics and their associated estimators, facilitating both analytic and simulated comparisons of their behaviors and utility in small sample and sparse large sample **categorical data analysis**.

Let the observed counts be n_{ij} for independent multinomial samples $i = 1, \ldots, s$ and response categories $j = 1, \ldots, r$, with n_{i+} the size of the i^{th} sample, $N = \sum_{i=1}^{s} n_{i+}$, and $p_{ij} = n_{ij}/n_{i+}$. Write a model for the cell probabilities $\pi_{ij} = E(p_{ij})$ as $\pi_{ij} = \pi_{ij}(\boldsymbol{\theta})$ for vector $\boldsymbol{\theta}$, with $\sum_{j=1}^{r} \pi_{ij}(\boldsymbol{\theta}) = 1$ for each i. The "**goodness-of-fit** statistics" most commonly used in this context are the log **likelihood ratio** statistic [33]

$$X_L^2 = 2\sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij} \log\left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right), \qquad (1)$$

Pearson's chi-square [36]

$$\begin{aligned} X_P^2 &= \sum_{i=1}^{s} n_{i+} \sum_{j=1}^{r} \frac{(p_{ij} - \pi_{ij}(\boldsymbol{\theta}))^2}{\pi_{ij}(\boldsymbol{\theta})} \\ &= \sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij} \left(\left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right) - 1\right) \\ &= \sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij} \left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right) - N, \end{aligned} \qquad (2)$$

and Neyman's modified chi-square [21, 31]

$$\begin{aligned} X_N^2 &= \sum_{i=1}^{s} n_{i+} \sum_{j=1}^{r} \frac{(p_{ij} - \pi_{ij}(\boldsymbol{\theta}))^2}{p_{ij}} \\ &= \sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij} \left(\left(\frac{\pi_{ij}(\boldsymbol{\theta})}{p_{ij}}\right)^2 - 1\right) \\ &= \sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij} \left(\frac{\pi_{ij}(\boldsymbol{\theta})}{p_{ij}}\right)^2 - N \end{aligned} \qquad (3)$$

(*see* **Chi-square Distribution**; **Chi-square Tests**). When the model specifies fixed $\boldsymbol{\theta} = \boldsymbol{\theta_0}$, each of (1)–(3) is used as an asymptotically $\chi^2_{s(r-1)}$ statistic to test the simple hypothesis (*see* **Hypothesis Testing**). Otherwise, X_L^2 and X_P^2 are evaluated at the **maximum likelihood** estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, in which case X_P^2 is Rao's likelihood score statistic [40]. In contrast, X_N^2 is evaluated at $\arg\min_{\boldsymbol{\theta}} X_N^2(\boldsymbol{\theta})$ which, for linear $\pi_{ij}(\boldsymbol{\theta})$, yields Wald's [55] statistic. The reference distribution for testing the composite hypothesis is $\chi^2_{s(r-1)-\dim(\boldsymbol{\theta})}$.

Note that each of (1)–(3) is the size-weighted sum of departures of the ratios of observed to expected cell proportions $p_{ij}/\pi_{ij}(\boldsymbol{\theta})$ from 1. However, the departures are measured on different scales: additive in (2), logarithmic in (1), and after exponentiation by -2 in (3). Other asymptotically chi-square statistics in the literature for this problem – for example, the "externally constrained" discrimination information statistic X_I^2 [12, 24] and the Freeman–Tukey chi-square, equivalently the Hellinger distance [4, 10, 41] – take form similar to (1) and (3) but with other exponents for the ratios. Since

$$\begin{aligned} &\lim_{\delta\to 0}\left(\frac{1}{\delta}\sum_{j=1}^{r} n_{ij}\left(\left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right)^{\delta} - 1\right)\right) \\ &= \sum_{j=1}^{r} n_{ij} \log\left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right), \end{aligned} \qquad (4)$$

X_L^2 is naturally addressable as the "exponent 0" member of this class.

Hence, Read and Cressie [8, 43, 45] defined and studied the class of power-divergence asymptotic test statistics, which, extending (2), can be expressed as

$$\begin{aligned} &2I^{\lambda}(\{n_{ij}\}, \{\pi_{ij}(\boldsymbol{\theta})\}) \\ &= \frac{2}{\lambda(\lambda+1)}\left[\sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij}\left(\left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right)^{\lambda} - 1\right)\right] \\ &= \frac{2}{\lambda(\lambda+1)}\left[\sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij}\left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right)^{\lambda} - N\right] \end{aligned} \qquad (5)$$

for $\lambda \neq -1, 0$, and as the limits of (5) for $\lambda = -1$ or 0. These limits are X_L^2 for $\lambda = 0$ and, for $\lambda = -1$, the discrimination information statistic

$$X_I^2 = 2I^{-1}(\{n_{ij}\}, \{\pi_{ij}(\boldsymbol{\theta})\}) = 2\sum_{i=1}^{s}\sum_{j=1}^{r} n_{i+}\pi_{ij}(\boldsymbol{\theta}) \log\left(\frac{\pi_{ij}(\boldsymbol{\theta})}{p_{ij}}\right), \quad (6)$$

which is a sample size-weighted sum across populations of **Kullback–Liebler information** numbers comparing fitted to observed distributions. By (1) and (5), for $\lambda \leq 0$, the statistic only exists when $n_{ij} > 0$ for all i and j. When $2I^{\delta}(\{n_{ij}\}, \{\pi_{ij}(\boldsymbol{\theta})\})$ exists and $\boldsymbol{\theta}$ must be estimated, $\widehat{\boldsymbol{\theta}}^{(\delta)} = \arg\min_{\boldsymbol{\theta}}(2I^{\delta}(\{n_{ij}\}, \{\pi_{ij}(\boldsymbol{\theta})\}))$ is the minimum power-divergence estimator (MPE) of order δ. In such a case, the power-divergence test statistics $2I^{\lambda}(\{n_{ij}\}, \{\pi_{ij}(\widehat{\boldsymbol{\theta}}^{(\delta)})\})$ may be used for $\delta \neq \lambda$ without affecting many asymptotic properties. Thus, as noted above, in common application X_P^2 corresponds to $\lambda = 1, \delta = 0$. For the usual Freeman–Tukey statistic, $\lambda = -1/2$ with $\delta = 0$. For unknown $\boldsymbol{\theta}$, the statistics are thus most appropriately doubly indexed. But most comparative studies of the power-divergence class have focused on relatively simple situations for which this is not necessary: either fixed $\boldsymbol{\theta}$ for a single multinomial, or $\widehat{\boldsymbol{\theta}}^{0}$ under the independence hypothesis for a two-way **contingency table**, for which the maximum likelihood estimates are typically used for tests of any order λ. We thus suppress δ here for simplicity (*see* **Independence of a Set of Variables, Tests of**).

Intuition may be gained from simple settings allowing analytic comparison of MPEs. For $\boldsymbol{\theta}$ of length t, the estimation equations are

$$\sum_{i=1}^{s}\sum_{j=1}^{r}\left(\frac{p_{ij}}{\pi_{ij}(\boldsymbol{\theta})}\right)^{\lambda+1}\frac{\partial\pi_{ij}(\boldsymbol{\theta})}{\partial\theta_k} = 0, \quad (7)$$

$k = 1, \ldots, t$. Consider a single multinomial under the constraint $\pi_j = \theta_1$ for $j = 1, \ldots, r^*$, $\pi_j = \theta_2$ for $j = r^* + 1, \ldots, r$. Then

$$(\widehat{\theta}_1^{(\lambda)}, \widehat{\theta}_2^{(\lambda)})' \sim \left(\left(\frac{1}{r^*}\sum_{j=1}^{r^*} p_j^{\lambda+1}\right)^{\frac{1}{\lambda+1}}, \left(\frac{1}{(r-r^*)}\sum_{j=r^*+1}^{r} p_j^{\lambda+1}\right)^{\frac{1}{\lambda+1}}\right)', \quad (8)$$

with proportionality constant enforcing the constraint $r^*\widehat{\theta}_1^{(\lambda)} + (r - r^*)\widehat{\theta}_2^{(\lambda)} = 1$. Similarly, for estimating $\theta' = (\pi_1, \ldots, \pi_r)'$, $\sum_{j=1}^{r}\pi_j = 1$ from s independent multinomial samples under the homogeneity hypothesis $\pi_{ij} = \pi_j, i = 1, \ldots, s, j = 1, \ldots, r$,

$$\widehat{\pi}_j^{(\lambda)} \sim \left(\sum_{i=1}^{s}\left(\frac{n_{i+}}{N}\right)p_{ij}^{\lambda+1}\right)^{\frac{1}{\lambda+1}}, \quad (9)$$

with proportionality constant such that $\sum_{j=1}^{r}\widehat{\pi}_j^{(\lambda)} = 1$ [20, 38, 45]. An extension of (9) is available for monotone **missing data**, for example, as with dropouts in **longitudinal** studies [20]. Thus, MPEs in simple cases are normalized and possibly weighted power means, with Pearson's chi-square leading to root mean squares, the Hellinger distance squared mean roots, X_I^2 the geometric mean, and X_N^2 the harmonic mean. The maximum likelihood estimator leads to arithmetic means, and consequently is the only MPE with the commonsense invariance property that estimators are unaffected by pooling, within and across populations, of categories with a common probability.

The form of (5) suggests that power-divergence statistics will be increasingly sensitive to cells with high observed to expected count ratios for increasingly large positive λ, and to cells with low ratios for increasingly large negative lambda. Read and Cressie [45] found this to be true in studies of statistical **power** against "spike" or "dip" alternatives to uniformity in a single multinomial, for which one or two cell counts substantially exceed or fall short of expectation. They recommended against the use of power-divergence statistics with $|\lambda| > 5$, on grounds that they are effectively dominated by unexpectedly low ($\lambda < 5$) or high ($\lambda > 5$) counts. Similarly, (5)–(7) suggest that MPEs will be more sensitive to the lowest among the observed proportions for increasingly negative λ. To illustrate, Figure 1 plots $\widehat{\pi}_1^{(\lambda)}$ from (9) for $\lambda = -6$ to 6, for equal size samples from two binomials with $p_1 = 0.1$, $p_2 = 0.6$. Over all λ, $\widehat{\pi}_1^{(\lambda)}$ ranges from 0.2 as $\lambda \to -\infty$ to 0.4 as $\lambda \to \infty$, as compared to $\widehat{\pi}_1^{(0)}$, the maximum likelihood estimator and pooled proportion, of 0.35.

It is somewhat surprising that the power-divergence unification was not accomplished thirty years earlier. X_N^2, X_L^2, and X_P^2 were all known before 1930 ([31, 33, 36], respectively). In a seminal unifying paper, Neyman [32] defined the

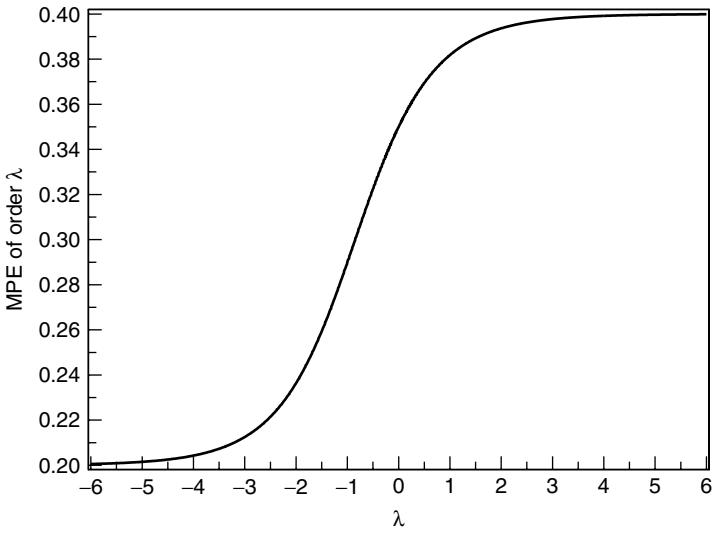

Figure 1 Minimum power-divergence estimates (MPEs) of π_1 under homogeneity of two binomials ($n_{1+} = n_{2+}$, $p_1 = .1$, $p_2 = .6$)

class of "best" asymptotically normal **(BAN) estimators** of $\boldsymbol{\theta}$: consistent, asymptotically normal continuous functions of the p_{ij}, with minimum asymptotic ($n_{i+} \to \infty, n_{i+}/n \to c_i < \infty$) variance among estimators having those properties (*see* **Efficiency and Efficient Estimators**). Neyman showed estimators minimizing any of X_L^2, X_P^2 or X_N^2 to be BAN. For testing, the minimized X_P^2 and X_N^2 were shown to share the same null chi-square distribution as X_L^2, or as X_P^2 evaluated at the MLE of $\boldsymbol{\theta}$, or as any of (1)–(3) evaluated at any BAN estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. This result applies to all power-divergence statistics and MPEs.

Haldane [13], based on conversation with C.A.B. Smith, credits M.C.K. Tweedie with the use of the estimating equations (7) in the 1940s; they may have been used informally even earlier (*see* **Estimating Functions**). Working with a single population and parameter θ, Haldane was looking for a class of fit statistics from which improved estimates might be found, and was clearly aware of $\sum_{j=1}^{r} n_j (p_j/\pi_j(\theta))^{\lambda}$ as a candidate kernel for such a class. However, to obtain test statistics and estimators fully defined even for samples with zero cell counts, he replaced n_j with $\lambda!\binom{n_j}{\lambda}$. The remainder of [13] shows the resulting estimators to be BAN, and compares and discusses bias adjustments (*see* **Unbiasedness**). Interestingly, one estimator is a version of X_N^2 for which each cell count is augmented by 1. Haldane's "divergence" measures and "minimum discrepancy" estimator did not, however, gain widespread acceptance.

Kullback [24] obtained the logarithmic functions X_L^2 and X_I^2 as *directed divergences* between observed and fitted multinomial distributions [42]. Imrey [20] applied Tweedie's estimating equations to monotone missing data, with a general solution implicit. Much later, Read and Cressie [8, 43, 45] recognized, crucially, that other goodness-of-fit chi-square statistics are multiples of nonadditive directed divergences [42]

$$I^{\alpha}(\pi : p) = c\left(\sum_{j=1}^{r} \pi_j^{\alpha} p_j^{1-\alpha} - 1\right) \qquad (10)$$

between fitted and empirical probability mass functions. The power-divergence class, as defined in (4), follows by taking $\alpha = -\lambda$ and normalizing to the chi-square distribution.

The distributions of power-divergence statistics and performance characteristics of the corresponding goodness-of-fit tests have been compared in several ways. For simplicity we restrict to $s = 1$. With $\boldsymbol{\theta}$ defined as the probability vector $\boldsymbol{\pi}$, when increasingly large samples are drawn from a null multinomial distribution with $\boldsymbol{\theta}=\boldsymbol{\pi}_{\mathbf{0}}$, all power-divergence statistics share the limiting χ_{r-1}^2 distribution. Under the same

scenario with composite null $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ of dimension t, and where the test statistic is calculated at a BAN estimate of $\boldsymbol{\theta}$, the degrees of freedom become instead $(r-1)-t$.

Under the alternative $\boldsymbol{\pi} = \boldsymbol{\pi_1}$, the power divergences with $\lambda > -1$ have limiting Gaussian distributions (*see* **Normal Distribution**) with respective means

$$\frac{2N}{\lambda(\lambda+1)}\left[\sum_{j=1}^{r}\pi_{1j}\left(\left(\frac{\pi_{1j}}{\pi_{0j}}\right)^{\lambda}-1\right)\right] \qquad (11)$$

and variances proportional to N. Consequently, all power-divergence statistics yield consistent $(\lim_{N\to\infty} Pr[2I^{\lambda}(\{n_j\},\{\pi_{1j}(\boldsymbol{\theta})\}) \geq c_{\alpha}] = 1)$ tests against fixed alternatives. The **Pitman efficiency** of test T_1 relative to test T_2 is the limiting ratio of the sample sizes required by T_2 and T_1 to maintain a specified power under a sequence of local alternatives converging to $\boldsymbol{\pi}_0$ and for which this limit exists, for example, $H_1 : \boldsymbol{\pi} = \boldsymbol{\pi}_0 + \boldsymbol{\delta}/\sqrt{N}$ with $\sum_{j=1}^{r}\delta_j = 0$, which converges at rate $N^{-1/2}$. The limiting distribution under this sequence is noncentral chi-square with noncentrality parameter $\sum_{j=1}^{r}\delta_j^2/\pi_{0j}$. In this setting, the Pitman efficiency is the ratio of noncentrality parameters of the limiting distributions of the corresponding test statistics under H_1. Since the limiting distribution does not depend on λ, all power-divergence tests have Pitman efficiency of 1 relative to X_L^2. Several power approximations for this situation have been compared for sample sizes in the range of $5r - 10r$, for which two based on **Edgeworth expansions** seem most effective [49].

For many tests, the attained significance level (i.e. ***P* value**) under a fixed **alternative hypothesis** almost surely declines exponentially to 0 with increasing sample size N, for N sufficiently large. Half the rate of exponential decline for such a test is known as its "exact slope", and the ratio of exact slopes of two tests is the limit of the ratio of sample sizes required, almost surely, for each of the tests to achieve statistical significance at level α, as $\alpha \to 0$. The Bahadur efficiency of T_1 relative to T_2 is thus defined as the ratio, where it exists, of the exact slope of T_2 to that of T_1. No power-divergence test has greater Bahadur efficiency than X_L^2, although others may have equal efficiency depending on the hypotheses being compared [37].

Efficiency results change considerably when one considers "sparse" asymptotics in which increasing sample size is associated with finer classification: as $N \to \infty$ the number of cells r_N in the multinomial expands as well, so that $r_N \to \infty$, $(N/r_N) \to c$, with $0 < c < \infty$. As suggested by the large-sample normality of chi-square distributions with many degrees of freedom, the limiting distributions of power-divergence statistics for $\lambda > -1$ are here Gaussian. The moments of these distributions are functions of λ, and hence, the behaviors of the statistics differ under both null and alternative models. For $\lambda \leq -1$ under sparse asymptotics, the probability that $2I^{\lambda}(\{n_{ij}\},\{\pi_{ij}(\boldsymbol{\theta})\})$ is undefined remains positive no matter how large the sample, so the limiting distributions can only be defined conditional on all positive cell counts. With $\lambda > -1$, for testing uniformity in a single multinomial, Pitman efficiency may be evaluated under a sequence $H_1 : \boldsymbol{\pi} = \boldsymbol{\pi}_0 + \boldsymbol{\delta}/N^{1/4}$. While dependence on λ is small for $-1 < \lambda \leq 3$, and declines with increasing cell expectations N/r_N, X_P^2 is optimally efficient. However, this optimality holds in the narrowest of circumstances, for it fails to generalize to testing a simple nonuniform null hypothesis, for which no statistic is clearly superior, or to nonlocal dip alternatives for which the relative performance of X_P^2 and X_L^2 appears dependent on the number of dips [23]. It has also been noted that limiting normal distributions under this local sequence are unchanged by the presence of **nuisance parameters**, suggesting that in this setting, power-divergence tests have low efficiency as a class [50].

Studies have compared accuracy in small sample settings of large-sample approximations to the distributions of power-divergence statistics. Read [44] notes that for testing a simple multinomial hypothesis, the standard chi-square approximation is not adequate for $\lambda < 1/3$ or $\lambda > 3/2$. Of the conventional goodness-of-fit statistics, only X_P^2 falls within the acceptable range. The poor performance of the approximation for X_L^2 relative to that for X_P^2 has been found consistently [11]. The power-divergence statistic with $\lambda = 2/3$ is sometimes known as the "Cressie–Read statistic". Read [44] and Read and Cressie [45] recommend this choice from the power-divergence class based on its generally good performance in studies of both conventional and sparse asymptotics, coupled with the convenient adequacy of the conventional null approximation for all but very small sample sizes. Recent work confirms the conventional practice of basing use of the asymptotic approximation on minimum expected count (*see*

Expectation) and shows that, for the Cressie–Read and Pearson statistics when $r > 2$, the approximations may be considerably more tolerant than has generally been appreciated [11].

While the above results on testing favor use of power-divergence tests for λ in a range no wider than perhaps 0.5–1.5, other studies indicate that MPEs for small negative $\lambda \varepsilon (-1, -0.5]$, for example, the minimum Hellinger distance MPE with $\lambda = -0.5$, may have substantial robustness advantages. Although this comes at a price in small sample efficiency, a penalized version that alters the influence of zero cell counts appears to improve the situation greatly [1, 47, 48].

Several data analytic techniques originally based upon individual power-divergence statistics have been generalized to the entire class. Thus, power-divergence tests for independence in the **2 x 2** contingency table have been compared [26]. Methods in the spirit of the "partition of chi-square" technique originally developed by Lancaster [25], who decomposed X_P^2 to localize sources of heterogeneity in contingency tables (*see* **Chi-square, Partition of**), and analysis of deviance for sequential reduction of hierarchical **loglinear models** [3], have been extended to "analysis of divergence" [7, 27–29]. **Akaike's information criterion** AIC, based upon X_L^2, has been generalized to the "power-divergence information criterion" PIC [6]. **Confidence intervals** derived by inverting Wald and score statistics have been compared to intervals obtained by inverting other power-divergence tests [2].

Recently, a class of statistics based on power divergences has been proposed for assessing fit of a categorical **time series regression** model [9]. Predictors may include external variables and the past history of the time series, which need not be stationary. The λ-order power-divergence based statistic is the sum, across observation times, of power divergences between the degenerate empirical distribution of the single observation at each time and the fitted distribution given by the predicted category probabilities at that time. The latter are based on the **partial likelihood** estimate $\widehat{\boldsymbol{\beta}}$ of the vector $\boldsymbol{\beta}$ of regression parameters. The centered power-divergence process using $\boldsymbol{\beta}$ is a zero-mean martingale, and its analogous process using $\widehat{\boldsymbol{\beta}}$ is approximated by a mean-square integrable martingale. From martingale **central limit theory**, the latter process is Gaussian for $\lambda > -1$, with variance a function of λ. It is also shown to be the partial likelihood score for expansion of the regression model in a λ-dependent direction. Notably, for $\lambda = 0$ the expanded model degenerates to the current model, so that the deviance is ineffective for detecting model inadequacy in this context. Values of λ from -1 to 1, excluding 0, are suggested as potentially useful.

MPEs have been considered for estimating parameters of continuous distributions using samples where the original observations have been grouped into categories and are no longer available. In this context, MPEs have been criticized as subject to substantial bias from low-level **misspecification** of the continuous distribution, particularly for certain distributions used commonly in econometric models [54].

Although the power-divergence class considerably generalizes classical goodness-of-fit statistics for categorical data, power divergence itself is a special case of a Csiszár's ϕ-divergences, a much broader class of directed divergences of form $\sum_{i=1}^{r} q_j \phi(p_j/q_j)$ for functions ϕ that are convex on $\Re_+$ with $\phi(1) = 0$, $\phi''(1) > 0$, and the conventions that $0\phi(0/0) \equiv 0$ and $0\phi(p/0) \equiv \lim_{u\to\infty} \phi(u)/u$. The power divergences arise from this formulation by taking

$$\phi_\lambda(u) = \frac{1}{\lambda(\lambda+1)}(x^{\lambda+1} - x + \lambda(1-x)) \qquad (12)$$

for $\lambda \neq -1$, 0, and limits at these two values. Considerable progress has been made in generalizing asymptotic results on power-divergence estimators and MPEs to ϕ-divergences and the corresponding MϕEs [30, 35]. Cressie and Pardo [7] have illustrated the increased flexibility by using a ϕ-divergence due to Renyi, evaluated at MPEs, for sequential testing of higher-order effects in hierarchical loglinear models.

Power divergences have also seen increasing use outside of the goodness-of-fit context, in their pure capacity as directed divergences representing the disparity between a set of probabilities and a reference distribution. Thus, for given λ, Tomizawa et al. [52] define a measure of asymmetry in square nominal contingency tables as a multiple of the order λ power divergence between the observed distribution among the off-diagonal cells and the reference symmetric distribution preserving the $n_{ij} + n_{ji}$ for all (i, j). For square tables with ordered categories, a similar measure is developed using power divergence applied to an artificial distribution constructed from cumulative probabilities calculated at all pairs of row and

column cut-points [51] (*see* **Square Contingency Table**). In classification trees with categorical variables, the sum of a power divergence of given order between each branch of a split and its source distribution, weighted proportionally to the observations in each branch, has been maximized over possible splits to determine the best split at each stage [46] (*see* **Tree-structured Statistical Methods**). Minimizing mean power divergence of observed from predicted observations in leave-one-out **cross-validation** has been used to determine the smoothing parameter in generalized **spline function**-based **logistic regression** [53]. Power divergence from a **uniform distribution** has been minimized subject to constraints in order to determine observation weights in unimodal kernel **density estimation** [15], monotone kernel regression [14], and monotone **hazard rate** estimation using a biased bootstrap [16] (*see* **Bootstrapping in Survival Analysis**).

Indeed, the task of minimizing divergence from uniformity is increasingly central to estimation. Empirical **likelihood** estimation [34, 39] views the empirical distribution as a multinomial. An unknown parameter θ is estimated subject to constraints by first associating with the observations a set of probabilities that maximizes the multinomial likelihood while preserving the constraints. The parameter estimate is then calculated from this weighted, or "maximum likelihood tilted," multinomial. The task of finding the tilted distribution is equivalent to minimizing the 0-order power divergence of tilted from uniform distribution, subject to the constraints. Estimation by exponential tilting [22] involves the same process, but the optimization criterion is minimization of the Kullback–Liebler information in the tilted distribution rather than maximization of the likelihood, again subject to constraints. The exponential tilting computation thus requires minimization of the order (-1) power divergence of tilted from uniform distribution. Generalized **method of moment** (GMM) estimation involves minimizing a quadratic form in a set of restrictions based on moment conditions [17]. In certain circumstances, an iterated version of GMM is equivalent to estimation from the distribution that minimizes the power divergence of order (-2) from uniformity [19]. Maximum likelihood and exponential tilting are also useful for increasing computational efficiency of **bootstrap** confidence intervals [18]. For parametric estimation, Choi et al. [5] suggest that **robustness** may be increased by a "biased bootstrap" tilted maximum likelihood estimator. The weights for tilting are obtained by maximizing the tilted loglikelihood on a constant contour of a power divergence from uniformity.

Thus, power divergence methods highlight connections between statistical inference and information theory, and provide a unifying framework for statistical procedures well beyond goodness-of-fit considerations for categorical data models.

References

[1] Basu, A. & Basu, S. (1998). Penalized minimum disparity methods for multinomial models, *Statistica Sinica* **8**, 841–860.

[2] Bedrick, E.J. (1987). A family of confidence intervals for the ratio of two binomial proportions, *Biometrics* **43**, 993–998.

[3] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge.

[4] Bhattacharyya, A. (1946). A measure of divergence between two multinomial populations, *Sankhyā* **7**, 401–406.

[5] Choi, E., Hall, P. & Presnell, B. (2000). Rendering parametric procedures more robust by empirically tilting the model, *Biometrika* **87**, 453–465.

[6] Cressie, N. (1996). PIC: Power-divergence information criterion, in *Statistical Theory and Applications: Papers in Honor of Herbert A. David*, H.N. Nagaraja, P.K. Sen & D.F. Morrison, eds. Springer, New York, pp. 3–14.

[7] Cressie, N. & Pardo, L. (2000). Minimum ϕ-divergence estimator and hierarchical testing in loglinear models, *Statistica Sinica* **10**, 867–884.

[8] Cressie, N.A.C. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society Series B* **46**, 440–464.

[9] Fokianos, K. (2002). Power divergence family of tests for categorical time series models, *Annals of the Institute of Statistical Mathematics* **54**, 543–564.

[10] Freeman, M.F. & Tukey, J.W. (1950). Transformations related to the angular and the square root, *Annals of Mathematical Statistics* **21**, 607–611.

[11] García-Pérez, M.A. & Núñez-Antón, V. (2001). Small-sample comparisons for power-divergence goodness-of-fit statistics for symmetric and skewed simple null hypotheses, *Journal of Applied Statistics* **28**, 855–874.

[12] Gokhale, D.V. & Kullback, S. (1978). *The Information in Contingency Tables*. Marcel Dekker, New York.

[13] Haldane, J.B.S. (1953). A class of efficient estimates of a parameter, *Bulletin of the International Statistical Institute* **33**, 231–248.

[14] Hall, P. & Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints, *The Annals of Statistics* **29**, 624–647.

[15] Hall, P. & Huang, L.-S. (2002). Unimodal density estimation using kernel methods, *Statistica Sinica* **12**, 965–990.

[16] Hall, P., Huang, L.-S., Gifford, J.A. & Gijbels, I. (2001). Nonparametric estimation of hazard rate under the constraint of monotonicity, *Journal of Computational and Graphical Statistics* **10**, 592–614.

[17] Hansen, L. (1982). Large-sample properties of generalized method of moments estimators, *Econometrica* **50**, 1029–1054.

[18] Hesterberg, Tim C. (1999). Bootstrap tilting confidence intervals and hypothesis tests, *Computing Science and Statistics* **31**, 389–393.

[19] Imbens, G.W., Spady, R.H. & Johnson, P. (1998). Information theoretic approaches to inference in moment condition models, *Econometrica* **66**, 333–357.

[20] Imrey, P.B. (1972). Linear Models Analysis of Incomplete Multivariate Categorical Data. Ph.D. Dissertation, School of Public Health, The University of North Carolina at Chapel Hill, Chapel Hill.

[21] Jeffreys, H. (1948). *Theory of Probability*, 2nd Ed. Clarendon Press, Oxford.

[22] Kitamura, Y. & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation, *Econometrica* **65**, 861–874.

[23] Koehler, K.J. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials, *Journal of the American Statistical Association* **75**, 336–344.

[24] Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons, New York.

[25] Lancaster, H.O. (1949). The derivation and partition of χ^2 in certain discrete distributions, *Biometrika* **36**, 117–129.

[26] Lee, C.-I.C. & Shen, S.-Y. (1994). Convergence rates and powers of six power-divergence statistics for testing independence in 2 by 2 contingency table, *Communications in Statistics. Theory and Methods* **23**, 2113–2126.

[27] Medak, F. & Cressie, N. (1991a). Confidence regions in ternary diagrams based on the power-divergence statistics, *Mathematical Geology* **23**, 1045–1057.

[28] Medak, F. & Cressie, N. (1991b). *Hierarchical Testing of Parametric Models Using the Power-divergence Family of Test Statistics*, Preprint Number 91-14, Iowa State University Statistical Laboratory, Ames.

[29] Medak, F. & Cressie, N. (1991c). *Hierarchical Testing for Homogeneity in Product-multinomial Distributions: Beyond the Likelihood Ratio Statistic*, Preprint Number 91-15, Iowa State University Statistical Laboratory, Ames.

[30] Morales, D., Pardo, L. & Vajda, I. (1995). Asymptotic divergences of estimates of discrete distributions, *Journal of Statistical Planning and Inference* **48**, 347–369.

[31] Neyman, J. (1929). Contribution to theory of certain test criteria, *Bulletin of the International Statistical Institute* **18**, 1–48.

[32] Neyman, J. (1949). Contributions to the theory of the χ^2 test, in *Proceedings of the First Berkeley Symposium in Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 239–273.

[33] Neyman, J., Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika* **20A**, 175–240, 263–294.

[34] Owens, A. (2001). *Empirical Likelihood*. Chapman & Hall, New York.

[35] Pardo, J.A., Pardo, L. & Zografos, K. (2001). Minimum ϕ-divergence estimators with constraints in multinomial populations, *Journal of Statistical Planning and Inference* **104**, 221–237.

[36] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling, *Philosophical Magazine* **5**(50), 157–175.

[37] Perng, S.K. (1982). Bahadur efficiency, in *Encyclopedia of Statistical Sciences*, Vol. 1, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, New York, pp. 178–181.

[38] Quade, D. & Salama, I.A. (1975). A note on minimum chi-square statistics in contingency tables, *Biometrics* **31**, 953–956.

[39] Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations, *Annals of Statistics* **22**, 300–325.

[40] Rao, C.R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

[41] Rao, C.R. (1982). Diversity and dissimilarity coefficients: A unified approach, *Theoretical Population Biology* **21**, 24–43.

[42] Rathie, P.N. & Kannappan, P. (1972). A directed divergence function of type β, *Information and Control* **20**, 38–45.

[43] Read, T.R.C. (1982). Choosing a Goodness-of-fit Test. Ph.D. Dissertation, School of Mathematical Sciences, The Flinders University of South Australia, Adelaide.

[44] Read, T.R.C. (1984). Small-sample comparisons for the power divergence goodness-of-fit statistics, *Journal of the American Statistical Association* **79**, 929–935.

[45] Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.

[46] Shih, Y.-S. (2001). Selecting the best splits for classification trees with categorical variables, *Statistics & Probability Letters* **54**, 341–345.

[47] Simpson, D.G. (1987). Minimum Hellinger distance estimation for the analysis of count data, *Journal of the American Statistical Association* **82**, 802–807.

[48] Simpson, D.G. (1989). Hellinger deviance test: efficiency, breakdown points, and examples, *Journal of the American Statistical Association* **84**, 107–113.

[49] Taneichi, N., Sekiya, Y. & Suzukawa, A. (2002). Asymptotic approximations for the distributions of the multinomial goodness-of-fit statistics under local alternatives, *Journal of Multivariate Analysis* **81**, 335–359.

[50] Tiwari, R.C. & Wells, M.T. (1992). On the power-divergence statistic in sparse multinomial models requiring parameter estimation, *Statistics & Probability Letters* **13**, 52–60.

[51] Tomizawa, S., Miyamoto, N. & Hatanaka, Y. (2001). Measure of asymmetry for square contingency tables having ordered categories, *Australian & New Zealand Journal of Statistics* **43**, 335–349.

[52] Tomizawa, S., Seo, T. & Yamamoto, H. (1998). Power-divergence-type measure of departure from symmetry for square contingency tables that have nominal categories, *Journal of Applied Statistics* **25**, 387–398.

[53] van der Linde, A. (2001). Estimating the smoothing parameter in generalized spline-based regression: I. Cross-validatory criteria for binary data using small sample sizes, *Computational Statistics* **16**, 209–219.

[54] Victoria-Fezer, M.-P. & Ronchetti, E. (1997). Robust estimation for grouped data, *Journal of the American Statistical Association* **92**, 333–340.

[55] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society* **54**, 426–482.

Further Reading

Good, I.J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.

PETER B. IMREY

Power Transformations

The assumptions in **regression** analysis include **normally distributed** errors of constant variance (*see* **Scedasticity**). Often these assumptions are more nearly satisfied not by the original **response variable** y, but by some **transformation** of y, $z(y)$. For nonnegative responses, one frequently used transformation is $\log y$. The original and transformed analyses can then be compared in a number of ways. **Residuals** can be plotted against fitted values, or assessed for normality by probability plots for various transformations (*see* **Normal Scores**). Another comparison is through analysis of the linear model using t or F tests – a correct transformation often yields a simple **linear regression** model, with no, or just a few, **interaction** or quadratic terms. A formal way of comparing transformations is to embed them in a parametric family and then to make inferences about the transformation parameter λ (*see* **Model, Choice of**). Transformations of three parts of the model are of differing complexity and importance. The most important is described in the following section.

Transformation of the Response: Box and Cox [5]. The logarithmic transformation is one special case of the normalized power transformation [5]

$$z(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y} \log y & \lambda = 0, \end{cases} \tag{1}$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i / n)$. The regression model to be fitted is then

$$z(\lambda) = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon. \tag{2}$$

For fixed λ, the value of $\boldsymbol{\beta}$ is estimated by **least squares** giving a residual sum of squares for the $z(\lambda)$ of $R(\lambda)$. The **maximum likelihood** estimate $\hat{\lambda}$ minimizing $R(\lambda)$ is found by numerical search (*see* **Optimization and Nonlinear Equations**), often over a grid of λ values. Exact minimization of $R(\lambda)$ is not required, since simple rational values of λ are customary in the analysis of data: $\lambda = 1$, no transformation: $\lambda = 1/2$, the square root; $\lambda = 0$, the logarithmic and $\lambda = -1$, the reciprocal being widely used. An approximate minimum of $R(\lambda)$ is, however, required to establish **confidence intervals** for and to test **hypotheses** about λ. It is important that these comparisons of $R(\lambda)$ do use the full form in (1) including the geometric mean. Omission of this term leads to meaningless comparisons – for most data, $\log y$ is very much smaller than y and so are the corresponding sums of squares, regardless of how well the regression models fit.

Transformation of Explanatory Variables: Box and Tidwell [6]. For a regression model with p terms, it sometimes makes sense to consider models in which one (or perhaps more) of the **explanatory variables** is transformed, when the model is

$$y = \sum_{j \neq k}^{p} \beta_j x_j + \beta_k x_k^\lambda + \varepsilon. \tag{3}$$

Again the maximizing value of λ has to be found numerically, but the calculations are more straightforward than those for the Box–Cox transformation, since the scale of the observations is unaffected by the transformation. The residual sums of squares of y can be compared directly as λ varies.

Transformation of Both Sides of the Model: Carroll and Ruppert [7]. The Box and Cox transformation often yields both approximately normal errors and a simple linear model. But sometimes the two transformations do not happen together. An example is the data on mandible length from Royston and Altman [10] plotted in **Goodness of Fit**. The analyses in **Diagnostics**, **Forward Search**, and **Residuals** showed that the log transformation yielded normal errors, but increased the evidence for the inclusion of a quadratic term in the linear model (*see* **Polynomial Regression**). If there is a simple model for y, the simplicity of the linear model can be maintained by subjecting both sides of the model to the same transformation. The purpose of the transformation is then to obtain normal errors of constant variance.

Let $\mathrm{E}(Y) = \eta = \boldsymbol{x}^T\boldsymbol{\beta}$ be the simple linear model. The transformation model is then

$$\begin{aligned} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} &= \frac{\eta^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} + \varepsilon \quad \lambda \neq 0, \\ \dot{y} \log y &= \dot{y} \log \eta + \varepsilon \quad \lambda = 0. \end{aligned} \tag{4}$$

The optimizing value of λ again minimizes the residual sum of squares of the transformed response. For given λ, estimating the parameters in general requires nonlinear estimation, although this is unaffected by

division by $\lambda\dot{y}^{\lambda-1}$. An example for the data on the volume of trees analyzed in **Residuals** is given by Atkinson [2].

The procedures using constructed variables described in **Residuals** provide tests for the value of λ, which avoid the need for calculation of $\hat{\lambda}$ whilst also giving information on the effect of individual observations on tests and parameter estimates. Numerous examples for the Box–Cox and Box–Tidwell transformations are given by Atkinson [1]. Cook and Weisberg [8, Chapter 13], describe interactive graphical methods for selecting a transformation. Diagnostic material on the transformation of both sides of the model is given by Hinkley [9], Shih [11], and by Atkinson [2]. Atkinson and Shephard [4] extend the Box–Cox transformation to time series analysis and describe the related diagnostic procedures. Diagnostic procedures for the effect of groups of observations via the **Forward Search** use the **Fan Plot**, described in greater detail in [3, Chapter 4].

References

[1] Atkinson, A.C. (1985). *Plots, Transformations, and Regression*. Oxford University Press, Oxford.

[2] Atkinson, A.C. (1994). Transforming both sides of a tree, *American Statistician* **48**, 307–313.

[3] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

[4] Atkinson, A.C. & Shephard, N. (1996). Deletion diagnostics for transformation of time series, *Journal of Forecasting* **5**, 1–17.

[5] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* **26**, 211–246.

[6] Box, G.E.P. & Tidwell, P.W. (1962). Transformations of the independent variables, *Technometrics* **4**, 531–550.

[7] Carroll, R.J. & Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, London.

[8] Cook, R.D. & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

[9] Hinkley, D.V. (1985). Transformation diagnostics for linear models, *Biometrika* **72**, 487–496.

[10] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[11] Shih, J.-Q. (1993). Regression transformation diagnostics in transform-both-sides model, *Statistics and Probability Letters* **16**, 411–420.

A.C. ATKINSON

Power

Traditionally, the *power* of a **hypothesis test** equals the probability of rejecting the **null hypothesis**, conditional on the falseness of the null. The definition implicitly depends on a number of mathematical assumptions. Equally importantly, the definition depends on a number of philosophical assumptions concerning the logic of science and the value system of the scientist. Hence, examining the concept of power requires at least a brief discussion of decision making.

The practice of statistics involves either creating **estimates** of population properties (using sample properties), or drawing **inferences** concerning population properties (using sample properties). **Probability theory** provides models of population properties and sampling processes. **Decision theory** blends mathematics, statistics, philosophy, and behavioral science to model and guide decision making. Although the systematic study of probability began centuries ago, a completely rigorous mathematical basis was not formalized until the beginning of the twentieth century. In turn, the development of estimation trailed probability, and inference (including hypothesis testing) trailed estimation.

The most common approach to statistical hypothesis testing stems from the work of **J. Neyman** and **E.S. Pearson**, around the middle of the twentieth century. No single approach has ever achieved universal approval. More frankly, statisticians have always disagreed on how to draw inferences. Currently, the most important alternative involves a **Bayesian** approach, which avoids the language and concepts of hypothesis testing.

In evaluating a scientific decision process, one may consider (i) the scientific goal, (ii) the scientist's payoff function (values), (iii) the information gathering mechanism, (iv) the mathematical model, and (v) the mathematical objective function. Our understanding of the human perceptual system depends critically on the separate consideration of perceptual performance and payoff function. For example, two radiologists examining a chest X-ray image may differ in their ability to detect lung cancer, in their willingness to declare a perceived anomaly as lung cancer, or both. Signal detection theory, as well as the closely related **receiver operating characteristic (ROC)** curve analysis, provide tools for disentangling performance from preference. Power computations provide the same ability for statistical hypothesis testing. Power analysis informs the scientist, who then chooses alternatives based upon goals and values in the setting at hand.

This broad, and therefore necessarily shallow, discussion should allow drawing three conclusions. First, to arrive at the doorstep of a power computation requires many assumptions, with most usually left unstated. Secondly, the concepts and machinery of power presented here derive from one particular approach to the problem of drawing inference in science. Thirdly, earnest and serious disagreement among statisticians suggests that achieving consensus will require a new approach, distinct from any currently available. Until then, the popularity of the Neyman–Pearson approach encourages understanding and extensive use of power analysis.

The most common approach to hypothesis testing centers on stating a null and an **alternative hypothesis** about parameters of a population. For example, consider a physician interested in mean response under standard treatment with mean response under an innovative treatment. A traditional classification of possible decisions (see, for example, Daniel [5, p. 195]) is summarized in Table 1. In such an interpretation, power equals the probability of rejecting the null, conditional on the alternative holding (*see* **Level of a Test**).

In contrast, mathematical statisticians now prefer to define power as the probability of rejecting the null, whether or not the alternative holds. The unconditional approach, used in the remainder of this article, treats the null hypothesis situation as a special case of the alternative. Statisticians describe a test which achieves minimum power at the null hypothesis as **unbiased**.

Ideally, a technique provides the **most powerful test** among a set of candidates. Typically, restrictions must hold in order to find a "best" test. One of the most appealing approaches arises from seeking a *uniformly most powerful unbiased test*. Among the class of all unbiased tests, one seeks that test, the power of which for any alternative never falls below the power of any other unbiased test. See Kendall & Stuart [11, Sections 22.16 and 23.24], and surrounding material) for further discussion.

Table 1 The traditional approach

		Decision	
		H_0: no effect	H_a: effect
Truth	No effect	Pr{Correct negative} = $(1-\alpha)$	Pr{False positive} = Pr{Type I error} = α
	Effect	Pr{False negative} = Pr{Type II error} = β	Pr{Correct positive} = $(1-\beta)$ = Power

Sensitivity Analysis for Power

Scientists encountering the concept of power sometimes ask a statistician "Just tell me how many subjects I need" (*see* **Sample Size Determination**). The question presumes the availability of defensible choices for unknown parameters, a precise size of a scientifically important difference, a clear specification of the costs of alternative designs, and a clear specification of costs and benefits of correct and incorrect decisions. The many dimensions of uncertainty imply the need to consider a wide range of alternatives. Considering a range of values creates a **sensitivity analysis** by examining the sensitivity of the study power to the assumptions.

The power of a t test as a function of mean difference is illustrated in Figure 1. The curve summarizes results of Warshauer et al. [25] who examined the value of a contrast agent in CT images of liver cancer. A subject's score equals the difference between a baseline and IV contrast enhanced CT. The horizontal axis represents the difference between group (difference) means in Hounsfield units (HU), a measure of brightness of an object in a CT. The scientists had used a t test to assess the value of adding a particular contrast agent to a patient's IV. The most important test was not significant. Hence, they computed the power curve in order to assess whether the study had possessed enough power to detect clinically relevant effects. A difference between the groups of 10 HU was deemed clinically significant. The computations assume five subjects in the control group, four in the experimental group, and $\alpha = 0.05$. The solid power curve uses the observed value of $\hat{\sigma}^2 = 6.78$ (based on seven degrees of freedom). As a function of a **random variable** (the variance estimate), a computed power value becomes a random variable. Taylor & Muller [23] described how to create an exact confidence interval for such a power value. The dashed line provides a one-sided 95% **confidence** region for the power values. The distance between the solid and dashed lines indicates the uncertainty in the power computation due to having used $\hat{\sigma}^2$ in lieu of σ_2. The 95% quantile of χ^2 (7) implies that $6.78 \times (14.067/7) \approx 6.78 \times 2$ provides an upper bound estimate of σ^2. Using this value in the power calculations would provide approximately the exact answer embodied in the dashed line. However, the inexact approximation should never be used because the exact answer requires no additional complexity to calculate (see Taylor & Muller [23, Section 2.2] for details). A similar looking plot would occur with other types of analyses, such as the comparison of groups on a **binary** response (such as in **logistic regression**) or with time-to-event **(survival) analysis**.

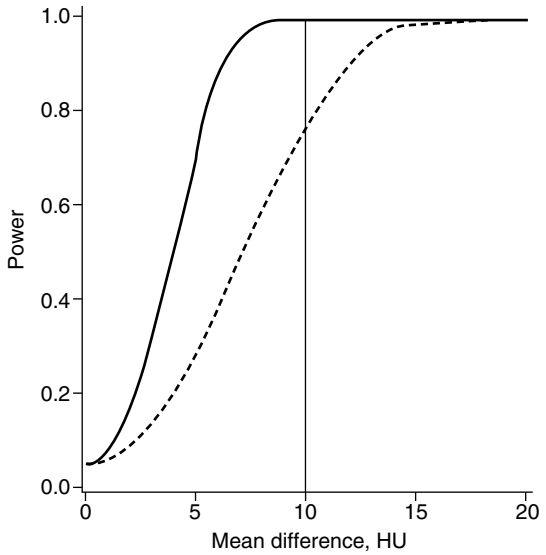

Figure 1 Power as a function of mean difference (after Warshauer et al. [25])

The example demonstrates many properties of a reasonable power analysis. For any Gaussian theory linear model analysis, such as the t test example, power varies only with (i) sample sizes, (ii) error variance, and (iii) mean differences, for a (iv) fixed α. More generally, power depends on (i) design properties, (ii) **nuisance parameters**, (iii) difference parameters, and (iv) test characteristics. Power tables and plots usually allow one or two dimensions to vary. The investigator controls and knows the values of design properties, such as total sample size and ratios of group sizes, as well as test characteristics. In contrast, the investigator does not know, with any certainty, the values of nuisance parameters and treatment differences. Hence, the latter provide the critical and interesting dimensions for plots and tables in a power analysis. A reasonable sensitivity analysis also involves varying the investigator-controlled dimensions, especially sample size.

Aligning Power

The potentially most damaging errors that commonly occur in power analysis involve misalignment. Describe a power analysis as *aligned* if the test examined in power analysis coincides with the test used in data analysis. Here, assume that the model has been correctly specified. Alignment errors may falsely inflate or deflate the power computed. Muller et al. [19] provided examples of both, in the context of using the power for **Student's *t* test** in lieu of that appropriate for **analysis of variance for longitudinal data**. Whether or not such simple power calculations apply must be defended in each particular application.

Although oversimplification of power analysis occurs most often, other sorts of misalignment also occur. Consider ignoring dropout in planning a clinical trial lasting five years. Data missing at random lead to the wrong sample size, while data missing not at random require distinct statistical methods (*see* **Nonignorable Dropout in Longitudinal Studies**). Obviously, the data analysis must align with the study, as well as the power analysis.

Ethical and monetary costs should dictate the effort expended on power analysis. Inexpensive, quick, and risk-free studies merit only limited power analysis. In contrast, expensive and lengthy studies with risky treatments merit thorough power analysis.

Increasing Power for a Fixed Sample Size

Increasing the type I error, α, automatically increases power. Scientific reviewers would frown on any blatant move of that sort. Of course, merely ignoring the issue of **multiple comparisons** (conducting many tests of the same idea with one set of data) will inflate α and hence have the same effect. The discussion here assumes that the scientist has balanced type I and II error rates by careful choice of the type and number of tests. See Muller et al. [18] for an extended discussion of the general issue in the context of toxicology.

The choice of analysis may increase power. For example, a repeated measures analysis may be more powerful than a necessarily **Bonferroni** corrected set of tests at each time, if the effect increases rapidly in time. However, a late developing effect may lead to the Bonferroni corrected set of tests being more powerful.

The choice of response variable may substantially affect power. Obviously, one variable may vary with the **explanatory variables** (or predictors) more than another. The acquisition of more knowledge and insight often rewards the scientist in this fashion. The scale of the response may play an equally important role. For example, the nature of clinical medicine pushes physicians to think in terms of binary or **categorical** responses. The practical need to assign one among many treatments naturally leads to the desire for categorical diagnosis. In studying treatments for a bacterial infection, one obvious response would be the presence or absence of disease. If the diagnosis derives from a blood level, more power would be expected from using the continuous measure of serum concentration.

Reducing error variance provides one of the simplest techniques to increase power. Merely ensuring scrupulous adherence to recruiting criteria, study protocol, and laboratory **assay** technique help. In the context of Gaussian theory linear models, cutting error variance in half has nearly the same impact as doubling the sample size.

The distribution of observations among design points affects power. Consider choosing the set of drug doses for an **analysis of variance** type design, with drug dose as the categorical predictor of interest. A design with one quarter of the subjects at each of four doses will usually have noticeably less power than a design with half of the subjects at

the two most extreme doses. With two groups, a balanced design (one with an equal number in each predictor value group) usually has more power than an unbalanced design. With a categorical outcome, such as cancer/no cancer, power tends to depend more on the distribution across predictors and number in the smaller of the two category counts, rather than on properties of the total sample.

Computing Power

Published theory may be combined with contemporary computers to facilitate computation of exact or approximate power for an extremely wide range of statistical methods. Unfortunately, currently available **software** does not fulfill the promise. Many commercial vendors provide user-friendly packages. At least until the middle of the 1990s, commercial products limited coverage only to the most common statistical techniques. Many statisticians have also written "freeware", with a similar target. Other statisticians have developed software for particular sets of complex, but narrow, applications.

Tables in books and especially journal articles may provide the most convenient current source for power values. Note that the available books tend to follow the same model as the commercial software, limiting coverage to the most popular techniques.

Using computer software and printed material for power requires a cautious approach. The issue of alignment must be addressed. More importantly, the conceptual complexity of power analysis, and the limited availability of training, leads to more mistakes than in data analysis. The user interface, whether labels for rows and columns of a table, or option names for a program, often proves to be the problem. The reader should also recognize that programs or associated documentation, whether sold for profit or free, may contain fundamental errors. For example, does the program need the variance or the standard deviation? Either provides plausible results when substituted for the other. Note also that power computation often proves more difficult numerically than data analysis. The competition in the market for data analysis software has gradually raised standards for numerical accuracy and ease of use. One rarely sees commercial data analysis programs printing negative variance estimates or probabilities greater than one. The market for power software has just started.

The limited bibliography provides sources for power analysis for many common statistical methods. Examine Muller et al. [18] and Muller & Benignus [16] for discussion of some of the philosophy of power use. Kraemer & Thiemann [12], Lipsey [15], and Cohen [3] provide general introductions to power methods. See Kupper & Hafner [13] and Gatsonis & Sampson [6] for an indication of limitations of using rough approximations. For a conceptual introduction to power for **general linear models**, see O'Brien & Muller [20]. Muller et al. [19] provide a more technical exposition, with particular attention to multivariate and repeated measures power. Odeh & Fox [21] provide extensive tables for univariate linear models. Gatsonis & Sampson [6] treat the **multiple linear regression** model with Gaussian predictors and response.

The reader should recognize that any reasonably thorough list of sources would cover many pages. Many widely used statistical methods in biostatistics and associated excellent papers have been omitted here for the sake of brevity. To find power information for a particular technique not covered in a general power book, consult texts describing the data analysis method of interest. For example, Agresti [1] describes power analysis methods for repeated measures categorical data (*see* **Longitudinal Data Analysis, Overview**). Next, consult the statistical literature. Be sure to examine primary statistical journals, not just scientific journals. The latter have many discussions of special case models, while the former may include more thorough treatments of general models.

Related Concepts

The concepts of power discussed here have generalizations in a wide range of scientific settings. In some cases, criteria other than power provide a superior measure of study quality. For example, in group sequential or purely **sequential** designs the expected sample size often proves more interesting. Power and sample size interact in nonobvious ways in problems centered on estimating **confidence intervals**. An ongoing debate in the study of **bioequivalence** surrounds the use of power for choosing sample size.

Prospective power analysis helps in planning studies to be conducted in the future. Retrospective power analysis may help evaluate studies previously completed. See Benignus et al. [2] for an example. Power

analysis also has value in **meta-analysis** [4, 7–10, 13, 21]. See Muller & Benignus [16], Taylor & Muller [23, 24], and Muller & Pasour [17] for further discussion.

References

[1] Agresti, A. (1991). *Categorical Data Analysis*. Wiley, New York.

[2] Benignus, V.A., Kafer, E.R., Muller, K.E. & Case, M.W. (1987). Absence of symptoms with carboxyhemoglobin levels of 16–23%, *Neurotoxicology and Teratology* **9**, 345–348.

[3] Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.

[4] Cooper, H.M. (1989). *Integrating Research: a Guide for Literature Reviews*, 2nd Ed. Sage, Newbury Park.

[5] Daniel, W.W. (1991). *Biostatistics: A Foundation for Analysis in the Health Sciences*, 5th Ed. Wiley, New York.

[6] Gatsonis, C. & Sampson, A.R. (1989). Multiple correlation: exact power and sample size calculations, *Psychological Bulletin* **106**, 516–524.

[7] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Researcher* **5**, 3–8.

[8] Glass, G.V., McGraw, B. & Smith, M.L. (1981). *Meta-Analysis in Social Research*. Sage, Newbury Park.

[9] Hedges, L.V. & Olkin, I. (1989). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando.

[10] Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage, Newbury Park.

[11] Kendall, M. & Stuart, A. (1979). *The Advanced Theory of Statistics*, Vol. 2, 4th Ed. Macmillan, New York.

[12] Kraemer, H.C. & Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Sage, Newbury Park.

[13] Kupper, L.L. & Hafner, K.B. (1989). How appropriate are popular sample size formulas? *American Statistician* **43**, 101–105.

[14] Light, R.J. & Pillemer, D.B. (1984). *Summing Up: the Science of Reviewing Research*. Harvard University Press, Cambridge, Mass.

[15] Lipsey, M.W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Sage, Newbury Park.

[16] Muller, K.E. & Benignus, V.A. (1992). Increasing scientific power with statistical power, *Neurotoxicology and Teratology* **14**, 211–219.

[17] Muller, K.E. & Pasour, V.B. (1997). Bias in linear model power and sample size due to estimating variance, *Communications in Statistics – Theory and Methods* **26**, 839–851.

[18] Muller, K.E., Barton, C.N. & Benignus, V.A. (1984). Recommendations for appropriate statistical practice in toxicology, *Neurotoxicology* **5**, 113–126.

[19] Muller, K.E., LaVange, L.M., Ramey, S.L. & Ramey, C.T. (1992). Power calculations for general linear multivariate models including repeated measures applications, *Journal of the American Statistical Association* **87**, 1209–1226.

[20] O'Brien, R.G. & Muller, K.E. (1993). A unified approach to statistical power for t-tests to multivariate models, *Applied Analysis of Variance in Behavioral Sciences*, L.K. Edwards, ed. Marcel Dekker, New York, pp. 297–344.

[21] Odeh, R.E. & Fox, M. (1991). *Sample Size Choice*. Marcel Dekker, New York.

[22] Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Sage, Newbury Park.

[23] Taylor, D.J. & Muller, K.E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model, *American Statistician* **49**, 43–47.

[24] Taylor, D.J. & Muller, K.E. (1996). Bias in linear model power and sample size calculations due to estimating noncentrality, *Communications in Statistics – Theory and Methods* **25**, 1595–1610.

[25] Warshauer, D.M., Wehmueller, M.D., Molina, P.L., Muller, K.E., DeLuca, M.C. & Lee, J.K.T. (1997). Hepatic contrast enhancement and metastatic lesion conspicuity: influence of IV glucagon and oral CT contrast, *Radiology* **292**, 394–398.

K.E. MULLER

Preclinical Treatment Evaluation

Every year the pharmaceutical industry develops thousands of new compounds in the hope that some of them will ultimately prove useful in the treatment of human disease. Before any prospective new drug can be tested in humans, it must undergo an extensive series of preclinical tests and assessments to ensure its safety and efficacy, to understand its **pharmacokinetics**, metabolism, and other chemical properties, and to determine safe and appropriate doses for human testing. Some of these tests are performed *in vitro* (that is, in test tubes and Petri dishes), while others involve laboratory animals (*in vivo* testing). The vast majority of compounds (some estimate more than 99%) are discarded long before reaching the steps of animal testing, and few eventually reach the marketplace.

Drug discovery sometimes begins with the observation that a compound from nature has some desirable action without knowing how the compound works. Penicillin is one such example. Large screening programs are still used to identify compounds active against bacteria and other disease-causing organisms. In the last few decades, the tools of combinatorial chemistry and molecular biology have become increasingly important. Combinatorial chemistry enables synthesis of many compounds to increase greatly the number available for screening. Molecular biology can sometimes be used to speed the search. For example, knowledge of how the **AIDS** (acquired immune deficiency syndrome) virus replicates may provide ideas for designing molecules that interfere with viral replication. These compounds are tested for activity against the virus *in vitro*, and, by a series of experiments, scientists attempt to find molecular structures with greatest potency.

Preclinical safety assessment requires a candidate drug to undergo a battery of tests to identify potential toxic effects. Statisticians are probably most involved in screening for *carcinogens* (cancer-causing agents) (*see* **Tumor Growth**; **Tumor Incidence Experiments**). However, procedures for identifying *teratogens* or *developmental toxicants* (agents that damage a developing fetus or child) (*see* **Teratology**) have become more quantitative in recent years, and many statisticians are now working in this area. While there are many other areas of toxicity testing; for example, neurotoxicity and genetic toxicology, this article focuses primarily on preclinical testing for carcinogenicity and developmental toxicity.

Historically, scientists have always been interested in devising *in vitro* tests that can assess drug safety quickly and cheaply without resorting to animal testing. One of the most well known is the Ames Salmonella assay for **mutagenicity**, wherein the test chemical is added to bacteria in a Petri dish. While the 1970s saw considerable optimism surrounding the Ames assay, most experts now agree that the test has fairly low **sensitivity** for predicting carcinogenicity in rodents. However, because its **specificity** is high, compounds found to be positive using the Ames assay are likely to be dropped from further development. *In vitro* assays are often supplemented with *in vivo* assays; for example, the mouse bone marrow cytogenetic assay. Recently, there has also been considerable interest in the use of computer-based algorithms that attempt to predict carcinogenicity on the basis of chemical structure and function, mutagenicity, 90-day toxicity, and other factors. However, the general consensus is that while these alternatives can provide useful insight into mechanisms of action, they cannot yet replace carcinogenicity bioassays (*see* **Biological Assay, Overview**) in rodents. Finding *in vitro* tests for teratogenicity and developmental toxicity is even more difficult, since less is known about their mechanisms of action than for carcinogenicity. Some tests (for example, assessing the motility and morphometry of sperm) show promise, but it is likely that *in vivo* tests will play a central role in drug testing for some time to come.

General Principles for Animal Testing

Regardless of whether the endpoint is carcinogenicity, teratogenicity, neurotoxicity, or some other kind of adverse effect, toxicologists for the most part concur on several general principles for designing a high-quality animal experiment. First, the experiment should be conducted in a familiar animal strain and under stable experimental conditions to avoid extra sources of variability that might confound study results. Secondly, the route of exposure should be the one that most closely mimics the intended clinical route(s) to humans. For example, a compound intended for intramuscular (IM) injection or intravenous (IV) administration to humans should usually

be given by the same route to laboratory animals. A drug intended for oral administration would be given to animals by gavage. A third and somewhat controversial principle is that the highest experimental dose should correspond to the maximum tolerated dose (MTD) for long-term studies (*see* **Phase I Trials**). Loosely speaking, this is the highest dose that can be administered without the experimental animals experiencing excessive systemic toxicity that could alter the test results. Precise definition of the MTD depends on the specific testing situation. The primary reason for using the MTD is to maximize the chance of detecting a carcinogenic effect, if one exists. If low doses were used instead, then experiments would need to be much larger to achieve adequate statistical **power** to detect effects. To illustrate, Table 1 shows the **sample sizes** that would be required in order to have a 95% chance of seeing at least one tumor, as a function of the true probability that an animal has a tumor:

Once pilot studies have been conducted to identify the MTD, the next choice is the number of dose groups to use, and the number of animals in each. While several different choices might be justified, most long-term experiments tend to include two or three dose groups between control and MTD, with the lowest at 1/4 (MTD) and the higher at 1/2 (MTD).

Caging is another important issue in experimental design. Specifics vary according to the type of experiment being run. For a carcinogenicity experiment, for example, some laboratories house multiple animals together in the same cage. While this practice is cost effective, it can lead to problems if the animals fight (male mice are particularly prone to this problem). Allocation of cages within the animal room is another important consideration. Because animals are sensitive to light, noise, heat, and humidity, **biases** could result if, for example, all the high-dose animals were placed on the top rack of cages. Possible approaches include completely random allocation (*see* **Randomized Treatment Assignment**) or cage rotation. However, there are practical constraints to both these strategies. Because light, heat, and humidity gradients, if they exist, are likely to be vertical, the most popular approach is a clustered block design (*see* **Blocking**) wherein each column of cages consists of just one dose group, and columns are randomly allocated to dose groups.

Table 1 Sample sizes required to observe at least one tumor with 95% probability

Pr(tumor)	Sample size needed
0.1	30
0.01	300
0.001	3000
0.0001	30000
0.00001	300000

Diet (or body weight) has received considerable attention in recent years, since food consumption is an important **confounder** in animal studies. Rodents given unlimited access to food can become obese, resulting in decreased life span, earlier onset of spontaneous tumors, and possibly altered drug metabolism and consequent incidence of treatment-induced tumors. Such considerations also add to the argument for single animal caging, because multiple animals per cage often lead to large variations in the food consumption of individual animals.

Carcinogen Bioassays

In a typical long-term rodent bioassay, control and exposed animals are followed over 18 to 24 months, and are examined at death or sacrifice for the presence of a variety of different tumors, as well as nonneoplastic lesions. Usually, the experiments are conducted in male and female animals from two different species (for example, Fischer-344 rats or $B6C3F_1$ mice), with 50 to 60 animals per dose group, sex and strain, for a total of approximately 800–960 animals (for a study with a **control** and three dosed groups). This long-term experiment is generally preceded by two shorter-term pilot studies whose main goals are to establish the doses to be used in the 24-month study and to study acute and subchronic effects.

In the early days of the carcinogen bioassay, statistical analysis was naively based on comparisons of the tumor counts observed in each dose group. For example, suppose that Table 2 represents the

Table 2 Summary data from a typical carcinogenicity study

	Dose level				
	d_0	d_1	...	d_I	Total
Number of animals with tumor	x_0	x_1	...	x_I	$x_.$
Number exposed	n_0	n_1	...	n_I	$n_.$

experimental data. Then, one could test for a dose effect using a Cochran–Armitage trend test (*see* **Trend Test for Counts and Proportions**), which can be written as

$$\chi^2_{\text{ca}} = \frac{\left[\sum_{i=0}^{I} d_i(x_i - \text{E}_i)\right]^2}{V}, \qquad (1)$$

where $\text{E}_i = n_i x_\cdot / n_\cdot$ is the expected number of tumors in the ith dose group under the **null hypothesis** of no dose effect (H_0), and

$$V = \left(\frac{x_\cdot}{n_\cdot}\right)\left(\frac{n_\cdot - x_\cdot}{n_\cdot}\right)\sum_{i=0}^{I} n_i(d_i - \overline{d})^2.$$

Under H_0, χ^2_{ca} has an asymptotic **chi-square distribution** with one **degree of freedom**. Eventually, it became apparent that such analyses based on *lifetime tumor incidence* could be biased when lifetimes were shortened due to toxicity associated with high experimental doses. Several age-adjusted tests were proposed in the early 1970s, including the *tumor prevalence test* [9], which involves stratifying time into three or four intervals (*see* **Stratification**) and then comparing exposed and controls with respect to $\pi_j = \Pr(\text{tumor-death in interval } j)$ using a **Mantel–Haenszel** test. Dinse & Lagakos [6] recognized that a prevalence analysis can be easily accomplished using **logistic regression** with tumor presence as the outcome and dose and time as **covariates**. Although the prevalence test is still widely used, many are concerned about its strong implicit assumption of nonlethality (that is, tumor onset does not change an animal's risk of death). Alternatives include a **survival analysis** treating death with tumor as the outcome [14]. However, this approach assumes that the tumor is instantly lethal. Peto et al. [15] proposed a method that relies on pathologists to assign *cause of death*. The problem with the Peto method, however, is that *cause-of-death* information is often unavailable, and can be unreliable. Much statistical research in the past 10 or 15 years has focused on the development of methods that apply to tumors of intermediate lethality. *Poly-3* is a promising method that replaces the n_i in the above table with an adjusted value, N_i, that downweights animals that died early with no tumors [1]. More precisely,

$$N_i = x_i + \sum\left(\frac{t_{ij}}{t_{\max}}\right)^3,$$

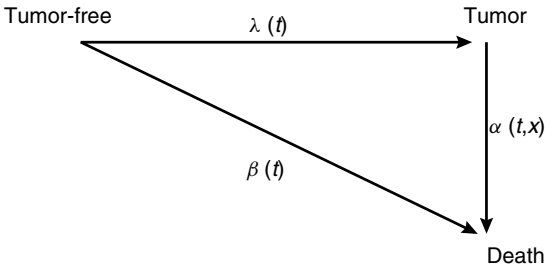

Figure 1 Three-state model for carcinogenicity

where $t_{\max}$ is the maximum death time observed in the study and the sum is over the subset of animals in the ith dose group that died with no tumor. The third power is chosen because generally it has been found that tumor **incidence rates** tend to occur as a third- or fourth-order function of time. Alternately, many have proposed methods of analysis based on fitting a *three-state model* as depicted in Figure 1, with $\lambda(t)$ representing the instantaneous rate of tumor onset at time t, $\beta(t)$ the instantaneous death rate at time t, and $\alpha(t, x)$ the instantaneous death rate at time t for an animal that developed a tumor at time x.

Many approaches to fitting the three-state model are impractical because they assume the availability of extensive interim sacrifice (wherein randomly selected animals are killed at prechosen times during the experiment and examined for tumors). Recently, Dinse [5] and Lindsey & Ryan [13] have advocated the use of **semiparametric regression** models that assume an additive $[\alpha(t, x) = \beta(t) + \Delta]$ or multiplicative relationship $[\alpha(t, x) = \beta(t)\text{e}^{\Delta}]$ between the **hazards** for death with and without tumor. These approaches are appealing from a biological perspective and can be applied to a typically sized experiment with only a single terminal sacrifice.

The carcinogen bioassay raises several other challenging statistical problems. One is the issue of **multiplicity**, wherein the chance of seeing statistically significant results is increased because many (usually over 50) different tumor sites are examined (*see* **Multiple Comparisons**). Some propose **multivariate analysis** methods that accommodate multiple tumor types [18]. Others argue that multiple testing is not a serious problem for tumor types that are rare in control animals, although this depends on sample size [8]. The role

of historical control information in the evaluation of data from a carcinogenicity study is also an important topic with interesting statistical questions that have been discussed by many authors in recent years [10].

Teratology Studies

The purpose of a teratology, or, more generally, a developmental toxicology, study is to assess adverse effects on reproduction and development. In response to the thalidomide episode and growing concern about prenatal effects of drugs, the US **Food and Drug Administration (FDA)** issued a set of guidelines in 1966 for conducting animal studies to assess the safety of drugs intended for human use. The guidelines established a three-stage testing procedure generally referred to as segments I, II and III. "Segments" refer to sequential periods during reproduction when the compound is administered. In a segment I (fertility) study, the compound is given prior to and during the mating phase and up to implantation of the embryo. In a segment II (teratology) study, the compound is administered during the major period of organogenesis. In a segment III (late gestation and lactation) study, exposure occurs from the end of the embryonic period through lactation. Subsequently, the FDA added a multigenerational protocol suitable for testing food additives and pesticides. More recently, the guidelines have been modified to allow more flexible study designs that combine segments, so long as all phases of reproduction, premating through lactation, are covered.

The remainder of this section focuses on the segment II design, since this is the one to which most of the published statistical work relates. Figure 2 summarizes the typical design, wherein pregnant animals (dams) are exposed during the critical period of major organogenesis (days 6–15, 17, or 19 for mice, rats and rabbits, respectively) and sacrificed just prior to normal delivery at which time the uterus is removed and the contents are thoroughly examined (note that "gd" denotes gestational day in Figure 2).

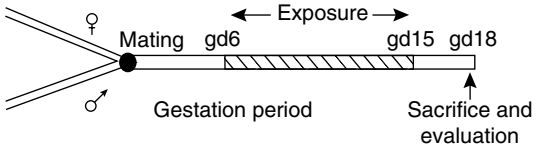

Figure 2 Chronological events in a developmental toxicity study

Ordinarily, between 20 and 30 pregnant dams are randomized to each dose group and control and typical control litter sizes (number of live-born offspring) range from eight to 10 in rabbits to 12 and 15 in mice and rats, respectively.

While the dam is the unit of randomization (*see* **Unit of Analysis**) and maternal toxicity is an important factor in these studies, interest centers primarily on fetal outcomes. Examination of the uterine contents produces several discrete and continuous fetus-level outcomes. In addition to the number of eggs that implant in the uterus, litter-specific counts of resorptions (early fetal deaths) and fetal deaths (deaths occurring later in gestation) are available. Viable fetuses (animals that would be born live if the dams were followed to term) are extensively examined for malformations. Body weight and, occasionally, other parameters, such as various body dimensions (e.g. crown to rump length) and organ weights, are also measured.

The statistical analysis of data from a teratology experiment must account for the *litter effect*, or the tendency for littermates to respond more similarly than offspring from different litters. One easy approach is to summarize data at the dam level; for example, by considering litter-specific percentages of resorptions or malformations as the endpoints for analysis. More recently, there has been considerable interest in the use of statistical methods for correlated data that allow for fetus-specific analyses. Some of these have been based on **likelihood**-based methods, such as the **beta-binomial** [3]. Others have been based on **generalized estimating equations**, which provide a relatively simple and often robust (*see* **Robustness**) approach to the analysis of **correlated binary data**. For example, Lefkopoulou & Ryan [11] propose an appealing generalization of the Cochran–Armitage trend test that applies to clustered binary data. To illustrate, suppose, for simplicity, that each fetus is classified as normal or abnormal. Let x_{ij} be the number abnormal among the n_{ij} pups in the jth litter of the ith dose group. Then, under the null hypothesis of no dose effect, the following test statistic has an asymptotic chi-square distribution with one

degree of freedom:

$$\frac{\left[\sum_i \sum_j d_i(x_{ij} - n_{ij}\tilde{\mu})\right]^2}{\sum_i \sum_j (d_i - \overline{d})^2 (x_{ij} - n_{ij}\tilde{\mu})^2},$$

where d_i is the dose level applied to the ith dose group,

$$\overline{d} = \frac{\sum_i \sum_j n_{ij} d_{ij}}{\sum_i \sum_j n_{ij}} \quad \text{and} \quad \tilde{\mu} = \frac{\sum_i \sum_j x_{ij}}{\sum_i \sum_j n_{ij}}.$$

Note that when $n_{ij} = 1$ (i.e. no litter effect), this test statistic is asymptotically equivalent to the Cochran–Armitage trend test, (1).

Handling multiple outcome data is another important issue that arises in the analysis of developmental toxicity data (*see* **Multiple Endpoints, *P* Level Procedures**). The issues here are slightly different from those encountered in carcinogenicity, since compounds that cause adverse effects are quite likely to affect a variety of organ systems in the developing fetus. Some authors [4, 16, 19] address the hierarchical nature of some outcomes (e.g. death and malformation), and suggest analyses based on **overdispersed multinomial** models. Others have suggested methods to deal with the analysis of multiple outcomes measured on the same fetus [12], as well as the challenge of analyzing the mixture of discrete and continuous outcomes that arise in developmental toxicity [2, 7].

Neurotoxicity and Other Studies

Neurotoxicity has traditionally been evaluated by histopathologic examination of slides taken at necropsy, much as other forms of toxicity. In recent years, assays that directly measure neurologic function in living animals have been increasingly used, since neurological damage may not be evident at necropsy, but may manifest later with behavioral problems [17]. For example, a swim maze may be used to measure problem-solving ability (time to find the exit), and memory (time to find the exit a few days later). An open field monitor records how a rat explores a new environment (an open box crisscrossed with electronic beams) by counting the number and location of beams broken over time. Other assays measure righting reflex, balance, or habituation to repeated loud noise. Such response data are often high-dimensional, and yet can show clear effects, suggesting a need for multivariate statistical methods (*see* **Multivariate Analysis, Overview**) to capture and summarize the information. At this time, however, little has been published on this topic.

Future Developments

Recent years have seen a number of exciting developments in toxicology. There is great interest in the development of *biologically based models* that take account of pharmacokinetics, metabolism, and physiological systems to characterize better dose response and improve the extrapolation from animals to humans. While such models are still in relatively early stages of development, it is likely that they will become more widely used in the future. The use of transgenic mouse models is also likely to become more widespread in the future.

References

[1] Bailer, A.J. & Portier, C.J. (1988). Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples, *Biometrics* **44**, 417–431.

[2] Catalano, P.J. & Ryan, L.M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes, *Journal of the American Statistical Association* **87**, 651–658.

[3] Chen, J.J. & Kodell, R.L. (1989). Quantitative risk assessment for teratologic effects, *Journal of the American Statistical Association* **84**, 966–971.

[4] Chen, J.J., Kodell, R.L., Howe, R.B. & Gaylor, D.W. (1991). Analysis of trinomial responses from reproductive and developmental toxicity experiments, *Biometrics* **47**, 1049–1058.

[5] Dinse, G.E. (1991). Constant risk differences in the analysis of animal tumorigenicity data, *Biometrics* **47**, 681–700.

[6] Dinse, G.E. & Lagakos, S.W. (1983). Regression analysis of tumour prevalence data, *Applied Statistics* **32**, 236–248.

[7] Fitzmaurice, G.M. & Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering, *Journal of the American Statistical Association* **90**, 845–852.

[8] Haseman, J.K. (1990). Use of statistical decision rules for evaluating animal carcinogenicity studies, *Fundamental and Applied Toxicology* **14**, 637–648.

[9] Hoel, D.G. & Walburg, H.E. (1972). Statistical analysis of survival experiments, *Journal of the National Cancer Institute* **49**, 361–372.

[10] Ibrahim, J.G. & Ryan, L.M. (1996). Use of historical controls in time-adjusted trend tests for carcinogenicity, *Biometrics* **52**, 1478–1485.

[11] Lefkopoulou, M. & Ryan, L. (1993). Global tests for multiple binary outcomes, *Biometrics* **49**, 975–988.

[12] Lefkopoulou, M., Moore, D. & Ryan, L. (1989). The analysis of multiple correlated binary outcomes: application to rodent teratology experiments, *Journal of the American Statistical Association* **84**, 810–815.

[13] Lindsey, J. & Ryan, L. (1993). A three-state multiplicative model for rodent tumorigenicity experiments, *Applied Statistics* **42**, 283–300.

[14] Peto, R. (1974). Guidelines on the analysis of tumor rates and death rates in experimental animals, *British Journal of Cancer* **29**, 101–105.

[15] Peto, R., Pike, M., Day, N., Graph, R., Lee, P., Parish, S., Peto, G., Richard, S. & Wahrendorf, J. (1980). Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments, *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, Suppl. 2: Long-term and Short-term Screening Assays for Carcinogens: a Critical Appraisal*. International Agency for Research on Cancer, Lyon, pp. 311–346.

[16] Ryan, L.M. (1992). Quantitative risk assessment for developmental toxicity, *Biometrics* **48**, 163–174.

[17] US EPA (1995). Proposed guidelines for neurotoxicity risk assessment, *Federal Register* **60**, 52032–52056.

[18] Westfall, P. & Young, S.S. (1992). *Resampling-Based Multiple Testing*. Wiley, New York.

[19] Zhu, Y., Krewski, D. & Ross, W.H. (1994). Dose-response models for correlated multinomial data from developmental toxicity studies, *Applied Statistics* **43**, 583–598.

Bibliography

The following books and articles provide an additional source of further reading on statistical methods for preclinical assessment of drug efficacy and safety.

Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research*. Vol. III: *The Design and Analysis of Long-Term Animal Experiments*. Oxford University Press, Oxford.

Holland, C.D. & Sielkin, R.L. (1993). *Quantitative Cancer Modeling and Risk Assessment*. Prentice-Hall, Englewood Cliffs.

Hood, R.D. (1996). *Handbook of Developmental Toxicology*. CRC Press, New York.

Hubert, J.J. (1996). *Environmental Risk Assessment*. University of Guelph Printing Services.

Krewski, D. & Franklin, C. (1991). *Statistics in Toxicology*. Gordon & Breach, New York.

Morgan, B.J. (1992). *Analysis of Quantal Response Data*. Chapman & Hall, London.

Morgan, B.J., ed. (1996). *Statistics in Toxicology*. Oxford University Press, London.

LOUISE M. RYAN & KEITH A. SOPER

Prediction

In statistical usage, the term *prediction* usually refers to the attempt to predict individual values of a random variable based on a statistical model, often one which relates a **response variable**, which is to be predicted, to a set of **explanatory variables**. The errors involved in prediction usually derive from uncertainty in the estimation of the mean, or some function of the mean, of the variable, and the "natural" variation of the observed value about its mean. The former is usually determined by the adequacy of an estimated statistical model, while the latter is not subject to manipulation. Prediction based on **multiple linear regression** provides an example.

Examination of the predictive ability of a statistical model is useful in a variety of situations. It is, for example, a natural aspect of **discriminant analysis** and the basis of **cross-validation**. Geisser [2, 3] has written extensively on prediction as a basis of inference procedures. A book-length treatment of prediction is [1].

In some other contexts, *prediction* takes on its more usual meaning of "predicting the future" (*see* **Forecasting**; **Predictive Modeling of Prognosis**).

References

[1] Aitchison, J. & Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.

[2] Geisser, S. (1971). The inferential use of predictive distributions, in *Foundations of Statistical Inference*, V.P. Godambe & D.A. Sprott, eds. Holt, Rinehart & Winston, Toronto, pp. 456–469.

[3] Geisser, S. (1980). A primer on predictivism, in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner, ed. North-Holland, Amsterdam, pp. 363–381.

VERN T. FAREWELL

Predictive Modeling of Prognosis

Changing **prognosis** over time is an essential feature of the course and treatment of chronic diseases such as cancer, HIV/**AIDS**, leukemia, and diabetes. Typically, when the patient presents initially he/she can be characterized as having a certain state of health, which the physician determines using various objective and subjective measurements. Whether the patient remains in the hospital for continuous monitoring or leaves and returns for periodic observation and treatment, many of his/her measurements will change over time. While some apparent clinical changes in the patient might be attributable to **measurement error**, others may reflect real beneficial treatment effects or a worsening of prognosis.

Having arrived at a diagnosis, the physician can assess the natural course of the disease, and, on the basis of this information, can make judgments about the patient's prognosis. Prognosis is defined in terms of the relative probability of developing one of several alternative outcomes in the natural history of the disease (*see* **Natural History Study of Prognosis**).

In recent years, **Markov models** have become important tools to describe and help understand the progression and regression of chronic diseases which are characterized by having well-defined stages. These models have been used to find possible markers for the transition from stable states to the accelerated phase and the irreversible (absorbing) state of a disease, and also to describe the natural course of these diseases. Aalen and Johansen [1] considered multistate continuous-time Markov processes assuming that transitions occurred just before the observation times (*see* **Aalen–Johansen Estimator**). Klein et al. [17] use a three-state **semi-Markov** model in a study of patients with chronic myelogenous leukemia to analyze the effect of elevated blood levels of adenosine deaminase as a marker for transition from stable disease to blast crisis, and then to death. Andersen and colleagues [2–4] introduce proportional intensity **regression** models for multistate Markov processes assuming that the transition times are known as in **proportional hazard** models [7]. Longini et al. [18] use a Markov model to describe the distribution of the **incubation period** for AIDS patients. In this model, only progression to the next higher state or a jump to the absorbing state are allowed.

Kay [16] proposed a methodology to fit a general k-disease-state Markov model in continuous time with application to the analysis of cancer markers in **survival** studies. Kay assumes that the transitions between states occur at unknown times between observations, except for transitions to the absorbing state. Andersen et al. [4] compare the approach of Aalen and Johansen [1], which assumes that the transition times are known, with the approach of Kay [16], which assumes that the transition times are unknown. Grüger et al. [13] looked at the interrelationship between the disease process and the patient's examination scheme (i.e. the study protocol).

This article discusses predictive models for prognosis using a general k-state Markov model in which the exact transition times are not observed (*see* **Transition Models for Longitudinal Data**). An important application of these models is discussed and analyzed. Data from a longitudinal study in young patients with diabetes from the Barbara Davis Center for Childhood Diabetes, University of Colorado Health Sciences Center, are used to determine factors affecting the natural course of diabetic retinopathy.

Three Examples

Hepatocellular Carcinoma

Kay [16] shows the results of a multistate model (*see* **Multistage Carcinogenesis Models**) applied to 81 patients with hepatocellular carcinoma. The main goal of this statistical analysis was to determine the usefulness of serum alpha-fetoprotein (AFP) as a marker in hepatocellular carcinoma. High levels of AFP are indicative of such cancers, particularly in subjects with cirrhosis of the liver, at the diagnosis state. Kay [16] shows that when the presence of cirrhosis is taken in account, AFP levels at the diagnosis time no longer predict the survival time. However, a multistate model showed that changes in AFP levels are related to risk of death.

The patients were treated mainly by Adriamycin, but a few cases underwent resection or transplantation. AFP levels were measured on all patients at the start of the treatment and subsequently at convenient time points during the follow-up period.

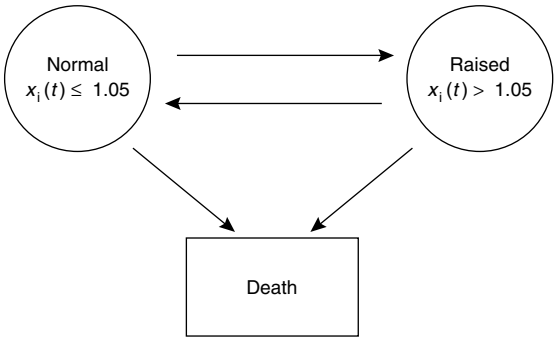

Figure 1 The three-state model for hepatocellular carcinoma

If $z_i(t)$ represents the AFP value recorded at time t for the ith patient, then the transient stages of the disease are defined based on 5% increase of $\log z_i(t)$ over the baseline value, as shown in Figure 1; this is

$$x_i(t) = \frac{\log z_i(t)}{\log z_i(0)}$$

HIV Infection

The understanding that individuals infected with human immunodeficiency virus (HIV) pass through various stages from infected but antibody-negative to acquired immune deficiency syndrome (AIDS) diagnosis implies that the mathematical modelling of the infection process is most naturally carried out using a multistate model.

To analyze the natural course of this disease, Longini et al. [18] use data on the progression of individuals to AIDS arranged in cohorts of persons recently infected whose serum specimens were found to be positive. The multistate model for HIV infection was defined using five states, as shown in Figure 2. The first stage is HIV infection but with antibody-negative status, the second is antibody-positive status but asymptomatic. The third stage occurs when an individual develops an abnormal hematologic indicator and/or prodromal illnesses (pre-AIDS symptoms), such as persistent generalized lymphadenopathy and oral candidiasis. The fourth stage is clinical AIDS. A fifth stage is included in the model to represent death due to AIDS.

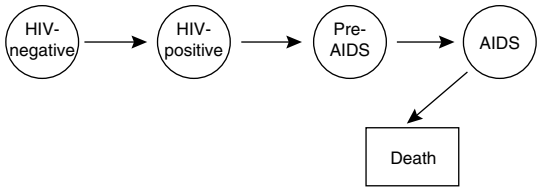

Figure 2 The multistate model for four stages of HIV infection

The Natural Course of Diabetic Retinopathy

Marshall and colleagues [19, 20] showed the results of implementing a multistate model for diabetic retinopathy in patients with type I diabetes. The study data came from the follow-up of 277 subjects who had type I diabetes for at least 5 years and ranged in age from 14 to 29 years when initially seen at the Eye-Kidney Clinic of the Barbara Davis Center for Childhood Diabetes at the University of Colorado Health Sciences Center. The Eye-Kidney Clinic was open to all patients 14 years of age or older who have had type I diabetes for at least three years. The average duration of insulin-dependent diabetes mellitus for this population was approximately 10 years, ranging from 3 to 28 years. The sex ratio was approximately one-to-one.

All subjects were seen longitudinally at least twice, with visits at an average of 1 year apart, for a mean follow-up of three years. A total of 882 patient visits occurred during the study period, an average of 3.2 visits per subject. At each visit, a retinal specialist graded retinal findings using a modified Airlie House classification [11] in which grade I indicates no retinopathy; grade II indicates microaneurysms only; grades III and IV indicate intermediate stages of background retinopathy; and grades V and VI indicate pre-proliferative and proliferative retinopathy, respectively. The worst eye grade for each visit was used to define the subject's state at the time of the visit. A four-state model was proposed to study the natural history of this disease, as shown in Figure 3.

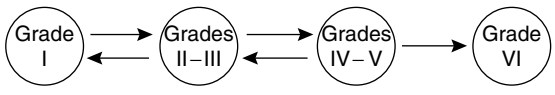

Figure 3 The multistate model for four stages of diabetic retinopathy defined on the basis of eye grades according to the Airlie House classification

Continuous-time Markov Processes

The theory of continuous-time, finite-state Markov processes is available from many sources (see, for example, [6] or [8]). Suppose we observe a continuous-time **stochastic process** $\{X(t), t \geq 0\}$ with a finite number of values in $E = \{1, 2, \ldots, k\}$ called states. We say that $\{X(t)\}$ is a continuous-time Markov process if for all times $t > s > u > 0$, and for any states i, j and $h \in E$,

$$\begin{aligned}\Pr\{X(t) = j | X(s) = i, X(u) = h\} \\ = \Pr\{X(t) = j | X(s) = i\}. \end{aligned} \quad (1)$$

This **conditional probability** represents the probability of a transition from the state i at time s to the state j at time t, and it is denoted as $p_{ij}(s, t)$. Clinically speaking, this property says that the clinical history of a patient with the disease is summarized entirely by the current stage of the disease. These transition probabilities have the basic properties, $0 \leq p_{ij}(s, t) \leq 1$, $p_{ii}(t, t) = 1$, $p_{ij}(t, t) = 0$ if $j \neq i$, and

$$\sum_{j=1}^{k} p_{ij}(s, t) = 1.$$

For any time τ in the interval (s, t) the transition probability $p_{ij}(s, t)$ can be written, using the Chapman–Kolmogorov equation, as

$$p_{ij}(s, t) = \sum_{v=1}^{k} p_{iv}(s, \tau) p_{vj}(\tau, t).$$

This equation, in matrix notation, is

$$\mathbf{P}(s, t) = \mathbf{P}(s, \tau)\mathbf{P}(\tau, t),$$

where $\mathbf{P}(s, t)$ is the transition probability matrix of dimension $k \times k$ with elements $p_{ij}(s, t)$.

The Markov process $X(t)$ is characterized by the transition intensities

$$\lambda_{ij}(t) = \lim_{\mathrm{d}t \to 0} \frac{\Pr\{X(t + \mathrm{d}t) = j | X(t) = i\}}{\mathrm{d}t}, \quad i \neq j$$

which represent instantaneous transition rates between the different states. For mathematical convenience,

$$\lambda_{ii}(t) = -\sum_{j \neq i}^{k} \lambda_{ij}(t).$$

The transition probability, $p_{ij}(s, t)$, satisfies the Kolmogorov forward differential equations

$$\frac{\mathrm{d}p_{ij}(s, t)}{\mathrm{d}t} = \sum_{v=1}^{k} p_{iv}(s, t)\lambda_{vj}(t),$$

or, in matrix notation,

$$\frac{\mathrm{d}\mathbf{P}(s, t)}{\mathrm{d}t} = \mathbf{P}(s, t)\mathbf{\Lambda}(t), \quad (2)$$

with the initial condition $\mathbf{P}(t, t) = \mathbf{I}$, where $\mathbf{I}$ is the $k \times k$ identity matrix.

Important mathematical simplifications are obtained by assuming that the process is homogeneous in time. The consequences of this assumption are that the transition intensities between the different states $\lambda_{ij}(t)$ are constant over time, and the transition probabilities $p_{ij}(s, t)$ depend only on the time differences $t - s$. Eq. (2) reduces to a system of differential equations with constant coefficients,

$$\frac{\mathrm{d}\mathbf{P}(t - s)}{\mathrm{d}t} = \mathbf{P}(t - s)\mathbf{\Lambda},$$

for which the closed-form solution is

$$\mathbf{P}(t - s) = \mathrm{e}^{\mathbf{\Lambda}(t-s)} = \sum_{n=0}^{\infty} \frac{\{\mathbf{\Lambda}(t - s)\}^n}{n!}.$$

If $\mathbf{\Lambda}$ has distinct **eigenvalues**, $\rho_1, \rho_2, \ldots, \rho_k$, and $\mathbf{A}$ is a square matrix where the jth column is the right **eigenvector** associated with ρ_j, then

$$\begin{aligned}\mathbf{P}(t - s) = \mathbf{A}\ \mathrm{diag}\{\exp[\rho_1(t - s)], \exp[\rho_2(t - s)], \\ \ldots, \exp[\rho_k(t - s)]\}\mathbf{A}^{-1}. \end{aligned} \quad (3)$$

Individual transition probabilities can be calculated, for any value of $t - s$, as

$$p_{ij}(t - s) = \sum_{v=1}^{k} a_{iv} \exp[\rho_v(t - s)]a^{vj}, \quad (4)$$

where a_{ij} and a^{ij} represent the elements (i, j) of the matrices $\mathbf{A}$ and $\mathbf{A}^{-1}$, respectively.

The Model

Suppose we have a model with $k - 1$ transient disease states $j = 1, \ldots, k - 1$ and a single absorbing state $j = k$, as in Figure 4. The transient states

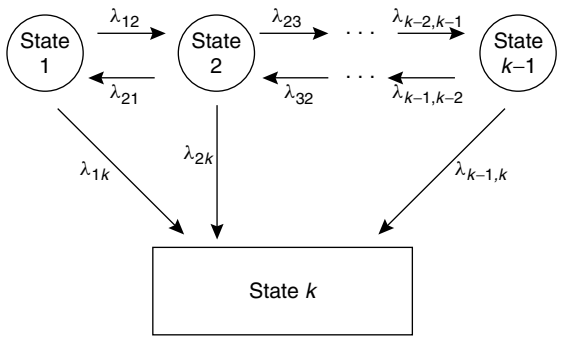

Figure 4 A multistate model with $k-1$ transient states and one absorbing state. The model has a total of $3k-5$ parameters

are assumed to be ordered according to j and instantaneous transitions can take place from state j to the adjoining states $j-1$ or $j+1$. Transitions can also take place from any of the transient states to the absorbing state k.

Submodels can be obtained by eliminating some of the parameters when some of the transitions are theoretically impossible or are unlikely to be observed during the study time. Longini et al. [18] use a submodel to describe the incubation period of AIDS with only progression transitions to the adjoining states (see Figure 2).

For the model in Figure 4 the transition intensity matrix $\mathbf{\Lambda}$ can be written as

$$\mathbf{\Lambda} = \begin{pmatrix} -(\lambda_{12}+\lambda_{1k}) & \lambda_{12} & \dots & 0 & \lambda_{1k} \\ \lambda_{21} & -(\lambda_{21}+\lambda_{23}+\lambda_{2k}) & \dots & 0 & \lambda_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -(\lambda_{k-1,k-2}+\lambda_{k-1,k}) & \lambda_{k-1,k} \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{5}$$

When equally spaced observations are available, and where a discrete time Markov model can be considered, a continuous-time Markov model will be more **parsimonious**. This model has $3k-5$ different parameters, in contrast to a discrete-time model with $k(k-1)$ independent parameters.

Regression Models

Suppose that each individual under study has an associated vector of **covariates** $\mathbf{z}' = (z_1, z_2, z_3, \dots, z_p)$ representing a set of characteristics and natural risk factors associated with the disease process. For a given value of this set of characteristics $\mathbf{z}$, we assume that the Markov process is homogeneous with an intensity matrix $\mathbf{\Lambda}(\mathbf{z})$ similar to the intensity matrix $\mathbf{\Lambda}$ in (5) with elements

$$\lambda_{ij}(\mathbf{z}) = \lambda_{ij} \exp(\beta_{ij}'\mathbf{z}), \tag{6}$$

where λ_{ij} represent the baseline transition rate, and β_{ij} represents the vector of regression coefficients associated with $\mathbf{z}$ for the transition from state i to state j. According to this model, for each possible transition between two states, we introduce a proportional intensity model similar to the **Cox regression model** with constant **hazard** function. Model (6) implies that each risk factor considered in the model may have a different effect in the progression and/or regression of the disease.

We must consider two types of **model** selection procedure in the context of this multistate Markov model. The first, more classical in statistical analysis, is the selection of covariates associated significantly with the progression and regression of the process. Given the large number of parameters associated with each covariate in model (6), it seems reasonable to consider a forward selection procedure. The second, more specific to this multistate model, relates to the selection of the most parsimonious representation of the association between each covariate and the disease process. Consider the case of a model with a single covariate. In the context of the four-state model for diabetic retinopathy (Figure 3), there are three natural models for representing the effect of the covariate. The first, named the saturated model, is defined as the model in which the effect of the covariate differs in each of the five disease transitions. In this model we have a total of 10 parameters, five baseline transition intensities, and five different regression coefficients.

The second model, named the progression and regression (PR) model, is defined as the model in which the effect of the covariate is the same for all progression transitions, and the same for all regression transitions. More formally, we formulate this model by assuming that the **null hypothesis** $H_0 : \beta_{j,j+1} = \beta_p$ and $\beta_{j,j-1} = \beta_r$ is true. Under this

hypothesis the number of parameters associated with each covariable reduces from five to only two. We can write the proportional intensity model (4) under this hypothesis as

$$\lambda_{ij}(z) = \begin{cases} \lambda_{ij} \exp(\beta_p' \mathbf{z}), & j = i+1, \\ \lambda_{ij} \exp(\beta_r' \mathbf{z}), & j = i-1. \end{cases} \quad (7)$$

Finally, the third model, called the progression minus regression (PMR) model, is where the effect of the covariate is the same for all progression transitions and the same, but with a sign change, for all regression transitions. Formally,

$$\lambda_{ij}(z) = \begin{cases} \lambda_{ij} \exp(\beta' \mathbf{z}), & j = i+1, \\ \lambda_{ij} \exp(-\beta' \mathbf{z}), & j = i-1. \end{cases} \quad (8)$$

Survival Analysis

A point of major interest in practical applications is the relationship between the Markov model and **survival analysis** functions, including the survival function, the hazard rate function, the median lifetime, the mean lifetime (*see* **Life Expectancy**) and the residual mean lifetime. Let T be a **random variable** which represents the lifetime of individuals in a homogeneous population. The survival function, given that the process is in state i at time $s = 0$, $S_i(t) = \Pr\{T > t | X(0) = i\}$, for a subject with covariables $\mathbf{z}$, is

$$S_i(t|\mathbf{z}) = 1 - p_{ik}(t; \mathbf{z}),$$

where $p_{ik}(t; \mathbf{z})$ is the (i, k)th element of the transition probability matrix $\mathbf{P}(t; \mathbf{z})$. The density function, expressed in terms of the survival function, $f_i(t) = -\,\mathrm{d}S_i(t)/\,\mathrm{d}t$, for a subject with covariables $\mathbf{z}$, is

$$f_i(t|\mathbf{z}) = \sum_{j=1}^{k-1} p_{ij}(t; \mathbf{z}) \lambda_{jk}(\mathbf{z}).$$

The hazard function $h_i(t|\mathbf{z}) = f_i(t|\mathbf{z})/S_i(t|\mathbf{z})$ as a function of the transition probabilities and intensities is

$$h_i(t|\mathbf{z}) = \sum_{j=1}^{k-1} \frac{p_{ij}(t; \mathbf{z})}{\sum_{v=1}^{k-1} p_{iv}(t; \mathbf{z})} \lambda_{jk}(\mathbf{z}),$$

which represents a weighted mean of the transition rates from the transient states to the absorbing state k.

The median lifetime from the transient state i to the absorbing state k is defined as the value of $t = \xi_i$ that satisfies $p_{ik}(\xi_i; \mathbf{z}) = 0.5$. The mean lifetime, $\mathrm{E}_i = \mathrm{E}\{T | X(0) = i\}$, is also a parameter of interest. Again, in terms of the transition probabilities, the mean lifetime is

$$\begin{aligned} \mathrm{E}_i(\mathbf{z}) &= \int_0^\infty S_i(t|\mathbf{z})\, \mathrm{d}t \\ &= \sum_{j=1}^{k-1} \sum_{v=1}^{k-1} a_{iv}(\mathbf{z}) \left(\frac{-1}{\rho_v(\mathbf{z})} \right) a^{vj}(\mathbf{z}), \end{aligned}$$

provided $\rho_v(\mathbf{z}) < 0$, for every state $v < k$. Finally, the residual mean lifetime, $m_i(t) = \mathrm{E}\{T - t | X(0) = i\}$, can be expressed in terms of the Markov model as

$$\begin{aligned} & m_i(t|\mathbf{z}) \\ &= \frac{\int_t^\infty S_i(u|\mathbf{z})\, \mathrm{d}u}{S_i(t|\mathbf{z})} \\ &= \frac{\sum_{j=1}^{k-1} \sum_{v=1}^{k-1} \left[\frac{-1}{\rho_v(\mathbf{z})} \right] a_{iv}(\mathbf{z}) \exp[\rho_v(\mathbf{z}) t] a^{vj}(\mathbf{z})}{1 - p_{ik}(t; \mathbf{z})}. \end{aligned}$$

The Data

The type of data collected will be different in each application and is directly dependent on the nature of the process. When the exact transition times of the process $\tau_1, \ldots, \tau_{m'}$ are available (see Figure 5), the statistical methods for estimating the parameters of the multistate model are straightforward.

Closed-form solutions for the **maximum likelihood** estimates can be derived for the basic model,

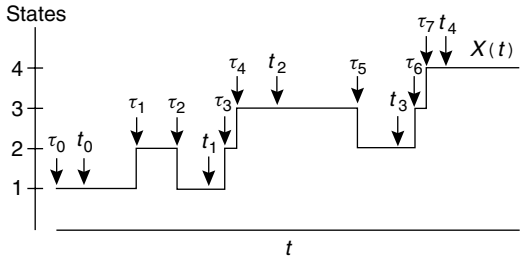

Figure 5 An example of a process with four states. The τ_i are the actual transition times and the t_i are the observation times of the process

Table 1

Patient	Data				
1	(0,2)	(41,2)	(78,1)	(95,3)	(104,4)
2	(0,1)	(17,1)	(52,4)		
3	(0,3)	(23,2)	(58,3)	(72,2)	

and an approach similar to fitting exponential regression models can be used for the model with covariables.

In clinical studies in which each realization of the process is a different patient, it is very unusual to observe the exact transition times. Data are typically only available at the times of visits to the clinic, $t_0, t_1, \ldots, t_{m+1}$, as shown also in Figure 5. We assume that the data obtained on each subject are unequally spaced in time, and that the exact transition times are not available. For a model of $k = 4$ states, the data shown in Table 1 correspond to weeks from the date of diagnosis and the state of the disease of the patient at that specific date [16]. At the date of diagnosis, patient 2 was in state 1, and 17 weeks later patient 2 was in state 1. This patient could have remained in state 1 for the whole 17 weeks, or could have transferred out of state 1 and back in again. Thirty-five weeks later, at week 52, the patient died. Survival times for these patients are 104 weeks for patient 1, 52 weeks for patient 2 and more than 72 weeks for patient 3. The data of patient 3 represent a **censored** observation.

The Likelihood Function

Suppose that the observation times for a subject are $t_0 < t_1 < \cdots < t_m < t_{m+1}$, and that $x(t_0) = i_0, x(t_1) = i_1, \ldots, x(t_{m+1}) = i_{m+1}$ represent the observed states of the process at these times. Then the joint distribution of $X(t_1), X(t_2), \ldots, X(t_{m+1})$ given $X(t_0)$ and the vector of covariables $\mathbf{z}$ can be represented, using the Markov property (1) and conditional probabilities, as

$$\prod_{j=1}^{m+1} p_{i_{j-1},i_j}(\mathbf{\Delta} t_j; \mathbf{z}), \tag{9}$$

where $\mathbf{\Delta} t_j = t_j - t_{j-1}$.

The above expression represents the contribution to the **likelihood** function for this subject if the arrival time in the absorbing state is interval-censored, i.e. if the time of transition to the absorbing state is unknown. In survival studies, t_{m+1} may represent the exact arrival time in the absorbing state k, which is the lifetime, or it may represent the end of the study for this subject, which is the censoring time.

Let $T = \tau$ be the exact arrival time in state k, and c be the censoring time for this subject. Then $t_{m+1} = \min(\tau, c)$ and

$$\delta = \begin{cases} 1, & \text{if } \tau \le c, \\ 0, & \text{if } \tau > c. \end{cases}$$

If $\delta = 1$, then the contribution of this last transition to the likelihood is

$$f_{i_m}(\mathbf{\Delta} t_{m+1}|\mathbf{z}) = \sum_{j=1}^{k-1} p_{i_m,j}(\mathbf{\Delta} t_{m+1}; \mathbf{z})\lambda_{jk}(\mathbf{z}),$$

and if $\delta = 0$ the contribution is $S_{i_m}(\mathbf{\Delta} t_{m+1}|\mathbf{z}) = 1 - p_{i_m,k}(\mathbf{\Delta} t_{m+1}; \mathbf{z})$. The above expression for $\delta = 1$ is a continuous-time version of Kay's likelihood contribution [16].

The likelihood function for this subject can then be written as

$$L_h(\theta) = \prod_{j=1}^{m} p_{i_{j-1},i_j}(\mathbf{\Delta} t_j; \mathbf{z})\{f_{i_m}(\mathbf{\Delta} t_{m+1}|\mathbf{z})\}^{\delta} \times \{S_{i_m}(\mathbf{\Delta} t_{m+1}|\mathbf{z})^{1-\delta}\}. \tag{10}$$

The full likelihood can be obtained from the product of the individual contributions. The subject subscript h has been omitted for $m, i_j, t_j, \mathbf{z}$, and δ in (10) and in the rest of this article, for simplicity.

With the expression of the first factor in (10), this likelihood function is equal to the likelihood for **parametric models in survival analysis** with censored observations. For a model with two states ($k = 2$), and with constant transition intensities, this Markov model reduces to a survival model using the **exponential distribution**. In particular, $p_{11}(t|\mathbf{z}) = \exp\{-\lambda_{12}(\mathbf{z})t\}$ and $p_{12}(t|\mathbf{z}) = 1 - \exp\{-\lambda_{12}(\mathbf{z})t\}$, and the contribution of this observation to the likelihood is $\{f(t_{m+1}|\mathbf{z})\}^{\delta}\{S(t_{m+1}|\mathbf{z})\}^{1-\delta}$.

The likelihood function (10) can be extended to the case of **time-dependent covariates** $\mathbf{z}(t)$ by replacing $\mathbf{z}$ by $\mathbf{z}_{j-1}$, where the covariates are assumed to be constant between two observations $\mathbf{z}(t) = \mathbf{z}_{j-1}$, for $t_{j-1} \le t < t_j$.

Parameter Estimation

Let $\theta_{ij} = (\log \lambda_{ij}, \beta_{ij1}, \ldots, \beta_{ijp})$ be the parameters associated with the transition between states i and j, and θ be a vector made up of the θ_{ij} vectors. A log **transformation** is used to prevent the baseline transition intensities from taking negative values during the iterative process of estimation. Let

$$\eta_{ij} = \log \lambda_{ij} + \beta_{ij1} z_1 + \cdots + \beta_{ijp} z_p$$

be the log transition intensity for a subject with covariables $\mathbf{z}$. Maximum likelihood estimates of θ can be found by maximizing the log-likelihood function $l(\theta) = \sum l_h(\theta)$ where

$$\begin{aligned} l_h(\theta) = &\sum_{j=1}^{m} \log\{p_{i_{j-1},i_j}(\boldsymbol{\Delta} t_j; \mathbf{z})\} \\ &+ \delta \log\{f_{i_m}(\boldsymbol{\Delta} t_{m+1} | \mathbf{z})\} \\ &+ (1-\delta) \log\{S_{i_m}(\boldsymbol{\Delta} t_{m+1} | \mathbf{z})\}. \end{aligned}$$

The first derivative of the log likelihood function with respect to θ_{uvl} is

$$\begin{aligned} \frac{\partial l_h(\theta)}{\partial \theta_{uvl}} = &\left\{ \sum_{j=1}^{m} \frac{1}{p_{i_{j-1},i_j}(\boldsymbol{\Delta} t_j; \theta)} \frac{\partial p_{i_{j-1},i_j}(\boldsymbol{\Delta} t_j; \theta)}{\partial \eta_{uv}} \right. \\ &+ \frac{\delta}{f_{i_m}(\boldsymbol{\Delta} t_{m+1} | \theta)} \frac{\partial f_{i_m}(\boldsymbol{\Delta} t_{m+1}; \theta)}{\partial \eta_{uv}} \\ &\left. + \frac{1-\delta}{S_{i_m}(\boldsymbol{\Delta} t_{m+1} | \theta)} \frac{\partial S_{i_m}(\boldsymbol{\Delta} t_{m+1}; \theta)}{\partial \eta_{uv}} \right\} \frac{\partial \eta_{uv}}{\partial \theta_{uvl}}, \end{aligned}$$

where the derivative of $f_i(t|\mathbf{z})$ with respect to η_{uv} is

$$\frac{\partial f_i(t|\mathbf{z})}{\partial \eta_{uv}} = \sum_{j=1}^{k-1} \left\{ \frac{\partial p_{ij}(t; \mathbf{z})}{\partial \eta_{uv}} e^{\eta_{jk}} + p_{ij}(t; \mathbf{z}) \frac{\partial e^{\eta_{jk}}}{\partial \eta_{uv}} \right\}$$

and where

$$\frac{\partial S_i(t|\mathbf{z})}{\partial \eta_{uv}} = -\frac{\partial p_{ik}(t; \mathbf{z})}{\partial \eta_{uv}}.$$

The derivative $\partial p_{ij}(t; \mathbf{z}) / \partial \eta_{uv}$ in the three expressions above is

$$\frac{\partial p_{ij}(t; \mathbf{z})}{\partial \eta_{uv}} = \sum_{r=1}^{k} \sum_{s=1}^{k} a_{ir}(\mathbf{z}) w_{rs}^{uv}(t; \mathbf{z}) a^{sj}(\mathbf{z}), \qquad (11)$$

where

$$w_{rs}^{uv}(t; \mathbf{z}) = \begin{cases} g_{rs}^{uv}(\mathbf{z}) \dfrac{\{\exp[\rho_r(\mathbf{z})t] - \exp[\rho_s(\mathbf{z})t]\}}{(\rho_r(\mathbf{z}) - \rho_s(\mathbf{z}))}, & \text{if } r \neq s, \\ g_{rr}^{uv}(\mathbf{z}) t \exp[\rho_r(\mathbf{z})t], & \text{if } r = s, \end{cases}$$

and $g_{rs}^{uv}(\mathbf{z})$ is the (r, s)th entry in

$$\mathbf{G}^{uv}(\mathbf{z}) = \mathbf{A}(\mathbf{z})^{-1} \frac{\partial \boldsymbol{\Lambda}(\mathbf{z})}{\partial \eta_{uv}} \mathbf{A}(\mathbf{z}).$$

The derivative of η_{uv} respect to θ_{uvl} is

$$\frac{\partial \eta_{uv}}{\partial \theta_{uvl}} = \begin{cases} 1, & \text{for } l = 1, \\ z_{l-1}, & \text{for } l = 2, \ldots, p+1. \end{cases}$$

The subscripts u and v refer to the transition between states u to v.

Quasi-Newton **algorithms** can be used to minimize $-2l(\theta)$ using only the likelihood function and finite differences to obtain numerical approximations of the derivatives, or the likelihood can be maximized using explicit expressions for the derivatives. A discussion of these two approaches can be found in Dennis & Schnabel [10] as well as a modular system of algorithms for unconstrained minimization. Dennis & Schnabel [10] also provide algorithms for computing numerical approximations to the second derivatives of the log likelihood using finite differences of the original function or the gradients if they are available.

Once we have the maximum likelihood estimates of the parameters of the transition intensity matrix $\boldsymbol{\Lambda}(\mathbf{z})$, we also have estimates of the transition probability matrix $\mathbf{P}(t : \mathbf{z} : \theta)$. In particular, an estimate of $p_{ij}(t; \mathbf{z}; \theta)$ can be obtained as $p_{ij}(t; \mathbf{z}; \hat{\theta})$.

An estimate of the asymptotic **covariance matrix** of $\hat{\theta}$ is obtained from the negative inverse of the empirical **information matrix**,

$$\hat{\mathbf{V}}(\hat{\theta}) = -\left\{ \frac{\partial^2 L(\theta)}{\partial \theta \partial \theta'} \right\}_{\theta=\hat{\theta}}^{-1}.$$

The estimate of the asymptotic **variance** of $p_{ij}(t; \mathbf{z}; \hat{\theta})$ can be found using the **delta method** as

$$\begin{aligned} \hat{\mathrm{V}}\{p_{ij}(t; \mathbf{z}; \hat{\theta})\} = &\left\{ \frac{\partial p_{ij}(t; \mathbf{z}; \theta)}{\partial \theta} \right\}_{\theta=\hat{\theta}}' \hat{\mathbf{V}}(\hat{\theta}) \\ &\times \left\{ \frac{\partial p_{ij}(t; \mathbf{z}; \theta)}{\partial \theta} \right\}_{\theta=\hat{\theta}}, \end{aligned}$$

where $\partial p_{ij}(t;\mathbf{z};\theta)/\partial\theta$ can be evaluated using (11). An estimate of the survival function can be obtained directly from the estimate of the transition probability matrix as $\hat{S}_i(t|\mathbf{z}) = 1 - p_{ik}(t;\mathbf{z};\hat{\theta})$. An approximate variance for $\hat{S}_i(t|\mathbf{z})$ is obtained as

$$\hat{\mathrm{V}}\{\hat{S}_i(t|\mathbf{z})\} = \hat{\mathrm{V}}\{p_{ik}(t;\mathbf{z};\hat{\theta})\}.$$

The Natural Course of Diabetic Retinopathy (Continued)

Diabetic retinopathy currently is the leading cause of new cases of blindness in people aged 20–74 years in the US, and is considered in most cases a progressive disease among people with insulin-dependent (type I) diabetes mellitus (IDDM).

Improvement of early stages of retinopathy as part of its natural history has been poorly understood. In the past, physicians believed diabetic retinopathy was a strictly progressive disease. Using a multistate Markov model, Garg, Marshall and colleagues [12, 20] have shown that the natural course of early diabetic retinopathy involves both progression and regression.

The course of early diabetic retinopathy in young subjects with type I diabetes was evaluated during 882 patient visits for 277 subjects over a mean of 2.7 years. All 277 subjects (138 males and 139 females) had direct ophthalmoscopy (with pupils dilated) by at least two examiners (one ophthalmologic and one pediatric), followed by color retinal photography, intravenous fluorescein angiography and slit-lamp examinations. The retinal specialist graded retinal findings as follows: a grade of I indicated no retinopathy; grades II–III microaneurysms or microaneurysms and one other finding; grades IV–V advanced background changes with intraretinal microvascular abnormalities; and grade VI proliferative retinopathy. The category assigned was that of the more severely involved eye.

Based on a modified Airlie House classification [11] of diabetic retinopathy, a four-state Markov model was used considering grades I, II–III and IV–V as transient states and grade VI as an absorbing state, as shown in Figure 3. In this case the exact arrival times at the absorbing state were interval-censored and the likelihood function (9) was used to estimate the parameters.

The influence of duration of diabetes, age, mean HbA_{1c}, diastolic and systolic blood pressure, HbA_{1c} at the visit, sex, smoking, cholesterol, and family history of hypertension on the transition intensities between various stages of diabetic retinopathy was evaluated. A model was fit for each of these factors for all transitions, as shown in Figure 3. All but gender, smoking and family history of hypertension are time-dependent covariates.

HbA_{1c} values were measured at each visit and blood pressure was measured at each visit after the patients rested in a sitting position for 5 minutes. Serum cholesterol levels were measured yearly. Cigarette smoking is a dichotomous indicator of ever having smoked. Family history of hypertension was considered positive if any first-degree relative had received medication for the treatment of hypertension before 50 years of age.

A single-covariate Markov model was used to assess the individual effects of factors associated with diabetic retinopathy using a custom-designed computer program [21]. The full model (6) with five regression coefficients, the PR model (7) with two regression coefficients, and the PMR model (8) with only one regression coefficient were fitted to each factor (Table 2). For each covariate, the most parsimonious model among these three was identified using the **likelihood ratio test**. If a given factor was found to be significantly associated with the disease process in the best model, the parsimonious representation of this factor was used later for **multiple regression** analysis.

Table 2 Likelihood ratio test of single-covariable Markov models for various factors associated with diabetic retinopathy using the full model (6), the PR model (7), and the PMR model (8). All tests are compared to a basic model without covariates

Factor	Full model $\chi^2(5)$	PR model $\chi^2(2)$	PMR model $\chi^2(1)$
Duration of diabetes	58.2	54.7[a]	47.5
Age	33.5[a]	26.3	20.2
Mean HbA_{1c}	27.2	22.2	22.2[a]
Diastolic blood pressure	12.0	10.9	10.5[a]
HbA_{1c} at the visit	10.7	9.0	8.9[a]
Gender	9.8[a]	3.6	1.6
Smoking	9.4	4.1	3.6[a]
Systolic blood pressure	6.7	6.4	6.1[a]
Cholesterol	5.0	4.5	4.4[a]
Family history of hypertension	4.5	4.3[a]	0.9

[a]Best model based on LR test.

The duration of diabetes, the age of the subject and the mean HbA_{1c} levels (mean of all assessments at or before visit time) were the factors most associated with transitions of diabetic retinopathy. Diastolic and systolic blood pressure and values of HbA_{1c} at visit times were also associated with the disease process. All other factors, including gender, mean cholesterol level (mean of all assessments at or before visit time), family history of hypertension, systolic blood pressure and a history of smoking, were not significantly associated with changes in diabetic retinopathy. The significance of the association between these factors and transition times was tested using the likelihood ratio test (Table 2). The only three factors in this study that are time-independent covariates are gender, family history of hypertension and a history of smoking.

Duration of diabetes shows similar effects in all progressive transitions and similar effects in all regressive transitions. Model (7) is chosen as the best representation for the association of this factor and diabetic retinopathy. The regression coefficient estimates for this model were $\hat{\beta}' = (0.0528, -0.2223)$, showing a significant departure from the assumption of model (8). Based on the **standard errors** of the estimates, (0.02774, 0.0456), and their **correlation** coefficient, $r = 0.5295$, we can construct a Wald test (*see* **Likelihood**) for the hypothesis, $H_0 : \beta_p = -\beta_r$ associated with model (8). By using, $\mathbf{L}' = (1, 1)$, the Wald statistic is

$$W = (\mathbf{L}'\hat{\beta})'(\mathbf{L}'\hat{\mathbf{V}}_{\hat{\beta}}\mathbf{L})^{-1}(\mathbf{L}'\hat{\beta}) = 6.80,$$

for which $p < 0.01$ based on the **chi-square distribution** with one **degree of freedom**. The equivalent likelihood ratio test for this hypothesis is $\chi^2_{(2)} - \chi^2_{(1)} = 54.7 - 47.5 = 7.2$ (Table 2). These two results confirm that the PMR model does not hold for duration of diabetes. **Confidence intervals** for the parameters in the model can be obtained by using a Wald-type test based on **normal** approximation.

Table 3 gives the estimates and the **standard errors** of the estimates for the parameters of the final multiple regression model. Duration of diabetes remained the most important factor for explaining changes in diabetic retinopathy. As expected, cumulative HbA_{1c} was the second most important clinical variable associated with transitions in retinopathy. The additional contribution of this factor in terms of the likelihood ratio chi-square test is slightly greater than the contribution without controlling for duration of diabetes. Diastolic blood pressure also remained in the model, showing that it is an independent factor associated with diabetic retinopathy.

Table 3 Parameter estimates and standard errors for the final multiple regression model

Factor	Parameter	Estimate	Standard error
Baseline	λ_{12}	0.0566	0.0075
Baseline	λ_{21}	0.0121	0.0024
Baseline	λ_{23}	0.0163	0.0035
Baseline	λ_{32}	0.0746	0.0243
Baseline	λ_{34}	0.0024	0.0011
Duration of diabetes	β_{p1}	0.0729	0.0283
Duration of diabetes	β_{r1}	−0.2084	0.0461
Mean HbA_{1c}	$\beta_{p2} = -\beta_{r2}$	0.2128	0.0386
Diastolic blood pressure	$\beta_{p3} = -\beta_{r3}$	0.0178	0.0056

The three covariates in the final model were centered in the mean values (in our study these are: 10.7 years of duration of diabetes, an HbA_{1c} value of 11.8 and a diastolic blood pressure of 70), therefore the baseline parameters represent the transition rates from one stage to another for a subject with average values for the risk factors for a given period of time, in this case 1 month. By multiplying the baseline transition estimate from state 3 to state 4 by 12 months and 100 subjects, we conclude that an average of $2.88 (= 0.0024 \times 12 \times 100)$ transitions will occur from stage 3 to stage 4 in a period of 1 year in a group of 100 subjects with average risk factors. Similar conclusions can be made from the remaining baseline transition estimates. The parameters associated with the covariates can be interpreted similarly to the regression coefficients in the Cox regression model. An increment of 1 year of duration of diabetes will increase the risk of progression of the disease process by 7.5% [$\exp(0.0729) = 1.075$] and reduce the chances of regression in the disease process by 19% [$\exp(-0.2084) = 0.81$].

Figure 6 shows estimated survival curves of the probability of remaining free of state 4 (grade VI) retinopathy for a subject with 8 years since onset of diabetes, 12% of HbA_{1c}, and a diastolic blood pressure of 70. The three curves represent the survival curves for starting in one of the three transient stages. Figure 6 shows that the probabilities of remaining free of state 4 (grade VI) retinopathy during a period of 5 years are 96%, 94%, and 86% starting from

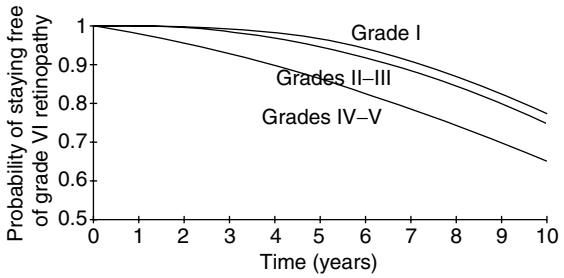

Figure 6 Survival curves for the probability of staying free of grade VI retinopathy

stages 1,2 and 3 at time 0, respectively. These probabilities dramatically decrease during a period of 10 years to 77%, 75%, and 65%, respectively.

These probabilities and Figure 6 also show that staying in stage 2 does not significantly increase the risk of progressing to diabetic retinopathy. However, stage 3 shows a significant reduction during the first 5 years of the probability of staying free of retinopathy and has similar reduction in the second 5-year period when compared to the probabilities of stages 1 and 2.

References

[1] Aalen, O.O. & Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics* **5**, 141–150.

[2] Andersen, P.K. (1986). Time-dependent covariates and Markov processes, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds., 82–103.

[3] Andersen, P.K. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes, *Statistics in Medicine* **7**, 661–670.

[4] Andersen, P.K., Hansen, L.S. & Keiding, N. (1991). Assessing the influence of reversible disease indicators on survival, *Statistics in Medicine* **10**, 1061–1067.

[5] Andersen, P.K., Hansen, L.S. & Keiding, N. (1991). Non- and semi-parametric estimation of transition probabilities from censored observations of a non-homogeneous Markov process, *Scandinavian Journal of Statistics* **18**, 153–167.

[6] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Krieger, Huntington.

[7] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[8] Cox, D.R. & Miller, H.D. (1965). *The Theory of Stochastic Processes*. Methuen, London.

[9] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[10] Dennis, J.E., Jr. & Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, Englewood Cliffs.

[11] Diabetic Retinopathy Study Research Group (1981). A modification of the Airlie House classification of diabetic retinopathy: report 7, *Invest Ophtalmol Vis Sci* **21**, 210–226.

[12] Garg, S.K., Marshall, G., Chase, H.P., Jackson, W., Archer, P. & Crews, M. (1990). The use of the Markov processes in describing the natural course of diabetic retinopathy, *Archives of Ophthalmology* **108**, 1245–1247.

[13] Grüger, J., Kay, R. & Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models, *Biometrics* **47**, 595–605.

[14] Kalbfleisch, J.D. & Lawless, J.F. (1985). The analysis of panel data under a Markov assumption, *Journal of the American Statistical Association* **80**, 832–871.

[15] Kalbfleisch, J.D., Lawless, J.F. & Vollmer, W.M. (1983). Estimation in Markov models from aggregate data, *Biometrics* **39**, 907–919.

[16] Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies, *Biometrics* **42**, 855–865.

[17] Klein, J.P., Klotz, J.H. & Grever, M.R. (1984). A biological marker model for predicting disease transitions, *Biometrics* **40**, 927–936.

[18] Longini, I.M. Jr., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F. & Hethcote, H.W. (1989). Statistical analysis of the stages of HIV infection using a Markov model, *Statistics in Medicine* **8**, 831–843.

[19] Marshall, G. & Jones, R.H. (1995). Multi-state Markov models and diabetic retinopathy, *Statistics in Medicine* **14**, 1975–1983.

[20] Marshall, G., Garg, S.K., Jackson, W.E., Holmes, D.L. & Chase, H.P. (1993). Factors influencing the onset and progression of diabetic retinopathy in subjects with insulin-dependent diabetes mellitus, *Ophthalmology* **100**, 1133–1139.

[21] Marshall, G., Guo, W. & Jones, R.H. (1995). Markov: a computer program for multi-state Markov models with covariables, *Computers, Methods and Programs in Biomedicine* **47**, 147–156.

(*See also* **Time-varying Treatment Effect**)

G. MARSHALL

Predictive Values

The positive predictive value of a diagnostic or **screening** test refers to the proportion of individuals with positive test results who actually have the target disease or disorder, i.e. Pr(disease|positive test result). In the diagram in the article on **Sensitivity**, the positive predictive value is $a/(a+b)$. A synonym is the post-test or posterior probability of disease, given a positive test result.

Correspondingly, the negative predictive value is the proportion of individuals with negative test results who do not actually have the disease or disorder in question. This is Pr(no disease|negative test result), or $d/(c+d)$ in the diagram. A synonym is the post-test or posterior probability of no disease, given a negative test result. There is also some interest in the complement of this quantity, $c/(c+d)$, which is the post-test probability of disease, given a negative test result.

The predictive values are useful clinically, and may influence therapeutic decisions. An individual with a positive result from a test with high positive predictive value is a relatively good candidate for therapeutic intervention. Conversely, there is relatively little justification for intervention for an individual with a negative result from a test that has high negative predictive value for no disease.

Predictive values are functions both of the test characteristics, in particular sensitivity and **specificity**, and of the **prevalence** of disease in the population. For instance, if the testing is done in a high-risk population, where the prevalence (or, equivalently, the pretest probability of disease) is high, then one would expect the predictive values of disease to be relatively high. Conversely, if the disease is rare, then the predictive values for disease are relatively low; in an extreme case, even the positive predictive value may be sufficiently small that therapy cannot be justified on the basis of the test alone, particularly if the therapy is expensive or hazardous.

(*See also* **Clinical Epidemiology**; **Diagnostic Tests, Evaluation of**)

STEPHEN D. WALTER

Prevalence of Disease, Estimation from Screening Data

Consider a large sample survey in which the investigators have used a simple, cheap but fallible indicator of morbidity in order to assess the **prevalence** of one or more illnesses. If morbidity is estimated solely on the basis of this fallible information, then there is a possibility that the results will be **biased**. A common refinement of this sample design (particularly in psychiatric epidemiology) is to use the fallible information as a *screen* to stratify the original sample into two or more groups (usually referred to as screen positives and screen negatives), and then to subsample from these strata for a more thorough diagnostic investigation that is assumed to reveal the correct status of the subject (the **gold standard**). The results of this second phase of investigation are then used to calibrate the fallible screening instrument and to develop an **unbiased** estimator of prevalence from the whole sample. This design is an example of *two-phase sampling* (*see* **Multistage Sampling**). An alternative, related design involves supplementing the large screened sample by a completely independent *calibration* or **validation** sample of comparable subjects who are assessed using both the screen and the gold standard (*see* **Screening Benefit, Evaluation of**; **Screening, Overview**; **Screening, Models of**).

Two-phase or multiphase sampling is also often referred to as **double sampling** by survey statisticians and it is extensively used in industry as one design for lot quality assurance sampling (LQAS). In the health field it is also commonly but less satisfactorily referred to as *two-stage* or *multistage sampling*. The latter terminology is more properly and more usually applied to circumstances in which the sampling units at each stage are different (i.e. *nested* sampling designs). For example, one might take a random sample of clinics or health regions in the first stage of the survey and then randomly sample patients from within each of the selected clinics or regions. In two-phase sampling the sampling units remain the same at both phases of the survey.

The use of a **randomization** procedure while "in the field" to determine second-phase sample membership offers the possibility of concatenating initial and later phases of data collection into a single assessment of each subject. However, it is more common for first and later phases of data collection to involve both different measurement techniques and different research staff, and to take place on different occasions and in different locations. Of increasing interest is the use of registry (*see* **Disease Registers**) or other bureaucratic information systems (*see* **Administrative Databases**), such as hospital birth records, as the source of first-phase "screening" data.

The terminology of first-phase "screening" implies the purposeful application of a screening test: a cheap, robust and often noninvasive test that is, however, recognized as being fallible or less accurate than the gold standard. This is how the idea of screening was introduced in the first paragraph of the present article. From a statistical perspective, however, this is a needless limitation to this sampling design. The statistical problem remains essentially the same whether a true screening test is used to assess morbidity, or whether the screen in fact attempts to measure risk factors rather than ill health, or whether the strata in the two-phase design arise "naturalistically", say, through the operation of a patient referral process. An example of the latter would be an epidemiological study of children sampled from all hospital birth records but stratified by hospital. The critical element of the design is not the use of a screening test, but the stratification of the sampling design on a variable or variables for which inference about the outcome variable should be made marginal to the strata rather than conditional on them (*see* **Stratified Sampling**). Simple prevalence **estimation**, being marginal to all other variables, is the obvious example, but the point also applies to more general analyses. To the extent that second-phase sampling results in data missing by design, the methods to be discussed also relate to methods for data *missing at random* [19] (*see* **Missing Data**) and to the use of *surrogate* or *auxiliary* measures [22].

One final note on terminology: here we use the term "screening" to indicate the provision of a simple fallible indicator of the subjects' state in order to stratify the first-phase sample prior to subsampling and further investigation. The term "screening" is also commonly used to describe the first phase of the process of *case detection*. Suspected cases, who have been identified by the screening procedure, are then further investigated with a view to treatment (in a clinical setting) or entry into a research study, a **clinical trial** for example. In this situation interest

in the screen negatives is more limited, typically to concerns relating to **false negative** patients who ought to be receiving treatment or other forms of help, and to the costs and possible risks associated with the screening process itself. Screening for caseness is not the subject of this article, but is discussed elsewhere (*see* **Screening, Overview**).

The Analysis of Two-phase ("Screened") Samples

Prevalence Estimation Using Simple Conditional Probabilities

Figure 1 illustrates the sample selection process. From a first-phase sample of size N, $N(\mathrm{s}+)$ are screen positive and $N(\mathrm{s}-)$ are screen negative (where $N(\mathrm{s}-) = N - N(\mathrm{s}+)$). Of those subjects who are screen positive, a subsample of size $N_{v+}(\mathrm{s}+)$ are selected at random for a full diagnostic assessment (validation). These subjects will have complete data on all variables. Of these, $N_{v+}(\mathrm{s}+, \mathrm{d}+)$ are found to have a positive diagnosis and $N_{v+}(\mathrm{s}+, \mathrm{d}-)$ to have a negative diagnosis. Similarly, among the $N(\mathrm{s}-)$ screen negatives a **random sample** of size $N_{v+}(\mathrm{s}-)$ is also selected for validation. Of these second-phase subjects, $N_{v+}(\mathrm{s}-, \mathrm{d}+)$ are found to have a positive diagnosis (despite the screening information) and $N_{v+}(\mathrm{s}-, \mathrm{d}-)$ are negative. Note that the $N_{v+}(\mathrm{s}+, \mathrm{d}-)$ subjects are **false positives** (with respect to the screening information) and the $N_{v+}(\mathrm{s}-, \mathrm{d}+)$ are false negatives. Finally there are $N_{v-}(\mathrm{s}+)$ and $N_{v-}(\mathrm{s}-)$ subjects for which there are missing second-phase diagnostic data.

The proportion of cases in the screen-positive and screen-negative strata are denoted by $\pi(\mathrm{s}+)$ and $\pi(\mathrm{s}-)$, respectively. The proportion of screen positives in the first phase is ω and the proportion of screen negatives is therefore $1 - \omega$. Using simple **conditional probabilities**, the overall prevalence, π, is given by

$$\pi = \omega\pi(\mathrm{s}+) + [1 - \omega]\pi(\mathrm{s}-) \tag{1}$$

and can be estimated by inserting sample estimates of each of the components on the right-hand side of the expression. That is, ω, $\pi(\mathrm{s}+)$ and $\pi(\mathrm{s}-)$ are estimated by $n(\mathrm{s}+)/n$, $n_{v+}(\mathrm{s}+, \mathrm{d}+)/n_{v+}(\mathrm{s}+)$ and $n_{v+}(\mathrm{s}-, \mathrm{d}+)/n_{v+}(\mathrm{s}-)$, respectively.

Using the well-known **delta method**, the **variance** of π is given by

$$\begin{aligned}\mathrm{var}(\pi) = {} & \frac{\omega^2\pi(\mathrm{s}+)[1 - \pi(\mathrm{s}+)]}{N_{v+}(\mathrm{s}+)} \\ & + \frac{(1 - \omega^2)\pi(\mathrm{s}-)[1 - \pi(\mathrm{s}-)]}{N_{v+}(\mathrm{s}-)} \\ & + \frac{[\pi(\mathrm{s}+) - \pi(\mathrm{s}-)]^2\omega(1 - \omega)}{N}.\end{aligned} \tag{2}$$

This variance is estimated by insertion of the sample estimates of ω, $\pi(\mathrm{s}+)$ and $\pi(\mathrm{s}-)$, as before.

This simple approach [6] is very straightforward and that most commonly used in two-phase prevalence estimation studies. If, however, the investigator wishes to consider multiple subpopulations and/or investigate the effects of risk exposures on the prevalence of disease, then there are less cumbersome methods.

Prevalence Estimation Using Inverse Probability (Expansion) Weights

For the ith subject with complete data a *probability weight* or *expansion weight*, w_i, can be defined as the reciprocal of the sampling fraction (probability of selection) for the second-phase sample. In the simplest design the weight takes on just two distinct values, one for the screen-positive subjects and another for the screen negatives. The estimates of the sampling fractions are $n_{v+}(\mathrm{s}+)/n(\mathrm{s}+)$ and $n_{v+}(\mathrm{s}-)/n(\mathrm{s}-)$, yielding inverse probability weights of $n(\mathrm{s}+)/n_{v+}(\mathrm{s}+)$ and $n(\mathrm{s}-)/n_{v+}(\mathrm{s}-)$, respectively. The sum of the weights over the subjects in the second-phase sample will be the first-phase sample size. The sum of the products, w_iY_i, where $Y_i = 1$ for validated cases and $Y_i = 0$ for the nonvalidated

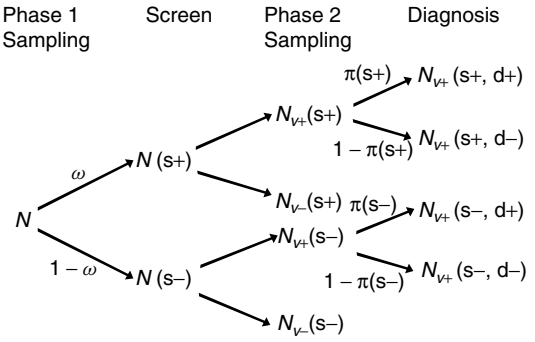

Figure 1 Two-phase data collection

cases, again over the subjects of the second phase, provides an estimate of the number of cases in the first-phase sample. An estimate of the prevalence of disorder is therefore provided by the ratio

$$\pi = \frac{\sum w_i Y_i}{\sum w_i}. \tag{3}$$

This estimator is the **Horvitz–Thompson estimator** familiar to survey methodologists (see, for example, [18]). As in the use of conditional probabilities, it is straightforward to use a *Taylor series* **linearization method** (the delta method) to estimate the variance of this ratio. Other methods of variance estimation, including the **jackknife** or **bootstrap** sampling, can also be used in the case of either of the prevalence estimation methods [18, Chapter 5].

Note that, although conditional probabilities and inverse probability weighting look superficially very different – the weighting method seeming to estimate directly the marginal model of interest while the conditional probability approach derives the marginal rate from the "uninteresting" conditional model for disease given the screen score – in this simple case the two methods of prevalence estimation are algebraically equivalent. The variance estimates, however, being based on different Taylor expansions, may differ slightly, but this is of no practical significance in large samples.

A More General Framework

Let X denote risk exposures assessed at phase 1, Z the screen score and Y the measure of true case status. If complete information were available on all subjects, the **maximum likelihood** estimator of $\boldsymbol{\beta}$, the vector of coefficients for the effects of risk exposures, would be found by maximizing the **log-likelihood** $\sum_i \log \Pr_\beta(Y_i|X_i)$. In the context of two-phase sampling, and letting i index subjects with complete data and j index subjects not selected for the second phase, then the **EM algorithm** for finding the maximum likelihood estimator would involve the iterative solution of

$$l(\boldsymbol{\beta}) = \sum_{i \in V+} \log \Pr_{\boldsymbol{\beta}}(Y_i|X_i) + \sum_{j \in V-} \mathrm{E}[\log \Pr_{\boldsymbol{\beta}}(Y|X_j)\}|\boldsymbol{\beta}^{\mathrm{c}}, X_j, Z_j], \tag{4}$$

where $\boldsymbol{\beta}^{\mathrm{c}}$ is the current estimate. Various authors have discussed the solution of such equations, primarily for categorical risk factors [4, 5, 10, 35]. Eq. (4) suggests a number of alternative estimators based on different approximations or sample-based estimates of the **expectation** term, including various forms of imputation (*see* **Multiple Imputation Methods**). In the *mean score* method of Pepe and others [22, 23, 26], the expectation is estimated by the sample average. This is equivalent to solving **score** equations of the form

$$\sum_{i \in V+} S_{\boldsymbol{\beta}}(Y_i|X_i) + \sum_{j \in V-} \times \left\{ \sum_{i \in \begin{cases} X_i = X_j \\ Z_i = Z_j \end{cases}} \frac{S_{\boldsymbol{\beta}}(Y_i|X_j)}{N_{V+}(X_i = X_j, Z_i = Z_j)} \right\} = 0, \tag{5}$$

which may be rewritten as

$$\sum_{i \in V+} \left[1 + \frac{N_{V-}(X_i, Z_i)}{N_{V+}(X_i, Z_i)} \right] S_{\boldsymbol{\beta}}(Y_i|X_i) = \sum_{i \in V+} w_i S_{\boldsymbol{\beta}}(Y_i|X_i) = 0. \tag{6}$$

The w_i here correspond to the expansion weight of the previous section. Hence, the estimating equations can be easily solved by means of weighted **regression**, with weighted **logistic regression** being the common choice.

Weighted regression estimators similar to (6) have been derived within the survey research field [2, 3]. Flanders & Greenland [11] derive the same estimator using **pseudo-likelihood** arguments, proposing the use of the variously termed "empirical", "robust", "information sandwich," "heteroscedastic consistent," and "Huberized" parameter **covariance matrix** [14]. In this case the parameter covariance matrix obtained assumes the weights W_i to be known. In practice, the weights are usually estimated and often adjusted using a variety of methods quite separate from the analysis for prevalence estimation. The weights can be **"poststratified"** [16] to conform to known population rates, or "raked" [9] to conform to some known margins, or they may

be smoothed using parametric or **nonparametric regression** methods [17], or even **trimmed** [25] or shrunken (*see* **Shrinkage**) [7]. Whatever the weights chosen, the subsequent analysis of the prevalence data can include any risk exposure measures, continuous or discrete, from either phase.

Perhaps contrary to intuition, Pepe et al. [23] illustrate how if weights are estimated by simple sample cell frequencies, then subjects in the nonvalidation set contribute information through the variation in these random weights, and worthwhile efficiency gains can be made by fully exploiting this variation rather than assuming the weights to be known. However, where some cells are empty or where one or more risk exposures or surrogate measures are continuous, the mean score method cannot be applied directly. Although under such circumstances some estimator based on smoothing might be attempted, a series of papers by Robins and others [27–29] argue that such estimators may not approach their asymptotic distribution in the moderate samples of typical studies. Instead they draw on the ideas of **semi-parametric regression** estimators to propose a modified score function of the form

$$\sum_{i \in V+} [w_i h(X_i)\varepsilon_i(\boldsymbol{\beta}) + (1 - w_i)\phi(W_i)] + \sum_{j \in V-} \phi(W_j) = 0, \quad (7)$$

where w_i is the usual expansion weight, $\varepsilon_i(\boldsymbol{\beta})$ is the simple observed minus model expected **residual**, and $h(X)$ and $\phi(W)$ are functions of the data and parameters that can be chosen to achieve optimal efficiency. Though of considerable interest, there is currently little practical experience with such estimators.

Bayesian Approaches

Many of the issues already discussed have parallels in **Bayesian methods** for multiphase sampling, including the choice as to the direction in which to model the graph [24]. Estimation is typically carried out using **Markov chain Monte Carlo** methods and *Gibbs sampling* in particular (see, for example, [12]). Although the ability to incorporate prior information may be of particular value where some screen strata may not have been subsampled, a particular focus of Bayesian work in this area has been in **model choice** and *model averaging* [36]. This typically arises where the data from the two phases are linked by happenstance rather than by formal design, for example as occurs in the overlapping samples obtained from bureaucratic systems and **capture–recapture** sampling methods, and thus where there are a number of plausible probability models that could describe the dependencies within the data. Recently, Robins & Ritov [27] have criticized the use of independent **prior distributions** for the parameters of the screen and of the diagnostic measure that have typically been used.

Comparisons of Methods

The Bayesian, conditional probability, weighting, and EM methods for a simple prevalence study are compared in Pickles et al. [24]. Schill & Drescher [30] present some comparisons for the related design where the screen is for risk exposure.

Statistical Software

In general, users of commercial statistical packages should take great care in the use of any weighting procedures provided. The use of weights within most packages (such as SPSS [20]; *see* **Software, Biostatistical**) will give the correct estimates of prevalence and **odds ratios**, but unfortunately, for the most part, will not use the appropriate variance estimator. Estimates for **confidence intervals**, for example, and associated significance tests will be invalid. This arises from the fact that the weights are typically interpreted as *frequency weights* (an indicator of the number of observations with identical data to that provided in a given record). The package accordingly treats the ith subject in the second-phase sample as if it had actually been recorded w_i times. The appropriate use of a probability weight, however, recognizes that the observation has only occurred once, but that the observed second-phase subject is representative of w_i first-phase subjects, all but one of which have not provided second-phase data. Programs such as SUDAAN [31] and STATA [34] will deal satisfactorily with weights assumed known. In general, though with substantial variation in ease of use, almost any statistical program can be persuaded to draw appropriate samples for bootstrap estimation of the variance of weighted estimates. Software for

mean score and semi-parametric regression methods is not widely available, although implementation is claimed to be straightforward. Bayesian models can now be easily fitted using Gibbs sampling programs such as BUGS [33].

Design Considerations

The relative advantages and disadvantages of screening in a two-phase survey, and also considerations of **optimal designs**, have been discussed by, among others, Deming [8] and Shrout & Newman [32]. Clearly the use of a preliminary screening instrument has potential advantages when it is both expensive, time-consuming and difficult to carry out an accurate diagnostic assessment. It is intuitively appealing (particularly for clinically trained researchers contemplating a series of long and detailed psychiatric interviews, for example) to consider some way of excluding the majority of subjects who do not have a problem in order that valuable diagnostic resources can be expended on those that do. The rarer the illness, the more appealing this idea becomes. The approach is only practically viable, however, if the potential screening procedures are cheap (relative to the full diagnostic assessment), accurate, easily administered, and accepted by the survey participants. Here accuracy implies both a high **sensitivity** and a high **specificity**, with high sensitivity being the more important of the two screening test characteristics. Clearly, we do not want a screening test which misses a relatively high proportion of our rare cases of illness. Hand [13] discusses the determination of the cut-off for a screening questionnaire that gives optimal sensitivity and specificity, and how different cut-offs may be appropriate for either prevalence estimation or case detection. Begg & Greenes [1] discuss the estimation of screen error rates from two-phase studies.

The costs involved in collecting survey data using a two-phase design come from three areas of activity: recruitment, screening, and diagnostic assessment. If it is difficult and/or expensive to recruit subjects, then the two-phase design becomes less attractive, and a better strategy might be to lower the cost of diagnostic assessment procedures. An example of the latter would be the development of fully structured interviews for use by lay interviewers. In many situations the relative gains in efficiency from two-phase sampling seem to be rather slight. In summary, in relation to single-phase designs, two-phase designs will be more efficient when the prevalence is low, and are likely to be less efficient if the screen costs more than half of the cost of the diagnostic assessment [32]. Deming [8] has argued that unless there are clear gains in terms of relative efficiency, many of the disadvantages of the design would lead us to decide not to adopt it in survey work. These include the extra administrative problems related to conducting a survey in two phases, the logistical problems in recontacting respondents selected for the second phase, scheduling their second-phase assessment and minimizing **nonresponse** and **noncompliance**. There are also the problems for the management of the more complex databases and the increased complexity in the analysis of the results. Despite all of these problems, the design seems to be growing in popularity. One suspects that investigators pay too little attention to them at the design stage of a study or, if they do, give them less weight than the obvious attractions of the design. Clinicians (and others) clearly consider that it is a waste of valuable resources to spend time assessing people without problems. The idea of only assessing a very small proportion of those likely to be well is possibly the strongest motivation for a two-phase study. Even if we accept this view we should be very wary of assessing too small a proportion of screen negatives and should definitely resist the temptation to assess none of them!

Little work has been done on the design of two-phase studies for estimating the effects of risk exposure. Design possibilities include **stratification** by potential risk factors. Palmgren [21] considers optimal design for estimating an **odds ratio** for comparing two prevalences. Reilly [26] also considers optimal two-phase designs.

Other Areas of Application

Although the purpose of this article is to discuss screening procedures for prevalence estimation, it is also useful to consider other areas of application in which either screening or two-phase sampling might be of potential value. The first is in studies specifically designed to evaluate the performance of new screening instruments or **diagnostic tests** [1]. In a prospective design all the first-phase subjects are assessed using the screen and then selected subsamples are evaluated using the gold standard. In a

retrospective study all of the first-phase subjects are given (or already have) a definitive diagnosis and in this design the selected subsamples are the ones who then get an assessment using the new test or screening procedure. Note that it is particularly important that the assessments made within the two phases (irrespective of whether the study is prospective or retrospective) are made independently, i.e. the assessors are blind to previous results (*see* **Blinding or Masking**). The value of maintaining blindness is another reason why both first-phase strata should be subsampled in a simple two-phase prevalence study (assuming, as should be the case, that first-phase results are not made available to the second-phase assessor). Similar designs might also be used in studies of diagnostic **agreement**: a trainee clinician or research student, for example, diagnosing all available patients and then their supervisor validating the diagnoses by independent assessment of subsamples of the first-phase diagnostic groups [15]. The final area of application to be mentioned here is in the use of **surrogate endpoint** measures (i.e. screens) in lengthy follow-up studies and, in particular, randomized controlled trials (*see* **Clinical Trials, Overview**). This is an area of application that is in infancy but has been discussed in the context of mean score [23] and regression methods [28].

References

[1] Begg, C.B. & Greenes, R.A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias, *Biometrics* **39**, 207–215.

[2] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review* **51**, 279–292.

[3] Binder, D.A. (1996). Linearization methods for single and two-phase samples: a cookbook approach, *Survey Methodology* **22**, 17–22.

[4] Chambless, L.E. & Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models, *Communications in Statistics – Theory and Methods* **14**, 1377–1392.

[5] Chen, T.T. (1979). Log-linear models for categorical data with misclassification and double-sampling, *Journal of the American Statistical Association* **74**, 481–487.

[6] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[7] Cohen, T. & Spencer, B.D. (1991). Shrinkage weights for unequal probability samples, in *American Statistical Association 1991 Proceedings of the Survey Research Methods Section*. American Statistical Association, Alexandria, pp. 625–630.

[8] Deming, W.E. (1977). An essay on screening, or two-phase sampling, applied to surveys of a community, *International Statistical Review* **45**, 29–37.

[9] Deming, W.E. & Stephan, F.F. (1940). On a least squares adjustment of a simple frequency table, when the expected margin totals are known, *Annals of Statistics* **11**, 427–444.

[10] Espeland, M.A. & Hui, S.L. (1987). A general approach to analyzing epidemiological data that contain misclassification errors, *Biometrics* **43**, 1001–1012.

[11] Flanders, W.D. & Greenland, S. (1991). Analytic methods for two stage case – control studies and other stratified designs, *Statistics in Medicine* **10**, 739–747.

[12] Gilks, W.R. Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. & Kirby, A.J. (1993). Modelling complexity: applications of Gibbs sampling in medicine (with discussion), *Journal of the Royal Statistical Society, Series B* **55**, 39–102.

[13] Hand, D. (1987). Screening versus prevalence estimation, *Applied Statistics* **36**, 1–7.

[14] Huber, P.J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, L. LeCam & J. Neyman, eds. University of California Press, Berkeley, pp. 221–233.

[15] Jannarone, R.J., Macera, C.A. & Garrison, C.Z. (1987). Evaluation of inter-rater agreement through "case – control" sampling, *Biometrics* **43**, 433–437.

[16] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[17] Lazzeroni, L.C. & Little, R.J.A. (1993). Models for smoothing post-stratification weights, in *American Statistical Association 1993 Proceedings of the Survey Research Methods Section*. American Statistical Association, Alexandria, pp. 764–769.

[18] Lehtonen, R. & Pakkinen, E.J. (1995). *Practical Methods for the Design and Analysis of Complex Surveys*. Wiley, Chichester.

[19] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, Chichester.

[20] Norusis, M.J. (1993). *SPSS for Windows*. SPSS Inc., Chicago.

[21] Palmgren, J. (1987). Precision of double sampling estimators for comparing two probabilities, *Biometrika* **74**, 687–694.

[22] Pepe. M.S. (1992). Inference using surrogate outcome data and a validation sample, *Biometrika* **79**, 355–365.

[23] Pepe, M.S., Reilly, M. & Fleming, T.R. (1994). Auxiliary outcome data and the mean score method, *Journal of Statistical Planning and Inference* **42**, 137–160.

[24] Pickles, A., Dunn, G. & Vazquez-Barquero, J.L. (1995). Screening for stratification in two-phase (two-stage) epidemiological surveys, *Statistical Methods in Medical Research* **4**, 73–89.

[25] Potter, F.J. (1990). A study of procedures to identify and trim extreme sampling weights, in *American*

Statistical Association 1990 Proceedings of the Survey Research Methods Section. American Statistical Association, Alexandria, pp. 225–230.

[26] Reilly, M. (1996). Optimal sampling strategies for two-stage studies, *American Journal of Epidemiology* **143**, 92–100.

[27] Robins, J.M. & Ritov, Y. (1997). A curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models, *Statistics in Medicine* **16**, 285–319.

[28] Rotnitzky, A. & Robins, J.M. (1995). Semiparametric regression estimation in the presence of dependent censoring, *Biometrika* **82**, 805–820.

[29] Rotnitzky, A. & Robins, J.M. (1997). Semi-parametric regression models with non-ignorable non-response, *Statistics in Medicine* **16**, 81–102.

[30] Schill, W. & Drescher, K. (1997). Logistic analysis of studies with two-stage sampling: a comparison of four approaches, *Statistics in Medicine* **16**, 117–132.

[31] Shah, B.V., Folsom, R.E., Lavange, L.M., Wheeless, S.C., Boyle, K.E. & Williams, R.L. (1995). *SUDAAN: Software for the Statistical Analysis of Correlated Data*. Research Triangle Institute, Research Triangle Park.

[32] Shrout, P.E. & Newman, S.C. (1989). Design of two-phase prevalence surveys of rare disorders, *Biometrics* **45**, 549–555.

[33] Spiegelhalter, D.J., Thomas, A., Best, N.G. & Gilks, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 5.0*. Medical Research Council Biostatistics Unit, Institute of Public Health, University Forvie Site, Cambridge.

[34] StataCorp (1997). *Stata Statistical Software: Release 5.0*. Stata Corporation, College Station.

[35] Tenenbaum, A. (1970). A double sampling scheme for estimating from binomial data with misclassification, *Journal of the American Statistical Association* **65**, 1350–1361.

[36] York, J., Madigan, D., Heuch, I. & Lie, R.T. (1995). Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty, *Applied Statistics* **44**, 227–242.

(*See also* **Diagnostic Tests, Evaluation of**)

ANDREW PICKLES & GRAHAM DUNN

Prevalence Rate or Ratio

The prevalence rate or ratio refers to the number of people who have a disease or condition at a given point in time in a given population divided by the number of people in that population. This quantity is also known as the *prevalence proportion*.

(*See also* **Prevalence**)

MITCHELL H. GAIL

Prevalence

Prevalence is the number of persons who have a disease or condition at a given point in time in a defined population. Sometimes prevalence refers to persons who either have currently or have previously had a disease, e.g. the prevalence of persons with a cancer diagnosis at any time up to the present. The term *prevalence* is used also to refer to the proportion of individuals with disease in the population, namely the **prevalence rate or ratio**.

MITCHELL H. GAIL

Prevalence–Incidence Bias

Bias in epidemiologic studies is a form of "systematic error in the design, conduct, or analysis of a study that results in a mistaken estimate of an exposure's effect on the risk of disease" [5]. While biases have been categorized in many ways, Rothman describes three general types: **selection bias,** *information bias*, and **confounding** [3]. Selection bias results from improper specification or selection of the study sample and leads to a distortion of the effect measured [3, 4].

Jerzy Neyman identified a form of bias in analytic epidemiologic studies which is in essence the result of the use of prevalent rather than incident cases of disease to assess etiologic relationships [2] (*see* **Causation**). *Prevalence–incidence bias* is a form of selection or sampling bias that results in part from evaluating the exposure–disease association well after the exposure first occurs. During that time interval, cases of short duration, due to death or short course of illness, cases mild in severity or asymptomatic, and cases in which the presence of disease alters or entirely removes the exposure, are missed [4].

Neyman developed a fictitious example that assessed the impact of differential survival on an etiologic association [2]. This example, which assessed the relationship between cigarette smoking and lung cancer, is shown in Table 1.

In this example, a comparison of the development of lung cancer between time 0 and time T in smokers and nonsmokers indicates a reduced risk among smokers (relative **odds ratio** = 0.44). If one assumes, however, that 95% of the lung cancer cases in nonsmokers died before time T, as opposed to only 10% of the lung cancer cases in smokers, and that lung cancer was the only source of mortality in this population, then a **case–control study** applied to those alive at time T would produce a large putative effect for smoking (relative odds = 8.0).

While clearly fictitious in terms of the known risk of smoking on lung cancer, this example serves to demonstrate the effect of using prevalent cases, at time T, to describe the risk of disease development. If an exposure, as in this example, results in selective survival and prevalent cases are used, then the estimate of the risk of disease is an overestimate. If an exposure results in selective mortality, then the estimate of the risk of disease associated with that exposure is an underestimate.

Prevalence–incidence bias is a particular problem in evaluating associations between selected risk factors and coronary heart disease [1, 4]. Many cases of coronary heart disease are rapidly fatal and some are clinically mild or asymptomatic. Each of these situations can lead a study of prevalent cases to a faulty conclusion due to an under-representation of cases. In addition, in coronary heart disease, and in perhaps other diseases, it is possible that the presence of the disease will result in a modification of the exposure being evaluated. This could make the association between an exposure and disease appear either larger or smaller than it really is. Such an example is shown in Table 2.

Here, the true risk of the development of coronary heart disease associated with high serum cholesterol, evident in the prospective study (relative odds = 2.40), is not seen in a case–control study using prevalent cases of coronary heart disease and assessment of serum cholesterol at the same time (relative odds = 1.16) [1, 4]. It is likely that persons with diagnosed coronary heart disease have altered their dietary patterns and reduced both their intake of foods high in cholesterol and their weight. It should be noted that this same change in relative odds was evident when analyses by Friedman et al. of the prospective component were restricted to persons who survived

Table 1 An example of prevalence–incidence bias using fictitious data on smoking and lung cancer [2]

	Number at time 0	Lung cancer by time T	Case–control study at time T	
			Number	Lung cancer cases
Smokers	10 000	1000[a]	9900	900
Nonsmokers	10 000	2000[b]	8100	100
Relative odds ratios	0.44		8.00	

[a]Includes 100 deaths from lung cancer.
[b]Includes 1900 deaths from lung cancer.

Table 2 Prospective versus retrospective study estimates of the relative odds of coronary heart disease among Framingham men with and without high cholesterol [1, 4]

	Prospective study[a]		Retrospective study[b]	
Cholesterol	CHD	NoCHD	CHD	NoCHD
High[c]	85	462	38	34
Low	116	1511	113	117
Relative odds ratio	2.40		1.16	

[a]Developed CHD by time T.
[b]CHD present at time T.
[c]Cholesterol measured at exam. 1 in prospective study and at exam. 6 in retrospective study.

the full time period, lending further credence to the possibility that prevalent cases of coronary heart disease may have altered their lifestyles, which resulted in lower serum cholesterol [1, 4].

In summary, prevalence–incidence bias is likely to exert its effect when a considerable amount of time elapses between exposure and selection of subjects for a study [4]. It is a bias that can result in either an overestimate or underestimate of the true risk of disease associated with a particular factor and is evident in case–control studies where prevalent, rather than incident, cases are selected as subjects.

The ability to correct, or even measure, the spurious effect caused by this bias is limited at best [4]. Avoidance of this form of bias is not simply achieved through the use of incident rather than prevalent cases in a case–control study. Ascertaining the true association between the exposure and the disease can only be achieved through the use of a prospective study of incident disease (*see* **Cohort Study**).

References

[1] Friedman, G.D., Kannel, W.B., Dawber, T.R. & McNamara, P.M. (1966). Comparison of prevalence, case history, and incidence data in assessing the potency of risk factors in coronary heart disease, *American Journal of Epidemiology* **83**, 366–377.
[2] Neyman, J. (1955). Statistics – servant of all sciences, *Science* **122**, 401–406.
[3] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company, Boston.
[4] Sackett, D.L. (1979). Bias in analytical research, *Journal of Chronic Diseases* **32**, 51–63.
[5] Schlesselman, J.J. (1982). *Case–Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.

SYLVIA. E. FURNER

Prevalent Case

A prevalent case is a subject with a given disease or condition who is alive in a defined population at a given time. Sometimes the condition may refer to the previous occurrence of an illness, as for persons who now have or who previously have had cancer. Thus, prevalent cases include subjects who developed disease previously as well as **incident cases** who just developed disease.

(*See also* **Biased Sampling of Cohorts**; **Case–Control Study, Prevalent**; **Cross-sectional Study**)

MITCHELL H. GAIL

Preventable Fraction

When considering a protective exposure or intervention, an intuitively appealing alternative to **attributable risk** (AR) is the preventable fraction (PF). The preventable or prevented fraction measures the impact of an **association** between a protective exposure and a disease at the population level. It is defined as the proportion of disease cases averted by a protective exposure or intervention (5). It can be formally written as:

$$PF = \frac{\Pr(D|\overline{E}) - \Pr(D)}{\Pr(D|\overline{E})}, \tag{1}$$

where $\Pr(D)$ is the probability of disease in the population, which may have some exposed (E) and some unexposed ($\overline{E}$) individuals, and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same population but with all (protective) exposure eliminated. Another formulation of PF is the proportion of cases prevented by the (protective) factor or intervention among the totality of cases that would have developed in the absence of the factor or intervention [5], which is why the denominator in (1) is the hypothetical probability of disease in the population in the absence of the protective factor.

PF can be rewritten as:

$$PF = \Pr(E)(1 - RR), \tag{2}$$

where $\Pr(E)$ denotes the **prevalence** of the protective exposure in the population and RR the **relative risk**. As is apparent from (2), PF depends both on the prevalence of the protective exposure and the strength of the association between the protective exposure and disease, in a similar fashion as for AR. A strong association between exposure and disease (marked by a low RR) may therefore correspond to a high or low value of PF, depending on the prevalence of exposure, and portability from population to population is not a common property of PF.

For a protective factor ($RR < 1$), PF lies between 0 and 1 and is usually expressed as a percentage. PF increases with the prevalence of exposure and the strength of the association between exposure and disease. PF is null in the absence of association between exposure and disease ($RR = 1$) and negative when the exposure factor is a risk factor ($RR > 1$), in which case there is no rationale for using PF as a measure of impact.

Counter to what intuition might suggest, PF is not a mere negative AR; that is, PF does not equal $-AR$ (unless $RR = 1$). The relationship between AR and PF was worked out by Walter [6] and is given by

$$1 - PF = \frac{1}{1 - AR}. \tag{3}$$

For a protective factor, AR is negative but can be made positive by reversing the coding of exposure. Under reverse coding, the exposed (protective) level is relabeled as the reference level and the unexposed level as the "exposed", which leads to a positive value of AR. This value is interpreted as the proportion of cases attributable to lack of exposure to the protective factor and which could therefore potentially be prevented by generalizing exposure in the population. This valid interpretation of AR under reverse coding has been used in the literature. For example, a positive AR for protective dietary factors was estimated in a **case–control study** of gastric cancer [3]. AR under this interpretation is sometimes called the preventable fraction [2, 4], which may introduce some confusion.

It is important to note that the value of AR under reverse coding differs from the value of PF. For instance, if RR and $\Pr(E)$ are both equal to 0.5, then PF is equal to 0.25 (or 25%), and the value of AR obtained by reversing the coding is 0.33 (or 33%). This difference is not surprising in view of the differing definitions of AR and PF. AR, with reverse coding, measures the potential reduction in disease cases if all subjects in the current population became exposed (if the absence of exposure were eliminated from the population), while PF measures the reduction in disease cases obtained by moving from a totally unexposed population to the current population with exposure prevalence given by $\Pr(E)$.

Given their close relationship, AR and PF share the same properties. Regarding the interpretation of PF, the above definition implies that PF measures the actual reduction in disease load due to exposure. This interpretation is fully warranted only under the same conditions that ensure the validity of the interpretation of AR as the actual reduction of disease corresponding to the elimination of exposure. These conditions are **unbiased** estimation of PF, a causal role for the exposure (*see* **Causation**), and

invariance (*see* **Sufficient Statistic**) of the distribution of the other factors influencing disease occurrence under a change in the exposure distribution. It might therefore be wiser to define PF as the *potential* reduction in disease load *associated with* exposure.

PF can not only be defined with regard to a protective exposure factor, but also to the *de novo* introduction of a prevention program in the target population (e.g. an intervention designed to reduce smoking). In such a case, since the prevalence of the program is equal to zero before its implementation, one can assess the impact of its introduction through the estimation of PF. Clearly, the impact depends both on the effectiveness of the program and its diffusion in the population (i.e. the prevalence of "exposure" to the program), which is reflected in the estimate of PF. This has been suggested as an analytic tool for assessing the effects of interventions [1].

It follows from (3) that estimability and **estimation** issues are similar for AR and PF. Adjusted PF estimates based on the **Mantel–Haenszel** approach have been derived for **cohort**, case–control and **cross-sectional studies** [2]. Unadjusted estimates and adjusted estimates based on the weighted-sum approach have been derived for cross-sectional studies, and corresponding sample size calculations are available for using a test that PF equals 0 to assess interventions [1].

References

[1] Gargiullo, P.M., Rothenberg, R. & Wilson, H.G. (1995). Confidence intervals, hypothesis tests, and sample sizes for the prevented fraction in cross-sectional studies, *Statistics in Medicine* **14**, 51–72. (Erratum in *Statistics in Medicine* **14**, 841, 1995).

[2] Greenland, S. (1987). Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data, *Statistics in Medicine* **6**, 701–708.

[3] La Vecchia C., D'Avanzo, B., Negri, E., Decarli, A. & Benichou, J. (1995). Attributable risk for stomach cancer in Northern Italy, *International Journal of Cancer* **60**, 748–752.

[4] Last, J.M., ed. (1983). *A Dictionary of Epidemiology*. Oxford University Press, New York.

[5] Miettinen, O.S. (1974). Proportion of disease caused or prevented by a given exposure, *American Journal of Epidemiology* **99**, 325–332.

[6] Walter, S.D. (1976). The estimation and interpretation of attributable risk in health research, *Biometrics* **32**, 829–849.

JACQUES BENICHOU

Prevention Trials

There is a considerable history of the use of randomized **clinical trials** to assess strategies for the primary prevention of disease. For example, in the US, major coronary heart disease prevention trials date from the 1960s [7, 9, 10] to the present, while a number of trials have been initiated over the past one to two decades that focus on the prevention of major cancers and other important diseases. Primary prevention trials generally focus on reducing the occurrence rate of one or more diseases, in contrast to **screening trials**, which aim to reduce mortality rates through early detection and effective treatment of disease.

However, the history of primary prevention trials is quite modest compared with that of therapeutic trials that assess strategies for the treatment of established disease. In fact, the role and place of primary prevention trials in relation to other research strategies remains controversial, and is an important topic for further methodologic research. It is useful to review some basic features of prevention trials to explain the reasons for controversy, and to highlight research needs.

First, consider the nature of the interventions or the treatments to be assessed or compared. In therapeutic research, these arise typically from basic biological research in conjunction with drug screening studies. While such sources may also generate primary disease prevention hypotheses, particularly for chemoprevention trials, observational epidemiologic sources (*see* **Observational Study**) also provide a common and important source of prevention trial hypotheses. For example, preventive interventions may involve such "lifestyle" maneuvers as physical activity increases, nutrient consumption reductions or supplementation, or modifications of sexual behavior.

A therapeutic trial among patients diagnosed with a serious disease will aim typically to identify effective treatments for the reduction of a frequent outcome, such as disease recurrence or death, and may need to be of only one or two years' duration. In contrast, a primary prevention trial of a common vascular disease or cancer will typically focus on the reduction of incidence of a disease typically occurring at a rate of 1% per year or less, and will assess interventions that may require several years to realize the most important of their hypothesized effects. Hence, therapeutic trials may require relatively small numbers of patients (*see* **Sample Size Determination for Clinical Trials**), perhaps only a few hundred, while primary prevention trials may require tens of thousands of subjects, with resulting logistical challenges and substantial costs. To cite a particular example, the Multiple Risk Factor Intervention Trial (MRFIT) [10] of combined hypertension treatment, blood cholesterol lowering and smoking cessation vs. control for the prevention of coronary heart disease, involved the **randomization** of over 12 000 middle-aged and older men thought to be at high risk of coronary heart disease, with an average follow-up duration of over seven years. Only about 2% of MRFIT men experienced the designated primary outcome (*see* **Outcome Measures in Clinical Trials**), coronary heart disease mortality, by the planned date of study completion. The cost of the trial was reported to be in the vicinity of US$100 million.

The size and cost of prevention trials may be increased further by the need to be conservative in establishing intervention goals to ensure the safety of ostensibly healthy study subjects for whom the ability to monitor adverse events carefully may be somewhat limited. In contrast, a therapeutic trial of frequently monitored patients may be able to risk toxicity to achieve efficacy. Furthermore, the long duration of prevention trial follow-up may imply a noteworthy reduction in adherence to intervention goals over time, resulting in important increases in the necessary sample size.

The interventions studied in a prevention trial, like the treatments assessed in a therapeutic trial, may have the potential to affect various disease processes, beneficially or adversely, in addition to those specifically targeted for reduction. In view of the typical dominance of the patient's disease, these other effects are often relatively unimportant in a therapeutic trial. In a prevention trial, however, overall benefit vs. risk assessments (*see* **Health Economics**) can be quite different from the assessment for the designated primary outcome alone. Even a fairly rare adverse effect can eliminate the public health utility of a preventive maneuver. The need to assess the interventions in terms of suitable measures of benefit vs. risk has implications for trial design and particularly for trial monitoring and reporting.

Role Among Possible Research Strategies

In view of the above litany of obstacles and challenges, it seems logical to take the viewpoint that a full-scale disease prevention trial is justified only if the interventions to be assessed have sufficient public health potential, and if alternate less costly research strategies appear unable to yield a sufficiently reliable assessment of intervention effects. If the intervention of interest falls outside the range of common human experience, as is often the case with chemopreventive interventions, there is little debate that randomized controlled trials constitute the research strategy of choice, and the discussion can focus on public health potential and research costs. However, if the intervention is already practiced in varying degrees by large numbers of persons, purely observational approaches may sometimes provide reliable disease prevention information at lesser cost, and perhaps in a shorter time, than can a randomized, controlled intervention trial. In fact, a single observational study; for example, a **cohort study**, may be able to assess a broader range of interventions than is practical to include in the design of a randomized prevention trial.

However, key determinants of observational study reliability include the ability to control **confounding** and the ability to measure accurately the level of intervention adopted. **Measurement error** in the "exposure" histories of interest, or in confounding factors histories, can invalidate observational study hypothesis tests and estimates of intervention effects. Furthermore, randomized trials have the major advantages that the **randomized treatment assignment** (i.e. intervention vs. control) is statistically independent of all prerandomization risk factors, whether or not these are even recognized, and that outcome comparisons among randomization groups (i.e. **intention to treat analysis**) typically will provide valid hypothesis tests, even if adherence to intervention varies among study subjects and is poorly measured (*see* **Compliance Assessment in Clinical Trials**).

Consider the specific context of the Women's Health Initiative (WHI) clinical trial [17, 21], which is enrolled 68 132 postmenopausal American women in the age range 50–79. This trial is designed to allow randomized controlled evaluation of three distinct interventions: a low-fat eating pattern, hypothesized to prevent breast cancer and colorectal cancer, and, secondarily, coronary heart disease; hormone replacement therapy, hypothesized to reduce the risk of coronary heart disease and other cardiovascular diseases, and, secondarily, to reduce the risk of hip and other fractures; and calcium and vitamin D supplementation, hypothesized to prevent hip fractures and, secondarily, other fractures and colorectal cancer. Each of these three interventions is already being practiced in some fashion by large numbers of postmenopausal American women. Important disease reductions can be hypothesized for each intervention, based on substantial observational studies, animal experiments (*see* **Preclinical Treatment Evaluation**) and randomized trials with intermediate outcomes (e.g. the Postmenopausal Estrogen/Progestin Intervention (PEPI) Trial [11]). In the case of hormone replacement therapy, a randomized trial is motivated by potential confounding in cohort and **case–control studies** as hormone users tend to be of higher socioeconomic status with fewer vascular disease risk factors, by the magnitude of the hypothesized benefits, and, importantly, by the need for reliable summary data on benefits vs. risks (*see* **Data Monitoring Committees**), particularly since breast cancer risk may be adversely affected by hormone replacement therapy. The dietary modification trial component is motivated by associations between international cancer incidence rates and per capita fat consumptions (*see* **Ecologic Study**), by **migrant study** data, and by rodent feeding experiments. A large number of case–control and cohort studies of dietary fat and various cancers have yielded mixed results. These latter studies rely exclusively on dietary self-reports, which are known from repeatability studies to involve substantial measurement error, though the absence of a **gold-standard** dietary measurement procedure precludes an assessment of measurement error characteristics as a function of actual dietary habits and of study subject characteristics, such as body mass. For example, Prentice [15] describes a plausible measurement model for dietary fat intake under which even the strong associations suggested by international disease rate comparisons would be essentially eliminated by random and systematic aspects of dietary assessment measurement error. Hence, the current observational studies of dietary fat in relation to cancer or other diseases may be uninterpretable, motivating the need for a randomized intervention trial to assess whether a change to a low-fat eating pattern during the middle decades of life can reduce the risk of selected cancers and cardiovascular diseases. This controversy over the interpretation of the

observational data on dietary fat points to the pressing need for objective measures of fat consumptions (i.e. biomarker measures) and for the development of flexible measurement models to allow self-report and objective exposure data to be combined in exposure–disease rate analyses. The issues of exposure measurement error, along with limited exposure variation within populations, and highly correlated exposure variables, also point to a possible greater role for aggregate (ecologic) study designs (e.g. [16]) among observational research strategies. The calcium and vitamin D component of the WHI is viewed as a comparatively inexpensive addition to the clinical trial. It is motivated by the public health potential of the intervention, as well as by observational data, and data from smaller clinical trials.

Prevention Trial Planning and Design

Suppose that an intervention having potential to prevent one or more diseases is to be subjected to a randomized controlled trial. The trial design should be responsive to the **target population** to which the intervention, if effective, might be applied. For example, the three interventions to be studied in the trial component of the WHI are all potentially applicable to the general population of postmenopausal women, and the trial will be open to women who are not otherwise practicing the interventions to any noteworthy degree. After identifying the potential target population for the intervention, there may still be reason to focus the trial on a subset of this population. There may be an identifiable subset at elevated risk for the primary outcome that could be chosen for trial participation, in order to reduce trial sample size. For example, it may be proposed that a colon cancer prevention strategy be assessed in subjects known to have had colonic polyps, or a breast cancer prevention strategy among women with a history of breast cancer among one or more first-degree relatives, even though it is hoped that the results will be applicable to a broader target population. There are several aspects to deciding on such an approach. First, although sample size may be reduced, trial logistics may be complicated and trial costs increased. For example, the costs of screening to identify eligible subjects will increase typically, and a larger number of participating clinical centers may be required. Depending on the intervention mechanism, high-risk study subjects may benefit less, by virtue of their stage in the targeted disease process, compared with other potential study subjects. Also, a focus on high-risk subjects may lead to a distorted view of the overall risks and benefits relative to the entire targeted population.

Within the target population, criteria may be needed to exclude study subjects with medical contraindications to either intervention or control regimens, study subjects who are already practicing the intervention to an unacceptable degree, or who may not adhere to intervention group requirements or to other protocol requirements (*see* **Eligibility and Exclusion Criteria**).

Even if study subjects are selected on the basis of elevated risk for the diseases that are targeted for prevention, primary outcome events may constitute a small minority of the disease events experienced by study subjects during the course of the trial, and perhaps even a small minority of disease events that may in some way be affected by intervention activities. Hence, there is an obligation to define sets of outcomes, to be carefully ascertained, including those that may plausibly be affected by intervention activities, in order to provide an opportunity to assess the overall risks and benefits in the target population.

The cost and logistics of a full-scale disease prevention trial may motivate a trial with some intermediate outcomes in place of the disease to be prevented. For example, a major trial was conducted in the US to prevent colonic polyps, rather than colon cancer, by means of a low-fat, high-fiber dietary pattern. This study makes the assumption that the formation of polyps is on the pathway between dietary habits and colon cancer occurrence, and that reduction in polyps occurrence will convey a corresponding reduction in colon cancer incidence. While the conditions for an intermediate outcome of this type to serve as a "surrogate" for the disease of interest are rather strict (see [4] and [13]), the benefits in terms of trial sample size, cost and duration may sometimes justify an intermediate outcome trial. In other circumstances, a trial with one or more intermediate outcomes may be conducted first to inform the decision concerning a trial with "harder" outcomes (*see* **Surrogate Endpoints**).

In some circumstances, the relationship between an intervention or behavior and a reduction in disease will be regarded as sufficiently well established that the research effort can shift logically to strategies to encourage the desired behavior change. Cigarette

smoking cessation or prevention in relation to lung cancer and heart disease, or breast screening by means of mammography and other techniques provide important examples. Randomized trials with such behaviors as outcomes can be classified as disease control research, rather than primary prevention research. In such studies, the intervention may sometimes be able to be delivered with particular economy to persons in natural groups, such as social groups, schools, or communities. In fact, the use of community organizations and media may even define the intervention strategy, as in the Community Intervention Trial for Smoking Cessation, which took place in 11 pairs of matched cities in the US. Such studies naturally involve **group randomization** and there is a range of interesting design and analysis issues [5, 6].

Returning to individually randomized prevention trials with disease outcomes, other design choices include the possible use of **factorial** designs, and intervention and control randomization fractions. Factorial designs have an obvious appeal in that they provide the potential to make two or more intervention comparisons in the same study population at a cost that will typically be considerably less than that for separate studies. For example, in the WHI trial, study subjects must be eligible and willing to participate in one or more of the hormone replacement or dietary intervention components, and, subsequently, are offered the opportunity to participate in the calcium and vitamin D component – a so-called partial factorial design. There is only a modest overlap between the hormone replacement and dietary intervention components due to component-specific exclusionary criteria, but a large overlap of either of these with the calcium and vitamin D component. As a result, the projected trial sample size is 68 132 rather than the 120 000 or more that would be required to assess the three interventions separately. The potential disadvantages to a factorial design are the possibility that the benefit associated with an intervention may be reduced by the presence of one or more other interventions, and the possibility that adherence to a given intervention may be reduced by the study demands or adverse effects that may arise from participation in the other interventions (*see* **Factorial Designs in Clinical Trials**).

The necessary sample size of a two-arm trial is approximately proportional to $[\gamma(1-\gamma)]^{-1}$, where γ is the fraction of the trial cohort assigned to the intervention group. Hence, if the average study costs associated with an intervention group subject are C times that for the corresponding control group, subject trial costs will be approximately proportional to $[C\gamma(1-\gamma)][\gamma(1-\gamma)]^{-1}$, which is minimized by setting $\gamma = (1 + C^{1/2})^{-1}$. For example, if study costs per intervention group subject are 2.25 times that per control group subject, then $\gamma = 0.4$, a randomization fraction that is used for the dietary intervention component of the WHI.

Upon selecting the interventions to be evaluated, the target population, and major trial outcomes, one needs to make a series of design assumptions that will determine the size of the trial cohort. Perhaps the most fundamental assumption concerns the anticipated primary endpoint intervention benefit, often expressed as a **relative risk** (hazard ratio) for fully adherent intervention vs. fully adherent control subjects as a function of time from randomization. Assumptions concerning primary endpoint disease rates in the absence of intervention, on intervention and control group adherence rates and accrual patterns, trial duration and **competing risks** can then be combined with the basic relative risk assumption to produce the sample size that will yield a significant result (e.g. at the 0.05 significance level) under design assumptions, with a specified probability or **power** (e.g. power of 90%). Various authors, including Self & Mauritsen [18], provide flexible sample size and power procedures that allow one to incorporate assumptions of this type. The WHI Study Group [21] details such assumptions for the WHI clinical trial. Corresponding primary endpoint power calculations played a major role in the specification of sample sizes of 48 000, 27 500, and 45 000 for the dietary modification, hormone replacement therapy, and calcium and vitamin D trial components, respectively.

Pilot and feasibility studies play a critical role in prevention trial planning. Such studies provide the opportunity to assess study subject recruitment rates, to evaluate the potential of a run-in period to identify and exclude study subjects who may not comply with study requirements, to observe biomarker changes that may help to establish the basic relative risk assumption, and to assess costs associated with all aspects of at least the early phases of trial operation. Information on these topics can be critically important to the development of an efficient trial design. See Urban et al. [19] for an example of the use of cost projections to inform the design choices for a low-fat

diet intervention trial, including eligibility criteria, average follow-up duration, randomization fraction and number of clinical centers. Careful consideration of subsampling rates for the collection and processing of baseline and follow-up data and specimens can also play an important role in controlling trial costs.

Conduct, Monitoring, and Analysis

A disease prevention trial requires a clear, concise protocol that describes trial objectives, design choices, performance goals and monitoring and analysis procedures. A detailed manual of procedures that describes how the goals will be achieved is necessary to ensure that the protocol is applied in a standardized fashion (*see* **Clinical Trials Protocols**). Carefully developed data collection and management tools and procedures, with as much automation as practical, can also enhance trial quality. Centralized training of key personnel may be required to ensure that the protocol is understood, and to enhance study subject recruitment, intervention adherence, and comparability of outcome ascertainment, possibly through **blinding** of the randomization assignment between intervention and control groups (*see* **Multicenter Trials**). A committee knowledgeable in the various aspects of the trial, and often external to the investigative group, typically will be required for the timely review of safety and clinical outcome data (*see* **Data Monitoring Committees**).

As mentioned previously, prevention trial monitoring for early stoppage (*see* **Data and Safety Monitoring**) will usually not only involve the designated primary outcome(s), but also some suitable measure of overall benefit vs. risks, as well as of important adverse effects. Some aspects of the proposed monitoring of the WHI clinical trial are described in Freedman et al. [3]. For example, early stoppage for benefit may be merited if the primary outcome incidence reduction is significant at customary levels ($p < 0.05$) and the summary benefit vs. risk measure is supportive (e.g. $p < 0.20$) without important adverse effects. Early stoppage based on harm may be indicated if an important adverse event is significant ($p < 0.05$) without suggestive evidence ($p > 0.20$) of benefit vs. risk. More sophisticated stopping criteria could also be considered and critical values that acknowledge the multiplicity of outcomes and of testing times need to be constructed (*see* **Multiplicity in Clinical Trials**). See Cook [1] for an example of such a construction for a bivariate response.

The basic test statistic to compare two randomization groups with respect to a failure time (disease) endpoint might often reasonably be defined to be of weighted **logrank** form:

$$\sum_{i=1}^{n} \delta_i g(t_i)[z_i - n_1(t_i)n(t_i)^{-1}],$$

where n is the total number of intervention and control study subjects, $n_1(t_i)$ and $n(t_i)$ are, respectively, the number of intervention subjects and total subjects at risk for failure at the failure time ($\delta_i = 1$) or **censoring** time ($\delta_i = 0$) for the ith subject, while z_i indicates whether the ith subject is assigned to intervention ($z_i = 1$) or control ($z_i = 0$). The test will have high efficiency if the "weight" $g(t)$ at time t from randomization is chosen to approximate the logarithm of the anticipated intervention vs. control group hazard ratio for the endpoint under test, taking account of anticipated adherence rates. For example, if this hazard ratio is expected to be approximately constant, then one might set $g(t) \equiv 1$, in which case one has the classical logrank test, while if the hazard ratio is expected to decline approximately exponentially over the follow-up period, then one might set $g(t) = t$. Adaptive versions in which the form of $g(t)$ is responsive to the evolving trial data may also be considered.

The above test can be generated as a **partial likelihood** [2] score test for $\beta = 0$ under a hazard ratio model $\exp[x(t)/\beta]$, where $x(t) = zg(t)$. This modeled regression vector can be extended to include other variables that may be intermediate between intervention activities and outcome events, in an attempt to explain intervention effects on disease outcomes. The trial monitoring process will have some effect on these tests and estimators, with typically larger effects if early stoppage occurs. The estimation of intervention effects may be biased (e.g. [20]) even for outcomes that do not contribute to early stoppage decisions. Analyses that attempt to explain intervention effects in terms of intermediate measures can often be based efficiently on case–control [8] or **case–cohort** [12] subsampling procedures, and should acknowledge measurement error in the intermediate variable assessment. See [22], which provides basic results from the WHI Clinical Trial component on combined hormones, following early

stoppage based on risks & feeding benefits as and example of important and unexpected trial results, and of the complexity of prevention trial reporting.

Acknowledgments

This work was supported by grant CA-53996 from the National Cancer Institute. This entry builds upon Prentice [14], which includes additional discussion of a number of the previously mentioned technical issues.

References

[1] Cook, R.J. (1996). Coupled error spending functions for parallel bivariate tests, *Biometrics* **52**, 422–450.

[2] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[3] Freedman, L.S., Anderson, G.A., Kipnis, V., Prentice, R.L., Wang, C.Y., Rossouw, J., Wittes, J. & DeMets, D.L. (1996). Approaches to monitoring the result of long-term disease incidence prevention trials: examples of the Women's Health Initiative, *Controlled Clinical Trials*, **17**, 509–525.

[4] Freedman, L.S., Graubard, B.I. & Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases, *Statistics in Medicine* **11**, 167–178.

[5] Freedman, L.S., Green, S.B. & Byar, D.P. (1990). Assessing the gain in efficiency due to matching in a Community Intervention Study, *Statistics in Medicine* **9**, 943–952.

[6] Gail, M.H., Byar, D.P., Pehacek, T.F. & Corle, D.K. (1992). Aspects of the statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT), *Controlled Clinical Trials* **13**, 6–21.

[7] Hypertension Detection and Follow-up Program Cooperative Group (1979). Five year findings of the Hypertension Detection and Follow-up Program 1. Reductions in mortality of persons with high blood pressure, including mild hypertension, *Journal of the American Medical Association* **242**, 2562–2571.

[8] Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977). Methods for cohort analysis: appraisal by application to asbestos mining data (with discussion), *Journal of the Royal Statistical Society, Series A* **140**, 469–490.

[9] Lipid Research Clinic Program (1984). The Lipid Research Clinic Coronary Primary Prevention Trial Results. I. Reduction in incidence of coronary heart disease, *Journal of the American Medical Association* **251**, 351–364.

[10] Multiple Risk Factor Intervention Trial (MRFIT) Research Group (1982). Multiple Risk Factor Intervention Trial: risk factor changes and mortality results, *Journal of the American Medical Association* **248**, 1465–1477.

[11] PEPI Trial Writing Group (1995). Effects of estrogen or estrogen/progestin regimens on heart disease risk factors in postmenopausal women: the Post-menopausal Estrogen/Progestin Intervention (PEPI) Trial, *Journal of the American Medical Association* **273**, 199–208.

[12] Prentice, R.L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[13] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: discussion, definition and operational criteria, *Statistics in Medicine* **8**, 431–440.

[14] Prentice, R.L. (1995). Experimental methods in cancer prevention research, in *Cancer Prevention and Control*, P. Greenwald, B.S. Krawer & D.L. Weed, eds. Marcel Dekker, New York, pp. 213–224.

[15] Prentice, R.L. (1996). Measurement error and results from analytic epidemiology: dietary fat and breast cancer, *Journal of the National Cancer Institute* **88**, 1738–1747.

[16] Prentice, R.L. & Sheppard, L. (1995). Aggregate data studies of disease risk factors, *Biometrika* **82**, 113–125.

[17] Rossouw, J.E., Finnegan, L.P., Harlan, W.R., Pinn, V.W., Clifford, C. & McGowan, J.A. (1995). The evolution of the Women's Health Initiative: prospectives from the NIII, *Journal of the American Medical Women's Association* **50**, 50–55.

[18] Self, S.G. & Mauritsen, R. (1988). Power/sample size calculations for generalized linear models, *Biometrics* **44**, 79–86.

[19] Urban, N., Self, S., Kessler, L., Prentice, R., Handerson, M., Iverson, D., Thompson, D.L., Byar, D., Insull, W., Gorach, S.G., Clifford, C. & Goldman, S. (1990). Analysis of the costs of a large prevention trial, *Controlled Clinical Trials* **11**, 129–146.

[20] Whitehead, J. (1986). Supplementary analysis at the conclusion of a sequential clinical trial, *Biometrics* **42**, 461–471.

[21] Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative Clinical Trial and Observational Study, *Controlled Clinical Trials* **19**, 61–109.

[22] Writing group for the Women's Health Initiative. (2002). Risk and benefit of estrogen plus progestron in healthy postmenopausal women, *Journal of the American Medical Association* **288**, 321–330.

ROSS L. PRENTICE

Preventive Medicine

Statistics plays a central role in epidemiology and preventive medicine, particularly in interpretation of behavior, physiology, and pathology in groups of people. Preventive medicine follows in a political setting from findings of epidemiology and **clinical trials**, and modes of disease causation and transmission (*see* **Communicable Diseases**). Emphasized are the classical strategies of hygiene and avoidance of contact with infectious agents, as well as more recently discovered community and personal actions which can reduce the burden of chronic diseases; for example, heart disease, stroke, cancer, **AIDS and HIV**, injury, and violence.

History and Context

Excellent reviews of the history of epidemiology are available [7, 16, 21], and of statistics in epidemiology [8]. In 1662, a **life table** approach used London birth and death records to study seasonal variations in **infant mortality**, and excess male mortality compared with females [9]. A rudimentary clinical trial in 1747 [16, 17] addressed epidemic scurvy on long ocean voyages. Lind allocated 12 sailors to six dietary treatments for six days, and observed a near cure in the two who ate citrus fruit each day, with no improvement in the other 10 men. Following this finding, Captain James Cook lost no one to scurvy in his voyages of 1769–1778 [10]. Despite these dramatic findings, for political reasons it was 1795 before the British Navy (the "limeys") routinely added limes to its diet. Ignaz Semmelweis, in 1840, introduced hand-washing in a Vienna hospital, and observed a reduced rate of puerperal fever in maternity wards [7]. In the 1849 London cholera epidemic, John Snow counted cases according to residence. The epidemic was restricted to a geographic area served by two water companies which used sewage-polluted water from the Thames River; access to this water was blocked as a preventive action [20].

However, the population approach was virtually derailed by a medical advance; namely, the Henle–Koch postulates of the late 1800s, that diseases are caused by specific living organisms [7, 21]. These postulates led epidemiologists to focus on laboratory-based methodology for studying infectious diseases [21]. It was not until the mid-1900s that population approaches again emerged, led by suspicions that an environmental factor, smoking, caused lung cancer (*see* **Smoking and Health**), [21], by the study of risk factors for cardiovascular disease in the **Framingham study** [4], and by reduction of dental decay by use of fluoridated water in one of two New York communities [1, 16]. Currently, the population approach is flourishing, and should be of substantial value should the threat of antibiotic-resistant infections materialize, due to evolution of micro-organisms. The study of populations has extended into such fields as establishment of population norms, control of fertility, and **population genetics** [21]. The mapping of the human genome will allow an epidemiologic approach to defining the **gene–environment interaction** in human disease.

Modern epidemiologic design was not described in a textbook until 1960 [18, 21]. Design considerations and basic statistical methods relied heavily on the work of **R.A. Fisher** [5, 6]. The seminal paper alerting epidemiologists to multivariate analysis was published by **Jerome Cornfield** in 1962, using Framingham data [3]. The importance of the community itself was illustrated in Ancel Keys' cross-cultural studies of coronary heart disease, beginning in the 1950s [14].

The use of statistics in epidemiology is strongly influenced by need for causal inference (*see* **Causation**). If epidemiologic associations are not causal, then the corresponding preventive strategy – for example, using safer sexual practices to avoid AIDS – will not reduce disease risk. Fundamental principles to be considered when judging causality of an association were published by **A. Bradford Hill** [11, 12], including strength of association, consistency with other related observations, temporality (presumed cause precedes presumed effect) biological gradient, biological plausibility, and coherence with other relevant knowledge (*see* **Hill's Criteria for Causality**). Experimental design and statistical methods used in epidemiology are often geared toward answering questions about causality.

Specific Uses of Statistics in Epidemiology

Many texts describe epidemiologic methods [15] and related statistical methods [13]. Computer software

has made statistics accessible to nonstatisticians. Central to all statistical methods used in epidemiology is counting cases, most simply in the **two-by-two table**, from which may be computed the disease rates or **risk** in exposed (r_E) and in unexposed (r_U) persons. The term "rate" pertains to prevalent disease (existent at the time of study) while the term "risk" refers to incident disease (develops during the period of study). Large **relative risk** (r_E/r_U) is seen for exposures which are etiologically related to the disease in question, while large risk differences ($r_E - r_U$) in highly prevalent diseases suggest a large population **burden of disease** due to exposure.

The sampling method influences the estimation of **absolute risk** and differential risk. In the longitudinal **cohort study**, exposed and unexposed persons may be sampled at different sampling fractions. Because incident disease develops within exposure categories, diseased and nondiseased people have the same sampling fraction. The clinical trial is similar, but exposure status is (randomly) allocated by the investigator (*see* **Randomization**), participant inclusion characteristics are typically much narrower than in the general longitudinal study (*see* **Eligibility and Exclusion Criteria**), and explicit participant informed consent for randomization is required. In these longitudinal designs, r_E is estimated as the number diseased divided by the number exposed, and analogously for r_U. In **cross-sectional studies**, where exposure and disease status are determined at the same time, the analogous estimator is of a rate rather than a risk.

A different estimator is used in the **case–control study** [2, 19], which is useful in limiting the numbers of nondiseased subjects needed to study rare diseases. Diseased persons are sampled at one sampling fraction and nondiseased at another (typically much smaller) sampling fraction. Within disease categories, exposed and unexposed people are sampled at the same rate. In this case r_E and r_U are **biased**, but r_E/r_U is well approximated for rare diseases by the **odds ratio** (the odds of exposure to nonexposure in diseased divided by the odds of exposure to nonexposure in nondiseased). Case–control studies may be biased by changes in exposure made after disease onset. Attempts to correct this problem include retrospective recall of predisease exposure, and nesting cases and controls in previously collected data (*see* **Case–Control Study, Nested**).

The generalization of these designs to several categories or to continuous exposure is mathematically straightforward, the former being analyzed with a series of risk estimates, and the latter with a linear or other continuous **regression** model of risk change per unit change in exposure; **life table** versions of regression account for length of exposure. "Disease" variables may also be continuous.

Problems encountered throughout epidemiologic statistics include within-person variation and **confounding**. Because population studies of humans do not perfectly control the measurement condition, it is rare that a measure represents the participant exactly. The resulting within-person variation tends to bias differential risk estimates falsely toward zero (*see* **Bias Toward the Null**). Confounding, on the other hand, occurs when variables are correlated for reasons unrelated to the causality of the association in question. Proper inference demands that "like be compared with like" [8], whether through deconfounding or randomization. In deconfounding for drinking, differential risk of lung cancer, for example, is estimated for smokers vs. nonsmokers, first within drinkers and secondly within nondrinkers. The adjusted estimate is pooled across drinkers and nondrinkers, using standard weights for the drinking strata. Regression methods for deconfounding extend this process to multiple continuous variables; in most such methods the stratified analysis is implicit (*see* **Stratification**).

Much effort is expended in epidemiologic statistics on the ***P* value**, the chance that a given observation would have arisen if risk did not vary across exposure categories. This method is fraught with interpretive difficulties, because *P* values depend on valid model specification, on the number of comparisons made, and on the context of the interpreter. An emerging methodology is **meta-analysis**, in which replications of a particular study are formally examined for consistency. Drawbacks are that meta-analysis loosens definitions of what is being studied, and is subject to bias because studies with null findings may be less available than studies with provocative findings. Strengths are that it tends to reduce dependence on *P* values because of larger sample sizes in the combined studies than in individual studies; it increases inferential dependence on consistency across studies.

References

[1] Ast, D.B. & Schlesinger, E.R. (1956). The conclusion of a ten-year study of water fluoridation, *American Journal of Public Health* **46**, 265–271.

[2] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1. *The Analysis of Case–Control Studies*. International Agency for Research on Cancer, WHO, Geneva.

[3] Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure; a discriminant function analysis, *Federation Proceedings* **2**, 58–61.

[4] Dawber, T.R. (1980). *The Framingham Study: the Epidemiology of Coronary Heart Disease*. Harvard University Press, Cambridge, Mass.

[5] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[6] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[7] Friis, R.H. & Sellers, T.A. (1996). *Epidemiology for Public Health Practice*. Aspen, Gaithersburg, Maryland.

[8] Gail, M.H. (1996). Statistics in action, *Journal of the American Statistical Association* **91**, 1–13.

[9] Graunt, J. (1939). *Natural and Political Observations Made Upon the Bills of Mortality*. Johns Hopkins University Press, Baltimore.

[10] Gray, W.R. (1981). *Voyages to Paradise: Exploring in the Wake of Captain Cook*. National Geographic Society, Washington, DC, p. 45.

[11] Greenland, S., ed. (1987). *Evolution of Epidemiologic Ideas, Annotated Readings on Concepts and Methods*. Epidemiology Resources, Inc., Chestnut Hill.

[12] Hill, B.A. (1965). The environment and disease: association or causation, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[13] Kahn, H.A. & Sempos, C.T. (1989). *Statistical Methods in Epidemiology*. Oxford University Press, Oxford.

[14] Keys, A. (1980). *Seven Countries: a Multivariate Analysis of Death and Coronary Heart Disease*. Harvard University Press, Cambridge, Mass.

[15] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research. Principles and Quantitative Methods*. Lifetime Learning Publications, Wadsworth, Belmont. California.

[16] Lilienfeld, A.M. (1994). *Foundations of Epidemiology*, 3rd Ed., revised by David E. Lilienfeld & Paul D. Stolley. Oxford University Press, New York.

[17] Lind, J. (1753). *A Treatise on the Scurvy*. Sands, Murray & Cochran, Edinburgh.

[18] MacMahon, B., Pugh, T.F. & Ipsen, J. (1960). *Epidemiological Methods*. Little, Brown, & Company, Boston.

[19] Schlesselman, J.J. (1982). *Case-Control Studies: Design Conduct, Analysis*. Oxford University Press, New York.

[20] Snow, J. (1965). *Snow on Cholera*, Harvard University Press, Cambridge, Mass.

[21] Susser, M. (1985). Epidemiology in the United States after World War II: the evolution of technique, *Epidemiologic Reviews* **7**, 147–177.

DAVID R. JACOBS, JR

Primary Factors

One of the aims of **factor analysis** is to display the configuration of variables in as simple a manner as possible. The **eigenvalue** technique was used to find a set of axes to which variables could be referred. However, the eigenanalysis leads to a solution such that the variance of the variables is distributed across all factors. Thus, the factors do not separate out independent clusters of variables (*see* **Cluster Analysis, Variables**). **Rotation of axes** is an attempt to place the factors so that each contains only a few highly loaded variables. *Primary axes* are the factor axes that are rotated to a good fit to the configurations of the variables involved [2]. They are unit vectors that lie in the intersection of the primary factors. In the two-factor case, primary axes coincide with the primary factors. The *primary factors* represent the "ideal" variables, each of which measures one and only one common factor, with no unique factors. They provide the effective boundaries for the variables involved. *Reference vectors* are defined as the axes that are perpendicular to each of the primary factors. With the **oblique rotation**, the reference vectors are so located as to maximize the number of near-zero loadings on each of the reference-vector factors, with at least as many near-zero loadings on each factor as the number of factors. They produce a rotated factor matrix of a **simple structure**. These axes are important concepts for the understanding of factor rotation and the interpretation of the different **factor loading matrices**. Following the discussion in Cureton & D'Agostino [1], we illustrate and describe these axes in terms of the geometric model.

In a two-factor geometric model, we can represent the effective common-factor vector by a semicircle with axes a (vertical) and b (horizontal) as the initial axes. This is obtained by reflecting any negative first-factor loadings of any variables. Figure 1 presents this two-dimensional geometric model. In Figure 1, the axes 0I and 0II are vectors of unit length extending from the origin to the perimeter of the semicircle and provide the effective boundaries for the configuration of variable vectors. Variable vectors are vectors that extend from the origin to the plotted points of the variables. The locations of boundary vectors are determined by the type of transformation we apply to the initial factor matrix. Ideally, most of the variable vectors should lie on or close to the boundary vectors 0I and 0II. In Figure 1, we have variables 1, 5, and 6 lying close to the boundary vector 0I. Therefore, boundary 0I is overdetermined by variables 1, 5, and 6. Similarly, boundary vector 0II is overdetermined by variables 2, 3, and 7. Even though variable 4 is not close to either vector, the structure is quite clear. The reference vectors, labeled RI and RII, are in Figure 1. RI is the reference vector that is orthogonal (*see* **Orthogonality**) to the axis 0I and RII is the reference vector that is orthogonal to the axis 0II. In other words, the axis 0I is **correlated** with the reference vector RII but uncorrelated with the reference vector RI. Similarly, the axis 0II is correlated with the reference vector RI but uncorrelated with the reference vector RII. By definition, each reference vector is correlated with the corresponding primary axis and uncorrelated with the other. Therefore, in the two-dimensional case the boundary vector 0II coincides with the primary axis PI, and the boundary vector 0I coincides with the primary axis PII. The primary axes correspond to the first and second primary factors and geometrically the axes represent the primary factors.

In higher-dimensional cases ($m \geq 3$), the effective boundaries become planes ($m = 3$) or hyperplanes ($m > 3$) and the primary axes of the primary factors are the vectors lying at the intersections of these planes or hyperplanes. For the three-dimensional case, we can represent the effective common-factor vector space by a hemisphere of unit radius lying on the top of the unit circle made by the axes b and c (see Figure 2). This effective vector space is obtained by reflecting the negative first-factor loadings. Figure 2 presents the geometric model for the three-factor case. The hyperplanes are the bounding planes labeled as HI, HII, and HIII. These planes can be thought of as an inverted triangular

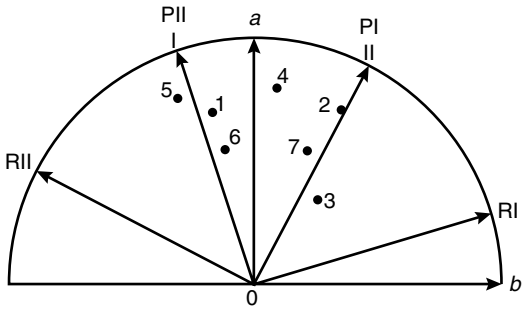

Figure 1 Geometric model for the two-factor case

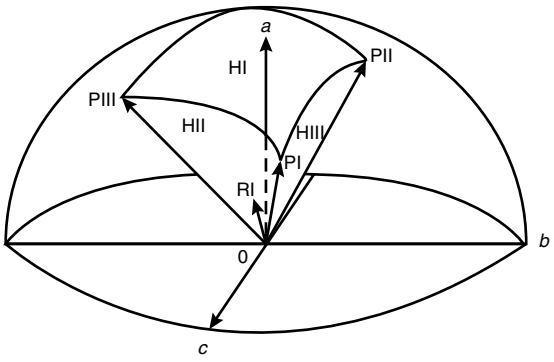

Figure 2 Geometric model for the three-factor case

pyramid with apex at the origin 0, and generate a spherical triangle at the top where they meet the surface of the hemisphere. Ideally, most of the variable vectors should lie on or close to the planes HI, HII, and HIII. The lines where these planes meet are called the primary axes. They are all unit vectors. We can see from Figure 2 that the primary axis 0PI is at the intersection of the planes HII and HIII. The primary axis 0PII is at the intersection of planes HI and HIII. The primary 0PIII is at the intersection of planes HI and HII. For reference vectors, it is not easy to locate them visually in this three-dimensional figure. Similar to the two-factor case, the reference vectors are the unit vectors that are orthogonal to the primary factors (here the planes HI, HII, and HIII). The reference vector RI is orthogonal to HI, which in turn is orthogonal to the primary axes PII and PIII since both primary axes lie in HI. Therefore, as defined, RI is correlated with the corresponding primary axis PI and uncorrelated with the other two. Similar descriptions apply to RII and RIII. The reference vector RI, greatly shortened, is contained in Figure 2.

In the m-dimensional case ($m > 3$), the sphere becomes a hypersphere. Again, by reflecting all the negative first-factor loadings, we obtain the effective vector space bounded by the part of hypersphere where all the a coordinates are positive. The three planes now become m hyperplanes of dimensionality $m - 1$. These m hyperplanes intersect at the origin. They form an inverted hyperpyramid and produce a hyperspherical triangle that meets the surface of the hyperhemisphere. Again, these hyperplanes provide effective boundaries for the variable vectors. The primary axes are the intersections of these hyperplanes, taken $m - 1$ at a time. As defined previously, the reference vector is correlated with the corresponding primary axis but uncorrelated with all other. So the reference vector RI is orthogonal to the hyperplane HI, which is therefore orthogonal to PII, . . . , PM because the primary axes PII, . . . , PM all lie in the hyperplane HI. In other words, RI is correlated with PI but uncorrelated with all other $m - 1$ primary axes. The same descriptions can be applied to the rest of the reference vectors.

As discussed in **Factor Loading Matrix**, different pattern and structure matrices can be obtained as the projections of variable vectors onto the different set of axes. They play an important role in the interpretation of the results.

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.
[2] Rummel, R.J. (1970). *Applied Factor Analysis*. Northwestern University Press, Evanstonm.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Principal Components Analysis

Principal components analysis is a method for transforming a set of n **correlated** variables, $X_1, X_2, \ldots, X_n$, to m uncorrelated variables, $Y_1, Y_2, \ldots, Y_m$, where $m \leq n$, and the variances of the Ys are in descending order with the sum of these m variances equal to the "salient" or nonrandom variance of the Xs. There are a number of related aims of principal components, some of which are the following:

1. To *produce n uncorrelated variables*. Transform the original n correlated variables to n uncorrelated linear functions of the original variables.
2. To *produce the best single composite variable*. Generate a single linear composite function of the original variables which has maximum variance among all possible linear functions of the original n variables (or which maximally discriminates the subjects in the data set).
3. To *explain the salient variance*. Transform the n correlated variables to m uncorrelated variables ($m < n$) which explain the salient variance of the n original variables.
4. To *reduce dimension*. Start with the n dimensions of the original variables, and identify m dimensions ($m < n$), where the m dimensions explain the salient variance of the original n variables.
5. To *produce clusters and cluster scores*. Cluster the original n variables into m ($m < n$) subsets (possibly exclusive) and generate simple scores for these m subsets.
6. *Battery reduction*. Identify m variables from the original n variables ($m < n$) which reproduce the salient variance of the original n variables.
7. To *produce scales*. Generate from the original variables composites or "scales" that measure the dimensions underlying the original data.
8. To produce *uncorrelated regressor variables*. Produce uncorrelated variables from a set of correlated variables and use the uncorrelated variables rather than the original variables as regressor variables in a **regression** analysis.

Development of the Principal Components

The principal components are linear functions of the original variables of the form

$$Y_j = e_{1j}X_1 + e_{2j}X_2 + \cdots + e_{nj}X_n, \qquad (1)$$

with

$$e_{1j}^2 + e_{2j}^2 + \cdots + e_{nj}^2 = 1, \qquad (2)$$

for $j = 1, \ldots, n$. We can understand them best if we view them as generated in a sequential manner. The first principal component is the linear function of the form (1) subject to (2) for $j = 1$, where the variance of Y_1 has the maximum variance over all possible linear functions of the original variables subject to (2). We call this variance λ_1.

The second principal component, Y_2, is the linear function, uncorrelated with Y_1 with the next largest variance $\lambda_2 \leq \lambda_1$. In a similar fashion, each principal component is uncorrelated with all others and has the largest possible variance λ_j subject to

$$\lambda_1 \geq \cdots \geq \lambda_n. \qquad (3)$$

In practice, usually there are strict inequalities in (3).

Thus, principal components are linear transformations of the original variables, uncorrelated with each other and with decreasing variance. One can perform the principal components analysis on the Xs in their original scale, i.e. on the raw data. However, we usually perform it on standardized variables, i.e. variables derived from the original variables by subtracting the mean and dividing by the standard deviation. The principal components derived from the raw data will be different from those obtained on the standardized data. If performed on the original variables, the variables with the largest variances can dominate the results. When the analysis is performed on standardized data, all variables are on equal footing.

The method of principal components requires no assumptions about the data. Usually the sample size is large. It can be applied to a **random sample** from some population. However, in practice this is often not the case, and frequently there is interest solely in understanding the sample data themselves.

Mathematical Derivation

Components of Original Variables

Assume a sample of size N is available and on each individual n variables $X_1, \ldots, X_n$ are measured. Let

S represent the sample variance–**covariance matrix**. The sample variance of a linear function of the form (1) of the original n variables is

$$\text{var}(Y_j) = \mathbf{e'}_j \mathbf{S} \mathbf{e}_j, \tag{4}$$

where $\mathbf{e'}_j$ is the vector of weights $(e_{1j}, e_{2j}, \ldots, e_{nj})$ of (1) subject to the restriction of (2) which, in vector notation, is

$$\mathbf{e'}_j \mathbf{e}_j = 1. \tag{5}$$

Condition (2) [or (5)] is arbitrary and only guarantees a unique solution. Any multiple of $\mathbf{e}_j$ produces basically the same component.

The problem of finding the $\mathbf{e}_j$ becomes that of finding the maximum of (4) subject to (5), or finding the maximum of

$$\mathbf{e'}_j \mathbf{S} \mathbf{e}_j - \lambda_j (\mathbf{e'}_j \mathbf{e}_j - 1). \tag{6}$$

Obtaining the solution to (6) is a simple **matrix algebra** exercise where the λ_j, for $j = 1, \ldots, n$, are the solutions of

$$|\mathbf{S} - \lambda \mathbf{I}| = 0. \tag{7}$$

These λ_j are the **eigenvalues** of **S**. The $\mathbf{e}_j$ are then found for each λ_j as the solution of

$$(\mathbf{S} - \lambda_j \mathbf{I}) \mathbf{e}_j = \mathbf{0}, \tag{8}$$

for $j = 1, \ldots, n$. These are the **eigenvectors** or characteristic vectors of **S**. Because of this, $\mathbf{e'}_j \mathbf{e}_t = 0$ for $j \neq t$, which ensures that the principal components are uncorrelated.

It should be noted that the sum of the eigenvalues equals the *trace* of **S**, which is the sum of the variances of the Xs. So the sum of the variances of the n principal components is equal to the sum of the variances of the original variables.

Geometrically, $\mathbf{e}_j$ is the vector of direction cosines of the line from the vector of means of the Xs through the direction of the jth greatest variation in the n-dimensional plot of the data. When the λ_j are distinct, the successive axes are perpendicular to one another. See [7, p. 9] for more details.

Components of Standardized Variables

Because the scales of the original variables are often different, the method of principal components is usually applied to the standardized variables. In this case consider the Xs as being standardized (mean 0 and variance 1), and replace the variance–covariance matrix **S** above with the sample correlation matrix **R**. The problem of finding the appropriate $\mathbf{e}_j$ then becomes one of finding the maximum of

$$\mathbf{e'}_j \mathbf{R} \mathbf{e}_j - \lambda_j (\mathbf{e'}_j \mathbf{e}_j - 1). \tag{9}$$

The λ_j and $\mathbf{e}_j$ of (9), for $j = 1, \ldots, n$, are the solutions of

$$|\mathbf{R} - \lambda \mathbf{I}| = 0 \quad \text{and} \quad (\mathbf{R} - \lambda_j \mathbf{I}) \mathbf{e}_j = \mathbf{0}. \tag{10}$$

The λ_j are the eigenvalues of **R** and the $\mathbf{e}_j$ are the eigenvectors of **R**. So the jth principal component is $Y_j = \mathbf{e'}_j \mathbf{X}$, where **X** is the vector of the standardized original variables and the variance of Y_j is λ_j. Here the sum of the eigenvalues (variances of Y_j) equals n, which is the trace of **R**. In most of what follows we assume we are dealing with standardized data.

One can perform principal components using standard procedures in readily available **software** packages such as the SAS [16] procedure PRINCOMP. This procedure will automatically standardize the variables.

Example

In the **Framingham Heart Study**, a 10-question depression scale was administered ($n = 10$) where the responses were No or Yes to the following (note that the terms in parentheses will be used in the following in reference to the variables);

1. I felt everything I did was an effort (EFFORT)
2. My sleep was restless (RESTLESS)
3. I felt depressed (DEPRESS)
4. I was happy (HAPPY)
5. I felt lonely (LONELY)
6. People were unfriendly (UNFRIEND)
7. I enjoy life (ENJOYLIF)
8. I felt sad (FELTSAD)
9. I felt that people dislike me (DISLIKED)
10. I could not get going (GETGOING)

A Yes was scored as 1 and No as 0, except for questions 4 and 7, where the scoring was reversed

Table 1 Correlation matrix for the Framingham Heart Study depression data

	EFFORT	REST-LESS	DEPRESS	HAPPY	LONELY	UNFRIEND	ENJOY-LIF	FELTSAD	DIS-LIKED	GET-GOING
EFFORT	1.000	0.218	0.348	0.320	0.235	0.173	0.228	0.273	0.171	0.440
RESTLESS		1.000	0.216	0.187	0.171	0.084	0.096	0.201	0.092	0.227
DEPRESS			1.000	0.514	0.453	0.192	0.291	0.558	0.175	0.333
HAPPY				1.000	0.347	0.118	0.417	0.396	0.112	0.314
LONELY					1.000	0.128	0.257	0.497	0.105	0.231
UNFRIEND						1.000	0.120	0.132	0.383	0.158
ENJOYLIF							1.000	0.328	0.051	0.157
FELTSAD								1.000	0.149	0.265
DISLIKED									1.000	0.197
GETGOING										1.000

Table 2 Eigenvector matrix and initial component matrix

Eigenvector matrix (**e**)

	PRIN1 ($\mathbf{e}_1$)	PRIN2 ($\mathbf{e}_2$)	PRIN3 ($\mathbf{e}_3$)	PRIN4 ($\mathbf{e}_4$)	PRIN5 ($\mathbf{e}_5$)	PRIN6 ($\mathbf{e}_6$)	PRIN7 ($\mathbf{e}_7$)	PRIN8 ($\mathbf{e}_8$)	PRIN9 ($\mathbf{e}_9$)	PRIN10 ($\mathbf{e}_{10}$)
EFFORT	0.327	0.135	0.406	−0.358	−0.182	0.321	0.192	−0.619	0.142	0.068
RESTLESS	0.213	0.058	0.540	0.516	0.626	−0.012	0.013	−0.004	0.016	−0.017
DEPRESS	0.419	−0.117	−0.102	0.141	−0.171	−0.166	−0.389	−0.230	−0.196	−0.693
HAPPY	0.381	−0.204	−0.059	−0.251	0.149	−0.405	−0.458	0.009	0.421	0.417
LONELY	0.348	−0.202	−0.207	0.380	−0.250	0.338	0.285	0.245	0.572	−0.070
UNFRIEND	0.191	0.593	−0.331	−0.012	0.198	0.518	−0.412	0.100	−0.063	0.104
ENJOYLIF	0.286	−0.239	−0.267	−0.481	0.543	0.092	0.395	0.154	−0.144	−0.233
FELTSAD	0.390	−0.198	−0.199	0.286	−0.154	−0.007	0.159	−0.122	−0.596	0.517
DISLIKED	0.185	0.635	−0.222	0.082	−0.022	−0.557	0.413	−0.088	0.125	−0.045
GETGOING	0.317	0.176	0.466	−0.240	−0.320	−0.051	0.000	0.669	−0.202	−0.029
Eigenvalue (λ_j)	3.358	1.290	1.022	0.872	0.795	0.627	0.590	0.552	0.508	0.386

Initial component matrix, **A**, elements $= \sqrt{(\lambda_j)}\mathbf{e}_j$

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	PRIN8	PRIN9	PRIN10
EFFORT	0.599	0.153	0.410	−0.334	−0.162	0.254	0.147	−0.460	0.101	0.042
RESTLESS	0.390	0.066	0.546	0.482	0.558	−0.010	0.010	−0.003	0.011	−0.011
DEPRESS	0.768	−0.133	−0.103	0.132	−0.152	−0.131	−0.299	−0.171	−0.140	−0.431
HAPPY	0.698	−0.232	−0.060	−0.234	0.133	−0.321	−0.352	0.007	0.300	0.259
LONELY	0.638	−0.229	−0.209	0.355	−0.223	0.268	0.219	0.182	0.408	−0.043
UNFRIEND	0.350	0.674	−0.335	−0.011	0.177	0.410	−0.316	0.074	−0.045	0.065
ENJOYLIF	0.524	−0.271	−0.270	−0.449	0.484	0.073	0.303	0.114	−0.103	−0.145
FELTSAD	0.715	−0.225	−0.201	0.267	−0.137	−0.006	0.122	−0.091	−0.425	0.321
DISLIKED	0.339	0.721	−0.224	0.077	−0.020	−0.441	0.317	−0.065	0.089	−0.028
GETGOING	0.581	0.200	0.471	−0.224	−0.285	−0.040	0.000	0.497	−0.144	−0.018

so that a score of 1 would indicated depression for all questions. Tables 1 and 2 contain the results of a principal components analysis of the *standardized* depression variables performed on $N = 1660$ observations. Table 1 contains the correlation matrix **R** (upper triangle matrix). The first section of Table 2

contains the eigenvectors $\mathbf{e}_j$ as columns of the eigenvector matrix. These are the coefficients of the principal components Y_j of (1). The eigenvalues or variances λ_j are given below this matrix.

The first principal component is

$$\begin{aligned} Y_1 = {} & 0.327 * \text{EFFORT} + 0.213 * \text{RESTLESS} \\ & + 0.419 * \text{DEPRESS} + 0.381 * \text{HAPPY} \\ & + 0.348 * \text{LONELY} + 0.191 * \text{UNFRIEND} \\ & + 0.286 * \text{ENJOYLIF} + 0.390 * \text{FELTSAD} \\ & + 0.185 * \text{DISLIKED} + 0.317 * \text{GETGOING}. \end{aligned} \tag{11}$$

Its variance is $\lambda_1 = 3.358$. We obtain the other principal components in an obvious fashion.

Initial Component Matrix

The matrix of eigenvectors contains the weights of the principal components. From these weights we can obtain a second matrix, labeled **A**, whose elements are defined as

$$a_{ij} = e_{ij}\sqrt{\lambda}_j, \tag{12}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. The elements a_{ij}, are the correlations of X_i with Y_j. The matrix **A** with these elements is called the *initial component matrix*. The elements are also often called *loadings*, and so **A** is also called the *initial loading matrix*.

Example

The second section of Table 2 contains the **A** matrix for the Framingham depression data. Notice how the first component contains large loadings, and the loadings of the other components are generally of much smaller magnitude. This reflects that the earlier components are more heavily correlated with the original Xs than are the later components.

Amount of Variance Explained

As we stated earlier, the sum of the variances (or eigenvalues) λ_j of the principal components equals the sum of the variance of the original variables, called the *total variance*. With standardized variables the total variance is the trace of the correlation matrix **R**, or n, which is the number of variables. Thus the sum of the variances of the first m principal components divided by n,

$$\frac{\lambda_1 + \cdots + \lambda_m}{n}, \tag{13}$$

is the cumulative proportion of the total variance "explained" by the first m components. Table 3 contains the λ_j values for $j = 1, \ldots, 10$, the differences in these, the proportions of the total variance explained by the individual components, and the cumulative proportion of total variance explained.

Number of Components Retained

In the above we started with n variables and ended with n principal components. We still have as many new components as original variables, but the components are uncorrelated. This is the solution to Aim 1 given at the beginning of this article. Unless the matrix **R** is singular, this will always be the case. Often we desire to retain only a smaller set, say $m < n$, of the principal components, possibly for use in later analyses. The question is to decide upon m. There are a number of approaches to this. First, if we desire to retain the "best" linear function or composite variable of the original variables that explains as much variance as possible, then the first principal component is the appropriate one to retain. This is the solution to Aim 2 given at the beginning of this article. For the Framingham depression data this is given as formula (11) above. See [4, Chapter 12] for more details on this best single composite variable.

Table 3 Eigenvalues of the correlation matrix

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	PRIN8	PRIN9	PRIN10
Eigenvalue	3.358	1.290	1.022	0.872	0.795	0.628	0.590	0.552	0.509	0.386
Difference	2.067	0.268	0.150	0.077	0.168	0.038	0.038	0.043	0.123	–
Proportion	0.336	0.129	0.102	0.087	0.080	0.063	0.059	0.055	0.051	0.039
Cumulative	0.336	0.465	0.567	0.654	0.734	0.797	0.855	0.911	0.961	1.000

Usually we desire not just one component, but rather the m components needed to explain the "salient" variance of the original n variables. Here salient means reproducing the important or nonrandom variance of the original data. Aims 3 and 4 above relate to this, where the number of components needed to explain the salient variance is also considered the meaningful dimension of the data.

There are a number of ways to determine m, most being data-analytic rather than formal procedures of statistical **inference**. Some of the popular data-analytic rules when dealing with the principal components obtained from the correlation matrix **R** are as follows (details on these are given in [4], [7–10]):

1. Retain all components with $\lambda_j \geq 1.0$ [13].
2. Retain all components with $\lambda_j \geq 0.7$ [8].
3. Cattell's **scree test** [3]. Often the differences in the eigenvalues decrease regularly up to a point, followed by a substantially larger difference, and then followed by even smaller differences (usually 0.10 or smaller). The scree test selects the number of components to retain as the number that corresponds to the component immediately preceding the substantially large difference. This test is equivalently performed by plotting the eigenvalues and retaining the number of components that come before a break in the plot (*see* **Scree Test** and [4, Chapter 5] for more details).
4. 80% rule. Retain all components needed to explain at least 80% of the total variance.
5. Broken stick rule [10]. Retain all components that correspond to eigenvalues whose proportion of the total variance explained is above what would be expected if all components were random.

The existing inferential rules for determining m are based on the assumption that the original X variables are normally distributed. Popular rules derive from Bartlett [2]. Others are due to Girshick [6] and Anderson [1].

Example

Applying the Kaiser $\lambda_j \geq 1.0$ rule to the Framingham depression data of Tables 2 and 3, we retain three components which explain 56.7% of the total variance. The other rules would have retained more components. The Jolliffe $\lambda_j \geq 0.7$ rule and the scree test would retain five components. Figure 1 displays graphically the scree test. The 80% total variance rule would retain six components.

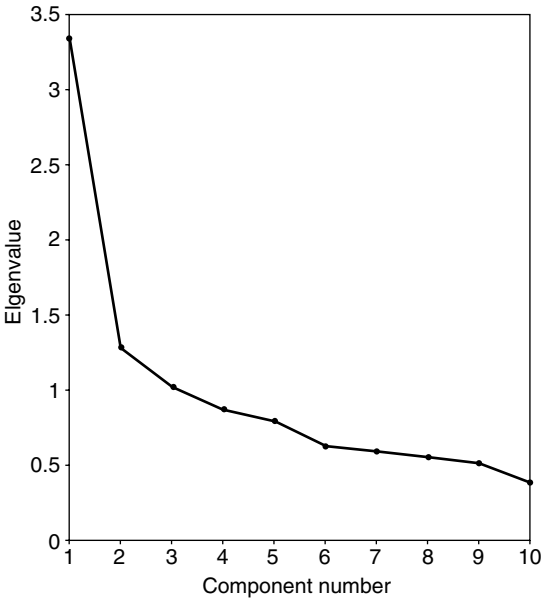

Figure 1 Scree plot of eigenvalues for Framingham depression data (principal components analysis)

For further examples we use the Kaiser rule and retain only three components. Table 4 contains the loadings for the three components. We still can designate it by **A**, but more appropriately we call it $\mathbf{A}_3$, where the subscript designates the number of components retained. Slight differences from Table 2 are due to rounding by the computer programs.

Some researchers use principal components analysis as a method to perform a **factor analysis**. In this context we can designate the reduced matrix **A** (or $\mathbf{A}_m$) with only m columns as **F** and call it the *initial factor matrix*, the initial **factor loading matrix**, or the initial *factor pattern matrix* (*see* **Factor Analysis, Overview**, and [4]). The matrix **F** is standard output from factor analysis software such as SAS Procedure FACTOR [16].

Rotations

After we decide upon the number of components to retain, the next step is to interpret them. The procedure often employed is to define the components

Table 4 Component matrices for Framingham depression data

	Initial component matrix			Varimax component matrix			Promax component matrix[a]		
	FACTOR1	FACTOR2	FACTOR3	FACTOR1	FACTOR2	FACTOR3	FACTOR1	FACTOR2	FACTOR3
EFFORT	0.600	0.153	0.411	0.250	0.684	0.146	0.080	0.604	0.056
RESTLESS	0.391	0.066	0.546	0.067	0.670	−0.043	−0.079	0.644	−0.116
DEPRESS	0.768	−0.133	−0.103	0.717	0.290	0.145	0.621	0.127	0.060
HAPPY	0.698	−0.232	−0.059	0.686	0.269	0.021	0.606	0.125	−0.059
LONELY	0.637	−0.230	−0.209	0.697	0.114	0.064	0.647	−0.033	−0.002
UNFRIEND	0.351	0.674	−0.334	0.117	0.039	0.821	0.036	−0.063	0.801
ENJOYLIF	0.525	−0.271	−0.270	0.649	−0.004	0.019	0.630	−0.134	−0.030
FELTSAD	0.715	−0.225	−0.201	0.754	0.162	0.088	0.690	0.000	0.013
DISLIKED	0.339	0.721	−0.224	0.043	0.134	0.816	−0.057	0.045	0.793
GETGOING	0.580	0.200	0.471	0.189	0.734	0.158	0.009	0.663	0.068
	Variance explained by each component			Variance explained by each component			Variance explained by each component taking into account other components		
	FACTOR1	FACTOR2	FACTOR3	FACTOR1	FACTOR2	FACTOR3	FACTOR1	FACTOR2	FACTOR3
	3.358	1.290	1.022	2.579	1.670	1.421	2.061	1.277	1.299

[a]This is the reference structure matrix; see Cureton & D'Agostino [4] for details.

in terms of the original variables that have high loadings (high correlations) with the components. Usually this does not work well with the retained principal components. For, as is typically the case, the first principal component is an average of all the variables with high loadings on all the variables (see Table 4), while the second component has, on average, smaller loadings, some positive and some negative. The third component usually has still smaller loadings. In general none of the principal components lends itself to simple interpretation.

For interpretation purposes, usually we rotate or transform the retained principal components in such a way that the rotated components have high loadings on a small set of variables, and zero or near zero loadings on the remaining variables. Often the hope is that for the rotated components the high loadings will form nearly exclusive sets. That is to say, a variable that has a high loading on one rotated variable will have a small (near zero) loading on all other rotated components. This is also called achieving a **simple structure**. The rotated components maintain the salient variance of the original variables explained by principal components, but spread it out "usually more evenly" across the rotated components. That is, the variances of the rotated components usually have variances more equal to each other than did the principal components (*see* **Rotation of Axes** and [4, Chapters 6, 8, and 9] for more details).

There are a number of procedures for performing the rotation (outlined in the article **Rotation of Axes**). Some are **orthogonal rotations**, such as the **varimax rotation** [11, 12], where the rotated components are uncorrelated, as are the principal components. Other rotations methods are **oblique rotations** where, to obtain better interpretation or, equivalently, a simpler structure, the rotated components are allowed to be correlated. Some major oblique rotations methods are **Oblimin rotation**, **Optres rotation**, **Orthoblique**, or Harris–Kaiser rotation, and the **Promax rotation**. The latter starts with the Varimax rotation results and employs a **Procrustes rotation** to obtain a simple structure. (See the above-mentioned articles and [4] for a detailed explanation of the details.)

Example

Table 4 also contains the results of two rotations of the three retained principal components of the depression data. It contains results of the orthogonal Varimax rotation and the oblique Promax rotation. The rotated components are easy to interpret. The first is defined by DEPRESS, HAPPY, LONELY, ENJOYLIF, and FELTSAD. The second is defined

by EFFORT, RESTLESS, and GETGOING. The last component is defined by UNFRIEND and DISLIKED.

Component Scores

Eq. (11) gives the first principal component function for the Framingham depression data. If we "plug" into that formula a subject's standardized values for the original data, we refer to the output as his *first principal component score*. We obtain, in a similar fashion, the other principal component scores from (1) and (2). The coefficients e_{ij} of the principal component scores for the retained components are in the first m columns of the eigenvector matrix, **e** (Table 2 for the depression data).

These principal components, Y_j have variances equal to λ_j for $j = 1, \ldots, m$. Many researchers prefer to have principal component scores with means zero and variances equal to unity (that is, they prefer to standardize them). The basic relationship of the standardized Xs to the standardized Ys (we call them Zs) is given by

$$\mathbf{X} = \mathbf{A}_m\mathbf{Z} + \mathbf{E}, \tag{14}$$

where **X** is an n-dimensional vector of standardized scores of the original data for a subject, $\mathbf{A}_m$ is the $n \times m$ initial component matrix of the retained components, **Z** is the m-dimensional vector of retained *standardized* principal components scores of the subject, and **E** is a "deviation" vector to acknowledge that the principal components do not reproduce exactly the initial data if we retain only m components ($m < n$).

In this case we obtain the **Z** as

$$\mathbf{Z} = \mathbf{D}_m^{-1}\mathbf{A'}_m\mathbf{X}. \tag{15}$$

Here D_m^{-1} is the $m \times m$ diagonal matrix with diagonal elements equal to $1/\lambda_j$, for $j = 1, \ldots, m$, and $\mathbf{A'}_m$ is the transpose of $\mathbf{A}_m$. Note this is an exact solution and not an estimate, even though there is a deviation vector **E** in (14). The first m principal component scores are the same whether we retain all n possible principal components or just the first m. This is not the case with **factor scores**.

For an orthogonal rotation the transformed component scores are simply

$$\mathbf{Z}_t = \mathbf{\Lambda'}_m(\mathbf{D}_m^{-1}\mathbf{A'}_m\mathbf{X}), \tag{16}$$

where $\mathbf{\Lambda}_m$ is the $m \times m$ transformation matrix that rotates the original principal components to the transformed component scores $\mathbf{Z}_t$. Here $\mathbf{Z}_t$ is the m-dimensional vector of transformed component scores. We often call these *component scores* and drop the term "transformed". The formula for the component scores for oblique rotated components is more complicated [4].

Example

Table 5 gives the weights for the component scores for the Varimax and Promax oblique rotation of the Framingham depression data. We apply these to the standardized variables (see [4, Chapter 12] for a discussion of the weights for the raw, unstandardized data).

Table 5 Component scoring coefficients for retained components Framingham depression data

	Rotation method: Varimax standardized scoring coefficients			Rotation method: Promax standardized scoring coefficients		
	FACTOR1	FACTOR2	FACTOR3	FACTOR1	FACTOR2	FACTOR3
EFFORT	−0.070	0.450	−0.002	0.025	0.419	0.044
RESTLESS	−0.142	0.514	−0.130	−0.045	0.449	−0.083
DEPRESS	0.270	0.016	0.015	0.267	0.079	0.047
HAPPY	0.270	0.025	−0.075	0.260	0.077	−0.040
LONELY	0.313	−0.106	−0.022	0.281	−0.033	0.001
UNFRIEND	−0.017	−0.113	0.614	0.020	−0.042	0.591
ENJOYLIF	0.324	−0.179	−0.036	0.275	−0.103	−0.020
FELTSAD	0.327	−0.087	−0.014	0.299	−0.011	0.013
DISLIKED	−0.079	−0.019	0.604	−0.022	0.034	0.584
GETGOING	−0.114	0.503	0.005	−0.006	0.461	0.052

Clusters and Cluster Scores

Yet another objective of principal components analysis (Aim 5 above) is to group the original variables together into clusters and generate simple summary variables or cluster scores. This can sometimes be accomplished by an "intuitive cluster analysis" achieved by examining either the rotated component matrices (Varimax or Promax component matrices of Table 4) or the weights of the rotated component scores (Table 5) and set the "small" loadings or weights to zero and the others to unity. The variables that have the new weights equal to unity for the same component comprise a cluster. This intuitive cluster analysis works well for the Framingham depression data. From either Table 4 or Table 5 three clusters emerge:

Cluster 1: DEPRESS, HAPPY, LONELY, ENJOYLIF, FELTSAD
Cluster 2: EFFORT, RESTLESS, GETGOING
Cluster 3: UNFRIEND, DISLIKED

Note that the clusters comprise mutually exclusive and exhaustive groups. We usually consider this ideal.

After the clusters are obtained, we then desire to produce scores or summary scores for them. Extending the intuitive clustering above, a simple solution is to define the cluster scores to be the sum of the standardized variables that comprise the clusters. For the Framingham depression data the three cluster scores are:
Cluster score 1:

$$\text{DEPRESS} + \text{HAPPY} + \text{LONELY} + \text{ENJOYLIF} + \text{FELTSAD}. \quad (17)$$

Cluster score 2:

$$\text{EFFORT} + \text{RESTLESS} + \text{GETGOING}. \quad (18)$$

Cluster score 3:

$$\text{UNFRIEND} + \text{DISLIKED}. \quad (19)$$

This intuitive clustering does not always work. Also, using the principal components solution as a starting point can cause trouble, for the clusters derived from principal components solutions may not be as well determined as those from the depression data (*see* **Cluster Analysis of Subjects, Nonhierarchical Methods**; **Cluster Score** for more details and discussions of other methods).

Cluster scores usually are more stable than components scores (or factor scores from a factor analysis) across various subgroups. We recommend them as summary measures.

Battery Reduction

Both component analysis and cluster analysis retain all the original variables and combine them either into component scores or cluster scores. At times we desire to reduce the number of variables. For example, if we have a questionnaire with, say, $n = 100$ questions and are concerned about the burden a long questionnaire places on a subject, we may want to find a subset of, say $m = 20$ or 25 questions that reproduces the salient variance of the original n variables. This is battery reduction (Aim 6 above). Again the rotated component solution can be a starting point for such an analysis. (*See* **Battery Reduction** for one method that involves **Gram–Schmidt** transformations of the initial component matrix for which a SAS macro is available [5] from `ralph@math.bu.edu`. Also see Cureton & D'Agostino [4, Chapter 12] for further elaboration and other methods.)

Scale Development

Given a set of variables a scale is simply a composite function of them. The terms *scale* and *scaling* are from the social sciences. There one objective of scaling variables is to use the resulting scales to summarize the status of the subjects from which the original data came. However, more frequently the objective is to use the original sample to generate the scales and then use the variables and their scales to investigate other populations and samples. For example, from the analysis of the Framingham depression data given in this article there are four possible scales that can be used in other populations. They are the sum of all 10 items and the three subscales (from the cluster scores) given by (17), (18), and (19). The sum is a summary of overall depression status and the three subscales ideally each measure a separate component or dimension of depression. Other typical examples from biostatistical research are scales to quantify symptoms, quality of life (*see* **Quality of Life and Health Status**), and comorbid status.

Principal components analysis offers a number of means for generating scales. In the context of scale production, the first principal component is the best linear scale from the original variables. Component scores and cluster scores are scales of the variables. Also, after a battery reduction exercise the retained smaller set of variables can be combined to a scale.

In the context of scale development there are usually data analysis steps and concerns beyond those of a "typical" principal components analysis. First, principal components analysis concerns itself with explaining the variance of the data. It may be more important in scale development to be concerned with the correlation of the variables within a scale. Factor analysis may be better suited for this. Secondly, there is the concern about **reliability** (or reproducibility) of the items or variables in the scales and the final scales themselves. And finally, the **validity** of the scales is always a major issue. This relates to determining if they measure what the developers "claim" they measure. Addressing the questions of reliability and validity are major issues in **psychometrics**.

Reduced Rank Regression

It is not uncommon to have a regression analysis where the number of potential **explanatory** (independent or regressor) variables is extremely large. Principal components analysis allows one to reduce the number of regressors to a smaller set of uncorrelated composite variables, which can be the regressor variables. The composite variables will tend to be more reliable and more valid than the individual variables. If performed carefully principal components regression (also called reduced rank regression) can offer an excellent way of examining the full set of potential regressors. (*See* **Reduced Rank Regression** for more details and Jackson [7] for an extensive discussion. Further discussion is also in [4].)

Conclusion

Principal components is a major practical statistical analysis technique. We have reviewed a number of traditional uses. It continues to be a major method in the understanding and development of important multivariate techniques such as **classification** techniques, **canonical correlation analysis**, factor analysis, **principal coordinate analysis**, **correspondence analysis**, and the identification of **multivariate outliers**. There are a number of excellent treatments going far beyond our present discussion. It is easy to implement with computer software that is readily available for its many uses [5, 16]. Almost any statistical software package will have a principal components analysis procedure. The reader will be well rewarded in examining the related articles in this Encyclopedia and a number of available books [4, 7, 10, 14, 15].

References

[1] Anderson, T.W. (1951). Classification by multivariate methods, *Psychometrika* **16**, 31–50.

[2] Bartlett, M.S. (1954). A note on the multiplying factors for various χ^2 approximations, *Journal of the Royal Statistical Society, Series B* **16**, 296–298.

[3] Cattell, R.B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research* **1**, 245–276.

[4] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[5] D'Agostino, R.B., Dukes, K.A., Massaro, J.M. & Zhang, Z. (1992). Data/variable reduction by principal components, battery reduction and variable clustering, in *Proceedings of the Fifth Annual Northeast SAS Users Group Conference*, Northeast SAS Users Group, pp. 464–474.

[6] Girshick, M.A. (1936). Principal components, *Journal of the American Statistical Association* **31**, 519–528.

[7] Jackson, J.E. (1991). *A User's Guide to Principal Components*. Wiley, New York.

[8] Jolliffe, I.T. (1972). Discarding variables in principal component analysis. I: Artificial data, *Applied Statistics* **21**, 160–173.

[9] Jolliffe, I.T. (1973). Discarding variables in principal components analysis. II: Real data, *Applied Statistics* **22**, 21–31.

[10] Jolliffe, I.T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

[11] Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

[12] Kaiser, H.F. (1959). Computer program for varimax rotation in factor analysis, *Education and Psychological Measurement* **19**, 413–420.

[13] Kaiser, H.F. (1960). The application of electronic computers to factor analysis, *Education and Psychological Measurement* **20**, 141–151.

[14] Krzanowski, W.J. (1988). *Principles of Multivariate Analysis: A User's Perspective*. Clarendon Press, Oxford.

[15] Morrison, D.F. (1976). *Multivariate Statistical Methods*, 2nd Ed. McGraw-Hill, New York.

[16] SAS Institute, Inc. (1990). *SAS/STAT User's Guide, Release 6.04* 4th Ed. SAS Inc., Cary.

(*See also* **Multivariate Analysis, Overview**)

RALPH B. D'AGOSTINO, SR

Principal Coordinates Analysis

In many applications, researchers collect information on n objects which can be accumulated into an $n \times n$ symmetric matrix $\mathbf{D}$ whose (i, j)th entry, d_{ij}, gives a measure of the relationship between the ith and jth objects. These entries may be measured directly, as with *confusion matrices*, which count the number of times out of t trials that object i is deemed to be the same as object j, or they may be derived from more fundamental multivariate observations on several variables, e.g. by calculating a *(dis)similarity matrix* or a *distance matrix* (*see* **Similarity, Dissimilarity, and Distance Measure**). When observed values depend on the order in which i and j are presented, the resulting square matrix is usually asymmetric. Then we may separate out the symmetric and skew-symmetric parts of the matrix. Here, we are concerned with the symmetric part, but there may be interesting structure in the skew part which also deserves separate analysis [1, 6]. The analysis of the symmetric part is the concern of **multidimensional scaling** of which principal coordinates analysis is computationally the most simple special case. The basic idea of all multidimensional scaling methods is that the symmetric matrix may be regarded as containing distance-like information and one seeks a set of n points in ρ dimensions whose $\binom{n}{2}$ interdistances approximate the matrix; we shall say that the n points *generate* approximations to the distances. If the diagonal of $\mathbf{D}$ is not already zero, then it will be ignored. Some preliminary transformation may be required to ensure a symmetric matrix in distance-like form. Thus, similarities might be converted into dissimilarities by subtracting from unity or probabilities p_{ij} might be transformed into $-\log p_{ij}$. When ρ is small, usually $\rho = 2$, the points may be plotted to give a graphic representation, or visualization, of the distances in the matrix (*see* **Graphical Displays**). A frequently used illustration is the recovery of the geographical map of a set of towns, given the road distances between them (see, for example, [5]); practical applications include "maps" of patients with differing symptoms, species, genotypes of agricultural crops, sets of psychological stimuli, kinships or other social structures, and many others. Inspection of these maps may indicate patterns suggesting interesting relationships or the maps may just be used as a convenient way of reporting complex multivariate relationships.

The relevant algebra is in two parts. The first part concerns the possibility of finding a set of points in *any* number of dimensions that generate the given distances exactly. The second part concerns *approximating* the distances in a few dimensions.

Exact Euclidean Representations

A Euclidean representation is desirable for graphic interpretation because, overwhelmingly, that is what most people are familiar with. One way of finding whether or not a Euclidean set of generating points exists is to proceed as follows. Put the first point at the origin, the second point a distance d_{12} from the origin. Put the third point in a plane containing the first two points and a distance d_{13} from the origin and a distance d_{23} from the second point. One may proceed in this way, at each stage adding a new point and, in general, requiring a new dimension. If the rows of $\mathbf{Y}$ give a set of coordinates for n points corresponding to the first n samples, then it may be shown [5] that the coordinates $\mathbf{y}$ of the $(n+1)$th point relative to the coordinate axes of $\mathbf{Y}$ are given by

$$\mathbf{y} = -\tfrac{1}{2}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'[\mathbf{d}_{n+1} - \mathbf{d} - (\mathbf{e}_1'\mathbf{d}_{n+1})\mathbf{1}] \qquad (1)$$

plus an extra coordinate y_{n+1}, in a new dimension orthogonal to those of $\mathbf{Y}$, given by

$$y_{n+1}^2 = \mathbf{e}_1'\mathbf{d}_{n+1} - \mathbf{y}'\mathbf{y}. \qquad (2)$$

In (1) and (2) $\mathbf{e}_1$ is a unit vector zero everywhere except for its first value of unity, $\mathbf{d}_{n+1}$ is a column vector containing all the squared distances of a new $(n+1)$th sample from the existing n samples, and $\mathbf{d}$ is the diagonal of $\mathbf{YY}'$ written as a column vector, giving the squared distances of the first n points from their origin. Thus, $\mathbf{e}_1'\mathbf{d}_{n+1}$ gives the squared distance of the $(n+1)$th point from the first point placed at the origin.

To see how (1) and (2) work, consider the set of squared distances among four points exhibited in the symmetric matrix:

$$\begin{pmatrix} 0 & 4 & 10 & 6 \\ \underline{4} & 0 & 10 & 6 \\ \underline{10} & 10 & 0 & 8 \\ \underline{6} & 6 & 8 & 0 \end{pmatrix}$$

from which we derive

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ \underline{0} & 4 & 2 & 2 \\ \underline{0} & \underline{2} & 5 & 4 \\ \underline{0} & \underline{2} & \underline{4} & 6 \end{pmatrix}$$

by replacing element a_{ij} of the first matrix by $-\frac{1}{2}(a_{ij} - a_{i1} - a_{j1})$. The underlined elements of the second matrix then turn out to be successive elements of the vector $-\frac{1}{2}[\mathbf{d}_{n+1} - \mathbf{d} - (\mathbf{e}_1'\mathbf{d}_{n+1})\mathbf{1}]$ needed in (1), while the underlined elements of the first matrix are the successive elements $\mathbf{e}_1'\mathbf{d}_{n+1}$ needed in (2). Thus, placing the first point at the origin and the second at 2 on the first axes gives our initial one-dimensional coordinates for the first two points as $\mathbf{Y}_1 = \binom{0}{2}$. The coordinates of the first dimension of the third point of $\mathbf{Y}_2$ are then obtained from (1) as $(4)^{-1}(0 \quad 2)\binom{0}{2} = 1$, where the column vector $\binom{0}{2}$ comes from the underlined elements in the third row of the derived matrix. The coordinate of the second dimension of the third point comes from the underlined element in the third row of the first matrix which, when substituted into (2), gives $y_3^2 = 10 - 1^2 = 3^2$ so that

$$\mathbf{Y}_2 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \\ 1 & 3 \end{pmatrix}.$$

We may now proceed to calculate the first two dimensions of the fourth point of $\mathbf{Y}_3$ from (1) to give

$$\begin{pmatrix} 5 & 3 \\ 3 & 9 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

while from (2), $y_4^2 = 6 - 1^2 - 1^2 = 2^2$ so that

$$\mathbf{Y}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 3 & 0 \\ 1 & 1 & 2 \end{pmatrix}$$

which, as is easily verified, generates the initial squared distances. Finally, if the last row of the given squared distances were 2,2,2 rather than 6,6,8, then it may be verified that the calculation of $\mathbf{Y}$ is replaced by

$$\begin{pmatrix} 5 & 3 \\ 3 & 9 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1\frac{1}{3} \end{pmatrix},$$

while from (2), $y_4^2 = 2 - 1^2 - \left(1\frac{1}{3}\right)^2 < 0$ and no real representation exists.

Thus, it is easy to generate the successive coordinates described above, ending up with the final version of $\mathbf{Y}$ in a lower-triangular matrix, the first value of which is zero. When the process terminates satisfactorily, the resulting configuration of n points generates the given distances, is Euclidean, and occupies, at most, $n - 1$ dimensions. If the distances do not have a Euclidean representation, then the process breaks down at some point, (2) requiring the square root of a negative number. This may be accepted but the resulting coordinates include imaginary numbers and the representation is not Euclidean. When the distances have a Euclidean representation, they are said to be *Euclidean embeddable*.

We are concerned with symmetric matrices of dissimilarities $\mathbf{D}$, with zero diagonal elements. Consider the *centered* form:

$$\mathbf{B} = (\mathbf{I} - \mathbf{1}\mathbf{s}')\mathbf{D}(\mathbf{I} - \mathbf{s}\mathbf{1}'), \tag{3}$$

where $\mathbf{s}'\mathbf{1} = 1$. When $\mathbf{B}$ is positive semidefinite, we may write

$$\mathbf{B} = \mathbf{Y}\mathbf{Y}', \tag{4}$$

nonuniquely. Regarding the ith row of $\mathbf{Y}$ as the coordinates of a point, the squared distance between the ith and jth points is given by $b_{ii} + b_{jj} - 2b_{ij}$. From (3) it follows after some elementary algebraic manipulations, that, provided we define $\mathbf{D} = \left\{-\frac{1}{2}d_{ij}^2\right\}$ then

$$b_{ii} + b_{jj} - 2b_{ij} = d_{ij}^2 \tag{5}$$

and hence $\mathbf{Y}$ generates the given distances. From now on we assume that $\mathbf{D}$ is defined in this way. Thus, that (3) be positive semidefinite guarantees the real decomposition (4), thus providing a sufficient condition for the distances d_{ij} to have a Euclidean representation. That the condition is also necessary for Euclideanarity was first proved by Schoenberg [10] for the case $\mathbf{s} = \mathbf{1}/n$ and by Householder & Young [9] for the case $\mathbf{s} = \mathbf{e}_1$, but the result is valid for all $\mathbf{s}$ where $\mathbf{s}'\mathbf{1} = 1$; an elementary proof is given by Gower [7]. Thus, $\mathbf{D} = \left\{-\frac{1}{2}d_{ij}^2\right\}$ is embeddable in a Euclidean space if and only if (3) is positive semidefinite.

From (3) we have that $\mathbf{s}'\mathbf{B}\mathbf{s} = 0$, so the vector $\mathbf{s}'\mathbf{Y} = 0$, showing that $\mathbf{s}$ determines the location of the origin of the generating points $\mathbf{Y}$; the different

solutions, $\mathbf{Y}$, that satisfy (4) represent different orientations around this origin. When $\mathbf{s} = \mathbf{e}_1$, the ith point of $\mathbf{Y}$ is at the origin, as in our example of four points where the first point is at the origin and $\mathbf{s} = \mathbf{e}_1$. Other choices of $\mathbf{s}$ are discussed by Gower [7] and by De Rooij and Gower [2]. The squares of distances of the n points from the origin are the elements in the diagonal of $\mathbf{B}$, which may be written as a column vector:

$$\mathbf{d} = (\mathbf{s}'\mathbf{D}\mathbf{s})\mathbf{1} - 2\mathbf{D}\mathbf{s} \tag{6}$$

and the sum of squares of the distances from the origin is $\mathbf{1}'\mathbf{d}$.

Of special importance is the choice $\mathbf{s} = \mathbf{1}/n$, which implies that $\mathbf{1}'\mathbf{Y} = 0$ and hence places the origin at the centroid of the generating points. The squares of the *centroid distances* are then given as

$$\mathbf{d} = \frac{1}{n^2}(\mathbf{1}'\mathbf{D}\mathbf{1})\mathbf{1} - \frac{2}{n}\mathbf{D}\mathbf{1}, \tag{7}$$

and the sum of squares about the mean (centroid) is

$$\mathbf{1}'\mathbf{d} = -\frac{1}{n}\mathbf{1}'\mathbf{D}\mathbf{1} = \frac{1}{n}\sum_{i<j}^{n} d_{ij}^2, \tag{8}$$

a classical result relating a sum of squares about the mean to the sum of squares of all $\binom{n}{2}$ intersample squared distances.

Approximate Euclidean Representations

We have seen that exact representations are multidimensional; for practical purposes low-dimensional approximations are required. Continuing with the setting $\mathbf{s} = \mathbf{1}/n$, and writing $\mathbf{N} = \mathbf{1}\mathbf{1}'/n$, (3) becomes

$$\mathbf{B} = (\mathbf{I} - \mathbf{N})\mathbf{D}(\mathbf{I} - \mathbf{N}), \tag{9}$$

and if we choose the spectral decomposition, then $\mathbf{Y}$ is given by

$$\mathbf{B} = \mathbf{Y}\mathbf{Y}', \qquad \mathbf{Y}'\mathbf{Y} = \mathbf{\Lambda}, \tag{10}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the **eigenvalues** of $\mathbf{B}$, assumed to be presented in nondecreasing order (*see* **Matrix Algebra**). Because $\mathbf{Y}'\mathbf{Y}$ is diagonal, $\mathbf{Y}$ is referred to its principal axes and it follows that $\mathbf{Y}_\rho$, the first ρ columns of $\mathbf{Y}$, gives a ρ-dimensional **principal components analysis** of $\mathbf{Y}$. Thus, with these settings, the method of approximation is that of principal components analysis so that the approximation $\mathbf{Y}_\rho$ is obtained as an orthogonal projection of the generating points with coordinates $\mathbf{Y}$ onto a ρ-dimensional subspace. It follows that the method could proceed by first computing any $\mathbf{Y}$ that satisfies (9) or, indeed, any $\mathbf{Y}$ that satisfies (4) and applying principal components analysis to it. However, the choice of (9) and (10) allows the two steps to be subsumed into one step. Because they are referred to principal axes, the rows of settings of $\mathbf{Y}$ that satisfy (9) and (10) are known as the *principal coordinates* of $\mathbf{D}$ and the method as *principal coordinates analysis* [4]. Alternatively, especially in the psychometric literature, the method is known as *classical scaling*, a terminology which stresses the links with multidimensional scaling [12]. The method could just as easily operate with other choices of $\mathbf{s}$ to give representations referred to principal axes through different origins, such as the circumcenter of the points (see [7]), but $\mathbf{s} = \mathbf{1}/n$ minimizes the sum of squares onto any ρ-dimensional subspace.

As with many statistical methods, there is interest in interpolating a new sample into an existing analysis. Of course, when interpolating a sample, we could operate on the total of $n + 1$ samples to find a ρ-dimensional principal coordinates approximation. With m separate samples we could do m separate analyses, but this would be both inefficient and, because the positions of the original n samples would change slightly in each analysis, would make it difficult to compare the different interpolations. So, sacrificing some accuracy, we seek to interpolate directly into the existing ρ-dimensional approximation; versions of (1) and (2) remain the key formulas for this operation. When the rows of $\mathbf{Y}$ are principal coordinates, $\mathbf{Y}'\mathbf{1} = 0$ and, using (7) and (10), the interpolation formula (1) simplifies to

$$\mathbf{y} = -\frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{Y}'\left(\mathbf{d}_{n+1} + \frac{2\mathbf{D}\mathbf{1}}{n}\right), \tag{11}$$

which gives coordinates in all $n - 1$ dimensions. Because $\mathbf{\Lambda}$ is diagonal the coordinates in the first ρ dimensions require only the first ρ columns, $\mathbf{Y}_\rho$ of $\mathbf{Y}$ to give

$$\mathbf{y}_\rho = -\frac{1}{2}\mathbf{\Lambda}_\rho^{-1}\mathbf{Y}_\rho'\left(\mathbf{d}_{n+1} + \frac{2\mathbf{D}\mathbf{1}}{n}\right), \tag{12}$$

which is very simple to apply. Note that the extra dimension (2) is orthogonal to the first ρ dimensions

and may be regarded as a residual, but does not enter into interpolation of further samples into the ρ-dimensional approximation.

A typical application of (11) and (12) is to add a new sample to the display, say a new town to the geographic map, without recalculating **eigenvectors**. Another application is to interpolate a point which represents a pseudo-sample that pertains to a value ξ_k of a kth variable, and then allowing ξ_k to vary while holding all other variables fixed, typically at zero. The resulting locus is a nonlinear *biplot axis* (*see* **Graphical Displays**) for the kth variable [8], unless dissimilarity is defined by the classical Pythagoras formula, when the locus is a linear axis. Such biplot axes have many of the familiar properties of Cartesian coordinate axes; in particular, the nonlinear axis may be marked at standard values of the kth variable and then the value of the variable pertaining to any point may be read off by orthogonal projection. Categorical variables may be handled similarly, replacing the kth axis by a set of category-level points, one for each permissible category level of the kth variable. Then the level of the kth variable pertaining to any point is that given by the nearest category-level point. Note that orthogonal projection is replaced by the more fundamental concept of finding the nearest point of a set, which becomes an orthogonal projection when the set forms a continuum.

When squared distances are additive, in the sense that $\mathbf{D} = \sum_1^p \mathbf{D}_k$, where $\mathbf{D}_k$ is a matrix of squared distances calculated solely from the values of the kth of p variables, the n interpolated points for the actual values of the kth variable in the sample are given by the columns of

$$\mathbf{Z} = \mathbf{\Lambda}^{-1}\mathbf{Y}'\mathbf{B}_k, \tag{13}$$

where $\mathbf{B}_k$ is calculated from $\mathbf{D}_k$ in the same way that $\mathbf{B}$ is calculated from $\mathbf{D}$ in (9). This formula is especially valuable for giving the category-level points, as all category levels will occur in a sample.

Distributional results associated with principal coordinates analysis are few. Indeed, in many applications it is difficult to see what null distribution might be reasonably assumed for the distance matrix or more fundamental variables from which it may be derived. Often, it may not even be realistic to assume that the variables are **random variables**. When stability of principal coordinates analysis, or indeed other multidimensional scaling analyses, is of interest **bootstrap methods** may be used. Sibson [11] provided a perturbation analysis for principal coordinates analysis, showing that if $\mathbf{D}$ is perturbed by a symmetric matrix $\varepsilon\mathbf{C}$, then the statistic for the Procrustean fit (*see* **Procrustes Rotation**) between the generating coordinates of $\mathbf{D}$ and its perturbed form, when expressed as a polynomial in ε, has no constant or linear terms, depending only on ε^2 and higher order terms. This suggests that principal coordinates analysis is a robust method and Sibson reports that the approximation is good, even when the perturbations are of the same order of magnitude as the given distances. Perturbation of a Euclidean distance matrix may induce non-Euclideanarity, manifesting itself as some negative eigenvalues in (9), as well as small positive eigenvalues representing noise. The principal coordinates analysis method may still be used if the negative eigenvalues are ignored, provided these are small. Sibson suggests a useful rule of thumb that small positive eigenvalues of the same, or lower, order of magnitude as the largest negative eigenvalue may be ignored. Ignoring small negative eigenvalues may also be justified by the Eckart–Young theorem [3] giving $\mathbf{Y}_\rho\mathbf{Y}_\rho'$ as a least-squares approximation to $\mathbf{B}$.

References

[1] Constantine, A.C. & Gower, J.C. (1978). Graphical representation of asymmetry, *Applied Statistics* **27**, 297–304.

[2] De Rooij, M. & Gower, J.C. (2003). The geometry of triadic distances, *Journal of Classification* **20**, 39.

[3] Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank, *Psychometrika* **1**, 211–218.

[4] Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**, 325–338.

[5] Gower, J.C. (1966). A Q-technique for the calculation of canonical variates, *Biometrika* **53**, 588–589.

[6] Gower, J.C. (1977). The analysis of asymmetry and orthogonality, in *Recent Developments in Statistics*, J.R. Barra, F. Brodeau, G. Romer & B. van Cutsem, eds. North-Holland, Amsterdam, pp. 109–123.

[7] Gower, J.C. (1982). Euclidean distance matrices, *Mathematical Scientist* **7**, 1–14.

[8] Gower, J.C. & Hand, D.J. (1996). *Biplots*. Chapman & Hall, London.

[9] Householder, A.S. & Young, G. (1938). Discussion of a set of points in terms of their mutual distances, *Psychometrika* **3**, 19–22.

[10] Schoenberg, I.J. (1935). Remarks to Maurice Fréchet's article "Sur la definition d'une classe d'éspaces vectoriéls applicable vectoriéllement sur l'éspace Hilbert", *Annals of Mathematics* **36**, 724–732.

[11] Sibson, R.R. (1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling, *Journal of the Royal Statistical Society, Series B* **41**, 217–229.

[12] Torgerson, W.S. (1958). *Theory and Methods of Scaling*. Wiley, New York.

(*See also* **Classification, Overview**; **Cluster Analysis, Variables**; **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**; **Pattern Recognition**; **Projection Pursuit**; **R- and Q-analysis**)

JOHN C. GOWER

Prior Distribution

Bayesian inference uses probability distributions to represent knowledge. In statistical settings, inferences are based on posterior distributions, which are computed from **Bayes' Theorem**:

$$p(\theta|y) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)\,\mathrm{d}\theta}, \qquad (1)$$

where θ is the unknown parameter, y represents the data, $L(\theta)$ is the **likelihood** function, $p(\theta|y)$ is the density (pdf) of the posterior distribution, and $\pi(\theta)$ is the density of the prior distribution. The appearance of the prior distribution on the right-hand side of (1) is at once a strength and a weakness of the Bayesian approach: a strength because it allows information beyond the data at hand to be used in making inferences, and a weakness because the inferences inevitably depend, at least to some degree, on the choice of the density $\pi(\theta)$.

It is useful to distinguish three kinds of prior distributions, *informative priors*, noninformative or *reference priors*, and hierarchical priors or priors used in *hierarchical models*.

Informative Priors

When there is substantial knowledge about a phenomenon under investigation, it is natural to consider incorporating that knowledge into a statistical analysis. The most advantageous case occurs when there is much relevant data that may be summarized in the form of a prior distribution. A wonderful case study is the analysis of *The Federalist* papers (a series of early American political pamphlets) by Mosteller & Wallace [15], in which papers of known authorship were used to form prior distributions in order to determine who wrote the papers of unknown authorship. Another domain in which prior information has been incorporated with great success is in medical **imaging** (see, for example, [14]), where specific features of anticipated images are introduced to aid reconstruction.

Alternatively, available knowledge may come from opinions expressed by one or more experts. In this situation, the opinions are formulated into prior distributions through a process called *elicitation*, with the resulting priors being called *subjective*. A general review of elicitation methods, and their application to **clinical trials**, is given by Chaloner [3]; see also [2, 6, 9, 11, 12, 16], and the many additional references cited by Chaloner. The formalization of expert opinion could be especially valuable in **experimental design**; see [4] and [5]. An important attendant issue in elicitation is the manner in which various expert judgments ought to be combined [8].

Reference Priors

When there is little reliable information, or when an analyst wishes to avoid subjective elicitation of prior distributions, standard forms for priors may be invoked. An example would be the use of a **uniform** prior on the interval (0, 1) for a **binomial** proportion. Although such uniform priors are useful in many settings, allowing formal Bayesian analyses to proceed, a variety of theoretical and practical concerns have been raised about this and related techniques for selecting prior distributions. Kass & Wasserman [13] have reviewed the extensive literature on the subject, and summarized its major findings.

Ideally, one would like to have a given prior distribution selected as "standard" for each specific problem, such as the uniform prior for a binomial proportion. As reviewed by Kass & Wasserman, this was the point of view articulated by **Harold Jeffreys** in his pioneering attempts to provide formal rules for selecting priors [10]. However, despite the efforts of Jeffreys and many others, no consensus exists as to the rules to be used (except possibly in a few problems, such as the binomial). Furthermore, important complications may arise when *improper priors* are used; that is, priors that do not integrate to one. For example, the standard prior for a single normal mean is uniform on the real line. Although in this case no substantial difficulties arise, in more complicated settings a variety of problems may occur, including the possibility that the posterior also may be improper, so that Bayesian inferences are no longer well defined. An additional serious concern is that seemingly innocuous prior distributions (such as uniform) on multidimensional parameter spaces may have unintended consequences for certain functions of the parameters. The general conclusion to be drawn from the literature is that one must be careful, and cognizant of attendant potential difficulties,

when selecting prior distributions in the absence of specific information. However, the worries are greatly diminished when posteriors are computed from large samples [13].

Hierarchical Models

One of the major contributions of Bayesian methods to data analysis involves what have come to be called hierarchical models, briefly described as follows. Suppose that we have a collection of observation vectors $y_1, y_2, \ldots, y_k$ that are assumed to come from the same family of distributions $p(y|\theta)$, but for different values of the parameter $\theta_1, \theta_2, \ldots, \theta_k$. If we also assume that the θ_is are drawn from some family of distributions $p(\theta|\lambda)$, indexed by a further parameter λ, then we obtain a hierarchical model. Formally, this second family plays the role of a prior for each parameter θ_i. When the parameter λ is estimated by maximum likelihood or related techniques, the resulting analyses are usually called **empirical Bayes** methods. When, instead, a prior distribution is introduced on λ (which is often called a *hyperparameter*, to signify its role as a parameter of a distribution on parameters), then inferences are considered to be *fully Bayesian*. For many examples and further discussion, see [3] and [7].

References

[1] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, New York.

[2] Carlin, B., Chaloner, K., Church, T., Matts, J.P. & Louis, T.A. (1995). Elicitation, monitoring and analysis of an AIDS clinical trial (with discussion), in *Case Studies in Bayesian Statistics*, Vol. II, C. Gatsonis, J.S. Hodges, R.E. Kass & N.D. Singpurwalla, eds. Springer-Verlag, New York, pp. 48–89.

[3] Chaloner, K. (1996). Elicitation of prior distributions, in *Bayesian Biostatistics*, D.A. Berry, & D.K. Stangl, eds. Marcel Dekker, New York, pp. 141–156.

[4] Chaloner, K. & Verdinelli, I. (1995). Bayesian experimental design: a review, *Statistical Science* **10**, 273–304.

[5] Flournoy, N. (1993). A clinical experiment in bone marrow transplantation: estimating a percentage point of a quantal response curve, in *Case Studies in Bayesian Statistics, I*, C. Gatsonis, J.S. Hodges, R.E. Kass & N.D. Singpurwalla, eds. Springer-Verlag, New York, pp. 324–336.

[6] Freedman, L.S. & Spiegelhalter, D.J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials, *Statistician* **32**, 153–162.

[7] Gelman, A., Carlin, J., Stern, H. & Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, New York.

[8] Genest, C. & Zidek, J.V. (1986). Combining probability distributions (with discussion), *Statistical Science* **1**, 114–148.

[9] Hogarth, R.M. (1987). *Judgment and Choice*, 2nd Ed. Wiley, New York.

[10] Jeffreys, H. (1961). *Theory of Probability*, 3rd Ed. Oxford University Press, Oxford.

[11] Kadane, J.B., & Wolfson, L.J. (1996). Priors for the design and analysis of clinical trials, in *Bayesian Biostatistics*, D.A. Berry & D.K. Stangl, eds. Marcel Dekker, New York, pp. 157–184.

[12] Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S. & Peters, S.C. (1980). Interactive elicitation of opinion for a normal linear model, *Journal of the American Statistical Association* **75**, 845–854.

[13] Kass, R.E. & Wasserman, L. (1996). The selection of prior distributions by formal rules, *Journal of the American Statistical Association* **91**, 1343–1370.

[14] Johnson, V., Bowsher, J., Jaszczak, R. & Turkington, T. (1995). Analysis and reconstruction of medical images using prior information (with discussion), in *Case Studies in Bayesian Statistics*, Vol. II, C. Gatsonis, J.S. Hodges, R.E. Kass & N.D. Singpurwalla, eds. Springer-Verlag, New York, pp. 149–240.

[15] Mosteller, F. & Wallace, D.L. (1964). *Inference and Disputed Authorship: the Federalist*; reprinted as *Applied Bayesian and Classical Inference: the Case of the Federalist Papers*. Springer-Verlag, New York, 1984.

[16] Spiegelhalter, D.J. & Freedman, L.S. (1988). Bayesian approaches to clinical trials, in *Bayesian Statistics*, Vol. 3, J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds. Oxford University Press, New York, pp. 453–477.

R.E. KASS

Privacy in Genetic Studies

The concept of privacy has come through a long evolution in Western thought. Early notions of privacy focused on the ownership and use of property, with the idea that an individual could exercise dominion over property, with the right to enjoy the benefits provided by property and the right to dispose of the property, either by destroying it or by transferring full or partial ownership to another party. Concepts of privacy expanded to include the immediate space around the person, so that individuals came to enjoy a personal space that cannot be entered without the consent of the person. Early applications of the idea of privacy of the person included the development of the law of assault and battery, both in the criminal law and the law of tort, as well as the evolution of respect for the person in matters of corporal punishment. The sphere of personal privacy was expanded in the late nineteenth century to include the "right to be let alone", so the individuals could expect not to have their lives and personal activities invaded and pursued by reporters, journalists, and the public media [17]. Early in the twentieth century, in the context of the physician–patient relationship, the idea of privacy was expanded to include support for an individual's right to decide personally what course to follow in questions of medical or surgical treatment [12].

At the turn of the twentieth century the world of medical and surgical practice was profoundly altered by the introduction of anesthesia and antisepsis. These new technical benefits rapidly expanded the possibilities for treating a variety of human illnesses. Patients became aware of an expanding array of options, and they began to assert their own right to make decisions about following medical treatment regimens and about submitting to surgical procedures [9]. Patients became more informed and less tolerant when they discovered that they had been denied information about health care options, and they became quick to allege legal negligence on the part of physicians for failing to convey complete information about choices of treatment [4]. Early support for personal autonomy in treatment decisions is found in the resounding holding by Justice Benjamin Cardozo that "[e]very human being of adult years and sound mind has a right to determine what shall be done with his own body. . ." [12]. This and other early decisions in the common law laid the foundation for building the Doctrine of Informed Consent as it governs interactions between physicians and patients in the practice of medicine and surgery. This doctrine has evolved to include the legal obligation of health care professionals to disclose all information that may influence the decision of a patient, including treatment options, the risks and benefits of the options, as well as the risk of choosing to decline treatment altogether [15]. Integral to the structure and function of the physician–patient relationship is the assurance that information shared between physician and patient will be held in confidence, so that both members of the relationship may enjoy reciprocal candor that should ultimately benefit the patient.

During the fifth and sixth decades of the twentieth century, respect for privacy of the person and the right of the individual to make autonomous decisions expanded into the world of biomedical research. The revelations of barbarous experimentation with human beings, before, during, and after the Second World War, both in the concentration camps in Nazi Germany [16] and in a variety of civilian and military arenas in the US [6], generated calls for protecting individuals who participate as subjects in biomedical research. These concerns were thoroughly studied by the federal government and became the subject of legislation and regulations that established safeguards to protect the interests of human subjects. These safeguards are now found in the structure of the nationwide system of Institutional Review Boards (IRBs) that supervise biomedical research at public and private research institutions that enjoy the benefits of federal funding. IRBs answer to the government and are responsible for the information that is conveyed to potential subjects, for their competent and voluntary consent, for their understanding of the risks and benefits of participating in research, and for their understanding of their rights as research subjects [10]. Integral to the consent process in biomedical research is consideration of how newly developed research information will be handled.

Both patients and research subjects have vital interests in the fate of information that is gathered about their own health and the health of their families. The interests of patients include candor and complete information about treatment options. The interests of research subjects include an option for access to information about research results and the implications of these results for their families. Both

patients and subjects are entitled to assurances that their personal medical and/or research information will be maintained in confidence and will not be disclosed to other parties without express consent. Indeed, the importance of confidentiality of information in these relationships is guarded by severe legal sanctions for unauthorized disclosures of personal medical or research information [7].

Over the past two decades the unprecedented expansion of knowledge of human heredity has opened significant new areas of medical practice and biomedical research. Technological innovations have driven the rapid dissection of the **human genome**, and they now permit precise determination of the **genotypes** of individuals and groups, with respect to both an array of single **gene** diseases and a growing number of complex genetic susceptibilities to future health problems. While exhaustive knowledge of the genetic endowment of the human species may remain elusive for years, the present compendium of genetic information is expanding rapidly. As more tests are developed, and as more and more questions about genetic disease and disorders are asked and answered, individuals and families are seeking information about their own genetic legacies and their genetic prospects for their own health and the health of their offspring. In response to the growing quest for genetic information, the practice of **genetic counseling** is expanding, albeit at a somewhat slower rate than the demand for counseling services [14].

The unique genetic constitution of each individual renders personal genetic information perhaps the most private information that may be generated for any human being. Genetic information is also, however, unique because it may have significant implications for members of an individual's immediate family and collateral relatives as well. Over the past two decades the power of genetic information to influence or disrupt the lives of individuals and families has been thoroughly documented [3] and has been the impetus behind numerous legislative efforts to protect the privacy and autonomy of individuals who have knowledge of their own genotypes. Most states now have enacted legislation that protects individuals from discrimination in employment and health insurance on the basis of genotype [11]. While numerous genetic discrimination bills have been introduced at the federal level, none has yet been passed [5]. However, the Health Insurance Portability and Accountability Act of 1996 [8], and the ensuing regulations, do address questions of genetic privacy in the context of protecting the health insurance status of individuals and families who move from one employer and health insurance provider to another.

During the 1960s the development of simple tests and effective treatments for a few genetic diseases resulted in the inception of newborn screening programs. These public health programs were initiated in all states after a reliable test for phenylketonuria became available, and screening quickly expanded to include testing for sickle cell anemia and congenital hypothyroidism. Some states have subsequently incorporated screening tests for several other genetic diseases as well. Newborn screening is mandated by law in most states, and screening is typically carried out just after birth, usually without the express consent of the parent(s). Screening without parental consent is an invasion of privacy, both personal and physical, that is justified because of the immense benefits that accrue from identifying and treating infants early in life: infants enjoy the prospect of a lifetime of health benefits, and public monies that would otherwise be spent on caring for severely damaged children can be directed toward other public health efforts [1].

With the exception of newborn screening, generating genetic information by genetic testing continues to be a matter of personal choice, an exercise of personal autonomy. Individuals may choose to be tested, either on their own initiative or in response to the suggestion of a professional. Parents may choose to have their children tested, although testing minors for some adult-onset genetic diseases has generated much debate and considerable caution [2]. Persons who participate in genetics research may also seek information about themselves that is recorded in research results, although disclosure of research results is far more guarded because of the tentative, or unconfirmed nature of research results and information. As the power of technology continues to expand, the possibilities of generation and disclosure of genetic information without consent have constituted a nucleus of federal concern that is now focused on the rights of individuals and their interactions with medical professionals and the research community. While the legislative and regulatory processes are slow, the depth of interest and concern for personal privacy and patient autonomy illustrates and reinforces the potential benefits from sequencing and understanding the human genome [13].

References

[1] Andrews, L.B., Fullerton, J.E., Holtzman, N.A. & Motulsky, A.G., eds. (1994). *Assessing Genetic Risks: Implications for Health and Social Policy*. National Academy Press, Washington.

[2] ASHG/ACMG Report (1995). Points to consider: ethical, legal, and psychosocial implications of genetic testing in children and adolescents, *American Journal of Human Genetics* **57**, 1233–1241.

[3] Billings, P.R., Kohn, M.A., de Cuevas, M., Beckwith, J., Alper, J.S. & Natowicz, M.A. (1992). Discrimination as a consequence of genetic testing, *American Journal of Human Genetics* **50**, 476–482.

[4] *Canterbury v. Spence* (1972). 464 F.2d 772 (D.C. Cir.).

[5] Colby, J.A. (1998). An analysis of genetic discrimination legislation proposed by the 105th congress, *American Journal of Law and Medicine* **24**, 443–480.

[6] Final Report of the President's Advisory Committee on Human Radiation Experiments (1996). *The Human Radiation Experiments*. Oxford University Press, New York.

[7] Furrow, B.R., Greaney, T.L., Johnson, S.H., Jost, T.S. & Schwartz, R.L. (1997). *Health Law: Cases, Materials, and Problems*, 3rd Ed. West Publishing Co., St. Paul.

[8] Health Insurance Portability and Accountability Act of 1996. Public Law 104–191. 45 *United States Code* §201 (1996).

[9] *Mohr v. Williams* (1905). 95 Minn. 261, 104 N.W. 12.

[10] OPRR Reports (1991). *Protection of Human Subjects*. Code of Federal Regulations, Title 45, Part 46.

[11] Rothenburg, K.H. (1995). Genetic information and health insurance: state legislative approaches, *Journal of Law, Medicine, and Ethics* **23**, 312–319.

[12] *Schloendorf v. Society of New York Hospital* (1914). 211 N.Y. 125, 105 N.E. 92.

[13] Shalala, D. (2000). Protecting research subjects – what must be done, *New England Journal of Medicine* **343**, 808–810.

[14] Thompson, M.W., McInnes, R.R. & Willard, H.F. (1991). *Thompson & Thompson Genetics in Medicine*, 5th Ed. W.B. Saunders, Philadelphia.

[15] *Truman v. Thomas* (1980). 27 Cal.3d 285, 165 Cal.Rptr. 308, 611 P.2d 902.

[16] *United States v. Karl Brandt, Trials of War Criminals Before the Nuremberg Military Tribunals under Control Law No. 10*, Vols. 1 and 3, "The Medical Case" (Military Tribunal I, 1947; Washington D.C.: U.S. Government Printing Office, 1948–49).

[17] Warren, S.D. & Brandeis, L.D. (1890). The right to privacy, *Harvard Law Review* **4**, 193–220.

MARY KAY PELIAS

Probability Sampling

Probability sampling is a process used to select a sample from a defined population with the characteristic that every element of the population has a known, nonzero probability of being included in the sample. In probability sampling we select the sample by a random mechanism under which each element of the population receives this known selection probability [1]. Nonprobability sampling, in contrast, does not have this feature.

Probability sampling allows us to evaluate statistical properties of estimators of population characteristics and to construct estimators with desired statistical properties, since the selection probability for every sample is known (*see* **Estimation**). For instance, we can evaluate how reproducible an estimator is over repetitions of the sampling process yielding the estimate (reliability) and whether the **expectation** of an estimator over repetitions yielding the estimate is the same as the true value of the parameter being estimated (**unbiasedness**).

Nonprobability sampling is almost always less expensive and easier to carry out than probability sampling. Examples of nonprobability sampling are selection of a predetermined number of individuals having specified characteristics, or selection of individuals believed to be representative (*see* **Quota, Representative, and Other Methods of Purposive Sampling**; **Snowball Sampling**). In nonprobability sampling we cannot evaluate statistical properties of estimators such as reliability or unbiasedness since there is no known selection probability for each element in the population.

Reference

[1] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York, p. 17.

JASON HSIA

Probability Theory

Probability theory is that part of mathematics that aims to provide insight into phenomena that depend on chance or on uncertainty. The most prevalent use of the theory comes through the frequentists' interpretation of probability in terms of the outcomes of repeated experiments, but probability is also used to provide a measure of subjective beliefs, especially as judged by one's willingness to place bets.

The roots of probability theory are not as ancient as those of many parts of mathematics, and only in the sixteenth and seventeenth centuries does one find the first glimmerings of the theory in the investigations made by Gerolamo Cardano, Pierre de Fermat, and Blaise Pascal into games of chance. Despite the luminous reputations of these famous mathematicians and philosophers, the subject of probability theory remained on the periphery of respectability, and for a long time development was halting and lugubrious. Through the first third of the twentieth century, the eighteenth century works of Jakob Bernoulli (*see* **Bernoulli Family**) and **Abraham De Moivre** continued to be viewed as the nearly definitive treatises of probability theory [3, 11].

Still, even in the early days of the twentieth century when probability theory clearly suffered from the lack of a widely accepted foundation, there were profound developments, most notably Albert Einstein's use of **Brownian motion** in 1905 to provide the first determination of Avagadro's number [7]. Nevertheless, in 1933 when **Andrey Nikolayevich Kolmogorov** published his elegant succinct volume *Foundations of the Theory of Probability* [10], the mathematical world was hungry for such a treatment, and the subsequent development of probability theory was explosive.

Firm Foundation

Central to Kolmogorov's foundation for probability theory was his introduction of the triple $(\Omega, \mathcal{F}, P)$ that we now call a probability space, or sometimes the "probabilist's trinity". The triple's first element, Ω, is required only to be a set. The second element is a collection of subsets of Ω about which more will be said later. The third element is a function that assigns a real number to each of the elements of $\mathcal{F}$. This function is called a probability measure P provided that it satisfies the three following axioms:

Axiom 1. For all $A \in \mathcal{F}$ we have $P(A) \geq 0$.

Axiom 2. For any countable collection $\{A_i \in \mathcal{F} : 1 \leq i < \infty\}$ for which $A_i \cap A_j = \emptyset$ for all $i \neq j$, we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Axiom 3. $P(\Omega) = 1$.

Axioms 1 and 3 are quite bland. Axiom 1 only captures our understanding that probabilities of events are nonnegative numbers, and Axiom 3 just echoes our assumption that Ω is a sensible representation for the universe of all possible outcomes of the chance experiment being modeled. Only about Axiom 2 can there be any quarrel, and at times arguments have been made for preferring a probability theory that only requires additivity of probabilities for finite collections of sets. Kolmogorov's decision to assume countable additivity is not the only possible choice, but it has been a fecund one that has proved to be appropriate in a wide variety of circumstances.

The mathematical benefit of Kolmogorov's second axiom is that it connects probability theory with the theory of measure as put forward by Borel, Lebesgue, Radon, and Fréchet in the early part of the twentieth century. It was in fact Fréchet who noted some 13 years after Lebesgue's famous 1902 thesis that the natural domain for a probability measure is a collection of sets that is closed under complementation and countable unions. Fréchet called such collections σ-algebras, and Kolmogorov required that the second term of his triple be just such a collection [4, 5].

Basic Quantities of the Theory

To the practical mind, Kolmogorov's axiomatization of probability may seem only to defer the problem of construction of probability models that serve to inform us about the physical and social world, but by putting the elusive probability function P on an axiomatic footing Kolmogorov did provide real assurance that one could study probability as sensibly as one could study measure theory, analysis, or algebra.

In particular, one could proceed with the investigation of the objects that had been of concern from probability's earliest days.

One of the most fundamental notions of probability theory is the **random variable**, and in Kolmogorov's framework a random variable is nothing more than a function from $X : \Omega \to \mathbb{R}$ with the property that for all t one has that the sets $\{\omega : X(\omega) \leq t\}$ are elements of the σ-algebra $\mathcal{F}$. With this definition we are on firm footing when we take the definition of the distribution function F of X to be

$$F(t) = P(X \leq t),$$

because the set $\{\omega : X(\omega) \leq t\}$ is in the domain of the set function P. In this framework the **expectation** $\mathrm{E}(X)$ of the random variable X can defined as the Lebesgue integral of X with regard to P, or as the Riemann–Stieltjes integral with respect to F, giving us

$$\mathrm{E}(X) = \int_{\Omega} X(\omega)\, \mathrm{d}P(\omega) = \int_{-\infty}^{\infty} x\, \mathrm{d}F(x).$$

The probability distribution function and the expectation operation provide us with the core language that is needed to express almost everything that one needs to say about individual random variables. For example, a basic measure of dispersion of a random variable is the **variance**, which one writes in terms of the expectation as

$$\mathrm{var}(X) = \mathrm{E}(X - \mu)^2,$$

where $\mu = \mathrm{E}(X)$ and the **standard deviation** of X is defined to be the square root of the variance.

Central Role of Independence

With expectations and distributions we recapture much of the most basic language of probability theory, but the real power of probability theory only emerges with the introduction of the central notion of *independence* of events, algebras, and random variables. To begin that development, one first defines elements A and B of $\mathcal{F}$ to be independent provided

$$P(A \cap B) = P(A)P(B).$$

This definition is then extended to sub-σ-algebras of $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{F}$ by calling $\mathcal{A}$ and $\mathcal{B}$ independent provided A and B are independent for all $A \in \mathcal{A}$ and all $B \in \mathcal{B}$. Finally, random variables X and Y are independent if $\mathcal{A}$ and $\mathcal{B}$ are independent when these are respectively the smallest σ-algebras containing all the sets $\{X \leq t\}$ and all the sets $\{Y \leq t\}$.

This definition of independence of random variables may look a little burdensome at first, but for many purposes it is much more convenient than the definition of independence that is sometimes given in elementary texts that call for the factorization of the joint density of X and Y. In fact, densities may not exist, but that is not the telling point. More to the heart of the matter is that with Kolmogorov's definition one clearly sees that the independence of X and Y implies the independence of $f(X)$ and $g(Y)$ for any monotone functions f and g, while this intuitive fact is cumbersome to check if one needs to verify a density factorization.

Theorems That Make the Theory

There are two theorems that live at the very heart of probability theory. The first is the **law of large numbers**, without which our most fundamental intuitions about the relationship of probability theory and the physical world would be at odds. The second is the **central limit theorem**, which is arguably the result that most clearly accounts for the practical utility of probability as a helpmate to statistics, as well as to the social and physical sciences [1, 2, 5, 6, 8, 9].

Theorem 1 (Law of Large Numbers). If $\{X_i : 1 \leq i < \infty\}$ is a sequence of independent random variables with the distribution function, F, and if $\mathrm{E}|X_i| < \infty$, then the event that the sequence

$$\frac{1}{n}\{X_1 + X_2 + \cdots + X_n\}$$

converges to $\mathrm{E}(X_1)$ has probability one.

Theorem 2 (Central Limit Theorem). If $\{X_i : 1 \leq i < \infty\}$ is a sequence of independent random variables with distribution function F, $\mathrm{E}(X_i) = \mu < \infty$, and $\mathrm{var}(X) = \sigma^2 < \infty$, then

$$\lim_{n\to\infty} P\left(\frac{1}{\sigma\sqrt{n}}\{X_1 + X_2 + \cdots + X_n - n\mu\} \leq x\right) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \mathrm{e}^{-u^{2/2}}\, \mathrm{d}u.$$

Beyond Independent Random Variables

While the purest view of the aims and accomplishments of probability theory may be found in the study of sums of independent random variables, the applications of probability theory require the development of structures that also capture aspects of dependence. To give the simplest illustration of a such a system, we consider a finite set $S = \{1, 2, \ldots, n\}$ which we will call the set of "states", and a matrix $\mathbf{P} = (p_{ij})$, where all of the matrix entries satisfy $0 \leq p_{ij} \leq 1$ and where the row sums $p_{i1} + p_{i2} + \cdots + p_{in}$ all equal one. We now consider a sequence of random variables X_n that are defined by sequential transitions according to the row of the matrix $\mathbf{P}$. Specifically, if $X_n = i$, then X_{n+1} is determined by making a choice from the set S in accordance with the probability masses (p_{ij}). Such a sequence of random variables is called a **Markov chain**, and the theory of such sequences offers an important first step from the core theory of independent random variables. The index of the sequence $\{X_n : n \geq 0\}$ is usually viewed as "time" and an important extension of the notion of a Markov chain is that of a **Markov Process** where the index is taken to be the whole positive real line and the state space is permitted to be $\mathbb{R}^d$ (or even a more complex space). The most important such process is *Brownian motion* [2, 9].

Another direction for the development of probability theory that goes beyond independence is provided by the theory of *martingales* [2, 4–6]. On one level, martingales capture the notion of a fair gambling game, and although this view is interesting (and loyal to the origins of probability theory), the theory of martingales turns out to be an appropriate tool for many kinds of investigation (*see* **Counting Process Methods in Survival Analysis**). In particular, the theory of martingales provides the key to profound connections between the theory of Markov processes and the classical theory of harmonic functions.

References

[1] Adams, W.J. (1974). *The Life and Times of the Central Limit Theorem*. Kaedmon Press, New York.

[2] Chung, K.L. (1974). *Elementary Probability with Stochastic Processes*. Springer-Verlag, New York.

[3] David, F.N. (1962). *Games, Gods, and Gambling: The Origins and History of Probability from the Earliest Times to the Newtonian Era*. Griffin, London.

[4] Doob, Joseph L. (1994). The development of rigor in mathematical probability (1900–1950). in *Development of Mathematics 1900–1950*, J.-P. Pier, ed. Birkhäuser-Verlag, Basel.

[5] Dudley, R.M. (1989). *Real Analysis and Probability*. Wadsworth-Brooks/Cole, Pacific Grove.

[6] Durrett, R. (1991). *Probability: Theory and Examples*. Wadsworth-Brooks/Cole, Pacific Grove.

[7] Einstein, A. (1905). On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat, *Annalen der Physik, Series 4* **17**, 549–560 (in German).

[8] Feller, W. (1968). *An Introduction to Probability and Its Applications*, Vol. I, 3rd Ed. Wiley, New York.

[9] Feller, W. (1971). *An Introduction to Probability and Its Applications*, Vol. II, 2nd Ed. Wiley, New York.

[10] Kolmogorov, A.N. (1933). *Grundbegriffe der Wahrscheinlichtkeitsrechnung*. Springer-Verlag, Berlin (English translation: N. Morrison (1956). *Foundations of the Theory of Probability*. Chelsea, New York.).

[11] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, Mass.

(*See also* **Axioms of Probability**; **Convergence in Distribution and in Probability**; **Foundations of Probability**; **Limit Theorems**; **Statistical Dependence and Independence**; **Stochastic Processes**)

J.M. STEELE

Procrustes Rotation

Most rotation procedures associated with **principal components analysis** and **factor analysis** are designed to rotate a set of vectors associated with the retained components or vectors into a new set of vectors whose associated transformed components or factors will attain a **simple structure** (*see* **Rotation of Axes**). In *Procrustes* rotation the procedure is reversed; one is furnished with the two sets of vectors and determines the matrix that will best transform one set into the other. Although generally associated with principal components and factor analysis, Procrustes rotation may be used to relate any two matrices of the same dimension.

The problem may be stated as follows. Given two $(n \times m)$ matrices, $\mathbf{A}$ and $\mathbf{B}$, what transformation matrix $\mathbf{T}$ of dimension $(m \times m)$ will best transform $\mathbf{A}$ into $\mathbf{B}$ (i.e. what matrix $\mathbf{T}$ will make $\mathbf{AT}$ most like $\mathbf{B}$)? Let $\mathbf{E} = \mathbf{AT} - \mathbf{B}$. Then $\mathbf{E}$ is the difference between $\mathbf{B}$ and the approximation for it, $\mathbf{AT}$. The object is to obtain $\mathbf{T}$ such that $\text{tr}(\mathbf{E}'\mathbf{E})$ is a minimum.

If $\mathbf{T}$ is orthonormal ($\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$) (*see* **Orthogonality**), this is called the *orthogonal Procrustes problem* and has a least squares solution [6], although the notion goes back to Mosier [4]. If $\mathbf{S} = \mathbf{A}'\mathbf{B}$, $\mathbf{U}$ are the orthonormal characteristic vectors of $\mathbf{S}'\mathbf{S}$, and $\mathbf{U}^*$ the characteristic vectors (*see* **Eigenvector**) of $\mathbf{SS}'$, then $\mathbf{T} = \mathbf{U}^*\mathbf{U}'$.

Table 1 contains the characteristic vectors for the audiometric example given in the article **Rotation of Axes**. Here, the vectors have been normalized to unit length and become the matrix $\mathbf{A}$. To facilitate the screening of large sets of medical records, these vectors were replaced by some simple approximations [2]. These are also normalized to unit length and included in Table 1 as matrix $\mathbf{B}$. The matrix $\mathbf{T}$ which best rotates $\mathbf{A}$ into $\mathbf{B}$ is:

$$\mathbf{T} = \begin{bmatrix} 0.996 & -0.062 & -0.041 & -0.057 \\ 0.058 & 0.993 & -0.099 & 0.009 \\ 0.048 & 0.096 & 0.994 & 0.021 \\ 0.056 & -0.015 & -0.022 & 0.998 \end{bmatrix}$$

$\text{tr}(\mathbf{E}'\mathbf{E}) = 0.150$.

The two-sided orthogonal Procrustes rotation [7] is used to test whether a $(p \times p)$ matrix $\mathbf{A}$ is a permutation of another $(p \times p)$ matrix $\mathbf{B}$. This is particularly useful in looking for particular patterns in matrices which may have been obscured by a permutation of the rows and/or columns. There are also solutions for *oblique Procrustes* rotation [3, 5, 8, 9]. For the audiometric example, an oblique solution reduced $\text{tr}(\mathbf{E}'\mathbf{E})$ to 0.133. There have been a number of other modifications and extensions of the Procrustes solution, including comparing three or more matrices (see [2]).

The name *Procrustes* was given to this procedure by Cattell (Hurley & Cattell [1]). In Greek mythology, Procrustes, the Stretcher, was an innkeeper who lured travelers with his "magical" bed which would fit anyone. If the guest was too short, he or she was stretched out to the length of the bed. If too long, he chopped off his or her feet. Hurley & Cattell felt this rotation procedure had about the same philosophy and urged caution in using this technique.

References

[1] Hurley, J.R. & Cattell, R.B. (1962). The PROCRUSTES program: producing direct rotation to test a hypothesized factor structure, *Behavioral Science* **7**, 258–262.

[2] Jackson, J.E. (1991). *A User's Guide to Principal Components*. Wiley, New York.

[3] Meredith, W. (1977). On weighted Procrustes and hyperplane fitting in factor analytic rotation, *Psychometrika* **42**, 491–522.

Table 1 Procrustes rotation: Audiometric example

Frequency	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4
500 Hz left	0.40	0.36	−0.32	−0.33	0.16	0.19	−0.33	−0.36
1000 Hz left	0.42	0.36	−0.23	−0.22	−0.05	0	−0.48	−0.36
2000 Hz left	0.37	0.36	0.24	0.22	−0.47	−0.57	−0.28	−0.36
4000 Hz left	0.28	0.36	0.47	0.55	0.43	0.38	−0.16	−0.36
500 Hz right	0.34	0.36	−0.39	−0.33	0.26	0.19	0.49	0.36
1000 Hz right	0.41	0.36	−0.23	−0.22	−0.03	0	0.37	0.36
2000 Hz right	0.31	0.36	0.32	0.22	−0.56	−0.57	0.39	0.36
4000 Hz right	0.25	0.36	0.51	0.55	0.43	0.38	0.16	0.36

[4] Mosier, C.I. (1939). Determining a simple structure when loadings for certain tests are known, *Psychometrika* **4**, 149–162.
[5] Nevels, K. (1979). On Meridith's solution for weighted Procrustes rotation, *Psychometrika* **44**, 121–122.
[6] Schönemann, P.H. (1966). A generalized solution of the orthogonal Procrustes problem, *Psychometrika* **31**, 1–10.
[7] Schönemann, P.H. (1968). On two-sided orthogonal Procrustes problems, *Psychometrika* **33**, 19–33.
[8] TenBerge, J.M.F. (1979). On the equivalence of two oblique congruence rotation methods and orthogonal approximations, *Psychometrika* **44**, 359–364.
[9] TenBerge, J.M.F. & Nevels, K. (1977). A general solution to Mosier's oblique Procrustes problem, *Psychometrika* **42**, 593–600.

(*See also* **Axes in Multivariate Analysis**)

J. EDWARD JACKSON

Product-integration

Product-integration was introduced more than 110 years ago by the Italian mathematician Vito Volterra, as a tool in the solution of a certain class of differential equations. It was studied intensively by mathematicians for half a century, but finally the subject became unfashionable and lapsed into obscurity. That is a pity, since ideas of product-integration make a very natural appearance in **survival analysis**, and the development of this subject (in particular, of the **Kaplan–Meier estimator**) could have been much smoother if product-integration had been a familiar topic from the start. The Kaplan–Meier estimator is the product-integral of the **Nelson–Aalen estimator** of the cumulative hazard function; these two estimators bear the same relation to one another as the actual survival function and the actual cumulative hazard function (*see* **Survival Distributions and Their Characteristics**). There are many other applications of product-integration in survival analysis, for instance in the study of multistate processes (connected to the theory of **Markov processes**), and in the theory of **partial likelihood**.

Ordinary integration is a generalization of summation, and properties of integrals are often easily guessed by thinking of them as sums of very, very many terms (all or most of them being very small). Similarly, product-integration generalizes the taking of products; a product-integral is a product of many, many terms (all or most of them being very close to the number 1). Thinking of product-integrals in this simplistic way is actually very helpful. Properties of product-integrals are easy to guess and to understand. The theory of product-integration can be a great help in studying the statistical properties of statistical quantities which explicitly or implicitly are defined in terms of product-integrals.

Before defining product-integrals in general and exhibiting some of their properties, we discuss the relation, in survival analysis, between the survival function and the hazard function. This leads us naturally to the notion of product-integration in the simplest of contexts.

Consider a survival time T with survival function $S(t) = \Pr(T > t), t \geq 0; S(0) = 1$. Suppose T is continuously distributed with density $f(t)$ and hazard rate $\alpha(t)$. These two functions have intuitive probabilistic meanings: for a small time interval $t, t + h$, the unconditional probability $\Pr(t \leq T \leq t + h) \approx f(t)h$, while the conditional probability $\Pr(t \leq T \leq t + h | T \geq t) \approx \alpha(t)h$. In fact, the probability density $f(t) = -(\mathrm{d}/\mathrm{d}t)S(t)$ while the hazard rate $\alpha(t) = f(t)/S(t)$. One can mathematically recover the distribution function $F(t) = 1 - S(t)$ from the density by integration; $F(t) = \int_0^t f(s)\,\mathrm{d}s$. Also, one can recover the survival function from the hazard rate. Noting that $\alpha(t) = -(\mathrm{d}/\mathrm{d}t)\log S(t)$, one finds by integration [and using $S(0) = 1$, hence $\log S(0) = 0$], that $-\log S(t) = \int_0^t \alpha(s)\,\mathrm{d}s$, and hence $S(t) = \exp[-\int_0^t \alpha(s)\,\mathrm{d}s]$. This is simple enough, but neither the result nor its derivation have a probabilistic interpretation.

If the survival time T had a discrete distribution, one would introduce the discrete density $f(t) = \Pr(T = t)$ and the discrete hazard $\alpha(t) = \Pr(T = t | T \geq t) = f(t)/S(t-)$. Still the survival function can be recovered from both density and hazard, but the formula in the latter case now seems quite different: $S(t) = \prod_0^t [1 - \alpha(s)]$. The continuous case formula $S(t) = \exp[-\int_0^t \alpha(s)\,\mathrm{d}s]$ has, therefore, two major defects: first, it does not have any intuitive interpretation; secondly, it gives the wrong generalization to the discrete case.

Here is how both formulas can be unified and made intuitively interpretable. Define the cumulative hazard $A(t)$ by, in the continuous case, $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$, and in the discrete case, $A(t) = \sum_0^t \alpha(s)$. [These two formulas are special cases of the completely general expression $A(t) = \int_0^t \mathrm{d}S(s)/S(s-)$.] Now we can write, both in the continuous and the discrete case,

$$S(t) = \prod_0^t [1 - \mathrm{d}A(s)], \tag{1}$$

which can be interpreted as the product over many small time intervals $s, s + \mathrm{d}s$ making up the interval $[0, t]$, of the probability $[1 - \mathrm{d}A(s)]$. Since the hazard $\mathrm{d}A(s)$ can be thought of as the probability of dying in the interval from s to $s + \mathrm{d}s$ given survival up to the beginning of that time interval, 1 minus the hazard is the probability of surviving through the small time interval given survival up to its start. Multiplying over the small time intervals making up $[0, t]$ yields the unconditional probability of surviving past t; in other words, (1) is just the limiting form of the

equality

$$\Pr(T > t) = \prod_{i=1}^{k} \Pr(T > t_i | T > t_{i-1})$$
$$= \prod_{i=1}^{k} [1 - \Pr(T \leq t_i | T > t_{i-1}],$$

where $0 = t_0 < t_1 < \cdots < t_k = t$ is a partition of the time interval $[0, t]$.

Consider now the statistical problem of estimating the survival curve $S(t)$ given a sample of independently censored survival times. Let $t_1 < t_2 < \cdots$ denote the distinct times when deaths are observed; let r_j denote the number of individuals at risk just before time t_j and let d_j denote the number of observed deaths at time t_j. We estimate the cumulative hazard function A corresponding to S with the Nelson–Aalen estimator

$$\hat{A}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j}.$$

This is a discrete cumulative hazard function, corresponding to the discrete estimated hazard $\hat{\alpha}(t_j) = d_j / r_j$, with $\hat{\alpha}(t)$ zero for t not an observed death time. The product-integral of $\hat{A}$ is then

$$\hat{S}(t) = \prod_0^t (1 - \mathrm{d}\hat{A}) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right),$$

which is nothing other than the Kaplan–Meier estimator.

The actual definition of the product-integral in (1) is the following:

$$\prod_0^t [1 - \mathrm{d}A(s)]$$
$$= \lim_{\max |t_i - t_{i-1}| \to 0} \prod \{1 - [A(t_i) - A(t_{i-1})]\},$$

where the limit is taken over a sequence of ever finer partitions $0 = t_0 < t_1 < \cdots < t_k = t$ of the time interval $[0, t]$.

From this point we can choose either to study properties of the product-integral or define it in greater generality. Both aspects are important in applications. Let us first give a more general definition. The important generalization is that we will define product-integrals of matrix-valued functions rather than just scalar-valued functions. The concept now really comes into its own, because when we multiply a sequence of matrices together the result will generally depend on the order in which the matrices are taken. Even in the continuous case there will not be a simple exponential formula expressing the result in terms of an ordinary integral. Multiplying products of matrices turns up in the theory of Markov processes, and this connects directly to the statistical analysis of multistate models in survival analysis.

Suppose $\mathbf{X}(t)$ is a $p \times p$ matrix-valued function of time t. Suppose also that $\mathbf{X}$ (or if you like, each component of X) is right continuous with left-hand limits. Let $\mathbf{I}$ denote the identity matrix. The product-integral of $\mathbf{X}$ over the interval $[0, t]$ is now defined as

$$\prod_0^t [I + \mathrm{d}\mathbf{X}(s)]$$
$$= \lim_{\max |t_i - t_{i-1}| \to 0} \prod \{I + [\mathbf{X}(t_i) - \mathbf{X}(t_{i-1})]\},$$

where, as always, the limit is taken over a sequence of ever finer partitions $0 = t_0 < t_1 < \cdots < t_k = t$ of the time interval $[0, t]$. For the limit to exist, $\mathbf{X}$ has to be of bounded variation; equivalently, each component of $\mathbf{X}$ is the difference of two increasing functions.

Application to Markov Processes

We briefly sketch the application of product-integration to Markov processes. Suppose an individual moves between p different states as time proceeds, staying in each state for some random length of time and then jumping to another. Suppose the individual has intensity $\alpha_{ij}(t)$ of jumping from state i to state j at time t, given the whole past history (in other words, the process is Markov: the intensity only depends on the present time and the present state). Define cumulative intensities $A_{ij}(t) = \int_0^t \alpha_{ij}(s)\,\mathrm{d}s$ and the negative total cumulative intensity of leaving a state $A_{ii} = -\sum_{j \neq i} A_{ij}$. Collect these into a square matrix valued function of time, $\mathbf{A}$. Then one can show that the matrix of transition probabilities $P(0, t)$, whose ij component is the probability of being in state j at time t given that the individual started at time 0 in state i, is given by a product-integral of $\mathbf{A}$:

$$P(0, t) = \prod_0^t [I + \mathrm{d}\mathbf{A}(s)].$$

This formula generalizes the usual formula for transition probabilities of a discrete time Markov chain, since the matrix $[\mathbf{I} + \mathrm{d}\mathbf{A}(s)]$ can be thought of as the transition probability matrix for the small time interval $s, s + \mathrm{d}s$.

Given possibly censored observations from a Markov process, one can estimate the elements of the matrix of cumulative intensities $\mathbf{A}$ by Nelson–Aalen estimators. The corresponding estimate of the transition probabilities, the **Aalen–Johansen estimator**, is found by taking the product-integral of $\hat{\mathbf{A}}$.

Mathematical Properties

A very obvious property of product-integration is its multiplicativity. Defining the product-integral over an arbitrary time interval in the natural way, we have for $0 < s < t$

$$\prod_0^t (I + \mathrm{d}X) = \prod_0^s (I + \mathrm{d}X) \prod_s^t (I + \mathrm{d}X).$$

We can guess many other useful properties of product-integrals by looking at various simple identities for finite products. For instance, it is often important to study the difference between two product-integrals. Now, if $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ are two sequences of numbers, then we have the identity:

$$\prod (1 + a_i) - \prod (1 + b_i) = \sum_j \prod_{i<j} (1 + a_i)(a_j - b_j) \times \prod_{i>j} (1 + b_i).$$

This can be easily proved by replacing the middle term on the right, $(a_j - b_j)$, by $(1 + a_i) - (1 + b_i)$. Expanding about this difference, the right-hand side becomes

$$\sum_j \left[\prod_{i \le j} (1 + a_i) \prod_{i>j} (1 + b_i) - \prod_{i \le j-1} (1 + a_i) \times \prod_{i>j-1} (1 + b_i) \right].$$

This is a telescoping sum; writing out the terms one by one, the whole expression collapses to the two outside products, giving the left-hand side of the identity. The same manipulations work for matrices. In general, it is therefore no surprise, replacing sums by integrals and products by product-integrals, that

$$\prod_0^t (I + \mathrm{d}X) - \prod_0^t (I + \mathrm{d}Y) = \int_{s=0}^t \prod_0^{s-} (I + \mathrm{d}X)[\mathrm{d}X(s) - \mathrm{d}Y(s)] \prod_{s+}^t (I + \mathrm{d}Y).$$

This valuable identity is called the Duhamel equation.

As an example, consider the scalar case; let A be a cumulative hazard function and $\hat{A}$ the Nelson–Aalen estimator based on a sample of censored survival times. Let S be the corresponding survival function and $\hat{S}$ the Kaplan–Meier estimator. The Duhamel equation then becomes the identity

$$\hat{S}(t) - S(t) = \int_{s=0}^t \hat{S}(s-)[\mathrm{d}\hat{A}(s) - \mathrm{d}A(s)] \frac{S(t)}{S(s)},$$

which can be exploited to get both small-sample and asymptotic results (*see* **Kaplan–Meier Estimator**).

We illustrate one other important identity in a similar manner. Note that

$$\prod_{i \le j} (1 + a_i) - \prod_{i \le j-1} (1 + a_i) = \prod_{i \le j-1} (1 + a_i) a_j.$$

Adding over j from 1 to k gives us

$$\prod_{i \le k} (1 + a_i) - 1 = \sum_j \prod_{i \le j-1} (1 + a_i) a_j.$$

Now we can guess the identity

$$\prod_0^t (I + \mathrm{d}X) - I = \int_{s=0}^t \prod_0^{s-} (I + \mathrm{d}X)\, \mathrm{d}X(s).$$

This is essentially Kolmogorov's forward equation from the theory of Markov processes (*see* **Brownian Motion and Diffusion Processes**), and it is the type of equation – solve $Y(t) = I + \int_0^t Y(s-)\, \mathrm{d}X(s)$ for unknown Y, given X – which originally motivated Volterra to invent product-integration. $Y(t) = \prod_0^t (I + \mathrm{d}X)$ is the unique solution of this equation. (It is also just a special case of the Duhamel equation when we take the second integrand Y identically equal to zero.)

Concluding Remarks

The product-integral seems first to have been used as a fundamental tool in modern survival analysis by

Aalen & Johansen [1], though it also appears in a more informal context in Cox [3] and in Kalbfleisch & Prentice [7]. Surveys of the theory of product-integration are given by Gill & Johansen [6] and Gill [5]. The former paper also pays attention to the earlier history of the subject. In particular, it is worth mentioning that a large variety of notations has been used for the product-integral, including large curly Ps, product-symbols, and the ordinary integral sign embellished with a half-circle over the top.

As well as playing a role in the theory of the Kaplan–Meier and the Aalen–Johansen estimators, the product-integral is also a useful way to write likelihoods and partial likelihoods in survival analysis, since these can be usefully thought of as continuous products of conditional likelihoods for the data in each new infinitesimal time interval given the past. The product-integral is also useful in **multivariate survival analysis**. In particular, Dabrowska's [4] multivariate product-limit estimator is based on a representation of a multivariate survival function in terms of product-integrals of a collection of higher-dimensional joint and conditional hazard functions. The book by Andersen et al. [2] gives a brief survey of the theory and many detailed applications, covering all the topics mentioned above.

References

[1] Aalen, O.O. & Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics* **5**, 141–150.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

[3] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] Dabrowska, D. (1988). Kaplan-Meier estimate on the plane, *Annals of Statistics* **16**, 1475–1489.

[5] Gill, R.D. (1994). Lectures on survival analysis, in *Lectures on Probability Theory (Ecole d'Été de Probabilités de Saint Flour XXII - 1992)*, D. Bakry, R.D. Gill, S.A. Molchanov & P. Bernard, eds. Springer-Verlag (SLNM 1581), Berlin, pp. 115–241.

[6] Gill, R.D. & Johansen, S. (1990). A survey of product-integration with a view towards application in survival analysis, *Annals of Statistics* **18**, 1501–1555.

[7] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

RICHARD D. GILL

Profile Likelihood

The profile likelihood is not a **likelihood**, but a likelihood maximized over **nuisance parameters** given the values of the parameters of interest. Let θ be the parameter(s) of interest and ϕ the nuisance parameter(s) of the statistical model $f(y|\theta, \phi)$ for data y. Once y is observed, the likelihood function is $L(\theta, \phi) = \Pr(y|\theta, \phi)$. Then the profile likelihood $P(\theta)$ is defined by $P(\theta) = L[\theta, \hat{\phi}(\theta)]$, where $\hat{\phi}(\theta)$ is the **maximum likelihood** estimate (MLE) of ϕ for given θ. The profile likelihood is thus the value of the likelihood generated as the nuisance parameter ϕ moves along the path $\phi = \hat{\phi}(\theta)$ through the parameter space. The name *profile* comes from the geometrical interpretation in three dimensions: if θ and ϕ are one-dimensional parameters, then the profile likelihood is the profile of the likelihood surface in θ as seen from a distance looking along the ϕ axis.

The profile likelihood conveniently formalizes the use of **likelihood ratio tests** to construct likelihood-based **confidence** regions for a parameter θ in the presence of nuisance parameters. For regular parametric models, the asymptotic distribution of the likelihood ratio test statistic (LRTS),

$$-2 \log \frac{L[\theta_0, \hat{\phi}(\theta_0)]}{L(\hat{\theta}, \hat{\phi})},$$

is χ^2_p, where p is the dimension of θ, under the hypothesis $\theta = \theta_0$ (*see* **Large-sample Theory**). Clearly,

$$\text{LRTS} = -2 \log \frac{P(\theta_0)}{P(\hat{\theta})},$$

and so the acceptance region for the hypothesis is the set of values of θ for which

$$\frac{P(\theta)}{P(\hat{\theta})} \geq \exp\left(-\frac{1}{2}\chi^2_{p,1-\alpha}\right),$$

where α is the size of the test. Define the profile *relative* likelihood by $PR(\theta) = P(\theta)/P(\hat{\theta})$; then the acceptance region is $[\theta : PR(\theta) \geq \exp(-\frac{1}{2}\chi^2_{p,1-\alpha})]$. Plotting the profile relative likelihood in single parameter-of-interest models provides a visual summary of the information about θ from this family of regions, and the regions themselves can be computed straightforwardly in many important models.

The main value of profile likelihoods is in constructing confidence regions for nonlinear functions of parameters, where methods based on MLEs and their asymptotic standard errors are often unsatisfactory.

A simple example is the toxicity test of a new drug on mice, described in [1]. The numbers r_i of mice dying out of n_i exposed at dose level x_i are

x_i:	422	744	948	2069
r_i:	0	1	3	5
n_i:	5	5	5	5

We fit the **logistic regression** model

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 \log x_i.$$

MLEs and standard errors for the two parameters are $\hat{\beta}_0 = -53.90(34.34)$ and $\hat{\beta}_1 = 7.93(5.08)$. For the null hypothesis $\beta_1 = 0$ of no regression on dose, the Wald test statistic $[\hat{\beta}_1/\text{se}(\hat{\beta}_1)]^2$ is 2.44, but the LRTS is 15.74, and the score test statistic is 11.29 (*see* **Likelihood**). The Wald test fails to detect the regression because the log likelihood in β_1 is far from quadratic, as is clear from the profile relative likelihood in Figure 1.

Conventional inference about the ED_{50} based on the parameter MLEs and their asymptotic covariance matrix is useless in this example because of the nonsignificance of the Wald test. However, the profile

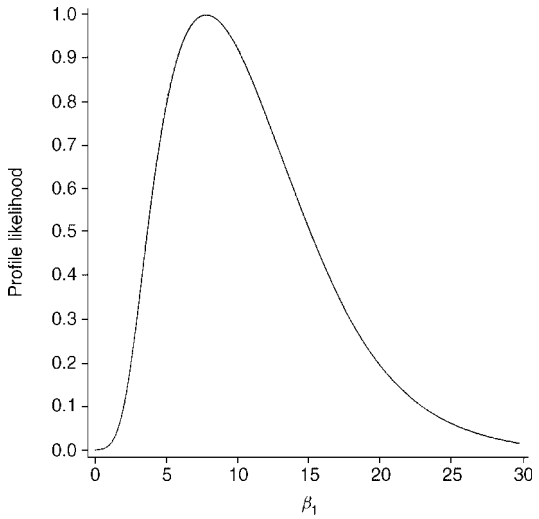

Figure 1 Profile relative likelihood for regression slope

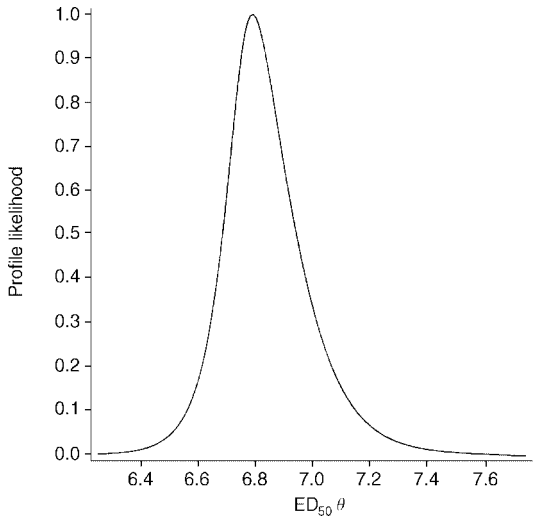

Figure 2 Profile relative likelihood for ED_{50}

likelihood is easily computed by reparameterizing the model: defining the ED_{50} by $\theta = -\beta_0/\beta_1$, the model becomes

$$\log \frac{p_i}{1-p_i} = \beta_1(\log x_i - \theta).$$

The likelihood is easily maximized over β_1 for fixed θ, generating the profile relative likelihood shown in Figure 2.

The MLE of θ is $-\beta_0/\beta_1 = 6.80$, and the 95% likelihood-based confidence interval for θ is (6.59, 7.11), corresponding to a dose interval of (728, 1224).

The advantages and disadvantages of the profile likelihood are clearly seen in the single-sample normal model $N(\mu, \sigma^2)$. Given a sample $y_1, \ldots, y_n$ with sample mean $\overline{y}$ and sum of squares $T = \sum(y_i - \overline{y})^2$, the likelihood is (omitting irrelevant constants)

$$L(\mu, \sigma) = \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}[T + n(\overline{y} - \mu)^2]\right\}.$$

Regarding μ as the parameter of interest and σ as the nuisance parameter gives

$$\begin{aligned}
\hat{\sigma}^2(\mu) &= \frac{1}{n}[T + n(\overline{y} - \mu)^2], \\
P(\mu) &= \frac{1}{\hat{\sigma}^n(\mu)} \exp\left(-\frac{n}{2}\right) \\
&= \frac{\exp(-n/2)n^{n/2}}{T^{n/2}} \left(1 + \frac{t^2}{n-1}\right)^{-n/2}, \\
PR(\mu) &= \left(1 + \frac{t^2}{n-1}\right)^{-n/2},
\end{aligned}$$

where $t = \sqrt{n}(\overline{y} - \mu)/s$ and $s^2 = T/(n-1)$.

Thus, $PR(\mu)$ is exactly the scaled t density (*see* **Student's *t* Distribution**), and a likelihood-ratio-test-based confidence interval for μ is exactly the usual t interval, though the exact distributional result improves on the asymptotic χ^2 distribution.

Regarding σ as the parameter of interest and μ as the nuisance parameter gives

$$\begin{aligned}
\hat{\mu}(\sigma) &= \overline{y}, \\
P(\sigma) &= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2}\frac{T}{\sigma^2}\right), \\
PR(\sigma) &= \left(\frac{\hat{\sigma}}{\sigma}\right)^n \exp\left[-\frac{1}{2}\left(\frac{T}{\sigma^2} - n\right)\right] \\
&= \left(\frac{T}{n\sigma^2}\right)^{n/2} \exp\left[-\frac{n}{2}\left(\frac{T}{n\sigma^2} - 1\right)\right],
\end{aligned}$$

where $\hat{\sigma}^2 = T/n = (n-1)s^2/n$. The profile likelihood here is less satisfactory because it treats μ as known to be $\overline{y}$ and so gains an extra degree of freedom, relative to the *marginal* or *restricted* likelihood based on the χ^2_ν distribution of T/σ^2, with $\nu = n - 1$. If we observed only T and not $\overline{y}$, then we would have a marginal likelihood for σ from T of

$$M(\sigma) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)\sigma^2} \left(\frac{T}{\sigma^2}\right)^{\nu/2-1} \exp\left(-\frac{1}{2}\frac{T}{\sigma^2}\right)$$

with a **restricted maximum likelihood** (REML) estimate $\tilde{\sigma}^2 = T/\nu$, giving

$$\begin{aligned}
MR(\sigma) &= \frac{M(\sigma)}{M(\tilde{\sigma})} \\
&= \left(\frac{T}{\nu\sigma^2}\right)^{\nu/2} \exp\left[-\frac{\nu}{2}\left(\frac{T}{\nu\sigma^2} - 1\right)\right].
\end{aligned}$$

Marginal likelihood-based confidence intervals based on the $\chi^2_{1,1-\alpha}$ critical value have very accurate confidence coverage and are shorter than the usual (biased) intervals based on the equal-tailed χ^2 test, while intervals based on the profile likelihood using the same critical value are even shorter because of the extra degree of freedom, but are offset to smaller values of σ and have reduced confidence coverage.

The unsatisfactory nature of the profile likelihood – its overprecision resulting from apparently

knowing the nuisance parameter as an explicit function of the data and the parameter of interest – has led to several proposals for *adjusting* or modifying it: this is a subject of continuing research interest. From a **Bayesian** viewpoint the nuisance parameter should be integrated out of the likelihood, not maximized over. Non-Bayesian adjustments have been proposed by Cox & Reid [4] and Barndorff-Nielsen [3] among others. In simple models these adjustments have an effect similar to integrating out the nuisance parameter with respect to a uniform **prior distribution**. In the normal model above when σ is the parameter of interest, integrating out μ with respect to a uniform prior gives the marginal likelihood. [A fully Bayesian analysis with the usual prior $d\sigma/\sigma$, however, gives the (normalized) *profile* likelihood as the posterior distribution of σ.] Profile likelihoods with a numerical *penalty* often appear as average (posterior mean) likelihoods in the likelihood approach of Aitkin [2].

References

[1] Aitkin, M. (1986). Statistical modelling: the likelihood approach, *Statistician* **35**, 103–113.

[2] Aitkin, M. (1991). Posterior Bayes factors (with discussion), *Journal of the Royal Statistical Society, Series B* **53**, 111–142.

[3] Barndorff-Nielsen, O. (1994). Adjusted versions of profile likelihood and directed likelihood, and extended likelihood, *Journal of the Royal Statistical Society, Series B* **56**, 125–140.

[4] Cox, D.R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion), *Journal of the Royal Statistical Society, Series B* **49**, 1–39.

M. AITKIN

Profiling Providers of Medical Care

The public debate on the cost and effectiveness of health care in recent years has brought much attention to the need for measuring and comparing performance of providers of medical care, such as physicians, clinics, hospitals, and health plans. Although comparative performance measures in health care were proposed as early as 1916, their use became widespread only in the late 1980s [1]. Prominent recent examples of assessing performance in the delivery of health care involved the analysis and publication of annual hospital mortality data by the Health Care Finance Administration and of heart surgery mortality rates by New York State [4, 5].

The comparison of measures of a provider's process of care, outcomes, or both, to normative or community standards is often called *profiling* [7]. The analysis may encompass several dimensions of care including **quality**, use of services (*see* **Health Care Utilization Data**), and cost (*see* **Health Economics**). After defining appropriate metrics and standards in each dimension of interest, the analyst proceeds to estimate provider-specific performance, to examine variations among providers and, ultimately, to identify possibly aberrant providers.

The results of profiling analyses are used to generate feedback for health care providers, to design educational and regulatory interventions by institutions and government agencies, to design marketing campaigns by hospitals and managed care organizations, and to select health care providers by individuals and managed care groups. Profiling information is often disseminated to the public in the form of "report cards" for health systems, hospitals, and individual health care practitioners [2]. A broad-based system for generating information for report cards is the Health Plan Employer Data and Information Set (see [2]).

Provider profiling entails a multifaceted set of analyses using data of varying quality, detail, and completeness. The process normally includes some form of **risk adjustment**, which is intended to account for possible differences in patient **case mix**. Commonly used methods for case mix adjustment are based on **regression** models for predicting a specific patient-level response, such as utilization, mortality, and cost. Many of these risk adjustment systems can be implemented using commercially available software such as APACHE and others. The adjusted responses are aggregated by provider, and comparisons are made to other providers or to normative data. Relative rankings and z-scores (*see* **Normal Scores**) for the difference between observed and expected performance are often used to compare providers [6]. More recent work approaches profiling analysis using a class of metrics of comparative and absolute performance, which are derived from the distribution of provider performance indices [8].

The methodologic challenges of **observational studies** generally apply to profiling analysis as well. In particular, a careful analysis would need to account for possible selection effects due to factors not reflected in the patient-level data (*see* **Propensity Score**; **Selection Bias**). In addition, the extent of **missing data** can be substantial and the pattern of missingness can vary across providers, even in well-designed studies. Recently developed methods for handling data not missing at random can be used to minimize bias and loss of efficiency (*see* **Multiple Imputation Methods**).

The statistical precision of provider-specific estimates and provider comparisons is seriously limited by the relatively small sample sizes of patients by provider involved in a typical profiling analysis. In addition, the analysis needs to account for **correlations** due to **clustering** in the data and to address the issue of **multiple comparisons** when **ranks** are reported and when performance indices are used to screen providers and identify "**outliers**". In response to these difficulties, recent methodologic work on profiling has made use of **hierarchical models**, with separate levels for modeling variation within and between providers [3, 7, 8]. The hierarchical model framework is suitable for accounting for sources of variation and clustering effects in the process of combining information across providers and for computing appropriate metrics of provider performance (*see* **Hierarchical Models in Health Service Research**; **Multilevel Models**).

As the process, cost, and outcomes of health care undergo close scrutiny, the need for careful profiling analysis is increasing. To meet this need successfully, further development of the statistical methods alone is not enough. Such work would have to be combined with considerable streamlining and improvement of the data collection mechanisms as well as with efforts

to formulate a consensus on standards for performing and reporting the results of profiling analysis.

References

[1] Codman, E. (1916). Hospital standardization, *Surgery, Gynecology, and Obstetrics* **22**, 119–120.

[2] Epstein, A. (1995). Performance reports on quality – prototypes, problems, and prospects, *New England Journal of Medicine* **333**, 57–61.

[3] Goldstein, H. & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical Society, Series A* **159**, 385–444.

[4] Green, J. & Wintfeld, N. (1995). Report cards on cardiac surgeons – assessing New York State's approach, *New England Journal of Medicine* **332**, 1129–1232.

[5] Health Care Financing Administration (1992). *Medicare Hospital Mortality Information: Technical Supplement*. US Government Printing Office, Washington.

[6] Localio, A.R., Hamory, B.H., Sharp, T.J., Weaver, S.L., Tenttave, T. & Landis, J.R. (1995). Comparing hospital mortality in adult patients with pneumonia, *Annals of Internal Medicine* **122**, 125–132.

[7] McNeil, B.J., Pederson, S. & Gatsonis, C. (1992). Current issues in profiling quality of care, *Inquiry* **29**, 298–307.

[8] Normand, S.L., Glickman, M. & Gatsonis, C. (1997). Statistical methods for profiling providers: issues and applications, *Journal of the American Statistical Association* **92**, 803–814.

CONSTANTINE GATSONIS

Prognosis

In clinical medicine, establishing a prognosis for an individual is the estimation of the relative probability of the various possible outcomes of a disease. Essentially, this involves **prediction** of the outcome probabilities for the natural history of a disease at some point after the diagnosis has been made. The outcome might be death, for example, in which case one is trying to estimate the risk of death according to characteristics of the patient, such as age and disease severity; other examples are the prognostic assessment for remission of symptoms, future **quality of life**, or permanent disability. Conventionally, this type of prediction extends the concept of natural history to include the effects of various treatment options.

(*See also* **Clinical Epidemiology**; **Natural History Study of Prognosis**; **Predictive Modeling of Prognosis**)

STEPHEN D. WALTER

Prognostic Factors for Survival

Prognostic assessment of time to an event (death) is important in many medical studies (*see* **Prognosis**). Applications are widely varied, and include:

1. predicting the outcome to assist in making treatment selection decisions
2. development of disease classification and staging systems
3. identification of biological factors that may help elucidate disease pathophysiology
4. analysis of treatment effects in randomized **clinical trials**.

Statistical modeling is a process of discovery. One usually wishes to determine what variables are associated with outcome, the nature of this association, and how the association is modulated by other variables. There is, however, a conflict between exploratory analysis for discovery and the ability to make statistically valid statements about the resulting model. Consequently, it is useful to distinguish exploratory prognostic investigations from confirmatory studies. In the latter, the form of the model, the identity of the variables, and the nature of their representations (e.g. **binary** with specified cutpoints, **categorical** with specified categories, linear, etc.) should be specified in advance. Under such conditions, valid statistical statements can be made about the model. Simon & Altman have defined three phases of prognostic factor studies and offered guidelines for phase III (confirmatory) prognostic factor studies [17].

When a statistical model is used for the analysis of randomized clinical trials, the need for making statistically valid statements is paramount. Consequently, this type of application should be dealt with as a confirmatory investigation. In some confirmatory prognostic studies even the regression coefficients of the model are specified in advance so that the data will provide a valid estimate of predictiveness of the model. Confirmatory prognostic studies are rarely performed. Consequently, the literature of prognostic factors in medicine is often contradictory and difficult for practicing physicians to utilize.

The process of model development and **validation** can be partially simulated by splitting the data into two parts. Investigators are usually reluctant to do this because most data sets are of limited size for assessing all the variables of interest. In many cases, however, published models have been overfitted to the data. Such models may feature spurious relationships and gross overestimates of predictiveness. When one has a very large data set, it is best to split it into a portion for model development and a portion for model validation. If the data set is too small for splitting, then it is best to limit the extent of model development to the variables of greatest interest, and to develop the model in a manner that will permit valid statistical inferences. The third alternative is to conduct the study as an exploratory process of discovery and to be emphatic about the need for validation before the model should be considered for adoption. Sample reuse methods such as the **bootstrap** or **cross-validation** can also be used if the model development process is algorithmic, although this is often not the case.

We focus attention here on factors for predicting a single outcome event, such as survival. Multistate and multievent models present special problems [1, 2].

Many types of regression models for survival data have been used. These include fully parametric models based on **exponential**, **Weibull**, **normal**, **log-normal**, **gamma**, and log-gamma distributions (*see* **Parametric Models in Survival Analysis**). **Proportional hazards** (PH) models, both parametric and nonparametric types, represent the hazard function as a product of a function of time and a function of the **covariate** vector. The most commonly used prognostic survival model is **Cox's regression** (proportional hazards) **model** [3]. Cox's model avoids parametric assumptions about the function of time. **Accelerated failure** (AF) **time** models are characterized by an assumption that the log of the survival time is the sum of a linear regression function ($\boldsymbol{\beta}'\mathbf{x}$) plus a random variable which represents the log survival of a subject with all covariate values equal to zero [12]. Exponential and Weibull regression models (with only the scale parameter depending on the covariates) are both parametric PH models and AF models.

The models mentioned above are all **fixed-effects** models in which the covariates are related to outcome through a linear functional $\theta = \boldsymbol{\beta}'\mathbf{x} = \beta_1 x_1 + \cdots + \beta_p x_p$, referred to as the *"prognostic index"*. The models differ in the assumptions they make about the underlying survival distribution and the parameter of the distribution which is determined by θ.

Random-effects models such as **frailty** models have also been studied but are not widely used yet [19].

Cox's PH model is widely used, primarily because it is a reasonably flexible model that does not require an assumption concerning the underlying survival distribution. In prognostic modeling attention is often focused on inference about specific regression coefficients or on the prognostic index. Cox's PH model is almost fully efficient for such inferences. A generalization of the PH model replaces the linear functional by $\theta = \sum f_i(x_i)$, where $f_i(\cdot)$ is a function with specified form such as a cubic **spline** [9].

For normal linear models it is well known that the inclusion of spurious or marginally important variables may increase the **mean square error** of prediction, although the residual mean square error will decrease [12]. This has led to the development of various stepwise **variable selection** algorithms. Unfortunately, the use of such methods invalidates the statistical inference statements usually associated with a regression model. Variable selection procedures are also frequently the basis for claims as to which are the "most important" covariates, when in fact there may be several models that predict about equally well. Use of stepwise procedures often makes it difficult to provide direct answers to questions such as whether use of a new **assay** provides improved predictiveness compared to use of a standard covariate alone [17]. For validity of inference it is best to avoid variable selection.

Prognostic modeling is often performed on the data at hand without much thought about sample size and the number of variables that can be studied reliably (*see* **Sample Size Determination**). Harrell [6] has suggested that the number of events should be at least ten times as large as the number of **degrees of freedom** for reliable modeling of survival data. If a continuous variable is represented by a restricted cubic spline with four knots, then two degrees of freedom are devoted to the variable. Similarly, a categorical variable with three categories is represented by two indicator variables and hence counts for two degrees of freedom. The number of variables modeled can be reduced by methods such as **principal components**, variable clustering, or using clinical summary indices that do not use the outcome data [8].

Continuous variables are most often represented linearly in prognostic modeling, but there is often interest in determining the nature of the relationship between an important continuous variable and outcome. Statisticians often approach this by using **polynomials** or by making the variable categorical (*see* **Categorizing Continuous Variables**). One of the most flexible and convenient approaches, however, is to use a restricted cubic regression spline to represent the relationship [5, 7]. This only involves defining new variables, based on the variable of interest and on the number and location of knots, and then using standard software to fit the model. As long as the number and location of the knots are specified in advance, standard significance tests (*see* **Hypothesis Testing**) and **confidence intervals** can be used to interpret the resulting relationship between the values of the covariate of interest and the outcome. Restricted cubic splines require the use of few degrees of freedom if the number of knots is kept small (e.g. 3–4). An alternative way of modeling a continuous variable is to dichotomize it using an "optimally selected" cutpoint. This approach is not recommended, however, because the optimizing of the cutpoint is a nonlinear process that is often ignored in subsequent inference using the model.

Once the form of the model is selected, the variables to be included are specified, and the manner of representing continuous or categorical variables is decided upon, the parameters of the model are generally determined to maximize a **likelihood** or **partial likelihood** function. The next step is assessment of **goodness of fit** of the model. Both analytic and graphic methods can be used. To test for the additivity of effects, models containing **interactions** can be fitted and **likelihood ratio tests** performed. To test for linearity of continuous variables, residual plots, such as martingale residual plots for the Cox model, can be used [18] (*see* **Residuals for Survival Analysis**). For parametric models there is a distributional assumption to be checked. For the PH model, the proportional hazards assumption should be checked either by examining Schoenfeld residuals or in other ways [13, 16]. One should also check for overly influential observations (*see* **Diagnostics**).

Once adequacy of fit is established, the statistical significance of individual regression coefficients can be examined using the Wald test based on the asymptotic normality of the maximum likelihood estimates (*see* **Likelihood**). Confidence intervals for individual regression coefficients or for linear combinations can be similarly determined. Risk groups can also

be established based on values of the prognostic index $\boldsymbol{\beta}'\mathbf{x}$.

In many cases it will be useful to quantify the proportion of variability explained by the model. Measures of explained variability have been proposed by several authors for survival models [11, 15] (*see* **Explained Variation Measures in Survival Analysis**). There is a bias in computing these measures on the same set of data used to fit the model. This bias can be reduced using the bootstrap or cross-validation [6, 10]. Even without bias correction, however, the proportion of variability explained often will be low, and this will temper the interpretation of highly statistically significant effects. The measures of explained variation are also quite useful for comparing different models.

Usually the covariates studied represent properties measured initially on subjects. In some applications, however, there is interest in determining whether a measurement made subsequently can be used as an early indicator of disease recurrence. This can be addressed by defining the measurement as a **time-dependent covariate**. Most survival regression models can accommodate time-dependent covariates. In using a survival model to evaluate treatment effect, however, time-dependent covariates should be generally avoided because they may be influenced by treatment and hence the regression coefficient for treatment may be misinterpreted.

It is common to analyze randomized clinical trials using survival models that account for important covariates. The model is used in order to increase power for detecting treatment effects. Since the trials are randomized there should be no **confounding** between treatment effect and prognostic factors. Since these clinical trials are often intended to be definitive bases for drug approval or public health policy, the prognostic factors to be included and the form in which they are modeled should be specified in advance. Usually only main effects are included in such models in spite of the fact that treatment by covariate interactions are of considerable interest to physicians. Incorporating such interactions, however, raises multiplicity questions which may be difficult to deal with satisfactorily (*see* **Simultaneous Inference**). Dixon & Simon [4] studied a Bayesian approach based on placing an **exchangeable** prior on the regression coefficients for treatment by covariate interactions with binary prognostic factors. Their approach is applicable to the PH and other survival models.

Survival models can also be used to evaluate treatment effects in nonrandomized settings. One approach is to use the model directly to reduce confounding between treatment and prognostic factors. An alternative approach is to develop a model; for example, a **logistic regression** model, to predict treatment assigned on the basis of prognostic factors. The prognostic index of that model, called a "propensity score", is then used as a covariate in a survival model to evaluate treatment effects [14].

References

[1] Andersen, P.K. (1988). Multistate models in survival analysis, *Statistics in Medicine* **7**, 661–670.

[2] Clayton, D. (1988). The analysis of event history data: a review of progress and outstanding problems, *Statistics in Medicine* **7**, 819–842.

[3] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] Dixon, D.O. & Simon, R. (1991). Bayesian subset analysis, *Biometrics* **47**, 871–881.

[5] Durrleman, S. & Simon, R. (1989). Flexible regression models with cubic splines, *Statistics in Medicine* **8**, 551–561.

[6] Harrell, F.E., Lee, K.L. & Mark, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine* **15**, 361–387.

[7] Harrell, F.E., Lee, K.L. & Pollock, B.G. (1988). Regression models in clinical studies: determining relationships between predictors and response, *Journal of the National Cancer Institute* **80**, 1198–1202.

[8] Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, K.B. & Rosati, R.A. (1984). Regression modeling strategies for improved prognostic prediction, *Statistics in Medicine* **3**, 143–152.

[9] Hastie, T. & Tibshirani, R. (1986). Generalized additive models (with discussion), *Statistical Science* **1**, 295–318.

[10] Houwelingen, J.C.V. & Cessie, S.L. (1990). Predictive value of statistical models, *Statistics in Medicine* **9**, 1303–1325.

[11] Korn, E.L. & Simon, R. (1990). Measures of explained variation for survival data, *Statistics in Medicine* **9**, 487–503.

[12] Miller, A.J. (1990). *Subset Selection in Regression*. Chapman & Hall, London.

[13] Pettitt, A.N. & Daud, I.B. (1990). Investigating time dependence in Cox's proportional hazards model, *Applied Statistics* **39**, 313–329.

[14] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

[15] Schemper, M. & Stare, J. (1996). Explained variation in survival analysis, *Statistics in Medicine* **15**, 1999–2012.

[16] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**, 239–241.

[17] Simon, R. & Altman, D.G. (1994). Statistical aspects of prognostic factor studies in oncology, *British Journal of Cancer* **69**, 979–985.

[18] Therneau, T.M., Grambsch, P.M. & Fleming, T.R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**, 147–160.

[19] Vaupel, J.W., Manton, K.G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **16**, 439–454.

RICHARD SIMON

Program Evaluation

Most societies are confronted by many complex health problems. A *health program* may be defined as an organized response to reduce one or more of these problems. In most cases, health programs are implemented to achieve specific objectives or outcomes by performing some type of service or intervention. Common examples include prenatal care programs to improve birth outcomes, fluoridation of public water supplies to improve community oral health, smoking cessation programs to reduce smoking behavior, school immunization programs to reduce morbidity and mortality, public insurance programs to increase use of health services, or medical treatments or preventive services to improve or maintain health.

Evaluation is the application of research methods to measure and explain the effects of a program against the objectives it set out to accomplish. Program evaluations help decision makers to understand the reasons for program performance, and to make informed judgments about improving a program, extending it to other sites, or cutting back or abolishing a program so that resources may be allocated elsewhere. In essence, evaluation is a management or decision-making tool for administrators, planners, policymakers, and other health officials.

Evaluation as a Profession

Program evaluation is a well-known, international profession. In many countries, evaluators have established associations (e.g. the American and Canadian Evaluation Associations) which hold annual conferences. The American Evaluation Association has developed guiding principles to promote the ethical conduct of evaluations by its members [1]. The *Evaluation Quarterly, Evaluation Review*, and *Evaluation and the Health Professions* contain findings of health program evaluations and advances in evaluation methods.

Evaluation Methods

Most evaluations of health programs consist of three basic steps. The first step occurs in the political realm, where evaluators work with decision makers to define the questions which the evaluation will answer about a program. While evaluations may be performed for a variety of reasons, most are conducted to answer two fundamental questions: "Did the program succeed in achieving its objectives?" and "Why is this the case?".

In general, program success or failure depends on the accuracy of its underlying "theory of cause and effect" and "theory of implementation" [7]. That is, all programs have either an explicit or implicit theory of cause and effect, which states that "if the program performs X, then Y will result" (*see* **Causation**). Programs also have a theory of implementation, which is usually defined by the protocols for implementing a program in the field. Programs may fail because of faulty implementation, or because of weaknesses in the program's implementation strategy. For example, a health promotion program to increase healthy lifestyles may be delivered by nurses as part of physical examinations, or identical services may be delivered by nurses conducting healthy lifestyle classes. The strategy for delivering the intervention – through physical examinations or classes – may affect the program's success in changing people's behavior. In short, programs are successful only when their underlying theories of implementation and cause and effect are sound. Faulty assumptions in either domain often undermine program performance.

In the second step, the evaluation is conducted to answer the questions about the program. *Impact evaluations* (also known as "outcome" or "summative" evaluations) address the first question and use experimental or **quasi-experimental designs** to estimate program effects [2, 4, 8] (*see* **Outcomes Research**). Experimental designs (*see* **Clinical Trials, Overview**) use **randomization** to determine whether observed outcomes are due to the program. However, in many programs randomization is not possible because laws prohibit excluding groups from the program, logistics prevent random assignment, the evaluation is performed after the program ends, or other reasons. In these cases, quasi-experimental designs (such as interrupted **time series**, regression–discontinuity analysis, and nonequivalent control group and static comparison group designs) are often used to estimate program effects [3]. In general, the greater the political controversy about a program in the first step, the greater the importance of having a rigorous impact design that can withstand

public scrutiny when the results of the evaluation are released to the public.

Process evaluations (also known as "evaluations of program implementation" or "formative evaluations") address the second question and attempt to explain why programs have positive, negative, or no effects by examining how they were implemented [6]. Process evaluations typically are designed to answer the following questions: (i) Was the program implemented as intended? (ii) Was the intervention strong enough to make a difference (**dose–response** relationship)? (iii) Did the **control** group receive a similar intervention from another source? (iv) Did external events weaken or reinforce the program's impact? To answer these questions, process evaluations use both quantitative methods (such as **surveys** and unobtrusive indicators) and qualitative methods (such as focus groups, ethnographic methods, and case studies). When an impact evaluation is based on a quasi-experimental design, evidence from process evaluations is useful for determining whether one or more threats to validity account for program effects.

In the third step the evaluation returns to the political realm, and findings are disseminated to decision makers, interest groups, and other constituents. A central assumption is that evaluations are useful only when their results are actually used to formulate new policy or improve program management. However, history indicates that this is often not the case [5, 9]. Formal dissemination plans may be developed to ensure that each group receives the information it wants about the program in a timely manner. Results also are more likely to be used when decision makers play an active role in creating the questions in step one.

In the end, the results of an evaluation are not the final determination of a program's worth, which is ultimately a political decision. An evaluation, however, can provide public evidence about a program to reduce uncertainties and clarify the gains and losses that different decisions might incur [9].

References

[1] American Evaluation Association, Task Force on Guiding Principles for Evaluators (1995). Guiding principles for evaluators, *Guiding Principles for Evaluators, New Directions for Program Evaluation, No. 66*, W.R. Shadish, D.L. Newman, M.A. Scheier & C. Wye, eds. Jossey-Bass, San Francisco, Chapter 2, pp. 19–26.

[2] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago.

[3] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton-Mifflin, Boston. p. 39–59.

[4] Mohr, L.B. (1995). *Impact Analysis for Program Evaluation*. Sage, Thousand Oaks.

[5] Rossi, P.H. & Freeman, H.E. (1993). *Evaluation: A Systematic Approach*. Sage, Thousand Oaks.

[6] Scheirer, M.A. (1994). Designing and using process evaluation, in *Handbook of Practical Program Evaluation*, J.S. Wholey, H.P. Hatry & K.E. Newcomer, eds. Jossey-Bass, San Francisco. pp. 40–68.

[7] Shortell, S.M. (1984). Steps for improving the study of health program implementation, *Health Services Research* **19**, 117–125.

[8] Shortell, S.M. & Richardson, W.C. (1978). *Health Program Evaluation*. C.V. Mosby, Saint Louis.

[9] Weiss, C. (1972). *Evaluation Research*. Prentice-Hall, Englewood Cliffs.

DAVID E. GREMBOWSKI

Projection Pursuit

Projection pursuit is a multivariate data analysis technique, the idea of which originates from Kruskal [15] and Switzer [29]. Its first successful implementation was by Friedman & Tukey [10] to **exploratory data analysis**, who also suggested the felicitous name *projection pursuit*. A unified notion of projection pursuit was introduced by Huber [12], which set a basis for further statistical research in the area.

As the name suggests, projection pursuit seeks interesting structures of a high p-dimensional data set by *searching* through *projections* of the data to lower k-dimensional spaces. The interesting structures include clusters, separations, or unexpected shapes or other "nonlinear structures". The nonlinear structures are opposite to linear structures. The linear structures are found via an analysis of sample mean and covariance, as attempted in classical **multivariate analysis**. The lower dimension k is usually 1 or 2 or maybe 3. A one-dimensional (projected) data set can be viewed via a histogram; a two-dimensional data set can be viewed via a scatter plot; and a three-dimensional data set may be inspected by spinning a three-dimensional scatter plot (*see* **Graphical Displays**). There are two key elements in implementing projection pursuit. One element is the *projection pursuit index*, which measures how important a projection is. The larger the index, the more interesting the projection. Another element is the *projection pursuit algorithm*, which is an **optimization** algorithm that searches stepwise over a k-dimensional space to maximize the index.

How do we choose or specify a projection pursuit index? Since it is easier to agree on what is an *uninteresting* projection, a projection pursuit index is an estimate of some distance between an uninteresting distribution and the distribution of projected data (or a projection). A natural uninteresting distribution is the standard **normal distribution**, N(0, 1), since it is the distribution that is the simplest and the one that is studied most thoroughly in statistics. Diaconis & Freedman [3] contains another argument for regarding the normal distribution as uninteresting: if the scale is fixed, then the normal distribution maximizes *entropy*, which is a standard measure of randomness (or unstructured distribution). Thus, heuristically the projection that maximizes the index (with the normal distribution as an uninteresting distribution) is most interesting (and may be called the *least normal projection*). Huber [12], Friedman [7], Jones & Sibson [14], Hall [11], Cook et al. [2], Li & Cheng [16], and Posse [21] all considered projection pursuit indices that use the standard normal distribution as the uninteresting distribution. However, the indices can easily be adapted to other uninteresting distributions. For example, Nason [18] proposed robust projection pursuit indices that measure divergence from student t-distribution.

Example (Hermite Index)

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a p-dimensional data set of size n. Consider one-dimensional projection pursuit. Then the dimension k of the lower dimensional projection space is 1 and a projection is specified by a (p-dimensional) vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = 1$. The projected data in this case are one-dimensional data points $Z_1, \ldots, Z_n$, where $Z_i = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i$ for $i = 1, \ldots, n$, the structures of which can be easily viewed via their histograms.

Assume for the moment that $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{X}_i) = \mathbf{I}_p$. Then the L_2 distance between the densities of projected data $Z_i = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i$ and N(0, 1) indicates how interesting the projection by $\boldsymbol{\alpha}$ is. Here, the L_2 distance between two functions f and g is $\int [f(z) - g(z)]^2 \, \mathrm{d}z$. Since polynomials are inexpensive to compute, Hall [11] defined his index on the basis of an estimate of the m-term Hermite expansion of the L_2 distance:

$$I_m^{\mathrm{H}}(\boldsymbol{\alpha}) = \sum_{j=1}^{m} \frac{\sqrt{\pi}}{j! \times 2^{j-1}} \left[\frac{1}{N} \sum_{i=1}^{N} H_j(\boldsymbol{\alpha}^{\tau}\mathbf{X}_i)\phi(\boldsymbol{\alpha}^{\tau}\mathbf{X}_i) \right]^2 + 2\pi^{1/2} \left[\frac{1}{N} \sum_{i=1}^{N} \phi(\boldsymbol{\alpha}^{\tau}\mathbf{X}_i) - \frac{1}{2\pi^{1/2}} \right]^2, \quad (1)$$

where ϕ is the probability density function (pdf) of N(0, 1) and the H_js are Hermite polynomials. In practice, $\mathrm{E}(\mathbf{X}_i)$ and $\mathrm{cov}(\mathbf{X}_i)$ are arbitrary and the distribution of $\mathbf{X}_i$ is unknown. Thus, a practical m-term Hermite index is simply $I_m^{\mathrm{H}}(\boldsymbol{\alpha})$ above with $\mathbf{X}_i$ replaced by the sphered data $\tilde{\mathbf{X}}_i$ of $\mathbf{X}_i$. The sphered data are invariant under scale, location, and rotation changes, and they satisfy

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{X}}_i = \mathbf{0}, \qquad \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^{\mathrm{T}} = \mathbf{I}_p. \quad (2)$$

For example, Friedman's [7] sphering scheme is $\tilde{\mathbf{X}}_i = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{U}}^\tau(\mathbf{X}_i - \overline{\mathbf{X}})$ for $i = 1, \ldots, n$, where $\hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{U}}^\tau = \hat{\mathbf{\Sigma}}$ is an **eigenvalue**–eigenvector decomposition of the sample **covariance matrix** $\hat{\mathbf{\Sigma}} = (1/n)\sum(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\tau$ and $\overline{\mathbf{X}} = \sum \mathbf{X}_i/n$ is the sample mean.

Of course, a different m will result in a different projection pursuit index. See remarks in the next two paragraphs for a discussion on choosing m. If the resulting $\boldsymbol{\alpha}$ is close to $(1, 0, \ldots, 0)$, then it indicates that the first coordinate or the first variable in the data may be interesting. However, if $\boldsymbol{\alpha}$ is close to $(1/\sqrt{2}, -1/\sqrt{2}, \ldots, 0)$, then it indicates that the difference of the first and second coordinates of the data may have an interesting structure.

Friedman's index [6] differs from Hall's in that Friedman applied a Legendre polynomial expansion of the L_2 distance between the densities of the transformed projection and a transformed standard normal random variable. The transformation is $R = 2\Phi(Z) - 1$, where $\Phi(z)$ is the cumulative distribution function (cdf) of N(0, 1). When $k = 2$, a two-dimensional Legendre index can be obtained in a similar way (cf. [7]). Other indices include those based on the Kolmogrov distance (*see* **Kolmogorov–Smirnov Test**) (Li & Cheng [16]) **skewness, kurtosis** and lower order polynomials [14], and Hermite functions [13]. Sensitivity to a particular nonlinear structure varies from one index to another; and even the sensitivity of one type index (Hermite index, say) changes as m changes. There are some preliminary studies on comparing sensitivities of the various indices, see [2] and [25].

Using sphered data in a projection pursuit procedure is very important in practice. It usually provides a large computational gain and may unmask the "linear structures" from the "nonlinear structures". The linear structures can often be found by examining the sample mean and covariance of data, or a study of **principal components** that are obtained by an analysis of the covariance matrix. See [12] for more discussions on the importance of sphering. Note that the sphered data satisfy the constraints (2), which makes the first two terms of Legendre, Hermite, and other polynomial or polynomial-based function (such as Hermite function) indices much smaller than the higher order terms. See Sun [25], who suggested using $3 \le m \le 6$ in these indices.

How do we choose a projection pursuit algorithm? It is important that the optimization algorithm provides solutions close to the global maximum, or that its first several solutions provide the most interesting projections. Friedman [7] gave a practical projection pursuit algorithm (available in STATLIB [23]) for his one and two-dimensional Legendre indices, in which he uses a clever idea called *structure removal*. The idea of structure removal is to transform data along an "interesting" direction (a solution) into standard normal data while keeping structures along the other orthogonal directions unchanged. It has the effect that interesting directions already found will not be rediscovered, because normal structures are uninteresting and will therefore minimize the index. Structure removal should be implemented in all projection pursuit algorithms.

The next question is to judge how interesting a solution is from a projection pursuit algorithm. Even if data were **multivariate normal**, a projection pursuit algorithm would provide some solutions or projections $\boldsymbol{\alpha}$. It is useful to have a significance test (*see* **Hypothesis Testing**) to decide whether the apparent structure is real or just the effect of noise. Sun [24] derived a theoretical approximation for the significance level (or ***P* value**) associated with Friedman's one-dimensional projection pursuit index [7]. See Sun [26] for a rule of thumb in using this formula. Sun's method is applicable in principle to all polynomial or polynomial function-based indices. See Posse [21] for a nontrivial application of the idea to his modified chi-square index, a two-dimensional projection pursuit index. A by-product of the P value is that it helps a user to decide when to stop a projection pursuit algorithm. Consecutive large P values indicate that there might not be much interesting structure left.

Projection Pursuit Regression

The idea of projection pursuit goes beyond exploratory data analysis. For example, Friedman et al. [9] and Friedman & Stuetzle [8] applied projection pursuit in **density estimation, classification**, and **regression** problems. More recent applications can be found in Duan [5], Maller [17], Li & Cheng [16], Zhu et al. [31], Ngai and Zhang [19], Renaud [22] and Polzehl [20].

Consider a regression problem where $\mathbf{X}$ is a p-dimensional predictor variable and Y is the response variable. Projection pursuit regression approximates the regression function $f(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x})$, by a

finite sum of ridge functions

$$f^{(m)}(\mathbf{x}) = \sum_{i=1}^{m} g_i(\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{x}), \tag{3}$$

where the ridge function g_i is defined on the domain of the projected data $\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\alpha}_i$ is the ith projection matrix such that $\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\alpha}_i = \mathbf{I}$, a $p \times k$ identity matrix. The projection pursuit regression model (3) is quite general. It can be used to approximate a large class of functions f well as $m \to \infty$ for suitable choices of $\boldsymbol{\alpha}_i$ and g_i (cf. Diaconis & Shahshahani [4]). For example, the simple **linear regression**, the **neural network** modeling with one hidden layer, and the **generalized additive model** can be viewed as special cases of (3), see [26]. There are many algorithms for fitting g_i and $\boldsymbol{\alpha}_i$. Friedman & Stuetzle [8] suggested a stepwise projection pursuit regression algorithm (available in **S-PLUS** software, `http://www.insightful.com/`), which allows a backfitting, where $\boldsymbol{\alpha}_i$ are found in a stepwise way by maximizing an updated index and g_i are estimated by super-smoother smoothing through the points represented by projected data $\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}$ and the ith level of residuals (*see* **Nonparametric Regression**).

Projection pursuit density estimation approximates a high p-dimensional data density $f(\mathbf{x})$ of $\mathbf{X}$ by

$$f^{(m)}(\mathbf{x}) = f_0(\mathbf{x}) \prod_{j=1}^{m} f_j(z_j),$$

where $z_j = \boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x}$ are projected data based on a suitable choice of projection operators $\boldsymbol{\alpha}_j$; f_0 is an uninteresting density or initial model; and f_i are *low* dimensional densities based on interesting projections of data. A stepwise projection pursuit density algorithm for estimating $f^{(m)}$ may be as follows. (i) Let $f^{(0)} = f_0$; (ii) given $f^{(l-1)}$, find $\hat{\boldsymbol{\alpha}}$ that maximizes the index

$$I(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{\hat{g}_{\boldsymbol{\alpha}}(Z_i)}{\hat{g}_{\boldsymbol{\alpha}}^{(l-1)}(Z_i)}\right], \tag{4}$$

where $\hat{g}_{\boldsymbol{\alpha}}$ is a marginal density estimate based on $Z_1, \ldots, Z_n$, and $\hat{g}_{\boldsymbol{\alpha}}^{(l-1)}$ is a density estimate based on the projection to $\boldsymbol{\alpha}$ of a **Monte Carlo** random sample from $f^{(l-1)}$; (iii) set

$$f_l(\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) = \frac{\hat{g}_{\hat{\boldsymbol{\alpha}}}(\hat{\boldsymbol{\alpha}}^T\mathbf{x})}{\hat{g}_{\hat{\boldsymbol{\alpha}}}^{(l-1)}(\hat{\boldsymbol{\alpha}}^T\mathbf{x})},$$

$$f^{(l)}(x) = f^{(l-1)}(x) \times f_l(\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}).$$

Then repeat steps (ii) and (iii) until max $I(\boldsymbol{\alpha})$ is less than a preselected small value.

Conclusions

The beauty of projection pursuit is its ability to lift lower dimensional techniques to higher dimensions, and hence it helps to overcome the curse of dimensionality of a high dimensional data set. A common feature of (exploratory) projection pursuit, projection pursuit classification, projection pursuit density estimation, and projection pursuit regression is to use an *algorithm* in a stepwise way to search over low dimensional projections to maximize a projection pursuit *index*, such as (1) and (4). Projection pursuit can be applied to other multivariate analysis, too. For example, Polzehl [20] recently developed a projection pursuit **discriminant analysis**. Just like any modern data analysis technique, projection pursuit should be used in conjunction with (after) some simple preliminary data analysis, such as finding simple structures using the summary statistics, performing a predimension reduction via principal components, and sphering the data.

Searching interesting projections via a projection pursuit algorithm (which maximizes a projection pursuit index) as described above may be called *automatic* projection pursuit. This is in contrast to the *manual* projection pursuit, where a dynamic graphic system is employed and the "interesting" projections are searched by the human eye via spinning the projected data, as in *Prim-9* [30] and *Grand Tour* [1]. Manual projection pursuit can be infeasible if the data dimension is large or if a user is not experienced. Xgobi [28] is a first attempt to combine properly automatic and manual projection pursuit. Xgobi contains functions beyond projection pursuit. The interactive dynamic projection pursuit by Sun, Fleischer and Loader [27] is another new development toward the same direction. It is an S-plus function that dynamically loads some C and Fortran algorithms for conducting both one- and two-dimensional projection pursuit; it is currently the only software that provides P-values of projection pursuit indices".

References

[1] Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data, *SIAM Journal on Scientific and Statistical Computing* **6**, 128–143.

[2] Cook, D., Buja, A. & Cabrera, J. (1993). Projection pursuit indexes based on orthonormal function expansions, *Journal of Computational and Graphical Statistics* **2**, 225–250.

[3] Diaconis, P. and Freedman, D. (1984). Asymptotics of Graphical Projection Pursuit. *Annals of Statistics* **12**, 793–815.

[4] Diaconis, P. & Shahshahani, M. (1984). On nonlinear functions of linear combinations, *SIAM Journal on Scientific and Statistical Computing* **5**, 175–191.

[5] Duan, N. (1990). The adjoint projection pursuit regression, *Journal of the American Statistical Association* **85**, 1029–1038.

[6] Friedman, J.H. (1984). Smart user's guide, Report LCM001, Department of Statistics, Stanford University.

[7] Friedman, J.H. (1987). Exploratory projection pursuit, *Journal of the American Statistical Association* **82**, 249–266.

[8] Friedman, J.H. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.

[9] Friedman, J.H., Stuetzle, W. & Schroeder, A. (1984). Projection pursuit density estimation, *Journal of the American Statistical Association* **79**, 599–608.

[10] Friedman, J.H. & Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **C-23**, 881–889.

[11] Hall, P. (1989). Polynomial based projection pursuit, *Annals of Statistics* **17**, 589–605.

[12] Huber, P. (1985). Projection pursuit (with discussion), *Annals of Statistics* **13**, 435–475.

[13] Johansen, S. & Johnstone, I. (1990). Hotelling's theorem on the volume of tubes: some illustrations in simultaneous inference and data analysis, *Annals of Statistics* **18**, 652–684.

[14] Jones, M.C. & Sibson, R. (1987). What is projection pursuit (with discussion)?, *Journal of the Royal Statistical Society, Series A* **150**, 1–36.

[15] Kruskal, J.B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation", in *Statistical Computation*, R. Milton & J. Nelder, eds. Academic Press, New York, pp. 427–440.

[16] Li, G. & Cheng, P. (1993). Some recent developments in projection pursuit in China, *Statistica Sinica* **3**, 35–51.

[17] Maller, R.A. (1989). Some consistency results on projection pursuit estimators of location and scale, *Canadian Journal of Statistics* **17**, 81–90.

[18] Nason, G.P. (2001). Robust projection indices, *Journal of Royal Statistical Society, Series B: Statistical Methodology* **63**(3), 551–567.

[19] Ngai, H. & Zhang, J. (2001). Multivariate cumulative sum control charts based on projection pursuit, *Statistical Sinica* **11**(3), 747–766.

[20] Polzehl, J. (1995). Projection pursuit discriminant analysis, *Computational Statistics and Data Analysis* **20**, 141–157.

[21] Posse, C. (1995). Projection pursuit exploratory data analysis, *Computational Statistics and Data Analysis* **20**, 669–687.

[22] Renaud, O. (2002). The discrimination power of projection pursuit with different density estimators, *Biometrika* **89**(1), 129–143.

[23] STATLIB is a system for distributing statistical software and data. To get Friedman's program, send statlib@temper.stat.cmu.edu an e-mail which contains send projpur from general in the text with no subject.

[24] Sun, J. (1991). Significance levels in exploratory projection pursuit, *Biometrika* **78**, 759–769.

[25] Sun, J. (1993). Some practical aspects of exploratory projection pursuit, *Journal of Scientific and Statistical Computing* **14**, 68–80.

[26] Sun, J. (1996). Projection pursuit, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz, C. Read, D. Banks & N. Johnson, eds. Wiley, New York.

[27] Sun, J., Fleischer, J. & Loader, C. (2001). Interactive Projection Pursuit. Version 3, `http://sun.cwru.edu/jiayang/nsf/ipp.html`

[28] Swayne, D.F., Cook, D. & Buja, A. (1992). XGobi: interactive dynamic graphics in the X window systems with a link to S, in *American Statistical Association 1992 Proceedings of the Section on Statistical Graphics*. American Statistical Association, Alexandria, pp. 1–8, `http://www.research.att.com/andreas/xgobi`

[29] Switzer, P. (1970). Numerical classification, in *Geostatistics*. Plenum, New York.

[30] Tukey, J.W., Friedman, J.H. & Fisherkeller, M.A. (1988). Prim-9, an interactive multidimensional data display and analysis system, in *Dynamic Graphics for Statistics*, W.S. Cleveland & M.E. McGill, eds. Wadsworth, Belmont, pp. 91–110.

[31] Zhu, L., Fang, K. & Li, R. (1998). A projection NT-type test of multinormality based on the skewness and kurtosis indices, *Biometrika*, unpublished manuscript.

(*See also* **Cluster Analysis, Variables**; **Factor Analysis, Overview**; **Multivariate Analysis of Variance**)

JIAYANG SUN

Projections: AIDS, Cancer, Smoking

Planning for, and allocation of, health services is very much a dynamic process which is highly dependent on data and information about the current extent and distribution of diseases and injuries in a population, and, equally importantly, about how these are changing. Decisions must be made today about future health resources allocation; these decisions are ideally based on informed judgments about future epidemiologic patterns.

A prerequisite for making projections of disease and injury rates, and of their determinants, is reliable knowledge about current rates of cause-specific illness, and, in the case of determinants of health status, reliable science linking various exposures with disease and injury outcomes. Ideally, projections should be made taking into account possible changes in risks or rates of disease and injury, based usually on recent trends and on probable or possible changes in various determinants of health status. In cases where information is not considered sufficiently reliable to project future disease or injury risks, then demographic projections are still possible provided that projections of the future size and age–sex composition of the population are available. These projections assume that the age–sex patterns and levels of mortality and morbidity will remain constant, and hence the results of such projections merely indicate the implications of projected demographic changes, assuming no change in the epidemiologic parameters. This is clearly an implausible assumption and hence these projections are likely to be of limited value for health planning.

More useful projections result from making assumptions about likely or even possible changes in disease and injury rates over the projection period. The shorter the period, the more confidence there will be in the results. Projections are commonly made over a period of 10–30 years and hence the accuracy of the projections is substantially lower than for shorter projection periods, for which uncertainty in disease trends is relatively minor. To provide some bounds on this uncertainty, longer-term projections are generally prepared on the basis of various "scenarios" of the future, each of which assume different developments in the underlying determinants of disease and injury. For example, bounds may be defined by so-called "optimistic" and "pessimistic" scenarios of change over the projection period, with intermediate assumptions defining a "baseline" or "best-guess" scenario (see, for example, [10]).

Projections of disease and injury rates, and of overall survival levels, are generally made using one of two broad approaches: either past trends in rates are projected using some mathematical model relating rates in the past to time (calendar year), or projections are made on the basis of projected changes in the determinants of disease and injury levels. Mathematical **extrapolation** typically assumes that a logarithmic or higher-order polynomial expression adequately describes a previous **time series** of rates. Conversely, deterministic models usually specify future rates as a function of distal or socioeconomic variables (e.g. educational levels), or in the case of diseases where the etiology is comparatively well understood, such as major vascular diseases and several sites of cancer, as a function of qualitative and quantitative changes in exposure levels. Such methods should also consider possible changes in the underlying effect of exposure on disease outcomes, where this can be ascertained. Much more detail on these and other approaches to projecting epidemiologic parameters can be found in various studies [1, 2, 8, 10].

Whatever method is used to make projections, the results will be more or less reliable depending on the validity of the methods and the quality of the data and science used to model the past. Rather than interpreting projections as predictions of the future, it is much more appropriate to view them merely as the numerical consequences of a set of assumptions which may, or may not, turn out to be valid.

Projections have been made for a variety of health outcome indicators and for other aspects of the health sector, including health services provision. Three particular types of projections are described below in view of their importance for global health status; namely, cancer, smoking, and AIDS.

Cancer

Cancer is not a single disease, but several diseases with very different causalities (*see* **Oncology**). The etiology of some important sites in some populations is well understood (e.g. the role of tobacco in lung

cancer and cancers of the upper-aerodigestive tracts (*see* **Smoking and Health**), and alcohol as a cause of esophageal and liver cancers), whereas for others, e.g. breast cancer, the science of **causation** is still inconclusive. To the extent that the natural history of certain cancers is well known, and can be clearly described in terms of cohort-exposure to underlying determinants, projections can be made with greater reliability. This is certainly the case for lung cancer, the leading site of the disease in the world, and to a lesser extent for stomach cancer, the second most common site, for which incidence and mortality have been progressively declining for decades in most populations where trends can be reliably documented.

When making cancer projections, therefore, it may be convenient to consider three broad clusters of sites depending on major known carcinogens and overall disease trends. These sites broadly include:

1. tobacco-related cancers, which, in populations where smoking has been widespread for decades, account for up to 50% of all cancer cases [11]
2. stomach cancer, which alone has been declining during the twentieth century, thought largely to be due to less use of salt, and
3. all other sites, for which either incidence and mortality have not changed greatly, or, where large changes have occurred, are relatively insignificant causes of overall cancer burden.

Projections of lung cancer can then be made using various models that relate lung cancer risk to previous cigarette consumption, appropriately allowing for the time lag between onset of persistent smoking and disease risk, sometimes expressed as the duration of smoking (*see* **Latent Period**). Doll & Peto [4] developed a risk function of the form

$$I = 0.273 \times 10^{-12}(\text{cigarettes per day} + 6)^2 \times (\text{age} - 22.5)^{4.5},$$

whereby lung cancer incidence, I, is primarily determined by duration of smoking. Other methods have projected lung cancer as a function of various scenarios for smoking cessation, or as a direct function of cohort exposure [6, 7].

Given the steady long-term decline in stomach cancer incidence and mortality (see Figure 1), the occurrence of the disease would appear to be strongly related to dietary factors, and, in particular, to the salting of foods, which change with development. Projections of stomach cancer mortality have been made using a model with literacy (HC), GNP per capita (Y), and time (T) (as a proxy for technology and treatment advances) of the form

$$\ln M_{a,k,i} = C_{a,k,i} + \beta_1 \ln Y + \beta_2 \ln HC + \beta_3 T,$$

where $C_{a,k,i}$ is a constant term and $M_{a,k,i}$ is the predicted mortality level for age group a, sex k, and cause i [10].

It has been estimated that cancer caused about 6 million deaths in 1990, 3.4 million of whom were men, and 2.4 million (of both sexes) in the developed world. Lung cancer is already the leading site of the disease, causing 945 000 deaths worldwide in 1990, followed by cancers of the stomach (752 000 deaths), liver (500 000), colon/rectum (472 000), esophagus (358 000), and breast (322 000). By the year 2020, lung cancer is projected to cause more than 2 million deaths, mostly in men, and stomach cancer is projected to decline further to 500 000 deaths. In that year, cancer is projected to be the cause of about 10 million deaths – 70% more than today [10].

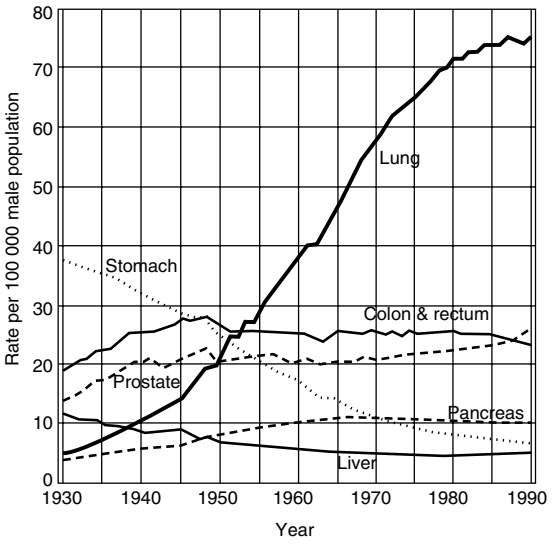

Figure 1 US male cancer trends adjusted for age to US 1970 population. Source: American Cancer Society. Reproduced from [11] by permission of Oxford University Press

Smoking

There are two aspects of particular interest in relation to projections of smoking; namely, projections of smoking patterns and levels in different populations and, much more importantly, projections of smoking-attributable mortality. Projections of smoking **prevalence** and consumption are, by themselves, of limited public health interest. However, projections of the future health effects of current (and past) smoking patterns are of very considerable public health concern, given the extensive health effects of smoking [5, 12].

Currently, about one in two men in the developing world smoke, as do 40% of men, on average, in the developed world. Large-scale prospective evidence suggests that one in two smokers will eventually, either in middle age or old age, be killed by smoking [5, 11]. If these risks were to apply globally, and if current smoking trends were to persist, then the massive increase in cigarette consumption in developing countries over the last few decades would eventually result in an annual global toll from tobacco (primarily smoking) of about 10 million deaths each year, expected to occur in the late 2020s or early 2030s [11]. In particular, of all children and young adults alive today, these prevalence and risk measures suggest that something like 200–250 million of them will eventually be killed by tobacco.

More precise age–sex–cause and region-specific projections of smoking-attributable mortality have been prepared by Murray & Lopez [10] by projecting a measure of smoking intensity originally proposed by Peto et al. [11] to estimate smoking-attributable deaths in developed countries. This methods yields similar projections to the classical **attributable risk** approach described above but has the advantage of greater specificity and detail for the projections.

AIDS

Acquired immune deficiency syndrome (**AIDS**) results from infection with the HIV virus. Epidemiologic **surveillance** of the epidemic includes monitoring the number of individuals who have contracted the HIV virus (i.e. are seropositive) as well as the number of those who develop AIDS. Follow-up cohorts of HIV-infected people suggest that virtually all will develop AIDS, on average about 10 years after infection, and that the disease will prove fatal within about two to three years from onset.

Scientists [3, 9] have used these observations to fit a **gamma distribution** (with parameter P) of the form

$$I(t) = \frac{t^{(P-1)}\mathrm{e}^{-t}}{(P-1)!}$$

to project future AIDS cases $[I(t)]$ and deaths, given estimated HIV infection rates in different populations. By the turn of the century, about 20 million adults and 1 million children worldwide are projected to be living with the HIV virus. Murray & Lopez [10] have used estimated infection rates since 1990 to project incidence and mortality from AIDS until 2020. These projections suggest that deaths from AIDS will rise from around 400 000 in 1990 to peak at around 1.7 million somewhere between 2005 and 2010. Since incidence in some regions (primarily developed) has peaked, the global death toll from the disease is expected to slowly decline beyond about 2010.

In contrast, tobacco-induced deaths are expected to continue to rise well into the twenty-first century, making tobacco by far the leading cause of death and disease by about 2020 [10].

References

[1] Benjamin, B. (1988). Mortality forecasting – the medical contribution needed, *Transactions of the Assurance Medical Society* **13**, 48–59.

[2] Caselli, G. (1996). Future longevity among the elderly, in *Health and Mortality Among Elderly Populations*, G. Caselli & A.D. Lopez, eds. Clarendon Press, Oxford, pp. 235–265.

[3] Chin, J. & Lwanga, S. (1991). Estimation and projection of adult AIDS cases: a simple epidemiological model, *Bulletin of the World Health Organization* **69**, 399–406.

[4] Doll, R. & Peto R. (1978). Cigarette smoking and bronchial carcinoma; close and time relationships among regular smokers and lifelong non-smokers, *Journal of Epidemiology and Community Health* **32**, 303–313.

[5] Doll, R., Peto, R., Hall, E., Wheatley, K. & Gray, R. (1994). Mortality in relation to smoking: 40 years' observations on male British doctors, *British Medical Journal* **309**, 901–911.

[6] Hakulinen, T. & Pukkala, E. (1981). Future incidence of lung cancer: forecasts based on hypothetical changes in the smoking habits of males, *International Journal of Epidemiology* **10**, 233–240.

[7] Harris, J. (1983). Cigarette smoking among successive birth cohorts of men and women in the United States during 1900-80, *Journal of the National Cancer Institute* **71**, 473–479.

[8] Lopez, A.D. & Hakama, M. (1986). Approaches to the projection of health statistics, in *Health Projections in Europe: Methods and Applications*. World Health Organization, Copenhagen, pp. 9–24.

[9] Mertens, T.E., Belsey, E., Stoneburner, R.L., Lowbeer, D., Sato, P., Burton, A. & Merson, M.H. (1995). Global estimates of HIV infections and AIDS: further heterogeneity in spread and impact, *Journal of Acquired Immune Deficiency Syndrome* **9**, Supplement 1, S251–S272.

[10] Murray, C.J.L. & Lopez, A.D. (1996). Alternative visions of the future: projecting mortality and disability, 1990–2020, in *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries and Risk Factors in 1990 and Projected to 2020*, C.J.L. Murray & A.D. Lopez, eds. Harvard University Press, Cambridge, MA, on behalf of the World Health Organization and the World Bank, pp. 325–395.

[11] Peto, R., Lopez, A.D., Boreham, J., Thun, M. & Heath, C., Jr (1994). *Mortality From Smoking in Developed Countries 1950–2000*. Oxford University Press, Oxford.

[12] US Department of Health and Human Services (1989). Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General, *DHHS Publication No. (CDC) 89–8411*, US Department Office on Smoking and Health.

ALAN D. LOPEZ

Promax Rotation

Promax rotation [1, 2] is a nonquartic method of performing an **oblique rotation** of a matrix $\mathbf{V}$ of dimension $(p \times k)$ associated with **principal components analysis** or **factor analysis** in order to transform these quantities into new variables by the relationship $\mathbf{B} = \mathbf{V}\boldsymbol{\Theta}$ such that $\mathbf{B}$ will approximate a **simple structure**. The matrix $\mathbf{B}$ is of dimension $(p \times k)$ and the matrix $\boldsymbol{\Theta}$ is of dimension $(k \times k)$. (*see* **Rotation of Axes**). Promax rotation is a two-stage procedure which first obtains an approximation to $\mathbf{B}$ by means of an **orthogonal rotation** such as **Varimax**. The vectors obtained from this rotation are raised to some power with negative signs restored if the power is even. This has the effect of reducing the size of small coefficients faster than larger coefficients and hence approaches a simple structure. If the result of the powering is $\mathbf{H}$, then by **Procrustes rotation**, a rotation of the original vectors $\mathbf{V}$ is performed which is a least squares approximation of $\mathbf{H}$. This becomes the Promax rotation.

The choice of power will have an effect on the results. Lower powers will approach an orthogonal rotation. (Not powering the Varimax results would leave the final solution unchanged.) Higher powers would increase the obliqueness of the solution. General practice is to use powers in the range of 2 to 4.

For the Decathlon example given in **Rotation of Axes** the Promax solutions for powers of 2 and 4 are given in Table 1 along with the original principal component characteristic vectors (*see* **Eigenvector**).

References

[1] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Earlbaum, Hillsdale.

[2] Hendrickson, A.E. & White, P.O. (1964). PROMAX: a quick method for rotation to oblique simple structure, *British Journal of Mathematical and Statistical Psychology* **17**, 65–70.

(*See also* **Axes in Multivariate Analysis**).

J. EDWARD JACKSON

Table 1 Decathlon data: characteristic and Promax rotated vectors

	Characteristic vectors				Promax rotation (power = 2)				Promax rotation (power = 4)			
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
100 m run	0.69	0.22	−0.52	−0.21	0.90	0.02	0.02	−0.09	0.94	−0.01	−0.07	−0.14
Long jump	0.79	0.18	−0.19	0.09	0.60	0.05	0.42	−0.01	0.56	0.02	0.40	−0.04
Shotput	0.70	−0.53	0.05	−0.18	0.16	0.80	0.08	−0.12	0.12	0.81	0.04	−0.13
High jump	0.67	0.13	0.14	0.40	0.17	0.01	0.73	0.05	0.07	−0.03	0.78	0.05
400 m run	0.62	0.55	−0.08	−0.42	0.82	0.01	−0.05	0.49	0.87	0.00	−0.12	0.44
110 m hurdle	0.69	0.04	−0.16	0.35	0.35	0.00	0.60	−0.19	0.28	−0.04	0.61	−0.20
Discus	0.62	−0.52	0.11	−0.23	0.11	0.81	0.01	−0.05	0.07	0.83	−0.03	−0.06
Pole vault	0.54	0.09	0.41	0.44	−0.12	0.08	0.77	0.19	−0.25	0.05	0.86	0.20
Javelin	0.43	−0.44	0.37	−0.24	−0.13	0.77	0.00	0.16	−0.17	0.80	−0.00	0.17
1500 m run	0.15	0.60	0.66	−0.28	0.05	0.00	0.06	0.94	0.05	0.01	0.10	0.93

Propensity Score

The propensity score is the **conditional probability** of exposure to treatment rather than control given observed **covariates** or, more generally, the conditional probability of selection given observed covariates. It is used to adjust for nonrandom treatment assignment or nonrandom selection. As a scalar summary of multidimensional covariates, the propensity score is often used for matching, **stratification**, or weighting adjustments. **Matching** and stratification are common in **observational studies**; that is, in studies of the effects of treatments not randomly assigned to subjects as they would be in a randomized **clinical trial**. Weighting adjustments are common in adjusting for **nonresponse** in surveys. Propensity scores have also been used as part of permutation inference and Bayesian inference (*see* **Bayesian Methods**), and have been incorporated as a variable in a model. Propensity scores were proposed by Rosenbaum & Rubin [25].

In thinking about what propensity scores may reasonably be expected to accomplish, it is useful to distinguish overt and hidden **biases**. An overt bias is visible in the data at hand. For instance, suppose that in comparing smokers and nonsmokers with recorded ages, one observes that smokers are somewhat older than nonsmokers. Then, a direct comparison of the mortality of smokers and nonsmokers ignoring age would be affected by an overt bias due to age. A hidden bias is similar, but it is not visible in the data at hand, though it may be suspected to exist. Adjustments for the propensity score reduce or eliminate overt biases, but the adjustments do little to address hidden biases, which must be investigated by other means; see Rosenbaum [21, Sections 4–9]).

In an observational study, N subjects are observed, of whom n receive the treatment and $N-n$ receive the control. Each subject has a vector $\mathbf{X}$ of observed covariates that describe the condition of subjects prior to treatment, so $\mathbf{X}$ is not affected by the treatment. Write $Z = 1$ if the subject receives the treatment and $Z = 0$ if the subject receives the control, so $n = \sum_{i=1}^{N} Z_i$. The propensity score $e(\mathbf{X})$ is the conditional probability of receiving the treatment given the covariates $\mathbf{X}$; that is, $e(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$. Although the propensity score will be discussed in the context of nonrandom treatment assignment in observational studies, it may be applied in other contexts, such as nonresponse in **sample surveys**; see [15].

The propensity score has several properties, the first of which is its balancing property. Pick a value of the propensity score $e(\mathbf{X})$, and sample at random treated subjects, those with $Z = 1$, and control subjects, those with $Z = 0$, having this same value of $e(\mathbf{X})$. Then these treated and control subjects with the same $e(\mathbf{X})$ have the same distribution of $\mathbf{X}$. To express this formally, follow Dawid [6] in writing $A \perp\!\!\!\perp B|C$ for A is conditionally independent of B given C. Then the balancing property of the propensity score says:

$$\mathbf{X} \perp\!\!\!\perp Z|e(\mathbf{X}) \tag{1}$$

or, equivalently,

$$\Pr[\mathbf{X}|Z = 1, e(\mathbf{X})] = \Pr[\mathbf{X}|Z = 0, e(\mathbf{X})].$$

The proof of (1) is straightforward; see [22, Theorem 1].

The balancing property is used in the following way. For each treated subject, find a **control** with approximately the same $e(\mathbf{X})$ forming a matched pair. Then the resulting matched sample will comprise a treated and control group with similar distributions for $\mathbf{X}$. If $\mathbf{X}$ is of high dimension, then matching exactly on $\mathbf{X}$ is difficult, but matching on a scalar $e(\mathbf{X})$ is straightforward. For instance, if $\mathbf{X}$ is a 20-dimensional vector of **binary** covariates, then there are 2^{20} or about a million possible values of $\mathbf{X}$, so finding controls that exactly match for all of $\mathbf{X}$ is infeasible with samples of reasonable size. In practice, one must estimate $e(\mathbf{X})$; for instance, using **logistic regression** of Z on $\mathbf{X}$. This is illustrated in [27] and [28]. There, 221 children with prenatal exposures to barbiturates were matched to 221 unexposed children drawn from a reservoir of 7027 potential controls. The vector $\mathbf{X}$ contained 20 covariates, such as gender, mother's socioeconomic status, mother's education, mother's cigarette use, etc. In this particular example, matching on the scalar estimate of the propensity score balanced all 20 covariates. Generally, after matching, the distributions of $\mathbf{X}$ in matched treated and control groups are compared to check that matching on the estimate of $e(\mathbf{X})$ has succeeded in balancing $\mathbf{X}$.

Alternatively, divide subjects into strata that are homogeneous in $e(\mathbf{X})$. Within strata, treated and control subjects tend to have similar distributions of $\mathbf{X}$.

This is illustrated in [26] where five strata formed from an estimated propensity score balance a 74-dimensional $\mathbf{X}$ of covariates in a study of coronary bypass surgery.

It is useful to compare and contrast the balancing property of propensity scores with the balancing property of **randomization** in a randomized experiment. Like randomization, strata that are homogeneous in the propensity score $e(\mathbf{X})$ may be quite heterogeneous in $\mathbf{X}$; however, within strata, the heterogeneity in $\mathbf{X}$ is not systematically related to the treatment group Z. For instance, in the coronary bypass example in [26], patients were denied bypass surgery if they were quite healthy and did not need it or if they were extremely ill and were unlikely to survive surgery. The stratum with the lowest probabilities of surgery, the lowest $e(\mathbf{X})$s, was extremely heterogeneous, containing both the healthiest and the sickest patients; however, within that stratum, the bypass patients and the controls had similar distributions of $\mathbf{X}$. Balancing does not eliminate heterogeneity; rather, it leaves the heterogeneity intact, but makes it nearly orthogonal to the treatment Z. Unlike randomization, adjustments for the propensity score balance the observed covariates $\mathbf{X}$ used in calculating $e(\mathbf{X})$, but they do not generally balance unobserved covariates. Randomization balances both observed and unobserved covariates, so randomization removes both overt and hidden biases, whereas adjustments for propensity scores reduce or eliminate only overt biases.

Each subject has two potential responses, a response r_{T} that would be observed if the subject received the treatment and a response r_{C} that would be observed if the subject received the control (see [18, 30, 31]). In fact, r_{T} is observed only if the subject receives the treatment, $Z = 1$, and r_{C} is observed only if the subject received the control, $Z = 0$. The effect caused by the treatment is a comparison of r_{T} and r_{C}, such as $r_{\mathrm{T}} - r_{\mathrm{C}}$; that is, a comparison of what the response would have been under treatment and under control. Because r_{T} and r_{C} cannot be observed jointly for one subject, causal effects cannot be calculated for individual subjects. Nonetheless, there are consistent and sometimes **unbiased** estimates of population summaries of causal effects under certain circumstances. For instance, in a randomized experiment the average response of treated subjects minus the average response of control subjects is an unbiased estimate of the average treatment effect, $\tau = \mathrm{E}(r_{\mathrm{T}} - r_{\mathrm{C}}) = \mathrm{E}(r_{\mathrm{T}}) - \mathrm{E}(r_{\mathrm{C}})$, and for binary responses, the sample **odds ratio** is a consistent estimate of the population odds ratio

$$\frac{\Psi = [\Pr(r_{\mathrm{T}} = 1)\Pr(r_{\mathrm{C}} = 0)]}{\Pr(r_{\mathrm{T}} = 0)\Pr(r_{\mathrm{C}} = 1)};$$

see Holland & Rubin [11]. Later in this discussion, it will be important to distinguish the odds ratio Ψ from the conditional odds ratio,

$$\omega_{\mathbf{X}} = \frac{\Pr(r_{\mathrm{T}} = 1|\mathbf{X} = \mathbf{x})\Pr(r_{\mathrm{C}} = 0|\mathbf{X} = \mathbf{x})}{\Pr(r_{\mathrm{T}} = 0|\mathbf{X} = \mathbf{x})\Pr(r_{\mathrm{C}} = 1|\mathbf{X} = \mathbf{x})}$$

given $\mathbf{X} = \mathbf{x}$,

which is generally a function of $\mathbf{x}$. Even when $\omega_{\mathbf{x}}$ is constant, not varying with $\mathbf{x}$, it does not typically equal the odds ratio Ψ that would be calculated in a randomized experiment without reference to covariates.

Adjustment for a covariate $\mathbf{X}$ in an observational study will yield **consistent** or unbiased estimates of population treatment effects if treatment assignment is ignorable, a term that will now be defined. Treatment assignment is said to be ignorable given $\mathbf{X}$ if

$$Z \perp\!\!\!\perp (r_{\mathrm{T}}, r_{\mathrm{C}})|\mathbf{X}$$

and (2)

$$0 < \Pr(Z = 1|\mathbf{X}) < 1, \quad \text{for all } \mathbf{X}.$$

This says that subjects may have unequal probabilities of receiving the treatment Z, but conditionally given the covariates $\mathbf{X}$, the assignment of subjects to treatment groups is equitable in the sense that it is unrelated to the responses subjects will later exhibit; that is, $\Pr(Z = 1|r_{\mathrm{T}}, r_{\mathrm{C}}, \mathbf{X}) = \Pr(Z = 1|\mathbf{X}) = e(\mathbf{X})$. It is not difficult to show that, if treatment assignment is ignorable, then exact matching or stratification or correct model-based adjustment for $\mathbf{X}$ can be used to estimate population causal effects such as τ or ψ; see [31] and [22, Theorem 4], with their $b(\mathbf{X}) = \mathbf{X}$. For instance, with ignorable assignment, an unbiased estimate of τ is obtained by sampling $\mathbf{X}$ at random, sampling a treated and a control subject at random with this same $\mathbf{X}$, and differencing their responses. The average of such differences is consistent for τ. With ignorable assignment, the odds ratio ψ is consistently estimated in an analogous way by first obtaining unbiased estimates of the four probabilities, $\Pr(r_{\mathrm{T}} = 1)$, $\Pr(r_{\mathrm{C}} = 0)$, $\Pr(r_{\mathrm{T}} =$

0), and $\Pr(r_C = 1)$, and, secondly, calculating the odds ratio of the unbiased estimates. Hamilton [10], Little [15], Sobel [36], and Stone [38] discuss related issues.

The second fact about the propensity score is that if treatment assignment is ignorable given $\mathbf{X}$, then it is also ignorable given just the propensity score $e(\mathbf{X})$, so if it suffices to adjust for $\mathbf{X}$ in estimating τ or ψ, then it suffices to adjust for the scalar $e(\mathbf{X})$. Formally, it may be shown [22, Theorem 3] that (2) implies

$$Z \perp\!\!\!\perp (r_T, r_C) | e(\mathbf{X}) \tag{3}$$

and

$$0 < \Pr[Z = 1 | e(\mathbf{X})] < 1, \quad \text{for all } e(\mathbf{X}).$$

In short, the propensity score tends to balance observed covariates $\mathbf{X}$ whether or not treatment assignment is ignorable. If treatment assignment is ignorable, then there is overt bias but no hidden bias, so it suffices to adjust for the observed covariates $\mathbf{X}$; but in this case, it suffices to adjust for the scalar $e(\mathbf{X})$. Specifically, if treatment assignment is ignorable, then parameters such as the average treatment effect τ or the odds ratio ψ can be consistently estimated by adjusting for the scalar $e(\mathbf{X})$ rather than the multivariate $\mathbf{X}$.

The performance of the propensity score in matching or stratification has been studied by **simulation**. Drake [7] compared propensity score adjustments and model-based adjustments when the propensity score or the model for the responses is incorrect. Gu & Rosenbaum [9] compared various matching techniques, in particular, concluding that matching on the propensity score was better than certain distance-based matching procedures when there are many covariates; say 20 or more covariates. Rubin & Thomas [34, 35] examine propensity score methods when covariates have **multivariate normal distributions**, ellipsoidal distributions, and empirical distributions derived from an example.

An alternative to matching or stratification for the propensity score is weighting adjustments. Although r_T is observed only if $Z = 1$ and r_C is observed only if $Z = 0$, the quantities Zr_T and $(1 - Z)r_C$ are always observed, although they are often zero. With a known propensity score, the quantity

$$\frac{Zr_T}{e(\mathbf{X})} - \frac{(1 - Z)r_C}{1 - e(\mathbf{X})} \tag{4}$$

may be computed for all N subjects. If treatment assignment is ignorable, then the average value of (4) may be shown to be unbiased for the average treatment effect τ; see [22]. In practice, estimated propensity scores are used in place of known propensity scores in (4), so the estimator becomes, in effect, the difference of two **Horvitz–Thompson estimators** [12] with estimated weights. As it turns out, the use of estimated weights is beneficial. The estimator based on (4) using estimated propensity scores is often superior to the analogous estimator based on true propensity scores for much the same reason that a **poststratified** estimator is often better than a sample **mean**; see [22] for specifics.

The propensity score is also used in various other ways. If the propensity score follows a linear logit model, then conditioning on a **sufficient statistic** for the parameters of the logit model eliminates unknown parameters in the propensity score, and yields exact permutation tests under the assumption that treatment assignment is ignorable; see [20, 24]. For instance, this may be used to adjust matched pairs for imbalances in covariates that were not fully controlled by the matching; see [23, 24]. The use of propensity scores in Bayesian inference is discussed by Rubin [33]. The propensity score may be used as a variable in a model-based adjustment; see [22, Corollary 4.3] for theory, and see [2, 37] for applications. Robins et al. [19] develop a novel semiparametric estimate of a regression coefficient using estimated propensity scores and various generalizations (*see* **Semiparametric Regression**). See also [3,4].

It is important to distinguish adjustment by balancing $\mathbf{X}$ from adjustment by control for $\mathbf{X}$. In a randomized experiment, random assignment of treatments would balance $\mathbf{X}$, while **blocking**, matching, or model-based adjustment, such as covariance adjustment, for $\mathbf{X}$ would control for $\mathbf{X}$. In experiments, balancing by randomization would permit estimation of the average treatment effect τ or the odds ratio ψ based on averaged proportions, but control for $\mathbf{X}$ would be required to estimate the conditional odds ratio $\omega_{\mathbf{x}}$ and other parameters that condition on a particular value of $\mathbf{X}$. Gail et al. [8] carefully develop some consequences of the distinction between balance and control in randomized experiments. Similarly, in an observational study with ignorable treatment assignment, matching or stratification for the propensity score $e(\mathbf{X})$ balances $\mathbf{X}$ permitting estimation of τ or ψ, but control for $\mathbf{X}$

would be required to estimate conditional parameters such as $\omega_{\mathbf{x}}$. Balancing and control may be combined. In experiments, this is done, for instance, using a **randomized complete blocks design**. In observational studies, this is done by combining matching or stratification $e(\mathbf{X})$ with further adjustments for $\mathbf{X}$. See [21, 23, 24, 26, 29, 32] for discussion of various aspects of combining matching or stratification with model-based adjustments.

When comparing several doses of treatment, rather than treated and control groups, a propensity score may sometimes be constructed using an ordinal logit model for the dose of treatment [13, 16]. When subjects begin as untreated and then sometimes switch to treatment after varied periods of time have elapsed, a time-dependent propensity score may be based on the hazard of receiving treatment given time-varying covariates [14]. The use of propensity scores in large case–cohort studies is discussed in [13].

For several applications of propensity scores, see [1, 2, 5, 17, 21, 27, 39].

References

[1] Aiken, L., Smith, H. & Lake, E. (1994). Lower medicare mortality among a set of hospitals known for good nursing care, *Medical Care* **32**, 771–787.

[2] Berk, R. & Newton, P. (1985). Does arrest really deter wife battery?, *American Sociological Review* **50**, 253–262.

[3] Cook, E. & Goldman, L. (1988). Asymmetric stratification. An outline for an efficient method for controlling confounding in cohort studies, *American Journal of Epidemiology* **127**, 626–639.

[4] Cook, E.F. & Goldman, L. (1989). Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score, *Journal of Clinical Epidemiology* **42**, 317–324.

[5] Czajka, J., Hirabayashi, S., Little, R.J.A. & Rubin, D.B. (1992). Projecting from advanced data using propensity modeling: an application to income tax statistics, *Journal of Business and Economic Statistics* **10**, 117–131.

[6] Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion), *Journal of the Royal Statistical Society, Series B* **41**, 1–31.

[7] Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect, *Biometrics* **49**, 1231–1236.

[8] Gail, M., Wieand, S. & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and missing covariates, *Biometrika* **71**, 431–444.

[9] Gu, X.S. & Rosenbaum, P.R. (1993). Comparison of multivariate matching methods: structures, distances and algorithms, *Journal of Computational and Graphical Statistics* **2**, 405–420.

[10] Hamilton, M. (1979). Choosing a parameter for 2o2 table or 2 o 2 o 2 table analysis, *American Journal of Epidemiology* **109**, 362–375.

[11] Holland, P. & Rubin, D. (1988). Causal inference in retrospective studies, *Evaluation Review* **12**, 203–231.

[12] Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**, 663–685.

[13] Joffe, M.M. & Rosenbaum, P.R. (1999). Propensity scores, *American Journal of Epidemiology* **150**, 327–333.

[14] Li, Y.P., Propert, K.J. & Rosenbuam, P.R. (2001). Balanced risk set matching, *Journal of the American Statistical association* **96**, 870–882.

[15] Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means, *International Statistical Review* **54**, 139–157.

[16] Lu, B., Zanutto, E., Hornik, R. & Rosenbaum, P.R. (2001). Matching with doses in an observational study of a media campaign against drug abuse, *Journal of the American Statistical Association* **96**, 1245–1253.

[17] Myers, W., Gersh, B., Fisher, L., Mock, M., Holmes, D., Schaff, H., Gillispie, S., Ryan, T. & Kaiser, G. (1987). Time to first new myocardial infarction in patients with mild angina and three-vessel disease comparing medicine and early surgery: a CASS registry study of survival, *Annals of Thoracic Surgery* **43**, 599–612.

[18] Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (in Polish). *Roczniki Nank Roiniczych Tom X*, 1–51; Reprinted in *Statistical Science* **5**, (1990). 463–480, with discussion by T. Speed and D. Rubin.

[19] Robins, J., Mark, S. & Newey, W. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders, *Biometrics* **48**, 479–495.

[20] Rosenbaum, P.R. (1984). Conditional permutation tests and the propensity score in observational studies, *Journal of the American Statistical Association* **79**, 565–574.

[21] Rosenbaum, P. (1986). Dropping out of high school in the United States: an observational study, *Journal of Educational Statistics* **11**, 207–224.

[22] Rosenbaum, P.R. (1987). Model-based direct adjustment, *Journal of the American Statistical Association* **82**, 387–394.

[23] Rosenbaum, P.R. (1988). Permutation tests for matched pairs with adjustments for covariates, *Applied Statistics* **37**, 401–411.

[24] Rosenbaum, P.R. (1995). *Observational Studies*. Springer-Verlag, New York.

[25] Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

[26] Rosenbaum, P. & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the

propensity score, *Journal of the American Statistical Association* **79**, 516–524.

[27] Rosenbaum, P. & Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *American Statistician* **39**, 33–38.

[28] Rosenbaum, P. & Rubin, D. (1985). The bias due to incomplete matching, *Biometrics* **41**, 106–116.

[29] Rubin, D.B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics* **29**, 185–203.

[30] Rubin, D.B. (1974). Estimating the causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.

[31] Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate, *Journal of Educational Statistics* **2**, 1–26.

[32] Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies, *Journal of the American Statistical Association* **74**, 318–328.

[33] Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference, in *Bayesian Statistics*, Vol. 2, J. Bernardo, M. DeGroot, D. Lindley & A. Smith, eds. Elsevier, New York, pp. 463–472.

[34] Rubin, D.B. & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with Normal distributions, *Biometrika* **79**, 797–809.

[35] Rubin, D.B. & Thomas, N. (1996). Matching using estimated propensity scores-relating theory to practice, *Biometrics* **52**, 249–264.

[36] Sobel, M. (1992). Causal inference in the social and behavioral sciences, in *A Handbook for Statistical Modelling in the Social and Behavioral Sciences*, G. Arminger, C. Clogg & M. Sobel, eds. Penum, New York Chapter 1, pp. 1–38.

[37] Solomon, P., Draine, J. & Mannion, E. (1996). The impact of individualized consultation and group workshop family education interventions on ill relative outcomes, *Journal of Nervous and Mental Disease* **184**, 252–254.

[38] Stone, R. (1993). The assumptions on which causal inference rest, *Journal of the Royal Statistical Society, Series B* **55**, 455–466.

[39] Stone, R., Obrosky, D., Singer, D., Kapoor, W., Fine, M., Hough, L., Karpf, M., Lave, J., Li, Y., Medsger, A., Redmond, C. & Ricci, E. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community acquired pneumonia, *Medical Care* **33**, 56–66.

PAUL R. ROSENBAUM

Proportional Hazards, Overview

In survival analysis, statistical models are frequently specified via the **hazard** function $\alpha(t)$. A simple model for the relation between the hazard functions in two groups (e.g. a treatment group 1 and a control group 0) is the *proportional hazards model*, where

$$\alpha_1(t) = \theta\alpha_0(t), \tag{1}$$

θ being the treatment effect. The relation (1) between the two hazards implies that the corresponding *survival functions* are related by the equation

$$S_1(t) = S_0(t)^\theta,$$

the so-called **Lehmann alternatives**.

To test the hypothesis $\alpha_1(t) = \alpha_0(t)$, several nonparametric tests are available. Among these, the **logrank test** is locally **most powerful** against proportional hazards alternatives.

In the **semiparametric** model (1), where "the baseline hazard" $\alpha_0(t)$ is left completely unspecified, there exist several estimators for the hazard ratio θ. Among these, the estimator based on the **Cox regression model** is the most frequently used; in fact, using this model, it is possible to adjust the treatment effect θ for effects of other **prognostic factors**. This proportional hazards model has gained widespread popularity in biostatistics.

In some epidemiologic applications, the hazard function $\alpha_1(t)$ in an exposed group is compared with a *known* hazard function; say, $\alpha_0(t)$. In this case, the **maximum likelihood** estimator $\hat{\theta}$ in the model (1) is the so-called *standardized mortality ratio*, which is the ratio between the observed number of deaths in the exposed group and the number of deaths one would expect if the hazard in the exposed group were $\alpha_0(t)$ (*see* **Standardization Methods**). Furthermore, the score test (*see* **Likelihood**) for the hypothesis $\theta = 1$ is the one-sample logrank test.

In conclusion, for survival data, proportional hazards has become the structure of choice, much like linearity in models for the mean of a quantitative outcome variable. However, other models are, indeed, available, including **additive hazard models** and **accelerated failure time models**.

PER KRAGH ANDERSEN

Proportional Mortality Ratio (PMR)

With mortality data classified by **cause of death**, the proportional mortality ratio (PMR) consists of a numerator that is the number of deaths from a particular cause and a denominator of total deaths from all causes. Thus, a PMR is simply the fraction (or percentage) of deaths from a particular cause.

In **descriptive epidemiology** involving characterization of PMRs by person, place, and time, it is customary to calculate standardized proportional mortality ratios (SPMRs) that use indirect standardization to adjust for age and often for age, gender, and race. See **standardization methods** for details of these calculations as well as for a historical description of the use of PMRs in vital statistics, the limitations of PMR analyses, and the relationships between PMRs and standardized mortality ratios (SMRs).

Particularly in **occupational epidemiology**, situations arise where the first and only available data consist of deaths classified by cause among persons who share a common occupational exposure. In these situations, a **proportional mortality study** is undertaken. Despite its major limitations and its often erroneous interpretation as an indicator of risk of mortality, PMR analysis can serve a useful role in descriptive epidemiology. A simple and low-cost proportional mortality study can suggest etiologic hypotheses and lead to a chain of increasingly complex analytic epidemiologic studies that ultimately establishes **causation**.

THEODORE COLTON

Proportional Mortality Study

Sometimes the only information available for an epidemiologic **observational study** consists of mortality records (usually death certificates) among a particular group of individuals who share some common exposure. This often occurs in occupational studies where the group consists of workers with a particular job classification, all workers in some industry, or employees at a particular installation or plant (*see* **Occupational Epidemiology**; **Occupational Mortality**). Examples of each of these are: nuclear shipyard workers, US Army veterans who served in Vietnam, and employees at a factory where inorganic arsenic compounds are handled. The intent is to see whether an exposure that the workers or group share is associated with increased risk of disease. Since the available information is only on deaths, it is mortality rather than morbidity that can be studied.

The key feature that characterizes a proportional mortality study is that there are no denominator data available that allow for the calculation of mortality risk or death rates. With information on causes of death among the group of decedents, one can calculate the proportion of *all* deaths due to a specific cause, namely the proportional mortality.

The essential strategy in a proportional mortality study is to compare the proportional mortality in the group of interest with proportional mortality in a comparison group. The comparison group can consist of an external **control** group, such as the general population for which **vital statistics** are available, or an internal control group; namely, a group of decedents from the same source population who do not share the exposure of interest. Thus, the essential comparison in a proportional mortality study is the ratio of proportional mortality in the study group to that in the comparison group, in other words, the **proportional mortality ratio (PMR)**.

The basic premise of a proportional mortality study is that if exposure is associated with increased risk of a specific disease, then with the available data on deaths by cause one should find proportionally more deaths from that cause among the exposed than among a comparison group of deaths, whether the comparison group consists of an external or an internal control group. Obviously, proportional mortality studies apply only to fatal diseases. For example, the common diseases of vision such as cataract, glaucoma, and age-related macular degeneration are each nonfatal and, hence, there is no place for proportional mortality studies of such disorders.

Of course, since mortality for virtually every disease has some relationship with age, there is the possibility that the decedents in the study and comparison groups have different age distributions. Consequently, age constitutes a potential **confounding** variable in virtually every proportional mortality study. To adjust for such potential confounding, it is customary to use **standardization methods** – in particular, indirect standardization – and to calculate a standardized proportional mortality ratio (SPMR). (Although standardization is usually by age, one can also standardize on other characteristics such as gender, race, and calendar time.)

Thus, using the three examples of occupational exposure described above, there have been proportional mortality studies conducted with calculation of SPMRs that have examined the following: (i) proportional mortality for leukemia and hematopoietic malignancies among shipyard nuclear workers vs. an internal comparison group of workers at these same shipyards but who were not involved in nuclear work [11]; (ii) proportional mortality from accidents and violent causes among US Army veterans who served in Vietnam vs. an internal comparison group of US Army veterans of the Vietnam era but who did not serve in Vietnam [2]; and (iii) proportional mortality from cancer – in particular, lung and skin – among employees at an arsenical factory in the UK vs. an external comparison group of population deaths (obtained from the death register of the area in which the factory was located) [7].

Interpretation of Proportional Mortality Studies

Proportional mortality studies need particularly cautious interpretation. A common and naive misinterpretation is to view the PMR or SPMR as equivalent to a **relative risk** as obtained in a **cohort study** or an **odds ratio** as obtained in a **case–control study**. Proportional mortality does *not* measure the *risk* of death from that cause. As defined, it measures only the relative frequency of that particular cause among all causes of death.

The basic limitation of a proportional mortality study is that one cannot determine whether an increase in proportional mortality for a particular cause of interest resulted from an increase in the risk of death from that cause (the basic premise underlying the study design), or from a deficit in mortality; namely, a lowering of risk of mortality, for various causes other than that of particular interest. For example, if a proportional mortality analysis resulted in an increased proportional mortality for some specific cancer, then one could not tell whether there was indeed an increased risk of death from that particular cancer, or whether the increased proportional mortality might have resulted from a lower risk of, say, mortality from cardiovascular diseases and accidents.

Even if there is increased risk of mortality from a particular cause, another limitation of proportional mortality is that one cannot distinguish whether such an increase resulted from an increase in the **incidence rate** of the disease or a worsening of the prognosis among existing (prevalent) cases of the disease.

The aforementioned basic limitations of the proportional mortality study severely limit its analytic potential. Proportional mortality studies are often the first, or an early, attempt to explore an association epidemiologically. Why are proportional mortality studies undertaken? In comparison with **analytic epidemiologic** studies, proportional mortality studies can be completed much more rapidly and with considerably less expenditure of resources. Often, it is the findings of a proportional mortality study that lead to a cascade of analytic epidemiologic studies of increasing complexity and cost that ultimately result in establishing a causal relationship between exposure and disease (*see* **Causation**). However, there have been proportional mortality studies with equivocal and controversial findings that have led to considerable efforts and expenditure of resources in subsequent analytic epidemiologic studies that ultimately result in the assurance to an alarmed public that there is no **association** between exposure and disease. An example of the latter is a proportional mortality study of the association of low-level occupational exposure to radiation among nuclear workers at Portsmouth Naval Shipyard, UK, with leukemia and hematopoietic cancers; a subsequent large-scale and expensive historical cohort study of mortality failed to find increased cancer mortality risks from radiation exposure [12]. In fact, a further analysis of more detailed and complete data at Portsmouth Naval Shipyard revealed **misclassification** bias in designating the causes of death in the initial proportional mortality study [5].

An important point in interpretation is that the same concerns with chance, bias (*see* **Bias in Observational Studies**), and confounding that apply to cohort and case–control studies apply also to proportional mortality studies. Of particular concern in proportional mortality studies are: the completeness of ascertainment of deaths; the accuracy of the coding of **causes of death**; the definition of exposure and its possible misclassification; and proper accounting for relevant confounding variables.

Most epidemiology textbooks, such as Henenkens & Buring [6] and Rothman & Greenland [14], describe proportional mortality studies, their characteristics, and limitations. More details about study design and analysis appear in occupational epidemiology texts such as Checkoway et al. [3] and Monson [10]. Considerable details on the statistical analysis and modeling of proportional mortality data appear in Breslow & Day [1]. One can also view a proportional mortality study as a special type of case–control study, as pointed out by Miettinen & Wang [9], and deploy the relevant design and analysis strategies for that design, in particular the calculation of a mortality odds ratio (MOR). This latter view of proportional mortality is discussed in the next section.

Design and Analysis of Proportional Mortality Studies

External Controls

The design of a proportional mortality study with external controls is indeed simple and straightforward. The basic material consists of deaths by cause among a particular group who share some exposure. One must take care that there is reasonably complete ascertainment of deaths in the series and that exposure classification is valid. As an example of an incomplete series, consider the situation of examining the employment records of an industrial plant and identifying all those instances where an employee had died. A proportional mortality analysis based on these existing records would likely find increased proportional mortality for diseases

such as acute myocardial infarction and accidents, deaths that are likely to occur during employment, and deficits in diseases such as cancer, which are likely to occur subsequent to employment and during retirement. Ideally, an appropriate proportional mortality study would include *all* deaths among those employed at the plant and would require ascertainment of deaths among former plant workers and retirees. Such ascertainment of postemployment deaths, presuming that there has not been widespread migration from the area, might entail searches of death certificates among the local health agencies, presuming that these sources can readily identify prior plant employees. Similarly, with the group of decedents assembled for analysis, one has to consider carefully the definition of exposure and have reasonable certainty that the group of deaths was indeed exposed to the agent under consideration. For example, a proportional mortality analysis of *all* employees at a plant would necessarily include office workers and others who might have minimal or no exposure to the agent under consideration. Such inclusions would tend to dilute the effects that exposure to the agent might have on proportional mortality.

The choice of the external reference standard population is also important, although one's possible choices are often extremely limited. If, for example, the group of decedents under study consists of almost exclusively white males who died during the period 1960–1975 (which was the case with the proportional mortality study of nuclear workers at Portsmouth Naval Shipyard), then one would ideally choose a reference population of white male decedents during this same calendar time. Anticipating age standardization, one needs a reference population of decedents classified by cause of death and by age, race, and calendar period. In this particular situation, although one might have preferred to use deaths from New England for the reference population, the only available proportional mortality data with this amount of detail were those for the entire US.

Once the reference population has been chosen, the calculation of SPMR is straightforward, with details of the calculations described in the article on **standardization methods**. Determination of the **standard error** of proportional mortality appears in Breslow & Day [1], although the authors warn against the conduct of statistical **inference** methods on PMRs.

Internal Controls

For a proportional mortality analysis with internal controls, one needs a group of deaths from unexposed subjects who are in other ways "comparable" with the exposed study group. For example, in a proportional mortality study of US soldiers who served in Vietnam, an obvious internal control group consists of deaths during the same time period among US soldiers who did not serve in Vietnam. In the example of shipyard nuclear workers at Portsmouth Naval Shipyard, the internal control group consisted of deaths among shipyard employees who were nonnuclear workers.

One common method for analysis of these data is to employ the methods for external controls described above to each of the exposed and unexposed groups of decedents and then to compare qualitatively the resulting SPMRs in the two groups. For example, in the initial proportional mortality study at Portsmouth Naval Shipyard [11], the SPMRs for leukemia were 5.62 for the nuclear workers (nearly a sixfold increase) and 0.71 for nonnuclear workers (close to the null value of 1.00). It is noted, however, that this initial study was particularly prone to bias in that there was (i) incomplete ascertainment of deaths; (ii) gross opportunity for misclassification of exposure (nuclear worker or not), since this was based on the recall of next of kin as to whether or not the decedent wore a radiation monitoring badge at work; and (iii) misclassification of cause of death, since the study's principal investigator determined, from his review of the data, the underlying cause of death for each decedent.

Another approach, as mentioned above, is to regard a proportional mortality study with internal controls as a variant of a case–control study. Cases consist of deaths from the cause of interest among both exposed and unexposed and controls consist of deaths from all other causes among both exposed and unexposed. Within this framework, one can analyze the data by the methods applicable to case–control studies; namely, to calculate an odds ratio. In this instance, such a calculation yields what is called a standardized mortality odds ratio (SMOR). Adjustment for confounding by age, or by other characteristics, can be accomplished with use of **Mantel–Haenszel methods**.

Viewed within the case–control framework, one might wish to be more careful in the choice of

controls in the design of such a proportional mortality study. Following the basic principle underlying the choice of controls in a case–control study; namely, that the controls should represent the source population from which the cases came, one would not necessarily choose all other deaths as controls. Instead, one would choose controls more carefully from a limited group of causes of death where there was a known lack of association of each cause with the exposure of interest. Thus, one would exclude from the control series those causes of death where there was a known or suspected association with the exposure of interest. Similar considerations apply in **hospital-based case–control studies**, in which one seeks to avoid selecting control diseases that may be associated with exposure.

Other Considerations

In some instances, one can categorize exposure according to duration and/or intensity and examine **dose–response** relationships in proportional mortality. Duration of employment often serves this purpose and, for the nuclear worker illustration, cumulative recorded radiation exposure by means of badge-monitoring constitutes an ideal dose measure. Methods for dose–response modeling with proportional mortality, based on **logistic regression**, are described by Breslow & Day [1].

There has been considerable investigation of the relationship between the PMR and the standardized mortality ratio (SMR) analyses of the same set of data. In fact, if one takes cause-specific SMRs and divides each by the all-causes SMR, then theoretically these should agree with the corresponding cause-specific PMRs. Kupper et al. [8] call this ratio of cause-specific to all-cause SMRs a relative standardized mortality ratio (RSMR). Zwerling et al. [15] give a recent example comparing the PMR and RSMR approaches with injury mortality among Iowa farmers. Decoufle et al. [4] and Roman et al. [13] also compared the PMR and SMR.

References

[1] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II. *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[2] Bullman, T.A., Kang, H.K. & Watanabe, K.K. (1990). Proportionate mortality among US Army Vietnam veterans who served in Military Region I, *American Journal of Epidemiology* **132**, 670–674.

[3] Checkoway, H., Pearce, N.E. & Crawford-Brown, D.J. (1989). *Research Methods in Occupational Epidemiology*. Oxford University Press, Oxford.

[4] Decoufle, P., Thomas, T.L. & Pickle, L.W. (1980). Comparison of the proportionate mortality ratio and standardized mortality ratio risk measures, *American Journal of Epidemiology* **111**, 263–269.

[5] Greenberg, R.G., Rosner, B., Hennekens, C., Rinsky, R. & Colton, T. (1985). An investigation of bias in a study of nuclear shipyard workers, *American Journal of Epidemiology* **121**, 1301–1308.

[6] Hennekens, C.H. & Buring, J. (1987). *Epidemiology in Medicine*. Little, Brown, & Company, Boston.

[7] Hill, A.B. & Fanning, E.L. (1948). Studies in the incidence of cancer in a factory handling inorganic compounds of arsenic. I. Mortality experience in the factory, *British Journal of Industrial Medicine* **5**, 1–6.

[8] Kupper, L.L., McMichael, A.J., Symons, M.J. & Most, B.M. (1978). On the utility of proportional mortality analysis, *Journal of Chronic Diseases* **31**, 15–22.

[9] Miettinen, O. & Wang, J.D. (1981). An alternative to the proportionate mortality ratio, *American Journal of Epidemiology* **114**, 144–148.

[10] Monson, R.R. (1990). *Occupational Epidemiology*, 2nd Ed. CRC Press, Boca Raton.

[11] Najarian, T. & Colton, T. (1978). Mortality from leukemia and cancer in shipyard nuclear workers, *Lancet* **i**, 1018–1020.

[12] Rinsky, R.A., Zumwalde, R.D., Waxweiler, R.J., Murray, W.E., Jr Bierbaum, P.J., Landrigan, P.J., Terpilak, M. & Cox, C. (1981). Cancer mortality at a naval nuclear shipyard, *Lancet* **i**, 231–235.

[13] Roman, E., Beral, V., Inskip, H., McDowall, M. & Adelstein, A. (1984). A comparison of standardized and proportional mortality ratios, *Statistics in Medicine* **3**, 7–14.

[14] Rothman, K.J. & Greenland, S. (1997). *Modern Epidemiology*, 2nd Ed. Raven-Lippincott, Philadelphia.

[15] Zwerling, C., Burmeister, L.F. & Jensen, C.M. (1995). Injury mortality among Iowa farmers, 1980–1988: comparison of PMR and SMR approaches, *American Journal of Epidemiology* **141**, 878–882.

THEODORE COLTON & R.W. CLAPP

Proportional-odds Model

The proportional odds model is a class of generalized linear models used for modeling the dependence of an ordinal response (*see* **Ordered Categorical Data**) on discrete or continuous covariates. Let Y denote the response category in the range $1, \ldots, k$, with $k \geq 2$, and let $\gamma_j = \Pr(Y \leq j|\mathbf{x})$ be the cumulative response probability when the covariate is held at $\mathbf{x}$. The most general form of the linear **logistic regression** model for the jth cumulative response probability,

$$\text{logit}(\gamma_j) = \alpha_j - \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}, \tag{1}$$

is one in which both the intercept α and the regression coefficient $\boldsymbol{\beta}$ depend on the category j. The proportional-odds model is a linear logistic model in which the intercepts depend on j, but the slopes are all equal. Thus, we arrive at the model

$$\text{logit}(\gamma_j) = \alpha_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \tag{2}$$

asserting that the graph of the $k-1$ cumulative logits against x is a series of parallel lines or planes with intercepts $\alpha_1, \ldots, \alpha_{k-1}$.

Ordinal response variables are common in a number of areas, notably survey research, food testing, industrial quality assurance, radiology, and clinical research. In a study of disease severity, for example, the degree of impairment might be described by one of a small collection of labels such as "none", "slight", "moderate", "severe", and "incapacitating". One of the most effective ways to construct a model for an ordinal response such as this is to invoke the concept of a latent, or unobserved, response Z. The actual recorded response Y is envisaged as a crude manifestation of the latent variable in such a way that the relationship is monotone:

$$\alpha_{j-1} < Z \leq \alpha_j \Longleftrightarrow Y = j. \tag{3}$$

The "cut-points" α_j are envisaged as unknown points on the latent scale. In the example described, the z interval $(-\infty, \alpha_1]$ is interpreted as no impairment; the interval $(\alpha_1, \alpha_2]$ as slight impairment, and so on. Unless the latent variable is close to one of the boundaries, similar values of the latent variable are not distinguished and give rise to identical responses.

This description of the model seems to require the observer to have a precisely measured latent variable Z available, if only to himself or herself, and to make the comparison (3) before reporting Y. Like all mathematical models of behavior, this is an idealization of what actually occurs, and is not to be taken literally, particularly at the edges. In fact, the model does not make these extreme demands on the observer. Although the model is capable of this mechanistic interpretation, it is not a necessary interpretation. What is important is not so much the mechanism but the prediction. If the model predictions are sufficiently close to observations and known limiting behavior, then all is well.

The dependence of the latent variable on the covariates may be specified by means of a linear or nonlinear model, as appropriate. In the case of a linear model, we have $Z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \varepsilon$, where ε is a random variable with cumulative distribution function F. Then the probability $\Pr(Z \leq z)$ is $F(z - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$. Relationship (3) between the latent variable and the response gives the implied model for Y in the form

$$\gamma_j = \Pr(Y \leq j) = \Pr(Z \leq \alpha_j) = F(\alpha_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}),$$

or in linearized form

$$F^{-1}(\gamma_j) = \alpha_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}.$$

If $F(z) = e^z/(1+e^z)$, implying that ε has the logistic distribution, this scheme produces the proportional-odds model (2) illustrated in Figure 1. Other choices for F produce generalized linear models of the same type. The cumulative probit model arises if ε is normal, and the grouped proportional hazards model, or complementary log–log model, arises if ε has the **extreme-value** distribution; in other words, if $\exp(\varepsilon)$ has the **exponential** or **Weibull distribution**. This derivation explains the unorthodox choice of sign for the regression coefficients in (1) and (2).

Various extensions of this scheme are possible. Suppose, for example, that the covariates affect both the location and scale of the latent variable according to the model

$$Z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{x})\varepsilon.$$

Models incorporating dispersion effects of this type are used in industrial quality assurance to detect factors whose effect is primarily on the variability of the product. The implied model for Y is then

$$F^{-1}(\gamma_j) = \frac{\alpha_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}}{\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{x})}, \tag{4}$$

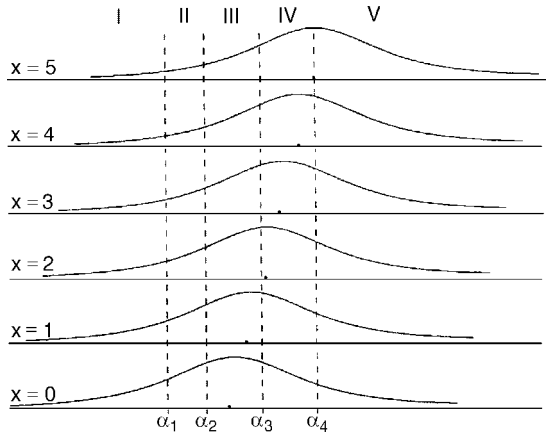

Figure 1 Diagram illustrating how the distribution of the latent variable Z changes with x in the proportional-odds model. The horizontal axis represents the latent variable, and the recorded categories are denoted by roman numerals attached to the five contiguous Z intervals. Over the range of x values shown, the probability for category IV is almost constant. By contrast, the probabilities for categories I and V vary by factors of 3–4 over the same range

which is no longer linearizable. Models (1), (2), and (4) have the limiting property for extreme covariate values, i.e. as $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \to \pm\infty$, all the probability accumulates in one of the extreme categories.

The class of models derivable in this way using a latent variable all have an important invariance, or closure, property connected with the amalgamation of adjacent response categories. Suppose that model (4) with $k > 2$ response categories is correct. If categories j and $j+1$ are combined into a single new response category, then model (4) still applies, but with k reduced to $k-1$ and with α_j deleted. In general, information is lost when categories are amalgamated, so the maximum likelihood estimate is affected. In extreme cases, the parameters might not be estimable from the reduced data. The model, however, is invariant, and the same regression parameters apply to the reduced data. By contrast, most of the competing models described at the end of this article are not closed under category amalgamation.

The term "proportional odds" stems from the fact that in model (3) the **odds** of the event $Y \leq j$ satisfies

$$\text{odds}(Y \leq j|\mathbf{x}) = \exp(\alpha_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}).$$

Consequently, the ratio of the odds of the event $Y \leq j$ for $\mathbf{x}_1$ and $\mathbf{x}_0$ is

$$\frac{\text{odds}(Y \leq j|\mathbf{x}_1)}{\text{odds}(Y \leq j|\mathbf{x}_0)} = \exp(-\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_0)),$$

which is a constant independent of j. If we arrange matters such that $x_0 = 0$ is the baseline value of the covariates, then it follows that $\exp(\alpha_j)$ is the baseline odds for the event $Y \leq j$. From this point of view, the proportional-odds model simply takes the baseline odds, which can be set arbitrarily, and multiplies by the factor $\exp(-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ to obtain the response odds at a nonbaseline covariate value. Neither of the extended models (1) nor (4) is a proportional-odds model in this sense.

By definition, the cumulative response probabilities are ordered $\gamma_1 \leq \cdots \leq \gamma_{k-1} \leq 1$. The logit transformation is strictly monotone from (0,1) to the real line. The proportional-odds model (2) must therefore satisfy the constraints $\alpha_1 \leq \cdots \leq \alpha_{k-1}$. This condition is both necessary and sufficient to ensure that the fitted response-category probabilities are nonnegative for all values of the covariate and for all values of the regression coefficient β. The same condition is necessary and sufficient for the nonlinear model (4).

The analogous condition to ensure nonnegative response probabilities in model (1) is much more complicated. Nonnegativity requires that

$$\alpha_1 + \beta_1 x \leq \cdots \leq \alpha_{k-1} + \beta_{k-1}x$$

for all values of x in some set, $\mathcal{X}$. At a minimum, $\mathcal{X}$ must include the observed covariates, but the set could be much larger, particularly if the model is to be used for extrapolation or prediction. Suppose, for simplicity, that there is a single covariate x taking values in the range $[0, \infty)$. Then, considering the logits at $x = 0$ and $x \to \infty$, we require both the intercepts and slopes to be nondecreasing in j. This condition is necessary and sufficient. Likewise, if the covariate space is bounded, say $\mathcal{X} = [-1, 1]$, then a necessary and sufficient condition is that

$$|\Delta\beta| \leq \Delta\alpha,$$

componentwise, where $\Delta\alpha$ is the difference vector with components $\alpha_{j+1} - \alpha_j$ for $j = 1, \ldots, k-2$. If there are p covariates, all in the interval $[-1, 1]$, then the necessary and sufficient condition $\sum|\Delta\beta_j| \leq \Delta\alpha$ suggests that most models in class (1) are close to (2) if p is large. Finally, if χ is a vector space,

the condition $(\Delta\beta)^T x \leq \Delta\alpha$, stating that the linear functional $\Delta\beta$ is bounded on χ, implies $\Delta\beta = 0$. The only models in (1) satisfying the nonnegativity condition for x in a vector space, are those in (2).

The proportional-odds model and related family (4) is only one of several families that are designed to be used for the analysis of ordinal data. The three main competing classes are as follows:

1. **Loglinear model** with pre-assigned category scores [6, 9].
2. Canonical regression models [2, 7, 8].
3. Continuation-ratio models [4, 5].

When the categories represent temporally ordered stages of development, such as educational attainment, it is natural to consider the conditional probability of failure at stage j conditional on survival up to stage j. The conditional probability of failure at stage j, or the hazard of stage j, or the attrition rate of stage j, is $\pi_j/(1-\gamma_{j-1})$. The natural linear logistic model, in this context called a continuation-ratio model or discrete-time **proportional-hazards** model, is

$$\text{logit}\left[\frac{\pi_j}{(1-\gamma_{j-1})}\right] = \log\left[\frac{\pi_j}{1-\gamma_j}\right] = \alpha_j - \beta_j x.$$

No order constraints are required on the parameters. However, depending on the context, it may be sensible to assume that $\beta_j = \beta$, or, less commonly, that $\alpha_j = \alpha$.

The adequacy of the proportional-odds model can in principle be tested by a generalized likelihood ratio test of model (2) against either (1) or (4). Readily available commercial software, such as SAS PROC LOGISTIC, is available for fitting the proportional-odds model. Regrettably, such software is rarely sufficiently flexible to fit alternatives such as (1) or (4), so likelihood ratio testing may require specially written computer programs. In the absence of special purpose programs, a feasible alternative for model testing is to compute the residuals, and to examine them for patterns, either by plotting or by visual inspection. However, particularly for ordinal data, the visual appearance of a residual plot can be drastically affected by the definition of residuals. Cumulative residuals seem to be more appropriate than cell residuals for many plots [11, Section 5.6].

The proportional-odds model goes back to the early work of Snell [13], Williams & Grizzle [14], and Simon [12]. Similar ideas, particularly the notion of a latent variable and its use for modeling an ordinal response, can be found in **Karl Pearson's** early work. For illustrations and numerical examples of the proportional-odds model, see [1, 3, 10], and [11, Chapter 5].

References

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

[2] Anderson, J.A. (1984). Regression and ordered categorical variables (with discussion), *Journal of the Royal Statistical Society, Series B* **46**, 1–30.

[3] Armstrong, B.G. & Sloan, M. (1989). Ordinal regression models for epidemiologic data, *American Journal of Epidemiology* **129**, 191–204.

[4] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[5] Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. MIT Press, Cambridge, Mass.

[6] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association* **74**, 537–552.

[7] Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories, *Journal of the American Statistical Association* **76**, 320–334.

[8] Greenland, S. (1994). Alternative models for ordinal logistic regression, *Statistics in Medicine* **13**, 1665–1677.

[9] Haberman, S.J. (1974). Log-linear models for frequency tables with ordered classifications, *Biometrics* **36**, 589–600.

[10] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[11] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[12] Simon, G. (1974). Alternate analyses for the singly ordered contingency table, *Journal of the American Statistical Association* **69**, 971–976.

[13] Snell, E.J. (1964). A scaling procedure for ordered categorical data, *Biometrics* **20**, 592–607.

[14] Williams, O.D. & Grizzle, J.E. (1972). Analysis of contingency tables having ordered response categories, *Journal of the American Statistical Association* **67**, 55–63.

(*See also* **Binary Data**; **Generalized Linear Model**; **Polytomous Data**; **Scores**)

PETER MCCULLAGH

Proportional-odds Regression

The well-known **proportional hazards** model can be expressed in terms of the hazard function $h(t; z)$ for a case with survival time t and **covariate** z by

$$h(t; z) = h_0(t)g(z), \tag{1}$$

where $h_0(t)$ is a baseline hazard function (*see* **Survival Distributions and Their Characteristics**). Plots of data, a priori information, and other circumstances might suggest that the proportional hazards assumption is inappropriate. An alternative model is to consider the **odds** for survival,

$$a(t; z) = \frac{[1 - S(t; z)]}{S(t; z)},$$

where $S(t; z)$ is the survival function, and assume a similar relationship to (1) above:

$$a(t; z) = a_0(t)b(z), \tag{2}$$

where $a_0(t)$ is the baseline odds of survival function. When $\log[a(t; z)]$ is plotted against t, parallel curves result, displaced by an amount $\log b(z)$ from the baseline log odds or survival function $\log a_0(t)$. Alternatively, if two cases are compared with different covariate values z_1 and z_2, then the ratio of the odds of survival functions is given by

$$\frac{a(t; z_1)}{a(t; z_2)} = \frac{b(z_1)}{b(z_2)},$$

which does not depend on t but depends only on the covariate values and the form of the "regression" function $b(\cdot)$. Eq. (2) defines the so-called *proportional-odds regression*. Particular examples are given below.

Typically, $a_0(t)$ is specified parametrically and $\log b(z)$ specified equal to a linear predictor $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$, with $\mathbf{z}$ a vector of known covariates and $\boldsymbol{\beta}$ an unknown regression parameter. For example, with $a_0(t) = t^{\phi}(\phi > 0)$, then $\phi \log t - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$ has the logistic density with $f(y) = \mathrm{e}^{y}\{1 + \mathrm{e}^{y}\}^{-2}, -\infty < y < \infty$ (*see* **Logistic Distribution**). This is also an example of **accelerated failure-time models**. If $a_0(t)$ is not specified parametrically, then estimation of $\boldsymbol{\beta}$ is still possible using a technique described by Bennett [3] based on estimating $a_0(t)$ at each failure time. Pettitt [6] develops some approximate estimates based on **ranks** of observations so that $a_0(t)$ need not be estimated explicitly. The scores used are given by Prentice [7] who developed tests for $\mathrm{H} : \beta = 0$ using linear rank statistics. Bennett [2, 3] and Pettitt [6] analyze a set of data referring to survival of lung cancer patients [5, pp. 89–90] and all authors find very similar estimates for the various parametric and nonparametric techniques. A possible explanation for this is that the signal-to-noise ratio for the data is small and therefore the ranks are almost fully efficient.

The proportional-odds model is popular in the analysis of **ordered categorical data** [1, pp. 322–324]. Hastie & Tibshirani [4, pp. 219–224] describe an additive proportional-odds regression model for ordered categorical data.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Bennett, S. (1983). Log-logistic regression models for survival data, *Applied Statistics* **32**, 165–171.

[3] Bennett, S. (1983). Analysis of survival data by the proportional odds model, *Statistics in Medicine* **2**, 273–277.

[4] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

[5] Kalbfleisch J.D. & Prentice R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[6] Pettitt, A.N. (1984). Proportional odds model for survival data and estimates using ranks, *Applied Statistics* **33**, 169–175.

[7] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika* **65**, 167–179.

ANTHONY N. PETTITT

Proportions, Inferences, and Comparisons

Binomially based inferences about one proportion, or about two proportions using data from independent samples, are among the most common tasks in statistical analysis, taught in every elementary course. However, despite the ease with which these tasks can be described and the frequency with which they are encountered, they remain controversial and inconsistently handled in statistical practice.

Numerous papers in theoretical and applied publications have covered binomial point **estimation**, interval estimation (*see* **Estimation, Interval**), and **hypothesis testing** using **exact**, approximate, and **Bayesian methods**. Yet, even with the advanced computational power now widely available, no single approach to this set of tasks has emerged as clearly preferable. The methodological choices regarding testing equality of two proportions, or estimating any disparity between them, are equally perplexing.

This article surveys, nonexhaustively, a range of methods for handling each of the problems above, based on underlying binomial or two-factor product-binomial distributions. A related problem, which is a comparison of two proportions using data from a matched-pairs design (*see* **Matched Pairs With Categorical Data**), can be placed within the framework of a single proportion inference by considering the **binomial distribution** of the number of discordant pairs of the (1,0) type after conditioning on the total number of both types ((1,0) and (0,1)) of discordant pairs (*see* **McNemar Test**). Unconditional approaches to such matched dichotomous data place the problem in the context of marginal symmetry of a 2×2 multinomial **contingency table**, requiring consideration of trinomial distributions, and are beyond our scope here (*see* **Multinomial Distribution**).

One-sample Case

We observe $X \sim \text{Bin}(N, p)$ where N is fixed, and wish to estimate or test hypotheses about the unknown parameter p.

Point Estimation

The observed proportion $\hat{p} = X/N$ is simultaneously the **method of moments**, **maximum likelihood**, and **minimum variance unbiased estimator** of p [22]. The Bayesian posterior mean, under a conjugate beta **prior** $p \sim \text{Beta}(\alpha, \beta)$, is $\hat{p}_B = (X + \alpha)/(N + \alpha + \beta)$ (*see* **Beta Distribution**; **Berkson's Fallacy**). In the special case of the **uniform** prior, Beta(1,1), the posterior mean thus reduces to $\hat{p}_B = (X + 1)/(N + 2)$, which is biased toward 0.5 compared with the maximum likelihood estimator (MLE) [32]. Another popular Bayesian choice is the Jeffreys prior, Beta$(1/2, 1/2)$, which yields $\hat{p}_B = (X + (1/2))/(N + 1)$ and produces, as will be discussed, well-behaved frequentist confidence intervals.

Interval Estimation

The discreteness of the binomial distribution – there are only $N + 1$ possible outcomes when $X \sim \text{Bin}(N, p)$ – sometimes leads to erratic and unpredictable behavior by confidence intervals for p. This is particularly apparent for intervals based on the asymptotic normal approximation $N(p, p(1 - p)/N)$ to the distribution of $\hat{p}$ (*see* **Normal Distribution**). On the basis of this approximation, nearly all entry level and many advanced courses teach the Wald $100(1 - \alpha)\%$ **confidence interval**, $\hat{p} \pm z_{1-(\alpha/2)}\sqrt{(\hat{p}(1 - \hat{p})/N)}$, for p, with z_γ the 100γth percentile of the standard normal, for example, $z_\gamma = 1.96$ when $\gamma = 0.975$. This interval offers intuitive and easily understandable properties for introductory level students. For fixed $\hat{p}$, the interval narrows as N increases while, for fixed N, the interval is widest when $\hat{p} = 0.5$ and narrows as p approaches 0 or 1.

The drawback, however, is that extremely large samples are necessary for the interval to achieve nominal $100(1 - \alpha)\%$ coverage, that is, for $100(1 - \alpha)\%$ of all intervals constructed to contain p. While this is particularly true for p near its extremes of 0 or 1, the coverage probability is low over the entire range of p. Moreover, due to the binomial's discreteness, coverage does not approach $100(1 - \alpha)\%$ monotonically as N increases [20], so that a larger N can yield poorer performance.

Many classic texts, in recognition of the asymptotic origin of the Wald interval, recommend it only

when min $(Np, N(1-p)) > 5$ or > 10 or, more stringently, when $Np(1-p) > 5$ or > 10. Yet even when $N = 40$ and $p = 0.5$ and this latter condition is thus met, the exact coverage of the Wald interval is only 91.9%. Even when N is 100, the portion of the range of p for which a 95% confidence interval achieves 95% coverage is negligible (Figure 1a).

A number of methods exist to ensure at least 100(1-α)% coverage for any fixed p [1, 3, 34]. These methods vary in computational complexity and associated software requirements. While some instructors and practitioners desire closed-form formulae, others believe that "simplicity and ease of computation have no roles to play in statistical practice" [27]. Some believe the confidence coefficient is meaningful only as a guaranteed minimum coverage probability at each use of an interval, while others find a method that guarantees only average coverage over a range of conditions to be quite satisfactory. Such varied opinions leave much room for

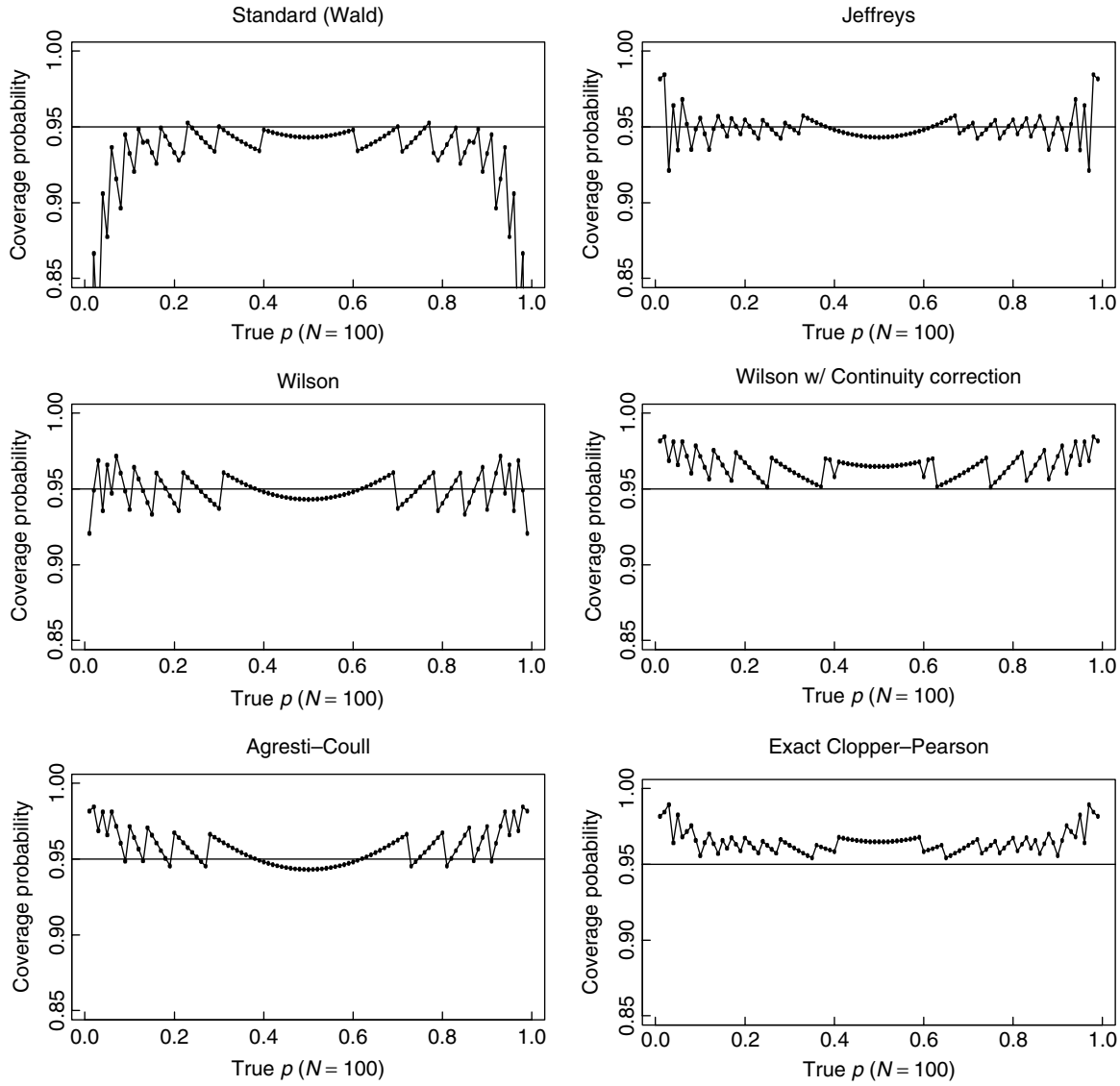

Figure 1 Coverage probabilities of six 95% confidence intervals for p, with $N = 100$

disagreement on practical recommendations. Further, to guarantee nominal coverage, one must generally form an interval by inverting the acceptance region of an exact hypothesis test. This requires computing binomial probabilities of the observed outcome and some unobserved outcomes over a range of possible values of p. Such computationally intensive calculations are now simple to perform using many statistical software packages (*see* **Software, Biostatistical**), but unsuitable for the elementary courses in which confidence intervals for proportions must be taught.

Below, we promote alternatives to the standard Wald interval that meet four criteria [18]. The confidence region must:

1. be one contiguous interval;
2. be invariant to $X \to N - X$ transformation, that is, the lower and upper endpoints of the interval for p based on X should respectively be the upper and lower endpoints of the interval for $(1 - p)$ based on $(N - X)$;
3. yield monotone endpoints in X, that is, for fixed N, $LB(X, N) < LB(X + 1, N)$ and $UB(X, N) < UB(X + 1, N)$; and
4. yield monotone endpoints in N, that is, for fixed X, $LB(X, N) > LB(X, N + 1)$, and $UB(X, N) > UB(X, N + 1)$.

We review three easily computable alternatives to the Wald interval and a more computationally demanding "exact" interval, comparing coverage properties with the Wald interval and with each other.

Jeffreys Interval

Assuming a Jeffreys prior $p \sim \text{Beta}(1/2, 1/2)$, the resulting posterior distribution is $p|X, N \sim \text{Beta}(X + (1/2), N - X + (1/2))$, and the Bayesian equal-tailed $100(1 - \alpha)\%$ credible set is formed by the $\alpha/2$ and $(1 - \alpha/2)$ quantiles. Interpreting this Bayesian credible set as a frequentist confidence interval offers desirable frequentist properties as will be demonstrated. While there are no closed-form expressions for the endpoints, all common statistical software packages (Excel, SAS, **S-PLUS**, etc.) include simple function calls for such Beta quantiles. Note that although error is allocated equally to each tail, the interval itself will be symmetric only when the posterior distribution itself is, requiring $X + (1/2) = N - X + (1/2)$ and hence $X = N/2$.

Wilson Interval

The Wald interval is formed by inverting the Wald test of $H_0 : p = p_0$. That test is based on a **Central Limit Theorem** (CLT) approximation to the distribution of $\hat{p}$ using the maximum likelihood binomial variance estimator $N\hat{p}\,\widehat{q}$, where $\widehat{q} = 1 - \hat{p}$. Wilson, in a 1927 *JASA* paper [55], introduced an interval with a similar relationship to what is now known as the score test (*see* **Likelihood**). Writing $q_0 = 1 - p_0$, the score test CLT approximation to the distribution of $\hat{p}$ uses the variance Np_0q_0 implied by the hypothesized p_0. The 100(1-α)% Wilson interval, also referred to as the score interval [2], is

$$\frac{X + \frac{z^2_{1-\alpha/2}}{2}}{N + z^2_{1-\alpha/2}} \pm \frac{z_{1-\alpha/2}\sqrt{N}}{N + z^2_{1-\alpha/2}} \sqrt{\widehat{p}\,\widehat{q} + \frac{z^2_{1-\alpha/2}}{4N}}. \quad (1)$$

As noted by Agresti and Caffo [4], this interval is centered about the pseudo-estimator $\tilde{p} = (X + z^2_{\alpha/2}/2)/(N + z^2_{\alpha/2})$, which may be viewed as a weighted average of the sample proportion and one-half or, equivalently, as obtained by adding $(z^2_{\alpha/2}/2)$ successes and $(z^2_{\alpha/2}/2)$ failures to the data. Similarly, the interval's width is a multiple of a pseudo **standard error** obtained from a weighted average of the maximum likelihood variance estimator used for the Wald interval and the true variance when $p = 1/2$.

Agresti–Coull Interval

The Agresti–Coull interval [5] takes the functional form of the standard asymptotic Wald interval but with minor adjustments in X, N, and p, to $\tilde{X} = X + (z^2_{\alpha/2}/2)$, $\tilde{N} = N + z^2_{\alpha/2}$, and $\tilde{p} = \tilde{X}/\tilde{N}$, respectively. Thus, the interval has the standard Wald-like form

$$\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{N}}}, \quad (2)$$

where $\tilde{q} = 1 - \tilde{p}$. The difference is that the whole experiment is treated as if there are $(z^2_{\alpha/2}/2)$ more successes and $(z^2_{\alpha/2}/2)$ more failures than were actually observed. For a 95% confidence interval, this has the affect of approximately adding two successes and two failures, or in a Bayesian sense, starting with a Beta(2, 2) prior for p. This prior has mean 1/2 and is concave with single mode 1/2, while the Jeffreys prior has mean 1/2 and is convex and bimodal at 0

and 1. Thus, in the Bayesian sense, the Agresti–Coull prior distribution on p is more informative.

The Agresti–Coull interval is never narrower than the Wilson interval, making it a more conservative choice. It offers a clear improvement over the Wald interval when $X = 0$ or $X = N$, for which the width of the Wald interval is zero, and corrects particularly well for the sometimes too narrow intervals and poor coverage of the Wilson method when p is close to 0 or 1. The Agresti–Coull and Wilson (score) intervals are very similar when p is near the center of its range. See [4] for a fine, highly accessible review of this interval.

Clopper–Pearson Interval

As always, it is easier to reach a performance goal on average than to guarantee such performance is always achieved across a range of conditions. The intervals above are superior to the Wald interval in providing approximately 100(1-α)% coverage averaged over the range of possible values of p, but coverage by each is below nominal for some combinations of p and N. In contrast, the Clopper–Pearson Interval [25], often called the Exact Interval, achieves at least nominal coverage for all combinations of p and N [1, 3, 25, 34]. For $X \neq 0$ or N, the 100(1-α)% Clopper–Pearson interval is

$$\left[\left(1+\frac{N-X+1}{XF_{2X,2(N-X+1)}(\frac{\alpha}{2})}\right)^{-1},\ \left(1+\frac{N-X}{(X+1)F_{2(X+1),2(N-X)}(1-\frac{\alpha}{2})}\right)^{-1}\right], \quad (3)$$

where $F_{df_1,df_2}(c)$ is the c **quantile** from the ***F*** **distribution** with df_1 and df_2 degrees of freedom. When $X = 0$ or N, the undefined bounds in (3) are respectively replaced by 0 and 1.

"Exact" in reference to the Clopper–Pearson interval refers to use of the exact binomial sampling distribution rather than using an asymptotic approximation to produce the interval. (The relevant exact binomial sums implicitly determine (3) through their relationship, and that of the F distributions, to the incomplete beta function.) This method, however, does not produce an exactly 100(1 − α)% interval, but rather one of at least 100(1 − α)% and sometimes much higher coverage. Thus, the price of guaranteeing at least 100 × (1 − α) coverage for each combination of N and p is loss of precision, in the sense that intervals are on average wider than necessary to achieve that coverage for most N, p combinations. Nevertheless, when preservation of nominal coverage is preferred despite this conservatism, the Clopper–Pearson interval accomplishes this objective and is widely used.

Other exact methods, for example, Blyth–Still [18], the Blaker [17] interval nested within it, and Blyth–Still–Casella intervals [21], are also available in specialized software. The **continuity correction** to the Wilson interval results in a wider, more conservative interval that better approximates the Clopper–Pearson interval. This frequently increases minimum coverage, as in Figure 1, to the nominal 100(1 − α)% [31, 38].

Coverage Comparison

For a fixed sample size of $N = 100$, and 101 possible values of the true p (0, 0.01, 0.02, . . . ,1.00), Figure 1 shows the true coverages of 95% Wald, Jeffreys, Wilson, continuity-corrected Wilson, Agresti–Coull, and Clopper–Pearson (*aka* Exact) intervals. Ideally, the coverage of each would be 95% for every value of p. While the discreteness of the binomial distribution prohibits this, one still desires coverage near 95%.

It is clear from Figure 1 that the Wald interval, even when $N = 100$, offers poor coverage. The Jeffreys interval offers better coverage properties than the interval obtained using a Uniform prior (not shown). Also, Ghosh [37] shows that the Wald interval is not only centered at the wrong place, but is also frequently wider than the Wilson interval.

Figure 2 shows interval widths for $N = 100$ and X from 1 to 50 (the plot is symmetric around 50). The Clopper–Pearson intervals are clearly wider for most values of X, and therefore for most values of p. The Exact approach produces the widest intervals except when p is near 0 or 1, when the Agresti–Coull intervals are wider. The Wilson and Agresti–Coull methods produce similar widths for p not near 0 or 1. Figure 2 also demonstrates that the Jeffreys interval is desirably narrow compared to other intervals (Figure 2) while producing coverages nearly 100(1 − α)% throughout the range of p (Figure 1).

There is philosophical debate over the relative merits of requiring *at least* 95% coverage for any value of p, as offered by exact methods or continuity

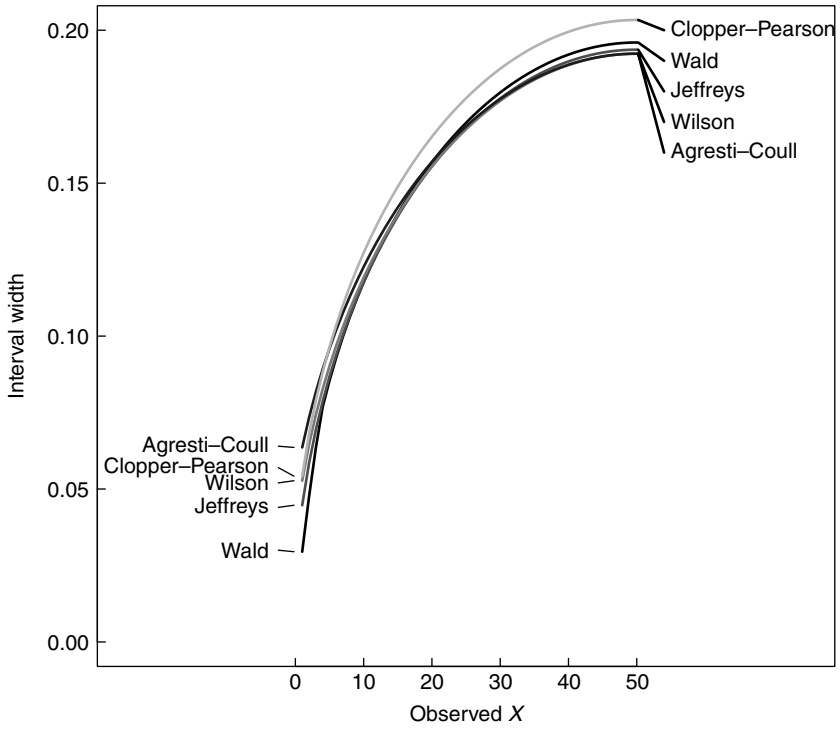

Figure 2 Widths of five 95% confidence intervals for p, $N = 100$

corrections, versus requiring average 95% coverage for all situations in which one might compute an interval. The practicing statistician must weigh the benefits and costs of each of the two competing approaches, and choose the most appropriate method for the given situation.

If guaranteed coverage is required, then the exact Clopper–Pearson interval is preferable. If average nominal coverage is satisfactory, then the Jeffreys, Wilson, and Agresti–Coull intervals offer sound frequentist properties. In the example shown, the Jeffreys interval offers tightest oscillation around 95%; this may differ, however, for other choices of $N \neq 100$. The Agresti–Coull interval may be the best compromise choice. It improves upon the Jeffreys and Wilson intervals by ensuring that coverage is not far below 100(1-α)% for values of p near 0 or 1. But it is not overly conservative, as are the continuity-corrected Wilson and the exact intervals, throughout the rest of the admissible range of p. The Agresti–Coull interval also offers the advantage of a form that is easy to remember and teach: for 95% confidence, just construct the simple Wald interval after adding two successes and two failures to the data. Various other references, for example, [1, 6, 20, 42, 53], offer similar graphical comparisons of the available choices of confidence intervals for a binomial proportion.

Hypothesis Testing

The score and Wald tests, respectively inverting the Wilson and Wald intervals, are commonly used, as is the likelihood ratio test [2]. The score statistic is equivalently Pearson's **goodness of fit** chi-square $X^2_{S1} = \sum_{i=1}^{2}(O_i - E_i)^2/E_i = (\hat{p} - p_o)^2/(p_o(1 - p_o)/N)$ (*see* **Chi-square Tests**), and the Wald statistic is Neyman's [44] modified chi-square statistic $X^2_{W1} = \sum_{i=1}^{2}(O_i - E_i)^2/O_i = (\hat{p} - p_o)^2/(\hat{p}(1 - \hat{p})/N)$, where the O_i and E_i, $i = 1, 2$ are respectively the observed counts of successes and failures and their **expectations** under H_0 : $p = p_0$. The corresponding forms of the **likelihood ratio** statistic are $X^2_{L1} = 2\sum_{i=1}^{2} O_i \log O_i/E_i = 2\,[X \log(\hat{p}/p_0) + (N - X)\,\log((1 - \hat{p})/(1 - p_0))]$.

Under H_0, all three statistics are asymptotically chi-square with one degree of freedom. As suggested by their denominator variances and the properties of their associated confidence intervals, convergence to this distribution is generally more rapid for X^2_{S1} than for X^2_{W1}, with departures for X^2_{W1} tending towards higher than nominal type I error rates (*see* **Hypothesis Testing**). The behavior of X^2_{L1} is intermediate [48].

The exact test, dual to the Clopper–Pearson interval, is easy to calculate (*see* **Exact Inference for Categorical Data**). For given p_0 and N, the binomial probability of observing $X = 0$ to $X = N$ under the null hypothesis is simply $\Pr(X = x) = \binom{N}{X} p_0^x (1 - p_0)^{N-x}$. Calculating the ***P* value**, the sum of probabilities of the observed and equally or less probable nonobserved outcomes, is straightforward.

The mid-P value, a 1961 innovation of Lancaster [40], has recently received renewed attention [2]. The mid-P value is the exact P value, as described above, less half the probability of the observed count. The mid-P value removes unattractive discrepancies between the properties of P values from discrete and continuous sampling distributions. Specifically, unlike conventional exact P values, the sum of mid-P values from two opposing one-sided exact tests equals 1.0, the mean of mid-P values is 0.5 under the null hypothesis, and the null distribution of mid-P values is closer to uniform than that of P values. However, while generally conservative, tests based on the mid-P value do not guarantee a type I error rate of less than α, nor do the corresponding confidence limits assure at least $100(1 - \alpha)\%$ coverage.

Hypothesis tests corresponding to the Jeffreys, Agresti–Coull, or another confidence interval for a binomial proportion can be performed by rejecting exactly when the interval excludes the hypothesized value. Such tests may well have type I error closer to nominal than one or more of the Wald, score, and likelihood ratio tests. To find the P value for a test formed by inverting a confidence interval, determine the confidence coefficient CC of the interval with p_0 on the boundary. Then, the P value is $1 - CC$.

For example, assuming $X = 38$ successes are observed in $N = 100$ trials with null hypothesis $p = 0.5$, the Jeffreys hypothesis test can be performed by determining which confidence coefficient provides an upper bound exactly equal to 0.5. A region of the Beta(38.5, 62.5) distribution bounded above at 0.5, and excluding equal probabilities on each tail, contains 98.40% of the distribution; hence $P = 0.0160$. Likewise, using the equation for the upper bound of the Agresti–Coull confidence interval and solving for α yields $P = 0.0165$.

Power and Sample Size Determination

Whether the endpoints of an interval or the rejection region of a test are based on exact binomial calculations or a large-sample approximation to the relevant binomial distribution(s), coverage of an interval or **power** of a test may be calculated either exactly using binomial distributions under the alternative, or approximately using a limiting normal distribution. Only direct calculation from the binomial distribution under an alternative gives true coverage or power, although Gaussian approximations to such calculations are generally used and often provide sufficient accuracy for practical purposes.

For example, the power of a two-tailed test of $H_0 : p = p_0$ under the fixed alternative of $H_A : p = p_1$ can be approximated as follows. First, find the acceptance region of the stipulated test, (T_L, T_U); for example, for X^2_{S1}, $X\varepsilon(T_L, T_U)$ with $(T_L, T_U) = Np_0 \mp z_{1-(\alpha/2)}\sqrt{(Np_0(1 - p_0))}$. Then, using normal theory under H_1, calculate the large-sample normal approximation to the probability that X falls outside that region, for example,

$$\text{Power} = \Pr\left(Z \le \frac{T_L - Np_1}{\sqrt{Np_1(1 - p_1)}}\right) + \Pr\left(Z \ge \frac{T_U - Np_1}{\sqrt{Np_1(1 - p_1)}}\right). \quad (4)$$

This approximates the true power, $\Pr(X \notin (T_L, T_U) | p = p_1)$ under $X \sim \text{Bin}(N, p_1)$, which may be calculated instead by summing the $\text{Bin}(N, p_1)$ probabilities for all values of X outside (T_L, T_U). This approach applies, with obvious modifications, to any other test procedure, specifically including exact tests and tests using the mid-P. Power, and the method for determining it, are based on the test's rejection region, not on how that rejection region is derived. Note that standard moment-based expressions in textbooks, and default power calculations in statistical software packages, are almost always asymptotic approximations to the true power of a test. The ease of inverting the moment-based expressions to yield

approximate sample size requirements has much to do with this. However, discrepancies in default sample size recommendations of software packages are common, owing to variations in defaults on the tests used and the specific power calculations inverted to obtain them (*see* **Sample Size Determination**).

While the asymptotically based counterparts to an exact test are generally more powerful, that is, type II error probabilities β for the asymptotic tests are lower than for the exact test, this gain comes at the expense of higher type I error rates, which are not guaranteed by the asymptotic tests. As shown in Figure 1 (with $\alpha = 1 -$ coverage probability), these may far exceed the nominal value used to (asymptotically) determine the rejection region. For fixed α, power as a function of sample size is also saw-toothed: counterintuitively, a small increase in sample size may slightly reduce power.

One-sample Summary

Numerous superior alternatives to the classic Wald interval exist. The Clopper–Pearson and, at least for 95% confidence, the continuity-corrected Wilson intervals, assure that coverage cannot fall below the nominal coefficient for any combination of N and p. Similarly, the tests based on inverting these intervals assure that type I error cannot exceed the nominal α. The cost is wider intervals and reduced power relative to procedures whose coverage roughly centers around, rather than below, the nominal confidence coefficient. Among such procedures are the relatively simple Jeffreys, Wilson, and Agresti–Coull intervals, the latter two with closed-form expressions, and their corresponding tests. The Agresti–Coull method produces intervals with generally sound frequentist properties, in a form easily taught and remembered.

For research design, software for sample size determination should be used cautiously. For hypothesis testing, for instance, such software typically requires specification of p_0 as well as a nominal α and a target or guess at p_1. Then, for fixed desired power, a single or a selection of sample size recommendations is provided from among the available choices. The researcher must recognize that power and sample size results for alternative tests can differ not only because one test is more efficient, but also because two tests of nominal size α may have different actual type I error rates, and/or because of approximations used in the calculations. This is particularly so since the power functions of the several tests, and even their relative performance, may be nonmonotonic with parameters and/or sample size in neighborhoods of their hypothesized or recommended values.

Two Independent Samples

Inference for two independent samples focuses on how much the relative frequency of an observed characteristic differs between two sampled populations. In general, the lessons of the one-sample case regarding (a) the liberality of the standard Wald procedure, (b) conservatism of the standard exact procedure, (c) availability of simple intermediate approaches that achieve closer to nominal type I error by sacrificing control of maximum type I error, and (d) the inherent trade-offs of coverage and power with different levels of type I error control, all continue to apply. However, the two-sample situation is more complex because (i) the **null hypothesis** of interest is typically the composite hypothesis of no difference between populations, with the common underlying proportion a **nuisance parameter**, and (ii) the disparity between populations may be parameterized in several ways, most commonly as the difference ("risk difference") [2–4, 8, 11, 43, 45, 49], ratio ("risk ratio") [2, 4, 12, 45, 49], and **odds ratio** [2, 3, 54], that are functionally dependent only for a fixed value of the nuisance parameter.

A consequence of (ii) is that there is no longer a one-to-one relationship of hypothesis tests to confidence intervals for any single parameter determined to be of primary interest, and the details of interval estimation vary depending upon the association parameter chosen. To simplify exposition, and in conformity with the historical development, the ensuing discussion will thus proceed primarily from a testing perspective, with the reader referred to the excellent reviews [1, 3] for additional detail on confidence intervals.

Comparing two binomial proportions has long occupied the field of statistics. In 1900, Karl **Pearson** introduced what became the "standard" chi-square test as a goodness-of-fit test to determine whether observed data were compatible with a proposed probability model [47]. Its proper application to contingency tables was clarified in 1922 by **Fisher** [33].

Hundreds of papers have since offered extensions, improvements, and adjustments to the test, which *Science 84*, a popular magazine of the American Association for the Advancement of Science (AAAS), called one of the 22 most important scientific breakthroughs of the twentieth century [7].

For the remainder of this section, we consider a **two-by-two** (2×2) table of observed counts under the probability model $n_{i1}|n_{i\cdot} \sim \text{Bin}(n_{i\cdot}, p_i)$,

	Response		
	Present	Absent	Total
Population 1	n_{11}	n_{12}	$n_{1\cdot}$
Population 2	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot} = N$

$i = 1, 2$. Such data, with this model, may arise from several experimental or observational research designs. The row totals $n_{1\cdot}$ and $n_{2\cdot}$ may be fixed by the conditions of a designed observational study or experiment or, in an observational study, only N may have been determined by the researcher, or $n_{11}, n_{12}, n_{21}, n_{22}$ may be independent **Poisson** counts. In these latter cases, the within-row binomial distributions arise by conditioning inference on the observed row totals $n_{1\cdot}, n_{2\cdot}$ of a multinomial or product-Poisson distribution, respectively. For such situations, we will discuss a general class of asymptotic procedures, exact inference, and Bayesian inference.

Asymptotic Methods

Read and Cressie [29, 48] defined the class of **power-divergence** asymptotic test statistics $T^{\lambda}(N, \widehat{m}_{ij})$ which, for $H_0 : p_1 = p_2$ as above, take the form

$$T^{\lambda}(N, \widehat{m}_{ij}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\left[\left(\frac{n_{ij}}{\widehat{m}_{ij}}\right)^{\lambda} - 1\right] \tag{5}$$

for $\lambda \neq -1, 0$, and the limiting forms $T^{-1}(N, \widehat{m}_{ij}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} \widehat{m}_{ij} \log(\widehat{m}_{ij}/n_{ij})$, $T^{0}(N, \widehat{m}_{ij}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \log(n_{ij}/\widehat{m}_{ij})$. The $\widehat{m}_{ij}$ are estimated expected values of the n_{ij} obtained by minimizing $T^{\zeta}(N, m_{ij})$, for some ζ (not necessarily equal to λ), under the constraint $p_1 = p_2$. In our notation, we suppress ζ, which does not affect the asymptotic distribution of the test statistics. This family includes the likelihood ratio test ($\lambda = \zeta = 0$), Pearson's chi-square ($\lambda = 1, \zeta = 0$), Neyman's minimum modified chi-square ($\lambda = \zeta = -2$), and others that may be conveniently studied within this unifying framework (*see* **Chi-square Tests**).

Under all of the experimental or observational designs given above, when $p_1 = p_2$ each member of the power-divergence family converges in distribution to chi-square with one degree of freedom as $n_{1\cdot}, n_{2\cdot} \to \infty$, and hence provides an asymptotically valid test of H_0 or, equivalently when $n_{1\cdot}$ and $n_{2\cdot}$ are random, of row by column independence.

Pearson's chi-square, which is the score test as in the one-sample case, is commonly written in each of the several forms

$$X^2_{S2} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \widehat{m}_{ij})^2}{\widehat{m}_{ij}}$$
$$= \frac{(\hat{p}_1 - \hat{p}_2)^2}{(n_{1\cdot}^{-1} + n_{2\cdot}^{-1})\hat{p}(1-\hat{p})} = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} \tag{6}$$

with $\widehat{m}_{ij} = n_{i\cdot}n_{\cdot j}$, $\hat{p}_1 = n_{11}/n_{1\cdot}$, $\hat{p}_2 = n_{21}/n_{2\cdot}$, $\hat{p} = n_{\cdot 1}/N$. In the third form it is easily extended to the score test for $H_{0\Delta} : p_1 - p_2 = \Delta$,

$$X^2_{S2\Delta} = \frac{((\hat{p}_1 - \hat{p}_2) - \Delta)^2}{(n_{1\cdot}^{-1} + n_{2\cdot}^{-1})\hat{p}(1-\hat{p})}. \tag{7}$$

The set of all Δ not rejected by this test forms an asymptotic confidence interval for $p_1 - p_2$ analogous to the Wilson interval in the one-sample case.

Neyman's minimum modified chi-square [44], which as earlier is the Wald statistic, replaces the denominator $E_{ij} = \widehat{m}_{ij}$ above with n_{ij}, yielding

$$X^2_{W2} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{O_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \widehat{m}_{ij})^2}{n_{ij}}$$
$$= \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}_1(1-\hat{p}_1)/n_{1\cdot} + \hat{p}_2(1-\hat{p}_2)/n_{2\cdot}}. \tag{8}$$

The set of all Δ not rejected by the corresponding Wald test of $H_{0\Delta}$ using

$$X^2_{W2\Delta} = \frac{((\hat{p}_1 - \hat{p}_2) - \Delta)^2}{\hat{p}_1(1-\hat{p}_1)/n_{1\cdot} + \hat{p}_2(1-\hat{p}_2)/n_{2\cdot}} \tag{9}$$

may be written directly as $(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{(\hat{p}_1(1-\hat{p}_1)/n_{1\cdot} + \hat{p}_2(1-\hat{p}_2)/n_{2\cdot})}$. This is the confidence interval traditionally presented in elementary statistics courses and texts, and most commonly used in practice. Unfortunately, this shares the propensity of the one-sample Wald interval to be too narrow to achieve nominal coverage. The performance of the same interval, however, substituting $\widetilde{p}_i = (n_{i1} + 1)/(n_{i\cdot} + 2)$ for $\hat{p}_i$, is much improved [3].

As also noted by Agresti [3], the Wald interval for the log odds ratio using the empirical asymptotic variance $\left(\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}^{-1}\right)$ derived by the **delta method**,

$$\log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) \pm 1.96\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}^{-1}} \qquad (10)$$

generally performs well, and its performance is improved if the n_{ij} are respectively replaced by $n_{ij} + 2n_{i\cdot}n_{\cdot j}/N^2$ and if, in addition, the intervals are extended to $\pm\infty$ respectively whenever $\min(n_{12}, n_{21}) = 0$, $\min(n_{11}, n_{22}) = 0$. Further, inverting an exact or asymptotic chi-square test of $H_0 : p_1/p_2 = \theta$ based on the score statistic

$$X_{S\theta} = \frac{n_{1\cdot}(\hat{p}_1 - \tilde{\tilde{p}}_1)^2}{\tilde{\tilde{p}}_1(1-\tilde{\tilde{p}}_1)} + \frac{n_{2\cdot}(\hat{p}_2 - \tilde{\tilde{p}}_2)^2}{\tilde{\tilde{p}}_2(1-\tilde{\tilde{p}}_2)}, \qquad (11)$$

where $\tilde{\tilde{p}}_i$ is the MLE of p_i under H_0, provides a generally well-behaved interval estimate for the risk ratio p_1/p_2.

The likelihood ratio statistic $X_{L2}^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log[n_{ij}/(n_{i\cdot}n_{\cdot j})]$ is also a commonly used power-divergence statistic for comparing two proportions. Although the various power-divergence statistics share the same limiting null chi-square distribution as both $n_{1\cdot}$ and $n_{2\cdot}$ increase, and the tests have the same **Pitman efficiency** under local alternatives in the nonnull case, they differ in performance under nonlocal alternatives (e.g. in Bahadur efficiency), in more general problems under nonstandard "sparse" asymptotics (in which the number of cells increases with N), and in small samples (*see* **Asymptotic Relative Efficiency (ARE)**). Cox and Groeneveld [28] provide a thorough comparison of X_{S2}^2 and X_{L2}^2, predicting when each will provide a higher, that is, more powerful, test statistic, under various null hypotheses. Read and Cressie [48] recommend $\lambda = 2/3$ as the best compromise between high power under a wide range of true alternative hypotheses and the ability of the chi-square distribution to approximate the distribution of the test statistic under the null hypothesis for small samples. Of the power-divergence methods in general practical use, the Pearson chi-square, with $\lambda = 1$, is closest to this.

Continuity Corrections

Owing to the discreteness of counts, the sampling distributions of power-divergence statistics from contingency tables are discrete, that is cdfs are step functions, and cannot generally be well-approximated by the continuous χ_1^2 distribution when $n_{1\cdot}$ or $n_{2\cdot}$ is small. In such circumstances, the actual type I error rate of a test may be well above or below the desired α level. Continuity corrections are modifications to the test statistic or the approximating distribution for the purpose of reducing or minimizing the effects of such approximation errors. Such a continuity correction can be used, for instance, to control excessive type I error of an asymptotic chi-square test by shrinking the test statistic, thus making the test more conservative. The classic continuity correction does this by replacing $(O - E)^2$ by $(|O - E| - (1/2))^2$ in the numerator of X_{S1}^2.

Continuity corrections are thoroughly covered in this volume and elsewhere [1, 2, 38, 43]. They are generally constructed to better approximate the behavior of exact methods for which statistical software is increasingly available. Thus, their current utility is primarily for situations when maintaining the nominal α is crucial and statistical software for exact testing is not handy.

Exact Methods

Statistical inferences using exact methods rely on computations from one or more completely specified discrete probability laws. A disadvantage is that discreteness makes it impossible to perform tests with a precisely predetermined type I error rate without using a supplemental **randomization** to decide some test results, a process most consider unsatisfactory for scientific discourse. While the exact test and interval are straightforward in the one-sample case, with two samples the null hypothesis is composite: rather than a specified p_0, under H_0: $p_1 = p_2 = p$,

p can assume any value in [0,1]. A simplification strategy is required to select from or combine over this universe of distributions compatible with H_0.

Fisher's Exact Test

R.A. Fisher proposed conditioning on fixed row and column marginal totals (*see* **Conditional Probability**; **Fisher's Exact Test**). This restriction on the sample space allows direct numerical calculation of the distribution of any test statistic from the 2×2 table. For instance, the distribution of any single cell count is **hypergeometric**, as in

$$\Pr(n_{11} = t) = \frac{\binom{n_{1\cdot}}{t}\binom{n_{2\cdot}}{n_{\cdot 1} - t}}{\binom{N}{n_{\cdot 1}}}, \qquad (12)$$

where t ranges from $\max(0, n_{1\cdot} + n_{\cdot 1} - N)$ to $\min(n_{1\cdot}, n_{\cdot 1})$. One can calculate the probability of each possible t, and then a P value by summing the probabilities of n_{11} and all t more compatible than n_{11} with the alternative hypothesis.

The highly constrained hypergeometric setting may lead to very few possible values of n_{11}, and hence very few possible tables and P values. In the tea tasting example through which Fisher introduced the test, $N = 8$ with $n_{1\cdot} = n_{\cdot 1} = n_{2\cdot} = n_{\cdot 2} = 4$. Since $0 \le n_{11} \le 4$, there are five possible P values, respectively, 0.014, 0.24, 0.76, 0.99, and 1.0 for $n_{11} = 0, \ldots, 4$ for his one-sided test, and three possible P values (0.029, 0.48, 1.0) for the two-sided test. Thus, for instance, the common *true* type I error rate of all one-sided tests with *desired* (nominal) type I error rates between 2 and 23% is actually 1.4%. This inherent conservatism, at times extreme, can be removed by supplemental randomization to precisely achieve any nominal α; Tocher [52] has shown such randomized tests to be uniformly most powerful among unbiased tests. The mathematical optimality of randomized tests has not, however, overcome the taint of arbitrariness that restricts their use.

Consequently, a variety of alternative nonrandomized exact methods have been developed that reduce the discreteness and hence conservatism of Fisher's approach. For instance, Agresti [3] provides a thorough and readable review of confidence interval options for the difference, ratio, and odds ratio of two binomial proportions that guarantee nominal coverage. See [6, 26, 50] for details of some intervals with good properties.

Using mid-P values with Fisher's exact test has gained considerable recent support. As in the one-sample case, this method does not guarantee preservation of the nominal level α, but performance of a test using mid-P is generally closer to nominal than that of the corresponding exact test, and power is inherently higher [1].

Unconditional Exact Tests

It is more common in experiments, and always the case in **observational** research, for at least one tabular margin to be random. Conditioning on only one set of margins, say $\{n_{1\cdot}, n_{2\cdot}\}$, produces a much larger sample space than conditioning on two, allowing many more possible tables and P values under H_0, and thus tests frequently yielding closer to the nominal type I error rates. However, the indeterminate nuisance parameter $p_1 = p_2 = p$ must still be removed. This problem may be overcome, in the context of an arbitrary test statistic T such as $T = X^2_{S2}$, by maximizing the exact P value over possible values of the nuisance parameter:

$$P \text{ value} = \sup_{0 \le p \le 1} \Pr_p(T \ge t_o | n_{1\cdot}, n_{2\cdot}), \qquad (13)$$

where t_0 is the observed test statistic [9, 10, 19].

This "unconditional exact" method has been criticized precisely because it maximizes over the full range of p; Fisher [35] and others [51] have argued that possible samples with far different total successes than were observed are irrelevant. In response, Berger and Boos [15] proposed maximizing the P value over a $100(1 - \beta)\%$ confidence set C_β for p, and adjusting the result for the restriction to the confidence set. Adding β to the maximum over the confidence set yields a valid P value, namely,

$$P \text{ value} = \sup_{p \in C_\beta} \Pr_p(T \ge t_o | n_{1\cdot}, n_{2\cdot}) + \beta. \qquad (14)$$

Since the approach becomes more conservative (P values increase) as the confidence interval narrows, these authors favor a high confidence coefficient, for example, $100(1 - \beta)\% = 99.9\%$ or $\beta = 0.001$, and hence a wide interval. This modification of the approach of Boschloo [19] also maintains the guaranteed level α and is "often uniformly more powerful than other tests" including Fisher's exact test [14]. At the time of

publication, this test and its associated confidence intervals were not widely commercially available. However, both have been implemented in **StatXact** Version 6 [30], and modified Boschloo P values are obtainable using software available from Berger at `http://www4.stat.ncsu.edu/~berger/software.html`. Several alternative unconditional exact intervals that also guarantee coverage have been proposed by Chan & Zhang [24].

Bayesian Methods

The reader is referred to [39] for a thorough review of Bayesian inference for 2×2 tables based on a variety of noninformative, subjective, and correlated priors; *see also* **Bayesian Methods for Contingency Tables**.

At the price of introducing a continuous subjective prior distribution for the two unknown parameters p_1 and p_2, the choice of which is open to debate, the Bayesian statistician bypasses many frequentist difficulties due to discreteness. In a hypothesis-testing framework, the frequentist's final result is a P value – the probability of the observed data and data more compatible with the alternative, conditional on a fixed null hypothesis. Owing to discreteness of the sample space, the discrete distribution of a test statistic may yield few choices of achievable type I error rates for a nonrandomized exact test, and may be poorly approximated by the continuous chi-square or other asymptotic distribution. The Bayesian, however, conditions upon the observed data – all of it rather than selected margins – and calculates the posterior probability of a particular hypothesis of interest: $\Pr(p_1 > p_2|\{n_{ij}\})$, $\Pr(p_1 < p_2|\{n_{ij}\})$, $\Pr(|p_1 - p_2| < \varepsilon|\{n_{ij}\})$, and so on [39]. Such probabilities are determined by integrating over the appropriate space in the joint posterior distribution of (p_1, p_2), which will be continuous whenever a continuous prior is selected. Circumstances are uncommon in which a discrete bivariate prior would be reasonable for the two unknown proportions.

A more general advantage of Bayesian methods is that they satisfy the **likelihood principle**, which asserts that inference about a parameter should depend only on the relative values of the likelihood function at the parameter's admissible values and not otherwise on the data collection method. As Berger and Berry clearly illustrate [13], frequentist hypothesis testing incorporates the probability of outcomes that never occur and therefore, two different research designs that yield the exact same data may provide different P values and hence, different inferences. Bayesian methods do not exhibit this somewhat counterintuitive behavior because, no matter what the research design, they condition on all of the data rather than on a particular choice of marginal totals [41]. In a Bayesian analysis, discreteness is still manifest in the distribution of the posterior probability of a specific hypothesis, considered as a random variable. The posterior can realize only as many values as there are possible tables under the study design, but this does not produce the interpretive complications presented by the frequentist context.

As an example, when independent Jeffreys priors are placed on p_1 and p_2, $\Pr(p_1 > p_2)$ is closely approximated by $\Phi^{-1}(z)$, where

$$z = (n_{11}n_{22} - n_{12}n_{21})\sqrt{\frac{n_{11} + n_{12} + n_{21} + n_{22}}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}. \tag{15}$$

This corresponds to the one-sided P value from X^2_{S2}. Numerical methods such as Gibbs sampling and **Markov Chain Monte Carlo** (MCMC) can be used to improve this approximation [23, 36]. Howard also considers the case of correlated priors for p_1 and p_2. Such a choice, with positive correlation, is one way to represent the subjective belief that p_1 and p_2 are unequal but not likely to differ substantially.

Study Design and Power

For **Student's *t*-test** and other tests based on continuous distributions, it is possible to choose a rejection region that guarantees precisely an arbitrary prespecified type I error rate, whatever the common value of the mean. This is not possible when comparing two proportions, regardless of the test used, without very large sample sizes or use of a supplementary randomization resisted by the scientific community. Consequently, the power of a test at a nominal α depends on the sample sizes in each group, the values of each of the two probabilities, and the true type I error rate achieved by the test in that combination of circumstances. The power is relatively increased when true type I error overshoots the nominal α, as for instance, with the Wald test based on the observed difference $\hat{p}_1 - \hat{p}_2$ in small to moderate samples, and relatively decreased when true type

I error falls below nominal α, as with Fisher's exact test in most of its uses. When designing a study, such trade offs should be kept in mind. The various widely published asymptotic formulae for approximate sample sizes or power for the chi-square, likelihood ratio, or Fisher's exact tests generally fail to capture the saw-tooth nature of this sample size/power trade off. As in the single proportion case, small increases in sample size may reduce power.

As a general rule, efficient use of data is promoted by the use of methods for which actual type I error rates are close to nominal. Methods based on power-divergence statistics and associated confidence intervals, slightly modified if needed to improve coverage, are generally useful for moderate to large samples, and for smaller samples when strict control of type I error is not essential. The power of any test against any fixed distribution in the space of alternatives is defined as the probability of the rejection region under the alternative distribution. This may generally be computed more accurately by exact calculations or **Monte Carlo methods** than by asymptotic formulae. Sample size may be chosen indirectly by this method which, although cumbersome, is reliable and avoids ambiguities associated with the use of asymptotic approximations in general purpose statistical software. Most available sample size/statistical power software packages calculate the power for Pearson, likelihood ratio, and Fisher exact tests [46], but may default to approximations.

When the true type I error rate must not be allowed to exceed the nominal α to any degree, as frequently occurs in the context of regulatory decision making, Fisher's exact test has been the conventional choice. However, unconditional exact methods generally offer more power, and under randomization, the unconditional framework for inference is no less valid than the conditional. The modified Boschloo unconditional test is usually more powerful than the original Boschloo test, which is uniformly more powerful than Fisher's exact test [14]. Similarly, inversion of the modified Boschloo test produces a narrower unconditional exact confidence interval for $p_1 - p_2$, while guaranteeing at least nominal coverage [15, 30].

We reiterate that commercially available sample-size software uses a variety of formulae and approximations that are not always well documented, and may or may not incorporate continuity corrections by default.

Two-sample Summary

The Pearson chi-square test is powerful and offers nearly α type I error rates for large samples. For smaller studies, however, test statistics are desirable that either do not require asymptotic assumptions or are adjusted to improve performance. The modified Boschloo unconditional exact test is a more powerful alternative to Fisher's exact test that strictly preserves nominal type I error when study design leaves at least one tabular margin random. If both margins are inherently fixed and preservation of nominal type I error is essential, then Fisher's exact test, with its inherent conservatism, is warranted. Bayesian tests that escape some discreteness problems and produce straightforward interpretations may also provide insight.

Confidence intervals for differences of proportions and ratios of proportions or odds are generally available by inverting hypothesis tests. These methods are discussed in a variety of texts and manuscripts [2–4, 8, 11, 12, 43, 45, 49, 54]. Bedrick provides thorough coverage of confidence intervals for ratios of two proportions within the power-divergence family of statistics. He concludes that $0.67 \leq \lambda \leq 1.25$ give intervals with the best coverage. This range includes the intervals based on the "Cressie–Read statistic" with $\lambda = 2/3$, and on the Pearson chi-square test.

For small samples, when strict preservation of coverage is not essential, confidence intervals can be constructed by inverting hypothesis tests based on mid-P values [16] to substitute for the overly-broad exact intervals. Closed forms do not exist but software such as Cytel's StatXact [30] computes these intervals.

For simple problems of inference about one and two proportions, research and expanded computing power have clarified deficiencies in methods that have formed the basis of statistical pedagogy and most scientific practice. Although improved methods have been identified, and their properties are generally well-understood, they have not yet seen widespread application. A key requirement for the acceptance of any modern statistical methodology is convenient availability in software. While many of the newer techniques discussed here are implemented in special purpose commercial software and/or can be readily programmed using SAS/IML, S-PLUS, or the freeware **R**, as of March 2004, we know of none employed as defaults, and support by the general purpose statistical packages used by most data analysts

is inconsistent. More frequent use of the improved methods, and hence better inferences for these scientifically ubiquitous situations, await more widespread and convenient implementation by the mass market software packages.

References

[1] Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies, *Statistics in Medicine* **20**, 2709–2722.

[2] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. Wiley, New York.

[3] Agresti, A. (2003). Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact, *Statistical Methods in Medical Research* **12**, 3–21.

[4] Agresti, A. & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions results from adding two successes and two failures, *The American Statistician* **54**, 280–288.

[5] Agresti, A. & Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions, *American Statistician* **52**, 119–126.

[6] Agresti, A. & Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions, *Biometrics* **57**, 963–971.

[7] American Association for the Advancement of Science. (1984). Science 84, Washington, November.

[8] Anbar, D. (1983). On estimating the difference between two probabilities, with special reference to clinical trials, *Biometrics* **39**, 257–262.

[9] Barnard, G.A. (1945). A new test for 2×2 tables, *Nature* **156**, 177.

[10] Barnard, G.A. (1947). Significance tests for 2×2 tables, *Biometrika* **34**, 123–138.

[11] Beal, S.L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples, *Biometrics* **43**, 941–950.

[12] Bedrick, E.J. (1987). A family of confidence intervals for the ratio of two binomial proportions, *Biometrics* **43**, 993–998.

[13] Berger, J.O. & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity, *American Scientist* **76**, 159–165.

[14] Berger, R.L. (1994). *Power Comparison of Exact Unconditional Tests for Comparing Two Binomial Proportions*, Institute of Statistics Mimeo Series No. 2266.

[15] Berger, R.L. & Boos, D.D. (1994). P values maximized over a confidence set for the nuisance parameter, *Journal of the American Statistical Association* **89**, 1012–1016.

[16] Berry, G. & Armitage, P. (1995). Mid-P confidence intervals: A brief review, *The Statistician* **44**, 417.

[17] Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions, *Canadian Journal of Statistics* **28**, 793–798.

[18] Blyth, C.R. & Still, H.A. (1983). Binomial confidence intervals, *Journal of the American Statistical Association* **78**, 108–116.

[19] Boschloo, R.D. (1970). Raised conditional level of significance for the 2×2 table when testing the equality of two probabilities, *Statistica Neerlandica* **24**, 1–35.

[20] Brown, D.L., Cai, T. & DasGupta, A. (2001). Interval estimation for a binomial proportion, *Statistical Science* **16**, 101–133.

[21] Casella, G. (1986). Refining binomial confidence intervals, *Canadian Journal of Statistics* **14**, 113–129.

[22] Casella, G. & Berger, R.L. (1990). *Statistical Inference*. Duxbury Press, Belmont.

[23] Carlin, B.P. & Lewis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC Press, Baco Raton.

[24] Chan, I.S.F. & Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions, *Biometrics* **55**, 1202–1209.

[25] Clopper, C.J. & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial, *Biometrika* **26**, 404–413.

[26] Coe, P.R. & Tamhane, A.C. (1993). Small sample confidence intervals for the difference, ratio, and odds ratio of two success probabilities, *Communications in Statistics, Part B – Simulation and Computation* **22**, 925–938.

[27] Corcoran, C. & Mehta, C. (2001). Comment on "Interval estimation for a binomial proportion, *Statistical Science* **16**, 122–123.

[28] Cox, C.P. & Groeneveld, R.A. (1986). Analytic results on the difference between G^2 and χ^2 test statistics in one degree of freedom cases, *The Statistician* **35**, 417–420.

[29] Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

[30] Cytel Software Corporation. (2003). *StatXact Version 6 with Cytel Studio*™, Vol. 2, Cytel Software Corporation, Cambridge, pp. 527–530.

[31] D'Agostino, R.B. (1990). Comment on "Yates's correction for continuity and the analysis of 2×2 contingency tables", *Statistics in Medicine* **9**, 367.

[32] DeGroot, M.H. & Schervish, M.J. (2001). *Probability and Statistics*, 3rd Ed. Wiley, New York.

[33] Fisher, R.A. (1922). On the interpretation of chi-square from contingency tables, and the calculation of P, *Journal of the Royal Statistical Society* **85**, 87–94.

[34] Fisher, R.A. (1935). *Design of Experiments*. Oliver and Boyd, Edinburgh, Chapter 2.

[35] Fisher, R.A. (1945). A new test for 2×2 tables (Letter to Editor), *Nature* **156**, 388.

[36] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

[37] Ghosh, B.K. (1979). A comparison of some approximate confidence intervals for the binomial parameter, *Journal of the American Statistical Association* **74**, 894–900.

[38] Haviland, M.G. (1990). Yates's correction for continuity and the analysis of 2×2 contingency tables, *Statistics in Medicine* **9**, 363–365.

[39] Howard, J.V. (1998). The 2×2 table: A discussion from a Bayesian viewpoint, *Statistical Science* **13**, 351–367.

[40] Lancaster, H.O. (1961). Significance tests in discrete distributions, *Journal of the American Statistical Association* **56**, 223–234.

[41] Little, R.J.A. (1989). Testing the equality of two independent binomial proportions, *The American Statistician* **43**, 283–288.

[42] Newcombe, R.G. (1998a). Two-sided confidence intervals for the single proportion: Comparison of seven methods, *Statistics in Medicine* **17**, 857–872.

[43] Newcombe, R.G. (1998b). Interval estimation for the difference between two independent proportions: comparison of eleven methods, *Statistics in Medicine* **17**, 873–890.

[44] Neyman, J. (1949). Contribution to the theory of the χ^2 test, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.

[45] Nurminem, M. (1986). Confidence intervals for the ratio and difference of two binomial proportions, *Biometrics* **42**, 675–676.

[46] O'Brien, R. (1998). A Tour of UnifyPow: A SAS module/macro for sample-size analysis, *Proceedings of the 23rd Annual SAS Users Group International Conference*, SAS Institute Inc., Cary, pp. 1346–1355.

[47] Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine* **5**(50), 157–175.

[48] Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York.

[49] Santner, S.J. & Snell, M.A. (1980). Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables, *Journal of the American Statistical Association* **75**, 386–394.

[50] Santner, T.J. & Yamagami, S. (1993). Invariant small sample confidence-intervals for the difference of two success probabilities, *Communications in Statistics, Part B – Simulation and Computation* **22**, 33–59.

[51] Sprott, D.A. (2000). *Statistical Inference in Science*. Springer-Verlag, New York.

[52] Tocher, K.D. (1950). Extension of Neyman-Pearson theory of tests to discontinuous variates, *Biometrika* **37**, 130–144.

[53] Vollset, S.E. (1993). Confidence intervals for a binomial proportion, *Statistics in Medicine* **12**, 809–824.

[54] Walter, S.D. & Cook, R.J. (1991). A comparison of several point estimators of the odds ratio in a single 2×2 contingency table, *Biometrics* **47**, 795–811.

[55] Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association* **22**, 209–212.

(*See also* **Categorical Data Analysis**)

JASON T. CONNOR & PETER B. IMREY

Proprietary Biostatistical Firms

A "proprietary biostatistical firm" is generally a private company offering statistical services to the pharmaceutical, biotechnology, and/or medical device industries. It might be privately owned or a publicly traded company. The more generic term, Contract Research Organization (CRO), is often used to describe such companies, but will also be used to refer to companies providing a wider range of services in pharmaceutical research and development.

Introduction

The fact that certain substances can prevent and/or relieve human conditions has played an important role in the history of mankind, spanning the centuries and cultures of the world. Today's pharmaceutical medicines; that is to say, products that have been synthesized and developed rather than naturally occurring, had their origins in the emergence of chemical companies in the nineteenth century. However, it was not until the synthesis of antibacterial substances in the 1930s that the chemotherapeutic revolution started. The outbreak of war at the end of the 1930s provoked a surge of research for other anti-infective compounds, and the therapeutic potential of penicillin, identified in 1929, was more fully realized. The modern research-based pharmaceutical industry was born.

It has been since the 1980s, and still is effectively, the exclusive right of the pharmaceutical and the biotechnology industries to research, develop, and deliver health care products to the populations of the world. As the pharmaceutical industry grew, so did the interest in, and concern for, the safety and efficacy of new medicines, and in the growth in expenditure for the health of the public. When the thalidomide misadventure happened in the early 1960s, society, through national governments, was ready to act. Regulations controlling drug development and subsequent marketing practices were born (*see* **Drug Approval and Regulation**).

With regulatory requirements established and increasing gradually since the 1960s, the drug development process has become more complex, resulting in increased costs and time pressures on the research-based pharmaceutical industry. The stimulus for reducing health care costs, the usage of generic drugs, and the emergence of the biotechnology industry in the 1980s added to the changes in the pharmaceutical environment. The industry began to consolidate and in some cases reduce the size of their staff. The concept of developmental research using sources outside the industry became a reality and the CRO industry was born.

The Beginnings in the 1970s

At first, the pharmaceutical companies were looking for services to manage peak periods in selected areas. With regulation had come the requirement for appropriate designs and statistical analyses of **clinical trials**. Although the pharmaceutical industry had rapidly expanded their in-house statistical staff in the 1970s, the specialist area of biostatistics lent itself to being contracted out. Some of this early consulting work was provided by academia (biostatistics/ statistics departments, medical schools, or other publicly funded organizations); indeed, many of today's CROs started life as consultancy services within academic departments.

As personal computing became established and continuously more powerful in the 1980s, and when commercial statistical software became more fully developed and reliable (*see* **Software, Biostatistical**), CROs began to function and offer their statistical services to the international pharmaceutical industry. Some of these companies flourished briefly and disappeared. Others expanded with the increased demand for data handling capabilities in a controlled environment and established themselves as complete data operations, providing comprehensive biostatistical and clinical data management services. Some companies expanded further by adding clinical operations to their services, such as expert medical advice in the design of studies and plans for drug development, drug packaging and distribution, monitoring of studies and adverse events (*see* **Data and Safety Monitoring**), strategies for regulatory affairs, design and development of CANDAs (Computer Assisted New Drug Applications), **post-marketing surveillance** and reporting, and economic cost–benefit evaluations of pharmacologic drugs (*see* **Pharmacoepidemiology, Adverse and Beneficial Effects**). Biostatistical services included

designs, methodology, and **sample size determination for clinical trials**, plans for statistical monitoring of the conduct of clinical trials, and analyses of studies including therapeutic **outcome measures** and **quality of life**.

Several CROs have expanded further to meet the changing needs arising from consolidations, the downsizing of pharmaceutical companies, and the growing biotechnology industry by offering other nonclinical services (e.g. analytical chemistry, toxicology, pharmacology, formulation development, laboratory services, and manufacturing). While the CROs specializing in the biostatistical aspects of clinical trials still exist, pharmaceutical companies often contract out biostatistical and other services, such as those associated with complete clinical investigations, whether they be single or multiple studies or indeed entire drug development programs.

The Biostatistical Consulting Operation

As already indicated, much biostatistical contract work is being done as part of more extensive services. In this section we outline the job and the career structures for biostatisticians, whether they are employed by a CRO specializing in biostatistics (and **data management**) or a CRO which offers biostatistics as one of their services.

The Career

The growth in biostatistical consultation and support of new drug development in the pharmaceutical industry resulted in an increased number of biostatisticians being employed during the 1970s and early 1980s. As the CROs grew in the 1980s and 1990s, they employed many biostatisticians, as may be noted by the membership in pharmaceutical statistics organizations such as The European Federation of Statisticians in the Pharmaceutical Industry (EFSPI and its national organizations) and the Biopharmaceutical Section of the **American Statistical Association**. For example, at the end of 1996, the UK national organization of EFSPI, PSI (**Statisticians in the Pharmaceutical Industry**), counted more than half of its 800 members as belonging to the contract research industry.

A career as a biostatistician in a CRO is not just a possibility but is a reality. Depending on the size of the organization and the variety of its services, as well as on the individual and his/her strengths and aspirations, careers in biostatistics span the full spectrum of both a technical and a managerial ladder. Professional and regulatory demands on the biostatistician require the highest level of technical skill, challenging any new graduate for years ahead. The academically trained biostatistician in the CRO must develop consulting skills (*see* **Statistical Consulting**) to meet the needs of varied clients, projects, therapeutic subject matter areas, and appropriate designs for clinical trials. Depending upon the responsibilities of the biostatistician, time is required to learn or enhance skills and teamwork, communicate effectively in client/customer relationships, and learn managerial skills, whether these skills are needed on a particular project or in leading and directing staff members. Biostatisticians in CROs may progress from junior/assistant to principal statistician, from leader of a team to manager of biostatistics and to a senior role as a business leader. The Chief Executive and Founder of the world's largest CRO is a biostatistician by training. A career in contract research offers a dynamic environment with opportunities for initiative and scope to develop, excel, succeed, and prosper.

The Job

The main responsibility of the biostatistician in a Biostatistics Department in a CRO is typically quite operational, i.e. writing a detailed statistical analysis plan from a prescribed study protocol, including tables, listings, and graphical analyses as well as a report outline (*see* **Clinical Trials Protocols**). The biostatistician participates in the generation and programming of analysis files, tables, listings, and graphs and prepares the statistical report or contributes to an integrated medical and statistical study report.

The work of the biostatistician often includes discussions with the client about the design, appropriate sample size, and methods of analysis of a study, interaction with the biostatistical personnel of the client concerning the statistical sections of the protocol, and consultation with appropriate clinical and/or data management departments concerning monitoring of the study and issues of data quality. Other aspects of the job include, for example, regulatory issues, drug packaging and distribution, preparation of patient **randomization** schedules (*see* **Randomized Treatment Assignment**), and

marketing/sales/business development in relation to contractual issues. In some circumstances, the biostatistician may work on-site at the client's company. Internally, work processes and standard operating procedures need constant attention. Statistical tools, especially software packages, need to be regularly reviewed and upgraded as necessary. Training is always an important concern. At a more advanced level the biostatistician may be called upon to advise the client on statistical issues and/or methodologies (*see* **Teaching Statistics to Physicians**), on regulatory statistical strategy, on drug development plans, and regulatory submission formats and documentation.

In conclusion, a successful biostatistician in a CRO will have skills in multiple areas, from the highest level of technical ability to knowledge of regulatory affairs, and be a highly effective communicator, able to function effectively with senior personnel in other disciplines.

Facts and Figures About the CRO Industry

The CRO industry, which counts biostatistical consulting as one of its components, is a burgeoning environment of industrial scientific life, which has seen sustained annual growth rates of over 50% in the last few years. The total dollars outsourced to CROs has been growing at a rapid pace for half a decade – 20% to 30% per year – and it is expected that this growth will continue at these levels for at least the next few years. In 1996, there was an estimated $15 billion spent on clinical development by pharmaceutical and biotechnology companies, of which about 18% ($2.7 billion) was contracted out to an estimated 500 CROs. Dollars for clinical development are expected to grow by 6% to 9% annually at least through 1998; with the expected increase in dollars spent on clinical development, it is estimated that approximately $3.5 billion will be outsourced in 1997 and $4.4 billion in 1998.

There are hundreds of CROs around the world competing for the drug development dollars. However, the top five to ten largest CROs currently receive over half of the outsourced business.

DIANE GEHAN & JORGEN SELDRUP

Pseudo-likelihood

The term *pseudo-likelihood* has been used by several authors (e.g. Besag [1], Suzuki [6], Prentice [3], Kalbfleisch & Lawless [2], Wild [7], and Scott & Wild [4] in a rather heuristic way, to describe a function of the data and the parameters of interest that has properties similar to those of the usual **likelihood** function. Often, the pseudo-likelihood arises as an estimate of an unobserved likelihood based on complete data; this is especially the case in the context of response-selective or response-biased sampling. In such contexts, the pseudo-likelihood function for the parameter gives rise to a pseudo-score function as its logarithmic derivative; typically, the pseudo-score function has expectation zero, though sometimes this only holds asymptotically. The primary use of the pseudo-likelihood, then, is to define an estimating equation (*see* **Estimating Functions**) through setting the pseudo-score to zero. In many instances, these heuristics lead to good inferential procedures, but there is no clear theoretical theme to develop for this article. What I shall do, however, is develop one simple example where pseudo-likelihoods have been suggested to give a general idea of the approach, and provide entry points to the literature through references to a number of papers.

Suppose that N independent individuals give rise to data (y_ℓ, x_ℓ), $\ell = 1, \ldots, N$, which have the joint probability density function (pdf)

$$f(y|\mathbf{x}, \boldsymbol{\theta})g(\mathbf{x}).$$

Here, y is a **response variable**, $\mathbf{x}$ is a vector of **explanatory variables**, and $\boldsymbol{\theta}$ is the vector of parameters of interest which describes, perhaps among other things, the regression of y on $\mathbf{x}$. It is assumed that $\mathbf{x}$ arises from some pdf $g(\mathbf{x})$ which does not depend on $\boldsymbol{\theta}$.

With complete data, inference on $\boldsymbol{\theta}$ would be based on the likelihood arising from the product of conditional densities. The log likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{\ell=1}^{N} \log f(y_\ell|\mathbf{x}_\ell; \boldsymbol{\theta}), \tag{1}$$

with the corresponding score function

$$s(\boldsymbol{\theta}) = \sum_{\ell=1}^{N} s_\ell(\boldsymbol{\theta}), \tag{2}$$

where $s_\ell(\boldsymbol{\theta}) = \partial \log f(y_\ell|\mathbf{x}_\ell, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$.

We consider, however, situations in which we have incomplete data and further where the data collection is response-selective. Specifically, we suppose that the range S of the response variable Y is partitioned into k subsets: $S = S_1 \cup S_2 \ldots \cup S_k$, where $S_i S_j = \phi$ for all $i \neq j$. The data collection is response-selective in that the probability that $(y_\ell, \mathbf{x}_\ell)$ is observed depends upon the class in which y_ℓ falls. Thus, $(y_\ell, \mathbf{x}_\ell)$ is observed with known probability p_j if $y_\ell \in S_j$ for all $l = 1, \ldots, N$ and $j = 1, \ldots, k$. Such a sampling scheme could arise in several different ways, of which we mention only two:

1. *Basic* **stratified samples** (*BSS*). The N units have been generated from (1), and N_j are observed to have response y_ℓ in S_j. A simple **random sample** of $n_j = p_j N_j$ items is selected and fully observed in the jth stratum.
2. *Variable probability sampling (VPS)*. As items arise, their stratum membership is identified. If the lth item falls in S_j it is independently selected for full observation with specified probability p_j. Selection continues until we have a sample of n items.

Variations on both of these schemes are possible and arise in practice.

For simplicity, we consider VPS. An estimate of the full (unobserved) log likelihood (1) is given by the log pseudo-likelihood function

$$l_p(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{1}{p_j} \sum_{l \in D_j} f(y_\ell|\mathbf{x}_\ell, \boldsymbol{\theta}), \tag{3}$$

where D_j is the set of items observed in the jth stratum. The corresponding pseudo-score function that estimates (2) is

$$s_p(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{1}{p_j} \sum_{l \in D_j} s_\ell(\boldsymbol{\theta}). \tag{4}$$

It can be seen that $s_p(\boldsymbol{\theta})$ has expectation 0 and that, under fairly general conditions, the estimator $\hat{\boldsymbol{\theta}}_p$ that satisfies $s_p(\boldsymbol{\theta}) = 0$ is asymptotically normal with variance estimated by $\hat{\boldsymbol{A}}^{-1} + \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-1}$, where

$$\hat{\boldsymbol{A}} = \frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\ \partial\boldsymbol{\theta}^{\mathrm{T}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

$$\hat{\boldsymbol{B}} = \sum (1 - p_j) p_j^{-2} \sum_{l \in D_j} s_\ell(\hat{\boldsymbol{\theta}}) s_\ell(\hat{\boldsymbol{\theta}})^{\mathrm{T}}.$$

It should be noted that there are many ways, in general, that the full log likelihood (1) could be estimated. The "weighted" pseudo-likelihood (3) and related pseudo-score (4)represent only one approach. Scott & Wild [4] compare a number of competing approaches to this problem.

One simple example of a design of this type is the **case–cohort** design of Prentice [3]. In this, N individuals are at risk of failure over an interval $(0, \tau)$, and interest centers on relating the failure rate $\lambda(t|\mathbf{z})$ to a vector of covariates $\mathbf{z}$ which can be measured on each individual under study. It is too expensive, however, to observe all individuals. Instead, the data comprise observations on all individuals who fail in $(0, \tau)$ and, in addition, a sample of those individuals who do not fail. Specifically, each individual in the cohort who does not fail is independently observed with probability p_2. Here, the response variable for the lth individual is T_ℓ and its range is divided into $S_1 = (0, \tau)$ and $S_2 = [\tau, \infty)$. The sampling rates for the two strata are $p_1 = 1$ and $p_2 < 1$, and the pseudo-likelihood described above can be applied directly for the case $\lambda(t|\mathbf{z}; \boldsymbol{\theta})$, a parametric model. Details are given in Kalbfleisch & Lawless [2], where various alternatives are also considered and compared.

Prentice [3] considers the case–cohort study in the context of a **proportional hazards model**

$$\lambda(t|\mathbf{z}) = \lambda_0(t) r(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}), \tag{5}$$

where $\mathbf{z}$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of regression parameters, and $r(x)$ is a **relative risk** function. He derives a pseudo-likelihood based on case–cohort data, and Self & Prentice [5] evaluate its asymptotic properties.

References

[1] Besag, J.E. (1977). Efficiency of pseudo-likelihood estimation for simple Gaussian fields, *Biometrika* **64**, 616–618.

[2] Kalbfleisch, J.D. & Lawless, J.F. (1988). Likelihood analysis for disease incidence and mortality, *Statistics in Medicine* **7**, 149–160.

[3] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.

[4] Scott, A.J. & Wild, C.J. (1997). Fitting regression models to case cohort data by maximum likelihood, *Biometrika*, **84**, 57–71.

[5] Self, S.G. & Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies, *Annals of Statistics* **16**, 64–81.

[6] Suzuki, K. (1985). Estimation of lifetime parameters from incomplete field data, *Technometrics* **27**, 263–272.

[7] Wild, C.J. (1991). Fitting prospective regression models to case-control data, *Biometrika* **78**, 705–717.

(*See also* **Bias from Nonresponse**; **Partial Likelihood**; **Profile Likelihood**)

JOHN D. KALBFLEISCH

Pseudo-random Number Generator

One of the most frequently called functions on a scientific computer is the random number generator. Random numbers are required for many purposes. A major use is in **simulations**. Large quantities of random numbers are needed to generate the **random samples** from distributions (theoretical and/or empirical) which are the basis of any stochastic simulation. Another use for random numbers is the **Monte Carlo** evaluation of multivariate integrals (*see* **Markov Chain Monte Carlo**). Ideally, what is required in a simulation is a stream of independently and uniformly distributed random variables taking values between 0 and 1 [i.e. a random sample from the **uniform distribution** on (0,1)]. At best this ideal can only be realized approximately.

Among the possible ways of generating random numbers are mechanical processes, such as those often used to select the winning numbers in lottery draws, and methods using electronic noise, as in ERNIE, the device used to select the winning numbers in the British premium bond scheme (another lottery). It is difficult to validate such machines and to ensure that they always behave correctly. Furthermore, if one needs to reproduce the stream of numbers used in a particular application, this can be done only by storing the complete stream as it is created. Such devices are also unsuitable for rapid access by the simulation program.

Nowadays, computers almost invariably use a "random number generator" program. These are, in fact, *pseudo-random number generators* (PRNGs), because it is impossible to construct a practicable program for computing truly random sequences. PRNGs are programs that produce a deterministic sequence that mimics a random sequence of numbers. Their advantages include speed, reliability, reproducibility, and portability. Because simulations may consume very large numbers of random numbers it is usually important that it takes very little time to produce a single one; PRNGs are usually highly efficient (often very short) programs and hence can be very fast. They are reliable sources of pseudo-random numbers in the sense that, *provided they are correctly implemented*, they will operate precisely as their theoretical specification predicts. There is no difficulty in producing a perfect replica of a sequence. Hence, if required, an entire simulation can be repeated exactly. PRNGs can be made extremely portable so that different users can get the identical random number stream, even on different types of computer.

Ripley [10] gives a formal definition: a sequence of pseudo-random numbers is a deterministic sequence of numbers in the interval [0, 1] having the same relevant statistical properties as a sequence of random numbers. The key word here is *relevant*. Before using a PRNG for a particular project it is essential to check whether it can be regarded as producing a sequence that is 'random enough' for the purpose in hand. For example, a sequence that is suitable for a fairly small **queueing** simulation may be unsatisfactory for a large **operations research** simulation of the ordering systems for a large health authority. It may be possible to predict suitability from a knowledge of the theoretical properties of the PRNG or it may be necessary to carry out a battery of statistical tests on trial sequences to check its suitability for various purposes. Some generators have been very well researched in both aspects. Some have been shown to have serious flaws but are, unfortunately, still in use.

The advantages and disadvantages of PRNGs can best be explained by considering a class of generators that has enjoyed wide popularity for most of the latter part of the century. These are the *multiplicative linear congruential generators* (MLCGs), often abbreviated to *multiplicative congruential generators* (MCGs). They are based on a very simple integer recurrence relationship. The **algorithm** starts with a "seed" value, say X_0. It then applies the congruency relationship,

$$X_i = aX_{i-1} (\text{mod } m), \tag{1}$$

to obtain the next member of the sequence (i.e. X_i is the remainder when the product aX_{i-1} is divided by m); X, a, and m are all integers. The pseudo-random numbers are obtained by setting

$$U_i = \frac{X_i}{m}. \tag{2}$$

At best, (1) produces each of the integers $1, 2, \ldots, m-1$ once, in some apparently random order, and then repeats that sequence exactly, i.e. the output is periodic. However, not all MCGs produce *all* the integers before starting to repeat the sequence.

The length of the period depends on a and m. A convenient choice for m, for ease of calculation on a binary computer, requires it to be a power of 2, say $m = 2^k$, but a necessary condition to achieve the maximum possible period is that m is prime. For computers with 32-bit word-length 2^{32}, 2^{31}, and especially $2^{31} - 1$ are the popular values. The MCG with $m = 2^{31} - 1$ has been particularly thoroughly researched because $2^{31} - 1$ is prime. Prime m is considered desirable for various reasons: The properties of generators with prime m are well understood. For a given m, there are many values of a which lead to a generator with the maximum possible period $(m - 1)$. Good (well-tested) implementations are available. Research has shown that, when $m = 2^{31} - 1$, one of the best choices for a is 16 807. The resultant generator has been in widespread use for many years and is still a popular choice for many applications.

In order that the actual output sequence from any generator program is precisely as theory predicts, the recurrence computations must be exact, i.e. there must be no rounding or other numerical errors. Owing to their finite word-length, computers can only represent real numbers approximately, whereas integers can be represented exactly (up to a certain size determined by word-length and the computer's arithmetic). This is the reason why the integer recurrence given in (1) is used, rather than a recurrence on the U_i directly.

The upper limit on exact integer representation causes implementation problems. For reasons of efficiency, m is often chosen to be very close to the upper limit. In this situation, aX can be very much larger than m and hence the upper limit. Thus, a naïve implementation of (1) will fail disastrously. Ways of overcoming this problem include using higher precision computer arithmetic (*see* **Floating Point Arithmetic**), or breaking up the problem and replacing the single (1) by several equations in each of which the upper limit is never exceeded.

The length of the period of the MCG with $m = 2^{31} - 1$ and $a = 16\,807$ is approximately 2×10^9. This is adequate for many problems, but the large-scale simulations that are increasingly common nowadays require PRNGs with much longer periods. One way of obtaining a generator of greater length, but which still uses only simple MCGs, is to combine the outputs of two or more different MCGs. The simplest way of doing this is by adding their outputs and taking the fractional part (as was originally proposed by Wichmann & Hill [12] for 16-bit computers). It is straightforward to compute the period of the combined generator: for suitable choices of m and a it is just the product of the individual periods. The increased period is not the only benefit. For well-chosen generators, the "randomness" of the combined generator is found to be better than that of the individual components. For example, L'Ecuyer [6] combined two particular MCGs, each of whose modulus was very close to $2^{31} - 1$. The combined generator has period 2×10^{18} and behaves very well. But period lengths of this order are still inadequate for many present-day applications. Adding more component MCGs would increase the period but it would also slow down the generator.

An alternative way of combining generators is *shuffling*: the order of the output sequence of one generator is varied "randomly" by the output from another generator. The theoretical properties of the resultant generator are difficult to determine. Such generators have not found favor generally.

Before considering other types of generator, a particular aspect of the randomness behavior of PRNGs in general needs to be mentioned. So far, we have only considered the generation of univariate uniform variables – these may be represented as random points on the unit line. However, many problems nowadays require the generation of multivariate random variables. This involves generating multivariate uniform variables – these may be regarded as random points in n-dimensional space, $n \geq 2$, depending on the specification of the random variable. The standard way of doing this is to use n consecutive numbers in the output from the PRNG as the coordinates of a pseudo-random point in n-dimensional space. Similarly, the next n numbers in the sequence form the next point, and so on. Marsaglia [8] found that for MCGs such points were located on a limited number of parallel hyperplanes instead of being distributed randomly in space. This lattice structure reflects the **serial correlation** structure (i.e. the lack of independence) in the output sequence from the MCG. For some generators the resultant regularity can easily be seen in two dimensions by plotting the output numbers against each other in pairs. Some generators are better than others, in the sense that the distance between the hyperplanes is less. Tests for lattice structure have been devised and can be applied, along with other tests, when validating a particular PRNG. Whether or not lattice structure is a serious problem in practice depends very much on the particular type

of application for which the generator is being used. It is known that combining MCGs by addition does not remove lattice structure problems.

MCGs are the special case of the *linear congruential generator* (LCG),

$$X_i = aX_{i-1} + c(\text{mod } m), \tag{3}$$

where $c = 0$. However, provided m is prime, incorporating a nonzero value of c confers no advantage over the corresponding MCG.

The LCGs are members of a much wider class of generator for which the current value X_i depends on more than one of the previous values:

$$X_i = a_1X_{i-1} + a_2X_{i-2} + \cdots + a_kX_{i-k}(\text{mod } m), \tag{4}$$

where some of the a_j may be zero.

The MCG of (1) uses large values of m. At the other extreme, we can set $m = 2$ and restrict a to either 0 or 1 in (4). The output stream is then a sequence of 0s and 1s. Viewing these as computer bits, b_j, we can construct pseudo-random numbers in the form of binary fractions, for example

$$U_i = 0 \cdot b_{i\ell}b_{i\ell+1} \ldots b_{i\ell+n}, \ell > n. \tag{5}$$

This type of generator is known as a *Tausworthe* or *feedback shift register* (FSR) generator because of the way in which it is implemented – the computer operations required to produce each pseudo-random number are very simple and fast. *Generalized feedback shift register* (GFSR) generators result if U_i is made up from nonconsecutive bits in the output stream. Instead of the simple seed that an MCG requires, FSR and particularly GFSR generators require special initialization procedures; they also use much more memory. However, given appropriate choices of constants and correct implementation, very fast generators with extremely large periods and good "randomness" properties (including their higher-dimensional behavior) can be constructed. Well-tested implementations are available. Combinations of GFSR generators by addition have also been implemented [11].

Other generators that have recently been introduced by Marsaglia & Zaman [9] are based on replacing the addition operation in special cases of (4) by other binary operators. These generators are known as *add-with-carry* (AWC) and *subtract-with-borrow* (SWB) generators. They are designed to be very fast and have extremely long periods. However recent theoretical analysis [1] suggests that their lattice structures are effectively equivalent to those of LCGs with very large moduli m.

A different approach to obtaining better lattice structure is to be found in the *inversive congruential generators* (e.g. [4]). They use the relationship

$$X_i = a\overline{X}_{i-1}(\text{mod } m), \tag{6}$$

where $\overline{X}$ (the *inverse* of X) is the solution of $X\overline{X} = 1(\text{mod } m)$. Such generators are known to have much better lattice structure than the normal MCGs, but suffer from the same constraints on period length. Also, the computation of $\overline{X}$ is time-consuming.

Properties such as period length are global (i.e. they describe overall features of the entire output sequence of a PRNG) and are derived theoretically. However, the practical question that faces any user is the behavior of samples from the sequence and, most important, the relevance of that behavior to particular applications. It is, therefore, extremely important that very thorough tests are carried out before a generator is released for general use. For the great majority of users, writing PRNG programs is not a recommended activity. However, it is not a good idea to place too much trust in the random number functions provided by computer operating systems (or indeed by various packages, particularly those that are not specifically designed for statistical use) without first checking their specifications and suitability for the application in hand. The reader is recommended to use one of the thoroughly researched and tested professional implementations (for example, in the IMSL and NAG libraries; *see* **Numerical Analysis**).

There is a very substantial research literature on the generation of pseudo-random numbers and this is increasing rapidly. General introductions to the standard methods can be found in the books by Dagpunar [2] (fairly elementary) and Ripley [10] (more mathematical). A good brief survey is given by L'Ecuyer [7]. Currently, a useful source of information on various aspects of random number generation is the World Wide Web site `http://random.mat.sbg.ac.at`, maintained by the Mathematics Department of the University of Salzberg, Austria.

The design and implementation of algorithms for converting pseudo-random numbers into pseudo-random values from various nonuniform distributions is best left to professionals. Some general methods are

available, but usually a substantial amount of work is needed to produce a workable algorithm for any particular distribution (*see* **Simulation**). For many distributions there is an *ad hoc* method based on the particular properties of that distribution. As in the case of PRNGs, the correct and efficient implementation of the algorithms is a skilled activity and the resulting programs require very thorough testing before adoption for general use. There is a very substantial research literature on nonuniform random variable generation. The books by Dagpunar [2] and Ripley [10] contain introductions to this topic as well. Devroye's encyclopedic book [3] concentrates on the theory leading to a vast range of algorithms, but contains very little practical information on implementation and efficiency. For a very recent general survey of the area and suggestions for further reading, see Kemp [5].

References

[1] Couture, R. & L'Ecuyer, P. (1994). On the lattice structure of certain linear congruential sequences related to AWC/SWB generators, *Mathematics of Computation* **62**, 799–808.

[2] Dagpunar, J. (1988). *Principles of Random Variate Generation*. Clarendon Press, Oxford.

[3] Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

[4] Eichenauer-Herrmann, J. (1992). Inversive congruential pseudorandom numbers: a tutorial, *International Statistical Review* **60**, 167–176.

[5] Kemp, C.D. (1997). Computer generation of random variables, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz, C.B. Read & D.L. Banks, eds. Wiley, New York.

[6] L'Ecuyer, P. (1988). Efficient and portable combined random number generators, *Communications of the Association for Computing Machinery* **31**, 742–779; 774.

[7] L'Ecuyer, P. (1990). Random numbers for simulation, *Communications of the Association for Computing Machinery* **33**, 86–97.

[8] Marsaglia, G. (1968). Random numbers fall mainly in the planes, *Proceedings of the National Academy of Sciences* **61**, 25–28.

[9] Marsaglia, G. & Zaman, A. (1991). A new class of random number generators, *Annals of Applied Probability* **1**, 462–480.

[10] Ripley, B.D. (1987). *Stochastic Simulation*. Wiley, New York.

[11] Tezuka, S. & L'Ecuyer, P. (1991). Efficient and portable combined Tausworthe random number generators, *Association for Computing Machinery Transactions on Modeling and Computer Simulation* **1**, 99–112.

[12] Wichmann, B.A. & Hill, I.D. (1982). An efficient and portable pseudo-random number generator, *Applied Statistics* **31**, 188–190.

(*See also* **Uniform Random Numbers**)

C.D. KEMP

Psychiatry

> The mentally ill have always been with us – to be feared, marveled at, laughed at, pitied or tortured, but all too seldom cured (Alexander & Selesnick, *The History of Psychiatry*, 1967).

In his *Dictionary of Psychology*, Professor Stuart Sutherland defines psychiatry as "the medical speciality that deals with mental disorders". An almost equally brief definition appears in Campbell's *Psychiatric Dictionary*, namely "the medical speciality concerned with the study, diagnosis, treatment and prevention of behaviour disorders". In terms of either definition psychiatry has a long history since, for example, Pythagoreans employed music therapy with emotionally ill patients [11], and Aretaeus (A.D. 50–130) observed mentally ill patients and did careful follow-up studies on them. As a result, he established the fact that manic and depressive states occur in the same individual and that lucid intervals exist between manic and depressive periods. He considered mental illness in terms of outcome, emphasizing the course of the disease and its prognosis. He also understood that not all persons with mental illness are destined to suffer intellectual deterioration, a fact not again adequately emphasized until the twentieth century.

A widely quoted remark of **Galton** is that until the phenomena of any branch of knowledge have been submitted to measurement and number, it cannot assume the dignity of a science. Psychiatry, during the twentieth century, has struggled to attain such scientific respectability by making quantitative observations on mentally ill patients, and psychiatrists have become increasingly aware that a strict scientific approach is required for their discipline to progress. Statistics has, during this period, become a most important basic science in psychiatry and more and more often psychiatric researchers resort to sophisticated and powerful statistical techniques to help them unravel the complexities of their data.

But in the nineteenth century and earlier, the use of statistics in psychiatry was largely restricted to simple descriptive measures, and it is only in the second half of this century that the use of inferential (*see* **Inference**) and other more complex methods has become widespread. However, the use of even simple descriptive statistics was important and their presentation often led to changes in policy if not to changes in attitude to the problems of lunacy. Table 1, for example, taken from Schull [22], shows the number of people officially identified as insane and the rate of insanity per 10 000 people in England and Wales at various times during the nineteenth century. The increase in lunacy as suggested by these figures became one of the main weapons in reformers' arguments for new legislation to deal with the insane, since they indicated that insanity was now a serious social problem, a view endorsed by the following from the 1844 Report of the Metropolitan Commissioners on Lunacy:

> Lunatics have unfortunately become so numerous throughout the whole kingdom, that the proper construction and cost of asylums for their use has ceased to be a subject which affects a few counties only, and has become a matter of national interest and importance.

Schull points out, however, that the achievement of reform (the construction of asylums and the employment of doctors to effect the cure of the insane) failed to bring a halt or even a diminution in the rapid upward spiral of cases of lunacy. Between 1844 and 1860, when the population as a whole grew by just over 20%, the number of lunatics almost doubled; and the growth in the number of the insane continued to far outstrip the rate of increase of the general population for the rest of the century. Schull discusses the various "official" explanations for the increase, one of which was that a large number of cases previously unreported had only recently been brought under observation because the method of gathering statistics on insanity had previously been slipshod and inadequate; the apparent increase was therefore dismissed as largely a statistical artefact. Alternative explanations for the increase assumed it to be real and attributable to stresses attendant upon life in a higher "mechanical" civilization. (The arguments described here appear not to have altered dramatically over the intervening 150 years!)

Examination of early issues of the *Journal of Mental Science* (the forerunner of today's *British Journal of Psychiatry*) provides an example of "statistical proof" in Matt's [18] investigation of the inheritability of insanity, although the proof amounts only to the presentation of a set of data rather than to the use of more formal inferential methods.

An indication of the psychiatrist's attitude to mathematical and statistical topics in the early part of the

Table 1 Total population, total number officially insane, and rate of insanity per 10 000 people in England and Wales

January 1	Population	Number deemed insane[a]	Rate per 10 000	Source of data on number insane
1807	9 960 000	2248	2.26	House of Commons, 1807
1819	11 106 000	6000	5.40	Burrows, 1820
1828	13 106 000	8000	6.10	Halliday, 1828
1829	13 370 000	16 500	12.34	Halliday, 1829
1836	14 900 000	13 667	9.18	Parliamentary Return, 1836
1844	16 480 000	20 893	12.6	Metropolitan Commissioners on Lunacy
1850			Not available	
1855	18 786 914	30 993[b]	16.49	Commissioners on Lunacy
1860	19 902 713	38 058	19.12	Annual Reports
1865	21 145 151	45 950	21.73	
1870	22 501 316	54 713	24.31	
1875	23 944 459	63 793	26.64	
1880	25 480 161	71 191	27.94	
1885	27 499 041	79 704	28.98	
1890	29 407 649	86 067	29.26	

[a]Includes lunatics in asylums, but also those in workhouses, at large in the community, etc.
[b]The Commissioners found 20 493 lunatics in asylums of all types in 1855; lacking a complete enumeration of all lunatics not so confined, they estimated that these amounted to 10 500 persons.

twentieth century may perhaps be gleaned from the comments made by Edward Mapother when reviewing Spearman's book *The Abilities of Man: Their Nature and Measurement* in the *Journal of Mental Science* in 1928:

> Doubtless most readers of this Journal, like the reviewer, will be content to take for granted the mathematics involved.

Freud, too, was not convinced about statistics. Fleiss reports that in the 1920s, Joseph Zubin and a few fellow graduate students undertook a study of 4-, 5-, 6-, and 7-year-old children to put to the test Freud's Oedipus theory. Data were collected and analyzed, and the statistical results seemed to confirm the master's theory. It was Zubin's task to prepare the tables, charts, and summary statistics and to send them to Freud. "Ganz amerikanisch" was his disparaging reply, implying that only in America was the need felt to test what was obvious.

Nevertheless, statistical methodology began to appear in psychiatric journals around this time. Cameron [1], for example, in a study of perseveration used the **correlation** coefficient and a test to assess its significance. A report by the Royal Medico–Psychological Association Committee on Mental Deficiency on the evidence of neuropathic conditions in the relatives of normal persons published in the *Journal of Mental Science* in 1937 used a *t* test to examine the difference in average family size for two groups of families, those who included a weak-minded person and those who did not. (The details of the *t* test (*see* **Student's *t* Distribution**) were confined to an appendix.) The same paper also contains an application of a one-way **analysis of variance**. Masserman & Carmichael [17] used Pearson's contingency coefficient to assess the relationship between diagnosis and prognosis in psychiatry.

The first **clinical trial** in psychiatry came in response to the focal infection theory as a cause of mental disorder [3]. Kopeloff & Cheney [14] and Kopeloff & Kirby [15] undertook to test this hypothesis by removing surgically all the foci of infection from one group of patients and comparing their outcome with an untreated control group. The only difference found was a higher mortality from the experimental group but no difference in outcome. No statistical tests of differences were made, and results were reported simply in terms of numbers and percentages of improved cases. It was only 50 years later that Zubin & Zubin [25] calculated the statistical significance of the difference in percentages and found it to be nonsignificant.

The first psychiatrist to advocate the use of **Fisher's** experimental methods in the evaluation of

physical treatments in psychiatry appears to have been Sir Aubrey Lewis [16]. In this article, he criticizes the past use of small series of cases and "the common lack of a coordinated plan for the therapeutic experiment". Of a controlled trial he concludes:

> An organized experiment would demand much that has not hitherto been practicable, including voluntary acceptance by independent hospitals and clinics of an agreed procedure for the selection, management, evaluation of mental state, and follow-up investigation of treated, as well as of control cases. Such an experiment, as R.A. Fisher has demonstrated, requires much forethought and self-discipline on the part of those who carry it out.

Lewis concludes with the following statement:

> For the most important psychiatric conditions, such trials are essential, unless we are prepared to go on taking decades to decide questions which could be settled in a few years.

Although small controlled trials of electro-convulsive therapy (ECT) and psychotropic drugs began to appear in the 1950s, it was some 20 years later that the results of the first trial of the sort envisaged by Lewis were published, when, in 1965, the Medical Research Council's **multicenter trial** of drug therapy and ECT in the treatment of depression was completed.

(It is interesting to note that this trial was not universally welcomed. In a letter to the *British Journal of Psychiatry*, Sargant wrote: "There is no psychiatric illness in which bedside knowledge and long clinical experience pays better dividends; and we are never going to learn how to treat depressions properly from double-blind sampling in an MRC statistician's office.")

Early studies of the inheritance of mental disorders also had reason to be grateful to Fisher. In his **twin studies** and family studies of manic depression, the British psychiatrist Elliot Slater was concerned with the problem of the use of age-of-incidence data in the estimation of life-time risk, and acknowledges early advice on appropriate statistical methods from Fisher; the following is taken from Slater's autobiographical sketch [23]:

> I had the temerity to write to R.A. Fisher to ask for his help. He gave it at once, and in a completely satisfying way; and it was in his journal that my paper was eventually published. Fisher helped me on many occasions, in investigations I carried out in later years, and never grudged time or trouble. I came to have for him the greatest admiration and affection.

Psychiatric epidemiology was at first hampered by inconsistencies in the reporting of the characteristics of the mentally ill, but in the US, at least, became more viable after 1917 when the American Medico–Psychological Association's Committee on Statistics urged all mental hospitals to adopt a uniform reporting system. With the assistance of the National Committee for Mental Hygiene, this Association produced the first uniform nomenclature of mental diseases in 1918. After World War II there was a literal explosion of community and demographic studies of the mentally ill. The National Mental Health Act was passed in the US in 1946. A result of this Act was that responsibility for gathering data on the mentally ill was transferred to the Public Health Service and the soon-to-be-created National Institute of Mental Health.

In England, the 1949 Millbank Memorial Fund Conference on the epidemiology of mental disorders produced general agreement on the importance of epidemiology for causal research and for administrative policy. Its relevance to clinical psychiatry, however, was disputed by many of the practising psychiatrists at the Conference who questioned how far epidemiological inquiry should be based on the conventional schemata of disease, which, in their opinion, were inapplicable to mental disorders. Francis reminded the Conference that epidemiology is basically dependent upon the accuracy of diagnosis and that until a valid base for **classification** can be generally employed, data from different areas cannot be properly compared.

The realization that, as in other medical and scientific disciplines, classification is also fundamental in psychiatry led many psychiatrists and statisticians in the 1950s, 1960s, and 1970s to devote much energy to using **multivariate analysis** techniques such as **principal components analysis**, **factor analysis**, and **cluster analysis** on various sets of psychiatric data in an effort to refine or even redefine diagnostic categories. Examples of such studies are those of Fleiss & Zubin [9] and Everitt et al. [8]. In many respects, this opportunity to apply complex multivariate techniques to an important practical problem provided an impetus to research in multivariate analysis in general and cluster analysis in particular. The reciprocal impact

Table 2 *American Journal of Psychiatry (AJP), British Journal of Psychiatry (BJP)* and *Archives of General Psychiatry (AGP)* during 1980 by categories of statistical usage

	Categories of statistical usage	*AJP* (1980) 339 papers	*BJP* (1980) 148 papers	*AGP* (1980) 110 papers
1.	Expository literature review, etc.	18 (5.3%)	6 (4.1%)	4 (3.6%)
2.	No statistical data, case reports, etc.	115 (33.0%)	12 (8.1%)	2 (1.9%)
3.	Descriptive statistics only tables, graphs, means, variances	65 (19.2%)	14 (9.5%)	11 (10.0%)
4.	Chi-squared and t tests, Fisher exact test; 1 or 2 samples, contingency tables	95 (28.0%)	75 (50.7%)	66 (50.0%)
5.	Product–moment correlations, rank correlations	42 (12.4%)	22 (14.9%)	30 (27.3%)
6.	Analysis of variance F tests: 1-, 2-, and higher way	25 (7.4%)	22 (14.9%)	32 (9.1%)
7.	Nonparametric rank methods (other than rank correlations)	9 (2.7%)	11 (11.5%)	10 (9.1%)
8.	Measures of association and agreement (other than correlation)	10 (2.9%)	13 (8.8%)	9 (8.2%)
9.	Regression analysis simple, multiple, polynomial, stepwise	6 (1.8%)	9 (6.1%)	10 (9.1%)
10.	Discriminant and factor analysis	4 (1.2%)	6 (4.1%)	7 (6.4%)
11.	Estimation: maximum likelihood interval estimation, etc.	0 (0.0%)	3 (2.0%)	2 (1.8%)
12.	Cluster analysis, classification	1 (0.3%)	0 (0.0%)	1 (0.9%)
13.	Life tables, life testing, survival analysis	1 (0.3%)	0 (0.0%)	2 (1.8%)
14.	Time series analysis spectral analysis	2 (0.6%)	0 (0.0%)	1 (0.9%)
15.	Classical experimental design: Latin squares, hierarchical models	0 (0.0%)	3 (2.0%)	1 (0.9%)
16.	Bayesian methods	0 (0.0%)	0 (0.0%)	1 (0.9%)

of the latter on psychiatry has, however, been limited. Certainly, neither of the current editions of the two diagnostic classifications systems in use today, the American Psychiatric Association's *Diagnostic and Statistical Manual* and the World Health Organization's **International Classification of Diseases**, appears to have benefited from any of the many cluster analysis studies that have been reported in the psychiatric literature.

In 1985, DeGroot & Mezzich [5] produced their Table 2 showing articles in three major psychiatric journals by category of statistical usage. By a large margin, the most common appearance of statistics in psychiatric journals is in the form of a significance test (*see* **Hypothesis Testing**), usually either a simple t or **chi-squared test**. But there is considerable evidence that not even such simple methods are always used wisely. White [24], for example, considered 12 issues of the *British Journal of Psychiatry* from July 1977 to June 1978, and found statistical errors that could potentially affect at least one conclusion in over a third of the papers. One common error was failure to use a dependent samples t test for matched data (*see* **Matched**

Analysis). A similar exercise carried out nearly 20 years later by McGuigan [19] showed that the problem had, if anything, become worse, since now 40% of papers examined contained serious statistical errors.

It is likely that repeating DeGroot & Mezzich's study today would give similar percentages to those shown in their Table 2, particularly with respect to simple significance tests. Psychiatrists (or at least editors of psychiatric journals) appear to have developed a fondness for the ***P* value**, which overcomes the many criticisms of its use that have been made in the psychological and medical literature – see, for example, Rozeboom [21], Oakes [20], and Gardner & Altman [10]. But despite its likely overall similarity, the 1995 version of DeGroot & Mezzich's Table 2 would contain a number of new categories. **Logistic regression**, for example, is now applied routinely in many psychiatric investigations. **Correspondence analysis** (see Greenacre [12]) has increased in popularity and is often used to display **contingency tables** graphically. A recent example is given by Corten et al. [2] in an investigation of subjective **quality-of-life** measures in assessing rehabilitation treatment in psychiatry. **Structural equation modeling** (see Dunn et al. [7]) has also begun to appear in the psychiatric and related literature. Hines et al. [13], for example, use such models to assess the influence of the corpus callosum on verbal fluency, language lateralization and visuospatial ability. Brain structure was modeled through correlations between measures taken of the cross-sectional surface area of the posterior fifth (splenum), the posterior third minus fifth (isthmus), the anterior fourth (genu) and the midregion lying between the isthmus and the genu using magnetic resonance imaging.

Meta analysis is now used frequently in medicine in general and psychiatry in particular for combining results from different studies of the same topic. A recent example involves the possible association between schizophrenia and birth complications. Many other categories of technique could almost certainly be added, although their number of occurrences would probably be rather small.

Research in psychiatry includes components from the biological, medical, behavioral, physical and social sciences. The mixture provides many opportunities for the statistician interested in diverse application experiences, and statistics has a central role to play in the continuing development of psychiatry as a scientific discipline. Statistical thinking should pervade every stage of a research investigation in psychiatry.

Two of the most exciting features of psychiatry in the last few years of the twentieth century are the availability of techniques such as *functional magnetic resonance imaging* (fMRI) with the promise of at least the possibility of new and powerful insights into the working of the human brain and the causes of mental illness, and the appearance of powerful new drugs capable of making the last part of the quotation that began this article obsolete. Vast amounts of data are generated by fMRI and related methods and the creation of appropriate statistical methodologies for their analysis will be a further challenge for statisticians working in psychiatry in the future. Longitudinal clinical trials to test new therapies generate many statistical problems and statisticians have, over the course of the last five years or so, developed a range of new techniques for analyzing the data from such trials (see, for example, Diggle et al. [6], and Crowder & Hand [4]) (*see* **Longitudinal Data Analysis, Overview**). As yet these developments have not had a major impact on psychiatric reports of clinical trials, and a further challenge for statisticians working with psychiatrists is to persuade them that the new methods offer real advantages.

The increasing collaboration of psychiatrists and statisticians over the last 20 years has had benefits for both and the future offers further challenges and possibilities in seeking the ultimate goal of overcoming the misery that is mental illness.

References

[1] Cameron, D.E. (1933). Studies in perseveration, *Journal of Mental Science* **79**, 735–745.

[2] Corten, P., Mercier, C. & Pelc, I. (1994). Subjective quality of life: clinical model for assessment of rehabilitation treatment in psychiatry, *Social Psychiatry and Psychiatric Epidemiology* **29**, 178–183.

[3] Cotton, H.A. (1922). The etiology and treatment of the so-called functional psychoses, *American Journal of Psychiatry* **2**, 157–210.

[4] Crowder, M. & Hand, D.J. (1996). *Practical Analysis of Longitudinal Data*. Chapman & Hall, London.

[5] DeGroot, M.H. & Mezzich, J.E. (1985). Psychiatric statistics, in *A Celebration of Statistics: The ISI Centenary Volume*, A.C. Atkinson & S.E. Fienberg, eds. Springer-Verlag, New York, pp. 145–165.

[6] Diggle, P.J., Liang, K.Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford Science Publications, Oxford.

[7] Dunn, G., Everitt, B.S. & Pickles, A. (1993). *The Analysis of Covariances and Latent Variables using EQS*. Chapman & Hall, London.

[8] Everitt, B.S., Gourlay, A.J. & Kendell, R.E. (1971). An attempt at validation of traditional psychiatric syndromes by cluster analysis, *British Journal of Psychiatry* **119**, 399–412.

[9] Fleiss, J.L. & Zubin, J. (1969). On the methods and theory of clustering, *Multivariate Behavioral Research* **4**, 235–250.

[10] Gardner, M.J. & Altman, D.G. (1986). Confidence intervals rather than P-values: estimation rather than hypothesis testing, *British Medical Journal* **292**, 746–750.

[11] Gordon, B.L. (1949). *Medicine Throughout Antiquity.* Davies, Philadelphia.

[12] Greenacre, M.J. (1992). Correspondence analysis in medical research, *Statistical Methods in Medical Research* **1**, 97–117.

[13] Hines, M., Chiu, L., McAdams, L.A., Bentler, P.M. & Lipeaman, J. (1992). Cognition and the corpus callosum: Verbal fluency, visuospatial ability and language lateralization related to midsagittal surface area of callosal subregions, *Behavioral Neuroscience* **106**, 3–14.

[14] Kopeloff, N. & Cheney, C.O. (1922). Studies in focal infection: its presence and elimination in the functional psychoses, *American Journal of Psychiatry* **2**, 139–156.

[15] Kopeloff, N. & Kirby, H.G. (1923). Focal infection and mental disease, *American Journal of Psychiatry* **3**, 149–198.

[16] Lewis, A.J. (1946). On the place of physical treatment in psychiatry, *British Medical Bulletin* **3**, 22–24.

[17] Masserman, J.H. & Carmichael, H.T. (1938). Diagnosis and prognosis in psychiatry, *Journal of Mental Science* **84**, 893–946.

[18] Matt, F.W. (1913). The neuropathic inheritance, *Journal of Mental Science* **59**, 222–263.

[19] McGuigan, S.M. (1995). The use of statistics in the British Journal of Psychiatry, *British Journal of Psychiatry* **167**, 683–688.

[20] Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences.* Wiley, Chichester.

[21] Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test, *Psychological Bulletin* **57**, 416–428.

[22] Schull, A.T. (1979). *Museums of Madness.* Allen Lane, London.

[23] Slater, E. (1971). Autobiographical sketch, in *Man, Mind, and Heredity*, G. Shields & I.I. Gottesman, eds. Johns Hopkins University Press, Baltimore, pp. 1–23.

[24] White, S.J. (1979). Statistical errors in the British Journal of Psychiatry, *British Journal of Psychiatry* **135**, 336–342.

[25] Zubin, D. & Zubin, J. (1977). From speculation to empiricism in the study of mental disorder: research at the New York State Psychiatric Institute in the first half of the 20th century, in *Roots of American Psychology: Historical Influences and Implications for the Future*, R.W. Rieber & K. Salzinger, eds. Annals of the New York Academy of Sciences, New York, pp. 104–135.

BRIAN S. EVERITT

Psychometrics, Overview

Gustav Fechner (1801–1887), the founder of psychophysics, was an optimist. Not only was he convinced that the mind coexisted with the body, but he believed that he could prove it. Initially, he enthusiastically studied "after images" by staring into the sun. Because these images lasted when he turned from the sun, that was initial proof for him that conscious perception was separate from physical perception. Fechner was almost blind from these studies when he came across the work of E.H. Weber, one of his colleagues at Leipzig.

Weber was interested in muscular sensation and noticed that when subjects lifted objects of varying weights, sometimes they could perceive the difference between objects, and sometimes they could not. Weber concluded that when a person distinguishes between two stimuli, it is not the difference between them that is perceived, but rather the ratio of this difference to the magnitude of the things being compared. That is, if subjects were comparing a 20 kg weight with a 20.1 kg weight, they may not perceive the two weights as different. However, when comparing a 1 kg weight with a 1.1 kg weight, the difference was readily perceived. Weber characterized his findings with the following formula:

$$\frac{\Delta R}{R} = K,$$

where ΔR was the "just noticeable difference" between a standard and comparison stimulus, K was a constant, and R was the magnitude of the standard stimulus. This formula states that a noticeable (threshold) stimulus increment divided by the magnitude of a standard stimulus gives a constant value. Weber demonstrated that this finding held true for virtually all physical stimuli, such as light, sound, and pressure.

Fechner was very excited by Weber's work and extended it to measure psychological sensation, rather than stimulus magnitude [13]. Fechner defined sensation as the conscious experience that accompanied the brain's perception of external stimuli. By evaluating subjects' perceptions of stimulus changes, he concluded that Weber's "just noticeable differences" were equal throughout the range of sensation, and so they could be used as units for measuring internal conscious experience. Fechner noted that the magnitude of the perceived sensation was proportional to the log of the stimulus magnitude; as one's sensation of a stimulus increased linearly, the value of the physical magnitude increased exponentially. Fechner's extension of Weber's formula became known as Fechner's Law, which stated

$$S = K \log(R),$$

where S represents sensation magnitude, and K and R are from Weber's formula. He used this model to develop the concept of a "sensation scale", which described the internal psychological processes underlying perception.

Fechner is sometimes referred to as the first psychometrician. *Psychometrics* is the study of the measurement of "psychological characteristics such as abilities, aptitudes, achievement, personality traits, skills, and knowledge" [2, p. 93]. Fechner's obstinate work on measuring sensation illustrated that it was possible to measure unobservable processes within the human psyche. Today, the field of psychometrics includes educational testing, measurement of attitudes and perceptions, personality testing, opinion surveys, health inventories, and other forms of psychological measurement. Psychometric techniques created for measuring the human mind have also proven useful in biometrics and other fields that deal with measuring the intangible.

Psychological Scaling

Measurement involves **measurement scales**. The scales involved in most physical measurement are so widely known and accepted (e.g. kilograms, meters, liters) that there is little controversy surrounding the measurement process. However, in psychological measurement, the scales are more nebulous. On what type of scale can we measure the human mind? Fechner's work provided the first psychological scale measured in units of "just noticeable differences". L.L. Thurstone, another early psychometrician, extended Fechner's work to measure qualities such as attitudes in units similar to these [48]. Later, Rensis Likert introduced the method of summated ratings that measured attitudes along ordered integer scales [24]. Today, psychological measurement involves a variety of scales including those used in educational testing, such as the SAT score scale; in intelligence testing such as the IQ scale; in

personality testing such as the Minnesota Multiphasic Personality Inventory (MMPI) [7]; and in health psychology such as the Jenkins Activity Survey [18]. These tests are measured on very different scales. To understand how psychological scales are developed, the concept of measurement itself must first be understood.

Measurement and Measurement Scales

In its most general sense, *measurement* is the process of assigning numbers to objects according to a set of rules [43]. The set of rules used to make these assignments is called the *measurement model*. The numbers used to classify objects within the measurement model comprise a *measurement scale*. There is a continuum of measurement scales. Selection of an appropriate scale is motivated by both the purpose of the study and the nature of the variable(s) being measured. For example, if the purpose of an assessment is to classify people into one of two stress behavior profiles such as "type A" (high-stress) or "type B" (low-stress), the nature of the measured variable required is dichotomous. However, if the purpose of the assessment is to determine the *degree* of stress experienced by an individual, the nature of the measured variable must be continuous. The variables measured in these two situations are both related to stress. However, because these variables differ qualitatively from one another, and because there are different purposes associated with each assessment, each situation requires a different measurement model and a different measurement scale. For example, the type A/type B classification might be accomplished using the Jenkins Activity Survey, while degree of stress experienced may be measured using galvanic skin response or heart rate.

Stevens [43] provided a taxonomy of measurement scales by describing in detail four specific levels on the measurement continuum. The lowest level, called the *nominal scale*, is used for assigning numbers to objects that form mutually distinct categories (*see* **Nominal Data**). No order relations exist between the categories to suggest that one category possesses more of the attribute measured than another. Ethnicity and political party affiliation are examples of attributes measured on a nominal scale. Asians and Latinos would be assigned different numbers in the measurement process; as would Democrats, Republicans, Independents, and Conservatives. Thus, the nominal scale of measurement is essentially numerical coding. The coding rule is to assign the same number to objects in the same category and different numbers to objects in different categories.

The second level of measurement is called the *ordinal scale*. These scales include an inherent logical order among the mutually exclusive categories (*see* **Ordered Categorical Data**). Unlike the nominal scale, the magnitude of the numbers assigned to the different categories correspond to meaningful differences in magnitudes of the attribute measured. For example, on a coma inventory, patients who are assigned higher scores exhibit superior physiological responses than patients who are assigned lower scores. School grades on a letter scale (A, B, C, D, E) provide another example of an ordinal scale. Similarly, higher scores on the Apgar assessment (a test of the health of newborns) indicate better physiological functioning among newborns than lower scores. However, an important feature missing from an ordinal scale is equal interval widths between categories. That is, we cannot say that the difference in physiological functioning between Apgar scores of seven and eight is the same as the difference in physiological functioning between Apgar scores of three and four. Similarly, on rating scale inventories, we cannot say the difference between "strongly agree" and "agree" is the same as the difference between "strongly disagree" and "disagree". Such statements are reserved for variables measured on interval or ratio scales.

Interval scales have the property of equality of intervals on the scale that correspond to distinct intervals of the characteristic measured. This property allows for differences between scores to be interpreted with respect to the magnitude of the measured variable. For example, when one measures temperature in degrees Centigrade one is using an interval scale. The difference between 10°C and 20°C is the same, in some physical sense, as the difference between 80°C and 90°C. Note that the same thing cannot be said of the perceived sensation of warmth. To determine equal intervals in that scale we would need to take a logarithmic transform of these temperatures (following Fechner's Law). This illustrates how an interval scale in one realm may not be in another. Note also that it is, in general, meaningless to make statements like "80°C is twice as warm as 40°C". In no sense is this true. If we need to make such inferences we need a stronger scale.

The fourth level of measurement, the *ratio scale*, describes those scales commonly used to measure physical attributes. Ratio scales possess all the qualities of the lower scales (mutually exclusive categories, inherent logical order, and equality of intervals) as well as an absolute zero point. This absolute zero point corresponds to the absence of the characteristic measured. For example, in measuring blood alcohol level, a reading of zero indicates the absence of alcohol in the bloodstream. Similarly, before a patient steps on a triple-beam scale, the scale should read zero, which corresponds to the absence of weight on the platform. The *absolute zero* property of a ratio scale is important because it allows for relative statements to be made across measurements. Examples of such statements are "the patient's heart rate has doubled since admission", "Alice is half as tall as Eduardo", and "Melanie's range of motion in her right knee is twice her range of motion in her left knee". Note how these types of statements could not be made with respect to a variable such as intellectual functioning, which is measured on a scale without an absolute zero point.

The nature of the measurement scale must be considered when analyzing and interpreting measurement data. As illustrated above, the level of the measurement scale places restrictions on the types of inferences that can be drawn from the measurement process. For example, when a nominal scale is used, the numbers assigned to the mutually exclusive categories provide no information regarding the magnitude of the variable being measured. Consider measuring ethnicity using a nominal scale where Blacks are coded "1", Hispanics are coded "2", and Whites are coded "3". In looking at these "scores", we could not infer that Whites have more ethnicity than the other groups just because a higher number was assigned. We could easily *transform* this scale by recording Blacks as "37", Hispanics as "12", and Whites as "-1". This transformation of the original scale is admissible because the fundamental property of the scale, mutually exclusive categories, holds across the transformation. All subjects who are in the same ethnic group are assigned the same number. However, calculation of statistics such as the "mean ethnicity" is affected by the scale transformation. This issue is not restricted to the nominal scale. Consider the measurement of health care satisfaction on a five-point Likert scale. The same order relations may not hold if the scale were transformed to a three- or 12-point scale. Different measurement scales have different types of admissible transformations. Thus, the nature of the measurement scale must be considered when analyzing measurement data and reporting the results (see [31], for a more comprehensive discussion of this issue).

The issue of an absolute zero point is another important scale feature that must be considered when interpreting measurement results. In most areas of psychometrics the measurement scales are below the ratio level, and so a zero point on the scale is either absent or arbitrary. Therefore, for these scores to be understood and interpreted appropriately, they must be referenced to external information. Incorporating external information into the score scale is an important area of psychometrics. However, before discussing how this is done, we first provide information regarding the purposes and models underlying psychological measurement.

Purposes and Uses of Tests and Inventories

Tests and inventories are systematic procedures for observing the characteristics of a group of objects and describing the characteristics using a numerical scale. There are many uses of tests and inventories; however, most are used for the purposes of making diagnoses or decisions about groups or individuals. For example, psychiatric inventories are used to place patients into treatment programs, public health surveys are used to monitor the quality of health of various groups of people, and medical tests are used to provide specific information regarding a patient's condition. Such tests can provide a great breadth of information, and so it is no surprise that their use is both pervasive and growing.

Most tests are standardized in some fashion so that extraneous factors do not interfere with the interpretation of test scores. *Standardized tests* refer to tests on which the content and administrative conditions are virtually the same for all test takers. Without standardization, test scores would vary depending on the specific content and characteristics of the testing situation. Thus, comparisons of groups and individuals would be difficult to make. Standardized tests provide a level playing field for making comparisons between individuals and groups who take the same test. For this reason, standardized tests and inventories provide more objective information than unstandardized measures. However, standardized tests and

inventories do not comprise a limited type of assessment. There is a plethora of testing formats and test questions (item types) that fall into this category. Standardized tests may differ markedly from one another on the basis of the item types and test formats used.

Item Types and Test Formats

There are several different types of items that appear on tests and inventories, as well as several different testing formats. A popular item type on surveys and educational tests is the multiple-choice item. This item type features a question or item stem, followed by a series of options. Only one option contains the correct answer, and so these items can be scored dichotomously (right/wrong). An advantage of multiple-choice items is that they can be scored mechanically, such as by using an optical scanner. Many other item types that are scored dichotomously, such as true/false items, matching items, and items requiring the "bubbling-in" of numerical responses can also be scored mechanically. In fact, the recent advent of computer-administered testing (described in more detail in the section "Contemporary Testing Practices" below) has allowed for mechanical scoring of many different item types on computer. For example, the US Medical Licensing Exam is scheduled to include "simulated patient" scenarios where medical licensure candidates are required to "treat" hypothetical patients on-line.

When the scoring of test takers' responses is automatic, the assessment is said to be scored "objectively". However, some item types can currently only be scored by human judges. These item types are said to be scored "subjectively". Examples of subjective item types are open-ended questions such as those found in structured interviews, essay questions, and some tasks presented in performance-based assessments (such as simulated patient tasks that are not computerized). The distinction between objectively-scored and subjectively-scored item types is often subtle. The key feature is whether an external grader is needed to interpret test takers' responses in order to assign a score. When external graders are used, the scores are typically assigned to responses according to a well-defined set of rules. The Apgar assessment for newborns presents an inventory that uses both objectively-scored and subjectively-scored items. The heart rate "item" of the Apgar is measured objectively by reading the baby's pulse, but the color "item" is measured subjectively by looking at the baby and discriminating between blue and pink hues.

Objectively-scored items, such as the multiple-choice item, reduce subjective elements that may contaminate scores associated with other item formats. For example, if a newborn's Apgar score depends on the perceptual acuity of the observing nurse, differences in acuity among nurses may become a source of measurement error. An additional advantage of many objectively-scored item types is that a broad range of content can be tested in a relatively short amount of time. Dozens of multiple-choice items, covering a variety of content areas, can typically be answered in a smaller amount of time than it takes to complete a performance task or write an essay. However, multiple-choice items and other objectively-scored formats also have disadvantages. First, they typically take longer to develop than open-ended items. Secondly, they do not allow test takers to provide creative responses. For this reason, open-ended item formats, which usually require external judges for scoring, are used to test skills that are not conducive to testing via objectively-scored formats.

Recent advances in computer technology and measurement are providing mechanisms for scoring new, complex item formats objectively. Within the near future it is likely that behavioral measurements such as interpersonal skills and psychological well-being, which were thought to be beyond objective measurement, will be scored mechanically. Thus, physical and monetary resources, such as access to these emerging technologies, is likely to be a key determinant in the selection of different item types and testing formats. Before discussing the relative utility of different test and item formats, fundamental measurement concepts and models that are critical to all item formats are presented. We return to item and test format differences in a later section.

Fundamental Measurement Concepts

In this section we discuss three critical notions: **reliability**, generalizability, and **validity**. Psychometricians have developed increasingly sophisticated theoretical conceptions of these notions as well as the statistical machinery to allow them to serve practical purposes.

Reliability

A key issue in all measurement is the stability of the particular measure under consideration. In psychometrics we are principally interested in test scores and their stability is usually denoted by the term *reliability*. The technical meaning of this term is closest to the dictionary meaning "giving the same result on successive trials" and refers to the degree to which test scores are free from errors of measurement. There are many ways to characterize the reliability of a measuring instrument, but a common one is as a correlation between two independent measurements of the same phenomenon. For example, suppose we are trying to ascertain the reliability of a yardstick as an instrument for use by humans in measuring objects of lengths ranging from an inch to a yard. We might choose a few dozen objects of different apparent lengths to be measured then go and measure them all. After measuring them we might then go and measure them all a second time and correlate the first set of measurements with the second. This correlation would be an estimate of the reliability of the measurement process utilizing the yardstick. See Feldt & Brennan [14] for the full story.

It is easy to see that there are a number of difficulties with this sort of operational approach to reliability. For example, how does the estimate vary with different individuals carrying out the measurement process or with a different set of objects to be measured? It was in reaction to these kinds of problems that Cronbach et al. [11] proposed generalizability theory.

Generalizability

Early approaches to estimating the reliability of a measure or instrument yielded a simple number, often termed the *reliability coefficient*. Consequently, none of the different sources of variability (i.e. unreliability) was distinguished from another. It is often the case, however, that a more detailed accounting is needed of the contributions made by these different sources of variability. This leads naturally to a decomposition of the variance into its component parts.

In the ruler example described previously, we would want to know what was the variance due to measurers as well as the variance due to the characteristics of the set of objects being measured. In other circumstances we might want to measure the variation associated with the use of different rulers, or at different times or temperatures. Each aspect or facet of the measurement process adds its own contribution to the total measurement variance. By examining this decomposition we can state more precisely just what we mean by measurement accuracy than is possible with a single reliability coefficient. Moreover, we can predict the extent to which the quality of the measurement process generalizes to other circumstances, e.g. different objects being measured or different people doing the measuring. The method derives its name from this capability. Note that by using this more powerful approach, it is possible to predict the effect on reliability if one or more sources of variation is eliminated.

To calculate these **variance components** requires a more extensive data-gathering effort than that required to calculate a reliability coefficient. To estimate the variance due to different measurers we will need to have a sample of measurers make measurements. To estimate the variance due to different-sized objects being measured, for example, we will need to have ratings on several samples of objects. Often the gathering of such information is difficult or expensive. Consequently, even though a variance components conception of variability is attractive, it may not always be practicable. See Brennan [5] for more details.

Validity

The basic notion is that a measurement process is valid if "it measures what it is supposed to measure". More specifically, the term *validity* refers to the appropriateness, meaningfulness, and usefulness of the specific inferences that are made from the measurement process. Note that the term validity does not refer to the process itself, but rather to *the inferences made from it*. This is a very important distinction. As an illustration, consider a process that yields a blood pressure measurement in its typical form of systolic over diastolic. An inference one might make on the basis of a particular measurement concerns the state of the individual's cardiovascular system. Another inference might be in regard to the individual's emotional state. Under ordinary circumstances one would expect the first kind of inference to have greater validity than the second kind.

Traditionally, validity of mental measurements has been treated within three distinct subcategories – content validity, predictive validity, and construct validity. They can be described briefly as follows:

1. *Content validity*. The extent to which the content of the measuring instrument samples the class of situations or subject matter about which conclusions are to be drawn. If we are interested in making inferences about a person's health, how many of the various components of health do we measure? And how well do we measure them?
2. *Criterion/predictive validity*. The extent to which an individual's future performance on one or more external variables (the criteria) is predicted from performance on the test. The predictive validity of parental height as a predictor of a child's adult height can be empirically determined.
3. *Construct validity*. The extent to which certain explanatory theories and concepts account for performance on the test. Construct validity refers to how well the inferences derived from test scores correspond to theoretical and practical notions of the construct presumably measured by the test. As an example, consider a mental test composed of mathematics items, verbal reasoning items, verbal analogies, reading and oral comprehension, and spatial visualizing. The extent to which all individuals align themselves identically on all of these various subtests is the extent to which the underlying unitary construct of "intelligence" is supported.

Messick [30] provides a more comprehensive description of validity theory, and is considered to be the seminal work in this area.

Formal Measurement Models

In the first section we described the various sorts of measurement scales and their associated characteristics. It is a common misconception that the strength of a measurement scale is determined by the experimental procedure used in the measurement. This is not the case. The strength of the scale is determined by the measurement model, and in particular by the model for parameter estimation that is applied to the data gathered from the experiment. Over the long history of psychometrics there have been many formal models proposed. In this section we describe two of them. The first, *true score theory*, is a straightforward linear model based on observed scores. The second, *item response theory* is nonlinear (but linear in the logit – see below) and bears a surface similarity to **dose–response** curves so familiar in **bioassay**. There is an important, indeed critical, difference between item response theory and bioassay. In bioassay, one typically plots survival rates against dosage. Since both quantities are observable, at least in principle, this is straightforward. Statistical issues arise as one postulates different models to describe (as parsimoniously as possible) the dose–response curve and then tries to estimate the free parameters of the model (*see* **Quantal Response Models**).

By contrast, in a typical application of item response theory one wants to plot performance on an item (say a question about the number of visits to a doctor made in the last year) against an overall health status variable. But the latter is not directly observable. Such variables are often referred to as latent variables (*see* **Path Analysis**). Each individual in the sample is assumed to possess a value of this latent variable and this (unknown) value is represented by a parameter to be estimated. In addition, models for the curve describing the relation between performance on the item and the latent variable are of interest. Progress can only be made when data on a number of such items are available – but the statistical problems are formidable because the "dose" values must be estimated at the same time as the parameters describing the set of "dose-response" curves for the different items.

True score theory is familiar and easy, but in most circumstances does not satisfy the requirements for an interval scale. Item response theory is more complex and may seem arcane, but provides much stronger measurement characteristics. At the moment at least this is the contemporary choice for most modern testing programs and it is being adopted for measuring medical outcomes [28, 38].

True Score Theory

In what follows we employ the terms "test" and "examinee" to conform with psychometric usage. However, the ideas apply equally well to any sort of assessment or inventory.

The fundamental idea of true score theory can be stated in a single simple equation:

$$\text{observed score} = \text{true score} + \text{error}. \tag{1}$$

This equation explicitly states that the score we observe on a test is composed of two components, the true score and an error term. The term *true score* has a very specific technical meaning. It is the average score that we would expect to obtain if the examinee retook exactly similar (parallel) forms of the exam very many times. The error term characterizes the difference between what is observed on a particular occasion and the unobserved long-term average. Such errors are considered to be random and hence unrelated to true score; that is, the distribution of the errors is the same regardless of the size of the true score. This definition requires the errors to have an average size of zero.

Repeating this same discussion in mathematical terms yields an analog to (1) for examinee i on test form j:

$$x_{ij} = \tau_i + e_{ij}, \tag{2}$$

where x_{ij} is the observed score for examinee i on test form j, τ_i is the true score for examinee i, and e_{ij} is the error for examinee i on test form j. These quantities have the following properties:

$$\mathrm{E}(x_{ij}) = \tau_i, \tag{3}$$

$$\mathrm{E}(e_{ij}) = 0, \tag{4}$$

$$\mathrm{cov}(\tau_i, e_{ij}) = 0. \tag{5}$$

In much of the discussion in the rest of this article it will be important to collect the scores for many examinees and to study the variability among those scores. It is commonly assumed that the scores of any examinee are uncorrelated with any other examinee. Utilizing (2) and these properties we can decompose the variance of the observed scores into two orthogonal components, true score variance and error variance, i.e.

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2. \tag{6}$$

Eq. (6) follows directly from the definitions of true score and error, but provides us with many tools to study test performance. Obviously we prefer tests whose error variance, σ_e^2, is small relative to observed score variance σ_x^2. A test with small error variance would measure an examinee's true score more reliably than one with a large error variance. We can characterize how reliably a test works by the ratio of error variance to observed score variance, σ_e^2/σ_x^2. If this ratio is close to zero, then the test is working well – the observed score has very little error in it. If it is close to one, then the test is working poorly – the variation in observed score is mostly just error. When the ratio is rescaled [see (7)] so that it takes the value one when there is no error and zero when it is all error, it is the test's *reliability*:

$$\text{reliability} = 1 - \left(\frac{\sigma_e^2}{\sigma_x^2}\right). \tag{7}$$

This representation of reliability is intuitively appealing, but is not in a useful form because it cannot be directly computed from observed data; although we can observe σ_x^2, we cannot observe σ_e^2. A slightly different conception, using the idea of parallel test forms, yields an observable quantity.

Before deriving this important equation we need a formal definition of parallel test forms. Specifically two forms, say form X and form X', are parallel if

$$\mathrm{E}(x) = \mathrm{E}(x') = \tau \quad \text{and} \quad \sigma_x^2 = \sigma_{x'}^2 \tag{8}$$

for all subpopulations taking the test, where x and x' are the scores on form X and form X', respectively.

The correlation of one parallel form with another, $\rho_{xx'}$, is

$$\begin{aligned}\rho_{xx'} &= \frac{\mathrm{cov}(x, x')}{\sigma_x \sigma_{x'}} \\ &= \frac{\mathrm{cov}(\tau + e, \tau + e')}{\sigma_x \sigma_{x'}} \\ &= \frac{\left[\begin{array}{c}\mathrm{cov}(\tau, \tau) + \mathrm{cov}(\tau, e) \\ +\mathrm{cov}(\tau, e') + \mathrm{cov}(e, e')\end{array}\right]}{\sigma_x \sigma_{x'}}.\end{aligned}$$

The last three terms in the numerator are zero, yielding

$$= \frac{\mathrm{cov}(\tau, \tau)}{\sigma_x \sigma_{x'}}.$$

But $\mathrm{cov}(\tau, \tau) = \sigma_\tau^2$, and from definition (8), $\sigma_x = \sigma_{x'}$ so that $\sigma_x \sigma_{x'} = \sigma_x^2$. Combining these results we obtain:

$$\rho_{xx'} = \frac{\sigma_\tau^2}{\sigma_x^2},$$

but since $\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2$ [from (6)], we can rewrite this as

$$\rho_{xx'} = 1 - \left(\frac{\sigma_e^2}{\sigma_x^2}\right). \tag{9}$$

Using a similar approach, it is possible to show that

$$\rho_{x\tau}^2 = 1 - \left(\frac{\sigma_e^2}{\sigma_x^2}\right)$$

so that

$$\rho_{xx'} = \rho_{x\tau}^2. \tag{10}$$

This result is important since $\rho_{xx'}$ is directly estimable from data, whereas $\rho_{x\tau}^2$ is not. How to estimate $\rho_{xx'}$ well is the subject of a great deal of work that we will only touch on here. One obvious way is to construct two parallel forms of a test, give them both to a reasonably large sample of appropriate people, and calculate the **correlation** between the two scores. That correlation is an estimate of the reliability of the test. But making up a second form of a test that is truly parallel to the first is a lot of work. An easier task is to take a single form, divide it randomly in half, consider each half a parallel form of the other, and correlate the scores obtained on the two halves. For obvious reasons such a measure of test reliability is called *split-half reliability*. This yields an estimate of reliability for a test similar to, but only half as long as, the test we actually gave. Some sort of adjustment is required. A second issue that must be resolved before using the split-half reliability operationally is to figure out how to split the test. Certainly all splits will not yield the same estimate, and we would not want to base our estimate on an unfortunate division. Let us consider each of these issues in turn.

The Spearman–Brown Formula for a Test of Double Length. Suppose we take a test X, containing n items and break it up into two half tests, say Y and Y', each with $n/2$ items. We can then calculate the correlation that exists between Y and Y' (call it $\rho_{yy'}$) but what we really want to know would have been the correlation between X and a hypothetical parallel form X' ($\rho_{xx'}$). A formula [(11)] for estimating this correlation was developed by Spearman [42] and Brown [6] and is named in their honor:

$$\rho_{xx'} = \frac{2\rho_{yy'}}{1 + \rho_{yy'}} \tag{11}$$

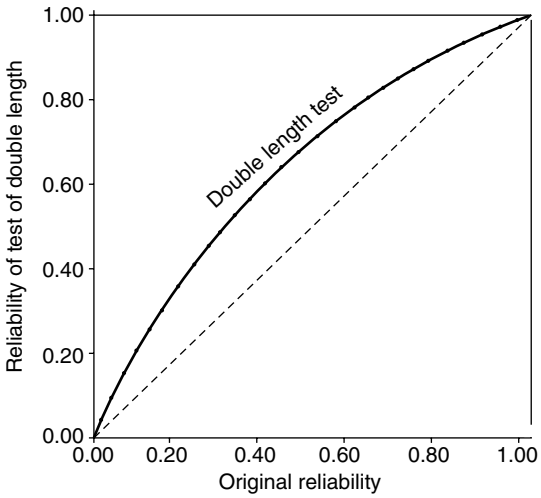

Figure 1 Reliability at double length as a function of reliability at unit length

A derivation of (11) follows directly from the characteristics of parallel tests and is given in [27, pp. 83–84].

To get an idea of how this expansion works, consider the two lines in Figure 1. The dashed diagonal line indicates equality of reliability between the original test and one of double length. The curved line shows the estimated reliability of a test of double length. Note that when reliability of the original test is extreme (0 or 1) doubling its length has no effect. The greatest effect occurs in the middle; a test whose reliability is 0.50 when made twice as long attains a reliability of 0.67.

Which Split Half? Cronbach's Coefficient α. There are many ways that we can split a test of n items in half (assuming that n is an even number). Specifically, there are

$$\frac{1}{2}\binom{n}{n/2} = \frac{n!}{2[(n/2)!]^2}$$

different ways that n items can be split in half. With all of these possible ways to divide the test in half it can be a difficult decision to determine which one we should pick to represent best the reliability of the test. One obvious way around this problem is to calculate all of them and use their mean as our best estimate. But calculating all of these for any test of nontrivial length is a very big deal. For example,

there are more than 63 trillion (63×10^{12}) split halves for 50 items! One way around this problem is due to Cronbach [10]. He derived a statistic that is a lower bound on the reliability of the test and is equivalent to taking the mean over all possible splits. Novick & Lewis [36] provided the conditions under which Cronbach's statistic is actually equal to the reliability of the test. Cronbach's statistic is usually called **Cronbach's α** in his honor, and is shown as the expression on the left-hand side of (12):

$$\frac{n}{n-1}\left[1 - \left(\frac{\sum_{i=1}^{n} \sigma_{yi}^2}{\sigma_x^2}\right)\right] = \alpha \leq \rho_{xx'}. \qquad (12)$$

A necessary and sufficient condition for Cronbach's α to be equal to the test's reliability and not just a lower bound, is that all the components making up the score have the same true score. The technical name for this condition is that all components be τ-equivalent.

To calculate α, we conceive of the test score x as being composed of the sum of n items Y_i. We calculate the variance of each item across all the examinees who took it and sum it up over all the items. Next we calculate the variance of the total score. The ratio of these two variances forms the core of Cronbach's α.

Estimating True Score. Under true score theory, the principal object of interest is the examinee's true score. A reasonable approach is to use the observed score as an estimate of the true score. Although this estimator has the virtue of being unbiased (that is, equal in expectation to its target) it can be improved upon.

The heuristic behind the improved estimator is similar to the familiar **"regression to the mean"** argument. Observed scores that fall far below the mean of the group are likely to have resulted, in part, from a relatively larger negative random error on that particular occasion. In a repeat test we would expect these observed scores still to fall below the mean, but not by nearly so great a margin. The greater the reliability of the test, the less the effect of the random errors.

The improved estimator, due to Kelley [19], requires three pieces of information: the observed score of the examinee, the mean score over the population of examinees, and a measure of the test's reliability. The estimate, $\hat{\tau}$, of the true score, τ, is given by

$$\hat{\tau} = \hat{\rho_{xx'}} x + (1 - \hat{\rho_{xx'}})\hat{\mu}, \qquad (13)$$

where $\hat{\rho_{xx'}}$ is an estimate of the reliability of the test, x is the observed score, and $\hat{\mu}$ is an estimate of the population mean. This result states that to estimate the true score from an observed score, the latter should be regressed toward the mean of the population by an amount related to the test's reliability. The amount of the "adjustment" to x increases as the reliability of the test decreases. Note too that for a fixed level of reliability, the amount of the adjustment is larger, the greater is the distance of the observed score from the population mean.

A Model for Error. The reliability coefficient that we have just discussed provides us with one measure of the stability of the measurement. It is often useful to have another measure that can be expressed on the scale of the score. For example, we would usually like to be able to present an estimate of someone's true score with error bounds: $\hat{\tau} = 1.2 \pm 0.3$. Such a statement raises two questions. The first is: What does $\pm$ 0.3 mean? The second is: How did you get 0.3?

There can be different answers to the first question depending on the situation. One common and reasonably useful answer is "it means that 95% of the time that someone whose estimated value of $\hat{\tau}$ is this value, her true score is within 0.3 of it".

The answer to the second question requires a little longer explanation. The explanation goes back to (9):

$$\rho_{xx'} = 1 - \left(\frac{\sigma_e^2}{\sigma_x^2}\right).$$

Simplifying this we can isolate error variance on one side of the equation and, by taking the square root, obtain an estimator for the *standard error of measurement*:

$$\sigma_e = \sigma_x(1 - \rho_{xx'})^{1/2} \qquad (14)$$

The uncertainty in our estimates of true score will be due to the variability of the error, e. To be able to provide a probability statement about the variability of the estimate we need to assume something about the distribution of the error. What sort of distribution we assume depends on the character of the scoring metric we use for the test. There are many choices

(see [27, Chapter 23] for an extended discussion of alternatives), but for the common situation of observed scores in the middle of the range of possible scores the assumption of a Gaussian distribution of errors is reasonable for most practical applications.

Item Response Theory

Item response theory is a family of mathematical descriptions of what happens when an examinee meets an item. It stems from early notions that test items all ought somehow to measure the same thing [26]. Item response theory formalizes this by explicitly positing a single dimension of knowledge or underlying trait on which all examinees rely, to some extent, for their correct response to all the test items. Examples of such traits are verbal proficiency, mathematical facility, jumping ability, or spatial memory. The position that each item occupies on this dimension is termed that item's *difficulty* (usually denoted b); the position of each examinee on this dimension is that examinee's *proficiency* (usually denoted θ). The item response theory model gives the probability of answering a question correctly in terms of the difference between b and θ (both of which are unobservable). The simplest model combines just these two elements within a logistic function (*see* **Logistic Distribution**). Because it characterizes each item with just a single parameter (difficulty $= b$) it is called the *one parameter logistic model*. This model was first developed and popularized by the Danish mathematician **Georg Rasch** (1901–1980) and so is often termed the **Rasch model** in his honor. We shall denote it the one parameter logistic model in this article to reinforce its position as a member of a parametric family of logistic models.

The one parameter logistic model is

$$p(\theta) = \frac{1}{1 + \exp[-(\theta - b)]}, \qquad (15)$$

where $p(\theta)$ is the probability of someone with proficiency θ responding correctly to an item of difficulty b. The interpretation of P is open for discussion. We tend to think of this P as arising from sampling. Specifically, if there is a large number of items all with the same difficulty b, a particular examinee would be able to answer some of them correctly, and some he would not. The proportion of items that a particular examinee with proficiency θ can answer correctly is given by (15). $P(\theta)$ is increasing in θ for a fixed b. For a fixed θ, $P(\theta)$ is smaller for larger values of b.

The structure of this model is most easily seen in a graph. Figure 2 shows a plot of what this function looks like for three items of different difficulty. These curves are called the *item characteristic curves* – they are also sometimes referred to as *trace lines or item response functions*. Note that the item characteristic curves for this model are parallel to one another. This is an important feature of the Rasch Model. It is informative to contrast the item characteristic curves shown in Figure 2 with those in Figure 3.

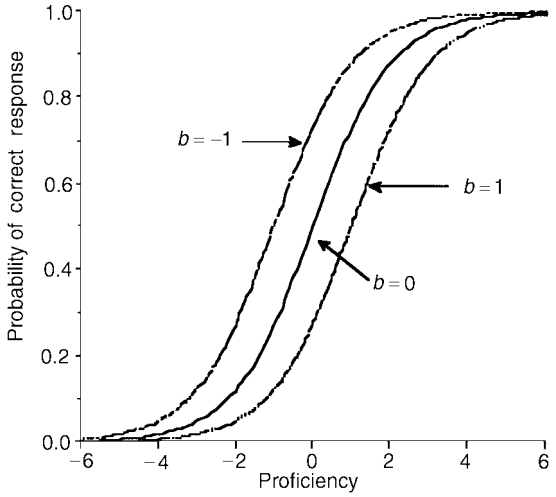

Figure 2 Item characteristic curves for the one parameterlogistic model at three levels of difficulty

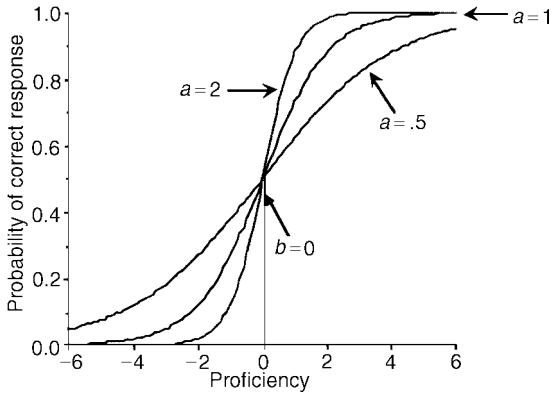

Figure 3 Typical item characteristic curves for the two parameter logistic model

The one parameter logistic model has many attractive features, and the interested reader is referred to the writings of Rasch [37] and of Wright [55] for convincing descriptions of its efficacy.

In many applications of item response theory to predetermined domains of items, it has been found that one does not get a good fit to the data with one parameter logistic model. A common cause of misfit is that the item characteristic curves of all items are not always parallel. When this occurs, there are two options. One is to delete items whose item characteristic curves show slopes that are divergent. The second is to generalize the model to allow for different slopes. This can be done through the addition of a second parameter for each item. This parameter, usually denoted a, characterizes the slope of the item characteristic curve, and is often called the *item's discrimination*. The resulting mathematical model, which now contains two parameters per item is called the two parameter logistic model and looks quite similar to the one parameter logistic model. Explicitly it is

$$P(\theta) = \frac{1}{1 + \exp[-a(\theta - b)]}. \qquad (16)$$

Once again our intuition is aided by seeing plots of the item characteristic curves achievable with this more general model. We have drawn three two parameter logistic model for items with the same b parameter ($b = 0$) in Figure 3, demonstrating the variation in slopes often seen in practice. Shown is an item that has rather high discrimination ($a = 2$), average discrimination ($a = 1$), and lower than average discrimination ($a = 0.5$).

The reason for calling the maximum value of the slope of the item characteristic curve the "discrimination" is that items with large slopes (i.e. high discrimination) are better able to distinguish between lower and higher proficiency examinees. For an item of high discrimination there is a relatively short interval along the proficiency scale where $P(\theta)$ moves from nearly zero to nearly one. Note that items of high discrimination are not very useful unless they are centered (i.e. have a b-value) in a region of the proficiency scale of interest to the examiner.

With the addition of the slope parameter, the two parameter logistic model greatly expanded the range of applicability item response theory. Many sets of items that could not fit under the strict equal slope assumption of the one parameter logistic model could be calibrated and scored with this more general model. However, this was not the end of the trail. So long as the multiple-choice item remains popular, the specter of an examinee getting an item correct through guessing remains not only a real possibility, but an event of substantial likelihood. Neither of the two models so far discussed allows for guessing – if an examinee gets an item right, it is assumed to provide evidence for greater proficiency. Yet, sometimes we see evidence in a response pattern that an examinee has not obtained the correct answer in a plausible fashion. Specifically, if someone gets a very difficult item correct, an item far beyond that examinee's estimated proficiency, then we can draw one obvious conclusion. The test is not unidimensional – the suspect item was answered using a skill or knowledge base other than the one we thought we were testing. This can be corrected by either modifying the test (removing the offending item) or generalizing the model to allow for guessing. The former fix is not likely to work, because different people may choose different items to guess on – eventually we will have to eliminate all difficult items. So, if we are to continue to use multiple-choice items, we have little choice but to use a more general model to describe examinees' performance on them. Such a model was fully explicated by Allan Birnbaum in Lord & Novick's [27] classic text. It adds a third parameter, c, which represents a binomial floor on the probability of getting an item correct. The resulting model, not surprisingly called the *three parameter model*, is shown explicitly below:

$$P(\theta) = c + \frac{1 - c}{1 + \exp[-a(\theta - b)]}. \qquad (17)$$

Once again we can get a better feel for the structure of the three parameter logistic model once we view a plot of a typical item characteristic curves. Such a plot is shown in Figure 4.

There is an indeterminacy in the estimation of the parameters of all these models that must be resolved one way or another. For example, suppose we define a new value of slope, say a^* as $a^* = a/A$, where a is the original value of the slope and A is some nonzero number. If we then define a new difficulty as $b^* = Ab + B$, where b is the original difficulty and B is some constant, a redefinition of proficiency as $\theta^* = A\theta + B$ would yield the result that $P(\theta, a, b, c) = P(\theta^*, a^*, b^*, c)$. Obviously there is no way to tell which set of parameters is better

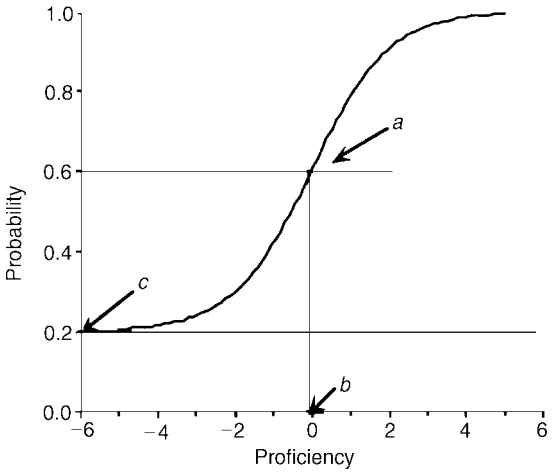

Figure 4 Typical item characteristic curve for the three parameter logistic model

because they produce identical estimates of probabilities of correct responses, and hence provide exactly the same fit to observed data. The usual way of resolving this indeterminacy is to scale proficiency so that θ has a mean of zero and a standard deviation of one in some reference population of examinees. This standardization allows us to understand at a glance the structure of our results. However, if we separately standardized with respect to two independent samples that were not randomly drawn from the same population, then we could not compare resulting item parameter estimates between them (because the first two moments of their proficiency distributions have been made to be identical artifactually). To do this we need to link these independent samples. A method for doing this is described in [50].

The three parameter logistic model is the item response theory model that is most commonly applied in large-scale testing applications, and so we confine the balance of our discussion to it. Estimating the parameters for a set of items under an item response theory model is usually called **calibration**. As indicated earlier, this is a difficult task since the examinees' proficiencies (θ values) must be estimated as well. Programs available for calibration are BILOG for **binary** (dichotomous) data [34] and MULTILOG for **polytomous data** [44].

Estimating Proficiency. In this section we assume that we have already calculated (somehow) the three parameters (a, b, and c) for each item, and have given this calibrated test to a sample of examinees. Our task is to estimate the proficiency, θ, for each of them. We do this using the method of **maximum likelihood**. In this discussion we alternate between the method of maximum likelihood and Bayes' modal estimates (*see* **Bayesian Methods**). In the way we use these terms, there is a clear connection between the two, because the *maximum likelihood estimator is just a Bayes' modal estimator with a uniform prior.*

To estimate proficiency we need to define three new symbols: $\mathbf{x}_i$ is the vector of item responses for examinee i, in which each response is coded 1 if correct, and 0 otherwise; it has elements $\{x_{ij}\}$, where the items are indexed by j. $\boldsymbol{\beta}_j$ is the item parameter vector (a_j, b_j, c_j) for item j and is a vector component of the matrix of all item parameters $\boldsymbol{\beta}$.

The conditional probability of $\mathbf{x}_i$ given θ and $\boldsymbol{\beta}$ is shown in (18):

$$P(\mathbf{x}_i|\theta_i, \boldsymbol{\beta}) = \prod_j P_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1-x_{ij}}, \qquad (18)$$

where $Q(\theta) = 1 - P(\theta)$.

The genesis of (18) should be obvious with a little reflection. It is merely the product of the model-generated probabilities (the item characteristic curves) for each item. The first term, $P(\theta)$, in the equation reflects the item characteristic curve for correct responses (when $x = 1$); the second term, $Q(\theta)$, for incorrect responses (when $x = 0$). Sometimes this is better understood graphically. Suppose we consider a two-item test in which an examinee gets the first item correct and the second item incorrect. The probabilities for each of these occurrences are shown in the top and middle panels of Figure 5, respectively.

In order for this product to represent validly the probability of a particular response vector, the model must be true and the item responses must be conditionally independent (*see* **Statistical Dependence and Independence**). Conditional independence is a basic assumption of most item response theory models. It means that the probability of answering a particular item correctly is independent of responses to any of the other items once we have conditioned on proficiency, θ. This assumption is testable [39], and when it is violated tends to yield overestimates of the accuracy of estimation [46].

If we know $\boldsymbol{\beta}$, the item parameters, we can look upon (18), for a fixed response pattern $\mathbf{x}_i$, as the likelihood function $L(\theta|\mathbf{x}_i)$ of θ given $\mathbf{x}_i$; its value at any value of θ indicates the relative likelihood that

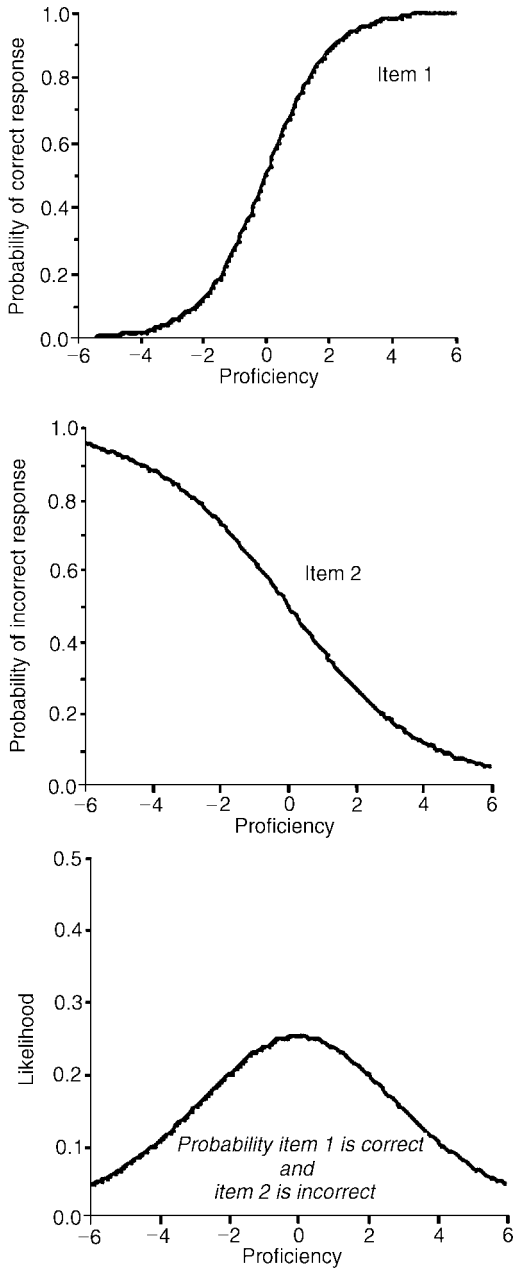

Figure 5 Example of how item characteristic curves multiply to yield posterior distribution of proficiency

$\mathbf{x}_i$ would be observed if θ were the true value. Eq. (18) thus conveys the information about θ contained in the data, and serves as a basis for estimating θ by means of maximum likelihood or Bayesian procedures. The *maximum likelihood estimate* of θ is merely the mode of the likelihood. Stated graphically, in terms of Figure 5 it is the value of θ associated with the highest point on the likelihood (the bottom panel in the figure). This likelihood was obtained by multiplying the curve in the top panel by the curve in the middle panel. The estimation methods commonly used are variations on this theme.

Another common method for estimating proficiency is the *Bayes modal estimate*. It is based upon the posterior distribution,

$$p(\theta|\mathbf{x}_i) \propto L(\theta|\mathbf{x}_i)p(\theta), \tag{19}$$

where $p(\theta)$ expresses knowledge about θ prior to the observation of $\mathbf{x}_i$. Thus, this is commonly called the **prior distribution** of θ. To accommodate the prior into the estimation scheme, we merely treat the prior as one more item and multiply it in, along with everything else. If we have no prior information whatsoever, then $p(\theta)$ has the same value for all θ – a "noninformative prior" – and the posterior distribution for θ is simply proportional to the likelihood function. Alternatively, an "informative prior" has a more profound effect; this is shown graphically in Figure 6. The prior used in this illustration is a standard normal distribution, and is shown in the top panel (labeled N). Note that in this example (taken from [47]), the examinee has taken a two-item test (labeled Items A and C), and has gotten both correct. The posterior is labeled P. If we multiply just these two item characteristic curves together to get the probability of occurrence of both events (their likelihood), then we would obtain a curve much like that shown at the bottom of Figure 5. This curve is an important component of the *posterior distribution* of proficiency because, when multiplied by the *prior distribution* of proficiency, it shows the likelihood of various values of proficiency after (posterior to) the examinee responds. The bottom panel of Figure 6 shows this posterior distribution.

Although finding the value of θ that maximizes (18) cannot be done in closed form, it is in principle straightforward using an iterative method like Newton–Raphson (*see* **Optimization and Nonlinear Equations**). In practice, many problems can arise. For example, if an examinee answers all items correctly (or all incorrectly) the estimate of proficiency will be infinite. In the three parameter logistic model there are a number of other response

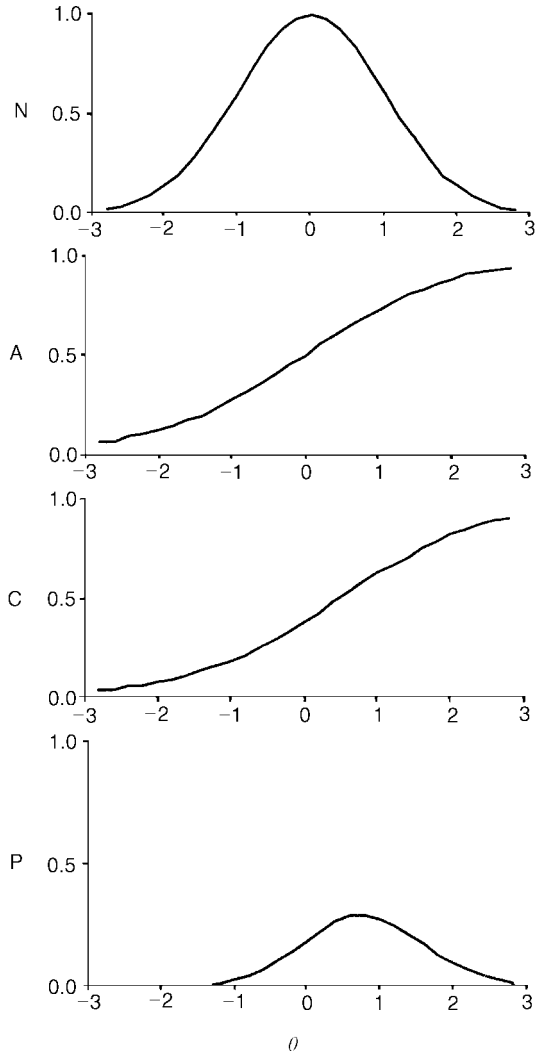

Figure 6 Schematic representation of Bayes' modal estimation, showing a normal prior (N), responses to two items (A and C), and the resulting posterior

patterns (e.g. many patterns at below chance levels) that would also yield infinite proficiency estimates. Also, the likelihood surface does not always yield a single mode; sometimes it can have a number of local extrema. In these cases, zeros may correspond to a local, but not the global, maximum of L, or even to a local minimum. The problems of infinite estimates are usually solved by utilizing a prior proficiency distribution (that is, using the kind of Bayesian estimator discussed next); those of local extrema are often resolved through the use of a "good" (i.e. close to the global maximum) starting value for the Newton–Raphson iterations (e.g. one based on a rescaled logit of percent correct).

On the Accuracy of the Proficiency Estimate

So far we have been concerned with obtaining a point estimate of an examinee's proficiency. The point we have adopted is the most likely value, the mode of the posterior distribution. But even a quick glance at the posterior in Figure 6 tells us that there is a substantial likelihood of other values. The width of the posterior distribution is commonly used to characterize the precision of the proficiency estimate. If the posterior is very narrow, then we are quite sure that the proficiency estimate we provide is a good one. If the posterior is broad, then we are less sure. In practice we can increase the accuracy of the estimate of proficiency by increasing the length of the test using *appropriately difficult* items. If we add an item that is much too easy for the examinees in question, then the item characteristic curve would essentially be a horizontal line in the neighborhood of these θs. Multiplying the posterior by a constant like this would do little to shrink the variance of the posterior. An identical thing happens if the item is much too hard. This is meant to emphasize the side condition of adding *appropriately difficult* items.

If the number of items an examinee has been administered is large, then the variance of the **likelihood** function can be approximated as the reciprocal of the **information** *function*:

$$I(\theta) = \sum_j \frac{(P_j')^2}{P_j(\theta)Q_j(\theta)}, \tag{20}$$

where P_j' is the first derivative of P_j with respect to θ. This expression has the attractive features of (i) being additive over items, and (ii) not depending on the values of the item responses. This means that for any given θ, one could calculate the contribution of information – and therefore to the precision of estimation of θ – from any item in an item pool. A typical information function is shown in Figure 7. The approximation for the estimation error variance of the maximum likelihood estimate of θ is less accurate for small numbers of items, but its advantages make it a popular and reasonable choice for practical work.

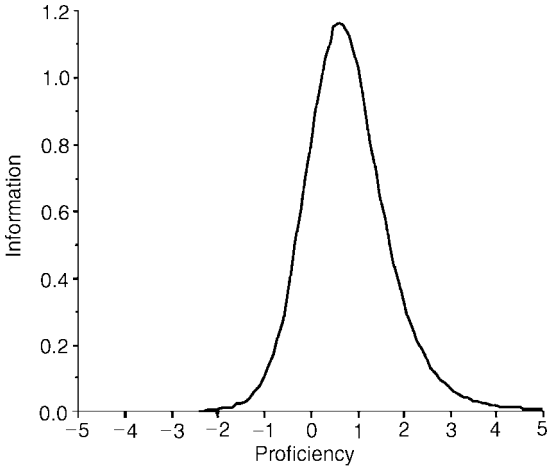

Figure 7 Typical three parameter logistic model information function

A similarly motivated expression can be used to indicate the precision of the Bayes model estimate of θ:

$$\text{var}^{-1}(\theta|\mathbf{x}_i) \approx \frac{I(\theta) - \partial^2 p(\theta)}{\partial\theta^2}.$$

The precision of the Bayes modal estimate, θ, typically exceeds that of the maximum likelihood estimate because the information from the item responses is augmented by a term that depends on the prior distribution. If $p(\theta)$ is normal with mean μ and variance σ^2, for example, then even before the first item is presented, one has at least this much knowledge about what an examinee's proficiency might be. In this case, $I(\theta) = 0$ since no items have been administered, but $\partial^2 p(\theta)/\partial\theta^2 = -\sigma^{-2}$. The impact of this prior information decreases as more items are given, however, because its contribution remains fixed while $I(\theta)$ increases with each item administered. Note also that if the prior is uniform, then the prior contributes nothing to measurement precision.

Estimating Item Parameters

Up to this point we have assumed that the item parameters were known. This is never the case in practice, when estimates must be used. Precise estimates with known properties are obviously desirable. In this section we outline a general framework for item calibration, briefly discuss the pros and cons of a variety of item calibration procedures that have been used in the past, and describe, in some detail, a Bayesian variation of the method of marginal maximum likelihood.

Notation and General Principles. The probability of observing the response matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ from a sample of N independently responding examinees can be represented as

$$P(\mathbf{X}|\boldsymbol{\theta}, \beta) = \prod_i P(\mathbf{x}_i|\theta_i, \boldsymbol{\beta}) = \prod_i \prod_j P(\mathbf{x}_{ij}|\theta_i, \beta_j), \tag{21}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ and $\boldsymbol{\beta} = (a_1, b_1, c_1, \ldots, a_n, b_n, c_n)$ are all considered unknown, fixed parameters. The continued product over items for each examinee is understood to run over only those items administered to that examinee. After responses have been observed, (18) is interpreted as a likelihood function for θ and β and serves as the foundation for item parameter estimation.

There are three commonly used approaches to item parameter estimation: *joint maximum likelihood* maximizes the likelihood depicted in (18). The other two approaches are *conditional maximum likelihood* and *marginal maximum likelihood*. The last is the method of choice. Conditional maximum likelihood is only possible for the one parameter logistic model, and even there is so computationally intensive as to be impracticable in many situations. It is never used for estimation in the three parameter logistic model and we do not dwell on it further. The interested reader is referred to [53] for details on this procedure. Before we go on to describe marginal maximum likelihood, we briefly discuss joint maximum likelihood.

Joint maximum likelihood estimates are obtained by finding the values of each β_j and each θ_i that together maximize (21). This is done by applying exactly the same ideas discussed earlier in the context of estimating $\boldsymbol{\theta}$. Direct maximization of (21) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ jointly often proves unsatisfactory for a number of reasons. First, for a fixed number of items, these estimates of $\boldsymbol{\beta}$ are not consistent in the number of examinees; that is, the expected values do not converge to their true values. Secondly, because each of the many $\boldsymbol{\theta}$ values is poorly determined when examinees take relatively few items, numerical instabilities can result. Maximizing values of item parameters may thus yield results that are unreasonable or even infinite.

Remedies leading to finite and reasonable estimates of $\boldsymbol{\beta}$ under the three parameter logistic model require information beyond that contained in the item responses, and structure beyond that implied by the item response theory model. Let us now discuss a Bayes' modal solution, extending ideas introduced in connection with Bayes' modal estimation of $\boldsymbol{\theta}$. Prior distributions for both examinee and item parameters are required. It is perhaps best to develop the solution in two stages.

Let $p(\boldsymbol{\theta})$ represent prior knowledge about the examinee distribution, assuming we have no additional information to lead us to different beliefs for different examinees. We treat $p(\boldsymbol{\theta})$ as known a priori, but it can be estimated from previous data, or even from the same data as those from which the item parameters are to be obtained (*see* **Empirical Bayes**). Consistency and increased stability follow if maximizing values for $\boldsymbol{\beta}$ are obtained after marginalization with respect to $p(\boldsymbol{\theta})$. That is, marginal maximum likelihood estimates of $\boldsymbol{\beta}$ maximize

$$L(\boldsymbol{\beta}|\mathbf{X}) = \int P(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta}, \tag{22}$$

or, more expansively,

$$L(\boldsymbol{\beta}|\mathbf{X}) = \prod_i \int p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}.$$

Numerical procedures for accomplishing marginal maximum likelihood estimation are described by Bock & Aitkin [3], Levine [22], and Samejima [40]. Without further precautions however, neither reasonable nor finite item parameter estimates are guaranteed.

Let $p(\boldsymbol{\beta})$ represent prior knowledge about the item parameter distribution, again assuming for the moment that we have no additional information that leads us to hold different expectations among them. We obtain a posterior distribution for item parameters by multiplying L by $p(\boldsymbol{\beta})$:

$$p(\boldsymbol{\beta}|\mathbf{X}) \propto L(\boldsymbol{\beta}|\mathbf{X}) p(\boldsymbol{\beta}). \tag{23}$$

Bayes' model estimates of $\boldsymbol{\beta}$ are the values that maximize (23) [32]. If a proper distribution $p(\boldsymbol{\theta})$ has been employed for examinees, and a proper and reasonable distribution $p(\boldsymbol{\beta})$ has been employed for items, then the resulting Bayes' modal estimates of $\boldsymbol{\beta}$ would be stable, reasonable, and consistent. Posterior means can also be employed, but modes are more often used because of their ease of calculation. When estimating item parameters, just as when estimating proficiency, one typically gets indices of the precision with which they have been estimated. Under ordinary marginal maximum likelihood, one gets standard errors of item parameter estimates; under Bayesian procedures, one gets posterior standard deviations.

The assumption of **exchangeability** among items (i.e. using the same prior distribution for all items) can be relaxed if some items have been administered previously. The prior distributions for these items can then be determined by the results of previous estimation procedures, taking the forms of distributions concentrated around previous point estimates.

Simulation studies have always been a prerequisite to using any estimation scheme. Asymptotic properties such as consistency do not necessarily characterize estimators' behavior in samples of the size and nature encountered in many specific applications. Moreover, although it is sometimes possible to obtain satisfactory parameter estimates with any of the procedures mentioned (see [35] for a comparison of two methods), the accumulation of evidence suggests that marginal maximum likelihood, with suitably chosen priors, is the best method currently available.

Interpreting Scores on Tests and Inventories

The selection of a measurement model is of critical importance for facilitating measurement precision. However, provision of an accurate score does not guarantee that the score will be interpreted appropriately. Thus, another important area of psychometrics is incorporating meaning into test scores and promoting accurate score interpretation. This section describes two common approaches for attaching meaning to test scores and describes how these different approaches affect score interpretation.

Norm-referenced and Criterion-referenced Approaches in Test Development

Consider a score of 112 obtained from an "adherence assessment" inventory, designed to predict whether a patient will stick to a critical treatment plan. Is this a "good" score? Does this score indicate that the patient is likely to adhere to her/his treatment regimen when

released from hospital? One way to make this score more informative would be to determine if it is above the average score of all patients who have taken this inventory. A related question is: Is the score at or above the average score of all patients who have stuck to their treatment plan? Interpreting scores in this fashion illustrates a *norm-referenced* interpretation. That is, norm-referenced scores are interpreted with respect to the scores attained by others.

Perhaps a more important question regarding the mysterious score of 112 is: Does this score signify that the patient should be retained in the hospital? Given this question, the performance of others on the inventory is much less relevant. The primary concern is whether the score is associated with the criterion of adherence to treatment. Interpretations attached to test scores on the basis of external criteria are called *criterion-referenced* interpretations. Meaning is attached to the scores by associating significant characteristics of the attribute measured with specific test scores. Both norm-referenced and criterion-referenced interpretations attach meaning to test scores, but they do so in different ways.

Norm-referenced interpretations describe test results in relation to the performance of one or more specific reference groups, called *norm groups*, who took the same test. Using our previous example, if we were told the average score on the adherence test was 100, we would know that a score of 112 is above average. But to what norm group does this average refer? If the average score of 100 refers to a population of patients who adhered to their treatment plan, then we may be impressed with a score of 112. Contrariwise, if the average score refers to a population of patients who did not adhere to their treatment plan, then a score of 112 may indicate that the patient is not likely to maintain treatment. Thus, in interpreting norm-referenced test scores, the appropriateness of the norm group for making inferences about a particular test taker must be considered.

Criterion-referenced interpretations describe test scores with respect to preestablished and well-defined relationships between test scores and external criteria associated with the characteristic measured. For example, the criterion variable "number of breaths taken per minute" may be used to measure respiration. Respiration is the underlying characteristic of interest, but meaningful standards of respiratory status are established on the basis of the number of breaths taken per minute. Using this criterion-referenced approach, interpretations such as "if patient takes less than seven breaths per minute, oxygen should be administered" can be made. In this situation, the breathing "scores" of other patients are irrelevant. All scores are interpreted with respect to the externally-defined standard of "less than seven breaths per minute".

Standards on criterion-referenced tests are developed using subject-matter experts to define carefully the variables to be measured, and to determine scores that reflect significant levels of the characteristic measured. These standards need not be static. For example, only a few years ago a cholesterol level of 240 was considered "normal". However, as a result of updated population analyses and reports of the incidents of coronary disease, medical experts revised their definition of normal cholesterol to correspond to a "score" of 200.

Both criterion-referenced and norm-referenced information can be used to interpret a test score. For example, if a patient's hypertension is measured by taking her/his blood pressure, it is informative to know how far the patient's blood pressure departs from the average of a comparable norm group (e.g. 120/80), as well as whether the score is above some maximum or minimum criterion of healthy blood pressure.

Criterion-referenced inventories often have several subscores indicating different factors related to the purpose of the assessment. For example, in measuring arterial blood gas, five parameters are obtained (pH level of blood, percentage of CO_2, percentage of O_2, saturation of the hemoglobin molecule, and amount of H_2CO_{3+}). Similarly, the General Health Questionnaire-28 [16], which is an inventory designed to identify nonpsychotic psychiatric disorders, provides four subscores describing somatic symptoms, anxiety and insomnia, social dysfunction, and severe depression. The provision of such subscores facilitates accurate diagnosis and the development of treatment plans.

The differences between norm-referenced and criterion-referenced scores are important for evaluating the appropriateness of a test for a particular purpose. In norm-referenced testing the character of the norm group is a primary concern when making inferences about test scores. In criterion-referenced testing it is critical that the criterion domain is appropriate and clearly defined.

Understanding and Interpreting Test Scores

The norm-referenced/criterion-referenced distinction is important when drawing inferences from test scores. However, to facilitate appropriate interpretations of test results, other distinctions and other "types" of test scores must be understood. The most straightforward test score is the raw score, which is simply the sum of the individual scores associated with each test item. For example, the score on an anxiety inventory may simply be the sum of the number of symptoms marked by a respondent. In intelligence and educational testing, items are typically scored "correct/incorrect". In these situations the raw score indicates the number of items answered correctly. An alternative way of reporting test scores is to report a *percent score*, or *percent correct score*, which is calculated by dividing the number of points earned on a test by the highest possible score on the test. For example, on a dichotomously scored intellectual functioning inventory, a percent correct score of 85 indicates that the person answered 85% of the items correctly.

Although raw and percent scores are simple to understand, they have some major limitations. First, in many assessment situations individuals take different forms of a test or take different tests altogether. For example, if an individual is given a test of academic skills as well as a performance test of various aspects of physical fitness, inferences about relative ability in the two realms are very limited if one can only use raw or percent scores. This same problem manifests itself even within a single test battery, which is comprised of many subscales. Secondly, raw and percent scores are not useful for making comparisons across subscales because the intensity, quality, and difficulty of the items comprising the subscales will be different for each scale (e.g. answering 50% of the vocabulary items correctly is unlikely to have the same interpretation as answering 50% of the arithmetic items correctly).

Norm-referenced and criterion-referenced interpretations can be used to facilitate comparisons across different tests and subscales. Percentile **rank** scores can be used to interpret an individual's relative standing on each subscale with respect to a specific norm group. A *percentile rank score* is a norm-referenced score that represents the percentage of test takers in a norm group who scored at or below the score attained by the individual. Some tests report more than one percentile rank score. Different percentile rank scores refer to different norm groups. An isometric strength measure would have separate percentile rank scores for males and females. Percentile ranks are also often calculated separately for different age groups. Percent correct scores are often enhanced using criterion-referenced interpretations such as a thorough description of the test items. For example, a percent correct score of 90 on a manual dexterity test may indicate that a physical rehabilitation patient was able individually to pick up nine of ten bolts, the thinnest of which was 2 mm; but she/he was unable to pick up the smallest bolt, which was 1 mm wide.

Although raw scores, percentile rank scores, and percent correct scores are meaningful, many test scores are reported on a standard score scale to facilitate further their interpretation. *Standard scores* are scores reported on a scale that has a predetermined mean and standard deviation. An individual's standard score represents her/his distance from the mean of a norm group in terms of the standard deviation of the score scale. Any raw score or percent correct score can be transformed to a standard score (using linear or nonlinear procedures as described earlier in the article). Standard score scales differ across tests as a result of the specific scale chosen by the test developers. For example, the SAT and ACT are both tests designed to assist colleges in making admissions decisions. However, these tests are reported on very different scales. Although two individuals taking these different tests could be compared with respect to their deviation from the mean of the score scale for each test, it must be remembered that different norm groups were used to derive these score scales, and so such comparisons are extremely limited. Standard score scales are designed to be most useful for interpreting scores associated with a single assessment instrument. However, when the same norm group is used across the different subscales comprising an assessment, relative comparisons across the subscales can be made with respect to this norm group.

Standard scores are also used to eliminate problems that may occur when comparing the performance of individuals who took different forms of a test. For example, to prevent cheating on a test that is used over a long period of time, or to prevent practice effects when tests must be administered to the same individual at different points in time, multiple test forms are needed. Unintended differences in difficulty or intensity between these forms would lead

to different raw scores or percent correct scores for the same individual. Thus, if raw or percent correct scores were used, then test takers would get different scores depending on the particular form they were administered. Through a process called *equating*, differences in test scores can be adjusted statistically so that "form effects" are eliminated when placed on the standard score scale (see [17] or [20] for descriptions of equating).

Percentile and Standard Score Bands

The standard error of measurement is another piece of information useful for interpreting test scores. As described in an earlier section, this index describes the average amount of error contained in a test score. By adding and subtracting one standard error of measurement from a particular score, a **confidence** band, or confidence interval, is formed around the test score. This band shows the probable limits of the score that would be obtained if the person were tested again. Many tests use this to report confidence bands for standard and percentile rank scores.

Setting Standards on Norm-referenced and Criterion-referenced Tests

As described previously, scores from criterion-referenced tests are usually interpreted with respect to well-defined aspects of the content domain. Many criterion-referenced tests incorporate preestablished standards of performance. Licensure tests are a conspicuous example. Test takers who score above the passing standard are awarded a license, while those who do not reach this standard are ineligible for licensure. Some tests have multiple standards. For example, in measuring brain waves, standards of electroencephalogram responses are used to classify patients into one of four states of excitement/relaxation (e.g. brain wave frequencies of 14 H_2–30 Hz signify an "excited" state [54]).

The process of determining the test scores that correspond to different categories of performance is called *standard setting*. Standard setting on norm-referenced tests is accomplished using percentile rank scores. For example, a scholarship may be awarded to the "top 10%" of examinees. The obvious disadvantage of norm-referenced passing standards is that classification decisions (e.g. pass/fail) vary primarily as a function of the characteristics of the norm group. How valuable is a bathroom scale to you if its estimate of your weight depended on who else used it that day?

Standards set on criterion-referenced tests circumvent this problem by using objectively established standards. One common criterion-referencing procedure uses subject-matter experts to scrutinize the items that comprise the test and estimate the probable performance of individuals representative of different diagnostic categories [8, 25]. These item judgments are summed over items and experts to derive a "cutscore" for each desired examinee classification.

A more empirical criterion-referenced approach is to use respondents' performance retrospectively on a criterion to establish standards on the test itself. For example, if a questionnaire is designed to predict relapse in a drug treatment program, then it could be given to all patients entering the program, and then one year later the average score of all patients who relapsed could be used as the cutscore for predicting relapse of patients entering the program in the future.

The procedures for determining criterion-referenced standards have their limitations. Standards developed using subject-matter experts are only as good as the particular group of experts employed. Standards established using external criteria are typically expensive of time and resources as well as being limited by the variation observed among the respondents used to establish the cutscores. Thus, there are tradeoffs among the different procedures for setting standards on tests. Criterion-referenced procedures have many advantages over norm-referenced procedures. Two obvious advantages are that attainment of a particular standard is not dependent on the performance of others, and also that score interpretation is especially straightforward.

Contemporary Testing Practices

Societal needs often evoke the potential inherent in science and technology. With respect to testing, there are current demands to measure attributes more thoroughly, and to measure characteristics that were previously considered untestable. Recent advances in computer technology, coupled with the newer measurement models such as item response theory, have allowed psychometricians to meet these demands. This section discusses some of these recent developments and their influence on testing in the health professions.

The two most consequential developments in psychometrics within the past decade are a renewed emphasis on "performance assessments" and delivering tests through computer-assisted technology. These two movements did not develop in complete isolation. In many cases, performance assessments became possible by capitalizing on innovations in computer technology. The new types of tests made possible by computer technology are called computer-based tests. These tests are redefining traditional testing practices along several dimensions.

Computer-based Testing

The idea of delivering tests by computer is not new. One of the earliest computer-based tests was the PLATO system designed to deliver both course material and tests to students at the University of Illinois and, eventually, around the country [1]. Like other early computer-based tests, the PLATO tests simply involved transferring existing paper-and-pencil multiple choice tests to an electronic database for later delivery via computer terminals. Two advantages of this immediately became obvious: (i) examinees could take the test whenever access to a terminal was available, and (ii) their scores were available immediately following conclusion of the test. Soon thereafter the idea of using a computer to "tailor" the test for each individual examinee emerged. This type of test, called a *computerized adaptive test*, is now widely accepted and is already implemented, or scheduled to be implemented, in many large-scale testing environments.

Computerized adaptive tests provide a test tailored to individual examinees by administering different questions to different examinees. In an administration of these an **algorithm** is programmed into the computer to select test questions that are most appropriate for the examinee currently taking the test. The NCLEX exam, which is the licensing examination for registered nurses in the US, provides a noteworthy example of how these work [56]. This computerized adaptive test involves an enormous pool of test items called an *item bank*. When a nursing candidate sits for this exam, a prespecified test item is administered. The next item administered depends on the examinee's response to the first item. If the candidate answers the item correctly, then a more difficult test item will be administered. However, if the candidate answers the test item incorrectly, then an easier item will be administered. The computer makes such item administration decisions on the basis of the candidate's responses to earlier items and other factors, such as coverage of the different content areas comprising the NCLEX. These decisions provide individualized tests, targeted to each candidate's ability level, that are essentially equivalent with respect to content coverage.

All computerized adaptive tests require a pool of content-valid items that are prearranged from simple to difficult. Typically, a middle difficulty item is administered first. If it is answered correctly, then the next item is chosen from among the items in the pool that are somewhat more difficult. If it is answered incorrectly, then the next item is chosen from among those items in the pool that are somewhat less difficult. If it too is answered incorrectly, then the third item is chosen from among those items that are even less difficult. The process continues until some suitable stopping point is reached. If the pool is sufficiently large, then each individual is likely to receive a unique sequence of items. The computer attempts to administer items that are of appropriate difficulty for each candidate (i.e. items that the candidate has about a 50–50 chance of answering correctly). Testing ceases either when a predetermined number of items have been presented or when the estimate of the uncertainty surrounding the current estimate of the candidate's ability falls below a prespecified threshold. From a psychometric perspective, items for which a candidate has about a 50–50 chance provide the most information about the candidate's skill, and so computerized adaptive tests increase testing efficiency by about 30%–40% [52]. Thus, these tend to be of much shorter duration than traditional paper-and-pencil tests of comparable accuracy.

The benefits of computerized adaptive tests include increased examinee motivation, reduced testing time, and increased precision of measurement. However, they increase the complexity of the scoring process. When examinees take such a test, they essentially take different tests. Therefore, they must include a method for placing these different test forms on a common score scale. Fortunately, item response theory methodology (described in an earlier section) provides the necessary foundation for an appropriate scoring model. The item response theory parameters are used to select items appropriate for

each candidate at a given level of proficiency, θ. Examinee's responses to early test questions are used to provide initial estimates, which govern the item selection process. The computer carries out these calculations rapidly so that the administer item $\rightarrow$ estimate $\theta \rightarrow$ administer new item cycle is transparent to the examinee. The θ scale onto which both items and examinees are mapped provides the common scale from which the test scores are derived.

In practice, computerized adaptive testing is more complicated to implement than described above. One reason is that usually there are multiple constraints operating on the process. For example, it may be required to ensure that each candidate is administered a specified number of items from different content domains during the course of the test. In that case the item selection algorithm must take into account both psychometric and substantive considerations in its choice. At present algorithms that handle nearly 200 constraints are operating on computerized adaptive tests such as the NCLEX and the Graduate Record Examination.

An additional advantage of computerized adaptive tests is that the tailored testing process lends itself naturally to diagnostic assessment. In this setting, real-time psychometric analyses identify potential examinee weaknesses and select future items to elucidate more specifically the nature of the potential weakness. In this manner a more detailed prescription for remediation is provided.

Many variants of computerized adaptive testing are currently operational. In the case of licensure tests such as the NCLEX and US Medical Licensure Examinations, interest centers on making a pass/fail decision at a particular cutscore. Some systems, such as the one implemented at Educational Testing Service (ETS) termed *computerized mastery testing*, involves building a library of mini-tests, called *testlets*, of the original full-scale paper-and-pencil examination [23, 49]. Candidates are administered two testlets, at which point a decision based on Bayesian analysis is made either to pass, fail, or continue testing. If the decision is to continue, then the candidate is administered another randomly chosen testlet and the Bayesian analysis is repeated using the additional information obtained from the testlet. Again there is a decision to pass, fail, or continue. A limit is set on the total number of testlets to be administered and a pass/fail decision is made at that point. Nearly 50% savings in testing time have been realized using this system.

Performance Assessment

With the advent of multimedia and other extensions of computer capabilities, the presentation of more complex stimuli, such as those involving video and audio, provided an impetus to move toward more complex item response formats. In most computer based test applications, multiple-choice items are typically administered. However, there is a current trend toward administering newer item formats that require the individual to construct a response rather than selecting from a number of predetermined choices [9]. These newer item formats fall under the rubric of "performance assessment". Performance assessments attempt to measure proficiency by having examinees perform a task, such as administering treatment to a simulated patient. As described earlier, multiple-choice and free-response item formats have both strengths and limitations. Thus, the features of each format are important before considering current innovations in performance assessment.

Despite recent criticisms [41], the multiple-choice item format has several attractive features. Multiple-choice items require less time to answer than free-response items, they tend to produce reliable scores, and they can be scored objectively (and inexpensively when automated scanning and sensing equipment is available). Because multiple-choice items take less time to answer than other item formats, they also provide greater ability to span the content domain tested in comparison with more time-consuming item formats. For example, in clinical settings there often is a need for instruments that can be administered quickly and do not demand much of the patient. For this reason, most psychological assessments (e.g. Beck Depression Inventory) and health inventories (e.g. MOS (Medical Outcome Study) Short-form Health Survey) rely on multiple-choice items.

However, reliance on multiple-choice items may limit the depth to which one can probe a domain. Thus, questions of test design are intimately tied to issues of validity. In addition to operational constraints, tradeoffs between broad (but perhaps superficial) coverage and in-depth analysis must be made within the context of the nature and the purpose of the assessment.

There is no universally accepted delineation of the response sets that qualify as raw material for performance assessments. One might consider a full range, extending from simple fill-in-the-blanks to short answer questions, to the kind of tasks involved in simulated patient examinations. As one moves to more extended and unconstrained response sets, accurate and reliable scoring is more difficult to achieve. It is necessary to develop scoring rubrics to guide the scorers or judges in their task. Constructing appropriate rubrics and effectively training judges to use them consistently is a demanding task that requires good judgment, experience, and considerable patience. Too often poorly executed scoring of performances has led to unreliability and reduced validity, thereby vitiating any of the hoped-for advantages of performance-based assessments [21].

Probes that are meant to elicit extended responses can present the investigator with other difficulties. One is the set of issues related to response style. To take a familiar example, a teacher may ask students to make a presentation to the class related to an assigned project. Some students may feel uncomfortable making such a presentation, either because of inexperience with the format, innate shyness, or cultural traditions. If the teacher is insensitive to these possibilities, then her assessment of student performance relative to standard learning goals may have substantially lowered validity. Similar considerations apply in clinical interviews. In other situations respondents may misinterpret the questions, may refuse to respond or respond in a perfunctory manner because of disinterest, fatigue, or frustration. While these factors may play a role when multiple-choice items are used, they are more likely to occur and to have more serious consequences when performance measures are used.

In some instances the response styles involved in performance assessments must be explicitly taken into account. For example, in scoring a Rorschach protocol, counts are taken of various kinds of responses. It is known that propensity to provide multiple responses is associated with various demographic characteristics (e.g. level of education) as well as with the psychological traits under study. Unless appropriate precautions are undertaken in the preparation and the analysis of data, misleading clinical interpretations could result. The Rorschach protocol is a good example of a response set so complex that at least five completely different scoring systems have arisen, all purporting to provide clinical indications for a comparable set of psychological states. Attempts to create a single comprehensive system have met with limited success [12].

Comparisons of the efficiency of information gathered between multiple-choice and performance assessment items typically reveal that the two formats are measuring the same construct, but that multiple-choice items are more efficient with respect to measurement precision [52]. However, when considering appropriate item formats, concerns of construct validity may be more heavily weighted than those of measurement efficiency.

It is often difficult to determine the relative worth of performance-based assessments empirically. As an example, consider a test administered to provide automobile licenses. In many states the test has two parts: a written multiple-choice section and a driving test. The latter part is a performance assessment. The two parts address different aspects of the standard set for granting a license. One aspect deals with knowledge of the rules of the road, appropriate laws, and so on; the other deals with the ability to drive the vehicle safely. It is certainly sensible to require that individuals meet both aspects of the standard. A factor analysis of the performance of a group of individuals on the two parts would very likely show a substantial positive correlation between the two. A positive correlation signifies that the ranking of individuals on the multiple-choice section is congruent with the ranking of individuals on the driving test. Nonetheless, the conclusion that one or the other of the tests is redundant is incorrect. It is important to ensure that the individual has met the standard for each aspect of the standard, which can be ascertained only by actually observing the performance on the two parts. In principle, one could use regression analysis (not correlation analysis) to estimate the level of performance on one part given performance on the other part and data from a large group on both parts. However in situations where meeting the standard carries some value, neither the test sponsor nor the majority of individuals would be likely to want to base the decision on just one component and a statistical prediction.

Despite their shortcomings, performance-based assessments are becoming increasingly popular [9]. Stringent test development procedures, appropriate scoring rubrics and procedures, and sophisticated scaling models have greatly improved the psychometric properties of many of these assessments [15]. It is

likely that the use of performance-based assessments will increase in the medical professions.

In terms of integrating performance assessment and computer based tests, two critical technologies must be in place before performance-based item formats can be delivered in practice and scored through computers. The first is the development of expert systems that can accurately score the responses. These systems, employing principles and techniques of **artificial intelligence**, exist for a number of different symbol systems including natural language, mathematics, engineering design, and graphic design [4]. In general, the cost and difficulty of building a suitable system is proportional to the complexity of the potential responses and inversely proportional to the number of constraints that are placed on the examinee. The second technology comprises the psychometric models that can be used to extract the maximum amount of information from these extended responses. Since constructed response items usually require considerable testing time, unless more useful information is obtained from each item the reliability and validity of the resulting test scores may be far below acceptable standards.

Unfortunately, measurement theory has not kept pace with the development of performance assessments. If responses to a probe can be assigned to one of a number of categories (ordered or unordered) then polytomous item response theory models [45] can be used as a measurement model. When a response set consists of a number of diverse performances (as might be the case on a certification or licensure examination), the measurement models employed typically do not take into account the multivariate character of the data. One exception is a class of multidimensional item response theory models [29] that were explicitly developed to deal with these types of data. However, these models are of increased complexity and have yet to be applied to operational test data. More recent developments employing Bayes inference networks [33] appear to hold great promise in providing a firmer foundation for the design and psychometric analysis of complex assessments.

With the problems and benefits associated with performance assessments, an appropriate conclusion is that greater thought should go into the type of application as well as the design of the probes so that more useful information can be generated. As Messick [30] argues in the context of educational assessment:

> There should be a guiding rationale akin to test specifications that ties the assessment of particular products or performances to the purposes of the testing, to the nature of the substantive domain at issue, and to construct theories of pertinent skills and knowledge (p. 27).

Thus, the decision to use performance-based assessment formats, multiple-choice items, or some combination of the two is dependent on the nature of the construct measured. The relative advantages and limitations of these item formats must be considered in the context of the reliability and validity of the scores they will provide as well as the practical constraints within which the assessment must exist.

The field of psychometrics has a short, but rich, history. Like many of the scientific fields that developed primarily during the twentieth century, psychometrics has experienced rapid growth. This article briefly presented the evolution of psychometric theory, described popular measurement models, and provided a snapshot of new directions in the field. Measuring psychological phenomena is a challenging, but rewarding endeavor. Galileo once proclaimed: "we must measure what is measurable and make measurable what cannot be measured". This goal continues to drive psychometricians to understand and measure abilities, proficiencies, attitudes, and other psychological constructs.

References

[1] Alpert, D. & Bitzer, D.L. (1970). Advances in computer-based education, *Science* **167**, 1582–1590.

[2] American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. American Psychological Association, Washington.

[3] Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm, *Psychometrika* **46**, 443–459.

[4] Braun, H.I. (1994). Assessing technology in assessment, in E.L. Baker & H.F. O'Neil, eds. *Technology Assessment in Education and Training*, Lawrence Erlbaum, Hillsdale, pp. 231–246.

[5] Brennan, R.L. (1983). *Elements of Generalizability Theory*. American College Testing Program, Iowa.

[6] Brown, W. (1910). Some experimental results in the correlation of mental abilities, *British Journal of Psychology* **3**, 296–322.

[7] Butcher, J.N. & Williams, C.L. (1992). *MMPI-2 and MMPI-A: Essentials of Clinical Interpretation*. University of Minnesota Press, Minneapolis.

[8] Cizek, G.J. (1996). Setting passing scores. [An NCME instructional module], *Educational Measurement: Issues and Practice* **15**, 20–31.

[9] Clauser, B.E., Subhiyah, R.G., Nungester, R.J., Ripkey, D.R., Clyman, S.G. & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts, *Journal of Educational Measurement* **32**, 397–415.

[10] Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* **16**, 297–334.

[11] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York.

[12] Exner, J.E. (1974). *The Rorschach: A Comprehensive System*. Wiley, New York.

[13] Fechner, G.T. (1860). *Elemente der Psychophysik*. Breitkopf & Hartel, Leipzig.

[14] Feldt, L.S. & Brennan, R.L. (1993). Reliability, in *Educational Measurement*, 3rd Ed., R.L. Linn, ed. Oryx Press, Phoenix, pp. 105–146.

[15] Fitzpatrick, A.R., Link, V.B., Yen, W.M., Burket, G.R., Ito, K. & Sykes, R.C. (1996). Scaling performance assessments: a comparison of one-parameter and two-parameter partial credit models, *Journal of Educational Measurement* **33**, 291–314.

[16] Goldberg, D. & Williams, P. (1988). *A User's Guide to the General Health Questionnaire*. Nfer-Nelson, Windsor.

[17] Holland, P.W. & Rubin, D.B. (1982). *Test Equating*. Academic Press, New York.

[18] Jenkins, C.D., Zyzanski, S.J. & Rosenman, R.H. (1979). *Manual for the Jenkins Activity Survey*. Psychological Corporation, New York.

[19] Kelley, T.L. (1947). *Fundamentals of Statistics*. Harvard University Press, Cambridge, Mass.

[20] Kolen, M.J. & Brennan, R.L. (1995). *Test Equating: Methods and Practices*. Springer-Verlag, New York.

[21] Koretz, D., Stecher, B., Klein, S. & McCaffrey, D. (1994). The Vermont portfolio assessment program: findings and implications. *Educational Measurement: Issues and Practice* **13**, 5–16.

[22] Levine, M. (1985). The trait in latent trait theory, in *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*, D.J. Weiss, ed., Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnesota, Minneapolis, pp. 41–65.

[23] Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test, *Applied Psychological Measurement* **14**, 367–386.

[24] Likert, R. (1932). A technique for the measurement of attitudes, *Archives of Psychology* **140**, 44–53.

[25] Livingston, S.A. & Zieky, M.J. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Educational Testing Service, Princeton.

[26] Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs* **61**, no. 4.

[27] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Addison Wesley, Reading.

[28] McHorney, C.A., Ware, J.E., Lu, J.F.R. & Sherbourne, C.D. (1994). The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions, and Reliability Across Diverse Patient Groups, *Medical Care* **32**, 40–66.

[29] McKinley, R.L. & Reckase, M.D. (1983). An Extension of the Two-Parameter Logistic Model to the Multidimensional Latent Space, *Research Report ONR83-2*. The American College Testing Program, Iowa.

[30] Messick, S. (1993). Validity, in *Educational Measurement*, 3rd Ed., R.L. Linn, ed. Oryx Press, Phoenix, pp. 13–103.

[31] Michell, J. (1986). Measurement scales and statistics: a clash of paradigms, *Psychological Bulletin* **100**, 398–407.

[32] Mislevy, R.J. (1986). Bayes modal estimation in item response models, *Psychometrika* **51**, 177–195.

[33] Mislevy, R.J. (1995). Probability-based inference in educational assessment, in P. Nichols, S. Chipman & R. Brennan, eds. *Cognitively Diagnostic Assessment*, Lawrence Erlbaum, Hillsdale, pp. 3–71.

[34] Mislevy, R.J. & Bock, R.D. (1983). *BILOG: Item and Test Scoring with Binary Logistic Models* (computer program). Scientific Software, Mooresville.

[35] Mislevy, R.J. & Stocking, M.L. (1987). A Consumer's guide to LOGIST and BILOG, *ETS Research Report 87-43*. Educational Testing Service, Princeton.

[36] Novick, M.R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements, *Psychometrika* **32**, 1–13.

[37] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Denmarks Paedagogiske Institut, Copenhagen.

[38] Roche, A.F., Wainer H. & Thissen, D. (1975). *Skeletal Maturity: The Knee Joint as a Biological Indicator*. Plenum, New York.

[39] Rosenbaum, P.R. (1988). A note on item bundles, *Psychometrika* **53**, 349–360.

[40] Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses, in *Principals of Modern Psychological Measurement*, H. Wainer & S. Messick, eds. Lawrence Erlbaum, Hillsdale, pp. 159–182.

[41] Shepard, L.A. (1992). Commentary: what policy-makers who mandate tests should know about the new psychology of intellectual ability and learning, in *Changing Assessments: Alternative Views of Aptitude, Achievement*

and Instruction, B.R. Gifford & M.C. O'Connor, eds. Kluwer, Boston.

[42] Spearman, C. (1910). Correlation calculated with faulty data, *British Journal of Psychology* **3**, 271–295.

[43] Stevens, S.S. (1946). On the theory and scales of measurement, *Science* **103**, 667–680.

[44] Thissen, D. (1991). *MULTILOG User's Guide* (Version 6). Scientific Software, Mooresville.

[45] Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models, *Psychometrika* **51**, 567–577.

[46] Thissen, D., Steinberg, L. & Mooney, J.A. (1989). Trace lines for testlets: a use of multiple-categorical-response models, *Journal of Educational Measurement* **26**, 247–260.

[47] Thissen, D., Steinberg, L. & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines, in *Test Validity*, H. Wainer & H. Braun, eds. Lawrence Erlbaum, Hillsdale, pp. 147–169.

[48] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review*, **34**, 273–286.

[49] Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: a case for testlets, *Journal of Educational Measurement* **24**, 185–202.

[50] Wainer, H. & Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation, in *Computerized Adaptive Testing: A Primer*, H. Wainer, D.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg & D. Thissen. Lawrence Erlbaum, Hillsdale, pp. 65–102.

[51] Wainer, H., Dorans, N., Flaugher, R., Green, B.F., Mislevy, R., Steinberg, L. & Thissen, D. (1990). *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum, Hillsdale.

[52] Wainer, H., Lukhele, R. & Thissen, D. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests, *Journal of Educational Measurement* **31**, 234–250.

[53] Wainer, H., Morgan, A. & Gustafsson, J.-E. (1980). A review of estimation procedures for the Rasch model with an eye toward longish tests, *Journal of Educational Statistics* **5**, 35–64.

[54] Williamson, D.A., McKenzie, S.J. & Duchman, E.G. (1988). Psychophysiological assessment, in *Handbook of Child Health Assessment*, P. Karoly, ed. Wiley, New York.

[55] Wright, B.D. & Stone, M.H. (1979). *Best Test Design*. MESA Press, Chicago.

[56] Zara, A. (1996). Overview of a successful conversion, in *Computer-Based Examinations for Board Certification: Proceedings from the 1996 American Board of Medical Specialties Educational Conference*, P.G. Bashook & E.C. Mancall, eds., American Board of Medical Specialties Education, Evanston, pp. 79–90.

STEPHEN G. SIRECI, HOWARD WAINER & HENRY BRAUN

Pulmonary Medicine

Respiratory disease (ICD 460–519), which excludes diseases of the lung not associated with respiration, such as lung cancer, accounts for about 15% of all deaths annually in the UK. The most common respiratory cause of death is pneumonia, accounting for 60% of all respiratory deaths, and this occurs mainly in those over 70 years old. The remainder are mainly due to chronic obstructive pulmonary disease (32% of all respiratory deaths). Other lung diseases are mainly associated with occupation, such as coal workers' pneumoconiosis, asbestosis, and byssinosis, but these are relatively rare. The principal obstructive diseases are chronic bronchitis, emphysema, and asthma. A patient can be assumed to have chronic bronchitis if sputum has been coughed up on most days of at least three consecutive months for more than two successive years, provided other causes have been excluded. It develops in response to long-term action of various types of irritant on the bronchial mucosa, the most important of which is cigarette smoke, but also as part of general air pollution. Emphysema means literally "inflation" in the sense of abnormal distension with air. Persistent inflammation in the airways, mainly as a result of smoking, causes irreparable damage to the alveolar septa. Most patients with pulmonary emphysema complain of exertional breathlessness. Asthma is associated with episodic attacks of wheeze and dyspnea. During an attack of acute asthma, the chest is held near the position of full inspiration, and expiration is difficult. Pneumoconiosis is an occupational disease that causes the lung to fibrose and is diagnosed from a chest radiograph by the profusion of small rounded opacities. It is graded into three main categories in accordance with a classification provided by the International Labor Office (ILO) [13]. Clinical diagnosis of respiratory diseases requires a clinical examination, and depends on a variety of signs and symptoms. However, simple questionnaires, such as the Medical Research Council (MRC) chronic bronchitis questionnaire [18, 19], and tests of lung function, are the main methods by which epidemiologic research in respiratory medicine has developed. Much useful information is given in [11] with many statistical issues covered in an appendix by Peto. Some general points on the presentation and analysis of data from respiratory studies are given in [2].

Lung Function Measurement

Historical Development

The early history of lung function measurement has been described by Cotes [6]. The volume of air that a person can inhale during a single deep breath was first measured by Borelli in 1679. In 1831, Thackrah showed the volume of air to be less in women than in men, and to be reduced among workers and other occupations due to the inhalation of dust. Hutchinson, in 1846, was the first to quantify lung volumes and designed a spirometer to measure *vital capacity* defined as "the greatest voluntary expiration following the deepest inspiration". He showed that vital capacity increased with height, decreased with age, and is reduced by excess weight and by disease of the lung.

Main Measures in Respiratory Medicine

There are a limited number of measures that are commonly used in **clinical trials** or epidemiologic studies. These include lung function tests such as *forced expiratory volume in one second (FEV_1), forced vital capacity (FVC), and peak expiratory flow rate (PEFR)*.

The FEV_1 is the volume of gas that is expelled from the lung over one second when the subject makes a maximal expiratory effort from a position of full inspiration. Other time periods are sometimes used, and indicated by the subscript. The *FVC* is the total volume of air expired until no more gas can be expelled from the chest after a full inspiration. Both are measured in liters using a spirometer. The peak expiratory flow rate (*PEFR*) is the maximum flow rate that can be sustained for a period of 10 ms and is measured in liters per second (l/s). It is another measure of ventilatory capacity [24] and is used widely in asthma studies.

Bronchial challenge tests give measures of bronchial irritability, related to asthma susceptibility, measured by the PD_{20} and PC_{20} that are derived from the FEV_1. A bronchoconstricting agent such as histamine, AMP or methacholine is inhaled in doubling doses until the FEV_1 falls 20% from baseline. The PD_{20} is the cumulative dose of the agent inhaled, whereas the PC_{20} is the particular concentration of the agent that caused the FEV_1 to cross the critical point.

Relationship of Disease to Lung Function

In chronic bronchitis, FEV_1 is reduced in many cases and the FEV_1/FVC ratio is subnormal. *PEFR* may be reduced but shows little diurnal variation. In emphysema, FEV_1 is reduced. The hallmark of asthma is diurnal variation in *PEFR*, with lowest values being recorded in the morning ("morning dipping"). Changes in FEV_1 due to bronchodilator drugs or corticosteroids are also an important indicator. The PD_{20} or PC_{20} is used widely in clinical and epidemiologic studies of asthma. It is useful because it is independent of diagnostic patterns; that is, the willingness of a physician to name the disease as asthma, and of symptom awareness, the importance placed on the severity of symptoms by the subject or the recognition of terms used in a questionnaire.

This article concentrates on some specific points of interest for the design and analysis of studies that involve measurements of respiratory function.

Particular Techniques in Respiratory Medicine

Estimating the FEV_1

Successive measurements of the FEV_1 of an individual will vary and a number of studies have been carried out to find what is the optimum combination of these. Because the test is a measure of maximal performance, one would have expected the maximum of a number of blows to be best. The Medical Research Council [18] recommended that the best procedure would be to require five blows, the first and second to be treated as practice attempts and to report the average of the third, fourth, and fifth blows as giving the individual's FEV_1.

Ferris [10], on the basis of three sets of data, suggested that the mean of the largest three of five measurements was best. Oldham & Cole [21] examined nine indices, including the two mentioned. From two data sets they concluded that in terms of repeatability for normal subjects there was little to choose between the various indices. However, differences were found in the measured rate of decline in FEV_1 using different indices over a period of 9.5 years. The index by Ferris gave values closest to the MRC definition over the 9.5 years and also gave the largest multiple **correlation** with age and degree of pneumoconiosis. They suggested that the mean of the largest three out of five blows was logically more sensible than the MRC definition and urged that it should replace the MRC index.

Estimation of the PD_{20} (PC_{20})

Bronchial reactivity is a measure of how the lung reacts to an inhaled agent such as histamine or methacholine. In general, people with asthma react by bronchoconstriction to lower doses than people without asthma, and so reactivity can be used to provide a standardized method of estimating the prevalence of asthma. Yan et al. [25] described a quick and simple method of measuring the PD_{20} that gave reproducible results under field conditions. The test stopped when the FEV_1 had fallen by more than 20% of the baseline value, or the 4.0 μmol dose had been given, or the subject asked to stop. The PD_{20} was estimated by interpolation. Chinn et al. [4] showed that linear interpolation was unsatisfactory and fitted a curve of the form:

$$\log_e(c - y) = a + bx,$$

where y is FEV_1, x is $\log_{10}$(dose), c represents mean FEV_1 before it is affected by histamine, b is the **regression** coefficient, and a is a curve position parameter. They estimated a, b, and c using a two-step procedure and showed that this exponential curve (*see* **Parametric Models in Survival Analysis**) fitted the data well. The advantage of estimating the PD_{20} by curve fitting is that it allows extrapolation of value above 4 μmol. Another parameter that can be derived from the above method is the slope b, and O'Connor et al. [20] showed how this could be useful in population studies. Peat et al. [22] described methods for measuring repeatability of the PD_{20} and explained methods for calculating changes for doubling dose, percentage change, or fold difference. The latter concept is explained by means of an example: a mean change in responsiveness from 1.96 to 0.98 μmol would be a change of 0.5-fold differences; that is, the mean dose at which the response occurred was half that at which it occurred initially.

Standardization of FEV_1

Lung function and, in particular, FEV_1 depends on age and height and, in any comparison of groups, differing age and height distributions must be accommodated. A common method is to use **linear regression**

models that have been fitted to large normal populations to obtain the predicted FEV_1 for a given age and height. Standard regression equations are available in Cotes [6] and Enright et al. [9]. The analysis may then proceed by analyzing the ratio of the observed to expected values, but this method is not recommended because there are issues of comparability of populations and **multiple linear regression** methods are usually easier to interpret.

Kory et al. [14] recommended the simple linear regression:

$$FEV_1 = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{height}. \quad (1)$$

Others have argued for an **interaction** term, age × height, in the model, or to standardize for height by calculating FEV_1/height^2 or FEV_1/height^3.

From the analysis of nine **cross-sectional** surveys, Cole [5] showed that a model used

$$\frac{FEV_1}{\text{height}^2} = \alpha_0 + \alpha_1 \times \text{age} \quad (2)$$

fitted the data better than a linear model. He also argued from allometric (*see* **Allometry**) principles that although a relationship with height^3 (or even height^4) may be the true model in which both FEV_1 and height are considered as correlated variables, when log(height) is thought of as a predictor of $\log(FEV_1)$, estimates from linear regression are shallower than the true slope and on the basis of the correlations in the data, the value of the slope would be shrunk to approximately 2, so that on a linear scale one would expect FEV_1 to be related to height^2.

However, Kronmal [15] has shown that if the true relationship is described by

$$FEV_1 = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{height}^2, \quad (3)$$

then the estimate for α_1 in (2) is a biased estimator of β_1 in (3). He argued that the coefficient for age in (2) is measuring the joint effect of varying age and height^2, which is an interaction, whereas the coefficient obtained in (3) from linear regression is measuring the effect of age, having adjusted for height^2. Prudent statistical practice does not fit an interaction term without fitting the main effects; in this case, terms for age and height^2 as well as age × height^2. These different models can, in fact, lead to different conclusions when two groups with different heights are compared; for example, a comparison of the rate of decline in lung function for men and women. For epidemiologic studies in which a large number of subjects with age and height measurements are available, it would seem better to adjust for age and height by including them (and possibly height squared) as terms in the regression equation. It is generally accepted that lung function declines linearly with age but the possibility of quadratic relationships with height should be investigated for each data set.

Statistical Methods Stimulated by Respiratory Data

The famous paper by Bland & Altman [1] on methods for assessing **agreement** between two methods of measurement originated in a comparison of two peak-flow meters. They dealt only with measurements on the same scale. However, it may be required to compare repeatability of different indices; for example, PD_{20}, in log(mg/ml) and the slope of the FEV_1 dose response curve in liters $(\text{mg/ml})^{-1}$. Dehaut et al. [7] and Chinn [3] suggested the use of the intraclass correlation coefficient (ICC) (*see* **Correlation**), defined as the ratio of the between-subject to total variation. Chinn [3] suggested that to be useful a measurement should have an intraclass correlation coefficient of at least 0.6. Baseline FEV_1 measured on two occasions 1–14 days apart were found to have an ICC of 0.88 [4]. However, repeated measurements of FEV_1 on the same day may give an ICC as high as 0.99 [21].

The paper by McCullagh [17] on *ordinal regression* was stimulated by a study of pneumoconiosis in coalminers. A typical study in this area would be to determine factors associated with the severity of pneumoconiosis using the ordinal ILO category as the dependent variable and exposure to dust as the main predictor variable, with other potential **confounding** variables such as cigarette smoking included in the equation.

When the dependent variable is PD_{20}, it may be truncated if the FEV_1 of a subject fails to drop to below 20% of baseline or if the subject asks to stop early. This leads to *truncated regression* as the method of analysis.

In truncated regression the model is

$$y_i = \alpha + \beta x_i + \varepsilon,$$

where ε is assumed $N(0, \sigma^2)$. If y_i is not censored, then the probability of observing y_i conditional on the

parameters is $f[y_i - x_i\beta)/\sigma]$, where f is the N(0,1) density. If y_i is right-censored at u_i (i.e. the true value is known to be greater than u_i), then the probability is $1 - F[(u_i - x_i\beta)/\sigma]$, where F is the cumulative normal density. The parameters can be estimated by **maximum likelihood** and the analysis is implemented in the statistical package STATA [23] (*see* **Software, Biostatistical**). The assumption of normality of the residuals assumes a much greater importance than in unconstrained multiple regression and should be investigated carefully. An example of truncated regression in respiratory medicine has been given by Devereaux et al. [8].

The whole area of the analysis of repeated measures has been stimulated by the fact that lung function measurements are made repeatedly on individuals. The classic paper by Laird & Ware [16] on **random effect** models considered 200 school children who had their FEV_1 measured on five successive occasions to examine the effect of air pollution. This has lead to a whole plethora of papers on repeated measurement analysis.

Discussion and Conclusions

Diseases of the respiratory system form a large part of the burden of disease in Western society and are the subject of much research. Asthma, particular in children, is increasing in prevalence, and there is intensive activity to determine the reasons for this. The FEV_1, originally devised to measure the effect of certain occupations on respiratory health, has emerged as one of the most important prognostic factors for future illness, comparable in power to serum cholesterol [12].

There are some interesting statistical problems that have arisen in the field of respiratory medicine, such as the standardization of results for age or height, comparison of methods, and ordinal regression which have proven fruitful elsewhere.

References

[1] Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **i**, 307–400.

[2] Campbell, M.J. (1993). Statistical issues: sample size, analysis and presentation of data, *Journal of Pharmaceutical Medicine* **3**, 223–230.

[3] Chinn, S. (1991). Statistics in respiratory medicine 2: repeatability and method comparison, *Thorax* **46**, 454–456.

[4] Chinn, S., Britton, J.R., Burney, P.G.J., Tattersfield, A.E. & Papacosta, A.O. (1987). Estimation and repeatability of the response to inhaled histamine in a community survey, *Thorax* **42**, 45–52.

[5] Cole, T.J. (1975). Linear and proportional regression models in the prediction of ventilatory function, *Journal of the Royal Statistical Society, Series A* **138**, 297–338.

[6] Cotes, J.E. (1979). *Lung Function: Assessment and Application in Medicine*, 3rd Ed. Blackwell Scientific, Oxford.

[7] Dehaut, P., Rachiele, A., Martin, R.R. & Malo, J.L. (1983). Histamine dose-response curves in asthma: reproducibility and sensitivity of different indices to assess response, *Thorax* **38**, 516–522.

[8] Devereux, G., Beach, J.P., Bromly, C., Avery, A.J., Ayatollahi, S.H.J., Williams, S.M., Stenton, S.C., Bourke, S.J. & Hendrick, D.J. (1995). Effect of dietary sodium on airways responsiveness and its importance in the epidemiology of asthma: an evaluation in areas of northern England, *Thorax* **50**, 941–947.

[9] Enright, P.L., Kronmal, R.A., Higgins, M., Schenker, M. & Haponik, E.F. (1993). Spirometry reference values for women and men 65 to 85 years of age: Cardiovascular Health Study, *American Review of Respiratory Disease* **147**, 125–133.

[10] Ferris, B.G. (1978). Epidemiology standardisation project, *American Review of Respiratory Disease* **118**, Suppl, 58–62.

[11] Fletcher, C., Peto, R., Tinker, C. & Speizer, F.E. (1976). *The Natural History of Chronic Bronchitis and Emphysema*. Oxford University Press, Oxford.

[12] Hole, D.J., Watt, G.C.M., Davey-Smith, G., Hart, C.L., Gillis, C.R. & Hawthorne, V.H. (1996). Impaired lung function and mortality risk in men and women: findings from the Renfew and Paisley prospective population study, *British Medical Journal* **313**, 711–715.

[13] International Labour Office (1980). ILO-U/C International classification of radiographs of pneumoconiosis. ILO, Geneva.

[14] Kory, R.C., Callahan, R., Boren, H.G. & Syner, J.C. (1961). The Veterans Administration-Army co-operative study of pulmonary function. I. Clinical spirometry in normal men, *American Journal of Medicine* **30**, 243–258.

[15] Kronmal, R.A. (1993). Spurious correlation and the fallacy of the ratio standard revisited, *Journal of the Royal Statistical Society, Series A* **156**, 379–392.

[16] Laird, N.M. & Ware, J.H. (1982). Random effects models for longitudinal data, *Biometrics* **38**, 963–974.

[17] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[18] Medical Research Council (1965). Definition and classification of chronic bronchitis for clinical and epidemiological purposes, *Lancet* **i**, 775–779.

[19] Medical Research Council (1966). *Instructions for Use of the Questionnaire on Respiratory Symptoms*. Holman, Dawlish.

[20] O'Connor, G., Sparrow, D., Taylor, D., Segal, M. & Weiss, S. (1987). Analysis of dose response curves to methacholine. An approach suitable to population studies, *American Review of Respiratory Disease* **136**, 1412–1417.

[21] Oldham, P.D. & Cole, T.J. (1983). Estimation of the FEV_1, *Thorax* **38**, 662–667.

[22] Peat, J.K., Unger, W.R. & Combe, D. (1994). Measuring changes in logarithmic data, with special reference to bronchial responsiveness, *Journal of Clinical Epidemiology* **47**, 1099–1108.

[23] Statacorp (1997). *Stata Statistical Software Release 5.0*. Stata Corporation, College Station.

[24] Wright, B.M. & McKerrow, C.W. (1959). Maximum forced expiratory flow rate as a measure of ventilatory capacity with a description of a new portable instrument for measuring it, *British Medical Journal* **2**, 1041–1047.

[25] Yan, K., Salome, C. & Woolcock, A.J. (1983). Rapid method for measurement of bronchial responsiveness, *Thorax* **38**, 760–765.

MICHAEL J. CAMPBELL

Quality Control in Laboratory Medicine

The statistical aspects of quality control (QC) in the field of clinical laboratory medicine are much more intricate than might be thought. Body fluids such as human blood are a molecular "soup" of extraordinary complexity, within which even a relatively simple constituent such as calcium can exist in a number of different states, for example, intracellular, compounded, ionized, or bound to carrier proteins. Laboratory assays encompass the blood gases, electrolytes, trace elements, vitamins, clotting factors, proteins, hormones, tumor markers, drugs and poisons, and more, and the assay of any one analyte has to accommodate, potentially at least, the presence of all (*see* **Biological Assay, Overview**). This is not to mention any one or more of the legion medications that the patient might be taking.

The vast majority of clinical laboratory assay methods rely upon the elicitation of a *specific* and observable property of the test analyte that varies in proportion to the concentration of the analyte in the biological test matrix. In order to visualize that relationship (the measurement function), standard solutions, or calibrators, of *known* concentration with respect to the analyte of interest are assayed in parallel with the test specimen. The level of the property (Y) recorded for each calibrator is plotted against the **calibrators** concentration value (X), the points so defined providing an estimate of the assay calibration function. The level of the property recorded for the test specimen is translated into units of concentration by reference to the inverse of the calibration function, under the critical assumption that calibration function and the measurement function are, to all intents and purposes, identical. Regular recalibration is a characteristic feature of laboratory assay systems.

Where possible, assay methods in routine use (field methods) should be supported by a clear link to a hierarchical program of *reference technology*. Such programs embrace a series of primary and secondary reference methods and standard reference materials, discussed in detail by Uriano & Cali [34], Tietz [32] and Boutwell [1]. The problems that beset attempts at assay standardization can prove formidable: Ekins [9] provides a very clear account of the subject from the perspective of immunoassay methods (*see* **Radioimmunoassay**).

The complexity of the biological test matrix poses a number of threats to the integrity of laboratory measurement. First, it is practically impossible to reproduce in full the complexity and properties of the biological test matrix in necessarily artificial calibrators. The primary impact of a failure in this regard is the very real possibility that the calibration function will no longer mirror the measurement function, thereby introducing a systematic error (or bias) into the measurement process. Second, the complexity of the analyte itself, or of its presentation in the biological test matrix, may also pose problems in respect of constructing plausible calibrators. For example, in its native state, the analyte may be intracellular, protein-bound, or conjugated with inorganic radicals as a prelude to excretion. In the laboratory, it may be a white powder in a bottle. Transforming the latter into a credible facsimile of the former is a prerequisite of meaningful measurement and failure in that endeavor is a primary source of systematic error. The reader is referred to [30, 31] for a more detailed account of the above problems. Third, the biological test matrix may, by virtue of its physical or chemical composition, elicit a signal in the measurement system that is indistinguishable from that generated by the target analyte (cross-reaction) or, it may modify the signal generated by the target analyte (interference). These effects, which reflect a lack of specificity in the measurement chemistry, are typically specimen dependent, manifesting as a bias on repeated assay of a single test specimen and as a random variable across different test specimens. They are referred to variously as matrix effects, specimen-dependent **biases**, random biases, or subject-by-method **interactions**. Summarizing, we have, for a given test result X_i:

$$X_i = \xi_i + \phi + \psi_i + \varepsilon_i \tag{1}$$

where: X_i = observed test result for sample i.
ξ_i = true test value for sample i.
ϕ = analytical bias, primarily calibration failure.
ψ_i = specimen-dependent bias.
ε_i = random measurement error (imprecision).

The ideal for a laboratory measurement system would be a degree of measurement specificity, standardization, and control that effectively eliminated

all bias. Under such ideal circumstances, (1) would reduce to (2):

$$X_i = \xi_i + \varepsilon_i \tag{2}$$

QC would appear to have little left to concern itself with beyond regulating the magnitude of the random error term ε There is, however, another dimension of complexity to the reported test value X. That dimension is time, whose passing reveals several new sources of variation in laboratory test results and several new challenges for QC to accommodate.

First, *all* living organisms change over time as a consequence of perfectly natural biological variation. Biological variation can take a number of forms; it may be nonrandom, for example, daily, monthly, or seasonal rhythms (*see* **Chronomedicine**), along with those changes that characterize our progress from cradle to grave, or it may be random in character. For the majority of analytes of clinical interest, biological variation can be construed as a random process over time (*see* **Stochastic Processes**), changes occurring around a homeostatic setting point μ, the latter being fixed for a given individual but differing in value across a set of individuals.

Second, the random measurement error term ε is itself time dependent. Laboratory assays are typically undertaken by assaying a number of test specimens in one batch or analytical run, for example, one run a day. A single analytical run will be characterized by fewer sources of variation than a number of consecutive analytical runs, for example, one run, one calibration of the measurement system, a common set of analytical reagents, a common set of all those factors beyond strict technical control. The time element in respect of assay imprecision is captured in the terms *within-run* and *between-run* imprecision. Incorporating the above factors into model (2) we have:

$$X_{ijk} = (\mu_i + \eta_{ij}) + \varepsilon_{ij} + \varepsilon_{ijk} \tag{3}$$

(true value ξ)

where X_{ijk} = the observed result from the kth replicate assay in the jth analytical run (or day) for the ith test subject.

μ_i = the subject's individual biologic mean or homeostatic setting point.

η_{ij} = a peturbation due to biologic variation in the jth analytical run having an expected value (*see* **Expectation**) of 0 and a variance σ_η^2.

ε_{ij} = a peturbation due to random measurement error in the jth analytical run having an expected value of zero and a variance σ_B^2.

ε_{ijk} = a peturbation due to random measurement error in the kth replication within the jth analytical run having an expected value of zero and a variance σ_W^2.

We have four components of variability associated with (3) which, in the laboratory sciences, are typically expressed in terms of estimated **standard deviations** (SD) or corresponding coefficients of variation (CV = SD as a % of the relevant mean value):

SD_G = Between-subject biological variation, reflecting the differences *between* individuals in a given population in respect of their homeostatic setting-points μ_ι. SD_G^2 is an estimate of σ_μ^2. $CV_G = SD_G$ as a percent of the population mean.

SD_I = Average within-subject biological variation, reflecting the changes in a given subject's test result over time. SD_I^2 is an estimate of σ_η^2. $CV_I = SD_I$ as a % of population mean.

SD_B = Between-run imprecision component. SD_B^2 is an estimate of σ_B^2. $CV_B = SD_B$ as percent of control sample mean.

SD_W = Within-run imprecision component. SD_W^2 is an estimate of σ_W^2. $CV_W = SD_W$ as percent of control sample mean.

We also have two derived quantities:

$$SD_{\text{biol}} = \textit{Total biological variation}.$$

$$SD_{\text{biol}} = (SD_I^2 + SD_G^2)^{1/2}$$

$$\text{or } CV_{\text{biol}} = (CV_I^2 + CV_G^2)^{1/2} \tag{4}$$

$$SD_T = \textit{Total imprecision}.$$

$$SD_T = (SD_W^2 + SD_B^2)^{1/2}$$

$$\text{or } CV_T = (CV_W^2 + CV_B^2)^{1/2} \tag{5}$$

Estimates of the random measurement error components can be obtained by assaying the same control or test specimen a minimum of $k = 2$ times on each of $j = 20$ days, following the recommendations of the National Committee for Clinical Laboratory

Standards (NCCLS) [24]. The error components are obtained from a one-stage nested **analysis of variance** (ANOVA). A worked example is presented by Kringle & Bogovitch [21].

Note that for (3) we have equated runs with days (one run = one day) simply in order to minimize model complexity. *If* multiple runs within a given working day were a feature of the test service under study, an additional level of subscripting would be required for the random error term ε in order to distinguish within-run, run-to-run, and day-to-day sources of random variation. The error components can be obtained using a two-stage nested ANOVA [21].

The majority of laboratory test methods exhibit some degree of dependence between the magnitude of the error variance and analyte concentration (heteroscedasticity; *see* **Scedasticity**). This dependence may take the form of a simple proportional relationship, that is, a constant CV_T, at least as an approximation, but this is by no means the rule. Heteroscedasticity necessitates obtaining several estimates of random error variation across the working range of the test method, preferably in proximity to medical decision points for the analyte in question. Given such estimates, imprecision profiles can be constructed for both within-run and total imprecision.

Coefficients of variation need careful handling in the context of summarizing analytical imprecision (CV_W or CV_T). Users may lose sight of the connection between a CV and its parent SD, for example, an analytical error CV of 6% implies a 95% error margin of plus or minus 12% (not 6%) at the relevant test concentration or activity level.

The factors to be considered in developing estimates of biological components of variation are reviewed in detail by Fraser and Harris [13]. The authors provide a clear account of the requisite **experimental design** and subsequent data preparation, prior to the estimation of the components of variation (*see* **Variance Components**) using a two-stage nested ANOVA. Estimates of biological variance components (CV_I and CV_{biol}) have already been published for a large number of laboratory analytes [12, 27]. A comprehensive data bank of such components, along with source references, is maintained online at `www.westgard.com`. The components play a key role in subsequent discussion of analytical goals for QC.

Quality Control in Context

Quality control (QC) refers to a set of procedures by which we seek to monitor the components of error associated with laboratory test procedures in furtherance of the practice of good medicine. QC is essentially a set of techniques adapted from industrial statistical process control (SPC). QC is, or should be, embedded within the larger framework of a total quality system embracing *all* aspects of the laboratory function. Total quality management (TQM) places the *user* of the laboratory center stage in respect of defining the levels of quality (and costs) that constitute an acceptable level of service.

Waiting in the wings, Six Sigma, a total quality management *system* that aims for an error (defect) rate of no more than four in a million in the "product" of a manufacturing or service delivery process. The Six Sigma goal is achieved through the pursuit and elimination of all sources of variation impacting upon process outcome. If the process in question starts with a request for a given test, and ends with the delivery of the required test result, every conceivable aspect of the journey from request to return is the subject of ongoing scrutiny and, where required, redesign or reeducation. A statistical overview of Six Sigma is provided by Hahn et al. [16]: a collection of lab-oriented essays on the topic is available at `www.westgard.com`.

Control Charts

The day-to-day practice of QC in laboratories (internal QC) is largely focused on charts, simple graphical displays of the results obtained for control samples against time of assay. The classic laboratory "control" chart is the single value modification of the "Levey–Jennings" chart, described by Henry and Segalove [17] in 1952. It is widely referred to as a "Shewhart" control chart in clinical laboratory literature, reflecting its origins in the work of Shewhart [29] in the 1930s. An example of the chart is illustrated in Figure 1.

For the construction of the chart, SD_T should be estimated from a minimum of 20 assays of the control sample over 20 different days at a time when the assay is judged to be in statistical control.

In operation, the results obtained for a single control specimen introduced into every analytical run should be checked against the $2s$ limits, or the 3s

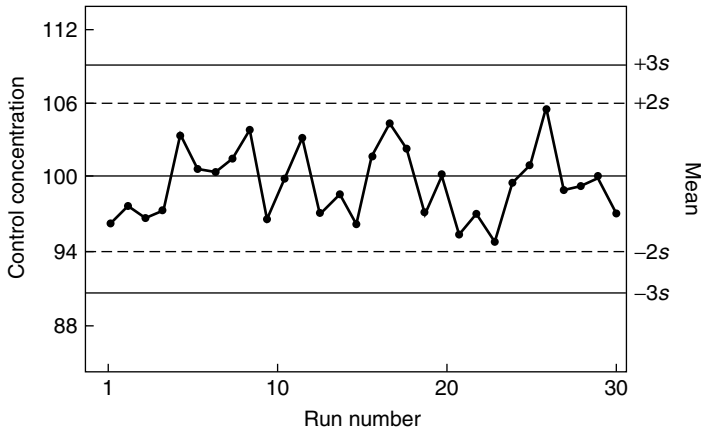

Figure 1 Levey–Jennings QC chart (s = total imprecision SD_T)

limits if there are two or more such controls. Any control value exceeding the appropriate threshold flags an out-of-control assay run, that is, no test results are to be reported from that run. In practice, many, if not most users of the above charting system rely on a simple $2s$ rule, irrespective of the number of control samples involved. Westgard & Groth [38] have published **power functions** for the above control rules, revealing the very high false-alarm rates that accompany misuse of the $2s$ rejection rule. A secondary problem with the basic Levey–Jennings chart is its relative insensitivity to systematic error (bias). To address these deficiencies, Westgard et al. [37] published their 'multi-rule' control procedure along with associated error-detection power curves. Essentially, it is an elaboration of the Levey–Jennings chart obtained by including control lines at the control mean plus or minus $1s$, $2s$, $3s$, and $4s$, preferably color-coded for clarity. Two control samples, or one control assayed twice, comprises the minimum configuration required to assess a given test run. The two control results are assessed against the following rules:

- 1_{2s} One control result exceeding the $2s$ limits. This is a warning rule that initiates reference to the remaining rules.
- 1_{3s} One control result exceeding the $3s$ limits. A REJECTION rule, primarily sensitive to random error.
- 2_{2s} Two consecutive control results both greater than $+2s$ or both less than $-2s$. A REJECTION rule, sensitive to bias.
- R_{4s} One observation exceeding $+2s$ limit and the other exceeding the $-2s$ limit. A REJECTION rule, sensitive to random error.
- 4_{1s} Four consecutive control results all exceeding the $+1s$ limit, or all exceeding the $-1s$ limit. A REJECTION rule, sensitive to bias.
- 10_m Ten consecutive control results all in excess of the mean, or all lower than the mean. A REJECTION rule, sensitive to bias.

Clinical laboratories are under increasing regulatory pressure from professional bodies and/or Government agencies in respect of meeting specified standards of performance, making it ever more important that an appropriate QC strategy is in place for every assay system in use. By appropriate, we mean a strategy that strikes a cost-effective balance between the probability of rejecting a bad run correctly, which probability we would like to be high, and the probability of rejecting a good run incorrectly, which probability we would like to be low. By "bad" we mean an assay whose *total error* (TE), a combination of both random and systematic error components, exceeds a specified threshold. What might constitute an appropriate threshold for TE is taken up in the section on analytical goals. Westgard [36, 40] provides a clear account of the steps involved in formulating an optimal QC strategy. A software package, EZ rules™, for automatic selection of statistical control rules is available through `www.westgard.com`, which also hosts a comprehensive inventory of resources in respect of laboratory QC. EZ rules™ has been reviewed by Linnet [11].

Cumulative Sum (Cusum) Control

A sensitive indicator for low-level bias in assay systems is the cumulative sum (cusum) chart, developed by Page [26] for manufacturing control. A cusum chart records the deviations of each control result in a series from a control set-point or mean μ_c, but instead of plotting the deviations individually, we plot the running, or cumulative sum of those deviations, that is, cusum value $= \Sigma(X_i - \mu_c)$. If the assay method is unbiased, we would expect an approximate balance between the positive and negative deviations, resulting in a cusum trace fluctuating around a horizontal line centered on zero. Any shift in the accuracy base (bias) will result in the accumulation of deviations of like sign, manifesting as a slope or ramp away from zero on the cusum chart.

Two control rule strategies are available to alert the user to significant bias, one visual, the V-mask template, and one numerical, the decision-interval rule. A practical account of both is presented in [30].

Westgard et al. [39] describe a combined Levey–Jennings cusum control chart, along with associated error-detection power curves. Cusum control rules are not without their problems, amongst which are the assumptions of statistical independence and constant error variance for the control results being charted. The consequences of violating these assumptions have been explored by Johnson & Bagshaw [20].

An interesting implementation of the CUSUM principle is to be found in the program "**Change-Point Analyzer**", (downloadable from `www.variation.com`). Essentially, it is a hybrid of the cusum principle and a **bootstrap** estimator for determining the probability of change (bias) in a given series. The program includes automatic checking for nonindependence and **outliers**, with a number of workarounds for both contingencies.

Alternative Control Procedures

Cembrowski et al. [6] and Neubauer [25] each describe control programs for detecting persistent systematic error, using exponentially weighted **moving averages** (EWMA), the former authors employing Trigg's tracking signal as a trend detection device [33]. Jay Smith & Myers [19] describe a procedure for detecting short-term trends in serial control results using a moving **least-squares** regression slope estimator. An original variation on the EWMA theme is described by Bull et al. [2, 3] utilizing the daily or batch mean for patient test results, for each of the erythrocyte indices, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, and mean corpuscular volume, to monitor calibration drift in automated hematology analyzers. The performance characteristics of Bull's **algorithm** have been studied by Lunetsky & Cembrowski [22].

Delta Checks

The use of patient test data to drive QC procedures has obvious attractions in terms of relevance (real test specimens rather than "artificial" controls), cost (controls cost money), and convenience. Bull's algorithm is an example. Another is the Delta check based on observing the difference between two separate test results on the same patient. The original Delta check rules were either empirical in nature, or based on studies of the observed distributions of differences in the population being served by the laboratory. The observation of unexpectedly large differences signals a possible problem, for example, mislabeling of specimens. Sheiner et al. [28] have reviewed the performance of several Delta check methods. Fraser [12] suggests an intellectually pleasing basis for Delta checking by defining the upper limit for expected change in terms of total imprecision (SD_T) and within-subject biological variation (SD_I). Fraser refers to such a limit as a reference change value (RCV):

$$RCV = 2^{0.5} * Z * (SD_T^2 + SD_I^2)^{1/2} \qquad (6)$$

where $Z =$ **standard normal deviate**, giving us:

$$RCV = 2.77 * (SD_T^2 + SD_I^2)^{1/2} \text{ for 95\% limit (2-sided).}$$
$$= 3.65 * (SD_T^2 + SD_I^2)^{1/2} \text{ for 99\% limit (2-sided).}$$

Bear in mind the possible dependence of SD_T on the level of the analyte, that is, select an appropriate value for SD_T from the assay imprecision profile. A 99% limit is recommended by Fraser for Delta checking. Large-scale Delta checking is easily incorporated into laboratory information management systems (LIMS).

As an aside, note that the RCV lends itself directly to assessing the significance of observed

changes between consecutive test results on the same patient, that is, can the observed change be ascribed to the combined effect of error plus perfectly natural biological variation, or is it so great as to signify something worthy of attention or clinical intervention?

Authoritative recommendations on the design and implementation of internal QC schemes have been published by the International Federation of Clinical Chemistry (IFCC); see [4], and by the NCCLS [23].

External Quality Control

The control procedures described above relate to internal QC, the laboratory monitoring itself on a day-to-day basis. External QC programs, run by commercial organizations or professional bodies, provide an invaluable link to the experience of hundreds of other laboratories, and they serve to alert any individual laboratory to the possibility that its performance might be inconsistent with that of the collective.

The basic operation of such schemes involves the distribution of the same "control" material to all participating laboratories, the subsequent results returned being subjected to statistical summary. Typically, the mean of results returned for a given analyte, stratified by assay methodology where this is appropriate, is viewed as a surrogate "true" value. The deviation of any one test result from the appropriate group mean is expressed in standardized form, usually based on an index of group dispersion. Schemes vary in the frequency of specimen distribution, in the definition of true or target values and in the criteria they employ for characterizing performance. All external QC schemes provide a necessarily retrospective view of performance. The topic is well covered by Buttner et al. [5], Westgard and Klee [40] and a well-informed discussion of outstanding problems from the perspective of the UK National External Quality Assessment Schemes (UK NEQASs) is provided by Hirst [18].

Analytical Goals for Laboratory Error

QC procedures are designed to monitor the performance characteristics of laboratory test methods (imprecision, bias, and blunders) and to alert us to changes in those characteristics that may impact upon the quality of the laboratory service. QC procedures do not of themselves define desirable quality standards. What standards? Defined by whom? Desirable to whom? Ask 20 different people with an interest in the laboratory (managing it, running it, paying for it, working in it, or using it) how they would define "quality" in relation to a laboratory test result. You will surely get 20 different answers, reflecting 20 different sets of priorities. And so it is with those who have busied their minds with this problem over the last two decades.

The subject is reviewed by Fraser [12, 14] who describes a hierarchy of objectivity in the definition of quality objectives. Paraphrasing from top down (from the most to the least desirable), we have quality standards defined by a specific medical need, by general medical needs in relation to case finding and case monitoring, by professional or expert recommendations, by legislation or regulation, or by what is technically possible.

The Aspen conference on Analytical Goals in Clinical Chemistry [10], organized by the College of American Pathologists in 1976, agreed guidelines for maximum tolerable imprecision in relation to two broad classes of test usage, (a) population screening, and (b), individual testing (e.g. case management). The guidelines, which drew on the work of Cotlove et al. [7], were framed in terms of a key recommendation. *Analytical variance should not exceed one-quarter of the relevant biological variance,* thereby insuring that the variability of result is increased by no more than 12% due to assay imprecision. The latter result follows from noting that the overall variability associated with an observed test result is made up of two (primary) components, one biological (within-subject biological variation) and the other analytical (total imprecision). The variability imparted to the observed test result by analytical imprecision, over and above that due to biology alone, is obtained from the following formula, adapted from [13]:

$$\left[1 + \left(\frac{CV_T^2}{CV_I^2}\right)\right]^{1/2} \tag{7}$$

If analytical variance $= 1/4$ within-subject biological variance, we have $(CV_T^2/CV_I^2) = 0.25$, so $[1 + 0.25]^{1/2} = 1.118$. That is an increase of 11.8% in observed variability due to analytical imprecision alone. The choice of 11.8% as a threshold for the inflationary effect of analytical imprecision is arbitrary. It follows from a memorably simple rule,

that is, $CV_T < 1/2\ CV_I$ (or 1/4 if using the corresponding CV^2), and from the fact that the impact of analytical imprecision on overall test variability grows appreciably larger once the Aspen threshold is exceeded.

A Working Party of the European Group for the Evaluation of Reagents and Analytical Systems in Laboratory Medicine [15] subsequently proposed quality specifications based on a pragmatic marriage of the Aspen "rule" and the current state of the art in laboratory testing as judged from external QC schemes:

The total imprecision threshold (CV_{limit}) should be:

(a) less than one-half of the average within-subject biological variation (CV_I), or
(b) less than the state of the art achieved by the best 0.20 fractile of laboratories,

whichever is the less stringent. The second approach to be used if the requisite biological data are unavailable.

The inaccuracy threshold (B_{limit}) should be:

(a) less than one-quarter of the total biological variation (CV_{biol}), or
(b) less than one-sixteenth of the clinical reference interval, when data on biological variation are unavailable, or
(c) less than twice the ideal imprecision, if the above specifications are too demanding.

Note the commonplace use of standard deviations (or corresponding CV) rather than variances (or CV^2) as descriptors of variability in clinical laboratory settings.

The user of the laboratory is unlikely to be very interested in the distinction between imprecision and inaccuracy. Of more relevance will be their combined effect in a given test result, the total error (TE), where TE% = total imprecision (%) + total bias (%). The bias prevailing in a given test system can be conveniently estimated using the cumulative performance statistics provided by external QC scheme organizers. The most commonly cited quality goal for TE is given by:

$$\text{Total error threshold} = TE_{\text{limit}} = B_{\text{limit}} + 1.65\ (CV_{\text{limit}}) \quad (8)$$

By way of illustration, we have the following basic biological reference values for serum copper, Table A1.3 of ref [12]: $CV_I = 4.9\%$ and $CV_G = 13.6\%$. It follows from (4) that the total biological variation is $CV_{\text{biol}} = 14.5\%$. The error thresholds for serum copper estimations are therefore:

for imprecision:
$CV_{\text{limit}} = 1/2\ (4.9) = 2.5\%$
for bias:
$B_{\text{limit}}\ (\%) = 1/4\ (14.5) = 3.6\%$
for total error:
$TE_{\text{limit}}\ (\%) = 3.62 + 1.65\ (2.5) = 7.7\%$

The quantity TE_{limit} plays a key role in formulating optimal internal QC strategies using Westgard's operating specifications charts [36, 40].

The logical basis of the above "European" quality goals is discussed in very accessible terms by Fraser [12]. The same publication also provides an extensive listing of the threshold values CV_{limit}, B_{limit}, and TE_{limit} for blood chemistry, immunology, urine chemistry, hematology, and hemostasis; see `www.westgard.com` for similar resources.

In the United States, quality goals have been enshrined in legislation as fixed upper limits for TE by the Clinical Laboratory Improvements Amendments of 1988 (CLIA'88), Final Rule [35]. The document is accessible online at `www.cms.hhs.gov/clia/`. In its current form, the CLIA'88 proposals call for the circulation, by approved proficiency-testing (PT) centers, of five test samples three times a year to US laboratories. In effect, a mandatory external QC program covering every pathology discipline. PT failure is getting two incorrect results out of five, in two out of three consecutive surveys.

The danger with establishing quality thresholds, however, derived or defined, is that they may be perceived to be targets to aim for, rather than lines that should not be crossed. With that in mind, Ehrmeyer et al. [8] have pointed out that any laboratory achieving a TE of one-third the CLIA thresholds will have something approaching a 100% chance of satisfying the CLIA'88 regulations. The "one-third" CLIA rule might therefore be regarded as a desirable analytical goal simply in terms of retaining a license to practice.

Complete listings of the CLIA'88 *TE* thresholds are available at `www.westgard.com`, along with a wealth of technical information and essays from

authorities in the field. European and US CLIA quality specifications are compared and contrasted by Westgard et al. [41].

References

[1] Boutwell, J.H. ed. (1978). *A National Understanding for the Development of Reference Materials and Methods for Clinical Chemistry – Proceedings of a Conference*. AACC Press, Washington.

[2] Bull, B.S. & Elashoff, R.M. (1974). The use of patient derived haematology data in quality control, *Proceedings of the San Diego Biomedical Symposium* **13**, 515–519.

[3] Bull, B.S., Elashoff, R.M., Heilbron, D.C. & Couperus, J. (1974). A study of various estimators for the derivation of quality control procedures from patient erythrocyte indices, *American Journal of Clinical Pathology* **61**, 473–481.

[4] Buttner, J., Borth, R., Boutwell, J.H. & Broughton, P.M.G. (1983). Federation of clinical chemistry: approved recommendation (1983) on quality control in clinical chemistry: IV. Internal quality control, *European Journal of Clinical Chemistry and Clinical Biochemistry* **21**, 877–884.

[5] Buttner, J., Borth, R., Boutwell, J.H., Broughton, P.M.G. & Bowyer, R.C. (1983). International federation of clinical chemistry: approved recommendation (1983) on quality control in clinical chemistry: V. External quality control, *European Journal of Clinical Chemistry and Clinical Biochemistry* **21**, 885–892.

[6] Cembrowski, G.S., Westgard, J.O., Eggert, A.A. & Toren, E.C. Jr. (1975). Trend detection in control data: optimization and interpretation of Trigg's technique for trend analysis, *Clinical Chemistry*, **21**, 1396–1405.

[7] Cotlove, E., Harris, E.K. & Williams, G.Z. (1970). Biological and analytic components of variation in long-term studies of serum constituents in normal subjects. III. Physiological and medical implications, *Clinical Chemistry* **16**, 1028–1032.

[8] Ehrmeyer, S.S., Laessig, R.H., Leinweber, J.E. & Oryall, J.J. (1990). Medicare/CLIA final rules for proficiency testing: minimum intralaboratory performance characteristics (CV and bias) needed to pass, *Clinical Chemistry* **36**, 1736–1740.

[9] Ekins, R. (1991). Immunoassay standardization, *Scandinavian Journal of Clinical Laboratory Investigation Supplement* **205**, 33–46.

[10] Elevitch, F.R. ed. (1977). *Proceedings of the 1976 Aspen Conference on Analytical Goals in Clinical Chemistry*, College of American Pathologists, Skokie.

[11] Linnet, K. (2002). EZ rules: automatic selection of statistical control rules for laboratory tests: software review, *Clinical Chemistry* **48**, 594–595.

[12] Fraser, C.G. (2001). *Biological Variation: from Principles to Practice*. AACC Press, Washington.

[13] Fraser, C.G. & Harris, E.K. (1989). Generation and application of data on biological variation in clinical chemistry, *Critical Reviews in Clinical Laboratory Sciences* **27**, 409–437.

[14] Fraser, C.G. & Hyltoft Peterson, P. (1999). Analytical performance characteristics should be judged against objective quality specifications, *Clinical Chemistry* **45**, 321–323.

[15] Fraser, C.G., Hyltoft Petersen, P., Ricos, C. & Haeckel, R. (1992). Proposed quality specifications for the imprecision and inaccuracy of analytical systems in clinical chemistry, *European Journal of Clinical Chemistry and Clinical Biochemistry* **30**, 311–317.

[16] Hahn, G.J., Hill, W.J., Hoerl, R.W. & Zinkgraft, S.A. (1999). The impact of six sigma improvement – a glimpse into the future of statistics, *The American Statistician* **532**, 208–215.

[17] Henry, R.J. & Segalove, M. (1952). The running of standards in clinical chemistry and the use of the control chart, *Journal of Clinical Pathology* **5**, 305–311.

[18] Hirst, A.D. (1998). External quality assurance, *Annals of Clinical Biochemistry* **35**, 12–18.

[19] Jay Smith, S. & Myers, G.L. (1991). Analyzing quality-control trends with moving slope charts, *Clinical Chemistry* **37**, 341–346.

[20] Johnson, R.A. & Bagshaw, M. (1974). The effect of serial correlation on the performance of CUSUM tests, *Technometrics* **16**, 103–112.

[21] Kringle, R.O. & Bogovitch, M. (1999). Statistical procedures, in *Tietz Textbook of Clinical Chemistry*, 3rd Ed, C.A. Burtis & E.R. Ashwood, eds. W.B. Saunders Co, Philadelphia, 288–294.

[22] Lunetsky, E.S. & Cembrowski, G.S. (1987). Performance characteristics of Bull's multirule algorithm for the quality control of multichannel hematology analyzers, *American Journal of Clinical Pathology* **33**, 634–638.

[23] NCCLS Document C24-A2 (1999). *Statistical Quality Control for Quantitative Measurements: Principles and Definitions; Approved Guidelines*, 2nd Ed. National Committee for Clinical Laboratory Standards, Wayne.

[24] NCCLS Document EP5-A (1999). *Evaluation of Precision Performance of Clinical Chemistry Devices: Tentative Guidelines*, 2nd Ed. National Committee for Clinical Laboratory Standards, Wayne.

[25] Neubauer, A.S. (1997). The EWMA control chart: properties and comparison with other quality-control procedures by computer simulation, *Clinical Chemistry* **43**, 594–601.

[26] Page, E.S. (1954). Continuous inspection schemes, *Biometrika* **14**, 100–115.

[27] Ricos, C., Alvarez, V., Cava, F., Garcia-Lario, J.V., Hernandez, A., Jimenez, C.V., Minchinela, J., Perich, C. & Simon, M. (1999). Current databases on biologic variation: pros, cons, and progress, *Scandinavian Journal of Clinical Laboratory Investigation* **59**, 491–500.

[28] Sheiner, L.B., Wheeler, L.A. & Moore, J.K. (1979). The performance of delta check methods, *Clinical Chemistry* **25**, 2034–2037.

[29] Shewhart, W.A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand, New York.

[30] Strike, P.W. (1991). *Statistical Methods in Laboratory Medicine.* Butterworth-Heinemann Ltd, Oxford.

[31] Strike, P.W. (1996). *Measurement and Control in Laboratory Medicine: A Primer on Control and Interpretation.* Butterworth-Heinemann Ltd, Oxford.

[32] Tietz, N.W. (1979). A model for a comprehensive measurement system in clinical chemistry, *Clinical Chemistry* **25**, 833–839.

[33] Trigg, D.W. (1964). Monitoring a forecasting system, *Operational Research Quarterly* **15**, 271–274.

[34] Uriano, G.A. & Cali, J.P. (1977). Role of reference materials and reference methods in the measurement process, in *Validation of the Measurement Process*, J.R. De Voe, ed. American Chemical Society, Symposium series, Washington, DC.

[35] U.S. Department of Health and Human Services: Clinical Laboratory Improvements Amendments of 1988. Final Rule. Laboratory Requirements, February 28, (1992). *Federal Register* **57**, 7002–7288.

[36] Westgard, J.O. (1992). Charts of operational process specifications ("OP-Specs charts") for assessing the precision, accuracy, and quality control needed to satisfy proficiency testing criteria, *Clinical Chemistry* **38**, 1226–1233.

[37] Westgard, J.O., Barry, P.L., Hunt, M.R. & Groth, T. (1981). A multi-rule Shewhart chart for quality control in clinical chemistry, *Clinical Chemistry* **27**, 493–501.

[38] Westgard, J.O. & Groth, T. (1979). Power functions for statistical control rules, *Clinical Chemistry* **25**, 863–869.

[39] Westgard, J.O., Groth, T., Aronsson, T. & de Verdier, C. (1977). Combined Shewhart-Cusum control chart for improved quality control in clinical chemistry, *Clinical Chemistry* **23**, 1881–1887.

[40] Westgard, J.O. & Klee, G.G. (1999). Quality management, in *Tietz Textbook of Clinical Chemistry,* 3rd Ed, C.A. Burtis & E.R. Ashwood, eds. W.B. Saunders Co, Philadelphia, 384–418.

[41] Westgard, J.O., Seehafer, J.J. & Barry, P.L. (1994). European specifications for imprecision and inaccuracy compared with operating specifications that assure the quality required by US CLIA proficiency-testing criteria, *Clinical Chemistry* **40**, 1228–1232.

P.W. STRIKE

Quality of Care

What is Quality of Care?

Quality of care refers to the attributes or characteristics of the delivery and subsequent outcomes of health care [4]. Health care includes the treatment of both physical and mental illnesses. High quality of care is defined by the Institute of Medicine [16] as the "degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge". This definition applies to various health providers, including physicians, nurses, and case managers; to all types of health care organizations such as preferred provider organizations, health maintenance organizations, and point-of-service plans; and to almost all settings of care such as hospitals, physician offices, nursing homes, and community sites.

Why Assess Quality of Care?

There are three general reasons for assessing quality of care delivered in the ordinary circumstances of routine practice: to improve care, to provide accountability, and to quantify information for market choices.

Improve Care

Empirical support for many treatment recommendations is surprisingly thin – most evidence arises from randomized trials (*see* **Clinical Trials, Overview**) where inclusion criteria can be stringent and in which selective practitioners participate. In contrast, routine practice consists of patients covering a wide spectrum of health status, from the very ill to the worried well. Well-collected data in these settings can provide information about the quality of treatments, of delivery strategies, and of organizational mechanisms. For example, if patients receive too much care, then this may result in the use of unnecessary interventions, unnecessary exposure to health risks, and resource waste. If patients receive too little care, either lacking in intensity or adequacy or both, then they are at risk of untoward outcomes, such as morbidity and mortality, which ultimately lead to higher costs. Identifying differences in health outcomes on the basis of quality measurement can lead to improved health outcomes.

Provide Accountability

Several constituencies have a stake in periodic disclosure of the quality of health care: consumers [1]; government [6]; professional societies [3, 13]; and purchasers [26]. Public disclosure of quality typically involves comparisons of health care providers or settings on the basis of quality measures in *report cards* [24] or **league tables** [11] (*see* **Profiling Providers of Medical Care**). Consumers of health care want to be able to identify high-quality practitioners and health plans. Professional societies want to ensure that specialists are kept informed of the latest advances to improve patient care, and may use quality measures as the basis for board certification. Government health programs wish to enforce compliance with health and safety requirements, and credentialing of medical staff. In the United States, the Joint Commission on Accreditation of Healthcare Organizations, for example, has statutory authority to certify hospitals, ambulatory surgical centers, clinical laboratories, home health agencies, and hospices. Employers who purchase health care want to avoid the effects of poor quality of care on productivity.

Market Choices

Public and private employers, business coalitions, and public programs, such as Medicaid and Medicare, have sought to redesign the health care system through negotiation with providers and insurers. These activities, denoted value-based purchasing [22], are designed to ensure and improve the quality of health programs. Purchasers contact selectively with health plans or provider organizations based on price and demonstrated quality [9]. Enhanced health benefit packages may help employers retain employees, increase employee satisfaction and productivity, decrease absenteeism, and hence reduce long-term health costs. Virtually all companies in the *Fortune 500* companies, for example, have reported collecting some information about health plan quality [18]. Some large companies, like Xerox, have used quality measures in making procurement decisions [17].

How to Measure Quality?

Prior to the mid-1980s, professional judgement was used to ensure patients received high quality of care.

This was usually accomplished by individual practitioners who would monitor care of their patients. In a 1980 article, Avedis Donabedian [7] suggested that quality of care should be measured by its structure, process, or outcome.

Structural measures include resources of the health care system that reflect the capacity of the provider to deliver good health care. The availability of diagnostic and therapeutic equipment for patient care may be assessed to determine whether quality of care is adequate. These organizational characteristics are often required by government programs through accreditation or certification requirements as a way to ensure some capacity for quality.

Process-based measures refer to what care givers do to and for patients. Typically, such measures focus on the diagnosis and management of disease, as well as preventive care such as **screening**. These measures should be under the control of the health professional. The idea is to define a population of patients who *should* get a therapy or test, and count how many actually got it. Eligible patients are identified using established guidelines or other explicit clinical criteria [20]. In Table 1, the **target population** consists of patients discharged alive or dead with a diagnosis of acute myocardial infarction (AMI). Because selected subsets of patients are used for **inference**, for example, those eligible for reperfusion therapy, achieving a sufficient sample size for valid inference and generalizing results to all patients may be difficult.

Outcomes-based measures refer to responses that characterize a patient's health status and quantifies whether the care received has improved the patient's health. Common examples include **risk-adjusted** mortality, occurrence of complications, relief of symptoms, and patient reports about their health. Thirty-day risk-adjusted mortality following an AMI is a common outcomes-based measure as the majority of treatments for care will have be given within one month of admission.

Good structural or process measures are those that, if modified, would lead to demonstrable changes in outcomes. Similarly, good outcome measures are those that, if the process measure was changed, changes in outcomes would occur.

Data Analyses

Analyses of health care quality data are challenging due to their nonexperimental nature and complex structure. Randomized studies are infrequently used

Table 1 AMI process and outcomes-based quality measures. Example measures for assessing hospital-specific quality of care for patients having a heart attack

	Process			Outcome
Measure	Reperfusion	Underuse of Angiography	Beta-blockers	30-day Mortality
What?	Therapy to open blocked vessels to restore blood to the heart	Invasive procedure to assess extent of heart disease	Medical therapy that lowers the heart's need for oxygen	All cause mortality
Definition	Angioplasty or thrombolysis within 12 hours of hospital arrival	Use within 12 hours after symptom onset & prior to discharge	Discharge R_x for beta-blockers	Died within 30-days of admission
Exclusions	No ST-elevation; onset of chest pain > 12 h prior to arrival; bleeding on arrival; hx of ulcer or chronic liver disease; CVA; warfarin on admission; age > 79; recent trauma; surgery within 2 months; aborted angiography	Not rated ACC Class I; died within 1 day of admission; underwent primary angioplasty	EF < 35%; pulmonary edema or CHF and EF < 50%; shock, systolic blood pressure < 100, or hypotension; hx of COPD, dementia; heart block; bradycardia; insulin or antidepressant trt	None but adjust for admission risk factors

hx = history; R_x = prescription; CVA = cerebrovascular accident; EF = ejection fraction; CHF = congestive heart failure; COPD = chronic obstructive pulmonary disease; ACC = American College of Cardiology.

to assess quality of care, although some have been used to identify interventions that lead to improvements in health quality [21, 25, 27]. Because the majority of quality studies are **observational**, the lack of randomization may lead to **confounding** when identifying factors that impact quality. Health services researchers have historically adopted one of two approaches to reduce **bias** [12]. The first approach involves modeling outcomes-based measures and utilizing **regression**-adjustment techniques to balance the observed confounders across groups defined by study factors, for example, hospital type. The idea is to adjust for the patient's condition prior to the initial contact with the health system. For example, to determine if patients receive poorer quality in one type of hospital compared to another type, the analysis needs to adjust for differences in coexisting conditions (*see* **Co-morbidity**) and severity of illness between the two groups . However, process-based measures are implicitly risk-adjusted by restricting the sample to those patients who are known to benefit from treatments. In estimating use of needed beta-blocker therapy, AMI patients whose ejection fraction is less than 35%, who have bradycardia, or several other observed comorbidities (Table 1), are eliminated from the analysis.

Regardless of the risk-adjustment strategy, researchers commonly invoke ignorability assumptions. That is, given the data, quality is independent of the factor (e.g. hospital type) conditional on the observed covariates. While this assumption is more likely to hold when detailed clinical data are available, researchers need to assess the validity of the assumption and adopt appropriate modeling techniques (*see* **Model, Choice of**) for causal inference with observational data, such as **propensity scores**.

A common mistake in analyzing health care quality data has been to ignore important sources of variation due to the natural **clustering** of data, such as hospital-level variation. Ignoring clustering of individuals within health care units leads to overstated precision estimates and inflated type I error rates (*see* **Hypothesis Testing**) [23]. This problem can be substantial if the objective is to make inferences about the quality of health care units. To circumvent these problems, hierarchical regression models can be used to accommodate individual-level and cluster-level **covariates**, as well as cluster-level variation (*see* **Hierarchical Models**). While a **generalized estimating equation** approach has the advantage of making no distributional assumptions about the joint distribution of quality measures over the clusters, this approach is not applicable if the primary objective is to make inference about each cluster.

A third important analytic issue relates to the fact that health services researchers collect multiple quality measures for a particular population. The multiple quality measures could be longitudinal, such as hospital-specific 30-day mortality over a 5-year period (*see* **Longitudinal Data Analysis, Overview**). Methods have been proposed to model longitudinal health care quality data [2, 5, 19] that separate longitudinal variation from sampling variation.

Alternatively, the multiple measures could consist of mixed responses, such as 30-day mortality, reperfusion therapy, needed angiography, and health-reported quality. Health services researchers have either modeled the multiple quality measures separately or pooled the measures in some fashion. Separate analyses of quality measures have several drawbacks. The numerous measures can often confuse consumers with the information [8]. If researchers are interested in identifying factors that impact quality, separate analyses provide no formal means of evaluating how similar the effects are across the various measures. Furthermore, separate analyses may be inefficient if interest centers on estimating the association between a particular factor, such as hospital type, and quality.

Pooling quality measures, for example, score the quality measure as met if the patient received any needed therapy, also have some disadvantages. Researchers are unable to isolate effects of specific factors on each quality measure. It is also difficult to arrive at a rule when the measures are made on different scales such as **binary**, continuous, or categorical. Lastly, the optimal method for combining the measures depends on the measurement process. To accommodate multiple mixed measures, latent variable models can be used to model the **correlation** among the measures [10, 14], even if the measures are made on different scales [15] (*see* **Path Analysis**).

References

[1] Agency for Healthcare Research and Quality. *Consumer Assessment of Health Plans Survey*. Available at `http://www.ahcpr.gov/qual/cahpfact.htm`. Accessed December 29, 2003.

[2] Aguilar, O. & West, M. (1998). Analysis of hospital quality monitors using hierarchical time series models,

in *Case Studies in Bayesian Statistics*, IV, C.A. Gatsonis, R.E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli & M. West, eds. Springer-Verlag, New York, pp. 287–302.

[3] American Medical Association. *Physician Consortium for Performance Improvement*. Available at `http://www.ama-assn.org/ama/pub/category/2946.html`. Accessed December 29, 2003.

[4] Blumenthal, D. (1996). Quality of care - what is it? *New England Journal of Medicine* **335**(12), 891–894.

[5] Bronskill, S.E., Normand, S.-L.T., Landrum, M.B. & Rosenheck, R.A. (2002). Longitudinal profiles of health care providers, *Statistics in Medicine* **21**, 1067–1088.

[6] Corrigan, J.M., Eden, J. & Smith, B.M. (2002). *Leadership by Example: Coordinating Government Roles in Improving Healthcare Quality*. The National Academies Press, Washington.

[7] Donabedian, A. (1980). *Explorations in Quality Assessment and Monitoring. Volume 1: The Definition of Quality and the Approaches to its Assessment*. Health Administration Press, Ann Arbor.

[8] Epstein, A.E. (1998). Rolling down the runway. The challenges ahead for quality report cards, *Journal of the American Medical Association* **279**, 1691–1696.

[9] Fraser, I., McNamara, P., Lehman, G.O., Isaacson S., & Moler, K. (1999). The pursuit of quality by business coalitions: a national survey, *Health Affairs* **18**, 158–165.

[10] Gibbons, R.D. & Wilcox-Gok, V. (1998). Health service utilization and insurance coverage: a multivariate probit analysis, *Journal of the American Statistical Association* **93**, 63–72.

[11] Goldstein, F. & Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical Society, A* **159**, 385–443.

[12] Jencks, S.M. (1995). Measuring quality of care under medicare and medicaid, *Health Care Financing Review* **Summer**, 39–54.

[13] Landon, B.E., Normand, S.-L.T., Blumenthal, D. & Daley, J. (2003). Physician clinical performance assessment: prospects and barriers, *Journal of the American Medical Association* **290**(9), 1183–1189.

[14] Landrum, M.B., Bronskill, S.E. & Normand, S.-L.T. (2000). Analytic methods for constructing cross-sectional profiles of health care providers, *Health Services and Outcomes Research Methodology* **1**, 23–47.

[15] Landrum, M.B., Normand, S.-L.T. & Rosenheck, R.A. (2003). Selection of related multivariate means: monitoring psychiatric care in the department of veterans affairs, *Journal of the American Statistical Association* **98**, 7–16.

[16] Lohr, K.N. ed. (1990). *Medicare: A Strategy for Quality Assurance*. National Academy Press, Washington.

[17] Maxwell, J., Briscoe, F., Davidson, S., Eisen L., Robbins M., Temin P., & Young C. (1998). Managed competition in practice: 'value purchasing' by fourteen employers, *Health Affairs* **17**, 216–226.

[18] Maxwell, J., Briscoe, F., Watts, C., Zaman S., & Temin P. (2001). *Corporate Health Care Purchasing Among The Furtune 500*. National Health Care Purchasing Institute, Washington.

[19] McClellan, M. & Staiger, D. (1999). The Quality of Health Care Providers. Working Paper #7327, Boston. National Bureau of Economic Research.

[20] McGlynn, E.A. (1997). Six challenges in measuring the quality of health care, *Health Affairs* **16**(3), 7–21.

[21] Mehta R.H., Montoye C.K., Gallogly M., Baker P., Blount A., Faul J., Roychoudhury C., Borzak S., Fox S., Franklin M., Freundl M., Kline-Rogers E., LaLonde T., Orza M., Parrish R., Satwicz M., Smith M.J., Sobotka P., Winston S., Riba A.A., & Eagle K.A., GAP Steering Committee of the American College of Cardiology. (2002). Improving quality of care for acute myocardial infarction, *Journal of the American Medical Association*, **287**, 1269–1276.

[22] Meyer, J., Rybowski, L. & Eichler, R. (1998). *Theory and Reality of Value-Based Purchasing: Lessons from the Pioneers*. Agency for Health Care Policy and Research 98-0004, Washington.

[23] Normand, S.-L.T. & Zou, K.H. (2002). Sample size considerations in observational health care quality studies, *Statistics in Medicine* **21**, 331–345.

[24] Shahian, D.M., Normand, S.-L., Torchiana, D.F., Lewis, S.M., Pastore, J.O., Kuntz, R.E. & Dreyer, P.I. (2001). Cardiac surgery report cards: comprehensive review and statistical critique, *Annals of Thoracic Surgery* **72**, 2155–2168.

[25] Soumerai, S.B., McLaughlin, T.J., Gurwitz, J.H., Guadagnoli E., Hauptman P.J., Borbas C., Morris N., McLaughlin B., Gao X., Willison D.J., Asinger R., & Gobel F. (1998). Effect of local opinion leaders on quality of care for acute myocardial infarction: a randomized controlled trial, *Journal of the American Medical Association*, **279**, 1358–1363.

[26] Tillmann, I.A. (2000). *The Health Care Consumer as Purchaser: Shifting Dynamics*. National Health Care Purchasing Institute, Washington, Accessed December 30, 2003 at `www.nhcpi.net/pdf/tillmann-brief.pdf`.

[27] Wells, K.B., Sherbourne, C., Schoenbaum, M., Duan N., Meredith L., Unutzer J., Miranda J., Carney M.F., & Rubenstein L.V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care. A randomized controlled trial, *Journal of the American Medical Association*, **283**, 212–220.

SHARON-LISE T. NORMAND

Quality of Life and Health Status

Introduction

Clinical studies, epidemiological investigations, population surveys, and clinical practices increasingly incorporate self-reported measures of health status and quality of life. These help to determine whether treatments are doing more good than harm, whether health and quality are improving or worsening, and to see if health status differs between groups. Self-reported outcomes are often compared to clinical or performance-based measurements that remain the primary endpoints for most **clinical trials** and are important markers of disease, injury, and their trajectories. Self-reported measures of health and quality of life, however, often have more meaning to the persons affected by disease or treatment. Because perceptions of health and illness influence what people do about their health (e.g. visit doctors, go to hospital, or ignore signs and symptoms), policy makers are also increasingly interested in self-reported outcomes.

Concepts

Self-reported outcomes, referred to as *patient-reported outcomes* (*PROs*) in the context of health care, include any report coming directly from the person or persons affected by their life, health condition(s), and treatment [1]. PROs address the source of the report rather than the content. PROs not only include health status and quality of life but also reports on satisfaction with treatment and care, adherence to prescribed regimens when directly related to end result outcomes, and any other treatment or outcome evaluation obtained directly from patients through interviews, self-completed questionnaires, diaries or other data collection.

Health status, functional status, well-being, quality of life, and *health-related quality of life* are concepts that are often used interchangeably in PROs-related discourse and measurement. There is no consensus and widely adopted definition of *quality of life* because it is used in different contexts by different people. One definition is unlikely to suit all uses or individuals. There is considerable agreement, nonetheless, that the quality of life construct is more comprehensive than health status and includes aspects of the environment that may or may not be affected by health or perceived health. The *health status* concept and its domains and constructs range from negatively valued aspects of life, including death, to the more positively valued aspects such as role function and happiness. Health status is a useful concept in the context of assessing health services and treatment effectiveness. *Functional status* measures usually refer to limitations in the performance of social roles or activity limitations. The status concept is highly dependent on the perspective of the assessor and the assessed. *Well-being* measures refer to subjective perceptions, including reports of unpleasant or pleasant sensations and global evaluations of health or subjective status. Symptoms may be included in well-being measures or considered separately. Well-being and quality of life may be distinguished by the level of evaluation; that is, quality of life contains more global evaluations of life position and perspectives, and well-being contains more domain-specific perspectives such as psychological or physical [10]. It is important to note that PROs and quality of life are sometimes equated with functional status, and this labeling can be erroneous and of particular concern to persons with disabilities. Persons with functional limitations may enjoy high quality of life through environmental supports or simply through their own life perspective and evaluation of their needs and desires. Although function may be important to many evaluations of their health, health-related quality of life or quality of life should not be used as synonyms and these concepts should be identified and labeled separately, particularly when using terminology of PROs context of health care.

The boundaries of concepts and their definition depend upon the measurement objectives, the funding sponsors' motives, the users' particular concerns, and most important of all, the evidence or data on the concept and constructs [11]. Investigators may be interested in defining the health of populations to discover or document unmet needs, to determine the effect of medical interventions, or to guide allocation of resources. Traditional measures of morbidity and mortality are limited in defining health status and leave the texture of peoples' lives unexplored. Physiological measures, along the traditional endpoints in clinical trials, often bear limited relation to how affected persons are feeling. Thus, if one wishes to determine the impact of interventions on

the outcomes of real interest to the persons most affected, it is necessary to assess persons' experience through subjective evaluation and reporting of that experience.

It is important to note that some widely valued aspects of human existence are not generally defined as health status, such as a safe environment, adequate housing, guaranteed income, and freedom. These global human concerns, however, sometimes adversely affect or are affected by health status. Thus, the term *health-related quality of life* is often used to indicate that the measure is concentrated on the health concept and the field of health outcomes. *Quality of life*, however, may include all aspects of life, including the environment or externalities outside the context of healthcare.

The World Health Organization Quality of Life (WHOQOL) group, a worldwide research group organized by the **World Health Organization** has defined quality of life as "individuals" perceptions of their position in life in the context of the culture and value systems in which they live, and in relation to their goals, expectations, standards, and concerns" [20–22]. This definition reflects the growing recognition that quality of life can be inherently subjective, although normative definitions have been proposed that include more objective standards as well as perceptions of objective conditions [3, 4]. Quality of life can be used as a descriptor (i.e. the presence or absence of a characteristic of life), an evaluative statement (i.e. some value is attached to the characteristics of an individual, population, or kind of human life), or a normative or prescriptive statement, (i.e. certain norms indicate, which characteristics must be present to have a life of quality).

The WHOQOL group places quality of life squarely in the two traditions of an internal psychological and physiological mechanism producing a sense of satisfaction or gratification with life [9] and those external conditions that trigger the internal mechanism [16]. Thus, quality of life is a broad ranging concept that incorporates in a complex way, individuals' physical health, psychological state, level of independence, social relationships, personal beliefs, and their relationships to salient features of the environment [21]. Figure 1 shows this relationship between *health concepts* and *general* quality of life, and how determinants from the internal (individual) as well as the external (social and cultural) environment influence the general quality of life.

The concepts and domains included in the measurement of quality of life help in making operational definitions. Table 1 contains core concepts and domains contained in many health and quality of life measures. Table 1 and Figure 1 also demonstrate the multidimensionality of the concepts of health status and health-related quality of life, as a result require multiple indicators to measure.

How the items are arranged, and how the domains are grouped and scored varies widely. It is generally agreed that the content validity of quality-of-life measures can be judged *only* by the persons or populations being assessed (*see* **Health Status Instruments, Measurement Properties of**).

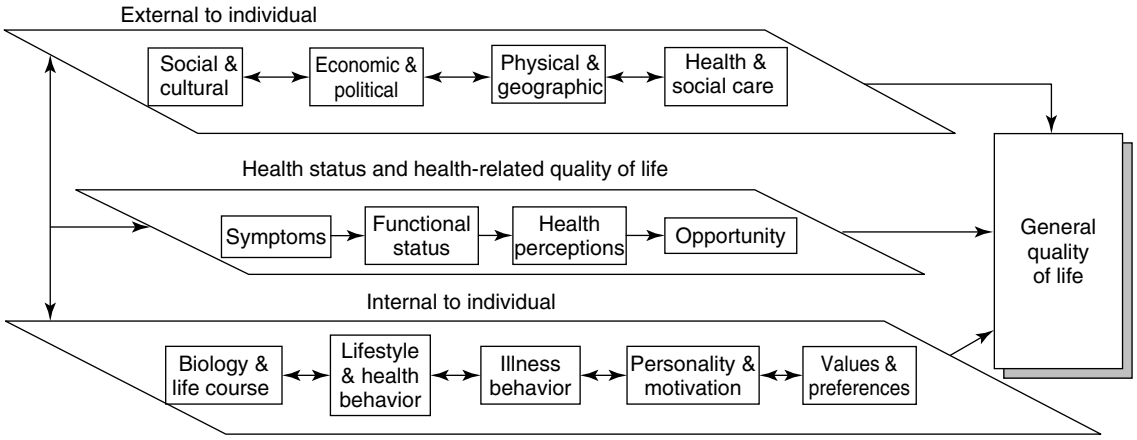

Figure 1 Relationship among quality of life and health concepts

Table 1 Concepts and domains used in defining self-reported health status, quality of life, and health-related quality of life

Concepts	Domains and *Attributes*
Symptoms	*Frequency, severity, bothersomeness* Reports of physical and psychological symptoms or sensations not directly observable, for example, energy and fatigue, nausea, irritability
Functional status	*Frequency, difficulty, severity, ability, with or without help*
Physical	Functional limitations and activity restrictions, for example, self care; walking, mobility, and sometimes sleep, and sexual function when construed narrowly
Psychological	Positive or negative affect and cognitive, for example, anger, alertness, self-esteem, sense of well-being, distress
Social	Engagement, limitations in work, school, play, household management, participation in the "community"
Health perceptions	*Frequency, severity/intensity, satisfaction*
Global	General ratings of health and quality of life, for example, satisfaction or overall well-being
Worries and concerns	About health, finances, the future
Spiritual	Meaning and purpose of life, connection to a deity, a belief system, or the universe
Disadvantage/Opportunity	*Frequency, impact* Perceptions of stigma or reports of discrimination because of health condition, reports of advantage
Resiliency	*Frequency, satisfaction, ability* Reports of ability to cope or withstand stress and illness
Environmental	*Satisfaction, importance* Evaluations of personal safety, adequacy of housing, respect, freedom, and so on
Satisfaction with treatment	*Expectations, importance, satisfaction* Reports of treatment and treatment experience
Adherence to prescribed or recommended treatment[a]	Behaviors directly linked to outcomes Reports of taking treatment, doses, attendance, or Routine behaviors like tooth-brushing intermediate to end results

[a]Included not as an end result of treatment but sometimes closely linked causally to treatment and thus a close proxy for a health status outcome. Reports of other behaviors like smoking, alcohol consumption, and so on are generally considered intermediate and not end result outcomes though this has been disputed by many.

Thus, the extent to which the domain of interest is comprehensively sampled by the items or questions in the measure can only be judged by representatives of the target population. If the target population is unable to speak for themselves, proxy judgments are considered acceptable, particularly if supported by rigorously controlled observational studies with inter-rater reliabilities. Before assuming that people cannot speak for themselves, however, they should be asked and every effort should be made to communicate with them directly. Proxy responses are not PROs.

In addition to content validity, the other **psychometric** properties of quality-of-life measures include the following: (1) specification of the measurement model including the instrument's scale and subscale structure and the conceptual and empirical basis for combining multiple items into a single score; (2) reliability, including the degree to which the instrument is free from **random error** either by testing the homogeneity of content on multi-item tests with internal consistency evaluation or testing the degree to which the instrument yields stable scores over time; (3) construct, criterion, and predictive validity wherein the logical relationships among different measures are examined; (4) responsiveness or the assessment of the ability of the measure to assess change over time when real change has occurred (longitudinal construct validation); and (5) interpretation of the effect

size, or the degree to which one can assign qualitative meaning to an instrument's quantitative scores [17].

Several concepts listed in Table 1 are considered important, but difficult to define and measure. Spirituality, often measured in end-of-life treatment evaluation and increasingly in other contexts, is rarely measured to determine generic health status even though life-threatening illness or orientation to health may be built on spiritual beliefs and practices. Resiliency can be measured as a self reported outcome of well-being, although physiologic measures may also be employed. Other concepts difficult to measure are capacity, stigma, disadvantage, and societal reaction, although self-report measures exist and major advances are being made in the analysis and interpretation of health disparities (`http://healthdisparities.nih.gov`). Both adherence to treatment and satisfaction with treatment can be measured by PROs. *Adherence* represents the patients' report of behaviors that coincide with medical or health advice, for example, to take medications, follow diets and exercise regimens, use medical devices when needed, or execute life style and behavioral change. Patient-reported measures of adherence are collected on self-reported questionnaires or interviews and often are compared to pill counts, physiological measures, or some other trace measure of adherence. Reports of adherence may be viewed both as predictors of health and treatment satisfaction or a function of health and satisfaction, and thus are not always "outcomes". Adherence is an intermediate outcome and not included as an end-result of treatment. It is discussed here to emphasize that it can be a PRO and is often considered in analyzing other outcomes.

Satisfaction with treatments and treatment experience, (i.e. care) represents salient aspects of treatment and overall evaluation of treatment experience. Ratings and reports of experience of treatment can be multidimensional or global. Global ratings of satisfaction are often positively skewed, that is patients report high levels of satisfaction, even in the face of other negative information. Rapid progress is being made to obtain patient reports of treatment satisfaction that also reflect the drivers of global satisfaction, such as expectations, outcomes, the gap between expectations and outcomes, and the importance of different attributes of treatment [14].

A large number of different examples of instruments and their development and validation process can be found in *Quality of Life Research, Journal of Clinical Epidemiology, Medical Care,* and many specialty journals. A detailed discussion of the psychometric properties and their evaluation are found in numerous places [2, 5, 6, 8, 13, 19].

Types of Measures

There are a number of ways of categorizing instruments designed to measure PROs such as health status and quality of life [7, 13]. Taxonomy of self-reported health status and quality of life measures is contained in Table 2. The measures are classified accordingly: (1) Source of the Report: information was gathered from the patient or proxy; (2) Mode of Data Collection: the data were collected through self administration, interviewer-administered, or computer administered tests; (3) Testing Content: use of adaptive or dynamic testing where the content varies for individuals and items are calibrated or standard content where everyone takes the same items; (4) Types of Scores: reflecting the level of aggregation across concepts and domains; (5) Range of Population and Concepts included or covered; and (6) Weighing System used in scoring items – whether an indicator, index, profile, or battery, instruments can be divided into two broad categories [12]. *Generic* instruments measure the full range of health and quality of life, without focusing on specific areas. They are designed for use across a wide variety of populations. *Specific* instruments are designed for application to individuals, conditions or diseases, domains, or populations. Generic and specific instruments may be *health profiles* or *utility measures* that are distinguished by having preference weights applied to the items and domains. Some utility measures, and indeed some profiles, also incorporate an index score or a single number for analyses. Utility measures, discussed more fully in **Utility in Health Studies**, are useful for economic applications, since they produce *quality-adjusted life years*, a combined measure of how long one lives as measured by survival or mortality and how well one lives, as measured by functional status and well-being.

Applications

Decision makers and analysts wanting to measure PROs should first identify the problem or application

Table 2 A taxonomy of self-reported health status and quality of life measures

Measure	Strengths	Weaknesses
Source of Report		
Person or patient	Sensations, feelings, evaluations known only to the person	When established that person cannot speak, write, or communicate to others
Proxy (Not a PRO)	Can observe and report behaviors only to patient	Cannot report feelings known
Mode of Collection		
Self-administered, with or without supervision	Privacy	Missing data, particularly by mail, and without supervision no assurance of who completed
Interviewer-administered	Control	Sensitive information like income sometimes difficult/cost
Computer-administered and/or computer-adapted	Flexibility	Cost/for persons not familiar or afraid of computer
Testing Strategy		
Dynamic or tailored content: based on health status, age, etc.	Items relevant to person More precise measurement	Requires item bank and item calibration
Standard or fixed-length content	Content same for all respondents Easy to administer	Many content items not relevant to individual Floor and ceiling effects
Types of Scores Produced		
Single *indicator* number	Global evaluation Sometimes easy to interpret	May be difficult to interpret trends
Single *index* number	Represents net impact Useful for cost effectiveness	Sometimes not possible to disaggregate contribution of domains to the overall score
Profile of interrelated scores	Single instrument Contribution of domains to overall score possible	Length may be a problem May not have overall score
Battery of independent scores	Wide range of relevant outcomes possible	Cannot relate different outcomes to common measurement scale May need to adjust for multiple comparisons May need to identify major outcome
Range of Populations and Concepts		
Generic: applied across diseases, conditions, populations, and concepts	Broadly applicable Summarizes range of concepts Detection of unanticipated effects possible	May not be responsive to change May not have focus of patient interest Length may be a problem Effects may be difficult to interpret
Specific: applied to individuals, diseases, conditions, populations, or concepts/domains	More acceptable to respondents May be more responsive to change	Cannot easily compare across conditions or populations Cannot detect unanticipated effects
Weighting System		
Utility: preference weights from patients, providers, or community	Interval scale Patient or consumer view incorporated	May have difficulty obtaining weights May not differ from statistical weights that are easier to obtain
Equal-weighting: items weighted equally or from frequency or responses	Self-weighting samples More familiar techniques Appears easier to use	May be influenced by prevalence Cannot incorporate tradeoffs

of the measure. With this information, one can then identify the desired characteristics of existing measures to be included in the assessment. For example, monitoring the health of populations and communities demands parsimonious instruments including global evaluations across a number of conditions and different population groups. For comprehensive evaluation in a clinical trial, health profiles or batteries are most appropriate according to the main effects intended and unintended or adverse consequences of treatment. For economic evaluation, utility measures are useful to produce a comparison across alternative treatments.

PROs directly from children and youth are taking greater prominence among all interested parties, following similar development as that for adults and older adults [18]. Children and youth represent a special case, however, given knowledge of variation in how and when children develop, the wide variation in willingness and ability to self-report across the age spectrum, and the "special" language of children and adolescents in different cultures of the world. Rapid progress is being made in Europe and North America, to be followed and informed by work in other parts of the world, often less accessible to Western parties.

Demand is also increasing for quality of life measures available for use in cross-cultural comparisons, which requires special attention to cultural adaptation and validation in each culture in which the measure is applied [15]. The most desirable means of development and validation is to have the goal of cross-cultural comparability in mind from the beginning. Measures developed simultaneously in different cultures have the advantage of identifying as early as possible those domains and items that are more or less valid in a particular culture or population. Translating instruments developed in one culture for use in another is more common, but the danger of this approach is the assumption that the conceptual structure, domains, and items are cross-cultural. For example, assessments of functional status that use examples such "ability to walk several blocks" run the danger that "blocks" have different if any meaning at all in different cultures. Response scale anchors, such as "quite a bit", also do not translate easily into different languages.

Testing of the psychometric properties of cross-cultural measures is similar to that for instruments used within one language or cultural group, although standards for aggregation across sites have not been rigorously applied. When it is and is not valid to use measures in different populations and to pool data across different cultures remains an area for further investigation and debate.

Conclusion

Quality of life measurements are important for measuring the impact of disease, treatment, health and social policies, and the progress of economic and social development. Developers and users should specify and label the content and type of measure for every application of a PRO and provide evidence of its appropriateness to the intended use, for validity of the measure as used in a particular case, and how to interpret results. A major challenge faces developers and user of these measures in establishing a testable theory of the expected and observed relationships among the different concepts and domains of quality of life. It is also important to establish a theory of how to link clinical variables with health-related quality of life as it is to link larger determinants of PROs such as political unrest, economic depression, inequalities, and sociocultural trends and processes [13, 23].

Researchers tend to approach the relationship inductively by collecting data and examining the correlations, but *a priori* hypotheses are important for developing systematic knowledge of how disease and treatment impacts different indicators of health outcome. The most appropriate approach to causal modeling, the use of health outcomes in **meta-analyses**, development and application of community-level indicators of health, and interpretation of observed health and quality-of-life measurements remain challenges for both the developers and uses of these measures.

References

[1] Acquadro, C., Berzon, R., Dubois, D., Leidy, N.K., Marquis, P., Revicki, D. & Rothman, M.; PRO Harmonization Group. (2003). Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the patient-reported outcomes (PRO) harmonization group meeting at the food and drug administration, *Value Health*, **6**(5), 522–531.

[2] Bowling, A. (1991). *Measuring Health: A Review of Quality of Life Measurement Scales*. Open University Press, Milton Keynes.

[3] Calman, K.C. (1987). Definitions and dimensions of quality of life, in *The quality of life of cancer patients*, N.K. Aaronson & J. Beckman, eds. Raven Press, New York, pp. 1–9.
[4] Campbell, A., Converse, P.E. & Rodgers, W.L. (1976). *Quality of American life: Perceptions, Evaluations and Satisfaction*. Russell Sage Foundation, New York.
[5] Fairclough, D.L. (2002). *Design and Analysis of Quality of Life Studies in Clinical Trials*. CRC Press, Boca Raton.
[6] Fayers, P. & Machin, D. (2000). *Quality of Life: Assessment, Analysis, and Interpretation*. John Wiley & Sons, New York.
[7] Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D. & Cox, D. (1992). Quality of life measures in health care. I: application and issues in assessment. *British Medical Journal* **305**, 1074–1077.
[8] Guyatt, G.H., Feeney, D.H. & Patrick, D.L. (1993). Measuring health-related quality of life, *Annals Of Internal Medicine* **118**, 622–629.
[9] Hornquist, J.O. (1982). The concept of quality of life, *Scandinavian Journal of Social Medicine* **10**(2), 57–61.
[10] Kahnemann, D. & Diener, E. & Schwartz, N. (1999). *Well-Being: The Foundations of Hedonic Psychology*. Russell Sage Foundation, New York.
[11] Patrick, D.L. & Bergner, M. (1990). Measurement of health status in the 1990s, *Annual Review of Public Health* **11**, 165–183.
[12] Patrick, D.L. & Deyo, R.A. (1989). Generic and disease-specific measures in assessing health status and quality and life, *Medical Care* **27**(Suppl. 3), S217–S232.
[13] Patrick, D.L. & Erickson, P. (1993). *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. Oxford University Press, New York.
[14] Patrick, D.L., Martin, M., Bushnell, D. & Pesa, J. (2003). Measuring satisfaction with migraine treatment: expectations, importance, outcomes, and global rating, *Clinical Therapeutics* **25**(11), 2920–2935.
[15] Patrick, D.L., Wild, D.J., Johnson, E.S., Wagner, T.H. & Martin, Ma. (1994). Cross-cultural validation of quality-of-life measures, in *Quality of Life Assessment: International Perspectives*, J. Orley & W. Kuyken, eds. Springer-Verlag, Berlin, Heidelberg, pp. 19–32.
[16] Rogerson, R.J. (1995). Environmental and health-related quality of life: conceptual and methodological similarities, *Social Science and Medicine* **41**(10), 1373–1382.
[17] Scientific Advisory Committee, Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria, *Quality of Life Research* **11**(3), 193–205.
[18] Starfield, B. (1996). Health status measurement: the special case of children and youth [editorial], *Injury Prevention* **2**(2), 86–87.
[19] Streiner, D.L. & Norman, G.R. (1995). *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford University Press, New York.
[20] Szabo, S. (1996). The world health organization quality of life (WHOQOL) assessment instrument, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, B. Spilker, ed. Lippincott-Raven Publishers, Philadelphia, pp. 355–362.
[21] The WHOQOL Group. (1994). The development of the world health organization quality of life assessment instrument (WHOQOL), in *Quality of Life Assessment: International Perspectives*, J. Orley & W. Kuyken, eds. Springer-Verlag, Berlin, Heidelberg, p. 41.
[22] The WHOQOL Group. (1995). The world health organization quality of life assessment (whoqol): position paper from the world health organization, *Social Science & Medicine* **41**(10), 1403.
[23] Wilson, I.B. & Cleary, P.D. (1995). Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes, *Journal of the American Medical Association* **273**(1), 59–65.

Further Reading

Patrick, D.L. & Chiang, Y.P. (2000). Measurement of health outcomes in treatment effectiveness evaluations: conceptual and methodological challenges, *Medical Care* **38**(Suppl. 9), II14–II25.

DONALD L. PATRICK

Quality of Life and Survival Analysis

In clinical research, interest often centers around the survival times of patients $T(> 0)$, or, more generally speaking, the times from a suitable starting point such as, for example, time of diagnosis or randomization to the occurrence of an end point such as death or disease recurrence. For these kind of data the methodology of **survival analysis** is usually applied.

In recent years an additional end point concerned with the subjective assessment of treatment by the individual patient has been introduced which is circumscribed with the term "(health-related) *quality of life* ((H)QoL)". A large body of literature is available covering the many aspects of how to adequately and sensibly evaluate an approximation of a patient's QoL (*see*, for example, [4–7, 15], and [16]; **Quality of Life and Health Status**).

The most commonly used approach of measuring QoL is by means of standardized, self-assessed questionnaires, popular examples being the EORTC QLQ-C30 of the European Organization for Research and Treatment of Cancer, the Rotterdam Symptom Checklist (RSCL), or the Functional Assessment of Cancer Therapy (FACT). The data emerging from assessments with one of these measuring instruments consist of patients' answers to about 20 to 50 questions with mostly preformulated, ordinal answering categories. For the analysis of QoL these questions are usually condensed by suitable methods, e.g. summation over correlated questions, to yield one global, unidimensional QoL score Q or several subscores Q_j corresponding to specific QoL dimensions (*see* **Psychometrics, Overview**).

As the evaluation of changes of QoL over time is of most importance in **clinical trials**, repeated measurements on the individual patient are essential for an adequate assessment of QoL. The least necessary requirement commonly accepted is to assess QoL at baseline before treatment starts, then during and shortly after treatment to account for short-term effects, and also some time after treatment has stopped to evaluate late effects. This describes the classical measurement situation of a QoL-incorporated clinical trial, considering in parallel the length of survival via clinical follow-up and the quality of this survival by continuing QoL assessment.

Two analysis approaches are principally applicable here: analyzing length and quality of survival as two separate end points, or jointly as one combined end point. In the following we assume for simplicity that our QoL measuring instrument will provide at times $t > 0$ a unidimensional QoL variable $Q(t)$ with values standardized between a lowest possible score of 0 and a highest possible score of 1.

Separate Analysis of QoL Data Over Time

When collecting QoL data over time, the adoption of the experimental design of a classical **analysis of variance** (ANOVA) accounting for repeated measurements within patients is warranted (*see* **Analysis of Variance for Longitudinal Data**). Depending on whether normality of QoL scores, possibly after suitable **transformation**, can be assumed or not, **general linear models** or **generalized linear models** are appropriate. For standard applications in this situation the statistical methodology has been worked out and software packages like **S-PLUS**, SAS, and BMDP can be used. A review of the most important procedures with special reference to QoL data can be found in [4, 12], and [17]. However, some problems specific to QoL data have to be encountered, of which the most influential departure from straightforward applications is to deal with the situation of missing QoL data. Although **missing data** can be regarded a common problem in longitudinal sampling plans, the assumption of missing at random is generally more questionable for QoL data (*see* **Nonignorable Dropout in Longitudinal Studies**; **Bias from Nonresponse**). It has been suggested in this context that patients with either an extremely good or an extremely poor QoL will have a higher likelihood of refusing to respond to a QoL questionnaire. In the extreme, questionnaires not available due to mortality clearly are not missing at random.

Estimates can be severely biased, especially when complete case analyses are routinely applied, but also when methods are used that account for data missing at random.

Combined Analysis of Length and Quality of Survival

The separate analysis of survival times and QoL may lead to situations where a univocal decision

about treatment efficacy appears problematic especially when mortality is not negligible.

In a combined analysis of both end points, QoL data will be used in a first step to define, implicitly or explicitly, a **stochastic process** describing patients' QoL levels over time. In a direct application, the parameters of the process using either continuous or discrete time, together with a continuous or a discrete state space incorporating an absorbing death state, could be estimated depending on the availability of corresponding QoL data. In an indirect way the states of the process could be used as QoL weights for the survival times spent in these states.

Time to QoL Defined Events

The most straightforward application of a combined analysis of time and quality is by definition of an additional QoL-oriented end point such as, for example, the reaching of a deterioration in QoL of a certain amount q_0. The time T_q until this state is reached may be measured with an appropriate sampling plan and then analyzed by the classical methods of survival analysis. In this way it is possible to assess a patient's time spent alive with an acceptable QoL.

In a comparison of two radiation strategies for brain metastases in lung cancer patients, Rosenman & Choi [14] defined as such an end point the time to the first occurrence of a Karnofsky index of 60% or below. They compared product-limit estimates (*see* **Kaplan–Meier Estimator**) for the overall survival probabilities with those obtained using the times to the QoL-oriented end point. Their analysis resulted in a nonsignificant trend of favoring one treatment over the other with regard to overall survival, which drastically reverses when survival with a "good" QoL is considered. This also represents the most interesting scenario for a QoL analysis, when treatments with comparable efficacy in the primary end points can be shown to differ in an important secondary end point.

However, the choice of suitable QoL values for the definition of states usually will be arbitrary, but nevertheless might be influential for the results.

Such an approach may be extended by defining more than one QoL state in between the optimal QoL state and death. Suitable stochastic processes may be defined by modeling the times spent in the different states and the transition probabilities from one state to another. Markov-type models have been proposed for comparable situations [12] (*see* **Markov Chains**). In a full Markov model it is assumed that transitions from one QoL state to another are independent and depend only on the previously occupied state. In practice, however, the methodology of stochastic processes is applied to QoL data only rarely. The major reason is that the continuous observation of patients' QoL required for obtaining exact transition times can be regarded as infeasible in practical clinical trial settings.

Quality-Adjusted Survival Times

From the viewpoint of survival analysis it seems appealing to combine length of survival and QoL into one single end point described as *quality-adjusted survival (QAS) time*. Originally, QAS times have been introduced in the field of decision analysis where they are usually called quality-adjusted life years (QUALYs) [16].

This approach is again, but now rather implicitly, based on the stochastic process formulation of the QoL process. J different QoL states are defined only for the purpose of producing weights q_j accompanying the time T_j the patient spent in that state. QAS times are then defined by multiplying each period of the individual survival time with the weight corresponding to the QoL assessment reported by the patient, or a general **utility** assessment, and then summing these weighted times:

$$\mathrm{QAS} = \sum_{j-1}^{J} q_j T_j.$$

In this way the number of different time variables T_j representing the transition times from one state to another are condensed into one time variable representing a new quality-adjusted time scale. The conventional survival time of a patient T can be regarded as a special case of a QAS time giving constant weight equal to 1 for all time spent in the alive state and 0 after death.

A comparable QAS approach was chosen by Korn [11], who allows individual QoL measurements $Q(t)$ at arbitrary points in time $0 = t_0 < t_1 < \cdots < t_J = T$ with $Q(t) = 0$ for $t > T$, and suggests plotting $Q(t)$ vs. time t to allow a graphical representation of the development of a patient's QoL over time. The area under this quality of life curve (AUC) created by

linear interpolation of QoL between adjacent QoL measures $Q(t_j)$ and $Q(t_{j+1})$ can be interpreted either as a weighted sum of partial survival times,

$$\text{AUC} = \sum_{j=1}^{J} \frac{(t_j - t_{j-1})[Q(t_j) - Q(t_{j-1})]}{2},$$

or equivalently as a time-weighted sum of QoL scores,

$$\text{AUC} = \frac{Q(t_0)(t_1 - t_0)}{2} + \sum_{j=2}^{J-1} \frac{Q(t_j)(t_{j+1} - t_{j-1})}{2}.$$

The most important application of QAS times has been introduced by Gelber et al. [8] with their definition of TWiST (time without symptoms and toxicity) or, later, Q-TWiST (quality-adjusted TWiST). TWiST was proposed as an additional new end point in a clinical trial of adjuvant therapy in patients with advanced breast cancer. The apparent benefit of chemo- and/or endocrine therapy with respect to disease-free and overall survival is thereby balanced against the treatments' toxicity and side-effects. Formally, both the times in which a patient suffered severe side-effects during the treatment period T_{TOX} and the times after a relapse T_{REL} are subtracted from his/her total survival time T. In the terminology of QAS this is equivalent to attaching a weight of 0 to both states:

$$\begin{aligned} T_{\text{TWiST}} &= 0 \times T_{\text{TOX}} + 1 \times T_{\text{TWiST}} + 0 \times T_{\text{REL}} \\ &= T - T_{\text{TOX}} - T_{\text{REL}} \end{aligned}$$

In Q-TWiST, weights q_j greater than 0 will allow positive partial QAS times from the states toxicity and relapse:

$$\begin{aligned} T_{\text{QTWiST}} &= q_{\text{TOX}} \times T_{\text{TOX}} + 1 \times T_{\text{TWiST}} \\ &\quad + q_{\text{REL}} \times T_{\text{REL}} \\ &= T - (1 - q_{\text{TOX}})T_{\text{TOX}} - (1 - q_{\text{REL}})T_{\text{REL}}. \end{aligned}$$

A natural approach for an analysis would be to use the QAS time of each patient instead of the conventional survival time and then apply the usual methods of survival analysis. A simple treatment comparison could, for example, be based on a comparison of the distribution functions of the QAS times for each treatment $\Pr(\text{QAS} > t)$. If all QAS times were observed, this would be unproblematic.

However, Gelber et al. [8] noted that, when censored observations occur, the use of QAS times can lead to seriously biased estimates of the corresponding QAS probabilities. They show that transforming the original survival time scale to a QAS time scale introduces informative **censoring**. A reason for this bias is the fact that patients with low QoL weights can only slowly accumulate their QAS time and will therefore more likely have earlier censoring than those with higher QoL weight. The inclusion of a relatively large proportion of patients with poor QoL in the "early" risk set leads to an underestimation of the corresponding hazard function for QAS time. In a simple, hypothetical, example, Glasziou et al. [9] illustrate the severity of the bias.

Korn [11] proposes a modified Kaplan–Meier estimate for QAS that aims to reduce this bias. His conditional independence estimator (CIE) is derived under the assumption that the censoring distribution in small time intervals is independent of the QoL score just before that interval. In practice, the more frequent QoL is assessed, the more bias reduction might be achieved. However, the CIE still remains asymptotically biased, but the bias will always be smaller than that of the naive product-limit estimate.

Another drawback of the CIE is that Korn's assumption of conditional independence will in general not be fulfilled, and it will remain difficult to quantify the amount of bias for a specific trial based on the trial data alone. Recently Zhao & Tsiatis [18] for the first time proposed an asymptotically **consistent** estimator for the distribution of QAS times, applying the method of weighted estimating equations (*see* **Estimating Functions**) according to Robins & Rotnitzky [13].

One solution that avoids the bias in the estimation of the distribution of QAS times introduced by the informative censoring is by a partitioned survival analysis [7]. This is principally applicable when QoL states can be defined such that patients may pass the QoL states only in descending order. In this application QAS times are not calculated on an individual basis, but, instead, mean marginal transition times for each health state are calculated by integrating over the corresponding survival functions. If censored observations are present, these means have to be restricted to some upper limit $M < t_{\max}$ being the largest censored survival time. In a second

step, weighted sums of these means are calculated according to the definition of the individual QAS times, again using the corresponding weight of each health state q_j. For Q-TWiST this leads to

$$E(T_{\text{Q-TWiST}}) = \int_0^M S_{\text{OS}}(t)\,\mathrm{d}t - (1 - q_{\text{TOX}}) \times \int_0^M S_{\text{TOX}}(t)\,\mathrm{d}t - (1 - q_{\text{REL}}) \times \int_0^M S_{\text{REL}}(t)\,\mathrm{d}t.$$

Estimates of these restricted means are obtained by inserting the corresponding product-limit estimators $\hat{S}_j(t) = \Pr(T_j > t)$. Mean QAS times estimated for different patient groups can be used for univariate treatment comparisons [9].

In a number of papers Cole et al. [1–3] have extended the methodology of partitioned survival analysis to adapt to further models used in survival analysis. To allow for the inclusion of **covariates** additional to treatment **Cox's** (proportional hazards) **regression models** have been proposed for Q-TWiST [2]. To overcome some of the limitation of the restricted means, parametric models have been proposed to predict Kaplan–Meier estimates beyond the restriction time M [1]. One of the advantages of the Q-TWiST methodology is the ability to perform **meta-analyses** over different trials without having used the same QoL questionnaire. The methodology and an example using data from eight trials of adjuvant chemotherapy trials in breast cancer is presented in [3].

If the choice of appropriate weights for the health states is in doubt, then a threshold utility analysis, originally called "inverse inference" by Hilden [10], provides an informative way to display how changes in the QoL weights q_j influence trial results based on the definition of corresponding QAS times. In the simple Q-TWiST model of two QoL states, treatment comparisons are calculated within a particular data set in dependence of all possible combinations of the QoL weights (q_{REL}, q_{TOX}). The results of this estimation process may be displayed graphically in a plane spanned by (q_{REL}, q_{TOX}). A threshold line identifying those values (q_{REL}, q_{TOX}) that yield treatment equivalence with regard to QAS, as well as areas of significant treatment differences, could be highlighted.

It is interesting to note in this context that the edges of the unit square of (q_{REL}, q_{TOX}) correspond to the situations usually considered in survival analysis: for $q_{\text{TOX}} = 1$ and $q_{\text{REL}} = 1$, QAS time reduces to the overall survival time, for $q_{\text{TOX}} = 1$ and $q_{\text{REL}} = 0$ to disease-free survival time and for $q_{\text{TOX}} = 0$ and $q_{\text{REL}} = 0$ to TWiST. This allows the conventional end points of survival analysis to be directly linked to all theoretically possible QAS times. The major advantage of a threshold utility analysis is that it allows individual patients or their physicians to estimate a preferable treatment according to their individual choices of weights for the QoL states.

Decisions on an overall treatment superiority might also be based on an optimality of one treatment over another for all possible weights, or at least in a markedly larger area.

Discussion

The subjective assessment of the impact of treatments on the individual patient using QoL questionnaires is becoming standard practice in clinical research. The adequate sampling plan for obtaining QoL data is that of a classical repeated measurement design (*see* **Longitudinal Data Analysis, Overview**). A suitable analysis of such data could rely on well-known procedures based on linear or generalized linear models if mortality or censoring did not occur. However, in the majority of clinical trials, censored observations will be inevitable, and mortality usually the primary end point. Treating QoL measures after disease recurrence or death as missing data, or assigning it an arbitrary low value, may lead to uninterpretable results. In this situation an integration of QoL measurements into the conventional survival analysis methodology that accounts for censoring seems sensible.

With the definition and use of QAS times, however, a number of statistical and conceptual problems arise. The obvious definition of a QAS time scale for an individual introduces informative censoring which prevents a direct adaption of methods of survival analysis to QAS time. At the moment the availability of suitable unbiased statistical estimation procedures is limited.

From a conceptual point of view the use of QAS times is the subject of much criticism, too. The major criticism is that this procedure leads to an explicit equation of qualitatively differing years of survival.

Assume the four different scenarios of: (i) three years spent with perfect QoL ($q_j = 1$); (ii) four years with a moderately impaired QoL ($q_j = 0.75$); (iii) six years with medium QoL ($q_j = 0.5$); and (iv) 15 years with severely impaired QoL ($q_j = 0.2$). Each of these scenarios leads formally to the same QAS time of three years. Forcing these scenarios of different life experiences into numerical equivalence can be regarded as extremely simplifying. Applied routinely, e.g. for allocating resources in health care programmes, QAS times may lead to far-reaching consequences [16] (*see* **Health Services Research, Overview**). It is therefore necessary that any QAS time analyses should always be accompanied by the corresponding, conventional survival analyses, and should not rely on one simple set of QoL weights, but be subject to additional **sensitivity analyses** [4].

References

[1] Cole, B.F., Gelber, R.D. & Anderson, K.M., for the International Breast Cancer Study Group (1994). Parametric approaches to quality adjusted survival analysis, *Biometrics* **50**, 621–631.

[2] Cole, B.F., Gelber, R.D. & Goldhirsch, A., for the International Breast Cancer Study Group (1993). Cox regression models for quality adjusted survival analysis, *Statistics in Medicine* **12**, 975–987.

[3] Cole, B.F., Gelber, R.D. & Goldhirsch, A. (1995). A quality-adjusted survival meta-analysis of adjuvant chemotherapy for premenopausal breast cancer, *Statistics in Medicine* **14**, 1771–1784.

[4] Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Spiegelhalter, D.J. & Jones, D.R. (1992). Quality of life assessment: can we keep it simple?, *Journal of the Royal Statistical Society, Series A* **155**, 353–393.

[5] Fairclough, D.L., (2002). *Design and analysis of quality of life studies in clinical trials*. Chapman and Hall/CRC Press, Boca Raton, Florida.

[6] Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D. & Cox, D. (1992). Quality of life measures in health care. I: Applications and issues in assessment, *British Medical Journal* **305**, 1074–1077.

[7] Fletcher, A., Gore, S., Jones, D., Fitzpatrick, R., Spiegelhalter, D. & Cox, D. (1992). Quality of life measures in health care. II: Design, analysis, and interpretation, *British Medical Journal* **305**, 1145–1148.

[8] Gelber, R.D., Gelman, R.S. & Goldhirsch, A. (1989). A quality-of-life oriented endpoint for comparing therapies, *Biometrics* **45**, 781–795.

[9] Glasziou, P.P., Simes, R.J. & Gelber, R.D. (1990). Quality adjusted survival analysis, *Statistics in Medicine* **9**, 1259–1276.

[10] Hilden, J. (1987). Reporting clinical trials from the viewpoint of a patient's choice of treatment, *Statistics in Medicine* **6**, 745–752.

[11] Korn E.L. (1993). On estimating the distribution function for quality of life in cancer clinical trials, *Biometrika* **80**, 535–542.

[12] Olschewski, M. & Schumacher, M. (1990). Statistical analysis of quality of life data in cancer clinical trials, *Statistics in Medicine* **9**, 749–763.

[13] Robins, J.M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, in *AIDS Epidemiology-Methodological Issues*, N. Jewell, K. Dietz & V. Farewell, eds. Birkhäuser, Boston, pp. 24–33.

[14] Rosenman, J. & Choi, N.C. (1982). Improved quality of life of patients with small-cell carcinoma of the lung by elective irradiation of the brain, *International Journal of Radiation Oncology Biology Physics* **8**, 1041–1043.

[15] Schumacher, M., Olschewski, M. & Schulgen, G. (1991). Assessment of quality of life in clinical trials, *Statistics in Medicine* **10**, 1915–1930.

[16] Spiegelhalter, D.J., Gore, S.M., Fitzpatrick, R., Fletcher, A.E., Jones, D.R. & Cox, D.R. (1992). Quality of life measures in health care. III: Resource allocation, *British Medical Journal* **305**, 1205–1209.

[17] Zee, B. & Pater, J. (1991). Statistical analysis of trials assessing quality of life, in *Effect of Cancer on Quality of Life*, D. Osoba, ed. CRC Press, Boca Raton, pp. 113–124.

[18] Zhao, H. & Tsiatis, A.A. (1997). A consistent estimator for the distribution of quality adjusted survival time, *Biometrika* **84**, 339–348.

MANFRED OLSCHEWSKI

Quality of Life

Background

The term *quality of life* (QoL) has been used in a wide variety of ways. In the broadest definition, the quality of our lives is influenced by our physical and social environment as well as our emotional and existential reactions to that environment. From a societal or global perspective, measures of QoL may include social and environmental indicators, such as whether there is affordable housing and how many days of air pollution there are each year in a particular location. These are general issues that concern everyone in a society. Kaplan and Bush [22] proposed the use of the term *health-related* **quality of life** (HRQoL) to focus on the specific role of health effects on the individual's perceptions of well-being, distinguishing these from job satisfaction and environmental factors. In the medical literature, the terms QoL and Health-related quality of life (HRQoL) have become interchangeable.

Health-related Quality of Life

The **World Health Organization** (WHO) defined *health* in 1948 [38, 39] as a "state of complete physical, mental, and social well-being and not merely the absence of infirmity and disease". This definition reflects the focus on a broader picture of health. Wilson and Cleary [37] propose a conceptual model of the relationships among health outcomes. There are five levels of outcomes that progress from biomedical measures to quality of life reflecting the WHO definition of health. The biological and physiological outcomes include the results of laboratory tests, radiological scans, and physical examination as well as diagnoses. Symptom status is defined as "a patient's perception of an abnormal physical, emotional, or cognitive state". Functional status includes four dimensions: physical, physiological, social, and role activity. General health perceptions include the patients' evaluation of past and current health, their future outlook, and concerns about health. All these factors subsequently influence the overall evaluation of quality of life (see Figure 1).

Although various definitions of HRQoL have been proposed during the past decade, there is general agreement that HRQoL is a multidimensional concept that focuses on the *impact* of disease and its treatment on the well-being of an individual. Cella and Bonomi [2] state

> "Health-related quality of life refers to the extent to which one's usual or expected physical, emotional and social well-being are *affected* by a medical condition or its treatment."

We may also include other aspects like economic and existential well-being. Patrick and Erickson [29] propose a more inclusive definition that combines quality and quantity.

> "the value assigned to duration of life as modified by the impairments, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment or policy."

All of these definitions emphasize the subjective nature of the evaluation of HRQoL, with a focus on its assessment by the individual. It is important to note that an individual's well-being or health status cannot be directly measured. We are only able to make inferences from measurable indicators of symptoms and reported perceptions.

Often the term *quality of life* is used when any *patient-reported outcome* is measured. This has led to both confusion and controversy. Side effects and symptoms are not equivalent to HRQoL, although clearly they influence an individual's evaluation of quality of life. While symptoms are often part of the assessment of HRQoL, solely assessing symptoms is a simple, convenient way of avoiding the more complex task of assessing HRQoL.

Measuring Health-related Quality of Life

Guyatt et al. [19] define an *instrument* to include the questionnaire, the method of administration, instructions for administration, the method of scoring and analysis, and interpretation for a **health status measure**. All these aspects are important when evaluating a measure of HRQoL.

Health Status versus Patient Preferences

There are two general types of HRQoL measures, health status assessment, and patient preference assessment [33, 40]. The development of these two forms is a result of the differences between the perspectives of two different disciplines:

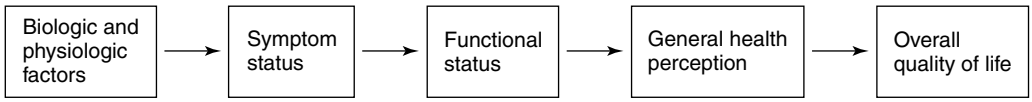

Figure 1

psychometrics and **econometrics**. In the health status assessment measures, multiple aspects of the patient's perceived well-being are self-assessed and a score is derived from the responses on a series of questions. This score reflects the patient's relative HRQoL compared with other patients and to the HRQoL of the same patient at other times. These measures are primarily designed to compare groups of patients receiving different treatments or to identify change over time within groups of patients. As a result, these measures have been used in clinical trials to facilitate the comparisons of therapeutic regimens. The assessments range from a single global question asking patients to rate their current quality of life to a series of questions about specific aspects of their daily life during a recent period of time. Among these health status measures, there is considerable range in the context of the questions with some measures focusing more on the perceived impact of the disease and therapy (How much are you bothered by hair loss?), other measures focusing on the frequency and severity of symptoms (How often do you experience pain?), and still others assessing general status (How would you rate your quality of life?).

Measures in the second group, patient preferences, are influenced strongly by the concept of *utility* (*see* **Utility in Health Studies**) borrowed from econometrics, which reflects individual decision making under uncertainty. These preference assessment measures are primarily used to evaluate the trade-off between the quality and quantity of life. Values of utilities are always between 0 and 1 with 0 generally associated with death and 1 with perfect health. Examples include **time trade offs** [24], **standard gamble** [32], and *multiattribute assessment measures* [5, 10]. Time trade-off utilities are measured by asking respondents how much of the time they expect to spend in their current state would they give up for a reduced period of time in perfect health. If, for example, a patient responded that he would trade five years in his current state for four years in perfect health (trading one year), the resulting utility is 0.8. Standard gamble utilities are measured by asking respondents to identify the point at which they become indifferent to the choices between two hypothetical situations. Suppose a patient is presented with two treatment alternatives, one option is a radical surgical procedure with no chance of relapse but significant impact on HRQoL and the other option is watchful waiting, with a chance of progressive disease and death. The chance of progressive disease and death is raised or lowered until the respondent considers the two options to be equivalent. Assessment of time trade-off and standard gamble utilities requires the presence of a trained interviewer or specialized computer program. Because of these resource needs, these approaches are generally too time- and resource-intensive to use in a large clinical trial. Multiattribute assessment measures combine the advantages of self-assessment with the conceptual advantages of utility scores. Their use is limited by the need to develop and validate the methods by which the multiattribute assessment scores are converted to utility scores for each of the possible health states defined by the multiattribute assessments. For example, the EuroQoL scale, also known as the EQ-5D, is a standardized non–disease-specific instrument for describing and evaluating HRQoL [3]. The EQ-5D covers five dimensions of health: mobility, self-care, role (or main) activity, family and leisure activities, pain and mood. Within each dimension, the respondent chooses one of three items that best describes his or her status. Weights are used in scoring the responses, reducing the 243 (3^5) possible health states to a single utility score. Utilities have traditionally been used in the calculation of quality-adjusted life years (QALYs) for economic evaluation (cost-effectiveness) and policy research as well as in analytic approaches such as Q-TwiST [14, 15]. It is important to note that the utility one gives to a hypothetical situation has been seen to vary from what the individual gives when the situation is real; results of any analysis should be interpreted carefully from that perspective.

Objective versus Subjective

Health status measures differ among themselves in the extent to which they assess observable phenomena

or require the respondent to make inferences. These measures may assess symptoms or functional benchmarks wherein individuals are asked about the frequency and severity of symptoms or whether they can perform certain tasks such as walking a mile. The measures may also, or instead, assess the impact of symptoms or conditions by asking individuals how much the symptoms *bother* them or *interfere* with their usual activities. Many instruments provide a combination. The value of each will depend on the research objectives: Is the focus to identify the intervention with the least severity of symptoms or to assess the impact of the disease and its treatment?

There has been considerable discussion of whether subjective assessments are less valid and reliable than objective measures. This misconception is generally based on the observation that patient ratings do not always agree with ratings of trained professionals. If we take the ratings of these professionals as constituting the **gold standard**, we are ignoring the valuable information of how the patient views his or her health and quality of life, especially the aspects of emotional and social functioning. There is measurement error in both subjective and objective assessments; neither is necessarily more accurate or precise in all circumstances. Most widely used measures of HRQoL are the product of careful development resulting in a measure that is highly reliable, sensitive to change with good **predictive** validity and minimal measurement error. In contrast, some of the biomedical endpoints that we consider objective can include a demonstrably high degree of measurement error (e.g. blood pressure), **misclassification** among experts, or have poor predictive and prognostic validity (e.g. pulmonary function tests) [36].

Generic versus Disease-specific Instruments

There are two basic types of health status measures – generic and disease-specific. The generic instrument is designed to assess HRQoL in individuals with and without active disease, and across disease types (e.g. heart disease, diabetes, depression, cancer). The Medical Outcomes Study Short Form (MOS SF-36) is an example of a generic instrument [35]. The broad item content of a generic instrument is an advantage when comparing vastly different groups of subjects or following subjects for extended periods after treatment has ended. Disease-specific instruments narrow the scope of assessment and address in a more detailed manner, the impact of a particular disease or treatment (e.g. joint pain and stiffness in patients with arthritis or treatment toxicities in patients with cancer) [12]. As a result, they may be more sensitive to smaller, but clinically significant changes induced by treatment [28].

Global Index versus Profile of Domain-specific Measures

HRQoL measures come in a variety of forms reflecting their intended use. The major distinction is between an index and a profile. *Profiles* consist of multiple scales that reflect the multiple dimensions of QoL such as the physical, emotional, functional, and social well-being of patients. In most instruments, each scale is constructed from the responses to multiple questions (often referred to as items). Two methods of construction are used for the creation of *indices*. In the first, a single question is used to assess the subject's assessment of quality of life. In the latter, developers provide methods to combine responses to multiple questions to provide a single index of QoL.

The advantage of the single index is that it provides a straightforward approach to decision making, which may be required in settings such as clinical trials where QoL is the primary outcome. Indices that are in the form of utilities are used in **cost-effectiveness** analyses performed in pharmacoeconomic research. On the other hand, a profile of the various domains reflects the multidimensional character of quality of life [1]. There are limitations that should be considered when using either approach. First, there is always a set of "values" being imposed when an index is constructed. These values may come from each individual's concept of what is meant by quality of life when a single question is asked, or the values that a developer assigns to the construction of the index. The values may be as arbitrary as the number of questions that are used to assess each domain, or statistically derived to maximize discrimination among different groups of patients. It is impossible to construct an index that aggregates the multiple dimensions of HRQoL that will be suitable in all contexts. The important point is that one should be aware of the weights (values) that are placed on the different domains in the interpretation of the results. A single index measuring HRQoL cannot capture changes in individual domains. For example, a particular intervention may produce benefits in one dimension and

deficits in another that cancel each other and are thus not observed in the aggregated score.

Response Format

Questionnaires may also differ in their response format. The most widely used format is the **Likert scale**, which contains a limited number of **ordered responses** that have a descriptive label associated with each level. Variations include scales in which only the extremes are anchored with a descriptive label. Individuals can discriminate at most seven to ten ordered categories [25, 31] and reliability and sensitivity to change drops off at five or fewer levels.

Dichotomous response formats and visual analog scales (VAS) (*see* **Pain**) are also used. The VAS consists of a line, generally 10 cm in length, with descriptive anchors at each end of the line. The respondent is instructed to place a mark on the line. The original motivation of the VAS was that the continuous measure could potentially discriminate more effectively than a Likert scale; this has not generally been true in most **validation studies** where both formats have been used. The VAS format has several limitations. It requires a level of eye–hand coordination that may be unrealistic for anyone with a neurological condition, those experiencing numbness and tingling side effects of chemotherapy, and for the elderly. VAS precludes telephone assessment and interview formats. Finally, it requires an additional data management step in which the position of the mark is measured. If forms are copied rather than printed, the full length of the line may vary, requiring two measurements and additional calculations. A compromise format is a numerical analog where patients provide a number between 0 and 100 (see Table 1).

Period of Recall

QoL scales often request individuals to base their evaluation over a specified period, such as the last seven days or the last four weeks. The time frame must be short enough to detect differences between treatments and long enough to minimize short-term fluctuations (noise) that do not represent real change [26]. In addition, the reliability with which individuals can rate aspects of their QoL beyond several weeks must be called into question. Scales specific to diseases or treatments where there can be rapid changes will have a shorter recall duration, whereas instruments designed for assessment of general populations will often have a longer recall duration. Longer time frames may also be appropriate when assessments are widely spaced (e.g. annually).

Scoring

The majority of HRQoL scales that are derived from a series of questions with a Likert response format are scored by summing or averaging the responses after reverse coding negatively worded questions. There are more complicated weighting schemes based on **factor analytic** weights; item response or **Rasch models** are rare but may become more common in **computer assisted testing**. To facilitate interpretation, there has been an increasing tendency to rescale this result so that the possible range of responses is 0 to 100 with 100 reflecting the best possible outcome. Most instruments also have an explicit strategy for scoring in the presence of **missing responses** to a small proportion of questions. The most common method is to impute the missing response using the average of the other responses in the specific subscale when at least half of the questions have been answered.

Table 1 Example of a Likert and visual analog scale

Likert Scale					
	Not at all	Slightly	Moderately	Quite a bit	Greatly
How bothered were you?	0	1	2	3	4
Visual analog scale					
How bothered were you?	|- -|				
	Not at all				Greatly

In contrast, utilities are always on a 0 to 1 scale. Scoring depends on the method used to elicit the preferences. Scores derived from multiattribute assessment are instrument specific.

Development and Validation of HRQoL Measures

Development

The effort and technical expertise required to develop a new instrument is generally underestimated, with most efforts taking three to five years (or more) rather than the couple of weeks initially expected. Researchers contemplating this step should research the existing instruments as well as the various methodologies for instrument development and validation including traditional psychometric theory and item response theory. To fully develop an instrument from the beginning requires multiple studies, hundreds of observations, years of testing and refinement.

Validation

There are numerous procedures for establishing the psychometric properties of an instrument. For a formal presentation, the reader is referred to one of the many available books. A partial list specific to HRQoL includes Streiner and Norman [31], McDowell and Newell [23], Juniper et al. [21] and Naughton et al. [27], Frank-Stromborg and Olsen [11], Staquet, Hays, and Fayers [30] and Fayers and Machin [9].

The validity of a measure in a particular setting is the most important and the most difficult aspect to establish. This is primarily because HRQoL is an unobservable latent variable (*see* **Path Analysis**) and there are no gold standards against which the empirical measures of validity can be compared. Nonetheless, we can learn a good deal about an instrument by examining the instrument itself and the empirical information that has been collected. For example, we can demonstrate that the measure behaves in a manner that is consistent with what we would expect and correlates with observable things that are believed to be related to HRQoL.

Face validity refers to the content of an instrument: Does the instrument measure what it proposes to measure? and Are the questions comprehensible and without ambiguity? The analogy is whether an archer has chosen the intended target. The wording of the questions should be examined to establish whether the content of the questions is relevant to the population of interest. Although experts (physicians, nurses) may make this evaluation, it is advisable to verify the face validity with patients as they may have a different perspective. *Criterion validity* is the strength of a relationship between the scale and a gold standard measure of the same construct. As there is no gold standard for the dimensions of quality of life, we rely on the demonstration of *construct validity*. This is the evidence that the instrument behaves as expected and shows similar relationships (*convergent validity)* and the lack of relationships (*divergent validity*) with other reliable measures for related and unrelated characteristics (*see* **Health Status Instruments, Measurement Properties of**). Confirmatory factor analysis **structural equation modeling** is one of the statistical methods used to support the construct validity or proposed structure (subscales) of an instrument. Application may be used to confirm that a scale is unidimensional. Results from exploratory factor analysis (*see* **Exploratory Data Analysis**) in selected populations should be interpreted cautiously especially when the sample is homogeneous with respect to stage of disease or treatment [8].

The next question is: Would a subject give the same response at another time, if they were experiencing the same HRQoL? This is referred to as *test–retest reliability* (*see* **Agreement, Measurement of**). If there is a lot of variation (noise) in responses for subjects experiencing the same level of HRQoL, then it is difficult to discriminate between subjects who are experiencing different levels of HRQoL or change in HRQoL over time. This is generally measured using Pearson or intraclass **correlations** when the data consists of two assessments. Finally, we ask: Does the instrument discriminate among subjects who are experiencing different levels of HRQoL? and Is the instrument sensitive to changes that are considered important to the patient? These characteristics are referred to as *discriminant validity* and *responsiveness*. Reliability can be characterized using the analogy of the archer's ability to hit the same target repeatedly with consistency. *Internal consistency* (*see* **Validity and Generalizability in Epidemiologic Studies**) refers to the extent to which items in the same scale (or subscale) are interrelated;

specifically the extent to which responses on a specific item increase as the responses to other items on the scale increase. **Cronbach's** coefficient α is typically reported. For the assessment of group differences, values above 0.7 are generally regarded as acceptable though values above 0.8 (good) are often recommended. For assessment of individual patients in clinical practice, it is recommended that the value should be above 0.9.

Responsiveness is the ability of a measure to detect changes that occur as the result of an intervention [16, 17]. Here the analogy is whether the archer can respond to change and hit various areas on the target consistently. One factor that can affect responsiveness is a *floor* or *ceiling effect*. If responses are clustered at either end of the scale, it may not be possible to detect change due to the intervention.

Translation/Cross-cultural Validation

When HRQoL is measured in diverse populations, attention needs to be paid to the methods of translation and cross-cultural validation. Backward and forward translations must be performed using the appropriate native language at each step. There are numerous examples where investigators have found problems with certain questions as questionnaires are validated in different languages and cultures. Techniques such as cognitive testing, with subjects describing verbally what they are thinking as they form their responses, have been very valuable when adapting a questionnaire to a new language or culture. This should be followed by formal validation studies designed to generate both standard reliability and validity statistics. Item response theory (IRT) methods (*see* **Rasch Models**) and Rasch models have facilitated the examination of differential item functioning across cultures or languages.

Item Banking and Computer-adaptive Testing

There has been considerable effort over the last decade to develop *item banks,* large databases of responses to individual questions from multiple questionnaires. One objective of this effort is to establish a method of translating results obtained on one scale to another scale. The second application, referred to as *computer-adaptive testing*, is an attempt to reduce the subject burden during testing by selectively presenting questions that will best discriminate in the range of function of that subject. Specifically, if an individual's response to the first question indicates a high level of functioning, that individual will not be presented with questions designed to discriminate among individuals with low levels of functioning. In both cases, IRT methods play a predominant role.

Use in Research Studies

> "Implicit in the use of measures of HRQoL, in clinical trials and in effectiveness research, is the concept that clinical interventions such as pharmacologic therapies, can affect parameters such as physical function, social function, or mental health." [37]

All principles of good design and analysis are applicable, but there are additional requirements specific to HRQoL. These include selection of an appropriate measure of HRQoL and the conduct of an assessment to minimize any bias. The HRQoL instruments should be selected carefully, ensuring that they are appropriate to the research question and the population under study. New instruments and questions should be considered only if all other options have been eliminated. Among the most common statistical problems are **multiple endpoints** and **missing data**.

Instrument Selection

Ware et al. [34] suggest two general principles to guide the selection of instruments to discriminate among subjects or detect change in the target population. "When studying general populations, consider using positively defined measures. Only some 15% of general population samples will have chronic physical limitations and some 10 to 20% will have substantial psychiatric impairment. Relying on negative definitions of health tells little or nothing about the health of the remaining 70 to 80% of general populations. By contrast, when studying severely ill populations, the best strategy may be to emphasize measures of the negative end of the health status continuum."

In cases where the population is experiencing periods of health and illness, very careful attention must be paid to the selection of the instrument balancing the ability to discriminate among subjects during different phases of their disease and treatment with appropriateness over the length of the study. One cannot assume that a questionnaire that works well in

one setting will work well in all settings. For example, questions about the ability to perform the tasks of daily living, which make sense to individuals who are living in their own homes, may be confusing when administered to a patient who has been in the hospital for the past week or is terminally ill and receiving hospice care. Similarly, questions about the amount of time spent in bed provide excellent discrimination among ambulatory subjects, but not among hospitalized patients.

There is a temptation to pick an HRQoL instrument, become familiar with it, and use it in all circumstances. Flexibility must be maintained in the choice of instrument to target the specific research or clinical setting, the specific population, the challenges associated with administration, and the problem of respondent burden.

Multiple Endpoints

Because QoL is a multidimensional concept that is generally measured using several scales that assess functional, physical, social, and emotional well-being, there are multiple endpoints associated with most QoL evaluations. **Longitudinal data** arise in most HRQoL investigations because we are interested in how a disease or an intervention affects an individual's well-being over time. Because of the multidimensional nature of HRQoL and repeated assessments over time, research objectives need to be explicitly specified and an analytic strategy developed for handling multiple endpoints. Although adequate for univariate outcomes such as survival, statements such as "To compare the quality-of-life of subjects on treatments A and B" are insufficient; details should include domains, population, and the time frame relevant to the research questions. Strategies addressing the **multiplicity** of endpoints include limiting confirmatory analyses, construction of summary measures/statistics [6] and **multiple comparison** procedures. Examples of summary measures that reduce multiplicity over time include area-under-the-curve (AUC) and average rates of change (slope); their interpretation is straightforward. Construction of these measures is complicated by the presence of missing data. QoL indices are used to reduce the multiplicity across domains.

Missing Data

Although analytic strategies exist for missing data, their use is much less satisfactory than initial prevention. Some missing data, such as that due to death, is not preventable; however, missing data should be minimized at both the design and implementation stages of a clinical trial [7, 26, 41].

The protocol and training materials should include specific procedures to minimize missing data. A practical schedule with HRQoL assessments linked to planned treatment or follow-up visits can decrease the number of missing HRQoL assessments. When possible, it is wise to link HRQoL assessments with other clinical assessments.

Interpretation/Clinical Significance

All new measures take time to become useful to clinicians or patients. This process requires that we define ranges of values that have clinical implications. When measures such as hemoglobin and blood pressure were first used, there was a period during which normal ranges were established; once the ranges were available, the readings became clinically useful. Nor are the rules that have been developed simple, since the benefits/risks of a change in either measure depends on where the individual started, age, gender, and current condition or lifestyle (pregnancy or about to run a marathon). Interpretation of measures of QoL is similarly complex.

Clinical significance has various meanings depending on the setting. When a treatment decision is required, there is an implied ordering of information into categories (often dichotomous) that correspond to various decisions. For example, one might consider a patient's current hemoglobin as well as gender and clinical history when making a decision to treat, monitor closely, or do nothing. In contrast, when the situation calls for evaluation of effectiveness of an intervention based on the information from a randomized clinical trial, the decision is generally based on continuous or ordered information such as grams/dL of hemoglobin. Thus, meaningful differences/changes in QoL measures will depend on whether a decision is being made for an individual or for a group of individuals. There are two general strategies used to define the *clinical significance* of QoL scores, distribution-based and anchor-based methods. There

is no single approach that is appropriate to all settings and none of the methods is without some limitations.

Distributional Methods

One general approach is based on the distribution of scores expressed as the relationship (ratio) between the magnitude of an effect and a measure of variability [3, 18]. The magnitude of the effect may be either the difference between two groups or the change within a group. Measures of variability include the **standard deviation** of a reference group, the standard deviation of change, and the **standard error** of measurement. A distributional method was used by Cohen [4] in his criteria for meaningful effect or "effect size" (*see* **Outcome Measures in Clinical Trials**) in psychosocial research. The major advantage is that values are relatively easy to generate from validation studies or clinical trials. There are a number of limitations. Many clinicians are unfamiliar with "effect size" and skeptical about defining meaningful differences solely on the basis of distributions. These values are generally applicable to groups. Measures of variability can differ across studies being affected by the selection criteria, which can influence the heterogeneity of the sample. Finally, one still needs to make a decision about the size of the effect that is relevant in any particular setting, requiring a value judgment of risks and benefits.

Anchor-based Methods

Anchor-based methods are based on the relationship between scores on the QoL measure and an independent measure or anchor. Examples of anchors are the patient's rating of health, disease status, and treatments with known efficacy. The anchor must be interpretable and there needs to be an appreciable association of the anchor with QoL. Within this group of methods, there are numerous approaches, none of which fits all needs. One concern is that the motivation for QoL measurement is to move beyond traditional clinical endpoints, but we appear to be using these same clinical endpoints to "justify" and interpret QoL measures.

One approach is to classify subjects into groups based on the anchor, and estimate differences in the QoL measures. For example, one might form three groups based on function corresponding to no, moderate, or severe limitations and observe average scores of 80, 70, and 50 respectively. The mirror image of that approach is to classify subjects using QoL measures and describe the outcomes in terms of either an external or internal anchor. In the first case, one might observe that a group of patients with a mean score of 80 experience 5% mortality, while another group of patients with a score of 60 experience 20% mortality. In the latter case, one might observe that 32% of those who score 50 on the SF-36 physical function scale can walk a block without difficulty, in contrast to 50% who score 60.

Another approach is to elicit a value, a minimum important difference (MID) from clinicians or patients; that is, the smallest difference in the scores that is perceived as important, either beneficial or harmful, and which would lead the clinician to consider a change in the patient's management [20]. Within-patient transitions are yet another approach that has been used. Individuals are asked to judge, during a specified time, whether they have improved, not changed, or gotten worse. The corresponding changes in QoL scores are then summarized within each group. The advantage of this approach is that it is easy to assess and appears simple to interpret. However, there is an accumulation of evidence that the retrospective assessment reflects the subject's current QoL rather than the change.

Conclusions

In the health sciences, QoL assessment is now an integral component of patient-focused research. Given the increasing complexity of health care, the extent of chronic illness, and the variety of therapies that generally do not improve survival but often only decrease morbidity, measurement of HRQoL outcomes provides an additional evaluation of treatment benefit. These developments have occurred in the latter half of the twentieth century, a period in which individual preferences and autonomy have been increasingly valued in many societies, especially in Europe and North America. In parallel, reliable and valid measurement strategies have evolved from social science research to make it possible to quantify subjective assessments of health status and QoL. Further, advances in statistical methodology have been integrated into research designs, making it possible to interpret these assessments in a variety of research and clinical settings. A major aspect of this

work has been to bridge the gap between psychometric/statistical theory and the language and realities of clinical practice. There are many who remain skeptical about the contributions of QoL assessments to treatment decisions and health care policies; however, the more these measurements are integrated into research, the greater the likelihood that the outcomes that matter to patients will ultimately be incorporated into medical care [13]. Many of the studies that have failed to detect changes have suffered from nonignorable missing data and the use of inappropriate analytic methods. Therefore, statisticians have an important role to play in the design and analysis of studies with QoL outcomes, if the studies are to produce interpretable results. In this article, we have provided a perspective on where we are in QoL assessment, as well as an honest evaluation of some of the limitations of this methodology. Much additional work needs to be done, and fortunately there is considerable international interest in addressing the challenges of this young measurement science.

References

[1] Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N.J., Filiberti, A., Flechtner H., Fleishman, S.B. & de Haes J.C. (1993). The European organization for research and treatment of cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology, *Journal of the National Cancer Institute* **85**, 365–376.

[2] Cella, D.F. & Bonomi, A.E. (1995). Measuring quality of life: 1995 update, *Oncology* **9**, 47–60.

[3] Cella, D., Bullinger, M., Scott, C. & Barofsky, I. (2002). Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life, *Mayo Clinic Proceedings* **77**, 384–392.

[4] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

[5] EuroQol Group (1990). EuroQol – A new facility for the measurement of health-related quality of life, *Health Policy* **16**, 199–208.

[6] Fairclough, D.L. (1997). Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy, *Statistics in Medicine* **16**, 1197–1209.

[7] Fairclough, D.L. & Cella, D.F. (1996). Eastern cooperative oncology group (ECOG), *Journal of the National Cancer Institute Monographs* **20**, 73–75.

[8] Fayers, P.M. & Hand, D.J. (1997). Factor analysis, causal indicators and quality of life, *Quality of Life Research* **6**, 139–150.

[9] Fayers, P.M. & Machin, D. (2000). *Quality of Life: Assessment, Analysis and Interpretation*. John Wiley & Sons, UK, Chapters 3–7.

[10] Feeny, D., Furlong, W., Barr, R.D., Torrance, G.W., Rosenbaum, P. & Weitzman, S. (1992). A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer, *Journal of Clinical Oncology* **10**, 923–928.

[11] Frank-Stromborg, M. & Olsen, S. (1997). *Instruments for Clinical Health-care Research*. Jones and Bartlett, Boston.

[12] Ganz, P.A. (1990). Methods of assessing the effect of drug therapy on quality of life, *Drug Safety* **5**, 233–242.

[13] Ganz, P.A. (2002). What outcomes matter to patients: a physician-researcher point of view, *Medical Care* **40**, III11–III19.

[14] Glasziou, P.P., Simes, R.J. & Gelber, R.D. (1990). Quality adjusted survival analysis, *Statistics in Medicine* **9**, 1259–1276.

[15] Goldhirsch, A., Gelber, R.D., Simes, R.J., Glasziou, P. & Coates, A.S. (1989). Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis, *Journal of Clinical Oncology* **7**, 36–44.

[16] Guyatt, G.H., Deyo, R.A., Charlson, M., Levine, M.N. & Mitchell, A. (1989). Responsiveness and validity in health status measurement: a clarification, *Journal of Clinical Epidemiology* **42**, 403–408.

[17] Guyatt, G.H., Kirshner, B. & Jaeschke, R. (1992). Measuring health status: What are the necessary measurement properties? *Journal of Clinical Epidemiology* **45**, 1341–1345.

[18] Guyatt, G.H., Osoba, D., Wu, A.W., Wyrwich, K.W. & Norman, G.R. (2002). Methods to explain the clinical significance of health status measures, *Mayo Clinic Proceedings* **77**, 371–383.

[19] Guyatt, G.H., Patrick, D., Feeny, D. (1991). Glossary, *Controlled Clinical Trials* **12**, 274S–280S.

[20] Jaeschke, R., Singer, J. & Guyatt, G.H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference, *Controlled Clinical Trials* **10**, 407–415.

[21] Juniper, E.F., Guyatt, D.H. & Jaeschke, R. (1996). How to develop and validate a new health-related quality of life instrument, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, B. Spilker, ed. Lippincott-Raven Publishers, Philadelphia, pp. 49–56.

[22] Kaplan, R. & Bush, J. (1982). Health-related quality of life measurement for evaluation research and policy analysis, *Health Psychology* **1**, 61–80.

[23] McDowell, I. & Newell, C. (1996). *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford University Press, New York.

[24] McNeil, B.J., Weichselbaum, R. & Pauker, S.G. (1981). Speech and survival: tradeoffs between quality and quantity of life in laryngeal cancer, *The New England Journal of Medicine* **305**, 982–987.

[25] Miller, G.A. (1956). The magic number seven plus or minus two: some limits on our capacity for information processing, *Psychological Bulletin* **63**, 81–97.

[26] Moinpour, C.M., Feigl, P., Metch, B., Hayden, K.A., Meyskens, F.L. Jr. and Crowley, J. (1989). Quality of life end points in cancer clinical trials: review and recommendations, *Journal of the National Cancer Institute* **81**, 485–495.

[27] Naughton, M.J., Shumaker, S.A., Anderson, R.T. & Czajkowski, S.M. (1996). Psychological aspects of health related quality of life measurement: test and scales, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, B. Spilker, ed. Lippincott-Raven Publishers, Philadelphia, pp. 117–132.

[28] Patrick, D.L. & Deyo, R.A. (1989). Generic and disease-specific measures in assessing health status and quality of life, *Medical Care* **27**, S217–S232.

[29] Patrick, D.L. & Erickson, P. (1993). *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. Oxford University Press, New York.

[30] Staquet, M.J., Hays, R.D. & Fayers, P.M. (1998). *Quality of Life Assessment in Clinical Trials: Methods and Practice Part II*. Oxford University Press, Oxford, New York, Chapters 2–5.

[31] Streiner, D.L. & Norman, G.R. (1995). *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford University Press, Oxford, New York.

[32] Torrance, G.W., Thomas, W.H. & Sackett, D.L. (1971). A utility maximizing model for evaluation of health care programs, *Health Services Research* **7**, 118–133.

[33] Tsevat, J., Weeks, J.C., Guadagnoli, E., Tosteson, A.N., Mangione, C.M., Pliskin, J.S., Weinstein M.C., Cleary, P.D. (1994). Using health-related quality-of-life information: clinical encounters, clinical trials, and health policy, *Journal of General Internal Medicine* **9**, 576–582.

[34] Ware, J.E. Jr., Brook, R.H., Davies, A.R. & Lohr, K.N. (1981). Choosing measures of health status for individuals in general populations, *American Journal of Public Health* **71**, 620–625.

[35] Ware, J.E. Jr. & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection, *Medical Care* **30**, 473–483.

[36] Wiklund, I., Dimenas, E. & Wahl, M. (1990). Factors of importance when evaluating quality of life in clinical trials, *Controlled Clinical Trials* **11**, 169–179.

[37] Wilson, I.B. & Cleary, P.D. (1995). Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes, *JAMA* **273**, 59–65.

[38] World Health Organization (1958). *The First Ten Years of the World Health Organization*. World Health Organization, Geneva.

[39] World Health Organization. *Constitution of the World Health Organization*. Basic Documents 48. World Health Organization, Geneva.

[40] Yabroff, K.R., Linas, B.P. & Schulman, K. (1996). Evaluation of quality of life for diverse patient populations, *Breast Cancer Research and Treatment* **40**, 87–104.

[41] Young, T. & Maher, J. (1999). Collecting quality of life data in EORTC clinical trials–what happens in practice? *Psycho-oncology* **8**, 260–263.

DIANE L. FAIRCLOUGH & PATRICIA A. GANZ

Quantal Response Models

In 1981 a new group of anti-parasite drugs, called avermectins, was successfully introduced to treat internal parasites in cattle. However, the drugs remain in low concentrations in the cattle dung, and there was concern that this would also kill the organisms that degrade the dung. The type of experiment needed to investigate this is a quantal response experiment: a particular organism was selected – the common yellow dung fly – and exposed in groups to different concentrations of the drug for a fixed time. The result of such an experiment is then a set of proportions of the kind displayed (for a different example) in Figure 1.

Quantal response data arise in many studies. If one is interested in an effect on human beings, such as a possible carcinogenic response, then typically the substance investigated would be applied to an alternative animal, in the hope that the conclusions there could be extrapolated to different animals of interest. Care is needed in the selection of appropriate animal models. Quantal data also occur in industrial reliability testing, and from observational studies, an example of which is shown in Figure 1. The response of interest here is the onset of menarche in young girls, grouped into a range of different age-bands. It is obviously far easier to design an experimental study than an observational study. Experiments need to be tailored to the type of organism involved, and can be limited by considerations of space, time, and expense – an experiment on insects would be conducted in a very different way from one on large mammals.

The quantal response experiment has obvious similarities with a standard regression setup, but the response variable at each dose/age-group, etc. is now discrete (*see* **Binary Data**). Frequently it can be assumed that individuals respond independently, both within and between the different dose levels, etc. resulting in a likelihood function which is a product of binomial probabilities. There are many variations on the simple quantal response experiment described here, and we examine several of these later. For instance, a response may be discrete but not binary, as when fetuses are classified as dead, alive, or deformed; animals may not respond independently, which could occur if they are housed in the same cage, or come from the same litter; doses administered may differ from the nominal level, as could arise when the substance is administered through food or in the field; organisms may respond "naturally" – for example handling mortality can be high in insect experiments, but negligible in studies of large animals; we may be interested in response over time as well as with respect to dose.

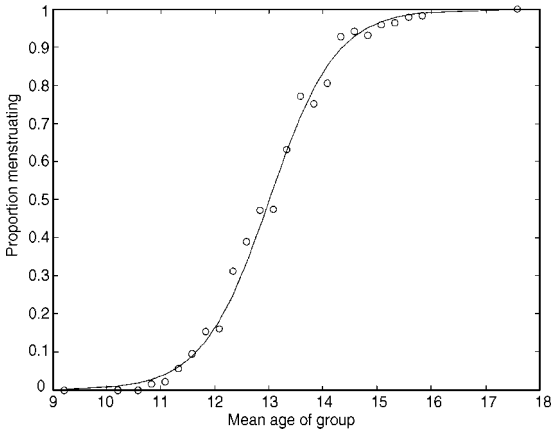

Figure 1 For each of a sample of 3918 Warsaw girls, taken in 1963, it was recorded in [29] whether or not they had reached menarche (started menstruating). The girls were divided into groups according to age, and plotted are the proportions in each group that had reached menarche v. mean age of group. The fitted curve results from a logit model, fitted by maximum likelihood

The routine testing of new substances for good behavior as potential new drugs to market regularly produces a sequence of quantal response experiments. Frequently a wide range of doses is then used, as response is usually uncertain before the experiment is conducted. However, although each experiment may be treated on its own, there is often an element of comparison involved, as when a new substance is a minor modification of an earlier one. Quantal response studies may be conducted explicitly in order to make a comparison, e.g. to investigate whether a rural environment results in an overall delay in onset of menarche, when compared with an urban environment (it does), or whether smoking advances the age of menopause (it does). To describe a set of quantal response data, or to make a comparison between two or more such sets of data, it is useful to think in terms of a simple summary of experiments.

The dung fly experiment mentioned above was thus described in the British newspaper *The Independent* (on September 23, 1996) as follows:

> ... half of the larvae of the common yellow dung fly died when exposed to just 0.05 parts of Ivermectin [the most effective avermectin] per million. Lower concentrations caused major disruption to the fly's life-cycle. Cattle dung from bolus-administered cattle contains 10 times this concentration of the drug.

The ED_{50} or **median effective dose** is the dose level that results in an expected 50% response under certain underlying model assumptions, and is the summary that has been used in the above newspaper article. In some applications, such as studies of the toxic effects of food additives, or in the evaluation of insecticides, other values such as $ED_{10^{-6}}$, or ED_{99} would, respectively, be more appropriate summaries.

Notation and the Spearman–Kärber Estimate

We assume that k doses are tested, that there are n_i individuals exposed to the ith dose, and r_i of those individuals respond.

Independence assumptions result in the **likelihood**

$$L = \prod_{i=1}^{k} \binom{n_i}{r_i} P_i^{r_i} (1 - P_i)^{n_i - r_i}, \qquad (1)$$

where P_i is the probability of response at the ith dose. In most cases, response increases as the dose level increases, and P_i is modeled by means of a cumulative distribution function, $P_i = F(d_i)$ (*see* **Random Variable**). This model is sometimes given a *threshold* or *tolerance* interpretation, as follows: any individual is assumed to have its own dose tolerance threshold, T, responding to dose d if and only if $d \geq T$. Thus if T is distributed throughout the population of individuals as a random variable with cumulative distribution function $F(t)$, then the probability that any individual responds at dose d is simply

$$\Pr(T \leq d) = F(d).$$

There are various situations where this interpretation is not appropriate, e.g. in situations where a response can result from infection by a single virus particle (*see* **Infectivity Titration**). However, the threshold model is often useful, and we see an example of this later when we discuss models for multivariate response.

The ED_{50} is now seen to be the median of $F(\cdot)$, and also its mean if $F(\cdot)$ is a symmetric function about the mean. A very simple estimate of the ED_{50} results if we make this symmetry assumption and if also the dose range has been chosen wide enough that $r_1 = 0$ and $r_k = n_k$. This is known as the Spearman–Kärber estimate of the ED_{50}, which takes its simplest form when the doses are equally spaced by an amount Δ, namely,

$$\widehat{\mathrm{ED}}_{50} = d_k + \frac{\Delta}{2} - \Delta \sum_{i=1}^{k} \left(\frac{r_i}{n_i} \right). \qquad (2)$$

For the case of $n_i \geq 2$ for all i, an unbiased estimate of $\mathrm{var}(\widehat{\mathrm{ED}}_{50})$ is then given by

$$\sum_{j=2}^{k-1} \frac{r_j (n_j - r_j) \Delta^2}{n_j^2 (n_j - 1)}. \qquad (3)$$

Extension and elaborations are described in [32].

It is convenient to possess an explicit estimate of the variance of a key quantity of interest. Experimental resources are usually limited by considerations not only of cost, space, and time, but also of limiting animal suffering and mortality. Experiments may then be designed in order to try to minimize the variance in the context of a fixed total number, $\sum_{i=1}^{k} n_i$, of animals. We discuss experimental design in more detail below. The Spearman–Kärber estimate can be made more robust by a small amount of suitable trimming, such as 5%; see [18] and [32]. The Spearman–Kärber approach is nonparametric, because it makes no assumptions about the form of the function $F(\cdot)$, other than the assumption of symmetry. Much more complex nonparametric methods now exist – see, for example, [22] for a method based on **density estimation**. We now consider parametric methods which historically came later than the Spearman–Kärber estimate of the ED_{50}. Parametric methods can make strong assumptions regarding the form of $F(\cdot)$ and this can in turn result in large increases in the precision with which, for example, the ED_{50} is estimated. This is graphically demonstrated in [16]. The advantage of a parametric approach is the flexibility with which departures from standard experimental procedures can be accommodated.

Models

Parametric models for quantal response data involve assuming a form for $F(\cdot)$. The simplest models involve the location and scale pair of parameters (α, β), usually through either $F(\alpha + \beta d)$ or $F(\alpha + \beta \log d)$. The latter form is useful if a wide dose-range has been adopted, as discussed above, in facilitating comparisons, or when natural mortality is present, as we shall discuss below. Historically the normal cumulative distribution function was assumed for $F(\cdot)$, resulting in *probit analysis*. Only the largest of data sets, and Figure 1 provides a rare example of this, permit discrimination between probit analysis and logit analysis, which is based upon

$$P(d) = \frac{1}{1 + \exp[-(\alpha + \beta d)]}, \tag{4}$$

or the same form using $\log(d)$, if appropriate. Logit analysis is the simplest example of **logistic regression**. It is usually now adopted in preference to probit analysis, but an interesting exception arises when dose levels are subject to error (as when the substance is administered through food which may only be partly eaten) – see [40]. In that case the probit model is computationally easier to handle than the logit model. The opposite is usually the case, owing to the normal cumulative distribution function lacking an explicit algebraic form. Both logit and probit models are simple examples of a **generalized linear model**.

Under probit and logit models, the likelihood is a function of α and β, and **maximum likelihood** estimates follow from routine numerical optimization, as explicit maximum likelihood estimates do not exist (*see* **Optimization and Nonlinear Equations**). This is easily accomplished using computers; computational aspects will be discussed below. When the model is parameterized in terms of α and β, the maximum likelihood estimate of the ED_{50}, when $F(\cdot)$ is symmetric, is given by $-\hat{\alpha}/\hat{\beta}$, and an estimate of its standard error is easily obtained using the **delta method** – see [6]. Alternatively, the model may be parameterized in terms of $F[\beta(\theta - d)]$, thereby providing a direct estimate of the ED_{50}, θ, and of its standard error.

The model fitted to the data illustrated in Figure 1 is the logit model, with maximum likelihood estimates of parameters. In this case the poor fit to the data in the tails is improved by switching to a probit model – the main difference between the two models is in the tail behavior. A simple graphical guide to the appropriateness of a model can be obtain from plotting $F^{-1}(r_i/n_i)$ against d_i, and checking for departures from linearity. The resulting ordinates for the logit and probit models, respectively, involve logits and probits. A wide range of more complex models have been studied to improve the fit of models to data, with particular reference to tail behavior. See, for example, [3, 43, 12], and [31].

One example of such a complex model has

$$P(d) = \begin{cases} 1 - [1 + \lambda \exp(\alpha + \beta d)]^{-1/\lambda}, \\ \qquad\qquad \text{for } \lambda \exp(\alpha + \beta d) > -1, \\ 0, \qquad \text{otherwise.} \end{cases} \tag{5}$$

The additional parameter, λ, provides a function with a more flexible shape. When $\lambda = 1$, this is simply the logit model, while if $\lambda \neq 1$, then the model has an asymmetric tolerance distribution. A likelihood ratio test, or a score test (*see* **Likelihood**) could be used to examine whether, for a particular data set, it was necessary to take $\lambda \neq 1$. In the limit as $\lambda \to 0$, the model becomes the complementary log–log model.

A simple way of modeling "natural" mortality or response also involves adding a parameter to the model, resulting in Abbott's formula:

$$P(d) = \lambda + (1 - \lambda) F(\alpha + \beta \log(d)). \tag{6}$$

Here λ is the probability of natural response. Control groups, which receive no dose, are often included as standard experimental procedure. Several papers have considered how historical controls, providing control response information from previous experiments, can be included in a current analysis; see, for example, [41] and [26].

Extensions

Making Comparisons

Assume tests of two substances result in the two fitted models

$$\Pr(\text{response to dose } d_j) = \frac{1}{[1 + \exp[-(\alpha_i + \beta_i \log d_j)]}, \quad \text{for } i = 1, 2. \tag{7}$$

Frequently we might expect parallelism (*see* **Parallel-line Assay**), which corresponds to $\beta_1 = \beta_2$, and be interested in a measure to represent the difference between the two substances. In this case the ratio d_1/d_2, of equally effective doses, may be used. It is called the *relative potency*. When both models produce the same probability of response at respective doses of d_1 and d_2 say, then

$$\alpha_1 + \beta \log d_1 = \alpha_2 + \beta \log d_2, \tag{8}$$

resulting in a constant value of the relative potency.

This approach was used in [19] to demonstrate an advancement of the age of menopause by 1.23 years due to smoking. In that case the logarithmic transformation was not used.

Mixtures of Drugs

As in certain treatments of **AIDS**, "cocktails" of drugs are sometimes found to perform especially well, in excess of expectations based on the separate performance of the components of the cocktails. The drugs are then said to exhibit synergy, the opposite effect to this being termed antagonism. A large literature exists on models for responses to drugs presented in combination – see, for example, [4] and [1] (*see* **Synergy of Exposure Effects**). An attractive application for quantal responses is described in [15].

Wadley's Problem

If the response variable has a **Poisson distribution** rather than a binomial distribution, then the resulting experiment is called Wadley's problem. It provides another example of a generalized linear model.

Polytomous Responses

When there are three responses, rather than two, at the ith dose there are probabilities (P_{i1}, P_{i2}, P_{i3}), with

$$\sum_{k=1}^{3} P_{ik} = 1.$$

If n_i individuals result in respective responses (r_{i1}, r_{i2}, r_{i3}), then the data follow a trinomial distribution:

$$\Pr(r_{i1}, r_{i2}, r_{i3}|n_i) = \frac{n_i!}{r_{i1}!r_{i2}!r_{i3}!} P_{i1}^{r_{i1}} P_{i2}^{r_{i2}} P_{i3}^{r_{i3}}. \tag{9}$$

In the case of a logit model we would then write

$$\begin{aligned} P_{i1} &= \{1 + \exp[-(\alpha_1 + \beta d_i)]\}^{-1} \\ 1 - P_{i3} &= P_{i1} + P_{i2} \\ &= \{1 + \exp[-(\alpha_2 + \beta d_i)]\}^{-1}, \\ &\quad \text{with } \alpha_1 < \alpha_2. \end{aligned} \tag{10}$$

The table of data is an example of a **contingency table** with **ordered categorical** responses. Once again, maximum likelihood estimation requires numerical iteration. This is a particular instance of the general framework established for the analysis of such data by McCullagh [27]. Simpler descriptions of data in contingency tables with ordered categories are provided by ridits [10, 48] or by rankits, if the categories can be ranked [24] (*see* **Polytomous Data**).

Multivariate Responses

In contrast, multivariate responses correspond to a single individual being classified in more than one way. An interesting example of multivariate response is provided by Ashford & Sowden [5], in which working coalminers are classified according to whether they are breathless or not and whether they wheeze or not. The model proposed for these data was a simple bivariate extension of the simple threshold model.

Bayesian Analysis

In a standard probit or logit analysis, we might have prior information relating to the pair of parameters, (α, β). This may have been obtained indirectly via prior opinions of the likely response levels to particular doses. It is shown in [39] how a classical **Bayesian** analysis can be carried out, and which would, for example, result in a posterior distribution for an ED_{50}. Toxicity classes – corresponding to ED_{50}s lying in particular intervals – can then be assigned probabilities directly. Further Bayesian work can be found in [17] and, from a design perspective, in [11].

Times to Response

Quantal response experiments are frequently run for fixed times before responses are recorded, and

popular times are one week or other multiples of days. Sometimes ED_{50} and similar dose levels are qualified by the duration of the experiment. In cases of prolonged exposure to the substance being tested, we might expect a tradeoff between dose and time, similar effects resulting from either a low dose for a long time or a high dose for a short time. In other experiments there may be a limit to the response at each dose, which would be approached as the duration of the experiment was increased. The former case is expressed in one form by Ostwald's equation [34], in which $XT^{\lambda} = \kappa_p$, for the constants λ and κ_p. Here X denotes concentration and T duration, corresponding to a level $p\%$.

Stochastic models which include a description of times to response are drawn from **survival analysis**. If $F(t;d)$ denotes the probability of response by time t to dose d, then following [36], the **Weibull** model,

$$F(t;d) = 1 - \exp[-t^{\gamma} \exp(\alpha + \beta \log d)], \quad (11)$$

can be shown to agree with Ostwald's equation. An alternative approach, which was used by Pack and Morgan [35] to describe data exhibiting limits to response at each dose, extended the standard threshold model: at each dose level there was a probability of response which was dose-related, as described by a logistic model. Responding individuals were then given a log logistic distribution, which was dose-independent.

A characteristic of time-to-response data is that it is usually interval-censored, typically with observations taken only once a day. For general discussion of such models, see [14].

Overdispersion

Data frequently exhibit more variation than that allowed for in simple binomial and Poisson models (*see* **Overdispersion**) of response. This can be due to factors such as a lack of independence of response, or heterogeneity of subjects. In a modeling context, the Poisson distribution might be replaced by the negative-binomial distribution, and the binomial by a beta-binomial distribution, though several alternatives also exist. In some cases it may be appropriate to fit mixtures of distributions, for which there may, for example, be a genetic justification – see [9] and [7]. A robust procedure results from adopting a quasi-likelihood approach, which makes use of just the means and variances of the responses – see, for example, [47] and [8].

The beta-binomial distribution has the probability function

$$\Pr(X = x|n) = \frac{\binom{n}{x} \prod_{r=0}^{x-1}(\mu + r\theta) \prod_{r=0}^{n-x-1}(1 - \mu + r\theta)}{\prod_{r=0}^{n-1}(1 + r\theta)}, \quad \text{for } 0 \le x \le n \quad (12)$$

(we interpret $\prod_{r=0}^{-1}$ as unity). We can see that $\theta = 0$ returns us to the binomial form. We have

$$\mathrm{E}[X] = n\mu,$$
$$\mathrm{var}(X) = n\mu(1-\mu)\left\{1 + \left(\frac{\theta}{1+\theta}\right)(n-1)\right\}. \quad (13)$$

Thus positive θ produces a larger variance than in the binomial case. Negative values of θ may also be used, allowing a straightforward likelihood ratio, or score, test of $\theta = 0$ [38, 44].

In the dose response context we can allow both μ and θ to be suitable functions of dose, and then use standard testing procedures to investigate whether such complexity is required – see, for example, [42, 13], and [30]. Mixed models, such as the logistic–normal and probit–normal, provide an alternative approach to dealing with overdispersion – see, for example, [20] and [37].

Design and Sequential Methods

Suppose our objective is to innoculate cattle against a disease. We might well wish to use a dose at about the PD_{90}, say, this being the dose that would protect 90% of animals. Higher doses might be thought to be too expensive to produce, or to result in undesired side effects. It may only be possible to use a small number of animals to estimate the PD_{90}, and the question then arises of how to allocate dose levels to the animals. Adopting an optimal design approach, we might, for example, focus on the PD_{90}, and try to estimate this as precisely as possible. We saw earlier an expression for the variance of the Spearman–Kärber estimate of

the ED_{50}. Prior estimates of parameters may be used to estimate numbers responding at any dose level, resulting in an expression to be optimized with regard to the allocation of a fixed number of individuals over doses. This is the basic idea of optimal design procedures. It relies on being able to produce prior estimates of parameters. The prescription can be very clear and simple – for example to distribute individuals over just two doses. However, poor performance can result from poor initial estimates of the parameters, and a more conservative approach is advisable in practice – see, for example, [2] and [21]. A two-stage design is studied in [2]; the first stage improved on prior estimates of parameters and informed the choice of doses at the second stage.

A formal Bayesian approach is adopted in [11].

Fully sequential methods involve allocating individuals to doses in a way that depends on previous responses. For example, in an up-and-down experiment, if a rat responds to a dose then the next rat is treated with a lower dose. Subsequent dose levels would then be reduced until a rat fails to respond, and the next dose level to be chosen would then increase, and so on. A form of **up-and-down** experiment results in the fixed dose procedure for evaluating toxicity – see [46]. In its simplest form, the up-and-down experiment produces dose levels according to the following sequence:

$$d_{i+1} = d_i - 2\Delta(r_i - 0.5),$$

where $r_i = 1$ if the individual exposed at dose d_i responds, and $r_i = 0$ otherwise. Here the doses are selected at intervals of Δ. For discussion of variants of this procedure, and how to analyze the resulting data, see [23] and [45]. The Robbins–Monro procedure of **stochastic approximation** for estimating an ED_{100p} value allows the step-length to decrease with increasing i, to produce the following sequence of doses:

$$d_{i+1} = d_i - \frac{c}{i}(r_i - p),$$

for a suitable constant c.

```
global x r n s
x = [9.21 10.21 10.58 10.83 11.08 11.33 11.58 11.83 12.08 12.33 12.58 ...
12.83 13.08 13.33 13.58 13.83 14.08 14.33 14.58 14.83 15.08 15.33 ...
15.58 15.83 17.58];
r = [0 0 0 2 2 5 10 17 16 29 39 51 47 67 81 88 79 90 113 95 117 107 92 ...
112 1049];
n = [376 200 93 120 90 88 105 111 100 93 100 108 99 106 105 117 98 97 120 ...
102 122 111 94 114 1049];
s = n-r;
x1 = fmins ('menarche', [0,1]);
plot (x,r ./n,'o'); hold; plot (x,1 ./(1+exp(-x1(1) - x1(2)*x)));
```

(a)

```
function l=menarche (paras)
global x r n s;
lin=paras(1)+paras(2) *x;
pinv=1+exp(-lin); cpinv=1+exp(lin);
l=r*(log(pinv))'+s*(log(cpinv))';
```

(b)

Figure 2 An example of computer code for fitting the model displayed in Figure 1, and producing the figure. The MATLAB language is used. The program of (a) reads in the data, minimizes the negative log likelihood using the *fmins* command, which uses a simplex method – see [33] – and plots the data and the results. The program of (b) establishes minus the log likelihood

Many variants have been devised – see [25] and [49]. Sequential optimization, in which optimal design ideas are used in a sequential context, has been considered in [28] and [21]. Sequential methods, even if only applied for a limited number of stages, have an obvious appeal, in limiting the use of experimental resources and individuals (*see* **Sequential Analysis**). In practice a nonsequential approach may prove to be more practicable. For example, in routine assays the experimenter will want to set up and conclude one experiment according to a laboratory schedule before starting another. In some cases responses may be slow to obtain, thereby ruling out a sequential approach. While it is simple to devise a system of new doses on paper, in practice it is usually much simpler to construct the appropriate dilutions on just one occasion, and then to select from these in future experiments.

Computing

There are several computer packages which may be used to fit models to quantal assay data (*see* **Software, Biostatistical**). A range of facilities exist in packages such as GLIM, GENSTAT, SPSS-X, SAS and S-Plus. In SAS, for example, PROC GENMOD can be used for fitting generalized linear models, and PROC LOGISTIC provides specialized software for logistic regression. The availability of powerful integrated computer languages means that for many statisticians it is a simple matter to program their own analyses. For instance, the fit and display of Figure 1 are produced by the MATLAB commands of Figure 2.

References

[1] Abdelbasit, K.M. & Plackett, R.L. (1982). Experimental design for joint action, *Biometrics* **38**, 171–179.

[2] Abdelbasit, K.M. & Plackett, R.L. (1983). Experimental design for binary data, *Journal of the American Statistical Association* **78**, 90–98.

[3] Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data, *Biometrika* **68**, 357–364. (Correction: *Biometrika* **70**, 303.)

[4] Ashford, J.R. (1981). General models for the joint action of mixtures of drugs, *Biometrics* **37**, 457–474.

[5] Ashford, J.R. & Sowden, R.R. (1970). Multi-variate probit analysis, *Biometrics* **26**, 535–546.

[6] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[7] Böhning, D., Schlattmann, P. & Lindsay, B. (1992). Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms, *Biometrics* **48**, 283–304.

[8] Breslow, N.E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics* **33**, 38–44.

[9] Brooks, S.P., Morgan, B.J.T., Ridout, M.S. & Pack, S.E. (1996). Finite mixture models for proportions, *Biometrics* (unpublished paper).

[10] Bross, I.D.J. (1958). How to use ridit analysis, *Biometrics* **14**, 18–38.

[11] Chaloner, K. & Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments, *Journal of Statistical Planning and Inference* **21**, 191–208.

[12] Copenhaver, T.W. & Mielke, P.W. (1977). Quantit analysis: a quantal assay refinement, *Biometrics* **33**, 175–187.

[13] Crowder, M.J. (1978). Beta-binomial ANOVA for proportions, *Applied Statistics* **27**, 34–37.

[14] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data, *Biometrics* **42**, 845–854.

[15] Giltinan, D.M., Capizzi, T.P. & Malani, H. (1988). Diagnostic tests for similar action of two compounds, *Applied Statistics* **37**, 39–50.

[16] Glasbey, C.A. (1987). Tolerance-distribution-free analyses of quantal dose-response data, *Applied Statistics* **36**, 251–259.

[17] Govindarajulu, Z. (1988). *Statistical Techniques in Bioassay*. Karger, New York.

[18] Hamilton, M.A. (1980). Inference about the ED_{50} using the trimmed Spearman–Kärber procedure – a Monte Carlo investigation, *Communications in Statistics – Simulation and Computation* **9**, 235–254.

[19] Healy, M.J.R. (1988). *GLIM: An Introduction*. Clarendon Press, Oxford.

[20] Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present, *Applied Statistics* **39**, 75–84.

[21] Kalish, L.A. (1990). Efficient design for estimation of median lethal dose and quantal dose-response curves, *Biometrics* **46**, 737–748.

[22] Kappenman, R.F. (1987). Nonparametric estimation of dose-response curves with application to ED_{50} estimation, *Journal of Statistical Computation and Simulation* **28**, 1–13.

[23] Kershaw, C.D. (1987). A comparison of estimators of the ED_{50} in up-and-down experiments, *Journal of Statistical Computation and Simulation* **27**, 175–184.

[24] Krewski, D. (1976). Distribution-free confidence intervals for quantile intervals, *Journal of the American Statistical Association* **71**, 420–422.

[25] Lai, T.L. & Robbins, H. (1979). Adaptive design and stochastic approximation, *Annals of Statistics* **7**, 1196–1221.

[26] Leroux, B.G., Fung, K.Y., Krewski, D. & Prentice, R.L. (1994). The use of historical control data in testing for trend in counts, *Statistica Sinica* **4**, 581–601.

[27] McCullagh, P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society* **42**, 109–142.

[28] McLeish, D.L. & Tosh, D.H. (1990). Sequential designs in bioassay, *Biometrics* **46**, 103–116.

[29] Milicier, M. & Szczotka, F. (1966). Age at menarche in Warsaw girls in 1965, *Human Biology* **38**, 199–203.

[30] Moore, D.F. (1987). Modelling the extraneous variance in the presence of extra-binomial variation, *Applied Statistics* **36**, 8–14.

[31] Morgan, B.J.T. (1988). Extended models for quantal response data, *Statistica Neerlandica* **42**, 253–272.

[32] Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. Chapman & Hall, London.

[33] Nelder, J.A. & Mead, R. (1965). A simplex method for function minimisation, *Computer Journal* **7**, 308–312.

[34] Ostwald, W. & Dernoschek, A. (1910). Über die Beziehungen zwischen Adsorption und Giftigkeit, *Kolloid-Zeitschrift* **6**, 297–307.

[35] Pack, S.E. & Morgan, B.J.T. (1988). A mixture model for interval-censored time-to-response quantal assay data, *Biometrics* **46**, 749–758.

[36] Petkau, A.J. & Sitter, R.R. (1989). Models for quantal response experiments over time, *Biometrics* **45**, 1299–1306.

[37] Preisler, H.K. (1989). Analysis of a toxicological experiment using a generalized linear model with nested random effects, *International Statistics Review* **57**, 145–159.

[38] Prentice, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, *Journal of the American Statistical Association* **81**, 321–327.

[39] Racine, A., Grieve, A.P., Fluhler, H. & Smith, A.F.M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry, *Applied Statistics* **35**, 93–150.

[40] Ridout, M.S. & Fenlon, J.S. (1991). Analysing dose-mortality data when doses are subject to error, *Annals of Applied Biology* **119**, 191–201.

[41] Ryan, L.M. (1993). Using historical controls in the analysis of developmental toxicity data, *Biometrics* **49**, 1126–1135.

[42] Segreti, A.C. & Munson, A.E. (1981). Estimation of the median lethal dose when responses within a litter are correlated, *Biometrics* **37**, 153–154.

[43] Stukel, T.A. (1990). A general model for estimating ED_{100p} for binary response dose-response data, *American Statistician* **44**, 19–22.

[44] Tarone, R.E. (1979). Testing the goodness of fit of the binomial distribution, *Biometrika* **66**, 585–590.

[45] Wetherill, G.B. & Glazebrook, K.D. (1986). *Sequential Methods in Statistics*, 3rd Ed. Chapman & Hall, London.

[46] Whitehead, A. & Curnow, R.N. (1992). Statistical evaluation of the fixed-dose procedure, *Food and Chemical Toxicology* **30**, 313–324.

[47] Williams, D.A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics* **31**, 144–148.

[48] Williams, O.D. & Grizzle, J.E. (1972). Analysis of contingency tables having ordered response categories, *Journal of the American Statistical Association* **67**, 55–63.

[49] Wu, C.F.J. (1985). Efficient sequential designs with binary data, *Journal of the American Statistical Association* **80**, 974–984.

(*See also* **Biological Assay, Overview**)

BYRON J.T. MORGAN

Quantile Regression

Consider the situation illustrated in Figure 1 which shows various **quantiles** of infant boys' weight plotted against age. Interest may be focused not only on the mean relationship between these two variables, but also on the relationship between the extreme quantiles and age. The quantiles may act as appropriate boundaries for indicating potential infant feeding problems, and therefore are useful when examining the progress of a child's weight gain over time. Clearly, in this and many similar situations, a standard regression model is not sufficient for predicting the relationship between all the quantiles of a dependent variable y and a vector $\mathbf{x}$ of **explanatory variables**.

With this fact in mind, Hogg [2] and Koenker & Bassett [4] generalized the method of minimum mean absolute deviation (MAD) regression (*see* **Robust Regression**) to quantile regression in which, under the assumption of linearity, they developed a technique for modeling the quantiles of the conditional distribution function $F(y|\mathbf{x})$.

To describe the method, let us examine how to specify ordinary quantiles as the solution to certain estimating equations (*see* **Generalized Estimating Equations**), which minimize a particular sample loss function. This will enable us to develop and estimate regression models for conditional quantiles. Given a sample of n univariate observations $\{y_1, y_2, \ldots, y_n\}$, the pth sample quantile $\hat{\mu}_p, 0 \leq p \leq 1$, satisfies the relationship

$$\sum_{i=1}^{n} [pI(y_i > \hat{\mu}_p) - (1-p)I(y_i < \hat{\mu}_p)] = 0, \quad (1)$$

where I is the indicator function taking the value 1 when its argument is true, and 0 otherwise. Note that $\hat{\mu}_p$ is an estimate of the corresponding population quantile μ_p, which satisfies $p[1 - F(\mu_p)] - (1-p)F(\mu_p) = 0$. Eq. (1) can be viewed as an estimating equation where positive **residuals** $r_i = y_i - \hat{\mu}_p$ are given weight p and negative residuals weight $(1-p)$, and value is measured by the sign of the residual. That is, (1) can be rewritten in the form $\sum_{i=1}^{n} \psi_p(r_i) = 0$, where

$$\psi_p(r) = \begin{cases} p\psi(r), & \text{if } r > 0, \\ (1-p)\psi(r), & \text{otherwise,} \end{cases} \quad (2)$$

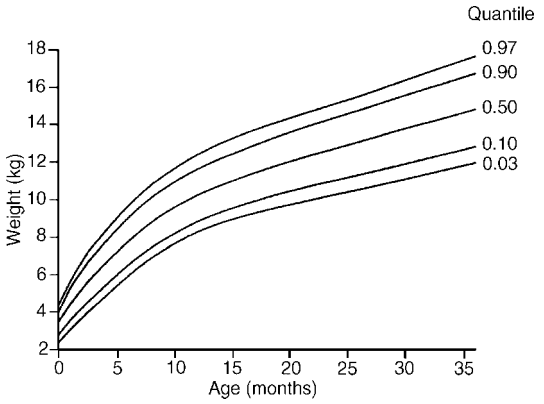

Figure 1 Quantiles of infant boys' weight against age. In this example the extreme quantile lines may be useful for indicating potential infant feeding problems

and $\psi(r) = \text{sign}(r)$. Sometimes ψ_p is referred to as the influence function (*see* **Diagnostics**) for estimating regression quantiles with corresponding asymmetrical **loss function**

$$\rho_p(r) = \begin{cases} p\rho(r), & \text{if } r > 0, \\ (1-p)\rho(r), & \text{otherwise,} \end{cases} \quad (3)$$

where $\rho(r) = |r|$. The estimating equation above, and hence $\hat{\mu}_p$, can be obtained by minimizing the corresponding mean sample loss function

$$n^{-1} \sum_{i=1}^{n} \rho_p(y_i - \mu_p). \quad (4)$$

To extend the method above to quantile regression, we proceed in a manner analogous to linear least squares regression and replace μ_p in (4) by $\mu_p(\mathbf{x}) = \mathbf{x}'\beta_p$, where β_p is a vector of regression coefficients for the pth conditional quantile. It is then fairly straightforward to see that β_p can be estimated by solving

$$\sum_{i=1}^{n} \psi_p(y_i - \mathbf{x}'_i\hat{\beta}_p)\mathbf{x}_i = \mathbf{0}. \quad (5)$$

In the special case $p = 0.5$, the fitted surface is the regression median of y on $\mathbf{x}$ that minimizes

$$n^{-1} \sum_{i=1}^{n} \left| y_i - \mathbf{x}'_i\hat{\beta}_{0.5} \right|, \quad (6)$$

the mean absolute deviation of residuals.

Expectile and M-Quantile Regression

Despite their apparent simplicity, regression quantiles suffer two distinct disadvantages in practice. First, they are usually not unique and hence are quite difficult to compute. Fairly sophisticated linear programming techniques are often used (see, for example, [8] and [6]). Secondly, and perhaps more importantly, they do not include ordinary **least squares** regression as a special case, which is generally the preferred technique for modeling average behavior.

In view of these difficulties, Newey & Powell [9] developed *expectile regression*. Expectiles are obtained by modifying the ordinary least squares criteria in the same way as (4) modifies the MAD criteria of (6). That is, they are obtained by solving (5) with $\psi(r) = r$. Clearly, expectile regression reduces to ordinary least squares regression when $p = 0.5$, and for any $0 < p < 1$, they are unique and can easily be computed using the method of iteratively reweighted least squares (*see* **Generalized Linear Model**).

Figure 2 shows the influence function ψ_p for expectiles when $p = 0.75$. One important feature worth noting is that its influence function is unbounded, which implies that estimates of expectile regression parameters may be quite sensitive to outliers. With this difficulty in mind, Breckling & Chambers [1] suggested that the influence function be modified by using the M-regression ψ-function,

$$\psi(r) = \psi(r, c) = \begin{cases} r, & \text{if } |r| < c, \\ c \ \text{sign}(r), & \text{otherwise,} \end{cases} \qquad (7)$$

in the definition of (2), where $c > 0$ (*see* **Robust Regression**). The resulting method is called M-quantile regression. As $c \to 0$ it approaches quantile regression, and as $c \to \infty$ it approaches expectile regression. In this regard it can be viewed as a compromise between the two, sharing the robustness properties of quantile regression and ease of computation of regression expectiles. In particular, when $p = 0.5$, M-quantile regression reduces to ordinary M-regression, a common choice for robust estimation of the conditional mean when the residual distribution is symmetrical around zero; see Huber [3].

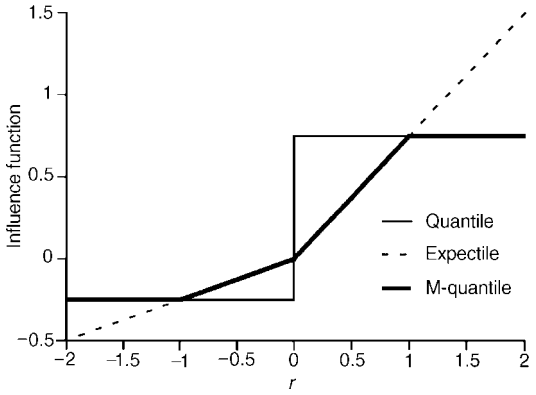

Figure 2 The influence curves ψ_p for quantiles, expectiles and M-quantiles ($c = 1$) in the case $p = 0.75$. The expectile influence curve is unbounded and hence may be sensitive to outliers in the data. The M-quantile influence curve is a compromise between the quantile and expectile influence curves

The Purpose of Modeling Quantiles

As illustrated in our introductory example, the relationship that holds between $\mathbf{x}$ and the conditional mean of y given $\mathbf{x}$ may not be appropriate or representative of those values of y not lying close to the conditional mean. One may well ask what circumstances lead to this situation.

It is often the case when modeling nonexperimental data that not all the explanatory information for the response variable y can be controlled, and only an incomplete set of covariates is available for the modeling process. There may be additional variables, $\mathbf{z}$, that also explain the variation in y, but which have not been measured. If there is an interaction between $\mathbf{z}$ and $\mathbf{x}$, then it is likely that the relationship between $\mathbf{x}$ and extreme y values differs from the mean relationship; see Figure 3. When it becomes necessary to make predictions based only on changes in the $\mathbf{x}$ values, then better forecasts could potentially be made by directly modeling the relationship between the extreme as well as mean values of y conditional upon $\mathbf{x}$. For example, it may be appropriate to use the actual quantile or M-quantile regression surface that passes through each individual observation when making forecasts. This is the basis of the prediction technique developed in [7].

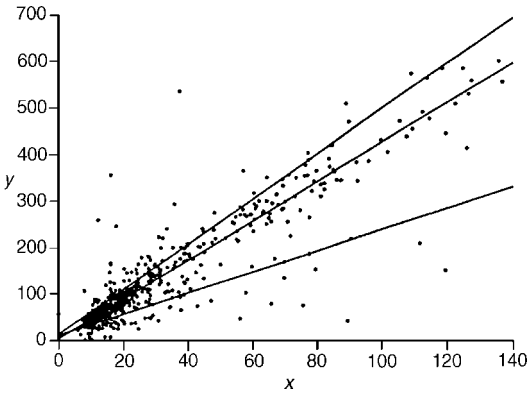

Figure 3 Regression quantiles for $p = 0.10$ (lower line), $p = 0.5$ (middle line) and $p = 0.90$ (upper line) of an indicative data set consisting of a single covariate **x** and dependent variable y. Note that the slope of the three regression lines differs significantly, indicating a varying relationship between y and **x** in the data

The varying relationship illustrated in Figure 3 of the regression quantiles with p can also be used to test for heteroscedasticity [5].

Acknowledgments

The author wishes to thank Ray Chambers and Vern Farewell for their constructive comments on this article.

References

[1] Breckling, J. & Chambers, R. (1988). M-quantiles, *Biometrika* **75**, 761–771.

[2] Hogg, R.V. (1975). Estimates of percentile regression lines using salary data, *Journal of the American Statistical Association* **70**, 56–59.

[3] Huber, P.J. (1981),. *Robust Statistics*. Wiley, New York.

[4] Koenker, R. & Bassett, G. (1978). Regression quantiles, *Econometrica* **46**, 33–50.

[5] Koenker, R. & Bassett, G. (1982). Robust tests for heteroscedasticity based on regression quantiles, *Econometrica* **50**, 43–62.

[6] Koenker, R.W. & D'Orey, V. (1987). Computing regression quantiles, *Applied Statistics* **36**, 383–393.

[7] Kokic, P.N., Beare, S., Topp, V. & Tulpule, V. (1993). Australian broadacre agriculture: forecasting supply at the farm level, *ABARE Research Report 93.7*. ABARE, Canberra, Commonwealth of Australia.

[8] Narula, S.C. & Wellington, J.F. (1990). An algorithm to find all regression quantiles using bicriteria optimization, *American Journal of Mathematical and Management Sciences* **10**, 229–259.

[9] Newey, W.K. & Powell, J.L. (1987). Asymmetric least squares estimation and testing, *Econometrica* **55**, 816–847.

P. KOKIC

Quantiles

Quantiles divide a statistical distribution (population) or a sample of data into *equal* and *ordered* parts. *Equal* means that all parts contain the same number of sample elements, or equal population mass. *Ordered* means that the parts are arranged so that all sample elements within a part are less than those in the part following it and greater than those in the part preceding it. The term has no numerical attributes until further specification is provided for its use, usually the number of equal parts. Division into N ordered parts requires $N-1$ points or quantiles. For certain N, the set of these $N-1$ points has a well-known name related to N: **median** ($N=2$), *tertiles* ($N=3$), *quartiles* ($N=4$), *quintiles* ($N=5$), *deciles* ($N=10$), and *percentiles* ($N=100$). Because, for example, one-fifth of the population or sample data is less than the lowest quintile, a more general naming of that quintile is the 0.2 quantile. Equivalently, this is the 20th percentile. Generalizations to expressions such as the 0.2143 quantile are possible but not very useful.

There is a limited number of quantiles defined for a discrete distribution. With sample data, various conventions are useful for approximating quantiles of interest. From a sample of size M odd, the median or 50th percentile is the $(M+1)/2$th ordered value; for even M, we use the average of the $M/2$th and $(M/2+1)$th ordered values. Similar ideas may be used for other quantiles, especially for summarization and visualization of data.

Quantiles are useful in a variety of biostatistical settings. In testing of statistical hypotheses and related estimation procedures, if we choose a level of significance α (*see* **Level of a Test**) or confidence $(1-\alpha)$, we must determine the $1-\alpha$ quantile, the point in the appropriate distribution exceeded by $100\alpha\%$.

Various sets of quantiles are useful in exploration and summarization of sample data (*see* **Exploratory Data Analysis**), especially visually through box plots (*see* **Graphical Displays**), to assess symmetry, dispersion, and location. For a biological measure for which extreme values are detrimental, such as blood pressure or cholesterol, the safety and efficacy of a new treatment may be assessed and compared to others with respect to specified upper quantiles. We may decide that an improved treatment is one which yields a 98th percentile no greater than the standard and also a 75th percentile (upper quartile) markedly less than the standard.

Quantiles are also useful when modeling risk relationships between predictor (or **explanatory**) and outcome (or **response**) variables, as in **logistic regression** or **generalized linear models**. For example, in studies of the effect of improved nutrition on reducing the incidence of disease, we often find that measures of nutritional status seem unrelated to outcome until the status reaches high levels or exceeds some threshold. A common approach to model this relationship, which requires few assumptions, is based on quantiles. The distribution of nutritional status is frequently **skewed**, and we may choose to transform it by the sample quintiles into a categorical variable. The four quintiles $\{Q1, Q2, Q3, Q4\}$ divide the sample into five ordered quintile groups $\{G1, G2, G3, G4, G5\}$. An observed value X of nutritional status receives the designation (transformed value) of the quintile group of which it is a member. Thus, if $Q2 < X < Q3$, then X has transformed value or category 3, or $G3$ if you wish. The lowest quintile group is typically selected as the reference group against which higher groups are compared, and we may find that only the upper category is significantly different from the reference category in relation to risk of disease.

Strictly speaking, one should not say "in the upper quintile", because quantiles are not groups but, rather, points that demarcate groups. The second quartile is also the median, and "in the median" highlights the inconsistency. The term *quantile group* seems appropriate and not overly onerous. Nevertheless, the simpler but imprecise usage such as "in the upper quintile" is commonly found in prestigious medical, epidemiology, and even biostatistics journals as *de facto* standard.

(*See also* **Order Statistics**)

ROY C. MILTON

Quartimax Rotation

Quartimax rotation [3, 4] is one of the earlier **orthogonal rotation** procedures for use with **principal components analysis** and **factor analysis**. Given a matrix $\mathbf{V}$ of dimension $p \times k$ consisting of a set of k vectors defining a set of principal components or factors, a new set of transformed variables is obtained by an orthogonal rotation of $\mathbf{V}$, namely $\mathbf{B} = \mathbf{V}\boldsymbol{\Theta}$. Here $\boldsymbol{\Theta}$ is a matrix of dimension $k \times k$, determined such that the coefficients of the resulting matrix $\mathbf{B}$ of dimension $p \times k$, containing the new vectors defining the transformed variables, will maximize the quantity

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{p} b_{ij}^4,$$

where p is the number of original variables and k is the number of retained components or factors. Quartimax rotation is a special case of *orthomax* rotation with $c = 0$ (*see* **Orthogonal Rotation**). In this procedure, the sums of squares of $\mathbf{B}$ are maximized *rowwise* as contrasted to **varimax rotation**, which maximizes them *columnwise*. There is a tendency for quartimax to produce a "general" rotated vector which has no small coefficients. Because of these properties, quartimax has generally been replaced by varimax rotation. The standard errors for quartimax loadings were given by Archer & Jennrich [1] and their asymptotic distribution also by Archer & Jennrich [2].

For the audiometric example introduced in the article on **Rotation of Axes**, $\mathbf{V}$ and $\mathbf{B}$ are given in Table 1. The results for the varimax rotation are included also, for comparison.

This is an example where the popular rotation procedures such as varimax did not produce useful results because the units of the original variables, Hz, represent points on a continuum rather than a set of qualitative variables which generally rotate much more favorably. Quartimax, on the other hand, summarizes the situation quite well. The first three rotated vectors define three groups of frequencies. The first rotated vector clusters both 500 Hz and 1000 Hz measurements, the second vector is associated with 4000 Hz, and the third vector with 2000 Hz. The fourth vector restates the left–right ear difference, as did the fourth characteristic vector.

References

[1] Archer, C.O. & Jennrich, R.I. (1973). Standard errors for rotated factor loadings, *Psychometrika* **38**, 581–592.

[2] Archer, C.O. & Jennrich, R.I. (1976). A look, by simulation, at the validity of some asymptotic distribution results for rotated loadings, *Psychometrika* **41**, 537–541.

[3] Ferguson, G.A. (1954). The concept of parsimony in factor analysis, *Psychometrika* **19**, 281–290.

[4] Neuhaus, J.O. & Wrigley, C. (1954). The quartimax method: an analytical approach to orthogonal simple structure, *British Journal of Statistical Psychology* **7**, 81–91.

(*See also* **Axes in Multivariate Analysis**)

J. EDWARD JACKSON

Table 1 Audiometric example: characteristic and rotated vectors

	Characteristic vectors				Varimax rotation				Quartimax rotation			
Frequency (Hz)	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
500 left	0.80	−0.40	0.16	−0.22	0.58	0.13	0.06	0.71	0.90	0.11	0.03	0.22
1000 left	0.83	−0.29	−0.05	−0.33	0.44	0.10	0.27	0.79	0.83	0.09	0.25	0.36
2000 left	0.73	0.30	−0.46	−0.19	0.05	0.22	0.78	0.45	0.35	0.22	0.79	0.28
4000 left	0.56	0.60	0.42	−0.11	0.04	0.89	0.15	0.20	0.18	0.89	0.15	0.11
500 right	0.68	−0.49	0.26	0.33	0.91	0.08	−0.00	0.23	0.87	0.05	−0.07	−0.35
1000 right	0.82	−0.29	−0.03	0.25	0.77	0.10	0.34	0.31	0.82	0.08	0.28	−0.23
2000 right	0.62	0.40	−0.56	0.27	0.17	0.19	0.93	−0.00	0.19	0.19	0.91	−0.17
4000 right	0.50	0.65	0.42	0.11	0.11	0.91	0.19	−0.02	0.11	0.90	0.17	−0.10

Quasi-experimental Design

Quasi-experiments have been defined as "experiments that have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons from which treatment-caused change is inferred" [6]. (*True experiments*, in contrast, are intervention studies that employ **randomization**.) The term *quasi-experiment* was first introduced in 1963 in Campbell & Stanley's classic *Experimental and Quasi-Experimental Designs for Research* [3]. These designs are often employed when random assignment to treatment groups (*see* **Randomized Treatment Assignment**) is not possible.

In principle, every randomized-trial design has a quasi-experimental counterpart that simply substitutes some other method of treatment allocation for random assignment. Thus, there are potentially a limitless number of quasi-experimental study designs, ranging from simple comparison of two parallel groups (*see* **Clinical Trials, Overview**) to **factorial** designs, repeated measures designs (*see* **Longitudinal Data Analysis, Overview**), **crossover designs, group-randomization designs,** and so forth. In addition, the term "quasi-experiment" has historically been considered to include certain single-group designs that have no randomized-trial counterpart, such as the interrupted **time series** design described below.

A quasi-experiment is designed around an intervention. Typically the main study goals are to estimate the size of the intervention effect on some outcome and to test whether it differs significantly from no effect. (Quasi-experiments can certainly consider multiple interventions and **multiple endpoints** simultaneously, but the theory is broadly similar, and for simplicity we shall consider only studies of a single intervention and a single main outcome.) At some point during the study, subjects are exposed to the intervention of interest, and outcomes in those subjects are observed. The crucial question in estimating the effect of the intervention is: What would have been observed on those subjects had they not been exposed to the intervention? Quasi-experimental designs differ chiefly with regard to how they seek to answer this question.

After discussing **confounding** as the central methodological issue in most quasi-experimental designs and describing a useful notation for study designs, the discussion below focuses on three quasi-experimental designs, each of which raises generic design and analysis issues that arise in a larger class of quasi-experimental studies. Cook & Campbell [6] and Campbell & Stanley [3] discuss several other specific quasi-experimental designs and comment on their strengths and weaknesses.

Confounding

Expectations about what would have happened in the experimental group in the absence of intervention are often based on outcomes observed in a non-exposed **control** group. Because quasi-experiments do not employ randomization to form comparison groups, the experimental and control groups can, and often do, differ with regard to other measured or unmeasured factors that influence the outcome. Hence observed differences in outcomes between the experimental and control groups can represent a mixture of effects of the intervention under study and the effects of these pre-existing differences between groups. Prevention or removal of **confounding** of this sort thus becomes an important methodological issue in **unbiased estimation** of intervention effects from quasi-experimental evidence [9].

Several techniques are available to overcome confounding. In the study-design stage, *restriction* or **matching** can be used. For example, confounding by gender can be prevented by restricting the study population to males alone or to females alone. Matching on one or more potential confounding factors can also be employed. However, restriction and matching are not always feasible or sufficient, particularly if the number of potential confounding factors is large. Accordingly, confounding must routinely be considered in the analysis stage of a quasi-experiment, usually by including potential confounding factors as **covariates** in an **analysis of covariance** (ANCOVA), **multiple regression** or other type of **multivariate analysis**.

Except for randomization, all of the methods for control of confounding require that the investigator know what the important confounding factors are and how to measure them. Omission of an important confounding factor from an analytic model results

in what is sometimes called a specification error or model **misspecification**, and the result can be substantial **bias** in estimation of the intervention effect.

Campbell–Stanley Notation

Campbell & Stanley [3] developed a useful notation for study designs that conveys several important features and helps distinguish one design from another at a glance. The notation applies to both randomized and nonrandomized study designs. The following symbols are used for every design:

X = exposure to an intervention;
O = observation or measurement.

Additional symbols are sometimes used to indicate the method by which subjects are allocated to treatment groups. These include:

R = random assignment;
C = assignment based on whether a subject's value on a certain score falls above or below a specified cutoff value.

These symbols are arranged in rows, each row representing a different group of study subjects. Within a row, the ordering of symbols from left to right indicates the temporal sequence in which steps are carried out. Symbols that are aligned vertically indicate performance at the same point in time. A simple example is:

$$\begin{array}{cccc} \mathrm{R} & \mathrm{O} & \mathrm{X} & \mathrm{O} \\ \mathrm{R} & \mathrm{O} & & \mathrm{O} \end{array} \qquad (1)$$

This design involves randomization (R) of subjects into two study groups, each of which is observed (O) after randomization. One group is then exposed to an intervention (X) while the other group is not. Lastly, both groups are observed (O) concurrently at follow-up.

When neither R nor C appears, then study groups are assumed to be formed on some other basis. For example, the quasi-experimental counterpart to design (1) above would be:

$$\begin{array}{ccc} \mathrm{O} & \mathrm{X} & \mathrm{O} \\ \mathrm{O} & & \mathrm{O} \end{array} \qquad (2)$$

Symbols may be subscripted as necessary to avoid ambiguity if there are multiple interventions, observation occasions, or allocation steps.

Pretest, Posttest Nonequivalent Control Group Design

In spite of its cumbersome name, design (2) above is among the most commonly used quasi-experimental designs, involving baseline ("pretest") and follow-up ("posttest") measurements on subjects in each of two groups, only one of which is exposed to the intervention. For example, Simon et al. [14] studied the effect of a $20/visit copayment for mental-health visits among government workers and their dependents who were enrolled in a large health-maintenance organization. The new copayment applied to families of state government employees, but not to families of federal government employees, who served as a no-intervention control group. Data were gathered on each enrollee's use of any mental-health services and annual visit rate during the year before the change and during the year after the change.

More generally, results from this type of quasi-experiment can be portrayed graphically as in Figure 1. Y is the outcome variable, and a value of Y is obtained for each subject in each of two treatment groups at each of two time points. (Unique identification of a particular Y-value would require three subscripts, but for simplicity we will omit the subscript that indexes individual subjects.) Baseline measures of Y are designated as Y_0 and those obtained at follow-up as Y_1. As shown in Figure 1, a

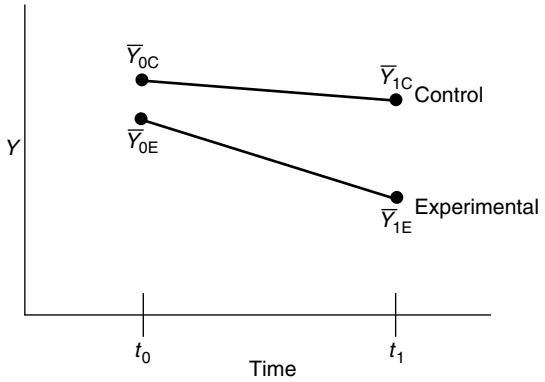

Figure 1 Diagram of results for the nonequivalent control group design

second subscript (E for experimental, C for control) is added when needed to distinguish between study groups. **Mean** values of Y are always taken across subjects within a treatment group and time point.

Despite the apparent simplicity of this design, approaches to data analysis have been varied and sometimes controversial. The parameter of main interest, τ, is the effect of the intervention. Conceptually, it can be defined as:

$$\tau = \overline{Y}_{1E} - \overline{Y}^*_{1E}, \tag{3}$$

where $\overline{Y}^*_{1E}$ is the mean value of Y that would have been observed in the experimental group at follow-up had the group not been exposed to the intervention. The alternative analytic approaches differ chiefly with regard to how they estimate $\overline{Y}^*_{1E}$. We compare three approaches here.

Separate Comparison of Baseline Means and Follow-Up Means

Under this approach, the experimental and control groups are first compared with regard to Y_0 and possibly other baseline characteristics. If no significant differences are found, the groups are treated as "essentially equivalent" (as if randomization had been performed), and τ is estimated from follow-up data alone:

$$\hat{\tau} = \overline{Y}_{1E} - \overline{Y}_{1C}. \tag{4}$$

In effect, this approach assumes that if the groups did not differ significantly at baseline, they would have been equivalent at follow-up in the absence of an intervention effect.

This approach has several shortcomings. Since the study groups were *not* formed by random assignment, they may well be found to differ significantly with regard to Y_0 or other potentially confounding variables, obligating the analyst to account for these differences in estimating τ. Even if the baseline differences are not statistically significant, the **power** to detect a true difference may be low, especially if sample sizes are small and/or if variation within groups is large. (Failing to reject the **null hypothesis** of equivalence at baseline does not necessarily mean that it is true.) Finally, ignoring baseline data when estimating τ may be a lost opportunity to increase the precision of $\hat{\tau}$, because Y_0 and Y_1 are likely to be **correlated** within the same subjects.

Difference in Change Scores

A second approach to analysis grants that the experimental and control groups may differ on Y_0 but posits that an effective intervention would produce larger (or smaller, depending on the **alternative hypothesis**) average *changes* in Y over time (i.e. $Y_1 - Y_0$) in the experimental group than in the control group. This approach leads to the following estimate of τ:

$$\hat{\tau} = (\overline{Y}_{1E} - \overline{Y}_{0E}) - (\overline{Y}_{1C} - \overline{Y}_{0C}). \tag{5}$$

Note that (5) can be rewritten as:

$$\hat{\tau} = \overline{Y}_{1E} - [\overline{Y}_{0E} + (\overline{Y}_{1C} - \overline{Y}_{0C})]. \tag{6}$$

Eq. (6) says that, in the absence of an intervention effect, one would expect the mean of Y at follow-up in the experimental group to be whatever the mean of Y was at baseline in that group, plus the average change observed in the control group from baseline to follow-up. With observations at only two time points, a test of the null hypothesis $H_0 : \tau = 0$ can be based either on an unpaired t-test of the difference in mean change scores between groups or (*see* **Student's *t* Statistics**), equivalently, on an **analysis of variance** (ANOVA) with repeated measures [1].

The main advantage of the change-score analysis is that it is a simple and easily understood method to accommodate baseline differences in Y_0 between groups. Perhaps surprisingly, however, it does not automatically improve the precision of $\hat{\tau}$ in comparison with the estimate from (4). Fleiss [7] shows that if σ_0 is the (within-group) **standard deviation** in Y_0, σ_1 is the standard deviation in Y_1, and ρ_{01} is the correlation between Y_0 and Y_1 in the same subjects, then the **standard error** of $\hat{\tau}$ from (5) is less than the standard error of $\hat{\tau}$ from (4) if and only if $\rho_{01} > (\sigma_0/\sigma_1)/2$.

A related shortcoming of change-score analysis applies when Y_0 is measured with error and/or is variable within subjects over time. Many physiologic or behavioral characteristics fit this description: systolic blood pressure, for example, is subject to both errors in measurement (*see* **Measurement Error in Epidemiologic Studies**) and short-term variability within individuals. Under these conditions, change scores tend to be negatively correlated with baseline values [7]: subjects with unusually low values of Y_0 tend to **regress toward the mean** by manifesting larger increases (or smaller decreases) in Y than do other subjects, while those with unusually high values of Y_0 show the opposite.

Finally, change-score analysis provides no way to examine possible **interaction** effects between values of Y_0 and the experimental treatment.

Analysis of Covariance

A third approach uses Y_1 as the dependent variable and treats Y_0 as one of possibly several covariates in an ANCOVA. Figure 2 illustrates the analytic model graphically. Each subject's pair of (Y_0, Y_1) values corresponds to a point on the graph. The two ovals indicate hypothetical clusterings of those points for the experimental and control groups. Within each group, Y_0 and Y_1 are positively correlated, but the data points for the experimental group form a cluster located to the left of those for the control group, implying that $\overline{Y}_{0E} < \overline{Y}_{0C}$. The unadjusted difference in $\overline{Y}_1$ between groups is the length of line segment ac, while the adjusted difference is the length of ab. The length of bc is thus the amount of bias in estimating τ if baseline differences in Y are ignored, as they are in (4).

As shown in [1], the ANCOVA formulation leads to a third estimate of τ:

$$\hat{\tau} = (\overline{Y}_{1E} - \overline{Y}_{1C}) - \beta(\overline{Y}_{0E} - \overline{Y}_{0C}) \qquad (7)$$

where β is the slope of the **regression** of Y_1 on Y_0. In Figure 2, $\overline{Y}^*_{1E}$ corresponds to the height of point b above the horizontal axis.

The relationship of this analysis to the change-score analysis is clarified by noting that (5) can be rewritten as

$$\hat{\tau} = (\overline{Y}_{1E} - \overline{Y}_{1C}) - (\overline{Y}_{0E} - \overline{Y}_{0C}). \qquad (8)$$

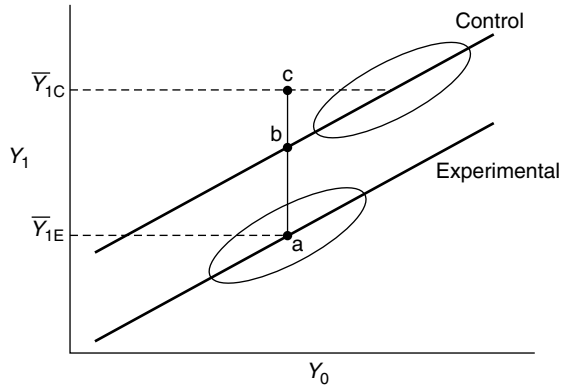

Figure 2 Analysis of covariance for the nonequivalent control group design, using Y_0 as the covariate

In other words, the change-score analysis corresponds to an ANCOVA analysis in which the slope coefficient β in (7) is constrained to be 1. In general, unless Y_1 and Y_0 are completely uncorrelated, $\hat{\tau}$ from (7) can be expected to be more precise than the corresponding estimates from (4) or (5).

Reichardt [13] notes that this basic ANCOVA model can be extended in several ways. The β parameter can be allowed to differ between experimental and control groups, thus allowing study of interaction effects between Y_0 and treatment group membership. Additional covariates besides Y_0 can also be introduced, some of which may be **transformations** of Y_0 to accommodate **nonlinear** relations between Y_1 and Y_0.

The ANCOVA model is not, however, immune to difficulties caused by extraneous **random error** in Y_0 due to measurement error or natural variation in Y over time. Cochran [4] shows that measurement variation in Y_0 leads to biased estimation of the slope parameter β toward 0, by a factor of $\sigma_{(\text{true}Y_0)}/\sigma_{(\text{measured}Y_0)}$, assuming that the measurement error is independent of true Y_0 values. This factor is sometimes termed the *reliability* of Y_0 measurements. Because of this attenuation of β, not all of the confounding by differences in $\overline{Y}_0$ between experimental and control groups is removed.

Because there is only a single baseline observation, pre-existing time trends in Y cannot be detected and taken into account in estimating $\overline{Y}^*_{1E}$. Such trends may arise, for example, from general historical changes that affect Y or from growth or maturation of study subjects over time.

Interrupted Time Series

An interrupted time series quasi-experiment can be diagrammed as follows:

$$\text{O} \quad \text{O} \quad \text{O} \quad \ldots \quad \text{O} \quad \text{X} \quad \text{O} \quad \text{O} \quad \ldots \quad \text{O} \qquad (9)$$

A single study group is examined on repeated occasions, then exposed to the intervention of interest, then examined repeatedly again thereafter. A major advantage of this design is the opportunity to recognize and characterize preintervention patterns of variation over time, which may be cyclical or monotone, and which provide a basis for predicting what would have been observed during the postintervention period in the absence of an intervention effect.

An important feature of time series data, however, is that observations closer to each other in time tend to be more correlated with each other than are observations separated more widely in time (temporal **autocorrelation**). Correlated errors violate a standard assumption of ordinary **least squares** regression and invalidate the usual tests of significance (*see* **Hypothesis Testing**) based on it.

Selection of an approach to analysis depends in part on whether individual subject or only group-level data are available for each of the time periods, and on the number of pre- and postintervention measurements. When a large number (say, 50 or more) of group-level observations are available over time, autoregressive integrated moving average (ARIMA) models, based on work by Box & Jenkins [2], offer several attractive features [10] (*see* **ARMA and ARIMA Models**). Briefly, ARIMA models seek first to model variability within the time series in terms of secular trends, cyclical variation (such as **seasonality**), autocorrelation between serial measurements, and persistence of random perturbations from one measurement to the next. The form of the ARIMA model can then be summarized as ARIMA (p, d, q), where p represents the number of autocorrelation parameters, d the number of differencing steps required to achieve **stationarity**, and q the number of **moving-average** parameters needed to accommodate persistence of random perturbations. A parameter for intervention is then added to the model in order to estimate a change in the series following the point of intervention. O'Carroll et al. [12] used ARIMA modeling to evaluate the effect of a new law in Detroit requiring a mandatory jail sentence for illegally carrying a gun in public. The evaluation was based on changes in monthly counts of firearm-related and non-firearm-related homicides occurring indoors or outdoors in public before and after the ordinance took effect.

Design (9) does not directly allow separation of intervention effects from the effects of extraneous historical factors that happen to coincide in time with the intervention. However, this shortcoming can sometimes be remedied by adding one or more no-intervention control groups followed over the same time period:

$$\begin{array}{cccccccccc} \mathrm{O} & \mathrm{O} & \mathrm{O} & \ldots & \mathrm{O} & \mathrm{X} & \mathrm{O} & \mathrm{O} & \ldots & \mathrm{O} \\ \mathrm{O} & \mathrm{O} & \mathrm{O} & \ldots & \mathrm{O} & & \mathrm{O} & \mathrm{O} & \ldots & \mathrm{O} \end{array} \tag{10}$$

Wagenaar & Holder [16], for example, used ARIMA modeling to assess the effect on alcohol consumption of privatizing wine sales in five US states, examining monthly sales of beer, wine, and spirits in those states before and after privatization and comparing the changes observed with concurrent changes in neighboring states and for the US as a whole.

When the Os in (9) and (10) represent individual-level measurements on many study subjects over time, the analyst can consider several relatively new statistical approaches for **analysis of longitudinal data**. Zeger & Liang [17] review three families of longitudinal analysis models (**marginal**, **transition**, and random effects). All of these approaches allow the use of familiar regression tools relating a response variable to several explanatory variables, while accounting properly for within-subject correlation over time.

Regression Discontinuity

The basic regression-discontinuity design can be denoted as follows:

$$\begin{array}{cccc} \mathrm{O} & \mathrm{C} & \mathrm{X} & \mathrm{O} \\ \mathrm{O} & \mathrm{C} & & \mathrm{O} \end{array} \tag{11}$$

As before, let Y represent a continuous outcome variable, measured at baseline (Y_0) and at follow-up (Y_1). The unique feature of this design is that Y_0 itself is used as the basis for allocating subjects to the experimental or control group, as indicated by the symbol C in (11). If a subject's value of Y_0 falls above (or below) a predefined cutoff value, the subject is assigned to the experimental group; if Y_0 falls on the other side of the cutoff value, the subject is assigned to the control group. Subjects in the experimental group are then exposed to the intervention, and Y_1 is measured at follow-up on everyone in both groups.

This subject-assignment scheme would at first seem to confound Y_0 and the intervention effect hopelessly. But Figure 3 illustrates the results expected if the intervention is effective. As usual, Y_1 and Y_0 are positively correlated within each group. However, the regression relation between them is discontinuous at the cutoff value: in this example, Y_1 scores on experimental-group subjects are shifted higher than Y_1 scores on control-group subjects, presumably because of the effect of the intervention. In particular, the size of the intervention effect is the

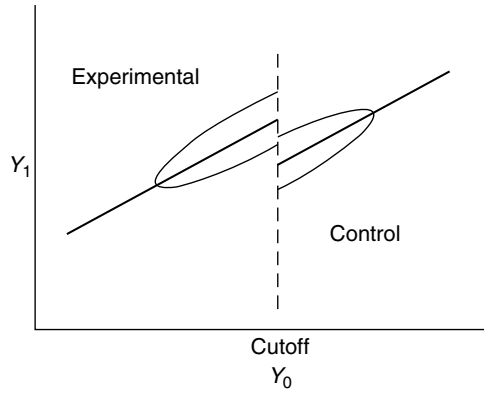

Figure 3 Illustration of the regression-discontinuity design

size of the offset between the two regression lines at the cutoff point.

The regression-discontinuity design has some very attractive features. Often resistance to randomization stems from concern that the neediest subjects would be deprived of a possibly beneficial intervention. Y_0 values may represent the severity of a disorder and may thus be good indicator of "need". The assignment scheme in the regression-discontinuity design assures that all subjects on the "needier" side of some cutoff value of Y_0 receive the intervention. Also, in contrast to many other quasi-experimental designs, in which the basis for nonequivalence of comparison groups may not always be clear, in the regression-discontinuity design the basis for assignment is known to depend completely on Y_0. Even though other potential confounding factors may be associated with Y_0 (and Y_1), they would not usually account for the sharp discontinuity illustrated in Figure 3.

An approach to data analysis from a regression-discontinuity study is described by Trochim [15] who first conceived the design. The basic regression model is:

$$Y_1 = b_0 + b_1(Y_0 - \text{cutoff}) + b_2 z + e \qquad (12)$$

where $z = 1$ for experimental-group subjects and $z = 0$ for control-group subjects, $b_0 = \overline{Y}_1$ in the control group when $Y_0 = \text{cutoff}$, and b_1 is the slope of the regression relation between Y_1 and Y_0 (assumed to be the same in both groups). The parameter of main interest is b_2, representing the intervention effect. (Subscripts indexing individual subjects have again been omitted for Y_1, Y_0, z, and e.)

Two key assumptions in (12) are a common value for b_1 in both groups and a linear relation between Y_1 and Y_0. The former assumption can be relaxed by introducing an interaction term, $b_3(Y_0 - \text{cutoff})z$. If this term improves the fit of the model significantly, it suggests that the intervention effect varies according to Y_0 and that a single summary value of effect may be insufficient. The assumption of a linear relation between Y_1 and Y_0 can be circumvented by adding pairs of terms representing quadratic, cubic, and higher-order powers of $Y_0 - \text{cutoff}$ and their corresponding interaction terms with z, as described by Trochim [15].

Correct inference about the key parameter b_2 turns out to depend heavily on correct specification of the form of the relation between Y_1, Y_0, and z. It also requires strict adherence to the assignment rule represented by C in (11). Moreover, the regression-continuity design has been shown to require about 2.75 times as many subjects as the corresponding randomized-trial design at fixed levels of power and hypothesized intervention effect [13].

Although the regression-discontinuity design has been applied for **program evaluation** in the social sciences and education, it appears to have been rarely used in health research: it is a solution awaiting a problem. However, Johnston et al. [8] describe useful potential applications in rehabilitation medicine, and the design's unique strengths seem worth exploiting in many other health-related contexts.

Conclusion

Quasi-experimental study designs are an attempt to apply some of the desirable features of controlled trials to research situations in which randomization is not possible. **Meta-analyses** that have compared the findings of randomized and nonrandomized studies have often found that the apparent benefits of new medical and surgical therapies over conventional ones tend to be larger in nonrandomized studies [5, 11]. While some of the differences may be due to features of randomized trials that confer a conservative bias (e.g. the **intention to treat** principle), it is widely believed that much is due to

confounding and **selection biases** in nonrandomized studies that cannot be completely overcome without randomization. Investigators would be prudent to continue seeking ways to use randomized designs whenever possible, be diligent in their attempts to measure and control for potential confounding factors when randomization is not possible, be appropriately modest about conclusions from quasi-experiments, and be open-minded as evidence from stronger study designs becomes available.

References

[1] Anderson, S., Auquier, A., Hauck, W.W., Oakes, D., Vandaele, W. & Weisberg, H.I. (1980). *Statistical Methods for Comparative Studies: Techniques for Bias Reduction*. Wiley, New York.

[2] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden Day, San Francisco.

[3] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin, Boston.

[4] Cochran, W.G. (1983). *Planning and Analysis of Observational Studies*. Wiley, New York.

[5] Colditz, G.A., Miller, J.N. & Mosteller, F. (1989). How study design affects outcomes in comparisons of therapy. I: Medical. *Statistics in Medicine* **8**, 441–454.

[6] Cook, T.D. & Campbell, D.T., eds (1979). *Quasi-Experimentation. Design and Analysis Issues for Field Settings*. Houghton Mifflin, Boston.

[7] Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

[8] Johnston, M.V., Ottenbacher, K.J. & Reichardt, C.S. (1995). Strong quasi-experimental designs for research on the effectiveness of rehabilitation, *American Journal of Physical Medicine and Rehabilitation* **74**, 383–392.

[9] Kaplan, R.M. & Berry, C.C. (1990). Adjusting for confounding variables, in *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*, Publication No. 90–3454, L. Sechrest, E. Perrin & J. Bunker, eds. US Department of Health and Human Services, Washington.

[10] McCain, L.J. & McCleary, R. (1979). The statistical analysis of the simple interrupted time-series quasi-experiment, in *Quasi-Experimentation. Design and Analysis Issues for Field Settings*, T.D. Cook & D.T. Campbell, eds. Houghton Mifflin, Boston.

[11] Miller, J.N., Colditz, G.A. & Mosteller, F. (1989). How study design affects outcomes in comparisons of therapy. II: Surgical. *Statistics in Medicine* **8**, 455–466.

[12] O'Carroll, P.W. Loftin, C., Waller, J.B. Jr, McDowall, D., Bukoff, A., Scott, R.O., Mercy, J.A. & Wiersema, B. (1991). Preventing homicide: an evaluation of the efficacy of a Detroit gun ordinance, *American Journal of Public Health* **81**, 576–581.

[13] Reichardt, C.S. (1979). The statistical analysis of data from nonequivalent group designs, in *Quasi-Experimentation. Design and Analysis Issues for Field Settings*, T.D. Cook & D.T. Campbell, eds. Houghton Mifflin, Boston.

[14] Simon, G.E., Grothaus, L., Durham, M.L., VonKorff, M. & Pabiniak, C. (1996). Impact of visit copayments on outpatient mental health utilization by members of a health maintenance organization, *American Journal of Psychiatry* **153**, 331–338.

[15] Trochim, W.M.K. (1990). The regression-discontinuity design, in *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*, Publication No. 90–3454, L. Sechrest, E. Perrin & J. Bunker, eds. US Department of Health and Human Services, Washington.

[16] Wagenaar, A.C. & Holder, H.D. (1996). Changes in alcohol consumption resulting from elimination of retail wine monopolies: results from five U.S. states, *Journal of Studies on Alcohol* **56**, 566–572.

[17] Zeger, S.L. & Liang, K.Y. (1992). An overview of methods for the analysis of longitudinal data, *Statistics in Medicine* **11**, 1825–1839.

THOMAS D. KOEPSELL

Quasi-independence

Many **contingency tables** contain cells whose counts are missing, unreliable, a priori structurally fixed, or ignored in certain hypotheses of interest. A contingency table with structurally fixed cell counts is said to be incomplete or truncated. A valid model for an incomplete contingency table must ignore the cell counts which are structurally fixed, and asymptotic tests of these models must have their degrees of freedom adjusted. The quasi-independence (QI) model is most commonly used to analyze incomplete contingency tables or used when our hypothesis of interest is focused on part of a complete table. For some models of a **square contingency table** only the off-diagonal cells are of interest. For complete contingency tables, QI is a form of independence *conditional* on the restriction of our interest to an incomplete portion of the table.

Incomplete contingency tables usually contain **structural zeros**, i.e. counts of zero in cells in which observations cannot occur. They are often found in genetics and medicine because of biological or logical constraints. Tables of chromosome combinations contain structural zeros for lethal genetic combinations. Cross classifications of diseases by sex contain structural zeros for sex-specific diseases, e.g. females cannot develop testicular cancer. Cross classifications of diseased patients according to birth order and sibship size yield triangular incomplete tables since birth order cannot exceed sibship size. See [5] for many examples.

A classic triangular table discussed by Bishop & Fienberg [4], among many others, is Table 1. Here, ratings of stroke patients, based on a five-point ordinal scale A to E of increasing severity, were obtained on admission to, and discharge from, hospital. As none of these patients had a second stroke and no patient was discharged if their condition worsened, then no discharged patient's final rating can be worse than their initial rating, and so for the given data there are 10 structural zeros. This cross-classification by initial and final rating yields what is termed a triangular table, since the structural zeros form a triangle.

Table 1 Initial and final ratings on disability of 121 stroke patients

		Final state				
		A	B	C	D	E
Initial state	E	11	23	12	15	8
	D	9	10	4	1	0
	C	6	4	4	0	0
	B	4	5	0	0	0
	A	5	0	0	0	0

The Quasi-Independence Model

The quasi-independence model is a useful generalization of the model of independence for a two-dimensional complete contingency table. Let Y_{ij} denote the count in cell (i, j) of a two-dimensional rectangular contingency table. Let S denote a set of cells. The QI model for S assumes that the expected value of Y_{ij}, $\mathrm{E}(Y_{ij})$, has the multiplicative form $\mathrm{E}(Y_{ij}) = \alpha_i \beta_j$ for all cells (i, j) in S, where α_i is a function only of row i and β_j only of column j. Note that the QI model for S makes no assumption about the expected counts in cells not in S, so that these expected counts are ignored by the QI model for S. If S is the set of all cells, then this is the definition of independence, but usually S will be a proper subset, namely the cells not structurally fixed. If the QI model for S holds, then the model of independence will hold conditionally for any rectangular subtable whose cells are all in S.

A loglinear form of the QI model is typically used when the QI model is fitted to complete tables, but the hypothesis of interest relates to only part of the table. The saturated **loglinear model** for the expected cell counts has the form of a constant term, a row-effects term depending only on i, a column-effects term depending only on j, and **interaction** terms for each cell in the table which depend on both i and j. The model of independence corresponds to all the interaction parameters equaling zero. The QI model for S corresponds to all the interaction parameters equaling zero for cells in S and taking nonzero values for cells not in S. In this case the fitted cell counts for cells not in S equal the observed counts, so that cells not in S are effectively ignored in the analysis. Hence, we may think of the QI model for S applied to a complete contingency table as a special model for interaction having a separate interaction parameter for each cell not in S.

The extension of the concept of QI to multidimensional contingency tables is straightforward [8]. Any

loglinear model for a complete multidimensional contingency table can be assumed to hold for a subset S of the cells in the complete table, and these models are called quasi-loglinear models.

Uses for Complete Tables

QI models can be used to analyze complete contingency tables. Sometimes when a loglinear model is fitted to a complete contingency table, a significant lack of fit can be caused by a few outlying cells, with the model fitting the remaining cells well. In two-dimensional tables, QI models have been used to detect outlying cells and to calculate deletion residuals (the **residual** from the expected count with the suspected outlying cells deleted) [16]. A related problem is that when the hypothesis of independence is rejected, the analyst may want to identify those cells that contribute most to the significant **goodness-of-fit** statistic [6].

QI models are often used in the analysis of square tables when both cross-classified variables have the same categories. A classic example is a table of interrater reliability (*see* **Observer Reliability and Agreement**; **Agreement, Measurement of**), where two observers rate a number of subjects on a nominal scale. After modeling patterns of agreement, which involves the main diagonal counts, it is natural to model patterns of disagreement, so that interest focuses on the off-diagonal cell counts. QI for the off-diagonal cells allows for differential agreement by each category and, given that the raters disagree, implies that their ratings are independent [2].

Goodman [11] used QI models to analyze the scalability of the observed response patterns for a set of three or more dichotomous items in order to estimate the proportion of intrinsically scalable (and unscalable) respondents.

Estimation

When the cell counts follow either independent Poisson, multinomial, or product-multinomial sampling, **maximum likelihood** estimation is typically used to fit QI models. Conditions for the existence of maximum likelihood estimates for incomplete tables is discussed in [13]. Explicit formulas for maximum likelihood estimates of the fitted counts usually only exist for contingency tables with very special structures, such as triangular tables [12]. Various iterative methods can be used to find maximum likelihood estimates of the parameters, including Newton–Raphson methods (*see* **Optimization and Nonlinear Equations**), some weighted least squares programs, and **iterative proportional fitting**. Iterative weighted least squares programs, which can be used for maximum likelihood estimation (*see* **Generalized Linear Model**) and which allow prior weights of zero, such as GLIM4, can be used for fitting QI models with a moderate number of parameters. In this case, the QI model for S is fitted by giving prior weights of zero to cells not in S. Alternatively, iterative proportional fitting can be used for fitting QI models with a large number of parameters when the program allows user-defined starting values for the expected counts. Starting values of zero are specified for expected counts of cells not in S, so that the fitted counts for cells not in S always equal zero at each stage of the iterative process. Both weighted least squares and iterative proportional fitting methods implicitly result in the cells not in S being ignored in the modeling.

Many computer packages can be used to fit QI models (*see* **Software, Biostatistical**). A guide to statistical software for categorical data analysis which describes fitting QI models is provided in [1, Appendix A].

Asymptotic Tests and Degrees of Freedom

Because certain cells in the contingency table are ignored when fitting QI models, the degrees of freedom (df) needed for carrying out asymptotic **chi-square tests** based on the **likelihood ratio** or Pearson chi-squared test statistics need adjustment. Procedures for calculating the correct df for asymptotic tests for many of the incomplete tables encountered in practice are given in [5, 10]. However, when the tables include sampling zeros (not structurally fixed), the calculation of the correct df are problematic even in the case of complete tables [14]. One way of diagnosing problems associated with df calculations is to use fitting methods, such as Newton–Raphson, that involve iterative matrix inversions. Typically, this problem will result in a rank problem in the iterative matrix inversions.

Exact Tests

When small cell counts in the contingency table cause concern about the validity of using asymptotic tests, such as in Table 1, exact or Monte Carlo exact tests may be used. For small contingency tables, complete enumeration may be used to calculate the exact ***P* value**. For moderate to large tables, **Monte Carlo methods** can be used to estimate the exact P value [15, 17]. One advantage of exact tests is that the concept of degrees of freedom is unnecessary, so that degrees of freedom calculations are also unnecessary.

Relationships with Other Models

Finally, we note that a number of models have been shown to be equivalent to a QI model. The **Bradley–Terry model** for **paired comparisons** is equivalent to a QI model when the data are rearranged into a specific contingency table format [9]. The model of **quasi-symmetry** or symmetric association for a square contingency table assumes that the interaction parameters in a loglinear model for symmetrically opposite cells are equal. Quasi-symmetry for a 3×3 table is equivalent to QI for the off-diagonal cells and QI for the off-diagonal cells implies quasi-symmetry for larger square tables. Certain **latent class** models are equivalent to QI models [7]. Tests of independence between two discrete or discretized random variables with random **censoring** are equivalent to testing the goodness of fit of a loglinear model applied either to complete or incomplete contingency tables [3].

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Agresti, A. (1992). Modelling patterns of agreement and disagreement, *Statistical Methods in Medical Research* **1**, 201–218.

[3] Akritas, M.G. & Clogg, C.C. (1991). Tests of independence for bivariate data with random censoring: a contingency table approach, *Biometrics* **47**, 1339–1354.

[4] Bishop, Y.M.M. & Fienberg, S.E. (1969). Incomplete two-dimensional contingency tables, *Biometrics* **25**, 119–128.

[5] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[6] Brown, M.B. (1974). Identification of the sources of significance in two-way contingency tables, *Applied Statistics* **23**, 405–413.

[7] Clogg, C.C. (1981). Latent structure models for mobility, *American Journal of Sociology* **86**, 836–868.

[8] Fienberg, S.E. (1972). The analysis of incomplete multiway contingency tables, *Biometrics* **28**, 177–202.

[9] Fienberg, S.E. & Larntz, K. (1976). Log linear representation for paired and multiple comparisons models, *Biometrika* **63**, 245–254.

[10] Goodman, L.A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries, *Journal of the American Statistical Association* **63**, 1091–1131.

[11] Goodman, L.A. (1975). A new model for scaling response patterns: an application of the quasi-independence concept, *Journal of the American Statistical Association* **70**, 755–768.

[12] Goodman, L.A. (1994). On quasi-independence and quasi-dependence in contingency tables, with special reference to ordinal triangular contingency tables, *Journal of the American Statistical Association* **89**, 1059–1063.

[13] Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

[14] Haslett, S.J. (1985). An algorithm for degree of freedom calculations in sparse complete contingency tables, in *GLIM85: Proceedings of the International Conference on Generalized Linear Models*, R. Gilchrist, B. Francis & J. Whittaker, eds. *Lecture Notes in Statistics*, Springer-Verlag, Berlin, pp. 56–65.

[15] McDonald, J.W. & Smith, P.W.F. (1995). Exact conditional tests of quasi-independence for triangular contingency tables: estimating attained significance levels, *Applied Statistics* **44**, 143–151.

[16] Simonoff, J.S. (1988). Detecting outlying cells in two-way contingency tables via backwards-stepping, *Technometrics* **30**, 339–345.

[17] Smith P.W.F., Forster, J.J. & McDonald, J.W. (1996). Monte Carlo exact tests for square contingency tables, *Journal of the Royal Statistical Society, Series A* **159**, 309–321.

JOHN W. MCDONALD

Quasi-likelihood

Quasi-likelihood (QL) is a method of estimation for **regression** analysis with discrete or continuous responses. Like weighted **least squares**, quasi-likelihood requires specification of only the first two **moments** of the response distribution. The quasi-likelihood also refers to the objective function, analogous to a **likelihood**, that is used for inference in this method. Quasi-likelihood (QL) is another successful example of a partially parametric approach to statistical modeling in which only that portion of the probability model of scientific interest is specified. Cox's **proportional hazards model** (*see* **Cox Regression Model**) and the **partial likelihood** [1] method of estimation for survival data are other leading examples.

QL is an extension of **generalized linear models** (GLMs). QL is to generalized linear models as least squares is to the **general linear model**. In the classical linear model, it is assumed that $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n$, where the ε_i are independent Gaussian random variables with mean zero and possibly different variances v_i, here assumed to be known. The maximum likelihood estimate of $\boldsymbol{\beta}$ is then given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y})$, where $\mathbf{X}$ is an $n \times p$ matrix with $\mathbf{x}_i$ as its ith row, $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ and $\mathbf{V} = \text{diag } (v_1, v_2, \ldots, v_n)$. But the same estimator can be arrived at by assuming only that $\text{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and that $\text{var}(\mathbf{Y}) = \mathbf{V}$, and then finding the **minimum variance unbiased estimator** of $\boldsymbol{\beta}$. Whereas the classical linear model specifies the entire probability distribution of the responses, weighted least squares estimators rely only on assumptions about the first two moments of the distribution.

The generalized linear model [16] is a powerful extension of the classical linear model that includes models for discrete (*see* **Categorical Data Analysis**) and non-Gaussian, continuous responses. In a GLM, the response variable is assumed to follow an **exponential family** distribution which includes the **normal**, **binomial**, **Poisson**, **gamma**, and other distributions as special cases. The expected value μ_i of the outcome y_i is assumed to depend on the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ through the *link function* g by $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$. Finally, the variance v_i of the response is a known function of the mean, $v_i = v(\mu_i)$. Most common regression models, including **linear regression, logistic regression**, probit analysis (*see* **Quantal Response Models**), **Poisson** regression and some **survival analysis** models, are special cases of GLMs. The advent of GLMs has unified regression methodology for the diverse types of responses encountered in biostatistical practice.

The score equation (*see* **Likelihood**) for the regression coefficients $\boldsymbol{\beta}$ in a GLM is

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v_i^{-1}(y_i - \mu_i(\boldsymbol{\beta})) = \mathbf{0}, \tag{1}$$

which is typically solved for $\hat{\boldsymbol{\beta}}$ through an iterative weighted least squares algorithm because the weights $v_i^{-1}(\mu_i)$ depend on $\boldsymbol{\beta}$ (*see* **Generalized Linear Model**). In 1974, Wedderburn [22] pointed out that only the mean and variance of the response appear in (1). He showed that the solution of this equation had desirable statistical properties whether or not the mean and variance functions derive from a particular likelihood. Hence, Wedderburn advocated an estimation method in which we specify only: the dependence of the mean μ on explanatory variables x; and the dependence of the variance v of a response on its mean μ, and then solve (1). He further introduced the integral of (1) as the *quasi-likelihood*, to be used as an objective function for **inference**, analogous to the likelihood function when the complete probability distribution of the data is specified. The QL is given by

$$Q(y_i, \mu_i) = \int_{-\infty}^{\mu_i} \frac{y_i - t}{v_i}\, \text{d}t + f(y_i), \tag{2}$$

where $f(y_i)$ is an arbitrary function of y_i.

Properties of QL Estimators

QL estimators have desirable finite sample and asymptotic statistical properties (*see* **Large-sample Theory**). Its finite sample optimality derives from the fact that the QL estimating equation is perhaps the most important example of an *optimal* **estimating function** as defined by Godambe [5, 6]. An optimal estimating function $g(Y, \boldsymbol{\theta}) = \sum_{i=1}^{n} g(y_i, \boldsymbol{\theta})$ is a random function with expectation 0 for all $\boldsymbol{\theta}$ which minimizes

$$S_n = \text{E}\left[\left(\frac{g(Y, \boldsymbol{\theta})}{\text{E}(\partial g(Y, \boldsymbol{\theta})/\partial \boldsymbol{\theta})}\right)^2\right]. \tag{3}$$

This criterion selects for unbiased estimating functions which have small variance (numerator) and steep gradients (denominator). It is an estimating equation analog of the Gauss–Markov criterion for linear unbiased estimators (*see* **Least Squares**). Score functions minimize this criterion. Godambe & Heyde [7] show that the QL estimating function minimizes (3) among unbiased functions that are linear in the data. A related line of theory has established that the QL is the projection of the true score function into a class of unbiased estimating functions [12]. Firth [4] presented specific evidence about the **efficiency** of QL estimators in finite samples.

QL estimators also have two desirable asymptotic properties. First, under regularity conditions, the solution of the QL estimating function is a **consistent** estimator of $\boldsymbol{\beta}$ given only that $\mathrm{E}(y_i) = \mu_i(\boldsymbol{\beta})$, regardless of whether the variance assumption $v_i = v(\mu_i)$ is correct. Hence, the QL estimator will converge to the correct value as long as the regression model for the mean is correct. This important property is illustrated in the example below. Secondly, McCullagh [15] has shown that the QL estimator is asymptotically **unbiased** and efficient among the class the estimating equation of which is a linear function of the data Y.

Example

The data set below [13] is from a teratologic experiment to test the effects of a chemical agent on the survivorship of rat pups. The fraction that survived 21 days of lactation for each dam for the control and treated groups are as follows:

control	13/13	12/12	9/9	9/9	8/8
	4/5	9/10	9/10	8/9	11/13
	8/8	12/13	11/12		
	5/7	7/10	7/10		
treated	12/12	11/11	10/10	9/9	5/10
	8/9	4/5	7/9	4/7	10/11
	3/6	3/10	0/7		
	9/10	9/10	8/9		

There is nearly four times as much variation among these survivor fractions as is expected under a binomial model, indicating that the probability of pup survival is not constant across dams; in other words, that there is a substantial *litter effect* (*see* **Overdispersion**).

The scientific question – Does the chemical exposure decrease the pups' chance of survival? – can be formulated as a logistic regression model. However, there are at least two approaches to account for the litter effect. First, the survival probabilities can be assumed to vary across dams beyond that due to the treatment. The conjugate prior for the binomial sampling distribution is the **beta distribution**, giving a **beta-binomial model** for the observed fractions [11, 23]. As an alternative, the QL approach can be used where the variance of the number of surviving pups is assumed to follow $v_i = \phi n_i \mu_i (1 - \mu_i)$, where n_i is the number of pups in litter i and ϕ is the overdispersion parameter which allows for extra-binomial variation. The QL estimator of the logistic regression coefficients is consistent given that only the logistic model for the expected survival rates is correct. Unfortunately, the same is not true for the beta-binomial model. For example, if the degree of extra-binomial variation (litter effect) is assumed to be the same in the treated and control groups, then the beta-binomial, maximum likelihood estimate of the treatment log odds ratio of survival is $-0.665(\mathrm{se} = 0.460)$. When the litter effects are allowed to differ for the two treatment groups, the treatment effect MLE is $-1.13(\mathrm{se} = 0.464)$, nearly twice as large. But the QL estimate is -0.961 whether or not the overdispersion is assumed to be the same in the two groups.

Hence QL has the important advantage that it unlinks the estimation of the regression coefficients in the mean model from the specification and estimation of the variance model [20].

Extensions of Quasi-Likelihood

Extensions and applications of QL, reaching from methods for **stochastic processes** [7, 24] to **missing data** problems [10], to estimation of **semiparametric regression models** [21], have broadened its influence on biometric research. We focus on four extensions. The first is to allow inference about variance functions. As originally proposed, the QL objective function (2) is useful for inference about the mean model: choosing among **explanatory variables** or for comparing link functions, but it is not useful for choosing among variance functions. Hence, Nelder & Pregibon [19] introduced an extended quasi-likelihood function that includes a second term

$-1/2\log[2\pi\phi v_{\boldsymbol{\theta}}(y)]$ where $\text{var}(y) = \phi \mathbf{v}_{\boldsymbol{\theta}}(\mu)$ is the variance function which is now assumed to be a member of a class, indexed by the parameter $\boldsymbol{\theta}$. Mean parameters that maximize the original QL also maximize the extended QL, but variance parameters can now also be estimated by maximizing the extended version. Finite sample and asymptotic studies of the extended QL have been reported in [3, 18].

The original QL estimating equation is linear in the data. A second interesting extension, proposed by Crowder [2] and Godambe & Thompson [8], is to define an estimating function analogous to (1) for the response $W = (y_1, y_2, \ldots, y_n, y_1^2, y_1y_2, y_1y_3, \ldots, y_{n-1}y_n, y_n^2)$, giving rise to a QL-like method with quadratic estimating functions. In quadratic QL estimation, the first and second moments of the distribution must be correctly specified to make consistent inferences about their parameters, third and fourth moments must be specified to increase the efficiency of estimation.

A third extension deals with the limitation that the original QL equation does not permit **nuisance parameters**. For example, in regression analysis with correlated responses, the regression parameters are often of scientific interest, but **correlation** among responses must be modeled to make valid and efficient inferences. Liang & Zeger [14] proposed a multivariate analog of (1), called the **generalized estimating equation** or GEE, which could include nuisance parameters. They showed that the asymptotic properties of the QL carry over to the GEE as well.

A final important line of QL research addresses the question of how to define the QL objective function for cases, such as the GEE, in which there is more than one path of integration available. Such a function is needed if the QL estimating function has multiple roots, for example. McLiesh & Small [17] project the **likelihood ratio** statistic on to the space of functions linear in the responses and then propose the log of this projection as an objective function. Li & McCullagh [12] obtain a unique quasi-likelihood function by projecting the true score equation, rather than log likelihood, on to a class of estimating functions, each member of which is linear in y_i and is required to be the derivative of an objective function. Hanfelt & Liang [9] have developed an alternate approach in which they do not limit the class of estimating functions, but rather allow the integral of the QL estimating function to be path-dependent. They showed that many useful properties of the log-likelihood, including identifying consistent roots, are preserved by these extended QL functions.

References

[1] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–200.

[2] Crowder, M.J. (1987). On linear and quadratic estimating functions, *Biometrika* **74**, 591–597.

[3] Davidian, M. & Carroll, R.J. (1988). A note on extended quasi-likelihood. *Journal of the Royal Statistical Society, Series B* **50**, 74–82.

[4] Firth, D. (1987). On the efficiency of quasi-likelihood estimation, *Biometrika* **74**, 233–245.

[5] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation, *Annals of Mathematical Statistics* **31**, 1208–1212.

[6] Godambe, V.P. (1985). The foundations of finite sample estimation in stochastic processes, *Biometrika* **72**, 419–428.

[7] Godambe, V.P. & Heyde, C.C. (1987). Quasi-likelihood and optimal estimation, *International Statistical Review* **55**, 231–244.

[8] Godambe, V.P. & Thompson, M.E. (1989). An extension of quasi-likelihood estimation, *Journal of Statistical Planning and Inference* **22**, 137–152.

[9] Hanfelt, J.J. & Liang, K.-Y. (1995). Approximate likelihood ratios for generalized estimating functions, *Biometrika* **82**, 461–477.

[10] Heyde, C.C. & Morton, R. (1996). Quasi-likelihood and generalizing the EM algorithm, *Journal of the Royal Statistical Society, Series B* **58**, 317–327.

[11] Kupper, L.L., Portier, C., Hogan, M. & Yamamoto, E. (1986). The impact of litter effects on dose-response data from certain toxicological experiments, *Biometrics* **42**, 85–98.

[12] Li, B. & McCullagh, P. (1994). Potential functions and conservative estimating function, *Annals of Statistics* **22**, 340–356.

[13] Liang, K.-Y. & Hanfelt, J. (1994). On the use of quasi-likelihood method in teratogical experiments, *Biometrics* **50**, 872–880.

[14] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[15] McCullagh, P. (1983). Quasi-likelihood functions, *Annals of Statistics* **11**, 59–67.

[16] McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models*. Chapman & Hall, London.

[17] McLeish, D.L. & Small, C.G. (1987). A projected likelihood function for semiparametric models, *Biometrika* **79**, 93–102.

[18] Nelder, J.A. & Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: some comparisons, *Journal of the Royal Statistical Society, Series B* **54**, 273–284.

[19] Nelder, J.A. & Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika* **74**, 221–232.

[20] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

[21] Severini, T.A. & Staniswalis, J.G. (1994). Quasi-likelihood, estimation in semiparametric models, *Journal of the American Statistical Association* **89**, 501–511.

[22] Wedderburn, R.W.M. (1974). Quasi-likelihood function, generalized linear models and the Gauss-Newton method, *Biometrika* **61**, 439–477.

[23] Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics* **31**, 949–952.

[24] Zeger, S.L. & Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach, *Biometrics* **44**, 1019–1031.

SCOTT L. ZEGER & KUNG-YEE LIANG

Quasi-symmetry

Quasi-symmetry arises most commonly in the analysis of dependent samples in **contingency tables**. In their simplest form, such tables have a two-dimensional square structure, where the variable for rows has the same categories as the variable for columns (*see* **Square Contingency Table**). Examples are: (i) when the responses of two matched subjects (such as mothers and daughters, or patients and controls) are classified according to some categorical variable (*see* **Matched Pairs With Categorical Data**); (ii) when subjects are categorized according to two essentially similar variables (such as visual faculty of the right and the left eye); and (iii) **panel studies**, where each subject is classified according to the same criterion at two different points in time, or according to the same criterion by two judges.

Three types of symmetry can occur in such situations: complete symmetry, marginal symmetry, and quasi-symmetry. These are introduced in the next Section for two-dimensional tables.

Two-Dimensional Tables

Tables 1–3 illustrate the three types of symmetry for a two-dimensional square table (cf. [5]). In that case, observed frequencies $\{n_{ij}, i, j = 1, \ldots, I\}$ in the cells (i, j) are formed by cross-classifying two categorical responses with a random sample of size n. The joint distribution of $\{n_{ij}\}$ is **multinomial** with parameters n and $\{\pi_{ij}\}$, where π_{ij} is the probability of the cell in row i and column j of the table. *Complete symmetry* is defined by $\pi_{ij} = \pi_{ji}$ for all i and j (see Table 1). The **maximum likelihood** estimates (MLEs) are $\hat{\pi}_{ij} = (n_{ij} + n_{ji})/2n$, and the number of independent constraints is $I(I-1)/2$.

Marginal symmetry, also referred to as *marginal homogeneity*, is defined by $\pi_{i+} = \pi_{+i}$, $i = 1, \ldots, I$, where $+$ denotes the sum, e.g. $\pi_{i+} = \sum_j \pi_{ij}$. Table 2 shows an example. The number of independent constraints is $I - 1$. Although no explicit solutions exists for MLEs, numerical solutions are easily obtained by iterative methods [7, p. 289; 13, p. 493] (*see* **Iterative Proportional Fitting**). Complete symmetry implies marginal symmetry. For 2×2 tables, marginal symmetry conversely implies complete symmetry.

Table 1 Example of complete symmetry

0.10	0.15	0.05	0.30
0.15	0.25	0.10	0.50
0.05	0.10	0.05	0.20
0.30	0.50	0.20	1.00

Table 2 Example of marginal symmetry

0.08	0.17	0.05	0.30
0.12	0.25	0.13	0.50
0.10	0.08	0.02	0.20
0.30	0.50	0.20	1.00

Table 3 Example of quasi-symmetry

0.10	0.05	0.15	0.30
0.05	0.40	0.05	0.50
0.06	0.02	0.12	0.20
0.21	0.47	0.32	1.00

A third kind of symmetry occurs when there are symmetric proportions in so far as it is possible, given asymmetric marginal distributions. *Quasi-symmetry* is a weaker condition than complete symmetry, meaning that the **odds ratios** describing the association structure, rather than the probabilities themselves, are symmetric. Specifically, for all i and j, the odds ratio $(\pi_{ik}\pi_{kj})/(\pi_{ki}\pi_{jk})$ is the same for all values of k. Thus, in Table 3, for $i = 1$, $j = 2$, and $k = 1, 2, 3$: $(\pi_{11}\pi_{12})/(\pi_{11}\pi_{21}) = (0.10 \times 0.05)/(0.10 \times 0.05) = (\pi_{12}\pi_{22})/(\pi_{21}\pi_{22}) = (0.05 \times 0.40)/(0.05 \times 0.40) = (\pi_{13}\pi_{32})/(\pi_{31}\pi_{23}) = (0.15 \times 0.02)/(0.06 \times 0.05) = 1$. Quasi-symmetry applies to square tables of size 3×3 and larger. The number of independent constraints is $(I-1)(I-2)/2$. For $I = 3$, quasi-symmetry is equivalent to **quasi-independence**.

The quasi-symmetry model was introduced by Caussinus [9] in multiplicative form as

$$\pi_{ij} = \alpha_i \beta_j \gamma_{ij},$$

where the α_i describe the rows, the β_j describe the columns, and the interaction parameters γ_{ij} satisfy $\gamma_{ij} = \gamma_{ji}$, for $i \neq j$. All parameters must be suitably constrained, such as $\prod_i \alpha_i = 1$. The model implies that there is a shift in the expected off-diagonal elements, relative to the symmetry model, caused by the differing main effects $\{\alpha_i\}$ of the rows and $\{\beta_j\}$ of the columns. For a **loglinear** representation, see [3, p. 235].

The **likelihood** equations have no direct solution but can be solved iteratively. Caussinus [9] used the method of iterative proportional fitting, Haberman [13] additionally described a method for obtaining ML solutions using the Newton–Raphson procedure (*see* **Optimization and Nonlinear Equations**), and Bishop et al. [7, p. 291] used an incomplete table representation. Using standard software such as SAS, one can fit the complete symmetry model as well as the quasi-symmetry model (cf. [3, p. 276]).

Example

Table 4, given by Stuart [18] and analyzed by Bishop et al. [7, p. 284], describes data on unaided distance vision for the right and left eyes. One would expect that most people have equivalent visual faculties in both eyes. Most observations in Table 4 concentrate on the main diagonal, supporting this hypothesis. Under complete symmetry, the estimated expected cell count is $\hat{m}_{ij} = (n_{ij} + n_{ji})/2$, and the Pearson **chi-square test** statistic for checking its fit (cf. [13, p. 489]) simplifies to

$$X^2 = \sum_{i=2}^{4} \sum_{j=1}^{3} \frac{(n_{ij} - n_{ji})^2}{(n_{ij} + n_{ji})}.$$

This, or the **likelihood ratio test** statistic G^2, has asymptotic χ^2 distributions with df = 6. The values $X^2 = 19.11$ and $G^2 = 19.25$ for Table 4 suggest that the hypothesis of complete symmetry is not tenable ($p = 0.004$).

For the quasi-symmetry (QS) model fitted to Table 4, the goodness of fit statistics equal $X^2 = 7.26$ and $G^2 = 7.27$, with corresponding p values of 0.064. Thus, the QS model provides a significant improvement in fit over the complete symmetry model. Table 5 shows the estimated expected frequencies $\hat{m}_{ij}$ for the QS model.

Table 4 Observed counts for unaided distance vision

Right eye grade	Left eye grade				Total
	Best	Second	Third	Worst	
Best	1520	266	124	66	1976
Second	234	1512	432	78	2256
Third	117	362	1772	205	2456
Worst	36	82	179	492	789
Total	1907	2222	2507	841	7477

Table 5 Expected values under quasi-symmetry for the unaided distance vision data of Table 4

Right eye grade	Left eye grade				Total
	Best	Second	Third	Worst	
Best	1520.00	263.38	133.58	59.04	1976
Second	236.62	1512.00	418.99	88.39	2256
Third	107.42	375.01	1772.00	201.57	2456
Worst	42.96	71.61	182.43	492.00	789
Total	1907	2222	2507	841	7477

The interpretation of quasi-symmetry requires some care, since the model refers to symmetrical odds ratios instead of symmetrical probabilities. For instance, the ML estimate of the odds of having second-degree rather than third-degree visual faculty on the left eye, given best faculty of the right eye, equal (263.38/133.58); compared to the odds of having second-degree rather than third-degree visual faculty on the left eye, given second-degree faculty of the right eye, which is (1512.00/418.99), we obtain an odds ratio of 0.55. Symmetrically, the odds of best grade to second grade on the left eye, given second-degree faculty on the right eye, compared to the odds given third-degree faculty on the right eye, equals (236.62/1512.00)/(107.42/375.01) = 0.55. Thus, the QS model implies a certain kind of symmetric relationship between the visual faculty of the left and the right eye, even though proportions of different grades of visual faculty differ for the right and the left eyes.

Three- and Higher-Dimensional Tables

More generally, consider three responses i_1, i_2, i_3, where each response i_g has I possible categories at any occasion g, $g = 1, 2, 3$. Then an I^3 contingency table summarizes the frequencies of the possible responses at the three occasions. A cell in the contingency table is denoted by $\mathbf{i} = (i_1, i_2, i_3)$ according to the responses. The probability that a randomly selected individual enters cell $\mathbf{i}$ in the contingency table is $\pi(\mathbf{i})$, the expected frequency for that cell is $m_{\mathbf{i}} = n\pi(\mathbf{i})$, and the **marginal probability** for response in category h at the gth occasion is denoted $\pi_g(h)$. Three types of symmetry can be defined as follows:

1. Complete symmetry occurs when $\pi_{\mathbf{i}} = \pi_{\mathbf{j}}$ for any permutation $\mathbf{j} = (j_1, j_2, j_3)$ of $\mathbf{i} = (i_1, i_2, i_3)$. For

instance, if $I = 3$, the responses (0,1,2), (0,2,1), (1,0,2), (1,2,0), (2,0,1), and (2,1,0) all have the same probability.

2. Marginal symmetry of order one is defined by $\pi_1(h) = \pi_2(h) = \pi_3(h)$, for all $h = 1, \ldots, I$. Marginal symmetry of order two additionally assumes equality of the joint marginal probabilities $\pi_{g,g'}(h, h'), = \pi_{g,g'}(h', h)$, for all $g \neq g'$, and $\pi_{12}(h, h') = \pi_{13}(h, h') = \pi_{23}(h, h')$. Marginal symmetry of order two implies that of order one.
3. In an I^3 table, quasi-symmetry of order two or order one is a property of **interactions** of π. Quasi-symmetry of order two is satisfied if there are three sets of parameters $\{\alpha_g(h), h = 1, \ldots, I\}$ and parameters $\{\gamma(\mathbf{i})\}$ that are identical for all permutations of $\mathbf{i} = (i_1, i_2, i_3)$, such that

$$\pi_{\mathbf{i}} = \alpha_1(i_1)\alpha_2(i_2)\alpha_3(i_3)\gamma(\mathbf{i}),$$

with suitable identifiability constraints. Quasi-symmetry of order one holds if there are $3(3-1)/2$ additional parameters $\beta_{gg'}(h, h')$ that satisfy $\beta_{g,g'}(h, h') = \beta_{g,g'}(h', h)$ and $\beta_{12}(h, h') = \beta_{13}(h, h') = \beta_{23}(h, h')$; that is,

$$\pi_{\mathbf{i}} = \alpha_1(i_1)\,\alpha_2(i_2)\,\alpha_3(i_3) \times \beta_{12}(i_1, i_2)\beta_{13}(i_1, i_3)\beta_{23}(i_2, i_3) \times \gamma(\mathbf{i}).$$

Quasi-symmetry of order one implies that of order two.

For a detailed discussion of this case, see [5] or [7]. MLEs are not directly available for either the marginal or the quasi-symmetry models. Bishop et al. [7] gave formulas for calculating the MLEs iteratively and thus the likelihood ratio statistics to test the corresponding hypotheses. Similar considerations apply to higher-dimensional tables, which can exhibit additional orders of quasi-symmetry. A more convenient representation of these models is given by a loglinear formulation (see [1, p. 388]). Computer packages for fitting loglinear models perform the necessary computations to obtain estimates for the symmetry and quasi-symmetry models of general order.

Tests Related to Quasi-Symmetry

An important property of quasi-symmetry is that it is complementary to marginal symmetry with respect to complete symmetry. The quasi-symmetry structure for expected cell counts with the addition of marginal symmetry is equivalent to complete symmetry (cf. [9]); that is,

$$\text{quasi-symmetry} + \text{marginal symmetry} = \text{complete symmetry}.$$

Thus, to test for marginal symmetry, one can test for complete symmetry assuming that quasi-symmetry holds. In terms of likelihood-ratio goodness of fit statistics,

$$G^2(\text{CS}|\text{QS}) = G^2(\text{CS}) - G^2(\text{QS}).$$

For $I \times I$ tables, this difference has on the **null hypothesis**, an approximate **chi-square** distribution with $I - 1$ **degrees of freedom**. For the example on unaided vision, $G^2_{\text{CS}} - G^2_{\text{QS}} = 19.25 - 7.27 = 11.98$ with df $= 3$, suggesting that the hypothesis of marginal symmetry does not hold. Note that complete symmetry implies both marginal symmetry and quasi-symmetry. In the absence of complete symmetry, one could have either marginal symmetry or quasi-symmetry, but not both. For T-dimensional tables Bhapkar & Darroch [6] proved that $H_{\text{CS}} = H^{(k)}_{\text{QS}} \cap H^{(k)}_{\text{MH}}$, where $H^{(k)}_{\text{QS}}$ and $H^{(k)}_{\text{MH}}$ are the hypothesis of quasi-symmetry and marginal symmetry of order k.

Several well-known tests, such as **McNemar's test**, Cochran's Q, or the **Mantel–Haenszel** procedure are special cases of tests for marginal symmetry. See, for instance, [2, p. 229]). Alternatively, one may use the "conditional likelihood score" statistics introduced by Darroch [10].

Other Applications of Quasi-Symmetry

Many useful models are special cases of the quasi-symmetry model. A **Markov chain** is a random process in which individuals move among a limited number of states at certain points in time. The probability of moving from state i to state j at any time point is called a transition probability. For a Markov chain of order one, these conditional probabilities depend only on what happened one time point previously. The considerations of complete, marginal, and quasi-symmetry apply to this model (see [15]). If the conditional probability of transition between events is the same in each direction, then there is

reversibility. In terms of the analysis of contingency tables, this is quasi-symmetry. If the margins do not change over time, then the process described by the Markov chain is in the equilibrium state; that is, the marginal distribution is stationary. In terms of contingency tables, this is marginal symmetry.

Other applications of quasi-symmetry arise in the context of **psychometric** models. The most basic item response model used in educational testing was proposed by Rasch [17] (*see* **Rasch Models**). It applies when a sample of subjects respond to a set of items, such that the observations are Bernoulli (*see* **Binary Data**) ("correct" or "incorrect" answer). In a certain sense, this model can be regarded as a quasi-symmetry model. For theoretical and computational aspects of this connection and for examples in the context of **generalized linear models**, see [14]. The **latent class model** is closely related to the Rasch model, but the latent variable representing the subjects is categorical rather than continuous. For quasi-symmetry in latent class models, see [3] and [4]. Another model closely related to the Rasch model is a model for **paired comparisons** formulated by Bradley & Terry [8] (see **Bradley–Terry Model**). This model has been shown to be equivalent to a quasi-symmetric loglinear model (Fienberg & Larntz [12]), and Dittrich et al. [11] describe how to fit the Bradley–Terry model using GLIM.

Conclusions

In practice, the rather restrictive hypothesis of complete symmetry is often rejected in favor of the more flexible structure of quasi-symmetry. Useful introductions to the topic are [1, 3, 7, and 13]. More advanced sources are [5, 6, 16].

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Agresti, A. (1994). Simple capture–recapture models permitting unequal catchability and variable sampling effort, *Biometrics* **50**, 494–500.

[3] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

[4] Agresti, A. & Lang, J.B. (1993). Quasi-symmetric latent class models, with application to rater agreement, *Biometrics* **49**, 131–139.

[5] Bhapkar, V.P. (1979). On tests of marginal symmetry and quasi-symmetry in two and three-dimensional contingency tables, *Biometrics* **35**, 417–426.

[6] Bhapkar, V.P. & Darroch, J.N. (1990). Marginal symmetry and quasi symmetry of general order, *Journal of Multivariate Analysis* **34**, 173–184.

[7] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis. Theory and Practice*. MIT Press, Cambridge, Mass.

[8] Bradley, R.A. & Terry, M.E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons, *Biometrika* **39**, 324–345.

[9] Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de correlation, *Annales de la Faculté des Sciences de l'Université de Toulouse* **19**, 77–182.

[10] Darroch, J.N. (1981). The Mantel–Haenszel test and tests of marginal symmetry; fixed effects and mixed models for a categorical response, *International Statistical Review* **49**, 285–307.

[11] Dittrich, R., Hatzinger, R. & Katzenbeisser, W. (1997). Fitting paired comparison models in GLIM, *GLIM Newsletter*, **27**.

[12] Fienberg, S.E. & Larntz, K. (1976). Log linear representation for paired and multiple comparison models, *Biometrika* **63**, 245–254.

[13] Haberman, S.J. (1979). *Analysis of Qualitative Data*, Vol. 2. Academic Press, New York.

[14] Hatzinger, R. (1995). *A GLM Framework for Item Response Theory Models*. Habilitationsschrift, Vienna University of Economics.

[15] Lindsay, J.K. (1993). *Models for Repeated Measurements*. Clarendon Press, Oxford.

[16] McCullagh, P. (1982). Some applications of quasi-symmetry, *Biometrika* **69**, 303–308.

[17] Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. The Danish Institute of Educational Research, Copenhagen. (Expanded Edition, The University of Chicago Press, 1980.)

[18] Stuart, A. (1955). A test for homogeneity of the marginal distribution in a two-way classification, *Biometrika* **42**, 412–416.

REINHOLD HATZINGER

Questionnaire Design

In responding to an ever-increasing demand for information gathering and analysis to assist in making policy decisions, and designing effective treatments, many health professionals find themselves in the position of developing and reviewing survey questions or relying on information based on survey data (*see* **Surveys, Health and Morbidity**). The survey research community has known for some time that poorly designed questions can result in poor data quality, particularly in household or general population surveys [15]. Since surveys are a necessary source of data, the importance of improving our understanding of the inherent flaws of survey questions remains critical.

Through the use of a survey interview, it is feasible to collect a wide array of meaningful data covering factors of possible etiologic significance for disease [1, 6, 9, 10]. Information concerning the social, psychological, and economic aspects of each individual's life situation, as well as past experiences and family medical history, can be determined with varying degrees of validity and can be related to the incidence or **prevalence** of disease [8]. Due to the usually retrospective character of such investigations and the difficulties in establishing validity for self-report measures, most researchers accept that answers to these kinds of questions rely in part on **inference**, which may be error-prone [2, 11]. Nevertheless, the broad range of factors that can be explored through surveys make them particularly suitable for sorting out a multitude of hypotheses about health issues, treatments, and outcomes.

Designing a questionnaire requires development of a set of questions used to obtain statistically useful information from an individual. While that task appears to be straightforward and clear cut, questionnaires are difficult to design for several reasons (see Schechter & Herrmann [16] for further discussion). First, each question must provide a valid and reliable measure. Secondly, the questions must clearly communicate the research intention to the survey respondent. Thirdly, the questions must be assembled into a logical, clear instrument that flows naturally and will keep the respondent sufficiently interested to continue cooperation (*see* **Interviewing Techniques**). To meet these challenges, designing a questionnaire should involve three distinct phases: initial questionnaire planning, development and testing of specific questions, and final construction of the data collection instrument as a whole.

Initial Questionnaire Planning

Low estimates of the time needed to develop early drafts of questions can have a serious effect on questionnaire quality. Typically, the person responsible for developing the survey instrument is one of many people involved in the survey process. However, since the survey cannot be fielded without a completed data collection instrument, pressure to develop a questionnaire quickly is often exerted. Experienced questionnaire designers joke that a questionnaire is finished when time runs out, not necessarily when it is the best it can be.

If the instrument needs to be constructed from scratch rather than modified from an existing one, a list of concepts or topics will first be generated. Translating general topics into series of survey questions is challenging and time-consuming. Consultations, meetings, and brainstorming sessions with subject matter experts, survey methodologists, and survey sponsors will be beneficial prior to writing any questions. Reviews of related surveys and data collection instruments that might have measured similar concepts will help facilitate the writing of new questions.

Once a draft of questions is developed, time is needed for iterative cycles of review and revision. This cyclical process is critical to the development of the instrument, because drafts of questions usually help researchers to refine and clarify their research objectives. It is through this process that the questionnaire will evolve, and the concepts included in the survey will become more focused. Reviews should insure that every item in the questionnaire is defensible as a meaningful contribution to the intended analyses. Most designers work with survey sponsors to develop table specifications that illustrate how the data from a series of questions will be analyzed. A critical examination of draft table shells is key to insuring a comprehensive, thoughtful research plan. A related approach that some find helpful is to document in written form the purpose or justification for each question.

Invariably, the length of the questionnaire becomes a topic of debate during the initial planning phase. A

lengthy instrument is cause for concern, because it increases both the cost of interviewer administration and the respondent burden. Survey sponsors and other interested researchers often view a questionnaire as a "once in a lifetime" opportunity to collect data that will contribute significantly to solving a unique problem. Reviewers of draft questions find that their topics of interest increase because they begin to see the rich potential that the data will offer. It is nevertheless necessary to limit the instrument scope for a number of reasons.

Beyond the sheer data collection expenses of long questionnaires is the impact that length has on data quality. Increased response burden may lead to cognitive problems in answering questions that can result in response error. For example, respondents may resort to guessing or less serious efforts at recalling information [13]. Respondents may begin to give wrong answers as a way to shorten their interview. They may ascertain, for example, that answering "No" to a screener question about a medical condition will eliminate a long series of follow-up questions about that condition. Interviewers also become tired and aware of the burden on respondents. In order to complete interviews successfully, they may read questions too quickly or may lead respondents to answers.

Lastly, the survey objectives and draft questions need to be reviewed together to insure that the right types of questions are being asked for a given topic. Questionnaires can contain a mix of questions that ask for reports of knowledge, behavior, practice, attitude, and opinion. In addition, many surveys ask one respondent for proxy reports of others. Consider the similarities and differences in the following questions:

1. Knowledge – How often does the National Cancer Institute recommend that women age 50 and older should get mammograms?
2. Opinion – How often do you think that women age 50 and older should get mammograms?
3. Behavior – Since turning 50, how many mammograms have you had?
4. Proxy – Since turning 50, how many mammograms has your wife had?

Sponsors sometimes suggest questions that will not measure what is needed. During the review process, categorizing questions by type (e.g. knowledge) and reviewing these categories against the list of topics can improve the instrument and subsequent analysis.

Development and Testing of Specific Questions

After draft questions have been developed, the next step is to refine these specific questions and test them with a variety of respondents. Testing questions in a cognitive laboratory offers an opportunity to identify serious question problems in a cost-effective and timely manner. Cognitive interviews are based on the theory that in answering a survey question, respondents mentally process the question in cognitive stages. Tourangeau [20], Willis et al. [21], and others (see Jobe & Herrmann [12] for further discussion) have presented cognitive models that describe four stages: comprehension of the question, retrieval of information, decision or judgment about the question and answer, and the actual response given. Applying principles of cognitive psychology when conducting laboratory interviews helps the researcher understand and resolve the problems that questions can create for respondents.

Terms and Concepts Should be Familiar and Easy to Understand

Problems arise when questions contain unfamiliar, vague, or ambiguous terms and concepts. The respondent's comprehension of the question may not coincide with the intentions of the questionnaire designer but, nevertheless, it serves as the internal frame from which the respondent develops an answer. As a general rule, questions should be written in the standard spoken version of the language being used. Wording that is specific and concrete is more apt to communicate uniform meaning. Furthermore, words that are nontechnical and shorter both in length and in syllables are preferred.

Laboratory interviews often reveal problems associated with question structure. For example, the question can be so brief that the respondent is unsure of the intent of the question. Similarly, a question can be too long and complex, so that the respondent is unsure what the question is really asking. These sorts of question problems can result in the respondent's individual interpretation of what is

being asked. If the respondent is to assume or know something specific before formulating an answer, the relevant information must be included in the question.

Cues and Ordering of Questions Should Serve to Stimulate Recall

Survey respondents are often asked to recall and access information from memory. This process may be fairly direct and a quick and accurate answer to the question may be provided. However, problems can occur when the respondent has to recall too much information, or has to recall information that is not readily accessible in memory. Recalling an event or behavior can be especially difficult if the event was relatively unimportant or trivial to the respondent, if the event happened long ago, or if the question requires recall of too much detail (*see* **Recall Bias**).

Recognizing some degree of error in reporting exact information, there are questionnaire design steps that can help the respondent's memory search. Sometimes, a short series of related questions will jog the respondent's memory and assist in more difficult recall tasks. Providing an anchor for the reference period (such as "Since last Christmas ...") can help place the memory in correct context [14]. In a face-to-face survey, recall can also be stimulated by the use of calendars, timelines, or other visual tools that may help the respondent to organize events in a time sequence. Questions should ask about information that is directly accessible to respondent memory. Recall delay should be kept to a minimum, as the recent past is generally easier to remember than more distant events. Focusing a question only on recent activity may reduce telescoping, the tendency for rare events occurring prior to the designated recall period to be erroneously included in the activity report [4, 5].

As the respondent evaluates retrieved information, he or she judges the accuracy of the response and may decide that the information is not complete or accurate. For certain questions, respondents may quickly decide that the task is too complicated to try to retrieve the information from memory and, rather, will answer with an estimate. The questionnaire designer should anticipate that respondents will rely on estimation strategies if the question requires too much detail or mental calculations.

Ordering and Format of Questions Should be Unbiased and Balanced

The ordering of questions can influence response strategies and response errors. Context effects occur when two or more questions deal with aspects of the same issue or with closely related issues. This takes on practical importance because questionnaires are ordinarily constructed by grouping together questions on the same subject matter which may, ironically, encourage context effects. General questions seem to be more sensitive to order effects than more specific questions. It may be that general questions are so broad that their frame of reference is more open to interpretation and, therefore, respondents are more likely to use the context from another question in their interpretation.

The format of the question itself can also cause cognitive problems. An open format allows the respondent to offer any answer and the interviewer records a verbatim response. However, questionnaires that are designed to collect significant amounts of data from a relatively large pool of respondents usually employ a closed format. Uniform, fixed response categories are provided and the respondent is instructed to select from them. The responses are coded, numbers are assigned to them, and they are counted and statistically analyzed. When determining the format of the responses, however, more than just an open and closed format needs to be addressed. What scales are being used (*see* **Measurement Scale**)? Are the response options consistent with the measure being sought? One typical finding in reviews of questionnaires is that the question and corresponding response categories ask for much more detail than is needed for analysis purposes [17]. This demand for detail can create an unnecessary burden on the respondent which could lead to "satisficing" (providing a quick answer, whether accurate or not, to satisfy the survey requirements [13]).

Assembling the Final Questionnaire

After the questions have been revised and tested, they must be integrated into a final data collection instrument. Sponsors and subject matter experts typically focus on the development and review of individual questions designed to measure specific items of interest. While each questionnaire item must be judged

carefully on its own merit, it is usually the responsibility of the questionnaire designer to fit the items together in a meaningful way so that the entire questionnaire is unified. The goal is for the respondent to sense that the progression between topics and questions is natural and interesting. The smooth flow of the questionnaire should not only increase the respondent's attention and cooperation, but has an important effect on the interviewer's ability to best do his/her job. Survey researchers who have observed field interviews often hear interviewers make comments such as "I know I already asked you this, but I have to ask again". This is troublesome both to the interviewer, who often worries that the respondent will break off the interview prior to completion, and to the respondent, who often becomes aggravated or irritated by questions that are repetitive, inappropriate, or abrupt (*see* **Interviewing Techniques**). It is useful to consider the following when reviewing the overall questionnaire:

1. Is there a logical progression of questions, so that the respondent is drawn into the interview by awakening interest in the topic?
2. Is the respondent first asked items that are simple to answer and then asked those which are more detailed or complex?
3. Is the respondent first asked objective, straightforward questions and then asked for more personal or sensitive information?
4. Is the respondent brought smoothly from one frame of reference to another by use of transitional statements or series of questions that fit well together?

When analyzing the questionnaire as a whole, one should experiment with moving sections of questions. The order of the question series is originally put together in a somewhat random fashion, and yet designers are hesitant to move series of questions around in order to increase logic and administration flow. One check on the flow of the questionnaire is to list, in order, the main topics of each series to determine if the ordering is appropriate and meaningful.

Questions and response formats should be reviewed for consistency. For example, it is undesirable for series of questions to shift abruptly from Yes/No to True/False to Agree/Disagree response alternatives. It is also confusing to vary the use of reference periods. For example, questions about events in the past 12 months should not be interspersed with questions about events in the past 30 days.

A thorough check of skip patterns is essential. Skip patterns insure that inappropriate questions are skipped, and also that appropriate items are not skipped. Both respondents and interviewers can quickly feel irritated by inappropriate, repetitive questions. These patterns should be tested by mock administrations of the questionnaire, following as many different branches of skip logic as possible.

Lastly, the mode of administration must be considered in the final review of the instrument. There are three common modes for data collection: (i) self-administered, where a respondent completes a form with no interviewer interaction; (ii) telephone, where an interviewer calls a respondent and asks survey questions over the phone (*see* **Telephone Sampling**); and (iii) face-to-face, where an interviewer meets with a respondent and asks the survey questions in person (*see* **Interviewer Bias**). For each mode, another level of complexity may be added by using a computer for data entry (*see* **Computer-assisted Interviewing**).

Choice of mode may seem primarily an administrative issue, but it can also affect data quality [3, 7, 18]. For self-administered questions, instructions must be clear, brief, and easy to follow without assistance. Since the respondent cannot ask for clarification, he or she will interpret questions and response alternatives from a personal frame of reference. And if the question or format is too complicated or technical, error may result as the respondent becomes frustrated in attempts to complete the questionnaire.

Personal interviews obviously offer visual communication between the interviewer and the respondent. This is important not only in motivating the respondent to answer accurately, but also in serving to troubleshoot comprehension and recall problems [19]. Interviewers can answer questions and provide detailed explanations as to what the questions really mean. In addition, face-to-face interviews provide excellent opportunities for respondents to use memory aides such as cards and calendars. When designing telephone surveys, questions should be more brief and direct than in face-to-face interviews [7]. Response alternatives should be shorter and easier to understand because the respondent must remember all the choices while selecting an answer.

Conclusions

In the 1980s, a shift in questionnaire design from a literary approach to a scientific approach was begun. Research during this period has resulted in an increased understanding of the mental processes that enable survey respondents to answer questions, and this understanding has produced a dramatic improvement in the quality of questionnaires. Clearly, a good questionnaire grows from research hypotheses that have been carefully studied and thought out. Discussion of the research problem with colleagues and subject matter experts is critical to developing good questions. Questions should be reviewed, revised, and tested on an iterative basis. Furthermore, examining the questionnaire as a whole is an essential element of good questionnaire design. Since health professionals, researchers, and policy makers rely heavily on survey data, current research on ways to reduce **nonsampling errors** in surveys will continue to have emphasis and importance.

References

[1] Aday, L. (1989). *Designing and Conducting Health Surveys*. Jossey-Bass, San Francisco.

[2] Aseltine, R., Carlson, K., Fowler, F. & Barry, M. (1995). Comparing prospective and retrospective measures of treatment outcomes, *Medical Care* **33**, AS67–AS76.

[3] Beatty, P. & Schechter, S. (1994). An examination of mode effects in cognitive laboratory research, in *American Statistical Association 1994 Proceedings of the Survey Research Methods Section*. American Statistical Association, Alexandria, pp. 1275–1280.

[4] Bradburn, N., Huttenlocher, J. & Hedges, L. (1993). Telescoping and temporal memory, in *Autobiographical Memory and the Validity of Retrospective Reports*, N. Schwarz & S. Sudman, eds. Springer-Verlag, New York.

[5] Bradburn, N., Rips, L. & Shevell, S. (1987). Answering autobiographical questions: the impact of memory and inference on surveys, *Science* **239**, 157–161.

[6] Brewer, M., Dull, V. & Jobe, J. (1989). Social cognition approach to reporting chronic conditions in health surveys, in *Vital and Health Statistics*, Series 6, No. 3 (DHHS Publication No. PHS 89–1078), US Government Printing Office, Washington.

[7] Dillman, D. (1978). *Mail and Telephone Surveys: the Total Design Method*. Wiley, New York.

[8] Feldman, J. (1960). The household interview survey as a technique for the collection of morbidity data, *Journal of Chronic Diseases* **11**, 535–557.

[9] Fowler, F. (1989). *Health Survey Research Methods*. DHHS Publication No. PHS 89-3447, US Government Printing Office, Washington.

[10] Friedenrich, C. (1994). Improving long-term recall in epidemiologic studies (editorial), *Epidemiology* **5**, 1–4.

[11] Herrmann, D. (1995). Reporting current, past, and changes in health status: what we know about distortion, *Medical Care* **33**, AS89–AS94.

[12] Jobe, J. & Herrmann, D. (1996). Comparison of survey cognition and models of memory, in *Basic and Applied Memory Research: New Findings*, D. Herrmann, M. Johnson, M. McEvoy, C. Herzog & P. Hertel, eds. Lawrence Erlbaum, Hillsdale, New Jersey.

[13] Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys, *Applied Cognitive Psychology* **5**, 213–236.

[14] Loftus, E. & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events, *Memory and Cognition* **11**, 114–120.

[15] Payne, S. (1951). *The Art of Asking Questions*. Princeton University Press, Princeton.

[16] Schechter, S. & Herrmann, D. (1997). The proper use of self-report questionnaires in effective measurement of health outcomes, *Evaluation and the Health Professions* **20**, 28–46.

[17] Schechter, S., Beatty, P. & Block, A. (1994). Cognitive issues and methodological implications in the development and testing of a traffic safety questionnaire, in *American Statistical Association 1994 Proceedings of the Survey Research Methods Section*. American Statistical Association, Alexandria, pp. 1215–1219.

[18] Schwarz, N., Strack, F., Hippler, H. & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement, *Applied Cognitive Psychology* **5**, 193–212.

[19] Suchman, S. & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews, *Journal of the American Statistical Association* **85**, 232–241.

[20] Tourangeau, R. (1984). Cognitive science and survey methods, in *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, T. Jabine, M. Straf, J. Tanur & R. Tourangeau, eds. National Academy Press, Washington.

[21] Willis, G., Royston, P. & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires, *Applied Cognitive Psychology* **5**, 251–267.

SUSAN SCHECHTER

Quetelet, Lambert–Adolphe–Jacques

Born: February 22, 1796, in Ghent, Belgium.
Died: February 17, 1874, in Brussels, Belgium.

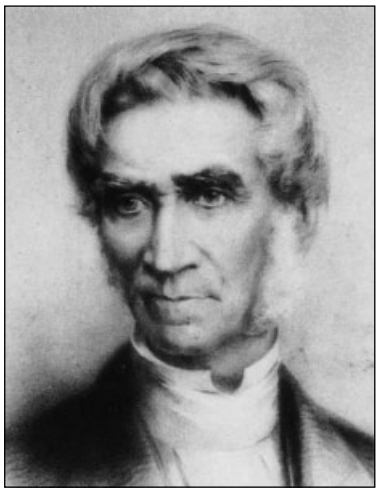

Adolphe Quetelet received his initial scientific training from the University of Ghent. After graduating with a doctor of science degree in 1819, he taught mathematics at a collège in Brussels. When he was elected to the Académie Royale des Sciences et Belles-Lettre de Bruxelles, he campaigned for the founding of an astronomical observatory in Belgium. In 1823, the Belgian government agreed to fund the project and to finance Quetelet's visit to Paris where he learned mathematical and astronomical methods from such scientific notables as Jean Baptiste Joseph Fourier and **Siméon-Denis Poisson** and, possibly, the aging **Pierre Simon Laplace**. In addition to these theoretical concerns, Quetelet's sojourn in Paris also sparked a lifelong interest in social statistics: he began to collect data on things like birth, death, marriage, crime, and suicide, and arranged his results according to age, sex, profession, and place of residence (*see* **Vital Statistics, Overview**).

In 1835, Quetelet combined his twin interests in statistics and the natural sciences in a book entitled, *Sur l'Homme et le Dévelopment de ses Facultés*. By developing an elaborate system of metaphors and similes, Quetelet attempted to describe the statistical regularities in society in terms of the theories of physics and astronomy; he dubbed his activity "social physics" with the central organizing construct being l'homme moyen, or the statistically "average man". The average man produced stability in the social order by serving as a kind of "center of gravity". In addition to this political role, Quetelet [2] also emphasized the centrality of the concept of the average man for medical diagnosis: "The consideration of the average man is so important in medical science, that it is almost impossible to judge of the state of an individual without comparing it to that of another imagined person, regarded as being in a normal condition."

Quetelet's specific analogies from physics and astronomy proved to have little resonance among his contemporaries; nevertheless, he still had a tremendous impact on the development of biostatistical thinking. As discussed in Lécuyer [1], his general belief in the orderliness of the social world as revealed through statistical data was exceedingly influential among some of the leading statistically minded physicians and demographers of the nineteenth century, such as Louis-Adolphe **Bertillon** (1821–1883), Jacques Bertillion (1851–1922), and **William Farr** (1807–1883). Also, Quetelet had a profound impact on the nineteenth-century statistical movement more generally by playing a decisive role in the founding of numerous statistical societies, such as Section F of the British Association for the Advancement of Science (1833), the Statistical Society of London (1834), and various international statistical congresses. Even today, some of Quetelet's statistical ideas are still in use; one prominent example is his measure of body mass (now called Quetelet's Index), which is formed by computing the weight of an individual and dividing it by the height of the individual squared.

References

[1] Lécuyer, B.-P. (1987). Probability in Vital and Social Statistics: Quetelet, Farr and the Bertillions, in *The Probabilistic Revolution: Ideas in History*, Vol. 1, L. Krüger, L.J. Daston & M. Heidelberger, eds. MIT Press, Cambridge, Mass., pp. 317–335.

[2] Quetelet, L.-A.-J. (1962). *A Treatise on Man and the Development of His Faculties*, R. Knox, trans., Burt Franklin Research Source Works Series #247, New York.

J. ROSSER MATTHEWS

Queuing Processes

A facility has s stations, or servers, for serving customers. If all the s stations are occupied, then newly arriving customers must form a queue, or a waiting line, until a station is available for service. The main features in analytic studies of queuing are the length of queue, the waiting time, and the duration of service.

Queuing phenomena can be observed in many practical situations where congestion problems exist. Queuing is apparent in everyday business transactions, in communications, in medical services, in transportation, and in industry. The objective of queuing is to resolve congestion, to optimize efficiency, to minimize waiting time or inconvenience to customers, to speed production, or even to save life.

The queuing concept was originally formulated by Erlang [3] in his study of telephone network congestion problems. In "The Life and Works of K Erlang," published in 1948, there is a good collection of articles of both practical importance and theoretical interest. However, it was the two papers by Kendall [6, 7] that brought queuing problems to the attention of theoretical people. As advances were made in **stochastic processes**, queuing theory also has flourished. Mathematicians, statisticians, engineers, and investigators in many other disciplines have made contributions to queuing theory and to the resolving of practical problems. Actual needs and the curiosity of theoretical investigators have elicited many variations of queuing problems. It has been observed, for example, that customers often arrive in groups of various sizes, rather than singly; unaccommodated incoming phone calls may be lost instead of waiting in a queue (balking); service time may vary from one station to another due to the difference in efficiency of servers (heterogeneous servers); the service provided in medical institutions, in assembly lines, and others, often consists of several phases, each requiring a separate waiting line (tandem queue). Most of the problems have been resolved with great mathematical ingenuity, some beyond the level of this article.

Generally, a queuing system is identified by: (i) the input process (arrivals may be random, planned, or patterned; (ii) the service time distribution; and (iii) the number of stations; or simply by

Input distribution/service time/number of stations.

Symbolically, G ("general") stands for an arbitrary distribution; M ("Markov") for arrivals by a **Poisson process**, **exponential** interarrival time, or exponential service time; and D ("deterministic") for constant interarrival time or service time. For example, we may denote a queuing system with Poisson arrivals, an exponential service time, and s stations by $M/M/s$, and denote a queue with arbitrary arrivals, a constant service time and one service station by $G/D/1$. In this entry, we consider $M/M/s$ queues, for $s > 1$, and discuss in some detail the number of customers in a queuing system, the length of the queue, the waiting time, the service time, and the total amount of time a customer will spend in the system, and their relationship. The corresponding formulas for an $M/M/1$ queue can be obtained from this article with the substitution of $s = 1$. The $M/M/\infty$ queue, in which there are an infinite number of stations, is also presented. The article closes with a brief account of $M/G/1$ queues. Queuing theory is a vast area with an enormous literature. This article provides just a brief introduction. Several important areas are omitted, such as networks of queues (see, for example, [5]) and queues with priorities (see, for example, [8]). The references include several sources for discovering more about this fascinating topic.

$M/M/s$ Queues

In an $M/M/s$ queue, there are s stations (servers), the arrival of customers follows a Poisson process with parameter λ; service time has an exponential distribution with parameter μ; and the service discipline is first come, first served. When all the s stations are occupied at time t, there is a probability $s\mu\Delta + o(\Delta)$ that one of the stations will be free for service within the time element $(t, t + \Delta)$. Let $X(t)$ be the number of customers in the system at time t, including those being served and those in the waiting line. If there are i customers present at time $t = 0$, so $X(0) = i$, let the transition probabilities be

$$P_{i,k}(0, t) = \Pr[X(t) = k | X(0) = i], \quad i, k = 0, 1, \ldots. \tag{1}$$

These satisfy the following system of differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,0}(0, t) = -\lambda p_{i,0}(0, t) + \mu p_{i,1}(0, t),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}p_{i,k}(0,t) = -(\lambda + k\mu)p_{i,k}(0,t) + \lambda p_{i,k-1}(0,t) + (k+1)\mu p_{i,k+1}(0,t), \quad k = 1, \ldots, s-1, \tag{2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}p_{i,k}(0,t) = -(\lambda + s\mu)p_{i,k}(0,t) + \lambda p_{i,k-1}(0,t) + s\mu p_{i,k+1}(0,t), \quad k = s, s+1, \ldots.$$

Instead of solving the above differential equations for every finite t, we consider the limiting case when $t \to \infty$, and let $X = X(\infty)$ be the corresponding limiting **random variable**. We shall call the possible values of $X(\infty)$ the *states* of the queuing system. The states in a system are said to be communicative if any state in the system can be reached by any other state in the system (*see* **Markov Chains**).

In this case, the limiting probabilities

$$\lim_{t\to\infty} \Pr[X(t) = k | X(0) = i] = \pi_k, \quad i, k = 0, 1, \ldots, \tag{3}$$

exist, independently of the initial state i. If, in addition,

$$\sum_k \pi_k = 1, \tag{4}$$

then these limiting probabilities give the stationary distribution of the Markov chain $\{X(t), t \geq 0\}$, otherwise there is no stationary distribution and $\pi_k = 0, k = 0, 1, \ldots$ (see [4, p. 261]).

Since the limiting probabilities are independent of time, their derivatives with respect to time vanish. This provides a series of difference equations, leading to the solution

$$\pi_k = \frac{(s\rho)^k}{k!}\pi_0, \quad k = 1, \ldots, s, \tag{5}$$

$$= \rho^k \frac{s^s}{s!}\pi_0, \quad k = s+1, \ldots, \tag{6}$$

or

$$\pi_k = \rho^{k-s}\pi_s, \quad k = s+1, \ldots, \tag{7}$$

where

$$\pi_0 = \left[\sum_{k=0}^{s} \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!}\frac{\rho}{1-\rho}\right]^{-1} \tag{8}$$

and

$$\rho = \frac{\lambda}{s\mu} \tag{9}$$

is called the *traffic intensity*.

Note that ρ is the ratio of the arrival rate to the service rate when all s servers are busy. The above solution requires that $\rho < 1$. If $\rho \geq 1$, then $\pi_k = 0$, $k = 0, 1, \ldots$, and no stationary distribution exists. It is intuitively clear that the queue eventually grows unboundedly if $\rho > 1$.

The mean and variance of X, the number of customers, are

$$\mathrm{E}(X) = s\rho + \frac{\rho\pi_s}{(1-\rho)^2} \tag{10}$$

and

$$\mathrm{var}(X) = s\rho + \frac{\rho\pi_s}{(1-\rho)^3}[(1+\rho) + s(1-\rho)^2] - \frac{\rho^2\pi_s^2}{(1-\rho)^4}. \tag{11}$$

Length of Queue

Let $Q = Q(\infty)$ be the number of customers in a queue in the limiting case, and let (q_n) be the probability distribution of Q:

$$q_n = \Pr(Q = n), \quad n = 0, 1, \ldots. \tag{12}$$

When there are s or fewer than s customers in the system, no one will be waiting in the line, and hence

$$q_0 = \Pr(X \leq s) = 1 - \frac{\rho}{1-\rho}\pi_s. \tag{13}$$

When there are $k = s + n$ customers in the system, the length of the queue is n, and the probability is

$$q_n = \Pr(X = s + n) = \rho^n \pi_s, \quad n = 1, 2, \ldots. \tag{14}$$

It is easy to verify that

$$\sum_{n=0}^{\infty} q_n = 1. \tag{15}$$

The mean and variance of Q are:

$$\mathrm{E}(Q) = \frac{\rho\pi_s}{(1-\rho)^2} \tag{16}$$

and

$$\mathrm{var}(Q) = \frac{\rho(\rho+1)\pi_s}{(1-\rho)^3} - \frac{\rho^2\pi_s^2}{(1-\rho)^4}. \tag{17}$$

Service Time

The service time t is the length of time needed to complete service to a customer. It is assumed here that all the s stations are equally efficient and have the same exponential distribution, with density function

$$h_t(\tau) = \mu \exp(-\mu\tau), \quad \tau > 0, \tag{18}$$

and distribution function

$$H_t(\tau) = 1 - \exp(-\mu\tau), \quad \tau \geq 0. \tag{19}$$

Let $t_1, t_2, \ldots, t_s$ be the service times of the s stations. They have the same density function in (18) and the same distribution function in (19). If the s stations begin to serve customers at the same time, then it is clear that their service times will have the same distribution. In practice, at a given moment, some of the stations may already have served customers for different lengths of time, and it seems inappropriate to speak of the same distribution when services are already in progress. However, under the present assumption of the exponential density function in (18) where μ is constant, the length of time required to complete service has the same probability distribution, regardless of when the service first started. Therefore it is justified to speak of the same service time distribution for $t_1, t_2, \ldots, t_s$.

Let us arrange the s service times in order of magnitude, and introduce the **order statistics**

$$t_{(1)} < t_{(2)} < \cdots < t_{(s)}.$$

Of course, we do not know in advance which station will have the shortest service time $t_{(1)}$, or which station will have the longest service time $t_{(s)}$. According to the theory of order statistics, when the exponential function (18) is the underlying distribution, the density function of $t_{(1)}$ is

$$f_{t_{(1)}}(\tau) = s\mu \exp(-s\mu\tau), \quad \tau > 0, \tag{20}$$

which has expectation

$$\mathrm{E}(t_{(1)}) = \frac{1}{s\mu} \tag{21}$$

and variance

$$\mathrm{var}(t_{(1)}) = \frac{1}{s^2\mu^2}. \tag{22}$$

It is the distribution of $t_{(1)}$ that is needed for studying the waiting time in a queuing process.

Waiting Time

The waiting time is the length of time that a customer has to wait before receiving service. Let W_n be the waiting time of the nth customer in a queue, $n = 1, 2, \ldots$. Since the first customer in a queue will get service as soon as one of the s stations becomes available, the waiting time W_1 has the same distribution as $t_{(1)}$. The density function of W_1 is

$$f_{W_1}(\tau) = f_{t_{(1)}}(\tau) = s\mu \exp(-s\mu\tau), \quad \tau > 0, \tag{23}$$

and the distribution function of W_1 is

$$F_{W_1}(\tau) = 1 - \exp(-s\mu\tau), \quad \tau \geq 0 \tag{24}$$

which is an exponential distribution with parameter $s\mu$, as noted earlier.

As soon as the first person in a queue leaves the queue to receive service, the second person in the queue becomes the first person in the queue, and he or she in turn will have to wait for another period of W_1 before receiving the service. In other words, the waiting time of the second person in the queue is

$$W_2 = W_{11} + W_{12}, \tag{25}$$

where both W_{11} and W_{12} stand for W_1, the second subscripts are added to differentiate the two periods of waiting. The density function of W_2, $f_{w_2}(\tau)$, is obtained from the density function of W_1:

$$f_{w_2}(\tau) = \int_0^\tau f_{w_1}(\xi)\, f_{w_1}(\tau - \xi)\, \mathrm{d}\xi. \tag{26}$$

Substituting (23) in (26) and integrating the resulting expression, we obtain an explicit formula of the density function of W_2:

$$f_{w_2}(\tau) = (s\mu)^2 \tau \exp(-s\mu\tau), \quad \tau > 0. \tag{27}$$

In general, the waiting time of the nth person in a queue is

$$W_n = W_{11} + W_{12} + \cdots + W_{1n}. \tag{25a}$$

Each of $(W_{11}, W_{12}, \ldots, W_{1n})$ has the same distribution as W_1. Following formula (26), we use the density function of W_1 in formula (23) and repeated integrations to find the following formula of the density function of W_n:

$$f_{W_n}(\tau) = (s\mu)^n \frac{\tau^{n-1}}{(n-1)!} \exp(-s\mu\tau), \quad \tau > 0. \tag{27a}$$

Now, consider a newly arriving customer and their waiting time W. If there are fewer than s customers in the system, then the new customer will receive service immediately, and W will have a value zero with probability

$$\Pr\{W = 0\} = \Pr\{X < s\}$$

or

$$\Pr\{W = 0\} = \sum_{k=0}^{s-1} \pi_k = 1 - \frac{\pi_s}{1-\rho}. \tag{28}$$

If there are $s + n - 1$ customers ahead of the new customer, then his or her waiting time is W_n. Therefore the density function of W is

$$f_W(\tau) = \sum_{n=1}^{\infty} \pi_{s+n-1} f_{W_n}(\tau). \tag{29}$$

Substituting formulas (7) and (27a) in (29), and simplifying the resulting expression, we find the density function of the waiting time W:

$$f_W(\tau) = \pi_s s\mu \exp[-(1-\rho)s\mu\tau], \quad \tau > 0. \tag{30}$$

We should emphasize that the waiting time W is not an ordinary continuous variable, as it has a point of discontinuity at $W = 0$. Hence,

$$\begin{aligned} F_W(\tau) &= \left(1 - \frac{\pi_s}{1-\rho}\right) + \int_0^{\tau} \pi_s s\mu \\ &\quad \times \{\exp[-(1-\rho)s\mu\xi]\}\,\mathrm{d}\xi \\ &= 1 - \frac{\pi_s}{1-\rho}\exp[-(1-\rho)s\mu\tau], \quad \tau \geq 0. \end{aligned} \tag{31}$$

The mean and variance of W are, respectively,

$$\mathrm{E}(W) = \frac{\pi_s}{(1-\rho)^2 s\mu} \tag{32}$$

and

$$\mathrm{var}(W) = \frac{2\pi_s}{(1-\rho)^3(s\mu)^2} - \frac{\pi_s^2}{(1-\rho)^4(s\mu)^2}. \tag{33}$$

Total Length of Time in the System

The total length of time T that a customer spends in a queuing system is the sum of the waiting time W and the service time t:

$$T = W + t, \tag{34}$$

with the density function $g_T(\tau)$ and the distribution function $G_T(\tau)$. While the waiting time W and the service time t are independent continuous random variables, the point of discontinuity of the distribution of W at $W = 0$ invalidates the usual relation of convolution between their density functions. The density function $g_T(\tau)$ may be obtained by way of the distribution function $G_T(\tau)$. Using the relationship among the three distribution functions of T, W, and t,

$$G_T(\tau) = \int_0^{\tau} F_W(\xi)\,\mathrm{d}H_t(\tau - \xi), \tag{35}$$

we find that

$$\begin{aligned} G_T(\tau) &= 1 - \exp(-\mu\tau) - \frac{\pi_s}{(1-\rho)[(1-\rho)s - 1]} \\ &\quad \times \{\exp(-\mu\tau) - \exp[-(1-\rho)s\mu\tau]\} \end{aligned} \tag{36}$$

from which

$$\begin{aligned} g_T(\tau) &= \mu\exp(-\mu\tau) + \frac{\pi_s\mu}{(1-\rho)[(1-\rho)s-1]} \\ &\quad \times \{\exp(-\mu\tau) - (1-\rho)s\mu\exp[-(1-\rho)s\mu\tau]\}, \\ &\quad \tau > 0. \end{aligned} \tag{37}$$

Since $G_T(\infty) = 1$, the distribution of T is proper. From (37) we derive the expectation

$$\mathrm{E}(T) = \frac{\pi_s}{(1-\rho)^2 s\mu} + \frac{1}{\mu} \tag{38}$$

and variance

$$\mathrm{var}(T) = \frac{2\pi_s}{(1-\rho)^3(s\mu)^2} - \frac{\pi_s^2}{(1-\rho)^4(s\mu)^2} + \frac{1}{\mu^2}. \tag{39}$$

Clearly, (38) and (39) can be derived also from the relations:

$$\mathrm{E}(T) = \mathrm{E}(W) + \mathrm{E}(t) \tag{40}$$

and

$$\mathrm{var}(T) = \mathrm{var}(W) + \mathrm{var}(t). \tag{41}$$

A Busy Station

The probability that a station is busy is equal to the traffic intensity ρ. Verification of this probability is straightforward. Since, if $X > s$, the probability that

a station is busy is one, and if $X = k$ for $k < s$, the probability that a station is busy is k/s,

$$\Pr\{\text{a station will be busy at a given moment}\}$$
$$= \sum_{k=0}^{s-1} \pi_k \frac{k}{s} + \sum_{k=s}^{\infty} \pi_k = \rho \left[1 - \frac{\pi_s}{\rho(1-\rho)} \right] + \frac{\pi_s}{1-\rho}$$
$$= \rho. \tag{42}$$

$M/M/\infty$ Queues

When there are infinitely many stations in a system, the number of stations in use is the number of customers present. There are no queues and no waiting times. For $t \geq 0$, let $X(t)$ denote the number of customers present at time t. Again we consider the case where the arrival of customers follows a Poisson process with rate λ and service time has an exponential distribution with parameters μ, so the queuing system under discussion is $M/M/\infty$. In that case, it is easily seen that $\{X(t), t \geq 0\}$ follows the simple migration process (with η replaced by λ) that is analyzed in the article entitled **Migration Processes**, where the distribution of $X(t)$ is determined for any $t \geq 0$ and any given initial number of customers $X(0)$. In particular, a stationary distribution always exists for $X(t)$, and is Poisson with mean λ/μ. Note that this corresponds to letting $s \to \infty$ in the solution (5)–(9) of the $M/M/s$ queue, as one would expect on intuitive grounds.

$M/G/1$ Queues

Whilst the assumption that the arrival of customers follows a Poisson process may be reasonable in many settings (although it precludes multiple arrivals), the assumption that a typical service time follows an exponential distribution is clearly unrealistic in most applications. Thus, we now study $M/G/1$ queues, so there is a single server, customers arrive at the points of a Poisson process with rate λ and the service times are independent and identically distributed, according to some specified distribution.

Method of Stages

The *method of stages* is a device invented by Erlang for determining the properties of $M/G/1$ queues for certain service time distributions. In its simplest form, the service time of a customer is assumed to consist of k stages, labelled 1, 2, ..., k, having independent exponential service times with density given by (18), so the total service time of a given customer follows the gamma distribution with density $f_S(t) = \mu^k t^{k-1} e^{-\mu t}/(k-1)!$ $(t > 0)$. For $t \geq 0$, let $X(t)$ denote the number of customers in the system at time t and, if $X(t) > 0$, let $I(t)$ denote the stage of the customer that is being served. Then the queuing system is completely specified by the process $\{(X(t), I(t)), t \geq 0\}$. Moreover, this process is Markov, owing to the lack-of-memory property of the Poisson process and the exponential distribution. Its stationary distribution can be determined using similar methods to that described for the $M/M/s$ queue (see [2, Section 5.2]) thus facilitating analysis of the queue. Note that the k stages need not correspond to physical stages during a typical service time. Observe that, in this model, service of a typical customer is described by a continuous time Markov chain with $k+1$ states (state $k+1$ corresponds to completion of service), with the service time being given by the time to absorption in state $k+1$. This framework clearly extends so that a typical service time is given by the time to absorption of an arbitrary but specified finite state space continuous time Markov chain; such an absorption time is said to have a **phase-type** distribution. Any given service time distribution can be approximated arbitrarily closely by a phase-type distribution, so this approach provides a flexible framework for computational analysis of queues (see, for example, [1] and [9]).

Embedding

Suppose now that a typical service time, S say, follows any arbitrary but specified distribution. For $t \geq 0$, let $X(t)$ denote the number of customers in the system at t. Now $\{X(t), t \geq 0\}$ is not a Markov process, unless S follows a negative exponential distribution, however, the long-term behavior of the queue can be studied by exploiting a Markov chain that is embedded in $\{X(t), t \geq 0\}$ ([7]), as is now described. For $n = 1, 2, \ldots$, let X_n be the number of customers in the system just after completion of the nth customer's service. The lack-of-memory property of the Poisson arrival process implies that $\{X_n, n \geq 1\}$ is a Markov chain, whose asymptotic properties are analyzed in, for example, Grimmett and

Stirzaker [4, Section 11.3]. Here we just outline the main results.

Let $\rho = \lambda E[S]$ denote the traffic intensity of the queue and suppose that $\rho < 1$. Then both $\{X(t); t \geq 0\}$ and $\{X_n, n \geq 1\}$ are asymptotically stationary and

$$\lim_{t\to\infty} \Pr[X(t) = k | X(0) = i] = \lim_{n\to\infty} \Pr[X_n = k | X_1 = i] = \pi_k, \quad i, k = 0, 1, \ldots, \tag{43}$$

where $\sum_k \pi_k = 1$. Moreover, the limiting probability that the queue is empty is $\pi_0 = 1 - \rho$ and the limiting mean number of customers in the system is

$$E[X] = \rho + \frac{\rho^2 + \lambda^2 \mathrm{var}(S)}{2(1-\rho)}, \tag{44}$$

which is known as the Pollaczek–Khintchine formula. Suppose that the queue is in equilibrium, and let W and $T = W + S$ denote, respectively, the waiting time and the total time spent in the system of a typical customer. Then $E[T]$ satisfies Little's formula

$$E[T] = \lambda^{-1} E[X]. \tag{45}$$

Note that for fixed mean service time $E[S]$, the equilibrium mean queue size, mean waiting time and mean total time spent in the system are all increasing with the variance of a typical service time, so these quantities are all minimized when the service times are constant.

Busy Periods

Suppose that the queue is empty just prior to time $t = 0$ and a customer arrives at $t = 0$. Let B denote the time that elapses until the queue is empty again, that is, the length of a typical busy period of the server. The distribution of B can be studied by considering an embedded branching process, in which individuals are customers in the queue and the offspring of a given customer are those customers that arrive whilst he/she is being served (see, for example, [4, Section 11.3]). In particular,

$$\begin{aligned} &\text{if} \quad \rho < 1 \quad \text{then} \quad E[B] = \frac{E[S]}{(1-\rho)}, \\ &\text{if} \quad \rho = 1 \quad \text{then} \quad E[B] = \infty \text{ and } \Pr\{B = \infty\} = 0, \\ &\text{if} \quad \rho > 1 \quad \text{then} \quad \Pr\{B = \infty\} > 0. \end{aligned} \tag{46}$$

Note that if $\rho > 1$ then with probability one the server is ultimately busy forever.

References

[1] Asmussen, S. (1987). *Applied Probability and Queues*. Wiley, Chichester.

[2] Cox, D.R. & Smith, W.L. (1961). *Queues*. Methurn, London; Wiley, New York.

[3] Erlang, A.K. (1909). Probability and telephone calls, *Nyt Tidsskr. Mat.* **B20**, 33–39.

[4] Grimmett, G.R. & Stirzaker, D.R. (2001). *Probability and Random Processes*, 3rd ed. University Press, Oxford.

[5] Kelly, F.P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester.

[6] Kendall, D.G. (1951). Some problems in the theory of queues, *J. R. Statistical Soc. B*, **13**, 151–185.

[7] Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by means of the imbedded Markov chain, *Ann. Math. Statist.* **24**, 338–354.

[8] Kleinrock, L. (1975). *Queuing Systems, Vol. 2: Computer Applications*. Wiley, New York.

[9] Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore.

Further Reading

Bremaud, P. (1981). *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, Berlin.

Brockmeyer, E., Halstrom, H.L. and Jensen, A. (1948). *The Life and Works of A K Erlang*. Danish Academy of Technical Sciences. Copenhagen.

Bruell, S.C. & Balbo, G. (1980). *Computational Algorithms for Closed Queuing Networks*. North-Holland, New York.

Chaudhry, M.L. & Templeton, J.G.C. (1983). *A First Course in Bulk Queues*. Wiley, New York.

Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. R E Kreiger, New York.

Cohen, J.W. (1982). *The Single Server Queue*, 2nd ed. North-Holland, Amsterdam.

Cooper, R.B. (1981). *Introduction to Queuing Theory*, 2nd ed. North-Holland, New York.

Doig, A. (1957). A bibliography on the theory of queues, *Biometrika*, **44**, 490–514.

Gross, D. & Harris, C.M. (1998). *Fundamentals of Queuing Theory*, 3rd ed. Wiley, Chichester.

Karlin, S. & McGregor, J. (1958). Many server queuing processes with Poisson input and exponential service times, *Pacific J. Math.* **8**, 87–118.

Kleinrock, L. (1975). *Queuing Systems, Vol. 1: Theory*. Wiley, New York.

Saaty, T.L. (1961). *Elements of Queuing Theory*. McGraw-Hill, New York.

Takacs, L. (1962). *Introduction of the Theory of Queue*. Oxford University Press, New York.
van Dijk, N.M. (1993). *Queuing Networks and Product Forms: A Systems Approach*. Wiley, New York.

CHIN LONG CHIANG & FRANK BALL

QUORUM

Like any research enterprise, particularly one that is observational, the meta-analysis of evidence can be flawed. Accordingly, the process by which meta-analyses are conducted has recently undergone scrutiny. A 1987 survey of 86 English-language meta-analyses [5] assessed each publication on 23 items from six content areas considered important in the conduct and reporting of a meta-analysis of randomized trials: study design, combinability, control of bias, statistical analysis, sensitivity analysis, and problems of applicability. The survey results indicated that only 24 (28%) of the 86 meta-analyses reported addressed all six content areas. The updated survey, which included more recently published meta-analyses, showed little improvement in the rigor with which they were reported [6].

To help overcome inadequate reporting several authors have suggested guidelines for reporting meta-analyses [2, 7]. However, a consensus across disciplines about how meta-analyses should be reported has not been developed. Following the Consolidated Standards of Reporting Trials (CONSORT) initiative to help improve the quality of reporting of randomized trials, the Quality of Reporting of Meta-analyses (QUOROM) conference was organized to address these issues as they relate to meta-analyses of randomized trials.

The QUOROM group comprised 30 clinical epidemiologists, clinicians, statisticians, editors and researchers. In conference, the group was asked to identify items they thought should be included in a checklist of standards. Whenever possible, checklist items were guided by research evidence suggesting that failure to adhere to the item proposed could lead to biased results. For example, authors are asked (under the "Methods" heading and "Searching" subheading) to be explicit about the publication status of reports included in a meta-analysis. Only about one-third of published meta-analyses report the inclusion of unpublished data [1].

The role of the grey literature (i.e. literature that is difficult to locate and/or retrieve) was examined in 39 meta-analyses that included 467 randomized controlled trials (RCTs), 102 of which were grey literature. On average, the exclusion of grey literature, compared with its inclusion, resulted in a statistically significant exaggeration of the effectiveness of an intervention by 15% [3].

A modified Delphi technique was used in assessing candidate items.

The conference resulted in the QUOROM statement, an 18-item checklist (Table 1) and a flow diagram (Figure 1). The checklist describes an optimal way to present the Abstract, Introduction, Methods, Results and Discussion sections of a report of a meta-analysis. It is organized into six headings and 14 subheadings. Subheadings in the Methods section include searches selection, validity assessment, data abstraction, study characteristics, and quantitative data synthesis. In addition, the Results section is broken into study characteristics and quantitative data synthesis. Research documentation was identified for nine of the 18 items. The flow diagram provides information about both the number of RCTs identified, included, and excluded and the reasons for excluding trials, throughout the meta-analytic process.

The QUOROM Statement was published in *The Lancet* in 1999 [4] and is available (checklist and flow diagram) on the CONSORT Internet site (`www.consort-statement.org`).

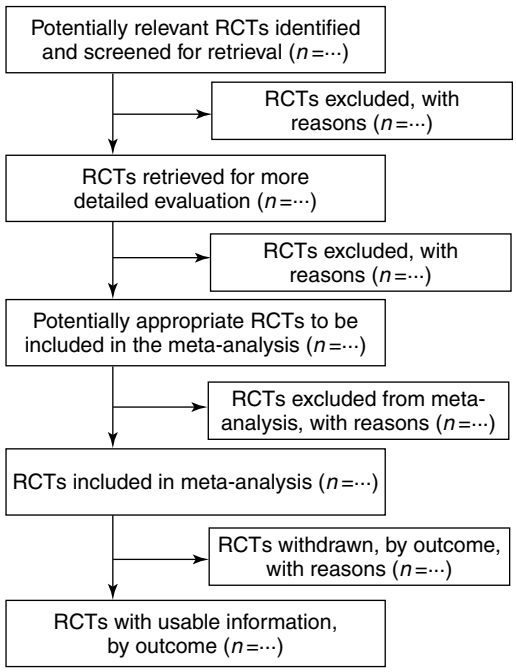

Figure 1 QUOROM flowchart

Table 1 QUOROM checklist

Heading	Subheading	Descriptor	Reported? [Y/N]
Title		Identify the report as a meta-analysis [or systematic review] of RCTs	
Abstract		Use a structured format	
		Describe	
	Objectives	The clinical question explicitly	
	Data sources	The databases, i.e. list and other information sources	
	Review methods	The selection criteria (i.e. population, intervention, outcome, and study design); methods for validity assessment, data abstraction, and study characteristics, and quantitative data synthesis in sufficient detail to permit replication	
	Results	Characteristics of the RCTs included and excluded; qualitative and quantitative findings (i.e. point estimates and confidence intervals); and subgroup analyses	
	Conclusion	The main results	
		Describe	
Introduction		The explicit clinical problem, biological rationale for the intervention, and rationale for review	
Methods	Searching	The information sources, in detail [e.g. databases, registers, personal files, expert informants, agencies, hand-searching], and any restrictions (years considered, publication status, language of publication)	
	Selection	The inclusion and exclusion criteria defining population, intervention, principal outcomes, and study design	
	Validity assessment	The criteria and process used [e.g. masked conditions, quality assessment, and their findings]	
	Data abstraction	The process or processes used [e.g. completed independently, in duplicate]	
	Study characteristics	The type of study design, participants' characteristics, details of intervention, outcome definitions, and how clinical heterogeneity was assessed	
	Quantitative data synthesis	The principal measures of effect [e.g. relative risk], method of combining results (statistical testing and confidence intervals), handling of missing data, how statistical heterogeneity was assessed, a rationale for any a priori sensitivity and subgroup analyses, and any assessment of publication bias	
Results	Trial flow	Provide a meta-analysis profile summarising trial flow (see Figure 1)	
	Study characteristics	Present descriptive data for each trial [e.g. age, sample size, intervention, dose, duration, follow-up period]	
	Quantitative data synthesis	Report agreement on the selection and validity assessment; present simple summary results (for each treatment group in each trial, for each primary outcome); present data needed to calculate effect sizes and confidence intervals in intention-to-treat analyses (e.g., 2×2 tables of counts, means and SDs, proportions)	
Discussion		Summarize key findings, discuss clinical inferences based on internal and external validity; interpret the results in light of the totality of available evidence; describe potential biases in the review process [e.g. publication bias]: and suggest a future research agenda	

At the time of this writing, 10 medical journals have participated in a randomized trial evaluating the impact of applying the QUOROM criteria on journal peer review. Accrual to this trial is now complete and the results should become available by the Spring of 2001.

References

[1] Cook, D.J., Guyatt, G.H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., McIlroy, W. & Oxman, A. (1993). Should unpublished data be included in meta-analyses? Current convictions and controversies, *Journal of the American Medical Association* **269**, 2749–2753.

[2] Cook, D.J., Sackett, D.L. & Spitzer, W. (1995). Methodologic guidelines for systematic reviews of randomized controlled trials in health care from the Potsdam consultation on meta-analysis, *Journal of Clinical Epidemiology* **48**, 167–171.

[3] McAuley, L., Pham, B., Tugwell, P. & Moher, D. (2000). Does the inclusion of Grey literature influence the estimates of intervention effectiveness reported in meta-analyses?, *Lancet* **356**, 1228–1231.

[4] Moher, D., Cook, D.J., Eastwood, S., Olkin, I, Rennie, D., Stroup, D.F., for the QUOROM Group (1999). Improving the quality of reports of meta-analyses of randomised controlled trials; the QUOROM statement, *Lancet* **354**, 1896–1899.

[5] Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A. & Chalmers, T.C. (1987). Meta-analyses of randomized controlled trials, *New England Journal of Medicine* **316**, 450–455.

[6] Sacks, H.S., Reitman, D., Pagano, D. & Kupelnick, B. (1996). Meta-analysis: an update, *Mount Sinai Journal of Medicine* **63**, 216–224.

[7] Shea, B., Dubé, C. & Moher, D. (2000). Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools, in *Systematic Reviews in Health Care: Meta-analysis in Context*, 2nd Ed., M. Egger, G. Davey-Smith & D.G. Altman, eds. British Medical Journal Publishing Group, London.

(*See also* **CONSORT**)

DAVID MOHER

Quota, Representative, and Other Methods of Purposive Sampling

The term *purposive* applies to any method of choosing a sample in which the probabilities of selection for the various sampling units cannot be calculated. For example, consider a **target population** of hospital patients. A full **probability sample** can be achieved by assigning each patient an identification label and selecting a sample wherein each subject has the same chance of selection (e.g. a **simple random sample**) using random numbers. Alternatively, the researcher can select a purposive sample based on personal judgment – literally, "this patient", "that female", "this elderly person over here", etc. The researcher might sincerely believe that by using judgment instead of letting the patients be chosen by **randomization** a more representative sample, or one that is better balanced on key variables, will result. The purposive sample, however, will always be vulnerable to charges of **selection bias** that never can be satisfactorily answered. Furthermore, without knowledge of the selection probabilities for each unit in the sample, estimation of the **mean square error** in purposive sampling is highly subjective, if not impossible.

Kruskal & Mosteller [8], and also Kish [7], have observed that a precise definition of "representative sampling" is elusive. Historically, the term was first used by Kiaer [4] to refer to sampling, in contrast to full enumeration. His concept of the "representative method", however, was broad enough to include purposive methods as well as what would be called probability sampling today. Indeed, this generality of meaning was carried down to the title of **Neyman's** seminal paper [12]. "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection". The term *representative* has since, however, disappeared from the technical lexicon of sampling. We still encounter the informal use of "representative sample" in newspaper articles and in occasional requests for proposal (RFPs) from various agencies. Whether they fully realize it or not, when prospective sponsors of survey research ask for a "representative sample" they mean one that will lead to valid statistical **inferences**, supported by estimated sampling errors that are small, and by evidence that **biases** due to selection, noncoverage, **nonresponse**, and various sources of **measurement error** are not serious. In other words, *representative* now implies probability sampling, hopefully, under conditions sufficiently optimal that the results will stand up under close scrutiny by the scientific community or statistically sophisticated courts of law.

A Situation Where Purposive Sampling is Permissible

We have defined purposive sampling as involving selection by judgment instead of randomization. One case in which a judgment selection may be preferable to strictly random sampling is that in which the survey budget permits replication in only a small number of sites. An example is a public health experiment for which probability samples of school children in more than one metropolitan area are desired. If the costs of administering the survey in a city are so high that only a few different metropolitan locations can be selected, it is better to use expert judgment to achieve "balance" in the selection of the cities rather than let the vagaries of simple randomization determine which areas are studied. An alternative, however, to selection of the sites by judgment is the method of controlled selection [6, Section 12.8]. Controlled selection, which involves random selection of a design from a set of multistratified layouts, might be called a combination of random and purposive selection since the designs to be selected are constructed purposively.

Quota Sampling

As practiced today, quota sampling is a purposive method that enables one to obtain a desired number of completed personal interviews in a relatively short period of time without the expense of call-backs (*see* **Call-backs and Mail-backs in Sample Surveys**). In the form with the least potential bias, referred to by Sudman [16] as "probability sampling with quotas", the method involves drawing sampling locations down to the block level with exactly the same technique of **multistage sampling** selection with **sampling probabilities proportional to size** as in full

probability area sampling. Then, instead of field listing of housing units and probability sampling of households and persons contained therein, interviewers are allowed to fill quotas of respondents according to the availability of qualified subjects and whatever personal judgment may enter into the selection process. (This is why the method must be called "purposive".) In the US, the predetermined quota controls, usually few in number for administrative feasibility, are set according to the most recent **census** counts for the smallest area that immediately surrounds the sampled location. For example, an interviewer may be instructed to begin calling on households at a particular corner of a city block and proceed from door to door until five adult females and five adult males are interviewed. Not-at-homes and refusals are ignored by continuing to the next household where a cooperating individual may be found. With this procedure, it is possible that a cluster of 10 interviews, balanced with respect to sex, can be completed in a day or two. If greater control over selection is desired, the sex strata can be broken down into age groups within sex, making the filling of quotas more time-consuming, but forcing the demographic characteristics of the cluster to be more congruent with those of the immediately surrounding area. It is these considerations that led Cochran [1, p. 136] to describe quota sampling as "stratified sampling with a more or less nonrandom selection of units within strata".

Criticism and Shortcomings of Quota Sampling

Other textbooks, in addition to [1], are critical of the use of quota sampling. Deming [2, p. 31] dismisses the technique in only a few words. While Yates [17] and Kish [6] can scarcely be said to endorse the method, they do allow for occasions when the results from quota samples may be useful. Kish [6, p. 565], for example, says that he believes that "a quota sample is more likely to represent the attitudes of the nation's young people than a probability sample of a college's students", but he is, of course, not so much praising quota sampling as he is deploring an inferential leap from the population of one college to the youth of a whole country. The fundamental problem with quota sampling is that, in the absence of a full understanding of the selection mechanism that determines which individuals shall participate in the survey, one can never be sure that serious biases are not working against valid **estimation** and **inference**. For example, tests of significance (*see* **Hypothesis Testing**) and **confidence interval** estimates, if done at all, must use estimated **standard errors** about biased **means**. These may be very different from the root mean square errors about **target population** parameters that are needed for meaningful inferences. As Yates [17, p. 84] observes, "the fact that a quota system has consistently given reliable results over a period of years is no guarantee that it will also do so in the future".

Research Comparing Quota Sampling and Full Probability Sampling

Immediately after the failure of the pre-election polls to forecast the victory of the incumbent President Harry S. Truman over Thomas E. Dewey in November 1948, the US Social Science Research Council appointed a committee of statisticians and sociologists to study the technical procedures and methods of interpretation that had been used. The method of selection used by the principal polling organizations was almost uniformly quota sampling because of its time and cost advantages. Although the full blame for the forecasting débâcle is not attributed to sampling – equally important was the failure of the pollsters to interpret the undecided vote and to detect the shift in voter attitudes that occurred near the end of the campaign – the committee's report [11] observes that quota methods tended to be biased against the lower educational attainment classes and against rural dwellers.

Other comparisons of quota and probability sampling are reported by Stephan & McCarthy [14], including the carefully controlled experiments of Moser & Stuart [10] in the UK. Sudman [16] and Stephenson [15] discuss similar studies performed at the National Opinion Research Center (NORC). In a theoretical paper, King [5], applying a **Bayesian** model of biased measurement to the allocation of fixed resources to either quota sampling, full probability sampling, or a combination of both methods, finds that quota sampling is only economically feasible when the prior **correlation** between the means of the unbiased and biased processes is very high. In plain language, one has to be fairly certain that the results between quota and full probability methods will be in agreement before quota sampling should be used.

Current Practice

Although widely used by American market research firms and polling organizations during the years in which the empirical studies described above were carried out, quota sampling has largely been supplanted by **random digit dialing** and telephone interviews (*see* **Telephone Sampling**). The extent to which the latter method achieves the equivalent of full probability sampling depends on the way in which refusals are dealt with and the thoroughness of call-back procedures, but it is probable that many of the concerns about bias that were present with quota samples have been alleviated. In the UK, however, quota methods still are used extensively. In fact, the public opinion polls – almost exclusively quota samples – conducted before the British 1992 general election were by everyone's reckoning a statistical disaster. Over 50 polls conducted in the month before the election gave an average lead of 1.5 percentage points to the Labour Party, but the final result was a Conservative victory by 7.6 percentage points. This polling calamity set off a flurry of investigative activity reminiscent of the American 1948 post-election experience mentioned above. Recent papers by Jowell et al. [3], Lynn & Jowell [9], and Smith [13] all discuss whether the incorrect results can be blamed (at least in part) on quota sampling and its inability to deal with refusals. Although almost 40 years have passed since their work was published, the following quote from Stephan & McCarthy [14] in commenting on the experiments of Moser & Stuart [10] would appear to be as relevant as ever:

> the fundamental status of quota sampling, as far as evidence derived from direct comparisons is concerned, is left more or less as outlined ... Instances of serious bias can be found; close agreement with check data or with probability sample results exists for many items; the sources of serious bias are frequently related to the socio-economic control; the actual allocation of bias among possible sources is extremely difficult; and it seems impossible to place quota sampling on a sound theoretical basis.

References

[1] Cochran, W.G. (1977). *Sampling Techniques*, Wiley, New York.

[2] Deming, W.E. (1960). *Sample Design in Business Research*. Wiley, New York.

[3] Jowell, R., Hedges, B., Lynn, P., Farrant, G. & Heath, A. (1993). The 1992 British Election: the failure of the polls, *Public Opinion Quarterly* **57**, 238–263.

[4] Kiaer, A.N. (1895–1896). Observations et expériences concernants des dénombrements représentatifs, *Bulletin of the International Statistical Institute* **9**, 176–183.

[5] King, B.F. (1985). Surveys combining probability and quota methods of sampling. *Journal of the American Statistical Association* **80**, 890–896.

[6] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[7] Kish, L. (1995). The hundred years' war of survey sampling, *Statistics in Transition* **2**, 813–830.

[8] Kruskal, W.H. & Mosteller, F. (1980). Representative Sampling IV: the history of the concept in statistics, 1895–1939, *International Statistical Review* **48**, 169–195.

[9] Lynn, P. & Jowell, R. (1996). How might opinion polls be improved? The case for probability sampling, *Journal of the Royal Statistical Society, Series A* **159**, 21–28.

[10] Moser, C.A. & Stuart, A. (1953). An experimental study of quota sampling, *Journal of the Royal Statistical Society, Series A* **116**, 349–394.

[11] Mosteller, F., Hyman, H., McCarthy, P.J., Marks, E.S. & Truman, D.B. (1949). *The Pre-Election Polls of 1948,* Bulletin 60. Social Science Research Council, New York.

[12] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society* **97**, 558–606.

[13] Smith, T.M.F. (1996). Public opinion polls: the UK general election, 1992, *Journal of the Royal Statistical Society, Series A* **159**, 535–545.

[14] Stephan, F.F. & McCarthy, P.J. (1958). *Sampling Opinions*. Wiley, New York.

[15] Stephenson, C.B. (1979). Probability sampling with quotas: an experiment, *Public Opinion Quarterly* **43**, 477–496.

[16] Sudman, S. (1967). *Reducing the Cost of Surveys*. Aldine, Chicago.

[17] Yates, F. (1981). *Sampling Methods for Censuses and Surveys*, 4th Ed. Macmillan, New York.

B. KING

R- and Q-analysis

This article discusses four alternative criteria for performing **principal components analysis** on a data matrix. The properties of these four criteria were given by Okamoto [6]. A numerical example illustrating all four criteria on a single data set may be found in Jackson [5].

R-Analysis

In principal component analysis, one generally begins with an $n \times p$ data matrix $\mathbf{X}$ representing n observations on p variables. From this, some type of $p \times p$ dispersion matrix is formed, usually a **covariance matrix** or its related **correlation** matrix. A set of linear transformations, utilizing the characteristic vectors of this matrix, is found which will transform the original correlated variables into a new set of variables. The variables in this new set are uncorrelated and are called *principal components*. The values of the transformed data are called *principal component scores*. Further analysis may be carried on in terms of these scores. (A subset of these transformed variables may be retained for this analysis.) This procedure is referred to as *R-analysis* and is the most common application of principal components analysis. Similar procedures are carried out in **factor analysis**.

Q-Analysis

The situation may arise where one may wish to reverse the process and study the relationships among the observations rather than the variables. This is referred to as *Q-analysis*. In this case, an $n \times n$ covariance or correlation matrix will be formed and the characteristic vectors and principal component scores obtained from these. Generally, $n > p$, so that covariance or correlation matrices will not have full rank and there will be a minimum of $n - p$ zero characteristic roots (*see* **Eigenvalue**). Q-analysis may be used in conjunction with clustering of the individuals in the data set. Some **multidimensional scaling** techniques are an extension of Q-analysis and are often used where the data are not homogeneous and require segmentation.

N-Analysis (Singular Value Decomposition)

With proper scaling or normalization, the characteristic vectors of R-analysis become the principal component scores of Q-analysis, and vice versa. These relationships can be extended to *N-analysis or singular value decomposition* [1, 4] (*See* **Correspondence Analysis**), where the characteristic roots and vectors (**eigenvectors**), as well as the principal component scores, may be determined directly from the data matrix.

Relationships Among R-, Q-, and N-Analysis

To show the relationships among R-, Q-, and N-analysis, assume that the $n \times p$ data matrix $\mathbf{X}$ has variable means equal to zero and that $\mathbf{X}$ has rank p. For R-analysis, the operations will be carried out on the $p \times p$ dispersion matrix $\mathbf{X}'\mathbf{X}$, whose unit characteristic vectors will be denoted by the $p \times p$ matrix $\mathbf{U}$ and whose characteristic roots will be the diagonal elements of the $p \times p$ matrix $\mathbf{L}$. $\mathbf{U}$ and $\mathbf{L}$ are used to obtain the principal component scores, $\mathbf{Y}$, for the observations, by the relationship $\mathbf{Y} = \mathbf{X}\mathbf{U}\mathbf{L}^{-1/2}$. For Q-analysis, the operations will be performed on the $n \times n$ dispersion matrix $\mathbf{X}\mathbf{X}'$. The nonzero characteristic roots will be the same $p \times p$ matrix $\mathbf{L}$. The corresponding p characteristic vectors will be denoted by the $n \times p$ matrix $\mathbf{U}^*$. $\mathbf{U}^*$ and $\mathbf{L}$ are used to obtain the principal component scores $\mathbf{Y}^*$ for the variables. It can be shown that the principal scores from the R-analysis are equal to the characteristic vectors from the Q-analysis, and vice versa. That is,

$$\mathbf{Y} = \mathbf{X}\mathbf{U}\mathbf{L}^{-1/2} = \mathbf{U}^* \text{ and } \mathbf{Y}^* = \mathbf{X}'\mathbf{U}^*\mathbf{L}^{-1/2} = \mathbf{U}.$$

In singular value decomposition, these quantities may be obtained for either R- or Q-analysis from the data matrix $\mathbf{X}$ directly, i.e.

$$\mathbf{X} = \mathbf{Y}\mathbf{L}^{1/2}\mathbf{U}' = \mathbf{U}^*\mathbf{L}^{1/2}\mathbf{Y}^*,$$

and, using the relationships above, it can be shown that

$$\mathbf{X} = \mathbf{U}^*\mathbf{L}^{1/2}\mathbf{U}'.$$

This relationship is employed in *dual-scaling* techniques where both variables and observations are being presented on the same chart. An example of such a technique is the biplot [2].

M-Analysis

Finally, a fourth technique, *M-analysis*, is used on a data matrix which has been corrected for both its column and row means (double-centering). This technique has been widely used for the two-way **analysis of variance** where there is no error term other than that included in the **interaction** term. The interaction sum of squares may be obtained directly from double-centered data. M-analysis is then employed on these data to detect instances of nonadditivity and/or obtain a better estimate of the true inherent variability. (See [5] for a summary of these techniques.) A version of M-analysis also used in multidimensional scaling is a method known as **principal coordinates** [3, 7, 8].

References

[1] Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank, *Psychometrika* **1**, 211–218.

[2] Gabriel, K.R. (1971). The biplot-graphic display of matrices with application to principal component analysis, *Biometrika* **58**, 453–467.

[3] Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**, 325–338.

[4] Householder, A.S. & Young, G. (1938). Matrix approximations and latent roots, *American Mathematical Monthly* **45**, 165–171.

[5] Jackson, J.E. (1991). *A User's Guide to Principal Components*. Wiley, New York.

[6] Okamoto, M. (1972). Four techniques of principal component analysis, *Journal of the Japanese Statistical Society* **2**, 63–69.

[7] Torgerson, W.S. (1952). Multidimensional scaling I: theory and method, *Psychometrika* **17**, 401–419.

[8] Torgerson, W.S. (1958). *Theory and Methods of Scaling*. Wiley, New York.

(*See also* **Cluster Analysis, Variables**)

J. EDWARD JACKSON

R

The R system is an independent, open source, freely available implementation of a dialect of the S language. The S language [1, 2] is the basis for the commercial **S-PLUS** software system for data analysis and graphics. Both R and S-PLUS are described and compared in [12].

Construction of the R system has been a major statistical computing research project. Ihaka & Gentleman [7], both at that time working at the University of Auckland, developed the initial version. The ongoing development of the R system is now coordinated by an international team of expert volunteers known as "R core" but much of the work is done by very many contributors of code from many countries around the world.

The R program operates by assembling collections of functions and data sets for use at any one time. These collections are known as "packages" in R, (although the analogous concept is known as a "library" in S-PLUS). R is particularly important for developing and promulgating new statistical, data analytical, or graphical techniques; this is mostly done by releasing a new package with a suite of functions for implementing the techniques in question.

The R system has, as key features:

- an interpreted, object oriented programming language with a C-like syntax (*see* **Computer Languages and Programs**),
- a command-line input, supplemented by a system of menus that is mainly useful for file operations, for accessing the internet or for finding and reading help pages,
- code that is open source, that is, freely available for scrutiny or modification,
- support for a wide variety of static and dynamic color graphics capabilities, making it easy to produce publication quality graphical output that can include mathematical text (*see* **Graphical Displays**),
- a close integration of data analysis and graphics,
- a base package that handles commonly required types of computations, and which is a basis for adding further packages,
- a large and rapidly growing range of packages, including some standard ones for "lattice" graphics and the classes and methods of S release 4 [3],
- various mechanisms, some offered via additional packages, for interfacing to other software systems – local **database systems**, networked database systems, other interpreted languages (e.g. Python), Geographical Information Systems GIS systems, and so on.

Experimental forms of graphical user interface (GUI) are available for limited component parts of R. There are several initiatives that are aimed at developing a GUI system or systems that might be used for the base package and perhaps for other packages that are supplied as part of the base installation. More details on this and other aspects of R can be found at the URL `http://cran.r-project.org`.

In recent years, there has been a major release of R at about six-monthly intervals. Because of the somewhat experimental and rapidly changing nature of the system, it has been impossible to ensure complete backwards compatibility. Package updates happen often as well and old packages may be updated to new versions or new packages installed from within a session of R itself, which is a particularly strong feature. In principle, anyone with access to the **internet** may have the latest version of R and all of its listed packages at any time.

The Sweave package is an interesting example of the linking of the abilities of two different software systems, an idea first proposed by Donald Knuth [8]. It allows the interleaving of code and text in a LaTeX document, which can then be processed into a final document that can include tables, graphics, computer output, and associated code. Changes in the code are immediately reflected in the output document, making it straightforward to ensure that published analyses, and modifications to those analyses, are reproducible.

The R system has close points of contact with the more experimental and specifically research-directed aims of the Omegahat project [9]. These connections also result in the development of code that is of direct use to R users as well as to Omegahat. More details on Omegahat can be found at the URL `http://www.omegahat.org`

The Bioconductor project [6], which is a major cooperative effort between professional statisticians who work with microarray data, is the most substantial of a number of R-based initiatives. For many statisticians working with microarray data, R is now

a very commonly used platform. Related computing challenges, which the Bioconductor packages address, include the development of data structures that facilitate work with microarray data, linkages into associated annotation information, and local and internet-based access to databases.

Official documentation for R includes an introductory manual, an R language manual, a guide to importing and exporting data from R, a manual on extensions to R, a guide to the installation and administration of R, and a document that gives answers to frequently asked questions. For the S language on which R is based, see [1, 2, 12]. Dalgaard [4] is a careful introductory text, both for the language and for using R for elementary statistical analysis; see also [14]. Maindonald & Braun [10] (intermediate level) and Venables & Ripley [13] (advanced level) both use examples as a basis for their exposition; see also Fox [5]. The mechanisms used by R for including mathematical annotation on graphs is discussed in [11].

References

[1] Becker, R.A. & Chambers, R.M. (1984). *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth, Belmont.

[2] Becker, R.A., Chambers, J.M. & Wilks, A.R. (1988). *The New S Language: Programming Environment for Data Analysis and Graphics*. Wadsworth, Belmont.

[3] Chambers, J.M. (1998). *Programming with Data: A Guide to the S Language*. Springer-Verlag, New York.

[4] Dalgaard, P. (2002). *Introductory Statistics with R*. Springer-Verlag, New York.

[5] Fox, J. (2002). *An R and S-PLUS Companion to Applied Regression*. Sage Books.

[6] Gentleman, R. & Carey, V. (2002). Bioconductor. Open source bioinformatics using R, *R News* **2**(1), 11–17. `http://cran.R-project.org/doc/Rnews`.

[7] Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**, 299–314.

[8] Knuth, D.E. (1992). *Literate Programming*. Center for Study of Language and Information, Stanford, California.

[9] Lang, D.T. (2000). The Omegahat environment: new possibilities for statistical computing, *Journal of Computational and Graphical Statistics* **9**, 423–451.

[10] Maindonald, J.H. & Braun, J.B. (2003). *Data Analysis & Graphics Using R. An Example-Based Approach*. Cambridge University Press, Cambridge.

[11] Murrell, P. & Ihaka, P. (2000). An approach to providing mathematical annotation in plots, *Journal of Computational and Graphical Statistics* **9**, 582–599.

[12] Venables, W.N. & Ripley, B.D. (2000). *S Programming*. Springer-Verlag, New York.

[13] Venables, W.N. & Ripley, B.D., (2002). *Modern Applied Statistics with S*, 4^{th} Ed. Springer-Verlag, New York.

[14] Venables, W.N. & Smith, D.M. (2002). *An Introduction to R*. Network Theory Ltd.

(*See also* **Software, Biostatistical**)

W.N. VENABLES & JOHN H. MAINDONALD

Radiation Epidemiology

Introduction

Radiation epidemiology is a specialized field of **epidemiology**, which seeks to characterize and quantify the health risks associated with exposure to ionizing and nonionizing **radiation**. One of the principal aims of radiation epidemiology is to provide the human data needed to recommend or set protection standards for workers and the general public. The data are also used to estimate levels of **risk** from diagnostic radiation procedures, to indicate how radiotherapy protocols can be improved to reduce acute and long-term side effects, to better understand individual susceptibility, and to learn more about disease mechanisms.

Radiation is a general term that includes both ionizing and nonionizing radiation. The many different types of radiation have a range of energy forming an electromagnetic spectrum. Ionizing radiation is located in the high frequency region of the electromagnetic spectrum. Nonionizing radiation includes optical radiation and electromagnetic fields, as well as acoustic fields. The nonionizing region of the spectrum has a wide range of frequencies and wavelengths. At the lowest frequencies there are static fields, for example, those used for magnetic resonance imaging (MRI). Further along the spectrum, nonionizing radiation includes time-varying electric and magnetic fields produced by power lines, electrical appliances, radiofrequencies, and microwaves, as well as radiation wavelengths in the infrared, visible, and ultraviolet ranges.

This article is concerned with the many types of ionizing radiation, which are emitted with different energies and penetration abilities. Ionizing radiation produces electrically charged particles (ions) that have enough energy to break chemical bonds and that can cause a range of biological damage. Exposure to ionizing radiation can come from external sources, such as X- or γ-rays, or from internal sources emanating from radioactive materials deposited through inhalation, ingestion, or rarely through dermal absorption. In radiation epidemiology, we are most concerned with X- and γ-rays, neutrons, and α-particles and other radionuclides because they are used in medicine, occur as a result of the nuclear bombings in Hiroshima and Nagasaki, are important components of fallout from nuclear testing or accidents, or occur in nature. Individuals or populations can receive one or more acute exposures to radiation or they can be exposed to protracted radiation over a long period of time, such as from occupational exposure. Unlike most other carcinogens, about 85% of radiation exposure to the general public comes from natural sources, largely radon, but also cosmic radiation and natural background radiation. The remaining public exposure comes from medical radiation (about 14%), with only about 1% coming from fallout, occupational exposure, radioactive discharges, and consumer products combined [117]. People are exposed to radiation in a variety of settings. Large numbers of the general population receives many low-dose radiologic examinations over their lifetime; whereas a relatively small percent of the population receives high-dose radiotherapy as treatment for cancer or a few benign diseases, for example, hyperthyroidism.

Radiation can cause a variety of acute and long-term biological effects at the molecular, cellular, tissue, and organ levels. Deleterious health effects occur when a sufficiently large number of cells die or are damaged so that tissue or organ function is no longer adequate, or when cells lose their proliferative capacity, transform or mutate. The important biologic consequences of radiation are thought to result primarily from direct damage to DNA [58, 117]. Recent experimental studies, however, suggest that radiation might also cause damage to the cytoplasm [121], and that radiation effects can occur in cells that are not directly exposed to radiation through bystander effects [7, 78] or genetic instability [49, 76]. Cataracts [85], neurological abnormalities [122], thyroid diseases [79, 99], cardiovascular diseases [102], benign and malignant tumors [113], and birth defects [123] are some of the more important long-term health effects associated with radiation exposure. Since radiation-related cancer has been studied in greatest detail and is of most concern to the public, it will be our major focus.

Wilhelm Röntgen discovered X-rays in 1895. They were used in medicine one year later to treat a nevus and the following year to treat cancer [34]. Owing to the very high early exposure of people working with radiation, it was recognized as a carcinogen very soon after it was discovered. Since then it has been studied intensively, and by 1920 it was already identified as the cause of

bone cancers occurring among young women who used radium to paint watch dials [25]. About two decades later, radiation was shown to increase the risk of leukemia among radiologists [69], and by the 1950s the first indication of radiation-induced malignancies was observed among the atomic bomb survivors [24]. Over the last fifty years, we have learned that most types of radiation can increase the risk of developing cancer, and that most cancers can be induced by radiation, but that there are differences in the magnitude and type of biologic effects manifested following equal doses of radiation. These differences are partly owing to the type of radiation and the duration of exposure, as well as host factors such as age, gender, and genetic susceptibility. In addition, some tissues and organs appear to be more radiosensitive than others.

The study of the long-term effects of radiation exposure began with astute observations of rare disease excesses among highly exposed individuals [20, 32]. As more was learned, larger studies of persons exposed to moderate doses of ionizing radiation were conducted and risks of disease were compared among exposed and nonexposed groups. When these studies unequivocally demonstrated that radiation was a carcinogen [81], the field moved toward more quantitative methods of describing risks. Because radiation exposure assessment (called dosimetry in radiation epidemiology) is detailed and relatively accurate compared with many other exposures, the field of radiation epidemiology has progressed to more precise quantification of risks, evaluation of the role of factors that might modify radiation risks, detecting risks at low doses, and characterizing the shape of the **dose–response** function. In parallel, the fields of radiation biology and biophysics were advancing and many relevant findings from experimental work were integrated into radiation epidemiology approaches [9]. The strong interactions between these disciplines has led to a general awareness of the importance of biological mechanisms in analyzing epidemiology results, and to a subfield of biological or mechanistic modeling of epidemiologic data [83, 84, 117]. Currently, radiation epidemiology employs sophisticated statistical methods for **risk assessment** and uncertainty analyses and has increasingly embraced a multidisciplinary approach to better understand how the physical aspects of radiation influence the biological effects observed in humans.

Major Issues in Radiation Epidemiology

Although there is a clear consensus that moderate to high radiation doses cause harmful effects in humans, a central issue in radiation today is quantifying the biological effects at low doses, that is, at doses below about 0.1 Gy (see "Radiation Measurements" section). Other important questions in radiation epidemiology include: how much cancer is caused by radiation; what is the shape of the dose–response relationship; does the increased risk associated with radiation exposure persist throughout life; how do age and gender modify effects (*see* **Effect Modification**); how does dose rate influence risk; how should known radiation risks from one population group be transferred to another population group; and to what extent do uncertainties in dose assessment influence dose–response analysis and inference on effect modifiers. As collaborations with radiobiologists, dosimetrists, and biophysicists expand, more intriguing questions are becoming the focus of attention. These include: how does radiation cause cancer; why do organs and tissues vary in sensitivity; what causes differences in individual susceptibility to radiation; and how does radiation interact with other disease-causing agents.

Types of Studies Used in Radiation Epidemiology

Epidemiology includes both observational and experimental methods, however, as in other subdisciplines of epidemiology, radiation epidemiology primarily relies on **observational studies**. Observational investigations use both **descriptive** and **analytic** study designs. In general, descriptive studies are used to generate hypotheses and analytic studies are used to test hypotheses.

Descriptive Epidemiology

Descriptive studies usually use publicly available data to describe a temporal or geographic pattern or trend in a population group. While useful observations have been made based on descriptive studies, the aggregated nature of the exposure and disease data, and the frequent lack of **confounding** information are serious drawbacks. Descriptive studies are sensitive to **biases** and can produce misleading results. Examples

of descriptive radiation studies include evaluations of cancer rates in geographic areas near nuclear power facilities compared with rates from similar areas not in proximity of a nuclear plant [47], and thyroid cancer rates in West European countries following the disastrous Chernobyl accident compared with rates before the accident [13]. Despite the rather inconsistent results from the various studies, some investigators have interpreted the positive findings as indicating an association between radiation exposure and cancer. Unfortunately, the lack of individual doses and information on potential confounding precludes concluding a cause and effect relationship between disease and exposure (*see* **Causation**).

A specific type of descriptive study, called **ecologic** (also known as correlational or geographic), has been used to a limited extent in radiation epidemiology. In this design, aggregate rates of disease are regressed on characteristics of population groups, but it is not known whether the people who are exposed are those who develop the health outcome of interest. Furthermore, because **covariate** data are limited, it is difficult to control for confounding. When the population groups are large, the statistical precision and stability is high, but when a strong confounding factor exists on the individual level, which cannot be adjusted for on the aggregate level adequately, the results can be erroneous [33, 77, 86]. (*see* **Ecologic Study** and **Ecologic Fallacy**.) A continuing controversy in radiation epidemiology concerns the role of residential radon in lung cancer etiology. Using an ecological approach, Cohen [12] aggregated lung cancer mortality rates with mean levels of radon exposure in 1600 counties in the United States. Using a surrogate smoking index to adjust for potential confounding from smoking, he reported a protective effect between lung cancer mortality and radon concentration levels indicating that radon is beneficial. Other investigators, however, have observed that a small correlation between smoking and radon concentration levels can sufficiently confound results so that a negative association is observed when the true association is positive [63]. In addition, Puskin [95] found that a variety of other smoking-related cancers were negatively correlated with radon, associations which were highly implausible, and concluded that the protective effect of radon on lung cancer was due to incomplete control of smoking. Moreover, analytic epidemiologic studies consistently show a small excess risk of lung cancer associated with residential radon.

Analytic Epidemiology

In analytic studies, such as **case–control** and **cohort** studies, the unit of observation is the individual, that is, information on outcome, exposure, and covariates is collected for individuals. In case–control studies, individuals with the disease of interest (cases) are drawn from a defined population and their exposure history is compared with appropriate nondiseased **controls** selected from a similar population with the same potential for radiation exposure. Questionnaires are most often used to collect data on exposure and covariates. Case–control studies are an efficient and cost-effective method for studying relatively rare diseases; they are, however, subject to **selection** and **recall** bias.

Studies of the carcinogenic risks associated with diagnostic radiography during early childhood have generally employed a case–control design [43, 70, 103, 104]. These studies are fraught with difficulties for a variety of reasons. Doses are generally low, and estimating individual organ doses is complicated because records of diagnostic examinations usually are not available. Dosimetry for diagnostic radiographs generally relies on patient and/or parent recall of the number, type, and date of exams and, therefore, is subject to the problem that cases and controls might remember or report their history of diagnostic examinations with different levels of accuracy. The potential for recall bias that is, cancer patients may remember their medical history better than a control who does not have a serious illness, can result in risk estimates that are artificially high. Nevertheless, despite exposure assessment weaknesses, case–control studies have been useful in highlighting the potential dangers of diagnostic X rays.

Cohort studies of selected populations exposed to radiation have been the mainstay of radiation epidemiology. In this method, exposed and nonexposed individuals are followed over time until a sufficient number of study subjects develop a particular disease outcome. A major advantage of prospective cohort studies or retrospective cohort studies (*see* **Cohort Study, Historical**) with good exposure records is that radiation exposure can be ascertained without relying on the memory of the study subjects. Fifty years after the Hiroshima and Nagasaki atomic bombings,

the Life Span Study (LSS) cohort of atomic bomb survivors continues to be the most informative study of the carcinogenic effects of radiation exposure in humans. It is the major source of epidemiologic data used for radiation risk assessment and has had the greatest impact on the development of radiation protection standards [83, 117]. Decades of dosimetry efforts, including periodic improvement in the dosimetry system, have resulted in well-characterized individual doses, for most organs and tissues, for close to 90 000 survivors in the cohort [97]. The good dosimetry and the relatively complete and unbiased cancer incidence and mortality ascertainment have helped make the LSS the "gold standard" in radiation epidemiology. Over time, it has become apparent that the LSS can provide valuable information on the nature of radiation-associated cancer risk following low doses, because about 35 000 survivors were exposed to doses between 5 and 200 mSv [89]. The latest cancer **incidence** and mortality reports demonstrate that the elevated cancer risk continues throughout life and that the shape of the dose response for solid cancers is linear with no evidence of a radiation dose threshold (*see* **Extrapolation, Low Dose**) below which there is no excess risk [89, 90, 113]. Indeed, even within the narrow low-dose range, cancer risks are well described by a simple linear dose response (*see* **Linear Regression, Simple**), and the best estimate of a threshold dose for cancer incidence, if it exists, is 0 mSv, with a 95% upper confidence bound (*see* **Confidence Intervals and Sets**) of about 60 mSv [89]. Cancer risks in the LSS are described in terms of both **excess relative risk** (ERR) and excess **absolute risk** (EAR) models. Using an ERR model, the risk per Sievert is higher for women than men and decreases with increasing age at exposure or attained age. The EAR per 10 000 person years per Sv is also higher among women, but although it tends to decrease with increasing age at exposure, the EAR has increased throughout the study period as the cohort ages. Evaluation of patterns of organ (or site) specific risks suggests that, with few exceptions, the excess risks for most solid cancers do not deviate significantly from the overall solid cancer risk estimate [88]. In the next 20 to 30 years, the number of cancer cases diagnosed in the LSS will nearly double. With the additional cases, the LSS will prove to be an important resource for evaluating **interactions** between radiation and other environmental or genetic risk factors.

Nested case–control (*see* **Case–Control Study, Nested**) and **case–cohort studies** are special types of cohort studies that are used in radiation epidemiology when detailed dosimetric data are needed, but are difficult or expensive to collect for the entire cohort. Nested case–control studies have been used very effectively to evaluate radiation treatment for a first primary cancer in the development of a second cancer. Detailed radiotherapy information is needed to estimate organ doses and chemotherapy and other data are needed to assess confounding and modifying effects, but it would be prohibitively expensive and time consuming to collect and abstract all of this information for a cohort large enough to have an adequate number of second cancers. In a population-based cohort of 19 046 patients with Hodgkin's disease, a nested case–control study was conducted to quantify the treatment-related risk of developing a second cancer of the lung [31, 115]. On the basis of the daily radiotherapy logs for each patient, dose to the subsite of the lung where cancer developed was estimated for 222 cases who developed a second lung cancer. Radiation doses to the same part of the lung as the case was estimated for the two matched controls for each case. Cumulative cytotoxic drug intake also was estimated for each study subject and information on smoking history was abstracted from patient records. This intense data collection effort allowed fairly precise radiation dose estimation and control for confounding. Lung cancer risk was found to increase significantly with radiation dose up to 40 or more Gy and also with increasing amounts of alkylating agents in an additive fashion, and tobacco appeared to multiply radiation risks.

A more recent development in radiation epidemiology is the use of meta and pooled analyses (thyroid, breast, radon, nuclear workers) to combine either published (meta analysis) or original (pooled analysis) data from several studies. Combined analyses, especially of rare diseases or exposures, can increase statistical **power** substantially [10, 14, 55, 64, 94, 99]. They are useful for detecting small risks, improving precision of risk estimates, formally comparing data from several studies with the goal of identifying consistent patterns of risk, and systematically assessing modifying factors and subgroup risks that could not be evaluated adequately in single studies. The thyroid gland of children is one of the more radiosensitive organs in humans [98], and therefore is of special interest in radiation epidemiology. But because

thyroid cancer is rare the number of cases in any individual study is likely to be small. A pooled analysis of seven studies, which included almost 120 000 people (about 58 000 exposed to radiation) and 700 thyroid cancers, demonstrated the linearity of the radiation dose response and the importance of age at exposure as a modifying effect, that is, people exposed during childhood had a very high excess relative risk, but for those exposed as adults there appeared to be little risk [99]. The ERR per Gy for people exposed before age 15 was 7.7 (95% $CI = 2.1; 28.7$), whereas for atomic bomb survivors exposed as adults it was 0.4 (95% $CI = -0.1; 1.2$). Following childhood exposure, the ERR remained elevated throughout the 40 year follow-up period, but appeared to decline somewhat after about 30 years. Linearity also described the dose response in a pooled analysis of 1502 breast cancers diagnosed among 77 527 women (almost half exposed) [94]. In this analysis, the age effects were more complicated, with age and age at exposure both being important in the excess absolute rate models, but only attained age in the excess relative risk models. This analysis also suggested that exposure at low-dose rates (protracted exposure) is less tumorigenic than acute or high-dose rate fractionated exposures. Pooled analyses of nuclear workers exposed to low levels of protracted radiation have provided more precise risk estimates than any of the individual studies [10]. While the number of excess cancers was still too small either to accept or to rule out the **null hypothesis**, the confidence bounds on the risk estimate indicated that the current occupational exposure standards are reasonable.

Radiation Measurements

Unlike many other environmental hazards, ionizing radiation can be measured with fairly good precision. The strength of a radioactive source is quantified by its activity, and the number of radioactive disintegrations per second. The unit of activity is the Becquerel (1 Bq = 1 disintegration per second). Absorbed dose deposited by ionizing radiation is quantified as the energy deposited per unit mass. The unit of absorbed dose is the Gray (1 Gy = 1 J/kg). Some types of ionizing radiation are more biologically effective (per unit dose) than others. To account for this, the quantity called equivalent dose is used, which is the absorbed dose multiplied by a radiation weighting factor. For example, the weighting factor is 20 for alpha particles. The unit of equivalent dose is the Sievert (1 Sv = 1 J/kg). To account for the effect of an inhomogeneous distribution of dose in the body, the effective dose is used. Effective dose is the sum over specified tissues of the product of the equivalent dose to the tissue, and a weighting factor for that tissue – this latter being a measure of its relative radiosensitivity. The unit of effective dose is also the Sievert (Sv). The SI units described above have replaced traditional units: 1 Bq = 2.7×10^{-11} Curie (Ci); 1 Gy = 100 rad; 1 Sv = 100 rem [44]. The concentration of radon in air is measured in Bq/m^3 (replaces the historical unit of pico Curies per liter). In studies of underground miners, exposure to alpha particles from radon (Rn-222) and its short-lived decay products, Po-218 and Po-214, is measured in working level months (WLM). One working level (WL) is any combination of the short-lived progeny of radon in 1 liter of air, under ambient temperature and pressure, that results in the ultimate emission of 1.3×10^5 MeV of alpha particle energy (1 WL = 2.08×10^{-5} Jm^{-3}). One WLM is a cumulative exposure equivalent to 1 WL for a working month of 170 hours (1 WLM = 3.5×10^{-3} Jhm^{-3}).

Retrospective Exposure Assessment

The crucial need to measure exposure from radiation carefully was clear very soon after the discovery of X rays. This need resulted in the development of the field of radiation dosimetry. The field has advanced substantially since then [51], so that today radiation dosimetry is probably the most sophisticated of epidemiologic exposure assessment methods. To quantify long-term radiation-related health risks adequately for populations exposed years ago, retrospective dose reconstruction is often performed in concert with radiation epidemiology. For epidemiologic studies, the ideal measure of exposure is the absorbed dose to the organ or tissue of interest for each study subject. Three basic methods are used in radiation dosimetry: physical measurements made in a laboratory or the environment; analytic model-based dose reconstruction, and biodosimetry. These methods have been applied to both external and internal sources of exposure. The quality of dose reconstruction will depend on the complexity of the exposure situation, the exposed population, the

quality and quantity of data relevant to the exposure, and the availability of physical and biological samples [45]. Identifying the sources of uncertainty in dose estimation is becoming more common in radiation epidemiology. With better understanding of the type and amount of these uncertainties, the influence of dose errors can be accounted for in the statistical analysis of radiation effects. This issue is discussed in detail in the section "Adjusting for errors in dosimetry".

Dosimetry for external radiotherapy largely relies on laboratory experiments, using an anthropomorphic phantom, or a water phantom in combination with a mathematical phantom, that simulate the actual exposure conditions recorded in a patient's medical record. The laboratory measurements along with computational models have been used to estimate absorbed doses and their uncertainties, to organs inside and outside the radiation field, for individual study subjects [108]. When detailed treatment parameters are known, individual organ dose estimates are fairly accurate. When treatment records are incomplete or not available, the dose estimates will have a large degree of uncertainty. Dosimetry for internal medical exposure is more complicated and, therefore, the level of uncertainty is much higher [51].

Dose estimation for past environmental exposures is difficult. The best environmental dose reconstruction is that for the Life Span Study of atomic bomb survivors in Japan. This major dosimetry program to estimate individual organ doses for most survivors is largely based on physical measurements from building materials found in the Hiroshima and Nagasaki surroundings, investigation of shielding factors made during mock-up tests in Nevada, and analytic modeling based on many variables, including what is known about the physical qualities and yields of the bombs, weather conditions at the time of the bombings, the location and shielding conditions of survivors at the time of the bombings, and the attenuation from different shielding configurations [97]. Several methods of biological dosimetry have been used to estimate doses in subgroups of the survivors [1, 6, 80, 106]. In general, there has been fair agreement with model-based dose estimates. Since the bombings, several investigators have assessed dosimetry errors in the LSS of atomic bomb survivors [27, 46, 91]. It has been estimated that there is an error of about $\pm 30\%$ in individual organ doses and that the error results in a 4 to 11% underestimate of the dose response [91]. To account for the error, an adjustment to the doses is made in the analyses.

Elaborate dose reconstruction projects have been conducted to estimate doses from fallout in the United States [8, 82, 114] and the former Soviet Union [105], from nuclear accidents [26, 57, 117], and from nuclear facility discharges [18, 19, 22]. In these situations, past environmental measurements and current measurements of soil and brick are often used in conjunction with computational modeling, especially for external exposure. Estimating radiation dose received as a consequence of intake of contaminated air, water, and food is more difficult because it not only requires knowing what radiation was released, but also past dietary intake for each individual. When possible, biodosimetric measurements are also performed. Combining several methods of exposure assessment can improve accuracy and reliability.

Studies of radiation workers provide an important source of information on the effects of low-dose protracted exposure. Occupational exposure assessment is based primarily on dosimeters worn by individual workers. The dosimeters provide a measurement of cumulative external exposure, but accurate dose estimates to an individual worker are dependent on a wide array of conditions, including the quality of the dosimeter, the precision in recording the measurements, the care in which the dosimeter is worn, and the uncertainties in the conversion from exposure to organ dose [10, 29, 118]. As in other situations, it is more complicated to estimate dose from internal occupational exposure [53].

Biodosimetry methods increasingly have been incorporated into epidemiologic studies to validate dose estimates derived from physical or model-based dosimetry methods or to provide dose estimates when other sources of dose data are missing or incomplete [116]. For many years, chromosome aberrations in lymphocytes have been evaluated as a measure of dose. For past radiation exposure, translocations can be assessed using fluorescence *in situ* hybridization (FISH) techniques. This method is one of the most accurate and sensitive of the biodosimetry methods, but is labor intensive and expensive [21, 48, 116]. More recently, the electron paramagnetic resonance (EPR) technique has been used to measure radiation dose accumulated in tooth enamel [11, 80]. New biodosimetry approaches currently are being developed.

Models for Disease Risk from Radiation Exposure

In radiation epidemiology, three general modeling approaches are used to analyze data and assess risk: the dosimetric approach; the empirical or descriptive approach; and the biologically motivated approach [84, 117]. (*see* **Model, Choice of**)

Dosimetric Risk Models

In the dosimetric approach, a model for disease risk is developed in studies with accurate disease outcome information and well-characterized dosimetry, then applied to a population with a possibly different radiation exposure experience. The dosimetric approach often requires knowledge of the relative consequences on disease outcomes of exposure to different types of radiation (specified by the relative biological effectiveness, RBE) and under different exposure regimes. For example, the application of risk models derived from the Japanese LSS of atomic bomb survivors to radon-exposed underground miners requires knowledge of differential effects of acute exposure to gamma radiation and neutrons relative to chronic exposure to alpha radiation. This process is complicated by evidence that RBE depends on dose level [71].

While the dosimetric approach has been applied historically in radiation risk assessment, it is no longer widely used because of the inherent difficulties of incorporating the wide range of exposure conditions encountered with diagnostic and therapeutic radiations and with natural radiations such as radon and cosmic rays. These diverse exposure conditions include very low and very high doses, direct and inverse dose rate effects (dose–response slope varying as a function of dose), and acute, chronic or highly fractionated doses. In addition, the need to apply risk models to populations with very different radiation exposures has diminished as more and larger populations with diverse radiation exposures are studied. Investigators, however, still conduct comparisons of risk models and risk estimates across diverse populations and types of radiation to gain insights into mechanisms of radiation effects.

Empirical or Descriptive Risk Models

The most common modeling approach with epidemiologic data uses empirical models. With the empirical approach, modeling starts with a relatively simple structure and few *a priori* assumptions. Models develop increasing complexity by including factors that statistically improve model fit, although added covariates typically require some level of biological plausibility. The approach is amenable to a broad-based exploration of diverse exposures and of factors that modify the exposure–response relationship. Although empirical models are less structured than biologically motivated models, these models still require assumptions on the basic shape of the dose–response relationship and on the functional form of factors that modify that relationship [84]. Although empirical models can be developed distinct from specific biologic assumptions, models are flexible enough to accommodate biologically hypotheses.

Data are typically analyzed under an ERR model,

$$\lambda(x_{\mathrm{bk}})[1 + \rho(d) \in (x_{\mathrm{em}})]$$

or EAR model,

$$\lambda(x_{\mathrm{bk}}) + \rho(d) \in^{*} (x_{\mathrm{em}})$$

where d is the radiation dose, x_{bk} and x_{em} are vectors of covariates used to specify the background rate of disease and effect modification, respectively, $\lambda(\bullet)$ is the background rate of disease in non-radiation-exposed individuals, $\rho(\bullet)$ is the dose–response function, and $\in(\bullet)$ and $\in^{*}(\bullet)$ are functions describing effect modification of the ERR and EAR, respectively.

In modeling radiation effects, the dose–response relationship, at least for cancer, is based on the radiobiological assumption that lesions are initiated as a result of one or two ionizing events, resulting in a molecular cascade of effects leading to malignancy, or to programmed cell death (apoptosis) [117]. With this biological interpretation as a framework, a general dose–response model is specified by the linear-quadratic-exponential form:

$$\rho(d) = (\beta_0 + \beta_1 d + \beta_2 d^2)\mathrm{e}^{-\alpha_1 d - \alpha_2 d^2} \qquad (1)$$

where the linear and quadratic terms in dose correspond to one or two ionizing events, and the exponential factor corresponds to cell death at high radiation doses. While this model is useful as a conceptual starting point for analysis, data are seldom sufficient in numbers of cases and ranges of doses to enable simultaneous estimation of all parameters (*see* **Estimation**).

In the Japanese LSS cohort, analysis is conducted under an assumption of piecewise exponential time to failure (*see* **Grouped Survival Times**), which is often referred to as **Poisson regression** due to the form of the **likelihood** function. ERR models are used to model disease rates for solid tumors, while EAR models are applied with leukemia, with dose given in Sv [83, 89, 93]. For these data, x_{bk} includes variables such as attained age, city, sex, calendar year of follow-up, and possibly other factors. For ERR models, the background disease rate, $\lambda(x_{\text{bk}})$, is often modeled **semi-parametrically** by categorizing the components of x_{bk} and jointly stratifying on all levels. Effect modifiers, x_{em}, may include variables used in the background as well as age at exposure and time since exposure. For solid tumors, the preferred risk model for ρ (d) is linear in dose, that is, only β_1 is nonzero,

$$\rho(d) = \beta_1 d$$

where β_1 is the excess relative risk per Sv (ERR/Sv). A variant of this model has been used to examine threshold relationships, where a dose below a given level is assumed to have no effect on disease rate, by setting $d' = d - d_o$ for $d > d_o$ and zero otherwise [41, 42, 62]. However, there has been no convincing evidence for the existence of such a threshold in the LSS cohort [89]. Variables to evaluated effect modification, $\in (x_{\text{em}})$, may be continuous or categorical (*see* **Random Variable**). The preferred risk model for the ERR of respiratory cancer is

$$\begin{aligned} \rho(d) &= \beta_1 d \\ \in (t, s) &= \exp\left[\alpha_1 \ln\left(\frac{t}{20}\right) + \alpha_2 I(s)\right] \end{aligned} \qquad (2)$$

where t is years after exposure and $I(\bullet)$ an indicator function (*see* **Dummy Variables**) taking value zero for females ($s = 0$) and one for males ($s = 1$) [83].

The dose–response parameter is sometimes reparameterized as

$$\rho(d) = \mathrm{e}^{\beta_1^*} d \qquad (3)$$

to remove range restrictions on β_1. However, under this formulation, the estimated parameter is constrained to be nonnegative and **standard errors** are proportional on the original dose scale.

Radiation from the Hiroshima atomic bomb includes a mixture of gamma and neutron radiation, with a neutron to gamma ratio which decreased with distance from the hypocenter [89]. For Hiroshima atomic bomb survivors, total dose is defined deterministically as $d = g + RBE \times n$, where g and n are the gamma and neutron dose, respectively, and the RBE is set to 10 (although 20 has also been used). To address the issue of neutron effects at low doses [100], Pierce and Preston apply a dose–response model of the form:

$$\rho(g, n) = \beta(g + \theta g^2 + \phi n) \qquad (4)$$

where θ defines a quadratic effect for gamma dose and ϕ specifies RBE [89]. While in theory θ and ϕ can be estimated directly along with β, data are not sufficient and results are examined for a variety of values of θ and ϕ [90].

Leukemia is one of the most radiogenic cancers. Excess cancer risks have been observed within two years of exposure. In the LSS cohort that was assembled five years following the bombings, excess risks are at their maximum 5 to 10 years after exposure then decline, although risks remain elevated 50 years after the bombings [117]. The preferred model for leukemia risk is an EAR model of the form:

$$\begin{aligned} \rho(d) &= \beta_1 d + \beta_2 d^2 \\ \in^*(t, a_0) &= \begin{cases} \exp[\alpha_1 I(t \le 15) + \alpha_2 I(15 < t \le 25)] & \text{if } a_0 \le 20 \\ \exp[\alpha_3 I(t \le 25) + \alpha_4 I(25 < t \le 30)] & \text{if } a_0 > 20 \end{cases} \end{aligned} \qquad (5)$$

where a_{o} is age at exposure. Here $I(\bullet)$ is an indicator function that takes value 1 when the argument is true and 0 otherwise.

ERR models are also applied to data from radon-exposed underground miners [66, 84]. Exposure–response has the form:

$$\rho(d) = [\beta(\theta_0 d_{5-14} + \theta_1 d_{15-29} + \theta_2 d_{30+})] \qquad (6)$$

where cumulative exposure, d, in units of Working Level Months (WLM) is redefined as "effective" cumulative exposure by weighting cumulative WLM incurred 5 to 14, 15 to 29, and 30 years or more prior to current age. The weights, θ_i, are estimated in the model fitting, with θ_0 fixed at one for **identifiability**. Exposure effects on lung cancer mortality in miners diminish over 50% after 25 years since exposure. Models for effect modification use indicator variables

and take the form

$$\in (a, z) = g_a(\alpha_1) g_z(\alpha_2) \tag{7}$$

with

$$\begin{aligned} g_a(\alpha_1) = \exp[&\alpha_{1,1} I(a < 55) + \alpha_{1,2} I(55 < a \leq 65) \\ &+ \alpha_{1,3} I(65 < a \leq 75) + \alpha_{1,4} I(75 < a)] \\ g_z(\alpha_2) = \exp[&\alpha_{2,1} I(z < 5) + \alpha_{2,2} I(5 < z \leq 15) \\ &+ \alpha_{2,3} I(15 < z \leq 25) \\ &+ \alpha_{2,4} I(25 < z \leq 35) + \alpha_{2,5} I(35 < z)] \end{aligned}$$

Here the exposure–response relationship $\rho(d)$ varies with attained age, a, and years of exposure, z. The $\alpha_{2,j}$ parameters represent variation of the exposure–response relationship with duration for fixed cumulative exposure. Parameter estimates are positive and monotonically increasing, suggesting an enhancement of radon effects with longer durations (and lower rates) of exposure. An alternative model includes categories of exposure rate, z, as an effect modifier.

Latency is the interval of time from an increment of exposure to a change in disease response. (*see* **Latent Period**) A general method based on B-splines (*see* **Spline Function**) has been used to explore the exposure-time-response patterns in Colorado Plateau uranium miners exposed to radon and its decay products [36, 35]. For each subject, suppose exposure history, $x(t)$, is given for each year $t, t = 1, \ldots, T$, where the index for subject is dropped for clarity. Cumulative exposure is $\sum x(t)$. If $w(\bullet)$ defines a weighting function for the contribution to risk of each exposure increment, then $\sum w(t)x(t)$ is the "effective" exposure. The logarithm of the relative risk is given as

$$\beta_1 \sum w(t)x(t) + \beta_2 z$$

where z is a vector of additional risk factors. Cubic B-splines, which are continuously differentiable, piecewise polynomial functions [17], are used to model the weighting function $w(t;\theta)$ by defining

$$w(t, \theta) = \sum_{j=-3}^{m} \theta_j B_j(t) \tag{8}$$

where θ_j are unknown parameters, and $B_j(t)$ are known basis functions with knots at $0 < t_1 < \ldots < t_m$, and with $\sum B_j(t) = 1$ for all t. For the Colorado miner data, Hauptmann et al. show that exposures 5 to 25 years prior to current age have the greatest impact on the relative risk of lung cancer [35]. These results are similar to a previous analysis of the same data using a bilinear latency model, which requires more restrictive assumptions [56].

Biologically Motivated Risk Models

Biologically motivated models attempt to use results from experimental animal, cellular, and molecular studies, and a detailed understanding of radiobiology to specify a structural form for the carcinogenic relationship between radiation dose (and other disease risk factors) and cancer outcome [83, 84, 117]. A number of biologically motivated models have been developed and applied to data from specific radiation studies, including experimental animal studies and epidemiologic studies. Biologically motivated models have not yet been used to develop general risk models for risk assessment and radiation protection.

It has long been recognized that age-incidence curves for many epithelial cancers vary approximately as a power of age, suggesting a multistage process for carcinogenesis (*see* **Multistage Carcinogenesis Models**). The most widely applied multistage model for carcinogenesis is the Armitage–Doll model [3] (*see* **Dose–Response Models in Risk Analysis**). The model assumes that a cell undergoes k distinct, ordered, heritable changes, and that the background rate of disease is independent of the age of the at-risk cells. An exposure acting at a single stage, say the ith stage, is postulated to modify the stage transition rate. The importance of multistage models to the analysis of epidemiological and experimental animal data lies in the assumption that age patterns of disease occurrence relative to the timing of exposure contains important information about whether exposure operates early and/or late in the carcinogenic process [16]. For many adult cancers, k generally ranges from four to six, implying five to seven stages [117]. Application of the Armitage–Doll model to the atomic bomb survivor data suggests the presence of five stages with radiation acting both at an early and at a late stage [83, 111], although others conclude incidence of solid tumors is consistent with three stages [60].

To explain age patterns of solid tumors in the LSS data and more generally, the Armitage–Doll model has been extended to allow radiation exposure to act as a general mutagen that may affect a particular

stage or all stages in the carcinogenesis pathway [87, 92]. Age–time patterns under this model are in concordance with model predictions, although some skepticism has been raised about this approach [117].

The biological and mathematical characteristics of the two-stage clonal expansion model have been discussed in the general context of carcinogenesis [72–74]. The model postulates that a pool of normal stem cells are transformed at a given rate into premalignant initiated cells, either spontaneously or in response to a specific carcinogenic agent. These initiated cells undergo a birth and death process, which may be modulated by a nongenotoxic "promoting" agent, and undergo clonal expansion. Initiated cells in the premalignant clone may undergo further genomic damages that lead to malignant transformation, growth, and malignant tumor. The two-stage clonal expansion model has been applied to radon exposure data from animal experiments [39, 67] and to radon-exposed underground miners [37, 75]. Analyses based on the two-stage clonal expansion model suggest that radon may affect both the first transformation rate of normal to initiated cells and the rate of clonal growth of the initiated cells. An alternate analysis using a generalized multistage clonal expansion model suggests that a three-stage model may offer a better fit to the Colorado data than the two-stage model [61]. Clonal expansion models have also been applied to data from the LSS cohort [40, 50, 59]. The most recent analysis suggests that a variety of models are consistent with risk patterns and that data are too limited to discriminate among the various models [40].

Adjusting for Errors in Dosimetry

Random errors in dose estimates for atomic bomb survivors and their potential impact on dose–response evaluation have long been investigated [27, 46, 91]. Errors arise principally from uncertainties in the precise location of survivors and the type of shielding, information that is used as input to a multiparameter dose prediction model. This information is based on extensive interviews with subjects, in person or by mail, several years following exposure [27]. Pierce et al. apply a regression calibration approach for error adjustment [91] (*see* **Multiplicative Model**). For true dose x and observed dose z, the regression calibration replaces z with $\mathrm{E}[x|z]$, where $\mathrm{E}[\bullet]$ denotes expected value. The density function for the true dose given the observed $f(x|z)$, is proportional to

$$f(x|z) \propto f(x)f(z|x)$$

The marginal distribution of x, $f(x)$, is assumed **Weibull**, with parameter values selected such that there is agreement between the observed z's and the theoretically values induced from $f(x)$ and $f(z|x)$ under the assumed error model. Four error models for log(z) are considered: (i) "**lognormal** with 30% error", that is, lognormal with geometric standard distribution (GSD) (*see* **Geometric Distribution**) of exp(0.30), (ii) "lognormal with 40% error", (iii) "contaminated lognormal with 40% error", that is, a mixed lognormal having GSD = exp(0.30) with probability 0.75 and GSD = exp(0.75) with probability 0.15, and (iv) z having a normal distribution with coefficient of variation 0.40. Models (iii) and (iv) are selected to evaluate a "heavy tailed" distribution and a nonsymmetric (on a log scale) distribution, respectively. The authors conclude that use of the unadjusted observed doses results in risk estimates, which are about 4 to 11% too small [91]. Most current analyses of the LSS data use adjusted doses.

Nuclear workers are studied to obtain direct estimates of the effects of radiation exposures at low doses and low dose rates. Exposure assessment is aided by annual doses estimated from dosimeters worn by workers and exchanged weekly or bi-weekly. However, exposure evaluation of nuclear workers may be influenced by both **systematic** errors and random errors [28, 29, 30]. Gilbert and Fix describe potential sources of measurement error for workers at the Hanford nuclear facility [29]. Cumulative exposures may be biased from an inability to identify workers who are truly nonexposed and have zero recorded doses, from workers who are exposed to low doses but have zero recorded doses due to the adjustment of dosimeters, which subtracts a fixed contribution from natural background radiation. Sensitivity of dosimeters over a range of photon energies and the ability to respond to radiations from a variety of directions vary over time and result in temporal changes in errors since all doses below detection limits are recorded as zero. Improvements in dosimeters result in temporal variation in the measurement error for recorded doses, with increased errors in earlier years. Risk estimates are based on dose to target tissue, while exposure assessment for workers is based on recorded doses from dosimeters. In

addition, no precise data are available for workers exposed at other nuclear facilities. Finally, computerized files include annual dose estimates and not individual dosimeter values, further potentiating underestimation of cumulative dose. Assuming errors are **multiplicative** and lognormally distributed, and that errors from different sources are independent, Gilbert presents a detailed **sensitivity analysis** on the effects of random and systematic errors for the Hanford nuclear workers [28]. Adjustment for errors in the analysis of all cancers, except leukemia are nearly unchanged, while adjustment for errors in the analysis of leukemia, excluding chronic lymphocytic leukemia, increase the upper confidence limit by a factor of 1.4. Since sampling uncertainty is large, adjustments do not fundamentally modify inference.

A cohort of children from southwestern Utah, southeastern Nevada, and southeastern Arizona who were exposed to iodine isotopes (principally ^{131}I and ^{133}I) from fallout from aboveground detonations of nuclear weapons tests conducted from 1951 through 1958 at the Nevada Test Site, were examined for incidence of thyroid disease. A positive but nonsignificant dose response was found for thyroid cancer and for thyroid nodules [52]. Dosimetry for this study is based on environmental modeling of ground deposition of radioiodine for each weapons test, interviews with subjects' parents to obtain information on dietary habits, prior medical irradiation, medical history, and lifestyle factors. A **path analysis** links the dispersion of radiation in the environment, the uptake of radioiodines into the food supply, the intake of milk and leafy vegetables, direct inhalation, and external exposure from radiation in the passing fallout cloud to create a radiation exposure estimate for each subject [114]. A dose conversion factor is then applied to convert exposure from radioiodine intake and from external sources to thyroid dose. A multiplicative error model and lognormal errors are specified and used to define a range of GSDs for uncertainty and a "subjective confidence interval" for true dose [114]. More formally, under a classical measurement error model and using a Gibbs sampling approach (*see* **Markov Chain Monte Carlo**), measurement error correction of large uncertainties in thyroid dose results in a three-fold increase in the dose–response estimate [112]. A semiparametric **Bayesian** approach that allows for a mixture of both classical and Berkson errors (*see* **Measurement Error in Epidemiologic Studies**) acting through a latent intermediate variable (*see* **Path Analysis**) has also been applied to these data [68]. The error-adjusted dose–response estimate is nearly double the unadjusted estimate, but has markedly widened confidence limits.

Diagnostic and treatment protocols for medical radiation procedures are typically well-documented in medical records, which suggests the possibility of a good characterization of errors in dosimetry. Studies using anthropomorphic phantoms, where radiation dosimeters are embedded in constructs of tissue-like material and exposed under realistic conditions, provide data for the development and testing of **prediction** equations relating radiation exposure from X ray machines to dose to target tissues. As an example, uncertainties in the estimation of dose to the thyroid are evaluated for their effects on risk estimates of thyroid cancer in a study of Israeli children aged 15 years and under, who are treated with cranial irradiation for tinea capitis (ringworm of the scalp) [101]. Data from studies of phantoms are used to develop a prediction equation on the basis of age at treatment, X ray machine output, and amount of additional machine filtration, and including estimates of within-patient error, between-patient error, and random error. This prediction equation is used with covariate information from patient records to provide individualized dose estimates. Estimation of parameters in the prediction equation results in classical error, while the need to impute unknown patient data on age at treatment and other factors results in Berkson error. Schafer et al. use **Poisson regression** methods for cohort data and compare a full likelihood approach with a regression calibration approach [101]. After adjustment for multiple sources of error, risk estimates, and their standard errors, as well as inference on effect modifiers, are not appreciably altered with either approach. The results suggest that classical error plays a relatively minor role and Berkson error is limited. Since the components of dose uncertainties in the tinea capitis study are likely present in other epidemiologic studies of patients treated with radiation, this analysis may provide a model for considering the potential role of uncertainties.

Cohort mortality studies of underground miners are an important source of information on the effects of inhalation of radon and its decay products on lung cancer occurrence. Uncertainties in miner data arise from a variety of sources [84]. While cause of death may be **misspecified** either by failure to

record deaths or miscoding of cause of death, the major source of uncertainty derives from characterization of exposures many years in the past. Errors in exposure assessment arise from limited numbers of measurements of radon in air, particularly in the early years of mine operations, the need to interpolate radon concentrations between unmeasured mines and time periods, and the need to extrapolate radon to time periods prior to measurements. Exposure measurement devices are typically not distributed throughout the mine, since measurements are often motivated by regulatory concerns. Finally, documentation on numbers of hours and precise locations of workers in mines is frequently incomplete. A precise assessment of errors is crucial for the evaluation of the shape of the exposure–response function, particularly at low exposure rates, which can influence the **extrapolation** of risks to the lower exposures in the general population. Results in miners suggest a greater exposure–response relationship for workers with increased exposure duration (or decreased exposure rate), that is, for a given total exposure, exposures for longer durations at lower rates are more deleterious than exposures of shorter durations at higher rates [66]. Uncertainties in exposures are greatest in the early mining years when mine exposures are highest and before widespread measurements are available. Thus, periods of highest exposure rates are also periods most prone to exposure measurement errors. Stram et al. address uncertainties for the complex exposure assessment in the study of Colorado plateau uranium miners [109, 110]. They fit a **multilevel** regression of the logarithms of radon concentration on mine, location, region, and year-specific data. The model defines a single imputation, which is used to estimate exposure, that is, the expected true exposure under the model given the observed measurements, for unmeasured time periods and unmeasured mines. A nested case–control study is selected from the cohort data, matching 40 controls to each case, and analyzed using conditional likelihood regression methods. They find a steeper exposure–response slope and a diminished inverse exposure-rate effect.

Studies of radon-exposed underground miners indicate that exposures up to 30 years and earlier may influence risk of lung cancer [64, 84]. For epidemiologic studies of residential radon, exposure histories must therefore be reconstructed up to 30 years prior, creating [84] substantial uncertainties in exposure assessment [65]. Uncertainties arise from contemporary measurements of radon in one or two rooms not precisely characterizing levels within a dwelling or not accurately reflecting past radon levels due to modified living patterns of the current occupant, structural alterations of the residence, or normal random variations within a room, between rooms of a house, and from year to year. The use of contemporary radon measurements to estimate concentrations in past years induce classical error, which tends to bias risk estimates towards the null [15, 96] (*see* **Bias Toward the Null**). Uncertainties also arise from gaps in the historical record, due to residences that no longer exist or are used for nonresidential purposes, are located outside the study area, are occupied only briefly and excluded by the measurement protocol, or are not measured owing to refusal of the current occupant. Investigators typically employ a single imputation approach for gaps, inserting the mean radon concentration of control houses for missing data in the calculation of time weighted average radon exposure over an exposure period of interest [120]. This single imputation induces Berkson error, which increases variance, but under a rare disease assumption, a linear model for the odds ratio and limited missing data does not appreciably bias risk estimates (*see* **Missing Data in Epidemiologic Studies**).

Precise characterization of all uncertainties in residential radon studies is problematic. Investigators have addressed uncertainties by developing models and directly adjusting risk estimates [15, 96]. Other investigators address uncertainties using simulations under realistic error models to conduct sensitivity analyses, with estimates of error based on previous measurement experience [54], or a **variance components analysis** of validation data for temporal and spatial variation [119]. These disparate approaches have consistently found that error-adjusted estimates of excess odds ratios for lung cancer incidence and radon were 50 to 100% greater than unadjusted estimates.

Investigators have also sought to limit uncertainties through study design by enrolling only long-term residents [4, 5, 23] or by analytically modeling radon levels for gaps in residential history [38]. Finally, an improved retrospective radon dosimeter has been developed that measures residual radiation embedded in glass artifacts, such as mirrors and picture frames, from radon in air. The dosimeter measures past

cumulative exposures and thereby reduce exposure uncertainty [2, 54, 107].

References

[1] Akiyama, M., Kyoizumi, S., Hirai, Y., Hakoda, M., Nakamura, N. & Awa, A. (1990). Studies on chromosome aberrations and mutations in lymphocytes and GPA mutation in erythrocytes of atomic bomb survivors, in *Mutation and the Environment*, M.L. Mendelsohn & R.J. Albertini, eds. Wiley-Liss Inc., New York, pp. 69–80.

[2] Alavanja, M.C., Lubin, J.H., Mahaffey, J.A. & Brownson, R.C. (1999). Residential radon exposure and risk of lung cancer in Missouri, *American Journal of Public Health* **89**, 1042–1048.

[3] Armitage, P. & Doll, R. (1961). Stochastic models for carcinogenesis, in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 4, J. Neyman, ed. University of California Press, Berkeley and Los Angeles, pp. 19–38.

[4] Auvinen, A. (1996). Lung cancer risk from indoor radon [letter], *Lancet* **348**, 1662–1663.

[5] Auvinen, A., Makelainen, I., Hakama, M., Castren, O., Pukkala, E., Reisbacka, H. & Rytomaa, T. (1996). Indoor radon exposure and risk of lung cancer: a nested case-control study in Finland [published erratum appears in J. Natl. Cancer Inst. 1998;90:401–2], *Journal of the National Cancer Institute* **88**, 966–972.

[6] Awa, A.A., Sofuni, T., Honda, T., Itch, T., Neriishi, S. & Otake, K. (1978). Relationship between the radiation dose and chromosome aberrations in atomic bomb survivors of Hiroshima and Nagasaki, *Journal of Radiation Research* **19**, 126–140.

[7] Belyakov, O.V., Folkard, M., Mothersill, C., Prise, K.M. & Michael, B.D. (2003). A proliferation-dependent bystander effect in primary porcine and human urothelial explants in response to targeted irradiation, *British Journal of Cancer* **88**, 767–774.

[8] Bouville, A., Simon, S.L., Miller, C.W., Beck, H.L., Anspaugh, L.R. & Bennett, B.G. (2002). Estimates of doses from global fallout, *Health Physics* **82**, 690–705.

[9] Brenner, D.J., Lubin, J.H. & Ron, E. (1998). Moving from under the lamppost: can epidemiologists and radiobiologists work together? *Nuclear Energy-Journal of the British Nuclear Energy Society* **37**, 25–31.

[10] Cardis, E., Gilbert, E.S., Carpenter, L., Howe, G., Kato, I., Armstrong, B.K., Beral, V., Cowper, G., Douglas, A., Fix, J., Fry, S.A., Kaldor, J., Lave, C., Salmon, L., Smith, P.G., Voelz, G.L. & Wiggs, L.D. (1995). Effects of low-doses and low-dose rates of external ionizing-radiation – cancer mortality among nuclear industry workers in 3 countries, *Radiation Research* **142**, 117–132.

[11] Chumak, V., Sholom, S. & Pasalskaya, L. (1999). Application of high precision EPR dosimetry with teeth for reconstruction of doses to Chernobyl populations, *Radiation Protection Dosimetry* **84**, 515–520.

[12] Cohen, B.L. (1995). Test of the linear-no threshold theory of radiation carcinogenesis for inhaled radon decay products, *Health Physics* **68**, 157–174.

[13] Cotterill, S.J., Pearce, M.S. & Parker, L. (2001). Thyroid cancer in children and young adults in the North of England. Is increasing incidence related to the Chernobyl accident? *European Journal of Cancer* **37**, 1020–1026.

[14] Darby, S.C., Whitley, E., Howe, G.R., Hutchings, S.J., Kusiak, R.A., Lubin, J.H., Morrison, H.I., Tirmarche, M., Tomasek, L. & Radford, E.P. (1995). Radon and cancers other than lung cancer in underground miners: a collaborative analysis of 11 studies, *Journal of the National Cancer Institute* **87**, 378–384.

[15] Darby, S., Whitley, E., Silcocks, P., Thakrar, B., Green, M., Lomas, P., Miles, J., Reeves, G., Fearn, T. & Doll, R. (1998). Risk of lung cancer associated with residential radon exposure in Southwest England: a case-control study, *British Journal of Cancer* **78**, 394–408.

[16] Day, N.E. & Brown, C.C. (1980). Multistage models and primary prevention of cancer, *Journal of the National Cancer Institute* **64**, 977–989.

[17] de Boor, C. (1978). *A Practical Guide to Splines*, Applied Mathematical Science, Springer, New York.

[18] Degteva, M., Anspaugh, L., Napier, B. & Bell, R. (2002). Dose reconstruction validation and epidemiological studies for the Russian extended Techa River cohort, *Health Physics* **82**, S163–S164.

[19] Degteva, M.O., Vorobiova, M.I., Kozheurov, V.P., Tolstykh, E.I., Anspaugh, L.R. & Napier, B.A. (2000). Dose reconstruction system for the exposed population living along the Techa River, *Health Physics* **78**, 542–554.

[20] Duffy, B.J. Jr., Fitzgerald, P.J. (1950). Cancer of the thyroid in children: a report of 28 cases, *Cancer* **3**, 1018–1032.

[21] Edwards, A.A. (1997). The use of chromosomal aberrations in human lymphocytes for biological dosimetry, *Radiation Research* **148**, S39–S44.

[22] Farris, W.T., Napier, B.A., Ikenberry, T.A. & Shipler, D.B. (1996). Radiation doses from Hanford site releases to the atmosphere and the Columbia River, *Health Physics* **71**, 588–601.

[23] Field, R.W., Steck, D.J., Smith, B.J., Brus, C.P., Fisher, E.L., Neuberger, J.S., Platz, C.E., Robinson, R.A., Woolson, R.F. & Lynch, C.F. (2000). Residential radon gas exposure and lung cancer: the Iowa radon lung cancer study, *American Journal of Epidemiology* **151**, 1091–1102.

[24] Folley, J.H., Borges, W. & Yamawaki, T. (1952). Incidence of leukemia in survivors of atomic bomb in Hiroshima and Nagasaki, Japan, *American Journal of Medicine* **13**, 311–321.

[25] Fry, S.A. (1998). Studies of US radium dial workers: an epidemiological classic, *Radiation Research* **150**, S21–S29.

[26] Gavrilin, Y.I., Khrouch, V.T., Shinkarev, S.M., Krysenko, N.A., Skryabin, A.M., Bouville, A. & Anspaugh, L.R. (1999). Chernobyl accident: reconstruction of thyroid dose for inhabitants of the Republic of Belarus, *Health Physics* **76**, 105–119.

[27] Gilbert, E.S. (1984). Some effects of random dose measurement errors on analyses of atomic-bomb survivor data, *Radiation Research* **98**, 591–605.

[28] Gilbert, E.S. (1998). Accounting for errors in dose estimates used in studies of workers exposed to external radiation, *Health Physics* **74**, 22–29.

[29] Gilbert, E.S. & Fix, J.J. (1995). Accounting for bias in dose estimates in analyses of data from nuclear worker mortality studies, *Health Physics* **68**, 650–660.

[30] Gilbert, E.S., Fix, J.J. & Baumgartner, W.V. (1996). An approach to evaluating bias and uncertainty in estimates of external dose obtained from personal dosimeters, *Health Physics* **70**, 336–345.

[31] Gilbert, E.S., Stovall, M., Gospodarowicz, M., van Leeuwen, F.E., Andersson, M., Glimelius, B., Joensuu, T., Lynch, C.F., Curtis, R.E., Holowaty, E., Storm, H., Pukkala, E., van't Veer, M.B., Fraumeni, J.E., Boice, J.D., Clarke, E.A. & Travis, L.B. (2003). Lung cancer after treatment for Hodgkin's disease: focus on radiation effects, *Radiation Research* **159**, 161–173.

[32] Goldstein, L. & Murphy, D.P. (1929). Etiology of the ill-health in children born after maternal pelvic irradiation. II. Defective children born after postconception pelvic irradiation, *American Journal of Roentgenology* **22**, 322–331.

[33] Greenland, S. & Robins, J. (1994). Invited commentary: ecologic studies-biases, misconceptions, and counterexamples, *American Journal of Epidemiology* **139**, 747–760.

[34] Hall, E.J. (2000). *Radiobiology for the radiologist*. Lippincott Williams & Wilkins, Philadelphia.

[35] Hauptmann, M., Berhane, K., Langholz, B. & Lubin, J.H. (2001). Using splines to analyse latency in the Colorado Plateau uranium miners cohort, *Journal of Epidemiology and Biostatistics* **6**, 417–424.

[36] Hauptmann, M., Wellmann, J., Lubin, J.H., Rosenberg, P.S. & Kreienbrock, L. (2000). Analysis of exposure-time-response relationships using a spline weight function, *Biometrics* **56**, 1105–1108.

[37] Hazelton, W.D., Luebeck, E.G., Heidenreich, W.E. & Moolgavkar, S.H. (2001). Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model, *Radiation Research* **156**, 78–94.

[38] Heid, I.M., Kuchenhoff, H., Wellmann, J., Gerken, M. & Kreienbrock, L. (2002). On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology, *Statistics in Medicine* **21**, 3261–3278.

[39] Heidenreich, W.F., Jacob, P., Paretzke, H.G., Cross, F.T. & Dagle, G.E. (1999). Two-step model for the risk of fatal and incidental lung tumors in rats exposed to radon, *Radiation Research* **151**, 209–217.

[40] Heidenreich, W.F., Luebeck, E.G., Hazelton, W.D., Paretzke, H.G. & Moolgavkar, S.H. (2002). Multistage models and the incidence of cancer in the cohort of atomic bomb survivors, *Radiation Research* **158**, 607–614.

[41] Hoel, D.G. (1999). Comments on "Threshold models in radiation carcinogenesis" – Response, *Health Physics* **76**, 434–435.

[42] Hoel, D.G. & Li, P. (1998). Threshold models in radiation carcinogenesis, *Health Physics* **75**, 241–250.

[43] Infante-Rivard, C., Mathonnet, G. & Sinnett, D. (2000). Risk of childhood leukemia associated with diagnostic irradiation and polymorphisms in DNA repair genes, *Environmental Health Perspectives* **108**, 495–498.

[44] International Commission on Radiation Units and Measurements. (1993). *Quantities and Units in Radiation Protection Dosimetry*, ICRU Report 51, International Commission on Radiation Units and Measurements, Bethesda.

[45] International Commission on Radiation Units and Measurements. (2002). *Retrospective Assessment of Exposure to Ionising Radiation*, ICRU Report 68, Vol. 2. International Commission on Radiation Units and Measurements, Bethesda.

[46] Jablon, S. (1971). *Atomic Bomb Radiation dose Estimate at ABCC*, Technical Report 23–71, Atomic Bomb Casuality Commission, Hiroshima.

[47] Jablon, S., Hrubec, Z. & Boice, J.D. (1991). Cancer in populations living near nuclear-facilities – a survey of mortality nationwide and incidence in 2 states, *Journal of the American Medical Association* **265**, 1403–1408.

[48] Jones, I.M., Tucker, J.D., Langlois, R.G., Mendelsohn, M.L., Pleshanov, P. & Nelson, D.O. (2001). Evaluation of three somatic genetic biomarkers as indicators of low dose radiation effects in clean-up workers of the Chernobyl nuclear reactor accident, *Radiation Protection Dosimetry* **97**, 61–67.

[49] Kadhim, M.A., Lorimore, S.A., Townsend, K.M.S., Goodhead, D.T., Buckle, V.J. & Wright, E.G. (1995). Radiation-induced genomic instability – delayed cytogenetic aberrations and apoptosis in primary human bone-marrow cells, *International Journal of Radiation Biology* **67**, 287–293.

[50] Kai, M., Luebeck, E.G. & Moolgavkar, S.H. (1997). Analysis of the incidence of solid cancer among atomic bomb survivors using a two-stage model of carcinogenesis, *Radiation Research* **148**, 348–358.

[51] Kase, K.R. & Sinclair, W. (1989). *Radiation Dosimetry-Past and Present*, Vol. 11. National Council on Radiation Protection, Bethesda, pp. 299–328; Radiation protection today–the NCRP at sixty years. 4-5-1989.

[52] Kerber, R.A., Till, J.E., Simon, S.L., Lyon, J.L., Thomas, D.C., Preston-martin, S., Rallison, M.L., Lloyd, R.D. & Stevens, W. (1993). A cohort study of thyroid-disease in relation to fallout from nuclear-weapons testing, *Jama-Journal of the American Medical Association* **270**, 2076–2082.

[53] Khokhryakov, V.F., Suslova, K.G., Vostrotin, V.V., Romanov, S.A., Menshikh, Z.S., Kudryavtseva, T.I., Filipy, R.E., Miller, S.C. & Krahenbuhl, M.P. (2002). The development of the plutonium lung clearance model for exposure estimation of the Mayak production association, nuclear plant workers, *Health Physics* **82**, 425–431.

[54] Lagarde, F., Falk, R., Almren, K., Nyberg, F., Svensson, H. & Pershagen, G. (2002). Glass-based radon-exposure assessment and lung cancer risk, *Journal of Exposure Analysis and Environmental Epidemiology* **12**, 344–354.

[55] Land, C.E., Boice, J.D., Shore, R.E., Norman, J.E. & Tokunaga, M. (1980). Breast-cancer risk from low-dose exposures to ionizing-radiation – results of parallel analysis of 3 exposed populations of women, *Journal of the National Cancer Institute* **65**, 353–376.

[56] Langholz, B., Thomas, D., Xiang, A. & Stram, D. (1999). Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado plateau uranium miners cohort, *American Journal of Industrial Medicine* **35**, 246–256.

[57] Likhtarev, I.A., Kovgan, L.N., Vavilov, S.E., Perevoznikov, O.N., Litvinets, L.N., Anspaugh, L.R., Jacob, P. & Prohl, G. (2000). Internal exposure from the ingestion of foods contaminated by Cs-137 after the Chernobyl accident – report 2. Ingestion doses of the rural population of Ukraine up to 12 Y after the accident (1986–1997), *Health Physics* **79**, 341–357.

[58] Little, J.B. (2000). Radiation carcinogenesis, *Carcinogenesis* **21**, 397–404.

[59] Little, M.P. (1995). Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon, and Knudson, and of the multistage model of Armitage and Doll, *Biometrics* **51**, 1278–1291.

[60] Little, M.P., Hawkins, M.M., Charles, M.W. & Hildreth, N.G. (1992). Fitting the Armitage-Doll model to radiation-exposed cohorts and implications for population cancer risks, *Radiation Research* **132**, 207–221. [published erratum appears in *Radiation Research* 137, 124–128, 1994].

[61] Little, M.P., Haylock, R.G.E. & Muirhead, C.R. (2002). Modelling lung tumour risk in radon-exposed uranium miners using generalizations of the two-mutation model of Moolgavkar, Venzon and Knudson, *International Journal of Radiation Biology* **78**, 49–68.

[62] Little, M.P. & Muirhead, C.R. (1998). Curvature in the cancer mortality dose response in Japanese atomic bomb survivors: absence of evidence of threshold, *International Journal of Radiation Biology* **74**, 471–480.

[63] Lubin, J.H. (1998). On the discrepancy between epidemiologic studies in individuals of lung cancer and residential radon and Cohen's ecologic regression, *Health Physics* **75**, 4–10.

[64] Lubin, J.H., Boice, J.D. Jr., Edling, C., Hornung, R.W., Howe, G.R., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P. & Samet, J.M. (1995). Lung cancer in radon-exposed miners and estimation of risk from indoor exposure, *Journal of the National Cancer Institute* **87**, 817–827.

[65] Lubin, J.H., Samet, J.M. & Weinberg, C. (1990). Design issues in epidemiologic studies of indoor exposure to Rn and risk of lung cancer, *Health Physics* **59**, 807–817.

[66] Lubin, J.H., Tomasek, L., Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Tirmarche, M., Woodward, A. & Yao, S.X. (1997). Estimating lung cancer mortality from residential radon using data for low exposures of miners, *Radiation Research* **147**, 126–134.

[67] Luebeck, E.G., Curtis, S.B., Cross, F.T. & Moolgavkar, S.H. (1996). Two-stage model of radon-induced malignant lung tumors in rats: effects of cell killing, *Radiation Research* **145**, 163–173.

[68] Mallick, B., Hoffman, F.O. & Carroll, R.J. (2002). Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site, *Biometrics* **58**, 13–20.

[69] March, H.C. (1944). Leukemia in radiologists, *Radiology* **43**, 275–278.

[70] Meinert, R., Kaletsch, U., Kaatsch, P., Schuz, J. & Michaelis, J. (1999). Associations between childhood cancer and ionizing radiation: results of a population-based case-control study in Germany, *Cancer Epidemiology Biomarkers & Prevention* **8**, 793–799.

[71] Miller, R.C., Marino, S.A., Brenner, D.J., Martin, S.G., Richards, M., Randers-Pehrson, G. & Hall, E.J. (1995). Biological effectiveness of radon-progeny alpha-particles. 2. Oncogenic transformation as a function of linear-energy – transfer, *Radiation Research* **142**, 54–60.

[72] Moolgavkar, S.H., Dewanji, A. & Venzon, D.J. (1988). A stochastic 2-stage model for cancer risk assessment. 1. the hazard function and the probability of tumor, *Risk Analysis* **8**, 383–392.

[73] Moolgavkar, S.H. & Knudson, A.G. (1981). Mutation and cancer – a model for human carcinogenesis, *Journal of the National Cancer Institute* **66**, 1037–1052.

[74] Moolgavkar, S.H. & Luebeck, G. (1990). 2-event model for carcinogenesis – biological, mathematical, and statistical considerations, *Risk Analysis* **10**, 323–341.

[75] Moolgavkar, S.H., Luebeck, E.G., Krewski, D. & Zielinski, J.M. (1993). Radon, cigarette smoke, and lung cancer: a re-analysis of the Colorado Plateau uranium miners' data, *Epidemiology* **4**, 204–217.

[76] Morgan, W.F., Day, J.P., Kaplan, M.I., McGhee, E.M. & Limoli, C.L. (1996). Genomic instability induced by ionizing radiation, *Radiation Research* **146**, 247–258.

[77] Morgenstern, H. (1995). Ecologic studies in epidemiology: concepts, principles, and methods, *Annual Review of Public Health* **16**, 61–81.

[78] Nagasawa, H. & Little, J.B. (1992). Induction of sister chromatid exchanges by extremely low-doses of alpha-particles, *Cancer Research* **52**, 6394–6396.

[79] Nagataki, S., Shibata, Y., Inoue, S., Yokoyama, N., Izumi, M. & Shimaoka, K. (1995). Thyroid-disease among atomic-bomb survivors in Nagasaki, *Journal of the American Medical Association* **273**, 288.

[80] Nakamura, N., Miyazawa, C., Sawada, S., Akiyama, M. & Awa, A.A. (1998). A close correlation between electron spin resonance (ESR) dosimetry from tooth enamel and cytogenetic dosimetry from lymphocytes of Hiroshima atomic-bomb survivors, *International Journal of Radiation Biology* **73**, 619–627.

[81] National Academy of Sciences. (1956). *Biologic Effects of Atomic Radiation*. National Academy of Sciences, Washington.

[82] National Cancer Institute. (1997). *Estimated Exposures and Thyroid Doses Received by the American People from Iodine-131 in Fallout Following Nevada Atmospheric Nuclear Bomb Tests*. U.S. Department of Health and Human Services, National Institutes of Health, National Cancer Institute, Bethesda.

[83] National Research Council. (1990). *Health Effects of Exposure to Low Levels of Ionizing Radiation (BEIR V)*. National Academy Press, Washington.

[84] National Research Council. (1999). *Health Effects of Exposure to Radon (BEIR VI)*. National Academy Press, Washington.

[85] Otake, M., Finch, S.C., Choshi, K., Takaku, I., Mishima, H. & Takase, T. (1992). Radiation-related opthalmological changes and aging among Hiroshima and Nagasaki A-bomb survivors – a reanalysis, *Radiation Research* **131**, 315–324.

[86] Piantadosi, S., Byar, D.P. & Green, S.B. (1988). The ecologic fallacy, *American Journal of Epidemiology* **127**, 893–904.

[87] Pierce, D.A. & Mendelsohn, M.L. (1999). A model for radiation-related cancer suggested by atomic bomb survivor data, *Radiation Research* **152**, 642–654.

[88] Pierce, D.A. & Preston, D.L. (1993). Joint analysis of site-specific cancer risks for the atomic-bomb survivors, *Radiation Research* **134**, 134–142.

[89] Pierce, D.A. & Preston, D.L. (2000). Radiation-related cancer risks at low doses among atomic bomb survivors, *Radiation Research* **154**, 178–186.

[90] Pierce, D.A., Shimizu, Y., Preston, D.L., Vaeth, M. & Mabuchi, K. (1996). Studies of the mortality of atomic bomb survivors. Report 12, part I. Cancer, *Radiation Research* **146**, 1–27.

[91] Pierce, D.A., Stram, D.O. & Vaeth, M. (1990). Allowing for random errors in radiation-dose estimates for the atomic-bomb survivor data, *Radiation Research* **123**, 275–284.

[92] Pierce, D.A. & Vaeth, M. (2000). Age-time distribution of cancer risks to be expected from acute or chronic exposures to general mutagens, *Radiation Research* **154**, 727–728.

[93] Preston, D.L., Kusumi, S., Tomonaga, M., Izumi, S., Ron, E., Kuramoto, A., Kamada, N., Dohy, H., Matsuo, T., Nonaka, H., Thompson, D.E., Soda, M. & Mabuchi, K. (1994). Cancer incidence in atomic bomb survivors. Part III: incidence of leukemia, lymphoma, and multiple myeloma, 1950–1987, RERF TR 24–92, *Radiation Research* **137**, S68–S97.

[94] Preston, D.L., Mattsson, A., Holmberg, E., Shore, R., Hildreth, N.G. & Boice, J.D. (2002). Radiation effects on breast cancer risk: a pooled analysis of eight cohorts, *Radiation Research* **158**, 220–235.

[95] Puskin, J.S. (2003). Smoking as a confounder in ecologic correlations of cancer mortality rates with average county radon levels, *Health Physics* **84**, 526–532.

[96] Reeves, G.K., Cox, D.R., Darby, S.C. & Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models, *Statistics in Medicine* **17**, 2157–2177.

[97] Roesch, W. ed. (1987). *Final report: US-Japan joint reassessment of atomic bomb radiation dosimetry in Hiroshima and Nagasaki*. Radiation Effects Research Foundation, Hiroshima.

[98] Ron, E. (1996). Thyroid Cancer, in *Cancer Epidemiology and Prevention*, D. Schottenfeld & J.F. Jr. Fraumeni, eds. Oxford University Press, New York, pp. 1000–1021.

[99] Ron, E., Lubin, J.H., Shore, R.E., Mabuchi, K., Modan, B., Pottern, L.M., Schneider, A.B., Tucker, M.A. & Boice, J.D. (1995). Thyroid-cancer after exposure to external radiation – a pooled analysis of 7 studies, *Radiation Research* **141**, 259–277.

[100] Rossi, H.H. & Zaider, M. (1996). Comment on the contribution of neutrons to the biological effect at Hiroshima, *Radiation Research* **146**, 590–591.

[101] Schafer, D.W., Lubin, J.H., Ron, E., Stovall, M. & Carroll, R.J. (2001). Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry, *Biometrics* **57**, 689–697.

[102] Shimizu, Y., Pierce, D.A., Preston, D.L. & Mabuchi, K. (1999). Studies of the mortality of atomic bomb survivors. Report 12, part II. Noncancer mortality: 1950–1990, *Radiation Research* **152**, 374–389.

[103] Shu, X.O., Jin, F., Linet, M.S., Zheng, W., Clemens, J., Mills, J. & Gao, Y.T. (1994). Diagnostic-X-ray and ultrasound exposure and risk of childhood- cancer, *British Journal of Cancer* **70**, 531–536.

[104] Shu, X.O., Potter, J.D., Linet, M.S., Severson, R.K., Han, D.H., Kersey, J.H., Neglia, J.P., Trigg, M.E. & Robison, L.L. (2002). Diagnostic X-rays and ultrasound exposure and risk of childhood acute lymphoblastic leukemia by immunophenotype, *Cancer Epidemiol. Biomarkers & Prev.* **11**, 177–185.

[105] Simon, S., Gordeev, K., Bouville, A., Luckyanov, N., Land, C. & Carr, C. (2002). Estimates of radiation doses to members of a cohort residing in villages near the Semipalatinsk nuclear test site, *Health Physics* **82**, S168–S169.

[106] Sposto, R., Stram, D.O. & Awa, A.A. (1991). An estimate of the magnitude of random errors in the DS86-Dosimetry from data on chromosome aberrations and severe epilation, *Radiation Research* **128**, 157–169.

[107] Steck, D.J., Field, R.W. & Lynch, C.F. (1999). Exposure to atmospheric radon, *Environmental Health Perspectives* **107**, 123–127.

[108] Stovall, M., Smith, S.A. & Rosenstein, M. (1989). Tissue doses from radiotherapy of cancer of the uterine cervix, *Medical Physics* **16**, 726–733.

[109] Stram, D.O., Langholz, B., Huberman, M. & Thomas, D.C. (1999). Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado plateau uranium miners cohort, *Health Physics* **77**, 265–275.

[110] Stram, D.O., Langholz, B. & Thomas, D.C. (1998). *Measurement Error Correction of Lung Cancer Risk Estimates in the Colorado Plateau Cohort. Part I: Dosimetry Analysis*, Technical Report #126, School of Medicine, Department of Preventive Medicine, Division of Biostatistics, Los Angeles.

[111] Thomas, D.C. (1990). A model for dose-rate and duration of exposure effects in radiation carcinogenesis, *Environmental Health Perspectives* **87**, 163–171.

[112] Thomas, D.C. (1999). The Utah fallout study: how uncertainty has affected estimates of dose-response, in *Uncertainties in Radiation Dosimetry and Their Impact on Dose-response Analyses*, E. Ron & F.O. Hoffman, eds., Publication No. 99–4541. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, Washington, pp. 217–224.

[113] Thompson, D.E., Mabuchi, K., Ron, E., Soda, M., Tokunaga, M., Ochikubo, S., Sugimoto, S., Ikeda, T., Terasaki, M., Izumi, S. & Preston, D.L. (1994). Cancer incidence in atomic-bomb survivors 2. Solid tumors, 1958–1987, *Radiation Research* **137**, S17–S67.

[114] Till, J.E., Simon, S.L., Kerber, R., Lloyd, R.D., Stevens, W., Thomas, D.C., Lyon, J.L. & Preston-martin, S. (1995). The Utah Thyroid Cohort Study – analysis of the dosimetry results, *Health Physics* **68**, 472–483.

[115] Travis, L.B., Gospodarowicz, M., Curtis, R.E., Clarke, E.A., Andersson, M., Glimelius, B., Joensuu, T., Lynch, C.F., van Leeuwen, F.E., Holowaty, E., Storm, H., Glimelius, I., Pukkala, E., Stovall, M., Fraumeni, J.F., Boice, J.D. & Gilbert, E. (2002). Lung cancer following chemotherapy and radiotherapy for Hodgkin's disease, *Journal of the National Cancer Institute* **94**, 182–192.

[116] Tucker, J.D. (2001). Fish cytogenetics and the future of radiation biodosimetry, *Radiation Protection Dosimetry* **97**, 55–60.

[117] United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). (2000). *Sources and Effects of Ionizing Radiation. Report to the General Assembly, with Scientific Annexes*. United Nations Publication, New York.

[118] Vasilenko, E., Khokhryakov, V., Miller, S. & Rabovsky, J. (2002). Determination of radiation doses received by workers at the Mayak production association, *Health Physics* **82**, S164.

[119] Wang, Z.Y., Lubin, J.H., Wang, L.D., Zhang, S.Z., Boice, J.D. Jr., Cui, H.Z., Zhang, S.R., Conrath, S., Xia, Y., Shang, B., Brenner, A., Lei, S.W., Metayer, C., Cao, J.S., Chen, K.W., Lei, S.J., & Kleinerman, R.A. (2002). Residential radon and lung cancer in a high exposure area in Gansu province, China, *American Journal of Epidemiology* **155**, 554–564.

[120] Weinberg, C.R., Moledor, E.S., Umbach, D.M. & Sandler, D.P. (1996). Imputation for exposure histories with gaps, under an excess relative risk model, *Epidemiology* **7**, 490–497.

[121] Wu, L.J., Randers-Pehrson, G., Xu, A., Waldren, C.A., Geard, C.R., Yu, Z.L. & Hei, T.K. (1999). Targeted cytoplasmic irradiation with alpha particles induces mutations in mammalian cells, *Proceedings of the National Academy of Sciences* **96**, 4959–4964.

[122] Yamazaki, J.N. & Schull, W.J. (1990). Perinatal loss and neurological abnormalities among children of the atomic-bomb – Nagasaki an Hiroshima revisited, 1949 to 1989, *Journal of the American Medical Association* **264**, 605–609.

[123] Yoshimaru, H., Otake, M., Schull, W.J. & Funamoto, S. (1995). Further observations on abnormal brain-development caused by prenatal A-bomb exposure to ionizing-radiation, *International Journal of Radiation Biology* **67**, 359–371.

ELAINE RON & JAY H. LUBIN

Radiation Hybrid Mapping

Radiation hybrid (RH) mapping is a somatic cell genetic technique for constructing maps of mammalian chromosomes. Originally developed in the 1970s, its current incarnation evolved in the early 1990s as a way to bridge the gap in map resolution between genetic linkage mapping, with a resolution of approximately 1 megabase (Mb) pairs of deoxyribonucleic acid (**DNA**), and physical mapping by pulse field gel electrophoresis, with a resolution of up to a few hundred kilobase (kb) pairs. RH maps can include nonpolymorphic markers (*see* **Genetic Markers**; **Polymorphism**) and are flexible and powerful enough to order polymorphic markers too closely linked to be easily mapped with **linkage** methods.

A set of hybrid clones ("hybrids") created in one experiment is called a RH panel. Early panels consisted of hybrids containing fragments from only a single human chromosome. This technology quickly gave way to whole-genome panels, which contain fragments from all human chromosomes, allowing one to build maps of any portion of the human genome.

Experimental Design

To develop a whole-genome RH panel for human mapping, one starts with a diploid human fibroblast cell line. The cells are irradiated with a dose of radiation strong enough to fragment the chromosomes into several pieces: the higher the radiation dose, the smaller the fragments, and the higher the mapping resolution. Cells containing the fragmented human chromosomes are recovered by fusion with a thymidine kinase (TK) deficient rodent cell line, and hybrid cells containing human chromosome fragments are selected for in hypoxanthine–aminopterin–thymine (HAT) medium. The hybrid cells retain loci near the human TK gene on chromosome 17 at high frequency, but retain loci on other human chromosomes nonselectively. The hybrids cells are propagated to create clones; each hybrid clone ("hybrid") contains cells with a unique set of fragments from the original human chromosomes, and can be tested for the presence of known human DNA markers.

Mapping begins by genotyping each hybrid for the markers to be mapped. A vector $\mathbf{X}_i$ denotes the observations for hybrid i under a given order, with a 1 indicating that a marker is present in the hybrid, and a 0 indicating absence. In reality, hybrids are often genotyped twice. If the **genotypes** at any marker are discordant, then the genotype for that hybrid at that locus is treated as missing. Whole-genome RHs may have retained 0, 1 or 2 copies of any marker. However, assays typically only determine the presence or absence of the marker, and not the number of times it is present in the hybrid. The patterns of presence and absence of the markers in the hybrid clones are used to infer marker order by exploiting the principle that the closer together two markers are on a chromosome, the fewer radiation-induced breaks are expected to occur between them, and the more likely it is that they will be retained or lost together in a hybrid. Figure 1 depicts a simple example of a single chromosome with eight markers and four hybrids.

Statistical Analysis

The goal of RH mapping is to determine the correct or most likely order of a set of markers and to estimate the distances between the markers. The analyst must choose both the criterion for evaluating and comparing individual locus orders, and a search method for traversing through the possible locus orders in order to find the optimal map.

Methods for Evaluating a Locus Order

Nonparametric. The closer together two loci lie on a chromosome, the less likely it is that a radiation-induced break will separate them, and thus the less likely it is that one will be lost and the other retained in a hybrid. A simple strategy is to find the order with the fewest total number of obligate chromosome breaks between adjacent markers (see, for example, [4]). An obligate chromosome break is observed between two markers in a hybrid if one of the markers is present and the other is absent from the hybrid. In counting the number of obligate breaks, untyped markers are ignored. For example, if hybrid 1 in Figure 1 has been typed for all eight markers, then the observation vector under the true order is $\mathbf{X}_1 = (10\,000\,111)$. This order requires two obligate

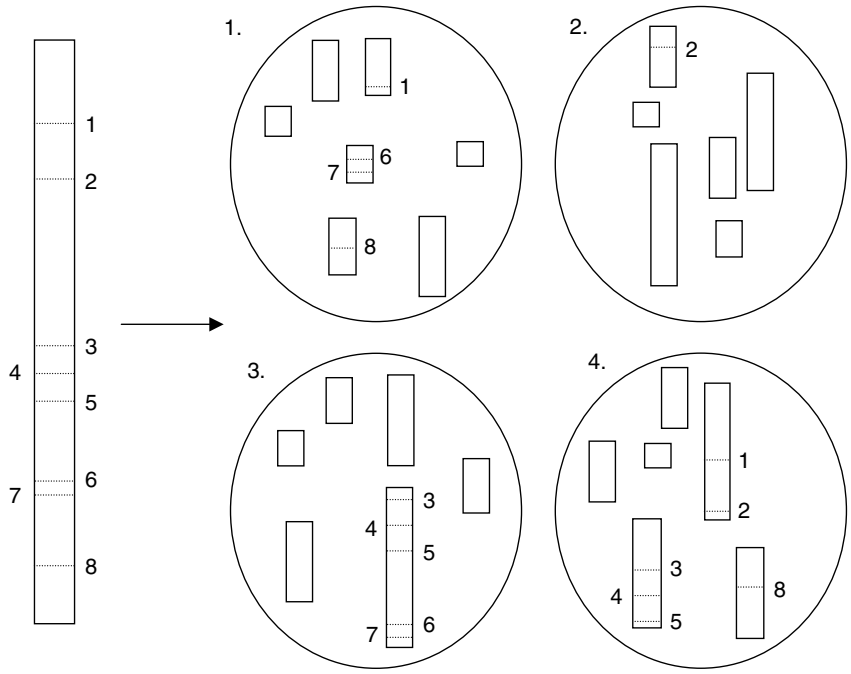

Figure 1 The figure at the left depicts the true (unknown in actual mapping experiments) locations of eight markers on a chromosome. The four circles represent four different hybrids. Within each are shown only the retained pieces of the chromosome of interest. In the first hybrid, three fragments have retained one or more markers from the chromosome. In hybrid 2, only one fragment containing a marker from this chromosome was retained. In hybrid 3, one fragment containing five markers was retained. In the last hybrid, three fragments contain six of the eight markers of interest. Using the correct chromosome order, the hybrid genotypes would be displayed as: $X_1 = (10\,000\,111)$; $X_2 = (01\,000\,000)$; $X_3 = (00\,111\,110)$; $X_4 = (11\,111\,001)$

breaks; one between the first and second markers, and one between the fifth and sixth. Other breaks may have occurred, and in fact the figure shows that there is also a break between the seventh and eight markers. However, only two breaks are required to explain the observed data for the true order. For this hybrid, any order in which the first marker is adjacent to one of the last three requires only one obligate break. Barrett [2] showed that under some common modeling assumptions the minimum breaks method is statistically consistent; as the number of hybrids increases, the probability of inferring the correct order converges to 1.0.

Maximum Likelihood. **Likelihood** methods produce an estimate of distances between markers as well as a statistic (the maximum likelihood) for evaluating a marker order. As with the obligate breaks method, we first choose a marker order. Under a given order, we estimate parameters via **maximum likelihood**. We consider the map order with the largest maximum likelihood the best order. Most researchers have parameterized their models in terms of fragment breakage and retention probabilities. Under a specific marker order $(A_1, A_2, \ldots, A_M)$, the breakage probability $\theta_i (1 \leq i < M)$ is the probability that at least one break occurs between markers A_i and A_{i+1}. The simplest assumption that can be made about fragment retention is that all fragments are equally likely to be retained. Actual data typically do not support this assumption, but it is usually adequate to order markers correctly, if not accurately estimate intermarker distances [11]. Many more realistic (and hence complex) retention models are available in RH mapping software (see, for example, [4], [10] and [15]).

Typically, we assume that breakage and retention are independent processes and that retention of one

fragment is independent of retention of any other (see, for example, [6]). Under the first assumption, breakage can be modeled as a Poisson process, and a breakage probability θ can be converted to an additive distance d using $d = -\ln(1-\theta)$, a formula analogous to Haldane's no-interference mapping function used in meiotic mapping (*see* **Genetic Map Functions**). The distance d is measured in Rays or centiRays (cR), which are analogous to Morgans and centiMorgans in meiotic mapping. A distance of 1 Ray corresponds to one expected break between the two loci per chromosome. The breakage probability between two markers, and therefore also the distance, depends on the radiation dose. Thus, it is best to state the X-ray dose with the distance, such as cR_{8000} for a dose of 8000 rads of radiation. The commercially available G3 whole-genome panel developed at Stanford [16] averages 24kb per $\mathrm{cR}_{10\,000}$. However, there is only a rough correspondence of breakage to physical length: the average ranges from 18 kb/cR for chromosome X to 33 kb/cR for chromosome Y.

For simplicity, assume a single retention rate r for all fragments. For a single hybrid and marker order $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_M$, we assume that the number of copies of each marker present in the hybrid $G_k, k = 1, \ldots, M$, form the states of a **Markov chain**. For whole-genome diploid (*see* **Human Genetics, Overview**) hybrids, only the presence or absence of markers is directly observable, and the states of the chain are partially hidden from view. The transition probabilities $t_{c,k}(i, j) = \Pr(G_{k+1} = j | G_k = i)$ are the probabilities that j copies of marker A_{k+1} are retained given that i copies of marker A_k are retained when the maximum number of copies (ploidy) is c.

To compute $t_{ck}(i, j)$, first consider a haploid hybrid. In this situation $c = 1$, and the transition probabilities are:

$$\begin{aligned} t_{1k}(0,0) &= 1 - \theta_k r, \\ t_{1k}(0,1) &= \theta_k r, \\ t_{1k}(1,0) &= \theta_k (1-r), \\ t_{1k}(1,1) &= 1 - \theta_k (1-r). \end{aligned} \quad (1)$$

The haploid transition probabilities (1) (*see* **Markov Chains**) can be used to construct the diploid transition matrix as follows:

$$\begin{pmatrix} t_{1k}(0,0)^2 & 2t_{1k}(0,0)t_{1k}(0,1) & t_{1k}(0,1)^2 \\ t_{1k}(1,0)t_{1k}(0,0) & \begin{matrix} t_{1k}(0,0)t_{1k}(1,1) \\ +t_{1k}(0,1)t_{1k}(1,0) \end{matrix} & t_{1k}(1,1)t_{1k}(0,1) \\ t_{1k}(1,0)^2 & 2t_{1k}(1,1)t_{1k}(1,0) & t_{1k}(1,1)^2 \end{pmatrix},$$

where the rows correspond to 0, 1, and 2 copies of marker A_k retained, and the columns to 0, 1, or 2 copies of marker A_{k+1} retained. For example, the probability of moving from a state where 0 copies of marker A_k are retained to a state where 2 copies of marker A_{k+1} are retained is found in the top right corner entry of the matrix.

For a single hybrid, let $\phi_i(x_i|g_i) = \Pr(X_i = x_i | G_i = g_i)$ be the probability that we observe marker A_i to be present ($X_i = 1$) or absent ($X_i = 0$) given that g_i copies of the marker are retained. $\phi_i(1|0)$ is the false positive error rate, and $\phi_i(0|g_i)$, $g_i = 1, 2$, are the false negative error rates for marker A_i. Here, assume that the error rates are known. Lange et al. [10] discuss aspects of error rate estimation, and Slonim et al. [15] discuss error detection. We assume that typing results for individual markers are independent given the $G_1, \ldots, G_M$; i.e.

$$\begin{aligned} &\Pr(X_1 = x_1, \ldots, X_M = x_M | G_1 = g_1, \ldots, G_M = g_M) \\ &\quad = \prod_{k=1}^{M} \phi_k(x_k|g_k). \end{aligned}$$

We further assume that the observations are independent given the underlying marker counts. Then the likelihood for the observations $(X_1, \ldots, X_M)$ from a single hybrid is

$$\begin{aligned} \Pr &= \Pr(X_1 = x_1, \ldots, X_M = x_M) \\ &= \sum_{g_1} \cdots \sum_{g_M} \binom{c}{g_1} r^g (1-r)^{c-g_1} \\ &\quad \times \prod_{k=1}^{M-1} t_{c,k}(g_k, g_{k+1}) \prod_{k=1}^{M} \phi_k(x_k|g_k). \end{aligned} \quad (2)$$

Since the hybrids are independent, their likelihoods are multiplied to get the full likelihood. For haploid hybrids with no typing error, there is no summation over the index g_k since $x_k = g_k$. The likelihood (2) can be evaluated as an iterated sum using Baum's algorithms from the theory of **hidden Markov** chains (see, for example, [10]), and the **EM algorithm** [7]

can be used to maximize the likelihood. Here, the retention patterns of the markers are the observed, "incomplete" data, and the "complete" data include the number of copies of each locus retained, the locations of the chromosome breaks, and the retention status for each fragment.

Bayesian. Bayesian methods yield posterior probabilities of locus orders that allow more easily interpretable comparisons of competing orders. However, the ease of interpretation comes at great expense in computational difficulty. Thus, the Bayesian methods that have been developed to date have not been widely used for map construction. One of the more promising methods uses a "simulated tempering" modified sampling scheme on the mixing characteristics of the Markov chain [8].

Map Building and Identifying the Best Order

For M markers, there are $M!/2$ possible marker permutations. When only a small number of markers are to be ordered, all possible orders may be evaluated and compared. For up to 10–12 markers, branch-and-bound algorithms (see, for example, [13]) also work well. Using branch and bound guarantees that the best order and all orders within a specified number of units (e.g. number of obligate breaks, or $\log_{10}$ likelihood) of the best order will be identified, while substantially decreasing the number of orders that must be evaluated. However, the number of orders to be evaluated still scales exponentially with the number of markers [4].

The locus ordering problem is a version of the classic traveling salesman problem (TSP) in combinatorial optimization. Thus, optimization algorithms that have been applied to the TSP are quite successful for building RH maps [1]. Simulated annealing [9] and genetic algorithms are explored by several authors [3, 4, 14].

For building maps across large segments or whole chromosomes, the common strategy is to first build a framework map, with markers that can be ordered with high confidence. Then, we put additional markers in "bins" along the framework using various strategies. Typically, these markers cannot be ordered among themselves with high certainty, but can be placed in a location relative to the framework (see, for example, [12], [15] and [16]). An alternative strategy for dealing with uncertainty in the order and location of markers is to calculate the posterior distribution of the position of each marker [17].

Software for RH Mapping

RHMAP: `http://www.sph.umich.edu/statgen/boehnke/rhmap.html` [5]
RHMAPPER: `http://www-genome.wi.mit.edu/ftp/pub/software/rhmapper/` [15]

References

[1] Agarwala, R., Applegate, D.L., Maglott, D., Schuler, G.D. & Schäffer, A.A. (2000). A fast and scalable radiation hybrid map construction and integration strategy, *Genome Research* **10**, 350–364.

[2] Barrett, J.H. (1992). Genetic mapping based on radiation hybrid data, *Genomics* **13**, 95–104.

[3] Ben-Dor, A., Chor, B. & Pelleg, D. (2000). RHO – Radiation Hybrid Ordering, *Genome Research* **10**, 365–378.

[4] Boehnke, M., Lange, K. & Cox, D.R. (1991). Statistical methods for multipoint radiation hybrid mapping, *American Journal of Human Genetics* **49**, 1174–1188.

[5] Boehnke, M., Lunetta, K., Hauser, E., Lange, K., Uro, J. & VanderStoep, J. (1996). *RHMAP: Statistical Package for Multipoint Radiation Version 3.0*.

[6] Cox, D.R., Burmeister, M., Price, E.R., Kim, S. & Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes, *Science* **250**, 245–250.

[7] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–22.

[8] Heath, S.C. (1997). Markov chain Monte Carlo methods for radiation hybrid mapping, *Journal of Computational Biology* **4**, 505–515.

[9] Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983). Optimization by simulated annealing, *Science* **220**, 671–680.

[10] Lange, K., Boehnke, M., Cox, D.R. & Lunetta, K.L. (1995). Statistical methods for polyploid radiation hybrid mapping, *Genome Research* **5**, 136–150.

[11] Lunetta, K.L. & Boehnke, M. (1994). Multipoint radiation hybrid mapping: comparison of methods, sample size requirements, and optimal study characteristics, *Genomics* **21**, 92–103.

[12] Lunetta, K.L., Boehnke, M., Lange, K. & Cox, D.R. (1995). Experimental design and error detection for polyploid radiation hybrid mapping, *Genome Research* **5**, 151–163.

[13] Nijenhuis, A. & Wilf, H.S. (1978). *Combinatorial algorithms*, 2nd Ed. Academic Press, New York, pp. 240–246.

[14] Slonim, D.K. (1996). Learning from imperfect data in theory and practice. PhD dissertation. MIT Laboratory for Computer Science. Technical Report MIT-LCS-TR-690.

[15] Slonim, D.K., Kruglyak, L., Stein, L. & Lander, E.S. (1997). Building human genome maps with radiation hybrids, *Journal of Computational Biology* **4**, 487–504.

[16] Stewart, E.A., McKusick, K.B., Aggarwal, A., Bajorek, E., Brady, S., Chu, A. et al. (1997). An STS-based radiation hybrid map of the human genome, *Genome Research* **7**, 422–433.

[17] Stringham, H.M., Boehnke, M. & Lange, K. (1999). Point and interval estimates of marker location in radiation hybrid mapping, *American Journal of Human Genetics* **65**, 545–553.

KATHRYN L. LUNETTA

Radiation

Biostatistical applications in the area of ionizing radiation are primarily addressed at understanding the biological consequences of radiation exposure, especially the health risks. The availability of data from a wide range of epidemiological and experimental studies, many with well-characterized estimates of dose or exposure, has led to sophisticated biostatistical modeling of risks that is perhaps unique to the radiation field. This article focuses on studies that have been aimed at identifying effects, understanding biological mechanisms, and estimating **dose–response** relationships for cancer risks, although other areas such as dosimetry and environmental clean-up have also required special statistical approaches.

During the 20 or so years following the discovery of X-rays in 1895, many of the effects of exposures to high doses of radiation, including acute tissue damage, risk of progressive anemia, and risk of cancer following doses sufficient to cause macroscopic tissue damage, were identified through clinical experience. Radiation exposure was also found to produce genetic changes in insects. Subsequently, following the atomic bombings of Hiroshima and Nagasaki, surveys quickly provided evidence that irradiation at high doses increased the risk of leukemia and cataracts in survivors, and of mental retardation among those exposed *in utero*. A program of genetic research on the effects of radiation was also set up in the late 1940s. However, in the mid-1950s, there were no data for obtaining the quantitative risk estimates needed for assessing the consequences of peacetime uses of radiation or for addressing concerns regarding world radioactive fallout from testing of hydrogen bombs.

As a result, large studies of the causes of death and the incidence of cancer among those who had been irradiated were initiated, first among the survivors of the Hiroshima and Nagasaki bombings and in several populations given medical irradiation, and later in several groups occupationally exposed to radiation including nuclear industry workers (*see* **Occupational Epidemiology**). Many of the studies included extensive efforts to estimate radiation doses. In addition to these studies, which were primarily of the effects of X- and gamma-radiation, studies of underground miners exposed to inhaled alpha particles in the form of radon and its progeny were also initiated. Together these studies have shown that the most important late effect of exposure to doses of irradiation too small to cause acute effects or macroscopic tissue damage is an increased risk of cancer in many organs and tissues in the body. Recently, evidence that exposure to ionizing radiation can increase the risk of cardiovascular, respiratory, and digestive diseases has also become available [37]. By contrast, hereditary effects of radiation exposure have not been clearly demonstrated in human populations. Atomic bomb survivor studies have also demonstrated effects on the developing brain, but absence of additional appropriate human data and uncertainty about the biological mechanisms have prevented extensive study of this topic. For a detailed historical review see Doll [12] and for a more recent review, see Ron et al. [34].

All human beings are exposed to low doses of radiation environmentally, many receive medical exposures, and some are also exposed occupationally. Because some of these exposures carry benefits to the individual or to society, and because reducing exposure can be costly or may carry alternative risks, quantifying risks from such exposures is of substantial interest to society. The effects of exposure to low doses of ionizing radiation in a population may not be negligible, and one estimate concluded that around 5%–6% of the 26 000 cancer deaths that occur in the US each year are likely to be due to natural ionizing radiation [6], but such calculations are subject to many difficulties, particularly that of extrapolating from high to low doses and low dose rates (*see* **Extrapolation, Low Dose**). Other difficulties in estimating radiation risks come about because experimental studies indicate that the effects of radiation exposure vary by the type of radiation involved (gamma rays, alpha particles, etc.) and in some cases by exposure rate, while epidemiologic studies indicate that radiation effects depend on characteristics of the exposed population such as age at exposure, sex and other factors. Also, to estimate lifetime risks and to evaluate detriment in terms of life-shortening, an understanding of the patterns of risk over time following exposure is required.

The Major Studies

The many settings in which radiation exposure has occurred, together with the ability to measure radiation dose, have led to a wealth of data on the

carcinogenic effects of radiation exposure so that, with the possible exception of smoking, more is known about the carcinogenic effects of exposure to radiation than of any other potential carcinogen.

Of particular importance is the Life Span Study of the survivors of the atomic bombings of Hiroshima and Nagasaki mentioned above, in which some 100 000 persons of all ages and both sexes were identified from the 1950 population **census**, and included in a **cohort study**. Dose estimates are available for about 86 000 subjects, of whom around 18 000 received doses of about 0.1 Sv or more to major internal body organs while a further 36 000 are thought to have been essentially unexposed, with doses of less than 0.005 Sv, and these form an internal **control** group (to get this in perspective, the average annual exposure to background radiation in the US excluding radon is roughly 0.001 Sv). To date, the cohort has been followed from 1950–1990 and a total of 4863 deaths have occurred from cancer in the exposed, of which 425 are estimated to be in excess (*see* **Excess Mortality**) and likely to be due to the radiation exposure; 85 of these excess deaths were due to leukemia [29]. For leukemia, most of the excess deaths occurred in the first 15 years after exposure, while for other cancers about 25% of the excess deaths occurred in the last 5 years of follow-up and the excess **absolute risk** per unit time is still increasing. There are clear dose–response relationships both for leukemia and for solid cancers (see Figure 1). Tumor registries

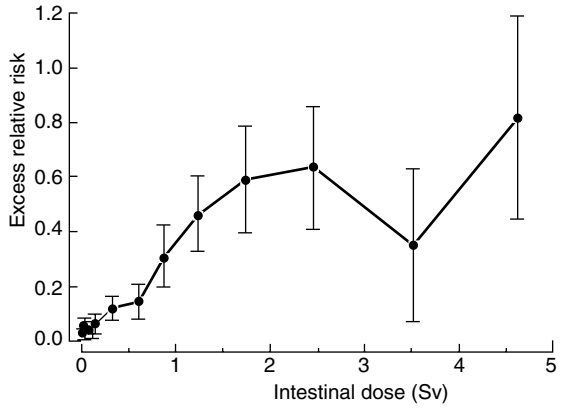

Figure 1 Shape of the dose–response curve for solid cancers among males aged 30 at exposure among survivors of the atomic bombings of Hiroshima and Nagasaki. Based on Pierce et al. [29]

in the cities of Hiroshima and Nagasaki also allow the study of cancer incidence in this cohort, with the most recent published results covering the period 1958–87 [41]. The tumor registries have also allowed detailed study of cancers of several specific sites.

As well as studies of the Japanese atomic bomb survivors, cohort studies of persons treated with generally high doses of external X-ray and gamma radiation for diseases such as ankylosing spondylitis, cervical cancer, tuberculosis, benign gynecological disease, peptic ulcer, skin hemangiomas, childhood cancers, mastitis, thymic reduction, and tinea capitis have been carried out. In addition, studies of persons exposed to internal alpha emitters have been conducted in radium dial painters and in patients with a variety of conditions injected with the contrast medium Thorotrast. In many instances these studies have reported results that are in accordance with those found in the Japanese atomic bomb survivors, but they have also provided important supplemental information. For example, they have demonstrated that, in marked contrast to other forms of leukemia, chronic lymphocytic leukemia, which occurs only very rarely among ethnic Japanese, is not readily inducible by radiation. In addition, a recent pooled analysis of several studies of breast cancer incidence have given support to the concept of a linear radiation dose response for breast cancer, have highlighted the importance of age and age at exposure on the risks, and have suggested a similarity in risks for acute and fractionated high-dose-rate exposures with much smaller effects from low-dose-rate protracted exposures. There is also a suggestion that women with some benign breast conditions may be at elevated risk of radiation-associated breast cancer Preston et al. [32].

Most of the above studies involve populations exposed to X-rays or gamma rays at doses and dose rates considerably higher than would normally be experienced and thus considerable downward extrapolation is involved in **risk assessment**. Several studies of much lower exposures have been carried out, including studies of nuclear workers [3, 26] and populations exposed to fallout [8, 38]. Where adequate dose estimates are available, such studies provide a direct evaluation of the risks at the actual levels of interest. Although these studies cover large populations, their low doses inevitably result in

reduced **power** to detect effects and substantial potential for **confounding**, and they cannot be expected to provide precise estimates of risk or much information on modifying factors (*see* **Effect Modification**). Nevertheless, they have provided reassurance that extrapolation from higher doses and dose rates has not seriously underestimated risks.

A promising source of new information comprises persons who were exposed in countries of the former Soviet Union, where exposures were generally much larger than those from similar activities in other countries. Studies include those exposed as a result of the Chornobyl accident in 1986 [43] and those exposed as a result of operations of the Mayak nuclear facility, which began operations in 1948 to produce plutonium for the Soviet Union's nuclear weapons program ([17, 18] – or cite whole issue – see references). Studies of Chornobyl clean-up workers, workers at the Mayak facility, and populations living near the Techa River downwind of the Mayak facility are unique in providing information on protracted whole body exposure at cumulative doses that are sufficiently large that risks can be estimated with some degree of precision. In addition, studies of Chornobyl exposures have clearly demonstrated that exposure to iodine-131 in childhood can increase the risk of thyroid cancer, and extensive efforts to estimate thyroid doses for individual subjects should make it possible to quantify risk as a function of dose. Studies of Mayak workers have provided the first direct demonstration in humans that exposure to plutonium increases risk of lung, liver, and bone cancers.

The studies discussed above are not directly relevant for one of the most important sources of population exposure to ionizing radiation, namely inhaled radon and its progeny. It had been known for several centuries that miners in the Erz Mountains had a high mortality from chest disease, but it was not until early in the twentieth century that it was appreciated that the disease was in fact lung cancer, and not until the 1950s that it was widely accepted that radon was likely to be the principal cause. About a dozen studies of radon-exposed miners have now been carried out including over 60 000 men. All the studies found high risks of lung cancer that were related to radon exposure [22] but the risk was very specific to this site of cancer and there was no material risk of mortality from other cancers [9].

The majority of the studies mentioned above are cohort studies. However, **case–control studies** have played an important role in documenting the association between risk of childhood cancer and obstetric X-rays [1], while **nested case–control** or **case–cohort studies** have sometimes been used to obtain information on doses [2, 44] or **covariates** [20] that would be too costly to obtain for the entire cohort. Case–control studies are also currently being used to estimate directly the risks of residential exposure to radon (e.g. [10, 28]). The need for this arises because residential radon is the largest source of exposure to ionizing radiation in many populations, often accounting for over half the population dose, whilst the available remedial measures are costly. Extrapolation of the risks from the miners' studies is particularly uncertain owing to the very different exposure conditions in mines compared with homes, and is complicated by the fact that there is evidence from the studies of radon-exposed miners that the risk per unit exposure varies with exposure rate, with a higher risk per unit exposure when the exposure is delivered at a lower compared with a higher rate [23, 42].

In addition to **observational studies** on humans, a large number of **experimental studies** have been conducted both *in vivo* and *in vitro*. These studies have increased our understanding of the carcinogenic process and have provided information in areas where human data are inadequate. Experimental studies have been especially important in developing our understanding of modifying factors such as exposure rate and type of radiation, and of the shape of the dose–response function.

The large and ever-growing number of studies has led to a sizable literature. This is reviewed periodically by a number of committees including the United Nations Scientific Committee on the Effects of Atomic Radiation UNSCEAR [43], the International Commission on Radiological Protection [16], and the Committee on Biological Effects of Ionizing Radiation of the US National Research Council [5, 27]. Their reports should be consulted for further details.

The Contribution of Biostatistics

The wealth of data and the availability of quantitative estimates of biologically relevant measures of dose

or exposure have led to the development of complex models for describing radiation risks that have drawn on information obtained from both epidemiologic and experimental studies. Statisticians have been stimulated to use a wide variety of methods to fit models that include the characterization of modifying factors and that address exposures that may be protracted over time.

As would be expected, the **Cox regression model** forms the basis of many applications. In a few cases (e.g. [14]) simple **proportional hazards** models have been fitted to data on individual subjects, but the complexity of fitting such models, especially when study sizes are large, often makes such an approach cumbersome. Furthermore, it is sometimes desirable to model absolute rather than relative risks, and this generally requires **parametric models** for the baseline risks. These difficulties can be avoided by the assumption of piecewise constant exponential hazards, enabling the data to be summarized in terms of tables of the numbers of events and **person-years at risk**. A wide variety of models can then be fitted by means of Poisson **regression**, and the flexible programme AMFIT [31] was developed to enable this type of analysis to be carried out easily (*see* **Additive Hazard Models**; **Poisson Regression in Epidemiology**).

The application of Poisson regression using AMFIT is illustrated by analyses of atomic bomb survivor mortality data in which both **excess relative risk** and excess absolute risk models were applied to data for leukemia and solid cancers with exploration of the shape of the dose–response functions and the dependence of risks on sex, city, age at exposure, and time since exposure [29]. Another application, which illustrates the feasibility of the approach for exposures that are protracted and variable over time, is an analysis of data from 11 underground miner cohorts [22], where risks were found to depend not only on cumulative exposure, but also on exposure rate, time since exposure, and age at risk. The shape of the exposure–response function and the effect of age at exposure were also explored in these analyses.

The radiobiological understanding that has come from experimental studies has sometimes affected the choice of models, and, in particular, has led to the use of models in which the relative risk is a linear or linear–quadratic function of dose. Biologically based models have also been applied to radiation data; for example, the two-stage clonal expansion model has been used to describe lung cancer risks in both Colorado miners [25] and rats [24] exposed to radon and radon progeny.

In the radiation epidemiology field there are several instances in which combined analyses of original data from several studies have been carried out. Early examples were analyses of data from three studies of radiation exposure and breast cancer [19] and analyses of data from the Life Span Study of Japanese atomic bomb survivors and patients given X-ray therapy for ankylosing spondylitis [7]. Subsequent combined analyses include those of 11 underground miner cohorts [9, 22], of several cohorts of nuclear workers in three countries [3], of several studies of radiation-induced thyroid cancer [35] and eight cohort studies of breast cancer [32]. A major reason for conducting combined analyses is to obtain more precise estimates of parameters than those based on individual studies, an advantage that is especially important for investigating modifying factors. Combined analyses also provide a more rigorous evaluation of consistency of results among studies. Some combined analyses have accounted for heterogeneity among the studies in the **confidence intervals** for the overall risk estimate [22, 35].

By contrast to the sophisticated analyses described above that have been applied to observational epidemiologic data, the statistical methods applied to experimental animal data have in many cases been fairly simple. However, in some cases the hazard has been modeled as a function of dose, age, and other factors, an approach that is similar to that applied to epidemiologic data although modifications have sometimes been required. In animals, where detailed examination of tissues is conducted after death, tumors are often not the cause of death, but are instead found incidentally to death from other causes, and appropriately modeling the hazard function requires different statistical treatment of fatal and incidental tumors. Although extensive statistical research on this statistical issue has been conducted, much of it has been addressed at appropriate tests of the **null hypotheses**, and methods need to be extended to allow the hazard modeling that is of interest in animal experiments involving exposure to radiation. Gart et al. [13] and Dewanji et al. [11] describe an approach in which two hazard functions (for fatal and incidental tumors) are modeled, and this general approach has been applied in analyses of

data on rats exposed to radon [15]. Other innovative analyses of animal data include those of Chmelevsky et al. [4] and those of Luebeck et al. [24] referred to above.

Recent Developments

Recently considerable effort has been expended on the pooled analysis of data from a number of different sources. The need for this arises in part from the desirability for further direct evaluation of risks to populations exposed at low levels, for example workers in the nuclear industry or populations exposed to environmental radon, where relative risks are small and the results from individual studies do not provide a clear answer (see Figure 2), and in part from the need to bring together data from a number of different sources, such as both high- and low-dose epidemiologic studies and experimental studies, to develop risk models. Already, some pooled analyses have revealed heterogeneity among the available studies (e.g. [22]), and in some cases there may be a need for the application of better methods to account for and describe such heterogeneity.

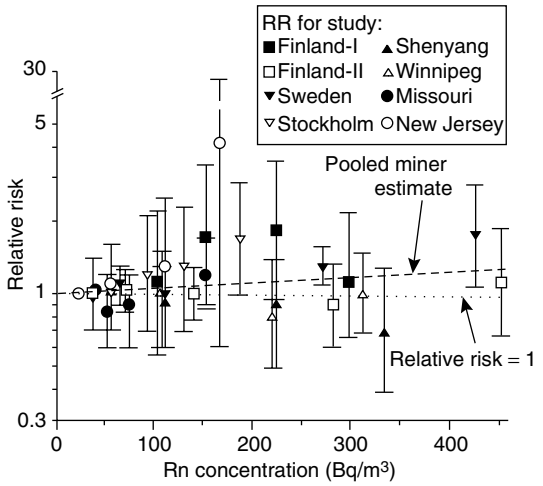

Figure 2 Relative risks for lung cancer by categories of radon (Rn) concentration for seven case–control studies of residential radon exposure. Also shown are the predicted risks from studies of underground miners exposed to radon, as estimated by the US Committee on the Biological Effects of Ionizing Radiations (BEIR IV). Based on Lubin et al. [21]

A considerable amount of effort has also been expended in the development and application of methods that account for uncertainties in the estimation of doses (or exposures) in several studies. The presence of **random errors** in individual dose estimates (*see* **Measurement Error in Epidemiologic Studies**) may cause underestimation of the effects of radiation in dose–response analyses, distort the shape of the dose–response curves, and cause confidence intervals and significance tests (*see* **Hypothesis Testing**) to give misleading answers. Some work on this topic has already been carried out in the radiation field. For example, Pierce et al. [30] estimated the distribution of true doses among Japanese atomic bomb survivors, and took an assumed coefficient of variation for the random dose errors, thereby aiming at a way to estimate E(true dose|estimated dose). Corrections based on this and related considerations are now being used in many analyses of atomic bomb survivor data. Another example is that of Thomas et al. who applied an **empirical Bayes** approach to a case–control study of leukemia in Utah residents exposed to radioactive fallout [40]. Recently, methods that model fully the effect of dose estimation uncertainties have been developed [33]. This work was carried out to provide methods for the analysis of case–control studies of indoor radon, but should also be useful for the analysis of a wide variety of case–control and cohort studies. Other recently published studies addressing measurement error include Schafer et al. [36] and by Stram et al. [39].

Finally, it seems likely that future scientific developments in radiation research and advances in understanding the biology and particularly the genetic basis of cancer will also stimulate future statistical developments.

References

[1] Bithell, J.F. (1989). Epidemiological studies of children irradiated in utero, in *Low Dose Radiation: Biological Bases of Risk Assessment*, K. Baverstock & J.W. Stather, eds. Taylor & Francis, London, pp. 77–87.

[2] Boice, J.D., Jr., Blettner, M., Kleinerman, R.A., Stovall, M., Moloney, W.C., Engholm, G., Austin, D.F., Bosch, A., Cookfair, D.L., Krementz, E.T., Latourette, H.B., Peters, L.J., Schulz, M.D., Lundell, M., Pettersson, F., Storm, H.H., Bell, C.M.J., Coleman, M.P., Fraser, P., Palmer, M., Prior, P., Choi, N.W., Hislop, T.G., Koch, M., Robb, D., Robson, D., Spengler, R.F., von Fournier,

D., Frischkorn, R., Lochmüller, H., Pompe-Kirn, V., Rimpela, A., Kjørstad, K., Pejovic, M.H., Sigurdsson, K., Pisani, P., Kucera, H. & Hutchison, G.B. (1987). Radiation dose and leukemia risk in patients treated for cancer of the cervix, *Journal of the National Cancer Institute* **79**, 1295–1311.

[3] Cardis, E., Gilbert, E.S., Carpenter, L., Howe, G., Kato, I., Armstrong, B.K., Beral, V., Cowper, G., Douglas, A., Fix, J., Fry, S.A., Kaldor, J., Lavé, C., Salmon, L., Smith, P.G., Voelz, G.L. & Wiggs, L.D. (1995). Effects of low doses and low dose rates of external ionizing radiation: cancer mortality among nuclear industry workers in three countries, *Radiation Research* **142**, 117–132.

[4] Chmelevsky, D., Kellerer, A.M., Lafuma, J. & Chameaud, J. (1982). Maximum likelihood estimation of the prevalence of nonlethal neoplasms – an application to radon-daughter inhalation studies, *Radiation Research* **91**, 589–614.

[5] Committee on Biological Effects of Ionizing Radiation (BEIR V) (1990). *Health Effects of Exposure to Low Levels of Ionizing Radiation*. National Academy of Sciences, National Research Council, National Academy Press, Washington.

[6] Darby, S.C. (1991). Contribution of natural ionizing radiation to cancer mortality in the United States, in *Origins of Human Cancer: A Comprehensive Review*, J. Brugge, T. Curran, E. Harlow & F. McCormick, eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 183–190.

[7] Darby, S.C., Nakashima, E. & Kato, H. (1985). A parallel analysis of cancer mortality among atomic bomb survivors and patients with ankylosing spondylitis given X-ray therapy, *Journal of the National Cancer Institute* **75**, 1–21.

[8] Darby, S.C., Olsen, J.H., Doll, R., Thakrar, B., de Nully Brown, P., Storm, H.H., Barlow, L., Langmark, F., Teppo, L. & Tulinius, H. (1992). Trends in childhood leukaemia in the Nordic countries in relation to fallout from atmospheric nuclear weapons testing, *British Medical Journal* **304**, 1005–1009.

[9] Darby, S.C., Whitley, E., Howe, G.R., Hutchings, S.J., Kusiak, R.A., Lubin, J.H., Morrison, H.I., Tirmarche, M., Tomásek, L., Radford, E.P., Roscoe, R.J., Samet, J.M. & Yao, S.X. (1995). Radon and cancers other than lung cancer in underground miners: a collaborative analysis of 11 studies, *Journal of the National Cancer Institute* **87**, 378–384.

[10] Darby, S., Whitley, E., Silcocks, P., Thakrar, B., Green, M., Lomas, P., Miles, J., Reeves, G., Fearn, T. & Doll, R. (1998). Risk of lung cancer associated with residential radon exposure in south-west England: a case-control study, *British Journal of Cancer* **78**, 394–408.

[11] Dewanji, A., Krewski, D. & Goddard, M.J. (1993). A Weibull model for the estimation of tumorigenic potency, *Biometrics* **49**, 367–377.

[12] Doll, R. (1995). Hazards of ionising radiation: 100 years of observation on man, *British Journal of Cancer* **72**, 1339–1349.

[13] Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research, Vol. III*, *The Design and Analysis of Long-Term Animal Experiments*, IARC Scientific Publications No. 79. International Agency for Research on Cancer, Lyon, France.

[14] Gilbert, E.S. (1989). Issues in analyzing the effects of occupational exposure to low levels of radiation, *Statistics in Medicine* **8**, 173–187.

[15] Gilbert, E.S., Cross, F.T. & Dagle, G.E. (1996). Analysis of lung tumor risks in rats exposed to radon, *Radiation Research* **145**, 330–360.

[16] International Commission on Radiological Protection, (1990). *Recommendations of the International Commission on Radiological Protection*, *Publication 60*. Pergamon Press, Oxford.

[17] Koshurnikova, N.A., Gilbert, E.S., Shilnikova, N.S., Sokolnikov, M., Preston, D.L., Kreisheimer, M., Ron, E., Okatenko, P. & Romanov, S.A. (2002),. Studies on the Mayak nuclear workers: health effects, *Radiation and Environmental Biophysix* **41**, 29–32.

[18] Kossenko, M.M., Preston, D.L., Krestinina, L.Y., Degteva, M.O., Startsev, N.V., Thomas, T., Vyushkova, O.V., Anapaugh, L.R., Napier, B.A., Kozheurov, V.P., Ron, E. & Akleyev, A.V. (2002). Studies on the extended Techa river cohort: cancer risk estimation, *Radiation and Environmental Biophysix* **41**, 29–32.

[19] Land, C.E., Boice, J.D., Jr, Shore, R.E., Norman, J.E. & Tokunaga, M. (1980). Breast cancer risk from low-dose exposures to ionizing radiation: results of parallel analysis of three exposed populations of women, *Journal of the National Cancer Institute* **65**, 353–376.

[20] Land, C.E., Hayakawa, N., Machado, S.G., Yamada, Y., Pike, M.C., Akiba, S. & Tokunaga, M. (1994). A case control interview study of breast cancer among Japanese A-bomb survivors: I. Main effects, *Cancer Causes & Control* **5**, 157–165.

[21] Lubin, J.H. & Boice, J.D., Jr (1997). Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies, *Journal of the National Cancer Institute* **89**, 49–57.

[22] Lubin, J.H., Boice, J.D., Jr, Edling, C., Hornung, R.W., Howe, G.R., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Tirmarche, M., Woodward, A., Yao, S.X. & Pierce, D.A. (1995). Lung cancer in radon-exposed miners and estimation of risk from indoor exposure, *Journal of the National Cancer Institute* **87**, 817–827.

[23] Lubin, J.H., Boice, J.D., Jr & Samet, J.M. (1995). Errors in exposure assessment, statistical power and the interpretation of residential radon studies, *Radiation Research* **144**, 329–341.

[24] Luebeck, E.G., Curtis, S.B., Cross, F.T. & Moolgavkar, S.H. (1996). Two-stage model of radon-induced

malignant tumors in rats: effects of cell killing, *Radiation Research* **145**, 163–173.

[25] Moolgavkar, S.H., Luebeck, E.G., Krewski, D. & Zielinski, J.M. (1993). Radon, cigarette smoke, and lung cancer: a re-analysis of the Colorado uranium miners' data, *Epidemiology* **4**, 204–217.

[26] Muirhead, C.R., Goodill, A.A., Haylock, R.G.E. et al. (1999). Occupational radiation exposure and mortality: second analysis of the National Registry for Radiation Workers, *Journal of Radiological Protection* **19**, 3–26.

[27] National Research Council, Committee on Health Risks of Exposure to Radon. (1999). *Health Effects of Exposure to Radon (BEIR VI)*. National Academy Press, Washington DC.

[28] Pershagen, G., Akerblom, G., Axelson, O., Clavensjö, B., Damber, L., Desai, G., Enflo, A., Lagarde, F., Mellander, H., Svartengren, M. & Swedjemark, G.A. (1994). Residential radon exposure and lung cancer in Sweden, *New England Journal of Medicine* **330**, 159–164.

[29] Pierce, D.A., Shimizu, Y., Preston, D.L., Vaeth, M. & Mabuchi, K. (1996). Studies of the mortality of atomic bomb survivors, Report 12, Part 1. Cancer: 1950–1990, *Radiation Research* **146**, 1–27.

[30] Pierce, D.A., Stram, D.O., Vaeth, M. & Schafer, D.W. (1992). The errors-in-variables problem: considerations provided by radiation dose-response analyses of the A-bomb survivor data, *Journal of the American Statistical Association* **87**, 351–359.

[31] Preston, D.L., Lubin, J.H. & Pierce, D.A. (1992). *Epicure: Risk Regression and Data Analysis Software*. Hirosoft International Corporation, Seattle, Washington.

[32] Preston, D.L., Mattsson, A., Holmberg, E., Shore, R., Hildreth, N.G. & Boice, J.D. Jr. (2002). Radiation effects on breast cancer risk: A pooled analysis of eight cohorts, *Radiation Research* **158**, 220–235.

[33] Reeves, G.K., Cox, D.R., Darby, S.C. & Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models, *Statistics in Medicine* **17**, 2157–2177.

[34] Ron, E. (1998). Ionizing radiation and cancer risk: evidence from epidemiology, *Radiation Research* **150**(Suppl.), S30–S41.

[35] Ron, E., Lubin, J.H., Shore, R.E., Mabuchi, K., Modan, B., Pottern, L.M., Schneider, A.B., Tucker, M.A. & Boice, J.D., Jr (1995). Thyroid cancer after exposure to external radiation: a pooled analysis of seven studies, *Radiation Research* **141**, 259–277.

[36] Schafer, D.W., Lubin, J.H., Ron, E., Stovall, M. & Carroll, R.J. (2001). Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry, *Biometrics* **57**, 689–697.

[37] Shimizu, Y., Pierce, D.A., Preston, D.L. & Mabuchi, K. (1999). Studies of the mortality of atomic bomb survivors: non-cancer mortality 1950–1990, *Radiation Research* **152**, 374–89. (Report 12, part 11.).

[38] Stevens, W., Thomas, D.C., Lyon, J.L., Till, J.E., Kerber, R.A., Simon, S.L., Lloyd, R.D., Elghany, N.A. & Preston-Martin, S. (1990). Leukemia in Utah and radioactive fallout from the Nevada test site, *Journal of the American Medical Association* **264**, 585–591.

[39] Stram, D.O., B. Langholz, M. Huberman, and D.C. Thomas (1999). Correction for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau uranium miners cohort, *Health Physics* **77**,(3) 265–275.

[40] Thomas, D.C., Gaudeman, J. & Kerber, R. (1990). A nonparametric Monte Carlo approach to adjustment for covariate measurement errors in regression analysis, *Department of Preventive Medicine Technical Report No. 15*. University of Southern California, Los Angeles.

[41] Thompson, D.E., Mabuchi, K., Ron, E. Soda, M., Tokunaga, M., Ochikubo, S., Sugimoto, S., Ikeda, T., Teraski, M., Izurni, S. & Preston, D.L. (1994). Cancer incidence in atomic bomb survivors. Part II: Solid tumors, 1958–1987, *Radiation Research* **137**, S17–S67. [RERF TR 5–92].

[42] Tomásek, L., Darby, S.C., Fearn, T., Swerdlow, A.J., Placek, V. & Kunz, E. (1994). Patterns of lung cancer mortality among uranium miners in West Bohemia with varying rates of exposure to radon and its progeny, *Radiation Research* **137**, 251–261.

[43] United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). (2000). *Sources and Effects of Ionizing Radiation*. United Nations, New York.

[44] Weiss, H.A., Darby, S.C., Fearn, T. & Doll, R., (1995). Leukemia mortality after X-ray treatment for ankylosing spondylitis, *Radiation Research* **142**, 1–11.

Further Reading

Blot, W.J., Xu, Z.Y., Boice, J.D., Jr, Zhao, D.-Z., Stone, B.J., Sun, J., Jing, L.-B. & Fraumeni, J.F. Jr (1990). Indoor radon and lung cancer in China, *Journal of the National Cancer Institute* **82**, 1025–1030.

Committee on Biological Effects of Ionizing Radiation (BEIR IV) (1988). *Health Risks of Radon and Other Internally Deposited Alpha-Emitters*. National Academy of Sciences, National Research Council, National Academy Press, Washington.

Pierce, D.A., Stram, D.O. & Vaeth, M. (1990). Allowing for random errors in radiation dose estimates for the atomic bomb survivor data, *Radiation Research* **123**, 275–284.

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (1994). Sources and Effects of Ionizing Radiation, *1994 Report to the General Assembly with Scientific Annexes*. United Nations, New York.

SARAH C. DARBY & ETHEL S. GILBERT

Radioimmunoassay

Radioimmunoassays (RIAs) are commonly performed in clinical and biomedical research laboratories to estimate the concentration of an antigen in a biological specimen. The RIA is a type of radioligand assay (RLA) in which antigens are labeled with radioisotopes. An analogous approach to radioligand assays in which antibodies are radiolabeled rather than antigens is referred to as an immunoradiometric assay (IRMA).

Radioimmunoassays differ from traditional bioassays (*see* **Biological Assay, Overview**) in that they arise from specific chemical reactions in biochemical systems that follow the law of mass action and do not have the errors associated with biological test systems (*see* **Pharmacokinetics and Pharmacodynamics**). Basically, the biochemical system modeled is the immune system which produces antigens (Ag) as a response to the presence of antibodies (Ab). The binding of antigens to antibodies can be represented as a two-way chemical reaction:

$$\mathrm{Ag} + \mathrm{Ab} \rightleftharpoons \mathrm{AgAb}.$$

The concentration of AgAb follows a curvilinear function of the initial concentrations of Ag and Ab. Therefore, by measuring the AgAb complex in the system, the initial concentrations of either the antibody or antigen can be inferred. Measuring AgAb is done by labeling part of the antigen with a radioactive tracer such as iodine-131. When the labeled antigen is introduced, the unlabeled antigen of unknown concentration in the system decreases the amount of labeled antigen that is bound to the antibody; thus, measuring either the unbound labeled antigen or the radioactive AgAb complex provides a method to estimate indirectly the unknown concentration of unlabeled antigen. As with other bioassays, a standard curve is developed, to which the specimens of unknown potency are compared. The techniques for formulating a standard curve, which are based on biochemical theory and the law of mass action, are described by McHugh & Meinert [6] and Meinert & McHugh [7].

The development of modern, virtually totally automated techniques, has enabled laboratories to move from the use of unsophisticated statistical procedures based on manual curve fitting to more appropriate methods for curve fitting which incorporate routine quality control checks for systematic errors and outliers. Weighted **least squares** estimation, or **maximum likelihood** estimation, based on linear or nonlinear functions, are often utilized, augmented by **analysis of variance** techniques to test the validity of the underlying model.

Statistical Design and Models

Much of the relevant statistical methodology for RIAs is found in clinical chemistry and biochemistry journals. Chapter 16 of Finney's classical text on bioassay methods [1] provides a useful summary of the statistical considerations and standard analysis techniques. The recent text by Govindarajulu [3] gives a more detailed description of the theoretical biochemical model and a useful list of references, that encompass laboratory quality control issues and **calibration** curve fitting.

The primary response measure in an RIA is a radiation count during a specified time interval for fixed dose levels. The counts of bound labeled or free ligand labeled antigen are designated B and F respectively, with T (total) $= B + F$. The symbol B_0 will be used for the expected count bound at zero dose level and N will be used for the expected count at "infinite (nonspecific)" dose. Direct measurements can be incorporated to estimate both "zero" and "infinite" dose levels. For statistical analysis, a logarithmic transformation of dose is typically used as a dose metameter (x), and a response metameter (y) is specified that is assumed to be linearly related to x, that is, $E(y) = \alpha + \beta x$. The expected count (U) at a given dose is expressed as

$$U = \mathrm{E}(u|x) = B_0 + (N - B_0)F(x),$$

where $F(x)$ can be any of a number of sigmoidally shaped curves, such as the **logistic** or cumulative **normal**, which will vary from one to zero for an RIA as x varies between $-\infty$ and $+\infty$ [1].

Suppose that the functional form used is

$$F(x) = \frac{1}{1 + \exp(-2Y)}.$$

(Note that some references use Y rather than $2Y$ in $F(x)$. If Y is used, then the estimates of α and β will be twice the value obtained when $2Y$ is used. Parameter estimates of B_0 and N are unchanged.) Then an

often satisfactory linearizing response metameter is

$$Y = \frac{1}{2}\ln\left(\frac{U-N}{B_0-U}\right) = \alpha + \beta x,$$

where $Y = \frac{1}{2}\log\text{it}(U-N)/(B_0-N)$. Iterative methods of estimation based on weighted least squares or maximum likelihood, analogous to those used in standard quantal bioassay designs, are utilized to estimate the parameters and to calculate the potency or relative potency.

An additional factor that must be considered for an RIA is the variance function of the observed counts, which is used for weighting the observations in the regression analysis. Analyses have shown that, in general, the variance of the counts is subject to extra-Poisson dispersion (*see* **Overdispersion**). (See, for example, Rodbard & Cooper [13] and Rodbard [11].) Rodbard [12] recommended a quadratic function for the variance, whereas Finney [1] has proposed that the variance ($\phi(U)$) be specified as

$$\phi(U) = VU^J,$$

where J is the rate of increase in variance with increase in count and U is assumed to be normally distributed. While the variance function could be estimated simultaneously within a single assay in conjunction with the regression parameters, greater reliability is obtained by pooling a large body of experiments, and assuming that J remains constant across assays. The careful analyst will reassess the value of J periodically. A flexible, comprehensive computer program is desirable to permit estimation of parameters assuming various forms for the weights (reciprocals of the variance function) (see, for example, [14]).

Potency Estimation

The parameters α, β, B_0, and N are generally estimated by minimizing the weighted sum of squares

$$\sum\left[\frac{(u-U)^2}{\phi(U)}\right]$$

or, alternatively, by maximizing the log likelihood

$$-\frac{1}{2}\sum\log[\phi(U)] - \frac{1}{2}\sum\left[\frac{(u-U)^2}{\phi(U)}\right].$$

Both approaches will give asymptotically equivalent results, but weighted least squares has more frequently been employed in estimation software primarily because of practical considerations.

The log potency is obtained by inserting a specified fraction bound (P) in

$$Y = \frac{1}{2}\ln\left(\frac{P}{1-P}\right)$$

to calculate Y and then solving for log dose (x) in

$$Y = \hat{\alpha} + \hat{\beta}x.$$

If $Y = 0$, then $\log(\text{ID}_{50}) = -\hat{\alpha}/\hat{\beta}$, the estimated log midrange (ID_{50}).

A common approach in radioimmunoassay is to compare a single dose level of a test preparation to a separately estimated standard response curve, where $\hat{\alpha}$, $\hat{\beta}$, $\hat{B}_0$, and $\hat{N}$ are the estimates obtained in the logit transformation analysis for the standard curve, and $\overline{u}$ is the mean value for n independent counts at log dose x of the test preparation (T). The logit is calculated as

$$Y = \frac{1}{2}\ln\left[\frac{\overline{u}-\hat{N}}{\hat{B}_0-\overline{u}}\right].$$

The potency estimate is then $\log(\hat{\rho}) = -x + (Y - \hat{\alpha})/\hat{\beta}$. The variance of Y depends not only on the variance of $\overline{u}$ but also on the variances of $\hat{\beta}_0$ and $\hat{N}$, and the covariance terms, which are generally estimated by first-order approximations for the variance of Y. The variance of $\overline{u}$ is estimated as

$$\phi(\overline{u}) = \frac{V\overline{u}^J}{n},$$

where V is the residual mean square of the least squares estimation based on the standard curve. Usual values of J in practice range from 1.0 to 1.5. The $(1-\alpha)\%$ **confidence intervals** are estimated through application of **Fieller's theorem** for ratio estimators.

Similarly, if observations are made within a single experiment on doses of standard (S) and test preparations (T), then the fundamental assumption of assay validity is the condition of similarity, which for radioimmunoassays leads to estimation closely analogous to that for **parallel-line assays**. The estimated log relative potency ($\hat{\rho}$) is

$$\log(\hat{\rho}) = \frac{\hat{\alpha}_T - \hat{\alpha}_S}{\hat{\beta}},$$

where $\hat{\beta}$, $\hat{B}_0$, and $\hat{N}$ are common estimates across both standard and test preparations. Fieller's theorem is employed to calculate confidence limits for $\log \rho$.

Validity Testing

Typically, as noted above, the dispersion of the counts increases directly with the mean response level; that is

$$\sigma_u^2 = \sigma^2 U^J,$$

where σ^2 is the weighted mean-square within replicate sets. A precision profile calculated from the within replicate "standard error" divided by the slope of the response curve (coefficient of variation) estimates the best that the assay can achieve. Tests relevant for evaluating the validity of the assay and the adequacy of the statistical model are [10]:

1. stability of assay system;
2. excessive within-replicate dispersion among sets of tubes;
3. **goodness of fit** of the assumed response curve; and
4. parallelism of the standard and test response curves.

Stability, or lack of systematic drift, is evaluated either (i) by checking one or two quality control samples at a few points throughout the assay or (ii) by including several standard curves at different times in the assay. The latter provides a more powerful tool for identifying drift. Healy [4] proposed a **robust** estimate of dispersion, based on each mean square following a multiple of a **chi-square** or **gamma** variate. If there are n values from which the $n - k$ outlier values are omitted, then each of the k mean squares will have r degrees of freedom if the number of replicates is $r + 1$. The proposed estimate is $\hat{\sigma}^2 = \sum_{i=1}^{k} y_i / nb$. The unbiasing factor, b, depends upon the slope of the linearized response curve, the fraction of the sample omitted, the degrees of freedom, r, and the total sample size, n (see [5, Table 13.4A]).

Finney [2] recommends that a reasonable way to evaluate goodness of fit, assuming that test doses are measured in duplicate, is to plot the apparent lack of fit, measured as the difference between the mean of the replicates of each dose and the fitted curve expressed as the percentage change in dose versus log dose on the same graph as the precision profile from the within-replicate dispersion versus log dose. A variance ratio test is used to compare the weighted sum of squares of the expected mean response for each curve to the value predicted if the curve provides a good fit.

In order to check for the validity of parallelism, the test preparations must be evaluated at two or more dose levels. It is customary to evaluate parallelism when an assay is under development, but parallelism may not be verified for subsequent assays. Inclusion of some samples at multiple dilutions is advisable, particularly when any change is made in the assay system, such as using a different reagent. Not infrequently, assays of multiple test preparations will find a lack of parallelism present for only a few of the test specimens. A combined test of parallelism is often employed which incorporates the traditional bioassay test of parallelism with the goodness of fit of the test samples to the assumed response curve.

Other Calibration Methods

Although the logistic model has been most widely used for calibration curves, it has the disadvantage that it does not always provide linearity over the full range of dose concentrations. Alternate approaches have included empirically based curve fitting using **splines**, polynomials, or polygonals, which often provide excellent fits in practice and models based on chemical reactions, some of which have been generalized to allow for multiple binding sites (e.g. [8]). Criticisms of the former are that "while" splines are "well-adapted for smoothing very good data in order to secure internal consistency of interpolation. . . , they are much less suited to estimation from points subject to appreciable experimental error, especially if the precision of that estimation is important to subsequent calculations" [2]. When the binding site concentration and the equilibrium constant are both small, a four- or five-parameter logistic model provides a good fit to the observed data for a wide range of doses, comparing well with the theoretically derived model of McHugh & Meinert [6] and Meinert & McHugh [7] based on biochemical theory. While the McHugh–Meinert theoretical model may generally give somewhat narrower confidence intervals than the corresponding logistic model, a comparison by Raab [9] of the four-parameter logistic model to the four-parameter mass action curve for ten databases found the logistic model to be more robust.

References

[1] Finney, D.J. (1978). *Statistical Methods in Biological Assay*, 3rd Ed. Charles Griffin, London, pp. 328–348.

[2] Finney, D.J. (1983). Response curves for radioimmunoassay, *Clinical Chemistry* **29**, 1562–1566.

[3] Govindarajulu, Z. (2001). *Statistical Techniques in Bioassay*. Karger, New York, pp. 189–206.

[4] Healy, M.J.R. (1979). Outliers in clinical chemistry quality control schemes, *Clinical Chemistry* **25**, 675–677.

[5] Healy, M.J.R. & Kimber, A.C. (1983). Robust estimation of variability in radioligand assays, in *Immunoassays for Clinical Chemistry*, W.M. Hunter & J.E.T. Corrie, eds. Churchill Livingstone, London, pp. 624–626.

[6] McHugh, R.B. & Meinert, C.L. (1970). A theoretical model for statistical inference in isotope displacement immunoassay, in *Statistics in Endocrinology*, J.W. McArthur & T. Colton, eds. MIT Press, Cambridge, Mass., pp. 399–410.

[7] Meinert, C.L. & McHugh, R.B. (1968). The biometry of an isotope displacement immunologic microassay, *Mathematical Biosciences* **2**, 319–338.

[8] Naus, A.J., Kuffens, P.S. & Borst, A. (1977). Calculation of radioimmunoassay standard curves, *Clinical Chemistry* **23**, 1624–1627.

[9] Raab, G.M. (1983). Comparison of logistic and a mass-action curve for radioimmunoassay data, *Clinical Chemistry* **29**, 1757–1761.

[10] Raab, G.M. (1983). Validity tests in the statistical analysis of immunoassay data, in *Immunoassays for Clinical Chemistry*, W.M. Hunter & J.E.T. Corrie, eds. Churchill Livingstone, London, pp. 614–623.

[11] Rodbard, D. (1971). Statistical aspects of radioimmunoassay, in *Principles of Competitive Protein Binding Assays*, W.D. Odell & W.H. Daughaday, eds. Lippincott, Philadelphia, pp. 204–259.

[12] Rodbard, D. (1974). Statistical quality control and routine data processing for radioimmunoassays and immunoradiometric assays, *Clinical Chemistry* **20**, 1255–1270.

[13] Rodbard, D. & Cooper, J.A. (1970). A model for prediction of confidence limits in radioimmunoassays and competitive protein binding assays, in *In Vitro Procedures with Radioisotopes in Medicine*. International Atomic Energy Agency, Vienna, pp. 659–674.

[14] Wallac Oy (1993). *User Guide to Multi Calc Functions*. Part 2: *Standard Curve Theory*. Wallac Oy, Finland.

CAROL K. REDMOND

Radon–Nikodym Theorem

If f is a real-valued function on a measure space $(\Omega, \mathcal{F}, \mu)$ for which the integral $\int_\Omega f \, d\mu$ exists (finite or infinite), then $\nu(B) = \int_B f \, d\mu$ defines another measure with the property that $\mu(B) = 0$ implies $\nu(B) = 0$, i.e. ν is absolutely continuous with respect to μ. When ν is a **probability** measure, the function f is called the density of ν with respect to μ; thus, the existence of a density implies absolute continuity. The Radon–Nikodým Theorem is essentially the converse of this statement, as follows. If $(\Omega, \mathscr{F}, \mu)$ is σ-finite and ν is a (possibly signed) measure on $(\Omega, \mathscr{F})$ which is absolutely continuous with respect to the measure μ, then there exists a real-valued measurable function f on Ω such that $\nu(B) = \int_B f \, d\mu$ for every $B \in \mathscr{F}$. More generally, then, for measurable functions g on Ω, $\int_\Omega g \, d\nu = \int_\Omega fg \, d\mu$ whenever either integral exists. The function f is called the *Radon–Nikodým derivative* of ν with respect to μ and is denoted $d\nu/d\mu$. It is unique up to changes on a set of μ-measure 0. For probability measures, a simplified statement of the theorem is that absolute continuity implies the existence of a density.

When μ is Lebesgue measure on the real line and ν is a probability measure with cdf F, absolute continuity of F (as a function) is equivalent to absolute continuity of ν (as a measure). By the Radon–Nikodým Theorem, this in turn implies that ν has a density.

The Lebesgue Decomposition Theorem gives a unique decomposition $\nu = \nu_{ac} + \nu_s$ of an arbitrary σ-finite measure ν on $(\Omega, \mathscr{F})$ into a measure ν_{ac} which is absolutely continuous with respect to μ and a measure ν_s which is singular with respect to μ [3]. Singularity means there is a set A with $\mu(A) = 0$ which supports all of ν_s, in the sense that $\nu_s(A^c) = 0$. Here A^c denotes the complement of A. The Radon–Nikodým Theorem then insures the existence of a Radon–Nikodým derivative for the absolutely continuous part ν_{ac}.

The theorem was proved by H. Lebesgue in the context of Euclidean space, and then generalized to abstract measure spaces by J. Radon and by O.M. Nikodým.

Applications in Probability and Statistics

Conditional Expectation

In elementary probability the conditional distribution (*see* **Conditional Probability**) of a discrete **random variable** X, given a discrete random variable Y, is given by $\Pr(X = x|Y = y) = \Pr(X = x, Y = y)/\Pr(Y = y)$. For jointly continuous X and Y with joint density $f(x, y)$, the conditional distribution is given by the conditional density $f_X(x|Y = y) = f(x, y)/f_Y(y)$, where f_Y is the marginal density of Y. These distributions determine conditional expectations $\mathrm{E}(X|Y = y)$. Taking Y random yields the random variable $\mathrm{E}(X|Y)$. Averaging the quantities $\mathrm{E}(X|Y = y)$ over y yields the overall average of X, i.e. $\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|Y)]$. More generally, averaging over a subset of the values y yields

$$\mathrm{E}[X1_B(Y)] = \mathrm{E}[\mathrm{E}(X|Y)1_B(Y)], \quad B \in \mathscr{B}, \quad (1)$$

where $\mathscr{B}$ denotes the Borel sets of the real line and 1_B denotes the indicator function of the set B. The Radon–Nikodým Theorem makes possible the definition of the conditional expectation $\mathrm{E}(X|Y)$ for more general joint distributions, where no elementary definition is possible, and without the intermediate construction of an explicit conditional distribution of X given $Y = y$. Specifically, provided that only $\mathrm{E}(X)$ exists, the measure $\nu(B) = \mathrm{E}[X1_B(Y)]$ on $\mathscr{B}$ is absolutely continuous with respect to the distribution μ of Y, and the defining property of $d\nu/d\mu$ states that

$$\mathrm{E}[X1_B(Y)] = \int_{-\infty}^{\infty} \left(\frac{d\nu}{d\mu} \right)(y) 1_B(y) \mu(dy), \quad B \in \mathscr{B}.$$

Therefore, defining $\mathrm{E}(X|Y = y) = (d\nu/d\mu)(y)$ ensures that the fundamental property (1) of conditional expectation still holds in the more general context. The uniqueness of the Radon–Nikodým derivative means that no other definition of conditional expectation could have property (1).

Likelihood Functions and Likelihood Ratios

Given a family $\{P_\theta, \theta \in \Theta\}$ of probability measures which are all absolutely continuous with respect to some underlying measure μ, the Radon–Nikodým

derivatives $dP_\theta/d\mu$ serve as **likelihood** functions, or as **likelihood ratios** if μ is also a probability measure. This is particularly useful when one has observations which are not elements of a finite-dimensional space, such as trajectories of a continuous-time **stochastic process**, where elementary definitions of the likelihood do not apply. The **Neyman–Pearson lemma** says that **hypothesis tests** based on such likelihood ratios are optimal. In comparing two probability measures P_1 and P_2, $\mu = P_1 + P_2$ can be used as the underlying measure.

Girsanov's Formula

When μ is Wiener measure on the space of continuous functions on an interval (i.e. the distribution of **Brownian motion**), and ν is the distribution of Brownian motion with a drift, Girsanov's formula provides an explicit expression for $d\nu/d\mu$. For signal detection in the presence of additive Gaussian white noise (*see* **Noise and White Noise**), this makes possible the construction of a **likelihood ratio test** for the presence of the signal, as the value of the likelihood ratio can be calculated from an observation of the trajectory of the process [5]. Girsanov's formula covers a wide family of processes beyond Brownian motion as well.

Spectral Measures

When ν is the spectral measure of a stationary process in continuous or discrete time, for example, a **time series**, properties of the spectral density $d\nu_{ac}/d\mu$ (*see* **Spectral Analysis**) are important in analyzing mixing properties, smoothness, interpolation, and **prediction** [2, 4]. Here, μ is Lebesgue measure on the real line or on the unit circle.

Local Time

If $X(t)$ is a real-valued stochastic process indexed by time $t \geq 0$, the corresponding occupation measure on $\mathbb{R}$ at time t is given by $\nu_t(B) = \mu(\{s \in [0, t] : X(s) \in B\})$, $B \in \mathscr{B}$, where μ denotes Lebesgue measure. If ν_t is absolutely continuous with respect to μ, the local time $(d\nu_t/d\mu)(x)$ at a point x then describes the relative amount of time the process X spends at x during $[0, t]$, even though the times t for which $X(t) = x$ form a set of Lebesgue measures 0 for each fixed x [1].

References

[1] Chung, K.L. & Williams, R.J. (1990). *Introduction to Stochastic Integration*. Birkhäuser, Boston.

[2] Dym, H. & McKean, H.P. (1976). *Gaussian Processes, Function Theory and the Inverse Spectral Problem*. Academic Press, New York.

[3] Halmos, P.R. (1974). *Measure Theory*. Springer-Verlag, New York.

[4] Rosenblatt, M. (1985). *Stationary Sequences and Random Fields*. Birkhäuser, Boston.

[5] Wong, E. (1979). *Stochastic Processes in Information and Dynamical Systems*. Krieger, Huntington, New York.

KENNETH S. ALEXANDER

Random Coefficient Repeated Measures Model

This topic is concerned with the modeling of data where measurements of one or more attributes are repeated on the same set of individuals over time. Typical applications are to the modeling of the anthropometric growth of children or animals (*see* **Nonlinear Growth Curve**). The model specification will be developed for the case where a single continuous measurement is made on several occasions for a sample. This will then be extended to consider the case of multiple measurements at each time point and mention will be made of extensions to latent variable models and to discrete response data. To begin with, we look at the simple, restricted, data structure where there are a fixed number of measurement occasions and each individual has a measurement on each occasion.

Multivariate Models

Consider the data matrix of responses:

Individual	Occ. 1	Occ. 2	Occ. 3	Occ. 4
1	y_{11}	y_{21}	y_{31}	y_{41}
2	y_{12}	y_{22}	y_{32}	y_{42}
3	y_{13}	y_{23}	y_{33}	y_{43}

The first subscript refers to occasion and the second to individual. We assume multivariate normality and so for the response vector we have initially

$$Y \sim \mathrm{N}(\mu, \Sigma) \tag{1}$$

This constitutes a null model and, in general, we will wish to include further variables, notably age or time. Suppose we wish to express the response; say, a weight measurement, as a linear function of time (t) measured at each occasion. We may then write

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \varepsilon_{ij}, \tag{2}$$

where we allow the intercept and average growth rate to vary across individuals. Suppose, further more, that

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\beta_0} & \\ \sigma_{\beta_{01}} & \sigma^2_{\beta_1} \end{pmatrix} \right] \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2_\varepsilon). \tag{3}$$

We have replaced the general mean and covariance structure given by (1) by the specific structure given by (3). Thus, for example, the **goodness of fit** of (3) can be judged and the model elaborated with suitable **explanatory variables**. Grizzle & Allen [9] provide details of estimation and test procedures.

This multivariate model cannot deal satisfactorily with the typical situation where the spacing and number of measurement occasions are variable and has generally been superseded, except in one or two special cases such as that of latent growth models, mentioned below. We now develop an alternative approach to fitting models such as (2), based upon a *multilevel* model.

The Two-Level Repeated Measures Model

Model (2) and the associated covariance structure (3), as they are written, make no particular assumptions about the number or spacing of measurement occasions and, in fact, constitute a special case of a two-level model (*see* **Multilevel Models**). Level 1 units are the measurement occasions and level 2 units are individuals. All the usual procedures for estimation and inference in such models are therefore available, including cases of multivariate responses, nonlinear models, etc. We can additionally consider individuals as nested within further hierarchies; say, animal litters or schools for students and cross-classifications may also occur.

A consequence of (2) is that measurements made on the same individual are correlated through the sharing of common intercept and slope parameters, and it is this dependency that leads to the inadequacy of simple estimation procedures, for example, based upon ordinary **least squares**. Furthermore, interest will usually lie just as much in the covariance matrix estimates as in the average growth parameters. We may also wish to form posterior mean estimates of the individual growth parameters (β_{0j}, β_{1j}) and we shall illustrate below how these can be used for efficient prediction. As in the general multilevel model case, we may have a **Bayesian** formulation for the model with prior distributions upon the parameters (see, for example, [1]).

In the following sections we shall consider in more detail nonlinear models, multivariate response

models with more than one response on each occasion and complex structures for the level 1 residuals. For a detailed exposition of further aspects of these topics and some alternative approaches, as well as a discussion of issues related to informatively missing data and transition-type models, the reader should consult Diggle et al. [4]. In particular, these authors consider the so-called population average model, where interest centers on the estimation of the fixed or average component of (2) (*see* **Marginal Models**). This often allows simplified estimation procedures to be used with no requirement for the separate estimation of the random components. This may be appropriate in certain circumstances, but, since it ignores the specific nature of repeated measurements data, is not considered further here.

Nonlinear and Generalized Linear Models

Most attempts to fit nonlinear models to repeated measurements have fitted separate curves to each individual's set of measurements and then combined these to describe the between-individual variation (*see* **Nonlinear Growth Curve**). A major problem with this approach is that it requires many measurements on each. Also, while nonlinear curves have been used successfully to describe change, for example, in **pharmacokinetic** studies, in other areas, such as growth, they can also impose inflexible relationships among growth events that are not empirically supported [5].

Bock et al. [2] describe a maximum likelihood analysis of a human growth model using the superimposition of three logistic functions. Lindstrom & Bates [13] describe an approximate estimation procedure for nonlinear models and Goldstein [7] gives an example using the so-called Jenss–Bayley [10] curve for children aged 5 to 10 years (*see* **Growth and Development**). Davidian & Giltinan [3] give a detailed discussion of different approaches to the fitting of nonlinear models to repeated measures data.

Where the response is discrete; for example, binary or ordered as in the case of recording developmental stages over time, then a generalized linear model will be appropriate. Consider the following example where each individual, j, is measured several times, t, and their nutritional state, y, at occasion i is classified as adequate (1) or inadequate (0). A standard model would be written as

$$\begin{aligned} \text{logit}\ \{\pi_{ij}\} &= a_j + b_j t_{ij}, \\ \pi_{ij} &= \Pr(y_{ij} = 1), \\ y_{ij} &\sim \text{bin}\ (\pi_{ij}, 1). \end{aligned} \tag{4}$$

This expresses the logit of the probability of having an adequate nutritional state as a linear function of time. Such a model might be appropriate, for example, in evaluating a nutritional intervention programme and further covariates for group membership, age, etc. can readily be introduced. We can also try alternative link functions and study the possibility of further random coefficients. For responses such as counts we would typically use a log link with a Poisson or related distributional assumption.

For many longitudinal data we are effectively measuring the *cumulative* probability of a response over time. Thus, when studying the onset of menarche the probability of occurrence is an increasing function of time and successive observations will consist of a string of zeros (nonoccurrences) followed by a string of ones (occurrences). More generally, we will have an ordered sequence of stages through which all individuals pass and (4) will be modified to reflect this. One such **proportional hazards** model can be written as

$$\gamma_{ij}^{(s)} = \{1 - \exp[-\exp(\beta_0 + \beta_1 t_s)]\}, \quad \beta_1 > 0, \tag{5}$$

where s indexes the stages and the cumulative probability is

$$\gamma_{ij}^{(s)} = \sum_{h=1}^{s} \pi_{ij}^{(h)}.$$

We can add further covariates and random coefficients as before.

An extension of both (4) and (5) is to the multivariate case where multiple responses are measured on each individual at each time point, with possibly missing responses on some occasions and where some responses are discrete and some continuous. A discussion of such models and their estimation is given in [7, Chapter 7] and the multivariate model with continuous-only responses is discussed in the next section.

In these models, so far, we have made the basic assumption that the level 1 errors are independent.

We shall deal with violations of this assumption for continuous responses below, but there are also many cases for discrete responses where this assumption is untenable and this gives rise to particular difficulties. As an example, consider a repeated survey of attitudes to abortion where we wish to study the characteristics of individual and group changes over time. For a large proportion, perhaps the majority, of the population there will be no change in their attitudes; thus, the probability that they will agree with a "pro-abortion" statement will be very close to one or zero. A model such as (4) would generally require such individuals to have extremely large positive or negative random effects, since it is unlikely that we would have covariates that could discriminate precisely among such individuals. This, then, poses severe distributional problems for parametric models.

An obvious way to avoid this difficulty is to consider the vector of, say, binary responses for each individual as a multivariate vector where the distribution on each occasion is binomial and the between-occasion covariances are estimated from the data. Lipsitz et al. [14] study such models with examples. While this approach is satisfactory for a number of fixed occasions, even with missing data, and while it can also be extended to other than binary responses, it is unable directly to handle the general case of arbitrary occasions. To do this requires an extension of the serial correlation models discussed below, but that is beyond the scope of this article.

Multivariate Continuous Responses

Where several responses are recorded on individuals on each occasion we will generally wish to model the average time relationship for each response and the covariance matrix among the responses as a function of time. This is readily done by considering the multivariate response structure as a further, lowest, level in the data hierarchy with measurements nested within occasions within individuals (*see* **Multilevel Models**).

There are several advantages to considering the joint modeling of several responses. The ability to estimate their covariance matrix as a function of time allows one to study the distribution of any function of the responses with respect to time. For example, when studying issues of prior determination it may be useful to see whether the correlation between two variables a given time apart is greater when one is the prior variable rather than the other. Likewise, it provides a general prediction procedure for one measurement, conditional on any set of observed prior measurements. We illustrate this with an example concerned with the prediction of adult height given a series of height measurements taken during a period of childhood growth. In this case, one of our response measurements, adult height, is made at the level of the individual and the others are made at the occasion level.

Consider the following extension to (2):

$$
\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j} t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \varepsilon_{ij}, \\
y_j &= \sum_k \gamma_{jk} x_{jk} + \alpha_j, \qquad (6) \\
\begin{pmatrix} \alpha_j \\ \beta_{0j} \\ \beta_{1j} \end{pmatrix} &\sim \mathrm{N}\left[\begin{pmatrix} \alpha \\ \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & & \\ \sigma_{\alpha\beta_0} & \sigma_{\beta_0}^2 & \\ \sigma_{\alpha\beta_1} & \sigma_{\beta_{01}} & \sigma_{\beta_1}^2 \end{pmatrix} \right],
\end{aligned}
$$

where the adult measurement, y_j, is allowed to depend on further covariates, and we may also wish to incorporate covariates into the growth period component of the model. The key feature of this model is that we have a joint covariance matrix for the adult height component and the growth curve parameters, all of which vary at the level of the individual. Given the model parameters and any set of growth measurements for an individual, say $Y_j^* = (y_{1j}^*, y_{2j}^*, \ldots, y_{pj}^*)$, we can estimate $\mathrm{E}(y_j | Y_j^*)$ together with an estimate of its standard error, etc. Details are given in [6]. A further development of the multivariate growth model is the so-called "latent growth model". In essence, this considers each of the sets of random coefficients β_{0j}, β_{1j}, etc. as a latent variable or factor. Each observed response is thus a linear function of factor scores where the coefficients are, for example, polynomials in time, or, more generally, may be estimated from the data. One restriction of such models is that they require the same set of discrete occasions for all measurements and thus lose the flexibility of the continuous time formulation. A full discussion can be found in [15].

Serial Correlation Models

For some kinds of repeated measurements, the structure implied by (2) or (4) is inadequate. For example, daily measurements of animal weights over a long period will not usually fluctuate completely randomly about a long-term smooth trend for each animal, the departure from such a trend on any one day being more like the departures on neighboring days than on days further distant. In a study of human growth, Goldstein et al. [8] found that residuals from measurements of height made on adolescent boys had a noticeable **serial correlation** when made less than three months apart. To incorporate such possibilities, we can extend (2) by adding the following covariance condition for two level 1 residuals s time units apart, where time is continuous:

$$\mathrm{cov}(\varepsilon_t, \varepsilon_{t-s}) = \sigma_\varepsilon^2 \exp(-g(s, z)). \qquad (7)$$

Here, g is a positive function and may depend on covariates, z, which may be measured at the individual or occasion level; thus, for example, allowing the exponential decay rate implied by (7) to vary with time.

One possible simple choice, which is the continuous time analog of a first-order autoregressive series, is $g = \alpha s$ and other possibilities are discussed by Goldstein et al. [8] and Diggle et al. [4, Chapter 5]. An alternative approach is via state-space modeling, which leads to similar, although not generally identical, models [11].

In Table 1, we give an example of a model with a simple correlation structure, together with the estimation of a seasonal effect for a set of three-monthly height measurements made on a sample of 26 boys aged between 11 and 14 years. Full details are given by Goldstein et al. [8]. A fourth-degree polynomial is fitted for the average growth curve with a cosine term representing seasonal growth. The first three coefficients are random at level 2 and the serial covariance structure is given by $g = \alpha s$ fitted at level 1.

The serial correlation parameter value of 6.9 implies that the residual correlation three months apart is 0.19 and that six months apart is 0.04. The existence of a seasonal effect implies an average difference between summer and winter of about 0.5 cm with no evidence of any variation between individuals.

Table 1 Height in centimeters as a fourth-degree polynomial on age, measured about 13.0 years. Standard errors in parentheses; correlations in parentheses for covariance terms. Serial correlation structure fitted for level 1 residuals

Parameter	Estimate (se)		
Fixed			
Intercept	148.9		
age	6.19(0.35)		
age^2	2.16(0.45)		
age^3	0.39(0.17)		
age^4	−1.55(0.43)		
cos (time)	−0.24(0.07)		
Random	Intercept	age	age^2
Level 2			
Intercept	61.5(17.1)		
age	7.9 (0.61)	2.7(0.70)	
age^2	1.5 (0.25)	0.9(0.68)	0.6(0.2)
Level 1			
σ_ε^2	0.23(0.04)		
α	6.90(2.07)		

Fitting this model, with an extra parameter to describe autocorrelation among the level 1 residuals, provides a more parsimonious model than attempting to fit, say, the cubic coefficient as random at level 2. In some cases, however, the data may be equally well explained either by such a random coefficient model with independent level 1 residuals or, alternatively, by a simpler between-individual covariance structure and a complex nonindependence structure at level 1. A choice between such models will then need to be made on grounds of substantive interpretation. In the view of the present author, substantive interpretations generally are best made by adopting a level 1 serial correlation structure only after fitting a suitably complex model using random coefficients alone. The use of various diagnostic tools for judging fit in multilevel models is discussed by Lewis & Langford [12].

Software

Some of the particular models described (for example, the nonlinear model of Bock et al. [2], the latent growth model or the Bayesian models) have specialized software, details of which can be found by consulting the references given at the end of the article. Some of the major **software** packages, most notably SAS, can handle many, although not all, of the

models and the **generalized estimating equations** procedures used by Diggle et al. [4] are available in S-PLUS. The general-purpose multilevel modeling package, MLn (Rasbash & Woodhouse [16]) uses both maximum likelihood and quasi-likelihood estimation and has facilities to analyze all the models described, although it can only handle the latent growth model indirectly by providing summary input for other structural equation software packages.

References

[1] Best, N.G., Spiegelhalter, D.J., Thomas, A. & Brayne, C.E.G. (1996). Bayesian analysis of realistically complex models, *Journal of the Royal Statistical Society, Series A* **159**, 323–342.

[2] Bock, R.D., Du Toit, S.H.C. & Thissen, D. (1994). *AUXAL: Auxological Analysis of Longitudinal Measurements of Human Stature*. Scientific Software International, Chicago.

[3] Davidian, M. & Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London.

[4] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

[5] Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies*. Academic Press, London.

[6] Goldstein, H. (1989). *Efficient prediction models for adult height, Auxology 88; Perspectives in the Science of Growth and Development*, J.M. Tanner, ed. Smith-Gordon, London, pp. 41–48.

[7] Goldstein, H. (1995). *Multilevel Statistical Models*. Arnold, London; Halsted Press, New York.

[8] Goldstein, H., Healy, M.J.R. & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data, *Statistics in Medicine* **13**, 1643–1655.

[9] Grizzle, J.C. & Allen, D.M. (1969). An analysis of growth and dose response curves, *Biometrics* **25**, 357–361.

[10] Jenss, R.M. & Bayley, N. (1937). A mathematical method for studying the growth of a child, *Human Biology* **9**, 556–563.

[11] Jones, R.M. (1993). *Longitudinal Data with Serial Correlation: A State-space Approach*. Chapman & Hall, London.

[12] Lewis, T. & Langford, I. (1996). Outliers in multilevel data (unpublished paper).

[13] Lindstrom, M.J. & Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics* **46**, 673–687.

[14] Lipsitz, S.R., Fitzmaurice, G.M., Sleeper, L. & Zhao, L.P. (1995). Estimation methods for the joint distribution of repeated binary observations, *Biometrics* **51**, 562–570.

[15] Muthen, B. (1995). *Longitudinal Studies of Achievement Growth Using Latent Variable Modeling*. University of California Press, Los Angeles.

[16] Rasbash, J. & Woodhouse, G. (1995). *MLn Command Reference*. Institute of Education, London.

(*See also* **Longitudinal Data Analysis, Overview**; **Linear Mixed Effects Models for Longitudinal Data**; **Nonlinear Mixed Effects Models for Longitudinal Data**)

HARVEY GOLDSTEIN

Random Digit Dialing Sampling for Case–Control Studies

Random digit dialing (RDD) is a method of sampling households through the selection of telephone numbers by a random choice of the digits in the telephone numbers. RDD was initially developed as a sampling method for household surveys, but it is now considered a useful tool in epidemiologic research, particularly for selecting **controls** in a **population-based case–control study**, i.e. studies using controls selected from the general population, as distinct from hospital controls or other specialized lists. In countries with high levels of telephone coverage, RDD can provide an almost **unbiased** sample of the household population for use as controls. Furthermore, once the households are contacted by telephone, there is a relatively low cost of screening to locate persons with the demographic and health characteristics that match the cases for the disease being studied.

RDD obviously omits residents of households that do not have telephones. In the US only about 5% are without telephones, but they are mainly very low-income households, and if the causal factors for the disease are believed to be heavily influenced by income, the results could be seriously **biased**. For such studies the researcher should consider whether RDD is appropriate. However, in most case–control studies, the exclusion of nontelephone households is not believed to have any appreciable effect. It is prudent to exclude cases who do not have home telephones from the analysis so that cases and controls have similar socioeconomic characteristics.

In the US each telephone contains 10 digits, consisting of a three-digit area code, a three-digit central office code (generally referred to as an exchange), and a four-digit suffix. There are presently 35 000–40 000 active area code/central office code combinations in use in the US. Since 10 000 possible number combinations exist in the four-digit suffixes attached to each area code/central office, there are a little over 350 000 000 possible telephone numbers that can be dialed in RDD, about four times the approximately 90 000 000 households with telephones. Most of the 75% of telephone numbers that do not connect to households are unassigned numbers; others are connected to businesses, government offices, institutions, pay phones, computers, faxes, etc.

With **simple random sampling**, about three-quarters of the calls will be completely unproductive, consisting of nonworking or nonhousehold numbers. The costs of such unproductive calls are quite high, and statisticians have examined the properties of a number of sampling methods designed to reduce the number of excess calls. The general consensus among sampling and survey statisticians is that currently one of two available methods should be used – what is referred to as the *Mitofsky–Waksberg Procedure* [13], or the *List-Assisted Method* [3, 8]. Both methods increase the proportion of household numbers in the sample from about 25% to 50%–60%. Brick et al. [2] note that although the list-assisted method is slightly biased, the bias is trivial for most practical purposes. The Mitofsky–Waksberg procedure provides a completely unbiased sample of households but has several operational complications that can affect the timing and the required record-keeping. (*See* **Telephone Sampling** for a description of the two sampling methods and a list of references.)

The cheapest method of sampling is to use *directory sampling*, that is, select a simple random sample from a **sampling frame** of all telephone numbers listed in the white pages of current telephone directories for the geographic area of interest. In the US, this procedure will produce a highly biased sample since only about two-thirds of telephone households have their current numbers listed. It is not recommended by knowledgeable statisticians. If used, the biases can be reduced, although not eliminated, by restricting the cases in the study to those whose telephone numbers are listed in telephone directories. It should be noted that although restricting both cases and controls to households listed in telephone directories appears to be similar to RDD in that the cases and controls come from similar populations, the population eligible for the study with RDD is about 95% of all households as compared with about 60% for directory sampling. The potential for bias is thus generally quite small for RDD studies but can be substantial for directory sampling. For this reason, directory sampling is usually not recommended.

There are, at present, at least two commercial firms that maintain up-to-date records on existing area codes/central office codes in the US (GENESYS Sampling Systems and Survey Sampling Inc.). These companies can select samples using either of the two

recommended methods, or following other procedures specified by the client. Both companies have extensive resources for geographic coding and can help identify the area codes/central office codes in the geographic area designated for a study. Statisticians and epidemiologists who have not had extensive experiences with RDD studies would probably benefit by consulting one of these companies for assistance in sampling.

Potthoff [9] has described the steps in selecting controls. We summarize and extend his discussion. The cases are all of the eligible ones that occur in a given geographical region in a given time period. RDD controls are sampled from the same region, typically by strata (demographic categories such as sex, age, race) (*see* **Stratified Sampling**). Frequently, geographic subareas are included in the matching criteria; the subareas may be defined by census tracts, zip codes, telephone exchanges, or broader geography such as counties or central cities vs. suburbs. The areas may serve as proxies for environmental or socioeconomic matching, but they should be broad enough to produce the required number of controls. In general, everyone within a stratum (often referred to as a matching cell) is given the same chance of selection. The sample sizes are designated in advance for the strata, the size being based on the desired number of controls per case, the available budget, and possibly logistical considerations. Most population-based case–control studies use **frequency matching** of controls to cases within strata rather than individual matching. Frequency matching involves defining matching cells, for example, white females 40–44 years of age, and locating the desired number of controls in each matching cell. RDD selection can be used for both individual and frequency matching.

Telephone contacts can be used to select controls and to conduct interviews, or only for the sample selection, in which case subsequent interviews may be carried out in personal visits to the households or, more rarely, by mail. The decision on which approach to use depends on the length and content of the interview, e.g. whether physical measurements or laboratory tests are necessary, whether observation of such items as prescription drugs is required, etc. The same general principles are used for both types of studies. If personal visits will be required, then the researcher should take this into account in establishing the geographic area in which the study will be conducted in order to avoid excessive interviewer travel. Methods of sampling that tend to cluster the sampling units, e.g. the Mitofsky–Waksberg method, should be considered.

With telephone interviewing, the selection and enlistment of controls and the interviewing can be done in a single telephone contact [5, 6]. Alternatively, a two-step process can be used in which the first call is restricted to taking a household census with sample selection carried out as an office operation, and a second telephone call made for a telephone interview or to arrange for a visit [6]. If the time schedule permits, then the two-step process is usually preferable since it provides a tighter control on sample size. There is difficulty in achieving the exact sample sizes in a one-step process without incurring biases from loss of persons who are infrequently at home and require multiple attempts to reach. However, some researchers prefer the one-step method because it usually produces a somewhat higher response rate. In addition, if there is a long lapse of time between the two contacts, then a non-trivial part of the sample will have moved, introducing the potential for important biases in the study as well as some uncertainty in the number of controls that will be interviewed. If a two-step process is used, then the time period between the two steps should be kept as short as possible.

A number of issues arise in the sample selection via RDD. They are discussed briefly below. More complete discussions can be found in the references, particularly [5–7], [11], and [12].

1. RDD covers only persons living in households with telephones. It is normal practice to exclude from the analysis the cases that do not have telephones or who do not reside in ordinary households, e.g. institutions or the military. A clear definition of households and household members needs to be established to treat such ambiguous cases as college students, persons in the military, etc.
2. The geographic areas that are used to determine the boundaries of the study, or the matching cells, frequently consist of political units (counties or cities) or census tracts. These areas may not conform to telephone exchanges, leading to a considerable amount of screening to identify persons within the designated geographic area. To avoid this excessive screening, the researcher should consider the possibility of defining the

study area as the set of telephone exchanges that approximates the geography desired. Matching cells can also be defined by telephone exchanges. The same geographic rules need to be applied to cases as to controls.

3. If the boundaries of the study area are political or census-defined geography, then it is necessary to include all telephone exchanges that cover the area. If the geographic location is not obvious from the telephone exchange, then each respondent should be asked whether the residence is within the boundaries of the study area.
4. In case–control studies, it is usually desirable to select controls with equal probability. Two problems occur in RDD sampling. The first is that households with two or more telephone numbers have greater chances of selection than those with one number. The higher probability of selection of such households can be avoided by subsampling them, i.e. retaining one-half of households with two telephone numbers, one-third of those with three numbers, etc. This adds somewhat to the number of households that need to be screened to reach the required sample size (about 4% of US, households have more than one telephone number), but the resulting simplification in the analyses usually makes it worthwhile. The second problem arises when it is considered undesirable to choose more than one person in a household, either to avoid a heavy response burden in any household or because the intraclass correlations (*see* **Correlation**) within a household would appear to complicate analyses of the data. Subsampling within households will create considerable variation in probabilities of selection and is not recommended. Many case–control studies are restricted to the adult population, or to a subset, such as ages 45–69. In such cases it is possible to reduce sharply the number of households with multiple controls by designating in advance half of the sample for male controls and half for female controls. In any household, only the males, or females, are then eligible for the sample. Depending on the distribution of the desired number of controls by sex, instead of designating 50% of the sample for male controls and 50% for females, a 60–40, 70–30, or some other ratio could be used. This sex designation of the sample greatly increases the amount of screening necessary to locate the controls, and a researcher should weigh this fact against the desirability of simpler analytic methods. Potthoff [9] has described another method of choosing only one control per household with an equal probability sample, but this method also increases the required screening.
5. With frequency matching, the amount of screening that will be required can be estimated by calculating the number of households that need to be screened to locate one required control in each of the matching cells (strata), and multiplying it by the number of desired controls. The stratum with the largest value determines the screening sample size. This result should be increased to account for the percentage of refusals or those who are out of scope (e.g. persons with some types of chronic diseases may be considered ineligible for the study). It may also be necessary to increase the screening if no more than one person per household is chosen and one of the devices described earlier for doing this is used.
6. It is frequently difficult to determine the total number of households that will need to be screened to provide the desired number of controls, and the subsampling rates for the various strata. Some of the difficulty comes from the small samples desired per matching cell and the consequent large sampling errors on the household yield. Another reason may be uncertainty of the population size of the various strata when geographic matching is required. When time permits, it is useful to do the sampling in waves, (e.g. divide the workload into monthly subsamples) with each wave a **random sample** of the population. The sample sizes and subsampling rates in each wave can be based on experience in previous waves. Since the waves are all random samples, the data can be pooled without the need to take the different sample sizes and rates into account.
7. It is necessary to inquire whether the telephone number reached is for a home, business, institution, or some other nonhousehold facility. This is normally one of the first questions asked. Some small businesses operate in residential units and the telephones are used both for business and personal use. The questions asked

should be able to identify such cases and retain them in the sample. Both companies previously referred to that supply telephone samples can match the sample numbers with yellow page directory listings to reduce the number of business numbers that need to be dialed. There is a small loss of households that operate small businesses, but the bias is generally considered trivial.

8. A policy needs to be established on how to treat answering machines, that is, whether to leave messages or simply to hang up and try again. Answering machines are quite common in the US, and the method of dealing with them could have an important effect on response rates.
9. The study should include provision for a considerable number of callbacks to insure that persons who are infrequently at home have the same chance of selection as the rest of the population. Many researchers make eight to 12 calls for the hard-to-get population spread throughout daylight and evening hours and weekdays and weekends, and some researchers use larger numbers. As a result of the large number of households in which all adults are employed and the number containing only a single adult, it is necessary to mount a major follow up operation to attain a reasonable response rate. There have been a number of studies of the differences between persons who can be easily reached by telephone and those requiring more effort, and they have revealed important differences in occupational status and life styles. There are also likely to be differences in exposure and in background variables affecting many diseases. Failure to achieve reasonably consistent response rates among persons who are easily reached and those who require more callbacks could lead to serious biases that are not possible to detect and correct for. Although a main purpose of the callbacks is to make first contact with the household residents, it can also be used to try to convince potential respondents who initially refused to cooperate and to change their minds. Conversion of about one-third of those who initially refused is not uncommon (*see* **Call-backs and Mail-backs in Sample Surveys**).
10. Evening calls are necessary in a large number of households, and researchers should recognize that most of the interviewing will have to be done after normal working hours. A common practice is to start off with one round of daytime calls to the entire sample which will usually identify almost all of the nonworking, business, and institutional numbers, and a minority of the households. The remaining calls – mostly to households – are then made in the evening.

Under some circumstances, weighting the data may be necessary to avoid biases in the analysis. Weighting will be needed if all persons within a matching cell did not receive the same chance of selection, e.g. if there was subsampling of household members within the same matching cell, or if all households with more than one telephone number were retained in the sample. For example, if households with multiple telephones were not subsampled, then households with two telephone numbers should be given a weight of one-half, those with three numbers given a weight of one-third, etc. Similarly, with subsampling within households so that only one person within a cell was chosen, the weight should be the number of household members within the matching cell. If the subsampling went further so that only one person per household was selected for the control regardless of the number of matching cells represented in a household, then a more complex system of weighting will be required. Such subsampling should be avoided if at all possible since it will complicate the study operations and may have a serious effect on the precision of the results.

Standard **odds ratio** analysis is not strictly applicable in case–control studies that use **cluster sampling** such as the Mitofsky–Waksberg method or that require weighting. Modifications in the analysis that account for clustering are described by Graubard et al. [4]. Appropriate **confidence intervals** for odds ratios can also be obtained through use of either of two software packages originally developed for analyses of surveys using complex sample designs but that also contain provision for estimating the precision of odds ratios, WESVAR [1] and SUDAAN [10]. If there are only minor deviations from an unclustered, equal-probability sample selection scheme, then it is probably satisfactory to use the more common methods of establishing confidence intervals around odds ratios.

References

[1] Brick, J.M., Broene, P., James, P. & Severynse, J. (1996). *A User's Guide to WesVarPC*. Westat, Rockville.

[2] Brick, J.M., Waksberg, J., Kulp, D. & Starer, A. (1995). Bias in list-assisted telephone samples, *Public Opinion Quarterly* **59**, 218–235.

[3] Casady, R.J. & Lepkowski, J.M. (1993). Stratified telephone survey designs, *Survey Methodology* **19**, 103–113.

[4] Graubard, B.I., Fears, T.R. & Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case–control studies, *Biometrics* **45**, 1053–1071.

[5] Harlow, D.L. & Davis, S. (1988). Two one-step methods for household screening and interviewing using random digit dialing, *American Journal of Epidemiology* **127**, 857–863.

[6] Hartge, P., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., Hoover, R.N. & Waksberg, J. (1984). Random digit dialing in selecting a population-based control group, *American Journal of Epidemiology* **120**, 825–833.

[7] Hartge, P., Cahill, J.I., West, D., Hauck, M., Austin, D., Silverman, D. & Hoover, R. (1984). Design and methods in a multi-center case–control interview study, *American Journal of Public Health* **74**, 52–56.

[8] Lepkowski, J.M. (1988). Telephone sampling methods in the United States, in *Telephone Survey Methodology*, R.M. Groves, P.P. Biemer, L.E. Lyberg et al., eds. Wiley, New York, pp. 73–98.

[9] Potthoff, R.F. (1994). Telephone sampling in epidemiologic research: to reap the benefits, avoid the pitfalls, *American Journal of Epidemiology* **139**, 967–978.

[10] Shah, B.V., Barnwell, B.G., Hunt, P.N. & LaVange, L.M. (1992). *SUDAAN User's Manual*. Research Triangle Institute, Research Triangle Park.

[11] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case–control studies. I. Principles, *American Journal of Epidemiology* **135**, 1019–1028.

[12] Wacholder, S., Silverman, D.T., McLaughlin, J.K. & Mandel, J.S. (1992). Selection of controls in case–control studies III. Design options, *American Journal of Epidemiology* **135**, 1042–1050.

[13] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **73**, 40–46.

JOSEPH WAKSBERG

Random Effects

Consider an **explanatory variable** which takes on k possible values in a particular data set and which is to be related to a **response variable** via a **regression** model. Assume that some function of the response variable is related to the linear predictor $\mu + \alpha_i$, $i = 1, \ldots, k-1$, or equivalently α_i, $i = 1, \ldots, k$, where i indexes the possible values of the explanatory variable (*see* **Dummy Variables**). Examples of such explanatory variables are an indicator for clinics in a multiclinic study, school classrooms in a study of school children, different studies in a **meta-analysis**, and blocking factors in **experimental design**. Other explanatory variables may be included in the linear predictor. For example, a single additional variable could be added to define a linear predictor $\mu + \alpha_i + \beta X, i = 1, \ldots, k-1$.

The α_is are referred to as random effects if they are assumed to arise as a **random sample** from a distribution of effects associated with a wider range of values for the explanatory variable. For example, it might be assumed that there is a distribution of effects, $f(\alpha, \theta)$, associated with all the possible clinics which could have been recruited for a **clinical trial** or perhaps, more generally, which might be expected to use the treatments under study in the trial. The parameter α might be some measure of central tendency, sometimes taken to be zero, and θ might represent a shape parameter such as a variance. The analysis would then not focus on the particular α_is but, rather, on the characteristics of the distribution f. This analysis contrasts with that which regards the α_is as **fixed effects** (for general discussion of the distinction and of its effect on the methods of analysis *see* **Analysis of Variance**). The choice of analysis can be quite influential in, for example, meta-analysis.

With an additional variable in the linear predictor, it is possible to extend the model further to have, say, $\alpha_i + \beta_i X, i = 1, \ldots, k$, where the regression coefficient for X varies with the value of the indicator variable. The β_is can be fixed or random. When they are assumed random, then the model is sometimes called a *random coefficient* model. Such an approach is used in **multilevel models**.

Random effects are also important in a variety of **Bayesian methods**.

(*See also* **Random Coefficient Repeated Measures Model**)

VERN T. FAREWELL

Random Error

Suppose one seeks to measure the width of a table by repeatedly applying a ruler to obtain a series of measurements. One assumes that the true table width, μ, remains constant and that any given measurement, $y_i = \mu + e_i$, represents the sum of a systematic component and an error component, $e_i = y_i - \mu$. In this simple model, μ is called the systematic part, and, if the expectation of e_i is zero, e_i is called the random error. If it is assumed that the random errors are independent, then the **mean** value $\overline{y} = \mu + n^{-1} \sum e_i$ converges (almost surely) to the true value (*see* **Convergence in Distribution and in Probability**). As the sample size, n, increases, the effects of random error diminish, and, in particular, if the e_i have a common **variance**, σ^2, $\text{var}(\overline{y}) = \sigma^2/n$. The diminishing influence of random error with increasing sample size is also found in more general models with systematic and random components (*see* **Generalized Linear Model**).

Suppose, however, we subsequently learn that our ruler had been worn down so that the putative interval [0, 1] cm was only 0.9 cm long. Then, the errors $e_i = y_i - \mu$ have expectation $(\mu - 0.1) - \mu = -0.1$ cm, and $\overline{y}$ gets closer and closer to the **biased** answer, $\mu - 0.1$, as n increases. Thus, increasing the sample size offers no protection against **systematic error** and only leads to more precise biased estimates (*see* **Estimation**).

MITCHELL H. GAIL

Random Mixing

The simplest models for epidemics of infectious diseases are nonlinear owing to their mass action terms. The mass action component of these models accounts for new infections through contacts between infected and susceptible persons, usually through the assumption that "... the chance of an infection is proportional to the number of infected on the one hand, and to the number not yet infected on the other" [3, p. 703]. Random mixing is an interpretation of contact patterns which justifies that assumption (*see* **Epidemic Models, Deterministic**).

In the elementary Kermack & McKendrick [3, p. 713] epidemic model for a closed population, the rate of new infections is written $\mathrm{d}x/\mathrm{d}t = -\kappa xy$, where x and y are the numbers susceptible and infected, respectively. Decomposing κ into β, the chance of infection per contact, times μ, the rate of contact per infected/susceptible pair per unit time, reveals the basic structure of the assumed contact process: there are xy pairs capable of producing a new infection in the population, each of which has the same rate of contact, μ, per unit time; a fraction β of those contacts produces new infections. Alternatively, define c as the contact rate per person per unit time and let p_y be the proportion of contacts that are with infected persons. Then $\mathrm{d}x/\mathrm{d}t = -\beta cxp_y$. If contacts are randomly distributed, then the proportion of contacts with infected persons (in the early stages of an epidemic when the population size $n \approx x + y$) is simply the proportion infected in the population, i.e. $p_y = y/n$. Since the contact rate per person is n times the contact rate per pair, $c = \mu n$, the two formulations of the mass action term are equivalent.

Epidemic models based on simple random mixing are often inconsistent with elementary facts about human populations. This is especially true with respect to diseases associated with sexual practices, including HIV/AIDS (*see* **AIDS and HIV**), where contacts that transmit disease are strongly influenced by variations in customs, preferences, and opportunities. To build models for such cases, epidemiologists have invented generalizations of random mixing on the basis of the idea that a population can be divided into groups characterized by different rates of contact and/or different patterns of mixing with other groups. In terms of the pair-rate notation, the group-specific infection rate in these more general epidemic models can be written as

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = -\beta \sum_j \mu_{ij} X_i Y_j,$$

where x_i and y_i are the number susceptible and infected in group i and μ_{ij} is the rate of contact per i, j pair per unit of time. This class of mixing models includes decidedly nonrandom patterns as well as generalizations of random mixing. Among the latter, we include the following:

1. Wiley & Herschkorn [7] propose *quasi-random mixing* for sexual contacts among male homosexuals at risk of HIV transmission. This model is specified by setting $\mu_{ij} = \mu$ for $i, j \in S$ and $\mu_{ij} = 0$ for $i, j \notin S$, where S is a set that defines admissible pairings. This specification is motivated by the concept of inconsistent roles in sexual intercourse.
2. Hethcote & Yorke [2] build epidemic models for gonorrhea transmission that incorporate *proportional random mixing* ($\mu_{ij} = \mu_i \mu_j$), allowing groups to mix randomly but with heterogeneous contact rates.
3. Koopman et al. [4] generalize proportional random mixing to allow for a bias toward within-group mixing. They call this pattern *reserved mixing*, and it is specified in the pair-rate notation by setting $\mu_{ij} = \mu_i \mu_j$ for $i \neq j$ and $\mu_{ij} = \mu_i^* + \mu_i^2$ for $i = j$.

More complex patterns of mixing can be formed as sums of random mixing patterns in different social or physical settings (see, for example, [6]). For further reading, see [1] and [5].

References

[1] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[2] Hethcote, H.W. & Yorke J.A. (1984). *Gonorrhea Transmission Dynamics and Control*, Lecture Notes in Biomathematics. Springer-Verlag, New York.

[3] Kermack, W.O. & McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society, Series A* **115**, 700–721.

[4] Koopman, J., Simon, C., Jacquez, J., Joseph, J., Sattenspiel, L. & Park, T. (1988). Sexual partner selectiveness effects on homosexual HIV transmission dynamics,

Journal of Acquired Immune Deficiency Syndromes **1**, 486–504.

[5] Mollison, D., ed. (1995). *Epidemic Models: Their Structure and Relation to Data*. Cambridge University Press, Cambridge.

[6] Sattenspiel, L. (1987). Population structure and the spread of disease, *Human Biology* **59**, 411–438.

[7] Wiley, J.A. & Herschkorn, S.J. (1989). Homosexual role separation and AIDS epidemics: insights from elementary models, *Journal of Sex Research* **26**, 434–449.

JAMES WILEY

Random Sample

A collection of units or observations generated by a probabilistic process is called a random sample. This apparently simple definition covers a multitude of situations.

In finite population sampling, the sample is selected from a fixed and finite population of units. Typically, the sample units are those for which data are collected. A random sample (or **"probability sample"**) is drawn by any rule that assigns a probability to every possible sample. A common additional requirement in this context is that every unit in the population has a nonzero probability of being included in the sample. One such sampling scheme is **simple random sampling** (SRS), in which the size of the sample, n, is predetermined and every sample of size n has equal probability. The term "random sampling" is loosely used for SRS, but in fact is more general, including sampling schemes in which different samples have different probabilities; furthermore, the probability of inclusion in the sample is not necessarily the same for every unit.

We may also speak of a sample from a probabilistic process capable of generating an infinite population of values. An independent and identically distributed (iid) random sample is a collection of observations on a **random variable**, each of which is independently generated from the same probability distribution; iid sampling is a natural model for repeated observations of the same process. In this context as well, however, "random sample" unmodified may also refer to samples which are not iid.

ALAN ZASLAVSKY

Random Variable

Let S be a *sample space*, the set of all possible outcomes of an experiment. A random variable is a real-valued function that assigns a real value to each outcome of the sample space S. For a more rigorous definition, see the article on **Probability Theory**.

Random variables are used to describe the uncertain outcomes of a study. Thus, in a hospital the number of patients that will be receiving treatments next year may not be known. Similarly, the survival time of a patient with a terminal disease is not known.

Let X denote a random variable and let x be a particular outcome of X. A random variable X is *discrete* if it can assume a finite or a countably infinite number of possible values. X is *continuous* if it can assume any value in some interval or intervals of real numbers and if the probability is zero that it will assume any specific value. That is, if for any real number x, $f(x) = \Pr(X = x)$ is zero.

We now consider properties of a discrete random variable X. Let $f(x) = \Pr(X = x)$. A function $f(x)$ is defined to be the *probability function* or the *probability mass function* if

$$0 \le f(x) \le 1 \tag{1}$$

and

$$\sum_x f(x) = 1, \tag{2}$$

where the summation is over all x. Here, $f(x)$ is defined for all x, and is a real-valued function. *The cumulative distribution function* (cdf) $F(x)$ is defined by $F(x) = \Pr(X \le x)$, for all real x. The *survival function* $S(x)$ is given by

$$S(x) = 1 - F(x) = \Pr(X > x). \tag{3}$$

Note $F(x)$ (and $S(x)$) are also real-valued functions, and

$$\Pr(a < x \le b) = F(b) - F(a) = S(a) - S(b). \tag{4}$$

For a continuous random variable we can use similar definitions. Let X be a continuous random variable. $f(x)$ is a *probability density function* (pdf) or density function if

$$f(x) \ge 0,$$
$$\int_{-\infty}^{\infty} f(x)\ \mathrm{d}x = 1,$$

and for any two real numbers a, b with $a < b$,

$$\Pr(a < X < b) = \Pr(a \le X \le b) = \int_a^b f(x)\ \mathrm{d}x.$$

The cdf $F(x)$ is defined by $F(x) = \Pr(X \le x)$. By definition,

$$F(x) = \int_{-\infty}^{x} f(x)\ \mathrm{d}x \tag{5}$$

and

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x). \tag{6}$$

The above definitions can be extended to two or more variables when several characters, such as height and weight, are of interest. Let us consider the bivariate case.

Let (X, Y) be a discrete bivariate random variable. Then $f(x, y) = \Pr(X = x, Y = y)$ is a *joint probability function* or *joint probability mass function*, where x and y are real numbers, if

$$f(x, y) \ge 0,$$
$$\sum_{\text{all } x} \sum_{\text{all } y} f(x, y) = 1.$$

The **bivariate distribution** function $F(x, y)$ is defined by $F(x, y) = \Pr(X \le x, Y \le y) = \sum_{X \le x} \sum_{Y \le y} f(x, y)$, where the sum is over all pairs (x, y) such that $X \le x$ and $Y \le y$. The **marginal probability** functions $f(x)$ of X and $g(y)$ of Y are given by

$$f(x) = \sum_{\text{all } y} f(x, y) \tag{7}$$

and

$$g(y) = \sum_{\text{all } x} f(x, y). \tag{8}$$

The **conditional probability** functions $f(x|y)$ of X given $Y = y$, and $g(y|x)$ of Y given $X = x$, are

defined as

$$f(x|y) = \frac{f(x, y)}{g(y)}, \quad \text{where} \quad g(y) \neq 0, \tag{9}$$

and

$$g(y|x) = \frac{f(x, y)}{f(x)}, \quad \text{where} \quad f(x) \neq 0. \tag{10}$$

X and Y are *independent* if and only if

$$f(x, y) = f(x)g(y) \tag{11}$$

for all real numbers x and y.

For continuous random variables there are similar definitions. $f(x, y)$ is the joint probability density function if

$$f(x, y) \geq 0,$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\ \mathrm{d}x\ \mathrm{d}y = 1,$$

$$\Pr(a \leq X \leq b \quad \text{and} \quad c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)\ \mathrm{d}x\ \mathrm{d}y,$$

for all real numbers a, b, c and d. The marginal densities $f(x)$, $g(y)$ of X and Y are given by

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\ \mathrm{d}y \tag{12}$$

and

$$g(y) = \int_{-\infty}^{\infty} f(x, y)\ \mathrm{d}x. \tag{13}$$

The conditional density $f(x|y)$ of X given $Y = y$ is

$$f(x|y) = \frac{f(x, y)}{g(y)}, \quad \text{where} \quad g(y) \neq 0, \tag{14}$$

and the conditional density $g(y|x)$ of Y given $X = x$ is similarly given by

$$g(y|x) = \frac{f(x, y)}{f(x)}, \quad \text{where} \quad f(x) \neq 0. \tag{15}$$

X and Y are independent if and only if

$$f(x, y) = f(x)g(y) \tag{16}$$

for all real numbers x and y.

Similar definitions in terms of distribution functions can be given. For three or more variables, similar definitions can be given in an analogous manner.

ASIT P. BASU

Randomization Tests

Using randomization to assign interventions in an experiment (such as a **clinical trial**) guarantees the validity of statistical tests of significance (*see* **Hypothesis Testing**), in that the process of randomization makes it possible to ascribe a probability distribution to the difference in outcome between groups under the **null hypothesis** [2, 7]. This can be implemented directly with randomization-based inference, wherein the outcome data are analyzed many times (once for each acceptable assignment that could have been employed) and then compared with the observed result, without dependence on additional distributional or model-based assumptions. Thus, hypothesis testing ("randomization tests" or "permutation tests") and corresponding test-based **confidence intervals** can be designed based on the randomization distribution. Often, the terms "randomization test" and "permutation test" are used interchangeably, but some authors make a distinction (for example, see [4]).

The randomization test must be selected based on the study design, to produce the appropriate randomization distribution; thus, it would differ among unrestricted, stratified (*see* **Stratification**), and pair-matched designs (*see* **Matched Analysis**). In the simplest case, with pair-matched designs, we estimate some quantity for each pair; for example, the difference in outcome between the two members of pair j, denoted $\hat{X}_{1j} - \hat{X}_{2j}$, $j = 1, 2, \ldots, J$. Then the expected value of the mean difference is 0 under the null hypothesis of no intervention (treatment) effect. For hypothesis testing, we want to know the probability that an estimate of the mean would be as large or larger than the observed estimate of the mean, by chance alone. We calculate the mean (of $\hat{X}_{1j} - \hat{X}_{2j}$) for each of the 2^J ways (permutations) that the intervention assignments could have occurred. The **rank** of the observed mean among all possible means provides the one-tailed significance level; for example, if it is in the top 1%, then it is significant at the 0.01 level. Ranking the absolute values of the means leads to a two-tailed significance level (*see* **Level of a Test**).

This randomization test, whether one-tailed or two-tailed, is conditional on the absolute differences $|\hat{X}_{1j} - \hat{X}_{2j}|$. It makes no modeling assumptions but does require that the outcome measures for the two members of a pair are **exchangeable** under the null hypothesis. As noted in [11], this requirement is equivalent to orthant symmetry on the paired differences, where, as defined by Efron [5], a random vector $\mathbf{U} = (U_1, U_2, \ldots, U_J)$ exhibits orthant symmetry if it has the same distribution as $\mathbf{U}_\partial = (\partial_1 U_1, \partial_2 U_2, \ldots, \partial_J U_J)$ for every choice of $\partial_j = \pm 1$, $j = 1, 2, \ldots, J$.

We can also use this randomization test to determine a test-based confidence interval for the between-group difference. If we shift the mean outcome of one of the groups by various amounts Δ, then the confidence interval is defined as those values of Δ for which a randomization test of $\hat{X}_{1j} - (\hat{X}_{2j} + \Delta)$ fails to reject the null hypothesis at the appropriate significance level [11–13]. Operationally, if we determine Δ_1 as the value of Δ at which the null hypothesis is just rejected at $P \leq \alpha$ in the upper tail, and Δ_2 the same for the lower tail, then Δ_1 and Δ_2 are, respectively, the lower and upper limits of a $100(1 - 2\alpha)\%$ confidence interval. Another approach, using a **bootstrap** distribution, has also been suggested for determining confidence intervals in this situation, particularly for large sample sizes, as discussed by Freedman et al. [8].

The randomization test described above applies to the pair-matched setting. For an unmatched, unstratified design in which $2n$ subjects are allocated to two equally sized groups, the approach is analogous but the number of permutations is larger; namely, $\binom{2n}{n}$. Here we assume simply that all of these reallocations of subjects to groups are equally likely. This approach can be generalized to a stratified design, the number of permutations depending on the randomization distribution.

The advantage of a randomization test in providing a robust test of significance can also be obtained in the presence of **covariates**. One situation is when baseline covariates are used to adjust the analysis of intervention effect (for an example, see [3]). Such covariates can be placed in a **regression** model to predict outcome under the null hypothesis of no intervention effect (i.e. no intervention term in the model), and then **residuals** between observed and predicted outcomes can be calculated. When differences in such residuals between groups are analyzed, a beneficial effect from intervention would lead to an expected mean difference that favored the intervention group. A randomization test using residuals is a valid test of the null hypothesis even if the model is misspecified [9, 10] (*see* **Misspecification**).

A second situation using covariates involves separate analyses in subsets defined by baseline covariates. In this situation, we use statistical tests for **interaction** to investigate whether the intervention effect differs according to the covariate value. As with the test for the main effect of intervention, tests for intervention–covariate interaction can be randomization tests; here, again, we permute the assignment to intervention group based on the randomization distribution (for an example, see [3]). Such interaction tests are of interest because of the real risk of finding spurious differences in subsets by chance alone [1] (*see* **Simultaneous Inference**).

Strictly speaking, the two-sample randomization test (described above) tests the strict null hypothesis that the two groups have the same distribution (i.e. that intervention has no effect on any of the observations), rather than the less strong hypothesis that the mean intervention effect is zero. This distinction is usually not a concern in practice, but it could affect the properties of the test in some situations; for a discussion of this issue, see [9] and [6].

Detailed discussion of randomization tests for various experimental designs can be found, for example, in [4].

References

[1] Byar, D.P. (1985). Assessing apparent treatment-covariate interactions in randomized clinical trials, *Statistics in Medicine* **4**, 255–263.

[2] Byar, D.P., Simon, R.M., Friedewald, W.T., Schlesselman, J.J., DeMets, D.L., Ellenberg, J.H., Gail, M.H. & Ware, J.H. (1976). Randomized clinical trials: perspectives on some recent ideas, *New England Journal of Medicine* **295**, 74–80.

[3] COMMIT Research Group (1995). Community Intervention Trial for Smoking Cessation (COMMIT): I. Cohort results from a four-year community intervention, *American Journal of Public Health* **85**, 183–192.

[4] Edgington, E.S. (1995). *Randomization Tests*, 3rd Ed. Marcel Dekker, New York.

[5] Efron, B. (1969). Student's t-test under symmetry conditions, *Journal of the American Statistical Association* **64**, 1278–1302.

[6] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[7] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

[8] Freedman, L., Sylvester, R. & Byar, D.P. (1989). Using permutation tests and bootstrap confidence limits to analyze repeated events data from clinical trials, *Controlled Clinical Trials* **10**, 129–141.

[9] Gail, M.H., Mark, S.D., Carroll, R.J., Green, S.B. & Pee, D. (1996). On design considerations and randomization based inference for community intervention trials, *Statistics in Medicine* **15**, 1069–1092.

[10] Gail, M.H., Tan, W.Y. & Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials, *Biometrika* **75**, 57–64.

[11] Green, S.B., Corle, D.K., Gail, M.H., Mark, S.D., Pee, D., Freedman, L.S., Graubard B.I. & Lynn, W.R. (1995). Interplay between design and analysis for behavioral intervention trials with community as the unit of randomization, *American Journal of Epidemiology* **142**, 587–593.

[12] Noether, G.E. (1985). Nonparametric confidence intervals, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, New York, pp. 319–324.

[13] Tukey, J.W. (1993). Tightening the clinical trial, *Controlled Clinical Trials* **14**, 266–285.

SYLVAN B. GREEN

Randomization

Randomization refers to the random assignment of experimental units to one of two or more treatments for the purpose of comparing the treatments on some outcome measure (*see* **Randomized Treatment Assignment**). Randomization prevents the existence of systematic differences between groups other than the treatments being compared. In statistical terms, randomization provides a sound objective basis for claiming a particular distribution for the outcome under the **null hypothesis** of no difference across treatments. In the absence of systematic differences between treatments, outcome differences across treatments are strictly a function of randomization when the null hypothesis is true.

The concept of randomization was originally made explicit and advocated in 1935 by **R.A. Fisher**, in his classic text, *The Design of Experiments* [2]. The argument for randomization is that it will prevent systematic differences of any kind, whether or not they can be identified by the researcher. This makes randomization preferable to systematic assignment of treatment groups (*see* **Systematic Sampling Methods**) to produce similar distributions on a set of recognized and measurable factors; as Fisher points out, ". . .the uncontrolled causes which may influence the result are always strictly innumerable".

Randomization may not result in exactly the same distribution of identified **confounding** factors across treatment groups in a particular sample, but does guarantee that "in the long run" distributions of any factor will be the same across treatment groups. The larger the sample size, the more similar will be the distributions of confounders across treatments. Randomization can be performed within blocks, where blocks are homogeneous with respect to identified confounders, thus guaranteeing exactly the same distribution of **blocking** factors across treatments while at the same time retaining the advantage of randomization (*see* **Randomized Complete Block Designs**). Much work has been devoted to randomization designs and to methods of implementing random assignment.

Randomization was originally used in agricultural experiments, where various plots in a field are randomly assigned to different experimental conditions. In the 1940s, the idea was adopted for use in **clinical trials**, where human subjects are randomly assigned to different experimental treatments. There has been argument against randomization in particular situations with human subjects due to practical and ethical considerations [1], for example (*see* **Ethics of Randomized Trials**), but still the statistical advantage of randomization is recognized and acknowledged as a limitation when randomization is not feasible.

References

[1] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton-Mifflin, Boston.

[2] Fisher, R.A. (1935). *The Design of Experiments*. Hafner, New York.

SALLY FREELS

Randomized Complete Block Designs

In the completely randomized design or one-way design (*see* **Experimental Design**), we assume that experimental units (EUs) are initially homogeneous and that subsequent differences are due to differences in applied treatments. In many situations, a large enough group of "uniform" EUs may not be available. If such nonhomogeneous EUs are used, then the error term for testing treatment differences and estimating **standard errors** of treatment means will be inflated and hence the **power** of tests will be decreased.

One technique for dealing with this situation uses **covariates** (regressor variables measured on each EU) to account for such nonhomogeneity (*see* **Analysis of Covariance**).

Alternatively, **blocking** may be used so that EUs are grouped ("blocked") into uniform subgroups called "blocks" or "reps". (The term "rep" is used in a wide variety of contexts, so we will only use the term "block" here.) For example, human subjects in a **clinical trial** may be blocked on the basis of factors such as age or weight if the researcher thinks such factors could influence a subject's response to treatment. We usually assume that a blocking factor is an environmental factor which is not itself of much interest but which may influence an EU's response and hence mask the effect of the treatments if uncontrolled.

The randomized complete block design (RCBD) is the simplest block design and has only one blocking factor. More complicated block designs include the **Latin square** and **Graeco-Latin square** designs which have two and three blocking factors, respectively. In addition to the number of blocking factors, designs can be distinguished depending on whether the blocks are "complete" or "incomplete". For example, in the RCBD, the number of EUs per block is the same as the number of treatments and so each treatment occurs in each block exactly once. In the generalized randomized complete block design, the number of EUs per block is some constant multiple (say, k) of the number of treatments and so each treatment occurs in each block k times. If there are not enough homogeneous EUs to obtain complete blocks, then we have the class of designs called **incomplete block designs**, which may be either **balanced** or **partially balanced** with respect to the number of times two treatments occur in the same block together.

In the RCBD, once the EUs have been grouped together into blocks, treatments are randomly assigned to EUs within each block; that is, there is a separate treatment **randomization** for each block. With more complicated designs, such as Latin square and incomplete block designs, there are further restrictions on randomization to achieve balance or partial balance.

Note that, as both the number of blocking factors and restrictions on randomization for balance or partial balance increase, complications in the statistical analysis due to missing cells also increase. (In the RCBD, a "cell" is a block-by-treatment combination and a "missing cell" is a block-by-treatment combination that should have been observed according to the original design but was not, for whatever reason.) If it is anticipated that missing cells are likely, then complicated block designs should be avoided. As the simplest block design, the RCBD suffers less from these problems than more complicated designs. However, the researcher should consider use of the one-way design if many missing cells are likely (*see* **Missing Data Estimation, "Hot Deck" and "Cold Deck"**; **Multiple Imputation Methods**).

These topics are covered in greater depth in the wide variety of experimental design texts and the reader is urged to consult these (see, for example, [1], [4], [6], and [7]).

The Usual Statistical Model: Linearity and Normality

The usual model that describes the RCBD is a linear model, i.e. a model that is additive in the parameters (*see* **General Linear Model**), as opposed to a nonlinear model. In addition, the main distributional assumption is that the data come from a **normal distribution**.

In what follows we write the RCBD model for p treatments and r blocks, for the two basic cases of blocks either **fixed** or **random**. Treatments are assumed fixed. We also assume that there are no missing cells.

In both cases we express each model in both a full-rank or cell-means formulation [5] and a non full-rank, effects formulation [9]. In addition, we use

the notational convention that lower-case Latin letters are **random variables** while Greek letters are fixed (constant) parameters.

Fixed Treatments and Fixed Blocks

In the development of the theory of linear models, the idea of "fixed" effects came first. The ideas of fixed, random, and mixed models were codified by Eisenhart in 1947 [2]. In recent usages of the RCBD, blocks are usually considered as random, but there are cases where blocks as fixed effects make sense. For example, consider an example from Neter et al. [8, Problem 24.10, p. 940] in which treatment is "fat content of diet" and block is "age groups". Three people were randomly selected from each of five different age groups and assigned to one of three diets. The age groups were 15–24, 25–34, and so on up to 55–64. Clearly, these age groups were not randomly selected from some larger population (the usual operational definition of a "random" effect).

The cell-means model for the RCBD with p fixed treatments and r fixed blocks is written as

$$y_{ij} = \mu_{ij} + e_{ij}, i = 1, \ldots, p; j = 1, \ldots, r, \quad (1)$$

$$\text{subject to } \mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'} = 0, \quad (2)$$

or equivalently

$$\text{subject to } \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \overline{\mu}_{\cdot\cdot} = 0. \quad (3)$$

Here, y_{ij} is the response on the experimental unit in treatment i and block j; $\mu_{ij} = \mathrm{E}(y_{ij})$ is the **mean** response of the (i, j)th treatment-by-block cell; and e_{ij} is the **random error** or **residual**. The e_{ij}s are assumed to be independently and identically distributed (iid) as normal random variables with mean zero and constant **variance** σ_e^2. The parameters $\overline{\mu}_{i\cdot}$, $\overline{\mu}_{\cdot j}$, and $\overline{\mu}_{\cdot\cdot}$ are marginal treatment, block, and overall means, respectively, and are obtained by averaging across block, treatment, and treatment-by-block levels, respectively.

The side condition (2) or (3) is that of "no block-by-treatment **interaction**". Because there is only one experimental unit per cell and hence no estimate of experimental error variance based on replication, this assumption is necessary to provide an estimate of σ_e^2. The side condition of no interaction means that, within a block, the difference between any pair of treatments is the same as that between the same two treatments in any other block. Thus a graph of the cell means μ_{ij} vs. the block subscript j will produce parallel lines when μ_{ij} with the same treatment subscript i are connected by lines. The condition of "no interaction" is also called **"additivity"**.

The effects model for the RCBD with blocks fixed is

$$y_{ij} = \mu + \tau_i + \beta_j + e_{ij}, i = 1, \ldots, p; j = 1, \ldots, r. \quad (4)$$

Here, μ is the overall mean, τ_i is the ith treatment effect, β_j is the jth block effect, and y_{ij} and e_{ij} are as before. The "no block-by-treatment interaction" condition is represented by the absence in the model of a term $(\tau\beta)_{ij}$; that is, setting $(\tau\beta)_{ij} = 0$ is equivalent to either of the side conditions (2) and (3).

Model (4) is "nonfull rank" or "overparameterized" and therefore additional conditions must be imposed on the τ_is and β_js to achieve full rank and thus obtain a solution to the normal equations (see below). In textbooks the most common conditions imposed are the so-called "dot" conditions: $\Sigma\tau_i = \tau_\cdot = 0$ and $\Sigma\beta_j = \beta_\cdot = 0$. Alternatively, a generalized inverse can be used to solve the normal equations. These two approaches are basically equivalent, with most computer routines opting for the latter. For the general user and for cases in which the data are balanced and there are no missing cells, these issues are somewhat unimportant. The interested reader is referred to [10] and [11] for thorough discussions of problems with nonfull rank, effects models, and unbalanced data.

Fixed Treatments and Random Blocks

The cell-means model for the RCBD with blocks random is

$$y_{ij} = \mu_i + b_j + e_{ij}, i = 1, \ldots, p; j = 1, \ldots, r. \quad (5)$$

Here, μ_i is the ith treatment mean, b_j is a random effect due to the jth block, and y_{ij} and e_{ij} are as before. The random block effects are assumed to be iid normal with mean zero and variance σ_b^2. In addition, the e_{ij}s and b_js are assumed to be mutually independent.

The effects model for the RCBD with blocks random is

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}, i = 1, \ldots, p; j = 1, \ldots, r. \quad (6)$$

Table 1 Summary of the two RCBD models

	Fixed model Treatments (i) and blocks (j) both fixed	Mixed model Treatments (i) fixed, blocks (j) random
Linear models		
Cell-means (full rank)	$y_{ij} = \mu_{ij} + e_{ij}, \quad i = 1, 2, \ldots, p;$ $j = 1, 2, \ldots, r$ subject to $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$ with e_{ij} iid $\mathrm{N}(0, \sigma_e^2)$	$y_{ij} = \mu_i + b_j + e_{ij}, \quad i = 1, 2, \ldots, p;$ $j = 1, 2, \ldots, r$ subject to $\sigma_{tb}^2 = 0$ with b_j iid $\mathrm{N}(0, \sigma_b^2)$, e_{ij} iid $\mathrm{N}(0, \sigma_e^2)$, and b_j and e_{ij} mutually independent
Effects (nonfull rank)	$y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$ subject to $(\tau\beta)_{ij} = 0$ with e_{ij} iid $\mathrm{N}(0, \sigma_e^2)$ and $\sum_i \tau_i = \sum_j \beta_j = 0$ (usually)	$y_{ij} = \mu + \tau_i + b_j + e_{ij}$ subject to $\sigma_{tb}^2 = 0$ with b_j iid $\mathrm{N}(0, \sigma_b^2)$ e_{ij} iid $\mathrm{N}(0, \sigma_e^2)$ and b_j and e_{ij} mutually independent
Mean, variance, and covariances of y_{ij}	(1) $\mathrm{E}(y_{ij}) = \mu_{ij}$ (2) $\mathrm{cov}(y_{ij}, y_{i'j'}) = \sigma_e^2, i = i', j = j'$ $0, \quad i = i', j \neq j'$ $0, \quad i \neq i', j = j'$ $0, \quad i \neq i', j \neq j'$	(1) $\mathrm{E}(y_{ij}) = \mu_i$ (2) $\mathrm{cov}(y_{ij}, y_{i'j'}) = \sigma_e^2 + \sigma_b^2, i = i', j = j'$ $0, \quad i = i', j \neq j'$ $\sigma_b^2, \quad i \neq i', j = j'$ $0, \quad i \neq i', j \neq j'$
	Source E(*MS*)	Source E(*MS*)
Expected mean squares for ANOVA	Treatments $\sigma_e^2 = \dfrac{r\sum_{i=1}^{p}(\overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot\cdot})^2}{p-1}$ Blocks $\sigma_e^2 = \dfrac{p\sum_{j=1}^{r}(\overline{\mu}_{\cdot j} - \overline{\mu}_{\cdot\cdot})^2}{r-1}$ Error σ_e^2	Treatments $\sigma_e^2 = \dfrac{r\sum_{i=1}^{p}(\mu_i - \overline{\mu}_{\cdot})^2}{p-1}$ Blocks $\sigma_e^2 + p\sigma_b^2$ Error σ_e^2

continued overleaf

Table 1 (*continued*)

	Fixed model Treatments (i) and blocks (j) both fixed			Mixed model Treatments (i) fixed, blocks (j) random		
	Parameter	Estimator	SE	Parameter	Estimator	SE
Fixed parameters, estimators and estimated standard errors	Cell mean μ_{ij}	$\overline{y}_{i.} + \overline{y}_{ij} - \overline{y}_{..}$	$[MS_e(p+r-1)/pr]^{1/2}$	–	–	–
	Treatment mean $\overline{\mu}_{i.}$	$\overline{y}_{i.}$	$(MS_e/r)^{1/2}$	Treatment mean μ_i	$\overline{y}_{i.}$	$\left\{\frac{1}{r}\left[MS_e + \frac{MS_b - MS_e}{p}\right]\right\}^{1/2}$
	Block mean $\overline{\mu}_{.j}$	$\overline{y}_{.j}$	$(MS_e/p)^{1/2}$	–	–	–
	Overall mean $\overline{\mu}_{..}$	$\overline{y}_{..}$	$(MS_e/pr)^{1/2}$	Overall mean $\overline{\mu}_{.}$	$\overline{y}_{..}$	$(MS_b/pr)^{1/2}$
	Treatment difference $\overline{\mu}_{i.} - \overline{\mu}_{i'.}$	$\overline{y}_{i.} - \overline{y}_{i'.}$	$(2MS_e/r)^{1/2}$	Treatment difference $\overline{\mu}_{i'} - \overline{\mu}_{i'}$	$\overline{y}_{i.} - \overline{y}_{i'.}$	$(2MS_e/r)^{1/2}$
	Variance component	REML or ANOVA estimator	SE	Variance component	Estimator	SE
Variance Component Estimators and estimated standard errors	σ_e^2	MS_e	$\left\{\frac{2(MS_e)^2}{(p-1)(r-1)}\right\}^{1/2}$	σ_e^2	MS_e	$[2(MS_e)^2/(p-1)(r-1)]^{1/2}$
				σ_b^2	$\frac{MS_b - MS_e}{p}$	$\left\{\frac{2}{p^2}\left[\frac{(MS_b)^2}{(r-1)} + \frac{(MS_e)^2}{(p-1)(r-1)}\right]\right\}^{1/2}$

The only change between (5) and (6) is that μ_i has been replaced with $\mu + \tau_i$, where again μ is the fixed overall mean and τ_i is the fixed treatment effect. Note that apart from the substitution of b_j for β_j, models (4) and (6) look identical. This resemblance is somewhat misleading, as the effect of blocking is described by quite different parameters in the two models: by means for fixed blocks and by the **variance component** σ_b^2 for random blocks. This difference is emphasized by comparing the mean, variance, and covariances of the y_{ij} for the two models (see Table 1). In addition, when blocks are random, the "no block-by-treatment interaction" assumption means that $\sigma_{tb}^2 = 0$.

The Usual Statistical Analysis

Estimation

Fixed parameters, such as treatment means or effects, are usually estimated by **least squares**, which involves minimizing the error sum of squares $SS_e = \sum_{i=1}^{p} \sum_{j=1}^{r} e_{ij}^2$. In the case of fixed blocks, this results in a set of linear equations called the normal equations. In the case of random blocks, we get a more complicated set of linear equations called the mixed model equations [3]. In the balanced case with no missing cells, formulas for the estimators are nicely intuitive, no matter whether blocks are considered fixed or random, and the estimators are, in fact, uniformly **minimum variance unbiased** (UMVU) under normality. Estimates for treatment means and pairwise treatment differences (*see* **Paired Comparisons**), along with their estimated **standard errors**, are summarized in Table 1.

There are many different techniques for estimating variance components. Three common ones are **maximum likelihood** (ML), **restricted maximum likelihood** (REML) and the **method of moments**, which is also referred to as **analysis of variance** (ANOVA) estimation. In the balanced case with no missing cells, the REML and ANOVA estimators are the same and are also UMVU under normality. These estimators are summarized in Table 1. In the balanced case, the ML estimators may be equal to the REML estimator or may differ by a simple constant, which is a function of the number of levels in the factor.

Tests of Hypothesis

Given that there are no missing cells, the data from an RCBD are usually analyzed by ANOVA, an arithmetic technique for partitioning an overall measure of variability ("sum of squares") into pieces that are associated with a particular set of effects, e.g. differences in treatment means or a variance component.

For the RCBD, the ANOVA partitions the total sum of squares into three pieces associated with treatments, blocks, and error. (Note that if the "no block-by-treatment interaction" assumption is not satisfied, the error source of variation is measuring the strength of that interaction.) These three sums of squares are then divided by their respective **degrees of freedom** to obtain mean squares, and ratios of the mean squares are then taken to obtain the F test statistics (*see* ***F* Distributions**). The ANOVA is the same whether blocks are fixed or random and is summarized in Table 2. The expected mean squares are used in mixed models to determine appropriate testing terms for the F tests. These do differ, depending on whether blocks are fixed or random, and are summarized in Table 1.

Table 2 Analysis of variance: fixed model and mixed model

Source	Degrees of freedom	Sums of squares	Mean squares	F ratio
Treatments	$p-1$	$SS_t = r\sum_{i=1}^{p}(\overline{y}_{i.} - \overline{y}_{..})^2$	$MS_t = SS_t/(p-1)$	$F_t = MS_t/MS_e$
Blocks	$r-1$	$SS_b = p\sum_{j=1}^{r}(\overline{y}_{.j} - \overline{y}_{..})^2$	$MS_b = SS_b/(r-1)$	$F_b = MS_b/MS_e$
Error	$(p-1)(r-1)$	$SS_e = \sum_{i=1}^{p}\sum_{j=1}^{r}(\overline{y}_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2$	$MS_e = SS_e/(p-1)(r-1)$	

References

[1] Cochran, W. & Cox, G. (1957). *Experimental Designs*. Wiley, New York.

[2] Eisenhart, C. (1947). The assumptions underlying the analysis of variance, *Biometrika* **3**, 1–21.

[3] Henderson, C. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph.

[4] Hinkelmann, K. & Kempthorne, O. (1994). *Design and Analysis of Experiments, Volume I: Introduction to Experimental Design*. Wiley, New York.

[5] Hocking, R. (1996). *Methods and Applications of Linear Models: Regression and Analysis of Variance*. Wiley, New York.

[6] Kuehl, R. (1994). *Statistical Principles of Research Design and Analysis*. Duxbury, Belmont.

[7] Mason, R., Gunst, R. & Hess, J. (1989). *Statistical Design and Analysis of Experiments with Applications to Engineering and Science*. Wiley, New York.

[8] Neter, J., Wasserman, W. & Kutner, M. (1990). *Applied Linear Statistical Models*, 3rd Ed. Irwin, Homewood.

[9] Searle, S. (1971). *Linear Models*. Wiley, New York.

[10] Speed, F. & Hocking, R. (1976). The use of $R()$-notation with unbalanced data, *American Statistician* **30**, 30–33.

[11] Speed, F., Hocking, R. & Hackney, O. (1978). Methods of analysis of linear models with unbalanced data, *Journal of the American Statistical Association* **3**, 105–112.

L. MURRAY

Randomized Response Techniques

Randomized response (RR) sampling owes its beginning to Stanley L. Warner [29] in 1965. He developed his method to address the need which often arises in medical, psychological, or sociological investigations for facts about highly sensitive matters. For example, most survey respondents would feel uncomfortable discussing their history of abortion, participation in drug use, sexual behavior, status relative to many medical conditions, extent of use of oral contraceptives, or whether they have defrauded some governmental agency. The standard approach for such questions has been to assure the respondent of the **confidentiality** or anonymity of the information he or she provides. Nevertheless, such attempts at open inquiry into sensitive issues often result in high **nonresponse** rates, willful misstatements, and outright lies, inducing a high degree of **bias** and error that cannot be overcome.

As conceived by Warner, the method of RR sampling would (by its nature) both guarantee the respondent's anonymity *and* convince the respondent of that protection. His original procedure was to offer to the respondent a choice of questions where one was the opposite or negation of the other. For example, the respondent could be offered two questions:

1. Have you smoked marijuana in the last 30 days?
2. Have you not smoked marijuana in the last 30 days?

To determine which question is to be answered, the respondent is given a "randomizing device" (*see* **Randomization**) of some kind, e.g. a deck of cards, some of which are marked with question 1 and the rest with question 2. Thus, the drawing of a card would "randomly" determine which question the respondent would answer. The outcome of the randomizing device is seen *only* by the respondent, *not* the interviewer. Thus the interviewer may record a "yes" response but never know if that answer meant "Yes, I have smoked marijuana in the last 30 days and I am answering question 1" or "Yes, I have not smoked marijuana in the last 30 days and I am answering question 2".

The nature of the sampling is explained by the interviewer carefully to the respondent and then the respondent is allowed to investigate the randomizing device to his or her satisfaction before participating in the survey. Thus (it was hoped that) the respondent would realize his or her "perfect anonymity". That is, not even the interviewer could know for sure to which group the respondent belonged. Nevertheless, by knowing the make-up of the randomizing device as well as the aggregate results, an estimate can be obtained of the proportion of respondents, in this example, who had smoked marijuana in the last 30 days without ever knowing the precise status of any single respondent. Such knowledge would presumably not only make respondents willing to participate in the survey but also persuade them to provide truthful responses.

There are a few disadvantages to the use of RR techniques. Implicit in the nature of RR is the assumption that the respondent is sufficiently cognizant, informed, and educated to recognize and appreciate his or her anonymity. This feature could escape an audience of poor education or low sophistication. Secondly, the estimates provided by RR techniques have larger variation than a standard estimator from the same-sized sample. This naturally results because of the "noise" or "extra variation" introduced by the randomizing device. However, one can easily compensate for this shortcoming by simply increasing the size of the sample to be taken by the RR techniques. Most statisticians feel this tradeoff is a small price to pay for both the increased cooperation and the increased truthfulness by respondents that result in more accurate estimates.

In summary, the underlying rationale for all RR techniques is that a randomizing device is used to select the question to be answered by the respondent. Such a device, it is hoped, will convince the respondent that even the interviewer will be unable to determine the respondent's true status with respect to the sensitive issue(s) being addressed.

Several papers on RR provide significant reviews: see [3], [14], [16], and [17]. However, by far the most recent and most comprehensive review of RR techniques which presents 424 references and an abstract of each publication is that of Daniel [5].

Truly from humble beginnings, RR techniques have flourished both in theoretical sophistication and in practical application to actual surveys. This article attempts to cite several significant papers and direct the reader to several important dimensions developed in RR techniques.

Warner's Original Model

Continuing with our marijuana example, let π be the proportion of individuals possessing the sensitive characteristic (those who had smoked marijuana in the last 30 days) and let p be the probability of the random device giving question 1 (where $p \neq \frac{1}{2}$). Then the probability of getting a "yes" is given by

$$\lambda = \Pr(\text{yes}) = \pi p + (1-\pi)(1-p). \qquad (1)$$

The πp term is due to the "yes" answers received from those "who had smoked marijuana in the last 30 days" *and* who got question 1. The $(1-\pi)(1-p)$ term is due to the "yes" answers received from those "who have *not* smoked marijuana in the last 30 days" *and* who got question 2. If we let n_1 be the total number of "yes" responses out of a sample taken of size n, then an **unbiased** estimate of λ is given by

$$\hat{\lambda} = \frac{n_1}{n}. \qquad (2)$$

Utilizing (1) and (2), an unbiased estimate of π is given by

$$\hat{\pi} = \frac{p-1}{2p-1} + \frac{n_1}{(2p-1)n}, \qquad (3)$$

where $p \neq \frac{1}{2}$. Since n_1 follows a **binomial distribution** with parameters n and λ, it can be shown that

$$\text{var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}, \qquad (4)$$

and an estimate of this **variance** can be found by using $\hat{\pi}$ for π in (4). The first term in (4) is the variance term due to binomial sampling, and the second is the variance term due to the randomizing device. Note that if $p = 0$ or if $p = 1$, RR sampling reduces to ordinary (binomial) sampling, and (3) and (4) simplify accordingly.

For sufficiently large samples this estimated variance can be used to construct a $(1-\alpha)$ 100% **confidence interval** for π:

$$\hat{\pi} \pm Z(\alpha/2)\sqrt{\widehat{\text{var}}(\hat{\pi})}.$$

In addition, Levy [18] describes how tests of hypothesis (*see* **Hypothesis Testing**) about π may naturally be done by inserting the hypothesized value of π in (4) to obtain an estimate of $\text{var}(\pi)$.

Continuing with our example, suppose $n = 400$ high school seniors were selected from a New York inner-city school and asked their marijuana use within the last 30 days by the RR method described. Let the randomizing device be a deck of 50 cards, where 15 have question 1 and 35 have question 2, so that $p = 15/50 = 0.30$. If $n_1 = 240$ answered "yes", then $\hat{\lambda} = n_1/n = 0.60$. Hence, by (3), Warner's original estimate is given by $\hat{\pi} = 0.25$ and (4) gives $\widehat{\text{var}}(\hat{\pi}) = 0.00375$.

If an earlier study seemed to indicate only 10% of these high school seniors had smoked marijuana in the last 30 days, a natural hypothesis test would be $H_0 : \pi = 0.10$ vs. $H_1 : \pi > 0.10$ at $\alpha = 0.05$. That is, does the RR method seem to indicate a proportion of marijuana use higher than 10%? Substituting the hypothesized proportion $\pi = 0.10$ into (4) gives $\text{var}(\hat{\pi}) = 0.00351$ with a resulting observed test statistic of

$$Z = \frac{\hat{\pi} - \pi_0}{[\text{var}(\hat{\pi})]^{1/2}} = \frac{0.25 - 0.10}{(0.00351)^{1/2}} = 2.53.$$

The critical value would be $+1.645$, and thus we would reject H_0 with an observed p value of $\Pr(Z > 2.53) = 0.0057$. A 95% confidence interval for π would be given by $\hat{\pi} \pm 1.96[\widehat{\text{var}}(\hat{\pi})]^{1/2} = 0.25 \pm 1.96(0.00375)^{1/2} = (0.13, 0.37)$.

Improving the Efficiency of RR Sampling Using Multiple Trials

It has already been noted that the introduction of the randomizing device induces extra variation into the estimators. This effectively decreases the sample size. So while in our example $n = 400$, the effect of randomizing might be an increase in variation so that the sample is "effectively" only 250. One obvious way to overcome this is simply to increase the sample size. Another less obvious solution is to have multiple trials performed on each respondent [19, 20]. While it does not make any sense to ask a respondent the same question repeatedly in ordinary sampling, the reader can see the potential benefit in RR sampling by considering an extreme case of our high school seniors example where $p = 0.30$. If a single respondent were asked 100 times, one could conceivably get 100 "yes" and 0 "no" answers or 0 "yes" and 100 "no" answers or anything in between. A large number of "yes" and few "no" answers (say,

70 "yes" and 30 "no") would tend to indicate that this respondent had *not* smoked marijuana in the last 30 days, since in 100 trials where $p = 0.30$, one would expect to get 30 answers to question 1 and 70 answers to question 2. Conversely, a large number of "no" and few "yes" answers (say, 70 "no" and 30 "yes") would tend to indicate that this respondent *had* smoked marijuana in the last 30 days. However, it would be expected that as a single respondent was asked the same question over and over the respondent might realize that he or she is progressively losing his or her anonymity. This could result in loss of truthfulness or random answering and a collapse of any benefit normally associated with RR sampling. Liu & Chow [20] explore the optimal number of trials per respondent and derive the model and estimators associated. Furthermore, they report on a field trial conducted in Taiwan where the randomizing device was a globe with a short neck that contained 3 red and 7 white balls. The globe was manipulated by a respondent so that a single ball fell at random into the neck. Depending on the color of the ball (which was hidden from the interviewer) the respondent answered a question related to abortion. Their study compared direct questioning to RR with one trial and to RR with three trials. Their results indicate a much higher rate obtained by both RR methods (with the three-trial estimate highest of all) and no apparent discomfort from the respondents who were asked to undergo the process three times.

Chow et al. [4] formulated a new RR model and randomizing device to obtain multiple responses without appearing to do so. Their device consisted of a glass urn with a long neck. The urn contained balls of two colors, one color associated with the sensitive category and the other to the nonsensitive. The urn is shaken by the respondent and inverted to allow balls to flow into the neck. The respondent then tells the number of balls in the neck that correspond to his or her category color. They develop estimates based on the **hypergeometric distribution**. Greenberg et al. [14] developed a **method of moments** estimators for this model and device.

Technical Concerns with Warner's Original Model

Several authors have noted some technical problems with Warner's original model. Moors [24, 25] notes that $1 - p \leq \lambda \leq p$ if $p > 1/2$ and $p \leq \lambda \leq 1 - p$ if $p < 1/2$ since $0 \leq \pi \leq 1$. Thus λ must be a truncated estimate and this prevents $\hat{\pi}$ from being a truly unbiased estimate of π. Also, although Warner originally claimed $\hat{\pi}$ was the **maximum likelihood** estimate (MLE) for π, a number of authors pointed out a flaw in his original claim [6, 8, 27, 28]. All recommend a truncated estimator that is the true MLE for π given by the following. For $p < 1/2$, the estimate is given by 1 if $\hat{\lambda} \leq p; \hat{\pi}$ if $p < \hat{\lambda} < 1 - p$; and 0 if $\hat{\lambda} \geq 1 - p$. But for $p > 1/2$, then the estimate is given by 0 if $\hat{\lambda} \leq 1 - p; \hat{\pi}$ if $1 - p < \hat{\lambda} < p$; and 1 if $\hat{\lambda} \geq p$. Without such truncations, sample estimates could give strange estimates of π that could be *negative* or *greater* than 1! For our illustrative high school example, suppose $n_1 = 300$ answered yes; then $\hat{\lambda} = n_1/n = 0.75$ and so $\hat{\pi} = -0.125$! Using the recommended truncated estimator, this value would be set to 0. Devore [6] offers a slightly modified procedure where n is an even number and stands for the number of cards in a deck, one-half of which are marked with the command to answer "yes". The other half of the cards are marked with the sensitive question of interest. The sample of n chosen persons is asked to select a card without replacement. This results in a true unbiased estimate of π that is also the MLE and is given by $\hat{\pi} = (2n_1/n) - 1$ with $\text{var}(\hat{\pi}) = (2/n)\pi(1 - \pi)$ and $\widehat{\text{var}}(\hat{\pi}) = [2/(n-2)]\hat{\pi}(1 - \hat{\pi})$.

Other Generalizations to Warner's Original Model

Abul-Ela et al. [1] extended Warner's model to consider a population of t mutually exclusive categories where at least one but not more than $t - 1$ of those categories is a sensitive category. Their study considered three categories of women who were known recently to have given birth to a child: those women who were already married when they became pregnant; those who got married during their pregnancy; and those who were still unmarried at the time of their delivery. Here, the last two categories are potentially sensitive ones. The randomizing device they used was a deck of cards with each card having one of three associated questions to which the respondent would answer "yes" or "no". The authors developed the model and estimators for this trichotomous situation and reported on some observations of a field trial

of the model. Note that (in general) a situation with t categories necessitates $t - 1$ independent samples, each of which has a different proportion of allocation of questions by the randomizing device.

Liu et al. [22] developed a new randomizing device to be used in the multiproportions environment. The device consists of a glass urn with a long neck. The urn contains beads of several colors, one color associated with each category of interest. The urn is inverted, allowing several balls of each color to flow into the neck which has locations sequentially numbered on it. The respondent is asked to reply according to the location of the first bead of his or her identifying color that appears in the neck.

Franklin [9] considered a dichotomous population (as Warner did) but used a randomizing device that used continuous distributions. Specifically, he presented a portable, programmable computerized device that presented respondents with two windows: one marked "yes" and the other marked "no". Whenever the respondent pressed a button on the device, two separate two-digit numbers appeared in these windows. The interviewer, after explaining what was to take place, then allowed the respondent as much time as he or she wished to "play with" the device and the button. Starting only after the respondent was content with the device, the interviewer then asked a sensitive question. The respondent then pushed the button on the device another time and reported the two-digit number that appeared in the appropriate "yes" or "no" window. This two digit response was recorded by the interviewer. The numbers from the "yes" window were distributed approximately normally with **mean** 40 and **standard deviation** 5, while the numbers from the "no" window were distributed approximately **normally** with mean 50 and standard deviation 5. Thus, a two-digit response such as "43" could have come from either distribution. Franklin incorporated multiple trials into his model by actually displaying six digits in each of the "yes" and "no" windows. This actually constituted three separate two-digit responses, unbeknown to the respondent who felt he or she was responding once. Franklin field-tested his device on university students, asking five sensitive questions. His article both derives the theory for his "continuous distribution randomizing device" and presents the results of his field trial which contrasts estimates of the five proportions of interest made by his RR technique with those made by direct sampling methods. His results indicate that all five RR estimates of the proportions were greater than the corresponding one from direct sampling. In addition, three of the RR proportions were significantly greater statistically than the corresponding one from direct sampling.

The Unrelated Question (RR) Model

The randomizing device that selects one of two questions to be answered is the crux of Warner's model; one question is the opposite of the other and hence they are "related". Greenberg, Horvitz and colleagues [11, 15] were the first to note that the two questions need not be related. They explored the use of two randomizing devices: one a deck of 50 cards and the other, a box with two colors of beads. Each card in the deck had one of two statements:

1. There was a baby born in this household after January 1, 1965, to an unmarried woman who was living here (sensitive).
2. I was born in North Carolina (nonsensitive).

The box of beads had a small window in which (after shaking) one bead would randomly appear. The color of the bead was associated with one of the two statements which were attached to the box. The window was not visible to the interviewer. Results of field studies undertaken in North Carolina are reported, and a "two trials per respondent" model developed for the unrelated question scenario.

The introduction of a second, unrelated and nonsensitive question appears to have both advantages and disadvantages. Presumably, the answering of two distinct questions by "yes" or "no" would not be as potentially confusing as Warner's original procedure. There, if a marijuana smoker obtained question 2, he or she must (if truthful) answer "no". This double negative answer could pose some difficulty in certain sampling situations.

However, the use of a second question introduces two proportions: the proportion of respondents who possess the sensitive characteristic (π_1) and the proportion of respondents who possess the nonsensitive characteristic (π_2). To solve for π_1, either outside or previous knowledge of the value of π_2 must be available so that the scenario reduces to solving one equation in one unknown, π_1; or, if π_2 is itself unknown, a second independent sample (with a different value of p for the randomizing device) must be obtained. This

produces a scenario of two equations obtained (one from each of the two independent samples) in two unknowns, π_1 and π_2. Whichever scenario occurs, estimators and estimates of their variances can be easily obtained.

In spite of the extra parameters introduced, Greenberg et al. [11] were able to show that under certain conditions and parameter choices, the unrelated model can be made more efficient than the original Warner model. In addition, Greenberg et al. [12] showed how the "unrelated but innocuous question" could be built into the randomizing device itself. Specifically, their randomizing device was a box with three types of colored ball: red, white, and blue.

The sensitive question they examined dealt with the occurrence of an emotional problem requiring professional help and was answered by the respondent if the ball which randomly appeared in the window of the box was red. Otherwise, the respondent answered the unrelated question of "The color of the ball in the window is blue". See [12] for derivation of estimates and their variances.

Moors [23] addressed the optimization of choices for samples and of the proportion for the randomizing devices for the unrelated question RR model.

The Use of RR Techniques in Obtaining Quantitative Data

All the RR techniques thus far considered are essentially categorical or qualitative in nature. That is, respondents fall into one of several mutually exclusive categories (at least one of which is sensitive) and the intent of the RR method is to estimate accurately the proportions associated with each category. But many scenarios may beg for a quantitative measure of the sensitive attributes. For example, instead of wishing to know "Have you smoked marijuana in the last 30 days?", one may really be interested in "How many times have you smoked marijuana in the last 30 days?".

Greenberg et al. [13] developed just such an RR technique for dealing with a quantitative response. They modified their earlier developed "unrelated question" model by having two unrelated questions (one of which was sensitive) but which both required a quantitative (numerical) response rather than a categorical (yes/no) response. The two specific questions they used were:

1. "How many abortions have you had in your lifetime?" (sensitive).
2. "If a woman has to work full-time to make a living, how many children do you think she should have?" (nonsensitive).

The randomizing device was a box with red and blue balls: The two questions were identified by "red" and "blue", respectively. The respondent shook and then tipped the box, allowing a single ball to appear randomly in a window (which was hidden from the interviewer). The respondent then answered the question which matched the color of the ball that appeared in the window.

This model allowed estimation of the means of the two distributions addressed by the questions. Since both of those means were unknown, the investigation necessitated two independent studies, each using a bead box with a different proportion of red and blue balls as the randomizing device. See [13] for derivation of estimates for the means and estimates of their respective variances.

Eriksson [7] introduced a model in which a finite set of values can be given as the response to the sensitive question. As an example, the sensitive question could be "In how many of the last 7 days have you smoked marijuana?" and thus the responses could be 0 through 7. The randomizing device he proposed was a deck of cards some of which were marked with this type of sensitive question and others were marked with statements like "Give as your answer _____!" where the blank would contain a number from 0 through 7. Knowing the proportion of the cards marked with "Give as your answer 0", "Give as your answer 1", etc., he derives estimates of the mean of the sensitive distribution.

Liu et al. [21] developed, apparently independently, a similar model and randomizing device. Their randomizing device consisted of a glass urn with a short neck that could contain precisely one ball. The urn contained balls of two colors: red and white. If after shaking, a red ball appeared in the neck of the urn, the respondent answered the sensitive, quantitative question. But the white balls all were numbered with a discrete number (such as from 0 through 7) and if a white ball was obtained, the respondent would give as his or her reply the number on the ball. Knowing the precise number of balls marked with 0s, with 1s, etc., they derived estimates

of the mean of the distribution for the quantitative, sensitive characteristic.

Poole [26] presented an RR technique capable of estimating an entire distribution function for a quantitative sensitive response. His randomizing method consisted in asking the respondent to select randomly a number from a table of random numbers, enter that number into a calculator and then multiply it by his or her salary (the sensitive issue). The resulting number is the value given by the respondent. Knowing the distribution of the random numbers, he derives an estimate of the distribution function associated with personal salaries.

Important Field Studies Using RR Techniques

Many field studies have been undertaken with a wide range of RR models and randomizing devices. While the various models and randomizing devices and field studies are not without controversy, nevertheless, it can be generally stated that RR techniques have performed at least as well as ordinary sampling methods. Four studies in particular have not only attempted to conduct a study by RR techniques but also simultaneously conducted a comparison study using ordinary sampling methods. Their results and observations are instructive.

Boruch [2] reported a study of marijuana use among college students and found no significant difference in the estimated means obtained by direct question methods, from an RR technique, and from his own contamination model. In fact, the results seem only to indicate greater variability present in the two indirect procedures than in the direct question method.

Goodstadt et al. [10] reported on a field study of drug usage involving 854 high school students and their drug usage of six separate substances over the preceding 3 months. They found that subjects were significantly more likely to respond to these questions when the RR method was used than with the direct method. Furthermore, for five of the six drugs (alcohol, marijuana, amphetamines ("speed"), tranquilizers, and heroin) the subjects claimed significantly more frequent drug use than with the traditional direct method. Only for hallucinogens was no significant difference found. One of their closing statements is insightful: "The most significant result of the reported research was the finding that the standard (direct) procedures for inquiring appear to provide a significant *underestimate* of drug use. Statistically significant differences in estimated mean drug use occurred *despite* the considerably higher variability associated with the randomized response estimates."

Liu et al. [20] presented the comparative results of rates of abortion in Taiwan by three methods: direct question; RR with one trial; and RR with three trials. The last two methods gave significantly higher estimates of the proportion of women who have had an abortion of 28.2% and 30.3% (respectively) as compared to several direct studies that never achieved over 19.5%. In addition, a matching study of 48 (successful) interviews with Taiwanese women known to have had an abortion was conducted by RR (multiple trial) and direct methods. The RR method estimated the proportion who had an abortion as 95.7% with a **standard error** of 8% so that a 95% confidence interval would contain the expected value of 100%. However, the direct method estimated a significantly lower percentage of 19.8% with a standard error of 10.5%. Thus, a 95% confidence interval estimate from the direct method would be no larger than 40%, which is considerably lower than the expected value of 100%. In their words, this "suggests that most of the abortion cases were willing to give truthful responses in this (RR) multiple trial model".

Finally, Franklin [9] reported on two studies of college students involving 473 who were interviewed by RR (using continuous randomizing distributions) and 477 who were interviewed by the direct method. The two studies were independent but conducted on one college campus within a period of a week. He deliberately chose five questions with what were judged to be varying degrees of "sensitivity". All five questions gave a higher estimated proportion by the RR method than by the direct method. However, two of the questions (of what was judged "moderate sensitivity") surprisingly showed no significant differences ("Have you smoked marijuana in the last 30 days?" and "Would you ever steal from an employer?"). There was a significantly greater estimate from the RR method than from the direct method for the other three questions, ("Have you ever participated in a homosexual act?", "Have you ever cheated on an exam here at this university?", and "Would you ever cheat on your income tax?") with all ***P* values**

less than 0.0004. It was not anticipated that the two questions on cheating would be significant as they were felt to be of "low sensitivity". However, the researcher noted that before the survey was actually conducted, the students were asked for their social security numbers, which were then used to screen "enrolled students" from "visitors on campus". The researcher reasoned that having those unique identifying numbers turned those questions about cheating into highly sensitive ones (for the "direct questioned respondents"). The article conjectures "that some of the confusion about the efficacy of the RR technique may be related to the 'true sensitivity' of the question for the interviewee as opposed to the 'perceived sensitivity' by the interviewer". The article also reported that 88.9% of the respondents who used the RR technique felt "their friends would be more likely to answer truthfully to sensitive questions by this RR technique". It should be noted also that the technique used by Franklin incorporated (unbeknown to the respondent) three trials per respondent along with sample sizes of near 500.

In summary, RR techniques seemed to establish themselves as being superior to direct question methods *if* certain reservations are understood. First, the question of interest must be "truly sensitive" to the population of interest. For example, it may well be the case that "Have you smoked marijuana in the last 30 days?" is a sensitive question for high schools students but not necessarily for college students. Secondly, RR techniques have greater variability that is inherently introduced by the randomizing device as compared to direct question methods. This additional variability needs to be factored into the circumstances and overcome *either* by multiple trial responses *or* by a careful choice of randomizing parameters *or* by a substantially increased sample size *or* by a combination of all three.

Odds and Ends

Winkler & Franklin [31] approached the original Warner dichotomous model from a **Bayesian** perspective. Taking a natural conjugate **beta** form for the **prior** density of π (the proportion of the population with the sensitive characteristic), they derived posterior density forms.

Warner [30] derived a **general linear model** for randomized response.

Chaudhuri & Mukerjee [3] present RR sampling from a finite population and develop some "unifying theory" that is very helpful and insightful.

References

[1] Abul-Ela, A.A., Greenberg, B.G. & Horvitz, D.G. (1967). A multi-proportions randomized response model, *Journal of the American Statistical Association* **62**, 990–1008.

[2] Boruch, R.F. (1972). Relations among statistical methods for assuring confidentiality of social research data, *Social Sciences Research* **1**, 403–414.

[3] Chaudhuri, A. & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. Marcel Dekker, New York.

[4] Chow, L.P., Liu, P.T. & Moseley, W.H. (1973). New randomized response technique for study of contemporary social problems. Paper presented to the Annual Meeting of the American Public Health Association, Statistics Section, San Francisco, November.

[5] Daniel, W.W. (1993). *Collecting Sensitive Data by Randomized Response: An Annotated Bibliography*, 2nd Ed., Research Monograph No. 107. Georgia State University Business Press, Atlanta.

[6] Devore, J.L. (1977). A note on the randomized response technique, *Communications in Statistics–Theory and Methods* **6**, 1525–1529.

[7] Eriksson, S.A. (1973). A new model for randomized response, *International Statistical Review* **41**, 101–113.

[8] Flingner, M.A., Policello II, G.E. & Singh, J. (1977). A comparison of two randomized response survey methods with consideration for the level of respondent protection, *Communications in Statistics–Theory and Methods* **6**, 1511–1524.

[9] Franklin, L.A. (1989). Randomized response sampling from dichotomous populations with continuous randomization, *Survey Methodology* **15**, 225–235.

[10] Goodstadt, M.S. & Gruson, V. (1975). The randomized response technique: a test on drug use, *Journal of the American Statistical Association* **70**, 814–818.

[11] Greenberg, B.G., Abdel-Latif, A.A., Simmons, W.R. & Horvitz, D.G. (1969). The unrelated question randomized response model: theoretical framework, *Journal of the American Statistical Association* **64**, 520–539.

[12] Greenberg, B.G., Abernathy, J.R. & Horvitz, D.G. (1970). A new survey technique and its application in the field of public health, *Milbank Memorial Fund Quarterly* **48**, 39–55.

[13] Greenberg, B.G., Kuebler, R.R., Jr, Abernathy, J.R. & Horvitz, D.G. (1971). Application of the randomized response techniques in obtaining quantitative data, *Journal of the American Statistical Association* **66**, 243–250.

[14] Greenberg, B.G., Horvitz, D.G. & Abernathy, J.R. (1974). Comparison of randomized response designs, in *Reliability and Biometry, Statistical Analysis of Life*

Length, F. Prochan & R.J. Serfling, eds. Society for Industrial and Applied Mathematics, Philadelphia, pp. 787–815.

[15] Horvitz, D.G., Shah, B.V. & Simmons, W.R. (1967). The unrelated question randomized response model, in *American Statistical Association 1967 Proceedings of the Social Statistics Section*. American Statistical Association, Alexandria, pp. 65–72.

[16] Horvitz, D.G., Greenberg, B.G. & Abernathy, J.R. (1975). Recent developments in randomized designs, in *A Survey of Statistical Design and Linear Models*, J.N. Srivastava, ed. North-Holland, New York, pp. 271–285.

[17] Horvitz, D.G., Greenberg, B.G. & Abernathy, J.R. (1976). Randomized response: a data gathering device for sensitive questions, *International Statistical Review* **44**, 181–196.

[18] Levy, K.J. (1976). Reducing the occurrence of omitted or untruthful responses when testing hypotheses concerning proportions, *Psychological Bulletin* **83**, 759–761.

[19] Liu, P.T. & Chow, L.P. (1973). A new randomized response technique: the multiple answer model. Department of Population Dynamics, Johns Hopkins University (unpublished paper).

[20] Liu, P.T. & Chow, L.P. (1976). The efficiency of the multiple trial randomized response technique, *Biometrics* **32**, 607–618.

[21] Liu, P.T. & Chow, L.P. (1976). A new discrete quantitative randomized response model, *Journal of the American Statistical Association* **71**, 72–73.

[22] Liu, P.T., Chow, L.P. & Mosley, W.H. (1975). Use of the randomized response technique with a new randomizing device, *Journal of the American Statistical Association* **70**, 329–332.

[23] Moors, J.J.A. (1971). Optimization of the unrelated question randomized response model, *Journal of the American Statistical Association* **66**, 627–629.

[24] Moors, J.J.A. (1981). Inadmissibility of linearly invariant estimators in truncated parameter spaces, *Journal of the American Statistical Association* **76**, 910–915.

[25] Moors, J.J.A. (1985). Estimation in truncated parameter spaces. *Doctoral dissertation*, Tilburg University, The Netherlands.

[26] Poole, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response, *Journal of the American Statistical Association* **69**, 1002–1005.

[27] Singh, J. (1976). A note on the randomized response technique, in *American Statistical Association 1976 Proceedings of the Social Statistics Section*. American Statistical Association, Alexandria, p. 772.

[28] Singh, J. (1978). A note on maximum likelihood estimation from randomized response models, in *American Statistical Association 1978 Proceedings of the Social Statistics Section*. American Statistical Association, Alexandria, pp. 282–283.

[29] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**, 63–69.

[30] Warner, S.L. (1971). The linear randomized response model, *Journal of the American Statistical Association* **66**, 884–888.

[31] Winkler, R.L. & Franklin, L.A. (1979). Warner's randomized response model: a Bayesian approach, *Journal of the American Statistical Association* **74**, 207–214.

L. FRANKLIN

Randomized Treatment Assignment

An important goal of clinical research is the development of therapies that improve the probability of a successful outcome in the ill or that prevent the onset of a disease in the healthy. The fundamental question asked in a **clinical trial**, "Does the therapy under investigation work," implies an interest in **causation**. Although frequently associations between therapy and outcome are observed, **association** alone does not imply causation. A rigorous answer to the question of whether a treatment actually benefits patients requires a direct concurrent comparison of those on the treatment to those not on it. Convincing evidence of effectiveness requires not only observing a difference between the two groups with respect to an outcome of interest, but also demonstrating that the therapy has most probably caused that difference. For example, patients undergoing a new surgical intervention may experience longer average survival time than patients who do not undergo the surgery. Whether that apparent benefit is due to the surgery itself or to the ability of surgeons to select patients of low surgical risk is relevant to the assessment of the effect of the intervention (*see* **Bias in Observational Studies**).

An experiment affords the best approach to inferring that an observed association reflects the actual effect of therapy, rather than a noncausal relationship. Experimentation, more rigorously than observation or a priori reasoning, isolates the effect of an intervention from systematic differences between the group of people being treated and the controls. To ensure an **unbiased** assessment of treatment, the study groups must be equivalent in all respects except for the treatment itself. In many clinical trials, the method used to render the groups equivalent is **randomization**, the allocation of people to treatment through a process governed by chance. Although lotteries, which reflect an understanding of the need for a chance mechanism to ensure unbiased selection, date back thousands of years, the application of random selection to experimentation is relatively recent. Stigler [29] credited the first application of randomization in experimentation to Pierce & Jastrow in 1885 [26]. **Fisher**, in a 1926 paper on agricultural experiments [8] and more generally in his *Statistical Methods for Research Workers* [10], stressed the centrality of randomization for statistical **inference**. He contended that statistical tests (*see* **Hypothesis Testing**) based on **normal** theory provided good approximations to the exact randomization distribution. While the literature [23] has cited Diehl's [7] clinical trial on vaccines for the common cold as the first clinical trial to randomize participants, Waller [31] has pointed out that Diehl probably used an alternating, not a random, sequence to assign treatments to participants. Armitage [1] states, "The successful implementation of randomized trials in medicine, in the 1940s, is largely due to the advocacy and example of Sir Austin Bradford Hill" (see, for example, [13]).

Randomization requires an active process of distributing experimental units by chance, not a passive, haphazard method of selection. According to Fisher,

> Apart ... from the avoidable error of the experimenter himself introducing with his test treatments, or subsequently, other differences in treatment, the effects of which the experiment is not intended to study, it may be said that the simple precaution of randomization will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged [9, p. 21].

Thus Fisher very early in his career stressed the importance of randomization to the valid calculation of experimental error.

Randomization has at least two purposes: it allows the deduction of causality and it removes **bias** from the selection of patients for specific treatments. The ability to deduce causality stems from the fact that randomization tends to balance the treatment groups with respect not only to known **prognostic factors**, but also with respect to unknown, unmeasured factors. By preventing investigators from consciously or unconsciously selecting patients for specific treatments, randomization removes investigators' bias from the process of the allocation of treatment groups.

Statistically, from the frequentist point of view, randomization has a third desirable feature: it allows construction of the sample space that renders statistical tests of significance valid.

While randomization is necessary for the unbiased comparison of treatments, it alone does not suffice to ensure fair comparisons. The actual assignment of participants to treatment must precisely reflect the randomization, and the formal analysis of the

data must respect the randomization; otherwise, the analysis, by violating the structural balance assured by the experimental technique, creates potential bias in the evaluation of the treatments (*see* **Intention to Treat Analysis**).

Necessity of Randomization in Clinical Trials

In many laboratory experiments, especially those in the physical sciences, the scientist has the tools to render the "subjects" equivalent. With the ability to exert exquisite control over the samples being compared, small experiments, fastidiously executed, may suffice to measure precisely the effects of stimuli. In biology, however, especially when the "subject" is the entire organism, the inherent variability among experimental units renders such tight control impossible. Even identical twins or genetic clones in the laboratory are not truly identical, because although they are genetically similar, they may differ considerably in experience and behavior. Unrelated people differ even more. Strategies to match experimental groups by controlling as much as possible the distribution of variables in the experimental groups help to achieve comparability, but the very large numbers of variables, measured and unmeasured, that characterize an individual make perfect balance impossible.

In addition to the difficulty of trying to select groups to make them comparable, in clinical trials the treating physician may have conscious or unconscious preference for a specific treatment for a given patient. Bias in construction of treatment groups can creep into a clinical trial quite subtly, because the treating physician may be reluctant to assign certain patients to certain arms. Thus an important feature of assignment of therapy to participant is unpredictability: the next treatment assignment must not be known in advance. Randomization offers the mechanism that typically achieves balance and unpredictability of assignment of treatment in experiments. Although any particular realization of a random allocation of subjects to groups may produce imbalance (otherwise, no one would ever be dealt a full house in poker!), the set of all possible random allocations does not in any way favor one group over another. Furthermore, a large sample size will yield groups in any specific experiment that will have a very high probability of being well balanced with respect both to baseline risk and intrinsic responsiveness to intervention.

Armitage [1] and Lachin et al. [21] present useful discussions of the need for randomization in clinical trials that aim to evaluate therapy (see also Friedman et al. [11]).

Randomization as the Basis of Statistical Inference

According to frequentist theory, randomization allows a direct test of cause and effect and permits construction of valid tests of statistical significance [14, 17]. A simplified description of the randomization model follows. Each patient in the study has a true value of the outcome variable. For example, in a trial that studies the effect of a new drug on the level of LDL-cholesterol (LDL) in the serum, the model conceptualizes each patient as characterized by a "true value" of LDL. The first step in the clinical trial is analogous to random shuffling of the deck of study participants into two piles. In the absence of assigned treatment, the shuffle is expected to produce two sets of cholesterol values with identical distributions.

A study with m patients in each of two treatment groups has $\binom{2m}{m}$ possible assignments of patients to the two groups. The mean levels of LDL in the two groups resulting from the random allocation are expected to be equal. Of course, in any specific allocation the means will differ from each other, but if the sample size is large, the random shuffling ensures a low probability of a large difference. To perform the clinical trial, one introduces a treatment, here a putative LDL-lowering medication, and assigns it to each member of one of the "piles". If the medication reduces LDL-cholesterol by 10 mg/dl, then the treated group is expected to have an average LDL-cholesterol level 10 mg/dl lower than the untreated group. To test the effect of treatment, one compares the observed **means** in the two populations. Under the randomization model, all possible ways the shuffle could have allocated participants to the treatment and control groups constitutes the sample space. A surprisingly large observed difference in the two groups leads to the claim that the treatment must have "caused" the difference because the only systematic characteristic that distinguishes

between the two groups is the treatment. To derive a statistical test of the **null hypothesis** of no effect of treatment, one can construct the sample space of all possible outcomes by enumerating each of the $\binom{2m}{m}$ combinations, computing the test statistic, and calculating in what position the observed sample sits in the space of all possible allocations ordered according to the **likelihood** of the occurrence under the null hypothesis. Even in relatively small samples, however, the sample space may be large enough to render this calculation intractable. In 1955, Kempthorne [17] proved Fisher's earlier contention that statistical tests based on normal theory were reasonable approximations to the randomization distributions (see also Hinkelmann & Kempthorne [14]). Lachin [18–21] presents discussions of the theory as applied specifically to randomized clinical trials. The act of randomization has provided the formal template that allows a statistical test of whether the treatment "caused" a difference in LDL. The answer in the study group leads to the inferential leap that if the treatment truly caused a difference in this specific group of people, then it should cause a difference in other people as well. The randomization procedure does not guarantee the legitimacy of this generalization of the trial's result to a larger population.

Arguments Against Randomization

While most clinical trialists view randomization as necessary to valid inference about the effects of therapy, some people object to randomization on ethical grounds (*see* **Ethics of Randomized Trials**). Some argue further that often randomization is difficult to perform, but according to Senn, "Contrary to what is sometimes claimed, randomization is not a nuisance in clinical trials: from the practical point of view it is one of the easiest allocation procedures to implement" [28]. Some hold that careful statistical modeling in the absence of randomization can in some cases discern the effect of therapy. For discussion of these arguments, see, for example, Basu [2], Levine [22], and Friedman et al. [11].

While randomization is necessary for the construction of statistical tests from the frequentist point of view, it is relevant to **Bayesian** and likelihood-based inference for another reason. Both of these methods of inference consider the data fixed after the experiment is completed; they, unlike in the frequentist approach, are not concerned with the experiments that might have occurred. Inference is based on the likelihood – the probability of the data given the parameters of interest – not the process that gave rise to the data. From the point of view of such inference, randomization enhances the validity of the likelihood. In fact, several Bayesian discussions of clinical trials urge randomization to ensure **blinding** and protection from **confounding** [16, 24]. Since the likelihood used in the Bayesian analysis is based on absence of confounding, randomization is essential.

When should Randomization Start and End?

The literature on clinical trials includes considerable discussion concerning the period during the life of development of a therapy that randomization is appropriate. Failure to randomize relatively early in the development of a new therapy or in the study of a new application of an accepted therapy may render interesting hypotheses untestable, for once the medical community regards a therapy to be safe and effective, physicians are reluctant to randomize to therapies they perceive as less effective. Chalmers [4] has therefore recommended randomizing from the first patient. On the other hand, if data from clinical trials have demonstrated the safety and effectiveness of a therapy, continuing to randomize to confirm a previous observation or to ask more refined questions about the therapy may raise ethical problems.

A reasonable practical approach is to start randomization quite early in the development of a new therapy, although not necessarily at the beginning, and to continue to randomize as long as legitimate uncertainty exists surrounding the safety and efficacy of the therapy [5] (*see* **Data and Safety Monitoring**).

Methods of Randomization

Randomization requires a mechanism governed by chance to assign treatments to people. The ideal allocation device is a perfectly unbiased coin tossed by an angel. Real clinical trials should use verifiable methods of randomization so that, after the study, the investigators can demonstrate that the allocation was free from bias. A coin flipped by a person is fallible chiefly because a less-than-honest coin flipper can fail to record tosses that land on the "wrong side".

Thus, the main problem with flipping coins is that the process permits no checks of the validity of the process.

In the past, many trials have used tables of random numbers to produce randomization lists. This method can be cumbersome for all but the simplest type of randomization. Furthermore, because published tables of random numbers are available to everyone, staff in a clinic can potentially find the sequence that is in use, thus permitting **selection bias** to affect the study.

Whenever possible, a computer should produce the randomization list. The person who generates the randomization list should be separate from the persons recruiting and treating participants. During the course of the study, the generator of the list should not divulge the details of the particular method used to construct the list to any of the clinic personnel involved with the patients.

Many studies use opaque sealed envelopes to hold the treatment assignment. This method, though quite standard, is subject to violation, especially in unmasked studies (*see* **Blinding or Masking**). An investigator intent on enrolling a patient into a specific treatment arm can hold the envelope to the light, or even open an envelope and not enter a patient if the "wrong" assignment is listed. In recent years, many clinical trials have adopted telephone, fax, or encrypted on-site computer codes for randomization. These methods allow a rigorous accounting of all persons entered into the study.

The randomization list itself should be held inviolate by the person or group controlling the assignment of treatments to participants. Because violations of randomization can invalidate the entire trial, the investigators should establish procedures for randomization and protection of the validity of the randomization before the trial begins.

For a study treatment that is viewed as a community-wide intervention, a trial might randomize one group of communities or clinics to receive the control treatment and another group to receive the study treatment. See, for example, a description of the COMMIT study [6]. Random allocation makes this approach theoretically acceptable provided that the **unit of statistical analysis** remains the community, not the individual person participating in the study. To have a **powerful** test of the effect of treatment, this method of allocation may require many communities. Moreover, the investigators must be careful to assess the impact that the intervention had on the entire community, not only on those people who participated in the program (*see* **Group-randomization Designs**).

So-called "randomized" clinical trials often use nonrandom approaches to assignment of treatment. In fact, people often use the term "random" loosely in the sense of apparent haphazardness, but strict randomization through a mechanism governed by chance prevents the treatment assignment from being predicted and protects against bias. The following paragraphs, which represent the control therapy by C and the experimental treatment by E, describe some common nonrandom ways of assigning patients. The considerations below apply with obvious modifications to the comparison of two active treatments or to the comparison of more than two study groups.

Some studies assign patients in alternating sequence. This scheme assigns the first participant to treatment C, the second to E, the third to C, and so forth. The argument adduced in favor of such alternating sequences is that because patients enroll in a "chance" order, a method that alternates patients to one treatment or other will result in groups of roughly equal risk. The flaw in this method stems from the fact that the treating physician who knows the sequence can choose which patients receive which therapy. Even if the therapy is blind, a single revelation of the treatment code unblinds the entire study. This type of assignment clearly violates the requirement of assigning treatment by chance in order to minimize physician bias. In addition, from a frequentist point of view, the sample space consists of only two possible realizations, ECECEC ... C and CECECE ... E. With only two possible outcomes, the exact two-sided ***P*-value** from the randomization test is 1.0, no matter how dramatically the two groups differ after treatment!

Another scheme allocates patients on alternating days to treatment C or E. This type of assignment has problems similar to the first one. From the frequentist view, the sample space again consists of two possible allocations, so that once more the exact P-value is always unity. From the point of view of biased selection, once the clinicians have deduced the scheme, they can control the allocation of patients to a particular therapy. Thus, this type of assignment is

subject to substantial bias in studies of nonemergency conditions. In true emergencies, practical exigencies may make such allocation necessary. If all patients are entered into the study, then the bias may be negligible since treatment cannot be delayed until the next day.

The medical literature includes many examples of studies that perform their randomization rigorously but fail to begin therapy until several hours, or even days, after randomization. Omitting from analysis the patients who drop out between the time of randomization and the time of therapy can introduce bias. Whether the dropouts in fact lead to bias depends on the reasons for failing to remain on the trial. If the treatments are unmasked, then bias should be highly suspected; if the treatments are masked, dropouts after randomization may not lead to bias. To avoid a biased comparison, the analysis should generally include all patients in the groups to which they were randomized (*see* **Intention to Treat Analysis**). When exceptions to this rule are made, the reasons for the exclusions must be strictly independent of treatment assignment.

Guidelines for Randomization

In a clinical trial that specifies the entire study sample before the experiment begins, the study statistician can construct the complete randomization schedule in advance and assign the treatments to the appropriate study participants. Phase I experiments of normal volunteers, experiments with dietary manipulations [15], or **vaccine trials** with closed populations [25] often have such prespecified study samples. The typical clinical trial, however, enters participants during a sometimes prolonged recruitment period. At the beginning of the trial, the people who will enroll are not known; perhaps some of them have not even yet manifested the disease to be studied. Thus, the assignment of treatments to participants must occur before identifying all participants. The process must be stepwise. First, the potential participant agrees to join the study, signs an informed consent (*see* **Ethics of Randomized Trials**) document, and is officially enrolled in the trial. Then the randomization process begins and the treatment is assigned. Once enrolled, a person is, except in certain special cases, part of the randomized trial, even if he or she fails to participate any further.

Some Problems with Simple Randomization

The primary purpose of randomization is to ensure comparability of the treatment and control groups. The type of randomization described above is simple randomization: each person arrives at the study, the study authority flips a theoretical coin that has probability p of assigning the participant to the control group, and, depending on the outcome, the patient receives treatment or control. On average, a proportion p, usually 1/2, of the assignments will be to treatment and $(1 - p)$ to **control**. A very large sample size guarantees that the proportions of people assigned each treatment will be arbitrarily close to the chosen values of p and $1 - p$. In practice, however, this kind of simple, unrestricted randomization will not produce exactly the prespecified proportions of patients in each study group.

Another important limitation of unrestricted randomization in clinical trials stems from the structure of the typical trial. A purely random process that generates a sequence of Cs and Es will produce, by chance, occasional long sequences of the same treatment assignment. Because participants enter trials over time, such sequences may produce unwanted homogeneity among patients entered at approximately the same time. Assignment to test and control therapies should ensure balance not only overall, but in addition, the length of sequences of the same treatment should be short and, at any particular time in the study, nearly the same number of patients should be on each therapy. Furthermore, unrestricted randomization can lead to some small clinics randomizing all, or nearly all, participants to the same treatment. Such an allocation is clearly undesirable.

Finally, some sequences of random numbers may by chance produce an unbalanced allocation with respect to a specific baseline variable. Therefore, if the primary outcome is strongly related to a specific prognostic variable, one might like to adopt a strategy that guarantees balance for that variable. Such constrained randomization is called **"blocking"** or **"stratification"**. While these words are technically synonymous, in randomized clinical trials "blocking" generally refers to randomization within small subsets without regard to specific prognostic variables, while "stratification" refers to randomization within subsets defined by categorical variables. To perform blocked randomization, one should use block sizes that are

multiples of the number of treatment groups, then assign at random the treatments within the blocks. For example, consider a study with three treatments A, B, and C to assign with equal probability. If the block size is six, then the assignment rule would allocate at random two each of A, B, and C to the six persons in each block. At the end of every six assignments, the numbers in the three treatment groups will necessarily be equal. The longest possible sequence, four of the same treatment, occurs when the last two participants in one block and the first two in the succeeding block are assigned the same treatment.

This more balanced allocation does not come without a price. Strictly, the fact that the data are blocked imposes extra complexity on any frequentist analysis, for the statistical analysis ought to reflect the randomization. Conventionally, however, the blocking is ignored in the analysis of clinical trials because the complexities incurred are generally not considered worth the very small gain in expected power. Lachin et al. [21] warn, ". . . if there is significant heterogeneity in some systematic way among the patients entering the trial, such as a change over time, then ignoring the stratification . . . may substantially distort the size of the test".

A potentially serious problem with simple blocking in unblinded studies is the fact that a clever investigator can deduce some of the allocations. For example, the first five assignments in each block of six always determine the last one. Smaller block sizes assure more balance but in unblinded trials they lead to a higher proportion of predictable allocations. In unblinded studies of treatment and control with blocks of size two, half of the allocations are known with certainty because the first allocation fully determines the configuration of the block. To prevent the clinic staff from knowing what patients are to be assigned what therapy, a study should take several precautions in selecting block sizes. First, the investigators should not know the block sizes. Secondly, the allocation should use an unpredictable mixture of block sizes, so that the sequence of assignments to therapy would confuse a person who tried to decode the system. Thirdly, block sizes can vary by clinic. In clinics that are expected to recruit many patients, the block size may be as high as 20. In clinics that are expected to recruit very small numbers of participants, the block size should be smaller to ensure reasonable balance of treatment assignments.

The type of simple or blocked randomization already described does not guarantee balance for specific baseline variables. Even in randomized clinical trials with large populations, simple unstratified randomization will lead to imbalance in some baseline variables. In fact, if tests are performed at a 5% level of significance, statistically significant imbalance is expected to occur in 1 out of 20 independent baseline variables. Although most observed imbalances will be dismissed as not germane to the question of effectiveness of therapy, sizable imbalance on important known prognostic factors can be unsettling. The statistical literature proposes several approaches to analysis of data when such imbalance occurs: ignoring the imbalance as an unlucky event but one within the range of outcomes in random assignment; "adjusting" through statistical models for the imbalance in specific variables if that imbalance is statistically significant; "adjusting" through statistical models for large imbalance in specific variables even if the degree of imbalance is not statistically significant; "adjusting" for all important prognostic variables when at least one important such variable shows considerable imbalance. The argument for the last choice is that the presence or absence of statistically significant imbalance should not be the driving force to ask whether actual allocation affects the conclusion (*see* **Covariate Imbalance, Adjustment for**).

Rather than relying on chance to avoid imbalance on important variables, one can create strata and then randomize within strata (or within blocks within strata). First, we consider precision. While heterogeneity does not generally affect the type I error (*see* **Hypothesis Testing**) – but see [21] – if there is considerable heterogeneity in the study population, then isolating sources of variation from the effect of treatment can sometimes lead to increased precision in the estimate of the treatment effect. One simple effective method to achieve such isolation is stratification of the study group into relatively homogeneous subgroups using a stratification variable strongly related to outcome. Gains in precision sometimes occur even if the sample is balanced overall.

In the absence of compelling reasons to the contrary, **multicenter trials** generally stratify randomization within clinic because of the potentially large differences among clinics. Different clinics may recruit from very different patient groups. Their approaches to treatment and concomitant therapies,

their threshold for aggressive therapy, the quality of their staff, and their available equipment may be quite heterogeneous. Furthermore, the degree to which the clinics adhere to the protocol may also vary and the quality of their data may differ markedly. Finally, even in large clinical trials the number of patients within individual clinics may be quite small. Thus allocation of treatment at random without regard to clinic leads to a high probability that some clinics will have very unequal numbers of patients assigned to the treatment groups.

Selection of the stratification variables presents a practical problem. Investigators often desire to stratify on many variables of interest. Sometimes these variables are highly **correlated**, so that stratification on one or two of them leads to near balance on the others. Thus, using a few stratifying variables can often achieve reasonable balance for all the variables of interest. A large number of strata may create so complex a routine in the clinic that the chance with which a patient receives the incorrect assignment increases. Too many strata may lead to excess costs. In drug trials, for instance, a sufficient quantity of the drug must be available for each stratum. Generally, the more strata, the more drug necessary, and often clinics will end up with unused drug. Depending on the cost of the drug and its phase of development, the need for excess drug can be very expensive. Finally, overstratification can lead in some extreme cases to *decreased* power. Too many strata can lead to incomplete strata. For example, suppose a trial to study the effect of LDL-lowering on the progression of atherosclerotic plaque uses only three important stratification variables: gender, smoking status (non-, former, or current smoker), and age (40–49, 50–64, >64). These three variables lead to $2 \times 3 \times 3 = 18$ possible strata. If all the strata were equally likely and all the clinics recruit the same number of people, a trial with 600 patients in 20 clinics would then have an expected stratum size of $600/(20 \times 18) = 1.7$, which is less than the number of treatments. Of course, some clinics will recruit more than others, and some strata are much more common than others. Nonetheless, many people will be the sole occupants of their randomization cells. A strict analysis according to the randomization would have to exclude those people. In practice, however, most data analyses would ignore the stratification [12].

Other Methods of Randomization

This discussion addresses only the usual methods of randomization in clinical trials. Other methods available include deterministic play-the-winner [33] or probabilistic urn model [32] (*see* **Ornstein–Uhlenbeck Process**) approaches to maximizing the probability of a participant's receiving the better therapy, adaptive allocation to achieve marginal balance of specific prognostic variables [3], and covariate adaptive procedures, both deterministic [30] and probabilistic [27] (*see* **Adaptive and Dynamic Methods of Treatment Assignment**). Another option is "prerandomization" to increase the pool of people willing to enter trials (*see* **Ethics of Randomized Trials**) [34].

Other Randomization in Clinical Trials

Randomization in clinical trials is not restricted to allocation of participants to treatments. Depending on the specific trial, a given study may randomize such items as the sequence of tests, the assignment of readers for diagnostic procedures, or the order in which a panel of judges views paired "before" and "after" measurements. Investigators should consider randomization to prevent bias at various critical points in the study.

Conclusion

Randomization, the allocation of treatments to study participants by an aleatory mechanism, ensures unbiased assignment of treatments to participants, guarantees the balance of treatment groups with respect to the expected distribution of measured and unmeasured baseline variables, and, under frequentist statistical theory, allows the calculation of experimental error. Coupled with complete follow-up of the study cohort and rigorous analytic strategies, randomization leads to unbiased tests of the null hypothesis of no difference between treatments.

References

[1] Armitage, P. (1982). The role of randomization in clinical trials, *Statistics in Medicine* **1**, 345–352.

[2] Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test, *Journal of the American Statistical Association* **75**, 575–582.

[3] Birkett, N. (1985). Adaptive allocation in randomized controlled trials, *Controlled Clinical Trials* **6**, 146–155.

[4] Chalmers, T. (1972). Randomization and coronary artery surgery, *Annals of Thoracic Surgery* **14**, 323–327.

[5] Collins, R. (1990). Discussion of papers on cost and efficiency, *Statistics in Medicine* **9**, 150.

[6] COMMIT Research Group (1991). Community Intervention Trial for Smoking Cessation (COMMIT): summary of design and intervention, *Journal of the National Cancer Institute* **83**, 1620–1628.

[7] Diehl, H.S., Baker, A.B. & Cowan, D.W. (1938). Cold vaccines: an evaluation based on a controlled study, *Journal of the American Medical Association* **111**, 1168–1173.

[8] Fisher, R.A. (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture of Great Britain* **33**, 503–513.

[9] Fisher, R.A. (1960). *The Design of Experiments*, 7th Ed. Hafner, New York.

[10] Fisher, R.A. (1970). *Statistical Methods for Research Workers*, 14th Ed. Hafner, New York.

[11] Friedman, L., Simon, R., Verter, J. & Wittes, J. (1984). Proceedings of the workshop on the evaluation of therapy, *Statistics in Medicine* **3**, 305–475.

[12] Hallstrom, A. & Davis, K. (1988). Imbalance in treatment assignments in stratified blocked randomization, *Controlled Clinical Trials* **9**, 375–382.

[13] Hill, A.B. (1962). *Statistical Methods in Clinical and Preventive Medicine*. Livingstone, London.

[14] Hinkelmann, K. & Kempthorne, O. (1994). *Design and Analysis of Experiments*, Vol. 1. Wiley, New York.

[15] Judd, J.T., Clevidence, B.A., Wittes, J., Muesing, R.A. & Podczasy, J.J. (1994). Dietary *trans* fatty acids: effects on plasma lipids and lipoproteins of healthy men and women, *American Journal of Clinical Nutrition* **59**, 861–868.

[16] Kadane, J.B. & Seidenfeld, T. (1996). Statistical issues in the analysis of data gathered in the new designs, in *Bayesian Methods and Ethics in a Clinical Trial Design*, J.B. Kadane, ed. Wiley, New York.

[17] Kempthorne, O. (1955). The randomization theory of experimental inference, *Journal of the American Statistical Association* **50**, 946–967.

[18] Lachin, J. (1988). Properties of randomization in clinical trials: foreword, *Controlled Clinical Trials* **9**, 287–288.

[19] Lachin, J. (1988). Properties of simple randomization in clinical trials, *Controlled Clinical Trials* **9**, 312–326.

[20] Lachin, J. (1988). Statistical properties of randomization in clinical trials, *Controlled Clinical Trials* **9**, 289–311.

[21] Lachin, J., Matts, J. & Wei, L. (1988). Randomization in clinical trials: conclusions and recommendations, *Controlled Clinical Trials* **9**, 365–374.

[22] Levine, R.J. (1986). *Ethics and Regulation of Clinical Research*, 2nd Ed. Yale University Press, New Haven.

[23] Lilienfeld, A.M. (1982). Ceteris paribus: the evolution of the clinical trial, *Bulletin of the History of Medicine* **56**, 1–18.

[24] Lindley, D.V. (1980). Comment on Basu, *Journal of the American Statistical Association* **75**, 589–590.

[25] Nosten, F., Luxemburger, C., Kyle, D., Ballou, W.R., Wittes, J., Wan, E., Chongsuphajaisiddhi, T., Gordon, D.M., White, N.J., Sadoff, J.C., Heppner, D.G. & the Shoklo SPf66 Malaria Vaccine Trial Group (1996). SPf66 randomized double-blind placebo-controlled trial of SPf66 malaria vaccine in children in Northwestern Thailand, *Lancet* **348**, 701–707.

[26] Peirce, C.S. & Jastrow, J. (1884). On small differences of sensation, *National Academy of Science Memoirs* **3**, 75–83.

[27] Pocock, S. & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial, *Biometrics* **31**, 103–115.

[28] Senn, S. (1994). Fisher's game with the devil, *Statistics in Medicine* **13**, 217–230.

[29] Stigler, S. (1986). *The History of Statistics*. The Belknap Press of Harvard University Press, Cambridge, Mass.

[30] Taves, D. (1974). Minimization: a new method of assigning patients to treatment and control groups, *Clinical Pharmacology and Therapeutics* **15**, 443–453.

[31] Waller, L. (1997). A note on Harold S. Diehl, randomization, and clinical trials, *Controlled Clinical Trials* **18**, 180–183.

[32] Wei, L.J. & Durham, S. (1978). The randomized play-the-winner rule in medical trials, *Journal of the American Statistical Association* **73**, 840–843.

[33] Zelen, M. (1969). Play the winner rule and the controlled clinical trial, *Journal of the American Statistical Association* **64**, 131–146.

[34] Zelen, M. (1979). A new design for randomized clinical trials, *New England Journal of Medicine* **300**, 1242–1245.

(*See also* **Randomization Tests**)

JANET WITTES

Randomness, Tests of

Tests of randomness are used to assess whether data are truly random or whether the data have some sort of pattern. For data that are collected over a period of time, such as blood chemistries for a patient, this pattern may be a relationship between the order in which the data were collected and the magnitude of a variable of interest. This may also be a relationship between a dichotomous characteristic of interest, such as gender, and the magnitude of another variable of interest. For data that are collected over geographic areas, such as mortality rates for a disease across counties, this may be a relationship between rate and location. Some tests of randomness are used to detect general nonrandomness, while other tests are used to detect a particular pattern of nonrandomness.

Many tests of randomness are based on the concept of runs. For example, suppose the outcome in each of a series of trials can be classified as either a success (S) or a failure (F). The results of nine trials are, in order,

F F F S S F S S S.

Each sequence of observations of the same type (success or failure) is called a "run", and in this case we have a total of four runs, two of S and two of F. The number of runs could be as few as two (F F F F S S S S S or S S S S S F F F F), if all failures occurred before all success or vice versa, or as many as nine (S F S F S F S F S), if successes alternated with failures. This illustrates the important point that "too few" or "too many" runs is evidence of a nonrandom relationship between the order of the experiments and the outcome, and we would reject the **null hypothesis** of randomness if either of these outcomes is the case. The Wald–Wolfowitz [16] runs test is the best known test that is based on the number of runs, but because this test compares the probability distribution of two populations and is not a test of randomness, the null hypothesis of identical distributions is rejected only if there are too few runs. Therefore, a test of randomness based on the number of runs can be thought of as a Wald–Wolfowitz runs test, but with a two-tailed rejection region (*see* **Alternative Hypothesis**).

Tests of randomness that focus on nonrandomness due to **clustering** can be based on the length of the runs. Mosteller [12] indicated that a test of randomness can be based on the length of the longest run, which is three in the above example. O'Brien & Dyck [13] developed a statistic using a weighted linear combination of the variances of the length of "success" runs and of "failure" runs. Larger variances suggest that the data are clustered. Agin & Godbole [1] developed a statistic based on the number of "success" runs of a given length, which can be used to detect cyclical clustering.

Tests of randomness that focus on nonrandomness due to trends can be based on the differences between values of successive observations, such as the von Neumann test [15]. This test is based on the *mean square successive difference*, the mean square of the difference of successive observations. Small squared differences between values of successive observations relative to squared difference between the values of the observations and their mean value are indicative of a trend. The rank von Neumann test of Bartels [2] is similar to the von Neumann test, except that this test uses the ranks of the observations. This test is also more powerful that the von Neumann test under certain conditions.

Daniels [5] and Mann [10] developed tests for nonrandomness due to trends that **correlate** the value of a variable with the time or order that the variable was measured, using **Spearman's** rho (Daniels) or Kendall's tau (Mann) (*see* **Rank Correlation**). A high positive or negative correlation suggests a trend. Dietz & Killeen [6] provided a multivariate extension to Mann's test which detects a trend in at least one variable.

An important application of tests of randomness involves spatial data. Some tests are used to assess whether the locations of point data are randomly distributed over a given region or whether the data are either aggregated or are distributed too regularly. These tests are primarily used in biological applications, and involve the use of techniques such as counting the number of observations in subregions, or quadrats, of the region of interest, or computing distances between each observation and its nearest neighbor [3] or its kth nearest neighbor.

Other tests involving spatial data are used to assess whether there is a relationship between a characteristic of a population and the location of that population (*see* **Geographic Epidemiology**). Cliff & Ord [4] explore the relationship between a dichotomous variable and geographic area based on the BW statistic, classifying areas as either black (B) or white (W) and

counting the number of times a black area borders, or is linked to, a white area. A small value of BW suggests that the data are aggregated, while a large value suggests that the data are too regularly distributed. This statistic is equivalent to an extension of the runs test to multidimensional data when observations are linked using orthogonal minimum spanning trees [7]. The multiresponse permutation procedure statistic discussed by Mielke et al. [11] may also be used to link data. Hubert et al. [9] discussed the relationship between unidimensional and multidimensional tests of randomness, showing that many tests, including some of those discussed above, have the same general form.

Tests of randomness can also be used to assess the randomness of a sequence of **pseudo-random numbers** generated by a given algorithm. This is important, because the validity of **simulation** techniques are dependent on whether generated random numbers are truly random. Some tests of randomness are discussed by Strube [14] and Gruenberger & Jaffray [8]. These include investigating the length of intervals between repetitions of the same number, the correlation between numbers that are close in the sequence, and the frequency of occurrence of certain numbers. In addition, goodness-of-fit tests can also be applied to a sequence of generated random numbers to test for randomness.

References

[1] Agin, M.A. & Godbole, A.P. (1990). A new exact runs test for randomness, in *Computer Science and Statistics: Proceedings of the Symposium on the Interface*, C. Page & R. LePage, eds. Springer-Verlag, New York, pp. 281–285.

[2] Bartels, R. (1982). The rank version of von Neumann's ratio test for randomness, *Journal of the American Statistical Association* **77**, 40–46.

[3] Clark, P.J. & Evans, F.C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations, *Ecology* **35**, 445–453.

[4] Cliff, A.D. & Ord, J.K. (1973). *Spatial Autocorrelation*. Pion, London, pp. 1–52.

[5] Daniels, H.E. (1950). Rank correlation and population models, *Journal of the Royal Statistical Society, Series B* **12**, 171–181.

[6] Dietz, E.J. & Killeen, T.J. (1981). Anonparametric multivariate test for monotone trend with pharmaceutical applications, *Journal of the American Statistical Association* **76**, 169–174.

[7] Friedman, J.H. & Rafsky, L.C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests, *Annals of Statistics* **7**, 697–717.

[8] Gruenberger, F. & Jaffray, G. (1965). *Problems for Computer Solution*. Wiley, New York.

[9] Hubert, L.J., Golledge, R.G., Costanzo, C.M. & Gale, N. (1985). Tests of randomness: unidimensional and multidimensional, *Environment and Planning, Series A* **17**, 373–385.

[10] Mann, H.B. (1945). Nonparametric tests against trend, *Econometrica* **13**, 245–259.

[11] Mielke, P.W., Berry, K.J. & Johnson, E.S. (1976). Multiresponse permutational procedures for a priori classifications, *Communications in Statistics – Theory and Methods* **5**, 1409–1424.

[12] Mosteller, F. (1941). Notes on an application of runs to quality control charts, *Annals of Mathematical Statistics* **12**, 228–232.

[13] O'Brien, P.C. & Dyck, P.J. (1985). A runs test based on run lengths, *Biometrics* **41**, 237–244.

[14] Strube, M.J. (1983). Tests of randomness for pseudorandom number generators, *Behavior Research Methods & Instrumentation* **15**, 536–537.

[15] von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance, *Annals of Mathematical Statistics* **12**, 367–395.

[16] Wald, A. & Wolfowitz, J. (1940). On a test of whether two samples are from the sample population, *Annals of Mathematical Statistics* **11**, 147–162.

F.S. WHALEY

Range

The range of a set of observations is defined to be the difference between the largest and the smallest values of the set. For **grouped data**, it is taken to be the difference between the upper limit of the last interval and the lower limit of the first interval. In practice, the **standard deviation** of a set of measurements is roughly equal to one-sixth of the range.

In life testing, when the observations are **exponentially distributed**, the traditional **likelihood ratio test** statistic cannot answer questions of how to compare location parameters or threshold values. In **multiple comparisons** among these location parameters, the range statistic is more appropriate.

The range statistic can also be used to test the hypothesis that a set of location parameters varies "within a small but negligible difference" [3] (*see* **Equivalence Trials**; **Bioequivalence**). This is equivalent to testing the hypothesis that the location parameters fall into the zone of indifference specified in advance. When observations follow a normal distribution, the range and **studentized** range statistics can be used to test the hypothesis that all the mutual differences among means are smaller than a negligible quantity.

Exponential Distribution

Let $\pi_1, \ldots, \pi_k$ denote $k \geq 2$ populations such that the n independent observations $X_{i1}, \ldots, X_{in}$ taken from population π_i are exponentially distributed with density

$$f(x; \alpha_i, \theta) = \left(\frac{1}{\theta}\right) \exp\left[-\frac{(x-\alpha_i)}{\theta}\right],$$
$$\alpha_i < x < \infty, \theta > 0,$$

zero elsewhere, where α_i is an unknown location parameter or guaranteed life span, and θ is a common but unknown scale parameter or standard deviation.

Let Y_i denote the first **order statistic** of the sample of size n from population π_i. Let $Y_{[1]} \leq \cdots \leq Y_{[k]}$ denote the ordered values of the k first-order statistics $Y_1, \ldots, Y_k$, and take as estimator of θ

$$\hat{\theta} = \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{(X_{ij} - Y_i)}{[k(n-1)]}.$$

Define the range statistic C to be $C = n(Y_{[k]} - Y_{[1]})/\hat{\theta}$. Under the null hypothesis $H_0 : \alpha_1 = \cdots = \alpha_k$, C has the density

$$g(c) = (k-1) \sum_{i=0}^{k-2} (-1)^i \binom{k-2}{i} \times \left[1 + \frac{c(i+1)}{v}\right]^{-(v+1)}, \quad c > 0,$$

where $v = k(n-1)$.

We note that, if $k = 2$, then C has an ***F* distribution** with $(2, 2v)$ **degrees of freedom**. Furthermore, the distribution of C **converges** to that of R as v goes to infinity, where R is the range of independent observations from the standard exponential distribution with the density $f(x) = e^{-x}$ for $x > 0$.

The range statistic C provides a quick test for $H_0 : \alpha_1 = \cdots = \alpha_k$ vs. $H_a : \alpha_i \neq \alpha_j$; one rejects H_0 at α level of significance if the computed value of C is larger than $c^{\alpha}_{k,v}$, where $c^{\alpha}_{k,v}$ was given by Chen [2] for $\alpha = 0.10, 0.05, 0.01$, $k = 2(1)5, 10(10)50, 100$, and $n = 2(1)6(2)10, 16, 30, 60, \infty$.

The range statistic C is not as **powerful** as the likelihood ratio test (LR). However, if multiple comparison among αs is of interest, the LR test is not applicable. Like Tukey's studentized range statistic for comparing normal means, the major role of the range statistic C rests on its extensive use in multiple comparisons of the location parameters. It is easy to find that a set of exact $1 - \alpha$ **simultaneous confidence intervals** for the difference $\alpha_i - \alpha_j$ is given by

$$\alpha_i - \alpha_j \in (Y_i - Y_j) \pm \frac{c^{\alpha}_{k,v} \hat{\theta}}{n}$$

for all $i \neq j = 1, \ldots, k$. When the sample sizes are not equal, we suggest replacing $1/n$ by $\frac{1}{2}(1/n_i + 1/n_j)$ and $k(n-1)$ by $v = \sum_i (n_i - 1)$ to obtain a conservative set of confidence intervals. Furthermore, a set of exact $1 - \alpha$ simultaneous confidence intervals for all linear contrasts of the location parameters is given by

$$\sum_{i=1}^{k} a_i \alpha_i \in \sum_{i=1}^{k} a_i Y_i \pm a^{\alpha}_{k,v} \frac{\hat{\theta}}{n} \sum_{i=1}^{k} \frac{1}{2} |a_i|,$$

where $\sum a_i = 0$.

Normal Distribution

It is well known that, given a large enough sample size, a point **null hypothesis** $\mu_1 = \mu_2 = \cdots = \mu_k$ will always be rejected. In applications, the point hypothesis is unrealistic; more appropriate is an "interval" null hypothesis: $H_0 : (\mu_{[k]} - \mu_{[1]})/\sigma \leq \delta$ vs. the **alternative hypothesis** $H_a : (\mu_{[k]} - \mu_{[1]})/\sigma > \delta$, where $\mu_{[1]} \leq \cdots \leq \mu_{[k]}$ denote the ordered population means, and $\delta(> 0)$ must be specified in advance (*see* **Ordered Alternatives**). The null hypothesis states that all standardized differences between means fall into a zone of indifference specified by a quantity δ, while the alternative describes the practically meaningful differences among means which are defined to fall in the preference zone.

Let there be k independent populations $\pi_1, \ldots, \pi_k$ such that observations obtained from π_i are independent and normally distributed with unknown mean μ_i and a common unknown variance $\sigma^2, i = 1, \ldots, k$. Our objective is to test the interval hypothesis H_0 using the range or the studentized range.

Let $X_{ij} (j = 1, \ldots, n)$ be an independent random sample of size n from population π_i. Define

$$\overline{X}_i = \sum_{j=1}^{n} \frac{X_{ij}}{n},$$

$$S = \left\{ \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2}{[k(n-1)]} \right\}^{1/2},$$

$$i = 1, 2, \ldots, k, \tag{1}$$

and let $\overline{X}_{[1]} \leq \cdots \leq \overline{X}_{[k]}$ be the order statistics of $\overline{X}_1, \ldots, \overline{X}_k$. The hypothesis H_0 is rejected at the α level of significance if

$$\overline{X}_{[k]} - \overline{X}_{[1]} > \frac{\gamma S}{\sqrt{n}}, \tag{2}$$

where γ is the solution to the equation

$$\begin{aligned} \alpha = 1 - l \int_0^\infty \int_{-\infty}^\infty & [\Phi(y + \gamma u) - \Phi(y)]^{l-1} \\ & \times [\Phi(y - \delta\sqrt{n} + \gamma u) - \Phi(y - \delta\sqrt{n})]^{k-l} \\ & \times \phi(y) q_v(u)\, \mathrm{d}y\, \mathrm{d}u - (k-l) \int_0^\infty \int_{-\infty}^\infty [\Phi(y \\ & + \delta\sqrt{n} + \gamma u) - \Phi(y + \delta\sqrt{n})]^l [\Phi(y + \gamma u) \\ & - \Phi(y)]^{k-l-1} \phi(y) q_v(u)\, \mathrm{d}y\, \mathrm{d}u, \end{aligned} \tag{3}$$

where $l = k/2$ for even k, and $l = (k-1)/2$ for odd k. The table of critical values γ was tabulated by Bau et al. [1] for $\alpha = 0.01, 0.05, \delta = 0.10, 0.20, 0.25, 1/3, 0.5$, $k = 2(1)10$, and $n = 2(1)20(2)30(10)60, 80, 100, 200$. If the sample sizes are not all equal, then we suggest replacing n by n_i and $k(n-1)$ by $\sum_i (n_i - 1)$ in (1) and n in (2) and (3) by the average sample size to obtain an approximate solution.

References

[1] Bau, J.J., Chen, H.J. & Xiong, M. (1993). Percentage points of the studentized range test for dispersion of normal means, *Journal of Statistical Computation and Simulation* **44**, 149–163.

[2] Chen, H.J. (1982). A new range statistic for comparisons of several exponential location parameters, *Biometrika* **69**, 257–260.

[3] Chen, H.J., Xiong, M. & Lam, K. (1993). Range test for the dispersion of several location parameters, *Journal of Statistical Planning and Inference* **36**, 15–25.

HUBERT J. CHEN

Rank Correlation

If X and Y are two observations on the same unit, both measured on an ordinal scale, then there are many circumstances in which it is important to assess the degree of **association** or **correlation** between X and Y. Such a measure of correlation may be used as a measure of **reliability** (when X and Y are two measurements of the same construct from independent observers or at two different times), or as a measure of **validity** (when X is a measurement of a construct and Y is a criterion defining the construct), or as a measure of **heritability** (when X and Y are two measurements taken from identical twins raised apart; *see* **Twin Analysis**), as well as a measure of more general types of association. The most common approach to assessing such association arises when, in a population sample, (X, Y) have a **bivariate normal distribution**, in which the correlation coefficient ρ is a natural parameter. Then the Pearson (product-moment) correlation coefficient r_{P} is the maximum likelihood estimator of ρ:

$$r_{\mathrm{P}} = \frac{\sum (X_i - M_X)(Y_i - M_Y)}{(N-1)s_X s_Y}.$$

Here the sample of size N from the population (X_i, Y_i), $i = 1, 2, \ldots, N$, has sample means M_X and M_Y, and sample standard deviations s_X and s_Y.

While r_{P} is reasonably robust to deviations from the assumptions of the bivariate normal distribution [6, 10], it can be misleading when the deviations are major. Consequently, an alternative approach, appropriate to measuring such association under less restrictive assumptions, is of value. This has given rise to proposals for rank correlation coefficients, which, like r_{P}, are measured on the interval from -1 to $+1$, are sensitive to monotonic association, take on a value of ± 1 for perfect positive or negative association, take on a value of 0 when X and Y are independent, but make no other distributional assumptions. The two most widely used such sample rank correlation coefficients were those proposed by Spearman [12] and by Kendall [3].

To compute the **Spearman rank correlation** coefficient, r_{S}, the set of observed Xs are rank-ordered from 1 to N, with tied observations assigned the average of the associated **ranks**. The same is done separately with the observed Ys. Thus, the sum of the ranks of either X or Y is always $N(N+1)/2$. Then the equation above for r_{P} is applied to the ranks rather than to the raw data. Clearly, r_{S} is invariant under strictly monotonic increasing transformations of either X or Y or both.

To compute Kendall's tau, τ, every pair of (X, Y) observations is assessed. For any pair of bivariate observations, say (X_i, Y_i) and (X_j, Y_j), a score of $+1$ is assigned to the pair if $\mathrm{sign}(X_i - Y_i) = \mathrm{sign}(X_j - Y_j)$ and -1 otherwise. The score of zero is assigned in the case of a tie. These scores are summed over the $N(N-1)/2$ possible pairs of observations to obtain S. Then,

$$\tau = \frac{S}{\left\{\left[\tfrac{1}{2}N(N-1) - U\right]\left[\tfrac{1}{2}N(N-1) - V\right]\right\}^{1/2}}.$$

If there are no ties in the X ranking, then $U = 0$; if there are no ties in the Y ranking, then $V = 0$. In that case, the denominator is simply $N(N-1)/2$. When there are ties, one counts the number of tied values in each set of tied values to obtain one value of u (for the X values) or v (for the Y values). Then,

$$U = \frac{1}{2}\sum u(u-1), \qquad V = \frac{1}{2}\sum v(v-1).$$

Once again, clearly, τ is invariant under strictly monotonic increasing transformations of either X or Y or both.

Both r_{S} and τ satisfy the general desiderata of a correlation coefficient. However, if (X, Y) were drawn from a bivariate normal population, then r_{P} would be a **consistent** estimator of ρ, r_{S} of $6/\pi[\sin^{-1}(\rho/2)]$, and τ of $2/\pi[\sin^{-1}(\rho)]$. Clearly, while the three sample correlation coefficients are measured on the same range, have the same random value, and values indicating perfect association, no two of them are measuring correlation in exactly the same way. In the case of the bivariate normal population, the differences between the parameters estimated by r_{P} and r_{S} are not major, at most a difference in the second decimal place. The parameter estimated by τ, however, may differ by as much as 0.2 from either of the other two. When the bivariate normal assumptions do not hold, the discrepancy between the r_{S} and τ may be even larger. The circumstances under which one would be preferred to the other have not been clearly enunciated. Most commonly, r_{S} is used, but largely because of its greater ease of computation and its closer correspondence to the familiar r_{P}.

In most contexts in which the correlation coefficient is used, it is clear from the outset that the population value is not exactly zero. Yet the most common statistical task is to test the **null hypothesis** that the correlation coefficient is zero. For small sample sizes, the exact distributions of both r_S and τ under the assumption of independence of X and Y are tabled [4].

For larger sample sizes, the distribution, under the null hypothesis of randomness, of each of these statistics is approximately normal. As a standard normal test statistic to test that hypothesis, one might use

$$\frac{3\tau[N(N-1)]^{1/2}}{[2(2N+5)]^{1/2}} \quad \text{or} \quad r_S(N-1)^{1/2}.$$

While it would be preferable to present **confidence intervals** rather than to perform a statistical test on a null hypothesis known a priori to be untrue, the exact non-null distributions of τ and r_S are in general unknown, since they depend on the parent distribution. If the parent distribution were bivariate normal (or the observed X and Y represented any monotonic **transformations** of X^* and Y^* drawn from a bivariate normal distribution), one might obtain approximate confidence intervals in the following way, using methods developed for r_P. Fisher's z-transformation for a correlation coefficient r is defined as $z(r) = \frac{1}{2}\ln[(1+r)/(1-r)]$. Then, two-tailed $100(1-\alpha)\%$ level confidence intervals for $z(\rho)$ are given approximately by [2]

$$z(r_S) \pm Z^{\alpha/2}\left[\frac{1.060}{(N-3)}\right]$$

$$\text{or} \quad z(\tau) \pm Z^{\alpha/2}\left[\frac{0.437}{(N-4)}\right],$$

where $Z^{\alpha/2}$ are the critical values of the standard normal distribution. **Bootstrap** methods have also proved useful for this purpose [9, 11].

The Spearman rank correlation coefficient concept can be extended to the situation where, for each unit sampled from the population, an m-dimensional vector $(X_1, X_2, \ldots, X_m)$ is observed, using Kendall's *coefficient of concordance* [5]. There are several equivalent ways to compute this coefficient, W, all based on rank ordering the units (averaging the ranks of ties) on each of $X_1, X_2, \ldots, X_m$ separately. The most revealing way is this: the Spearman rank correlation coefficient between each pair of variables is computed, and r_{S-ave} is the average Spearman rank correlation coefficient over all $m(m-1)/2$ possible pairs. The coefficient of concordance, W, equals

$$W = \frac{[1+(m-1)r_{S-ave}]}{m}.$$

It should be noted that the coefficient of concordance, W, equals $1/m$, rather than zero, when all the Xs are independent, but that it does equal $+1$ for perfect positive association. Thus, W itself does not satisfy the general qualities required of a rank correlation coefficient as stated above. Moreover, logically one cannot have perfect negative association when $m > 2$, for if X_i and X_j were perfectly negatively associated, and if X_j and X_k also were, then that would automatically mean that X_i and X_k would have to be perfectly positively associated. For this reason, the most common situation for application of this approach is that of reliability or validity assessment, where it is assumed that the pairwise correlations are all positive.

An alternative method of computation is based on applying the formula for an intraclass correlation coefficient (itself based on two-way **analysis of variance** for N units by m observations [1]; *see* **Correlation**) to the ranks. Again, for small samples, the distributions under the null hypothesis of total independence have been tabled [4]. However, once again, given the context of use, there is seldom any a priori doubt that the true correlation exceeds zero. Thus, confidence interval estimation would be preferable. It has been shown that when the distribution of the m-dimensional vector is **multivariate normal**, with equal correlation coefficients between each pair of observations, with large sample size, the distribution of r_{S-ave} is approximately that of the intraclass correlation coefficient based on the actual observed values [7, 8]. Again, since r_{S-ave} is invariant under any strictly increasing transformation of any or all of the m-variables involved, the observed values themselves does not require a multivariate normal distribution in order for the approach to yield a good approximation. Both tests and confidence interval estimation may be based on this approximation [8]. In more general circumstances, bootstrap estimation might be used.

References

[1] Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability, *Psychological Reports* **19**, 3–11.

[2] Fieller, E.C., Hartley, H.O. & Pearson, E.S. (1957). Tests for rank correlation coefficients, I, *Biometrika* **44**, 470–481.

[3] Kendall, M.G. (1938). A new measure of rank correlation, *Biometrika* **30**, 91–93.

[4] Kendall, M. & Gibbons, J.D. (1990). *Rank Correlation Methods*, 5th Ed. Oxford University Press, New York.

[5] Kendall, M.G. & Stuart, A. (1939). The problem of *m* rankings, *Annals of Mathematical Statistics* **10**, 275–287.

[6] Kowalski, C.J. (1972). On the effects on non-normality on the distribution of the sample, *Journal of the Royal Statistical Society* **21**, 1–12.

[7] Kraemer, H.C. (1976). The small sample non-null properties of Kendall's Coefficient of Concordance for normal populations, *Journal of the American Statistical Association* **71**, 608–613.

[8] Kraemer, H.C. & Korner, A.F. (1976). Statistical alternatives in assessing reliability, consistency or individual differences for quantitative measures: application to behavioral measures of neonates, *Psychological Bulletin* **83**, 914–921.

[9] Lunneforg, C.E. (1985). Estimating the correlation coefficient: the bootstrap, *Psychological Bulletin* **98**, 209–215.

[10] Norris, R.C. & Hjelm, H.F. (1961). Nonnormality and product moment correlation, *Journal of Experimental Education* **29**, 261–270.

[11] Rasmussen, J.L. (1987). Estimating correlation coefficients: bootstrap and parametric, *Psychological Bulletin* **101**, 136–139.

[12] Spearman, C. (1910). Correlation calculated from faulty data, *British Journal of Psychology* **3**, 271–295.

HELENA CHMURA KRAEMER

Rank Regression

The term *rank regression* was coined by Cuzick [7] to denote a regression model in which the **ranks** of the dependent variable were regressed on a set of covariates. The approach can be viewed as a hybrid between M-estimation and R-estimation (*see* **Robustness**). It differs from M-estimation in that the dependent variable is replaced by a score based on ranks and from R-estimation in that the ranks of the dependent variable itself are used, not those of the **residuals**. Since the ranks are unaffected by strictly increasing transformations, this approach can be formally specified by the model

$$g(t) = \boldsymbol{\beta}'\mathbf{z} + \varepsilon, \tag{1}$$

where g is any strictly increasing transformation of the dependent variable t, $\mathbf{z}$ is a vector of covariates, $\boldsymbol{\beta}$ are the corresponding regression coefficients and ε is the error term, usually taken to be mean zero. These models have also been called *transformation models*. The Box–Cox [3] model in which g is restricted to be a power law, or the logarithm, forms a well-known parametric submodel (*see* **Power Transformations**).

In general, the intercept term is undefined for this model, since it can be incorporated in g, and, for the same reason, information on scale is relative to the error term. Thus, when $\mathrm{E}(\varepsilon^2) < \infty$, one usually takes $\mathrm{var}(\varepsilon) = 1$ and interprets the regression coefficients in terms of number of the standard errors.

The most well-known rank regression model is the **proportional hazards** model

$$\lambda(t|z) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}),$$

where $\lambda_0(t)$ is the unknown and completely unspecified baseline hazard function, and $\exp(\boldsymbol{\beta}'\mathbf{z})$ is the relative risk term indicating how covariates affect the hazard. This model can be rephrased in terms of (1) by letting

$$g(t) = \log \int_0^t \lambda_0(s)\,\mathrm{d}s,$$

taking ε to have (minus) an **extreme value** distribution:

$$p(\varepsilon \geq x) = \exp(-\mathrm{e}^x),$$

and changing the sign of $\boldsymbol{\beta}$. The motivation for rank regression models has come from survival analysis, but their use is not restricted to this area. More general rank regression models allow ε to have different distributional forms (e.g. **normal** or **Pareto**). The Pareto form has an interesting interpretation in terms of a proportional hazards model with an unknown (or unmeasured) regression parameter that is assumed to have a log **gamma** distribution and to be independent of the error term and the other covariates [6]. A useful special case is when the **frailty** also has a log-exponential distribution, leading to a symmetric error distribution (**logistic**). This model has an interpretation in terms of proportional **odds** [1], i.e.

$$\frac{F(t|\mathbf{z})}{1 - F(t|\mathbf{z})} = \exp(\boldsymbol{\beta}\mathbf{z})\frac{F_0(t)}{1 - F_0(t)},$$

where $F_0(t)$ is the baseline distribution of t and $F(t|\mathbf{z})$ is the distribution when the covariates take the value $\mathbf{z}$. Fully efficient and computationally feasible methods for estimating $\boldsymbol{\beta}$ (and g) are only known for the proportional hazards model, although Wu [18] has an efficient estimator for the proportional odds model in the two-sample case. Dabrowska & Doksum [8] have constructed a class of $n^{1/2}$-consistent and asymptotically normal estimates for the general Pareto model in the two-sample case. Klaassen [12] has investigated methods based on solving Sturm–Liouville equations for the Pareto model, and Magaluri [15] has some general theoretical results. Cuzick [7] proposes a method for estimating $\boldsymbol{\beta}$ for known general error distributions that are shown to be $n^{1/2}$-consistent and asymptotically normal, provided a consistent initial estimate exists. In essence, this method consists of replacing $g(t)$ by an estimate based on the ranks of the observed values of t and the marginal distribution of $\boldsymbol{\beta}'\mathbf{z} + \varepsilon$ (which depends on $\boldsymbol{\beta}$), and then using the **maximum likelihood** (ML) estimating equation for $\boldsymbol{\beta}$ corresponding to the distribution of the error ε. Specifically, for independent identically distributed (iid) samples $\{t_i, \mathbf{z}_i, i = 1, \ldots, n\}$ and error distribution F, let

$$F_b(t) = \frac{1}{n}\sum_{i=1}^{n} F(t - b\mathbf{z}_i)$$

for general b, and define

$$\bar{t}_i^b = F_b^{-1}\left(\frac{R_i}{n+1}\right),$$

where R_i is the rank of t_i. If $\phi = (\log F')'$ is the influence function for ε, then solve

$$\sum_{i=1}^{n} \mathbf{z}_i \phi(\overline{t}_i^b - b\mathbf{z}_i) = 0$$

for $\hat{b}$. The method also provides an estimate of the covariance matrix for $\hat{b}$, and $g(t)$ can be estimated in a $n^{1/2}$-consistent manner by

$$\hat{g}(t_i) = \overline{t}_i^{\hat{b}}(t_i),$$

with interpolation between the t_i. Weak convergence of this estimator was established. The estimate can be extended to deal with censoring by replacing $(R_i/n + 1)$ with the **Kaplan–Meier estimator** and replacing ϕ with $(\log F)'$ when t_i is censored. Further details can be found in [2].

Cheng et al. [5] have proposed a similar approach that is more easily analyzed when there is censoring. They define

$$\begin{aligned}\xi(\mathbf{z}_{ij}'\boldsymbol{\beta}) &= \Pr(\varepsilon_i - \varepsilon_j > \mathbf{z}_{ij}'\boldsymbol{\beta}) \\ &= \mathrm{E}(I\{g(t_i) - g(t_j)\}|z_i, z_j),\end{aligned}$$

where $\mathbf{z}_{ij} = \mathbf{z}_i - \mathbf{z}_j$. Note that $\xi(\mathbf{z}_{ij}'\boldsymbol{\beta})$ depends only on the error distribution for ε, which is assumed known so that this can be computed. They then define the **estimating function**

$$U(\boldsymbol{\beta}) \equiv \sum_{j=1}^{n}\sum_{i=1}^{n} w(\mathbf{z}_{ij}'\boldsymbol{\beta})\mathbf{z}_{ij}\{I\{t_i > t_j\} - \xi(\mathbf{z}_{ij}', \boldsymbol{\beta})\},$$

where w is a weight function, and choose a root $\hat{\boldsymbol{\beta}}$ of $U(\boldsymbol{\beta}) = 0$ as the estimate. To mimic the quasi-likelihood approach, the weight function is taken as

$$w(0) = \frac{\xi'(0)}{\{\xi(0)[1 - \xi(0)]\}}.$$

When censoring is present, and the potential censoring times are assumed to be iid with survival function $G(t) = P(c \geq t)$, they note that

$$\mathrm{E}\left(\left.\frac{\Delta_j I\{t_i \geq t_j\}}{G^2(t_j)}\right|_{z_i, z_j}\right) = \xi(\mathbf{z}_{ij}'\boldsymbol{\beta}_0),$$

where $\Delta_j = I\{T_j \geq C_j\}$ is the censoring indicator. The estimating function is modified to become

$$\begin{aligned}U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{n} & w(\mathbf{z}_{ij}'\boldsymbol{\beta})\mathbf{z}_{ij} \\ & \times \left\{\frac{\Delta_j I(t_i \geq t_j)}{\hat{G}^2(t_j)} - \xi(\mathbf{z}_{ij}'\boldsymbol{\beta})\right\},\end{aligned}$$

where $\hat{G}(\cdot)$ is the Kaplan–Meier estimator for the survival function G of the **censoring** distribution. Cheng et al. [5] show that if the weights $w(\cdot)$ are positive, then $U(\boldsymbol{\beta}) = 0$ asymptotically has a unique solution and that when $w \equiv 1$ and $\sum\sum \mathbf{z}_{ij}'\mathbf{z}_{ij}$ is positive definite, the above equation has a unique solution for all n. Asymptotic normality is established and an expression for the variance is given that can be approximated from the data.

Lai & Ying [13] have also used the term *rank regression* to refer to models based on ranks of residuals in censored regression models. These models have a very different character and are based on the "aligned-rank" or R-estimator methods in [10, 16], and [11] for noncensored data. The classical approach is based on solving the estimating equation

$$\sum_{i=1}^{n} \mathbf{z}_i \phi(t_i - b\mathbf{z}_i) = 0,$$

where ϕ is the influence function for the chosen error distribution (see also [9]). Note that here t_i is used directly, not after transformation into $\overline{t}_i^b$, and the main goal of this approach is to provide robustness against misspecification of the error distribution. Early results for the two-sample problem with censoring were given in [14] and [17]. Another early approach along these lines for censored data was explored by Buckley & James [4]. They chose $\phi(t) = t$ corresponding to **least squares** regression, and developed an extension for censored data in which censored observations were replaced by their expected values based on current estimates of $\boldsymbol{\beta}$ and the Kaplan–Meier estimator. For dealing with censoring (and truncation), an estimating equation based on a weighted **logrank test** and martingale theory has proved more analytically tractable but is technically very demanding [19].

References

[1] Bennett, G. (1983). Analysis of survival data by the proportional odds model, *Statistics in Medicine* **2**, 273–277.

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[3] Box, G.E. & Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

[4] Buckley, J. & James, I. (1979). Linear regression with censored data, *Biometrika* **66**, 429–436.

[5] Cheng, S.C., Wei, L.J. & Ying, Z. (1995). Analysis of transformation models with censored data, *Biometrika* **82**, 835–845.

[6] Clayton, D. & Cuzick, J. (1985). The semi-parametric Pareto model for regression analysis of survival times, in *Proceedings of the Forty-fifth Session of the International Statistical Institute* 23.3-1–23.3-18. International Statistical Institute, Amsterdam. (Also in *Papers on Semiparametric Models at the ISI Centenary Session*, R.D. Gill and M.N. Voors, eds. Report MS-R8169. Centre for Mathematics and Computer Science, Amsterdam, pp. 19–30).

[7] Cuzick, J. (1988). Rank regression, *Annals of Statistics* **16**, 1369–1389.

[8] Dabrowska, D.M. & Doksum, J.A. (1988). Estimation and testing in a two-sample generalized odds-rate model, *Journal of the American Statistical Association* **83**, 747–749.

[9] Hájek, J. & Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

[10] Hodges, L.J., Jr & Lehmann, E.L. (1963). Estimates of location based on rank tests, *Annals of Mathematics and Statistics* **34**, 598–611.

[11] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

[12] Klaassen, C.A.J. (1988). Efficient estimation in the Clayton-Cuzick model for survival data. Preprint, University of Leiden.

[13] Lai, T.L. & Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data, *Annals of Statistics* **19**, 531–556.

[14] Louis, T.A. (1981). Nonparametric analysis of an accelerated failure time model, *Biometrika* **68**, 381–390.

[15] Magaluri, G. (1993). Semiparametric estimation of association in a bivariate survival function, *Annals of Statistics* **21**, 1648–1662.

[16] Sen, P.K. (1963). On the estimation of relative potency in dilution (-direct) assays by distribution-free methods, *Biometrics* **19**, 532–552.

[17] Wei, L.J. & Gail, M.H. (1983). Nonparametric estimation for a scale-change with censored observations, *Journal of the American Statistical Association* **78**, 382–388.

[18] Wu, C.O. (1995). Estimating the real parameter in a two-sample proportional odds model, *Annals of Statistics* **23**, 376–395.

[19] Ying, Z.L. (1993). A large sample study of rank estimation for censored regression data, *Annals of Statistics* **21**, 76–99.

(*See also* **Survival Distributions and Their Characteristics**)

JACK CUZICK

Rank Transformation

If a set of univariate observations are ordered from the smallest to the largest, then the position of an observation in the ordering is termed its **rank**. The smallest observation has rank 1, the next smallest rank 2, and the largest observation has rank equal to the number of observations. If for a set of observations on a variable X, denoted $x_1, x_2, \ldots, x_n$, $r(x_i)$ denotes the rank of the ith observation, then a rank transformation of X generates the set of ranks $r(x_1), r(x_2), \ldots, r(x_n)$.

Rank transformations of response variables play a fundamental role in many **nonparametric methods**. More generally, the **response variable** measurements for a set of observations are uniquely represented by their ranks and their **order statistics**. Model-based inference based on ranks typically derives from the **marginal likelihood** generated from the marginal distribution of the ranks. For example, the **partial likelihood** used for the **Cox regression model** in **survival analysis** corresponds to a marginal distribution based on ranks for uncensored survival data with no ties.

Rank transformations of explanatory variables in regression models may also be used. Their use might be motivated by a reluctance to rely too heavily on the measured values of the explanatory variable and can be regarded as a more general procedure than grouping the variable into a small number of classes. It has been suggested that rank transformations are particularly useful for **variable selection** [2].

A discussion of the link between standard parametric analysis procedures on rank transformed data and nonparametric procedures is provided in [1]. Ranking is also an essential component of statistical procedures based on **scores**.

References

[1] Conover, W.J. & Iman, R.J. (1981). Rank transformations as a bridge between parametric and nonparametric statistics, *American Statistician* **35**, 124–133.

[2] O'Gorman, T.W. & Woolson, R.F. (1993). On the efficacy of the rank transformation in stepwise logistic and discriminant analysis, *Statistics in Medicine* **12**, 143–151.

(*See also* **Rank Correlation**; **Rank Regression**)

VERN T. FAREWELL

Ranks

We are all familiar with ranks and rankings as they occur in everyday life. Ranks arise naturally in situations where performance or some other quality possesses a natural ordering from the "best" or "first" to the "worst" or "last", as, for example, the ordering of finishers in a race. If the n objects to be ranked are represented by the symbols $(o_1, o_2, \ldots, o_n)$, we denote by $r_i = r(o_i)$ the rank assigned to object o_i. Thus,

$$\mu = [r(o_1), r(o_2), \ldots, r(o_n)] = (r_1, r_2, \ldots, r_n)$$

is a *ranking* of the n objects, and is a permutation or rearrangement of the n integers $(1, 2, 3, \ldots, n)$. For example, if a judge is asked to rank four contestants (objects), then a ranking of these may be displayed as

$$\begin{array}{cccccc} \text{Object} & 1 & 2 & 3 & 4, \\ \text{Rank} & 3 & 2 & 4 & 1, \end{array}$$

where the object subscripts appear in the first line, so that in this case

$$r_1 = r(o_1) = 3, \qquad r_2 = r(o_2) = 2,$$
$$r_3 = r(o_3) = 4, \qquad r_4 = r(o_4) = 1.$$

Sometimes we may find it more convenient to think in terms of an *ordering* of the objects, in which case the above example may be represented as

$$\begin{array}{cccccc} \text{Rank} & 1 & 2 & 3 & 4, \\ \text{Object} & 4 & 2 & 1 & 3, \end{array}$$

i.e. rank 1 goes to o_4, rank 2 goes to o_2, rank 3 goes to o_1, and rank 4 goes to o_3.

The space of all possible rankings consists of the $n!$ permutations of the integers $(1, 2, 3, \ldots, n)$, so that, for example, with $n = 4$ objects we find that the $4! = 4 \times 3 \times 2 \times 1 = 24$ possible rankings are given by

$$\begin{array}{cccccc} 1234 & 1243 & 1324 & 1342 & 1423 & 1432, \\ 2134 & 2143 & 2314 & 2341 & 2413 & 2431, \\ 3124 & 3142 & 3214 & 3241 & 3412 & 3421, \\ 4123 & 4132 & 4213 & 4231 & 4312 & 4321. \end{array}$$

In such spaces of rankings one may be interested in probability models and in various statistical questions. One model is the so-called "null" or uniform model, which assumes that every possible ranking in the space is equally likely. Under such a model each object can equally achieve any of the n ranks, so that the expected rank of a given object is simply the sum of the first n integers divided by n,

$$\mathrm{E}(r_i) = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{(n+1)}{2}.$$

Using the fact that the sum of squares of the first n integers is $n(n+1)(2n+1)/6$, we can also easily compute the variances and covariances of the ranks under the uniform model,

$$\mathrm{var}(r_i) = \frac{(n^2 - 1)}{12},$$
$$\mathrm{cov}(r_i, r_j) = -\frac{(n+1)}{12}, \quad 1 \le i \ne j \le n.$$

The properties of ranks under the uniform model are fundamental to the development of many nonparametric tests (*see* **Nonparametric Methods**), in that many such tests can be represented as *linear rank statistics*, i.e. as linear combinations of functions of the ranks.

In practice, it may be difficult for judges to rank a large number of objects, as, for example, if one is asked to taste test many varieties of apples. In such a situation we may prefer to present for comparison every one of the $\frac{1}{2}n(n-1)$ pairs of objects separately. **Paired comparisons** are discussed in [5] and [12]. Note that, although every ranking uniquely determines the outcome of every paired comparison, not every set of $\frac{1}{2}n(n-1)$ paired comparisons can be resolved into a ranking. Thus, in comparing three objects, a judge may prefer o_1 to o_2, o_2 to o_3, and yet prefer o_3 over o_1, creating what Kendall refers to as a "circular triad".

Models for the mechanism by which individuals generate rankings and models for nonnull distributions on the space of rankings can incorporate paired comparisons as long as only those preferences which can produce a ranking are permitted. Such an approach was originally introduced by Babington Smith [16]. Other approaches utilize **order statistics** and distances between rankings, and these originate with Thurstone [17] and Mallows [13]. A model which decomposes the ranking process into stages is discussed in Fligner & Verducci [10].

If, in fact, we actually observe a **random sample** $x_1, x_2, \ldots, x_n$ from a continuous distribution, then the order statistics are an ordering of these variables

from the smallest to the largest, denoted by $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, so that the rank of the ith order statistic $x_{(i)}$ is $r(x_{(i)}) = i, i = 1, \ldots, n$. The rank of the observation x_i is the value $r(x_i) = r_i$ equal to the number of observations $\leq x_i$.

Distances between rankings arise naturally in any situation where we need to measure how "close" two different rankings are. Various such distances can be defined and some of the more useful ones are studied in Diaconis & Graham [7]. If we denote the distance between the rankings μ_1 and μ_2 of judges 1 and 2 by $d(\mu_1, \mu_2)$, then the **rank correlation** between the two rankings can be defined as

$$\alpha(\mu_1, \mu_2) = 1 - \frac{2d(\mu_1, \mu_2)}{M},$$

where M is the largest possible distance between two rankings. By choosing an appropriate distance we can generate both **Spearman's** and Kendall's rank correlation, among others. In the case of Spearman, the appropriate distance is the squared Euclidean distance between the two rankings, which we call Spearman's distance, defined as

$$d_s(\mu_1, \mu_2) = \frac{1}{2} \sum_{i=1}^{n} [r_1(o_i) - r_2(o_i)]^2,$$

for which the maximum is $M_S = n(n^2 - 1)/6$.

In a **randomized block design**, where m judges (blocks) rank n objects (treatments), the average distance between all $\frac{1}{2}m(m-1)$ pairs of rankings measures the level of agreement or *concordance* between the judges and generates a test statistic for treatment effects (*see* **Agreement, Measurement of**). When the distance is Spearman, this measure is essentially Kendall's coefficient of concordance W, [12], and the test reduces to Friedman's test. Other distances will generate different test statistics, so that, for example, the distance which generates Kendall's rank correlation yields a concordance statistic introduced by Ehrenberg [9] and studied in Alvo et al. [3].

Up to this point we have assumed that all objects are to be ordered so that there is a strict preference. However, either by design or by circumstance, there may not be a preference between two or more objects. In such situations the space of possible rankings is no longer the space of permutations of $(1, 2, 3, \ldots, n)$.

For example, a judge may be asked to pick from a list of eight possible qualities of a mate the three qualities that best describe an ideal partner. If the best three are to be ordered, then the ranks would be 1, 2, 3, for the top three and a 4 for each of the remaining five least preferred qualities. An example in which the best three qualities are o_4, o_2, and o_1, in that order, would give the *partial ranking*

Quality	1	2	3	4	5	6	7	8,
Rank	3	2	4	1	4	4	4	4.

If, on the other hand, the top three are not required to be ordered, then the partial ranking will contain only two distinct numbers, a 1 for each of the top three, and a 2 for each of the remaining seven. The above example then becomes

Quality	1	2	3	4	5	6	7	8,
Rank	1	1	2	1	2	2	2	2.

Distances between partial rankings may be defined in various ways and used to approach various statistical questions. An important work in this area is Critchlow [4].

A similar problem arises when a distinct ranking is expected but for some reason the ranking has *ties*. For example, the ranking may have been generated from a random sample $x_1, x_2, \ldots, x_n$ in which not all the x_i values are distinct. How we assign values to the tied ranks depends on the situation. In the context of certain nonparametric tests the usual approach is to assign each set of tied objects the *midrank*, i.e. the average of the ranks they would have received had they not been tied. This is done partly to ensure that $\mathrm{E}(r_i)$, the expected rank of observation i, remains $(n+1)/2$. The example above, in which the first three are ordered and the rest and tied, would now become

Quality	1	2	3	4	5	6	7	8,
Rank	3	2	6	1	6	6	6	6.

A very good discussion on dealing with ties may be found in Pratt & Gibbons [15]. Note that the presence of ties has various implications on the null distributions of many nonparametric tests.

Certain situations may arise as in the apple tasting example above, where it is much easier for the judges to rank a small set of objects. In such a case we may want to present to the judges only a subset of k of the possible n objects for their consideration. The ranking of this subset is known as an *incomplete ranking*. Such rankings may also arise by chance if, for example, the ranking arises from a random

sample $x_1, x_2, \ldots, x_n$ in which some of the x_i values are missing. Distances between incomplete rankings and corresponding measures of rank correlation are discussed in Alvo & Cabilio [2].

The pattern of k out of n objects (treatments) presented to each of m judges may be designed to follow a **balanced incomplete block design**. The classical test for treatment effect in such a situation is a generalization of Friedman's test due to Durbin [8]. The Durbin statistic turns out to be essentially the average of the Spearman distances between all $\frac{1}{2}m(m-1)$ pairs of incomplete rankings, and an analogous statistic can be defined using the Kendall distance (see [1]).

There has been a resurgence of interest in the area of ranking models and rank-based statistical methods in recent years. Some noteworthy books are that by Diaconis [6] and a very complete coverage of the subject by Marden [14]. Also of some interest is a collection of papers in this area edited by Fligner & Verducci [11].

References

[1] Alvo, M. & Cabilio, P. (1991). On the balanced incomplete block design for rankings, *Annals of Statistics* **19**, 1597–1613.

[2] Alvo, M. & Cabilio, P. (1995). Rank correlation methods for missing data, *Canadian Journal of Statistics* **23**, 345–358.

[3] Alvo, M., Cabilio, P. & Feigin, P. (1982). Asymptotic theory for measures of concordance with special reference to Kendall's tau, *Annals of Statistics* **10**, 1269–1276.

[4] Critchlow, D.E. (1985). *Metric Methods for Analyzing Partially Ranked Data*. Springer-Verlag, New York.

[5] David, H.A. (1988). *The Method of Paired Comparisons*. Oxford University Press, New York.

[6] Diaconis, P. (1988). *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics, Hayward.

[7] Diaconis, P. & Graham, R.L. (1977). Spearman's footrule as a measure of disarray, *Journal of the Royal Statistical Society, Series B* **39**, 262–268.

[8] Durbin, J. (1951). Incomplete blocks in ranking experiments, *British Journal of Psychology* **4**, 85–90.

[9] Ehrenberg, A.S.C. (1952). On sampling from a population of rankers, *Biometrika* **39**, 82–87.

[10] Fligner, M.A. & Verducci, J.S. (1988). Multistage ranking models, *Journal of the American Statistical Association* **83**, 892–901.

[11] Fligner, M.A. & Verducci, J.S., eds. (1993). *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, New York.

[12] Kendall, M. & Gibbons, J.D. (1990). *Rank Correlation Methods*. Oxford University Press, New York.

[13] Mallows, C.L. (1957). Non-null ranking models I, *Biometrika* **44**, 114–130.

[14] Marden, J.I. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall, New York.

[15] Pratt, J.W. & Gibbons, J.D. (1981). *Concepts of Nonparametric Theory*. Springer-Verlag, New York.

[16] Smith, Babington B. (1950). Discussion on symposium on ranking methods, *Journal of the Royal Statistical Society, Series B* **12**, 183–187.

[17] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Reviews* **34**, 273–286.

(*See also* **Signed-rank Statistics**; **Wilcoxon Signed-rank Test**; **Wilcoxon–Mann–Whitney Test**; **Wilcoxon-type Scale Tests**)

PAUL CABILIO

Rao–Blackwell Theorem

This theorem gives a general way of reducing the variance of an **unbiased** estimator when a **sufficient statistic** is available. It also brings about the relevance of sufficiency to the question of seeking a **minimum variance unbiased estimator** (see [3] and [5]).

The most familiar version of the Rao–Blackwell theorem is stated in the following form.

Theorem A. Let $(P_\theta : \theta \in \Theta)$ be a family of probability distributions on a sample space X and suppose that $\tilde{\tau}$ is an unbiased estimator of a real-valued function τ of θ. Then if $T = T(X)$ is a sufficient statistic for θ, $\hat{\tau} = \mathrm{E}_\theta[\tilde{\tau}|T]$ is also an unbiased estimator of $\tau(\theta)$, and $\mathrm{var}_\theta(\hat{\tau}) \leq \mathrm{var}_\theta(\tilde{\tau})$ for all θ; that is, $\hat{\tau}$ is a uniformly better unbiased estimator.

An important use of this theorem is to establish the existence of uniformly minimum variance unbiased estimators. This is done by demanding the existence of a sufficient statistic T with an additional property such as *completeness* (*see* **Sufficient Statistic**). For example, let $X_1, X_2, \ldots, X_n$ be independently, identically distributed (iid) in a **Poisson distribution** with mean θ, and suppose that $\tau(\theta) = \mathrm{e}^{-\theta}$ is the parameter of interest. Let $\hat{\tau} = 1$ if $X_1 = 0$, and $\hat{\tau} = 0$ if $X_1 \neq 0$. This is obviously an unbiased estimator of $\tau(\theta)$. It is clear that, if we consider the statistic $\hat{\tau} = \mathrm{E}_\theta(\tilde{\tau}|T)$, where $T = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic, then $\hat{\tau} = [1 - 1/n]^{\mathrm{T}}$ and is the unique minimum variance unbiased estimator of $\tau(\theta)$ (see [5, Sections 2.5 and 2.6]).

In the event that $\tau(\theta)$ is a vector or $\tilde{\tau}$ is not unbiased, Theorem A does not apply. We give below a more general version of the Rao–Blackwell theorem formulated in terms of **decision theory**.

Theorem B. Let the action space A be a convex subset of R^k, and suppose that the **loss function** $L(\theta, a)$ is a convex function of $a \in A$ for each $\theta \in \Theta$. Suppose that T is a sufficient statistic for θ. If $\tilde{\tau}(X)$ is a decision rule, then the decision rule based on T, defined by

$$\hat{\tau}(T) = \mathrm{E}[\tilde{\tau}(X)|T],$$

is a decision rule that is as good as $\tilde{\tau}$ provided that this expectation exists; that is, the **risk** of $\hat{\tau}$ is no larger than that of $\tilde{\tau}$.

The Rao–Blackwell theorem was proved by Rao [4] and Blackwell [1] for unbiased estimators with squared error loss in the form of Theorem A. The general version, Theorem B, can be found in [3] or [2].

References

[1] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation, *Annals of Mathematical Statistics* **18**, 105–110.

[2] Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

[3] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[4] Rao, C.R. (1945). Information and accuracy attainable in estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society* **37**, 81–91.

[5] Silvey, S.D. (1975). *Statistical Inference*. Chapman & Hall, London.

S.H. Lo

Rasch Models

In psychological tests (*see* **Psychometrics, Overview**) or attitude studies, we are often interested in quantifying the value of an unobservable *latent trait*, such as mathematical ability or manual dexterity, on a sample of individuals (*see* **Path Analysis**). While latent traits are not directly measurable, we assume that we can assess indirectly a person's value for the latent trait from his/her responses to a set of well-chosen items on a test. In some studies, interest centers on traits associated with the items in the test (for example, difficulty of each item) rather than on individual traits.

Initial attempts at modeling latent variables included *formal measurement models*, two examples of which are the standard factor models, and what are known as Thurstone models for attitude scaling (see, for example, Molenaar [32]; **Factor Analysis, Overview**; **Latent Class Analysis**; **LISREL**). It is now widely recognized that a better alternative to classical test theory is what is known as *Item Response Theory* (IRT) – see, for example, [21] and [7]. Item response theory states that when a person is confronted with an item (for example, a question in a test), the probability of a certain response is a function of the person's position on the latent trait, plus one or more parameters associated to the particular item. For each item, the *item response function* (IRF) is the probability of a certain answer given as a function of the latent trait value.

A fundamental component of IRT is the family of Rasch models (RM), introduced by Danish statistician **Georg Rasch** [33]. Given responses from n individuals to k items in a test, the RM permits the estimation of parameters associated with individuals and with items, as well as prediction of the person's behavior when confronted with a different set of items from the same domain. The individual parameter is often referred to as *ability*, while item parameters refer to the *difficulty* (or simplicity) of each item. When the ability parameter is high and the difficulty parameter is low, the probability of a correct answer to the item increases. Here, we view "persons", "items", and "responses", in a general context, with "persons", for example, representing perhaps laboratory animals or households. In the example given later, "persons" represent individuals, "items" represent different influenza outbreaks, and "responses" are binary (individuals get sick or not in each outbreak).

Rasch models can be of at least two kinds. If the number of possible answers to each item is two, then the RM is called *dichotomous* (*see* **Binary Data**); otherwise, the model is said to be **polytomous**. In the dichotomous RM, positive answers are indicative of a high position in the latent trait scale.

The Rasch model from psychological testing is a simple but very important logistic response model (*see* **Logistic Regression**) that allows for the incorporation of individual effects. It has important technical links to a subclass of the better-known **loglinear models** involving **quasi-symmetry**. As a consequence we can think of the heterogeneity resulting from the Rasch model's individual effects as inducing a specific form of dependence in the loglinear model used to describe the cross-classification of responses to several variables when we aggregate across individuals. The Rasch model thus serves as a heuristic for interpreting this special kind of heterogeneity in more general loglinear model settings. There are, of course, alternative approaches to heterogeneity such as stratification and the use of regression-like components utilizing additional explanatory variables that "account for" heterogeneity, but we do not pursue these here. Among regression-like approaches to modeling heterogeneity, *generalized mixed linear (or nonlinear) models* (*see* **Generalized Linear Model**) have received increased attention recently, in particular in the context of **longitudinal data analysis**; see, for example, [28] and [14], and the references given therein (*see* **Generalized Linear Models for Longitudinal Data**).

This article introduces the Rasch model, describes the important link to loglinear models, and illustrates the Rasch model approach to interpreting interactions in the context of an example on infection in response to a series of influenza outbreaks.

The Rasch Model

We focus on the RM for the dichotomous case. Let $S_1, S_2, \ldots, S_n$ denote the n individuals providing binary responses to k items $I_1, \ldots, I_k$ that measure the same latent trait θ. It is assumed that each individual S_i has a value θ_i that reflects his/her position on the latent trait scale (ability). Furthermore, each item

I_j has a parameter α_j associated with it that denotes the difficulty of the item. (Generalizations of the simple RM allow for more than one parameter associated with an item. For example, an additional item parameter may reflect a change in the difficulty of the item for an individual who takes the same test at two different times. For more details, refer to Fischer [22, 23] and Embretson [17].)

If we let $\mathbf{X}$ denote the $n \times k$ matrix of responses, and $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ denote the vectors of item and individual parameters, respectively, the simple dichotomous RM states that

$$\Pr(X_{ij} = x_{ij} | \boldsymbol{\theta}_i, \boldsymbol{\alpha}_j) = \frac{\exp[x_{ij}(\theta_i - \alpha_j)]}{1 + \exp(\theta_i - \alpha_j)}, \qquad (1)$$

where the entries x_{ij} are either 0 or 1. Thus

$$\log\left[\frac{\Pr(X_{ij} = 1 | \theta_i, \alpha_j)}{\Pr(X_{ij} = 0 | \theta_i, \alpha_j)}\right] = \theta_i - \alpha_j, \qquad (2)$$

so the RM is evocative of the *logit* model for the log **odds** for $X_{ij} = 1$ vs. $X_{ij} = 0$ (*see* **Logistic Regression**).

A standard approach to modeling the matrix of responses $\mathbf{X}$ is to assume independence of items *and* of individuals. The assumption of independence among persons is one we typically make in multivariate problems (*see* **Multivariate Analysis, Overview**), and when it fails to hold we often turn to dependence structures described by loglinear models. The assumption of independence of answers within an individual, however, deserves some explanation. Intuitively, given a sample of individuals with different values for the latent trait, we would expect a positive correlation between the value of θ and the number of positive answers. For a given individual the model assumes, however, that all systematic variation between items is explained by the value of θ. Thus, we can use a conditional independence argument and say that, given θ, the responses of an individual to different items are independent. This is referred to as *local independence*. Under these assumptions, the likelihood function for the matrix of responses $\mathbf{X} = \mathbf{x}$ is given by

$$\Pr(\mathbf{X} = \mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \frac{\exp[x_{ij}(\theta_i - \alpha_j)]}{1 + \exp(\theta_i - \alpha_j)}$$

$$= \frac{\prod_{i=1}^{n} \exp\left[\theta_i x_{i+} - \sum_{j=1}^{k} \alpha_j x_{ij}\right]}{\prod_{i=1}^{n} \prod_{j=1}^{k} (1 + \exp[\theta_i - \alpha_j])}, \qquad (3)$$

where $x_{i+} = \sum_{j=1}^{k} x_{ij}$ are the individual *scores* or total number of correct answers for each individual. Model (3) is overparameterized, since for any constant c and for $\theta_i^* = \theta_i + c$ and $\alpha_j^* = \alpha_j + c$ we have that $\theta_i^* - \alpha_j^* = \theta_i - \alpha_j$. We can estimate the parameters in the model, however, by imposing a restriction that fixes the origin of the scale. Typically, we use the restriction $\alpha_+ = \sum_j \alpha_j = 0$ but, equivalently, we could also set $\theta_+ = 0$.

One very appealing aspect of the model is the existence of very simple **sufficient statistics** for both θ_i and α_j. Note that model (3) has an **exponential family** form, and so it is simple to show that the sum of correct answers x_{i+} for an individual is sufficient for the individual parameter θ_i, and the sum of correct answers for an item across individuals, $x_{+j} = \sum_i x_{ij}$, is sufficient for the item parameter α_j (see, for example, [5], [6], [30], and [31]). Because these sufficient statistics are in fact the "margins" of the matrix $\mathbf{X}$, we should not be surprised to discover that there are interesting links between methods of estimation for RMs and certain loglinear models.

Several estimation methods for the parameters in (3) have been proposed – for example, see [21] and [24]. Essentially, unrestricted **maximum likelihood** (ML) estimation has problematic asymptotic properties. Since the model includes one parameter for each individual in the sample, as $n \to \infty$ the number of parameters also goes to infinity, and thus ML estimators of individual and of item parameters are inconsistent. We refer the reader to [5] for a detailed discussion of ML estimation in the RM and to [27] for an interesting Bayesian treatment. The method of *conditional maximum likelihood estimation* (CML), first proposed by Rasch [33] and based on maximizing the likelihood function conditional on the individual scores, is an alternative to unrestricted estimation and gives rise to consistent estimators for item parameters. Further, such a conditional approach to estimation provides some direct links to maximum likelihood estimation for *loglinear models* applied to derived contingency tables. We describe this linkage below.

RM and its Relation to Loglinear Models

The conditional approach to likelihood estimation (CML) was suggested initially by Rasch, who noted that the conditional distribution of $\mathbf{X}$ given the individual marginal totals $\{X_{i+} = x_{i+}\}$ depends only on the item parameters, $\boldsymbol{\alpha}$. Each of the row sums $\{X_{i+}\}$ can take only $k+1$ distinct values corresponding to the number of correct responses. Next, we recall the alternate representation of the data in the form of an $n \times 2^k$ array, $\{\mathbf{W}\} = \{W_{ij_1 j_2,\ldots,j_k}\}$, where $W_{ij_1 j_2,\ldots,j_k} = 1$ if individual i responds $\{j_1, j_2, \ldots, j_k\}$ to the k items, and $= 0$ otherwise. Adding across individuals we create a 2^k contingency table, $\mathbf{Y}$, with entries $Y_{j_1 j_2,\ldots,j_k} = W_{+j_1 j_2,\ldots,j_k}$.

We can work with this collapsed array since all of the information we need is the response pattern, i.e. $\{j_1, j_2, \ldots, j_k\}$, and the number of "correct" responses that correspond to that pattern. Such information allows us to completely reconstruct the original matrix of responses, $\mathbf{X}$, except for the labeling of individuals, and thus we can use the 2^k array $\mathbf{Y}$ to represent the conditional distribution of $\mathbf{X}$ given $\{X_{i+} = x_{i+}\}$.

Duncan [16] and Tjur [34] independently noted that we can estimate the item parameters for the *conditional Rasch model* that arises from expression (1) using the 2^k array $\mathbf{Y}$, and certain loglinear models. The conditional RM is obtained by conditioning on the value of the individual scores $\{x_{i+}\}$. The resulting expression is

$$\Pr(X_{ij} = x_{ij},\, j = 1, 2, \ldots, k | \{x_{i+}\}, \boldsymbol{\alpha}) = \frac{\exp\left(-\sum_{j=1}^{k} \alpha_j x_{+j}\right)}{\gamma_i(\alpha_1, \ldots, \alpha_k)}, \quad (4)$$

where $\gamma_i(\alpha_1, \ldots, \alpha_k)$ are elementary symmetric functions given by

$$\gamma_{x_{i+}}(\alpha_1, \ldots, \alpha_k) = \sum_{x_1} \cdots \sum_{x_k} \exp\left(-\sum_{j} \alpha_j x_{+j}\right), \quad (5)$$

subject to $\sum_j x_{ij} = x_{i+}$.

More specifically, Tjur [34] shows that maximum likelihood estimation of the 2^k contingency table of expected values, $\mathbf{m} = \{m_{j_1 j_2,\ldots,j_k}\}$, using a **Poisson** sampling scheme and the loglinear model

$$\log m_{j_1 j_2,\ldots,j_k} = \omega + \sum_{j=1}^{k} \delta_j - \sum_{i=1}^{n} \log \gamma_{x_{i+}}(\alpha_1, \ldots, \alpha_k), \quad (6)$$

with $\omega = \log n$ and $\delta_j = -\alpha_j x_{+j}$, leads to consistent estimators of item parameters in the conditional RM (4). The sums of elementary symmetric functions in (6) turn out to depend only on the values for the totals $\{x_{i+}\}$, and thus there are only $k+1$ distinct values. Tjur proves this equivalence by: (i) assuming that the individual parameters are independent identically distributed random variables from some completely unknown distribution, π; (ii) integrating the conditional distribution of $\mathbf{X}$ given $\{X_{i+} = x_{i+}\}$ over the mixing distribution, π; (iii) embedding this **"random-effects"** model in an "extended random model"; and (iv) noting that the likelihood for the extended model is equivalent to that for expression (6) applied to $\mathbf{Y}$. An important technical issue, not explored further here, is the set of moment inequalities that must be satisfied for π (see [10]). Kelderman [29] gives a step-by-step derivation of the loglinear models that correspond to both the unconditional RM of expression (1) and the conditional RM of expression (4), and Fienberg & Meyer [19] present a related description and also an equivalent representation in the form of a **multiplicative model**, which we reproduce here in Table 1.

The *multiplicative parameters* a, b, and c in this Table correspond to δ_1, δ_2, and δ_3, and the multiplicative parameters $\{S_i\}$ correspond to the $k+1$ distinct values of the sums of elementary symmetric functions in (6). The minimal sufficient statistics are

$$\{y_{i++}\}, \{y_{+j+}\}, \{y_{++k}\}$$

Table 1 Multiplicative form for the expected values $\mathbf{m}$ in the Rasch model for the 2^3 table

		Item C			
		Yes		No	
		Item A		Item A	
		Yes	No	Yes	No
Item B	Yes	$abcS_3$	bcS_2	abS_2	bS_1
	No	acS_2	cS_1	aS_1	S_0

and

$$\{y_{111}, y_{110} + y_{101} + y_{011}, y_{100} + y_{010} + y_{001}, y_{000}\}.$$

Note that these are the minimal sufficient statistics of the model of *quasi-symmetry* preserving one-dimensional marginal totals.

A 2^k contingency table is *symmetric* if the expected counts under all possible permutations of subscripts are equal. Symmetry implies that all of the r-way sets of marginal totals are equal to one another for $r = 1, 2, \ldots, k-1$. Quasi-symmetry generalizes this notion by allowing sets of lower-order marginal totals to differ (see [8, Chapter 8]). The quasi-symmetry model preserving one-dimensional marginal totals for a 2^k table is equivalent to that of expression (6). Additional simplifications ensue here because

$$\hat{m}_{111} = y_{111}, \qquad \hat{m}_{000} = y_{000}.$$

For more details on this relationship between quasi-symmetry and the Rasch model, see [18, 20]. In particular, they point out the relevance of the moment constraints described by Cressie and Holland [10], that are not encompassed in the quasi-symmetry structure (see also [9]). Agresti [1–4], Darroch et al. [13], and Kelderman [29] all explore further aspects of these loglinear representations for the Rasch model, as we do in a limited fashion in the example below. A parallel literature in the 1980s linked models for individual heterogeneity, similar to the Rasch model, with models of symmetry and quasi-symmetry (for example [11]). Darroch & McCloud [12] give an especially interesting application of this approach to an example involving the separation of sources of dependence in a series of four influenza outbreaks. Their model is similar to a generalized Rasch model but uses a **fixed-effects** version for heterogeneity rather than the random-effects version described above. We revisit this example in the next section.

Example: Influenza Outbreaks in Michigan

Table 2 contains infection frequencies to four influenza outbreaks for a sample of 263 individuals in Tecumseh, MI during the winters of 1977/78 to 1980/81. These data were first reported by Haber [25] and later analyzed in depth by Darroch & McCloud [12]. One interesting aspect of these data is that outbreaks 1 and 4 were caused by the same type of virus, and thus responses (conditional on individuals) to these two outbreaks are potentially dependent. Since we have already collapsed over individuals, the data form a 2^k contingency table, where $k = 4$ for the four influenza outbreaks and corresponds to the number of items for a standard Rasch model.

Table 2 Infection profiles and frequency of infection for the influenza example

j_1, j_2, j_3, j_4	Frequency	j_1, j_2, j_3, j_4	Frequency
0 0 0 0	140	1 0 0 0	20
0 0 0 1	31	1 0 0 1	2
0 0 1 0	16	1 0 1 0	9
0 0 1 1	3	1 0 1 1	0
0 1 0 0	17	1 1 0 0	12
0 1 0 1	2	1 1 0 1	1
0 1 1 0	5	1 1 1 0	4
0 1 1 1	1	1 1 1 1	0

To understand the different sources of heterogeneity in this example, consider an individual who, by virtue of her personal characteristics, is highly susceptible to influenza. She is likely to succumb to outbreaks 1, 2, and 3, as expected, but is unlikely to get sick during outbreak 4. Since outbreaks 1 and 4 are caused by the same type of organism, individuals who get sick during outbreak 1 develop some degree of immunity to the virus, thus inducing a negative dependence between outbreaks 1 and 4.

We can fit RM such as (1) to these data in a straightforward manner if we assume independence among individuals and among influenza outbreaks. This assumption might be reasonable if all influenza outbreaks were caused by different types of virus. The loglinear model arising from the conditional representation (4) of the RM (in the u-term notation of Bishop et al. [8], and used in the article on **Loglinear Models**) is

$$\log m_{j_1 j_2 j_3 j_4} = u + \sum_{j=1}^{4} u_j(x_j) + u_5(x_{i+}), \qquad (7)$$

where the $u_5(x_{i+})$ and the $u_j(x_j)$ factors correspond to individual scores and influenza parameters, respectively, and is simply the u-term representation for the loglinear model of expression (6).

The model in (7) corresponds to a standard RM, and therefore does not accommodate the negative correlation between outbreaks 1 and 4. Thus, we do not expect it to fit these data well. More appealing models for these data would include one or more interaction terms to represent the dependence between outbreaks 1 and 4. One such model, suggested by the Darroch & McCloud analysis, includes an interaction between outbreaks 1 and 4 and is given by

$$\log m_{j_1 j_2 j_3 j_4} = u + \sum_{j=1}^{4} u_j(x_j) + u_{14}(x_1 x_4) + u_5(x_{i+}), \tag{8}$$

where the $u_{14}(x_1 x_4)$ are associated with the four possible infection profiles when considering only outbreaks 1 and 4.

Model (8) may still be underparameterized, insofar as it does not include an **interaction** term for individual scores × outbreaks parameters. From an intuitive viewpoint, the interaction between scores x_{i+} and the first outbreak x_1 should be included, since the value of the score might well depend on the individual's response to x_1 and, conditional on x_1, also on the response to x_4. Another way to think about this is the following. Consider the ways an individual can obtain a score $x_{i+} = 2$. If he does not get sick during outbreak 1, then he can obtain the score of 2 in three different ways, each with the same probability of occurrence given independence among outbreaks (2, 3, 4). If he gets sick during outbreak 1, however, then he needs one more infection to round up the score, but now each of the three possible infection patterns does not have equal probability, in light of the dependence between outbreaks 1 and 4. To capture such an effect, we consider a third model closely linked to the analysis of Darroch & McCloud with an additional interaction term:

$$\log m_{j_1 j_2 j_3 j_4} = u + \sum_{j=1}^{4} u_j(x_j) + u_{14}(x_1 x_4) + u_5(x_{i+}) + u_{15}(x_1, x_{i+}). \tag{9}$$

We fitted models (7), (8), and (9) to the influenza infection data in Table 2, and compared the fit of the models using the usual likelihood-ratio chi-square statistic, G^2 (*see* **Likelihood Ratio Tests**), as reported in Table 3.

Table 3 Degrees of freedom and deviance statistics for the three models fitted to the influenza data

Model	Degrees of freedom	Deviance G^2
Rasch model (no interactions)	8	25.79
RM + outbreaks 1 and 4 interaction	7	16.11
RM + outbreak (1, 4) + score by outbreak 1 interactions	5	5.52

From Table 3 we see that, as expected, the standard RM did not fit the data well. The fit of the model improved considerably when we added the first interaction term, between the first and fourth outbreaks. While this model accounts for the fact that both outbreaks 1 and 4 are caused by the same type of virus, it ignores the effect that the response to the first outbreak has on individual scores. The last model we fitted includes this effect, and its fit is superior to that of the other two models.

Note that, while model (7) corresponds to the standard RM for this problem, neither loglinear models (8) or (9) have an RM representation. In fact, the latter two models are not RM at all, due to the presence of the interaction terms. Nonetheless it is interesting to see that reasonable loglinear models can be obtained by starting with the RM as a basis, and then adding terms as needed to model potential dependences not accommodated by the simple RM.

Finally, Table 4 gives the estimated cell counts for the 2^4 contingency table under the three fitted models. Given its better fit, model (9) produces values of cell counts that are in good agreement with observed counts. We can also compare the estimated values with observed frequencies in Table 4. Estimated infection frequencies computed from the RM (7) over- and underestimate observed cell counts in the expected direction. For example, the RM model tends to overestimate infection frequencies in cells $\{0, j_2, j_3, 0\}$ and $\{1, j_2, j_3, 1\}$, which is consistent with the assumption of independence among outbreaks.

Summary

The Rasch model comes to general statistical practice, and biostatistics in particular, from psychological

Table 4 Infection profiles, observed frequencies, and estimated frequencies under three loglinear models of infections during the four influenza outbreaks

j_1, j_2, j_3, j_4	Obs. freq.	$\hat{\mathbf{m}}$, model (7)	$\hat{\mathbf{m}}$, model (8)	$\hat{\mathbf{m}}$, model (9)
0 0 0 0	140	140.0	140.0	140.0
0 0 0 1	31	19.7	22.8	28.4
0 0 1 0	16	18.5	15.7	16.7
0 0 1 1	3	4.6	5.7	3.6
0 1 0 0	17	20.9	17.9	18.9
0 1 0 1	2	5.2	6.4	4.0
0 1 1 0	5	4.9	4.4	2.4
0 1 1 1	1	1.3	2.1	1.0
1 0 0 0	20	24.8	27.7	20.0
1 0 0 1	2	6.1	1.8	1.8
1 0 1 0	9	5.8	6.9	10.0
1 0 1 1	0	1.5	0.6	0.6
1 1 0 0	12	6.5	7.8	11.3
1 1 0 1	1	1.7	0.7	0.7
1 1 1 0	4	1.6	2.6	3.8
1 1 1 1	0	0.0	0.0	0.0

testing, and it generalizes the model of independence among a set of response variables by allowing for the incorporation of individual effects. Thus we speak of local independence, conditional on the individual. In this article we have shown an important connection between the Rasch model and a special loglinear model for the usual contingency table representation of the relevant response variables, that of quasi-symmetry. As a consequence we can think of quasi-symmetry as providing a representation for dependence introduced by Rasch-model-like heterogeneity. Through an example dealing with influenza epidemics, we have illustrated here how loglinear generalizations of this Rasch model representation are useful in biostatistical contexts involving such forms of heterogeneity.

The most extensive area of application of this type of approach to heterogeneity to date has come in the area of **capture–recapture** modeling – for example, see [9, 13, 15, 20] and [26]. As we noted at the outset, there are alternative approaches to heterogeneity such as stratification and the use of regression-like components utilizing additional explanatory variables that "account for" heterogeneity. The principal example in [26], dealing with the ascertainment of diabetes in a region of Italy, contrasts the results of generalized-Rasch-like loglinear models and separate loglinear models for separate strata. Fienberg, Johnson and Junker [20] revisit this example and provide an alternative Bayesian analysis based on the Rasch model and some generalizations of it.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters, *Scandinavian Journal of Statistics* **20**, 63–72.

[3] Agresti, A. (1993). Distribution-free fitting of logit models with random effects for repeated categorical responses, *Statistics in Medicine* **12**, 1969–1987.

[4] Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort, *Biometrics* **50**, 494–500.

[5] Andersen, E.B. (1980). *Discrete Statistical Models with Social Sciences Applications*. North-Holland, Amsterdam.

[6] Andersen, E.B. (1990). *The Statistical Analysis of Categorical Data*. Springer-Verlag, Heidelberg.

[7] Baker, F.B. (1992). *Item Response Theory. Parameter Estimation Techniques*. Marcel Dekker, New York.

[8] Bishop, Y.M.M., Fienberg, S.E. & Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[9] Coull, B.A. & Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies, *Biometrics* **55**, 294–301.

[10] Cressie, N.E. & Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models, *Psychometrics* **48**, 129–141.

[11] Darroch, J.N. (1986). Quasi-symmetry, In *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 469–473.

[12] Darroch, J.N. & McCloud, P.I. (1990). Separating two sources of dependence in repeated influenza outbreaks, *Biometrika* **77**, 237–243.

[13] Darroch, J.N., Fienberg, S.E., Glonek, G. & Junker, B. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability, *Journal of the American Statistical Association* **88**, 1137–1148.

[14] Diggle, P.J., Liang, K.Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[15] Dobra, A. & Fienberg, S.E. (2001). How big is the world wide web? in *Computing Science and Statistics, Volume 33 – Proceedings of Interface 2001*, California.

[16] Duncan, O.D. (1983). Rasch measurement: Further examples and discussion, in: *Survey Measurement of Subjective Phenomena*, Vol. 2. C.F. Turner & E. Martin, eds. Russell Sage, New York, pp. 367–403.

[17] Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change, *Psychometrika* **56**, 495–515.

[18] Erosheva, E.A., Fienberg, S.E. & Junker, B.W. (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables, *Annales de la Facult(e) des Sciences de l'Universite de Toulouse Math(e)matiques* **11**, in press.

[19] Fienberg, S.E. & Meyer, M. (1983). Loglinear models and categorical data analysis with psychometric and econometric applications, *Journal of Econometrics* **22**, 191–214.

[20] Fienberg, S.E., Johnson, M. & Junker, B. (1999). Classical multi-level and Bayesian approaches to population size estimation using data from multiple lists, *Journal of the Royal Statistical Society, Series A* **162**, 383–406.

[21] Fischer, G.H. (1976). Some probabilistic models for measuring change, in *Advances in Psychological and Educational Methods*, D.N.M. De Gruijter & L.J.Th. Van der Kamp, eds. Wiley, New York, pp. 97–110.

[22] Fischer, G.H. (1977). Linear logistic trait models: theory and applications, in *Structural Models of Thinking and Learning*, H. Spada & W.H. Kempf, eds. Huber, Berne, pp. 203–225.

[23] Fischer, G.H. (1989). An IRT-based model for dichotomous longitudinal data, *Psychometrika* **54**, 599–624.

[24] Gustafsson, G.E. (1980). Testing and obtaining the fit of data to the Rasch model, *British Journal of Mathematical and Statistical Psychology* **33**, 205–233.

[25] Haber, M. (1986). Testing for pairwise independence, *Biometrics* **42**, 429–435.

[26] International Working Group for Disease Monitoring and Forecasting (1995). Capture–recapture and multiple-record systems estimation. I: History and theoretical development, *American Journal of Epidemiology* **142**, 1047–1058.

[27] Johnson, V.E. & Albert, J.H. (1999). *Ordinal Data Modeling*. Springer-Verlag, New York.

[28] Jones, R.H. (1993). *Longitudinal Data With Serial Correlation: A State-Space Approach*. Chapman & Hall, London.

[29] Kelderman, H. (1984). Loglinear Rasch model tests, *Psychometrika* **49**, 223–245.

[30] Kelderman, H. & Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items, *Psychometrika* **59**, 149–170.

[31] Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change, *Psychometrika* **61**, 629–645.

[32] Molenaar, I.W. (1995). Estimation of item parameters, in *Rasch Models: Foundations, Recent Developments, and Applications*, G.H. Fischer & I.W. Molenaar, eds. Springer-Verlag, New York, pp. 39–52.

[33] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The Danish Institute of Educational Research. (Expanded edition, The University of Chicago Press, 1980).

[34] Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model, *Scandinavian Journal of Statistics* **9**, 23–30.

(*See also* **Random Coefficient Repeated Measures Model**)

ALICIA L. CARRIQUIRY &
STEPHEN E. FIENBERG

Rasch, Georg

Born: September 21, 1901, in Odense, Denmark.
Died: October 19, 1980, in Byrum, Laesø, Denmark.

Georg Rasch was Professor of Statistics from 1962 to 1972 at the University of Copenhagen. He received his degree in mathematics from the University of Copenhagen in 1925 and worked as a mathematician at the university until 1930, when at age 29, he became a doctor of science on a dissertation concerned with matrix calculations and its applications in differential and difference equation theory [1]. At the time, he was considered to be one of the most talented of the new generation of Danish mathematicians. But as no satisfactory position was created for him as a mathematician, he chose to work as a consultant in applied mathematics, primarily data analysis and statistics. In the 1930s he worked with problems in medicine and biology, but he later added education, psychology and sociology as fields of interest (*see* **Social Sciences**).

Between 1935 and 1936, he visited University College in London, primarily to work with **R.A. Fisher**. He was much impressed by Fisher's ideas on the foundations of mathematical statistics and introduced them in Denmark after his return. In the following years, he worked primarily at the State Serum Institute, where he founded the Biostatistics Department and was its director from 1940 to 1956. In this capacity, he made many contributions to new developments in biology and medicine, primarily as a consultant for doctoral dissertations by the scientists at the Institute and many other medical doctors. He had, however, a much more lasting influence on the development of statistics, in both theory and applications, through the fact that most, if not all, of the next generation of Danish statisticians worked as his assistants at the Serum Institute. For example, Professor A. Hald started his career as an assistant to Rasch.

In the 1940s and 1950s he had various part-time teaching assignments at the university, but it was not until 1961, when he was almost 60 years old, that he was appointed to a chair in statistics at the University of Copenhagen. It may seem surprising, but it is nevertheless a fact, that he did not work with applications in education and psychology until the mid-1950s, when he was into his own fifties. These disciplines occupied most of his thinking in the 1960s and 1970s, and it was here that he made his most original contributions. As a consultant to the Ministry of Social Affairs, to the Office of Military Psychology, and to the Danish Educational Research Institute, he was faced with the task of extracting information on individuals from intelligence and ability tests. He rejected the traditional statistical methods, primarily based on various factor analytic techniques (*see* **Factor Analysis, Overview**), and developed new and more exact methods based on latent trait models as we know them today. The most simple and elegant of these models was fully developed by Rasch in 1960 and now bears his name: the **Rasch model**. The model was not invented as a new theoretical development, but was established through careful study of the empirical data with which he worked. He also realized that the model required a new statistical methodology based on the use of **conditional probabilities**. In 1960, in his famous book [2] and in an important paper read at the Berkeley Symposium on Probability and Statistics [3], he presented both a new revolutionary model and an associated fascinating new statistical methodology. The model was developed further in the following years and he made many important applications of it, but to a remarkable degree the theory was developed within a span of three to four years. In the 1960s and 1970s there followed a few papers in which he tried to extend his discoveries from 1960 to a more general theory of measurement primarily directed toward the social sciences. It was these ideas that occupied his thinking for the rest of his life. In his scientific works, Rasch combined mathematical skill and a skill for reading empirical evidence in a unique way. He used mathematics to make ideas precise and to formulate the theoretical framework for the analysis of data in an exact way. But data from real life were the main source for all his theoretical developments and model formulations. Rasch was thus an early and eager advocate of checking the fit of a model by statistical and/or graphical methods (*see* **Model Checking**). Georg Rasch was a knight of the Danish order of Dannebrog and an honorary member of the Danish Statistical Society.

References

[1] Rasch, G. (1930). *Om Matrixregningog dens Anvendelse på Differensog Differential-ligninger*. Levin og Munksgård, Copenhagen.

[2] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Pædagogiske Institut, Copenhagen.

[3] Rasch, G. (1961). *Proceedings of the Fourth Berkeley Symposium on the Mathematics of Statistics and Probability*, Vol. 5. University of California Press, Berkeley, pp. 321–333.

ERLING B. ANDERSEN

Rate

Rate refers to a limiting ratio of the changes in two quantities as these changes tend to zero. The denominator often involves time, t, as for the **hazard rate**, $\lambda(t) = \lim_{\Delta\downarrow} \Delta^{-1} \Pr$ (disease first occurs in $[t, t+\Delta)$|disease first occurs at or after t). The notation $[t, t+\Delta)$ means that the event occurs at or after t but before $t+\Delta$.

Sometimes rate refers to a proportion, as in neonatal mortality rate and **prevalence rate**.

MITCHELL H. GAIL

Ratio and Regression Estimates

Ratio and regression estimators are used to improve the precision of estimates of a population total or **mean**, by exploiting the relationship between an outcome y, and an auxiliary measurement x. Estimates of ratios themselves are also often made – these include **rates** and proportions for the population and for targeted subgroups. Such estimates are widely used in biostatistical research; indeed most public-use data tapes are issued with **poststratified** ratio adjustments (based on sex, race, and age) to the sampling weights. Domain estimates can be viewed as ratio estimates as well, so that epidemiologic **prevalence** estimates for population subgroups are essentially ratio estimates. In the following, estimators, along with **variance** estimates for population totals, are discussed; for estimators of population means, divide by N, and adjust the variance estimate by N^{-2}.

Ratio Estimators

The ratio estimator is used when two numbers are associated with each of N units in a finite population. One of these is a positive known quantity x, and the other is an unknown y, the outcome measure of interest. Letting s be the set of n units in a sample from the population, the ratio estimate for the population total, $T = \sum_{i=1}^{N} y_i$, is

$$\hat{T}_R = \frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i} \sum_{i=1}^{N} x_i = \hat{R}X, \qquad (1)$$

where $\sum_{i\in s}$ denotes the sum over the sample units, $\hat{R} = \sum_{i\in s} y_i \big/ \sum_{i\in s} x_i$ is the sample ratio, and X is the population x-total. The sample ratio can also be written

$$\hat{R} = \frac{\overline{y}_s}{\overline{x}_s} = \frac{N\overline{y}_s}{N\overline{x}_s} = \frac{\hat{Y}}{\hat{X}},$$

where $\overline{y}_s = n^{-1}\sum_{i\in s} y_i$ and $\overline{x}_s = n^{-1}\sum_{i\in s} x_i$ are the sample means, $\hat{Y} = N\overline{y}_s$ and $\hat{X} = N\overline{x}_s$ are the simple expansion estimators of the population y-total, T, and the population x-total, X, respectively. This makes $\hat{R}$ a natural estimate of the population ratio $R = \sum_{i=1}^{N} y_i \big/ \sum_{i=1}^{N} x_i$.

A domain is defined by a sample characteristic which is generally not available until the survey has been executed. Suppose we are interested in the average number of physician visits in the past year for Latinas aged 18 to 64. Letting $z_i = 1$ if the respondent is in this category and $z_i = 0$ otherwise, and measuring the number of physician visits with y, the population characteristic we want to estimate is a population ratio $R = \sum_{i=1}^{N} y_i z_i \big/ \sum_{i=1}^{N} z_i = \sum_{i=1}^{N} y_i' \big/ \sum_{i=1}^{N} z_i$, which we estimate with $\hat{R} = \Sigma_{i\in s} y_i z_i \big/ \Sigma_{i\in s} z_i$. Thus estimates for population subgroups can be made from survey data by appropriately defining z_i and calculating a ratio estimate.

An equivalent form for the ratio estimator (1) of the y-total is $\hat{T}_R = N\overline{y}_s(\overline{X}/\overline{x}_s)$, where $\overline{X} = N^{-1}\sum_{i=1}^{N} x_i = X/N$ is the known population average of the xs. Writing $\hat{T}_R$ this way illustrates the intuitive appeal of the ratio estimator as a multiplicative adjustment to the simple expansion estimator $\hat{Y} = N\hat{y}_s$ by the factor $\overline{X}/\overline{x}_s$. This adjustment is upward if $\overline{x}_s < \overline{X}$ and downward if $\overline{x}_s > \overline{X}$; intuition supports such an adjustment for if the sample mean of the xs is smaller than the population mean, one might suspect that the sample mean of the ys would be low as an estimate of the population average, justifying an upward adjustment of the simple expansion estimate, $N\overline{y}_s$.

An example where the ratio estimator would be valuable is in estimation of the total number of hospital discharges over a given period, based on a sample of n hospitals from N. Here the auxiliary information x, known for all hospitals in the population, is bed size. If the hospitals selected for the sample are smaller (as measured by bed size) on the average than those in the population, an upwards adjustment of the simple expansion estimate of the number of discharges will be in order.

Of course intuition alone is not enough to justify use of the ratio estimator. The properties of this estimator have been extensively studied from two distinct points of view. One treats the ys as unknown constants and develops properties of the estimator with respect to the **random sampling** plan used to select the units for s. Cochran [1] gives a comprehensive treatment of this **probability sampling** theory approach to **estimation**. The other theory which treats $y_1, \ldots, y_N$ as realizations of **random**

variables $Y_1, \ldots, Y_N$, is developed using a statistical model which relates Y_i and x_i. There has been considerable contention over the role each of these theories should play in **inference** for finite populations. A number of **simulation** studies have been done where both theories apply, and situations where model-based inference has important messages are noted in the following.

Cochran [1] delineated the properties of the ratio estimator assuming a **simple random sample** (without replacement; *see* **Sampling With and Without Replacement**) of size n from the N units in the population. The properties so developed are often used as well when the sampling method is **systematic** rather than simple random, even though there is considerable evidence that appropriately chosen systematic samples can be superior to random samples for inference. Averaging over all possible samples, the ratio estimate has a **bias** of order $1/n$ and thus is negligible for large n. This bias is generally ignored even for moderate samples sizes. An empirical study by Kish et al. [4] showed that the bias relative to the root **mean square error** is small unless n is very small.

Beginning with the simple ratio, $\hat{R}$, a common formula for the estimated variance is

$$v(\hat{R}) = \hat{R}^2\{\mathrm{cv}^2(\overline{y}_s) + \mathrm{cv}^2(\overline{x}_s) - 2r\mathrm{cv}(\overline{x}_s)\mathrm{cv}(\overline{y}_s)\} \tag{2}$$

where $\mathrm{cv}^2(\overline{y}_s) = \widehat{\mathrm{var}}(\overline{y}_s)/\overline{y}_s^2 = \{(1-f)/n(n-1)\}$ $\sum_{i\in s}(y_i - \overline{y}_s)^2/\overline{y}_s^2$ so $\mathrm{cv}(\overline{y}_s)$ is the estimated coefficient of variation (*see* **Standard Deviation**) of $\overline{y}_s$, $\mathrm{cv}(\overline{x}_s)$ of $\overline{x}_s$, and r is the estimate of the **correlation** between $\overline{y}_s$ and $\overline{x}_s$, which for simple random sampling is just the ordinary sample correlation calculated from the data, $r = [\sum_{i\in s}(x_i - \overline{x}_s)y_i]/[\sum_{i\in s}(x_i - \overline{x}_s)^2 \sum_{i\in s}(y_i - \overline{y}_s)^2]^{1/2}$; f is the sampling fraction n/N. This formula follows from a simple Taylor series for $\hat{R}$, often referred to in survey research as the **linearization method**. Unfortunately, linearization has not always produced useful variance estimates in survey research. For example, this method, when applied to $\hat{T}_R$, leads to a variance estimator for $\hat{T}_R$ equal to $v_0 = N^2\{(1-f)/n(n-1)\}\sum_{i\in s}(y_i - \hat{R}x_i)^2$, which has been shown to be a poor estimator of the variability in $\hat{T}_R$.

Wu [22] and Wu & Deng [24] studied a general class of variance estimators for use with $\hat{T}_R$ of the form

$$v_g(\hat{T}_R) = \left(\frac{\overline{X}}{\overline{x}_s}\right)^g v_0.$$

The two special cases of this estimator corresponding to $g = 0$ and $g = 2$ are the variance estimators recommended in many sampling texts for use with the ratio estimate of a population total. Cochran [1] listed both, deriving v_0 via a linearization argument, and v_2 from the relation $\hat{T}_R = N\overline{X}\hat{R}$ which implies $v_2 = (N\overline{X})^2 v(\hat{R})$, where $v(\hat{R})$ is found as in (2).

When applying the ratio method of estimation, the data should show a straight-line **regression** through the origin with variance increasing with x. Its use in other situations can lead to large inefficiencies and/or large biases in estimation. Royall & Cumberland [12] studied the ratio estimator and estimators of its variance under such a model: $Y_i = \beta x_i + \varepsilon_i\sqrt{x_i}$, where $\mathrm{E}(\varepsilon_i) = 0, \mathrm{E}(\varepsilon_i^2) = \sigma^2$, and $\mathrm{E}(\varepsilon_i\varepsilon_j) = 0$ for $i \neq j$. This model assumes variance proportional to x; Royall & Cumberland [11], following Royall & Eberhardt [15], developed a class of variance estimators whose performance is **robust** against failure of this assumption and which have good properties as an estimator of the mean square error of $\hat{T}_R$ when viewed conditionally on sample characteristics. This conditional analysis allows us to see properties of the estimators which are concealed in theoretical developments which average over all possible samples. In an empirical study of the ratio estimator with different populations all of which appear to conform well to this model, Royall & Cumberland [12] concluded that v_0 should not be used to estimate the variance of the ratio estimator, that the robust variance estimators they studied did a good job of tracking the mean square error of $\hat{T}_R$, and that the ratio estimator itself can show a large conditional bias due to failure of the specification $\mathrm{E}(Y_i) = \beta x_i$ in badly balanced samples – those where $\overline{x}_s$ and $\overline{X}$ differ substantially. The use of **stratification** on x, and/or systemic sampling after ordering the population on x, can help avoid the selection of badly balanced samples, thus providing some protection against bias in $\hat{T}_R$. One of the robust variance estimators Royall & Cumberland [12] studied was very nearly v_2, hence on these grounds this estimator can be recommended. Wu and Deng [24], in an unconditional analysis of v_g, suggest using the data to choose g optimally, but in an empirical study drew the conclusion that v_2 was a good

performer in populations where variance increases with x. More recently published sampling texts [8, 18] generally recommend v_2 for use with the ratio estimator, but this is not universal among sampling texts. A simple computing formula for this variance estimator is

$$v_2(\hat{T}_R) = (N\overline{X})^2\hat{R}^2\{\mathrm{cv}^2(\overline{y}_s) + \mathrm{cv}^2(\overline{x}_s) - 2r\mathrm{cv}(\overline{x}_s)\mathrm{cv}(\overline{y}_s)\}. \quad (3)$$

The **jackknife** variance estimator for $\hat{T}_R$ has also been studied extensively [6, 7, 10–12, 22–24]. The jackknife variance estimate [3] is

$$v_J(\hat{T}_R) = (N\overline{X})^2\frac{1-f}{n}(n-1)\sum_{j\in s}(\hat{R}_{(j)} - \hat{R}_{(\cdot)})^2$$

where for every $j \in s$, $\hat{R}_{(j)} = [n\overline{y}_s - y_j]/[n\overline{x}_s - x_j]$ is the ratio estimate found after deleting unit j from the sample and $\hat{R}_{(\cdot)}$ is the average of these n estimates. Royall & Cumberland [12] showed v_J is asymptotically equivalent to v_2 and in their empirical study noted that it generally was larger than other robust variance estimators, leading to more conservative confidence intervals. Wu & Deng [24] drew similar conclusions about v_J and cautiously recommended its use with $\hat{T}_R$.

Confidence intervals for a population total $\hat{T}_R \pm 1.96\sqrt{v}$ rely on **large-sample normality** of the estimates. Conditions for asymptotic normality of the ratio and regression estimators were given in Scott & Wu [19]. Such asymptotic results should be used with caution in moderate samples; Royall & Cumberland [14] and Valliant [20] note from empirical studies poor coverage properties of confidence intervals for some populations, even though the variance estimates were appropriately tracking the mean square error.

Stratified Ratio Estimates

Sampling from strata is much more common than simple random sampling from a population. Stratification improves the efficiency of the estimators and helps to avoid conditions of extreme imbalance which can give rise to large conditional biases in the ratio estimator. Royall & Herson [16] discussed the use of stratification on x with the ratio estimator and showed the value of stratification in protecting inferences from bias due to imbalance.

Suppose the N population units are divided into H strata of sizes $N_1, \ldots, N_H$ so that $\sum_{h=1}^{H} N_h = N$. **Stratified random sampling** consists of independently choosing simple random samples of size n_h from each stratum with $\sum_{h=1}^{H} n_h = n$ the total sample size. Denote the measurements on the N population units by y_{hi} and x_{hi} for $i = 1, \ldots, N_h$ and $h = 1, \ldots, H$, and let the set of sampled units from stratum h be s_h. Two ratio estimators of the population total are commonly used with this plan, the separate ratio $\hat{T}_{RS}$, and the combined ratio $\hat{T}_{RC}$. These are

$$\hat{T}_{RS} = \sum_{h=1}^{H} N_h\overline{X}_h\left(\frac{\overline{y}_{s_h}}{\overline{x}_{s_h}}\right),$$

$$\hat{T}_{RC} = \sum_{h=1}^{H} N_h\overline{X}_h(\hat{R}_C),$$

where $\hat{R}_C = \sum_{h=1}^{H} N_h\overline{y}_{s_h} \Big/ \sum_{h=1}^{H} N_h\overline{x}_{s_h}$ is the ratio of the stratified expansion estimators for the y-total and the x-total, $\overline{y}_{s_h} = n_h^{-1}\sum_{i\in s_h} y_{hi}$ is the sample mean of the ys in stratum h, $\overline{x}_{s_h} = n_h^{-1}\sum_{i\in s_h} x_{hi}$ that of the xs, and $\overline{X}_h = N_h^{-1}\sum_{i=1}^{N_h} x_{hi}$ is the known stratum mean of the xs in stratum h.

The separate ratio estimate uses a different ratio estimate for each stratum and, if the n_h are not too small, will be a better estimator than the combined ratio estimate. The estimator $\hat{R}_C$ is used by itself with stratified designs to estimate rates and prevalences for subgroups of the population, such as disease rates among well-defined racial subgroups.

The variance estimator for $\hat{T}_{RS}$ comes directly from the single-sample case, since $\hat{T}_{RS}$ is the sum of H independent ratio estimates, $\hat{T}_{Rh} = N_h\overline{X}_h(\overline{y}_{s_h}/\overline{x}_{s_h})$. Using the robust variance estimates for $\hat{T}_{Rh}$, we have

$$v_2(\hat{T}_{RS}) = \sum_{h=1}^{H} v_2(\hat{T}_{Rh}) = \sum_{h=1}^{H}(N_h\overline{X}_h)^2\left(\frac{\overline{y}_{s_h}}{\overline{x}_{s_h}}\right)^2\{\mathrm{cv}^2(\overline{y}_{s_h}) + \mathrm{cv}^2(\overline{x}_{s_h}) - 2r_h\mathrm{cv}(\overline{x}_{s_h})\mathrm{cv}(\overline{y}_{s_h})\}$$

and each $v_2(\hat{T}_{RS})$ is calculated using only the data from stratum h, exactly as in the single-sample case[(3)]. A robust variance estimate for the combined ratio estimate can be written in a

parallel fashion. First, for the ratio $\hat{R}_{\mathrm{C}}$ itself we have

$$v(\hat{R}_{\mathrm{C}}) = \hat{R}_{\mathrm{C}}^2\{\mathrm{cv}^2(\hat{Y}) + \mathrm{cv}^2(\hat{X}) - 2r\mathrm{cv}(\hat{X})\mathrm{cv}(\hat{Y})\},$$

where $\hat{Y} = \sum_{h=1}^{H} N_h \overline{y}_{s_h}$ is the simple stratified expansion estimate of the y-total, $\mathrm{cv}^2(\hat{Y}) = \widehat{\mathrm{var}}(\hat{Y})/\hat{Y}^2$, and $\widehat{\mathrm{var}}(\hat{Y}) = \sum_{h=1}^{H} N_h^2\{(1-f_h)/n_h(n_h-1)\}\Sigma_{i\in s_h}$ $(y_{hi} - \overline{y}_{s_h})^2$ (with analogous formulas for $\mathrm{cv}^2(\hat{X})$), $f_h = n_h/N_h$, and $r = \widehat{\mathrm{cov}}(\hat{X},\hat{Y})/[\widehat{\mathrm{var}}(\hat{X})\widehat{\mathrm{var}}(\hat{Y})]^{1/2}$, where $\widehat{\mathrm{cov}}(\hat{X},\hat{Y}) = \sum_{h=1}^{H} N_h^2\{(1-f_h)/n_h(n_h-1)\}$ $\sum_{i\in s_h}(y_{hi} - \overline{y}_{s_h})(x_{hi} - \overline{x}_{s_h})$. For $\hat{T}_{RC}$ a variance estimate is

$$v_2(\hat{T}_{RC}) = X^2\hat{R}_{\mathrm{C}}^2\{\mathrm{cv}^2(\hat{Y}) + \mathrm{cv}^2(\hat{X}) - 2r\mathrm{cv}(\hat{X})\mathrm{cv}(\hat{Y})\},$$

where $X = \sum_{h=1}^{H}\sum_{i=1}^{N_h} x_{hi} = \sum_{h=1}^{H}(N_h\overline{X}_h)$ is the population total of the xs. The variance estimators $v_2(\hat{T}_{RS})$ and $v_2(\hat{T}_{RC})$ (or close variants of them) have been studied empirically by Valliant [20], Wu [23], and Deng & Wu [2], and they generally perform quite well. Both Valliant [20] & Wu [23] also considered the stratified jackknife variance estimator defined generally [3] as

$$v_{\mathrm{J}} = \sum_{h=1}^{H} \frac{1-f_h}{n_h}(n_h-1)\sum_{j\in s_h}(\hat{T}_{(hj)} - \hat{T}_{(h)})^2,$$

where $\hat{T}_{(hj)}$ is the estimate calculated without the hjth unit and $\hat{T}_{(h)} = \sum_{j\in s_h}\hat{T}_{(hj)}/n_h$. Rao & Wu [9], Krewski & Rao [6], and Lemeshow & Levy [7] considered other versions of the jackknife variance estimator. Wu [23] argued that $v_2(\hat{T}_{RC})$ and v_{J} applied to $\hat{T}_{RC}$ should have similar performance. In a simulation study Valliant [20] compared the performances of several variance estimators for $\hat{T}_{RC}$ and $\hat{T}_{RS}$; among these were stratified versions of robust variance estimators, one with similar properties to $v_2(\hat{T}_{RC})$, and another the jackknife variance estimator. With respect to conditional coverage properties, their performance was better than the other estimators considered. Valliant further pointed out in his study that stratification on x alone was not sufficient to protect oneself from a conditional bias due to imbalance on x, although it does guard against gross imbalances that can occur with simple random sampling. He suggested a combination of stratification on x with systematic sampling within strata ordered on x, and the use of a robust variance estimator (like v_{J} or v_2) for reliable inference.

Regression Estimators

From a simple random sample s from a population, the **simple linear regression** estimator for a population total can be calculated as

$$\hat{T}_{\mathrm{L}} = N\overline{y}_s + b(N\overline{X} - N\overline{x}_s)$$

where $b = \sum_{i\in s}(x_i - \overline{x}_s)(y_i - \overline{y}_s)\big/\sum_{i\in s}(x_i - \overline{x}_s)^2$, which is the usual estimate of a slope in a simple linear regression. Like the ratio estimator, the regression estimator has considerable intuitive appeal as an adjustment to the simple expansion estimator, $N\overline{y}_s$. If a plot of the data indicates a straight-line regression of y on x with constant variance, then a linear regression estimator is appropriate and will be superior to a ratio estimator when the intercept is not the origin. Under simple random sampling, the regression estimator is biased, but, as in the case of ratio estimation, this is usually ignored when n is not small. A variance estimator found in most sampling textbooks [1] is $v_{\mathrm{C}} = N^2\{(1-f)/n(n-2)\}\sum_{i\in s} d_i^2$, where $d_i = y_i - \overline{y}_s - b(x_i - \overline{x}_s)$ are the **residuals**. In an empirical study of the regression estimator, Royall & Cumberland [13] showed that v_{C} was seriously flawed as an estimator of the variance of $\hat{T}_{\mathrm{L}}$, and suggested several superior variance estimates. One choice was

$$v_{\mathrm{D}}(\hat{T}_{\mathrm{L}}) = N^2\sum_{i\in s}\frac{a_i d_i^2}{(1-p_i)}, \tag{4}$$

where

$$a_i = \left(\frac{1-f}{n} + \frac{(x_i - \overline{x}_s)(\overline{X} - \overline{x}_s)}{\sum_{j\in s}(x_j - \overline{x}_s)^2}\right)^2 + \frac{1-f}{nN}$$

and $p_i = 1/n + (x_i - \overline{x}_s)^2/\sum_{j\in s}(x_j - \overline{x}_s)^2$ is the ith diagonal element of the "hat" matrix in a standard linear regression. Another choice was the jackknife variance estimator

$$v_{\mathrm{J}}(\hat{T}_{\mathrm{L}}) = \frac{1-f}{n}(n-1)\sum_{j\in s}(\hat{T}_{\mathrm{L}(j)} - \hat{T}_{\mathrm{L}(\cdot)})^2,$$

where $\hat{T}_{\mathrm{L}(j)}$ is the regression estimator based on the sample obtained by deleting unit j from the sample, and $\hat{T}_{\mathrm{L}(\cdot)}$ is the average of the $n\hat{T}_{\mathrm{L}(j)}$s. Royall & Cumberland [13] indicated that the jackknife variance estimate and v_{D} should perform similarly

as they were asymptotically equivalent. Deng & Wu [2] studied v_J, v_D, and a class of adjusted estimators $v_g(\hat{T}_L) = (\overline{X}/\overline{x}_s)^g v_C$, where g is to be chosen optimally from the sample. They also caution against the use of v_C, and note, with respect to both conditional and unconditional coverage of confidence intervals, that v_J, v_D, and v_2 are much better than v_C. Särndal et al. [17] proposed a variance estimator for $\hat{T}_L$ which, like (4), is calculated from weighted squared residuals.

Stratified Regression Estimators

Regression estimators of a population mean or total can also be defined when stratified random sampling is used, and an auxiliary variable x is available. Analogous to the case of ratio estimation, there is the separate regression estimator $\hat{T}_{LS}$, and the combined regression estimator $\hat{T}_{LC}$. The separate regression estimate of a population total is

$$\hat{T}_{LS} = \sum_{h=1}^{H} \hat{T}_{Lh},$$

where $\hat{T}_{Lh} = N_h \overline{y}_{s_h} + b_h(N_h \overline{X}_h - N_h \overline{x}_{s_h})$ is the linear regression estimator of the stratum total, $\sum_{i=1}^{N_h} y_{hi}$, using data only from that stratum. Here

$$b_h = \frac{\sum_{i \in s_h} (x_{hi} - \overline{x}_{s_h})(y_{hi} - \overline{y}_{s_h})}{\sum_{i \in s_h} (x_{hi} - \overline{x}_{s_h})^2}$$

is the estimated slope in stratum h. The combined regression estimator uses a single slope estimate, combining information across strata,

$$b_c = \frac{\widehat{\mathrm{cov}}(\hat{X}, \hat{Y})}{\widehat{\mathrm{var}}(\hat{X})} = \frac{\sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h(n_h - 1)} \sum_{i \in s_h} (x_{hi} - \overline{x}_{s_h})(y_{hi} - \overline{y}_{s_h})}{\sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h(n_h - 1)} \sum_{i \in s_h} (x_{hi} - \overline{x}_{s_h})^2},$$

where $\hat{X}$ and $\hat{Y}$ are the simple stratified expansion estimators of the x-total and the y-total, respectively.

Variance estimation for the separate linear regression estimator is straightforward, since it is a sum of H independent regression estimates. Hence a robust estimator is $v_D(\hat{T}_{LS}) = \sum_{h=1}^{H} v_D(\hat{T}_{Lh})$, where each $v_D(\hat{T}_{Lh})$ is calculated as described earlier in (4) for the unstratified case. A similar calculation can be done summing the jackknife variance estimates. For the combined regression estimator $\hat{T}_{LC}$, variance estimation is more problematical. Valliant [20] studied several variance estimators for $\hat{T}_{LC}$, and recommended a robust version analogous to v_D (4). He also found the performance of the traditional linearization estimators unacceptable. Valliant showed empirically that the robust estimator and a jackknife variance estimate were generally superior to the other choices of variance estimator. Because the formula for v_D is unwieldy, the jackknife variance estimate for $\hat{T}_{LC}$ is a good choice for variance estimation for the combined regression estimator. Its calculation is

$$v_J(\hat{T}_{RC}) = \sum_{h=1}^{H} \frac{1 - f_h}{n_h} (n_h - 1) \sum_{j \in s_h} (\hat{T}_{(hj)} - \hat{T}_{(h)})^2,$$

where $\hat{T}_{(hj)}$ is the estimate calculated without the hjth unit and $\hat{T}_{(h)} = \sum_{j \in s_h} \hat{T}_{(hj)}/n_h$. Valliant [20] indicated some simplifications to v_J that considerably lessen the computational burden of the jackknife in stratified sampling.

Stratified Systematic Sampling

We have already indicated some of the benefits of using stratification on x or systematic sampling from strata ordered on x, with the ratio or regression estimators. Kott [5] noted that systematic sampling is one way of protecting against certain kinds of model failure. Valliant [21] studied stratification on x and, using systematic sampling within strata, compared this plan with simple random sampling within strata. He considered the ratio and regression estimators (both separate and combined) and a number of variance estimators. From theoretical considerations and an empirical study, he recommended for moderate n_h the separate regression estimator, $\hat{T}_{LS}$, combined with stratified systematic sampling, and suggested using either a jackknife variance estimator or the robust variance estimate $v_D(\hat{T}_{LS})$ (4). This combination provided the most reliable inferences for the population total.

References

[1] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[2] Deng, L.Y. & Wu, C.F.J. (1987). Estimation of variance of the regression estimator, *Journal of the American Statistical Association* **82**, 568–576.

[3] Jones, H.L. (1974). Jackknife estimation of functions of stratum means, *Biometrika* **61**, 343–348.

[4] Kish, L., Namboodiri, N.K. & Pillai, R.K. (1962). The ratio bias in surveys, *Journal of the Royal Statistical Society, Series B* **36**, 1–37.

[5] Kott, P.S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population, *Biometrika* **73**, 485–491.

[6] Krewski, D. & Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods, *Annals of Statistics* **9**, 1010–1019.

[7] Lemeshow, S. & Levy, P.S. (1978). Estimating the variance of the ratio estimates in complex sample surveys with two primary units per stratum – a comparison of balanced replication and jackknife techniques, *Journal of Statistical Computing and Simulation* **8**, 191–195.

[8] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations – Methods and Applications*, 2nd Ed. Wiley, New York.

[9] Rao, J.N.K. & Wu, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics, *Journal of the American Statistical Association* **80**, 620–630.

[10] Rao, P.S.R.S. & Rao, J.N.K. (1971). Small sample results for ratio estimators, *Biometrika* **58**, 625–630.

[11] Royall, R.M. & Cumberland, W.G. (1978). Variance estimation in finite population sampling, *Journal of the American Statistical Association* **73**, 351–358.

[12] Royall, R.M. & Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance (with discussion), *Journal of the American Statistical Association* **76**, 66–88.

[13] Royall, R.M. & Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance – an empirical study, *Journal of the American Statistical Association* **76**, 924–930.

[14] Royall, R.M. & Cumberland, W.G. (1985). Conditional coverage properties of finite population confidence intervals, *Journal of the American Statistical Association* **80**, 355–359.

[15] Royall, R.M. & Eberhardt, K.R. (1975). Variance estimates for the ratio estimator, *Sankhyā, Series C* **37**, 43–52.

[16] Royall, R.M. & Herson, J. (1973). Robust estimation in finite populations II: stratification on a size variable, *Journal of the American Statistical Association* **68**, 890–893.

[17] Särndal, C.E., Swensson, B. & Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator, *Biometrika* **76**, 527–537.

[18] Särndal, C.E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[19] Scott, A. & Wu, C.F. (1981). On the asymptotic distribution of ratio and regression estimators, *Journal of the American Statistical Association* **76**, 98–102.

[20] Valliant, R. (1987). Conditional properties of some estimators in stratified sampling, *Journal of the American Statistical Association* **82**, 509–519.

[21] Valliant, R. (1990). Comparisons of variance estimators in stratified random and systematic sampling, *Journal of Official Statistics* **6**, 115–131.

[22] Wu, C.F. (1982). Estimation of variance of the ratio estimator, *Biometrika* **69**, 183–189.

[23] Wu, C.F. (1985). Variance estimation for the combined ratio and combined regression estimators, *Journal of the Royal Statistical Society, Series B* **47**, 147–154.

[24] Wu, C.F.J. & Deng, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study, in *Scientific Inference, Data Analysis and Robustness*, G.E.P. Box, T. Leonard & C.F. Wu, eds. Academic Press, New York, pp. 245–277.

W.G. CUMBERLAND

Real Time Approach in Survival Analysis

The *real time approach* in survival analysis (or, more generally, event history analysis) means that the statistical modeling respects the original order in calendar time in which the events recorded in the data took place.

This principle is realized naturally in parametric **likelihood** inference and in **Bayesian** inference, where, if the modeling is based on conditional intensities **(hazards)** conditioned at each point in calendar time on the events in the past, the likelihood expression will always assume the same simple canonical form.

In more explicit terms, we can express the data from an observation interval $(0, t]$ in the form of $N(t)$ *marked points* (T_i, X_i), where $0 < T_1 < T_2 < \cdots < T_{N(t)} \leq t$ and where X_i is a description of the event (such as the index or label of a failed individual) which occurred at time T_i. If $\lambda_t(x)$ denotes the conditional intensity of an event indexed by x occurring at time t, and $\overline{\lambda}_t = \sum_x \lambda_t(x)$ is the corresponding "crude" intensity of an event regardless of its index, then the likelihood expression will be of the well-known canonical form

$$L_t = \prod_{(i:T_i \leq t)} \lambda_{T_i}(X_i) \times \exp\left(-\int_0^t \overline{\lambda}_s \, \mathrm{d}s\right).$$

Depending on the studied context, the intensity $\lambda_t(x)$ can correspond, for example, to the failure of individual x, and then depend on the recorded pre-t history through (possibly time-dependent) internal and/or external **covariates**, such as the age of individual x at time t, time elapsed from a treatment, if any, type of the treatment received, and, possibly, calendar time itself. Such dependencies, when modeled explicitly, will then lead to the model parameters appearing in some particular functional form in the likelihood expression. The real time approach can accommodate, without any additional difficulty, study designs involving **staggered entry** of individuals or general noninformative **censoring** schemes. Noninformative censoring will simply result in multiplicative contributions to the likelihood expression which do not depend on the model parameters of interest, and which therefore can be ignored in likelihood-based inference. For a concrete example, see [3]. From the point of view of asymptotic theory, the real time approach has the additional advantage that the *score* $\partial \log L_t / \partial \theta$, evaluated at the "correct" parameter value $\theta = \theta_0$, and viewed as a **stochastic process** in time parameter t, will always be a martingale with respect to the recorded pre-t histories and the probability P_{θ_0} [1, 2] (*see* **Counting Process Methods in Survival Analysis**). This can be used as a convenient technical device in proving **consistency** and asymptotic normality of the parameter estimators under weak mathematical conditions (see, for example, [4]).

The real time approach can be compared and, to some extent, contrasted with more commonly used **nonparametric** and **semiparametric** estimation techniques in survival analysis, such as the **Nelson–Aalen**, **Kaplan–Meier**, and the **Cox regression** estimators. In these methods, the individuals are first aligned according to some time reading which is used as a baseline, and statistical estimators are formed by comparing, for each individual failure, the intensity of the failing individual to the crude intensity of all individuals who were then at risk simultaneously (according to the baseline). If the baseline time reading does not match with calendar time (e.g. if the study design involves staggered entry and age or time from treatment is used as a baseline), then the original sequencing of the events recorded in the data will be changed in the realignment of the individuals. As a result, the natural notion of pre-t "past" may be lost, and this in turn may violate the assumptions underlying the statistical survival model, under consideration.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, pp. 681–682.

[2] Arjas, E. (1985). Contribution to the discussion on the paper by P.K. Andersen and Ø. Borgan, *Scandinavian Journal of Statistics* **12**, 150–153.

[3] Arjas, E. (1986). Stanford heart transplantation data revisited: a real time approach, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds. Wiley, New York, pp. 65–81.

[4] Arjas, E. & Haara, P. (1987). A logistic regression model for hazard: asymptotic results, *Scandinavian Journal of Statistics* **14**, 1–18.

E. ARJAS

Recall Bias

Recall bias occurs in **case–control studies** and refers to the **bias** that results when cases with the disease of interest tend to over- or underestimate their previous exposures, compared with controls without disease. For example, a woman who has just given birth to a malformed infant may more assiduously recall her antenatal drug exposures than a **control** woman who had a normal infant. Recall bias induces **differential error** and can seriously distort the results of a case–control study.

(*See also* **Bias in Case–Control Studies**; **Bias in Observational Studies**; **Bias, Overview**)

MITCHELL H. GAIL

Receiver Operating Characteristic (ROC) Curves

Receiver operating characteristic (ROC) analysis was developed to summarize data from signal detection experiments in psychophysics [21]. Today, the term refers to

> a method of quantifying how accurately experimental subjects, professional diagnosticians and prognosticators (and their various tools: tests or instruments yielding numerical results, combinations of data-collection and data-display devices, different amounts and types of information ...) perform when they are required to make a series of fine discriminations or to say which of two conditions or states of nature, confusable at the moment of decision, exists or will exist [50].

In biomedical applications, the two states are often referred to as diseased and nondiseased, or D+ and D− for short. Central to this analysis is the ROC curve, which displays diagnostic accuracy as *a series of pairs* of performance measures. Each pair consists of a true positive fraction (TPF) and the corresponding **false positive** fraction (FPF) for a given definition of "test" (t) positivity, t+. These fractions are calculated from the D+ and D− groups respectively, TPF as the proportion of (t+, D+) among those D+, and FPF as the proportion of (t+, D−) among those D−. In the medical literature, the term TPF is called **sensitivity** and the complement of the FPF is called **specificity**. For the performance of statistical tests, the term **power**, rather than sensitivity, tends to be used. Equivalently, one can use its complement, the **false negative** fraction (FNF, the complement of TPF) or the frequency (β) of a so-called type II error. There is no direct statistical term for specificity. Instead, statisticians again focus on the complement, using the false positive fraction (FPF, the complement of the true negative fraction, TNF) to denote the frequency (α) of what they call type I error.

Diagnostic performance is sometimes naively characterized using a *single* overall index of "accuracy", calculated as the sum of two proportions, i.e. the proportion of (t+, D+) instances plus the proportion of (t−, D−) instances, where the proportions are based on all patients undergoing the test. This index is a weighted average of sensitivity and specificity, using as weights the (particularistic) relative frequencies of the D+ and D− states. Using two measures, namely a (TPF, FPF) *pair*, avoids this arbitrariness.

Although a (TPF, FPF) pair is a big improvement over an overall accuracy index, it is often not sufficient. A single (TPF, FPF) pair still does not allow meaningful comparison of the performance of one diagnostic test with another, or even with the same test performed in another setting or by another observer, when different criteria for test positivity are used in the two instances compared. The ROC curve, in the form of a *series* of (TPF, FPF) pairs (see Figure 1), isolates a test's capacity to discriminate between a given disease and its absence, from the **confounding** influence of the decision criterion (confidence level or cutting score) that is adopted for test positivity [37, 52, 58]. A more accurate test will be located on an ROC curve closer to the top left corner than a less accurate one. A noninformative test will have an ROC curve that lies along the diagonal.

Statistical techniques to handle the full range of ROC study designs continue to be developed [3, 5, 9, 26]. Analyses can vary in complexity from deriving an ROC curve for a single diagnostic test involving numerical values derived from patients at a single institution, to complex multi-institution studies to compare two or more imaging modalities. The complexity also depends on the purpose of the discrimination test, the setting and context to which it refers, whether in the study interpretations are performed individually in real time [18] or later in "batch" mode [52], and whether the tests under study and the procedures for independent definitive determination of the true state of nature (the **gold standard**) are costly, invasive, uncomfortable or dangerous.

This latter issue can create special problems since ROC curves are strongly influenced by the source of the test material used [6]. Distortions occur when the result of the test being studied affects the subsequent work-up needed to establish a definitive diagnosis. Information available on the distribution of test results and clinical indicants in the source population can be used to remove quantitatively this "verification bias" from ROC curves [20, 30]. Other **biases** in the assessment of diagnostic tests and guidelines for circumventing the problems in prospective studies have been described [4, 7].

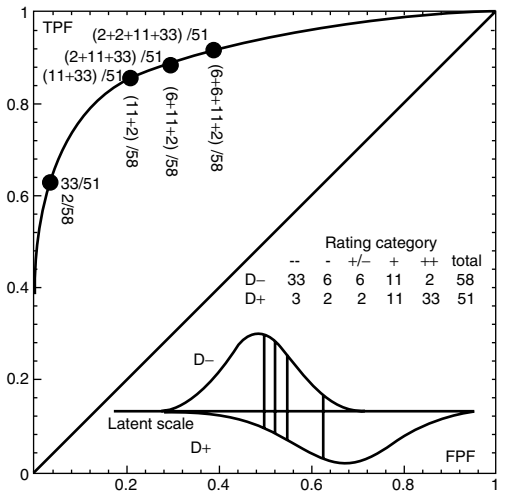

Figure 1 Example of empirical ROC points and smooth curve fitted to them. The empirical points are calculated from successively more liberal definitions of test positivity applied to the 2×5 table (inset) of disease status (D+ or D−) and rating category (−− to ++). The smooth ROC curve is derived from the fitted binormal model (inset, lower right, with parameters $a = 1.657$ and $b = 0.713$ on a continuous latent scale) by using all possible scale values for test positivity. The fitted parameters a and b, together with the four estimated cutpoints, produce fitted frequencies of $\{32.9, 6.4, 5.9, 10.7, 2.1\}$ and $\{3.2, 1.5, 2.1, 11.2, 32.9\}$ for the D− and D+ rows of the 2×5 table. Note that a monotonic transformation of the latent axis may produce overlapping distributions with nonbinormal shapes, but will yield the same multinomial distributions and the same fitted ROC curve

The complexity of the test material can have an important bearing on the ability of a study to compare tests. Cases resulting in an ROC curve that is midway between the diagonal (subtle or completely obscure ones) and the upper left corner (all 'obvious') allow for sizable differences in performance; however, the closer the curve is to the upper left corner, the narrower is the **sampling distribution** of the various indices derived from the curve [39].

For clinical imaging studies involving interpretations, the most economical method of collecting a reader's impression of each case is through the use of a rating scale, i.e. graded levels of confidence that the case is D+. A discrete five-point scale – 1 = "definitely not diseased", 2 = "probably not diseased", 3 = "possibly diseased", 4 = "probably diseased", 5 = "definitely diseased" – is commonly used. Getting a reader to use all of the rating categories provided yields a more stable ROC curve estimate, but is not always easy to accomplish without causing other problems [23]. Use of ratings from the continuous 0–100% confidence scale [31, 49] has several advantages: it more closely resembles reader's clinical thinking and reporting; its use of a finer scale leads to somewhat smaller **standard errors** of estimated indices of accuracy; and it increases the possibility that the data will allow parametric curve fitting.

Obtaining an ROC Curve and Summary Indices Derived from it

For *rating scale data*, the 2 (D states) $\times$ k (rating categories) frequency table of the ratings yields $k - 1$ empirical (TPF, FPF) ROC points. As shown in Figure 1, these are obtained from the $k - 1$ possible **two-by-two tables** formed by different re-expressions of the $2 \times k$ data table. After TPF = 0 at FPF = 0, the lowest leftmost ROC point is derived using the strictest cutpoint, where only the most positive category would be regarded as positive; each subsequent point towards the top right ROC corner (TPF = 1, FPF = 1) is obtained by employing successively laxer criteria for test positivity. For objective tests that yield *numerical data*, the same procedure – with each distinct observed numerical test value as a category boundary and with k no longer fixed a priori but rather determined by the numbers of 'runs' of D+ and D− in the aggregated data – is used to calculate the series of empirical ROC data points. The sequence of points can then be joined to form the empirical ROC curve or a smooth curve can be fitted.

As a summary measure of accuracy, one can use: (i) TPF[FPF], the TPF corresponding to a single selected FPF; (ii) the area under the ROC curve; or (iii) the area under a selected portion of the curve, often called the partial area. Summary (i) is readily understood and most clinically pertinent. However, reported TPFs are often in reference to different FPF values, and it may be unclear whether a reference FPF was chosen in advance or after inspection of the curve. Moreover, the statistical reliability tends to be lower than that of other summary indices.

Summary (ii) has been recommended as an alternative [52]. It has an interpretation in signal detection theory as the proportion of correct choices in a two-alternative forced choice experiment [21], i.e. an

experiment where in each trial the subject is presented with a pair of stimuli, one from a randomly chosen D+ and one from a randomly chosen D−, and is asked to decide which derives from which. This method of reporting judgments is common in psychophysics and has statistical advantages when using synthetic images, where observer time is the limiting factor [8].

To some, the area index has a serious limitation. Since a large part of the area comes from the rightmost part of the curve, it includes FPFs of no clinical relevance, and so can be insensitive when used to compare the performance of two tests. One curve may have higher TPFs than another in the region of relevant FPFs, but they could conceivably cross. Since the area under the entire curve averages the sensitivity over the full (0, 1) range of FPFs, any superiority in the relevant FPF region may be lost, or even reversed, when the curves are ranked on the basis of the entire area. The average sensitivity (TPF) over a range of relevant FPFs, summary index (iii) [34, 53, 60], is a compromise between (i) and (ii).

All three indices can be calculated either from the empirical or a (parametrically fitted) smooth curve. The statistical precision of nonparametric estimates can be calculated using a general method applicable to all three indices [60]; in the case of (ii) other essentially equivalent but less cumbersome methods, based on ***U*-statistics**, are also available [11, 28]. The case of (i) is more subtle than most realize: the standard error of an estimated TPF must include, in addition to its own obvious **binomial** variation, the uncertainty associated with determining the position of the FPF point [32].

A smooth ROC curve can be fitted to rating scale data by fitting two overlapping distributions on a continuous but "latent" scale underlying the results for D− and D+ cases [12, 36] In the most commonly used, "binormal", model, the two distributions are taken to be, without loss of generality, N(0, 1) for D−, and N(μ, σ) for D+. The distributions of the ratings are thus **multinomial**, with **expectations** that are functions of the $k-1$ cutpoints and the two parameters μ and σ, allowing the $k+1$ (two relevant and $k-1$ **nuisance**) parameters to be fitted to the observed data table ($2k-2$ **degrees of freedom** in total, leaving $k-3$ degrees of freedom to test the fit) using the criterion of **maximum likelihood**. Small additions to empty cells can be used to avoid "degenerate" situations [13]. The extent to which the normal deviate (*see* **Normal Scores**) **transformations** (z[TPF], z[FPF]) of the empirical (TPF, FPF) pairs are linear provides a visual test of the fit, since under the binormal model their expectations satisfy

$$z(\text{TPF}) = (1/\sigma)z(\text{FPF}) - \mu/\sigma = bz(\text{FPF}) - a.$$

When $b = 1$, the curve in (TPF, FPF) space is symmetric about the negative diagonal, while $b < 1$ produces a curve which rises more steeply at first and "flattens out" at the end.

Since the various summary indices derived from the fitted curve are functions of the estimates of a and b, their statistical precision – used in tests (*see* **Hypothesis Testing**) and **confidence intervals** – can be calculated from the corresponding variance/covariances provided by the maximum likelihood procedure. Confidence intervals can also be calculated for the entire curve [33].

For rating scale data, the popularity of the binormal model over bilogistic [22, 47] or other competitors [16] is more historical than theoretical. Use of a binormal model for rating data does not imply that if one could observe the latent distributions, they would have this exact form [38]. Rather, the working assumption is that the two overlapping multinomial distributions can be mathematically predicted from the discretization of two **normal distributions** on some unspecified latent scale. Whereas any two overlapping distributions will uniquely determine a specific ROC curve, the reverse is not true: the "binormal" assumption concerns only the functional form of the ROC curve, which can always be examined empirically, and not the form of the underlying distributions themselves, which cannot be determined in many applications of ROC analysis [38]. Use of a small number of rating categories, with few degrees of freedom, to distinguish the fit of one specific form over another, leaves considerable freedom to fit different distributional forms. This freedom is not a function of sample sizes (numbers of cases) but of the number of rating categories [25].

More important than the choice of distributional family seems to be the need to allow for unequal **variances** ($b \neq 1$). Empirically, b tends to be less than 1 [51], possibly because of the presence of unidentified subtypes in the D+ sample. Thus, whereas one-parameter models, with $b = 1$, would have practical advantages, particularly for **meta-analyses** and for fitting an entire (but symmetric)

ROC curve to a single empirical (TPF, FPF) data point, they are not supported by empirical findings.

Care must be taken in the fitting of parametric curves to results recorded on a numerical scale: directly fitting $N(\mu_1, \sigma_1)$ for D− and $N(\mu_2, \sigma_2)$ for D+ can yield severe distortions when the data do not arise from normal distributions [19]. The method in which the raw data are first categorized and the categorized data analyzed as if they were rating data [41] – with the assumptions of overlapping normal distributions on an unspecified transform of the actual measurement scale – is a much more **robust** approach.

Comparison of ROC Curves

Because of the need to account for large differences in case difficulty, compared curves are usually based on the same set of cases, and one must therefore take the **correlation** of the estimated curves – and summaries derived from them – into account when calculating the standard error of differences.

Parametric methods for comparing two curves are based on the estimates of the binormal (or bilogistic, or other model) parameters (a, b) associated with each curve, and their variances and covariances [34, 42]. The equality of two curves can be assessed by testing the equality of the two vectors (a_1, b_1) and (a_2, b_2). Comparisons involving a summary index are made by computing, for each curve, the appropriate function of the parameter estimates, then using the **delta method** to calculate the standard error of the difference in indices.

A **nonparametric method** is now available to compare two curves based on continuous data from the same set of cases [59]. A criterion for positivity that is common for the two tests is induced using the **ranks** in the combined D+ and D− data for each test. Using this calibration, one first computes the difference in the numbers of errors made by the two tests at each possible level of test positivity and then calculates the average of the absolute differences over the different levels. The test statistic is referred to the permutation distribution obtained by randomly interchanging pairs of ranks. Nonparametric comparisons of the areas under two curves are based on correlated U-statistics [11] or equivalently on the **jackknife method** [27], while partial areas, and – ultimately – sensitivity at a single specificity value can be compared with a more general method [60].

Guidelines for **sample size determination** and **power** calculations are available for both parametric [software program ROCPWR from Charles Metz at the University of Chicago, or the article by Obuchowski & McClish [43]] and nonparametric approaches [29].

Comparison of Accuracy of Imaging Procedures

The methods just described deal only with simple comparisons of two tests that yield objective numerical results, and are not sufficient for imaging studies (*see* **Image Analysis and Tomography**) which produce interpretations of each case by multiple readers. For example, the performance with conventional versus laser printed films might be studied by having several readers interpret each image; a comparison of computed tomography, magnetic resonance and ultrasound images might involve different readers for each modality [46]. Several refinements and some alternatives to the method initially proposed for dealing with these more complex comparisons have been suggested. The challenge is to include and estimate properly each of the several relevant **components of variance** and covariance since the comparison is necessarily an average over cases, readers and (possibly) rereadings [40, 52]. Two methods [15, 44] deal with the problem by modeling the variation in the summary index in question, while another [54] models the raw rating data responses. From the investigations thus far [14, 55], both modeling approaches appear to give comparable answers, but commentators [48, 35] have called for some further work to investigate the performance of analysis strategies that use statistical tests to decide what is the appropriate error term and denominator **degrees of freedom** when reader x modality **interactions** are involved.

When studies of imaging procedures involve multiple centers, complex procedures, ethical concerns, "real-time" readings, and different experts in the different imaging modalities [18], the data can quickly become imbalanced and/or incomplete. Analysis problems are thus aggravated by the subjective (and thus possibly nonpoolable) nature of the ratings, the often large numbers of case–reader sets, each containing too few observations to allow parametric fitting of separate ROC curves, and the fact that,

unlike the usual response measures in **clinical trials**, the elemental ROC data are not absolute numbers that can be easily averaged or displayed individually in a descriptive way. If all available data are to be used, the only logical approach is to use **regression** methods. Since the first regression work in this area [57], there has been considerable activity in developing **parsimonious** approaches to the problem, including **random effects** models [2], Gibbs sampling (*see* **Markov Chain Monte Carlo**) [17], and **generalized estimating equations** [56, 61].

In some situations, the study material may involve more than one region of interest on an image. Sample reuse methods can help to calculate the precision with which statistical contrasts are made [1, 10, 24, 45] (*see* **Bootstrap Method**).

Software

Programs for parametric estimation are available from the WWW location `www-radiology.uchicago.edu/cgi-bin/software.cgi` maintained by the developer, Charles Metz. Special-purpose programs for nonparametric inference are more numerous, but most of the tasks can be accomplished using a spreadsheet [27]. Software for the multireader, multicase approach to imaging data is available from the authors of ref. [15]; software for the other approaches is still evolving and interested users should contact the various authors cited above.

Future Developments

Whereas methods for comparing ROC curves associated with tests that yield objective test results have now become routine, solutions to the complex analytic problems involved in the comprehensive comparison of accuracy of imaging procedures have not been completely achieved. The methods proposed in the last 5 years need further testing; the links between them need to be better understood; and user-friendly software to implement these newest approaches remains to be developed.

References

[1] Baum, R.A., Rutter, C.M., Sunshine, J.H., Blebea, J.S., Carpenter, J.P., Dickey, K.W., Quinn, S.F., Gomes, A.S., Grist, T.M. et al. (1995). Multicenter trial to evaluate vascular magnetic resonance angiography of the lower extremity. American College of Radiology Rapid Technology Assessment Group, *Journal of the American Medical Association* **274**, 875–880.

[2] Beam, C. (1995). Random effects models in the ROC curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches and issues, *Academic Radiology* **2**, Supplement 1, S4–S13.

[3] Begg, C.B. (1986). Statistical methods in medical diagnosis, *Critical Reviews in Medical Informatics* **1**, 1–22.

[4] Begg, C.B. (1987). Biases in the assessment of diagnostic tests, *Statistics in Medicine* **6**, 411–424.

[5] Begg, C.B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980s, *Statistics in Medicine* **10**, 1887–1895.

[6] Begg, C.B. & Greenes, R.A. (1983). Assessment of diagnostic tests when disease verification is subject to verification bias, *Biometrics* **39**, 207–215.

[7] Begg, C.B. & McNeil, B.J. (1988). Assessment of radiologic tests: control of bias and other design considerations, *Radiology* **167**, 565–569.

[8] Burgess, A.E. (1995). Comparison of receiver operating characteristic and forced choice observer performance measurement methods, *Medical Physics* **22**, 643–655.

[9] Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests, *Statistics in Medicine* **13**, 499–508.

[10] Dagirmanjian, A., Ross, J.S., Obuchowski, N., Lewin, J.S., Tkach, J.A., Ruggieri, P.M. & Masaryk, T.J. (1995). High resolution, magnetization transfer saturation, variable flip angle, time-of-flight MRA in the detection of intracranial vascular stenoses, *Journal of Computer Assisted Tomography* **19**, 700–706.

[11] DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver-operating characteristic curves: a nonparametric approach, *Biometrics* **44**, 837–845.

[12] Dorfman, D.D. & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data, *Journal of Mathematical Psychology* **6**, 487–496.

[13] Dorfman, D.D. & Berbaum, K.S. (1995). Degeneracy and discrete receiver operating characteristic rating data, *Academic Radiology* **2**, 907–915.

[14] Dorfman, D.D. & Metz, C.E. (1995). Rejoinder in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S76–S78.

[15] Dorfman, D.D., Berbaum, K.S. & Metz, C.E. (1992). Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method, *Investigative Radiology* **27**, 723–731.

[16] Egan, J.P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press, New York.

[17] Gatsonis, C. (1995). Random-effects models for diagnostic accuracy data, *Academic Radiology* **2**, Supplement 1, S14–S21.

[18] Gatsonis, C. & McNeil, B.J. (1990). Collaborative evaluation of diagnostic tests: experience of the Radiologic Diagnostic Oncology Group, *Radiology* **175**, 571–575.

[19] Goddard, M.J. & Hinberg I (1990). Receiver operating characteristic (ROC) curves and non-normal data: an empirical study, *Statistics in Medicine* **9**, 325–337.

[20] Gray, R. & Begg C.B. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias, *Medical Decision Making* **4**, 151–164.

[21] Green, D.M. & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

[22] Grey, D.R. & Morgan, B.J.T. (1972). Some aspects of ROC curve fitting: normal and logistic models, *Journal of Mathematical Psychology* **9**, 128–139.

[23] Gur, D., Rockette, H.E., Good, W.F., Slasky, B.S., Cooperstein, L.A., Straub, W.H., Obuchowski, N.A. & Metz, C.E. (1990). Effect of observer instruction on ROC study of chest images, *Investigative Radiology* **25**, 230–234.

[24] Hajian-Tilaki, K.O., Hanley, J.A., Joseph, L. & Collet, J.P. (1997). Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks, *Academic Radiology* **4**, 222–229.

[25] Hanley, J.A. (1988). The robustness of the binormal model used to fit ROC curves, *Medical Decision Making* **8**, 197–203.

[26] Hanley J.A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art, *Critical Reviews in Diagnostic Imaging* **29**, 307–335.

[27] Hanley, J.A. & Hajian-Tilaki K.O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update, *Academic Radiology* **4**, 49–58.

[28] Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under an ROC curve, *Radiology* **143**, 129–133.

[29] Hanley, J.A. & McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same set of cases, *Radiology* **148**, 839–843.

[30] Hunink, M.G., Richardson, D.K., Doubilet, P.M. & Begg, C.B. (1990). Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing, *Medical Decision Making* **10**, 201–211.

[31] King, J.L., Britton, C.A., Gur, D., Rockette, H.E. & Davis, P.L. (1993). On the validity of the continuous and discrete confidence rating scales in receiver operating characteristic studies, *Investigative Radiology* **28**, 962–963.

[32] Linnet, K. (1987). Comparison of quantitative diagnostic tests: type I error, power and sample size, *Statistics in Medicine* **6**, 147–158.

[33] Ma, G. & Hall, W.J. (1993). Confidence bands for receiver operating characteristic curves, *Medical Decision Making* **13**, 191–197.

[34] McClish, D.K. (1989). Analyzing a portion of the ROC curve, *Medical Decision Making* **9**, 190–195.

[35] McClish, D.K. (1995). Invited discussion in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S61–S64.

[36] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[37] Metz, C.E. (1978). Basic principles of ROC analysis, *Seminars in Nuclear Medicine* **8**, 283–298.

[38] Metz, C.E. (1986). ROC methodology in radiological imaging, *Investigative Radiology* **21**, 720–733.

[39] Metz, C.E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies, *Investigative Radiology* **24**, 234–245.

[40] Metz, C.E. & Shen, J.H. (1992). Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis, *Medical Decision Making* **12**, 60–75.

[41] Metz, C.E., Shen, J.H. & Herman, B.A. (1990). New methods for estimating a binormal ROC curve from continuously distributed test results. Paper presented at the annual meeting of the American Statistical Association, Anaheim.

[42] Metz, C.E., Wang, P-L. & Kronman, H.B. (1984). A new approach for testing the significance of differences between ROC curves from correlated data, in *Information Processing in Medical imaging*, F. Deconink, ed. Martinus Nijhoff, The Hague, pp. 432–445.

[43] Obuchowski, N.A. & McClish, D.K. (1997). Sample size determination for diagnostic accuracy studies involving binormal r.o.c. curve indices, *Statistics in Medicine* **16**, 1529–1542.

[44] Obuchowski, N.A. (1995). Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using analysis of variance with dependent observations, *Academic Radiology* **2**, Supplement 1, S22–S29.

[45] Obuchowski, N.A. (1996). Nonparametric analysis of clustered ROC data. Presentation at Eastern North American Biometrics Meeting.

[46] Obuchowski, NA. & Zepp, RC. (1996). Simple steps for improving multiple-reader studies in radiology, *American Journal of Roentgenology* **166**, 517–521.

[47] Ogilvie, J.C. & Creelman, C.D. (1968). Maximum likelihood estimation of ROC curve parameters, *Journal of Mathematical Psychology* **5**, 377–391.

[48] Rockette, H.E. (1995). Contributed comments in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S70–S71.

[49] Rockette, H.E., Gur, D. & Metz, C.E. (1992). The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques, *Investigative Radiology* **27**, 169–172.

[50] Swets, J.A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models, *Psychological Bulletin* **99**, 100–117.

[51] Swets, J.A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance, *Psychological Bulletin* **99**, 181–198.

[52] Swets, J.A. & Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

[53] Thompson, M.L. & Zucchini W. (1989). On the statistical analyses of ROC curves, *Statistics in Medicine* **8**, 1277–1290.

[54] Toledano A. & Gatsonis, C.A. (1995). Regression analysis of correlated receiver operating characteristic data, *Academic Radiology* **2**, Supplement 1, S30–S36.

[55] Toledano A. & Gatsonis, C.A. (1995). Rejoinder in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S81–S82.

[56] Toledano, A.Y. & Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data, *Statistics in Medicine* **15**, 1807–1826.

[57] Tosteson, A.N.A. & Begg, C.B. (1988). A general regression methodology for ROC curve estimation, *Medical Decision Making*, **8**, 204–215.

[58] Turner, D.A. (1978). An intuitive approach to receiver operating characteristic curve analysis, *Journal of Nuclear Medicine* **19**, 213–220.

[59] Venkatraman, E.S. & Begg, C.B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment, *Biometrika* **83**, 835–848.

[60] Wieand, S., Gail, M.H., James, K.L. & James, B.R. (1988). A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data, *Biometrika* **76**, 585–592.

[61] Zhou, X.H. (1996). Empirical Bayes combination of estimated areas under ROC curves using estimating equations, *Medical Decision Making* **16**, 24–28.

(*See also* **Diagnostic Test Evaluation Without a Gold Standard**; **Diagnostic Tests, Evaluation of**)

J.A. HANLEY

Record Linkage

At the core of all **descriptive epidemiology** studies lies a data set, with many variables, which has been gathered to answer a specific hypothesis. Often it is only as the project develops that the researcher realizes the potential of exploring alternate study endpoints by adding in other data about the same respondents. The tried and tested technique is for a clerk to look at the individual records, sorted in some logical order, and put the records together, applying intuitive decision rules based on human judgment. As record systems have been computerized over the past 20 years, one of the greatest impacts of increased processing power has been to facilitate linkages between related data sets, even when they do not share a unique identifier.

Three main techniques are used for record linkage, Newcombe [13] and Jamieson et al. [9] describe in detail the technical issues relating to exact matching and probability linkage (*see* **Matching, Probabilistic**). They can be summarized as follows:

1. *Unique*. Records are linked together where unique identifiers such as insurance number or health service number match exactly. The files of records are computer sorted into the same order, and matched together within blocks. It is a fairly simple process, but may only identify 80%–85% of true matches due to errors in recording of identifiers.
2. *Fuzzy*. For data sets which do not have unique identifiers, key identifiers such as surname, date of birth, sex, date of interview/treatment, and postal district are used for linkage. To cope with coding errors, fuzzy matching identifies records which are "almost" the same, such as surname spelling incorrect, or year and month of birth correct, but day wrong. Computer programs either present a choice of matches for the user to choose the best match, or have incorporated a simple scoring system and determine the best match from the score. Computer **algorithms** are well developed for matching on individual variables, and this technique provides 85%–90% of true matches. It requires human intervention and there may be operator **bias**.
3. *Probability*. This is the most sophisticated form of linkage, in which decision rules on records matching are programmed based on the probability of two records being from different people having the same identifier. These probabilities are aggregated to a score and checked against a threshold to determine whether a match is made. The computer system needs to be tailored for the data sets to be matched and is processor intensive, but provides linkages of 95%–99% true matches with false positive rates of 1%–2%.

The following are examples of the uses made of record linkage within healthcare systems and demonstrate the value of this powerful technique. Many of the examples come from uses made of the Scottish Medical Record Linkage Database [7], which contains morbidity records from Scottish hospitals, and mortality records from the General Register Office (Scotland) from 1968 onwards – almost 4 million people with 12 million episodes.

Medical record linkage poses problems of data **confidentiality** and privacy, because the linked data are comprehensive and the techniques use personal identifying data. Most analytic studies do not require access to patient identifiable data once the linkages have been made. For administrative data sets, strict controls need to be in place to ensure that the data are not released to individuals and used for purposes other than those registered in government legislation.

The issue of infringing civil liberties, by invasion of privacy through wrongful use of information, is currently taxing most governments. Researchers need to be aware of legislation and appropriate use of data. For example, in Scotland, access to identifiable data is controlled by medically qualified data holders, and a Privacy Advisory Committee [10] has been established, with membership drawn from senior medical officers, legal professions, and the public, to ensure that ethical approval is in place for record linkage studies.

Evidence-Based Medicine

The perception remains that descriptive epidemiology has little to contribute to the development of **evidence-based medicine** with its focus on randomized **clinical trials** (see McPherson [12]). Probability-based record linkage techniques can make a major contribution in assessing the efficacy of treatment regimes at the macro level. For example, using exact

matching on health service administration numbers, Evans et al. [4] at the Tayside Medicines Monitoring Unit are able to use **case–control** methodology to review the association of topical nonsteroidal anti-inflammatory drugs with hospital admission for upper gastrointestinal bleeding and perforation. The Scottish Health Service are now automating the linkage between prescribing data, patient hospitalization, and death profiles. This will establish a facility for **post-marketing surveillance**, in which possible adverse drug reactions can be quickly analyzed and assessed before the public are alarmed by the media (*see* **Pharmacoepidemiology, Overview**).

Another area in which linkage is being used effectively is the follow-up of very low birthweight children and the impact of their improved survival on health care costs. In California [2], probability-linked data from the California Birth Cohort and Medicaid claims in years after birth have been used to evaluate competing hypotheses for racial and ethnic differences (*see* **Ethnic Groups**) in mortality and health care costs, and to assess the need for hospital services from the improved neonatal survival of these children.

Outcome Measurement

Evaluating the effects of medical care is not a new idea, but it has received increased emphasis over the past decade because of concerns for the quality and cost of medical care (*see* **Quality of Care**). While most attention has been placed on determining the effectiveness of new treatment regimes through randomized control trials, the inclusion criterion for patients can be so selective (*see* **Eligibility and Exclusion Criteria**) that the true efficacy of the treatment can only be assessed when it comes into general usage. Application of record linkage techniques using **administrative databases** for follow-up of cohorts of patients with specific disease patterns, or procedures, permits analysis of outcomes measures which would otherwise be prohibitively expensive.

The Clinical Resource and Audit Group of the Scottish Office Department of Health have pioneered the publication of routine clinical outcome measures in the UK since 1993. The three reports [17] to date have been produced following detailed consultation with health service professionals, to gain consensus on the measures and to assess the feasibility of using them to monitor the effectiveness and appropriateness of health purchasing strategies. Without a unique patient identifier, probability linkage is the key to determining readmission rates, including to other institutions, and postoperation survival after discharge.

While the measures tend to be presented as interval estimates (*see* **Estimation, Interval**), standardized for **confounding** factors, such as age, sex, deprivation, and co-morbidities, administrative data do not yet contain **robust** measures of severity of disease. As with all descriptive epidemiology techniques, the outcome measures highlight topics for more detailed investigation via randomized controlled trials or clinical audit.

Survival Rates

As the search continues for new, meaningful outcome measures, one of the main uses of record linkage has been analysis of survival patterns for disease, especially for cancer (*see* **Survival Analysis, Overview**). Most civil registration authorities provide an exact matching service for bona fide researchers. However, increased computing power has meant that this process can now be automated to include probability or "fuzzy" matching techniques, which increase the reliability of the links. Within the Scottish Cancer Registry, we found that exact matching with manual techniques under-ascertained almost 5% of deaths [16], because the procedures were built on zero tolerance of **false positive** rates. This resulted in one study for the nuclear industry showing a "healthy worker" effect (*see* **Occupational Epidemiology**), until another three deaths were determined by automated probability linkage among the cohort of employees.

The availability of population-based data in specific **diseases registers**, such as cancer, diabetes, and renal failure, with linkage to death registrations enables the development of survival tables [16] (*see* **Life Table**), which are of use not only to the professional dealing with individual patients but also to the patients and their carers. Society is becoming more attuned to the concepts of **risk**, and one of the most common questions asked when life-threatening disease is diagnosed is "What is my chance of surviving 1 year, 5 years, or 10 years?". Insurance companies are very interested in improved estimates of

actuarial risk (*see* **Actuarial Methods**) for health care policies.

Changing Treatment Patterns in Hospital Care

Access to databases containing linked patient episodes over long time periods (15+ years) helps to identify the changing treatment patterns and use of hospital resources by cohorts of patients with specific disease. For example, the protocol for clinical treatment of asthma in children has changed considerably over the past 20 years, and Strachan et al. [18] used linked data for children with their first hospital admission for asthma between 1980 and 1984 to explore the increase in subsequent emergency admissions for the disease during the following 10 years (*see* **Health Care Utilization Data**).

In mental health (*see* **Psychiatry**) the impact of policies for shifting care from the acute sector to the community can be monitored from linked data sets. Geddes & Juszczak [5] argue that trends in increased suicide rates for recently discharged female psychiatric patients may well be related to changes in discharge protocols due to implementation of government policy and new clinical practice. In a similar vein, the Scottish Health Service is currently linking mental health discharge records to the general hospital patient database, to investigate if the policy of early discharge from psychiatric institutions has resulted in psychiatric patients being readmitted to acute care after a short period in the community.

The effect on emergency readmission rates from early hospital discharge and associated **quality of care** have been identified by Henderson et al. [8] and Thomas & Holloway [19]. Investigation of Scottish data demonstrates "like" Trusts which have significantly different medical emergency readmission rates, with inversely related bed occupancy rates and lengths of continuous inpatient stay. Instead of focusing on efficiency (high throughput, and short length of hospital stay) commissioners of health care can consider the effect of a 50% variation in the risk of emergency readmissions: Does this provide acceptable levels of value for money versus quality of care for the patient?

One of the roles of descriptive epidemiology is to aid the understanding of uncertainty. McPherson [12] uses the example of 5 year mortalities following treatments for prostatectomy, reported from analysis of large linked databases, to highlight the impact of changing clinical practice without the rigors of assessment trials, and the concerns raised amongst patients when consensus cannot be reached amongst clinicians on the effect of different treatments.

Health Service Planning

The potential of taking data created routinely as part of a government's system for paying for medical care and turning it into information on health needs and the health of the population is well demonstrated by the work of Roos & Shapiro [14] and Roos et al. [15] and his team at the Manitoba Center for Health Policy and Evaluation (*see* **Health Services Research, Overview**). The research attempts to move beyond medical care policy initiatives (e.g. insuring availability, quality of care, and efficiency) to health policy initiatives of improving longevity and **quality of life**. As pressures grow, throughout Europe and North America, to contain the costs of health care by reducing investment in acute sectors, health planners are looking to such population health information systems for quantitative trend data on which to base their decisions. Evidence is needed to answer questions such as:

1. Are high risk populations poorly served or do they have poor health outcomes despite availability of services?
2. Can we shift resources from acute care to primary care?
3. What services can be rationed without jeopardizing at-risk populations?

Lack of population morbidity information leads epidemiologists to use hospitalization rates as proxy measures for underlying morbidity. Improved availability of **general practice** diagnosis and treatment data from administrative systems in the surgery allows estimation of the true level of demand in communities, which can be linked with hospital discharge data at a patient level. This is invaluable for needs assessment work in public health, where commissioners of services attempt to balance supply with demand within small geographic areas (*see* **Small Area Variation Analysis**).

Longitudinal and Cohort Studies

Many of the most renowned epidemiologic studies, such as **Framingham** [20] and Whitehall [11], used the manual tracking systems available from civil death registries. The advent of computer technology has made automated follow-up of survey data for alternative endpoints other than death a simple process – provided that informed consent is obtained by the subject for access to computerized medical records.

In 1973–76, a cohort of the population in the west of Scotland towns of Paisley and Renfrew, aged 45–64 years, took part in a cardiovascular survey of mid-life health, the MIDSPAN [6] project. Participants gave written consent to their medical records being used for follow-up, and the data set now includes details of all episodes of hospital care and death registrations for participants as they have aged over the past 20 years. It can be used for prospective **case–control studies** to investigate determinants of good and poor health from baseline lifestyle variables and clinical measurements.

The West of Scotland Coronary Prevention Study (WOSCOPS) [21] demonstrated the value of automated linkage in a randomized–controlled trial compared to prospective follow-up using direct contact with patients. An accuracy check of the study's own independent records of deaths and hospitalizations for the study population with data available from the Scottish Medical Record Linkage Database showed that while almost 100% accuracy was achieved for deaths, the study records under-ascertained hospitalizations for cardiovascular disease.

In the US, the Veterans Affairs database [1] has formed the source for **multicenter** randomized and quasi-randomized health service trials, which are much easier to plan and conduct in a centralized state system than in the private sector.

The UK Case–Control Study for Childhood Cancers will report findings in 1997. One of the methodological issues which has arisen from the study, access to case notes, has demonstrated the value of conducting applied research within the health administration system of the country rather than solely in an academic environment. The study is investigating hypotheses for cause and effect of cancer in children, covering ionizing radiation, chemical exposure, preconception and *in utero* exposure, parental occupational hazards, electromagnetic fields, and infectious exposure.

Summary

This article has described the main applications of record linkage techniques in epidemiology: from follow-up studies in randomized controlled trials and surveys to outcome measurement and survival following hospital care in the general population.

The Chief Medical Officer of the UK government [3] recently identified the need to use better descriptions of public health risk. We have demonstrated that, within health administration systems, much of the data already exists. When integrated using linkage techniques, these data can be used to build the knowledge base to identify these risks and to quantify them in both relative and attributable ways.

The use of such linked databases for descriptive epidemiology brings the following benefits:

1. data ascertainment, validity, and quality are documented within the administrative system;
2. linkage can be performed by computer at low cost relative to staff costs;
3. completeness of linkage is greater than by manual methods;
4. research efforts can focus on the analyses and interpretation of the data rather than data collection.

Provided that privacy and confidentiality rules are strictly applied, and users remember that in any linkage system, be it exact match or probability-based, one cannot be 100% certain that the correct data have been linked, there are vast data repositories available waiting to unlock the answers to key epidemiologic questions.

References

[1] Ashton, C.M., Menke, T.J., Deykin, D., Camberg, L.C. & Charns, M.P. (1996). A state-of-the-art conference on databases pertaining to veterans' health, *Medical Care* **34**, MS1–MS234.

[2] Bell, R.M., Keesey, J. & Richards, T. (1994). The urge to merge: linking vital statistics records and medicaid claims, *Medical Care* **32**, 1004–1018.

[3] Department of Health (1996). *"On the State of Public Health" – Annual Report of the Chief Medical Officer*. HMSO, London.

[4] Evans, J.M.M., McMahon, A.D., McGilchrist, A.D., White, G., Murray, F.E., McDevitt, D.G. & McDonald, T.M. (1995). Topical non-steroidal anti-inflammatory drugs and admission to hospital for upper gastrointestinal bleeding and perforation: a record-linkage case control study, *British Medical Journal* **311**, 22–26.

[5] Geddes, J.R. & Juszczak, E. (1995). Period trends in rates of suicide in first 28 days after discharge from psychiatric hospital in Scotland, 1968–92, *British Medical Journal* **311**, 357–360.

[6] Hawthorne, V.M., Beevers, D.G. & Greaves, D.A. (1974). Blood pressure in a Scottish town, *British Medical Journal* **3**, 600–603.

[7] Heasman, M.A. (1968). The use of record linkage in long-term prospective studies, in *Record Linkage in Medicine, Proceedings of the International Symposium, Oxford*, E.D. Acheson, ed. Oxford University Press, Oxford.

[8] Henderson, J., Evans, J.G. & Goldacre, M.J. (1991). Use of medical records linkage to study readmission rates, *British Medical Journal* **303**, 389–393.

[9] Jamieson, E., Roberts, J. & Browne, G. (1995). The feasibility and accuracy of anonymised record linkage to estimate shared clientele among three health and social service agencies, *Methods of Information in Medicine* **34**, 371–377.

[10] Kendrick, S. & Clarke, J. (1993). The Scottish Record Linkage System, *Health Bulletin* **51**, 72–79.

[11] Marmot, M.G., Rose, G., Shipley, M. & Hamilton, P.J.S. (1978). Employment grade and coronary heart disease in British civil servants, *Journal of Epidemiology and Community Health* **32**, 244–249.

[12] McPherson, K. (1994). The best and the enemy of the good: randomized controlled trials, uncertainty, and assessing the role of patient choice in medical decision making, *Journal of Epidemiology and Community Health* **48**, 6–15.

[13] Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford University Press, Oxford.

[14] Roos, N.P. & Shapiro, E. (1995). A productive experiment with administrative data, *Medical Care* **33**, DS7–DS12.

[15] Roos, N.P., Black, C.D., Frehlich, N., Decoster, C., Cohen, N., Tataryn, D., Mustard, C.A., Toil, F., Carriere, K.C., Burchill, C.A., MacWilliam, M. & Bogdanovic, B. (1995). A population-based health information system, *Medical Care* **33**, DS13–DS20.

[16] Scottish Health Service, Information & Statistics Division (1993). *Trends in Cancer Survival in Scotland: 1968–1990*. Common Services Agency, Edinburgh.

[17] Scottish Office Department of Health (1996). *Clinical Outcome Measures – CRAG Report*. HMSO, Edinburgh.

[18] Strachan, D.P., Seagroatt, V. & Cook, D.G. (1994). Chest illness in infancy and chronic respiratory disease in later life – an analysis of month of birth, *International Journal of Epidemiology* **23**, 1060–1068.

[19] Thomas, J.W. & Holloway, J.J. (1991). Investigating early readmission as an indicator for quality of care studies, *Medical Care* **29**, 377–394.

[20] Truett, J., Cornfield, J. & Kannel, W.A. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham, *Journal of Chronic Diseases* **20**, 167–179.

[21] West of Scotland Coronary Prevention Study Group (1995). Computerized records linkage compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study, *Journal of Clinical Epidemiology* **48**, 1441–1452.

MARY SMALLS & STEVE KENDRICK

Reduced Rank Regression

Given a set of predictors (**explanatory variables**), $x_1, x_2, \ldots, x_p$, and a response variable, y, the **multiple linear regression** procedure finds the best **least squares** linear prediction of the response variable given all of the predictors. If $\mathbf{X}$ is an $n \times p$ matrix of n observations on the predictor variables and $\mathbf{y}$ is an $n \times 1$ vector of observations on the response variable, the multiple regression model would be

$$\mathbf{y} = \mathbf{Xb},$$

where $\mathbf{b}$ is a $p \times 1$ vector of regression coefficients. The ordinary least squares estimator for $\mathbf{b}$ is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

and the lack-of-fit of this estimator is given by the standard error of estimate,

$$\left(\frac{\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}}{n - p - 1}\right)^{1/2}.$$

Among the many problems associated with multiple regression is the difficulty of *multicollinearity* of the predictor variables (*see* **Collinearity**). Multicollinearity means that two or more predictor variables are highly correlated (*see* **Correlation**) or that there are one or more linear constraints on these variables. In the latter case, $\mathbf{X}'\mathbf{X}$ does not have an inverse. Even if these variables are not completely collinear but are still correlated, the following problems may arise:

1. obtaining a stable inverse for $\mathbf{X}'\mathbf{X}$ may be difficult.
2. as the predictor variables become more correlated, the standard errors of the regression coefficients increase in size and the regression coefficients become more and more correlated. These conditions make it difficult to interpret these coefficients.

A number of procedures have been designed to deal with these problems. Early solutions include elimination of predictor variables in the model by various sequential procedures such as stepwise regression or the investigation of combinations of subsets of variables (*see* **Variable Selection**). Another procedure is **ridge regression**, which enhances the chances of getting a good inverse of $\mathbf{X}'\mathbf{X}$ at the expense of introducing bias into the regression coefficients. Most of these are rather ad hoc procedures. There are also some straightforward multivariate solutions.

Principal Components Regression

Principal components regression is a technique which requires the predictor variables to be transformed into **principal components** which, in turn, become the predictors in the least squares solution. Principal components regression deals directly with the problem of multicollinearity. If some of the predictors are perfectly correlated and/or other linear constraints exist, then the principal components analysis will produce one or more characteristic roots (**eigenvalues**) equal to zero. The characteristic vectors (**eigenvectors**) associated with these roots may be used to identify these situations (see, for example, [1]). This may suggest the deletion of some variables so that a stable regression solution can still be obtained. In many cases, the ordinary least squares solution may experience near-multicollinearity. Because the principal components are uncorrelated, there is no problem in principal components regression with matrix inversion. The regression coefficients are also uncorrelated and there is no inflation of their standard errors. This regression equation may then be restated back in terms of the original correlated predictor variables.

In addition to transforming correlated variables into uncorrelated ones, principal components analysis also allows one to approximate the original data with $k < p$ components to obtain a more **parsimonious** description of the structure of the original variables. Principal components regression may be carried out with this reduced set of components but the resultant regression coefficients, in terms of the original variables, would only be estimates of the ordinary least squares solution, and the corresponding standard error of estimate would be larger. However, if there is not much increase in this quantity and the principal components are interpretable, then this could be a useful prediction equation, particularly if k is small relative to p. This is the goal of principal components regression.

Principal components regression has been widely used in many fields, particularly with a reduced set of components. However, caution should be used with this technique. The principal components are

obtained sequentially in order by the amount of variability of the original variables they account for. Various stopping rules are available to determine how many components to retain, the general assumption being that the variability unaccounted for is inherent variability. There is no reason to assume that the components accounting for the most variability will be the best predictors of the response variable, and there are many examples in the literature where this is not so. This would suggest the use of *stepwise* principal components regression (see, for instance, [2]) but more effort has been directed towards some other methods.

Latent Root Regression

Latent root regression [5, 11] differs from principal components regression in that the response variable is included in the principal component analysis. Any component that has a zero characteristic root has a linear constraint among the variables, and the corresponding vector should indicate what that would be. Latent root regression is recommended for this purpose as well as to select variables which would make the best predictors. Reviews of these techniques are given by Gunst [4] and Mason [7].

Partial Least Squares Regression

The main criticism of principal components regression is that there is no guarantee that the larger components will be the best predictors, so some ad hoc scheme must be employed. For this reason, there has been considerable use in the last few years of a technique called *partial least squares regression* [12]. This technique is similar to principal components regression in that it produces a set of vectors for the predictor variables but does take the response variable into account. As each vector is obtained, it is immediately related to the response and the reduction in variability among the predictors. The estimation of the next vector takes that information into account. The very nature of partial least squares regression would indicate that it should do at least as well as principal components regression for the same number of retained components.

At the present time, partial least squares regression has been most widely used by analytical chemists, and most of the relevant information on applications is in the **chemometrics** literature. An algorithm for performing partial least squares regression may be found in [3]. Stone & Brooks [8] proposed a technique called *continuum regression* in which ordinary least squares, principal components regression, and partial least squares regression all fall out as special cases.

Multiple Responses

In addition to multiple predictors, there may also be multiple responses. Ordinary least squares and principal components regression treat each response as a separate regression problem, neither of them taking into account the relationships among the response variables. Partial least squares regression does take these relationships into account. As partial least squares regression sequentially establishes a set of vectors for the predictor variables, it simultaneously establishes a corresponding set of vectors for the response variables. For this reason, partial least squares regression has been referred to as "criss-cross" regression.

Maximum Redundancy

Being confronted with *sets* of both predictor and response variables might suggest the use of **canonical correlation**, a technique which obtains sets of vectors for each set of variables whose corresponding components will have maximum correlation. That technique will not produce an optimum prediction equation but an optimum solution can be obtained by a similar technique, *maximum redundancy* [9, 10], where *redundancy* is defined as the trace of the explained **covariance matrix** of the responses divided by the trace of the covariance matrix of the responses. Most of the applications of maximum redundancy have been in the fields of psychology and education.

Summary

A number of procedures have been suggested as alternatives to ordinary least squares to enhance the interpretation of the relationship between the predictor and response variables and/or to resolve problems associated with multicollinearity among the predictor variables. Although the most popular alternative, historically, has been principal components

regression, it does have some weaknesses and is being replaced by techniques such as partial least squares regression and maximum redundancy. Because partial least squares regression and maximum redundancy have been developed widely in different fields of application, it has not yet been established what the relative merits of them are. A unified treatment of the methods discussed in this article along with computational details may be found in [6].

References

[1] Box, G.E.P., Hunter, W.G., MacGregor, J.F. & Erjavac, J. (1973). Some problems associated with the analysis of multiresponse data, *Technometrics* **15**, 33–51.

[2] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[3] Geladi, P. & Kowalski, B. (1986). Partial least squares regression: a tutorial, *Analytica Chimica Acta* **185**, 1–17.

[4] Gunst, R.F. (1983). Latent root regression, in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 495–497.

[5] Hawkins, D.M. (1973). On the investigation of alternative regressions by principal component analysis, *Applied Statistics* **22**, 275–286.

[6] Jackson, J.E. (1991). *A User's Guide to Principal Components*. Wiley, New York.

[7] Mason, R.L. (1986). Latent root regression: a biased regression method for use with collinear prediction variables, *Communications in Statistics–Theory and Methods* **15**, 2651–2678.

[8] Stone, M. & Brooks, R.J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion), *Journal of the Royal Statistical Society, Series B* **52**, 237–269.

[9] Tyler, D.E. (1982). On the optimality of the simultaneous redundancy transformations, *Psychometrika* **47**, 77–86.

[10] Van den Wollenberg, A.L. (1977). Redundancy analysis: an alternative for canonical correlation analysis, *Psychometrika* **42**, 207–219.

[11] Webster, J.T., Gunst, R.F. & Mason, R.L. (1974). Latent root regression analysis, *Technometrics* **16**, 513–522.

[12] Wold, H. (1982). Soft modeling, in *Systems under Indirect Observation*, Vol. 11, K.G. Joreskog & H. Wold, eds. North-Holland, New York, pp. 1–54.

(*See also* **Battery Reduction**)

J. EDWARD JACKSON

Regression to the Mean

Regression to the mean (RTM) can broadly be described as the tendency of observations that are extreme by chance to move closer to the **mean** when repeated. The importance of this in biostatistics is that causality (*see* **Causation**) – rather than RTM – may erroneously be inferred when there is improvement after a treatment or change after an intervention.

Examples of RTM are common in clinical medicine. For example, a patient is noted to have a higher than average blood cholesterol on an initial screening exam. A repeat measurement will also likely be high but will on average be closer to normal even if the patient was not given medication. As Turner et al. [5] note in the context of pain problems, most acute and some chronic pain will resolve regardless of treatment. Since many individuals seek treatment or agree to enroll in trials when symptoms are more extreme, any change is likely to be an improvement, making any treatment appear effective – even if it is useless.

To define RTM further suppose we are interested in the ability of a drug to lower blood cholesterol and compare a patient's initial value x and posttreatment value y. Figure 1 is a hypothetical graph of x and y assuming a pretreatment mean μ. For simplicity, we assume that x and y are both **normally distributed** with common **variance** σ^2 and **correlation** ρ and that the drug has no effect.

A patient with a high pretreatment value, x, will have an expected posttreatment cholesterol level that is also high. The expected value, $\mathrm{E}(Y|x)$, will not be x but instead will be the lower value $\mathrm{E}(Y|x) = x - (1-\rho)(x-\mu)$. The difference, $(1-\rho)(x-\mu)$, is the effect of RTM, and this increases for both larger values of $x-\mu$ and for smaller values of ρ. In other words, the magnitude of the RTM varies directly with the variability of the process being observed and the correlation between the two measurements.

When there is a treatment effect, τ,

$$\mathrm{E}(Y - x|x) = \tau - (1-\rho)(x-\mu).$$

In this case the apparent treatment effect, $\mathrm{E}(Y-x|x)$ will be due in part to the actual treatment effect, τ, and to the relative size of both $1-\rho$ and $x-\mu$.

In **clinical trials**, regression to the mean can **bias** the estimate of the treatment effect. Suppose that we now conduct a trial in which a sample of n patients is taken from the population and compare the average posttreatment cholesterol values, $\overline{y}$, with the pretreatment values, $\overline{x}$. If the sample is a **simple random sample**, the difference $\overline{y} - \overline{x}$ is an **unbiased** estimate of τ because $\overline{x}$ will have expected value μ.

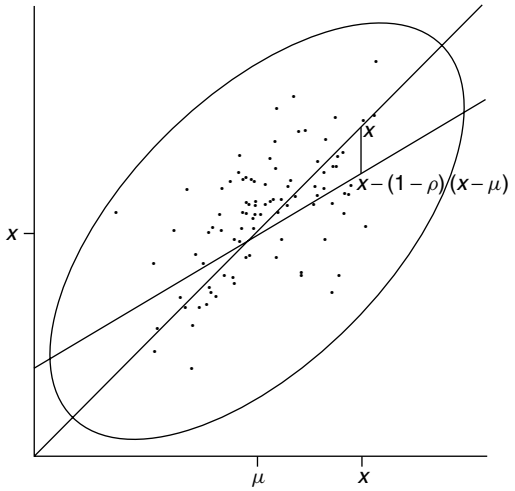

Figure 1 Hypothetical graph of posttreatment vs. blood cholesterol level pretreatment

However, most clinical trials are different in that we select from only a subset of the population – those with disease or, in this case, those with a cholesterol level above some cutoff, say c. In this case $\overline{y} - \overline{x}$ does not estimate the true treatment effect, but instead

$$\mathrm{E}(\overline{Y} - \overline{X}|X \geq c) = \tau - (1-\rho)\left(\frac{\sigma\phi\left(\frac{c-\mu}{\sigma}\right)}{1-\Phi\left(\frac{c-\mu}{\sigma}\right)}\right),$$

where ϕ and Φ are the density and cumulative density function for the normal distribution. Again, for extreme values of $c-u$ or small values of ρ the unadjusted treatment effect is biased. Results using this formula are presented in Table 1, which presents the effect of regression to the mean for various combinations of selection percentile, x, and correlation, ρ.

Suppose that patients are selected to be enrolled in a drug trial if their cholesterol levels are high, say above the 75th percentile. The average percentile for such patients is the 90th percentile. Assuming that pre- and posttreatment values have a correlation coefficient of 0.8, the posttreatment average will have dropped to the 85th percentile even in the absence

Table 1 Expected posttreatment percentile for a given pretreatment cutoff percentile and ρ (see text)

	Selection percentile			
ρ	50th	75th	90th	95th
0.0	0.50	0.50	0.50	0.50
0.2	0.56	0.60	0.64	0.66
0.4	0.63	0.69	0.76	0.80
0.6	0.68	0.78	0.85	0.89
0.8	0.74	0.85	0.92	0.95
1.0	0.79	0.90	0.96	0.98

of treatment effect. If $\rho = 0.6$, then the average will drop to the 78th percentile. If $\rho = 0$, then the expected posttreatment value will be at the 50th percentile for any pretreatment selection percentile. If $\rho = 1$, then there is no bias from RTM.

The term "placebo effect" has been defined by Turner et al. [5] as the nonspecific effects of treatment attributable to factors other than the active drug, including physician attention, patient expectations, changes in behavior, etc. Benefits from taking a placebo are often attributed to these factors, but RTM alone can produce such apparent benefits. Regression to the mean, then, is distinct from a "placebo effect". McDonald & Mazzuca [3] reviewed 30 randomly selected clinical trials in which the outcome was either a biologic, physiologic, or anatomic measurement. The authors noted that the improvement observed in placebo-treated patients and that of the estimate that would occur in biochemical variables due to regression to the mean were "remarkably similar".

Since RTM is a consequence of a correlation between two measurements, x and y, strategies consist of either eliminating the correlation or correcting it. McDonald & Mazzuca [3] suggest repeating the pretreatment measurement. In the cholesterol example, we would admit patients into the trial if x_1 were above the 90th percentile. A repeat pretreatment cholesterol, x_2, would be compared with a posttreatment y_1. Senn [4] shows that under restricted conditions, $\mathrm{E}(y_1 - x_2|x_1) = \tau$ is independent of x_1. A second possibility is to adjust for the correlation. When the underlying population from which the selected sample is drawn is large, Chen & Cox [1] suggest using the population to estimate the correlation and thereby adjust the treatment estimate. Hayes [2] examines various graphical methods. A third strategy is to conduct a randomized placebo-controlled clinical trial. This option will yield the most accurate results, since it involves no conditions or distributional assumptions.

In an experiment in which before and after measurements are made in order to evaluate some intervention, regression to the mean can account for some and possibly all of the estimated treatment effect. The extent of **confounding** will depend on the actual treatment effect, the sampling procedure used for patient selection, and the correlation between the two measurements. Methods of analysis and sampling techniques are available to adjust for regression to the mean, although direct estimation of treatment effects in trials is probably best.

References

[1] Chen, S. & Cox, C. (1992). Use of baseline data for estimation of treatment effects in the presence of regression to the mean, *Biometrics* **48**, 593–598.

[2] Hayes, R.J. (1988). Methods for assessing whether change depends on initial value, *Statistics in Medicine* **7**, 915–927.

[3] McDonald, C.J. & Mazzuca, S.A. (1983). How much of the placebo "effect" is really statistical regression?, *Statistics in Medicine* **2**, 417–427.

[4] Senn, S.J. (1988). How much of the placebo "effect" is really statistical regression?, *Statistics in Medicine* **7**, 1203.

[5] Turner, J.A., Deyo, R.A., Loeser, J.D., Von Korff, M. & Fordyce, W.E. (1994). The importance of placebo effects in pain treatment and research, *Journal of the American Medical Association* **271**, 1609–1614.

MICHAEL L. BEACH & JOHN BARON

Regression

The use of the term *regression* in statistics originated with **Francis Galton**, to describe a tendency to mediocrity in the offspring of parent seeds, and was used by **Karl Pearson** in a study of the heights of fathers and sons. The sons' heights tended on average to be less extreme than the fathers, demonstrating a so-called "**regression towards the mean**" effect (for details and a description of the most widely used form of regression analysis, *see* **Linear Regression, Simple**; **Multiple Linear Regression**). The term is now used in a wide variety of analysis techniques which examine the relationship between a **response variable** and a set of **explanatory variables**. The nature of the response variable usually determines the type of regression that is most natural (for a very general formulation of regression models and examples *see* **Generalized Linear Model**).

(*See also* **Correlation**; **Cox Regression Model**; **Logistic Regression**; **Poisson Regression**; **Proportional-odds Model**)

VERN T. FAREWELL

Regressive Models

Regressive models are designed for the analysis of correlated and naturally ordered data. They are **regression** models with **explanatory variables** including functions of preceding outcomes. Autoregressive models (*see* **ARMA and ARIMA Models**) are special cases with regression on the immediately preceding outcomes and are therefore conveniently indexed by the order, p say, indicating how far back the lagged values of the series itself should be regressed on. Thus, a **Markov process** is a first-order autoregressive model including only the regression on the first most immediately preceding outcome; a **Yule process** is second-order involving the two immediately preceding outcomes; and so on. Autoregressive models have been extensively studied in the context of long time, or spatial, series in which **stationarity**, seasonality, and the possibility of change points (*see* **Change-point Problem**) are of most interest. In the more general case of regressive models, the regression may involve just the first or all the preceding outcomes, as in a study of successive pregnancy outcomes, or the outcomes of natural links in branching structures, as, for example, parental disease status in human pedigree studies.

The emphasis in the present summary is on **likelihood** models for studying biological phenomena allowing for dependence among the outcomes. The nature of the measured outcome largely determines the measure of dependence used. Thus, for continuous outcomes, the **correlation** coefficient is natural, whereas for binary outcomes a regression coefficient that measures the change in the logarithm of the odds is a more natural measure.

Continuous Outcomes

The regressive model for n correlated and naturally ordered continuous outcomes $y_1, y_2, \ldots, y_n$ can be constructed following [6] using the **Gram–Schmidt** orthogonalization

$$\begin{aligned} z_1 &= y_1, \\ z_2 &= y_2 - b_{21}y_1, \\ z_3 &= y_3 - b_{31}y_1 - b_{32}y_2, \\ &\vdots \\ z_n &= y_n - b_{n1}y_1 - b_{n2}y_2 \cdots - b_{n,n-1}y_{n-1}, \end{aligned} \tag{1}$$

i.e. $\mathbf{z} = \mathbf{B}\mathbf{y}$ where $\mathbf{B}$ is a lower triangular matrix with 1s along the diagonal and is chosen so that the $\mathbf{z}$s are uncorrelated. Letting $\mathrm{var}(\mathbf{y}) = \mathbf{V}$, the covariance matrix of z is the diagonal matrix $\mathbf{W} = \mathbf{B}\mathbf{V}\mathbf{B}' = \mathbf{D}[w_i]$, w_i scalar. The Jacobian of the transformation is unity. Let the density function of a p-variate normal distribution (*see* **Multivariate Normal Distribution**) with mean zero and covariance matrix $\mathbf{\Sigma}$ be written as

$$\phi(\mathbf{t}, \mathbf{\Sigma}) = (2\pi)^{-p/2}|\mathbf{\Sigma}|^{-1/2} \exp\left(-\tfrac{1}{2}\mathbf{t}'\mathbf{\Sigma}^{-1}\mathbf{t}\right).$$

The density function of $\mathbf{y}$ can thus be written as a product of univariate normal densities:

$$\phi(\mathbf{y}, \mathbf{V}) = \phi(\mathbf{z}, \mathbf{W}) = \prod_{i=1}^{n} \phi(z_i, w_i). \tag{2}$$

In the language of **multiple linear regression**, for $i = 1, 2, \ldots, n$, z_i is y_i adjusted for $y_1, y_2, \ldots, y_{i-1}$; $b_{i1}, b_{i2}, \ldots, b_{i,i-1}$ are the partial regression coefficients; and w_i is the conditional variance of y_i given $y_1, y_2, \ldots, y_{i-1}$. Let $\mathbf{V}_{i-1} = (\sigma_{st})$ and $\mathbf{R}_{i-1} = (\rho_{st})$ denote the variance matrix and correlation matrix of $y_1, y_2, \ldots, y_{i-1}$, with respective inverses $\mathbf{V}_{i-1}^{-1} = (\sigma^{st})$ and $\mathbf{R}_{i-1}^{-1} = (\rho^{st})$. Then

$$\begin{aligned} b_{ij} &= \sum_{s=1}^{i-1} \sigma_{is}\sigma^{sj} = \left(\frac{\sigma_{ii}}{\sigma_{jj}}\right)^{1/2} \sum_{s=1}^{i-1} \rho_{is}\rho^{sj}, \\ w_i &= \sigma_{ii} - \sum_{j=1}^{i-1} \sigma_{ij}b_{ij} = \sigma_{ii}\left(1 - \sum_{j=1}^{i-1} \rho_{ij}b_{ij}^*\right) = \sigma_{ii}w_i^*, \end{aligned} \tag{3}$$

where

$$b_{ij}^* = \left(\frac{\sigma_{ji}}{\sigma_{ii}}\right)^{1/2} b_{ij} \quad \text{and} \quad w_i^* = \frac{1}{\sigma_{ii}}w_i.$$

The computation of b_{ij} and w_i becomes trivial once $\mathbf{R}_{i-1}^{-1}$ is computed, yielding explicit formulas for b_{ij}^* and w_i^*.

Consider the case in which $y_1, y_2, \ldots, y_{i-1}$ can be put into exactly two subgroups, A and B. Let the correlation of y_i and the elements of class A be η, the correlation of y_i and elements of class B be α, all the ys in A have the same correlation ρ, all those in B have the same correlation γ, and every y in A have a correlation of τ with every y in B. Using the notation $\mathbf{I}_i$ for an identity matrix of order n_i, $\mathbf{1}_i$ for a column

vector of n_i ones, and $\mathbf{J}_{ij}$ for an $n_i \times n_j$ matrix of ones, the correlation matrix of $y_1, y_2, \ldots, y_i$ is

$$\mathbf{R}_i = \begin{bmatrix} \mathbf{R}_{i-1} & | & \eta\mathbf{1}_1 \\ & | & \alpha\mathbf{1}_2 \\ \eta\mathbf{1}_1' \quad | \quad \alpha\mathbf{1}_2' & | & 1 \end{bmatrix}$$

$$= \begin{bmatrix} (1-\rho)\mathbf{I}_1 + \rho\mathbf{J}_{11} & | & \tau\mathbf{J}_{12} & | & \eta\mathbf{1}_1 \\ \tau\mathbf{J}_{21} & | & (1-\gamma)\mathbf{I}_2 + \gamma\mathbf{J}_{22} & | & \alpha\mathbf{1}_2 \\ \eta\mathbf{1}_1' & | & \alpha\mathbf{1}_2' & | & 1 \end{bmatrix}.$$

The inverse of $\mathbf{R}_{i-1}$ is

$$\mathbf{R}_{i-1}^{-1} = \begin{bmatrix} p_{11}\mathbf{I}_{11} + q_{11}\mathbf{J}_{11} & p_{12}\mathbf{J}_{12} \\ p_{21}\mathbf{J}_{21} & p_{22}\mathbf{I}_{22} + q_{22}\mathbf{J}_{22} \end{bmatrix},$$

where

$$p_{11} = \frac{1}{1-\rho}, \quad q_{11} = -\frac{\lambda_2\rho - n_2\tau^2}{(1-\rho)(\lambda_1\lambda_2 - n_1n_2\tau^2)}$$

$$p_{12} = p_{21} = -\frac{\tau}{\lambda_1\lambda_2 - n_1n_2\tau^2},$$

$$p_{22} = \frac{1}{1-\gamma}, \quad q_{22} = -\frac{\lambda_1\gamma - n_1\tau^2}{(1-\gamma)(\lambda_1\lambda_2 - n_1n_2\tau^2)},$$

$$\lambda_1 = 1 + (n_1 - 1)\rho \quad \text{and} \quad \lambda_2 = 1 + (n_2 - 1)\gamma.$$

Then

$$b_{ij}^* = \begin{cases} \dfrac{\lambda_2\eta - n_2\tau\alpha}{\lambda_1\lambda_2 - n_1n_2\tau^2}, \\ \qquad \text{if } y_j \text{ is in subgroup A} \\ \dfrac{\lambda_1\alpha - n_1\tau\eta}{\lambda_1\lambda_2 - n_1n_2\tau^2}, \\ \qquad \text{if } y_j \text{ is in subgroup B,} \end{cases}$$

and

$$w_i^* = 1 - \frac{n_1\eta(\lambda_2\eta - n_2\tau\alpha) + n_2\alpha(\lambda_1\alpha - n_1\tau\eta)}{\lambda_1\lambda_2 - n_1n_2\tau^2}.$$

Some special cases are as follows:

1. $n_1 = n_2 = 1, \eta = \alpha$. Then

$$b_{ij}^* = \frac{\eta}{1+\tau} \quad \text{and} \quad w_i^* = 1 - \frac{2\eta^2}{1+\tau}.$$

2. If $y_1, y_2, \ldots, y_i$ belong to class A and class B is empty, then $n_1 = i - 1, n_2 = 0, \gamma = 0, \eta = \rho$, so that

$$b_{ij}^* = \frac{\rho}{1 + (i-2)\rho} \quad \text{and}$$

$$w_i^* = 1 - \frac{(i-1)\rho^2}{1 + (i-2)\rho}.$$

3. If $\alpha = n_1\tau\eta/\gamma_1$, then

$$b_{ij}^* = \begin{cases} \dfrac{\eta}{\lambda_1}, & \text{if } y_j \text{ is in class A,} \\ 0, & \text{if } y_j \text{ is in class B,} \end{cases}$$

and

$$w_i^* = 1 - \frac{n_1\eta^2}{\lambda_1}.$$

The orthogonal variate z_i in (1) is then a linear combination of y_i and the ys for class A members only. Here, class A may be taken as all those measurements immediately preceding the ith measurement. The results are then relevant to the study of autoregressive models. For example, $n_1 = 1$ and $\tau = \eta = \rho$ give $\alpha = \rho^2$, the well-known condition for a Markov type of dependence. If $n_1 = 2$, we have a Yule type of dependence.

Regressive Logistic Models for Correlated Binary Outcomes

In the case of **binary** outcomes $Y_j = 1, 0$ with associated covariates $\mathbf{X}_j$, one can consider the regression on the preceding outcome $Y_{j-1}^* = 2Y_j - 1$, or the first outcome (Y_1^*), and on the sum of preceding positives $S_{j-1}^+ = \sum_{s=1}^{j-1} Y_s$, preceding negatives $S_{j-1}^- = \sum_{s=1}^{j-1}(1 - Y_s)$, and the other covariates $\mathbf{X}_j$. The model is defined by a **logistic** function; thus

$$\theta_j = \log[\Pr(Y_j = 1|Y_1, \ldots, Y_{j-1}, \mathbf{X}_j)/ \Pr(Y_j = 0|Y_1, \ldots, Y_{j-1}, \mathbf{X}_j)]$$
$$= \gamma_0 + \gamma_1 Y_1^* + \gamma_2 Y_2^* + \cdots + \gamma_{j-1} Y_{j-1}^* + \beta X_j,$$

where for convenience we have included only one component of $\mathbf{X}$. To postulate parsimonious versions so that only a few γs need to be estimated, consider the following model introduced by Bonney [7]:

$$\gamma_t = \begin{cases} \gamma^+ + \gamma, & \text{if } t = j-1, Y_{j-1}^* = 1, \\ \gamma^- + \gamma, & \text{if } t = j-1, Y_{j-1}^* = -1, \\ \gamma^+, & \text{if } t < j-1, Y_{j-1}^* = 1, \\ \gamma^-, & \text{if } t < j-1, Y_{j-1}^* = -1. \end{cases}$$

This parameterization allows the immediately preceding outcome to increase or decrease the logarithm of the odds of the current outcome by an amount γ

more than an increase or decrease from a more remote outcome. The logit of the current outcome becomes

$$\theta_j = \gamma_0 + \gamma Y^*_{j-1} + \gamma^+ S^+_{j-1} + \gamma^-(-S^-_{j-1}) + \beta X_j.$$

The modeling problem is therefore reduced to that of regression on the immediately preceding outcome, the number of preceding positives (1s), the number of preceding negatives (0s), and other covariates. There is no problem replacing Y^*_{j-1} by Y^*_1 so that the first outcome has a different effect on the current outcome. Some old approaches are special cases. Thus, for Markov dependence of order 1 as in [14, 15], and [18], $\gamma^+ = \gamma^- = 0$. The suggestion in [14] of regressing on the cumulative sum of preceding successes corresponds to $\gamma^- = 0$. Therefore we have a sufficiently general model against which the simpler models may be tested. It incorporates the natural explanatory variables one would consider in analyzing a sequence of binary outcomes. If it fits the data, then we have accounted for covariates of interest as well as the dependence in the sequence of observations in terms of a model that is easy to fit and interpret. Moreover,

$$S^+_{j-1} + S^-_{j-1} = j - 1,$$

so S^+_{j-1}, S^-_{j-1}, and the serial order j are perfectly **collinear**. Hence, any two of them are sufficient in the regression analysis. If, on the other hand, only one of them is included, then we have omitted a critical covariate that may show up as residual dependence. Table 1 displays an example.

Furthermore, in the case of equally spaced outcomes, regression on time is the same as regression on the serial order. In unequally spaced studies, actual time may replace serial order, as is commonly done. However, it is likely that significant collinearity with S^+_{j-1} and S^-_{j-1} may remain. Moreover, many unequally spaced serial observations are made at convenient periods, e.g. baseline, six months, 12 months, and 24 months. In such cases a few **dummy variables** describing the periods may be better than the more common linear function of time.

The results for binary outcomes generalize naturally to **polytomous** outcomes [2, 9].

With regard to the performance of the regressive models, a large computer **simulation** study [13] showed the following:

1. If the regression on preceding outcomes is ignored as in standard **logistic regression**, then the bias in the estimation of the regression coefficient of a covariate is increased by at least 12% in samples of sizes 200–500 and dependent groups of size two, and can be as bad as 25% in groups of size five if the true state of nature is Markov of order 1, and 30%–60% in the equally predictive case. In samples of sizes greater than 500, the bias is negligible only if the regression

Table 1 Frequencies of spontaneous abortion[a]

Number of pregnancies	Sequence of outcomes[b]	Sample		
		MRH	Leridon	Roman
1	0	1435	948	2651
	1	201	103	417
2	00	1238	752	1914
	01	197	72	261
	10	156	62	295
	11	45	21	53
3	000	827	590	853
	001	176	64	129
	010	128	42	188
	011	31	8	34
	100	100	44	216
	101	20	8	30
	110	27	12	29
	111	8	5	14
4	0000	405	466	240
	0001	57	40	41
	0010	68	45	64
	0011	30	9	18
	0100	85	28	65
	0101	18	11	26
	1000	19	31	103
	1001	6	6	18
	0110	19	8	23
	0111	6	0	5
	1010	15	6	24
	1011	2	0	1
	1100	13	8	15
	1101	9	3	7
	1110	5	1	5
	1111	2	3	5

[a]Extracted from [22].

[b]0 = live birth; 1 = spontaneous abortion.

Equivalent models [7]: $\theta_i = 5.9157 + 1.5408 S^+_{i-1} - 0.0835 S^-_{i-1} + 0.5322 X_2 - 0.3124 X_3$, and $\theta_i = -6.6443 + 0.8121 S_{i-1} + 0.7285 X_1 + 0.5322 X_2 - 0.3123 X_3$, where S^+_{i-1} is the number of spontaneous previous abortions, S^-_{i-1} is the number of previous live births, X_1 is the serial pregnancy number, $X_2 = 1$, $X_3 = 0$ for Leridon, $X_2 = 0$, $X_3 = 1$ for Roman, $X_2 = 0$, $X_3 = 0$ for MRH, and $S_{i-1} = S^+_{i-1} + S^-_{i-1}$.

coefficient of the covariate and the dependence are both positive or both negative.
2. Modeling the dependence leads to more robust estimates of regression parameters, with the equally predictive model being generally more robust than the Markov model.
3. Caution should be exercised when the sample size is small because, while the regressive model can correct the bias of desired regression estimates, the additional parameters can increase problems of noncovergence, infinite variances, and nonunique estimates. Thus, for a sample size of 50 and group size two, the mean square error associated with the regressive model that fits an extra parameter can often be larger than that of the standard logistic.

Regressive Models in Family Studies

The design of the regressive models for family data [3, 4, 7, 8] was based on the following observations on pedigrees (see Figure 1):

1. Dependence among relatives arises from the biologic relationships, whether genetic or environmental in origin.
2. Pedigrees are structures evolving in time: grandparents appear before parents, who in turn appear before grandchildren, and so on. The time ordering along vertical lines of descent is evolutionary.
3. A pedigree, however complex, is made up of distinct sibships joined through the common parents. Thus, it is natural to consider models in which different sibships are conditionally independent given the intervening parents.

In view of these observations, the regressive models account for patterns of dependence in family data by specifying a regression relationship between a person's phenotype (Y), the phenotypes of ancestors and older relatives ($\mathbf{Y}_A$), the **genotype** (g), if any, at some postulated genetic loci, and other explanatory variables ($\mathbf{X}$). In essence the classical **penetrance** function is generalized to include ancestral phenotypes. The generalized penetrance function is the conditional probability of Y given g, $\mathbf{Y}_A$, and $\mathbf{X}$, i.e.

$$\Pr(Y|g, \mathbf{Y}_A, \mathbf{X}).$$

For a continuous trait, this is replaced by the probability density function. Including $\mathbf{Y}_A$ in the regression allows for unspecified (or residual) factors, such as spouse correlations, unspecified genes, and cultural and other environmental effects. Broad patterns of dependence are possible with $\mathbf{Y}_A$ including only a few relatives, which simplifies the calculations greatly. In the class A pattern of dependence parents alone account for residual correlations among sibs, so that $\mathbf{Y}_A$ includes only the phenotypes of mother ($\mathbf{Y}_M$), father ($\mathbf{Y}_F$), and spouse ($\mathbf{Y}_S$) if spouse correlations are not zero; class B adds the oldest sibs, class C adds the immediately preceding sibs, and class D includes all preceding sibs. An example of the penetrance function is given in Figure 2.

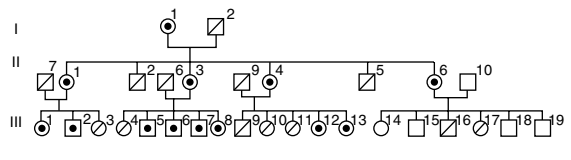

Figure 1 A three generational human family or pedigree

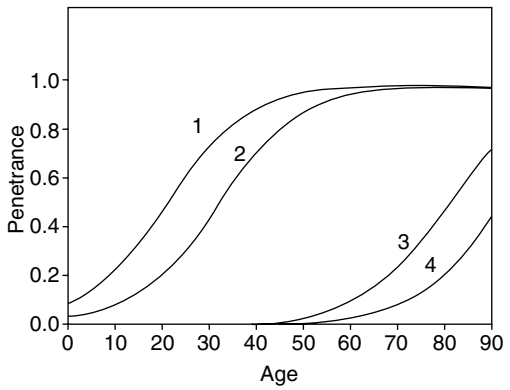

Figure 2 Estimated penetrance functions for a single gene in a family study of chronic atrophic gastritis [5]: 1. homozygous recessive (AA), mother affected; 2. homozygous recessive (AA), mother unaffected; 3. carrier (AB) or noncarrier (BB), mother affected; 4. carrier (AB) or noncarrier (BB), mother unaffected, where penetrance $= [1 + \exp[\theta(g)]]^{-1}$, $\theta(\mathrm{AA}) = -3.33 + 0.65Z_M + 0.12X$ for AA persons, $\theta(\mathrm{AB/A'B}) = -10.25 + 0.65Z_M + 0.12X$ for AB or BB persons, $Z_M = 1$ if mother is affected and $Z_M = -1$ if mother is unaffected, and $X =$ age at examination. Reproduced from Bonney et al. [5] by permission of John Wiley & Sons Ltd

Using Correlations to Describe Familial Dependence [6]

Suppose that the vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ comprises n measurements on members of a nuclear family: father, mother, and $n-2$ offspring (sibs), in that order. We use the following notation:

μ_i = mean of x_i,
σ_i^2 = variance of x_i,
ρ_{FM} = father–mother correlation,
ρ_{PO} = parent–offspring correlation, and
ρ_{SS} = sib–sib correlation.

Usually the means, μ_i, and sometimes the variances, σ_i^2, are modeled as linear functions of explanatory variables including major genotypes, if any are postulated. The orthogonalization process (1) yields

$$z_1 = x_1 - \mu_1, \quad \text{and}$$

$$z_i = (x_i - \mu_i) - \sum_{j=1}^{i-1} b_{ij}(x_j - \mu_j)$$

$$= (x_i - \mu_i) - \sum_{j=1}^{i-1} \frac{\sigma_i}{\sigma_j} b_{ij}^*(x_j - \mu_j), \; i = 2, \ldots, n,$$

and $\text{var}(z_i) = w_i = \sigma_i^2 w_i^*$, where

$$w_i^* = 1,$$

$$b_{21}^* = \rho_{\text{FM}}, \qquad w_2^* = 1 - \rho_{\text{FM}}^2,$$

$$b_{31}^* = b_{32}^* = \frac{\rho_{\text{PO}}}{1+\rho_{\text{FM}}}, \qquad w_3^* = 1 - \frac{2\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}}.$$

For $i > 3$, we are adjusting x_i for xs from individuals who could be either parents or sibs, i.e. we have a two-class case. Letting $n_1 = 2$, $n_2 = i-3$, $\rho = \rho_{\text{FM}}$, $\gamma = \alpha = \rho_{\text{SS}}$, and $\tau = \eta = \rho_{\text{PO}}$, we obtain:

$$b_{ij}^* = \begin{cases} \dfrac{\rho_{\text{PO}}}{1+\rho_{\text{FM}}}\left[1 + \dfrac{i-3}{1-\rho_{\text{SS}}}\left(\rho_{\text{SS}} - \dfrac{2\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}}\right)\right]^{-1}, \\ \quad \text{if } j \text{ is parent}, \\ \dfrac{1}{1-\rho_{\text{SS}}}\left(\rho_{\text{SS}} - \dfrac{2\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}}\right) \\ \times \left[1 + \dfrac{i-3}{1-\rho_{\text{SS}}}\left(\rho_{\text{SS}} - \dfrac{\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}}\right)\right]^{-1}, \\ \quad \text{if } j \text{ is a sib}, \end{cases}$$

and

$$w_i^* = 1 - \left[\frac{2\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}} + \frac{(i-3)\rho_{\text{SS}}}{1-\rho_{\text{SS}}}\left(\rho_{\text{SS}} - \frac{2\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}}\right)\right] \times \left[1 + \frac{i-3}{1-\rho_{\text{SS}}}\left(\rho_{\text{SS}} - \frac{2\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}}\right)\right]^{-1}.$$

For a class A regressive model, Bonney [3] gave the correlation among sibs due to common parentage as $2\rho_{\text{PO}}^2/(1+\rho_{\text{FM}})$, so that in the absence of other sources of correlation, $\rho_{\text{SS}} = 2\rho_{\text{PO}}^2/(1+\rho_{\text{FM}})$. Then the formulas reduce to

$$b_{ij}^* = \begin{cases} \dfrac{\rho_{\text{PO}}}{1+\rho_{\text{FM}}}, & \text{if } j \text{ is a parent}, \\ 0, & \text{if } j \text{ is a sib}, \end{cases}$$

and

$$w_i^* = 1 - \frac{2\rho_{\text{PO}}^2}{1+\rho_{\text{FM}}}.$$

Thus, the orthogonal variate z_i in (1) for a sib depends only on the ys for the parents. See [3] for the formulas for classes B, C, and D.

Using Variance Components to Describe Dependence

A popular example of the use of **variance components** with human family data was developed by Morton & MacLean [21], who proposed a model for polygenic inheritance in which variances (σ^2) are partitioned into $\sigma^2 = \sigma_{\text{a}}^2 + \sigma_{\text{e}}^2$ for parents and $\sigma^2 = \sigma_{\text{a}}^2 + \sigma_{\text{c}}^2 + \sigma_{\text{r}}^2$ for sibs, where

σ_{a}^2 = variance due to additive polygenic factors,
σ_{e}^2 = variance due to environmental factors,
σ_{c}^2 = variance due to common sibling environmental factors, and
σ_{r}^2 = variance due to random environmental factors on sibs such that $\sigma_{\text{e}}^2 = \sigma_{\text{c}}^2 + \sigma_{\text{r}}^2$.

Boyle & Elston [12] reviewed and extended the model to include other sources of variation. For this model of inheritance, the parent–offspring correlation is $\rho_{\text{PO}} = \frac{1}{2}\sigma_a^2/\sigma^2$ and the sib–sib correlation is $\rho_{\text{SS}} = (\frac{1}{2}\sigma_{\text{a}}^2 + \sigma_{\text{c}}^2)/\sigma^2$. Morton & MacLean assumed zero spouse correlations ($\rho_{\text{FM}} = 0$) and $\sigma_i^2 = \sigma^2$ for all i. Thus, in terms of the regressive models (1), the variance components model leads to

$$z_1 = x_1 - \mu_1, \, w_1^* = 1,$$

$$z_2 = x_2 - \mu_2, \, w_2^* = 1,$$

and

$$z_1 = (x_i - \mu_1) - \sum_{j=1}^{i-1} b_{ij}^*(x_j - \mu_j), \quad i \geq 3,$$

where

$$b_{ij}^* = \begin{cases} \left[\dfrac{\sigma^2}{\frac{1}{2}\sigma_a^2} + (i-3)\left(\dfrac{\sigma_c^2+\sigma_r^2}{\frac{1}{2}\sigma_a^2+\sigma_r^2} + \dfrac{\sigma^2\sigma_c^2}{\frac{1}{2}\sigma_a^2\left(\frac{1}{2}\sigma_a^2+\sigma_r^2\right)}\right)\right]^{-1}, \\ \quad \text{if } j \text{ is a parent,} \\ \left[(i-3) + \dfrac{\sigma^2\left(\frac{1}{2}\sigma_a^2+\sigma_r^2\right)}{\sigma^2\sigma_c^2 + \frac{1}{2}\sigma_a^2\left(\sigma_c^2+\sigma_r^2\right)}\right], \\ \quad \text{if } j \text{ is a sib,} \end{cases}$$

and

$$w_i^* = 1 - \left[\frac{1}{2}\left(\frac{\sigma_a^2}{\sigma^2}\right)^2 + (i-3) \times \frac{\frac{1}{2}\sigma_a^2+\sigma_c^2}{\frac{1}{2}\sigma_a^2+\sigma_r^2}\frac{\frac{1}{2}\sigma_a^2(\sigma_c^2+\sigma_r^2)+\sigma^2\sigma_c^2}{(\sigma^2)^2}\right] \times \left[1 + (i-3)\frac{\frac{1}{2}\sigma_a^2(\sigma_c^2+\sigma_r^2)+\sigma^2\sigma_c^2}{\sigma^2\left(\frac{1}{2}\sigma_a^2+\sigma_r^2\right)}\right]^{-1},$$

if $i \geq 3$.

Using these in (2) leads to a **likelihood** function which is a product of univariate normal densities.

Scope of the Regressive Models. Table 2 summarizes the scope of the regressive models for family studies and the relevant references for further reading.

Table 2 Scope of regressive models regression setup

Dependent variable		Explanatory variables		
	Phenotype Y	Genotype (specific) or major g	Ancestral phenotypes Y_A	Other covariates X
1. No causal scheme	Y		Y_A	X
2. Polygenic causal scheme	Y		Y_A	X
3. Genetic association (measured genotype)	Y	g	Y_A	X
4. Segregation analysis	Y	g		
Oligogenic complete penetrance	Y	g		
Incomplete penetrance	Y	g		X
More general (includes mixed models)	Y	g	Y_A	X
5. Linkage	Y	g (two or more loci)	Y_A	X
One trait One or more markers				
Two or more traits	**Y** (vector)	g (two or more loci)		
6. Pleiotropy	**Y** (vector)	g (one locus)	Y_A	X

Some pertinent publications: continuous traits [3, 10]; binary traits [4]; variable age-of-onset in familial disease [1, 17]; liability models [16]; path analytic models [19, 20]; genetic linkage analysis [8, 11] (*see* **Linkage Analysis, Model-based**).

References

[1] Able, L. & Bonney, G.E. (1990). A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases, *Genetic Epidemiology* **7**, 391–407.

[2] Amfoh, K., Shaw, R. & Bonney, G. (1994). The use of logistic models for the analysis of codon frequencies of DNA sequences in terms of explanatory variables, *Biometrics* **50**, 1054–1063.

[3] Bonney, G.E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: regressive models, *American Journal of Medical Genetics* **18**, 731–749.

[4] Bonney, G.E. (1986). Regressive logistic models for familial disease and other binary traits, *Biometrics* **42**, 611–625.

[5] Bonney, G.E., Elston, R.C., Correa, P., Haenszel, W., Zavala, D.E., Zarama, G., Collazos, T. & Cuello, C. (1986). Genetic etiology of gastric carcinoma: I. Chronic atrophic gastritis, *Genetic Epidemiology* **3**, 213–224.

[6] Bonney, G.E. & Kissling, G.E. (1986). Gram-Schmidt orthogonalization of multinormal variates: applications in genetics, *Biometrical Journal* **28**, 417–425.

[7] Bonney, G.E. (1987). Logistic regression of dependent observations, *Biometrics* **43**, 951–973.

[8] Bonney, G.E., Lathrop, M. & Lalouel, J.M. (1988). Combined linkage and segregation analysis using regressive models, *American Journal of Human Genetics* **43**, 24–37.

[9] Bonney, G.E., Dunston, G. & Wilson, J. (1989). Regressive logistic models for ordered and unordered polychotomous traits: application to affective disorders, *Genetic Epidemiology* **6**, 211–215.

[10] Bonney, G.E. (1992). Compound regressive models for family data, *Human Heredity* **42**, 28–41.

[11] Borecki, I., Bonney, G.E., Rice, T., Bouchard, C. & Rao, D.C. (1993). Influence of genotype-dependent effects of covariates on the outcome of segregation analysis of the body mass index, *American Journal of Human Genetics* **53**, 676–687.

[12] Boyle, C.R. & Elston, R.C. (1979). Multifactorial genetic models for quantitative traits in humans, *Biometrics* **35**, 55–68.

[13] Brooks, C.A. & Bonney, G.E. (1989). A simulation study of properties of a regressive logistic model, *Journal of Statistical Computation and Simulation* **32**, 31–43.

[14] Cox, D.R. (1958). The regression analysis of binary sequences (with discussion), *Journal of the Royal Statistical Society, Series B* **20**, 215–242.

[15] Cox, D.R. (1970). *The Analysis of Binary Data*. Methuen, London.

[16] Demenais, F.M. (1991). Regressive logistic models for complex familial diseases: A formulation assuming an underlying liability model, *American Journal of Human Genetics* **46**, 773–785.

[17] Elston, R.C. & George, V.T. (1989). Age of onset, age at examination, and other covariates in the analysis of family data, *Genetic Epidemiology* **6**, 217–220.

[18] Haldane, J.B.S. & Smith, C.A.B. (1948). A simple exact test for birth order effect, *Annals of Eugenics* **14**, 117–124.

[19] Li, Z., Bonney, G.E. & Rao, D.C. (1994). Genetic analysis combining path analysis with regressive models: the BETA path model of polygenic and familial environmental transmission, *Genetic Epidemiology* **11**, 431–442.

[20] Li, Z., Bonney, G.E., Lathrop, G.M. & Rao, D.C. (1994). Genetic analysis combining path analysis with regressive models: the TAU model of multifactorial transmission, *Human Heredity* **44**, 305–311.

[21] Morton, N.E. & MacLean, C.J. (1974). Analysis of family resemblances. III. Complex segregation of quantitative traits, *American Journal of Human Genetics* **26**, 489–503.

[22] Wilcox, A.J. & Gladen, B.C. (1982). Spontaneous abortion: the role of heterogeneous risk and selective fertility, *Early Human Development* **7**, 165–178.

(*See also* **Correlated Binary Data**; **Familial Correlations**)

GEORGE E. BONNEY

Relationship Testing

Methods of genetic analysis generally assume relationships are known without error within families. Relationship misclassification due to factors such as false paternity, unknown adoption, and sample switches or duplications can compromise genetic analyses. Therefore, methods have been developed to detect misspecified relationships within a family and infer true relationships when putative ones are incorrect.

Depending on the pedigree structure, analysts often can detect relationship misspecification by observing Mendelian inconsistencies within a family: for example, a parent–offspring pair failing to share an allele (*see* **Gene**), or a sibship exhibiting more than four alleles. Identifying such inconsistencies often results from calculating a zero **likelihood** for pedigree data, or visually detecting logical inconsistencies in pedigree drawings. Algorithms also exist for detecting Mendelian inconsistencies and identifying the individual(s) most likely responsible for these errors [13, 15, 18]. Given smaller family subsets such as sibling pairs, it may not be possible to identify misspecified relationships with certainty. Still, with sufficient marker data, they may be identified probabilistically.

Many relationship inference methods analyze different pairs of relatives. In what follows, we focus on these relative-pair-based methods. These methods can be divided into three general categories: likelihood-based methods, expected allele-sharing methods, and continuous-data methods. We compare and contrast these categories of methods. All three categories distinguish different relationships using the pattern of alleles shared by the relative pair based on the principle that closely related pairs in general share more alleles than distantly related pairs.

Assumptions and Notation

We assume a noninbred relative pair is typed for a collection of M codominant **markers**. Assume marker k has n_k alleles with population frequencies $q_1, q_2, \ldots, q_{n_k}$. Let θ_k be the recombination fraction between markers k and $k+1$ $(1 \le k \le M-1)$. We assume θ_k is known without error and is the same for both sexes. If ***X*-linked** data are included, then we assume the sex of each individual is known. Let X_k be the pair of genotypes at marker k, and $X = (X_1, X_2, \ldots, X_M)$ be all the genotype data for the pair. Let ibd_k and ibs_k denote the number of alleles shared identical by descent (ibd) and identical by state (ibs), respectively, by the pair at marker k. Finally, let R_{Put} and R denote the putative relationship and true relationship for a relative pair (*see* **Linkage Analysis, Model-free** for more detail on ibd and ibs).

Likelihood-Based Methods for Relationship Inference

Likelihood-based methods for relationship inference are based on work by Thompson [19], who evaluated the probability, $\Pr(X_k|R)$, of a pair's data at marker k conditional on the pair's relationship. Using Bayes' rule,

$$\Pr(X_k|R) = \sum_{i=0}^{2} \Pr(X_k|\text{ibd}_k = i)\Pr(\text{ibd}_k = i|R). \tag{1}$$

$\Pr(X_k|\text{ibd}_k = i)$ is the conditional probability of the data X_k at marker k, given the pair shares i alleles ibd at that marker. This probability is a simple function of ibd_k and the allele frequencies $q_1, q_2, \ldots, q_{n_k}$, and is independent of the relationship R. These probabilities were evaluated for an autosomal marker by Thompson [19] and for an X-linked marker by Epstein et al. [8].

$\Pr(\text{ibd}_k = i|R)$ is the probability a pair of relationship R shares i alleles ibd at marker k; it is a simple function of R. For example, for an autosomal marker and $i = (0, 1, 2)$,

$$\Pr(\text{ibd}_k = i|R = \text{full sibs}) = \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\right)$$

and

$$\Pr(\text{ibd}_k = i|R = \text{parent–offspring}) = (0, 1, 0).$$

Similar probabilities may be calculated for other relationships and for X-linked data.

The joint probability for M unlinked markers is then [19]

$$\Pr(X|R) = \prod_{k=1}^{M} \Pr(X_k|R)$$

$$= \prod_{k=1}^{M} \left[\sum_{i=0}^{2} \Pr(X_k | \text{ibd}_k = i) \Pr(\text{ibd}_k = i | R) \right]. \quad (2)$$

One can determine the most likely relationship of a pair by evaluating $\Pr(X|R)$ under different relationships R and inferring the relationship that maximizes the joint probability (2). One can also obtain **maximum likelihood** estimates of $\Pr(\text{ibd}_k = i|R)$ for $i = 0, 1, 2$, and infer the relationship that is most consistent with these estimates [12, 19].

Equation (2) only applies to unlinked markers, which limits marker number and reduces the ability to distinguish different relationships. Göring & Ott [10] and Boehnke & Cox [2] independently extended the work of Thompson [19] to allow for linked markers. In particular, they noted that under the assumption of no genetic interference, $\text{ibd}_1, \text{ibd}_2, \ldots, \text{ibd}_M$ represent a **hidden Markov** chain for many relationships R. They then used Baum's algorithms [1] to evaluate $\Pr(X|R)$. Let $\alpha_k(i|R) = \Pr(X_1, X_2, \ldots, X_{k-1}, \text{ibd}_k = i|R)$ be the joint probability of the marker data at the first $k-1$ markers and that the pair shares i alleles ibd at marker k given the pair's relationship is R. For the first marker, $\alpha_1(i|R) = \Pr(\text{ibd}_1 = i|R)$, which is evaluated as before. For subsequent markers,

$$\alpha_{k+1}(j|R) = \sum_{i=0}^{2} \alpha_k(i|R) \Pr(X_k|\text{ibd}_k = i) \times \Pr(\text{ibd}_{k+1} = j|\text{ibd}_k = i; R). \quad (3)$$

$\Pr(\text{ibd}_{k+1} = j|\text{ibd}_k = i; R)$ is the transition probability that a pair of relationship R shares j alleles ibd at marker $k+1$ given they share i alleles ibd at marker k. These probabilities depend on θ_k and R. The transition probabilities for different relationships were presented for autosomal data by Risch [16] and for X-linked data by Epstein et al. [8]. One obtains $P(X|R)$ by the final summation

$$\Pr(X|R) = \sum_{i=0}^{2} \alpha_M(i|R) \Pr(X_M|\text{ibd}_M = i). \quad (4)$$

Boehnke & Cox [2] calculated $\Pr(X|R)$ under four different relationships [monozygotic (MZ) twin, full sib, half sib, and unrelated] and inferred the relationship that maximized this probability. Göring & Ott [10] assumed prior probabilities for full sibs, half sibs, and unrelated pairs in the study population and calculated the posterior probability, $\Pr(R|X)$, of each of these relationships using Bayes' rule.

McPeek & Sun [12] and Epstein et al. [8] extended the method of Göring & Ott [10] and Boehnke & Cox [2] to test additional relationships including parent–offspring, grandparent–grandchild, avuncular (e.g. aunt–niece) and first cousins. McPeek & Sun [12] evaluated $\Pr(X|R)$ and then used a **likelihood ratio** statistic to test the **null hypothesis** that the putative relationship of a pair is correctly specified against the **alternative hypothesis** that the putative relationship is misspecified. Under the alternative, they maximized the **likelihood** as a function of the probability from among the other tested relationships. Their likelihood ratio statistic is skewed, so they estimated significance by **simulation**. For more distant relationships such as avuncular and first cousins, McPeek & Sun [12] noted that $\{\text{ibd}_k\}$ no longer form a Markov chain [9], which complicates the likelihood calculation. They remedied this problem by calculating the likelihood for these more distant relationships under an augmented ibd Markov chain.

Epstein et al. [8] calculated $\Pr(X|R)$ for a variety of relationships and inferred the relationship that maximized this probability. Unlike McPeek & Sun [12], they chose to approximate $\Pr(X|R)$ for avuncular and first-cousin relationships by incorrectly assuming $\{\text{ibd}_k\}$ are Markovian. This approximate likelihood is an adequate substitute for the true likelihood in the inference of avuncular pairs and first cousins [12].

Broman & Weber [3] and Epstein et al. [8] extended these likelihood-based methods to allow for genotyping error. Failure to allow for genotyping error has only a modest impact for many relationships but often leads to incorrect classification of MZ twins and parent–offspring pairs. To model genotyping error, they let ε denote the probability that a genotype is chosen at random according to population genotype frequencies and let $1-\varepsilon$ be the probability that genotyping is done correctly with certainty. For this genotyping-error model, the only component altered is $\Pr(X_k|\text{ibd}_k = i)$. If either member of the pair is correctly genotyped for marker k, then $\Pr(X_k|\text{ibd}_k = i)$ remains the same. If either member is randomly genotyped, then the pair is effectively unrelated. Hence,

$$\Pr(X_k|\text{ibd}_k = i; \varepsilon)$$

$$= (1-\varepsilon)^2 \Pr(X_k | \text{ibd}_k = i; \varepsilon = 0) + \left[1 - (1-\varepsilon)^2\right] \Pr(X_k | \text{ibd}_k = 0; \varepsilon = 0). \quad (5)$$

While this random-genotype model is not a true representation of how genotype error occurs, it is computationally simple and can detect errors generated by more realistic error mechanisms [6].

Expected Allele-Sharing Methods for Relationship Inference

In contrast to likelihood-based methods, allele-sharing methods directly compare the observed allele sharing of a relative pair with that expected given the pair's putative relationship. These methods originated with Chakraborty & Jin [5]. Ehm & Wagner [7] and McPeek & Sun [12] proposed ibs-based test statistics to detect misspecified relationships based on a set of linked autosomal markers. Stivers et al. [17] derived an analogous ibs-based test statistic limited to unlinked markers.

Ehm & Wagner [7] only tested pairs of putative full sibs, although one could easily extend their method to test other putative relationships. They based their test on half the total number of alleles shared ibs by a pair across a series of M autosomal markers:

$$S_{\text{EW}} = \frac{1}{2} \sum_{k=1}^{M} \text{ibs}_k.$$

The authors standardized S_{EW} by subtracting the mean of S_{EW} and dividing by its standard deviation, where they calculated both the mean and variance assuming the putative relationship of full sibs.

McPeek & Sun [12] described essentially the same statistic generalized to any arbitrary putative relationship. Their test statistic relied on the average number of alleles shared ibs by the pair across a series of markers:

$$S_{\text{MS}} = \frac{1}{M} \sum_{k=1}^{M} \text{ibs}_k.$$

They also standardized the sum, again based on the putative relationship for the pair. For large M, the resulting standardized statistics Z_{EW} and Z_{MS} are approximately distributed as standard normal if the putative relationship is true.

Since ibd-based methods often are more powerful than ibs-based ones, McPeek & Sun [12] derived an analogous ibd-based allele-sharing statistic. This statistic, which they called the expected identical by descent (Eibd) statistic, is the average estimated number of alleles shared ibd by a pair at a series of markers, conditional on the marker genotypes and the putative relationship; it takes the form

$$\text{Eibd} = \frac{1}{M} \sum_{k=1}^{M} \text{E}(\text{ibd}_k | X_k; R_{\text{Put}}).$$

Eibd is approximately distributed as normal if the putative relationship is true.

McPeek & Sun [12] also created an ibd-based method that is conditional on ibs sharing. This statistic, which they called the adjusted ibs-sharing statistic (Aibs), has the form

$$\text{Aibs} = \frac{1}{M} \sum_{k=1}^{M} \text{Aibs}_k.$$

For marker k, the author calculated Aibs_k by summing over all four possible draws of an allele from each member of the relative pair. For a given draw, the authors evaluated the probability that the two alleles drawn from the pair are shared ibd given they are shared ibs and the putative relationship of the pair. Aibs approximately follows a normal distribution under the putative relationship.

For putative full sibs, Olson [14] presented an ibd-sharing statistic that infers the most likely relationship of a pair when the putative relationship is rejected. The method has some similarities with the Eibd statistic since both calculate the expected number of alleles shared ibd by a relative pair. However, while one calculates Eibd at a series of markers, Olson [14] considered any location along the genome. Let $\widehat{\text{ibd}}_s$ denote the estimated number of alleles shared ibd by the pair at some genomic location s, where $\widehat{\text{ibd}}_s$ is calculated using existing multipoint methods such as those described in Kruglyak et al. [11]. For full sibs, Olson [14] standardized $\widehat{\text{ibd}}_s$ as $Z_s = \sqrt{2}(\widehat{\text{ibd}}_s - 1)$, and calculated the expected ibd sharing along a chromosome as

$$Y_c = \frac{1}{L_c} \int_0^{L_c} Z_s \, \text{d}s,$$

where c denotes an autosomal chromosome of interest and L_c denotes the length of that chromosome. She

then calculated an overall statistic across all 22 autosomes as

$$Y = \left(\sum_{c=1}^{22} Y_c\right) \Big/ \left[\sum_{c=1}^{22} \mathrm{var}(Y_c)\right]^{\frac{1}{2}}.$$

Y approximately follows a standard normal distribution under the putative relationship of full sibs.

Since one does not estimate ibd sharing continuously along a chromosome, but rather at a number of points spaced at equal intervals, Olson [14] approximated Y_c by another statistic

$$\widehat{Y}_c = \frac{d\sqrt{2}\sum_{p=1}^{P}\left(\widehat{\mathrm{ibd}}_p - 1\right)}{P},$$

where P denotes the number of points where one estimates ibd and d denotes the distance between adjacent points. Then, she calculated the overall approximate $\widehat{Y}$ in the same analogous fashion as Y. Her simulations demonstrated that $\widehat{Y}$ often has smaller tail probabilities than a standard normal distribution. Therefore, Olson [14] calculated relationship critical values for $\widehat{Y}$, which are functions of the genome length and the average marker information content. Her method inferred the relationship based on these critical values.

Continuous-Data Methods for Relationship Inference

Both likelihood-based and expected allele-sharing methods use the allele sharing of a relative pair at discrete points along the genome to infer relationship. Continuous-data methods assume that, with advances in sequencing technology, it may soon be possible to observe ibd allele sharing continuously along a chromosome. Relationship inference methods based on continuous data should be more powerful than methods based on discrete data, since the patterns and lengths of the sharing provide information for distinguishing different relationships.

Browning [4] used **Monte Carlo methods** to estimate the likelihood of a particular relationship of a pair using simulated continuous-ibd gamete data. She calculated the likelihood for the observed data weighted by multiple crossover processes consistent with the data simulated from the particular relationship. She evaluated this likelihood under both the putative relationship and the alternative one and then constructed a likelihood ratio statistic for inference. As with other Monte Carlo procedures, this method is computationally intensive.

Zhao & Liang [20] developed a method to evaluate the exact likelihood for the continuous-ibd gamete data using the theory of continuous-time **Markov chains**. The authors developed a relationship-specific intensity matrix that denotes the transitions from ibd sharing to nonsharing and vice-versa along the genome. Using this intensity matrix and the observed lengths of ibd sharing and nonsharing along the chromosome, the authors calculated the exact likelihood for the data under the given relationship. The authors then constructed a likelihood ratio statistic to test the putative relationship against an alternative one.

Discussion

Likelihood-based and expected allele-sharing methods provide an efficient means of distinguishing a variety of relationships. Simulation studies have shown that likelihood-based methods tend to have greater power to reject an incorrect putative relationship than expected allele-sharing methods under a variety of marker and map situations [7, 12]. Likelihood-based methods can also infer the actual relationship of the relative pair if the putative one is rejected. With the exception of Olson [14], none of the expected allele-sharing methods infer the actual relationship of a pair if the putative one is rejected.

The likelihood-based and expected allele-sharing methods rely on several assumptions. They assume that intermarker recombination fractions and population marker allele frequencies are known and assume that crossover interference (*see* **Genetic Map Functions**) is absent. In principle, the violation of these assumptions might lead to biased results; in fact, this appears not to be the case. Epstein et al. [8] explored the effect of recombination fraction misspecification and found their likelihood-based method was robust to such error. McPeek & Sun [12] investigated the effect of misspecified allele frequencies and found their ibs-based statistics were more sensitive to error than their likelihood-based method and Eibd statistic. The authors also investigated the effect of crossover

interference and found that it had a trivial effect on their results.

In theory, continuous-data methods should have more power than the other methods to infer the most likely relationship, since they use more data. However, as of yet, scientists cannot readily apply these methods because continuous data with known ibd sharing and nonsharing are unavailable. The power of the continuous-data methods to differentiate commonly tested relationships such as full sibs, half sibs, and grandparent–grandchild is unknown. Browning [4] and Zhao & Liang [21] only applied their methods to relatively distant relationships, including greatgrandparent–greatgrandchild and first cousins.

This work concentrated on inference methods for pairwise relationships. Recently, Sieberts et al. [17] developed a likelihood-based method for relationship inference for trios of relatives. The addition of a third relative is valuable since it can increase the power to infer relationships correctly. The authors constructed their likelihood by extending the Baum algorithms [1] in (3) and (4) to accommodate trios. Their method also allowed for a general error model that makes no assumptions about the relationship between the observed marker phenotype and the true underlying marker genotype. The previous error models of Broman and Weber [3] and Epstein et al. [8] both made the unrealistic assumption that the observed marker phenotype and true marker genotype were independent conditional on the occurrence of an error. Sieberts et al. [17] applied their trio-based method to real data examples and showed their method has an increased ability to distinguish relationships and requires less marker data for proper inference compared to pairwise methods. While their method can be more computationally intense than pairwise methods (depending on the chosen error model), the extra amount of computer time generally should not prevent efficient analysis.

Conclusions

Relationship testing of relative pairs within families is important to ensure the validity of analysis results. We have presented a variety of inference methods that can be used to test the putative relationship of a pair. Some methods presented here can also infer the most likely relationship of the pair when the putative one is incorrect. Unless otherwise noted, these methods are accurate, computationally fast, and robust to model misspecifications. Software for these methods has been developed and can be downloaded, usually for free, from the World Wide Web [7, 8, 10, 14, 17].

Acknowledgments

This work was supported by National Institutes of Health grants T32 HG00040 (to M.P.E.) and R01 HG00376 (to M.B.).

References

[1] Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, *Inequalities* **3**, 1–8.

[2] Boehnke, M. & Cox, N.J. (1997). Accurate inference of relationships in sib-pair linkage studies, *American Journal of Human Genetics* **61**, 423–429.

[3] Broman, K.W. & Weber, J.L. (1998). Estimation of pairwise relationships in the presence of genotyping errors, *American Journal of Human Genetics* **63**, 1563–1564.

[4] Browning, S. (1998). Relationship identification contained in gamete identity by descent data, *Journal of Computational Biology* **5**, 323–334.

[5] Chakraborty, R. & Jin, L. (1993). Determination of relatedness between individuals using DNA fingerprinting, *Human Biology* **65**, 875–895.

[6] Douglas, J.A., Boehnke, M. & Lange, K. (2000). A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data, *American Journal of Human Genetics* **66**, 1287–1297.

[7] Ehm, M.G. & Wagner, M. (1998). A test statistic to detect errors in sib-pair relationships, *American Journal of Human Genetics* **62**, 181–188.

[8] Epstein, M.P., Duren, W.L. & Boehnke, M. (2000). Improved inference of relationship for pairs of individuals, *American Journal of Human Genetics* **67**, 1219–1231.

[9] Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method, *Journal of Applied Probability* **30**, 766–779.

[10] Göring, H.H.H. & Ott, J. (1997). Relationship estimation in affected sib pair analysis of late-onset diseases, *European Journal of Human Genetics* **5**, 69–77.

[11] Kruglyak, L., Daly, M., Reeve-Daly, M. & Lander, E. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.

[12] McPeek, M.S. & Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data, *American Journal of Human Genetics* **66**, 1076–1094.

[13] O'Connell, J.R. & Weeks, D.E. (1998). Pedcheck: a program for identification of genotype incompatabilities in linkage analysis, *American Journal of Human Genetics* **63**, 259–266.

[14] Olson, J.M. (1999). Relationship estimation by Markov-process models in a sib-pair linkage study, *American Journal of Human Genetics* **64**, 1464–1472.

[15] Ott, J. (1993). Detecting marker inconsistencies in human gene mapping, *Human Heredity* **43**, 25–30.

[16] Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs, *American Journal of Human Genetics* **46**, 229–241.

[17] Stivers, D.N., Zhong, Y., Hanis, C.L. & Chakraborty, R. (1996). RELTYPE: a computer program for determining biological relatedness between individuals based on allele sharing at microsatellite loci, *American Journal of Human Genetics*, Supplement, **59**, A190.

[18] Stringham, H.M. & Boehnke, M. (1996). Identifying marker typing incompatibilities in linkage analysis, *American Journal of Human Genetics* **59**, 946–950.

[19] Thompson, E.A. (1975). The estimation of pairwise relationships, *Annals of Human Genetics* **39**, 173–188.

[20] Zhao, H. & Liang, F. (2001). On relationship inference using gamete identity by descent data, *Journal of Computational Biology* **8**, 191–200.

MICHAEL P. EPSTEIN & MICHAEL BOEHNKE

Relative Hazard

The relative hazard is the ratio of two **hazard rate** functions at a given time. If this hazard ratio is constant, as is assumed in the **proportional hazards model**, it can be **consistently estimated** both from **cohort** and from time-matched **case–control studies** (*see* **Density Sampling**). Over a small time interval, the relative hazard can be estimated as the **incidence density** ratio, also known as the **incidence rate** ratio.

MITCHELL H. GAIL

Relative Odds

The relative odds, or **odds ratio**, is the ratio of the **odds** of disease in an exposed cohort divided by that in an unexposed cohort. The relative odds can be estimated not only from **cohort** data but also from **case–control** data, because the relative odds of exposure comparing cases with disease-free controls equals the relative odds of disease comparing exposed with unexposed [1]. For rare diseases, the odds ratio approximates the **relative risk**.

Reference

[1] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix, *Journal of the National Cancer Institute* **11**, 1269–1275.

MITCHELL H. GAIL

Relative Risk Modeling

Risk models are used to describe the hazard function (*see* **Hazard Rate**) $\lambda(t, z)$ for time-to-failure data as a function of time t and **covariates** $\mathbf{Z} = Z_1, \ldots, Z_p$, which may themselves be time dependent. The term "relative risk models" is used to refer to the covariate part $r(\cdot)$ of a risk model in a **proportional hazards** form

$$\lambda(t, \mathbf{Z}) = \lambda_0(t)\, r[\mathbf{Z}(t); \boldsymbol{\beta}], \tag{1}$$

where $\boldsymbol{\beta}$ represents a vector of parameters to be estimated. In the standard proportional hazards model, the **relative risk** term takes the **loglinear** form $r(\mathbf{Z}, \boldsymbol{\beta}) = \exp(\mathbf{Z}'\boldsymbol{\beta})$. This has the convenient property that it is positive for all possible covariate and parameter values, since the hazard rate itself must be nonnegative. However, in particular applications, some alternative form of relative risk model may be more appropriate.

First, an aside on the subject of time is warranted. Time can be measured on a number of different scales, such as age, calendar time, or time since start of observation. One of these must be selected as the time axis t for use of the proportional hazards model. In **clinical trials**, time since diagnosis or start of treatment is commonly used for this purpose, since one of the major objectives of such studies is to make statements about **prognosis**. In epidemiologic studies, however, age is the preferred time axis, because it is usually a powerful determinant of disease rates, but it is not of primary interest; thus, it is essential that its **confounding** effects be eliminated. However, other temporal factors, such as calendar date, or time since exposure began may still be relevant and can be handled either by treating them as covariates or by **stratification**.

Why Model Relative Risks?

Before proceeding further, it is worth pausing to inquire why one might wish to adopt the proportional hazards model at all. Certainly, there are examples of situations where some other form of model provides a better description of the underlying biologic process. Two alternative models that have received some attention are the **additive hazards model** $\lambda(t, \mathbf{Z}) = \lambda_0(t) + \mathbf{Z}'\boldsymbol{\beta}$ and the **accelerated failure-time model** $S(t, \mathbf{Z}) = S_0[t \exp(\mathbf{Z}'\boldsymbol{\beta})]$, where $S(t) = \exp[-\int_0^t \lambda(u)\, du]$ is the survival function. Although any risk model can be reparameterized in proportional hazards form, it may be that a more **parsimonious** model can be found using some alternative formulation. For example, the additive risk model could be written as $\lambda(t, \mathbf{Z}) = \lambda_0(t)(1 + \tilde{\mathbf{Z}}'\boldsymbol{\beta})$, where $\tilde{\mathbf{Z}} = \mathbf{Z}/\lambda_0(t)$ if the baseline hazard $\lambda_0(t)$ were some known parametric function, such as a set of external rates for an unexposed population. In this case, whether the proportional hazards or additive hazards model provides a more parsimonious description of the data depends on whether relative risk or the **excess risk** is more nearly constant over time (or requires the fewest time-dependent **interaction** effects).

The advantages of relative risk models are both mathematical and empirical. Mathematically, the proportional hazards model allows "**semiparametric**" estimation of covariate effects via **partial likelihood** without requiring parametric assumptions about the form of the baseline hazard. Furthermore, at least with the standard loglinear form of the relative risk model, asymptotic **normality** seems to be achieved faster in many applications than for most alternative models. Empirically, it appears that many failure-time processes do indeed show rough proportionality of the hazard to time and covariate effects, at least with appropriate specification of the covariates. Evidence of this phenomenon for cancer incidence is reviewed in Breslow & Day [2, Chapter 2]: age-specific **incidence rates** from a variety of populations have more nearly constant ratios than differences.

Data Structures and Likelihoods

Failure-time data arise in many situations in biology and medicine. In clinical trials, time-to-death or time-to-disease-recurrence are frequently used endpoints. In epidemiology, **cohort studies** are often concerned with disease incidence or mortality in some exposed population, and **case–control studies** can be viewed as a form of sampling within a general population cohort. All these designs involve the collection of a set of data for each individual $i = 1, \ldots, I$ comprising a failure or censoring time t_i (*see* **Censored Data**), a censoring indicator $d_i = 1$ if the failure time is observed (i.e. the subject is affected), zero otherwise, and a vector of covariates $\mathbf{z}_i$, possibly time dependent.

The appropriate **likelihood** depends on the sampling design and data structure. For a clinical trial or cohort study with the same period of observation for all subjects, but where only the disease status, not the failure-time itself, is observed, a **logistic model** for the probability of failure of the form $\Pr(D = 0|\mathbf{Z}) = [1 + \alpha r(\mathbf{Z}, \boldsymbol{\beta})]^{-1}$ might be used, where α is the odds of failure for a subject with $\mathbf{Z} \equiv 0$. Again, the standard form is obtained using $r(\mathbf{Z}, \boldsymbol{\beta}) = \exp(\mathbf{Z}'\boldsymbol{\beta})$. The likelihood for this design would then be

$$L(\alpha, \boldsymbol{\beta}) = \prod_i \Pr(D = d_i|\mathbf{Z} = \mathbf{z}_i; \alpha, \boldsymbol{\beta}) = \prod_i \frac{[\alpha r(\mathbf{z}_i; \boldsymbol{\beta})]^{d_i}}{1 + \alpha r(\mathbf{z}_i; \boldsymbol{\beta})}. \quad (2)$$

The same model and likelihood function would be used for an unmatched case–control study, except that α now involves the control sampling fractions as well as the baseline disease risk.

In a clinical trial or cohort study in which the failure times are observed, the proportional hazards model (1) leads to a full likelihood of the form

$$L[\lambda_0(\cdot), \boldsymbol{\beta}] = \prod_i \lambda_0(t_i)^{d_i} r[\mathbf{z}_i(t_i); \boldsymbol{\beta}]^{d_i} \times \exp\left\{-\int_{s_i}^{t_i} \lambda_0(t)\, r[\mathbf{z}_i(t); \boldsymbol{\beta}]\, \mathrm{d}t\right\}, \quad (3)$$

where s_i denotes the entry time of subject i. Use of the full likelihood requires specification of the form of the baseline hazard. Cox [6] proposed instead a "partial likelihood" of the form

$$L(\boldsymbol{\beta}) = \prod_{n=1}^{N} \frac{r[z_{i_n}(t_n); \boldsymbol{\beta}]}{\sum_{j \in R_n} r[z_j(t_n); \boldsymbol{\beta}]}, \quad (4)$$

where $n = 1, \ldots, N$ indexes the observed failure times, i_n denotes the individual who fails at time t_n and R_n denotes the set of subjects at risk at time t_n. This likelihood does not require any specification of the form of the baseline hazard; the estimation of $\boldsymbol{\beta}$ is said to be "semiparametric", as the relative risk factor is still specified parametrically (e.g. the **loglinear model** in the standard form). This partial likelihood can also be used to fit relative risk models for matched case–control studies (including nested case–control studies within a cohort), where n now indexes the cases and R_n indicates the set comprising the nth case and his matched controls.

For very large data sets, it may be more convenient to analyze the data in grouped form using **Poisson regression**. For this purpose, the total person-time of follow-up is grouped into $k = 1, \ldots, K$ categories on the basis of time and covariates, and the number of events N_k and person-time T_k in each category is recorded, together with the corresponding values of the (average) time t_k and covariates $\mathbf{z}_k$. The proportional hazards model now leads to a Poisson likelihood for the grouped data of the form

$$L(\lambda, \boldsymbol{\beta}) = \prod_{k=1}^{K} [\lambda_k T_k r(\mathbf{z}_k; \boldsymbol{\beta})]^{N_k} \times \frac{\exp[-\lambda_k T_k r(\mathbf{z}_k; \boldsymbol{\beta})]}{N_k!}, \quad (5)$$

where $\lambda_k = \lambda_0(t_k)$ denotes a set of baseline hazard parameters that must be estimated together with $\boldsymbol{\beta}$.

Approaches to Model Specification

For any of these likelihoods, it suffices to substitute some appropriate function for $r(\mathbf{Z}; \boldsymbol{\beta})$ and then use the standard methods of **maximum likelihood** to estimate its parameters (*see* **Estimation**) and test hypotheses (*see* **Hypothesis Testing**). In the remainder of this article, we discuss various approaches to specifying this function. The major distinction we make is between empiric and mechanistic approaches. Empiric models are not based on any particular biologic theory for the underlying failure process, but simply attempt to provide a parsimonious description of it, particularly to identify and quantify the effects of covariates that affect the relative hazard. Perhaps the best known empiric model is the loglinear model for relative risks, but other forms may be appropriate for testing particular hypotheses or for more parsimonious modeling in particular data sets, as discussed in the following section. With a small number of covariates, it may also be possible to model the relative risk **nonparametrically**. Mechanistic models, on the other hand, aim to describe the observed data in terms of some unobservable underlying disease process, such as the **multistage theory of carcinogenesis**. We touch on such models briefly at the end.

Before proceeding further, it should be noted that what follows is predicated on the assumption that the covariates $\mathbf{Z}$ are accurately measured (or that the exposure–response relationship that will be estimated refers to the measured value of the covariates, not to their true values). There is a large and growing literature on methods of adjustment of relative risk models for measurement error (*see* **Measurement Error in Epidemiologic Studies**).

Empiric Models

The loglinear model, $\ln r(\mathbf{Z};\boldsymbol{\beta}) = \mathbf{Z}'\boldsymbol{\beta}$, is probably the most widely used empiric model and is the standard form included in all statistical packages for logistic, Cox, and Poisson regression (*see* **Software, Biostatistical**). As noted earlier, it is nonnegative and it produces a nonzero likelihood for all possible parameter values, which doubtless contributes to the observation that in most applications, parameter estimates are reasonably normally distributed, even with relatively sparse data. However, the model involves two key assumptions that merit testing in any particular application:

1. for a continuous covariate Z, the relative risk depends exponentially on the value of Z; and
2. for a pair of covariates, Z_1 and Z_2, the relative risk depends multiplicatively on the marginal risks from each covariate separately (i.e. $r(\mathbf{Z};\boldsymbol{\beta}) = r(Z_1;\beta_1)r(Z_2;\beta_2)$).

Neither of these assumptions is relevant for a single categorical covariate with K levels, for which one forms a set of $K-1$ indicator variables corresponding to all levels other than the "referent" category. In other cases, the two assumptions can be tested by nesting the model in some more general model that includes the fitted model as a special case. This test can be accomplished without leaving the general class of loglinear models. For example, to test the first assumption, it may suffice to add one or more **transformations** of the covariate (such as its square) to the model and test the significance of its additional contribution. To test the second assumption, one could add a single product term (for two continuous or binary covariates) or a set of $(K-1)(L-1)$ products for two categorical variables with K and L levels respectively.

If these tests reveal significant lack of fit of the original model, one might nevertheless be satisfied with the expanded model as a reasonable description of the data (after appropriately testing the fit of that expanded model). However, one should then also consider the possibility that the data might be more parsimoniously described by some completely different form of model. In choosing such an alternative, one would naturally be guided by what the tests of fit of the earlier models had revealed, as well as by categorical analyses. For example, if a quadratic term produced a negative estimate, that might suggest that a linear rather than loglinear model might fit better; similarly, a negative estimate for an interaction term might suggest an **additive** rather than **multiplicative** form of model for joint effects. In this case, one might consider fitting a model of the form $r(\mathbf{Z};\boldsymbol{\beta}) = 1 + \mathbf{Z}'\boldsymbol{\beta}$. Alternatively, one might prefer a model that is linear in each component, but multiplicative in their joint effects, $r(\mathbf{Z};\boldsymbol{\beta}) = \prod_p (1 + Z_p\beta_p)$, or one that is loglinear in each component but additive jointly, $r(\mathbf{Z};\boldsymbol{\beta}) = 1 + \sum_p [\exp(Z_p\beta_p) - 1]$.

In a rich data set, the number of possible alternative models can quickly get out of hand, so some structured approach to model building is needed. The key is to adopt a general class of models that would include all the alternatives one might be interested in as special cases, allowing specific submodels to be tested within nested alternatives. A general model that has achieved some popularity recently consists of a mixture of linear and loglinear terms of the form

$$r(\mathbf{Z},\mathbf{W};\boldsymbol{\beta},\boldsymbol{\gamma}) = \exp(\mathbf{W}_0'\boldsymbol{\gamma}_0)\left[1 + \sum_{m=1}^{M} \mathbf{Z}_m'\boldsymbol{\beta}_m \exp(\mathbf{W}_m'\boldsymbol{\gamma}_m)\right], \quad (6)$$

where $\boldsymbol{\beta}_m$ and $\boldsymbol{\gamma}_m$ denote vectors of regression coefficients corresponding to the subsets of covariates $\mathbf{Z}_m$ and $\mathbf{W}_m$ included in the mth linear and loglinear terms, respectively. Thus, for example, the standard loglinear model would comprise the single term $m = 0$, while the linear model would comprise a single term $m = 1$ with no covariates in the loglinear terms. A special case that has been widely used in radiobiology (*see* **Radiation**) is of the form

$$r(\mathbf{Z},\mathbf{W};\boldsymbol{\beta};\boldsymbol{\gamma}) = 1 + (\beta_1 Z + \beta_2 Z^2) \times \exp(-\beta_3 Z + \mathbf{W}'\boldsymbol{\gamma}),$$

where Z represents radiation dose (believed from microdosimetry considerations to have a linear-quadratic effect on mutation rates at low doses multiplied

by a negative exponential survival term to account for cell killing at high doses) and **W** comprises modifiers of the slope of the **dose–response** relationship, such as attained age, sex, latency, or age at exposure. For example, including the log of latency and its square in **W** allows for a **lognormal** dependence of excess relative risk on latency (*see* **Poisson Regression in Epidemiology** for a discussion of software for fitting such models).

Combining linear and loglinear terms, using the same p covariates, would produce a model of the form $r(\mathbf{Z};\boldsymbol{\beta},\boldsymbol{\gamma}) = \exp(\mathbf{Z}'\boldsymbol{\gamma})(1+\mathbf{Z}'\boldsymbol{\beta})$ against which the fit of the linear and loglinear models could be tested with p df. Although useful as a test of fit of these two specific models, the interpretation of the parameters is not straightforward since the effect of the covariates is essentially split between the two components. It would be of greater interest to form a model with a single set of regression coefficients and an additional mixing parameter for the combination of the submodels. Conceptually the simplest such model is the exponential mixture [28]

$$r(\mathbf{Z};\boldsymbol{\beta};\theta) = (1+\mathbf{Z}'\boldsymbol{\beta})^{1-\theta}\exp(\theta\mathbf{Z}'\boldsymbol{\beta}), \qquad (7)$$

which produces the linear model when $\theta = 0$ and the loglinear model with $\theta = 1$. An alternative, based on the Box–Cox transformation, was proposed by Breslow & Storer [3], which also includes the linear and loglinear models as special cases. However, Moolgavkar & Venzon [20] pointed out both of these mixture models are sensitive to the coding of the covariates: for example, for binary covariates, relabelling the two possible values leads to different models, leading to different inferences both about the mixing parameter and the relative importance of the component risk factors. Guerro & Johnson [12] developed a variant of the Box–Cox model of the form

$$r(\mathbf{Z};\boldsymbol{\beta},\theta) = \begin{cases} \exp(\mathbf{Z}'\boldsymbol{\beta}), & \theta = 0, \\ (1+\theta\mathbf{Z}'\boldsymbol{\beta})^{1/\theta}, & \theta \neq 0, \end{cases} \qquad (8)$$

which appears to be the only model in the literature to date that does not suffer from this difficulty. These kinds of mixtures could in principle also be used to compare relative risk with additive (excess) risk models, although the interpretation of the β coefficient becomes problematic because it has different dimensions under the different submodels.

Although suitable for testing multiplicativity vs. additivity with multidimensional categorical data, these mixtures are less useful for continuous covariates because they combine two quite different comparisons (the form of the dose–response relationship for each covariate and the form of their joint effects) into a single mixing parameter. One way around this difficulty is to compare linear and loglinear models for each covariate separately first to determine the best form of model, then to fit joint models, testing additivity vs. multiplicativity. Alternatively, one could form mixtures of more than two submodels with different mixing parameters for the different aspects.

A word of warning is needed concerning **inference** on the parameters of most nonstandard models. Moolgavkar & Venzon [20] pointed out that for nonstandard models, convergence to asymptotic normality can be very slow indeed. Thus, the log-likelihoods are generally far from quadratic, leading to highly skewed confidence regions. For this reason, Wald tests and **confidence limits** should generally be avoided. Furthermore, as the parameter moves away from the null, the **standard error** increases more quickly than the mean, so that the Wald test can appear to become less and less significant the larger the value of the parameter [13, 34]. These problems are particularly important for the mixing parameters θ, for which inference should be based on the **likelihood ratio test** and likelihood-based confidence limits. For example, Lubin & Gaffey [15] describe an application of the exponential mixture of linear-additive and linear-multiplicative models [28] to testing the joint effect of radon and smoking on lung cancer risk in uranium miners; the point estimate of θ was 0.4, apparently closer to additivity than multiplicativity, but the likelihood ratio tests rejected the additive model ($\chi^2_1 = 9.8$) but not the multiplicative model ($\chi^2_1 = 1.1$). A linear mixture showed an even more skewed likelihood, with $\hat{\theta} = 0.1$ (apparently nearly additive) but with very similar likelihood ratio tests that rejected the additive but not the multiplicative model.

Models for Extended Exposure Histories

Chronic disease epidemiology often involves measurement of an entire history of exposure $\{X(u), u < t\}$ which we wish to incorporate into a relative risk model through one or more time-dependent covariates $\mathbf{Z}(t)$. How this is done depends upon one's assumptions about the underlying disease mechanism.

We defer for the moment the possibility of modeling such a disease process directly and instead continue in the vein of empiric modeling, now focusing on eliciting information about the temporal modifiers of the exposure–response relationship.

Most approaches to exposure–response modeling in epidemiology are based on an implicit assumption of *dose additivity*, i.e. that the excess relative risk at time t is a sum of independent contributions from each increment of exposure at earlier times u, possibly modified in some fashion by temporal factors. This hypothesis can be expressed mathematically as

$$r[t, X(\cdot); \boldsymbol{\beta}; \boldsymbol{\gamma}] = R[Z(t); \boldsymbol{\beta}],$$

where

$$Z(t) = \int_0^t f[X(u); \alpha]\, g(t, u; \boldsymbol{\gamma})\, \mathrm{d}u, \qquad (9)$$

and where $R(Z; \boldsymbol{\beta})$ is some known relative risk function such as the linear or loglinear models discussed above, f is a known function describing the modifying effect of dose rate, and g is a known function describing the modifying effect of temporal factors. The simplest weighting functions would be $f(X) = X$ and $g(t, u) = 1$, for which $Z(t)$ becomes cumulative exposure, probably the most widely used exposure index in epidemiology. For many disease with long latency, such as cancer, it is common to use lagged cumulative exposure, corresponding to a weighting function of the form $g(t, u; \gamma) = 1$ if $t - u > \gamma$, zero otherwise. Other simple exposure indices might include time-weighted exposure $\int_0^t X(u)\,(t-u)\,\mathrm{d}u$ or age-weighted exposure $\int_0^{t-\gamma} X(u)\,u\,\mathrm{d}u$, which could be added as additional covariates to $R(\mathbf{Z}; \boldsymbol{\beta})$ to test the modifying effects of latency or age at exposure. The function f can be used to test dose-rate effects (the phenomenon that a long, low-intensity exposure has a different risk from a short, high-intensity exposure for the same cumulative dose). For example, one might adopt a model of the form $f(X) = X^\alpha$ or $f(X) = X \exp(-\alpha X)$ for this purpose.

Models that do not involve unknown parameters α and $\boldsymbol{\gamma}$ are easily fitted using standard software by the device of computing the time-dependent covariate(s) for each subject in advance. Relatively simple functions of γ (such as the choice of lagging interval in the simple latency model) might be fitted by evaluating the likelihood over a grid of values of the parameter. For more complex functions $g(t, u; \boldsymbol{\gamma})$, such as a lognormal density in $t - u$ with unknown mean and variance (and perhaps additional dependence of these parameters on age, exposure rate, or other factors), it is preferable to use a package with the capability of computing $Z(t; \alpha, \boldsymbol{\gamma})$ at each iteration. This generally requires some programming by the user, whereas most of the likelihood calculations and iterative estimation are handled by the package. For example, using SAS procedure NLIN, one can recompute the covariates at each iteration by the appropriate commands inside the procedure.

Unfortunately, the additivity assumption has seldom been tested. In principle, this could be done by nesting the dose-additive model in some more general model that includes interactive effects between the dose increments received at different times. The obvious alternative model would simply add further covariates of the form

$$Z^*(t) = \int_0^t \int_0^u F[X(u)X(v); \alpha] \times G(t, u, v; \boldsymbol{\gamma}, \delta)\, \mathrm{d}v\, \mathrm{d}u, \qquad (10)$$

where F and G are some known weighting functions. However, one should take care to see that the dose-additive model is well fitted first before testing the additivity assumption (e.g. by testing for nonlinearities and temporal modifiers).

Nonparametric Models

The appeal of Cox's partial likelihood is that no assumptions are needed about the form of the dependence of risk on time, but it remains parametric in modeling covariate effects. Even more appealing would be a nonparametric model for both time and covariate effects. For categorical data, no parametric assumptions are needed, of course, although the effects of multiple covariates are commonly estimated using the loglinear (i.e. multiplicative) model, with additional interaction terms as needed. Similarly, continuous covariates are frequently categorized to provide a visual impression of the exposure–response relationship, but the choice of cutpoints is arbitrary. However, nonparametric smoothing techniques are now available to allow covariate effects to be estimated without such arbitrary grouping.

One approach relies only on an assumption of monotonicity. Thomas [29] adapted the technique of

isotonic regression to relative risk modeling, and showed that the MLE of the exposure–response relationship under this constraint was a step function with jumps at the observed covariate values of a subset of the cases. The technique has been extended to two dimensions by Ulm [33], but in higher dimensions the resulting function is difficult to visualize and can be quite unstable.

Cubic **splines** and other means of smoothing provide attractive alternatives which produce smooth, but not necessarily monotonic, relationships. The **generalized additive model** [14] has been widely used for this purpose. For example, Schwartz [26] described the effect of air pollution on daily mortality rates using a generalized additive model, after controlling for weather variables and other factors using similar models. A complex dependence on dew point temperature was found, with multiple maxima and minima, whereas the smoothed plot of the particulate air pollution was seen to be almost perfectly linear over the entire rate of concentrations.

With the advent of **Markov chain Monte Carlo methods**, **Bayesian techniques** for model selection and smoothing have become feasible and are currently an active area of research. A full treatment of these methods is beyond the scope of this article; see Gilks et al. [11] for recent reviews of this literature.

Mechanistic Models

In contrast with the empiric models discussed above, there are circumstances where the underlying disease process is well enough understood to allow it to be characterized mathematically. Probably the greatest activity along these lines has been in the field of cancer epidemiology. Two models in particular have dominated this development, the multistage model of Armitage & Doll [1] and the two-event model of Moolgavkar & Knudson [18] (*see* **Multistage Carcinogenesis Models**). For thorough reviews of this literature, see [17], [31], and [36]; here, we merely sketch the basic ideas.

The Armitage–Doll multistage model postulates that cancer arises from a single cell that undergoes a sequence of k heritable changes, such as point mutations, chromosomal rearrangements, or deletions, in a particular sequence. The model further postulates that the rate of one or more of these changes may depend on exposure to carcinogens. Then the model predicts that the hazard rate for the incidence of cancer (or more precisely, the appearance of the first truly malignant cell) following continuous exposure at rate X is of the form

$$\lambda(t, Z) = \alpha t^{k-1} \prod_{i=1}^{k} (1 + \beta_i X). \tag{11}$$

Thus, the hazard has a **power** function dependence on age and a polynomial dependence on exposure rate with order equal to the number of dose-dependent stages. It further implies that two carcinogens would produce an additive effect if they act at the same stage and a multiplicative effect if they act at different stages. If exposure is instantaneous with intensity $X(u)$, its effect is modified by the age at and time since exposure: if it acts at a single stage i, then the excess relative risk at time t is proportional to $Z_{ik}(t) = X(u)u^{i-1}(t-u)^{k-i-1}/t^{k-1}$ and for an extended exposure at varying dose rates, the **excess relative risk** is obtained by integrating this expression over u [8, 35]. More complex expressions are available for time-dependent exposures to multiple agents acting at multiple stages [30]. These models can be fitted relatively easily using standard software by first evaluating the covariates $Z_{ik}(t)$ for each possible combination of $i < k$ and then fitting the linear relative risk model, as described above. Note, however, that the expressions given above are only approximations to the exact solution of the stochastic differential equations [16], which are valid when the mutation rates are all small.

The Moolgavkar–Knudson two-stage model postulates that cancer results from a clone of cells of which one descendant has undergone two mutational events, either or both of which may depend on exposure to carcinogens. The clone of intermediate cells is subject to a birth-and-death process (*see* **Stochastic Processes**) with rates that may also depend on carcinogenic exposures. The number of normal stem cells at risk varies with age, depending on the development of the particular tissue. Finally, in genetically susceptible individuals, all cells carry the first mutation at birth. The predicted risk under this model (in nonsusceptible individuals) is then approximately

$$\lambda[t, X(u)] = \mu_1\mu_2[1 + \beta_2 X(t)] \int_0^t [1 + \beta_1 X(u)] \times \exp[\rho(t-u)]\,\mathrm{d}u, \tag{12}$$

where μ_k are the baseline rates of the first and second mutations, β_k are the slope of the dependence of the mutation rates on exposure, and ρ is the net proliferation rate (birth minus death rates) of intermediate cells. For the more complex exact solution, see [24].

There have been a number of interesting applications of these models to various carcinogenic exposures. For example, the multistage model has been fitted to data on lung cancer in relation to asbestos and smoking [30], arsenic [4], coke oven emissions [9], and smoking [5, 10], as well as to data on leukemia and benzene [7] and nonleukemic cancers and radiation [32]. The two-stage model has been fitted to data on lung cancer in relation to smoking [23], radon [21, 25], and cadmium [27], as well as to data on breast [22] and colon cancers [19]. For further discussion of some of these applications, see [31].

As in any other form of statistical modeling, the analyst should be cautious in interpretation. A good fit to a particular model does not of course establish the truth of the model. Instead the value of models, whether descriptive or mechanistic, lies in their ability to organize a range of hypotheses into a systematic framework in which simpler models can be tested against more complex alternatives. The usefulness of the multistage model of carcinogenesis, for example, lies not in our belief that it is an accurate description of the process, but rather in its ability to distinguish whether a carcinogen appears to act early or late in the process or at more than one stage. Similarly, the importance of the Moolgavkar–Knudson model lies in its ability to test whether a carcinogen acts as an "initiator" (i.e. on the mutation rates) or a "promoter" (i.e. on proliferation rates). Such inferences can be valuable, even if the model itself is an incomplete description of the process, as must always be the case.

References

[1] Armitage, P. & Doll, R. (1961). Stochastic models of carcinogenesis, in *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, pp. 18–32.

[2] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. I. *The Analysis of Case–Control Studies*. *IARC Scientific Publications*, No. 32, Lyon.

[3] Breslow, N.E. & Storer, B.E. (1985). General relative risk functions for case–control studies, *American Journal of Epidemiology* **122**, 149–162.

[4] Brown, C.C. & Chu, K. (1983). A new method for the analysis of cohort studies: implications of the multistage theory of carcinogenesis applied to occupational arsenic exposure, *Environmental Health Perspectives* **50**, 293–308.

[5] Brown, C.C. & Chu, K. (1987). Use of multistage models to infer stage affected by carcinogenic exposure: example of lung cancer and cigarette smoking, *Journal of Chronic Diseases* **40**, 171–179.

[6] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[7] Crump, K.S., Allen, B.C., Howe, R.B. & Crockett, P.W. (1987). Time factors in quantitative risk assessment, *Journal of Chronic Diseases* **40**, 101–111.

[8] Day, N.E. & Brown, C.C. (1980). Multistage models and primary prevention of cancer. *Journal of the National Cancer Institute* **64**, 977–89.

[9] Dong, M.H., Redmond, C.K., Maxumdar, S. & Costantion, J.P. (1988). A multistage approach to the cohort analysis of lifetime lung cancer risk among steelworkers exposed to coke oven emission, *American Journal of Epidemiology* **128**, 860–873.

[10] Freedman, D.A. & Navidi, W.C. (1989). Multistage models for carcinogenesis, *Environmental Health Perspectives* **81**, 169–188.

[11] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[12] Guerro, V.M. & Johnson, R.A. (1982). Use of the Box–Cox transformation with binary response models, *Biometrika* **69**, 309–314.

[13] Hauck, W.W. & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis, *Journal of the American Statistical Association* **72**, 851–853.

[14] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall, New York.

[15] Lubin, J.H. & Gaffey, W. (1988). Relative risk models for assessing the joint effects of multiple factors, *American Journal of Industrial Medicine* **13**, 149–167.

[16] Moolgavkar, S.H. (1978). The multistage theory of carcinogenesis and the age distribution of cancer in man, *Journal of the National Cancer Institute* **61**, 49–52.

[17] Moolgavkar, S.H. (1986). Carcinogenesis modelling: from molecular biology to epidemiology, *Annual Review of Public Health* **7**, 151–169.

[18] Moolgavkar, S. & Knudson, A. (1980). Mutation and cancer: a model for human carcinogenesis, *Journal of the National Cancer Institute* **66**, 1037–1052.

[19] Moolgavkar, S.H. & Luebeck, E.G. (1992). Multistage carcinogenesis: population-based model for colon cancer, *Journal of the National Cancer Institute* **84**, 610–618.

[20] Moolgavkar, S. & Venzon, D.J. (1987). General relative risk regression models for epidemiologic studies, *American Journal of Epidemiology* **126**, 949–961.

[21] Moolgavkar, S.H., Cross, F.T., Luebeck, G. & Dagle, G.D. (1990). A two-mutation model for radon-induced lung tumors in rats, *Radiation Research* **121**, 28–37.

[22] Moolgavkar, S.H., Day, N.E. & Stevens, R.G. (1980). Two-stage model for carcinogenesis: epidemiology of breast cancer in females, *Journal of the National Cancer Institute* **65**, 559–569.

[23] Moolgavkar, S.H., Dewanji, A. & Luebeck, G. (1989). Cigarette smoking and lung cancer: reanalysis of the British doctors' data, *Journal of the National Cancer Institute* **81**, 415–420.

[24] Moolgavkar, S.H., Dewanji, A. & Venzon, D.J. (1988). A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor, *Risk Analysis* **8**, 383–392.

[25] Moolgavkar, S.H., Luebeck, E.G., Krewski, D. & Zielinski, J.M. (1993). Radon, cigarette smoke, and lung cancer: a re-analysis of the Colorado Plateau uranium miners' data, *Epidemiology* **4**, 204–217.

[26] Schwartz, J. (1993). Air pollution and daily mortality in Birmingham, Alabama, *American Journal of Epidemiology* **137**, 1136–1147.

[27] Stayner, L., Smith, R., Bailer, A.J., Luebeck, E.G. & Moolgavkar, S.H. (1995). Modeling epidemiologic studies of occupational cohorts for the quantitative assessment of carcinogenic hazards, *American Journal of Industrial Medicine* **27**, 155–170.

[28] Thomas, D.C. (1981). General relative risk models for survival time and matched case–control studies, *Biometrics* **37**, 673–686.

[29] Thomas, D.C. (1983). Nonparametric estimation and tests of fit for dose–response relations, *Biometrics* **39**, 263–268.

[30] Thomas, D.C. (1983). Statistical methods for analyzing effects of temporal patterns of exposure on cancer risks, *Scandinavian Journal of Work and Environmental Health* **9**, 353–366.

[31] Thomas, D.C. (1988). Models for exposure–time–response relationships with applications in cancer epidemiology, *Annual Review of Public Health* **9**, 451–482.

[32] Thomas, D.C. (1990). A model for dose rate and duration of exposure effects in radiation carcinogenesis, *Environmental Health Perspectives* **87**, 163–171.

[33] Ulm, K. (1983). Dose–response-models in epidemiology, in *Mathematics in Biology and Medicine: An International Conference*. Bari, Italy.

[34] Vaeth, M. (1985). On the use of Wald's test in exponential families, *International Statistical Review* **53**, 199–214.

[35] Whittemore, A.S. (1977). The age distribution of human cancers for carcinogenic exposures of varying intensity, *American Journal of Epidemiology* **106**, 418–32.

[36] Whittemore, A. & Keller, J.B. (1978). Quantitative theories of carcinogenesis, *SIAM Review* **20**, 1–30.

DUNCAN C. THOMAS

Relative Risk

The relative risk is the ratio of the **risk** of disease in an exposed **cohort** over a defined time interval to the risk of disease in an unexposed cohort over this same time interval. Relative risk is synonymous with **cumulative incidence ratio**. Relative risk can be estimated both from cohort studies, and, for rare diseases, from **case–control studies**.

MITCHELL H. GAIL

Reliability Study

Reliability studies and **validation studies** provide information on **measurement error** in exposures or other **covariates** used in epidemiologic studies. Such information on the measurement error process is needed to obtain valid estimates and **inference** using methods such as regression calibration or **maximum likelihood** (*see* **Misclassification Error**). Reliability studies are based on repeating an error-prone measurement, and the validity of this method depends on a model for the errors given by (1) below. Validation studies are applicable to a broader class of error models, including models admitting **differential error**, but validation studies require that one be able to measure correct ("**gold standard**") covariate values on some subjects.

To define reliability sampling plans more precisely, let $\mathbf{Y}$ be the response variable, and let $\mathbf{X}$ be the true values of the variable which may be misclassified or measured with error. In some cases, $\mathbf{X}$ can never be observed and can be thought of as a *latent* variable. In other cases, $\mathbf{X}$ is a "gold standard" method of covariate assessment which is infeasible and/or expensive to administer to large numbers of study participants. Instead of observing $\mathbf{X}$, we observe $\mathbf{W}$, which is subject to error. Finally, there may be covariates $\mathbf{Z}$ upon which the model for response depends that are measured without error. In main study/reliability study designs, the main study consists of the data $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i), i = 1, \ldots, n_1$. If the reliability study is *internal*, it consists of $(\mathbf{Y}_i, \mathbf{W}_{ij}, \mathbf{Z}_i)$, $j = 1, \ldots, n_i, i = n_1 + 1, \ldots, n_1 + n_2$ observations, and if the reliability study is *external*, it consists of $(\mathbf{W}_{ij})$, $j = 1, \ldots, n_i, i = n_1 + 1, \ldots, n_1 + n_2$ observations. Thus, there is only a single measurement for each main study subject, but replicate measurements for each subject in the reliability study.

The measurement error model for which a reliability study can be used is

$$\mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{1}$$

where $\mathbf{U}$ is a mean zero error term with some variance–covariance Σ. The error $\mathbf{U}$ is assumed independent of $\mathbf{X}$. Note that model (1) implies that the error is **nondifferential**, not only with respect to $\mathbf{Y}$ but also with respect to $\mathbf{Z}$ because $f(\mathbf{W}|\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = f(\mathbf{W}|\mathbf{X})$.

Under model (1), replicate data from a reliability study can be used for valid **estimation** and inference. This model has been applied to the analysis of blood pressure, serum hormones, and other serum biomarkers such as vitamin concentrations, viral load measurements, and CD4 cell counts.

We assume that subjects in an internal reliability study are selected completely at random. That is, if V is an indicator variable that equals 1 if a participant is in the validation study and 0 otherwise, then $\Pr(V = 1|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{W}) = \Pr(V = 1) = \pi$.

To correct point and interval estimates relating $\mathbf{Y}$ to $\mathbf{X}$ for **bias** from measurement error in $\mathbf{W}$, it is necessary to estimate Σ and $\text{var}(\mathbf{X}) = \Sigma_{\mathbf{X}}$ using model (1). Estimates of the quantities, Σ and $\Sigma_{\mathbf{X}}$, are needed to correct the estimate of the parameter of interest describing the association between $\mathbf{Y}$ and $\mathbf{X}$, β, for bias due to measurement error. If an internal reliability sample is used one can estimate Σ from it. The quantity $\text{var}(\mathbf{W}) = \Sigma_{\mathbf{W}}$ can be estimated from the combined main study/internal reliability study data, and $\Sigma_{\mathbf{X}}$ can be estimated by $\hat{\Sigma}_{\mathbf{X}} = \hat{\Sigma}_{\mathbf{W}} - \hat{\Sigma}$. The same approach can be used if an external reliability sample is used except, in this case, $\Sigma_{\mathbf{W}}$ should be estimated from the main study only. This is because, under model (1), it is reasonable to assume that Σ may be transportable from one population to another, whereas $\Sigma_{\mathbf{X}}$, and hence $\Sigma_{\mathbf{W}}$, are likely to vary across populations. Because an internal reliability study ensures that Σ is correctly estimated and yields more efficient estimates of $\Sigma_{\mathbf{X}}$, it is preferred to an external reliability study.

In some applications, the goal of the research is simply estimation of the reliability coefficient, also known as the intraclass correlation coefficient, ρ, equal to $\Sigma_{\mathbf{X}}[\Sigma_{\mathbf{W}}]^{-1}$ (*see* **Correlation**). These applications arise, for example, in the evaluation of new medical diagnostic procedures such as new technology for ascertaining load of HIV in body tissue, or in assessing the consistency of different clinicians in evaluating the functional status of their patients. Designs of studies whose purpose is to estimate the reliability coefficient have $n_1 = 0$ and no data on $\mathbf{Y}$. In what follows, we will first discuss design of such reliability studies. Then, we will discuss the main study/reliability study design, where $n_1 > 0$ and $\mathbf{Y}$ is observed in the main study and possibly in the reliability study.

Design of Reliability Studies

A nontechnical introduction to reliability study design considerations appeared in a recent epidemiology

textbook by Armstong et al. [1]. A series of papers by Donner and colleagues [3, 4, 8] investigated design of reliability studies in considerable detail. The first and last of these provided formulas for the **power** to test $H_0 : \rho = \rho_0$, vs. $H_a : \rho = \rho_A$, where ρ is the intraclass correlation or reliability coefficient, equal to $\Sigma_{\mathbf{X}}/\Sigma_{\mathbf{W}}$, for a given (n_2, R), and where R is the number of replicates per subject. In addition, tables were given for power for fixed values of n_2 and R. The first paper was based upon exact calculations, and the last paper developed a less computationally intensive approximation to the exact formula which appears to work quite well. The total number of observations $(n_2 \times R)$ is minimized with a relatively small value for R, as long as the reliability is 40% or higher. In these cases, $R = 2$ or 3 is sufficient. Eliasziw & Donner [4] minimized reliability study cost with respect to n_2 and R, subject to fixed power to test H_0 vs. H_a as given above using the formula for power derived in [3]. Cost was taken as a function of the unit cost of replicating data within subjects, the unit cost of accruing subjects, and the unit cost related jointly to the number of replicates and the number of subjects. Tables were given for the optimal values of n_2 and R, for different unit cost ratios and different values of ρ_0. They found that for $\rho > 0.2$, the cost per subject is more influential than the cost per measurement. In addition, they found that the optimal n_2 and R were highly stable despite moderate changes in unit cost ratios.

Freedman et al. [5] investigated the design of reliability studies when **X** and **W** are **binary**. Reliability of W as a **surrogate** for X was parameterized by the probability of disagreement between the two replicate measures of X, W_1, and W_2, corresponding to the values obtained from two different raters (*see* **Agreement, Measurement of**). They gave tables for n_2 which assured a fixed **confidence interval** width around the estimated probability of disagreement when $R = 2$. For probability of disagreement between 0.05 and 0.40 and confidence interval widths of 0.1 to 0.2, sample sizes between 50 and 350 are needed. These authors also considered study design when the goal is to estimate the within-rater probability of disagreement as well as the between-rater probability of disagreement, and provided tables of power for scenarios in which there are two raters and two replicates per rater.

Design of Main Study/Reliability Studies

One can select various main study sizes (n_1), reliability study sizes (n_2), and numbers of replicate measurements (R) for each subject in the reliability substudy. An "optimal" main study/reliability study design will find (n_1, n_2, R) to achieve some design goal. One may wish to minimize the **variance** of an important parameter estimate, such as the log **relative risk**, β, subject to a fixed total cost. Alternatively, one may wish to minimize the overall cost of the study, subject to specified power constraints on the parameter of interest (*see* **Validation Study** for further discussion of choices of design optimization criteria). Liu & Liang [6] considered the optimal choice of R for internal reliability designs with $n_1 = 0$, that is, designs in which all subjects are in the reliability study. They studied **generalized linear models** for $f(\mathbf{Y}|\mathbf{X}, \mathbf{Z}; \beta)$ with the identity, log, probit, and logit link functions. They assumed the measurement error model for **X** described by (1) with **X** following a **multivariate normal distribution** $\text{MVN}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}})$. The validity of their results required an additional approximation in the case of the logistic link function, which is the link function most commonly used in epidemiology. For scalar **X** and **W**, these authors derived a formula for **asymptotic relative efficiency** of β^*, the measurement-error corrected parameter describing the relationship between **Y** and **X**, as a function of $\Sigma/\Sigma_{\mathbf{X}}$ and R. They found that the precision of β^*, relative to the precision which would be obtained for estimating β if **X** were never measured with error, is little improved by increasing R above 4.

Rosner et al. [7] investigated the effect of changing n_2 and R on the variance of elements of a nine-dimensional vector β, where β is the log **odds ratio** relating coronary heart disease incidence to the model covariates in data from the **Framingham Heart Study** [2]. Four of the model covariates were measured with error (Figure 1). In this figure, n_1 was 1731, and Σ and Σ_x were assigned the values estimated in the analysis. When n_2 was greater than or equal to 100, the **standard error** of the four measurement-error corrected estimates reached an asymptote, indicating little or no gain in efficiency from increasing n_2 beyond that value. At that point, the gain in efficiency ranges between a 10%–20% reduction in the variance for the three variables measured with some error (BMI has little error, as

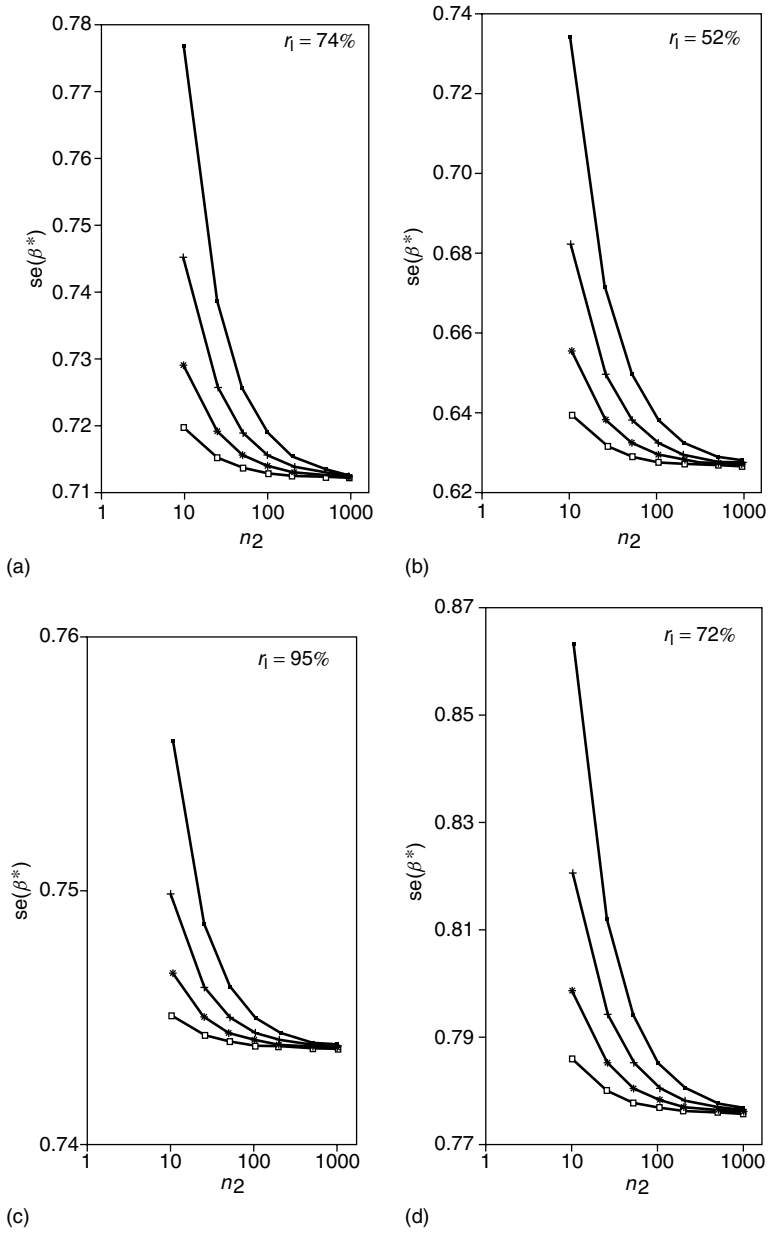

Figure 1 The relationship between the sample size (n_2) and the number of replicates per subject (R) in a reliability study, and the standard error of the measurement-error corrected logistic regression coefficient, β^*. Abbreviations and symbols used are: se for standard error and r_1 for the reliability coefficient $\text{var}(x)/\text{var}(w)$. Number of replicates, R: $\bullet = 2$; $+ = 3$; $* = 5$; $\square = 10$. (a) Cholesterol; (b) glucose; (c) body mass index; (d) systolic blood pressure

evidenced by the high reliability coefficient, $r_1 = 95\%$). Increasing the number of replicates decreased the standard errors of the estimates substantially when n_2 was small, but made little difference for larger reliability studies. For the three variables measured with error (cholesterol, glucose, and systolic blood pressure), the design ($n_2 = 10$, $R = 10$) was equally efficient as the design ($n_2 = 100$, $R = 2$). Although

the former requires fewer measurements, the latter may be more feasible, as it only requires two visits per subject.

Conclusion

Although model (1) is restrictive, there are many instances in biomedical research where it is considered reasonable. Methods of analysis under this model are well developed, but more research is needed on **optimal design**, and there is a need for user-friendly software for finding optimal designs.

Acknowledgments

This work was supported by National Cancer Institute grants CA50587 and CA03416.

References

[1] Armstong, B.K., White, E. & Saracci, R. (1992). *Principles of Exposure Measurement in Epidemiology*. Oxford University Press, Oxford, pp. 89–94.

[2] Dawber, T.R. (1980). *The Framingham Study*. Harvard University Press, Cambridge, Mass.

[3] Donner, A.P. & Eliasziw, M. (1987). Sample size requirements for reliability studies, *Statistics in Medicine* **6**, 441–448.

[4] Eliasziw, M. & Donner, A.P. (1987). A cost-function approach to the design of reliability studies, *Statistics in Medicine* **6**, 647–655.

[5] Freedman, L.S., Parmar, M.K.B. & Baker, S.G. (1993). The design of observer agreement studies with binary assessments, *Statistics in Medicine* **12**, 165–179.

[6] Liu, X. & Liang, K.Y. (1992). Efficacy of repeated measurements in regression models with measurement error, *Biometrics* **48**, 645–654.

[7] Rosner, B., Spiegelman, D. & Willett, W. (1992). Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error, *American Journal of Epidemiology* **136**, 1400–1413.

[8] Walter, S.D., Eliasziw, M. & Donner, A.P. (1998). Sample size and optimal designs for reliability studies, *Statistics in Medicine* **17**, 101–110.

DONNA SPIEGELMAN

Remington, Richard D.

Born: August 2, 1931, in Nampa, Idaho.
Died: July 26, 1992, in Iowa City, Iowa.

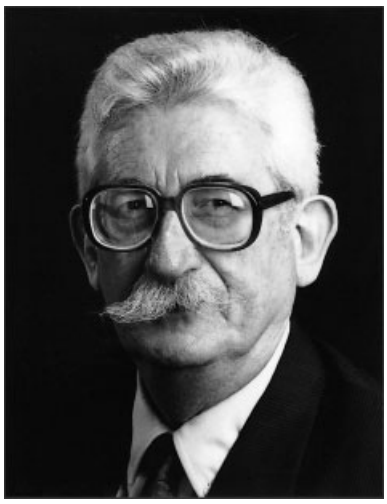

As a leading public health statistician, Richard D. Remington played a major role in linking the quantitative fields of biostatistics and epidemiology to public health research and policy. His distinguished academic career reflects an extraordinary commitment to public health, particularly his advocacy of the vital role that biostatistical and epidemiologic thinking and research must play in the resolution of significant public health problems and in the development of national public health policy. He was a tireless, articulate advocate for his profession.

Richard Remington began his university studies pursuing the field of mathematics. He received a B.A. degree in mathematics with honors in 1952 and an M.A. in mathematics in 1954, both from the University of Montana. He entered the field of public health, first receiving an M.P.H. degree in 1957, then a Ph.D. in Biostatistics in 1958, both from the University of Michigan.

On receiving his doctoral degree, he was appointed to the faculty of the Department of Biostatistics at the University of Michigan School of Public Health, serving there through 1969, and earning the rank of full professor in 1965. During a sabbatical year in 1966, he was a Visiting Scholar at the London School of Hygiene and Tropical Medicine.

In 1969 he joined the faculty of the University of Texas School of Public Health, Houston, where he served through 1974 as Associate Dean for Research, Professor and Head of Biometry. In 1975, he returned to the University of Michigan as Professor of Biostatistics and Dean of the School of Public Health. In 1982, he was appointed as the University of Iowa Foundation Distinguished Professor of Preventive Medicine and Environmental Health, and from 1982 to 1988 served as Vice President for Academic Affairs and Dean of the Faculty at the University of Iowa, also serving as Interim President of the University of Iowa in 1987–1988. Following a sabbatical year in 1989 at the University of Texas School of Public Health, Professor Remington returned to the University of Iowa, and founded the Institute for Health, Behavior, and Environmental Policy, serving as the Institute's Director until his death in 1992.

During his tenure at the University of Texas School of Public Health, Professor Remington served as the Scientific Director of the Data Management and Analysis Coordinating Center for the National Heart, Lung, and Blood Institute (NHLBI)'s landmark Hypertension Detection and Follow-up Program (HDFP). HDFP, in a national, community-based, randomized controlled trial (*see* **Clinical Trials, Overview**) involving 10 940 persons with high blood pressure, compared the effects on five-year mortality of systematic antihypertensive treatment and referral to usual community medical therapy. In recognition of the significant research advances made by HDFP, Professor Remington, his research colleagues, and HDFP staff received the Albert and Mary Lasker Special Public Health Award, the highest award in the field of public health.

Several decades of Professor Remington's career were involved with fundamental inquiries into the public health impact of hypertension, the importance of its early detection and treatment, and, more broadly, its prevention. For this work, Professor Remington was recognized in 1992 as the Lewis Conner Memorial Lecturer of the American Heart Association, and the following year he received the Golden Heart Award, the most prestigious honor bestowed by the American Heart Association.

Professor Remington was an effective voice for public health research, policy, and training through his leadership of dozens of significant national and international committees, including membership on the NHLBI Clinical Trials Review Committee, one of

the most distinguished review groups of the **National Institutes of Health**, and chairmanship of the Committee for the Study of the Future of Public Health, a committee appointed by the Institute of Medicine of the National Academy of Sciences. As Chairman, Professor Remington was chief architect of the Institute of Medicine report, *The Future of Public Health* [1].

Professor Remington's contributions to biostatistics, epidemiology, and public health have been recognized by honors and his election to memberships in scientific academies. He was elected a fellow of the **American Public Health Association**, the **American Statistical Association**, the UK **Royal Statistical Society**, an elected member of the Institute of Medicine of the National Academy of Sciences, and was awarded an honorary doctor of science degree from the University of Montana in 1984.

Professor Remington was a prolific author, producing over 80 scholarly articles and two books. His text, coauthored with Professor M. Anthony Schork, [2] is widely used in introductory biostatistics and statistics courses.

Second only to public health was his love of music. As Dick Remington he was a virtuoso on the double bass and jazz tuba, always in demand to record with well-known Dixieland musicians and Dixieland bands.

Perhaps even more enduring than his contributions, publications, and music is his colleagues' memory of his altruism, warm embrace, and humanity. No matter what the occasion, Dick Remington always found time to listen to his colleagues and students, urge them on, and share with them his insight, zest, and dedication to the public health profession.

A memorial, *Tribute to Richard D. Remington*, was read in the Senate of the United States by the Honorable Tom Harkin, Iowa Senator, on the Legislative day of Tuesday, September 8, 1992.

References

[1] Institute of Medicine Committee for the Study of the Future of Public Health (1988). *The Future of Public Health*. National Academy Press, Washington.

[2] Remington R. & Schork M.A. (1970). *Statistics: With Applications to the Biological and Health Sciences*, 2nd Ed. Prentice-Hall, Englewood Cliffs, 1985.

R.F. WOOLSON

Renewal Processes

In a renewal process we study the occurrences, or recurrences, of an event "E" of interest, and the distribution of the corresponding **random variables**. Generally, a process depends on the complexity of the distributions of the random variables, and the distribution may change following the occurrence of an event. In the development of the theory of renewal processes, we make use of the repetitive pattern in many practical situations, and consider an event E to be a "renewal event" only if the underlying conditions of a process remain unchanged following an occurrence of E. Formally, a renewal event is defined by the following two conditions:

1. the occurrence or nonoccurrence of an event E in any given time interval is uniquely determined; and
2. the process following an occurrence of event E is a complete (independent) replica of the process following any other occurrence of E.

Let t_r be a time interval following the $(r-1)$th occurrence of E and including the rth occurrence of E, for $r = 1, 2, \ldots$. According to condition 2 above, $(t_1, t_2, \ldots)$ will be independent and identical distributed random variables. The time in the process may be discrete as in a random walk (*see* **Stochastic Processes**), or continuous as in a **Poisson process**. In this article we briefly review the distributions of the renewal processes for both the discrete and continuous cases. We also consider the number of renewals $N(t)$ within a given time interval $(0, t]$, and the age, excess life, and total lifetime of the component in use at time t. Some extensions of the basic ordinary renewal process are also outlined.

We illustrate the concept of renewal events with a few examples.

Example 1. Success in Bernoulli Trials

Let event E be "success" in a sequence of Bernoulli trials (*see* **Binary Data**). The result of a sequence of trials may appear as follows: FFFSFSSFFS, so that the event E occurs at the fourth, the sixth, the seventh, and the tenth trials, and $t_1 = 4, t_2 = 2, t_3 = 1$, and $t_4 = 3$.

Example 2. Return to Origin

Consider a one-dimensional random walk of a particle starting from the origin. The particle moves one step to the right or one step to the left after each trial. Let event E be "return to origin". Suppose that the result of a sequence of trials is: RL LR RRLL LLRR .... Event E occurs at the second, the fourth, the eighth, and the 12th trials.

Remark. The event "success" in Example 1 is a "single" event; its occurrence at a trial is independent of the preceding trials. In Example 2, "return to origin", RL, LR, RRLL, LLRR, etc. represents a pattern. The occurrence of E at a particular trial is dependent on the outcomes of the preceding trials. Nevertheless, both examples satisfy the conditions underlying a renewal process. So far as occurrence of an event is concerned, trials need not be independent. Sequences of trials following occurrences of an event are independent sequences.

Example 3. Failure and Renewal

When an electric bulb, an automobile tire, or a mechanical component fails, it is replaced with a new one. The renewal times are continuous random variables, and are assumed independent and identically distributed.

When a renewal is a result of a failure, the term *failure time* or *lifetime* is often used for *renewal time*.

The theory of renewal processes was developed mainly by Feller [5–8]. Other contributions to the theory of renewal processes include those in [1, 4, 8, 10, 12], and [13]. This article is based on the material in [2] and [11].

Discrete Time Renewal Processes

In discrete time renewal processes, the units of time t may be called "trials". The occurrence of event E at the nth trial is denoted by "$t = n$". For the ith trial, we define a random variable X_i such that $X_i = 1$ if event E occurs at the ith trial, and $X_i = 0$ if not. We let $f(n)$ be the (first) renewal probability, defined as follows:

$$f(n) = \Pr\{X_n = 1 \quad \text{and} \quad X_m = 0; m = 1, 2, \ldots, n-1 | X_0 = 1\}, \quad n = 1, 2, \ldots. \tag{1}$$

It is clear that

$$f(n) = \Pr\{t = n\}.$$

The sum

$$\sum_{n=1}^{\infty} f(n) = f(\cdot)$$

is the probability that event E will eventually occur.

Classification of Events

A renewal event may or may not be recurrent, depending on the probability $f(\cdot)$.

Transient Event. A renewal event E is a transient event if $f(\cdot) < 1$. In this case there is a positive probability $1 - f(\cdot)$ that event E will not occur in a finite number of trials.

Recurrent Event. A renewal event E is a recurrent event if $f(\cdot) = 1$. In this case, the sequence $\{f(n)\}$ forms a proper probability distribution. The expectation

$$\sum_{n=1}^{\infty} nf(n) = \mathrm{E}(t) = \lambda \tag{2}$$

is the mean recurrent time.

Recurrent Null Event. A renewal event E is a recurrent null event if $f(\cdot) = 1$ and $\lambda = \infty$.

Recurrent Nonnull Event. A renewal event E is a recurrent nonnull event if $f(\cdot) = 1$ and $\lambda < \infty$.

Periodic Event and Aperiodic Event. A renewal event E is a periodic event if there exists an integer $\alpha > 1$ such that E can occur only at trials $\alpha, 2\alpha, 3\alpha, \ldots$. The largest α with this property is the period of E. A renewal event E is aperiodic if $\alpha = 1$.

Let t_r be the number of trials following the $(r-1)$th occurrence of E and including the rth occurrence of E, for $r = 1, 2, \ldots$, as defined above. Each random variable t_r has the probability distribution:

$$\Pr\{t_r = n\} = f(n), \quad n = 1, 2, \ldots.$$

The sum $T_r = t_1 + t_2 + \cdots + t_r$ is the length of time required for r occurrences of E, and its probability distribution,

$$\Pr\{T_r = n\} = f_r(n) \quad \text{for} \quad n = r, r+1, \ldots \tag{3}$$

is the r-fold convolution of $f(n)$ with itself, or, in convolution notation,

$$\{f_r(n)\} = \{f(n)\}^{r*}.$$

The sum

$$f_r(\cdot) = \sum_{n=1}^{\infty} f_r(n)$$

is the probability of r occurrences of E in an infinite sequence of trials. For a transient event E where $f(\cdot) < 1$, the probability $f_r(\cdot)$ tends to zero as r becomes infinitely large. This means that, in an infinite sequence of trials, a transient event occurs a finite number of times, while a recurrent event occurs infinitely often.

For a renewal event E, there are two types of probabilities associated with the occurrence of E. In addition to the probability $f(n)$ introduced above, there is a probability $p(n)$ that the event E will occur at the nth trial regardless of its occurrences in the preceding trials. Formally, we define $p(n)$ as follows:

$$p(n) = \Pr\{X_n = 1 | X_0 = 1\}, \quad n = 1, 2, \ldots,$$

with $p(0) = 1$. These two types of probabilities have the relationship

$$p(n) = \sum_{j=1}^{n} f(j)p(n-j), \tag{4}$$

so that

$$f(n) = p(n) - \sum_{j=1}^{n-1} f(j)p(n-j).$$

Theorem 1. The sum of $p(n)$ and the sum of $f(n)$ are related by formulas

$$\sum_{n=0}^{\infty} p(n) = \frac{1}{1 - f(\cdot)}, \tag{5}$$

and

$$f(\cdot) = 1 - \frac{1}{\sum_{n=0}^{\infty} p(n)}.$$

Thus, $f(\cdot) = 1$ if the infinite sum $\sum_{n=0}^{\infty} p(n)$ diverges, and $f(\cdot) < 1$ if the infinite sum of $p(n)$ converges.

Corollary. The event E is recurrent if

$$\sum_{n=0}^{\infty} p(n) = \infty,$$

and E is transient if

$$\sum_{n=0}^{\infty} p(n) < \infty.$$

Theorem 2. If event E is recurrent non-null with period α, then

$$\lim_{n\to\infty} p(\alpha n) = \frac{\alpha}{\lambda},$$

where λ is the mean renewal time. If E is recurrent non-null and aperiodic, then

$$\lim_{n\to\infty} p(n) = \frac{1}{\lambda}. \tag{6}$$

If E is recurrent null or transient, then $p(n) \to 0$ as $n \to \infty$.

We illustrate the above results with an example.

Example 4. "Success" in Bernoulli Trials

Let E be "success" in an infinite sequence of Bernoulli trials, with the probability of success in a single trial denoted by π. Then $p(n) = \pi$ and $p(0) = 1$ by definition. Clearly, the event E is recurrent nonnull and aperiodic, and

$$f(n) = [1-\pi]^{n-1}\pi.$$

According to (4),

$$\pi = \sum_{j=1}^{n-1}[(1-\pi)^{j-1}\pi]\pi + (1-\pi)^{n-1}\pi,$$

which obviously is true. Furthermore, the infinite sum

$$\sum_{n=0}^{\infty} p(n) = \sum_{n=0}^{\infty} \pi = \infty,$$

and

$$f(\cdot) = \sum_{n=1}^{\infty} f(n) = \sum_{n=1}^{\infty} (1-\pi)^{n-1}\pi = 1,$$

consistent with the fact that E is recurrent. To verify (6) in Theorem 2, note that $p(n) = \pi$ is independent of n, so that

$$\lim_{n\to\infty} p(n) = \pi,$$

while the mean renewal time is

$$\lambda = \sum_{n=1}^{\infty} nf(n) = \sum_{n=1}^{\infty} n[1-\pi]^{n-1}\pi = \frac{1}{\pi}.$$

Thus, λ and π are reciprocal, as required by (6).

Delayed Renewal Processes

In many practical situations the renewal process is already in progress when a first observation is made. This type of renewal processes is called the *delayed renewal process*. The probability distribution of t_1, denoted by $k(n)$, is different from $f(n)$, the common distribution of t_2 or t_3, etc. Suppose that at the time of first observation, n_0 trials have taken place since the last occurrence of E. Then $k(n)$ is the conditional probability that E will occur for the first time at the $(n_0 + n)$th trial given that E does not occur in the first n_0 trials. This means that $k(n)$ and $f(n)$ satisfy

$$k(n) = \frac{f(n_0+n)}{1 - \sum_{j=1}^{n_0} f(j)}.$$

The main feature of a delayed renewal process is the first occurrence of E. After that, the ordinary renewal process returns.

Continuous Time Renewal Processes

In a continuous renewal process, each renewal time t_r of a system has the same density function $f(\tau)$ and the same distribution function $F(\tau)$, for $r = 1, 2, \ldots$. The sum $T_r = t_1 + t_2 + \cdots + t_r$ is the length of time needed for r renewals of the system. The density function of T_r is obtained from

$$f_r(\tau) = \int_0^{\tau} f_{r-1}(\tau - x) f(x)\, \mathrm{d}x, \quad r = 2, 3, \ldots, \tag{7}$$

and the distribution function of T_r from

$$F_r(\tau) = \int_0^{\tau} F_{r-1}(\tau - x)\, \mathrm{d}F(x), \quad r = 2, 3, \ldots, \tag{8}$$

with repeated integrations beginning with $r = 2$.

Delayed Renewal Processes

If, at the initiation of a study, the system has been in operation for a period τ_0, then we have a delayed renewal process. The first random variable t_1 is the residual lifetime of the system beyond "age" τ_0. The distribution function of t_1 is

$$K(\tau) = \frac{F(\tau_0 + \tau) - F(\tau_0)}{1 - F(\tau_0)}. \tag{9}$$

The density function of t_1 is

$$k(\tau) = \frac{f(\tau_0 + \tau)}{1 - F(\tau_0)}.$$

The formulas of the distribution function and the density function of T_r are the same as those in (7) and (26), except that $F_1(\tau) = K(\tau)$ and $f_1(\tau) = k(\tau)$.

Example 5. Exponential Distribution I

Suppose each t_r has an **exponential distribution** with

$$f(\tau) = \mu \exp(-\mu\tau) \quad \text{and} \quad F(\tau) = 1 - \exp(-\mu\tau). \tag{10}$$

The expectation and the variance of t_r are, respectively,

$$\mathrm{E}(t_r) = \frac{1}{\mu} \quad \text{and} \quad \mathrm{var}(t_r) = \frac{1}{\mu^2},$$

where the parameter μ is known as the force of mortality in **life table** analysis and the failure rate or **hazard rate** in **survival analysis**. Using (7) and (8) we find the density function and the distribution of T_r:

$$f_r(\tau) = \mu^r \frac{\tau^{r-1}}{\Gamma(r)} \exp(-\mu\tau),$$

and

$$F_r(\tau) = 1 - \exp(-\mu\tau) \sum_{i=0}^{r-1} \frac{(\mu\tau)^i}{i!}, \tag{11}$$

which is a **gamma distribution** with parameters μ and r. Note that when each t_r has the exponential distribution (10), the process of renewal times is a one-dimensional Poisson process with rate μ.

In the delayed renewal process where t_1 is the residual lifetime of the system beyond τ_0, the distribution function of t_1 is, from (9),

$$\begin{aligned} K(\tau) &= \frac{\{1 - \exp[-\mu(\tau_0 + \tau)]\} - [1 - \exp(-\mu\tau_0)]}{1 - [1 - \exp(-\mu\tau_0)]} \\ &= 1 - \exp(-\mu\tau) = F(\tau). \end{aligned}$$

Thus the residual lifetime t_1 has the same distribution as the total lifetime t_2 or t_3. This, however, is a special case. Here, the force of mortality, or the failure rate μ, is independent of the "age" of the system; the probability that the system will fail in a time element $(\tau, \tau + \mathrm{d}\tau)$ remains the same regardless of the length of time it has been in operation. A system that has been in operation up to time τ_0 will last as long as when it is new. The equality $K(\tau) = F(\tau)$ is justified.

A System with Components

Suppose that a system has s components which have the renewal times (or failure times) $(\tau_1, \tau_2, \ldots, \tau_s)$. Each τ_i has the same density function $h(\tau)$ and the distribution function $H(\tau)$. Let T be the renewal time of the system with the density function $f(\tau)$ and the distribution function $F(\tau)$. Obviously, the distribution of T is a function of $H(\tau)$ and the number s, but their exact relationship depends on the definition of the failure of the system. The following are two definitions of failure and the corresponding distributions.

Minimum Length of Life

An electric circuit with light bulbs connected in series fails as soon as one of the bulbs burns out. A chain is broken when its weakest link fails. Generally, a system with s components operating concurrently fails when the component with the shortest lifetime fails. Let us arrange the lifetimes of the components $(\tau_1, \tau_2, \ldots, \tau_s)$ in the order of magnitude: $[\tau_{(1)} < \tau_{(2)} < \cdots < \tau_{(s)}]$. The renewal time of the system is the first **order statistic** $\tau_{(1)}$. Therefore the density function of the renewal time of the system T is

$$f(\tau) = h_{(1)}(\tau) = s[1 - H(\tau)]^{s-1} h(\tau),$$

and the distribution function of T is

$$F(\tau) = H_{(1)}(\tau) = 1 - [1 - H(\tau)]^s.$$

Example 6. Exponential Distribution II

Suppose the common distribution of the s components is exponential as given in (10). The renewal time of the system has the density function

$$f(\tau) = h_{(1)}(\tau) = s\mu \exp(-s\mu\tau),$$

and the distribution function

$$F(\tau) = H_{(1)}(\tau) = 1 - \exp(-s\mu\tau),$$

which is an exponential distribution with parameter $s\mu$. The expected renewal time of the system is

$$\mathrm{E}(T) = \mathrm{E}[t_{(1)}] = \frac{1}{s\mu},$$

and the variance is

$$\mathrm{var}(T) = \mathrm{var}[t_{(1)}] = \frac{1}{(s\mu)^2}.$$

Maximum Length of Life

An electric circuit with s light bulbs connected in parallel fails only when all the bulbs are burned out. A room with s lights will not be dark unless all the lights are out. A system is still in operation so long as one of the components is working. In such cases, the renewal time of a system T is the maximum lifetime of $(\tau_1, \tau_2, \ldots, \tau_s)$, or the sth order statistic $\tau_{(s)}$. It follows that the distribution of the renewal time T is

$$F(\tau) = [H(\tau)]^s,$$

and the density function of t is

$$f(\tau) = s[H(\tau)]^{s-1} h(\tau)$$

Example 7. Exponential Distribution III

Suppose that the common distribution of the s components is exponential as in (10). Then the renewal time of the system has density function

$$f(\tau) = h_{(s)}(\tau) = s[1 - \exp(-\mu\tau)]^{s-1} \mu \exp(-\mu\tau),$$

and distribution function

$$F(\tau) = H_{(s)}(\tau) = [1 - \exp(-\mu\tau)]^s.$$

The expected renewal time of the system is

$$\mathrm{E}(T) = \mathrm{E}[\tau_{(s)}] = \frac{1}{\mu}\left(1 + \frac{1}{2} + \cdots + \frac{1}{s}\right),$$

and the variance is

$$\mathrm{var}(T) = \mathrm{var}[\tau_{(s)}] = \frac{1}{\mu^2}\left(1 + \frac{1}{2^2} + \cdots + \frac{1}{s^2}\right).$$

Number of Renewals *N(t)*

Thus far we have been discussing the length of time $\{t_r\}$ and $\{T_r\}$ needed for a given number of renewals to take place. Another aspect of renewal processes of considerable interest is the number of renewals $N(t)$ occurring within a given time interval $(0, t]$. In this case, the time interval $(0, t]$ is fixed, while the number of renewals $N(t)$ is a random variable. The main purpose here is to derive a formula for the probability distribution of $N(t)$,

$$\Pr\{N(t) = r\} = P_r(t), \quad r = 1, 2, \ldots.$$

It can be shown that

$$\sum_{r=1}^{\infty} \Pr\{N(t) = r\} = 1 \tag{12}$$

and $N(t)$ is a proper random variable.

Theorem 3. The probability distribution and the expectation of $N(t)$ are related to the distribution function $F_r(t)$ of the time of r renewals T_r as follows:

$$\Pr\{N(t) = 0\} = 1 - F_1(t) \tag{13}$$

$$\Pr\{N(t) = r\} = F_r(t) - F_{r+1}(t), \quad r = 1, 2, \ldots \tag{14}$$

and

$$\mathrm{E}[N(t)] = \sum_{r=1}^{\infty} F_r(t). \tag{15}$$

Proof. If the number of renewals occurring within the time interval $(0, t]$ is greater than or equal to r, then the length of time required for r renewals must be less than or equal to t, and vice versa. In other words,

$$N(t) \geq r \quad \text{and} \quad T_r \leq t$$

are equivalent, and the corresponding probabilities must be equal:

$$\Pr\{N(t) \geq r\} = \Pr\{T_r \leq t\} = F_r(t).$$

Therefore,

$$\begin{aligned}\Pr\{N(t) = r\} &= \Pr\{N(t) \geq r\} - \Pr\{N(t) \geq r+1\} \\ &= F_r(t) - F_{r+1}(t).\end{aligned}$$

When $r = 0$, (14) implies (13). To find the expectation $\mathrm{E}[N(t)]$, we write

$$\begin{aligned}\mathrm{E}[N(t)] &= \sum_{r=1}^{\infty} r \Pr\{N(t) = r\} \\ &= \sum_{r=1}^{\infty} r[F_r(t) - F_{r+1}(t)] \\ &= \sum_{r=1}^{\infty} F_r(t),\end{aligned}$$

and complete the proof.

Theorem 3 provides us with a simple method of obtaining the probability distribution and the expectation of $N(t)$ directly from the distribution function of the renewal time $F_r(t)$.

The function $M(t) = \mathrm{E}[N(t)]$ is called the *renewal function* and, when it exists, its derivative $m(t) = M'(t)$ is called the *renewal density*. Note from (15) that $m(t) = \sum_{r=1}^{\infty} f_r(t)$, so that, for small Δt, the probability that there is a renewal in the time interval $(t, t + \Delta t)$ is given by $m(t)\Delta t + o(\Delta t)$.

Theorem 4. If in an ordinary renewal process the renewal time t_r has a finite expectation $\mathrm{E}(t_r) = \lambda$ and a finite variance σ^2, then, as $t \to \infty$, the number of renewals $N(t)$ has an asymptotic normal distribution which has a mean t/λ and a variance $t\sigma^2/\lambda^3$.

Feller [6] originally established the theorem for the discrete case, where t is the number of trials. Takacs [13] proved the theorem for time - continuous processes. The following theorem is due to Smith [12].

Theorem 5. If the renewal time has a finite mean $E(t_r) = \lambda$ and a finite variance σ^2, then

$$\lim_{t\to\infty} \frac{\mathrm{E}[N(t)]}{t} = \frac{1}{\lambda}, \tag{16}$$

and

$$\lim_{t\to\infty} \frac{\mathrm{var}[N(t)]}{t} = \frac{\sigma^2}{\lambda^3}. \tag{17}$$

There is a fine difference between Theorems 4 and 5. Theorem 4 is regarding the asymptotic distribution of $N(t)$, whereas Theorem 5 concerns the moments of $N(t)$ as $t \to \infty$. Since convergence of a distribution does not imply convergence of moments, the two theorems are addressing two different issues.

To appreciate the above theorems, let us consider a specific distribution.

Example 8. Exponential Distribution IV

When the renewal time of a system t_r has an exponential distribution as given in (10) and the expectation and the variance of t_r are

$$\mathrm{E}(t_r) = \frac{1}{\mu} \quad \text{and} \quad \sigma^2 = \frac{1}{\mu^2}, \tag{18}$$

the interval of r renewals $T_r = t_1 + t_2 + \cdots + t_r$ has a gamma distribution with the distribution function given in (11):

$$F_r(t) = 1 - \exp(-\mu t) \sum_{i=0}^{r-1} \frac{(\mu t)^i}{i!}. \tag{11a}$$

According to Theorem 3, the probability distribution of the number of renewals in $(0, t)$ is

$$\Pr\{N(t) = r\} = F_r(t) - F_{r+1}(t). \tag{14}$$

Substituting (11) in (14) yields

$$\Pr\{N(t) = r\} = \frac{(\mu t)^r}{r!} \exp(-\mu t), \tag{19}$$

which is a **Poisson distribution** with parameter μt. It follows that the expectation and the variance of $N(t)$ are, respectively,

$$\mathrm{E}[N(t)] = \mu t \quad \text{and} \quad \sigma^2_{N(t)} = \mu t. \tag{20}$$

According to Theorem 5,

$$\mathrm{E}[N(t)] \longrightarrow \frac{t}{\lambda} \tag{16a}$$

and

$$\mathrm{var}[N(t)] \longrightarrow \frac{t\sigma^2}{\lambda^3}. \tag{17a}$$

Substituting formulas (18) for the expectation and the variance of t_r in (16a) and (17a), we recover the formulas of the expectation and the variance of $N(t)$ in (20). Verification of the theorems is complete.

Finally, the reader may wish to prove as an exercise the equation

$$\sum_{r=1}^{\infty} F_r(t) = \mu t,$$

where $F_r(t)$ is given in (11).

Backward and Forward Recurrence Times

Consider a continuous time renewal process with lifetimes t_r having $E(t_r) = \lambda < \infty$, density $f(\tau)$, and distribution function $F(\tau)$. Let $t > 0$ be a fixed point in time. The *backward recurrence time* A_t is the time since the last renewal, that is, the age of the component in use at time t, and the *forward recurrence time* E_t is the time until the next renewal, that is, the excess life of the component in use at time t. For most choices of density $f(\tau)$, the distributions of A_t and E_t cannot be determined explicitly, however, their asymptotic distributions are equal and given by (e.g. [11, Proposition 3.4.5])

$$\lim_{t\to\infty} \Pr\{E_t \le x\} = \lim_{t\to\infty} \Pr\{A_t \le x\} = \frac{1}{\lambda}\int_0^x [1 - F(u)]\,\mathrm{d}u. \quad (21)$$

Moreover, provided $\sigma^2 = \mathrm{var}(t_r) < \infty$, then (e.g. [11, Proposition 3.4.6])

$$\lim_{t\to\infty} E[E_t] = \lim_{t\to\infty} E[A_t] = \frac{\sigma^2 + \lambda^2}{2\lambda}. \quad (22)$$

It is easily verified that the mean of the distribution given by the right-hand side of (21) is $(\sigma^2 + \lambda^2)/(2\lambda)$. For $t > 0$, let $T_t = A_t + E_t$ be the *total lifetime* of the component in use at time t. Then, using (22),

$$\lim_{t\to\infty} E[T_t] = \lim_{t\to\infty} E[E_t] + \lim_{t\to\infty} E[A_t] = \frac{\sigma^2 + \lambda^2}{\lambda}. \quad (23)$$

Example 9. Exponential Distribution V

When the renewal time of a system t_r has the exponential distribution given by (10), it is easily verified that the limiting distributions of E_t and A_t are also given by (10) and that $\lim_{t\to\infty} E[T_t] = 2/\mu$. Thus, the limiting mean total lifetime of the component in use at time t is $2/\mu$ while the mean lifetime of a component is $1/\mu$! At first sight, this result, which is known as the *inspection paradox*, may seem surprising. However, it has a fairly simple explanation, which is given below in the general setting.

Arguing informally, considering the interval that covers the point t and letting $t \to \infty$ is equivalent to considering the interval covering a point chosen uniformly in $[0, t]$ and letting $t \to \infty$. In the latter, it is clear that an interval of length y is y times more likely to be sampled than an interval of length 1, so the limiting distribution of T_t as $t \to \infty$ has density $f_T(y)$ that is directly proportional to $yf(y)$. Since $\int_0^\infty f_T(y)\,\mathrm{d}y = 1$ it follows that $f_T(y) = \lambda^{-1} y f(y)$ $(y > 0)$. Note that

$$E[T] = \int_0^\infty y f_T(y)\,\mathrm{d}x = \int_0^\infty \lambda^{-1} y^2 f(x)\,\mathrm{d}x = \frac{\sigma^2 + \lambda^2}{\lambda},$$

agreeing with (23). Note also, that in the limit as $t \to \infty$, the sampled point is uniformly distributed within the interval containing it. Thus, denoting the limiting excess and total lifetimes by E and T, respectively, for $x > 0$, $\Pr\{E > x | T = y\} = (y - x)/y$ if $y > x$ and zero otherwise, so conditioning on T,

$$P(E > x) = \int_0^\infty \Pr\{E > x | T = y\} f_T(y)\,\mathrm{d}y = \int_x^\infty \lambda^{-1} f(y)(y - x)\,\mathrm{d}y. \quad (24)$$

Differentiating (24) with respect to x shows that E has density

$$f_E(x) = \lambda^{-1}\int_x^\infty f(y)\,\mathrm{d}y = \lambda^{-1}(1 - F(x)),$$

which corresponds to the right-hand side of (21).

Equilibrium Renewal Processes

Consider a delayed renewal process in which the first lifetime t_1 has density function given by

$$k(\tau) = \lambda^{-1}(1 - F(\tau)) \quad (25)$$

and distribution function given by

$$K(\tau) = \lambda^{-1} \int_0^\tau [1 - F(u)]\, du, \qquad (26)$$

where F is the distribution function of the subsequent lifetimes $t_2, t_3, \ldots$. Such a process is called an *equilibrium renewal process*. Note that if we start observing an ordinary renewal process at some fixed time $t > 0$, then the resulting process is a delayed renewal process in which the initial lifetime t_1 is given by the excess life E_t. Thus, for t large, it follows from (21) that the observed process will be very close to the equilibrium renewal process, and indeed equal to it in the limit as $t \to \infty$. It can be shown that the equilibrium renewal process is stationary, that for any $t > 0$, the excess life of the component in use at time t has the same distribution as t_1, (i.e. given by (25) and (26)), and that its renewal function and renewal density are respectively $\lambda^{-1}t$ and λ^{-1} ($t \geq 0$) (see, for example, [11, Theorem 3.5.2]).

Alternating Renewal Process

In an *alternating renewal process*, the even numbered renewal times $t_{2r}, r = 1, 2, \ldots$, have a different distribution to the odd numbered renewal times $t_{2r-1}, r = 1, 2, \ldots$. For example, consider a single machine that breaks down repeatedly. Suppose that the machine has just been repaired at time $t = 0$. Then $t_1, t_3, t_5, \ldots$ denote the lengths of successive working periods of the machine and $t_2, t_4, t_6, \ldots$ denote the lengths of successive repair times. Suppose that $t_1, t_2, \ldots$ are independent, $t_1, t_3, \ldots$ each have density $f_1(\tau)$ and mean λ_1, and $t_2, t_4, \ldots$ each have density $f_2(\tau)$ and mean λ_2. Then

$$\lim_{t\to\infty} \Pr\{\text{machine is working at time } t\} = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \qquad (27)$$

In fact, (27) holds under the weaker assumption that (t_1, t_2), (t_3, t_4), ... are independent, that is, allowing for dependence between the lengths of a working period and the subsequent repair period (see, for example, [11, Theorem 3.4.4]). Further characteristics of the system, such as the limiting excess life distribution given that the machine is working at time t, can also be computed; (see, for example, [3, Chapter 7], and [11, Section 3.4.1].

Renewal Reward Process

Suppose that in a renewal process there is an award, R_i say, associated with the ith renewal. Then the total reward earned by time t is

$$R(t) = \sum_{i=1}^{N(t)} R_i,$$

where $N(t)$ is the number of renewals occurring in the time interval $(0, t]$ and the sum is zero if $N(t) = 0$. For example, suppose that claims come into an insurance company at the points of a renewal process $\{N(t), t \geq 0\}$. Then, if R_i denotes the value of the ith claim, $R(t)$ is the total value of all claims over $(0, t]$. Suppose that (t_1, R_1), (t_2, R_2), ... are independent and identically distributed, that $\lambda = E[t_1] < \infty$ and $\lambda_R = E[R_1] < \infty$. Then

$$\frac{R(t)}{t} \longrightarrow \frac{\lambda_R}{\lambda} \quad \text{almost surely as} \quad t \to \infty,$$

and

$$\frac{E[R(t)]}{t} \longrightarrow \frac{\lambda_R}{\lambda} \quad \text{as} \quad t \to \infty,$$

(see, for example, [11, Theorem 3.6.1]). Note that t_i and R_i are not assumed to be independent. The above results also hold if, instead of being earned en masse at the end of a lifetime the reward is earned gradually over the lifetime. As well as occurring naturally in many practical settings, renewal reward processes can also be used to analyze more complicated stochastic processes, such as queueing processes, where they arise as embedded processes, see, for example, [9, Section 10.5].

References

[1] Blackwell, D. (1948). A renewal theorem, *Duke Mathematical Journal* **15**, 145–150.

[2] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Krieger, New York.

[3] Cox, D.R. (1962). *Renewal Theory*. Methuen, London.

[4] Doob, J. (1948). Renewal theory from the point of view of the theory of probability, *Transactions of the American Mathematical Society* **63**, 422–438.

[5] Feller, W. (1941). On the integral theory of renewal theory, *Annals of Mathematical Statistics* **12**, 243–267.

[6] Feller, W. (1949). Fluctuation theory of recurrent events, *Transactions of the American Mathematical Society* **67**, 98–119.

[7] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol II. Wiley, New York.

[8] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd Ed. Wiley, New York.

[9] Grimmett, G.R. & Stirzaker, D.R. (2001). *Probability and Random Processes*, 3rd Ed. University Press, Oxford.

[10] Pyke, R. (1961). Markov renewal processes, *Annals of Mathematical Statistics* **32**, 1231–1242.

[11] Ross, S.M. (1996). *Stochastic Processes*, 2nd Ed. Wiley, Chichester.

[12] Smith, W.L. (1958). Renewal theory and its ramifications, *Journal of the Royal Statistical Society, Series B* **20**, 243–302.

[13] Takacs, L. (1954). Some investigations concerning recurrent stochastic processes of a certain type (in Hungarian), *Magyar Tud. Akad. Alaklm. Mat. Int. Kozl.* **3**, 115–128.

CHIN LONG CHIANG & FRANK BALL

Repeated Events

Repeated or recurrent events often arise in **longitudinal** studies involving multiple subjects. Some examples are the occurrence of epileptic seizures [2], the recurrence of tumors in cancer patients or laboratory animals [5, 12], and coughing or wheezing episodes in persons with bronchial asthma [6]. Broad objectives in analyzing repeated events include (i) understanding and characterizing event occurrence processes for individual subjects, (ii) characterizing subject-to-subject variability and relating it to **covariates** or treatments, and (iii) assessing the relationship of **time-dependent covariates** or other processes to event occurrence. Several types of repeated events may be of interest in a single application but throughout most of this discussion, we assume that one specific type of event is being considered.

To set up a framework for analyzing repeated events, suppose that subjects $i = 1, \ldots, m$ are observed and that the times of occurrence of events are recorded. Assume that subject i is observed over the time period $(0, \tau_i]$; τ_i is sometimes referred to as a **censoring** or termination time. Let $t_{i1} \leq t_{i2} \leq \cdots$ denote the times of event occurrence for subject i and let $N_i(t)$ represent the number of events over $(0, t]$. Finally, let $t_{i0} = 0$ and $x_{ij} = t_{ij} - t_{i,j-1}$ $(j = 1, 2, \ldots)$.

In the following sections, we describe the modeling of repeated events, discuss methods of analysis, and present an example.

Models for Repeated Events

Let us temporarily drop the subscript i and consider events that occur in continuous time for an arbitrary subject. A full probability model for an orderly (no coincident events) process $\{N(t) : t \geq 0\}$ may be specified in terms of the complete intensity function [9, p. 9]; (*see* **Point Processes**). Define $H_t = \{N(s) : s < t\}$ as the "history" of the process up to time t, and let $\mathrm{d}N(t)$ denote the number of events over the small interval $[t, t + \Delta t)$. The intensity $\lambda(t; H_t)$ is defined by

$$\lambda(t; H_t) = \lim_{\Delta t \to 0} \frac{\Pr\{\mathrm{d}N(t) = 1 | H_t\}}{\Delta t}. \qquad (1)$$

The probability distribution of the point process $\{N(t) : t \geq 0\}$ over $(0, \tau]$ can be given in terms of $\lambda(t; H_t)$; see [3, pp. 57–58] or [4]. In particular, the probability density that exactly n events occur over the specified interval (τ_0, τ), and at times $t_1 < \cdots < t_n$, is

$$\prod_{j=1}^{n} \lambda(t_j; H_{t_j}) \exp\left[- \int_{\tau_0}^{\tau} \lambda(t; H_t)\, \mathrm{d}t \right] \qquad (2)$$

The formula (2) provides **likelihood** functions for **maximum likelihood** estimation and associated inference procedures for models in which the intensity is specified in terms of unknown parameters. When τ_o and τ are random, it also gives **partial likelihoods**, which can be used in the same way as likelihoods, provided that τ_o and τ are determined according to a process depending only on prior event history or covariates (see [1] or [3, p.59]).

Poisson processes, where the complete intensity (hereafter just the "intensity", for brevity) is of the form $\lambda(t; H_t) = \rho(t)$, or **renewal processes**, where it is of the form $h(t - t_{N(t^-)})$, are common models. The former is convenient when counts (i.e. numbers of events in various time intervals) are emphasized; for a Poisson process, the number of events occurring in time interval $(s, t]$ has a **Poisson distribution** with mean $\int_s^t \rho(u)\, \mathrm{d}u$, and the numbers of events in disjoint intervals are independent. Renewal processes are convenient when the intervals between events are of more direct interest. For a renewal process, the times $X_1, X_2, \ldots$ between successive events are independent and identically distributed with **hazard** function $h(x)$.

Fixed or time-varying covariates may be incorporated into the intensity function. We consider here only "exogeneous" covariates that are not affected by the event process, and condition on their realized values. If z is a vector of fixed covariates, then **multiplicative models** in which $\lambda(t; H_t, z) = \lambda_o(t; H_t)\phi(z)$ are often useful. Models with

$$\lambda(t; H_t, z) = \rho(t)\phi(z) \qquad (3)$$

$$\lambda(t; H_t, z) = h(t - t_{N(t^-)})\phi(z) \qquad (4)$$

for the Poisson and renewal cases, respectively, are easy to interpret and handle statistically. In (3) and (4), $\phi(z)$ is some positive-valued function that specifies the effect of z on the intensity. Time-varying covariates $z(t)$ may be incorporated in the same way.

When we specify an intensity function (1), we have a complete model for the repeated event process. As in ordinary **regression** or longitudinal data analysis, simple analyses of means or rates that avoid strong assumptions about the recurrent event process are often attractive, however. The rate of occurrence function is defined by $r(t) = \mathrm{d}E\{N(t)\}/\mathrm{d}t$, and the mean or expected count function $R(t) = \mathrm{E}\{N(t)\}$ is thus

$$R(t) = \int_0^t r(u)\,\mathrm{d}u. \tag{5}$$

For a Poisson process, the function $r(t)$ is also the intensity function but, in general, they are quite distinct, and $r(t)$ does not fully specify the process. Covariates z can be introduced into the rate or mean functions, the simplest way being by the multiplicative specification

$$r(t; z) = r_0(t)\phi(z) \tag{6}$$

A distinction should be made between fully parametric models that are specified solely in terms of a finite-dimensional parameter θ and **semiparametric** models, which involve arbitrary rate, mean, or intensity functions. For example, if in (3), we specify $\rho(t) = \exp(\gamma_0 + \gamma_1 t)$ and $\phi(z) = \exp(\beta z)$, the model is fully parametric; if we specify $\phi(z)$ thus but leave $\rho(t)$ an arbitrary positive-valued function, the model is semiparametric.

Poisson and renewal processes are fundamental in their simplicity and ease of interpretation but in many situations fail to represent event processes satisfactorily. Three broad approaches to modeling and analysis in such cases are (i) to seek satisfactory specification of the intensity processes by building in aspects of the event and covariate histories, (ii) to add **random effects** or time-varying covariates to Poisson or renewal models, and (iii) to generalize renewal models by incorporating serial dependence among the times between events.

Other types of models are discussed in books on point processes (e.g. [10] and [20]) and on **counting processes** (e.g. [3]).

Random effects are sometimes incorporated into Poisson or renewal models in order to represent unobservable heterogeneity among subjects. For example, if events for individual i occur according to a Poisson process, the counts $N_i(t)$ have Poisson distributions with means $R_i(t)$ and variances also equal to $R_i(t)$. Larger variances than means are often observed across individuals, however, and this **"overdispersion"** may signal unobserved heterogeneity.

One way of dealing with this (e.g. [6, 14]) is to assume that given (unobservable) random effects α_i, the processes $\{N_i(t), t \geq 0\}$ are Poisson with intensity functions $\alpha_i r_i(t)$. The α_i's are assumed to be independent random variables with mean 1 and variance σ_α^2. It is easily seen that $\mathrm{E}\{N_i(t)\} = R_i(t)$ and $\mathrm{Var}\{N_i(t)\} = R_i(t) + \sigma_\alpha^2 R_i(t)^2$. In addition, the event counts in nonoverlapping time intervals (s_1, t_1) and (s_2, t_2) for the individual are now **correlated**, with covariance $\sigma_\alpha^2 R_i(s_1, t_1) R_i(s_2, t_2)$, where $R_i(s, t) = R_i(t) - R_i(s)$. It is often convenient to assume that the α_i's have **gamma distributions**, in which case the $N_i(t)$'s have **negative binomial distributions**.

Aalen and Husebye [1], Lawless and Fong [16], and others discuss the incorporation of random effects into renewal processes, and Andersen et al. [3, Chapter 9] consider general counting processes.

The modeling approach taken with repeated events necessarily depends on one's objectives. Even so, more than one approach may be supportable in a given setting. A particularly interesting situation is when the comparison of treatments is a main objective. In this case, intensity-based comparisons may mask the treatment effects, and comparisons based on marginal rate or mean functions may be preferred. However, one might be interested in the relationship of treatment to patterns in events, and not just their frequency. For discussions along these lines, see [6] and [13, Chapter 9].

Statistical Methods

Likelihood and Partial Likelihood Methods

For models with a full probability specification, maximum likelihood is in principle available, although in some circumstances, it may be difficult or indeed impossible to implement. From (2), the likelihood based on n independent subjects observed over intervals (τ_{oi}, τ_i) is of the form (suppressing dependence on covariates in the notation and writing $H_i(t)$ for the history of subject i, including covariates)

$$\prod_{i=1}^{m} \left\{ \prod_{j=1}^{n_i} \lambda_i(t_{ij}; H_i(t_{ij})) \right\} \exp\left\{ -\int_{\tau_{oi}}^{\tau_i} \lambda_i(t; H_i(t))\,\mathrm{d}t \right\} \tag{7}$$

If the information necessary to evaluate the integrals in (7) is not available (e.g. values of covariates at all time points may not be available) or if the integrals are intractable, then they have to be approximated somehow for maximum likelihood to be feasible. For simple models, maximum likelihood is discussed in many references; for example, see [10] and [20] for Poisson and renewal processes, [1] and [14] for Poisson and renewal processes with random effects, and [3, 4], and [6] for general discussion.

Maximum likelihood based on (7) is most attractive when the model is fully parametric, in which case, standard large sample results concerning estimators, score (*see* **Likelihood**), and **likelihood ratio test** statistics apply as $m \to \infty$, and these can be used to obtain **hypothesis tests** or **confidence intervals**. Semiparametric methods may in many cases also be obtained from (7), although delicate mathematical points arise in a rigorous treatment. For semiparametric models, the most common approach is via multiplicative intensity Poisson models, which are analogous to **proportional hazards** models in survival analysis. In this case, the complete intensity is assumed to be of the form

$$\lambda(t; H_t) = \lambda_0(t)\phi[z(t); \boldsymbol{\beta}], \tag{8}$$

where $\lambda_0(t)$ is an arbitrary "baseline" intensity function and $z(t)$ is a vector of fixed or time-varying covariates.

The partial likelihood method of Cox [7, 8] (see **Cox Regression Model**) applies to models of the form (8). Define $\delta_i(t) = I(\tau_i \geq t \geq \tau_{oi})$ and note that if an event is observed at time t, then under (8), the probability it occurs for subject j is

$$\frac{\delta_j(t)\phi[z_j(t); \boldsymbol{\beta}]}{\sum_{i=1}^{m} \delta_i(t)\phi[z_i(t); \boldsymbol{\beta}]}, \tag{9}$$

which is independent of $\lambda_0(t)$. The partial likelihood for $\boldsymbol{\beta}$ is the product across all event times of terms of the form (9), and can be maximized to give an estimate $\hat{\boldsymbol{\beta}}$. Tests or confidence intervals for $\boldsymbol{\beta}$ follow by standard maximum likelihood methods. The cumulative baseline intensity function $\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$ can be estimated by

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\mathrm{d}N_{\cdot}(u)}{\sum_{i=1}^{m} \delta_i(u)\phi[z_i(u); \hat{\boldsymbol{\beta}}]}, \tag{10}$$

where $\mathrm{d}N_{\cdot}(u)$ is the total number of events observed at time u. Andersen et al. [3] give a comprehensive account of these methods.

In the special case in which there are no covariates, the right side of (8) is just $\lambda_0(t)$ so the process is Poisson. Then, the estimate (10) becomes

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\mathrm{d}N_{\cdot}(u)}{\sum_{i=1}^{m} \delta_i(u)}, \tag{11}$$

which is termed the **Nelson–Aalen estimator**.

Owing to a connection with ordinary maximum likelihood, partial likelihood methodology can also be used with semiparametric renewal process models where

$$\lambda(t; H_t) = h_o[B(t)]\phi[z(t); \boldsymbol{\beta}], \tag{12}$$

where $B(t) = t - T_{N(t-)}$ is the time since the last event before t, and $h_o(s)$ is a baseline hazard function; see [11] and [19]. The models (8) and (12) can also be extended by allowing $z(t)$ to include functions of previous event history; these are sometimes referred to as modulated Poisson and renewal processes.

Estimating Function Methods for Rate and Mean Functions

The methods above are based on full probability models. Sometimes, it is attractive to model only the mean and rate functions for the repeated event, as in (5) and (6). It turns out that if the start and termination times τ_{oi} and τ_i for observation of individuals are determined independently of their event processes, then simple **robust** methods that are closely related to analysis under the Poisson model (8) can be used.

This approach follows from an observation of Nelson [23], who noticed that the Nelson–Aalen estimator (11) is a generally valid **nonparametric** estimate of a common mean function, $R(t) = \mathrm{E}[N_i(t)]$ for $i = 1, \ldots, m$. Lawless and Nadeau [17] subsequently provided methodology for parametric and semiparametric models including covariates; Pepe and Cai [25] consider related methods. Surveys of these and later developments are given in [6] and [18]. We outline here the methodology for the case of multiplicative semiparametric models, where the rate function (6) is given by

$$r(t; z) = r_o(t)\phi[z; \boldsymbol{\beta}] \tag{13}$$

with $r_o(t)$ an unspecified baseline rate function. Proofs for results stated below can be found in [17] and [21].

The key idea is to recognize that the maximum likelihood **estimating equations** for (13), which come from a Poisson process (in which case (13) is also the process intensity) are unbiased more generally, provided that the observation intervals (τ_{oi}, τ_i) are independent of the event processes. Thus, the estimates coming from these equations are consistent under the model (13), regardless of the underlying event process, and estimating equation theory may be used to give robust variance estimates and other tools for inference.

The Poisson maximum likelihood estimating equations for the semiparametric model (13) are [17]

$$\sum_{i=1}^{m} \int_o^{\tau} \delta_i(t)[\,\mathrm{d}N_i(t) - \mathrm{d}R_o(t)\phi(\boldsymbol{z}_i;\boldsymbol{\beta})] \times \frac{\partial \log \phi(\boldsymbol{z}_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \tag{14}$$

and

$$\mathrm{d}R_o(t) = \frac{\mathrm{d}N_{\cdot}(t)}{\sum_{i=1}^{m} \delta_i(t)\phi(\boldsymbol{z}_i;\boldsymbol{\beta})} \quad 0 \le t \le \tau \tag{15}$$

where $\tau = \max(\tau_i)$, we assume each $\tau_{io} = 0$ for simplicity, and $R_o(t) = \int_o^t r_o(s)ds$ is the baseline mean function. Inserting $\mathrm{d}R_o(t)$ from (15) into (14) produces the same estimating equation for $\boldsymbol{\beta}$ as given by the Cox partial likelihood for $\boldsymbol{\beta}$ based on the terms (9) in the Poisson model. Furthermore, insertion of the resulting estimate $\hat{\boldsymbol{\beta}}$ into (15) gives the generalized Nelson–Aalen estimate (10) as the estimate $\hat{R}_o(t)$. Thus, any software that implements the Cox Poisson regression model (8) for repeated events can be used to obtain the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{R}_o(t)$ arising from (14) and (15). For example, for the common model in which

$$\phi(\boldsymbol{z};\boldsymbol{\beta}) = \mathrm{e}^{\boldsymbol{\beta}'\boldsymbol{z}}, \tag{16}$$

with $\boldsymbol{\beta}$ and $\boldsymbol{z}$ both $p \times 1$ vectors, the **S-Plus** function *coxph* with the "cluster" option for individuals can be employed.

Sandwich-type robust variance estimates for $\hat{\boldsymbol{\beta}}$, $\hat{R}_o(t)$, and estimated mean functions

$$\hat{R}(t;\boldsymbol{z}) = \phi(\boldsymbol{z};\hat{\boldsymbol{\beta}})\hat{R}_o(t) \tag{17}$$

can also be given ([17, 21]). Current **software** packages give the robust variance estimates for $\hat{\boldsymbol{\beta}}$ only, but more is planned for a forthcoming version of SAS.

In the case where there are no covariates, the common mean function $R_o(t) = \mathrm{E}[N_i(t)]$ is estimated from (15) with $\phi(\boldsymbol{z};\boldsymbol{\beta}) = 1$, which is the Nelson–Aalen estimator (11). For non-Poisson processes, the Poisson variance estimate for $\hat{R}_o(t)$ should be replaced with the robust estimate

$$\widehat{\mathrm{V}}\mathrm{ar}\{\sqrt{m}[\hat{R}_o(t) - R_o(t)]\} = \sum_{i=1}^{m} \left\{ \int_o^t \frac{\delta_i(s)}{\delta_{\cdot}(s)} \left[\mathrm{d}N_i(s) - \frac{\mathrm{d}N_{\cdot}(s)}{\delta_{\cdot}(s)} \right] \right\}^2, \tag{18}$$

where $\delta_{\cdot}(s) = \sum_{i=1}^{m} \delta_i(s)$.

Additional Considerations

It is, in general, important to consider the process that determines the observation periods $(0, \tau_i]$ for the various subjects. It has been noted that the robust methods require that the τ_i's are determined independently of the event processes, whereas for maximum likelihood methods based on a full intensity specification, the τ_i's may be determined randomly in a way that depends upon past event history but not on future events. Andersen et al. [3, Section 2.7] and Aalen and Husebye [1] give precise requirements for the observation periods (τ_{oi}, τ_i) of individual subjects. A practical requirement for the use of (7) is that the history $H_i(t)$ needed to compute the intensity function at time points in (τ_{oi}, τ_i) be available. Finally, the methods above do not require that the censoring or observation processes for individuals be modeled. If censoring can be modeled satisfactorily, it may be possible to use censoring weights (e.g. see [6] and [26]) to modify estimating functions for rate and mean functions so as to allow history-dependent censoring.

We now mention a few other topics of practical interest. First, events are sometimes recorded only at intermittent times, resulting in counts for specified time intervals rather than exact event times; see [6, Section 2.3] for a review and references to methodology. Secondly, in some settings, the repeated events process ceases with the occurrence of some terminating event, with which it may be associated; for example, events may be associated with the treatment of a specific illness and terminate when the illness ends or the patient dies. This area is reviewed in

[6, Section 5]. A third point is that we have discussed the analysis of one specific type of event. Multiple event types $j = 1, \ldots, J$ can be handled via intensity-based modeling by setting up counting processes $N_j(t)$ and complete intensities $\lambda_j(t; H_t)$ for each event type [3]. This allows association between event types through dependence of the intensities on prior events in H_t. Other approaches discussed in this article can also be extended to handle multiple events, in particular, the robust methodology for rate and mean functions [24].

We have emphasized **multiplicative models** like (8), (12), and (13). Multiplicative models for event occurrence are convenient, flexible, and easily interpreted, but other types are also useful in many settings. Fully parametric models of other types are easily handled, but semiparametric models can require some additional development beyond what has been given here. For example, see [13, p.289] and [22] for the case of **accelerated failure-time** models for rate and mean functions.

Finally, we have not discussed the very important topic of **model checking**. Specific methodology depends to some extent on the types of models under consideration, but a universal approach is to test base models against "expanded" models having additional parametric structure. **Graphical displays** involving the raw data, nonparametric and parametric estimates, and suitably defined **residuals** are also important. Andersen et al. [3] and Therneau and Grambsch [27] provide numerous examples for multiplicative intensity models.

The following rather simple example illustrates several of the points about modeling and methodology that have been discussed earlier. For more complex settings involving multiple covariates and more complicated observational patterns, see for example, illustrations and discussion in [6, 13, Chapter 9] and [27, Chapter 8].

An Illustrative Example

Gail et al. [12] presented data on the times to development of mammary tumors for 48 female rats in a carcinogenicity experiment (*see* **Tumor Incidence Experiments**). The animals were assigned randomly to two groups, treatment (23 animals) and control (25 animals), and the days on which new tumors occurred for each animal were recorded. All animals were observed over the time period $(0, \tau] = (0, 122]$ days. Although the data are given in discrete time units (days) and, in fact, animals were inspected for the presence of new tumors every two to four days, we will, for simplicity, treat the occurrence times as continuous; this does not affect the main conclusions; see [17] for methodology written in terms of discrete time processes. The main objective of the experiment was to compare the frequency of tumor occurrence for treatment and control animals.

A useful place to start is to plot nonparametric estimates (11) of the mean functions for animals in the two treatment groups. Figure 1 shows this and indicates that (i) there is no pronounced trend in the

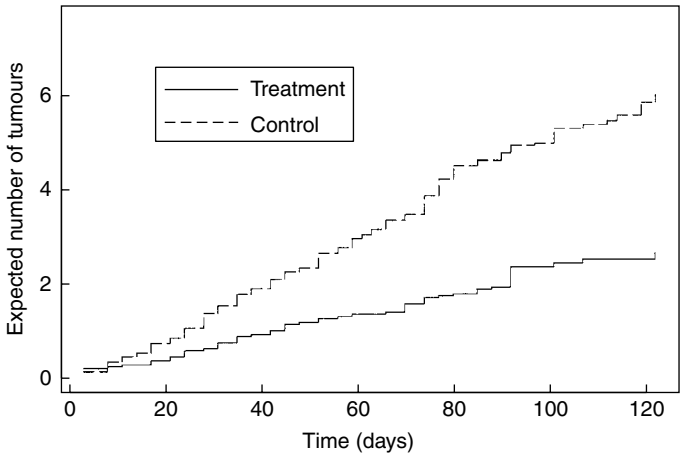

Figure 1 Estimates of mean functions for treatment and control groups

tumor occurrence rate for either group, and (ii) the rate and the expected number of tumors for control animals is about twice that for treatment animals.

To make a more thorough comparison, we might consider fitting Poisson processes to the events for the individual animals. However, there is evidence of overdispersion, particularly in the control group. Under a Poisson model, the total number of tumors $N_i(122)$ for each animal has a Poisson distribution. However, the sample means and variances for the treatment and control group animals are $\overline{N}_T = 2.65$, $s_T^2 = 3.62$ and $\overline{N}_C = 6.04$, $s_C^2 = 14.92$, respectively. Formal tests [14] provide evidence against the Poisson model.

To compare the rate functions for the two groups, Figure 1 suggests the multiplicative model (6) with $\phi(z) = \exp(\beta z)$, where $z_i = 0$ if animal i is in the control group and 1 if it is in the treatment group. Robust estimation of β, as described above (see (14) and (15)) and in [15] and [17] yields $\hat{\beta} = -0.82$ with a standard error of 0.21. This indicates a rather strong treatment effect; the tumor rate for treatment animals is estimated to be $\exp(-0.82) = .44$ times that for control animals. Inference based on a Poisson model with individual random effects for each animal [14] yields the same result. However, if we ignored the overdispersion and fitted Poisson processes to the two groups, we would obtain the same estimate $\hat{\beta} = -0.82$, but a smaller standard error of .15, thus overstating the strength of the treatment effect. A partial likelihood analysis yields the same inference for β as the Poisson model, and thus also overstates the treatment effect.

Other models may be considered for these data. For example, Gail et al. [12] fit renewal models in which times between successive events (tumors) for an animal are considered independent. Graphical and more formal methods may be used to check models that assume a specific probabilistic structure for the event processes (e.g. see [3, 10, 14, 27]). Plots involving both residuals and the raw data (e.g. [10]) are especially helpful.

References

[1] Aalen, O.O. & Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes, *Statistics in Medicine* **10**, 1227–1240.

[2] Albert, P.S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts, *Biometrics* **47**, 1371–1381.

[3] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Berman, M. & Turner, T.R. (1992). Approximating point process likelihoods with GLIM, *Applied Statistics* **41**, 31–38.

[5] Cheuvart, B. (1988). A nonparametric model for multiple recurrences, *Applied Statistics* **37**, 157–168.

[6] Cook, R.J. & Lawless, J.F. (2002). Analysis of repeated events, *Statistical Methods in Medical Research* **11**, 141–166.

[7] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society B* **34**, 187–202.

[8] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[9] Cox, D.R. & Isham, V. (1980). *Point Processes*. Chapman & Hall, London.

[10] Cox, D.R. & Lewis, P.A.W. (1966). *The Statistical Analysis of Series of Events*. Methuen, London.

[11] Dabrowska, D.M., Sun, G. & Horowitz, M.M. (1994). Cox regression in a Markov renewal model: an application to the analysis of bone marrow transplant data, *Journal of the American Statistical Association* **89**, 867–877.

[12] Gail, M.H., Santner, T.J. & Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor, *Biometrics* **36**, 255–266.

[13] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. John Wiley & Sons, New York.

[14] Lawless, J.F. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association* **82**, 808–815.

[15] Lawless, J.F. (1995). The analysis of recurrent events for multiple subjects, *Applied Statistics* **44**, 487–498.

[16] Lawless, J.F. & Fong, D. (1999). State duration models in clinical and observational studies, *Statistics in Medicine* **18**, 2365–2376.

[17] Lawless, J.F. & Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events, *Technometrics* **37**, 158–168.

[18] Lawless, J.F., Nadeau, C. & Cook, R.J. (1997). Analysis of mean and rate functions for recurrent events, in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D.Y. Lin & T.R. Fleming, eds. Springer-Verlag, New York, 37–49.

[19] Lawless, J.F., Wigg, M.B., Tuli, S., Drake, J. & Lamberti-Pasculli, M. (2001). Analysis of repeated failures or durations, with application to shunt failures for patients with paediatric hydrocephalus, *Applied Statistics* **50**, 449–465.

[20] Lewis, P.A.W. ed. (1972). *Stochastic Point Processes*. John Wiley & Sons, New York.

[21] Lin, D.Y., Wei, L.J., Yang, I. & Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events, *Journal of the Royal Statistical Society B* **62**, 711–730.

[22] Lin, D.Y., Wei, L.J. & Ying, Z. (1998). Accelerated failure time models for counting processes, *Biometrika* **85**, 605–618.

[23] Nelson, W.B. (1988). Graphical analysis of system repair data, *Journal of Quality Technology* **20**, 24–35.

[24] Ng, E. & Cook, R.J. (1999). Robust inference for bivariate point processes, *Canadian Journal of Statistics* **27**, 509–524.

[25] Pepe, M.S. & Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates, *Journal of the American Statistical Association* **88**, 811–820.

[26] Strawderman, R. (2000). Estimating the mean of an increasing stochastic process at a censored stopping time, *Journal of the American Statistical Association* **95**, 1192–1208.

[27] Therneau, T.M. & Grambsch, P.M. (1999). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

J.F. LAWLESS

Reproduction Number

The basic reproduction number (or ratio), R_0, is the expected number of secondary cases of an infection which one typical case could generate during its infectious period in a completely susceptible population. It is the most important theoretical concept in communicable disease epidemiology. If R_0 is greater than one, then a newly introduced infection may lead to a large epidemic in a completely susceptible population and a stable endemic level may persist if the host population is large enough such that new susceptibles are born into the population at a sufficiently high rate. If R_0 is less than one, then the total size of a newly introduced outbreak will remain small (*see* **Epidemic Thresholds**). In addition to this qualitative threshold property, R_0 has an important quantitative interpretation because it allows one to determine the amount of effort needed to reduce the incidence of a disease to zero: in a homogeneously mixing population the lower bound, c^*, for the necessary coverage with a vaccine that is 100% efficacious is given by the simple expression $1 - 1/R_0$, because for this coverage the number of secondary cases of one infective is reduced to the critical value of one $[R_0(1 - c^*) = 1]$ (*see* **Vaccine Studies**). This practical implication motivates epidemiologists to estimate R_0 from data about epidemics or endemic equilibria.

The notion, which has its origin in **demography**, was first introduced into the epidemiology of infectious diseases by Macdonald [13] in the context of malaria. Smith [15] applied it to the control of arboviruses, i.e. viruses which are transmitted from man to man or from animal to man by arthropods (e.g. yellow fever by mosquitoes). There is a parallel sequence of papers, started by Bharucha-Reid [5], followed by Neyman & Scott [14] and Bartoszýnski [3], in which the spread of infectious diseases in large populations is approximated by **branching processes**. In 1975 the concept was introduced independently and under different names in the context of directly transmitted virus diseases by Becker [4], Dietz [7], and Hethcote [12]. Currently, only the symbol has been standardized in epidemic theory, but R_0 is still being called by several names. The most natural one would be "basic reproduction *ratio*" or "basic reproduction *number*", since R_0 is a dimensionless quantity and does not deserve the affix "rate", which suggests a dimension "time^{-1}". A Dahlem Workshop [1] helped tremendously to popularize R_0 and the seminal paper by Diekmann et al. [6] provides a mathematically rigorous framework for its definition. The key reference is the dissertation of Heesterbeek [9], with its remarkably short title: "R_0". The most recent survey article is Heesterbeek & Dietz [11]. For an excellent description of the history of R_0, see review article by Heesterbeek [10].

Because of its important implications, it is essential to estimate R_0 on the basis of epidemiological data. A survey of estimation methods is given by Dietz [8]. R_0 depends on three parameters: (i) the contact rate, (ii) the duration of the infectious period, and (iii) the probability that a contact between an infective and a susceptible individual leads to an infection. The last parameter is sometimes broken down into two factors describing infectivity and susceptibility. For a homogeneously mixing population (*see* **Random Mixing**), the equilibrium proportion of susceptibles equals the inverse of the basic reproduction number, because in this situation on average one case produces one secondary case (*see* **Secondary Attack Rate**). For infections with lifelong immunity, the average age at first infection divided by life expectancy equals the proportion of susceptibles in the population, because this equals the fraction of life before the infection. Therefore, one can estimate a basic reproduction number by the ratio of life expectancy over the average age at first infection [7]. The comprehensive reference work on the mathematical approach to the epidemiology and control of human infectious diseases by Anderson & May [2] contains numerous estimates of R_0 for a wide variety of diseases in different geographical regions.

One can also estimate R_0 from the final size of an epidemic if one knows the number of individuals infected during the epidemic and the final proportion of individuals still susceptible. The following formula provides an estimate of R_0 for a homogeneously mixing population:

$$R_0 = (u_0 - u_\infty)^{-1}(\ln u_0 - \ln u_\infty),$$

where u_0 denotes the initial and u_∞ the final proportion of susceptibles.

The estimates of R_0 are highly dependent upon the assumptions about the contact structure in the population and the effects of immunity. If one only has

age-specific antibody prevalence data, it is impossible in principle to deduce the underlying mixing matrix. One frequent simplifying assumption is proportional mixing, which associates with each individual a certain contact rate. The probability that a contacted individual is infectious is then a weighted average of the prevalence using the contact rates as weights.

In spite of impressive theoretical progress, the practical application of R_0 is still in its infancy because of the intrinsic difficulties in applying this theoretical concept to real-life situations.

References

[1] Anderson, R.M. & May, R.M., eds. (1982). *Population Biology of Infectious Diseases: Dahlem Konferenzen*. Springer-Verlag, Berlin.

[2] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[3] Bartoszýnski, R. (1969). *Branching Processes and Models of Epidemics*, Dissertationes Mathematicae. Panstwowe Wydawnictwo Naukowe, Warsaw.

[4] Becker, N.G. (1975). The use of mathematical models in determining vaccination policies, *Bulletin of the International Statistical Institute* **46**, 478–490.

[5] Bharucha-Reid, A.T. (1956). On the stochastic theory of epidemics, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 5: *Biology and Problems of Health*, J. Neyman, ed. University of California Press, Berkeley, pp. 111–119.

[6] Diekmann, O., Heesterbeek, J.A.P. & Metz, J.A.J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations, *Journal of Mathematical Biology* **28**, 365–382.

[7] Dietz, K. (1975). Transmission and control of arboviruses, in *Epidemiology: Proceedings of a SIMS Conference*, D. Ludwig & K.L. Cooke, eds. SIAM, Philadelphia, pp. 104–121.

[8] Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases, *Statistical Methods in Medical Research* **2**, 23–41.

[9] Heesterbeek, J.A.P. (1992). R_0. Centrum voor Wiskunde en Informatica, Amsterdam.

[10] Heesterbeek, J.A.P. (2002). A brief history of R_0 and a recipe for its calculation, *Acta Biotheoretica* **50**, 189–204.

[11] Heesterbeek, J.A.P. & Dietz, K. (1996). The concept of R_0 in epidemic theory, *Statistica Neerlandica* **50**, 89–110.

[12] Hethcote, H.W. (1975). Mathematical models for the spread of infectious diseases, in *Epidemiology: Proceedings of a SIMS Conference*, D. Ludwig & K.L. Cooke, eds. SIAM, Philadelphia, pp. 122–131.

[13] Macdonald, G. (1952). The analysis of equilibrium in malaria, *Tropical Diseases Bulletin* **49**, 813–829.

[14] Neyman, J. & Scott, E.L. (1964). A stochastic model of epidemics, in *Stochastic Models in Medicine and Biology*, J. Gurland, ed. University of Wisconsin Press, Madison, pp. 45–83.

[15] Smith, C.E.G. (1964). Factors in the transmission of virus infections from animals to man, *Scientific Basis of Medicine, Annual Review*, 125–150.

(*See also* **Epidemic Models, Deterministic**; **Epidemic Models, Stochastic**)

K. DIETZ

Reproduction

Human reproduction, fertility, sexuality, and family planning are central to our lives, and research into these areas has been under way for years. While this research uses methodologies and techniques that are common to many other branches of behavioral and biomedical science, there are a number of biostatistical challenges that are unique to the field. The main biostatistical problem is the assessment of fertility and pregnancy rates in different groups of users – ranging from men and women using highly effective methods to reduce their fertility, those using less effective methods, those living in "natural fertility" situations in which no fertility regulation methods are used at all, or to infertile couples who may be trying to increase their fertility. With a greater understanding of the factors that affect human fertility and the different ways in which these factors might be modified, clearer ideas have emerged on the most appropriate statistical measures to summarize information. Similarly, we now appreciate better the limitations of different research designs, the information generated by such designs and the types of data that need to be recorded.

The main biostatistical challenge is to summarize information on the effectiveness of different contraceptive methods in preventing pregnancy, the adverse and beneficial secondary effects associated with their use, and the reasons why people stop using their chosen method, switch to another method, or stop using any contraceptive method at all. Ideally, this information would be presented in such a way that an individual could freely choose the most appropriate method for his or her situation and reproductive intentions, and that this choice could be made with the fullest breadth and depth of knowledge available. In addition, the provider needs to know the risks, benefits, advantages, and disadvantages of different methods of fertility regulation in order to give appropriate advice and counseling on the most suitable methods to use. Similarly, policy makers, responsible for ensuring reproductive health and family planning services, need to have information relevant to their country or populations. Thus, not only is comprehensive information required on the characteristics of different methods, their reliability under ideal and typical conditions of use, their side effects, the type of advice, and the counseling to be given to potential or existing users, but also the extent to which this information can be generalized to women and men in different personal and social situations.

Comparisons between contraceptive methods within the same broad class are in general more straightforward and reliable than comparisons between widely different methods. For example, the comparative efficacy and rates of side effects of two types of intrauterine device (IUD) are relatively easy to establish, and to generalize to populations other than those studied. By contrast, comparative statements about methods that are widely different, for example periodic abstinence or withdrawal (coitus interruptus) and long-term methods such as five-year implants or sterilization, are much more problematic. However, this is exactly the information required.

To appreciate the biostatistical problems encountered in the study of human reproduction, fertility, and family planning, an understanding is necessary of the complex social, behavioral, and physiological processes that determine human fertility. A great deal is known about these and has been presented from different perspectives, and the following briefly summarizes those aspects that give rise to the unique and challenging problems for the biostatistician. Further details can be found in many good textbooks or reviews (see, for example, Gray et al. [8], Hatcher et al. [11], and Nieschlag & Behre [14]).

Key Factors in Male and Female Reproductive Physiology

The Female Partner

The physiological factors which determine whether a woman becomes pregnant include her ability to conceive (fecundity), the fertilizing capacity of her partner, the timing and frequency of intercourse, and the chances that the fertilized ovum will be implanted. Her fecundity is mainly determined by ovulation – whether or not it occurs and its regularity. These can be affected by many factors: ovulation only occurs naturally during the reproductive years, except during and soon after a pregnancy or during full breast feeding, and its regularity decreases with age. It can also be affected by extreme malnutrition or obesity, other physiological disorders, smoking, and strenuous exercise.

Successful fusion of oocyte and sperm and subsequent implantation require patent fallopian tubes which effectively transport the ovum from the ovary to the uterus. These may be adversely affected by pollutants such as smoking or damage caused by sexually transmitted diseases or other infections.

The regular human menstrual cycle has usual length of 28–30 days, although there is considerable variation in length both between women and within the same woman. The normal cycle is divided into the follicular phase, which leads up to ovulation around 14–18 days after the onset of menses, and the luteal phase, which lasts about 14 days. The endometrium (the lining of the uterus) is prepared to receive the fertilized ovum, and, in the absence of a successful implantation, is sloughed off at the onset of the next menstrual cycle. Typically an ovum can be fertilized up to one or two days following ovulation. A number of physiological changes, some of which can be comparatively easily observed, occur around the time of ovulation as the hormonal balance shifts from the follicular to the luteal phase. Examples include changes in cervical mucus quantity and quality around ovulation, a rise in basal body temperature (BBT) between 0.2 °C and 0.4 °C, or changes in the ratios of urinary metabolites that reflect circulating hormone levels.

There is a delay from the actual time of fertilization to the recognition of a pregnancy. Implantation occurs about six days following fertilization, at which time it is possible with very sensitive urinary assays to detect early signs of pregnancy [2]. Less sensitive tests, such as home pregnancy urinary testing kits, are only reliable two or more weeks after fertilization. Usually, the first recognizable sign of pregnancy is a delay in the expected onset of menstruation, although this is not a very specific indicator, as there are many other factors that can be responsible for menstrual delay. Clinical signs of pregnancy other than lack of menstruation are not apparent before six weeks.

The Male Partner

By contrast to the woman, who only ovulates once in each cycle, sperm production is a continuous process. The sperm are formed in the testes and stored in the epididymis ready to be released during ejaculation. A typical ejaculate contains between 100 million and 400 million sperm, only a small proportion of which are viable, motile, and able to swim towards the ovum. Only a few hundred actually reach the ovum and only a single sperm is required for fertilization. Sperm can survive in the female genital tract up to four or five days following ejaculation. The spermatogenic cycle (time from the formation of spermatogonia to the production of mature sperm capable of fertilization) takes about 90 days.

The factors that affect male fertility are less well understood than for the female. Male fertility declines naturally with age, and a number of environmental pollutants and toxic exposures can reduce sperm production and quality. In addition, there can be partial or complete blockage of the epididymis or vas deferens due to sexually transmitted diseases or other infections.

Opportunities for Contraceptive Methods

There are many factors that can disrupt the normal menstrual cycle, introduce variability, or prevent ovulation, fertilization, or implantation. Methods of fertility regulation target different stages of the process. For example, the method of periodic abstinence restricts intercourse to the infertile phases of the cycle; female hormonal methods reduce fecundity by blocking ovulation or implantation or changing the quality of cervical mucus to prevent the sperm ascending the genital tract. Barrier methods prevent contact between sperm and the ovum, while male hormonal methods reduce the production of spermatozoa. Sperm production is a complex process of precisely timed stages, any of which can be targets for the development of new methods of fertility regulation (see, for example, Hamilton & Saling [9]).

Some contraceptive methods act continuously and are permanent (e.g. male and female sterilization), some act continuously but are reversible (e.g. IUDs and hormonal methods), while others are used only around the time of intercourse (e.g. spermicides or barrier methods such as condoms and diaphragms). The coitus dependent methods require a high degree of compliance with correct method use to prevent pregnancy reliably, and it is difficult to separate the intrinsic efficacy of the methods from those factors that determine whether the method is used at all, and, if used, whether it is used correctly.

Biostatistical Challenges

The delay in recognition of pregnancy and the high rate of early pregnancy loss cause problems for estimating pregnancy rates and the efficacy of different contraceptive methods in **cohort studies**. First, when the decision to stop using a method is made, the woman may already be pregnant, although the pregnancy may not be recognized until later. It is unclear whether such an event should be counted as a contraceptive failure, or according to the expressed reason why the woman stopped using the method, or both. Some studies require an additional visit six to eight weeks after discontinuation of the method or release from the study to ensure that all pregnancies, including those not recognized at the time of discontinuation, are recorded. The second, though related, problem is to determine when fertilization actually occurred. In a cycle that does not result in pregnancy, menstruation usually starts about 14 days after ovulation, so the timing of ovulation can be made retrospectively. However, in a cycle in which fertilization and subsequent implantation occur, the exact time of ovulation cannot be determined accurately, unless there is daily monitoring of follicular growth or urinary metabolites [31]. In the absence of such methods, 14 days are usually added to the date of last menstrual period to give an "estimated date of conception", although this can be substantially in error if menstrual cycles are not regular. The third problem concerns the distinction between early "chemical" pregnancies detected by sensitive assays and clinically recognized pregnancies which can only be detected much later. About 20% of fertilized ova never reach the stage of implantation and a further 10% are lost before they result in clinically recognizable pregnancies. Thus, without a clear and consistent definition of pregnancy, rates reported from different studies are difficult to compare.

Some methods of fertility regulation are immediately effective while others require a delay before they become effective. Barrier methods, such as condoms or spermicides, which prevent contact between the sperm and ovum, are immediately effective and can be immediately reversed. Similarly, the IUD is effective from the moment it is inserted and its effect is rapidly reversed following removal. By contrast, currently available methods of male fertility regulation have a delayed effectiveness as the sperm production is stopped and the extra-testicular reserves of sperm are exhausted. This occurs with vasectomy, which prevents the sperm reaching the ejaculate, as well as with hormonal methods, which can reversibly suppress sperm production.

Many social, environmental, and personal factors determine which contraceptive methods are chosen and whether they are used correctly and consistently. These factors include the user's age, marital status, educational level, socioeconomic status, reproductive intentions, actual and perceived risks associated with use of different methods, access to appropriate family planning services, attitudes to unwanted pregnancy, and access to safe and appropriate methods for the termination of pregnancy. The interactions between these factors are complex and unique to each individual and vary according to time, personal circumstances, and place. While the goal is to develop biostatistical methods to assess contraceptive efficacy that can be generalized to other users, these factors are critically important in determining overall performance of a method. Moreover, the advantages and disadvantages of available methods will be weighed differently as individual circumstances evolve.

Historical Development

Effectiveness of a Method which is not Coitus-Related

The first comprehensive undertaking to record information systematically on large scale use of a contraceptive method was the Cooperative Statistical Program (CSP) for the Evaluation of Intrauterine Devices, which was established in the USA by the National Committee on Maternal Health in 1963 and supported by the Population Council. Although the IUD had been known and used since the 1930s, the CSP was the first attempt to evaluate the safety, effectiveness, and acceptability of a newly introduced contraceptive method by analysis of pooled data using uniform procedures and a systematic statistical approach. In 1970, the CSP reported on 23 917 insertions with 10 different devices [23] and showed features now known to be common to many different contraceptive methods. In particular, pregnancy and expulsion rates of the devices were higher in the first compared with subsequent years of use, pregnancy rates were higher among younger women, and removals due to medical reasons (primarily

complaints of increased pain and/or menstrual bleeding) were more common than removals for personal reasons. Thus, even with the early IUDs, their contraceptive efficacy was high (cumulative two-year pregnancy rates in the range 3–17%, compared with over 90% among women not using any method to avoid pregnancy), there were technical problems with the devices remaining in place (cumulative expulsion rates up to 30% at two years) but that "side effects" were the main factors leading to device removal (cumulative rates 15–40% at two years). About 60% of women still had the device in place two years after insertion.

Effectiveness of a Coitus-Dependent Method

The IUD is a device that is continuously effective (provided that it remains in place), does not interfere with sexual intercourse, and does not require any particular action by the user for its effectiveness. Thus the method effectiveness is the same as its effectiveness under typical conditions of use. By contrast, periodic abstinence methods require no external technology or devices. All rely for their effectiveness on rules of when to abstain from sexual intercourse in order to avoid pregnancy, and users may consciously or unconsciously depart from these rules. If the rules are not followed correctly, the efficacy of such coitus-dependent methods is much reduced, and this difference is reflected in two distinct measures of contraceptive efficacy – the method effectiveness under conditions of perfect use, and the method effectiveness under conditions of typical use. The distinction is important for methods which require a high degree of user compliance.

The recognition that fertile and infertile phases of a woman's menstrual cycle could be used to avoid pregnancy had been known since the 1930s [12, 15], but little work was done prior to 1970 with natural family planning (NFP) methods to study their effectiveness or investigate how their reliability and ease of use could be improved. In 1976, the **World Health Organization** initiated a five-country study of the ovulation method, which is based on recognizing changes in quantity and quality of cervical mucus to identify the fertile period [32]. A total of 869 volunteers of proven fertility with regular menstrual cycles kept daily records of menstruation, cervical mucus secretions, and acts of intercourse. Couples were instructed to abstain from intercourse during menstruation (because of possible early onset of the fertile period during the last days of menstrual bleeding), on alternate "dry days" prior to the onset of the fertile period (to minimize the difficulty of recognizing the onset of mucus secretion because of the presence of seminal fluid), and during the fertile period, which began on the first day of mucus secretion or the sensation of dampness or wetness detectable at the vulva. The "peak day" was defined as the last day on which fertile type mucus was recognized, and intercourse could be resumed on the fourth day after the peak day. Couples were thus required to abstain from intercourse for about half the menstrual cycle.

Almost 95% of women were able to identify accurately the fertile period of the menstrual cycle after an initial three-month training phase, and these entered a 12-month effectiveness phase. The cumulative discontinuation rate at the end of the effectiveness phase was 35.6%, the most common reasons for discontinuation being pregnancy (cumulative rate 19.6%) or desire for pregnancy (6.6%). Pregnancies were classified as method-related (all the rules for the method had been followed, and the peak day had been correctly identified), inadequate teaching or application of instructions (the record was only partially completed, the woman did not fully understand the method, had difficulty in recognizing the onset of wet days, or was confused as to which day following the peak was safe for resumption of intercourse), conscious departure from the rules (the couple knowingly made a decision to have intercourse, despite indications of fertility), or of uncertain classification. A total of 130 pregnancies were observed over 7514 cycles (22.5 pregnancies per 100 woman-years (*see* **Person-years at Risk**), assuming 13 cycles per year), of which 16 (2.8 per 100 woman-years) were method related, 22 (3.8 per 100 woman-years) were due to inadequate teaching or application of the method, 89 (15.4 per 100 woman-years) were due to conscious departure from the rules, and the remaining three were unclassified. The low pregnancy rate when the method was correctly used has stimulated further research into improved methods of identifying the fertile phase of the cycle, based for example on urinary metabolites [5, 13]. These would not require the couple to abstain from intercourse on alternate days prior to the onset of mucus secretion, and may be able to detect more accurately the timing of ovulation. They require less training in the recognition of

cervical mucus changes and no genital touching and therefore may be more accessible to a wider range of users. They may also require fewer days of abstinence and thus be less likely to result in departures from the rules and subsequent "user failures".

Method Failure or User Failure

Unfortunately, the method of analysis of pregnancy rates and assessment of contraceptive efficacy in this pioneering report were incorrect. The authors used the same denominator (total number of cycles of exposure) to compute all rates, but they divided pregnancies according to the different reported patterns of intercourse. The correct analysis requires each cycle also to be classified according to the reported pattern of intercourse. The number of pregnancies that occurred during correct method use should be compared with the number of cycles during which the method was correctly used, to derive the pregnancy or failure rate for correct method use. Similarly, the pregnancies that occurred following conscious departure from the rules need to be compared with the number of cycles when the rules were not followed, and not with the total number of cycles (*see* **Denominator Difficulties**). The only correct result in the original report is the overall pregnancy rate (22.5 pregnancies per 100 woman-years), the other rates given all being underestimates since the denominator is overestimated. However, the overall pregnancy rate is the most difficult to interpret and to generalize to other potential users of the method.

A major problem encountered with classifying each cycle according to the pattern of intercourse and rule breaking is that couples may be more likely to report acts of intercourse contrary to the rules in cycles in which a pregnancy in fact occurred, which would result in an overestimate of the pregnancy rate. For example, suppose there were 100 cycles during which intercourse occurred on a given fertile day, and these resulted in 30 pregnancies. If there is correct reporting of acts of intercourse in the pregnancy cycles, but only half of the acts in cycles that did not result in pregnancy are reported, the 30 pregnancies would apparently have occurred in 65 instead of 100 cycles. It is impossible to judge the extent to which this problem occurred with the WHO study, but the data collection methods were designed to minimize any such underreporting – volunteers were asked to maintain records on a daily basis and were visited by study staff monthly. In many cycles the visits would have taken place and the charts reviewed before any pregnancy had been recognized.

A subsequent reanalysis of the WHO study effectiveness phase data [24] showed that 16 pregnancies occurred in 6683 cycles of correct use, resulting in a "method failure" rate of 3.1 per 100 woman-years, about 12% higher than the originally computed method failure rate. However, the remaining 114 pregnancies occurred in 801 cycles in which rule breaking had been recorded, resulting in a pregnancy rate of 14.2% *per cycle* or 185.0 per 100 woman-years. Compare this result with the original report of 15.4 pregnancies per 100 woman-years for conscious departure from the rules. It is clear that the ovulation method does result in a low pregnancy rate when used correctly, but it is very unforgiving of any departure from the rules.

Overall Pregnancy Rate

The overall pregnancy rate is a weighted average of method failure rate and user failure rate, with weights proportional to the number of cycles of each type observed in the study. It is therefore difficult to interpret and generalize to other groups who may have a different proportions of perfect and imperfect use cycles. Better counseling of users, better teaching of how to use the method, different personal circumstances, and attitudes to an unplanned pregnancy would all change these proportions. Moreover, in the same population or group of volunteers these proportions may also be different, as the consequences of departures from the rules are known to them. Thus, not only can the overall pregnancy rate not be generalized to other groups, it does not even apply to a further study in the same cohort of volunteers! The other measures of failure – the pregnancy rate during perfect use, or the pregnancy rate when the rules were not followed – may be more applicable to other groups of users *when they adopt that particular pattern of intercourse*. The only assumption required to generalize these particular pregnancy rates to other groups is that the women have similar fecundity rates and their partners' fertility is comparable. We cannot be assured that this will be the case, and a rough assessment can only be made using indirect indicators of fertility. The women were of proven fertility, had regular menstrual cycles of length 23–35 days, were aged 34.5

(sd 6.2) years and their partners were aged 30.1 (sd 4.6) years. Thus the volunteers in the WHO study were selected so as to insure that only fertile couples were included, although the fertility of those who did not get pregnant during the study can only be presumed.

Pregnancy Rate Summary

During correct use of the ovulation method there were 16 pregnancies in 6683 cycles. This can either be expressed as a Pearl rate (usually pregnancies per 100 woman-years) or as the percentage of women who conceive within one year. Both statistics assume that the risk of conception in each cycle is constant (implying a **geometric distribution** for the number of cycles to conception) and when the **incidence rate** is low give very similar results. It is conventional to consider 13 cycles per year – in the WHO study the **median** cycle length was 27.7 days, corresponding to an average of 13.2 cycles per year. Thus the Pearl rate for correct method use is $16/6683 \times 1300$ or 3.11 pregnancies per 100 woman-years, with 95% **confidence limits** from the **Poisson distribution** (1.78, 5.05) pregnancies per 100 woman-years. The proportion of women conceiving within 13 cycles is $1 - (1 - p)^{13}$, where p is the probability of conception per cycle and gives the cumulative rate 3.07% (1.76%, 4.94%). When the incidence rate is low there is very little difference between the two statistics and it is often incorrectly assumed, by analogy with percentages, that the Pearl rate must lie in the range 0–100. However, when events are not rare, the two statistics give very different results. In the 801 cycles during which the rules were broken there were 114 pregnancies. Thus 14.2% (11.9%, 16.8%) of cycles resulted in pregnancy with corresponding Pearl rate 185 (155, 219) pregnancies per 100 woman-years. By contrast, the cumulative percentage of women conceiving within 13 cycles is 86.4% (80.7%, 90.9%). Thus the choice of statistic is important and the cumulative proportion conceiving within a specified period is preferred. It is also similar to the cumulative **life table** rate which is used to assess method failure and method discontinuation rates. However, the Pearl rate is simple to calculate and interpret, can be generalized to different types of exposure, and gives similar results when the incidence is rare. It thus has its uses, particularly when the incidence is low.

Constant or Decreasing Risks

Both statistics (Pearl rate and cumulative proportion conceiving within one year) assume that the **risk** (or **hazard rate**) of pregnancy in each cycle is constant, but, while this assumption may be appropriate for an individual, it is in general not true for a **cohort** of users. The cohort can be considered to consist of two types of women – those who adhere to the rules and those who do not. Since those who break the rules have a higher pregnancy rate than the others, they will drop from the risk set at a faster rate than those who adhere to the rules. After a number of cycles, the remaining cohort or risk set will have a smaller proportion of rule breakers and thus the cohort pregnancy risk will be lower. The problem of heterogeneity in time to event data is discussed in detail by Aalen [1]. This decreasing incidence rate applies not only to pregnancies in NFP studies, but also to other endpoints and other contraceptive methods. For example, in a cohort of IUD users, the younger, higher fertility women will become pregnant earlier, those prone to expel the device will drop from the cohort earlier, and those intolerant of or susceptible to side effects, such as menstrual disturbances, will discontinue method use earlier. Thus the study cohort changes in composition as the study progresses and event rates for all types of events decline with time. Within a limited time interval the cohort incidence rate may be constant, and the Pearl rate can provide a good summary measure in each interval. Similarly, annual cumulative life table rates, conditional on being at risk in the interval, can be computed to show how the incidence rate changes with time [33].

In the WHO NFP study, the overall pregnancy rate did decline with time, but the incidence rates were more nearly constant in the subgroups of women who adhered to the rules and among those who departed from the rules [24], illustrating how the changing composition of the cohort affects the overall incidence rate.

Possible Underreporting of Acts of Intercourse

The WHO study required daily records of mucus symptoms and menstruation as well as acts of intercourse. As a minimum, volunteers were required to record the last act before, all acts during, and the first act after the fertile period, with a note indicating whether all acts during the cycle had been

reported. A simpler design and instructions would have been to insist on all acts of intercourse being reported, irrespective of their timing relative to the fertile period. Although there were monthly reviews of the records by the study monitors and an assessment of the reliability of the information, there are strong suggestions in the patterns of pregnancy rates that substantial underreporting did occur. The percentage of imperfect use cycles which resulted in pregnancy were 12.6%, 7.7%, and 3.9% for women aged under 28 years, 28–34 years or 35–38 years, respectively, in Ireland, New Zealand, and the Philippines, while in India the corresponding pregnancy rates were 46.0%, 30.7%, and 16.8%, and in El Salvador they were 86.9%, 70.1%, and 45.4%, respectively. It is not plausible that the underlying fecundity of the volunteers is so different by country since the **eligibility criteria** were the same in all countries. Adjustment for other factors that might be related to pregnancy rates (such as reported frequency of intercourse) could not explain these large differences between countries. Moreover, the pregnancy rates among perfect use cycles were much more homogeneous between countries, and Trussell & Grummer-Strawn [24] concluded that acts of intercourse which occurred contrary to the rules were less likely to be reported when the cycle did not result in pregnancy, and the rate of such underreporting was higher in India and El Salvador than the other countries.

It is impossible to assess correctly perfect and imperfect use failure rates for methods which require a high degree of user compliance, such as the ovulation method, or male or female condoms, or the diaphragm, unless there is accurate, **unbiased** recording of all acts of intercourse and other features of coital behavior. The only statistic that does not require accurate records of coitus and method use is the overall pregnancy rate, but we have seen above that this rate is the most difficult to interpret. By contrast, the assessment of pregnancy rates for methods which are not coitus- or user-dependent, such as the IUD, can be made without the need for diaries, since the perfect use and overall pregnancy rates are the same.

The Steiner Model for Assessing Efficacy and Effectiveness

The importance of distinguishing between the efficacy of the contraceptive method when used correctly and the behavioral factors that determine whether the method is used correctly or at all was recognized by the authors of *Contraceptive Technology* [10]. A comprehensive review of the literature on contraceptive effectiveness [25] was undertaken and two main statistics were used to summarize the efficacy of different methods – the perfect use and typical use pregnancy rates.

This distinction has been widely used in subsequent publications (see, for example, Hatcher et al. [11]). More recently, Steiner et al. [21] proposed a theoretic model of contraceptive efficacy and contraceptive effectiveness that distinguishes clearly the different factors that govern the assessment of contraceptive methods. The couple's ability to conceive (fecundity of the female and fertilizing capacity of her partner) combined with the timing and frequency of intercourse determine the (unobservable) expected pregnancy rate R_{Exp} in the absence of contraception (see Figure 1). This expected pregnancy rate is reduced by the protection due to the contraceptive method under conditions of perfect use to yield the "perfect use pregnancy rate", $R_{\text{Obs,Perf}}$. The *efficacy* of the contraceptive method is defined as the

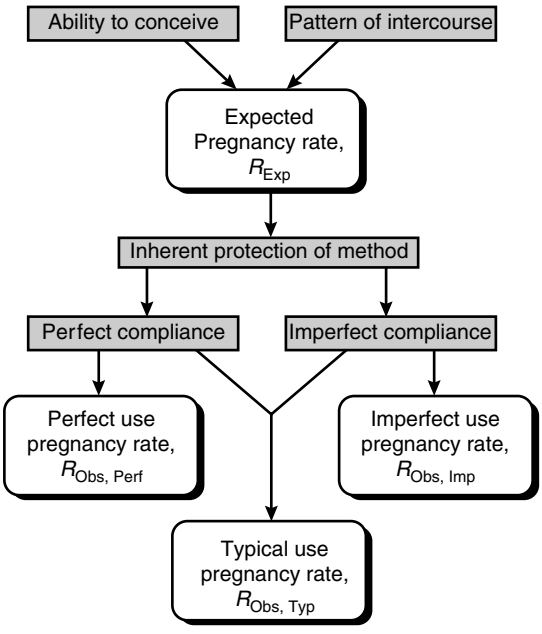

Figure 1 A conceptual model for contraceptive efficacy and contraceptive effectiveness (adapted from Steiner et al. [21])

preventable fraction under conditions of perfect use $1 - R_{\text{Obs,Perf}}/R_{\text{Exp}}$. When the method is used imperfectly, we observe the "imperfect use pregnancy rate", $R_{\text{Obs,Imp}}$. The difference between the two pregnancy rates is a measure of how unforgiving the method is of imperfect use. Similarly, the preventable fraction can be computed to give the efficacy of the method under conditions of imperfect use. In practice, there may be degrees of imperfect use, each with a different impact on the pregnancy rate, but only one type of imperfect use is shown in Figure 1 for simplicity. Most users or groups of users will have a mixture of perfect and imperfect use and we observe the "typical use pregnancy rate", $R_{\text{Obs,Typ}}$. The *effectiveness* of the method, the preventable fraction under conditions of typical use, is $1 - R_{\text{Obs,Typ}}/R_{\text{Exp}}$. Effectiveness is thus a measure that includes the degree of compliance with correct method use. The difference between efficacy and effectiveness rates depends not only on the pregnancy rates under conditions of perfect or imperfect use, but is also a function of the proportion of users who use the method perfectly, or the proportion of cycles in which the method is used perfectly. These proportions provide information on the degree of difficulty in using the method according to the rules for perfect use (assuming that subjects are trying to use the method perfectly to avoid pregnancy and do not deliberately use it imperfectly). Methods that do not require any particular intervention by the user, such as the IUD or sterilization, have the same typical and perfect use pregnancy rates, since the method cannot be used imperfectly. Thus IUD effectiveness is the same as IUD efficacy. However, oral contraceptives, which must be taken according to a fixed schedule, or coitus-dependent methods such as the condom, have a lower perfect than typical use pregnancy rate, and there can be a considerable difference between the efficacy and effectiveness of the method.

Contraceptive efficacy measures the inherent protection of the method and can thus be readily generalized to other populations and groups of users. In theory, the preventable fraction would be the same for users with different fecundity and patterns of intercourse. Similarly, the imperfect use pregnancy rate and preventable fraction under conditions of imperfect use can be generalized to other groups of users and can demonstrate the implications of imperfect use on pregnancy rates. By contrast, contraceptive effectiveness is very difficult to generalize, as the degree of compliance with correct method use (or proportion of perfect compliance users or cycles) depends on many factors that differ from one group to another, according to personal circumstances, and may also vary within the same couple from cycle to cycle. Although the typical use pregnancy rate is the easiest to observe, it is the most difficult to generalize.

The value of the model introduced by Steiner et al. [21] is that it focuses on what information is necessary to assess the different measures of a contraceptive method and generalize these to other users. Only in studies in which there is accurate recording of all acts of intercourse relative to the time of ovulation can the expected number of pregnancies be computed. Moreover, to distinguish between the effectiveness of the method under different patterns and types of use, accurate records of each use are required in an unbiased manner. In the NFP study, it was possible to classify cycles according to whether or not the rules for abstinence were correctly followed, and if not, according to the type of departure. However, a more complex example is the assessment of a barrier contraceptive method such as the diaphragm. If the day of ovulation and all acts of intercourse are accurately recorded, then the expected pregnancy rate R_{Exp} can be computed, but exact details of how the diaphragm was used for each act of intercourse (e.g. how long before intercourse it was inserted, how long it was left in place after intercourse, and whether spermicide was also used) are necessary to distinguish the different types of imperfect use, their associated pregnancy rates, and the efficacy of the method according to these different types of imperfect use.

Conceptions According to Different Times of Intercourse

To apply the Steiner model and estimate contraceptive efficacy, we need to know the probability of conception among couples not using any contraceptive method according to different patterns of frequency and timing of intercourse. Estimates of such conception probabilities have been derived from pregnancies among women who received a single insemination with donor sperm [19], from records of menstrual cycles, intercourse, and pregnancies among couples practicing the calendar method of NFP [4], and from records of couples planning pregnancies [31]. The analysis and interpretation of these conception probabilities is not straightforward, and there is currently

no consensus on the exact values for couples of normal fertility.

Donor Insemination

Schwartz et al. [19] reported on the success rate of artificial insemination with frozen donor semen among 529 presumed fertile women from infertile couples in which the male partner was either azoospermic or oligozoospermic (no or few sperm in the ejaculate). The day of ovulation was estimated from the basal body temperature (BBT) chart as the last day of hypothermia before the postovulatory rise. Data are available from 631 cycles with interpretable charts that resulted in 82 pregnancies. The proportion of cycles that resulted in a pregnancy (defined as at least three weeks of sustained hyperthermia from the BBT chart) is shown in Table 1 according to cycle day and show the highest pregnancy rate for inseminations on day -1. These results probably underestimate the chances of conception following coitus in couples of normal fertility, since the majority of infertile couples have some degree of fertility impairment in both partners [22]. While only couples with documented male factor infertility were included in the donor insemination series, reduced fecundity in the female partner is to be expected. Moreover, frozen donor semen may have a lower fertilizing capacity than fresh semen deposited during coitus.

The Barrett–Marshall Model

Conception probabilities according to cycle day among couples of normal fertility were obtained by Barrett & Marshall [4], who studied records of menstrual cycles, acts of intercourse and daily BBT charts for 241 couples, the majority of whom were practicing the calendar method of fertility regulation. Some couples also continued to record acts of intercourse and daily BBT charts in cycles in which they were attempting to achieve a pregnancy. By contrast to the donor insemination data, where there was only a single insemination per cycle, multiple acts of intercourse occurred in many cycles. Barrett & Marshall introduced a simple probability model for the probability of conception in a given cycle

$$P = 1 - \prod_i (1 - \pi_i)^{x_i}, \qquad (1)$$

where π_i is the probability of fertilization on day i and x_i is an indicator variable, which takes value 1 if intercourse takes place on day i and 0 otherwise. The results of the **maximum likelihood** fit are shown in Table 1 (using Schwartz' renumbering of the days before ovulation in preference to Barrett & Marshall's notation) and show reasonable agreement with the artificial insemination data.

A third set of conception probabilities was obtained by Vollman [29] from 74 couples who had been using periodic abstinence to avoid conception and then "agreed to have intercourse only once in the cycle for the next planned pregnancy". Although the data included cycles with more than a single act of intercourse, Vollman did not use model (1) but computed the conception probabilities directly from the number of conceptions that resulted from intercourse on a particular day by the number of cycles with intercourse on that day. This will have underestimated the conception probabilities.

These three sets of conception probabilities were smoothed by Dixon et al. [6] using a weighted **moving average** to reflect uncertainty in the exact time of ovulation which had been indirectly estimated from the BBT charts (Table 1). There is no theoretic reason or evidence that such smoothed estimates are more appropriate, and the highest chance of conception on day -1 is substantially less than that obtained from the original daily conception probabilities.

The Extended Schwartz–Barrett–Marshall Model

A limitation of model (1) is the assumption that the probabilities of conception from different coital acts in the same cycle are independent. If P_i is the probability of conception during a cycle with intercourse on day i, then the model implies that the probability P_{ij} of conception in a cycle with intercourse on days i and j is given by

$$P_{ij} = P_i + P_j - P_i P_j. \qquad (2)$$

Application of the independence model to cycles with multiple acts of intercourse leads to conception probabilities as high as 68% if intercourse takes place every day [18]. This appears to be too high, particularly in view of the rate of fetal loss within the first six weeks of pregnancy (the endpoint used by Barrett & Marshall). The model can be generalized to include

Table 1 Estimated conception probabilities relative to day of ovulation

Source	Data and model	Day of cycle (relative to day of ovulation)								
		−5	−4	−3	−2	−1	0	1	2	3
Schwartz [19]	Donor insemination (single insemination per cycle)		8%	20%	13%	21%	15%	11%	9%	
	Natural Family Planning cycles									
Barrett & Marshall [4]	Model (1)		13%	20%	17%	30%	14%	7%		
Dixon et al. [6]	Combined estimate (smoothed)	6%	10%	15%	17%	17%	14%	9%	5%	2%
Schwartz et al. [18]	Model (3)	4%	14%	20%	20%	34%	14%	7%		
	Direct estimates (cycles with a single coitus)	4%	20%	26%	15%	27%	15%	7%		
Royston [17]	Model (4)	4%	8%	16%	22%	22%	16%	7%	3%	1%
	Early Pregnancy Study									
Wilcox et al. [31]	Model (3)	10%	16%	14%	27%	31%	33%			
	Direct estimates (cycles with a single coitus)	8%	17%	8%	36%	34%	36%			

these other factors by writing the overall probability of pregnancy in a given cycle as $P = P_o P_f P_v$, where P_o is the probability of ovulation, P_f the probability that the ovum is fertilized, and P_v the probability that the fertilized ovum successfully implants and survives to the time of observation. Assuming that coital acts within the cycle are independent, we have, as before,

$$P_f = 1 - \prod_i (1 - \pi_i)^{x_i}.$$

The extended Schwartz–Barrett–Marshall model for the probability of conception in a cycle with coital pattern **x** is

$$P(\mathbf{x}) = kP_f = k\left[1 - \prod_i (1 - \pi_i)^{x_i}\right], \qquad (3)$$

where only the product $k = P_o P_v$ is estimable (*see* **Estimation**), since the probabilities of ovulation and ovum viability cannot be estimated separately from the observed pregnancies and coital patterns. Note that the probability of pregnancy in a cycle with intercourse on days i and j is now given by $P_{ij} = P_i + P_j - P_i P_j / k$ instead of (2) from the Barrett–Marshall model. The parameter estimates from model (3) obtained by Schwartz et al. to Barrett & Marshall's NFP data (extended by a small number of additional cycles; Table 1) are comparable to those from model (1). The combined ovulation/viability factor k was estimated to be 0.52, and the estimated probability of conception for a cycle in which intercourse occurs every day during the fertile period is reduced from 68% to 49%. Since all cycles included in the analysis had an ovulatory BBT pattern, the parameter k provides an estimate of P_v, the probability that the fertilized ovum survives to six weeks of pregnancy. There is also good agreement with the direct estimates of the conception probabilities from those cycles in which a single act of intercourse occurred in the fertile period (Table 1).

The model was further extended by Royston [17] who postulated that the probability of fertilization resulting from a single act of intercourse on the day of ovulation was 1 and declined **exponentially** according to the survival capacity of sperm and ovum for single acts before or after the day of ovulation. He also assumed that the viability of the fertilized ovum decreased linearly with the age of the woman. In addition, he allowed for uncertainty in the exact time of ovulation (indirectly estimated from the day of BBT shift) by averaging over an assumed **normal distribution** with **mean** 2 and **standard deviation** 1.25 days. The fitted conception probabilities (Table 1) show a maximum around days −2 and −1 with nonzero probabilities as far as day 3. Note that we have renumbered the days of the cycle so that day 0 corresponds to the last day before the BBT rise, instead of the first day of hyperthermia used in Royston's original paper. The viability of the fertilized ovum was estimated to be 0.48, close to that obtained in model (3), and declined by 0.022 for each additional year of the woman's age. The mean lifetime of the sperm was 1.47 days, twice as long as that of the ovum (0.70 days).

The Early Pregnancy Study

The fourth source of data on conception probabilities according to cycle day is records of couples enrolled in the Early Pregnancy Study [31]. This was a prospective study of 221 women planning pregnancy who kept coital and menstrual diaries and provided daily early morning urine samples for estimation of ovarian steroid metabolites. The exact day of ovulation was estimated from the rapid drop in the estrogen-to-progesterone ratio that occurs just before ovulation [3]. The primary objective of the study was to determine the **risk** of early pregnancy loss among healthy women, and thus a highly sensitive assay was used to detect pregnancies within six days of fertilization, around the time of implantation ("chemical pregnancies"). From a total of 199 chemical pregnancies, 48 ended within six weeks of last menstrual period and the remaining 76% were recognized clinically. The estimated conception probabilities obtained from these data for cycles in which a single act of intercourse took place are similar to those obtained from model (3) using all observed cycles (Table 1). The peak conception probabilities around the day of ovulation were in general higher than those seen with the donor insemination and NFP series, particularly for coitus on day 0 where the rates were more than twice as great. This is to be expected, since the study included chemical pregnancies. Interestingly, no pregnancies were observed for coital acts on the day after ovulation, although the authors could not exclude conception probabilities as

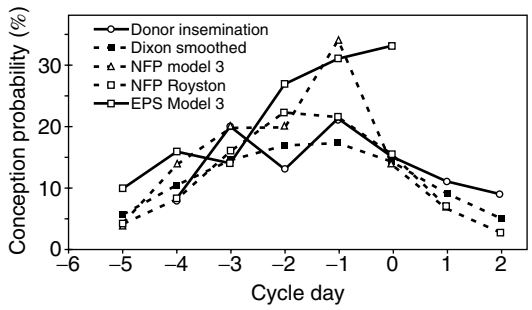

Figure 2 Estimated probabilities of conception arising from single acts of intercourse relative to the day of ovulation

high as 12% owing to the small number of total cycles observed.

Summary

The data described above are the only currently available sources of information on conception probabilities for acts of intercourse on different days of the cycle and some are plotted in Figure 2 for comparison. The artificial insemination data probably underestimate the true rates due to the potential for subfertility in the female. Similarly, the NFP data of Barrett & Marshall and Vollman are also probable underestimates, since they refer to couples who had been successfully using the method for some time and thus the highest fecundity couples will have either achieved a pregnancy or have changed to another contraceptive method. By contrast, the results from the Early Pregnancy Study are overestimates due to the inclusion of chemical pregnancies. A further analysis of these data considering only the clinically recognized pregnancies has recently been completed [27].

Studies of Emergency Postcoital Contraception

Emergency postcoital contraceptive methods protect against pregnancy after unprotected intercourse by reducing the viability of any fertilized ovum and preventing implantation. Although a variety of different hormonal regimens have been studied, the most widely used method is the Yuzpe regimen [34], which involves two high doses of estrogen and progestogen, the first taken within 72 hours of intercourse and the second 12 hours later. Each dose contains 500 μg of levonorgestrel and 100 μg of ethinylestradiol, which is about three times higher than the usual daily doses of hormones in the most widely used combined oral contraceptive pills (150 μg of levonorgestrel and 30 μg of ethinylestradiol). The majority of women request emergency contraception soon after a single act of unprotected intercourse during the fertile period, and thus the expected number of pregnancies can be estimated using the daily conception probabilities, either directly for a single act, or using model (3) if there were multiple acts. In general, the exact day of ovulation is not known and must be estimated from the usual menstrual cycle length and the date of onset of the previous menses. Any uncertainty in this estimate is not a problem for properly randomized comparative studies (*see* **Clinical Trials, Overview**) [30], since the comparison between treatment arms will not be biased. Similarly, differences between treatment arms will not be biased by calculations based on daily conception probabilities that are too low or too high (see, for example, Dixon et al. [6]), and the simple comparison of the number of observed pregnancies in the treatment arms [7] also provides a valid estimate of differences between the groups.

However, it is less easy to estimate the absolute efficacy of postcoital contraceptive methods. The overall pregnancy rate is a poor measure of the performance of such methods, since many cycles with unprotected intercourse will not result in pregnancy. A better measure is the preventable fraction, or efficacy rate, which has been estimated to be approximately 75% (95% confidence interval 68%–79%) by Trussell et al. [26], who reanalyzed data from ten studies of the Yuzpe regimen. However, the estimated efficacy rate for one study ranged from 55.3% to 67.1% according to which estimates of daily conception probabilities were used.

Estimating the Efficacy of Other Contraceptive Methods

While the conceptual model (Figure 1) and the estimates of conception probabilities (Table 2) clarify the information necessary to assess efficacy, these cannot provide estimates of efficacy for all contraceptive methods. Estimating the efficacy of emergency

contraception is straightforward, since the necessary information can be obtained by interview when the woman requests the method. Most users are exposed to a single act of unprotected intercourse and the method must be used within three days so that the timing of intercourse, usual length of menstrual cycle, and date of onset of the most recent menses can be accurately provided. This is by contrast to other coitus-dependent methods, for which information on coitus and menstruation is seldom sufficient.

Withdrawal (coitus interruptus) presents particular difficulties, since the method is used in ways that are difficult to observe – either it is used in emergency as a last resort when no other contraceptive methods are available, or it is used as a deliberate strategy for the prevention of pregnancy. The former type of user cannot be identified in advance, and any retrospective information collected on such users will be hopelessly biased by a greater reporting rate for cycles where pregnancy actually occurred. The second type of user will usually have poor access to reproductive health care services and will thus be difficult to identify. Nevertheless, any cohort of regular withdrawal users would be biased toward the better and more reliable users of the method. Information on the typical pregnancy rates of withdrawal must be obtained from other sources, such as **population based studies** or cohorts assembled for other reasons. For example, as part of the Romania 1993 Reproductive Health Care Survey, women provided information retrospectively on contraceptive and pregnancy history. These were combined to reconstruct periods of use of different contraceptive methods, the timing of any pregnancies, and the times and reasons for changing to another method. The 12-month cumulative life table pregnancy rate was found to be 30% [20]. This rate is more than four times higher than that estimated from a cohort of women participating in a long term prospective study conducted in England and Scotland by the Oxford Family Planning Association, which enrolled 17 000 British women aged 25–39 years using oral contraceptives, a diaphragm, or an IUD. In this cohort, the estimated pregnancy rate for withdrawal was 6.7 per 100 woman-years [28].

The wide discrepancy in estimated pregnancy rates from the UK cohort and the Romanian study illustrates the difficulties in assessing typical use pregnancy rates. The reproductive health care facilities in the two countries were widely different and the Oxford FPA cohort contained a large proportion of careful users. Indeed, a more representative group of married users in the UK yielded an estimated pregnancy rate of 21.9 per 100 woman-years [16] that is closer to the estimate from Romania. Note that it is almost impossible to obtain sufficient accurate information on menstruation and acts of intercourse among a group of withdrawal users to estimate the expected number of pregnancies. Thus the contraceptive efficacy of the method cannot be estimated, but only the typical use pregnancy rate.

Unresolved Problems

The first unresolved problem concerns estimates of daily conception probabilities. The extended Schwartz–Barrett–Marshall model has been shown to provide a good fit to data on conceptions and patterns of intercourse, and has proven its value for estimating conception probabilities according to different days relative to ovulation. It can also be applied to observed coital patterns to estimate the expected number of pregnancies in the absence of contraception and hence the contraceptive efficacy of different methods. The only outstanding issue is whether better or more data can be obtained on which to apply the model and derive more reliable estimates. Data on pregnancies arising from donor insemination or from users of natural family planning methods have their limitations, while studies of couples planning pregnancy, as in the Early Pregnancy Study, may yield more representative data. It is important to collect such information from a wide range of couples (different countries, age ranges, personal circumstances) to understand better the factors related to fecundity and fertilization probabilities. The availability of simple home testing methods to record urinary metabolites (see, for example, May [13]) would greatly simplify data collection and estimation of the day of ovulation.

The second unresolved problem concerns the efficacy of other coitus-dependent methods and the simultaneous use of combinations of methods. As we have seen, typical use pregnancy rates can be obtained for coitus-dependent methods, but the requirement to record coital acts and estimate the day of ovulation make it very difficult to collect unbiased data that would permit calculation of the efficacy of perfect use and of various types of imperfect use. Such efficacy estimates will be valuable

in counseling prospective users and predicting pregnancy rates among different types of user. An additional challenge is to establish typical use pregnancy rates and contraceptive efficacy of mixed method use; for example, use of a barrier method only during the fertile phase of the cycle. Theoretic typical use pregnancy rates can be estimated using the extended Schwartz–Barrett–Marshall and the contraceptive efficacy of barrier and natural family planning methods, but validation against observed data would be essential.

Key References

The following references are considered of particular importance in preparing this article: Barrett & Marshall [4], Royston [17], Schwartz et al. [18], Steiner et al. [21], Trussell & Grummer-Strawn [24], Trussell et al. [26], and Wilcox et al. [31].

References

[1] Aalen, O.O. (1988). Heterogeneity in survival analysis, *Statistics in Medicine* **7**, 1121–1138.

[2] Armstrong, E.G., Ehrlich, P.H., Birken, S., Schlatterer, J.P., Siris, E., Hembree, W.C. et al. (1984). Use of a highly sensitive and specific immunoradiometric assay for detection of human chorionic gonadotropin in urine of normal, nonpregnant, and pregnant individuals, *Journal of Clinical Endocrinology and Metabolism* **59**, 867–874.

[3] Baird, D.D., Weinberg, C.R., Wilcox, A.J. & McConnaughey, D.R. (1991). Using the ratio of urinary oestrogen and progesterone metabolites to estimate day of ovulation, *Statistics in Medicine* **10**, 255–266.

[4] Barrett, J.C. & Marshall, J. (1969). The risk of conception on different days of the menstrual cycle, *Population Studies* **23**, 455–461.

[5] Bonnar, J., Freundl, G., Royston, P., Flynn, A., Snowden, R. & Kirkman, R. (1999). Personal hormone monitoring for contraception *British Journal of Family Planning* **24**, 128–134.

[6] Dixon, G.W., Schlesselman, J.J., Ory, H.W. & Blye, R.P. (1980). Ethinyl estradiol and conjugated estrogens as postcoital contraceptives, *Journal of the American Medical Association* **244**, 1336–1339.

[7] Glasier, A., Thong, K.J., Dewar, M., Mackie, M. & Baird, D.T. (1992). Mifepristone (RU486) compared with high-dose estrogen and progestogen for emergency postcoital contraception, *New England Journal of Medicine* **327**, 1041–1044.

[8] Gray, R.H., Leridon, H. & Spira, A., eds. (1993). *Biomedical and Demographic Determinants of Reproduction*. Oxford University Press, Oxford.

[9] Hamilton, D.W. & Saling, P.M. (1996). Male methods, in *Contraceptive Research and Development: Looking to the Future*, P.F. Harrison & A. Rosenfield, eds. National Academy Press, Washington, pp. 381–400.

[10] Hatcher, R.A., Guest, F., Stewart, F., Stewart, G.K., Trussell, J., Cerel, S. & Cates, W. (1988). *Contraceptive Technology 1988–1989*, 1st Ed. Printed Matter, Atlanta.

[11] Hatcher, R.A., Trussell, J., Stewart, F., Stewart, G.K., Kowal, D., Guest, F., Bowen, S., Cates, W. & Policar, M.S. (1996). *Contraceptive Technology*, 16th Ed. Irvingston, New York.

[12] Latz, L.J. (1934). *The Rhythm of Sterility and Fertility in Women*, 4th Ed. Latz Foundation, Chicago.

[13] May, K. (1991). Home tests to monitor fertility, *American Journal of Obstetrics and Gynecology* **165**, 2000–2002.

[14] Nieschlag, E. & Behre, H.M., eds. (1997). *Andrology: Male Reproductive Health and Dysfunction*. Springer-Verlag, Berlin.

[15] Ogino, K. (1930). Ovulationstermin and Konzeptionstermin (Ovulation day and conception day), *Zentralblatt für Gynäkologie* **54**, 464–479.

[16] Peel, J. (1972). The Hull family survey: II. Family planning in the first five years of marriage, *Journal of Biosocial Science* **4**, 333–346.

[17] Royston, J.P. (1982). Basal body temperature, ovulation and the risk of conception, with special reference to the lifetimes of sperm and egg, *Biometrics* **38**, 397–406.

[18] Schwartz, D., Macdonald, P.D.M. & Heuchel, V. (1980). Fecundability, coital frequency and the viability of ova, *Population Studies* **34**, 397–400.

[19] Schwartz, D., Mayaux, M.-J., Martin-Boyce, A., Czyglik, F. & David, G. (1979). Donor insemination: conception rate according to cycle day in a series of 821 cycles with a single insemination, *Fertility and Sterility* **31**, 226–229.

[20] Serbanescu, F. & Morris, L. (1995). Contraception, in *Reproductive Health Survey Romania 1993*. Institute for Mother and Child Health Care, Bucharest, Romania, pp. 61–89.

[21] Steiner, M., Dominik, R., Trussell, J. & Hertz-Picciotto, I. (1996). Measuring contraceptive effectiveness: a conceptual framework, *Obstetrics and Gynecology* **88**, 24S–30S.

[22] The ESHRE Capri Workshop (1996). Guidelines to the prevalence, diagnosis, treatment and management of infertility, *Human Reproduction* **11**, 1775–1807.

[23] Tietze, C. & Lewit, S. (1970). Evaluation of intrauterine devices: ninth progress report of the Cooperative Statistical Program, *Studies in Family Planning* **55**, 1–40.

[24] Trussell, J. & Grummer-Strawn, L. (1990). Contraceptive failure of the ovulation method of periodic abstinence, *Family Planning Perspectives* **22**, 65–75.

[25] Trussell, J. & Kost, K. (1987). Contraceptive failure in the United States: a critical review of the literature, *Studies in Family Planning* **18**, 237–283.

[26] Trussell, J., Ellerston, C. & Stewart, F. (1996). The effectiveness of the Yuzpe regimen of emergency contraception, *Family Planning Perspectives* **28**, 58–64, 87.

[27] Trussell, J., Rodriguez, G. & Ellerston, C. (1998). New estimates of the effectiveness of the Yuzpe regimen of emergency contraception *Contraception* **57**, 363–369.

[28] Vessey, M., Lawless, M. & Yeates, D. (1982). Efficacy of different contraceptive methods, *Lancet* **i**, 841–842.

[29] Vollman, R.F. (1977). Assessment of the fertile and sterile phases of the menstrual cycle, *International Review of Natural Family Planning* **1**, 40–47.

[30] Webb, A.M.C., Russell, J. & Elstein, M. (1992). Comparison of Yuzpe regimen, danazol, and mifepristone (RU486) in oral postcoital contraception, *British Medical Journal* **305**, 927–931.

[31] Wilcox, A.J., Weinberg, C.R. & Baird, D.D. (1995). Timing of sexual intercourse in relation to ovulation: effects on the probability of conception, survival of the pregnancy, and sex of the baby, *New England Journal of Medicine* **333**, 1517–1521.

[32] World Health Organization (1981). A prospective multicentre trial of the ovulation method of natural family planning II. The effectiveness phase, *Fertility and Sterility* **36**, 591–598.

[33] World Health Organization, Special Programme of Research, Development and Research Training in Human Reproduction: Task Force on the Safety and Efficacy of Fertility Regulating Methods (1990). The TCu380A, TCu220C, Multiload 250 and Nova T IUDs at 3, 5 and 7 years of use – results from three randomized multicentre trials, *Contraception* **42**, 141–158.

[34] Yuzpe, A.A. & Lancee, W.J. (1977). Ethinylestradiol and dl-norgestrel as a postcoital contraceptive, *Fertility and Sterility* **28**, 932–936.

TIMOTHY M.M. FARLEY

Resampling Procedures for Sample Surveys

One of the major statistical challenges in the development and application of complex **probability sample** designs is the valid **estimation** of sampling errors. Probability sampling theory and practice permits various departures from the **simple random, with or without replacement**, sampling model. These departures, which often take the form of **stratification**, **cluster sampling**, **multistage sampling**, and unequal probability of selection, are often used in order to produce sample designs that are both feasible and cost efficient. One of the drawbacks to these practical and efficient probability sample designs is the complexity and possible intractability that occurs with respect to the estimation of sampling errors. There are several basic methods linked to the general concept of resampling that have been developed for the estimation of sampling errors from complex (clustered and stratified) sample designs. These methods are discussed under three general headings: Simple Replication, **Jackknife** Repeated Replication, and Balanced Repeated Replication. It is interesting to note that while the use of a resampling technique known as the **bootstrap** has enjoyed wide use and acceptance in the general statistical literature, there have not been many attempts to apply simple bootstrap methods to complex applied probability sample designs. There has been some work by Rao and Wu [8], but they conclude that for certain classes of complex designs "the bootstrap variances estimators are less stable than those based on the linearization or the jackknife."

The basic approach of repeated replication in survey design was developed at the US Census [1] and built on the basic concepts of Mahalanobis [6]. McCarthy [7] introduced the idea of orthogonal balancing. It should be noted, however, that the use of replication in the form of "split-samples" was probably in use by psychologists prior to the development of probability sampling.

Simple Replication

The basic strategy of Simple Replication (*see* **Interpenetrating Samples**) or Replicated Subsamples involves four basic steps.

Assuming that the total sample is to consist of a primary selection (which will produce the desired sample size of n elements), a sample design is developed that will involve the selection of a/K primary sampling units. The value of K must be some integer greater than one and less than n (Deming [2] recommends the use of $K = 10$). A probability sampling design is formulated so that it may be repeated K times. Full flexibility is allowed in the sample design as long as the conditions required for probability sampling are satisfied. The sample design may be as simple as simple random or **systematic** selection of a/K elements with no stratification. It may be quite complex and involve stratification, clusters of unequal size, unequal probability of selection, and/or multiple stages of sampling.

Once the sample design has been specified, the actual sample selection process is carried out separately and independently a total of K times. Each repetition produces a replication or replicate. Let R_k denote the Kth replication or replicate. The set consisting of all K replicates constitutes the total sample $S : S = \{R_1, \ldots, R_k\}$.

Application of the estimation function $g(\cdot)$ for the particular survey estimate produces the total sample estimate $g(S)$. Let $g(R_k)$ denote the survey estimate produced from the kth replicate.

The simple replicated estimate of the sampling **variance** of $g(S)$ is

$$\text{var}_{\text{rep}}[g(S)] = \frac{1}{K(K-1)} \sum_{k=1}^{K} [g(R_k) - g(S)]^2. \quad (1)$$

The **standard error** of $g(S)$ is estimated as

$$\text{se}[g(S)] = \{\text{var}[g(S)]\}^{1/2}. \quad (2)$$

Confidence intervals based on this estimated standard error se$[g(S)]$ generally use the **Student's *t* distribution** with $K - 1$ "degrees of freedom".

It should be noted that (1) is an **unbiased** estimator of the sampling variance of $\overline{g(R_k)} = \sum g(R_k)/K$, the mean of the K estimates $g(R_1), \ldots, g(R_k)$. The use of (1) as a variance estimator for $g(S)$, the estimate derived from the total sample, depends upon the assumption that the **sampling distribution** of $g(S)$ is approximately equal to the sampling distribution of $\overline{g(R_k)}$. For certain simple statistics and simple sample designs, the two estimates $g(S)$ and $\overline{g(R_k)}$ are algebraically identical. This is the case for simple means

and proportions from simple random samples of fixed size n elements. For other more complex estimates (e.g. **regression** and **correlation** coefficients) and/or more complex sample designs (e.g. designs based on clusters of unequal size), these estimates may be different. For certain types of estimates based on **order statistics** (e.g. **medians** and percentiles (*see* **Quantiles**)) the impact of departures from this assumption may be substantial.

When first proposed, Simple Replicated Subsampling was seen as a sample design tool: that is, it was a model for sample designs that would permit simple and straightforward estimation of standard errors. The resampling took place in terms of generating K replications from the "sampling distribution". In many practical sample design situations, particularly those involving the presentation of legal evidence in either administrative or legal proceedings, the method has proven to be both simple and intuitively appealing for both statisticians and nonstatisticians. However, the method does have its limitations. The greatest limitation of the model is the limits that are placed on the complexity and efficiency of design. For example, a sample of 100 elements may, in its fullest complexity, utilize a stratification structure of 50 or even 100 strata. In a Simple Replicated Subsample Design which utilizes $K = 10$ replicates, the maximum number of strata is equal to 10.

The Simple Replicated Subsampling model is often used as a model for the computation of sampling errors, even in those situations in which it is not actually used in the sample design. That is, even when the sample is not selected in accord with the model, a pseudoselection model is formulated that reformulates the actual sampling process into a similar replicated subsampling process. This reformulation often involves a collapsing or combining of substrata within primary strata, after sample selection, for the purpose of standard error estimation. For example, a sample design might specify the selection of 100 elements from a population that is partitioned into 50 equal sized strata, with two elements selected per stratum. This design might be viewed as a replicated sample based on $K = 2$ or it might be reformulated, for purposes of sampling error estimation, into a design consisting of 10 strata with $K = 10$. The actual 50 strata would be collapsed or combined into 10 computational strata.

In most situations in which a design is reformulated for the purpose of standard error estimation, it is generally the case that the estimate of sampling error will be "conservative".

Jackknife Repeated Replication

Jackknife Repeated Replication (JRR) and Balanced Repeated Replication (BRR) are methods for standard error estimation that involve actual "resampling" or reuse of sample observations.

In their simplest forms, BRR and JRR assume a sample selection model based on a stratified design with two independent selections per stratum. This "paired selection" model assumes that the population is partitioned into $H = a/2$ strata, where a represents the total number of primary selections. Within each of these strata, it is assumed that there will be two independent primary selections. Following the first stage of sampling, there may be any number of subsequent stages, and selection may involve equal or unequal final probabilities for elements.

Jackknife Repeated Replication (JRR) estimates of sampling variance and standard error [3] are constructed as follows: we assume $H = a/2$ strata, each consisting of two primary stet units.

Let S denote the entire sample along with any weights that have been applied to the sample observations (including poststratification) associated with the full set of a primary selections.

Let J_h denote the hth jackknife replicate formed by including all sample observations not in the hth stratum, removing all sample observations associated with one of the two primary selections in the hth stratum, and including *twice* all sample observations associated with the other primary selection in the hth stratum.

Let CJ_h denote the hth complement jackknife replicate formed in the same way as the hth jackknife replicate J_h, *except* that the eliminated and doubled primary selections are interchanged.

Let $g(S)$ denote the total sample derived estimate for which a sampling variance is sought. Let $g(J_h)$ and $g(CJ_h)$ denote the same estimator applied to the hth jackknife replicate and complement jackknife replicate respectively. Note that it is assumed that any weighting process that has been applied to the total sample is applied to each jackknife and complement jackknife replicate as if they constituted the sample that was being used for estimation. (In those cases in which "reweighting" of each jackknife replicate is

not feasible, the original weights may be used, but this may result in some estimation **bias**. In practice, the magnitude of this bias is often negligibly small.)

There are two jackknife repeated replication estimates that are used to estimate the variance of $g(S)$ and corresponding standard error $g(S)$. These are defined as follows:

$$\text{var}_{\text{JRR-S}}[g(S)] = \frac{1-f}{2}\sum_{h=1}^{H}[g(J_h) - g(S)]^2 + \frac{1-f}{2}\sum_{h=1}^{H}[g(CJ_h) - g(S)]^2, \tag{3}$$

with

$$\text{se}_{\text{JRR-S}}[g(S)] = \{\text{var}J_{\text{BRR-S}}[g(S)]\}^{1/2}; \tag{4}$$

and

$$\text{var}_{\text{JRR-D}}[g(S)] = \frac{1-f}{4}\sum_{h=1}^{H}[g(J_h) - g(CJ_h)]^2, \tag{5}$$

with

$$\text{se}_{\text{JRR-D}}[g(S)] = \{\text{var}_{\text{JRR-D}}[g(S)]\}. \tag{6}$$

Confidence intervals based on these Jackknife Repeated Replication estimates of standard error generally use the Student's t distribution with H "degrees of freedom". The form $\text{se}_{\text{JRR}-\text{D}}[g(S)]$ given in (6) provides a more conservative estimate of standard error and is generally preferred [5].

Balanced Repeated Replication

Balanced Repeated Replication (BRR) estimates of sampling variance and standard error are constructed as follows: we assume $H = a/2$ strata, each consisting of two primary selections units.

Let S denote the entire sample along with any weights that have been applied to the sample observations (including **poststratification**) associated with the full set of a primary selections.

Let HS_i denote the ith half-sample formed by including all of the observations associated with one of the two primary selections from each of the strata; and let CHS_i denote the ith complement half-sample formed by all of the observations associated with the primary selections in S not in HS_i. The method used for choosing the pattern of primary units that form the half-samples HS_i and complement half-samples CHS_i is known as "full-orthogonal balance". In general, to achieve full-orthogonal balance it is necessary to form K half and complement half samples, where K is the smallest multiple of 4 that is equal to or greater than H. Given the K half- and complement half-samples, the two forms for BRR estimates of sampling variance and standard error are [4, 7]:

$$\text{var}_{\text{BRR-S}}[g(S)] = \frac{1-f}{2K}\sum_{i=1}^{K}[g(HS_i) - g(S)]^2 + \frac{1-f}{2K}\sum_{i=1}^{K}[g(CHS_i) - g(S)]^2, \tag{7}$$

with

$$\text{se}_{\text{BRR-S}}[g(S)] = \{\text{var}_{\text{BRR-S}}[g(S)]\}^{1/2}; \tag{8}$$

and

$$\text{var}_{\text{BRR-D}}[g(S)] = \frac{1-f}{4K}\sum_{i=1}^{K}[g(HS_i) - g(CHS_i)]^2, \tag{9}$$

with

$$\text{se}_{\text{BRR-D}}[g(S)] = \{\text{var}_{\text{BRR-D}}[g(S)]\}^{1/2}. \tag{10}$$

Confidence intervals based on these balanced repeated replication estimates of standard error generally use the Student's t distribution with H "degrees of freedom". The form $\text{se}_{\text{BRR-D}}[g(S)]$ given by (6) provides a more conservative estimate of standard error and is generally preferred [5]. For a more complete discussion of variance estimation in survey samples, readers should consult [9].

References

[1] Deming, W.E. (1956). On simplification of sampling design through replication with equal probabilities and without stages, *Journal of the American Statistical Association* **51**, 24–53.

[2] Deming, W.E. (1960). *Sample Design in Business Research*. Wiley, New York.

[3] Frankel, M.R. (1971). *Inference from Survey Samples*. Institute for Social Research, University of Michigan, Ann Arbor.

[4] Kish, L. & Frankel, M.R. (1970). Balanced repeated replication for standard errors, *Journal of the American Statistical Association* **65**, 1071–1094.

[5] Kish, L. & Frankel, M.R. (1974). Inference from complex samples, *Journal of the Royal Statistical Society, Series B* **36**, 1–37.

[6] Mahalanobis, P.C. (1944). On large-scale sample surveys, *Philosophical Transactions of the Royal Society, Series B* **231**, 329–451.

[7] McCarthy, P.J. (1966). *Replication: an Approach to the Analysis of Data from Complex Surveys*. National Center for Health Statistics, Series 2, No. 14, Washington.

[8] Rao, J.N.K. & Wu, C.F.J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association* **83**, 231–241.

[9] Wolter, K.M. (1986). *Introduction to Variance Estimation*. Springer-Verlag, New York.

MARTIN R. FRANKEL

Residuals for Survival Analysis

The standard definition of a **residual** is the observed datum minus its expected value estimated from a model. Right-censoring precludes its direct application to survival data; a **censored** observation provides incomplete information on the failure time. Alternative definitions of residuals are needed. Three major ones are the generalized residuals of Cox & Snell [13], residuals based on **counting process** martingales and their transforms, and residuals from the **generalized linear regression model** [27] for loglinear **Poisson regression**. The first two are general, while the third is specific to the loglinear **proportional hazards** model of Cox [12] (*see* **Cox Regression Model**). In each case, there are interesting analogies between survival residuals and residuals for normal theory **linear regression**.

The proportional hazards (Cox regression) model is the most frequently used model for survival data, and we emphasize it. The notation and setup are as follows. On each of n independent individuals, one has observed a p-dimensional vector, $\mathbf{X}_i, i = 1, \ldots, n$, of predictor values (or **explanatory variables**), a non-negative random variable, T_i, the duration of follow-up time, and a binary indicator, δ_i, taking the value 1 if the follow-up terminates in the failure event of interest (e.g. death) and 0 if not. Underlying the variables T_i and δ_i is a pair of latent random variables: D_i, the time to failure of person i, and C_i, the time to censoring, where D_i and C_i are conditionally independent given $\mathbf{X}_i$; $T_i = \min(D_i, C_i)$ and $\delta_i = I(D_i \leq C_i)$, with $I(A)$ as the indicator function for event A. Also D_i is assumed to be an absolutely continuous random variable with probability density function $f_D(t)$ and survivor function $S_D(t) = \Pr(D > t)$. Interest centers on modeling the **hazard** function $\lambda_i(t) = f_{Di}(t)/S_{Di}(t)$. Under the Cox model, we have

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{X}_i). \tag{1}$$

The model is parametric if $\lambda_0(t)$ is a specified function of time or **semiparametric** if $\lambda_0(t)$ is unspecified. The covariate vectors may be functions of time; for simplicity, we initially assume that they do not vary over time.

Generalized Residuals

Suppose for the ith individual there exists a function h_i of data $\mathbf{Z}_i$ and parameter vector $\boldsymbol{\theta}$ such that $h_i(\mathbf{Z}_i, \boldsymbol{\theta}) = e_i$, where the e_is are independent and identically distributed of known distribution. Then the Cox & Snell *generalized residual* is $R_i = h_i(\mathbf{Z}_i, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the **maximum likelihood** estimator of $\boldsymbol{\theta}$.

Crowley & Hu [14] and Kay [22] were the first to apply this definition to censored survival data regression. The cumulative hazards $\Lambda_i(T_i)[= -\ln S_i(T_i)], i = 1, \ldots, n$, are distributed as a censored sample of independent unit **exponentials**, with $1 - \delta_i$ serving as a censoring indicator. Under the Cox model, $\Lambda_i(T_i) = \exp(\boldsymbol{\beta}' \mathbf{X}_i) \int_0^{T_i} \lambda_0(s)\, \mathrm{d}s$, and the Cox–Snell generalized residual is $\hat{\Lambda}_i(T_i) = \exp(\hat{\boldsymbol{\beta}}' \mathbf{X}_i) \hat{\Lambda}_0(T_i)$. Many investigators have proposed standard **hypothesis tests** and plots for assessing the exponentiality of these residuals as global checks for model **goodness of fit** [18]. A common plot is a graph of the ordered residuals on the abscissa with their **Nelson–Aalen** cumulative hazard estimator [5, pp. 445, 556] on the ordinate. If failure times are completely observed, it is just a graph of the ordered residuals against the expected values of exponential **order statistics**. Thus, this plot is the exponential analogue of the normal Q–Q plot for residuals from linear regression (*see* **Normal Scores**).

Unfortunately, the validity of these techniques is highly questionable for semiparametric models [7]. An illustrative extreme case occurs when there are no **covariates**. If the failure times are completely observed, the generalized residuals are *precisely* the expected order statistics of a unit exponential sample [15], and the Nelson–Aalen plot is exactly a 45° line through the origin. **Monte Carlo** simulations show that when there is censoring and one covariate, the appearance of the Nelson–Aalen plot depends on the variance of $|\boldsymbol{\beta}' \mathbf{X}_i|$, with the deviation from the 45° line increasing with the variance [7]. Thus, a model containing covariates may appear to fit less well than one with no covariates, even when the covariates have a substantial impact on the hazard. For parametric models, on the other hand, the generalized residuals behave like standardized residuals from **least squares** regression. The ordered residuals have the correct means but substantially smaller variances than the order statistics from the reference distribution, a unit exponential sample for survival data [7], and a standard Gaussian sample for least squares. The

assessment of exponentiality of generalized survival residuals provides a valid indicator of model appropriateness only for parametric models.

Generalized Linear Model Residuals

The formal similarity between likelihoods for proportional hazards models for survival data and **loglinear models** for **Poisson** data [4, 23, 33] has inspired the use of residuals developed for the generalized linear model which has Poisson regression as an important special case. Let $Z_i, i = 1, \ldots, n$, be independent Poisson variates with means $\mu_i = \exp(\boldsymbol{\beta}'\mathbf{X}_i)$, where $\mathbf{X}_i$ is a p-dimensional covariate vector. The basic "observed minus expected" residual is $z_i - \hat{\mu}_i$, but transformed residuals have proved more useful for model checking. Three of these, the Pearson, deviance, and partial residual have been adapted to the semiparametric Cox model. The Pearson residual is $(z_i - \hat{\mu}_i)/\hat{\mu}_i^{1/2}$, the deviance residual is $\text{sgn}(z_i - \hat{\mu}_i)\{2z_i ln(z_i/\hat{\mu}_i) - z_i + \hat{\mu}_i\}^{1/2}$ [27, p. 39], and the partial residual for the jth covariate is $(z_i - \hat{\mu}_i)/\hat{\mu}_i + \hat{\beta}_j x_{ij}$ [27, p. 402], $j = 1, \ldots, p$.

In the Cox model, the unit of analysis can be either each individual or, more finely, each individual at each death time. The residual definition depends on the choice of unit. For the finer level of analysis, let $T_1 < T_2 < \cdots < T_Q$ denote the ordered death times. Suppose r_q individuals are at risk of death at T_q^-. Let $\delta_{kq} = 1$ if individual $k, k = 1, \ldots, r_q$, at risk at T_q^-, died at T_q and 0 otherwise, for $q = 1, \ldots, Q$. Then, using the Whitehead [33] Poisson formulation, δ_{kq} is the analog of z_i. Conditioning on the **risk set** and assuming no **tied survival times**, the probability of death at time t_q for an individual at risk with covariate $\mathbf{X}_i$ is $p_q(\mathbf{X}_i) = \exp(\boldsymbol{\beta}'\mathbf{X}_i)/\sum_{k=1}^{r_q}\exp(\boldsymbol{\beta}'\mathbf{X}_k)$. The analog of μ_i is $p_q(\mathbf{X}_i)$, a conditional mean. Hall et al. [21] extend to Cox regression the adjusted variable plot from **multiple linear regression**. Their plot for the jth covariate contains $\sum_{q=1}^{Q} r_q$ points (one for each individual at each death time) whose coordinates have complicated formulas but can be interpreted as "adjusted z" versus "adjusted x_j", where the adjustment is for the other covariates, the Poisson weights, and the loglink function. A least squares line through the origin has slope $\hat{\beta}_j$, the maximum partial likelihood estimate from the Cox regression, and residuals equal to the Pearson residuals, $\{\delta_{jk} - \hat{p}_q(\mathbf{X}_k)\}\big/\hat{p}_q(\mathbf{X}_k)^{1/2}$, where $\hat{p}_q(\mathbf{X}_k) = \exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_k)\big/\sum_{j=1}^{r}\exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_j)$. As in linear regression, these plots show the partial leverage and influence of each unit on each β coefficient.

At the level of analysis where each individual is the unit, δ_i is analogous to z_i and $\hat{\Lambda}_0(T_i)\exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_i)$ is analogous to $\hat{\mu}_i$. The analogy is purely formal because $\Lambda_0(T_i)\exp(\boldsymbol{\beta}'\mathbf{X}_i)$, itself a random variable, cannot be the expected value of δ_i nor can it be a conditional mean, since the correct conditional mean is given by

$$\mathrm{E}(\delta|T, \mathbf{X}) = \frac{f_{D|\mathbf{X}}(T)S_{C|\mathbf{X}}(T)}{f_{D|\mathbf{X}}(T)S_{C|\mathbf{X}}(T) + S_{D|\mathbf{X}}(T)f_{C|\mathbf{X}}(T)}.$$

The counting process martingale approach (below) gives a more probabilistic justification for the Poisson residual $\delta_i - \hat{\Lambda}_0(T_i)\exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_i)$. The deviance transformation symmetrizes the distribution of the Poisson residuals and therefore the deviance residual can offer improvement for detecting **outliers** in some cases [31]. A plot of the partial residuals for the jth covariate against X_{ij} (see Figure 1) suggests

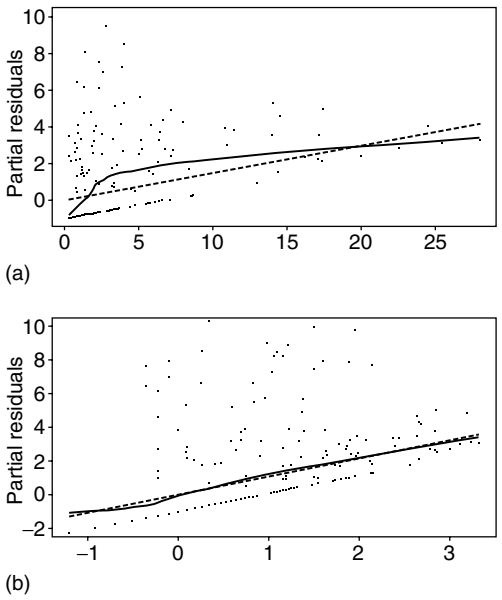

Figure 1 Partial residual plots for primary biliary cirrhosis data, comparing a model with bilirubin (a) to one with log bilirubin (b). The solid line is a weighted smooth, quadratic loess with span of 60% [10] and the dotted line is the fitted functional form from the model. Roughly 10% of the partial residuals are too large to fit on these plots; the vertical scale was chosen to enhance clarity

the correct functional form for the covariates in the linear predictor. As with Poisson regression, a scatterplot smooth superimposed on the residual plot is a useful guide; the smooth should weight each point proportional to $\hat{\mu}_i$ to stabilize variance and reduce bias. If the model is correct, then the smooth will give roughly a straight line with slope $\hat{\beta}_j$. If incorrect, the smooth will suggest the correct functional form. For example, a concave curve would suggest using $\log X_{ij}$ rather than X_{ij}. The primary biliary cirrhosis data [17, Appendix D] provides an illustrative example. Primary biliary cirrhosis (PBC) is a fatal liver disease. A survival model for PBC patients ($n = 312$) was developed and the bilirubin level in the blood was found to be the most important risk factor. The Figure shows the partial residuals from a Cox model with bilirubin as the sole predictor. The weighted smooth in Figure 1(a) is a concave curve, suggesting lack of fit for bilirubin with no transformation. When log bilirubin is used instead (Figure 1(b)), the weighted smooth approximates the line $\hat{\beta}$ log bilirubin quite closely, indicating a good fit for the log transformation. Analogous plots for more extensive models with multiple predictors have a similar appearance [20] and concur in supporting the log transformation as the appropriate functional form for bilirubin. If the covariates are highly correlated, an incorrect functional form for one covariate may influence the appearance of partial residual plots for other covariates. However, the augmented partial residual plot [26] for highly correlated predictors in linear regression works for Cox regression as well. In this technique, the regressors are quadratic polynomials instead of linear terms, and the partial residual is $[(\delta_i - \hat{\mu}_i)/\hat{\mu}_i] + \hat{\beta}_j x_{ij} + \hat{\gamma} x_{ij}^2$ plotted against x_{ij} [20].

Counting Process Martingale Residuals

A counting process $N(t)$ is a **stochastic process** with $N(0) = 0$, and with sample paths that are right-continuous step functions having jumps of size one. Typically $N(t)$ counts the number of events in $[0, t], t \leq \tau$, where τ is the prespecified time for the end of the study. Given a filtration, a sequence of increasing sigma-fields $\{\mathcal{F}_t, t \leq \tau\}$ to which $\{N(t), t \leq \tau\}$ is adapted, the Doob–Meyer decomposition gives

$$N(t) = A(t) + M(t),$$

where $A(t)$ is an increasing, predictable process called the compensator, and $M(t)$ is a mean-zero martingale. Usually $\mathcal{F}_t$ contains all available information on the counting process and any covariates through time t. Heuristically, $\mathrm{E}[\,\mathrm{d}N(t)|\mathcal{F}_{t^-}] = \mathrm{d}A(t)$, so the differential of the compensator gives the conditional probability of an event in the next instant of time, given the preceding history. Because $\mathrm{E}[M(t+s)|\mathcal{F}_t] = M(t)$, the martingale is a process without drift. It has uncorrelated increments and is a natural generalization of a white noise process. Thus, the Doob–Meyer theorem decomposes the counting process into two stochastic processes – a statistical model and a residual process.

Barlow & Prentice [8] laid down the framework for martingale residuals in survival analysis. Allowing for time-varying covariate processes, the data consist of independent triples $\{N_i(t), Y_i(t), \mathbf{X}_i(t); t \leq \tau, i = 1, \ldots, n\}$ and a filtration specified by

$$\mathcal{F}_t = \sigma\{N_i(u), Y_i(u^+), \mathbf{X}_i(u^+); \\ 0 \leq u \leq t, i = 1, \ldots, n\}.$$

In the single-event setting, emphasized here, $N_i(t)$ is 0 prior to the observed death of individual i and 1 at and after the death. For recurrent event data, $N_i(t)$ counts the events in $[0, t]$ occurring for individual i; $Y_i(t)$ is an adapted left-continuous at risk process, which takes the value 1 when individual i is at risk for an observed event and 0 otherwise. $\mathbf{X}_i(t)$ is a vector process, giving the value of p covariates over time for individual i. Suppose that time to death is independent of censoring, and the covariate processes are adapted and have sample paths that are left-continuous step functions with right-hand limits [17, Lemmas 1.4.1 and 2.3.1 and Theorem 4.2.3]. Then, the compensator for N_i is $A_i(t) = \int_0^t Y_i(u)\lambda_i(u)\,\mathrm{d}u$, where $\lambda_i(u)$ is the hazard function for individual i. The subject-specific martingale process is $M_i(t) = \int_0^t \mathrm{d}N_i(u) - \int_0^t Y_i(u)\lambda_i(u)\,\mathrm{d}u$. Transforms of these martingales provide a family of residual processes. Consider a predictable, locally bounded, possibly vector-valued process $H_i(t)$, defined in terms of data on the ith and possibly other subjects prior to t. The martingale transform $R_i(t) = \int_0^t H_i(u)\,\mathrm{d}M_i(u)$ is itself a mean-zero martingale. Furthermore, $\mathrm{cov}[R_i(s), R_j(t)] = 0$ for $i \neq j$, although R_i and R_j are not independent unless H_i and H_j are independent; $\mathrm{var}[R_i(t)] = \mathrm{E}\int_0^t Y_i(u)H_i(u)^{\otimes^2}\lambda_i(u)\,\mathrm{d}u$, where $a^{\otimes^2}$ is the outer

product of vector **a** and has (j, k) element = $a_j a_k$. Having fitted a model, one has estimates $\hat{\beta}$ and $\hat{\Lambda}_0(\cdot)$ and residual processes $\hat{R}_i(\cdot)$. The residual processes evaluated at $t = \tau$ have properties similar to the residuals from multiple linear regression. They sum to zero and are negatively correlated with asymptotic **correlation** $-1/n$. If the correct model has been fitted, then the residual processes have the patternless structure of martingales. Various choices of H_i can be used to detect different aspects of model inadequacy.

In the semiparametric Cox model, two choices of H have proved useful. $H = 1$ gives the estimated subject-specific martingale, $\widehat{M}_i(\cdot)$, which estimates the difference between the number of events observed up to t and the integrated conditional expectation under an assumed model. In the case of time-fixed covariates, $\widehat{M}_i(\tau) = \delta_i - \hat{\Lambda}_0(T_i)\exp(\hat{\boldsymbol{\beta}}' X_i)$, the Poisson residual discussed above. For the second choice, let $H_i(t, \boldsymbol{\beta}) = \mathbf{X}_i(t) - \mathrm{E}\mathbf{X}(t, \boldsymbol{\beta})$, where $\mathrm{E}\mathbf{X}(t, \boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{X}_i(t) Y_i(t) \exp[\boldsymbol{\beta}'\mathbf{X}_i(t)] / \sum_{i=1}^n Y_i(t) \exp[\boldsymbol{\beta}'\mathbf{X}_i(t)]$, the weighted mean covariate vector of those at risk at time t. The score statistic (first partial derivative of the log likelihood) for the Cox partial likelihood can be written as $\sum_{i=1}^n \int_0^\tau H_i(t, \beta)\, \mathrm{d}M_i(t, \beta)$, and $\int_0^\tau \hat{H}_i(t)\, \mathrm{d}\widehat{M}_i(t)$ is the ith score residual. Let $\mathcal{I}(t, \beta)$, a $p \times p$ matrix, denote the negative Hessian of the Cox partial log likelihood; $\mathcal{I}(\tau, \hat{\beta})$ is the observed Fisher **information matrix**. The scaled score residuals, $\mathcal{I}(\tau, \hat{\beta})^{-1} \int_0^\tau [\mathbf{X}_i(t) - \mathrm{E}\mathbf{X}(t, \hat{\beta})]\, \mathrm{d}\widehat{M}_i(t)$, are infinitesimal **jackknife** measures of influence and approximate the change in $\hat{\beta}$ that would occur if the ith individual were deleted. They are the analogs of the dfbeta residuals from linear regression, which have a very similar formula, being proportional to $(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}_i\hat{r}_i$, where $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ is the information matrix [9, p. 13]. The sum of the outer products of the dfbeta residuals gives the famous sandwich estimator for the variance of $\boldsymbol{\beta}$ [25].

Suppose the individuals are not independent but come in independent clusters. Examples include time to blindness, where the cluster is the pair of eyes, and studies of households, where the individual lifetimes are clustered into families. A simple approach is to fit a model assuming independence of individuals and then correct the estimated variance of $\hat{\boldsymbol{\beta}}$ for within-cluster covariance. The sum of the dfbeta residuals within each cluster measures the influence of the cluster on β. The sum of the outer products of these summed cluster dfbeta residuals provides a consistent variance estimator which takes into account the within-cluster correlation. This approach provides a simple computational method for the marginal models of Wei et al. [32] as one example.

The integrands of the score residuals are themselves useful residuals. Suppose we have time-fixed covariates and let $\mathbf{X}_{(k)}$ denote the covariate vector of the individual with the event at t_k. Then $\mathbf{X}_{(k)} - \mathrm{E}\mathbf{X}(t_k, \hat{\beta})$ is the Schoenfeld residual [30], useful for detecting nonproportional hazards. An alternative to proportional hazards is time-varying coefficients. Suppose the hazard function is

$$\lambda_i(t) = \lambda_0(t) \exp\left[\sum_{j=1}^p \beta_j(t) X_{ij}\right],$$

where $\beta_j(t) = \beta_j + \theta_j g_j(t)$, with $g_j(t)$ a predictable process for $j = 1, \ldots, p$. Proportional hazards holds if $\theta_j = 0$ for all j. Let

$$\mathbf{V}(t, \boldsymbol{\beta}) = \left[\sum Y_i(t) \exp(\boldsymbol{\beta}'\mathbf{X}_i)\mathbf{X}_i^{\otimes^2} \Big/ \sum Y_i(t) \times \exp \boldsymbol{\beta}'\mathbf{X}_i\right] - \mathrm{E}\mathbf{X}(t, \boldsymbol{\beta})^{\otimes^2},$$

the weighted covariate variance at t, where summation is over the n individuals. Then a Taylor's series expansion shows that

$$\mathrm{E}[\mathbf{X}_{(k)} - \mathrm{E}\mathbf{X}(t_k, \boldsymbol{\beta})] \simeq \mathbf{V}(t_k, \boldsymbol{\beta})\ \mathrm{diag}\ [\theta_j g_j(t_k)].$$

[19, 29, 30,] and a plot of the jth component of the scaled Schoenfeld residuals, $\mathbf{V}(t_k, \hat{\boldsymbol{\beta}})^{-1}[\mathbf{X}_{(k)} - \mathrm{E}\mathbf{X}(t_k, \hat{\boldsymbol{\beta}})] + \hat{\boldsymbol{\beta}}$, against event times suggest the functional form of $\beta_j(t)$, particularly when enhanced by a superimposed scatterplot smooth. A horizontal line is indicative of proportional hazards.

Extensions

Of the three survival residual definitions, the counting process martingale transformations have proved the most successful and have recently been extended to other survival models, beyond the semiparametric Cox model where they were originally developed. Lin & Spiekerman [24] consider a broad class of parametric regression models including proportional hazards and **accelerated failure time**. They suggest plotting martingale residuals against each covariate as an informal check on the correctness of functional

form for the covariate and plotting cumulative sums of martingale residuals, with several realizations simulated from the asymptotic distribution assuming correct functional form superimposed, to assess visually how unusual the observed residual pattern is. They suggest another cumulative residual plot for assessing goodness of link.

As discussed earlier, martingale residual transforms are also useful in assessing leverage and influence. Suppose the hazard has a general form as a parametric function of time and covariate vector, $\lambda_i(t) = \lambda(t, \mathbf{X}_i, \boldsymbol{\theta})$. The contribution of the ith individual to the score statistic is

$$\int_0^\tau \frac{\partial \ln \lambda(s, X_i, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \, \mathrm{d}\widehat{M}_i(s),$$

where $\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s, \mathbf{X}_i, \hat{\boldsymbol{\theta}}) \, \mathrm{d}s$. This is proportional to the infinitesimal jackknife measure of influence. Escobar & Meeker [16] suggest a quadratic form in these martingale transform residuals as a local influence statistic [11], an approximation to case deletion influence statistics, for parametric accelerated failure time models.

Martingale residuals are also useful in assessing goodness of fit in the linear regression model [1, 2]. The hazard is

$$\lambda_i(t) = \beta_0(t) + \boldsymbol{\beta}'(t)\mathbf{X}_i(t), \tag{2}$$

a linear function of the covariates with time-varying regression coefficients. Because the hazard model is linear, the estimated martingale residual processes $\widehat{M}_i(\cdot)$ are martingales rather than approximations to martingales, as with the loglinear Cox model [3]. Aalen [3] recommends two residual plots. The individuals in the data set are grouped, usually on the basis of similar covariate values, and an overall counting process, compensator, and martingale residual process are computed for each group, by summing the individual processes. The martingale residual process plot graphs each group's martingale residual against time. If the model fits well, then the plot fluctuates around the zero line and thus can identify groups or time intervals for which the model fits poorly, as shown by large deviations from zero. The Arjas plot [6] graphs the overall counting process against the estimated compensator for the group at each event time, thus comparing the "observed" and "expected" number of events (*see* **Real Time Approach in Survival Analysis**). The null configuration is the 45° line. Because the model is linear, the martingale residuals, $\widehat{M}_i(\tau)$, are directly useful in evaluating the correct functional form of covariates, unlike the Cox model, where the transformed residual $\widehat{M}_i(\tau)/\hat{\mu}_i$ is helpful for the partial residual plot. Aalen [3] recommends linear regression of the martingale residuals on curvilinear transformations of the covariates, such as low-order polynomials, to detect nonlinear covariate effects. Martingale transforms are useful for assessing influence. McKeague & Sasieni [28] consider a partially parametric linear hazard model with time-fixed covariates and hazard function

$$\lambda_i(t) = \boldsymbol{\beta}_1(t)'\mathbf{X}_{1i} + \boldsymbol{\beta}_2'\mathbf{X}_{2i},$$

where $\mathbf{X}_{1j}$ and $\mathbf{X}_{2i}$ are q- and p-dimensional covariates. They show that the infinitesimal jackknife influence measure for person i on $\hat{\boldsymbol{\beta}}_2$ is $\mathrm{var}(\hat{\boldsymbol{\beta}}_2) \int \{X_{2i} - \mathrm{E}[X_{2i}|X_{1i}(t)]\} W_{ii} \, \mathrm{d}\widehat{M}_i(t)$, where W_{ii} is the ith diagonal element of a user-defined weight matrix.

It is clear that martingale transform residuals provide a wide variety of diagnostic techniques for a broad class of survival models. As with linear regression, there is no single "one-size-fits-all" general-purpose residual.

References

[1] Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes, *Springer Lecture Notes in Statistics* **2**, 1–25.

[2] Aalen, O.O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**, 907–925.

[3] Aalen, O.O. (1993). Further results on the non-parametric linear regression model in survival analysis, *Statistics in Medicine* **12**, 1569–1588.

[4] Aitkin, M. & Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM, *Applied Statistics* **29**, 156–163.

[5] Andersen, P., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, London.

[6] Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model, *Journal of the American Statistical Association* **83**, 204–212.

[7] Baltazar-Aban, I. & Pena, E.A. (1995). Properties of hazard-based residuals and implications in model diagnostics, *Journal of the American Statistical Association* **90**, 185–197.

[8] Barlow, W.E. & Prentice, R.L. (1988). Residuals for relative risk regression, *Biometrika* **75**, 65–74.

[9] Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

[10] Cleveland, W.S., Grosse, E. & Shyu, W.M. (1992). Local regression models, in *Statistical Models in S*, J.M. Chambers & J.J. Hastie, eds. Wadsworth and Brooks, Pacific Grove, pp. 309–376.

[11] Cook, R.D. (1986). Assessment of local influence (with discussion), *Journal of the Royal Statistical Society, Series B* **48**, 133–169.

[12] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[13] Cox, D.R. & Snell, E. (1968). A general definition of residuals (with discussion), *Journal of the Royal Statistical Society, Series B* **30**, 248–275.

[14] Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data, *Journal of the American Statistical Association* **72**, 27–36.

[15] Crowley, J. & Storer, B. (1983). Comment on "A reanalysis of the Stanford heart transplant data", by M. Aitkin, N. Laird, and B. Francis, *Journal of the American Statistical Association* **78**, 277–281.

[16] Escobar, L.A. & Meeker, W.Q. (1992). Assessing influence in regression analysis with censored data, *Biometrics* **48**, 507–528.

[17] Fleming, T. & Harrington, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[18] Grambsch, P.M. (1995). Goodness-of-fit and diagnostics for proportional hazards regression models, in *Recent Advances in Clinical Trial Design and Analysis*, P.F. Thall, ed. Kluwer, Boston, pp. 95–112.

[19] Grambsch, P.M. & Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika* **81**, 515–526.

[20] Grambsch, P.M., Therneau, T.M. & Fleming, T.R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models, *Biometrics* **51**, 1469–1482.

[21] Hall, C.B., Zeger, S.L. & Bandeen-Roche, K.J. (1996). Adjusted variable plots for Cox's proportional hazards regression model, *Lifetime Data Analysis* **2**, 73–90.

[22] Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data, *Applied Statistics* **26**, 227–237.

[23] Laird, N. & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques, *Journal of the American Statistical Association* **76**, 231–240.

[24] Lin, D.Y. & Spiekerman, C.T. (1996). Model checking techniques for parametric regression with censored data, *Scandinavian Journal of Statistics* **23**, 157–179.

[25] Lin, D.Y. & Wei, L.J. (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association* **84**, 1074–1079.

[26] Mallows, C.L. (1986). Augmented partial residuals, *Technometrics* **28**, 313–319.

[27] McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[28] McKeague, I.W. & Sasieni, P.D. (1994). A partly parametric additive risk model, *Biometrika* **81**, 501–514.

[29] Pettitt, A.N. & Bin Daud, I. (1990). Investigating time dependence in Cox's proportional hazards model, *Applied Statistics* **39**, 313–329.

[30] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**, 239–241.

[31] Therneau, T., Grambsch, P. & Fleming, T.R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**, 147–160.

[32] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.

[33] Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM, *Applied Statistics* **29**, 268–275.

(*See also* **Diagnostics**; **Survival Analysis, Overview**)

P.M. GRAMBSCH, THOMAS R. FLEMING & T.M. THERNEAU

Residuals

For the linear **regression** model with response $\boldsymbol{y}$ and fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, the residuals are the vector of differences $\boldsymbol{y} - \hat{\boldsymbol{y}}$. Their great importance in the analysis of data lies in their use for checking agreement between the fitted model and the data. With a few exceptions, any pattern in the residuals is evidence either of an inadequate model or of irregularities in the data, such as **outliers**. The pattern suggests how the model may be improved.

In the next section, we describe some general purpose plots for regression residuals, which are helpful in detecting systematic differences between the fitted model and the data. Normal probability plots and simulation envelopes for their interpretation are described in the following section, before a discussion of deletion residuals. The succeeding two sections demonstrate two useful **graphical** procedures involving residuals: added variable plots for the inclusion of a new **explanatory variable** and the related constructed variable plots, demonstrated for a **power transformation** of the response. Residuals for **generalized linear models** and the constructed variable plot for examining the goodness of the link function conclude the entry.

Attention throughout is on the patterns made by residuals in suitable plots. Related material on influence of individual observations is in **Diagnostics**. The extension to the effect of groups of observations is described in the article on the **Forward Search**. Statistics based on aggregations of the residuals over the data are discussed in **Goodness of Fit**.

Least-Squares Residuals

In the **multiple regression** model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

$\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters and it is assumed that the additive errors of observation ε are independently distributed with constant variance σ^2. Also in (1) $\boldsymbol{X}$ is the $n \times p$ **matrix** of carriers, that is of explanatory variables and perhaps functions of them, such as quadratics (*see* **Polynomial Regression**) and interaction terms. The observation y_i together with $\boldsymbol{x}_i^T$, the ith row of $\boldsymbol{X}$, form the ith case. The **least-squares** estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{2}$$

The least-squares residuals are given by

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}. \end{aligned} \tag{3}$$

In (3) $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\boldsymbol{H}$ is the "hat" matrix, so called because $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$.

Before fitting any model, the data should be plotted to reveal the structure and suggest appropriate models. Once a multiple regression model has been fitted useful plots of residuals include the following:

- Residuals against fitted values $\hat{\boldsymbol{y}}$, to check for constancy of variance. If the variance of the residuals seems to increase with $\hat{\boldsymbol{y}}$, either weighted regression or a transformation of the data may be appropriate.
- Residuals against a variable $\mathbf{x}_{\text{out}}$ not currently included in the model. Any relationship between the two suggests including x_{out} in the model.
- Similarly, the residuals can be plotted against the variables $\boldsymbol{x}_{\text{in}}$ already in the equation. Any structure suggests that either a higher-order term should be included in the model, for example, $\boldsymbol{x}_{\text{in}}^2$ if $\boldsymbol{x}_{\text{in}}$ appears linearly in the model, or that $\boldsymbol{x}_{\text{in}}$ be replaced by a function $f(\boldsymbol{x}_{\text{in}})$, for example, $\log(\boldsymbol{x}_{\text{in}})$.
- Residuals e_i against lagged residuals e_{i-1}. If the observations are in time (or space) order, any pattern in this plot would suggest correlation between the observations, when ordinary least squares is no longer applicable. **Time series** methods should be used, for example the **structural time series** models described by Harvey [11].

Often, it does not make much difference whether the residuals used in plotting are the least-squares residuals $\boldsymbol{e}$ or the studentized or deletion residuals defined in the next two sections.

As a first example of the usefulness of these plots, we take the data from Royston and Altman [13] on mandible length as a function of gestational age in 167 fetuses with ages from 12 weeks. The data, plotted in Figure 2 of **Goodness of Fit**, show a clear

linear relationship, which is statistically highly significant. The plot of the residuals $\boldsymbol{e}$ from the regression of length on age against $\hat{y}$ m, given in Figure 1, shows that the variance of the observations is not constant and partly increases with fitted value. This is not surprising with nonnegative observations ranging from 8 to 45. If the percentage accuracy of the measurements is constant, a logarithmic **transformation** of the response $\boldsymbol{y}$ would yield a response with errors of constant variance. An analysis of these data with $\log(\boldsymbol{y})$ as response is in **Diagnostics**.

The second example is of data on the volume $\boldsymbol{y}$ of 70 shortleaf pine as a function of tree diameter x_1 and of height x_2. The data are tabulated by Atkinson [2] who discusses appropriate models. As a first analysis $\boldsymbol{y}$ was regressed on x_1 and x_2. Comparison of the geometry of the trunk of a pine tree with that of a cone suggests that a term in x_1^2 might also have to be included in the model. To help investigate this suggestion Figure 2 shows the plot of $\boldsymbol{e}$ against x_1. This interesting plot shows strong curvature, suggesting indeed that a term in x_1^2 should be included. The increasing scatter in the plot also suggests that the data may need transformation – the volumes range from 2.0 to 163.5 cubic feet, so measurement errors of constant variance are implausible.

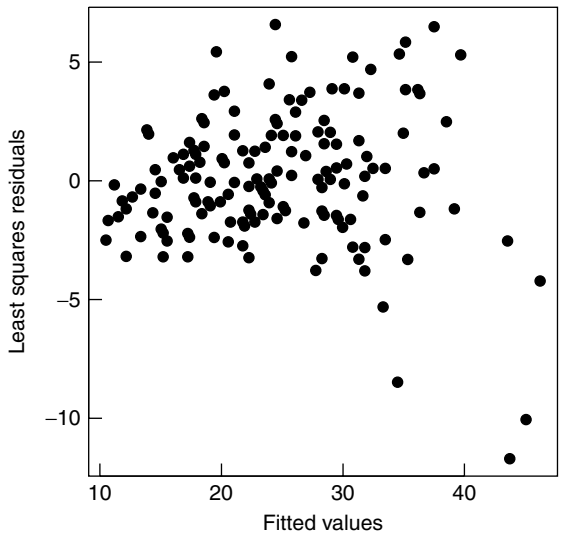

Figure 1 Mandible length data. Residuals $\boldsymbol{e}$ and fitted values $\hat{\boldsymbol{y}}$ from regression of length on gestational age. The increasing variance with $\hat{\boldsymbol{y}}$ suggests a transformation of the response

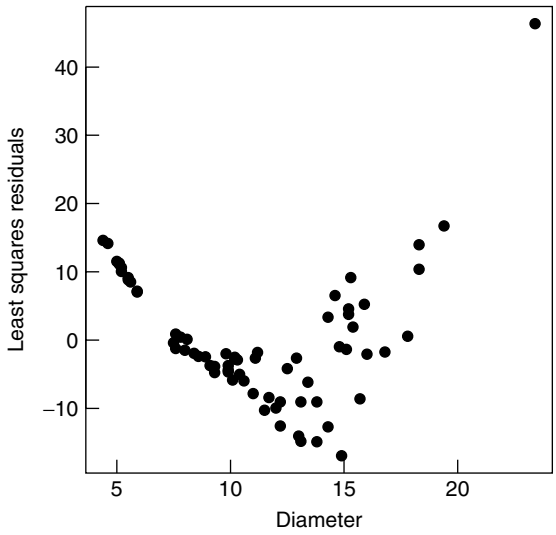

Figure 2 Pine data. Residuals $\boldsymbol{e}$ from regression of volume on diameter and height against diameter, x_1. The curvature in the plot suggests inclusion of a term in x_1^2. The heteroscedasticity again suggests a transformation of the response

Studentized Residuals and Envelope Plots

The plots of the preceding section indicate in a general way which, if any, departures are present from a model. In this and the following sections, descriptions are given of methods involving residuals, which are specific for particular problems, sometimes being derived from score tests. But we start with further consideration of residuals for the regression model.

The residuals $\boldsymbol{e}$ are not independent, nor do they have the same variance. It follows from (2) that they have covariance matrix $(\boldsymbol{I} - \boldsymbol{H})\sigma^2$, so that $\text{var}(e_i) = \sigma^2(1 - h_i)$, with h_i the ith diagonal element of $\boldsymbol{H}$. The *studentized residuals* r_i are given by

$$r_i = \frac{e_i}{s\sqrt{1 - h_i}}, \tag{4}$$

where $s^2 = \sum e_i^2/(n - p)$ is used to estimate σ^2. Although they are not uncorrelated, the r_i all do have variance one. Unfortunately, the nomenclature for residuals is not standard. The r_i are sometimes known as standardized residuals as well as *(internally) studentized residuals*.

The **normality** of the errors ε_i can be checked by a normal probability, or Q–Q, plot of the residuals e_i or

of the studentized residuals r_i. Figure 3 of **Goodness of Fit** gives a normal probability plot of the residuals e_i for the data on mandible length used in Figure 1. This plot is curved and suggests that the residuals are far from normal. However, the plot of the residuals from a fit of a quadratic in age to $\log(\boldsymbol{y})$ in Figure 1 of **Diagnostics** is much straighter, indicating that the transformation helps achieve normality.

The interpretation of residual plots is aided by an indication of how straight the plot can be expected to be. This guidance can be provided by **simulation**. One way is to simulate m sets of data using the fitted model, to calculate the n values of r_i for each simulation and then to produce m Q–Q plots, one per simulation. A comparison is made by eye to see whether the plot of the observed r_i differs in any systematic way from the m simulated plots. A more objective comparison, following Atkinson [1], is to use a simulation envelope. An envelope with 95% content can be constructed by taking the maximum and minimum, at each of the n observational points, of $m = 39$ simulated probability plots. The probability is then 1/40 that the observed value is the largest of itself and the 39 simulated values and 1/40 that it is the smallest, making the probability 5% in all that the observed value lies outside the envelope at each of the n plotting positions. The probability that, for example, at least one point lies outside the envelope is larger. However, the envelopes provide a useful calibration of probability plots, even if the exact probability of the observed plot lying in part outside the envelope is not known.

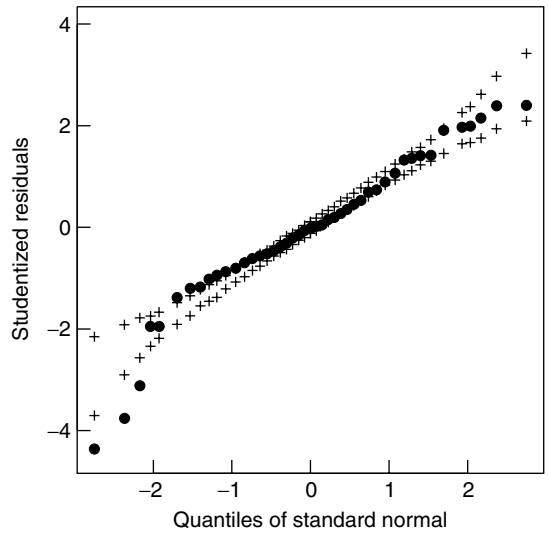

Figure 3 Mandible length data. Normal Q–Q plot of studentized residuals r_i from regression on age: •, residuals; +, simulation envelope. There is clear evidence of departure from normality

For many models, the simulations require estimation of the parameters of the model, in order to generate the simulated values of the data. However, the parameter estimates $\hat{\boldsymbol{\beta}}$ and s^2 are not required for the studentized residuals r_i from linear regression, as their distribution does not depend on the mean and variance of the y_i, but only on the **correlation** induced in the fitting process. The studentized residuals from fitting the linear model with the same matrix of carriers $\boldsymbol{X}$ as the data to a random normal sample therefore have the required distribution.

Figure 3 shows a normal Q–Q plot of the studentized residuals for the mandible length data. The residuals clearly show signs of nonnormality, both in the lower tail of the distribution and in the more moderate values where both positive and negative residuals fall outside the envelope. A smooth envelope was obtained by increasing the number of simulations to 119 and taking the third largest and third smallest values, again giving an envelope with pointwise 95% content. Because of the number of observations, 167, the central part of the plot is congested unless some points are omitted. Here, away from the tails of the distribution, every third or fifth point has been plotted.

Although this example is for studentized residuals from the **normal distribution**, the procedure can be used for other quantities, such as the Cook statistic for influence described in **Diagnostics**. Nonnormal data can also be simulated, for example, for generalized linear models, although, as was mentioned above, the parameters of the model fitted to the data may have to be used in the simulation. Examples for regression models are given by Atkinson [1] and by Venables and Ripley [18].

Deletion Residuals

Even if the errors ε_i have a normal distribution, the distribution of the studentized residuals is not normal, the distribution of r_i^2 being a scaled **beta**. In this section, we discuss the *deletion* or *(externally) studentized residual* r_i^*, which has a **Student's *t* distribution**. This is obtained from (4) on replacing

s^2 as an estimate of σ^2 by the deletion estimate $s^2_{(i)}$, that is, the mean square estimate from the $n-1$ observations excluding case i.Then

$$r_i^* = \frac{e_i}{\sqrt{s^2_{(i)}(1-h_i)}}, \tag{5}$$

has a t distribution on $n-p-1$ **degrees of freedom**.

In common with other deletion quantities for linear regression, the value of $s^2_{(i)}$ can be found exactly from the fit to all the data, the e_i and the leverage measures h_i. If $S(\hat{\beta})$ is the residual sum of squares for all n cases and $S(\hat{\beta}_{(i)})$ is the same without case i, the two estimates of σ^2 are

$$s^2 = \frac{S(\hat{\beta})}{n-p} \quad \text{and} \quad s^2_{(i)} = \frac{S(\hat{\beta}_{(i)})}{n-p-1}. \tag{6}$$

The relationship

$$S(\hat{\beta}_{(i)}) = S(\hat{\beta}) - \frac{e_i^2}{1-h_i} \tag{7}$$

provides a means of calculating $s^2_{(i)}$ and so r_i^*.

A discussion of deletion quantities such as r_i^* is given in **diagnostics**. If there is some systematic departure from normality, as is indicated by Figure 3, there is little to choose between plotting the studentized residuals r_i and the deletion residuals r_i^*, particularly if a simulation envelope is used to aid interpretation of the plot. However, if an outlier is present, particularly, at a leverage point, that is with a value of h_i close to one, the outlier will be revealed by the deletion residuals: the value of $s^2_{(i)}$ will be small, due to deletion of the outlier and so the value of the deletion residual will be large. An informative alternative derivation of $r^*_{(i)}$ is as the t statistic for the presence of an outlier, that is

$$r_i^* = \frac{y_i - \hat{y}_{(i)}}{\text{s.e. } (y_i - \hat{y}_{(i)})} = \frac{y_i - x_i^T \hat{\boldsymbol{\beta}}_{(i)}}{\text{s.e. } (y_i - \hat{y}_{(i)})}. \tag{8}$$

Q–Q plots, with envelopes, of deletion residuals are given by Atkinson [1]. Discussions of deletion diagnostics are also given by Belsley et al. [4], by Cook and Weisberg [7] and by Ryan [14].

Added Variable Plots

Figure 2 is a plot, for the pine data, of the residuals from the regression of volume on x_1 (diameter) and x_2 (height). One implication is that a term in x_1^2 should be considered for addition to the model. This implication can, of course, be examined by fitting the augmented model including x_1^2 and testing the extra term. To find out how the result of this test depends on individual observations an added variable plot can be used, which is a plot of two sets of residuals.

In general, the model $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ has been fitted and we are interested in the augmented model

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} + \boldsymbol{w}\gamma, \tag{9}$$

where $\boldsymbol{w}$ is $n \times 1$ and γ is scalar. The least-squares estimate $\hat{\gamma}$ can be found in the usual way by fitting (9) or by using the formulation of multiple regression as a series of linear regressions. For this, it is helpful to extend the notation for the least-squares residuals, (3) and let

$$\boldsymbol{e}(y) = \boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}. \tag{10}$$

Similarly, the residuals from regression of $\boldsymbol{w}$ on $\boldsymbol{X}$ are

$$\boldsymbol{e}(w) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{w} = \boldsymbol{A}\boldsymbol{w}. \tag{11}$$

Then $\hat{\gamma}$ is found by the regression, through the origin, of $\boldsymbol{e}(y)$ on $\boldsymbol{e}(w)$, that is

$$\hat{\gamma} = \frac{\boldsymbol{e}^T(w)\boldsymbol{e}(y)}{\boldsymbol{e}^T(w)\boldsymbol{e}(w)}. \tag{12}$$

Further, the t test for γ when (9) is fitted is exactly that from (12), where the estimate of σ^2 is on $n-p-1$ degrees of freedom, rather than the apparent $n-1$. The output of statistical **software** may need adjustment on this point.

The added variable plot is the residual plot of $\boldsymbol{e}(y)$ against $\boldsymbol{e}(w)$. An advantage of the plot over that of $\boldsymbol{e}(y)$ against $\boldsymbol{w}$ itself is that the use of residuals allows for the effect of other variables in the model, with which $\boldsymbol{w}$ might be highly correlated. Since the plot has a direct regression interpretation, it can provide insight into the effect of individual cases on the evidence for including $\boldsymbol{w}$.

As an example, we return to the pine data. When volume is regressed on x_1 and x_2, the regression on x_2 is not significant. If x_2 is dropped and, following the indication of Figure 2, x_1^2 is included, the new variable is highly significant, with a t value of 13.49. It is then sensible, since much of the variability in the data has been accounted for, to check whether x_2 should now be included. The added variable plot

of Figure 4 accordingly shows the residuals from regressions on x_1 and x_1^2. There is a clear trend in the plot, suggesting the inclusion of x_2 and, in fact, the t value is 9.54. The scatter of points in the plot suggests that evidence for this regression is not confined to a few points, but is supported by all the data.

Further analysis of the pine data is given by Atkinson [2], including comparison with models for the Minitab tree data described by Ryan et al. [15]. Further examples are given by Cook and Weisberg [7] and by Atkinson [1]. In addition, both describe the use of *partial residuals* to provide a complement to the added variable plot, the partial residual plot which again has slope $\hat{\gamma}$. The procedure can be useful for suggesting functional forms $f(\boldsymbol{x})$ to replace regression on $\boldsymbol{x}$.

Constructed Variable Plots

The constructed variable plot extends the idea of the added variable plot to tests of nonlinear aspects of model specification, often through a Taylor series expansion. As an example, we take the Box–Cox power transformation of the response in a regression model.

The analysis of the data on mandible length shows appreciable evidence not only of the nonnormality

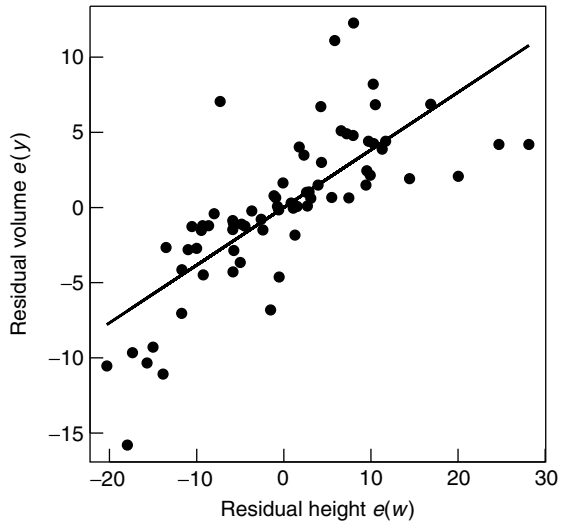

Figure 4 Pine data. Added variable plot. Residuals of $\boldsymbol{y}$ (volume) and height $\boldsymbol{w}$ after regression on diameter, x_1, and on x_1^2. Evidence that height should be included in the regression

of the residuals, Figure 3, but also of increasing variance with fitted value, Figure 1. Often, normality and constant variance can be achieved by fitting the regression model not to $\boldsymbol{y}$ but to a function of $\boldsymbol{y}$, many times $\log(\boldsymbol{y})$. The appropriate transformation frequently also leads to a simple linear model, without quadratic or interaction terms.

The logarithmic transformation is one special case of the normalized power transformation (Box and Cox [6])

$$z(\lambda) = \begin{cases} (\boldsymbol{y}^\lambda - 1)/\lambda \dot{y}^{\lambda-1} & \lambda \neq 0 \\ \dot{y} \log \boldsymbol{y} & \lambda = 0, \end{cases} \tag{13}$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i / n)$. If the residual sum of squares of the $z(\lambda)$ is $R(\lambda)$, the **profile** loglikelihood of the observations, maximized over $\boldsymbol{\beta}$ and λ, is

$$L_{\max}(\lambda) = -\left(\frac{n}{2}\right) \log \left\{ \frac{R(\lambda)}{n-p} \right\} \tag{14}$$

so that $\hat{\lambda}$ minimizes $R(\lambda)$.

For inference about the transformation parameter λ, Box and Cox suggest likelihood ratio tests using (14). A disadvantage of this likelihood ratio test is that a numerical maximization is required to find the value of $\hat{\lambda}$. For regression models, a computationally simpler alternative test is the approximate score statistic derived by Taylor series expansion of (13) as

$$\begin{aligned} z(\lambda) &\doteq z(\lambda_0) + (\lambda - \lambda_0) \frac{\partial z(\lambda)}{\partial \lambda}|_{\lambda=\lambda_0} \\ &= z(\lambda_0) + (\lambda - \lambda_0) \boldsymbol{w}(\lambda_0). \end{aligned} \tag{15}$$

In (15) $\boldsymbol{w}(\lambda_0)$ is the "constructed variable" for the transformation and can be treated as is the extra explanatory variable in (9). The approximate score statistic (*see* **Likelihood**) for testing the transformation, $T_p(\lambda_0)$, is then the t statistic for regression on $\boldsymbol{w}(\lambda_0)$ in (9). For the power transformation (13), the constructed variable is

$$\boldsymbol{w}(\lambda) = \frac{\boldsymbol{y}^\lambda \log \boldsymbol{y}}{\lambda \dot{y}^{\lambda-1}} - z(\lambda) \left(\frac{1}{\lambda +} \right) \log \dot{y}. \tag{16}$$

Provided the model for $z(\lambda)$ contains a constant, regression on (16) is equivalent, in the special cases of $\lambda = 1$ and 0, to regression on

$$\boldsymbol{w}(1) = \boldsymbol{y} \left\{ \log \left(\frac{\boldsymbol{y}}{\dot{y}} \right) - 1 \right\} \quad (\lambda = 1) \tag{17}$$

and

$$\boldsymbol{w}(0) = \dot{y} \log \boldsymbol{y} \left(\frac{\log \boldsymbol{y}}{0.5 \log y - \log \dot{y}} \right) \quad (\lambda = 0). \tag{18}$$

Box and Cox analyze a set of 48 observations on the survival times of animals in a 3×4 factorial experiment and show that a simple additive model is obtained for the reciprocal transformation, that is, for $\lambda = -1$. Rate of death, rather than survival time, is the quantity with a simple structure. If one of the observations for poison II, treatment A is changed from 0.23 to 0.13, the log transformation is indicated, rather than the reciprocal. Figure 5 shows the constructed variable plot for the altered data and the log transformation. The t value for regression on $\boldsymbol{w}$ is 1.15, so that there is no evidence, from this aggregate statistic, that the log transformation is not satisfactory. But, as the figure shows, the evidence for regression provided by the majority of the data is being annulled by the altered observation. When the observation is corrected, further transformation is indicated by the slope of the plot. The constructed variable plot for $\lambda = -1$ then shows no significant patterns.

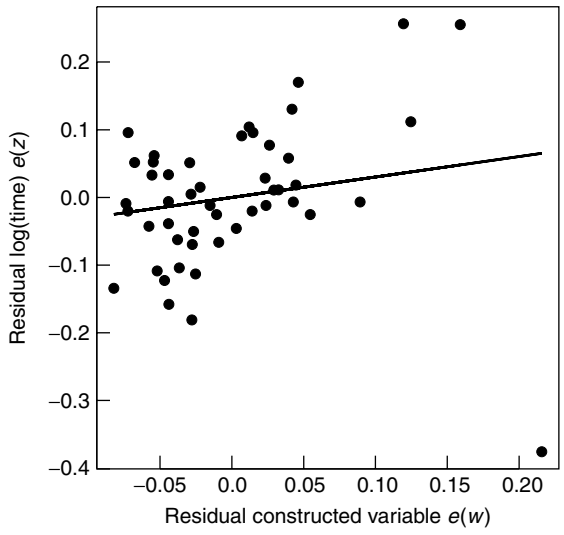

Figure 5 Poison data. Constructed variable plot for the log transformation of the altered data. The line is the regression of $\boldsymbol{e}(y)$ on $\boldsymbol{e}(w)$, which is rendered not significant by the one altered case

This example illustrates the use of the constructed variable plot of residuals against residuals in pinpointing the effect of this one observation on inference about the transformation parameter. Other plots, such as those of $\boldsymbol{e}(y)$ against $\hat{\boldsymbol{y}}$ also suggest that a transformation might be beneficial, but do not indicate the effect of individual cases on the estimated value of λ. An analysis of the poison data, including the altered data, is given by Atkinson [1] with a more complete analysis using the **forward search** in Atkinson and Riani [3]. Cook and Weisberg [8] and [9] describe the use of graphical methods in which λ can be varied interactively: the effects of the value of λ on the straightness of the Q–Q plot of residuals, or on the constructed variable plot, amongst others, can be assessed visually as a complement to the numerical choice of $\hat{\lambda}$. Similar methods can be applied to transformations of the explanatory variables or of both sides of the model mentioned in **Power Transformations**.

Generalized Linear Models

For the generalized linear model there are several definitions of residuals, with slightly different properties, which all reduce to the same quantity for linear regression. In the nomenclature of McCullagh and Nelder [12] the response y_i has expectation $E(Y_i) = \mu_i$, where μ_i is related to the linear predictor $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ by the link function $g(\mu_i) = \eta_i$. The variance of Y_i is given by $\text{var}(Y_i) = \phi V(\mu_i)$, where ϕ is the scale factor and $V(\mu_i)$ is the variance function. The **likelihood ratio** statistics for testing hypotheses about the parameters of the linear predictor are based on the differences of the scaled deviances $D(\beta)/\phi$. For the normal theory regression model of earlier sections, the link is the identity, $g(\mu_i) = \mu_i$, $\phi = \sigma^2$ and the deviance $D(\hat{\beta}) = R(\hat{\beta})$, the residual sum of squares. A slightly fuller description is given in **Goodness of Fit**. Three residuals are as follows:

- **Pearson Residual.**

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V}(\hat{\mu})}, \tag{19}$$

so named since

$$\sum r_{Pi}^2 = \phi X^2, \tag{20}$$

where X^2 is Pearson's Goodness-of-Fit test (*see* **Chi-square Tests**).

- **Deviance Residual.** Since the observations are independent, the deviance D can be written as

$$D = \sum_{i=1}^{n} d_i^2. \tag{21}$$

The deviance residual is then

$$r_{Di} = d_i \text{sign } (y_i - \hat{\mu}_i), \tag{22}$$

so that r_{Di} and r_{Pi} have the same sign. The distribution of r_{Di} is closer to normal than that of r_{Pi}, although neither will be at all normal for small counts.

- **Deletion Residual.** For least-squares regression, the change in the residual sum of squares on the deletion of the ith case is given by (7), which is the square of the unscaled version of the deletion residual r_i^* (5). Taylor expansion of the change in deviance yields the deletion residual

$$r_{Gi} = \left\{ r_{Di}^2 + \frac{h_i}{1 - h_i} r_{Pi}^2 \right\}^{1/2} \text{sign } (y_i - \hat{\mu}_i), \tag{23}$$

which reduces to $e_i / \sqrt{(1 - h_i)}$ for linear regression. Unless points of high leverage are present (that is some h_i are near one), the deletion residual is close to the deviance residual.

The Pearson and deviance residuals can be studentized by division by $\sqrt{\{\hat{\phi}(1 - h_i)\}}$. The deletion residual can be scaled by the deletion estimate $\hat{\phi}_{(i)}$. In addition, the *working residuals* come from the last stage of the iteratively reweighted least-squares fitting algorithm used for generalized linear models.

Although residuals for generalized linear models can be informative, they are not so helpful as those for normal regression models. One problem is that the discreteness of the data, for **Poisson** and **binomial** models, can induce patterns that have nothing to do with good or bad fit. For example, Pearson residuals for binary data can only take two values, being proportional to either $1 - \hat{\mu}_i$ or $-\hat{\mu}_i$.

As an example where residuals are informative for binomial data with large n_i we turn to the data on the mortality of beetles from Bliss [5], analyzed in **Goodness of Fit**. There are readings at eight dose levels. When a **logistic model** is fitted with linear predictor $\eta_i = \beta_o + \beta_1 x_i$, the residual deviance is 11.23. Since $\phi = 1$ for the binomial distribution, the value is large although not significant at the 5% level when compared with χ_6^2. Figure 6 is a plot, against dose, of the observed and fitted proportions of insects killed. The figure suggests some systematic difference between the two sets of values, which is revealed in Figure 7. This shows the deviance residuals r_{Di} against x_i. There appears to be a U-shaped relationship between residuals and dose, suggesting perhaps that a quadratic term in x should be included. Inclusion of x^2 in the logistic model reduces the residual deviance to 3.19, a clear improvement in the model.

Goodness of Link Plot

In **Goodness of Fit**, a goodness of link test was derived for Bliss's beetle data using the constructed variable $\hat{\boldsymbol{\eta}}^2$. This, like the addition of x^2 in the logistic model, was significant, with a t value of 2.70. It is however possible that this aggregate value is being caused by one or a few cases, as happened in the analysis of the altered survival data. We therefore prepare a constructed variable plot for this test.

The plot is of the Pearson residuals from the fit of the logistic model against the residuals of $\hat{\boldsymbol{\eta}}^2$

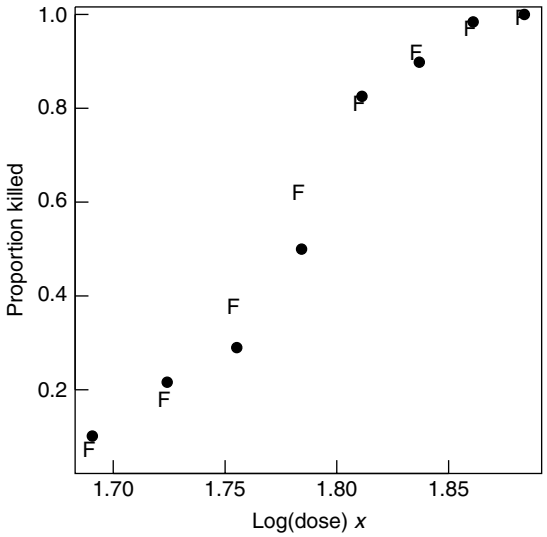

Figure 6 Beetle data. Observed and fitted values for a logistic model in logdose: ●, observed proportions; *F*, fitted. There seems to be a systematic difference between the two

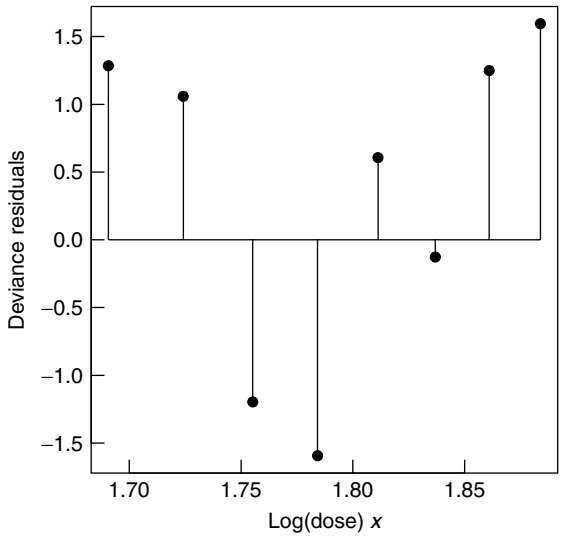

Figure 7 Beetle data. Deviance residuals against x. The curved pattern suggests including a term in x^2 in the model

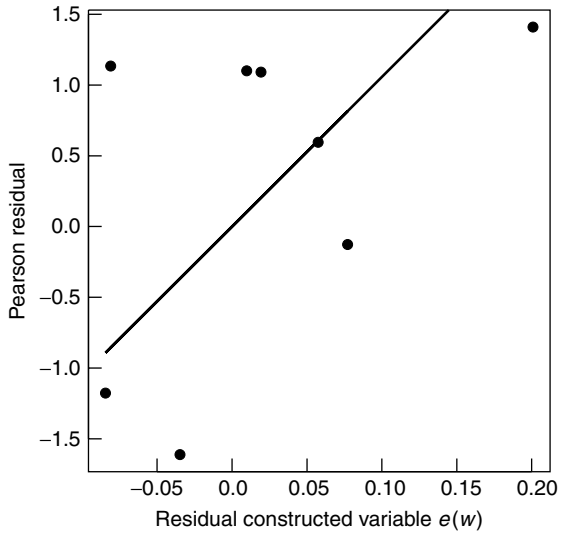

Figure 8 Beetle data. Goodness of link plot for the logistic model in $\boldsymbol{x}$, suggesting that the link is unsatisfactory

from weighted regression on x, the weights being those from fitting the generalized linear model. The plot in Figure 8 shows a trend shared by nearly all points, supporting the conclusion that the logistic link is generally unsatisfactory when combined with a simple linear model. If the complementary log–log model (*see* **Generalized Linear Model**) is used, the residual deviance is 3.45, close to the 3.19 of the quadratic model with the logistic link, but with one less parameter. In this case, the complementary log–log model appears preferable.

Discussion

A fuller discussion of residuals, especially for generalized linear models, is given by Davison and Snell [10]. Seber and Nyangoma [16] present more recent results on residuals for **multinomial** models. Therneau and Grambsch [17] is a book-length treatment of modeling **survival data** when **censoring** is present, including a chapter on residuals with SAS and **S-Plus** code. Other examples of the computation of residuals are given by Venables and Ripley [18]. Like the residuals described in this entry, these residuals are all based on least-squares or maximum likelihood estimation. Often, information on the adequacy of a model can be obtained by comparing such residuals with those from a very **robust** fit such as least **trimmed** squares or least **median** of squares. An example is given in **Diagnostics**.

The successful extraction of the information contained in residuals is aided by good graphical procedures. The flexible environment for data analysis provided by S-Plus permits the calculation and plotting of many kinds of residual (Venables and Ripley [18]). Cook and Weisberg [8] illustrate incisive graphical procedures, using their *Arc* package based on the *Xlisp-Stat* language. An introduction, with an emphasis on regression, is given by the same authors in [9]. Plots of residuals from robust fits during the **forward search** are described by Atkinson and Riani [3] for regression, response transformations, and generalized linear models.

References

[1] Atkinson, A.C. (1985). *Plots, Transformations, and Regression*. Oxford University Press, Oxford.

[2] Atkinson, A.C. (1994). Transforming both sides of a tree, *American Statistician* **48**, 307–313.

[3] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

[4] Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics*. Wiley, New York.

[5] Bliss, C.I. (1935). The calculation of the dosage-mortality curve, *Annals of Applied Biology* **22**, 134–167.

[6] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* **26**, 211–246.
[7] Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London.
[8] Cook, R.D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.
[9] Cook, R.D. & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.
[10] Davison, A.C. & Snell, E.J. (1991). Residuals and diagnostics, in *Statistical Theory and Modelling*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, pp. 83–106.
[11] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.
[12] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.
[13] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.
[14] Ryan, T.P. (1997). *Modern Regression Methods*. Wiley, New York.
[15] Ryan, B.F., Joiner, B.L. & Ryan, T.A. (1985). *Minitab Handbook*, 2nd Ed. Duxbury Press, Boston.
[16] Seber, G.A.F. & Nyangoma, S.O. (2000). Residuals for multinomial models, *Biometrika* **87**, 183–191.
[17] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.
[18] Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*, 4th Ed. Springer-Verlag, New York.

(*See also* **Model Checking**; **Model, Choice of**)

A.C. ATKINSON

Response Effects in Sample Surveys

Sample surveys are the most widely used method for obtaining information from populations of interest because of their flexibility, but they are, of course, subject to **measurement error**, as are all forms of data collection. Response effects, as distinguished from sample **biases**, are those that are caused by the methodology for obtaining answers from respondents: the questions, the context of the questionnaire, the method of data collection, and the characteristics and behavior of interviewers.

For behavioral questions, where presumably there is a "true" answer, response effects are synonymous with response errors. For attitudinal questions, there is now a general recognition that there is not a single "true" answer, but that answers are context dependent. The differences in responses that depend on the survey context are called response effects.

Respondents' Tasks

In the past decade, a major thrust in understanding response effects has been to understand better the cognitive tasks faced by a respondent in answering a question. In general, there is wide agreement among researchers regarding the substantive nature of these tasks, although different researchers use somewhat different labels (see, for example, Groves [15], Strack & Martin [25], Tourangeau [30–32], and Tourangeau & Rasinski [33]).

As a first step, respondents have to interpret the question to understand what is meant. If the question is an opinion question, they may either retrieve a previously formed opinion from memory, or they may "compute" an opinion on the spot. To do so, they need to retrieve relevant information from memory to form a mental representation of the target that they are to evaluate. In most cases, they will also need to retrieve or construct some standard against which the target is evaluated. Once a "private" judgment is formed in their mind, respondents have to communicate it to the researcher. To do so, they may need to format their judgment to fit the response alternatives provided as part of the question. Moreover, respondents may wish to edit their response before they communicate it, due to influences of social desirability and situational adequacy.

Similar considerations apply to behavioral questions. Again, respondents first need to understand what the question refers to, and which behavior they are supposed to report. Next, they have to recall or reconstruct relevant instances of this behavior from memory. If the question specifies a reference period, then they must also determine if these instances occurred during this reference period or not. Similarly, if the question refers to their "usual" behavior, then respondents have to determine if the recalled or reconstructed instances are reasonably representative or if they reflect a deviation from their usual behavior. If they cannot recall or reconstruct specific instances of the behavior, or are not sufficiently motivated to engage in this effort, then respondents may rely on their general knowledge or other salient information that may bear on their task to compute an estimate. Finally, respondents have to provide their estimate to the researcher. They may need to map their estimate on to a response scale provided to them, and they may want to edit it for reasons of social desirability.

Accordingly, interpreting the question, generating an opinion or a representation of the relevant behavior, formatting the response, and editing the answer are the main components of a process that starts with respondents' exposure to a survey question and ends with their overt report. Each of these steps may influence the direction and magnitude of response effects.

Question Comprehension

The key issue at the question comprehension stage is whether or not the respondent's understanding of the question matches what the researcher had in mind: Is the attitude object, or the behavior, that the respondent identifies as the target of the question the one that the researcher intended? Does the respondent's understanding tap the same facet of the issue and the same evaluative dimension?

Belson [1–4] asked respondents after the interview to define what the question meant and clearly demonstrated that many respondents defined terms differently than the researcher intended. In addition, many of the terms used in public opinion research do not have clearly defined lexical meanings to begin with [6, 13, 14], and respondents who ask the interviewer to provide a definition are usually instructed

to define the concept for themselves. As a result, it remains often unclear what a term meant to a respondent. In a study by Fee [12] it was observed that there were at least nine different meanings for the term "energy crisis". Similarly, the term "big government" elicited at least four distinct representations: one referred to "big government" in terms of welfare and overspending; one in terms of big business and government for the wealthy; another one in terms of a combination of federal control and diminished states' rights: and a fourth in terms of bureaucracy and a lack of democratic process. Needless to say, it is nearly impossible to interpret the responses to a question without knowing which interpretation respondents chose.

Aside from the different meanings that individual respondents attach to questions, there are cultural differences as well. This means that cross-cultural comparisons between ethnic groups, whether in the same or different countries, are always subject to differences in interpretation. Finally, the meaning of a term may change over time, posing considerable problems for trend analyses.

Recalling or Computing a Judgment

Once respondents determine what the researcher is interested in, they need to recall relevant information from memory. In some cases, respondents may have direct access to a previously formed relevant judgment that they can offer as an answer. In most cases, however, they will not find an appropriate answer readily stored in memory and will need to compute a judgment on the spot.

Whether respondents can recall a previously formed relevant judgment from memory depends on whether such a judgment has been formed in the first place, and on whether it is accessible at the time of the interview. In the case of *attitude questions*, one of the key determinants is the personal importance of the issue and the degree of respondents' personal experience with the attitude object. Not surprisingly, issues of personal importance are more likely to elicit spontaneous judgments than less important ones. Moreover, some daily activities, such as major purchasing decisions, require the evaluation of different objects and, if the decision was made recently, the evaluations formed at that point may still be accessible in memory. In addition, the likelihood that a respondent has access to evaluative judgments of an attitude object increases with the degree of the respondent's personal experience with the object [11]. Finally, if respondents have been asked a related question before, the judgment formed at that time may still be accessible in memory, provided that little interfering information has been activated in the meantime.

In the case of *behavioral questions*, a relevant answer is most likely to be directly accessible if the behavior is of personal importance and has a low frequency of occurrence [8, 22, 26]. If the behavior is a frequent one, respondents are only likely to have direct access to a judgment if the behavior is highly regular, in which case they may remember a rate of occurrence, such as "once a week" [17]. For infrequent, irregular behavior, forgetting is the most serious problem. Forgetting takes two forms: forgetting that an event occurred at all, leading to under-reporting; and remembering that the event occurred, but mis-remembering when the event occurred, called telescoping [7, 27]. Telescoping usually results in over-reporting of events, because respondents recall events as occurring in the reference period whereas they actually occurred in an earlier period. It is possible to reduce omissions by using aided recall and other cueing methods. It is also possible to reduce telescoping by use of bounded recall methods [18, 29].

Most frequently, however, answers to survey questions are not stored in memory, and respondents will need to compute a judgment when asked. In the case of attitude questions, this is because issues are complex whereas survey questions are necessarily simple, as Schuman & Kalton [21] noted. Thus, even under conditions in which respondents can retrieve an opinion on the issue from memory, this opinion may not exactly match the facet tapped in the question. Similarly, respondents are unlikely to have an appropriate answer to most behavioral questions stored in memory. Even if they can recall relevant instances, they will still need to determine if these instances fit the reference period, and so on. As a result, most of the answers that we record in surveys reflect judgments that respondents generate on the spot, in the specific context of the specific interview. They are therefore strongly influenced by the information that is accessible at that time, which is in part a function of the preceding questions.

For behavioral questions, if the behavior is frequent, but irregular, respondents will estimate by computing a rate for a short time period and extrapolating [5]. The computed rate may be too high or too low because of forgetting, telescoping, or arithmetic errors (*see* **Recall Bias**).

Formatting the Response

Once respondents have formed a judgment, they cannot typically report it in their own words. Rather, they are supposed to report their judgment by endorsing one of the response alternatives provided by the researcher. This requires that they format their response in line with the options given. Accordingly, the researcher's choice of response alternatives may strongly affect survey results [23]. Respondents use the response alternatives as reflecting the researcher's knowledge of the population distribution for an item, with the middle response category reflecting the middle of the distribution. Then, they select an answer category based on whether they believe they are above average, below average, or average relative to the population. One obvious way to avoid this problem is to omit the answer categories (*see* **Questionnaire Design**).

Editing the Response

Finally, respondents may want to edit their response before they communicate it, reflecting considerations of social desirability and self-presentation. Not surprisingly, the impact of these considerations is more pronounced in face-to-face interviews, lower in telephone interviews, and lowest in self-administered questionnaires [10, 24].

Socially desirable behavior that has been studied includes voting, donating to charity, reading, and exercising [19]. Record checks generally reveal significant over-reporting for such behavior that is reduced as the method becomes less personal.

On the other hand, reports of illegal or socially undesirable behavior, such as traffic violations, alcohol consumption, and drug use, are substantially under-reported, and this under-reporting is not usually improved by using self-administered questionnaires. Some improvements are possible by making the questions less threatening and using **randomized response** procedures, but no method has been shown to produce accurate reports of highly threatening behavior such as drunken driving [9, 34].

Interviewer Effects

The visible characteristics of interviewers have been shown to impact responses, particularly to attitude questions. The most notable examples of this have been responses to questions about racial attitudes. Questions asked of white respondents by black interviewers result in more positive attitudes toward blacks than do questions of white respondents asked by white interviewers. Similarly, questions of black respondents by black interviewers result in more militant attitudes toward whites than when the interviewers are white [16, 20].

Similar effects are noted for other ethnic groups when the ethnicity of the interviewer is visible or is revealed by the interviewer's name. Also, interviewer effects are observed based on the gender of the interviewer for questions of men and women dealing with women's rights. No effects are observed when the visible characteristics of the interviewer are unrelated to the topic of the questions. Interviewer effects may also be observed if respondents give ambiguous answers to questions. Then, interviewer **variance** may be observed in how the answer is probed and recorded (*see* **Interviewer Bias**; **Interviewing Techniques**).

The Magnitude of Response Effects

We have discussed the causes and direction of response effects, but no attempt has been made here to quantify their magnitude (see Sudman & Bradburn [27]). For specific surveys, it is difficult to predict the magnitude of effects, especially since there will often be multiple effects, sometimes operating in different directions. Generally, however, response effects are often larger than sampling biases and, for reasonable sample sizes, much larger than sampling variance.

It is often possible, using recently developed cognitive methods, to reduce response effects significantly by careful design and testing of survey instruments and methods [28]. If this is possible, then it is almost always a better use of resources to attempt to do so, or at least to measure response effects, than simply to increase sample size to reduce sampling variances.

References

[1] Belson, W.A. (1966). The effects of reversing presentation order of verbal rating scales, *Journal of Advertising Research* **6**, 30–37.

[2] Belson, W.A. (1968). Respondent understanding of survey questions, *Polls* **3**, 1–13.

[3] Belson, W.A. (1981). *The Design and Understanding of Survey Questions*. Gower, Aldershot.

[4] Belson, W.A. (1986). *Validity in Survey Research*, Gower, Brookfield.

[5] Blair, E.A. & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions, *Journal of Consumer Research* **14**, 280–288.

[6] Bolton, R.N. & Bronkhorst, T.M. (1995). Questionnaire pretesting: computer assisted coding of concurrent protocols, in *Answering Questions*, N. Schwarz & S. Sudman, eds. Jossey-Bass, San Francisco.

[7] Bradburn, N.M., Huttenlocher, J. & Hedges, L.V. (1994). Telescoping and temporal memory, in *Autobiographical Memory and the Validity of Retrospective Reports*, N. Schwarz & S. Sudman, eds. Springer-Verlag, New York.

[8] Bradburn, N.M., Rips, L.J. & Shevell, S.K. (1987). Answering autobiographical questions: the impact of memory and inference on surveys, *Science* **236**, 157–161.

[9] Bradburn, N.M., Sudman, S. and associates (1979). *Improving Interview Method and Questionnaire Design*. Jossey-Bass, San Francisco.

[10] DeMaio, T.J. (1984). Social desirability and survey measurement: a review, in *Surveying Subjective Phenomena*, Vol. 2, C.F. Turner & E. Martin, eds, Russell Sage, New York, pp. 257–281.

[11] Fazio, R.H. (1989). On the power and functionality of attitudes: the role of attitude accessibility, in *Attitude Structure and Function*, A.R. Pratkanis, S.J. Breckler & A.G. Greenwald, eds. Lawrence Erlbaum, Hillsdale.

[12] Fee, J. (1979). Symbols and attitudes *Unpublished Doctoral Dissertation*. University of Chicago.

[13] Fowler, F.J. (1989). The effect of unclear terms on survey-based estimates, in *Conference Proceedings, Health Survey Research Methods*, F.J. Fowler, ed. National Center for Health Services Research, Washington, pp. 9–12.

[14] Fowler, F.J. & Cannell, C.F. (1995). Using behavioral coding to identify cognitive problems with survey questions, in *Answering Questions*, N. Schwarz & S. Sudman, eds. Jossey-Bass, San Francisco.

[15] Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

[16] Hyman, H.A. (1954). *Interviewing in Social Research*. University of Chicago Press, Chicago.

[17] Menon, G. (1994). Judgments of behavioral frequencies: memory search and retrieval strategies, in *Autobiographical Memory and the Validity of Retrospective Reports*, N. Schwarz & S. Sudman, eds. Springer-Verlag, New York, pp. 161–172.

[18] Neter, J. & Waksberg, J. (1964). A study of response errors in expenditure data from household surveys, *Journal of the American Statistical Association* **59**, 18–55.

[19] Parry, H.J. & Crossley, H.M. (1950). Validity of responses to survey questions, *Public Opinion Quarterly* **14**, 61–80.

[20] Schuman, H. & Converse, J.M. (1971). The effects of black and white interviewers on black responses in 1968, *Public Opinion Quarterly* **35**, 44–68.

[21] Schuman, H. & Kalton, G. (1985). Survey methods, in *Handbook of Social Psychology*, Vol. I, G. Lindzey & E. Aronson, eds. Random House, New York.

[22] Schwarz, N. (1990). Assessing frequency reports of mundane behaviors: contributions of cognitive psychology to questionnaire construction, in *Research Methods in Personality and Social Psychology (Review of Personality and Social Psychology)*, Vol. 11, C. Hendrick & M.S. Clark, eds. Sage, Beverly Hills, pp. 98–119.

[23] Schwarz, N. & Hippler, H.J. (1991). Response alternatives: the impact of their choice and ordering, in *Measurement Error in Surveys*, P. Biemer, R. Groves, N. Mathiowetz & S. Sudman, eds. Wiley, Chichester, pp. 41–56.

[24] Smith, T.W. (1979). Happiness: time trends, seasonal variations, intersurvey differences, and other mysteries, *Social Psychology Quarterly* **42**, 18–30.

[25] Strack, F. & Martin, L. (1987). Thinking, judging, and communicating: a process account of context effects in attitude surveys, in *Social Information Processing and Survey Methodology*, H.J. Hippler, N. Schwarz & S. Sudman, eds. Springer-Verlag, New York, pp. 123–148.

[26] Strube, G. (1987). Answering survey questions: the role of memory, in *Social Information Processing and Survey Methodology*, H.J. Hippler, N. Schwarz & S. Sudman, eds. Springer-Verlag, New York, pp. 86–101.

[27] Sudman, S. & Bradburn, N.M. (1974). *Response Effects in Surveys: a Review and Synthesis*. Aldine, Chicago.

[28] Sudman, S., Bradburn, N. & Schwarz, N. (1995). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco.

[29] Sudman, S., Finn, A. & Lannom, L. (1984). The use of bounded recall procedures in single interviews, *Public Opinion Quarterly* **48**, 520–524.

[30] Tourangeau, R. (1984). Cognitive science and survey methods: a cognitive perspective, in *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, T. Jabine, M. Straf, J. Tanur & R. Tourangeau, eds. National Academy Press, Washington, pp. 73–100.

[31] Tourangeau, R. (1987). Attitude measurement: a cognitive Perspective, in *Social Information Processing and Survey Methodology*, H.J. Hippler, N. Schwarz & S. Sudman, eds. Springer-Verlag, New York, pp. 149–162.

[32] Tourangeau, R. (1992). Attitudes as memory structures: belief sampling and context effects, in *Context Effects in Social and Psychological Research*, N. Schwarz & S. Sudman, eds. Springer-Verlag, New York, pp. 35–47.

[33] Tourangeau, R. & Rasinski, K.A. (1988). Cognitive Processes underlying context effects in attitude measurement, *Psychological Bulletin* **103**, 299–314.

[34] Warner, S.L. (1965). Randomized response: a survey technique for eliminating error answer bias, *Journal of the American Statistical Association* **60**, 63–69.

SEYMOUR SUDMAN

Response Surface Methodology

The main purpose of this methodology is to model the response based on a group of experimental factors presumed to affect the response, and to determine the optimal setting of the experimental factors that maximize or minimize the response. The factors are all quantitative and the objective is achieved through a series of experiments.

Let $F_1, F_2, \ldots, F_k$ be k factors affecting the response y and let $\mathrm{E}(y) = f(X_1, X_2, \ldots, X_k)$, where $X_1, X_2, \ldots, X_k$ are the levels of $F_1, F_2, \ldots, F_k$, and $\mathrm{E}(y)$ is the expected response. We assume f to be a polynomial of degree d. A k-dimensional design of order d is said to be constituted of n runs of the k factors $(X_{i1}, X_{i2}, \ldots, X_{ik})$, $i = 1, 2, \ldots, n$, if from the responses recorded at the n points all of the coefficients in the dth degree polynomial are estimable.

First-Order Design

Initially a first-order design will be used to fit the model $\mathrm{E}(y) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i$. For this, one usually uses Plackett & Burman designs [15]. A Hadamard matrix $\mathbf{H}_m$ is an $m \times m$ matrix of ± 1 such that $\mathbf{H}_m' \mathbf{H}_m = m\mathbf{I}_m$, where $\mathbf{I}_m$ is the identity matrix of order m. A necessary condition for the existence of $\mathbf{H}_m$ is $m = 2$ or $m \equiv 0 \pmod 4$. If $4t - 5 \le k < 4t - 1$, in an $\mathbf{H}_{4t}$, the first column will be converted to have all ones, and any k columns of the last $4t - 1$ columns of $\mathbf{H}_{4t}$ will be identified with the coded levels of the k factors in $4t$ runs. These $4t$ runs and several central points $(0, 0, \ldots, 0)$ in coded levels constitute the design. If the lack of fit is significant, then one plans a second-order design at that center. Otherwise, one moves away from the center by the method of steepest ascent to determine a new center to plan a second-order design.

Method of Steepest Ascent

One maximizes the estimated response $\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i$ from the initial design on the contours $\sum_{i=1}^{k} X_i^2 = R^2$. The maximum occurs when $X_i \propto \hat{\beta}_i$. One decides desirable increments to proceed for factor F_i, determines the proportionality constant, and determines X_j for $j = 1, 2, \ldots, k$, $j \neq i$. In this way all coordinates in k dimensions are determined, to obtain Δ. Moving the center by incrementing $\Delta, 2\Delta, 3\Delta, \ldots$, one determines the expected $\hat{y}$. If $\hat{y}$ shows a maximum or minimum in the experimental region, then one moves the center to the setting at which $\hat{y}$ is optimum and carries out a second-order experiment. Otherwise, at a reasonable distance away from the original center, one performs another first-order experiment to determine a new path along which a center in the second-order experiment will be determined.

Second-Order Experiment

Let the design consist of F noncentral points and n_0 central points and let $n = F + n_0$. In the coded doses, without loss of generality, one assumes that:

$$\sum X_{i\alpha} = 0, \qquad \sum X_{i\alpha} X_{i\beta} = 0,$$

$$\sum X_{i\alpha} X_{i\beta}^2 = 0, \qquad \sum X_{i\alpha}^3 = 0,$$

$$\sum X_{i\alpha} X_{i\beta}^3 = 0, \qquad \sum X_{i\alpha} X_{i\beta} X_{i\gamma} = 0,$$

$$\sum X_{i\alpha} X_{i\beta} X_{i\gamma}^2 = 0, \tag{1}$$

$$\sum X_{i\alpha} X_{i\beta} X_{i\gamma} X_{i\delta} = 0, \quad \text{for } \alpha \neq \beta \neq \gamma \neq \delta;$$

$$\sum X_{i\alpha}^2 = n; \tag{2}$$

$$\sum X_{i\alpha}^4 = na,$$

$$\sum X_{i\alpha}^2 X_{i\beta}^2 = n\lambda_4, \quad \text{for } \alpha \neq \beta. \tag{3}$$

If $\mathbf{X}$ is the design matrix, then $\mathbf{X}'\mathbf{X}$ cannot be made diagonal in original parameters. However, in orthogonal polynomials of the factors settings, one may obtain **orthogonality** if $\lambda_4 = 1$. A second-order design is called orthogonal when $\sum X_{i\alpha}^2 X_{i\beta}^2 = n$.

A second-order design is said to be rotatable if the **variance** of the estimated response at $(x_1, x_2, \ldots, x_k)$ is a function of $\rho = \sum_{i=1}^{k} x_i^2$. For a rotatable design, we have $a = 3\lambda_4$ and $\lambda_4 > k/(k+2)$. The last inequality is needed to make $\mathbf{X}'\mathbf{X}$ nonsingular. By making $\mathrm{var}(\hat{y})$ the same at the settings at which $\rho = 1$ and $\rho = 0$, we obtain uniformly precise rotatable designs and, for them, λ_4 for different values of k are as shown in Table 1.

Table 1

k	λ_4
2	0.7844
3	0.8385
4	0.8704
5	0.8918

The nonzero central points are usually taken as follows:

1. A 3^n experiment or a fractional replication of a 3^n experiment in which main effects and two-factor interactions are not aliased with each other; with factor levels $-g$, 0, and g (*see* **Fractional Factorial Designs**).
2. A Central Composite Design (CCD), in which the factorial points form a 2^n experiment or a resolution 5 fractional replication with levels g and $-g$ and $2n$ axial points $(\pm\alpha, 0, \ldots, 0)$, $(0, \pm\alpha, \ldots, 0), \ldots (0, 0, \ldots, \pm\alpha)$.
3. A Box–Behnken design [3], in which the $v \times b$ incidence matrix of a **balanced incomplete block design** with parameters v, b, r, k, and λ, where $r = 3\lambda$, is used, in which the ones in each column are replaced by $\pm g$, so that the v factors are experimented in $b(2^k)$ runs.

The factor levels in the runs are determined so that the design is orthogonal, or rotatable, or uniform precision.

Let us illustrate using a CCD, which is orthogonal and rotatable in $k = 3$ factors. The noncentral points are $F = 8 + 6 = 14$ of the form $(\pm g, \pm g, \pm g)$, $(\pm\alpha, 0, 0)$, $(0, \pm\alpha, 0)$, $(0, 0, \pm\alpha)$. Let n_0 be the number of central points and let $n = 14 + n_0$. If one wants the CCD to be orthogonal and rotatable, one must have

$$8g^4 + 2\alpha^4 = 24g^4, \qquad 8g^4 = 14 + n_0.$$

Furthermore, condition (2) implies that

$$8g^2 + 2\alpha^2 = 14 + n_0.$$

An approximate solution is

$$\alpha = 2.197, \qquad g = 1.306, \qquad n_0 = 9.$$

Canonical and Ridge Analysis

Using a second-order design, one conducts an experiment and, using the data, fits a second-degree regression equation,

$$\hat{y} = \hat{\beta}_0 + \mathbf{X}'\hat{\boldsymbol{\beta}} + \mathbf{X}'\hat{B}\mathbf{X},$$

where $\mathbf{X}' = (X_1, X_2, \ldots, X_k)$ is the vector of the k factors settings, $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$, and

$$\hat{B} = \begin{bmatrix} \hat{\beta}_{11} & \frac{1}{2}\hat{\beta}_{12} & \ldots & \frac{1}{2}\hat{\beta}_{1k} \\ \frac{1}{2}\hat{\beta}_{21} & \hat{\beta}_{22} & \ldots & \frac{1}{2}\hat{\beta}_{2k} \\ \frac{1}{2}\hat{\beta}_{k1} & \frac{1}{2}\hat{\beta}_{k2} & \ldots & \hat{\beta}_{kk} \end{bmatrix}.$$

The critical point at which the derivative of $\hat{y}$ with respect to $\mathbf{X}$ is zero is given by $\mathbf{x}_0$, where $2\hat{B}\mathbf{x}_0 = -\hat{\boldsymbol{\beta}}$. Letting $\mathbf{z} = \mathbf{X} - \mathbf{x}_0$, and $\hat{y}_0 = \hat{\beta}_0 + \mathbf{x}_0'\hat{\boldsymbol{\beta}} + \mathbf{x}_0'\hat{B}\mathbf{x}_0'$, the regression equation can be rewritten as $\hat{y} = \hat{y}_0 + \mathbf{z}'\hat{B}\mathbf{z}$.

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\kappa$ be the **eigenvalues** of $\hat{B}$, and let D be the diagonal matrix with elements $\lambda_1, \lambda_2, \ldots, \lambda_k$. Let M be an orthogonal matrix such that $D = M'\hat{B}M$, and let $\mathbf{w} = M'\mathbf{z}$. Then $\hat{y} = \hat{y}_0 + \sum_{i=1}^{k} \lambda_i w_i^2$, where $\mathbf{w} = (w_1, w_2, \ldots, w_k)$. This implies that at the critical value $\mathbf{x}_0$ local maximum is attained when $\lambda_1 \leq 0$, and a local minimum is attained when $\lambda_k \geq 0$. When the inequalities are strict, $\mathbf{x}_0$ is the unique critical value, whereas when the inequalities are not strict, $\mathbf{x}_0$ is a point at which a local maximum or minimum is attained. When some λ_i are positive and some negative, one may find an absolute maximum (or minimum) of $\hat{y}$ at concentric spheres of varying radii R_i. The estimated regression function $\hat{y} = \hat{\beta}_0 + \mathbf{X}'\hat{\boldsymbol{\beta}} + \mathbf{X}'\hat{B}\mathbf{X}$ is maximized (or minimized) such that $\mathbf{X'X} = \mathbf{R}^2$, and $\mathbf{x}^*$ satisfying $2(\hat{B} - \mu I_n)\mathbf{x}^* = \hat{\boldsymbol{\beta}}$ maximizes (or minimizes) $\hat{y}$ if $\mu > \lambda_k$ (or $\mu < \lambda_1$). For different choices of μ depending on the objective, $\mathbf{x}^*$ and R^2 will be determined. The $\hat{y}$ values at those $\mathbf{x}^*$ values will be determined, and $\hat{y}$ will be plotted against R^2 to find the absolute maximum or minimum in the region of experimentation.

Further Reading

1. For some of the original ideas in this methodology, the interested reader is referred to the papers of G.E.P. Box and his co-authors (see Box & Wilson [7], Box [1, 2], Box & Youle [8], Box & Hunter [6], and Box & Draper [4].
2. For more details of this methodology, the interested reader is referred to the books by Box & Draper [5], Khuri & Cornell [10], and Myers & Montgomery [13]. For review articles

on this methodology, see Herzberg & Cox [9], Mead & Pike [11], and Myers et al. [14]. For other constructions of rotatable designs, see Raghavarao [16].

3. For the construction of Hadamard matrices, which are Plackett & Burman designs, see Raghavarao [16].
4. Taguchi and his co-workers developed different ideas to optimize responses using orthogonal arrays [17]. See also Vining & Myers [18] for combining Taguchi and response surface philosophies.
5. For handling dual responses in this methodology, see Myers & Carter [12].
6. For basic ideas of orthogonal blocking of second-order experiments, see Box & Hunter [6].

References

[1] Box, G.E.P. (1952). Multifactor designs of first order, *Biometrika* **39**, 49–57.

[2] Box, G.E.P. (1954). The exploration and exploitation of response surfaces: some general considerations and examples, *Biometrics* **10**, 16–60.

[3] Box, G.E.P. & Behnken, D.W. (1960). Some new three-level designs for the study of quantitative variables, *Technometrics* **2**, 455–475.

[4] Box, G.E.P. & Draper, N.R. (1963). The choice of a second order rotatable design, *Biometrika* **50**, 335–352.

[5] Box, G.E.P. & Draper, N.R. (1987). *Empirical Model Building and Response Surfaces*. Wiley, New York.

[6] Box, G.E.P. & Hunter, J.S. (1957). Multifactor experimental designs for exploring response surfaces, *Annals of Mathematical Statistics* **28**, 195–241.

[7] Box, G.E.P. & Wilson, K.B. (1951). On the experimental attainment of optimum conditions, *Journal of the Royal Statistical Society, Series B* **13**, 1–45.

[8] Box, G.E.P. & Youle, P.V. (1955). The exploration and exploitation of response surfaces: an example of the link between the fitted surface and the basic mechanism of the system, *Biometrics* **11**, 287–322.

[9] Herzberg, A.M. & Cox, D.R. (1969). Recent work on the design of experiments: a bibliography and a review, *Journal of the Royal Statistical Society, Series A* **132**, 29–67.

[10] Khuri, A.I. & Cornell, J.A. (1987). *Response Surfaces: Designs and Analyses*. Marcel Dekker, New York.

[11] Mead, R. & Pike, D.J. (1975). A review of response surface methodology from a biometric viewpoint, *Biometrics* **31**, 803–851.

[12] Myers, R.H. & Carter, W.H., Jr (1973). Response surface techniques for dual response systems, *Technometrics* **15**, 301–317.

[13] Myers, R.H. & Montgomery, D.C. (1995). *Response Surface Methodology: Process and Product Optimization using Designed Experiments*. Wiley, New York.

[14] Myers, R.H., Khuri, A.I. & Carter, W.H., Jr (1989). Response surface methodology: 1966–1988, *Technometrics* **3**, 137–157.

[15] Plackett, R.L. & Burman, J.P. (1946). The design of optimum multifactorial experiments, *Biometrika* **33**, 305–325.

[16] Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. Wiley, New York.

[17] Taguchi, G. (1987). *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost*. UNIPUB/Kraus International, White Plains, New York.

[18] Vining, G.G. & Myers, R.H. (1990). Combining Taguchi and response surface philosophies: dual response approach, *Journal of Quality Technology* **22**, 38–45.

D. RAGHAVARAO & S. ALTAN

Response Variable

In many data sets, there is a distinction between *response variables* and **explanatory variables**. If so, the primary questions of the analysis relate to how the response variables depend on the explanatory variables. The term *dependent variable* is often used as an alternative name for a response variable. Cox & Snell [1] indicate that response variables are the primary properties of interest and that explanatory variables hopefully explain systematic variation in the response variables.

Response variables are naturally defined when explanatory variables are fixed by an experimenter and the aim is to assess the effect of the explanatory variables on the subsequently measured response. However, it may be simply that the response variable is regarded as dependent on the explanatory variables. Alternatively, the aim of the analysis may be to predict the response on the basis of the explanatory variables. In all such cases, it is generally helpful if the distinction between variables is maintained in the analysis. The most natural example of this arises in the use of **regression** models.

There may be a number of response variables, perhaps regarded as a multivariate response variable. The formation of a single combined response variable or the separate consideration of the **univariate response** variables will often simplify the analysis. Otherwise, techniques of **multivariate analysis** become appropriate.

Cox & Snell [1] also distinguish intermediate response variables which may, for some purposes, also be treated as explanatory variables. For example, in a **clinical trial** in HIV-infected individuals, the primary response may be the time to the development of **AIDS**. The level of CD4 counts at some time after treatment could be used as an intermediate response to determine the effect of treatment on CD4 counts. The effect of treatment on the time to AIDS would also be investigated but, additionally, this could be examined when information on the changes in CD4 counts is used to define an additional explanatory variable. The latter analysis examines how much of the treatment effect on the primary response variable is accounted for by changes in the intermediate variable.

Reference

[1] Cox, D.R. & Snell, E.J. (1981). *Applied Statistics*. Chapman & Hall, London.

VERN T. FAREWELL

Restricted Maximum Likelihood

Restricted maximum likelihood estimation (REML) is an approach to estimation that maximizes the **likelihood** over a restricted parameter space. While applicable to more general models, it has most often been applied to the estimation of **variance components** in a **general linear model** with a **multivariate normal distribution**. It is an alternative to **maximum likelihood** (ML) estimation which leads to **unbiased** estimators. This method was first proposed by Patterson & Thompson [12] for a simple balanced data setting, and has been developed by Corbeil & Searle [2], Harville [8], and Dempster et al. [4]. Essentially, the procedure "adjusts" for the fact that the fixed effects are unknown when estimating components of variance. In balanced analysis of variance settings, this takes the form of an adjustment in **degrees of freedom**. In these settings, the REML estimators of the variances are the familiar unbiased **least squares** estimators.

In the case of the ordinary univariate **analysis of variance** (ANOVA) or **multiple linear regression** model, we have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is an $N \times 1$ observed data vector, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of fixed-effects parameters with $\mathbf{X}$ as its associated $N \times p$ design matrix, and $\boldsymbol{\varepsilon}$ is an $N \times 1$ vector of error terms. In this simple case we assume that the error terms are independent and distributed as $\mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The ML estimator of σ^2 is sse/N, where sse is the residual sum of squares. This, however, is a biased estimator of σ^2, since its expectation is $\sigma^2(N-p)/N$. The REML estimator of σ^2 is, instead, sse/$(N-p)$, the usual least squares estimator of the residual variance, which is unbiased.

Now suppose that the $N \times 1$ vector $\mathbf{Y}$ follows a general linear model of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of fixed-effects parameters, $\mathbf{b}$ is a $q \times 1$ vector of **random effects**, $\mathbf{X}$ and $\mathbf{Z}$ are design matrices of dimension $N \times p$ and $N \times q$, respectively, and $\boldsymbol{\varepsilon}$ is an $N \times 1$ error vector. This error term is assumed to have an $\mathrm{N}(\mathbf{0}, \mathbf{R})$ distribution, the random effects $\mathbf{b}$ are assumed to be distributed as $\mathrm{N}(\mathbf{0}, \mathbf{D})$, and $\boldsymbol{\varepsilon}$ and $\mathbf{b}$ are assumed to be independent. The variance of $\mathbf{y}$ is then $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$. The elements of $\mathbf{D}$ and $\mathbf{R}$ may be taken to be functions of a $k \times 1$ unobservable parameter vector $\boldsymbol{\theta}$. The variance of the error terms is often assumed to be $\mathbf{R} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, although restrictions on $\mathbf{R}$ may be less severe. When the variance matrix $\mathbf{V}$ is known, the maximum likelihood estimator, $\boldsymbol{\alpha}_{\mathrm{M}}$, of the fixed-effects parameters is the least squares estimator with

$$\boldsymbol{\alpha}_{\mathrm{M}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{3}$$

If $\mathbf{V}$ is unknown, then the estimator takes the same form, with maximum likelihood estimates of $\mathbf{D}$ and $\mathbf{R}$ substituted for the unknown parameters. As in the univariate case, however, these variance estimates may be biased, sometimes severely so.

Patterson & Thompson suggested dividing the data into two independent parts, each represented by an appropriate transformation of the $\mathbf{Y}$ vector. One of these is the set of error contrasts; the other is a set of linear functions of the fixed effects. Some candidates for suitable transformations are $\mathbf{SY}$ and $\mathbf{QY}$, where

$$\mathbf{S} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

$$\mathbf{Q} = \mathbf{X}'(\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R})^{-1}.$$

In univariate regression, as in model (1), $\mathbf{SY}$ is the $N \times 1$ vector of residuals using the least squares estimate of $\boldsymbol{\alpha}$, and $\mathbf{QY}$ is the $p \times 1$ vector of sums of cross-products of $\mathbf{X}$ and $\mathbf{Y}$, divided by the residual variance σ^2. The matrices $\mathbf{S}$ and $\mathbf{Q}$ are not restricted to these forms; any full-rank matrix $\mathbf{S}$ with the property that $\mathrm{E}(\mathbf{SY}) = \mathbf{0}$ for all $\boldsymbol{\alpha}$ may be used. The log likelihood of the data, L, is also divided into the two corresponding parts L' and L'', with $L = L' + L''$. Patterson & Thompson use only the log likelihood of $\mathbf{SY}$ to estimate the variances $\mathbf{D}$ and $\mathbf{R}$. They claim that when the $\boldsymbol{\alpha}$ are regarded as fixed and unknown, linear functions of these, as in $\mathbf{QY}$, cannot provide information about variance parameters. By basing the variance estimators on the error **contrasts** only, the loss of degrees of freedom due to estimating the fixed effects is accounted for. Rao [14] also breaks down the data into the same component parts, but does not use this for estimation. He shows instead that the test for adequacy of the model can be performed using the error contrasts only. The next step is to use the likelihood for $\mathbf{QY}$ to estimate $\boldsymbol{\alpha}$, substituting the variance

estimates already found. The REML estimator of $\boldsymbol{\alpha}$ takes the form of (3) above, substituting the REML estimates for the unknown variances.

The REML estimators of $\boldsymbol{\alpha}$ and its variance are, therefore,

$$\boldsymbol{\alpha}_{\mathrm{R}} = (\mathbf{X}'\mathbf{V}_{\mathrm{R}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_{\mathrm{R}}^{-1}\mathbf{Y} \tag{4}$$

and

$$\mathrm{var}_{\mathrm{R}}(\boldsymbol{\alpha}) = (\mathbf{X}'\mathbf{V}_{\mathrm{R}}^{-1}\mathbf{X}^{-1}),$$

where

$$\mathbf{V}_{\mathrm{R}} = \mathbf{Z}\mathbf{D}_{\mathrm{R}}\mathbf{Z}' + \mathbf{R}_{\mathrm{R}},$$

and $\mathbf{D}_{\mathrm{R}}$ and $\mathbf{R}_{\mathrm{R}}$ are the REML estimates. The ML estimator of $\boldsymbol{\alpha}$ and its variance have the same form as (4) above, except that ML estimates of the variances $\mathbf{D}$ and $\mathbf{R}$ are substituted in place of the REML estimates. In both cases the variance of the estimator of $\boldsymbol{\alpha}$ is underestimated since it does not account for the variability in the estimates of $\mathbf{D}$ and $\mathbf{R}$. Kacker & Harville [10] and Prasad & Rao [13] have proposed alternative forms for the variance which account for this extra variability but are computationally intensive.

The REML estimator of $\boldsymbol{\theta}$, the unknown components of $\mathbf{V}$, can be shown to maximize the log **likelihood** $L'(\boldsymbol{\theta})$, where

$$\begin{aligned} L'(\boldsymbol{\theta}) = &-\tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| \\ &- \tfrac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\alpha}_{\mathrm{R}})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\alpha}_{\mathrm{R}}) \end{aligned}$$

and $\boldsymbol{\alpha}_{\mathrm{R}}$ is the REML estimator of $\boldsymbol{\alpha}$. In contrast, the ML estimator of $\boldsymbol{\theta}$ maximizes the log likelihood $L(\boldsymbol{\theta})$, where

$$L(\boldsymbol{\theta}) = -\tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\alpha}_{\mathrm{M}})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\alpha}_{\mathrm{M}})$$

and $\boldsymbol{\alpha}_{\mathrm{M}}$ is the ML estimator of $\boldsymbol{\alpha}$, defined above. The two likelihoods differ by a single term, usually of order p.

The same REML estimators for $\boldsymbol{\alpha}$, $\mathbf{D}$, and $\mathbf{R}$ can be derived in another way, based on **Bayesian** principles. To do this, assume a vague **prior distribution** for $\boldsymbol{\alpha}$. Specifically, let $\boldsymbol{\alpha}$ have the limiting distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}^{-1} \to \mathbf{0}$. We then find the posterior distribution for $\boldsymbol{\alpha}$, and base inference upon this. We may estimate $\boldsymbol{\alpha}$ by the mean of its posterior distribution. When $\mathbf{D}$ and $\mathbf{R}$ are known, this posterior distribution is normal, with mean and variance

$$\mathrm{E}(\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{D}, \mathbf{R}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \tag{5}$$

and

$$\mathrm{var}(\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{D}, \mathbf{R}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

When the variances are unknown, we must substitute estimates of $\mathbf{D}$ and $\mathbf{R}$ for these parameters. We now compute them from the marginal likelihood of $\mathbf{Y}$, or $f(\mathbf{Y}; \mathbf{D}, \mathbf{R})$. The REML estimators are the maximum likelihood estimators of $\mathbf{D}$ and $\mathbf{R}$ based on this marginal likelihood. In contrast, the usual ML estimators are based on the complete likelihood $f(\mathbf{Y}; \boldsymbol{\alpha}, \mathbf{D}, \mathbf{R})$.

Note that in (4) the variance $\mathrm{var}_{\mathrm{R}}(\boldsymbol{\alpha})$ is not really the variance of the estimator $\boldsymbol{\alpha}_{\mathrm{R}}$; it is the variance in the posterior distribution of $\boldsymbol{\alpha}$. In the Bayesian approach we treat $\boldsymbol{\alpha}$ as a random variable, and should base inference on the posterior distribution. Alternatively, for strict sampling theorists, $\mathrm{var}_{\mathrm{R}}(\boldsymbol{\alpha})$ is a proper estimator of the variance if we use $\mathbf{R}_{\mathrm{R}}$ and $\mathbf{D}_{\mathrm{R}}$ to estimate $\mathbf{R}$ and $\mathbf{D}$ and ignore the variation in $\mathbf{R}_{\mathrm{R}}$ and $\mathbf{D}_{\mathrm{R}}$ themselves. This is true since

$$\begin{aligned} \mathrm{var}(\boldsymbol{\alpha}_{\mathrm{R}}) &\approx (\mathbf{X}'\mathbf{V}_{\mathrm{R}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_{\mathrm{R}}^{-1}\mathbf{V}\mathbf{V}_{\mathrm{R}}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_{\mathrm{R}}^{-1}\mathbf{X})^{-1} \\ &\simeq \mathrm{var}_{\mathrm{R}}(\boldsymbol{\alpha}). \end{aligned}$$

Likelihood ratio tests for $\boldsymbol{\alpha}$, however, are inappropriate, since the likelihood maximized is actually $f(\mathbf{Y}; \mathbf{D}, \mathbf{R})$ and not $f(\mathbf{Y}; \boldsymbol{\alpha}, \mathbf{D}, \mathbf{R})$.

Harville [7] shows the connection between Patterson & Thompson's REML approach to estimating $\mathbf{D}$ and $\mathbf{R}$ based on the error contrasts and the Bayesian approach which gives $\boldsymbol{\alpha}$ a vague prior normal distribution. Harville first shows that the likelihood for the error contrasts is proportional to the marginal likelihood of $\boldsymbol{\theta}$ based on the full data, $\mathbf{Y}$, or

$$f_S(\mathbf{SY}; \boldsymbol{\theta}) \propto \int f_Y(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{\alpha}, \tag{6}$$

where f_S and f_Y are the probability density functions of $\mathbf{SY}$ and $\mathbf{Y}$, respectively. If we assign an improper prior, $g(\boldsymbol{\alpha})$, to $\boldsymbol{\alpha}$ that is independent of $\boldsymbol{\theta}$, as do Dempster et al. [4], then

$$\begin{aligned} \int f_Y(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{\alpha} &\approx \int f_Y(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\alpha}) g(\boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{\alpha} \\ &= f_Y(\mathbf{Y}; \boldsymbol{\theta}). \end{aligned}$$

We see from (6) that our inferences about $\boldsymbol{\theta}$ will be the same as in REML. Also, as long as the joint prior density for $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is proportional to the prior for $\boldsymbol{\theta}$ alone, the marginal density for $\boldsymbol{\theta}$ based on $\mathbf{Y}$ is proportional to that based on the error contrasts. We therefore do not lose information by using only the error contrasts.

Secondly, suppose that the joint prior for $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ is flat relative to the data's likelihood, so that the posterior density is proportional to the likelihood. Then the ML estimator of $\boldsymbol{\theta}$ is the $\boldsymbol{\theta}$ component of the mode of the joint posterior distribution. Eq. (6) shows that, in contrast, the REML estimator of $\boldsymbol{\theta}$ based on $f_S(\mathbf{SY};\boldsymbol{\theta})$ is the mode of the marginal posterior distribution of $\boldsymbol{\theta}$.

REML estimation for a **random-coefficients model** for longitudinal data has been presented by Laird & Ware [11]. It has been described for arbitrary structural models for the within-person **covariance matrix** by Jennrich & Schluchter [9]. Both of these works use the **EM algorithm** of Dempster et al. [3] to estimate the parameters with REML. Diggle et al. [5] give an overview of REML estimation for longitudinal data. Suppose that m repeated measures are observed on each of N individuals, with Nm total observations. The distinction between REML and ML estimation is particularly important when the number of fixed parameters p is large relative to Nm, and the ML estimators are more severely biased. REML estimators also perform better when the variance matrix is near-singular, as in an example presented by Tunnicliffe-Wilson [15].

Although largely developed for the case of multivariate normality, REML estimation can be applied to other situations in which the parameter space is restricted. The term has been used when particular restrictions, such as order restrictions, are placed on the parameters (e.g. [1]; *see* **Isotonic Inference**). Many of these problems lead to the technique of **isotonic regression**. Dykstra & Madsen [6], for example, discuss some restricted estimators for the **Poisson distribution**.

References

[1] Barlow, R.E., Bartholomew, D.J., Bremer, J.M. & Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

[2] Corbeil, R.R. & Searle, S.R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model, *Technometrics* **18**, 31–38.

[3] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[4] Dempster, A.P., Rubin, D.B. & Tsutakawa, R.K. (1981). Estimation in covariance components models, *Journal of the American Statistical Association* **76**, 341–353.

[5] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1995). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[6] Dykstra, R.L. & Madsen, R.W. (1976). Restricted maximum likelihood estimators for Poisson parameters, *Journal of the American Statistical Association* **71**, 711–718.

[7] Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**, 383–385.

[8] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**, 320–337.

[9] Jennrich, R.I. & Schluchter, M.D. (1986). Unbalanced repeated measures with structured covariance matrices, *Biometrics* **42**, 805–820.

[10] Kacker, R.N. & Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *Journal of the American Statistical Association* **79**, 853–862.

[11] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

[12] Patterson, D.V. & Thompson, R. (1971). Recovery of inter-block information when the block sizes are unequal, *Biometrika* **58**, 545–554.

[13] Prasad, N.G.N. & Rao, J.N.K. (1990). The estimation of mean squared error of small-area estimators, *Journal of the American Statistical Association* **85**, 163–171.

[14] Rao, C.R. (1959). Some problems involving linear hypotheses in multivariate analysis, *Biometrika* **46**, 49–58.

[15] Tunnicliffe-Wilson, G. (1989). On the use of marginal likelihood in time series model estimation, *Journal of the Royal Statistical Society, Series B* **51**, 15–27.

NANCY R. COOK

Retrospective Study

Retrospective study is a term originally used to describe a **case–control study**, in which the previous exposures and other characteristics of cases with the disease of interest are compared with the previous exposures and other characteristics of disease-free **controls**. More generally, the term is applied to studies in which the relevant exposures and/or disease incidences have occurred before the time of the study data collection. For example, a **historical cohort study**, in which historical records of occupational exposures (*see* **Occupational Epidemiology**) and disease occurrence are analyzed just as in a prospective follow-up study, is sometimes called a retrospective study.

MITCHELL H. GAIL

Reverse Arrangement Test

The reverse arrangement test is used for evaluating whether a sequence of ordered data is derived from independent observations of the same random variable by detecting whether a significant trend underlies the observations. It is a **nonparametric** test, making no assumptions about the distribution of the input data and about a model for the possible trend.

Given a sequence of N observed values of a random variable, $x_1, x_2, \ldots, x_N$, the ith reverse arrangement, A_i, is the number of times that $x_i > x_j$ for $i > j$. The total number of reverse arrangements is $\mathbf{A} = \sum_{i=1}^{N-1} A_i$. The number $\mathbf{A}$ may range between 0 and $N \times (N-1)/2$. If $\{x_i\}$ are N independent observations of the same random variable, then A is a random variable with mean value:

$$\mu_A = N\left(\frac{N-1}{4}\right) \tag{1}$$

and variance

$$\sigma_A^2 = N(N-1)\left(\frac{2N+5}{72}\right) \tag{2}$$

If an increasing or a decreasing trend underlies the data, we may expect $\mathbf{A}$ to be respectively greater or lower than μ_A. The distribution of $\mathbf{A}$ is derived in [2] and that for N between 10 and 100 is tabulated in [1]. The values up to $N = 20$ are shown in Table 1. However, the tendency to normality is extremely rapid: when $N \geq 14$, the variable

$$z = \frac{\mathbf{A} - \mu_A}{\sigma_A} \tag{3}$$

approximately follows the standard normal distribution and can be used to reject the null hypothesis with little loss of accuracy.

To clarify the test, consider the following series of diastolic blood pressure values measured daily in a patient during a two-week monitoring period:

$$\{75; 90; 85; 82; 68; 82; 69; 64; 75; 63; 60; 73; 73; 70\}.$$

Table 1 Reverse arrangement distribution. Values of A_α^N such that the probability $(A > A_\alpha^N) = \alpha$, with N = number of observations

	α					
N	0.99	0.975	0.95	0.05	0.025	0.01
10	9	11	13	31	33	35
12	16	18	21	44	47	49
14	24	27	30	60	63	66
16	34	38	41	78	81	85
18	45	50	54	98	102	107
20	59	64	69	120	125	130

Source: [1]. Reproduced with permission of John Wiley & Sons Ltd. 1986 © John Wiley & Sons Ltd

In this sequence of $N = 14$ measures, $x_1 = 75$ is greater than eight of the following values, and thus $A_1 = 8$; $x_2 = 90$ is greater than all the next 12 observations, and $A_2 = 12$. The value of each reverse arrangement, A_i is:

$$\begin{array}{llll} A_1 = 8 & A_5 = 3 & A_8 = 2 & A_{11} = 0; \\ A_2 = 12 & A_6 = 8 & A_9 = 5 & A_{12} = 1; \\ A_3 = 11 & A_7 = 3 & A_{10} = 1 & A_{13} = 1 \\ A_4 = 9 & & & \end{array}$$

and $\mathbf{A} = A_1 + A_2 + \cdots + A_{13} = 64$. As shown in Table 1, we should reject the null hypothesis at the 5% significance level α because $\mathbf{A}$ does not fall within the range from 27 to 63. Alternatively, we could use (1–3), which give $\mu_A = 45.5$, $\sigma_A^2 = 83.417$, and $z = 2.026$. The two-tails significance of z is $p = 0.043$, and also, in this case, we should reject the hypothesis of independence at $\alpha = 5\%$.

References

[1] Bendat, J.S. & Piersol, A.G. (1986). *Random Data–Analysis and Measurement Procedures*, 2nd Ed. John Wiley & Sons, New York.

[2] Kendall, M.G. & Stuart, A. (1967). *The Advanced Theory of Statistics*, Vol. 2: Inference and Relationship, 2nd Ed. Charles Griffin & Co., London.

PAOLO CASTIGLIONI

Reversibility

Reversibility in the context of **time series** analysis implies that all modeling and statistical analysis can equally be performed on the reversed time-ordered values as on the original sequence. A reversible sequence will have a statistical time symmetry in its appearance, meaning that features of the series will not change after the time series is presented in reversed order. It only makes sense to consider reversibility in respect of stationary series, and not to consider deterministic aspects such as trend and seasonality. Irreversible or directional time series are very common and can often be identified visually, but most standard methods of time series analysis do not react to directionality. Directionality may be considered as an asymmetric assessment of time series dependence, not detected by autocorrelation or other second-order methods. Thus, models based on second-order properties, such as Gaussian linear time series models, are reversible [12]. However, if the standard innovation variables are not Gaussian, then these models exhibit directionality. Likewise, there are various nonlinear time series models that are unavoidably directional. Directionality can be incorporated in the assessment of suitability of a proposed model, and is particularly relevant when predictive use is anticipated. The classical time series of sunspot numbers and Canadian lynx data are often given as examples with directionality. The literature contains scattered references, and there is an assessment of the area by Lawrance [8].

Technical Definitions

A time series in discrete time, modeled by the random variables $[X(t), t = 0, \pm 1, \pm 2, \ldots]$ is said to be *reversible* when, for all $r = 1, 2, \ldots$, and $t = 0, \pm 1, \pm 2, \ldots$, the joint distribution of $[X(t), X(t+1), \ldots, X(t+r)]$ is equal to the joint distribution of $[X(t+r), X(t+r-1), \ldots, X(t)]$. This definition appears to be by Brillinger & Rosenblatt [3], although the first mention of the idea may be due to Daniels [5]. More limited definitions have been given in [8]. For instance, *first-order reversibility* was proposed as the case $r = 1$ of the general definition; it implies marginal **stationarity**. The term *lag reversibility* was used to denote that the joint bivariate distribution of $[X(t), X(t+r)]$ is the same as that of $[X(t+r), X(t)]$ for $r = 1, 2, \ldots$ and $t = 0, \pm 1, \pm 2, \ldots$. A particular aspect of this type of reversibility is that $\text{corr}[X(t)^2, X(t+r)] = \text{corr}[X(t), X(t+r)^2]$. These conditions can very easily be verified with actual data to give a statistical assessment of directionality.

Theoretical Results

Linear Models

A key contribution of Weiss [12] proved the result that linear autoregressive processes, with or without a moving-average component (*see* **ARMA and ARIMA Models**) are reversible if and only if they are Gaussian. For a moving-average linear model component of the form

$$b_0\varepsilon(t) + b_1\varepsilon(t-1) + \cdots + b_q\varepsilon(t-q),$$

where $\varepsilon(t), t = 0, \pm 1, \pm 2, \ldots$, is an independent and identically distributed innovations sequence of random variables, and $b_0, b_1, b_2, \ldots$ is a sequence of constants, a general condition giving reversibility is that $(b_j = b_{q-j}, j = 0, 1, \ldots, q)$; if the innovations sequence is symmetrically distributed $(b_j = -b_{q-j}, j = 0, 1, \ldots, q)$ also gives reversibility. There are direct consequences of this reversibility to the *invertibility* of linear time series models. Box & Jenkins [2] refer to this as being able to express the innovation term $\varepsilon(t)$ as a linear function of the present and past $X(t)$. If this is to be so, roots of the equation

$$1 - b_1x - b_2x^2 - \cdots - b_qx^q = 0,$$

inside or on the complex unit circle, are not allowed. With odd-order moving-average components, there is always at least one root on the unit circle; with even-order components, there is at least one root inside the unit circle. Thus, in both cases, reversibility precludes invertibility. Moving away from reversible models, a number of directional linear models have been studied. For first-order linear autoregressive models, Gaver & Lewis [6] consider exponential and gamma marginal distributions, and Lawrance [7] gives their compound Poisson innovation distribution (*see* **Contagious Distributions**); Rao et al. [11] give results for such models with self-decomposable marginal

distributions. A striking result for first-order linear models with uniform marginals is the connection between reversibility, chaos and congruential random number generators (*see* **Pseudo-random Number Generator**) [1]. The reversed uniform process is the multibranch generalization of the chaotic shift-map process. This is the continuous-valued version of the congruential random number generator.

Nonlinear Models

Most of the available results for nonlinear models concern first-order autoregressions, although in nonstandard forms. Chernick et al. [4] show that the reversed form of the Gaver & Lewis exponential model was a nonlinear model with minimization replacing addition for combining terms. Lewis & McKenzie [9] construct an extended class of processes based on the minimization operation, modeling **uniform distribution**, **Weibull** and **Pareto distribution**, among others. McKenzie [10] also develops nonreversible models in **negative binomial** variables using a thinning operation in place of multiplication. A variety of other nonlinear models have since appeared in the literature and most are irreversible.

References

[1] Bartlett, M.S. (1990). Chance or chaos? (with discussion), *Journal of the Royal Statistical Society, Series A* **153**, 321–347.

[2] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

[3] Brillinger, D.R. & Rosenblatt, M. (1967). Computation and interpretation of kth order spectra, in *Spectral Analysis of Time Series*, B. Harris, ed. Wiley, New York, pp. 189–232.

[4] Chernick, M.R., Daley, D.J. & Littlejohn, R.P. (1988). A time reversibility relationship between two Markov chains with exponential stationary distributions, *Journal of Applied Probability* **25**, 418–422.

[5] Daniels, H.E. (1946). Discussion to Symposium on Autocorrelation in Time Series, *Journal of the Royal Statistical Society* **8**, Supplement, 29–97.

[6] Gaver, D.P. & Lewis, P.A.W. (1980). First order autoregressive sequences and point processes, *Advances in Applied Probability* **12**, 727–745.

[7] Lawrance, A.J. (1982). The innovation distribution of a gamma distributed autoregressive process, *Scandinavian Journal of Statistics* **9**, 234–236.

[8] Lawrance, A.J. (1991). Directionality and reversibility in time series, *International Statistical Review* **59**, 67–79.

[9] Lewis, P.A.W. & McKenzie, E. (1991). Minification processes and their transformations, *Journal of Applied Probability* **28**, 45–57.

[10] McKenzie, E. (1986). Autoregressive moving-average processes with negative binomial and geometric marginal distributions, *Advances in Applied Probability* **18**, 679–705.

[11] Rao, P.S., Johnson, D.H. & Becker, D.D. (1992). Generation and analysis of non-Gaussian Markov time series, IEEE *Transactions on Signal Processing* **40**, 845–856.

[12] Weiss, G. (1975). Time reversibility of linear stochastic processes, *Journal of Applied Probability* **12**, 831–836.

A.J. LAWRANCE

Rheumatology

The field of rheumatology deals with clinical disorders which involve the musculoskeletal system. Rheumatology includes inflammatory and noninflammatory diseases of joints, bones, muscles, and connective tissues. The inflammatory disorders are characterized by the features of inflammation, including pain, stiffness, redness, swelling, and reduced function in the affected areas. Generally, these conditions improve with activity and are worsened by rest. The noninflammatory disorders are characterized by pain which is made worse with activity and improves with rest.

The inflammatory disorders include various forms of arthritis and the collagen vascular disorders. The prototype of inflammatory arthritis is rheumatoid arthritis, which presents with joint pain and swelling associated with stiffness, involving the small joints of the hands and feet as well as the larger joints, in a symmetric distribution. The cause of most forms of inflammatory arthritis is unknown, but it is recognized that the main problem is an inflammatory process in the lining of the joint, the synovium. These conditions tend to be chronic, with periods of exacerbation and remission. With time, the chronic persistent inflammation in the synovium leads to joint destruction, deformity, and disability.

Some forms of inflammatory arthritis have known causes. For example, septic arthritis is caused by infective agents, and some viruses are associated with arthritis. Gout, which is a crystal induced arthritis, results from uric acid deposition in the joints, and pseudo-gout, also a crystal induced arthritis, results from calcium pyrophosphate dihydrate deposition in the joints.

The collagen vascular diseases include a number of inflammatory conditions which affect the connective tissues, including the joints. These conditions are usually multisystem, affecting many organs and systems in the body. Systemic lupus erythematosus, the commonest of these conditions, affects young women, with a variable presentation, course, and prognosis. Scleroderma, also known as systemic sclerosis, also affects women, is not as common, and has an important vascular component. The vasculidites are a group of diseases characterized by inflammation in the blood vessels, and the symptoms and signs depend on the blood vessel involved. Polymyositis is a rarer condition which affects primarily the muscles: when the skin is involved it is called dermatomyositis.

The noninflammatory conditions include degenerative arthropathies, which are the most common form of arthritis, nonarticular rheumatism, and traumatic injuries. Osteoporosis is also included in the noninflammatory rheumatological disorders.

Historical Development

The term “rheuma” was introduced in the first century AD, and indicates a substance that flows. Rheumatic diseases such as gout were recognized as early as the fourth century BC, but the concept of systemic rheumatic disease was introduced in the sixteenth century. The term “rheumatologist”, to refer to the physician dealing with rheumatic diseases, is recent [21].

Gout was the first rheumatic condition to be clinically described and urate was found to be the causative factor in the nineteenth century. Until the late 1940s, almost all patients with arthritis were thought to have either gout or rheumatoid arthritis. However, in 1948 rheumatoid factor was discovered, as an antibody present in the sera of more than 80% of patients with rheumatoid arthritis, but not in patients with osteoarthritis. This further resulted in the differentiation of inflammatory forms of arthritis into those which were rheumatoid factor positive (seropositive), such as rheumatoid arthritis, and those which were rheumatoid factor negative (seronegative), which include psoriatic arthritis, which is associated with psoriasis, ankylosing spondylitis, which affects primarily the back, Reiter’s disease, which affects the back, peripheral joints and skin, and the arthritis of inflammatory bowel disease. In addition to being seronegative, the latter group of conditions have other features in common: they affect the back, are associated with skin and mucous membrane lesions, and are associated with certain genetic factors, namely HLA-B27. In the same year, the diagnosis of systemic lupus erythematosus was facilitated by the discovery of the lupus erythematosus cell preparation. Subsequently, the antinuclear factor was described and the relationship of autoantibodies to the collagen vascular diseases became clear. About 140 arthritic conditions are recognized by the **International Classification of Diseases** (ICD) codes.

With the recognition of these various forms of arthritis, it became necessary to develop criteria for

classification and diagnosis [22]. Criteria were established for the diagnosis of rheumatoid arthritis, systemic lupus erythematosus (SLE), the arteridites, and other conditions. These have facilitated research into the mechanisms involved in these conditions, as well as the inclusion of patients into therapeutic trials (*see* **Clinical Trials, Overview**). Comparisons of patients from different centers can now be achieved, and **multicenter trials** are feasible since patients with similar characteristics may be recruited. A consensus on criteria for response to treatment in rheumatoid arthritis has also been developed [10].

Types of Studies

The classical epidemiologic studies, such as **cohort** and **case–control studies** are used in rheumatological research, although the extent of epidemiologic investigation into risk factors is less than for many other diseases. **Genetic epidemiology** has been and should continue to be an area of particular interest. Types of investigations in this area include those based on sibling pairs, family studies, and the search for candidate genes through disequilibrium testing. Multicenter efforts may be needed to provide adequate sample sizes [14]. Similarly, there are many clinical trials in rheumatology which aid in the evaluation and licensing of treatments. Other investigations are based on clinical databases devoted to patients with a particular rheumatological disease, specific investigations on a specially recruited group of patients, and comparisons of different groups of patients (*see* **Administrative Databases**).

A wide variety of clinical and radiological measures in rheumatology derive from expert assessment. **Reliability studies** are therefore of particular importance in rheumatological research. Reliability of measurement within a single center or across multiple centers may be required, depending on the nature of the investigations under consideration. Reliability and validity studies are also of considerable importance for **quality of life** measurements, which are increasingly being used in research activities.

Historically, classification studies have aided in the definition of different rheumatological conditions. The approach of the collection of **"gold standard"** cases and the comparison of these to individuals with other conditions led to some progress. Recursive partitioning methodology [2] (*see* **Tree-structured Statistical Methods**) has been used in the validation of criteria. These studies continue to be useful, but the methodological requirements, such as the definition of a gold standard, become increasingly challenging as disease subgroups are defined on the basis of immunologic abnormalities, genetic factors, and so on. A related area is the evaluation of new **diagnostic tests** which involves the classic considerations of **sensitivity**, **specificity**, and **predictive value**.

Treatment

The therapeutic approach to arthritis is control of **pain** and inflammation. The choice of anti-inflammatory medication depends on the severity of the inflammatory process, and its possible effect on vital organs. Thus, for mild cases of joint inflammation, nonsteroidal anti-inflammatory drugs (NSAIDs) are used. These have been shown effective in controling inflammation in both animal models of inflammation and in the human disease. An increasing number of NSAIDs have become available, primarily because of their side-effects which motivate attempts to develop newer drugs with less toxicity. When NSAIDs are ineffective, disease modifying drugs are used. These include medications such as gold compounds, penicillamine, antimalarials, salazopyrine, methotrexate, azathioprine, cyclophosphamide, and cyclosporin (see [18, Section VI]).

Corticosteroids are used to control inflammation in the more systemic conditions, as well as in the resistant cases of joint inflammation. These drugs appeared so effective when first introduced in the late 1940s that, although there have not been controlled trials to prove their efficacy, new modalities are now measured against corticosteroids.

On the basis of the current understanding of mechanisms of inflammation, new approaches have been developed including agents that interfere with Tumor Necrosis Factor such as the anti-TNF, antibody infliximab, and the TNF receptor fusion protein etanercept. In RCTs, these later agents have been shown effective in RA, PsA, and ankylosing spondylitis and are being tested in vasculitis. Other agents directed against B cells and T cells are at various stages of development and testing.

Methodological Issues

The nature of rheumatological diseases and their medical management creates some particular methodological difficulties. An obvious initial one is consideration of the time of origin for studies of disease. Patient recall is a source for information on the onset of symptoms. But, for example, there are groups of patients with some features of SLE who have not developed full blown disease even after more than 10 years of follow-up. Similarly, referral patterns related to medical care are widely varied influencing both the time of treatment initiation and the type of clinical center within which treatment is sought. Clinical trials and database studies thus must almost inevitably have time origins which are variable and somewhat arbitrary. However, on balance, many investigators support the notion that the time of diagnosis should be used as the baseline for disease origin [19], sometimes taking it to be the time the diagnosis might have been made had an appropriate physician seen the patient.

At the "other end" of the studies, there is wide variation in the **outcome measures** used by different researchers. Short-term outcomes are common in clinical trials related to drug licensing activities (*see* **Drug Approval and Regulation**) but longer-term outcomes are often of more clinical interest. The identification of a single outcome of interest is usually difficult, as evidenced by the selection of a set of outcomes as the consensus requirements for the reporting of clinical trials in rheumatoid arthritis [11]. Other conditions have not yet reached even this stage of standardization but efforts are being made, for example, in psoriatic arthritis [16] and myositis [15, 17].

The evaluation of these multiple outcomes requires careful consideration. Many classical **multiple comparison** procedures are not relevant, based as they are on an experimental false positive error defined as one or more false positives on individual outcomes. It is difficult to provide a mathematical characterization of the "clinical insight" which jointly evaluates disease progression in terms of these multiple outcomes [4, 5].

As mentioned earlier, there is a subjective evaluative component to many of these outcomes, some evaluated by the patient, some by physicians. In addition to the reliability issue already discussed, it is also important to consider expectation bias which can arise in randomized, but unblinded, clinical trials. Epstein [8] illustrates the dramatic differences which can result when evaluation of a "promising" new treatment is done in a blinded versus unblinded manner.

The classic assumption in most statistical methods to deal with time to event data is that censoring is independent of the outcome (*see* **Censored Data**). Thus, individuals who are lost to follow-up are assumed to be no different, in terms of subsequent outcomes, to individuals who remain under follow-up. Because of the long course of most rheumatic diseases, follow-up is particularly problematic. Specific attention should be paid to the nature of lost-to-follow-up patients before undertaking studies which involve long-term outcomes. There is recent interest in mortality studies which consider the survival of arthritis patients compared with the general population, motivated partly by the **burden of disease** hypothesis. Loss-to-follow-up information is critical in these studies, which must be based on databases accrued over a considerable period of time [9].

The course of rheumatic diseases means that not only are there multiple types of outcomes (*see* **Multiple Endpoints, *P* Level Procedures**) but also multiple measurements of outcomes over time (*see* **Longitudinal Data Analysis, Overview**). A simple example is the number of inflamed joints in a patient with rheumatoid arthritis which could be measured at each clinic visit. Statistical methods for the analysis of times to a single event are therefore not appropriate in many cases. Fortunately, there is increasing methodological interest in models for longitudinal data, and much of this work is relevant to studies in rheumatology.

A particular approach which may be useful is to model disease progression as a Markov multistate model (*see* **Marker Processes**). Not only will this make use of repeated measurements on outcome variables, but it will also allow individuals to enter follow-up at different states and may help to deal with the lack of a well characterized time of disease onset. This approach has been used to study disease progression in psoriatic arthritis [13]. Other approaches should be explored [6, 14] and, of course, **goodness of fit** investigations will be required [1].

In these long-term studies of disease progression, prognostic markers of interest may also be measured at several points in time. Thus the ability to incorporate time-dependent explanatory variables would be a

useful feature of any method of analysis. The incorporation of such variables will reflect what occurs in the clinical monitoring of patients. There are, of course, other time-independent variables such as demographics and genetic factors which must also be considered.

Landmark Studies

Prognosis

The study of survival in SLE performed by Merrell & Shulman [19] provided a framework for subsequent studies of prognosis in SLE and other rheumatic diseases. They defined the onset of the disease as the time of diagnosis, and provide a rationale for its use. They also took appropriate account of patients lost to follow-up who were regarded as censored at the time at which they were last observed.

Treatment

Gold was the first antirheumatic drug to be investigated through a controlled clinical trial in RA [12], although a high placebo effect was noted. A subsequent trial by the Empire Rheumatism Council is considered the landmark article both for the efficacy of gold, and for the standard of clinical trials in rheumatology [7]. This was a multicenter trial involving 24 centers.

Disease Associations

The HLA region of chromosome 6 in man is the major histocompatibility locus. Because of its role in the immune response, it was thought to be related to the susceptibility and/or resistance to disease. It was not until landmark articles from the UK [3] and US [20] that the role of this region in disease susceptibility was confirmed. These studies provided the framework for subsequent studies of **HLA** and disease associations.

References

[1] Acquirre-Hernandez, R. & Farewell, V.T. (2002). A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models, *Statistics in Medicine* **21**, 1899–1911.

[2] Block, D.A., Moses, L.E. & Michel, B.A. (1990). Statistical approaches to classification: methods for developing classification criteria and other criteria rules, *Arthritis and Rheumatology* **33**, 1137–1144.

[3] Brewerton, S.A., Hart, F.D., Nicholis, A., James, D.C.O. & Sturrock, R.D. (1973). Ankylosing spondylitis and HLA-27, *Lancet* **i**, 904–907.

[4] Cook, R.J. & Farewell, V.T. (1996). Multiplicity considerations in the design and analysis of clinical trials, *Journal of the Royal Statistical Society, Series A* **159**, 93–110.

[5] Cook, R.J., Farewell, V.T. & Gladman, D.D. (1997). Methodology for clinical trials in rheumatology: logical foundations, *Journal of Rheumatology* **24**, 1861–1865.

[6] Cook, R.J., Yi, G.Y., Gladman, D.D. (2004). A conditionally Markov model for multivariate processes under incomplete observation, *Biometrics*, in press.

[7] Empire Rheumatism Council (1961). Gold therapy in rheumatoid arthritis: final report of a multicentre controlled trial, *Annals of Rheumatologic Diseases* **20**, 315–334.

[8] Epstein, W.V. (1996). Expectation bias in rheumatoid arthritis clinical trials, *Arthritis and Rheumatism* **39**, 1773–1780.

[9] Farewell, V.T., Lawless, J.F., Gladman, D.D. & Urowitz, M.B. (2003). Tracing studies and analysis of the effect of loss to follow-up on mortality estimation from patient registry data, *Applied Statistics* **52**, 445–456.

[10] Felson, D.T., Anderson, J.J., Boers, M., Bombardier, C., Furst, D., Goldsmith, C., Katz, L.M., Lightfoot, R. Jr., Paulus, H., Strand, V., Tugwell, P., Weinblatt, M., Williams, H.J., Wolfe, F. & Kieszak, S. (1995). American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis, *Arthritis and Rheumatism* **38**, 727–735.

[11] Felson, D.T., Anderson, J.J., Boers, M., Bombardier, C., Chernoff, M., Fried, B., Furst, D., Goldsmith, C., Kieszak, S., Lightfoot, R., Paulus, H., Tugwell, P., Weinblatt, M., Widmark, R., Williams, H.J. & Wolfe, F. (1993). The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials, *Arthritis and Rheumatism* **36**, 729–740.

[12] Fraser, T.N. (1945). Gold treatment in rheumatoid arthritis, *Annals of Rheumatic Diseases* **4**, 71–75.

[13] Gladman, D. & Farewell, V. (1995). The role of HLA antigens as indicators of disease progression in psoriatic arthritis: multivariate relative risk model, *Arthritis and Rheumatism* **38**, 845–850.

[14] Gladman, D.D. & Farewell, V.T. (1999). Progression in psoriatic arthritis (PSA): Role of time varying clinical indicators, *Journal of Rheumatology* **38**, 1130–1137.

[15] Gladman, D.D. & Farewell, V.T. (2003). HLA studies in psoriatic arthritis: Current situation and future needs, *Journal of Rheumatology* **30**, 4–6.

[16] Gladman, D.D., Helliwell, P., Mease, P.J., Nash, P., Ritchlin, C. & Taylor, W. (2004). Assessment of patients

with psoriatic arthritis, *Arthritis and Rheumatism* **50**, 24–35.

[17] Isenberg, D.A., Allen, E., Farewell, V.T., Ehrenstein, M. et al. (2004). International consensus outcome measures for patients with idiopathic inflammatory myopathies, *Rheumatology* **43**, 49–54.

[18] Kelley, W.M., Harris, E.D., Ruddy, S. & Sledge, C.B., eds. (1997). *Textbook of Rheumatology*, Vol. 1, 5th Ed. Saunders, Philadelphia, pp. 707–849.

[19] Merrell, M. & Shulman, L.E. (1995). Determination of prognosis in chronic disease, illustrated by systemic lupus erythematosus, *Journal of Chronic Diseases* **1**, 12–32.

[20] Schlosstein, L., Terasaki, P.I., Bluestone, R. & Pearson, C.M. (1973). High association of an HLA antigen, w27, with ankylosing spondylitis, *New England Journal of Medicine* **288**, 704–706.

[21] Schumacher, H.R., Klippel, J.H. & Koopman, W.J., eds. (1993). *Primer on the Rheumatic Diseases*, 9th Ed. The Arthritis Foundation, Atlanta.

[22] Silman, A.S. & Symmons, D.P.M., guest eds. (1995). *Bailliere's Clinical Rheumatology*, International Practice and Research. *Classification and Assessment of Rheumatic Diseases*, Part I. 9(2), Part II. 9(3).

VERN T. FAREWELL & D.D. GLADMAN

Ridge Regression

Ridge regression was initially promoted by Hoerl & Kennard [8, 9] within the multiple regression model:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X}$ is an $(n \times p)$ matrix of n observations on p explanatory variables, the columns of $\mathbf{X}$ have mean zero (i.e. centered), $\mathbf{1}$ is an $(n \times 1)$ vector of ones allied to the mean coefficient β_0; $\mathbf{Y}$ is the n-vector of responses, and $\boldsymbol{\varepsilon}$ the n-vector of random errors with constant variance σ^2. The class of ridge estimators of the p-vector $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, \tag{2}$$

with the class indexed by the hyperparameter k. The mean coefficient β_0 is usually estimated by $\bar{y}$, the sample mean of the observed n-vector $\mathbf{y}$; see [3] for alternative forms. The idea of regularization by adding a small constant to the diagonal of $\mathbf{X}'\mathbf{X}$ in (2) had appeared earlier in the context of function minimization, the Levenberg–Marquardt modification of Newton–Raphson, as described, for example, in [14] (*see* **Optimization and Nonlinear Equations**).

Hoerl & Kennard suggested plotting the p-coefficients, $\hat{\beta}_i(k)$, from (2) as functions of k, ranging from $k = 0$ (**least squares**) to $k = \infty[\hat{\beta}_i(\infty) = 0, i = 1, \ldots, p]$. This *ridge trace* typically shows large changes in at least some of the coefficient estimates as k increases from zero. The squared length of the vector, $\hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{\beta}}(k)$, monotonically decreases as k increases. At the same time the residual sum of squares may only show a modest increase from the least squares value at $k = 0$. One argument of Hoerl & Kennard was that the least squares estimator of $\boldsymbol{\beta}$ is unnaturally long, especially when the $\mathbf{X}$-matrix is nearly collinear (*see* **Collinearity**), and dramatic reductions in length are possible with small increases in k from zero.

From a slightly different perspective, the effect of the estimator may be judged from the amalgamated **mean squared error** of estimation,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta}).$$

This can be decomposed as the sum of variance and a squared bias term. Initially at $k = 0$ the bias is zero, but the variance may be large, especially if $\boldsymbol{X}$ is near collinear. The variance of the ridge estimator reduces as k increases from zero monotonically, the squared bias term increases, and typically the mean squared error, the sum of the two terms, decreases initially and then starts to increase as bias takes over for larger values of k. Hoerl & Kennard provided a theorem proving the existence of $k > 0$ for which the mean squared error was less than that of least squares. Unfortunately this does not provide guidance on the choice of k, which was left to the "elbowing" out of the ridge trace.

Nowadays the ridge trace has been largely abandoned in favor of explicit choices of estimator $\hat{k}$. Whatever value is chosen the improvement in mean squared error of estimation cannot be realized at all values in the parameter space of $\boldsymbol{\beta}$, when near-collinearity is present; see [4, Chapter 4]. The situation is not so critical for prediction squared error loss at the design points, where ill-estimated directions are down-weighted.

Estimators of k are often motivated by **Bayesian** roots. If a priori $\beta_i, i = 1, \ldots, p$, are independent normal with mean zero and variance τ^2 and a vague prior is taken for β_0, then, with normality of the errors in the model in (1), (2) is the Bayes estimate of $\boldsymbol{\beta}$ (mean of the posterior distribution), where $k = \sigma^2/\tau^2$. Such estimators range from simple **empirical Bayes** plug-in values for the hyperparameter τ^2 as in [10], to maximum **marginal likelihood** [1].

Ridge regression may be viewed as a continuous version of **variable selection**, where unimportant variables have their coefficients shrunken towards zero. **Shrinkage** to zero is evident from both the Bayesian roots and the form of the the ridge estimator. In canonical form (2) becomes $[\lambda_i/(\lambda_i + k)]\hat{\alpha}_i$, where $\boldsymbol{\theta}$ is the orthogonally transformed $\boldsymbol{\beta}$, $\hat{\boldsymbol{\theta}}$ its least squares estimate, and $\lambda_1, \ldots, \lambda_p$ the eigenvalues of $\mathbf{X}'\mathbf{X}$ (*see* **Eigenvalue**). Thus, the shrinkage is more pronounced for directions of small eigenvalues of $\mathbf{X}'\mathbf{X}$, corresponding to a large variance of the least squares estimator. Shrinkage to zero is also evident from the augmented form given in [12]. This form is also useful for computation. If the n-vector $\mathbf{Y}$ is augmented by p zeros, to form an $(n + p) \times 1$ vector $\mathbf{Y}^*$, and $\mathbf{X}^*$ is the $(n + p) \times p$ matrix $\mathbf{X}$ augmented by $\sqrt{k}$ times a $p \times p$ identity matrix, then least squares applied to $\mathbf{X}^*$ and $\mathbf{Y}^*$ gives the ridge regression estimator, (2).

For any $k > 0$, the estimator in (2) is not invariant to changes in scale. The Bayesian roots also attest

to the importance of the relative scales of the p explanatory variables, since changing from, say, days to weeks will increase x by a factor of 7 and consequently decrease the corresponding β-coefficient by a factor of 7. The required **exchangeability** of coefficients underpinning ridge regression will not necessarily match the common approach of standardization of explanatory variables to all have variance 1 (autoscaling). This would also imply that it is not necessarily sensible to rescale the explanatory variables each time an observation or observations are left out in **cross-validation** or **bootstrap** approaches to estimation of k.

For reviews of ridge regression and other shrinkage estimators in the context of multiple regression see [7] and [4, Chapter 4]. The simple ridge form of regularization of statistical problems involving many parameters has become commonplace. In the **neural network** literature it comes in the form of a quadratic penalization, a particular example of "weight decay"; see, for example, [13]. The Bayesian form of the technique may motivate adaptations to nonnormal models, as in [2]. Le Cessie & Van Houwelingen [11] adapt ridge to **logistic regression** and apply it to the diagnosis of ovarian cancer using the groups in a histogram of DNA content as explanatory variables. Multivariate forms of ridge regression have been developed by Brown & Zidek [5, 6] (*see* **Multivariate Multiple Regression**).

References

[1] Anderssen, R.S. & Bloomfield, P. (1974). A time series approach to numerical differentiation, *Technometrics* **16**, 69–75.

[2] Askin, R.G. & Montgomery, D.C. (1980). Augmented robust estimators, *Technometrics* **22**, 333–341.

[3] Brown, P.J. (1977). Centering and scaling in ridge regression, *Technometrics* **19**, 35–36.

[4] Brown, P.J. (1993). *Measurement, Regression, and Calibration*. Clarendon Press, Oxford.

[5] Brown, P.J. & Zidek, J.V. (1980). Adaptive multivariate ridge regression, *Annals of Statistics* **8**, 64–74.

[6] Brown, P.J. & Zidek, J.V. (1982). Multivariate regression shrinkage estimators with unknown covariance matrix, *Scandinavian Journal of Statistics* **9**, 209–215.

[7] Frank, I.E. & Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics* **35**, 109–147.

[8] Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: applications to nonorthogonal problems, *Technometrics* **12**, 69–82.

[9] Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* **12**, 55–67.

[10] Hoerl, A.E., Kennard, R.W. & Baldwin, K.F. (1975). Ridge regression: some simulations, *Communications in Statistics – Theory and Methods* **4**, 105–123.

[11] Le Cessi, S. & Van Houwelingen, J.C. (1992). Ridge estimators in logistic regression, *Applied Statistics* **41**, 191–201.

[12] Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, *Technometrics* **12**, 591–612.

[13] Ripley, B.D. (1994). Neural networks and related methods for classification (with discussion), *Journal of the Royal Statistical Society, Series B* **56**, 409–456.

[14] Thisted, R.A. (1988). *Elements of Statistical Computing*. Chapman & Hall, London.

P.J. BROWN

Risk Adjustment

Some cases are more difficult than others. For example, an 80-year-old hospital patient with a heart attack is more likely to die than one aged 50, while almost anyone admitted for cataract surgery is substantially less likely to die than either of these two. Clearly, when comparing hospitals on patient outcomes (such as mortality) or system outcomes (such as how much of what services are used), comparisons should be adjusted for intrinsic differences in patient risk (see **Case Mix**).

When calculating and adjusting for "risk", it is important to stay focused on "risk of what". For example, patients at a moderate risk of death from coronary artery disease may have the highest risk for receiving coronary artery bypass graft (CABG) surgery, since those who are less sick may not need this aggressive therapy, while the very sickest patients may be deemed too frail to survive it.

Raw comparisons of outcomes can be misleading. For example, in the mid-1980s, reporters used the Freedom of Information Act to force the Health Care Financing Administration (HCFA, called CMS) to release its data on hospital mortality for Medicare patients. Ominously, the "worst" facility had 87.6% of its Medicare patients die; however, this rate seemed less strange when the facility was revealed to be a hospice program. Over the years, HCFA made great strides in the sophistication of the risk-adjustment methods used to produce its annual hospital mortality reports. However, the HCFA administrator's stated reason for abandoning these annual reports in 1993 was that the methodology appeared to unfairly penalize inner city public facilities, whose patients may well be at greater risk of dying than is captured in the data available for modeling.

Thus, differences in patient risk often do not "average out" and failure to adequately adjust for differences between the patients seen by different health care providers can produce seriously misleading comparisons. Much research supports this concern, and supports the use of risk adjustment (*see* **Adverse Selection**).

Thus, for example, instead of reporting an observed rate, O, of mortality following CABG surgery, for a surgeon or a hospital, E, an expected rate based on patient characteristics, is also computed, and some measure of the discrepancy is reported (e.g. the risk difference, O minus E, or the risk ratio, O divided by E). Typically, the expected rate is produced using standard **multivariable modeling** techniques applied to large databases, and all caveats for the care required in developing and using such models apply.

One way to produce a "risk-adjusted" rate for a particular facility is to multiply its risk ratio by a broadly defined average rate. For example, if nationally, 2% of nursing home patients develop bedsores in a six-month period, a facility with O/E equal to 1.10 would have its risk-adjusted problem development rate equal to 2.2%. If either the number of cases seen or the number of expected events at a facility is small, the risk ratio calculation becomes unstable; **empirical Bayes** modeling is one way to produce more stable estimates. For further information, see [1].

Reference

[1] Iezzoni, L.I., ed. (1994). *Risk Adjustment for Measuring Health Care Outcomes*, 3rd Ed., 2003 Health Administration Press, Ann Arbor, Chapters 1 and 12–17.

ARLENE S. ASH

Risk Assessment for Environmental Chemicals

In 1983, the National Academy of Sciences (NAS) published a seminal report, *Risk Assessment in the Federal Government: Managing the Process* [8], on the theory and practice of human health risk assessment for chemicals in the environment.

Even though dictionaries define risk in terms of the *probability* of injury, damage, or loss [19], and even though other professions have long since adopted probabilistic frameworks using the **Monte Carlo method** developed in 1946, most human health risk assessments for chemicals in the environment still use deterministic methods that employ point values for all variables. Probabilistic methods have three key advantages over the deterministic ones that they replace. First, probabilistic methods use all the information available about the variability and the uncertainty inherent in the assessment, while deterministic assessments discard most of the information. Secondly, by using probability distributions to represent the range of exposure and/or toxicity, probabilistic methods reveal the compounded conservatisms inherent in deterministic methods. Thirdly, probabilistic methods – relying as they do on the full range of values that a variable may assume – reestablish the now blurred boundary between **risk assessment** and risk management.

We note that risk assessment and *epidemiology* have some goals in common, but that risk assessment differs from epidemiology by its central focus on the prediction of future events.

Risk Assessment vs. Risk Management

The NAS report defined two roles for individuals and stressed the need to keep these roles and associated activities well separated from each other:

1. A *risk assessor* is an analyst – perhaps an engineer or scientist – who uses facts and quantitative reasoning to estimate or bound the exposures and health effects, if any, to a person exposed to chemicals in the environment. The risk assessor may also analyze different technical options for remediation of the property.
2. A *risk manager* is a different person – perhaps a legislator, judge, member of a jury, a regulator, or the public itself – who then weighs the health risks in light of other social, political, and economic factors. The risk manager(s) then decide(s) the actions necessary or appropriate for a given situation. The actions may range from continuing the "do nothing" or "no action" alternative, to restrictions on land use, to complete excavation and removal. Risk management is the process of weighing policy alternatives and selecting the most appropriate regulatory action by integrating the results of risk assessment with engineering data and with social, economic, and political concerns to reach a decision.

Considering a contaminated property as an example, a risk assessor usually (i) performs a "baseline risk assessment", i.e. she or he analyzes the health risks to people who use the property under both the current and reasonable foreseeable use for reducing the risk and then (ii) estimates cleanup targets, as appropriate. A risk assessor may also complete a "verification risk assessment" *ex post* remediation. A risk assessor may perform the same general types of studies to assess or compare the effects of pesticide residues in foods, the operation of a proposed incinerator, new methods to disinfect public water supplies, the reliability of a manufacturing plant, or even the transportation of hazardous materials.

Risk Assessment: The Five-Step Process

Risk assessors usually follow a five-step process when completing a risk assessment. The typical five-step process used today is shown in Figure 1.

In what follows we use the assessment of a contaminated property as an example.

Hazard Identification

In this first step, the risk assessor reviews information about the property, such as location, land use, and abutting land use, and plans for future development. For a hazardous waste site, the risk assessor would review and analyze all of the monitoring data, including, for example, (i) chemical measurements in soils, ground water, sediments, surface water and/or air, and (ii) physical measurements, such as wind speed,

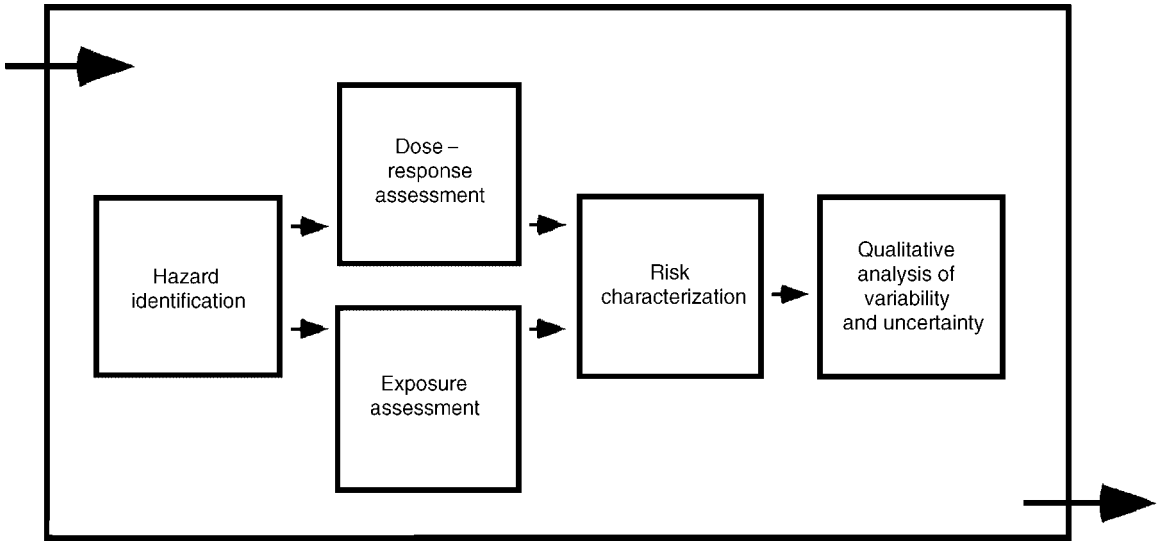

Figure 1 The five-step process for a risk assessment

temperature, hydraulic gradients, and turbulence. In the hazard identification step, the risk assessor must define the nature and extent of the problem, especially the lateral and vertical extent of the contamination. The resulting "study area" need not conform to property boundaries of ownership or even to state lines.

In hazard identification, the risk assessor must also select a list of "study chemicals" or "chemicals of concern", a subset of organic and/or inorganic compounds reported at the site that entails the highest exposure and the highest risk. The risk assessor usually selects the study chemicals for consideration according to these (sometimes competing) factors:

1. high average and/or maximum concentrations;
2. long persistence in the environment;
3. high toxicity, especially carcinogenicity, teratogenicity, or reproductive toxicity;
4. high frequency of detection;
5. great mobility in the environment; and/or
6. high public concern or awareness.

The risk assessor may also consider other factors as well, including whether the chemical is related to human activities or naturally occurring, whether the chemical is reported in concentrations above either natural or anthropogenic background concentrations, and whether the chemical is an essential nutrient for plants or animals.

Dose–Response Assessment

The risk assessor assembles information on the acute, subchronic, and chronic toxicities of the study chemicals selected in the previous step. Toxicologists distinguish between (i) potential carcinogens, which are chemicals that may initiate or promote the development of cancer, and (ii) noncarcinogens, which are chemicals that cause damage other than cancer to cells, tissues, organs, or organ systems in the exposed individual (*see* **Dose–Response Models in Risk Analysis**). More often now, toxicologists also identify neurotoxins, mutagens (agents that may cause somatic or genetic mutations; *see* **Mutagenicity Study**), and teratogens (agents that may cause birth defects in newborns; *see* **Teratology**). Of course, some chemicals may cause multiple effects at different times and doses. For example, in high doses, a particular dioxin called 2,3,7,8-TCDD may cause immediate tissue damage and a skin disease (effects of acute exposure); in lower doses, and over time, it may increase the probability that a few different types of cancer will develop (an effect of low-dose chronic exposure). (For general references, see [7] and [9].)

Developing Reference Doses (RfDs). Studies have shown that many people incorrectly believe "No dose of a chemical is safe". In fact, for any chemical,

there is a dose low enough that it does not cause significant clinical effects and a dose high enough that it probably does. We safely eat small amounts of cyanide in almonds, accidentally swallow apple seeds, and other foods; on the other hand, ingestion of less than a half pound of salt all at once will likely kill an adult. Mammals (including humans) have evolved to eat safely a multitude of naturally occurring toxic chemicals in foods, and, as a result, mammals have excellent defenses against many types of chemicals (including many industrial chemicals) and can readily detoxify and excrete them.

Identifying a dose of a chemical that is unlikely to cause effects is challenging. Experience has taught us that humans do not respond the same way as any particular experimental animal species. Sometimes humans are more sensitive to a certain chemical than rats are but less sensitive than, say, guinea pigs. Different animals (including humans) sometimes even respond to a chemical or drug exposure with different types of responses altogether. For example, some tranquilizer drugs used with dogs and horses will make cats more excited instead of depressing them. And humans themselves vary from person to person in their response: one person can drink alcohol seemingly all evening, while another is quite drunk after one glass of wine. Given a new chemical, a toxicologist cannot predict whether humans are more or less sensitive than the test animal, or whether they will respond with the same toxic effect, or how much variability different people will exhibit in their response. If this were not trouble enough, most animal toxicity tests last for six months to two years, while a human exposure might occur over 30–40 years or more.

The solution to these questions has been to apply limits on the uncertainty. Experience over many years with thousands of chemicals has allowed toxicologists to develop rules for estimating a "safe" dose. The US Environmental Protection Agency (EPA) has developed a standardized method for developing "reference doses" (RfDs) for several hundred of the solvents, pesticides, metals, and other chemicals most commonly encountered in the environment. The method does not identify the highest dose of a chemical that is safe, but it does identify a dose unlikely to cause effects; that is, the highest safe dose is probably not lower than the RfD.

The current EPA method has three main steps (although changes are proposed for the future). First, test animal studies and human observations (if available) are reviewed for study quality and to find the lowest dose at which adverse effects were observed (the "LOAEL" or lowest adverse effect level), and the next lower dose at which these effects are not seen (the "NOAEL," or no observed adverse effect level). Secondly, depending on the quality and findings of the experimental data, several uncertainty factors are applied to the LOAEL (or the NOAEL, if it is available) to estimate the RfD. If good dose information is available from accidental human exposures (as for mercury) one can use a NOAEL directly as an RfD, or one can divide it by a factor of 1–10 to account for variations among humans. For most chemicals we must use animal studies. If we have available only a LOAEL from an animal study, we divide by a factor of 10 to estimate a NOAEL. The factor of 10 accounts for potential differences between humans and experimental animals. The result may be divided by a factor between one and 10 if the experiment lasted less than the animals' lifetime (rats and mice live about two years) to estimate a safe dose for a human lifetime. We can use an additional modifying factor if we have reason for additional concern, such as that the only effects observed in the test are very serious ones. Sometimes we make other modifications in the dose estimate to represent more closely the dose that we expect that humans will receive. This process has been used both for ingested and inhaled doses of chemicals. There is no comparable approach developed for dermal exposure to chemicals; generally we use the oral RfD to evaluate a dermally absorbed dose, in spite of the obvious uncertainties.

The final step in the derivation of an RfD is to identify qualitatively the scientific confidence in it. Good, plentiful, studies with consistent findings result in high confidence, while if the available studies are poor or few, confidence is low. Note that low confidence usually results in a lower RfD because the greater uncertainty is reflected in higher uncertainty factors. So the RfD is no less "safe" if confidence is low; in fact, the true highest safe dose may be 1000 or more times higher.

Most scientists recognize that this approach is not optimal; it may result in overprotection at unnecessary expense. The current approach produces some odd results, such as an RfD for phenol that is below the dose received when one uses a common over-the-counter sore throat medication as directed (phenol is its active ingredient). The EPA and other groups

are working on more effective approaches, using **pharmacokinetics,** modeling, sophisticated statistics, quantitative analysis of variability and uncertainty, and other methods. One such approach is used for lead and is being developed for other metals such as arsenic and mercury. The approach is to use a model to evaluate several sources of exposure to lead and to predict the expected values in a population of a particular biomarker, in this case blood concentrations. (This approach has had only limited success so far.) Meanwhile, risk managers need to be mindful of how RfDs are developed and understand that if a risk assessment shows that a certain chemical concentration "exceeds a level of concern", that this does not mean we shall see toxic effects in exposed people. For most chemicals, there is a large margin of safety built into the toxicity value used in the risk calculation. Also, the target concentrations of concern are so low that the effects cannot be detected in small or medium sized populations.

Developing Cancer Slope Factors (CSFs). Toxicologists agree that there is a safe dose with respect to toxic effects other than cancer. The EPA has assumed that there is not a safe dose for carcinogens; there is, however, a dose that poses an acceptably low risk. Cancer is considered a dose-related example of a **stochastic process**; that is, there is always a chance of acquiring cancer from exposure to a carcinogen, no matter how small the dose; however, the smaller the dose, the smaller the chance. For some cancer causing agents, such as **radiation**, this appears to be true. Cancer (really a group of diseases) differs fundamentally from other toxic effects such as liver damage, in which a certain threshold of damage is necessary before the damage makes any practical difference, since the body can live with or replace many damaged cells. It is thought that cancer can result from a single mutation in DNA. This mutation can theoretically be caused by a single molecule of a carcinogenic chemical [1]. Therefore, there is theoretically always a possibility that a single molecule could cause the damage that leads to cancer. In reality, this is probably not true for most chemicals, but the EPA currently regulates carcinogens on this basis.

There are only about 30 known chemicals or industrial processes that cause cancer in humans. All other chemicals considered carcinogenic are only suspected to cause cancer in humans, perhaps because of weak evidence from human studies, or because the chemical tested positive in an animal test system. Groups of test animals (usually rats, mice, or dogs) receive the highest dose or half the highest dose of the test chemical that one expects they can tolerate for their lifetime without significant toxic effects other than cancer, and at the end of that time (about two years for rats and mice) the test animals are necropsied and examined for tumors. If the test groups have more tumors than the control group, then one can employ tumor data in a model to predict possible cancer rates in humans exposed over a much longer lifetime to much lower doses. The model most commonly used by the EPA produces a "cancer slope factor" (CSF) that one can use to estimate cancer risk in exposed humans. The CSF assumes a linear relationship between exposure dose and carcinogenic response. For example, with a person exposed to 0.02 mg alachlor per kg body weight every day for a lifetime and where the CSF is 0.08 mg/(kg d), an estimate of the cancer risk is $(0.02 \times 0.08) = 0.0016 = 1.6$ in 1000.

Often, results from the animal studies fail to show a clear effect. In any case, toxicologists must evaluate all available evidence for a chemical, including studies among different animal species, and epidemiologic studies in humans, to determine whether the weight of evidence indicates whether a chemical may be a human carcinogen. Under the pre-1996 guidelines [10], the EPA classified carcinogens according to the weight of evidence into one of five groups that ranged from "Known human carcinogen" to "No evidence of carcinogenicity". All but perhaps 35–40 chemicals ever studied fall somewhere between these two categories; their carcinogenicity in humans is uncertain. Guidelines published for review in 1996 replace the five groups with three categories and a narrative to describe the weight of evidence [18].

Much of this uncertainty concerning carcinogenicity comes from the process of identifying and evaluating potential carcinogens. Toxicologists know that there are serious questions about the validity of animal carcinogenicity studies. Some scientists believe cancer in test animals is actually a byproduct of the cell damage and subsequent cell reproduction and tissue repair induced in the experimental animals due to the toxic dose levels typically used in the tests. With humans exposed to extremely low doses of the same chemical in the environment, the cell damage

is too minor to induce tissue repair, and the chemical probably does not cause cancer. On this basis, many people believe that the true risk from such low exposures may be as low as zero. A second argument applies to chemicals that only cause tumors in specific situations. For example, many chemicals appear to cause tumors only in mice – not in other experimental animals – and only in the livers of those mice, rather than in other organs. Chlordane is such a chemical. We would not expect these chemicals to cause cancer in humans: experts disagree, however, so the cancer risk is uncertain. The same issues discussed for noncancer reference doses apply here, as well; toxicologists must extrapolate from high doses to low, short-term experiments to long-term environmental exposure, and animals to humans. Finally, uncertainty about the models used to develop the CSF is large. While the true cancer slope factor is not likely higher than the published CSF, it may be much lower, and perhaps as low as zero (*see* **Extrapolation, Low Dose**).

Sources of Chemical Toxicity Information. The EPA has already established and published the physical, chemical, and toxicological properties of many chemicals found at hazardous waste sites. These evaluations, based on the results from studies conducted by the National Cancer Institute (NCI) and from articles in refereed journals, are often reduced to either "chemical profiles" or simply a handful of numbers that represent toxic potencies. Published by the EPA, the findings and opinions on these chemicals are usually listed in two widely available resources: (i) the Integrated Risk Information System (known as IRIS [17]), a database updated monthly and available over several wide-area computer networks and (ii) the Health Effects Assessment Summary Tables (known as HEAST [16]), a database updated quarterly or semiannually and available in print from the National Technical Information Service (NTIS) in Springfield, Virginia. These two databases typically give several toxicity values for a single chemical, depending whether the exposure occurs via ingestion (e.g. via food or water) or via inhalation (e.g. via gases or particulates). Neither of these databases currently includes toxicity values for dermal exposures.

For use in a numerical example later in this article, we note that EPA's IRIS database recently listed the CSF for the ingestion of benzene as 2.9×10^{-2} mg/(kg d).

Exposure Assessment

The risk assessor first determines if complete "exposure pathways" exist and then estimates the doses delivered along those pathways [11].

Exposure Pathways. An exposure pathway is any route that a chemical may travel from an environmental source (e.g. an abandoned dry sludge lagoon) to a receptor (also called the exposed individual), such as a child living nearby. An exposure pathway has five main parts:

1. a chemical source;
2. a release mechanism (e.g. leaking, leaching, wind erosion);
3. a transport and/or exposure medium (e.g. air, water, soil, sediment, food);
4. an exposure point with receptors present or potentially present (actual location where exposure is possible); and
5. a route of entry (inhalation, ingestion, dermal contact).

A complete exposure pathway is one that has no functional barrier that prevents an exposure. The pathway may be completed (i) by the chemical moving from the source to the receptor or (ii) by the receptor moving to the source. In this example, fugitive dust that carries the chemicals may blow from the lagoon to the child's house, or the child may play in or near the old lagoon. Either way, with a completed pathway, the receptor comes into contact with some of the chemical from the source. If no exposure pathway is complete, there is no exposure and subsequently no risk.

In exposure assessment, the risk assessor usually considers three different exposure routes into the body: inhalation, ingestion, and dermal contact. With ingestion, the risk assessor considers whether a person may deliberately or inadvertently swallow some liquid or solid, including food, beverages, or soils (say, by hand-to-mouth movement). With inhalation, the risk assessor considers whether a person breathes toxic materials as gases, vapors, aerosols, or particles. With dermal contact, a risk assessor considers whether a person may touch gases, liquids, or solids that contain toxic materials and possibly absorb the chemical through the skin. Of course, in some situations, all three exposure routes may convey meaningful amounts of a chemical.

Exposure pathways are often categorized as "direct" or "indirect" pathways. Although no strict definitions exist, a direct pathway exists when the exposed person experiences the chemical in the same medium as that in which it is present in the source. In our example, the child may inadvertently ingest contaminated soil near the abandoned lagoon and thus experience a direct pathway. Alternately, an indirect pathway exists when the exposed person experiences the contaminants in a different medium, often at a distance from the source. In a different example, the child may drink cow's milk containing dioxin that traveled from an incinerator via this complicated route: formation in the incinerator, emission into the atmosphere as a gas, adsorption on to a particle in the atmosphere, wet or dry deposition on to grass in a pasture, ingestion by the cow, secretion in the cow's milk, and ingestion by the child drinking milk. For some types of facilities – for example, some incinerators – risk assessments show that these indirect pathways, although challenging and difficult to measure or analyze, may cause greater exposures than do direct exposures for some chemicals. In some areas of the US and in many other countries, homegrown vegetables may be an important part of the diet, and transfer of chemicals from soil, water, or air to vegetables may be highly important. In other cases, high exposure estimates for indirect pathways may result from compounding many conservative assumptions.

Estimation of Dose. When reading a risk assessment or a research publication, one should always check the definition of dose used by the authors, because different concepts and different units of measurement are common in the literature [1].

Here we distinguish three primary concepts of dose, noting that the first is most commonly used in risk assessments at hazardous waste sites.

1. *Exposure dose* is the mass of chemical that enters a person's body via ingestion, inhalation, or dermal contact. No allowance is made for excretion or exhalation of the chemical before absorption or metabolism. The conventional measure of exposure dose in units of milligrams of chemical per kilogram of body weight per day, mg/(kg d), explicitly scales by body weight, because a larger person needs greater exposure than does a smaller person to have a comparable effect. A milligram of chemical theoretically causes more harm to a child than to an adult.
2. *Absorbed dose* is the mass of chemical absorbed or metabolized by the receptor's body. Although measured in the same units, mg/(kg d), absorbed dose is always smaller than exposure dose, because some of the chemical is excreted or exhaled from the body before absorption in the lungs or gastrointestinal tract. On scientific merit, absorbed dose is always preferable to exposure dose, but it is also more difficult to measure or estimate because the (relative) absorption of the chemical may depend on many factors, including the age, health, and health status of the exposed person. With some chemicals, such as lead and other metals found in soils, identification of the absorbed dose may be very important in estimating risk. It may also be used in extrapolating from one exposure route to another, such as from an ingestion to a dermal dose.
3. The third concept of dose – *biologically effective dose* (BED) – is rarely used today directly in practical risk assessments for hazardous waste sites. BED is the mass of chemical (or sometimes the concentration of chemical) that reaches the target organ or tissue and causes the physiological or genetic damage. Of central importance in laboratory studies, BED is rarely used in practical risk assessments, because it is so difficult to determine even the identity of the chemical or the metabolite that causes the damage at the molecular level in the body. Usually, laboratory scientists study BED using radio-labeled compounds or other chemical measurement and pharmacokinetic models of absorption, metabolism, and excretion. This type of information can be helpful in extrapolating from animal studies to predicted human effects, or from one exposure situation to another. The toxicity values used in risk assessments may be refined as a result of such studies.

The duration of exposure clearly plays a central role in toxicology. Sometimes a brief, relatively large dose may cause less damage to an organism than does a much lower total dose sustained over a longer period of time, and sometimes the opposite is true. While occupational hygienists, police, or fire officials focus on exposures that occur over minutes or hours, risk assessors usually focus on exposures that range

in duration from a year or so to a full lifetime (usually assumed to be 70 years). Thus, risk assessors rarely consider acute health effects (from exposures of a few seconds to a few weeks) and instead tend to focus on subchronic, chronic, or lifetime exposures (here taken to mean, respectively, a few months, a few to many years, or a full lifetime). The exception to this is if very high exposures over a short time are possible, such as during site remediation. When reading an article or report, it is important to understand what time frames are included and what are excluded from the analyses.

We estimate exposure dose from the measurements or models of the exposure point concentration (e.g. the concentration in soil in a residential yard) and the contact rate (e.g. the amount of soil ingested from the yard). The exposure point concentration is the steady or time-varying concentration of the chemical in the medium to which a person has been exposed. In a particular situation, a single exposed person may have several exposures to a single compound. For example, at work, a person may breathe air that contains a chlorinated solvent, while at home the person may ingest water that contains the same compound. If the contaminant comes from a single source, the analysis is usually much easier than if the compound comes from multiple sources. A risk assessor may rely upon measurements or models to estimate the exposure point concentration. In most instances, concentration measurements are more reliable than modeled concentrations. However, models are frequently used if there are no cost-effective or realistic ways to measure the exposure point concentrations directly. Also, models are often used to predict the fate and transport of chemicals in air or ground water.

When using a model to estimate the exposure point concentrations of a chemical, especially a multimedia model, a risk assessor should remember the words of George Box, "All models are wrong but some are useful" [2]. Although originally penned for another purpose, these words suggest the need for great caution in using models to estimate exposure point concentrations and movement of chemicals in the environment. Before one can rely upon any concentrations modeled by oneself or others, we urge that one attempts to understand all of the limitations of the model(s) and to make a reality check for the values before proceeding.

When estimating the contact rate, a risk assessor is really estimating 'in the order of' intensity, frequency, and duration of exposure. A typical adult may drink daily $\sim 1-3$ liters of water, some at home and some at work. Those exposures may last five, 20, or more years, depending on changes in residence, employment, or occupation. The same person may have other exposures also via ingestion, inhalation, or dermal contact. Some exposures happen intermittently in time or space; for example, recreational use of a park.

The EPA has published numerous guidance manuals on the selection of the exposure factors needed to estimate contact rate [15]. For example, the Agency's *Exposure Factors Handbook* [12] lists point values for many physiological or behavioral variables for children and adults. The Agency's *Exposure Factors Handbook* does contain many of the standard values widely quoted and mandated in risk assessments. According to the EPA, each adult is assumed to weigh 70 kg, to ingest 2 l/d of drinking water, to breath $20-24\,m^3$/d of air, and to live in the same residence for 30 years. The Agency chose some of the values as "conservative" or upper bound values (e.g. 2 l/d of drinking water for each adult) and chose others as typical or average values (e.g. 70 kg as the average weight of an adult).

While these numbers are simple to memorize and easy to apply, it is important to realize that they misrepresent conditions that most people experience. People vary in many attributes. Not everyone weighs the same amount, or has the same diet, or lives in a home as long as 30 years. Furthermore, the actual values for some assumptions such as soil ingestion rates are simply unknown. Each of these exposure factors is better represented by a range or distribution of values – a topic discussed later.

As a practical matter, we recommend that risk assessors first estimate the exposure dose to a person on a day during which exposure is known to occur. For example, if a child has exposure to a toxic chemical when trespassing on an industrial property, the risk assessor should first estimate the dose on that particular day of trespass. If the behavior occurs every day of a year, then the average daily dose on a day of exposure equals the average daily dose during that year of exposure. However, if the exposure does not happen every day, then the average daily dose for the year is less than the average daily dose on the day of exposure. All doses need evaluation against

an appropriate measure of toxicity. For example, an adult could drink one Martini every night for two weeks without serious adverse effect, but drinking 14 Martinis in one night would have dangerous health effects, even though the average daily dose over two weeks is the same. The risk assessor must account for high-dose, short-term exposure potential, even during one year or less.

As a numerical example, we estimate the average daily dose (*ADD*) averaged over a lifetime of exposure for an adult who (unwittingly) drank water from a contaminated well at a vacation home. To make the calculation, we estimate the exposure dose to this person who weighed 70 kg, drank 2 l/d of water, and visited the vacation home two days per week (the weekend) for 10 weeks per year (the summer). This person owned the house for 20 years. The well water contained 115 μg/l of benzene (a known human carcinogen). We use this formula:

$$\langle ADD \rangle_{\text{life}} = \frac{Conc \times IngR \times CF}{BW} \frac{D}{7} \frac{W}{52} \frac{Y}{70},$$

where $\langle ADD \rangle_{\text{life}}$ is the average daily dose, averaged over a lifetime (mg/(kg d)), $Conc$ is the concentration in drinking water (μg/l), $IngR$ is the ingestion rate (l/d), CF is the conversion factor (mg/μg), BW is the body weight (kg) D is the number of days of exposure per week, W is the number of weeks of exposure per year, and Y is the number of years of exposure in a lifetime of 70 years.

Substituting the values with $CF = 10^{-3}$, we find $\langle ADD \rangle_{\text{life}} \sim 5.16 \times 10^{-5}$ mg/(kg d).

While doses of carcinogens may be averaged over a lifetime, doses of noncarcinogens are averaged over a shorter time (usually the duration of exposure). In the example above, it is appropriate to average exposure to a noncarcinogen over 20 years.

Risk Characterization

The risk assessor combines all the information gathered in the preceding three steps to estimate quantitatively the health risk. In practice, this step usually culminates (i) in a numerical estimate of noncarcinogenic health effects as measured by a summary statistic called the total *hazard index* (*HI*) and also (ii) in a numerical estimate of the carcinogenic potential as measured by a summary statistic called the total *incremental lifetime cancer risk* (*ILCR*) (see, for example, [13]).

For exposure to a single noncarcinogenic chemical via a single exposure pathway, the hazard index is usually defined as the average daily dose averaged over one year of exposure divided by the reference dose for that chemical via that exposure pathway.

$$HQ_{ij} = \frac{\langle ADD \rangle_{\text{year}}}{RfD},$$

where HQ_{ij} denotes the *hazard quotient* for that combination of chemical and, exposure pathway (as a ratio, it has no units); $\langle ADD \rangle_{\text{year}}$ denotes the average daily dose averaged over one year, expressed in mg of chemical per kg of body weight per day; and RfD denotes the reference dose for that chemical and route of exposure (e.g. inhalation), also expressed in mg of chemical per kg of body weight per day.

In a full study, the risk assessor estimates the hazard index (HI) by summing the hazard quotients over all chemicals and exposure pathways [14]:

$$HI = \sum_i \sum_j HQ_{ij}.$$

Given its definition as a ratio of positive numbers, the HI may range from zero to infinity. If the $HI > 1$, then the risk assessor may disaggregate it into those components that act on a common organ system or by a single molecular mechanism. The reasoning behind this is that the body can handle multiple chemical stresses through multiple defenses. For example, at a mining site, people may be exposed to lead, zinc, arsenic, manganese, and copper. For practical purposes, at low doses, these metals act independently on different organ systems, and regulatory agencies often treat the risks from these metals separately instead of adding them together. The risk manager may become increasingly concerned as the HI disaggregated by organ system or molecular mechanism exceeds unity.

For exposure to a single carcinogenic chemical via a single exposure pathway, the incremental lifetime cancer risk is usually defined as the average daily dose averaged over a lifetime multiplied by the cancer slope factor [10]:

$$ILCR_{ij} = \langle ADD \rangle_{\text{life}} \times CSF,$$

where $ILCR_{ij}$ denotes the incremental lifetime cancer risk for that combination of chemical and exposure pathway (as a probability, it has no units);

$\langle ADD \rangle_{\text{life}}$ denotes the average daily dose averaged over a full life, usually taken as 70 years, expressed in mg of chemical per kg of body weight per day; and CSF denotes the cancer slope factor for that chemical and route of exposure, expressed in the inverse of (mg of chemical per kg of body weight per day).

Continuing the numerical example from above, the adult who drank the water contaminated with benzene ($CSF = 2.9 \times 10^{-2}$ mg/(kg d)) while staying at a vacation home has an estimated $ILCR \sim 1.5 \times 10^{-6}$.

In a full study, the risk assessor estimates the total incremental lifetime cancer risk by summing over all chemicals and exposure pathways [13]:

$$\text{Total } ILCR = \sum_i \sum_j ILCR_{ij}.$$

Before estimating the total incremental lifetime cancer risk, the risk assessor may estimate subtotals by chemical source and by route of exposure to help clarify which exposure pathways cause the greatest risk. Of course, for a contaminated site, the risk assessor must take care to add risks from exposures related to the site.

Because the incremental lifetime cancer risk is formulated and interpreted as a probability of developing cancer at some time during a lifetime, the value (in theory) ranges from zero to one. As a practical matter, the probability of developing cancer is not linear at high doses (as the CSF implies). Even very heavy cigarette smokers are not guaranteed to develop lung cancer. Probabilities that exceed 1 in 100 are usually highly inaccurate.

Analysis of Variability and Uncertainty

The risk assessor names the sources, magnitudes, and likely effects of variability and the uncertainty in the analysis. We define variability and uncertainty as follows:

1. *Variability* represents diversity or heterogeneity in a well characterized population of plants, animals, or people. Fundamentally a property of nature, variability is usually not reducible through further measurement or study. For example, different adults drink different volumes of tap water each day, no matter how carefully or how often we measure their diets.
2. *Uncertainty* represents partial ignorance or lack of perfect information about poorly characterized phenomena or models. Fundamentally a property of the risk analyst, uncertainty is sometimes reducible through further measurement or study. For example, a risk assessor may not now know how much soil each adult ingests per day, but she or he may be able to design experiments to gain additional (but still imperfect) information.

Few risk assessments contain more than a few paragraphs of text acknowledging the variabilities and uncertainties in the methods and results. Some risk assessments go further and include **sensitivity analysis**. For example, a risk assessment may include several calculations of the same result, but each one predicated on a different set of input values chosen within the range of the variability or the uncertainty inherent in the analysis. A sensitivity analysis might indicate that the risk estimate is sensitive to the rate of fish ingestion, and may suggest that a survey of the affected population might tighten the upper and lower bounds on the risk estimate.

Risk assessments may include information on model uncertainty; in other words, the uncertainty inherent in the mathematical formulation of the models used in exposure assessment, dose–response assessment, or risk characterization.

New Directions in Human Health Risk Assessment

The practice of human health risk assessment for exposure to chemicals in the environment is shifting from a deterministic to a probabilistic paradigm. Two features distinguish the two paradigms. In the probabilistic paradigm we consider (i) all variables as **random variables** instead of point values, and (ii) quantitative analyses of variability and uncertainty now become an integral part of risk characterization.

Most people understand intuitively that exposure variables such as body weight, the daily ingestion rate of drinking water, and the number of days a person visits a park are random variables. Similarly, most people understand intuitively that toxicity values, such as the RfD or the CSF, are also random variables (by exhibiting inter-individual variability).

However, some people are uncomfortable with the logical consequence from these facts; namely, if all of the input variables in a risk assessment are random variables, then the output variable – the estimated incremental lifetime cancer risk – is also a random variable. In other words, the estimated risk for a situation is not a point value but a range of values (see, for example, [6] and [3]).

When establishing the probability distributions for the input variables for a risk assessment, it is instructive to distinguish two driving forces behind the need for probability distributions – variability and uncertainty – as defined earlier. For example, with body weight, the random variable captures mostly the known and well measured inter-individual variability in a population. As a second example, for the number of days a person swims in a local pond, the random variable may capture mostly the unknown or poorly measured inter-individual behavior in a population. So in a risk assessment, the random variables inevitably capture different combinations of variability and uncertainty for each different input variable [4].

For the probabilistic risk assessment paradigm, in any given situation, the risk assessor should focus on choice of (i) accurate input distributions and (ii) accurate exposure, toxicity, and risk characterization formulas. In probabilistic risk assessment, the analyst has to select input distributions based on the facts of the situation.

In the probabilistic paradigm, the risk manager receives more information than in the deterministic paradigm; namely, she or he receives distributions for exposure and risk instead of merely point values. While risk management in the deterministic paradigm consists of comparison of point values for estimated and acceptable risks as a so-called "bright line test", risk management in the probabilistic paradigm consists of comparison of estimated and acceptable distributions of risk.

To continue the numerical example from above, we learn that the fixed values used earlier oversimplified the history. After discussion with the exposed individual and with further field testing, we find that we can better represent some of the variables in the equations as probability distributions than as point values, precisely because variability was an intrinsic part of the person's behavior and also of the aquifer from which the person consumed ground water. With this new information, we find that these probability distributions better describe the situation than do the point values that they replace:

1. The variability in *Conc* (concentration) is well described by a triangular probability distribution with a minimum of 80, a mode of 85, and a maximum of 120, in units of μg/l.
2. The variability in *IngR* (ingestion rate) is well described by a **normal** or Gaussian probability distribution with a **mean** of 1.60 and a **standard deviation** of 0.20, in units of l/d.
3. The variability in *BW* (body weight) is well described by a normal or Gaussian probability distribution with a mean of 70 and a standard deviation of 10 in units of kg.
4. The variability in *D* (days per week) is well described by a **uniform distribution** with a minimum of 1.0 and a maximum of 2.5.
5. The variability of *W* (weeks per year) is well described by a uniform distribution with a minimum of 7 and a maximum of 11.

All the other variables and conversions in the equation have the same point values as before.

With this new information, we use a commercial software package named Crystal Ball® [5] to convolve (an operation analogous to ordinary multiplication) the probability distributions and the point values in the equations for estimating $\langle ADD \rangle_{\text{life}}$ and *ILCR*. The results of the convolution (as done by 5000 repetitions of a Monte Carlo simulation in the software package) are shown in Figure 2. Figure 2 now more fully expresses the variability inherent in the situation as a range of values from $\sim 5.0 \times 10^{-7}$ to $\sim 2.0 \times 10^{-6}$. The point estimate calculated earlier ($\sim 1.5 \times 10^{-6}$) occurs well above

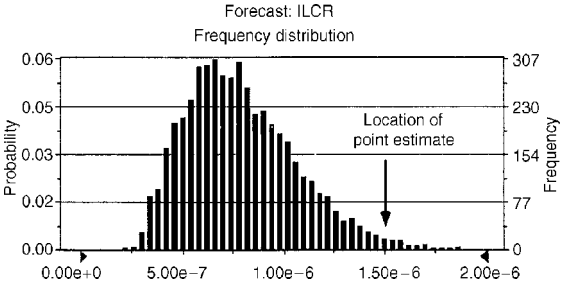

Figure 2 Estimated distribution of risk based on estimated distribution of exposure

the 95th percentile of the estimated distribution. This graph conveys considerably more information than did the point value calculated earlier in this article.

Given an estimated distribution for risk, the risk manager might use decision rules along these lines to render an opinion on the acceptability of the estimated risk: (i) Is the **median** of the risk distribution less than 1 in a million? (ii) Is the average of the distribution less than 1 in 100 000? (iii) Is the 95th percentile of the risk distribution less than 1 in 10 000? If the answer to *all* three questions is "yes", then the risk manager might decide that the risk is acceptable. In other words, the risk manager may only look at selected percentiles or summary statistics when deciding if a risk is acceptable for a population.

The new paradigm does require more effort to specify the input variables and more computation to estimate the distribution of the output variable, namely risk. The probabilistic paradigm still contains unquantified uncertainty. Entire exposure pathways may have been overlooked. Laboratory analyses may have been **biased**. Statistical data analysis may have been inappropriate. Conservative assumptions may have been incorporated in the risk assessment unnoticed. The risk paradigm itself may over- or underestimate risk due to the averaging of exposure, the interactions of chemicals, or other reasons. Overall, the uncertainties for some exposure variables such as soil ingestion rates or toxicity values may have been underestimated in the past due to a focus on specific studies without consideration of fundamental biological principles or other information. One should acknowledge these and other uncertainties to allow for proper interpretation of the distribution of risk. Even considering these remaining uncertainties, the output distribution conveys much more fully the full range and probabilities of plausible health risks – and provides enormous benefit over the simple point estimates that result from a deterministic calculation.

The probabilistic paradigm builds on the fundamental definition of risk as the probability of adverse outcome. It reestablishes the now blurred lines between risk management and risk assessment.

Acknowledgment

We thank Wendy H. Koch, Ph.D., for helpful comments and suggestions. Crystal Ball is a registered trademark of Decisioneering, Inc., Denver, Colorado.

References

[1] Amdur, M.O., Doull, J. & Klaassen, C.D. (1991). *Casarett and Doull's Toxicology*, 4th Ed. Pergamon Press, New York.

[2] Box, G.E.P. (1979). Robustness is the strategy of scientific model building, in *Robustness in Statistics*, R.L. Launer & G.N. Wilkinson, eds. Academic Press, New York.

[3] Burmaster, D.E. & Harris, R.H. (1993). The magnitude of compounding conservatisms in Superfund risk assessments, *Risk Analysis* **13**, 131–134.

[4] Burmaster, D.E. & Wilson, A.M. (1996). An introduction to second-order random variables in human health risk assessment, *Human and Ecological Risk Assessment* **2**, 892–919.

[5] Decisioneering, Inc. (1994). *User's Manual for Crystal Ball*. Denver.

[6] Harris, R.H. & Burmaster, D.E. (1992). Restoring science to Superfund risk assessment, *Toxics Law Reporter*, Bureau of National Affairs, Washington, March 25.

[7] Kamrin, M.A. (1988). *Toxicology*. Lewis, Chelsea, Michigan.

[8] National Academy of Sciences (1983). *Risk Assessment in the Federal Government: Managing the Process*. National Academy Press, Washington.

[9] Ottoboni, M.A. (1984). *The Dose Makes the Poison*. Vincente Books, Berkeley.

[10] US Environmental Protection Agency (1986). Guidelines for Carcinogen Risk Assessment, *51 FR 33992–34003*, September 24.

[11] US Environmental Protection Agency (1988). Superfund Exposure Assessment Manual, *Office of Remedial Response, OSWER Directive 9285.5-1, EPA/540/1-88/001*. April.

[12] US Environmental Protection Agency (1989). Final Report Exposure Factors Handbook, *Office of Health and Environmental Assessment, EPA/600/8-89/043*. May.

[13] US Environmental Protection Agency (1989). Risk Assessment Guidance for Superfund, Vol. I, Human Health Evaluation Manual (Part A), Interim Final, *EPA/540/1-89-002*. December.

[14] US Environmental Protection Agency (1991). Risk Assessment Guidance for Superfund, Vol. I, Human Health Evaluation Manual (Part B, Development of Risk-based Preliminary Remediation Goals), Interim Final, *OSWER 9285.7-01B*. December.

[15] US Environmental Protection Agency (1992). Guidelines for Exposure Assessment, *57 FR 22888 et seq.* May 29.

[16] US Environmental Protection Agency (1993). Health Effects Assessment Summary Tables, Annual FY 1993, *OERR 9200.6-303(91-1)*. March 1993 (or latest update).

[17] US Environmental Protection Agency (1994). US EPA's database documented in: Integrated Risk Information System, Vol. 1, and Electronic Information

System, *Office of Health and Environmental Assessment, EPA/600/8-86/032a*. March.

[18] US Environmental Protection Agency (1996). Guidelines for Carcinogen Risk Assessment, *Office of Research and Development, EPA/600/P-92-003C*. Washington, April.

[19] *Webster's New World Dictionary* (1970). Guaralnik, D.B., (Editor in Chief). World Publishing, New York.

(*See also* **Risk**; **Risk Assessment in Clinical Decision Making**)

DAVID E. BURMASTER & JEANNE C. WILLSON

Risk Assessment in Clinical Decision Making

Risk has been defined technically as "the probability that a particular adverse event occurs during a stated period of time, or results from a particular challenge" [5, 21]. In popular usage, the term is usually broader, and can refer simply to exposures that may cause or lead to adverse events, but without reference to probability. For example, the *Oxford English Dictionary* gives as its primary definition: "hazard, danger; exposure to mischance or peril".

In this article, we will use the technical definition of risk and will focus on risks to patients in the clinical setting, particularly of events incurring physical harm due to therapeutic interventions. There are also possible risks to caregivers, e.g. loss of professional reputation or being sued following a poor decision – or even to a health care system or insurance company responsible for paying the costs of decisions made by caregivers and their patients. Similar general principles can be applied to other types of risk, but the issues discussed in this article will relate directly to clinicians and patients, and to the statisticians who provide them with the estimates of risk.

Factors that Modulate Risk Assessment and Clinical Decisions

Clinical decisions are made because it is believed that the actions that follow them will do more good than harm. **Risk assessment**, therefore, is one important component of such decisions. The assessment of risk (like clinical decisions in general) can be modulated by the circumstances or context in which they are made, the values and preferences of the decision makers, and the information available [7, 12]. While some values are already held by each of the decision makers, their ability to assess risk efficiently in different circumstances (in conjunction with their values) depends on whether they are aware of the relevant information needed to make the decisions; they have access to such information; the information is available to them in a manner that is intellectually accessible, unbiased and appealing; they are able to interpret the information; and they have the skills to incorporate the information into their decisions. Information to guide clinical decisions can come from multiple sources, some of which are more reliable than others (*see* **Utility in Health Studies**; **Decision Analysis in Diagnosis and Treatment Choice**).

Ultimately, clinical decisions should take account of both risks and benefits, and should consider alternative actions (including doing nothing). A fully informed decision, therefore, requires knowledge of both the benefits and the risks of the available options. While the beneficial effects of an intervention may be known from randomized controlled trials (RCTs) (*see* **Clinical Trials, Overview**), however, frequently RCTs are not appropriate or have paid less attention to the harm that may result from the same intervention. For example, RCTs (usually of short duration) may be unable to capture all the relevant information on risks associated with drugs that may have long-term side-effects. On those occasions, decision makers are forced to look at other types of study (such as **cohort** or **case–control studies**) or other sources of information (usually less valid, such as anecdotes or **case series**) to guide their decisions.

Components of Risk Assessment

Several aspects make up risk assessment. These aspects include risk estimation, risk communication, risk perception, and risk acceptance.

Risk Estimation

Risk estimation involves the use of statistical techniques to obtain a numerical value of risk. Since this is dealt with in detail elsewhere, we will give here only a brief overview.

Several measures of risk are available: **absolute risk** (and absolute risk difference); **relative risk**; and the **number needed to treat** (NNT) or the number of treatment-years to produce a single adverse outcome. The value of these measures to decision makers depends on their precision, validity (**unbiasedness**), and reliability. A precise estimate may be difficult to achieve if the study generating the measure has small sample size. This is particularly relevant when evaluating interventions with rare outcomes. For example, Miller et al. [15] estimated the risk of neurologic sequelae in previously normal children persisting 1 year after pertussis vaccination to be one in 310 000. In such circumstances, the **confidence interval** might be used to estimate an upper bound on

the risk involved. (In this case, the upper 95% limit was one in 54 000.) (*See* **Pharmacoepidemiology, Adverse and Beneficial Effects**.)

In addition, the validity of the estimate may be threatened if the design of the study that has generated the estimates is **biased**. An unbiased estimate may be difficult to achieve, particularly if the outcome is recorded using subjective measures, the intervention cannot be studied under double-blind conditions (*see* **Blinding or Masking**) or the allocation to study groups is neither concealed nor randomized [1, 2, 22] (*see* **Randomization**). Reliability, in turn, can be affected by limitations in the definition, attribution, recording, identification, classification, reporting, measurement, and analysis of information related to adverse effects of treatments [3, 8, 11]. On occasions, a further complication may be introduced by the need to obtain an exposure–response relationship. Different formulations of a given drug, for example, may contain different amounts of the drug, leading to different risk estimates. (*See* **Dose–Response Models in Risk Analysis**; **Dose-response in Pharmacoepidemiology**.)

Risk Communication

Even when the risk estimates are accurate (precise and unbiased), it would be unrealistic to expect that such estimates will lead automatically to risk assessment. It has been shown, for instance, that the type of risk estimate presented to decision makers (i.e. absolute event data versus **relative risk** reductions) can affect their assessments of risk [16]. The way in which risk estimates are described (or "framed") can also influence the reaction of those given the information. It has been shown repeatedly, for instance, that if the information is worded in a way that emphasizes the negative rather than the positive aspects of the same outcome (e.g. when describing the effects of an anticancer drug in terms of deaths rather than survivors), decision makers may alter their perception of risk and their preferences [18, 23].

In addition to the selection of the outcomes and the way in which the information is framed, the format in which the risk estimates are shown to decision makers can also affect their reactions. **Graphical displays**, for instance, are used frequently to represent risk with the hope that they will facilitate data interpretation. There is little empirical evidence, however, that could be used to guide the selection of graphical displays to communicate risk information. Some of the existing evidence suggests that patients and clinicians may interpret information from the same displays differently. When results of clinical trials are presented as survival curves, for instance, patients appear to focus more on the endpoints, while clinicians pay more attention to intermediate points [14]. Different shading and plotting symbols can also have strong effects on the visual perception of data [24]. More research is required to assess the impact of different visual displays on decision makers. A recent review, for instance, identified 13 methods to display the results of **meta-analyses** graphically, but did not find a single study on the effects of such displays on decisions [9]. Other aspects of risk communication also require more research. Little is known, for instance, on the effect of risk perception of different media to communicate risk (e.g. face-to-face contact, paper, videotapes, audiotapes, CD-ROM, Internet).

In summary, there appears to be no simple best method of presenting information to decision makers. One approach is to provide data to decision makers in several different forms, trying to explain any differences that may arise in the interpretation of the information across the methods.

Risk Perception

Risk estimation and communication play an important role in the way in which risk is perceived. However, they are not the only factors that determine risk perception. Even if decision makers are presented with accurate estimates in multiple forms, their perception of risk could be affected by factors such as: the probability value associated with the particular events and the nonlinearity of decision weights; their prior beliefs and experience; their ability to interpret probabilistic information; their intuitive rules of thumb (heuristics); and the suspicion of vested interest in those generating risk estimates (risks evaluated by those who might have a vested interest in the results may be viewed as understated compared with risks evaluated by independent sources). We consider each of these in turn.

Probability Value. It has been shown that patients and healthy volunteers can be strongly influenced

by different levels of probability of adverse events [18]. When the probability of survival given to cancer patients in that study dropped below 0.5, patients adopted a "dying mode" in which **quality of life** became more salient than quantity of life in decision making [18]. In addition, it has been suggested that decision makers weight different levels of probability in a nonlinear fashion. Moderate and high probabilities, for instance, tend to be underweighted relative to outcomes that are certain, low probabilities tend to be overweighted, and very low probabilities are either severely overweighted or neglected altogether [10]. This has important implications for risk acceptance (see below).

Prior Beliefs and Experience. Prior beliefs of decision makers can also affect their perception of risk. Clinical decision makers, like "lay" people, tend to form opinions rather quickly, usually in the absence of strong supporting evidence. These opinions, once formed, are slow to change in response to new evidence. New evidence is usually handled in a very asymmetric way by decision makers: supportive evidence tends to be considered more convincing than opposing evidence, regardless of the rigor with which it has been gathered [17]. One important factor that can influence dramatically the prior beliefs of decision makers is a vivid experience [17]. A tragic outcome with the last patient can change the way in which a clinician will perceive the risk that the same outcome will occur in the next patient. Similarly, a patient's perception of risk may be influenced more by the experience of a close friend than by evidence collected from thousands of patients in well-controlled clinical trials, leading the patient to reject such evidence – especially if it contradicts the patient's preconceived theories or challenges his/her hopes [20]. Alternatively, if an intervention has been used by clinicians or patients for a long time, the hazards associated with a "new" or unfamiliar intervention may be perceived as worse than those which are more familiar to them [5, 21].

Ability to Interpret Probabilistic Information. The ability of decision makers to interpret probabilistic information can also influence their perception of risk. In part because of their exaggerated reliance on vivid experiences or anecdotes, but also because of their lack of formal training in statistics, patients (and other lay members of the public) have a limited ability to interpret probabilistic information or any other type of scientific evidence that could help them assess risk. This "lack of training" limits their ability to build hypotheses, to assess covariation and **causation** and to predict events [17]. There is little research in relation to the understanding of statistical principles by clinicians. In a recent study, however, doctors who said that they were confident about their ability to evaluate risk of coronary heart disease consistently overestimated such risks in individual patients as well as the absolute benefits of modification of coronary risk factors [6]. In another study, the same tendency of physicians to overestimate risk was identified [19]. Furthermore, the investigators taught the physicians to make better judgments of disease probability, but such an improvement in risk assessment did not result in changes in their treatment decisions [19].

Heuristics. Even if clinicians and patients were equipped with the skills required to evaluate probabilistic information on risks, there are few choices in the clinical setting that could be informed fully with evidence generated in RCTs or **observational studies**. Despite this lack of evidence, most clinicians do make decisions. The way in which they do so in the absence of evidence has been explained by the use of intuitive *ad hoc* rules of thumb, also called "heuristics", to guide their choices [13]. Patients also use heuristics, particularly to reduce complex inferential tasks to simple judgmental operations during the evaluation of event frequency, probability and causality [17]. Many of these strategies are poorly understood and potentially problematic. There have been recent calls for the systematic study of the heuristics of medicine, hoping that the more uniform use of explicit, refined and better heuristics could lead to more efficient medical care [13]. Similar efforts should be made to understand and refine patient heuristics.

Risk Acceptance

Whether a particular risk is accepted or not depends, at least in part, on each of the factors described so far. In addition, there are other elements that could also affect risk acceptance:

Implications of the Decision. A risk may be more acceptable when the adverse event, if it occurs, has

a minor impact on the person making the decision than if it has a major impact. For instance, in cases of mental depression, the risk of experiencing the relatively minor adverse effects associated with tricyclic antidepressants or lithium may be more acceptable than the risks of suicide if the depressive disorder is left untreated.

Type of Outcome. An increase in resource utilization might be regarded as more acceptable to a patient with cancer, particularly if the costs of care are covered by an insurance company, than a reduction in the likelihood of survival.

Timing Between Decision and Outcome. Surgical mortality, an immediate hazard, may be considered worse than later death caused by the toxic effects of a drug, a deferred hazard [5].

Circumstances. Risks imposed for the benefit of others may be less acceptable than risks undertaken for self-protection. For example, compulsory whooping cough vaccinations may be imposed on older children primarily for the benefit of younger age groups [5, 20, 21].

Role. A hypothetical risk is likely to be accepted more easily by a healthy volunteer than a real risk by a patient [4].

Formal Decision Making Procedures

There exist a number of procedures for helping a patient reach a decision which attempt to incorporate the values and preferences of the patient. These include: the **standard gamble**, **time tradeoff**, **utilities**, and **willingness to pay**.

Conclusion

Biostatistics is primarily concerned with providing accurate estimates of risk in clinical situations. What we have shown in this article is that decision making following the estimation of risk depends on more than just the risks (and benefits) involved. Risk communication, perception and acceptance have subtle characteristics that play important roles. Statisticians, health care practitioners and others must be aware of these to help patients reach decisions in the clinical setting.

References

[1] Chalmers, T.C., Celano, P., Sacks, H.S. & Smith, H. (1983). Bias in treatment assignment in controlled clinical trials, *New England Journal of Medicine* **309**, 1359–1361.

[2] Colditz, G.A., Miller, J.N. & Mosteller, F. (1989). How study design affects outcomes in comparisons of therapy. I. Therapy, *Statistics in Medicine* **8**, 441–454.

[3] Cook, M. & Ferner, R.E. (1977). Adverse drug reactions: who is to know?, *British Medical Journal* **307**, 480–481.

[4] Degner, L.F. & Sloan, J.A. (1992). Decision making during serious illness: what role do patients really want to play, *Journal of Clinical Epidemiology* **45**, 941–950.

[5] Department of the Environment & Health and Safety Executive (1979). Risk Assessment and the Acceptability of Risk. Summary Joint Seminar held by DOE and HSE, Sunningdale, January 3–4, 1979 (unpublished), quoted in Royal Society Study Group (1983).

[6] Grover, S.A., Lowensteyn, I., Esrey, K.L., Steinert, Y., Joseph, L. & Abrahamowicz, M. (1995). Do doctors accurately assess coronary risk in their patients? Preliminary results of the coronary health assessment study, *British Medical Journal* **310**, 975–978.

[7] Haynes, R.B., Sackett, D.L., Gray, J.A.M., Cook, D.J. & Guyatt, G.H. (1996). Transferring evidence from research to practice: 1. The role of clinical care research evidence in clinical decisions, *ACP Journal Club* **125**, No. 3, A14–A15.

[8] Jadad, A.R. (1994). Meta-Analysis of Randomised Clinical Trials in Pain Relief, D. Phil. thesis. University of Oxford.

[9] Jadad, A.R., Raina, P., Cook, D.J., Walter, S.D., Torrance, V., Krueger, P.D. & Chambers, L.W. (1996). Selecting methods to display graphically the results of systematic reviews: where is the evidence? Paper presented to the 4th Cochrane Colloquium, Adelaide, Australia.

[10] Kahneman, D. & Tversky A. (1984). Choices, values and frames, *American Psychologist* **39**, 341–350.

[11] Koch-Weser, J., Sellers, E.M. & Zacest, R. (1977). The ambiguity of adverse drug reactions, *European Journal of Pharmacology* **11**, 75–78.

[12] Llewellyn-Thomas, H. (1995). Patients' health-care decision making: A framework for descriptive and experimental investigations, *Medical Decision Making* **15**, 101–106.

[13] McDonald, C.J. (1996). Medical heuristics: the silent adjudicators of clinical practice, *Annals of Internal Medicine* **124**, 56–62.

[14] Mazur, D.G. & Hickam, D.H. (1993). Patients' and physicians' interpretations of graphic data displays, *Medical Decision Making* **13**, 59–63.

[15] Miller, D.L., Adderslade, R. & Ross, E.M. (1982). Whooping cough and whooping cough vaccine: the risks and benefits debate, *Epidemiologic Reviews* **4**, 1–24.

[16] Naylor, C.D., Chen, E. & Strauss, B. (1992). Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Annals of Internal Medicine* **117**, 916–921.

[17] Nesbitt, R. & Ross, L. (1981). *Human Inference: Strategies and Shortcomings of Social Judgement*. Prentice-Hall, Englewood Cliffs.

[18] O'Connor, A.M. (1989). Effects of framing and level of probability on patients' preferences for cancer chemotherapy, *Journal of Clinical Epidemiology* **42**, 119–126.

[19] Poses, R.M., Cebul, R.D. & Wigton, R.S. (1995). You can lead a horse to water – improving physicians' knowledge of probabilities may not affect their decisions, *Medical Decision Making* **15**, 65–75.

[20] Redelmeier, D.A., Rozin, P. & Kahneman D. (1993). Understanding patients' decisions - cognitive and emotional perspectives, *Journal of the American Medical Association* **270**, 72–76.

[21] Royal Society Study Group (1983). *Risk Assessment: Report of a Royal Society Study Group*. The Royal Society, London.

[22] Schulz, K.F., Chalmers, I., Hayes, R.J. & Altman, D.G. (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effect in controlled clinical trials, *Journal of the American Medical Association* **273**, 408–412.

[23] Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice, *Science* **211**, 453–458.

[24] Walter, S.D. (1993). Visual and statistical assessment of spatial clustering in mapped data, *Statistics in Medicine* **12**, 1275–1291.

(*See also* **Clinical Epidemiology**; **Pharmacoepidemiology, Overview**).

H.S. SHANNON & A. JADAD

Risk Assessment

The various usages of the term *risk* all concern the possible occurrence of events or situations, called *hazards*, the consequences of which are uncertain, but may be harmful. Informal usages of **risk** may indicate the nature, or merely the existence, of the possible danger ("There is a risk of post-operative infection"; "I never take risks"). In technical discussions the term is used quantitatively, but even there the usage is not standard.

There are two principal, and mutually incompatible, interpretations:

1. The probability, or chance, of an adverse event. Clearly, this must be put in context: it should refer to a defined set of circumstances, and, for hazards continuing over time, the rate per time unit, or for a unit of exposure, is normally used.
2. A combination of the chance of an adverse effect and its severity. There are obvious difficulties with this type of definition: How is severity measured, and how are the two components combined?

The extensive literature on risk covers many aspects, which are commonly collectively termed *risk assessment* or *risk analysis*. The first of these terms is sometimes used more restrictively, to include the concepts of *risk estimation, risk evaluation*, and *risk perception*, as defined below. The study of risk brings together engineers, behavioral and social scientists, statisticians, and others, and to some extent usage of terms varies amongst these groups. For example, engineers and other technologists tend to favor approach 2, statisticians and biologists tend to favor 1, and behavioral and social scientists tend often to use a mutifaceted approach. Reference [6] contains chapters by groups of writers from different backgrounds, and has extensive bibliographies. See also [1] for a popular exposition.

Risk theory has a specialized meaning, being concerned with the financial integrity of an insurance company in the light of random fluctuations in claims. It forms an application of the theory of **stochastic processes** [7].

Statisticians will note that usage 2 above is closely related to the concept of a *risk function* in **decision theory**. There, uncertain events, the distribution of which depends on an unknown scenario, have consequences measured by a **loss function**; a particular decision function, defining the action to be taken when the event is observed, has an average loss for any given scenario; and the *risk* (or *integrated risk*) is the mean of the average loss when taken over the **prior distribution** of the scenarios. Application of this approach is hampered by the difficulty of determining losses in financial terms, and of defining the various probability distributions.

Attention has been focused on various interrelated aspects of risk, including the following:

1. *Risk estimation*: the estimation of the probabilities of the adverse outcomes, and of the nature and magnitude of their consequences.
2. *Risk evaluation*: determination of the significance of the hazards for individuals and communities. This depends importantly on the next aspect.
3. *Risk perception*: the extent to which individuals assess risks and the severity of possible outcomes, assessments that may differ from those made by "experts".
4. *Risk management*: the measures taken by individuals and societies to prevent the adverse effects of hazards and to ameliorate their consequences.

We deal briefly with these topics in turn. The articles in [6], and their bibliographies, provide a much broader picture. Many of these topics are discussed fully elsewhere, in relation to **risk assessment for environmental chemicals** (also, *see* **Risk Assessment in Clinical Decision Making**).

Health Hazards

There are several clearly distinct categories of hazards that give rise to health risks.

First, there are hazards that arise from the physical and biological environment. Many of the hazards in the physical environment are man-made. It is our own choice, collectively, to pollute the atmosphere with emissions from domestic fires, power stations, or burning oil wells, and to treat water supplies with disrespect. These are examples of damage *to* the environment, and damage to ourselves *from* the environment. The biological environment presents a hazard to us mainly in the form of microorganisms causing infectious disease.

Other hazards arise from personal, rather than societal, choice. These include habits with adverse consequences, such as the consumption of tobacco, alcohol, and narcotic drugs. The category includes also indulgence in sport and travel; and our often unwise dietary choices. We tend to shrug off the hazards that we ourselves incur, by understating the risks or overstating the benefits, while deprecating the folly of others.

Finally, there are hazards that cannot be prevented by personal decisions. They follow inexorably from our innate or ingrained characteristics – our genetic makeup or our experiences in early life. In some instances, medical science can reduce the risk to which susceptible individuals are subject: in others, the burden has to be endured.

Risk Estimation

The risks from many prominent health hazards can be estimated reliably from objective statistical information. In other instances, in which numeric information is lacking, risks may be guessed by informed experts (as in the setting of insurance premiums for nonstandard risks; *see* **Actuarial Methods**). There are, for instance, no reliable data on the frequency of explosions at nuclear power installations, and estimates of risk would have to rely on expert judgments, or on careful estimation of risks of failure at individual links in the chain of connected events.

Even when statistical information is available, an individual may argue that his or her risk is not properly represented by the population estimate. There is a long-standing debate as to whether medical statistical information necessarily applies to an individual in the population concerned; in the nineteenth century, for instance, opposite views were held by **P.C.A. Louis** and by **C. Bernard**. Clearly, if the individual has known characteristics that can be shown to affect the risk, they should be taken into account. If no such characteristics can be identified, it seems reasonable to apply the population estimate to the individual. The point is important, in emphasizing that risk estimation is far from being an objective matter.

Statistical information on the risk of mortality from different diseases is widely available, for individuals of each sex at different ages, in different occupational (*see* **Occupational Epidemiology**) and social groups and for different countries (*see* **Mortality, International Comparisons**). Information on the risks of morbidity is less comprehensive. Such information, based on data for large populations, is of some value for the estimation of risks for random members of the populations, but gives little or no indication of risks for individuals exposed to certain specified hazards; such as environmental pollution (*see* **Environmental Epidemiology**), social habits, or the onset of disease.

For questions of this type, special investigations are required. The whole range of types of epidemiologic study is available, including **case–control studies**, **cohort studies**, and **case–cohort studies**. The risks of adverse progression of disease may be estimated by a study of **prognosis**. See [3] for an example of various investigatory methods employed in a study of the apparent excess risks of childhood leukemia (*see* **Leukemia Clusters**) due to contamination of water supplies in a town.

In many "high-profile" public health problems, it is not possible to mount epidemiologic studies to give unambiguous estimates of risk. The mechanism giving rise to the risk may not be fully understood, or the dangers may arise from a complex chain, the risks for which are difficult to measure. In such instances, the risks may be estimable only within very broad bands. For instance, in the crisis in the British beef industry, due to the outbreak of bovine spongiform encephalopathy (**BSE**), leading to an apparent risk of Creutzfeldt–Jakob disease (CJD), it was very difficult to estimate precisely the risk of CJD to a person eating beef. Since the cessation of use of suspect cattle feed, and the culling of relevant herds, it is probably reasonable to say that the risk is "extremely low", and perhaps to put some upper bound on it, but such estimates would rely on somewhat shaky data, and on the personal judgments of experts.

Another example is that of prolonged exposure to low levels of possibly carcinogenic chemicals. Carcinogenicity experiments, with the administration of high doses to animals (*see* **Animal Screening Systems**; **Serial-sacrifice Experiments**), may give quite precise estimates of a **median effective dose**. However, risk estimation for low-level exposure to humans involves **extrapolation to low doses** (using models that are not necessarily correct [5]), and from the animal to the human species (*see* **Dose–Response Models in Risk Analysis**). The result of such extrapolation may well be reassuring, but it is unlikely to be quantitatively precise.

Risk Evaluation and Perception

The evaluation of risk, either by individuals or by societies, should in principle involve a balancing of the costs and benefits: the potential occurrence of adverse effects, arising from exposure to a hazard, should be balanced against the potential benefits in physical or psychological rewards. Cost–benefit analysis (*see* **Health Economics**) is, however, a somewhat idealized concept. Apart from the difficulties of risk estimation, outlined above, both the potentially adverse effects and the supposed benefits may be difficult to evaluate on commensurate scales.

The benefits may in part be assessable as direct economic gains to a community. They may also include amenities, such as palatable food or attractive cosmetics, the value of which may be estimable by enquiry as to the prices that people are willing to pay for them.

The costs may be even more elusive. They include direct financial losses; for instance, in productivity. They include also disbenefits of pain and other symptoms. One might enquire how much people would be willing to pay to avoid such discomforts, but this would be a difficult exercise for people who had never experienced the symptoms in question.

Then, there is the crucial question of the value of human life. There are various approaches to this task, such as: (i) calculation of lost earning capacity; (ii) implicit evaluation based on societal practice, such as compensation awards or expenditure on specific safety measures; or (iii) the size of insurance premiums.

None of these possible approaches is likely to be simple, but it seems important to encourage further discussion and research, especially for the evaluation of risks for which community decisions, such as the imposition of government regulations, are required.

Evaluation by individuals of risks incurred by possible individual choices, again in principle involves the balancing of costs and benefits, but these may be very subjective and even more difficult to quantify than those involved in community action. In a sense, the decisions actually taken by individuals, sometimes without appreciable introspection, carry implications about the values attached by those individuals to the various elements in the equation. From this point of view, the relevant estimates of risk may be the subjective perceptions of the individuals themselves, rather than more "objective" estimates provided by experts. These two forms of estimate may be quite disparate. We tend to be more concerned about infrequent but dramatic events, such as major air crashes, than about frequent but less dramatic series of events such as the regular toll of road accident deaths. In one study [4], people thought that accidents caused as many deaths as disease, whereas in fact disease causes 15 times as many. The incidences of death from spectacular causes such as murder, botulism, tornadoes, and floods were all overestimated, whereas those for cancer, stroke, and heart disease were underestimated.

The importance of the "benefit" side of the equation is illustrated by the varying acceptability of activities with comparable risks. People are generally prepared to accept much higher risks of death from activities in which they participate voluntarily, such as sports, than from those encountered involuntarily.

Risk Management

This term covers the decisions, taken by individuals and communities, to accept or forego hazardous situations after assessment of risks, or to reduce exposure to the hazards and/or their adverse consequences.

As noted above, decisions by individuals are highly personal, and to a detached observer they may often seem irrational. A rational study of teenage smoking may conclude that the hazardous practice should be avoided, but its conclusions may carry little weight with a young person who is ill-informed about risks, and whose "benefits" include the pleasures of conformity with peer practice. Nevertheless, in such situations, improved information about risks and adverse consequences is highly desirable, and the provision of risk information forms one of the major roles of government and other public bodies concerned with risks.

Institutions with a role in risk management include international, national, and regional governments, and a variety of public and private organizations. Apart from the provision of information, governments may issue regulations to reduce or control the use of hazardous substances. Their decisions may be guided by advisory committees, perhaps internationally based. For instance, in the assessment of evidence of carcinogenicity of chemicals, authoritative advice is provided by the program of the International Agency

for Research on Cancer (IARC) Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans [2].

Institutions concerned with mitigation of the adverse effects of hazards include the judiciary (through compensation awarded in the law courts), insurance companies, and a variety of community bodies concerned with social welfare.

Conclusions

The interdisciplinary nature of all the aspects of risk assessment discussed here has encouraged lively discussion and research. Biostatistics forms only one component in the mixture, but it is an essential ingredient. Publications are spread widely in the technical press, but special note should be taken of the journal *Risk Analysis*.

References

[1] British Medical Association (1987, 1990). *The BMA Guide to Living with Risk*. Penguin, London.

[2] International Agency for Research on Cancer (IARC) (1982). *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, Supplement 4, Chemicals, Industrial Processes and Industries Associated with Cancer in Humans*. *IARC Monographs*, Vols. 1–29. International Agency for Research on Cancer, Lyon.

[3] Lagakos, S.W., Wessen, B.J. & Zelen, M. (1986). An analysis of contaminated well water and health effects in Woburn, Massachusetts (with discussion), *Journal of the American Statistical Association* **81**, 583–614.

[4] Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M. & Combs, B. (1978). Judged frequency of lethal events, *Journal of Experimental Psychology: Human Learning and Memory* **4**, 551–578.

[5] Lovell, D.P. & Thomas, G. (1996). Quantitative risk assessment and the limitations of the linearized multistage model, *Human and Experimental Toxicity* **15**, 87–104.

[6] Royal Society (1992). *Risk: Analysis, Perception and Management*. Report of a Royal Society Study Group. Royal Society, London.

[7] Seal, H.L. (1988). Risk theory, in *Encyclopedia of Statistical Sciences*, Vol. 8, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 152–156.

(*See also* **Postmarketing Surveillance of New Drugs and Assessment of Risk**)

PETER ARMITAGE

Risk Factor

A factor whose presence is associated with an increase in the probability of developing a disease is called a risk factor for that disease. It is a generic term widely used in epidemiology (*see* **Epidemiology, Overview**) that can stand for genetic traits, sociodemographic characteristics as well as occupational, environmental, or any other types of exposures. This definition implies that a risk factor for a given disease must be present before disease occurrence. By analogy to risk factors that refer to the development of a disease, prognostic factors are defined as factors whose presence is associated with an increase in the probability of patients developing a certain outcome (e.g. recurrence or death) during the course of disease (*see* **Clinical Epidemiology**; **Prognosis**). An association between a risk factor and a disease can be quantified by various measures such as the **relative risk**, **hazard ratio**, **odds ratio**, or risk difference. Risk factors are usually identified from **observational** epidemiologic studies so that the association between a risk factor and a disease is not necessarily of a causal nature (*see* **Causation**; **Hill's Criteria for Causality**), which is why Miettinen [1] recommended the use of the term *risk indicator* rather than risk factor. In contrast with risk factors, factors whose presence is associated with a decrease in the probability of developing a disease are usually called *protective factors*.

Reference

[1] Miettinen, O.S. (1985). *Theoretical Epidemiology*. Delmar Publishers, Albany, p. 10.

JACQUES BENICHOU

Risk Set

In survival analysis, one of the most frequently used methods is the **Kaplan–Meier** estimator for nonparametric estimation of the survival distribution function $S(t)$ based on right-**censored** data. Thus, let $X_1, \ldots, X_n$ be independent identically distributed lifetimes with $S(t) = \Pr(X_i \geq t)$ and assume that $(\tilde{X}_i, D_i)$, $i = 1, \ldots, n$, are observed. Here, $\tilde{X}_i = X_i$ and $D_i = 1$ if individual i is observed to die and $\tilde{X}_i = U_i$, a right-censoring time, and $D_i = 0$ if the lifetime of individual i is censored at time U_i, in which case the only information on X_i is that $X_i > U_i$ ("independent censoring"). Then the Kaplan–Meier estimator is given by

$$\hat{S}(t) = \prod_{\tilde{X}_i \leq t} \left(1 - \frac{D_i}{r(\tilde{X}_i)}\right), \tag{1}$$

where $r(s) = \#R(s)$ and

$$R(s) = \{i : \tilde{X}_i \geq s\} \tag{2}$$

is the *risk set* at time s; that is, the set of individuals alive and uncensored just before time s. Thus, the concept of a risk set is fundamental in nonparametric **survival analysis**.

Also, when the survival times are subject to left-truncation (**delayed entry**) – that is, individual i is followed, not necessarily from time 0, but maybe from a later entry time $V_i \geq 0$ – the survival distribution function may be estimated by (1) by redefining the risk set to be the set

$$R(s) = \{i : V_i < s \leq \tilde{X}_i\} \tag{3}$$

of individuals alive and uncensored just before time s and with entry times before time s.

The **Nelson–Aalen estimator**

$$\hat{A}(t) = \sum_{\tilde{X}_i \leq t} \frac{D_i}{r(\tilde{X}_i)}, \tag{4}$$

of the cumulative **hazard** function $A(t) = -\log S(t)$ also involves the risk set in an explicit way, and since most linear nonparametric test statistics for comparison of survival distributions are based on sums of weighted differences of the Nelson–Aalen estimators, the risk sets, again, play a fundamental role.

The **Cox regression model** states that the hazard function for the conditional distribution of X_i given covariates $\mathbf{Z}_i$ is given by

$$\alpha_i(t|\mathbf{Z}_i) = \alpha_0(t) \exp(\boldsymbol{\beta}'\mathbf{Z}_i),$$

and the unknown regression coefficients β are estimated by maximizing the Cox **partial likelihood**

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_i)}{\sum_{j \in R(\tilde{X}_i)} \exp(\boldsymbol{\beta}'\mathbf{Z}_j)} \right)^{D_i}. \tag{5}$$

Thus, at every failure time, the risk score $\exp(\boldsymbol{\beta}'\mathbf{Z}_i)$ for the individual failing at that time is compared with the sum of the corresponding risk scores $\exp(\boldsymbol{\beta}'\mathbf{Z}_j)$ for all individuals, j, in the risk set (2) or (3) at that time.

If the sample size, n, is large and, in particular, if the Cox regression includes time-dependent covariates, then the calculation of the sum over the risk set in (5) may be time consuming. In such cases, sampling from the risk set may be advantageous. This amounts to replacing $R(\cdot)$ in (5) by a sampled subset, say $\tilde{R}(\cdot)$, of the risk set. This kind of sampling is frequently used in epidemiology in so-called nested case–control designs.

In multistate survival models based on, for example, nonhomogeneous **Markov processes**, the Nelson–Aalen estimator (4) carries over in a rather straightforward manner. Thus, for the intensity of transition $\alpha_{hj}(t)$ from *state* h to *state* j one may estimate the integral $A_{hj}(t) = \int_0^t \alpha_{hj}(s)\,\mathrm{d}s$ by the sum

$$\widehat{A_{hj}}(t) = \sum_{X_{hji} \leq t} \frac{D_{hji}}{r_h(X_{hji})}$$

over all the observed $h \to j$ transition times, $X_{hji} \leq t$. Here, D_{hji} is the number of such transitions at X_{hji} and $r_h(s)$ is the number of individuals *at risk for making an* $h \to j$ *transition* just before time s; that is, the number of individuals in state h at time $s-$ or, stated slightly differently, *the size of the type* h *risk set* at time $s-$. So, also in multistate models, the concept of a risk set is important.

PER KRAGH ANDERSEN

Risk

Risk is the probability that an individual without disease will develop disease over a defined age or time interval. If risk is estimated as the proportion of members of a fixed cohort who develop disease in a defined time period, it corresponds to an average individual-specific risk (*see* **Cumulative Incidence**). This proportion is an estimate of a **crude risk** or **absolute risk** because it is reduced by the chance that subjects will die of other diseases before they develop the disease of interest (*see* **Competing Risks**).

Often the risk for the interval $[0, t)$ is calculated from $1 - \exp(-\int_0^t \lambda(u)\,\mathrm{d}u)$, where $\lambda(u)$ is the cause-specific **hazard rate**. For small hazard rates, this expression is approximately equal to the **cumulative hazard**, $\int_0^t \lambda(u)\,\mathrm{d}u$. These expressions correspond to pure probabilities of disease and estimate the probability of developing disease in the absence of other causes of death under the assumption that the various causes of death act independently.

MITCHELL H. GAIL

Robust Methods in Time Series Analysis

In general, a statistical procedure might be said to be robust if it is not overly sensitive to departures from any assumptions upon which it depends. **Robustness** was introduced by Box [4] when he considered the properties of tests of equality of variances. Robust *estimators*, usually of *location* and *scale* parameters, are not sensitive to the presence of **outliers**. The three main kinds of robust estimator are those defined as L-estimators (linear functions of **order statistics**), R-estimators (based on tests involving **ranks**), and M-estimators (maximization of some function of the data and parameter). There exists an extensive literature concerned with the robust estimation of location and scale parameters in statistical analysis [2, 15, 17, pp. 157–162].

Associated with the notion of robustness, particularly in the context of time series, is the procedure of smoothing – where the purpose is to "remove or suppress", in some way, the possible effects of *contamination* or *spurious* information. This occurs commonly with biomedical time series data. Tiao & Xu [30] bring these ideas together and extend methods first introduced by Cox [12], to produce robust exponential smoothing in multistep forecasts with autoregressive integrated moving average (ARIMA) models (*see* **ARMA and ARIMA Models**).

Many biostatistical investigations use the class of Gaussian ARIMA models to analyze time series data. If $\{Y_t\}$ is a time series, then the standard ARIMA (p, d, q) model is written

$$\Phi(B)\nabla^d Y_t = \Theta(B)a_t, \qquad (1)$$

where B is the backshift operator $B(Y_t) = Y_{t-1}$ (*see* **Backward and Forward Shift Operators**); $\nabla \equiv 1 - B$, the differencing operator; $\Phi(B) \equiv (1 - \phi_1 B - \cdots - \phi_p B^p)$, $\Theta \equiv (\theta_0 - \theta_1 B - \cdots - \theta_q B^q)$ the autoregressive and moving average polynomials in B, respectively; $\{a_t\}$ is an assumed white-noise process (*see* **Noise and White Noise**); and, $\phi_1, \ldots, \phi_p$ and $\theta_0, \theta_1, \ldots, \theta_q$, unknown parameters to be estimated. The specification and estimation of p, d, and q also present robustness considerations.

To fit the model, the investigator carries out a sequence of steps: tentatively specifying a model; using maximum likelihood to estimate the parameters; and then, performing various diagnostic tests to check the adequacy of the model. If the model is found to be deficient in some respect, than the sequence is repeated.

This general procedure is an important example where *robustification* [17, pp. 176–181] is often necessary to remove the unwarranted effects of contamination or the breakdown of assumptions, or both. A detailed review and summary of the many different aspects and considerations of robust time series techniques (up to 1987) is provided by Stockinger & Dutter [27].

Time Series Outliers

Outliers in time series are usually classified as either innovation or additive [14]. For a variety of reasons, *gross* values can occur quite "naturally" in sets of biological and medical observations – for example, in the continuous monitoring of heart-rate, sudden and unusual measurements can often be observed, attributable to well-known characteristics of the body system. In fact, it is the nature of many feed-back mechanisms in the human body to allow and adjust for such events.

In general, there are two main ways of handling time series outliers. Detection followed by removal, or robust modeling and estimation. Details of the detection approach can be found in [8], in which outliers are identified on a one-by-one basis and the contaminated series adjusted accordingly.

Any algorithm used to estimate the parameters of a model must be able to protect these estimates from outliers. Because the estimation problem is nonlinear, these *robust algorithms* are almost always *iterative*. Sejling et al. [26] discuss various such algorithms, and, building upon previous work, introduce a general method for obtaining recursive robust parameter estimates in autoregressive (AR) models. Two algorithms derived from this model are compared with the use of a recursive *least squares* estimation algorithm, in which the outliers are treated as *missing*. McDougall [22] extends these ideas and produces methods applicable to general ARIMA models, where both kinds of outliers, innovation and additive, might be present.

With the knowledge that any time series $\{Y_t; t = 0, \ldots, N-1\}$ can be represented as the sum of N sines and cosines at the Fourier frequencies $\{\omega_k =$

$2\pi k/N; k = 0, \ldots, N/2\}$, an approach to the analysis of real data is to produce initially a robust form of the discrete Fourier transform (*see* **Fast Fourier Transform (FFT)**). The coefficients can then be inverse Fourier transformed to produce a *filter* for the data. The filtered time series data can then be used for standard analysis. Tatum & Hurvich [29] give details of a filter that can handle large amounts of contamination and outliers. Details of other non-Gaussian filters that deal with both types of outlier can be found in [1, 16], and [9].

Other methods for robust filtering and smoothing include the use of the robust Kalman filter [10], *M* estimate smoothers, and robustified **splines** [27].

Robust Bayesian Estimation

Bayesian ideas and methods have been used by a number of authors to obtain time series models robust to outliers. Le et al. [18] use the ratio of posterior to prior odds, known as *Bayes factors*, to compare autoregressive models on a pairwise basis. This comparison is made, in the presence of outliers, using a *robust likelihood* procedure following the techniques of Martin [24]; see also [28].

Neural Networks

The application of an artificial **neural network** to nonlinear systems has been a major area of research in recent years. Wu [32] gives details of an economic example, where the performances of ARIMA models and neural networks are compared in terms of their robustness. Connor et al. [11] developed a robust learning algorithm, applied it to *recurrent* neural networks in order to filter outliers from time series data, and assessed the sensitivity of the procedure.

Statistical Tests

A well-known statistic for testing the adequacy of a time series model is the Box–Pierce [5] or so-called *portmanteau statistic*. Li [19], Chan [7], and Wong & Li [31] have produced robust versions of this statistic, and examined their performance in the presence of outliers. Li & Hui [20] have produced a robust multivariate version of the portmanteau statistic for use in multiple time series modeling. Large-sample properties of robust M-estimates used in the testing of hypotheses in autoregressive models can be found, for example, in [3]. Li & Hui [21] have also developed robust tests for lagged relations between two time series.

Examples

A discussion of many of the issues, including that of robustness, associated with autoregressive and **spectral analysis** models in heart rate variability studies is given in [6]. The robust smoothing and filtering of psychophysiological time series data has been considered by Schmitz et al. [25]. Details of methods, originally developed for engineering applications but applicable to many problems in the medical and life sciences, to identify and model time-varying biological systems can be found in [23]. The authors show, in **simulation** studies, that these methods are robust to the presence of general noise in the system, and apply the techniques to a study of ankle stiffness.

An investigator must not lose sight of the fact that what constitutes an outlier, contamination, or spurious information is often a matter of judgment and degree. Understanding of the system (e.g. biological or medical) that has given rise to the data is important in all aspects of the modeling of biostatistical time series data [13].

References

[1] Alspach, D.L. & Sorenson, H.W. (1971). Recursive Bayesian estimation using Gaussian sums, *Automatica* **6**, 465–479.

[2] Barnett, V.M. & Lewis, T. (1977). *Rejection of Outliers*. Wiley, New York.

[3] Basawa, I.V., Huggins, R.M. & Staudte, R.G. (1985). Robust tests for time series with an application to first-order autoregressive processes, *Biometrika* **72**, 559–571.

[4] Box, G.E.P. (1953). Non-normality and tests on variances, *Biometrika* **40**, 318–335.

[5] Box, G.E.P. & Pierce, D.A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *Journal of the American Statistical Association* **65**, 1509–1526.

[6] Burr, R.L. & Cowan, M.J. (1993). Autoregressive spectral models of heart rate variability, *Journal of Electrocardiology* **25**, Supplement, 224–233.

[7] Chan, W.S. (1994). On portmanteau goodness-of-fit tests in robust time series modelling, *Computational Statistics* **9**, 301–310.

[8] Chang, I., Tiao, G.C. & Chen, C. (1988). Estimation of time series parameters in the presence of outliers, *Technometrics* **30**, 193–204.

[9] Chow, H.-K. (1994). Robust estimation in time series: An approximation to the Gaussian sum filter, *Communications in Statistics – Theory and Methods* **23**, 3491–3505.

[10] Cipra, T. & Romera, R. (1991). Robust Kalman filter and its application in time series analysis, *Kybernetika* **27**, 481–494.

[11] Connor, J.T., Martin, R.D. & Atlas, L.E. (1994). Recurrent neural networks and robust time series prediction, *IEEE Transactions on Neural Networks* **5**, 240–254.

[12] Cox, D.R. (1961). Prediction by exponentially weighted moving averages and related methods, *Journal of the Royal Statistical Society, Series B* **23**, 414–422.

[13] Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

[14] Fox, A.J. (1972). Outliers in time series, *Journal of the Royal Statistical Society, Series B* **34**, 350–363.

[15] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

[16] Kitagawa, G. (1988). Non-Gaussian state space modelling of non stationary time series, *Journal of the American Statistical Association* **82**, 1032–1063.

[17] Kotz, S. & Johnson, N.L. (1992). *Encyclopedia of Statistical Sciences*, Vol. 8. Wiley, New York.

[18] Le, N.D., Raftery, A.E. & Martin, R.D. (1996). Robust Bayesian model selection for autoregressive processes with additive outliers, *Journal of the American Statistical Association* **91**, 123–131.

[19] Li, W.K. (1988). A goodness-of-fit test in robust time series modelling, *Biometrika* **75**, 355–361.

[20] Li, W.K. & Hui, Y.V. (1989). Robust multiple time series modelling, *Biometrika* **76**, 309–315.

[21] Li, W.K. & Hui, Y.V. (1994). Robust residual cross correlation tests for lagged relations in time series, *Journal of Statistical Computation and Simulation* **49**, 103–109.

[22] MacDougall, A.J. (1994). Robust methods for recursive autoregressive moving average estimation, *Journal of the Royal Statistical Society, Series B* **56**, 189–207.

[23] MacNeil, J.B., Kearney, R.E. & Hunter, I.W. (1992). Identification of time-varying biological systems from ensemble data, *IEEE Transactions on Biomedical Engineering*, **39**, 1213–1225.

[24] Martin, R.D. (1981). Robust methods for time series, in *Applied Time Series II*, D.F. Findley, ed. Academic Press, New York.

[25] Schmitz, N., Kugler, W., Neumann, W. & Krüskemper, G. (1993). Robust smoothing and filtering of psychophysiological time series, *International Journal of Psychophysiology* **14**, 148.

[26] Sejling, K., Madsen, H., Holst, J., Holst, U. & Englund, J.-E. (1994). Methods for recursive robust estimation of AR parameters, *Computational Statistics & Data Analysis* **17**, 509–536.

[27] Stockinger, N. & Dutter, R. (1987). *Robust Time Series Analysis: A Survey*. Acadmia, Praha.

[28] Taplin, R.H. (1993). Robust likelihood calculation for time series, *Journal of the Royal Statistical Society, Series B* **55**, 829–836.

[29] Tatum, L.G. & Hurvich, C.M. (1993). High breakdown methods of time series analysis, *Journal of the Royal Statistical Society, Series B* **55**, 881–896.

[30] Tiao, G.C. & Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case, *Biometrika* **80**, 623–641.

[31] Wong, H. & Li, W.K. (1995). Portmanteau test for conditional heteroscedasticity, using ranks of squared residuals, *Journal of Applied Statistics* **22**, 121–134.

[32] Wu, B. (1995). Model-free forecasting for nonlinear time series (with application to exchange rates), *Computational Statistics & Data Analysis* **19**, 433–459.

CLIVE J. LAWRENCE

Robust Regression

The term *robust regression* refers to a collection of procedures and a body of theory associated with the application of robust or resistant procedures to **regression models**.

Robustness and resistance are related properties which differ slightly in their foundations: the definition of robustness is grounded in distributional assumptions while that of resistance is based on the numerical properties of an algorithm. The idea of robustness has many different tight mathematical definitions, all capturing some characteristic of a procedure the properties of which are relatively insensitive to small changes in the assumptions specifying a model. The property of resistance, on the other hand, is the insensitivity of the results of a procedure to changes in a small fraction of the data.

Here, a regression model will denote a statistical model relating a response y to a collection of explanatory variables $\mathbf{x} = (x_1, \ldots, x_p)'$. The regression models most commonly used include: **additive models** of the form

$$y = \mathbf{x}'\boldsymbol{\beta} + \sigma e,$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, σ is a scale parameter, and e denotes an error variable, typically of mean 0; **logistic regression**, where y is assumed to be **binomial** with mean $n\exp(\mathbf{x}'\boldsymbol{\beta})/[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]$; and **loglinear** models, where y is assumed to be **Poisson** with mean $\exp(\mathbf{x}'\boldsymbol{\beta})$.

Regression models are very widely applied. Their structure has implications for the robustness and resistance of any estimation procedure, and has given rise to many different procedures for estimation and testing. This is because the observations, although typically assumed to be independent, are not identically distributed and, as a result, have differing influences on the estimates. The ideas of robustness and resistance are illustrated by the mean and median in a location model (*see* **Location–Scale Family**).

Location Models

Location models are the simplest regression models. These involve only one parameter and have the form $y = \mu + \sigma e$. The parameter may be estimated by the **mean** $\overline{y}$ or by the **median** $\tilde{y}$ (and a great many other estimators). Small changes in the tail length of the distribution can make much greater changes in the variance of the mean than of the median. The median is more robust than the mean. Large changes in only one observation produce much greater changes in the mean than in the median. The median is more resistant.

Least Squares Regression

Least squares estimates are **maximum likelihood** estimates if the errors e are assumed to be independent **normally distributed** or Gaussian random variables with mean 0 and common variance. The Gauss–Markov theorem (*see* **Least Squares**) states that the resulting estimates have minimum variance among the class of linear, unbiased estimates. However, because the estimate is linear, it is not resistant: one observation y can change the estimate by an arbitrarily large amount. And, as in the case of location, the variance of the parameter estimate depends heavily on the assumed tail length of the distribution. The estimates are not robust.

These properties are exacerbated if some observations are potentially highly influential. This arises when the explanatory variables of the observations are very different from the rest. This is easily seen by considering the variances of the **residuals**. If $\mathbf{X}$ denotes the matrix with rows the explanatory variables $\mathbf{x}'$, then the variance matrix of the residuals is proportional to $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{H}$. The diagonals of this matrix, $1 - h_{ii}$, are the variances of individual residuals. Some of these can be arbitrarily small. If this is the case, then the associated residuals must be small and the fitted values $\mathbf{x}'\hat{\boldsymbol{\beta}}$ very close to the observation. This implies that observations associated with small residual variances will have a determining effect on the estimation. Such points are called leverage points (*see* **Diagnostics**).

L_p Regression Estimation

For the location model, least squares, minimizing the sum of squares of the residuals, $y - \hat{\mu}$, led to the estimate $\hat{\mu} = \overline{y}$. The median may be shown to minimize the sum of absolute residuals. These are special cases of L_p estimates, which are defined to minimize the sum of pth powers of the residuals. The robustness and resistance of the median might

suggest that the same properties would hold for L_p estimates with $p < 2$ and for $p = 1$ in particular. However, this is not the case if points of high leverage exist. An observation with very unusual explanatory variables will have fitted values very dependent on the parameter estimates. These differences will have a large, determining effect on L_p estimates, including L_1 estimates.

M-Estimation for Regression

Maximum likelihood estimates for regression models are defined to maximize $\sum \psi((y - \mathbf{x}'\hat{\boldsymbol{\beta}})/\sigma)$, where $\psi(\cdot)$ is the log density of e. If ψ is differentiable, the estimates may be defined by the system of p equations

$$0 = \sum x_i \phi\left(\frac{y - \mathbf{x}'\hat{\boldsymbol{\beta}}}{\sigma}\right) \quad \text{for } i = 1, \ldots, p, \quad (1)$$

where ϕ is the derivative of the log density.

M-estimates are the generalization of (1) that arises when ϕ is not restricted by a relation to a density function. Different choices of ϕ lead to different estimators with different properties of resistance and robustness.

For example, if ϕ is unbounded, a single observation may contribute an arbitrarily large component to the sum in (1). Bounded functions, ϕ, may be expected to have greater resistance to individual observations. If ϕ is 0 outside of a finite interval, observations outside of this interval will not contribute to the sum defining the estimate. Estimates defined by such functions will be insensitive to the length of the extreme tails of the distribution. Such estimates will be robust to differences in the extreme tails.

Huber [3], studying the location model, used a specific measure of distance between distributions and found the form of ϕ to minimize the maximum variance of the estimate in a neighborhood of the normal or Gaussian distribution. The resulting ϕ was the derivative of the density of the least favorable distribution in the neighborhood. The function ϕ is given by

$$\begin{aligned} \phi(u) &= -c, \quad u < -c, \\ &= u, \quad -c < u < c, \\ &= c, \quad u > c. \end{aligned} \quad (2)$$

The bounded nature of the function leads to resistance of the estimate to single observations.

The estimating equations (1) may be expressed in the form of the equations defining weighted least squares estimates:

$$0 = \sum w\left(\frac{y - \mathbf{x}'\hat{\boldsymbol{\beta}}}{\sigma}\right) x(y - \mathbf{x}'\hat{\boldsymbol{\beta}}),$$

where $w(u) = (1 - u^2)^2$. This is equivalent to (1) with

$$\phi(u) = uw(u). \quad (3)$$

The weights are often called the bi-square weights. The Princeton robustness study [1] presents properties of a large variety of M-estimates for the location model.

The M-estimates described above involve the parameter σ. Estimates are produced by replacing this parameter with an estimate of scale. The median of the absolute residuals,

$$MAD = \text{median}(y - \mathbf{x}'\hat{\boldsymbol{\beta}}) \quad (4)$$

may be rescaled to yield a robust estimate of scale. The regression estimates may then be computed by iteration, successively estimating scale by *MAD* and regression parameters by weighted least squares. Beginning with an initial estimate of the regression parameters, the scale may be estimated. The regression estimates are then updated using weighted least squares with weights given by (3).

Alternatively, a defining estimating equation for σ may be added to (1) and the larger system of equations solved iteratively.

The solution is unique if the system of equations corresponds to the derivatives of a convex function. In other cases, the value of the estimate will depend on the initial values iteration. Huber [4] contains a more detailed discussion of iteration and convergence.

Downweighting Leverage Points

Observations with unusual values of the independent variables may remain highly influential for the M-estimate. For this reason, many authors have proposed the use of iterated weighted least squares, where the weights include a factor that diminishes the influence of such observations. Many of these

proposals are based on the least squares estimate of the variance of residuals: $1 - h_{ii}$. Huber [4] suggests replacing ϕ in (1) with $(1 - h_{ii})^{1/2}\phi$. Observations with high leverage and therefore with small residual variance will have little influence in the equation defining the estimate. Krasker & Welsch [5] review more extensively the handling of leverage points.

R and S Estimators

Two other classes of robust estimators have been proposed. Hettmansperger [2] develops R-estimates based on the **ranks** of **residuals**. Estimates are defined to minimize $\sum a(R_i)r_i$, where r_i denotes a residual and R_i its rank. The function a is bounded and monotonic.

Because the defining equation is linear rather than quadratic in the residuals (as in the case of least squares), the estimates are more resistant and robust.

Least squares estimates may also be considered as those which minimize s^2, an estimate of scale. Any estimate of scale may be used to define an estimate of the regression parameters. Such estimates are called S-estimates [11, 12].

Rousseeuw [10] proposed defining estimates to minimize the *MAD* (4). Although the computational problems involved in this minimization are extremely difficult, the resulting estimate is very resistant.

Robust Logistic Regression

Logistic regression is used to model binomial data. Since such data are bounded, insensitivity to gross outliers is not a concern. However, even in this case, individual observations can greatly influence the maximum likelihood estimates.

The maximum likelihood estimates are found by minimizing the sum of deviances, essentially negative components of the log-**likelihood**. These deviance components are the logistic analogs of squared residuals. Pregibon [7] proposed estimates found by dampening the contribution of large deviances to the minimization. The proposed form corresponds to the function ϕ in (2). The resulting estimates are not unbiased. Kunsch et al. [6] proposed alternative estimators which are conditionally unbiased. Morgenthaler [9] proposed estimates of parameters in **generalized linear models** [8], and for logistic models in particular, based on least absolute deviations of residuals.

References

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*, University of Toronto Technical Report. Princeton University Press, Princeton.

[2] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

[3] Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

[4] Huber, P.J. (1980). *Robust Statistics*. Wiley, New York.

[5] Krasker, W.S. & Welsch, R.E. (1980). Efficient bounded-influence regression estimation, *Journal of the American Statistical Association* **77**, 595–604.

[6] Kunsch, H.R., Stefanski, L.A. & Carroll, R.J. (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models, *Journal of the American Statistical Association* **84**, 460–466.

[7] Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications, *Biometrics* **38**, 485–498.

[8] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd. Ed. Chapman & Hall, London.

[9] Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models, *Biometrika* **79**, 747–754.

[10] Rousseeuw, P.J. (1984). Least median squares regression, *Journal of the American Statistical Association* **79**, 871–880.

[11] Rousseeuw, P.J. & Yohai, V. (1984). Robust regression by means of S-estimators, in *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics, Vol. 26. Springer-Verlag, New York, pp. 642–656.

[12] Yohai, V.J. (1987). High breakdown point and high efficiency robust estimates for regression, *Annals of Statistics* **15**, 642–656.

DAVID F. ANDREWS

Robustness

Robust procedures are generally considered to be statistical methods which are insensitive to small deviations from the underlying assumptions. In particular, if the optimal procedures require the assumption of normality, then the corresponding robust procedures would not be influenced by departures from normality of the form of slightly longer or shorter tails or slight skewness in the underlying distribution. Such departures from normality could result from the presence of a small proportion of **outliers** or spurious values in the observations. Robust procedures are ones such that these outliers, if they occurred, would have little effect on the analysis of the data.

Historical Development

Huber [17], in his extensive review article, points out that in 1821 **Gauss** [9] specifically introduced the **normal distribution** to suit the sample **mean**, and Huber suggests that a misunderstanding of the Gauss–Markov theorem and the **central limit theorem** by contemporary nineteenth-century researchers, led to the almost exclusive use of the arithmetic mean in practical applications throughout that time (*see* **Least Squares**). These theorems refer to the arithmetic mean as the best linear unbiased estimator of the expected value of a population, and to its distribution being approximately normal. If the observations are independent with a common normal distribution, then the sample mean is the least squares (**maximum likelihood**) estimator, the best unbiased estimator (*see* **Unbiasedness**), the **minimax** estimator, and is asymptotically efficient (*see* **Efficiency and Efficient Estimators**), so it is best in a variety of ways. However, the sample mean is not robust against quite small departures from the assumption of normality; in particular, being seriously affected by outliers or long-tailedness. The occurrence of discordant values or of distributions with longer tails than normal was recognized by researchers in the nineteenth century, but the full significance of the effect of these on the behavior of the sample mean as an estimator was not fully appreciated, except by a few.

Huber's historical review gives details of a few instances of anxieties caused by the problems of dealing with outlying values and long-tailed distributions. He cites, for example, an early case of the routine application of a 5% trimmed mean (*see* **Trimming and Winsorization**) for estimating land yields in France from 20 consecutive years of observations with the highest and lowest values removed, and refers to the development of procedures for detecting and rejecting grossly discordant observations by Peirce [24] and Chauvenet [4]. More details of the heated controversy that Peirce's criterion for rejecting observations generated are given by Stigler [26], who also describes many overlooked contributions to the development of robust estimators between 1885 and 1920. In particular, Stigler discusses Newcomb's [22] use of subjectively weighted astronomical observations and long-tailed distributions, produced from mixtures of normal distributions, to derive a Bayes-type estimator (*see* **Bayesian Methods**) which effectively gives lower weight to the extreme observations. The computational efforts involved in evaluating some of these alternative estimators was a major deterrent to their general acceptance. **Edgeworth** [7] investigated properties of the **median** and later, in 1893, he suggested an estimator based on weighted quartiles. Eddington [6] and **Jeffreys** [19], amongst others, proposed alternative probability models for errors with long-tailed distributions and used robust alternatives to the usual estimators, remarkably similar to some of the procedures rediscovered more recently. Stigler [27] gives an account of an original suggestion, by Smith [25], of a robust estimator of location which is surprisingly similar to a biweighted M-estimator.

The effects of nonnormality on estimates of **variance** and the sensitivity of tests involving sample variances, such as F tests and t tests, were noted by E. S. **Pearson** and by R.A. **Fisher**, but there was little formal development of robust methods during that time. The word "robust" was initially used by Box [3] as a technical term in connection with the comparison of variances problem.

It was not until around 1960 that J.W. Tukey and the Statistical Research Group at Princeton began to investigate seriously the properties of robust estimators of location and scale. Tukey [28, 29] provided a survey of this work. He dramatically illustrated the sensitivity of the mean square deviation (*see* **Standard Deviation**)

$$s_n = \left[\frac{1}{n}\sum (x_i - \overline{x})^2\right]^{1/2},$$

based on a sample of n observations $x_i, i = 1, \ldots, n$, with sample mean $\overline{x} = \sum x_i/n$, to very slight departures from normality. He considered the **asymptotic relative efficiency (ARE)** of this estimator of scale, relative to the **mean** (absolute) **deviation**

$$d_n = \frac{1}{n} \sum |x_i - \overline{x}|,$$

for samples from a contaminated normal distribution with distribution function

$$F(x) = (1 - \varepsilon)\Phi\left(\frac{x - \mu}{\sigma}\right) + \varepsilon\Phi\left(\frac{x - \mu}{3\sigma}\right), \quad (1)$$

where $\Phi(x)$ is the standard normal cumulative distribution and ε is the proportion of contamination of the standard normal by the normal with three times its standard deviation. When ε equals zero (or one), the relative efficiency of d_n to s_n is nearly 88%, so that there is a 12% loss in efficiency if d_n is used instead of s_n with normal samples. However, ε need be only 0.002 for the relative efficiency of d_n to s_n to exceed 100%. Only two observations in 1000 need to come from the wider distribution for the mean absolute deviation to be a better estimate of scale than the mean square deviation. This was quite a surprising result and does not even imply that these two observations are outliers. They need only to have come from the contaminating distribution which produces a population that has slightly longer tails than the normal distribution. This work prompted further investigation into alternative robust estimators for both location and scale, with important early contributions from Huber [16] and Hampel [10].

Criteria of Robustness

One of the major problems in the application of a robust procedure is the consideration of an appropriate set of criteria that the procedure should satisfy. Different criteria have led to the development of different robust methods. The Princeton Robustness Study investigated the behavior of 68 estimators of location over a wide variety of nonnormal distributions. The six authors [1] agreed that many of these alternative estimators were more robust than the sample mean, but they were not able to recommend a specific estimator which would meet the various requirements of robustness, because they did not agree on the criteria to be used.

Huber [17], in his Wald lecture, asks "What is a robust procedure?" and points out that there are several conflicting aims which make it difficult to choose, in a rational manner, between different robust competitors. The assumptions, such as normality, underlying any procedure are required to allow calculation of certain probabilities such as the type I error for tests or the **confidence** levels for estimates, or to show that the method is efficient or has high **power**. These probabilities will change under departures from the assumptions. Procedures which maintain the type I error or the actual confidence levels are regarded as *validity robust* while those that maintain high power or size of confidence interval are regarded as *efficiency robust*. Even within these categories it is possible to consider different kinds of departure from the underlying assumptions and different characteristics of the estimators or test statistics involved. The characteristics used for judging the competing methods could include the asymptotic variance, the absolute efficiency or the relative efficiency. If the distribution function is represented by F, then the range of alternative distributions could include all smooth F, or all F belonging to a selected finite set of F_i such as the normal, **Cauchy**, two-sided **exponential** and rectangular distributions, or all F in the neighborhood of a specific F. Huber [16] considers procedures with a small asymptotic variance over distributions in a neighborhood of the normal distribution, while Hampel [10, 11] considers estimates whose distribution changes little under arbitrary small variations of F. The various objectives of robustness have resulted in the development of a variety of robust estimation methods broadly classified as M-estimators (based on maximum likelihood methods), L-estimators (based on linear functions of order statistics) and R-estimators (based on ranking methods).

M-Estimators

If $f(x - \theta)$ is the density function of a **random variable** x with unknown location parameter θ, then the log **likelihood** function, $l(\theta)$, for a sample $x_i, i = 1, \ldots, n$, is

$$l(\theta) = \sum_{i=1}^{n} \ln f(x_i - \theta) = -\sum_{i=1}^{n} \rho(x_i - \theta), \quad (2)$$

where $\rho(x) = -\ln f(x)$.

If possible, the maximum likelihood estimator of θ is found by differentiation of $l(\theta)$, which gives

$$\frac{\mathrm{d}[l(\theta)]}{\mathrm{d}\theta} = -\sum_{i=1}^{n} \frac{f'(x_i - \theta)}{f(x_i - \theta)} = \sum_{i=1}^{n} \varphi(x_i - \theta), \quad (3)$$

where $\rho'(x) = \varphi(x)$. The solution of

$$\sum_{i=1}^{n} \varphi(x_i - \theta) = 0 \quad (4)$$

that maximizes $l(\theta)$ is called the maximum likelihood estimator, or M-estimator, of θ. The form of the M-estimator depends on the shape of the function ρ, or equivalently the shape of the function φ. For the normal distribution $\rho(x) = x^2/2$ (apart from a constant), from which $\varphi(x) = x$, and the solution to (4) is the sample mean $\overline{x}$. For the two-sided exponential distribution, which has longer tails than the normal distribution, $\varphi(x) = -1$ for $x < 0$ and $\varphi(x) = 1$ for $x > 0$, leading to the median as the M-estimator. Other distributions with long tails have φ functions which are bounded or which "descend" to zero for $|x| > k$, for some k. Huber's [16] robust estimator, proposed to minimize the asymptotic variance over a class of distributions in the neighborhood of the normal, is one of these M-estimators with $\varphi(x) = x$ for $|x| \leq k$, and $\varphi(x) = k \text{ sign}(x)$ for $|x| > k$. Eq. (4) would in this case need to be solved by an iterative method. One feature of this and other estimators based on bounded or redescending φ functions, is that they are not scale invariant. A scale invariant version may be obtained if

$$\sum_{i=1}^{n} \varphi\left[\frac{(x_i - \theta)}{d}\right] = 0 \quad (5)$$

is solved instead, where d is a robust estimate of scale. One possible robust estimate of scale which could be used for d is

$$d = \frac{\text{median}|x_i - \text{median}(x_i)|}{0.6745}. \quad (6)$$

This form of robust scale estimate is known as the mean absolute deviation (from the median), and includes the factor 0.6745 so that d approaches the population standard deviation for large samples. The sample standard deviation is not used for d since it is not robust to outliers. The value of k may be determined so that the asymptotic efficiency of the estimator reaches a satisfactory level under normal assumptions. For instance, when $k = 1.5$ the asymptotic efficiency was shown by Huber [16] to be greater than 95%.

Other forms of M estimators have been proposed, including Hampel's redescending φ function, Andrews' sine wave and Tukey's "biweight". These are respectively defined as follows:

1. Hampel's redescending φ:

$$\begin{aligned} &\varphi(x) = && \text{if} \quad -k_1 < x < k_1, \\ &\varphi(x) = k_1 \text{ sign}(x) && \text{if} \quad k_1 \leq |x| < k_2, \\ &\text{and} \\ &\varphi(x) = k_1 \left(\frac{k_3 - |x|}{k_3 - k_2}\right) \text{sign}(x) \\ &&& \text{if} \quad k_2 \leq |x| < k_3, \end{aligned}$$

   with $k_1 = 1.7$, $k_2 = 3.4$, and $k_3 = 8.5$.
2. Andrews' sine wave: $\varphi(x) = \sin(x/k)$ if $|x| \leq k\pi$, and $\varphi(x) = 0$ if $|x| > k\pi$, with $k = 1.5$.
3. Tukey's biweight: $\varphi(x) = x[1 - (x/k)^2]^2$ if $|x| \leq k$, and $\varphi(x) = 0$ if $|x| > k$, with $k = 5$.

These values of the constants give reasonable performance of the estimators when the distribution is normal, but alternative values may be used instead. Further comments about the convergence properties of the iterative procedures involved in deriving these M-estimators are given in the review article by Hogg [14]. In his article, Hogg also discusses the extension of this approach, using M-estimators based on robust ρ and φ functions, to the estimation of the coefficients in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. This leads to the concept of **robust regression**.

L-Estimators

An L-estimator is one that is based on a linear combination of the ordered sample values. Examples of such estimators are the median and the α-symmetrically trimmed means, defined as

$$\overline{x}_{\alpha\mathrm{T}} = \frac{\sum_{i=r+1}^{n-r} x_{(i)}}{n - 2r}, \quad (7)$$

where $r = [n\alpha]$ is the largest integer less than or equal to $n\alpha$ and $x_{(i)}$ is the ith ordered observation. A symmetrically Winsorized mean is obtained in a

similar way except that, instead of the r smallest and largest observations being deleted, they are replaced by the values of the smallest and largest untrimmed observations, $x_{(r+1)}$ and $x_{(n-r)}$ respectively, so that the α-symmetrically Winsorized means are defined as

$$\overline{x}_{\alpha\mathrm{W}} = \left(r x_{(r+1)} + \sum_{i=r+1}^{n-r} x_{(i)} + r x_{(n-r)} \right) \Big/ n. \quad (8)$$

For samples from a symmetric population, the symmetrically **trimmed and Winsorized** means are both unbiased estimators of the population mean. For estimating from asymmetrical distributions, it would be natural to define asymmetrically trimmed or Winsorized means in an obvious manner.

This class of robust L-estimators also includes those based on selected percentiles, such as Gastwirth's [8] weighted average of the $33\frac{1}{3}$rd, 50th (median) and $66\frac{2}{3}$rd percentiles, with weights 0.3, 0.4, and 0.3, and the trimean using the 25th, 50th, and 75th percentiles with weights 0.25, 0.5, and 0.25. Patel et al. [23] considered that the "trimmed means, the tri-mean and Gastwirth's estimator are perhaps the simplest among reasonably good robust estimators" of location. David [5] provides a detailed review of linear order-statistic estimators (*see* **Order Statistics**). The link between these estimators and M-estimators through the influence function (*see* **Diagnostics**) is discussed by Hampel [12].

R-Estimators

The development of nonparametric procedures such as the Wilcoxon and Mann–Whitney tests (*see* **Wilcoxon–Mann–Whitney Test**), which are suitable under more general conditions than the corresponding t tests, provided some respite for those worried about nonnormality. For normal populations the one- and two-sample Wilcoxon tests have efficiency of more than 95% compared with the t tests, and can be substantially more powerful if the samples come from long-tailed distributions. Highly efficient Hodges & Lehmann [13] estimates may be developed from the Wilcoxon tests. These are the median of all pairwise averages, med $[(x_i + x_j)/2]$, for the location of a single population, and the median of all between-sample pairwise differences, med $(y_i - x_j)$, for the difference between the locations of the two populations. These estimates may be used to provide suitable confidence intervals in fairly general situations. An extensive introduction to rank statistics is given by Lehmann [20], and further discussion of R-estimates may be found in Huber [18].

Adaptive Procedures

The basic idea of adaptive estimation is that the estimation procedure is selected after observing the data. For example, the form of the estimator is dictated by the sample values themselves or some characteristic of the sample such as **skewness** or **kurtosis**, so that samples with large kurtosis (heavy tailed) could use the median as a location estimator, while those with kurtosis near zero (normal) would use the sample mean. More generally, an adaptive trimmed mean uses some characteristic of the sample, such as a ratio of linear functions of order statistics, to determine the proportion of trimming applied to the ordered sample. Hogg & Lenth [15] give a detailed review of full and partial adaptive procedures used in estimating location and include illustrations involving the extension of these procedures to **regression** analyses and **analysis of variance**.

Resistant Procedures

A statistical procedure is called *resistant* if the estimate or test statistic has a value which is insensitive to small changes in the underlying sample [21]. The underlying distribution does not really come into this definition, but Hampel's theorem, linking continuity of estimators and robustness in a neighborhood of the underlying distribution, suggests that although conceptually different, resistance and robustness are, for practical purposes, the same.

Extension to Other Problems

Over the last 30 years there has been extensive research into robust methods applicable to a variety of statistical problems. The ideas of robust estimation of location using M-, L-, and R-estimators, have been extended to multivariate data and to **robust regression**. There is now an extensive literature on robust estimation in the presence of **outliers**,

robust regression **diagnostics**, robust estimation of **correlation matrices** for use in multivariate techniques, robustness in scientific modeling, in **time series** modeling and in **experimental design**. Barnett & Lewis [2] provide a comprehensive coverage of all aspects of dealing with outliers in statistical data.

References

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972). *Robust Estimation of Location: Survey and Advances*. Princeton University Press, Princeton.

[2] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Ed. Wiley, Chichester.

[3] Box, G.E.P. (1953). Non-normality and tests on variances, *Biometrika* **40**, 318–335.

[4] Chauvenet, W. (1863). *Manual of Spherical and Practical Astronomy*. Philadelphia.

[5] David, H.A. (1981). *Order Statistics*, 2nd Ed. Wiley, New York.

[6] Eddington, A.S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.

[7] Edgeworth, F.Y. (1886). Problems in probabilities, *Philosophical Magazine* **22**, Series 5, 371–384.

[8] Gastwirth, J.L. (1960). On robust procedures, *Journal of the American Statistical Association* **61**, 929–948.

[9] Gauss, C.F. (1821). *Göttingische gelehrte Anzeigen*, pp. 321–327 (reprinted in *Werke* **4**, 98).

[10] Hampel, F.R. (1968). Contributions to the Theory of Robust Estimation, *Ph.D. dissertation*. University of California, Berkeley.

[11] Hampel, F.R. (1971). A general qualitative definition of robustness, *Annals of Mathematical Statistics* **42**, 1887–1896.

[12] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**, 383–393.

[13] Hodges, J.L. & Lehmann, E.L. (1963). Estimates of location based on rank tests, *Annals of Mathematical Statistics* **34**, 598–611.

[14] Hogg, R.V. (1979). Statistical robustness: one view of its use in applications today, *American Statistician* **33**, 108–115.

[15] Hogg, R.V. & Lenth, R.V. (1984). A review of some adaptive statistical techniques, *Communications in Statistics – Theory and Methods* **13**, 1551–1579.

[16] Huber, P.J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**, 73–101.

[17] Huber, P.J. (1972). Robust statistics: a review, *Annals of Mathematical Statistics* **43**, 1041–1067.

[18] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

[19] Jeffreys, H. (1932). An alternative to the rejection of outliers, *Proceedings of the Royal Society, Series A* **137**, 78–87.

[20] Lehmann, E.L. (1975). *Nonparametric Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

[21] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Linear Regression*. Addison-Wesley, Reading, Mass.

[22] Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result, *American Journal of Mathematics* **8**, 343–366.

[23] Patel, K.R., Mudholkar, G.S. & Fernando, J.L.I. (1988). Student's t approximations for three simple robust estimators, *Journal of the American Statistical Association* **83**, 1203–1210.

[24] Peirce, B. (1852). Criterion for the rejection of doubtful observations, *Astronomical Journal* **2**, 161–163.

[25] Smith, R.H. (1888). True average of observations? *Nature*, March 15, p. 464.

[26] Stigler, S.M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920, *Journal of the American Statistical Association* **68**, 872–879.

[27] Stigler, S.M. (1980). Studies in the history of probability and statistics XXXVIII. R.H. Smith, a Victorian interested in robustness, *Biometrika* **67**, 217–221.

[28] Tukey, J.W. (1960). A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics*, I. Olkin, ed. Stanford University Press, Stanford, pp. 448–485.

[29] Tukey, J.W. (1962). The future of data analysis, *Annals of Mathematical Statistics* **33**, 1–67.

PHILIP PRESCOTT

Rotation of Axes

Both **principal components analysis** and **factor analysis** are procedures which transform a set of p correlated variables into a set of k new variables. In the case of principal components analysis, the new variables, the principal components, are uncorrelated (*see* **Correlation**). In the case of factor analysis, the new variables, the factors, are uncorrelated within a reduced space defined by the reduction in total variability due to residual variability associated with each variable, independent of the others. For both procedures, these transformations are generally linear. In most cases, $k < p$, since a principal aim of both procedures is to reduce the dimensionality required to describe a multivariate situation. The coefficients defining these transformations, often referred to as *loadings*, are used to aid in identifying the nature of the new variables. Sometimes, the new variables will be difficult to interpret. At other times, the procedure has been employed chiefly as an intermediate step in determining a reduced *set* of the original variables which satisfactorily account for the variability of the deleted variables. In either case, a second linear transformation may be useful. This latter transformation is generally referred to as a *rotation* of the components or factors.

If there are k components or factors to be rotated, then there will be k new rotated variables. The variables produced by rotation may be correlated. They will account for the same amount of variability of the original variables as the components or factors from which they were derived. Most texts on principal components and/or factor analysis will have some material on rotation. Among those with detailed explanation of the philosophy and general mechanics of rotation are [3] and [4]. Other articles in this Encyclopedia deal extensively with principal components analysis, factor analysis, and specific rotation methods.

Simple Structure

Most of the methods described in this article are designed to attain the properties of what Thurstone [7] called **simple structure**. Its purpose is to produce rotated vectors whose coefficients are either relatively large or close to zero. If the matrix of dimension $p \times k$ of rotated vectors is $\mathbf{B}$, each column defining a transformed variable, then simple structure requires that:

1. each row of $\mathbf{B}$ should contain at least one zero. This means that each of the original variables should be uncorrelated with at least one of the rotated components or factors
2. if there are k components or factors, then each column of $\mathbf{B}$ should have at least k zeros. This specifies a goal; for interpretation, the more zeros the better
3. for each pair of columns of $\mathbf{B}$, there should be several variables that have zeros in one column but not the other and, if $k \geq 4$, a large number of variables with zeros in both columns and a smaller number of variables with nonzero coefficients in both columns. This is an attempt to obtain some independence among the variables produced by the rotated vectors.

The objective of simple structure is to produce a set of new vectors, each involving primarily a subset of the original variables with as little overlap as possible so that the original variables are divided into groups somewhat independent of each other. This is, in essence, a method of clustering the original variables (*see* **Cluster Analysis, Variables**), and some computer packages (*see* **Software, Biostatistical**) employ this method to do it. Most statistical computer packages that include principal components and factor analysis will also have a number of rotation options.

In practice, it is nearly impossible to obtain the number of zeros required for a simple structure, but rotated vectors containing very small coefficients will suffice for most problems of interpretation.

Algebraically, rotations may be described as follows: Let the transformation of the original p variables, $\mathbf{x}$, into k components or factors, $\mathbf{y}$, be $\mathbf{y} = \mathbf{V}\mathbf{x}$, where $\mathbf{V}$ is a matrix of dimension $p \times k$ and consists of a set of vectors relating one set of variables, $\mathbf{x}$, to the other, $\mathbf{y}$ (*see* **Principal Components Analysis** and **Factor Analysis, Overview** for a discussion of the mathematical models underlying these transformations.) Then the rotation of the vectors $\mathbf{V}$ into a new set $\mathbf{B}$ is done by the relationship $\mathbf{B} = \mathbf{V}\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a matrix of dimension $k \times k$ consisting of angles defining the rotation. This rotation takes place only in the subspace defined by the k retained components or

factors. Because of this, the amount of variability of the original variables accounted for by the variables obtained by this rotation will be exactly the same as that accounted for by the original components or factors. If the number of retained components or factors is changed, then the rotated results will also change. Early rotation procedures were graphical solutions which, while quite sophisticated, could handle only problems of limited size. Present day solutions are generally performed by means of computerized mathematical **algorithms**.

Numerical Example

Table 1 displays a correlation matrix related to some physical measurements on 305 girls [4]. Note that the first set of four variables – measurements of "lankiness" – are highly correlated, as are the second set, representing "stockiness". The intercorrelations among the two sets are smaller. The characteristic vectors defining the first two principal components are displayed in Table 2, where the vectors are normalized to their corresponding characteristic roots. The first principal component, accounting for 58% of the total variability, is a measure of overall size. The second component (22%) represents a contrast between the two types of measurement.

Orthogonal Rotation

An **orthogonal rotation** is a rotation which preserves the **orthogonality** of the transformed component or factor. For the above example, one possible orthogonal rotation has the matrix defining the angle of rotation as

$$\Theta = \begin{bmatrix} 0.771 & 0.636 \\ -0.636 & 0.771 \end{bmatrix}.$$

This particular orthogonal rotation is a **varimax rotation**. The rotated vectors are shown in Table 2. The first four coefficients of the first vector, representing the lankiness measurements, are quite large relative to the remaining four. The situation for the second vector is reversed. The conclusion is that there are two groups of variables: the first four and the second four. While this seems obvious when rotating a pair of vectors, examples requiring higher dimensionality are apt to be more difficult to interpret, particularly if one wishes to cluster variables.

The rotated axes are shown in Figure 1 and are represented by the dashed lines. The points represent the original eight variables in terms of the loadings of the two characteristic vectors. If these same points are projected against the new axes, then the coefficients or loadings of **B** will result. This is an *orthogonal rotation*, in that the new axes are still at right angles to each other.

There are a number of methods for producing orthogonal rotations, each having their strengths and weaknesses. A more detailed description of some of them appears in the article **Orthogonal Rotation**, as well as in separate entries for some specific methods.

Oblique Rotation

It is possible to obtain an even greater differentiation between the two sets of physical measurements by performing an **oblique rotation**, one in which the resultant axes are not at right angles to each other. One such oblique rotation is the **orthoblique** or Harris–Kaiser rotation, which is also shown in Table 2.

This produces greater differences between the two sets of coefficients but at the cost of loss of orthogonality. The amount of variability of the original variables attributed to these rotated variables will be the same as that due to the components or factors. The improvement towards the simple structure is shown in Figure 2. There are now *two* sets of rotated vectors. Those labeled P_1 and P_2 are called *primary* vectors, which pass through the clusters of points. R_1 and R_2 are *reference* vectors, R_i and P_i being orthogonal to each other. It is the projection of the points on the reference vectors that produce the rotated vectors **B**. In orthogonal rotations, this situation does not exist. Note, in Table 2, that the amount of variability attributable to the two orthogonally rotated components equals the amount attributable to the principal components, but those of the oblique rotation do not. These latter quantities must also include the joint contribution of the rotated components pairwise; these, added to that of the rotated components themselves, will add to the total.

There are a number of methods for producing oblique rotations which will appear in the article **Oblique Rotations**, as well as separate entries for some specific methods. A number of both

Table 1 Physical measurements

	Correlation matrix							
Height	1	0.85	0.80	0.86	0.47	0.40	0.30	0.38
Arm span	0.85	1	0.88	0.83	0.38	0.33	0.28	0.42
Length of forearm	0.80	0.88	1	0.80	0.38	0.32	0.24	0.34
Length of lower leg	0.86	0.83	0.80	1	0.44	0.33	0.33	0.36
Weight	0.47	0.38	0.38	0.44	1	0.76	0.73	0.63
Bitrochanteric diameter	0.40	0.33	0.32	0.33	0.76	1	0.58	0.58
Chest girth	0.30	0.28	0.24	0.33	0.73	0.58	1	0.54
Chest width	0.38	0.42	0.34	0.36	0.63	0.58	0.54	1

Reproduced from [4] by permission of the University of Chicago Press.

Table 2 Physical measurements: characteristic and rotated vectors

	Characteristic vectors		Orthogonal rotation		Oblique rotation	
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_1$	$\mathbf{b}_2$
Height	0.86	−0.37	0.90	0.26	0.91	0.06
Arm span	0.84	−0.44	0.93	0.20	0.96	−0.02
Length of forearm	0.81	−0.46	0.92	0.16	0.96	−0.06
Length of lower leg	0.84	−0.40	0.90	0.23	0.92	0.02
Weight	0.76	0.52	0.25	0.89	0.05	0.90
Bitrochanteric diameter	0.67	0.53	0.18	0.84	−0.02	0.87
Chest girth	0.62	0.58	0.11	0.84	−0.10	0.89
Chest width	0.67	0.42	0.25	0.75	0.08	0.75
Var. explained	4.67	1.77	3.50	2.95	2.84	2.35

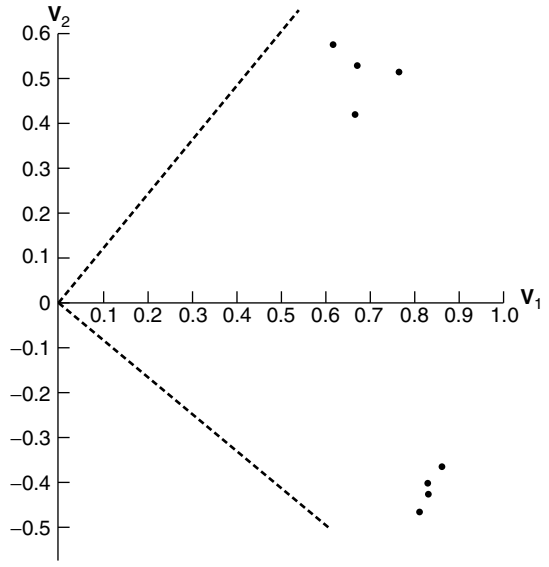

Figure 1 Physical measurements: orthogonal (varimax) rotation

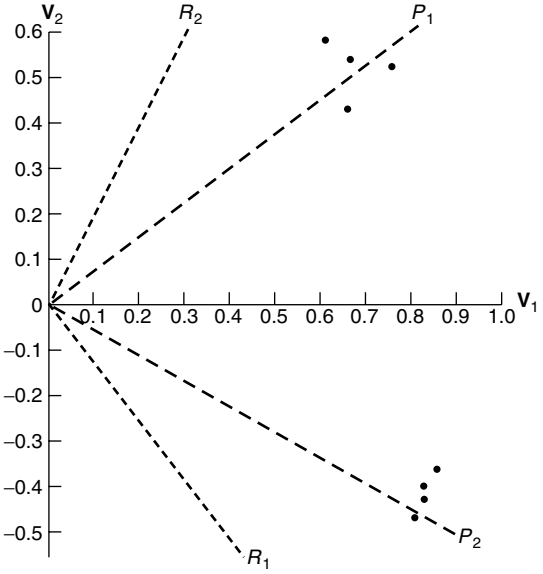

Figure 2 Physical measurements: oblique (orthoblique) rotation

orthogonal and oblique rotations can be expressed in a generalized form known as *quartic* rotation procedures. An overview of these procedures may be found in [2].

Procrustes Rotation

The occasion may arise where one already has two sets of vectors and wishes to find the matrix which will best "rotate" one set into the other. The method which attempts to determine this relationship is called **Procrustes rotation**, and is described in that article.

Applicability

Rotational procedures may be quite useful when one wishes to cluster variables, and as such have found widespread use in such diverse fields as psychology, education, anthropology, biology, and market research. These procedures have somewhat less application in sciences such as chemistry or physics. The physical measurements example of Table 1 is a typical application in the field of multivariate biostatistics.

Numerical Examples

The numerical example above is useful in producing a graphical explanation of rotation but will not be as useful in illustrating the various rotation techniques because only two vectors are being rotated. In the case of $k = 2$, many of these techniques produce the same results. We need $k > 2$ to show the difference. For this reason, two additional examples will be given. One of these will be an example where rotation was appropriate and the other where it was not. These examples will also be employed in some of the separate articles in this Encyclopedia dealing with specific procedures.

The first example deals with Decathlon data for 160 individual records from the first eight Olympics after World War II [6]. Table 3 includes both the characteristic vectors **(eigenvectors)** associated with the four retained principal components of these data and the corresponding varimax rotation. (The correlation matrix is included in the original reference and also in [1].) The other common rotation techniques all showed similar results.

The first rotated vector has high loadings for 100 m run and 400 m run along with somewhat lower loadings for long jump (which requires a short run before execution) and the 110 m hurdles. This would imply a cluster involving short distance running and associated jumping. The second rotated vector has high loadings for shot put, discus, and javelin, all of which are throwing events. The third rotated vector has high loadings for the long jump, high jump, 110 m hurdles, and the pole vault, all events requiring jumping. The final rotated vector has a high loading for the 1500 m run and a lower loading for the 400 m run, indicating a cluster for long distance running. The 1500 m run had a correlation of 0.39 with the 400 m run and very low correlations with anything else. In interpreting rotation results, reference to the original correlation matrix will also be of use. Although most of the results obtained by rotation probably could have been deduced from the characteristic vectors, the rotated results may seem much clearer.

The second example deals with the audiometric examinations of 100 39-year-old men [5]. Table 4 includes both the characteristic vectors associated with the four retained principal components of these data and the corresponding varimax rotation. The variables represent hearing loss from a standard at four different frequencies for both left and right ears.

The first principal component represents an overall shift in hearing for all frequencies in both ears. The second component represents a contrast between high and low frequencies and can be used as an early warning of hearing loss. The third component is another contrast, primarily between the two higher frequencies, and the fourth component represents the difference between left and right ears. The variability unexplained by these four components is a measure of testing and measurement variability. The results from the varimax rotation are not as distinct as those obtained in the other example. One could probably deduce that the 500 Hz and 1000 Hz measurements formed one cluster, the 2000 Hz a second, and the 4000 Hz a third. There is a suggestion of ear differences also, but the results are not very distinct. In this case, the characteristic vectors furnish all the

Table 3 Decathlon data: characteristic and rotated vectors

	Characteristic vectors				Varimax rotation			
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
100 m run	0.69	0.22	−0.52	−0.21	0.88	0.14	0.16	−0.12
Long jump	0.79	0.18	−0.19	0.09	0.63	0.19	0.52	−0.01
Shot put	0.70	−0.53	0.05	−0.18	0.24	0.82	0.22	−0.15
High jump	0.67	0.13	0.14	0.40	0.24	0.15	0.75	0.08
400 m run	0.62	0.55	−0.08	−0.42	0.80	0.07	0.10	0.47
110 m hurdle	0.69	0.04	−0.16	0.35	0.40	0.15	0.64	−0.17
Discus	0.62	−0.52	0.11	−0.23	0.19	0.81	0.15	−0.08
Pole vault	0.54	0.09	0.41	0.44	−0.04	0.18	0.76	0.22
Javelin	0.43	−0.44	0.37	−0.24	−0.05	0.74	0.11	0.14
1500 m run	0.15	0.60	0.66	−0.28	0.05	−0.04	0.11	0.93

Reproduced from [6] by permission of *Research Quarterly for Exercise and Sports*.

Table 4 Audiometric example: characteristic and rotated vectors

Frequency	Characteristic vectors				Varimax rotation			
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
500 Hz left	0.80	−0.40	0.16	−0.22	0.58	0.13	0.06	0.71
1000 Hz left	0.83	−0.29	−0.05	−0.33	0.44	0.10	0.27	0.79
2000 Hz left	0.73	0.30	−0.46	−0.19	0.05	0.22	0.78	0.45
4000 Hz left	0.56	0.60	0.42	−0.11	0.04	0.89	0.15	0.20
500 Hz right	0.68	−0.49	0.26	0.33	0.91	0.08	−0.00	0.23
1000 Hz right	0.82	−0.29	−0.03	0.25	0.77	0.10	0.34	0.31
2000 Hz right	0.62	0.40	−0.56	0.27	0.17	0.19	0.93	−0.00
4000 Hz right	0.50	0.65	0.42	0.11	0.11	0.91	0.19	−0.02

Reproduced from [5] by permission of the American Society for Quality Control.

information needed, and no rotation is needed. The frequencies chosen for the audiometric procedure are from a continuum of frequencies that could have been chosen, and, when the original variables are of this form, rotation is less likely to be useful than the decathlon example where the variables are specific events. All of the common rotation procedures had similar problems with these data except for **quartimax**. That exception is dealt with in the article on quartimax.

References

[1] Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*. Wiley, New York.

[2] Clarkson, D.B. & Jennrich, R.J. (1988). Quartic rotation criteria and algorithms, *Psychometrika* **53**, 251–259.

[3] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[4] Harmon, H.H. (1976). *Modern Factor Analysis*, 3rd Ed. University of Chicago Press, Chicago.

[5] Jackson, J.E. (1991). *A User's Guide to Principal Components*. Wiley, New York.

[6] Linden, M. (1977). Factor analytical study of Olympic Decathlon Data, *Research Quarterly* **48**, 562–568.

[7] Thurstone, L.L. (1947). *Modern Factor Analysis*. University of Chicago Press, Chicago.

(*See also* **Axes in Multivariate Analysis**)

J. EDWARD JACKSON

Roy's Maximum Root Criteria

To illustrate his **union–intersection principle**, Roy [37, 38] proposed the maximum characteristic root **eigenvalue** test statistic for each of the following problems: (i) testing the equality of k p-variate normal distributions (*see* **Multivariate Normal Distribution**) with the same but unknown **covariance matrix**; (ii) testing the independence between two sets of variates jointly distributed as a normal distribution with unknown mean vector; (iii) testing the equality of covariance matrices of two p-variate normal distributions with unknown mean vectors; and (iv) testing that the covariance matrix of a p-variate normal distribution with unknown mean vector equals a specified matrix $\mathbf{\Sigma}_0$.

For problems (i)–(iii), the test statistic can be expressed as the largest characteristic root b_1 of a random matrix $\mathbf{B} \equiv \mathbf{S}_1(\mathbf{S}_1 + \mathbf{S}_2)^{-1}$. For problem (i), which can be considered as a version of **multivariate analysis of variance** (MANOVA), the matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the sums of squares and cross-products matrices "due to hypothesis" and "due to error" with **degrees of freedom** (df) ν_1 and ν_2, respectively. For problem (ii), $\mathbf{S}_1 = \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$, $\mathbf{S}_2 = \mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{22}$, where $(\mathbf{S}_{ij}; i, j = 1, 2)$ is the partitioned sums of squares and cross-products (SP) matrix corresponding to the two sets. Lastly, $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the SP matrices corresponding to random samples from the two populations for problem (iii). Furthermore, Roy also considered the largest characteristic root t_1 of the matrix $\mathbf{S}$ as a test statistic for problem (iv), $\mathbf{S}/\nu_1$ being the sample covariance matrix with df ν_1.

The null distribution of b_1 takes the same form for each of the problems (i)–(iii). Moreover, the limiting distribution of $\nu_2 b_1$ is the same as the distribution of t_1 defined for (iv) with $\mathbf{\Sigma}_0 = \mathbf{I}$, as $\nu_2 \to \infty$. We shall assume that the number s of nonzero roots of $\mathbf{B}$ equals p, in the sequel; for $s < p$, see [22] for an appropriate change of parameters for (i).

The literature on this topic is rather extensive. We cite only the major references which give information on many other related works. In a pioneering paper, Roy [36] considered the maximum root statistic for problem (i) and derived its null distribution with an explicit expression for $s = 2$, 3, and 4. Later, the null distribution of b_1 has been derived in a series form following the reduction method of Roy, the Pfaffian method of Mehta [21], and through zonal polynomials introduced by Constantine [8] for the distribution of roots. See [1, 18, 19, 28], and [40] for more details.

To make the null distribution of b_1 amenable to evaluation of upper percentage points, Pillai [24] has suggested an approximation. Using this approximation as well as using the exact distribution, tables have been constructed for the upper percentage points of b_1 for various values of the parameters (see [19, 25, 26, 31], and [44].

The nonnull distribution of b_1 in the multivariate analysis of variance problem has been obtained by Khatri & Pillai [16], DeWaal [11], Pillai & Sugiyama [35], Krishnaiah & Chang [20], and Khatri [15], among others. The noncentral distribution of the largest **canonical correlation** coefficient has been derived by Pillai & Sugiyama [41], DeWaal [11], Pillai [27], and Khatri [15], in particular. For problem (iii), the nonnull distribution of b_1 has been obtained by Khatri [14], Pillai & Sugiyama [35], and Chang [5] in the general case. For this problem, Chang [5] has obtained a beta-type asymptotic expansion of the distribution of b_1. For asymptotic distribution of the maximum root for problems (i) and (ii) see [1] and [22]. Also see the review paper by Pillai [29].

8The distribution of the largest characteristic root of a **Wishart** matrix follows from the result in Constantine [8]; it has been obtained also by Sugiyama [41], Krishnaiah & Chang [20], and Khatri [15]; see [22] for its asymptotic distribution. Tables for the upper percentage points of t_1 are given in Clemm et al. [7] and Krishnaiah [19]. The distribution of the maximum root of a noncentral Wishart matrix has been obtained by Hayakawa [13]. For an approximation to the cumulative distribution function of the largest root of the covariance matrix, see [32].

The maximum root test for MANOVA has been shown to be admissible (*see* **Decision Theory**) by Ghosh [12] and Anderson & Takemura [4]. The monotonicity of the **power** function of Roy's maximum root test in terms of the corresponding noncentrality parameters has been shown by Das Gupta et al. [10] for MANOVA, by Anderson & Das Gupta [2] for the test of independence, and by Anderson & Das Gupta [3] for problems (iii) and (iv) with one-sided alternatives.

It has been observed that Roy's maximum root test for (i) and (ii) has relatively (in comparison with

other standard tests) lower power for local alternatives and nonlinear alternatives; see [33, 34], and [39]. Based on a **Monte Carlo** study, Olson [23] has observed that Roy's test for MANOVA is (relatively) most affected by deviations from normality and homoscedasticity; see Korin [17] for a similar study. For testing the equality of covariance matrices, Chu & Pillai [6] have observed that Roy's two-sided test based on the maximum root performed best locally among all standard two-sided tests. Pillai & Hsu [34] have studied robustness of the test of independence based on Roy's maximum root criterion along with three other criteria.

For **simultaneous confidence intervals** based on the maximum root, see Roy [38] and Srivastava & Khatri [40]; properties of such confidence intervals have been studied by Wijsman [42, 43]. The relative efficiency of these confidence intervals has been studied by Cox et al. [9].

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] Anderson, T.W. & Das Gupta, S. (1964). Monotonicity of power functions of some tests of independence between two sets of variates, *Annals of Mathematical Statistics* **35**, 206–208.

[3] Anderson, T.W. & Das Gupta, S. (1964). Monotonicity property of the power functions of some tests of the equality of two covariance matrices, *Annals of Mathematical Statistics* **35**, 1059–1063.

[4] Anderson, T.W. & Takemura, A. (1982). A new proof of admissibility of tests in multivariate analysis, *Journal of Multivariate Analysis* **12**, 457–468.

[5] Chang, T.C. (1970). On asymptotic representation of the distribution of the characteristic roots of $S_1S_2^{-1}$, *Annals of Mathematical Statistics* **41**, 440–444.

[6] Chu, S.S. & Pillai, K.C.S. (1979). Power comparisons of two-sided tests of equality of covariance matrices based on six criteria, *Annals of the Institute of Statistical Mathematics* **31**, 185–200.

[7] Clemm, D.S., Krishnaiah, P.R. & Waikar, V.B. (1973). Tables for the extreme roots of the Wishart matrix, *Journal of Statistical Computation and Simulation* **2**, 65–92.

[8] Constantine, A.G. (1963). Some noncentral distribution problems in multivariate analysis, *Annals of Mathematical Statistics* **34**, 1270–1285.

[9] Cox, C.M., Krishnaiah, P.R., Lee, J.C., Reising, J. & Schuurmann, F.J. (1980). A study on finite intersection tests for multiple comparison of means, in *Multivariate Analysis*, Vol. V, P.R. Krishnaiah, ed. North-Holland, New York, pp. 435–466.

[10] Das Gupta, S., Anderson, T.W. & Mudholkar, G.S. (1964). Monotonicity of the power functions of some tests of the multivariate linear hypothesis, *Annals of Mathematical Statistics* **35**, 200–205.

[11] DeWaal, D.J. (1969). On the noncentral distribution of the largest canonical correlation coefficients, *South Africa Statistical Journal* **3**, 91–93.

[12] Ghosh, M.N. (1964). On the admissibility of some tests of MANOVA, *Annals of Mathematical Statistics* **35**, 789–794.

[13] Hayakawa, T. (1969). On the distribution of the latent roots of a positive definite symmetric matrix, *Annals of the Institute of Statistical Mathematics* **21**, 1–21.

[14] Khatri, C.G. (1967). Some distribution problems connected with the characteristic roots of $S_1S_2^{-1}$, *Annals of Mathematical Statistics* **38**, 944–948.

[15] Khatri, C.G. (1972). On the exact finite series distribution of the smallest or largest root of matrices in three situations, *Journal of Multivariate Analysis* **2**, 201–207.

[16] Khatri, C.G. & Pillai, K.C.S. (1968). On the noncentral distributions of two test criteria in multivariate analysis of variance, *Annals of Mathematical Statistics* **39**, 215–226.

[17] Korin, B.P. (1972). Some comments on the homoscedasticity criterion and the multivariate analysis of variance tests T^2, W and R, *Biometrika* **59**, 215–216.

[18] Krishnaiah, P.R. (1978). Some recent developments in real multivariate analysis, in *Development in Statistics*, Vol. I, P.R. Krishnaiah, ed. Academic Press, New York, pp. 135–169.

[19] Krishnaiah, P.R. (1980). Computations of some multivariate distributions. *Handbook of Statistics*, Vol. I, P.R. Krishnaiah, ed. North-Holland, New York.

[20] Krishnaiah, P.R. & Chang, T.C. (1971). On the exact distributions of the extreme roots of the Wishart and MANOVA matrices, *Journal of Multivariate Analysis* **1**, 108–117.

[21] Mehta, M.L. (1967). *Random Matrices and the Statistical Theory of Energy Levels*. Academic Press, New York.

[22] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

[23] Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* **69**, 894–908.

[24] Pillai, K.C.S. (1956). On the distribution of the largest or smallest root of a matrix in multivariate analysis, *Biometrika* **43**, 122–127.

[25] Pillai, K.C.S. (1957). *Concise Statistical Tables*. The Statistical Center, University of Philippines.

[26] Pillai, K.C.S. (1960). *Statistical Tables for Tests of Multivariate Hypothesis*. Statistical Center, University of Philippines.

[27] Pillai, K.C.S. (1970). On the noncentral distributions of the largest roots of two matrices in multivariate analysis, in *Essays in Probability and Statistics*, R.C. Bose, I.M. Chakravarti, P.C. Mahalanobis, R. Rao & K.J.C. Smith, eds. University of North Carolina Press, Chapel Hill.

[28] Pillai, K.C.S. (1976). Distributions of characteristic roots in multivariate analysis. Part I: Null distributions, *Canadian Journal of Statistics* **4**, 154–184.

[29] Pillai, K.C.S. (1977). Distributions of characteristic roots in multivariate analysis. Part II: Non-null distributions, *Canadian Journal of Statistics* **5**, 1–62.

[30] Pillai, K.C.S. & Chang, T.C. (1970). An approximation to the cdf of the largest root of the covariance matrix, *Annals of the Institute of Statistical Mathematics* **6**, 115–124.

[31] Pillai, K.C.S. & Flury, B.N. (1984). Percentage points of the largest characteristic roots of the multivariate beta matrix, *Communications in Statistics – Theory and Methods* **13**, 2199–2237.

[32] Pillai, K.C.S. & Hsu, T. (1979). Exact robustness studies of the test of independence based on four multivariate criteria and their distribution problems, *Annals of the Institute of Statistical Mathematics* **31**, 85–101.

[33] Pillai, K.C.S. & Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypothesis based on four criteria, *Biometrika* **44**, 195–210.

[34] Pillai, K.C.S. & Sudjana (1975). Exact robustness studies of tests of two multivariate hypothesis based on four criteria and their distribution problems under violations, *Annals of Statistics* **3**, 617–638.

[35] Pillai, K.C.S. & Sugiyama, T. (1969). Noncentral distributions of the largest latent roots of three matrices in multivariate analysis, *Annals of the Institute of Statistical Mathematics* **21**, 321–327.

[36] Roy, S.N. (1945). The individual sampling distribution of the maximum, the minimum and any intermediate of the p-statistics on the null hypothesis, *Sankhyā*, **7**, 133–158.

[37] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[38] Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

[39] Schatzoff, M. (1966). Sensitivity comparisons among tests of the general linear hypothesis, *Journal of the American Statistical Association* **61**, 415–435.

[40] Srivastava, M.S. & Khatri, C.G. (1979). *An Introduction to Multivariate Statistics*. North-Holland, New York.

[41] Sugiyama, T. (1967). On the distribution of the largest latent root of the covariance matrix, *Annals of Mathematical Statistics* **38**, 1148–1151.

[42] Wijsman, R.A. (1979). Constructing all simultaneous confidence sets in a given class, with applications to MANOVA, *Annals of Statistics* **7**, 1003–1018.

[43] Wijsman, R.A. (1980). Smallest simultaneous confidence sets with applications in multivariate analysis, in *Multivariate Analysis*, Vol. V, P.R. Krishnaiah, ed. North-Holland, New York.

[44] Yamamuti, Z. (1977). *Concise Statistical Tables*. Japanese Standards Association.

(*See also* **Lambda Criterion, Wilks'**; **Lawley–Hotelling Trace**; **Multivariate Analysis, Overview**; **Pillai's Trace Test**)

SOMESH DASGUPTA

Royal Statistical Society

The Statistical Society of London was established in 1834, with the stated purpose of procuring, arranging, and publishing "Facts calculated to illustrate the Condition and Prospects of Society". Four main classes of study were specified in the founding prospectus: economical statistics, political statistics, medical statistics, and moral and intellectual statistics. Although the term "medical" was then primarily applied to issues of public health, the Royal Statistical Society (as it was to become in 1887), is still strongly associated with what is now known as **"biostatistics"**. This article has a biostatistical perspective; for more general historical articles, see [2, 7, 13, 14, 16].

Early volumes of the *Journal of the Statistical Society* (*see* ***Journal of The Royal Statistical Society***) are full of worthy statistical data on national and international trade and economics, the empire, transport, mortality, and social investigations. The latter strongly reflect the Victorian concern, if not obsession, with insanity, crime, and the ills of the lower classes. Typical examples include: Report upon the Mortality of Lunatics (1841), Sanitary Condition of Borough of Reading (1847), Rate of Mortality among Persons of Intemperate Habits (1851), Duration of Life among the Clergy (1851), Statistics of the Insane, Blind, Deaf and Dumb, and Lepers, of Norway (1852), and Vital and Medical Statistics of Chittagong (1862). Sadly, the content of such papers does not always live up to the interest aroused by their titles; there is a general dullness of presentation, a lack of graphics, and an absence of incisive interpretation. The latter merely puts into practice the strictures expressed in the original prospectus in 1834: "The Statistical Society will consider it to be the first and most essential rule of its conduct to exclude carefully all opinions from its transactions and publications, – to confine its attention rigorously to facts, – and, so far as it may be found possible, to facts which can be stated numerically and arranged in tables."

An exception to this self-imposed role is Florence **Nightingale**, who returned in 1857 from her revolutionary work in the Crimean War, determined to use all possible statistical tools as polemical weapons in her crusade to reform military and civil hospitals. She enlisted the help of William **Farr**, who was then the dominant force in the statistical analysis of public health data; he contributed many articles to the *Journal* and became President of the Society in 1871. Florence Nightingale was elected as the first woman Fellow of the Society in 1858, and embarked in 1859, on a campaign for uniform hospital and surgical statistics, which would "enable us to ascertain the relative mortality of different hospitals. . .". In 1860, the International Statistical Congress in London decided that "Miss Nightingale's Scheme for Uniform Hospital Statistics should be conveyed to all governments represented". The *Journal of the Statistical Society* published summaries of these statistics for London and provincial hospitals for five years, as a series of tables without comment. These were summarized in 1867 by **Guy** [6], who specifically denied that any variability in outcome could be due to **quality of care** given by the staff "chosen, as it is, from among those members of the profession who have already given proofs of sound training, ability, and skill in practice".

The Society, which became the Royal Statistical Society with the award of the Royal Charter in 1887, was slow to embrace the dramatic methodological and applied developments in statistics that began near the end of the nineteenth century; only **Edgeworth** and **Yule** seriously contributed to the methodological content of the *Journal*, beginning with a classic exposition in the 1885 Jubilee issue [4]. Here, Edgeworth discusses the variability of the **mean** of a set of observations, the **central limit theorem**, and introduces the term "insignificant", while Yule [17] displays a remarkable use of modern statistical method: model fitting (*see* **Model, Choice of**), estimates of error, tests of **goodness of fit**, and interpretation of a large data set. There were also mathematical papers by Karl **Pearson** and **Galton**, but these were largely summaries of longer papers given elsewhere; Karl Pearson was never a Fellow of the Society, and the new Biometric school primarily relied on its own new journal ***Biometrika*** as an outlet. The hugely influential work of **Fisher** carried out at Rothamsted Experimental Station was also not reflected in the *Journal* until the 1930s.

The Society began publishing the discussion of its read papers in 1873, but it was not until the 1930s that a notorious series of papers established its continuing reputation for public statistical disputes. **Neyman** presented his development of **confidence intervals** to the Society on June 19, 1934, in which he sought to make **inferences** about parameters without having

to use a **prior distribution** [10]. Arthur Bowley, a strong advocate of **Bayesian methods**, or "inverse probability" as it was then known, was given the task of proposing the vote of thanks. He started by saying "I am not at all sure that the 'confidence' is not a 'confidence trick"', and added "Does that really take us any further? Does it take us beyond Karl Pearson and Edgeworth? Does it really lead us towards what we need – the chance that in the universe which we are sampling the proportion is within certain limits? I think it does not." He finished with "The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity."

R.A. Fisher, however, was generous in his praise of the paper, interpreting it as direct support for his own "**fiducial** theory". This good-natured alliance against the inverse probabilists was soon to break down. Later that year, Fisher was strongly attacked when he presented his work on **likelihood** [5], and in 1935, he rounded on Neyman: ". . .were it not for the persistent efforts which Dr Neyman and Dr [E.S.] **Pearson** had made to treat what they speak of as problems of estimation, by means merely of tests of significance, he had no doubt that Dr Neyman would not have been in any danger of falling into the series of misunderstandings which his paper revealed" [11]. Published continuations of this personal dispute were still appearing 20 years later, and it might be still claimed to be reflected in arguments concerning the focus of regulatory bodies on ***P* values**.

The war focused attention on industrial statistics, but in the 1950s, the Society began to turn its attention to modern biostatistics. Austin Bradford **Hill** became President in 1950, Sir Ronald Fisher in 1952, and in October 1955, the Study Circle on Medical Statistics became a Section. The computer was enthusiastically embraced: Bartlett [1] records early work on "the Manchester computer" in simulating epidemics (*see* **Epidemic Models, Stochastic**), and in the discussion, Norman Bailey generously states that "provided they are not made an excuse for avoiding difficult mathematics, I think there is great scope for such computers in biometrical work". Hollingsworth [8] is the first published example of **"computer-aided diagnosis"**.

The year 1972 turned out to be somewhat of an *annus mirabilis* for methods applicable in biostatistics: Nelder & Wedderburn [9] described the basic algebraic and computational framework for **generalized linear models**, Peto & Peto [12] established the theoretic framework for the **logrank test**, and Cox [3] introduced the **proportional hazards** model with an arbitrary underlying **hazard rate** function, and hence made possible the introduction of multiple **covariates** into **survival analysis**. Consulting the Science Citation Index in April 1997, we find Nelder & Wedderburn [9] had over 600 citations and Peto & Peto [12] over 1100, but these are paltry compared with the 6000 citations of Cox [3].

Biostatistical activity in the Society has continued to grow in line with the increasing importance of the subject in medical research. The 1960 regulations state that the Medical Section may "arrange periodical meetings or conferences of the Section for the reading of papers, discussion or demonstrations. . .. The Section Committee may recommend that any papers read before the Section shall be published by the Society." This has led to special meetings and collected articles on such topics as HIV/**AIDS** (1988), cancer near nuclear installations (1989) (*see* **Leukemia Clusters**), institutional **"league tables"** (1996), **BSE/CJD** (1997) and ethics, integrity, and clinical trials (2002) (*see* **Medical Ethics and Statistics**).

Guy Medals are awarded in gold (since 1892), silver (since 1893), and bronze (since 1936). In 1999, the Guy Gold medal was awarded to Michael Healy for his extensive statistical contributions to agriculture, medicine, and a wide variety of other applications in which he has had a significant influence. Discussion papers cited in awards of the Guy Silver Medal often feature biostatistical themes; they include examples on repeated significance tests (*see* **Sequential Analysis**), multivariate proportional hazards (*see* **Multivariate Survival Analysis**), **DNA sequencing**, Bayesian methods in the **pharmaceutical industry**, medical expert systems (*see* **Artificial Intelligence**), and so on. The Bradford Hill Medal for medical statistics was inaugurated in 1994, with a posthumous award to Martin **Gardner**.

The Institute of Statisticians was for many years the professional body for statisticians in the United Kingdom, but after many years of attempted negotiations, the Society finally merged with the Institute on January 1, 1993. The increased emphasis on professional matters led to the introduction of Chartered Statistician (CStat) status, and the formation of a Professional Affairs Committee. There is a strong relationship with the pharmaceutical industry: the Society was instrumental in bringing full-time

statisticians into the United Kingdom drug regulatory framework [15] (*see* **Drug Approval and Regulation**), and there is continued dialog with statisticians in the pharmaceutical industry (*see* **Statisticians in the Pharmaceutical Industry (PSI)**).

The merger with the Institute of Statisticians raised membership of the Society to over 6000. Sections currently include Medical, Research, General Applications, Social Statistics, Business and Industrial, Quality Improvement, Official Statistics, and Statistical Computing, and there are study groups in Environmental Statistics and Primary Health Care. General conferences are held every two years, alternating with specialist meetings on topics such as Practical Bayesian Statistics and **teaching statistics**. In 2003, the theme of the Society's conference was statistical genetics and **bioinformatics** and in 2005, the Society and PSI will be organizing a joint conference in statistics and health care.

In March 2004, *Significance* was launched as the Society's new quarterly magazine for anyone interested in statistics and the analysis and interpretation of data. Its aim is to communicate and demonstrate in an entertaining and thought-provoking way, the practical use of statistics in all walks of life and to show how statistics benefit society. Articles are largely nontechnical and hence accessible and appealing, not only to members of the profession, but also to all users of statistics. It is intended that the magazine should be relevant to people working, for instance, in central and local government, medicine and health care, administration, economics, business and commerce, industry, social studies, survey research, science, and the environment.

After many years without a base in which general meetings could be held, in 1995, the Society finally moved into excellent premises at 12 Errol Street in London. In the last 10 years, the Society has increasingly sought to raise its own profile and that of statistics. The Society is committed to a policy of outreach – disseminating and promoting the use and understanding of statistics to advance the welfare of society. In line with this policy, it has, for instance, taken a close interest in statistical education at all levels and has established a Centre for Statistical Education, currently based at Nottingham Trent University. The Society has produced a report on the use of performance indicators and is now working with stakeholders in this area to develop good practice not only in target setting, but also in the design, analysis, and reporting of performance indicators. It is similarly working with stakeholders within the legal and associated professions to ensure the appropriate collection and use of forensic statistical evidence (*see* **Medico–Legal Cases and Statistics**; **Statistical Forensics**). Since 1990, the Society has argued that an independent statistical service, free from political interference, is essential to the maintenance of a healthy democracy and has paid close attention to the UK Government's measures to implement National Statistics. Comprehensive information about the Society, its purposes and activities, can be found at `www.rss.org.uk`.

With a thriving Medical Section, the Royal Statistical Society continues to keep its traditional balance between biostatistical methodology and practice. It now stands in a very strong position to further the role of statistics in public life in general, and in the health field in particular.

References

[1] Bartlett, M. (1957). Measles periodicity and community size (with discussion), *Journal of the Royal Statistical Society, Series A* **120**, 48–70.

[2] Bonar, J. & Macrosty, H.W. (1934). *Annals of the Royal Statistical Society*. Royal Statistical Society, London.

[3] Cox, D.R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] Edgeworth, F.Y. (1885). Methods of statistics, in *Jubilee Volume. Journal of the Statistical Society*, Royal Statistical Society, London, pp. 181–217.

[5] Fisher, R.A. (1935). The logic of inductive inference, *Journal of the Royal Statistical Society* **98**, 39–82.

[6] Guy, W.A. (1867). On the mortality of London hospitals: and incidentally on the deaths in the prisons and public institutions of the metropolis, *Journal of the Statistical Society* **30**, 293–322.

[7] Hill, I.D. (1984). Statistical society of London–royal statistical society: the first 100 years: 1834–1934, *Journal of the Royal Statistical Society, Series A* **147**, 130–139.

[8] Hollingsworth, T.H. (1959). Using an electronic computer in a problem of medical diagnosis, *Journal of the Royal Statistical Society, Series A* **122**, 221–332.

[9] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

[10] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society* **97**, 558–625.

[11] Neyman, J. (1935). Statistical problems in agricultural experimentation, *Journal of the Royal Statistical Society* **98**, 107–180.

[12] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series A* **135**, 185–207.

[13] Plackett, R.L. (1984). Royal statistical society: the last 50 years: 1934–1984, *Journal of the Royal Statistical Society, Series A* **147**, 140–150.

[14] Rosenbaum, S. (1984). The growth of the royal statistical society, *Journal of the Royal Statistical Society, Series A* **147**, 375–388.

[15] RSS Working Party. (1991). Statistics and statisticians in drug regulation in the United Kingdom, *Journal of the Royal Statistical Society, Series A* **154**, 413–419.

[16] Tippett, L.H.C. (1972). Annals of the royal statistical society 1934–1971, *Journal of the Royal Statistical Society, Series A* **135**, 545–568.

[17] Yule, G.U. (1896). Notes on the history of pauperism in England and Wales from 1850, treated by the method of frequency-curves; with an introduction on the method, *Journal of the Royal Statistical Society* **59**, 318–349.

DAVID J. SPIEGELHALTER & IVOR J. GODDARD

Saddlepoint Approximation

Saddlepoint approximations are a class of asymptotic approximations to a density or tail probability of a statistic. McCullagh [5], Jensen [8], and Kolassa [6] discuss these methods in detail. Suppose that $T = \sum_{j=1}^{n} Y_j/n$, for Y_j independent and identically distributed. Suppose that the **cumulant generating function** $\mathcal{K}(\beta) = \log \mathrm{E}\left[\exp(\beta Y_j)\right]$ exists for β in an open set about 0. Let $f_T(t)$ represent the density of T. Then $f_T(t)\exp([t\beta - \mathcal{K}(\beta)])$ represents an **exponential family** for T at a potential value t. Note that $\mathrm{E}_\beta[T] = \mathcal{K}'(\beta)$ and $\mathrm{Var}_\beta[T] = \mathcal{K}''(\beta)$. Let $\hat{\beta}$ solve $t = \mathrm{E}_{\hat{\beta}}[T]$, or

$$\mathcal{K}'(\hat{\beta}) = t. \tag{1}$$

Applying the Gaussian approximation (*see* **Normal Distribution**) to $f_T(t)\exp(n[t\hat{\beta} - \mathcal{K}(\hat{\beta})])$, one obtains

$$f_T(t)\exp(n[\hat{\beta}t - \mathcal{K}(\hat{\beta})]) = \frac{(\exp(0)/\sqrt{2\pi})}{\sqrt{\mathcal{K}''(\hat{\beta})}}\left(1 + O\left(\frac{1}{n}\right)\right). \tag{2}$$

The notation $O(1/n)$ indicates a quantity that, when multiplied by n, is bounded as $n \to \infty$, and represents a relative error; that is, it reflects the difference between the density approximation and the true density, divided by the true density. Daniels [2] showed that

$$f_T(t) = \frac{\sqrt{n}\exp(n[\mathcal{K}(\hat{\beta}) - \hat{\beta}t])}{\sqrt{2\pi\mathcal{K}''(\hat{\beta})}}\left(1 + O\left(\frac{1}{n}\right)\right). \tag{3}$$

Approximation (3) is preferred to direct application of the **central limit theorem**, because the bound on the relative error is uniform for t in compact regions in the range of $\mathcal{K}'$; a uniform bound on the relative errors of more direct Gaussian-based approximations generally exist only for $t/\sqrt{n}$ bounded. The other approximations given below share this relative error behavior. Approximations discussed in this article are called saddlepoint approximations, since they might also be derived using complex integration techniques, in which the integrand near $\hat{\beta}$ is shaped like a saddle; $\hat{\beta}$ is known as the saddlepoint.

When $\boldsymbol{T} = (T^1, \ldots, T^d)$ is the mean of independent and identically distributed random vectors, each with d components, the above logic motivates the approximation

$$f_{\boldsymbol{T}}(\mathbf{t}) = \frac{n^{d/2}\exp(n[\mathcal{K}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}\mathbf{t}])}{\left[(2\pi)^{d/2}\sqrt{|\mathcal{K}''(\hat{\boldsymbol{\beta}})|}\right]}\left(1 + O\left(\frac{1}{n}\right)\right), \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ satisfies

$$\mathcal{K}'(\hat{\boldsymbol{\beta}}) = \mathbf{t}. \tag{5}$$

Statistical applications typically require approximate tail probabilities rather than approximate densities, since tail probabilities may be used to construct **hypothesis tests**, and hypothesis tests may be inverted to construct confidence intervals; see [3]. When $d = 1$, one might integrate (3) to approximate $\mathrm{P}[T \geq t]$. Typically, one cannot perform these integrations exactly. One might integrate by parts, and demonstrate that an omitted term is sufficiently small. Let $\hat{w} = \sqrt{(2[\hat{\beta}t - \mathcal{K}(\hat{\beta})])}\,\mathrm{sgn}(\hat{\beta})$. Lugannani and Rice [7] show that

$$\mathrm{P}[T \geq t] = 1 - \Phi(\sqrt{n}\hat{w}) + \frac{\phi(\sqrt{n}\hat{w})(1/\hat{w} - 1/\hat{z})}{\sqrt{n}}\left(1 + O\left(\frac{1}{n}\right)\right), \tag{6}$$

where $\hat{z} = \hat{\beta}\sqrt{(K''(\hat{\beta}))}$. [1] derives the r^* form of the approximation

$$\mathrm{P}[T \geq t] = 1 - \Phi(\sqrt{n}\hat{w}^*)\left(1 + O\left(\frac{1}{n}\right)\right) \text{ for } \hat{w}^* = \hat{w} + (n\hat{w})^{-1}\log\left(\frac{\hat{z}}{\hat{w}}\right). \tag{7}$$

Conditional probability densities may be approximated by calculating (4) for both the joint and the marginal densities (*see* **Marginal Probability**), and dividing the results. Applying this technique to the distribution of the last component T^d, conditional on all other components $\boldsymbol{T}_{-d} = (T^1, \ldots, T^{d-1})$, yields

$$f_{T^d|\boldsymbol{T}_{-d}}(t^d|\mathbf{t}_{-d}) = \frac{\sqrt{n}\phi(\sqrt{n}\hat{w})}{\sqrt{|\mathcal{K}''(\hat{\boldsymbol{\beta}})|/|\mathcal{K}''_{-d}(\tilde{\boldsymbol{\beta}})|}} \times \left(1 + O\left(\frac{1}{n}\right)\right). \tag{8}$$

Here $\hat{\boldsymbol{\beta}}$ solves (4) and $\tilde{\boldsymbol{\beta}}$ solves $\tilde{\beta}_d = 0$ and $\mathcal{K}'_{-d}(\tilde{\boldsymbol{\beta}}) = \mathbf{t}_{-d}$, $\mathcal{K}''_{-d}$ is the matrix $\mathcal{K}''$, omitting the row and column corresponding to d, $\mathcal{K}'_{-d}$ represents $\mathcal{K}'$ with component d omitted, and $\hat{w} = \sqrt{(2[\mathcal{K}(\tilde{\boldsymbol{\beta}}) - \mathcal{K}(\hat{\boldsymbol{\beta}}) + [\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}]\mathbf{t}])}$. Again, $\mathrm{P}\left[T^d \geq t^d | \boldsymbol{T}_{-d} = \mathbf{t}_{-d}\right]$ may again be approximated by (6) or (7), for $\hat{z} = \hat{\beta}_d \sqrt{(|\mathcal{K}''(\hat{\boldsymbol{\beta}})|/|\mathcal{K}''_{-d}(\tilde{\boldsymbol{\beta}})|)}$. Skovgaard [9] derived this approximation, using (6). Approximations (6) and (7) are called double saddlepoint approximations, to distinguish them from an application of the univariate saddlepoint distribution function approximation based on the cumulant generating function of the conditional distribution, which might be called a single saddlepoint conditional probability approximation. The single saddlepoint approach is useful when the conditional cumulant generating function is available, or when, like [4], one employs an approximation to the conditional cumulant generating function.

The preceding development is for continuous **random variables**. If possible values of components of $\boldsymbol{T}$ are separated by a constant, say $1/n$, then probability masses for $\boldsymbol{T}$ are approximated by (3) or (4). When $d = 1$, one might add (3) over the tail region to show that (6) and (7) still hold, with $\hat{z} = 2n \sinh(\hat{\beta}/(2n))\sqrt{(\mathcal{K}''(\hat{\beta}))}$, and with $\hat{\beta}$ satisfying $\mathcal{K}'(\hat{\beta}) = t - 1/(2n)$. For $d > 1$, when the possible values of T^d are separated by $1/n$, (6) and (7) hold as approximations to $\mathrm{P}\left[T^d \geq t^d | \boldsymbol{T}_{-d} = t_{-d}\right]$, where $\hat{z} = 2n \sinh(\hat{\beta}_d/(2n))\sqrt{(|\mathcal{K}''(\hat{\boldsymbol{\beta}})|/|\mathcal{K}''_{-d}(\tilde{\boldsymbol{\beta}})|)}$, and t_d is corrected for **continuity** before applying (5).

References

[1] Barndorff–Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio, *Biometrika* **73**(2), 307–322.

[2] Daniels, H.E. (1954). Saddlepoint approximations in statistics, *Annals of Mathematical Statistics* **25**(0), 614–649.

[3] Davison, A.C. (1988). Approximate conditional inference in generalized linear models, *Journal of the Royal Statistical Society Series B* **50**(3), 445–461.

[4] Fraser, D.A.S., Reid, N. & Wong, A. (1991). Exponential linear models: a two–pass procedure for saddlepoint approximation, *Journal of the Royal Statistical Society Series B* **53**(2), 483–492.

[5] Jensen, J.L. (1995). *Saddlepoint Approximations*. Oxford Science Publications, Oxford.

[6] Kolassa, J.E. (1997). *Series Approximation Methods in Statistics*, 2nd Ed. Springer-Verlag, New York.

[7] Lugannani, R. & Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables, *Advances in Applied Probability* **12**, 475–490.

[8] McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall, London.

[9] Skovgaard, I.M. (1987). Saddlepoint expansions for conditional distributions, *Journal of Applied Probability* **24**, 875–887.

JOHN E. KOLASSA

Salk Vaccine

The largest and, until the 1980s, the most expensive medical experiment in history was carried out in 1954. Well over a million young children participated, and the immediate direct costs were over 5 million mid-century dollars. The experiment was carried out to assess the effectiveness, if any, of the Salk vaccine as a protection against paralysis or death from poliomyelitis. The study was elaborate in many respects, most prominently in the use of placebo controls (children who were inoculated with simple salt solution) assigned at random (that is, by a carefully applied chance process that gave each volunteer an equal probability of getting vaccine or salt solution) (*see* **Randomization**) and subjected to a double-blind evaluation (that is, an arrangement under which neither the children nor the physicians who evaluated their subsequent state of health knew who had been given the vaccine and who the salt solution; *see* **Blinding or Masking**).

Why was such elaboration necessary? Did it really result in more or better knowledge than could have been obtained from much simpler studies? These are the questions on which this discussion is focused.

Background

Polio was never a common disease, but it certainly was one of the most frightening and, in many ways, one of the most inexplicable in its behavior. It struck hardest at young children, and, although it was responsible for only about 6% of the deaths in the age group 5 to 9 in the early 1950s, it left many helpless cripples, including some who could survive only on a respirator. It appeared in epidemic waves, leading to summer seasons in which some communities felt compelled to close swimming pools and restrict public gatherings as cases increased markedly from week to week; other communities, escaping an epidemic one year, waited in trepidation for the year in which their turn would come.

The determination to mount a major research effort to eradicate polio arose in no small part from the involvement of President Franklin D. Roosevelt, who was struck down by polio when a successful young politician. His determination to overcome his paralytic handicap enabled a great deal of attention, effort, and money to be expended on the care and rehabilitation of polio victims and – in the end, more importantly – on research into the causes and prevention of the disease.

During the course of this research, it was discovered that polio is caused by a virus. Although clinical manifestations of polio are rare, it was discovered that the virus itself was not rare, but common, and that most adults had experienced a polio infection sometime in their lives without ever being aware of it. This finding helped to explain the otherwise peculiar circumstance that polio epidemics seemed to hit hardest those who were better off hygienically (that is, those who had the best nutrition, most favorable housing conditions, and were otherwise apparently most favorably situated). Indeed, the disease seemed to be virtually unknown in those countries with the poorest hygiene. The explanation is that because there was plenty of polio virus in the less-favored populations, almost every infant was exposed to the disease early in life while still protected by the immunity passed on from the mother. As a result, everyone had polio, but under protected circumstances, and, thereby, everyone had developed immunity.

As with many other virus diseases, an individual who has been infected by polio and recovered is usually immune to another attack (at least by a virus strain of the same type). The reason for this is that the body, in order to fight the infection, develops antibodies in response to the presence of the protein part of the polio virus. Once the body has learned how to make antibodies to a particular kind of virus, it is able to make them again very rapidly, if the virus should attack a second time. This ability to make antibodies rapidly to fight against subsequent viral attacks is part of what makes people immune.

Smallpox and influenza illustrate two different approaches to the preparation of an effective vaccine. For smallpox, which has long been controlled by a vaccine, we use for the vaccine a closely related virus, cowpox, which is ordinarily incapable of causing serious disease in humans, but which gives rise to antibodies that also protect against smallpox. (In a very few individuals this vaccine is capable of causing a severe, and occasionally fatal, reaction. The risk is small enough, however, so that before smallpox was conquered we did not hesitate to expose all our school children to it in order to

protect them from smallpox.) In the case of influenza, however, instead of a closely related live virus, the vaccine is a solution of the influenza virus itself, prepared with a virus that has been killed by treatment with formaldehyde. Provided that the treatment is not too prolonged, the dead virus still has enough antigenic activity to produce the required antibodies so that, although it can no longer infect, it is sufficiently like the live virus to be a satisfactory vaccine.

For polio, both of these methods were explored. A live-virus vaccine would have the advantage of reproducing in the vaccinated individual and, hopefully, giving rise to a strong reaction that would produce a high level of long-lasting antibodies. With such a vaccine, however, there might be a risk that a vaccine virus so similar to the virulent polio virus could mutate into a virulent form and itself be the cause of paralytic or fatal disease. A killed-virus vaccine should be safe because it presumably could not infect, but it might fail to give rise to an adequate antibody response. These and other problems stood in the way of the rapid development of a successful vaccine. Some unfortunate prior experience also contributed to the cautious approach of the researchers. In the 1930s, attempts had been made to develop vaccines against polio; two of these were actually in use for a time. Evidence that at least one of these vaccines had been responsible for cases of paralytic polio soon caused both to be promptly withdrawn from use. This experience was very much in the minds of polio researchers, and they had no wish to risk a repetition.

Research to develop both live and killed vaccines was stimulated in the late 1940s by the development of a tissue culture technique for growing polio virus. Those working with live preparations developed harmless strains from virulent ones by growing them for many generations in suitable tissue culture media. There was, of course, considerable worry lest these strains, when used as a vaccine in humans, might revert to virulence and cause paralysis or death. (It is now clear that the strains developed are indeed safe – a live-virus preparation taken orally is the vaccine presently in widespread use throughout the world.) Those working with killed preparations, notably Jonas Salk, had the problem of treating the virus (with formaldehyde) sufficiently to eliminate its infectiousness, but not so long as to destroy its antigenic effect. This was more difficult than expected, and some early lots of the vaccine proved to contain live virus capable of causing paralysis and death. There are statistical issues in the safety story [1], but our concern here is with the evaluation of effectiveness.

Evaluation of Effectiveness

In the early 1950s the Advisory Committee convened by the National Foundation for Infantile Paralysis (NFIP) decided that the killed-virus vaccine developed by Jonas Salk at the University of Pittsburgh had been shown to be both safe and capable of inducing high levels of the antibody in children on whom it had been tested. This made the vaccine a promising candidate for general use, but it remained to prove that the vaccine actually would prevent polio in exposed individuals. It would be unjustified to release such a vaccine for general use without convincing proof of its effectiveness, so it was determined that a large-scale "field trial" should be undertaken (*see* **Vaccine Studies**).

That the trial had to be carried out on a very large scale is clear. For suppose we wanted the trial to be convincing if indeed the vaccine were 50% effective (for various reasons, 100% effectiveness could not be expected). Assume that, during the trial, the rate of occurrence of polio would be about 50 per 100 000 (which was about the average incidence in the US during the 1950s). With 40 000 in the **control** group and 40 000 in the vaccinated group, we would find about 20 control cases and about 10 vaccinated cases, and a difference of this magnitude could fairly easily be attributed to random variation. It would suggest that the vaccine might be effective, but it would not be persuasive. With 100 000 in each group, the expected numbers of polio cases would be 50 and 25, and such a result would be persuasive. In practice, a much larger study was clearly required because it was important to get definitive results as soon as possible, and if there were relatively few cases of polio in the test area, the expected number of cases might be well under 50. It seemed likely, also, for reasons we discuss later, that paralytic polio, rather than all polio, would be a better criterion of disease, and only about half the diagnosed cases are classified "paralytic". Thus the relatively low incidence of the disease, and its great variability from place to place and time to time, required that the trial involve a huge number of subjects – as it turned out,

over a million (*see* **Sample Size Determination for Clinical Trials**).

The Vital Statistics Approach

Many modern therapies and vaccines, including some of the most effective ones such as smallpox vaccine, were introduced because preliminary studies suggested their value. Large-scale use subsequently provided clear evidence of efficacy. A natural and simple approach to the evaluation of the Salk vaccine would have been to distribute it as widely as possible, through the schools, to see whether the rate of reported polio was appreciably less than usual during the subsequent season. Alternately, distribution might be limited to one or a few areas because limitations of supply would preclude effective coverage of the entire country. There is even a fairly good chance that were one to try out an effective vaccine against the common cold, convincing evidence might be obtained in this way.

In the case of polio – and, indeed, in most cases – so simple an approach would almost surely fail to produce clear-cut evidence. First and foremost, we must consider how much polio incidence varies from season to season, even without any attempts to modify it. From Figure 1, which shows the annual reported incidence from 1930 through 1955, we see that, had a trial been conducted in this way in 1931, the drop in incidence from 1931 to 1932 would have been strongly suggestive of a highly effective vaccine because the incidence dropped to less than a third of its previous level. Similar misinterpretation would have been made in 1935, 1937, and 1952. One might suppose that such mistakes could be avoided by using the vaccine in one area, say, New York State, and comparing the rate of incidence there with that of an unvaccinated area, say, Illinois. Unfortunately, an epidemic of polio might well occur in Chicago – as it did in 1956 – during a season in which New York had a very low incidence.

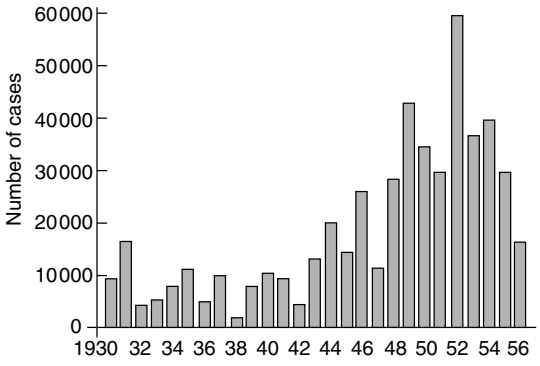

Figure 1 Poliomyelitis in the US, 1930–1956. *Source*: Meier [2]. Estimated. Figures complete through December 8

Another problem, more subtle, but equally burdensome, relates to the vagaries of diagnosis and reporting. There is no difficulty, of course, in diagnosing the classic respirator case of polio, but the overwhelming majority of cases are less clear-cut. Fever and weakness are common symptoms of many illnesses, including polio, and the distinction between weakness and slight transitory paralysis will be made differently by different observers. Thus the decision to diagnose a case as nonparalytic polio instead of some other disease might well be influenced by a physician's general knowledge or feeling about how widespread polio is in his or her community at the time.

These difficulties can be mitigated to some extent by setting down very precise criteria for diagnosis, but it is virtually impossible to obviate them completely when, as would be the case after the widespread introduction of a new vaccine, there is a marked shift in what the physician expects to find (*see* **Outcome Measures in Clinical Trials**).

The Observed Control Approach

The difficulties of the **vital statistics** approach were recognized by all concerned, and the initial study plan, although not judged entirely satisfactory, circumvented many of the problems by introducing a control group similar in characteristics to the vaccinated group. More specifically, the idea was to offer vaccination to all children in the second grade of participating schools and to follow the polio experience not only in these children but in the first- and third-grade children as well. Thus the vaccinated second graders would constitute the treated group, and the first- and third-graders would constitute the control group. This plan follows what we call the observed control approach.

It is clear that this plan avoids many of the difficulties listed above. The three grades all would be drawn from the same geographic location so that an epidemic affecting the second grade in a

given school would certainly affect the first and third grades as well. Of course, all subjects would be observed concurrently in time. The grades, naturally, would be of different ages, and polio incidence does vary with age. Not much variation from grade to grade was expected, however, so it seemed reasonable to assume that the average of first and third grades would provide a good control for the second grade.

Despite the relative attractiveness of this plan and its acceptance by the NFIP advisory committee, serious objections were raised by certain health departments that were expected to participate. In their judgment, the results of such a study were likely to be insufficiently convincing for two important reasons. One is the uncertainty in the diagnostic process mentioned earlier and its liability to be influenced by the physician's expectations, and the other is the selective effect of using volunteers.

Under the proposed study design, physicians in the study areas would have been aware of the fact that only second-graders were offered vaccine, and in making a diagnosis for any such child, they would naturally and properly have inquired whether the child had been vaccinated. Any tendency to decide a difficult diagnosis in favor of nonpolio when the child was known to have been vaccinated would have resulted in a spurious piece of evidence favoring the vaccine. Whether or not such an effect was really operating would have been almost impossible to judge with assurance, and the results, if favorable, would have been forever clouded by uncertainty (*see* **Bias in Observational Studies**).

A less conjectural difficulty lies in the difference between those families who volunteer their children for participation in such a trial and those who do not. Not at all surprisingly, it was later found that those who do volunteer tend to be better educated and, generally, more well-to-do than those who do not participate. There was also evidence that those who agree to participate tend to be absent from school with a noticeably higher frequency than others. The direction of effect of such selection on the incidence of diagnosed polio is by no means clear before the fact, and this important difference between the treated group and the control group also would have clouded the interpretation of the results (*see* **Selection Bias**).

Randomization and the Placebo Control Approach

The position of critics of the NFIP plan was that the issue of vaccine effectiveness was far too important to be studied in a manner that would leave uncertainties in the minds of reasonable observers. No doubt, if the vaccine should appear to have fairly high effectiveness, most public health officials and the general public would accept it, despite the reservations. If, however, the observed control scheme were used, a number of qualified public health scientists would have remained unconvinced, and the value of the vaccine would be uncertain. Therefore, the critics proposed that the study be run as a scientific experiment with the use of appropriate randomization procedures to assign subjects to treatment or to control and with a maximum effort to eliminate observer bias (*see* **Randomized Treatment Assignment**). This plan follows what we call the placebo control approach.

The chief objection to this plan was that parents of school children could not reasonably be expected to permit their children to participate in an experiment in which they might be getting only an ineffective salt solution instead of a probably helpful vaccine. It was argued further that the injection of placebo might not be ethically sound since a placebo injection carries a small risk, especially if the child is unknowingly already infected with polio (*see* **Ethics of Randomized Trials**).

The proponents of the placebo control approach maintained that, if properly approached, parents would consent to their children's participation in such an experiment, and they judged that because the injections would not be given during the polio season, the risk associated with the placebo injection was vanishingly small. Certain health departments took a firm stand: they would participate in the trial only if it were such a well-designed experiment. The consequence was that, in approximately half the areas, the randomized placebo control method was used, and in the remaining areas, the alternating-grade observed control method was used.

A major effort was put forth to eliminate any possibility of the placebo control results being contaminated by subtle observer **biases**. The only firm way to accomplish this was to ensure that neither the subject, nor the parents or the diagnostic personnel could know which children had gotten the

vaccine until all diagnostic decisions had been made. The method for achieving this result was to prepare placebo material that looked just like the vaccine but was without any antigenic activity, so that the controls might be inoculated and otherwise treated in just the same fashion as were the vaccinated.

Each vial of injection fluid was identified only by a code number, so that no one involved in the vaccination or the diagnostic evaluation process could know which children had gotten the vaccine. Because no one knew, no one could be influenced to diagnose differently for vaccinated cases and for controls. An experiment in which both the subject getting the treatment and the diagnosticians who will evaluate the outcome are kept in ignorance of the treatment given each individual is called a double-blind experiment. Experience in clinical research has shown the double-blind experiment to be the only satisfactory way to avoid potentially serious observer bias when the final evaluation is in part a matter of judgment.

For most of us, it is something of a shock to be told that competent and dedicated physicians must be kept in ignorance lest their judgments be colored by knowledge of treatment status. We should keep in mind that it is not deliberate distortion of findings by the physician that concern the medical experimenter. It is rather the extreme difficulty in many cases of making an uncertain decision that, experience has shown, leads the best of investigators to be subtly influenced by information of this kind. For example, in the study of drugs used to relieve postoperative pain, it has been found that it is quite impossible to get an unbiased judgment of the quality of pain relief, even from highly qualified investigators, unless the judge is kept in ignorance of which patients were given the drugs.

The second major feature of the experimental method was the assignment of subjects to treatments by a careful randomization procedure. As we observed earlier, the chance of coming down with a diagnosed case of polio varies with a great many factors, including age, socioeconomic status, etc. If we were to make a deliberate effort to match up the treatment and control groups as closely as possible, we should have to take care to balance these and many other factors, and, even so, we might miss some important ones. Therefore, perhaps surprisingly, we leave the balancing to a carefully applied equivalent of coin tossing: we arrange that each individual has an equal chance of getting vaccine or placebo, but we eliminate our own judgment entirely from the individual decision and leave the matter to chance.

The gain from doing this is twofold. First, a chance mechanism usually will do a good job of evening out all the variables – those we did not recognize in advance as well as those we did recognize. Secondly, if we use a chance mechanism in assigning treatments, we may be confident about the use of the theory of chance (that is, **probability theory**) to judge the results. We can then calculate the probability that so large a difference as that observed could reasonably be due solely to the way in which subjects were assigned to treatments, or whether, on the contrary, it is really an effect due to a true difference in treatments.

To be sure, there are situations in which a skilled experimenter can balance the groups more effectively than a random-selection procedure typically would. When some factors may have a large effect on the outcome of an experiment, it may be desirable, or even necessary, to use a more complex experimental design that takes account of these factors. However, if we intend to use probability theory to guide us in our judgment about the results, we can be confident about the accuracy of our conclusions only if we have used randomization at some appropriate level in the experimental design.

The final determinations of diagnosed polio proceeded along the following lines. All cases of polio-like illness reported by local physicians were subjected to special examination, and a report of history, symptoms, and laboratory findings was made. A special diagnostic group then evaluated each case and classified it as nonpolio, doubtful polio, or definite polio. The last group was subdivided into nonparalytic and paralytic, with paralytic further divided into nonfatal and fatal polio. Only after this process was complete was the code broken and identification made for each case as to whether vaccine or placebo had been administered.

Results of the Trial

The main results are shown in Table 1, which shows the size of the study populations, the number of cases classified as polio, and the disease rates; that is, the number of cases per 100 000 population. For example, the second line shows that in the placebo

Table 1 Summary of study cases by diagnostic class and vaccination status (rates per 100 000)

Study group	Study population	All reported cases		Poliomyelitis cases								Not polio	
				Total		Paralytic		Nonparalytic		Fatal polio			
		No.	Rate	No.	Rate	No.	Rate	No.	Rate	No.	Rate	No.	Rate
All areas: total	1 829 916	1013	55	863	47	685	37	178	10	15	1	150	8
Placebo control areas: total	749 236	428	57	358	48	270	36	88	12	4	1	70	9
Vaccinated	200 745	82	41	57	28	33	16	24	12	–	–	25	12
Placebo	201 229	162	81	142	71	115	57	27	13	4	2	20	10
Not inoculated[a]	338 778	182	54	157	46	121	36	36	11	–	–	25	7
Incomplete vaccinations	8 484	2	24	2	24	1	12	1	12	–	–	–	–
Observed control areas: total	1 080 680	585	54	505	47	415	38	90	8	11	1	80	7
Vaccinated	221 998	76	34	56	25	38	17	18	8	–	–	20	9
Controls[b]	725 173	439	61	391	54	330	46	61	8	11	2	48	6
Grade 2 not inoculated	123 605	66	53	54	44	43	35	11	9	–	–	12	10
Incomplete vaccinations	9 904	4	40	4	40	4	40	–	–	–	–	–	–

Source: Adapted from Francis [1, Tables 2 and 3].

[a]Includes 8577 children who received one or two injections of placebo.

[b]First- and third-grade total population.

control area there were 428 reported cases, of which 358 were confirmed as polio, and, among these, 270 were classified as paralytic (including four that were fatal). The third and fourth rows show corresponding entries for those who were vaccinated and those who received placebo, respectively. Beside each of these numbers is the corresponding rate. Using the simplest measure – all reported cases – the rate in the vaccinated group is seen to be half that in the control group (compare the boxed rates in Table 1) for the placebo control areas. This difference is greater than could reasonably be ascribed to chance, according to the appropriate probability calculation (*see* **Hypothesis Testing**). The apparent effectiveness of the vaccine is more marked as we move from reported to paralytic cases to fatal cases, but the numbers are small and it would be unwise to make too much of the apparent very high effectiveness in protecting against fatal cases. The main point is that the vaccine was a success; it demonstrated sufficient effectiveness in preventing serious polio to warrant its introduction as a standard public health procedure.

Not surprisingly, the observed control areas provided results that were, in general, consistent with those found in the placebo control areas. The volunteer effect discussed earlier, however, is clearly evident (note that the rates for those not inoculated differ from the rates for controls in both areas). Were the observed control information alone available, considerable doubt would have remained about the proper interpretation of the results [3].

Although there had been wide differences of opinion about the necessity or desirability of the placebo control design before, there was great satisfaction with the method after the event. The difference between the two groups, although substantial and definite, was not so large as to preclude doubts had there been no placebo controls. Indeed, there were many surprises in the detailed data. It was known, for example, that some lots of vaccine had greater antigenic power than did others, and it might be supposed that they should have shown a greater protective effect. This was not the case; lots judged inferior in potency did just as well as those judged superior. Another surprise was the rather high frequency with which apparently typical cases of paralytic polio were not confirmed by laboratory test. Nonetheless, there were no surprises of a character to cast serious doubt on the main conclusion. The favorable reaction of those most expert in research on polio was expressed soon after the results were reported. By carrying out this kind of study before introducing the vaccine, it was noted, we had facts about the Salk vaccine that we still lack about the typhoid vaccine, and about the tuberculosis vaccine after many decades of use.

Epilogue

It would be pleasant to report an unblemished record of success for the Salk vaccine following so expert and successful an appraisal of its effectiveness, but it is more realistic to recognize that such success is but one step in the continuing development of public health science. The report of the field trial was followed by widespread release of the vaccine for general use, and it was discovered very quickly that a few of these lots had actually caused serious cases of polio. Distribution of the vaccine was then halted while the process for making the vaccine was reevaluated. Distribution was reinitiated a few months later, but the momentum of acceptance had been broken and the prompt disappearance of polio that researchers had hoped for did not come about. Meanwhile, research on a more highly purified killed-virus vaccine and on several live-virus vaccines progressed, and within a few years the Salk vaccine was displaced in the US (but not in Sweden) by live-virus vaccines.

References

[1] Francis, T., Jr et al. (1955). An evaluation of the 1954 poliomyelitis vaccine trials-summary report, *American Journal of Public Health* **45**, 1–63.

[2] Meier, P. (1957). Safety testing of poliomyelitis vaccine, *Science* **125**, 1067–1071.

[3] Rutstein, D.D. (1957). How good is polio vaccine?, *Atlantic Monthly* **199**, 48.

PAUL MEIER & REBECCA PRINGLE SMITH

Sample Size Adequacy in Surveys

Introduction

The Cambridge Dictionary of Statistics defines a **sample survey** as "A study that collects planned information from a *sample* of individuals about their history, habits, knowledge, attitudes, or behavior in order to *estimate* particular *population* [Italics ours] characteristics" [5]. While the general concepts involved in the determination of **sample size** adequacy are essentially the same for sample surveys as they are for other designs, the actual process is generally driven by the following three features:

Consideration of Sample Design

In most other study designs, the data are assumed to arise from an unrestricted **random sampling** process (i.e. random sample from an infinite universe). In sample surveys, the sampling process can involve **stratification**, **clustering**, **multistage** selection, and other complications such as **missing data**; and methodology has to take these into consideration. Also, the chances of being selected in the survey may not be the same for each population unit (i.e. unequal selection probability).

Emphasis on Estimation

In sample survey work, the primary objectives generally focus on construction of point and interval estimates of characteristics of the population (*see* **Estimation**; **Estimation, Interval**). Hypothesis testing is often a secondary objective. Thus, sample size assessment is generally stated in terms of the widths of symmetric α **confidence intervals** achievable with a particular n (number of units or number of clusters) rather than the statistical **power** of rejecting a specified **null hypothesis**.

Population

Sample surveys involve samples of n units from a population of N units, where N is some finite number. As noted above, this contrasts with the usual (often implied) assumptions made in statistical **inference**; namely, that the units are selected by random sampling from a universe assumed to be infinite. It also allows estimation of population totals or aggregates, which are not possible if populations are considered infinite.

Objectives

With this in mind, our objective in this entry is to formulate an approach to sample size assessment, which is accessible and can be readily used in a wide variety of situations for point estimation of entities such as means, totals, proportions, and ratios. We are aiming this discussion at researchers who need to assess adequacy of sample size for purposes of a grant or contract application or some similar activity. The methods described below are not "cutting edge" but are widely known, widely used, and applicable to many situations found in practice.

Approach

The approach we are taking is very similar to that used by Thompson in his sampling book [12]. Let us suppose we are assessing the adequacy of a sample survey design based on a **simple random sample** of n enumeration units from a population containing N enumeration units, and we want the resulting half-width of the α confidence interval for the estimated mean to be no wider than ε units on the scale of the variable being estimated. Let us first ignore the fact that we are sampling from a finite population and assume that we are using unrestricted random sampling from an infinite universe and that the variable of interest is **normally distributed** about its mean with **variance** equal to σ^2, where σ^2 is known. Since the **standard error** of an estimated mean under unrestricted random sampling is equal to $\sigma/\sqrt{n}$, the half-width of the α confidence interval for the estimated mean is equal to $z_{1-\alpha/2}\sigma/\sqrt{n}$. Then from classical theory, the sample size, n, satisfies the specifications if the following holds:

$$z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \varepsilon$$

or

$$n \geq \frac{z_{1-\alpha/2}^2 \sigma^2}{\varepsilon^2} \tag{1}$$

where

$z_{1-\alpha/2}$ = the $100 \times (1-\alpha/2)$ percentile of the normal distribution

σ^2 = the variance of the variable being analyzed

For finite population sampling, if $\hat{d}$ is an estimate of a parameter, d, then the specifications are that we want n large enough that the half-widths of the resulting $100 \times (1-\alpha/2)\%$ confidence intervals are no wider than ε. However, the variance of $\hat{d}$ is dependent not only on the variance, σ^2, and the sample size, n, but also on the sample design being used (denoted by Ψ) and the population size, N.

Let $\mathrm{Var}(\hat{d}) = f_\Psi(\sigma, n, N)$. Then the specification is satisfied if the following inequality holds:

$$z_{1-\alpha/2}\sqrt{f_\Psi(\sigma, n, N)} \leq \varepsilon \tag{2}$$

Frequently Used Sample Designs

Simple Random Sampling

If the design is simple random sampling of n units from a population of N units, then the inequality (2) is given by

$$z_{1-\alpha/2}\sqrt{\left(\frac{\sigma^2}{n}\right)\left(\frac{N-n}{N-1}\right)} \leq \varepsilon \tag{3}$$

Solving inequality (3) for n, we see that the specification is satisfied if

$$n \geq \frac{N\sigma^2 z_{1-\alpha/2}^2}{\sigma^2 z_{1-\alpha/2}^2 + (N-1)\varepsilon^2} \tag{4}$$

If n is very much smaller than N (for practical purposes, $n \leq 10\%$ of N), then the right hand side of relation (4) reduces numerically to the right hand side of relation (3).

The following example will illustrate the procedures discussed above for determining the adequacy of sample size. Suppose we guess that approximately two-thirds of the clients of a chain of weight loss clinics will lose between 2 and 14 pounds during the first month, and that we wish to take a simple random sample of 25 clients from the 750 registered clients. We wish to be 95% confident of estimating the true mean weight loss to within 3 pounds of its true value. Is the proposed sample of 25 clients large enough to meet this specification?

In this example, $N = 750$, $n = 25$, $\varepsilon = 3$, $\sigma = 8$ (assuming a normal distribution, approximately two-thirds of values encompass 2 standard deviations from the mean), and $z_{1-\alpha/2} = 1.96$. From relation (4), it would require a sample of 26.39 (which rounds to 27) persons to satisfy the specification, so $n = 25$ is too small a sample size.

Stratified Random Sampling

A major rationale for stratification is its potential property as a variance lowering design tool. Thus, if effective, it should yield an estimate that has a lower standard error than that obtained by a simple random sampling design having the same number of observations. Thus, it should require a smaller n to obtain the same specifications of precision.

The variance of an estimate from **stratified random sampling** depends on the strata variances within, denoted $\sigma_h : h = 1, \ldots, L$; the strata population sizes, $N_h : h = 1, \ldots, L$, and the sampling allocations, $\pi_1 = n_1/n, \ldots, \pi_L = n_L/n$ within each stratum (see [9], p. 147 for the specific form of the variance). Using this and solving the resulting expression in relation (2), we would obtain the following approximate expression for the total sample size, n, that satisfies the specifications:

$$n \approx \frac{\left(\frac{z_{1-\alpha/2}^2}{N^2}\right)\left(\sum_{h=1}^{L}\frac{N_h^2\sigma_{hx}^2}{\pi_h}\right)}{\varepsilon^2 + \left(\frac{z_{1-\alpha/2}^2}{N^2}\right)\left(\sum_{h=1}^{L}N_h\sigma_{hx}^2\right)} \tag{5}$$

If estimates of the components of (5) are available from pilot studies, previous or similar surveys, or can be "guesstimated", then (5) can be used to assess the adequacy of the sample size.

Two-stage Cluster Sampling

In many situations, it may not be either feasible or possible to compile **sampling frames** that enumerate all units for the entire population. In such cases, one may be able to construct a sampling frame that identifies groups or *clusters* of enumeration units without listing explicitly all individual units. One can perform sampling from such frames by taking a sample of clusters, obtaining a list of individual

units within each selected cluster, and then selecting a sample of the enumerated units.

While the approach for sample size estimation is similar for two-stage **cluster sampling** designs as for the more simple designs considered above, the variance formulas are often more complex, especially if the clusters vary with respect to the number of sampling units and are selected with unequal probabilities at the first stage of sampling. One approach is to use the following approximate formula for the variance of an estimate from a two-stage cluster sample and substitute it into relation (2):

$$\mathrm{Var}(\overline{\overline{x}}_{\mathrm{clu}}) = \mathrm{Var}(\overline{x}_{\mathrm{srs}}) \times DEFF, \qquad (6)$$

where

$\mathrm{Var}(\overline{\overline{x}}_{\mathrm{clu}})$ = the variance of an estimated mean per unit from a two-stage cluster sample,

$\mathrm{Var}(\overline{x}_{\mathrm{srs}})$ = the variance of an estimated mean per unit from a simple random sample of the same number, n, of observational units,

and

$$DEFF = \frac{\mathrm{Var}(\overline{\overline{x}}_{\mathrm{clu}})}{\mathrm{Var}(\overline{x}_{\mathrm{srs}})}$$

DEFF is a quantity known as the **design effect**, which reflects the inflation in variance (mostly due to the sampling of clusters and the unequal weighting of observations intrinsic in most multistage designs). Sometimes the design effect can be estimated as a product of two factors: one representing the inefficiency of the unequal weighting used in the sample design and the other representing the intraclass correlation coefficient (*see* **Correlation**), which represents, in essence, the inefficiency in sampling more than 1 unit from the same cluster. Specific formulas are shown in [9].

Some Practical Guidelines for Assessing Sample Size Adequacy

The authors highly recommend utilizing the following guidelines to help ensure both appropriateness and accuracy in determining sample size adequacy.

Required Input from Client

1. Specification of key variables (and their approximate value or range of values) on the basis of which sample sizes will be determined.
2. Specification of needed precision of the estimates and "credibility level" of obtaining that level of precision. Note: different clients may request precision in a variety a ways (e.g. standard error of the estimate, half-width of confidence interval, coefficient of variation (CV) (*see* **Standard Deviation**), and effective sample size). Sometimes differences are of interest, so aspects such as power, detectable differences or effect size (*see* **Design Effects**) might be what are specified.
3. Available budget (if applicable) or cost components.
4. Sample Design (sampling plan and estimation procedure).
5. "Guesstimates" of variance of relevant variables, intracluster correlations or a range of possible correlations.
6. Will the survey be designed to satisfy multiple constraints?
7. Response and eligibility definitions and rate assumptions.

Computation of Required Sample Sizes

1. Search for available **algorithms/software** that is appropriate to obtaining the required sample size based on the particular sample design and estimation procedure being used.
2. On the basis of client input (3. above), run appropriate statistical algorithm. Though a theoretical algorithm may exist, an applicable one may not and hence some programming may be necessary.
3. Perform quality control (QC) procedures on the resulting sample size determination.

Items/Procedures for QC Checks

1. Verify with client investigators that determined sample sizes have both "face validity" (*see* **Health Status Instruments, Measurement Properties of**) and are likely to be obtained given the available budget.
2. Verify that the sample size satisfies the study specifications (i.e., using the obtained sample size, compute the precision of the estimate). This should be performed for several key estimates (at least for all on which sample sizes are based), if the survey is required to satisfy multiple constraints.

Appendix

References to Methodology and Formulas for sample sizes appropriate for common specific sample designs (denoted by number in reference section) are given below:

1. **Simple random sampling** [4, 6, 7, 9]
   a. Linear estimates
   b. **Ratio estimates**
2. **Systematic sampling** with a random start [1, 9]
3. **Stratified random sampling** [2, 4, 6, 7, 9, 10]
   a. Equal allocation
   b. Proportional allocation
   c. Optimal allocation
4. **Multistage Cluster sampling** [4, 6, 7, 11]
5. **Two-Phase sampling** [9]
6. **List-assisted telephone** surveys [7–9]
7. Multiple variance constraints [3].

References

[1] Bellhouse, D.R. (1998). Systematic sampling, in *Encyclopedia of Biostatistics*, P.A. Armitage & T. Colton, eds. Wiley, Chichester.

[2] Brewer, K.R.W. (1984). Stratified designs, in *Encyclopedia of Statistical Sciences*, N. Johnson & S. Kotz, eds. Wiley, New York.

[3] Chromy, J.R. (1987). Design optimization with multiple objectives, *Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, VA.

[4] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[5] Everitt, B.S. (2002). *The Cambridge Dictionary of Statistics*, 2nd Ed. Cambridge University Press, Cambridge, UK.

[6] Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vols. 1 and 2. Wiley, New York.

[7] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[8] Lepkowski, J.M. (1988). Telephone sampling methods in the United States, in *Telephone Survey Methodology*, R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, II & J. Waksberg, eds. Wiley, New York.

[9] Levy, P.S. & Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*, 3rd Ed. Wiley, New York.

[10] Parsons, V. (1998). Stratified sampling, in *Encyclopedia of Biostatistics*, P.A. Armitage & T. Colton, eds. Wiley, Chichester.

[11] Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

[12] Thompson, S.K. (1992). *Sampling*. Wiley, New York.

Further Reading

Castelloe, J.M. (2001). Power and sample size determination for linear models, *Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference*, Paper 240–26. SAS Institute Inc, Cary.

Elashoff, J.D. (2000). *nQuery Advisor registered Version 4.0 User's Guide*. Dixon Associates, Los Angeles.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

Jensen, R.J. (1978). *Statistical Survey Techniques*. Wiley, New York.

Rice, J.A. (1995). *Mathematical Statistics and Data Analyses*, 2nd Ed. Duxbury Press, Belmont.

MICHAEL A. PENNE & PAUL S. LEVY

Sample Size Determination for Clinical Trials

A fundamental rule of **sample size determination** is that the method of calculation should be based on the planned method of analysis. For **clinical trials**, the array of analytic methods is large. Rather than produce a catalog of methods, this article will provide the reader with a discussion of the conceptual issues behind sample size determination that arise specifically in clinical trials.

One defining characteristic of a clinical trial is that observations are made on human subjects. This impacts on the sample size calculation of clinical trials in the form of three primary distinguishing features. First, however well-intentioned, patients cannot always be cooperative with the planned conduct of the study. In dealing with plots of land or laboratory animals, one is usually less concerned about the experimental "units" receiving only part of the assigned treatment, or receiving another treatment, or failing to be present for a scheduled measurement. Factors such as these are a significant part of everyday clinical trials, and their impact on sample size can be substantial.

Secondly, because clinical trials deal with the treatment of humans, ethical issues (*see* **Ethics of Randomized Trials**) raise the importance of the sample size calculation. A sample size that is too small can lead to a failure to detect a treatment effect and consequently the abandonment of what may be a very promising treatment. This in turn represents the breach of an implicit contract with the study patients that the trial, as designed, is of sufficient size to detect a useful improvement in treatment. If it is not, then the patients may have needlessly donated their cooperation and needlessly been exposed to an experimental therapy whose **risks** and benefits are as yet not clearly known.

Finally, clinical trials tend to be costly. The Systolic Hypertension in the Elderly Program [1] trial was budgeted at 50 million dollars, and the Women's Health Initiative at about ten times that amount. With such large expenditures involved, one needs to balance carefully the extra cost of requiring a larger sample size with the danger of failing to detect a useful treatment difference. To repeat a very complex or large trial with another of adequate sample size will often be difficult to justify, even if the sponsor's objectives are still worth pursuing.

Clinical trials can be designed in many ways, but the setting most frequently encountered is the comparative trial, in which two or more treatments are compared. It will be assumed here that the primary objective is **hypothesis testing**, and initially, the discussion will be restricted to the two-group case.

Sample size formulas frequently take the form

$$N_{\text{tot}} = \frac{c^2(z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_{\text{A}})^2}{\delta^2}, \qquad (1)$$

where N_{tot} is the total sample size, c is a constant, $z_{1-\alpha}$ is the **standard normal deviate** whose probability of being exceeded is α, σ_0 and σ_{A} are the **standard deviations** under the **null** and **alternative hypotheses**, and δ is the treatment effect. The variables on the right-hand side of this formula are parameters whose values depend on the design of the trial. Values for all parameters must be obtained to determine the sample size. In most cases, however, one cannot be confident of knowing all of these values with precision. Arriving at a chosen sample size often involves calculating sample size for a range of values of the uncertain parameters and choosing a size that seems feasible and reasonably close to meeting the scientific requirements of the study. Sample size calculations usually involve other considerations, such as non **compliance** and **missing data**, which directly impinge on the parameters in (1).

Choice of Outcome Measure

Primary and Secondary Outcomes

Often in clinical trials there are many variables that can be used to assess the success of the treatment. These are called outcome or endpoint variables (*see* **Outcome Measures in Clinical Trials**). Under the **null hypothesis** of no treatment effect, the formal testing of many outcome variables (*see* **Multiplicity in Clinical Trials**) increases the probability of falsely discovering a significant effect for at least one of them. To maintain the desired probability of a type I error in the presence of many outcomes, adjustment for **multiple comparisons** is necessary. One consequence of such adjustment is that the **power** is reduced for demonstrating that a specific one of those

outcomes is significantly altered by the treatment. In order to strike a balance between the desirability of assessing the effect of treatment on many variables, while maintaining adequate power for specific variables, outcomes in clinical trials are often classified as primary or secondary. Usually the primary endpoint is limited to one outcome, although there are exceptions. The efficacy of the treatment is then formally based on statistical significance only with respect to the primary outcome. If more than one primary outcome is chosen, then a multiple comparisons adjustment is usually applied; this is reflected in the sample size calculations through the significance levels assigned to the individual primary outcomes. The sample size is then chosen so that there is adequate power for each primary outcome. Note that multiple comparisons are appropriate only if one will claim success when any of the primary outcomes is significant. If significance is required for all of the primary outcomes, then the type I error becomes much smaller, rather than larger, than the nominal significance level. The usual practice in this case, however, is to use the nominal significance level.

Repeated Measurements

Now assume that a single primary outcome has been chosen, and attention is at first restricted to continuous measurements. For purposes of illustration, consider a trial of hypertensive patients in which blood pressure has been identified as the primary outcome. There are many possible times at which this measurement can be taken. Before **randomization** it can be used to identify patients who are eligible for the trial, i.e. hypertensive patients. In a long trial, blood pressures may be measured frequently to assure that the blood pressure of hypertensive patients is adequately controlled. It also could be measured frequently to study the time course of response. Furthermore, since the variability of the measurement is a crucial factor for sample size, one can use/repeated measurements (*see* **Longitudinal Data Analysis, Overview**) to reduce the variability. For example, one can average all of the postrandomization measurements for an individual. For blood pressure, this is rarely done, because the initial response may wane, owing to biologic changes or diminishing compliance with the treatment regimen. Alternatively, the **variance** can still be reduced by taking multiple measurements near the end of a specified period of interest. In deciding whether to take multiple measurements, many factors must be considered. It is well known that the variability of blood pressure measurements has between-patient and within-patient variability, and within a patient, there are also within-visit and between-visit components of variability. While it is less expensive to measure a patient repeatedly within a single visit, repeating the within-visit measurement beyond two or three times is generally not done because at this point the between-visit variability becomes the dominant source of variation. The cost of increasing the number of visits has to be weighed against the cost of enrolling more patients. Additionally, increasing the number of visits per patient may result in increasing rates of failure to attend some of these visits.

Change from Baseline

Should one use change from baseline or just the follow-up assessment? The choice here hinges on the **correlation** between baseline and follow-up: if the correlation is less than 0.5, then only the follow-up measurement should be used, otherwise the difference is preferred. This correlation is a function of the length of time between baseline and follow-up. The same considerations regarding the **variance components** of the final measurement also apply to the baseline measurement (*see* **Baseline Adjustment in Longitudinal Studies**).

Survival

If the outcome is an "event" that may happen over time, then the trial is frequently called a survival trial (*see* **Survival Analysis, Overview**). Although survival trials bring to mind outcomes such as death or heart attack, for a drug designed to relieve pain the event could be meaningful pain relief. Patients who do not have an "event" during a specified period are said to have **censored** observations. With a survival-type event, an important consideration is whether one is interested in comparing the entire survival curve (i.e. the times to an event) or only the proportions surviving at a specified time. In sepsis trials, patients have virulent infections with high mortality rates within 30 days (*see* **Clinical Trials of Antibacterial Agents**). Suppose, at the end of 30 days, the drug is unable to keep alive more patients than the placebo. While a comparison of survival curves may detect a difference in time to death, this may amount to a

few extra days of life for unwell patients who still succumb within the 30 days. Thus, it is argued, only the proportions surviving at 30 days are important.

When there is censoring, comparing proportions becomes more difficult. In a trial such as the sepsis trial, where each patient is observed for exactly 30 days, and mortality from any cause is the primary outcome, censoring is not expected to be a problem, since the outcome information can presumably be determined for all patients. However, in a trial where the outcome is fatal or nonfatal heart attack, death from a noncardiac disease would censor the data. In long-term trials, where patient entry is staggered over a recruitment period, and each patient is followed until the trial is closed out, censoring is very common. In these situations, each patient has a different length of follow-up. It is difficult to assign meaning to the term "proportion surviving" without specifying a fixed period of observation common to all patients. One can estimate the probability of surviving when patients have differential lengths of follow-up using the **Kaplan–Meier** method. Kaplan–Meier estimates of the probability of surviving to the end of such a trial are usually very unreliable because of heavy censoring towards the end of the trial. Often in these long-term trials, survival curves are compared and the **logrank test**, which compares the entire survival experience, is usually preferred. Sample size methodology for the logrank statistic has been developed by Schoenfeld [12], Freedman [4], and Lakatos [7]. The Kaplan–Meier method and logrank statistic are designed only for noninformative censoring, and the sample size methods just referenced assume noninformative censoring as well (*see* **Sample Size Determination in Survival Analysis**).

Selecting the Parameters

Frequently, identifying a method or formula which is reasonable for determining the appropriate sample size is the easy part of the problem. With every sample size formula, the statistician must choose values for the parameters. Unfortunately, the values of most of the parameters are only poorly known before the trial. In specific applications, the values of some of the parameters may be fairly well agreed upon. Some statisticians rely on the clinicians to provide values for the parameters. However, the statistician is best served by judging his or her own estimates for all parameters, and negotiating these judgments with the clinicians. It should be understood that the ability to negotiate successfully for adequate sample size may be greatly diminished if the statistician does not have a firm position on the clinical background and the choice of parameters.

Type I and Type II Error Rates (Significance and Power)

Type I Error (Corresponding to $z_{1-\alpha}$)

The **null hypothesis**, H_0, is that the experimental therapy has no different effect on the outcome compared with the **control**. Associated with every test is the type I error rate, or significance level, which is the probability of falsely rejecting the null hypothesis. When testing an experimental treatment, it is usually assumed that one has introduced this new treatment because it is thought to be superior to control, and that the trial is being carried out to "prove" this. There is rarely any interest in proving that it is worse; often sponsors will abandon a drug that is not much better than the control (although see the section "Equivalence Trials" below). According to most authors, the standard by which a new treatment is judged a success is whether the equivalent of the **standard normal deviate** value of the primary outcome statistic exceeds 1.96. The implied level of significance corresponding to 1.96 is generally felt to be a reasonable standard against which all trials should be judged. The 1.96 can correspond to a one-sided hypothesis with significance level 0.025 or a two-sided hypothesis of significance level 0.05. If one is truly interested in also knowing whether the treatment is worse than control (a recent trial of digitalis is an example), then 1.96 can also be used as the standard for testing this. A minority opinion is that because the researcher is not interested in testing whether the treatment is worse than control, a reduced standard, such as 1.645, is appropriate for judging whether the experimental therapy is better. However, most biostatisticians consider that the standards by which an experimental therapy is judged better than control should not depend on whether one was interested in proving that it is worse than control.

Type II Error Rate (Corresponding to $z_{1-\beta}$)

The choice of the **alternative hypothesis**, H_a, is an area of frequent contention (see below). The type II

error rate is the probability of failing to reject the null hypothesis, given that the alternative hypothesis is true. The power is one minus the type II error rate. In other words, if the experimental treatment is truly better than control by the amount postulated in the alternative hypothesis, the power is the probability of showing a significant difference. When trials are designed, a good deal of attention may be placed on whether the power is adequate, with 90% power the generally accepted standard for major Phase III trials. However, when a trial is analyzed, there is often little or no attention paid to the power, with the significance level dominating the spotlight. If the trial fails to reach the conventional criterion for statistical significance, then it is often concluded that the experimental treatment was no better than control. Such a conclusion may not be true – the lack of significance may simply be a matter of inadequate power. In the analysis of the trial, lack of power is properly expressed in a wide **confidence interval** for the treatment difference.

A Balance Between Type I and Type II Errors

The error of falsely concluding benefit of an ineffective therapy (type I error) has traditionally been thought of as the more critical of the two types of error. If ineffective therapies were given the stamp of approval, then physicians would end up prescribing ineffective medications, and patients would spend large amounts of money on worthless treatments. More importantly, patients might be denied superior therapies because physicians believed an approved inferior was effective. Furthermore, research in a disease might be impeded because a beneficial therapy appeared to exist.

However, if trials are designed with inadequate power, then there are also important consequences. Some of these were discussed at the beginning of the article. Inadequate power can result in failure to detect a treatment that is truly effective. If the therapy is subsequently abandoned, then patients are denied therapy that could provide relief or even save lives. Alternatively, if there is still hope and conviction that the treatment is sufficiently promising, then trials will have to be rerun, this time with adequate power; this process leads to loss of time and money.

On balance, primary importance is still given to the type I error. Additionally, some trials are felt to be of sufficient importance to warrant very large sample sizes and consequently high power. Even then, the type I error rate is generally chosen to be at least as small as the type II.

The Treatment Effect (δ)

There are a number of different criteria that may be used to arrive at the treatment effect for the calculation, i.e. the alternative hypothesis H_a. Two criteria that are very important are (1) the smallest clinically meaningful difference (SCMD), and (2) the anticipated treatment effect (ATE).

As sample sizes increase, the size of an estimated treatment effect that can be detected as statistically significant decreases towards 0. However, very small treatment effects may not be meaningful clinically, and trials need not be designed to detect such small differences. But if the sample size of a trial is based on a rather large treatment effect which is considerably larger than some meaningful difference then, at the final analysis, statistical significance may very well not be achieved because the treatment effect is smaller than the large assumed treatment effect, but still clinically meaningful. Had the original trial been powered to detect this smaller effect, statistical significance of this clinically meaningful difference would have been achieved. Thus, it is important to consider what is the SCMD, even if its exact value is often elusive (see below).

The decision as to what is a clinically meaningful difference is one that requires the input of the statistician, the clinician and the sponsor. Consider first, mortality. While "clinically meaningful" seems to imply a physician's opinion, one could argue that patients are often in the best position to judge whether a decrease in mortality is meaningful in a specific situation. Some people feel that any decrease in mortality is meaningful. Others might consider a minimum of 10% decrease in the probability of mortality necessary to offset the toxic side effects of a given chemotherapy regimen. A sponsor may consider the **prevalence** of a disease to be so low that if the drug produces only a 10% reduction in mortality, then it would not be worth the cost of development. With so many diverse considerations, it should not be surprising that agreement on an SCMD in mortality is often difficult.

Now suppose the primary endpoint of the trial is diastolic blood pressure. For an individual patient,

a physician might feel that a 10 mmHg decrease is the smallest that is clinically meaningful. The desired decrease might be larger for patients with higher initial blood pressures. In many trials, however, the targeted treatment effect compared with placebo is chosen to be in the range of 3–5 mmHg reduction. Why is there this discrepancy? One reason may be that the anticipated treatment effect (ATE) is much smaller than 10 mmHg. This arises because the observed reduction will be the average of a population of individual changes, with some of the distribution of changes being well below 10 mmHg. One could consider using a drug that achieved a 10 mmHg or larger reduction in only a portion of the patients. Physicians routinely monitor the blood pressure of hypertensive patients, and if a patient's blood pressure remains excessive with the current prescribed medication, the physician will usually change the prescription.

If the ATE is considerably larger than the SCMD, then it is unrealistic to expect the sponsor to support the sample size needed to detect an effect much smaller than is likely to occur. However, it has been common to conduct trials that attempt to detect treatment effects that are far larger than could realistically be anticipated. In summary, the chosen treatment effect should reflect both of these considerations: it must be achievable and it must be meaningful.

Equivalence Trials

Sometimes treatments are introduced which are expected to have no better efficacy, or minimally poorer efficacy, than an already accepted treatment. Some reasons for introducing such a therapy may be because this new treatment may be less toxic, less expensive, or have fewer side effects (*see* **Equivalence Trials**). One of the key differences between testing for superiority and equivalence is that the latter requires the designers to specify the largest acceptable treatment difference as part of the null hypothesis. This factor is critical, because this declared difference plays a large role in the ***P* value** obtained at the end of the trial, and whether or not the nominal significance level is attained. A central issue for equivalence testing is agreeing upon how much one is willing to accept reduced efficacy in exchange for other benefits such as reduced toxicity. If a new therapy offers very visibly reduced side effects as compared with a standard highly toxic chemotherapy, a patient or his physician may be willing to accept up to perhaps a 5% or 10% increase in mortality. With an antihypertensive medication, where the side effects are much less severe, such an increase in mortality may appear less acceptable. If, however, the fact that the patient must take the antihypertensive medication with these side effects for the rest of his life is factored into the picture, then a mortality increase may appear less objectionable. A sample size formula for this situation is

$$N_{\text{tot}} = \frac{c^2(z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_{\text{A}})^2}{(\delta_{\text{a}} - \delta_0)^2}$$

Here, δ_{a} is the difference that one is willing to accept, and δ_0 is the expected or true difference. For equivalence, we usually set $\delta_0 = 0$. Confidence intervals are often the preferred approach for analyzing the results of equivalence trials. The above formula is valid for confidence intervals.

Sample Size Adjustment

There are many factors which influence the sample size which are not considered in the sample size formulas above. These factors should be accounted for in the method of calculation. A few common factors are now discussed.

A Simple Adjustment for Noncompliance

One of the factors that distinguishes clinical trials from other experiments is the difficulty in getting patients to adhere to their assigned treatment regimens. This is particularly true in the longer **prevention trials**, since the motivation of an acute condition is not present. The **intention to treat** paradigm dictates that patients be analyzed with respect to their initial treatment assignment. The usual philosophy in sample size determination is that a treatment difference that is both meaningful and likely to occur is identified. Then, recognizing that many patients may take less than their assigned treatment regimen, the statistician models the impact of noncompliance on the treatment effect and in turn on the probability of rejecting the null hypothesis. Many models have been proposed for accounting for noncompliance in sample size calculation [5, 6, 14]. A simple approach is discussed here that can be applied regardless of the type of outcome

variable. A more complex procedure is discussed later under survival analysis. In some trials, most of the noncompliance is likely to occur near the time of randomization. If there are difficult side effects, then these are likely to occur soon after initiation of therapy, and most patients will not endure these for long periods. Furthermore, those patients who enroll but have little interest in the therapy will stop taking medications soon after randomization. Perhaps the simplest model for noncompliance is for a new treatment vs. placebo trial, in which a proportion of the patients on the new treatment discontinue their medication immediately, while the remainder adhere. In this case, a proportion will receive no benefit of therapy, while the remainder receive full benefit. This model is considered conservative. One approach to account for this is to calculate what is referred to as the "observed treatment effect". If δ is the treatment effect under full compliance, then the observed treatment effect is $d = (1 - p_m)\delta$, where p_m is the proportion of noncompliers. A sample size adjustment factor based on this observed treatment effect d would be $n_{adj} = n/(1 - p_m)^2$, since

$$\frac{n_{adj}}{n} = \frac{4(z_\alpha + z_\beta)^2\sigma^2}{(1 - p_m)^2\delta^2} \Big/ \frac{4(z_\alpha + z_\beta)^2\sigma^2}{\delta^2}$$

Table 1 presents some inflation factors using this model and various proportions of noncompliance.

A Simple Adjustment for Nondifferential Loss to Follow-up

The term "loss to follow-up" refers to a type of censoring applied to any patient whose status with respect to the final endpoint cannot be determined at the time of the analysis. Thus, if the endpoint is stroke, and the patient dies of cancer prior to having a stroke, then this is considered loss to follow-up. If the reason for censoring is not related to the endpoint, then the censoring is referred to as noninformative or nondifferential censoring. If the probability of nondifferential loss to follow-up is l, then a simple adjustment is $n_{adj} = n/(1 - l)$. This is equivalent to assuming that all patients who are lost will be lost at the time of randomization; generally this is conservative.

Table 1 Inflation factors for a simple, conservative noncompliance adjustment

Proportion noncomplying	0.05	0.10	0.20	0.30	0.50
Inflation factor $= 1/(1 - p_m)^2$	1.11	1.23	1.56	2.04	4.00

Adjusting Survival Analyses for Noncompliance and Other Factors

Many models have been proposed for analyzing survival data. The desirability of a comparison of proportions vs. a comparison of survival curves was discussed in the section "Choice of Outcome Measure" earlier in this article. In those situations for which survival curves should be compared, the curves may have a variety of shapes. A common pattern is for events to happen soon after randomization, and then taper off. The pattern is dramatic in congestive heart failure (e.g. Study of Left Ventricular Dysfunction [13] and CONSENSUS [2]), and in angioplasty. In these nonconstant hazard situations, **proportional hazard** models should be used in preference to **exponential** models, since the latter assume that the **hazard rate** will be constant throughout the trial. Additionally, the treatment effect may vary during the trial. In angioplasty, Reopro, which is administered intravenously, is usually given for less than a day. The effects on mortality appear to continue for perhaps 6 months. However, the largest benefit is around the time of administration, with the benefit tapering rapidly over the first 30 days. A similar tapering of benefit occurs on a group basis when there is noncompliance. In these cases, nonproportional hazards models should be used in preference to proportional hazard models. Lakatos [7] provided a general method for calculating sample size for the logrank statistic which allows a great deal of flexibility in simultaneously specifying nonuniform accrual patterns, nonconstant and nonproportional hazard functions, lags in treatment effects, loss to follow-up, noncompliance, and dropout. The Lakatos method employs nonstationary **Markov** models; a simple computer program is needed for implementation. Exponential models are not appropriate if noncompliance, drop-in or treatment lag are to be modeled, and the nonstationarity of the Markov model is essential.

Determining Duration and Treatment Effect for Survival Trials

Increasing the duration of a trial increases the overall failure rate and, in turn, can reduce the required

sample size, especially if failure is a rare event. The relationship between required sample size and duration may be complex in survival trials, particularly if there is noncompliance or nonconstant hazards. Although a parameter for duration does not appear explicitly in the sample size formula, the duration which provides a prespecified power can be determined by fixing all other parameters (including sample size). A simple iterative procedure is then used, in which power is calculated for a series of fixed trial durations which are successively adjusted to bring the power as close as desired to the target power.

Similarly, for survival trials, the treatment effect may be a complex function of other factors. Noncompliance, for example, can cause a diminution of the treatment effect over time. As with duration, an iterative procedure can be used.

Adjusting for Group Sequential Designs

Methods for calculating sample size when group sequential procedures will be used (*see* **Sequential Analysis**) in the presence of trial complexities such as noncompliance, drop-in, loss-to-follow-up and the like have been developed by Lakatos [8]. Here a few issues are discussed. Suppose there is a boundary z_1, $z_2, \ldots, z_k$, and the tests are successively performed at predetermined fractions of the total "information" (for continuous endpoints the information fraction is roughly the proportion of patients in the current analysis, and for survival trials, the information fraction is roughly the proportion of events (see Lan & Zucker [10] for a detailed account). The alpha level is the cumulative probability under successive testing, and under the null hypothesis, of the calculated z value exceeding the boundary z values. In other words, it is the accumulated probability of rejecting the null hypothesis after the sequence of tests has been performed *assuming the null*. Similarly, the power is the accumulated probability of rejecting the null hypothesis after the sequence of tests has been performed *assuming the alternative*. What is important here is that to calculate the power, one generally needs to integrate, numerically, under the alternative hypothesis. It is also important to note that a lower bound on the power can be calculated without numerical integration. This is because the probability of rejecting the null at the final analysis must be less than or equal to the combined probability of rejecting it over all analyses. Thus, if the final boundary value for a group sequential procedure is $z_k = 2.10$, then an upper bound on the sample size is $n_{\text{adj}} = n[(1.28 + 2.10)/(1.28 + 1.96)]^2 = 1.088n$; here the 1.28 corresponds to 90% power, and the 1.96 to a two-sided 0.05 level test. Since the very popular O'Brien–Fleming final boundary [11] value is usually less than 2.10 (*see* **Data and Safety Monitoring**), the sample size inflation factor is usually less than 8.8%. In contrast, the sample size adjustment factor when noncompliance, drop-in, lag in treatment effect, and loss to follow-up are expected may be 100% or more (cf. Lakatos & Lan [9]). Therefore, when using the O'Brien–Fleming procedure, if the exact sample size incorporating group sequential and all other expected effects would be difficult to calculate, it is generally better to use a sophisticated procedure for calculating the effects of noncompliance and the like, and a *post hoc* inflation for the group sequential testing, than to use a sophisticated group sequential sample size method and a *post hoc* inflation for noncompliance.

Unequal Allocation between Treatment Groups

While most trials allocate patients to the treatment groups in equal proportions (e.g. 50% to group A, 50% to group B), this rule is not universally applied. If there are two or more active groups and the primary objective is to compare each active group with control, then optimal power is achieved with unequal allocation. This occurs because a disproportionately large fraction allocated to the control group will benefit all tests. Formulas for optimal allocation can be found in [3]. Frequently, in drug trials there may be a disproportionately high allocation to the active groups, simply to obtain more safety experience with a new investigational drug.

References

[1] Borhani, N.O., Applegate, W.B., Cutler, J.A., Davis, B.R., Furberg, C.D., Lakatos, E., Perry, H.M., Smith, W.M. & Probstfield, J.L. (1991). Systolic hypertension in the elderly program: Part I. Rationale and design, *Hypertension* **17S**, II-2–II-15.

[2] CONSENSUS Trial Study Group (1987). The effects of enalapril on mortality in severe congestive heart failure, *New England Journal of Medicine* **316**, 1429–1435.

[3] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. Wiley, New York.

[4] Freedman, L.S. (1982). Tables of the number of patients required in clinical trials using the log-rank test, *Statistics in Medicine* **1**, 121–129.

[5] Halperin, M., Rogot, E., Gurian, J. & Ederer, F. (1968). Sample sizes for medical trials with special reference to long-term medical therapy, *Journal of Chronic Diseases* **27**, 15–24.

[6] Lakatos, E. (1986). Sample sizes for clinical trials with time-dependent rates of losses and noncompliance, *Controlled Clinical Trials* **7**, 189–199.

[7] Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials, *Biometrics* **44**, 229–241.

[8] Lakatos, E. (2002). Designing complex group sequential survival trials, *Statistics in Medicine* **21**, 1969–1989.

[9] Lakatos, E. & Lan, K.K.G. (1992). A comparison of sample size methods for the logrank statistic, *Statistics in Medicine* **11**, 179–191.

[10] Lan, K.K.G. & Zucker, D.M. (1993). Sequential monitoring of clinical trials: the role of information and Brownian motion, *Statistics in Medicine* **12**, 753–765.

[11] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[12] Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distribution, *Biometrika* **68**, 316–318.

[13] SOLVD Investigators (1991). Effect of enalapril on survival in patients with reduced left ventricular fractions and congestive heart failure, *New England Journal of Medicine* **325**, 293–302.

[14] Wu, M., Fisher, M. & DeMets, D.L. (1980). Sample sizes for long term medical trials with time-dependent noncompliance and event rates, *Controlled Clinical Trials* **1**, 109–121.

(*See also* **Sample Size Determination**; **Sample Size Determination in Survival Analysis**)

E. LAKATOS

Sample Size Determination in Survival Analysis

Many survival studies are designed to compare two alternative treatments, but information on the values of certain **explanatory variables** may also be available. It has been shown by Schoenfeld [6] that the expression for calculating the required number of deaths is the same whether or not account is taken of supplementary explanatory variables. For this reason, an efficacy study to compare the survival times of individuals who receive a new treatment with those who receive a standard will be the focus for this article.

Suppose that there are two groups of individuals, and that the standard treatment is allocated to the individuals in Group I, while the new treatment is allocated to those in Group II. Assuming a **proportional hazards model** for the survival times, the hazard of death at time t for an individual on the new treatment, $h_{\mathrm{N}}(t)$, can be written as

$$h_{\mathrm{N}}(t) = \psi h_{\mathrm{S}}(t),$$

where $h_{\mathrm{S}}(t)$ is the **hazard** function at t for an individual on the standard treatment and ψ is the unknown **hazard ratio**. We will also define $\theta = \log \psi$ to be the log-hazard ratio. If θ is zero, then there is no treatment difference. On the other hand, negative values of θ indicate that survival is longer under the new treatment, while positive values of θ indicate that individuals survive longer on the standard treatment.

In a survival study, the occurrence of **censoring** means that it is not usually possible to measure the actual survival times of all individuals in the study. However, it is the number of actual deaths that is important in the analysis, rather than the total number of subjects. Accordingly, the first step in determining the required number of individuals in a study is to calculate the number of deaths that must be observed.

Calculating the Required Number of Deaths

To determine the sample size requirement for a study (*see* **Sample Size Determination for Clinical Trials**), we calculate the number of individuals needed for there to be a certain chance of declaring θ to be significantly different from zero when the true, but unknown, log-hazard ratio is θ_{R}. Here, θ_{R} is the *reference value* of θ. It will be a reflection of the magnitude of the treatment difference that it is important to detect, using the test of significance (*see* **Hypothesis Testing**). In practice, θ_{R} might be chosen on the basis of the increase in the **median survival time** that is to be detected, or in terms of the probability of survival beyond some specific time.

The required number of deaths is taken to be such that there is a probability of $1 - \beta$ of declaring the observed log-hazard ratio to be significantly different from zero, using a hypothesis test with a specified significance level of α, when in fact $\theta = \theta_{\mathrm{R}}$ (*see* **Level of a Test**). The quantity $1 - \beta$ is the probability of rejecting the null hypothesis when it is in fact false, and is the **power** of the test. Both α and β are taken to be small, and the values chosen will depend on the circumstances of the study; typical values are $\alpha = 0.05$ and $\beta = 0.1$.

The required total number of deaths, d, can be found using

$$d = \frac{4(z_{\alpha/2} + z_{\beta})^2}{\theta_{\mathrm{R}}^2}, \qquad (1)$$

where $z_{\alpha/2}$ and z_{β} are the upper $\alpha/2$- and upper β-points, respectively, of the standard normal distribution.

This result appears in many papers, although the assumptions on which the result is based can be different. For example, Bernstein & Lagakos [1] obtain (1) on the assumption that the survival times in each of the two groups have **exponential distributions**. However, Schoenfeld [5] obtains the same result when the **logrank test** is used as a basis for comparing the treatments, without making the assumption of exponentiality; this derivation is included in [2].

A variant on the formula for the required number of deaths is given by Freedman [3], who has $[(1 + \mathrm{e}^{\theta_{\mathrm{R}}})/(1 - \mathrm{e}^{\theta_{\mathrm{R}}})]^2$, in place of $4/\theta_{\mathrm{R}}^2$. However, when θ_{R} is small,

$$\left(\frac{1 + \mathrm{e}^{\theta_{\mathrm{R}}}}{1 - \mathrm{e}^{\theta_{\mathrm{R}}}}\right)^2 \approx \left(\frac{2 + \theta_{\mathrm{R}}}{\theta_{\mathrm{R}} + \theta_{\mathrm{R}}^2/2}\right)^2 = \frac{4}{\theta_{\mathrm{R}}^2},$$

and so the two expressions will tend to give similar results. Freedman's expression is the basis for the extensive tables of sample size requirements in Machin & Campbell [4].

The derivation of the result in (1) assumes that the same number of individuals is to be assigned to each treatment group. If this is not the case, a modification has to be made. In particular, if the proportions of individuals to be allocated to Groups I and II are π_1 and π_2, respectively, then the required total number of deaths becomes

$$d = \frac{(z_{\alpha/2} + z_\beta)^2}{\pi_1 \pi_2 \theta_R^2}.$$

Notice that an imbalance in the number of individuals in the two treatment groups leads to an increase in the total number of deaths required. The derivation also includes an approximation which means that the calculated number of deaths could be an underestimate. Some judicious rounding up of the calculated value is therefore recommended to compensate for this.

Example: Comparison of Two Treatments

A clinical trial is to be designed to compare a new form of chemotherapy with a standard for the treatment of patients with ovarian carcinoma. The time from randomization to death is the response variable of interest.

As a first step, information is obtained on the survival times, in months, of patients who have received the standard treatment. The **Kaplan–Meier** estimate of the survivor function derived from such data is shown as the step function drawn with a solid line in Figure 1.

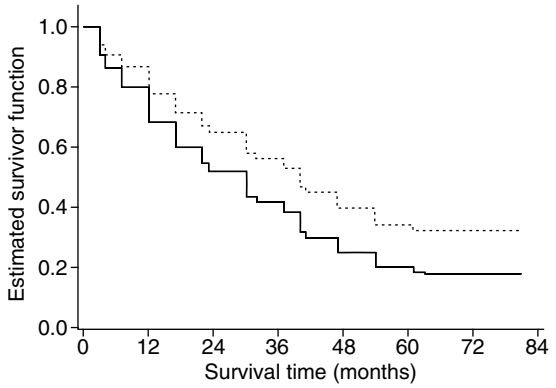

Figure 1 Estimated survivor functions for individuals on the standard treatment (—) and the new treatment (- - -)

From this estimate of the survivor function, the median survival time is 30 months, and the survival rates at 1, 3, and 5 years can be taken to be given by $S(12) = 0.68$, $S(36) = 0.42$, and $S(60) = 0.20$.

The new treatment is expected to increase the survival rate at 4 years from 0.25, the value under the standard treatment, to 0.40. This information can be used to calculate a value for θ_R. To do this, we use the result that, if the hazard functions are assumed to be proportional, then the survivor function for an individual on the new treatment at time t is

$$S_N(t) = [S_S(t)]^\psi, \tag{2}$$

where $S_S(t)$ is the survivor function for an individual on the standard treatment at t and ψ is the hazard ratio. Therefore,

$$\psi = \frac{\log S_N(t)}{\log S_S(t)},$$

and so the value of ψ corresponding to an increase in $S(t)$ from 0.25 to 0.40 is

$$\psi_R = \frac{\log(0.40)}{\log(0.25)} = 0.66.$$

With this information, the survivor function for an individual on the new treatment can be estimated by $[S_S(t)]^{\psi_R}$. In particular, $S_N(12) = 0.76$, $S_N(36) = 0.56$, and $S_N(60) = 0.35$. The estimated survivor function for the new treatment is shown as a dotted line in Figure 1.

The median survival time under the new treatment can be found from this estimate of the survivor function. Using Figure 1, the median survival time under the new treatment is estimated to be about 40 months.

To calculate the number of deaths that would be required in a study to compare the two treatments, we take $\alpha = 0.05$ and $\beta = 0.10$. With these values of α and β, $z_{\alpha/2} = 1.96$ and $z_\beta = 1.28$, and taking $\theta_R = \log \psi_R = \log 0.66 = -0.416$, the number of deaths required to have a 90% chance of detecting a hazard ratio of 0.66 to be significant at the 5% level is then given by

$$d = \frac{4(1.96 + 1.28)^2}{0.416^2} = 243.$$

Allowing for possible underestimation, this can be rounded up to 250 deaths in total.

Calculations such as those used in this example are only going to be of direct use when a study is to be continued until all patients entered into the study have died. In most trials, the analysis will take place before everyone has experienced the endpoint, so that some observations will be censored. This has to be taken into account when designing the study, and so we now examine how the required number of individuals can be obtained.

Calculating the Required Number of Individuals

To calculate the actual number of individuals that are required in a survival study, we need to consider the probability of death over the duration of a study. Typically, individuals are recruited over an *accrual period* of length a. After recruitment is complete, there is an additional *follow-up period* of length f. The total duration of a study will therefore be of length $a + f$. Notice that, if f is small, or even zero, then there will need to be correspondingly more individuals recruited in order to achieve a specific number of deaths.

Once the probability of an individual dying in the study has been evaluated, the required total number of individuals will be found from

$$n = \frac{d}{\Pr(\text{death})}, \tag{3}$$

where d is the required number of deaths found from (1). The probability of death can be taken as

$$\Pr(\text{death}) = 1 - \frac{1}{6}[\overline{S}(f) + 4\overline{S}(0.5a + f) + \overline{S}(a + f)], \tag{4}$$

where

$$\overline{S}(t) = \frac{S_S(t) + S_N(t)}{2}, \tag{5}$$

and $S_S(t)$ and $S_N(t)$ are the estimated values of the survivor functions for individuals on the standard and new treatments, respectively, at time t. This result is similar to that given by Schoenfeld [6], and full details of the derivation are included in Collett [2].

A simpler result is based on the assumption that survival times are exponentially distributed. If the mean survival times under the standard and new treatments are λ_S^{-1} and λ_N^{-1} respectively, then the average survivor function, $\overline{S}(t)$, is given by

$$\overline{S}(t) = \frac{e^{-\lambda_S t} + e^{-\lambda_N t}}{2}. \tag{6}$$

Estimates of λ_S and λ_N can be obtained from the corresponding median survival times for each treatment group, using the result that the median, t_m, of an exponential distribution with mean λ^{-1} is such that $\lambda = (\log 2)/t_m$.

Although (6) is based on more restrictive assumptions about the distribution of survival times than (5), it will often lead to quite similar results. Schoenfeld & Richter [7] give nomograms that enable the required number of individuals to be determined on the assumption of exponential survival times.

The tables in [4] follow [3] in assuming that each individual in a study is followed up for some time τ after randomization, and that analysis takes place at time τ after the last person has been recruited. The proportion of individuals expected to survive in each group is then $S_S(\tau)$ and $S_N(\tau)$, and so the probability of death is $1 - [S_S(\tau) + S_N(\tau)]/2$. In many situations, individuals are followed up until the end of the study, rather than for a fixed time. In this case, the value of τ could rather be taken to be the average length of follow-up, given by $\tau_0 = f + (a/z)$. However, this approach does not take account of patient follow-up extending beyond time τ_0, and so the required number of individuals will tend to be overestimated.

The result in (3) shows how the required number of individuals can be calculated for a trial with an accrual period of a and a follow-up period of f. Of course, the duration of the accrual period and follow-up period will depend on the recruitment rate. So suppose that the recruitment rate is expected to be m individuals per month and that d deaths are required. The number recruited in an accrual period of length a is then ma, and so the expected number of deaths in the study is

$$ma \times \Pr(\text{death}).$$

Values of a and f which make this value close to the number of deaths required can then be found numerically – for example, by trying out different values of a and f. This **algorithm** could be computerized and an optimization method used to find the value of a

that makes

$$d - [ma \times \Pr(\text{death})] \tag{7}$$

close to zero for a range of values of f. Alternatively, the value of f that yields the result in (7) for a range of values of a can be found. A two-way table giving the required number of individuals for different combinations of values of a and f will be particularly useful in planning a study. This process is facilitated by software packages for calculating sample size requirements in a survival study, such as nQuery Advisor.

Example: Comparison of Two Treatments (Continued)

Previously, it was shown that 250 deaths needed to be observed for the study on ovarian cancer to have sufficient power to detect a hazard ratio of 0.66 as significant. Suppose that individuals are to be recruited to the study over an 18-month accrual period and that there is to be a subsequent follow-up period of 24 months. From (4), the probability of death in the 42 months of the study will then be given by

$$\Pr(\text{death}) = 1 - \frac{1}{6}[\overline{S}(24) + 4\overline{S}(33) + \overline{S}(42)].$$

Now, using the estimated survivor functions shown in Figure 1,

$$\overline{S}(24) = \frac{S_S(24) + S_N(24)}{2} = \frac{0.52 + 0.64}{2} = 0.58,$$

$$\overline{S}(33) = \frac{S_S(33) + S_N(33)}{2} = \frac{0.42 + 0.56}{2} = 0.49,$$

$$\overline{S}(42) = \frac{S_S(42) + S_N(42)}{2} = \frac{0.30 + 0.44}{2} = 0.37,$$

and so the probability of death is

$$1 - \frac{1}{6}[0.58 + (4 \times 0.49) + 0.37] = 0.515.$$

From (3), the required number of individuals is

$$n = \frac{250}{0.515} = 486,$$

and so nearly 500 individuals will need to be recruited to the study over the accrual period of 18 months. This demands a recruitment rate of about 28 individuals per month.

If it is only expected that 15 individuals can be found each month, the accrual period will need to be extended to ensure that there is a sufficient number of individuals to give the required number of deaths. The number of individuals that could be recruited in a period of a months would be $15a$. Various values of a can then be tried in order to make this approximately equal to the value obtained from (3). For example, if we take $a = 30$ and continue with $f = 24$, then the probability of death during the study is

$$\Pr(\text{death}) = 1 - \frac{1}{6}[\overline{S}(24) + 4\overline{S}(39) + \overline{S}(54)].$$

From Figure 1, the survivor functions for individuals on each treatment at 24, 39, and 54 months can be estimated, and we find that $\overline{S}(24) = 0.58$, $\overline{S}(39) = 0.46$, and $\overline{S}(54) = 0.27$. The probability of death now turns out to be 0.552, and the required number of individuals to give 250 deaths is 453. This would be consistent with an estimated recruitment rate of 15 per month.

Now suppose that it is decided that the study will not have a follow-up period, so that the accrual period is equal to the duration of the study. If the accrual period is taken to be 30 months, so that $a = 30$ and $f = 0$, then the probability of death is given by

$$\Pr(\text{death}) = 1 - \frac{1}{6}[\overline{S}(0) + 4\overline{S}(15) + \overline{S}(30)].$$

Now, $\overline{S}(0) = 1.00$, $\overline{S}(15) = 0.73$, and $\overline{S}(30) = 0.51$, and the probability of death is 0.262. The required number of individuals is now $250/0.262 = 955$, and this would just about be met by a recruitment rate of 32 individuals per month. This shows that the absence of a follow-up period leads to an increase in the number of individuals that must be entered into the study.

References

[1] Bernstein, D. & Lagakos, S.W. (1978). Sample size and power determination for stratified clinical trials, *Journal of Statistical Computation and Simulation* **8**, 65–73.

[2] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd ed., Chapman & Hall/CRC, Boca Raton.

[3] Freedman, L.S. (1982). Tables of the number of patients required in clinical trials using the logrank test, *Statistics in Medicine* **1**, 121–129.

[4] Machin, D. & Campbell, M.J. (1987). *Statistical Tables for the Design of Clinical Trials*. Blackwell, Oxford.

[5] Schoenfeld, D.A. (1981). The asymptotic properties of comparative tests for comparing survival distributions, *Biometrika* **68**, 316–319.

[6] Schoenfeld, D.A. (1983). Sample size formula for the proportional-hazards regression model, *Biometrics* **39**, 499–503.

[7] Schoenfeld, D.A. & Richter, J.R. (1982). Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint, *Biometrics* **38**, 163–170.

D. COLLETT

Sample Size Determination

Sample size determination refers to the evaluation of the sample size desired or required for a study during the design stage, before data are collected. Such evaluations are based on a consideration of the operating characteristics of the statistical procedures to be employed in the ultimate statistical analysis of the study data. Invariably these operating characteristics depend in part on the sample size.

For example, a **cross-sectional** survey of the prevalence of diabetes (diagnosed or undiagnosed) among native Americans would require a sample size of 1421 to allow **estimation** of the **prevalence** to within a precision of ± 0.02 with 90% confidence, assuming a true prevalence no larger than 30%. Also, a randomized **clinical trial** of a new drug treatment vs. placebo for congestive heart failure would require 652 patients for a two-sided test at $\alpha = 0.05$ to provide 90% **power** to detect a 30% reduction in the risk of mortality after 1 year of follow-up, assumed to be no greater than 40% in the population (*see* **Alternative Hypothesis**; **Level of a Test**; **Null Hypothesis**).

In the first example, the sample size was determined based on the desired *precision* of the **confidence interval** estimate of the prevalence (probability) of the disease on sampling from a large population. In the second example, the sample size determination was based on the desired *power* of the test for the difference between two proportions (probabilities).

In general, the required sample size for any study can be based on the operating characteristics of the statistical procedures to be applied in the analysis of the data. For simple statistical estimates and tests, these relationships can be defined explicitly, for example by the derivation of the expression for the power function of a particular statistical test. In more complex analyses, such as **regression** models, these characteristics can be assessed indirectly through **simulation**, assuming an appropriate population model under which the sampling in the actual study is expected to be performed.

In all cases, the adequacy of the sample size determined depends on the accuracy of the initial specifications of the assumed parameters in the population. For example, if the true prevalence of diabetes in the target population is closer to 50%, then a sample size of 1421 will provide 87% confidence for a precision of ± 0.02. For the purposes of planning a study, therefore, it is always advisable that one consider a range of population parameters over which the operating characteristics are assessed for specific sample sizes.

Since the application of procedures for sample size evaluation is universal to all realms of statistical methods, the literature on this topic is indeed vast. McHugh & Le [11] provide a review of sample size determination for commonly used statistical procedures in biostatistical practice from the perspective of the precision of an estimator. Lachin [7] and Donner [3], among others, likewise provide a review of sample size determination for commonly used statistical tests from the perspective of the power of the test. General texts on the topic include [10, 1, 2, 15], and [12], among others. Many reference texts on general statistical methods, epidemiologic methods, and clinical trials include descriptions of sample size evaluation, or of the operating characteristics of an estimator or test based on sample size. The following presents an introduction to the basic concepts for statistical procedures which are commonly used in biostatistical practice.

Estimation Precision

Consider that we wish to estimate a parameter θ in a large (infinite) population based on a simple **random sample**. Assume that we plan to employ an estimator $\hat{\theta}$ which is normally distributed, at least asymptotically, as $\hat{\theta} \sim \mathrm{N}(\theta, \Sigma^2)$ where Σ^2 is some function of sample size N, such as $\Sigma^2 = \sigma^2/N$, where σ^2 is a **variance component**. Then the following apply:

1. The $1 - \alpha$ confidence interval (CI) for θ is of the form $\hat{\theta} \pm e_\alpha$, where

$$e_\alpha = z_{1-\alpha/2}\Sigma \tag{1}$$

   is the precision of the estimate at level $1 - \alpha$ and $z_{1-\alpha/2}$ is the upper two-sided **standard normal deviate** at level $1 - \alpha/2$. Since Σ is a function of the sample size N, then so also is the precision of the estimate e_α.

2. For a confidence interval $\hat{\theta} \pm e$ with a given degree of precision e, the corresponding level of confidence $1 - \alpha$ is provided by the standardized deviate $z_{1-\alpha/2} = e/\Sigma = \sqrt{N}e/\sigma$. This allows one to evaluate the relationship between the level of confidence and the degree of precision of the estimate for different sample sizes.
3. Solving for N, the sample size required to provide an interval estimate with precision $\pm e$ at confidence level $1 - \alpha$ for a given variance component σ is given by

$$N = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2. \tag{2}$$

For example, the simple proportion with a characteristic of interest, say p, from a sample of N observations is asymptotically distributed as $P \sim \mathrm{N}[\pi, \pi(1-\pi)/N]$, where π is the probability of the characteristic in the population and the **large-sample** variance of p is $\Sigma^2 = \pi(1-\pi)/N$ with $\sigma^2 = \pi(1-\pi)$ (*see* **Binomial Distribution**). The sample size required to estimate a probability assumed to be less than $\pi = 0.3$ (or greater than $\pi = 0.7$) with precision $e = 0.02$ at 90% confidence is provided as

$$N = \left[\frac{z_{1-\alpha/2}[\pi(1-\pi)]^{1/2}}{e}\right]^2$$
$$= \left[\frac{(1.645)[(0.3)(0.7)]^{1/2}}{0.02}\right]^2 = 1421.$$

(Throughout, all calculations of N are rounded up to the next whole integer.)

It may be judged that this sample size is too large. However, using the first relationship, one can determine that a 90% confidence interval with $N = 1000$ provides a degree of precision of $e = 1.645[(0.3)(0.7)]^{1/2}/\sqrt{1000} = 0.024$. Alternatively, using the second relationship, one could show that a sample size of $N = 1000$ provides 83% confidence of estimating π with a precision of ± 0.02, where the corresponding normal deviate is $1.38 = 0.02\sqrt{1000}/[(0.3)(0.7)]^{1/2}$.

Such computations, however, assume that the variance is known a priori, or in this case that the probability is known. These developments can be generalized to allow for sampling variation in the estimated variance, and thus in the precision of the estimate. This leads to expressions which determine the sample size needed to provide probability $1 - \beta$ that the realized $1 - \alpha$ confidence interval will have a precision of no greater than e (cf. [5]).

Clearly, such computations could be applied to a variety of estimation problems. Sample size determination based on the precision of an estimate will not be considered further, in part because the above expressions are analogous to those obtained from a consideration of the power of the corresponding test (see below). Those interested are referred to the article by McHugh & Le [11], or to texts on sampling.

Power

The power of a statistical test refers to the probability that a statistically significant test statistic will be obtained in a study under a specific hypothesis. In any statistical test, one sets out to assess the probability of the data under the null hypothesis H_0, commonly termed the ***P* value**. Usually "the data" are summarized in the form of an estimate of a parameter or a **sufficient statistic** for a parameter, and statistical significance is declared if the resulting P value is less than the a priori stated significance level.

Normal Theory Population Model

The simplest case is that of a statistic, say T, which is normally distributed, at least asymptotically, and for which the mean, μ, and variance, Σ^2, can be specified under the null hypothesis $\mathrm{H}_0 : \mu = \mu_0$ and under an alternative hypothesis $\mathrm{H}_1 : \mu = \mu_1 \neq \mu_0$. In many cases, the variance of the statistic will depend on the mean value, such that the variance under the **null hypothesis**, Σ_0^2, differs from that under the **alternative hypothesis**, Σ_1^2. Thus $T \sim \mathrm{N}(\mu_0, \Sigma_0^2)$ under H_0 and $T \sim \mathrm{N}(\mu_1, \Sigma_1^2)$ under H_1, with Σ_1^2 possibly $\neq \Sigma_0^2$.

The test of $\mathrm{H}_0 : \mu = \mu_0$ is based on the usual z-test $z = (\mathrm{T} - \mu_0/\Sigma_0$, where $z \sim \mathrm{N}(0, 1)$ under H_0. The null hypothesis is rejected against a one-sided alternative in the lower tail ($\mathrm{H}_1 : \mu_1 < \mu_0$) if the obtained z-test value has a corresponding P value $= \Phi(z) \leq \alpha$ at significance level α (one-sided). Likewise, H_0 is rejected against a one-sided alternative in the upper tail ($\mathrm{H}_1 : \mu_1 > \mu_0$) if the corresponding P value $= 1 - \Phi(z) < \alpha$; H_0 is rejected against a two-sided alternative when the two-sided P value $= 2[1 - \Phi(|z|)] \leq \alpha$. The choice of the alternative, one- vs. two-sided, is based on the nature of the questions to be addressed or the desired information, not

the expected direction of the result (*see* **Alternative Hypothesis**).

General Expressions for Power

In this setting, basic expressions for the power function of the statistical test are readily derived. Under $H_1 : T \sim N(\mu_1, \Sigma_1^2)$, and thus

$$Z = \frac{T - \mu_0}{\Sigma_0} \sim N\left[\frac{\mu_1 - \mu_0}{\Sigma_0}, \frac{\Sigma_1^2}{\Sigma_0^2}\right]. \quad (3)$$

The expected value of the test statistic under the alternative is termed the noncentrality parameter, Δ, where in this case $\Delta = (\mu_1 - \mu_0)/\Sigma_0$.

From the distribution under the alternative, the power of the test is obtained as follows. Power $1 - \beta$ is the complement of the probability of a type II error, β, of failing to reject the null hypothesis H_0 when an alternative hypothesis is true. Consider the case where $\mu_1 > \mu_0$ and a one-sided upper-tail test is conducted. Then $\beta = \Pr(z < z_{1-\alpha}|\mu_1, \Sigma_1^2)$ and

$$\beta = \Phi\left[\frac{z_{1-\alpha} - \left((\mu_1 - \mu_0) \;/\; \Sigma_0\right)}{\Sigma_1 \;/\; \Sigma_0}\right]. \quad (4)$$

Thus, $\beta = \Phi(z_\beta)$, where z_β is the standard normal deviate corresponding to the probability of a type II error β:

$$z_\beta = \frac{z_{1-\alpha}\Sigma_0 - (\mu_1 - \mu_0)}{\Sigma_1}. \quad (5)$$

Since this is a lower-tail probability, then the upper-area probability corresponding to the level of power is obtained from $z_{1-\beta} = -z_\beta$, so that

$$z_{1-\beta} = \frac{(\mu_1 - \mu_0) - z_{1-\alpha}\Sigma_0}{\Sigma_1}. \quad (6)$$

To allow for a one-sided alternative of the form $H_1 : \mu_1 < \mu_0$, or a two-sided alternative, we employ $|\mu_1 - \mu_0|$. This leads to the general expression for the relationship of power to the noncentral distribution of the test statistic:

$$|\mu_1 - \mu_0| = z_{1-\alpha}\Sigma_0 + z_{1-\beta}\Sigma_1, \quad (7)$$

where $z_{1-\alpha/2}$ is employed for a two-sided test. Some authors have used an abbreviated notation z_α and z_β for $z_{1-\alpha}$ and $z_{1-\beta}$, respectively.

In this formulation, T can usually be defined such that the variances can be factored of the form $\Sigma_i^2 = \sigma_i^2/N$, where σ_i^2 is the variance component ($i = 0, 1$) and N is the total sample size. Substituting into the above yields

$$\sqrt{N}|\mu_1 - \mu_0| = z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_1. \quad (8)$$

From this we can derive expressions to perform the following types of computations:

1. The total *sample size* N required to ensure a power of $1 - \beta$ of detecting a relevant difference $\mu_1 - \mu_0$ with a test at level α (or $\alpha/2$ if two-sided) is
$$N = \left(\frac{z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_1}{\mu_1 - \mu_0}\right)^2. \quad (9)$$
2. The *power* $1 - \beta$ to detect a difference $\mu_1 - \mu_0$ with a test at level α with a specific sample size N is provided by $\Phi(z_{1-\beta})$, where
$$z_{1-\beta} = \frac{\sqrt{N}|\mu_1 - \mu_0| - z_{1-\alpha}\sigma_0}{\sigma_1}. \quad (10)$$
Note that power $1 - \beta$ is 0.50 for $z_{1-\beta} = 0$, > 0.5 for $z_{1-\beta} > 0$, and < 0.5 for $z_{1-\beta} < 0$.
3. The *difference* $\mu_1 - \mu_0$ which can be detected with power $1 - \beta$ with a specified sample size N is provided by
$$|\mu_1 - \mu_0| = \frac{(z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_1)}{\sqrt{N}}. \quad (11)$$

In general, all three relationships may be used in planning the sample size required for a study. The latter relationships are also useful for the assessment of the power of a study post hoc, after it has been completed, especially in the event of a nonsignificant result.

Each of these relationships is a representation of the power function of the statistical test derived from (10). If the variance components under the null and alternative hypotheses are equal (at least approximately so), then (10) reduces to

$$z_{1-\beta} = \frac{\sqrt{N}|\mu_1 - \mu_0|}{\sigma} - z_{1-\alpha} = \sqrt{N}\theta - z_{1-\alpha}, \quad (12)$$

where $\theta = |\mu_1 - \mu_0|/\sigma$ is the standardized difference, termed the *noncentrality factor*.

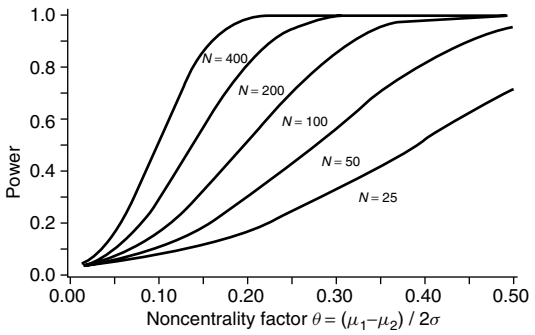

Figure 1 Power curves for the test of a difference in means for $\alpha = 0.05$ (two-sided)

Figure 1 presents the power to detect a difference for a test statistic with various sample sizes as a function of θ. In all cases, for a difference of 0, the probability of a significant test result is simply α, the significance level under the null hypothesis. As the magnitude of the difference increases under the alternative, the power increases, reaching an asymptote at 1.0 as θ approaches ∞. The rate at which power increases is a function of the sample size. Similar power function curves could be computed for any test statistic based on the more precise equation (10) employing the variance components under the null and alternative hypotheses if they differ.

The equation for the determination of sample size (9) can be viewed as the determination of the N for which the corresponding power function intersects the point $(\theta, 1-\beta)$ for a given level α and standardized difference θ, or the point $(\mu_1 - \mu_0, 1-\beta)$ for a specific difference $\mu_1 - \mu_0$ with variance components σ_1 and σ_0. For example, suppose we wish to detect a difference of $|\mu_1 - \mu_0| = 2$ with $\sigma = 8.7 (\sigma_0^2 \cong \sigma_1^2)$, yielding $\theta = 0.23$. Referring to Figure 1, the power function for $N = 200$ provides $\sim 90\%$ power to detect this difference. This is the same sample size that would be provided by a direct computation using (9).

Sample size evaluation in practice, however, is usually an iterative process. In this example it may be that an N of 100 would be financially feasible. The power function (Figure 1) shows that $N = 100$ provides 63% power to detect the originally specified difference $\theta = 0.23$, but that it also provides 90% power to detect a difference of $\theta = 0.33$. If these operating characteristics are considered acceptable, then the smaller N of 100 might be employed.

Such power function curves are also important in evaluating the interpretation of a negative, nonsignificant test result after a study has been completed. For example, suppose that a sample size of only 50 was employed in this study, and that the final result is not significant. Since the power function for $N = 50$ (Figure 1) shows that this sample size has power of 0.90 or greater to detect a standardized difference $\theta = 0.46$, then one can safely conclude that a difference of this magnitude likely does not exist. However, there was less power to detect smaller differences, such as a power of 0.369 for $\theta = 0.23$. Thus, one can only conclude that the study may have failed to detect such differences due to lack of power.

Power and Precision

Most of the literature on sample size evaluation is described in terms of the power function of a statistical test rather than the precision of an estimate. However, there is a simple correspondence between the two approaches. Most z-tests can also be expressed as an estimation-based test of the form $Z = (T - \mu_0)/\Sigma_0$, where T is an estimate of the parameter of interest ($\theta = \mu_1 - \mu_0$, $\hat{\theta} = T$).

Thus, the expression relating N to the precision of the estimate is approximately the same as that relating N to the power of the test, where $\mu_1 - \mu_0$ is replaced by the precision e and $z_{1-\beta} = 0$, so that $e = z_{1-\alpha/2}\Sigma_0$. Therefore, the N derived using a $1-\alpha$ confidence interval to ensure a precision e is approximately the same as the N needed to detect a difference $\mu_1 - \mu_0 = e$ with power = 0.50 using a two-sided test.

However, when constructing a confidence interval, the variance under the alternative is employed, so that (1) is actually $e = z_{1-\alpha/2}\Sigma_1$. To derive the equivalent equations for the precision of an estimate from those for the power of a test, we would employ $z_{1-\alpha/2}\Sigma_1$ when $\Sigma_0 \neq \Sigma_1$, where Σ_1 (or σ_1) is the multiplier of $z_{1-\beta}$ in the sample size equation.

Example: Test for Two Proportions

For a simple **2 x 2 table**, the usual large-sample test is the Pearson contingency **chi-square test** on 1 **degree of freedom** (df) which is directed towards a two-sided alternative. The chi-square value is equal to the square of the usual z-test for two proportions, which is based on the large-sample normal

approximation to the binomial. Either a one- or a two-sided z-test can be performed. Given two samples of sizes n_e and n_c for the experimental and control treatments, respectively, in which x_e and x_c are positive for the characteristic of interest, the corresponding simple proportions are $p_e = x_e/n_e$ and $p_c = x_c/n_c$. Each is the maximum likelihood estimate of the corresponding population probabilities such that $E(p_e) = \pi_e$ and $E(p_c) = \pi_c$, and the large-sample variances of each proportion are $\text{var}(p_e) = \pi_e(1-\pi_e)/n_e$ and $\text{var}(p_c) = \pi_c(1-\pi_c)/n_c$. In the following, it is convenient to present these expressions in terms of the total sample size N and the sample fractions $Q_e = n_e/N$ and $Q_c = n_c/N$, where $Q_e + Q_c = 1.0$.

The null hypothesis is $H_0 : \mu = \mu_0 = (\pi_e - \pi_c) = 0$ such that $\pi_e = \pi_c = \overline{\pi}$. Under H_0 the MLE of $\overline{\pi}$ is $\overline{p} = (x_e + x_c)/(n_e + n_c)$. The test then is based on $T = p_e - p_c$, where $\mu_0 = 0$, and

$$\begin{aligned}\Sigma_0^2 &= \text{var}[(p_e - p_c)|H_0] \\ &= \frac{\overline{\pi}(1-\overline{\pi})}{N}\left(\frac{1}{Q_e} + \frac{1}{Q_c}\right) = \frac{\sigma_0^2}{N}, \end{aligned} \quad (13)$$

which for large samples, can be estimated as

$$\hat{\Sigma}_0^2 = \frac{\overline{p}(1-\overline{p})}{N}\left(\frac{1}{Q_e} + \frac{1}{Q_c}\right).$$

Thus the test for two proportions is of the form $Z = (p_e - p_c)/\hat{\Sigma}_0$.

Under an alternative hypothesis, $H_1 : \mu = \mu_1 = (\pi_e - \pi_c) \neq 0$, and thus

$$\begin{aligned}\Sigma_1^2 &= \text{var}[(p_e - p_c)|H_1] \\ &= \frac{1}{N}\left[\frac{\pi_e(1-\pi_e)}{Q_e} + \frac{\pi_c(1-\pi_c)}{Q_c}\right] \\ &= \frac{\sigma_1^2}{N}. \end{aligned} \quad (14)$$

Substituting these expressions into the general equations relating sample size to power yields

$$\begin{aligned}&\sqrt{N}|\pi_e - \pi_c| \\ &= z_{1-\alpha}\left[\overline{\pi}(1-\overline{\pi})\left(\frac{1}{Q_e} + \frac{1}{Q_c}\right)\right]^{1/2} \\ &\quad + z_{1-\beta}\left[\frac{\pi_e(1-\pi_e)}{Q_e} + \frac{\pi_c(1-\pi_c)}{Q_c}\right]^{1/2}. \end{aligned} \quad (15)$$

Solving for N, $z_{1-\beta}$ or $(\pi_e - \pi_c)$ provides the expressions needed to address the three types of questions described above.

When $n_e = n_c = N/2 (Q_e = Q_c = 1/2)$, this expression simplifies to

$$\begin{aligned}\sqrt{N}|\pi_e - \pi_c| &= z_{1-\alpha}[4\overline{\pi}(1-\overline{\pi})]^{1/2} \\ &\quad + z_{1-\beta}[2\pi_e(1-\pi_e) \\ &\quad + 2\pi_c(1-\pi_c)]^{1/2}. \end{aligned} \quad (16)$$

Furthermore, Lachin [7] shows that $\Sigma_0^2 \geq \Sigma_1^2$. Thus, it is conservative to use

$$\sqrt{N}|\pi_e - \pi_c| = (z_{1-\alpha} + z_{1-\beta})[4\overline{\pi}(1-\overline{\pi})]^{1/2}. \quad (17)$$

For example, suppose we wish to plan a study with two equal-sized groups ($n_e = n_c$) to detect a 30% reduction in mortality associated with congestive heart failure, where the 1-year mortality in the control group is assumed to be no greater than 0.40. Thus, $\pi_c = 0.40$ and $\pi_e = 0.28$ ($= 0.70 \times 0.40$). Under the null hypothesis $\overline{\pi} = 0.34$. We desire 90% power for a two-sided test for two proportions at $\alpha = 0.05$. Using (9) the required total N is obtained as

$$\begin{aligned}N &= \big[\{1.96[4(0.34)(0.66)]^{1/2} + 1.282[2(0.28)(0.72) \\ &\quad + 2(0.4)(0.6)]^{1/2}\} \,/\, (0.4 - 0.28)\big]^2 \\ &= 652.\end{aligned}$$

Using the simplification which employs only the null variance component yields:

$$N = \left[\frac{[1.96 + 1.282][4(0.34)(0.66)]^{1/2}}{0.4 - 0.28}\right]^2 = 656.$$

Alternatively one could solve for $z_{1-\beta}$ to determine the power to detect a difference with a specified sample size, or the magnitude of the difference which could be detected with a given power for a specific sample size. For example, the power to detect this same difference with the smaller sample size $N = 500$ using the more complete equation (10) is provided by

$$\begin{aligned}z_{1-\beta} &= \frac{(500)^{1/2}(0.4 - 0.28) - 1.96[4(0.34)(0.66)]^{1/2}}{[2(0.28)(0.72) + 2(0.4)(0.6)]^{1/2}} \\ &= 0.879,\end{aligned}$$

yielding 81% power.

Finally, we note that the above expressions also provide the basic elements for the corresponding expression relating sample size to the precision of the estimate of the difference in probabilities. For example, to provide a 95% $(1-\alpha)$ confidence interval for $\pi_e - \pi_c$ with precision ± 0.05 requires a total sample size

$$N = \left(\frac{z_{1-\alpha/2}\sigma_1}{e}\right)^2$$
$$= \left[\frac{(1.96)[2(0.28)(0.72) + 2(0.4)(0.6)]^{1/2}}{0.05}\right]^2$$
$$= 1358.$$

Simplifications

As suggested by (17), the expressions for sample size and power can be simplified by using only one variance component rather than that under the null and under the alternative in the cases where the two differ, as where the variance depends on the expectation. Referring to (8), if $\sigma_0 \cong \sigma_1 \cong \sigma$, then the basic equation becomes

$$\sqrt{N}\frac{|\mu_1 - \mu_0|}{\sigma} = z_{1-\alpha} + z_{1-\beta}. \tag{18}$$

The left-hand side is the noncentrality parameter of the noncentral distribution of the test statistic; see (12). Factoring $\sqrt{N}$, the remainder is termed the noncentral factor θ,

$$\theta = \frac{|\mu_1 - \mu_0|}{\sigma}, \tag{19}$$

such that

$$\sqrt{N}\theta = z_{1-\alpha} + z_{1-\beta}. \tag{20}$$

This leads to simple equations for sample size and power as a function of the noncentral factor:

$$N = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\theta}\right)^2, \tag{21}$$

$$z_{1-\beta} = \theta\sqrt{N} - z_{1-\alpha}. \tag{22}$$

Using these expressions, Lachin [7] presents tables which give the sample sizes required to detect a range of values of θ for different levels of α and β, and the power to detect specific values of θ for a range of sample sizes. Other articles and texts likewise present various tables for sample size and power. In practice, however, direct computation using the appropriate expressions is readily performed.

Variance Components and Noncentral Factors

Within this framework, it is relatively straightforward then to derive the basic equations for sample size or power from the expected value and variance of the test statistic under the null and alternative hypotheses for a test statistic which is normally distributed, at least asymptotically. Expressions can be further simplified by noting that, for two independent group problems with equal sample sizes ($Q_e = Q_c = 0.5$), the term in Σ_0 involving the sample fractions is $(Q_c^{-1} + Q_e^{-1})^{1/2} = 2$. The following is a summary for many common tests, assuming equal sample sizes in the two groups. Below, a simple adjustment is described for the case of unequal sample sizes. In each of the following cases, $\mu_0 = 0$ under H_0.

For a z-test of means between two groups, the test statistic is based on the sample mean, which is assumed to be normally distributed as $N(\nu_i, \gamma^2/N)$ within each group $i = e, c$ with common variance γ^2 between subjects in each group. Then $\mu_1 = (\nu_e - \nu_c)$, $\sigma_0^2 = \sigma_1^2 = 4\gamma^2$, and $\theta = |\nu_e - \nu_c|/2\gamma$. For the paired z-test for means, the statistic is based on the mean of N paired differences assumed to be distributed as $N(\nu, \gamma^2/N)$. Then $\mu_1 = \nu \neq 0$ and $\theta = |\nu|/\gamma$.

The expressions for the z-test for proportions in two independent groups are presented in (15)–(17). From (17), the noncentral factor is $\theta = |\pi_e - \pi_c|/[4\overline{\pi}(1-\overline{\pi})]^{1/2}$. In the case of paired or matched observations, the test (**McNemar test**) is based on the **multinomial** parameters for the 2×2 table with discordant probabilities π_{01} and π_{10} for the matched assessments in N pairs with binary measurements (0 or 1) for each pair member. Thus, for example, π_{01} is the proportion of the N pairs where the first pair member has measurement 0 and the second member a 1. Then $H_0 : \pi_{10} = \pi_{01} = \overline{\pi}$, and under H_1, $\mu_1 = \pi_{10} - \pi_{01} \neq 0$. Various expressions for the power of this test have been proposed for the unconditional case where the number of discordant pairs is not fixed a priori by design. Lachin [8] concludes that it is conservative to use that obtained from the underlying multinomial probabilities where

$$\sigma_1^2 = [(\pi_{10} + \pi_{01}) - (\pi_{10} - \pi_{01})^2] \tag{23}$$

and $\sigma_0^2 = 2\overline{\pi}$, where $\sigma_0 \geq \sigma_1$. Thus the noncentral factor is $\theta = |\pi_{10} - \pi_{01}|/(2\overline{\pi})^{1/2}$.

Although nonparametric rank tests are generally used for the analysis of survival (event-time) data (*see* **Survival Analysis, Overview**), calculation of sample size and power are often performed assuming some simple parametric model. The Mantel or **logrank test** is the most commonly used test in this setting, which is asymptotically fully efficient against a **proportional hazards** or **Lehmann alternative**. The simplest parametric form of this model is the **exponential** model with constant hazard rates λ_e and λ_c over time in each group. Some authors have derived expressions for the power function of the test for exponential hazards using the normal approximation to the distribution of $\hat{\lambda}$ (e.g. [9]), while others have used that for $\ln \hat{\lambda}$ (e.g. [14]). The resulting computations are nearly equivalent, but those using the distribution of $\ln \hat{\lambda}$ are preferred.

Asymptotically, the sample estimate of the natural log **hazard rate** $\ln \hat{\lambda}$ is distributed as $N[\ln \lambda, 1/E(D|\lambda)]$. Thus, the power of the test depends on the expected total number of events $E(D|\lambda)$ to be observed during the study. Here $E(D|\lambda) = NE(\delta|\lambda)$, where δ is a binary variable (*see* **Dummy Variables**) representing observation of the event ($\delta = 1$) vs. not (right **censoring** of the event time, $\delta = 0$), and $E(\delta|\lambda)$ is the probability that the event will be observed as a function of λ and the total exposure of the cohort (patient years of follow-up). Under $H_0 : \lambda_e = \lambda_c = \overline{\lambda}$, or $\ln(\lambda_e/\lambda_c) = 0$, while under H_1, $\mu_1 = (\ln \lambda_e - \ln \lambda_c)$ and

$$\sigma_1^2 = \left[\frac{2}{E(\delta|\lambda_e)} + \frac{2}{E(\delta|\lambda_c)}\right]. \tag{24}$$

Thus $\sigma_0^2 = 4/E(\delta|\overline{\lambda})$. Lachin [7] shows that in this case $\sigma_1 \geq \sigma_0$, so that it is conservative to use $\theta = [\ln(\lambda_e/\lambda_c)]/\{2[E(\delta|\lambda_e)^{-1} + E(\delta|\lambda_c)^{-1}]\}^{1/2}$.

In a study with no censoring of event times, where the event times of all subjects are observed, then $E(\delta|\lambda) = 1$. In the case where each subject is followed for T years of exposure, then $E(\delta|\lambda) = 1 - \exp(-\lambda T)$. In a study with uniform entry over a recruitment interval of R years (*see* **Staggered Entry**), and a total study duration of T years ($T \geq R$),

$$E(\delta|\lambda) = \left[1 - \frac{\exp[-\lambda(T-R)] - \exp(-\lambda T)}{\lambda R}\right] \tag{25}$$

(*see* **Sample Size Determination in Survival Analysis**).

Unequal Sample Fractions

In each of the above cases, the variance component under the null or alternative hypothesis is used to define the noncentrality parameter. For two independent group problems with unequal sampling fractions $Q_e \neq Q_c$, the term in Σ involving the sample fractions is $(Q_c^{-1} + Q_e^{-1})^{1/2} \neq 2$. Thus, the Σ for unequal sample sizes equals $C = (Q_c^{-1} + Q_e^{-1})^{1/2}/2$ times that for equal sample sizes, so that the factor θ with equal sample sizes should be multiplied by $1/C$. Thus, from (21), the sample size for the unbalanced design is $C^2 = (Q_c^{-1} + Q_e^{-1})/4$ times that for equal sample fractions. For example, for Q_e (or Q_c) $= 0.7$, the sample size required for the unbalanced design is $(1/0.7 + 1/0.3)/4 = 1.19$ times the N for the balanced design, so that the unbalanced design requires approximately 19% larger sample size to provide the same level of power as the balanced design.

Power for t, χ^2 and F-Tests

For test statistics which follow the **Student's *t***, **chi-square** or ***F* distributions**, among others, determination of sample size or power is conducted through identification of the noncentrality parameter, $\Delta = N\theta^2$, of the noncentral distribution of the test statistic. For a 1 df χ^2 test statistic, the noncentrality parameter is $N\theta^2$, where θ is the noncentral factor for the test. Thus, from (20), the value of the noncentrality parameter which provides power $1 - \beta$ for a $u = 1$ df two-sided test at level α, designated as $\Delta(1, \alpha, \beta) = N\theta(1, \alpha, \beta)^2 = (z_{1-\alpha/2} + z_{1-\beta})^2$. For example, $\Delta(1, 0.05, 0.10) = (1.96 + 1.2816)^2 = 10.507$.

Values of the noncentrality parameter providing various levels of power for the noncentral chi-square and **noncentral *t* distributions** on u df, $\Delta(u, \alpha, \beta)$, and for the F-distribution on u and v df, $\Delta(u, v, \alpha, \beta)$, are widely tabulated. Programs are also available, such as the SAS functions PROBCHI for the cumulative probabilities and CINV for quantiles of the chi-square distribution, both of which provide computations under the noncentral distribution. Equivalent functions provide these computations for the t distribution (PROBT and TINV), and the F distribution (PROBF and FINV). SAS functions CNONCT, TNONCT, and FNONCT, now available in release 6.07, provide the values of the required noncentrality parameter Δ for specific levels of α and β [4] (*see* **Software, Biostatistical**).

To determine sample size using this approach, one first obtains the value of the noncentrality parameter which will provide the desired level of power, e.g. the value $\Delta(u, \alpha, \beta)$ for the noncentral chi-square distribution. One then evaluates the value of the noncentrality factor θ under the alternative hypothesis. The noncentrality factor is usually defined using the variance under the null hypothesis, often because the expected value of the statistic (the noncentrality parameter) is derived under a sequence of local alternatives. Given the value of θ, the N required to provide power $1 - \beta$ is that value for which $\Delta(u, \alpha, \beta) = N\theta^2$, yielding $N = \Delta(u, \alpha, \beta)/\theta^2$.

Example: Contingency Chi-Square Test

Consider the case of a random sample of N subjects divided among K mutually exclusive categories (cells) with cell frequency x_i in the ith category $l = 1, \ldots, K$. These frequencies are distributed as multinomial with cell probabilities $\pi_l, l = 1, \ldots, K$. These cell probabilities, in turn, can be expressed as a function of M underlying parameters, expressed as $\pi_l(\boldsymbol{\alpha})$, for $\boldsymbol{\alpha} = (\alpha_1 \ldots \alpha_M)$. Under a null hypothesis stated in terms of the $\boldsymbol{\alpha}$, $\mathrm{H}_0 : \boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, with corresponding multinomial probabilities $\pi_l(\boldsymbol{\alpha}_0)$, where the parameters $\boldsymbol{\alpha}$ are estimated from the sample, then the Pearson contingency chi-square statistic

$$\chi^2 = \sum_l \frac{[x_l - N\pi_l(\hat{\boldsymbol{\alpha}}_0)]^2}{N\pi_l(\hat{\boldsymbol{\alpha}}_0)} \tag{26}$$

is distributed as chi-square on $K - M - 1$ df. Under an alternative hypothesis, $\mathrm{H}_1 : \boldsymbol{\alpha} = \boldsymbol{\alpha}_1$, χ^2 is distributed as noncentral chi-square with noncentrality parameter $N\theta^2$, where

$$\theta^2 = \sum_l \frac{[\pi_l(\boldsymbol{\alpha}_1) - \pi_l(\boldsymbol{\alpha}_0)]^2}{\pi_l(\boldsymbol{\alpha}_0)} \tag{27}$$

The most common instance is the $r \times c$ **contingency table** ($K = rc$) under the hypothesis of independence, where $\sum_l$ denotes summation over rows $i = 1, \ldots, r$ and columns $j = 1, \ldots, c$. Then π_{ij} is the probability associated with the ijth cell, with marginal probabilities $\pi_{i.}$ and $\pi_{.j}$ for the ith row and jth column, respectively. The hypothesis of independence then implies the null hypothesis that $\pi_{ij} = \pi_{i.}\pi_{.j}$. Thus, the parameters under the null hypothesis are $\boldsymbol{\alpha} = [\pi_{1.} \ldots \pi_{(r-1).}\pi_{.1} \ldots \pi_{.(c-1)}]$, consisting of $M = (r - 1) + (c - 1)$ parameters to be estimated from the sample. The resulting test is based on $K - M - 1 = rc - (r - 1) - (c - 1) - 1 = (r - 1)(c - 1)$ df.

One way that such a contingency table might arise is the $r \times c$ comparative trial in which r independent groups of sample sizes $n_i = Q_i N$ are compared with respect to the proportions $p_{j(i)}$ in each of c categories, where $p_{j(i)} = x_{ij}/n_i$ and $\mathrm{E}(p_{j(i)}) = \pi_{j(i)}$, $\sum_j p_{j(i)} = \sum_j \pi_{j(i)} = 1.0$ for each group $i = 1, \ldots, r$. The null hypothesis of homogeneity is $\mathrm{H}_0 : \pi_{j(i)} = \alpha_j$ for all i. Under the alternative, these conditional probabilities differ across groups such that under $\mathrm{H}_1 : \pi_{j(i)} = \alpha_j + \delta_{ij}$, where $\delta_{ij} \neq 0$ for some (i, j). Then, Lachin [6] shows that

$$\theta^2 = \sum_j \frac{1}{\alpha_j} \left[\sum_i Q_i \delta_{ij}^2 - \left(\sum_i Q_i \delta_{ij} \right)^2 \right]. \tag{28}$$

For example, let the following be the pattern of conditional probabilities $\{\alpha_j\}$ expected under H_0 for a planned clinical trial comparing three equal-sized ($Q_i = 1/3$) treatment groups with respect to three categories of recovery:

	Complete recovery	Partial recovery	No recovery
$\alpha_j = \sum_i Q_i \pi_{j(i)} =$	0.10	0.15	0.75

Under the alternative, assume we wish to detect the following pattern of differences (δ_{ij}) among the three groups, where $\pi_{j(i)} = \alpha_j + \delta_{ij}$:

$$\{\delta_{ij}\} = \begin{cases} \text{Placebo} & -0.09 & -0.11 & 0.20 \\ \text{Low dose} & -0.01 & 0.01 & 0.00 \\ \text{High dose} & 0.10 & 0.10 & -0.20 \end{cases}$$

Substituting these values into (28) yields $\theta^2 = 0.1456$. From tables of the noncentral chi-square distribution for $\alpha = 0.05$, $\beta = 0.10$, and 4 df, or using the SAS function CNONCT, we require $\Delta(4, 0.05, 0.10) = 15.405$. Solving for $N = \Delta/\theta^2$ yields $N = 106$.

To determine power based on a given sample size N, one simply determines the value of the noncentrality parameter $\Delta = N\theta^2$ and then evaluates the cumulative probability at the critical value under the noncentral chi-square distribution. This approach is employed in the following example.

Example: K-Group ANOVA

Consider the case of the one-way **analysis of variance** (ANOVA) for the test of equality of K independent group means, $H_0 : \nu_1 = \nu_2 = \cdots \nu_K$ vs. $H_1 : \nu_i \neq \nu_j$ for some two groups $1 \leq i < j \leq K$, assuming a common variance σ^2 within groups under the null and alternative hypotheses. Also assume equal sample size n within each group, with a total sample size $N = nK$. Then the K group ANOVA F-test $= MSB/MSE$ on $u = K - 1$ and $v = K(n-1)$ df follows a noncentral F-distribution, where MSB and MSE are the mean squares between and within groups, respectively.

The expected mean squares between groups $\mathrm{E}(MSB)$ and within groups or for error $\mathrm{E}(MSE)$ under the alternative hypothesis that some differences exist among the K population means are:

$$\mathrm{E}(MSB) = \sigma_\varepsilon^2 + \frac{n\sum_i (\nu_i - \overline{\nu})^2}{K-1} = \sigma_\varepsilon^2 + n\sigma_\nu^2, \quad (29)$$

$$\mathrm{E}(MSE) = \sigma_\varepsilon^2.$$

Under the null hypothesis, $\sigma_\nu^2 = 0$ and $\mathrm{E}(MSB) = \sigma_\varepsilon^2$, so that $(K-1)MSB/\sigma_\varepsilon^2$ is distributed as central chi-square. Under the alternative, however, $(K-1)MSB/\sigma_\varepsilon^2$ is distributed as noncentral chi-square with noncentrality parameter $\Delta = n\theta^2$, where

$$\theta^2 = \frac{\sum_i (\nu_i - \overline{\nu})^2}{\sigma_\varepsilon^2}. \quad (30)$$

Thus, under the alternative, the F-test is also distributed as noncentral F with noncentrality parameter $\Delta = n\theta^2$. The values of Δ for given (k, n, α, β) have been tabulated. Some charts and tables present these relationships in terms of the reparameterization $\phi = [\Delta/(u+1)]^{1/2}$, where u is the numerator df$(= K - 1)$.

For example, for $K = 3$ groups, assume that the population means under the alternative hypothesis are $(\nu_1 = 2,\ \nu_2 = 4,\ \nu_3 = 6)$, so that $\sum_i (\nu_i - \overline{\nu})^2 = 8$. Also assume that $\sigma_\varepsilon^2 = 15$. Then $\theta^2 = 8/15 = 0.533$. For $n = 25$ per group, the critical value for the F-test at the 0.05 level on 2 and 72 df, obtained as FINV (0.95, 2, 72), is 3.12391. The noncentrality parameter is $\Delta = (25)(8/15) = 13.333$. The type II error β is then the probability of an Fvalue $<$ 3.12391 on (2, 72) df with $\Delta = 13.33$. Using FPROB (3.12391, 2, 72, 13.33) yields $\beta = 0.09691$ and power $= 0.90309$.

For sample-size determination for an F-test or a t-test, since the denominator df depends on n, an iterative procedure is required. Charts are also widely available relating sample size to the noncentrality factor θ.

Example: Multiple Regression Model

In a **multiple linear regression** model with m **covariates** or **explanatory variables**, a variety of different tests may be conducted, such as an overall model test on p df and tests of each of the individual regression coefficients, each of which will have a different power function. The power for these tests depends on the total N, the residual variance σ_ε^2, and on the joint distribution of the m covariates. For example, in the homoscedastic normal errors model with m covariates (plus the constant term), $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}_\varepsilon^2$, where $\hat{\sigma}_\varepsilon^2$ is the MSE on $N - m - 1$ df. Then the test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$ (including the intercept) is provided by the quadratic form $F = \hat{\boldsymbol{\beta}}'(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})^{-1}\hat{\boldsymbol{\beta}}$ on $(m+1, N-m-1)$ df. Under $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$, F is distributed as noncentral F with noncentrality parameter $\Delta = \boldsymbol{\beta}'(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}})^{-1}\boldsymbol{\beta} = \mathrm{E}[\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}]/\sigma_\varepsilon^2 = \mathrm{E}(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y})/\sigma_\varepsilon^2$. Thus, to evaluate power or sample size a priori, it is necessary to specify the covariance matrix of $(\mathbf{Y}|\mathbf{X})$ to determine the noncentrality parameter for the test of regression for the model.

Similarly, the test for the jth individual coefficient in the model is $t = \hat{\beta}_j/\hat{\gamma}_j$, which is distributed as t on $N - m - 1$ df, where $\mathrm{var}(\hat{\beta}_j) = \sigma_j^2 = (\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}})_{jj}$, which involves the jj element of $(\mathbf{X}'\mathbf{X})^{-1}$. To determine the noncentrality parameter for the test again requires specification of the joint distribution of $\mathbf{X}$. For analyses involving quantitative covariates, therefore, it is rare that there is adequate prior information on the joint distribution of $(\mathbf{Y}|\mathbf{X})$ to evaluate the size of the noncentrality parameter. For this reason, some authors, e.g. Cohen [1], discuss power and sample size for arbitrarily defined "small" to "large" effect sizes in terms of the values of Δ.

Example: Logit Model

One case in which it is tractable to consider the evaluation of the noncentrality parameter is the logit

(**logistic regression**) model for binary covariates using the power function of the large sample Wald chi-square test (cf. [13]; *see* **Likelihood**). The model relates the logit of the probability of an index characteristic, π, to a linear function of the covariate vector $\mathbf{X}$, including a constant, of the form $\ln[\pi/(1-\pi)] = \mathbf{X}'\boldsymbol{\beta}$, where π is obtained from the inverse logit

$$\pi = \frac{\exp(X'\beta)}{[1+\exp(X'\beta)]}. \tag{31}$$

Suppose we wish to estimate this relationship for $i = 1, \ldots, K$ cells, where the ith cell has sample size $n_i = Nq_i$ (N the total sample size) and covariate vector $\mathbf{X}_i$ consisting of the constant and m discrete covariates with coefficients $(\beta_0, \beta_1, \ldots, \beta_m)$, β_0 being the intercept. For example, for s binary covariates, $K = 2^s$ and $m \leq K$. Within the ith cell, the observed proportion with the index characteristic is p_i. Since $\operatorname{var}\ln[p/(1-p)] = 1/[n\pi(1-\pi)]$ the parameters can be estimated through weighted **least squares** such that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y}$ and $\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_{\hat{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}/N$, where

$$\boldsymbol{\Omega} = \operatorname{diag}\left\{\frac{1}{[q_i\pi_i(1-\pi_i)]}\right\} \quad \text{and}$$

$$\boldsymbol{\Omega}^{-1} = \operatorname{diag}[q_i\pi_i(1-\pi_i)]. \tag{32}$$

Note that $\boldsymbol{\Omega}$ is directly obtained as a function of $\mathbf{X}_i$ and $\boldsymbol{\beta}$ through (31).

The Wald test for a linear hypothesis of the form $\mathrm{H}_0 : \mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$ for an $r \times (m+1)$ matrix $\mathbf{L}'$ is of the form

$$\chi^2 = \hat{\boldsymbol{\beta}}'\mathbf{L}(\mathbf{L}'\boldsymbol{\Sigma}\mathbf{L})^{-1}\mathbf{L}'\hat{\boldsymbol{\beta}} \tag{33}$$

on r df with noncentrality parameter

$$\begin{aligned}\Delta &= \boldsymbol{\beta}'\mathbf{L}(\mathbf{L}'\boldsymbol{\Sigma}L)^{-1}\mathbf{L}'\boldsymbol{\beta} \qquad (34)\\ &= N\boldsymbol{\beta}'\mathbf{L}[\mathbf{L}'(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{L}]^{-1}\mathbf{L}'\boldsymbol{\beta} = N\theta^2.\end{aligned}$$

Since θ^2 is a function of $\mathbf{L}$, $\mathbf{X}$, and $\boldsymbol{\beta}$, sample size and power can readily be obtained as described above.

For example, consider the design matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix},$$

representing effects for the intercept, three strata (2 df), and two treatments. The expected cell fractions are specified to be $\{q_i\} = (0.075, 0.075, 0.25, 0.25, 0.175, 0.175)$, which assume that there are equal sample sizes for each of the two treatment groups within the three strata, comprising 15%, 50%, and 35% of the total sample, respectively. Under the alternative hypothesis we specify $\boldsymbol{\beta}' = (1.099, -0.251, -0.480, 0.925)$. These are the values of β_j which correspond to a model where the odds ratio for treatment 1 vs. treatment 0 is $\exp(\beta_3) \cong 2.5$ and the associated probabilities for each cell are $\{\pi_i\} = (0.75, 0.923, 0.70, 0.882, 0.65, 0.770)$, such that the odds ratios within each stratum are 4.0, 3.2, and 1.8, respectively. For the test of $\mathrm{H}_0 : \beta_3 = 0$, with vector $\mathbf{L}' = (0 \quad 0 \quad 0 \quad 1)$, the value of the noncentral factor is $\theta^2 = 0.0342$. For a 1 df chi-square test at $\alpha = 0.05$, the noncentrality parameter $\Delta(0.05, 0.10, 1) = 10.5074$ provides power = 0.90. Thus, the total N required to provide 90% power for this test is $N = 10.5074/0.0342 = 307$.

For this and similar examples, one approach to specifying the model under the alternative is first to specify the $\{\pi_i\}$ and then determine the values of β which satisfy the model. This can readily be obtained by generating a set of cell frequencies summing to a large number, say 10 000 – the frequencies being proportional to the corresponding probabilities – and then fitting the logit model to obtain the values of β_j.

Factors which Affect Sample Size

In addition to the variance components and noncentral factors described herein, other features of the observed data may affect precision or power. Two of the more important are missing data and measurement errors.

Missing Data

Precision and power are directly related to the amount of information in the data. In many cases, **information** in the Fisherian sense is proportional to the total sample size. Thus, if one expects $M \times 100\%$ of the observations to be missing completely at random (purely by chance), then the sample size required should be adjusted upwards by the factor $1/(1-M)$. In some cases, however, such as rank tests for survival data, the information is not directly proportional

to N alone, but to other factors, such as the pattern of losses to follow-up over time. In such cases it is necessary to consider the extent to which the process by which the missing data is generated will impact on the power of the test (*see* **Missing Data in Clinical Trials**; **Missing Data in Epidemiologic Studies**).

Reliability of Measurements

In a simple measurement error model one can express the observed measurement as $X_i = \eta_i + \varepsilon_i$, where η_i is the true measurement and ε_i is a random measurement error with expectation 0 and variance σ_ε^2 independent of η_i. Thus, $\sigma_X^2 = \sigma_\eta^2 + \sigma_\varepsilon^2$ and the **reliability** of the measurements is reflected by $\rho = \sigma_\eta^2/\sigma_X^2$. Since $\sigma_X^2 = \sigma_\eta^2/\rho$, power decreases as ρ decreases and the required N increases. For example, if N is needed to provide a desired level of power to detect a given difference in the mean values of the true measurements, with variance component σ_η^2, then N/ρ is required to detect the same difference in the observed measurements. For example, for $\rho = 0.8$, the sample size required to detect a difference between population means is 25% greater than that for measures without error. This is often an important consideration since in many cases the reliability of measurements can be controlled within limits, such as for laboratory assessments (*see* **Measurement Error in Epidemiologic Studies**).

References

[1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. Laurence Erlbaum, Hillsdale.

[2] Desu, M.M. & Raghavarao, D. (1990). *Sample Size Methodology*. Academic Press, New York.

[3] Donner A. (1984). Approaches to sample size estimation in the design of clinical trials – a review, *Statistics in Medicine* **3**, 199–214.

[4] Hardison C.D., Quade D. & Langston R.D. (1986). Nine functions for probability distributions, in *SUGI Supplemental Library User's Guide, Version 5 Edition*, R.P. Hastings, ed. SAS Institute, Inc., Cary, pp. 385–393.

[5] Kupper, L.L. & Hafner, K.B. (1989). How appropriate are popular sample size formulas?, *American Statistician* **43**, 101–105.

[6] Lachin, J.M. (1977). Sample size determination for $r \times c$ comparative trials, *Biometrics* **33**, 315–324.

[7] Lachin, J.M. (1981). Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials* **2**, 93–114.

[8] Lachin, J.M. (1992). Power and sample size evaluation for the McNemar test with application to matched case-control studies, *Statistics in Medicine* **11**, 1239–1251.

[9] Lachin, J.M. & Foulkes, M.A. (1986). Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, non-compliance and stratification, *Biometrics* **42**, 507–519.

[10] Machin, D. & Campbell, M.J. (1987). *Statistical Tables for the Design of Clinical Trials*. Blackwell Scientific, Oxford.

[11] McHugh, R.B. & Le, C.T. (1984). Confidence estimation and the size of a clinical trial, *Controlled Clinical Trials* **5**, 157–163.

[12] Odeh, R.E. & Fox, M. (1991). *Sample Size Choice: Charts for Experiments with Linear Models*, 2nd Ed. Marcel Dekker, New York.

[13] Rochon, J. (1989). Application of the GSK method to the determination of minimum sample sizes, *Biometrics* **45**, 193–205.

[14] Rubinstein, L.V., Gail, M.H. & Santner, T.J. (1981). Planning the duration of a comparative clinical trial with losses to follow-up and a period of continued observation, *Journal of Chronic Diseases* **34**, 469–479.

[15] Schuster, J.J. (1990). *CRC Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, Boca Raton.

JOHN M. LACHIN

Sample Size in Epidemiologic Studies

Determining the number of subjects to be included in a study is a crucial step in designing the study and writing the protocol. To determine sample size, the study objectives have to be clearly defined, the general design (e.g. cohort, case–control) (*see* **Case–Control Study**; **Cohort Study**) and specific design options have to be selected, the main outcome and exposure variables have to be specified, the planned analysis strategy (i.e. hypothesis testing or estimation) and statistical methods have to be determined. Therefore, sample size determination is a very important aspect of design and cannot be carried out without a thorough and quantitative understanding of the planned study.

The study sample size should be large enough that the estimates will be sufficiently precise and the difference of interest is likely to be detected. It is usually true that the more subjects that are included in the study, the better the precision of the estimates, and the more likely the difference of interest will be detected. However, an oversized study may not always be the best choice because of economic and study time (sometimes ethical) considerations.

Hypothesis Testing and Power: The Case of a Normally Distributed Outcome

The principle of sample size and power considerations can be illustrated by a simple hypothesis test for normally distributed data. The normal case study will build the concepts and form the mathematical basis for most other sample size procedures, which will be discussed in later sections.

Suppose n samples are drawn from a population that has a normal distribution $\mathrm{N}(\mu, \sigma^2)$. To test a null hypothesis $\mathrm{H_0}$: $\mu = \mu_0$ vs. an alternative hypothesis $\mathrm{H_a}$: $\mu = \mu_a (> \mu_0)$, we use a test statistic $Z = \sqrt{n}(\overline{X} - \mu_0)/\sigma_0$, which follows a standard normal distribution under $\mathrm{H_0}$. For simplicity, we assume that the variances are known: under $\mathrm{H_0}$, $\sigma^2 = \sigma_0^2$ and under $\mathrm{H_a}$, $\sigma^2 = \sigma_a^2$. The null hypothesis will be rejected if the observed value of Z falls in an extreme region, i.e. $Z > c$, where c is a constant to be determined (see below).

Two types of error can be made with the test. First, a type I error is that the null hypothesis is true but the observed Z falls in the rejection region (i.e. $Z > c$) such that the null hypothesis is rejected. The type I error is also called the significance level of the test. It is often protected by setting an upper limit, e.g. 0.10 or 0.05, for the significance level. Secondly, a type II error is the probability of failing to reject the null hypothesis when the alternative hypothesis is true. Both error rates depend on the sample size, test statistic Z, and the critical value c. In the normal distribution, for a given $\mathrm{H_a}$: $\mu = \mu_a > \mu_0$ (one-sided test),

$$\text{Type I error} : \alpha = P(Z > c|\mathrm{H_0}) = 1 - \Phi(c),$$

$$\text{Type II error} : \beta = P(Z \le c|\mathrm{H_a}),$$

where $\Phi(\cdot)$ is the standard normal distribution function. Solving the first equation, we have $c = z_{1-\alpha}$, the $(1-\alpha)$th percentile of the standard normal distribution.

The power of a statistical test is defined as the probability that a statistically significant test statistic will be obtained (i.e. reject the null hypothesis), given that the alternative hypothesis is true. It equals one minus the type II error. In the above example,

$$\begin{aligned}\text{Power} = 1 - \beta &= P(Z > c|\mathrm{H_a}) \\ &= 1 - \Phi\left(\frac{z_{1-\alpha}\sigma_0 - \sqrt{n}(\mu_a - \mu_0)}{\sigma_a}\right).\end{aligned}$$

Solving the equation, we obtain the required sample size to ensure a $1 - \beta$ power on detecting the difference of $\mu_a - \mu_0$; that is,

$$n = \left(\frac{z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_a}{\mu_a - \mu_0}\right)^2.$$

This equation can also be used to find the minimum detectable difference for given statistical power $1 - \beta$ and sample size n; that is,

$$\Delta = \mu_a - \mu_0 = \frac{z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_a}{\sqrt{n}}.$$

This is the smallest difference that can be detected with given sample size and power.

Although the sample size formula is obtained from a normal distribution, the relationship among the parameters of sample size, significance level α,

power $1-\beta$, variances, and effect-size Δ is generally true for any type of study and distribution. As the significance level α is getting smaller, or the power is getting higher, or the variances are getting larger, or the effect size is getting smaller, a larger sample size will be required. While the significance level is usually 5% and the power is usually 80% or 90%, it is very important to select suitable values for effect size and variances in designing a study. Investigators should not use overoptimistic values for the effect size or variances to avoid having an underpowered study.

The calculations of sample size, power and minimum detectable difference can be generalized to the case of $\mu_a < \mu_0$ by replacing $\mu_a - \mu_0$ with $|\mu_a - \mu_0|$ in the above formula. The above formulas are based on one-sided tests. For a two-sided alternative hypothesis, H_a: $\mu_a \neq \mu_0$, the sample size formula is essentially the same but $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$.

The choice of a one-sided or two-sided test will depend on the problem of interest. If one wants to test only one direction, e.g. $\mu = \mu_0$ against $\mu > \mu_0$, or test whether the population mean is greater than μ_0, then a one-sided test is appropriate. If the interest is to test the deviation from μ_0 from either direction, then a two-sided test is to be used. In this case, the null hypothesis will be rejected when the population mean is either too small or too large statistically compared with μ_0.

Estimation and Precision

In epidemiologic studies, researchers may be interested in estimating the magnitude of the effect from exposure instead of testing the hypothesis of no effect. The precision of an estimate can be measured by the width of a confidence interval that is designed to cover the true parameter of interest with a specified probability (coverage probability), $1-\alpha$. In the normal distribution example, if we want to estimate the mean μ with known variance σ^2, then the $1-\alpha$ confidence interval is $(\overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n})$. With regard to replications of sampling, we have a $1-\alpha$ probability that the true mean μ will be included in this interval. The larger the sample size, the narrower the confidence interval, and therefore the higher the precision of the parameter estimate (*see* **Random Error**).

Over the last decades some researchers have been stressing the advantage of using confidence intervals rather than testing p-values to present study results and make statistical inferences. One reason is that a confidence interval conveys not only information on the point estimation but also an impression of the precision of the estimate. Some further discussions of the pros and cons of hypothesis testing and confidence intervals can be found in Rothman & Greenland [33].

The sample size calculated from the confidence interval viewpoint will depend on the objective of the study. If a study is solely to estimate the effect of a parameter of interest with a given precision, then the sample size can be calculated from the width of the confidence interval. For the normal mean example, if one wants the width of the $1-\alpha$ confidence interval to be no more than 2δ, that is $2z_{1-\alpha/2}\sigma/\sqrt{n} \leq 2\delta$, then $n \geq (z_{1-\alpha/2}\sigma/\delta)^2$.

When the objective of a study is effectively to distinguish the parameter of interest from a specified value or distinguish among specified values, the sample size calculation should consider the expected location of the confidence interval. The sample size based on the width of the confidence interval alone will be insufficient [15]. In this case, the sample size obtained from confidence interval estimation will be similar to the sample size from hypothesis testing. In fact, hypothesis testing (two-sided) and confidence interval estimation are closely related. A study that yields a test p-value (two-sided) of precisely α for testing H_0: $\mu = \mu_0$ will have a $1-\alpha$ confidence interval that has one end at μ_0. In other words, if the $1-\alpha$ confidence interval of μ contains μ_0, then the null hypothesis H_0 will not be rejected with the significance level of α (two-sided).

Practical Considerations and Outline

In this article, we will focus our presentation on the determination of sample size from the traditional hypothesis testing approach and present some limited results based on estimation and precision. For the actual study, the sample size, test significance level, variability, power and minimum detectable difference (effect size) have to be considered at the design stage. The relationship among these parameters will allow investigators to calculate one parameter given the others.

In practice, the test significance level is often 5% and the statistical power is usually 80% or 90%. The variability and expected difference are often obtained from previous studies. Special consideration may be given to the choice of the minimal difference to be detected. The difference should be reasonable and suitable, such that it is practically meaningful and the study can be planned and conducted feasibly. If the assumed difference is too large and the sample size is consequently underestimated, then the study may fail to detect the true difference due to insufficient power. On the other hand, an oversized study may be costly to conduct and sometimes may detect a "tiny" difference that is not practically meaningful. The factors of time, cost and recruitment of subjects from the study population should all be considered, together with the selection of the study sample size, power and expected difference.

In general, the required sample size for a study depends not only on the parameters such as α, power $1-\beta$, and minimal detectable difference but also on the statistical procedures to be applied in the analysis and the study design. For epidemiologic studies, typically cohort studies and case–control studies, many papers and review articles have been published for sample size and power determinations (e.g. [2, 24, 25, 36, 43]). In this article we give a broad overview of sample size determination methods for a variety of epidemiologic designs. The methodologic details are skipped and relegated to the references. In the second section, sample size determination methods for studies with a binomial outcome (in **cohort** and unmatched **case–control studies**) are discussed. Methods for matched case–control studies are discussed in the next section. For cohort studies with Poisson outcomes, the sample size determination methods are presented in the fourth section. The fifth section discusses sample size determination for cohort studies when the outcome of interest is time to event. The sample size required for cohort studies with longitudinal or correlated outcomes is discussed in the sixth section, while the following section highlights some sample size calculations when the problem of interest is estimation and precision. The final section presents some further considerations on sample size and power determination followed by a discussion.

Without loss of generality, we will present the sample size formulas for one-sided tests in the following sections. The two-sided formulas can be obtained by replacing $z_{1-\alpha}$ with $z_{1-\alpha/2}$ in all instances unless otherwise specified.

Studies with Binomial Outcomes

Dichotomous Exposure: Cohort Study

In **cohort studies** with a fixed follow-up time, the main outcome is typically disease occurrence. Let p_0 and p_1 be the proportion of subjects who develop the disease in the unexposed and exposed populations, respectively. For a one-sided test H_0: $p_1 = p_0$ vs. H_a: $p_1 > p_0$, assuming an equal number of subjects n in the exposed and unexposed groups, the required sample size is given as follows (e.g. [36]):

$$n = \frac{\left(z_{1-\alpha}\sqrt{(2\overline{pq})} + z_{1-\beta}\sqrt{(p_1q_1 + p_0q_0)}\right)^2}{(p_1 - p_0)^2}, \tag{1}$$

where $q_1 = 1 - p_1$, $q_0 = 1 - p_0$, $\overline{p} = (p_0 + p_1)/2$, and $\overline{q} = 1 - \overline{p}$. The above formula is obtained from a normal approximation to the test statistic for comparing two binomial proportions. The formula can be represented as

$$n = \frac{\left[\begin{array}{c} z_{1-\alpha}\sqrt{(2\overline{pq})} \\ +z_{1-\beta}\sqrt{(p_0(1+r) - p_0(1+r^2))}\end{array}\right]^2}{[p_0(1-r)]^2} \tag{2}$$

in terms of the risk ratio $r = p_1/p_0$ (*see* **Relative Risk**) to test H_0: $r = 1$ vs. H_a: $r > 1$.

In general, let π_e be the proportion of subjects in the exposed group. Then the total sample size required is given by

$$N = \frac{\left[\begin{array}{c} z_{1-\alpha}\sqrt{(\overline{pq}/\pi_e(1-\pi_e))} \\ +z_{1-\beta}\sqrt{(p_1q_1/\pi_e + p_0q_0/(1-\pi_e))}\end{array}\right]^2}{(p_1 - p_0)^2}. \tag{3}$$

Dichotomous Exposure: Case–Control Study

The sample size required for an unmatched **case–control study** is similar to that for a cohort study. Let p_1 and p_0 now denote the proportions of subjects exposed in the case and control groups, respectively. For a study with one control per case, the sample size required can be calculated by eqs (1)

and (2). For a study with k controls per case, the number of cases required is given by [36]

$$n = \frac{\left(\begin{matrix} z_{1-\alpha}\sqrt{((1+1/k)\overline{p}'\overline{q}')} \\ +z_{1-\beta}\sqrt{(p_1q_1+p_0q_0)}\end{matrix}\right)^2}{(p_1-p_0)^2},$$

where $\overline{p}' = (p_1 + kp_0)/(1+k)$ and $\overline{q}' = 1 - \overline{p}'$.

In practice, the exposure rate among controls, p_0, is usually obtained from previous studies and estimated from the general population. The **relative risk** $r = p_1/p_0$ or the **odds ratio** $(\mathrm{OR}) = p_1q_0/p_0q_1$ is then specified under the alternative hypothesis to calculate the power of the study. For example, consider a case–control study of a potential association between congenital heart defects and oral contraceptives used around the time of conception. An estimate of the exposure rate among controls is 30%. Given $\alpha = 0.05$ and $\beta = 0.10$, in order to detect a relative risk of $r = 1.5$ (so $p_1 = 0.45$), and based on (2), the sample size required is $n = 177$ (per group).

When an OR is specified, the exposure rate among cases can be solved as follows:

$$p_1 = \frac{p_0\mathrm{OR}}{[1 + p_0(\mathrm{OR} - 1)]}.$$

For the above example, the sample size required to detect an $\mathrm{OR} = 2$ (so $p_1 = 0.462$), given $\alpha = 0.05$ (two-sided) and $\beta = 0.10$, will be $n = 153$ (per group).

Continuous Exposure: Cohort Study

For a continuous exposure variable, sample size estimation methods for cohort studies have been derived by several authors (e.g. [24] and [44]). Let $p(x)$ be the probability of developing a disease with exposure level $X = x$ over a fixed follow-up time. Within the framework of **logistic regression**, the association between $p(x)$ and the continuous exposure variable X can be modeled as

$$\log\left[\frac{p(x)}{(1-p(x))}\right] = \delta + \theta x, \qquad (4)$$

where θ is the log OR for a unit increase in X. Testing the null hypothesis of no association is equivalent to testing H_0: $\theta = 0$.

Whittemore [44] derived sample size requirements for Wald tests based on maximum likelihood methods. To approximate the variance of the maximum likelihood estimate of θ, the disease probability is assumed to be small, i.e. $p(x) \approx 0$. Under this assumption, the total sample size for testing $\theta = 0$ with significant level α and power $1 - \beta$ is estimated by

$$N = \frac{\left[z_{1-\alpha}\sqrt{\nu(0)} + z_{1-\beta}\sqrt{\nu(\theta_\mathrm{a})}\right]^2}{[\theta_\mathrm{a}^2 \mathrm{e}^{\delta}]}. \qquad (5)$$

In formula (5), $\nu(\theta) = [m/(mm_{11} - m_1^2)](\theta)$ and $m(t) = \mathrm{E}[\exp(tX)]$ is the moment-generating function of X, and m_1 and m_{11} are the first and second partial derivatives of $m(t)$ with respect to t, respectively. The term e^{δ} is the odds of disease corresponding to $X = 0$ and θ_a is the log OR under the alternative hypothesis for a unit increase in the exposure X. Formula (5) is suitable to use when the sample size N is large. Tables for various distributions of exposure are given in Whittemore [44]. For example, when X has an N(0, 1) distribution, with $p(0) = 0.07$ (the disease probability in controls), $\exp(\delta) = p(0)/(1 - p(0)) = 0.075$, $\alpha = 0.05$ and power $= 0.90$, approximately $N = 543$ observations are needed to detect an OR of $\exp(0.5) = 1.65$ for a unit increase in the exposure X.

Lubin & Gail [24] studied a general method based on the score test statistic

$$U(\theta_0) = \partial \log \frac{L}{\partial\theta} = \sum_i x_\mathrm{i}(d_\mathrm{i} - p(x_\mathrm{i})) \qquad (6)$$

evaluated under the null hypothesis H_0: $\theta = \theta_0$, where L is the likelihood function for the logistic model (4) and d_i is a disease indicator (i.e. $d_\mathrm{i} = 1$ and $d_\mathrm{i} = 0$ for disease and nondisease, respectively). The total sample size required to test the null hypothesis with a significance level α and power $1 - \beta$ is given by

$$N = \frac{\left[z_{1-\alpha}\sqrt{\nu_0(U(\theta_0))} + z_{1-\beta}\sqrt{\nu_\mathrm{a}(U(\theta_0))}\right]^2}{[\Delta(\theta)]^2}, \qquad (7)$$

where $\nu_0(U(\theta_0))$ and $\nu_\mathrm{a}(U(\theta_0))$ are the variances of $U(\theta_0)$ under H_0: $\theta = \theta_0$ and H_a: $\theta = \theta_\mathrm{a}$, respectively; and $\Delta(\theta) = E_\mathrm{a}[U(\theta_0)]/N$. The evaluation will depend on the hypothesized parameters θ_0 and θ_a, as well as the statistical distribution of the exposure variable X. In general, special numerical calculation is needed to estimate the sample size. Details and some examples are given in Lubin & Gail [24].

Continuous Exposure: Case–Control Study

Lubin et al. [25] have derived sample size formulas for **case–control studies** with continuous exposure variables based on the logistic model (4) and score test (6). Let $F_1(x)$ and $F_0(x)$ be the distributions of exposure among cases and **controls**, respectively. It is shown that the number of cases required for a one-sided test with significance level α and power $1-\beta$ is given by

$$n = \frac{(k+1)}{k} \frac{\left[z_{1-\alpha}\sigma_x + z_{1-\beta}\sqrt{\left(\frac{k\sigma_1^2 + \sigma_0^2}{k+1}\right)} \right]^2}{(\mu_1 - \mu_0)^2}, \tag{8}$$

where k is the number of controls for each case (so total sample size $N = (k+1)n$), and (μ_1, σ_1^2) and (μ_0, σ_0^2) are the mean and variance of the exposure variable under F_1 and F_0, respectively; and

$$\sigma_x = \frac{(\sigma_1^2 + k\sigma_0^2)}{(k+1)} + \frac{(\mu_1 - \mu_0)^2 k}{(k+1)^2}.$$

The quantities μ_0, μ_1, σ_0^2 and σ_1^2 may be obtained from preliminary data such as previous studies that give estimates of the distributions of the exposure variable among cases and controls, or calculated from specifying the distribution functions of exposure in cases and controls, in which case numerical integration may be needed. The details are given in Lubin et al. [25].

When continuous exposure variables are dichotomized in the study (cohort or case–control), the required sample size will be increased due to loss of information from dichotomization. The efficacy losses will depend on the nature of the exposure distribution and the choice of the cutoff points for the dichotomization, as discussed in Lubin et al. [25].

Adjustment for Confounding Variables

Whittemore [44] extended the sample size formula (5) to adjust for **confounding** variables. When the joint multivariate variables $\mathbf{X}$ (a vector of the variable of interest X_1 and the confounding variables $X_2, \ldots, X_k$) follow an exponential family distribution, the variance term $v(\boldsymbol{\theta})$ can be obtained from the moment-generating function for $\mathbf{X}$. For example, when $\mathbf{X}$ has a multivariate normal distribution with mean μ and positive covariance matrix Σ, Whittemore showed that

$$v(\boldsymbol{\theta}) = \left[\mathrm{var}(X_1) \exp\left(\boldsymbol{\theta}'\boldsymbol{\mu} + \frac{\boldsymbol{\theta}'\Sigma\boldsymbol{\theta}}{2} \right) (1 - \rho_{1\cdot 2\ldots k}^2) \right]^{-1}$$

where X_1 is the exposure variable of interest, $\rho_{1\cdot 2\ldots k}$ is the multiple correlation coefficient relating X_1 to $X_2, \ldots, X_k$, and k is the total number of variables.

Lubin & Gail [24] proposed a general method for determining the sample size required to test whether exposure is associated with disease outcome, while adjusting for potential confounding variables. The method is based on a regression model and can be applied to both **cohort studies** and **case–control studies**. Let $\mathbf{X}$ be the joint multivariate variables of exposure and potential **confounders**. It is assumed that the probability of disease for $\mathbf{X} = \mathbf{x}$ is given by

$$p(\mathbf{x}; \delta, \theta, \lambda) = \frac{r(\mathbf{x})}{[1 + r(\mathbf{x})]} \text{ and } r(\mathbf{x}) = \mathrm{e}^{\delta} R(\mathbf{x}; \theta, \lambda),$$

where $R(\mathbf{x}; \theta, \lambda)$ is a smooth, positive function satisfying $R(\mathbf{0}; \theta, \lambda) = 1$, θ is the parameter of interest and λ is a vector of parameters (nuisance parameters) associated with the confounding variables. Thus, the **logistic regression model** (4) is a special case of this model when $R(\mathbf{x}; \theta, \lambda) = \exp(\mathbf{x}\theta)$. In fact, the function $R(\mathbf{x}; \theta, \lambda)$ can assume a multiplicative or additive form (*see* **Additive Model**; **Multiplicative Model**). For a one-sided test of the null hypothesis $\theta = \theta_0$ vs. the alternative $\theta = \theta_a$, a score test statistic is again given in (6). The total sample size required for a significance level α and power $1-\beta$ is given in (7). However, $v_0(U(\theta_0))$, $v_a(U(\theta_0))$ and $\Delta(\theta) = E_a[U(\theta_0)]/N$ contain the nuisance parameters δ and λ. For statistical analysis after data are collected, these parameters can be estimated by maximum likelihood. For sample size evaluation at the design stage, these parameters are replaced by δ_0 and λ_0 to which the maximum likelihood estimates $\hat{\delta}$ and $\hat{\lambda}$ (obtained under H_0: $\theta = \theta_0$) converge when the alternative hypothesis is true. The method works in general for both continuous or categorical variables as long as a joint distribution of $\mathbf{X}$ is specified. However, the sample size formula needs specialized numerical calculations based on the model specification for $R(\mathbf{x}; \theta, \lambda)$ and the joint distribution of $\mathbf{X}$.

An example discussed in Lubin & Gail [24] is to test the effect of radon exposure on lung cancer after

adjusting for smoking. They considered a multiplicative joint OR model,

$$R(\text{WLM, SMOK}) = (1 + \theta\ \text{WLM})(1 + \lambda\ \text{SMOK}),$$

where SMOK denotes the mean number of cigarettes smoked per day and WLM denotes exposure to radon decay products measured in working level months. They illustrated the use of this model to test the null hypothesis $\theta_0 = 0$, against the alternative hypothesis $\theta_a = 0.015$ for a five-year study assuming $\lambda_a = 0.3$. Based on prior knowledge, a five-year lung cancer mortality rate among nonsmokers is 0.00471. Therefore, $\delta = \log[0.00471/(1 - 0.00471)] = -5.354$. For a case–control study with $k = 5$ (number of **controls** per case), $\alpha = 0.05$ and power $1 - \beta = 0.90$, they showed that $n = 251$ cases are needed (number of controls $= 5n = 1255$). This is slightly larger than the 218 cases and 1086 controls required when the SMOK factor is ignored. This is because the confounding variable SMOK brings additional variability into the **multiplicative model**. The calculation assumes independence of the radon and smoking exposure distributions. Lubin & Gail [24] showed that the required sample size would be much larger if the radon and smoking exposures were highly and negatively correlated.

Adjustment for Stratification Factors

When samples are drawn from several strata, the **stratification** factors should be considered in the data analysis as well as in the sample size calculation at the design stage. When the probability of disease can be modeled in a **logistic regression**, the method for adjusting for **confounding** variables described above [24] can be used to estimate sample size. Smith & Day [40] provided extensive tabulation for the required sample size. They concluded that if the stratification factor is not strongly related to the exposure or disease status, then an increase of more than 10% in the sample size is unlikely to be needed.

Logistic regression is a special case of generalized linear models. Therefore, the methods proposed by Self & Mauritsen [38] and Self et al. [39] for generalized linear models can be used for sample size and power calculations. The methods are based on a score test and a likelihood ratio statistic, respectively. The sample size is estimated by treating the **stratification** factors as nuisance parameters. Their methods require, in general, specialized numerical calculations based on a specified joint distribution for the covariates in the model. See the section on a **cohort study** with Poisson outcomes for further discussion.

Other methods for sample size estimation for stratified studies can be found in the literature. For testing unity of a common OR for a collection of several 2×2 tables, Munoz & Rosner [28] studied sample size determination based on the Mantel–Haenszel test for stratified data (*see* **Mantel–Haenszel Methods**). Their method is appropriate when all margins of each table are fixed. Woolson et al. [45] and Nam [29] considered sample size calculations based on Cochran's test that do not require fixed margins of the 2×2 tables in each stratum. Nam used a continuity correction to guarantee that the actual type I error rate of the test does not exceed the nominal level.

Test for Interaction and Trend

When the question of interest is whether the relative risks among different strata (or levels of **confounding** factors) are equal, the problem becomes a test of **interaction** between exposure and the **stratification** factors (or confounding factors). Smith & Day [40] presented methods for evaluating power and sample size for testing the interaction between a dichotomous stratification variable and a categorical exposure. They showed that in order to detect an interaction effect of the same magnitude as a specified main effect, a sample size at least four times as large as for testing the main effect is required.

In general, tests for interaction can be addressed by testing appropriate coefficients in a **logistic regression** model. Within this framework, the methods developed by Lubin & Gail [24] and Self et al. [38, 39] can be used to estimate the required sample size. The methods will treat the interaction terms as the parameters of interest and treat all other factors as nuisance parameters. Usually, the methods need specialized numerical calculations. Garcia-Closas & Lubin [11] compared several methods for sample size calculations on testing gene-environmental interactions.

Testing for trend can also be addressed (*see* **Dose–Response**) within the framework of generalized linear models. For example, when an exposure variable X is ordered categorical, a trend test is to test whether the disease odds are proportionally increased as the exposure level increases. This is

equivalent to testing for a nonzero value of parameter θ in the logistic model (4). The methods developed by Lubin & Gail [24] and Self et al. [38, 39] can be used to estimate the required sample size.

Matched Case–Control Study

To improve comparability and efficiency in **case–control studies**, one may match **controls** with cases on potential **confounders**. A pair-matched study matches one control with each case, which is the simplest matched case–control design. When the number of cases is limited, more than one control can be matched to each case to increase statistical precision. For the matched case–control studies, the matching should be considered in the data analysis as well as in the sample size and power calculations (*see* **Matched Analysis**).

Pair-Matched Study

The probabilities of outcomes of a pair-matched case–control study are laid out in the following 2×2 table:

	Control		
Case	+	−	
+	π_{11}	π_{10}	$\pi_{1.}$
−	π_{01}	π_{00}	$\pi_{0.}$
	$\pi_{.1}$	$\pi_{.0}$	1

The "+" and "−" signs denote exposure and nonexposure status, respectively. Let m_{01} and m_{10} denote the number of (−+) and (+−) pairs for (case, control), and $m = m_{01} + m_{10}$ be the total number of discordant pairs. Conditional on m, the observation m_{10} has a binary distribution with $p = \pi_{10}/(\pi_{01} + \pi_{10}) = \psi/(1 + \psi)$, where $\psi = \pi_{10}/\pi_{01}$ denotes the disease–exposure OR.

The test of no disease–exposure association, i.e. H_0: $\pi_{01} = \pi_{10}$, is equivalent to testing H_0: $p = 1/2$. Schlesselman [36] gave a formula for the total number of discordant pairs m required to detect a relative risk R based on a normal approximation to McNemar's test,

$$m = \frac{\left[z_{1-\alpha}/2 + z_{1-\beta}\sqrt{(p(1-p))}\right]^2}{[p - 1/2]^2}. \qquad (9)$$

Suppose $\pi_d = \pi_{01} + \pi_{10}$ is the probability of obtaining discordant pairs. Then the total number of pairs for the study is estimated by

$$n = \frac{m}{\pi_d}. \qquad (10)$$

To estimate π_d, some additional information is required other than the OR ψ and the marginal exposure rate for **controls**, $\pi_{.1}$. When exposure for cases and controls within each pair is statistically independent, Schlesselman gives the following estimate:

$$\pi_d = \pi_{.1}(1 - \pi_{1.}) + \pi_{1.}(1 - \pi_{.1}),$$

where $\pi_{1.} = \psi\pi_{.1}/[1 + (\psi - 1)\pi_{.1}]$.

For example, Schlesselman shows that to detect an OR of $\psi = 2$, from (9) one requires $m = 90.3$ discordant pairs for $\alpha = 0.05$ (two-sided) and $\beta = 0.10$. For a control exposure rate $\pi_{.1} = 0.30$, assuming independent exposures, $\pi_{1.} = 0.46$, $\pi_d = 0.485$, and the total number of pairs required for the study is $n = 187$.

When the independence assumption does not hold, Schlesselman's sample size estimate may be severely biased (often underestimated). Several corrections have been proposed to allow for correlation between the exposure status within pairs that is often induced in the case of efficient matching. To allow for exposure association, Fleiss & Levin [7] used the exposure OR $\omega = \pi_{11}\pi_{00}/\pi_{01}\pi_{10}$ and corrected the estimation of discordant probability π_d given by Schlesselman. Under the independence assumption, $\omega = 1$. In the case of $\omega \neq 1$, the corrected estimate of the discordant probability is

$$\pi'_d = \pi_d \frac{\sqrt{(1 + 4(\omega - 1)\pi_1(1 - \pi_{1.}))} - 1}{2(\omega - 1)\pi_{1.}(1 - \pi_{1.})}.$$

In the above example, if $\omega = 2.5$, then the corrected $\pi_d = 0.376$. Therefore, the required number of pairs for the study is $n = 241$ rather than 187.

Dupont [6] presented another correction based on the contingency coefficient

$$\phi = \frac{\pi_{11}\pi_{00} - \pi_{10}\pi_{01}}{\sqrt{(\pi_{1.}(1 - \pi_{1.})\pi_{.1}(1 - \pi_{.1}))}}$$

for the case–control exposure association within a pair. The discordant probability π_d is adjusted as follows for specified ψ, $\pi_{.1}$ and ϕ,

$$\pi'_d = \pi_d - 2\phi\sqrt{(\pi_{1.}(1 - \pi_{1.})\pi_{.1}(1 - \pi_{.1}))},$$

where

$$\pi_{1\cdot} = \frac{\left[\begin{array}{c} 2\psi\pi_{\cdot 1}(\psi\pi_{\cdot 1} + 1 - \pi_{\cdot 1}) \\ +(\psi-1)^2\pi_{\cdot 1}(1-\pi_{\cdot 1})\phi^2 \\ -(\psi-1)\pi_{\cdot 1}(1-\pi_{\cdot 1})\phi\sqrt{(\phi^2(\psi-1)^2+4\psi)} \end{array}\right]}{2[(\psi\pi_{\cdot 1}+1-\pi_{\cdot 1})^2+(\psi-1)^2\pi_{\cdot 1}(1-\pi_{\cdot 1})\phi^2]}.$$

Several other methods are discussed and reviewed by Lachin [17] and Wickramaratne [39]. Lachin recommended always using a corrected procedure for sample size estimation. He pointed out that Dupont's correction and Fleiss & Levin's correction give very similar results. Qiu et. al. [32] proposed a sample size calculation method to test interaction between a specific exposure and a risk factor in a pair-matched case–control study.

Multiple Control per Case Study

For studies calling for k matched **controls** for each case (i.e. $1:k$ **matching**), Schlesselman gave an approximation for the required number of matched sets, $n' = (k+1)n/2k$, where n is calculated from (10) with given $\pi_{\cdot 1}$, ψ, test level α and power $1-\beta$. This approximation is valid when the exposure rates in cases and controls are similar.

Taylor [41] and Lui [26] studied other approximations that do not assume that the exposure rates in cases and controls are similar. Lui provided simulation results showing that when the OR of exposure between cases and controls is small ($\leq$4), his method gives more accurate results than that of Taylor; when ORs are large, the formula given by Taylor is recommended.

All three approximations (i.e. Schlesselman, Taylor and Lui) are based on the assumption of homogeneity of exposure among different matched sets. That is, the probabilities of exposure for cases and controls are constant across matched sets. Tables for the required number of case–control sets are given in Breslow & Day [2] for different values of power, significance level, relative risk and different matching ratios.

Two remarks are given as follows. First, a matched **case–control study** can be regarded as a special case of general stratified study in which each matching category is treated as a unique stratum. The methods discussed in the previous sections for unmatched case–control studies with confounding or **stratification** variables can therefore be used for sample size estimation. However, the methods may break down when the number of strata is large and data in each strata are sparse. Secondly, selection of the number of controls per case is more of a practical consideration (e.g. availability, time and cost) than a statistical power concern. In fact, the power gain is diminished when the number of controls is increased to beyond four controls per case (see [2]).

Cohort Study with Poisson Outcomes

When the number of events for a **cohort study** (e.g. diseases or deaths) is relatively small compared with the total number of subjects, the probability of events occurring may be modeled by a Poisson distribution (*see* **Poisson Regression**). In such studies, it is often interesting to test the rate of event incidence rather than the overall probability of events.

Dichotomous Exposure

In a cohort study with dichotomous exposure, Gail [9] presents methods to calculate power for studies with Poisson outcomes when the number of exposed and unexposed samples are equal. Let μ_1 and μ_0 be the incidence rate per unit time for the exposed and unexposed groups, respectively. We want to test H_0: $\mu_0 = \mu_1$ vs. H_a: $\mu_1 > \mu_0$. Two study designs are considered by Gail. Design A is to follow subjects until a predetermined total number of events is observed. Design B is to follow subjects up to a predetermined length of time. For design A, Gail provides a table for the total number of events for given **relative risk** $r = \mu_1/\mu_0$, significance level α and power $1-\beta$. Brown & Green [4] extend the method to the case of unequal group sizes. Tables are provided for the total number of events. For Design B, the event rate for the unexposed group, λ_0, is estimated from given relative risk, significance level α and power $1-\beta$. Tables are provided in Brown & Green [4]. Then, the expected duration of a study can be estimated as $t = \lambda_0/(\mu_0 n_0)$, where n_0 is the number of subjects in the unexposed group for the study that is specified by investigators.

When the expected number of events in the study is large, approximation formulas are presented in Gail [9], Brown & Green [4], as well as in

Breslow & Day [2]. Let k be the ratio between the numbers of subjects in the exposed and unexposed groups. For design A, subjects are followed until a certain number of events, say m, is observed. Conditioning on m, the number of events observed in the exposed group, follows a binomial distribution with probability $k\mu_1/(k\mu_1 + \mu_0) = rk/(rk + 1)$. On the basis of a normal approximation to the arcsine transformation of the square root of a binomial proportion, Brown & Green [4] give

$$m = \frac{[z_{1-\alpha} + z_{1-\beta}]^2}{4\left[\sin^{-1}\sqrt{(rk/(rk+1))} - \sin^{-1}\sqrt{(k/(k+1))}\right]^2},$$

where r is the relative risk between exposed and unexposed groups.

On the basis of a normal approximation to a binomial proportion, Breslow & Day [2] give

$$m = \frac{\left[z_{1-\alpha}\sqrt{k}/(k+1) + z_{1-\beta}\sqrt{(rk)}/(rk+1)\right]^2}{[rk/(rk+1) - k/(k+1)]^2}. \tag{11}$$

A more accurate estimate is obtained by using Yates' correction to the chi-squared significance test [42], which results in multiplying the right-hand side of (11) by $(1 + \sqrt{(1 + A)})^2/4$, where

$$A = \frac{2[rk/(rk+1) - k/(k+1)]}{\left[z_{1-\alpha}\sqrt{k}/(k+1) + z_{1-\beta}\sqrt{(rk)}/(rk+1)\right]^2}.$$

Brown & Green [4] present an example of a study to compare **incidence rates** of congenital malformations among children born in a specific town. A control population is identified that is twice as large as the town under study ($k = 0.5$). To have 90% power to detect a fourfold relative risk, $r = 4$, it is estimated that a total of $m = 20$ events will be needed (based on the table given in Brown & Green with $\alpha = 0.05$). Similar results are obtained from the approximation formulas. With Brown & Green's approximation, one obtains $m = 19$. With Breslow & Day's approximation, one obtains $m = 18$ without Yates' correction, and $m = 20$ with Yates' correction.

Loglinear Models

A loglinear model may be used to associate the event rate $\lambda(x)$ with the exposure level x; namely

$$\log[\lambda(x)] = \delta + \theta x. \tag{12}$$

A test of no association is equivalent to testing H_0: $\theta = 0$.

Sample size and power calculations for this model have been studied by Self et al. [38, 39]. They developed methods for generalized linear models based on the score test in their first paper and based on the likelihood ratio test in their second paper. The loglinear model for Poisson outcomes is a special case of the models they discussed. For a categorical exposure variable X (or a finite number of categorized configurations for a continuous variable), the required sample size is estimated from a noncentral chi-square approximation to the test statistic. However, there is no explicit sample size or power formula in general. Sample size determination is performed numerically for given nuisance parameters and distribution of X. Simulations are recommended to check the accuracy of the estimated sample size. Simulation results [39] show that the method based on the likelihood ratio statistic usually gives better results than that based on the score test.

Test for Trend with Categorical Variable

When the exposure variable X is ordered categorical with K levels ($K > 2$), a trend test can be used to assess whether the event rate $\lambda(x)$ changes monotonically with exposure (*see* **Dose–Response**). Breslow & Day [2] presented a method to estimate the sample size based on a chi-square trend test statistic. The test is based on a score statistic under a loglinear model for a Poisson distribution (*see* **Poisson Regression**). It contrasts the observed number of events, O_k, with the expected number of events, E_k, calculated from external rates. Let x_k be the kth exposure level of X. Then the power of the test is the probability such that

$$\sum_{k=1}^{k} O_k \left[x_k - \frac{\sum_j x_j E_j}{\sum_j E_j} \right] - z_{1-\alpha}\sqrt{V} \geq 0,$$

where $z_{1-\alpha}$ is the $(1-\alpha)$th percentile of the standard normal distribution and

$$V = W \sum_k O_k$$

and

$$alW = \frac{\left[\sum_k x_k^2 E_k - \left(\sum_k x_k E_k\right)^2\right]}{\sum_k E_k}.$$

Under an alternative of a linear trend in relative risk such that $r(x) = 1 + \theta x$, the left-hand side of the probability equation will have mean μ approximated by

$$\mu = \sum_k \theta x_k E_k \left[x_k - \frac{\sum_j x_j E_j}{\sum_j E_j}\right] - z_{1-\alpha}\sqrt{\left(W \sum_k (E_k + \theta x_k E_k)\right)}$$

and variance σ^2 approximated by

$$\sigma^2 = \sum_k (1 + \theta x_k) E_k \left[x_k - \frac{\sum_j x_j E_j}{\sum_j E_j}\right]^2 - z_{1-\alpha} \sum_k \theta x_k E_k \left[x_k - \frac{\sum_j x_j E_j}{\sum_j E_j}\right] \times \sqrt{\frac{W}{\sum_k E_k} + \frac{z_{1-\alpha}^2 W}{4}}.$$

Therefore, the expected number of events for given test level α and power $1-\beta$ should satisfy $\mu = \sigma z_{1-\beta}$. This equation can be used to solve for the number of events required for the study under a given distribution for the exposure variable X and alternative hypothesis $\theta = \theta_a$. Numerical methods are needed, in general, except for some special cases; see [2] for details and examples.

Survival Study

In a cohort study, when the time to an event (e.g. disease or death) is observed exactly or within a certain interval, survival analysis can be used for the comparison of incidence rates (*see* **Survival Analysis, Overview**; **Proportional Hazards, Overview**; **Cox Regression Model**). Survival analysis uses not only the number of events but also the time when an event occurs. This often brings more information for comparing event rates than a method using the number of events alone. The survival function $S(t)$ is a probability function that an individual will survive or be disease free up to a certain time point t. The hazard function is defined as

$$\lambda(t) = -\frac{\mathrm{d}S(t)/\mathrm{d}t}{S(t)}.$$

It measures an instantaneous mortality or morbidity risk relative to a survival probability at that time. A commonly used model for survival analysis is the proportional hazard model, which assumes that hazard functions for two groups satisfy

$$\lambda_1(t) = \psi \lambda_2(t).$$

It is of interest to test the constant hazard ratio $\psi = 1$ or the constant log hazard ratio $\theta = \log(\psi) = 0$ (*see* **Hazard Rate**; **Relative Hazard**).

In a survival study, it is uncommon to observe the actual event time for all subjects. Usually, for some subjects, time to event is censored by the end of the study or at some time point when the subject is lost to follow-up. The statistical power of testing $\theta = 0$ depends principally on the number of events actually observed.

Calculating the Number of Events

For a given alternative on the log hazard ratio, $\theta = \theta_a$, test significance level α and statistical power $1-\beta$, the required total number of events is estimated as

$$D = \frac{4(z_{1-\alpha} + z_\beta)^2}{\theta_a^2} \qquad (13)$$

based on a score test under the proportional hazard model [37]. For example, to detect a hazard ratio of 1.5 with $\alpha = 0.05$ (two-sided) and $1 - \beta = 0.90$, the required number of events is $D = 256$.

An alternative approximation to the required total number of events based on a normal approximation to the log rank statistic, assuming constant hazard ratio $\psi_a = \exp(\theta_a)$, is given by Freedman [8] as

$$D = \frac{(z_{1-\alpha} + z_\beta)^2(\psi_a + 1)^2}{(\psi_a - 1)^2}. \qquad (14)$$

Simulation results show that this approximation usually provides a slight overestimate for the total number of events [8]. When θ_a is small (close to 0), expressions (13) and (14) will give similar estimates for the total number of events. For the above example, $\psi_a = 1.5$, the required number of events based on (14) is $D = 263$ instead of $D = 256$ from formula (13).

The sample size formulas (13) and (14) are obtained assuming an equal number of subjects allocated in the two study groups. If this is not the case, then the required total number of events corresponding to formula (13) is

$$D = \frac{(z_{1-\alpha} + z_\beta)^2}{\pi_1 \pi_0 \theta_a^2},$$

where π_1 and $\pi_0 = 1 - \pi_1$ are the proportions of subjects to be allocated to the two groups. The corresponding formula for (14) under unequal number of subjects in the two groups is

$$D = \frac{(z_{1-\alpha} + z_\beta)^2(\pi_1 \psi_a + \pi_0)^2}{\pi_1(\psi_a - 1)^2}.$$

Calculating the Number of Subjects

Suppose P is the average probability of an individual having an event in the study population. Then the total number of subjects required for the study is approximately $N = D/P$, where D is the total number of events required for the study.

If a study enrolls all subjects at once (e.g. a **cohort study**) and follows every subject up to time f, then the probability P can be estimated by $P = 1 - S(f)$, where S is an average of the two survival functions for the two study groups; that is, $S(t) = [S_0(t) + S_1(t)]/2$ at any time t.

In many cases, the survival function is approximated by an exponential distribution; that is, $S_0(t) = \exp(-\lambda_0 t)$ and $S_1(t) = \exp(-\lambda_1 t)$. The parameters λ_0 and λ_1 can be obtained from the specified median survival time for the corresponding survival functions. Continuing from the example above, where the hazard ratio to be detected is 1.5, suppose the median survival times for the unexposed and exposed subjects are two and three years, respectively. Then $\lambda_0 = 0.347$, $\lambda_1 = 0.231$ and $S(t) = [\exp(-0.347t) + \exp(-0.231t)]/2$. If the study has a three-year follow-up, then the required total sample size can be estimated as $N = 256/0.573 = 447$ based on (13), and $N = 263/0.573 = 459$ based on (14).

In experimental and cohort studies, subjects may be enrolled within a period of time, say from 0 to time T (*see* **Cohort Study**; **Experimental Study**). The follow-up time for subjects in the study can be anywhere between f (for the last recruited subject) and $T + f$ (for the first subject in the study). The probability of events for the study can be approximated by an average using Simpson's rule [37]; that is,

$$P = 1 - \frac{1}{6}[S(f) + 4S(0.5T + f) + S(T + f)],$$

where $S(t)$ is again the average of the two survival functions.

Lachin & Foulkes [18] presented a sample size and power calculation method based on exponential distributions. Their method allows for adjustment on the staggered entry of subjects, loss to follow-up (including deaths from **competing risks**), **stratification**, drop-in and lag in the effectiveness (or exposure effect) during the course of study. Lakatos [19] extended the method using the log rank statistic. Simulation studies show that Lakatos's method [19] is robust even when the proportionality assumption is not satisfied [20]. Other factors can influence sample size, including lack of **sensitivity** in making diagnoses and alternative methods of analysis such as comparison of two Kaplan–Meier curves (see, for example, [10]).

Longitudinal Studies

Longitudinal studies have become popular as the methods for longitudinal data analysis became available (see, for example, [5]). In a longitudinal study, repeated measures are taken from a subject over

a period of time. A longitudinal study will provide information (e.g. time effect for individuals) that cannot be obtained by a **cross-sectional study**. This type of study can be used for time-dependent exposures (e.g. smoking, alcohol use, diet, stress, blood pressure) and/or recurrence outcomes (e.g. pain, allergy, asthma, depression). Data from a longitudinal study require special statistical methods because the repeated measures from the same subject tend to be correlated. It is therefore required to consider the correlation among repeated measures while planning the sample size and power for a longitudinal study.

Dichotomous Exposure

The impact of repeated measures on sample size calculations can be illustrated by comparing the average differences in a continuous response. Suppose k repeated measures are taken from each subject, assume the correlation coefficient between any two measures is ρ, then the average of the k measures has variance

$$\text{var}(\overline{Y}) = \frac{1}{k}[1 + (k-1)\rho]\sigma^2,$$

where σ^2 is the variance of each response measure. Assume the variance and correlation are the same under the null and alternative hypotheses. Then the sample size required to detect a difference Δ with significance level α (one-sided) and power $1 - \beta$ is

$$n = \frac{2(z_{1-\alpha} + z_{1-\beta})^2[1 + (k-1)\rho]\sigma^2}{k\Delta^2} \qquad (15)$$

for each group [5, Chapter 2]. The correlation ρ will usually be positive. Therefore, the larger the value of ρ, the larger the required sample size for the study. This is because there is less independent information gained from each repeated measurement as ρ approaches 1. When $\rho = 1$, the number of subjects required is the same as in a study with one measurement per subject.

For example, consider a **cohort** (or **experimental**) **study** to investigate the association of a certain diet with total cholesterol level. Subjects in the study will have their total cholesterol measured quarterly for a period of a year. Suppose a standard deviation of 80 is assumed for each measure and the correlation between any two measures is $\rho = 0.5$. With $\alpha = 0.05$ and power = 90% to detect a 20-point difference, we will need $n = 172$ per group based on (15).

For a binary response variable, the sample size for a longitudinal study with repeated measures can be obtained similarly. Suppose p_0 and p_1 are the proportions of subjects who develop the disease in the unexposed and exposed groups, respectively. Then the required sample size for a longitudinal study with k repeated measures is estimated as [5, Chapter 2]

$$n = \left(z_{1-\alpha}\sqrt{(2\overline{pq})} + z_{1-\beta}\sqrt{(p_1q_1 + p_0q_0)}\right)^2 \times \frac{[1 + (k-1)\rho]}{[k(p_1 - p_0)^2]}, \qquad (16)$$

where $q_1 = 1 - p_1$, $q_0 = 1 - p_0$, $\overline{p} = (p_0 + p_1)/2$, and $\overline{q} = 1 - \overline{p}$. The quantity ρ is the correlation coefficient of the binomial response variable between any two repeated measures. It is assumed to be the same for all subjects under the null and alternative hypotheses. Because the binomial response takes values of 0 or 1, unlike for normal distributed data, the correlation coefficient ρ is constrained in complicated ways (see [5]). For repeated measures, the correlation ρ is usually positive and its value may range in $[0, b]$, where b can be less than 1.

In the above example, if the total cholesterol is dichotomized and the threshold for elevated total cholesterol is 200, then the response will be binomial (yes/no). If one assumes a correlation of 0.5 between any two responses, with $k = 4$ repeated measures, $p_1 = 60\%$ and $p_0 = 70\%$, a sample size of $n = 243$ per group is required to detect a 10 percentage point difference in elevated total cholesterol, with $\alpha = 0.05$ and power = 90%.

Generalized Estimating Equation Method

In general, Liu & Liang [23] presented a method for computing sample size and power for studies with longitudinal observations using the generalized estimating equation (GEE) method. Suppose μ_{ij} is the mean of the ith subject at the jth measure. The generalized linear model with a link function of g is given as

$$g(\mu_{ij}) = X_{ij}\boldsymbol{\theta} + Z_{ij}\boldsymbol{\lambda},$$

where $\boldsymbol{\theta}$ is a vector of parameters of interest, $\boldsymbol{\lambda}$ is a vector of nuisance parameters, and X_{ij} and Z_{ij} are covariates related to study design. The sample

size required for testing H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. H_a: $\boldsymbol{\theta} = \boldsymbol{\theta}_a$ is derived based on a quasi-score test statistic.

The method can be applied to generalized linear models with longitudinal observations such as linear models for continuous responses, logistic models for binomial response, and loglinear models for Poisson responses. However, there is no explicit sample size formula except for some simple special cases, for which the sample size formulas are provided in Liu & Liang [23]. The sample size or power values are estimated numerically, in general.

To calculate sample size and power, we have to specify the values of the parameters of interest as well as the nuisance parameters under the alternative hypothesis. In addition, the element of the correlation matrix for the longitudinal observations has to be specified, and this is usually the most difficult part. Information about correlation may be obtained from previous studies. In the case of no prior information about the correlation, a sensitivity analysis may be performed using various correlation values based on the investigator's judgment. The sample size for the study may be taken conservatively based on the sensitivity analysis.

On the basis of the GEE method, explicit sample size formulas are obtained by Liu et al. [22] for comparing two means, two slopes and two proportions under several simple correlation structure.

Estimation and Precision

The sample size determinations discussed so far are based on achieving a certain probability (power) to detect a given alternative for a specified statistical hypothesis test. When the problem of interest is to estimate the magnitude of the effect, e.g. the disease–exposure **odds ratio**, the study planning must focus on the precision of the estimate, which is often measured by the width of a confidence interval. In general, the sample size required for the study should be increased if the confidence level is increased, the variability associated with the measure is increased, or the total width of the confidence interval is reduced.

There are two different approaches to determining sample size based on the width of a confidence interval. One is to have the expected width of the confidence interval sufficiently small. Another is to have a tolerance level to guarantee that the width of the confidence interval will be within a given precision limit. In the latter case, the width of a confidence interval is regarded as a random variable. The latter approach usually requires larger sample sizes than the former. In this section we illustrate the two approaches to determining sample size for estimating ORs and standardized mortality ratios.

OR

Consider a 2×2 table generated from an unmatched case–control study:

Exposure	Case	Control	
Yes	n_{11}	n_{10}	$n_{1.}$
No	n_{01}	n_{00}	$n_{0.}$
	$n_{0.1}$	$n_{0.0}$	n

The OR is estimated by

$$\mathrm{OR} = \frac{n_{11}n_{00}}{n_{01}n_{10}}.$$

For the first approach, O'Neill [31] derived a sample size equation by replacing the cell counts with their expected values and then using a logit method to calculate the confidence interval width. To have the expected width of the $1-\alpha$ confidence interval less than 2δ, for given values of true OR, exposure rate π_0 for the control group, and the case–control ratio k (i.e. $k = n_{.0}/n_{.1}$), the number of cases required is

$$n_{.1} = \left[\frac{1}{\pi_1(1-\pi_1)} + \frac{1}{\pi_0(1-\pi_0)}\right]\left[\frac{z_{1-\alpha/2}}{\delta}\right]^2,$$

where π_1 is the exposure rate for the case group, which can be calculated as $\pi_1 = \pi_0\mathrm{OR}/[1 + \pi_0(\mathrm{OR} - 1)]$.

For the second approach, a $1-\alpha$ confidence interval for ln(OR) is obtained from the normal approximation (e.g. [1, Chapter IV]),

$$\ln(\mathrm{OR}) \pm z_{1-\alpha/2}\left[\frac{1}{n_{11}} + \frac{1}{n_{00}} + \frac{1}{n_{10}} + \frac{1}{n_{01}}\right]^{1/2}.$$

The precision of the estimate can be evaluated by the probability of the confidence interval being within $[-\delta, \delta]$. That is,

$$P\left(z_{1-\alpha/2}\left[\frac{1}{n_{11}} + \frac{1}{n_{00}} + \frac{1}{n_{10}} + \frac{1}{n_{01}}\right]^{1/2} \le \delta\right) = 1 - \beta.$$

The probability $1 - \beta$ is called the tolerance probability. The sample size required for the study is obtained by solving this equation for given values of the true OR, α, β, δ, the exposure rate π_0 for the control group, and the case–control ratio k. Numerical computation is needed to solve the equation for n. Satten & Kupper [35] provided tables (as well as a computer program) for the minimum sample size required to produce a 95% confidence interval of total width not greater than 2δ with probability $1 - \beta$ for various values of δ, k, OR, and π_0.

Satten & Kupper [35] also gave an example for a case–control study. The anticipated OR for the study population is no smaller than 3. For a probability of exposure in controls $\pi_0 = 0.05$, with tolerance probability $1 - \beta = 0.90$, $k = 1$, and in order to have the width of the 95% confidence interval no more than 1.5, $n = 271$ cases are required (with $k = 1$, the number of controls is the same). With O'Neill's method, the required number of cases is $n = 202$.

Standardized Mortality Ratio

The standardized mortality ratio (SMR) is the ratio between the observed number of events and the expected number of events. The latter quantity is usually derived from external studies or vital statistics and is assumed to be known and fixed (*see* **Vital Statistics, Overview**). The SMR is a common measure of **relative risk** in occupational epidemiologic studies (*see* **Occupational Epidemiology**).

Assume the observed number of events d in a cohort study follows a Poisson distribution with mean λ. The confidence interval of the SMR is obtained by finding the corresponding confidence limits for λ and then dividing these limits by the expected number of events. Using the relation between the Poisson distribution and the chi-squared distribution, Gordon [12] derived the $1 - \alpha$ confidence interval for λ as

$$(\lambda_L, \lambda_U) = \left[\tfrac{1}{2}c_{\alpha/2}(2d), \tfrac{1}{2}c_{1-\alpha/2}(2d+1)\right],$$

where $c_\alpha(k)$ is the αth quantile of a chi-squared distribution with k degrees of freedom.

The expectations of λ_L and λ_U, and the expected width of the interval are

$$E(\lambda_L) = \sum_{d=1}^{\infty} \frac{1}{2} c_{\alpha/2}(2d) \frac{\lambda^d}{d!} \exp(-\lambda),$$

$$E(\lambda_U) = \sum_{d=0}^{\infty} \frac{1}{2} c_{1-\alpha/2}(2d+1) \frac{\lambda^d}{d!} \exp(-\lambda),$$

$$E(w) = E(\lambda_U) - E(\lambda_L).$$

Tables are given by Gordon [12] for these expected values for various values of λ, which can be used to determine the expected number of events given the true SMR and an upper limit, a lower limit or the width of the confidence interval.

The sample size obtained by Gordon is based on the first approach without having a tolerance level to guarantee the precision. The estimated sample size will be too small to distinguish the SMR from a specified value [15]. For example, to have the upper bound of the 95% confidence interval no more than 1.0 when the true SMR is 0.7, the expected number of events estimated by Gordon's method is 42.9. However, this sample size will have only a 50% chance of yielding a 95% confidence interval, which will exclude the value 1.0 when the true SMR is 0.7.

On the basis of the second approach, Greenland [15] provided a method to estimate the sample size needed so that the confidence interval can reliably distinguish between two different values of the SMR. He concluded that the sample size obtained will be similar to that based on a hypothesis test to compare the two specified SMR values. For the above example, in order to have a 90% chance that the upper 95% confidence interval be smaller than 1.0 when the true SMR is 0.7, the required number of events is 98.4. This sample size can also be obtained from a hypothesis test that compares SMR = 0.7 vs. SMR = 1.0, with significance level 0.05 and power of 90%.

Further Issues and Discussion

Exact Methods

As computing technology advanced in recent years, exact methods were developed for analyses of studies with categorical observations. These methods provide exact test p-values that may be quite different from the asymptotic p-values when the sample size and/or the probability of events is small. If it is planned to use exact methods for the analysis, then it is preferable to estimate the sample size for the study based on the same exact approach.

For pair-matched **case–control studies**, a number of papers have been published for calculating sample size and power based on McNemar's test. Lachin [17] compared several unconditional sample size and power calculation methods (*see* **Matched Analysis**). Royston [34] published tables of sample sizes based on the conditional and unconditional approaches.

For ordered categorical data, Hilton & Mehta [16] developed an algorithm for computing sample size and exact power. They also considered a Monte Carlo method for power estimation. They pointed out that the asymptotic power function works well when the number of categories is not too small (e.g. more than five categories).

Lui [27] considered sample size estimation for the exact conditional test under inverse sampling, in which one continues to sample subjects until one obtains a predetermined number of index subjects (e.g. events). Under inverse sampling, the number of events to be observed for each exposure group is fixed, and the number of subjects to be sampled follows a negative binomial distribution. Conditional on the total number of subjects, a conditional test is used to compare event rates for the two groups. On the basis of a numerical approximation, tables are provided for the minimum required number of events given events rates, significance level = 0.05 and power = 0.80 or 0.90.

Adjustment for Loss to Follow-up and Missing Data

In studies where loss to follow-up or **missing data** occur, the planned sample size should reflect the information loss to maintain the desired statistical power. If loss to follow-up or missing data are purely by chance (missing completely at random), then a simple adjustment on the sample size is to enlarge the required sample size by the proportion of information loss. For example, suppose the required sample size for a cohort study is 400 subjects per group and a 20% loss to follow-up rate is assumed, then the total number of subjects for the study may be enlarged to 500 (= 400/(1−0.2)) per group. This simple adjustment provides a conservative sample size estimate if the partial data obtained from the lost to follow-up subjects can be used in the statistical analysis (e.g. in longitudinal analysis models). If we know that 25% data from the 20% lost to follow-up subjects can be used in the analysis, then a refined sample size adjustment is $n = 400/(1-0.15) = 471$.

This simple adjustment can also be used for the case when missing data are missing at random, i.e. the probability of missingness depends on at most the data already observed but not on the missing data [21]. However, when data are not missing at random (nonignorable missingness), the probability of missingness will depend on the missing data. In this case, it can be difficult to compute the proportion of missing information. Further discussion of missing data is beyond the scope of this article and can be found in Little & Rubin [21] (*see* **Missing Data in Epidemiologic Studies**).

Computation and Software

Several commercial software packages are currently available for sample size and power calculations (*see* **Software, Biostatistical**; **Software, Epidemiological**). They include EGRET-SIZ by Cytel Software Corporation, SamplePower by SPSS Inc., nQuery Advisor by Statistical Solutions, and PASS by Number Cruncher Statistical Systems. EGRET-SIZ is the only package for sample size and power calculations with the main focus on epidemiologic studies. It provides sample size estimates for four specialized statistical models including **logistic regression** (for **cohort** and unmatched **case–control studies**), **Poisson regression**, **conditional logistic regression** and **Cox** proportional hazards **regression**. A good feature of EGRET-SIZ is that there is a Monte Carlo procedure for one to verify the estimated sample size and obtain empirical power. SamplePower, nQuery Advisor and PASS provide sample size estimates for a broad range of statistical models including tests for means, proportions, analysis of variance (ANOVA), regression and survival analysis. Most of the software packages can be used to estimate one of the parameters among sample size, power, minimum detectable effect, variances, and test significance level, provided the other parameters are specified. Some of the packages (e.g. nQuery, SamplePower) have modules to estimate sample size from the given width of a confidence interval.

Other "freeware" may be found in public health service organizations or from individual statisticians. For example, "Epi info" from the Centers for Disease Control and Prevention provides some sample size estimation routines for cohort and case–control studies. A SAS module/macro for sample size analysis, "UnifyPow", has been developed and distributed

by O'Brien [30]. Many computer codes for some specific and complex statistical methods may be obtained directly from the authors who developed these methods. For example, a power and sample size module for testing **interaction** has been developed by Lubin and colleagues at the National Cancer Institute based on a paper by Lubin & Gail [24]. A sample size calculation module is included in EPITOME, an epidemiologic data analysis package developed at the National Cancer Institute.

Although many approaches and software packages can be used to calculate sample size, care should always be taken to formulate the study problem in the framework that the computer software requires. It is necessary to follow the program instructions or user's manual to provide input parameters for the computer packages. In all cases, it is important to understand the statistical procedures for which the computer package is calculating the sample size or power. Otherwise, the calculated sample size can be erroneous and lead to an underpowered or overpowered study.

Discussion

Sample size determination is a very important aspect of the design of any study. It often helps to clarify important features of a study protocol. Investigators will not be able to calculate sample size without fully understanding the nature of the measurements to be taken, specifying the planned analysis (test statistic or estimation procedure, significance level and study power), and obtaining preliminary information on the effect to be detected and the variability of the measurements. Some important information can often be obtained from reviewing the literature or discussions with other investigators who conducted previous studies.

In practice, sample size and power evaluation may be an iterative learning process. The parameters and distribution characteristics obtained from previous studies can be used to estimate the required sample size for planning the current study. If there is no previous information for a new investigation, then a pilot study may be designed to gain some knowledge about the parameters and distribution. The results from this pilot study will then provide information for designing the main study. The iterative process may be integrated to form a two-stage design in which the results from the pilot study (first stage) will be used to guide the sample size estimation for the second stage. The data from both stages will be used for the final data analysis. For instance, the impact of additional follow-up in cohort studies was investigated by Brookmeyer et al. [3]. In randomized clinical trials, Gould [13] and Gould & Shih [14] presented methods to adjust the sample size during the course of a study.

Group sequential procedures have been considered for experimental studies by health researchers, in which investigators are allowed to have several interim analyses. Each interim analysis not only provides a preliminary estimation of the parameters of interest before the completion of the study, but also offers a chance to terminate the study early when there is sufficient evidence to reach a conclusion (either positive or negative). Although this sequential approach has mainly been used and advocated for **experimental** and **case–control studies** (*see* **Case–Control Study, Sequential**), the idea can be used in **cohort studies** by performing interim analyses when partial follow-up data are available. However, there is usually less ethical pressure for early termination of **observational** epidemiologic studies than **experimental studies** such as randomized clinical trials. Furthermore, the need for a large sample size is often stressed in epidemiologic studies in order to estimate the parameters of interest with precision. For these reasons, sequential methods have not gained widespread acceptance and use in epidemiologic studies.

Usually, new statistical methods are developed for data analysis before they are considered in sample size estimation for designing studies. For some complicated study designs and statistical models, sample size estimation methods may not be available (e.g. case–cohort design, two-stage case–control studies, and structural models for causal inference; see [33]). Further research will be needed on sample size estimation for these specialized study designs and statistical models (*see* **Case–Cohort Study**; **Case–Control Study, Two-phase**).

Two approaches may be considered to estimate sample size for a complicated study when there is no sample size estimation method available. First, a Monte Carlo simulation may be conducted to estimate the power empirically for several fixed sample sizes and thus to find the sample size that yields the required power. Today's powerful computation tools make this approach feasible. The second approach is

to approximate the sample size by using a simplified study design and/or statistical model for which a sample size estimation method is available. The simple model should be chosen so that it requires at least as large a sample size as would be required by the more complex and presumably more efficient analysis. There is a tradeoff between using simpler methods and using more sophisticated models for sample size estimation. At the design stage, prior information about the parameters of the statistical models may be limited. Sample size estimation requires fewer assumptions with simplified statistical models than with more sophisticated models. Unless the assumed values of parameters and forms of distributions are accurate for the designed study, the sample size estimated from the sophisticated models may be inaccurate due to misleading assumptions. Simplified methods, on the other hand, can be more robust because of fewer assumptions.

Acknowledgment

The author would like to thank Dr Jacques Benichou for his detailed review, comments and suggestions. Thanks also to Dr Mitchell Gail for his valuable comments and some additional references.

References

[1] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research, Volume I – The Analysis of Case–Control Studies*. International Agency for Research on Cancer, Lyon.

[2] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research, Volume II – The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[3] Brookmeyer, R., Day, N. & Pompe-Kirn, V. (1980). Assessing the impact of additional follow-up in cohort studies, *American Journal of Epidemiology* **121**, 611–619.

[4] Brown, C.C. & Green, S.B. (1980). Additional power computations for designing comparative Poisson trials, *American Journal of Epidemiology* **115**, 752–758.

[5] Diggle, P.J., Liang, K.Y. & Zeger, S.L. (1980). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[6] Dupont, W.D. (1980). Power calculations for matched case–control studies, *Biometrics* **44**, 1157–1168.

[7] Fleiss, J.L. & Levin, B. (1980). Sample size determination in studies with matched pairs, *Journal of Clinical Epidemiology* **41**, 727–730.

[8] Freedman, L.S. (1980). Tables of the number of patients required in clinical trials using the logrank test, *Statistics in Medicine* **1**, 121–129.

[9] Gail, M. (1980). Power computations for designing comparative Poisson trials, *Biometrics* **30**, 231–237.

[10] Gail, M. (1980). Sample size estimation when time-to-event is the primary endpoint, *Drug Information Journal* **28**, 865–877.

[11] Garcia-Closas M. & Lubin J. (1999). Power and sample size calculations in case–control studies of gene-environmental interactions: Comments on different approaches, *American Journal of Epidemiology* **149**, 689–693.

[12] Gordon, I. (1980). Sample size estimation in occupational mortality studies with use of confidence interval theory, *American Journal of Epidemiology* **125**, 158–162.

[13] Gould, A.L. (2001). Sample size re-estimation: recent developments and practical considerations, *Statistics in Medicine* **20**, 2625–2643.

[14] Gould, A.L. & Shih, W.J. (1980). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance, *Communications in Statistics A – Theory and Methods* **21**, 2833–2853.

[15] Greenland, S. (1980). On sample-size and power calculations for studies using confidence intervals, *American Journal of Epidemiology* **128**, 231–237.

[16] Hilton, J.F. & Mehta, C.R. (1980). Power and sample size calculations for exact conditional tests with ordered categorical data, *Biometrics* **49**, 609–616.

[17] Lachin, J.M. (1980). Power and sample size evaluation for the McNemar test with application to matched case–control studies, *Statistics in Medicine* **11**, 1239–1251.

[18] Lachin, J.M. & Foulkes, M.A. (1980). Evaluation of sample size and power for analysis of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification, *Biometrics* **42**, 507–519.

[19] Lakatos, E. (1980). Sample sizes based on the log-rank statistics in complex clinical trials, *Biometrics* **44**, 229–241.

[20] Lakatos, E. & Lan, K.K.G. (1980). A comparison of sample size methods for the logrank statistic. *Statistics in Medicine* **11**, 179–191.

[21] Little, R.J. & Rubin, D.B. (1980). *Statistical Analysis with Missing Data*. Wiley, New York.

[22] Liu, A., Shih, W.J. & Gehan, E. (2002). Sample size and power determination for clustered repeated measurements, *Statistics in Medicine* **21**, 1787–1801.

[23] Liu, G. & Liang, K.Y. (1980). Sample size calculations for studies with correlated observations, *Biometrics* **53**, 937–947.

[24] Lubin, J.H. & Gail, M.H. (1980). On power and sample size for studying features of the relative odds of disease, *American Journal of Epidemiology* **131**, 552–566.

[25] Lubin, J.H., Gail, M.H. & Ershow, A.G. (1980). Sample size and power for case–control studies when exposures are continuous, *Statistics in Medicine* **7**, 363–376.

[26] Lui, K.J. (1980). Estimation of sample sizes in case–control studies with multiple controls per case,

dichotomous data, *American Journal of Epidemiology* **127**, 1064–1070.

[27] Lui, K.J. (1980). Sample size for the exact conditional test under inverse sampling, *Statistics in Medicine* **15**, 671–678.

[28] Munoz, A. & Rosner, B. (1980). Power and sample size for a collection of 2×2 tables, *Biometrics* **40**, 995–1004.

[29] Nam, J. (1980). Sample size determination for case–control studies and the comparison of stratified and unstratified analyses, *Biometrics* **48**, 389–395.

[30] O'Brien, R.G. (1980). A tour of UnifyPow: A SAS module/macro for sample size analysis, *Proceedings of the Twenty-third SAS Users Group International Conference*, Cary, NC, 1346–1355.

[31] O'Neill, R.R. (1980). Sample sizes for estimation of the odds ratio in unmatched case–control studies, *American Journal of Epidemiology* **120**, 145–153.

[32] Qiu, P., Moeschberger, M.L., Cooke, G.E. & Goldschmidt-Clermont, P.J. (2000). Sample size to test for interaction between a specific exposure and a second risk factor in a pair-matched case–control study, *Statistics in Medicine* **19**, 923–935.

[33] Rothman, K.J. & Greenland, S. (1980). *Modern Epidemiology*. Lippincott-Raven, Philadelphia.

[34] Royston, P. (1980). Exact conditional and unconditional sample size for pair-matched studies with binary outcome: a practical guide, *Statistics in Medicine* **12**, 699–712.

[35] Satten, G.A. & Kupper, L.L. (1980). Sample size requirements for interval estimation of the odds ratio, *American Journal of Epidemiology* **131**, 177–184.

[36] Schlesselman, J.J. (1980). *Case–control Studies: Design, Conduct, Analysis*. Oxford University Press, Oxford.

[37] Schoenfeld, D.A. (1980). Sample size formula for the proportional-hazards regression model, *Biometrics* **39**, 499–503.

[38] Self, S.G. & Mauritsen, R.H. (1980). Power/sample size calculations for generalized linear models, *Biometrics* **44**, 79–86.

[39] Self, S.G., Mauritsen, R.H. & Ohara, J. (1980). Power calculations for likelihood ratio tests in generalized linear models, *Biometrics* **48**, 31–39.

[40] Smith, P.G. & Day, N.E. (1980). The design of case–control studies: the influence of confounding and interaction, *International Journal of Epidemiology* **13**, 87–93.

[41] Taylor, J.M.G. (1980). Choosing the number of controls in a matched case–control study: some sample size, power and efficiency calculations, *Statistics in Medicine* **5**, 29–36.

[42] Ury, H.K. & Fleiss, J.L. (1980). On approximate sample sizes for comparing two independent proportions with the use of Yates' correction, *Biometrics* **36**, 347–351.

[43] Wickramaratne, P.J. (1980). Sample size determination in epidemiologic studies, *Statistical Methods in Medical Research* **4**, 311–337.

[44] Whittemore, A.S. (1980). Sample size for logistic regression with small response probability, *Journal of the American Statistical Association* **76**, 27–32.

[45] Woolson, R.F., Bean, J.A. & Rojas P.B. (1980). Sample size for case–control studies using Cochran's statistic, *Biometrics* **42**, 927–932.

GUANGHAN LIU

Sample Surveys in the Health Sciences

Reasons for Conducting Sample Surveys

Perhaps the most compelling argument for using a sample survey rather than complete enumeration is that for the same cost a sample survey can provide results more accurately, with greater scope, and faster. Studying a well-chosen sample can increase accuracy by reducing **bias** and by increasing precision of the results. By reducing the number of people to be studied, a sample survey can devote more resources to finding and persuading nonresponders to participate, thus reducing **nonresponse**. Fewer interviewers are required, so that more effort can be devoted to training and monitoring them for accuracy. Replicate measurements may be made to increase the precision. Sample surveys can also take advantage of the smaller sample size and more intensive effort for each person to ask more questions or make more detailed measurements, thus broadening the scope of the questions addressed. Finally, sample surveys may be able to reduce the total time to collect and analyze the data, thus providing more timely answers to important questions.

A second important reason for the use of sample surveys is feasibility. When the **target population** is very large, such as the entire population of the United States, a moderate-sized sample survey, if well designed, can provide highly accurate results at substantially less cost than a complete enumeration. The cost of a complete enumeration can be substantial both for the researchers and for the participants; one motivation for using sample surveys is when the respondents are institutions such as hospitals.

The desire to study certain subgroups in more depth also leads people to use sample surveys. Policy issues may require valid estimates for children, people aged 65 and older, women, African-Americans, Hispanics, rural residents, or other subgroups in the population. A sampling design can oversample from important subgroups to increase precision and to ensure that their health can be characterized accurately.

Recently, epidemiologic studies have begun to take advantage of the ability of sample survey designs to increase **power** and reduce bias in studies of risk factors for disease onset and progression (*see* **Observational Study**). The power to conduct comparative analyses of risk factors is largely determined by the number of **prevalent** or **incident** disease cases identified for study. To increase the number of cases, researchers can increase the sample size or the length of follow-up, an expensive strategy, or try to increase the proportion of disease cases in the sample by oversampling groups at high risk. A clever design can improve power substantially. Sample surveys also help reduce bias for epidemiologic studies by providing a truly comparable group of unaffected people, sampled from exactly the same population in which the affected group was identified (*see* **Controls**). A recent study of Alzheimer's disease in people aged 65 and older in the community in East Boston, Massachusetts, used a **stratified sample** design to oversample from the oldest age groups and those with poor performance on a simple memory test. When a neurologist examined the resulting sample, about 35% were found to have clinical Alzheimer's disease, compared with an estimated prevalence of 10% in the community. In addition, the **mean** age of the unaffected comparison group was much closer to that of the Alzheimer's group in this sample than in the community, and both diseased and disease-free participants had received the same interviews and clinical evaluation [13].

Some History of Sample Surveys

Early Developments at the US Bureau of the Census

In the early part of the twentieth century, the importance of using **random sampling** in surveys was not generally recognized. Units to be canvassed were still selected purposively, with the survey managers deciding which units would be most "representative" of the population of interest (*see* **Quota, Representative, and Other Methods of Purposive Sampling**). **Probability theory** had yet to be applied. In the 1930s, more attention was being paid to **R. A. Fisher's** work on the importance of **randomization** in **experimental design** In 1934, **J. Neyman's** paper [43] arguing in favor of random sampling and establishing the theoretical foundation for it, appeared in the *Journal of the Royal Statistical Society* (see also [55]). These developments in statistical theory laid the groundwork for modern sample surveys. However, the key

catalyst for the incorporation of probabilistic sampling into the selection of units in sample surveys (*see* **Probability Theory**) was the increasing need for reliable estimates to use in policy-making.

At the US Bureau of the Census, the first application of probabilistic sampling was in a 1937 "check census" of unemployment. The Bureau used postmen as canvassers, choosing two out of every 100 postal routes. The success of this early foray into probabilistic sampling was possibly a decisive moment for those decision makers who were skeptical of the value of nonpurposive methods [11].

In the 1940 decennial **census** of population and housing, the first "long form" sample was introduced: 5% of the population was asked a set of questions in addition to the key census battery. This approach is still an integral part of the decennial census, enabling the Census Bureau to collect more detailed data without overburdening all respondents, and at a moderate cost [1, 18, 49].

In 1942, the Bureau was given the Work Projects Administration's Monthly Report on the Labor Force (MRLF), from which were derived unemployment estimates. In order to develop an efficient sample design for this survey, the Bureau statisticians under the direction of Morris Hansen and William Hurwitz found that they had to develop totally new theory for the design of sample surveys. Major new developments included sampling with unequal probabilities, **cluster sampling**, optimization in **multistage sampling**, and **estimation** methods. These are now considered standard sampling methods for face-to-face household surveys. Under the direction of Hansen and Hurwitz, a relatively small research staff made further contributions to sampling methods. One of the more public products of this work, a two-volume text by Hansen, Hurwitz, and William G. Madow published in 1953 [22], still stands as a classic work in statistics.

The MRLF, later named the Current Population Survey, provided a laboratory for research into sampling and other statistical aspects of survey methods. It also became a model for household surveys all over the world. Other current US demographic surveys have used the same basic design. These include: The National Crime Survey (now The National Crime Victimization Survey), The National Health Interview Survey, The Survey of Income and Program Participation, The American Housing Survey (formerly The Annual Housing Survey), and The Consumer Expenditure Quarterly and Diary Surveys. Data for all of these surveys are collected by the Bureau of the Census, typically for sponsoring agencies who publish the estimates from the surveys. Most of these surveys use as their basic frame (*see* **Sampling Frames**) the list of housing units obtained from the decennial census; this list is supplemented by frames for new construction and other special categories. Characteristics of those housing units and their occupants are used to stratify units within counties and to group counties within primary strata.

Extensions of the Survey Methodology to Other Fields of Study

The combination of the success of the sample survey method developed and used by the Census Bureau in the 1940s and 1950s in providing population, housing and economic data, and the need for new data on health characteristics of the population led to the adoption of sampling techniques in conducting national health surveys, beginning in 1957 and continuing until the present time (*see* **Surveys, Health and Morbidity**).

It had been established in the 1920s that community studies of illness and disability were feasible, and a major health survey to obtain data on diseases, injuries, and impairments in the general population of the United States was conducted from 1935 to 1936 [41]. Following this survey, additional community studies on morbidity led to the formation of the US National Committee on Vital and Health Statistics in 1949, which ultimately recommended "That a continuing national morbidity survey be conducted... Its purpose would be to obtain data on the prevalence and incidence of disease, injuries and impairments, on the nature and duration of the resulting disability, and on the amount and type of medical care received. The data would be obtained from a probability sample of households" [41]. Thus, what is now known as the National Health Interview Survey (NHIS) was begun.

The data from these continuing surveys have been used extensively by the US Federal government in setting policy and developing programs for the continued benefit of the population. Perhaps, the largest such program ever enacted is the national health insurance program for the elderly, known as **Medicare**, which came into existence some eight years after the NHIS began producing relevant data on the health of the general population. In addition, data

derived from the NHIS were used in developing the 1964 recommendations of the US Surgeon General regarding **smoking and health**.

Today, data from both national surveys and community population studies are being used to understand causes and prevention of disease. For example, most of what we know about the **risk factors** for coronary heart disease came from the results of the long-standing community study known as the **Framingham Heart Study** Currently, community studies utilizing sampling techniques are being conducted in Hawaii, Illinois, Washington (state), and other areas to determine risk factors for Alzheimer's disease and other dementing illnesses, which may aid researchers in developing preventive strategies for the future. Data from a longitudinal supplement to the NHIS and from a national long-term care survey have suggested that a decline in disability among older people has been taking place over time [33]. These data will be useful for planning health care services for the elderly for the future.

Design and Objectives of Sample Surveys in Health and Medical Studies

The specification of the design and objectives of a sample survey form a very important part of the planning of the study. A clear statement of objectives is essential for the researcher to be able to stay on track and design the study effectively. One can become so engrossed in the details of planning that one loses sight of the overall purpose for the study, and perhaps makes decisions that may contradict the objectives that were originally set.

Objectives of Sample Surveys

As surveys are usually of two types, **descriptive** and **analytic**, the statement of objectives should include the main purpose. Most large-scale surveys are usually of the descriptive type, although analytic uses of the data may ultimately be made. An example of a statement of objectives that follows comes from a publication of the **National Center for Health Statistics** (NCHS) regarding the redesign of the NHIS in 1995 [2]:

> Improving the reliability of estimates for Hispanic persons
>
> Improving the reliability of estimates for subnational areas, including states
>
> Continuing to have NHIS serve as a sampling frame for follow-on surveys

These objectives along with others were used in developing the criteria for the redesign of the survey, and, while they serve as an example for us, the careful statements enabled the NCHS statisticians to work through the design phase without losing sight of where they should be going.

Major Design Features of Sample Surveys

While there are numerous features of sample surveys that could be discussed in this article, space dictates that we limit our discussion to only a few of the main features seen in most sample surveys executed in practice. Recall the textbook definition of **simple random sampling** in a finite population, namely, a method of selecting n units out of N such that each of the possible combinations of N units taken n at a time has an equal chance of being chosen (see, for example, [5]). In addition to simple random sampling, survey statisticians often employ methods of sampling including, but not limited to, stratification (*see* **Stratified Sampling**) and clustering (*see* **Cluster Sampling**), and compute estimates using techniques such as ratio estimation (*see* **Ratio and Regression Estimates**) and **poststratification** Following discussion of these topics, we will conclude the section with some information about sources of error in surveys.

Stratification. Use of **stratification** in sample surveys involves first the division of the population of N units into L nonoverlapping subpopulations, called strata, of size $N_1, N_2, \ldots, N_L$. In order to obtain maximum benefit from selecting a stratified sample, the number of units, N_i, in each stratum must be known. The sample is then drawn independently from each stratum.

There are numerous reasons for using stratification in sample surveys. As stated in the introductory section, one reason for stratification is to insure that one can make estimates of a certain level of precision for subgroups of the population under study, by fixing sample sizes separately for each subgroup. Secondly, if it is known in advance of conducting the survey that characteristics to be estimated from

the survey vary at substantially different rates from one subgroup to another, it is possible to achieve increased precision in the overall population estimates by creating strata in which within-stratum variability is small and between-stratum variability is large. For example, in the National Hospital Discharge Survey [9], larger hospitals (in terms of number of beds) tend to be more alike among themselves than they are like smaller hospitals, and similarly for the smaller hospitals. The **variance** of within-stratum estimates of a discharge rate may be quite small compared to the variance of rates among strata. A third reason for stratification is that the population may have natural divisions that lend themselves to stratification, or layering. For example, field offices for a given survey may be scattered throughout a large geographic area, thus dictating that the sample be stratified according to the geographic breakdowns inherent in the total population.

To illustrate the computations involved in estimating a population mean from a stratified sample and its sampling variance, consider a population that is divided into two strata with N_1 units in the first stratum and N_2 units in the second. The **unbiased** estimate of the mean is $\overline{x}_{\text{st}} = F\overline{x}_1 + (1 - F)\overline{x}_2$, where $\overline{x}_1$ and $\overline{x}_2$ are the respective within-stratum sample means and $F = N_1/N$ and $1 - F = N_2/N$. The variance of the estimate of the sample mean is

$$s_{\overline{x}_{\text{st}}}{}^2 = F^2 s_1{}^2 \left(\frac{1}{n_1} - \frac{1}{N_1}\right) + (1 - F)^2 s_2{}^2 \left(\frac{1}{n_2} - \frac{1}{N_2}\right), \qquad (1)$$

where $s_1{}^2$ and $s_2{}^2$ are the sample variances within the two strata, and n_1 and n_2 are sample sizes within the strata. If the within-stratum variances are sufficiently small, the overall variance of the estimated mean could turn out to be smaller than the mean of a simple random sample. Finally, the results shown here can be extended to estimation of population totals, proportions, or other characteristics by algebraically manipulating the formulas given here for the mean.

As an example of the increased precision that can be achieved through stratification, consider the following data set taken from the Honolulu Asia Aging Study (HAAS), a study of dementing illness among Japanese-American men living in Hawaii [57]. The sample to be studied was selected from three groups based on the individuals' performance on a screening test for cognitive function. Those persons who performed the poorest – and therefore were at highest risk of dementia – were chosen at the highest rate. The "good" performers were selected at the lowest rate, and those in the middle at an intermediate rate, the objective being to obtain the largest number of diseased cases possible while not ruling out the possibility that at least a few of the "good" performers may also have been at risk for dementia. The overall estimate of the prevalence of dementia among these men was 9.3% with a **standard error** 1.6% when a simple **binomial** model was used to calculate the variance. However, when the stratified sampling assumptions were taken into account, as above, the standard error estimate was reduced to 0.83%. Thus, the stratified design in this case led to a standard error slightly more than half that of a simple random sample.

Clustering. Clustering, or sampling in which the units sampled are chosen in groups or *clusters* of smaller units, called elements, is used for two reasons. First, for many surveys a *sampling frame*, or list of population units to be sampled, does not exist. For example, if one were asked to design a sampling plan for estimating the number of trees in a given geographic area, say the state of California, clearly no list of population units exists. Furthermore, the construction of such lists might be impossible for some studies, while for others it might be feasible but prohibitively expensive. However, from maps of geographic regions or lakes or whatever areas are to be sampled, it is possible to divide the region into subregions with definable boundaries. These subregions, which contain clusters of the sampling units of interest, are selected for study because they solve the problem of constructing a list of sampling units.

The second reason for selecting cluster samples is purely economic. Suppose that one were interested in studying the characteristics of physicians in office-based practices in the United States. It is known that the American Medical Association maintains an up-to-date listing of all such physicians for the United States. However, if one were to select a simple random sample of these physicians and send interviewers to collect the data from them, the interviewers would be traveling all over the country at tremendous expense to reach what would undoubtedly be a widely scattered sample of physicians. A more practical approach to conducting the study would be to select a relatively small sample of geographic areas

around the nation and conduct interviews with a sample of physicians limited to those selected areas. Clearly, a simple random sample of 3000 physicians would cover the nation more evenly than 100 counties or metropolitan areas containing an average of 30 physicians each, but greater field costs in locating the doctors and traveling from place to place to interview them would outweigh the precision obtainable with the simple random sample.

The choice of cluster size involves balancing costs versus precision for a given survey and can become rather complicated, especially if the design involves several stages of sampling. Rather than becoming engrossed in an overly complicated analysis regarding cluster sampling, let us look at a simple design in which a simple random sample of n clusters is selected from N clusters in the population, with simple random selection of m elements (out of M) within each cluster. Let x_{ij} represent the observed value for the jth element in the ith cluster and let x_i be the cluster total. Here we need to distinguish between two kinds of means: the mean per cluster $\overline{X} = \sum x_i/N$ and the mean per element $\overline{\overline{X}}/M = \sum x_i/NM$. Sampling at two levels thus introduces **variance components** for the effect of sampling clusters and elements within clusters. The between-cluster component can be calculated from the sample as

$$s_b^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} \tag{2}$$

and the within-cluster component as

$$s_w^2 = \sum_{i=1}^{n} \frac{1}{m-1} \sum_{j=1}^{m} (x_{ij} - \overline{x_i})^2, \tag{3}$$

where $\overline{x_i}$ is the cluster mean for the ith cluster. Then an unbiased estimator of the variance among all elements in the population is

$$V = \frac{(N-1)s_b^2 + N(M-1)s_w^2}{NM-1}. \tag{4}$$

As stated earlier, these considerations can be extended to multiple levels of sampling, unequal numbers of elements within the clusters and stratification of clusters prior to sampling, and stratification of elements within clusters before sampling.

One other aspect of cluster sampling deserves mention here, namely, the concept of intracluster, or intraclass, **correlation** Characteristics of individuals occupying the same cluster are often likely to be correlated. For example, in a household health survey, it would not be unusual to find correlated responses among members of a household. The intracluster correlation coefficient is defined to be

$$\rho = \frac{\mathrm{E}(x_{ij} - \overline{x})(x_{ik} - \overline{x})}{\mathrm{E}(x_{ij} - \overline{x})^2}. \tag{5}$$

Then, for the sampling design described above, the variance of the sample mean per element can be written in terms of the intracluster correlation coefficient as

$$V(\overline{x}) = \frac{1-f}{n} \frac{NM-1}{M^2(N-1)} V[1 + (M-1)\rho], \tag{6}$$

where $f = n/N$, the sampling fraction, and ρ is the intracluster correlation coefficient. As a final note, intracluster correlation is closely related to the idea of **overdispersion** For further information on cluster sampling and related topics, the reader is referred to classic sampling texts such as [5] or [22].

Ratio Estimation. Ratio estimation is a method in which the statistician takes advantage of known correlation between a characteristic to be estimated from a survey and an auxiliary variable available from a source independent of the survey in order to increase the precision of the survey estimate. Suppose that a variable y_i is available for every unit in the sample and that y_i is correlated with x_i, the variable of interest in the survey. Suppose further that the population total Y of the auxiliary variable is known. Then the ratio estimator of X, the population total of the x's, is

$$\hat{X}_R = \frac{x}{y} Y = \frac{\overline{x}}{\overline{y}} Y, \tag{7}$$

where x and y are the totals of the x_i and y_i, respectively. Similarly, the population mean could be estimated by replacing the total Y by the population mean of the auxiliary variable. The gain in precision obtained by calculating a ratio estimate can be seen in the approximate formula for the variance of the ratio, given by

$$V(\hat{X}_R) = \frac{N^2(1-f)}{n}(S_x^2 + R^2 S_y^2 - 2R\rho S_x S_y), \tag{8}$$

where ρ is the correlation between x and y and $R = X/Y$. Notice that if x and y are highly correlated, the

variance of the ratio estimator is diminished by the large value of that correlation. The ratio estimator is also biased in most applications, but in large samples the bias is negligible.

Sources of Error in Surveys. In what we have presented so far, the only errors ascribed to sample surveys have been those arising from the fact that only a sample of units are measured instead of the entire population. However, in complex surveys that involve multiple measures of quantities, which may be difficult to measure, additional errors not related to sampling may be present. In what follows, we describe four sources of such errors, sometimes referred to as **nonsampling errors**.

First, the sample may not adequately cover the universe of units of interest. Secondly, it may not be possible to measure some of the units in the population chosen for the sample. This may occur for a variety of reasons, including inability of the fieldwork team to locate certain individuals selected for the sample, or the respondents' refusal to answer some or all of the questions being asked in the survey. Thirdly, the measuring device may not be able to determine accurately the characteristics being measured, or sample individuals may not understand the question or may not know the correct answer to the question. Finally, errors may arise in the recording, coding, editing, and tabulation of the data (*see* **Data Management and Coordination**). The statistician may find it necessary to modify standard statistical procedures to account for the occurrence of such errors in order to make valid inferences from the data when nonsampling errors are present.

Errors of *coverage* arise when the sampling frame does not fully cover the universe of units to which the sample estimates are to be generalized (i.e. the target population). For example, suppose that in 1996 one uses the list of housing units from the 1990 decennial census as a frame for a sample of households for a US survey. Without supplemental frames, this list would exclude housing units constructed since the census. Thus, any estimates based on the sample would be generalizable only to the list of units, whereas the true population of interest is "all housing units in the US in 1996". Such coverage error could bias the results, because the people living in the newly constructed units might be different, on the variables of interest, than those living in the older housing units. Even if we supplemented the list with a frame that captured new construction since the census (for example, based on an ongoing survey of building permits or new construction), we might still have coverage error. This is because even the best lists are subject to errors. Lists constructed from door-to-door canvassing and listing of units could be incomplete if some unusual housing units were missed (e.g. a carriage house turned into a rental unit). Some lists, such as those constructed from a census, may not be complete if there was nonresponse to the census. And commercial providers of lists, such as professional associations, will only be able to supply lists that are as complete as the information their members provide.

A related problem is the potential discrepancy between the true population of inference, such as "all people in the US over the age of 18", and the target population, which might be "all people in the US over the age of 18 at a particular point in time". Although not a coverage error *per se*, it is an important issue to consider when defining the research question.

Both types of problems are related to the concept of "external validity".

Nonresponse, or failure to measure some of the units in the sample, is probably the most common nonsampling error incurred in survey practice. There are few, if any, surveys that do not experience at least some level of nonresponse. In the case of household surveys or surveys involving human respondents, one way of dealing with nonresponse is recontact with the nonrespondents in an attempt to obtain the required data. This may take the form of repeated visits to the household or repeated telephone calls (*see* **Callbacks and Mail-backs in Sample Surveys**). In some cases, it is possible to make a valid estimate of the characteristic under study by recontacting a sample of the nonrespondents. Whatever method is used to obtain complete data, it is likely that a "hard core" of nonrespondents will persist in failing to provide the requested data. If the percentage of nonresponse is relatively large, say greater than 5%, it is quite likely that the results of the study will be biased an unknown amount by the exclusion of those individuals who did not provide complete data. As an example, consider a disability survey in which persons are asked about their ability to perform certain activities of daily living. Studies have shown that the people who have the most difficulty in performing those activities are the ones who are most likely to refuse to answer the questions. Therefore, estimates of the prevalence of disability based on complete

responses to the questions are lower than they would be if the more disabled individuals had answered the questions. Also, because the sample size is smaller than if complete response had been obtained, the standard errors of the estimates will be correspondingly larger. This, however, can be remedied if the statistician anticipates the loss of sample size and increases the sample size accordingly at the design stage.

A considerable body of literature on the adjustment of survey estimates for **missing data** exists. These techniques mostly involve weighting adjustments and so-called imputation procedures, which have been studied by numerous authors (*see* **Multiple Imputation Methods**). For a useful summary of these methods, see [32].

Analysis of variance models have been applied in the study of **measurement errors** in sample surveys in much the same way as in **experimental studies** The simplest models assume that a measurement includes the true value of what is being measured plus an error term. However, when the errors depend in some way on the value of the characteristic or are correlated with the item being measured, the models must necessarily become more complicated. One way to attempt to determine the correct value for an item is to remeasure it by an independent method that is more accurate than the original method. In many surveys, for example, a subsample of respondents will be reinterviewed by the best interviewers to assess the correctness of the data obtained in the original interview. Other methods might involve embedding controlled experiments in surveys or subdividing the sample into groups so that there is no correlation between the groups. Many of these topics have been carefully reviewed in articles in the literature, such as [20, 21].

The fourth area of nonsampling errors dealing with recording, coding, editing, and tabulating data will be discussed in the section on data management, later in this article.

Examples of Large-scale Surveys in Current Use in Health Research

Surveys of the US National Center for Health Statistics

The National Health Interview Survey. As stated previously in this article, NHIS was begun by the US Public Health Service in 1957, and has continued on an annual basis since that time. It is one of the major components of the National Center for Health Statistics of the **Centers for Disease Control and Prevention** The NHIS produces information on the health of the US civilian noninstitutionalized population, collected by the US Bureau of the Census in household interviews throughout the United States. The sample design for this study has been evaluated and modified after each succeeding census during the survey's existence, but currently available data do not yet reflect the redesign completed following the 2000 census. The description that follows pertains to the design used to collect the current data, a design developed following the 1990 decennial census.

In concept, the design of the NHIS has remained essentially the same since 1957. That is, the sampling plan follows a stratified multistage probability design, which permits continuous sampling of the target population. The sample of households interviewed each week is representative of the nation and the weekly samples are additive over time. This allows great flexibility in the agency's ability to respond to rapidly changing data needs.

The basic features of the design include a first-stage selection of a large number of primary sampling units (PSUs) (*see* **Sampling in Developing Countries**). This is accomplished by **sampling with probability proportional to size** (**pps**) from an area frame supplemented by a frame of building permits to enable the inclusion of housing units constructed since the completion of the previous census. Approximately one-quarter of the PSUs are self-representing; that is, they are chosen with certainty. These PSUs are primarily metropolitan statistical areas, which usually consist of a large city and its suburban areas. The non-self-representing PSUs are single counties, or groups of contiguous counties. Within the PSUs, clusters of approximately eight households are selected in the area frame and four households in the permit frame. This results in a yearly expected number of interviewed households of about 40 000 and about 110 000 interviewed persons.

The respondent rules for the NHIS allow a single individual over the age of 17 to respond for all persons dwelling in the household. However, if other persons over age 17 are available, they are invited to respond for themselves. Historically, between 65 and 70% of adults have been self-respondents. They answer questions for a set of basic health and demographic items. In addition, one or more sets of

questions on current health topics are typically asked. Also, a random subsample of adult respondents is generally asked to respond to additional questions on current health topics, which vary from year to year. Questionnaire topics include demographic characteristics such as age, sex, race, education, marital status, and family income. Health characteristics measured include disability days, physician visits, acute and chronic conditions, long-term limitations of activity, and short-stay hospital utilization. In addition, subsets of households are asked about selected chronic conditions. In the supplements that vary from year to year, special health items are asked in such areas as alcohol use, dental care, health insurance, aging, health promotion and disease prevention, vitamin and mineral intake, functional limitations, and risk factors for certain chronic diseases. Data from the survey are regularly published in the NCHS Vital and Health Statistics Series 10 reports, as well as in professional journals in the scientific literature. Standardized public-use micro-data tapes and CD-ROMs are also made available for purchase.

As a final word on the NHIS, it should be pointed out that the redesign developed and implemented in 1985 and continued in 1995 includes a feature that enables NCHS to integrate the survey designs of several of its population surveys. This was accomplished by using the NHIS sample as a sampling frame for the other surveys. In this way, the surveys could be linked analytically and possibly duplication of data collection could be avoided. Also, NHIS information could be used to oversample subgroups of the population in order to achieve sufficient sample size for studying groups, which otherwise would have been underrepresented in the surveys. The successful application of this method to the design of the NHIS has led to considerably increased efficiency in the overall designs of NCHS surveys in recent years. For more details on the research leading to the current NHIS design, the reader is referred to [2]. The updated design that was put in place in 1995 will be used until 2004.

The National Health and Nutrition Examination Survey. The National Health and Nutrition Examination Survey (NHANES) is one of the NCHS surveys now linked to the NHIS through the integrated survey design concept. At its inception in 1971, however, the NHANES was conducted using an independent design. The purpose of the initial cycle of NHANES, now known as NHANES I, was to measure the nutritional status of the US population and monitor changes in that status over time. The nutrition component represented an expansion of a previous series of three cycles of national health examination surveys, which had been completed on subsets of the US population between 1959 and 1970. As in the previous cycles and in the NHIS, the target population was the civilian noninstitutionalized population of the US, only for this survey, the population was limited to ages from 1 to 74 because of a belief that older individuals would not respond to an examination survey as readily as younger people. It was also determined at the outset that emphasis should be placed on studying individuals believed to be at increased risk of having poor nutritional status, including segments of the population classified as at or below the poverty level, young children, and the aged. Hence, oversampling of these segments of the population yielded a sample with sufficient numbers to study these characteristics.

Examinations were carried out in three mobile examination centers that traveled to the primary sampling units (PSUs) chosen in the first stage of sampling for the survey. Within each PSU, a sample of households was drawn – as in the NHIS – but a single individual from each household was selected to be examined in the mobile clinic. Because of the limited time frame of two years for completing a cycle of the NHANES, the number of PSUs was limited to 65. Approximately 30 000 persons were selected to be examined.

Data collection included both questionnaires and examinations. All sample persons received general medical history and dietary intake (both 24-hour recall and food frequency) questionnaires. A subsample received supplementary questionnaires on selected medical conditions, health care needs, and general well-being. The nutritional component examination included general medical and dental examinations, dermatological and ophthalmic examinations, anthropometric measurements, hand–wrist X rays, and an extensive battery of laboratory determinations. In addition, a subset of sample persons in the so-called "detailed" component received an extended medical examination, X rays of major joints, audiometry, electrocardiography, goniometry, spirometry, pulmonary diffusion, a tuberculin test, and additional laboratory determinations. Additional details of the design and content of NHANES I are available in Miller [37].

A second cycle of the NHANES was conducted between 1976 and 1980. A major purpose for NHANES II was to monitor changes in health and nutritional status since the first cycle. The assessment of nutritional status was carried out using methods that were essentially the same as those used in the first cycle, with some modification. Again high-risk segments of the population were oversampled. The most important change in nutritional assessment in NHANES II concerned anemia, which was discovered to be a significant health problem for the US population. The approach included additional questionnaire items on symptoms, signs and causes of anemia, and additional laboratory measurements.

In the realm of the detailed health examination, new emphases were placed on diabetes, kidney pathology, liver disease, osteoarthritis and disk degeneration, cardiovascular conditions, and the effects of environmental exposures on health, as measured by pulmonary function and blood levels of carbon monoxide, lead, and pesticides. As with NHANES I, details of the design and content of this cycle are available in [36].

A third cycle of NHANES was completed between 1988 and 1994, but since 1999, the NHANES has been conducted on a continuous basis. More information on NHANES III is contained in [42] and [24].

One other aspect of the NHANES deserves mention here. Several components of the **National Institutes of Health** led by the National Institute on Aging, combined resources to fund a recontact of the original NHANES I respondents in 1982, thus invoking a longitudinal component to the study. Of approximately 21 000 sample persons examined in NHANES I, some 14 000 were either located and reinterviewed or their vital status was determined. This longitudinal follow-up allowed one of the first nationally representative epidemiologic studies of its type to be conducted. A wide variety of data analyses have been completed and published in both the epidemiologic journal literature and in government publications. For more information concerning the design and objectives of this follow-up, see [6].

The National Hospital Discharge Survey. The NCHS has conducted the The National Hospital Discharge Survey(NHDS) continuously since 1965. The original sample was selected from a sampling frame consisting of a listing of health facilities known as the National Master Facility Inventory. The basic design of the NHDS, with minor periodic updates, was followed until a major redesign in 1988. Hospitals were stratified by bed size and were sampled with probabilities ranging from certainty in the largest hospitals to 1 in 40 in the smallest. Eligible hospitals included those with an average length of stay of less than 30 days and excluded Federal, military, and Veterans Administration hospitals. Within each hospital, discharges were selected using a **systematic random sampling** plan. Information was abstracted manually at each sample facility, until 1985, at which time the NCHS began using dual methods for collecting the in-hospital data. This involved purchasing data tapes from commercial abstracting services, sampling from those data tapes for hospitals where such services were used, and continuing the manual abstracting of the data in those hospitals that did not use the abstracting services.

The NHDS was redesigned in 1988 to conform to the integrated survey design paradigm described earlier for the National Health Interview Survey. The new sampling frame consisted of hospitals listed in the SMG Hospital Market Data Tape [9, 19]. As in the past, large hospitals (i.e. those with 1000 or more beds or 40 000 or more discharges per year) were sampled with certainty. The remaining strata were sampled using a three-stage design, which began with a sample of PSUs as in the NHIS, selected proportional to the projected 1985 population in the PSU, and a subsample of hospitals within the PSUs. Hospitals in the PSUs were then stratified by geographic region and ordered by PSU, abstracting service status, and hospital specialty-size group. A systematic sample was selected with a probability proportional to SMG annual numbers of discharges for the most recently available year. Finally, a systematic random sample of discharges was selected according to the hospital's stratum, and to whether the manual or automated abstracting system was used in the hospital. This procedure resulted in a 2001 sample of 504 hospitals, of which 448 were eligible and responded to the survey. The number of patient records in the 2001 sample was approximately 330 000 discharge medical record abstracts. Currently, the NHDS has been merged with other record-based surveys and expanded into one integrated survey of health care providers, including ambulatory surgical centers, hospital outpatient departments, emergency rooms, hospices, and home health agencies.

Data collected in the survey include personal characteristics of the patients, including date of birth, sex, race, ethnicity, marital status, and expected sources of payment; administrative data, including admission and discharge dates, and discharge status; and medical information, including diagnoses, surgical and nonsurgical operations and procedures, and dates of surgery. Medical information is coded using the **International Classification of Diseases** 9th Revision, Clinical Modification (ICD-9-CM). The large number of medical records in the sample each year makes possible the study of relatively rare conditions, particularly when it is feasible to combine several years of NHDS data.

The National Ambulatory Medical Care Survey. The NCHS conducted the National Ambulatory Medical Care Survey (NAMCS) from 1973 to 1990 as a survey of nonfederal office-based physicians in private and group practices throughout the United States. The purpose of the surveys was to provide national estimates of the characteristics of patient visits to physicians' offices, where the overwhelming majority of ambulatory care is rendered. A multistage stratified probability sample of approximately 3000 physicians was selected to be interviewed each year. The design consisted of a first-stage selection of geographically defined primary sampling units, as in the NHIS and other surveys. Within each PSU, a sample of physicians was selected from frames provided by the American Medical Association and the American Osteopathic Association, stratified by specialty type. During a randomly selected week of the year, an interviewer visited the physicians assigned to that week and selected a sample of patients seen that week. Typically, about 65 000 patient records were sampled for the year. Interviewers – sometimes with the aid of the physician's staff – collected information on such topics as the patient's reason for the visit, relevant diagnoses made in the office, laboratory procedures performed, treatment(s) received, and disposition of the visit. National estimates of the characteristics of interest were computed by weighting the weekly estimates from the sample physicians and aggregating over time.

One of the limitations of the NAMCS is that it does not cover visits to hospital emergency and outpatient departments, the second largest segment of the ambulatory care system. In 1991, the NCHS began the National Hospital Ambulatory Medical Care Survey (NHAMCS) to fill a gap in the coverage of ambulatory medical care data. It is known, for example, that hospital ambulatory patients differ from office patients not only in their demographic characteristics but likely in their medical characteristics as well. The need for the new study was also related to increased efforts at medical care cost containment, the burgeoning aging population, large numbers of persons without health insurance, and emerging medical technologies. The result of a series of planning efforts by the NCHS and its contractors was a sample design that involved four stages of sampling. The first stage was a subsample of the NHIS PSUs chosen for the integrated sample design described above. Within PSUs, samples of hospitals were selected, then clinics within hospitals, and finally patient visits within clinics. The resulting sample included 474 eligible hospitals, 854 clinics from outpatient departments, 462 emergency service areas, 35 114 outpatient visits, and 36 271 emergency service visits. Data were collected by hospital staff, who had been trained by survey field staff. They recorded the information on one of two patient record forms designed to account for the differences in emergency and outpatient care. The items on the forms included the demographic characteristics of the patient and medical items relating to the patient's reason for the visit and physician diagnoses. The outpatient form resembles that used in the original NAMCS, whereas the emergency service form was designed to reflect the types of services provided in that setting. Finally, medical coding follows the ICD-9-CM classification, and reason for visit is coded according to the NAMCS reason for visit classification. For further information on the NHAMCS, see [34].

The National Nursing Home Survey. What is now known as the National Nursing Home Survey (NNHS) began in 1963 with the first Institutional Population Survey conducted by the NCHS. This was originally intended to complement the NHIS, which covered only the noninstitutional population. The sampling frame for the study was the 1962 Master Facility Inventory (MFI) maintained by the NCHS and described earlier in this article. The sample contained institutions of four types: nursing care homes, personal care homes with nursing, personal care homes without nursing, and domiciliary care homes. The sample design was a multistage stratified design on which strata were defined by type of service

and bed size. The sample was selected systematically within each of the basic strata. The second stage of the sample was a systematic selection of residents or patients living in the sample establishments. A number of published reports have provided information on the characteristics of the homes and of the residents. For a more complete reading of the data, see, for example, [4].

A second cycle of the National Nursing Home Survey was conducted from 1973 to 1974 by the NCHS. The design was similar in nature to that of the first cycle, but emphasis was placed on the certification status of the homes. Certification status was determined by whether the facility was allowed to admit patients whose care was covered by the Medicare or Medicaid programs, which were not in existence at the time of the first survey. Thus, the design was changed to include certification status as a stratification variable. As one might expect, only nursing care homes were certified by Medicare for reimbursement.

The third cycle of the NNHS was conducted in 1977. Again, the design was constructed to reflect Medicare and Medicaid certification status. Many of the reports published using the data from this cycle involved trends in characteristics of the homes as well as the patients, comparing results from 1977 and the 1973 to 1974 cycle. A fourth cycle was completed in 1985 and provided additional trend data on the use of long-term care.

New cycles of the NNHS were conducted in 1995 [54] and 1997 [17]. Data from these suggest that a movement away from institutional care has begun, with more older and disabled persons utilizing newer forms of long-term care, including home-based care, visiting nurses, and the like. However, as the aging population continues to grow, it is expected that additional demands on the long-term care delivery system will grow as well. For further information on these survey results, see [17] and [54].

Surveys of Other Health Agencies

The National Medical Expenditure Surveys. To meet the growing demand for data on current health policy issues, the US government has sponsored three national household surveys of the utilization of health care services received and the expenditures related to use of those services (*see* **Health Care Utilization Data**). First, in 1977 the National Center for Health Services Research (later named the Agency for Health Care Policy and Research and subsequently the Agency for Health Research and Quality) and the NCHS conducted a National Medical Care Expenditure Survey (NMCES). A second survey, cosponsored by the NCHS and the Health Care Financing Administration and named the National Medical Care Utilization and Expenditure Survey (NMCUES), was completed in 1980. Both surveys were based on multistage stratified probability designs, and both were **panel surveys** in the sense that the data were collected by a series of periodic interviews with the initial sample of households during the year of interest. The principal data items of interest included each dental, doctor, clinic, or emergency room visit, and each hospital stay. These data include dates and services received; charges for the services received; prescribed medicines purchased and their costs; other medical expenses, and finally sources of payment, including out-of-pocket and insurance amounts, both public and private. For more detailed information on the methodological issues involved in conducting these surveys, see [27].

A third survey, the National Medical Expenditure Survey (NMES), was conducted by the Agency for Health Care Policy and Research in 1987. Many of its characteristics were similar to the NMCES and NMCUES (see above), but a second component was added to include information on the population residing in or admitted to nursing homes and facilities for the mentally retarded. Furthermore, oversampling was used to insure greater representation of population groups of special policy interest including poor and low income families, the elderly, the functionally impaired, and black and Hispanic minorities. A detailed description of the design of this survey is given in [7].

Finally, this series of expenditure surveys has once again expanded to a survey now known as the Medical Expenditure Panel Survey (MEPS). In addition to many of the features of the previous surveys, the MEPS has now incorporated measures of the quality of care received by the participants, as indicated by the change of name of the agency. For further information on specific statistical features of the MEPS, see [52].

The National Long Term Care Survey. The 1982, 1984, 1989, and 1994 National Long Term Care Surveys (NLTCS) were designed to measure the point

prevalence of chronic (90 days or more) disability in the US elderly Medicare enrolled population, as well as changes in chronic disability and institutionalization over time. The 1982 design was a list sample randomly drawn from Medicare administrative files. Screening interviews identified 6393 individuals, each with at least one chronic impairment, in seven Instrumental Activities of Daily Living (IADL) or nine Activities of Daily Living (ADL). Interviews were completed with 95% of the sample individuals.

The 1984, 1989, and 1994 surveys included both **cross-sectional** and longitudinal components because new samples of persons who had reached age 65 and survived since the last interview were drawn from the Medicare files and screened. The three later surveys also included an institutionalization component for those persons who were admitted to nursing homes during the course of follow-up. A striking and somewhat unexpected finding from these surveys was that a slight decline in age-standardized disability and mortality (*see* **Standardization Methods**) was observed between the 1982 and 1989 surveys [33].

The National Longitudinal Mortality Study (NLMS). The NLMS is a long-term prospective study of mortality in the United States. The study is funded and directed by the National Heart, Lung, and Blood Institute, and is carried out with the help of the Bureau of the Census and the National Center for Health Statistics. The basic objective of the study is to investigate socioeconomic, demographic, and occupational differentials in mortality within the United States.

The main study population consists of 13 **cohorts** of data of over two million records, drawn from the Census Bureau's Current Population Survey and from the 1980 census. The data records are periodically matched to the National Death Index (NDI), a centralized, computerized index of death records in the United States. The NDI is maintained by the National Center for Health Statistics and was begun in 1979. A public-use file of the NLMS data is available. For more information, see [46, 47].

Sample Surveys in Other Countries

Sample surveys in health research are not limited to the US, although many of the methodologies presented here stemmed from work done in the United States. Probably, the best known international effort in sample survey work is the World Fertility Survey, conducted in several countries, including developing countries beginning in the mid-1970s and continuing into the 1980s. The studies have been described in several publications, including a large number dealing with the results of individual countries themselves. For an overall description, see [59]. A description of the use of hand-held computers in conducting surveys in developing countries is provided by Forster & Snow [16]. Other aspects of the World Fertility Survey, including implications for future such studies, are discussed by Cornelius [8].

Health surveys are also conducted in developed countries. Two examples from Canada include the heart health surveys [40] and the Canadian Health Survey [25, 26]. Other studies relate to drinking behavior [10]. Still other countries, such as the United Kingdom, Sweden, the Netherlands, Israel, and others, maintain national central bureaus of statistics, many of which are responsible for designing, conducting, and analyzing social surveys, which often include questionnaire items pertaining to health and well-being.

Data Collection and Management

In many studies, biostatisticians play a very active role in the collection and management of data. At a minimum, the statistical group needs to be represented as the data collection and **data management** systems are designed and implemented, so that the statisticians know how the data reached them for analysis. Decisions made at the data collection stage can have substantial impact on the analytic process.

Data Collection

Two key decisions about data collection will affect the statistician directly: how will the participant communicate responses to the researcher, and how will the researcher record and transmit the responses? Sample surveys can collect data by mail, by telephone, or in person; each approach has both advantages and drawbacks. Furthermore, data can be recorded on paper forms and keyed in later, or directly entered into a computer at the time of the interview. Other means of recording data that has been in more common use in recent years for some surveys is to have the respondents record answers

to questions directly using a touchtone telephone or the Internet.

Contact with Participant. Mail questionnaires are used for some sample surveys because they offer a simple and economical way to request data. However, they have serious drawbacks. The most crucial one is the possibility of sampling **bias**. Bias can occur because the mailing list used as a frame may not adequately reflect the target population, or because of poor response rates in some or all of the communities, or because of difficulty in interpreting and filling out the forms. A second serious problem is that some health issues cannot be addressed without in-person assessment of the participant. Blood pressure, for example, needs to be measured in person, and self-reported diagnosis of hypertension is a far less reliable alternative. Mail questionnaires may be useful for some highly motivated and sophisticated populations; mail surveys of physicians and nurses have been used successfully to study many chronic diseases of major public health importance (stroke, breast cancer, myocardial infarction, and so on). In addition, mail surveys may be useful for interim tracking of participants in longitudinal studies.

Telephone interviews avoid some of the problems of mail questionnaires, in that direct conversation may help clarify concerns or confusion of the participants (*see* **Telephone Sampling**). Bias remains a problem, both because some people in the population do not have telephones and because response rates may be low and may be different for important subgroups. In addition, data requiring direct measurement of the participant cannot be collected over the telephone. In one particular survey conducted by the National Institute on Aging, the Survey of the Last Days of Life, the use of the telephone was instrumental in securing an acceptable overall response rate for the study. Many of the participants – recently bereaved individuals following the death of a spouse or other family member – were reluctant to be interviewed in person. In spite of a sometimes lengthy interview, lasting as long as 45 minutes to an hour, the response rate and the overall quality of the data remained good [3].

In-person interviews allow the greatest variety of data to be collected on a participant. Direct measurements, performance tests, and blood samples for laboratory work can all be carried out even in participants' homes. Developing a suitable frame may be challenging, and participation rates may be low. In addition, in-person interviews are the most costly to conduct. Some studies use a combination of telephone interviews with in-person interviews of a subsample (*see* **Interviewing Techniques**).

Data Recording and Transfer. Recording the data and transferring to the computer are key steps, on which much of the data quality will depend. Perhaps, the simplest approach is to record the answers on a preprinted form and have the forms keyed in to the computer at a later date. In this procedure, there is no way to check data at the time of collection and prompt for correction of implausible answers. Some data checking can (and probably should) be programmed into the data-entry keying program. Turnaround time depends on the data-entry service. If the number of questions is small and the possible responses are simple, a scanner form can be filled out by the participant or the interviewer, but this is practical only for the briefest of questionnaires. Careful design of a paper form is essential to make it clear and easy to fill out, and to key for data entry. It is useful to have a standardized header for paper forms, identifying the study, the batch, and sequence to record when a form was sent to data entry for keying, the staff member filling out the form and the date on which it was filled out. This information can be vital for data management (*see* **Questionnaire Design**).

Computer-assisted interviews for in-person and telephone interviews (CAPI and CATI) are becoming more widely used, especially for large-scale sample surveys, where they offer preprogrammed checks for accuracy and rapid turnaround of data. The greatest drawbacks are the cost of the equipment and its support and maintenance, the need to train interviewers in use of the computer and the program, and the initial investment in time and effort for programming. It is important to recognize that a CAPI or CATI instrument is a program, and as such needs close attention to the overall architecture as well as to the details of branches, range checks, and logic checks. Developing a computer-assisted data collection instrument requires close collaboration between the programmer, the subject-matter specialist, and someone who is familiar both with computing and with forms design.

Data Management

Data management is the process that takes the data for the study from the point of its entry into the

computer system to the time of analysis. The goal in data management is to build a system that handles quality control, tracking of study progress, linking of study components, and access for statistical analysis. In addition, the data management system needs to protect the data against loss, corruption, and unauthorized access. Each participant in the study should be assigned a study Identification Number (ID) at the time of sample selection, and the data management process should track and refer to participants by ID rather than by any personal identifier such as name, social security number, or address.

Quality Control. The goal of quality control is to ensure that the data to be analyzed are as faithful as possible a representation of the participant's true responses, and that any errors in responding, transcribing, uploading, or processing the data are identified and corrected. One common and potentially disastrous error is having a wrong ID on a form. IDs can be generated to include one or more "check digits", so that an invalid ID can be caught at the time of data keying or, if computer-assisted data collection is used, at the time of the interview. Other checks that can be carried out at the time of the first computer entry of the form are range checks (Is the value of the variable within the permitted limits for that question?), logic checks (Is the value for this variable logically consistent with the value entered for a previous related variable?), and branch checks (Has an answer been given for a question that should not have been asked or, conversely, has a question been skipped that should not have been skipped?). CATI and CAPI systems can be programmed to prompt the interviewer to correct the error at the time of data entry or, in some cases, to override the prompt if the response was unlikely but nonetheless correct. However, these checks can only be carried out within a single form collected at the same time – not across multiple forms. Thus, additional checks are probably needed at the time the data are uploaded into the main computer in which they are to be stored for analysis.

The quality control process also needs to include a standardized procedure for error correction. This should include both global corrections, where all records with a given value are changed to a new value, and person-specific corrections. It is useful to keep a system log of corrections, including the staff ID of the person who made the correction.

Database Management. Until fairly recently, data were usually stored on the computer in ASCII files or flat files, which were simply records in which each variable was identified by the columns that it occupied (a legacy from an era when data were stored on punch cards.) More sophisticated options now range from spreadsheets, to add-ons for statistical packages, to relational databases (*see* **Database Systems**). The greatest advantage of a relational database is the ability to link data across studies and across forms within a study. The database chosen should be able to meet both operational and analytic needs, as well as being large and flexible enough to handle all the data collected in a given study. Operational needs include tracking completion of data for participants, tracking performance of interviewers, generating routine reports, and providing authorized people with interim access to the data. For statistical analysis, a friendly link to the statistical package is helpful (*see* **Software for Sample Survey Data**). Some relational databases have a feature permitting some statistical packages to access the study directly, including variable labels. Again, it is crucial that the participant be identified by a study ID across all forms in the study.

Use of the database requires that the statistician know all the forms used in the study, the variable names for each form and the question to which they correspond, and the meaning of all possible values of the response, including missing value codes. One useful format for this information is a codebook – having on-line codebooks can be extremely helpful. The importance of good documentation of the database and the management process cannot be overemphasized.

A final word on maintaining and managing data files for a survey involves keeping backup files for each type of record created for the study. An example of a disaster that occurred at one statistical agency serves as a reminder of the importance of keeping the data properly backed up. Some years ago, seven cartons containing data tapes from a multi-million-dollar survey were being moved from one location to another for "safe keeping". In the course of this movement, the cartons were inadvertently left on the building's loading dock and were taken by trash collectors to the city's sanitary landfill. In spite of a valiant effort on the part of the agency to recover, clean, and reprocess the tapes, more than half of the total data set could not be recovered. Had the data been properly backed up prior to the movement of

the tapes, such a disaster could have been avoided. Attention to such details is of utmost importance.

Analysis

Two considerations should determine the analytic strategy: What is the scientific question to be addressed? How was the sample obtained? Surveys conducted primarily for policy purposes often have as their primary goal the description of the health status of the country or state and important component groups. Examples would include the types of statistics described in the section on the NHIS: **frequency distributions** and cross tabulations of disability days, prevalence and incidence of acute and chronic conditions, physician visits, hospital utilization, and so on. Other examples might include national norms for certain measured quantities such as cholesterol level, blood pressure, height, and weight (*see* **Normal Values of Biological Characteristics**). Epidemiologic surveys, however, often are designed with the goal of analyzing the relationship between characteristics of the population and the risk of prevalent or incident disease or disease prognosis. The analytic strategies for accounting for the sample design in epidemiologic studies may differ from those in studies where the main goal is to characterize a specific population (*see* **Epidemiology, Overview**).

Descriptive Analyses

These analyses usually consist of the presentation of population estimates of the characteristics under study and some indication of the sampling variability of the estimates. Most standard texts on survey sampling (e.g. Cochran [5]) provide the necessary information to construct the desired estimates. Typical analytic reports from descriptive surveys include a variety of standard tables containing estimates of means, totals and percentages, or proportions. If the survey designs are relatively simple, estimates of sampling variance can be computed using algebraic formulas. For more complicated designs, however, approximation techniques such as Taylor series representations (see, for example, [50]) (*see* **Linearization Methods of Variance Estimation**) or pseudoreplication methods [12, 29, 39] (*see* **Resampling Procedures for Sample Surveys**) are usually used to estimate sampling variability. In the estimation of variances, one needs to be aware of the possible necessity of applying a **finite population correction** factor, if the sampling rate for the survey is, say, more than 5 to 10% in a given stratum, even though the overall sampling rate is much lower than that.

Two other areas of descriptive analysis deserve mention. First, a considerable amount of work has been done on the topic of small domain, or **small area estimation** Here, one is interested in providing estimates of health characteristics for either small strata (such as a subgroup of the population at high risk for disease) or for small geographic areas, which are subgroups of the larger area covered by a given survey. When the sample sizes in the small area are too small to allow the computation of reliable direct estimates from the survey, some researchers have proposed the use of so-called synthetic estimators, based on **regression** relationships between the characteristic of interest and ancillary variables available from the survey. Others have proposed the use of **composite** or **"shrinkage" estimators** that combine direct and synthetic components. For additional information on this topic, see [15, 44].

The final topic on descriptive analysis concerns the description of change from one time period to another. Estimates of change are usually desired to study the effects of forces that are known to have acted on the population under study. For example, if a hypertension intervention is initiated in a community, we would like to know whether the intervention has influenced the prevalence of hypertension in the community. In such an instance, it is necessary to estimate the prevalence both before and after the intervention, and it is best to retain the same sample for both occasions. However, if the goal is to estimate the aggregate average blood pressure level at the two occasions, it is best to select a new sample each time. Each of these alternatives has advantages and drawbacks. The first alternative requires careful maintenance of the sample over the time period for the study and the statistician must deal with dropouts and other losses to follow-up. The second alternative necessitates the drawing of a second sample and all the work required to recruit and to inform new sample members. More information on this topic is available in [5].

Epidemiologic Studies

Studies of potential risk factors for the onset and progression of health problems, in contrast to

descriptive studies, are likely to rely on regression models to measure the **association** between **risk factors** and disease and to adjust for other factors thought to affect the risk. There is general agreement that descriptive summaries based on survey data must adjust for the sample design, but there has been less of a consensus on how to estimate regression parameters and calculate standard errors for regression analysis of survey data. One possible approach for estimation is to analyze the data by a standard approach such as **maximum likelihood** treating them as if they arose from a **simple random sample** (i.e. ignoring the sampling design). A second approach is to assume that the design influences the results primarily through key variables related to the sampling design, such as age, sex, or race, and to adjust for those by including coefficients for those variables as predictors in the model. In a stratified design, the variables chosen are typically the stratum definitions, at minimum. A third approach, more complicated to implement and thus used less frequently, is to modify the basic approach to reflect all features of the design, for example, maximizing the likelihood over the complete probability distribution associated with the sampling design (design-based analysis) (*see* **Superpopulation Models in Survey Sampling**). Finally, a more widely-used approach is to modify the estimation approach to reflect the sampling weights, in particular, for example, by weighting the score function components to obtain so-called **pseudo-maximum-likelihood** estimates (model-based analysis). The estimation of standard errors associated with the point estimates is challenging and usually relies on some asymptotic assumptions. In the past, the ability of some researchers to adjust for complex sampling was limited by the lack of commercial software. Software is now available, however, to adjust for complex samples for many different kinds of regression models (see, for example, [50, 56] (*see* **Software for Sample Survey Data**; **Software for Sample Survey Data, Misuse of Standard Packages**).

Some researchers have argued that adjustment is important when describing the parent population from which the survey sample was drawn, but not for estimating regression parameters for comparing risk groups [53]. For example, for standard **linear regression** the usual estimator

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{9}$$

is shown in standard texts to have commendable properties. If the sampling design is ignored, the estimator is the best linear unbiased estimator (*see* **Least Squares**). In the survey data setting, however, as Sarndal et al. [48] point out, there are distinct theoretical drawbacks to the usual estimator. First, its optimal properties only hold if the model is correct. Secondly, obtaining standard error estimates that reflect the true variability from the sampling design is difficult. These authors recommend using a sample-weighted estimator. Other authors [23] have also stated that "the design is relevant, including especially the effects of intraclass correlations from cluster sampling, and perhaps also variable sampling fractions and other aspects of design. Failure to recognize such effects may lead to serious understatement of confidence intervals and overstatements of precision in inferences to the causal system".

The sample-weighted estimator for **likelihood**-based estimators from regression models makes use of the sampling weights, reflecting how much larger a segment of the population an individual would represent than in the sample. In a simple random sample, the log likelihood would just be the sum of the **score** contributions from each individual in the sample. The sample-weighted estimators assume that each individual should contribute to the total log likelihood by an amount reflecting the composition of the whole population; thus, the estimated log likelihood is a weighted sum of the individuals' contributions. This weighted sum is not typically the exact log likelihood for the full sample design. In fact, the design likelihood can rarely be calculated explicitly. However, the weighted sum can be thought of as an unbiased estimate of the log likelihood for the population from which the sample was drawn, and thus its root is called the **pseudo-maximum-likelihood** estimator [51]. For linear regression, the sample-weighted estimator or pseudo MLE is given by

$$\hat{\beta}_w = (X'WX)^{-1}X'WY. \tag{10}$$

Pseudo-maximum-likelihood estimators have been worked out for a number of standard procedures including **logistic regression** The correct point estimators can be obtained by using weights in standard software packages (*see* **Software, Biostatistical**), but the standard error estimates obtained by simply adding weights to a procedure for simple random samples will not be correct. More complicated

standard error estimates, described below, must be used for pseudo MLEs.

The sample-weighted estimates typically give different parameter estimates for stratified designs with unequal sampling weights than do the unweighted estimates, but are not affected by clustering. The conventional standard errors based on the usual estimates, however, may systematically underestimate the true sampling variability in the presence of clustering if there is within-cluster correlation. In this case, the usual regression estimators assume independent and identically distributed errors, and overestimate the effective sample size. A more conservative approach is to take account of the clustering by using a Taylor series approximation to the design-adjusted variance, as described in [51] or [48]. Such estimates can also take account of the sample weights used in the sample-weighted estimators. Commercial software is now available that calculates sample-weighted regression estimators, and calculates the Taylor series approximation for the standard error for linear regression, logistic regression, and so on [50].

Another approach to standard error estimation for sample surveys is **resampling** or replication. Balanced repeated replication and **jackknife methods** have been used for some time in survey sampling [29, 35]; more recently, these methods have been extended and, additionally, **bootstrap methods** have been applied [12, 28, 31, 45, 58]. These methods differ from the Taylor series approach in two key ways: First, unlike the Taylor series approximation, it is not necessary to write down an explicit differentiable expression for the variance. Second, replication methods may perform better in small samples than the Taylor series approximation [48]. Owing to recent improvements in computing, software is now available to implement replication-based methods (e.g. WesVar [38, 39, 56]; VPLX [14]). A widely used statistical software package, SAS (SAS Institute, Cary, NC), has added new procedures for analyzing survey data in its most recent release.

The effect of clustering on the variance can be substantial if the number of primary sampling units is not large relative to the number of strata, but clustering does not affect the parameter estimates. Sample weighting can affect both the parameter estimates and the variance. If sampling weights are very unequal, the standard error of the sample-weighted estimates is typically substantially larger than that of the unweighted estimates, reflecting the uncertainty in weighting a small number of observations very heavily in the analysis. Korn & Graubard [30] have examined these effects for a study based on the NHANES I survey, and found that different analyses led to very different conclusions. They suggest that the clustering should generally not be ignored. However, if extremely unequal sampling fractions were used, one way to obtain reasonable point estimates without reducing the **power** of the study is to include those factors related both to the design and to the regression variables as **covariates** in the analysis. Korn and Graubard note in conclusion, however, that a better solution might be to plan studies in advance to have adequate sample sizes in all strata (*see* **Sample Size Adequacy in Surveys**). This would permit regression models to use the design in the analysis – the more conservative approach, and one that addresses directly the difficulties of making inferences about risk factors in a population using data from a complex survey design.

References

[1] Anderson, M. (1988). *The American Census: A Social History*. Yale University Press, New Haven.

[2] Botman, S.L., Moore, T.F., Moriarity, C.L. & Parsons, V.L. (2000). Design and estimation for the national health interview survey, 1995–2004, *Vital and Health Statistics*, Series 2, No. 130. National Center for Health Statistics, Hyattsville, p. 5.

[3] Brock, D.B., Holmes, M.B., Foley, D.J. & Holmes, D. (1992). Methodological issues in a survey of the last days of life, in *The Epidemiologic Study of the Elderly*, R.B. Wallace & R.F. Woolson, eds. Oxford University Press, New York, pp. 315–332.

[4] Bryant, E.E. (1965). Institutions for the aged and chronically III, *Vital and Health Statistics*, Series 12, No. 1. U.S. Government Printing Office, Washington, pp. 1–46.

[5] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[6] Cohen, B.B., Barbano, H.E., Cox, C.S., Feldman, J.J., Finucane, F.F., Kleinman, J.C. & Madans, J.H. (1987). Plan and operation of the NHANES I epidemiologic followup study: 1982–1984, *Vital and Health Statistics*, Series 1, No. 22. National Center for Health Statistics, Hyattsville, pp. 5–7.

[7] Cohen, S., DiGaetano, R. & Waksberg, J. (1991). *Sample Design of the 1987 Household Survey*, Publication No. 91–0037. Agency for Health Care Policy and Research, Rockville.

[8] Cornelius, R.M. (1985). The world fertility survey and its implications for future surveys, *Journal of Official Statistics (Sweden)* **1**, 427–433.

[9] Dennison, C.F. & Pokras, R. (2000). Design and operation of the national hospital discharge survey: 1988 redesign, *Vital and Health Statistics*, Series 1, No. 39. National Center for Health Statistics, Hyattsville, p. 2.

[10] Duffy, J.C. (1985). Questionnaire measurement of drinking behavior in sample surveys, *Journal of Official Statistics (Sweden)* **1**, 229–234.

[11] Eckler, A.R. (1972). *The Bureau of the Census*. Praeger, New York.

[12] Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM, Philadelphia.

[13] Evans, D.A., Funkenstein, H.H., Albert, M.S., Scherr, P.A., Cook, N.R., Chown, M.J., Hebert, L.E., Hennekens, C.H. & Taylor, J.O. (1989). Prevalence of Alzheimer's disease in a community population of older persons: higher than previously reported, *Journal of the American Medical Association* **262**, 2251–2256.

[14] Fay, R.E. (1990). VPLX: variance estimates from complex samples, *American Statistical Association 1990 Proceedings of the Survey Research Methods Section*. American Statistical Association, Alexandria, pp. 266–271.

[15] Fay, R.E. & Herriott, R.A. (1979). Estimates of income for small places: an application of James–Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269–277.

[16] Forster, D. & Snow, R.W. (1995). An assessment of the use of hand-held computers during demographic surveys in developing countries, *Survey Methodology* **21**, 179–184.

[17] Gabrel, C.S. (2000). An overview of nursing home facilities: data from the 1997 national nursing home survey, *Advance Data from Vital and Health Statistics*, No. 311. National Center for Health Statistics, Hyattsville, pp. 1–12.

[18] Halacy, D. (1980). *Census: 190 Years of Counting America*. Elsevier/Nelson Books, New York.

[19] Hall, M.J. & DeFrances, C.J. (2003). 2001 national hospital discharge survey, *Advance Data from Vital and Health Statistics*, No. 332. National Center for Health Statistics, Hyattsville, p. 4.

[20] Hansen, M.H., Hurwitz, W.N. & Bershad, M. (1961). Measurement errors in censuses and surveys, *Bulletin of the International Statistical Institute* **38**, 359–374.

[21] Hansen, M.H., Hurwitz, W.N., Marks, E.S. & Mauldin, W.P. (1951). Response errors in surveys, *Journal of the American Statistical Association* **46**, 147–190.

[22] Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vols. I & II. Wiley, New York.

[23] Hansen, M.H., Madow, W.G. & Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association* **78**, 776–808.

[24] Harris, T., Woteki, C., Briefel, R.R. & Kleinman, J.C. (1989). NHANES III for older persons: nutrition content and methodological considerations, *American Journal of Clinical Nutrition* **50**, 1145–1149.

[25] Hidiroglou, M.A. & Rao, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys: part I – simple goodness-of-fit, homogeneity and independence in a two-way table with applications to the Canada health survey (1978–1979), *Journal of Official Statistics (Sweden)* **3**, 117–132.

[26] Hidiroglou, M.A. & Rao, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys: part II – independence in a three-way table with applications to the Canada health survey (1978–1979), *Journal of Official Statistics (Sweden)* **3**, 133–140.

[27] Horvitz, D.G. & Folsom, R.E. (1980). Methodological issues in medical care expenditure surveys, *American Statistical Association 1980 Proceedings of the Survey Research Methods Section*. American Statistical Association, Alexandria, pp. 21–27.

[28] Judkins, D. (1990). Fay's method for variance estimation, *Journal of Official Statistics (Sweden)* **6**, 223–240.

[29] Kish, L. & Frankel, M. (1974). Inference from complex samples (with discussion), *Journal of the Royal Statistical Society, Series B* **36**, 1–37.

[30] Korn, E.L. & Graubard, B.I. (1991). Epidemiologic studies utilizing surveys: accounting for the sampling design, *American Journal of Public Health* **81**, 1166–1173.

[31] Kovar, J.G., Rao, J.N.K. & Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates, *Canadian Journal of Statistics* **16**(Suppl), 25–45.

[32] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[33] Manton, K.G., Corder, L.S. & Stallard, E. (1993). Estimates of change in chronic disability and institutional incidence and prevalence rates in the U.S. elderly population from the 1982, 1984 and 1989 national long term care survey, *Journal of Gerontology* **48**, S153–S164.

[34] McCaig, L.F. & McLemore, T. (1994). Plan and operation of the national hospital ambulatory medical care survey, *Vital and Health Statistics*, Series 1, No. 34. National Center for Health Statistics, Hyattsville, pp. 1–78.

[35] McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys, *Vital and Health Statistics*, Series 2, No. 14. U.S. Government Printing Office, Washington, pp. 10–24.

[36] McDowell, A., Engel, A., Massey, J.T. & Maurer, K. (1981). Plan and operation of the second national health and nutrition examination survey: 1976–1980, *Vital and Health Statistics*, Series 1, No. 15. National Center for Health Statistics, Hyattsville, pp. 4–25.

[37] Miller, H.W. (1973). Plan and operation of the health and nutrition examination survey, *Vital and Health Statistics*, Series 1, No. 10a. National Center for Health Statistics, Hyattsville, pp. 4–38.

[38] Morganstein, D.R. & Brick, J.M. (1996). WesVarPC: software for computing variance estimates from complex designs, *Proceedings of the Bureau of the Census 1996 Annual Research Conference*. Bureau of the Census, Washington, pp. 861–866.

[39] Morganstein, D.R., Brick, J.M., Broene, P. & Nixon, M.G. (1998). *The Replication Method for Estimating Sampling Errors: Creating Replicates*. Eustat – The Basque Statistics Institute, Vitoria-Gasteiz, Spain.

[40] Nargundkar, M.S., Balram, C., Hogan, K., Joffres, M., MacLean, D., MacLeod, E.B., O'Connor, B. Petrasovits, A., Reeder, B. & Stechenko, S. (1990). Heart health surveys in Canada, *American Statistical Association 1990 Proceedings of the Social Statistics Section*. American Statistical Association, Alexandria, pp. 61–65.

[41] National Center for Health Statistics. (1964). Health survey procedure, *Vital and Health Statistics*, Series 1, No. 2. U.S. Government Printing Office, Washington, p. 2.

[42] National Center for Health Statistics. (1994). Plan and operation of the third national health and nutrition examination survey, 1988–1994, *Vital and Health Statistics*, Series 1, No. 32. Hyattsville, pp. 20–35.

[43] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society* **97**, 558–606.

[44] Purcell, N.J. & Kish, L. (1979). Estimation for small domains, *Biometrics* **35**, 365–384.

[45] Rao, J.N.K., Wu, C.F.J. & Yue, K. (1992). Some recent work on resampling methods for complex surveys, *Survey Methodology* **18**, 209–217.

[46] Rogot, E., Sorlie, P.D., Johnson, N.J., Glover, C.S. & Treasure, D.W. (1988). *A Mortality Study of One Million Persons by Demographic, Social, and Economic Factors: 1979–1981 Follow-up*, NIH Publication No. 88-2896. National Institutes of Health, National Heart, Lung & Blood Institute, Bethesda.

[47] Rogot, E., Sorlie, P.D., Johnson, N.J. & Schmitt, C. (1992). *A Mortality Study of One Million Persons by Demographic, Social, and Economic Factors: 1979–1985 Follow-up*, NIH Publication No. 92-3297. National Institutes of Health, National Heart, Lung & Blood Institute, Bethesda.

[48] Sarndal, C.E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[49] Scott, A.H. (1968). *Census, U.S.A.: Fact Finding for the American People, 1790–1970*. Seabury Press, New York.

[50] Shah, B.V., Folsom, R.A. & LaVange, L. (1991). *SUDAAN User's Manual*. Research Triangle Institute, Research Triangle Park.

[51] Skinner, C.J., Holt, D. & Smith, T.M.F. eds. (1989). *Analysis of Complex Surveys*. Wiley, New York.

[52] Sommers, J.P. (2000). Methods to produce establishment and firm level estimates for an economic survey, *Proceedings of the International Conference on Establishment Surveys II*. American Statistical Association, Alexandria.

[53] Stevens, R.G., Jones, D.Y., Micozzi, M.S. & Taylor, P.R. (1989). Body iron stores and the risk of cancer (comment), *New England Journal of Medicine* **320**, 1012–1014.

[54] Strahan, G.W. (1997). An overview of nursing homes and their current residents: data from the 1995 national nursing home survey, *Advance Data*, No. 280. Centers for Disease Control and Prevention/National Center for Health Statistics, Hyattsville, pp. 1–12.

[55] Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, *Metron* **2**, 461–493, 646–683.

[56] Westat, Inc. (2000). *WesVar 4.0 User's Guide*. Westat, Rockville.

[57] White, L.R., Petrovich, H., Ross, G.W., Masaki, K.H., Abbott, R.D., Teng, E.L., Rodriguez, B.L., Blanchette, P.L., Havlik, R.J., Wergowske, G., Chiu, D., Foley, D.J., Murdaugh, C. & Curb, J.D. (1996). Prevalence of dementia in older Japanese-American men in Hawaii: the Honolulu–Asia aging study, *Journal of the American Medical Association* **276**, 955–960.

[58] Wolter, K. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

[59] World Fertility Survey. (1986). *Final Report*. International Statistical Institute, Voorburgh, Netherlands.

Further Reading

Wright, J.D., Wang, C.-Y., Kennedy-Stephenson, J. & Ervin, R.B. (2003). Dietary intake of ten key nutrients for public health, U.S.: 1999–2000, *Advance Data from Vital and Health Statistics*, No. 334. National Center for Health Statistics, Hyattsville, pp. 1–4.

DWIGHT B. BROCK, LAUREL A. BECKETT & JULIA L. BIENIAS

Sampling Distributions

The concept of a sampling distribution is an essential feature of statistical inference. A sampling distribution is a **probability** distribution that describes the behavior of a statistic calculated from a **random sample** of a particular size. To understand it, we need to consider briefly the issues of sampling variability and parameter estimation. We will then see that the sampling distribution provides the all-important connection between probability models and statistical inference.

Data collected for research studies in medicine, public health, and other fields are used for the making of inferences about unknown parameters. Consider two examples. Researchers in obstetrics may wish to evaluate the effect of maternal smoking on infant birthweight among babies born at full term to mothers between the ages of 18 and 40. Research oncologists might need to determine the proportion of lung cancer patients who experience tumor shrinkage when administered a chemotherapeutic drug. As we set out to study these problems, it is often assumed that the parameter of interest (average birthweight or probability of tumor response) has some "true" value in the population as a whole. It is not practicable (and is usually impossible) to attain the "true" parameter value by assessing the entire population, since its number is essentially infinite in most instances. Instead, we choose individuals at random from the population under study, and make our assessments on the members of our sample (*see* **Simple Random Sampling**). We use the data from this sample of subjects in order to guess at the true parameter value. The population or "true" value of the parameter is unknown and unknowable. What can be obtained is the sample value of the parameter, estimated from the data collected during the study. The two values (population and sample) will almost certainly not be identical, but we hope that they will be close. One of the major determinants of the accuracy of a sample statistic is whether the subjects selected for the study are representative of the subjects in the population. Other determinants include the size of the sample and the probability distribution of the original measurement in the population.

In the frequentist's approach to statistical **inference**, the population (true) parameter value is assumed to be a fixed quantity. In contrast, its sample value is a random quantity, since it is a function of the data collected. For example, suppose we record the birthweights of 100 infants born at full term. From these data we compute the mean birthweight, say $\overline{X}$. Provided that our sample is representative of the population we wish to characterize, the sample mean, $\overline{X}$, represents a reasonable guess as to the average birthweight in the population. Suppose that another research group assembles a random sample of 100 full-term newborns, and computes the mean birthweight. Surely the two mean values will differ, simply due to chance. If we conducted the same study a third time, a fourth time, and so on, then we would obtain a new estimate of the sample mean for each group of 100 randomly selected subjects. This feature is known as *sampling variability*. Because the individual birthweight measurements are random, any function of them (i.e. any sample statistic) is also random. The behavior of a sample statistic, then, can be characterized by a probability distribution, just as the behavior of the individual variable under study can be described by a probability distribution.

To continue, let us assume that birthweights in the specified population (babies born at full term to mothers aged 18–40) are well characterized by the normal distribution, with a mean of 3500 g and a standard deviation of 400 g. Remember that these population parameter values are unavailable to us in practice; we must instead rely on estimates from a random sample. How accurate an estimate of the average birthweight of full-term babies born to mothers aged 18–40 could we obtain from, say, a random sample of size 100? Because of sampling variability, it is very unlikely that the sample mean based on 100 subjects would be identical to the overall population mean of 3500 g. It would be useful to know, however, how precise we can expect our estimate to be. This is where the sampling distribution comes into play. It tells us how variable a sample mean is expected to be in successive samples of a specified size, and allows us to make statements about the likelihood of the true mean value falling within a certain range of the sample mean.

The sampling distribution of a statistic calculated from n observations is derived from mathematical principles. This distribution tells us not about the behavior of the individual observations (e.g. birthweight or occurrence of tumor shrinkage), but about the behavior of the summary statistic based on these values. If we were to take random samples of size n

repeatedly and compute a particular summary statistic (say, the **mean**, **median**, proportion of positive responses, or **standard deviation**), then the distribution of the summary statistic would be described by a sampling distribution. Of course, we do not, in practice, repeat the same study again and again. This idea is conceptually important as a means of interpreting the distribution of the resulting summary statistic from a single sample of size n. If we consider the sample mean, then there is an important statistical result which tells us that the sampling distribution of $\overline{X}$ is approximately normal, with mean equal to the true population mean, and variance depending on the sample size (n) and the variability of the measurement in the population (*see* **Central Limit Theory**). This result holds regardless of the probability distribution of the original measurements, so long as n is sufficiently large.

If the population mean of the variable of interest is μ and its standard deviation is σ, then $\overline{X}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. The standard deviation of the sample mean (referred to as the **standard error** of the mean) is clearly a smaller quantity than the standard deviation of the original measurements, and decreases as n increases. Thus, the larger the sample size, the less variable (i.e. more precise) our estimate of the sample mean. Knowledge of the sampling distribution allows us to appropriately design a study, as well as to generate confidence intervals (*see* **Estimation, Interval**) and to conduct **hypothesis testing** for the parameters under consideration once the data have been collected.

Consider again the birthweight example. We assumed earlier that the unknown mean birthweight for all full-term babies born to mothers aged 18–40 to be 3500 g, with a standard deviation of 400 g. If we plan to collect birthweight data from a random sample of size 100, then there is a 95% probability before the study is conducted that the true mean will be within about 80 g of the sample mean. If we increase the sample size to 500, then there is a 95% chance before the study is conducted that the true mean will be within 36 g of the sample mean. The result of increasing the sample size is to tighten the sampling distribution around its mean, allowing us to make more precise inferences about the population mean.

Properties of the sampling distribution also permit meaningful comparisons of parameters across different groups. For example, we might want to compare the probability of tumor shrinkage for a standard chemotherapy regimen vs. an experimental one in patients with lung cancer. Even if the true, group-specific parameter values were the same in the two populations, we would not expect to obtain identical sample statistics across sample comparison groups because of sampling variability. However, we do need some way of deciding when a difference in sample statistics between groups is large enough for us to conclude that the population parameters are probably different. The sampling distribution of the statistic of interest helps us to answer this question. The difference in parameter estimates is gauged against the expected variability of the sample statistics, giving us a formal method by which to make inferences about the parameter values in the population. Without knowledge of the sampling distribution, this would be impossible.

The sampling distribution, then, is a key element in the conduct of statistical inference. It describes, in probabilistic terms, the behavior of a statistic computed from a random sample of size n. Among the most commonly encountered sampling distributions are the normal distribution, **Student's *t* distribution**, **chi-square distribution**, and ***F* distribution**.

MELISSA D. BEGG

Sampling Frames

Probability sampling allows one to make **inferences** about large, sometimes infinite, populations without observing every member. In a probability sample every member of the population has a known, nonzero probability of selection. Knowing these probabilities, it is possible to select a subset of the population from which to make estimates (*see* **Estimation**) about the entire population with specific degrees of precision. To draw a probability sample from a population it is necessary to have a list or other selection process, called a *sampling frame*, that ensures some probability of selection for each element in the population. The sampling frame defines the portion of the population from which the sample is selected. Hence, the quality, completeness, and availability of possible sample frames are major considerations when selecting a population for study using statistical inference.

Frames are usually defined by geographic listings of blocks or other topographic units, maps, directories, membership, or other kinds of lists, or they may be defined from telephone or other electronic formats. The United Nations Statistical Office [11] defines the frame content as maps, lists, directories, and other sources that permit the construction and selection of sample units. A frame's specifications "should define the geographic scope of the survey; categories of material covered; and include the date [the frame was constructed] and the source of the frame" [11]. Wright & Tsao [14, p. 26] also recommend that the frame should include "any auxiliary information (measures of size, demographic information) that might be used for (i) special sampling techniques such as stratification or selection with probabilities proportionate to size sample selections or for (ii) special estimation techniques such as ratio or regression estimation". In other words, a frame contains listings or other relevant demarcations of the population from which sampling units can be selected and provides related documentation that helps describe the selection process. Sampling units may be individual elements or they may be clusters of elements [6].

In the US or other countries where there is extensive telephone coverage, telephone interviewing has become the data collection mode of choice, particularly in urban areas, because of cost and related problems with accessibility of respondents for face-to-face interviews (*see* **Telephone Sampling**). Although telephone directories are used sometimes for frames, usually they are incomplete and **random digit dialing** (RDD) [12] surveys are now commonly used. The frames for RDD studies are constructed by random selection of ten-digit numbers comprised of the area code, prefix, and suffix of individual telephone numbers, or lists that have been screened to eliminate banks defined by area code and prefix that include only business listing, or that contained unassigned numbers. (Note that in the US the 10 digit format is the sampling unit. In other countries the unit may be some other combination of numbers that constitutes the full telephone number.) In the US it is possible to purchase lists that have been purged of most business numbers and are adjusted for the proportions of households by county consistent with an equal probability of selection of element model (epsem) [13].

Frame Coverage

While there should be a one-to-one correspondence between the list and population, few lists meet this requirement. There are two potential sources of variation between the population of inference and the sample frame described in the literature [3, 6, 11]. One source of variation results from how the *eligible* population is defined [3]. This source of variation is often deliberately introduced for reasons of feasibility and cost. It amounts to a redefinition of the population about which inferences are to be made, by specifically excluding subsegments of the population that – for reasons of cost or efficiency – may not be included in the final sampling frame. For example, a RDD sample for a telephone survey deliberately redefines the population to include only those households with telephones.

Deliberate differences between what Groves [3] calls the *population of inference* and the actual population elements that comprise the frame might be introduced for several reasons. For example, if inferences are to be drawn about the general US population, there may be subgroups within that population that may be difficult or impossible to interview, or who differ from the general population in ways that make them nonrepresentative of the population that is relevant for the study (i.e. they may be cognitively impaired or foreign nationals), or they may be housed or located in ways that make them never available for

interview during the study period (i.e. they may be in prison, in college, or permanent residence abroad) or they may be very rare groups or may reside in inaccessible places, making the cost of interviewing them prohibitive.

Individuals in the population of inference, who might otherwise be available, might not be available during a particular study because the sample is selected and interviewed at a specific point in time in order to fix a point of estimation; for example, the Current Population Survey is conducted during the week containing the 19th of the month [3], or the US Decennial Census is carried out every tenth year ending in zero during which all residents of the US are enumerated at their residences as of 1 April. In these cases, elements of the population not available during the study period have zero probability of selection. Groves [3] refers to the population resulting after excluding these groups as the **target population**, which represents the actual population about which inferences can be drawn based on the available sampling frame.

The second source of variation results from flaws in the actual listings of the eligible population that comprise the frame. A perfect frame is one in which there is a one-to-one correspondence between the listing and each element; that is, "every element appears on the list separately, once, only once, and nothing else appears on the list" [6]. Perfect frames are rare. Most often, there are problems which must be detected and modified, or the frame will produce a **biased** sample. In fact, many times there is no existing frame and one has to be created from other, imperfect, sources. In such cases, the date the frame is created and the methods by which the frame is modified or constructed are included as elements of the documentation of the frame [11]. However, whether the variations between the two populations are deliberately introduced for reasons of efficiency or result from flaws in the correspondence between the elements and the listing, these variations need to be ascertained and decisions made about how to treat them.

Coverage error, according to Groves [3], is the difference between statistics based on the population defined by the sampling frame and statistics based on the target population. In other words, coverage error occurs most often in population estimates when there is a lack of correspondence between the frame population and the target population. One of the clearest examples would be differences in population statistics calculated based on data from a population defined by telephone numbers and those calculated based on household data, including those without telephones.

Kish [6] describes four basic frame problems that might lead to coverage error in survey data. These include: (i) missing elements (*see* **Missing Data**), noncoverage, or an incomplete frame; (ii) **clustering** in which more than one element is in a single unit; (iii) blank or foreign elements (ineligible units); and (iv) duplicate listings. Overall, Kish [6] suggests three strategies. (i) The problems can be ignored if their potential impact is small compared with other sources of error, and correcting them would be too costly given the value of the corrections in reducing bias. (ii) As noted above, the target population can be redefined to fit the frame, assuming that the redefinition still permits accurate estimates of the parameters of interest in the population. (iii) The frame can be corrected by splitting clusters or selecting individual units and weighting the clusters, deleting blanks, duplicates, or ineligible or foreign elements. In some cases, such as with duplicates or blanks, the corrections need to be made in advance. The effects of clustering or of missing or under-represented elements can sometimes be corrected through weighting the data after the interviewing is completed.

Until the advent of telephone interviewing (*see* **Telephone Sampling**), frame coverage problems were pretty much as described above. However, telephone frames have inherent coverage problems that require special consideration before they are used. One problem mentioned above is the fact that telephone frames contain many foreign elements in the form of business numbers, telephone company service numbers, nonworking numbers, and computer and FAX modems [3, 13]. As noted earlier, it is now common to purchase prescreened lists of "seed" telephone numbers from which RDD samples can be generated by deleting the last one to four digits of the number's suffix and replacing them with one to four numbers selected from a table of random numbers. This strategy is increasingly being employed in the US as an inexpensive way of creating a RDD frame [13]. According to the vendors, the resulting biases in such samples are about 3% due to missed households. However, the use of these prescreened banks is an instance in which the cost in terms of potential bias is very low compared with

the costs of creating a RDD sample by randomly selecting and screening full 10-digit numbers using the Waksberg–Mitofsky method [12].

Groves also notes that telephone frames also suffer from "over coverage" because some households may have more than one listing, such as a business telephone, an adolescent telephone, or a second household number. His estimate, however, was published almost 20 years ago and suggested that 3%–4% of US households have second telephones [4]. This estimate is probably low by current standards, particularly if households with computer modems are included in the estimates.

However, by far the greatest concern with telephone frames is the amount of under-coverage due to the number of households that do not have working telephones. Under-coverage is a source of considerable bias, because it is correlated with sociodemographic characteristics such as race and income [9]. Although the number of households without a telephone has varied, it could be as high as 15%–20% among minority and rural households, and even higher among households in which the income is very low. Clearly, there are indications that households without telephones differ substantially from those with telephones, and these differences are often correlated with other important substantive variables [3, 4, 9]. Thus, the potential bias cannot always be ignored by using a telephone sample frame and designating only households with telephones as members of the target population, although this is commonly done. An alternative solution might be to obtain data on the demographic characteristics of households without telephones and then post-weight the sample to accommodate the potential bias. In the US the Decennial **Census** is the most easily accessible source for such weights, but it too has biases.

Clustering and Stratification Effects

In most large-scale, national household surveys, the frame includes clusters and strata at various levels. The effects of these characteristics of the frame are taken into account by weighting and/or by calculating design effects on the variances of the key variables. These effects are included as part of the documentation of the frame.

When the sampling unit is a household, then usually the differences between the population of inference and the target population also concern definitions of eligibility of individual members of the household. In most cases the target population will include residents of households in the geographic area of the survey at the time of the survey. Thus, following Groves' [3] conceptualization, the *frame population* would be those in the target population who can be enumerated prior to the survey. The *survey population* would be those in the frame population; "who, if they were selected for the survey, would be respondents"; that is, they would be accessible to an interviewer. This is, as Groves notes, a hypothetical population, since if they are not asked to participate in an interview there is no way to know whether they would consent. However, those not asked might also exclude as potential members of the frame persons who reside in the frame population but who – due to incapacity, continuing absence from any enumeration unit (i.e. a student residing in a dormitory), speaking a foreign language that has not been translated for interview, or for some other reason – might not be able to consent to an interview. The survey population is always determined after the household is contacted and those in the frame population who are excluded from the survey population would usually be those designated as ineligible [3].

Ineligibles affect the probability of selection of the eligibles and, therefore, fall into the second category of coverage problems described above (clustering of units). By defining residents of households as the target population, other residents of the population of inference with no permanent household residence or those who are institutionalized would be excluded. Also excluded might be residents of households in remote areas, where the cost of obtaining an interview would be prohibitive relative to the potential bias due to failing to interview them. Including those who do not consent to be interviewed, the difference between the target population and the survey population is **nonresponse** [3].

Once the household is established as the enumeration unit, it becomes necessary to define what constitutes a household and, beyond that, who is a member of the household once it is identified. The definition of a household and its membership are key elements of the sampling frame. The elements of the target population are individuals, whereas the elements of the frame population are households. This means there is not a one-to-one correspondence between the frame population and the target population. Hence, a rule is required to

establish correspondence between the elements in the two populations. Groves describes two alternative correspondence rules [3]. The *de jure* rule attaches each person to a single housing unit based on where they live; and the *de facto* rule attaches each person to the residence where they are staying at the time of the survey. The distinction between these two rules focuses on what constitutes a dwelling unit or household and the difference between "living" vs. "staying somewhere" at the time of the interview. Failure to make these decisions results in coverage error due to poor frame specification.

The process of selecting the member of the target population has also been the subject of much discussion. There are several ways of going about this process of defining the eligible respondents in the target population. The most rigorous is prescribed by Kish [6], in which all members of the household are listed in order of age in a table and then one is randomly selected for interview based on a prespecified selection procedure. This procedure generally works well in face-to-face interviewing situations, but does not work well on the telephone, where it requires respondents to name all members of the household to an anonymous telephone caller. Other techniques which seem more acceptable for use with the telephone have been proposed by Trodahl & Carter [10], Salmon & Nichols [7], and Czaja et al. [1]. These methods require less information about the entire household and rely on selection by most recent birth date and, sometimes, gender.

In the cases in which respondents are rare elements in the target population, techniques designed to improve the probability of locating a respondent have been proposed. These are generally described as multiplicity or **network sampling** methods [2, 8]. This technique takes advantage of clustering and depends on linkages between members of a family or close friends for locating rare respondents. Initial respondents are asked to list members of the household plus other selected members of the family living at other addresses. Then each member listed is considered part of a cluster of persons and the probability of selection is determined by the probability of being listed. The use of weights allows for unequal probabilities of selection based on family size.

Where frames are incomplete or do not otherwise match the population of inference, it is sometimes an option to select multiple frames. The use of multiple frames is sometimes used to allow for uneven telephone coverage in which a telephone and an area frame are used. In the instance described by Groves & Lepkowski [5], a random-digit dialed sample was drawn and screened for household members. Interviews were conducted with an eligible respondent by telephone. A second area frame survey was also conducted face-to-face. During these interviews, information was obtained about telephone coverage in the sample. The overall mean for each variable was estimated by a weighted average of the mean values obtained from households with telephones and those without them.

This strategy does offer some reduction in coverage error, but also poses some administrative problems that can affect cost and themselves produce error if not appropriately addressed. Most obviously, there will be duplicates in the frames. If attempts are made to identify and eliminate duplicates before data collection, there is the risk of increased costs and error due to mismatching. On the other hand, if no matching is attempted prior to the interviewing, then duplicates may not be detected, thereby increasing interviewing costs and error. A second problem could result from the necessity to initiate more than one data collection procedure. A list frame may contain more information about the elements than an area frame, since the latter is derived from listing. Thus, the use of area frames often requires greater expense. Finally, duplicates are usually not purged but, rather, they are dealt with by weighting. Thus, the estimates from multiple frames require calculations based on different combinations of weights because individuals may be covered by different frames – hence the calculations are more complex than those from a single frame survey.

Finally, under certain circumstances, coverage error may be dealt with through post-weighting the data to adjust for coverage error. However, this strategy is only useful if there is some way of estimating the extent of coverage error. As noted above, national censuses, if they exist, may be the best source, but biases are also likely in any census listing, and they are probably the same as those an investigator might be trying to correct by weighting. The cost of arriving at those estimates has to be weighed against the gain in precision of estimates derived from adjusting the data. Also, the benefits of adjusting are generally greatest for simple linear statistics such as a mean or proportion [3].

References

[1] Czaja, R., Blair, J. & Sebestick, J. (1982). Respondent selection in a telephone survey: a comparison of three techniques, *Journal of Marketing Research* **19**, 381–385.

[2] Czaja, R., Casady, R.J. & Snowden, C.B. (1986). Reporting bias and sampling errors in a survey of a rare population using multiplicity counting rules, *Journal of the American Statistical Association* **81**, 411–419.

[3] Groves, R.M. (1989). *Surveys Errors and Survey Costs*. Wiley, New York.

[4] Groves, R.M. & Kahn, R.L. (1979). *Surveys by Telephone*. Wiley, New York.

[5] Groves, R.M. & Lepkowski, J.M. (1985). Dual frame, mixed mode survey designs, *Journal of Official Statistics* **1**, 263–286.

[6] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[7] Salmon, C.T. & Nichols, J.T. (1983). The next birthday method of respondent selection, *Public Opinion Quarterly* **47**, 270–276.

[8] Sirken, M. (1970). Household surveys with multiplicity, *Journal of the American Statistical Association* **65**, 257–266.

[9] Thornberry, O.T. & Massey, J.T. (1988). Trends in United States telephone coverage across time and subgroups, in *Telephone Survey Methodology*, R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey & W.L. Nicholls, II, eds. Wiley, New York.

[10] Trodahl, V.C. & Carter, Jr, R.E. (1964). Random selection of respondents within households in phone surveys, *Journal of Marketing Research* **1**, 71–76.

[11] United Nations (1950). *The Preparation of Sampling Survey Reports, Series C, No. 1*. United Nations, New York.

[12] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **73**, 40–46.

[13] Wiggins, B. (1995). *NNSP Newsletter* **21**. National Network of State Polls, Institute for Research in Social Science, Chapel Hill.

[14] Wright, T. & Tsao, H.O. (1983). A frame on frames: annotated bibliography, in *Statistical Methods and the Improvement of Data Quality*, T. Wright, ed. Academic Press, New York.

RICHARD B. WARNECKE

Sampling in Developing Countries

The objective of this article is to discuss issues that are important in the design and implementation of sample surveys that are undertaken in developing countries. Our approach is to discuss the major elements that anyone planning a sample survey in a developing country should take into consideration. For some of these elements, there are no special differences between planning a sample survey in the Third World and planning one in a developed country. Where there are differences, we will attempt to outline the special problems that are likely to occur in Third World surveys.

Objectives of the Survey

From a sampling point of view, it is not enough to state that a survey aims to provide information about, for example, potential users of family planning. The survey objectives must state precisely who will be sampled (sampling or selection unit), who will be interviewed (measurement or observation unit), and about what population will inferences be drawn (analysis unit) (*see* **Target Population**).

For example, a household survey might sample households, interview all female household members of childbearing age, and draw national inferences about children born.

Universe definition is also a very basic concept in sampling. For example, the objective might be to draw national inferences, but certain regions of the country might be considered unsafe or inaccessible because of reasons such as war, guerrilla activity, political unrest, disease, and so on. Under these circumstances, the universe might be redefined to include the country without certain regions.

Sample Size

Sample size refers to the total number of completed interviews. Because of a variety of factors such as **nonresponse**, it is almost always necessary to select an original sample larger than the targeted number of final completed interviews. The basic three issues that need to be considered in setting the sample size are:

1. available resources;
2. desired level of precision; and
3. analysis plans.

The relationship between *available resources* and sample size is the easiest to grasp. If, for example, the desired precision calls for a sample size of 1000 and the budget allows for 800, then the choice is a simple one of sampling the maximum permitted by the budget.

With respect to *desired precision*, there is fortunately a very simple "family" of statistical formulae that can be used to determine the sample size. These formulae permit one to establish a desired level of precision in terms of a **confidence interval** and use this target to derive the necessary sample size. It is important to note that, in most real-life situations, the *sampling precision does not depend on the population size* (*see* **Sample Size Determination**). It should also be noted that paying for a larger sample size has the benefit of reducing the **standard error** and increasing the precision of the results.

The third and final consideration with respect to sample size is the *analysis plan*; most importantly, the subclasses to be analyzed. The researcher first needs to decide for which major subclasses will results and **inferences** be desired. For example, in a national survey of condom sale and usage, it might be important to draw conclusions about individual cities. In this case, it will be necessary to establish sample sizes separately for each city and then cumulate these individual samples to arrive at the total sample size.

A second part of the analysis plan involves anticipating cross-classifications that will be carried out for each population of interest. For example, let us consider a survey whose objective is to study the relationship between education and condom usage. This goal suggests a cross-classification of the two variables, and it is important that the expected cell sizes be sufficiently large to permit the required analyses. A useful rule of thumb is to aim for minimum cell sizes of 20.

Sampling Frame

The first step in any sample selection is to produce a list of the population elements from which the sample is to be drawn. This is known as the **sampling frame**. Whereas samples based on telephone lists and **random digit dialing** schemes have become

part of the sampler's repertoire, it is highly unlikely that these methods could be used effectively in most developing countries. Not only is it often true that large portions of these countries' areas simply do not have telephone service, but even in areas where telephones are prevalent, the coverage is far from complete, and hardly a source for reaching targeted groups such as urban residents.

This leaves the sampler in a developing country to choose between a *list sample* and an *area sample*. However, again, it is doubtful that accurate and complete lists of population elements exist in these countries. Samplers therefore resort to designs based on area frames, whereby the first stage of sampling involves selecting relatively large geographic areas such as counties, districts, zones, municipalities, and so on. In order to implement this sample design, lists are required for the total number of such areas in the country.

Following are some criteria for evaluating sampling frames.

Frame Coverage, Accuracy, and Duplication

Starting with a good sampling frame is extremely important for a successful sample design. How complete is the frame? Are there elements of the population that are missing? If there are known missing elements, then the frame suffers from *undercoverage*. For example, a list of tobacco retail outlets, no matter how small, in a large city in a developing country is highly likely to be incomplete.

Are there foreign elements in the frame? That is, are there elements on the frame that should not be there? If so, then the frame suffers from *overcoverage*. For example, a list of schools in an urban area should not contain buildings that are not schools, for example, teacher colleges, kindergartens, and so on.

The information on the frame should be *accurate* and up-to-date. A sample drawn from an error-prone frame will not reflect the population that is the target.

Finally, the frame should be free of duplicates. If duplicates are known to be present, they should be removed or the selection probabilities should be adjusted.

In many developing countries it is often a major challenge to locate a reliable source to serve as a sampling frame. Recent **censuses** can often provide one solution but, if this source is unavailable, one must define and select large geographic units based on the most accurate measures of size that are available. Once these units are selected, one has to update constantly their counts (e.g. population, households, schools).

Not only are sampling frame unit counts often unavailable or unreliable, but the cartographic details can also be missing or defective. Accurate limits of the areas have to be specified and mapped, and the first visit to these areas necessitates careful cartographic work, which needs to be maintained and updated as part of future surveys in these areas.

With respect to establishment surveys in developing countries, it is worth pointing out that there is usually a less-skewed distribution of the units by size than is found in more industrialized nations. For example, there is likely to be a higher proportion of small farms in developing countries as well as lower concentration of total land and revenue in the larger agricultural holdings.

Sample Design

Probability sampling designs can include any or a combination of the following sampling techniques: **simple random sampling, systematic sampling**, **stratified sampling, cluster sampling**, or **multistage sampling**. At any stage in the sampling process, survey units may be selected with either equal or unequal probabilities.

Simple random sampling (SRS) and systematic sampling are two basic methods for randomly selecting samples. Both lead to equal probabilities of selection for every unit in the population, but systematic sampling is easier to implement.

In a single stage design, every sampled unit is surveyed. In a multistage design, selections are made first of larger units and then, in subsequent stages, subselections are made within these first-stage units. The units selected in the first stage are called Primary Sampling Units (PSUs); units selected in the second stage are called Secondary Sampling Units (SSUs). A typical household survey employs two to five stages of selection, the last of which is usually the selection of one eligible member from the household. The overall probability of selection of a multistage sample is equal to the product of the probabilities of selection at each sampling stage.

One popular sample design leading to equal probabilities of selection is known as **sampling with**

probability proportional to size (PPS) by which first-stage units (PSUs) are selected with probabilities proportional to some measure of size (e.g. number of households) and a constant number of second stage units are selected within each selected PSU. This approach is often particularly effective for developing countries where PSUs can vary considerably in size. Unfortunately, the drawback is that measures of size are often unknown or known poorly. The solution is to estimate and record accurately the measures during the first visit to the PSUs.

Stratification and Clustering

Stratification and clustering are two techniques utilized to group population elements before sample selection for the purpose of improving the practicality or the efficiency of the sample design.

Stratification involves the division of a population into parts called strata. Ideally, each stratum contains units that are homogeneous with respect to the survey variables of interest. A **random sample** is selected from each stratum. Examples of strata are large geographic areas such as districts, urban and rural areas in a country, and high and low socioeconomic status areas in a city.

Clustering also involves the division of the population into groups but, unlike strata, clusters are groups of heterogeneous population units. Ideally, each cluster should be a microcosm of the population. Some naturally occurring clusters are provinces, city blocks, and classrooms.

Briefly, stratification is "good" and clustering is "bad". Stratification tends to increase precision and to ensure representativity. Clustering decreases precision.

Implicit stratification can be very useful when the sample is being drawn from a list of units. For example, suppose that one has a list of administrative areas in a country and for each we have a reasonably accurate estimate of the total population. Stratification could be implemented by simply ordering the list of areas from smallest to largest and then systematically sampling every *n*th area starting at the top. Since small areas are at the beginning of the file and largest at the end, it is very unlikely that only small or only large areas will be in the sample. This procedure is known as implicit stratification, since the objective of stratification has been achieved without creating explicit strata.

Another useful concept is the *self-representing stratum*, which refers to strata in which no subsampling takes place at the first stage. For example, it is typical in a national area-based sample for large cities to be defined as self-representing strata. Self-representing strata are usually large densely populated areas. Once a city has been defined as a self-representing stratum, sampling is carried out in that city during subsequent stages of the sample design.

Allocation

Once the strata are created, it is necessary to decide what size sample to allocate to each (*see* **Stratified Sampling, Allocation in**). One approach is called *equal allocation* which consists of assigning to each stratum an equal sample size. For example, a total sample of 200 over 10 strata would require 20 sample elements per stratum. Another approach is *proportionate allocation*, which calls for allocation of the total sample to the strata in proportion to the stratum population sizes.

In practice, there is often a need for *oversampling* in certain strata. This would call for a strategy that is neither equal nor proportionate. This approach can be justified especially if the analysis calls for separate reports for the strata being oversampled.

Selecting Households and Respondents in Household Surveys

One of the most important types of survey is the household survey, in which households are usually selected in the penultimate stage and a respondent or several respondents in the last stage. It should be pointed out that in many developing countries the definition of what constitutes a household and family differs sometimes markedly from concepts commonly used in more industrialized nations. Care must be taken that the definition of household is specific, unambiguous, and easy to implement in the field.

At some stage in the survey process it is necessary to select households to interview. This can be done basically in two ways. One approach would be for interviewers (or *listers*) to visit each selected block and make a complete listing of all households in those blocks. Clearly, this is a time-consuming and expensive operation. However, once completed, the

lists can be used to draw accurate samples, to monitor field progress, and to draw samples for subsequent surveys.

A less resource-intensive strategy and one very often employed in sample surveys in developing countries is called *co-listing*, whereby interviewers both list and interview in one operation. They are given a map of the block, they are told where to start, in what direction to proceed (usually "serpentine"), and with what frequency to select households.

Once a household has been selected, it is often required to interview one adult in the household, and this person has to be selected at random. Many procedures have been devised to implement this procedure, the most rigorous one involving a complete listing of all persons in the household.

Execution

All sampling procedures need to be tested before production. Part of the design (PSU creation and selection) can often be implemented in the head office, but other stages – for example, selection of final sampling units – have to be tested and carried out in the field.

Even the best planned design will be subject to unpredictable deviations during the execution stage. Reasons are numerous: nonresponse, imperfect frame and population data, real change, human error, and so on. There are at least two corrective actions that can mitigate the effect of these problems. First, a focused attempt should be made to predict as many of these factors as possible and to make contingency plans. The second part of the solution is to keep meticulous, accurate, well-organized, and preferably computer-based quantitative records of all steps in the sampling process.

As with any large-scale and complex operation involving teamwork, it is imperative that rigorous monitoring and quality control be included as an integral and important part of the survey process. The work of all field staff, including interviewers – especially new ones – should be regularly checked for errors, both accidental and intentional.

The question that is asked almost as often as the one about sample size is "What level of response rate is considered acceptable?" First of all, it is not the response rate *per se* that is the danger – it is the nonresponse bias; that is, the extent to which responders differ from nonrespondents (*see* **Bias from Nonresponse**). It is conceivable, albeit unlikely, that a 20% response rate sample manifests little bias.

Secondly, there is no substitute for high response rates, both at the level of the respondent and at the level of the individual questions. A large part of the survey resources, training, and overall effort should be directed to increasing the response rate as much as possible. Many statistical procedures (weighting and imputation (*see* **Multiple Imputation Methods**)) have been developed to overcome nonresponse, but these should be considered imperfect solutions to a problem that is much better solved in the field.

At the risk of sounding arbitrary, we would suggest that surveys that are to be conducted in developing countries strive to achieve at least a 50% response rate and preferably a minimum of 60%.

Data Processing

Once data are collected in the field, they have to be processed; that is, edited, coded, keyed, cleaned, and analyzed (*see* **Data Management and Coordination**). Between the editing and analysis phases, it is necessary to calculate survey weights.

Weighting

Before the analysis phases, it is necessary to calculate survey weights to account for unequal probabilities of selection (planned or otherwise), other deviations from the design, and to make final adjustments that bring the sample results in line with known population distributions.

Unequal probabilities may arise at any stage of the sampling process. A separate weight is required for each stage of the sampling process in which differing probabilities are used.

The sampling weight is further adjusted to account for nonresponse. This entails knowing something about nonresponding outlets, information which can be gleaned from looking at response rates across various subclasses.

In the final stage in the weighting process a **post stratification** weight is derived by comparing the sample and population distributions for basic, known variables, such as region, age, race, sex, and income.

Analysis

It would be convenient, although unrealistic, if all surveys could be based on simple random samples with no stratification or clustering. However, this is almost never the case. It then becomes an important, although often overlooked, requirement of analyses based on sample survey data to include the complexity of the design in the analysis. The two components of complexity are stratification and clustering. They work in opposite directions, the first raising precision and the second lowering it. Unfortunately, in typical area-based surveys, the effect of clustering is dominant, and the overall result is that ignoring the complexity of the design will result in an underestimation of the true sampling **variance**. What this means is that results that might seem statistically significant are in fact not, if one takes into account the true sampling variability.

The complexity of the design can be incorporated by running specific software (*see* **Software for Sample Survey Data**) or by including design effects in the calculation of sampling errors.

Dissemination

As part of the dissemination plan, in addition to standard analyses, tables, and other products, it is very important to include a complete description of the sample design and all deviations that occurred in the field (e.g. nonresponse rates). This will help readers to determine how much confidence they can have in the results and the conclusions.

As a matter of course, it is also advisable to include standard errors (correctly calculated) in the tables.

Further Reading

Clairin, R. & Brion, P. (1996). *Manuel de Sondages – Applications aux pays en developpement*, Documents et manuels due CEPED No. 3. (This publication is a recent, comprehensive, yet nontechnical treatment of basic survey design issues with emphasis on topics that arise in developing countries.

Cochran, W. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York. (This is a classic, one of the standard texts about sampling with a heavily mathematical bent.

Food and Agriculture Organization (1989). *Sampling Methods for Agricultural Surveys*. FAO Statistical Development Series No. 3. FAO, Rome. (This document was prepared by Dr L. Kish and is intended to guide statisticians in their design of agricultural surveys. Many of the principles covered in this text can be transferred to other domains of study.)

Kish, L. (1965). *Survey Sampling*. Wiley, New York. (Another classic, but more applied and practical than Cochran (1977).)

Lohr, S. (1999). *Sampling Design and Analysis*, Duxbury Press, California. (Very complete, up-to-date, well-organized, and an excellent introduction to sampling as well as a reference source.)

Seijas, F. (1981). *Investigación por Muestreo*. Universidad Central de Venezuela, Caracas. (A fundamental text in the basics of survey research from the perspective of a statistician whose experience is drawn largely from his work in Latin America.)

Seijas, F. (1987). *Encuesta de Hogares por Muestreo*. Oficina Central de Estadística e Informática, Caracas, Venezuela. (A primer in household sample surveys with emphasis on challenges faced by statisticians in Latin America.)

United Nations (1986). *Sampling Frames and Sample Designs for Integrated Household Survey Programmes*. National Household Survey Capability Programme. United Nations, New York. (This excellent document represents a detailed and far-reaching treatment of sampling issues for developing countries especially for those trying to set up a permanent statistical system.)

United Nations (1989). *Household Income and Expenditure Surveys: a Technical Study*. National Household Survey Capability Programme. United Nations, New York. (Part of the same series as United Nations (1986), this text provides a wealth of information and experience for statisticians designing household income and expenditure surveys in developing countries.)

KAROL P. KRÓTKI

Sampling With and Without Replacement

Sampling with replacement is a class of sampling procedures in which the selected elements are replaced in the selection pool following each "draw" and may be reselected on subsequent draws. In contrast, *sampling without replacement* is a class of procedures in which the already selected elements are not replaced in the pool and cannot be selected again [1, p. 37].

Under **simple random sampling** without replacement of a sample of size n from a population of size N, the first element of the sample is selected with a probability of $1/N$; the second element of the sample is then selected from the remaining $N-1$ elements with a probability of $1/(N-1)$; and so on. Finally, the nth element of the sample is selected with a probability of $1/(N-n+1)$. It can also be shown from elementary combinatorial theory that the probability of the kth element of the population being included in the sample, under simple random sampling without replacement is n/N.

Under simple random sampling with replacement of a sample of size n from a population of size N, n independent draws are made such that in each draw, each element of the population has an equal probability, $1/N$, of being selected. Since, after each draw, the selected element is replaced into the population, some elements in the sample may be drawn more than once. Thus, the probability that the kth element of the population is *not* included in the sample is $(1-1/N)^n$ and its inclusion probability is $1-(1-1/N)^n$ [2, p. 49].

Some further points concerning the relationship between sampling with and without replacement are as follows:

1. The **variance** of conventional estimators (e.g. means, total, or ratios) under simple random sampling without replacement is $(N-n)/(N-1)$ times its value under simple random sampling with replacement (*see* **Finite Population Correction**), when conventional estimators are applied [3, pp. 29–30]. That is, the variance in sampling without replacement is smaller than that in sampling with replacement. However, the variances under the two sampling schemes are nearly equal when the population size, N, is large and the sample size, n, is small.
2. In practice, almost all sampling is conducted without replacement, and the "classical" methods of finite population sampling theory were developed primarily for this class of sampling designs.
3. Under certain conditions, results derived under the assumptions of sampling with replacement are approximately equivalent to those that would have been obtained under sampling without replacement. One example is **sampling with probability proportional to size**.

References

[1] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[2] Sarndal, C.E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[3] Cochran, W.G. (1977). *Sampling Techniques*. Wiley, New York.

JASON HSIA

Sampling With Probability Proportional to Size

There are a number of reasons for selecting a sample with probabilities proportionate to size (PPS). As the name "proportionate to size" implies, the sampling elements are aggregated or arranged into groups or units of different sizes. Examples include students in schools, people in cities or counties, and patients served by hospitals. Sampling PPS is an efficient design when: (i) a list of the population is not available but ancillary information is; (ii) the elements are in clusters or groups which differ greatly in size; (iii) it is important to control the sample size; (iv) costs and equalizing interviewer workloads are important; (v) the use of natural divisions is possible and supplementary information is available by the natural divisions; and (vi) the sample must be drawn in multiple stages.

The quintessential example of PPS sampling is a multistage area probability sample (*see* **Multistage Sampling**). A typical study might be of individuals that reside in households and the research objectives might be to estimate a health characteristic, action, or belief. An area probability sample satisfies all of the above conditions. In the US, for example, no complete list of the general population exists. Lists are available of telephone subscribers, registered voters, property tax payers, people with driver's licenses, and others, but all of these have major omissions/exclusions and have varying degrees of currency and turnover. As a result, they are not acceptable **sampling frames** for most studies. The most comprehensive listing of the US population is the count of the population in various types of domiciles every 10 years by the US Census Bureau. The decennial **census** provides counts by various civil and political areas: in other words, states; counties, parishes, or boroughs; Indian reservations, towns, or townships; census tracts or block numbering areas; and blocks. This information is used to construct the initial phases of a PPS sampling frame. Social and economic characteristics of the population are used to stratify (*see* **Stratified Sampling**) and order the geographic entities before selection.

Because no list of the population exists, the sample is drawn in stages using population or housing unit count data. These data indicate how many people reside in a defined geographic area and how many households there are in the same area, but they do not tell us where and what types of housing unit structures are located within the area, how many people reside in each household, or which households are occupied or vacant. The sample design is nested in that subsequent selections are made within areas selected in previous stages. At the next to last stage of selection, which is typically blocks or segments of blocks, a person trained in the listing of housing units and households is sent to each selected area. A comprehensive list of all dwelling units within every commercial and residential structure is made. The final sampling rates are applied to these lists of dwelling units.

An advantage of PPS sampling over **simple random sampling** is control of the final sample size and the workload for interviewers. We will illustrate the first point after a brief explanation of the latter point. Selecting and training interviewers is usually handled in one of two ways. One approach is to hire and train in one location and then send the interviewers to each sample location. This works well when travel costs and distances are not very great. When the sample covers a large geographic area (e.g. several states or the entire US) a better approach is to hire interviewers in multiple locations. Training is then carried out at a number of regional locations to reduce travel costs. With either approach, it clearly makes sense to equalize the workload across interviewers. It is not cost effective to hire and train an interviewer to conduct only a few interviews if the average workload is 25–40 cases.

In densely populated areas, interviewers are usually assigned cases in their home counties. In small towns and rural areas, they are given cases not only in their home county but also in adjacent or nearby counties. The travel distances are longer but the travel times are comparable to large urban areas. Interviewers are typically paid for both travel time and expense from their homes to the sample locations; the actual time to conduct the interviews is estimated to be only 25%–40% of the total interviewing costs [7]. A major portion of the total cost is attributable to interviewer travel time and expenses. Thus, selecting more than one household at each sample location is a way to reduce interviewer costs; however, it typically increases sample **variances**. Considerations for addressing this issue are discussed later.

With the above points in mind, we can now illustrate why using a PPS design is more efficient than using an equal probability of selection method (epsem) at all stages of selection in a multistage design. Assume that we want to select a sample of 1200 individuals in the State of North Carolina. Further assume that we want the design to satisfy the following conditions: (i) the sample should be epsem or self-weighting; (ii) the sample size should yield approximately 1200 interviews; and (iii) the workload among interviewers should be fairly equal. There are two important design components in a multistage epsem sample: (i) the overall probability of selection, f, where $f = n/N$, with n being the sample size and N being the population size; and (ii) determining the number of units selected at the first stage of sampling – the primary sampling units (PSUs) – or choosing the average number of sampling elements to be interviewed at the last stage of selection – which is usually the block level. When the number of PSUs is large and expensive to subdivide, the number of PSU selections should not exceed one-fifth of the total number of units; otherwise, the expense for this stage of selection cannot be justified [3]. Counties are frequently used as PSUs because there are a large number of them (more than 3000 in the continental US), they are important administrative and governmental units, they are stable, and measures of size and other useful sampling information for them are readily available [3].

North Carolina contains 100 counties. If we want 1200 interviews from a population of approximately 8 million people, the overall probability of selection $\Pr_T = n/N = 1200/8\,049\,313 = 0.000149$, or a sampling rate of 1/6707. If we select an epsem sample of 20 counties at the first stage of selection, the sampling rate within counties will be 1 in 1341 people, because $1/6707 = (20/100)(1/1341)$.

Table 1 presents the results from a systematic random sample of 20 counties and the number of cases to be selected within each county using a rate of 1 in 1341. The results show that instead of 1200 cases being selected, there are only 881: a shortage of 319 cases or approximately 27%. The number of cases selected per county varies widely, from a high of 110 cases in Davidson County to a low of three cases in Tyrrell.

A related inefficiency of this design is the estimate of variance of the sample **mean**. If we assume equal population variances by county, a PPS sample of 20 counties and 60 persons per county would give the following estimate [7]:

$$\sum_{i=1}^{n} \frac{\sigma^2}{n_i} = \frac{20\sigma^2}{60} = 0.33\sigma^2.$$

The estimate from Table 1, however, is $0.9638\sigma^2$, or 2.92 times greater.

A few additional comments about these results, and about using an epsem when PSUs vary in size, are in order. The population of North Carolina counties ranges from a low of 4149 in Tyrrell County to a high of 695 454 in Mecklenburg; the latter being more than 167 times larger than the former. Among the 100 counties, five have a population of 250 000 or more and 23 have a population greater than 100 000. The important points to note about the results in Table 1 are that none of the five largest counties was selected, only three of the counties with a population over 100 000 were selected, and counties with the fewest sampled cases contribute disproportionately to the higher estimated variance. Sampling with probabilities proportionate to size at all stages except the last stage in a multistage design can correct these design deficiencies.

Table 1 Expected sample sizes and $1/n_i$ from 20 randomly selected counties.

North Carolina counties	Population	Sample size (population/1341)	$1/n_i$
Anson	25 275	19	0.0530
Bladen	32 278	24	0.0415
Caldwell	77 415	58	0.0173
Chatham	49 329	37	0.0272
Columbus	54 749	41	0.0245
Davidson	147 246	110	0.0091
Franklin	47 260	35	0.0284
Granville	48 498	36	0.0277
Haywood	54 033	40	0.0248
Iredell	122 660	91	0.0109
Lenoir	59 648	44	0.0225
Martin	25 593	19	0.0524
Nash	87 420	65	0.0153
Pamlico	12 934	10	0.1037
Pitt	133 798	100	0.0100
Rockingham	91 928	69	0.0146
Stanly	58 100	43	0.0231
Tyrrell	4149	3	0.3232
Washington	13 723	10	0.0977
Yadkin	36 348	27	0.0369
Total		881	0.9638

Table 2 PPS selection of 20 counties arranged by median household income

No. of Co.	County	Income	Population	Cum Pop	Sampling Int.
1	Wake	54 988	627 846	627 846	r.s.126 599; 529 031
2	Union	50 638	123 677	751 523	
3	Mecklenburg	50 579	695 454	1 446 977	931 529; 1 333 994
4	Cabarrus	46 140	131 063	1 578 040	
5	Durham	43 337	223 314	1 801 354	1 736 459
6	Chatham	42 851	49 329	1 850 683	
7	Guilford	42 618	421 048	2 271 731	2 138 924
8	Dare	42 411	29 967	2 301 698	
9	Orange	42 372	118 227	2 419 925	
10	Forsyth	42 097	306 067	2 725 992	2 541 389
11	Iredell	41 920	122 660	2 848 652	
12	Lincoln	41 421	63 780	2 912 432	
13	Moore	41 240	74 769	2 987 201	2 943 854
14	Johnston	40 872	121 965	3 109 166	
15	Currituck	40 822	18 190	3 127 356	
...	...	...	...	...	
76	Avery	30 627	17 167	7 366 906	
77	Mitchell	30 508	15 687	7 382 593	7 370 969
78	Pasquotank	30 444	34 897	7 417 490	
79	Duplin	29 890	49 063	7 466 553	
80	Anson	29 849	25 275	7 491 828	
81	Yancey	29 674	17 774	7 509 602	
82	Perquimans	29 538	11 368	7 520 970	
83	Alleghany	29 244	10 677	7 531 647	
84	Washington	28 865	13 723	7 545 370	
85	Richmond	28 830	46 564	7 591 934	
86	Ashe	28 824	24 384	7 616 318	
87	Martin	28 793	25 593	7 641 911	
88	Swain	28 608	12 968	7 654 879	
89	Hyde	28 444	5826	7 660 705	
90	Warren	28 351	19 972	7 680 677	
91	Robeson	28 202	123 339	7 804 016	7 773 434
92	Cherokee	27 992	24 298	7 828 314	
93	Bladen	26 877	32 278	7 860 592	
94	Columbus	26 805	54 749	7 915 341	
95	Northampton	26 652	22 086	7 937 427	
96	Graham	26 645	7993	7 945 420	
97	Halifax	26 459	57 370	8 002 790	
98	Hertford	26 422	22 601	8 025 391	
99	Tyrell	25 684	4149	8 029 540	
100	Bertie	25 177	19 773	8 049 313	

Table 2 illustrates one method of preparing the sample frame for the selection of PSUs in a PPS sample. To conserve space, only 40 of the 100 counties are listed. The counties are arranged in descending order by a **stratification** variable, **median** household income. The number of people in each county is listed and the cumulative total is calculated. We can see, for example, that Wake County has the highest median household income of 54 988 and a population of 627 846, which is the second largest. The cumulated subtotals are required because at this stage of selection each county receives a probability of selection which is proportionate to its population size.

Table 2 shows county population and median household income data for the 15 counties with the highest median household incomes and the 25 counties with the lowest median household incomes. As in Table 1, we want to select 20 counties systematically, but this time by probabilities proportionate to their population sizes. To do this we divide the total

cumulative population (8 049 313) by the desired number of PSUs, $a = 20$. This gives a sampling interval of 402 465. A random start between 1 and 402 465 is selected (126 599), and the sampling interval is added to the random start until the cumulated total is exceeded.

The results in Table 2 indicate that four of the five counties (Cumberland is not shown and was not selected) with populations exceeding 250 000 were selected in the sample and that only two of the last 25 counties were selected. There is a moderate **correlation** between population size and household income and all but two of the 23 largest counties (population of 1 000 000+) are in the first 55 counties listed. Note that the sampling interval falls twice within Mecklenburg and Wake Counties. If we select approximately 60 cases per county, a double sample or 120 cases would be selected in these two counties. Also, because Mecklenburg and Wake were selected twice, only 18 different counties were selected. Counties such as Mecklenburg, Wake, and Guilford are selected into the sample "with certainty" because their population sizes exceed the sampling interval.

In situations in which some of the PSUs are larger than the sampling interval, each certainty PSU should be treated as a separate stratum. The overall sampling rate is applied to each certainty PSU and the number of selected cases per PSU is allowed to vary depending on the size of the PSU. The noncertainty counties are treated as a separate group. With this approach, the certainty PSUs are sampled using the initial overall selection rate and the noncertainty counties are allocated the remaining cases and sampled PPS with a new sampling interval.

To illustrate this approach we proceed as we did in Table 2. We want to select 20 PSUs and a sample size of 1200. We calculate the sampling interval as before, but before we select a random start, we look over the list of counties to see how many will fall into the sample with certainty; that is, have a population greater than 402 465. There are three counties: Wake, Mecklenburg and Guilford. We now subtract the total population of these three counties (1 744 348) from the state total (8 049 313) and divide the new total (6 304 965) by the number of PSUs that we still want to select (17). This gives a new interval of 1 in 370 880. Again, we check the list of 97 counties to see if any are now certainty counties and we see that none are. If any counties were certainty, we would deal with them as we did the first three. This process would continue until no counties were certainty selections. When that occurs, the final PSUs are selected.

We can see that at this stage in the selection process the probabilities of selection among the PSUs are unequal and, therefore, a fixed sampling rate must be used at the final stage of selection [6]. Since the PSUs were selected proportionate to size, the final elements must be selected inversely proportionate to size in order to give each element the same overall probability of selection. This is demonstrated as follows [7]:

$$\Pr_{\mathrm{T}} = \frac{n}{N}$$

is the overall probability of selection;

$$\Pr_{\mathrm{PSU}} = \frac{MOS_{\mathrm{PSU}}}{(N/a)}$$

is the probability of a PSU being selected, where MOS_{PSU} is the population count for the county and a is the desired number of PSUs to be selected;

$$\Pr_{\mathrm{W}} = \frac{\Pr_{\mathrm{T}}}{\Pr_{\mathrm{PSU}}}$$

is the probability of selection within the PSU.

$\Pr_{\mathrm{W}}$ is known as the selection equation [2]. Essentially, it is the probability of a PSU being selected multiplied by the probability of an element within the PSU being selected if its PSU was selected at the initial stage. This can be seen more clearly by solving for $\Pr_{\mathrm{W}}$:

$$\Pr_{\mathrm{W}} = \frac{\Pr_{\mathrm{T}}}{\Pr_{\mathrm{PSU}}} = \frac{n}{N} \times \frac{N}{aMOS_{\mathrm{PSU}}} = \frac{n}{aMOS_{\mathrm{PSU}}}.$$

The number of selected elements in a PSU is determined by the PSU size multiplied by the probability of selection within:

$$n_i = MOS_{\mathrm{PSU}} \left(\frac{n}{aMOS_{\mathrm{PSU}}} \right) = \frac{n}{a}.$$

Thus, the same number of elements will be selected from each noncertainty PSU regardless of the size of the PSU. For certainty PSUs, $\Pr_{\mathrm{W}} = \Pr_{\mathrm{T}}$ because $\Pr_{\mathrm{PSU}} = 1$. The number of elements to be selected within a certainty PSU is $MOS_{\mathrm{PSU}} \times n/N$. Assume that Wake County and Mitchell County are selected. Wake is a certainty and Mitchell is a noncertainty county. For Wake, the number of selected elements is determined by $\Pr_{\mathrm{T}} = n/N = 0.000149 \times$

$MOS_{\text{PSU}}(627\,846) = 93.6$. The sample size not allocated to the certainty counties will be allocated equally among Mitchell and the other noncertainty counties. Thus, 940 cases allocated to 17 counties is approximately 55 per county. We can illustrate this with Mitchell County. The noncertainty selection interval is 370 880. Mitchell County's *MOS* and probability of selection are $15\,687/370\,880 = 0.042297$.

$$\text{Pr}_{\text{W}} = \frac{\text{Pr}_{\text{T}}}{\text{Pr}_{\text{PSU}}} = \frac{0.000149}{0.042297} = 0.00352.$$

The estimated number of selected elements in Mitchell County is $15\,687 \times 0.00352 = 55.2$ or 55, or 57.

The same PPS procedures are used when there are two or more stages to the selection process. This is illustrated as follows [7]:

$$\text{Pr}_{\text{PLACE}} = \frac{MOS_{\text{PLACE}}}{MOS_{\text{PSU}}/b}$$

is the probability of a place within a PSU being selected;

$$\text{Pr}_{\text{BLOCK}} = \frac{MOS_{\text{BLOCK}}}{MOS_{\text{PLACE}}/c}$$

is the probability of a block within a place being selected;

$$\begin{aligned}\text{Pr}_{\text{HOUSE}} &= \frac{\text{Pr}_{\text{T}}}{\text{Pr}_{\text{PSU}}\text{Pr}_{\text{BLOCK}}\text{Pr}_{\text{PLACE}}} \\ &= \frac{n}{N}\left(\frac{MOS_{\text{PSU}}}{N/a} \times \frac{MOS_{\text{PLACE}}}{MOS_{\text{PSU}}/b}\right. \\ &\quad \left. \times \frac{MOS_{\text{BLOCK}}}{MOS_{\text{PLACE}}/c}\right)^{-1} = \frac{n}{abcMOS_{\text{BLOCK}}}\end{aligned}$$

is the probability of a household within a block or block area being selected. Here, a is the desired number of PSUs, b is the desired number of places to be selected per PSU, and c is the desired number of blocks to be selected per place. The expected number of cases per block or block area is

$$\begin{aligned}MOS_{\text{BLOCK}} \times \text{Pr}_{\text{HOUSE}} &= \frac{MOS_{\text{BLOCK}} \times n}{abc \times MOS_{\text{BLOCK}}} \\ &= \frac{n}{abc}.\end{aligned}$$

The number of places and blocks selected within a PSU is based on costs and the degree of homogeneity within clusters for the variable being estimated. These factors will be discussed shortly. If we had determined that in noncertainty PSUs seven places and two blocks per place should have been selected, the number of households selected per block would be approximately: $n/abc = 940/17 \times 7 \times 2 = 3.95$. For certainty PSUs, where the sample sizes are larger, the number of places selected is increased while the number of households selected per block remains the same.

A number of design considerations from the foregoing example need to be addressed. Many of the points raised in the following five items apply to **cluster sampling** in general.

1. Very seldom do we know the exact sizes of our sampling units. This is especially true when dealing with human populations because areas change in size due to growth, migration, or demolition of housing units. In most situations, good estimates are available from a recent census, real estate or neighborhood groups, city and county governments, and other groups that monitor or are involved in population change. It is important that measures of size be reasonably accurate so that serious selection problems are avoided. A serious problem would be selecting a block with an expected 25 housing units and, upon visiting the block, finding that a 500-unit condominium has been built. For a sample to remain epsem, the cluster size at the final stage of sampling must be allowed to vary to reflect the differences between the expected and the actual sizes. Detailed discussions of listing blocks and areas and how to deal with high growth and zero household blocks can be found in Kish [3] and Sudman [7].
2. When we use estimated measures of size, the sample size becomes a random variable. The sample mean is estimated by $r = y/x$, where y is the sample total for the variable being estimated and x is the total sample size. This estimator is not unbiased, but the **bias** can be ignored when the coefficient of variation of x (*see* **Standard Deviation**) is less than 0.2 [2]. The general form of the variance estimator is

$$v(r) = \frac{[v(y) + r^2 v(x) - 2rc(x, y)]}{x^2},$$

where $c(x, y)$ is the sample covariance between x and y, and $v(y)$ and $v(x)$ are sample variances. This formula holds whether or not PSUs are

stratified and the method of subsampling within PSUs does not matter [2, 3] (*see* **Ratio and Regression Estimates**).

3. When elements are selected in clusters, two things occur. The elements are not selected independently of each other as in simple random sampling and elements that are grouped together usually have some degree of similarity or homogeneity. The degree of homogeneity within clusters and the average cluster size typically cause an increase in sampling error. Levy & Lemeshow [5] provide the following formula for the standard error of the element mean from a two-stage cluster sample:

$$\text{se}(\overline{\overline{x}}_{\text{CLU}}) = \left(\frac{\sigma_x}{\sqrt{n}}\right)[1 + \delta_x(\overline{c} - 1)]^{1/2}.$$

The only difference between the **standard error** for a simple random sample and the above formula is the expression in the square brackets. This is known as the **design effect**. It is the ratio of the variance from a cluster sample to that from a simple random sample of the same size. $\overline{c}$ is the average number of households sampled per block and δ_x is the intraclass **correlation** coefficient. δ_x is a measure of the homogeneity among all possible pairs of elements. This parameter can take values between $+1$, when all the elements are similar on the characteristic, and $-1/(\overline{c} - 1)$, which would indicate that there is more variability among the elements within a cluster than in a simple random sample. The intraclass correlation is typically small, positive and below 0.15 [2].

4. Costs need to be considered when determining the number of PSUs to select and the average cluster size per block. A cost function for the sampling and field costs for a two-stage design is

$$C_{\text{T}} = aC_1 + nC_2,$$

where C_{T} is the total cost of sampling and fieldwork; a is the number of PSUs; C_1 is the average cost of sampling, listing, and hiring, selecting, training, and supervising interviewers at each PSU; n is the total sample size; and C_2 is the average cost of each interview. The optimum cluster size per block can be determined from

$$\overline{c}_{\text{opt}} = \left[\frac{C_1}{C_2}\left(\frac{1-\delta}{\delta}\right)\right]^{1/2}.$$

This expression shows that when C_1 is much larger than C_2 and δ is small, the optimum cluster size can be large. Sudman [7] indicates that the optimum cluster sizes per block for many social science variables ranges from three to eight. The number of PSU selections can be determined from [1]:

$$a_{\text{opt}} = \frac{C_{\text{T}}}{(C_1 + C_2\overline{c}_{\text{opt}})},$$

where C_{T}, C_1, C_2, and $\overline{c}_{\text{opt}}$ are defined above.

5. The same measure of size information need not be used at every stage of selection in a multistage design. For example, PSUs can be selected using population count data and places within PSUs can be selected using household count data. This is possible because there is a very high correlation (0.97) between these items of information [7].

Sampling PPS for Telephone Surveys

One of the most innovative **random digit dialing** (RDD) telephone sample designs uses PPS sampling [8] (*see* **Telephone Sampling**). Early national RDD samples were designed to select a **systematic sample** of area code/prefix combinations from a national list frame. Four random digits were added to these six-digit numbers to form a complete ten-digit telephone number. The problem with the design is that 75%–80% of the telephone numbers generated are not residential household numbers. Waksberg [8] proposed a two-stage procedure which increases the proportion of working residential household numbers by subsampling within banks of numbers where a household is found at stage 1. Essentially, ten-digit numbers are formed as described above; however, if a household is contacted, a new sample number is created based on the initial number. The Mitofsky–Waksberg procedure suggests keeping the first eight digits (the area code + prefix + first two random digits) and substituting two random digits for the last two digits to create a new ten-digit telephone number. New numbers are created at the second stage of sampling until a fixed number of households are contacted. This is a PPS design because the first-stage selection probabilities are determined by the number of residential telephone households in the 100-number bank selected and, since a fixed number of households are called at the second stage,

the probabilities at this stage are inversely proportional to size. With this design, about 60% of the numbers generated at the second stage are residential households. Lepkowski [4] and others [5] discuss difficulties encountered with this design and a much simpler design, list-assisted RDD, that has, in general, near complete telephone population coverage and a small bias (*see* **Telephone Sampling**).

References

[1] Foreman, E.K. (1991). *Survey Sampling Principles*. Marcel Dekker, New York.

[2] Kalton, G. (1983). *Introduction to Survey Sampling*. Sage, Beverly Hills.

[3] Kish, L. (1967). *Survey Sampling*. Wiley, New York.

[4] Lepkowski, J.M. (1988). Telephone sampling methods in the United States, in *Telephone Survey Methodology*, R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Niclolls, II & J. Waksberg, eds. Wiley, New York.

[5] Levy, P.S. & Lemeshow, S. (1999). *Sampling of Populations: methods and applications*. Wiley, NewYork.

[6] Moser, C.A. & Kalton, G. (1972). *Survey Methods in Social Investigation*, 2nd Ed. Basic Books, New York.

[7] Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

[8] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **73**, 40–46.

RONALD CZAJA

Savage, Leonard Jimmie

Born: November 20, 1917, in Detroit.
Died: November 1, 1971, in New Haven.

Leonard Savage lived all his life with very bad eyesight. As a result, whenever he read anything, and he read a lot and widely, he read with deep concentration and understanding. Not for him the superficial appreciation that the modern information revolution often imposes. He gained a Ph.D. in mathematics from the University of Michigan in 1941. His future career was determined by working in the Statistical Research Group at Columbia University. This he described as "one of the greatest hotbeds statistics has ever had". Thereafter, his main professionalism was in statistics. From 1946 to 1960 he was at the University of Chicago. After a brief period at the University of Michigan, he moved to Yale in 1964.

He gained an international reputation as a result of the publication in 1954 of his book *The Foundations of Statistics* [1]. Prior to 1954, statistics had been a collection of separate techniques. He felt that it should be treated like any other branch of mathematics, with its axioms, proofs and theorems. Inspired by the success of von Neumann in developing an axiom system for **utility**, Savage was able to construct a system for **inference** and decision making (*see* **Decision Theory**). In particular, he showed that uncertainty must be described probabilistically. The axioms of Kolmogorov became theorems in the new system. He has been described as the Euclid of statistics.

Savage set out to justify the separate techniques of statistics, like **confidence intervals**. They were to appear as theorems in his deductive system. While the first half of the book is a triumph, the second, in which he attempts to justify contemporary statistics, is a failure. By 1960 he had realized that what he had done was to destroy frequentist statistics. In its place, he had constructed what we now call Bayesian statistics (*see* **Bayesian Methods**). In particular, the confidence interval (a probability statement about an interval, given the value of a parameter) is replaced by a probability statement directly about the parameter. Savage, thus unwittingly, produced a revolutionary paradigm for inference and decision making. This revolution is, almost 50 years later, affecting the whole of statistics and much decision making. The plethora of significance tests (*see* **Hypothesis Testing**) is being replaced by probabilities of hypotheses. These probabilities, as Savage and **de Finetti** showed, are personal, or subjective. Objectivity in science arises only through the accumulation of data and the bringing together of different opinions.

Savage was a scholar of the old school. He was meticulous in his appreciation of the work of others. He was the first to understand the work of de Finetti and Frank Ramsey, done in the 1930s. He discovered independently some of their results and, by his work, was the first fully to understand what they had done. Although Bayesian statistics is often in disagreement with the ideas of **Fisher**, Savage's article on his work [2] is perhaps the best from which to understand the genius of Fisher. Savage's collected works, with further biographical details, appeared in 1981 [3]. He was a superb writer and lecturer who spoke and wrote with the scrupulous care that he devoted to his reading.

References

[1] Savage, L.J. (1954). *The Foundations of Statistics*. Dover, New York.

[2] Savage, L.J. (1976). On Rereading R.A. Fisher, *Annals of Statistics* **4**, 441–500 (J.W. Pratt,. ed., with discussion).

[3] Savage, L.J. (1981). *The Writings of Leonard Jimmie Savage: A Memorial Selection*. American Statistical Association, Washington.

DENNIS V. LINDLEY

Scan Statistics for Disease Surveillance

Epidemiologists investigating disease incidence are drawn to clusters of cases occurring within a short period of time (*see* **Clustering**). Public health specialists as well as the media focus on clusters of cases of birth defects, cancer or suicides. Investigators seek to determine whether such clusters are more than just chance bunching, and search for common causative factors. Statistical tests are used to provide the researcher with a measure of the unusualness (statistical significance) of the cluster relative to chance bunching (*see* **Hypothesis Testing**). A popular statistic used in **disease surveillance**, first investigated in detail by Naus [37], is the scan statistic, the maximum number of events in a window of predetermined width, w. Formally, if we rescale time to [0, 1], and let $Y_w(t)$ be the number of events in time $[t, t + w]$, then the scan statistic, S_w, is

$$S_w = \max_{0 \le t < 1-w} Y_w(t). \tag{1}$$

(Alternatively, without rescaling time, set T to be the time frame of interest, r the duration of the window in actual time units, $w = r/T$, and let S_w be the maximum of $Y_r(t)$ over $0 \le t \le T - r$.)

The most common application of the scan statistic is testing for clustering conditional on N, the total number of events observed. Under the **null hypothesis**, H_0, the times of the N events have a **uniform distribution** on the unit interval. Rejection of H_0 for large values of the scan statistics is a generalized **likelihood ratio test** [38] for testing against the pulse alternative H_a: that for some unknown τ, representing the start of the increase, $0 \le \tau \le 1 - w$, and some **relative risk** $\theta > 1$, the density is given by

$$f(t) = \begin{cases} \theta/(1 - w + w\theta), & \tau \le t < \tau + w, \\ 1/(1 - w + w\theta), & \text{otherwise.} \end{cases} \tag{2}$$

The scan statistic has been used to detect clustering of a wide variety of reproductive and other outcomes, including congenital heart disease [41] and poisonings [20], and is used routinely in the Ontario Cancer Registry [27]. Applications to a wide variety of topics including cancer clusters, clusters of inflammatory bowel disease, parasuicide clustering, clusters of HIV in dialysis patients, and visual perceptions are described in [18].

The unconditional version of the statistic could be used to sound an alarm in real time in prospective surveillance applications, and requires the investigator to specify, a priori, the duration of the interval under study, and λ, the expected number of events over the entire time period. The null hypothesis for this model assumes that events occur at random according to a **Poisson process**. An analogous pulse-type alternative is that for some unknown τ, and unknown $\theta > 1$, $\mathrm{E}[Y_w(\tau)] = \theta\lambda w$, while for $t < \tau - w$ or $t > \tau + w$, $\mathrm{E}[Y_w(t)] = \lambda w$. However, since, in practice, the purpose of monitoring is to stop when a cluster is observed, the test could be applied when there are two intensities with an unknown change point Kulldorff [30] describes a generalization of the scan statistic for surveillance, in which w, the interval of the presumed increase, need not be specified in advance.

Often the scan statistic cannot be exploited fully since the precise times of the events are not known, but rather the data are grouped in discrete intervals. The most frequent application is when data are tabulated monthly, but clustering over 3, 6 or 12 months is of interest. The ratchet scan statistic [29, 51] maximizes $Y_w(t)$ when t can only take on values starting at the beginning of a calendar month.

Weinstock [54] modifies the scan so it can be used even when there is some underlying temporal trend to the disease specified by the density $f_0(t)$, or if the population at risk changes. He tests the hypothesis H_0: $f(t) = f_0(t)$ by replacing the constant window, w, by a variable window width $\omega(t)$, where

$$\int_t^{t+\omega(t)} f_0(s)\,\mathrm{d}s = w.$$

The statistic, however, thus loses its simple interpretation, and the associated optimal properties related to detecting pulses of length w.

A defect of the scan statistic is that w must be specified before the data are observed and should not be based on examination of the data. (For the surveillance model, both w and the time frame must be specified in advance.) Cressie [9] notes that it is better to choose an interval slightly larger than

the true pulse rather than one slightly smaller. Some protection against missing a cluster can be achieved by choosing two window widths and utilizing the **Bonferroni bounds** to test each at the $\alpha/2$ level. Loader [35] and Nagarwalla [36] present a statistic for testing H_0 against (2) in the case when w is unknown.

The Exact Distribution

Since the scan statistic does not have a **normal distribution** even for large N, and, furthermore, moments are unavailable, most of the literature has focused on finding critical values, k, so that under H_0 the occurrence of k or more events in any window of width w is unlikely to be due to chance. Naus [37] calculates $\Pr(S_{1/2} \geq k)$, $\Pr(S_{1/3} \geq k)$, and $\Pr(S_w \geq k)$, $w \geq 1/2$, under the null distribution of no clustering. General formulas for the exact distribution under both the null and the alternative are based on a generalized ballot problem [6] dealing with the amount of lead among L candidates. Naus [38] applies the result to express the distribution under the null, when $w = 1/L$, L an integer, as the sum of $L \times L$ determinants. The result was extended to arbitrary w [23], and to the distribution under a pulse alternative [9]. In general, these exact formulas are difficult to implement for moderate or large samples and small w, except in specialized cases, because they involve a large number of summations over many large determinants.

Approximations and Bounds

Many approximations or bounds are based on generalized Bonferroni-type inequalities involving intersections of up to J events, or on tighter versions of these inequalities for $J = 1$ [22] or $J = 2$ [33]. Especially for $J = 1$ and 2, these methods yield good approximations for small values of $\Pr(S_w \geq k)$, but are generally poorer for approximating the median of S_w or upper tail probabilities. Wallenstein [48] applies the simple bounds with $J = 1$ and 2 to $D_i = \sup_{0 \leq s \leq w} Y_w(iw + s) \geq k$, $i = 0, 1, \ldots, [1/w]$, to tabulate probabilities for a range of values of w common in disease surveillance applications. Berman & Eagelson [7] apply the upper bound with $J = 1$ based on $E_i = \{X_{(k+i-1)} - X_{(i)} < w\}$, $i = 1, \ldots N - k$, where $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ are the **order statistics** for the N events. Glaz [14, 15] derives tighter, but computationally more difficult approximations, by applying Bonferroni-type inequalities with larger J. An approximation that is both simple and accurate, involving sums and alternating sum of binomial coefficients, is given in [17].

Naus [40] develops a highly accurate formula for type I error for several types of scan statistic, by noting that conditioning on the recent past is approximately the same as conditioning on the entire past, or formally that $\Pr(D_i^c | D_1^c, \ldots, D_{i-1}^c) \cong \Pr(D_i^c | D_{i-1}^c)$. Thus, $\Pr(S_w \geq k)$ can be approximated based only on $\Pr(S_{1/2} \geq k)$ and $\Pr(S_{1/3} \geq k)$. Huffer & Lin [21] define M_k to be the number of k-clusters of length w, note that $\Pr(S_w \geq k) = \Pr(M_k \geq 1)$, and approximate this probability using the **method of moments**.

Applying the Hunter [22] bounds to $\Pr(\cup D_i)$, and then performing further approximations, a simple approximation for the null distribution of the scan statistic is [49]

$$\Pr(S_w \geq k) \cong (k/w - N + 1)\Pr(Z = k) + 2\Pr(Z > k), \quad (3)$$

where $Z \sim \text{bin}(N, w)$, i.e. Z is a **binomial random variable** based on N trials with probability of success, w. The first term (with coefficient $(k/w - N)$) in this approximation is implicit in asymptotic work by Cressie [10], while Loader [35], based on large deviation theory, gives an analogous but slightly more complicated approximation which, at least for $w = 1/2$, improves precision.

For the conditional scan, Alm [2] uses asymptotic theory to find that under the null hypothesis, given $E(N) = \lambda$,

$$\Pr(S_w \leq k) \cong \Pr(Z \leq k) \exp\{-\lambda(k + 1 - \lambda w) \times (1 - w)\Pr(Z = k)/(k + 1)\}, \quad (4)$$

where Z has a **Poisson distribution** with mean value λw.

Wallenstein et al. [52] approximate the **power** of the scan statistic against a pulse alternative for the conditional and unconditional cases. A further approximation for the conditional case yields that for relative risk $\theta \gg 1$,

$$\text{power} \cong \Pr(Z \geq k) + 2\Pr(Z = k)/(\theta - 1), \quad (5)$$

where $Z \sim \text{bin}(N, \theta w/(1 - w + \theta w))$. Based on **simulation** results, Sahu et al. [43] suggest sample sizes to achieve adequate power and compare power for triangular and rectangular pulses.

Space–Time Clustering, Moments

The scan statistic can be modified [50] to test for space–time clustering, for the case where "space" consists of g discrete geographical areas (towns, schools, cities, etc.) and there is no overall time trend. The statistic can be viewed as a variant of the Ederer–Myers–Mantel [11] statistic, where the maximum number of events within a calendar year is replaced by the maximum within any 365-day interval. The numerator of the statistic is the difference between the scan statistic for each geographical area and its expected value, and the denominator is the square root of the sum of the **variances**. When the number of geographic regions is moderately large, the **central limit theorem** indicates that the statistic has approximately a normal distribution.

Using both exact probabilities ($N < 19$) and simulation, the first two moments of the scan are tabulated [50] for six window widths and a range of Ns from 2 to 1000. Values of N not tabulated can be obtained from interpolation, or by use of the suggested linear approximations $\text{E}[S_w(N)] = wN + b_w N$, where the coefficients, b_w, are estimated from the tabulated data. Anderson and Titterington [3] extend the theoretical concept of the two-dimensional scan to geographic clustering by stretching or contracting geographic regions so that they have equal density. They display some empirical critical values and power for the spatial scan statistic, and compare it to a statistic based on squaring differences between the observed and expected densities and then integrating over the region. Kulldorff and Nagarwalla [32] describe a generalization of the scan, which is applicable for arbitrary spatial distributions and does not require specification of an a priori critical distance. The concept is extended by Kulldorff et al. [31] to detecting space–time clustering.

Seasonal Clustering

For detecting seasonal clustering (*see* **Seasonal Time Series**), data from several years are merged. The resulting 365-day period is viewed as a circle, with December 31 adjacent to January 1. The circular scan statistic, C_w, is the maximum number of events in a fraction, w, of the year. Ajne [1] finds $\Pr(C_{1/2} \geq k)$ in terms of an infinite sum, in contrast to the much simpler expression for the line [37]. He also points out that $C_{1/2}$ is the most powerful invariant test for the pulse alternative as θ approaches infinity, while Cressie [9] extends this result to general w and finds some interesting asymptotic results.

The pulse alternative differs from the sinusoidal alternative (peak followed by a trough 6 months later) for which Edwards' statistic [12] is often used as a test of seasonal clustering. The statistic $C_{1/2}$ is related to Hewitt's statistic [19] in which the monthly totals are replaced by their **ranks**, and the test statistic is the maximum over the sums of six consecutive monthly ranks. Rogerson [42] compares the statistics $C_{1/4}, C_{1/3}$, and $C_{5/12}$ with his generalizations of Hewitt's statistic based on the maximum of the sum of the ranks over 3, 4 or 5 consecutive months.

Except for special cases, the exact distribution of the circular scan statistic is very difficult to obtain since it cannot be cast in the form of a ballot problem, as the first and last "candidates" are the same. Nevertheless, the computation of $\Pr(C_w > k)$ can be reformulated so that the single probability that cannot be derived, involving an intersection of $[1/w]$ events, is very small, and approximations [51] can be obtained using methods similar to those described for the line.

Wallenstein et al. [51] propose a modification of the circular scan, termed the ratchet scan, which is applicable when only monthly totals are available and give a figure plotting ***P* values** against values of the statistic. Krauth [28] gives bounds based on the Bonferroni inequality, so that the P values need not be read off a figure.

A Generalized Scan Statistic, with Application to Assessment of Inhomogeneities in DNA Sequence Data

Glaz & Naus [16] generalize the scan statistic to the case of N independent random variables, $X_1, X_2, \ldots, X_N$, where N could be fixed or a random variable. They let $Y_m(t)$ be the sum of m consecutive random variables X_t to X_{t+m-1}, and define the scan statistic, S_m, as the maximum of $Y_m(t)$ over the integers $t = 1$ to $N - m + 1$. The special case $\{S_m = m\}$ is related to the runs test (*see* **Clustering**).

The special case where X is a **binary** random variable could be applied in the context of clustering of disease. For example, letting $X = 1$ denote the event that a birth is associated with a congenital malformation, and $X = 0$ otherwise, the statistic S_m is the maximum number of cases of congenital malformations in a series of m consecutive births. Fu & Curnow [13] show that S_m is a function of the log **likelihood ratio** for testing the hypothesis of a constant probability of disease, against the pulse alternative of a higher probability for m consecutive trials and a lower one elsewhere. Under both the null and pulse alternatives, they give a method to obtain exact probabilities, which, however, is difficult to implement for $m > 20$. Saperstein [44] and Naus [39] relate the distribution of S_m to a generalization of the birthday problem, and give results concerning the null distribution conditional on N. Chen and Glaz [8] describe and compare several approximations for the distributions of discrete scan statistic for one and two dimensions Wallenstein et al. [53] give an approximation for the power against a pulse alternative.

Recently, the generalized scan statistic has been applied to problems in **DNA sequencing** in which DNA can be viewed as a sequence of letters from a four-letter alphabet of nucleotides, a 20-letter alphabet of amino acids derived from triplets of these four nucleotides, or a three-letter alphabet of charges of amino acids. Exact or approximate probabilities for the length of the longest almost matching subsequence, or the largest net charge within any series of m consecutive amino acids, are given by Glaz & Naus [16], Sheng & Naus [46], and Karwe & Naus [26]. Asymptotic results for the distribution of the generalized scan statistic, based on methods such as the Chen–Stein method of Poisson approximation, are given by Arratia et al. [4, 5], Karlin & Macken [25], and Karlin & Brendel [24].

These results are often phrased in term of r-scans, the width of the smallest interval containing $\mathrm{r} + 1$ events, or equivalently the sum of r interarrival times. Su, Wallenstein and Bishop [48] use a compound Poisson approximation of Glaz et al. [17], as well as a modified binomial approximation to approximate the number of nonoverlapping r scans, and use the procedure for identifying gene regulatory regions. Leung and Yamashita [34] describe applications of r-scans to DNA sequence analysis focusing on identifying nonrandom clusters of palindromes. Segal and Wiemels [45] compare the scan statistic, bandwidth tests, and gap statistics for detection of translocation breakpoints.

References

[1] Ajne, B. (1968). A simple test for uniformity of a circular distribution, *Biometrika* **55**, 343–354.

[2] Alm, S.E. (1983). On the distribution of the scan statistic of a Poisson process, in *Probability and Mathematical Statistics. Essays in Honour of Carl-Gustave Esseen*, A. Gut & L. Holst, eds. Uppsala University Press, Uppsala, pp. 1–10.

[3] Anderson, N.H. & Titterington, D.M. (1997). Some methods for investigating spatial clustering, with epidemiological applications, *Journal of Royal Statistical Society* **160**, 87–105.

[4] Arratia, R., Goldstein, L. & Gordon, L. (1990). Poisson approximation and the Chen Stein method, *Statistical Science* **5**, 403–434.

[5] Arratia, R., Gordon, L. & Waterman, M.S. (1990). The Erdös Rényi law in distribution for coin tossing and sequence matching, *Annals of Statistics* **18**, 539–570.

[6] Barton, D.E. & Mallows, C.L. (1965). Some aspects of the random sequence, *Annals of Mathematical Statistics* **36**, 236–260.

[7] Berman, M. & Eagelson, G.K. (1985). A useful upper bound for the tail probabilities of the scan statistic when the sample size is large, *Journal of the American Statistical Association* **80**, 886–889.

[8] Chen, J.- & Glaz, J. (1999). Approximations for the distributions of the moments of discrete scan statistics, in *Scan Statistics and Applications*, J. Glaz & N. Balakrishnan, eds. Birkhäuser, Boston, pp. 27–66.

[9] Cressie, N. (1977). On some properties of the scan statistic on the circle and the line, *Journal of Applied Probability* **14**, 272–283.

[10] Cressie, N. (1980). The asymptotic distribution of the scan statistic under uniformity, *Annals of Probability* **8**, 828–840.

[11] Ederer, F., Myers, M.H. & Mantel, N. (1964). A statistical problem in space and time: do leukemia cases come in clusters?, *Biometrics* **20**, 626–636.

[12] Edwards, J.H. (1961). The recognition and estimation of cyclic trends, *Annals of Human Genetics* **25**, 83–86.

[13] Fu, Y. & Curnow R.N. (1990). Locating a changed segment in a sequence of Bernoulli variables, *Biometrika* **77**, 295–304.

[14] Glaz, J. (1989). Approximations and bounds for the distribution of the scan statistic, *Journal of the American Statistical Association* **84**, 560–566.

[15] Glaz, J. (1992). Approximations for tail probabilities and moments of the scan statistic, *Computational Statistics and Data Analysis* **14**, 213–227.

[16] Glaz, J. & Naus J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data, *Annals of Applied Probability* **1**, 306–318.

[17] Glaz, J., Naus, J., Roos, M. & Wallenstein, S. (1994). Poisson approximations for the distribution and moments of ordered m-spacings, *Journal of Applied Probability* **31a**, 271–281.

[18] Glaz, J., Naus, J. & Wallenstein, S. (2001). *Scan Statistics*, Springer-Verlag, New York.

[19] Hewitt, D., Milner, J., Csima, A. & Pakuyla, A. (1971). On Edwards' criterion of seasonality and a nonparametric alternative, *British Journal of Preventive and Social Medicine* **25**, 174–176.

[20] Hryhorczuk, D.O., Frateschi, L.J., Lipscomb, J.W. & Zhang, R. (1992). Use of the scan statistic to detect temporal clustering of poisonings, *Journal of Toxicology – Clinical Toxicology* **30**, 459–465.

[21] Huffer, F.W. & Lin C.T. (1997). Approximating the distribution of the scan statistic using moments of the number of clumps, *Journal of the American Statistical Association* **92**, 1466–1475.

[22] Hunter, D. (1976). An upper bound for the probability of a union, *Journal of Applied Probability* **13**, 597–603.

[23] Huntington, R. & Naus, J.I. (1975). A simpler expression for kth nearest neighbor coincidence probabilities, *Annals of Probability* **3**, 894–896.

[24] Karlin, S. & Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis, *Science* **257**, 39–49.

[25] Karlin, S. & Macken C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data, *Journal of the American Statistical Association* **86**, 27–35.

[26] Karwe, V.V. & Naus, J. (1997). New recursive methods for scan statistic probabilities, *Computational Statistics and Data Analysis* **23**, 389–402.

[27] King, W.D., Darlington, G.A., Kreiger, N. & Fehringer, G. (1993). Response of a cancer registry to reports of disease clusters, *European Journal of Cancer, Series A* **29**, 1414–1418.

[28] Krauth, J. (1991). Bounds for the linear probabilities of the linear ratchet scan statistic, in *Analyzing and Modeling Data and Knowledge*, M. Schader, ed. Springer-Verlag, Berlin, pp. 55–61.

[29] Krauth, J. (1992). Bounds for the upper tail probabilities of the circular ratchet scan statistic, *Biometrics* **48**, 1177–1185.

[30] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society, Series A* **164**, 61–72.

[31] Kulldorff, M., Athas, W.F., Feuer, E.J., Miller, B.A. & Key, C.R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, *American Journal of Public Health* **88**, 1377–1380.

[32] Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine* **14**, 799–810.

[33] Kwerl, S.M. (1975). Most stringent bounds on aggregated probabilities of partially specified dependent probability systems, *Journal of the American Statistical Association* **70**, 472–479.

[34] Leung, M.Y. & Yamashita, T.E. (1999). Applications of the scan statistic in DNA sequence analysis, in *Scan Statistics and Applications*, J. Glaz and N. Balakrishnan, eds. Birkhäuser, Boston, pp. 269–286.

[35] Loader, C. (1991). Large deviation approximations to the distribution of scan statistics, *Advances in Applied Probability* **23**, 751–771.

[36] Nagarwalla, N. (1996). A scan statistic with a variable window, *Statistics in Medicine* **15**, 845–850.

[37] Naus, J. (1965). The distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association* **60**, 532–538.

[38] Naus, J. (1966). Some probabilities, expectations, and variances for the size of the largest clusters and smallest intervals, *Journal of the American Statistical Association* **61**, 1191–1199.

[39] Naus, J. (1974). Probabilities for a generalized birthday problem, *Journal of the American Statistical Association* **69**, 810–815.

[40] Naus, J. (1982). Approximations for distributions of scan statistics, *Journal of the American Statistical Association* **77**, 177–183.

[41] Paneth, N., Kiely, M., Hegyi, T. & Hiatt, I. (1984). Investigation of a temporal cluster of congenital heart disease, *Journal of Epidemiology and Community Health* **38**, 340–344.

[42] Rogerson, P.A. (1996). A generalization of Hewitt's test for seasonality, *International Journal of Epidemiology* **25**, 644–648.

[43] Sahu, S.K., Bendel, R.B. & Sison, C.P. (1993). Effect of relative risk and cluster configuration on the power of the one dimensional scan statistic, *Statistics in Medicine* **12**, 1853–1865.

[44] Saperstein, B. (1972). The generalized birthday problem, *Journal of the American Statistical Association* **67**, 425–428.

[45] Segal, M.R. & Wiemels, J.L. (2002). Clustering of translocation breakpoints, *Journal of the American Statistical Association* **97**, 66–76.

[46] Sheng, K. & Naus J. (1994). Pattern matching between two nonaligned random sequences, *Bulletin of Mathematical Biology* **56**, 1143–1162.

[47] Su, X., Wallenstein, S. & Bishop, D. (2001). Non-overlapping clusters: approximate distributions and application to molecular biology, *Biometrics* **57**, 420–426.

[48] Wallenstein, S. (1980). A test for detection of clustering over time, *American Journal of Epidemiology* **111**, 367–372.

[49] Wallenstein, S. & Neff, N. (1987). An approximation for the distribution of the scan statistic, *Statistics in Medicine* **6**, 197–207.

[50] Wallenstein, S., Gould, M.S. & Kleinman, M. (1989). Use of the scan statistic to detect time-space clustering, *American Journal of Epidemiology* **130**, 1057–1064.

[51] Wallenstein, S., Weinberg, C.R. & Gould, M. (1989). Testing for a pulse in seasonal event data, *Biometrics* **45**, 817–830.

[52] Wallenstein, S., Naus, J. & Glaz, J. (1993). Power of the scan statistic for the detection of clustering, *Statistics in Medicine* **12**, 1829–1843.

[53] Wallenstein, S., Naus, J. & Glaz, J. (1994). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence, *Biometrika* **81**, 595–601.

[54] Weinstock, M. (1981). A generalized scan statistic for the detection of clusters, *International Journal of Epidemiology* **10**, 289–293.

Further Reading

Hoh, J. & Ott, J. (2000). Scan statistics to scan markers for susceptibility genes, *Proceedings of National Academy of Sciences* **97**, 9615–9617.

S. WALLENSTEIN

Scedasticity

Homoscedasticity is the condition in which a random variable or its observed values have the same degree of variation for all sampling units in the mathematical model or data set. By "constant variation" we usually mean a common variance for all sampling units. When that condition is not met, the random variable or data are called *heteroscedastic*. The origin of these terms is from the Greek word $\sigma\kappa\varepsilon\delta\acute{\alpha}\nu\nu\upsilon\mu\iota$, "to scatter or disperse", and the Greek words for "same" and "different".

Linear Regression Analysis

Homoscedasticity is an essential assumption in the **linear regression** model and its fit by **least squares**. For the simplest case of a single dependent variable Y and one independent (or **explanatory**) variable X, the model is

$$Y_i = \alpha + \beta X_i + e_i, \qquad i = 1, \ldots, N.$$

The e_i are random variables with $\mathrm{E}(e_i) = 0$, and if the homoscedasticity condition holds, $\mathrm{var}(e_1) = \cdots = \mathrm{var}(e_N) = \sigma^2$. If the random disturbance terms are heteroscedastic and have different variances ${\sigma_i}^2 = \mathrm{var}(e_i)$, the ordinary least squares estimators of α and β, while still unbiased, no longer have the Gauss–Markov property of minimum variance (*see* **Least Squares**). If the individual variances are known up to a proportionality constant, a weighted least squares fit of the linear model can be obtained by using the scaled data

$$\frac{Y_i}{\sigma_i} = \frac{\alpha}{\sigma_i} + \frac{\beta X_i}{\sigma_i} + \frac{e_i}{\sigma_i}.$$

The usual ordinary least squares estimation process is applied to the scaled values, with the intercept also scaled by the known standard deviations σ_i.

The weighted least squares estimators are easily displayed by the matrix form of the multiple regression model with an intercept and p independent variables:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

$\mathbf{Y}$ is the $N \times 1$ vector of observations on the dependent variable. $\mathbf{X}$ is the $N \times (p+1)$ matrix of predictor variable values of full rank $p+1$, with a first column of ones for the intercept term. $\boldsymbol{\beta}$ is the $(p+1) \times 1$ parameter vector with the intercept $\boldsymbol{\alpha}$ in the first position. The $N \times 1$ vector of random disturbances $\mathbf{e}$ has $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and diagonal covariance matrix

$$\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{e}, \mathbf{e}') = \begin{bmatrix} {\sigma_1}^2 & \cdots & 0 \\ \cdot & \cdots & \cdot \\ 0 & \cdots & \sigma_N^2 \end{bmatrix}.$$

The weighted least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y},$$

where $\boldsymbol{\Sigma}^{-1}$ is the $N \times N$ diagonal matrix, the diagonal elements of which are the reciprocals $1/\sigma_i^2$ of the corresponding elements in $\boldsymbol{\Sigma}$. Unfortunately, since the variances of the disturbance terms are rarely known, these results are mainly only of theoretical interest.

Two-Sample *t* Test

The hypothesis that two independent normal distribution means are equal frequently arises in practical data analysis. The Student–Fisher t test of that hypothesis requires that the two populations have a common unknown variance (*see* **Student's *t* Statistics**). That assumption is easily tested by the ratio $F = {s_1}^2/{s_2}^2$ of the independent sample variances. Hsu [3] calculated the true type I error probabilities for selected ratios of the population variances $\theta = {\sigma_1}^2/{\sigma_2}^2$ and sample sizes $R = N_1/N_2$ for a 0.05 level test (*see* **Level of a Test**). If θ is less than one and R is larger than one, the true type I error rate will be larger than the nominal 0.05 value. If θ and R are both larger than one, the true type I probability will be less than 0.05. These and other properties of the t test in the presence of unequal variances have been described by Scheffé [4].

The test of the equality of two means of independent normal populations is called the **Behrens–Fisher problem**, after its original investigators. Welch [5, 6] has proposed an alternative test with an approximate t distribution when the population variances are unequal (*see* **Aspin–Welch Test**).

The Analysis of Variance

Box [1, 2] has investigated the effect of heteroscedastic disturbance terms on the one- and two-way **analysis of variance** type I error rates. As in the two-sample t test, the effect of unequal variances in the

one-way layout is exacerbated by unequal sample sizes. When the sample sizes are equal, the true type I error rates are only slightly higher than the nominal 0.05 value. If the treatments with the smaller variances have larger sample sizes, the type I error rate may be much greater than 0.05. If the opposite condition holds, or if the ratios of the variances roughly follow those of the sample sizes, the true type I error rate may be smaller than the nominal 0.05. For the two-way layout with a single observation in each cell, Box showed that unequal column variances led to row test type I error rates slightly below the nominal 0.05 value, and to column test type I error rates slightly above 0.05. Further remarks on heteroscedasticity in the analysis of variance have been given by Scheffé [4].

References

[1] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification, *Annals of Mathematical Statistics* **25**, 290–302.

[2] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification, *Annals of Mathematical Statistics* **25**, 484–498.

[3] Hsu, P.L. (1938). Contribution to the theory of Student's *t* test as applied to the problem of two samples, *Statistical Research Memoirs* **2**, 1–24.

[4] Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

[5] Welch, B.L. (1937). The significance of the difference between two means when the population variances are unequal, *Biometrika* **29**, 350–362.

[6] Welch, B.L. (1947). The generalization of "Student's" problem when several population variances are involved, *Biometrika* **34**, 28–35.

DONALD F. MORRISON

Schneiderman, Marvin Arthur

Born: December 25, 1918, in Brooklyn, New York.
Died: April 1, 1997, in Bethesda, Maryland.

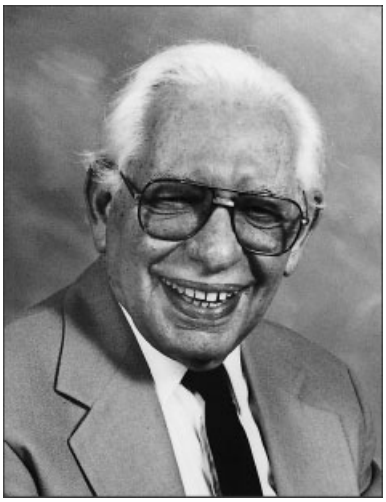

Marvin A. Schneiderman had a distinguished career as a statistician and scientific administrator at the National Cancer Institute (NCI) from 1948 to 1980. During this period, he made several notable contributions to statistical methods and applications, including the early development of the **cooperative cancer clinical trials** program at NCI, some of the first work on closed sequential boundaries (*see* **Sequential Analysis**) for **clinical trials**, the evaluation and application of the Coulter counter and related methods for quantifying peripheral blood elements, **low dose extrapolation** for the establishment of "safe" levels of exposure to potential carcinogens, and the analysis of cancer trends (*see* **Morbidity and Mortality, Changing Patterns in the Twentieth Century**). He made an even greater contribution, however, in increasing awareness of the utility of biostatistics in many areas of application, and in fostering an environment in which statisticians could thrive and become independent scientists and respected collaborators.

After graduation from New Utrecht High School in Brooklyn, Schneiderman attended City College of New York, where he received his B.S. degree in Mathematics and Statistics in 1939, graduating as a member of Phi Beta Kappa. After a short period of work at the National Container Corporation, he joined the US Census Bureau in April 1940 and served as a clerk in the 1940 census of agriculture. During this period, he also studied sampling theory and statistics with **Jerome Cornfield** and Duane Evans at the Department of Agriculture Graduate School.

From 1940 to 1944, he worked in the War Department as a quality control officer in the Office of the Quartermaster General of the Army. From June 1944 to December 1945, he served as a member of the US Army Air Corps, rising to the rank of Second Lieutenant. While in the Air Corps, he studied management sciences, economics, and statistics at the Harvard University Graduate School of Business. Following his military service, he held a civilian position as a Statistical Control Officer in the Army Air Corps until 1948. During this period (1946–1947), while stationed at Wright Field, he studied economics at the Ohio State University Graduate School.

In 1948, Jerome Cornfield left the Bureau of Labor Statistics to work at the National Cancer Institute (NCI) for **Harold Dorn**, the demographer and population expert. Cornfield hired Schneiderman as a (junior) consulting statistician in the NCI's Biometrics Section. At that time, the attitude of many laboratory research workers was "if I need statistics to show that something has had an effect, then the effect is too small to be bothered with". Schneiderman began to demonstrate the use of statistics to scientists who until that time, did not see its practical use. His work as a statistical consultant to such basic research workers as George Brecher and Fred Stohlman on problems of counting blood cell elements or in measuring the production of erythropoietin, the red blood cell stimulating hormone [20, 24], was reflected in hematologic research world-wide. He was one of the first statisticians on the editorial board of *Blood*. He was one of many statisticians who made it respectable for scientists to consult statisticians. Work with Walter Heston and Michael Shimkin led to models of **dose–response** that formed the basis for the Environmental Protection Agency's (EPA's) regulation of toxic materials. While in the Biometry Branch, he received the M.S. in Statistics from American University in 1953.

Statisticians count things. As a counting specialist, Schneiderman was involved in the evaluation and

propagation of a simple device to count platelets [3], the blood element involved in blood clotting. The visual counting of blood cells by a laboratory technician eventually gave way to the Coulter counter, a device for measuring the flow of electric current across a narrow aperture while a blood sample was passed across the aperture [4]. He also worked in other areas such as industrial carcinogenesis [6] and drug toxicity.

In 1954, C. Gordon Zubrod joined the NCI. Under his leadership and with Schneiderman as chief statistician, plans were begun for a national program of cooperative clinical trials in cancer. The US Congress at this time created a Cancer Chemotherapy National Service Center (CCNSC), [11], which served as the major organizational unit for the development and conduct of the first randomized clinical trials in cancer. The clinical trials program at NCI enlarged rapidly from 1955 to 1960, requiring statisticians for each of the cooperative groups that were established. Part of Schneiderman's responsibility included recruitment of statisticians all across the country. He was able to bring to these trials Irwin Bross, **Bernard Greenberg**, and James Grizzle (North Carolina); Will Dixon (UCLA); **Donald Mainland** (NYU); and several others. The North Carolina connection was especially productive. Several graduates of Greenberg's program, including Edmund Gehan, joined the NCI group.

From 1955 to 1959, Schneiderman served as Acting Chief of the NCI Therapeutic Trials Section and coauthored the paper reporting the first randomized control trial in cancer research [10] and similar subsequent studies [28]. In cooperation with a young visiting scientist at the NCI, Peter Armitage, a multistage sequential scheme was developed for screening candidate materials using mice as a biological model. This was the first of the so-called "rational" **screening** programs and could be modified to produce a minimum of **false negatives** (while producing some **false positives** – which would be discarded when tested in humans), or a minimum of false positives – so that the very ill patients on whom these drugs were tested would be most likely to receive active, effective agents [1].

From 1959 to 1960, Schneiderman attended the London School of Hygiene and Tropical Medicine under a Rockefeller Public Service Award, continuing his work with Peter Armitage on epidemiology, medical research involving humans, controlled trials, and biostatistics [19]. During his military service in World War II, he had come upon the work of **Abraham Wald** and his student Milton Sobel on sequential testing. Their aim was to minimize destructive testing of military supplies (often ammunition) while attempting to insure the quality and potential effectiveness of the batches that "passed". Recognizing a related problem in humans with strong ethical overtones, Schneiderman and his colleagues wished to test, in a controlled way, potential anticancer drugs, while minimizing the number of patients exposed to the less effective drug when making a comparison of two drugs. They considered it unethical to delay ending a trial long after enough evidence had been developed to convince that one treatment was superior to another. The Wald–Sobel schemes provided a starting place for such trials – with the drawback that the potential existed for quite long trials if the evidence indicated that no firm decision had yet been reached. What was needed were closed sequential designs. I.D.J. Bross developed a small number of closed sequential designs. Armitage modified the Wald–Sobel designs to give an upper limit to the number of participants in a trial. Under the mentorship of Armitage at the London School of Hygiene, Schneiderman developed a family of sequential schemes that encompassed the Armitage designs at one extreme and the Wald designs at the other. Theodore Colton, as a summer fellow at NCI soon thereafter, developed another scheme for the early termination of a comparative trial, as part of a continuing treatment-development program, which are sometimes referred to as "horizon" trials.

Returning from London, Schneiderman received his Ph.D. in Statistics from American University in 1961 and served as Associate Chief of the Biometry Branch under William Haenszel until 1970. There he worked with many physicians and statisticians on a wide range of topics [5, 7, 13–16, 21, 26, 27]. The Biometry Branch at one time or another included Marvin Zelen, Sally Fand, Nancy Brombacher, Polly Feigl, and Emanuel Landau.

The positive effects of having MD-statisticians working on controlled trials led to recruiting additional MDs. These included John Bailar (who in turn recruited **David Byar**), Sylvan Green, David Levin, Robert Huse, Mitchell Gail, and Elia Kazam. The use of these physician–statisticians epitomized the structure and process whereby research in the

statistics-related fields proceeded at NCI. The process involved recruiting very bright young people, having them work for a short time as juniors and then giving them their heads – with a minimum of supervision. When they worked as collaborators or consultants, Schneiderman was insistent that they be included as coauthors in the publications that followed. Publication was extremely important for advancement at the **National Institutes of Health** (NIH).

In 1970, Schneiderman was appointed to head NCI's Field Studies and Statistics Program, which consisted of three branches: the Biometry Branch with William Haenszel as Chief, the Environmental Epidemiology Branch with Joseph Fraumeni, Jr. as Chief, and the Clinical Epidemiology Branch with Robert Miller as Chief. Schneiderman's philosophy was to hire good people and to give them the freedom and authority to succeed. He provided resources and support for the Surveillance, Epidemiology, and End Results (SEER) Program that grew out of the Third National Cancer Survey, under the guidance of William Haenszel, John Bailar, and **Sidney Cutler**. He needed to defend this program vigorously, as it appeared to be quite expensive to those accustomed to funding laboratory research.

Schneiderman encouraged the extension of statistical methods to many areas and he continued to write in a variety of areas [2, 9, 12, 17, 18, 22, 23, 25]. He was among the first to support the work of a young Commissioned Officer, Fred Burbank, who pioneered the use of computer-generated maps and **regression** methods to study US cancer rates. He supported work on **risk assessment**, including statistical methods for analyzing animal studies by Drs John Gart and Robert Tarone, and methods for low dose extrapolation and interpretation by Dr Charles Brown and Mr Nathan Mantel. He supported methodological work and "hands on" experience for statisticians such as David Byar and Mitchell Gail on clinical trials and on the **evaluation of diagnostic tests**. He helped to develop the randomized trial to evaluate the usefulness of mammography screening to reduce breast cancer mortality (the "HIP" study; *see* **Screening Trials**). He supported epidemiologic work leading to pathbreaking findings in cancer genetics (for example, the Li–Fraumeni syndrome; *see* **Genetic Epidemiology**) and **environmental epidemiology**.

From 1978 until he retired from the NIH in 1980, Schneiderman served as Associate Director for Science Policy in the Office of the Director of NCI. There, he continued to advocate the wide application and support of statistical methods, and was a vigorous spokesman for this cause among scientific administrators and decision-makers, on numerous advisory panels, and in the media.

After leaving the NCI, Schneiderman joined the Environmental Law Institute, where he helped to develop environmental legislation and prepared several papers with Devra Davis, an environmental activist [8]. After a brief excursion into the work of an environmental consulting firm, he followed Dr Davis into a position with the National Academy of Sciences, National Research Council Board on Environmental Studies and Toxicology. He had in the past been a member of many NAS/NRC committees. While on one of the committees, and again as a staff member, his emphasis was on how a statistician viewed the problem. He continued to be active until his death in 1997.

Perhaps Marvin Schneiderman's view of his professional life can best be summarized in the following, which he wrote just before his death:

> What pleased me most in my 50 or so years as a statistician – with its occasional side excursions into epidemiology – was that it was fun. It was exciting. It seemed to me that it had a positive impact on this country's health and environment. First my work brought me face-to-face with some of mankind's most important and intractable problems – problems of disease and problems of life and death on which I believe I had a positive effect. Second it enabled me to work with persons exploding with intellectual fire and originality. Some were other statisticians and some were the laboratory research workers for whom I was a consultant and at times a collaborator. Third it enabled me to merge my own concepts of appropriate human behavior with scientific needs so that I could participate in ethical research on humans – and make it possible (or perhaps imperative) that those who followed me would treat their patients who happened to be seriously ill as humans and not only as research subjects. Finally it brought into my sphere brilliant young men and women whose lives and professional careers I influenced, and – I hope – enlarged.

References

[1] Armitage, P. & Schneiderman, M.A. (1958). Statistical problems in a mass screening program, *Annals of the New York Academy of Science* **76**, 896–908.

[2] Blokhin, H.N. & Schneiderman, M.A., eds (1979). *Epidemiology of Cancer in the USSR and USA*. Meditsina, Moscow (in Russian).

[3] Brecher, G., Schneiderman, M.A. & Cronkite, E.P. (1953). The reproducibility and constancy of platelet counts, *American Journal of Clinical Pathology* **23**, 15–26.

[4] Brecher, G., Schneiderman, M.A. & Williams, G.Z. (1956). Evaluation of an electronic red blood cell counter, *American Journal of Clinical Pathology* **26**, 1439–1449.

[5] Carbone, P.P., Spurr, C., Schneiderman, M.A., Scotto, J., Holland, J.F. & Shnider, B. (1968). Management of patients with malignant lymphoma: a comparative study with cyclophosphamide and vinca alkaloids, *Cancer Research* **28**, 811–822.

[6] Cutler, S.J., Schneiderman, M.A. & Greenhouse, S.W. (1954). Some statistical considerations in the study of cancer in industry, *American Journal of Public Health* **44**, 1159–1166.

[7] Cutler, S.J., Greenhouse, S.W., Cornfield, J. & Schneiderman, M.A. (1966). The role of hypothesis testing in clinical trials, *Journal of Chronic Diseases* **19**, 857–882.

[8] Davis, D.L., Bridbord, K. & Schneiderman, M.A. (1982). Cancer prevention: assessing cause, exposure, and recent trends in mortality for U.S. males, 1968–1978, *Teratogenesis, Carcinogenesis and Mutagenesis* **2**, 105–135.

[9] Devesa, S.S. & Schneiderman, M.A. (1977). Increase in the number of cancer deaths in the United States, *American Journal of Epidemiology* **106**, 1–5.

[10] Frei, E. III, Holland, J.F., Schneiderman, M.A., Pinkel, D., Selkirk, G., Freireich, E.J., Silver, R.T., Gold, G.L. & Regelson, W.A. (1958). A comparative study of two regimens of combination chemotherapy in acute leukemia, *Blood* **13**, 1126–1148.

[11] Gehan, E.A. & Schneiderman, M.A. (1990). Historical and methodological developments in clinical trials at the National Cancer Institute, *Statistics in Medicine* **9**, 871–880.

[12] Hoel, D.G., Gaylor, D.W., Kirschstein, R.L., Saffiotti, U. & Schneiderman, M.A. (1975). Estimation of risks of irreversible, delayed toxicity. *Journal of Toxicology and Environmental Health* **1**, 133–151.

[13] Schneiderman, M.A. (1961). Controlled clinical trials: Monday's count-down for Tuesday's launching, *Journal of New Drugs* **1**, 250–255.

[14] Schneiderman, M.A. (1963). Is it really bad? A proposal for the toxicity-testing of drugs, *Journal of the Society of Cosmetic Chemistry* **14**, 227–232.

[15] Schneiderman, M.A. (1964). The proper size of a clinical trial: "Grandma's strudel" method, *Journal of New Drugs* **4**, 3–11.

[16] Schneiderman, M.A., (1969). Quantitative thinking in medicine – biostatistics (using numbers to mark the route from cause to effect and back), in *Traumatic Medicine and Surgery for the Attorney*, P. Cantor, ed. Matthew Bender, New York, pp. 419–477.

[17] Schneiderman, M.A. (1978). Environmental factors and cancer prevention, in *Third National Symposium on Detection and Prevention of Cancer, New York (April 26–30, 1976)*. Marcel Dekker, New York.

[18] Schneiderman, M.A. (1978). Legislative possibilities to reduce the impact of cancer, *Preventive Medicine* **7**, 424–438.

[19] Schneiderman, M.A. & Armitage, P. (1962). A family of closed sequential procedures, *Biometrika* **49**, 41–56.

[20] Schneiderman, M. & Brecher, G. (1950). The relative frequency of sparse cell elements – an application to reticulocyte blood counts, *Biometrics* **6**, 390–394.

[21] Schneiderman, M.A. & Levin, D.L. (1972). Trends in lung cancer: mortality, incidence, diagnosis, treatment, smoking and urbanization, *Cancer* **30**, 1320–1325.

[22] Schneiderman, M.A. & Mantel, N. (1973). The Delaney clause and a scheme for rewarding good experimentation, *Preventive Medicine* **2**, 165–170.

[23] Schneiderman, M.A., DeCoufle, P. & Brown, C.C. (1979). Thresholds for environmental cancer: biologic and statistical considerations, *Annals of the New York Academy of Science* **329**, 92–130.

[24] Schneiderman, M.A., Mantel, N. & Brecher, G. (1951). The effect of rejection procedures on the accuracy of blood counts, *American Journal of Clinical Pathology* **21**, 973–978.

[25] Schneiderman, M.A., Mantel, N. & Brown, C.C. (1975). From mouse to man – or how to get from the laboratory to Park Avenue and 59th Street, *Annals of the New York Academy of Science* **246**, 237–248.

[26] Schneiderman, M.A., Myers, M.H., Sathe, Y.S. & Koffsky, P. (1964). Toxicity, the therapeutic index, and the ranking of drugs, *Science* **144**, 1212–1214.

[27] Scotto, J. & Schneiderman, M.A. (1972). Predicting survival in terminal cancer, *British Medical Journal* **4**, 50.

[28] Zubrod, C.G., Schneiderman, M., Frei, E. III, Brindley, C., Gold, G.L., Shnider, B., Oviedo, R., Gorman, J., Jones, R. Jr., Jonsson, U., Colsky, J., Chalmers, T., Ferguson, B., Dederick, M., Holland, J., Selawry, O., Regelson, W., Lasagna, L. & Owens, A.H. Jr. (1960). Appraisal of methods for the study of chemotherapy of cancer in man: comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide, *Journal of Chronic Diseases* **11**, 7–33.

DAVID L. LEVIN

Scientific Method and Statistics

The word "statistics" has multiple meanings. For example, it can refer to numeric summaries obtained from a body of data, and it can also refer to methods for analyzing data. This article is chiefly concerned with the second of these two meanings, and in particular with how these methods relate to the scientific enterprise in general. To explore this relationship, we begin by examining the idea of "scientific method", and later see how statistical methods fit in. We shall see that it is no accident that statistics has been described as "the science of doing science" [20].

The notion of scientific method has changed over time but, like the models used within science, models of the way in which scientists work are idealizations. They skip such things as serendipity and focus on the pattern of behavior. Some models are descriptive, seeking to show how scientific advance occurs in practice, while others are proscriptive, seeking to show how things should be done. Early views – a perspective variously attributed to William of Ockham, John Herschel, and John Stuart Mill – saw the growth of (scientific) knowledge as an essentially inductive process. That is, they saw science as a process of accumulating facts by observation, classifying these facts according to observed regularities, and hence generalizing them into scientific laws. This view was formalized in various ways. For example, Mill's *Canon of Agreement* conveys ideas familiar to all statisticians: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur have every circumstance in common save one, that one occurring only in the former, the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause of the phenomenon." [13, Book III, Chapter 8].

Although experimentation featured in these views, the emphasis put on its central and fundamental role is normally attributed to Sir Francis Bacon [8]. The twin pillars of observation and experiment that Bacon emphasized presented a practical alternative to the notion that truth lay in ancient authority – whose conclusions were tacitly assumed to be based on thought rather than observation. At the time, Bacon's approach was even termed "the experimental philosophy" to contrast it with theorizing.

These views regard observation and experiment as essentially *confirmatory* exercises: that is, each supporting result is regarded as adding weight to the hypothesis under investigation. From this perspective, science generalizes or abstracts from many particular observations. We note how long an iron ball takes to strike the ground when released from various heights and hence formulate a general *law* which fits the observations very well: that is, it is a process of induction (different from *mathematical induction*, which is a more formal procedure), from the particular to the general.

Inductivism is all very well, but it misses something important. It leads to a *description* of phenomena, not an *explanation*. Moreover, as **Popper** [16, 17] pointed out, one cannot in fact logically *prove* anything by induction: the mere fact that all the swans you may have seen are white does not imply that all are white (in fact, black swans were discovered in Australia at a time when all known swans were white). On the other hand, one can *disprove* a universal statement by example: observation of a black swan disproves the statement that all swans are white. Thus theories, in general, can be disproven, but not proven. Scientific method, Popper argued, thus seeks the *falsification*, not the *verification*, of scientific theories. One successively formulates hypotheses or theories and tests their predictions against experimental observations. When an hypothesis fails a test, it is modified, refined, or replaced. Advance occurs because a new hypothesis must pass not only those tests passed by its predecessor, but also tests that the predecessor fails. This means that very radical revisions become more difficult as time passes, because of the mass of accumulated evidence. A key aspect of this Popperian view is the alternation of theory and observation: theories are postulated and experiments/observations test them. Like the chicken and the egg, neither is pre-eminent.

Note that this Popperian view of how science does and should work is, as we have remarked above, like all scientific models, an idealization. In reality, given a conflict between theory and data, we do not necessarily reject the theory. Data may be subject to error. The lack of a match may simply be due to error in measurement, not because of a faulty theory. Thomas Kuhn [10, p. 81] gives a nice example of this:

> ... during the sixty years after Newton's original computation, the predicted motion of the moon's perigee remained only half of that observed. As Europe's best mathematical physicists continued to wrestle unsuccessfully with the well-known discrepancy, there were occasional proposals for a modification of Newton's inverse square law. But no one took these proposals very seriously, and in practice this patience with a major anomaly proved justified. Clairaut in 1750 was able to show that only the mathematics of the application had been wrong and that Newtonian theory could stand as before.

If scientific theories themselves become more difficult to replace by radical alternatives as time passes (because the theories have already passed more tests), so also do theories of science. So, Kuhn's notion of *paradigm shifts*, in which the scientific constructs are jettisoned in favor of a complete reformulation, is really a refinement of Popper's notion of refutations. Kuhn [10, p. 10] defines "normal science" as "research, firmly based upon one or more past scientific achievements, achievements that some particular scientific community acknowledges for a time as supplying the foundation for its further practice". Paradigms are [10, p. viii] "universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners". A paradigm shift describes the process of moving from one paradigm to another. Kuhn has emphasized that most scientists do not find such shifts, but instead are concerned with normal science, working within a dominant paradigm.

Statistics relates to all this in various ways. Statistical methods assist in the precise formulation and selection of theories or models, the quantification of errors, and in examining the match between the theory and the data. Formulation of models includes such things as **variable selection** and parameter **estimation**. Implicit within this are comparisons between alternative models. Both falsification and accumulation of supportive evidence occur through the blurring spectacles of the necessary simplifications in the modeling process. It is in helping to control and remove this blurring or error that statistics plays one of its key roles in science. Finally, the match between theory and data may be examined informally, perhaps by graphical methods such as **residual** plots, or more formally, perhaps via significance tests (*see* **Hypothesis Testing**). Again, for the reasons mentioned above, we cannot expect perfect matches or perfect mismatches – and again statistical methods enable us to judge the quality of a match.

Perspectives on the role and purpose of statistical methods seem to parallel the development of ideas about how science progresses. For example, early statisticians such as **Karl Pearson** adopted an inductivist perspective: "The classification of facts and the formation of absolute judgments upon the basis of this classification – judgments independent of the idiosyncrasies of the individual – essentially sum up the *aim and method of modern science*" ([14, p. 11], his italics). Modern **exploratory data analysis**, techniques aimed at examining data in order to discern patterns and structure, without specifying too closely exactly what sort of patterns are sought, might naturally be thought of as an inductive process. Moving on, however, we find inductivism combined with the hypothetico-deductive strategy. **Fisher** [4], for example, described how statistical analysis is essentially what we would regard as the Popperian view of scientific method writ small: "The statistical examination of a body of data is thus logically similar to the general alternation of inductive and deductive methods throughout the sciences. A hypothesis is conceived and defined with all necessary exactitude; its logical consequences are compared with the available observations; if these are completely in accord with the deductions, the hypothesis is justified, at least until fresh and more stringent observations are available" [4, 4th Ed., p. 9]. Similarly, Box [2, p. 383] describes science as proceeding in what is essentially a Baconian/Popperian way and statistics as playing a central role in this:

> It seems that scientific knowledge advances by a practice–theory iteration. Known facts (data) suggest a tentative theory or model, implicit or explicit, which in turn suggests a particular examination and analysis of data and/or the need to acquire further data; analysis may then suggest a modified model that may require further practical illumination and so on .... New knowledge thus evolves by an interplay between *dual* processes of induction and deduction in which the model is not fixed but is continually developing .... The statistician's role is to assist this evolution .... In doing so he [*sic*] employs two inferential devices: *Criticism* and *Estimation*.

The Bayesian strategy (*see* **Bayesian Methods**) may also be seen as fitting the Popperian framework. Essentially, the Bayesian approach takes prior beliefs and modifies them in the light of the data. This

modification may be rather more subtle than a crude hypothesis test, in that a distribution of strength of posterior belief results, so that an initial theory is not "rejected" or "refuted", but the basic idea is the same.

If the statistical model fitting process provides a parallel to the scientific process, it can also shed light on that process. One of the driving forces in selecting between scientific theories is simplicity. From a set of explanations, each equally effective at explaining a collection of facts, the most simple is to be preferred. This principle goes under various names, such as *Ockham's razor* or the *principle of* **parsimony**. In science, a simple theory is sometimes said to be *elegant*, and to be a preferred explanation because of its elegance. In statistical model building (*see* **Model, Choice of**), a similar preference is accorded to simpler models. For example, a set of ten different points in a plane could be modeled by an infinite number of curves that pass through all the points. However, if one of these curves is a straight line, then that, being the simplest, will often be the preferred model (unless, for example, there is some extra aspect to the theory, leading to some more complicated curve being preferred). Some approaches to statistical inference, such as the *Minimum Message Length* [22] and *Minimum Description Length* approaches [18, 19] are based directly on this principle. They essentially provide a formalism for combining measures of the complexity of a theory and the complexity of the data (the facts) in terms of that theory. Current work on computational learning theory and machine learning [9, 21] is directly concerned with these issues of how well models, based on a finite set of observations *(a training set)* will generalize to other data.

Statistical models come in (at least) two types: *mechanistic* and *descriptive*. The former are based on some underlying theory or mechanism which purports to explain the observed phenomena. The latter seek merely to summarize the data in a convenient way. One can argue that the term "model" should be restricted to the former, but common usage applies it to both. Descriptive models have a role in theory formulation, and also in pragmatic situations. For example, a **prediction** rule may be based on empiric observation of relationships between variables, without there being any underlying theory or explanation for why those relationships should exist. Bacon distinguishes models that give light from those that bear fruit (see [8, p. ix]), and this seems very close to our mechanistic and descriptive distinction.

If mechanistic models aim to represent some underlying process, they might be regarded as "true" or "false" according as they are or are not faithful representations of that process. In both science and statistics, this view is no longer generally held. For example, in discussing this issue in statistics, Durbin [3] says that "Undoubtedly most applied workers have always been aware that any statistical model is at best an approximation to reality. There is in real life no such thing as a 'true model' ". He goes on to criticize debates about the different schools of statistical **inference** because of this misconception:

> . . . much of the discussion of the foundations of statistical inference that has taken place over the past half century has been predicated on the assumption that the model is "true". The alternative formulations of the inference problems that have been considered relate mainly to the properties of models and there has been too little discussion of the interaction with the underlying statistical reality. Statements about parameter values have been discussed as if parameters had a clearly-defined tangible existence, whereas in most cases they are at best mathematical artifacts introduced only in order to provide the most useful approximation available to the behavior of the underlying reality. It is all too easy to lose sight of the fact that the real purpose of the analysis is to make statements about this reality rather than about the models that approximate it.

Durbin [3] continues:

> Of course I appreciate that a standard procedure in science is to postulate a model and make inferences about the behavior of the phenomena under study on the assumption that the model is an accurate one. My point is that because of the manifest imperfections of many statistical models as descriptions of the reality under investigation, this process has been carried a bit too far. The obsessional desire to make "best possible" statements about parameter values in artificially small models has been over-indulged to an extent that seems out of proportion to the true interests of users of statistical models.

Just as other sciences are not static, but progress as new theories and new discoveries require their modification, so statistics, as science or technology [6], is not static. A glance back over recent decades shows new problems, new methods, and new ideas changing the way statistical methods are thought about

and applied (**survival analysis**, **generalized linear models**, and **multidimensional scaling** provide just a few clear examples). Much of this development, in recent decades, has been the consequence of the growth in readily available computational resources. This has had the further consequence that modern statistics might more naturally be thought of as a computational science than as a mathematical science. And, of course, these developments continue. Examples of current new perspectives stimulating the development of new methods are not hard to find. One is the impact of feedforward **neural networks**, originally introduced as complex structures of individually simple interacting components, but now seen as a highly parameterized and very flexible models for function estimation. Another is the growing interest in issues of multiple data sets, in contrast to the earlier situation, in which the focus was on the single data set to hand (see, example, [7]). **Meta-analysis** is one manifestation of this in the statistical literature. A third, also arising as a consequence of computational advances, is a concern with very large data sets, perhaps with many millions or even billions of observations; here significance tests tend to lose their relevance.

While clearly (one would hope!) statistics has had a positive impact on the development of science via its methods of data collection and analysis, it has also had a more subtle, and not necessarily beneficial impact. The theories of scientific method outlined above say nothing of the social and cultural environment in which the scientist works. In particular, modern scientists communicate their results via published papers and there is evidence that papers have sometimes been accepted for publication with a less than rigorous attitude to the statistical methods employed. Misapplication of statistical methods in medical research has been examined by several authors (see, for example, [1, 5, 11, 12, 15], and [23]) – the depressing results being perhaps a partial consequence of the fact that doctors are primarily trained in the technology of medicine, rather than the principles of science. From another perspective, sometimes in research standards are imposed and practices adopted which most statisticians would regard as dubious (the classic example is the tendency to require significance tests – and then to favor papers that show significant results in those tests). We have already commented about how the computer has changed, and continues to change, the face of statistical practice. In the present context, we should also add the remark that electronic communication, in the form of the **Internet**, is already beginning to change scientific practice.

References

[1] Altman, D.G. (1994). The scandal of poor medical research, *British Medical Journal* **308**, 283–284.

[2] Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness, *Journal of the Royal Statistical Society, Series A* **143**, 383–430.

[3] Durbin, J. (1987). Statistics and statistical science, *Journal of the Royal Statistical Society, Series A* **150**, 177–191.

[4] Fisher, R.A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[5] Glantz, S.A. (1980). How to detect, correct and prevent errors in the medical literature, *Circulation* **61**, 1–7.

[6] Healy, M.J.R. (1978). Is statistics a science?, *Journal of the Royal Statistical Society, Series A* **141**, 385–393.

[7] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-analysis*. Academic Press, Orlando.

[8] Jones, R.F. (1961). *Ancients and Moderns: a Study of the Rise of the Scientific Movement in Seventeenth Century England*. Dover, New York.

[9] Kearns, M.J. & Vazirani, U.V. (1994). *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Mass.

[10] Kuhn, T.S. (1970). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

[11] MacArthur, R.D. & Jackson, G.G. (1984). An evaluation of the use of statistical methodology in the *Journal of Infectious Diseases*, *Journal of Infectious Diseases*, **149**, 349–354.

[12] McGuigan, S.M. (1995). The use of statistics in the *British Journal of Psychiatry*, *British Journal of Psychiatry*, **167**, 683–688.

[13] Mill, J.S. (1879). *A System of Logic*. Longmans Green, London.

[14] Pearson, K. (1892). *The Grammar of Science*, 1937 Ed. Dent, London.

[15] Pocock, S.J., Hughes, M.D. & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials: a survey of three medical journals, *New England Journal of Medicine* **317**, 426–432.

[16] Popper, K. (1959). *Logic of Scientific Discovery*. Hutchinson, London.

[17] Popper, K. (1963). *Conjectures and Refutations*. Routledge and Kegan Paul, London.

[18] Rissanen, J. (1987). Stochastic complexity, *Journal of the Royal Statistical Society, Series B* **49**, 223–239.

[19] Rissanen, J. (1989). *Stochastic Complexity in Statistical Enquiry*. World Scientific, Singapore.

[20] Smith, A.F.M. (1996). Mad cows and ecstasy: chance and choice in an evidence-based society, *Journal of the Royal Statistical Society, Series A* **159**, 367–383.
[21] Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.
[22] Wallace, C.S. & Freeman, P.R. (1987). Estimation and inference by compact coding, *Journal of the Royal Statistical Society, Series B* **49**, 240–252.
[23] White, S.J. (1979). Statistical errors in papers in the *British Journal of Psychiatry*, *British Journal of Psychiatry* **135**, 336–342.

(*See also* **Statistics, Overview**)

DAVID J. HAND

Scores

We primarily consider the situation where both the dependent (or **response**) variable and independent (**explanatory** or predictor) variable are **categorical**. When either the independent or dependent variable is ordered, or when both variables are ordered, use of the standard χ^2 test is inappropriate because it does not exploit the ordering. To increase the power of tests when the alternative is related to the ordering, we assign numbers, called scores, to the ordered categories. The scores should reflect the scientific meaning of the categories.

Predictor Variable Scores

Scores for the predictor variable (if it is ordered) arise from the underlying process, i.e. from the model generating the data. For example, Graubard & Korn [14] considered the use of midpoints of the category boundaries, midranks, and equally spaced scores for the predictor variable (alcohol consumption in mothers) in the analysis of the data in Table 1 on occurrence of congenital malformation in offspring. The midpoint scores for these data are 0.0, 0.5, 1.5, 4.0, and 7.0. Midranks, or average category **ranks**, may be computed by ranking the observations as if they were ungrouped, and then taking the mean of the ranks within each category. Suppose there are K categories with N_i observations in category $i, i = 1, \ldots, K$. Then the midrank of the ith category is $\sum_{j=1}^{i-1} N_j + (N_i + 1)/2$. For the data in Table 1, the midranks are 8557.5, 24 365.5, 32 013.0, 32 473.0, and 32 555.5. The equally spaced scores are 1.0, 2.0, 3.0, 4.0, and 5.0. The standardized scores, i.e scores linearly transformed so that they have zero mean and unit variance, are plotted in Figure 1. The closeness of the midrank scores for the three heaviest drinking categories is inappropriate from the underlying science which suggests greater relative effect with higher levels of alcohol consumption. Hence midrank scores should not be used to analyze the data [14]. Indeed, the one-sided significance levels varied considerably according to the scores used (midpoints, $P = 0.02$; midranks, $P = 0.29$; equally spaced, $P = 0.10$). In general, one should assign reasonable scores based on the substantive meaning of the categories [1, 3, 8, 14].

If there are two or more sets of plausible scores, then a single procedure that combines the tests based on each set seems appropriate. One such procedure is the maximin efficiency robust test (MERT) [12, 23]. For example, the standardized scores for the MERT based on the midpoints, midranks, and equally spaced scores for the above data are also plotted in Figure 1. The combination property of the MERT is clearly visible (and its associated P value is 0.05).

Table 1 Occurrence of congenital sex organ malformation categorized by alcohol consumption of the mother [14]

	Alcoholconsumption (average numberof drinks per day)				
Malformation	0	<1	1–2	3–5	≥6
Absent	17 066	14 464	788	126	37
Present	48	38	5	1	1
Total	17 114	14 502	793	127	38

Response Variable Scores

Response variables can be ordered quantitatively and qualitatively. The scale of measurement of these variables (*see* **Measurement Scale**) is also referred to as interval and ordinal, respectively [1, 21]. Quantitatively ordered variables arise when an underlying continuous variable, e.g. blood pressure, is grouped into ordered categories. Data that are ordered but

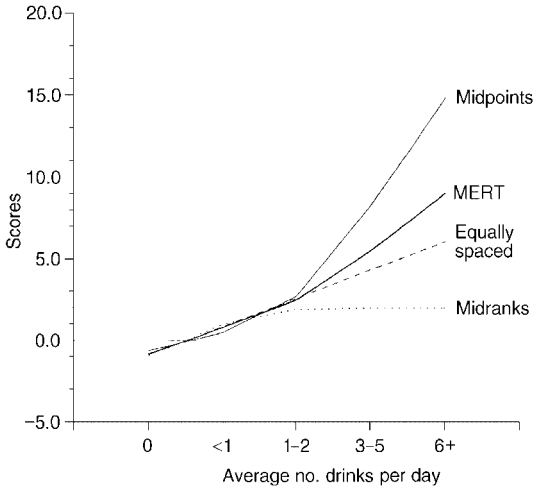

Figure 1 Standardized scores for several scoring systems for the Graubard & Korn [14] data (Source: [23])

without an apparent underlying continuous measurement scale are qualitatively ordered. An example is patient response to choices: "How do you feel: poor, fair, good, or excellent?" (*see* **Ordered Categorical Data**).

Quantitatively Ordered Variables

Quantitatively ordered variables arise from data typically reported with category boundaries such that N_i observations are in the interval $[b_{i-1}, b_i)$, $i = 1, \ldots, K$. We first discuss the assignment of optimal scores if the investigator has some knowledge of the underlying distribution of the data. We then discuss the perhaps more usual case where less formal assignment methods are employed. Optimal category scores are derived from general rank test theory for grouped data [10, 11, 25, 27]. Consider the linear rank statistic $S = \sum_{l=1}^{N} c_l a_l$, where the c_l are regression constants and the a_l are scores. For example, in the two-sample problem with continuous data, $c_l = 1$ when the lth order statistic is a Y and $c_l = 0$ otherwise. The scores are derived from $a_l = J[l/(N+1)]$, where $J(\cdot)$ is a score function that may be selected by the investigator. In particular, if we know the underlying distribution F (with density f), and we want to test for a shift in location, then the optimal score function [7, 15, 24] is $J(u) = -f'[F^{-1}(u)]/f[F^{-1}(u)]$. For grouped data the assigned scores are the average of the rank test scores in each category. Suppose there is grouping of the N outcome observations into K categories with N_i observations in category i. Then we may estimate the ith fractile, u_i, by $\sum_{j=1}^{i} N_j/N$, letting $u_0 = 0$, and we estimate $F^{-1}(u_i)$ by b_i. We obtain the estimated scores a_i, $i = 0, 1, 2, \ldots, K-1$, from

$$a_i = \frac{f(F^{-1}(u_i)) - f(F^{-1}(u_{i+1}))}{u_{i+1} - u_i} = \frac{1}{u_{i+1} - u_i} \int_{u_i}^{u_{i+1}} J(u)\, \mathrm{d}u. \quad (1)$$

Note that Eq. (1) corresponds to the mean of the scores that would be assigned to the individual observations in the interval $[b_i, b_{i+1})$ if we had them.

Table 2 presents ELISA absorbance ratios in AIDS patients and healthy blood donors [5, 13, 29] and the estimated optimal **Wilcoxon–Mann–Whitney**, $J(u) = 2u - 1$, and inverse Savage (or exponential or **logrank**), $J(u) = -1 - \ln(1-u)$, scores. To illustrate the use of formula (1) we calculate the first Wilcoxon score, a_0, in Table 2: $u_0 = 0$, $u_1 = 202/385 = 0.5247$, the value of the integral is $u(u-1)|_0^{0.5247} = -0.2494$, and therefore $a_0 = -0.475$. Both sets of scores yield highly significant results ($Z = 15.1$ using Wilcoxon scores and $Z = 16.0$ using inverse Savage scores).

The above procedure generates asymptotically **most powerful rank tests**. For small samples, one might use scores derived from locally most powerful tests (LMPRT) [16, 24]. Now the score given to the observations in the ith group is the mean of the LMPRT scores if the individual observations were available.

Nevertheless, some investigators may prefer to assign scores based on less formal methods. For example, as in the discussion above for ordered predictor variables, midpoints of the category boundaries or equally spaced scores may be assigned. If the investigator has some idea of the relative distances between categories, which are perhaps only loosely

Table 2 Distribution of ELISA absorbance ratios in healthy blood donors and AIDS patients [5, 13, 29] and the estimated optimal Wilcoxon and inverse Savage scores

	Absorbanceratio							
	<2	2–2.99	3–3.99	4–4.99	5–5.99	6–11.99	≥12	Total
AIDS patients	0	2	7	7	15	36	21	88
Healthy blood donors	202	73	15	3	2	2	0	297
Total	202	75	22	10	17	38	21	385
Wilcoxon scores	−0.475	0.244	0.496	0.579	0.649	0.792	0.945	
Inverse Savage scores	−0.674	−0.016	0.381	0.559	0.744	1.30	2.91	

dependent on the actual scale of measurement, then he or she may be able to assign reasonable scores [2]. These are valid procedures, if scores are assigned prior to examining the data [8]. That is to say, one must avoid choosing scores to obtain a desired result. If more than one scoring procedure seems plausible, then a combination procedure, such as the MERT mentioned above, is a reasonable approach.

Qualitatively Ordered Variables

Regardless of the form of the underlying distribution, when the data are continuous, a measure of the difference between two distributions F_X and F_Y is $\Pr(X < Y)$. This is the probability that a randomly selected X is less than a randomly selected Y and is the basis of the Mann–Whitney form of the Wilcoxon rank sum test (*see* **Wilcoxon–Mann–Whitney Test**). For data grouped into K categories, let $q_k = \Pr(X = k)$, $p_k = \Pr(Y = k)$, and $r_k = \sum_{j=0}^{k-1} q_j + \frac{1}{2}q_k$, where $q_0 = 0$. Then [4],

$$\Pr(X < Y) + \frac{1}{2}\Pr(X = Y) = \sum_{k=1}^{K} r_k p_k. \qquad (2)$$

When the underlying data are qualitatively ordered, Bross [6] used (2) to measure the difference in the two distributions. He introduced the term ridit for the partial sum r_k and called (2) the mean ridit. The measure has been used to compare one sample with a known larger population or to compare two samples. Formulas for appropriate variances and large sample normal approximation to the distribution are given in [4]. Worked examples are given in [6] and [26].

Selvin [26] demonstrated the relationship between the mean ridit (2) and the Wilcoxon rank sum test (which uses midranks). Let w_i be the midrank of category i. If n_i is the number of Y observations in category i, then the Wilcoxon statistic $W = \sum_{i=1}^{K} w_i n_i$. Selvin showed that the mean ridit (2) is equal to $[W - n_{\cdot}(n_{\cdot} + 1)/2]/[(N - n_{\cdot})n_{\cdot}]$, where $n_{\cdot} = \sum_{i=1}^{K} n_i$. Indeed, since the right-hand side of (2) is in the form of a linear rank statistic, with the ridit r_k playing the role of a score assigned to the kth category, some researchers have used it to analyze quantitatively ordered data. Although this is a valid procedure, one might be able to obtain more power via the methods of the previous section, i.e. using scores derived from grouped data theory or from the scientific judgment of the investigator.

Related Topics and Recent Developments

While we have separately considered ordered predictor variables and ordered outcome variables, we note that the procedures outlined above may be applied simultaneously when both are ordered. Here one would employ the linear-by-linear association procedure [1].

Analytic procedures employing scores are discussed more fully in the entry on ordered categorical data. Binary outcome data arise in dose–response settings (*see* **Quantal Response Models**), and, more generally, when the predictor variable is ordered (*see* **Trend Test for Counts and Proportions**), for example, in epidemiologic studies.

Suppose covariate adjustment is needed in the two-sample problem with ordered outcome data or in dose–response data with a binary outcome. In either case, one can form strata based on covariate levels, where each stratum is a $2 \times C$ table [19, 28]. For the stratified two-sample problem, a test is constructed by computing stratum-specific linear rank tests and summing them to obtain a combined test. One may assign scores independently within each stratum [17] or, if the category boundaries are the same across all strata, pool the strata and assign scores to the pooled observations [19, 20]. In the continuous case for moderate to large strata the first assignment method may have power advantages over the second [22], and this is likely to carry over to the grouped data situation. In the dose–response setup, Tarone & Gart [28] show that the optimum test depends on the underlying response function, e.g. logistic, probit, or extreme value. If there are several plausible underlying response functions, then one can develop an efficiency robust procedure [12].

While our focus is scores in categorical data analysis, many practitioners assign scores to qualitatively ordered responses as though the scores were observations from continuous distributions. Lipsitz [18] discusses parameter estimation under a **general linear model** using **maximum likelihood**, ordinary **least squares**, and generalized least squares with estimated weights when the scores are used that way. His suggestion that the applied statistician use ordinary least squares is based in part on its relative simplicity.

The occurrence of data where the observations are correlated, e.g. blood lead levels in children within households, has stimulated the development of appropriate analytic methods. Fay & Gennings [9]

present a ridit permutation test and a permutation test based on means using predefined scores for clustered ordinal response data.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

[3] Armitage, P. (1955). Tests for linear trends in proportions and frequencies, *Biometrics* **11**, 375–385.

[4] Beder, J.H. & Heim, R.C. (1990). On the use of ridit analysis, *Psychometrika*, **55**, 603–616.

[5] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, New York.

[6] Bross, I.D.J. (1958). How to use ridit analysis, *Biometrics* **14**, 18–38.

[7] Chernoff, H. & Savage, I.R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics, *Annals of Mathematical Statistics* **29**, 972–994.

[8] Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.

[9] Fay, M.P. & Gennings, C. (1996). Non-parametric two-sample tests for repeated ordinal responses, *Statistics in Medicine* **15**, 429–442.

[10] Gastwirth, J.L. (1965). Asymptotically most powerful rank tests for the two-sample problem with censored data, *Annals of Mathematical Statistics* **36**, 1243–1247.

[11] Gastwirth, J.L. (1966). On robust procedures, *Journal of the American Statistical Association* **61**, 929–948.

[12] Gastwirth, J.L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis, *Journal of the American Statistical Association* **80**, 380–384.

[13] Gastwirth, J.L. (1987). The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data, *Statistical Science* **2**, 213–238.

[14] Graubard, B.I. & Korn, E.L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ contingency tables, *Biometrics* **43**, 471–476.

[15] Hájek, J. & Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

[16] Hoeffding, W. (1951). 'Optimum' nonparametric tests, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California, Berkeley, pp. 83–92.

[17] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

[18] Lipsitz, S.R. (1992). Methods for estimating the parameters of a linear model for ordered categorical data, *Biometrics* **48**, 271–281.

[19] Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, *Journal of the American Statistical Association* **58**, 690–700.

[20] Mantel, N. & Ciminera, J.L. (1979). Use of logrank scores in the analysis of litter-matched data on time to tumor appearance, *Cancer Research* **39**, 4308–4315.

[21] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[22] Podgor, M.J. & Gastwirth, J.L. (1994). A cautionary note on applying scores in stratified data, *Biometrics* **50**, 1215–1218.

[23] Podgor, M.J., Gastwirth, J.L. & Mehta, C.R. (1996). Efficiency robust tests of independence in contingency tables with ordered classifications, *Statistics in Medicine* **15**, 2095–2105.

[24] Randles, R.H. & Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

[25] Saleh, A.K.Md.E. & Dionne, J.-P. (1977). On a further generalization of the Savage test, *Communications in Statistics – Theory and Methods, A* **6**, 1213–1221.

[26] Selvin, S. (1977). A further note on the interpretation of ridit analysis, *American Journal of Epidemiology* **105**, 16–20.

[27] Sen, P.K. (1967). Asymptotically most powerful rank order tests for grouped data, *Annals of Mathematical Statistics* **38**, 1229–1239.

[28] Tarone, R.E. & Gart, J.J. (1980). On the robustness of combined tests for trends in proportions, *Journal of the American Statistical Association* **75**, 110–116.

[29] Weiss, S.H., Goedert, J.J., Sarngadharan, M.G., Bodner, A.J., The AIDS Sereoepidemiology Working Group, Gallo, R.C. & Blattner, A. (1985). Screening test for HTLV-III (AIDS agent) antibodies: specificity, sensitivity and applications, *Journal of the American Medical Association* **253**, 221–225.

(*See also* **Categorizing Continuous Variables**)

MARVIN J. PODGOR & JOSEPH L. GASTWIRTH

Scree Test

The scree test is a technique for determining the number of factors to retain in a **factor analysis** or a **principal components analysis**. It was proposed by Cattell [1] in 1966. He noticed from practical observation that the variance of the factors levels off when the factors are mainly measuring random error. The scree test consists of plotting the **eigenvalues** (in descending order of their magnitude) against their factor numbers and determining this "leveling off". In particular, the scree plot typically shows a distinct break between the steep slope of the larger factors and the gradual trailing off of the rest of the factors. The name "scree test" comes from the resemblance of such a plot to the rubble that accumulates at the foot of a mountain. Two examples of scree plots are given in Figures 1 and 2.

The eigenvalues presented in Figures 1 and 2 are obtained, respectively, from applying a principal components analysis and a factor analysis to the Framingham depression data (for data description, *see* **Principal Components Analysis**). Figure 1 (principal components analysis) suggests retention of five factors, while Figure 2 (factor analysis) suggests retention of four factors. In general, a factor analysis provides a better scree plot solution than a principal components analysis.

Instead of examining the scree plot, an alternative version of the scree test involves an examination of the eigenvalues and their differences [2]. This also provides excellent evidence for the number of salient factors. In this approach, we first set up a table having the eigenvalues (in descending order of their magnitude) as the first row. Then, we calculate the successive differences of the eigenvalues and put them in the second row. When the differences decrease consistently up to a point, followed by a substantially larger difference, and followed then by later differences that are all small (usually less than 0.1), this version of the scree test suggests that the last nonrandom factor is the one immediately preceding the substantially larger difference. There is no precise definition of "substantially larger difference". Tables of eigenvalues and their successive differences are shown in Tables 1 and 2. The eigenvalues from Tables 1 and 2 are obtained, respectively, from performing a principal components analysis and a factor analysis on the Framingham depression data.

In Table 1 (principal components solution), the differences decrease regularly from 2.067 to 0.077, then there is a substantially larger difference (0.168), and all the later differences are well below 0.1, except for the last difference. This suggests retaining the first

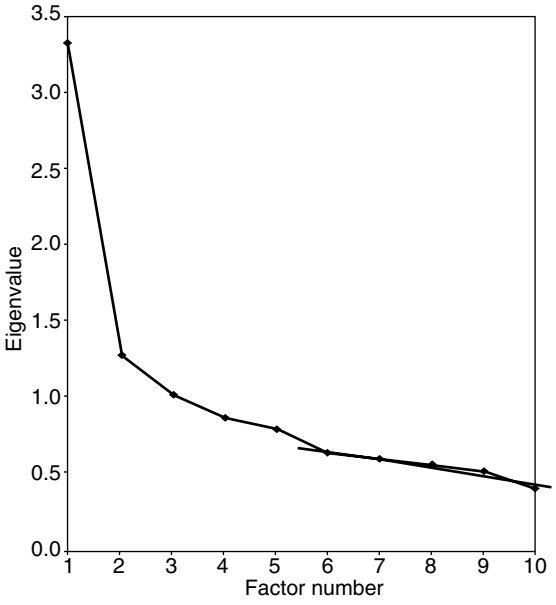

Figure 1 A scree plot from principal components analysis

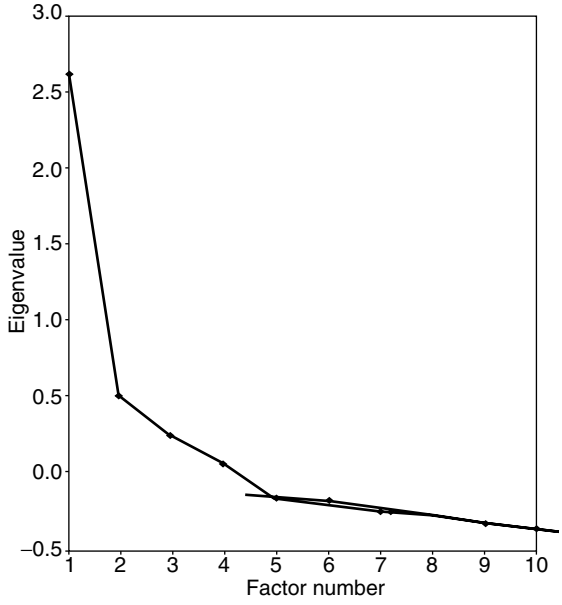

Figure 2 A scree plot from factor analysis

Table 1 Eigenvalues and differences (principal component analysis solution)

	prin1	prin2	prin3	prin4	prin5	prin6	prin7	prin8	prin9	prin10
Eigenvalue	3.358	1.290	1.022	0.872	0.795	0.628	0.590	0.552	0.509	0.386
Difference	2.067	0.268	0.150	0.077	0.168	0.038	0.038	0.043	0.123	
Proportion	0.336	0.129	0.102	0.087	0.080	0.063	0.059	0.055	0.051	0.039
Cumulative	0.336	0.465	0.567	0.654	0.734	0.797	0.855	0.911	0.961	1.000

Table 2 Eigenvalues and differences (factor analysis solution)

	FACT1	FACT2	FACT3	FACT4	FACT5	FACT6	FACT7	FACT8	FACT9	FACT10
Eigenvalue	2.686	0.505	0.254	0.131	−0.042	−0.053	−0.128	−0.155	−0.209	−0.247
Difference	2.182	0.250	0.123	0.173	0.011	0.075	0.027	0.054	0.039	
Proportion	0.979	0.184	0.093	0.048	−0.015	−0.019	−0.047	−0.057	−0.076	−0.090
Cumulative	0.979	1.163	1.256	1.304	1.288	1.269	1.223	1.166	1.090	1.000

five components. In Table 2 (factor analysis solution), the differences decrease from 2.182 to 0.123, then there is a substantial larger jump (0.173), and all the later differences are well below 0.1. Therefore, Table 2 suggests retaining four factors.

The scree test does not always provide a clear solution to the number of retained factors. Sometimes, the break in the scree plot is not as distinct as is shown in Figure 1, or the substantially larger difference is not followed by all small differences in a table of successive eigenvalue differences. In that case, we need to rely on other methods to determine the number of factors to retain (for procedures to determine the number of factors, *see* **Factor Analysis, Overview**).

References

[1] Cattell, R.B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research* **1**, 245–276.

[2] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Screening Benefit, Evaluation of

Screening is the testing of apparently healthy individuals from a population for the purpose of separating them into groups with high and low probabilities of having a given disease such as cancer. The screening is usually initiated by those who offer the tests, and there is thus an implicit promise that those who are screened will benefit. Therefore those individuals screened positive should receive diagnostic follow-up and treatment of proven efficacy if they are diagnosed to have disease [43]. The early detection of cancer and other chronic diseases through screening has long been viewed as a worthwhile public health goal. Many believe that diagnosing disease earlier means that the treatment will be more effective than treatment occurring at the usual time. Unfortunately, the presumption of benefit may not be correct, and the value of a screening program must be demonstrated. This article examines various issues encountered in assessing the benefit of screening. The discussion is in the context of cancer screening.

There are two basic types of studies that can be used to evaluate cancer screening programs, **experimental** and **observational** [8, 29, 35]. The experimental study, commonly termed the randomized controlled trial (RCT), is the method of choice, as it alone produces an **unbiased** assessment of effect (*see* **Clinical Trials, Overview**; **Screening Trials**). When an RCT is not possible, observational or **quasi-experimental designs** may be used. They may be similar to experimental studies in many respects but, because they lack **randomization**, they are generally difficult to analyze and interpret. Nevertheless, both **cohort** and **case–control** observational studies have been used for evaluation of screening for several types of cancer, and the design and interpretation of these studies have been discussed [28, 29, 37, 38, 44, 51].

In the evaluation of cancer screening programs, both the effectiveness of offering screening and the efficacy of screening are important measures. The effectiveness of offering screening is the ratio of the cancer mortality rate for those offered screening to what their cancer mortality rate would have been had they not been offered screening. Effectiveness is of particular importance in considering public health policy. The efficacy of screening is the ratio of the cancer mortality rate for those actually screened to what their cancer mortality rate would have been had they not been screened. Efficacy is of interest in considering the value of screening to those who accept the screening test.

The effectiveness of offering screening can be estimated directly from a comparison of the mortality rates between the randomized groups in an RCT. An indirect estimate of efficacy can also be obtained [10]. Observational study designs typically provide estimates of efficacy. To achieve this goal, the design features typically used in an RCT, such as clearly established criteria for inclusion or exclusion (*see* **Eligibility and Exclusion Criteria**), a well-defined **target population**, a carefully defined intervention protocol with quality control provisions (*see* **Clinical Trials Protocols**), and a clear definition of the study end point (*see* **Outcome Measures in Clinical Trials**) with careful (possibly blind) ascertainment, should be met by an observational study just as for an experimental study.

Selection Bias and Observational Studies

The deficiency of all observational studies is the lack of a **control** group constructed by a chance mechanism. The purposes and advantages of randomization are well known [6]. Use of a control group chosen by any method other than randomization requires the assumption either that the control and intervention groups are identical in all important variables except the intervention under study, or that one can correct for all relevant differences. In the latter case, one must further assume that all factors affecting the course of the disease are identified and measured. These assumptions are rarely, if ever, justified.

In the screening setting, the ability to make valid **inferences** inferences from studies of patient groups or observations of changes in community rates is severely limited by the difficulty of defining appropriate comparison groups, by the lack of detailed knowledge of disease natural history, and by the inadequate understanding of reasons for changes over time in incidence, survival, and mortality rates for various diseases (*see* **Morbidity and Mortality, Changing Patterns in the Twentieth Century**).

The essence of the problem with observational studies is the potential noncomparability of the populations being compared. In screening, a strong **selection bias** may operate with regard to the characteristics of individuals who agree to participate in a screening program compared to those who refuse to participate. This has been demonstrated in the HIP study [46]. The HIP study was designed to evaluate screening with clinical examination plus mammography by comparing breast cancer mortality between a control group and a total study group, the latter including those who were screened and those who refused screening. Examination of the data revealed substantial differences in disease characteristics between the respondents and the refusers. Overall, among study screened women over a five-year period, the rate of case detection was about 2.3 per 1000. By contrast, the incidence rate among study women who refused screening was 1.59 per 1000 per year, while the rate in the control group was 1.95 per 1000 per year. This suggests that women with a higher risk of breast cancer tended to select themselves for screening. In addition, the death rates from all causes excluding breast cancer per 10 000 per year after five years of follow-up were: control group 56.4; total study group 55.1; study screened group, 42.4; study refused screening group, 81.0. Thus, general mortality excluding breast cancer was far lower among the respondents. Given this self-selection bias, construction of an appropriate comparison group for the women who elected to be screened was not possible [4, 46]. In contrast, it is generally recognized in North America that women who participate in cervical cancer screening are at lower risk for cervical cancer incidence and death.

Lead Time and Length Bias

There is only one outcome variable known to be valid in cancer screening studies: the cancer mortality rate. This is the number of cancer deaths per unit time per unit population at risk (*see* **Person-years at Risk**) [8, 29, 35, 37, 38, 42]. For some screening procedures which detect presumed precursor lesions, such as the Pap smear for cervical cancer, a reduction in cancer incidence is also a useful outcome to assess, but it is still important to know that the prevented cancers were those that would have led to death. Obtaining an accurate estimate of the mortality measure requires a careful study design and long-term follow-up of large populations, which is usually a costly undertaking. Consequently, intermediate or surrogate outcome measures have been sought, such as a shift to a more favorable stage at diagnosis distribution (*see* **Surrogate Endpoints**). There are, however, critical shortcomings associated with these endpoints [29, 35, 37, 42]. The shortcomings can be traced to lack of knowledge about the preclinical **natural history** of disease and well-known **biases** which occur in screening programs, lead time bias, and **length bias**.

If an individual participates in a screening program and has disease detected earlier than it would have been in the absence of screening, then the amount of time by which diagnosis is advanced is termed the *lead time*. Because of the lead time phenomenon, the point of diagnosis is advanced and survival as measured from diagnosis is automatically lengthened for cases detected by screening, even if length of life is not increased. This is referred to as *lead time bias* and renders the case survival endpoint invalid [23, 29, 35, 37, 38, 42, 55].

Length bias refers to the phenomenon that cases of disease detected by a screening program are not a random sample from the general distribution of cases of preclinical disease in the screened population. Instead, the longer-duration preclinical disease cases are overrepresented among the detected cases [1, 29, 35, 37, 54]. The importance of this bias is that if disease with long preclinical duration is slow-growing preclinical disease, which then progresses to slow-growing clinical disease, cases of disease with more favorable progression rates are the ones more likely to be detected by screening. Thus, screen-detected cases will tend to have characteristics of good **prognosis**, such as lack of involvement of regional lymph nodes and a more favorable outcome even in the absence of screening.

Related to the concepts of lead time bias and length bias is overdiagnosis bias. There exists the possibility of a nonprogressive or regressive preclinical disease state in which some cases of the disease are detectable by the screening test but would not progress to clinical disease during the individuals' lifetimes in the absence of screening. This is potentially a major problem in screening for prostate cancer, where the autopsy **prevalence** of prostate cancer in elderly men can approach 50% [25]. Clearly,

the detection of such a case is of no benefit to the individual, but such cases remain preclinical over a long time and with repeated screenings are therefore more likely to be detected. Because the counterparts to these lesions do not surface in a control population, a screened population will contain a higher proportion of early-stage cases even if there is no mortality effect from screening.

Study Endpoints

Three of the most frequently proposed alternative endpoints are case-finding rate or yield, stage of disease, and case–survival rate. Consider first the case-finding rate or the **incidence rate**. This quantity may provide an early clue as to whether or not screening might be doing something, in the sense that more cases should be detected in the presence than in the absence of screening. However, this rate yields no information on the effect of the screening on disease outcome. One would expect case finding to increase in a screened population, at least initially, relative to an unscreened population, because of lead time bias. This can happen in the presence or absence of a mortality effect. Furthermore, care must be taken about the definition of discovered cancer in an early detection program. Many borderline lesions found by screening may not be progressive disease. This results in overdiagnosis bias, as noted above. If so, individuals may be unnecessarily treated and exposed to other possible risks of screening.

Stage

The stage of a disease at diagnosis can also be used as an early indicator that screening might be accomplishing something, but it can be misleading and is unsatisfactory as a final endpoint for various reasons. The measurement and definition of stage can be subjective, and its proper use requires strict guidelines and tight control over pathology, each of which may be difficult to implement in practice. More importantly, the relationship of stage to survival or mortality has not been generally established in the screening setting.

The problem is likely to be most pronounced for Stage I or localized cases, particularly if a study has a cutoff point after which new cases are not accrued. In this circumstance, lead time and length bias can result in slow-growing, even nonprogressive, cases being detected in Stage I in a screened group to a greater extent than in a control group. Their counterparts in the control group may not surface by the cutoff point, if ever, and as a result the screened group will contain a higher proportion of Stage I cases even if screening has no effect on mortality. If there is a mortality effect, the magnitude could be exaggerated by confining an analysis to stage of disease. Thus, while observation of a stage shift in a screened group is a sign of early detection, it is insufficient evidence of an effect on disease outcome.

Survival

A further alternative endpoint is survival, specifically the case–survival rate. In contrast to mortality, which is a population measure, the case–survival rate refers only to cancer cases within a population, being the proportion of cases alive after some time period. Because there are losses to follow-up, this endpoint is ordinarily calculated using **life table** methods (*see* **Survival Analysis, Overview**). While this endpoint does address the final outcome of disease and gives suggestive evidence of screening effectiveness, it cannot be relied upon to reflect mortality accurately. Because diagnosis occurs earlier in a screening program, any observed increase in survival from time of diagnosis is, at least in part, simply a reflection of lead time. For any given case of the disease, it is impossible to distinguish between a true increase in survival time and an artificial increase due to lead time because lead time cannot be observed directly for ethical reasons. Further, there is as yet no universally accepted procedure available to estimate lead time or to adjust survival for lead time, although research in this area has appeared [50, 53]. Consequently, case survival is not a valid measure of screening effectiveness.

Furthermore, even if one could adjust for lead time, the problem of length bias would still exist in making survival comparisons. For example, the length bias effect may be different between two subgroups of cases detected by different screening modalities. That is, the cases in one subgroup may have a different distribution of natural histories than the cases in another subgroup because of a modality-dependent sampling effect. Thus, even if one could adjust for lead time, any remaining

survival difference could be real, or could simply be a consequence of the difference in disease natural history between the groups, or a combination of the two factors. Unfortunately, no general methodology exists to either estimate the magnitude of a length bias effect or to adjust the survival for length bias.

An alternative approach is to consider survival time measured from entry into a study instead of from time of diagnosis. In this way, all cases of disease have their survival time starting from the same time origin, thereby eliminating the lead time bias. Survival times and survival distributions calculated in this way are not comparable to survival calculated in the usual way, but their use can lead to a valid comparison of time in study between the cases in an intervention group and the cases in a control group. While this procedure avoids the lead time problem, one must still be concerned about length bias [2].

Rate of Advanced-Stage Disease

Another measure which has received increasing attention as an endpoint in screening studies is the population incidence rate of advanced-stage disease [7, 11, 34, 41, 49]. The overall incidence rate or the rate of early-stage disease should increase with screening, and could be artificially inflated because of length bias and overdiagnosis bias as discussed above, rendering these measures invalid as endpoints. However, if screening reduces the rate of advanced disease or disease which has metastasized and/or is likely to lead to death, then it is reasonable to expect that the death rate from the disease will also be reduced. Whether this is a valid substitute for mortality must be established for each cancer separately, by first defining advanced-stage disease for a particular cancer and then assessing the relationship between advanced disease and mortality in properly designed studies.

Randomized Controlled Trial

In the typical RCT of screening, individuals or groups are randomized to either a study (screening) group or a **control** group. Screening is offered to those in the study group and no screening is offered to those in the control group. Alternatively, the groups may be offered different screening modalities. At the conclusion of the trial, the difference in the cancer mortality from entry to the end of follow-up for the two groups is assessed. If screening does detect preclinical disease and if treatment initiated earlier than usual is more effective than treatment given at the usual time of diagnosis, then the intervention group should have fewer cancer deaths than the control group.

Several authors [9, 26, 27, 40] maintain that the screening RCT with mortality endpoint is the only way to ensure that inferences are not subject to selection, length, and lead time biases, and that all other designs are suspect in this regard. Familiarity with basic screening RCT designs is therefore important. Four basic RCT designs are described and their relative advantages and disadvantages discussed. In principle, each of these designs can be used to address a single question in a two-arm design or multiple questions in a multiple-arm design [16, 36].

Continuous Screen Design

A natural design for a cancer-screening RCT is to randomize individuals either to an intervention or a control arm, with the intervention consisting of periodic screening throughout the trial. Those in the control arm are not offered the periodic screening; they follow their usual medical care practices. This is called the "Continuous screen" design since screening continues for the duration of the study. The NCI Cooperative Lung Cancer Screening RCT done in the mid 1970s to the mid 1980s essentially followed this design [17]. The major goal of the study was to determine whether screening for lung cancer with sputum cytology and chest X-ray was more effective in reducing lung cancer mortality than screening using chest X-ray. One drawback of the continuous screen design is that the cost involved in screening all intervention group participants for the duration of the trial may be prohibitive. With this in mind, an alternative, namely the "stop screen" design, has been proposed.

Stop Screen Design

The "stop screen" design is similar to the continuous screen design, except that screening is offered for only a limited time in the intervention group. However, both arms are followed for disease incidence and mortality until the end of the trial. This design is suggested when it is anticipated that a long follow-up will be required before a reduction in mortality can be

expected to emerge, and when it would be expensive or difficult to continue the periodic screening for the entire trial period. The HIP study [47] followed this design. Sixty-two thousand women aged 40–64 from the HIP population were randomized. The intervention arm was offered four annual screens consisting of two-view mammography and clinical breast examinations. The screens were offered at entry and for the next 3 years. Women in the control arm followed their usual medical practices. Evaluations were done at 5 and 10 years, but women in both arms were followed for 15 years to assess long-term effects of screening.

By stopping screening, the stop screen design can result in a considerable saving in cost and effort relative to the continuous screen design. However, analysis of the stop screen design can be more complex than that of the continuous screen design because the difference in disease-specific mortality between the two arms may be diluted by deaths among the cancers that develop in the intervention arm after screening stops. In addition, the stop screen design is the only one that allows for assessment of overdiagnosis by screening because, by stopping screening and continuing follow-up, one can determine if any excess of cases existing at the time screening stops persists (overdiagnosis) or disappears.

Split Screen Design

The "split screen" design is a variant of the stop screen design. The difference is that at the time the last screen is offered to the intervention arm, a screen is also offered to all those in the control arm. The Stockholm Breast Cancer screening trial is an example of this design [20]. Women were randomized to intervention or control, beginning in 1981. The intervention consisted of two single-view mammograms, performed roughly 28 months apart. The control group was offered a single screen, at approximately 4.5 years after study entry.

One advantage of the split screen design is that there is greater potential to identify comparable groups of cancer cases in the control and intervention arms for the analysis, since the screen in the control arm presumably identifies the counterparts to the cases previously identified in the screened arm. However, at least some of the control arm cancers detected by screening may benefit from being screened and, if so, this benefit may cause some dilution of effect.

Delayed Screen Design

The "delayed screen" design is a variant of the continuous screen design. The difference is that periodic screening is offered to the control arm starting at some time after the start of the study and continuing until the end of the study. This design allows one to estimate the marginal effect of introducing screening at an early time or age, relative to starting the screening at a later time or age. The UK Breast Cancer Screening Trial of women under 50 is basically following this design [30]. Specifically, women in the intervention arm are offered annual screening starting at age 40–41 and continuing to age 47–48, then at age 50 all women in both arms are offered periodic screening as part of the National Health Care Program in the UK. This study is being conducted to evaluate starting periodic screening at age 40–41 relative to waiting until age 50 to start the screening.

The delayed screen design is particularly well suited for the situation where screening is already the standard of care in an older population, and the research question concerns the benefit of introducing screening at an earlier age. Otherwise, the additional costs associated with implementing this design may render it infeasible.

Observational Study Designs

One useful observational study design involves a comparison of cancer incidence and mortality in a defined population before and after the introduction of a screening program. An alternative is to compare geographic regions, established to be as comparable as possible with respect to disease mortality. Time trends in incidence and mortality can be examined and interarea comparisons of intensively screened areas with nonscreened areas can be made. Both approaches require rapid introduction of the screening program and virtually full coverage of the population at risk. For cancer screening, reliable incidence and mortality data, which are available for at least a 10-year period prior to the start of screening and which are predictable for the future, should be available. Ideally, this would be total incidence and mortality data, not simply for the cancer of interest. For other diseases, shorter time spans might apply. Another requirement is the capability for accurate, long-term follow-up of the entire population

at risk. Such an observational study may require substantially larger populations than an RCT and may eventually prove more costly than the more rigorous design. Apart from the potential biases involved in such an observational study, this design offers the possibility of estimating either screening effectiveness or efficacy, depending upon whether the intervention population involves only screened individuals or is a population offered screening in which some individuals accept the testing and some refuse. This approach has been used to evaluate cervical cancer screening [21].

The case–control study is another observational design which can serve as an approach to screening evaluation [8, 28, 44, 51]. In principle, such studies can be used to evaluate the efficacy of screening in the prevention of death, or of invasive disease and consequently death in situations where the screening test detects precursor lesions. A number of such studies have been undertaken, mainly aimed at evaluation of cervical and breast cancer screening [10, 14, 19, 33, 45, 48]. In addition, papers have appeared which address methodologic issues that arise in the screening case–control design [3, 5, 10, 13, 15, 22, 28, 29, 31, 44, 51, 52].

For this purpose, appropriate definition of cases and controls is required. As the primary measure of cancer screening efficacy is mortality, eligible cases should be deaths from the disease of interest in the population under study, irrespective of the means of diagnosis. Eligible controls are all living individuals in the population from which the cases were derived, including individuals with the disease. One then determines whether or not exposure to screening is associated with a reduction in the risk of death from the disease.

This approach also has potential for evaluating the **sensitivity** of screening tests, and may yield information on the relative effectiveness of different screening strategies. However, the approach is only applicable for screening tests which have been in use for several years because screening histories of cases and controls are required.

A particular concern with both the cohort and case–control designs involves selection bias among individuals who choose to be screened as against those who do not, and the impact of this bias on the validity of the inference drawn [10, 18, 24, 32]. The case–control design and the related cohort design in which screened individuals are compared with refusers both assess whether or not screening reduces the mortality of those individuals who elect to be screened relative to those who do not so elect. This comparison does not estimate efficacy since the self-selected comparison group does not in general provide an estimate of the mortality rate of the screenees if they had not been screened. However, both designs can provide a valid efficacy comparison if those refusing to be screened have the same cancer mortality rate as those accepting screening would have had had there been no screening, i.e. if there is no self-selection bias [10].

Information Requirements for Evaluation

Several key data items must be defined and carefully collected in order to achieve a complete evaluation of a screening program [7, 11, 12, 36].

1. *Population characteristics*. The demographic, socioeconomic, and risk characteristics of the target population or study population should be ascertained. In some circumstances, dietary and occupational history may also be pertinent, such as in a study of colorectal cancer, where diet may play an etiologic role, or of bladder cancer, where occupational exposure to carcinogens may influence the evaluation of screening.
2. *Coverage and compliance*. The proportion of the population offered screening who actually undergo the initial and subsequent screening tests should be determined. This indicates the level of interest in, and acceptability of, the screening procedure, and whether or not the level is high enough to have a chance of achieving an impact in the population.
3. *Test yield*. The number and proportion of cases found by screening, particularly in relation to the cases not discovered by screening (the so-called interval cases) is important for evaluating how successful the screening test is in finding the disease.
4. *Stage of disease*. This should be ascertained for each case of the disease in the population under surveillance, whether detected by screening or diagnosed clinically. This information can be used to compare the stage distribution of screen detected cases vs. other case subsets to determine if screening might have an impact on mortality,

and is necessary for defining stage-specific incidence rates.

5. *Case survival*. Survival time should be determined for each case of disease. As with stage information, the survival distribution of screen-detected cases can be compared with that of other case groups to seek some indication that screening might have an impact on disease outcome.
6. *Incidence and prevalence rates*. As noted above, the rate of advanced stage disease may be a good intermediate indicator of the impact of screening on mortality. The ratio of prevalence to incidence yields an estimate of the preclinical duration of disease and can be used to estimate the average lead time gained by screening [23] (*see* **Incidence–Prevalence Relationships**).
7. *Mortality rates*. Mortality rates provide the primary evidence on the effectiveness of screening. Mortality rates for the disease of interest can be calculated and compared between a screened group and a control group to measure the impact of screening. In addition, the death rates from other causes should be scrutinized to assess the comparability of the groups with regard to causes of death other than the one of interest.
8. *Therapy*. The therapy used for each case of the disease should be recorded. At a minimum this should be the initial therapy, but adjuvant therapy or treatment for recurrence could be noted as well. This information should be recorded in the same way in the screened and control populations within each disease stage, and is relevant for separating the early detection effect from the treatment component of any screening impact.
9. *Procedures and costs*. To perform an assessment of the cost or cost-effectiveness (*see* **Health Economics**) of a screening program, it necessary to collect data on the costs of all phases of the program. Alternatively, one can record the procedures done in each phase so that costs can be assigned at a later date. Procedures to be included are efforts to recruit the population, the screening tests, all diagnostic procedures following a positive screen or those used to diagnose a case clinically, all treatment procedures, and any efforts expended to follow the population [39].

References

[1] Albert, A., Gertman, P.M. & Louis, T. (1978). Screening for the early detection of cancer – I: the temporal natural history of a progressive disease state, *Mathematical Biosciences* **40**, 1–59.

[2] Aron, J.L. & Prorok, P.C. (1986). An analysis of the mortality effect in a breast cancer screening study, *International Journal of Epidemiology* **15**, 36–43.

[3] Baum, M. & MacRae, K.D. (1984). Screening for breast cancer (letter), *Lancet* **i**, 462.

[4] Bearhs, O.H., Shapiro, S. & Smart, C. (1979). Report of the working group to review the National Cancer Institute/American Cancer Society breast cancer detection demonstration projects, *Journal of the National Cancer Institute* **62**, 639–709.

[5] Berrino, F., Gatta, G., D'Alto, M., Crosignani, P. & Riboli, E. (1984). Use of case-control studies in evaluation of screening programmes, in *Screening for Cancer I – General Principles on Evaluation of Screening for Cancer and Screening for Lung, Bladder and Oral Cancer*. P.C. Prorok & A.B. Miller, eds, *UICC Technical Report Series*, Vol. 78. International Union Against Cancer, Geneva, pp. 29–43.

[6] Byar, D.P., Simon, R.M., Friedewald, W.T., Schlesselman, J.J., DeMets, D.L., Ellenberg, J.H., Gail, M.H. & Ware, J.H. (1976). Randomized clinical trials: perspectives on some recent ideas, *New England Journal of Medicine* **295**, 74–80.

[7] Chamberlain, J. (1984). Planning of screening programs for evaluation and non-randomized approaches to evaluation, in *Screening for Cancer. I – General Principles on Evaluation of Screening for Cancer and Screening for Lung, Bladder and Oral Cancer*, P.C. Prorok & A.B. Miller, eds, *UICC Technical Report Series*, Vol. 78. International Union Against Cancer, Geneva, pp. 5–17.

[8] Cole, P. & Morrison, A.S. (1980). Basic issues in population screening for cancer, *Journal of the National Cancer Institute* **64**, 1263–1272.

[9] Connor, R.J. & Prorok, P.C. (1994). Issues in the mortality analysis of randomized controlled trials of cancer screening, *Controlled Clinical Trials* **15**, 81–99.

[10] Connor, R.J., Prorok, P.C. & Weed, D.L. (1991). The case-control design and the assessment of the efficacy of cancer screening, *Journal of Clinical Epidemiology* **44**, 1215–1221.

[11] Day, N.E., Williams, D.R.R. & Khaw, K.T. (1989). Breast cancer screening programs: the development of a monitoring and evaluation system, *British Journal of Cancer* **59**, 954–958.

[12] Draper, G.J. (1986). Information requirements for cervical cancer screening programs, in *Screening for Cancer of the Uterine Cervix*, M. Hakama, A.B. Miller & N.E. Day, eds. *IARC Scientific Publication* No. 76. International Agency for Research on Cancer, Lyon, pp. 171–181.

[13] Dubin, N., Friedman, D.R., Toniola, P.G. & Pasternack, B.S. (1987). Breast cancer detection centers and case-control studies of the efficacy of screening, *Journal of Chronic Diseases* **40**, 1041–1050.

[14] Ebeling, K. & Nischan, P. (1987). Screening for lung cancer – results from a case-control study, *International Journal of Cancer* **40**, 141–144.

[15] Editorial (1984). Breast screening: new evidence, *Lancet* **i**, 1217–1218.

[16] Etzioni, R.D., Connor, R.J., Prorok, P.C. & Self, S.G. (1995). Design and analysis of cancer screening trials, *Statistical Methods in Medical Research* **4**, 3–17.

[17] Fontana, R.S. (1986). Screening for lung cancer: recent experience in the United States, in *Lung Cancer: Basic and Clinical Aspects*, H.H. Hansen, ed. Martinus Nijhoff, Boston, pp. 91–111.

[18] Friedman, D.R. & Dubin, N. (1991). Case-control evaluation of breast cancer screening efficacy, *American Journal of Epidemiology* **133**, 974–984.

[19] Friedman, G.D., Hiatt, R.A., Quesenberry, C.P. & Selby, J.V. (1991). Case-control study of screening for prostatic cancer by digital rectal examinations, *Lancet* **337**, 1526–1529.

[20] Frisell, J., Eklund, G., Hellstrom, L. et al. (1989). The Stockholm breast cancer screening trial 5-year results and stage at discovery, *Breast Cancer Research and Treatment* **13**, 79–87.

[21] Hakama, M., Miller, A.B., Day, N.E., Glas, U. & Somell, A. (1986). *Screening for Cancer of the Uterine Cervix. IARC Scientific Publications* No. 76. International Agency for Research on Cancer, Lyon.

[22] Hosek, R.S., Flanders, W.D. & Sasco, A.J. (1996). Bias in case-control studies of screening effectiveness, *American Journal of Epidemiology* **143**, 193–201.

[23] Hutchison, G.B. & Shapiro, S. (1968). Lead time gained by diagnostic screening for breast cancer, *Journal of the National Cancer Institute* **41**, 665–681.

[24] Janzon, L. & Andersson, I. (1991). The Malmo mammographic screening trial, in *Cancer Screening*, A.B. Miller, J. Chamberlain, N.E. Day, M. Hakama & P.C. Prorok, eds. Cambridge University Press, Cambridge, pp. 37–44.

[25] Kramer, B.S., Brown, M.L., Prorok, P.C., Potosky, A.L. & Gohagan, J.K. (1993). Prostate cancer screening: what we know and what we need to know, *Annals of Internal Medicine* **119**, 914–922.

[26] Miller, A.B. & Bulbrook, R.D. (1982). Screening, detection, and diagnosis of breast cancer, *Lancet* **1**, 1109–1111.

[27] Morrison, A. (1982). The effects of early treatment, lead time, and length bias on the mortality experienced by cases detected by screening, *International Journal of Epidemiology* **11**, 261–267.

[28] Morrison, A.S. (1982). Case definition in case-control studies of the efficacy of screening, *American Journal of Epidemiology* **115**, 6–8.

[29] Morrison, A.S. (1985). *Screening in Chronic Disease*. Oxford University Press, New York.

[30] Moss, S. (1994). Personal communication.

[31] Moss, S.M. (1990). Case-control studies of screening, in J. Chamberlain, N.E. Day, M. Hakama & P.C. Prorok eds. *Screening For Cancer*, A.B. Miller, Cambridge University Press, Cambridge, pp. 419–428.

[32] Moss, S.M. (1991). Case-control studies of screening, *International Journal of Cancer* **20**, 1–6.

[33] Oshima, A., Hirata, N., Ubukata, T., Urneda, K. & Fujimato, I. (1986). Evaluation of a mass screening program for stomach cancer with a case-control design, *International Journal of Cancer* **38**, 829–833.

[34] Paci, E., Ciatto, S., Buiatti, E., Cecchini, S., Palli, D. & Rosselli del Turco, M. (1990). Early indicators of efficacy of breast cancer screening programs. Results of the Florence district program, *International Journal of Cancer* **46**, 198–202.

[35] Prorok, P.C. (1984). Evaluation of screening programs for the early detection of cancer, in *Statistical Methods for Cancer Studies*, R.G. Cornell, ed. Marcel Dekker, New York, pp. 267–328.

[36] Prorok, P.C. (1995). Screening studies, in *Cancer Prevention and Control*, P. Greenwald, B.S. Kramer & D.L. Weed, eds. Marcel Dekker, New York, pp. 225–242.

[37] Prorok, P.C. & Connor, R.J. (1986). Screening for the early detection of cancer, *Cancer Investigation* **4**, 225–238.

[38] Prorok, P.C. & Miller, A.B., eds (1984). General principles on evaluation of screening for cancer in *Screening for Cancer I – General Principles on Evaluation of Screening for Cancer and Screening for Lung, Bladder and Oral Cancer. UICC Technical Report Series*, Vol. 78, Geneva, pp. 3–4.

[39] Prorok, P.C., Connor, R.J. & Baker, S.G. (1990). Statistical considerations in cancer screening programs, *Urologic Clinics of North America* **17**, 699–708.

[40] Prorok, P.C., Hankey, B.F. & Bundy, B.N. (1981). Concepts and problems in the evaluation of screening programs, *Journal of Chronic Diseases* **34**, 159–171.

[41] Roberts, M.M., Alexander, F.E., Anderson, T.J., Chetty, U., Donnan, P.T., Forrest, P., Hepburn, W., Huggins, A., Kirkpatrick, A.E., Lamb, J., Muir, B.B. & Prescott, R.J. (1990). Edinburgh trial of screening for breast cancer: mortality at seven years, *Lancet* **335**, 241–246.

[42] Sackett, D.L. (1975). Periodic examination of patients at risk, in *Cancer Epidemiology and Prevention, Current Concepts*, D. Schottenfeld, ed. Charles C. Thomas, Springfield, pp. 437–454.

[43] Sackett, D.L. & Holland, W.W. (1975). Controversy in the detection of disease, *Lancet* **23**, 357–359.

[44] Sasco, A.J., Day, N.E. & Walter, S.D. (1986). Case-control studies for the evaluation of screening, *Journal of Chronic Diseases* **39**, 399–405.

[45] Selby, J.V., Friedman, G.D., Quesenberry, C.P. & Weiss, N.S. (1992). A case-control study of screening sigmoidoscopy and mortality from colorectal cancer, *New England Journal of Medicine* **326**, 653–657.

[46] Shapiro, S., Venet, W., Strax, P., Venet, L. & Roeser, R. (1985). Selection, follow-up, and analysis in the Health Insurance Plan study: a randomized trial with breast cancer screening, *National Cancer Institute Monographs* **67**, 65–74.

[47] Shapiro, S., Venet, W., Strax, P. & Venet, L. (1988). *Periodic Screening for Breast Cancer. The Health Insurance Plan Project and its Sequelae, 1963–1986*. The Johns Hopkins University Press, Baltimore.

[48] Sobue, T., Suzuki, T., Naruke, T. & the Japanese Lung Cancer Screening Research Group (1992). A case-control study for evaluating lung-cancer screening in Japan, *International Journal of Cancer* **50**, 230–237.

[49] Tabar, L., Fagerberg, C.J.G., Gad, A., Baldetorp, L., Holmberg, L.H., Grontoft, O., Ljungquist, U., Lundstrom, B., Manson, J.C., Eklund, G., Day, N.E. & Pettersson, F. (1985). Reduction in mortality from breast cancer after mass screening with mammography, *Lancet* **i**, 829–832.

[50] Walter, S.D. & Stitt, L.W. (1987). Evaluating the survival of cancer cases detected by screening, *Statistics in Medicine* **6**, 885–900.

[51] Weiss, N.S. (1983). Control definition in case-control studies of the efficacy of screening and diagnostic testing, *American Journal of Epidemiology* **118**, 457–460.

[52] Weiss, N.S., McKnight, B. & Stevens, N.G. (1992). Approaches to the analysis of case-control studies of the efficacy of screening for cancer, *American Journal of Epidemiology* **135**, 817–823.

[53] Xu, J.L. & Prorok, P.C. (1995). Non-parametric estimation of the post-lead time survival distribution of screen detected cancer cases, *Statistics in Medicine* **14**, 2715–2725.

[54] Zelen, M. (1976). Theory of early detection of breast cancer in the general population, in *Breast Cancer: Trends in Research and Treatment*, W.H. Mattheiem & M. Rozenzweig, eds, J.C. Henson, Raven Press, New York, pp. 287–300.

[55] Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic disease, *Biometrika* **56**, 601–614.

(*See also* **Prevalence of Disease, Estimation from Screening Data**; **Screening, Models of**; **Screening, Sojourn Time**)

PHILIP C. PROROK

Screening Trials

The early detection of cancer and other chronic diseases has long been a goal of medical scientists. Many believe that by moving the point of diagnosis backward in time so that the disease is diagnosed earlier than usual, treatment will be more effective than treatment given at the usual time. However, this presumption may not be correct and the effect of any screening program must be evaluated. **Cohort studies** and **case–control studies** have been used for evaluating **screening** for several types of cancer, and the design and interpretation of these studies have recently been the topic of increasing discussion (*see* **Screening Benefit, Evaluation of**). However, an **observational study** rarely yields definitive answers or permits solid conclusions with regard to the public health consequences of cancer screening. The most rigorous approach is the randomized **clinical trial**. There are special design and analysis issues for such screening trials.

Design Issues

The randomized controlled trial involves the prospective testing and long-term follow-up of defined populations according to a protocol (*see* **Clinical Trials Protocols**). There are several major design and implementation aspects that should be considered. First, the target disease(s), the screening test(s), and the diagnostic and therapeutic regimens must be determined. Then, the appropriate outcome variable (*see* **Outcome Measures in Clinical Trials**) must be chosen and the sampling unit (*see* **Unit of Analysis**) (individual or group) selected. Next, the admission and exclusion criteria need to be established (*see* **Eligibility and Exclusion Criteria**) and a **randomization** procedure chosen to allocate eligibles to the study and **control** groups (*see* **Randomized Treatment Assignment**). The study and control groups should be followed up with equal intensity and in the same time frame, with the outcome variable measured in a blind fashion (*see* **Blinding or Masking**), if possible. Every effort should be made to maximize adherence to the study protocol for both groups (*see* **Compliance Assessment in Clinical Trials**). It is also important in the analysis that all individuals in the control group be compared to all individuals in the study group, including both individuals accepting the offer of screening and those rejecting the offer (*see* **Intention to Treat Analysis**).

A decision on the number of screening examinations and the interval between examinations (screens) must be made. The number of screens depends on the tradeoff between a sufficient number to realize an effect, if there is one, and the cost of additional screens. Trials may incorporate screening for essentially the entire follow-up period [35], or employ an abbreviated screening period typically involving four or five screening rounds, with a subsequent follow-up period devoid of screening [22, 32]. Several modeling efforts have addressed these issues [6, 16, 18, 19].

Another design problem involves the relationship between study duration, sample size, and the expected timing of any effect. Sample size and study duration are inversely related. If these two parameters were the only ones to consider, the relationship between follow-up cost, on the one hand, and recruitment and screening cost, on the other, would determine the design. However, the time at which a reduction in mortality may occur must also be considered. For those cancer screening trials that have demonstrated a reduction in mortality, a separation between the mortality rates in the screened and control groups did not occur until four to five years or more after randomization [32, 35]. Furthermore, the difference may continue to increase with time, even after screening stops [32]. Thus, even with a very large sample size, follow-up may have to continue for many years to observe the full effect of the screening. A follow-up period of at least 10 years is appropriate, but a longer period may be required if the screening effect is manifested primarily among a subset of patients with slowly growing cancer (*see* **Sample Size Determination for Clinical Trials**).

Determination of the appropriate endpoint in a cancer screening study is intimately related to the disease natural history. For a screening trial, the relevant natural history is from the time the cancer is screen-detected to death. This natural history is usually not well understood, and potential early indicators of outcome such as a shift in disease stage or a lengthening of survival among cases, which depend on knowledge of this natural history for their validity, cannot provide a definitive assessment of screening in the absence of this knowledge.

There is only one outcome variable known to be valid in a cancer screening trial, namely the population cancer mortality rate. This is the number of cancer deaths per unit time per unit population at risk [23, 27]. The mortality rate provides a combined assessment of early detection plus therapy. No improvement in mortality will be seen if either the screening does not lead to earlier detection or therapy at the time of earlier detection confers no extra benefit.

Intermediate or surrogate outcome measures have also been considered (*see* **Surrogate Endpoints**). However, these have critical shortcomings that can be traced to the well-known **biases** that occur in screening programs: lead time bias, **length bias**, and over-diagnosis bias [23, 27] (*see* **Screening Benefit, Evaluation of**). Among the most frequently proposed alternative endpoints are the case–finding rate or yield, stage of disease, and case–survival rate.

Another measure that has been proposed as an endpoint in screening studies is the population **incidence rate** of advanced stage disease [1, 5], since, if screening reduces the rate of advanced disease, disease that has metastasized, or is likely to lead to death, then it is reasonable to expect that the death rate from the disease will also be reduced.

Sample Size

In the **hypothesis-testing** framework, the sample size can be calculated from the appropriate formula if one knows the event rate, effect size, and statistical procedure, all of which depend on the choice of endpoint for the study (*see* **Sample Size Determination**). In cancer screening trials, this is the cancer mortality. Since the analysis involves a comparison of the numbers or rates of deaths, methods for **Poisson-distributed** data can be used [34, 36]. Other factors that must be taken into consideration are noncompliance in the screened and control groups, randomized groups of different sizes, and lower than expected event rates among the individuals who participate in the study. Several approaches have been formulated to address these problems [23, 25].

One approach to sample size estimation is based on the method suggested by Taylor & Fontana [36], modified to allow for an arbitrary magnitude of screening impact, an arbitrary sample size ratio between the screened and control groups, and arbitrary levels of compliance in the screened and control groups. Let N_c be the number of individuals randomized to the control group, and N_s the number randomized to the screened group, with $N_s = fN_c$. Assume the study is designed to detect a $(1-r) \times 100\%$ reduction $(0 \le r \le 1)$ in the cumulative disease-specific death rate over the duration of the trial. Also, let P_c be the proportion of individuals in the control group who comply with the control group intervention and P_s be the proportion of individuals in the screened group who comply with the screened group intervention.

Using a model in which the death rate in the presence of noncompliance in the screened group is a linear combination of the screened and control group death rates, weighted by the compliance levels, one finds that the total number of disease-specific deaths, D, needed for a one-sided α-level significance test with power $1-\beta$ is given by

$$D = \frac{[(\Theta_1 + f\Theta_2)Z_{1-\alpha} - (\Theta_1\Theta_2)^{1/2}(1+f)Z\beta]^2}{f(\Theta_1 - \Theta_2)^2},$$

where $\Theta_1 = r + (1-r)P_c$, $\Theta_2 = 1 - (1-r)P_s$, and $Z_{1-\alpha}$ and Z_β are the $1-\alpha$ and β **quantiles** of the standard **normal distribution**, respectively. The number of participants required in the control group is $N_c = D/(\Theta_1 + f\Theta_2)R_cY$, where Y is the duration of the trial from entry to end of follow-up in years, and R_c is the average annual disease-specific death rate in the control group expressed in deaths per person per year.

Calculation of N_c requires an estimate of R_c. Individuals recruited for a screening trial are expected to be healthier than the general population due to **selection** factors and eligibility criteria. Hence, the usual cancer mortality rate obtained from national or registry (*see* **Disease Registers**) data is likely to overestimate the mortality rate of the participants, at least for the early part of a trial. An ad hoc approach to this problem is to use the relationship between the observed and expected death rates in previous screening trials. Alternatively, one can calculate an expected event rate using the age-specific incidence rates of a cancer-free population combined with the survival rates of these incident cases to arrive at the expected mortality [23, 25, 26].

Study Designs

Classic Two-Arm Trial that Addresses a Single Question

In this design, the study population is randomized to a group offered screening according to a protocol and a control group not offered screening. At the end of follow-up, the mortality rates in the two groups are compared [28]. The prototype trial for this design is the Health Insurance Plan (HIP) trial of breast cancer screening [32].

Designs for Investigating more than One Question in the Same Study

Extensions of the classic two-arm design have been used or suggested for cancer screening trials to answer more than one question in the same study. This topic has also been discussed for cancer **prevention trials** [11]. For example, the National Study of Breast Cancer Screening in Canada involves two different study populations, but under the same administrative and scientific umbrella (i) to determine in women aged 40–49 at entry whether annual screening by mammography and physical examination, when used as an adjunct to the highest standard of care in the Canadian health care system, can reduce mortality from breast cancer, and (ii) to evaluate in women aged 50–59 at entry the additional contribution of routine annual mammographic screening to screening by physical examination alone in reducing breast cancer mortality. This involves separate randomizations of women in the two age groups [22] (*see* **Randomized Treatment Assignment**). In some circumstances, several related questions can be addressed by including additional randomized groups in the trial. An example is the colon cancer screening trial at the University of Minnesota [14]. Two basic issues are being addressed; namely, whether screening can reduce mortality, and whether there is a different effect at different screening frequencies. Three randomized groups were formed: a control group, a group offered annual screening with a test to detect occult blood in the stool, and a group offered the occult blood test every two years. Another extension of the basic design is a two-group trial in which the intervention group includes multiple interventions, known as the all-versus-none design [11]. One version of this design involves several interventions, with each intervention aimed at early detection of a different type of cancer. Use of this design requires two assumptions: first, that the test for any given cancer does not affect the case detection or mortality of any other cancer site, and, secondly, that disease-specific mortality is independent among the cancers under study. An example is the Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial sponsored by the National Cancer Institute [15], the objectives of which are to determine whether: (i) in females and males, screening with flexible sigmoidoscopy can reduce mortality from colorectal cancer, and screening with chest X-ray can reduce mortality from lung cancer; (ii) in males, screening with digital rectal examination plus serum prostate-specific antigen (PSA) can reduce mortality from prostate cancer; and (iii) in females, screening with pelvic examination plus CA 125 and transvaginal ultrasound can reduce mortality from ovarian cancer. Another design option to answer more than one question at a time is the reciprocal control design [11]. In this design, the participants in each arm of a trial receive an intervention, but also serve as controls for an intervention in another arm of the trial. This requires the assumption that the intervention aimed at a given cancer does not affect any of the other cancers under study.

Within each of the above design types there are options for the relationship between screening and follow-up [9]. A natural design is to randomize individuals either to an intervention or a control group, with the intervention consisting of periodic screening throughout the trial. Those in the control arm are not offered the periodic screening; they follow their usual medical care practices. This is called the continuous-screen design. The NCI Cooperative Lung Cancer Screening RCT, conducted in the mid 1970s to the mid 1980s, essentially followed this design [10].

One drawback of the continuous-screen design is that the cost involved in screening all intervention group participants for the duration of the trial may be prohibitive. An alternative is the stop-screen design in which screening is offered for a limited time in the intervention group and both groups are followed for disease incidence and mortality until the end of the trial. This design is used when it is anticipated that a long follow-up will be required before a reduction in mortality can be expected to emerge, and when it would be expensive or difficult to continue

the periodic screening for the entire trial period. The Health Insurance Plan (HIP) of Greater New York Breast Cancer Screening Study followed this design [32]. The stop-screen design can result in a considerable saving in cost. However, the analysis can be more complex than that of the continuous-screen design, because the difference in disease-specific mortality between the two groups may be diluted by deaths from cancers that develop in the intervention group after screening stops.

The split-screen design is a variant of the stop-screen design in which a screen is also offered to all those in the control group at the time the last screen is offered to the intervention group. The Stockholm Breast Cancer screening trial, conducted in the 1980s, is an example of this design [13]. An advantage of the split-screen design is that there is greater potential to identify comparable sets of cancer cases for the analysis (discussed below).

The delayed-screen design is a variant of the continuous-screen design in which periodic screening is offered to the control group starting at some time after the start of the study and continuing until the end of the study. This design allows the estimation of the marginal effect of introducing screening at some standard time or age, relative to starting the screening at a later time or age. The current UK Breast Cancer Screening Trial of women under 50 is basically following this design to evaluate the effect of beginning screening before the age of 50 years [24].

Analysis

Screening trials involve special issues in their analysis, both for the continuous-screen and stop-screen designs, and for primary and secondary analyses. Primary analyses are concerned with evaluating whether there is a statistically significant difference in disease-specific mortality between the control and intervention groups. Secondary analyses are concerned with ascertaining the magnitude of the mortality difference, and with gaining a deeper understanding of the underlying mechanisms [9].

Primary Analysis

Proposed statistical methods for primary analysis include a Poisson test statistic comparing the observed death rates [33], a **Fisher exact test** comparing the observed proportions of cancer deaths [2], and a **logrank test** comparing disease-specific death rates over time in the two groups [2, 3]. For example, the **Poisson process** test statistic for comparing cumulative mortality rates is

$$\frac{Z_r = (PY_S D_C - PY_C D_S)}{[PY_C PY_S (D_C + D_S)]^{1/2}},$$

where D_C = the number of deaths from the cancer of interest in the control group through the time of analysis, D_S = the corresponding number of deaths in the screened group, PY_C = the number of **person years at risk** of death from the cancer of interest in the control group through the time of analysis, and PY_S = the corresponding number of person years in the screened group [34].

Logrank test analysis may be based on the disease-specific mortality experience of all randomized participants, termed the overall mortality analysis, or it may be based on the mortality experience of comparable groups of cancer cases in the two arms of the trial, in which case it is termed the limited mortality analysis [3].

Overall Mortality Analysis

Overall mortality analyses possess the advantage of comparability of comparison groups formed by randomization. However, logrank tests comparing disease-specific mortality can be relatively inefficient. This is because the logrank test is optimal under proportionality of the disease-specific mortality hazards in the two groups, whereas, in cancer screening trials, there is generally a delay from the beginning of the intervention program to the time that effects on cancer mortality can be observed, and the magnitude of any effect may vary over time. In addition, in stop-screen designs, cases continue to accrue in both groups after screening stops. The cancer deaths in the intervention group that are due to cancers developing after screening dilute the screening effect. Thus, the ratio of hazards decreases with time after some point in the trial. If the specific form of departure from proportional death rates is known, then efficiency can be gained by use of a weighted logrank statistic instead of the usual (unweighted) logrank statistic [8].

Zucker & Lakatos [42] propose a method to accommodate a possible lag until full screening effect within a continuous-screen design. They specify a

range of plausible lag times to full screening effect and then identify the weighted logrank statistic that minimizes the worst possible efficiency loss over this range. Self [29] and Self & Etzioni [30] propose **adaptive** testing methods for stop-screen designs. Their suggestion is to use the observed departures from constancy of the relative hazards to improve the efficiency of the test procedure. These methods are also weighted logrank tests, but the weights are identified in a data-dependent fashion. Sequential versions of these procedures are not yet available.

Limited Mortality Analysis

In a limited mortality analysis, one restricts analysis to comparable sets of cancers, one set consisting of cancers from the intervention group diagnosed through some designated time interval after the start of the study, and the other consisting of their counterparts in the control group diagnosed during the same interval. Limited mortality analyses are typically only applied to split-screen or stop-screen designs. The split-screen design leads naturally to two presumably comparable case sets; namely, those diagnosed up to and including the final screen offered. In the stop-screen design, however, determination of comparable case sets is less straightforward. The main question is how to choose the time interval for ascertainment of cases for analysis. The end of the case ascertainment period should not be too long after screening stops, because the continued accrual of clinically detected cases in both groups may lead to dilution of the observed screening effect as described previously. If, however, we exclude all cases diagnosed after screening has stopped, another form of dilution can arise. Among the cases diagnosed in the control group after screening has stopped, some may correspond to cases in the intervention group that were screen-detected and therefore diagnosed earlier than they would have been without screening. If this set of cancers benefits from the earlier diagnosis due to the screening, then excluding the control group counterparts to these cancers also dilutes the screening effect [8].

Comparability

Whatever the method used to select comparable case sets, the true comparability of the sets selected must be fully investigated. Both the cases in the selected sets and the cases that are diagnosed after the time used to define the selected sets must be evaluated. Methods have been proposed for assessing the comparability of case sets in a stop-screen study [3]. They consider the numerical as well as the biological comparability of the sets. Numerical comparability concerns the numbers of cancers in the case sets. Biologic or qualitative comparability concerns the composition of the cancer case sets with regard to their natural history, and especially their survival characteristics in the absence of screening. Qualitative comparability of candidate sets is assessed by **covariates** defined at randomization associated with the cancer cases.

The identification of comparable case sets is not straightforward. In a stop-screen design, it may be impossible to identify comparable sets if screening is available outside of the trial, and the use of outside screening differs between the two arms after trial screening stops. In a continuous-screen design, it is unlikely that equalization will ever occur. In such cases, appropriately weighted overall mortality analysis may be the only valid option.

In summary, the overall analysis is the most **unbiased** as it compares all randomized individuals. However, this approach may assess a diluted relative effect of screening in a stop-screen design and it requires follow-up of all randomized trial participants. Alternatively, the limited analysis requires follow-up of only selected case sets after a certain point in time and so is less costly. However, the approach is subject to **bias** if the case sets are not truly comparable.

Secondary Analyses

Secondary analyses of cancer screening trials typically involve information related to the outcome of cancer cases captured in survival data, and indications of earlier diagnosis, captured by estimates of the screening program's lead time, **sensitivity**, and the degree of shifting to an earlier clinical stage at diagnosis in the screened group.

Estimates of survival differential are based on the postdiagnosis survival curves in the two case sets (*see* **Survival Analysis, Overview**). The postdiagnosis survival of screen-detected cases includes lead time, which must be explicitly removed to avoid lead-time bias. Initial approaches were developed for the HIP trial assuming a fixed lead time of one year and considering the k-year actuarial survival from

diagnosis of control group cases and interval cases (cases diagnosed in the intervals between screens because of signs or symptoms) as equivalent to the $k + 1$-year survival from diagnosis of screen-detected cases [33]. Walter & Stitt [39] allowed lead time to be a **random variable** with a known distribution. This approach was extended to **nonparametric** estimation procedures by Xu & Prorok [40].

Explicit adjustment for lead time requires knowledge of its probability distribution. Direct estimates of mean lead time have been presented by Shapiro et al. [31], Morrison [23], and Kafadar & Prorok [17]. The approach of Shapiro et al. and Morrison yields crude estimates of average lead time based on comparing disease incidence in the control and intervention groups. Kafadar & Prorok used differences in survival from entry and from diagnosis between screened and control group cases to estimate benefit time and lead time assuming two comparable case sets like those identified for a limited mortality analysis.

Other methods for estimating lead time have been developed by Zelen & Feinleib [41] and Walter & Day [38]. These approaches may be thought of as statistical modeling efforts (*see* **Model, Choice of**). **Simulation** modeling is also being increasingly employed to estimate screening program properties and disease natural history, and to project the costs and benefits (*see* **Health Economics**) of alternative screening strategies [7, 37].

The information on shifts in the distribution of clinical stage at diagnosis should be interpreted with caution, since shifts may be due to overdiagnosis or length bias and therefore need not imply disease-specific mortality benefit. However, a stage-shift model has been developed that allows the estimation of the amount of shift between and within stages due to screening, as well as the associated mortality benefits. The model requires comparable case sets [4].

Trial Monitoring

Various categories of data and information become available at successive stages of a screening trial. These relate to the population under study, acceptance of the screening test by the population, outcomes and characteristics of the screening test, and intermediate and final effect measures or endpoints used for determining the value of screening. These variables can be examined on a regular basis for evidence to alter the protocol or stop the trial, and are also valuable in assessing the consistency of findings or conclusions. More specifically, the data that can be used for monitoring include: population descriptors such as demographic, socioeconomic, and risk characteristics of the population; the proportion of the study population offered screening who undergo the initial screening, the level of compliance with scheduled repeat screens, and the level of screening contamination in the control group; the yield of cancers as a result of screening, the interval cancer rate, and screening test characteristics including sensitivity, **specificity** and **predictive value**; diagnostic and therapeutic follow-up among individuals designated suspicious or positive by the screening test and the costs involved in these procedures; cancer case characteristics such as stage, histologic type, grade, and nodal involvement; survival of cancer cases; incidence and prevalence rates of the cancer of interest; the incidence rate of advanced stage cancer; and mortality rates from the cancer of interest and other causes [28].

Another aspect of the monitoring process of a trial is the use of formal statistical stopping rules. These include various methods aimed at accounting for repeated looks at the data such as the Lan–DeMets technique and stochastic curtailment procedures, as well as **Bayesian** approaches [12, 20, 21] (*see* **Data and Safety Monitoring**).

References

[1] Chamberlain, J. (1984). Planning of screening programs for evaluation and non-randomized approaches to evaluation, in *Screening For Cancer. I-General Principles on Evaluation of Screening for Cancer and Screening for Lung, Bladder and Oral Cancer*, P.C. Prorok & A.B. Miller, eds. UICC Technical Report Series, Vol. 78, International Union Against Cancer, Geneva, pp. 5–17.

[2] Chu, K.C., Smart, C.R. & Tarone, R.E. (1988). Analysis of breast cancer mortality and stage distribution by age for the Health Insurance Plan clinical trial, *Journal of the National Cancer Institute* **80**, 1125–1132.

[3] Connor, R.J. & Prorok, P.C. (1994). Issues in the mortality analysis of randomized controlled trials of cancer screening, *Controlled Clinical Trials* **15**, 81–99.

[4] Connor, R.J., Chu, K.C. & Smart, C.R. (1989). Stage-shift cancer screening model, *Journal of Clinical Epidemiology* **42**, 1083–1095.

[5] Day, N.E., Williams, D.R.R. & Khaw, K.T. (1989). Breast cancer screening programs: the development of

a monitoring and evaluation system, *British Journal of Cancer* **59**, 954–958.

[6] Eddy, D.M. (1980). *Screening for Cancer – Theory, Analysis and Design*. Prentice-Hall, Englewood Cliffs.

[7] Eddy, D.M., Hasselblad, V., McGivney, W. & Hendee, W. (1988). The value of mammography screening in women under age 50 years, *Journal of the American Medical Association* **259**, 1512–1519.

[8] Etzioni, R. & Self, S.G. (1995). On the catch-up time method for analyzing cancer screening trials, *Biometrics* **51**, 31–43.

[9] Etzioni, R.D., Connor, R.J., Prorok, P.C. & Self, S.G. (1995). Design and analysis of cancer screening trials, *Statistical Methods in Medical Research* **4**, 3–17.

[10] Fontana, R.S. (1986). Screening for lung cancer: recent experience in the United States, in *Lung Cancer: Basic and Clinical Aspects*, H.H. Hansen, ed. Martinus Nijhoff, Boston, pp. 91–111.

[11] Freedman, L.S. & Green, S.B. (1990). Statistical designs for investigating several interventions in the same study: methods for cancer prevention trials, *Journal of the National Cancer Institute* **82**, 910–914.

[12] Freedman, L.S. & Spiegelhalter, D.J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials, *Controlled Clinical Trials* **10**, 357–367.

[13] Frisell, J., Eklund, G., Hellstrom, L., Glas, U. & Somell, A. (1989). The Stockholm breast cancer screening trial – 5-year results and stage at discovery, *Breast Cancer Research Treatment* **13**, 79–87.

[14] Gilbertsen, V.A., Church, T.R., Grewe, F.A., Mandel, J.S., McHugh, R.M., Schuman, L.M. & Williams, S.E. (1980). The design of a study to assess occult-blood screening for colon cancer, *Journal of Chronic Diseases* **33**, 107–114.

[15] Gohagan, J.K., Prorok, P.C., Kramer, B.S., Cornett, J.E. (1994). Prostate cancer screening in the prostate, lung, colorectal and ovarian cancer screening trial of the National Cancer Institute. *Journal of Urology* **152**, 1905–1909.

[16] Habbema, J.D.F., Lubbe, J.T.N., Van der Maas, P.J. & Van Oortmarssen, G.J. (1983). A computer simulation approach to the evaluation of mass screening, in *Medinfo-83*, Van Bemmel, Ball & Wigertz, eds. IPIF-IMIA, North-Holland, Amsterdam, pp. 1222–1225.

[17] Kafadar, K. & Prorok, P.C. (1994). A data-analytic approach for estimating lead time and screening benefit based on survival curves in randomized cancer screening trials, *Statistics in Medicine* **13**, 569–586.

[18] Kirch, R.L.A. & Klein, M. (1974). Surveillance schedules for medical examinations, *Management Science* **20**, 1403–1409.

[19] Knox, E.G. (1973). A simulation system for screening procedures, in *The Future and Present Indicatives, Problems and Progress in Medical Care*, G. McLachlan, ed. Ninth Series, Nuffield Provincial Hospitals Trust. Oxford University Press, London, pp. 17–55.

[20] Lan, K.K.G. & De Mets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.

[21] Lan, K.K.G., Simon, R. & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials, *Communications in Statistics – Sequential Analysis* **1**, 207–219.

[22] Miller, A.B., Howe, G.R. & Wall, C. (1981). The national study of breast cancer screening, *Clinical and Investigative Medicine* **4**, 227–258.

[23] Morrison, A.S. (1985). *Screening in Chronic Disease*. Oxford University Press, New York.

[24] Moss, S. (1994). Personal communication.

[25] Moss, S., Draper, G.J., Hardcastle, J.D. & Chamberlain, J. (1987). Calculation of sample size in trials of screening for early diagnosis of disease, *International Journal of Epidemiology* **16**, 104–110.

[26] Petronella, P.G.M., Verbeek, A.L.M. & Straatman, H. (1995). Sample size determination for a trial of breast cancer screening under age 50: population versus case mortality approach, *Journal of Medical Screening* **2**, 90–93.

[27] Prorok, P.C. (1984). Evaluation of screening programs for the early detection of cancer, in *Statistical Methods for Cancer Studies*, R.G. Cornell, ed. Marcel Dekker, New York, pp. 267–328.

[28] Prorok, P.C. (1995). Screening studies, in P. Greenwald, B.S. Kramer & D.L. Weed, eds. *Cancer Prevention and Control*, Marcel Dekker, New York, pp. 225–242.

[29] Self, S.G. (1991). An adaptive weighted logrank test with application to cancer prevention and screening trials, *Biometrics* **47**, 975–986.

[30] Self, S.G. & Etzioni, R. (1995). A likelihood ratio test for cancer screening trials, *Biometrics* **51**, 44–50.

[31] Shapiro, S., Goldberg, J.D. & Hutchison, G.B. (1974). Lead time in breast cancer detection and implications for periodicity of screening, *American Journal of Epidemiology* **100**, 357–366.

[32] Shapiro, S., Venet, W., Strax, P. & Venet, L. (1988). *Periodic Screening for Breast Cancer. The Health Insurance Plan Project and Its Sequelae, 1963–1986*. The Johns Hopkins University Press, Baltimore.

[33] Shapiro, S., Venet, W., Strax, P., Venet, L. & Roeser, R. (1982). Ten- to fourteen-year effect of screening on breast cancer mortality, *Journal of the National Cancer Institute* **69**, 349–355.

[34] Shiue, W.K. & Bain, L.J. (1982). Experiment size and power comparisons for two-sample Poisson tests. *Applied Statistics* **31**, 130–134.

[35] Tabar, L., Fagerberg, G., Duffy, S.W., Day, N.E., Gad, A. & Grontoft, O. (1992). Update of the Swedish two-county program of mammographic screening for breast cancer, *Radiologic Clinics of North America* **30**, 187–210.

[36] Taylor, W.F. & Fontana, R.S. (1972). Biometric design of the Mayo lung project for early detection and localization of bronchogenic carcinoma, *Cancer* **30**, 1344–1347.

[37] van Oortmarssen, G.J., Habbema, J.D.F., van der Maas, P.J., de Koning, H.J., Collette, H.J.A., Verbeek, A.L.M., Geerts, A.T. & Lubbe, K.T.N. (1990). A model for breast cancer screening, *Cancer* **66**, 1601–1612.

[38] Walter, S.D. & Day, N.E. (1983). Estimation of the duration of a preclinical disease state using screening data, *American Journal of Epidemiology* **118**, 865–886.

[39] Walter, S.D. & Stitt, L.W. (1987). Evaluating the survival of cancer cases detected by screening, *Statistics in Medicine* **6**, 885–900.

[40] Xu, J.L. & Prorok, P.C. (1995). Non-parametric estimation of the postlead time survival distribution of screen detected cancer cases, *Statistics in Medicine* **14**, 2715–2725.

[41] Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic diseases, *Biometrika* **56**, 601–613.

[42] Zucker, D.M. & Lakatos, E. (1990). Weighted logrank-type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment, *Biometrika* **77**, 853–864.

PHILIP C. PROROK

Screening, Models of

Screening asymptomatic people to allow the early detection and treatment of chronic diseases is an important part of modern medicine and public health. For screening to be both an efficient and cost-effective medical intervention, it must be carefully targeted and evaluated. Mathematical models of disease screening constitute one of the major tools in the design and evaluation of screening programs.

The purpose of this article is to describe models for disease screening and how they have developed in recent years. The discussion will focus on screening for cancer, because most of the methodologic advances in screening design and evaluation have concerned cancer screening. In the first part of the article we will describe the characteristics of these models and illustrate them with a discussion of a simple screening model. In the second part we will describe the development of the two main types of model. In the third part we will discuss model fitting and validation, and in the final part we will briefly describe models for diseases other than cancer and discuss the current state and possible future directions for models of disease screening.

This is not intended to be an exhaustive study of all modeling of disease screening. Rather, it is intended to be a description of the main approaches used and their strengths and weaknesses. For more detailed reviews of modeling disease screening, see Eddy & Shwartz [30], Shwartz & Plough [56], Prorok [50, 51], Alexander [5], and Baker et al. [9].

What is Screening?

Screening for disease control can be defined as the examination of asymptomatic people in order to classify them as likely or unlikely to have the disease that is the object of screening. People identified by a screening test as likely to have the disease are then further investigated to arrive at a final diagnosis [45]. The objective of screening is the early detection of a disease where early treatment is either easier or more effective than later treatment.

Figure 1 is a schematic representation of the main features of the natural history of a disease which are relevant to screening. The *preclinical* phase of the disease is the phase in which a person has the disease

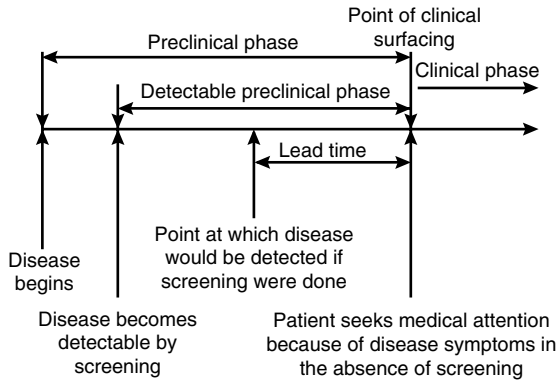

Figure 1 The natural history of a disease with and without screening

but does not have any clinical symptoms and is not yet aware of having it. Screening aims to detect the disease during this phase. In principle, the preclinical phase starts with the beginning of the disease, but, in practice, modeling focuses on the phase commencing at the earliest point at which the disease is detectable with a screening test. This is known as the *detectable preclinical phase*.

The preclinical phase finishes with the *clinical surfacing* of the disease. This is the point at which the person develops clinical symptoms of the disease, seeks medical attention for these symptoms, and the disease is diagnosed. The disease then enters the *clinical phase*, where the person has a diagnosable case of the disease.

The outcome of a screening test is designated either *positive*, if the person is identified as likely to have the disease, or *negative* if they are not. All screening tests are open to error either from the test itself or its interpretation. These errors are designated as **false positive**, where a person without the disease has a positive screening result, and **false negative**, where a person with the disease has a negative screening result. The **sensitivity** of a screening test is the probability that a person with the disease has a positive screening result. The **specificity** of a screening test is the probability that a person without the disease has a negative screening result. Cases of the disease which clinically surface following a false negative result (i.e. where the screening test missed the disease) are known as *interval cases*.

It is important to note that sensitivity and specificity are not properties of the test alone. For example,

mammography is used to screen for breast cancer in women. In this case the sensitivity and specificity will depend on characteristics of the test, such as the nature of the mammography machine and the number of views taken, as well as on factors such as the skill of the person interpreting the mammogram, the size of any tumor in the woman being screened, the density of her breast tissue, and so on.

The *reliability* of a test is its capacity to give the same result, either positive or negative, on repeated application in a person with a given level of the disease. The *survival time* is the length of time between disease diagnosis, either by clinical surfacing or detection by screening, and death. The *lead time* is the time between the detection of a disease by screening and the point at which it would have clinically surfaced in the absence of screening.

The lead time is an important issue in the examination of screening benefits. The immediate focus of screening is to detect an early form of the disease. Hence the lead time can be used as an index of benefit in its own right. It is also important in examining survival benefits conferred by screening. A simple comparison of survival times between screened and unscreened populations is likely to show spurious screening benefits, since the survival time for a screen detected disease includes the lead time while that for a disease which surfaced clinically does not.

There is another, more subtle, reason why such survival comparisons may be spurious, even if adjusted for lead time. Screening will tend to detect people with a longer preclinical phase. This is known as **length-biased** sampling. Usually this will equate to a more slowly progressing disease. Since the disease behavior before clinically surfacing is likely to be correlated with that after surfacing, this is likely to result in screen detected diseases having a longer survival time than clinically surfacing diseases.

Why Use Modeling?

The evaluation of screening usually focuses on whether or not the screening program has led to a fall in mortality from the disease in question. As with most medical interventions, randomized controlled trials (RCT) (*see* **Clinical Trials, Overview**) provide the most satisfactory empirical basis for evaluating screening programs. However, they do have significant limitations.

RCTs for screening are expensive and time-consuming to run – typically requiring very large sample sizes and having long time lags until benefits are apparent (*see* **Screening Trials**). For example, the RCT of mammography screening carried out in the two Swedish counties of Kopparberg and Ostergotland had a total sample size of 134 867. A statistically significant mortality differential between the control and study groups did not appear until after six years of follow-up, with a further four years of follow-up before the results could be considered definitive [58]. Twenty years of data would be required to yield results on some aspects of screening program design [21].

Any one trial cannot address all the issues involved in designing a screening program. For example, the Minnesota Colon Cancer Control Study used an RCT to demonstrate a statistically significant fall in mortality due to screening with a Fecal Occult Blood Test (FOBT), followed by colonoscopy in those with a positive screen [43]. However, Lang & Ransohoff [39] have subsequently suggested that the sensitivity of FOBT is considerably less than that reported in the Minnesota study. FOBT has a high false positive rate, and they argue that one-third to one-half of the fall in mortality could be due to chance selection for colonoscopy where an early cancer or large adenomatous polyp is present but not bleeding and the FOBT is positive for other reasons. The original RCT provides no basis for deciding on the role of FOBT separately from that of colonoscopy.

Models are one way in which the information on the disease and screening tests from a number of different sources – including RCTs and other clinical and epidemiologic research – can be combined with known and hypothesized features of the specific population to be screened. They can be used to investigate the effect of different screening regimes on different subgroups of the population, both on disease mortality and program costs. For example, one use of modeling has been to investigate the inclusion of different age groups in the population to be screened. They can also be used to project the future course of the disease and screening program, to evaluate the changes in costs and benefits over time.

The modeling approach does have limitations. The extra information is obtained from models only by imposing assumptions about the screening process. These include assumptions about the natural history of the disease, about the characteristics of the

screening test and about the behavior of the population under study. These assumptions can only rarely be verified, although they can be evaluated as part of the modeling process.

A further complication in making these assumptions is that the natural history of most diseases is not completely understood, particularly in the asymptomatic preclinical phase, which is the main focus of screening. This means that one may hypothesize a disease model that meets the constraints of current knowledge but which is still ultimately misleading.

Characteristics of Screening Models

Types of Model

Bross et al. [14] proposed a classification of models used to analyze screening strategies into two types: *surface models* and *deep models*. Surface models consider only those events that can be directly observed, such as disease incidence, prevalence, and mortality. Deep models, on the other hand, incorporate hypotheses about the disease process that generates the observed events. Their intent is to use the surface events as a basis for understanding the underlying disease dynamics. This implies models that explicitly describe the disease natural history underlying the observed incidence and mortality.

Deep modeling permits generalization from the particular set of circumstances that generated the surface events. As a result, whereas surface models provide a basis for interpreting the observable effects of screening, deep models provide an explicit basis for determining the outcomes of screening scenarios that have not been directly studied in clinical trials [56]. This article will focus on the application of deep models to population screening.

These models can be further grouped into two broad categories – those that describe the system dynamics mathematically and those that entail computer simulation. The first of these, designated *analytic* models, uses a model of the disease to derive direct estimates of characteristics of the screening procedure and its consequent benefits. The second, designated **simulation** models, uses the disease model to simulate the course of the disease in a hypothetical population with and without screening and derives measures of the benefit of screening from the simulation outcomes.

Markov Framework for Modeling

Most screening models use an illness–death model for the disease which is developed within the framework of a **Markov chain**. A sequence of **random variables** $\{X_k, k = 0, 1, \ldots\}$ is called a Markov chain if, for every collection of integers $k_0 < k_1 < \cdots < k_n < \nu$,

$$\Pr(X_\nu = i | X_{k_0}, \ldots, X_{k_n}) = \Pr(X_\nu = i | X_{k_n}), \quad \text{for all } i. \tag{1}$$

In other words, given the present state (X_{k_n}), the outcome in the future $(X_\nu = i)$ is not dependent on the past $(X_{k_0}, \ldots, X_{k_{n-1}})$.

The Markov chain formulation is applied to an illness–death model in the following way [16]. The population under study is classified into n states, the first m of which are *illness states* and the remaining $n - m$ of which are *death states*. An *illness* state can be broadly defined to be the absence of illness (a *healthy state*), a single specific disease or stage of disease, or any combination of diseases. In modeling screening, these states typically refer to a healthy state and preclinical and clinical phases of the disease.

A *death* state is defined by **cause of death**, either single or multiple. Emigration or loss to follow-up may also be treated as a death state. In modeling screening, typically there will be one death state due to death from the disease and another due to death from any other competing cause (*see* **Competing Risks**). Entry to a terminal stage of the disease is also sometimes treated as a death state. Transition from one state to another is determined by the *transition probabilities*, p_{ij}, where

$$p_{ij} = \Pr(X_{k+1} = j | X_k = i), \quad i, j = 1, 2, \ldots, n; \quad k = 1, 2, \ldots. \tag{2}$$

Death states are *absorbing* states, since once one reaches that state, transition to any other state is impossible (i.e. $p_{ij} = 0$, for $i = m + 1, \ldots, n$, and $j \neq i$). The disease model is said to be *progressive* if, once one enters the first stage of the disease, in the absence of interventions (such as screening) and competing risks, the only valid transitions are through the remaining disease stages. Because the disease is modeled using a Markov chain, the future path of an individual through the illness and death states depends only on his or her current state, and the future distribution of individuals between illness and death

states depends only on the present distribution and not on any past distributions.

This basic model can be varied in a number of ways. The Markov chain treats time as increasing in discrete steps corresponding to the index k. Thus a transition between states can only occur at discrete time intervals. Most screening models extend this to allow transitions to occur in continuous time. In this case, the transition probabilities for any two points in time t_1 and t_2 are

$$p_{ij}(t_1, t_2) = \Pr(X(t_2) = j | X(t_1) = i), \quad i, j = 1, 2, \ldots, n. \tag{3}$$

If $p_{ij}(t_1, t_2)$ only depends on the difference $t_2 - t_1$ but not on t_1 or t_2 separately, the model is *time homogeneous*. The simple Markov chain described above is time homogeneous. This can be varied to allow the transition probabilities to vary with time. The probabilities can also be allowed to vary with age and other relevant characteristics of the individual. Some of the model formulations also allow the probability of transition out of a state to depend on the sojourn time in that state.

A Simple Disease and Screening Model

In this section we describe a simple model presented (and discussed in greater detail) by Shwartz & Plough [56], based on a characterization of the disease process proposed by Zelen & Feinleib [65]. We assume that a person can be in one of three states – a healthy state, the preclinical phase of the disease, or its clinical phase. This characterization also implicitly assumes a death state following the clinical phase, but since the focus of the analysis is on the preclinical phase, the death state is not explicitly used.

The model is progressive in that once a person enters the preclinical state, in the absence of screening or death from another cause, the disease will ultimately surface and enter the clinical phase. If the person is screened while in the preclinical state, then the disease may be detected with a probability depending on the sensitivity of the screening test.

The main assumption underlying this model (and the whole screening process) is that the earlier in the preclinical phase the disease is found, the better will be the prognosis. Hence, the screening benefit is directly related to the lead time.

For this model we define the following:

1. L is the lead time;
2. $g(y)$ is the **hazard rate** for entering the preclinical state at age y;
3. $p(t)$ is the hazard rate for clinical surfacing after the disease has been in the preclinical phase for time t;
4. $f(t)$ is the false negative rate of the screen when the disease has been present for time t; and
5. $b(t)$ is the probability of ultimately dying from the disease if it is detected when it has been present for time t.

For simplicity, we ignore the possibility of death from other causes.

If we let m and σ^2 be the **mean** and **variance** of the sojourn time distribution, then Zelen & Feinleib [65] show that if we assume a constant hazard rate for disease initiation (i.e. $g(t) = g$) we obtain the following expression for the mean lead time:

$$\mathrm{E}(L) = \frac{m^2 + \sigma^2}{2m} = \frac{m}{2}\left[1 + \left(\frac{\sigma^2}{m}\right)\right]. \tag{4}$$

Note that $\mathrm{E}(L) > m/2$ for $\sigma^2 > 0$. This illustrates the effect of length-biased sampling, since, if the screen detected cases were selected at random from all of the cases, one would expect the mean lead time to be $m/2$.

This expression also illustrates one of the central difficulties with this form of modeling. The lead time, which is the main index of screening benefit, is a function of the distribution of the sojourn time in the preclinical phase (*see* **Screening, Sojourn Time**). However, the preclinical phase is, by definition, unobservable. The question of how to estimate characteristics of the sojourn time distribution has been at the center of most of the work done in this area.

For a person to be in the preclinical state at age a, then they must have entered the preclinical state before age a and not leave it until after age a. Hence the probability of this is a function of the hazard rates $g(\cdot)$ and $p(\cdot)$. Thus

$$\Pr(\text{preclinical phase at age } a) = \int_0^a g(u) \exp[-G(u)] \exp[-P(a-u)]\, \mathrm{d}u. \tag{5}$$

Furthermore, the probability that the disease clinically surfaces in some time interval δa following a is

$$\Pr(\text{clinical surfacing in } (a, \delta a)) = \int_0^a g(u)\exp[-G(u)]\exp[-P(a-u)] \times p(a-u)\delta a\,\mathrm{d}u. \quad (6)$$

We combine this with the prognosis measure $b(\cdot)$ to calculate a baseline probability of death from the disease in the absence of screening:

$$\Pr(\text{death in the absence of screening}) = \int_0^\infty \int_0^a g(u)\exp[-G(u)]\exp[-P(a-u)] \times p(a-u)\delta a b(a-u)\,\mathrm{d}u\,\mathrm{d}a. \quad (7)$$

Now we introduce the effect of screening. We will consider the case of one screening test performed at age s. There are four possibilities:

1. the disease is detected by the screening test;
2. the disease clinically surfaces before the test (i.e. at age $a < s$);
3. the disease is missed by the screening test and clinically surfaces after the screen (i.e. it is an interval case); or
4. the disease both enters the preclinical phase and clinically surfaces after the screening test.

For the disease to be detected by this test, it must be in the preclinical phase and the test must not give rise to a false negative. The probability of this is

$$\Pr(\text{disease detection at age } s) = \int_0^s g(u)\exp[-G(u)]\exp[-P(s-u)] \times (1 - f(s-u))\,\mathrm{d}u. \quad (8)$$

We have already calculated the probability that the disease clinically surfaces at age $a < s$ in (6). For the disease to have been missed by the screen, the person must be in the preclinical state at age s, the test must have produced a false negative, and the disease must have clinically surfaced after the screen. The probability of this is

$$\Pr(\text{disease missed by test}) = \int_s^\infty \int_0^s g(u)\exp[-G(u)]\exp[-P(a-u)] \times f(s-u)p(a-u)\delta a\,\mathrm{d}u\,\mathrm{d}a. \quad (9)$$

The probability that the disease both enters the preclinical phase and clinically surfaces after the screening test is

$$\Pr(\text{disease both develops and surfaces after the screen}) = \int_s^\infty \int_s^\infty g(u)\exp[-G(u)] \times \exp[-P(a-u)]p(a-u)\delta a\,\mathrm{d}u\,\mathrm{d}a. \quad (10)$$

Once again we can combine these probabilities with our prognosis measure to obtain the probability of death from the disease in the presence of screening:

$$\begin{aligned}\Pr(\text{death in the presence of screening}) &= \int_0^s g(u)\exp[-G(u)]\exp[-P(s-u)] \times (1 - f(s-u))b(s-u)\,\mathrm{d}u \\ &+ \int_0^s \int_0^a g(u)\exp[-G(u)]\exp[-P(a-u)] \times p(a-u)\delta a b(a-u)\,\mathrm{d}u\,\mathrm{d}a \\ &+ \int_s^\infty \int_0^s g(u)\exp[-G(u)]\exp[-P(a-u)] \times f(s-u)p(a-u)\delta a b(a-u)\,\mathrm{d}u\,\mathrm{d}a \\ &+ \int_s^\infty \int_s^\infty g(u)\exp[-G(u)]\exp[-P(a-u)] \times p(a-u)\delta a b(a-u)\,\mathrm{d}u\,\mathrm{d}a. \quad (11)\end{aligned}$$

This expression gives us our screening figure to compare with the baseline figure in (7).

Although none of the models used for disease screening is exactly like the simple model presented here, they all incorporate its fundamental ideas. In particular, they all depend on knowing in one form or another the transition probabilities into and out of the preclinical state, the distribution of the sojourn time in the preclinical state, the sensitivity of the screening test, and the disease prognosis as a function of the development of the disease (*see* **Natural History Study of Prognosis**).

Analytic Models for Cancer

A mathematical disease model with two states was first proposed by Du Pasquier [24], but it was Fix & Neyman [32] who introduced the stochastic version and resolved many problems associated with the model (*see* **Fix–Neyman Process**). Their model has

two illness states – the state of "leading a normal life" and the state of being under treatment for cancer – and two death states – deaths from cancer and deaths from other causes or cases lost to observation. Chiang [15] subsequently developed a general illness–death stochastic model which could accommodate any finite number of illness and death states (*see* **Stochastic Processes**). Some of the major analytic models developed for cancer screening are listed in Table 1.

Lincoln & Weiss [41] were the first to propose a model of cancer as a basis for analyzing serial screening, in this case screening for cervical cancer. They did not explicitly use a Markov framework, but their model implicitly uses a classification of the disease into illness states.

Zelen & Feinleib [65] proposed the simple three-state, continuous-time, progressive disease characterization described in the previous section and used it in model screening for breast cancer. In a modification to this basic model, the authors further divide the preclinical state into two parts, defined as:

1. a preclinical state in which the disease never progresses to the clinical state (i.e. the sojourn time is allowed to be infinite); and
2. a preclinical state in which the disease is progressive and will eventually progress to the clinical state.

These are used to allow for the possibility that some individuals with the disease in a preclinical state will never have the disease progressing to a clinical state. This approach has been generalized in a number of ways by subsequent authors, with most focusing on simple disease models and the estimation of specific screening characteristics.

Prorok [48, 49] extended the lead time estimation to multiple screens. Blumenson [10–12] calculated the probability of terminal disease as a function of disease duration to date, and used this as a prognostic measure to evaluate screening strategies. Shwartz [54, 55] modeled disease progression for breast cancer using tumor size and number of axillary lymph nodes involved to define the preclinical and clinical states. He then determined screening benefit measures (*see* **Screening Benefit, Evaluation of**), from data on five year survival rate and five year disease recurrence rate for patients, as a function of tumor size and lymph node involvement.

Albert and his co-workers [3, 4, 42] developed a comprehensive model for the evolution of the natural history of cancer in a population subject to screening and natural demographic forces. In its general formulation, the model uses Zelen & Feinleib's classification of the disease into preclinical and clinical phases, but divides the preclinical phase into states corresponding with prognostic tumor staging schemes. It also has two death states which correspond to clinical surfacing of the disease or death from a competing risk. The model is progressive, but allowance is made for staying indefinitely in any given state.

This model is then applied to breast and cervical cancer. Breast cancer is modeled with two illness states, state 1 corresponding to disease with no lymphatic involvement and state 2 corresponding to disseminated disease (the contrary case). Cervical cancer is modeled with three illness states, state 1 corresponding to neoplasms *in situ*, state 2 corresponding to occult invasive lesions, and state 3 corresponding to frankly invasive lesions. The authors then impose on this model a screening strategy with a particular probability of a positive screen, depending on a person's age and disease state. Using this, they derive equations describing how the natural history of cancer (depicted by the distribution of numbers in each state and associated sojourn times) evolves over time in the presence of screening. These, in turn, are used to derive equations for measures of benefit from screening in terms of the disease status. These benefit measures include the percentage reduction in the cumulative number of observed cases of late disease due to screening and the percentage decrease in lost "salvageables" due to screening. A salvageable is a person who would have benefited from screening but who, in the absence of screening, progresses to a late stage of the disease before discovery.

Dubin [25, 26] developed a general multistage disease model similar to that of Chiang [15], and applied this to breast cancer using the same two stage classification as Albert et al. [4]. He noted the difficulty in estimating parameter values for detailed disease models from existing data from screening programs. His model aimed to avoid these difficulties by maintaining comparability between the model and the observable characteristics of a screened population. He did this by focusing on age and stage-specific incidence and survival times in the presence and absence of screening. He derived formulas for the proportion

Table 1 Selected analytic models of cancer screening

Literature references for model	Model inputs	Key features	Model output/measures of screening benefit
Lincoln & Weiss [41]	The probability density for the beginning of the detectable preclinical phase and the probability of a false negative screen at time t after entering the detectable preclinical phase – calculated by assuming specific functional forms rather than by direct estimation	Two illness states – a "healthy" state, in which the disease is not detectable, and a state covering the time between when the disease is first detectable and when it is actually detected by a screening examination	Distribution of time to discovery of tumor
		Symptoms assumed never to appear, with all disease detected by screening	
		Model applied to cervical cancer screening	
Zelen & Feinleib [65]	Disease prevalence and incidence data	Progressive three-state illness model – a healthy state, the preclinical phase, and the clinical phase	Mean lead time
		Assumes single screen	
		Applied to breast cancer screening	
Prorok [48, 49]	Preclinical state sojourn time distribution and disease prevalence and incidence data	Uses the Zelen & Feinleib illness model and develops theory for application to multiple screens	Mean lead time and proportion of preclinical cases detected
Blumenson [10–12]	Disease incidence data	Similar three-state illness model to Zelen & Feinleib, with a point occurring in either the preclinical or clinical phase where the disease becomes incurable	Number of cases of diseases becoming incurable before detection
	Preclinical state sojourn time distribution	Applied to breast cancer screening	
	Screening parameters including age at first screen, screening sensitivity, and screening interval		

(*continued overleaf*)

Table 1 (*continued*)

Literature references for model	Model inputs	Key features	Model output/measures of screening benefit
Shwartz [54, 55]	Specific functional forms and associated parameters governing tumor growth rate and lymph node involvement – chosen to be consistent with published results and available data	Model developed specifically for breast cancer	Changes in life expectancy as a result of screening
	Breast cancer incidence and death rates and death rates from other causes	Progressive illness model with a healthy state, 21 disease states defined in terms of the tumor size and lymph node involvement and two death states – death from breast cancer and death from any other cause	The probability that there will be no disease recurrence and the probability of detection before nodal involvement
	Screening parameters	Transition from preclinical phase to clinical phase possible in any disease state, with probability dependent on tumor size and tumor rate of growth	The probability of disease detection before death from other causes
		Model predictions validated against independent data source (third-order validation)	
Albert et al. [3, 4], Louis et al. [42]	Maximum likelihood estimation of model parameters based on screening data and model assumption	Progressive illness model with preclinical phase classified into states corresponding with prognostic tumor staging schemes, and two "death" states – one corresponding to clinical surfacing and one to death from a competing cause	Percentage reduction due to screening in observed cases of late disease
		Model of screening strategy with probability of positive screen depending on person's age and disease state	Percentage decrease in lost salvageables due to screening
		Applied to breast and cervical cancer screening	
Dubin [25, 26]	Age and stage-specific disease incidence	Aimed at maintaining comparability between the model and observable characteristics of a screened population	Increase in life expectancy
	Survival times in the presence and absence of screening – derived from screening data by assuming particular functional forms for survival distributions	Progressive illness model consisting of disease stages corresponding with prognostic tumor staging schemes	Reduction in probability of dying of breast cancer
		Applied to breast cancer	Reduction in life years lost to women dying of breast cancer
			Mean lead time

Day & Walter [22], Walter & Day [63], Walter & Stitt [64]	Disease incidence derived from screening data	Progressive three-state illness model – a "healthy" state with no detectable disease, the detectable preclinical phase, and the clinical phase	Lead time
	Probability distribution specified for preclinical state sojourn time and survival time	Focus on sojourn time in detectable preclinical phase and survival times after detection	Survival time after detection by screening
		Applied to breast cancer	
Coppleson & Brown [20]	Age-specific clinical incidence data and prevalence data derived from detection rates at first Pap smear	Four-state illness model developed for cervical cancer	Not applicable (model focused on examination of disease natural history)
		Found that observed data could not be explained without allowing for cancer regression in the illness model	
Albert [2]	Transition probability matrix for movement between model states estimated from numbers of cancers detected for each stage in a screening program	Four-state illness model developed for cervical cancer – a healthy state, two preclinical states, and a clinical state.	Not applicable (model focused on examination of disease natural history)
		Cancer regression allowed in the two preclinical states	
Brookmeyer & Day [13]	Parameters of sojourn distribution estimated from screening data and data on interval cancer cases	Extends Day & Walter model	Total preclinical phase sojourn distribution
		Preclinical phase divided into two states – one in which the disease may progress or regress and a second in which the disease always progresses	Screening false negative rate
		Applied to cervical cancer	
van Oortmarssen & Habbema [59]	Parameters of sojourn distribution estimated from screening data and data on interval cancer cases	Similar illness model to Brookmeyer & Day	Not applicable (model focused on examination of disease natural history)
		Applied to cervical cancer	

(*continued overleaf*)

Table 1 (*continued*)

Literature references for model	Model inputs	Key features	Model output/measures of screening benefit
Eddy [27, 29, 30]	Model parameters derived from published results of disease studies and screening programs	Five-stage combined disease and screening model – one healthy state, three preclinical states defined by detectability by screening, and one clinical state	Probability of disease detection
			Probability of death following detection
			Increase of life expectancy due to screening
		Assumes that once a disease is detectable by a screening modality, then any screen using that modality will detect the disease	
		Applied to breast, cervix, lung, bladder, and colon cancer	
Connor et al. [19], Chu & Connor [17]	Stage shifts estimated from analysis of a randomized controlled trial of screening	Multistage progressive disease model	Reduction of deaths at a given stage due to screening
		Focus is on estimation of the shift of the disease at detection to an earlier stage or an earlier point in the same stage as a result of screening	Death prognosis of screen detected cancers
		Applied to breast cancer screening	
Baker et al. [9]	Peak time period for mortality comparison selected from results of a randomized controlled trial of screening.	Focuses analysis on period when screening has maximum effect and hence analysis of screening trial results gives rise to more powerful statistical tests	Ratio of cancer mortality between screened and control groups
		Applied to proportional hazards model for survival analysis	
Day & Duffy [21]	Uses known prognostic factors which are available early in a randomized controlled trial of screening to predict subsequent mortality differentials	Users surrogate endpoints for randomized controlled trial of screening to shorten the duration of the trial and increase its power	Tumor size at cancer detection used as a basis for predicting subsequent mortality differentials
		Applied to breast cancer	

of disease incidence which had been diagnosed earlier due to screening than it would have been in the absence of screening, and used these to derive various measures of screening benefit. Dubin's model is not strictly a deep model as defined above. However, although he makes no explicit hypotheses about the rate of disease progression, such hypotheses are implicit in his model.

Day & Walter [22] developed a variation on the simple three-stage model which has been used extensively. The focus of this model is the sojourn time in the detectable preclinical phase, for which a probability distribution is specified. For example, Walter & Day [63], in applying the model to breast cancer, used several alternate distributions, including the **exponential**, the **Weibull**, and a **nonparametric** step function. Under the model assumptions, one may derive expressions for the anticipated **incidence rates** of clinical disease among groups with particular screening histories and for the anticipated **prevalence** of preclinical disease found at the various screening times. One advantage of this model is that it is relatively simple to obtain approximate **confidence intervals** for parameter values. The model was extended by Walter & Stitt [64] to permit evaluation of survival of cancer cases detected by screening.

A useful synthesis of the analytic models described above applied to breast cancer is presented by O'Neill et al. [46].

All of the above are progressive models, but there are some forms of cancer for which the assumption of progression is not appropriate and for which some form of regression is required. These are cancers, such as large bowel cancer and particularly cervical cancer, where screening detects preinvasive or even precancerous lesions [13].

A number of models have attempted to address this. Coppleson & Brown [20] developed a model for cervical cancer and found that the observed data could not be explained without allowing for regression. Albert [2] developed a variation of his earlier model for cervical cancer which allowed for regression from the carcinoma *in situ* stage back to the healthy state. Brookmeyer & Day [13] and van Oortmarssen & Habbema [59] both developed similar extensions to the Day & Walter model to divide the preclinical phase into two. The first stage allows regression to a healthy state, but once a cancer reaches the second stage only progression is allowed.

The Coppleson & Brown, Albert, and van Oortmarssen & Habbema studies provide an interesting variation on the use of these models, in that the aim of the model was not to study cancer screening directly. Rather, the model was used to study the disease dynamics and, in particular, to examine the epidemiologic evidence for the existence of regression in preinvasive cervical cancer.

The models described above follow a common theme of characterizing the disease as a series of states (corresponding to health, the various disease stages, and death), with people moving between the states with certain transition probabilities and/or certain sojourn times. Screening is then evaluated by superimposing on the disease process a screening process with particular screening regimes and screen sensitivity. This is in contrast to the next model, due to Eddy [27], which uses a different strategy.

Eddy's modeling strategy uses a time varying Markov framework. However, he models the **interaction** between the screen and the disease in his basic model. This is a five-stage model defined in terms of three time points. The first is a reference time point t_p. The way in which this is defined varies with the cancer under discussion but, as an example, for breast cancer it is the point at which the disease can first be detected by physical examination. The *occult interval* is then defined as the time interval between this and the point t_M at which the disease is first detectable by screening (e.g. by mammography). The *patient interval* is defined as the time between t_p and the time t_Π at which the patient would actually seek medical care for the lesion. With Eddy's model, t_Π, t_p, and t_M can occur in any order. The important assumption is that once a disease is detectable by a screening modality (i.e. after t_M), then any screen using that modality will always detect the disease. This assumption replaces the assumption commonly made in models of screening that successive screens are independent.

The other two states are a "healthy" state (which includes any preclinical disease which is still undetectable by screening) and a clinical disease state. Eddy models the probability distributions of the occult and patient intervals and uses these to derive formulas for the probabilities of discovering a malignant lesion by screening and by other methods. Eddy's model has been applied to several breast cancer screening data sets as well as to cervical, gastrointestinal, lung, and bladder cancer. It has also

been extended to the case in which there is more than one type of screening test [31].

Finally, there are three recent analytic models which provide interesting variations on screening modeling.

The first of these is the stage shift model [19]. This assumes that the effect of screening is to shift the diagnosis of a cancer from a higher to a lower stage or within a given stage to an earlier time of diagnosis. Connor et al. develop the theory for a randomized controlled trial with equal sized intervention and control groups, but the equations can be modified to allow for proportional number of cases if unequal groups are used. The method of fitting this model requires a completed trial with follow-up that has reached the point at which comparable sets of cancer cases have accumulated in the study and control groups. For most of the discussion, Connor et al. ignore variability associated with the estimation process and the determination of the point at which comparability is reached in order to emphasize the exploratory nature of the analysis. However, they do present simple variance estimates based on the assumption that their data follow a **Poisson distribution**. The need for a completed trial and long follow-up period limits the model's applicability, but it has been used to analyze breast cancer screening data [17].

The second is the peak analysis model [9]. This uses data from a randomized trial to determine the time period during which screening has the maximum effect on mortality. The results of the trial can then be analyzed restricting attention to that time period, providing more powerful statistical tests. For breast cancer screening, for example, this could mean excluding the mortality experience of the first few years after the initiation of screening. A disadvantage of this model is that the selection of the peak time period for the mortality comparison could be regarded as "data-driven" and subject to the usual problems of a *post hoc* analysis [44].

The third is the use of **surrogate endpoints** for RCTs to shorten the duration of the trial and to increase the **power** [21]. Day & Duffy apply this approach to a study comparing breast cancer screening at three yearly and one yearly intervals. Tumor size is the most important variable in predicting survival from breast cancer in the screening context, so they consider the difference in tumor size distribution between the study groups. They show that using this as an index of benefit and projecting expected mortality allows a result after only five years, compared with the 15–20 years required for a trial based on observed mortality. Furthermore, they demonstrate the rather surprising result that the use of surrogate endpoints leads to an increase in the power of the RCT compared with using the observed mortality. While completed trials remain necessary to establish the primary benefits of screening, this approach allows faster and more efficient resolution of subsidiary issues.

Simulation Models for Cancer

Some of the major simulation models developed for cancer screening are listed in Table 2. Knox [34] developed the earliest and most comprehensive simulation model. As with the analytic models, Knox uses a healthy state, a number of illness states and two death states. However, the model involves considerably more illness states, including classifying the disease as a preclinical, early clinical, or late clinical cancer, and further classifying each of these as treated or not treated and each cancer as high or low grade.

Knox defines a transition matrix containing the estimated transfer rates between the various pathological states, modified according to the age of the individual or the duration of the state. He then simulates the evolution of the disease in a hypothetical cohort of study subjects which has similar characteristics to the population that he wishes to study (which, in this case, is the adult female population of England and Wales) using the transition matrix and a standard **life table** to provide the risks of competing causes of death.

Finally, he adds details of the screening procedures to be considered, specifying the clinico-pathological states to which they apply, and their sensitivities and specificities in relation to each, and the transfers between model states which will occur following detection or nondetection. The screening policies are arranged in incremental series, and the results compared with each other and with the results of providing no screening at all. This allows the appraisal of benefits and costs in both absolute and marginal terms.

This model has been applied to both cervical cancer [34] and breast cancer [35]. It illustrates one

Table 2 Selected simulation models of cancer screening

Literature references for model	Model inputs	Key features	Model output/measures of screening benefit
Knox [34, 35]	Model parameters derived from published results of disease studies and screening programs and from the known characteristics of population under study	Cohort simulation model	Simulated mortality and morbidity in the presence of screening under various screening regimes
		Illness model with 26 defined states	
		Transition matrix defined for movements between these states following detection or nondetection of disease in the presence of specified screening procedures	
		Model applied to a hypothetical cohort of study subjects with similar characteristics to the population under study	
		Model applied to both breast and cervical cancer screening	
Knox [36], Knox & Woodman [37]	Model parameters derived from published results of disease studies and screening programs and from the known characteristics of population under study	Cohort simulation model	Simulated reduction in mortality due to screening
		Illness model with two disease states – one in which the disease is susceptible to early detection and full or partial cure and a second in which the disease is incurable	
		Model applied to subjects who have died from cancer but may have been saved if screening had been offered	
		Applied to breast and cervical cancer screening	

(*continued overleaf*)

Table 2 (*continued*)

Literature references for model	Model inputs	Key features	Model output/measures of screening benefit
Parkin [47]	Model parameters derived from published results of disease studies and screening programs and from the known characteristics of population under study	Microsimulation model developed for cervical cancer screening	Simulated mortality and morbidity in the presence of screening under various screening regimes
		Illness model has nine states – a healthy state, three preclinical states, one clinical state, two death states, and a hysterectomy state (in which a woman is no longer at risk of cervical cancer)	
		Model applied to a hypothetical population with age structure similar to that of the population under study	
Habbema et al. [33], van Oortmarssen et al. [61]	Model parameters derived from published results of disease studies and screening programs and from the known characteristics of population under study	General framework for microsimulation modeling	Simulated mortality and morbidity in the presence of screening under various screening regimes
		Follows similar approach to Parkin	
		Applied to breast and cervical cancer screening	

major difference between the analytic and simulation approaches – the greater complexity of the disease and screening models in the simulation case. However, this extra complexity requires more detailed information on the disease dynamics in order to specify the model and this information is often not readily available. Knox [36] says of his earlier work that

> The chief problem of applying the predictions stemmed from uncertainties about the clinical course of the early stages of cancer.

In this and in all his subsequent analyses, he simplified his model to one with only two illness states. This two-state model is worth discussing in detail because of its different approach to the population under study. Whereas the usual approach is to consider all people at risk of a cancer and to use the model to project mortality with and without screening, Knox's approach is to consider only those who have died from cancer, and to use the model to estimate how many would have been saved if screening had been offered. He refers to it as "tearing down" a graph of age-distributed deaths in successive steps through the insertion of screening procedures at selected ages [37]. This means that Knox does not need to consider variations in the course of the disease, such as lesions which never clinically surface or which regress to a healthy state, because all members of his population have, by definition, a progressive form of the disease.

The two illness states are designated A and B. During state A the disease is susceptible to early detection and full or partial cure. During state B, the disease is incurable. The sojourn time in each state varies around an age-specific mean. The screening procedure has a probability of detecting the lesion which rises linearly during period A, while the probability of curing the disease falls linearly during A.

This model has the advantage of simplicity, which means that it is relatively easy to find plausible parameter values for it. However, this simplicity has disadvantages. The model only considers the situation of a fully established screening program, so that it cannot be used to investigate issues surrounding setting up a new program. Also, because it is focused on mortality reduction, it cannot be used to consider issues relating to costs of screening programs.

Researchers at the Australian Institute of Health and Welfare have extended this approach by combining Knox's disease model with a costs model to evaluate the introduction of breast and cervical cancer screening programs in Australia [6, 7]. They have also combined the disease model with mortality projections to investigate the timing of mortality reductions due to the introduction of a breast cancer screening program [8].

Parkin [47] identifies a number of advantages of the *cohort simulation* approach of transferring year by year specified proportions of a single cohort in a deterministic fashion between model states. These include the model's ability to:

1. demonstrate the relationships between variables;
2. explore the effects of different acceptance rates and test characteristics on outcome measures;
3. examine the net cost-effectiveness of different screening policies by imputing costs to the different outcomes of screening tests (*see* **Health Economics**); and
4. explore the effect of different theoretic natural histories on the outcome of screening.

However, he also identifies some of the disadvantages of this approach. First, services have to be planned, not for a single cohort over an entire lifespan, but for a very heterogenous population over relatively short time periods. When a screening program providing for testing at certain fixed ages is introduced into a community, only people younger than the starting age for the screening policy can possibly receive the full schedule of tests. Thus, benefits from screening will at first be small, but will increase progressively as more of the population receives a series of examinations. Furthermore, many people will have already had previous examinations, so the results of the screening policy will depend on the existing screening status of the population. This cannot be simulated by a single cohort model; nor can differences in the risk of disease in different birth cohorts.

Secondly, it may be desirable to use characteristics other than age to identify subgroups of the population for selective screening. This is less often of practical use, since such subgroups are usually not readily identifiable, but a planning model should be able to explore the effectiveness of policies involving differential screening of such subpopulations. In addition, population subgroups often have different rates of attendance at screening programs which may be correlated with different disease risks.

Finally, screening programs do not exist in isolation from the rest of the health care system. Much screening activity can take place outside a screening program. Most models usually treat this activity as "diagnostic" and ignore it. However, a planning model should take account of all relevant screening activity.

Parkin proposes instead a *microsimulation* approach. Here, the life histories of individual members of a population are simulated. The population in his model has the demographic make-up of that of England and Wales and its size is governed by two considerations: (i) the computer time involved in microsimulation of very large populations; and (ii) the need for reliable results in a stochastic simulation of relatively rare events.

Each individual is characterized by his or her values for a set of variables which will be used in simulating demographic events, the disease natural history, or screening programs. The values of these variables are updated annually using sets of **conditional transition probabilities** (e.g. the probability of childbirth given age, marital status, and initial parity). The occurrence of a transition is decided by comparing the relevant probability against a randomly generated number. There is considerable flexibility in modeling screening programs and, since the model follows individuals, it is possible to simulate contacts with the health care system and the "incidental" screening which occurs on such occasions.

Parkin's microsimulation model was developed specifically for cervical cancer screening, but a group working at Erasmus University in the Netherlands has developed a general modeling framework for microsimulation modeling of cancer screening called MISCAN (MIcrosimulation SCreening ANalysis) [33, 61]. Strictly speaking, MISCAN is not itself a model, but rather a model generator – a package that can generate and calculate a variety of these microsimulation models.

The MISCAN approach, like Parkin's model, is based on the actual structure of a population as it develops in a given country at a particular time. The mass screening program under consideration is taken as starting in a particular year and finishing in a particular year. Standard demographic techniques (*see* **Demography**) are used to project the study population to a year well after the nominated end of the program. This allows for both the introduction of the program to be modeled and the effects, after the end of the program, to be followed up.

The basic structure of the cancer model is similar to Knox's earlier model with a detailed classification of clinical and preclinical cancer states, although it uses a smaller number of states. The interaction between the disease model and the screening program is designed to allow projection of screening and treatment costs as well as cancer mortality and morbidity. MISCAN has been widely used to analyze breast and cervical cancer screening programs.

Model Fitting and Validation

Eddy [28] proposed four levels of validation for mathematical models (*see* **Model Checking**):

1. First-order validation: this requires that the structure of the model makes sense to people who have a good knowledge of the problem.
2. Second-order validation: this involves comparing estimates made by the model with the data that were used to fit the model.
3. Third-order validation: this involves comparing the predictions of the model with data that were available when the model was fitted but that were not used in the estimation of model parameters.
4. Fourth-order validation: this involves comparing the outcomes of the model with observed data when applied to data generated and collected after the model was built (for example, data from a previously unobserved screening program).

In this section we discuss model fitting and validation for cancer screening in the framework of these levels.

First-order validation is generally not difficult to accomplish. The conceptualization of cancer as a series of preclinical and clinical stages is virtually universally accepted as a reasonable characterization of the disease. Problems may arise when the details of the disease stages are specified, but generally a wide variety of model formulations are plausible within the constraints of the limited knowledge of preclinical cancer.

Second-order validation highlights one of the central problems with this sort of deep model. This is the difficulty of directly relating available data to model parameters. The mismatch between the data available, either from screening trials or other sources,

and the model data requirements for parameter estimation has been recognized from the beginning of this type of modeling. Lincoln & Weiss [41, p. 188] note, for example, that

> Here we can do no more than introduce plausible forms for the different functions involved and plausible values for the parameters.

They go on to describe the difficulties in relating available data to the mathematical functions on which their model is based. This is a recurring problem in modeling cancer for screening, and to some extent affects all of the models described in this article.

Some of the analytic models have developed methods of estimating model parameters using standard statistical **estimation** approaches. Dubin [26], for example, structured his model so that it could directly use the data from screening trials, although as a consequence his model relates less to the disease natural history than do the others. Louis et al. [42] derived nonparametric models for the probability distributions specified in their model and proposed the use of **maximum likelihood** methods to fit them. Day & Walter [22] used both parametric and nonparametric functions for their preclinical sojourn time and suggested either maximum likelihood methods or **least squares** criteria to fit them. However, many of the analytic models and all of the simulation models proceed in a more *ad hoc* fashion by varying their disease natural history and model parameters until their models closely reproduce existing data.

Knox [35], pp. 17–18 gives an example of how this *ad hoc* fitting operates, in fitting his earlier model to breast cancer screening data. He describes fitting the natural history data thus:

> A statement of the natural history of the disease process must be provided in the form of a "transition matrix" which gives estimated transfer rates between the various pathological states, modified suitably according to the age of the woman or the duration of the state. This set of values is adjusted iteratively until an output is produced which matches available data on incidence, prevalence and mortality. If, as sometimes happens, more than one natural history statement is capable of mimicking these facts, then the natural history will have to be treated as one of the uncertainties. Subsequent runs will then have to be repeated for a range of natural history alternatives, and each prediction of results will be conditional upon the accuracy of the natural history used.

Parkin [47] provides an example of just such an uncertainty about natural history, with the final model including three different natural histories as alternatives.

This approach to model fitting has the disadvantage that, particularly for models with a large number of unknown parameters, the fit of the predicted values may be close to the observed data whether or not the model is in any sense valid. However, fitting the model to a number of independent data sets simultaneously and validating it against each of these data sets, as was done by van Oortmarssen et al. [61], provides some protection against this possibility.

Third-order validation is usually made difficult by the lack of data. Generally, most available data are used in determining the parameters of the model [56]. Breast cancer models are a good example of this. The only real data sources for fitting models for breast cancer screening are the screening studies, and in particular the RCTs. The first major study was the Health Insurance Plan of New York study (HIP) [53]. This program started screening in 1963. Subsequent studies were not started for another ten years, with the Utrecht Screening Program [18] starting screening in 1974 and the Swedish Two-county Randomized Trial starting in 1977 [58]. This means that many of the models only had access to the HIP data. Screening technology has changed significantly since the HIP program began [61], so when later studies became available they could not be directly compared with the HIP program and, in any case, it is questionable whether models based only on HIP data are directly relevant to modern screening. Because of the long time before mortality benefits from screening are fully apparent, models fitted using solely data from later studies have only appeared relatively recently [61] and, at least in their published form, have generally not addressed the issue of third-order validation. However, as more screening programs are implemented, more data should become available for third-order validation [6].

Eddy [28] recognized that fourth-order validation is only possible in rare cases. However, there are at least two examples of studies which use models in a way that could be called fourth-order validation, coincidentally both using Eddy's own model. Verbeek et al. [62] compare predictions from Eddy's

model for breast cancer to data from a mammography screening program in Nijmegen. The authors note that the comparison does not suggest too good a fit. However, this is only a preliminary study, and further validation work remains to be done. Eddy [29] compares his model for cervical cancer with a later independent analysis of empirical data. In this case the model appears to predict accurately the effect of different cervical cancer screening policies on outcomes that are important for policy decisions.

The best way to see how these models are fitted and used in practice is to examine examples. The following three sections describe an example of fitting a model followed by a description of its application.

An Example of Model Fitting

This section describes the analysis by van Oortmarssen et al. [61] of breast cancer screening based on the MISCAN computer simulation package. This model is designed to reproduce the detection rates and incidence of interval cancers as observed in the screening projects in Utrecht and Nijmegen in the Netherlands.

The basic model structure is shown in Figure 2. The first state is the state of no breast cancer. Women stay in this state until a transition occurs to one of the preclinical states that is detectable by screening (either mammography or clinical examination). The preclinical phase is divided into four states. There is one preinvasive state, intraductal carcinoma *in situ* (dCIS), and three screen detectable invasive states subdivided according to the diameter of the tumor: <10 mm, 10–19 mm, and ≥ 20 mm.

The subdivision applied to the preclinical invasive states is also used for the clinical phase and for screen detected tumors. The state "false positives" refers to women with a positive screening examination in whom no breast cancer is found at further assessment. The two end states of the model are "death from breast cancer" and "death from other causes". Transitions into the "death from other causes" state (not shown in the figure) are possible from every other model state and are governed by the Dutch life table, which is corrected for death from breast cancer. The values of the key parameters of the model are summarized in Table 3.

Parameters relating to clinical breast cancer and survival can usually be taken directly from available data. In this case, the preclinical incidence was estimated from the reported Dutch clinical incidence figures shifted to younger ages according to the model's assumptions about the transitions and durations in the preclinical stages. The distribution of the tumor diameters for clinically diagnosed cancers was obtained directly from data on cancers diagnosed outside the screening program in Utrecht and Nijmegen. Survival is described by a fraction cured and a survival time distribution for women who are at risk of dying from breast cancer. The survival time distribution is based on the **lognormal**, with mean and variance taken from a published analysis of the Swedish Cancer Registry data [52]. The fraction cured was estimated from the Utrecht data on clinically diagnosed cancers and varied with age according to another published analysis of Swedish data on age-specific breast cancer survival [1]. The combination of model assumptions on clinical incidence, stage distribution, and survival result in a good fit for the mortality rate for breast cancer in the Netherlands at all ages.

Parameters relating to the preclinical phase are less easily specified. Parameter estimation was done by comparing simulated results from the model with data from the Utrecht and Nijmegen projects. An initial set of parameter values, partly taken from an earlier analysis of the HIP screening trial [60], resulted in many discrepancies between the simulated and observed data. The model parameters were systematically varied until a set of model specifications was found which gave an adequate overall fit to the Utrecht and Nijmegen data. Finally, the improvement in prognosis due to screen detection was calculated from the results of the Swedish Two-county screening study [58].

This model passes both first- and second-order validation, in that it is consistent with what is known about the natural history of breast cancer and with previous models developed in the literature, and its results are consistent with the Utrecht and Nijmegen data used in its fitting. Third-order validation is more problematic. As noted above, the HIP data are not directly comparable with those considered here and the authors used all the other available data in fitting the model. Similarly, fourth-order validation is not possible in this case, since published results from other breast cancer RCTs were not available at the time this analysis was carried out.

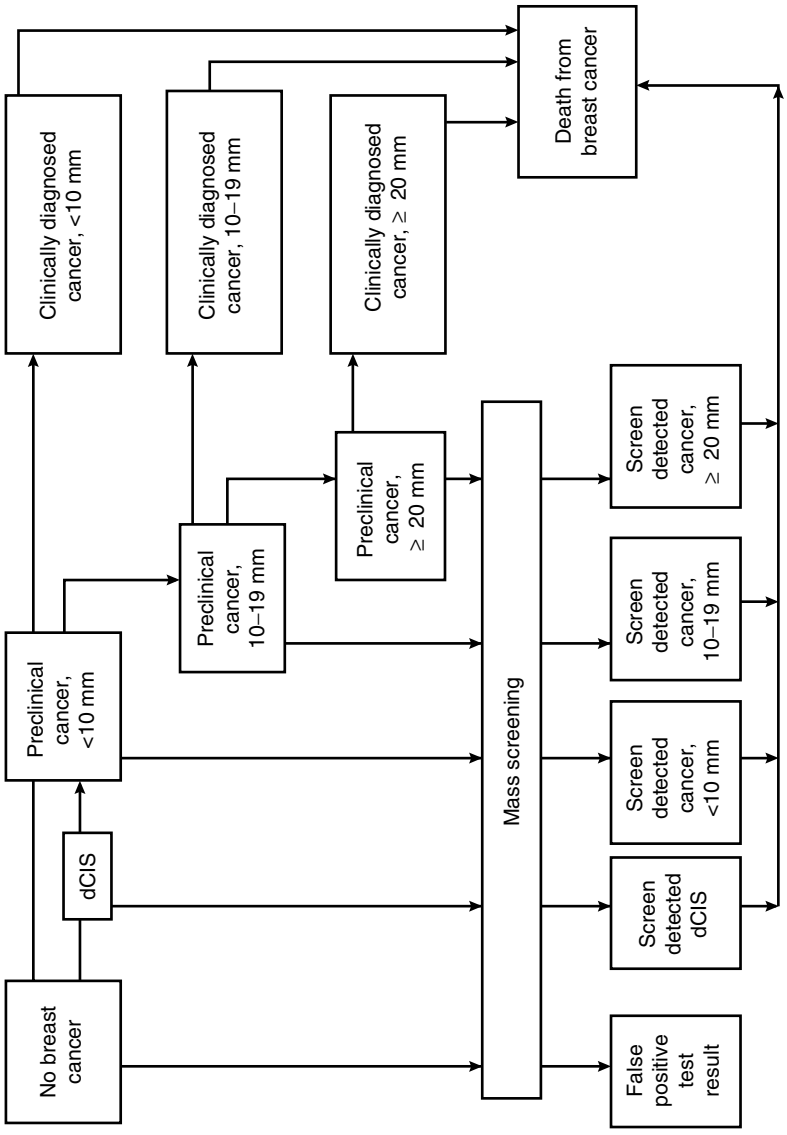

Figure 2 The structure of the disease and screening model for breast cancer developed by van Oortmarssen et al. The state "death from other causes" is not shown. It may be reached from all other states. *Source:* van Oortmarssen et al. [61]

Table 3 Key assumptions of the van Oortmarssen et al. breast cancer screening model

Parameter	Assumption
Preclinical incidence	Based on Dutch clinical incidence, 1977–82
Clinical stage distribution	*Independent of age*
<10 mm	10%
10–19 mm	22%
≥20 mm	68%
20 year survival of clinically diagnosed breast cancer (diagnosis at age 55)	*Age-dependent*
<10 mm	83%
10–19 mm	68%
≥20 mm	51%
Duration of preclinical invasive stages	*Average duration (years)*
Age 40 years	1.6
Age 50 years	2.1
Age 60 years	3.0
Age 70 years	4.7
Sensitivity of mammography	*Independent of age*
dCIS	70%
<10 mm	70%
≥10 mm	95%
Impact of early detection	
Mortality reduction for screen detected cancers	52%

Source: van Oortmarssen et al. [61].

An Application of the Model to Breast Cancer Screening

The breast cancer disease model described above was applied to Australian data by Stevenson et al. [57] to simulate the introduction of a breast cancer screening program. Australian breast cancer data and life table data were used to estimate cancer incidence and population **life expectancies**. Pilot testing of screening programs suggested that a screening participation rate of 70% was a reasonable target [6]. All other model parameters were taken from the van Oortmarssen et al. model.

The model was applied to five different screening options defined in terms of the age group offered screening and the interval between successive screens. These are listed in Table 4. Taking 1990 as the nominal starting year, the analysis simulated the introduction of a screening program phased in over five years and running for a further 25 years. The simulated total life years lost in the absence of a screening program and the life years saved by screening for each of the screening options are listed in Table 5. These results show a clear benefit in including women aged 40–49 in the screening program and of a two year interval over a three year interval. However, they also show that decreasing the interval to one year for women aged 40–49 makes only a marginal improvement.

Table 4 Breast cancer screening options

Option number	Age group screened (years)	Screening interval (years)
1	50–69	2
2	50–69	3
3	40–49	1
	50–69	2
4	40–49	2
	50–69	3
5	40–69	2

An analysis of screening should include consideration of costs as well as benefits. A complete discussion of estimating costs is beyond the scope of this article, but generally they will be based on both current screening experience (with, for example, screening pilot projects in the location under study) and model based projections. These estimates

Table 5 Number and proportion of life years saved among Australian women by mammography screening over a 30 year screening period, as estimated from the van Oortmarssen et al. simulation model

Screening option	Total life years lost in the absence of a screening program ('000s)	Number of life years saved as a result of the screening program ('000s)	Life years saved as a percentage of total life years lost
1	3766.6	250.5	6.7
2	3767.0	202.4	5.4
3	3741.2	324.3	8.7
4	3743.1	258.1	6.9
5	3755.6	309.4	8.2

Source: Stevenson et al. [57].
Note: These results are based on the simulation of individual life histories, with the outcomes for each individual being determined randomly by applying the probabilities of developing the disease and of surviving the disease. This means that the outcome for each individual may vary between simulations. This accounts for the small variation in the simulated total life years lost figures.

Table 6 Relative cost-effectiveness of screening at different screening intervals for women aged 40–69

Screening option	Net present value of costs to service providers and women ($ million)	Net present value of life years saved ('000s)	Average cost per life year saved ($)
3	1917.8	622.2	3082.3
4	1097.5	628.6	1745.9
5	1374.6	620.6	2215.0

Source: Costs data taken from Australian Health Ministers' Advisory Council report on breast cancer screening [6]. Projected life years saved data taken from Stevenson et al. [57].
Note: Net present value calculated by applying an annual discount rate of 5%.

Table 7 Percentage of total life years saved among Australian women by mammography screening as estimated by two simulation models

Screening option	Life years saved as a percentage of total life years lost – van Oortmarssen et al. model	Life years saved as a percentage of total life years lost – Knox two-stage model
1	6.7	12.6
2	5.4	11.1
3	8.7	12.9
4	6.9	11.1
5	8.2	12.8

Source: Stevenson et al. [57].

are usually reported as the present value of the costs. This involves applying an annual discount rate to costs projected for future years. Hence, where costs are compared with benefits, the benefits are usually also presented in present value terms by applying the same annual discount rate.

The estimated total costs and costs per life year saved for the three screening options which include women aged 40–49 are presented in Table 6. This shows that the small increase in life years saved gained by moving to a one year screening interval for women aged 40–49 is offset by a substantial increase in the cost per life year saved.

A Comparison of Two Models for Breast Cancer

Stevenson et al. [57] also simulated the introduction of an Australian breast cancer screening program using Knox's two-state disease model described above. In Table 7 is presented a comparison between the

percentage life years saved for each screening option derived from both this model and the van Oortmarssen et al. model. There are clear differences between the two models, with the Knox model estimates consistently higher for all screening options. Furthermore, the evidence from the Knox model for including women aged 40–49 is more equivocal.

It is tempting to ask which model is right but, while there is some reason for preferring the van Oortmarssen model (because of its more extensive validation), a more relevant question is which model more correctly addresses the issue under study. The Knox model applies to a steady state situation, in which the screening program has been operating for long enough so that no one in the target population is too old to have participated in the full program. The van Oortmarssen et al. model makes allowance for the start of the program excluding some women from fully participating. The effect of this is that the Knox model will overstate the gains in life years saved at the start of the program. The difference in the results for including women aged 40–49 years arise from more realistic assumptions in the van Oortmarssen et al. model about the effect of screening at those ages on subsequent mortality.

Models for Other Diseases

Models for screening can be applied to diseases other than cancer. For example, screening tests exist for diabetes and there is a clear value in its early detection. Undiagnosed diabetes could be considered as a preclinical phase of the disease and modeling techniques applied to investigating its characteristics. Similarly, a disease such as hypertension could be modeled either for its own sake or as a preclinical form of cardiovascular disease.

Some work has been done on simulation modeling for coronary heart disease [38]. This model used **logistic regression** to estimate transition probabilities between risk factor states and heart disease. It focused on the effects of risk factor reduction, but did not address details of screening programs. Hence, it avoided having to model details of the preclinical phase. There have to date been no significant published attempts at modeling the preclinical phase to investigate specific screening programs for chronic diseases other than cancer.

On the other hand, modeling of infectious diseases has a long history in biostatistics (*see* **Communicable Diseases**; **Infectious Disease Models**). Most recently considerable work has been done on disease models of **AIDS and HIV**, although most of this effort has focused on projecting the spread of the disease rather than modeling screening programs (see, for example, Day et al. [23]). However, there has been some work on modeling screening for infectious diseases.

Lee & Pierskalla's model [40] is a good illustration of the similarities and differences in modeling infectious diseases for mass screening. In this model, the preclinical phase equates to the period during which a disease is infectious but without symptoms and the clinical phase to the period during which symptoms develop, the person seeks treatment, and is isolated or removed from the population. The main quantities used in the modeling are:

1. the number of infected people at a given time;
2. the natural incidence rate of the disease;
3. the rate of transmission of the disease from a contagious unit to a susceptible;
4. the rate of infected units ending the infectious period (i.e. clinical surfacing); and
5. the probability that an infected unit will not be detected by a screening test (i.e. the probability of a false negative).

The crucial difference here is that disease is initiated by spread from one unit to another, as well as by its natural incidence rate. Hence, in addition to the lead time, the main index of benefit is the removal of infected units from the population. Indeed, Lee & Pierskalla show that defining the measure of screening benefit as the average lead time across the population under study is equivalent to defining it as the average number of infected units per time period in the population.

Taking treatment as the endpoint of the model, rather than ultimate mortality, has the advantage of avoiding the necessity of modeling survival in the presence of screening. However, these models still have the difficulty of specifying parameters for an unobserved preclinical phase. For example, Lee & Pierskalla note that their model is an oversimplification, because it assumes that the sensitivity of the screening test is constant and independent of how long the person has been infected with the disease.

They also note that varying this assumption is of little practical use, since data on transmission rates at the various disease states are almost nonexistent.

Current State and Future Directions

The problem of model validation and its effect on the credibility of model based results is still a barrier to their wider use. Nevertheless, there are a number of areas in which modeling can make a uniquely important contribution to our current understanding of screening.

In the absence of specific RCTs, modeling remains the only effective way of evaluating different screening regimes. For example, the inclusion of women aged between 40 and 50 in a mammography screening program is still a contentious issue, with no international consensus on the effectiveness of screening at these ages [6]. While it could be argued that decisions on screening these women should not be made in the absence of reliable evidence on the presence or absence of the benefits, in practice, governments are already developing screening programs and modeling plays an important role in guiding policy-makers.

Modeling also has a crucial role to play in assessing the cost-effectiveness of screening programs. Even for cheap and easily available screening technologies, organized mass screening programs are the best way to insure that the benefits of screening are fully realized [7]. Modeling is not only necessary in order to plan these programs, but funding bodies are unlikely to fund such programs without at least initial cost–effectiveness studies, and modeling is the only practical way to derive the necessary estimates of future benefits and costs.

Miller et al. [44], p. 768 best summarize the current situation when, in discussing some recent models, they say

> It is clear that these, and other models already developed or under consideration, may enhance our understanding of the natural history of screen-detected lesions and the process of screening. However, they require validation with the best available data, which is preferably derived from randomized trials, before they could be extrapolated in ways that might guide policy decisions. As such data become available, assumption-based models need to be modified to incorporate this extra information, in order to improve the extrapolations needed to make policy.

While analytic models have a role in investigating specific facets of the disease and screening process (see, for example, [59]), the more comprehensive simulation models, and particularly the microsimulation models, seem best suited to the overall assessment of costs and effectiveness in screening programs and the investigation of different screening regimes. However, the challenge in using the simulation approach is to derive disease and screening models which are sufficiently complex to model all relevant aspects of screening but sufficiently simple to enable interpretable second-order validation.

References

[1] Adami, H.A., Malker, B., Holmberg, L., Persson, I. & Stone, B. (1986). The relationship between survival and age at diagnosis in breast cancer, *New England Journal of Medicine* **315**, 559–563.

[2] Albert, A. (1981). Estimated cervical cancer disease state incidence and transition rates, *Journal of the National Cancer Institute* **67**, 571–576.

[3] Albert, A., Gertman, P.M. & Louis, T.A. (1978). Screening for the early detection of cancer – I. The temporal natural history of a progressive disease state, *Mathematical Biosciences* **40**, 1–59.

[4] Albert, A., Gertman, P.M., Louis, T.A. & Liu, S.-I. (1978). Screening for the early detection of cancer – II. The impact of the screening on the natural history of the disease, *Mathematical Biosciences* **40**, 61–109.

[5] Alexander, F.E. (1989). Statistical analysis of population screening, *Medical Laboratory Science* **46**, 255–267.

[6] Australian Health Ministers' Advisory Council (1990). *Breast Cancer Screening in Australia: Future Directions*. *Australian Institute of Health: Prevention Program Evaluation Series* No 1. AGPS, Canberra.

[7] Australian Health Ministers' Advisory Council (1991). *Cervical Cancer Screening in Australia: Options for Change*. *Australian Institute of Health: Prevention Program Evaluation Series* No 2. AGPS, Canberra.

[8] Australian Institute of Health and Welfare (1992). *Australia's Health 1992: the Third Biennial Report of the Australian Institute of Health and Welfare*. AGPS, Canberra.

[9] Baker, S.G., Connor, R.J. & Prorok, P.C. (1991). Recent developments in cancer screening modeling, in *Cancer Screening*, A.B. Miller, J. Chamberlain, N.E. Day, M. Hakama & P.C. Prorok, eds. Cambridge University Press, Cambridge, pp. 404–418.

[10] Blumenson, L.E. (1976). When is screening effective in reducing the death rate? *Mathematical Biosciences* **30**, 273–303.

[11] Blumenson, L.E. (1977). Compromise screening strategies for chronic disease, *Mathematical Biosciences* **34**, 79–94.

[12] Blumenson, L.E. (1977). Detection of disease with periodic screening: Transient analysis and application to mammography examination, *Mathematical Biosciences* **33**, 73–106.

[13] Brookmeyer, R. & Day, N.E. (1987). Two-stage models for the analysis of cancer screening data, *Biometrics* **43**, 657–669.

[14] Bross, I.D.J., Blumenson, L.E., Slack, N.H. & Priore, R.L. (1968). A two disease model for breast cancer, in *Prognostic Factors in Breast Cancer*, A.P.M. Forrest & P.B. Bunkler, eds. Williams & Wilkins, Baltimore, pp. 288–300.

[15] Chiang, C.L. (1964). A stochastic model of competing risks of illness and competing risks of death, in *Stochastic Models in Medicine and Biology*, J. Gurland, ed. University of Wisconsin Press, Madison, pp. 323–354.

[16] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and their Applications*. Krieger, Huntington, New York.

[17] Chu, K.C. & Connor, R.J. (1991). Analysis of the temporal patterns of benefits in the Health Insurance Plan of Greater New York Trial by stage and age, *American Journal of Epidemiology* **133**, 1039–1049.

[18] Collette, H.J.A., Day, N.E., Rombach, J.J. & de Waard, F. (1984). Evaluation of screening for breast cancer in a non-randomized study (the DOM project) by means of a case control study, *Lancet* **i**, 1224–1226.

[19] Connor, R.J., Chu, K.C. & Smart, C.R. (1989). Stage-shift cancer screening model, *Journal of Clinical Epidemiology* **42**, 1083–1095.

[20] Coppleson, L.W. & Brown, B. (1975). Observations on a model of the biology of carcinoma of the cervix: a poor fit between observations and theory, *American Journal of Obstetrics and Gynecology* **122**, 127–136.

[21] Day, N.E. & Duffy, S.W. (1996). Trial design based on surrogate end points – application to comparison of different breast screening frequencies, *Journal of the Royal Statistical Society, Series A* **159**, 49–60.

[22] Day, N.E. & Walter, S.D. (1984). Simplified models of screening for chronic disease: estimation procedures from mass screening programmes, *Biometrics* **40**, 1–14.

[23] Day, N.E., Gore, S.M. & De Angelis, D. (1995). Acquired immune deficiency syndrome predictions for England and Wales (1992–97): sensitivity analysis, information, decision, *Journal of the Royal Statistical Society, Series A* **158**, 505–524.

[24] Du Pasquier, (1913). Mathematische theorie der Invaliditatsversicherung, *Milt. Verein. Schweiz. Versich.-Math.* **8**, 1–153.

[25] Dubin, N. (1979). Benefits of screening for breast cancer: application of a probabilistic model to a breast cancer detection project, *Journal of Chronic Diseases* **32**, 145–151.

[26] Dubin, N. (1981). Predicting the benefit of screening for disease, *Journal of Applied Probability* **18**, 348–360.

[27] Eddy, D.M. (1980). *Screening for Cancer: Theory, Analysis and Design*. Prentice-Hall, Englewood Cliffs.

[28] Eddy, D.M. (1985). Technology assessment: the role of mathematical modeling, in *Assessing Medical Technologies*, Institute of Medicine, ed. National Academy Press, Washington, pp. 144–153.

[29] Eddy, D.M. (1987). The frequency of cervical cancer screening: comparison of a mathematical model with empirical data, *Cancer* **60**, 1117–1122.

[30] Eddy, D.M. & Shwartz, M. (1982). Mathematical models in screening, in *Cancer Epidemiology and Prevention*, D. Schottenfeld & J.F. Fraumeni, eds. Saunders, Philadelphia, pp. 1075–1090.

[31] Eddy, D.M., Nugent, F.W., Eddy, J.F., Coller, J., Gilbertsen, V., Gottleib, L.S., Rice, R., Sherlock, P. & Winawer, S. (1987). Screening for colorectal cancer in a high-risk population, *Gastroenterology* **92**, 682–692.

[32] Fix, E. & Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients, *Human Biology* **23**, 205–241.

[33] Habbema, J.D.F., Lubbe, J.Th.N., van der Maas, P.J. & van Oortmarssen, G.J. (1983). A computer simulation approach to the evaluation of mass screening, in *MEDINFO 83. Proceedings of the 4th World Conference on Medical Informatics*, van Bemmel et al., eds. North-Holland, Amsterdam.

[34] Knox, E.G. (1973). A simulation system for screening procedures, in *Future and Present Indicatives, Problems and Progress in Medical Care, Ninth Series*, G. McLachlan, ed. Nuffield Provincial Hospitals Trust, Oxford, pp. 17–55.

[35] Knox, E.G. (1975). Simulation studies of breast cancer screening programmes, in *Probes for Health*, G. McLachlan, ed. Oxford University Press, London, pp. 13–44.

[36] Knox, E.G. (1988). Evaluation of a proposed breast cancer screening regimen, *British Medical Journal* **297**, 650–654.

[37] Knox, E.G. & Woodman, C.B.J. (1988). Effectiveness of a cancer control programme, *Cancer Surveys* **7**, 379–401.

[38] Kottke, T.E., Gatewood, L.C., Wu, S.C. & Park, H.A. (1988). Preventing heart disease: is treating the high risk sufficient? *Journal of Clinical Epidemiology* **41**, 1083–1093.

[39] Lang, C.A. & Ransohoff, D.F. (1994). Fecal occult blood screening for colorectal cancer – is mortality reduced by chance selection for screening colonoscopy? *Journal of the American Medical Association* **271**, 1011–1013.

[40] Lee, H.L. & Pierskalla, W.P. (1988). Mass screening models for contagious diseases with no latent period, *Operations Research* **36**, 917–928.

[41] Lincoln, T. & Weiss, G.H. (1964). A statistical evaluation of recurrent medical examination, *Operations Research* **12**, 187–205.

[42] Louis, T.A., Albert, A. & Heghinian, S. (1978). Screening for the early detection of cancer – III. Estimation of disease natural history, *Mathematical Biosciences* **40**, 111–144.

[43] Mandel, J.S., Bond, J.H., Church, T.R., Snover, D.C., Bradley, G.M., Schuman, L.M. & Ederer, F. (1993). Reducing mortality from colorectal cancer by screening for fecal occult blood, *New England Journal of Medicine* **328**, 1365–1371.

[44] Miller, A.B., Chamberlain, J., Day, N.E., Hakama, M. & Prorok, P.C. (1990). Report on a workshop of the UICC Project on evaluation of screening for cancer, *International Journal of Cancer* **46**, 761–769.

[45] Morrison, A.S. (1985). *Screening in Chronic Disease*. Oxford University Press, New York.

[46] O'Neill, T.J., Tallis, G.M. & Leppard, P. (1995). A review of the technical features of breast cancer screening illustrated by a specific model using South Australian cancer registry data, *Statistical Methods in Medical Research* **4**, 55–72.

[47] Parkin, D.M. (1985). A computer simulation model for the practical planning of cervical cancer screening programmes, *British Journal of Cancer* **51**, 551–568.

[48] Prorok, P.C. (1976). The theory of periodic screening I: lead time and proportion detected, *Advances in Applied Probability* **8**, 127–143.

[49] Prorok, P.C. (1976). The theory of periodic screening II: doubly bounded recurrence times and mean lead time and detection probability estimation, *Advances in Applied Probability* **8**, 460–476.

[50] Prorok, P.C. (1986). Mathematical models and natural history in cervical cancer screening, in *Screening for Cancer of the Uterine Cervix*, M. Hakama, A.B. Miller & N.E. Day, eds. IARC Scientific Publication, Vol. 76, pp. 185–198.

[51] Prorok, P.C. (1988). Mathematical models of breast cancer screening, in *Screening for Breast Cancer*, N.E. Day & A.B. Miller, eds. Hans Huber, Toronto, pp. 95–109.

[52] Rutqvist, L.E. (1985). On the utility of the lognormal model for analysis of breast cancer survival in Sweden 1961–1973, *British Journal of Cancer* **52**, 875–883.

[53] Shapiro, S., Venet, W., Strax, P., Venet, L. & Roeser, R. (1982). Ten to fourteen year effect of screening on breast cancer mortality, *Journal of the National Cancer Institute* **69**, 349–355.

[54] Shwartz, M. (1978). A mathematical model used to analyse breast cancer screening strategies, *Operations Research* **26**, 937–955.

[55] Shwartz, M. (1978). An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease, *Cancer* **41**, 1550–1564.

[56] Shwartz, M. & Plough, A. (1984). Models to aid in planning cancer screening programs, in *Statistical Methods for Cancer Studies*, R.G. Cornell, ed. Marcel Dekker, New York, pp. 239–416.

[57] Stevenson, C.E., Glasziou, P., Carter, R., Fett, M.J. & van Oortmarssen, G.J. (1990). Using Computer Modelling to Estimate Person Years of Life Saved by Mammography Screening in Australia, Paper presented at the 1990 Annual Conference of the Public Health Association of Australia.

[58] Tabar, L., Fagerberg, G., Duffy, S. & Day, N.E. (1989). The Swedish two county trial of mammography screening for breast cancer: recent results and calculation of benefit, *Journal of Community Health* **43**, 107–114.

[59] van Oortmarssen, G.J. & Habbema, J.D. (1991). Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer, *British Journal of Cancer* **64**, 559–565.

[60] van Oortmarssen, G.J., Habbema, J.D., Lubbe, K.T. & van der Maas, P.J. (1990). A model-based analysis of the HIP project for breast cancer screening, *International Journal of Cancer* **46**, 207–213.

[61] van Oortmarssen, G.J., Habbema, J.D., van der Maas, P.J., de Koning, H.J., Collette, H.J., Verbeek, A.L., Geerts, A.T. & Lubbe, K.T. (1990). A model for breast cancer screening, *Cancer* **66**, 1601–1612.

[62] Verbeek, A.L.M., Straatman, H. & Hendriks, J.H.C.L. (1988). Sensitivity of mammography in Nijmegen women under age 50: some trials with the Eddy model, in *Screening for Breast Cancer*, N.E. Day & A.B. Miller, eds. Hans Huber, Toronto, pp. 29–38.

[63] Walter, S.D. & Day, N.E. (1983). Estimation of the duration of a preclinical state using screening data, *American Journal of Epidemiology* **118**, 865–886.

[64] Walter, S.D. & Stitt, L.W. (1987). Evaluating the survival of cancer cases detected by screening, *Statistics in Medicine* **6**, 885–900.

[65] Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic diseases, *Biometrika* **56**, 601–614.

(*See also* **Incubation Period of Infectious Diseases**; **Prevalence of Disease, Estimation from Screening Data**)

CHRIS STEVENSON

Screening, Overview

The term "screening" is used to denote a variety of procedures in medicine and epidemiology. "Screening for disease" is mostly used to denote "the examination of asymptomatic people in order to classify them as likely, or unlikely, to have the disease that is the object of screening. People who appear likely to have disease are investigated further to arrive at a final diagnosis. Those people who are found to have the disease are then treated" (Morrison [18]).

This definition describes three distinguishing characteristics of screening for disease. First, screening is targeted at apparently healthy persons who are not aware of symptoms for which medical help would be sought. The **prevalence** of the disease in these persons will in general be (very) low. Secondly, the screening examination will give a crude distinction between persons with a normal test result, who do not receive further special attention, and persons in whom abnormalities are found in the screening test. Different grades of abnormalities may lead to more or less intensive follow-up, ranging from a repeat screening test to immediate treatment. Thirdly, appropriate treatment of disease which is detected early is expected to have a favorable impact on prognosis. The public health goal of screening for disease, "To reduce mortality or morbidity or to improve the quality of life" [12], is achieved by the more favorable outcome of early treatment in the cases identified by the screening test in comparison with similar cases that have been diagnosed on the basis of symptoms.

However, Table 1 shows that this benefit of screening is accrued by a very small proportion (group D+) of the persons screened. Inevitably screening will have a negative impact for other persons (groups C and D−). The small **risks** of some screening tests – for example, the increased risk of miscarriage following amniocentesis as part of antenatal screening, or the **radiation** risk of mammographic screening – cannot be disregarded completely given the very large number of tests performed (groups ABCD). Participation in screening for a serious disease will lead to anxiety, followed by relief when the result appears to be negative (groups A and B). Although this may sometimes be a relatively small effect, it cannot be neglected given the large number of persons involved.

False positive test results (group C) might lead to a (sometimes serious) burden of follow-up diagnostic tests needed to exclude disease. A true positive test result may still turn out to give adverse effects when early treatment does not improve the prognosis (group D−), but the person has been made aware of the disease for a longer period of time. Lack of improvement may occur when the outcome of early treatment remains unfavorable, but also when a person would have had a very good **prognosis** without screening.

Several extensions and modifications to the rather strict definition of Morrison are being used; see, for example, Holland & Stewart [13] or Wald [22]. For example, the term screening is also used to describe identification of people at high *risk* of disease (for example, high cholesterol or blood pressure levels) instead of early detection of the disease itself. The public health goal of screening for a disease may also be achieved indirectly, e.g. by preventing morbidity and mortality in other persons than those being screened. For example, specific groups of individuals, such as employees in the food industry, persons applying for a driving licence, or a circumscribed population in which an outbreak of an infectious disease occurred, may be screened to protect the general population.

Performing screening tests in an asymptomatic population can also have a scientific aim, such as to estimate the population prevalence of certain conditions, for example HIV infection (*see* **Prevalence of Disease, Estimation from Screening Data**). This extended usage of the term screening is reflected in the more general definition of McKeown [17]: "A medical investigation which does not arise from a patient's request for advice for a specific complaint".

In medicine and epidemiology, usage of the term screening is not necessarily related to testing of (asymptomatic) individuals, but is often used as a synonym for "testing". In clinical medicine, screening is used to denote testing of symptomatic patients to establish a diagnosis. It is also used in laboratory testing of donor blood for HIV infections, for example, and in medical research (laboratory, epidemiologic surveillance), the term screening is used to denote testing of chemical agents to identify toxic substances.

Discussion of screening for disease in this Section is confined to screening adhering to the strict definition of Morrison, while admitting that

Table 1 The different outcomes of screening, their usual impact and frequency. Extended version of table given by Morrison [18]

	Outcome	Impact	Proportion of target population
O	Nonparticipation	?	medium–large
A	True negative	anxiety, relief, side-effects of test	large
B	False negative	anxiety, false relief, side-effects of test	very small
C	False positive	moderate adverse effects	(very) small
D−	True positive, serious condition not postponed	adverse	(very) small
D+	True positive, serious condition postponed	large benefit	very small

it is difficult to draw exact boundaries for this definition. Examples will mostly be derived from cancer screening studies. (See the series *Screening Brief* in the *Journal of Medical Screening* for up-to-date information on screening programs for specific diseases.)

A broad distinction can be made between, on the one hand, genetic screening (*see* **Genetic Counseling**; **Genetic Markers**), and most antenatal and neonatal screening procedures which involve a single screening examination and, on the other hand, screening for chronic diseases, including screening for problems during growth of children and screening for cancer in adults, which typically involve repeated screening tests with intervals between several months to years.

Genetic screening can be done at different times throughout life. It can be performed prior to conception to inform persons of a high risk of conceiving a child with a severe disorder, during pregnancy for early detection and elective termination of pregnancy, shortly after birth to detect treatable disorders, or later in life to enable preventive measures which reduce the risk of developing serious disorders. In antenatal and neonatal screening, optimal timing of the test is important because the gestational age or age of the child determine the **sensitivity** and **specificity** of the test(s) and the possibilities for intervention.

In the case of repeated screening examinations for early detection of diseases such as (breast, cervical, or colorectal) cancer, proper timing of tests is even more complicated because of various time-related factors involved: incidence and prevalence of the detectable preclinical phase (DPCP) varies with age, the **sojourn time** of the DPCP varies between persons, and test characteristics and the outcome of early treatment vary during the course of the DPCP. The number of factors involved, and the dynamic interrelations between factors, complicate the design, analysis and evaluation of such screening programs (*see* **Screening Benefit, Evaluation of**; **Screening, Models of**).

Important questions in analysis and evaluation of screening for disease are:

1. Will screening indeed reduce the mortality and/or morbidity in the population and, if so, what is the estimated magnitude of the reduction?
2. What are the favorable and adverse effects and costs of different screening policies, and what will be the impact on existing health care? A policy is characterized by the recommended age(-range) to be screened, the screening test(s) used, the intervals between examinations in the case of repeated screening, and the diagnostic follow-up and subsequent treatment to be applied.
3. What are efficient policies, and is screening worthwhile?
4. Does the screening program, when implemented as part of routine care, perform adequately?

These issues are discussed in turn in the following Sections.

Establishing the Effectiveness of Screening

The effect of a certain screening policy on mortality and morbidity depends on several factors, such as the screening test, the natural history of the disease, the

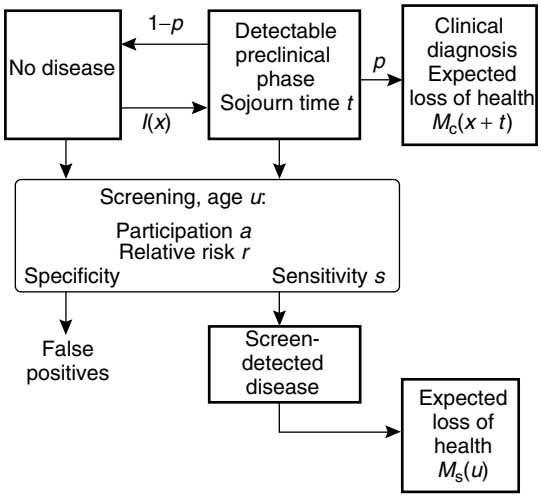

Figure 1 A single screening examination for a disease with a detectable preclinical phase starting at age x. Without screening, the disease will remain unnoticed until progressive cases are diagnosed clinically at age $z = x + t$. Screening at age u will lead to false positive test results, but also to early detection and treatment for part of the prevalent cases at age u, i.e. in participants for which $x < u < x + t$ and the test result is correct. These true positive cases will include persons with nonprogressive disease that would never have been diagnosed in the absence of screening

x = age at onset of the detectable preclinical phase
$I(x)$ = incidence density of the detectable preclinical phase
t = sojourn time in the detectable preclinical phase, probability density $f(t)$
p = proportion with progressive disease among incidence $I(x)$
a = proportion participating in screening at age u
r = relative risk of participants
s = sensitivity of the screening test
$M_s(u)$ = expected loss of health following detection by screening at age u
$M_c(z)$ = expected loss of health following clinical diagnosis at age $z = x + t$

diagnostic and treatment options for the disease, the improvement in prognosis resulting from early treatment, and, at the level of the population, the degree of participation in the screening program, including possible selective participation of high risk groups. Figure 1 shows these factors for a single screening examination, one screening test, and a disease with a fixed duration for the detectable preclinical phase

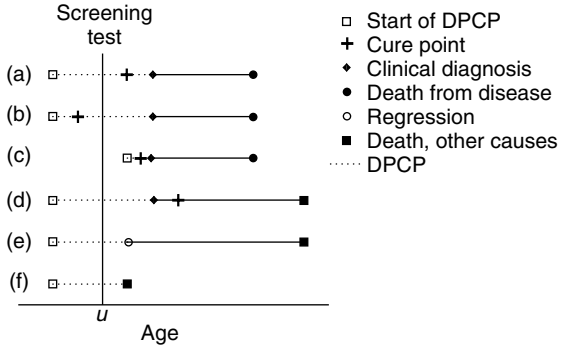

Figure 2 Example disease histories that are missed, diagnosed without benefit, or diagnosed with favorable effect, by a single screening examination at age u. The cure point indicates a hypothetical moment in the disease history where treatment ceases to be effective. The thick line indicates the detectable preclinical phase (DPCP)

(*see* **Decision Analysis in Diagnosis and Treatment Choice**; **Natural History Study of Prognosis**; **Risk Assessment in Clinical Decision Making**).

Together, these factors determine not only the positive health effect of screening, but also its negative effects and its costs. Figure 2 shows examples of disease histories, starting from the onset of the detectable preclinical phase (DPCP), for which screening has favorable or adverse effects. In these examples it is assumed that the main goal of screening is to prevent death from the disease, such as, for example, in cancer screening. A cure point is indicated, denoting a hypothetical point in the history where treatment ceases to be effective [9].

In history (a), the disease is detected by a true positive screening test result, and death from the disease is to be prevented by screening since detection occurs before the cure point. Death from the disease will not be prevented in the case of a **false negative** test result or inadequate follow-up in this history. Screening will also be ineffective and have merely adverse effects when it is too late (history (b)), or when the DPCP has a short sojourn time and is missed by the screen (history (c)). Screening has no impact on mortality and clear negative effects in history (d), where clinical diagnosis would also have occurred before the cure point, in history (e) of regressive disease where the DPCP would never have been detected in the absence of screening, and in history (f) where the person dies from other causes

before the disease would have been diagnosed clinically. The increasing likelihood of the last type of history at older ages should be kept in mind in evaluating screening for diseases such as cancer of the prostate that mainly occur in the elderly.

The lead time is the length of the time interval between the moments of detection by screening and clinical diagnosis without screening. The average lead time is indicative of the potential positive effect of screening, since a longer average lead time means that more cases can be detected before the cure point is reached. But in histories (b) and (d), the lead time only represents a negative effect: the person is merely aware of having serious disease for a longer period of time. Histories (a) and (c) illustrate the **length-biased** sampling phenomenon, which means that screening tends to pick up histories with long sojourn times selectively. Zelen & Feinleib [26] have pointed out the consequences for the **mean** lead time ($\overline{L}$) of screen-detected cases at a single screening examination:

$$\overline{L} = \frac{1}{2}\left(\overline{t} + \frac{\sigma_t^2}{\overline{t}}\right), \qquad (1)$$

indicating that the mean lead time is longer than the mean sojourn time $\overline{t}$ when the **variance** σ_t^2 of the sojourn time distribution exceeds its mean value.

For the simplified situation of Figure 1 with a single screening test at age u, the true prevalence $D(u)$ of preclinical disease is given by

$$D(u) = \int_{x=0}^{u}\int_{t=u-x}^{\infty} I(x)\,f(t)\,\mathrm{d}t\,\mathrm{d}x, \qquad (2)$$

where $I(x)$ denotes the incidence density of the detectable preclinical phase and $f(t)$ the probability density function of the sojourn time t in the detectable preclinical phase.

The expected health effects $G(u)$ for the population in which this single screening examination is taking place accrue to the screen-detected cases. This is a subset of the prevalent cases, including persons who participate in screening and have a certain associated **relative risk**, and for whom the test has a true positive result. The health effect for these cases is obtained by subtracting the loss of health M_c in a situation without screening for the fraction p of progressive cases, which would have been diagnosed clinically at age $x + t$, from the loss of health M_s in all screen-detected cases at age u in the situation with screening:

$$G(u) = \int_{x=0}^{u}\int_{t=u-x}^{\infty} I(x)\,f(t)ars \times [M_s(u) - pM_c(x+t)]\,\mathrm{d}t\,\mathrm{d}x, \qquad (3)$$

where $I(x)$ denotes the **incidence density** of the detectable preclinical phase, $f(t)$ the probability density of the sojourn time t in the detectable preclinical phase, a the proportion participating in screening at age u, r the relative risk of participants, and s the sensitivity of the screening test.

In (3), the lead time for screen-detected progressive cases is $x + t - u$, the time interval between detection at screening and clinical diagnosis in the absence of screening. In the example of a potentially lethal disease as presented in Figure 2, the measures M_c and M_s for the loss of health might be taken to represent the lethality from the disease following diagnosis and treatment. Comparison of M_c and M_s on the individual level is, of course, impossible, because the exact clinically diagnosed counterparts of screen-detected cases will always remain unknown.

In general it will be difficult to obtain direct estimates of the components of (3), except for $M_c(\cdot)$ and $M_s(\cdot)$, which may be based on follow-up registries (*see* **Disease Registers**) of clinically diagnosed and screen-detected patients. Although it is tempting to compare $M_c(\cdot)$ and $M_s(\cdot)$ directly, this will give rise to incorrect conclusions because of four sources of sampling **bias** that all tend to lead to a too favorable estimate for the effect of screening.

If not all preclinical stages progress ($p < 1.0$), then the comparison will yield a too optimistic estimate because of overdiagnosis bias (see history (e) in Figure 2.) Self-selection bias (*see* **Selection Bias**) with respect to survival occurs when participants in screening would have had a better prognosis anyhow, for example because of self-selection of health-conscious persons. Self-selection may also be related to the risk of developing the disease. Participants have been observed to have a higher than average risk in breast cancer screening [23], but the opposite has been observed in cervical cancer screening [1]. Lead time bias occurs when cumulative lethality is compared for equal durations of follow-up after diagnosis, giving screen-detected cases an advantage equal to the duration of the lead time even in the absence

of a real effect of screening. Length-biased sampling will also lead to biased comparisons if the sojourn time and M_c are **correlated**, for example when slowly developing preclinical disease also has a better-than-average prognosis.

Lead time bias and length biased sampling are specific for screening, and different approaches have been proposed to correct these biases [24]. However, these correction methods are always based on assumptions about the sojourn time distribution, and will not lead to unambiguous evidence about the effect of screening.

The four biases can only be avoided by conducting a randomized controlled trial (RCT) (*see* **Clinical Trials, Overview**; **Screening Trials**). In its basic form, a population involved in an RCT is randomly divided (*see* **Randomization**) in a study group in which persons are invited to be screened, and a **control** group in which no screening is offered. The endpoint to be compared between the two groups is the condition that is to be prevented by early detection and treatment, for example mortality in cancer screening (*see* **Outcome Measures in Clinical Trials**). Use of other endpoints – for example, diagnosis of malignancy in cervical cancer screening – might lead to biased estimates of the effect of screening when, on average, screen-detected cases are less severe than clinically diagnosed cases. A huge population will be needed to obtain sufficient **power**, and several design variants have been proposed to limit trial costs [9]. The impact of trials has been considerable, both in diminishing the use of screening for lung cancer, and in speeding up implementation of breast cancer screening program in several countries. Screening for lung cancer did not turn out to be effective according to RCT results, despite clear differences in survival between screen-detected and clinically diagnosed patients [10]. In most RCTs conducted thus far, screening for breast cancer has been found to reduce breast cancer mortality in women above age 50, but results for women below age 50 are still not conclusive [5, 21].

Use of disease-specific outcome measures in an RCT is controversial. Critics state that the beneficial effect should be checked from overall mortality and morbidity. But this would require an enormous trial size. Even the combined results of four Swedish breast cancer trials, with a relative risk for breast cancer death of 0.80 (95% CI 0.70–0.92) did not show a discernable effect on overall mortality in the trial population [20].

Some screening tests that are widely used have never been rigorously tested in an RCT. One example is cervical cancer screening, for which a very large RCT would be required. In such a situation, estimates of the effectiveness of screening can only be obtained from nonexperimental designs, such as **cohort studies, case–control studies**, and **ecologic studies**.

In cohort studies, a comparison of morbidity and mortality is made between persons with different screening experience in the cohort. Case–control studies for testing the effectiveness of screening have become increasingly popular, but the outcomes are highly sensitive to several kinds of bias such as overdiagnosis bias, self-selection bias, and healthy screenee bias (see Morrison [18] or Weiss [25]). This has been demonstrated empirically by performing a case–control study on data from an RCT [11].

In ecologic studies, an investigation is made into the association between the morbidity or mortality and the screening intensity in different populations. Cervical cancer screening is now generally considered to be effective on the basis of the findings of many **observational studies:** cohort studies and case–control studies (IARC Working Group [14]) and ecologic analyses (see, for example, Läärä et al. [16].

The Favorable and Adverse Effects and Costs of Alternative Policies

When screening is being considered in a country or region, usually different screening policies are being considered which might differ in their effects and costs. Only a limited number of policies have been rigorously tested in RCTs, which are typically carried out in countries or regions with marked differences with respect to, for example, incidence or mortality rate, participation in screening, and specific characteristics and quality of the screening procedures. Furthermore, the long follow-up in many screening trials implies that their results typically pertain to screening technology from the past. These observations complicate both combined analysis of trial results, and **extrapolation** of these results to other situations such as the near future in a new

country in which screening is being considered. Also, RCTs tend to focus on the serious health effects to be prevented by screening (category D+ in Table 1), paying less attention to the adverse effects of screening.

Two methods are used to resolve (partially) these problems – modeling and use of **surrogate endpoint** measures.

In building a model, assumptions have to be made about the factors listed in Figure 1. These assumptions can be checked (*see* **Model Checking**) by fitting the model to available data from RCTs and to data from nonexperimental screening studies. For example, outcomes of cancer screening models are compared with observed detection rates at successive screening rounds, incidence of clinical disease in the interval between screening exams, and with the stage-distribution of these different types of cases. Characteristics of, and trends in, background variables can be taken into account explicitly in a model. In this way, known differences between trials can be incorporated in combined analyses of RCT results (see [5]). Models are increasingly being used to make predictions about other screening policies than those tested in trials, and to transfer outcomes from trials to specific characteristics in other areas (see, for example, [6]).

Introducing screening for a disease in a population will change the type and amount of diagnostic and treatment procedures for this disease. In a first screening round, a relatively large pool of prevalent cases will have a positive test result, leading to a temporary increase in demand for assessment and treatment, followed by a decrease in later screening rounds. Models can make quite detailed predictions of these changes, which is useful in planning of equipment, recruitment, and training of personnel, and in anticipation of changes in clinical procedures [4].

The number of assumptions used in screening models may easily become quite large, and some of these assumptions lack a thorough foundation. Use of surrogate outcome measures can be regarded as an attempt to obtain a more direct empirical basis in comparing different policies, without the lengthy follow-up period needed in large RCTs [2]. A surrogate outcome measure is a short-term observed result of a screening program that is known (or suspected) to be closely associated with the long-term impact on morbidity and mortality. Such a relation can be estimated from RCT data. For example, mortality reduction after breast cancer screening has been shown to be closely associated with particular changes in the stage distribution of diagnosed cancer cases.

Surrogate outcome measures can be used to evaluate various small adaptations of policies that have already been proven to be effective. The surrogate outcome measures of the policy variants can be compared directly, and the observed differences can also be translated into predicted differences in reduction of mortality and morbidity.

Efficient Policies

In deciding about screening policies an important criterion is the ratio of the health effects of a screening program and its incremental costs, i.e. the difference in (medical) costs between the situation with and without screening. The incremental costs of a screening policy are directly related to the cost of administering and assessing the screening test, but may also be influenced by the impact of screening on the demand for diagnostic and treatment procedures. For each level of the incremental costs, an optimal screening policy exists which gives the highest effectiveness (net health benefit) or the best cost-effectiveness ratio (*see* **Health Economics**). Figure 3 shows a typical intensity–response relationship which emerges when

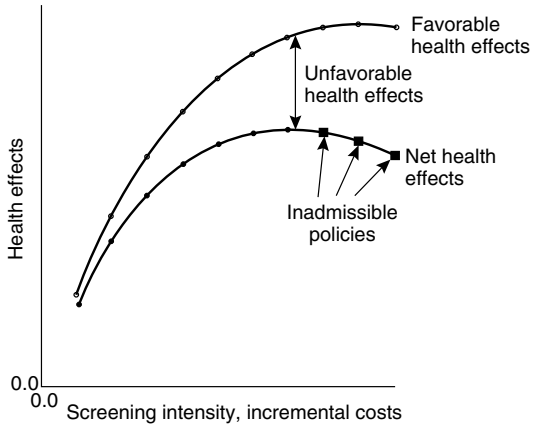

Figure 3 Relation between intensity of screening and its favorable and unfavorable health effects (intensity–response curve) or similarly between the incremental costs of screening and health effects (efficient frontier). The net health benefit summarizes the favorable and adverse effects of screening

different numbers of screenings are compared. The net health benefit is relatively high for a single screening test. The extra benefit decreases for each additional screening examination. When the number of screening tests per person increases, the adverse effects of screening might well become larger than the favorable effects. For example, extending screening to older ages will increasingly lead to detection of cases that would not have been diagnosed in the absence of screening, leading to reduction in **quality of life** that is no longer compensated for by a decrease in mortality.

Although the top of the curve in Figure 3 represents the policy with the maximum effectiveness, the marginal cost-effectiveness of this policy is very poor. If the X-axis of Figure 3 represents costs, the curve can be interpreted as the efficient frontier, i.e. the set of all **Pareto**-optimal screening policies – policies for which no alternative can be found that give both higher health benefits and lower costs [8]. A good policy on this frontier would be the one which still shows an acceptable marginal cost-effectiveness ratio.

Simulation models have been applied to derive the efficient frontier of screening policies [7, 8, 15].

Monitoring of Screening Program

If screening is not conducted properly, effectiveness and cost-effectiveness can easily be impaired. A well-known example is the cervical cancer screening program in the UK, which has not been able to prevent increasing mortality rates in young women [19]. Each of the stages in the screening process could give rise to loss of potential benefits, and sometimes also to excessive negative effects. Sufficient coverage of the population at risk, adequate administration of the test and interpretation of test results, compliance with follow-up in the case of suspicious results, and proper treatment of early stages of the disease are all necessary to achieve the health benefits at a reasonable cost. Quality assurance and monitoring are important in this respect.

Measures of performance can be defined for each component of a screening program – e.g. the participation rate and **predictive value** of a positive test result – and should encompass diagnostic and treatment procedures applied to screen-detected disease. Target values for the short-term results of a screening program (such as detection rates and incidence of interval cases) can be specified on the basis of experience from randomized trials and pilot projects [3]. Regular evaluation of the screening results and of the costs of the screening program may than lead to timely revision of the screening policy.

References

[1] Boyes, D.A., Morrison, B., Knox, E.G., Draper, G.J. & Miller, A.B. (1982). A cohort study of cervical cancer screening in British Columbia, *Clinical Investigations in Medicine* **5**, 1–29.

[2] Day, N.E. (1991). Surrogate measures in the design of breast cancer screening trials, in *Cancer Screening*, A.B. Miller, J. Chamberlain, N.E. Day, M. Hakama & P.C. Prorok, eds. Cambridge University Press, Cambridge, pp. 392–403.

[3] Day, N.E., Williams, D.R.R. & Khaw, K.T. (1989). Breast cancer screening programmes: the development of a monitoring and evaluation system, *British Journal of Cancer* **59**, 954–958.

[4] De Koning, H.J., Van Oortmarssen, G.J., Van Ineveld, B.M. & Van der Maas, P.J. (1990). Breast cancer screening: its impact on clinical medicine, *British Journal of Cancer* **61**, 292–297.

[5] De Koning, H.J., Boer, R., Warmerdam, P.G., Beemsterboer, P.M.M. & Van der Maas, P.J. (1995). Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials, *Journal of the National Cancer Institute* **87**, 1217–1223.

[6] Eddy, D.M. (1980). *Screening for Cancer: Theory, Analysis and Design*. Prentice-Hall, Englewood Cliffs.

[7] Eddy, D.M. (1990). Screening for cervical cancer, *Annals of Internal Medicine* **113**, 214–226.

[8] Eddy, D.M. (1990). Screening for colorectal cancer, *Annals of Internal Medicine* **113**, 373–384.

[9] Etzioni, R.D., Connor, R.J., Prorok, P.C. & Self, S.G. (1995). Design and analysis of cancer screening trials, *Statistical Methods in Medical Research* **4**, 3–17.

[10] Flehinger, B.J., Kimmel, M., Polyak, T. & Melamed, M.R. (1993). Screening for lung cancer, *Cancer* **72**, 1573–1580.

[11] Gullberg, B., Andersson, I., Janzon, L. & Ranstam, J. (1991). Screening mammography, *Lancet* **337**, 244.

[12] Hakama, M. (1991). Screening, in *Oxford Textbook of Public Health*, Vol. 3. Oxford University Press, Oxford.

[13] Holland, W.W. & Stewart, S. (1990). *Screening in Health Care – Benefit or Bane?* The Nuffield Provincial Hospitals Trust, London.

[14] IARC Working Group on Evaluation of Cervical Cancer Screening Programmes (1986). Screening for squamous cervical cancer: duration of low risk after negative results of cervical cytology and its implication for screening policies, *British Medical Journal* **293**, 659–664.

[15] Koopmanschap, M.A., Lubbe, J.Th.N., Van Oortmarssen, G.J., Van Agt, H.M.E., Van Ballegooijen, M. & Habbema, J.D.F. (1990). Economic aspects of cervical cancer screening. *Social Science and Medicine* **30**, 1081–1087.

[16] Läärä, E., Day, N.E. & Hakama, M. (1987). Trends in mortality from cervical cancer in the Nordic countries: association with organized screening programmes, *Lancet* **i**, 1247–1249.

[17] McKeown, T. (1968). Validation of screening procedures, in *Screening in Medical Care: Reviewing the Evidence*. The Nuffield Provincial Hospitals Trust, Oxford.

[18] Morrison, A.S. (1992). *Screening in Chronic Disease*, 2nd Ed. Oxford University Press, New York.

[19] Murphy, M.F.G., Campbell, M.J. & Goldblatt, P.O. (1988). Twenty years' screening for cervical cancer of the uterine cervix in Great Britain, 1964–84: further evidence for its ineffectiveness, *Journal of Epidemiology and Community Health* **42**, 49–53.

[20] Nyström, L., Larsson, L.-G., Wall, S., Rutqvist, L.E., Andersson, I., Bjurstam, N., Fagerberg, G., Frisell, J. & Tabár, L. (1996). An overview of the Swedish randomized mammography trials: total mortality pattern and the representativity of the study cohorts, *Journal of Medical Screening* **3**, 85–87.

[21] Nyström, L., Rutqvist, L.E., Wall, S., Lindgren, A., Lindqvist, M., Rydén, S., Andersson, I., Bjurstam, N., Fagerberg, G., Frisell, J., Tabár, L., & Larsson, L.-G. (1993). Breast cancer screening with mammography: overview of Swedish randomized trials, *Lancet* **341**, 973–978.

[22] Wald, N.J. (1994). Guidance on Terminology, *Journal of Medical Screening* **1**, 76.

[23] Walter, S.D. & Day, N.E. (1983). Estimation of the duration of the pre-clinical state using screening data, *American Journal of Epidemiology* **118**, 865–886.

[24] Walter, S.D. & Stitt, L.W. (1987). Evaluating the survival of cancer cases detected by screening, *Statistics in Medicine* **6**, 885–900.

[25] Weiss, N.S. (1994). Application of the case-control method in the evaluation of screening, *Epidemiologic Reviews* **16**, 102–108.

[26] Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic diseases, *Biometrika* **56**, 601–613.

(*See also* **Diagnostic Tests, Evaluation of**; **Natural History Study of Prognosis**).

GERRIT J. VAN OORTMARSSEN

Screening, Sojourn Time

Screening for disease can take place in a variety of contexts, notably the control of transmission of infectious disease, as part of an immunization program or to advance the time of diagnosis to facilitate effective treatment. Central to the potential suitability of a disease for screening is the period during which the disease does not produce symptoms leading to medical consultation and diagnosis, but is detectable by a screening test. This is the time between the point at which a disease case becomes detectable by screening and the point at which it would become clinically apparent in the absence of screening. This period is called the *preclinical phase* and its duration is referred to as the *sojourn time* [4, 18]. One example is the period during which a breast tumor is not palpable and has no symptoms but is detectable by mammography. Another is the period of HIV antibody positivity after seroconversion but before onset of **AIDS**-related illnesses.

The sojourn time is an important parameter of the potential effectiveness of a screening program. It is an upper limit on the lead time, the advance in the time of diagnosis achieved by screen detection. In practice, the parameter estimated is the average sojourn time of all disease cases (average time from becoming screen detectable to becoming clinically apparent in the absence of screening), usually referred to as the mean sojourn time and often abbreviated to MST. A long mean sojourn time will indicate a good potential for screening. The shorter the mean sojourn time, the more frequently screening will have to take place in order to be effective [2]. If mean sojourn time is very short, then it may not be worth screening at all.

In terms of the definition above, the sojourn time seems a relatively simple concept. There are, however, various complexities, due either to the varying effectiveness of the screening tool or the nature of the disease screened for [8, 18]. To take the first problem, suppose that in one breast cancer screening program in women aged 50–69 a mean sojourn time of 2.5 years is observed and that in a neighboring region the breast screening program has a mean sojourn time of 3.3 years. There are two possible explanations for this: (i) that in the second region tumors take longer to come to clinical attention, and (ii) that in the second region the sojourn time begins earlier in tumor development, due to better **sensitivity** (relating to better perceptive skills on the part of screening staff or better image quality of the X-ray films). Thus the sojourn time is closely related to the sensitivity of the screening instrument.

Another complication relates to nomenclature and the disease screened for. In breast cancer screening this problem does not arise. The aim of a breast cancer screening program is to diagnose breast cancers while they are small and before they have spread to the regional lymph nodes, in order to facilitate successful curative therapy and thus prevent death from the disease. In what is nominally a cervical cancer screening program, the main aim is to detect premalignant dysplasia and to take action which will avoid even a diagnosis of invasive cervical cancer in the future. The sojourn time in this case is not that of cervical cancer alone but of both cancer and dysplasia which may or may not progress to cancer if left untreated. The meaning of sojourn time becomes even more nebulous if we consider the use of hepatitis B virus testing in a liver cancer screening program or Epstein–Barr virus testing in screening for nasopharyngeal cancer (*see* **Incubation Period of Infectious Diseases**).

Development and Estimation

Although the above complications may arise, the concept is still a crucial one in screening for many diseases, notably cancers, and considerable research effort has been expended on methodology for its estimation, in conjunction with estimation of screening sensitivity. The seminal work on the theory and modeling is by Zelen & Feinleib [19], Prorok [12], and Walter & Day [18]. Other early research on the subject includes that of Eddy [8], Shapiro et al. [14], and Shwartz [15]. Underlying the work is the basic acknowledgment that the sojourn time varies from individual to individual. In turn, it is clear that the mean sojourn time is fundamentally important to the screening process, and that a postulated distribution of sojourn time which fits with observed disease incidence and screening data is desirable. An **exponential distribution** has many attractive qualities for temporal data, and Day & Walter [4] found it to be a good fit to breast cancer screening data.

Given a mathematical model for the sojourn time, for estimation of the mean sojourn time, it is also

necessary to model the incidence of preclinical disease. Also, because of the relationship with sensitivity mentioned above, it is preferable to take sensitivity into account when estimating mean sojourn time. Data on both screen-detected cancers (with diagnosis taking place by definition before the sojourn time expires) and cancers diagnosed in the interval between screens (with diagnosis at the expiry of the sojourn time) are necessary to estimate both parameters at once. Clearly a high **prevalence** at screening in comparison with expected annual incidence in the absence of screening suggests a long sojourn time, as does a low incidence of disease after a negative screen, although both of these are also affected by the sensitivity of the screening instrument. Early work tended to approximate the incidence of preclinical disease by the **incidence rate** of disease in the absence of screening, estimated from randomized trials (*see* **Clinical Trials, Overview**) or from historical control data [4, 19]. **Estimation** procedures ranged from simple formulae [14, 6] to results from complex models, involving computer-intensive **optimization** [4] or algebraically complicated **moment** calculations [19].

Current Approaches

Advances in computing power in recent years have enabled a wider variety of modeling techniques to be used in the estimation of mean sojourn time. Alexander [1] has adapted the model of Day & Walter [4] to estimation of the sojourn time when two screening techniques are used in the same program. Paci & Duffy [10] have used **generalized linear models** to estimate the mean sojourn time. This has the advantage of being programmable in a few lines of code using generally available statistical software (*see* **Software, Biostatistical**), but has the disadvantage of having to assume a sensitivity of 100% while estimating the MST. Van Oortmarssen et al. [17] have developed powerful computer **simulation** techniques. This approach is very versatile, but careful attention has to be paid to assumptions made in the simulations. More recently, explicit **Markov chain** modeling of both entry to and exit from the preclinical phase has been performed, with **generalized estimating equations** used as the estimation technique [7, 2]. This approach requires laborious computer programming, but has considerable potential for estimation of progression rates with respect to the disease stage as well as from the preclinical to the clinical phase.

In the context of HIV disease, there has been considerable recent activity in estimation of the sojourn time (referred to as the incubation period in this context), as an aid to prediction of the course of the epidemic. Examples abound in the literature, but perhaps the most important models and methods are typified by **back calculation** and use of **Weibull** models [5] and Markov chain models [13].

An interesting example of application is in the evaluation of breast cancer screening by age. It has long been known that screening for breast cancer is less effective in women aged under 50 years than in women aged 50 or more [16]. Breast tissue is denser in premenopausal women, and it is thought that this leads to reduced sensitivity of mammographic screening. Table 1 below shows simultaneous estimates of sensitivity and MST from Markov chain models applied to data from the Swedish Two-County Trial of breast cancer screening [2]. The results indicate that more rapid progression, i.e. shorter sojourn time, is also an important factor in the differential effectiveness of screening by age.

Likely Future Developments

One notable gap in the available methodology is the estimation of mean sojourn time and other screening parameters in the presence of informative attendance (i.e. subjects at greater or lesser risk of disease are more likely to attend for screening). This area seems ripe for future research. Another topic in the modeling field which is likely to be increasingly addressed in the future is the inclusion of attributes of preclinical disease (e.g. tumor size in cancer screening), possibly multidimensional, in models for the sojourn time. Already the simple model of the preclinical phase has been expanded to include some measure of stage of development of disease [17, 16, 3]. It is likely

Table 1 Mean sojourn time and sensitivity by age; Swedish Two-County Trial of breast cancer screening

Parameter	40–49	50–59	60–69
MST (years)	2.35	3.75	4.23
Sensitivity (%)	90	100	100

that more formal and more complex models will be developed in the future.

The increasing complexity of disease progression models may require new methods of estimation. A likely solution is the series of **Markov chain Monte Carlo** techniques, which are becoming increasingly common in biostatistical modeling [9]. One application, in a single model of colorectal cancer screening, has already been published and doubtlessly others will follow [11].

References

[1] Alexander, F.E. (1989). Estimation of sojourn time distributions and false negative rates in screening programmes which use two modalities, *Statistics in Medicine* **8**, 743–755.

[2] Chen, H.H., Duffy, S.W. & Tabar, L. (1996). A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening, *Statistician* **45**, 307–317.

[3] Connor, R.J., Chu, K.C. & Smart, C.R. (1989). Stage-shift cancer screening model, *Journal of Clinical Epidemiology* **42**, 1083–1095.

[4] Day, N.E. & Walter, S.D. (1984). Simplified models of screening for chronic disease: estimation procedures from mass screening programmes, *Biometrics* **40**, 1–13.

[5] Day, N.E., Gore S.M. & de Angelis, D. (1995). Acquired immune deficiency syndrome predictions for England and Wales (1992-97): sensitivity analysis, information, decision, *Journal of the Royal Statistical Society, Series A* **158**, 505–524.

[6] Day, N.E., Walter, S.D. & Collette, B. (1984). Statistical models of disease natural history: their use in the evaluation of screening programmes, in *Screening for Cancer I – General Principles on Evaluation of Screening for Cancer and Screening for Lung, Bladder and Oral Cancer*, P.C. Prorok & A.B. Miller, eds. International Union Against Cancer, Geneva, pp. 55–70.

[7] Duffy, S.W., Chen, H.H., Tabar, L. & Day, N.E. (1995). Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase, *Statistics in Medicine* **14**, 1531–1543.

[8] Eddy, D. (1980). *Screening for Cancer: Theory, Analysis and Design*. Prentice-Hall, Englewood Cliffs.

[9] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[10] Paci, E. & Duffy S.W. (1991). Modelling the analysis of breast cancer screening programmes: sensitivity, lead time and predictive value in the Florence District Programme (1975–1986), *International Journal of Epidemiology* **20**, 852–858.

[11] Prevost, T.C., Lannoy, G., Duffy, S.W. & Chen H.H. (1998). Estimating sensitivity and sojourn time in screening for colorectal cancer: a comparison of statistical approaches, *American Journal of Epidemiology* **148**, 609–619.

[12] Prorok, P.C. (1976). The theory of periodic screening II: Doubly bounded recurrence times and mean lead time and detection probability estimation, *Advances in Applied Probability* **8**, 460–476.

[13] Satten, G.A. & Longini, I.M. (1996). Markov chains with measurement error: estimating the "true" course of the human immunodeficiency virus disease, *Applied Statistics* **45**, 275–295.

[14] Shapiro, S., Goldberg, J.D. & Hutchison, G.B. (1974). Lead time in breast cancer detection and implications for periodicity of screening, *American Journal of Epidemiology* **100**, 357–366.

[15] Shwartz, M. (1978). An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease, *Cancer* **41**, 1550–1564.

[16] Tabar, L., Fagerberg, G., Chen, H.H., Duffy, S.W., Smart, C.R., Gad, A. & Smith, R.A. (1995). Efficacy of breast cancer screening by age: new results from the Swedish two-county trial, *Cancer* **75**, 2507–2517.

[17] Van Oortmarssen, G.J., Habbema, J.D.F., van der Maas, P.J., de Koning, H.J., Collette, H.J.A., Verbeek, A.L.M., Geerts, A.T. & Lubbe, K.T.N. (1990). A model for breast cancer screening, *Cancer* **66**, 1601–1612.

[18] Walter, S.D. & Day, N.E. (1983). Estimation of the duration of a preclinical disease state using screening data, *American Journal of Epidemiology* **118**, 865–886.

[19] Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic diseases, *Biometrika* **56**, 601–613.

(*See also* **Length Bias**; **Screening Benefit, Evaluation of**; **Screening, Models of**)

STEPHEN W. DUFFY

Seasonal Time Series

Literally, seasonality is a cyclic variation (*see* **Circadian Variation**) driven by the seasons within the year and, traditionally, seasonal models have been of most importance for economic or business applications, so the methodologies so far favor that direction. In such applications, prediction and seasonal adjustment of mostly univariate series have been the primary objectives. With many biological measurements, daily cycles are probably the most natural, but the methodology of seasonal adjustment would still play a role, so it will be convenient to view periodic time series as being seasonal. However, unlike in economic applications, it will be quite easy in biostatistical applications to get replicated time series from different individuals. In such applications, the analysis objectives of seasonal time series data are likely to be (a) a comparison of different groups of individuals, each having a time series record or (b) a comparison of before-and-after conditions. A general methodology to address these questions is not commonly available in the current literature. More relevant data sets need to be available to encourage methodological development. The rest of the article is a description of the current state of seasonal modeling.

Following the above discussion, to develop a general model of seasonality it is more useful and interesting to include any type of periodic patterns such as a day-and-night or a weekly cycle, or even a periodicity other than that shown by natural cycles. A famous example of the latter is the Canadian lynx time series (the annual number of lynx trapped in the Mackenzie River district of north-west Canada) with a periodicity of around 10 years [7]. Figure 1 shows

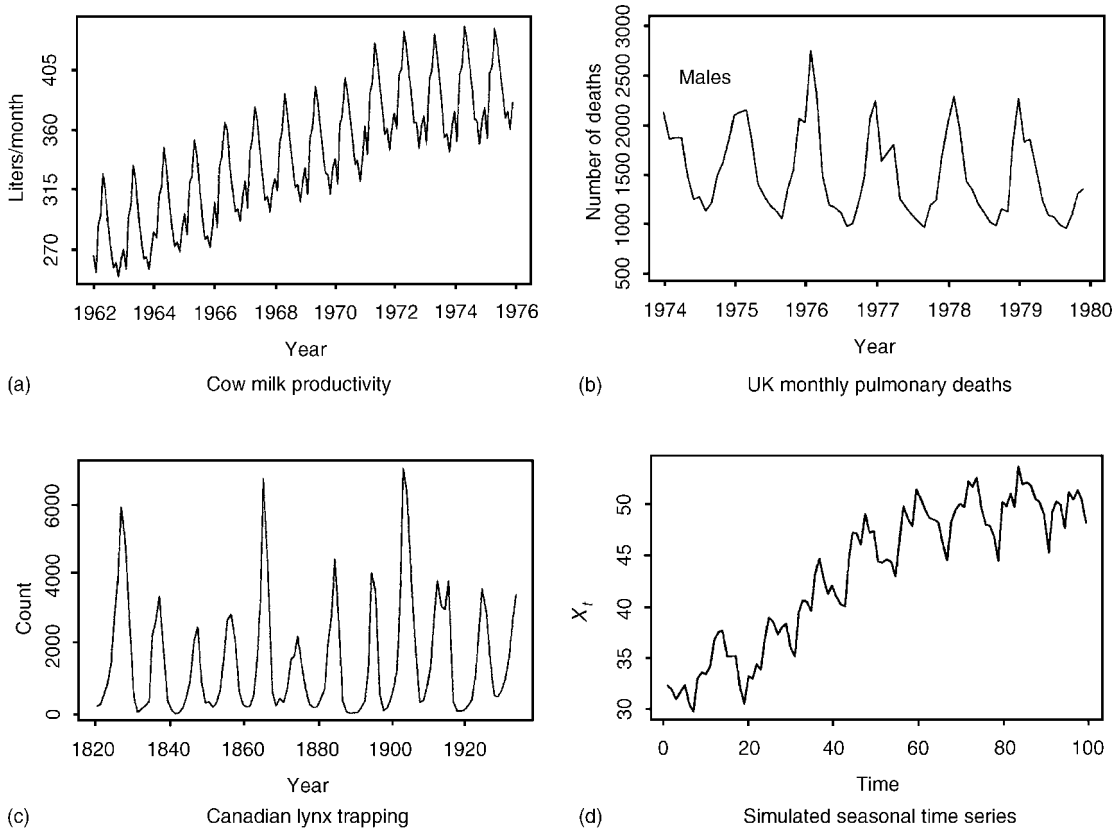

Figure 1 Examples of seasonal time series. In most cases, the seasonality is driven by some natural cycles, except in the Canadian lynx series. See Cryer [4] for the milk production data and Diggle [6] for the UK pulmonary data. The simulated series is ARIMA $(0, 1, 1) \times (0, 1, 1)_{12}$ with Gaussian noise

several examples of time series data that exhibit a common feature of seasonality or periodicity.

The interest in analyzing a seasonal time series may lie in the seasonality itself if it is an indication of some unknown underlying process; this is especially true when the seasonality does not coincide with some natural cycle. Another interest, usually associated with economic time series, is the *removal* of the seasonality. In this case, the cause of the seasonality is typically known and is not of analytical interest. This so-called *seasonal adjustment* allows us to gauge effects other than the known seasonality. For example, we know that the temperature is higher during the summer, so we must remove the general summer effect if we wish to know whether a particular summer has an unusually low or high temperature.

Models for Seasonality: Static

An intuitive and empirical model for a seasonal time series X_t is of the form

$$X_t = T_t + S_t + I_t, \tag{1}$$

where T_t is a long-term trend, S_t a seasonal factor, and I_t an irregular component. The model is static in the sense that it does not specify how the process evolves from one point to the next. In most applications, it is common to model I_t as random noise, and T_t and S_t as nonparametric functions of time. Low-dimensional parametric regression models may also be specified for T_t and S_t: for example, polynomial and trigonometric functions, but generally they are too rigid for most time series data. A simple nonparametric model for S_t may be based on any periodic repetition of a function p_t defined on $t = 1, \ldots, s$, where s is the seasonality. For monthly data, we expect that $s = 12$; for quarterly data, $s = 4$, etc. Then the seasonal component is $S_t = p_{t(\mathrm{mod}\ s)}$, where t (mod s) is the integer remainder of t/s; for example 13 (mod 12) is 1. The function p_t may be estimated from the data by the simple averaging of appropriate times; for example, we can simply compute a January average, a February average, etc.

Figure 2 shows an analysis of the UK pulmonary deaths between 1974 and 1980. In this example, one may be interested in explaining the seasonal variation of the pulmonary deaths, which is high during the winter and low during the summer. If the interest is in prediction, it may be computed simply by extending the periodic component p_t beyond the last measurement. As one might expect, a strongly seasonal series is highly predictable. The residual series is the seasonally adjusted values $X_t - S_t$, centered at the mean number of deaths of each group. These adjusted series show, for example, that (a) the male deaths are decreasing over time, (b) the number of deaths during winter 1976 was unusually high and (c) the pattern is similar for males and females, which suggests a common cause in addition to the seasonal variation.

The periodic repetition model above is generally not satisfactory for many time series data as it does not allow any variation of the function p_t over time. Thus, a straightforward extension is to allow p_t to change slowly over time. One might say that such time series exhibit a stochastic seasonality. This is the basis of many seasonal adjustment procedures currently in use, such as the so-called X-11 program used by the US Bureau of the Census [10]. Instead of computing, say, a January average from the whole time series, the assumption of a slowly varying p_t suggests a local averaging of the nearby January values. In practice, some weights are usually applied when computing the average.

Models for Seasonality: Dynamic

The development of autoregressive and integrated moving-average (ARIMA) modeling by Box & Jenkins [1] (*see* **ARMA and ARIMA Models**) provides a rich class of dynamic linear models that also include stochastic seasonal models. Figure 1(d) shows a simulated ARIMA $(0, 1, 1) \times (0, 1, 1)_{12}$, i.e. a nonstationary seasonal moving-average process of the form

$$(1 - B)(1 - B^{12})X_t = (1 - 0.25B)(1 - 0.5B^{12})a_t, \tag{2}$$

where $BX_t \equiv X_{t-1}$, so the model specifies

$$X_t = X_{t-1} + X_{t-12} - X_{t-13} + a_t - 0.25a_{t-1} - 0.5a_{t-12} + 0.125a_{t-13}, \tag{3}$$

where a_t is an uncorrelated Gaussian series. This is, in fact, a fitted model for the milk production series (with slightly different parameter values). Note that the model generates a stochastic seasonality,

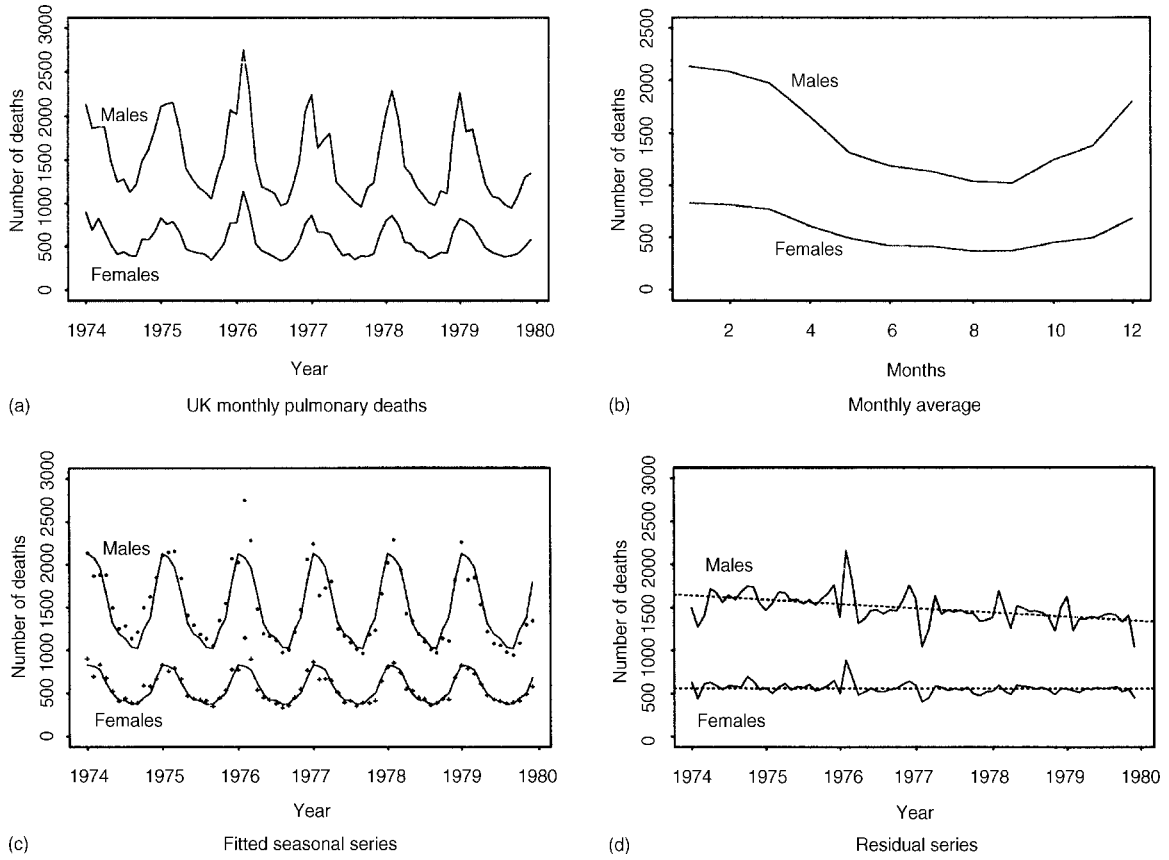

Figure 2 Analysis of UK pulmonary deaths. The seasonal component is computed by simple averaging, e.g. January average, February average, etc. The residual series is the so-called seasonally adjusted series, centered at the mean number of deaths

one where the periodic function p_t changes slowly over time. This has been the basis for a modified X-11 procedure called X-11-ARIMA developed by Statistics Canada. The advantage of the dynamic modeling is apparent for the estimation of S_t at the beginning and ending periods of observations [5]. Another approach is via a separate ARIMA modeling of the trend and seasonal components. Using a linear estimation theory, one can estimate S_t from the time series data X_t; see, for example, [2]. The set of weights generated by this method is, in fact, similar to that used in the X-11 program.

There is currently quite a large literature on seasonal models; a keyword search on "seasonal" in the Current Index of Statistics up to 1992 yields more than 500 records, most of which are related to seasonal modeling and adjustments. A recent reference is Hylleberg [8]. A literature study would indicate that most of the methodology for seasonal adjustment currently in use is either of the nonparametric type described above, or some ARIMA-based modification; a review article by Pierse [9] is still relevant. For a general statistical program, the X-11 procedure in SAS implements the Bureau of the Census X-11 program as well as X-11-ARIMA. The function stl() in **S-PLUS** performs nonparametric seasonal adjustment with some robustness capability as described in Cleveland et al. [3]. (*see* **Software, Biostatistical**).

Recent development of dynamic nonlinear models may also provide another wide class of periodic time series; Tong [11] is the main reference in this area. One of the most important properties of these models is the possible existence of limit cycles, a

phenomena that is absent within the linear models context. For series that exhibit a periodicity other than the natural cycles, a nonlinear model may yield a more satisfactory scientific description of the pattern; see, for example, [11, Chapter 7] for an extensive analysis of the Canadian lynx series.

References

[1] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

[2] Cleveland, W.P. & Tiao, G. (1976). Decomposition of seasonal time series: A model for the Census X-11 program, *Journal of the American Statistical Association* **71**, 581–587.

[3] Cleveland, R.B., Cleveland, W.S., McRae, J.E. & Terpening, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess, *Journal of Official Statistics* **6**, 3–73.

[4] Cryer, J.D. (1986). *Time Series Analysis*. Duxbury Press, Boston.

[5] Dagum, E.B. (1982). The effects of asymmetric filters on seasonal factor revisions, *Journal of the American Statistical Association* **77**, 732–738.

[6] Diggle, P. (1990). *Time Series: A Biostatistical Introduction*. Oxford Science Publications, Oxford.

[7] Elton, C. & Nicholson, M. (1942). The ten-year cycle in numbers of the lynx in Canada, *Journal of Animal Ecology* **11**, 215–244.

[8] Hylleberg, S. (1992). *Modelling Seasonality*. Oxford University Press, Oxford.

[9] Pierse, D. (1980). A survey of recent developments in seasonal adjustment, *American Statistician* **34**, 125–134.

[10] Shiskin, J., Young, A.H. & Musgrave, J.C. (1967). *The X-11 variant of the Census method II seasonal adjustment program*. Technical Paper No. 15, US Bureau of the Census.

[11] Tong, H. (1990). *Non-linear Time Series*. Oxford Science Publications, Oxford.

YUDI PAWITAN

Secondary Attack Rate

The secondary attack rate (SAR) is the probability that infection occurs among susceptible persons within a reasonable **incubation period** following known contact with an infectious person or another infectious source [7]. The SAR is conditional on the contact between an infectious source and a susceptible host, as opposed to the usual unconditional parameters of epidemiology such as the **incidence rate**, **hazard rate**, or **cumulative incidence** [10]. It is a special form of the transmission probability (*see* **Communicable Diseases**; **Infectious Disease Models**).

The term SAR is a misnomer, because it is actually a proportion, not a **rate**. The magnitude of the SAR depends on **covariates** of the infectious person and the susceptible person, the type of contact between the two, and the infectivity of the infectious agent. For given types of susceptibles, infectives, and contacts, it provides an epidemiologic measure of the infectivity of an infectious agent. Measles [1] and chicken pox [24] have relatively high household SARs of 86% or higher, while mumps has a lower household SAR of about 43% [13].

Study Designs to Estimate SAR

The most common study design to estimate the conventional SAR is first to identify infectious persons, and then to identify the susceptible people who make contact with them. The initially identified infectious persons are called the *primary* or *index cases*. The conventional SAR is estimated as the probability of the occurrence of disease among known (or presumed) susceptible persons following contact with a primary case:

$$\mathrm{SAR} = \frac{\begin{array}{c}\text{number of persons exposed}\\ \text{who develop disease}\end{array}}{\begin{array}{c}\text{total number of susceptible}\\ \text{exposed persons}\end{array}}. \qquad (1)$$

A clear definition of what is a contact is important in designing a study. It can vary from study to study for the same infectious agent. A potentially infective contact in a whooping cough study could be defined as being in a school on one day with someone with culture-proven whooping cough. Alternatively, it could be defined as living in the same house during the presumed period of infectiousness of a person with clinically diagnosed whooping cough. In a study of HIV transmission, a potentially infectious contact could be defined as each sex act between two sexual partners in a steady relationship, or as simply being in a partnership with someone who is infectious. The SAR is often defined for exposure to an infective of the susceptibles within some small population unit, such as a household, classroom, or school bus. Within any given unit, mixing and exposure of the susceptible persons to infection are usually assumed to be homogeneous.

Another approach to estimating the secondary attack rate is contact tracing. For example, upon identification of an infectious person with tuberculosis, public health officials locate people who have made contact with the infectious case and test them for whether or not they have become infected. The pooled estimate of the proportion who have become infected is an estimate of the SAR. The SAR can also be estimated in **experimental studies**. In studies of the infectivity of malaria in humans for mosquitoes, groups of 20–30 mosquitoes are fed experimentally on an infectious person. After 2 weeks, the mosquitoes are dissected to see if they have become infected and how many parasites have developed. The proportion of mosquitoes becoming infected is an estimate of the SAR.

The SAR has no explicit time dimension. However, the time interval during which the infected person is presumed to be infectious determines which of the people making contact were potentially exposed. Therefore, the time interval of infectiousness affects the determination of the denominator of the SAR. If infection is the outcome of interest, the minimum and maximum **latent periods** of the infectious agent define the time interval after exposure in which exposed people can develop infection and have been infected by the index case. If disease is the outcome of interest, the minimum and maximum incubation periods determine the time interval after exposure in which exposed people can develop disease and have been infected by the index case. Thus, either the latent period or incubation period enters into the determination of the numerator of the SAR.

Ratios of SARs

The ratio of two SARs can be used to estimate the relative infectivity or susceptibility of two types

of people, different infectious agents, or types of contacts. Let 0 and 1 represent two levels of a risk factor, such as vaccination status, gender, or age. Then there are four possible secondary attack rates for a given infectious agent and definition of contact – SAR_{00}, SAR_{11}, SAR_{10}, and SAR_{01} – where the first subscript represents the infectious person and the second subscript represents the susceptible person. The relative susceptibility of a person with risk factor level 1 compared to risk level 0 conditional on a given exposure to infection is SAR_{01}/SAR_{00} or SAR_{11}/SAR_{10}. The relative infectivity of a person with risk factor level 1 compared to risk factor level 0 is estimated by SAR_{10}/SAR_{00} or SAR_{11}/SAR_{01}. The relative transmissibility between persons of risk factor level 1 compared to that between persons with risk factor level 0 is SAR_{11}/SAR_{00}.

Vaccine efficacy can be estimated using the SAR (*see* **Vaccine Studies**). If 0 and 1 represent the unvaccinated and vaccinated people, respectively, then vaccine efficacy for susceptibility, VE_S, and infectiousness, VE_I, are estimated by

$$VE_S = 1 - \frac{SAR_{01}}{SAR_{00}}$$

and

$$VE_I = 1 - \frac{SAR_{10}}{SAR_{00}}. \quad (2)$$

The relative SAR between two vaccinated people compared to two unvaccinated people, i.e. the ratio SAR_{11}/SAR_{00}, can be thought of as the *naive susceptible equivalent* of a vaccinated compared to an unvaccinated person [11]. It gives the relative contribution of a vaccinated person to the basic **reproduction number**, R_0, compared to that of an unvaccinated person, and thus information about the effect of widespread vaccination on reducing the spread of an infectious agent in a population.

Household Secondary Attack Rate

The *household secondary attack rate* (SAR) is the probability that a susceptible individual living within the same household as an infectious person during their period of infectiousness will become infected. The household SAR is a parameter commonly used for estimating the protective efficacy of a vaccine in directly transmitted infections, such as pertussis, mumps, chicken pox, and measles [6, 22]. It can also be used to estimate the efficacy of a vaccine in reducing infectiousness [9, 10]. The data required are the time of onset of disease for each case in the household, as well as knowledge of who is susceptible. Estimates or assumptions about the minimum and maximum incubation periods, E_1 and E_2, respectively, and the maximum time I that a person remains infectious are also required and are sometimes obtained from other studies. One sometimes assumes that the onset of symptoms coincides with the onset of infectiousness, and that there are no asymptomatic cases.

The first step in assessing SAR is to define for the disease under study the time interval after the index case that would include secondary cases. The presumed beginning of infectiousness of the index case is defined as time 0 for each household. *Secondary cases* are those with time of onset between the end of the minimum incubation period E_1 relative to the beginning of infectiousness of the index case ($t = 0$) and the end of the maximum incubation period E_2 relative to the time of the maximum infectious period of the primary case, $t = I$. Thus, secondary cases are those occurring in the interval $(E_1, I + E_2)$. A case with recorded onset time less than one minimum incubation period, E_1, after that of the index case, was presumably not infected by the index case and is called a *co-primary case*. Tertiary and higher cases are those occurring after the maximum allowable time interval for the secondary cases.

Example

For an early efficacy study of pertussis vaccines, Kendrick & Eldering [12] estimated the infectious period for the bacteria from studies of throat cultures. In those studies, nearly everyone had a positive culture up to 21 days after onset of symptoms. They defined a definite exposure (potentially infective contact) as living in the same house as the index case or being indoors in another house with the index case for at least 30 min within $I_d = 21$ days of onset of symptoms of the index case. The **mean** incubation period of pertussis from two other studies was estimated to be 13 ± 7.6 days and 15.4 ± 1.3 days. Based on this information, Kendrick & Eldering somewhat arbitrarily set the minimum incubation period to $E_1 = 10$ days and the maximum incubation

period to $E_2 = 30$ days. Under the definition of definite exposure, secondary cases were those occurring between $E_1 = 10$ and $I_d + E_2 = 21 + 30 = 51$ days after the onset of symptoms in the index case.

Kendrick & Eldering had a second, less stringent, definition for a potentially infective contact that included outdoor contacts. Based on the observation that between 21 and 35 days after onset of symptoms, throat cultures were less often positive, someone exposed up to $I_i = 35$ days after onset of symptoms in the index case was defined as an indefinite exposure. For indefinite exposures, secondary cases were those occurring between day $E_1 = 10$ and $I_i + E_2 = 35 + 30 = 65$ days after the onset of symptoms of the index case.

The second step in assessing the SAR is to determine for each ascertained case within each household whether it is a co-primary, secondary, tertiary, or higher generation case. The estimated household SAR is the total number of secondary cases in all households divided by the total number of at-risk susceptibles in all households as in (1). Co-primary cases are excluded from the denominator. Tertiary or higher cases are excluded from the numerator but are included in the denominator.

Difficulties in estimating the conventional SAR include determination of the latent, incubation, and infectious periods, ascertainment of onset times of cases, and determining when an exposure to infection has taken place.

Inference

Possible **correlation** of responses among susceptibles exposed to the same infectious source need to be taken into account in making **inferences**. **Generalized estimating equations** (GEES) [15] using a logit model or the nested **bootstrap** [5] can be used for inference when estimating secondary attack rates [4]. The GEE approach is usually the preferred method.

Model-Based Approaches to Estimating SAR

One problem with the conventional SAR is that it does not take into account that susceptibles exposed to the index case could become infected from infectious sources other than the index case. The estimated SAR would be too high if there is a substantial possibility of becoming infected from other sources in the community. An alternative approach is a model for transmission of an infectious disease in a community of households that allows joint estimation of the SAR as well as the probability of becoming infected within the community, CPI, by the end of an outbreak [16]. In this model, the probability that during an outbreak exactly j persons become infected in a household with s susceptibles is

$$\begin{aligned}\pi_{js} &= \binom{s}{j}\pi_{jj}(1-\mathrm{CPI})^{(s-j)} \\ &\quad \times (1-\mathrm{SAR})^{j(s-j)}, \quad 0 \le j < s, \\ \pi_{ss} &= 1 - \sum_{j=0}^{s-1}\pi_{js}, \end{aligned} \tag{3}$$

where the **likelihood** function is the product of the probabilities π_{js}, $j = 0, 1, \ldots, s$, over all the households in the sample. Maximization of the likelihood provides **maximum likelihood** estimates of the SAR as well as the CPI. The advantage of this model over the conventional secondary attack rate method is that it requires only final value data, namely, who is in which household and who becomes infected during the outbreak. Disadvantages include relatively large **standard errors** and strong modeling assumptions.

Similar models have been used by others [2], as well as for estimating the effects of covariates [8, 19], including vaccine efficacy [14, 17, 18]. Another approach to estimation [20] is to use the **EM algorithm** [3] and a **generalized linear model** [21]. A discrete time model has been developed that allows estimation of the SAR from the time of onset data while adjusting for the possibility of infection from outside the household [23]. The disadvantage of the approach is that it requires the user to specify probability distributions for the latent, incubation, and infectious periods. **Chain binomial models** can also be used to estimate the SAR.

Summary

The secondary attack rate is defined as the probability that infection occurs among susceptible persons within a reasonable incubation period following

known contact with an infectious person or other infectious source. It is a key epidemiologic parameter in infectious diseases that are transmitted by contact. It can be estimated using a variety of epidemiologic study designs, models, and methods of estimation. Inference needs to take into account correlation of susceptibles exposed to the same infectious source.

References

[1] Bailey, N.T.J. (1957). *The Mathematical Theory of Epidemics*. Griffin, London.

[2] DeGruttola, V., Seage, G.R., Mayer, K.H. & Horsburgh, C.R. (1989). Infectiousness of HIV between male homosexual partners, *Journal of Clinical Epidemiology* **42**, 849–856.

[3] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[4] Dunson, D.B. & Halloran, M.E. (1996). Estimating Transmission Blocking Efficacy of Malaria Vaccines, *Technical Report 96–16*. Department of Biostatistics, Emory University, Atlanta.

[5] Efron, B. & Tibshirani, R.J., (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[6] Fine, P.E.M., Clarkson, J.A. & Miller, E. (1988). The efficacy of pertussis vaccines under conditions of household exposure: further analysis of the 1978–80 PHLS-ERL study in 21 area health authorities in England, *International Journal of Epidemiology* **17**, 635–642.

[7] Fox, J.P., Hall, C.E. & Elveback, L.R. (1970). *Epidemiology: Man and Disease*. Macmillan, New York.

[8] Haber, M., Longini, I.M., & Cotsonis, G.A. (1988). Models for the statistical analysis of infectious disease data, *Biometrics* **44**, 163–173.

[9] Halloran, M.E. (1996). Evaluating HIV vaccines: discussion, *Statistics in Medicine* **15**, 2405–2412.

[10] Halloran, M.E. & Struchiner, C.J. (1995). Causal inference for infectious diseases, *Epidemiology* **6**, 142–151.

[11] Halloran, M.E., Cochi, S., Lieu, T., Wharton, M. & Fehrs, L.J. (1994). Theoretical epidemiologic and morbidity effects of routine immunization of preschool children with live-virus varicella vaccine in the U.S., *American Journal of Epidemiology* **140**, 81–104.

[12] Kendrick, P. & Eldering, G. (1939). A study in active immunization against pertussis, *American Journal of Hygiene, Section B* **38**, 133.

[13] Kim-Farley, R., Bart, S., Stetler, H. et al. (1985). Clinical mumps vaccine efficacy, *American Journal of Epidemiology* **121**, 593–597.

[14] Koopman, J.S. & Little, R.J. (1995). Assessing HIV vaccine effects, *American Journal of Epidemiology* **142**, 1113–1120.

[15] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[16] Longini, I.M. & Koopman, J.S. (1982). Household and community transmission parameters from final distributions of infections in households, *Biometrics* **38**, 115–126.

[17] Longini, I.M., Datta, S. & Halloran, M.E. (1997). Measuring vaccine efficacy for both susceptibility to infection and reduction in infectiousness for prophylactic HIV-1 vaccines, *Journal of AIDS and HR*, to appear.

[18] Longini, I.M., Halloran, M.E., Haber, M.J. & Chen, R.T. (1993). Measuring vaccine efficacy from epidemics of acute infectious agents, *Statistics in Medicine* **12**, 249–263.

[19] Longini, I.M., Koopman, J.S., Haber, M. & Cotsonis, G.A. (1988). Statistical inference for infectious diseases: risk-specified household and community transmission parameters, *American Journal of Epidemiology* **128**, 845–859.

[20] Magder, L. & Brookmeyer, R. (1993). Analysis of infectious disease data from partners studies with unknown source of infection, *Biometrics* **49**, 1110–1116.

[21] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, London.

[22] Orenstein, W.A., Bernier, R.H. & Hinman, A.R. (1988). Assessing vaccine efficacy in the field: further observations, *Epidemiologic Reviews* **10**, 212–241.

[23] Rampey, A.H., Longini, I.M., Haber, M.J. & Monto, A.S. (1992). A discrete-time model for the statistical analysis of infectious disease incidence data, *Biometrics* **48**, 117–128.

[24] Ross, A.H. (1962). Modification of chicken pox in family contacts by administration of gamma globulin, *New England Journal of Medicine* **267**, 369–376.

M. ELIZABETH HALLORAN

Segregation Analysis, Classical

To determine the mode of inheritance of a genetic disease (or a trait), dichotomous (affected and unaffected) phenotypic data on members of families are generally collected. Given the phenotypes of parents, the probability of an offspring being affected depends on whether the **gene** responsible for the disease is dominant or recessive. Therefore, inference regarding the mode of inheritance of the disease is possible from an estimate of this probability. Statistical analysis of family data to determine the mode of inheritance is called *segregation analysis*.

The probability of an offspring being affected given parental phenotypes can be estimated from the proportion of affected offspring in a family. Consider, for example, a disease determined by a recessive allele (D) at a single autosomal biallelic locus with alleles D and d. Suppose in a family one parent is affected (**genotype** dd) and the other parent is unaffected (DD or Dd). Let R denote the number of affected offspring and S the total number of offspring in this family. If $R \neq 0$, then the unaffected parent is of genotype Dd. Hence, an estimator of $\theta = \Pr$ (an offspring is affected | parental genotypes) is R/S. θ is called the *segregation probability* or *segregation ratio*. R/S is actually the **maximum likelihood** estimator (MLE) of θ because the likelihood $L_\theta(s, r)$ of the observations $S = s$, $R = r$, conditional on parental genotypes is $\text{bin}(s, \theta)$. If, however, $R = 0$ (a *nonsegregating family*), then the genotype of the unaffected parent cannot be uniquely determined. If the parental mating is d d × DD, then the family is incapable of producing an affected offspring; hence, in this family, $\theta = 0$. Thus, ascertaining a family through the presence of an affected offspring (a *segregating family*) helps uniquely determine the genotypes of both parents for a simple recessive disease. If the disease allele D is dominant, even when the family is ascertained through an affected offspring, the genotypes of both parents cannot be uniquely determined; the affected parent can be of either genotype DD or Dd. If the mating is d d × DD, then the segregation probability is $\theta_1 = 1$. If the mating is d d × Dd, then this probability is $\theta_2 = 1/2$. The **likelihood** for the observations on s offspring of whom $r(\neq 0)$ are affected is a mixture of likelihoods

$$\omega_1 L_{\theta_1}(s, r) + \omega_2 L_{\theta_2}(s, r), \tag{1}$$

where $\omega_1 = \Pr(\text{d d} \times \text{DD} \mid \text{unaffected} \times \text{affected})$ and $\omega_2 = \Pr(\text{d d} \times \text{Dd} \mid \text{unaffected} \times \text{affected})$. ω_1 and ω_2 are functions of the D allele frequency in the population from which the family is drawn.

To avoid the complications arising from a mixture of likelihoods we shall make an assumption that enables unique identification of parental genotypes from parental and offspring phenotypes. *Classical segregation analysis* relates to analysis of such data on offspring for which the likelihood is not a mixture of likelihoods (*see* **Segregation Analysis, Complex**). We shall assume that the probability of the disease-causing allele in the population is low (say, 0.01 or 0.001). Under this assumption, in the example of the autosomal dominant disease given earlier, the parental mating will virtually always be d d × Dd. (For a recessive disease this assumption is not required to identify parental genotypes uniquely.)

For a rare genetic disease, random sampling of families is not a method of choice because all members of most randomly sampled families will be unaffected and hence will provide no information for estimating θ. A commonly used method is to ascertain families through the presence of at least one affected offspring. An affected individual in a family through whom the family can be ascertained is called a *proband*; the *probability of* **ascertainment** of a proband is defined as $\pi = \Pr$ (individual is a proband | individual is affected). Ascertainment can also be made through an affected parent. The dependent nature of familial data and adoption of nonrandom sampling schemes require specialized statistical methodology for analysis of such data. If an appropriate correction for bias of ascertainment (bias due to nonrandom sampling) is not made, the estimate of θ will be positively biased and inference on the mode of inheritance will be incorrect.

Statistical Methodology

Estimation of θ

If R denotes the number of affected offspring in a sibship of size S, then its probability density function

(pdf) is

$$p_\theta(S, r) = \binom{S}{r} \theta^r (1-\theta)^{S-r}, \tag{2}$$

$r = 0, 1, \ldots, S, 0 \le \theta \le 1$. However, if a family is ascertained through an affected offspring, then families with $R = 0$ will be excluded. Let $w(r, \pi)$ denote the probability of ascertaining a family with $R = r$. Hence, the pdf of R in ascertained families will be [10]

$$p^*_\theta(r, \pi, S) = \frac{w(r, \pi) p_\theta(S, r)}{\mathrm{E}[w(R, \pi)]}. \tag{3}$$

In the present case, since each ascertained family has at least one proband [9],

$$w(r, \pi) = 1 - (1-\pi)^r, \tag{4}$$

$$\begin{aligned} \mathrm{E}[w(R, \pi)] &= \sum_{r=1}^{S} w(r, \pi) p_\theta(S, r) \\ &= 1 - (1 - \pi\theta)^S. \end{aligned} \tag{5}$$

If $\pi = 1$, *complete ascertainment*, then $p^*_\theta(r, \pi, S)$ reduces to a truncated bin(S, θ) distribution, truncated at zero. If $\pi \simeq 0$, *single ascertainment*, then $w(r, \pi) \approx r\pi$, and $p^*_\theta(r, \pi, S)$ is a bin$(S-1, \theta)$ distribution.

In practice, ascertained families will have different numbers of offspring, both total and affected. If a_{rs} denotes the observed number of independently ascertained families each with $s(= 1, 2, \ldots, S)$ offspring of whom $r(= 1, 2, \ldots, s)$ are affected, then the joint likelihood for observations on all sibships will be

$$L_s(\theta) = \frac{n_s!}{\prod_{r=1}^{s} a_{rs}!} \prod_{r=1}^{s} [p^*_\theta(r, \pi, s)]^{a_{rs}}, \tag{6}$$

where $n_s = \sum_{r=1}^{s} a_{rs}$. Let $A = \sum_{s=1}^{S} \sum_{r=1}^{s} r a_{rs} =$ observed total number of affected offspring in all families. Then, it is easy to show [4] that the MLE of θ is obtained iteratively from the equation

$$\frac{A}{\theta} = \sum_{s=1}^{S} \frac{[1 - (1-\pi)(1-\pi\theta)^{s-1}] s n_s}{1 - (1 - \pi\theta)^s}, \tag{7}$$

where $\pi \in (0, 1]$ is assumed to be known. For $\pi \simeq 0$, an explicit solution of θ is easily obtained as

$$\hat{\theta} = \frac{A - N}{T - N}, \tag{8}$$

where $N = \sum_{s=1}^{S} n_s =$ total number of ascertained families, and $T = \sum_{s=1}^{S} s n_s =$ total number of offspring in all ascertained families. For $\pi = 1$, (7) reduces to

$$\frac{A}{N} = \sum_{s=1}^{S} \frac{s n_s}{1 - (1-\theta)^s}. \tag{9}$$

We note that the value of π may be unknown and may need to be estimated simultaneously with θ. The maximum likelihood estimation procedure, score vectors and **information matrices** are derived in [2] and [4] for both cases, π known and unknown. (See, however, some corrections in [1].) A simpler and computationally more efficient method of estimating θ (or, θ and π) using the **EM algorithm** is given in [1].

While for a rare recessive disease both parents in most families will be unaffected, for a relatively common disease at least one of the two parents may be affected, and thus families can be ascertained through an affected parent. If the disease is recessive, then to identify the genotype of the unaffected parent uniquely, we restrict ourselves to data on only those families in which there is at least one affected offspring. Then, the pdf of R (= number of affected offspring) in a sibship of size S will be

$$p_\theta(S, r | r > 0) = \frac{\binom{S}{r} \theta^r (1-\theta)^{S-r}}{1 - (1-\theta)^S}, \tag{10}$$

$r = 1, 2, \ldots, S$, which is identical to the pdf [truncated bin(S, θ) distribution] for complete ascertainment through offspring.

If the disease is dominant, then instead of analyzing data on affected persons, we can analyze data on unaffected persons, which is a recessive trait.

Thus, simple segregation analysis for families with at least one recessive offspring ascertained through an affected parent is the same as that for complete ascertainment through offspring.

Hypothesis Testing

In classical segregation analysis, we generally wish to test whether a disease is dominant or recessive. This **null hypothesis** $\mathrm{H}_0 : \theta = \theta_0$ is easily framed, because θ_0 is known for a specific parental genotypic mating under a model. For example, if the disease is

recessive, we know that in Dd × Dd families, $\theta_0 = 1/4$. Thus, from unaffected × unaffected families each ascertained through an affected offspring, we can obtain $\hat{\theta}$ and calculate the test statistic

$$X^2 = \frac{(\hat{\theta} - \theta_0)^2}{V(\theta_0)}, \tag{11}$$

where $V(\theta_0)$ is the variance of θ evaluated at θ_0. X^2 follows a **chi-square distribution** with 1 df.

Efficient Approximate Methods

For complete ascertainment ($\pi = 1$), Li & Mantel [8] suggested a simple method, the *singles method*. The estimator of θ is

$$\hat{\theta}_{\mathrm{LM}} = \frac{A - J_1}{T - J_1}, \tag{12}$$

where J_1 denotes the number of families with only one affected offspring. Gart [6], who independently proposed the method, has shown that $\hat{\theta}_{\mathrm{LM}}$ is almost fully efficient as the MLE for all realistic values of S.

It may be noted that for single ascertainment ($\pi \simeq 0$), Weinberg [11] proposed a method, the *proband method*, and derived the simple estimator given by (8), which was shown to be the MLE by Haldane [7].

For *incomplete ascertainment* ($0 < \pi < 1$), Weinberg [12] had initially suggested an estimator that was modified by Fisher [5]. However, Davie [3] showed that these estimators are not very efficient and proposed another estimator which he showed to be very efficient at all levels of ascertainment. This estimator is

$$\hat{\theta}_{\mathrm{D}} = \frac{R - J}{T - J}, \tag{13}$$

where J denotes the number of families having exactly one proband. Not only does $\hat{\theta}_{\mathrm{D}}$ becomes identical to $\hat{\theta}_{\mathrm{LM}}$ for complete ascertainment and to $\hat{\theta}$ of (8) for single ascertainment, the large sample variance of $\hat{\theta}_{\mathrm{D}}$ also becomes identical to the variances of the corresponding estimates at these two extremes of ascertainment.

References

[1] Achuthan, N.R. & Krishnan, T. (1992). EM algorithm for segregation analysis, *Biometrical Journal* **8**, 971–988.

[2] Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Clarendon Press, Oxford.

[3] Davie, A.M. (1979). The singles method for segregation analysis under incomplete ascertainment, *Annals of Human Genetics* **42**, 507–512.

[4] Elandt-Johnson, R.C. (1971). *Probability Models and Statistical Methods in Genetics*. Wiley, New York.

[5] Fisher, R.A. (1934). The effect of methods of ascertainment upon estimation of frequencies, *Annals of Eugenics* **6**, 13–25.

[6] Gart, J.J. (1967). A simple nearly efficient alternative to the simple sib method in the complete ascertainment case, *Annals of Human Genetics* **31**, 283–291.

[7] Haldane, J.B.S. (1938). The estimation of the frequencies of recessive conditions in man, *Annals of Eugenics* **8**, 255–262.

[8] Li, C.C. & Mantel, N. (1968). A simple method of estimating the segregation ratio under complete ascertainment, *American Journal of Human Genetics* **20**, 61–81.

[9] Morton, N.E. (1959). Genetic tests under incomplete ascertainment, *American Journal of Human Genetics* **11**, 1–16.

[10] Rao, C.R. (1965). On discrete distributions arising out of methods of ascertainment, in *Classical and Contagious Discrete Distributions*, G.P. Patil, ed. Statistical Publishing Society, Calcutta, pp. 320–333.

[11] Weinberg, W. (1912). Weitere Beiträge zur Theorie der Vererbung. 4. Über Methode und Fehlerquellen der Untersuchung und Mendelsche Zahlen beim Menschen, *Archiv für Rassen- und Gesellschafts-Biologie* **2**, 165–174.

[12] Weinberg, W. (1912). Weitere Beiträge zur Theorie der Vererbung. 5. Zur Vererbung der Anlage zur Blutenkranheit mit methodologischen Ergänzungen meiner Geschwistermethode, *Archiv für Rassen- und Gesellschafts-Biologie* **2**, 694–709.

PARTHA P. MAJUMDER

Segregation Analysis, Complex

Complex segregation analysis refers to a statistical genetic method that focuses on the detection and characterization of the unobserved **genes** that influence phenotypic variation. It was originally termed "complex" because, unlike **"classical" segregation analysis**, it considers the situation in which more than one mating type is possible for a given sibship, thus requiring summation over mating types in the **likelihood** [19]. The adjective "complex" is equally appropriate because the phenotypes that it considers have multiple determinants (e.g. multiple loci, environmental factors, etc.). Mendel first discovered the transmission laws that bear his name by the examination of continuous characters of pea morphology (*see* **Mendel's Laws**). He was fortunate that the underlying **genotypic** distributions were essentially nonoverlapping and thus could easily be discretized in a biologically meaningful way and that the leading factors determining them were simple two-allele **polymorphisms**. However, for most characters, the form of inheritance is more elaborate. Multiple genes interact to influence most traits of relevance to biomedicine. Additionally, most of the physiological characters that are important in normal and pathological variation are also influenced by the environment. Ultimately, genes and environment interact (*see* **Gene-environment Interaction**) to determine the phenotype.

The primary goal of modern **human genetics** is to disentangle this complex web of interacting variables and to determine the role of genetic variation in health and disease. Most of the genes influencing complex phenotypes exhibit small effects, although in aggregate their joint effect may be large. However, it is clear that some of the genes influencing phenotypic variation have moderate to large effects and that the signal of their Mendelian transmission pattern may be seen in the examination of variation within and between families. It is the goal of segregation analysis to detect these leading genetic factors and to provide statistical descriptions of their essential features. The formal application of complex segregation analysis is used to infer the inheritance model for a trait and to provide estimates of the underlying genetic parameters. Most often, complex segregation analysis is used in an attempt to determine whether the transmission pattern of a phenotype within a family is consistent with Mendelian expectations (*see* **Genetic Transition Probabilities**). The basic strategy of complex segregation analysis is to fit a series of inheritance models, including nongenetic models, to family data, and then to select the model that best explains the observed data.

There are three basic models of genetic inheritance. Their distinctions are largely quantitative, yet inexact. When a phenotype is dominated by the effects of a single locus, and there is no evidence for any residual genetic effects, the inheritance pattern is called monogenic. If the trait in question is influenced by a few loci, then the inheritance is said to be oligogenic. Finally, when the trait is influenced by a large number of loci each with small effects, we have **polygenic inheritance**. In our search to elucidate the genetic architecture of complex phenotypes, we are primarily interested in finding large genetic effects due to specific loci. Therefore, for those traits whose inheritance pattern is monogenic or oligogenic, we hope to be able to characterize the primary loci influencing observed phenotypic variability.

The methods of complex segregation analysis are used to analyze many different types of phenotypes from continuous characters to meristic traits to discrete phenotypic states, although some latent continuous distribution is usually associated with discrete traits. In the following, we primarily focus on the analysis of continuous traits (*see* **Genetic Liability Model** for discrete traits).

The Model

In this Section, the basic model utilized in complex segregation analysis of quantitative traits is described. Table 1 presents the definition of the canonical parameters of quantitative trait segregation analysis to be explicated. The reader should bear in mind that there are many equivalent alternative parameterizations.

Modeling the Genotype

Since the goal of segregation analysis is to determine whether or not there is sufficient evidence for the effects of a specific locus influencing variation in a trait, it is necessary to consider more general alternative models in which mixtures of environmental

Table 1 Basic parameters of complex segregation analysis for a quantitative trait

Parameter	Definition
p_A	Frequency of A allele (factor)
$\mu_{AA}, \mu_{Aa}, \mu_{aa}$	Genotypic (ousiotypic) means
h_r^2	Residual heritability
σ_r	Residual phenotypic standard deviation
$\tau_{AA}, \tau_{Aa}, \tau_{aa}$	Transmission probabilities

origin are allowed. In most applications of complex segregation analysis, the number of components in the mixture distribution is limited to three. These three distributions can be related to unobservable genotypes or, more generally, ousiotypes [13], with or without genetic inheritance.

Ousiotypes are the product of two discrete factors, A or a. Upper case letters (e.g. A) represent factors associated with lower levels of the quantitative trait, and lower case letters represent factors associated with higher levels. The three genotypes can be denoted as AA, Aa, and aa. To simplify computations, the frequencies of the genotypes are usually assumed to follow **Hardy–Weinberg** proportions $\boldsymbol{\psi} = [\psi_{AA}, \psi_{Aa}, \psi_{aa}]' = [p_A^2, 2p_A(1-p_A), (1-p_A)^2]'$, where p_A is the frequency of the A factor (allele); thus only one **admixture** parameter (p_A) is required. The elements of $\boldsymbol{\psi}$ provide the probabilities that an individual in the founding (i.e. parental) population has a particular genotype. As an alternative parameterization, frequencies of genotypes can be directly estimated, requiring two parameters (ψ_{AA}, and ψ_{Aa}, since $\psi_{aa} = 1 - \psi_{AA} - \psi_{Aa}$). This relaxation of the Hardy–Weinberg equilibrium assumption may be more appropriate in populations where we know that random mating does not hold (e.g. small isolated populations). Regardless, it is assumed that the probability of a mating between two individuals with genotypes ij and kl, respectively, is simply $\psi_{ij}\psi_{kl}$ [20].

Modeling Transmission of the Genotype

The focal source of nonindependence among relatives in complex segregation analysis is due to the transmission of factors (alleles) between generations from parents to offspring. Mendelian segregation provides a decidedly nonrandom and systematic form of transmission that is unlikely to be mimicked by environmental agents. Therefore, in the absence of having a directly measurable genotype, the inference that a gene is influencing a particular phenotype is based on testing whether the observed pattern of phenotypic variation within families is consistent with Mendelian transmission. If there is no transmission, then there is no nonindependence due to the genotype and offspring genotypes will not be a function of parental genotypes. All of the essential features of genotype transmission are contained in a set of probabilities that are a function of a vector of arbitrary transmission parameters written as $\boldsymbol{\tau} = (\tau_{AA}, \tau_{Aa}, \tau_{aa})'$, whose elements denote the probability that an individual of a given genotype transmits factor "A" to an offspring [22]. Conversely, the probability of transmitting factor "a" is given by $\mathbf{1} - \boldsymbol{\tau}$. When Mendelian transmission holds, $\boldsymbol{\tau} = (1, 1/2, 0)$. With these three basic parameters, all of the **genetic transition probabilities** for offspring can be obtained. For the $m^2 = 9$ possible mating types, we can obtain the 9×3 matrix of probabilities for offspring genotypes by

$$\mathbf{T} = \begin{bmatrix} \boldsymbol{\tau} \otimes \boldsymbol{\tau} \\ \boldsymbol{\tau} \otimes (\mathbf{1} - \boldsymbol{\tau}) + (\mathbf{1} - \boldsymbol{\tau}) \otimes \boldsymbol{\tau} \\ (\mathbf{1} - \boldsymbol{\tau}) \otimes (\mathbf{1} - \boldsymbol{\tau}) \end{bmatrix}', \qquad (1)$$

where $\otimes$ is a Kronecker product operator.

Modeling the Phenotype

The phenotype of an individual is usually assumed to be a linear function of a set of **fixed effects** and **random effects**. Under the **mixed model**, which includes both a major factor and a residual polygenic component [22, 36, 38], the phenotype of the jth individual with genotype i is

$$(y_j | o_j = i) = \mu_i + g_j + \boldsymbol{\beta}'(\mathbf{x}_j - \mathbf{s}) + e_j, \qquad (2)$$

where o is the genotype, $\mu_i (i = AA, Aa, aa)$ is the mean associated with the ith genotype, and $\mathbf{x}_j$ is the jth individual's vector of **covariates** (i.e. the jth row of $\mathbf{X}$) scaled to some baseline $\mathbf{s}$. In this model, the genotype represents an unobservable discrete random effect and the covariates represent fixed effects. The g and e terms represent random effects, with g being an additive polygenotypic effect and e being a random environmental deviation or random error. Assuming that $E(g) = E(e) = 0$, the expectation of (2) is

$$E(y_j | o_j = i) = \mu_i + \boldsymbol{\beta}'(\mathbf{x}_j - \mathbf{s}). \qquad (3)$$

It is generally assumed that the random effects g and e are **normally distributed** so that the polygenotypic effect can be thought of as the cumulative sum of additive genetic effects at a large number of loci [25]. It is also assumed that the two random factors are independent so that $\text{cov}(g, e) = 0$. Thus, there is no genotype–environment correlation. Following these assumptions, the conditional variance of the y is written as

$$\text{var}(y_j|o_j = i) = \sigma_g^2 + \sigma_e^2. \tag{4}$$

The genetic component (σ_g^2) of the conditional variance given in (4) represents the residual additive genetic variance. This residual polygenic component is useful to absorb genetic effects other than the potential major locus. Residual nonindependence among relatives due to biological kinship is allowed by including this genetic component (*see* **Genetic Correlations and Covariances**), the residual **heritability** parameter (h_r^2) refers to the proportion of phenotypic variance due to additive genetic variance within each genotype's phenotypic distribution, and $\sigma_r = (\sigma_e^2 + \sigma_g^2)^{1/2}$ is the within-genotype standard deviation. Therefore, the proportion of phenotypic variance attributable to random environmental variation within genotypes is given by $1 - h_r^2$.

Since g and e are assumed to be normally distributed and the convolution of two normal densities is again normal, the conditional density of y is normal. This density is also known as the **penetrance** function in genetic terminology. It provides the probability density for having a specific phenotypic value given a specific genotype.

The unconditional variance of y has an additional variance component attributable to the effect of the major factor:

$$\text{var}(y) = \sigma_o^2 + \sigma_g^2 + \sigma_e^2, \tag{5}$$

where $\sigma_o^2 = \sum \psi_i(\mu_i - \sum \psi_i \mu_i)^2$ is the variance due to the major factor (locus), so that the relative proportion of phenotypic variance (after controlling for covariates) that is due to the major factor is $h_o^2 = \sigma_o^2/(\sigma_o^2 + \sigma_g^2 + \sigma_e^2)$. If a gene is the source of this genotypic variation, then h_o^2 is the heritability due to this gene. Generally, if the relative variance due to a locus is approximately 15% or larger, then it is termed a major gene. However, rare alleles with large displacements (i.e. the difference between the means of contrasting genotypes) of several standard deviations (so-called megaphenic effects) may also be called major genes, regardless of their relative importance at the population level.

The Likelihood

The three basic parts of the model outlined above provide the necessary components for developing the likelihood function of a phenotypic vector in a group of relatives. Generally, the term pedigree can be used for any set of biologically related individuals and their mates. Useful pedigrees for segregation analysis may be as simple as nuclear families or as complex as multigenerational extended kindreds. The key relationship in segregation analysis is that of parent–offspring, which is the fount of information regarding Mendelian segregation. As multiple generations are added, segregations can be followed further down the descendent chain, thus improving the ability to detect Mendelian transmission. In this Section, the general form of the likelihood for a pedigree of arbitrary size and complexity is described.

Joint Density Function of Genotypes and Quantitative Phenotypes

Let $\mathbf{O}$ denote the $n \times 3^n$ matrix containing all possible genotypic combinations for a given pedigree of size n. Given a vector of genotypes $\mathbf{o}_j$ (which denotes the jth column of the $\mathbf{O}$ matrix), the joint probability density of genotypes and quantitative phenotypes can be written as

$$f(\mathbf{o}_j, \mathbf{y}) = f(\mathbf{o}_j) f(\mathbf{y}|\mathbf{o}_j). \tag{6}$$

The first factor on the right-hand side of (6) gives the probability of observing the genotypic vector. This probability is a function of the frequency of the major factors (alleles) and the transmission probabilities, and is written

$$f(\mathbf{o}_j) = \prod_{i=1}^{n_F} \psi_{o_{ij}} \prod_{k=n_F+1}^{n} \Pr(o_{kj}|o_{fj}, o_{mj}, \boldsymbol{\tau}). \tag{7}$$

In (7), $\psi_{o_{ij}}$ is the probability of observing the genotype of individual i, which is equal to the population frequency of the genotype. The term $\Pr(o_{kj}|o_{fj}, o_{mj}, \boldsymbol{\tau})$ is the probability of an individual exhibiting genotype k given his father's and mother's genotypes, which is simply the appropriate element

of the **T** matrix obtained in (1). The first product in (7) is over the n_F founders (individuals whose parents are not represented in the pedigree), while the second product is over the $n - n_F$ nonfounders. Eq. (7) shows that the probability of the genotypic vector can be decomposed into a series of univariate densities, one for each individual. This is possible because of the **Markov** pattern of dependence in which the genotype of the individual depends at most on those of his/her parents.

The second factor on the right-hand side of (6) is the conditional density of **y** given the genotypic vector, and is assumed to take the following **multivariate normal** form:

$$f(\mathbf{y}|\mathbf{o}_j) = 2\pi^{-n/2}|\boldsymbol{\Omega}|^{-1/2} \times \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{Z}_j\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{Z}_j\boldsymbol{\mu})\right], \quad (8)$$

where $\mathbf{Z}_j$ is an $n \times 3$ indicator matrix whose ith row is $[\delta_{AA}, \delta_{Aa}, \delta_{aa}]$, with δ_{kl} equal to 1 only if the ith individual has genotype kl and equal to 0 otherwise. In the above density, the phenotypic **covariance matrix** of **y** conditional on knowledge of $\mathbf{o}_j$ is $\boldsymbol{\Omega}$. For the standard mixed model, this residual phenotypic covariance matrix for the pedigree can be written by

$$\text{var}(\mathbf{y}|\mathbf{o}_j) = \boldsymbol{\Omega}, \quad (9)$$

$$= 2\boldsymbol{\Phi}h_r^2\sigma_r^2 + \mathbf{I}_n(1 - h_r^2)\sigma_r^2, \quad (10)$$

where $\boldsymbol{\Phi}$ is a matrix whose ijth element is twice the coefficient of kinship between members i and j (*see* **Inbreeding**), and $\mathbf{I}_n$ is an $n \times n$ identity matrix.

From the conditional joint density in (6), we can now write the likelihood function for a pedigree as

$$L(\boldsymbol{\mu}, \sigma_r, h_r^2, p_A, \boldsymbol{\tau}|\mathbf{y}, \mathbf{O}) = f(\mathbf{O}, \mathbf{y}) \quad (11)$$

$$= \sum_{j=1}^{3^n} f(\mathbf{o}_j) f(\mathbf{y}|\mathbf{o}_j), \quad (12)$$

where the summation is over all possible genotypic vectors. Eq. (12) shows that the underlying distribution of the quantitative trait is made up of a mixture of multivariate normal distributions. It also shows that the model allows for statistical nonindependence among pedigree members due to both the transmission of the oligogenes and the residual polygenic background. The practical utility of the above exact likelihood formulation is rather limited since it requires repeated inversions of the potentially large $n \times n$ matrix, $\boldsymbol{\Omega}$, and the evaluation of all 3^n possible multivariate conditional likelihoods. For large pedigrees, such a direct approach is likely to be intractable. For example, a moderately sized human pedigree with 20 members requires the evaluation of approximately 3.5 billion genotypic vectors. Therefore, for practical applications, it is necessary to consider some algorithmic improvements and likelihood approximations.

Calculating the Likelihood: Approximations and Alternatives

The model presented above poses formidable computational difficulties. A great deal of research has been oriented towards reducing this burden. A number of recursive algorithms, alternative model and likelihood formulations, and likelihood approximations have been suggested [5, 7–9, 17, 22–24, 26–29, 31, 42] (*see* **Elston–Stewart Algorithm** for a discussion of probability calculations on pedigrees).

Approximating the Mixed Model Likelihood

When there is a residual polygenic component, it is impossible to write the likelihood in a simple form. We can write the density of y given by (11) and (8) as

$$f(\mathbf{O}, Y) = \sum_{o_1}\sum_{o_2}\cdots\sum_{o_n}\prod_{i=1}^{n_F}\psi_{o_{ij}}\prod_{k=n_F+1}^{n}\Pr(o_{kj}|o_{fj}, o_{mj}, \boldsymbol{\tau}) \times 2\pi^{-1/2}|\boldsymbol{\Omega}|^{-1/2}\prod_{i=1}^{n}\exp\left[-\frac{\omega^{ii}}{2}(y_i - \mu_{o_i})^2\right] \times \prod_{i=1}^{n-1}\prod_{j=j+1}^{n}\exp\left[-\omega^{ij}(y_i - \mu_{o_i})(y_j - \mu_{o_j})\right],$$

where ω^{ij} is the ijth element of the inverse of the residual covariance matrix $\boldsymbol{\Omega}$ [24, 28]. The complexity lies in the cross-product terms since there will not be complete knowledge of the genotype for both members of any relative pair, other than parent–offspring and spousal pairs, during the appropriate step in the peeling process. There is no simple solution to this problem. Therefore, in practice, approximation is employed to reduce computations.

While this density can also be written in integral form [5, 22, 23, 27], most approximations implicitly involve weighting the cross-product terms using the current approximate probabilities of genotypes [17, 27, 28].

Alternative Formulations

There are several alternative formulations of the mixed model (if we define the mixed model loosely as any model allowing both focal genotypic and residual genetic effects). One option is to employ residual genotypic effects directly. For example, Morton [35, 37] has argued for absorbing residual genetic variation using an oligogenic model incorporating a second locus with two alleles. Although this is a crude approximation to a residual polygenic component, it will tend to absorb much of the residual genetic variation. Additionally, the likelihood of such a model can be calculated exactly and rapidly. Other available alternatives include the finite polygenic mixed model (*see* **Polygenic Inheritance**) and the **regressive models**.

A disadvantage of all of these alternatives is that they lack the potential generality that is implicit in the **variance component** form of the likelihood. This variance component model can be extended easily to allow additional variance components [23, 28] and complexities such as genotype × environment interaction [3] that can lead to violation of the assumptions regarding conditional independence among relatives given parental genotypes.

Likelihood Corrections for Nonrandom Ascertainment of Pedigrees

The likelihood framework sketched above is only appropriate when pedigrees have been ascertained randomly (i.e. sampled without regard to particular phenotypic configurations). Typically, when dealing with diseases that may be rare, it is necessary to sample pedigrees nonrandomly so that they are enriched for the disease or for higher (or lower) values of a quantitative phenotype that is a concomitant of the disease. In such cases, the sampling mechanism by which the selection of a pedigree has occurred needs to be taken into account by the likelihood model. If nonrandom ascertainment is ignored, then the estimation of some of the parameters may be severely **biased**. This is particularly true for allele frequencies and relative variance components due to genetic factors. Therefore, ascertainment corrections to the likelihood typically require that the selection mechanism be known (*see* **Ascertainment**; **Pedigrees, Sequential Sampling** for a more detailed discussion).

Estimation

Maximum Likelihood Estimation

Given that the likelihood can be calculated exactly or approximately, estimates of the parameters of complex segregation analysis are usually obtained by the standard **maximum likelihood** method. Maximization of such a complicated likelihood represents a considerable numerical problem, with the primary difficulty being the high probability of observing multiple maxima (a problem associated with finite mixture models in general). One approach to this problem is to search the likelihood surface for all existing maxima by using random initial parameter estimates in a **Monte Carlo** procedure [1]. While we are primarily interested in the global maximum, sometimes local maxima appear to supply important additional information about the underlying genetic model [1, 4, 6].

Markov Chain Monte Carlo Estimation

One relatively recent approach to parameter estimation in complex segregation analysis involves the use of **Markov chain Monte Carlo** (MCMC) methods [26]. In this approach, the latent genotypes and polygenotypes are imputed conditional upon the phenotypic information and current parameter values using a Monte Carlo **algorithm**. After this "missing" information is completely filled in for a pedigree, estimation using the augmented data is straightforward since it utilizes closed-form estimators. The process is iterated until convergence. The strength of this method is that it can accommodate models of great complexity in which it would be difficult, if not impossible, to calculate the likelihood directly. The MCMC approach, therefore, opens the door for the application of more realistic models in segregation analysis. Also, this method allows the possibility of performing **Bayesian** estimation and inference, if reasonable functional forms for the **prior distributions** of the parameters can be specified [43]. Using this method, it is also possible either

to estimate **likelihood ratio test** statistics to compare models [44] or alternatively to estimate the likelihood directly [43].

Estimation Using Generalized Estimating Equations

Another recent approach to parameter estimation in complex segregation analysis involves the application of **generalized estimating equations** (GEE) [33, 46–48]. This method of estimation makes fewer assumptions about the distributional form that the mixture of genotypes takes and therefore may be more **robust**. It involves the estimation of parameters using the first few **moments** of the observed distribution by relating observed estimates of moments to those expected under the model. A set of potentially high-dimensional linear equations are solved to obtain estimates of the parameters. It has been shown that evaluation of the first four moments is required to achieve statistical identification of the parameters of the mixed model [48]. Since these moments involve all possible tricovariances and quadricovariances within pedigrees, the dimension of the matrices requiring inversion increases nonlinearly as pedigree size is increased.

Comparing Models of Inheritance

Competing Transmission Models

In practice, complex segregation analysis involves the evaluation of several possible models of inheritance and their comparison with a general model in which transmission of the genotype is allowed to take an arbitrary form (i.e. the transmission probability parameters ($\boldsymbol{\tau}$) are allowed to vary freely). Each of the models to be compared to the general model represents a nested submodel of the general model.

Several classes of restricted models can be tested against the most general model using the unified approach of Lalouel et al. [32]. The simplest models generally considered include sporadic models which allow only random environmental effects. In this type of model, all individual trait values are independent of one another. Therefore, there are no genetic factors acting on the trait.

When multiple distributions are considered, a sporadic model becomes a simple finite mixture (or **commingling**) model. It is obtained by forcing the admixture parameter (p_A) to equal the transmission parameters ($p_A = \tau_{AA} = \tau_{Aa} = \tau_{aa}$). This transmission model preserves the assumption of equilibrium since the expected genotypic frequencies do not change from generation to generation. By allowing p_A to vary independently of the τs, this model permits heterogeneity of mixture proportions between generations.

A closely related class of model, the environmental transmission model, assumes random environmental effects for major factors, but also permits residual polygenic inheritance (i.e. it is a finite mixture model extended to allow nonindependence due to genetic kinship among individuals). Again, two different constraints on the transmission probabilities are possible: $p_A = \tau_{AA} = \tau_{Aa} = \tau_{aa}$ or $p_A \neq (\tau_{AA} = \tau_{Aa} = \tau_{aa})$, depending on whether it is decided to force equilibrium between generations. A model with only one underlying phenotypic distribution allowing for a polygenic component of variation reduces to the classical additive polygenic model of quantitative genetics.

The Mendelian models considered incorporate transmission probabilities fixed at their Mendelian expectations ($\tau_{AA} = 1, \tau_{Aa} = 1/2, \tau_{aa} = 0$). Mixed Mendelian models additionally allow for a residual polygenic background. Subsets of Mendelian models include: (i) additive models in which the allelic effects act additively to determine the mean of a genotype ($\mu_{Aa} = [\mu_{AA} + \mu_{aa}]/2$); (ii) recessive models in which the "a" allele is recessive, leading to $\mu_{AA} = \mu_{Aa}$; and (iii) dominant models in which the "a" allele is dominant so that $\mu_{Aa} = \mu_{aa}$. Analogous constraints on the genotypic means can be made for both the environmental and general models.

One potential problem with the general model is that it does not necessarily preserve the equilibrium between generations [16]. While this provides some flexibility with regard to model fitting, it may not always be biologically relevant. Therefore, if it is desirable to maintain the assumption of equilibrium, then the following nonlinear constraint can be placed on the transmission probability for the heterozygote:

$$\tau_{Aa} = \frac{p_A - p_A^2\tau_{AA} - (1 - p_A)^2\tau_{aa}}{2p_A(1 - p_A)}. \qquad (13)$$

Likelihood Ratio Tests

Depending upon the availability of a sufficient data structure, all parameters can be estimated by

numerical maximization of the likelihood for the data given the assumed transmission model. Model comparison is usually performed using likelihood ratio tests in which the test statistic is defined by

$$\Lambda = 2\left[\text{Sup}\ln L(\hat{\theta}_{H_0+H_A}|\mathbf{y}) - \text{Sup}\ln L(\hat{\theta}_{H_0}|\mathbf{y})\right],$$

where $\hat{\theta}_{H_A+H_0}$ refers to the maximum likelihood parameter estimates under the alternative hypothesis, and $\hat{\theta}_{H_0}$ refers to the estimates under the **null hypothesis**. This test statistic is simply twice the difference between the $\log_e$ likelihoods of the unrestricted and restricted models. In some cases, these test statistics are asymptotically distributed as **chi-square** variates with degrees of freedom equal to the difference in the number of parameters between the two competing models. However, for the main contrast that we are interested in, that of the general transmission model ($\hat{\tau}_{AA}, \hat{\tau}_{Aa}, \hat{\tau}_{aa}$) against the Mendelian mixed model ($\tau_{AA} = 1, \tau_{Aa} = 1/2, \tau_{aa} = 0$), two of the constrained transmission parameters of H_0 fall on a boundary of the parameter space, $\tau_{AA} = 1$ and $\tau_{aa} = 0$. The resulting Λ is not distributed as a χ^2 variable but as a complex mixture of χ^2 distributions [15, 41]. If the comparison involves the equilibrium-constrained general model in which τ_{Aa} is estimated by (13), then $\Lambda \sim \frac{1}{4} + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$ and the ***P* value** can be calculated easily [41].

Inferring the Presence of a Major Gene

The inference that a major gene is influencing a trait requires the sequential elimination of a number of competing hypotheses. When compared against the general transmission model, each of the nested submodels (e.g. the sporadic model, the polygenic model, and the environmental model) should be rejected, while the Mendelian mixed model should not exhibit a significantly worse likelihood than the general transmission model. Rejection of the environmental model is particularly important since such a test effectively guards against simple distributional **skewness** being interpreted as a major gene effect. This conservative testing framework protects against spurious findings of major loci but unfortunately tends to lead to diminished **power** to detect major genes [11, 12].

Comparing Nonnested Inheritance Models

Sometimes it is useful to compare pairs of nonnested models. In this case, there is no asymptotic theory for the distribution of the likelihood ratio test. Generally, when nonnested models must be compared, **Akaike's criterion** (AIC) can be used to choose the most parsimonious model, or the distribution of the likelihood ratio test statistic found empirically using a Monte Carlo procedure such as the parametric **bootstrap** [40].

Extensions of Complex Segregation Analysis

Genotype-Specific Regressions and Genotype × Environment Interaction

The model for the phenotype given in (2) contains a number of simplifying assumptions that can be removed. For example, different genotypes may exhibit different relationships with covariates. A number of authors have used genotype-specific **regressions** on covariates to model genotype–environment interaction [3, 18, 30, 34, 39]. Under the genotype-specific regression model, the phenotype of the jth individual (with genotype $o_j = i$) is given by

$$(y_j|o_j = i) = \mu_i + g_j + \boldsymbol{\beta}_i'(\mathbf{x}_j - \mathbf{s}) + e_j, \qquad (14)$$

where the vector of regression coefficients $\boldsymbol{\beta}_i$ is now a function of the major genotype. The conditional variance of y in this model is the same as that in (4). The unconditional variance has an additional component due to this **interaction** and is given by

$$\text{var}(y) = \sigma_o^2 + \sigma_{o\times c}^2 + \sigma_e^2 + \sigma_g^2, \qquad (15)$$

where $\sigma_{o\times c}^2$ denotes the variance due to major genotype by covariate interaction. Tests of heterogeneity of these effects among genotypes can be performed using likelihood ratio statistics. The test of $\beta_{AA} = \beta_{Aa} = \beta_{aa}$ is a direct test of genotype × environment interaction. Rejecting this null hypothesis leads to the inference that there is a major locus component in the response of the phenotype to the environment [3]. However, to safeguard against falsely inferring the presence of genotype × environment, it is necessary to also consider tests of polygenotype × environment interaction [3]. These latter tests require application of the fully general variance component model. Ignoring genotype × environment interaction may severely compromise the ability to detect major loci [3, 30, 45].

Two-Locus Models

The model can also be extended to allow two latent factors yielding two sets of genotypes/genotypes. If we assume (or have determined statistically) that both factors represent genetic loci, we can also allow these two genotypes to exhibit epistasis (i.e. the two loci need not have additive effects on genotypic means). The phenotype of an individual can be modeled as a function of both loci (o_A and o_B) as

$$(y_j|o_{Aj} = i, o_{Bj} = k) = \mu_{ik} + g_j + \boldsymbol{\beta}'(\mathbf{x}_j - \mathbf{s}) + e_j, \tag{16}$$

where μ is now a function of both loci and $k =$ BB, Bb, bb at the second locus. For two loci, the unconditional variance of y is given by

$$\text{var}(y) = \sigma^2_{oA} + \sigma^2_{oB} + \sigma^2_{oA\times oB} + \sigma^2_g + \sigma^2_e, \tag{17}$$

where $\sigma^2_{oA\times oB}$ is the variance due to the interaction (i.e. epistasis) between the two loci.

A reparameterization of the two-locus means allows direct tests of epistasis [6]. We can classify a number of mean effect models into two categories (epistatic vs. additive) which are based on how the two-locus genotypes map to phenotypes. To simplify interpretation, we define the mean genotypic vector as

$$\begin{bmatrix} \mu_{AABB} \\ \mu_{AABb} \\ \mu_{AAbb} \\ \mu_{AaBB} \\ \mu_{AaBb} \\ \mu_{Aabb} \\ \mu_{aaBB} \\ \mu_{aaBb} \\ \mu_{aabb} \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} m \\ a_A \\ a_B \\ d_A \\ d_B \\ aa_{AB} \\ ad_{AB} \\ ad_{BA} \\ dd_{AB} \end{bmatrix}, \tag{18}$$

or, in matrix notation, as

$$\boldsymbol{\mu} = \mathbf{D}\boldsymbol{\gamma},$$

where $\boldsymbol{\mu}$ is the vector of genotypic means, $\mathbf{D}$ is a design matrix, and $\boldsymbol{\gamma}$ is the vector of two-locus effects. The backtransformation is given by

$$\boldsymbol{\gamma} = \mathbf{D}^{-1}\boldsymbol{\mu}.$$

The components of $\boldsymbol{\gamma}$ have simple interpretations based on gene action. The parameter m represents the unweighted mean of the four double homozygotes and serves as a baseline component of all nine two-locus genotypes. The parameters a_A and a_B represent the additive genetic effects of the A and B loci, respectively. Similarly, d_A and d_B are defined as locus-specific dominance effects. The remaining parameters represent different forms of epistatic interactions. The aa_{AB} term is a digenic additive $\times$ additive interaction, whereas dd_{AB} represents dominance $\times$ dominance interaction. The terms ad_{AB} and ad_{BA} measure the two possible types of additive $\times$ dominance interactions.

This method of defining two-locus effects permits intuitive specification of a number of hypotheses that can be tested using the likelihood ratio. For example, the null hypothesis of no epistasis (i.e. a model in which effects are additive among loci) can be tested by comparing a model in which the four digenic interaction terms are forced to be zero, with the more general model in which the interactions are estimated.

Multivariate Complex Segregation Analysis

The basic model for the phenotype can be extended to the multivariate case [2, 5, 6, 10, 21]. Multivariate segregation analysis can be used to examine the pleiotropic (multivariate) effects of major loci and polygenes. For t phenotypes, the expected conditional covariance matrix for the ith genotype is given by

$$\begin{aligned} \text{var}(\mathbf{y}_j|o_j = i) &= \mathbf{P}_w \\ &= \mathbf{G} + \mathbf{E}, \end{aligned}$$

where $\mathbf{P}_w$ refers to the $t \times t$ within-genotype phenotypic covariance matrix, $\mathbf{G}$ is the residual additive genetic covariance matrix, and $\mathbf{E}$ is the residual environmental covariance matrix.

For the multivariate mixed model, the total phenotypic covariance matrix is

$$\begin{aligned} \text{var}(\mathbf{y}) &= \mathbf{P}_T \\ &= \mathbf{M} + \mathbf{G} + \mathbf{E}, \end{aligned}$$

where **M** is the genetic covariance matrix due to the major locus. There are general matrix formulas for calculating **M**, **G**, **E**, and other decompositions of $\mathbf{P}_T$ [5]. One fast method for multivariate segregation analysis uses the approximate mixed model of Hasstedt [27] and is based on a simplification of the multivariate likelihood via a transformation that simultaneously orthogonalizes $\mathbf{P}_w$, **G**, and **E**. The transformation makes the traits genetically (for the residual polygenic component) and environmentally uncorrelated. Multivariate conditional likelihoods then reduce to the product of t univariate conditional likelihoods. An alternative **bivariate** approach using the Class D regressive model also is available [2, 7].

Given well-known results from finite mixture distribution theory [14] showing that the asymptotic variance of the estimator of the mixing weights always decreases as the number of traits (showing evidence of the mixture) is increased, multivariate segregation analysis is likely to be more powerful than univariate segregation analysis for detecting major genes. The inclusion of multiple traits, that are influenced by the pleiotropic effects of a major locus, provides additional information on the underlying major locus and leads to increased precision in the estimation of the parameters of major locus transmission.

Discrete Phenotypes

All of the above discussion has been oriented around continuous phenotypes. However, much of the research in genetic epidemiology involves dichotomous traits such as affection status or ordered **polytomous** variables such as discrete indicators of disease severity. Such traits are usually assumed to be the outward indicators of some unknown continuous process. The assumption of an underlying continuous liability makes the analysis of these traits completely analogous to that of quantitative phenotypes (*see* **Genetic Liability Model**).

Examples

Genetics of Thyroxine in Baboons

We recently obtained preliminary evidence for a major gene influencing quantitative variation of circulating levels of the thyroid hormone, thyroxine (T_4), in the baboon. Thyroid hormones influence many aspects of physiology, yet little is known regarding the genetic basis of normal quantitative variability. Total T_4 levels were measured by **radioimmunoassay** in the frozen sera of 248 pedigreed baboons (*Papio hamadryas anubis*). After correcting for sex and age effects, we performed complex segregation analysis on these data.

A number of competing transmission hypotheses were evaluated. A series of models of varying complexity were examined, including: (i) a general model that allows arbitrary transmission probabilities (τs); (ii) a finite mixture model in which there is no transmission of the major factor; (iii) a reduced general model (the free τ_{Aa} model) in which the transmission probability of the Aa heterozygote is allowed to take a non-Mendelian value; (iv) a Mendelian recessive model; (v) a polygenic model in which there is no major factor; and (vi) a sporadic model in which there is no resemblance among relatives. Models incorporating a major factor (gene) were also allowed to exhibit a residual polygenic component. Extensive model comparisons revealed that only two component distributions were required to account adequately for observed variation in baboon T_4. Therefore, only two-distribution models are presented in which the heterozygote mean μ_{Aa} is constrained to be equal to that of the low homozygote, μ_{AA}. The best fitting, most parsimonious, model was chosen as the one that was not significantly different from the fully parameterized most general model and which also exhibited the minimum AIC (a measure of both model parsimony and fit). Table 2 shows the results of this analysis.

Table 2 shows that all models are significantly worse than the general model except for the free τ_{Aa} model and the recessive model. Additionally, the recessive model is not significantly worse fitting than the free τ_{Aa} model ($\Lambda_1 = 1.49$, $P = 0.222$) and also exhibits the minimum AIC value (AIC = 13.54). Therefore, the recessive Mendelian model represents the best-fitting, most parsimonious, model for transmission of quantitative T_4 levels. The recessive model exhibits two distributions. Individuals in the lower one comprise both AA homozygotes and Aa heterozygotes and have a mean T_4 level of 4.50 μg/dl, while aa homozygotes exhibit a mean of 6.58 μg/dl. There was no evidence for a residual polygenic effect ($h_r^2 = 0.000$), and the common within-genotype phenotypic standard deviation was estimated as 1.02.

Table 2 Segregation analysis of thyroxine in 248 baboons: maximum likelihood estimates, AICs, and Λ statistics. Parentheses identify noniterated constrained parameters; brackets denote parameters that are constrained as functions of other parameters

Parameter	General	Environmental	Free τ_{Aa}	Recessive	Polygenic	Sporadic
p_A	0.339	0.460	0.387	0.457	–	–
τ_{AA}	1.000	[0.460]	(1)	(1)	–	–
τ_{Aa}	0.659	[0.460]	0.676	(1/2)	–	–
τ_{aa}	0.153	[0.460]	(0)	(0)	–	–
$\mu_{AA} = \mu_{Aa}$	4.476	4.513	4.420	4.504	5.144	5.141
μ_{aa}	6.592	6.668	6.544	6.580	[5.144]	[5.141]
σ	0.991	0.994	1.017	1.015	1.396	1.395
h_r^2	0.000	0.000	0.000	0.000	0.204	(0)
AIC	16.00	18.81	14.06	13.54	20.16	21.13
Λ	–	8.81	2.06	3.54	14.16	17.13
df	–	3	2	2	4	5
P	–	0.032	0.357	0.316	0.007	0.004

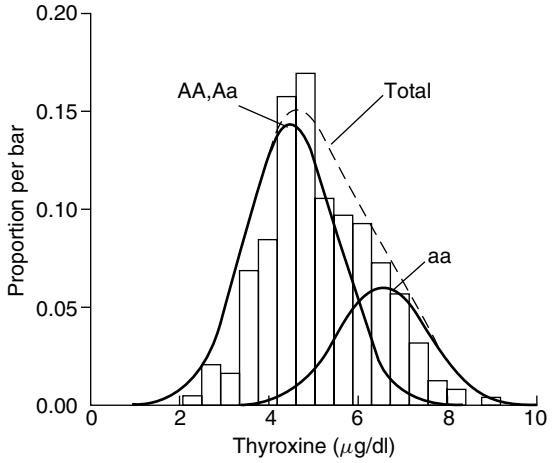

Figure 1 Expected distribution of T_4 given the best fitting major locus model. The bars of the histogram show the observed distribution

The major locus accounted for 47% of the total phenotypic variation in baboon T_4 variation. Figure 1 shows the expected distribution implied by the fitted major locus model against the observed distribution of T_4 concentration.

Analysis of IGF-I in Mexican Americans

The second example of complex segregation analysis involves the analysis of insulin-like growth factor-I (IGF-I) concentrations. IGF-I is an important regulator of cell growth/differentiation and exhibits insulin-like metabolic effects on glucose homeostasis.

Table 3 Segregation analysis of IGF-I in Mexican Americans: model comparisons. Parentheses identify noniterated constrained parameters; brackets denote parameters that are constrained as functions of other parameters

Model	Major factor transmission			Λ	df	P
General	τ_{AA}	$[\tau_{Aa}]$	τ_{aa}	–	–	–
Environmental	$[p_A]$	$[p_A]$	$[p_A]$	7.31	2	0.026
Mendelian	(1)	(1/2)	(0)	0.03	2	0.687
Polygenic	–	–	–	106.23	10	<0.001
Sporadic	–	–	–	113.36	11	<0.001

Serum levels of IGF-I decrease markedly with age and are partly regulated by nutritional factors such as protein intake. To understand better the role of genes and genotype × environment interaction in the determination of normal IGF-I variation, we measured IGF-I serum levels in 422 Mexican Americans distributed in 24 pedigrees. Information on dietary intake was obtained for each individual using a food frequency questionnaire (*see* **Nutritional Exposure Measures**). Using an extension of segregation analysis that allows for genotype × environment interaction [3], we found evidence for the effect of a major gene influencing IGF-I levels. Table 3 shows the results of the comparisons among competing transmission models. For this analysis, the equilibrium-constrained general model was used in which τ_{Aa} is given by (13). All restricted transmission models exhibited significantly worse likelihoods than the general transmission model except for the Mendelian model. The estimated frequency of the A

allele associated with lowered IGF-I levels was 0.54 ± 0.05. The A allele/factor also appeared to be dominant to the a allele/factor in all analyses.

Likelihood ratio tests of the heterogeneity in genotype-specific regression coefficients revealed evidence for genotype × sex, geno-type × age, and genotype × diet interaction. These tests are shown in Table 4. When these interactions were not considered, evidence for a major locus was attenuated.

Figure 2 shows the form of the genotype-specific regressions for two covariates – age and dietary composition. As can be seen in Figure 2(a), individuals with genotypes AA and Aa showed a less marked decline in IGF-I levels with age than that observed for the aa genotype. Similarly, as shown in Figure 2(b), IGF-I concentration exhibited a stronger positive relationship with the relative intake of dietary protein and carbohydrate in aa individuals than that observed in the other genotypes. Because of this genotype × environment interaction, the relative phenotypic variance attributable to the major gene is highly dependent upon age and diet.

Table 4 Tests of MG × E interaction: IGF-I

Interaction	Λ	df	P
Dietary composition	4.70	1	0.030
Sex	6.35	1	0.012
Age	39.93	2	<0.001

Acknowledgments

This research was supported by National Institutes of Health grants HL28972, HL45522, GM31575, and DK44297.

References

[1] Atwood, L.D., Kammerer, C.M. & Mitchell, B.D. (1995). Exploring the HDL likelihood surface, *Genetic Epidemiology* **10**, 641–645.

[2] Bagchi, P., Jiang, O. & Bonney, G.E. (1993). Compound regressive models for quantitative multivariate phenotypes: application to lipid and lipoprotein data, *Genetic Epidemiology* **12**, 647–651.

[3] Blangero, J. (1993). Statistical genetic approaches to human adaptability, *Human Biology* **65**, 941–966.

[4] Blangero, J. (1995). Genetic analysis of a common oligogenic trait with quantitative correlates: summary of GAW9 results, *Genetic Epidemiology* **12**, 689–706.

[5] Blangero, J. & Konigsberg, L.W. (1991). Multivariate segregation analysis using the mixed model, *Genetic Epidemiology* **8**, 299–316.

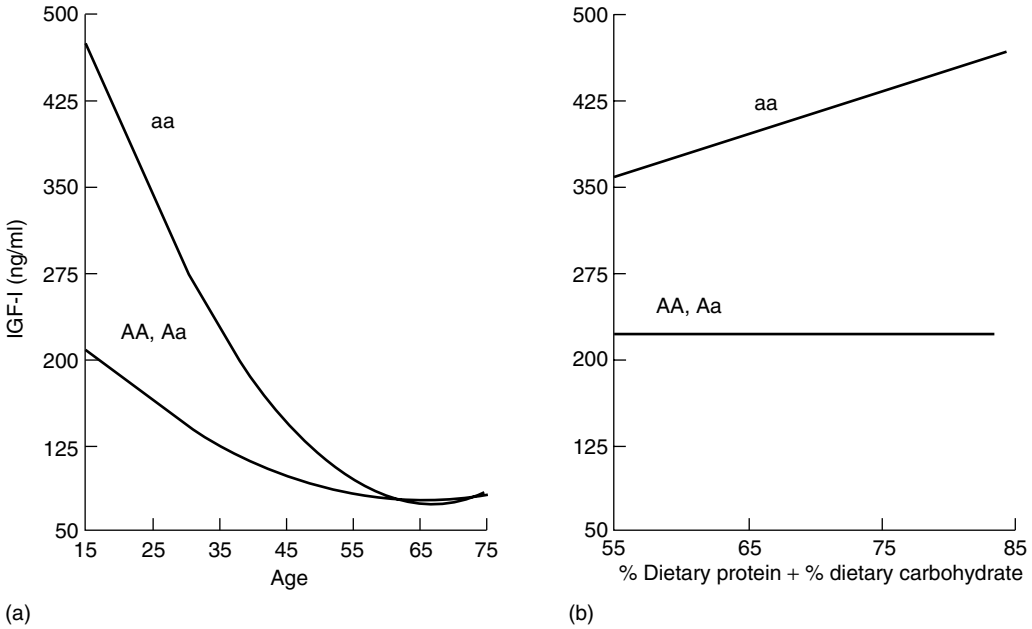

Figure 2 Genotype-specific regressions of IGF-I on covariates in Mexican American families: (a) genotype × age interaction; (b) genotype × diet composition interaction

[6] Blangero, J., MacCluer, J.W., Kammerer, C.M., Mott, G.E., Dyer, T.D. & McGill, H.C., Jr (1990). Genetic analysis of apolipoprotein A-I in two dietary environments, *American Journal of Human Genetics* **47**, 414–428.

[7] Bonney, G.E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: regressive models, *American Journal of Medical Genetics* **18**, 731–749.

[8] Bonney, G.E. (1986). Regressive logistic models for familial disease and other dependent binary traits, *Biometrics* **42**, 611–625.

[9] Bonney, G.E. (1992). Compound regressive models for family data, *Human Heredity* **42**, 28–41.

[10] Bonney, G.E., Lathrop, G.M. & Lalouel, J.M. (1988). Combined linkage and segregation analysis using regressive models, *American Journal of Human Genetics* **43**, 29–37.

[11] Borecki, I.B., Province, M.A. & Rao, D.C. (1994). Power of segregation analysis for detection of major gene effects on quantitative traits, *Genetic Epidemiology* **11**, 409–418.

[12] Borecki, I.B., Province, M.A. & Rao, D.C. (1995). Inferring a major gene for quantitative traits by using segregation analysis with tests on transmission probabilities: How often do we miss?, *American Journal of Human Genetics* **56**, 319–326.

[13] Cannings, C., Thompson, E.A. & Skolnick, M.H. (1978). Probability functions on complex pedigress, *Advances in Applied Probability* **10**, 26–61.

[14] Chang, W.C. (1976). The effects of adding a variable in dissecting a mixture of two normal populations with a common covariance matrix, *Biometrika* **63**, 676–678.

[15] Chernoff, H. (1954). On the distribution of the likelihood ratio, *Annals of Mathematical Statistics* **25**, 573–578.

[16] Demenais, F. & Elston, R.C. (1981). A general transmission probability model for pedigree data, *Human Heredity* **31**, 93–99.

[17] Demenais, F., Murigande, C. & Bonney, G. (1990). Search for faster methods of fitting the regressive models for genetic analysis. I. Continuous traits, *Genetic Epidemiology* **7**, 319–334.

[18] Eaves, L.J. (1984). The resolution of genotype × environment interaction in segregation analysis of nuclear families, *Genetic Epidemiology* **1**, 215–228.

[19] Elandt-Johnson, R.C. (1971). *Probability Models and Statistical Methods in Genetics*. Wiley, New York.

[20] Elston, R.C. (1981). Segregation analysis, *Advances in Human Genetics* **11**, 63–120.

[21] Elston, R.C. (1991). Genetic analysis of multivariate traits, *Epilepsy Research* **S4**, 161–171.

[22] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[23] Elston, R.C., George, V.T. & Severtson, F. (1992). The Elston-Stewart algorithm for continuous genotypes and environmental factors, *Human Heredity* **42**, 16–27.

[24] Fernando, R.L., Stricker, C. & Elston, R.C. (1994). The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance, *Theoretical and Applied Genetics* **88**, 573–580.

[25] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[26] Guo S.W. & Thompson, E.A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees, *Biometrics* **50**, 417–432.

[27] Hasstedt, S.J. (1982). A mixed model likelihood approximation for large pedigrees, *Computers and Biomedical Research* **15**, 295–307.

[28] Hasstedt, S.J. (1991). A variance components/major locus likelihood approximation on quantitative data, *Genetic Epidemiology* **8**, 113–125.

[29] Hasstedt, S.J. (1993). Variance components/major locus likelihood approximation for quantitative, polychotomous, and multivariate data, *Genetic Epidemiology* **10**, 145–158.

[30] Konigsberg, L.W., Blangero, J., Kammerer, C.M. & Mott, G.E. (1991). Mixed model segregation analysis of LDL-C concentration with genotype-covariate interaction, *Genetic Epidemiology* **8**, 69–80.

[31] Lalouel, J.M. (1980). Probability calculations in pedigrees under complex modes of inheritance, *Human Heredity* **30**, 320–323.

[32] Lalouel, J.M., Rao, D.C., Morton, N.E. & Elston, R.C. (1983). A unified model for complex segregation analysis, *American Journal of Human Genetics* **35**, 816–826.

[33] Lee, H. & Stram. D.O. (1996). Segregation analysis of continuous phenotypes by using higher sample moments, *American Journal of Human Genetics* **58**, 213–224.

[34] Moll, P.P., Sing, C.F., Ussier-Cacan, S. & Davignon, J. (1984). An application of a model for a genotype-dependent relationship between a concomitant (age) and a quantitative trait (LDL cholesterol) in pedigree data, *Genetic Epidemiology* **1**, 301–314.

[35] Morton, N.E. (1982). *Outline of Genetic Epidemiology*. Karger, Basel.

[36] Morton, N.E. & MacClean, C.J. (1974). Analysis of familial resemblance. III. Complex segregation analysis of quantitative traits, *American Journal of Human Genetics* **26**, 489–503.

[37] Morton, N.E., Shields, D.C. & Collins, A. (1991). Genetic epidemiology of complex phenotypes, *Annals of Human Genetics* **55**, 301–314.

[38] Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees, *American Journal of Human Genetics* **31**, 161–175.

[39] Pérusse, L., Moll, P.P. & Sing, C.F. (1991). Evidence that a single gene with gender- and age-dependent effects influences systolic blood pressure determination in a population-based sample, *American Journal of Human Genetics* **49**, 94–105.

[40] Schork, N.J. & Schork, M.A. (1989). Testing separate families of segregation hypotheses: bootstrap methods, *American Journal of Human Genetics* **45**, 803–813.

[41] Self, S.G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association* **82**, 605–610.

[42] Stricker, C., Fernando, R.L. & Elston, R.C. (1993). Segregation analysis under an alternative formulation for the mixed model, *Genetic Epidemiology* **10**, 653–658.

[43] Thomas, D.C. & Gauderman, W.J. (1996). Gibbs sampling methods in genetics, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 419–440.

[44] Thompson, E.A. & Guo, S.W. (1991). Evaluation of likelihood ratios for complex genetic models, *IMA Journal of Mathematical Applications in Medicine and Biology* **8**, 149–169.

[45] Tiret, L., Abel, L. & Rakotovao, R. (1993). Effect of ignoring genotype-environment interaction on segregation analysis of quantitative traits, *Genetic Epidemiology* **10**, 581–586.

[46] Whittemore, A.S. & Gong, G. (1994). Segregation analysis of case-control data using generalized estimating equations, *Biometrics* **50**, 1073–1087.

[47] Zhao, L.P. (1995). Segregation analysis of human pedigrees using estimating equations, *Biometrika* **81**, 197–209.

[48] Zhao, L.P. & Grove, J.S. (1995). Identifiability of segregation parameters using estimating equations, *Human Heredity* **45**, 286–300.

JOHN BLANGERO

Segregation Analysis, Mixed Models

The term "mixed model" is used often in the biostatistical literature and basically refers to statistical models that incorporate parameters that quantify a wide variety of factors thought to influence a specific outcome. For example, extensions of linear and **variance component** models that incorporate parameters that estimate the degree to which *fixed* factors, i.e. directly measured items such as gender or age (*see* **Fixed Effects**), and *random* factors, i.e. unmeasured or hypothetical items whose presence can only be dealt with probabilistically (*see* **Random Effects**), contribute to some outcome are often referred to as mixed models [18]. In parametric pedigree **segregation** and **linkage analysis** contexts the term "mixed model" refers to models that incorporate parameters that quantify the degree to which both a single locus (or few loci) with a large, individually measurable (but as yet unmeasured) effect, i.e. a "major" locus, and collective or aggregated loci with small, individually unmeasurable effects contribute to a particular phenotype [4]. Segregation analysis mixed models may also incorporate parameters that quantify the effects of measured **covariates** such as age, gender, and race, or include parameters that quantify fixed or random environmental factors shared by the pedigree members. Thus, pedigree analysis mixed models often "mix" more than two types of effect. The design of efficient, reliable, and computationally feasible mixed models for segregation and linkage analysis purposes has been notoriously difficult and presents quite a history for students of statistical genetics (*see* **Genetic Epidemiology**), especially as these models relate to the analysis of human quantitative variation. Although mixed models for categorical or discrete traits have been devised [2], they have not received as much attention as models for quantitative trait analysis. In this brief review, mixed models for quantitative traits are given exclusive attention.

Virtually all derivations of pedigree analysis mixed models for quantitative traits have as a foundation a classical single-locus, two-allele, mixture distribution-based segregation analysis model [17]. As with all parametric pedigree analysis models, this single-locus model typically involves two basic modeling components: a **penetrance** function which characterizes the probability that an individual with a certain genotype will have a certain quantitative trait value; and a transmission function. Where various mixed model formulations differ, however, is in the way in which they model and parameterize the aggregate effects of polygenes acting over and above the major locus (see, for example, [1, 3, 5, 7, 9, 10, 13, 14, 16], and [19]). In the following, a formulation of a mixed model originally attributed to Ott [14] is described. A description of this model can offer insight into the modeling and computational issues that make mixed models difficult to implement.

To introduce Ott's segregation analysis mixed model, an overview of how the modeling of the collective action of polygenes that influence the trait over and above a locus with large and measurable effect should be discussed. Consider a pedigree with N members whose trait values can be collected in a vector $\mathbf{Y} = [y_1, \ldots, y_N]$. **Multivariate normality** of the trait values among the pedigree members is assumed (see [12] for a discussion), with an N-dimensional mean vector whose elements consist of a common mean parameter $\boldsymbol{\mu} = [\mu, \ldots, \mu]$ and an $N \times N$ **covariance matrix**, Ψ, which can be partitioned, for example, in the following way:

$$\boldsymbol{\Psi} = 2\mathbf{A}\sigma_{\mathrm{a}}^2 + \mathbf{D}\sigma_{\mathrm{d}}^2 + \mathbf{H}\sigma_{\mathrm{h}}^2 + \mathbf{I}\sigma_{\mathrm{r}}^2, \qquad (1)$$

where σ_{a}^2 is a **variance component** associated with additive genetic factors, σ_{d}^2 is a variance component associated with dominance genetic factors, σ_{h}^2 is variance component associated with (unmeasured) shared household factors, and σ_{r}^2 is a variance component associated with random or individual-specific factors. The coefficient terms in (1) are $N \times N$ matrices that relate the variance components to the individuals in the pedigree: $\mathbf{A}$ is the *kinship coefficient* matrix and characterizes the degree to which related individuals share genes in expectation [11, 20]; $\mathbf{D}$ is Jacquard's Δ_7 matrix and characterizes the degree to which related individuals share both alleles at a locus [11] (*see* **Identity Coefficients**); $\mathbf{H}$ is a matrix that characterizes household sharing, whose ijth element could be either 1 or 0 depending on whether or not persons i and j share the same household; and $\mathbf{I}$ is the identity matrix. The **likelihood** of the parameters can be

written as

$$L(\boldsymbol{\mu}, \sigma_a^2, \sigma_d^2, \sigma_h^2, \sigma_r^2|\mathbf{Y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Psi}|^{1/2}} \exp\left[\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right]. \quad (2)$$

The next step in Ott's model is the addition of parameters accommodating a specific major genetic locus effect within the framework of (2). Once this component is added to the model in (2), the variance component parameters characterize "residual" variation, or variation not explained by the major locus. As with a simple monogenic segregation analysis model, Ott's mixed model considers all possible **genotype** configurations at this major locus for the pedigree members. Ott's model assumes that a locus with two alleles (three genotypes) influences a trait, and produces an extended version of (2) whose formulation for a pedigree becomes:

$$L(p_a, \mu_{AA}, \mu_{Aa}, \mu_{aa}, \sigma_a^2, \sigma_d^2, \sigma_h^2, \sigma_r^2|\mathbf{Y}) = \sum_g^{G(n)} \tau(g) \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Psi}|^{1/2}} \times \exp\left[-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu}_g)'\boldsymbol{\Psi}^{-1}(\mathbf{Y}-\boldsymbol{\mu}_g)\right], \quad (3)$$

where the sum is over all possible genotype configurations the pedigree members can have and where p_a is a major locus allele frequency (the other allele having frequency $1-p_a$) and μ_{AA}, μ_{Aa}, and μ_{aa} are mean effect parameters for each of the three major locus genotypes. $\tau(g)$ is the probability of the genotype arrangement g and is a function of the allele frequency p_a for persons without parents in the pedigree and is determined by **Mendel's Laws** for those with parents in the pedigree. Consider a nuclear family with five members, where the two parents are identified as individuals 1 and 2 and their three offspring are identified as individuals 3, 4, and 5. One possible genotype configuration for this family is: $g = [AA, Aa, AA, Aa, Aa]$ (where the genotypes are given in pedigree member number order). The probability of this configuration, given a simple allele frequency parameter, p, and Mendel's laws, is $\tau(g) = (1-p)^2 \times 2p(1-p) \times 1/2 \times 1/2 \times 1/2$. During the evaluation of the likelihood, the mean vector will change depending on the genotype configuration assessed for the family. Thus, for the genotype configuration in question, the mean vector would be $\mu_g = [\mu_{AA}, \mu_{Aa}, \mu_{AA}, \mu_{Aa}, \mu_{Aa}]$.

Evaluation of (3) for large pedigrees is extremely difficult computationally. The total number of possible genotype configurations for a family of size N is 3^N. However, not all of these genotype configurations are compatible with Mendelian theory; for example, the configuration $g = [AA, AA, aa, aa, aa]$ is incompatible with Mendelian theory, barring mutation, since it suggests that parents without a alleles transmit them to offspring. Ott [14] has shown that for a nuclear family of size N, there are $4 + 3^{N-2} + 2^N$ Mendelian compatible genotype configurations for a two-allele, three-genotype model. Although this number is considerably less than 3^N for large N, it can still be quite large. As a result, the computational burden associated with segregation analysis of quantitative traits in large pedigrees via Ott's mixed model can be heavy, and this is especially true if one is analyzing a number of pedigrees or families, since in this situation the likelihood is given by the product of the likelihoods associated with each family. In addition, the fact that the evaluation of the residual variation and covariation involves all the pedigree members does not permit the likelihood to be broken up in a way that could accommodate the **Elston–Stewart algorithm**. Schork [16], however, has described a version of Ott's model that allows the use of the Elston–Stewart algorithm, but at the sacrifice of parameters that specify polygenic covariation among distantly related individuals. Other researchers have tried to formulate the mixed model in ways that permit computational ease by eliminating terms in this manner and have met with varying degrees of success [1, 9, 10]. Attempts to incorporate residual parameter terms simply in ways that do not involve variance components as in (2) suffer from enormous computational difficulties for large families [15].

The computational complexity associated with likelihood evaluation is not the only problem plaguing mixed models. For example, one must often estimate a large number of parameters that are highly correlated. Since most mixed model likelihoods are unwieldy analytically, it is often difficult, if not impossible, to obtain analytic derivatives of the likelihood. Thus, maximization of the relevant likelihoods

to obtain parameter estimates often requires numerical methods whose reliability can be hard to gauge in pedigree analysis contexts. In addition, inclusion of covariate and other fixed or random effects information only exacerbates computational and estimation difficulties. Also, since the mixed model (as with the standard major locus segregation model for quantitative traits) often uses a mixture distribution to model the effects of the major locus genotypes, **likelihood ratio tests** of relevant hypotheses suffer a number of problems whose nature is beyond the scope of this review (see [6] and [8]). As a result of these and other problems, the derivation and implementation of reliable and easily implemented mixed models will likely continue to present challenges to statistical geneticists for some time.

References

[1] Bonney, G.E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: regressive models, *American Journal of Medical Genetics* **35**, 816–826.

[2] Bonney, G.E. (1986). Regressive logistic models for familial disease and other binary traits, *Biometrics* **42**, 611–625.

[3] Bonney, G. (1992). Compound regressive models for family data, *Human Heredity* **42**, 28–41.

[4] Elston, R.C. (1981). Segregation analysis, in *Advances in Human Genetics*, H. Harris & K. Hirshhorn, eds. Plenum, New York, pp. 63–120.

[5] Elston, R., George, V. & Severtson, F. (1992). The Elston-Stewart algorithm for continuous genotypes and environmental factors, *Human Heredity* **42**, 16–27.

[6] Ghosh, J. & Sen, P. (1985). On the asymptotic properties of the log likelihood ratio statistic for the mixture model and related issues, in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Keifer*, L. LeCam & R. Olshen, eds. Wadsworth, Monterey, pp. 789–806.

[7] Guo, S.W. & Thompson, E.A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees, *Biometrics* **50**, 417–432.

[8] Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures, in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Keifer*, L. LeCam & R. Olshen, eds. Wadsworth, Monterey, pp. 807–813.

[9] Hasstedt, S.J. (1992). A mixed-model likelihood approximation on large pedigrees, *Computers and Biomedical Research* **15**, 295–307.

[10] Hasstedt, S. (1993). Variance components/major locus likelihood approximation for quantitative, polychotomous, and multivariate data, *Genetic Epidemiology* **10**, 145–158.

[11] Jacquard, A. (1974). *The Genetic Structure of Populations*. Springer-Verlag, New York.

[12] Lange, K. (1978). Central limit theorems for pedigrees, *Journal of Mathematical Biology* **6**, 59–66.

[13] Morton, N.E. & MacLean, C.J. (1974). Analysis of family resemblance. III. Complex segregation analysis of complex traits, *American Journal of Human Genetics* **26**, 489–503.

[14] Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigree analysis. *American Journal of Human Genetics* **31**, 161–175.

[15] Schork, N. (1991). Efficient computation of patterned covariance matrix mixed models in quantitative segregation analysis, *Genetic Epidemiology* **8**, 29–46.

[16] Schork, N.J. (1992). Extended pedigree patterned covariance matrix mixed models for quantitative phenotype analysis, *Genetic Epidemiology* **9**, 73–86.

[17] Schork, N.J., Allison, D.B. & Thiel, B. (1996). Mixture distributions in human genetics research, *Statistical Methods in Medical Research* **5**, 155–178.

[18] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

[19] Stricker, C., Fernando, R.L. and Elston, R.C. (1993). Segregation analysis under an alternative formulation for the mixed model, *Genetic Epidemiology* **10**, 653–658.

[20] Thompson, E.A. (1986). *Pedigree Analysis in Human Genetics*. Johns Hopkins University Press, Baltimore.

(*See also* **Commingling Analysis**; **Linear Mixed Effects Models for Longitudinal Data**; **Segregation Analysis, Complex**)

NICHOLAS J. SCHORK

Segregation Ratios

When parents with known **genotypes** for a particular trait have children, the expected ratios of the different phenotype classes of offspring are called segregation ratios. If, for example, two parents each have genotype Aa at a particular locus, and the phenotypes corresponding to the three genotypes AA, Aa, and aa are distinct, then these three classes of offspring are produced in the segregation ratios $1:2:1$. However, if the allele A is dominant to the allele a with respect to the observed phenotype, then the two classes of offspring (AA/Aa and aa) are produced in the segregation ratio $3:1$; the same information can also be expressed by saying that for the mating Aa $\times$ Aa the segregation ratio of the recessive phenotype is 1/4.

More generally, the genotypes of the parents may not be known, and the segregation ratios are then quoted conditional on the parent's phenotypes, e.g. affected $\times$ affected, affected $\times$ unaffected, or unaffected $\times$ unaffected mating types. **Segregation analysis** is concerned with determining whether empirical segregation ratios from such mating types are consistent with simple modes of inheritance, such as one-locus dominant or recessive inheritance. Segregation ratios refer to the distribution of offspring phenotype classes, whereas **genetic transition probabilities** refer to the distribution of offspring genotype classes.

ROBERT C. ELSTON

Selection Bias

Selection bias is a **bias** that arises when individuals included in a study are not representative of the **target population** for the study. Selection bias can arise because an inappropriate **sampling frame** is used, because inappropriate sampling methods are applied (*see* **Probability Sampling**), or because some of those sampled refuse to participate in the study (*see* **Bias from Nonresponse**). In studies relying on samples of convenience, such as **hospital-based case–control studies** or **clinical trials** in which patients volunteer for particular treatments, it is difficult to rule out the possibility of selection bias (*see* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**; **Bias, Overview**; **Missing Data in Clinical Trials**; **Missing Data in Epidemiologic Studies**; **Validity and Generalizability in Epidemiologic Studies**).

MITCHELL H. GAIL

Semi-Markov Processes

Semi-Markov processes are useful generalizations of a class of **stochastic processes** commonly referred to as a Markov jump process in continuous time. One of the most famous examples of a Markov jump process in biostatistics is the illness–death process, which has been discussed extensively by Chiang [1] and subsequent editions (*see* **Fix–Neyman Process**). To illustrate the ideas, suppose one takes observations on the histories of patients who visit a clinic from time to time. As time passes, any patient may be in one of a set of states. Among these states are E_1, indicating that a patient has died, and another E_2, indicating that a patient has been lost to follow-up. From the point of view of an observer in the clinic, no further data would be available on patients in states E_1 and E_2, so in this sense the process terminates. In general, such states of termination will be called *absorbing states*.

From the point of view of health care workers in a clinic, two other states would be of compelling interest. One of these states could be labeled E_3, indicating that a patient is ill and undergoing treatment, and another state could be labeled E_4, indicating that the patient is well, or, in the case of cancer, the patient is in remission. Moreover, as time passes, after an episode in state E_3, a patient may move to state E_4; after an episode in state E_4, a patient may move back to state E_3, and so his or her sample path continues until one of the absorbing states, E_1 or E_2, is entered. In general, sets of states such as E_3 and E_4, among which a process may move prior to termination, will be called *transient*.

If an investigator were to formulate a model of the situation just described as a Markov jump process in continuous time with stationary laws of evolution, then from the sample path perspective, a very simple picture for the evolution of the process emerges. Suppose that the process begins in transient state E_3 at time $t = 0$. Then the length of the episode in this state follows an **exponential distribution** with parameter $\theta_3 > 0$, i.e. the probability that the process is still in state E_3 at time $t > 0$ is $\exp(-\theta_3 t)$. Given that an exit from state E_3 occurs, a jump to one of the states E_1, E_2, or E_4 occurs with probabilities p_{31}, p_{32}, or p_{34}, respectively. Similar statements hold for the transient state E_4.

A limitation of the exponential distribution as a model of the waiting time for the occurrence of some biological event is that it has the memoryless or nonaging property. To illustrate this property, suppose it is known that the length of an episode in state E_3 is $s > 0$ time units. Then, given this event, the past is forgotten in the sense that $\exp(-\theta_3 t)$ is the **conditional probability** that the process is still in this state at time $s + t$. Among other things, semi-Markov processes were introduced to remove the restriction that the length of an episode in a state necessarily follows an exponential distribution. As we shall see, formulating a model as a semi-Markov process not only removes this restriction but also provides useful alternatives for viewing Markov jump processes in continuous time. Some authors, such as Cinlar [2], also refer to these generalized models as Markov **renewal processes**.

An Overview of the Structure of Semi-Markov Processes

A first step in formulating a model as a semi-Markov process is to define a state space $\mathfrak{S}$, consisting of two disjoint sets of absorbing and transient states denoted by $\mathfrak{S}_1$ and $\mathfrak{S}_2$, respectively. For the sake of simplicity, attention will be confined to the case where the state space is finite. By way of another illustrative example, suppose an investigator is considering the evolution of a cohort of patients infected with HIV, the causal agent of AIDS (*see* **AIDS and HIV**). Then, in light of the effects of protease inhibitors in controlling HIV reported recently in the literature, it would seem plausible to consider a set $\mathfrak{S}_1$ of three absorbing states E_{11}, E_{12}, and E_{13}, denoting death from a cause other than AIDS, death due to an AIDS defining disease, and a case in which patients are cleared of the virus through treatment, respectively. The elements of the set $\mathfrak{S}_2$ of transient states could be symbolized by E_{2k}, $k = 1, 2, \ldots, 6$, representing classes of $CD4^+$ counts of T-lymphocytes used to define clinical stages of HIV disease.

In general, let $\mathfrak{S}_1$ denote the set of $r_1 \geq 1$ absorbing states and $\mathfrak{S}_2$ the set of $r_2 \geq 1$ transient states so that the state space $\mathfrak{S}$ has $r = r_1 + r_2$ elements. The evolution of the process is defined by two sequences of random variables. Let $X_0 = i \in \mathfrak{S}_2$ denote the initial transient state, and let the random variable X_n denote the state in $\mathfrak{S}$ entered at the nth jump

for $n \geq 1$. The evolution of the process in time is accounted for by the increasing sequence of random variables

$$0 = T_0 \leq T_1 \leq T_2 \leq \cdots \leq T_n \leq \cdots,$$

taking values in $[0, \infty)$ and denoting the random times jumps occur, where, for $n \geq 1$, the random variable T_n is the time that the nth jump occurs. If $X_n = j$, a transient state, then $T_{n+1} - T_n$ is the random length of the episode in state $j \in \mathfrak{S}_2$.

A another basic step in formulating a model as a semi-Markov process is the construction of an $r_2 \times r$ matrix,

$$(a_{ij}(t)|i \in \mathfrak{S}_2, j \in \mathfrak{S}), \tag{1}$$

of continuous one-step density functions defined for $t \in [0, \infty)$. In the next section, methods for constructing such densities will be described, but, for the time being, attention will be focused on the structure of the process. In describing this structure, functions defined by

$$A_{ij}(t) = \int_0^t a_{ij}(s)\,\mathrm{d}s \tag{2}$$

will play a basic role.

The sequence of pairs

$$\{(X_n, T_n)|n = 0, 1, 2, \ldots\}$$

will be said to have the semi-Markov property if the condition

$$\begin{aligned} &\Pr[X_{n+1} = j, T_{n+1} - T_n \leq t|(X_k, T_k), \\ &\qquad k = 0, 1, \ldots, n] \\ &= \Pr[X_{n+1} = j, T_{n+1} - T_n \leq t|X_n] \end{aligned} \tag{3}$$

holds for all $n = 0, 1, 2, \ldots$ and $t > 0$. When $X_n = i \in \mathfrak{S}_2$, the conditional probability on the right will be identified as

$$\Pr[X_{n+1} = j, T_{n+1} - T_n \leq t|X_n = i] = A_{ij}(t). \tag{4}$$

According to this formulation, the laws of evolution of the process are stationary in the sense that the right-hand side of (4) does not depend on n.

By considering marginal distributions in (4), it can be seen that the sequence of random variables $(X_n|n = 0, 1, 2, \ldots)$ is a discrete time **Markov chain** embedded in a continuous time semi-Markov process with a one-step transition matrix determined for all $n \geq 0$ by

$$\Pr[X_{n+1} = j|X_n = i] = \lim_{t \uparrow \infty} A_{ij}(t) = p_{ij}, \tag{5}$$

where $i \in \mathfrak{S}_2$ and $j \in \mathfrak{S}$. Similarly, the distribution function on the random length of time for any episode in state i is given by

$$A_i(t) = \Pr[T_{n+1} - T_n \leq t|X_n = i] = \sum_{j \in \mathfrak{S}} A_{ij}(t) \tag{6}$$

for all $n \geq 0$.

When working with a semi-Markov process, it is often possible to avoid much formidable formalism by focusing on a matrix representation of the basic functions governing the evolution of the process. To this end, one may extend the definition of the one-step density matrix in (1) to the case $i \in \mathfrak{S}_1$, the set of absorbing states. For every $i \in \mathfrak{S}_1$, let $a_{ii}(0) = 1$ and let $a_{ii}(t) = 0$ for $t > 0$, and for $i \neq j \in \mathfrak{S}$, let $a_{ij}(t) = 0$ for all $t \geq 0$ to indicate that any transition from an absorbing state occurs with probability zero. Then, for $t > 0$, the matrix of one-step transition densities may be represented in the partitioned form

$$\mathbf{a}(t) = (a_{ij}(t)) = \begin{bmatrix} \mathbf{0}_{r_1,r_1} & \mathbf{0}_{r_1,r_2} \\ \mathbf{r}_{r_2,r_1}(t) & \mathbf{q}_{r_2,r_2}(t) \end{bmatrix}, \tag{7}$$

corresponding to the sets of r_1 and r_2 absorbing and transient states, respectively. Having identified the dimensions of the submatrices in (7), from now on, to lighten the notation, subscripts on such matrices will be dropped. Briefly, the matrix $\mathbf{r}(t)$ governs one-step transitions from transient states to absorbing states, while the matrix $\mathbf{q}(t)$ governs transitions among transient states.

With the $r \times r$ matrix $\mathbf{a}(t)$ defined as in (7), and the atom at $t = 0$ for absorbing states properly accommodated in the integrals, the matrix of functions in (4) may be represented in the partitioned form

$$\mathbf{A}(t) = \int_0^t \mathbf{a}(s)\,\mathrm{d}s = \begin{bmatrix} \mathbf{I}_{r_1} & \mathbf{0} \\ \mathbf{R}(t) & \mathbf{Q}(t) \end{bmatrix} \tag{8}$$

for $t > 0$, where $\mathbf{I}_{r_1}$ is an identity matrix of order r_1. From this representation it can be seen that the transition matrix for the embedded Markov chain may be represented in the partitioned form

$$\mathbf{P} = \lim_{t \uparrow \infty} \mathbf{A}(t) = \begin{bmatrix} \mathbf{I}_{r_1} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{bmatrix}. \tag{9}$$

In their pioneering book on finite Markov chains, Kemeny & Snell [3] used this partitioned form extensively. As we shall see, numerical answers to many questions of interest may be expressed in terms of the matrices **R** and **Q**.

On Constructing Transition Densities

As will be shown by illustrative examples, several approaches may be used in constructing matrices of these densities. In one approach, based on the classical theory of **competing risks**, a point of departure could be the specification of a matrix of latent risk, **hazard**, or rate functions. If, for example, an illness–death process were being considered, then a 4×4 matrix of continuous latent risk functions could take the form

$$\mathbf{\Theta}(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \theta_{31}(t) & \theta_{32}(t) & 0 & \theta_{34}(t) \\ \theta_{41}(t) & \theta_{42}(t) & \theta_{43}(t) & 0 \end{bmatrix}, \quad (10)$$

where $t \in (0, \infty)$. Because the states E_1 and E_2 are absorbing, the latent risks governing transitions out of these states are 0. But, if a patient is in state E_3 undergoing treatment, then the function governing his or her rate of entrance into the well state, E_4, is $\theta_{34}(t)$. Analogous interpretations may be attached to the other nonzero risk functions in (10). Also, observe that all diagonal elements of $\mathbf{\Theta}(t)$ corresponding to transient states are zero, because in jump processes a state cannot make a transition into itself.

In general, let $\mathbf{\Theta}(t) = (\theta_{ij}(t))$ be an $r \times r$ matrix of latent risk functions. For any transient state $i \in \mathfrak{S}_2$, the total risk function is

$$\theta_i(t) = \sum_{j \in \mathfrak{S}_2} \theta_{ij}(t). \quad (11)$$

Thus, by appealing to well-known formulas for expressing survival functions in terms of risk functions (*see* **Survival Distributions and Their Characteristics**), it can be seen that if transient state i is entered at time $t = 0$, then

$$S_i(t) = \exp\left[-\int_0^t \theta_i(x)\,\mathrm{d}x\right] \quad (12)$$

is the conditional probability the process is still in state i at time $t > 0$. As in the second section, let $A_{ij}(t)$ be the conditional probability that a transition to state $j \neq i$ occurs sometime during the time interval $(0, t]$, $t > 0$, given that state i was entered at $t = 0$. Then, by appealing to the classical theory of competing risks, it follows that

$$A_{ij}(t) = \int_0^t S_i(x)\theta_{ij}(x)\,\mathrm{d}x = \int_0^t a_{ij}(x)\,\mathrm{d}x. \quad (13)$$

A simple and useful case arises when all nonzero latent risk functions are positive constants. For in this case, (13) takes the form

$$A_{ij}(t) = [1 - \exp(-\theta_i t)]\frac{\theta_{ij}}{\theta_i}. \quad (14)$$

Consequently, when all latent risk functions are constant, the distribution of the length of any episode in transient state i is that of a random variable following an exponential distribution with parameter $\theta_i > 0$.

Because of this property, it can be shown that a model formulated as a semi-Markov process with constant latent risks is an alternative way of viewing a Markov jump process in continuous time. Furthermore, it can be seen from (14) that the transition probabilities for the embedded Markov chain have the form

$$p_{ij} = \lim_{t \uparrow \infty} A_{ij}(t) = \frac{\theta_{ij}}{\theta_i}. \quad (15)$$

If an investigator is inclined to suspect that latent risk functions would not be constant, then among the alternative choices would be **Weibull** risk functions of the form

$$\theta_{ij}(t) = \alpha_{ij}\beta_{ij}t^{\alpha_{ij}-1}, \quad (16)$$

where $t > 0$, $\alpha_{ij} > 0$, and $\beta_{ij} > 0$. If $0 < \alpha_{ij} < 1$, then $\theta_{ij}(t)$ decreases as t increases, but if $\alpha_{ij} > 1$, then $\theta_{ij}(t)$ increases as t increases, indicating, in the latter case, that the longer the episode in state i, the greater is the risk of a transition to state $j \neq i$. As software becomes more user-friendly and as desk-top computers increase in power and speed, the computation of integrals of form (13) become increasingly feasible.

An alternative approach to the classical theory of competing risks is to start with "latent" distribution functions $G_{ij}(t)$ with a finite expectation, μ_{ij}, governing the waiting time for the transition $i \to j$ in the "absence" of other transitions. When there is

"competition" for transitions out of state i, it seems plausible that the larger the value of μ_{ij}, the smaller the conditional probability, p_{ij}, of an eventual transition from state i to j. Hence, it seems reasonable to consider transition probabilities for the embedded Markov chain of the form

$$p_{ij} = \frac{c_i}{\mu_{ij}}, \tag{17}$$

where c_i is a normalizing constant chosen so that

$$\sum_{j \in \mathfrak{S}} p_{ij} = 1. \tag{18}$$

In this formulation the function $A_{ij}(t)$ would be chosen as

$$A_{ij}(t) = p_{ij} G_{ij}(t) \tag{19}$$

so that the distribution function of the length of any episode in state i would be the mixture

$$A_i(t) = \sum_{j \in \mathfrak{S}} p_{ij} G_{ij}(t). \tag{20}$$

The construction just described would be most useful for those families of distributions with nonelementary risk functions. An example of a family of such distributions is the **gamma**, which has simple Laplace transforms that can be useful in the numerical analysis of semi-Markov models with transition densities constructed by this method.

Renewal-Type Integral Equations

Unlike models based on continuous time Markov jump processes, in which some version of the forward Kolmogorov differential equations is the focus of primary attention, renewal-type integral equations play a fundamental role in the analysis of semi-Markov processes. Trains of thought, known as *renewal type arguments*, are used repeatedly in derivations of these equations. For example, suppose, at time $t = 0$, a process starts in some transient state $i \in \mathfrak{S}_2$ and let $f_{ij}(t)$ be the density function of the waiting time for the termination of the process in some absorbing state $j \in \mathfrak{S}_1$. Either the process enters state j on the first step or there is a jump to some other transient state $k \neq j$ at some point $s \in (0, t]$, $t > 0$, with probability $a_{ik}(s)\,\mathrm{d}s$. At time s the process "renews" and $f_{kj}(t - s)$ is then the density of the waiting time to absorption in state j. By integrating and summing over all possibilities, the following renewal-type integral equation arises:

$$f_{ij}(t) = a_{ij}(t) + \sum_{k \in \mathfrak{S}_2} \int_0^t a_{ik}(s) f_{kj}(t - s)\,\mathrm{d}s. \tag{21}$$

As an aid to understanding the structure of the process, it will be helpful to cast (21) in matrix form. Let

$$\mathbf{f}(t) = (f_{ij}(t) | i \in \mathfrak{S}_2, j \in \mathfrak{S}_1) \tag{22}$$

be an $r_2 \times r_1$ matrix of absorption densities. Then, from an inspection of (7), it can be seen that (21) may be written in the compact matrix form

$$\mathbf{f}(t) = \mathbf{r}(t) + \int_0^t \mathbf{q}(s)\mathbf{f}(t - s)\,\mathrm{d}s \tag{23}$$

for $t > 0$.

Laplace transforms can be very useful tools in deducing formulas of interest to understanding semi-Markov processes. Accordingly, let

$$\hat{\mathbf{f}}(s) = \left[\hat{f}_{ij}(s) = \int_0^\infty \exp(-st),\, f_{ij}(t)\,\mathrm{d}t \right], \tag{24}$$

defined for $s \geq 0$, be the $r_2 \times r_1$ matrix of Laplace transforms of the absorption densities, and let the matrices $\hat{\mathbf{r}}(s)$ and $\hat{\mathbf{q}}(s)$ be defined similarly. Then, from (23), it follows that these matrices of Laplace transforms satisfy the following matrix linear equation:

$$\hat{\mathbf{f}}(s) = \hat{\mathbf{r}}(s) + \hat{\mathbf{q}}(s)\hat{\mathbf{f}}(s). \tag{25}$$

Given that the process starts in transient state $i \in \mathfrak{S}_2$ at time $t = 0$, let b_{ij} be the conditional probability that the process eventually terminates in absorbing state $j \in \mathfrak{S}_1$. Then,

$$b_{ij} = \int_0^\infty f_{ij}(t)\,\mathrm{d}t = \lim_{s \downarrow 0} \hat{f}_{ij}(s) \tag{26}$$

connects the elements of the matrix $\mathbf{B} = (b_{ij})$ with the Laplace transforms in (25). But, from (9), it can be seen that

$$\lim_{s \downarrow 0} \hat{\mathbf{r}}(s) = \mathbf{R} \tag{27}$$

and

$$\lim_{s \downarrow 0} \hat{\mathbf{q}}(s) = \mathbf{Q} \tag{28}$$

connect the Laplace transforms with the transition matrix of the embedded Markov chain. Therefore, by letting $s \downarrow 0$ in (25) and solving for the matrix $\mathbf{B}$, it can be seen that

$$\mathbf{B} = (\mathbf{I}_{r_2} - \mathbf{Q})^{-1}\mathbf{R}, \tag{29}$$

provided that the matrix inverse exists.

For many processes it can be shown under rather general conditions that $\mathbf{Q}^n \to \mathbf{0}$, an $r_2 \times r_2$ zero matrix as $n \uparrow \infty$. Therefore, the inverse matrix in (29) is the sum of a matrix geometric series:

$$\mathbf{M} = (m_{ij}) = \mathbf{I}_{r_2} + \mathbf{Q} + \mathbf{Q}^2 + \cdots = (\mathbf{I}_{r_2} - \mathbf{Q})^{-1}. \tag{30}$$

Moreover, m_{ij} is the conditional expectation of the number of episodes in transient state j prior to termination of the process in some absorbing state, given that the process starts in transient state i.

A given row of the matrix $\mathbf{B}$ of absorption probabilities is often of fundamental interest. For example, for the case of a model for the progression of patients with HIV disease, given that a patient is first observed in transient state i with some $CD4^+$ count, the ith row of the matrix $\mathbf{B}$ could be interpreted as the conditional probabilities that a patient eventually dies from a cause other than AIDS, an AIDS-defining disease, or is eventually cleared of the virus by treatment with drugs.

Another useful perspective for viewing the evolution of a semi-Markov process is the state of the process at time $t \in [0, \infty)$. Let the random function $Z(t)$ denote the state of the process at time t. For $i \in \mathfrak{S}_2$, a transient state, the conditional probabilities

$$\Pr[Z(t) = j | Z(0) = i] = P_{ij}(t), \tag{31}$$

which are sometimes referred to as the *current state probabilities*, are often the focus of attention. When working with models formulated on Markov jump processes in continuous time, it is these probabilities that are the desired solution to the forward Kolmogorov differential equations.

If $j \in \mathfrak{S}_1$, an absorbing state, then for $t > 0$, the equation

$$P_{ij}(t) = \int_0^t f_{ij}(s)\,\mathrm{d}s, \tag{32}$$

connecting current state probabilities with absorption densities, is valid. But, if $j \in \mathfrak{S}_2$, then it can be shown that the $r_2 \times r_2$ matrix,

$$\mathbf{P}(t) = (P_{ij}(t)), \tag{33}$$

of current state probabilities for transient states satisfies a renewal-type integral equation.

Let $S_i(t) = 1 - A_i(t)$ be the survival function for transient state i and let $\mathbf{D}(t)$ be the $r_2 \times r_2$ diagonal matrix defined by

$$\mathbf{D}(t) = (\delta_{ij} S_i(t)), \tag{34}$$

where δ_{ij} is the Kronecker delta. Then, by a renewal argument similar to that used in the derivation of (21), it can be seen that the matrix in (33) satisfies the following matrix renewal-type integral equation

$$\mathbf{P}(t) = \mathbf{D}(t) + \int_0^t \mathbf{q}(s)\mathbf{P}(t - s)\,\mathrm{d}s \tag{35}$$

for $t > 0$. If there were no absorbing states, then this integral equation would be a primary focus of attention.

Just as in (23), the passage to Laplace transforms in (35) can yield useful and interesting results. Let the random variable V_j denote the total time spent in transient state $j \in \mathfrak{S}_2$ prior to the termination of the process in some absorbing state. The expected length of any episode in state j is

$$\eta_j = \int_0^\infty S_i(t)\,\mathrm{d}t. \tag{36}$$

By passing to Laplace transforms in (35) it can be shown, after some analysis, that for any initial transient state $i \in \mathfrak{S}_2$

$$E[V_j | Z(0) = i] = \int_0^\infty P_{ij}(t)\,\mathrm{d}t = \lim_{s \downarrow 0} \hat{P}_{ij}(s) = m_{ij}\eta_j. \tag{37}$$

This result has a clear and simple interpretation. If the process starts in transient state i, then m_{ij} is the expected number of episodes in transient state j prior to the termination of the process in some absorbing state, and the expected length of each episode in state j is η_j. Therefore, (37) is valid. If, for example, state j indicates a patient is in a hospital and the cost per unit time is known, then (37) could be used to estimate the expected total cost of an illness.

The ease with which the formulas of this section could be implemented numerically would depend to

some extent on the tractability of the Laplace transforms of the transition densities $a_{ij}(t)$. If all these densities belonged to the gamma family and the construction in (33) were used, then all Laplace transforms would be tractable and all formulas presented in this section could be evaluated numerically with relative ease, provided the state space is not too large. When all latent risks are constant, it is also possible to use the exponential matrix as a solution of the Kolmogorov differential equations, which will require the construction an $r \times r$ rate matrix $\mathbf{Q}$.

To construct this matrix, let $\mathbf{\Theta}$ be a constant matrix of latent risks and let θ_i be the sum of the elements of the ith row of $\mathbf{\Theta}$. Then, the matrix $\mathbf{Q}$ has the form

$$\mathbf{Q} = \mathbf{\Theta} - \text{diag}(\theta_1, \theta_2, \ldots, \theta_r). \tag{38}$$

As is well known, for $t > 0$ the solution of the Kolmogorov differential equations is the exponential matrix

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) \tag{39}$$

Many computer packages have software designed to evaluate an exponential matrix either symbolically or numerically, (*see* **Matrix Computations**).

Further Reading

Whenever a model is formulated as a semi-Markov process, it is usually difficult to use **maximum likelihood** to estimate unknown parameters from data, because the **likelihood** function is difficult to derive and compute. Thompson et al. [6] have proposed an interesting method for estimating parameters on the basis of **simulating** realizations of the process and minimizing a **goodness-of-fit** criterion such as a **chi-square test** statistic. Such well-known computer software packages as MATLAB have built-in programs for computing the exponential matrix, Packages with a capability for doing **computer algebra**, such as MAPLE, may also be used to produce symbolic forms of the exponential matrix. Methods for solving discrete time versions of renewal-type integral equations numerically have been discussed and used extensively in Mode [4]. Tan [5] has presented an extensive array of applications and references to Markov processes with time inhomogeneous laws of evolution in cancer research. Many of these models could also be viewed within a semi-Markov framework with time inhomogeneous laws of evolution. See [4] for applications of related processes in **demography**.

References

[1] Chiang, C.L. (1964). *Introduction to Stochastic Processes in Biostatistics*. Wiley, New York.

[2] Cinlar, E. (1975). *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs.

[3] Kemeny, J.G. & Snell, J.L. (1976). *Finite Markov Chains*. Springer-Verlag, New York.

[4] Mode, C.J. (1985). *Stochastic Processes in Demography and Their Computer Implementation*, *Springer Biomathematics Series*, Vol. 14. Springer-Verlag, Berlin.

[5] Tan, W.Y. (1991). *Stochastic Models of Carcinogenesis*. Marcel Dekker, New York.

[6] Thompson, J.R., Stivers, D.N. & Ensor, K.B. (1991). SIMEST – technique of model aggregation with considerations of chaos in *Mathematical Population Dynamics*, O. Arino, D.E. Axelrod & M. Kimmel, eds. *Marcel Dekker Lecture Notes in Pure and Applied Mathematics*, Vol. 131. Marcel Dekker, New York, pp. 483–510.

(*See also* **Markov Processes**)

CHARLES J. MODE

Semiparametric Regression

Semiparametric regression models are a compromise between parametric and nonparametric models. The idea is to retain a certain amount of the **parsimony** and structure of a parametric model while gaining some of the flexibility of a nonparametric model.

There is no widely accepted rigorous definition of a semiparametric model. Informally we will call a model semiparametric if it is not fully parametric but has a finite dimensional parameter of interest. The most widely used semiparametric regression model is the proportional hazards (*see* **Proportional Hazards, Overview**) model of Cox [6] (*see* **Cox Regression Model**), in which the conditional hazard at time t, given **explanatory variable z**, is $\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z})$, where λ_0 is an unknown function of t (the nonparametric component) and $\boldsymbol{\beta}$ is an unknown (finite dimensional) vector (the parametric component). It is often helpful to think of λ_0 as an infinite-dimensional parameter rather than as a nonparametric component.

This informal definition of a semiparametric model excludes structured **nonparametric regression** models such as the **generalized additive model**:

$$Y_i = \alpha_0 + \alpha_1(X_{1i}) + \cdots + \alpha_p(X_{pi}) + \varepsilon_i,$$

in which $\alpha_1, \ldots, \alpha_p$ are mean zero smooth functions of the variables $X_1, \ldots, X_p$; ε_i is a mean zero "error term" and $\alpha_0 = \mathrm{E}(Y_i)$ for all i. Although α_0 is a one-dimensional parameter, we assume that it is not of primary interest in the model.

Here we describe a large class of semiparametric regression models and identify some special cases of biostatistical interest. We also discuss the sort of problems posed by semiparametric regression. Details of specific models are discussed elsewhere.

Consider the generic model

$$\psi(Y) = r(\mathbf{X}) + \varepsilon, \tag{1}$$

in which ε is a random element with distribution F. For a full specification we would have to consider the distribution of $\mathbf{X}$, but we prefer to view (1) as a model for Y given X. The model in (1) has three components: the transformation function ψ, the regression function r, and the distribution function F. For the regression model to be termed semiparametric, r must have a finite dimensional component of interest, and at least one of ψ, r, and F must have an infinite-dimensional component.

Example 1. Semiparametric Regression Functions

1. Partly parametric additive model:

$$\psi(Y) = Y, \varepsilon \sim \mathrm{N}(0, \sigma^2), r(\mathbf{X}) = \boldsymbol{\beta}'\mathbf{X}_1 + s(\mathbf{X}_2).$$

2. **Projection pursuit** regression. As in (1) but with $r(\mathbf{X}) = s(\boldsymbol{\beta}'\mathbf{X})$, where $s : \mathcal{R} \to \mathcal{R}$ is an unknown function. Only the direction of $\boldsymbol{\beta}$ is identifiable.

Example 2. Transformation Model

A general transformation model has an arbitrary nondecreasing ψ and $r(\mathbf{X}) = -\boldsymbol{\beta}'\mathbf{X}$. The error distribution F may be parametric or nonparametric. When the distribution of ε is Gaussian, one has a semiparametric extension of the Box–Cox model [3] (*see* **Power Transformations**). The model leads naturally to **rank regression**. Estimation of $\boldsymbol{\beta}$ with **censored data** has been studied by Cheng et al. [4]. Horowitz [8] considers estimation of ψ and F.

1. The Cox model. The Cox model is a special case in which $F(t) = 1 - \exp(-e^t)$, the **extreme value** (minimum) distribution, $\psi(Y) = \log \Lambda_0(Y)$, and $r(\mathbf{X}) = -\boldsymbol{\beta}'\mathbf{X}$. Here Λ_0 is a cumulative hazard function; in particular, it is nonnegative and nondecreasing.
2. The Clayton–Cuzick model. A multivariate generalization of the Cox model proposed by Clayton & Cuzick [5] has received much attention. The model assumes that individuals within the same "cluster" share a common **frailty**. The frailty may be regarded as an unobserved covariate. The hazard for the ith individual is

$$Z_i\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{X}_i)$$

where the frailty Z_i is common to all individuals in the cluster, and is assumed to have a **gamma distribution** with mean 1.

Example 3. Accelerated Failure-Time Models

Accelerated failure-time models are simply linear models for the logarithms of survival times. Such a

model will usually permit censored survival times. If the error distribution F is nonparametric, then the model is semiparametric. One has

$$\log Y = \boldsymbol{\beta}'\mathbf{X} + \varepsilon,$$

where the ε has distribution function $F \in \mathcal{F}$, the space of all distribution functions on the real line.

Schick [11] considers efficient estimation in the model in (1) when ψ is the identity function. His approach takes into account the unknown distribution of the covariates $\mathbf{X}$. Not all semiparametric regression models can be described by (1).

Example 4. Partly Parametric Aalen model

McKeague & Sasieni [10] added a parametric component to the model introduced by Aalen [1] (*see* **Aalen's Additive Regression Model**). The model for the hazard function conditional on vectors $\mathbf{x}$ and $\mathbf{z}$ is

$$\lambda(t|\mathbf{x},\mathbf{z}) = \alpha_0(t) + \alpha_1(t)x_1 + \cdots + \alpha_p(t)x_p + \boldsymbol{\beta}'\mathbf{z}.$$

It has $p+1$ infinite-dimensional components $\alpha_0, \alpha_1, \ldots, \alpha_p$ and a finite-dimensional parameter $\boldsymbol{\beta}$.

Example 5. Cox Model with Time-Dependent Coefficients

Cox Model with Time-Dependent Coefficients

A closely related model is a generalization of the Cox model which permits the relative hazards to change over time:

$$\lambda(t|x,z) = \exp[\alpha_0(t) + \alpha_1(t)x_1 + \cdots + \alpha_p(t)x_p + \boldsymbol{\beta}'\mathbf{z}].$$

Here the baseline hazard function is given by $\exp[\alpha_0(t)]$. Whereas no explicit smoothing is required to obtain $n^{1/2}$-consistent estimators of $\boldsymbol{\beta}$ in the partly parametric Aalen model, all estimators for this model require smoothing [13, 7].

Example 6. Conditionally Parametric Models

Severini & Wong [12] studied estimation of the parametric component of a semiparametric model via the **profile likelihood**. The method is particularly useful in conditionally parametric models, in which, conditional on an explanatory variable X, the model is parametric, but the dependence on X is nonparametric. For instance, given $X = x$, the parameters are θ and η_x, but $\eta_x = s(x)$ is a smooth function of x. The model is closely related to the varying coefficient models of [7].

Example 7. Semiparametric Regression Functionals

LeBlanc & Crowley [9] considered semiparametric models in which a certain linear functional T of the conditional distribution function F_X of Y given X is assumed to be linear in X:

$$T[F_X(y)] = \mathbf{X}'\boldsymbol{\beta}.$$

It is further assumed that $F_{X=x}$ is a smooth function of x. Applications include **quantile regression**, for which one would solve the estimating equation

$$\int \psi(y, \mathbf{x}'\boldsymbol{\beta})\, \mathrm{d}\hat{F}_n(y) = 0,$$

where $\hat{F}_n(y)$ is an estimate of the conditional distribution function and $\psi(y,\eta) = -q/(1-q)$ if $y \geq \eta$ and 1 otherwise.

The book by Bickel et al. [2] on semiparametric models concentrates on asymptotic bounds for estimation. It discusses how well, in theory, one can estimate the parameters (both finite- and infinite-dimensional) in a given model. Some consideration is also given to construction of estimators, but different methods seem best for different problems. Of course, all the problems that statisticians have studied on parametric models may be posed for semiparametric models. Most researchers have concentrated on the problem of efficient estimation (*see* **Efficiency and Efficient Estimators**), but more interest is now being given to robust estimation (*see* **Robustness**). Other topics that will doubtless be studied in greater depth include: **Hypothesis testing** – in particular, testing between semiparametric models and parametric submodels (*see* **Model, Choice of**); **bootstrap** and **jackknife methods**, and whether these can be used to perform tests and provide confidence bands; **goodness of fit**, **model checking** and **diagnostics**; computing and finite-sample consideration; **Bayesian methods**. Additionally, several authors have begun to relax the independent and identically distributed assumption common to most papers on semiparametric models.

References

[1] Aalen, O.O. (1980). A model for nonparametric regression analysis of counting processes, in *Lecture Notes in Statistics*, Vol. 2. Springer-Verlag, New York, pp. 1–25.

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[3] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B* **26**, 211–243.

[4] Cheng, S.C., Wei, L.J. & Ying, Z. (1995). Analysis of transformation models with censored data, *Biometrika* **82**, 835–845.

[5] Clayton, D. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion), *Journal of the Royal Statistical Society, Series A* **148**, 82–117.

[6] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[7] Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models (with discussion), *Journal of the Royal Statistical Society, Series B* **55**, 757–796.

[8] Horowitz, J.L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, *Econometrica* **64**, 103–137.

[9] LeBlanc, M. & Crowley, J. (1995). Semiparametric regression functionals, *Journal of the American Statistical Association* **90**, 95–105.

[10] McKeague, I. & Sasieni, P. (1994). A partly parametric additive risk model, *Biometrika* **81**, 501–514.

[11] Schick, A. (1993). On efficient estimation in regression models, *Annals of Statistics* **21**, 1486–1521.

[12] Severini, T.A. & Wong, W.H. (1992). Profile likelihood and conditionally parametric models, *Annals of Statistics* **20**, 1768–1802.

[13] Zucker, D.M. & Karr, A.F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach, *Annals of Statistics* **18**, 329–353.

PETER SASIENI

Sensitivity Analysis

Sensitivity analysis is a reassessment of the model used for data summary that attempts to detect whether changing any of the assumptions used in a model to derive the analysis leads to different interpretations of the outcome.

Sensitivity analysis is commonly used by taking estimates of parameters from models and varying them over possible values to detect whether this variation leads to different interpretation of the response. For example, in a cost–effectiveness analysis (*see* **Health Economics**), the discount rate for the cost of future events may be set at the conventional level of 5%. (The cost of a future event in present-day currency is cheaper than a current unit because the delay of k years could generate compound interest at the rate r, so the value of a unit compounded for k years would be $(1 + r)^k$. Hence, the cost of a future event is said to be discounted at $r\%$ in terms of present currency values. The usual or conventional discount rate is $r = 5\%$.) The sensitivity analysis might be varied over the levels, 0%, 5%, and 10% to decide whether changing the discount rate over these three levels had any impact of final conclusions drawn from the study. If the final conclusion was insensitive to varying of the factor over its multiple levels, the conclusion is said to be insensitive to the choice of discount rate. A sensitivity analysis that varies one factor at a time is called a "one-way" sensitivity analysis, while a sensitivity analysis that simultaneously varies two factors in the model is called a "two-way" sensitivity analysis; these can be extended to "n-way" sensitivity analysis, where n is a positive integer.

A sensitivity analysis can be extended to a **Monte Carlo** sensitivity analysis providing the analyst can get computer-generated samples with random number **algorithms** to select amongst the various factor level combinations (cases) to be used in the sensitivity analysis. A Monte Carlo sensitivity analysis generally takes into account all of the factors that are being varied in the sensitivity analysis. A distribution of results or a set of summary statistics computed from the distribution are then used to convey the results.

Readers who are familiar with **factorial experiments** will recognize that information on multiple factors as well as changes in the structure of the model can be detected by using factorial experimentation rather than "one factor at a time" studies. Indeed, factorial designs have the possibility of detecting whether **interactions** between factors play a role and can detect smaller effects by increasing precision using the hidden replication property of factorial design. (Hidden replication is the term used in the experimental design literature to characterize the efficiency of a factorial design to estimate the effects of multiple factors with a sample size used for a single factor.) However, in the health sciences literature, one does not usually find factorial experiments conducted as sensitivity analyses; these sensitivity analyses tend to deal with one factor at a time.

An explicit definition of sensitivity analysis in equation form is difficult to find in the literature. However, there are many applications of these principles in the health economics literature. For example, a two-way sensitivity analysis showed that the use of the drug Misoprostol to prevent gastric bleeding in rheumatoid arthritis patients who were taking nonsteroidal anti-inflammatory drugs was an economically sound decision, provided the background ulcer complication rate was at least 1.5%. This conclusion was derived from a sensitivity analysis. The authors also show the results of a Monte Carlo sensitivity analysis, without changing the conclusions of the study [4].

Some authors have attempted to provide multiple factor approaches to using the principles of **experimental design** in suggesting how a sensitivity analysis can be helpful to provide conclusions about studies where judgments have to be made about more than one factor. One reference has drawn together examples that use the **general linear model** from a factorial experiment to suggest how this may be used to answer multiple sensitivity questions on the same set of data. They include applications to pneumonoccol vaccination, neonatal intensive care, and prevention of pulmonary emboli [6].

If one examines the references in the health sciences literature, there are many examples of sensitivity analysis applications to clinical and biological problems. In *Current Index to Statistics* [5], there are 13 entries suggesting applications of sensitivity analysis to a variety of problems. These include applications to growth models (*see* **Bacterial Growth, Division, and Mutation**), stochastic flow networks, (*see* **Stochastic Processes**) and **structural equation models**.

Sensitivity analyses tend *not* to challenge: (i) the form of the models; (ii) the underlying distribution of the errors; or (iii) the link between the model and the error distribution that one conventionally finds in **generalized linear models**. Hence, these may be a fruitful areas of research to determine whether improvements can be made in the application of sensitivity analysis to a variety of health care problems by exploiting some of the recent innovations in generalized linear modeling of data [8].

Sensitivity analyses have been used more recently in the field of **meta-analysis**. Meta-analysis is a technique for combining outcomes from multiple studies simultaneously with the hope of increased precision and increased **power** to detect important clinical effects [2, 3]. Sensitivity analysis has been used by systematically dropping studies, performing subgroup analyses (*see* **Treatment-covariate Interaction**) and generally using all the conventional statistical techniques to understand the **robustness** of the analyses [2, 3].

Petitti [9] discusses applications of sensitivity analyses to gallstone surgery and isoniazid prophylaxis in HIV patients, choices of factors for sensitivity analyses, and how sensitivity analysis can be used in meta-analysis.

Bailar & Mosteller [1] also define sensitivity analysis and show various **graphical displays** that can help in the interpretation of the findings from multiway sensitivity analyses, including the use of estrogens to prevent osteoporosis in postmenopausal women.

Sensitivity analysis can be described as a measure and then applied to different **estimation** methods of growth rates of algae, heart catheterization data, and the atomic weight of iodine [10].

Sensitivity analysis can help to understand whether living near a nuclear waste facility relates to childhood leukemia incidence (*see* **Leukemia Clusters**). Here the authors varied the cluster test method, reference rates, time, and age of children [12].

Smith et al. [11] used sensitivity analysis to help describe the similarities and differences between using **fixed effects** and **random effects** models in a meta-analysis.

Hunter [7] applied the principles of sensitivity analysis to product design by employing the principles of experimental design. These principles could be applied to diagnostic test development (*see* **Diagnostic Tests, Evaluation of**), drug and therapy evaluation.

References

[1] Bailar, J.C., III & Mosteller, F., eds (1992). *Medical Uses of Statistics*, 2nd Ed. NEJM Books, Boston, pp. 166–170.

[2] Cooper, H. & Hedges, L.V., eds (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York, p. 540.

[3] Eddy, D.M., Hasselblad, V. & Shachter, R. (1992). *Meta-Analysis by the Confidence Profile Method – The Statistical Synthesis of Evidence*. Academic Press, San Diego, pp. 309–310, 365.

[4] Gabriel, S.E., Campion, M.E. & O'Fallon, M. (1994). A cost-utility analysis of Misoprostol prophylaxis for rheumatoid arthritis patients receiving nonsteroidal antiinflammatory drugs, *Arthritis and Rheumatism* **37**, 333–341.

[5] Gbur, E.E., Jr, ed. (1994). *Current Index to Statistics: Applications, Methods and Theory*, Vol. 19, 1993 articles. American Statistical Association, Institute for Mathematical Statistics, Alexandria, p. 647.

[6] Goldsmith, C.H., Gafni, A., Drummond, M.F., Torrance, G.W. & Stoddart, G.L. (1986). Sensitivity Analysis and Experimental Design: The Case of Economic Evaluation of Health Care Programmes, *Proceedings of the Third Canadian Conference on Health Economics*, pp. 129–148.

[7] Hunter, J.S. (1985). Statistical design applied to product design, *Journal of Quality Technology* **17**, 210–221.

[8] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, New York.

[9] Petitti, D.B. (1994). *Meta-Analysis Decision Analysis and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. Oxford University Press, Oxford.

[10] Severini, T.A. (1986). Measures of the sensitivity of regression estimates to the choice of estimator, *Journal of the American Statistical Association* **91**, 1651–1658.

[11] Smith, T.C., Spiegelhalter, D.J. & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study, *Statistics in Medicine* **14**, 2685–2699.

[12] Viel, J.-F. & Pobel, D. (1995). Incidence of leukaemia in young people around the La Hague nuclear waste reprocessing plant: a sensitivity analysis, *Statistics in Medicine* **14**, 2459–2472.

(*See also* **Model Checking**; **Model, Choice of**; **Simulation**)

CHARLES H. GOLDSMITH

Sensitivity

In the context of diagnostic testing or disease **screening**, sensitivity refers to the proportion of individuals with the target disease who have a positive test result. In other words, it is the probability that an actual case of disease will be correctly diagnosed by the test. In probability terms, sensitivity is Pr(positive test|disease). Consider Table 1 for the general relationship between the test results and the true disease state. Then, the sensitivity is given by $a/(a+c)$. A synonym is the *true positive rate*, invoking the proportion of positive test results among the denominator of true disease cases.

Achievement of high sensitivity is important when case detection is important, specifically where the implied costs of missing disease cases (i.e. giving **false negative** test results to cases) are high relative to the costs of incorrectly assigning positive test results to individuals without the disease (the so-called **false positive** results). Typically, if the test is designed for high sensitivity, the false positive rate $b/(b+d)$ will also be high and the test **specificity** $d/(b+d)$ will be low.

A test with high sensitivity is useful clinically for the purpose of ruling out possible disease; a negative result from such a test implies a relatively high chance of not having the disease.

Therapeutic decisions are often considered using a **likelihood ratio** calculation. The likelihood ratio, LR, here is

$$LR = \frac{\text{Pr(positive test|disease)}}{\text{Pr(positive test|no disease)}}$$

or, equivalently the ratio of sensitivity to (1 − specificity). The positive and negative **predictive values** are also relevant.

In simple formulations, sensitivity and specificity are often assumed to be independent of the **prevalence** of disease in the population. In practice, these test characteristics may actually depend on prevalence, for a variety of reasons. For instance, if testing is carried out in a population in which the prevalence is higher because of a greater proportion of mild disease, one would expect sensitivity to be lower. Such effects may occur artifactually on occasion; for instance, because of different clinical definitions of disease operative in various populations.

Table 1

	Disease present	Disease absent
Test positive	a	b
Test negative	c	d

A second meaning of sensitivity is used in the context of describing the measurement properties of a device. The sensitivity here refers to the smallest stimulus that the device can detect, or the smallest input required so that the device can provide an appropriate output. This idea is commonly used to describe electronic components, but is also used to characterize the lowest concentration or amount of a substance that is detectable by the device, such as in clinical chemistry testing. A similar interpretation pertains to the sensitivity of an individual, organism or biological system, depending on the context.

In the context of data analysis, sensitivity may refer to the degree of dependence of the results to assumptions invoked by the particular techniques employed (e.g. an assumption of **normality** in a **regression** calculation), or to features in the data (e.g. the presence and position of an outlier observation). Lack of sensitivity to such characteristics is known as **robustness**. For further details of methods to examine sensitivity (in this sense), see **sensitivity analysis**.

Further Reading

Altman, D.G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall, London, Chapter 14.

Sackett, D.L., Haynes, R.B., Guyatt, G.H. & Tugwell, P. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*, 2nd Ed. Little, Brown, & Company Boston, Chapter 4.

(*See also* **Clinical Epidemiology**; **Diagnostic Tests, Evaluation of**; **Diagnostic Tests, Multiple**; **Gold Standard Test**; **Receiver Operating Characteristic (ROC) Curves**)

STEPHEN D. WALTER

Separate Families of Hypotheses

A fundamental problem in statistical analysis is that of the choice between alternate statistical models. In this context, the following questions may arise.

1. Is there any evidence that different models give significantly different fits to the data?
2. If it is assumed that one model is true, what is the evidence provided by the data as to which is the true one? (This question is often the basis of a **Bayesian** formulation of model choice.)
3. If one model represents the currently maintained hypothesis, is there any evidence of a departure from it in the direction of another model?

To compare models, the Neyman–Pearson theory of **hypothesis testing** may be used if the models belong to the same family of distributions and the relevant comparisons involve **hierarchical** (or nested) models. However, special procedures are needed if the models belong to families that are separate, in the sense that an arbitrary member of one family cannot be obtained as a limit of members of the other.

A considerable amount of research on separate families of hypothesis has been done since the fundamental work of Cox [7, 8], who first dealt with the problem. For previous reviews and references, see [11, 15–18, 21, 24].

This article first presents the results of Cox and some alternatives. Then the Bayesian approach is introduced. Finally, some references to applications of this work are given.

Cox Procedure and Alternatives

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be independent observations from some unknown distribution. Suppose that there are **null** and **alternative hypotheses** H_f and H_g specifying parametric densities $f(\mathbf{y}, \boldsymbol{\alpha})$ and $g(\mathbf{y}, \boldsymbol{\beta})$ for the random vector $\mathbf{y}$. Hence $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown vector parameters and it is assumed that the families are separate in the sense defined above. Formal definitions of separate or nonnested hypotheses are given in [10] and [27].

The asymptotic tests (*see* **Large-sample Theory**) developed by Cox [7, 8] were based on a modification of the Neyman–Pearson maximum **likelihood ratio**. If H_f is the null hypothesis and H_g the alternative hypothesis, the test statistic considered was

$$T_{fg} = lr_{fg}\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) - \mathrm{E}_{\hat{\boldsymbol{\alpha}}}\{lr_{fg}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\alpha}})\},$$

where for a random sample of size n, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively, $lr_{fg}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = l_f(\boldsymbol{\alpha}) - l_g(\boldsymbol{\beta})$ is the log-likelihood ratio, $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$ is the probability limit, as $n \to \infty$, of $\hat{\boldsymbol{\beta}}$ under H_f, and the subscript $\boldsymbol{\alpha}$ means that expectations and so on are calculated under H_f.

Cox showed that, asymptotically, under the alternative hypothesis T_{fg} has a negative mean and that under the null hypothesis T_{fg} is normally distributed with mean zero and variance

$$V_{\boldsymbol{\alpha}}(T_{fg}) = V_{\boldsymbol{\alpha}}\{lr_{fg}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\alpha}})\} - \mathbf{C}'_{\boldsymbol{\alpha}}\mathbf{I}^{-1}\mathbf{C}_{\boldsymbol{\alpha}},$$

where $\mathbf{C}_{\boldsymbol{\alpha}} = \partial \mathrm{E}_{\boldsymbol{\alpha}}\{lr_{fg}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\alpha}})\}/\partial\boldsymbol{\alpha}$, and $\mathbf{I}_{\boldsymbol{\alpha}}$ is the **information matrix** of $\boldsymbol{\alpha}$.

When H_g is the null hypothesis and H_f is the alternative hypothesis, analogous results are obtained for a statistic T_{gf}. Therefore $T^*_{fg} = T_{fg}\{V_{\boldsymbol{\alpha}}(T_{fg})\}^{-1/2}$ and $T^*_{gf} = T_{gf}\{V_{\boldsymbol{\beta}}(T_{gf})\}^{-1/2}$, under H_f and H_g respectively, can be considered as approximately standard normal variates, and two-tailed tests can be performed. For example, if T^*_{fg} is significantly negative, there is evidence of a departure from H_f in the direction of H_g. If T^*_{fg} is significantly positive, there is evidence of a departure from H_f in the opposite direction to H_g. The possible outcomes when both tests are undertaken are shown in Table 1. The decision-related terms of accept and reject are used for simplicity. Rejection of both hypotheses suggests that it is necessary to look elsewhere for an appropriate model. Acceptance of both implies that there is no evidence with which to choose between the models. Possible acceptance suggests that further testing is required, since while one model is not rejected, the other is rejected in favor of alternatives in a direction opposite to that of the model which is not rejected.

As an illustration suppose that H_f specifies that the distribution is **lognormal** and H_g specifies that it is **Weibull**; that is,

$$\mathrm{H}_f : \frac{1}{y_i(2\pi\alpha_2)^{1/2}} \exp\left\{\frac{-(\log y_i - \alpha_1)^2}{2\alpha_2}\right\}$$

Table 1 Possible outcomes from a pair of separate family significance tests

		T_{fg}		
		Significantly negative	Not significant	Significantly positive
T_{gf}	Significantly negative	Reject both	Accept H_f	Reject both
	Not significant	Accept H_g	Accept both	Possible acceptance of H_g
	Significantly positive	Reject both	Possible acceptance of H_f	Reject both

and

$$H_g : \left(\frac{\beta_2}{y_i}\right)\left(\frac{y_i}{\beta_1}\right)^{\beta_2} \exp\left\{-\left(\frac{y_i}{\beta_1}\right)^{\beta_2}\right\}.$$

We then have [23]

$$T_{fg} = n\{\hat{\beta_2}\log\hat{\beta_1} - \beta_{2\hat{\alpha}}\log\beta_{1\hat{\alpha}} - \log\hat{\beta_2} + \log\beta_{2\hat{\alpha}} - \hat{\alpha_1}(\hat{\beta_2} - \beta_{2\hat{\alpha}})\}$$

and

$$V_\alpha\{T_{fg}\} = 0.2183\ n,$$

where

$$\beta_{1\hat{\alpha}} = \exp\left\{\hat{\alpha_1} + \frac{1}{2}(\alpha_2)^{1/2}\right\} \text{ and } \beta_{2\hat{\alpha}} = \frac{1}{(\hat{\alpha_2})^{1/2}}.$$

Also,

$$T_{gf} = n\left\{\hat{\beta_2}(\hat{\alpha_1} - \alpha_{1\hat{\beta}}) + \frac{1}{2}\log\frac{\hat{\alpha_2}}{\alpha_{2\hat{\beta}}}\right\}$$

and

$$V_\beta\{T_{gf}\} = 0.2834\ n,$$

where

$$\alpha_{1\hat{\beta}} = \frac{-0.5772}{\hat{\beta_2}} + \log\hat{\beta_1} \text{ and } \alpha_{2\hat{\beta}} = \frac{1.6449}{\hat{\beta}_2^2}.$$

As an alternative to this approach, Cox [7] suggested the combination of the two models in a general model of which they would be both special cases. The density could be taken to be proportional to the exponential mixture [3]

$$\{f(\mathbf{y}, \boldsymbol{\alpha})\}^\lambda\{g(\mathbf{y}, \boldsymbol{\beta})\}^{1-\lambda},$$

or a linear mixture distribution [28]

$$\lambda f(\mathbf{y}, \boldsymbol{\alpha}) + (1-\lambda)g(\mathbf{y}, \boldsymbol{\beta}),$$

and inferences concerning λ are possible. These mixtures can be generalized for testing more than two models. The exponential mixture is the base of much of the econometric work. Cox also outlined a general formulation from the point of view of Bayesian decision theory.

Likelihood inference was used by Lindsey [13, 14] to compare models. His approach is based on the relative likelihoods

$$R_f(\tilde{P}_f) = \prod_j \left(\frac{\tilde{P}_{fj}}{\hat{P}_j}\right)^{n_j}$$

and

$$R_g(\tilde{P}_g) = \prod_j \left(\frac{\tilde{P}_{gj}}{\hat{P}_j}\right)^{n_j},$$

where $\tilde{P}_{fj} = \int_\delta^\gamma f(y_j,\ \hat{\boldsymbol{\alpha}})\ \mathrm{d}y_j$, $\tilde{P}_{gj} = \int_\delta^\gamma g(y_j,\ \hat{\boldsymbol{\beta}})\ \mathrm{d}y_j$, and $\hat{P}_j = n_j \big/ \sum_i n_i$, in which $\hat{P}_j$ is the **maximum likelihood** estimator obtained from a **multinomial** with endpoints of the jth interval defined by $\delta = y_j - 1/2\Delta_j$ and $\gamma = y_j + 1/2\Delta y_j$ and where Δy_j is the width of the jth interval. The higher the relative likelihood, the higher is the plausibility of that model.

An **information** criterion was used by Sawyer [29], who proposed the alternative to Cox statistics,

$$S_{fg}(\hat{\boldsymbol{\alpha}}) = \mathrm{E}_{\hat{\boldsymbol{\beta}}}\{lr_{fg}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})\} - E_{\hat{\boldsymbol{\alpha}}}\{E_{\hat{\boldsymbol{\beta}}}[lr_{fg}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})]\},$$

with the analogous definition for a statistic $S_{gf}(\boldsymbol{\beta})$. These statistics are asymptotically normally distributed.

Shen [34], also using **Kullback–Liebler** measure of direct divergence of a density function $g(\mathbf{y}, \boldsymbol{\beta})$ from a target density function $f(\mathbf{y}, \boldsymbol{\alpha})$, proposed a test of H_f against H_g based on a classical chi-square result for the likelihood ratio (*see* **Likelihood Ratio Tests**) by testing $g(\mathbf{y}, \boldsymbol{\beta}_{\boldsymbol{\alpha}})$ against $g(\mathbf{y}, \boldsymbol{\beta})$ since $g(\mathbf{y}, \boldsymbol{\beta}_{\boldsymbol{\alpha}})$ is the closest member to $f(\mathbf{y}, \boldsymbol{\alpha})$.

A test based on the empirical **moment generating function** was studied by Epps et al. [9]. The statistic is based on the asymptotically normal distribution of

$$G_{fg} = M(t) - M_f(t, \hat{\boldsymbol{\alpha}}),$$

where $M(t) = n^{-1} \sum \exp(ty_j)$ and $M_f(t, \boldsymbol{\alpha}) = \mathrm{E}_{\boldsymbol{\alpha}}\{\exp(ty)\}$, are, respectively, the empirical moment generating function and the moment generating function of $f(y, \boldsymbol{\alpha})$.

Sawyer [30], using the distributional results in [8], proposed a multiple test applied when K alternate models are under consideration. Let $f_i(y, \boldsymbol{\alpha}_i), i = 1, \ldots, K$, be the densities considered, and T_{ij}, the $K - 1$ Cox statistics for testing the null hypothesis H_i against each alternative hypothesis H_j, $j \neq i$, and let $\mathbf{T}'_i = (T_{i1}, \ldots, T_{i,i-1}, T_{i,i+1}, \ldots, T_{ik})$. The statistics for testing H_i against all others H_j, $j \neq i$, is

$$\mathbf{T}'_i \boldsymbol{\Sigma}^{-1} \mathbf{T}_i,$$

which is asymptotically χ^2_{k-1} under H_i. Here $\boldsymbol{\Sigma}$ is the covariance matrix $\mathbf{C}_i(T_{ij}, T_{i1})$ which is obtained from the results of Cox. An analogous test can be obtained from the exponential mixture as a Lagrange multiplier test.

Comparison among these alternative statistics does not suggest any general preferences.

Finally, other alternative tests for separate families are based on most powerful invariant statistics [20], the generalized **method of moments** [2] and **bootstrap** methods [31–33, 36].

Bayesian Analysis

Another general approach suggested by Cox [7] used Bayesian inference. The posterior odds for H_f vs. H_g are

$$\frac{\pi_f \int f(\mathbf{y}, \boldsymbol{\alpha}) \pi_f(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha}}{\pi_g \int g(\mathbf{y}, \boldsymbol{\beta}) \pi_g(\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\beta}} = \frac{\pi_f}{\pi_g} B_{fg}(y),$$

where π_f and π_g are the **prior probabilities** of H_f and H_g, respectively, and $\pi_f(\boldsymbol{\alpha})$ and $\pi_g(\boldsymbol{\beta})$ are the prior probabilities for the parameters conditional on H_f and H_g. $B_{fg}(y)$ is the Bayes factor and represents the weight of evidence in the data for H_f over H_g. Cox also gives a general expression when **loss functions** are involved, and a large-sample approximation.

One difficulty with this approach lies in the fact that the prior knowledge expressed by π_f and $\pi_f(\boldsymbol{\alpha})$ must be coherent with that of π_g and $\pi_g(\boldsymbol{\beta})$. If the parameter spaces have different dimensions and there is no simple relation between the parameters, the problems are not simple. When prior information is weak and improper priors are used there are also difficulties and paradoxes with the use of Bayes factors which are unspecified (see [1, 19]). To overcome these difficulties due to improper priors the following alternatives have been proposed recently.

Posterior Bayes Factor [1]

The posterior density $\pi_f(\boldsymbol{\alpha}|y)$ under H_f is, by **Bayes' theorem**,

$$\pi_f(\boldsymbol{\alpha}|\mathbf{y}) = \frac{f(\mathbf{y}, \boldsymbol{\alpha}) \pi_f(\boldsymbol{\alpha})}{\int f(\mathbf{y}, \boldsymbol{\alpha}) \pi_f(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha}}.$$

Let

$$\bar{l}_f(\boldsymbol{\alpha}) = \int f(\mathbf{y}|\boldsymbol{\alpha}) \pi(\boldsymbol{\alpha}|\mathbf{y}) \, \mathrm{d}\boldsymbol{\alpha}$$

be the posterior mean of the likelihood function under H_f. Define $\bar{l}_g(\boldsymbol{\beta})$ similarly. The ratio of the posterior means $PB_{fg} = \bar{l}_f(\boldsymbol{\alpha}) / \bar{l}_g(\boldsymbol{\beta})$ is called the posterior Bayes factor.

Partial Bayes Factor [19]

Here, the sample is divided in two parts $(y, \ldots, y_i)$ $(y_{i+1}, \ldots, y_n) = (\mathbf{x}, \mathbf{z})$. The first part is used as a training sample to obtain a proper posterior $\pi_f(\boldsymbol{\alpha}|\mathbf{x})$ which is taken as a prior distribution to be used with the second part $\mathbf{z}$ of the data. Similarly, $\pi_g(\boldsymbol{\beta}|\mathbf{x})$ is obtained. The partial Bayes factor is defined as

$$PB_{fg}(\mathbf{z}|\mathbf{x}) = \frac{\int f(\mathbf{z}, \boldsymbol{\alpha}) \pi_f(\boldsymbol{\alpha}|\mathbf{x}) \, \mathrm{d}\boldsymbol{\alpha}}{\int g(\mathbf{z}, \boldsymbol{\beta}) \pi_g(\boldsymbol{\beta}|\mathbf{x}) \, \mathrm{d}\boldsymbol{\beta}}$$

Intrinsic Bayes Factor [5]

Suppose that $y = (\mathbf{x}, \mathbf{z})$ and that $\mathbf{x}$ is a minimal training sample for the comparison of H_f and H_g if the posteriors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are proper and no subset of $\mathbf{x}$ gives a proper posterior. There are usually many training samples. Let N be the number of training

samples $\mathbf{x}$. The idea of an intrinsic Bayes factor is to use the median or average of all the partial Bayes factors obtained for the N training samples. If N is too large the suggestion is to take a random sample from the collection of possible training samples.

The geometric intrinsic Bayes factor is

$$IB_{fg}^{G}(\mathbf{y}) = \left\{ \prod_{i=1}^{N} PB_{fg}(\mathbf{z}_i|\mathbf{x}_i) \right\}^{1/N},$$

the arithmetic Bayes factor is

$$IB_{fg}^{A}(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} PB_{fg}(\mathbf{z}_i|\mathbf{x}_i),$$

and the median Bayes factor is $IB_{fg}^{M}(\mathbf{y}) = \text{med}\{PB_{fg}(\mathbf{z}_i|\mathbf{x}_i)/i = 1, \ldots, N\}$. Other measures of location could also be defined.

Fractional Bayes Factor [19]

The Fractional Bayes Factor with training fraction b is defined by

$$B_{fg}^{b}(\mathbf{y}) = \frac{q_f(\mathbf{y})}{q_g(\mathbf{y})},$$

where

$$q_f(\mathbf{y}) = \frac{\int f(\mathbf{y}, \boldsymbol{\alpha}) \pi_f(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha}}{\int f^b(\mathbf{y}, \boldsymbol{\alpha}) \pi_f(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha}}$$

and there is an analogous expression for $q_g(\mathbf{y})$.

These different Bayes factors can be interpreted using Jeffreys' rule, but their properties are still under investigation.

Applications

Applications of the procedures of this article are reviewed in [21, 24], and [17], where the focus is primarily on econometric research. In econometrics the emphasis is on testing, $\mathrm{H}_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0\lambda \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_0^2)$ against $\mathrm{H}_i : \mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_i^2)$, $i = 1, \ldots, m$, where $\mathbf{X}$ and $\mathbf{Z}_i$ represent separate explanatory variables. Pereira [26] tests H_0 vs. H_1, when the $\boldsymbol{\varepsilon}_i$ follow a Weibull distribution. He shows that the results of [22] hold also when the alternative hypotheses specify alternative regressors with alternative error distributions. McAller's paper [17] also reviews the linear $\times$ loglinear and the **time series** hypotheses.

Some biostatistical applications can be found in [4, 6, 8, 12–14, 25, 32, 33, 35], and [36].

References

[1] Aitkin, M. (1991). Posterior Bayes factors (with discussion), *Journal of the Royal Statistical Society, Series B* **53**, 111–142.

[2] Arkonac, S.Z. & Higgins, M.L. (1995). A Monte Carlo study of tests for non-nested models estimated by generalized method of moments, *Communications in Statistics – Simulation and Computation* **24**, 745–763.

[3] Atkinson, A.C. (1970). A method for discriminating between models (with discussion), *Journal of the Royal Statistical Society, Series B* **32**, 323–353.

[4] Bain, L.J. & Engelhardt, M. (1980). Probability of correct selection of Weibull versus gamma based on likelihood ratio, *Communications in Statistics – Theory and Methods* **9**, 375–381.

[5] Berger, J.O. & Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association* **91**, 109–122.

[6] Cole, T.J. (1975). Linear and proportional regression models in the prediction of ventilatory function (with discussion), *Journal of the Royal Statistical Society, Series A* **138**, 297–337.

[7] Cox, D.R. (1961). Tests of separate families of hypotheses, *Proceedings of the Fourth Berkeley Symposium*, Vol. 1. University of California Press, Berkeley, pp. 105–123.

[8] Cox, D.R. (1962). Further results on tests of separate families of hypotheses, *Journal of the Royal Statistical Society, Series B* **24**, 406–423.

[9] Epps, T.W., Singleton, K.J. & Pulley, L.B. (1982). A test of separate families of distributions based on the empirical moment generating function, *Biometrika* **69**, 391–399.

[10] Ghosh, J.K. & Subramanyan, K. (1975). Inference about separated families in large samples, *Sankhyā, Series A* **37**, 502–513.

[11] Gourieroux, C. & Monfort, A. (1994). Testing non-nested hypothesis, in *Handbook of Econometrics*, Vol. IV R. Engle & D.L. Mcfadden, eds. Elsevier, London, pp. 2585–2637.

[12] Kotz, S. (1973). Normality vs. lognormality with applications, *Communications in Statistics* **1**, 113–132.

[13] Lindsey, J.K. (1974). Comparison of probability distributions, *Journal of the Royal Statistical Society, Series B* **36**, 38–47.

[14] Lindsey, J.K. (1974). Construction and comparison of statistical models, *Journal of the Royal Statistical Society, Series B* **36**, 419–425.

[15] MacKinnon, J.G. (1983). Model specification tests against non-nested alternatives, *Econometric Reviews* **2**, 85–110.

[16] McAller, M. (1987). Specification tests for separate models: a survey, in *Specification Analysis in the Linear Model*, M.L. King & D.E. Giles, eds. Routledge & Kegan Paul, London, pp. 146–196.

[17] McAller, M. (1995). The significance of testing empirical non-nested models, *Journal of Econometrics* **65**, 149–171.

[18] McAller, M. & Pesaran, M.H. (1986). Statistical inference in non-nested econometric models, *Applied Mathematics and Computation* **20**, 271–311.

[19] O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion), *Journal of the Royal Statistical Society, Series B* **57**, 99–138.

[20] Pandey, M., Ferdous, J. & Udden, M.B. (1991). Selection of probability distributions for life testing data, *Communications in Statistics – Theory and Methods* **20**, 1373–1388.

[21] Pereira, B. de B. (1977). Discriminating among separate models: a bibliography, *International Statistical Review* **45**, 163–172.

[22] Pereira, B. de B. (1978). Tests and efficiencies of separate regression models, *Biometrika* **65**, 319–327; also *Biometrika* **68**, (1981). 34.

[23] Pereira, B. de B. (1978). Empirical comparisons of some tests of separate families of hypotheses, *Metrika* **25**, 219–239.

[24] Pereira, B. de B. (1981). Discriminating among separate models: an additional bibliography, *International Statistical Information* **6**, 3; Reprinted in Katti, S. K. (1982) On the Preliminary Test for the CEAS Model versus the Thompson Model for Predicting Soybean Production, *Technical Report 125*, Department of Statistics, University of Missouri, Columbia.

[25] Pereira, B. de B. (1981). Choice of a survival model for patients with a brain tumour, *Metrika* **28**, 53–61.

[26] Pereira, B. de B. (1984). On the choice of a Weibull model, *Journal of the American Statistical Institute* **26**, 157–163.

[27] Pesaran, M.H. (1987). Global and partial non-nested hypotheses and asymptotic local power, *Econometric Theory* **3**, 69–97.

[28] Quandt, R.E. (1974). A comparison of methods for testing nonnested hypotheses, *Review of Economics and Statistics* **56**, 92–99.

[29] Sawyer, K.R. (1983). Testing separate families of hypotheses: an information criterion, *Journal of the Royal Statistical Society, Series B* **45**, 89–99.

[30] Sawyer, K.R. (1984). Multiple hypotheses testing, *Journal of the Royal Statistical Society, Series B* **46**, 419–424.

[31] Schork, N. (1993). Combining Monte Carlo and Cox tests of non-nested hypotheses, *Communications in Statistics – Simulation and Computation* **22**, 939–954.

[32] Schork, N. & Schork, M.A. (1989). Testing separate families of segregation hypotheses: bootstrap methods, *American Journal of Human Genetics* **45**, 803–813.

[33] Schork, N., Weder, A.B. & Schork, M.A. (1990). On the asymmetry of biological frequency distribution, *Genetic Epidemiology* **7**, 417–446.

[34] Shen, S.M. (1982). A method for discriminating between models describing compositional data, *Biometrika* **69**, 587–595.

[35] Thomas, D.G. (1972). Tests of fit for a one-hit vs. two-hit curve, *Applied Statistics* **21**, 103–112.

[36] Wahrendorf, J., Becher, H. & Brown, C.C. (1987). Bootstrap comparison of non-nested generalized linear models: applications in survival analysis and epidemiology, *Applied Statistics* **36**, 72–81.

Further Reading

Pesaran, M.H. & Weeks, M. (2001). Nonnested hypothesis testing: an overview, in *Companion in Theoretical Econometrics*, B.H. Baltagi, eds. Basil Blackwell, Oxford.

BASILIO DE BRAGANCA PEREIRA

Sequence Analysis

Owing to the abundant and rapidly increasing availability of genomic sequence data, we are confronted with many new methodologic problems, most of which are statistical in nature. Statistical significance estimates play a critical role in biologic sequence comparisons, or, more specifically, in pairwise alignment. The most basic sequence analysis task is to ask if two genomic or protein sequences are homologous. That is, are they derived from the same evolutionary ancestor (*see* **Cladistic Analysis**)? This can be achieved by first aligning the sequences to measure the sequence similarity by using some scoring function, and then using a test for statistical significance of the similarity measures to infer whether alignment is due to chance alone.

We assume an evolutionary model where the two sequences have diverged from a common ancestor by the process of **mutation** and selection (*see* **Population Genetics**). Mutations can be classified by the type of changes caused by substitution, which changes one nucleotide to another in a **DNA sequence** or changes one amino acid to another in a protein. Natural selection works as a screening process to weed out certain deleterious mutations and favors neutral or advantageous mutations. Comparisons of two sequences usually cannot determine whether a deletion has occurred in one sequence or an insertion has occurred in the other. Insertions and deletions are together referred to as *gaps*. When scoring the alignment, the gaps are penalized by assigning a cost to a gap by a function of a gap length. There is a well-established theory for generating a substitution matrix for every pair of nucleotides or pair of amino acids.

Let $x = x_1, x_2, \ldots, x_n$ and $y = y_1, y_2, \ldots, y_m$ be two sequences of length n and m, respectively, where x_i and y_j represent the elements of the set {A, G, C, T} in the case of a DNA sequence, and they assume values from the set of 20 amino acids in the case of proteins. For simplicity, suppose we are given two aligned sequences of equal length, i.e. $n = m$. How do we test for a significantly good match? The null hypothesis is that x and y do not diverge from the common ancestor. So, the hypothesis assumes an underlying random model R where x_i in x and y_j in y occur independently and hence the probability of two sequences is just the product of the probabilities of each letter in the sequences:

$$\Pr(x, y|\mathrm{R}) = \prod p_{x_i} \prod p_{y_j}. \tag{1}$$

The alternative hypothesis A is that x and y have diverged from same ancestor, i.e. x and y have each independently been derived from some unknown ancestor sequence z. Let $p_{x_i y_j}$ be the probability that x_i and y_j are derived from z_k. So, the probability for the whole alignment is

$$\Pr(x, y|\mathrm{A}) = \prod p_{x_i y_j}. \tag{2}$$

The ratio of these two **likelihoods** gives a reasonable score for the alignment, since it compares the alternative hypothesis that is based on evolution with the **null hypothesis** of the random model. To achieve an additive scoring system, we take the logarithm of this ratio by defining the entries of the substitution score matrix by

$$s(x_i, y_j) = \log\left(\frac{p_{x_i y_j}}{p_{x_i} p_{y_j}}\right). \tag{3}$$

For proteins, we have a 20×20 matrix, with $s(x_i, y_j)$ in the position i, j in the matrix, where x_i and y_j correspond to ith and jth residues. The most commonly used scoring matrices are PAM matrices [2] and BLOSUM matrices [6].

Now, given the scoring system, we need to have an algorithm to find an optimal alignment for a pair of sequences. There are several methods for finding alignment, depending on whether interest focuses on global alignment, i.e. involving entire sequences, or on local alignment, i.e. involving just some part of the sequences. One approach would be to find all the alignments and then pick the best one. However, the number of alignments between two sequences is exponential, and such an approach would result in an extremely slow algorithm. To find the global alignment of protein sequences, Needleman & Wunch [9] developed a dynamic programming algorithm, and there are a number of extensions to the original algorithm, most notably by Gotoh [4].

The basic idea of the Needleman & Wunch algorithm is to build up the optimal alignment recursively from the previously aligned subsequences. We construct a matrix M indexed by i and j corresponding to x_i and y_j in the sequences x and y, respectively. Let d_{ij} be the score of the optimal alignment between the subsequence of x up to x_i and

subsequence of y up to y_j. If $d_{i-1,j-1}$, $d_{i-1,j}$, and $d_{i,j-1}$ are known, then it is possible to calculate d_{ij}. There are three possible ways that the d_{ij} could be obtained. The x_i could be aligned to y_j, in this case, $d_{ij} = d_{i-1,j-1} + s(x_i, y_j)$; or x_i is aligned to a gap, in this case $d_{ij} = d_{i-1,j}$ + gap penalty; or y_j is aligned to a gap, in this case $d_{ij} = d_{i,j-1}$ + gap penalty. The optimal score will be the largest of these three values, therefore

$$d_{ij} = \max \begin{cases} d_{i-1,j-1} + s(x_i, y_j), \\ d_{i-1,j} + \text{gap penalty}, \\ d_{i,j-1} + \text{gap penalty}. \end{cases} \tag{4}$$

We fill the entries of the matrix M recursively, and keep a pointer in each cell back to the cell from which it was derived. The value in the final cell is, by definition, the best score for an alignment of x and y. To find the alignment itself, we trace back through the matrix M in the usual manner, setting our pointer back from the final cell, and retracing back through the cell from which it was derived. At the end, we will reach the start of the matrix, $i = j = 0$.

Local alignment algorithms are very similar to the preceding method, except that the goal is different. Instead of trying to find similarity between the sequences, we are now trying to find the best alignment between subsequences of x and y. The highest scoring alignment of subsequences of x and y is called the best local alignment. The algorithm by Smith & Waterman [11] is commonly used, and is

$$d_{ij} = \max \begin{cases} 0, \\ d_{i-1,j-1} + s(x_i, y_j), \\ d_{i-1,j} + \text{gap penalty}, \\ d_{i,j-1} + \text{gap penalty}. \end{cases} \tag{5}$$

The only difference between this and the global optimal score (4) given above is that d_{ij} can take on the value 0 if all other options are negative. Because of the 0 values in the matrix, the score can never become negative, and hence we will obtain areas of similarity even if there are long mismatches or gaps in between them.

Suppose we have optimal alignment. How do we assess the significance of this alignment statistically? The statistical theory has been well developed by Karlin & Altschul [7, 8] and Dembo & Karlin [3] and is implemented in the basic local alignment search tool (BLAST) [1]. BLAST returns a list of high-scoring matched subsequences between the query sequence and sequences in the database. We can determine the distribution of the maximum of N match scores compared with independent random sequences. The score of a match to a random sequence is the sum of many similar random variables, so can be approximated by a normal distribution, and the limiting distribution of the maximum of N identical independent normal variables is known to be the extreme value distribution (EVD) [5]. So, we can use the EVD to calculate the probability that the optimal match from the search of a large number of unrelated sequences has a score greater than our observed maximal score. If this is less than some small value, then the observation is considered significant. FASTA [10] is another heuristic sequence searching package widely used for sequence database search.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool, *Journal of Molecular Biology* **215**, 403–410.

[2] Dayoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978). model of evolutionary change in proteins, in *Atlas of Protein Sequence and Structure*, Vol. 5, Supplement 3, M.O. Dayoff, ed. National Biomedical Foundation, Washington, pp. 345–352.

[3] Dembo, A. & Karlin, S. (1991). Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables, *Annals of Probability* **19**, 1737–1755.

[4] Gotoh, O. (1982). An improved algorithm for matching biological sequences, *Journal of Molecular Biology* **162**, 705–708.

[5] Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

[6] Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks, *Proceedings of the National Academy of Sciences* **89**, 10915–10919.

[7] Karlin, S. & Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proceedings of the National Academy of Sciences* **87**, 2264–2268.

[8] Karlin, S. & Altschul, S.F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences, *Proceedings of the National Academy of Sciences* **90**, 5873–5877.

[9] Needleman, S.B. & Wunch, C.D. (1970). A general method applicable to search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology* **48**, 443–453.

[10] Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison, *Proceedings of the National Academy of Sciences* **4**, 2444–2448.

[11] Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences, *Journal of Molecular Biology* **147**, 195–197.

(*See also* **Bioinformatics**)

HEMANT K. TIWARI

Sequential Analysis

The subject of sequential analysis was initiated by **Abraham Wald** [39, 40] in response to demands for more efficient sampling inspection procedures during World War II.

Sequential Tests of Hypotheses

Wald introduced the sequential probability test (SPRT) of a simple **null hypothesis** $H_0 : f = f_0$ vs. a simple **alternative hypothesis** $H_1 : f = f_1$ based on independent observations $X_1, X_2, \ldots$ having a common density function f (*see* **Hypothesis Testing**). The test stops sampling at stage

$$N = \text{first } n \geq 1 \text{ such that } r_n \leq A \text{ or } r_n \geq B,$$

where $0 < A < 1 < B$, $r_n = \prod_{i=1}^{n} [f_1(X_i)/f_0(X_i)]$ and N is defined to be ∞ if $A < r_n < B$ for all n. The SPRT rejects H_0 if $r_N \leq A$ and rejects H_1 if $r_N \geq B$. Wald & Wolfowitz [41] showed that it has the following optimality property: among all tests whose expected sample sizes under H_0 and H_1 are finite and whose type I and type II error probabilities (*see* **Level of a Test**) are less than or equal to those of the SPRT, denoted by α and β, respectively, the SPRT minimizes the expected sample sizes under H_0 and H_1. Moreover, Wald [39] showed that

$$\alpha \leq \frac{(1-\beta)}{B}, \qquad \beta \leq A(1-\alpha),$$

and that these inequalities become equalities if r_N does not "overshoot" the boundary B or A. Ignoring overshoots, Wald treated these inequalities as equalities and arrived at the approximation $A \simeq \beta/(1-\alpha)$, $B \simeq (1-\beta)/\alpha$ to determine the boundaries of the SPRT from prescribed error probabilities α and β.

Within a few years after Wald's introduction of the subject, it was recognized that sequential hypothesis testing might provide a useful tool in biomedical studies. In particular, making use of Wald's theory of the SPRT, Morton [26] developed a standard for proving genetic linkage (*see* **Linkage Analysis, Model-based**). A number of papers appeared during the 1950s on modifications of the SPRT for the design of sequential clinical trials, and an overview of these developments was given in the first edition of Armitage's book [3] in 1960. Subsequently, Armitage et al. [4] proposed a new alternative to the SPRT and its variants. This is the "repeated significance test" (RST), a detailed treatment of which appeared in the second edition of Armitage's book in 1975. The underlying motivation is that, since the strength of evidence in favor of a treatment from a clinical trial is conveniently indicated by the results of a conventional significance test, it is appealing to apply such a test, with nominal significance level α, repeatedly during the trial. However, the overall significance level α^*, which is the probability that the nominal significance level is attained at some stage, may be substantially larger than α.

For example, suppose that $X_1, X_2, \ldots$ are independent normal with unknown mean μ and known variance σ^2. Let $S_n = X_1 + \cdots + X_n$. The conventional significance test of $H_0 : \mu = 0$ based on $X_1, \ldots, X_n$ rejects H_0 if $|S_n| \geq a\sigma\sqrt{n}$, where $1 - \Phi(a) = \alpha/2$. The RST, with a maximum sample size M, stops sampling and rejects H_0 at stage

$$T = \text{first } n \geq 1 \text{ with } n \leq M \text{ such that } |S_n| \geq a\sigma\sqrt{n}.$$

If $|S_n| < a\sigma\sqrt{n}$ for all $1 \leq n \leq M$, then the RST does not reject H_0. The overall significance level of the test is

$$\alpha^* = \Pr_{\mu=0}(|S_n| \geq a\sigma\sqrt{n} \text{ for some } 1 \leq n \leq M)$$

$$= 1 - \Phi(a) + \sum_{n=2}^{M} p_n(a),$$

where $p_n(a) = \Pr_{\mu=0}(|S_n| \geq a\sigma\sqrt{n}$ and $|S_j| < a\sigma\sqrt{j}$ for $1 \leq j < n)$. Armitage et al. [4] developed a recursive numerical integration algorithm to evaluate $p_n(a)$. The choice of a is such that the overall significance level α^* (instead of the nominal significance level) is equal to some prescribed number. For example, for $\alpha^* = 0.05$ and $M = 71$, Table 5.5 of [3] gives $a = 2.84$, which corresponds to a nominal significance level of $\alpha = 0.005 = \alpha^*/10$. Note that $a = 2.84$ is considerably larger than the value 1.96 associated with a 5% level significance test with fixed sample size M. The price of the smaller expected sample size of the RST is, therefore, a loss of power compared to a fixed sample size test with the same significance level and the same M.

Haybittle [17], Peto et al. [28], and Siegmund [33] proposed the following modification of the RST to increase its power. The stopping rule has the same

form as the preceding RST but the rejection region is modified to

$$T \leq M-1 \quad \text{or} \quad |S_M| \geq c\sigma\sqrt{M},$$

where $a \geq c$ are so chosen that the overall significance level is equal to some prescribed number. In particular, $a = \infty$ gives the fixed sample size test and $a = c$ gives the RST.

Pocock [29] introduced another modification of the RST. Noting that in practice it is difficult to arrange for continuous examination of the data as they accumulate to perform the RST, he considered a "group sequential" (*see* **Data and Safety Monitoring**) version in which X_n above represents an approximately normally distributed statistic of the data in the nth group (instead of the nth observation) and M represents the maximum number of groups. Instead of the square-root boundary $a\sigma\sqrt{n}$ for $|S_n|$ in the group sequential RST, O'Brien & Fleming [27] proposed to use a constant stopping boundary b that does not change with n. Siegmund [34] gives an extensive treatment of the theory of truncated sequential tests, and in particular of the RST and its modifications.

The problem of group sequential testing for the mean of a normal distribution with known variance discussed above serves as a prototype for more complex situations. Note that a group sequential test for a normal mean, assuming M equally sized groups, involves a stopping rule for $(S_1, \ldots, S_M)$ which has a **multivariate normal distribution** with $\text{var}(S_n) = n\sigma^2 = \text{cov}(S_i, S_n)$ for $i \geq n$. For more complicated statistics U_n in more general situations, one has an asymptotically normal distribution for $(U_1, \ldots, U_M)$ whose covariance matrix is not known in advance and has to be estimated from the data. Flexible methods to construct stopping boundaries of group sequential tests in these situations have been proposed by Slud & Wei [35], Lan & DeMets [25], Fleming et al. [13], and Jennison & Turnbull [18] (*see* **Data and Safety Monitoring**).

For example, consider a **clinical trial** whose primary objective is to compare survival times (times to failure) between two treatment groups. Patients enter the trial serially and are randomized to either treatment and then followed until they fail or withdraw from the study, or until the trial is terminated. The trial is scheduled to end by a certain time t_M and there are also $M-1$ periodic reviews at calendar times $t_1, \ldots, t_{M-1}$ prior to t_M (*see* **Data and Safety Monitoring**). Let U_i be the **logrank** statistic calculated at calendar time t_i. Then under the null hypothesis that the two treatment groups have the same survival distribution $(U_1, \ldots, U_M)$ is asymptotically normal, as shown by Tsiatis [38], who considered more general **rank** statistics including the logrank statistics as a special case. It is also shown in [38] how the asymptotic covariances of the U_i can be estimated to perform group sequential testing. For the logrank and many other rank statistics, U_i and $U_j - U_i$ are asymptotically independent for $j > i$, so one needs only estimate $\text{var}(U_i)$ in this case.

Whitehead [42] gives a comprehensive overview of these and other methods for sequential hypothesis testing in clinical trials. He considers the case where $(U_1, \ldots, U_M)$ is asymptotically normal under the null hypothesis, with $U_j - U_i$ asymptotically independent of U_i for $j > i$. Letting V_i denote a consistent estimate of the null variance of U_i for $i = 1, \ldots, M$, he advocates the use of certain triangular stopping boundaries in the (V_i, U_i) plane and has developed a computer package, PEST, for their implementation (*see* **Software, Biostatistical**). These triangular boundaries are associated with the problem of minimizing the maximum expected sample size of sequential tests for the mean μ of a normal distribution subject to constraints on the type I error at $\mu = 0$ and type II error at some $\mu \neq 0$, as shown by Lai [20].

Sequential Estimation

Analysis of the data at the conclusion of a clinical study typically not only permits testing of the null hypothesis but also provides **estimates** of parameters associated with the primary and secondary end points. The use of a stopping rule whose distribution depends on these parameters introduces substantial difficulties in constructing valid **confidence intervals** for the parameters at the conclusion of the study. For example, consider the simple example of independent normal X_i with unknown mean μ and known variance σ^2. For a sample of fixed size n, the sample mean $\overline{X}_n$ is an **unbiased** estimate of μ and has a normal distribution with variance σ^2/n, yielding the classical confidence interval $\overline{X}_n \pm z_{1-\alpha}\sigma/\sqrt{n}$ with coverage probability $1-2\alpha$ for μ, where z_α is the α-**quantile** of the standard normal distribution. If n is replaced by a stopping rule T whose distribution depends on μ, then $\overline{X}_T$ is typically biased and

$\sqrt{T}(\overline{X}_T - \mu)/\sigma$ is no longer standard normal but has a distribution that depends on μ.

Rosner & Tsiatis [31] proposed the following method to construct a $1 - 2\alpha$ confidence interval for μ. For every value of μ, find the quantiles $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$ of $\sqrt{T}(\overline{X}_T - \mu)$, i.e.

$$\Pr_\mu[\sqrt{T}(\overline{X}_T - \mu) < u_\alpha(\mu)] = \alpha$$
$$= \Pr_\mu[\sqrt{T}(\overline{X}_T - \mu) > u_{1-\alpha}(\mu)].$$

These probabilities can be computed by the recursive numerical integration algorithm of Armitage et al. [4] when T is bounded by M. Hence the confidence region $\{\mu : u_\alpha(\mu) \le \sqrt{T}(\overline{X}_T - \mu) \le u_{1-\alpha}(\mu)\}$ has coverage probability $1 - 2\alpha$. Note that this confidence region reduces to an interval whose end points are found by intersecting the line $\sqrt{T}(\overline{X}_T - \mu)$ with the curves $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$ if there is only one intersection with each curve, which is the case commonly encountered in practice.

Siegmund [33] proposed another approach based on ordering the sample space in a certain way, following an earlier proposal of Armitage [2]. Chapter 5 of Whitehead's monograph [42] gives a comprehensive treatment of this ordering approach. It also discusses the construction and properties of bias-adjusted estimates following sequential tests. Emerson & Fleming [12] proposed an alternative ordering and used it to construct bias-adjusted estimates and confidence intervals.

A considerably simpler class of sequential estimation problems deals with estimation of a parameter θ with prescribed accuracy using a randomly stopped statistic $\hat{\theta}_N$, whose stopping rule N is targeted towards achieving the prescribed accuracy. In these problems, Anscombe's [1] **central limit theorem** for randomly stopped sums typically yields adequate normal approximations for the distribution of $\sqrt{N}(\hat{\theta}_N - \theta)$, which can be used to construct fixed-width confidence intervals for θ. For example, let $X_1, X_2, \ldots$ be independent random variables from a population with mean μ and variance σ^2. The variance of the estimate $\overline{X}_n$ of μ is σ^2/n and an approximate $1 - 2\alpha$ confidence interval for μ is $\overline{X}_n \pm z_{1-\alpha}\sigma/\sqrt{n}$, which can be made to have width $2d$ by choosing n to be the smallest integer $\ge (z_{1-\alpha}\sigma/d)^2$, assuming σ to be known. When σ is unknown, Chow & Robbins [7] proposed to replace it by the sample variances at successive stages, leading to the stopping rule

$$N = \text{first } n \ge m \text{ such that } nd^2/z_{1-\alpha}^2 \ge (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + n^{-1}.$$

The confidence interval is taken to be $\overline{X}_N \pm d$. This has approximate coverage probability $1 - 2\alpha$ when d is small, since $\sqrt{N}(\overline{X}_N - \mu)/\sigma$ has a limiting standard normal distribution as $d \to 0$ by Anscombe's theorem. Schmidt et al. [32] used this procedure to construct fixed-width confidence intervals for the concentrations of enzymes in the normal human pancreas. Two-stage and three-stage analogs of this fully sequential procedure were developed by Stein [37] and Hall [16].

Adaptive Allocation, Sequential Design, and Decision Theory

Other topics in the field of sequential analysis of interest to biomedical studies are adaptive treatment allocation (*see* **Adaptive and Dynamic Methods of Treatment Assignment**) and sequential design of experiments. The "**up-and-down**" (staircase) method in bioassay and dosage determination is an example of sequential experimentation. A traditional nonsequential method for performing a **bioassay** experiment is to test a prescribed number of animals at each of several fixed dose levels. In the up-and-down method, one chooses a series of test levels with equal spacing (on an appropriate scale, usually log dose) between doses, and carries out a series of trials using the following rule: use the next higher dose following a negative response and use the next lower dose following a positive response. Details of implementation of the design and applications to estimation of the LD_{50} ("lethal dose 50" – the dose producing response on 50% of the subjects) are given by Dixon & Moood [10] and Dixon [9].

Stochastic approximation, introduced by Robbins & Monro [30], is another example of sequential experimentation. In the context of quantal bioassay, Cochran & Davis [8] considered the following version of the Robbins–Monro scheme. To start the experiment, an initial guess x_1 of LD_{50} is made and m animals are given dose x_1. Let $c > 0$. For $n \ge 1$, let $\hat{p}_n$ be the observed proportion of deaths

in the group of m animals assigned dose x_n, and define

$$x_{n+1} = x_n - cn^{-1}\left(\hat{p}_n - \tfrac{1}{2}\right)$$

to be the dose level at which another group of m animals are tested. The basic idea behind stochastic approximation is to use the recursive scheme $x_{n+1} = x_n - a_n(Y_n - h)$ to find the solution θ of the equation $M(\theta) = h$, in which $M(x)$ is not observable and all that can be observed for each x is a random variable $Y(x)$ with $\mathrm{E}[Y(x)] = M(x)$ and in which $Y_n = Y(x_n)$. Thus, in the quantal bioassay application above, $h = \frac{1}{2}$ and $Y_n = \hat{p}_n$. Under certain regularity conditions, the best choice of a_n is $(\beta n)^{-1}$, where β is the derivative of M at θ. Adaptive stochastic approximation schemes that replace the unknown β in the optimal choice $a_n = (\beta n)^{-1}$ by simple recursive estimates b_n have been proposed and analyzed by Lai & Robbins [24] and Frees & Ruppert [14].

In classical fixed-sample **decision theory**, one has a parameter space containing all possible values of the unknown parameter θ, an action space consisting of all possible actions a, and a **loss function** $L(\theta, a)$ representing the loss when the true parameter is θ and action a is taken. In sequential decision theory, one has a sequence of actions $a_1, a_2, \ldots$ and loss $L_n(\theta, a_n)$ at stage n. For sequential experimental design problems of the type described above, a_n is the choice of the design level x_n at stage n. For sequential hypothesis testing (or estimation) problems, a_n represents whether stopping occurs at stage n and also acceptance of the null or alternative hypothesis (or the estimate of the unknown parameter) when stopping indeed occurs at stage n. In this case, it is more convenient to represent the action sequence $(a_1, a_2, \ldots)$ by a stopping rule denoting when stopping occurs and a terminal decision rule denoting the action taken upon stopping. Given successive observations $Z_1, Z_2, \ldots$, whose joint distribution depends on θ, a finite-horizon sequential decision problem, with horizon M, is to choose action $d_n = d_n(Z_1, \ldots, Z_n)$ at stage n on the basis of the current and past observations, for $1 \le n \le M$ (*see* **Adaptive and Dynamic Methods of Treatment Assignment**). The overall risk of $(d_1, \ldots, d_M)$ is $R(\theta) = \mathrm{E}_\theta[\sum_{n=1}^{M} L_n(\theta, d_n)]$. In particular, putting a **prior distribution** G on the parameter space, one can consider the Bayes rule that minimizes $\int R(\theta)\,\mathrm{d}G(\theta)$ (*see* **Bayesian Methods**). The solution can be found by the backward induction algorithm of dynamic programming. Applications of the algorithm to determine optimal stopping boundaries of group sequential tests have been given by Berry & Ho [5] and Eales & Jennison [11]. Chernoff's monograph [6] gives a comprehensive treatment of optimal stopping problems in sequential analysis. Spiegelhalter et al. [36] discuss Bayesian approaches to monitoring clinical trials.

The handbook edited by Ghosh & Sen [15] gives extensive references and survey articles on a wide variety of topics in sequential analysis including those covered in the present brief review which is oriented towards biomedical applications. The monograph by Jennison & Turnbull [19] on group sequential tests and the review articles by Lai [21, 22, 23] describe important developments in stochastic approximation, interim and terminal analyses of clinical trials with failure-time endpoints, and other areas of sequential analysis following the publication of the First Edition and provide updated lists of references.

References

[1] Anscombe, F.J. (1992). Large sample theory of sequential estimation, *Proceedings of the Cambridge Philosophical Society* **48**, 600–607.

[2] Armitage, P. (1958). Numerical studies in the sequential estimation of a binomial parameter *Biometrika* **45**, 1–15.

[3] Armitage, P. (1975). *Sequential Medical Trials*, 2nd Ed. Blackwell, Oxford.

[4] Armitage, P., McPherson, C.K. & Rowe, B.C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

[5] Berry, D.A. & Ho, C.H. (1988). One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach, *Biometrics* **44**, 219–227.

[6] Chernoff, H. (1972). *Sequential Analysis and Optimal Design*. SIAM, Philadelphia.

[7] Chow, Y.S. & Robbins, H. (1965). On the asymptotic theory of fixed width sequential confidence intervals for the mean, *Annals of Mathematical Statistics* **36**, 457–462.

[8] Cochran, W.G. & Davis, M. (1965). The Robbins–Monro method for estimating the median lethal dose, *Journal of the Royal Statistical Society, Series B* **27**, 28–44.

[9] Dixon, W.J. (1965). The up-and-down method for small samples, *Journal of the American Statistical Association* **60**, 967–978.

[10] Dixon, W.J. & Mood, A.M. (1948). A method for obtaining and analyzing sensitivity data, *Journal of the American Statistical Association* **43**, 109–126.

[11] Eales, J.D. & Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests, *Biometrika* **79**, 13–24.

[12] Emerson, S.S. & Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing, *Biometrika* **77**, 875–892.

[13] Fleming, T.R., Harrington, D.P. & O'Brien, P.C. (1984). Designs for group sequential tests, *Controlled Clinical Trials* **5**, 348–361.

[14] Frees, E.W. & Ruppert, D. (1990). Estimation following a Robbins–Monro designed experiment, *Journal of the American Statistical Association* **85**, 1123–1129.

[15] Ghosh, B.K. & Sen, P.K. (1991). *Handbook of Sequential Analysis*. Marcel Dekker, New York.

[16] Hall, P. (1981). Asymptotic theory of triple sampling for sequential estimation of a mean, *Annals of Statistics* **9**, 1229–1238.

[17] Haybittle, J.L. (1971). Repeated assessments of results in clinical trials of cancer treatment, *British Journal of Radiology* **44**, 793–797.

[18] Jennison, C. & Turnbull, B.W. (1989). Interim analysis: the repeated confidence interval approach (with discussion), *Journal of the Royal Statistical Society, Series B* **51**, 306–361.

[19] Jennison, C. & Turnbull, B.W. (2001). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall & CRC, Boca Raton & London.

[20] Lai, T.L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *Annals of Statistics* **1**, 659–673.

[21] Lai, T.L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica* **11**, 303–408.

[22] Lai, T.L. (2003). Stochastic approximation. *Annals of Statistics* **31**, 391–406.

[23] Lai, T.L. (2003). Interim and terminal analyses of clinical trials with failure-time endpoints and related group sequential designs. In *Applications of Sequential Methodologies*, N. Mukhopadhyay, S. Datta & S. Chattopadhyay, eds. Marcel Dekker, New York, in press.

[24] Lai, T.L. & Robbins, H. (1979). Adaptive design and stochastic approximation, *Annals of Statistics* **7**, 1196–1221.

[25] Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.

[26] Morton, N.E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.

[27] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[28] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design, *British Journal of Cancer* **34**, 585–712.

[29] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.

[30] Robbins, H. & Monro, S. (1951). A stochastic approximation method, *Annals of Mathematical Statistics* **22**, 400–407.

[31] Rosner, G.L. & Tsiatis, A.A. (1988). Exact confidence intervals following sequential tests, *Biometrika* **65**, 341–349.

[32] Schmidt, B., Cornée, J. & Delachaume-Salem, E. (1970). Application de procédures statistiques séquentielles à l'étude des concentrations enzymatiques du suc pancréatiques humain normal, *Comptes Rendus des Séances de la Societé de Biologie Paris* **164**, 1813–1818.

[33] Siegmund, D. (1978). Estimation following sequential tests, *Biometrika* **65**, 341–349.

[34] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.

[35] Slud, E.V. & Wei, L.J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic, *Journal of the American Statistical Association* **157**, 357–416.

[36] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials (with discussion), *Journal of the Royal Statistical Society, Series A* **157**, 357–416.

[37] Stein, C. (1945). A two sample test for a linear hypothesis whose power is independent of the variance, *Annals of Mathematical Statistics* **16**, 243–258.

[38] Tsiatis, A.A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis, *Journal of the American Statistical Association* **77**, 855–861.

[39] Wald, A. (1945). Sequential tests of statistical hypotheses, *Annals of Mathematical Statistics* **16**, 117–186.

[40] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

[41] Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test, *Annals of Mathematical Statistics* **19**, 326–339.

[42] Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*, 2nd Ed. Ellis Horwood, Chichester.

(*See also* **Wald's Identity**)

T.L. LAI

Sequential Linkage Analysis

The term "sequential **linkage analysis**" has several possible meanings to genetic epidemiologists. One arises from extending an already sampled pedigree, by sequentially adding additional contiguous segments to it based upon the presence of the phenotype (*see* **Genotype**) of interest. This sequential extension procedure was originally developed for **segregation analysis**, but is now sometimes used in linkage analysis (e.g. [12]). A second meaning arises from serially adding linked loci, as in the *variance component linkage method* implemented in the program Sequential Oligogenic Linkage Analysis Routines (SOLAR), which is actually closer to a stepwise analysis procedure, rather than a sequential sampling one (e.g. [2]). But the primary meaning (and the one we discuss here) arises from the classical development of the lod score by Morton [8], in which sequential sampling of independent family units is performed until prespecified levels for or against linkage are reached. A very good overview of the history of sequential designs in genetic linkage studies can be found in Böddeker & Ziegler [3].

Sequential Probability Ratio Test (SPRT) and the Lod Score

The roots of the lod score method, like all sequential ones, go back to Wald's Sequential Probability Ratio Test (SPRT) [13]. Wald developed this theory during the Second World War as a way to minimize measurement costs when testing meant destruction of expensive samples (e.g. firing high-tech proximity-fused antiaircraft ammunition). Morton recognized that an analogy existed in genetic linkage, when genotyping was both limited and expensive, being confined to a small number of blood **markers**, with relatively few **genetic epidemiology** groups working on any given problem. He derived the lod score method for *post hoc* combining of linkage results across investigations which, if invented today, might be termed a **meta-analysis** or "retrospective collaboration" [7]. Each pedigree is added one at a time, followed by formal analyses to see if we continue sampling or have reached a conclusion in favor or against linkage, in which case pedigree sampling stops.

The SPRT approach starts from the recognition that a hypothesis test depends upon four interconnected quantities:

1. type I error (α) = the probability of concluding for H_1 if H_0 is true;
2. type II error (β) = (1 − power) = the probability of concluding for H_0 if H_1 is true;
3. effect size (D) = a measure of how far H_1 is from H_0; and
4. sample size (N).

In traditional (fixed sampling) theory, we hold both α and N constant, and take whatever relationship we find in the data between the other two factors (β, D) [usually, N is chosen to achieve a target (β, D) relationship]. By contrast, the SPRT a priori fixes the first three quantities (α, β and D), allowing N to vary with the experiment. It therefore automatically accounts for sampling fluctuations, as they occur at every stage of the process, requiring just the right amount of additional samples to achieve the target precision in type I and II errors for the target effect size, D. However, the "price" we pay for being able to fix these three factors is that N could theoretically extend indefinitely. Fortunately, sequential theory demonstrates that:

- The sequential process *will* terminate at a finite N with probability "1" (almost surely).
- On average, the N required under sequential sampling will be smaller than that for the "best" fixed sample test that gives the same power [13].

For the SPRT applied to the linkage problem, if θ is the recombination fraction and $f(x_n|\theta)$ is the **likelihood** function of the cumulative pedigree data x_n, up to the nth family unit, then to test the simple hypothesis H_0: $\theta = \frac{1}{2}$ (no linkage) against any simple alternative H_1: $\theta = \theta_1 \neq \frac{1}{2}$ (such as $\theta_1 = 0$, tight linkage, for instance), the SPRT procedure defines the ratio of the probabilities:

$$Z_n = \ln \frac{f(x_n|\theta = \theta_1)}{f\left(x_n|\theta = \frac{1}{2}\right)} \quad \text{for } n = 1, 2, \ldots.$$

By the **maximum likelihood** principle, when data are more compatible with H_1, then this ratio will tend to grow larger with increasing n, while, if the data are more compatible with H_0, then Z_n will tend to become more negative. Before sampling, prespecified

limits are defined, $a^* < 0 < b^*$, which are functions of the target α, and β. Analysis of additional data stops when the first Z_n falls outside these limits, with a decision for the corresponding hypothesis.

For convenience, Morton used the base 10 log instead of the natural log in defining his lod score, but this just applies a scalar multiplier to the SPRT statistic. An important theoretical result is that very simple bounds exist on a^* and b^* which can be used to define slightly wider (more liberal) stopping criteria, but that are much easier to compute, namely

$$A = \frac{1-\beta}{\alpha} \leq a^* \quad \text{and} \quad b^* \leq B = \frac{\beta}{1-\alpha}.$$

If we take type I error $= \alpha = 0.001$ and power $= (1-\beta) = 0.99$, we have $\log(A) = +3$ and $\log(B) = -2$, the values proposed by Morton [8].

Beyond the Lod Score

The SPRT method can be readily extended to both one-sided as well as two-sided compound hypothesis tests in many statistical frameworks (e.g. [5, 11, 15], and [16]). In linkage, sequential adaptations have been extended beyond strongly model-based linkage tests in pedigrees to model-free linkage tests, and to smaller sampling units (e.g. [14]) as well as to joint linkage/association tests of disequilibrium (*see* **Linkage Disequilibrium**) using group sequential designs [6]. The sequential sampling philosophy has also been advocated as an approach to reduce genotyping costs in linkage. Boehnke & Moll [4] demonstrated via simulation that considerable savings can be produced by ranking pedigrees by the evidence for segregation at a locus and sequentially genotyping for linkage accordingly.

Sequential Sampling vs. Sequential Analysis

While sequential testing methods have been widely available for the past 50 years, they are still relegated to a relatively small universe of devoted followers, and are largely ignored by the "fixed sampling" world of investigators. This is partly due to the analytic difficulty of obtaining some of the solutions, and partly to the practical difficulty of conducting a truly sequentially sampled study. Genotyping in small family or relative pair units is not very cost efficient, and the prospect of having quickly to clean, transform, and reanalyze data at every sampling point is daunting. Even the more practical block sequential designs [16] require a higher degree of organization and immediate response to data than most investigators are willing or able to commit, although there is increasing enthusiasm for the promise of such approaches [6]. But perhaps the most unappealing aspect of true sequential sampling is the idea that sample size for a given study should be so completely dictated by the test of any single hypothesis. Usually, there are many phenotypic outcomes and many hypotheses to be tested. This is particularly true in the context of a genome scan, when one may want to use the same family data to search for genes for many phenotypes using a large spanning set of linkage markers. The requirement of true sequential sampling would be prespecifying exactly one primary hypothesis on which to make sampling decisions to the exclusion of all others, and could very easily leave one with an inadequate sample for all other hypotheses of interest. While actual sequential *sampling* may not be very practicable, sequential *analysis* of fixed-sample data is not only practicable but can be quite efficient, at least in the special case of genome-wide linkage or **association** scans. The theory of sequential sampling predicts that the same power can be achieved at a substantially lower average sample number using sequential analysis. Fixed sample advocates usually argue that this "saving" is meaningless if one has already decided to use a fixed sampling scheme, since it is too late to make use of the savings. But in the special case of genome-wide scans, we can gain a lot if we envision a genome scan as a two-step process, an initial hypothesis-generation or "training" phase, in which promising regions are suggested, followed by a confirmatory hypothesis testing phase in which only those few regions are formally tested in independent samples. A special class of sequential procedures, called sequential multiple decision procedures (SMDPs), allow us to formalize and optimize this concept.

SMDPs

SMDPs [1] provide a powerful generalization of the traditional two-hypothesis paradigm to allow U mutually exclusive and exhaustive hypotheses, H_i, where $i = 1, 2, \ldots, U \geq 2$, of which we want to

select one. In the particular case of a genome-wide scan [9, 10], we form every possible subset of markers, (U total) and try to select the one subset that contains only the truly linked (or associated) ones. In traditional hypothesis testing, we have only two hypotheses, H_0 and H_1, two corresponding decisions, D_0 and D_1, and two types of errors, α and β, with relationship for any test procedure, $\wp$:

- $(1-\alpha) = \Pr_{\wp}[\text{make decision } D_0 | H_0 \text{ is true}]$.
- $(1-\beta) = \Pr_{\wp}[\text{make decision } D_1 | H_1 \text{ is true}]$.

But in the SMDP framework, we have U types of errors, α_i, where

- $(1-\alpha_i) = \Pr_{\wp}[\text{make decision } D_i | H_i \text{ is true}]$ for $i = 1, 2, \ldots, U$.

We wish to minimize the probability of any incorrect decision, or conversely maximize the probability of a correct decision (PCD), denoted by P^*, on condition that the "distance" between hypotheses (using an appropriately defined metric) is at least at a certain prespecified "effect size". In the case of a genome-wide scan [9, 10], we compare the *relative* evidence for linkage (or association) among the markers, instead of looking for *absolute* evidence at each marker. As evidence accumulates with each successive data point, the few "signal" markers will eventually separate from the more prevalent background "noise" markers as a distinct subgroup, thus terminating the sequential analysis and identifying the hypothesis to select. For example, using one of the variations on the Haseman–Elston (H–E) method (*see* **Linkage Analysis, Model-free**) on a large number of markers, M, we first *rank* the linkage evidence, $\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[M]}$, where the $[i]$ denotes index of the ith ranked marker, using the (sequential estimate of the) error **variance** from the H–E regression, $\sigma^2_{e[1]} \leq \sigma^2_{e[2]} \leq \cdots \leq \sigma^2_{e[M]}$. Intuitively, this makes sense, as the regressions showing the smallest error variances will be the most significant ones, while the nonsignificant "noise" markers should have error variances nearly equal to the total variance of the response variable. Next we try to divide the markers into two subsets: those highest t showing "nonsignificant" linkage and the lowest $(M-t)$ "significant" ones, by splitting between rank $[M-t]$ and rank $[M-t+1]$. This is equivalent to choosing the one partition that correctly separates the $(M-t)$ truly linked ones from the t unlinked out of all possible ways (U) to select t from M populations, [$U = M!/(t!M-t)!$) hypotheses]. The sufficient statistics for this procedure, $\wp_B$, at sib pair $h+1$, will be the (transformed) sequential sums of squared residuals, using the prediction (H–E) equation for all the *previous* sib pairs. Then, the target effect size, D^*, will be characterized in terms of a minimum "distance" between two adjacently ranked error variances at the critical juncture between $[M-t]$, and $[M-t+1]$, using the distance metric

$$D_{ij} = \left| \frac{1}{2\sigma_i^2} - \frac{1}{2\sigma_j^2} \right|.$$

Using these definitions, the SMDP theory guarantees that:

$$\Pr_{\wp_B}\{\text{correct selection}\} > P^*, \quad \text{whenever } D > D^*.$$

Since it is sequential, the SMDP zeros in on the hit regions with predefined, analyst-specified type I and type II errors, using (on average) a smaller sample than the corresponding fixed sampling test. Also, since it is a single test for all regions simultaneously, questions about the differences between the locus-wise and genome-wise type I and type II errors do not arise (as they do when one conducts multiple fixed sampling marker by marker tests) (*see* **Genome-wide Significance**). Because of these very compelling advantages, the sequential analysis research continues to be of interest as a method to dissect the genetic nature of complex traits.

References

[1] Bechhoffer, R.E., Kiefer, J. & Sobel, M. (1968). *Sequential Identification and Ranking Procedures*. University of Chicago Press, Chicago.

[2] Blangero, J. & Almasy, L. (1997). Multipoint oligogenic linkage analysis of quantitative traits, *Genetic Epidemiology* **14**, 959–964.

[3] Böddeker, I.R. & Ziegler, A. (2001). Sequential designs for genetic epidemiological linkage or association studies: a review of the literature, *Biometrical Journal* **43**, 501–525.

[4] Boehnke, M. & Moll, P.P. (1989). Identifying pedigrees segregating at a major locus for a quantitative trait: an efficient strategy for linkage analysis, *American Journal of Human Genetics* **44**, 216–224.

[5] Ghosh, B.K. (1970). *Sequential Tests of Sequential Hypotheses*. Addison-Wesley, Reading.

[6] Konig, I.R., Schafer, H., Muller, H.H. & Ziegler, A. (2001). Optimized group sequential study designs for

tests of genetic linkage and association in complex diseases, *American Journal of Human Genetics* **69**, 590–600.

[7] Lonjou, C., Barnes, K., Chen, H., Cookson, W.O., Deichmann, K.A., Hall, I.P., Holloway, J.W., Laitinen, T., Palmer, L.J., Wjst, M. & Morton, N.E. (2000). A first trial of retrospective collaboration for positional cloning in complex inheritance: assay of the cytokine region on chromosome 5 by the consortium on asthma genetics (COAG), *Proceedings of the National Academy of Sciences* **97**, 10 942–10 947.

[8] Morton, N.E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.

[9] Province, M.A. (2000). A single, sequential, genome-wide test to simultaneously identify all promising areas in a linkage scan, *Genetic Epidemiology* **19**, 301–322.

[10] Province, M.A. (2001). Sequential methods of analysis for genome scans, in *Genetic Dissection of Complex Traits Advances in Genetics: 42*, D.C. Rao & M.A. Province, eds. Academic Press, San Diego, pp. 499–514.

[11] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.

[12] Thompson, E.A. & Guo, S.W. (1991). Evaluation of likelihood ratios for complex genetic models, *IMA Journal of Mathematical Applications in Medicine and Biology* **8**, 149–169.

[13] Wald, A. (1947). *Sequential Analysis*. Dover, New York.

[14] Weeks, D.E. & Harby, L.D. (1995). The affected-pedigree-member method: power to detect linkage, *Human Heredity* **45**, 13–24.

[15] Wetherill, G.B. (1966). *Sequential Methods in Statistics*. Wiley, New York.

[16] Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials*. Wiley, New York.

M.A. PROVINCE

Sequential Methods for Clinical Trials

For the purpose of this article, a "sequential method" is any approach to the conduct of a clinical trial in which:

(i) there is the potential to perform a series of analyses on the accumulating data at different times during the conduct of the trial;
(ii) each analysis includes a comparison of the treatments featuring in the trial; and
(iii) each analysis has the potential to lead to stopping the trial.

In (i), the word "potential" is included because the very first analysis might lead to stopping, the point being that the number of analyses actually performed is not fixed in advance. Sample size reviews, also known as "internal pilot studies" [10, 25, 26, 69] are excluded from this definition as no treatment comparison is involved. The number of potential analyses may be just two, or may be one after every new patient response. Thus, the definition is intended to encompass the use of a single interim analysis or of group sequential methods (*see* **Data and Safety Monitoring**), as well as earlier approaches such as the sequential probability ratio test.

Part (iii) of the definition excludes purely administrative looks at the data with no potential for stopping. However, it is questionable whether comparisons of the treatments can be made which do not have the potential to lead to stopping.

Each of the series of analyses will be referred to as an "interim analysis". Here only the treatment comparison and its use in deciding whether to stop the trial will be considered, although other calculations might be performed at the same time. Once the trial has stopped, a "final analysis" will be performed in which the significance level of the test of treatment effect will be calculated together with a point estimate and confidence limits for its magnitude.

In most of what follows it will be assumed that a Phase III clinical trial (*see* **Clinical Trials, Overview**) is being conducted to establish whether a single experimental treatment is more efficacious than some control treatment in respect of a single primary endpoint (*see* **Outcome Measures in Clinical Trials**). The extension of sequential methods to other forms of trial will be considered briefly at the end of the article. The exposition will be frequentist throughout: equivalent **Bayesian** procedures have been described in [50].

The Past

Traditional statistical approaches to scientific investigations separate out the phases of design, conduct and analysis. During the design phase, the sample size is fixed, and the method for allocating treatments to experimental units determined. Once the data have been collected, the analysis is conducted. In agricultural applications, which were so influential in the early development of statistical methods, the seasonal nature of farming makes this approach both natural and perfectly satisfactory.

It was in the context of quality control inspections that the above statistical pattern was first broken. Manufactured items are inspected, one at a time, with a view to accepting or rejecting a batch in terms of its quality. Double sampling, in our terms the use of a single interim analysis, was introduced to quality control by Dodge & Romig [16]. The Second World War provided the impetus for the development of the sequential probability ratio test, by Wald [60] in the US and by Barnard [5] in the UK. In this procedure, an "interim analysis" is conducted after the inspection of every individual item.

In quality control, a single sequence of observations is made, modeled as independent, identically distributed random variables from some parametric distribution. The way in which sequential methods could be applied far more generally, by plotting against Fisher's information rather than sample size, was indicated early in the development of the subject by Bartlett [6].

The advantages of sequential methods for clinical trials were also soon noticed. In the medical context the benefits of stopping early can be ethical as well as purely economic. Bross [11] introduced sequential medical plans for comparing two sets of binary responses, and Kilpatrick & Oldham [32] applied the sequential t-test to a comparison of bronchial dilators. The latter design was perhaps too effective: the study was stopped after only four responses. By 1960 there was already enough accumulated theory and practice to guarantee an audience for the first edition of Armitage's book on *Sequential Medical Trials* [3].

The principal limitation of the early theory of sequential analysis for implementation in clinical trials was the need to perform an interim analysis after every patient response. To be more precise, a plot of a comparative sample statistic against sample size had to be approximated as a continuous Brownian motion, and its final position after crossing a stopping boundary had to be treated as being on the boundary. This "no overshoot" assumption could be justified if interim analyses were conducted after every response, or in larger trials, at frequent intervals. Even after decades of talk of immediate on-line computer data entry, the reality of such continuous data monitoring in clinical trials appears to be as far away as ever.

The no overshoot assumption was eventually overcome in two differing ways, one involving computing power, the other relying on mathematical sophistication. The method of recursive numerical integration introduced by Armitage et al. [4] utilizes the independent increments of a Brownian motion to compare the distribution of the final position of a sequential sample path, given a series of upper and lower stopping limits. This allows a more precise evaluation of the properties of designs such as the sequential probability ratio test applied with infrequent looks as well as the specification of alternative designs based on different criteria. The method requires considerable computing input, which even today can be prohibitive if there are to be very large numbers of interim analyses.

The stopping criteria, which became popular during the 1970s, involved considering each interim analysis as a miniature fixed sample analysis, conducted with a type I error rate referred to as a "nominal significance level". It was well known that for the procedure as a whole to comply with an overall type I error rate of (say) 0.05, each of these nominal significance levels had to be less than 0.05. Various schemes were devised for setting the nominal significance levels. Pocock [42] suggested constant values, and then in 1982 recommended certain varying sequences [43]. O'Brien & Fleming [38] considered an increasing sequence of levels for which early stopping was extremely unlikely, so that the final nominal level would be close to 0.05 (or some other chosen overall value). These methods were referred to as "group sequential methods", a title which unfortunately has led to a false distinction from the wider family of sequential methods of which they are a part.

The mathematical answer to the overshoot problem was developed from renewal theory by Siegmund [47]. Earlier theory assumed that the final position of the sample path was on the boundary. Siegmund evaluated the expected overshoot of the boundary. This allows the boundary to be moved inwards by an amount equal to the expected overshoot. Now the expected final position of the sample path is on the original boundary. The boundary is changed, and the "no overshoot" theory is left intact and now more accurate. Whitehead & Stratton [68] and Whitehead [65] explain how this device can be used to allow "group sequential methods" to be based on straight-line boundaries such as the sequential probability ratio test and its modifications due to Anderson [1]. The resulting "Christmas tree correction" is extremely accurate in the case of the triangular test [51], but it can be less successful for other straight-line boundaries if interim analyses are few.

The Christmas tree correction offers complete flexibility over the frequency and timing of interim analyses while guaranteeing that when they occur they are conducted according to preordained rules. By contrast, the early "group sequential methods" relied on tabulations of critical values which were accurate only if a prespecified schedule of looks was followed. The α-spending function method of Lan & DeMets [34] brought that same flexibility into "group sequential methods".

The timing of the interim analysis no longer had to be prespecified or, more importantly, the amount of information available at each interim did not have to be anticipated. Instead, when an interim analysis was conducted, the amount of information available could be calculated, and the stopping criteria deduced in order to ensure that the total null probability of stopping up to and including the current look achieved some desired value, $\alpha(t)$. Here t is the ratio of the information accumulated to date and the maximum possible amount of information. The function $\alpha(t)$ is defined in advance for all t, and is called the α-spending function. When $t = 1$ (information is at its maximum value), $\alpha(1)$ is equal to the overall type I error rate (perhaps 0.05). Various α-spending functions have been devised, by Lan & DeMets [34], Kim & DeMets [33] and Hwang et al. [28] amongst others. Some are similar in spirit to older rules, such as those of Pocock and O'Brien and Fleming.

During the 1980s research attention focused on how to analyze a sequential clinical trial once it had stopped. Unfortunately, some of the language used to describe sequential methods obscured the main task. Interim analyses are not interpretations of the data in the way that conventional statistical analyses are; they serve only to determine whether the trial should stop. The term "nominal significance level" is potentially misleading, as it relates to the significance level appropriate to a design which was not in fact used. "Adjusting P values" or "paying a penalty on α" suggest rather *ad hoc* procedures, whereas appropriate analysis methods can be quite precise.

The monitoring of a sequential clinical trial is in fact best viewed as part of the design phase rather than being part of the analysis. It is a flexible **sample size computation** that avoids unnecessary sampling while guaranteeing the desired power. Once this sample size has been reached, and the trial has been stopped, the analysis phase begins. Frequentist analyses require consideration of those study outcomes *supporting the alternative hypothesis as strongly or more strongly than was observed*. For ease of interpretation it is best to restrict attention to one-sided alternatives of the form "experimental treatment is better than control", and to perform the appropriate multiplication by 2 to obtain two-sided P values. In a fixed sample study, evidence concerning treatment difference is usually expressed in terms of a one-dimensional test statistic, and the phrase above in italics is taken to refer to test statistics as large or larger than that observed. The frequentist analysis of a sequential design necessitates contemplation of repeated runs of the same design, and identification of which outcomes support experimental superiority more than others.

Armitage [2] and Siegmund [46] addressed the ordering of outcomes by degree of support for the alternative hypothesis in the case of continuous monitoring. In that situation the sample path will end on the boundary, and an anticlockwise ordering, from sample paths plunging down to the lower boundary (least support for superiority), via lengthy horizontal sample paths, through to sample paths shooting up to the upper boundary (most support), is natural. This leads to P values being defined for trials ending positively on the upper boundary, as the null probability of earlier stopping on the upper boundary.

Although the anticlockwise ordering will serve to provide a good approximation when interim analyses are frequent, it is incomplete in the more common case in which only a few looks at the data are planned. Here, a two-dimensional set of possible outcomes comprising the value of the final test statistic and the identity of the look at which it was observed have somehow to be resolved into a single ordering. The original, and most successful form of ordering in this case was introduced by Fairbanks & Madsen [20] and explored further by Tsiatis et al. [59]. For trials stopping on the upper boundary, in favor of the experimental treatment, the earlier the stopping, the stronger the support for superiority. Outcomes with the same final look are ordered in terms of the value of the test statistic. On the lower boundary, the later the final look, the better the support for experimental superiority (or at least, the less bad). This retains an anticlockwise element, in common with the ordering of Armitage and Siegmund.

The ordering of Fairbanks & Madsen [20] is *truncation adaptable*, to use a phrase later introduced by Liu & Hall [35]. Computations of analyses are made conditionally on how much information turned out to be available at each look, up to the final one. However, they require no knowledge of whether, when and how any future interim analyses would have been conducted, had the trial not been stopped. This means that analyses based on two designs that share the same criteria for (say) the first three looks, and then diverge, will be identical if stopping occurs at one of the first three looks. In particular, if a sequential trial is stopped at the first look, then the Fairbanks & Madsen [20] ordering leads to an analysis which is identical to the conventional fixed sample analysis. This is not to be confused with the type I error rates of the two procedures which, being properties of the whole design, are definitely different.

Once an ordering of possible outcomes has been identified, a full frequentist analysis becomes possible. Let θ denote the advantage of the experimental treatment over control. Define the P value function $P(\theta)$ as the probability of obtaining evidence supporting experimental superiority as strongly or more strongly than observed, according to the ordering, when the treatment advantage is θ. Then the P value against the one-sided alternative of experimental superiority will be $P(0)$, and against the one-sided alternative of experimental inferiority it will be $1 - P(0)$. Taking the smaller of these two and doubling it will give the P value against the two-sided alternative. A 95% confidence interval for

θ is given by (θ_L, θ_U), where $P(\theta_L) = 0.025$ and $P(\theta_U) = 0.975$, and a median unbiased estimator of θ by θ_M, where $P(\theta_M) = 0.05$

Other orderings of the sample space have been suggested by Rosner & Tsiatis [45], Chang [12] and Emerson & Fleming [18]. The last of these, for example, consists of ordering by the magnitude of the final maximum likelihood estimate $\hat{\theta}$ of θ. These orderings are not truncation adaptable. For example, if stopping occurs at the second look, it might have been possible for a later look to yield a higher value of $\hat{\theta}$. Thus, computations of P values and estimates have to take into account what might have happened later. Worse still, certain of the orderings of Rosner & Tsiatis [45] lead to confidence regions which are not intervals, but instead consist of disjointed separate intervals. Such flaws would appear to render them unsuitable for practical use.

Alternative methods for post-trial analysis include bias-adjusted maximum likelihood estimates [62] and Woodroofe confidence intervals [58, 70]. These methods modify certain conventional analysis procedures (maximum likelihood estimation and pivot-based confidence intervals) for use after a sequential trial. The former estimates are, to quite an accurate extent, unbiased in the conventional expectation sense. These methods also share the need to integrate over all possible final outcomes in order to compare expectations of various combinations of test statistic and information measure. Unfortunately, this need prevents them from being truncation adaptable. In practice, various scenarios for the inspection schedule that would have taken place after stopping can be imposed, and they make little difference to the numerical results. All the same, the mere need to speculate about what might have been is a drawback to these methods.

Emerson [17] has devised a truncation adaptable method of computing an unbiased estimate of treatment effect using the method, due to Rao [44] and Blackwell [8], of taking the expected value of a simple unbiased estimate conditional on a sufficient statistic. The simple unbiased estimate used is the maximum likelihood estimate of θ computed at the first look, and the sufficient statistic is the bivariate combination of the test statistic and the information measure at termination. The mathematics of the method were explored by Ferebee [21], and the resulting estimate has been shown to give the uniformly minimum variance unbiased estimate within the class of truncation adaptable estimates by Liu & Hall [35].

Sequential methods are now widely used in practice, especially in large-scale studies in serious and life-threatening diseases. The book edited by Peace [41] provides a collection of case studies, and Whitehead [67] provides references to several more.

The Present

Investigators planning a clinical trial today have a wide range of sequential methods available to choose from, with **software** packages such as PEST 4 [37], EaSt 2000 [15] and the S-PLUS module S + SeqTrial [36] to facilitate calculations. If the trial is to be a comparison of two treatments in respect of a single primary endpoint, with the objective of discovering whether one treatment is superior to the other, then it is extremely likely that a suitable method already exists. This means that infeasibility or unfamiliarity are no longer excuses for avoiding interim analyses and stopping rules in such trials, when ethical or economic purposes would be served by them.

Sample size determination in any trial begins with specification of a power requirement. If a certain treatment advantage ($\theta = \theta_R$) is present, then significance at level α should be achieved with power $1 - \beta$. In a fixed sample study there will inevitably be an interplay between resource limitations and the setting of θ_R and $1 - \beta$, although it is unwise to proceed with a sample size that is underpowered for credible and worthwhile treatment advantages.

Sequential designs offer more scope, first in the power requirements available, and second in how these are to be attained. In a fixed sample design, assuming a suitable measure θ of treatment difference, its power is set at $1 - \beta$ for $\theta = \theta_R$; it will also be $1 - \beta$ for $\theta = -\theta_R$. This need not be so in sequential studies. It is possible to set the power to be $1 - \beta$ for $\theta = \theta_R$ while accepting a much lower power at $\theta = -\theta_R$. This is entirely appropriate if the investigator has no need to distinguish the experimental treatment as being inferior to the control from the case of no effect. If the experimental treatment is novel and expensive, then either of these situations will lead to its development being abandoned, and so it would not be proper to recruit patients merely to determine which is true. This sort of specification is called

Power Requirement I by Whitehead [65], and it leads to asymmetric designs such as the triangular test. The opposite situation of requiring a power of $1 - \beta$ for $\theta = -\theta_R$, but not needing to distinguish between superiority and **equivalence**, arises in a noninferiority comparison of a cheaper and safer alternative with an established active control. (Sequential equivalence and noninferiority designs are discussed by Whitehead [64].) Sequential methods can be devised, of course, for the same symmetric power requirement as the fixed sample design (known as Power Requirement II).

Having chosen a power requirement, the shape of the plot of the expected terminal sample size against θ should be considered. This is known as the average sample number (ASN), within quality control applications. For a fixed sample size trial the ASN is a horizontal line. For a triangular test, it rises from low values over $\theta < 0$, to a maximum around $0.5\ \theta_R$ and then falls again. A truncated sequential probability ratio test has lower expected sample sizes for $\theta < 0$ and $\theta > \theta_R$ than the triangular test, but rises to a higher maximum in between. This then is the design question: For the power requirement set, for what values of θ should the expected sample size be large, and for what values should it be small?

Generally speaking, sample sizes should be small when θ is distant from zero, as in those situations when one group of patients is being seriously disadvantaged relative to the other. For designs constructed from Power Requirement I, the ASN will be asymmetric, with low values whenever $\theta < 0$. When Power Requirement II is specified, the main choice concerns whether expected sample sizes should be small for $\theta = 0$. They can be made so by choosing a design such as the double triangular test in PEST or opting for "Early rejection of H_0" in EaSt. Such designs will generally reduce sample size, and may be suitable for establishing equivalence. In some trials it is desirable to allow larger sample sizes when $\theta = 0$. There are seldom ethical concerns about continuation under H_0, and the larger sample size allows scope for eventual investigation of secondary endpoints and subgroup effects (*see* **Treatment-covariate Interaction**). This option may also provide insurance against concerns over model fit. Designs such as the restricted procedure within PEST or methods for two-sided alternatives and without early rejection of H_0 in EaSt satisfy these objectives. Within the desired class of procedures, fine tuning can be achieved by varying parameters such as the slope in PEST or Δ in EaSt and investigating the effect on the ASN. Other properties of final sample size, such as the median or 90th percentile, can also be used to facilitate choice.

The conduct of a sequential clinical trial is often within the context of a **Data and Safety Monitoring Board** (DSMB). The timings of the interim analyses are usually fixed in advance, and the plots of the test statistic against information measure form part of the report to the DSMB. As far as the mathematical model of the trial is concerned, it is the prespecified sequential plan that controls the stopping of the trial; there is no deviation from the prespecified protocol. The analysis computations developed over the years all rely on this being true, and any lack of adherence to plan will lead to inaccuracy in the computation of significance levels, point estimates and confidence limits. However, this is a clinical trial involving human subjects, and so the mathematical model of the trial cannot be the whole story. The DSMB will be presented with data other than the formal sequential plot, and these data may lead to stopping the trial earlier than the formal sequential rule or continuing it for longer. The mathematical modeling may involve formal interim analyses starting some time after the trial began, becoming more frequent thereafter. The DSMB, however, may receive safety data that lead to stopping even before the first formal interim analysis has taken place. Part of the art of designing sequential studies is to make the formal rule as close as possible to the likely actions of the DSMB. This involves careful choice of the primary efficacy endpoint, a detailed presentation of the sequential plan to the DSMB before the trial begins and, if necessary, pretrial revision of the plan to make it fit more closely to the DSMB's view of safety issues. An account of the role of the statistician in the DSMB, with particular reference to sequential studies, is given by Whitehead [66].

The most likely direction of discrepancy between a formal sequential plan and the actions of the DSMB is towards stopping for safety. The Board may see problems sooner than they are picked up by the sequential analysis, or concerning patient outcomes that are not part of the formal procedure. It is far less likely that the Board will wish to stop the study earlier than the plan allows in order to claim increased efficacy. This is because they know that a positive finding has to be accepted as valid, either by drug regulatory authorities (*see* **Drug Approval**

and Regulation) or by clinical opinion, before the new treatment receives widespread use. Departure from the plan might weaken the authority of the trial findings and result in delay in their acceptance. Having set in place a formal plan which does react to early strong positive evidence, it is most likely that the DSMB will respect it. The DSMB is also likely to respect rules governing stopping because the null hypothesis appears to be true: sometimes this is called stopping for futility. As scientists, they may wish to see the accrual of large and informative data sets, but this is not a safety issue, and in this instance they are acting to conserve the sponsor's resources.

The only result of unplanned stopping for safety can be the negative one of not claiming advantage and so its only consequence for trials seeking to demonstrate superior efficacy which do stop positively is to cause their P values to be conservative and their point estimates to be biased downwards. This is because, potentially, true claims of efficacy might be lost due to these "unofficial" stopping rules. The effect of such unofficial stopping does, however, have more serious consequences for equivalence and noninferiority trials. It is possible to formulate additional safety monitoring rules to be used in addition to efficacy sequential plans or within a fixed sample trial; for examples see Bolland & Whitehead [9].

The final analysis, including computation of the P value, point estimate and confidence interval, must be done in a way that is consistent with the sequential designs used. This need is now being recognized by regulators [19]. Sometimes a design has been chosen deliberately to ensure that a conventional analysis is essentially valid, and an argument to that effect may be acceptable. However, a numerical demonstration of the adequacy of the conventional analysis removes any residual doubt, and it is unwise anyway to constrain the choice of design in order to avoid the need to conduct an appropriate analysis.

Between the termination of the study and the final analysis, extra data may become available that did not feature in the last interim analysis. These data may be the result of inevitable delays in reporting, or they might be responses taken after some weeks or months of follow-up that could not have been available any earlier. Provided that these observations are collected under **protocol** conditions they should form part of the final analysis. The inclusion of such results in an "overrunning analysis" is discussed by Whitehead [63].

The Future

Current methodological research in sequential analysis is being directed towards extending its utility beyond comparisons of two treatments in respect of a single endpoint.

The multiple treatment problem has long been of interest, with the general methodology going back to the elimination procedures of Paulson [39, 40]. Sequential χ^2 and F tests have been developed by Siegmund [48] and Jennison & Turnbull [29]. These procedures are of limited utility in clinical applications. Elimination procedures allow treatments to be dropped at each of a series of interim analyses, and aim to select the best. However, although the treatment remaining at the end is indeed likely to be the best, evidence of significant superiority over any competitor is not guaranteed. Sequential χ^2 and F tests stop when it is evident that the treatments under comparison differ, but this may not mean that any individual treatment has yet demonstrated superiority over others.

More recent work has concerned allowing inferior treatments to be eliminated at interim analyses while preserving the error rates associated with recommending effective treatments. Follmann et al. [23] present a general procedure based on pairwise comparisons of k treatments with conservative preservation of error rates. Thall et al. [55, 56] present two-stage procedures for binary outcomes in which one of several experimental treatments is selected at the end of the first stage, and its comparison with a control is completed during the second. Data from both stages feature in the final analysis, and the overall type I error is controlled. Stallard & Todd [52, 53] take a similar approach, considering general responses rather than just the binary case, incorporating extra looks into the second pairwise comparison stage, and allowing the selection to be based on a **surrogate** rather than the primary response.

Methods that allow for the simultaneous monitoring of more than one endpoint are also being developed. Jennison & Turnbull [30], Cook & Farewell [13] and Thall & Cheng [54] have devised procedures in which both efficacy and safety endpoints are considered at each interim analysis. Todd [57] has looked at more general bivariate procedures, considering cases in which superiority has to be demonstrated in each of two efficacy endpoints

and also cases in which superiority in either will be sufficient.

A special case of multiple endpoints arises when subjects are assessed repeatedly in terms of the same response during the course of follow-up. The inclusion of such longitudinal data in interim analyses has been considered by Gange & DeMets [24] and Cook & Lawless [14] amongst others. In each case it is necessary to make modeling assumptions that allow the predominantly early follow-up data available at the first one or two interim analyses to be used to decide whether to stop a study intended to investigate longer-term follow-up. This dependence on extrapolated model assumptions is shared with long follow-up survival studies, in which context it is discussed by Gregory et al. [27] and Sooriyarachchi & Whitehead [49].

The theoretical developments cited above are ongoing, and have as yet received little implementation. However, within the pharmaceutical industry a desire to reduce drug development time and a need to satisfy regulatory conditions concerning multiple aspects of treatment safety and effectiveness are likely to lead to practical exploitation of this work. The selection procedures can be viewed as ways of combining Phases II and III, and the multiple endpoint designs may ensure that stopping does not occur until all trial objectives have been realized. On-line model checking is an important issue, especially for long-term survival or longitudinal studies, and the new methods may be able to address the problems of making repeated goodness-of-fit assessments alongside the repeated treatment comparisons.

Another current area of research concerns "adaptive designs". These allow greater freedom of action as a result of the findings of interim analyses. Examples include the methods of Bauer & Köhne [7] and Wassmer [61] as well as the "self-designing clinical trials" of Fisher [22]. Such approaches have the potential to introduce a great amount of freedom into the conduct of clinical trials, allowing them to follow more closely the instinctive learning processes of scientific enquiry, without losing control over error rates. However, two dangers are present. The first is that the new found freedom leads to a progressive reduction of the targeted "clinically relevant difference" as a trial reveals less and less promise of a new treatment. This can result in the greatest resources being devoted to the least important treatments, being the counterpart of scientific desperation rather than self-critical enquiry. The second danger lies in the conflict between the need to limit access to the results of interim analyses and the desire to act upon them. Confidentiality of interim results is a necessary precaution to avoid operational bias; that is, the conduct of the trial being a result of its findings rather than the other way round. If actions affecting the trial or the development of the drug involved are taken as a result of interim findings, then inferences about the nature of those findings will be drawn. It may be that mathematical preservation of error rates is the easy part of developing adaptive procedures.

The future of sequential methods in clinical trials appears to be assured. The recent appearance of two up-to-date texts on the subject [31, 65] and the simultaneous release of three new software packages for their implementation [15, 36, 37] ensures that the procedures developed in the past can be routinely and accurately implemented. The flourishing interest in methodological research will serve to ensure that the remit of sequential methods continues to widen.

References

[1] Anderson, T.W. (1960). A modification of the sequential probability ratio test to reduce sample size, *Annals of Mathematical Statistics* **31**, 165–197.

[2] Armitage, P. (1957). Restricted sequential procedures, *Biometrika* **44**, 9–26.

[3] Armitage, P. (1960). *Sequential Medical Trials*, 1st Ed. Blackwell, Oxford.

[4] Armitage, P., McPherson, C.K. & Rowe, B.C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

[5] Barnard, G.A. (1946). Sequential tests in industrial statistics, *Journal of the Royal Statistical Society*, Supplement **8**, 1–26.

[6] Bartlett, M.S. (1946). The large sample theory of sequential tests, *Proceedings of the Cambridge Philosophical Society* **42**, 239–244.

[7] Bauer, P. & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses, *Biometrics* **50**, 1029–1041.

[8] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation, *Annals of Mathematical Statistics* **18**, 105–110.

[9] Bolland, K. & Whitehead, J. (2000). Formal approaches to safety monitoring of clinical trials in life-threatening conditions, *Statistics in Medicine* **19**, 2899–2917.

[10] Bolland, K., Sooriyarachchi, M.R. & Whitehead, J. (1998). Sample size review in a head injury trial with ordered categorical responses, *Statistics in Medicine* **17**, 2835–2847.

[11] Bross, I. (1952). Sequential medical plans, *Biometrics* **8**, 188–205.

[12] Chang, M.N. (1989). Confidence intervals for a normal mean following a group sequential test, *Biometrics* **45**, 247–254.

[13] Cook, R.J. & Farewell, V.T. (1994). Guidelines for monitoring efficacy and toxicity responses in clinical trials, *Biometrics* **50**, 1146–1152.

[14] Cook, R.J. & Lawless, J.F. (1996). Interim monitoring of longitudinal comparative studies with recurrent event responses, *Biometrics* **52**, 1311–1323.

[15] Cytel Software Corporation (2000). *EaSt 2000: A Software Package for the Design and Interim Monitoring of Group-Sequential Clinical Trials*. Cytel, Cambridge.

[16] Dodge, H.F. & Romig, H.G. (1926). A method of sampling inspection, *The Bell System Technical Journal* **8**, 613–631.

[17] Emerson, S.S. (1993). Computation of the uniform minimum variance unbiased estimator of a normal mean following a group sequential trial, *Computers and Biomedical Research* **26**, 68–73.

[18] Emerson, S.S. & Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing, *Biometrika* **77**, 875–892.

[19] Facey, K.M. & Lewis, J.A. (1998). The management of interim analyses in drug development, *Statistics in Medicine* **17**, 1801–1809.

[20] Fairbanks, K. & Madsen, R. (1982). P values for tests using a repeated significance test design, *Biometrika* **69**, 69–74.

[21] Ferebee, B. (1983). An unbiased estimator for the drift of a stopped Wiener process, *Journal of Applied Probability* **20**, 94–102.

[22] Fisher L.D. (1998). Self-designing clinical trials, *Statistics in Medicine* **17**, 1551–1562.

[23] Follmann, D.A., Proschan, M.A. & Geller, N.L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials, *Biometrics* **50**, 325–336.

[24] Gange, S.J. & DeMets, D.L. (1996). Sequential monitoring of clinical trials with correlated responses, *Biometrika* **83**, 157–167.

[25] Gould, A.L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate, *Statistics in Medicine* **11**, 55–66.

[26] Gould, A.L. (1995). Planning and revising the sample size for a trial, *Statistics in Medicine* **14**, 1039–1051.

[27] Gregory, W.M., Bolland, K., Whitehead, J. & Souhami, R.L. (1997). Cautionary tales of survival analysis: conflicting analyses from a clinical trial in breast cancer, *British Journal of Cancer* **76**, 551–558.

[28] Hwang, I.K., Shih, W.J. & DeCani, J.S. (1990). Group sequential designs using a family of type I error probability spending functions, *Statistics in Medicine* **9**, 1439–1445.

[29] Jennison, C. & Turnbull, B.W. (1991). Exact calculations for sequential t, χ^2 and F tests, *Biometrika* **78**, 133–141.

[30] Jennison, C. & Turnbull, B.W. (1993). Group sequential tests for bivariate response: interim analysis of clinical trials with both efficacy and safety endpoints, *Biometrics* **49**, 741–752.

[31] Jennison, C. & Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton.

[32] Kilpatrick, G.S. & Oldham, P.D. (1954). Calcium chloride and adrenaline as bronchial dilators compared by sequential analysis, *British Medical Journal* **ii**, 1388–1391.

[33] Kim, K. & DeMets, D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function, *Biometrika* **74**, 149–154.

[34] Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika* **70**, 659–663.

[35] Liu, A. & Hall W.J. (1999). Unbiased estimation following a group sequential test, *Biometrika* **86**, 71–78.

[36] Mathsoft Inc. (2000). *S-Plus 2000*. MathSoft, Seattle.

[37] MPS Research Unit (2000). *PEST 4: Operating Manual*. The University of Reading.

[38] O'Brien, P.C. & Fleming. T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[39] Paulson, E. (1962). A sequential procedure for comparing several experimental categories with a standard or control, *Annals of Mathematical Statistics* **33**, 438–443.

[40] Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations, *Annals of Mathematical Statistics* **35**, 174–180.

[41] Peace, K.E., ed. (1992). *Biopharmaceutical Sequential Statistical Applications*. Dekker, New York.

[42] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.

[43] Pocock, S.J. (1982). Interim analysis for randomized clinical trials: the group sequential approach, *Biometrics* **38**, 153–162.

[44] Rao, C.R. (1945). Information and accuracy attainable in estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society* **37**, 81–91.

[45] Rosner, G.L. & Tsiatis, A.A. (1988). Exact confidence limits following group sequential tests, *Biometrika* **75**, 723–729.

[46] Siegmund, D. (1978). Estimation following sequential tests, *Biometrika* **65**, 341–349.

[47] Siegmund, D. (1979). Corrected diffusion approximations in certain random walk problems, *Advances in Applied Probability* **11**, 701–719.

[48] Siegmund, D. (1980). Sequential χ^2 and F tests and the related confidence intervals, *Biometrika* **67**, 387–402.

[49] Sooriyarachchi, M.R. & Whitehead, J. (1998). The sequential analysis of survival data with non-proportional hazards, *Biometrics* **54**, 1072–1084.

[50] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society, Series A* **157**, 357–416.

[51] Stallard, N. & Facey, K.M. (1996). Comparison of the spending function method and the Christmas tree correction for group sequential trials, *Journal of Biopharmaceutical Statistics* **6**, 361–373.

[52] Stallard, N. & Todd, S. (1999). Sequential designs for phase III clinical trials incorporating treatment selection. *Technical Report 99/1*, Department of Applied Statistics, University of Reading.

[53] Stallard, N. & Todd, S. (1999). Calculations for sequential designs incorporating treatment selection using spending functions. *Technical Report 99/10*, Department of Applied Statistics, University of Reading.

[54] Thall, P.F. & Cheng, S.C. (1999). Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials, *Biometrics* **55**, 746–753.

[55] Thall, P.F., Simon, R. & Ellenberg, S.S. (1988). Two-stage selection and testing designs for comparative clinical trials, *Biometrika* **75**, 303–310.

[56] Thall, P.F., Simon, R. & Ellenberg, S.S. (1989). A two-stage design for choosing among several experimental treatments and a control in clinical trials, *Biometrics* **45**, 537–547.

[57] Todd, S. (1999). Sequential designs for monitoring two endpoints in a clinical trial, *Drug Information Journal* **33**, 417–426.

[58] Todd, S., Whitehead, J. & Facey, K.M. (1996). Point and interval estimation following a sequential trial, *Biometrika* **83**, 453–461.

[59] Tsiatis, A.A., Rosner, G.L. & Mehta, C.R. (1984). Exact confidence intervals following a group sequential test, *Biometrics* **40**, 797–803.

[60] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

[61] Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials, *Biometrics* **54**, 696–705.

[62] Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test, *Biometrika* **73**, 573–581.

[63] Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials, *Controlled Clinical Trials* **13**, 106–121.

[64] Whitehead, J. (1996). Sequential designs for equivalence studies, *Statistics in Medicine* **15**, 2703–2715.

[65] Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, revised 2nd Ed. Wiley, Chichester.

[66] Whitehead, J. (1999). On being the statistician on a data and safety monitoring board, *Statistics in Medicine* **18**, 3424–3434.

[67] Whitehead, J. (2001). Use of the triangular test in sequential clinical trials, in *Handbook of Statistics in Clinical Oncology*, J. Crowley, ed. Dekker, New York.

[68] Whitehead, J. & Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions, *Biometrics* **39**, 227–236.

[69] Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials, *Statistics in Medicine* **9**, 65–72.

[70] Woodroofe, M. (1992). Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test, *Biometrika* **79**, 347–353.

JOHN WHITEHEAD

Serial Correlation

Many biostatistical investigations involve longitudinal data in which subjects, possibly in naturally occurring or experimentally determined groups, are observed on several different occasions over some particular period of time (*see* **Longitudinal Data Analysis, Overview**). A characteristic of this type of data is a correlation between pairs of measurements on the same subject, the magnitude of which usually depends on the time separation of the measurements – typically, the correlation becomes weaker as the time separation increases. This *serial correlation* needs to be properly accounted for in the analysis of such data if appropriate inferences are to be made.

(*See also* **Analysis of Variance for Longitudinal Data**; **Durbin–Watson Test**; **Generalized Linear Models for Longitudinal Data**)

BRIAN S. EVERITT

Serial Dilution Assay

Serial dilution assays were originally developed to estimate the concentration of viable bacteria in liquid suspension, from observations of the presence or absence of the organism in samples at different dilutions (*see* **Dilution Method for Bacterial Density Estimation**). Let λ denote the concentration of bacteria per unit volume; this is the parameter of interest. A key assumption is that the bacteria are distributed at random throughout the suspension, with no tendency to aggregate. This implies that the number of bacteria in a sample of volume v follows a **Poisson distribution** with mean λv. In particular, the probability that the sample contains one or more bacteria is $1 - \exp(-\lambda v)$.

The original suspension is diluted m times and, at the ith dilution, n_i samples are taken from the diluted suspension, each sample containing a volume v_i of the original suspension. The samples are deposited in tubes containing a suitable culture medium. Following incubation, Y_i, the number of samples that show evidence of bacterial growth, is recorded. The second key assumption is that observable growth will result from any sample that contains at least one viable organism, implying that the data indicate reliably the presence or absence of bacteria in each sample.

The probability of observing bacterial growth at the ith dilution is therefore

$$\pi_i = 1 - \exp(-\lambda v_i) \tag{1}$$

and, provided that samples can be assumed to be independent, Y_i will have the **binomial distribution** $\text{bin}(n_i, \pi_i)$. This model forms the basis of the statistical analysis.

Bacterial concentrations can also be estimated by culturing samples on petri dishes and counting the number of colonies that grow. Colony counting clearly gives more precision per sample than simply recording presence or absence of the organism in the sample. However, the gain in precision per sample must be balanced against the increased time needed for counting. Moreover, if the sample is not sufficiently diluted, then the bacterial colonies will coalesce and counting becomes impossible.

Serial dilution assays are used routinely to estimate concentrations of micro-organisms in, for example, foods, water, and soils. Experimental procedures differ in detail, but the key assumptions remain that the number of particles in a sample has a Poisson distribution and that the observed positive or negative responses indicate reliably the presence or absence of the organism in the sample.

Dilution assays are also used to estimate the proportion of individuals in a population that have some characteristic, when it is not feasible, for economic or other reasons, to test individuals and estimate the proportion directly. Such assays are often called *limiting dilution assays*. For example, in immunology, limiting dilution assays are used to estimate the proportion of cells, θ say, that are immunocompetent. The procedure is to test *random* samples consisting of different numbers of cells to ascertain whether each sample does or does not contain immunocompetent cells. As in bacterial estimation, it is assumed that the presence/absence test is completely reliable. It follows that the probability that a sample of r_i cells contains at least one immunocompetent cell is

$$\pi_i = 1 - (1 - \theta)^{r_i}, \tag{2}$$

which is equivalent to (1) with $v_i = r_i$ and $\lambda = -\log(1 - \theta)$. In practice, however, at least in immunology, limiting dilution assays are often analyzed on the basis of (1) with $v_i = r_i$ and $\lambda = \theta$. There are two slightly different justifications for this. First, if θ is small, as it often is in immunologic applications, then $-\log(1 - \theta) \approx \theta$. More importantly, in immunologic applications, the number of cells tested, r_i, is not usually known exactly. Instead, the number is assumed to be a Poisson variable with mean r_i. The unconditional probability that the sample contains immunocompetent cells is then given exactly by (1) with $v_i = r_i$ and $\lambda = \theta$.

This method of estimating a proportion is also known as *group testing* [6], although this term is also used to describe a method of *repeatedly* testing individuals in groups to determine precisely which individuals have the attribute of interest [12].

Point Estimation of λ

Since the observations Y_i are binomially distributed, the log-likelihood for the data, ignoring some terms that do not depend on λ, is

$$l(\lambda) = \sum_{i=1}^{m} Y_i \log[1 - \exp(-\lambda v_i)] - (n_i - Y_i)\lambda v_i.$$

From this one can obtain the **maximum likelihood** estimator of λ, $\hat{\lambda}$ say, which is often known as the *most probable number* (MPN), a terminology introduced by McCrady [24]. Except in special cases, iterative methods, such as the Newton–Raphson method (*see* **Optimization and Nonlinear Equations**), are needed for the calculation of $\hat{\lambda}$, and various special-purpose computer programs are available, e.g. [22]. Alternatively, the maximum likelihood estimator of $\log\lambda$ can be obtained as the constant term in a binomial **generalized linear model** with complementary–log–log link function and with $\log(v_i)$ as an offset variable [25, Section 1.2.4]. Thus, estimation is possible in any package that caters for generalized linear models, though ill-fitting data sets can cause problems occasionally in packages that estimate parameters by Fisher's method of scoring [29].

When all samples are positive, $\hat{\lambda} = \infty$ and the best that can be done is to estimate a lower confidence limit for λ. Sometimes this may be sufficient, e.g. in food safety testing where it is sufficient to know that the bacterium is present in large numbers. Alternatively, if it is important to have a more precise estimate of concentration, then it will be necessary to do further dilutions until some samples give a negative result. Conversely, when all samples are negative, $\hat{\lambda} = 0$. In all other circumstances the likelihood function is unimodal with a finite, nonzero maximum.

Fisher [15] used the serial dilution assay as one of the examples in his original paper on maximum likelihood estimation. However, for practical application he suggested a computationally simpler method in which the observed number of negative results is equated to the expected number. He showed that this method, fully described by Fisher & Yates [17], has an **asymptotic relative efficiency** of 88% compared with maximum likelihood estimation, although Best & Raynor [3] found the estimators to be very similar for sample sizes typically used in practice. Historically, various other methods of estimation have been proposed, but maximum likelihood estimation remains the method most commonly used in practice.

Bias Correction

The maximum likelihood estimator of λ is, however, positively biased, i.e. it tends to *overestimate* the true value of λ (*see* **Unbiasedness**). The magnitude of the bias depends on the value of λ, on the dilutions used, and on the number of replicates at the different dilutions, but can exceed 10% in small assays.

Several authors have therefore considered bias corrections. Gart [18] derives a bias-corrected estimator as a particular instance of the general asymptotic theory given in [8, Section 9.2]. Mehrabi & Matthews [26] note that this is equivalent to an estimator derived, using a different method, by Salama et al. [32]. Mehrabi & Matthews [26] also derive an alternative bias-corrected estimator, based on the approach of Firth [14].

Strijbosch & Does [34] compare various jackknife and bootstrap estimators intended to improve bias. They show in particular that a jackknife estimator based on omitting each individual sample in turn succeeds in reducing bias. However, Mehrabi & Matthews [26] prefer the likelihood-based estimators because their variance can be estimated reliably, whereas the variance of the jackknife estimator cannot.

Garthright [19] has derived a bias-corrected estimator of $\log\lambda$, arguing that, for many purposes, it is more important to have an unbiased estimator of $\log\lambda$ than of λ itself.

Interval Estimation of λ

The distribution of $\hat{\lambda}$ is often quite skewed, particularly in small assays, and **confidence intervals** based on the standard error of $\hat{\lambda}$ are unreliable. However, the distribution of $\log\hat{\lambda}$ is more nearly symmetrical [7], and approximate $100(1-\alpha)\%$ confidence intervals for $\log\lambda$ can be calculated as

$$\log\hat{\lambda} \pm z_{\alpha/2}\ \text{se}\ [\log(\hat{\lambda})],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard **normal distribution**. These limits can then be back-transformed to give limits for λ itself.

Other "large-sample" methods of forming confidence intervals, based on the score statistic (*see* **Likelihood**) or the **likelihood ratio** statistic, are discussed by Gart [18] and Ridout [30]. Cyr et al. [11] give some alternative large sample methods. For all of these methods the coverage (the probability that the interval contains λ) fluctuates around the nominal confidence level, and depends on λ.

Alternatively, "exact" confidence intervals can be constructed. "Exact" here signifies that the intervals are based on the exact binomial probabilities that determine the outcome of the assay. The intervals themselves are *conservative*, i.e. their coverage is never below, and generally exceeds, the nominal level, though this conservatism can be reduced by the use of mid-P values [21]. The various methods are well described by Loyer & Hamilton [23]. Roughly speaking their preferred method chooses the confidence interval as those values of λ for which the observed outcome of the assay is not "improbable". A complication is that some assay outcomes are improbable, whatever the value of λ, and so the method can lead to empty confidence intervals. It can also occasionally lead to disjoint intervals. Hepworth [21] gives an alternative method for which these complications do not arise. It is based on ordering outcomes according to their associated maximum likelihood estimates, and was proposed originally by Woodward [36]. Myers et al. [28] describe a third method based on an exact likelihood ratio test.

Exact intervals can require large amounts of computation and are perhaps best suited to small assays with a standard design, so that confidence intervals can be tabulated for all possible outcomes. Ridout [30] compares exact intervals with large sample intervals for some small assays and recommends intervals based on the large sample distribution of the likelihood ratio statistic.

Basu et al. [2] describe a bootstrap method of constructing confidence intervals for λ. Their method tends to give shorter intervals than the likelihood ratio method, but the coverage is often below the nominal level.

Testing the Validity of the Assumptions

Goodness of fit is often assessed in practice by comparing observed and fitted values using the standard χ^2 statistic for binomial data. However, because this is a general purpose test statistic it tends to have low power against specific alternatives [4].

An idealized assay involving an infinite number of dilutions with constant dilution factor and five replicates at each dilution might yield the following results:

$$\ldots 5\ 5\ 5\ 3\ 0\ 0\ 2\ 0\ 0\ 0 \ldots$$

The transition from all positive to all negative results spans four dilutions (with outcomes 3,0,0,2). Stevens [33] suggested using the length of the transition as a goodness-of-fit statistic, calling it the *range*. Large values of the range indicate a poor fit. The same test statistic was proposed independently by Moran [27]. Haas & Heller [20] develop the test for short series involving only three or four distinct dilutions (*see* **Infectivity Titration**).

When micro-organisms are not distributed randomly this is almost always because they are aggregated or clustered. Consequently, the distribution of micro-organisms in a sample is overdispersed relative to the Poisson distribution. Suppose, for example, that for a sample of volume v the distribution is not Poisson with mean λv but a **negative binomial distribution** with the same mean and with variance $\lambda v + \tau \lambda v^2$. Then the probability that the sample contains one or more micro-organisms is

$$1 - (1 + \tau \lambda v)^{-1/\tau},$$

which approaches the standard model in the limit as $\tau \to 0$. Testing for this particular departure from randomness is therefore equivalent to testing the null hypothesis that $\tau = 0$. A likelihood ratio test could be used, but this would involve fitting the alternative model. Cyr & Singh [9] suggest instead a score test for which it is only necessary to fit the standard model. This score test is robust insofar as the same test statistic results from some other distributions that are overdispersed relative to the Poisson.

However, some types of overdispersion are not detectable. For example, if the number of micro-organisms in a sample of volume v again has a negative binomial distribution with mean λv, but now with variance $(1 + \phi)\lambda v$ (where $\phi > 0$), then it can be shown that the standard analysis will incorrectly estimate $\lambda' = \lambda \log(1 + \phi)/\phi$ instead of λ, but there will be no apparent lack of fit.

The other principal assumption is that the presence of at least one bacterium in a sample is sufficient to produce an observable response. Any departures from this assumption will result in underestimation of λ. Particular types of departure that might arise depend on the specific application. Cyr & Singh [10] develop score tests for several alternatives that are important in immunology.

Planning a Serial Dilution Assay

Fisher [16] showed that to minimize the asymptotic variance of the maximum likelihood estimator $\hat{\lambda}$, the volume of the original suspension present in every sample should be $v = 1.59/\lambda$. This volume is such that each sample has a probability of 0.80 of giving a positive response. This result is not of much practical use unless a good prior estimate of λ is available.

Most work on planning of serial dilution assays has assumed instead that the experimenter can indicate an interval (λ_L, λ_U) within which the true value of λ is believed to lie. Finney [13, Section 20.8], for example, suggests choosing a range of dilutions such that the expected number of organisms per sample (λv) is at least 2 at the first dilution and at most 0.5 at the last dilution. Cochran [7] suggests slightly different limits. Dilutions are then made with a constant dilution factor, d, to cover the required range of volumes. The usual recommendation is that d should be as small as is practicable. For example, two-fold dilutions would be preferred to five- or 10-fold. This is because the smaller the value of d, the more nearly constant is the standard error of $\log \lambda$, when the same number of samples are tested in total [7]. One disadvantage of a small dilution factor is that there will be a greater cumulative effect of any pipetting errors. The results of Chase & Hoel [5] suggest that small errors of dilution can increase the variance of $\hat{\lambda}$ substantially.

For $d \leq 5$, Cochran [7] notes that the standard error of $\log_{10}(\hat{\lambda})$ is approximately independent of λ over the interval (λ_L, λ_U) and is given by

$$0.55 \left(\frac{\log_{10} d}{n} \right)^{1/2},$$

where n is the number of samples at each dilution (assumed equal). For $d \geq 10$ the approximation is better if the constant 0.55 is increased to 0.58. This approximate formula is useful in choosing a suitable value of n.

Strijbosch et al. [35] give an alternative procedure which ensures that, for any value of λ in the interval (λ_L, λ_U), there are some dilutions that are "informative", in the sense that the probability of a positive result is not too close to zero or one.

Another approach is to construct Bayesian optimal designs, based on a prior distribution for λ. Zacks [37] and Ridout [31] have given designs based on a gamma prior for λ and a uniform prior for $\log \lambda$ respectively. These authors also discuss two- and three-stage Bayesian designs.

Multistage designs are also considered by Abdelbasit & Plackett [1]. Given an initial estimate of λ, λ_0 say, the first stage tests samples with volume $v_1 = 1.59/\lambda_0$. The second stage tests samples with volume $v_2 = 1.59/\hat{\lambda}_1$, where $\hat{\lambda}_1$ is the maximum likelihood estimator of λ based on data from the first stage. At subsequent stages the volume is determined by the maximum likelihood estimator of λ based on data from all previous stages. Abdelbasit & Plackett present results on efficiency of designs with up to five stages. The optimal number of stages depends on how good the initial estimate λ_0 is and on how many samples are to be tested in total.

References

[1] Abdelbasit, K.M. & Plackett, R.L. (1983). Experimental design for binary data, *Journal of the American Statistical Association* **78**, 90–98.

[2] Basu, S., Guerra, R. & Read, R. (1996). Bootstrap confidence intervals for concentration parameters in dilution assays, *Journal of Agricultural, Biological and Environmental Statistics* **1**, 454–466.

[3] Best, D.J. & Raynor, J.C.W. (1985). A comparison of the MPN and Fisher–Yates estimators for the density of organisms, *Biometrical Journal* **27**, 167–172.

[4] Bonnefoix, T. & Sotto, J.-J. (1994). The standard χ^2 test used in limiting dilution assays is insufficient for estimating the goodness-of-fit to the single-hit Poisson model, *Journal of Immunological Methods* **167**, 21–33.

[5] Chase, G.R. & Hoel, D.G. (1975). Serial dilutions: error effects and optimal designs, *Biometrika* **62**, 329–334.

[6] Chen, C.L. & Swallow, W.H. (1990). Using group testing to estimate a proportion, and to test the binomial model, *Biometrics* **46**, 1035–1046.

[7] Cochran, W.G. (1950). Estimation of bacterial densities by means of the "most probable number", *Biometrics* **6**, 105–116.

[8] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[9] Cyr, L. & Singh, K.P. (1991). Validity tests for the single-hit Poisson model in serial dilution experiments. Presented at *Third International Conference on Environmetrics*, 7–10 October, Madison, Wisconsin.

[10] Cyr, L. & Singh, K.P. (1993). Score tests for the single-hit Poisson model in limiting dilution assays, *Environmetrics* **4**, 105–121.

[11] Cyr, L., Rust, P.F., Peters, J.R., Schmehl, M.K. & Bank, H.L. (1993). Confidence intervals for the relative frequency of responding cells in limiting dilution assays, *Biometrics* **49**, 491–498.

[12] Dorfman, R. (1943). The detection of defective members of large populations, *Annals of Mathematical Statistics* **14**, 436–440.

[13] Finney, D.J. (1978). *Statistical Method in Biological Assay*, 2nd Ed. Griffin, London.

[14] Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika* **80**, 27–38. Correction: **82** (1995) 667.

[15] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, Series A* **222**, 309–368.

[16] Fisher, R.A. (1928). *Statistical Methods for Research Workers*, 2nd Ed. Oliver & Boyd, Edinburgh.

[17] Fisher, R.A. & Yates, F. (1970). *Statistical Tables for Biological, Agricultural and Medical Research*, 6th Ed. Oliver & Boyd, Edinburgh.

[18] Gart, J.J. (1991). An application of score methodology: confidence intervals and tests of fit for one-hit curves, in *Handbook of Statistics*, Vol. 8, C.R. Rao & R. Chakraborty, eds. North-Holland, Amsterdam, pp. 395–406.

[19] Garthright, W.E. (1993). Bias in the logarithm of microbial density estimates from serial dilutions, *Biometrical Journal* **35**, 299–314.

[20] Haas, C.N. & Heller, B. (1988). Test of the validity of the Poisson assumption for analysis of most-probable-number results, *Applied and Environmental Microbiology* **54**, 2996–3002.

[21] Hepworth, G. (1996). Exact confidence intervals for proportions estimated by group testing, *Biometrics* **52**, 1134–1146.

[22] Klee, A.J. (1993). A computer program for the determination of most probable number and its confidence limits, *Journal of Microbiological Methods*, **18**, 91–98.

[23] Loyer, M.W. & Hamilton, M.A. (1984). Interval estimation of the density of organisms using a serial-dilution experiment, *Biometrics* **40**, 907–916.

[24] McCrady, M.H. (1915). The numerical interpretation of fermentation-tube results, *Journal of Infectious Diseases* **17**, 183–212.

[25] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[26] Mehrabi, Y. & Matthews, J.N.S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays, *Biometrics* **51**, 1543–1549.

[27] Moran, P.A.P. (1958). Another test for heterogeneity of host resistance in dilution assays, *Journal of Hygiene* **56**, 319–322.

[28] Myers, L.E., McQuay, L.J. & Hollinger, F.B. (1994). Dilution assay statistics, *Journal of Clinical Microbiology* **32**, 732–739.

[29] Ridout, M.S. (1990). Non-convergence of Fisher's method of scoring – a simple example, *GLIM Newsletter* **20**, 8–11.

[30] Ridout, M.S. (1994). A comparison of confidence interval methods for dilution series experiments, *Biometrics* **50**, 289–296.

[31] Ridout, M.S. (1995). Three-stage designs for seed testing experiments, *Applied Statistics* **44**, 153–162.

[32] Salama, I.A., Koch, G.G. & Tolley, H.D. (1978). On the estimation of the most probable number in a serial dilution experiment, *Communications in Statistics – Theory and Methods*, **A7**, 1267–1281.

[33] Stevens, W.L. (1958). Dilution series: a statistical test of technique, *Journal of the Royal Statistical Society, Series B* **20**, 205–214.

[34] Strijbosch, L.W.G. & Does, R.J.M.M. (1988). Comparison of bias-reducing methods for estimating the parameter in dilution series, *Communications in Statistics – Simulation and Computation* **17**, 1173–1190.

[35] Strijbosch, L.W.G., Buurman, W.A., Does, R.J.M.M., Zinken, P.H. & Groenewegen, G. (1987). Limiting dilution assays: experimental design and statistical analysis, *Journal of Immunological Methods* **97**, 133–140.

[36] Woodward, R.L. (1957). How probable is the most probable number?, *Journal of the American Waterworks Association* **49**, 1060–1068.

[37] Zacks, S. (1977). Problems and approaches in design of experiments for estimation and testing in nonlinear models, in *Multivariate Analysis*, Vol. 4, P. Krishnaiah, ed. North-Holland, Amsterdam, pp. 209–223.

MARTIN RIDOUT

Serial-sacrifice Experiments

In simple survival experiments with animals to study the development of one or more particular disease(s), one observes the age at death along with the presence or absence of the disease(s) and also, in studies with **competing risks**, the cause of death for each animal. At the end of the experiment (which may be prefixed by design), there may be a provision for killing all the animals surviving at that time point, which is called *terminal sacrifice*. Such data give little information on the progress of disease(s), if they can be detected only at death, and possible interaction between different disease(s). For example, in many animal carcinogenicity experiments, development of cancer (occurrence of tumor) is of particular interest along with its contribution to the final death and possible interaction with other causes of death. But simple survival experiments give insufficient information to study this aspect, although with the introduction of *conceptual* survival times due to different causes of death and strong assumption such as independence between incidence of tumor and subsequent death and death from other causes, some analyses have been carried out [16, 20, 23, 27, 42, 52] to estimate the distribution of time to tumor incidence. One can also circumvent this problem by assuming strong parametric models ([10, 14, 26] (*see* **Parametric Models in Survival Analysis**); see also Borgan et al. [7], who consider piecewise-constant intensity parameters to demonstrate that serial-sacrifice experiments are moderately efficient with respect to complete observation, whereas simple survival experiments have very low efficiency). There is also a semiparametric approach wherein parametric assumptions are made for a part (tumor incidence rate) of the model, thereby requiring few or only one (terminal) sacrifices [2, 15, 30, 44, 46, 47]. In general also, data from simple survival experiments are analyzed by introducing conceptual survival times due to different causes of death (assumed mutually exclusive and exhaustive) which are then assumed independent using parametric or nonparametric methods [19, 22, 40].

Tsiatis [50] has shown that, with data only from simple survival experiment, **identifiability** problems may arise for the distribution of the conceptual survival times; it is impossible to distinguish the independent model from some dependence models. However, the point is that a simple survival experiment is not an appropriate one to study development of different diseases and possible interactions between them (leading to dependence between the conceptual survival times).

Serial-sacrifice experiments are an improvement over simple survival experiment toward achieving this goal, as indicated above, by allowing interim observation on some individuals before death occurs. Intuitively, this will allow one to probe the disease process and the complex interrelationships between different diseases before death occurs. Ideally, in serial-sacrifice experiments, individual animals are randomly selected and killed (sacrificed) at fixed or adaptively selected time points, allowing the examination of presence or absence of different diseases (which is done also for naturally dying animals); cause of death for each case may or may not be recorded [4, 9, 38, 42, 51].

The above serial-sacrifice experiment, in principle, permits one to test if the different diseases are independent and to determine the nature of interaction if it exists, as considered in a complicated illness–death model by Berlin et al. [4] assuming Markov transition rates (*see* **Fix–Neyman Process**). In animal carcinogenicity experiments, as mentioned earlier, serial sacrifice allows the estimation of tumor occurrence time distribution under very general conditions (see the next section for more clarification).

Identifiability and Analysis

Clearly, a serial-sacrifice experiment, as described above, embodies two sampling mechanisms in operation – one corresponding to death of an animal and the other to sacrifice. The latter is unbiased in the sense that all the animals alive at time t have equal chance of being sacrificed, thus providing an estimate for $p_{\mathcal{A}}(t)$, probability of being alive and having a particular disease combination $\mathcal{A}$ at time t. On the other hand, death is a biased sampling mechanism, the probability of death at any time depending on the illness state history of the animal. This will, thus, provide an estimate of probability of death with the particular disease combination $\mathcal{A}$, given by $\mu_{\mathcal{A}}(t)p_{\mathcal{A}}(t)$, where $\mu_{\mathcal{A}}(t)$ denotes the mortality rate from disease combination $\mathcal{A}$. Note that $S(t)$, the probability of

being alive at time t, can be estimated either in the **Kaplan–Meier** style [24] using both death and sacrifice information, or simply summing estimates of $p_{\mathcal{A}}(t)$ over all possible $\mathcal{A}$s. Hence, the prevalence of disease combination $\mathcal{A}$ (probability of having disease combination A given alive) at time t, $p_A(t)/S(t)$, can also be estimated. The **likelihood** function of the observable quantities is, therefore, a product of terms such as $p_A(t)$ and $\mu_A(t)p_A(t)$ for different disease combinations and different sacrifice and death times, respectively. Thus, only terms such as the above two and functions thereof will be estimable (identifiable) from serial-sacrifice experiments. Clearly, this is more than one can do from a simple survival experiment with different diseases. However, as pointed out by Clifford [9], assuming a progressive Markov illness–death model, quantities related to transition from one illness state (alive with a particular disease combination) to another, and hence quantities related to events subsequent to an illness state (e.g. mean residual life after being in an illness state), are not identifiable even from a serial sacrifice experiment; that is, these quantities are not expressible in terms of the above two quantities.

Berlin et al. [4] formulate different cases of independence between different diseases leading to relationships between different μs and ps from relationships between different transition rates. For example, if disease a is independent of disease b, then, besides many others, we should have the relationship $\mu_{(a,b)} = \mu_a + \mu_b$ (notation having the usual meaning). These relationships allow identifiability to some extent. Thus, hypotheses of disease independence, in some cases, can be rejected when they are false. One has to take note of the number of estimable independent parameters while calculating **degrees of freedom** for an asymptotic **likelihood ratio test**. However, failure to reject the hypotheses when they are false may be due not only to a type II error (*see* **Level of a Test**), but also to the nonidentifiability of the model parameters (i.e. the fact that relationships implied by hypotheses of independence may not necessarily imply independence).

In this context, two papers by Turnbull and Mitchell (see [38, 51]) also merit mention. Here, the authors write the likelihood function in terms of the estimable parameters of the type $p_{\mathcal{A}}$s and $\mu_{\mathcal{A}}$s, and then suggest estimation of these parameters using the **EM algorithm** [11]. They also parameterize the model in **loglinear** form by taking *treatment, time*, and *illness states* as different factors for reparameterizing the "prevalence" and "lethality" parameters corresponding to observations from sacrificed and dead animals, respectively. A generalized EM algorithm is used, by making use of the **iterative proportional fitting** algorithm [6] in the M-step, to obtain maximum likelihood estimates of the parameters for a broad class of unsaturated models. Tests based on relative likelihoods are proposed to investigate the effects of treatment, time, and the presence of other diseases on the prevalence and lethality of a particular disease of interest. Mitchell & Turnbull [39] develop a computer program for this analysis.

In animal carcinogenicity experiments with serial sacrifice, the tumor occurrence time distribution can be expressed, at least approximately, in terms of the estimable quantities as mentioned above ([12, 33, 37]; see [36] for a review). Thus, the tumor occurrence time distribution becomes estimable under very general conditions without assuming independence between different event times (tumor occurrence, death with tumor present, and death with no tumor) and having no cause of death information. McKnight & Crowley [37] give a closed form estimator for a function which approximates the tumor incidence rate; this, however, may take a negative value. A test for differences in tumor incidence rates in two groups is suggested, by comparing estimates of the approximate tumor incidence rates. Dewanji & Kalbfleisch [12] develop an EM algorithm for estimating the tumor incidence rates in a discrete framework. A score test (*see* **Likelihood**) is developed for comparison of two groups with respect to tumor incidence, assuming a polychotomous logistic model for the tumor incidence rate and death rate with no tumor. Malani & Van Ryzin [33] give closed form estimates of the tumor incidence rates, also in a discrete framework, but these may fail to satisfy the nonnegativity condition in some cases. Malani & Van Ryzin [34] extend this work to the problem of comparing two treatment groups.

Design Issues

In designing serial-sacrifice experiments, optimal sacrifice schedules have received some attention. The choice of an **optimal design** depends on the specific criterion to be optimized. Thus, a design which is optimal for one criterion may not be so for a different

one. Sometimes procedural simplicity dictates selection of a specific criterion.

Berry [5] considers the determination of an optimal time to terminate an experiment by sacrificing all surviving animals in terms of maximum Fisher's **information** per unit cost, assuming a **Weibull distribution** for tumor incidence. It turns out that the optimum strategy is to allow all animals to live out their lives. Portier [43] addresses the question of finding optimal fixed sacrifice times for small carcinogenesis experiments using different optimal criteria including optimal **goodness of fit**. The study is very limited, as it considers only four underlying tumor incidence distributions and six designs.

It is argued [3] that when no prior knowledge on tumor incidence is available, information on an optimum sacrifice schedule should be obtained only from the data. Otherwise, if sacrifices are carried out too early there will be few animals with tumor and if sacrifices are too late nearly all animals will have tumor, leading to very little information on tumor incidence in either case. Such consideration led to modification of serial sacrifice schedule in the ED01 study [21, 49] when too few tumors were found at the beginning of the study. Also, in the FD & C Red No. 40 mouse experiment [28], unexpected findings of early reticulo-endothelial (RE) tumors led to an acceleration of the sacrifice rate. All this suggests consideration of sequential or adaptive approach for choosing an optimal design although that requires a quick histopathological examination of all the sacrificed/dead animals.

Bergman & Turnbull [3] consider this problem for nonlethal tumors assuming an **exponential distribution** for tumor incidence. For a fixed sequence of times $t_1 < \cdots < t_M$, at which one or more sacrifice could be made, their procedure suggests how many to sacrifice at each time point so that estimation of the exponential parameter is asymptotically efficient in the sense that the Fisher information approaches that as obtained by using the true value of the parameter, as the total number of animals increases. At time t_i, animals are selected one by one at random and sacrificed until either a stopping rule $\mathcal{R}_i$ is satisfied or there are no animals left for sacrifice, whichever occurs first. If there are still more animals left after $\mathcal{R}_i$ is satisfied, the next sacrifice is made at time t_{i+1}, and so on. At the last time t_M, all the remaining animals are sacrificed and the experiment is terminated. The stopping rule they suggest for achieving asymptotic efficiency stops sacrificing at time t_i, when the ratio of the number of animals with tumor *plus* a specified constant to the number of animals without tumor, found at time t_i, is sufficiently small.

In order to find optimum sacrifice times, for experimenting with nonlethal tumors, Bergman & Turnbull [3] (see also [8] and [18] in a different context) note, assuming exponential distribution with mean $1/\theta$ for tumor incidence, that Fisher's information is maximized by choosing only one sacrifice time at $1.5936 \times \theta^{-1}$. This result is of no immediate use as the optimal design depends on unknown θ. Chernoff [8] suggests a sequential procedure by which the ith sacrifice will be at time $1.5936 \times \hat{\theta}_i^{-1}$, where $\hat{\theta}_i$ is the current best estimate of θ based on observation from the $i-1$ animals sacrificed so far (see also [1]). However, this optimal design does not ensure that the ith sacrifice time will always be later than the $(i-1)$th sacrifice time. Louis [31] proposes an asymptotically efficient suboptimal rule for time-ordered sequential design which has favorable small sample properties (see also [32] and [41]).

Portier [45] gives a brief review of the different design issues discussed here.

Examples

The earliest example (found in the literature) of serial-sacrifice experiment is that of Upton et al. [53], conducted at the Oak Ridge National Laboratory. Groups of RFM female mice were given various doses of γ-radiation. Each group included about 4000 mice, of which about 300 were sacrificed at different times ranging from 150 to 800 days. For all animals, postmortem examination records presence or absence of up to eight diseases. Berlin et al. [4] consider two groups (control and irradiated to $300R$) with three disease categories from the eight for their analysis. Parts of these data have been considered for analysis by many other authors [12, 38, 44, 51].

The ED01 study with 2-acetylamino-fluorene (2-AAF) conducted by the US National Center for Toxicological Research (NCTR) ([49]; see also a series of 19 papers in a special issue of *Journal of Environmental Pathology and Toxicology* [21] – Littlefield et al., pp. 17–34, in this series may be mentioned as one example) is probably the largest serial-sacrifice experiment so far. This study involved

over 24 000 mice exposed to one of eight different doses varying from 0 to 150 ppm in the diet. Serial sacrifices were carried out at 9, 12, 14, 15, 16, 17, 18, and 24 months, with a terminal sacrifice at 33 months. In addition to presence/absence of tumor at different sites, cause of death was also ascertained. The data from this study have been extensively analyzed by many authors [2, 13–15, 30, 35].

Another experiment by NCTR with benzidine dihydrochloride in mice has been considered for analysis by many authors [16, 26, 27, 30, 34, 46, 55, 56]. Information on cause of death and presence/absence of liver tumor was available. Sacrifices were designed at specific time points (280, 420, and 560 days).

Red 40 data [28] considered the incidence of RE tumors due to dosing of Red 40. The experiment used three treated groups and a control in both sexes of CD-1 HAM/ICR mice with about 50 animals in each group. A single interim sacrifice of between 14 and 20 animals per group was done at 42 weeks and a terminal sacrifice at 104 weeks.

There are few other examples of serial-sacrifice experiments found in the literature. Levitt et al. [29] studied morphogenesis of pancreatic adenocarcinoma in Syrian golden hamster induced by *N*-nitroso-*bis*(2-hydroxypropyl)amine. Kennedy et al. [25] considered ^{210}Po-induced tumor in the peripheral lung of Syrian golden hamster. Borgan et al. [7] mentioned an experiment for studying lymphatic leukemia in mice (see also [17]). Schuller et al. [48] studied pulmonary toxicity induced by the anticancer drug 1,3-*bis*(2-chloroethyl)-1-nitrosourea (BCNU), in F344 rats. Van Nesselrooij et al. [54] considered blood-filled cavities in estrogen-induced anterior pituitary tumors in male Sprague–Dawley rats with two treated and two control animals sacrificed at each of 15 time points ranging from 7 to 272 days.

References

[1] Abdelbasit, K.M. & Plackett, R.L. (1983). Experimental design for binary data, *Journal of the American Statistical Association* **78**, 90–98.

[2] Archer, L. & Ryan, L. (1989). Accounting for misclassification in the cause-of-death test for carcinogenicity, *Journal of the American Statistical Association* **84**, 787–791.

[3] Bergman, S.W. & Turnbull, B.W. (1983). Efficient sequential designs for destructive life testing with application to animal serial sacrifice experiments, *Biometrika* **70**, 305–314.

[4] Berlin, B., Brodsky, J. & Clifford, P. (1979). Testing disease dependence in survival experiments with serial sacrifice, *Journal of the American Statistical Association* **74**, 5–14.

[5] Berry, G. (1975). Design of carcinogenesis experiments using Weibull distribution, *Biometrika* **62**, 321–328.

[6] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[7] Borgan, Ø., Liestøl, K. & Ebbesen, P. (1984). Efficiencies of experimental designs for an illness - death model, *Biometrics* **40**, 627–638.

[8] Chernoff, H. (1972). *Sequential Analysis and Optimal Design*. SIAM, Philadelphia.

[9] Clifford, P. (1977). Nonidentifiability in stochastic models of illness and death, *Proceedings of the National Academy of Sciences* **74**, 1338–1340.

[10] David, H.A. (1974). Parametric approaches to the theory of competing risks, in *Reliability and Biometry, Statistical Analysis of Lifelength*, F. Proschan & R.J. Serfling, eds. SIAM, Philadelphia, pp. 275–290.

[11] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–22.

[12] Dewanji, A. & Kalbfleisch, J.D. (1986). Nonparametric methods for survival/sacrifice experiments, *Biometrics* **42**, 325–341.

[13] Dewanji, A., Krewski, D. & Goddard, M.J. (1993). A Weibull model for the estimation of tumorigenic potency, *Biometrics* **49**, 367–377.

[14] Dinse, G.E. (1988). Simple parametric analysis of animal tumorigenicity data, *Journal of the American Statistical Association* **83**, 638–649.

[15] Dinse, G.E. (1991). Constant risk differences in the analysis of animal tumorigenicity data, *Biometrics* **47**, 681–700.

[16] Dinse, G.E. & Lagakos, S.W. (1982). Nonparametric estimation of lifetime and disease onset distributions from incomplete observations, *Biometrics* **38**, 921–932.

[17] Ebbesen, P., Borgan, Ø. & Liestøl, K. (1983). Decreasing leukemia risk in old AKR mice, *Experimental Gerontology* **18**, 347–353.

[18] Fisher, R.A. (1971). *The Design of Experiments*, 9th Ed. Oliver & Boyd, Edinburgh.

[19] Gail, M. (1975). A review and critique of some models used in competing risk analysis, *Biometrics* **31**, 209–222.

[20] Hoel, D.G. & Walburg, H.E. (1972). Statistical analysis of survival experiments, *Journal of the National Cancer Institute* **49**, 361–372.

[21] *Journal of Environmental Pathology and Toxicology* (1980). **3**, 1–246.

[22] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[23] Kalbfleisch, J.D., Krewski, D. & Van Ryzin, J. (1983). Dose - response models for time-to-response toxicity data, *Canadian Journal of Statistics* **11**, 25–49.

[24] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[25] Kennedy, A.R., McGundy, R.B. & Little, J.B. (1978). Serial sacrifice study of pathogenesis of ^{210}Po-induced lung tumors in Syrian golden hamsters, *Cancer Research* **38**, 1127–1135.

[26] Kodell, R.L. & Nelson, C.J. (1980). An illness - death model for the study of the carcinogenic process using survival/sacrifice data, *Biometrics* **36**, 267–277.

[27] Kodell, R.L., Shaw, G.W. & Johnson, A.M. (1982). Nonparametric joint estimators for disease resistance and survival functions in survival/sacrifice experiments, *Biometrics* **38**, 43–58.

[28] Lagakos, S.W. & Mosteller, F. (1981). A case study of statistics in the regulatory process: the FD & C Red No. 40 experiments, *Journal of the National Cancer Institute* **66**, 197–212.

[29] Levitt, M.H., Harris, C.C., Squire, R., Springer, S., Wenk, M., Mollelo, C., Thomas, D., Kingsbury, E. & Newkirk, C. (1977). Experimental pancreatic carcinogenesis. I. Morphogenesis of pancreatic adenocarcinoma in the Syrian golden hamster induced by *N*-nitroso-*bis*(2-hydroxypropyl)amine, *American Journal of Pathology* **88**, 5–28.

[30] Lindsay, J.C. & Ryan, L.M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments, *Applied Statistics* **42**, 283–300.

[31] Louis, T.A. (1987). Efficient monotone sequential design, *Research Report No. MS-R8705*. Centre for Mathematics and Computer Science, Amsterdam.

[32] Louis, T.A. & Orav, E.J. (1985). Sacrifice plans for the carcinogen bioassay, in *Proceedings of Long Term Animal Carcinogenicity Studies: A Statistical Perspective*. American Statistical Association, Washington DC, pp. 36–41.

[33] Malani, H. & Van Ryzin, J. (1986). Nonparametric estimates of the tumor incidence rate, the tumor lethality rate and the mortality rate in absence of tumors in animal carcinogenicity experiments, *Columbia University Statistical Reports, Technical Report B-56*. Department of Biostatistics, Columbia University, New York.

[34] Malani, H. & Van Ryzin, J. (1988). Comparison of two treatments in animal carcinogenicity experiments, *Journal of the American Statistical Association* **83**, 1171–1177.

[35] Malani, H.M. & Lu, Y. (1993). Animal carcinogenicity experiments with and without serial sacrifice, *Communications in Statistics – Theory and Methods* **22**, 1557–1584.

[36] McKnight, B. (1988). A guide to the statistical analysis of long-term carcinogenicity assays, *Fundamental and Applied Toxicology* **10**, 335–364.

[37] McKnight, B. & Crowley, J. (1984). Tests for differences in tumor incidence based on animal carcinogenesis experiments, *Journal of the American Statistical Association* **79**, 639–648.

[38] Mitchell, T.J. & Turnbull, B.W. (1979). Log-linear models in the analysis of disease prevalence data from survival/sacrifice experiments, *Biometrics* **35**, 221–234.

[39] Mitchell, T.J. & Turnbull, B.W. (1983). A computer program for the statistical analysis of disease prevalence data from survival/sacrifice experiments, *Computer Methods and Programs in Biomedicine* **17**, 45–64.

[40] Moeschberger, M.L. & David, H.A. (1971). Life tests under competing causes of failure and the theory of competing risks, *Biometrics* **27**, 909–933.

[41] Orav, E.J. & Louis, T.A. (1985). Adaptive terminal sacrifice plans for the rodent bioassay, *Harvard Biostatistics Research Report*.

[42] Peto, R., Pike, M., Day, N., Gray, R., Lee, P., Parish, S., Peto, J., Richards, S. & Wahrendorf, J. (1980). Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments, in *IARC Monograph on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, Supplement 2, Long-term and Short-term Screening Assays for Carcinogens: a Critical Appraisal*. IARC, Lyon, pp. 311–346.

[43] Portier, C. (1985). Optimal dose/animal allocation for terminal sacrifice carcinogenicity studies, in *Proceedings of the ASA Conference on Long-Term Animal Carcinogenicity Studies*. American Statistical Association, Washington, pp. 45–50.

[44] Portier, C.J. (1986). Estimating the tumor onset distribution in animal carcinogenesis experiments, *Biometrika* **73**, 371–378.

[45] Portier, C.J. (1991). Design of two-year carcinogenicity experiments: dose allocation, animal allocation and sacrifice times, in *Statistics in Toxicology*. D. Krewski & C. Franklin, eds. Gordon & Breach, New York, pp. 457–469.

[46] Portier, C.J. & Dinse, G.E. (1987). Semiparametric analysis of tumor incidence rates in survival/sacrifice experiments, *Biometrics* **43**, 107–114.

[47] Ryan, L. & Orav, E.J. (1988). On the use of covariates for rodent bioassay and screening experiments, *Biometrika* **75**, 631–637.

[48] Schuller, H.M., Smith, A.C., Gregg, M. & Boyd, M.R. (1985). Sequential pathological changes induced in rats with the anticancer drug 1,3-*bis*(2-chloroethyl)-1-nitrosourea (BCNU), *Experimental Lung Research* **9**, 327–339.

[49] Staffa, J.A. & Mehlman, M.A. (1979). *Innovations in Cancer Risk Assessment (ED01 Study)*. Pathotox, Park Forest South.

[50] Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences* **72**, 20–22.

[51] Turnbull, B.W. & Mitchell, T.J. (1978). Exploratory analysis of disease prevalence data from survival/sacrifice experiments, *Biometrics* **34**, 555–570.

[52] Turnbull, B.W. & Mitchell, T.J. (1984). Nonparametric estimation of distributions of time to onset and time to death for specific diseases in survival/sacrifice experiments, *Biometrics* **40**, 41–50.

[53] Upton, A.C., Allen, R.C., Brown, R.C., Clapp, N.K., Conlin, J.W., Cosgrove, G.E., Darden, E.B. Jr, Kastenbaum, M.A., O'dell, P.T. Jr, Serrano, L.J., Tyndall, R.L. & Walburg, H.A. Jr (1969). Quantitative experimental study of low-level radiation carcinogenesis, in *Radiation Induced Cancer*. International Atomic Energy Agency, Vienna, pp. 425–438.

[54] Van Nesselrooij, J.H., Hendriksen, G.J., Feron, V.J. & Bosland, M.C. (1992). Pathogenesis of blood-filled cavities in estrogen-induced anterior pituitary tumors in male Sprague - Dawley rats, *Toxicologic Pathology* **20**, 71–80.

[55] Williams, P.L. & Portier, C.J. (1992). Analytic expressions for maximum likelihood estimators in a nonparametric model of tumor incidence and death, *Communications in Statistics – Theory and Methods* **21**, 711–732.

[56] Williams, P.L. & Portier, C.J. (1992). Explicit solutions for constrained maximum likelihood estimators in survival/sacrifice experiments, *Biometrika* **79**, 717–729.

(*See also* **Animal Screening Systems**; **Markov Chains**; **Tumor Incidence Experiments**)

A. DEWANJI

Sex Ratio at Birth

Most human populations contain approximately equal numbers of males and females. This apparent equality can clearly be only approximate, since the mortality and migration rates of the two sexes vary in different ways with age and with time. In most communities, the males slightly outnumber the females at birth, a ratio of about 1.05 being typical, but are subject to higher mortality rates, so that females are in excess at higher ages and have the higher expectations of life. Studies of the variation in the sex ratio are predominantly concerned with the sex ratio at birth, conventionally expressed as the ratio of males to females (sometimes multiplied by 100 or 1000). In some research studies, it is more convenient to use the proportion of males, denoted here by p. More formally, the ratio at birth is the *secondary sex ratio*, the *primary sex ratio* being the ratio at conception. Reference is sometimes made to the less well-defined *tertiary sex ratio*, the ratio of males to females at an age at which children become independent of their parents, or perhaps at the onset of reproductive capacity.

The primary sex ratio is especially difficult to estimate, since many early spontaneous abortions are not observed. Some workers have suggested that the ratio is very high, the sperm bearing the Y chromosome perhaps being more successful in achieving fertilization. However, Hytten [20] argues that in spontaneous abortions during the first trimester, females considerably outnumber males, suggesting that the primary sex ratio is quite low. Later in pregnancy, there are more male than female fetuses, but more males than females miscarry or are stillborn. McKeown & Lowe [30] reported that sex ratios in early stillbirths (after 28 weeks' gestation) were not greatly in excess of unity, but that they increased slightly with **gestational age**.

The secondary sex ratio (for which we shall use merely the term "sex ratio") has been a topic of statistical interest for over 200 years. Statisticians and probabilists in the eighteenth and nineteenth centuries seized on the data emerging from birth registers to exemplify the developing theory of **binary** events; we summarize their work below. During the twentieth century, most studies have been motivated by **demographic**, **epidemiologic**, or **genetic** considerations. From the demographic point of view, the sex ratio contributes to an understanding of the age–sex composition of a population, although its effect is very limited because it varies between populations to such a small extent. The epidemiologist is interested in even small differences in the sex ratio between population groups, or in trends over time, because they may point to the possible effects of environmental agents or differing lifestyles.

Much interest has been focused on surveys of the sex distributions in families of various sizes. These studies are interesting from a general point of view and in relation to the effects of family limitation. Perhaps a more important purpose is to examine the evidence for variability in the sex ratio, in order to shed light on the **heritability** of the sex ratio [9]. If the tendency to produce an excess of offspring of one or the other sex is genetically determined, it will be subject to natural selection. Fisher [13] and others [21, 22] have offered theories as to how such a mechanism might act. It is therefore of some interest to see whether distributions of the sex composition in different families provide any evidence for such variation. As we shall note later, such evidence is hard to find.

Early Studies

The following summaries rely heavily on the much fuller accounts given by Hald [18, 19]. The earliest statistical study appears to be that of **Graunt** [17], who noted the slight excess of males at both christenings and burials, in both London and Romsey. Graunt regarded the near equality as a justification for the practice of monogamy rather than polygamy.

Arbuthnot's study [1] is celebrated as the first recorded example of a statistical significance test (*see* **Hypothesis Testing**). Arbuthnot noted that for each of the 82 years between 1629 and 1710, there were more male than female christenings in London. A sign test gives a ***P*** **value** of $(1/2)^{82}$. Arbuthnot regarded the excess of male christenings (and hence, presumably, of births) as evidence of divine providence in compensating for the higher mortality risks encountered by males. The calculations of Nicholas **Bernoulli** [3] showed that the 82 proportions of males are **overdispersed** relative to the **binomial distribution**, although Bernoulli himself appears to have accepted the hypothesis of homogeneity, the mean proportion of males being 0.516.

Daniel Bernoulli [2] used the **normal** approximation to the binomial to study a related data set of christenings in London between 1664 and 1758. He noted that the overall estimate of p was 0.513, but that the value fell to 0.510 during the decade 1721 to 1730. Taking the 10 values during that decade, he showed that they agreed with normal variation for $p = 0.510$ (as judged by the number of years for which the value fell within probable error limits), but less satisfactorily for $p = 0.513$. The demographers Struyck (1687–1769) and Süssmilch (1707–1767), in the meantime, had extended Graunt's descriptive work to cover data sets from various regions, but avoiding any probabilistic analyses.

Laplace [25] used the numbers of male and female births in Paris during 1745 to 1770 to illustrate the calculation of a posterior probability using the normal approximation to the binomial, finding, of course, that the probability of a ratio favoring females was extremely low. Cournot [4] used a similar approach to compare sex ratios in different subgroups, showing an awareness of the danger of **multiple comparisons**. **Poisson** [32] examined births in France from 1817 to 1826, again using the normal approximation and found no evidence for overdispersion between these 10 values. He then examined the yearly figures for different administrative areas, and found too high a frequency of instances where female births were in excess, suggesting that variation between years and between areas had been obscured in the overall picture.

Recent Epidemiologic Work

Although the eighteenth- and nineteenth-century workers were mainly concerned about illustrating theoretical results by using conveniently extensive sets of binary data, many of them were interested in variations in p between different population groups. Explanations of these differences were less easy to find.

Modern workers have access to even more extensive data sets, and publications exploring group differences abound. Yet, explanations remain contentious, and different studies often seem contradictory.

Time trends in national data sets are easily established. Between 1838 and 1997, the sex ratio for live births in England and Wales showed a roughly sinusoidal curve, with smoothed values falling from about 1.05 to 1.04 around 1895, rising to about 1.06 between 1945 and 1975, and falling to a little over 1.05 in 1995 [28, Figure 6.5]. A similar trend in Japan between 1900 and 1995, with a peak at about 1.07 around 1970, and a subsequent fall to about 1.05 in 1995, is reported by Ohmi et al. [31]. Similar recent falls have been reported for other countries [5].

Many other associations have been reported. Black populations have frequently been shown to have low ratios, but extensive ethnic comparisons are hampered by the lack of reliable data in many developing countries. The sex ratio tends to decline with maternal age [27]. A negative relationship with the sex ratio has been reported for a very wide range of factors, including maternal smoking, maternal schizophrenia, fathers who fly extensively, births in late autumn and winter, and multiple births, with a complex effect of parents' hormonal levels and time of conception during the menstrual cycle [20, 28]. The sex ratio tends to decline with increases in the stillbirth rate, a natural consequence of the higher ratio for stillbirths.

Claims for associations need to be replicated with sufficiently large studies to eliminate random variation and serendipitous selection. However, there seems to be a general finding that low sex ratios are associated with deprivation of some sort. Unexpected changes in the sex ratio may therefore provide a form of monitoring to detect adverse environmental effects. This must, however, be a rather blunt instrument, as the change in sex ratio may not become evident for some time after the causative event, and the effect will be nonspecific. In 1978, the sex ratio in Northern Ireland (and in three adjacent counties of the Republic of Ireland) was unusually low (about 1.01), but no explanation has been found.

Family Studies

Geissler [14] published data on the distributions of boys and girls in families of various sizes for about four million births in Saxony from 1876 to 1885. The data have various deficiencies, being derived from statements by parents at the time of birth registration, recording the numbers of previous children of each sex. They have, nevertheless, been the subject of extensive analyses by, among others, Gini [15, 16], Fisher [12], Lancaster [23] and Edwards [6]. Much of the research has been concerned with possible deviations from the binomial distributions to be

expected from purely random sampling with constant p. One difficulty here is that parents may be less likely to stop having children if the current sex distribution is unbalanced; so the sex distributions of completed families of any given size above two might be expected to show excess frequencies for the extreme categories. However, this effect of family limitation may have been less pronounced in Geissler's data than in some later data sets of this type.

Lancaster's [23] doubts about the reliability of Geissler's data were dismissed by Gini [16] and Edwards [6], but were reiterated by Lancaster [24]. Edwards [6] concluded that the data showed excess variability of p between families, which they represented by the **beta-binomial distribution**, and Lindsey and Altham [26] fitted a variety of models for overdispersion, suggesting that the effect increased with family size.

Edwards and Fraccaro have analyzed several later data sets in which the sequence of boys and girls in each family is recorded: 14 230 French families [7], 5477 Swedish families [10, 11], and 60 334 Finnish families [8]. The Swedish data show no evidence of any form of heterogeneity. The French and Finnish data support the hypothesis of a positive **correlation** between the sexes of adjacent births as distinct from the fixed correlation between *all* members of the family that would be expected if there were merely a between-family variance component.

The possible effect of family limitation on these data sets is unclear. As Edwards & Fraccaro observe, the effect may be mitigated by considering the frequencies of different ordered sex combinations in the first N children in families with N births or more, but this is not a complete solution to the problem. An alternative approach is to estimate p directly from observations of the Nth birth, separately for each ordered combination of outcomes for the first $N-1$. Maconochie & Roman [29] examined in this way a large collection of 549 048 singleton births in Scotland from 1975 to 1988, derived from linked records of maternity discharges. They found no evidence of heterogeneity between families, or for the effect of birth order, maternal age, maternal height, paternal or maternal social class, year of delivery, or season of birth, the estimated sex ratio being 1.06.

The evidence for heterogeneity between families is thus, at best, equivocal. If heterogeneity were to be clearly established, it would remain to be shown that this was genetic rather than environmental. As Edwards [9] remarks, "... if genetic variability exists, it is of a very low order of magnitude".

Acknowledgments

I am grateful to Drs Anthony Edwards and Alison Macfarlane for suggesting many of the references used in this article.

References

[1] Arbuthnot, J. (1712). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes, *Philosophical Transactions of the Royal Society of London* **27**, 186–190; Reprinted in *Studies in the History of Statistics and Probability*, Vol. 2, M.G. Kendall & R.L. Plackett, eds. Griffin, London, 1977.

[2] Bernoulli, D. (1770–1771). Mensura sortis ad fortuitam successionem rerum naturaliter contingentium applicata, *Novi Comment. Acad. Sci. Imp. Petrop.* **14**, 26–45; **15**, 3–28.

[3] Bernoulli, N. (1713). Letter to P.R. de Montmort, in *Essay d'Analyse sur les Jeux de Hazard*, 2nd Ed., P.R. de Montmort, ed. Quillau, Paris; reprinted by Chelsea, New York, 1980.

[4] Cournot, A.A. (1843). *Exposition de la Théorie des Chances et des Probabilités*. Hachette, Paris.

[5] Davis, D.L., Gottlieb, M.B. & Stampnitzky, J.R. (1998). Reduced ratio of male to female births in several industrial countries: a sentinel health indicator? *Journal of the American Medical Association* **279**, 1018–1023.

[6] Edwards, A.W.F. (1958). An analysis of Geissler's data on the human sex ratio, *Annals of Human Genetics* **23**, 6–15.

[7] Edwards, A.W.F. (1959). Some comments on Schützenberger's analysis of data on the human sex ratio, *Annals of Human Genetics* **23**, 233–238.

[8] Edwards, A.W.F. (1961). A factorial analysis of sex ratio data, *Annals of Human Genetics* **25**, 117–121.

[9] Edwards, A.W.F. (1962). Genetics and the human sex ratio, *Advances in Genetics* **11**, 239–272.

[10] Edwards, A.W.F. & Fraccaro, M. (1958). The sex distribution in the offspring of 5,477 Swedish ministers of religion, 1585–1920, *Separat ur Hereditas* **44**, 447–450.

[11] Edwards, A.W.F. & Fraccaro, M. (1960). Distribution and sequences of sexes in a selected sample of Swedish families, *Annals of Human Genetics* **24**, 245–252.

[12] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[13] Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, London, New York.

[14] Geissler, A. (1889). Beiträge zur Frage des Geschlechtsverhältnisses der Geborenen, *Zeitschrift des Königlichen Sächsischen Statistichen Bureaus* **35**, 1–24, 56.

[15] Gini, C. (1908). *Il Sesso dal Punto di Vista Statistico*. Sandron, Milan.
[16] Gini, C. (1951). Combinations and sequences of sexes in human families and mammal litters, *Acta Genetica et Statistica Medica* **2**, 220–244.
[17] Graunt, J. (1662). *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*. Martin, Allestry & Dicas, London.
[18] Hald, A. (1990). *A History of Probability and Statistics and their Applications before 1750*. Wiley, New York.
[19] Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York.
[20] Hytten, F.E. (1982). Boys and girls, *British Journal of Obstetrics and Gynaecology* **89**, 97–99.
[21] Kalmus, H. & Smith, C.A.B. (1960). Evolutionary origin of sexual differentiation and the sex ratio, *Nature* **186**, 1004–1006.
[22] Karlin, S. (1986). *Theoretical Studies on Sex Ratio Evolution*. Princeton University Press, Princeton.
[23] Lancaster, H.O. (1950). The sex ratio in sibships, with special reference to Geissler's data, *Annals of Eugenics* **15**, 153–158.
[24] Lancaster, H.O. (1994). *Quantitative Methods in Biological and Medical Sciences: A Historical Essay*. Springer-Verlag, New York.
[25] Laplace, P.S. (1781). Mémoire sur les probabilités, *Mémoires de l'Académie Royale des Sciences de Paris* **1778**, 227–332.
[26] Lindsey, J.K. & Altham, P.M.E. (1998). Analysis of the sex ratio by using overdispersion models, *Applied Statistics* **47**, 149–157.
[27] Lowe, C.R. & McKeown, T. (1950). The sex ratio of human births related to maternal age, *British Journal of Preventive and Social Medicine* **4**, 75–85.
[28] Macfarlane, A.J. & Mugford, M. (1999). *Birth Counts: Statistics of Pregnancy and Childbirth*, Vol. 1, 2nd Ed. Stationery Office, London.
[29] Maconochie, N. & Roman, E. (1997). Sex ratios: Are there natural variations within the human population? *British Journal of Obstetrics and Gynaecology* **104**, 1050–1053.
[30] McKeown, T. & Lowe, C.R. (1951). The sex ratio of stillbirths related to cause and duration of gestation. An investigation of 7,066 stillbirths, *Human Biology* **23**, 41–60.
[31] Ohmi, H., Hirooka, K. & Mochizuki, Y. (1999). Reduced ratio of male to female births in Japan, *International Journal of Epidemiology* **28**, 597–598.
[32] Poisson, S.D. (1830). Mémoire sur la proportion des naissances des filles et des garçons, *Mém. R. Acad. Sci. Inst. Fr.* **9**, 239–308.

Further Reading

James, W.H. (1987). The human sex ratio. Part I: a review of the literature, *Human Biology* **59**, 721–752.

PETER ARMITAGE

Shape Analysis

Shape is an essential ingredient of biology and medicine. The geometrical description of an object can be separated into two parts: (a) the registration information and (b) the "shape" (which is invariant under registration transformations). By registration, we mean a basic geometrical transformation of an object, for example, translation, rotation, and rescaling. Objects can be registered into a standard reference frame or with respect to each other. Equivalent names for registration include superimposition, superposition, transformation, pose, and matching.

Shape has been studied for centuries in medicine and biology, for example, Galileo's [16] study of bone shape (see Figure 1). Most studies have relied on the location of anatomical landmarks, and then **multivariate analysis** is carried out on collections of angles and ratios of lengths [14, pp. 6–7].

Geometrical shape analysis began with the independent work of Kendall [28, 30], Bookstein [4, 5] and Ziezold [51]. Subsequent developments have led to a deep differential geometric theory of shape spaces [31], as well as practical statistical approaches to analyzing objects using probability distributions of shape and **likelihood**-based inference. Summaries of the field are given by Bookstein [7], Goodall [20], Small [46], Dryden and Mardia [14], Kendall et al. [31], and Lele and Richtsmeier [37], and the main emphasis is on the shapes of labeled point set configurations.

Shape and Shape Space

General Shape

Shape is defined to be all the geometrical information that is invariant under registration transformations. Depending on the application at hand, the registration transformations may be of little interest (e.g. in comparing the shapes of bones); the registration and shape may be equally important (e.g. in object recognition (*see* **Pattern Recognition**) in **image analysis**); or the registration parameters are the primary interest (e.g. in medical image registration).

There are several common types of registration invariance that are encountered in shape analysis. The most common examples include the Euclidean similarity transformations (translation, rotation, and scale), the rigid body transformations (translation and rotation) and affine transformations (translation, rotation, and shears).

The types of objects under study are either point sets of landmarks, curves, surfaces, or solid objects. The study of shape is particularly well developed for the study of *landmarks*, which are points of meaningful biological or geometrical correspondence.

Two-dimensional Point Sets

When landmarks are available in two dimensions, the shape space and statistical analysis of shapes are relatively straightforward (e.g. [5, 14, 30]). Consider $k \geq 3$ points in a plane and use complex notation: $z_j \in \mathbb{C},\ j = 1, \ldots, k$. We remove location (by centering) $z_j - \overline{z}$, $j = 1, \ldots, k$, where $\overline{z} = \sum_{j=1}^{k} z_j / k$ is the centroid. We then identify scaled and rotated versions as an equivalence class, which is the shape of z:

$$[z] = \{\lambda(z_j - \overline{z}),\ j = 1, \ldots, k, : \\ \lambda = re^{i\theta} \in \mathbb{C} \setminus \{0\}\}. \qquad (1)$$

The shape space is, therefore, the complex projective space $\mathbb{C}P^{k-2}$ [30], which is the space of complex lines through the origin (but not including it). So, the challenge from the statistical point of view is to provide models and inferential procedures, which are appropriate for the non-Euclidean shape space.

There are several choices of shape distance that could be used and a natural choice is the Riemannian shape distance between two landmark configurations $z = (z_1, \ldots, z_k)^T$, $w = (w_1, \ldots, w_k)^T$:

$$\rho = \arccos \frac{\sum (z_j - \overline{z})^* (w_j - \overline{w})}{(\sum |z_j - \overline{z}|^2 \sum |w_j - \overline{w}|^2)^{1/2}}, \qquad (2)$$

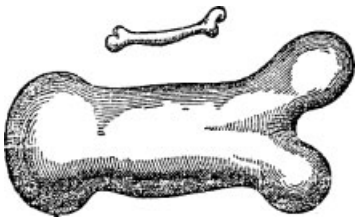

Figure 1 From Galileo (1638), illustrating the differences in shapes of the bones of small and large animals. Reproduced from [14]

where z^* is the complex conjugate of z^T. The triangle case is special since $\mathbb{C}P^1 \equiv S^2$ [29]. The shapes of triangles are represented by points on a sphere of radius $\frac{1}{2}$ and in this case, ρ is the great circle distance.

The size of a configuration is often taken to be the centroid size:

$$S(z) = \sqrt{\sum_{j=1}^{k} |z_j - \overline{z}|^2} = \|Cz\|, \tag{3}$$

where $C = I_k - 1_k 1_k{}^T/k$ is the $k \times k$ centering **matrix**, with I_k the $k \times k$ identity matrix and 1_k the column k-vector of ones. Other choices such as square root of area could be used, but are not so convenient to work with statistically. In order to represent shape, it is often convenient to specify suitable shape coordinates, for example, Bookstein shape coordinates

$$u_j^B = \frac{z_j - z_1}{z_2 - z_1} - \frac{1}{2}, \quad j = 3, \ldots, k, \tag{4}$$

where u_j^B are the complex coordinates of the landmarks after translating, rotating, and rescaling so that point 1 is sent to $-1/2 + 0i$ and point 2 is sent to $1/2 + 0i$. When shape variability is small, one can work in a tangent space to shape space, and hence use tangent space coordinates (see [14, p. 71]). For small variations, Bookstein coordinates and tangent space coordinates are approximately linearly related, and hence **multivariate normal**–based inference is approximately equivalent using either shape coordinate system [32].

Higher-dimensional Point Sets

For higher than two-dimensional point sets, the geometry is not so straightforward. Kendall et al. [31] discuss the differential geometry of shape spaces in detail, and one particular problem with the higher-dimensional shape spaces is that the spaces are not homogeneous and there are singularities. Nevertheless, we can obtain distances and work with care with such higher-dimensional spaces.

Let X be a $k \times m$ matrix of the Cartesian coordinates of the k points in m real dimensions ($k \geq m + 1$). We can consider three steps to obtaining the shape:

1. Remove location (center)
$$X_C = CX \tag{5}$$
2. Remove size (rescale)
$$Z = \frac{X_C}{S(X)} = \frac{CX}{\|CX\|}, \tag{6}$$
which is the preshape that lies on a sphere ($Z \in S^{(k-1)m-1}$).
3. Finally, the shape is obtained by identifying all rotated versions as an equivalence class, that is,
$$[X] = \{Z\Gamma : \Gamma \in SO(m)\}, \tag{7}$$
is the shape of X, where $SO(m)$ denotes the special orthogonal group of rotation matrices (i.e. orthogonal matrices with determinant 1).

Statistical analysis of shapes can be carried out on the preshape sphere subject to invariance under rotations.

Shape Distances

There are a variety of choices of shape distances for shape analysis. Some possible choices are Procrustes distances, Riemannian distance, and **Mahalanobis distance** in the Procrustes tangent space. For the m-dimensional case, some specific shape distances are:

Partial Procrustes distance:

$$\mathrm{d}_P(X_1, X_2) = \inf_{\Gamma \in SO(m)} \|Z_2 - Z_1\Gamma\|, \quad (0 \leq \mathrm{d}_P \leq \sqrt{2}). \tag{8}$$

Riemannian distance:

$$\rho(X_1, X_2) = 2 \arcsin\left(\frac{\mathrm{d}_P}{2}\right), \quad (0 \leq \rho \leq \pi/2). \tag{9}$$

Full Procrustes distance:

$$\mathrm{d}_F(X_1, X_2) = \inf_{r>0, \Gamma} \|Z_2 - rZ_1\Gamma\| = \sin\rho(X_1, X_2), \quad (0 \leq \mathrm{d}_F \leq 1). \tag{10}$$

Note that ρ reduces to equation (2) in the two-dimensional case. These distances are all quite similar for shapes, which are close together, in particular, $\mathrm{d}_F = \rho + O(\rho^3) = \mathrm{d}_P + O(\mathrm{d}_P^3)$ for small ρ, d_P.

Sometimes we may have registration transformations G, which are not a group: for example, a smoothing **spline** deformation. We can still obtain a discrepancy measure in such cases, for example, a shape discrepancy measure between Y and T:

$$D(Y, T) = \inf_{g \in G} dist(Y, g(T)), \qquad (11)$$

which is not symmetric in T and Y if g is a smoothing thin-plate spline for example.

Shape Variability

Shape Distributions

There are various approaches for modeling shape variability in biomedical objects, for example, (1) marginal/offset shape distributions, (2) distributions in preshape space with rotational symmetry, (3) distributions in shape space, and (4) distributions in a tangent space. Specifying distributions of shapes is also an important component of high-level **Bayesian** image analysis, where the shape distributions form part of the prior model.

Marginal/offset Distributions

We first consider a model for a configuration in the original space of the landmarks. In particular, we take the mean configuration μ with independent isotropic zero mean normal perturbations with variance σ^2. The marginal or offset normal model is the marginal distribution of shape after integrating out the location, rotation, and scale information (see Figure 2).

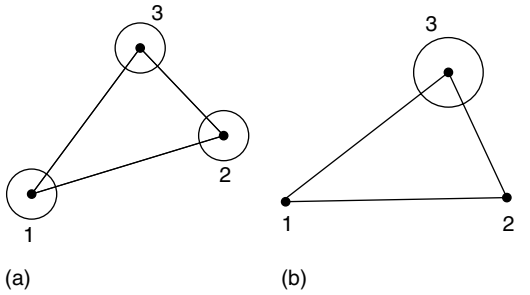

Figure 2 The offset normal shape model involves independent circular Gaussian perturbations about the mean landmarks (a), and then transforming to the shape variables such as Bookstein shape variables where the shape variability is transformed to the third vertex after fixing points 1 and 2 (b)

The offset normal shape density (with respect to uniform measure) is [12, 13, 42, 43]

$$\mathcal{L}_{k-2}(-2\kappa \cos^2 \rho(X, \mu))\exp(-2\kappa \sin^2 \rho(X, \mu))$$

where $\kappa = S(\mu)^2/(4\sigma^2)$, $S(\mu)$ is the centroid size of μ and $\mathcal{L}_j(-x) = \sum_{i=1}^{k-2} \binom{j}{i} \frac{x^i}{i!}$ is the Laguerre polynomial. The parameters are the Shape(μ): $2k-4$ mean shape parameters and κ: concentration parameter. For triangles ($k = 3$), the shape density is

$$\{1 + 2\kappa \cos^2 \rho(x, \mu)\} \exp\{-2\kappa \sin^2 \rho(x, \mu)\}.$$

General covariance matrices and higher dimensions have also be considered by Dryden and Mardia [12] and Goodall and Mardia [21].

Inference with marginal shape models can be carried out, such as testing for mean shape difference between two groups (see [14, p. 144]), although inference is not straightforward for general **covariance** structures due to overparameterization. We consider a particular test in the section "Hypothesis Testing", which assumes an isotropic covariance structure.

Distributions in Preshape and Shape Space

Other shape distributions include the complex Watson distribution [44] with density $f(z)$ proportional to

$$\exp\{-2\kappa \sin^2 \rho\}$$

and the complex Bingham distribution [32] with density $f(z)$ proportional to

$$\exp(z^* A z), \ \in \mathbb{C}S^{k-1},$$

where A is Hermitian. Both distributions are specified on the preshape sphere and have rotational symmetry, that is, $f(z) = f(e^{i\theta}z)$, and hence, they are suitable for shape modeling. The complex Watson distribution is a special case of the complex Bingham distribution, where A has just two distinct eigenvalues. The models on the preshape sphere have the advantage of being simple and tractable, but they do impose rather restrictive symmetries–isotropy for the complex Watson and complex symmetry for the complex Bingham distribution.

If the rotational information is integrated out, then the complex Watson and complex Bingham distributions can be regarded as distributions in the shape space. In the $k = 3$ triangle case, both distributions

reduce to the Fisher–von-Mises distribution on the shape sphere [39].

Procrustes Tangent Space Models

Another practical approach to specifying shape variability is to examine **principal components** from **least squares** matching of geometrical objects. Consider n objects of k landmarks in m real dimensions, that is, $T_1, \ldots, T_n$ are $k \times m$ matrices and $T_i \in \mathbb{R}^{km}$. **Procrustes** matching involves least squares matching to give $\hat{T}_j$:

$$\hat{\mu} = \arg \inf_{\mu: S(\mu)=1} \inf_{r_j>0, \Gamma_j \in SO(m), b_j} \sum_j \|\mu - r_j T_j \Gamma_j - 1_{kb_j}^T\|^2, \tag{12}$$

where the fitted configurations are $\hat{T}_j = \hat{r}_j T_j \hat{\Gamma}_j + 1_k \hat{b}_j^T$. Note that if variations are small, the shapes lie approximately in a linear space (a tangent space to shape space). Kent and Mardia [35] give a thorough description of Procrustes tangent space.

After the matching procedure is carried out, the Procrustes mean is $\hat{\mu} = \sum \hat{T}_i / n$ and the estimated covariance matrix is

$$\hat{\Sigma} = \frac{1}{n} \sum V(\hat{T}_i - \hat{\mu}) \{V(\hat{T}_i - \hat{\mu})\}^T, \tag{13}$$

where $V(T) = \text{vec}(T)$ is the stacked column vector of the columns of T. For two-dimensional objects, the Procrustes matching can be carried out using a complex eigendecomposition [32], but for higher-dimensional cases an iterative procedure such as Generalized Procrustes Analysis [22] must be carried out.

The structure of variability in the objects can be examined through the principal components of the Procrustes matched configurations, that is, through the eigendecomposition of $\hat{\Sigma}$. We can formulate the point distribution model for a two-dimensional configuration matrix $X(2k \times 1)$ of k landmarks in $\mathbb{R}^2$ based on the first p PCs as [40]

$$X = \mu + \sum_{j=1}^{p} y_j \gamma_j + \varepsilon, \tag{14}$$

where $y_j \sim N(0, \lambda_j)$, $\varepsilon \sim N_{2k}(0, \sigma^2 I)$, independently and the vectors γ_i satisfy

$$\mu^{\mathrm{T}} \gamma_j = 0, \gamma_j^{\mathrm{T}} \gamma_j = 1, \gamma_i^{\mathrm{T}} \gamma_j = 0, \quad i \neq j, \tag{15}$$

and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. In addition, for invariance under rotation and for translation, the vectors γ_i satisfy respectively

$$\gamma_j^{\mathrm{T}} \nu = 0 \text{ and } \gamma_j^{\mathrm{T}}(1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}} = 0,$$
$$\gamma_j^{\mathrm{T}}(0, \ldots, 0, 1, \ldots, 1)^{\mathrm{T}} = 0, \tag{16}$$

where $\nu = (-\beta_1, \ldots, -\beta_k, \alpha_1, \ldots, \alpha_k)^{\mathrm{T}}$ with $\mu = (\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k)^{\mathrm{T}}$. Here $p \leq \min(n-1, 2k-4)$ and p is preferably taken to be quite small, for a **parsimonious** model.

This method of shape modeling has been used to great success by Cootes et al. [10, 11] and Kent [32]. Effectively, models are specified in the tangent space to the estimated mean, and hence, they are appropriate for small variations.

Example: T2 vertebrae In Figure 3, we see an example dataset of 60 landmarks on the outline of T2 mouse vertebrae, which have been matched together using Procrustes analysis. The first two principal components of the T2 vertebrae are given in Figure 4. PC1 and PC2 explain 65% and 9% of the shape variability, and PC1 includes the effect of protrusion at the topmost part of the bone, and PC2 includes the effect of asymmetry in this part of the bone. Although the interpretation is relatively straightforward here, the PCs can be difficult to interpret in some applications and may consist of multiple effects.

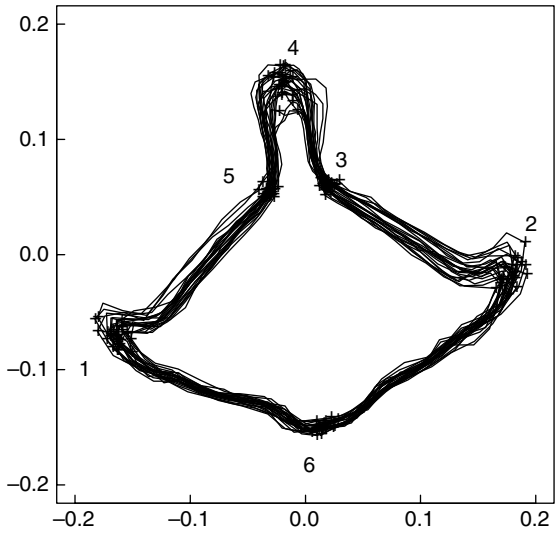

Figure 3 A dataset of 23 second thoracic mouse vertebrae that have been matched using Procrustes registration

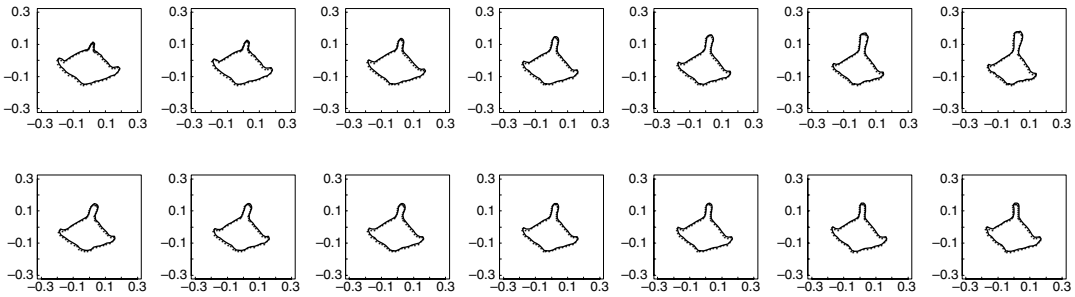

Figure 4 The first two principal components of the mouse vertebrae data. The jth row shows PCj, with the ith column displaying $\hat{\mu} + (i-4)\hat{\lambda}_j^{1/2}\hat{\gamma}_j$ where $\hat{\mu}$ is the Procrustes mean, $\hat{\gamma}_j$ is the jth PC and $\hat{\lambda}_j$ is the j eigenvalue of the tangent space covariance matrix

Hypothesis Testing

Hypothesis tests can be constructed using Procrustes methods, for example, testing for equal mean shapes in two independent groups. Consider an isotropic normal model with mean μ and transformed by an additional location, rotation and scale, that is,

$$X = \beta(\mu + E)\Gamma + 1_k\gamma^{\mathrm{T}}, \quad \mathrm{vec}(E) \sim N(0, \sigma^2 I_{km}), \tag{17}$$

where $\beta > 0$ (scale), $\Gamma \in SO(m)$ (rotation) and $\gamma \in \mathbb{R}^m$ (translation), and σ is small. We take two independent random samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ from Model (17) with means μ_1 and μ_2 respectively, and arbitrary transformation parameters for each observation. Both populations are assumed to have a common variance for each coordinate σ^2. We wish to test H_0 : Shape(μ_1) = Shape(μ_2){= Shape(μ_0)}, say, against H_1 : Shape(μ_1) $\neq$ Shape(μ_2). Let $\hat{\mu}_1$ and $\hat{\mu}_2$ be the full Procrustes means of each sample. Under H_0, with σ small, the Procrustes distances are approximately distributed as

$$\sum_{i=1}^{n_1} \mathrm{d}_F^2(X_i, \hat{\mu}_1) \sim \tau_0^2 \chi^2_{(n_1-1)M},$$

$$\sum_{i=1}^{n_2} \mathrm{d}_F^2(Y_i, \hat{\mu}_2) \sim \tau_0^2 \chi^2_{(n_2-1)M},$$

$$\mathrm{d}_F^2(\hat{\mu}_1, \hat{\mu}_2) \sim \tau_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \chi^2_M,$$

where $\tau_0 = \sigma/\delta_0$ and $\delta_0 = S(\mu_0)$ (*see* **Chi-square Distribution**). Proofs of the results can be obtained using Taylor series expansions. In addition, these statistics are approximately mutually independent (exactly in the case of the first two expressions). Hence, under H_0, we have the approximate distribution

$$F = \frac{n_1 + n_2 - 2}{{n_1}^{-1} + {n_2}^{-1}} \times \frac{\mathrm{d}_F^2(\hat{\mu}_1, \hat{\mu}_2)}{\sum_{i=1}^{n_1} \mathrm{d}_F^2(X_i, \hat{\mu}_1) + \sum_{i=1}^{n_2} \mathrm{d}_F^2(Y_i, \hat{\mu}_2)} \sim F_{M,(n_1+n_2-2)M}, \tag{18}$$

and this result is valid for small σ (*see* ***F*** **Distributions**). We reject H_0 for large values of this test statistic. This test is the two independent sample Goodall's [20] test. An alternative test is a **Hotelling's T^2** test in the tangent space to a pooled mean shape estimator, which can have general (but equal) covariance matrices for each group. Dryden and Mardia [14], Chapter 7 provide more examples of hypothesis testing using tangent space approximations.

Shape and Images

Registration and Matching

In many applications, it is important to match images or objects. An early example was D'Arcy Thompson's [48] work on describing differences between species using simple geometrical transformations. (see Figure 5)

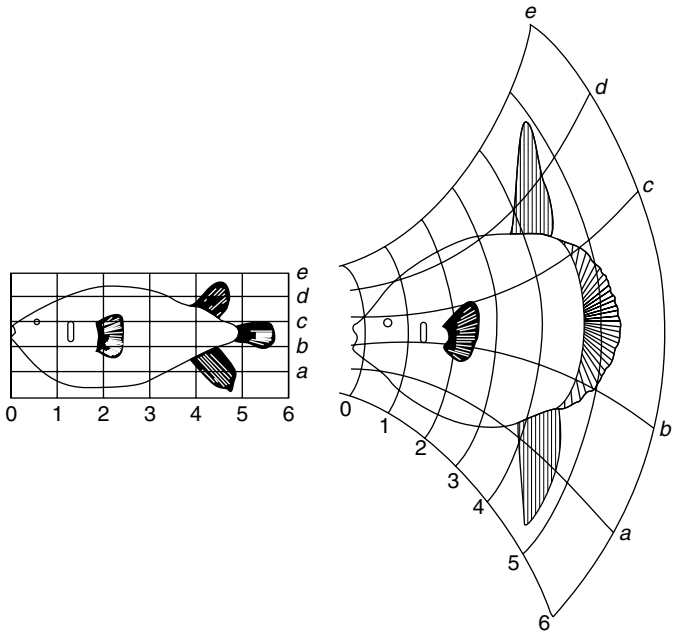

Figure 5 D'Arcy Thompson's [47] famous example of a species of fish *Diodon* being geometrically transformed into another species *Orthagoriscus*

Recent examples include the matching of electrophoresis gel images (e.g. [15]) to assess for differences between species, and the registration of a medical image to an atlas [25]. An algebraically simple but effective method for matching is the use of thin-plate splines [6, 8]. Bookstein's thin-plate spline transformations are also used in an alternative to Procrustes PCA called *relative warps*, where different aspects of bending (large and small scale) can be emphasized (e.g. see [14, Chapter 10]). More sophisticated approaches to image warping include Christensen et al. ([9]).

High-level Image Analysis

An appropriate method for high-level Bayesian image analysis is the use of deformable templates, pioneered by Grenander and colleagues [23, 24]. In many applications one has prior knowledge on the composition of the scene, and we can formulate parsimonious geometric descriptions for objects in the images. For example, in medical imaging, we can expect to know *a priori* the main subject of the image, for example, a heart or a brain. Consider our prior knowledge about the objects under study to be represented by a template S_0. Note that S_0 could be a template of a single object or many objects in a scene. A probability distribution is assigned to the parameters with density (or probability function) $\pi(S)$, which models the allowed variations S of S_0. Hence, S is a random vector representing all possible templates with associated density $\pi(S)$. Here S is a function of a finite number of parameters, say $\theta_1, \ldots, \theta_p$.

In addition to the prior model, we require an image model. Let the observed image I be the matrix of gray levels and the image model (or likelihood) be the joint probability density function of the gray levels given the parameterized objects S, written as $L(I|S)$. The likelihood expresses the dependence of the observed image on the deformed template.

By **Bayes' Theorem**, the posterior density $\pi(S|I)$ of the deformed template S given the observed image I is

$$\pi(S|I) \propto L(I|S)\pi(S). \tag{19}$$

An estimate of the true scene can be obtained from the posterior mode (the maximum *a posteriori* or MAP estimate) or the posterior mean. The posterior mode is found either by a global search, gradient descent (which is often impracticable due to the

large number of parameters) or by techniques such as simulated annealing (*see* **Computer-intensive Methods**) [17] or iterative conditional modes (ICM) [1]. Alternatively, **Markov chain Monte Carlo** (MCMC) algorithms (see, for example, [2, 18]) provide techniques for simulating from the posterior density.

There is a wide variety of possible template parameterizations that we could consider, including geometrical parameter templates, landmarks/point distribution models, graphical templates, continuous outline templates, and continuous deformation models (e.g. see [14, Chapter 11]).

Some possibilities for the image model include (i) a scientific model based on the mode of image capture (e.g. [26]), (ii) a model based on spatial smoothness (e.g. Gaussian Markov random field), (iii) a model based on measurement noise assumptions, (iv) a feature density, where particular weight is given to certain features in the image (e.g. [45]) or (v) combinations of the above. It is often convenient to also include a blurring term in the model.

The use of image analysis in various branches of medicine and biology is becoming increasingly common. Applications include the analysis of cell shapes to detect malignant versus benign tumors, the assessment of tumor volume in an MR image, and the analysis of MRI signals to relate brain activity with a repeated performed task (see [19], and the **Image Analysis and Tomography** entry in this Encyclopedia).

Discussion

An alternative but complementary approach to the use of shape space–based methods is to use methods based on Euclidean distance matrices (*see* **Similarity, Dissimilarity, and Distance Measure**) and **Multidimensional Scaling**. Visualization is more problematic with this approach, but estimation and testing procedures can be carried out with similar results for small shape variability situations (as often encountered in biology and medicine); see [37] for a review of this work.

There are many other examples of the use of shape in high-level image analysis but we have described the main ingredients of landmark or point set models; see, for example, [38, 49] for general reviews of shape measures in pattern recognition.

There is also a wide variety of work on nonlandmark shape models, such as snakes [27], continuous outline models such as the circular Gaussian Markov random field model [33], and Younes' [50] approach to continuous shape analysis with applications in high-level image analysis.

Although we have mainly concentrated on scale invariance, there are many situations in which the relationship between size and shape is of major importance, for example, in the study of **growth**. **Allometry** involves the study of the relationships between shape and size, and, in particular, in the manner in which shape depends on size. Sprent [47] provides a summary of traditional applications, and Dryden and Mardia [14, Chapter 8] provide discussion of the joint modeling and analysis of shape and size.

The notion of symmetry and bilateral symmetry in particular is very important in biology. Mardia et al. [41] and Kent and Mardia [35] explore decompositions of shape variability, which are useful for investigating bilateral symmetry.

There has been much recent discussion about the properties of shape estimators and statistical inference on shape spaces. Consistency issues for Procrustes estimators have been addressed by Kent and Mardia [34] following work by Lele [36]. Also, statistical properties of estimators of intrinsic and extrinsic mean shapes have been investigated by Bhattacharya and Patrangenaru [3], and these estimators are consistent under a very general class of models.

Finally, we conclude with remarks about the future opportunities for shape and registration methodology to be used in many applications. Geometrical invariances and registration are increasingly commonly found in medicine and biology, such as in the analysis of functions (EEG traces, mass spectrometry; *see* **Clinical Signals**), the analysis and prediction of protein structures, the analysis of electrophoretic gel images, and the analysis of medical image sequences. One of the key difficulties with many applications is that the labeling and correspondence between points/curves/surfaces is unknown. The future development of shape and registration analysis in medicine and biology for the effective analysis of these and other geometrical data is of great importance.

References

[1] Besag, J.E. (1986). On the statistical analysis of dirty pictures (with discussion), *Journal of the Royal Statistical Society, Series B* **48**, 259–302.

[2] Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems, *Statistical Science* **10**(1), 3–66. With comments and a reply by the authors.

[3] Bhattacharya, R. & Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds - I, *Annals of Statistics* **31**, 1–29.

[4] Bookstein, F.L. (1978). *The Measurement of Biological Shape and Shape Change*, Lecture Notes on Biomathematics, Vol. 24, Springer-Verlag, New York.

[5] Bookstein, F.L. (1986). Size and shape spaces for landmark data in two dimensions (with discussion), *Statistical Science* **1**, 181–242.

[6] Bookstein, F.L. (1989). Principal warps: thin-plate splines and the decomposition of deformations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 567–585.

[7] Bookstein, F.L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge.

[8] Bookstein, F.L. (1997). Shape and the information in medical images: a decade of the morphometric synthesis, *Computer Vision and Image Understanding* **66**, 97–118.

[9] Christensen, G., Rabbitt, R.D. & Miller, M.I. (1996). Deformable templates using large deformation kinematics, *IEEE Transactions on Image Processing* **5**, 1435–1447.

[10] Cootes, T.F., Taylor, C.J., Cooper, D.H. & Graham, J. (1992). Training models of shape from sets of examples, in *British Machine Vision Conference*, D.C. Hogg & R.D. Boyle, eds. Springer-Verlag, Berlin, pp. 9–18.

[11] Cootes, T.F., Taylor, C.J., Cooper, D.H. & Graham, J. (1994). Image search using flexible shape models generated from sets of examples, in *Statistics and Images: Vol. 2*, K.V. Mardia, ed. Carfax, Oxford, pp. 111–139.

[12] Dryden, I.L. & Mardia, K.V. (1991). General shape distributions in a plane, *Advances in Applied Probability* **23**, 259–276.

[13] Dryden, I.L. & Mardia, K.V. (1992). Size and shape analysis of landmark data, *Biometrika* **79**, 57–68.

[14] Dryden, I.L. & Mardia, K.V. (1998). *Statistical Shape Analysis*. Wiley, Chichester.

[15] Dryden, I.L. & Walker, G. (1999). Highly resistant regression and object matching, *Biometrics* **55**, 820–825.

[16] Galileo (1638). *Discorsi e dimostrazioni matematiche, informo a due nuoue scienze attenti alla mecanica i movimenti localli*, appresso gli Elsevirii; Opere VIII.

[17] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions of Pattern Analysis and Machine Intelligence* **6**, 721–741.

[18] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

[19] Glasbey, C.A. & Horgan, G.W. (1995). *Image Analysis for the Biological Sciences*. Wiley, Chichester.

[20] Goodall, C.R. (1991). Procrustes methods in the statistical analysis of shape (with discussion), *Journal of the Royal Statistical Society, Series B* **53**, 285–339.

[21] Goodall, C.R. & Mardia, K.V. (1993). Multivariate aspects of shape theory, *Annals of Statistics* **21**, 848–866.

[22] Gower, J.C. (1975). Generalized Procrustes analysis, *Psychometrika* **40**, 33–50.

[23] Grenander, U. (1994). *General Pattern Theory*. Clarendon Press, Oxford.

[24] Grenander, U. & Keenan, D.M. (1993). Towards automated image understanding, in *Statistics and Images: Vol. 1*, K.V. Mardia & G.K. Kanji, eds. Carfax, Oxford, pp. 89–103.

[25] Grenander, U. & Miller, M.I. (1994). Representations of knowledge in complex systems (with discussion), *Journal of the Royal Statistical Society, Series B* **56**, 549–603.

[26] Husby, O., Lie, T., Lango, T. & Rue, H. (2001). Bayesian 2-D convolution: a model for diffuse ultrasound scattering, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **48**, 121–130.

[27] Kass, M., Witkin, A. & Terzopoulos, D. (1988). Snakes: active contour models, *International Journal of Computer Vision* **1**, 321–331.

[28] Kendall, D.G. (1977). The diffusion of shape, *Advances in Applied Probability* **9**, 428–430.

[29] Kendall, D.G. (1983). The shape of Poisson-Delaunay triangles, in *Studies in Probability and Related Topics*, M.C. Demetrescu & M. Iosifescu, eds. Nagard, Montreal, pp. 321–330.

[30] Kendall, D.G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces, *Bulletin of the London Mathematical Society* **16**, 81–121.

[31] Kendall, D.G., Barden, D., Carne, T.K. & Le, H. (1999). *Shape and Shape Theory*. Wiley, Chichester.

[32] Kent, J.T. (1994). The complex Bingham distribution and shape analysis, *Journal of the Royal Statistical Society, Series B* **56**, 285–299.

[33] Kent, J.T., Dryden, I.L. & Anderson, C.R. (2000). Using circulant symmetry to model featureless objects, *Biometrika* **87**(3), 527–544.

[34] Kent, J.T. & Mardia, K.V. (1997). Consistency of Procrustes estimators, *Journal of the Royal Statistical Society, Series B* **59**, 281–290.

[35] Kent, J.T. & Mardia, K.V. (2001). Shape, tangent projections, and bilateral symmetry, *Biometrika* **88**, 469–485.

[36] Lele, S. (1993). Euclidean distance matrix analysis (EDMA): estimation of mean form and mean form difference, *Mathematical Geology* **25**, 573–602.

[37] Lele, S.R. & Richtsmeier, J.T. (2001). *An Invariant Approach to the Statistical Analysis of Shapes*. Chapman & Hall/CRC, Boca Raton.

[38] Loncaric, S. (1998). A survey of shape analysis techniques, *Pattern Recognition* **31**, 983–1001.

[39] Mardia, K.V. (1989). Shape analysis of triangles through directional techniques, *Journal of the Royal Statistical Society, Series B* **51**, 449–458.

[40] Mardia, K.V. (1997). Bayesian image analysis, *Journal of Theoretical Medicine* **1**, 63–77.

[41] Mardia, K.V., Bookstein, F.L. & Moreton, I.J. (2000). Statistical assessment of bilateral symmetry of shapes, *Biometrika* **87**(2), 285–300.

[42] Mardia, K.V. & Dryden, I.L. (1989a). Shape distributions for landmark data, *Advances in Applied Probability* **21**, 742–755.

[43] Mardia, K.V. & Dryden, I.L. (1989b). The statistical analysis of shape data, *Biometrika* **76**, 71–282.

[44] Mardia, K.V. & Dryden, I.L. (1999). The complex Watson distribution and shape analysis, *Journal Of The Royal Statistical Society Series B-Statistical Methodology* **61**(4), 913–926.

[45] Mardia, K.V., McCulloch, C., Dryden, I.L. & Johnson, V. (1997). Automatic scale-space method of landmark detection, in *Proceedings of the Leeds Annual Statistics Research Workshop*, K.V. Mardia, C.A. Gill & R.G. Aykroyd, eds. University of Leeds Press, Leeds, pp. 17–29.

[46] Small, C.G. (1996). *The Statistical Theory of Shape*. Springer, New York.

[47] Sprent, P. (1972). The mathematics of size and shape, *Biometrics* **28**, 23–37.

[48] Thompson, D.W. (1917). *On Growth and Form*. Cambridge University Press, Cambridge.

[49] Veltkamp, R.C. (2001). Shape Matching: Similarity Measures and Algorithms. UU-CS, (Ext. r. no. 2001–03). Utrecht, The Netherlands: Utrecht University: Information and Computing Sciences.

[50] Younes, L. (1998). Computable elastic distances between shapes, *SIAM Journal On Applied Mathematics* **58**(2), 565–586.

[51] Ziezold, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces, *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*, Vol. A, Academia: Czechoslovak Academy of Sciences, Prague, pp. 591–602.

IAN L. DRYDEN

Sheppard's Corrections

Suppose that the discrete variable Y is a continuous variable X rounded to the nearest multiple of some positive h. Generally, the sample **moments** of Y are biased estimates of the population moments of X. Sheppard's corrections [2] are simple formulas, valid for small h, that eliminate this bias (*see* **Unbiasedness**).

For example, suppose that X is normal with mean μ and variance σ^2. If the sample average and variance of Y are $\overline{y}$ and s^2, then the Sheppard-corrected estimates are $\hat{\mu}_{\mathrm{SC}} = \overline{y}$ and $\hat{\sigma}^2_{\mathrm{SC}} = s^2 - h^2/12$. That is, the rounding has negligible effect on the mean but causes a positive bias in the variance.

One derivation of Sheppard's corrections relates the moments of Y to the moments of X. Letting $\mu_k = \mathrm{E}[X^k]$ and $\nu_k = \mathrm{E}[Y^k]$, the Euler–Maclaurin quadrature theorem implies

$$\nu_k = \sum_{m=0}^{[k/2]} \left(\frac{h}{2}\right)^{2m} \binom{k}{2m} \frac{\mu_{k-2m}}{2m+1} + R,$$

where $[\cdot]$ is the greatest integer function. The remainder R becomes small as $h \to 0$ if the density and several of its derivatives go to zero at the limits of the range of X, or at some pseudo-limits that contain most of the probability. The normal distribution is one special case.

The corrected mean and variance are also approximate **maximum likelihood** estimators. The **likelihood** based on Y has a series representation in terms of the density of X (and its derivatives) evaluated at the observed y. Starting from $\mu = \overline{y}$ and $\sigma^2 = s^2$, executing one Newton–Raphson or **EM** step leads again to Sheppard's corrections (*see* **Optimization and Nonlinear Equations**). This result holds specifically for normal X, but also for other distributions satisfying certain regularity conditions.

Although the uncorrected s^2 overestimates the variance of X, the sampling variance of $\hat{\mu}_{\mathrm{SC}}$ is not $\hat{\sigma}^2_{\mathrm{SC}}/n$ but s^2/n. Thus the corrected variance is of little value unless one specifically wants to estimate σ^2. The basic formulas are valid for the normal but possibly not for other distributions. For example, with **uniform** data, the proper correction to s^2 is to *add* $h^2/12$.

For further details and a historical review, see [1].

References

[1] Heitjan, D.F. (1989). Inference from grouped continuous data: a review, *Statistical Science* **4**, 164–183.

[2] Sheppard, W.F. (1897). On the calculation of the average square, cube &c., of a large number of magnitudes, *Journal of the Royal Statistical Society* **60**, 698–703.

(*See also* **Grouped Data**)

DANIEL F. HEITJAN

Shrinkage Estimation

Starting from the work of Stein [32] (*see* **James–Stein Estimator**), the topic of shrinkage estimation has received an enormous amount of attention in the statistical literature. The original shrinkage estimators were developed for the case of estimating the mean of a **multivariate normal distribution** under squared error loss, based on observing $\mathbf{X} = \mathbf{x}$, with $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\theta}, \mathbf{I})$, a p-dimensional normal random variable. However, results on shrinkage have been generalized to the extent that these estimators can now be applied routinely to actual problems.

In terms of practical applicability, the direction pointed out by Lindley [29] has proved quite fruitful. Lindley showed that one could shrink toward a point chosen by the data, and demonstrated, for $p \geq 4$, the minimaxity of the estimator

$$\mathbf{d}^{\mathrm{L}}(\mathbf{x}) = \overline{x}\mathbf{1} + \left(1 - \frac{p-3}{|\mathbf{x} - \overline{x}\mathbf{1}|^2}\right)(\mathbf{x} - \overline{x}\mathbf{1}),$$

where $\mathbf{1}$ is a column vector of 1s and $|\mathbf{x} - \overline{x}\mathbf{1}|^2 = \sum(x_i - \overline{x})^2$. A fourth dimension is needed here, rather than the three dimensions needed for the minimaxity of the James–Stein estimator, because we are now shrinking to a one-dimensional subspace, rather than the zero-dimensional point toward which the James–Stein estimator shrinks. The idea of shrinking toward a subspace has enhanced the applicability of shrinkage estimators, and has connected them with **empirical Bayes** estimation. Much of this topic was developed in a sequence of papers by Efron & Morris [13–15], where the connection with **minimax** estimation is explored thoroughly. A comprehensive treatment of theory and applications of empirical Bayes methods is given by Morris [30], and less technical introductions are given by Casella [9, 10].

On the more theoretical side, in the normal case, Strawderman [34] was the first to exhibit *proper Bayes minimax* estimators – estimators that not only dominated $\mathbf{X}$, but were themselves proper Bayes and admissible (*see* **Bayesian Methods**). These estimators have the form of Baranchik's estimators (*see* **James–Stein Estimator**), and a particular one is given by

$$\mathbf{d}^{\mathrm{S}}(\mathbf{x}) = \left(1 - \frac{c(|\mathbf{x}|)}{|\mathbf{x}|^2}\right)\mathbf{x}, \tag{1}$$

where

$$c(|\mathbf{x}|) = p + 2 - \frac{2\exp\left(-\frac{1}{2}|\mathbf{x}|^2\right)}{\int_0^1 \lambda^{p/2}\exp(-\lambda|\mathbf{x}|^2/2)\,\mathrm{d}\lambda}.$$

The estimator (1) can be derived from the Bayes model

$$\begin{aligned}
&\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\theta}, \mathbf{I}),\\
&\boldsymbol{\theta} \sim \mathrm{N}_p[\mathbf{0}, \lambda^{-1}(1-\lambda)\mathbf{I}],\\
&\lambda \sim \textbf{uniform}\ (0, 1),
\end{aligned}$$

which is a proper Bayes model if $p \geq 5$.

Thus far we have discussed only the normal distribution; however, domination of the usual estimator by a shrinkage estimator occurs in many other situations, even in discrete families. For example, if $X_i \sim \textbf{Poisson}(\lambda_i)$, $i = 1, \ldots, p$, $p \geq 2$, are independent, and the **loss function** is given by

$$L(\boldsymbol{\lambda}, \mathbf{d}) = \sum_{i=1}^{r} \frac{(\lambda_i - d_i)^2}{\lambda_i},$$

then, as Clevensen & Zidek have shown [12], the estimator

$$\mathbf{d}^{\mathrm{CZ}}(\mathbf{x}) = \left[1 - \frac{c\left(\sum x_i\right)}{\sum x_i + b}\right]\mathbf{x}$$

is minimax if

1. $c(\cdot)$ is nondecreasing,
2. $0 \leq c(\cdot) \leq 2(p-1)$, and
3. $b \geq p - 1$.

This result highlights two differences between the normal and Poisson cases. First, domination only requires $p \geq 2$, and the loss is now scaled squared error, instead of ordinary squared error. (The fact that we only now require $p \geq 2$ is discussed by Brown [6], who described it as a "dimension doubling" phenomenon; see also [27].) Shrinkage estimators continue to dominate in many other discrete families. Using a different method of proof from that of Clevensen & Zidek [12], Hwang [24] (see also [19]) demonstrated dominance of shrinkage estimators in many discrete families.

An interesting exception is the **binomial distribution**, where Johnson [26] demonstrated that no

shrinkage estimator will dominate the usual estimator. This result was extended by Brown [8], and later Guttmann [21, 22] established the somewhat surprising result that shrinkage estimators can never dominate in any problem with a finite sample space. (Domination by shrinkage is often referred to as *the Stein effect*, so there is no Stein effect in problems with finite sample spaces.)

Even with this limitation from finite sample spaces, shrinkage estimation has played a large role in developments in both theory and practice. On the practical side, the previously mentioned connection with empirical Bayes methods (and also hierarchical Bayes methods) has allowed the application of shrinkage estimators in a wide variety of problems. The theoretical developments have also been numerous, and have sometimes been accompanied by advances in the mathematical attack on the problem.

In the normal case all restrictions on the **covariance matrix** can be removed (see, for example, [20]). Outside the normal case, shrinkage estimators exist for spherically symmetric distributions [11, 4], and some results apply to the entire **exponential family** [23]. For the case of estimating a **gamma** scale parameter, Berger [2] obtained some interesting domination results, including domination by some "expanders" rather than shrinkers. The implications of this are further discussed by Brown [7].

The theory of *superharmonic functions*, a type of multivariate concave function, which was originally applied to minimax estimation by Stein [33], has also been valuable in extending shrinkage domination. George [17, 18] used it to establish dominion by *multiple shrinkage* estimators – estimators that can shrink to more than one target. More recently, Fourdrinier et al. [16] applied it to construct new families of proper Bayes minimax estimators based on **Cauchy** prior distributions.

Although the use of squared error loss is analytically convenient, shrinkage domination extends to other losses as well. For example, variations on squared error loss that allow weight matrices can easily be accommodated. Domination under an entire class of weighted squared error loss functions can be achieved [5, 31], as well as more general universal domination [25]. Other results include those of Brandwein & Strawderman [3], who established domination results for concave losses, and Berger [1], who derived necessary conditions for dominance under a wide variety of losses. A more complete discussion of this, and many other aspects of shrinkage estimation, can be found in [28].

References

[1] Berger, J. (1976). Tail minimaxity in location vector problems and its applications, *Annals of Statistics* **4**, 33–50.

[2] Berger, J. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters, *Annals of Statistics* **8**, 545–571.

[3] Brandwein, A.C. & Strawderman, W.E. (1980). Minimax estimation of location parameters for spherically symmetric distributions with concave loss, *Annals of Statistics* **8**, 279–284.

[4] Brandwein, A.C. & Strawderman, W.E. (1990). Stein estimation: the spherically symmetric case, *Statistical Science* **5**, 356–369.

[5] Brown, L.D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters), *Journal of the American Statistical Association* **70**, 417–427.

[6] Brown, L.D. (1979). A heuristic method for determining admissibility of estimators – with applications, *Annals of Statistics* **7**, 960–994.

[7] Brown, L.D. (1980). Examples of Berger's phenomenon in the estimation of independent normal means, *Annals of Statistics* **8**, 572–585.

[8] Brown, L.D. (1981). A complete class theorem for statistical problems with finite sample spaces, *Annals of Statistics* **9**, 1289–1300.

[9] Casella, G. (1985). An introduction to empirical Bayes data analysis, *American Statistician* **39**, 83–87.

[10] Casella, G. (1992). Illustrating empirical Bayes methods, *Chemolab* **16**, 107–125.

[11] Cellier, D., Fourdrinier, D. & Robert, C. (1989). Robust shrinkage estimators of the location parameter for elliptically symmetric distributions, *Journal of Multivariate Analysis* **29**, 39–42.

[12] Clevensen, M.L. & Zidek, J. (1975). Simultaneous estimation of the mean of independent Poisson laws, *Journal of the American Statistical Association* **70**, 698–705.

[13] Efron, B. & Morris, C.N. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach, *Journal of the American Statistical Association* **68**, 117–130.

[14] Efron, B. & Morris, C. (1973). Combining possibly related estimation problems (with discussion), *Journal of the Royal Statistical Society, Series B* **35**, 379–421.

[15] Efron, B. & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations, *Journal of the American Statistical Association* **70**, 311–319.

[16] Fourdrinier, D., Strawderman, W.E. & Wells, M.T. (1996). On the Construction of Proper Bayes Minimax Estimators, *Technical Report*, Statistics Center, Cornell University.

[17] George, E.I. (1986). Minimax multiple shrinkage estimators, *Annals of Statistics* **14**, 188–205.

[18] George, E.I. (1986). Combining minimax shrinkage estimators, *Journal of the American Statistical Association* **81**, 437–445.

[19] Ghosh, M., Hwang, J.T. & Tsui, K.-W. (1983). Construction of improved estimators in multiparameter estimation for discrete exponential families, *Annals of Statistics* **11**, 351–367.

[20] Gleser, L.J. (1986). Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix, *Annals of Statistics* **14**, 1625–1633.

[21] Guttman, S. (1982). Stein's paradox is impossible in problems with finite parameter spaces, *Annals of Statistics* **10**, 1017–1020.

[22] Guttman, S. (1982). Stein's paradox is impossible in the nonanticipative context, *Journal of the American Statistical Association* **77**, 934–935.

[23] Hudson, H.M. (1978). A natural identity for exponential families with applications in multiparameter estimation, *Annals of Statistics* **6**, 473–484.

[24] Hwang, J.T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases, *Annals of Statistics* **10**, 857–867.

[25] Hwang, J.T. (1985). Universal domination and stochastic domination: estimation simultaneously under a broad class of loss functions, *Annals of Statistics* **13**, 295–314.

[26] Johnson, B. McK. (1971). On the admissible estimators for certain fixed sample binomial problems, *Annals of Mathematical Statistics* **42**, 1579–1587.

[27] Johnstone, I. & MacGibbon, K.B. (1992). Minimax estimation of a constrained Poisson vector, *Annals of Statistics* **20**, 807–831.

[28] Lehmann, E.L. & Casella, G. (1997). *Theory of Point Estimation*, 2nd Ed. Springer-Verlag, New York.

[29] Lindley, D.V. (1962). Discussion of the paper by Stein, *Journal of the Royal Statistical Society, Series B* **24**, 265–296.

[30] Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion) *Journal of the American Statistical Association* **78**, 47–65.

[31] Shinozaki, N. (1980). Estimation of a multivariate normal mean with a class of quadratic loss functions, *Journal of the American Statistical Association* **75**, 973–976.

[32] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution, in *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 197–206.

[33] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *Annals of Statistics* **9**, 1135–1151.

[34] Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean, *Annals of Mathematical Statistics* **42**, 385–388.

(*See also* **Decision Theory**; **Shrinkage**)

GEORGE CASELLA

Shrinkage

The problem of estimating the mean of a **normal distribution** is central to the practice of statistics. This simple problem is at the heart of many of the most common procedures used today, such as the **analysis of variance** or **regression.** If we have a **random sample** $X_1, \ldots, X_n$ from a normal population with mean μ and variance σ^2, then the natural estimator of μ is the sample mean $\overline{X} = (1/n)\sum_i X_i$. A question of interest is whether this estimator is the *best* estimator of the parameter μ.

When assessing the performance of an estimator; in particular, whether it is best, it is necessary to have a criterion against which to measure it. A most popular measure is *squared error loss*, where we measure the performance of an estimator d of a parameter θ by the function

$$L(\theta, d) = (\theta - d)^2, \tag{1}$$

which is called a **loss function**.

Under the loss function (1), $\overline{X}$ has many optimality properties. For example, it is a **minimax** *estimator* of μ, meaning that of all estimators of μ, its loss has the smallest maximum value. There are other properties that $\overline{X}$ enjoys, including the property of *admissibility*. An estimator d of a parameter θ is an *admissible estimator* of θ under the loss $L(\theta, d)$ if there is no other estimator d' that satisfies

$$\mathrm{E}_\theta[L(\theta, d)] \geq \mathrm{E}_\theta[L(\theta, d')], \quad \text{for all } \theta,$$

with strict inequality for some values of θ.

Is $\overline{X}$ an admissible estimator of θ? Hodges & Lehmann [3] and Blyth [1] showed that it was. That is, there is no estimator that is uniformly better. However, if the problem is made slightly more complex, then an interesting result unfolds. Suppose that, instead of estimating the mean of one normal population, we are interested in estimating the mean of many normal populations; that is, we observe $\overline{X}_k$, $k = 1, \ldots, p$, where $\overline{X}_k$ is the mean of n observations from a normal population with mean μ_k and variance σ^2, and we want to estimate $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$. The loss of an estimator $\mathbf{d} = (d_1, \ldots, d_p)$ is measured by the sum of squared errors, that is

$$L(\boldsymbol{\mu}, \mathbf{d}) = \sum_{k=1}^{p} (\mu_k - d_k)^2, \tag{2}$$

and we ask if $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_p)$ is still an admissible estimator of $\boldsymbol{\mu}$. For $p = 2$, Stein [6] showed that the answer is Yes, but he also showed that, if $p > 2$, then the answer is No. Using arguments based on the idea that, for estimating more than 2 means, $\overline{\mathbf{X}}$ tends to be "too long", Stein demonstrated the existence of a better estimator – a *shrinkage* estimator. Such an estimator shrinks the vector $(\overline{X}_1, \ldots, \overline{X}_p)$ toward a specific point in the parameter space. In [4], it was shown that the estimator

$$\mathbf{d}^{\mathrm{JS}}(\overline{\mathbf{X}}) = \left[1 - \frac{(p-2)\sigma^2}{|\overline{\mathbf{X}}|^2}\right]\overline{\mathbf{X}},$$

which shrinks $\overline{\mathbf{X}}$ toward $\mathbf{0}$, uniformly dominates $\overline{\mathbf{X}}$ as an estimator of $\boldsymbol{\mu}$ under the loss (2), so $\overline{\mathbf{X}}$ is not an admissible estimator.

This extremely surprising result has resulted in an enormous amount of research in areas such as **decision theory** and **empirical Bayes** analysis. Many superior procedures have been derived since. See the review article by Brandwein & Strawderman [2], or the book by Lehmann & Casella [5].

References

[1] Blyth, C.R. (1951). On minimax statistical decision procedures and their admissibility, *Annals of Mathematical Statistics* **22**, 22–42.

[2] Brandwein, A.C. & Strawderman, W.E. (1990). Stein estimation: the spherically symmetric case, *Statistical Science* **5**, 356–369.

[3] Hodges, J.L., Jr & Lehmann, E.L. (1951). Some applications of the Cramér-Rao inequality, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 13–22.

[4] James, W. & Stein, C. (1961). Estimation with quadratic loss, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 311–319.

[5] Lehmann, E.L. & Casella, G. (1997). *Theory of Point Estimation*, 2nd Ed. Springer-Verlag, New York.

[6] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 197–206.

(*See also* **James–Stein Estimator**; **Shrinkage Estimation**)

GEORGE CASELLA

Sign Tests

The sign test is a simple distribution-free test (*see* **Nonparametric Methods**) that can be applied easily in a variety of situations. Observations, which may be difference scores, are replaced by their positive or negative signs and analyzed using a binomial test. The specific applications discussed here are a location test for the median, the Cox–Stuart test for trend, and the McNemar test for correlated proportions. Bradley [2] provides a survey of these and other sign tests. These tests are also discussed in more detail in many books on nonparametric statistics, including [1, 3, 5, 7–10, 13–15].

Binomial Test

Consider data consisting of n independent dichotomous (**binary**) trials, where the outcome of each trial is classified as a success or a failure. If the probability of success, p, remains constant from trial to trial, then B, the total number of successes, has the **binomial distribution** with parameters n and p:

$$\Pr(B = b) = \binom{n}{b} p^b (1-p)^{n-b}, \quad b = 0, \ldots, n. \tag{1}$$

In the sign test, we test the **null hypothesis** $H_0 : p = 0.5$ against one- or two-sided **alternatives**. For $H_a : p > 0.5$, the ***P* value** of the test is $\Pr(B \geq b)$ for the observed value b. For $H_a : p < 0.5$, the P value is $\Pr(B \leq b)$; for $H_a : p \neq 0.5$, we double the smaller of the tail probabilities corresponding to the observed b.

For large n, the P value can be approximated using the fact that the null distribution of $Z = (B - n/2)/[(n/4)]^{1/2}$ is approximately standard normal. The approximation can be improved by incorporating a continuity correction of ± 0.5 in the numerator of Z; the sign of the correction is chosen to increase the probability being calculated.

Location Test for the Median

Let $Z_1, \ldots, Z_n$ be independent observations from a distribution with **median** θ. Assume that the distribution is continuous at θ, so that $\Pr(Z_i = \theta) = 0$. Then, $\Pr(Z_i > \theta) = \Pr(Z_i < \theta) = 0.5$. The Z_is may be either a single sample or differences of paired data, $Z_i = Y_i - X_i$.

To test $H_0 : \theta = \theta_0$, define a success as $Z_i > \theta_0$ and a failure as $Z_i < \theta_0$. Under H_0, the number of successes has a binomial distribution with parameters n and $p = 0.5$. The sign test for the median consists of a binomial test of $p = 0.5$ based on the signs of the $Z_i - \theta_0$. The alternative hypothesis $H_a : \theta > \theta_0$ corresponds to $H_a : p > 0.5$, and $H_a : \theta < \theta_0$ corresponds to $H_a : p < 0.5$.

Although continuity at the median implies that $\Pr(Z_i - \theta = 0) = 0$, in practice, $Z_i - \theta = 0$ can occur. A commonly recommended solution is to discard the zero values and reduce n accordingly [9]. This and other methods of handling zeros are discussed by Bradley [1, 2] and Emerson & Simon [6].

The exact **power** of the sign test can be computed using the binomial distribution with parameters n and $p = \Pr(Z_i > \theta_0)$. Randles & Wolfe [14] present results of a Monte Carlo study comparing the small-sample power of the sign test, **Wilcoxon signed-rank test**, and **Student's *t* test** for the **uniform**, **normal**, **logistic**, double **exponential**, and **Cauchy** distributions. The sign test is the most powerful of the three tests for the Cauchy distribution and is superior to the t test for the double exponential. Otherwise, the sign test has the poorest performance in the situations studied.

Noether [12] gives a formula for determining the sample size required to obtain the desired power against specified alternatives (*see* **Sample Size Determination**). The **asymptotic relative efficiency** of the sign test relative to Student's t test is 0.637 for the normal distribution, 0.333 for the uniform, 0.822 for the logistic, 2.0 for the double exponential, and at least 1/3 for any continuous, unimodal symmetric distribution [14, p. 168].

The sign test can be inverted to construct a **confidence interval** for the median θ. The lower and upper limits of the confidence interval are appropriately chosen **order statistics** of the Z_is [9].

Other Sign Tests

Sign Test for Trend

Cox & Stuart [4] proposed a sign test for upward or downward trend in a sequence $X_1, \ldots, X_{n'}$

of continuous observations. The observations are grouped into pairs $(X_1, X_{1+c}), (X_2, X_{2+c}), \ldots, (X_{n'-c}, X_{n'})$, where $c = n'/2$ if n' is even, and $c = (n'+1)/2$ if n' is odd. (If n' is odd, then the middle observation is discarded.) A pair (X_i, X_{i+c}) with $X_i < X_{i+c}$ is a success; a pair with $X_i > X_{i+c}$ is a failure. A preponderance of successes suggests an upward trend, and a preponderance of failures, a downward trend. Under the null hypothesis of no trend, the number of successes is binomial with parameters n and $p = 0.5$, where n is the number of pairs formed.

A modification of the Cox–Stuart test can be used to test for **correlation** between continuous variables X and Y. The bivariate observations (X, Y) are ordered with respect to increasing values of X (or Y), and the trend test is applied to the corresponding values of Y (or X).

McNemar Test

The **McNemar** [11] **test** for correlated proportions is usually considered a sign test, although no signs are actually involved. The test is used to analyze paired dichotomous data (*see* **Matched Pairs With Categorical Data**). For example, subjects and matched controls might be classified according to the presence or absence of a symptom, or subjects might be asked a Yes/No question both before and after an intervention. In the latter example, those subjects who answer Yes at both times and those who answer No at both times contribute no information about the direction of any change caused by the intervention. Thus, the test is based on only those n subjects who change their answers. Under the null hypothesis of no effect, the number of subjects who change from No to Yes is binomial with parameters n and $p = 0.5$.

References

[1] Bradley, J.V. (1968). *Distribution-Free Statistical Tests*. Prentice-Hall, Englewood Cliffs.

[2] Bradley, J.V. (1969). A survey of sign tests based on the binomial distribution, *Journal of Quality Technology* **1**, 89–101.

[3] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Ed. Wiley, New York.

[4] Cox, D.R. & Stuart, A. (1955). Some quick tests for trend in location and dispersion, *Biometrika* **42**, 80–95.

[5] Daniel, W.W. (1990). *Applied Nonparametric Statistics*, 2nd Ed. PWS-Kent, Boston.

[6] Emerson, J.D. & Simon, G.A. (1979). Another look at the sign test when ties are present: the problem of confidence intervals, *American Statistician* **33**, 140–142.

[7] Gibbons, J.D. & Chakraborti, S. (2003). *Nonparametric Statistical Inference*, 4th Ed. Marcel Dekker, New York.

[8] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

[9] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Ed. Wiley, New York.

[10] Lehmann, E.L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, Rev. 1st Ed. Prentice Hall, Upper Saddle River, NJ.

[11] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–157.

[12] Noether, G.E. (1987). Sample size determination for some common nonparametric tests, *Journal of the American Statistical Association* **82**, 645–647.

[13] Pratt, J.W. & Gibbons, J.D. (1981). *Concepts of Nonparametric Theory*. Springer-Verlag, New York.

[14] Randles, R.H. & Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*, Reprint Ed. with Corrections. Krieger, Malabar, FL.

[15] Sprent, P. & Smeeton, N.C. (2001). *Applied Nonparametric Statistical Methods*, 3rd Ed. Chapman & Hall/CRC, Boca Raton, FL.

E. JACQUELIN DIETZ

Signed-rank Statistics

Signed-rank statistics are commonly employed for the analysis of a single sample of data, or for the analysis of matched pairs. Procedures based on signed-rank statistics are nonparametric as they permit valid statistical inferences for data from broad families of probability distributions, such as all continuous, symmetric distributions.

The most typical biostatistical setting in which signed-rank statistics are constructed is for the matched-pairs study. In this instance intrapair differences are calculated, and these differences are **ranked** according to their absolute (unsigned) values using the integers $1, 2, \ldots, N$. Here, 1 is assigned to the smallest absolute difference, 2 is assigned to the next largest absolute difference, and so forth, until the largest absolute difference is assigned the rank of N. After such ranking of the absolute differences, the sign of the difference is restored to the rank. Statistics calculated from these signed-ranks are termed signed-rank statistics. A well-known statistic in this class is the Wilcoxon signed-rank statistic, on which the corresponding **Wilcoxon signed-rank test** procedure is based. This test procedure was introduced by Wilcoxon [4], and is a simple and powerful competitor to the **paired *t* test**.

To illustrate the calculation of signed rank, consider the following study in which resting heart rates are recorded for nine healthy persons. Measurements are made both before and six months after initiation of an aerobic exercise program. The data are displayed in Table 1. Each heart rate in the Table is the mean of five resting heart rates for each person at each time period. The effectiveness of the exercise program in reducing the resting heart rate can be assessed by examination of the direction and the magnitude of the intrapair differences. The positive signed ranks are those associated with the positive differences which occur for all subjects, except subjects 3 and 5. Both of these persons have negative signed ranks.

Two problems can arise in constructing signed ranks; these are zero differences and differences equal to one another in absolute value. It is customary in signed-rank analyses to discard the zero differences prior to ranking, and then reduce the sample size accordingly by the number of such zero differences; see [3] and [1] for further discussion of the effects of this convention. Tied absolute differences are usually resolved by midranking, wherein the tied values each receive the mean value of the ranks they would have been assigned had they been slightly different from one another. For example, suppose the data for subject 2 in Table 1 had been 76 and 72 for the baseline and six-month heart rates, respectively. The difference would then be $+4$, which would be tied in absolute value with the -4 difference calculated for subject 3. In midranking, the rank of 2.5, i.e. $(2 + 3)/2$, would be assigned to each of these absolute differences. The rank of 4 would be assigned to the next largest absolute difference, etc. Data with many such ties require special treatment and adjustments in the analyses; standard text books on nonparametric methods, such as Hollander & Wolfe [2], detail the specifics of these adjustments.

A most common distributional assumption in the construction of signed-rank statistics is that the

Table 1 Resting heart rate of nine people before and after initiation of an exercise regimen

Subject	Heart rate at baseline (y_i)	Heart rate at six months (x_i)	Difference $d_i = y_i - x_i$	Absolute value of difference $\lvert d_i \rvert$	Rank of absolute difference (sign)
1	80	72	+8	8	5(+)
2	76	70	+6	6	3(+)
3	78	82	−4	4	2(−)
4	90	76	+14	14	9(+)
5	84	86	−2	2	1(−)
6	86	76	+10	10	7(+)
7	81	74	+7	7	4(+)
8	84	75	+9	9	6(+)
9	88	76	+12	12	8(+)

differences (such as those calculated in Table 1) arise from a continuous, symmetric distribution. The null hypothesis in settings like Table 1 is that the distribution of these differences is centered at zero (although one can specify any constant, and apply signed ranking to those quantities). Under the null hypothesis, each assignment of signs (+ or −) to each of the N ranks is equally likely, and this fact is utilized to construct formal **hypothesis testing** procedures from these signed-rank statistics. The most common of these tests is the Wilcoxon signed-rank test [4], described in a separate article.

References

[1] Cureton, E.E. (1967). The normal approximation to the signed-rank sampling distribution when zero differences are present, *Journal of the American Statistical Association* **62**, 1068–1069.

[2] Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods*. Wiley, New York.

[3] Pratt, J.W. (1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures, *Journal of the American Statistical Association* **54**, 655–667.

[4] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

R.F. WOOLSON

Similarity, Dissimilarity, and Distance Measure

Statisticians are familiar with the concept of the **correlation**, or other measures of **association**, between two variables, but the concept of the similarity between two samples seems to have originated outside mainstream statistics, in ecology, taxonomy, psychology, and entomology (see, for example, Sneath & Sokal [6]) (*see* **Numerical Taxonomy**). Because of the taxonomic origins, **binary** variables are often termed characters or features; similar terms drawn from other fields of application are often encountered. Similarly, the term "sample" used here may be replaced by case, object, subject, OTU (Operational Taxonomic Unit), or unit. The most simple measures of similarity are concerned with binary variables denoting the presence or absence of characteristics, or perhaps two forms of a character, such as red and white. We shall see that the logical difference between these different types of binary variable affects the algebraic forms of acceptable coefficients. With two samples labeled i and j, say, we can count the number of agreements and disagreements among p binary variables to give Table 1, in which the two values of a binary variable are denoted by 1 and 0; thus 0/1 may refer to absence/presence or to red/black.

Thus, Table 1 shows that there are a agreements among form 1 of the p variables and d agreements among form 0. There are b (c) cases in which sample i has form 1 (0) and sample j has form 0 (1). Many similarity coefficients are simple functions of a, b, c, and d, where $p = a + b + c + d$; Hubálek [4] lists 43 similarity coefficients of this type. Two of the most common are the Simple Matching coefficient, defined by

$$S_{\mathrm{SM}} = \frac{a+d}{a+b+c+d} = \frac{a+d}{p}, \qquad (1)$$

and the Jaccard coefficient, defined by

$$S_{\mathrm{J}} = \frac{a}{a+b+c} = \frac{a}{p-d}. \qquad (2)$$

When $a = b = c = 0$, we define $S_{\mathrm{J}} = 0$. Note that S_{SM} and S_{J} typify the two major classes of similarity coefficients, those which include "negative matches" given by d and those which do not. By a negative match we mean that both samples lack a character and a shared missing character may be deemed no useful evidence of agreement. Thus, that two people both speak French as a mother tongue is an indication of a shared characteristic; but two people, neither of whom speaks French, is no such indication – there is a lack of symmetry between what is meant by a score of 1 (presence or +) and a score of 0 (absence or −). However, two samples that are both black or both white would, unless there is evidence to support differential weighting (see below), usually be deemed of equal similarity in respect of color; note however, that in genetics "white" may indicate the lack of a gene controlling the black state. Clearly, there can be major problems in deciding what is and what is not a negative match. These introductory remarks have touched on some of the issues associated with similarity coefficients, and we shall expand on these below (*see* **Agreement, Measurement of**).

Table 1 Numbers of agreements and disagreements between binary variables for two samples

		Sample j	
		1	0
Sample i	1	a	b
	0	c	d

The important problem of matching nucleic or amino acid sequences is related to that of constructing general purpose similarity coefficients, but because **DNA sequences** require alignment, there are special problems, with a large literature.

General Properties of Similarity Coefficients

We have discussed the similarity between samples i and j, and so, strictly speaking, the entries in Table 1 should be written a_{ij}, b_{ij}, c_{ij}, and d_{ij}, with a corresponding similarity coefficient; for simplicity, and when there is no ambiguity, it is customary to drop the suffices i and j. As for the Simple Matching and Jaccard coefficients, most similarity coefficients S_{ij} are positive, symmetric, and bounded by zero and unity. A value of zero indicates no matches ($a_{ij} = 0$, and where necessary also $d_{ij} = 0$). A value of unity implies a complete

match ($a_{ij} = p$, and where necessary, $a_{ij} + d_{ij} = p$). The property of symmetry implies that $a_{ij} = a_{ji}$, $d_{ij} = d_{ji}$ and $b_{ij} = c_{ji}$. A few coefficients of inter-sample correlational form satisfy $-1 \leq S_{ij} \leq 1$, and then the above comments need corresponding changes. With n samples, the $\binom{n}{2}$ pairs of similarity coefficients may be assembled into a nonnegative symmetric matrix $\mathbf{S}$ with unit diagonal elements. $\mathbf{S}$ is termed a similarity matrix and may be analyzed directly (*see* **Classification, Overview**; **Cluster Analysis of Subjects, Hierarchical Methods**) or may be converted into a dissimilarity matrix $\mathbf{D} = \mathbf{11}' - \mathbf{S}$, which has a zero diagonal and values $1 - S_{ij}$ off the diagonal. The obvious question to ask is to what extent do dissimilarities have the same properties as distances, usually Euclidean distances? Gower & Legendre [3] examine the following questions:

1. Do the dissimilarities satisfy the metric (triangle) inequality?
2. Are the dissimilarities Euclidean embeddable for an explanation, (*see* **Principal Coordinates Analysis**)?
3. As above, but for the square root of dissimilarity.

They also provide tools to help answer these questions, and answers for many specific similarity coefficients. The metric property is important because if it is not true, then the dissimilarity between samples i and j can be greater than the sum of the dissimilarities between i and k and j and k, which does not satisfy intuitive ideas of similarity. Euclidean embeddability is of interest in **multidimensional scaling** and, writing $\mathbf{N}$ for the matrix all of the values of which are $1/n$, we have (see (3) in the article on **Principal Coordinates Analysis**) that necessary and sufficient conditions for $(1 - S_{ij})^{1/2}$ to be embeddable is that the centred matrix $(\mathbf{I} - \mathbf{N})\mathbf{S}(\mathbf{I} - \mathbf{N})$ be positive semi-definite; it is sufficient for $\mathbf{S}$ itself to be positive semi-definite, which is usually more easily established than for the centred form. Thus, answers to question 3 relate directly to the properties of $\mathbf{S}$ and, it turns out, are easier to provide than those to question 2. We know that if $\mathbf{S}$ is positive semi-definite then so is the Hadamard (or element by element) product $\mathbf{S} * \mathbf{S}$, which would imply that $(1 - S_{ij}^2)^{1/2}$ is embeddable, but the answer to question 2 requires an investigation of the definiteness of the centred form of the similarity matrix with elements $1 - (1 - S_{ij})^2$. Gower & Legendre [3] answer some of these questions, and the results of special interest – shown in Table 2 – pertain to the following generalizations of the Simple Matching coefficient, defined by

$$S_{\mathrm{SM}}(\theta) = \frac{a + d}{a + d + \theta(b + c)}, \quad \theta > 0, \tag{3}$$

and of the Jaccard coefficient,

$$S_{\mathrm{J}}(\theta) = \frac{a}{a + \theta(b + c)} \quad \theta > 0. \tag{4}$$

Table 2 How the metric and Euclidean properties of the families $S_{\mathrm{J}}(\theta)$ and $S_{\mathrm{SM}}(\theta)$ change with positive values of θ

Family	Coefficient	Nonmetric	Metric	Euclidean
$S_{\mathrm{J}}(\theta)$	$1 - S_{ij}$	$0 < \theta < 1$	$\theta \geq 1$	*
	$(1 - S_{ij}^2)^{1/2}$	$0 < \theta < \frac{1}{3}$	$\frac{1}{3} \leq \theta < \frac{1}{2}$	$\frac{1}{2} \leq \theta$
$S_{\mathrm{SM}}(\theta)$	$1 - S_{ij}$	$0 < \theta < 1$	$\theta \geq 1$	*
	$(1 - S_{ij}^2)^{1/2}$	$0 < \theta < \frac{1}{3}$	$\frac{1}{3} \leq \theta < 1$	$1 \leq \theta$

The entries in Table 2 are worst-case scenarios. Thus, "Nonmetric" means that for the values of θ shown, nonmetric configurations can always be found, although some configurations may be metric or even Euclidean. "Metric" means that all configurations for the indicated values of θ are metric, although Euclidean representations will exist for some sets of data. "Euclidean" means that all configurations for the indicated values of θ are Euclidean. The asterisk indicates that, for these settings, data which generate non-Euclidean representations exist for all values of θ. Note that the case $\theta = 0$ represents a degenerate set of points coincident at the origin, and is not of interest.

The families $S_{\mathrm{J}}(\theta)$ and $S_{\mathrm{SM}}(\theta)$ include many popular similarity coefficients, and the results of Table 2 give some guidance on choices. For both families, we may note that if $S_{ij}(\theta) > S_{pq}(\theta)$, then $S_{ij}(\phi) > S_{pq}(\phi)$ for arbitrary θ and ϕ; that is, the coefficients in each family are monotonically related. It follows that if they are analyzed by any monotonically invariant method, such as nonmetric multidimensional scaling or single-linkage cluster analysis (*see* **Classification, Overview**), then the results do not depend on θ.

Multilevel qualitative variables may be dealt with similarly to binary variables. The simplest method is

to give a score $0 \le s_{ijk} \le 1$ for the match between the ith and jth samples for the kth variable. Then we may define similarity by

$$S_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk}. \tag{5}$$

When $s_{ijk} = 1$ for a match, otherwise $s_{ijk} = 0$, (5) is termed the Extended Matching Coefficient S_{ESM}, and becomes S_{SM} when all variables are binary. If there are a positive matches and a total of L levels over all p variables, then $S_{\mathrm{ESM}} = a/L$. An alternative approach for handling multilevel qualitative variables is to code each qualitative state as a separate binary variable and then use one of the binary coefficients, now based on L binary variables. Assuming that every character must occur in one of its states, then there are $p - a$ mismatches, each occurring twice, so that $b + c = 2(p - a)$ and $d = L + a - 2p$. Then $S_{\mathrm{SM}} = (L + 2a - 2p)/[(L + 2a - 2p + 2\theta(p - a)]$ and $S_{\mathrm{J}} = a/[a + 2\theta(p - a)]$, both of which are monotonically related to S_{ESM}. Thus, for many purposes, S_{ESM} is equivalent to all members of both families.

Similarity coefficients may also be defined for quantitative variables. Denoting the ith value of the kth variable by x_{ik}, most similarity coefficients of this type have the general form

$$S_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_k(x_{ik}, x_{jk}), \tag{6}$$

in which each variable contributes independently to overall similarity. Here $s_k(x_{ik}, x_{jk})$ denotes a function that may differ for each variable but, in practice, is nearly always the same function in any one study. Typical choices are based on one of the Minkowski metrics to give for positive values of t:

$$s_k(x_{ik}, x_{jk}) = 1 - \left\{\frac{|x_{ik} - x_{jk}|t}{r_k^t}\right\}^{1/t}$$

where r_k is a normalizer that eliminates the effects of scales of measurement and is chosen to ensure that $0 \le s_k(x_{ik}, x_{jk}) \le 1$. The usual choice is $t = 1, 2$ giving, respectively, similarities based on absolute differences and Euclidean distance. Note that for ratio scales, a logarithmic transform eliminates scale effects in differences. A simple choice is to set r_k equal to the range of the kth variable, possibly after transformation, in the sample. Other choices of similarity coefficients for quantitative variables and of normalizers and their metric and Euclidean properties are discussed by Gower [2] and Gower & Legendre [3].

The different choices of θ in (3) and (4) may be regarded as simple examples of character-weighting. In the field of classification, there has been much controversy over the desirability or otherwise of weighting, but if one decides to do so, it is simply done. A general coefficient that includes much of the above as special cases is to define (see Gower [1])

$$S_{ij} = \frac{\sum_{k=1}^{p} w_k(x_{ik}, x_{jk}) s_k(x_{ik}, x_{jk})}{\sum_{k=1}^{p} w_k(x_{ik}, x_{jk})}, \tag{7}$$

where $w_k(x_{ik}, x_{jk})$ is a weighting function. We may obtain $S_{\mathrm{SM}}(\theta)$ by setting $w_k(x_{ik}, x_{jk}) = 1$ when x_{ik} and x_{jk} match, and $w_k(x_{ik}, x_{jk}) = \theta$ when x_{ik} and x_{jk} do not match. Similarly, we obtain $S_j(\theta)$ by setting $w_k(x_{ik}, x_{jk}) = 1$ when x_{ik} and x_{jk} match positively, $w_k(x_{ik}, x_{jk}) = 0$ when x_{ik} and x_{jk} match negatively, and $w_k(x_{ik}, x_{jk}) = \theta$ when x_{ik} and x_{jk} do not match. We may also set $w_k(x_{ik}, x_{jk}) = 0$ when at least one of x_{ik} and x_{jk} is missing. Among other coefficients included are recursive similarity coefficients, where a hierarchy of primary, secondary, tertiary and so on, characters is recognized. Then the similarity among the secondary characters may be used to weight the primary characters, and similarly for characters at higher levels.

Metrics on Graphs

The Overview of Classification mentions minimal link trees, leading to ultrametrics, and additive distances that may be associated with hierarchical representations. More generally, distances may be defined on any connected network: the shortest route between two nodes, the minimal or maximal link on the route between two nodes, or measures that have physical interpretations. For example, one may associate unit resistance with every link and use Kirchoff's law to determine have found applications in molecular chemistry (see [5]).

Similarity Between Populations

In the above, we have referred to the similarity between pairs of samples. In taxonomy, samples typically refer to whole biological populations and characters are chosen that have little or no variation within the populations. This is often acceptable for qualitative characters – for example, all cats have claws and all dandelions are yellow – but quantitative variables will nearly always have a distribution within populations. Taxonomists try to find quantitative characters with little overlap between biological populations, which may be represented by their average or some other typical value. Then, each population may continue to be represented by a single sample, and similarity computed as discussed above. In applications that study closely related populations, the overlap cannot be ignored and inter-population distances, or other measures of inter-population overlap, rather than similarities must be used. These are discussed elsewhere (*see* **Discriminant Analysis, Linear**; **Mahalanobis Distance**). When, perhaps as a **null hypothesis**, the samples can be viewed as random drawn from a single population, the usual statistical questions arise concerning the distribution and joint distribution of the S_{ij}. Because of the nature of the variables, theoretical results are few (see, for example, Snijders et al. [7]) and therefore applications tend to use **jackknife** and other data-resampling techniques.

References

[1] Gower, J.C. (1971). A general coefficient of similarity and some of its properties, *Biometrics* **27**, 857–871.

[2] Gower, J.C. (1985). Measures of similarity, dissimilarity and distance, in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, New York, pp. 397–405.

[3] Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification* **3**, 5–48.

[4] Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence–absence data): an evaluation, *Biological Reviews* **57**, 669–689.

[5] Klein, D.J. (1997). Graph geometry, graph metrics and Wiener, *Comm. Math. Chem.* **35**, 7–27.

[6] Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

[7] Snijders, T.A.B., Dormaar, M., van Schuur, W.H., Dijkman-Caes, C. & Driessen, G. (1990). Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes, *Journal of Classification* **7**, 5–31.

(*See also* **Cluster Analysis of Subjects, Nonhierarchical Methods**; **Cluster Analysis, Variables**; **Pattern Recognition**; **Projection Pursuit**; **R- and Q-analysis**)

JOHN C. GOWER

Simple Structure

In **factor analysis**, a simple structure is the "ideal" structure, in which each factor is defined by a subset of the original variables with as little overlap as possible. In other words, ideally, we want to find a final **factor loading matrix** such that each column has a considerable number of near-zero loadings and a small number of large loadings, and each row has only one or a very small number of large entries. While orthogonal transformations (*see* **Orthogonal Rotation**) are often used, **oblique rotations** are generally better for finding a matrix that exhibits a simple structure. In general, the reference-vector structure matrix, **V**, is used to determine simple structure (*see* **Factor Loading Matrix**; **Primary Factors**). This matrix usually has at least as many near-zero loadings on each factor as the number of factors. Also, this reference structure matrix, which contains the correlations between the variables and the reference-vector factors, is often used for interpretation of the factors. The five criteria of simple structure proposed by Thurstone [4] are often too rigid to be applied to the real data. We prefer the following two criteria, discussed in Cureton & D'Agostino [2].

The first condition of simple structure is given as a criterion of overdetermination. The condition states that:

> There should be at least m near-zero loadings and usually several more than m in each column of the reference-vector structure matrix, where m is the number of the retained factors (i.e. the column of the matrix).

This first condition is used to overdetermine the location of a **primary factor** by at least m variables, and usually more than m. This is to make sure the locations of the primary factors are well determined. This first condition allows a few variables with nonzero loadings on all factors as long as there are enough variables that have near-zero loadings to determine the locations of the primary factors.

The second condition is based on the idea that the factors should be maximally distinct from one another, and that the factorial structure should be unique. The condition is given as follows:

> Among the subset of m or more rows of the reference-vector structure matrix having near-zero loadings in any one column, there is at least one and usually more than one nonzero loading in every other column. Every one of these rows must have at least one nonzero loading, and these nonzero loadings are distributed over all the other columns. Also in every column of the reference structure matrix, the number of negative nonzero loadings should be a minimum.

Alternative to the criteria of simple structure, a procedure based on a criterion of simplicity has been proposed by Bentler [1]. The results given by his procedure are not very different from those based on simple structure.

Since the conditions of the simple structure are given in qualitative terms, it is evident that subjective judgments are required to determine terms such as "near-zero" or "large" loadings, and "subset" of variables. Consequently, various analytic procedures have been developed for computing a simple structure. A review of this development can be found in Harman [3].

References

[1] Bentler, P.M. (1977). Factor simplicity index and transformations, *Psychometrika* **42**, 277–295.

[2] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: an Applied Approach*. Lawrence Erlbaum, Hillsdale.

[3] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd Ed. University of Chicago Press, Chicago.

[4] Thurstone, L.L. (1947). *Multiple Factor Analysis*. University of Chicago Press, Chicago.

RALPH B. D'AGOSTINO, SR & HEIDY K. RUSSELL

Simplex Models

Simplex models are employed for the analysis of relationships among variables that can be arranged according to a logical ordering. In some situations the ordering is known beforehand. For example, the ordering can reflect the sequence of trials in a learning experiment, the age of subjects in a longitudinal study, or the length of preparatory intervals in a stimulus response task. In other situations, the ordering refers to some unobservable property of the variable and has to be inferred from the data. For example, in the original formulation of the simplex model due to Guttman [5], the ordering was according to the complexity of ability tests. A simplex analysis is applicable only when all intercorrelations are positive. The primary aim of the analysis is to investigate the degree of similarity or closeness between successive variables.

Relationships of the simplex model with Markov and Wiener stochastic processes were discussed by Anderson [1]. Jöreskog [6] presented a comprehensive review of different types of simplex models and showed how these models can be fitted by **maximum likelihood**. Most methods for fitting the simplex require a prior knowledge of the ordering of variables, although two-stage procedures in which the ordering of variables is estimated in the first stage have been suggested by Kaiser [8] and by Cureton & D'Agostino [4, Chapter 15]. A reformulation of the simplex model that does not require a prior specification of the ordering of variables was suggested by Schönemann [10]. He also provided a procedure for fitting the order free simplex model using an approach based on **multidimensional scaling**. The method of maximum likelihood can also be used for fitting the order-free formulation of the simplex model [2, p. 132].

The Perfect Simplex

A classic example of a **correlation** matrix exhibiting a simplex pattern, originally presented by Guttman [5, Table 5] and frequently used since, is given in Table 1 (*see* **Guttman Scale**).

This shows correlations between six verbal ability tests applied to 1046 Bucknell sophomores. A correlation matrix is considered, since different tests are on different scales so that variances and covariances would not be meaningful. All tests are from the same domain and all intercorrelations are positive. There is a noticeable inequality pattern in the correlation coefficients. Those next to the diagonal are largest and they taper off as the lower left-hand corner is approached. This pattern is characteristic of the simplex. It is dependent on the ordering of the tests but not on direction. A similar pattern will occur if tests are listed in reverse order, with Test 6 listed first and Test 1 listed last. In Guttman's original work, the fundamental ordering represented complexity, but other fundamental orderings, such as time, may be considered.

Table 1 Six verbal ability tests: Bucknell College sophomores, $N = 1046$

Spelling	1	1.000					
Punctuation	2	0.621	1.000				
Grammar	3	0.564	0.742	1.000			
Vocabulary	4	0.476	0.503	0.577	1.000		
Literature	5	0.394	0.461	0.472	0.688	1.000	
Foreign literature	6	0.389	0.411	0.429	0.548	0.639	1.000

We distinguish between two types of simplex; a perfect simplex where no allowance is made for error of measurement in observing variables, and a quasi-simplex where measurement error is taken into account.

Guttman's Conceptualization of the Perfect Simplex

Let **P** represent the correlation matrix of the variates, $X_1, X_2, \ldots, X_p$. The basic assumption in Guttman's conceptualization is that if **P** satisfies a perfect simplex, all partial correlations between pairs of nonadjacent variables given any variable that is intermediate in the fundamental ordering will be zero:

$$\rho_{ik \cdot j} = 0, \quad 1 \leq i < j < k \leq p. \tag{1}$$

A consequence of the assumption in (1) is that a typical element of **P** may be expressed as

$$\rho_{ij} = \frac{\alpha_i}{\alpha_j}, \quad i < j, \tag{2}$$

where α_i is a parameter that may be interpreted as a measure of the degree of complexity of X_i. Guttman named the α_i "complexity loadings" and regarded them as correlation coefficients of the X_i with a

hypothetical endpoint of the fundamental continuum. They are defined only up to a constant of proportionality, so that an identification condition should be imposed. A suitable identification condition is

$$\alpha_p = 1. \tag{3}$$

If (3) is imposed, α_i may be interpreted as the correlation coefficient of X_i with the most complex available variable, X_p.

The basic assumption (1) also implies (cf. [1, p. 209]) that the regression equation of any variable on the remaining $p-1$ variables will have nonzero partial regression weights only on the adjacent variables in the fundamental ordering (*see* **Multiple Linear Regression**). Consequently, if $p = 5$ for example, the array of row vectors of regression weights of each variable on the remaining variables will have the pattern

$$\begin{bmatrix} - & \beta_{1.2} & 0 & 0 & 0 \\ \beta_{2.1} & - & \beta_{2.3} & 0 & 0 \\ 0 & \beta_{3.2} & - & \beta_{3.4} & 0 \\ 0 & 0 & \beta_{4.3} & - & \beta_{4.5} \\ 0 & 0 & 0 & \beta_{5.4} & - \end{bmatrix},$$

where $\beta_{i \cdot j}$ denotes the partial regression weight of the ith variable on the jth variable.

Since the density of a sample correlation matrix cannot be expressed in closed form, it is common practice to treat a correlation structure as a covariance structure,

$$\mathbf{\Sigma}_X = \mathbf{D}_\sigma \mathbf{P} \mathbf{D}_\sigma, \tag{4}$$

where the diagonal elements, $\sigma_1, \ldots, \sigma_p$, of $\mathbf{D}_\sigma$ represent standard deviations regarded as **nuisance parameters**, and $\mathbf{P}$ is a function of the parameters of interest. Estimates are then obtained by maximizing the Wishart likelihood function (*see* **Wishart Distribution**).

The covariance structure for a perfect simplex model involves $2p-1$ parameters. Maximum Wishart likelihood estimates may be expressed in closed form [9, Section 8.11; 6, Section 2.3]. Let $\mathbf{S}$ represent a sample covariance matrix, with typical element s_{ij}, based on a sample of size N, and let $\mathbf{R}$ be the corresponding correlation matrix, with typical element r_{ij}. **Maximum likelihood** estimates, subject to the identification conditions (3), are given by

$$\hat{\sigma}_i = (s_{ii})^{1/2}, \quad i = 1, \ldots, p$$
$$\hat{\alpha}_i = r_{i,i+1}\hat{\alpha}_{i+1}, \quad i = 1, \ldots, p, \ \hat{\alpha}_{p-1} = 1, \tag{5}$$

and the corresponding -2 log **likelihood ratio test** statistic is [6, Section 2.4]

$$G = (N-1)\left[\sum_{i=1}^{p-1} \ln(1 - r_{i,i+1}^2) - \ln|\mathbf{R}|\right].$$

Under the null hypothesis that the perfect simplex model holds, the asymptotic distribution of G is **chi-square** with $\frac{1}{2}p(p-3)+1$ **degrees of freedom**.

The maximum likelihood estimates (5) are particularly easy to calculate and only make use of the $p-1$ correlation coefficients adjacent to the main diagonal. These estimates are, however, dependent on the ordering of variables. This is generally not known in advance.

Schönemann [10] suggested that the (2) for the elements of a perfect simplex be expressed in the alternative order-free form

$$\rho_{ij} = \frac{\min(\alpha_i, \alpha_j)}{\max(\alpha_i, \alpha_j)}. \tag{6}$$

The direction of complexity in (6) is indeterminate, as $\alpha_1^{-1}, \ldots, \alpha_p^{-1}$ may be regarded as points on an equivalent fundamental continuum, since

$$\frac{\min(\alpha_i^{-1}, \alpha_j^{-1})}{\max(\alpha_i^{-1}, \alpha_j^{-1})} = \frac{\min(\alpha_i, \alpha_j)}{\max(\alpha_i, \alpha_j)}.$$

Schönemann [10] pointed out that the elements of the symmetric matrix $\mathbf{P}^*$, with typical element

$$\rho_{ij}^* = -\ln \rho_{ij} = |\ln \alpha_i - \ln \alpha_j|,$$

s represent Euclidean distances between points $\alpha_i^* = \ln \alpha_i, i = 1, \ldots, p$, on the real line and showed how classical **multidimensional scaling** can be applied to the corresponding matrix of distance estimates to obtain estimates, $\hat{\alpha}_i^*$, of these points. Estimates of the α_i are then obtained by taking antilogarithms: $\hat{\alpha}_i = \exp(\hat{\alpha}_i^*)$. The classical scaling procedure involves the extraction of the largest **eigenvalue** and corresponding **eigenvector** of a symmetric matrix, and the resulting estimates are invariant under reordering of the variables.

Stochastic Process Interpretation

A **stochastic process** is a family of **random variables** $X(t)$ indexed by a continuous parameter $t, 0 \leq t < \infty$, often regarded as a time parameter. Variables

are considered to be observations of the process at the points $t_1, \ldots, t_p$. Correlation coefficients between variables, $X(t_i)$ and $X(t_j)$, are defined by a function of t_i and t_j known as a correlation function.

Anderson [1] considered the simplex from the viewpoint of the Markov and Wiener stochastic processes and pointed out that the partial correlation property in (1) underlying the simplex is a known property of a **Markov process**. The correlation function of a Markov process is

$$\rho_{ij} = \rho(X(t_i), X(t_j)) = \rho^{|t_i - t_j|}, \quad \text{for all } i, j. \quad (7)$$

where ρ is a correlation parameter. This correlation function may be shown to be equivalent to expression (6) for the elements of a perfect simplex by means of the substitution

$$\alpha_i = \rho^{-t_i}. \quad (8)$$

Thus (7) may be regarded as a reparameterization of (6) involving $p + 1$ parameters, $\rho, t_1, \ldots, t_p$ instead of the p parameters $\alpha_1, \ldots, \alpha_p$.

In some situations the time points are known in advance and ρ is estimable. Here, the t_i are not known in advance and the correlation parameter is indeterminate. Two identification conditions must be imposed. Suitable identification conditions are

$$t_1 = 1, \qquad t_p = p. \quad (9)$$

If an order-free approach is employed, t_1 and t_p denote the smallest and largest points respectively, and need not be the first and last points.

These identification conditions fix the endpoints of the time scale and serve to identify ρ and the remaining "time" points, which need not assume integral values. The t_i may be regarded as alternate parameters to be interpreted instead of the complexity loadings, α_i, and values satisfying the identification conditions in (9) may be calculated from

$$t_i = 1 + (p-1)\frac{\ln \alpha_i - \ln \alpha_1}{\ln \alpha_p - \ln \alpha_1}$$

where, again, α_1 and α_p refer to the smallest and largest complexity loadings if they are not given in increasing order. The correlation parameter then is given by

$$\rho = \exp\left[-\left(\frac{\ln \alpha_p - \ln \alpha_1}{p-1}\right)\right].$$

An *equally spaced* simplex [5] is one where the time points, t_i, are equally spaced on the real line.

A data model (cf. [6, Section 5.6; 7, Section 4.1]) that generates a Markov process is the first order autoregressive (AR1) time series (*see* **ARMA and ARIMA Models**) with nonhomogeneous autoregression weights, β_i, and nonhomogeneous white noise variances, ψ_{ii}:

$$(X_{i+1} - \mu_{i+1}) = \beta_i(X_i - \mu_i) + Z_{i+1}, \quad i = 1, \ldots, p-1,$$

where $\mathcal{E}(X_i) = \mu_i$ and the Z_i are mutually independently distributed white noise terms with variances $\psi_{ii}, i \geq 2$. We define $\psi_{11} = \text{var}(X_1)$. This data model generates the covariance structure

$$\mathbf{\Sigma}_X = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{D}_\psi(\mathbf{I} - \mathbf{B}')^{-1}, \quad (10)$$

where $\mathbf{D}_\psi$ is a diagonal matrix with typical diagonal element ψ_{ii} and $\mathbf{B}$ is a matrix with zero elements except for those just below the main diagonal. When $p = 5$ for example,

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \beta_1 & 0 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 0 \\ 0 & 0 & \beta_3 & 0 & 0 \\ 0 & 0 & 0 & \beta_4 & 0 \end{bmatrix}.$$

The model of (10) is equivalent (cf. [6, section 5.6]) to the perfect simplex given by (4) and (2). There are $2p - 1$ parameters: $\beta_1, \ldots, \beta_{p-1}, \psi_{11}, \ldots, \psi_{pp}$. In order to see the relationship between the autoregression weights and the complexity loadings, we define the standardized autoregression weight

$$\beta_i^* = \left(\frac{\sigma_{ii}}{\sigma_{i+1,i+1}}\right)^{1/2}\beta_i,$$

where σ_{ii} denotes the ith diagonal element of $\mathbf{\Sigma}_X$ in (10). Thus $\beta_i^* = \beta_i$ if the X_i have been standardized to have unit variances so that $\mathbf{\Sigma}_X$ in (10) has unit diagonals and is therefore a correlation matrix. The relationship between the complexity loadings and the standardized autoregression weights then is

$$\alpha_i = \prod_{j=i}^{p-1}\beta_j^*, \quad i = 1, \ldots, p-1$$

Prior knowledge of the ordering of variables is necessary for use of the formulation of the simplex model in (10).

A Wiener stochastic process has structured variances (*see* **Brownian Motion and Diffusion Processes**). Its indexing parameter will be referred to as $s, 0 \leq s < \infty$, with scale points $s_1, \ldots, s_p$. The Wiener process has the covariance function

$$\sigma_{ji} = \sigma_{ij} = \text{cov}[X(s_i), X(s_j)] = s_i, \quad s_i \leq s_j, \tag{11}$$

so that the variances $\sigma_{ii} = s_i$ are constrained. It can be shown that the correlations obtained from the variances and covariances of (11) have the simplex structure (2) (cf. [1, pp. 207–208; 6, Section 3.1; 9, Section 8.11]) but the relationship between the complexity loading, α_i, and scale point, s_i, is

$$\alpha_i = (s_i)^{1/2}$$

and differs from the corresponding relationship (8) between the complexity loading and the scale point, t_i, of a Markov process. A matrix expression for the covariance matrix with typical element given by (11) is [6, Section 3.1]

$$\boldsymbol{\Sigma}_X = \mathbf{T}\mathbf{D}_\varphi\mathbf{T}', \tag{12}$$

where $\mathbf{D}_\varphi$ is a diagonal matrix with diagonal elements $\varphi_{11} = s_1$ and $\varphi_{ii} = s_i - s_{i-1}, i = 2, \ldots, p$, and $\mathbf{T}$ is a lower triangular matrix with elements on and below the diagonal equal to 1. If $p = 5$, for example,

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Jöreskog [6] referred to the covariance structure of (12) as a Wiener simplex and to the covariance structure defined equivalently by (4) with (6) or (7) or by (10) as a Markov simplex. The Wiener simplex is more restrictive than the Markov simplex as it has p parameters, instead of $2p - 1$, and imposes a structure on variances. Unlike the Markov simplex, the Wiener simplex is scale dependent.

Maximum likelihood estimates in closed form of the φ_{ii} of the Wiener simplex are provided in Jöreskog [6, Section 3.2]. Since the Wiener simplex is both scale-dependent and order-dependent, it is applicable mainly in situations in which repeated measurements are taken over time on a number of subjects. An example is provided in Jöreskog [6, Section 4.8].

A data model that generates the Wiener simplex has been suggested by Guttman [5, p. 310]. If

$$X_i = \mu_i + \sum_{j=1}^{i} V_j,$$

where the V_j are mutually independently distributed with $\mathcal{E}(V_j) = 0$ and $\text{var}(V_j) = \varphi_{jj}$, then $\boldsymbol{\Sigma}_X$ is given by (12). Each V_j is interpreted as representing the increase in complexity between X_{j-1} and X_j.

An alternate parameterization of the Markov simplex may be obtained by applying a scaling transformation to the Wiener simplex to allow variances to be arbitrary [6]:

$$\boldsymbol{\Sigma}_X = \mathbf{D}_\gamma\mathbf{T}\mathbf{D}_\varphi\mathbf{T}'\mathbf{D}_\gamma, \tag{13}$$

where $\mathbf{D}_\gamma$ is a diagonal matrix with scaling factors, γ_i, as diagonal elements. A single identification condition is required; for example,

$$\varphi_{11} = 1.$$

The relationship between the complexity loadings, α_i, and incremental variances, φ_{jj}, is then given by

$$\alpha_i = \left(\frac{\sum_{j=1}^{i} \varphi_{jj}}{\sum_{j=1}^{p} \varphi_{jj}} \right)^{1/2}.$$

We thus have four alternate but equivalent parameterizations of the covariance structure of a perfect Markov simplex involving parameters with different interpretations. In (4) with (6), the complexity parameters, α_i, represent correlation coefficients with the last variable on the complexity scale; in (4) with (7), the parameters, t_i, are analogous to time points; in (10), the parameters, β_i, are regression weights on the preceding variable; and in (13) the parameters, φ_{ii}, represent incremental variances. Maximum likelihood estimates of the parameters of each of these parameterizations of the simplex applied to the data of Table 1 are shown in Table 2. Since the maximum likelihood estimates

Table 2 Perfect simplex maximum likelihood estimates: six verbal ability test data

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$\hat{\rho}$
$\hat{\alpha}_i$	0.12	0.19	0.25	0.44	0.64	1.00[a]	
$\hat{t}_i$	1.00[a]	2.11	2.80	4.09	4.96	6.00[a]	0.65
$\hat{\beta}_i^*$	0.62	0.74	0.58	0.69	0.64		
$\hat{\varphi}_{ii}$	1.00[a]	1.59	2.12	9.44	15.74	43.30	

[a]Parameter value fixed for identification purposes

reported are invariant under scale changes, analysis of a correlation matrix instead of a covariance matrix is in order.

Since the four parameterizations of the perfect simplex model are equivalent they yield the same **goodness of fit**, provided that the same ordering is employed in each. This is the case in the present example, since the order-free parameterizations of the model in (6) or (7) with (4) yield the same ordering of variables as Guttman's original ordering in Table 1, used for the parameterizations of (10) and (13). Fit of the perfect simplex to the data of Table 1 is not satisfactory. The largest absolute residual between a sample correlation coefficient and the corresponding correlation coefficient reproduced from the model is 0.27 and the value of the likelihood ratio goodness of fit test statistic is 202.6 with associated degrees of freedom equal to 10.

The Quasi-Simplex

The perfect simplex model is quite restrictive and seldom fits well in practice. Guttman [5] suggested several possible modifications that allow for error and referred to them as quasi-simplex models. One, Guttman's δ-simplex, has become generally accepted as the quasi-simplex.

Suppose now that the variates $X_1, \ldots, X_p$, that have a correlation matrix $\mathbf{P}$ satisfying the condition (1) for a perfect simplex, are unobservable and that, instead, variates $Y_1, \ldots, Y_p$ may be observed with

$$Y_i = X_i + E_i,$$

where the E_i are errors, distributed mutually independently and independently of the X_i with diagonal covariance matrix $\mathbf{D}_\theta$. It follows that the covariance matrix, $\mathbf{\Sigma}_Y$, of the observable variables, $Y_1, \ldots, Y_p$, is related to the covariance matrix, $\mathbf{\Sigma}_X$, of the simplex variables, $X_1, \ldots, X_p$, by

$$\mathbf{\Sigma}_Y = \mathbf{\Sigma}_X + \mathbf{D}_\theta, \tag{14}$$

where $\mathbf{\Sigma}_X$ is defined by any one of the four parameterizations for a perfect Markov simplex considered earlier, or by the perfect Wiener simplex (12).

The introduction of $\mathbf{D}_\theta$ is accompanied by additional indeterminacy in the model. In the case of the quasi-Wiener simplex, defined by (12) with (14), there is a single indeterminacy involving φ_{11} and θ_{11} [6, Section 4.2]. A suitable identification condition is $\varphi_{11} = \theta_{11}$. Two additional identification conditions are required when $\mathbf{D}_\theta$ is introduced in the quasi-Markov simplex: one involves the error variance, θ_{11} for the least complex variable and the other, the error variance, θ_{pp}, for the most complex variable [1; 6, Section 5.1]. Suitable identification conditions (IC) are as follows:

$$\text{IC1:} \quad \theta_{11} = \theta_{22}, \quad \theta_{pp} = \theta_{p-1,p-1} \tag{15a}$$

or

$$\text{IC2:} \quad \theta_{11} = 0, \quad \theta_{pp} = 0. \tag{15b}$$

The choice of identification conditions does not affect the values of $\theta_{22}, \ldots, \theta_{p-1,p-1}$, but does affect other parameters in the model that define the structure of the simplex correlation matrix $\mathbf{P}$ in (4). If variables are ordered according to complexity, the first row (column) and last row (column) of $\mathbf{P}$ are affected. This implies that hypotheses concerning relationships of the least complex simplex variable, X_1, and most complex simplex variable, X_p, to the other simplex variables are not testable.

Maximum likelihood estimates of parameters in the quasi-simplex model cannot be expressed in closed form and an iterative computational procedure is required. When the order-dependent parameterizations, (10), (13) and (12) of $\mathbf{\Sigma}_X$ are incorporated in (14), standard **structural equation modeling** computer programs may be employed (see e.g., [6, 7]). If the order-free formulations of (6) or (7) are employed, a computer program that allows a flexible specification of the model (see e.g., [3]) is necessary.

In Table 3 are shown maximum likelihood estimates of the error variances and complexity loadings when a quasi-Markov simplex, using (6), is fitted to the data of Table 1 under both identification conditions IC1 and IC2 in (15). It is apparent that the

Table 3 Quasi-simplex. Maximum likelihood estimates: six verbal ability test data

		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$
IC 1	$\hat{\theta}_{ii}$	0.212[a]	0.212[a]	0.212	0.215	0.218[b]	0.218[b]
	$\hat{\alpha}_i$	0.381	0.483	0.517	0.713	0.817	1[c]
IC 2	$\hat{\theta}_{ii}$	0[c]	0.212	0.212	0.215	0.218	0[c]
	$\hat{\alpha}_i$	0.299	0.428	0.457	0.631	0.722	1[c]

[a,b]Parameter values equal for identification purposes
[c]Parameter value fixed for identification purposes

Table 4 Quasi-simplex. Simplex variate correlations: six verbal ability tests

	1	2	3	4	5	6
1	1	*0.79*	*0.74*	*0.53*	*0.47*	*0.38*
2	*0.70*	1	0.93	0.68	0.59	*0.48*
3	*0.65*	0.93	1	0.73	0.63	*0.52*
4	*0.47*	0.68	0.73	1	0.87	*0.71*
5	*0.41*	0.59	0.63	0.87	1	*0.82*
6	*0.30*	*0.43*	*0.46*	*0.63*	*0.72*	1

IC 1 above diagonal; IC 2 below diagonal

choice of identification conditions affects the estimates of complexity loadings. The transition between the two sets of complexity loadings is accomplished by multiplying $\hat{\alpha}_2, \ldots, \hat{\alpha}_5$ by the same constant and $\hat{\alpha}_1$ by a different constant. Consequently, simplex variable correlations that involve the first simplex variable or last simplex variable are dependent on the identification conditions employed. This is apparent in Table 4, in which simplex variate intercorrelation estimates under IC1 are shown above the main diagonal and those under IC2 below. Correlations affected by the identification conditions are shown in italics. While the choice of identification conditions is arbitrary in the sense that the fit of the model is not affected, it seems that IC1 are more plausible because of the closeness of $\hat{\theta}_{22}, \ldots, \hat{\theta}_{55}$ in Table 3.

The fit of the quasi-simplex model is reasonably satisfactory. The largest absolute residual is now 0.09 and the likelihood ratio test statistic is 43.81 with six degrees of freedom.

References

[1] Anderson, T.W. (1960). Some stochastic process models for intelligence test scores, in *Mathematical Methods in the Social Sciences*, K.J. Arrow, S. Karlin & P. Suppes. eds. Stanford University Press, Stanford, pp. 205–220.

[2] Browne, M.W. (1982). Covariance structures, in *Topics in Applied Multivariate Analysis*, D.M. Hawkins, ed. Cambridge University Press, Cambridge, pp. 72–141.

[3] Browne, M.W. & Du Toit, S.H.C. (1992). Automated fitting of nonstandard models, *Multivariate Behavioral Research* **27**, 269–300.

[4] Cureton, E.E. & D'Agostino, R. (1983). *Factor Analysis, an Applied Approach*. Lawrence Erlbaum, Hillsdale, Chapter 15.

[5] Guttman, L. (1954). A new approach to factor analysis: the radex, in *Mathematical Thinking in the Social Sciences*, P.F. Lazarsfeld, ed. Columbia University Press, New York, pp. 258–348.

[6] Jöreskog, K.G. (1970). Estimation and testing of simplex models, *British Journal of Mathematical and Statistical Psychology* **23**, 121–145.

[7] Jöreskog, K.G. & Sörbom, D. (1977). Statistical models and methods for analysis of longitudinal data, in *Latent Variables in Socioeconomic Models*, D. Aigner & A. Goldberger, eds. North-Holland, Amsterdam, pp. 285–325.

[8] Kaiser, H.F. (1962). Scaling a simplex, *Psychometrika* **27**, 155–162.

[9] Morrison, D.F. (1967). *Multivariate Statistical Methods*. McGraw-Hill, New York.

[10] Schönemann P.H. (1970). Fitting a simplex symmetrically, *Psychometrika* **35**, 1–21.

(*See also* **Cronbach's Alpha**; **Factor Analysis, Overview**; **Likert Scale**; **Psychometrics, Overview**)

MICHAEL W. BROWNE

Simpson's Paradox

Simpson's paradox can be illustrated with an example, as Simpson did in his 1951 paper [8]. In that example, 40 patients are treated for a certain disease, whereas the control group consisted of 12 people. The results are given in the following $2 \times 2 \times 2$ **contingency table**, which shows the relationship between treatment and response (whether alive or dead) separately for males and females. Table 1 also shows the marginal table obtained by collapsing over the sex variable.

The odds ratios (*see* **Odds Ratio**) for the first two 2×2 tables (male/female) are both 5/6, which implies that the treatment is effective for both males and females. However, the odds ratio of the combined table is 1.0, which means the treatment is not effective. This phenomenon is called Simpson's paradox, which states that the direction of association between variables X (untreated/treated) and Y (alive/dead) may reverse after pooling over a covariate Z (male/female). The paradox can occur because pooling can lead to inappropriate weighting of the different subgroups [2].

Yule [11] first discovered this phenomenon, and it is also called the Yule–Simpson paradox. Numerous real-life examples of Simpson's paradox have been reported in many areas, including epidemiology, physics, social science, psychology, and sports. For example, Cohen et al. [3] compared tuberculosis deaths in New York City and Richmond, Virginia, in 1910. If the population was divided into racial groups, then Richmond had a lower death rate in both white and nonwhite categories, but the overall death rate was lower in New York.

In the context of contingency tables, Simpson's paradox is restricted neither to one association measure – the odds ratio, nor to $2 \times 2 \times 2$ tables. Let $[a_i, b_i; c_i, d_i]$, $i = 1, \ldots, K$, denote cell counts in a $2 \times 2 \times K$ table, and let $[a = \sum a_i, b = \sum b_i; c = \sum c_i, d = \sum d_i]$ be the corresponding marginal table. Let $\alpha(a_i, b_i; c_i, d_i)$ represent a measure of the association, such as the odds ratio $\alpha(a_i, b_i; c_i, d_i) = (a_i d_i)/(b_i c_i)$. *Simpson's paradox* occurs if $(a_i d_i)/(b_i c_i) > 1 (< 1)$ for all i, and $(ad)/(bc) \leq 1 (\geq 1)$. The paradox is also called *association reversal* by Samuels [7]. Good et al. [6] extended Simpson's paradox to an *amalgamation paradox*, which is defined as follows:

$$\alpha(a, b; c, d) > \max \alpha(a_i, b_i; c_i, d_i) \quad \text{or}$$

$$\alpha(a, b; c, d) < \min \alpha(a_i, b_i; c_i, d_i),$$

where the measure of association α can be the odds ratio or some other measure, such as the relative risk.

It is frequently helpful to collapse high-dimensional contingency tables, since the collapsed table has larger cell frequencies, fewer parameters, and is easier to interpret (*see* **Collapsibility**). It is then of interest to know when a table can be safely collapsed, avoiding the paradox. Let $X \perp Y$ and $X \perp Y|Z$ denote the independence and conditional independence of X and Y, respectively. Wermuth [9] showed that for a $2 \times 2 \times 2$ table we can meaningfully pool over a covariate Z and expect to find the same odds ratio in the marginal table and the partial tables (*strict collapsibility*) if and only if $X \perp Z|Y$ or $Y \perp Z|X$ (see also Bishop et al. [1]). Strict collapsibility implies that Simpson's paradox does not occur. A similar result was obtained for the relative risk [9]. Whittemore [10] showed that an $I \times J \times 2$ table is strictly collapsible if and only if at least one of the two-factor interactions of Z with X or Y in a loglinear model is zero. A table is called *strongly collapsible* (Ducharme et al. [5]) if it remains strictly collapsible no matter how it is partially collapsed (the definition was generalized to n-dimensional contingency tables by Whittemore). In this case, the odds ratio is totally independent of the level of the covariate. Ducharme et al. [5] provided a necessary and sufficient condition for a table to be strongly collapsible.

The paradox can also be avoided by a carefully designed experiment. For a $2 \times 2 \times K$ table, if the ratio of the sums of the two rows of each of the 2×2 partial tables remains constant, the design is called *row-uniform*. A *column-uniform* design can be similarly defined. If a design is both row-uniform and column-uniform, then the odds ratio of the combined table falls between the maximum and minimum of that of the individual tables [6], so the amalgamation paradox is avoided. Samuels [7] gave a necessary and sufficient condition to avoid association reversal, and also obtained a similar result in a regression setting.

When Simpson's paradox does occur in practice, what will the appropriate conclusion be? For example, if a disease is unrelated to a **genotype** for both whites and nonwhites but they are related when the two racial groups are combined, one is interested

Table 1 Survival by treatment among males and among females

	Male		Female		Combined	
	Untreated	Treated	Untreated	Treated	Untreated	Treated
Alive	4	8	2	12	6	20
Dead	3	5	3	15	6	20

in knowing whether the disease is related to the genotype. Suppose X takes value G (genotype A) or $\overline{G}$ (other than genotype A), Y takes value D (disease) or $\overline{D}$ (no disease), and Z is a covariate indexing race. If $X \perp Y|Z$ without $X \perp Y$, Simpson's paradox occurs. Dawid [4] defined Z as a sufficient set of covariates if $Y \perp I|(X, Z)$, where I contains the labels of individual units of the population. In the above example, race is a sufficient covariate if given a person's race and genotype; whether the disease occurs does not depend on an individual person. If Z is sufficient and $X \perp Y|Z$, then the disease is unrelated to the genotype even though they look related when the tables are collapsed over Z.

References

[1] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

[2] Blyth, C.R. (1972). On Simpson's paradox and the sure-thing principle, *Journal of the American Statistical Association* **67**, 364–366.

[3] Cohen, M.R. & Nagel, E. (1934). *An Introduction to Logic and Scientific Method*. Harcourt Brace, New York.

[4] Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion), *Journal of the Royal Statistical Society, Series B* **41**, 1–31.

[5] Ducharme, G.R. & Lepage, Y. (1986). Testing collapsibility in contingency tables, *Journal of the Royal Statistical Society, Series B* **48**, 197–205.

[6] Good, I.J. & Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables, *Annals of Statistics* **15**, 694–711.

[7] Samuels, M.L. (1993). Simpson's paradox and related phenomena, *Journal of the American Statistical Association* **88**, 81–88.

[8] Simpson, E.H. (1951). The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society, Series B* **13**, 238–241.

[9] Wermuth, N. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable, *Journal of the Royal Statistical Society, Series B* **49**, 353–364.

[10] Whittemore, A.S. (1978). Collapsibility of multidimensional contingency tables, *Journal of the Royal Statistical Society, Series B* **40**, 328–340.

[11] Yule, G.U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* **2**, 121–134 (Reprinted in *Statistical Papers of George Udny Yule*, Griffin, London, pp. 71–84).

Further Reading

Dong, J. (1998). On avoiding association paradoxes in contingency tables, *Journal of Systems Science and Mathematical Sciences* **11**(3), 272–279.

(*See also* **Contingency Table**; **Loglinear Model**)

JIANPING DONG

Simulation

To check his derivation of **Student's *t* distribution**, W.S. **Gossett** ("Student") conducted a simulation experiment:

> The material used was a ... table containing the height and left middle finger measurements of 3000 criminals ... The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random ..., each consecutive set of 4 was taken as a sample ... and the mean [and] standard deviation of each sample determined ... This provides us with two sets of ... 750 [values] on which to test the theoretical results arrived at. The height and left middle finger ... table was chosen because the distribution of both was approximately normal ...

The pieces of cardboard of 1908 are redundant in the computer era. The result of Figure 1 is readily obtained by the MATLAB code of Figure 2 (*see* **Software, Biostatistical**). Computers have effectively made available easily obtained streams of **random variables** from any distribution. This has revolutionized the whole of statistics, and in particular **Bayesian methods**. Gossett's use of simulation to verify a theoretical result has now become standard practice. The fitting of nonlinear models to data, for example by **maximum likelihood**, usually requires numerical iteration procedures, carried out by computer (*see* **Optimization and Nonlinear Equations**). The correct operation of these procedures should be checked by first applying them to data simulated from the model, using known parameter values.

Classical numerical optimization employs deterministic search **algorithms**. Stochastic search methods, which allow random excursions over the surface to be optimized, can be achieved using simulated annealing techniques – see for example Brooks & Morgan [6] – which are less likely to be trapped in local optima. This is just one of the many ways in which simulation has greatly increased the tools available to statisticians. Note also for instance the use of genetic algorithms [34], the requirements of **randomization tests** and permutation tests [15, 28], the use of the **bootstrap** [11, 16, 21, 41], and **Markov chain Monte Carlo methods** (MCMC) [4, 37]. To demonstrate the wide-ranging influence of simulation on statistics, we outline later **Monte Carlo** inference and Monte Carlo testing. We start with a description of how computers produce the random number streams required.

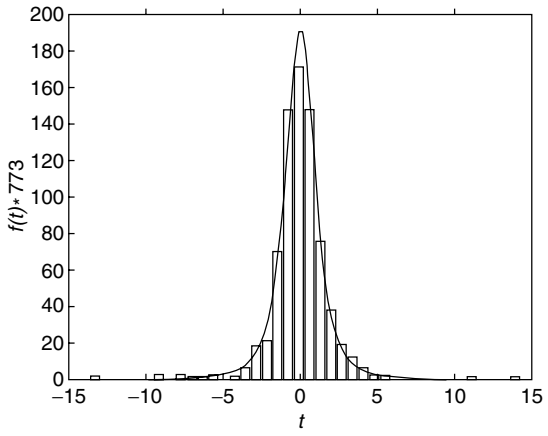

Figure 1 Histogram of 750 realizations of $2\overline{x}/s$, where $\overline{x}$ and s are respectively the mean and standard deviation of a sample of size 4 from a normal distribution of mean zero and variance unity. Also plotted is the underlying probability density function, of a t_3 random variable, given by

$$f(t) = \frac{2}{\pi\sqrt{3}(1 + t^2/3)^2}, \quad -\infty < t < \infty$$

```
nsim=750;n=4;nbins=40;
x=randn(n,nsim);
z=2*mean(x)./std(x);r=max(z)-min(z);s=r/nbins;
hist(z,nbins);hold on
t=linspace(-10,10,100);
f=2*nsim*s./(pi*sqrt(3)*(1+t.*t/3).^2);
plot(t,f)
title(`Figure 1')
xlabel(`t')
ylabel(`f(t)*773')
```

Figure 2 The MATLAB program which produces and plots Figure 1

Pseudo-Random Numbers

Computers generate streams of **pseudo-random numbers**, rather than strictly random ones, and they do this by means of recursion formulae. One which is frequently adopted is the congruential generator:

$$x_{n+1} = ax_n + b (\text{mod } m), \quad \text{for } n \geq 0. \qquad (1)$$

Here a, b and m are suitably chosen fixed integer constants, and the stream of numbers is initiated from a seed, x_0.

The integers resulting from (1) all lie in the range 0 to $(m-1)$, and approximations to random variables which are **uniformly distributed** over (0, 1) are obtained from setting $u_i = x_i/m$. Care is needed in how the fixed integers are chosen (see [33]). The MATLAB uniform random number generator, for example, uses $a = 7^5, b = 0$, and $m = 2^{31} - 1$. In particular, it is important to obtain a long sequence before x_0 reappears and the previous sequence cycles again and again. Cycling is not what we expect from random sequences. To put this in perspective, Wichmann & Hill [39] produced a generator that would take more than 800 years to cycle if 1000 of the numbers were used each second. The connection of the sequence of (1) with **chaos** is shown by Bartlett [2] and Lawrance [27]. However MCMC methods, for example, make intensive use of random variables, and it is advisable to exhibit a degree of caution when using pseudo-random numbers. A wide range of tests are available to check that pseudo-random numbers satisfy many necessary requirements of random variables. These range from graphical procedures (we can check Figure 1 by eye for obvious discrepancies between the histogram and the probability density function), through other empirical tests carried out on generated numbers, to theoretical tests such as the spectral test [17, 26] (*see* **Spectral Analysis**).

Alternative methods have been devised to improve on the performance of congruential generators. Generalized Feedback Shift Register (GFSR) methods provide examples that are widely used. These are described by Fishman [17], who also emphasizes that generators need to be easy to implement and portable, in that they produce the same results in different computing environments.

Generating Nonuniform Random Variables

The random variable $X = -\log_e U$, where U is uniformly distributed over the range 0–1, has an **exponential distribution**. If U_1 and U_2 are two independent such uniform random variables, then the pair of random variables given by

$$N_1 = (-2\log_e U_1)^{1/2}\cos(2\pi U_2),$$
$$N_2 = (-2\log_e U_1)^{1/2}\sin(2\pi U_2) \qquad (2),$$

are independent standard normal variables.

In simulations, such as that producing Figure 1, we need to be able to generate realizations of any random variables. The starting point is a stream of pseudo-random uniform random variables. Random variable simulation can be done using particular relationships, as in the Box–Müller [5] method of (2), or through one of a range of general procedures, such as the inversion method, the rejection method, and the composition method, all of which are described in [32]. For discrete random variables, the analog of the inversion method is the "table look-up" method. To take the inversion method as an illustration, if U is a U (0, 1) random variable and we wish to simulate a continuous random variable with cumulative distribution function $F(x)$, then it suffices to set

$$X = F^{-1}(U),$$

and this demonstrates immediately why $X = -\log_e U$ has an exponential distribution.

The use of trigonometric functions in (2) can be avoided by setting

$$N_1 = V_1\left(\frac{-2\log_e W}{W}\right)^{1/2},$$
$$N_2 = V_2\left(\frac{-2\log_e W}{W}\right)^{1/2}, \qquad (3)$$

subject to $W = V_1^2 + V_2^2 \leq 1$, where V_1 and V_2 are independent uniform random variables over the range -1 to 1. If $W > 1$ then the pair of values (V_1, V_2) is rejected and a new pair selected. Thus rejection occurs for a proportion $(1 - \pi/4)$ of the pairs, but results in the computational efficiency gain of using (3), rather than (2). We can see from (3) that the ratio V_1/V_2 has a **Cauchy distribution**. A general simulation method based on a ratio of uniformly distributed random variables has been proposed by Kinderman & Monahan [24]. It is well suited to adaptation to simulate from a probability density function which is specified only up to proportionality, and is therefore very useful in **Bayesian** computations (see [38]).

The general rejection method for any continuous random variable requires a uniform scatter of points over the area underneath the probability density function, $f(x)$, of the random variable. The abscissae of the points then provide realizations of the random variable. The required scatter is obtained by simulating from a density function $h(x)$, chosen both for

its similarity to $f(x)$, and relative ease of variate simulation. The density $f(x)$ is then enveloped by $g(x) = kh(x)$ for a suitable constant $k \geq 1$. The rejection probability is $(1 - k^{-1})$. Typically, selecting k involves solving an optimization problem. Examples are given by Morgan [32].

Random variables may be simulated in many different ways. It is important to use efficient methods, especially in cases of intensive use, as in MCMC work and **bootstrap** sampling. Detailed descriptions of alternative methods for many univariate and **multivariate distributions** are to be found in [10, 12, 17], and [35]. Efficiency can often be increased by the use of variance reduction methods such as importance sampling (*see* **Numerical Integration**), **stratification**, and using **antithetic** or control variates (see [32]).

Monte Carlo Inference

When a likelihood is difficult to construct, simulation techniques may be used to produce an approximation to the **likelihood**. As a simple illustration, suppose the random variable Y is given by the convolution:

$$Y = X_1 + X_2, \tag{4}$$

where X_1 has a **gamma distribution** and X_2 has an independent normal distribution. It is easy to simulate from both X_1 and X_2, and hence from Y, but it is generally not straightforward to write down the density function of Y, and hence the likelihood function. Many examples of this nature occur in statistics – for example in the theory of **queues** (see [20]).

Suppose we observe a **random sample** $\{y_i, 1 \leq i \leq n\}$, from a model with probability density function $f(y_i; \boldsymbol{\theta})$. To obtain the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$, we need to form,

$$l(\boldsymbol{\theta}; y) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}),$$

and then maximize this with respect to θ.

In Monte Carlo inference we use

$$l^*(\boldsymbol{\theta}; y) = \sum_{i=1}^{n} \log \hat{f}(y_i; \boldsymbol{\theta}),$$

where simulation has been used to form the density estimate $\hat{f}$.

Diggle & Gratton [13] used a kernel approach (*see* **Density Estimation**), resulting in:

$$\hat{f}(y) = \frac{1}{(sh)} \sum_{k=1}^{s} K\left(\frac{y - x_k}{h}\right), \tag{5}$$

where $\{x_k, 1 \leq k \leq s\}$ is a simulated sample from $f(y; \boldsymbol{\theta})$, $K(u)$ is a kernel function, given by

$$K(u) = \begin{cases} 0.75(1 - u^2), & -1 \leq u \leq 1, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

and h determines the smoothness of the approximation.

Diggle & Gratton [13] discuss the choice of s, h, and K. For other applications, see [9] and [19]. In the latter case, in the context of dependent data, MCMC methods are used.

Monte Carlo Testing

After a model has been fitted to a data set, multiple samples can be obtained by simulating from the fitted model, and the model may be fitted in turn to each of these samples. The resulting sets of parameter estimates may be used for inference. For example, **confidence intervals** may be obtained using the percentile method, which selects parameter cutoff points with a percentage, such as 5%, of simulated values lying outside the resulting interval. This general approach is called the parametric bootstrap, as replicate samples are obtained from a fitted model.

The **goodness of fit** of a model to data may be measured in a variety of ways – for example by means of a deviance (*see* **Generalized Linear Model**) or a Pearson chi-square statistic (*see* **Chi-square Tests**). Whatever goodness-of-fit statistic is selected, it may also be calculated for each of the samples simulated from the model fitted to the original data in the parametric bootstrap. In these cases we know that the model is correct, as we simulate from it. The values of the statistic from the simulated samples therefore provide a benchmark set against which to compare the value obtained from the original data. Due to Barnard [1], this approach is considered further by Hope [22] and Marriott [29], who discuss **power**. It may also be used to compare nonnested models, [7, 40] (*see* **Separate Families of Hypotheses**). Monte Carlo tests have been especially useful in spatial analysis (*see* **Epidemic Models, Spatial**).

An extension to dependent data is considered by Besag & Clifford [3]. Monte Carlo exact tests are derived by Forster et al. [18] using Gibbs sampling (*see* **Markov Chain Monte Carlo**).

Computing and Analysis

Computer simulation is widely utilized to mimic the rules of complex systems. The resulting simulation models may then be used to study the effect of changes to those systems. Usually sensitivity studies (*see* **Sensitivity Analysis**) also need to be carried out, in which predictions are investigated for perturbations of the parameter values adopted in the model. Two examples are provided by Duncan & Curnow [14] and Byrom & Gettinby [8]. Such models may be programmed in languages such as C, FORTRAN, and MATLAB. However the simulations regularly require standard bookkeeping operations, and specialized languages exist to simplify such tasks, such as GPSS, SIMSCRIPT, and SIMULA (see [30]). Simulated sequences need examination in order to decide whether they have reached equilibrium, and this is also a problem in MCMC work (see [23]). Analysis is frequently complicated by the presence of **serial correlation**. Moran [31] investigated the use of "batching", in which a sequence is divided into batches, and analysis then proceeds using batch means. Overviews are provided by Fishman [17] and Kleijnen & Groenendaal [25]. Simulation experiments allow statisticians to make use of their design skills (see, for example, [36]).

References

[1] Barnard, G.A. (1963). Discussion of Professor Bartlett's paper, *Journal of the Royal Statistical Society, Series B* **25**, 294.

[2] Bartlett, M.S. (1990). Chance or chaos?, *Journal of the Royal Statistical Society, Series A* **153**, 321–348.

[3] Besag, J. & Clifford, P. (1989). Generalized Monte Carlo significance tests, *Biometrika* **76**, 633–642.

[4] Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems, *Statistical Science* **10**, 3–66.

[5] Box, G.E.P. & Müller, M.E. (1958). A note on the generation of random normal deviates, *Annals of Mathematical Statistics* **29**, 610–611.

[6] Brooks, S.P. & Morgan, B.J.T. (1995). Optimization using simulated annealing, *Statistician* **44**, 241–257.

[7] Brooks, S.P., Morgan, B.J.T., Ridout, M.S. & Pack, S.E. (1997). Finite mixture models for proportions, *Biometrics* **53**, 1097–1115.

[8] Byrom, W. & Gettinby, G. (1992). Using the computer model ECFXPERT to study ticks and East Coast Fever. *Insect Science and its Applications* **13**, 527–535.

[9] Crowder, M. (1994). Least squares with simulated means for a problem in fibre strength testing, *Applied Statistics* **43**, 109–116.

[10] Dagpunar, J. (1988). *Principles of Random Variate Generation*. Clarendon Press, Oxford.

[11] Davison, A.C., Hinkley, D.V. & Schechtman, E. (1987). Efficient bootstrap simulation, *Biometrika* **74**, 555–566.

[12] Devroye, L. (1986). *Non-uniform Random Variate Generation*. Springer-Verlag, New York.

[13] Diggle, P.J. & Gratton, R. (1984). Monte Carlo methods of inference for implicit statistical models, *Journal of the Royal Statistical Society, Series B* **46**, 193–227.

[14] Duncan, I.B. & Curnow, R.N. (1978). Operational research in the health and social services. *Journal of the Royal Statistical Society, Series A* **141**, 153–194.

[15] Edgington, E.S. (1987). *Randomization Tests*, 2nd Ed. Marcel Dekker, New York.

[16] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[17] Fishman, G.S. (1996). *Monte Carlo Concepts, Algorithms and Applications*. Springer-Verlag, New York.

[18] Forster, J.J., McDonald, J.W. & Smith, P.W.F. (1996). Monte Carlo exact conditional tests for log-linear and logistic models, *Journal of the Royal Statistical Society, Series B* **58**, 445–454.

[19] Geyer, C.J. & Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data, *Journal of the Royal Statistical Society, Series B* **54**, 657–700.

[20] Gross, D. & Harris, C.M. (1974). *Fundamentals of Queueing Theory*. Wiley, Toronto.

[21] Hinkley, D.V. (1988). Bootstrap methods, *Journal of the Royal Statistical Society, Series B* **50**, 321–337.

[22] Hope, A.C.A. (1968). A simplified Monte Carlo significance test procedure, *Journal of the Royal Statistical Society, Series B* **30**, 582–598.

[23] Jennison, C. (1993). Discussion on Gibbs Sampler and other MCMC methods, *Journal of the Royal Statistical Society, Series B* **55**, 54–56.

[24] Kinderman, A.J. & Monahan, J.F. (1977). Computer generation of random variables using the ratio of normal deviates, *ACM Transactions on Mathematical Software* **3**, 257–260.

[25] Kleijnen, J.P.C. & Groenendaal, W. (1992). *Simulation. A Statistical Perspective*. Wiley, New York.

[26] Knuth, D.E. (1981). The Art of Computer Programming, Vol. 1, *Fundamental Algorithms*. Addison-Wesley, Reading.

[27] Lawrance, A.J. (1990). Discussion of the paper by Bartlett, *Journal of the Royal Statistical Society, Series A* **153**, 335–336.

[28] Manly, B.F.J. (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman & Hall, London.
[29] Marriott, F.H.C. (1979). Barnard's Monte Carlo tests: How many simulations?, *Applied Statistics* **28**, 75–77.
[30] Mitrani, I. (1982). *Simulation Techniques for Discrete Event Systems*. Cambridge University Press, Cambridge.
[31] Moran, P.A.P. (1975). The estimation of standard errors in Monte Carlo simulation experiments, *Biometrika* **62**, 1–4.
[32] Morgan, B.J.T. (1984). *Elements of Simulation*. Chapman & Hall, London.
[33] Park, S.K. & Miller, K.W. (1988). Random Number Generators – Good ones are hard to find, *Communications of the Association for Computing Machinery* **32**, 1192–1201.
[34] Reeves, C.R. (1993). *Modern Heuristic Techniques*. Blackwell Scientific, Oxford.
[35] Ripley, B.D. (1987). *Stochastic Simulation*. Wiley, New York.
[36] Schruben, L.W. & Margolin, B.H. (1978). Pseudo random number assignment in statistically designed simulation and distribution sampling experiments, *Journal of the American Statistical Association* **73**, 504–525.
[37] Smith, A.F.M. & Roberts. G.O. (1993). Bayesian Computation via the Gibbs sampler and related Markov Chain Monte Carlo methods, *Journal of the Royal Statistical Society, Series B* **55**, 3–23.
[38] Wakefield, J.C., Gelfand, A.E. & Smith, A.F.M. (1991). Efficient generation of random variables via the ratio-of-uniforms method, *Statistics and Computing* **1**, 129–133.
[39] Wichmann, B.A. & Hill, I.D. (1982). Algorithm AS183: an efficient and portable pseudo-random number generator, *Applied Statistics* **31**, 188–190.
[40] Williams, D.A. (1982). GLIM and Hirayama's data, *Royal Statistical Society News and Notes* **9**, 7.
[41] Young, G.A. (1994). Bootstrap: more than a stab in the dark?, *Statistical Science* **9**, 382–415.

(*See also* **Computer-intensive Methods**)

BYRON J.T. MORGAN

Simultaneous Confidence Intervals

The terminology *simultaneous confidence intervals* (SCI) refers to a **confidence** region for a multivariate parameter $\boldsymbol{\phi}$, comprised of individual confidence intervals for the components of $\boldsymbol{\phi}$. For example, in a medical application whose objective is to compare k treatment means with a control from $\mathrm{N}(\mu_i, \sigma^2)$, $1 \leq i \leq k$, and $\mathrm{N}(\mu_0, \sigma^2)$ populations respectively, $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta} = (\mu_0, \mu_1, \ldots, \mu_k, \sigma)$, while $\phi_i = \mu_i - \mu_0$, $1 \leq i \leq k$. SCI provide an overall assurance at a specified confidence level of the simultaneous correctness of all the statements concerning the differences ϕ_i, $1 \leq i \leq k$. Typically, $\boldsymbol{\phi}$ has a finite number of components $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ but sometimes an infinite number as in a confidence band for a **regression** curve.

Denote by $\mathbf{x}$ the data collected in an experiment. A parametric model $P_{\boldsymbol{\theta}}$ is a family of distributions describing a random vector $\mathbf{X}$ which models the data collection process. For a fixed $0 < \alpha < 1$, a $(1-\alpha)100\%$ confidence region $D(\mathbf{X})$ for $\boldsymbol{\phi}$ is a random region satisfying

$$P_{\boldsymbol{\theta}}[D(\mathbf{X}) \ni \boldsymbol{\phi}] \geq 1 - \alpha, \tag{1}$$

no matter what the value of $\boldsymbol{\theta}$. Some authors use equality in (1). An SCI for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ is a special case of (1) of the form

$$P_{\boldsymbol{\theta}}[l_i(\mathbf{X}) \leq \phi_i \leq u_i(\mathbf{X}), \text{ for all } 1 \leq i \leq k] \geq 1 - \alpha. \tag{2}$$

The left and right endpoints in (2) provide interval estimates for $\boldsymbol{\phi}$ endowed with the frequentist interpretation that in replicated experiments, in the long run, these intervals, constructed from the data, will cover every corresponding parameter simultaneously $(1-\alpha)100\%$ of the time.

The most universally valid (and "ancient" to quote Miller [5, p. 67]) approach relies on Boole's inequality (or the first **Bonferroni inequality**),

$$P\left(\bigcap_{i=1}^{k} A_i\right) = 1 - P\left(\bigcup_{i=1}^{k} \overline{A}_i\right) \geq 1 - \sum_{i=1}^{k} P(\overline{A}_i) \geq 1 - \alpha,$$

applied to a set of univariate $(1-\alpha/k)100\%$ confidence intervals

$$P_{\boldsymbol{\theta}}[l_i(\mathbf{X}) \leq \phi_i \leq u_i(\mathbf{X})] \geq \frac{1-\alpha}{k},$$

which yields (2) from the choice $A_i = \{l_i(\mathbf{X}) \leq \phi_i \leq u_i(\mathbf{X})\}$ and $\overline{A}_i$, the complement of A_i.

Perhaps the most widely studied SCI are those for **multiple comparisons** of treatment means for which there are three (among others) well-known competitors to the Boole–Bonferroni approach, each having their own advantages depending on the context. Scheffé's [6, 7] method as an SCI valid for all contrasts in an **analysis of variance** (ANOVA) was originally proposed as a follow-up to provide insight when the ***F*-test** rejects the null hypothesis (*see* **Multiple Comparisons**). Tukey's [8] approach, based on the **studentized range** distribution, was derived for comparing the special contrasts of pairwise treatment mean differences. Dunnett's [1] elegant method used the **multivariate *t* distribution** for the further special case of comparing treatments with a control, rather than all differences in means. The natural question of which method provides the shortest intervals has been studied, for instance, by Ury [9] and Einot & Gabriel [2] (*see also* **Multiple Comparisons**, for reference to Duncan's multiple range test).

As a numerical illustration, we present an example adapted from Dunnett [1] of blood count measurements on three groups of animals (see Table 1). The sample standard deviation is $s = 1.175$ and the corresponding multivariate t critical value with 12 df is 2.50. Based on Dunnett's method, the 95% SCI for $\mu_{\mathrm{A}} - \mu_{\mathrm{C}}$ and $\mu_{\mathrm{B}} - \mu_{\mathrm{C}}$ are $(-1.25, 2.55)$ and $(0.85, 4.41)$ respectively, while the Boole–Bonferroni intervals are $(-1.30, 2.60)$ and $(0.80, 4.46)$, which are

Table 1 Blood counts (millions of cells per cubic millimeter)

	Control	Drug A	Drug B
	7.40	9.76	12.80
	8.50	8.80	9.68
	7.20	7.68	12.16
	8.24	9.36	9.20
	9.84		10.55
	8.32		
Means	8.25	8.90	10.88

a little wider. Thus, with 95% confidence we may conclude that drug A raises the blood count by an amount between -1.25 and 2.55 million cells per cubic millimeter, while drug B raises the blood count by an amount between 0.85 and 4.41 million cells. For comparison, Scheffé's intervals, $(-1.47, 2.77)$ and (0.64, 4.62), are wider, but have higher confidence than 95% since we are dealing only with two contrasts, not all possible contrasts.

Applications of SCI have been developed in many other settings, including ANOVA, **multivariate analysis of variance** (MANOVA), **analysis of covariance, Hotelling's T^2** test, regression coefficients (*see* **Multiple Linear Regression**), growth curve analysis (*see* **Nonlinear Growth Curve**), and **variance components**. References may be found in the extensive bibliographies in the books by Miller [5], Hochberg & Tamhane [3], and Hoppe [4].

References

[1] Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control, *Journal of the American Statistical Association* **50**, 1096–1121.

[2] Einot, I. & Gabriel, K.R. (1975). A study of the powers of several methods of multiple comparisons, *Journal of the American Statistical Association* **70**, 574–583.

[3] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[4] Hoppe, F.M. (1993). *Multiple Comparisons, Selection and Applications in Biometry*. Marcel Dekker, New York.

[5] Miller, R.G., Jr (1981). *Simultaneous Statistical Inference*, 2nd Ed. Springer-Verlag, New York.

[6] Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika* **40**, 87–104.

[7] Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

[8] Tukey, J.W. (1953). *The Problem of Multiple Comparisons*, Mimeographed Notes, Princeton University. Reprinted in *The Collected Works of John W. Tukey*, Vol. VIII – *Multiple Comparisons: 1948–1983*, H.I. Braun, ed. Chapman & Hall, New York, 1994.

[9] Ury, H.K. (1979). A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary sample sizes, *Technometrics* **18**, 89–97.

(*See also* **Estimation, Interval**; **Experiment-wise Error Rate**; **Simultaneous Inference**; **Tolerance Interval**; **Tolerance Region**)

TUHAO CHEN & FRED M. HOPPE

Simultaneous Inference

In a broad sense, *simultaneous inference* or *the multiplicity problem* includes statistical procedures that assess either more than one parameter in the course of an experiment or one parameter repeatedly in the course of an experiment. The problems studied include **multiple comparisons** (several parameters, one for each treatment group), multiple endpoints (several parameters for each treatment group), **sequential** methods or group sequential methods (one parameter assessed as the data accumulate during the experiment), and **longitudinal data analysis** (one parameter assessed at different time points). Recent work has focused on hypothesis testing problems in combination; for example, group sequential methods for multiple comparisons or multiple endpoints. Simultaneous inference includes point and interval estimation (*see* **Estimation**) as well as **hypothesis testing** and other methods of inference, such as ranking and selection. We give examples of hypothesis testing for multiple comparisons and multiple endpoints, simultaneous confidence intervals, as well as a more complex example involving group sequential hypothesis testing with multiple comparisons. For more detailed information we refer the reader to a number of related articles in this Encyclopedia: **Multiplicity in Clinical Trials; Multiple Comparisons; Multiple Endpoints, Multivariate Global Tests; Multiple Endpoints, *P* Level Procedures; Data and Safety Monitoring;** and **Longitudinal Data Analysis, Overview**.

Philosophies of Simultaneous Inference

When is it appropriate to consider several parameters simultaneously? In his classic book, Miller [12] considered a family to be those statements about parameters resulting from an individual experiment of a single researcher in a majority of instances. He went on to say: "There are no hard and fast rules for where family lines should be drawn, and the statistician must rely on his own judgment for the problem at hand." Hochberg & Tamhane [6] defined a family as any collection of inferences for which it is meaningful to take into account some combined measure of errors. This definition still leaves it to the judgment of the investigators which inferences should be taken as a family.

For a family F of inferences about a set of parameters, let $N(F)$ denote its cardinality and let P be the set of statistical procedures used to decide if each statement in F is true. In undertaking P, we want some protection against falsely rejecting statements in F when they are, indeed, true. Let M (a function of both F and P and the data) be the (random) number of false positives. The familywise (experimentwise) error rate is the probability of making at least one false positive conclusion, i.e. $P\{M > 0\}$. The per-family (per-experiment) error rate is the expected number of false inferences, i.e. $E\{M\}$. The per-comparison (comparisonwise, per-statement) error rate is the expected number of false inferences divided by the number of inferences, i.e. $E\{M\}/N(F)$. [The per-comparison error rate cannot be defined in many cases when $N(F)$ is infinite.] Using elementary probability, it is clear that

$$\frac{E\{M\}}{N(F)} \leq P\{M > 0\} \leq E\{M\},$$

so that control over the per-family error rate is stronger than control over the familywise error rate, which is, in turn, stronger than control over the per-comparison error rate. The three error rates depend on the true configuration of the parameters.

In hypothesis testing, the family of inferences, F, is taken to be the set of statements included in the **null hypothesis**. Procedures that control the familywise error rate for F and all subsets of F are said to provide *strong control*. Procedures that control the familywise error rate for F alone are said to provide *weak control*.

The difference between weak and strong control of the familywise error rate may be illustrated by one-way **analysis of variance** (ANOVA) with $A(> 3)$ groups. Suppose the first $A - 1$ means are equal and the Ath mean is far different. Then a procedure that provides weak control of the familywise error would protect against the rejection of the global hypothesis that all of the means are equal, but would provide no protection against false rejection of the equality of the first $A - 1$ means. Strong protection would protect against both.

While many authors argue that control over the familywise error rate is appropriate [6, 12, 23], others support controlling the per-family error rate for

finite families [18] or even the per-comparison error rate [2, 13, 17]. The relative value of each type of error rate control will depend on what the investigators are trying to accomplish. By way of illustration, suppose four experimenters could have collected exactly the same data in a one-way ANOVA but be interested in different error rates because they are pursuing different goals. The first investigator might have several prespecified hypotheses in mind involving pairwise comparisons and may want to address each one individually. This investigator might be primarily interested in controlling the per-comparison error rate and test each comparison at level α. The second investigator might be searching for effective drugs in a pilot development program (e.g. a **Phase II trial**). In this case a significant initial F test (*see* **Analysis of Variance**) might lead to a larger **clinical trial**, whereas a nonsignificant F test and a significant result on a single drug using a per-comparison error rate might lead to another pilot study on that drug, and no significant per-comparison tests would lead to a negative conclusion. In this situation the investigator is interested in *both* the familywise and per-comparison error rates. The third investigator might test the global hypothesis of the equality of all of the means (controlling the familywise error rate), but having done that, perform pairwise treatment comparisons controlling the per-comparison error rate. The fourth investigator might be performing a definitive trial comparing several treatments and consider it important to control the type I error rate on both the entire family of null hypotheses and on every subset of that family.

The first and second investigators above required per-comparison control of the type I error rate, although with differing families of inferences. The third investigator required weak control of the familywise error and the fourth required strong control of the familywise error rate. Thus, depending on the situation at hand, different degrees of type I error rate control will be considered appropriate. When seeking new drug approval, a manufacturer may conduct a trial of several doses vs. a placebo and argue that the per-comparison error rate with a placebo is of interest, but those with the power to approve the new drug may well disagree.

Prior to data analysis, careful specification of the medical question(s) that the analysis is intended to answer will help investigators focus on what their family of inferences should be and what kind of error control is appropriate in a given situation. Hypotheses that are data driven (i.e. that are tested without prior specification) require tentative conclusions which need further, definitive test no matter what ***P* value** results. Tukey [24] says, "these give us *hints*".

Examples of Simultaneous Inference

Henceforth we assume that investigators have agreed that certain inferences constitute a family. Many procedures have been proposed for various simultaneous inference situations. Here we give three examples and outline several procedures which might be used.

Example 1. Simultaneous Confidence Intervals for all Treatment Versus a Control

Suppose we have a clinical trial with A treatment arms. Let $X_{ij}, i = 1, \ldots, A, j = 1, \ldots, M$ be the jth observation of the ith treatment arm. Assume X_{ij} has mean μ_i and unknown variance σ^2. Suppose that μ_1 is a control treatment and we are interested in simultaneous, two-sided $100(1-\alpha)\%$ confidence intervals for

$$\mu_j - \mu_1, \quad j = 2, 3, \ldots, A. \tag{1}$$

F is a set of $N(F) = A - 1$ confidence intervals of the pairwise comparisons with μ_1. If we assume that the underlying data have a normal distribution, then we may use the procedure of Dunnett [3]. The **simultaneous** two-sided $100(1-\alpha)\%$ confidence intervals for treatment minus control differences, $\mu_j - \mu_1$, are given by

$$(\mu_j - \mu_1) \in \left[\overline{Y}_j - \overline{Y}_1 \pm |T|^{(\alpha)}_{A-1,\nu,1/2} S \left(\frac{2}{M} \right)^{1/2} \right], \quad j = 2, 3, \ldots, A, \tag{2}$$

where S^2 is the mean squared error estimate and $|T|^{(\alpha)}_{A-1,\nu,1/2}$ is the upper α point of the maximum of the absolute value of an A-variate **multivariate *t* distribution** with $\nu = A(M-1)$ **degrees of freedom** and **correlation** coefficients 1/2. The critical values of $|T^{(\alpha)}|_{A-1,\nu,1/2}$ are tabulated, c.f. [[6],

Table 5]. When the sample sizes are not equal, the correlations between the $\overline{Y}_j - \overline{Y}_1$ are no longer 1/2 so that calculations of the critical values are difficult and approximations must be used. The joint confidence intervals have the property that when many experiments are undertaken, in $100(1-\alpha)\%$ of them, *all* of the confidence intervals will contain the true parameters.

Example 2. Hypothesis Testing for all Pairwise Comparisons

A classic example of simultaneous statistical inference considers the A-sample multiple comparisons problem. Suppose we have a clinical trial with A treatment arms. Let $X_{ij}, i = 1, \ldots, A, j = 1, \ldots, M$, be the jth observation of the ith treatment arm. Assume X_{ij} has mean μ_i and unknown variance σ^2. Suppose we are interested in testing

$$\mathrm{H}_0 : \mu_i = \mu_j, \quad i \neq j, i, j = 1, 2, \ldots, A, \quad (3)$$

vs.

$$\mathrm{H}_1 : \mu_i \neq \mu_j, \quad \text{for some } i \neq j, \quad i, j = 1, 2, \ldots, A.$$

Then F is a set of $N(F) = A(A-1)/2$ statements of inferences on all of the pairwise comparisons.

If we assume that the underlying data have a normal distribution, then we may first perform a one-way ANOVA F test to test (3). If $A = 3$, then strong control and weak control coincide, so assume $A > 3$. If the F test rejects H_0 at level α, we may follow with pairwise t tests each at level α, using the pooled estimate of σ^2, a procedure known as Fisher's Least Significant Difference (LSD) for weak control of the familywise error rate [4].

If we wished strong control of the familywise error rate, then we would have to enlarge the family F and consider all statements of equality of means of all subsets in the null hypothesis. That is, we would need to consider the null hypothesis

$$\mathrm{H}_0 : \mu_1 = \cdots = \mu_A \quad (3')$$

and all alternatives, beginning with $A-1$ means equal and one mean different, $A-2$ means equal and the other two means equal but different from the first $A-2$ means, $A-2$ means equal and the other two means each different, etc. We could then use the closed F procedure or the closed Newman–Keuls procedure of Begun & Gabriel [1]. In these procedures, if the hypothesis $(3')$ is not rejected at level α, then we stop and say there are no treatment differences at level α. If $(3')$ is rejected, then we consider all subsets of the null hypothesis in $(3')$ in a particular stepdown manner. If there is no significant difference at level α when testing the null hypothesis of equality of $p(= A-1, A-2, \ldots, 3, 2)$ of the means, then we say the p treatments are homogeneous and do not test further subsets of these. Strong control of α is obtained by simultaneously testing, in addition, disjoint subsets of the null hypothesis in a particular manner. For example, if $A = 4$, then the simultaneous test of $\{\mu_1 = \mu_2$ and $\mu_3 = \mu_4\}$ is undertaken by testing each pair of hypotheses at level $1-(1-\alpha)^{1/2}$. If one of the hypotheses in the intersection set is not rejected at level $1-(1-\alpha)^{1/2}$, then *neither* is rejected. This implies that in determining whether there are pairwise differences based on this procedure, the homogeneity of certain means depends on the homogeneity of other means. In the case $A = 4$, if we do not find a difference between μ_1 and μ_2 at level $1-(1-\alpha)^{1/2}$, then we would also not find μ_3 and μ_4 to differ. While this initially seems strange, it is the price that this procedure pays for strong control of α within the family of hypotheses. For A arms, the tests on the disjoint subsets are performed at level $1-(1-\alpha)^{p/A}$, where p is the number of hypotheses undergoing test.

Example 3. Hypothesis Testing for Multiple Endpoints in a Clinical Trial

Suppose we have a two-armed clinical trial comparing the efficacy of an experimental and a standard treatment and efficacy is reflected by multiple patient characteristics. A statistical formulation of this problem is

$$\mathrm{H}_0 : \boldsymbol{\mu}_\mathrm{E} = \boldsymbol{\mu}_\mathrm{S} \quad (4)$$

vs.

$$\mathrm{H}_1 : \boldsymbol{\mu}_\mathrm{E} - \boldsymbol{\mu}_\mathrm{S} = \boldsymbol{\delta}(> \mathbf{0}),$$

where $\boldsymbol{\mu}_\mathrm{E}$ is the mean of the vector-valued observations of the experimental group, $\boldsymbol{\mu}_\mathrm{S}$ is the mean of the vector-valued observations of the standard treatment, and $\boldsymbol{\delta}$ is a fixed vector of relative treatment differences. The formulation of H_1 is specific to clinical

trials, for which we are interested in the **alternative hypothesis** that one treatment is better than the other with respect to k multiple endpoints. F is a set of statements about the equality of the components of the vectors $\boldsymbol{\mu}_i$ and has cardinality $N(F) = k$.

The rationale for the formulation of H_0 in (4) is that we will control the type I error when there is no difference between the experimental and standard treatments. The formulation of H_1 serves to direct power to an alternative of particular interest, where all endpoints derive a meaningful benefit, a situation in which high power would be desired. In this context, testing efficacy with multiple endpoints need not be viewed as a simultaneous inference problem; the test statistic may be viewed as a univariate measure of efficacy. (For further discussion, see [15].) However, a family of hypotheses becomes of interest when we attempt to identify the individual endpoints that benefit from treatment.

If we assume that the underlying data have a normal distribution, we may test (4) using one of the linear combination test statistics proposed by O'Brien [14] and Tang et al. [20]. Following rejection of the global null hypothesis (4), we may make inferences on the individual endpoints by considering the family F. Lehmacher et al. [10] proposed a stepdown procedure which maintains strong control of α. Following rejection of the global null hypothesis (4), proceed by testing all subsets of size $A - 1$ endpoints, with each test performed at the α level using the same procedure P as used on the A endpoints. If a subset of $A - 1$ endpoints does not differ (at the α level), then conclude that the treatments do not differ with respect to each of those $A - 1$ endpoints. If, however, the test on $A - 1$ endpoints is significant at level α, continue with all subsets of $A - 2$ endpoints. Once a set of endpoints does not differ, no further tests are done on those endpoints and this avoids contradictions. The procedure is continued until it is determined whether the treatments differ with respect to each endpoint. An alternative approach, conferring only weak protection of α, is to follow a significant global test by tests only on the individual endpoints, each conducted at the same α level as used for the global test.

It is worth noting that obtaining strong control in the multiple endpoint problem in clinical trials (Example 3) is easier than obtaining it in the multiple comparisons problem (Example 2) because there is no need to simultaneously test disjoint sets of subset hypotheses in the multiple endpoint case.

An alternative approach which does not require the normality assumption, is the rank sum statistic proposed by O'Brien [14] and the stepdown procedures are analogous. The rank procedure does not consider the correlation between endpoints. This may or may not be desirable, but it should be noted that weighting endpoints according to the correlation structure produces a different measure of overall efficacy.

Other procedures for multiple endpoints have been proposed by Westfall & Young [25], Lefkopoulou & Ryan [9], and Tang et al. [21]. O'Brien & Geller [15] discuss the importance of the specific question being asked when choosing a procedure.

An Example of a More Complex Multiplicity Problem

Recent developments in simultaneous inference entail theoretical results combining classical multiple comparison problems, multiple endpoints, or longitudinal data analysis along with group sequential monitoring (cf. [8, 11, 19], and [22]). We give one conceptual example.

The following trial design was initially proposed for a Raynaud's treatment study. Raynaud's disease is characterized by episodes or "attacks" in cold weather of decreases in blood flow through the veins in the extremities, resulting in extreme cold and pain (and possible loss of function) in the fingers. Patients would be randomized to receive a pharmacologic agent (a long-acting calcium channel blocker), a thermal biofeedback treatment, or a placebo pill. The primary outcome would be change from baseline in the number of Raynaud's episodes per day. Baseline incidence would be assessed over a month period in winter and outcome incidence would be assessed over the same month, a year later. The trial would be undertaken on two cohorts of patients during two successive winters and would be monitored for efficacy at the end of the first season so that if there were a strong treatment effect, early stopping could be considered (*see* **Data and Safety Monitoring**).

The major question of the trial may be formulated in several ways; depending on the formulation, an appropriate statistical procedure could be applied. The global null hypothesis of no difference in the treatment effects overall could be tested using a

group sequential F test [7, 16]. The question of whether either of the active treatments differs from the placebo requires group sequential monitoring of treatments vs. a control. The question of whether any of the pairwise treatment comparisons differ requires group sequential methods for all pairwise comparisons (cf. [5]). The relevant error rate to be controlled would need to be specified in the study protocol.

References

[1] Begun, J.M. & Gabriel, K.R. (1981). Closure of the Newman-Keuls multiple comparisons procedure, *Journal of the American Statistical Association* **76**, 241–245.

[2] Duncan, D. (1955). Multiple range and multiple F tests, *Biometrics* **11**, 1–42.

[3] Dunnett, C. (1955). A multiple comparison procedure for comparing several treatments with a control, *Journal of the American Statistical Association* **50**, 1096–1121.

[4] Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh/London.

[5] Follmann, D.A., Proschan, M.A. & Geller, N.L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials, *Biometrics* **50**, 325–336.

[6] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[7] Jennison, C. & Turnbull, B. (1991). Exact calculations for sequential t, χ^2 and F tests, *Biometrika* **78**, 133–141.

[8] Lee, J.W. & DeMets, D. (1991). Sequential comparison of changes with repeated measurements data, *Journal of the American Statistical Association* **86**, 757–762.

[9] Lefkopoulou, M. & Ryan, L. (1993). Global tests for multiple binary outcomes, *Biometrics* **49**, 975–988.

[10] Lehmacher, W., Wassmer, G. & Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate, *Biometrics* **47**, 511–521.

[11] Lin, D.Y. (1991). Nonparametric sequential testing in clinical trials with incomplete multivariate observations, *Biometrika* **78**, 123–131.

[12] Miller, R.G. (1961). *Simultaneous Statistical Inference*, Rev. Ed. 1985 Springer-Verlag. New York.

[13] O'Brien, P.C. (1983). The appropriateness of analysis of variance and multiple comparison procedures, *Biometrics* **39**, 787–789.

[14] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1087.

[15] O'Brien, P.C. & Geller, N.L. (1997). Interpreting tests for efficacy in clinical trials with multiple endpoints, *Controlled Clinical Trials*, **18**, 222–227.

[16] Proschan, M.A., Follmann, D.A. & Geller, N.L. (1994). Monitoring multi-armed trials, *Statistics in Medicine* **13**, 1441–1452.

[17] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company, Boston.

[18] Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures, *Annals of Mathematical Statistics* **43**, 398–411.

[19] Su, J.Q. & Lachin, J.M. (1992). Group sequential distribution-free methods for the analysis of multivariate observations, *Biometrics* **48**, 1033–1042.

[20] Tang, D.-I., Geller, N.L. & Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints, *Biometrics* **49**, 23–30.

[21] Tang, D.-I., Gnecco, C. & Geller, N.L. (1989). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials, *Biometrika* **76**, 577–583.

[22] Tang, D.-I., Gnecco, C. & Geller, N.L. (1989). Design of group sequential clinical trials with multiple endpoints, *Journal of the American Statistical Association* **84**, 776–778.

[23] Tukey, J.W. (1953). *The Problem of Multiple Comparisons*. Mimeographed monograph.

[24] Tukey, J.W. (1991). The philosophy of multiple comparisons, *Statistical Science* **6**, 100–116.

[25] Westfall, P.H. & Young, S.S. (1993). *Resampling-Based Multiple Testing*. Wiley, New York.

[26] Wu, M. & Lan, K.K.G. (1992). Sequential monitoring for comparison of changes in a response variable in clinical studies, *Biometrics* **48**, 765–779.

NANCY L. GELLER & PETER C. O'BRIEN

Simple Random Sampling

The following definition of *simple random sampling* implies that a sample of n enumeration units is selected from a population of N enumeration units *without replacement* (*see* **Sampling With and Without Replacement**).

A simple random sample of n enumeration units from a population of N enumeration units is one in which each of the $\binom{N}{n}$ possible samples has the same probability of selection; namely, $1/\binom{N}{n}$ (cf. [1]).

For example, if a particular population consists of 15 households and we wish to take a simple random sample of 10 households, then there are $\binom{15}{10} = 15!/(10!5!) = 3003$ possible samples, with each having a probability of being selected of $1/3003 = 0.000333$. The above definition is much more restrictive than the term **random sampling**, which can have many different meanings depending on the specific context in which it is being used.

In simple random sampling, no prior knowledge concerning characteristics of the enumeration units other than their labels or identification numbers is used in selecting the sample. Thus, other sampling designs such as **stratified sampling** which do make use of such knowledge generally yield estimates that have lower sampling variability than those obtained from a simple random sample of the same number of enumeration units. Likewise, sampling designs such as **cluster sampling** and **multistage sampling** can be accomplished at lower field costs than simple random sampling. Thus, sample surveys, especially those involving sampling of human populations over large geographic areas, are rarely based on simple random sampling.

Table 1

Population characteristic	Estimate	Standard error of estimate
Mean	$\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$	$\mathrm{se}(\overline{x}) = \dfrac{\sigma_x}{\sqrt{n}}\left(\dfrac{N-n}{N-1}\right)^{1/2}$
Total	$x' = N\overline{x}$	$\mathrm{se}(x') = \dfrac{N\sigma_x}{\sqrt{n}}\left(\dfrac{N-n}{N-1}\right)^{1/2}$

Estimates of population totals and **means** are shown in Table 1 along with their **standard errors**. These are appropriate for simple random sampling when the **unit of analysis** is the enumeration unit. In Table 1, the term σ_x is the **standard deviation** of the distribution of the variable x in the population. Also, the term $[(N-n)/(N-1)]^{1/2}$ is known as the **finite population correction** and is close to unity when n is considerably smaller than N.

Reference

[1] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

PAUL S. LEVY

SIR Epidemic Models

SIR epidemic models describe the spread of infectious diseases that follow the scheme Susceptible $\rightarrow$ Infected $\rightarrow$ Removed. This scheme means that a susceptible, if ever "adequately" contacted by an infective, becomes infected for some period of time (the infectious period) after which it recovers and is immune to the disease (other removal states are possible after adjustment).

Our review is only concerned with the stochastic approach, which is more realistic, but also more complex, than the deterministic approach (*see* **Epidemic Models, Stochastic**; **Epidemic Models, Deterministic**). Standard and recent treatises on the subject are [1, 3, 11, 12, 16, 31] (see also [32]). For deterministic models, the reader is referred to [2, 18].

The central situation is when the population is closed (there is neither birth/death nor immigration/emigration), homogeneous (all S, I, or R individuals have similar behaviors), and independently mixing (meetings between any pair of individuals occur independently of each other). In the course of time, the population state is represented by the random vector giving the numbers S_t of susceptibles, I_t of infectives, and Z_t of removed cases present at time $t \geq 0$. Initially, $S_0 = n$, $I_0 = m$, $Z_0 = 0$ say, and the population being closed, $S_t + I_t + Z_t = n + m$ for all t (thus, any two of these variables specify the epidemic process). The epidemic ceases as soon as there are no more infectives in the population, which arises with probability one after a finite time T. Then, S_T is the ultimate number of susceptibles escaping the disease, and $Z_T = n + m - S_T$ (the final size) is the total number of infected cases, including the m initial ones. The statistics S_T is of great interest in theory and practice, and it has received by far the most attention in the literature. Moreover, it is generally very difficult to evaluate the epidemic process in transient condition.

To begin with, we will examine SIR models that are built with a Markovian structure (*see* **Markov Chains**). In their majority, the proposed models constitute variants or extensions of two well-known epidemic models, named the Reed–Frost epidemic (*see* **Chain Binomial Model**) and the general epidemic. A Markovian modeling has the advantage to make possible a study of the temporal evolution of the epidemic process. A severe drawback, however, is the associated assumption that the infectious periods are **exponentially distributed**, which is unrealistic for many diseases.

Next, we will turn to SIR models, no longer necessarily Markovian, that precisely allow an arbitrary distribution for the infectious periods. The basic model, named the generalized epidemic, is the direct corresponding extension of the general epidemic. It is a particular case of the so-called collective epidemic in which infectives contact susceptibles by sampling of random size. Studying the temporal behavior of these models is quite complex (or even irrelevant), and the main purpose is to analyze the ultimate epidemic state.

Finally, we will discuss multitype versions of these models that allow us to incorporate heterogeneities in susceptibility, infectivity, and/or mixing behaviors. Such factors play an important role in the mechanism of spread of infection and for the evaluation of control policies (*see* **Epidemic Models, Control**).

The list of references is very partial (by necessity). Additional references can be found in the treatises mentioned above (and in [21] for a review of work prior to 1990).

Markovian Epidemic Models

The Reed-Frost Epidemic

The model assumes that the periods of time between the receipt of infection and the onset of infectiousness (the latent periods) are of fixed length and the subsequent infectious periods are contracted to a single point. Each infective is able to contact any given susceptible with the probability $p = 1 - q$, all these events being independent. A discrete timescale $t = 0, 1, 2, \ldots$ is then used, which corresponds to the successive generations of infections (separated by the latent periods). In other words, if at time t there are $I_t = i$ infectives, any of the $S_t = s$ susceptibles will remain susceptible at $t + 1$ with the probability q^i. Thus, the epidemic process $\{(S_t, I_t), t = 0, 1, \ldots\}$ is a Markov chain and the transition law is of **binomial** form:

$$P(S_{t+1} = s - j, I_{t+1} = j \mid S_t = s, I_t = i) = \binom{s}{j} q^{i(s-j)}(1 - q^i)^j, \quad j = 0, \ldots, s. \quad (1)$$

This chain-binomial mechanism is very convenient for computations with small populations. For

example, with initially two susceptibles and two infectives, four different paths lead to the end of infection: $(2,2) \to (2,0)$ with probability q^4, $(2,2) \to (1,1) \to (1,0)$ with probability $2q^3(1-q^2)$, $(2,2) \to (1,1) \to (0,1) \to (0,0)$ with probability $2q^2(1-q^2)p$, and $(2,2) \to (0,2) \to (0,0)$ with probability $(1-q^2)^2$. In particular, the distribution of S_T is given by $P(S_T = 2) = q^4$, $P(S_T = 1) = 2q^3(1-q^2)$ and the remainder for $P(S_T = 0)$; the law of T also follows.

Enumerating all the transient paths becomes cumbersome with larger populations. The distribution of S_T, however, can be determined recursively. Putting $x_{[k]} = x(x-1)\dots(x-k+1)$ for any naturals x, k, we have the n relations:

$$\mathrm{E}\{S_{T,[k]}q^{kS_T}\} = n_{[k]}q^{k(n+m)}, \quad k = 1, \dots, n. \tag{2}$$

This constitutes a system of n linear equations in the n unknown probabilities $P(S_T = s)$, $s = 1, \dots, n$. It is solved recursively for $k = n, \dots, 1$; then $P(S_T = 0)$ follows [23]. Somewhat surprisingly, the relation (2) can be extended to a variety of SIR models (see below).

The General Epidemic

The model neglects the effects of latency and assumes that the infectious periods are independent and exponentially distributed with parameter μ. When infected, an individual makes contacts with any given susceptible at the time points of a **Poisson process** with rate β, all these events being independent. Thus, the epidemic process $\{(S_t, I_t), t \geq 0\}$ is a Markov process and the infinitesimal transition probabilities are

$$\begin{aligned} &P(S_{t+\mathrm{d}t} = s-1, I_{t+\mathrm{d}t} = i+1 \mid S_t = s, I_t = i) \\ &\qquad = \beta s i \,\mathrm{d}t + o(\mathrm{d}t), \end{aligned} \tag{3}$$

$$\begin{aligned} &P(S_{t+\mathrm{d}t} = s, I_{t+\mathrm{d}t} = i-1 \mid S_t = s, I_t = i) \\ &\qquad = \mu i \,\mathrm{d}t + o(\mathrm{d}t). \end{aligned} \tag{4}$$

The state probabilities $p_{s,i}(t) = P(S_t = s, I_t = i)$ satisfy the forward Kolmogorov differential equations (*see* **Stochastic Processes**):

$$\begin{aligned} \frac{\mathrm{d}p_{s,i}(t)}{\mathrm{d}t} &= \beta(s+1)(i-1)p_{s+1,i-1}(t) \\ &\quad + \mu(i+1)p_{s,i+1}(t) - (\beta s + \mu)i p_{s,i}(t), \end{aligned} \tag{5}$$

for $s = 0, \dots, n$ and $i = 0, \dots, n+m-s$, and with $p_{s,i}(t) = 0$ outside this range; initially, $p_{n,m}(0) = 1$. A Laplace transform solution is given in [19], and a simpler method by recursion is developed in [14, 41]. In [38], the algebraic structure of the solution is exhibited, which allows us to highlight and improve the recursive technique.

For the end of the epidemic, the distribution of S_T is provided by a system of n linear equations similar to (2):

$$\mathrm{E}\{S_{T,[k]}q_k^{S_T}\} = n_{[k]}q_k^{n+m}, \quad k = 1, \dots, n, \tag{6}$$

where $q_k = \mu/(\mu + \beta k)$ (in place of q^k).

Varying Susceptibilities

The hypothesis of a common infection rate for all pairs of susceptible and infective is a simplification. To account for differences between susceptibles, the general epidemic is extended by splitting up the susceptible class into h homogeneous groups, labelled $l = 1, \dots, h$, with initial sizes n_l. The infectives form a single class and act independently as before. Within the susceptible group l, $l = 1, \dots, h$, each susceptible can be contacted by any given infective at the rate β_l; let $\beta^S = (\beta_1, \dots, \beta_1, \dots, \beta_h, \dots, \beta_h)$ be the row vector of the susceptibility rates for the $n = n_1 + \cdots + n_h$ susceptibles.

It seems to be intuitive that more heterogeneous susceptibilities decrease the damage caused to the whole susceptible class. Indeed, this can be established using the concepts of majorization (denoted by $\prec$) between real vectors of n elements with equal sum [29] and of usual stochastic order (denoted by $\leq_{st}$) between random variables [42]. Consider an identical epidemic model but built with another vector β'^S of susceptibility rates; let S'_t be the associated total number of susceptibles at time t. Then, it can be proved [5, 22] that $\beta^S \prec \beta'^S$ (a more diverse vector of susceptibilities) implies $n - S_t \geq_{st} n - S'_t$ (a smaller total infection), for all t and at the end of the epidemic. In particular, the worst situation arises when the susceptibles form a homogeneous group with rate $\overline{\beta}^S = (n_1\beta_1 + \cdots + n_h\beta_h)/n$.

Some Other Epidemics

Various adaptations or generalizations of the previous SIR models have been proposed to account for specificities in the spread of certain infectious diseases. We present two of them in continuous-time.

With *fatal epidemics*, the removal of an infective is inevitably by death. So, at time t, the surviving population being of size $S_t + I_t$, the contact rate per pair of susceptible and infective is no longer constant but equal to $\beta/(S_t + I_t)$. This means that the infection probability (3) is modified as

$$P(S_{t+\mathrm{d}t} = s - 1, I_{t+\mathrm{d}t} = i + 1 \mid S_t = s, I_t = i) = \frac{\beta si}{s + i}\,\mathrm{d}t + o(\mathrm{d}t). \tag{7}$$

Transient and final behaviors of the epidemic are investigated in [9, 37]. In particular, a system of n linear equations still exists for the distribution of S_T:

$$\mathrm{E}\left\{ S_{T,[k]} \prod_{j=1}^{S_T - k} q_{k,j} \right\} = n_{[k]} \prod_{j=1}^{n+m-k} q_{k,j}, \quad k = 1, \ldots, n, \tag{8}$$

where $q_{k,j} = \mu(k + j)/[\mu(k + j) + \beta k]$.

With *carrier epidemics*, infectives are immediately detected and removed, and the disease is propagated by carriers that do not display any symptom. If contacted, a susceptible either becomes a carrier with probability π, or is recognized as an infective, thus removed, with probability $1 - \pi$. Denoting the number of carriers at t by I_t, the infection probability (3) is replaced by

$$P(S_{t+\mathrm{d}t} = s - 1, I_{t+\mathrm{d}t} = i + 1 \mid S_t = s, I_t = i) = \beta\pi si\,\mathrm{d}t + o(\mathrm{d}t), \tag{9}$$

$$P(S_{t+\mathrm{d}t} = s - 1, I_{t+\mathrm{d}t} = i \mid S_t = s, I_t = i) = \beta(1 - \pi)si\,\mathrm{d}t + o(\mathrm{d}t). \tag{10}$$

The model with $\pi = 1$ is equivalent to the general epidemic; $\pi = 0$ is a separate case, simpler because there is no transition from susceptible to carrier. An analysis of the epidemic is carried out in [35, 36]. For the distribution of S_T, the system (6) is changed as:

$$\mathrm{E}\{S_{T,[k]} q_k'^{\,S_T}\} = n_{[k]} q_k'^{\,n} q_k^m, \quad k = 1, \ldots, n. \tag{11}$$

where $q_k = \mu/(\mu + \beta k)$ (as before), and $q_k' = \pi q_k + 1 - \pi$.

Notice that in general, many Markovian epidemics can be viewed as **compartmental models** of *right-shift* type in which the population is subdivided into a finite number of cells (here S, I, R) and transitions are shiftings of one unit from a cell to some other to its right [20, 38].

We also mention that demographic forces can generate recurrent epidemic outbreaks for diseases that confer immunity. A general epidemic model accounting for *demography*, examined in [34], assumes that new susceptibles arrive at a constant rate θn and all individuals die at a rate θ (thus, the population size will fluctuate around n). The study is concerned with the time to extinction as a measure of the persistence of infection; it relies on the concept of quasi-stationary distribution (*see* **Stationarity**).

Collective Epidemic Model

A Common Structure

As shown in [28] and subsequent works, the Markovian approach can be relaxed to some extent when only the final states S_T or Z_T are under investigation. A flexible model in this context is the collective epidemic presented in [25, 36] (its appellation underlines the focus on the final outcome).

The model assumes that the infectives act independently and their infectious periods are independent and identically distributed with an arbitrary distribution. The fates of the susceptibles in front of the risk of infection are similar and interdependent (probabilistically interchangeable). Specifically, each infective fails to contact anyone in any given set of k susceptibles, $k = 1, \ldots, n$, with a probability q_k which is a function of the set size k (and not of the set itself).

These q_k's are the parameters of the model. They can be expressed under the form

$$q_k = \mathrm{E}\left[\binom{n-k}{R} \Big/ \binom{n}{R} \right], \quad k = 1, \ldots, n, \tag{12}$$

for some random variable R valued in $\{0, \ldots, n\}$. An interpretation for (12) is that any infective, j say, contacts susceptibles by drawing a sample of random size R_j without replacement among the n initial susceptibles; all the random variables R_j are independent and distributed as R (see also [30] for this formulation).

A standard special model is the *generalized epidemic*. It assumes, as in the general epidemic, that infectious contacts are ruled by independent Poisson

processes with rate β (i.e. (3) is kept). The novelty is that the infectious periods D_j are independent but with any fixed distribution, that of a random variable D say. Thus, for this model, $q_k = \mathrm{E}[\exp(-\beta k D)]$. In particular, for the general epidemic, D is exponentially distributed with parameter μ, yielding $q_k = \mu/(\mu + \beta k)$ as indicated in (6).

The Reed–Frost epidemic too is a collective model with $q_k = q^k$. This is not true for the fatal epidemic (since the infection rate depends on the surviving population size). The carrier epidemic is an extension of the collective model with two possible types of infection: the carriers, who are removed after a period distributed as D, and the detected infectives, who are directly removed; thus, for the initial carriers, the probabilities of nontransmission are $q_k = \mathrm{E}[\exp(-\beta k D)] = \mu/(\mu + \beta k)$, while for each new infection, the probabilities of nontransmission are $q_k' = \mathrm{E}[\pi \exp(-\beta k D) + 1 - \pi] = \pi q_k + 1 - \pi$.

The Final State

Different representations of this model can be built, which allow the study of its final state S_T. So, let us label the infectives according to their order of removal in the epidemic, and denote by $\hat{S}_j$, $j = 1, 2, \ldots$, the number of susceptibles that escape contact with the first j infectives; put $\hat{S}_0 = n$. It is easily seen that the process $\{\hat{S}_j, j = 0, 1, \ldots\}$ is a Markov chain that has the same final state as the collective epidemic. One can then show that the exact distribution of S_T is provided by the system (6) with q_k defined as above [25, 36].

When considering large populations, however, there is a need for approximation methods. Let $n \to \infty$, and write $R = R_n$ depending on n; for clarity, m is fixed here. Roughly, three different asymptotic behaviors can arise.

(a) A minor infection (S_T is near n). Suppose that $R_n \to R^*$ in distribution with $\mathrm{E}(R^*) \leq 1$. Then, the final size Z_T **convergence in distribution** to Y_∞, where $P(Y_\infty < \infty) = 1$ and Y_∞ is the total progeny in a **branching process** having m ancestors and with all offspring sizes independent and distributed as R^*. The **generating function** of Y_∞ is given by $[\phi(z)]^m$, where $\phi(z)$ satisfies the equation $\phi(z) = f[z\phi(z)]$, $f(z)$ being the generating function of R^*.

(b) A possible major infection (with S_T around a positive fraction of n). Suppose that $R_n \to R^*$ in distribution with $\mathrm{E}(R^*) > 1$. Then, Z_T still converges to Y_∞ but now $P(Y_\infty < \infty) = \rho^m < 1$ where ρ^m is the extinction probability of the branching process (ρ is the unique root in $(0, 1)$ of the equation $\rho = f(\rho)$). With probability $1 - \rho^m$, a true epidemic occurs that infects infinitely many susceptibles, that is, $S_T/n \to \sigma < 1$ in probability (σ is the unique root in $(0, 1)$ of the equation $\sigma = \exp[-\mathrm{E}(R^*)(1 - \sigma)]$); moreover, $\sqrt{n}(S_T - n\sigma)$ has a normal limit distribution with mean 0 and variance $\sigma(1 - \sigma)\{1 + [\mathrm{var}(R^*) - \mathrm{E}(R^*)]\sigma\}/[1 - \mathrm{E}(R^*)\sigma]^2$.

(c) A drastic infection (S_T is bounded). Suppose that $R_n \to \infty$ in probability. Then, convergence in distribution of S_T to a law nondegenerate in 0 can occur under some conditions on the asymptotic behavior of R_n, and the limit distribution is necessarily a **Poisson** law with random mean. It reduces to a Poisson law with fixed mean b if and only if $n[1 - \mathrm{E}(R_n)/n]^n \to b$ and $\mathrm{var}(R_n)/n \to 0$.

The qualitative difference between the behaviors (a) and (b) depicts a threshold phenomenon (*see* **Epidemic Thresholds**) in which the threshold parameter is $\mathrm{E}(R^*)$ and the critical value is equal to 1. The parameter $\mathrm{E}(R^*)$ is named the basic **reproduction number**; it is usually denoted by R_0 and is interpreted as the expected number of infectious contacts which one infective would make in a large completely susceptible population [17]. Considerable work has been devoted to these questions [4, 8, 27, 30, 39, 44]. The behavior (c) is peculiar to highly infectious diseases; it is studied in [6, 24, 26].

As a special case, let us examine the generalized epidemic, where $\beta = \beta_n$ is function of n (D being fixed). Here, R_n in (12) has a binomial law with n trials and random success probability $1 - \exp(-\beta_n D)$. For (a) and (b), a standard situation is when $\beta_n = \beta/n$ (each infective meets, on the average, a limited number of susceptibles); then, R^* has a Poisson law with random mean βD, with $R_0 \equiv \mathrm{E}(R^*) = \beta \mathrm{E}(D)$. For (c), convergence can occur under some conditions on the tail distribution of D and for a sequence β_n that necessarily satisfies $n\{1 - \mathrm{E}[\exp(-\beta_n D)]\} - ln(n/b_n) \to 0$ with $b_n \to b$; then, the limit distribution is either a Poisson law with fixed parameter b or a Poisson law with a random mean based on an asymmetric **Cauchy** stable law.

We point out that the alternative situation where $m_n \to \infty$ as $n \to \infty$ can also lead to different asymptotic regimes. Results are given in [26, 27, 30]; an extensive analysis for the generalized epidemic is carried out in [43].

Varying Infectivities

It is intuitively clear that a lower infectivity power decreases the total damage caused by the epidemic. To make this precise, we use the concept of s-increasing convex order between nonnegative random variables (denoted by $\leq_s$ in the arithmetic case and by $\leq_{s*}$ in the continuous case). When $s = 1$, it corresponds to the usual stochastic order and when $s = 2$, to the classical increasing convex order [25, 42]. Consider a new collective model with parameters q'_k given by (12) with some random variable R' substituted for R; let $S'_{T'}$ be the associated final susceptible size. Then, it can be proved [25] that $n - R \leq_s n - R'$ (a smaller number of contacts per infective) implies $S_T \leq_s S'_{T'}$ (a larger final susceptible state).

For the generalized epidemic, we get that $\exp(-\beta D) \leq_{s*} \exp(-\beta D')$ (a shorter infectious period) yields $S_T \leq_s S'_{T'}$ (a better protection for the susceptibles). This result is of practical interest when the law of D is not known except a few moments. For instance, suppose that $E(D) = d_1$ and $\mathrm{var}(D) = d_2$ are given. One shows that $\exp(-\beta D'') \leq_{2*} \exp(-\beta D) \leq_{3*} \exp(-\beta D')$ where D'' is a random variable degenerate in d_1, while D' is a two-points random variable of values $(d_1^2 + d_2)/d_1$ with probability $d_1^2/(d_1^2 + d_2)$ and 0 otherwise. Thus, bounds for S_T are $S''_{T''} \leq_2 S_T \leq_3 S'_{T'}$. Note that replacing D by its mean d_1 leads to underestimate, in the increasing convex sense, the ultimate number of susceptibles.

Multitype Epidemic Models

Individual Heterogeneities

Variability in individual susceptibilities and/or infectivities is a first kind of potential heterogeneity. This factor is easily incorporated in the previous models. For that, we subdivide the whole population into h homogeneous groups of individuals, labeled $l = 1, \ldots, h$, with initially n_l susceptibles and m_l infectives. Going back to Markovian epidemics, let us consider the *general multitype epidemic*. In group l, each infective is removed at the rate μ_l, and while infected, it contacts any given susceptible within group l' at the rate $\beta_{l,l'}$. In particular, the case called *proportionate mixing* is when $\beta_{l,l'}$ is of product form $\beta_l \gamma_{l'}$. The *collective multitype epidemic* is constructed in a similar way. In group l, each infective fails to transmit infection within any given set of k_1 susceptibles in group $1, \ldots, k_h$ susceptibles in group h, with a probability $q^{(l)}_{k_1,\ldots,k_h}$, where $k_1 = 0, \ldots, n_1, \ldots, k_h = 0, \ldots, n_h$ with $k_1 + \cdots + k_h \geq 1$. In particular, for the *generalized multitype epidemic*, then $q^{(l)}_{k_1,\ldots,k_h} = \mathrm{E}[\exp(-\beta_{l,1}k_1 + \cdots + \beta_{l,h}k_h)D^{(l)}]$ where $D^{(l)}$ is the length of an infectious period in group l.

The study of these multitype models is, roughly, similar to the homogeneous case. Let us concentrate on the ultimate numbers of susceptibles $S_T^{(l)}$, $l = 1, \ldots, h$, that escape the disease in the different groups. For their exact joint distribution, we have the relations:

$$\mathrm{E}\left\{\prod_{l=1}^{h} S^{(l)}_{T,[k_l]}[q^{(l)}_{k_1,\ldots,k_h}]^{S_T^{(l)}}\right\} = \prod_{l=1}^{h} n_{l,[k_l]}[q^{(l)}_{k_1,\ldots,k_h}]^{n_l+m_l}, \tag{13}$$

where $k_1 = 0, \ldots, n_1, \ldots, k_h = 0, \ldots, n_h$ with $k_1 + \cdots + k_h \geq 1$. This is a system of $(n_1 + 1)\ldots(n_h + 1) - 1$ linear equations in the final state probabilities $P[S_T^{(1)} = s_1, \ldots, S_T^{(h)} = s_h]$, for $s_1 = 0, \ldots, n_1, \ldots, s_h = 0, \ldots, n_h$, with $s_1 + \cdots + s_h \geq 1$; $P[S_T^{(1)} = \cdots = S_T^{(h)} = 0]$ follows [36].

Large population limits still hold true. So, putting $n = n_1 + \cdots + n_h$, consider the generalized epidemic where the contact rates are of normalized form $\beta_{l,l'}/n$ and each group size is large with $n_l/n \to \nu_l > 0$; $D^{(l)}$ is fixed, and m_l too say. Then, the final sizes vector $(Z_T^{(1)}, \ldots, Z_T^{(h)})$ converges in distribution to the total progeny vector $(Y_\infty^{(1)}, \ldots, Y_\infty^{(h)})$ in a multitype branching process having $(m_1, \ldots, m_l)$ ancestors, with independent lifetimes distributed as $(D^{(1)}, \ldots, D^{(h)})$ and **matrix** of birth rates $\{\beta_{l,l'}\nu_{l'}\}$, $1 \leq l, l' \leq h$. Thus, there exists a threshold parameter (basic reproduction number) R_0 which is the largest **eigenvalue** of the matrix of mean offspring $\{\mathrm{E}(D^{(l)})\beta_{l,l'}\nu_{l'}\}$, $1 \leq l, l' \leq h$. If $R_0 \leq 1$, the total progeny of the branching is finite with probability one: the epidemic is minor. If $R_0 > 1$, the total progeny can explode with a strictly positive probability; on this part, the epidemic is major and the final sizes vector has asymptotically a **multivariate normal distribution** [7]; see also [40].

Structural Heterogeneities

Nonuniform mixing caused by the social or geographical structure of the population is another kind of

potential heterogeneity. A simple case with two levels of mixing, local and global, is motivated by *epidemics among households* in which the infection power of an infective is much higher within its own household than outside. Extending in this sense the above generalized epidemic, one obtains again a threshold behavior for the asymptotic situation in which the number of households tends to infinity [10]. Such a model is useful especially when evaluating potential strategies for the control of disease transmission [13] (*see* **Epidemic Models, Control**).

In *spatial epidemics*, individuals are located in different sites, and infectives contact others according to a given spatial distribution (*see* **Epidemic Models, Spatial**). Important aspects are the threshold phenomenon and the velocity of the spread of the disease. Results can be obtained for some spatial extensions of the generalized epidemic [15, 33].

References

[1] Andersson, H. & Britton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*. Lecture Notes in Statistics 151, Springer-Verlag, New York.

[2] Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.

[3] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

[4] Ball, F.G. (1983). The threshold behaviour of epidemic models, *Journal of Applied Probability* **20**, 227–241.

[5] Ball, F.G. (1985). Deterministic and stochastic epidemics with several kinds of susceptibles, *Advances in Applied Probability* **17**, 1–22.

[6] Ball, F.G. & Barbour, A.D. (1990). Poisson approximation for some epidemic models, *Journal of Applied Probability* **27**, 479–490.

[7] Ball, F.G. & Clancy, D. (1993). The final size and severity of a generalised stochastic multitype epidemic model, *Advances in Applied Probability* **25**, 721–736.

[8] Ball, F.G. & Donnelly, P.J. (1995). Strong approximations for epidemic models, *Stochastic Processes and their Applications* **55**, 1–21.

[9] Ball, F.G. & O'Neill, P.D. (1993). A modification of the general stochastic epidemic motivated by AIDS modelling, *Advances in Applied Probability* **25**, 39–62.

[10] Ball, F.G., Mollison, D. & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing, *Annals of Applied Probability* **7**, 46–89.

[11] Bartlett, M.S. (1960). *Stochastic Population Models in Ecology and Epidemiology*. Methuen, London.

[12] Becker, N.G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

[13] Becker, N.G. & Utev, S. (1998). The effect of community structure on the immunity coverage required to prevent epidemics, *Mathematical Biosciences* **147**, 23–39.

[14] Billard, L. (1973). Factorial moments and probabilities for the general stochastic epidemic, *Journal of Applied Probability* **10**, 277–288.

[15] Cox, J.T. & Durrett, R. (1988). Limit theorems for the spread of epidemics and forest fires, *Stochastic Processes and their Applications* **30**, 1171–1191.

[16] Daley, D.J. & Gani, J. (1999). *Epidemic Modelling: an Introduction*. Cambridge University Press, Cambridge.

[17] Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases, *Statistical Methods in Medical Research* **2**, 23–41.

[18] Diekmann, O. & Heesterbeek, J.A.P. (2000). *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. John Wiley, New York.

[19] Gani, J. (1967). On the general stochastic epidemic, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, University of California Press, Berkeley, pp. 271–279.

[20] Kryscio, R.J. & Severo, N.C. (1975). Computational and estimation procedures in multidimensional right-shift processes and some applications, *Advances in Applied Probability* **7**, 349–382.

[21] Lefèvre, Cl. (1990). Stochastic epidemic models for SIR infectious diseases: a brief survey of the recent general theory, in *Stochastic Processes in Epidemic Theory*, J.-P. Gabriel, Cl. Lefèvre & Ph. Picard, eds. Lecture Notes in Biomathematics 86, Springer-Verlag, New York, pp. 1–12.

[22] Lefèvre, Cl. (1994). Stochastic ordering of epidemics, in *Stochastic Orders and their Applications*, M. Shaked & J.G. Shanthikumar, eds. Academic Press, San Diego, pp. 323–348.

[23] Lefèvre, Cl. & Picard, Ph. (1990). A non-standard family of polynomials and the final size distribution of Reed-Frost epidemic processes, *Advances in Applied Probability* **22**, 25–48.

[24] Lefèvre, Cl. & Utev, S. (1995). Poisson approximation for the final state of a generalized epidemic process, *Annals of Probability* **23**, 1139–1162.

[25] Lefèvre, Cl. & Utev, S. (1996). Comparing sums of exchangeable Bernoulli random variables, *Journal of Applied Probability* **33**, 285–310.

[26] Lefèvre, Cl. & Utev, S. (1997). Mixed Poisson approximation in the collective epidemic model, *Stochastic Processes and their Applications* **69**, 217–246.

[27] Lefèvre, Cl. & Utev, S. (1999). Branching approximation for the collective epidemic model, *Methodology and Computing in Applied Probability* **1**, 211–228.

[28] Ludwig, D. (1975). Final size distributions for epidemics, *Mathematical Biosciences* **23**, 33–46.

[29] Marshall, A.W. & Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications*. Academic Press, New York.

[30] Martin-Löf, A. (1986). Symmetric sampling procedures, general epidemic processes and their threshold limit theorems, *Journal of Applied Probability* **23**, 265–282.
[31] Mode, C.J. & Sleeman, C.K. (2000). *Stochastic Processes in Epidemiology*. World Scientific, Singapore.
[32] Mollison, D. ed. (1995). *Epidemic Models: Their Structure and Relation to Data*. Cambridge University Press, Cambridge.
[33] Mollison, D. (1977). Spatial contact models for ecological and epidemic spread, *Journal of the Royal Statistical Society, Series B* **39**, 283–326.
[34] Nåsell, I. (1999). On the time to extinction in recurrent epidemics, *Journal of the Royal Statistical Society, Series B* **61**, 309–330.
[35] Picard, Ph. (1980). Applications of martingale theory to some epidemic models, *Journal of Applied Probability* **17**, 583–599.
[36] Picard, Ph. & Lefèvre, Cl. (1990). A unified analysis of the final size and severity distribution in collective Reed-Frost epidemic processes, *Advances in Applied Probability* **22**, 269–294.
[37] Picard, Ph. & Lefèvre, Cl. (1993). Distribution of the final state and severity of epidemics with fatal risk, *Stochastic Processes and their Applications* **48**, 277–294.
[38] Picard, Ph. & Lefèvre, Cl. (1999). On the algebraic structure in Markovian processes of death and epidemic types, *Advances in Applied Probability* **31**, 742–757.
[39] Scalia-Tomba, G. (1985). Asymptotic final size distribution for some chain-binomial processes, *Advances in Applied Probability* **17**, 477–495.
[40] Scalia-Tomba, G. (1990). On the asymptotic final size distribution of epidemics in heterogeneous populations, in *Stochastic Processes in Epidemic Theory*, J.-P. Gabriel, Cl. Lefèvre & Ph. Picard, Lecture Notes in Biomathematics 86, Springer-Verlag, New York, pp. 189–196.
[41] Severo, N.C. (1969). The probabilities of some epidemic models, *Biometrika* **56**, 197–201.
[42] Shaked, M. & Shanthikumar, J.G. (1994). *Stochastic Orders and their Applications*. Academic Press, San Diego.
[43] Startsev, A.N. (2001). Asymptotic analysis of the general stochastic epidemic with variable infectious periods, *Journal of Applied Probability* **38**, 18–35.
[44] Von Bahr, B. & Martin-Löf, A. (1980). Threshold limit theorems for some epidemic processes, *Advances in Applied Probability* **12**, 319–349.

(*See also* **Epidemic Models, Inference**; **Mathematical Biology, Overview**)

CLAUDE LEFÈVRE

Skewness

A statistical distribution which is not symmetric about some point is said to be skew. The extent of skewness may be measured in a variety of ways. For example, the mean, median, and mode generally do not coincide for a skew distribution, and so their relationships may be used to quantify skewness. Two candidates are:

$$\text{skew}_1 = \frac{\text{mean} - \text{mode}}{\text{standard deviation}},$$
$$\text{skew}_2 = \frac{\text{mean} - \text{median}}{\text{standard deviation}}.$$

Both were considered by Pearson [6], who noted the empirical fact that for many distributions, $\text{skew}_1 \approx 3\text{skew}_2$ (Pearson's law of skewness). Pearson gave an explanation in terms of his family of type III distributions (*see* **Gamma Distribution**). The definition skew_1 is unattractive in the case of multimodal distributions, and since the main motivation for skew_2 is via Pearson's law, which also fails in multimodal cases, neither skew_1 nor skew_2 has general application (*see* **Unimodality**).

Today, skewness is usually defined in terms of the third cumulant of a distribution (*see* **Characteristic Function**), rather than through relationships among its mean, median, and mode. Nevertheless, the two approaches are linked; see, for example, Haldane [1] and Hall [2], who provided an explanation for Pearson's law in terms of the third cumulant and the **central limit theorem**.

Given a random variable X with mean $\mu = \text{E}(X)$, let σ^2 denote its variance, let $\mu_3 = \text{E}(X - \mu)^3$ be its third cumulant, and define

$$\beta_1 = \frac{\mu_3^2}{\sigma^6}, \qquad \text{skew}_3 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma_3}.$$

The latter is sometimes referred to simply as "the skewness of the distribution of X". Its multivariate analog, for a d-vector X, is the set of all $\frac{1}{6}d(d^2 + 3d + 2)$ third **moments** of components of $\mathbf{Y} = \Sigma^{-1/2}\mathbf{X}$, where $\mathbf{\Sigma}$ denotes the **covariance matrix** of $\mathbf{X}$. (There are d terms of the form $\text{E}(Y^{(i)^3})$, $d(d-1)$ terms like $\text{E}(Y^{(i)}Y^{(j)^2})$, and $\frac{1}{6}d(d-1)(d-2)$ terms like $\text{E}(Y^{(i)}Y^{(j)}Y^{(k)})$.)

If the distribution of X has been standardized for location and scale (i.e. $\mu = 0$ and $\sigma = 1$), then the first terms in Edgeworth and Gram–Charlier expansions of the distribution or density function of X are proportional to $\sqrt{\beta_1}$; see, for example, [3, pp. 28, 30]. This indicates the central role which $\sqrt{\beta_1}$ plays in determining properties of "regular" distributions, relative to other measures of skewness.

Van Zwet [7] proposed a partial ordering of distributions in terms of skewness, and suggested defining a distribution with distribution function F to be *skewed to the right* if $F^{-1}\{1 - F(x)\}$ is convex in x. Oja [5] noted that some skewness measures (e.g. skew_3) preserve van Zwet's ordering, while others (e.g. skew_1) do not.

Descriptions of skewness in terms of moments are not always well defined. More robust definitions include that attributed to F. Galton, whose statistical work stressed the importance of quartiles (*see* **Quantiles**):

$$\text{skew}_4 = \frac{\text{lower quartile} + \text{upper quartile} - 2 \times \text{median}}{\text{upper quartile} - \text{lower quartile}}.$$

Each of $\text{skew}_1, \ldots, \text{skew}_4$ is estimable, although the difficulty of estimating the mode makes skew_1 unattractive for data analysis. The sample skewness, for data $X_1, \ldots, X_n$, is often defined as

$$\sqrt{b_1} = \frac{n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left\{n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\}^{3/2}},$$

and is an estimate of $\sqrt{\beta_1}$. In normal samples, $\sqrt{b_1}$ is approximately normally distributed with zero mean and variance $6/n$. More extensive properties of its moments are addressed in [4, pp. 316–318].

References

[1] Haldane, J.B.S. (1942). The mode and median of a nearly normal distribution with given cumulants, *Biometrika* **32**, 294–299.

[2] Hall, P. (1980). On the limiting behaviour of the mode and median of a sum of independent random variables, *Annals of Probability* **8**, 419–430.

[3] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1. Wiley, New York.

[4] Kendall, M. & Stuart, A. (1977). *The Advanced Theory of Statistics*, Vol. 1. Griffin, London.

[5] Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions, *Scandinavian Journal of Statistics* **8**, 154–168.

[6] Pearson, K. (1895). Contributions to the mathematical theory of evolution, II. Skew variation in homogeneous material, *Philosophical Transactions* **186**, 343–414.

[7] Van Zwet, W.R. (1964). *Convex Transformations of Random Variables.* Mathematical Centre Tracts No. 7, Mathematisch Centrum, Amsterdam.

PETER HALL

Slope–Ratio Assay

Slope-ratio bioassays are analytic dilution assays (*see* **Biological Assay, Overview**) that arise mainly from microbiologic applications. The subject is usually an inoculum of specified amount of a bacterial culture. The response is typically either a measure of the bacterial growth during a fixed time interval or the amount of a base (alkali) needed to neutralize the acid that is formed during growth of the bacteria following application of fixed doses (d_i) of test and standard preparations.

The statistical model relating fixed doses of the standard and test preparations to the response leads to straight lines that intersect at zero dose. Generally, the expected response of the standard is assumed to be linear in some known power of dose. The regression line for the standard is then

$$\mathrm{E}[y_\mathrm{S}|d_\mathrm{S}] = \alpha + \beta_\mathrm{S} x,$$

where $x = d_\mathrm{S}^\lambda$ is the dose metameter. In practice, setting the power parameter (λ) equal to one often provides an adequate approximation [2]. The dosage of the test preparation (d_T) that provides an equivalent response to a specified dose of the standard is $d_\mathrm{S} = \rho d_\mathrm{T}$, where ρ designates the relative potency. The regression line for the test preparation is thus

$$\begin{aligned}\mathrm{E}(y_\mathrm{T}|d_\mathrm{T}) &= \mathrm{E}[y_\mathrm{S}|\rho d_\mathrm{T}] = \alpha + \beta_\mathrm{S}(\rho d_\mathrm{T})^\lambda \\ &= \alpha + \beta_\mathrm{S}\rho^\lambda d_\mathrm{T}^\lambda.\end{aligned}$$

When λ equals one and $x = d_\mathrm{T}$, then $\beta_\mathrm{T} = \beta_\mathrm{S}\rho$ (Figure 1).

The fundamental assumption for validity (condition of similarity) in a slope-ratio assay implies that the standard and test preparations intersect at zero dose. Therefore, the relationship can alternatively be expressed as a multiple regression equation, such that

$$\mathrm{E}(y) = \alpha + \beta_\mathrm{S} x_\mathrm{S} + \beta_\mathrm{T} x_\mathrm{T}.$$

The solution for the relative potency is the ratio of the two slopes from the multiple regression:

$$\rho = \left(\frac{\beta_\mathrm{T}}{\beta_\mathrm{S}}\right)^{1/\lambda}.$$

When $\lambda = 1$, the relative potency is

$$\rho = \frac{\beta_\mathrm{T}}{\beta_\mathrm{S}}.$$

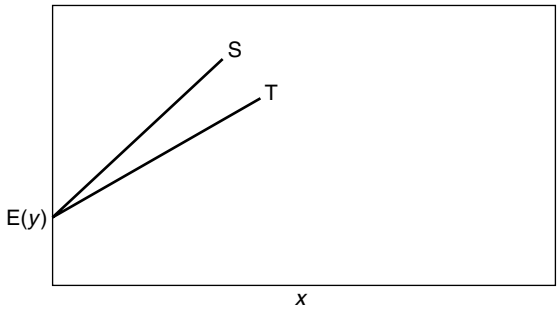

Figure 1 Expected response (E(y)) vs. dose metameter (x) for standard (S) and test (T) preparations in a slope-ratio assay

Relative Potency Estimation and Validity Tests

Estimates of the slopes for the standard and test preparations are readily calculated using conventional **multiple linear regression** techniques. To evaluate fully the model's assumptions, measurements of the response at zero dose (blanks) need to be incorporated in the experiment. For computational purposes, it is convenient to regard the doses as N triplets $\{x_0, x_\mathrm{S}, x_\mathrm{T}\}$ with associated response y, where $N = N_0 + N_\mathrm{S} + N_\mathrm{T}$ (total observations across blanks, standard, and test preparations). Here, dose levels are coded as "0" in the triplet for other than the preparation specified; in other records, (1, 0, 0) corresponds to blanks, $\{0, x_\mathrm{S}, 0\}$ to standard, and $\{0, 0, x_\mathrm{T}\}$ to test doses, respectively. A comprehensive analysis, including estimation of the relative potency, and validity tests entails computation of three regression analyses: (i) estimation of three regression coefficients ($\hat{\beta}_{03}, \hat{\beta}_{\mathrm{S}3}, \hat{\beta}_{\mathrm{T}3}$) simultaneously using all N observations; where $\hat{\beta}_{03}$ estimates the response at zero dose; (ii) estimation of two regression coefficients $\hat{\beta}_{\mathrm{S}2}$ and $\hat{\beta}_{\mathrm{T}2}$, simultaneously for the N observations including the blanks; and (iii) estimation of $\hat{\beta}_{\mathrm{S}1}$ and $\hat{\beta}_{\mathrm{T}1}$ as separate lines based, respectively, on the N_S and N_T observations of the standard and test preparations. Generally, the slopes from (ii) are utilized to estimate relative potency; that is,

$$\hat{\rho} = \frac{\hat{\beta}_\mathrm{T}}{\hat{\beta}_\mathrm{S}} = \frac{\hat{\beta}_{\mathrm{T}2}}{\hat{\beta}_{\mathrm{S}2}},$$

assuming that $\lambda = 1$. However, when validity tests indicate a significant F test for blanks (see below)

in the absence of other forms of invalidity, suggesting curvature at very low doses, it may be advisable to use $\hat{\beta}_{T3}$ and $\hat{\beta}_{S3}$ from (i) as well as their corresponding variance terms for calculating **confidence intervals** (see below).

Before proceeding to calculate the confidence intervals for ρ, verification that there is no serious evidence of invalidity of the assumptions of the design model is important. The general form of the **analysis of variance** formulas needed for slope ratio assay validation is shown in Table 1. Calculated F tests are conventionally compared to tabulated values of the ***F* distribution**, $F_{(df_1,df_2)}$, at a 5% significance level. The specific hypotheses of interest and the corresponding variance ratio statistics are as follows:

1. $H_1 : \alpha = \alpha_S = \alpha_T$; reject if
$$F = \frac{MS_I}{MS_E} > F_{(1,N-K_S-K_T-1)}.$$
Failure to reject indicates that the fundamental assumption of validity (the condition of similarity) is not seriously violated. It is analogous to the test for parallelism in a parallel-line assay.
2. $H_2 : \alpha =$ average response for blanks; reject if
$$F = \frac{MS_B}{MS_E} > F_{(1,N-K_S-K_T-1)}.$$
Rejection of this test suggests the presence of curvature at very low doses.
3. $H_3 : \beta_S = \beta_T = 0$; reject if
$$F = \frac{MS_R}{MS_E} > F_{(2,N-K_S-K_T-1)}.$$
A valid assay will have a highly significant test for regression.
4. H_4 : test of deviation of lines from linearity; reject if
$$F = \frac{MS_{NL}}{MS_E} > F_{(K_S+K_T-4,N-K_S-K_T-1)}.$$
Rejection indicates invalidity of the statistical assumption of linearity. If H_4 is rejected, there is a need to consider alternative approaches.

When no evidence of invalidity is present, confidence intervals for
$$\hat{\rho} = \frac{\hat{\beta}_T}{\hat{\beta}_S}$$
are calculated through direct application of **Fieller's theorem** for a ratio estimator [1]. In the slope-ratio assay, the $(1-\alpha) \times 100\%$ confidence intervals are

$$\hat{\rho}_L, \hat{\rho}_U = \left[\hat{\rho} - \frac{g\nu_{12}}{\nu_{11}} \pm \frac{t\hat{\sigma}}{\hat{\beta}_S}\left\{\nu_{22} - 2\hat{\rho}\nu_{12} + \hat{\rho}^2_{\nu_{22}} - g(\nu_{22} - \nu_{12}/\nu_{11})^{1/2}\right\}\right] \Big/ 1 - g,$$

where

$$g = \frac{t^2\hat{\sigma}^2\nu_{11}}{\hat{\beta}_S^2}, \qquad \hat{\sigma} = (MS_E)^{1/2}$$

and $t_{(1-\alpha/2)}$ is based on $N - K_S - K_T - 1$ df.

The terms ν_{11}, ν_{12}, and ν_{22} are the coefficients in the variance–**covariance matrix** from which the slopes are estimated. In practice, g is often small in slope-ratio assays and thus has little effect on the computations.

For more detailed presentation of the design and analysis of slope-ratio assays, Chapters 7 and 8 in Finney's text [2] and Chapter 3 in Hubert [3] are useful references.

Additional Remarks

The formulas presented above are general in form and do not depend on having a symmetrical, balanced design for the assay or equal spacing between doses. While a symmetric design with equal numbers of subjects at each dose level, including the blanks, is preferable on the basis of efficiency considerations, modern computing tools obviate the need for the simplified formulas that accompany such designs. At least three dose levels of the test and standard preparations and inclusion of blanks are needed to test validity assumptions fully. Such a design would usually be referred to as a $(3K + 1)$ design.

The methodology can easily be extended to accommodate designs in which multiple test preparations are simultaneously compared to the same standard. For designs with more than one test preparation, consideration of optimal allocation among preparations given a total number of subjects is relevant. If N_0 is prespecified, then optimal allocation of the remaining

Table 1 Slope-ratio assay analysis of variance

Source	df	Sums of squares	Mean squares	F
Total	$N-1$	$\mathrm{SS}_y = \sum_{0,\mathrm{S,T}} \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (y_{pij} - \overline{y})^2$	$\mathrm{MS}_y = \mathrm{SS}_y/(N-1)$	
Among doses	$K_\mathrm{S} + K_\mathrm{T}$	$\mathrm{SS}_\mathrm{D} = \sum_{0,\mathrm{S,T}} \sum_{i=1}^{K_p} n_{pi}(\overline{y}_{pi} - \overline{y})^2$	$\mathrm{MS}_\mathrm{D} = \mathrm{SS}_\mathrm{D}/(K_\mathrm{S} + K_\mathrm{T})$	$\mathrm{MS}_\mathrm{D}/\mathrm{MS}_\mathrm{E}$
Regression	2	$\mathrm{SS}_\mathrm{R} = S_{x_\mathrm{S} y}\hat{\beta}_{S2} + S_{x_\mathrm{T} y}\hat{\beta}_{\mathrm{T}2}$	$\mathrm{MS}_\mathrm{R} = \mathrm{SS}_\mathrm{R}/2$	$\mathrm{MS}_\mathrm{P}/\mathrm{MS}_\mathrm{E}$
Blanks (control)	1	$\mathrm{SS}_\mathrm{B} = (S_{x_0 y}\hat{\beta}_{03} + S_{x_\mathrm{S} y}\hat{\beta}_{S3} + S_{x_\mathrm{T} y}\hat{\beta}_{\mathrm{T}3}) - \mathrm{SS}_\mathrm{R}$	$\mathrm{MS}_\mathrm{B} = \mathrm{SS}_\mathrm{B}$	$\mathrm{MS}_\mathrm{B}/\mathrm{MS}_\mathrm{E}$
Intersection	1	$\mathrm{SS}_\mathrm{I} = \mathrm{SS}_\mathrm{D} - \mathrm{SS}_\mathrm{R} - \mathrm{SS}_\mathrm{B} - \mathrm{SS}_\mathrm{NL}$	$\mathrm{MS}_\mathrm{I} = \mathrm{SS}_\mathrm{I}$	$\mathrm{MS}_\mathrm{I}/\mathrm{MS}_\mathrm{E}$
Nonlinearity	$K_\mathrm{S} + K_\mathrm{T} - 4$	$\mathrm{SS}_\mathrm{NL} = \sum_{\mathrm{S,T}} \left[\sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (y_{pij} - \overline{y}_p)^2 \right] - \sum_{\mathrm{S,T}} \left(\frac{S^2_{x_p y_p}}{S_{x_p x_p}} \right)$	$\mathrm{MS}_\mathrm{NL} = \mathrm{SS}_\mathrm{NL}/(K_\mathrm{S} + K_\mathrm{T} - 4)$	$\mathrm{MS}_\mathrm{NL}/\mathrm{MS}_\mathrm{E}$
Within doses (error)	$N - K_\mathrm{S} - K_\mathrm{T} - 1$	$\mathrm{SS}_\mathrm{E} = \mathrm{SS}_y - \mathrm{SS}_\mathrm{D}$	$\mathrm{MS}_\mathrm{E} = \mathrm{SS}_\mathrm{E}/(N - K_\mathrm{S} - K_\mathrm{T} - 1) = \hat{\sigma}^2$	

Notation (adapted from [2]):

n_{pi} = number observations at dose i of preparation p,
y_{pij} = response of subject j to dose i of preparation p,
x_{pij} = dose metameter for subject j to dose i of preparation p,

where

$i = 1, 2, \ldots, K_p$, $p = 0$ (blanks), S (standard), or T (test)

and

$$K_p = \text{number dose levels of preparation } p,$$

$$N_p = \sum_{i=1}^{K_p} n_{pi}, \qquad N = \sum_{\mathrm{O,S,T}} N_p,$$

$$S_{y_p y_p} = \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (y_{pij} - \overline{y}_p)^2;$$

where

$$\overline{y}_p = \left(\sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} y_{pij} \right) \Big/ N_p,$$

$$S_{x_p x_p} = \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (x_{pij} - \overline{x}_p)^2,$$

where

$$\overline{x}_p = \left(\sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} x_{pij} \right) \Big/ N_p,$$

$$S_{x_p y_p} = \sum_{i=1}^{K_p} \sum_{j=1}^{n_{pi}} (x_{pij} - \overline{x}_p)(y_{pij} - \overline{y}_p).$$

$N_S + N_T$ subjects among r test preparations and the standard would be

$$N_S = r^{1/2} N_T,$$

assuming that the variance of $\hat{\rho}$ is approximately proportional to $1/N_S + 1/N_T$.

Relative potency estimation also assumes both homoscedasticity and normality for the distribution of the random errors in y at each dose level. Whenever sufficient data are available, appropriate tests should be conducted to assess whether these assumptions are violated. As noted in **Biological Assay, Overview** the statistical properties are suspect for the small to moderate sample sizes that characterize most bioassays. Finney [2] demonstrates that the analyses described above may provide very similar estimates of the relative potency and its confidence intervals even when the normality assumption is not fulfilled, but other model assumptions are not seriously violated. The choice of an appropriate response metameter should rely not primarily on the statistical evidence within a single assay, but should reflect information derived from evaluation of validity across a related class of independent assays.

References

[1] Fieller, E.C. (1940). The biological standardization of insulin, *Journal of the Royal Statistical Society Supplement* **7**, 1–64.

[2] Finney, D.J. (1978). *Statistical Methods in Biological Assay*, 3rd Ed. Griffin, London, pp. 148–178, 297–315.

[3] Hubert, J.J. (1984). *Bioassay*, 2nd Ed. Kendall–Hunt, Dubuque, pp. 26–39.

(*See also* **Parallel-line Assay**; **Radioimmunoassay**)

CAROL K. REDMOND

Slutzky–Yule Effect

The practice of smoothing a **time series** by forming a **moving average** of some kind is common, and is, provided there is no seasonality (*see* **Seasonal Time Series**), a very useful tool; if there is (or may be) seasonality, it remains useful, but the span and other details need more careful attention.

It is, however, not without some drawbacks, first discovered, independently, by Slutzky [2] and Yule [3].

As a very simple example, suppose that $\{X_t\}$ is a time series of completely independent observations, with a common distribution, *white noise* (*see* **Noise and White Noise**), and that $\{Y_t\}$ is the result of applying a moving average of order 3,

$$Y_t = \frac{(X_{t-1} + X_t + X_{t+1})}{3}. \tag{1}$$

Then $\{Y_t\}$ consists of correlated observations, even though $\{X_t\}$ was made up of uncorrelated observations: it is easily calculated that the autocorrelation (*see* **Autocorrelation Function**) of $\{Y_t\}$ is 2/3 at lag 1, 1/3 at lag 2, and 0 at greater lags. In other words, the process of averaging, and thereby smoothing, has introduced correlation. (This is not surprising: a series will appear smooth(er) if values nearby in time are close(r) together, which is another way of saying that they are (more) correlated.)

The same is true if $\{X_t\}$ is operated on by *any* linear filter (*see* **ARMA and ARIMA Models**): if

$$Y_t = \sum_{i=-c}^{d} g_i X_{t-i}, \tag{2}$$

say, then again $\{Y_t\}$ consists of correlated observations, even though $\{X_t\}$ was made up of uncorrelated observations.

A different perspective may be had by considering the spectral approach (*see* **Spectral Analysis**). The spectral density of $\{X_t\}$, as white noise, is constant for all frequencies. The spectral density of $\{Y_t\}$ is then given by the product of this constant and the transfer function of the linear filter (*see* **ARMA and ARIMA Models**); unless the filter is the trivial one which corresponds to multiplication by a constant the transfer function will not be constant as a function of frequency, hence neither will the spectral density of $\{Y_t\}$ be, and consequently $\{Y_t\}$ will have nonvanishing autocorrelations. In the special case of the moving average above, the spectral density of $\{Y_t\}$ will be proportional to $(1 + 2\cos\omega)^2$ at circular frequency ω.

In fact this effect can have rather more serious consequences than might appear at first sight. Suppose that the transfer function of the filter takes its maximum (modulus) at a single nonzero frequency ω_0, and suppose the filter is applied several (many) times in succession: the resulting combined filter will strongly emphasize ω_0 relative to the rest – in other words $\{Y_t\}$ will show some approximately periodic behavior; this is the essence of the Sinusoidal Limit Theorem. The above moving average cannot directly produce this effect, but suppose the moving average is used to estimate the trend, which is then subtracted from the original series to produce a "detrended" series for further analysis: the resulting filter is of the form

$$Y_t = \frac{(2X_t - X_{t-1} - X_{t+1})}{3},$$

whose transfer function is proportional to $(1 - \cos\omega)$, with a maximum at $\omega = \pi$, or equivalently at frequency (in the usual sense) 1/2 and period 2.

Thus we can produce some oscillatory behavior in a completely nonperiodic series by operating on it with a linear filter. Slutzky and Yule noted that the analysis of data (or even its collection, if in the collection process we necessarily subject it to a linear filter) can, for this reason, produce periodicities which have no real basis and thus some of those actually observed could be artifacts caused by data processing or collection.

The effect in practice, with filters that are likely to be used, will be less immediately striking – for example, the oscillations will not be regular, either in period or in amplitude – but the overall impression may nevertheless suggest the presence of an oscillatory component in the process, and a statistical analysis is then quite likely to confirm its existence. Kendall & Stuart [1] give some results about, and illustrations of, the average distance apart of peaks and of upcrossings of the axis, showing the effect of applying filters. The oscillations observed in real economic time series, and very likely in those arising in other fields of application, do have an appearance similar to those which can be generated in this way.

The detailed discussion has assumed that $\{X_t\}$ is white noise, but clearly the argument applies for any $\{X_t\}$.

References

[1] Kendall, M.G. & Stuart, A. (1976). *The Advanced Theory of Statistics*, Vol. 3 (3-volume edition). Griffin, London, Chapter 46.

[2] Slutzky, E.E. (1927). *Problems of Economic Conditions (in Russian)*, English translation in *Econometrica* **5** (1937) 105–146.

[3] Yule, G.U. (1926). Why do we sometimes get nonsense-correlations between time-series? – A study in sampling and the nature of time-series, *Journal of the Royal Statististical Society* **89**, 1–69.

(*See also* **Serial Correlation**)

R.M. LOYNES

Small Area Estimation

Federally sponsored sample surveys have been designed to meet the needs of government agencies, legislative bodies, and health professionals for the comprehensive national estimates needed in the formulation and analysis of national policy initiatives directed to social, economic, or health related issues. These national data collection efforts are generally limited, however, in their capacity to produce reliable estimates (*see* **Estimation**) at the subnational or small area level. While **unbiased** direct estimates can often be derived at the **census** region or census division level for most nationally based survey efforts, sample size requirements and budget constraints preclude the capacity for the derivation of state level estimates and for geographic areas at the sub-state level such as counties. In spite of these budgetary constraints, a strong demand persists for accurate and reliable estimates of sociodemographic, economic, and health parameters at the subnational level.

Small area estimation can be defined as the application of model-based or indirect estimators, using data from surveys primarily designed to produce estimates of criterion measures at the national or regional level, to derive comparable estimates at more geographically disaggregated levels such as counties or other small areas. These techniques are typically characterized by prediction models or indirect estimators that make use of available survey data at the national or regional level, data on population characteristics at the state or local level, and available auxiliary (predictor) data at the local level that are related to the criterion measure of interest [21, 22, 26]. A survey design with sufficient sample size in every state or local area for which separate estimates are desired would permit the derivation of unbiased direct estimates of core survey measures with acceptable levels of precision to satisfy underlying analytic objectives. Budget and logistical considerations, however, make sample sizes too small to support estimates of criterion variables at the subnational level in most federally sponsored surveys. This has resulted in the development of indirect small area estimators. Such estimators use auxiliary survey data and population information that characterize the state or small area, together with national survey data, to develop a model-based small area estimate. The level of accuracy and reliability achieved by these small area estimation strategies is completely dependent on the degree to which underlying model assumptions are satisfied. The type of predictor information available, the functional relationship of the predictor information to the specified criterion variables, and the assumptions underlying each prediction model narrow the number of procedures that are appropriate for a particular application.

As a consequence of empirical tests of validity and widespread usage, a set of small area estimation strategies has gained respectability under certain qualifying assumptions. This article provides a review of these alternative small area estimation techniques that have been developed and are currently being implemented. Attention is given to the underlying assumptions and data requirements to operationalize the respective estimation strategies. Furthermore, examples of specific applications of the small area estimation techniques are also provided to illustrate their pervasive utility.

Small Area Estimation Techniques

The NCHS Synthetic Estimator

The NCHS synthetic estimator is an approach formalized by the **National Center for Health Statistics** [20, 21, 26]. The underlying assumption of the model is that within a demographic subgroup, the estimate of the criterion variable for the small area is equivalent to that obtained for the nation or the census region in which the small area is located.

Sociodemographic information such as age, race, ethnicity, sex, and income must be available both for the sample survey and for the small areas. D domains are formed by cross classification of these demographic variables. To estimate the **mean** of the characteristic Y for the small area l, the estimate of $\overline{Y}(d)$ for each of the D domains is calculated from the survey data. The small area estimate, $\overline{Y}_s(l)$, for area l is the sum of the weighted average of $\overline{Y}(d)$ across all domains, where the weight $(P(ld))$ is the proportion of the population of small area l that is in each domain. More specifically,

$$\overline{Y}_s(l) = \sum_{d \in l} P(ld)\overline{Y}(d),$$

where $\overline{Y}_s(l)$ is the NCHS small area estimator of the mean for criterion variable Y in small area l,

$P(ld)$ is the proportion of the lth area's population that belongs to domain d, and $\overline{Y}(d)$ is a national or regional survey estimate of the mean value of the criterion variable Y for domain d.

The Sample Regression Estimator

The sample regression estimator is based on a **regression** model using selected predictor (symptomatic) variables as independent variables and sample data for the variable of interest as the criterion or dependent variable. This approach is generally attributed to Ericksen [8, 9]. Criterion variable data must be available for a sample of n small areas selected from the set of N small areas in the total population. These small areas are referred to as primary sampling units (PSUs) and the n small areas in the sample as sampled PSUs. Estimates of the criterion variable are computed for these sampled PSUs. For most national household surveys, the PSUs are defined as counties or groups of contiguous counties. Predictor information is also needed for the sample PSUs and the nonsampled PSUs for which small area estimates are desired. Using predictor data for these PSUs, a regression model is developed to predict Y_r. More specifically,

$$Y_r = X\beta + \varepsilon,$$

where Y_r is an $n \times 1$ vector of values for the criterion variable in the n sample PSUs, X is an $n \times (p+1)$ matrix containing the set of p symptomatic indicators for the n sampled PSUs and an indicator for an intercept term, β is a $(p+1) \times 1$ vector of regression coefficients, and ε is an $n \times 1$ vector of stochastic errors.

The values of the symptomatic indicators for small areas, defined at the same level of geographic aggregation as the PSUs, are then substituted into the estimated regression model to derive the estimate of the criterion variable for the small areas. Specifically, the sample regression estimator for small area l (e.g. county level) is obtained as

$$Y_r(l) = X(l)\hat{\beta},$$

where $Y_r(l)$ is the sample regression estimator of the mean for the criterion variable in small area l, $X(l)$ is the $(p+1)$ vector of symptomatic information for local area l, and $\hat{\beta}$ is the regression estimate obtained in fitting the model for the data from the sample PSUs.

If the small area of interest is at the state level (S), the sample regression estimator developed at the county level would be applied in the following manner:

$$\overline{Y}(S) = \sum_{l \in S} P(lS) Y_r(l),$$

where $P(lS)$ is the proportion of the population in state S in small area (county) l.

The Base Unit Estimator

For those situations in which the linearity assumption of the sample regression model is suspect, an alternative strategy has been developed, which is referred to as the base unit or **poststratified** estimator [5, 14]. This small area estimation technique divides the small area of interest into constituent geographic sectors or base units, which might be counties, enumeration districts, or other geographic subunits. The small area l for which a criterion variable estimate is to be derived is referred to as the target area and further subdivided into target area base units. Counties would be the base units within target areas such as states. Unlike other methods that use symptomatic information directly for estimation, this procedure uses the symptomatic information to group the base units. G groups are formed using a suitable clustering **algorithm** or by a minimum variance **stratification** method. All target base units belonging to the small area of interest are assigned to one of the G poststrata based upon the symptomatic information. An estimate of the criterion variable for each of the target base units is obtained from the sample base units in the poststratum to which it has been assigned. In essence, each target base unit estimate can be considered as a small area estimate [4, 6].

An estimate of $\overline{Y}(g)$, the mean of the criterion variable of interest, is calculated for each of the G groups or poststrata by taking a weighted average of the estimates of the criterion variable across the sample base units that comprise each group. The mean estimate of the criterion variable for the gth group ($g = 1, 2, \ldots, G$) is given by

$$\overline{Y}(g) = \sum_{i \in g} W(i)\overline{Y}(i),$$

where $W(i)$ estimates the proportion of the total population of sample base units in group g that is represented by base unit i, and $\overline{Y}(i)$ is an estimate

of the criterion variable for the ith sample base unit. The base unit estimate of the criterion variable for each small area l is calculated as

$$\overline{Y}_b(l) = \sum_{g \in l} P(lg)\overline{Y}(g),$$

where $P(lg)$ is the proportion of the population of small area l that is classified in group g.

The base unit estimator bears a striking resemblance to the NCHS synthetic estimator. The primary difference lies in the method of poststrata construction vs. domain formation [24]. The base unit estimator links all individual observations within a sample base unit to a particular group or poststrata, based on symptomatic information for the entire unit.

The Composite Small Area Estimator

The composite small area estimator takes the form of a weighted average of two component small area estimators for small area l,

$$\overline{Y}_c(l) = C(l)\overline{Y}_1(l) + [1 - C(l)]\overline{Y}_2(l),$$

where $C(l)$ is an appropriately chosen weight and $\overline{Y}_1$ and $\overline{Y}_2$ are alternative small area estimators for local area l.

Schaible [25] has demonstrated that, with a judicious selection of composite weights, the composite estimator will have a **mean square error** that is smaller than the mean square error of the individual estimators. When the expected value of $\mathrm{E}[\overline{Y}_1(l) - \overline{Y}(l)][\overline{Y}_2(l) - \overline{Y}(l)]$ is small relative to the mean square error of $\overline{Y}_2(l)$, the weight that will minimize the composite estimator's mean square error can be approximated as

$$\frac{C(l) = 1}{[1 + R(l)]},$$

where

$$\frac{R(l) = \mathrm{MSE}[\overline{Y}_1]}{\mathrm{MSE}[\overline{Y}_2]}.$$

Bayesian Methods Using Hierarchical Models

A **hierarchical model** has also been utilized to model the geographic variation of health care measures expressed in terms of **binary** outcomes. The primary objective of this approach is to account for the **small area variation** that is generally ignored by the other small area estimation strategies [17]. As part of the estimation scheme, available **covariates** at the local level are incorporated in the model specification to improve on the predictive capacity for small areas. In addition, the variable of response due to local effects is explicitly incorporated into the model. Estimates and their accuracy are derived using **Bayesian** predictive **inference** [17].

Using this approach, the small area estimator, $Y_{h,s}$, of a mean for criterion measure of interest at the state level (s) is based on the posterior mean of $Y_{h,s}$,

$$\overline{Y}_{h,s} = \sum_{cbk \in sp} \frac{Y_{cbk}}{N} + \sum_{cb \in sp} (N_{cb} - n_{cb}) \times \frac{\mathrm{E}(p_{cb}|Y_{sp})}{N},$$

where b represents a demographic class; c represents a county in state s; k represents an individual in demographic class b and county c; sp indicates selection in the sample; N_{cb} and n_{cb} are the population and sample sizes in class b of county c; p_{cb} denotes the probability than an individual in demographic class b, county c, has criterion measure outcome of interest (Y has binary measure: $Y = 1$ indicates yes; $Y = 0$ indicates no); and $\mathrm{E}(p_{cb}|Y_{sp})$ represents the posterior mean of p_{cb}, where

$$\ln\left[\frac{p_{cb}}{(1 - p_{cb})}\right] = X_{cb}\beta_c$$

is defined as a **logistic regression** model, so that X_{cb} is a vector of covariates which characterizes the demographic class within county c.

In this setting the logistic parameter is allowed to vary across counties to incorporate local error in the model [17]. Efforts to approximate the mean squared error for this type of small area estimator have been made by Prasad & Rao [23].

Applications of Small Area Estimation Techniques

These small area estimation strategies, and variants of the techniques described, have been applied to a widespread set of social, demographic, economic, and health related criterion variables for which local area estimates are desired. With respect to health

specific applications, the NCHS synthetic estimator has been used to produce state specific estimates of mortality [15], disability [15, 18, 19, 20], utilization of medical services [15], infant and maternal health characteristics [11], and functional dependency [7]. In addition, the base unit method has been used to derive state specific estimates of health insurance coverage [3].

Postcensal estimates of population growth have been derived at the state, county, and local level, using variants of the sample regression small area estimator [8, 16]. In addition, indirect regression type small area estimators have been used to produce state and county level estimates of personal income and annual income by the Bureau of Economic Analysis [1]. A composite type small area estimator has also been used by the Bureau of the Census to produce state estimates of **median** annual income for four-person families [10].

The Bureau of Labor Statistics produces state and local area employment and unemployment estimates under a federal–state cooperative program using indirect regression type estimators [27]. The NCHS synthetic estimator has also been used to produce small area unemployment and housing estimates [12]. In addition, the Department of Agriculture has implemented regression type small area estimators in the derivation of county estimates of crop acreage [2] and has used composite type estimators to derive county estimates of crop production and livestock inventories [13].

References

[1] Bailey, W., Hazen, L. & Zabronsky, D. (1996). State, metropolitan area, and county income estimation, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[2] Bellow, M., Graham, M. & Iwig, W. (1996). County estimation of crop acreage using satellite data, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[3] Braden, J.J. & Cohen, S.B. (1994). An application of small area estimation techniques to derive state level estimates of health insurance coverage from the 1987 National Medical Expenditure Survey, *Journal of Economic and Social Measurement* **20**, 193–213.

[4] Cohen, S.B. (1979). A modified approach to small area estimation, in *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse Research Monograph 24. DHEW Publication No. (ADM) 79–801. Health Resources Administration. US Government Printing Office, Washington, pp. 98–134.

[5] Cohen, S.B. & Kalsbeek, W.D. (1977). An alternative strategy for estimating the parameters of local areas, in *American Statistical Association 1977 Proceedings of the Social Statistics Section*. American Statistical Association, Alexandria, pp. 781–786.

[6] Cox, B.G. & Cohen, S.B. (1985). *Methodological Issues for Health Care Surveys*. Marcel Dekker, New York.

[7] Elston, J.M., Koch, G.G. & Weissert, W.G. (1991). Regression-adjusted small area estimates of functional dependency in the noninstitutionalized American population age 65 and over, *Journal of the American Public Health Association* **81**, 335–343.

[8] Ericksen, E.P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas, *Demography* **10**, 137–159.

[9] Ericksen, E.P. (1974). A regression method for estimating population changes of local areas, *Journal of the American Statistical Association* **69**, 867–875.

[10] Fay, R. & Nelson, C. (1996). Estimation of median income for 4-person families by State, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[11] Gonzalez, J.E., Placek, P.J. & Scott, C. (1996). Synthetic estimation in followback surveys at the National Center for Health Statistics, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[12] Gonzalez, M.E. & Hoza, C. (1978). Small area estimation with application to unemployment and housing estimates, *Journal of the American Statistical Association* **73**, 7–15.

[13] Iwig, W. (1996). The National Agricultural Statistical Service County Estimates Program, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[14] Kalsbeek, W.D. (1973). A Method for Obtaining Local Postcensal Estimates for Several Types of Variables, *Ph.D. dissertation*. University of Michigan, Ann Arbor.

[15] Levy, P.S. & French, D.K. (1977). *Synthetic Estimation of State Characteristics Based on the Health Interview Survey*. Vital and Health Statistics: Series 2, No. 75, DHEW Publication (PHS) 78–1349 US Government Printing Office, Washington.

[16] Long, J.F. (1996). Postcensal population estimates: states, counties, and places, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[17] Malec, D. (1996). Model based state estimates from the National Health Interview Survey, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[18] Malec, D. & Sedransk, J. (1993). Bayesian predictive inference for units with small sample sizes: the case of binary random variables, *Medical Care* **5**, 66–70.

[19] Namekata, T. Levy, P.S. & O'Rourke, T.W. (1975). Synthetic estimates of work loss disability for each state and the District of Columbia, *Public Health Reports* **90**, 532–538.

[20] National Center for Health Statistics (1968). *Synthetic Estimates of Disability*, PHS Publication No. 1759. US Government Printing Office, Washington.

[21] NIDA Research Monograph 24 (1979). *Synthetic Estimates for Small Areas*, DHEW Publication No. (ADM) 79–801. Health Resources Administration. US Government Printing Office, Washington.

[22] Office of Management and Budget (1993). *Statistical Policy Working Paper 21: Indirect Estimators in Federal Programs*, Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology. US Government Printing Office, Washington.

[23] Prasad, N.G.N. & Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators, *Journal of the American Statistical Association* **85**, 163–171.

[24] Purcell, N.J. & Kish, L. (1979). Estimation for small domains, *Biometrics* **35**, 365–384.

[25] Schaible, W.L. (1979). A composite estimator for small area statistics, in *Synthetic Estimates for Small Areas* (National Institute on Drug Abuse Research Monograph 24), DHEW Publication No. (ADM) 79–801. Health Resources Administration. US Government Printing Office, Washington, pp. 36–53.

[26] Schaible, W.L. (1996). *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

[27] Tiller, R., Brown, S. & Tupek, A. (1996). Bureau of Labor Statistics' state and local area estimates of employment, in *Lecture Notes in Statistics 108: Indirect Estimators in U.S. Federal Programs*. Springer-Verlag, New York.

STEVEN B. COHEN

Small Area Variation Analysis

This area of investigation in health services research seeks to detect and explain variation in the amount of health care consumed by residents in different "small areas" [7] (*see* **Small Area Estimation**). Areas in which the utilization is higher than average may be investigated further in the interests of lowering the costs of health care by preventing unnecessary services. Areas with particularly low utilization are of interest as places in which access to health care may be inadequate. Ideally, one would study variation about the desirable utilization rate, but the desirable or appropriate rate at which a procedure should be performed is not usually known. Small area variation analysis was pioneered by John Wennberg and Alan Gittelsohn, who showed that the probability of having a tonsillectomy, hysterectomy, or prostatectomy varied substantially among small geographic areas in the northeast United States, while rates for cholecystectomy, appendectomy, and herniorrhaphy showed substantially less variation [9]. Such studies were the impetus for the growth in **outcomes research**, and a variations analysis is one of the first steps recommended in outcomes research.

A small area may be a zip (postal) code area, a county, a state, or at times a country. (Related studies treat all the patients in a hospital or a dental practice as a "small area".) Some small areas are created specifically for variations analysis, e.g. a **hospital market area**.

A variations analysis begins with calculation of a utilization rate, in which the denominator is the number of residents in the small area and the numerator is the number of procedures utilized by the *residents* of the small area, no matter where they actually received the procedure. The data for calculating such rates usually come from hospital billing data, which are available from some Canadian provinces and, in the US, from some states and also from the Health Care Financing Agency for patients enrolled in Medicare. The rates are usually standardized by age and sex, often using the method of indirect **standardization** because the number of procedures in some strata may be very small.

Statistical analysis of area variations is often informal, consisting of graphs of the admission rates by area, and descriptive statistics. Descriptive statistics usually include the extremal coefficient (maximum rate divided by minimum rate), and the weighted or unweighted *coefficient of variation* (*see* **Standard Deviation**). These statistics are unsatisfying unless the numbers of procedures per area are very large, because they do not distinguish variation among the areas from variation within the area. The systematic component of variation (SCV) [6] does make this distinction, under the assumption that no individual has more than one procedure. The square root of the SCV is an estimate of the coefficient of variation under this assumption. The variance among areas can be calculated from a mixed model **analysis of variance**. An estimate of the coefficient of variation based on a **moment** estimate of the **variance** among areas is the coefficient of variation from analysis of variance (CVA) [4]. Of all the simple statistics, only the CVA incorporates a **confidence interval** and a significance test (*see* **Hypothesis Testing**).

Simulation work has shown that the most popular descriptive statistics are sensitive to such factors as the **prevalence** of the procedure under study, the number of small areas being considered, the likelihood of multiple admissions per person, and the population sizes and relative population sizes of the small areas [2]. The unweighted coefficient of variation and the extremal quotient are also less likely than other statistics to demonstrate true variation when it occurs [3].

"Variation" is not well defined, and graphs of admission rates for small areas do not always agree with the usual descriptive statistics, because the usual statistics adjust the variance among areas by the prevalence, permitting a highly prevalent procedure to have more variation than a low prevalence procedure. The rationale for this has never been formally justified. The CVA has been shown to be the best of the usual descriptive statistics that adjust for prevalence, in that it is relatively uncorrelated with the prevalence of the procedure under study [4].

Multivariable statistical methods have been used to estimate or display the area variation, and importantly to allow the incorporation of **covariates**. One method models the number of admissions per area as following a **Poisson distribution** with extra-Poisson variation (*see* **Overdispersion**) [10]. The Poisson assumption effectively assumes that there are no multiple admissions [10]. If there are multiple admissions per person, the extra-Poisson variability can be due

to that fact, as well as to true variation among the small areas. Formal **hierarchical modeling** permits assessment of trends at the person level as well as at the area level [5, 8].

Investigators often wish to answer whether there is "too much" variation, since this could signal a diagnosis, procedure, hospital, or small area that merits further study. If there are not multiple admissions per person, one can test whether the observed variation is significantly different from zero, using a simple **chi-square test**. Although the **alternative hypothesis** is unlikely to be true, some very small data sets will not pass this test. A second approach is to compare the variation for (say) the diagnosis of interest to the variation of another "standard" diagnosis. Hernia surgery has been suggested as a standard surgical diagnosis because it usually shows low variation. Confidence intervals for the CVA can be used in this way, and there is an associated F-test (*see* ***F*** **Distributions**) for the SCV statistic [4, 6]. Similarly, one can compare the variation in one geographic region to the variation in another for the same diagnosis.

Once the existence of variation has been established, it is common to test each area for significant differences from the standard (usually the average rate), with the **outlier** areas subjected to further study and perhaps intervention. A chi-square test is usually employed, without accounting either for the possibility of multiple admissions per person [1] or for **multiple comparisons**. Such adjustments should probably be made, however, in situations where conservative results are important, such as finding differences in mortality rates among hospitals. Attempts to explain why rates are too high or too low often fail to find any inappropriate utilization of services, and often find that variation in coding practices is responsible for the observed discrepancies.

Other statistical issues arise when investigators wish to perform variations analyses in very small areas, in which a substantial number of areas will have zero events. Investigators often combine several years of data in order to increase the number of events, but this could cause other problems in that the probability of multiple admissions will increase. Expanding the procedure or diagnosis of interest to include additional events may cloud the interpretation of the findings. Hierarchical modeling is another approach for handling very small numbers.

References

[1] Cain, K. & Diehr, P. (1992). Testing the null hypothesis in small area analysis, *Health Services Research* **27**, 267–294.

[2] Diehr, P., Cain, K., Connell, F. & Volinn, E. (1990). What is too much variation? The null hypothesis in small area analysis, *Health Services Research* **24**, 741–771.

[3] Diehr, P., Cain, K., Kreuter, W. & Rosenkranz, S. (1992). Can small-area analysis detect variation in surgery rates? The power of small-area variation analysis, *Medical Care* **30**, 484–502.

[4] Diehr, P., Cain, K., Ye, Z. & Abdul-Salam, F. (1993). Small-area variation statistics: methods for comparing several DRGs, *Medical Care* **31**, YS45–YS53.

[5] Gatsonis, C., Normand, S.L., Liu, C. & Morris, C. (1993). Geographic variation of procedure utilization. A hierarchical model approach, *Medical Care* **31**(Supplement 5), YS54–YS59.

[6] McPherson, K., Wennberg, J., Hovind, O. & Clifford, P. (1982). Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway, *New England Journal of Medicine* **307**, 1310–1314.

[7] Paul-Shaheen, P., Clark, J. & Williams, D. (1987). Small area analysis: a review and analysis of the North American literature, *Journal of Health Politics, Policy, and Law* **12**, 741–809.

[8] Shwartz, M., Ash, A.S., Anderson, J., Iezzoni, L.I., Payne, S.M. & Restuccia, J.D. (1994). Small area variations in hospitalization rates: how much you see depends on how you look, *Medical Care* **32**, 189–201.

[9] Wennberg, J. & Gittelsohn, A. (1982). Variations in medical care among small areas, *Scientific American* **246**, 120.

[10] Wolfe, R.A., Petroni, G.R., McLaughlin, C.G. & McMahon, L.F., Jr (1991). Empirical evaluation of statistical models for counts or rates, *Statistics in Medicine* **10**, 1405–1416.

PAULA DIEHR

Smith, Cedric Austen Bardell

Born: February 5, 1917, in Leicester, UK.
Died: January 16, 2002, in London, UK.

Cedric Smith, Weldon Professor of Biometry at the Galton Laboratory, University College London, from 1964 to 1982, was a leading exponent of biostatistical human genetics in the United Kingdom. His genetic work centered on the detection and measurement of human linkage, and he was influential in the growing use of Bayesian methods. As Edwards [1] noted, he was "the principal link between the modern development and the 1930 pioneers F. Bernstein, L. **Hogben**, L.S. Penrose, Julia Bell, R.A. **Fisher** and J.B.S. Haldane", but his major work was done before the computer-based revolution.

Smith was born on February 5, 1917 in Leicester, and after his schooling there and subsequently in London, he went in 1935 to Trinity College, Cambridge. Here, he obtained first-class honors in Part II of the Mathematical Tripos and a distinction in Part III. He started graduate studies in statistics in 1938 and obtained a doctorate in 1942. Undergraduates of this period will remember a fascinating talk by Smith on recurrent functions, entitled "On growing fish from seed" (these being invented names for specific functions), as an early example of his quirky humor and curious ingenuity. He was a member of the Society of Friends and worked during the war as a hospital porter.

In 1946, he joined the Galton Laboratory at University College London. He remained there throughout his career and became Weldon Professor of Biometry in 1964. J.B.S. Haldane was a strong influence and encouraged Smith to work on problems in the testing and estimation of **linkage** in human genetics. Smith was attracted to the use of the **likelihood** function to provide a test of the **null hypothesis** and as a basis for the estimation of the recombination fraction. In 1953, he applied the word "lods" (log **odds**) used earlier by G.A. **Barnard**. In 1955, N.E. Morton used lods in applying Wald's **sequential** probability ratio test to the testing for linkage, but Smith disliked this approach, having moved firmly toward the **Bayesian** position. His Bayesianism, though, was not entirely orthodox. For instance, he investigated the possibility of using ranges of approximation to prior probabilities, related to the range of betting odds that the subject would accept. In 1957, in joint work, he introduced counting methods for estimation of **gene frequencies** and **segregation ratios**, which were early examples of the **EM algorithm** for **maximum likelihood** estimation.

In 1954, Smith published *Biomathematics*, nominally the third edition of a book by W.M. Feldman first published in 1923, but essentially a new and highly original work. This was followed in 1966 and 1969 by a fourth edition in two volumes. As Edwards [1] remarks, "its charming idiosyncrasies endeared it to its admirers but rather distracted the orthodox student."

Smith was an active member of the **Royal Statistical Society**, particularly in the Research Section and on the Editorial Board of the **Journal**, *Series B*; of the British Region of the **International Biometric Society**, serving as President from 1971 to 1972; and of the Genetical Society. He was a notable coeditor of the *Annals of Human Genetics*, frequently contributing personally to the flow of important statistical papers published in that journal.

Reference

[1] Edwards, A.W.F. (2002). Professor C.A.B. Smith, 1917–2002, *Statistician* **51**, 404–405.

PETER ARMITAGE

Smoking and Health

Tobacco was introduced to Spain and England in the sixteenth century by explorers returning from the New World. In England an import duty of 2 pence a pound was imposed in 1590. Tobacco attained rapid popularity and was even praised as a prophylactic against many ills. Nonetheless, from early times it was condemned as a "noxious vice", foul smelling, loathsome custom, and harmful to the brain and lungs.

In 1900 an increase in cancer of the lung was noted by vital statisticians but definite trends in mortality and disease incidence really only became apparent after 1930. Two studies in Germany [24, 32] were particularly notable. Müller [24] documented the rise in the proportion of cases of lung cancer at postmortem from 0.2% to 1% over the period 1918 to 1937 in his institute in Cologne. He associated this with the 82% increase in tobacco consumption in Germany over this period, and contrasted this to a fall in alcohol and coffee consumption. He emphasized that over the years the preparation of cigarettes had changed–over time more and more of the coarse materials (e.g. stems) from tobacco leaves were incorporated in the final product. The relation between experimental production of cancer in animals by tobacco tar was described. Müller then went on to describe a clinical investigation in which he questioned the surviving relatives of 96 patients who had died from lung cancer. Only 3% of these patients had not smoked, while 75% were classified as heavy smokers. He compared this with a group of healthy men of the same ages and stated that 16% of these were nonsmokers and 36% heavy smokers. The latter group were estimated to smoke 1259 g tobacco per day, while the cancer of the lung group smoked 2900 g tobacco per day. He investigated the patients' exposure to other carcinogenic agents, and found that 17 had worked in industries where they were exposed to possible risk, e.g. painters, printers, and lead workers. He also investigated their experience of other respiratory illnesses, particularly influenza, but found no records of such illness in about half the cases. He concluded that tobacco consumption was the major cause of lung cancer.

Schairer & Schöniger [32], referring to Müller's study, went further, since similar increases in postmortems from lung cancer had been shown in their institute in Jena. They posted questionnaires on smoking habits, including quantity, illnesses, and occupational exposure to polluted air, to the relatives of 195 cases of cancer of the lung who had died between 1930 and 1941. They posted similar questionnaires to the relatives of individuals who had died of stomach, colon, esophagus, and tongue cancer as controls. Completed questionnaires were returned by 50–60% of those posted. In addition they mailed 700 similar questionnaires to living male residents of Jena aged 53–54 years (the average age of death of the cancer of the lung patients). Of these 39% were satisfactorily completed. Of the cancer of the lung cases, 3% were nonsmokers and 52% were heavy smokers. Amongst the other groups between 11% and 16% were nonsmokers and 21%–38% heavy smokers. They were unable to find any association of cancer of the lung with air pollution or previous respiratory illnesses. They concluded that one of the causes of the rise in lung cancer deaths was due to heavy smoking, but that this could not be the only cause.

Case–Control Studies

Following these early clinicopathologic investigations, a series of more carefully controlled studies were designed and undertaken independently by Wynder & Graham [40], Levin et al. [21], and Doll & **Bradford Hill** [6], and published almost simultaneously. They all commented on the increase in mortality of cancer of the lung. Doll & Hill discussed possible reasons for this increase, in particular general atmospheric pollution, e.g. from industry, coal fires and traffic, and from smoking tobacco. They concluded that, in view of the studies described above, the most likely association was with tobacco.

The British studies were promoted by the **Medical Research Council**, which also initiated the studies by Lawther [19] on bronchitis and air pollution. All three of these studies followed a similar pattern. Patients diagnosed as having cancer of the lung and admitted to hospital were notified to the investigators and interviewed. In all studies a group of **controls** from either the same or similar hospitals was also interviewed (*see* **Case–Control Study, Hospital-based**).

The study by Doll & Hill [6] has been considered as the model of a **case–control study**, and so will be described in some detail. Twenty London hospitals notified all patients admitted to them with carcinoma

of the lung, stomach, colon, or rectum. "On receipt of the notification a research almoner (medical social worker) visited the hospital to interview the patient using a set questionnary". In addition the almoners were required to make similar inquiries of a group of "noncancer control" patients of the same sex, within the same 5-year age group, and at the same hospital at or about the same time.

The interviewers had not been told of the hypothesis that was being tested; thus, apart from questions on smoking habits there were others on, for example, place of residence, exposure to war gases, occupational hazards, etc. Obviously the interviewers could not be blinded by the illnesses of the patients they were interviewing (*see* **Blinding or Masking**). However, a number of the patients thought to have carcinoma at the time of the interview were subsequently found not to have the condition. The smoking habits of the patients incorrectly considered to have lung cancer were "sharply distinguished from the habits of those patients who did in fact have carcinoma of the lung", but did not differ from those of the other patients interviewed. Thus the authors concluded that the results could not be attributed to the results of **interviewer bias**.

All these case–control studies showed that patients with cancer of the lung were more likely to be smokers than those with other cancers or noncancer patients, and that those who smoked, smoked more cigarettes. None of the other possible suspected agents, e.g. air pollution, area of residence, or exposure to war gases, showed such a clear differentiation between the cases and controls. Doll & Hill, in particular, looked at other possible sources of **bias**, e.g. selection of the patients and controls, and were unable to find any particular bias in their groups (*see* **Bias, Overview**; **Bias in Case–Control Studies**; **Bias in Observational Studies**).

All three groups of authors concluded that the smoking of cigarettes was associated with cancer of the lung. Levin went so far as to state "the data suggest, although do not establish, a causal relation between cigarette and pipe smoking and cancer of the lung and lip respectively". When these results were reported to the main board of the Imperial Tobacco Company, according to a participant, the chairman was so appalled by the findings that he turned to his board and said "surely these results cannot be correct since we produce a clean, hygienic product"!

Prospective Surveys

The results of these **retrospective studies** were confirmed by a series of prospective studies (*see* **Cohort Study**). In these the smoking habits of a defined group were first ascertained and then the **causes of death** during several years' observations recorded. Different populations and strategies were used in these investigations. Doll & Hill [7, 8] sent a simple questionnaire to doctors on the medical register. They were notified of all deaths in the 34 000 who provided usable replies by the Registrar General. The group was enrolled in October 1951, aged 35 years or more. This group has been followed since that time. Hammond & Horn [12] enrolled a large number of American Cancer Society Volunteers, each of whom was asked to have a questionnaire on smoking completed by 10 white men aged 50–69 years. These were followed for 44 months only. **Dorn** [9] questioned 248 000 men who had served in the armed forces between 1917 and 1940 and who held US Government Life Insurance polices. Dunn et al. [10] questioned 67 000 men aged 35–64 years in nine occupations in California who were suspected of being subject to a higher than usual occupational risk of lung cancer, and followed them for about 48 months. Best et al. [4] questioned 78 000 Canadian veterans and their dependents (pensioners) aged 35 or more years, and followed them for 72 months.

All these studies showed a consistent gradient in the total mortality ratio (after adjusting for age) (*see* **Standardization Methods**) from nonsmoker to heavy smokers, ranging from 1.06 to 1.55 for those smoking less than 10 cigarettes per day, to 1.85 to 2.5 for those smoking 40 or more cigarettes per day. For lung cancer the ratios for these amounts varied from 4.4 to 8.4, and 15.1 to 43.7.

The investigators looked at causes of death other than cancer of the lung, and consistently demonstrated an association with smoking for bronchitis and emphysema, cancer of the larynx, cancer of the oral cavity, cancer of the esophagus, stomach and duodenal ulcers, other circulatory diseases, and coronary artery disease.

Other aspects of smoking, such as pipes, cigars, inhalation and stopping smoking, were investigated. The studies all showed that the **risks** of developing cancer of the lung were much less for "pure" cigar and pipe smokers, and the risks of lung cancer

diminished with the length of time since the individual had stopped smoking.

The subjects in these studies were selected by the answering of a questionnaire, which might have introduced bias by the inclusion of more or fewer smokers who are in ill-health at the beginning of the study. However, in all studies it was shown that the association between deaths from lung cancer and smoking was more evident in the later than the earlier part of the observation period, the reverse of what would have been expected if there had been **selection bias**. A further source of bias might have been the accuracy of diagnosis, for example, if smokers were likely to be "overdiagnosed" as having cancer of the lung. But all studies showed an excess of total mortality associated with smoking, hence if cancer of the lung had been overdiagnosed, other causes would have been "underdiagnosed" (*see* **Bias in Cohort Studies**).

These retrospective and prospective studies of smoking and cancer of the lung were the seminal works which laid the benchmark for good case–control and prospective studies. The use of defined, comparable groups of cases and controls, clearly defined questions, blinding of interviewers, and demonstrating that patients originally included as "cases", but not confirmed, resembled controls are a model of a good study. The prospective investigations which showed the need for very large defined groups, adequate follow-up, and consistent diagnostic criteria have illustrated how the imaginative use of certain groups, for example, doctors, veterans, and voluntary organizations, makes such large studies feasible both financially and operationally.

Policy Toward Smoking

As a result of these studies a variety of bodies were set up between 7 and 14 years after the initial studies to examine the evidence of the relationship between smoking cigarettes and cancer of the lung, e.g. Medical Research Council [22], Ministry of Health [23], National Cancer Institute of Canada [27], and Netherlands Ministry of Social Affairs and Public Health [28]. All agreed that the relationship was established. However, the most thorough reviews of the evidence were those of the Royal College of Physicians, London [31], in 1962, and of the US Surgeon-General in 1964 [33].

The former considered a number of possible explanations of the association of cancer of the lung and smoking:

1. Years before cancer of the lung becomes manifest some early process produces the desire to smoke. This was considered improbable.
2. Smoking may not cause cancer but only determine the site at which it appears. This is disproved as other forms of cancer are not less common among smokers than nonsmokers.
3. The rising death rate from lung cancer was a consequence of the falling death rate from tuberculosis. There was no evidence for this, and the gender effects were not consistent.
4. Some factors might be independently associated with both lung cancer and smoking (*see* **Confounding**). This hypothesis was contradicted by studies, for example, in Seventh Day Adventists (all nonsmokers), when the only cases of lung cancer were in converts who were ex-smokers.
5. **Berkson** [2, 3] suggested that nonsmokers are a highly selected group who are "biologically self-protective", and endowed with "robustness in meeting mortal stress from disease generally" while Eysenck [11] stressed the "accelerated rate of living" of cigarette smokers as a possible explanation of their higher death rates. This hypothesis fails to account for the disproportionate increase in death rates among smokers from lung cancer as compared with other causes. Berkson's objections were refuted by the finding that the 1952 London smog episode increased the death rate for a number of causes, in particular bronchitis and coronary heart disease, but no one doubted its importance as a cause of mortality.
6. Since heavy smoking is associated with heavy drinking the latter was incriminated – but all studies showed that the effect of smoking was independent of alcohol consumption.
7. The possibility that motor vehicle exhausts might be an important cause of the rise in deaths of lung cancer was dismissed because road haulage workers and those living near roads did not show an excess risk independent of smoking.
8. The role of general air pollution was more complex, but since the relation of smoking and lung cancer could be shown in both urban and rural

areas this was considered of lesser importance that the smoking of cigarettes. The differences between mortality of rural and urban dwellers were explained, in part, by the duration of smoking.

The Surgeon-General's Committee judged the causal significance of the association on a number of criteria which had to include:

1. Consistency of the association
2. Strength of the association
3. Specificity of the association
4. Temporal relationships of the association
5. Coherence of the association.

Hill [14] expanded these criteria to:

1. Strength of the association
2. Consistency
3. Specificity
4. Relationship in time
5. Biological gradient
6. Biological plausibility
7. Coherence
8. Experiment
9. Reasoning by analogy (*see* **Hill's Criteria for Causality**).

This account does not deal with all the other evidence, pathological, biochemical, physiological, etc. considered by the Surgeon-General's and RCP Committees. Both concluded that cigarette smoking was causally related to lung cancer (*see* **Causation**), and that cigarette smoking far outweighs all other factors.

As a result of these reports, and others, governments and most members of the public began to take the problem of smoking more seriously. Tobacco companies reduced the amount of tobacco in cigarettes (from 1 g per cigarette to 0.75 g per cigarette) and introduced filters in most brands. They also became involved in a search for what they hoped would prove to be less harmful products.

Less Harmful Cigarettes

The government in the UK reacted to these initiatives by setting up an Independent Scientific Committee under the chairmanship of Lord Hunter [16] in 1973. The failure of these efforts is chronicled in Holland & Wood [15] and Waller & Froggatt [37]. A very large randomized controlled trial (*see* **Clinical Trials, Overview**) was mounted by Withey et al. [38, 39] in 1985 to determine the efficacy of reducing the tar and nicotine levels in cigarettes. They found that lowering the tar intake did not lead to any improvement in respiratory health, and there was some evidence that smokers of the low-tar cigarettes compensated in the way they smoked these cigarettes in order to absorb more tar and nicotine. In the US a workshop held at the World Conference on Smoking and Health [26] had recommended a number of measures, such as labeling of cigarettes as being harmful, addition of filters, and redesigning the cigarette to reduce inhalation. The main recommendations, however, were to reduce the tar and nicotine content of cigarettes in the hope that this would reduce harmful effects. The trial by Withey et al.demonstrated that this strategy was unlikely to be effective – and thus that only stopping smoking is likely to lead to any reduction of risk.

Prevention

In view of the failure to reduce the harm from cigarettes we are left with the options of stopping people from smoking, or better still stopping them from taking it up.

The attempt to stop people from smoking through providing an alternative source of nicotine, such as nicotine-containing skin patches or nicotine chewing gum, medicalizes the problem of tobacco addiction. It does lead to some benefit and governments and others have supported such efforts. Other approaches, for example the use of counseling, have had some success – but none is effective in more than a small proportion of individuals [17, 30].

Discouraging children from taking up smoking is perhaps even more important. But this must involve parents and teachers as well as others such as the children's peers. Several studies endeavored to determine the major factors that lead children to take up smoking. One of the largest of these studies by Swan et al. [25, 34] followed about 6000 children from entry to secondary school at age 11 to young adulthood, age 21, in Derbyshire from 1972 to 1983. The results showed that 30% of smoking was attributable to peer pressure and to exposure to parental smoking at 11–12 years of age. More than 70% of children will try smoking before age 16 irrespective of their attitudes and circumstances, but

only about 3.5% try for the first time after this age. There was a maximum incidence of regular smoking of about 20% per annum as children moved into the third year of secondary school, and their smoking behavior changed after they become dismissive of the health hazards. Thus to prevent regular smoking in adults one needs to create programs before age 11 which demonstrate and maintain awareness of the health hazards. This implies changes to the way education is organized.

Passive Smoking or Environmental Tobacco Smoke

That cigarette smoking contaminates the atmosphere and that people exposed to others smoking may experience nose and eye irritation, cough, and headache has always been recognized. But there was little evidence of any objective harm to health until two papers appeared in March and November 1974 in the *Lancet*. Harlap & Davies [13] showed in a study of over 10 000 infants studied prospectively that the infants of mothers who smoked had significantly more admissions to hospital for bronchitis and pneumonia, especially in the winter, and more injuries. This was dose-related and independent of birthweight, social class or birth order. Colley et al. [5, 20] confirmed the greater incidence of bronchitis and pneumonia in children aged less than 1 year, and also demonstrated that it was related to the amount of exposure to smoking by the parents of the baby. These effects were independent of sex, parental symptoms and disease, and number of siblings. Many studies have since been published of the effects of passive smoking on health. The most recent authoritative reviews in the UK [18] and US [35] have confirmed that environmental tobacco smoke increases the risk of chronic respiratory disease in adults by about 25%, the risk of respiratory illness in children by 50%–100%, and of lung cancer by about 24% (95% **confidence interval**: 11% to 38%). It may also increase the risk of ischemic heart disease, and exposure in pregnancy may lower **birthweight**.

Conditions Associated with Smoking

Apart from cancer of the lung, a number of other conditions are also related to smoking. It has been estimated that about 3 million people die each year from smoking-related diseases in developed countries [36]. The number dying in the whole world is of course, much greater, and will rise considerably [29]. Other cancers which have been considered as satisfying the Surgeon-General's and Hill's criteria are upper respiratory, bladder, pancreas, esophagus, stomach, kidney, and leukemia. Respiratory heart disease, chronic obstructive lung disease, stroke, pneumonia, aortic aneurysm, ischemic heart disease, peripheral vascular disease, cataracts, hip fracture, and periodontal disease also satisfy these criteria. In pregnancy, smoking increases the risk of limb reduction defects, spontaneous abortion, ectopic pregnancy, and low birthweight [36]. Some suggestions have been made that smoking protects from the occurrence of uterine fibroids, endometriosis, hypertensive disorders, vomiting in pregnancy, ulcerative colitis, and Parkinson's disease [1].

Current Concerns

The importance of smoking as a health hazard has now gained universal acceptance amongst the health professions and most governments. It has become far less acceptable in society now than it was 20 years ago. The suppression of smoking on transport, in offices, and in restaurants has had a profound influence in some countries. However, there have been counter-pressures. Although some governments appreciate the effect of a ban on cigarette advertising, or other governmental measures, there has been a consistent lack of willingness to suppress the promotion of this health hazard by most Western governments. Some even continue to subsidize the growth of tobacco. The problem of smoking in developing countries, in spite of valiant efforts by the **World Health Organization** (WHO) are even greater. Here smoking is still considered a status symbol and tobacco companies intent on preserving their market export to these countries cigarettes with much higher tar and nicotine levels than permitted in Western countries in order to foster habituation.

Concluding Comments

The epidemiologic investigation of smoking has been responsible for the development of case–control and prospective methods of investigation and for the formulation of acceptable criteria for assessing

cause–effect relationships. The studies over the past 50 years have provided superb examples of how to perform descriptive and experimental epidemiologic studies. The problem has also illuminated the relations between epidemiology, medical statistics, and health policy, and has led to the development of appropriate preventive strategies.

Although the proportion of individuals who smoke in the UK and US is now less than half of what it was in the 1950s, smoking is still the most important health hazard in these countries. More worrying is the changes in those who smoke. Whereas up to about 20 years ago the great majority of smokers were men, now more and more women have adopted this habit. In the early years of this century smoking was practiced by upper social class groups; now it is most common in the poor and deprived, who are least able to afford it. A further major concern is the targeting of developing countries in tobacco promotion, whose populations are only too ready to mimic the more developed countries.

References

[1] Baron, J.A. (1996). Beneficial effects of nicotine and cigarette smoking: the real, the possible and the spurious, *British Medical Bulletin* **52**, 58–73.

[2] Berkson, J. (1958). Smoking and lung cancer. Some observations on two recent reports, *Journal of the American Statistical Association* **53**, 28–38.

[3] Berkson, J. (1959). The statistical investigation of smoking and cancer of the lung, *Proceedings of the Mayo Clinic* **34**, 206–224a.

[4] Best, E.W.R., Josie, G.H. & Walker, C.B. (1961). A Canadian study of mortality in relation to smoking habits, a preliminary report, *Canadian Journal of Public Health* **52**, 99–106.

[5] Colley, J.R.T., Holland, W.W. & Corkhill, R.T. (1974). Influence of passive smoking and parental phlegm on pneumonia and bronchitis in childhood, *Lancet* **2**, 1031–1034.

[6] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung. A preliminary report, *British Medical Journal* **2**, 739–748.

[7] Doll, W.R. & Hill, A.B. (1954). The mortality of doctors in relation to their smoking habits. A preliminary report, *British Medical Journal* **1**, 1451–1455.

[8] Doll, W.R. & Hill, A.B. (1956). Lung cancer and other causes of death in relation to smoking, *British Medical Journal* **2**, 1071–1081.

[9] Dorn, H.F. (1958). The mortality of smokers and nonsmokers, in *American Statistical Association 1958 Proceedings of the Social Statistics Section*. American Statistical Association, Alexandria, pp. 34–71.

[10] Dunn, J.E., Linden, G. & Breslow, L. (1960). Lung cancer mortality of men in certain occupations in California, *American Journal of Public Health* **50**, 1475–1487.

[11] Eysenck, H.J., Tarrant, M., Woolf, M. & England, L. (1960). Smoking and personality, *British Medical Journal* **1**, 1456–1460.

[12] Hammond, E.C. & Horn, D. (1958). Smoking and death rates – report on forty-four months follow-up on 187,783 men. Part I. Total mortality. Part II. Death rates by cause, *Journal of the American Medical Association* **166**, 1159–1175; 1294–1308.

[13] Harlap, S. & Davies, M. (1974). Infant admissions to hospital and maternal smoking, *Lancet* **1**, 529–532.

[14] Hill, A.B. (1965). The environment and disease: association or causation, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[15] Holland, W.W. & Wood, R. (1995). Policies on prevention: the hazards of politics, *Proceedings of the Royal College of Physicians of Edinburgh* **25**, 189–203.

[16] Independent Scientific Committee on Smoking and Health (1975). First Report, Chairman R.B. Hunter, *Tobacco Substitutes and Additives in Tobacco Products: Their Testing and Marketing in the United Kingdom*. HMSO, London.

[17] Kunze, M. & Wood, M. (1984). Guidelines on Smoking Cessation, *UICC Technical Report Series* **79**. International Union Against Cancer, Geneva.

[18] Law, M.R. & Hackshaw, A.K. (1996). Environmental tobacco smoke. Tobacco and health, R. Doll, & J. Crofton, eds. *British Medical Bulletin* **52**, 22–34.

[19] Lawther, P.J. (1958). Climate, air pollution and chronic bronchitis, *Proceedings of the Royal Society of Medicine* **51**, 262–264.

[20] Leeder, S.R., Corkhill, R.T., Irwig, L.M., Holland, W.W. & Colley, J.R.T. (1976). Influence of family factors on the incidence of lower respiratory illness during the first year of life, *British Journal of Preventive and Social Medicine* **30**, 203–212.

[21] Levin, M.L., Goldstein, H. & Gerhardt, P.R. (1950). Cancer and tobacco smoking. A preliminary report, *Journal of the American Medical Association* **143**, 336–338.

[22] Medical Research Council (1957). *Tobacco Smoking and Cancer of the Lung*, Cmd 8387. HMSO, London.

[23] Minister of Health (1954). Smoking and lung cancer, *British Medical Journal* **1**, 465.

[24] Müller, F.H. (1939). Tabakmissbrauch und lungencarcinom, *Zeitschrift Krebsforschung* **49**, 57–85.

[25] Murray, M., Jarrett, L. & Swan, A.V. (1988). *Smoking Among Young Adults*. Gower, Aldershot.

[26] National Cancer Institute (1968). *Toward a Less Harmful Cigarette*, Monograph No. 28, E.L. Wynder & D. Hoffman, eds. US Department of Health, Education, and Welfare, PHS, National Cancer Institute, Bethesda.

[27] National Cancer Institute of Canada (1958). Lung cancer and smoking, *Canadian Medical Association Journal* **79**, 566.

[28] Netherlands Ministry of Social Affairs and Public Health (1957). *Nederlands Transaktion Geneeskundee* **101**, 459.
[29] Peto, R., Lopez, A.D., Boreham, J., Thun, M., Heath, C.W., Jr & Doll, R. (1996). Mortality from smoking worldwide, *British Medical Bulletin* **52**, 12–21.
[30] Royal College of Physicians (1991). *Preventive Medicine*. Royal College of Physicians, London, Chapter 2, pp. 13–26.
[31] Royal College of Physicians of London (1962). *A Report on Smoking and Health*. Pitman Medical, London.
[32] Schairer, E. & Schöniger, E. (1943). Lungenkrebs und tabakverbrauch, *Zeitschrift Krebsforschung* **54**, 261–269.
[33] Surgeon-General of the Public Health Service (1964). Report of the Advisory Committee, *Smoking and Health*, Publication No. 1103. US Department of Health, Education, and Welfare, PHS, Washington.
[34] Swan, A.V., Murray, M. & Jarrett, L. (1991). *Smoking Behaviour from Pre-adolescence to Young Adulthood*. Gower, Aldershot.
[35] US Environment Protection Agency (1993). *Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders*, NIH Publication No. 93–3605. US Department of Health and Human Services, PHS, Bethesda.
[36] Wald, N.J. & Hackshaw, A.K. (1996). Cigarette smoking: an epidemiological overview, *British Medical Bulletin* **52**, 3–11.
[37] Waller, R.E. & Froggatt, P. (1996). Product modification, *British Medical Bulletin* **52**, 193–205.
[38] Withey, C.H., Papacosta, A.O., Swan, A.V. et al. (1992). Respiratory effect of lowering tar and nicotine levels of cigarettes smoked by young male middle tar smokers. I. Design of a randomized controlled trial, *Journal of Epidemiology and Community Health* **46**, 274–280.
[39] Withey, C.H., Papacosta, A.O., Swan, A.V. et al. (1992). Respiratory effects of lowering tar and nicotine levels of cigarettes smoked by young male middle tar smokers. II. Results of a randomized controlled trial, *Journal of Epidemiology and Community Health* **46**, 281–285.
[40] Wynder, E.L. & Graham, E.A. (1950). Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma, *Journal of the American Medical Association* **143**, 329–336.

W.W. HOLLAND

Smoothing Hazard Rates

Introduction

In the analysis of lifetime data or time-to-event data, a primary interest is to assess the risk of an individual at certain times (or ages) (*see* **Survival Analysis, Overview**). Let T denote a lifetime variable with distribution function $F(t) = \Pr(T \leq t)$ and probability density function $f(t) = \mathrm{d}F(t)/\mathrm{d}t$. The risk of an individual at age t can be measured by the so-called **"hazard rate"** or "hazard function", which is defined as:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}, \quad \text{for } F(t) < 1. \tag{1}$$

That is, $\lambda(t)\,\mathrm{d}t$ represents the instantaneous chance that an individual will die in the interval $(t, t + \mathrm{d}t)$ given that this individual is alive at age t. The hazard rate provides the trajectory of risk and is widely used also in other fields. Engineers refer to it as "failure rate function" and demographers refer to it as "force of mortality function". The term "lifetime" simply denotes the time until the occurrence of an event of interest.

While parametric models provide convenient ways to analyze lifetime data, the necessary model assumptions, when violated, can lead to erroneous analyses and thus need to be checked carefully (*see* **Parametric Models in Survival Analysis**). We give a brief survey on hazard rate estimation in this article. No shape restriction on the hazard rate is assumed except for smoothness. Such a model-free approach is data driven and can be used for parametric model checking. The nonparametric approach of hazard rate estimation typically involves the smoothing of an initial hazard estimate. The brief survey of various smoothing hazard rate estimators provided here covers grouped lifetime data on the one hand and continuously observed lifetime data on the other.

For grouped data, the observations occur in the form of scatter plots (t_i, q_i), where q_i is an initial hazard estimate at the midpoint t_i of the ith time interval. Smoothing for such data corresponds to a scatter-plot smoothing or **nonparametric regression** step. As for continuously observed data, hazard rate estimation resembles **density estimation** (smoothing the increments of a cumulative function estimate). Almost any density estimation method can be adapted for hazard rate smoothing. The simplest such method is the kernel method, which should however be employed with care in the boundary region. More details are given later in the section "More on Kernel Hazard Estimators for Continuously Observed Data".

Smoothing Hazard Rates for Grouped Data: Nonparametric Graduation of Lifetables

The earliest nonparametric hazard rate estimate was the **life table** estimate based on grouped lifetimes (*see* **Grouped Survival Times**), which has been known for centuries. Assume for simplicity that lifetimes are grouped into intervals of unit length with midpoints $t_1, \ldots, t_p$. Let n_i denote the number of individuals alive (or at risk) at the beginning of interval i, and d_i denote the number of observed deaths during this interval. An ad hoc estimate of the hazard rate for the ith interval is the so called death rate, $q_i = d_i/n_i$ (for intervals of length Δ the death rate is replaced by $d_i/(\Delta n_i)$). A plot of the raw death rates at various times t_i typically yields a curve that is ragged, indicating high variability; see Figure 1 for an example concerning the death rates of 1000 female Mediterranean fruit flies. Dead flies were counted daily, and q_i is the death rate at day i.

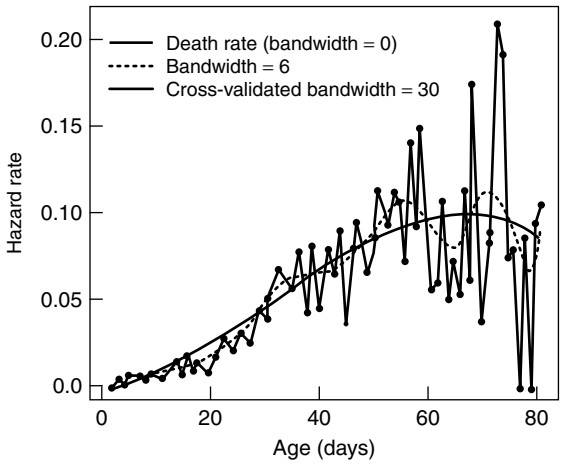

Figure 1 Three hazard rate estimates for the survival of 1000 female Mediterranean fruit flies. (a) death rates (thin line) (b) smoothed hazard rate with fixed bandwidth $b = 6$ (solid line) (c) smoothed hazard rate with least-squares cross-validated bandwidth choice $b = 30$ (bold line)

Since the actual hazard rate λ is typically assumed to be a smooth function, smoothing the death rates provides an aesthetically improved estimate (see Figure 1 for two versions of smoothed death rates). A smoothing procedure, when applied properly, also improves the statistical performance of the resulting hazard rate estimator.

For example, the smoothed death rates typically have a faster convergence rate than the unsmoothed death rates. The smoothing of death rates was pioneered by actuaries who referred to these smoothing methods as "linear graduation" or "nonparametric graduation", in contrast to "analytic graduation" based on parametric models (*see* **Actuarial Methods**). The term "linear" refers to the fact that these nonparametric graduation methods yield hazard estimates of the form

$$\hat{\lambda}(t) = \sum_{i=1}^{p} c_i(t) q_i, \text{ where}$$

$$\sum_{i=1}^{p} c_i(t) = 1, \text{ at each time } t. \qquad (2)$$

That is, the resulting hazard estimate at age t is a weighted average of the death rates with weights $c_i(t)$ specified by the method of graduation and adjusted locally at each age t.

The graduation (or smoothing) process typically reduces the variance of the resulting hazard estimates at the expense of introducing biases. The graduated or smoothed hazard estimate converges to the true hazard rate at a slower rate than the $\sqrt{n}$ rate, which holds for a parametric (or analytic) graduated hazard estimate.

Moving averages, local weighted **least-squares** methods and the so-called Whittaker–Henderson estimates have been the earliest proposals among a variety of different possible graduation methods, and are commonly adopted by actuaries (see [6, 32]). Any nonparametric regression method can be used to graduate **life tables** in order to obtain a smooth hazard rate estimate. One just applies the chosen smoother, which could be a spline or kernel method, to the scatter plot $\{(t_i, q_i),\ i = 1, \ldots, p\}$. The Whittaker–Henderson estimate resembles a spline estimate. The kernel method for graduation (see [4, 8]) is conceptually simple but needs to be applied with caution in the boundary region of the data, owing to its large bias there.

For the graduation of grouped data, we recommend the local polynomial method, which is also called the locally weighted least-squares method. This graduation method has been credited to the famous mathematician J.P. Gram, perhaps best known for his contributions to **Gram–Schmidt** orthogonalization; see [31, 51] for historical reviews. Specifically, in his doctoral dissertation, Gram [21] suggested a weighted least-squares method to fit a smooth curve locally by polynomials. The explicit form of Gram's estimate using a local linear fit is given in (4) below.

The local polynomial method is well suited for graduating initial hazard estimates based on life tables. As a least squares based procedure, it is simple to interpret, and automatically includes boundary corrections. For the kernel method, boundary corrections require the implementation of special boundary kernels. Both kernel and local polynomial methods are theoretically more tractable than the spline method, especially for lifetime data, which are often incomplete. Some asymptotic results for the local polynomial estimator are reviewed in the next section.

We note that the death rate q_i can be replaced by any initial estimate of the hazard rate. For example, the central death rate, $q_{c_i} = 2d_i/(n_i + n_{i+1})$, is a good alternative. If death rates are used in (2), it is recommended (see (13) of next section and [60]) to include a transformation of the smoothed death rates $\hat{\lambda}(t)$, and to use $-\log(1 - \hat{\lambda}(t))$ as the final hazard estimate. This transformation reduces the bias resulting from grouping the data. This bias can be substantial at extreme ages (i.e. for large t) and may result in inconsistent estimates of the hazard rate. If the central death rates are used in (2), another transformation (see (15) of next section and [42]) of the smoothed central death rates is recommended instead.

As for the choice of the smoother in (2), it is a judgment call, and typically, the choice of an adequate smoothing parameter is more important. The sampling or asymptotic properties of the resulting hazard rate estimator are much more complicated than in the standard regression setting, as the q_i or other initial hazard estimates are not independent of each other. The incompleteness of lifetime data further complicates theoretical analysis. Therefore, much is yet to be explored in hazard rate estimation based on smoothing life tables.

For an overview and details of the kernel smoothing method, see [59]; for the spline method, [25]; and for the local polynomial method, [14].

More on Local Polynomial Hazard Smoothing for Grouped Data

In addition to the grouping, we shall assume that the lifetimes $T_1, T_2, \ldots, T_n$, based on a cohort of n individuals, are subject to random censoring by $C_1, C_2, \cdots, C_n$. Let $I_1, I_2, \ldots, I_p$ denote a partition of p ordered intervals over a time interval of length L. For the jth individual, the value of $\delta_j = 1_{\{X_j = T_j\}}$ is known but not the actual value of $X_j = \min(T_j, C_j)$. It is only known that $X_j \in I_i$ for some i. Observed are (d_i, n_i), where $d_i = \sum_{j=1}^{n} 1_{\{X_j \in I_i, \delta_j = 1\}}$ is the number of observed deaths in the interval I_i, and $n_i = \sum_{j=1}^{n} 1_{\{X_j \in I_k, \text{ for some } k \geq i\}}$ is the number of individuals at risk at the beginning of the interval I_i.

For simplicity of presentation, we shall assume that the intervals I_i are of equal length Δ and that the first interval starts at zero. The nonequal length case can be handled similarly as in nonparametric regression with non-equidistant design points and will not be discussed here. The grouped data can thus be summarized in life table form, which consists of data pairs (t_i, q_i), $i = 1, \ldots, p$. Here, $t_i = \Delta(i - 1/2)$ is the midpoint of the ith interval I_i and $q_i = \tilde{q}(t_i) = d_i/(\Delta n_i)$ is the death rate (out of those alive) for interval I_i. A closer look at $\hat{q}$ reveals that it is an empirical estimate of the population death rate defined by

$$q(t) = \Delta^{-1} \Pr\left(T \in \left(t - \frac{\Delta}{2}, t + \frac{\Delta}{2}\right) \middle| T > t - \frac{\Delta}{2}\right), \tag{3}$$

and one expects $q(t)$ to be close to the true hazard function $\lambda(t)$, provided that Δ is small.

The local polynomial smoother due to Gram [21, 22], is based on smoothing the lifetable data $\{(t_i, q_i), i = 1, \ldots, p\}$ by locally fitting a polynomial of fixed degree r. Thus, given a bandwidth or window of size $b = b_n$, for estimation at age t, a polynomial $g(x - t)$ of degree r is fitted to all life table data points (t_i, q_i) for which $|t - t_i| \leq b$. The coefficients of the polynomial $g(\cdot)$ are obtained via the weighted least-squares criterion and the value of the fitted polynomial at t (i.e. the intercept) is the hazard estimate. A common choice is to fit local linear polynomials (i.e. $r = 1$).

For $r = 1$, this estimate, denoted by $\hat{q}(t)$, is equal to the minimizer for a_0 of

$$\sum_{i=1}^{p} w_i K\left(\frac{t - t_i}{b}\right) \{q_i - [a_0 + a_1(t_i - t)]\}^2. \tag{4}$$

Here w_i are case weights, typically chosen as $w_i = n_i$, and K is a nonnegative kernel function satisfying

$$V = \int K^2(x)\,\mathrm{d}x < \infty. \tag{5}$$

We recommend using either the Epanechnikov kernel

$$K(x) = .75(1 - x^2), \quad -1 \leq x \leq 1,$$
$$\text{or the Gaussian kernel } K(x) = (2\pi)^{-1/2} \mathrm{e}^{-x^2/2}. \tag{6}$$

The bandwidths should satisfy

$$b_n \to 0 \text{ and } nb_n \to \infty; \tag{7}$$

The weighted least-squares method is used for two reasons. First, in the spirit of smoothing methods, it gives remote observations less influence in a way that can be controlled by choice of bandwidth and kernel in (4). Second, it allows to address the high degree of heteroscedasticity (*see* **Scedasticity**) of the lifetable estimate q_i, through the choice of the case weights w_i in (4). Bias and variance expressions are derived in [60] and summarized below.

First, we define a constant that appears in the leading bias term:

$$B = \frac{1}{2} \int x^2 K(x)\,\mathrm{d}x \tag{8}$$

Under the kernel and bandwidth conditions (5) and (7), and if in addition

$$\Delta \to 0, \text{ and } \Delta \log n/b \to 0 \text{ as } n \to \infty, \tag{9}$$

we have for t with $F(t) < 1$ and $G(t) < 1$, and B and V as in (5), (8),

$$\begin{aligned} \mathrm{bias}(\hat{q}(t)) &= -\frac{\Delta}{2}\lambda^2(t) + \frac{\Delta^2}{24}[\lambda^{(2)}(t) + 4\lambda^3(t)] \\ &\quad + b^2\lambda^{(2)}(t)B + o(b^2) + o(\Delta^2) \end{aligned} \tag{10}$$

$$\mathrm{var}(\hat{q}(t)) = \frac{1}{nb}\left\{\frac{\lambda(t)}{[1 - F(t)]\,[1 - G(t)]}V + o(1)\right\}. \tag{11}$$

Bias Reduction Transformation

Note that the leading term of the variance in (11) is the same as for the kernel estimate for continuously observed data in (24). The terms in (10) involving b correspond to the bias due to smoothing and are also the same as for continuously observed data with $k = 2$ in (23). The terms involving Δ in (10) correspond to an additional bias due to the grouping of the data. This additional bias can be improved by the transformation $\phi(x) = -\log(1 - \Delta x)/\Delta$, which is motivated by the relation

$$\begin{aligned}\Delta q(t) &= 1 - \frac{1 - F\left(t + \frac{\Delta}{2}\right)}{1 - F\left(t - \frac{\Delta}{2}\right)} \\ &= 1 - \exp\left[-\int_{t-\frac{\Delta}{2}}^{t+\frac{\Delta}{2}} \lambda(x)\,\mathrm{d}x\right] \approx 1 - \mathrm{e}^{-\Delta\lambda(t)}.\end{aligned} \tag{12}$$

Thus, we propose the transformed estimate

$$\phi(\hat{q}(t)) = \frac{-\log(1 - \Delta\hat{q}(t))}{\Delta}, \tag{13}$$

which has the same variance expression (11) as $\hat{q}$ has, but a bias of smaller order:

$$\begin{aligned}\mathrm{bias}(\phi(\hat{q}(t))) = \frac{\Delta^2}{24}\lambda^{(2)}(t) + b^2\lambda^{(2)}(t)B \\ + o(b^2) + o(\Delta^2).\end{aligned} \tag{14}$$

Comparing (10) and (14), we see that $\hat{q}(t)$ has an additional bias, $-\frac{\Delta}{2}\lambda^2(t) + \frac{\Delta^2}{6}\lambda^3(t)$, as compared to $\phi(\hat{q}(t))$. In addition to this bias reduction there are other advantages in using $\phi(\hat{q}(t))$ rather than $\hat{q}(t)$, especially when hazards at extreme ages are of primary interest (see [60] for details). If the central death rate, q_{c_i}, is used in (4) instead of the death rate, q_i, a different transformation is proposed in [42], given by:

$$\psi(\hat{q}_c(t)) = \frac{1}{\Delta}\log\frac{2 + \Delta\hat{q}_c(t)}{2 - \Delta\hat{q}_c(t)} \tag{15}$$

We close this section by pointing out that the rate of convergence of $\hat{q}(t)$, $\phi(\hat{q}(t))$, $\hat{q}_c(t)$ or $\psi(\hat{q}_c(t))$, and the choice of the bandwidth b can be derived analogous to that of the kernel estimate $\hat{\lambda}$ in the section "More on Kernel Hazard Estimators for Continuously Observed Data", with Δ playing a role in the asymptotic bias term. The program to compute $\hat{q}(t)$ in (4) or $\hat{q}_c(t)$ and their corresponding transformed estimates, $\phi(\hat{q}(t))$ in (13) or $\psi(\hat{q}_c(t))$ in (15) is very simple, and so is the computation of the cross-validated bandwidths as employed in [42] and [60].

The hazard rate estimate, based on the least-squares cross-validated bandwidth, calculated from the lifetimes for 1000 female Mediterranean fruit flies is plotted in Figure 1. The lifetimes are grouped into days. Here the cross-validated bandwidth is fairly large ($b = 30$), owing to the large variation of the death rates after day 60. The hazard plot was truncated at day 81 when there were only 10 flies left.

Smoothing Hazard Rates for Continuously Observed Data

The grouped data situation discussed in the previous section is common for demographic data that were observed at fixed time points or grouped for convenience. The estimation of hazard rates for continuously observed data is conceptually close to **density estimation**. To see this, consider, instead of (1), the hazard rate function as the derivative of the cumulative hazard function $\Lambda(t) = \int_0^t \lambda(x)\,\mathrm{d}x$. A hazard rate estimate can thus be obtained, analogous to a density estimate, by smoothing the increments of an estimate of $\Lambda(t)$.

Watson and Leadbetter [62, 63] were the first to propose and study such a smoothed hazard estimator using the empirical cumulative hazard estimate $\Lambda_n(t)$ based on an independent and identically distributed (i.i.d.) sample of lifetimes (that is, the $\Lambda_n(t)$ in (18) with all $\delta_{[j]} = 1$). They propose the following convolution type hazard estimator.

$$\hat{\lambda}_n(t) = \int W_n(t - x)\,\mathrm{d}\Lambda_n(t), \tag{16}$$

where W_n is a sequence of smooth functions approaching the Dirac delta function for large n. This delta-sequence method is quite general and covers several types of smoothing methods, including the kernel method (with $W_n(x) = b_n^{-1}K(x/b_n)$). Another type of hazard estimator proposed in [63] is of a ratio

type,

$$\tilde{\lambda}_n(t) = \frac{\hat{f}_n(t)}{1 - \hat{F}_n(t)}, \tag{17}$$

where $\hat{f}_n$ can be any density estimate of the lifetime density f and $\hat{F}_n$ is an empirical estimate of the lifetime distribution function F. Both types of hazard estimators have the same asymptotic variance but different asymptotic biases [49]. The convolution type estimator $\hat{\lambda}_n$ has prevailed owing to its theoretical tractability (exact **mean square errors** available) and aesthetic superiority over the ratio type estimator $\tilde{\lambda}_n$.

A complete **random sample** of lifetimes as assumed above is often unavailable. In reality, lifetime data are often incomplete owing to **staggered entry**, loss to follow-up, or early termination of a study. For simplicity of presentation, we focus on the random **censoring** case for the rest of the entry. Basic references for hazard estimation for other incomplete data such as left-truncated and right-censored data can be found in [26, 58]. The related problem of estimating transition intensities for a two-state Markov Process was explored in [34].

Under the random censorship model, the actual lifetime T_i of an individual may be censored by another random variable C_i. One observes instead (X_i, δ_i), where $X_i = \min(T_i, C_i)$, the minimum of the lifetime and censoring time of the ith individual, and $\delta_i = 1_{\{X_i = T_i\}}$, which is one if the actual lifetime is observed and zero otherwise. We shall assume that the censoring times $C_1, C_2, \ldots, C_n$ have a common distribution function G and that they are independent of the lifetimes $T_1, \ldots, T_n$. Let $(X_{(i)}, \delta_{[i]})$, $i = 1, 2, \ldots, n$, be the ordered sample with respect to X_i's (that is, $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, and $\delta_{[i]}$ is the corresponding censoring indicator of $X_{(i)}$).

Hazard estimators in this situation are ordinarily obtained by smoothing the increments of the **Nelson–Aalen estimator** $\Lambda_n(\cdot)$ for the cumulative hazard function $\Lambda(t)$. Let $N_n(t) = \sum_{i=1}^n 1_{\{X_i \le t, \delta_i = 1\}}$, and $Y_n(t) = \sum_{i=1}^n 1_{\{X_i \ge t\}}$. The Nelson–Aalen estimator $\Lambda_n(\cdot)$, which is instrumental in survival analysis for censored data, is defined as

$$\begin{aligned}\Lambda_n(t) &= \int_0^t \frac{1_{\{Y_n(s) > 0\}}}{Y_n(s)}\, \mathrm{d}N_n(s) \\ &= \sum_{i=1}^n \frac{\delta_{[i]} 1_{\{X_{(i)} \le t\}}}{n - i + 1}\end{aligned} \tag{18}$$

if there are no tied observations. Properties of the random step function $\Lambda_n(t)$ have been studied extensively; see, for example, [1, Section IV.1] for details.

Kernel Estimators

Substituting the Λ_n in (18) into (16) and choosing $W_n(x) = b^{-1} K((t - x)/b)$, for a particular choice of kernel K and bandwidth $b = b_n$, we arrive at the kernel hazard estimator:

$$\begin{aligned}\hat{\lambda}(t) &= \int \frac{1}{b} K\left(\frac{t - x}{b}\right) \Lambda_n(x), \\ &= \sum_{i=1}^n \frac{1}{b} K\left(\frac{t - X_{(i)}}{b}\right) \frac{\delta_{[i]}}{n - i + 1},\end{aligned} \tag{19}$$

if there are no tied observations.

Asymptotic properties on consistency are typically obtained under the following assumptions: (i) the true hazard rate is k-times differentiable for a $k \ge 0$; (ii) the bandwidths satisfy (7); and (iii) the kernel is of order k, defined as:

$$\begin{aligned}&\int K(x)\, \mathrm{d}x = 1, \quad \int K^2(x)\, \mathrm{d}x < \infty, \\ &\int x^j K(x)\, \mathrm{d}x = 0 \text{ for } 1 < j < k, \\ &\int x^k K(x)\, \mathrm{d}x \text{ is finite but nonzero.}\end{aligned} \tag{20}$$

The choice of the bandwidth is of crucial importance and regulates the trade off between the bias and variance of the estimator in (19). A small bandwidth yields a less smooth curve, with smaller bias but larger variance, as compared to a larger bandwidth (see (23) and (24)). Bandwidth choice is particularly crucial for hazard estimation near the right boundary of the data as the variance increases to infinity there. More discussions on bandwidth choice is provided in the next section.

As for the choice of the kernel, smoothness of the kernel determines the smoothness of the corresponding kernel estimate, and the order of the kernel determines the order of the bias (see (23)) and thus the rate of convergence. Often, nonnegative kernels are used in practice, and the Epanechnikov kernel in (6) has certain optimality properties (see [39]).

The kernel hazard estimate is the simplest and thus a widely adopted smooth hazard estimator. It

has been studied extensively in the literature, for example, by Ramlau–Hansen [47, 48], Yandell [65], Tanner and Wong [57], Burke and Horváth [7], Diehl and Stute [11] and Müller and Wang [40].

Spline Estimators

Another commonly adopted smoothing method is the spline method. There are several types of spline methods. The most widely investigated spline method for hazard smoothing is the penalized likelihood approach. Let $\eta(t) = \log \lambda(t)$ be the log hazard function. The log-likelihood function for censored data is

$$\ell(\eta) = \sum_{i=1}^{n} \left\{ \delta_i \eta(X_i) - \int_0^{X_i} e^{\eta} \right\}, \qquad (21)$$

which is unbounded if no shape restriction on η is imposed. A penalty $J(\eta)$, measuring the roughness of η, is therefore incorporated and the penalized likelihood estimate $\hat{\eta}$ of η is the maximizer of the penalized log likelihood

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i \eta(X_i) - \int_0^{X_i} e^{\eta} \right\} - \frac{\alpha}{2} J(\eta), \qquad (22)$$

among all η in a Hilbert space. Here α is a smoothing parameter. Smaller α yields a better fit but a more variable (rough) curve. A typical choice of $J(\eta)$ is $\int [\eta^{(2)}(x)]^2 \, dx$, which leads to a cubic spline with knots at all $X's$. More specifically, $\hat{\eta}$ is two-times continuously differentiable and is a piecewise cubic polynomial between any two consecutive $X's$. The smoothing parameter α plays a similar role to that of the bandwidth b in a kernel estimate. **Cross-validation** is a common way to determine the value of α; see [43, 44] for computational details and [26] for asymptotic results.

In (22), the roughness of $\log \lambda(t)$ is penalized so as to avoid nonnegative constraints on the hazard function. Other forms of penalty functions were proposed in [2, 3, 52]. The penalty function J determines the kind of spline resulted from (22). For example, the penalty $J(\eta) = \int [\lambda'(X)]^2 \, dx$ is employed in [2], and the resulting hazard estimate is a piecewise quadratic spline. Note that this hazard estimate may yield negative values under heavy censoring.

The above spline estimates have knots at each of the observed X values and are called smoothing splines in the literature (see [25, Chapter 2]). Another type of spline method is regression splines or B-Splines, which adopt a fixed number of knots and basis functions; see [35, 50] for details and ways to select the number and location of knots. A hazard function estimate with flexible tails, called HEFT, is proposed in [35] by estimating the log-hazard function using cubic splines.

Other Hazard Rate Estimators

The ratio type hazard estimator in (17), also due to Watson–Leadbetter, has been extended to censored data as well and was studied by Blum and Susarla [5], Földes, Rejtö and Winter [16] and Lo, Mack, and Wang [37].

Hjort [30] advocated the use of semiparametric approaches to estimate hazard rates. The approach is to start with a possibly crude parametric estimate and to improve it via some nonparametric procedures. The motivation is to reduce the bias of a parametric estimate via nonparametric correction locally, and yet to arrive at an estimate that is less variable than a fully nonparametric one.

For reviews of earlier results on hazard rate estimation, see [18, 45], and [54] for uncensored data.

More on Kernel Hazard Estimators for Continuously Observed Data

The rate of convergence of the kernel hazard estimate (19) depends on the order of the kernel, the bandwidth, and the differentiability of the hazard function. Typically, the order k of the kernel is chosen to be an even number with $k = 2$ being the standard choice. The resulting bias and variance are respectively

$$\text{bias}(\hat{\lambda}(t)) = b^k [\lambda^{(k)}(t) B_k + o(1)], \qquad (23)$$

$$\text{var}(\hat{\lambda}(t)) = \frac{1}{nb} \left\{ \frac{\lambda(t)}{[1 - F(t)][1 - G(t)]} V + o(1) \right\}, \qquad (24)$$

where $B_k = (-1)^k / k! \int x^k K(x) \, dx$ and V is as in (5).

The influence of the bandwidth b and the trade off between the bias and variance is seen from (23) and (24). The optimal rate for the mean squared error (MSE) of $\hat{\lambda}(t)$ is attained when the $(bias)^2$ and variance are of the same order. This results in an optimal MSE rate of convergence of $n^{2k/(2k+1)}$, which is $n^{4/5}$ for the standard choice of $k = 2$. This rate is

slower than the usual parametric rate of n regardless of the order of k. For the asymptotic distribution, we further assume that $d = \lim_{n\to\infty} nb^{2k+1}$ exists for some $0 \le d < \infty$. Then

$$(nb)^{1/2}(\hat{\lambda}(t) - \lambda(t)) \xrightarrow{\mathcal{D}} N \times \left(d^{1/2}\lambda^{(k)}(x)B_k, \frac{\lambda(t)}{[1 - F(t)][1 - G(t)]}V\right). \quad (25)$$

Extensions to the estimation of derivatives of hazard functions have been considered as well [40]. These essentially involve a change in the kernel. Derivatives are of interest to detect rapid changes in hazard rates or for data based bandwidth choices, as the optimal bandwidths in (26) or (27) depend on the derivatives of the hazard rates. Again, the order k of the kernel affects the convergence rate and also asymptotic constants.

Bandwidth Choice

The bandwidth for a kernel hazard estimate can be fixed at all points (global bandwidth b) or can vary for different points (local bandwidth $b(t)$). Usually, a global bandwidth is employed for a smooth density or regression estimate owing to its simplicity. However, for the hazard estimation situation discussed here, there are compelling reasons to adopt local rather than global bandwidth choices. According to (24), the variance of the kernel estimate $\hat{\lambda}(t)$ eplodes to infinity as t approaches the right boundary of the data. Thus, the variance tends to dominate the bias in the right tail and this needs to be compensated for by a larger bandwidth.

The optimal local bandwidth of $\hat{\lambda}(t)$, which minimizes the leading term of $MSE(\hat{\lambda}(t))$ is

$$b^*(t) = n^{-1/(2k+1)} \times \left\{\frac{1}{2k}\frac{\lambda(t)}{[1 - F(t)][1 - G(t)]}\frac{V}{[\lambda^{(k)}(t)B_k]^2}\right\}^{1/(2k+1)} \quad (26)$$

To find the optimal global bandwidth, we have to restrict the range of t to a compact interval $[0, \tau]$ with $F(\tau) < 1$ and $G(\tau) < 1$. The global optimal bandwidth which minimizes the leading term of $MISE(\hat{\lambda}) = E\int_0^\tau [\hat{\lambda}(x) - \lambda(x)]^2\,\mathrm{d}x$ is

$$b_{\text{opt}} = n^{-1/(2k+1)}\left\{\frac{1}{2k}\int_0^\tau \frac{\lambda(x)}{[1 - F(x)][1 - G(x)]}\,\mathrm{d}x \times \frac{V}{B_k^2\int_0^\tau [\lambda^{(k)}(y)]^2\,\mathrm{d}y}\right\}^{1/(2k+1)}. \quad (27)$$

Note that both the local and global optimal bandwidths in (26) and (27) involve unknown quantities. In practice, one has to find alternatives. There is an extensive literature on bandwidth selection and "cross-validation" and "plug-in" techniques are popular; see [40, 41, 46] for details. A bootstrap method to select the global bandwidth has been advocated in [20] as an alternative. In addition to the local bandwidth choice in (26), which adopts different bandwidths at different time point t, choosing bandwidths as the distance of t to its kth nearest neighbor among the remaining uncensored observations is a convenient way to adapt to the data by allowing for varying degrees of smoothing; see [17, 56, 57] for detailed descriptions. Other data-adaptive local or global bandwidth choices for hazard estimates can be derived analogously to the density estimation case as discussed in [53, Section 3.4] and [59, Chapter 3].

Boundary Effects

We close this section with a cautionary remark that the kernel smoothing method needs to be employed very carefully near the boundary as there is a bias problem in such regions, usually referred to in the literature as boundary effects. Boundary effects may be attributed to the fact that the support of the kernel exceeds the available range of data and are not unique to hazard estimates.

An unmodified kernel estimate is unreliable in the boundary region, which is the region within one bandwidth of the largest or smallest observations. To remedy the boundary effects, different kernels, referred to as "boundary kernels" can be used within the boundary region. As a consequence, varying kernels are employed at each location t and the bandwidths are affected accordingly. The resulting kernel estimate with varying kernels and varying local bandwidths takes the form

$$\hat{\lambda}(t) = \int \frac{1}{b(t)}K_t\left(\frac{t - x}{b(t)}\right)\mathrm{d}\Lambda_n(x), \quad (28)$$

where both the bandwidth $b = b(t)$ as well as the kernel $K = K_t$ depend on the point t. Details for the choices of the kernel K_t and bandwidths $b(t)$ can be found in [41].

Simulation Comparison of Hazard Estimators and Software

A very informative and extensive simulation study was carried out in [29] to compare the aforementioned kernel-based hazard estimators with various local and global bandwidth choices and boundary corrections, the kernel-based hazard estimators in [17] with varying bandwidth methods based on kth nearest-neighbor, and the spline-based estimators in [35]. The results indicated advantages of using HADES, the aforementioned local optimal bandwidth choice and boundary correction in [41]. There is significant improvement (over 50% on the average) in mean square error over the global bandwidth choice if a local optimal bandwidth is employed. Boundary corrections will lend additional efficiency. The locally optimal bandwidth estimators in [41] with only left boundary correction also outperformed two publicly available procedures, the spline estimator in [35] and the nearest-neighbor estimator in [17]. The latter is based on the procedures in [56, 57].

A library of Fortran and S-Plus programs for the HADES estimator in [41] and for the nearest-neighbor estimator in [17] is available under a package called "muhaz" at the website of the authors of [29] `:http://odin.mdacc.tmc.edu/anonftp/` To get the S-code follow the link: `ftp://odin.mdacc.tmc.edu/pub/S/muhaz.tar.gz` The corresponding R program for **muhaz** is also publicly available at: `cran.r-project.org/doc/packages/muhaz`

The S-plus code of the spline estimator in [35] called, **HEFT** is publicly available from the StatLib software library.

Hazard Regression

Estimating a Baseline Hazard Function

So far, we discussed hazard smoothing for a homogeneous population. Often the risk of an individual varies according to the values of some **covariates**. Thus, the hazard function of an individual with covariate $Z \in \Re^d$ is $\lambda(t, Z)$ and regression techniques are required. A semiparametric approach with a regression parameter β and a nonparametric baseline hazard function $\lambda_0(t)$ is often adopted. Examples include Cox's **proportional hazards regression model**, where $\lambda(t, Z) = \lambda_0(t) \exp(\beta^T Z)$, and the **accelerated failure-time model**, where $\lambda(t, Z) = \lambda_0(\exp(\beta^T Z)t) \cdot \exp(\beta^T Z)$.

A smooth estimate of the baseline hazard is preferable and often necessary to obtain consistent estimates of $\lambda(t, Z)$. Anderson and Senthilselvan [2] applied the penalized maximum likelihood approach, and Gray [23] and Wells [64] applied the kernel method to estimate the baseline hazard function in Cox's proportional hazard model. Andersen et al. [1, Section VII.2.5] give several examples of estimated baseline hazard functions.

The Cox proportional model has been extended in [9] to allow covariate dependent baseline hazard function. The model is $\lambda(t, Z) = \lambda_0(t, X_t) \exp[\beta^T Z_t]$, where X_t and Z_t are predictable covariate processes or covariate vectors. Another type of extension is to employ, as in [61], an unknown link function in the proportional model, where $\lambda(t, Z) = \lambda_0(t) g(\beta^T Z)$ with g completely unknown and estimated via local **partial likelihood** method. Etezadi-Amoli and Ciampi [13] also investigated another extension of Cox's proportional hazards and accelerated failure-time models of the form: $\lambda(t, Z) = \lambda_0(g_1(\alpha^T Z)t) g_2(\beta^T Z)\cdot$, where $\lambda_0(t)$ denotes the baseline hazard function, which is estimated by the regression spline method.

Generalized Additive Proportional Hazards Model

Another type of **proportional hazards** model allows an arbitrary covariate effect of the form:

$$\lambda(t, Z) = \lambda_0(t) \exp[g(Z)], \tag{29}$$

where g is an unspecified smooth function of Z. LeBlanc and Crowley [36] use the CART (Classification and Regression Trees) algorithm to estimate the relative risk g (*see* **Tree-structured Statistical Methods**), Gentleman and Crowley [19] and Fan, Gijbels, and King [15] use local full or partial likelihood methods to estimate g. Although this is the most general proportional hazards model, it is difficult to estimate $g(Z)$ when the covariate Z is of high

dimension, say $d \geq 3$. An extremely large sample size would be needed. This is called the "curse of dimensionality". Dimension reduction models and methods are thus called for. Among these, the additive regression model is a promising alternative to (29).

Under the additional assumption that g is additive in (29), that is, $g(z) = \sum_{i=1}^{d} g_i(z_i)$, Hastie and Tibshirani [28] and O'Sullivan [43, 44] use smoothing splines to estimate g (*see* **Generalized Additive Model**). Sleeper and Harrington [55] use B-splines, and Gray [24] uses penalized splines with fixed knots to estimate g and incorporate time-varying coefficients. Apart from the minor differences in the various spline methods, all the aforementioned methods adopt the partial likelihood approach with a penalty for each g_i to be estimated.

Let (X_i, Z_i, δ_i), $i = 1, \ldots, n$ denote the observed data and $Y_1 < \cdots < Y_k$ denote the k distinct failure times with d_i failures at time Y_i. The penalized log partial likelihood with smoothing parameters $\alpha_1, \ldots, \alpha_d$ is:

$$\ell(g_1, \ldots, g_d) = \sum_{i=1}^{k} \delta_i \left\{ \sum_{j \in D_i} g(Z_j) - d_i \log \left(\sum_{j \in R_i} e^{g(Z_j)} \right) \right\} - \frac{1}{2} \sum_{i=1}^{d} \alpha_i \int \left[g_i^{(2)}(t) \right]^2 dt, \quad (30)$$

where D_i is the set of indices of the failures at observed failure time X_i, and R_i is the set of indices of individuals at risk at time X_i. Minimizing $\ell(g_1, \ldots, g_d)$ then yields the smoothing spline estimates $(\hat{g}_1, \ldots, \hat{g}_d)$. Calculations of the estimates can be very time-consuming; see [27, Section 8.3] for computational issues.

Nonparametric Hazard Regression

A completely nonparametric approach to estimate $\lambda(t, Z)$ is desirable sometimes. Kooperberg, Stone, and Truong [35] used **loglinear regression** splines and their tensor products to estimate $\log \lambda(t, Z)$. Gu [26] considered the penalized likelihood approach. Doss and Li [12] used linear polynomials in Z to fit $\lambda(t, Z)$ locally in a neighborhood of Z. Martingale convergence theory for **counting processes** was used to derive the weak convergence of their hazard estimate.

For continuously observed lifetimes, one can obtain a hazard regression estimate for $\lambda(t, Z)$ by smoothing the increments of any cumulative hazard estimate $\Lambda(t, Z)$. Such a cumulative hazard estimate can be found in [10] and is further studied by McKeague and Utikal [38]. Again, any of the smoothing methods discussed so far can be extended to a nonparametric hazard regression estimate.

Note that by grouping the data along the time axis and the covariate axis, one can also apply any nonparametric regression smoother to grouped data. Gray [24] illustrates this grouping method through a local linear polynomial smoother and kernel regression.

Lexis Diagram

An interesting application of nonparametric hazard regression is the **Lexis diagram** in which individual lifelines are represented as line segments between (time at birth, 0) and (time, age) of death. Here time at birth can be used in a broad sense, that is, as the onset time of a disease. If mortality of individuals varies according to time of birth, a covariate Z based on an individual's calendar time of birth can be incorporated to model individual risks at age t represented by $\lambda(t, Z)$. Keiding [33] suggests using bivariate versions of nonparametric smoothing methods, as discussed above, to estimate $\lambda(t, Z)$, provided that the influence of Z on the hazard function is continuous in Z.

References

[1] Andersen, P.K., Borgan, Ø, Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Anderson, J. & Senthilselvan, A. (1980). Smooth estimates for the hazard function, *Journal of the Royal Statistical Society B* **42**, 322–327.

[3] Antoniadis, A. & Grégoire, G. (1990). Penalized likelihood estimation for rates with censored survival data, *Scandinavian Journal of Statistics* **17**, 43–63.

[4] Bloomfield, D. & Haberman, S. (1987). Graduation: Some experiments with kernel methods, *Journal of the Institute of Actuaries* **114**, 339–369.

[5] Blum, J.R. & Susarla, V. (1980). Maximal deviation theory of density and failure rate function estimates based on censored data, in *Multivariate Analysis*, Vol. V, P.R. Krishnaiah ed. North Holland, New York, pp. 213–222.

[6] Borgan, Ø. (1979). On the theory of moving average graduation, *Scandinavian Actuarial Journal* **1979**, 83–105.

[7] Burke, M.D. & Horváth, L. (1984). Density and failure rate estimation in a competing risks model, *Sankhya-The Indian Journal of Statistics Series A* **46**, 135–154.

[8] Copas, J. & Haberman, S. (1983). Nonparametric graduation using kernel methods, *Journal of the Institute of Actuaries* **110**, 135–156.

[9] Dabrowska, D. (1997). Smoothed Cox regression, *Annals of Statistics* **25**, 1510–1540.

[10] Dabrowska, D.M. (1987). Non-parametric regression with censored survival time data, *Scandinavian Journal of Statistics* **14**, 181–197.

[11] Diehl, S. & Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring, *Journal of Multivariate Analysis* **25**, 299–310.

[12] Doss, H. & Li, G. (1995). An approach to nonparametric regression for life history data using local linear fitting, *Annals of Statistics* **23**, 787–823.

[13] Etezadi-Amoli, J. & Ciampi, A. (1987). Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function, *Biometrics* **43**, 181–192.

[14] Fan, J. & Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

[15] Fan, J., Gijbels, I. & King, M. (1997). Local likelihood and local partial likelihood in hazard regression, *Annals of Statistics* **25**, 1661–2690.

[16] Földes, A., Rejtö, L. & Winter, B.B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data. II: Estimation of density and failure rate, *Periodica Mathematica Hungarica* **12**, 15–29.

[17] Gefeller, O. & Dette, H. (1992). Nearest neighbor kernel estimation of the hazard function from censored data, *Journal of Statistical and Computational Simulations* **43**, 93–101.

[18] Gefeller, O. & Michels, P. (1992). A review on smoothing methods for the estimation of the hazard rate based on kernel functions, in *Computational Statistics*, Y. Dodge & J. Whittaker eds. Physica-Verlaz, Switzerland, pp. 459–464.

[19] Gentleman, R. & Crowley, J. (1991). Local full likelihood estimation for the proportional hazard model, *Biometrics* **47**, 1283–1296.

[20] González-Manteiga, W., Cao, R. & Marron, J.S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation, *Journal of The American Statistical Association* **91**, 1130–1140.

[21] Gram, J.P. (1879). *Om Rækkeudviklinger, bestemte ved Hjælp af de mindste Kvadraters Methode*. A.F. H Øst & SØn, Copenhagen.

[22] Gram, J.P. (1883). Ueber Entwickelung reeller Functionen in Reihen mittelst der Methode der Kleinsten Quadrate, *Journal of Mathematics* **94**, 41–73.

[23] Gray, R. (1990). Some diagnostic methods for Cox regression models through hazard smoothing, *Biometrics* **46**, 93–102.

[24] Gray, R. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis, *Journal of the American Statistical Association* **87**, 942–951.

[25] Green, P.J. & Silverman, B.W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

[26] Gu, C. (1996). Penalized likelihood hazard estimation: A general procedure, *Statistica Sinica* **6**, 861–876.

[27] Hastie, T. & Tibshirani, R. (1990a). *Generalized Additive Models*. Chapman and Hall, London.

[28] Hastie, T. & Tibshirani, R. (1990b). Exploring the nature of covariate effects in the proportional hazards model, *Biometrics* **46**, 1005–1016.

[29] Hess, K.R., Serachitopol, D.M. & Brown, B.W. (1999). Hazard function estimators: a simulation study, *Statistics in Medicine* **18**, 3075–3088.

[30] Hjort, N. (1991). Semiparametric estimation of parametric hazard rates, in *Survival Analysis: State of the Art*, J.P. Klein & P.K. Goel eds. Kluwer, Dordrecht, pp. 211–236.

[31] Hoem, J. (1983). The reticent trio: Some little-known discoveries in life insurance mathematics by L.H.F. Oppermann, T.N. Thiele, and J.P. Gram, *International Statistical Review* **51**, 213–221.

[32] Hoem, J. (1984). A contribution to the statistical theory of linear graduation, *Insurance, Mathematics, and Economics* **3**, 1–17.

[33] Keiding, N. (1990). Statistical inference in the Lexis diagram, *Philosophical Transactions of the Royal Society of London A* **332**, 487–509.

[34] Keiding, N. & Andersen, P.K. (1989). Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process, *Applied Statistics* **38**, 319–329.

[35] Kooperberg, C., Stone, C.J. & Truong, Y.K. (1995). Hazard regression, *Journal of the American Statistical Association* **90**, 78–94.

[36] LeBlanc, M. & Crowley, J. (1992). Relative risk trees for censored survival data, *Biometrics* **48**, 411–425.

[37] Lo, S.-H., Mack, Y.P. & Wang, J.L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator, *Probability Theory and Related Fields* **80**, 461–473.

[38] McKeague, I.W. & Utikal, K.J. (1990). Inference for a nonlinear counting process regression model, *Annals of Statistics* **18**, 1172–1187.

[39] Müller, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer, New York.

[40] Müller, H.G. & Wang, J.L. (1990). Locally adaptive hazard smoothing, *Probability Theory and Related Fields* **85**, 523–538.

[41] Müller, H.G. & Wang, J.L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths, *Biometrics* **50**, 61–76.

[42] Müller, H.G., Wang, J.L. & Capra, W.B. (1997). From lifetables to hazard rates: The transformation approach, *Biometrika* **84**, 881–892.

[43] O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation, *SIAM Journal of Science and Statistical Computation* **9**, 531–542.

[44] O'Sullivan, F. (1988a). Fast computation of fully automated log-density and log-hazard estimators, *SIAM Journal of Science and Statistical Computation* **9**, 363–379.

[45] Padgett, W.J. (1988b). Nonparametric estimation of density and hazard rate functions when samples are censored, in *Handbook of Statistics*, Vol. 7, P.R. Krishnaiah & C.R. Rao eds. North-Holland, New York, pp. 313–331.

[46] Patil, P.N. (1993). Bandwidth choice for nonparametric hazard rate estimation, *Journal of Statistical Planning and Inference* **35**, 15–30.

[47] Ramlau-Hansen, H. (1983a). Smoothing counting process intensities by means of kernel functions, *Annals of Statistics* **11**, 453–466.

[48] Ramlau-Hansen, H. (1983b). The choice of a kernel function in the graduation of counting process intensities, *Scandinavian Actuarial Journal* **10**, 165–182.

[49] Rice, J. & Rosenblatt, M. (1976). Estimation of the log survivor function and hazard function, *Sankhya-The Indian Journal of Statistics Series A* **38**, 60–78.

[50] Rosenberg, P.S. (1995). Hazard function estimation using B-splines, *Biometrics* **51**, 874–887.

[51] Seal, H.L. (1981). Graduation by piecewise cubic polynomials: a historical review. Blätter, *Deutsche Gesellshaft für Versicherungsmathematik* **15**, 89–114.

[52] Senthilselvan, A. (1987). Penalized likelihood estimation of hazard and intensity functions, *Journal of the Royal Statistical Society B* **49**, 170–174.

[53] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[54] Singpurwalla, N.D. & Wong, M.-Y. (1983). Estimation of the failure rate – a survey of nonparametric methods. Part I: Non-Bayesian methods, *Communications in Statistics-Theory and Methods* **12**, 559–588.

[55] Sleeper, L.A. & Harrington, D.P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease, *Journal of the American Statistical Association* **85**, 941–949.

[56] Tanner, M.A. (1983). A note on the variable kernel estimator of the hazard function from randomly censored data, *Annals of Statistics* **11**, 994–998.

[57] Tanner, M.A. & Wong, W.H. (1983). The estimation of the hazard function from randomly censored data by the kernel method, *Annals of Statistics* **11**, 989–993.

[58] Uzunogullari, U. & Wang, J.-L. (1992). A comparison of hazard rate estimators for left truncated and right censored data, *Biometrika* **79**, 297–310.

[59] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

[60] Wang, J.L., Müller, H.G. & Capra, W.B. (1998). Analysis of oldest-old mortality: Lifetables revisited, *Annals of Statistics* **26**, 126–163.

[61] Wang, W. (2001). *Proportional hazard regression model with unknown link function and applications to longitudinal time-to-event data*. Ph.D. Thesis, University of California, Davis.

[62] Watson, G.S. & Leadbetter, M.R. (1964). Hazard analysis. I, *Biometrika* **51**, 175–184.

[63] Watson, G.S. & Leadbetter, M.R. (1964). Hazard analysis. II, *Sankhya-The Indian Journal of Statistics Series A* **26**, 101–116.

[64] Wells, M.T. (1994). Nonparametric kernel estimation in counting processes with explanatory variables, *Biometrika* **81**, 759–801.

[65] Yandell, B.S. (1983). Nonparametric inference for rates with censored survival data, *Annals of Statistics* **11**, 1119–1135.

Further Reading

Gray, R. (1996). Hazard regression using ordinary nonparametric regression smoothers, *Journal of Computational and Graphical Statistics* **5**, 190–207.

JANE-LING WANG

Smoothing Methods in Epidemiology

Introduction

Most observational data in the health sciences are generated by poorly understood and largely nonrandom mechanisms of exposure assignment, subject selection, and measurement error, and are analyzed to estimate causal structures that are latent under these mechanisms. In contrast, most statistical methods presume the data are generated by **identifiable** random processes (i.e. distributions known up to a parameter vector of dimension no greater than the data), and that the structures of interest are estimable under the assumed distributions. Until recently, this chasm between observational reality and statistical theory was formally addressed in no general statistics textbook. As a result, statistical methods for identifiable random processes are used routinely for inferences about unidentified structure.

The validity of these inferences depends on latent independencies (often left implicit or cast as ignorability assumptions) that identify the parameters of interest. In **observational studies** of causation, however, large departures from identifying assumptions are a distinct possibility. Most reports deal with these possibilities in a narrative fashion based on (often incorrect) intuitions about the biasing effects of departures, rather than on quantitative reasoning. Quantification of departure effects, when done at all, is usually by **sensitivity analyses** (e.g. [50], Chapter 19). By embedding assumptions in a larger parametric framework, these analyses show how departures would affect summary statistics. They do not, however, quantify the net uncertainty one should have in light of uncertainties about the departures. Furthermore, because the departures are not identified, the sensitivity results have no inferential interpretation without reference to **prior distributions** for the departures [23].

Several approaches have been developed for inference about parameters (such as causal effects) that are not identified given the uncertainties about biases in observational data. Examples include methods for inference under nonignorable data-generating mechanisms [15, 42], **Bayesian methods** for nonexperimental data analysis (e.g. [12, 18, 27, 33, 39)], and **Monte-Carlo** sensitivity analyses [29, 38, 48]. In some situations, these latent-structure analyses may be essential for coherent inferences about causation in the absence of informative experiments. But between initial data description and these structural analyses, one can mark out a preliminary smoothing or filtering step concerned with estimating expectations for the observed data, rather than with making inferences about deeper and latent structure [20].

In the health sciences, this filtering step is usually skipped or dealt with by fairly primitive means (such as adding $1/2$ to each cell of a table). The present article outlines alternatives in which saturated **hierarchical models** for random variation are used to smooth or filter out extraneous noise before structural exploration (*see* **Generalized Linear Model**). It also provides a brief overview and illustration of basic methods that can be applied using ordinary regression software. For further information on those methods, see **penalized maximum likelihood**; for other smoothing methods, see **density estimation**, **generalized additive model**, **geographical analysis**, Kalman filtering and smoothing, **nonparametric regression**, **smoothing hazard rates**, and **spline function**. At certain points, an analogy will be drawn between smoothing methods and methods for imputing missing data: Both are intended to "clean up" the data before developing inferences about the population structures of interest.

Example Data

The observed counts in Table 1 are taken from an analysis of **bias** in 14 **case–control studies** of exposure to residential magnetic fields and childhood leukemia [29]. On the basis of the earlier work [33], a cutpoint of 3 milligauss (3 mG, equal to 0.3 microtesla) was used for all the studies except the United Kingdom Childhood Cancer (UKCC) study. The latter published only categorizations at 1,2, and 4 mG, hence its estimate compares >4 mG versus ≤ 2 mG; this estimate appears consistent with the other studies, however, and reanalysis of the other studies using a 4 mG cutpoint changed the pooled estimate by only 5% [32]. A study by Green et al. based analyses on quartile categories, resulting in upper cutpoints of only 1.3 to 1.5 mG, and is excluded here because the use of such low cutpoints strongly influenced estimates from earlier studies [32]; it did however report

Table 1 Summary data from 14 case–control studies of magnetic fields and childhood leukemia, smoothed counts for numbers >3 mG (milligauss), and odds ratios. The smoother preserves all totals so totals are not repeated

First author	Country	No. cases		No. controls		Odds ratio (95% limits)
		>3 mG	Total	>3 mG	Total	
[Coghill 8]	England	1	56	0	56	∞
Smoothed		0.72		0.28		2.63 (0.52, 13.4)
[Dockerty 11]	NZ	3	87	0	82	∞
Smoothed		2.33		0.67		3.31 (0.74, 14.6)
[Feychting 14]	Sweden[b]	6	38	22	554	4.53 (1.72, 12.0)
Smoothed		4.79		23.21		3.30 (1.30, 8.34)
[Kabuto 35]	Japan	11	312	13	603	1.66 (0.73, 3.75)
Smoothed		11.39		12.61		1.78 (0.84, 3.73)
[Linet 41]	US[a]	42	638	28	620	1.49 (0.91, 2.44)
Smoothed		42.22		27.78		1.51 (0.94, 2.43)
[London 43]	US[a]	17	162	10	143	1.56 (0.69, 3.53)
Smoothed		17.13		9.87		1.59 (0.76, 3.36)
[McBride 44]	Canada[a]	14	297	11	329	1.43 (0.64, 3.20)
Smoothed		14.25		10.75		1.49 (0.71, 3.12)
[Michaelis 45]	Germany	6	176	6	414	2.40 (0.76, 7.55)
Smoothed		5.92		6.08		2.35 (0.89, 6.15)
[Olsen 46]	Denmark[b]	3	833	3	1666	2.00 (0.40, 9.95)
Smoothed		2.83		3.17		1.80 (0.52, 6.20)
[Savitz 52]	US[a]	3	36	5	198	3.51 (0.80, 15.4)
Smoothed		2.40		5.60		2.45 (0.76, 7.90)
[Tomenius 56]	Sweden	3	153	9	698	1.53 (0.41, 5.72)
Smoothed		3.37		8.63		1.80 (0.64, 5.08)
[Tynes 57]	Norway[b]	0	148	31	2004	0
Smoothed		1.49		29.51		0.68 (0.19, 2.40)
UKCC (1999)	UK[c]	5	1057	3	1053	1.66 (0.40, 6.98)
Smoothed		5.26		2.74		1.92 (0.62, 5.96)
[Verkasalo 58]	Finland[b]	1	32	5	320	2.03 (0.23, 18.0)
Smoothed		0.88		5.12		1.74 (0.40, 7.57)
Totals[d]		115	4025	146	8740	1.69 (1.28, 2.23)

[a] 120v 60 Hz systems, $V = 1$ (others are 220v 50 Hz, $V = 0$).
[b] Calculated fields, $D = 0$ (others are direct measurement, $D = 1$).
[c] Comparison of >4 mG versus ≤2 mG, excluding 16 cases and 20 controls at 2–4 mG.
[d] Final entry is MLE of common odds ratio (lower deviance $P = 0.0001$, homogeneity $P = 0.24$), which is unchanged by smoother to three digits past decimal point.

positive associations upon contrasting the top and bottom quartiles. Finally, a study by Schüz et al. [53] with only three highly exposed cases was excluded because of evidence of severe upward sparse-data bias [31] in the reported estimates (odds ratios of 5 to 11), and insufficient reporting of data to allow further evaluation.

Leukemia is a very rare disease and the usual justifications for interpreting the observed **odds ratios** as rate-ratio estimates apply [50, Chapter 7]. The odds ratios are very consistent across studies (with homogeneity $P = 0.24$). The pooled **maximum likelihood** estimate (MLE) of a common odds ratio is 1.69 with 95% confidence limits of 1.28, 2.23, nearly identical to the **Mantel–Haenszel** (MH) odds ratio of 1.68 with 95% confidence limits of 1.27, 2.22. The association is not explained or modified by any known study characteristic or feature of the available data, such as information source (e.g. measurement method), location, or current type (120v 60 Hz versus 220v 50 Hz). **Covariate** adjustment had almost no impact on the estimates. Results are unchanged using finer categories (e.g. contrasting >3 mG vs. ≤ 1 mG) or continuous field measurements, and there is no evidence of publication bias [32] (*see* **Meta-analysis of Clinical Trials**).

Nonetheless, taking the statistics in Table 1 as **unbiased** for the field effect is equivalent to assuming that each study reported an experiment in which children randomized to known residential field levels (*see* **Randomization**), were never switched from their initial assignment, and were followed until either leukemia or selection as a control or random censoring (*see* **Censored Data**) occurred. These assumptions are unacceptable, and so analyses of the impact of departures (i.e. bias analyses) were carried out [27, 29]. However, some of the methods used for these analyses require that all input counts be nonzero; hence, preliminary smoothing of the counts was done to eliminate the zeros in the observed table. The goal of this step is to smooth out obvious "holes" in the data without disturbing data summaries or patterns.

Some Basic Theory for Smoothing

Notation and Definitions

Let Y be the portion of the data to be treated as having a random component, that is, Y is a **random variable** whose range of possible values is the sample space. Let X be the portion to be treated as fixed, a set of known conditions upon which we condition our **expectation** for Y. In most of the notation that follows, X conditioning will be left implicit; in particular, $\mathrm{E} \equiv \mathrm{E}(Y|X)$ will denote the regression of Y on X, whose possible values define an expectation space Ω. This setup includes multivariate outcomes by making Y an N-row array, and further extends to random regressors by treating those regressors as part of the Y array and including only known fixed regressors in X. The present development, however, will be limited to univariate Y and fixed regressors.

Defining the residual vector as $\varepsilon \equiv Y - \mathrm{E}$, we have $Y = \mathrm{E} + \varepsilon$, where $\mathrm{E}(\varepsilon|\mathrm{X}) = 0$, even if Y has discrete components. Typically, $Y = (Y_1, \ldots, Y_N)'$ is a random vector of outcomes observed on N subjects or groups indexed by $i = 1, \ldots, N$, Ω is a subset of R^N, X is an N-by-J observed-covariate matrix with rows X_i corresponding to the Y_i, and X may include a constant. For the case-control data in Table 1, Y will be the numbers exposed above 3 mG within the $N = 2(14) = 28$ groups defined by disease status and study, X will contain design variables based on these groups, and E will be the expected exposed counts for the groups, given X and the group totals.

Parsimony, Information Loss, and Distortion

Most conventional modeling has the goal of estimating some **parsimonious** summary β for E, usually in the form of the parameter vector in a model $M_{\mathrm{E}}(\beta)$ for the regression E, coupled with a model $p(\varepsilon; \upsilon)$ for the distribution of ε, where β and υ have much lower dimension than E and may overlap. A paradigmatic example is normal **linear regression** in which $\mathrm{E} = M_{\mathrm{E}}(\beta) = X\beta$ and $p(\varepsilon; \upsilon) = \{2\pi \det(\upsilon) \exp(\varepsilon' \upsilon^{-1} \varepsilon)\}^{-1/2}$, with υ a simple **covariance matrix** for ε, for example, $\upsilon = \sigma^2 \mathrm{I}$ with σ^2 functionally independent of β. Another example is binomial **logistic regression** in which Y contains counts and hence ε is discrete (as in the example), and in which $\upsilon = \beta$ because the variance is a fixed function of the mean.

Parsimony also arises from **estimation** issues. Nothing in the above setup requires that X has independent columns or that $J < N$. Nonetheless, classical estimation procedures require independent X columns, and may exhibit many problems unless the number of X columns J is much less than the number of observations N. These procedures thus require parsimonious choice of X columns, otherwise known as **variable selection**. That selection should be heavily guided by contextual theory, such as causal ordering, but unfortunately is often left to mechanical testing procedures that distort significance levels and coverage rates of conventional tests and confidence intervals (*see* **Level of a Test**).

For the present purposes, parsimony gives rise to information loss. Indeed, conventional model-based estimates $\tilde{\mathrm{E}}$ of E (more often written $\hat{\mathrm{Y}}$) must lie within the final model manifold (i.e. the image of the β-space traced by $M_{\mathrm{E}}(\beta)$ in Ω), which has dimension no greater than J and which cannot capture any pattern not visible in this manifold. For example, if X contains only a single term for a treatment, the resulting estimates will be unable to capture (say) three-dimensional patterns in the relation of that treatment to Y. Thus, parsimony in modeling E (limiting the number of parameters) entails distortion along at least some dimensions of E, and thus seems a questionable goal in an exploratory or smoothing context.

Filtering

Consider instead the rather different goal of "filtering out" ε from Y to get and estimate $\tilde{\mathrm{E}}$ of E, without the

objective of fitting a particular structure $M_E(\beta)$ to E. We might wish to do this because we regard ε and E as representing the components of Y that are purely random and purely systematic, with the E component determined both by biases and by effects of interest. Because it is free of **random error**, E is preferable to Y for making inferences about the structure of systematic variation in Y. This does *not* mean that E is sufficient for (say) causal inference (*see* **Causation**), because causal models are unidentified (latent) structures if there is no identifiable distribution for **confounding** effects (such as a randomization distribution) [3, 19, 47, 49, 51]. Nonetheless, E would admit more straightforward modeling of latent structures than would the original Y. The idea here is akin to imputation of missing values with a goal of providing a completed data set useful to any subsequent analyst, regardless of that analyst's modeling objective [42].

To express this idea in a Bayesian formalism, suppose Y and hence ε is uninformative about any latent structure once E is known, that is, for any structure of interest $M_E(\beta)$, $p(\beta|E, Y) = p(\beta|E, \varepsilon) = p(\beta|E)$. For an estimator $\tilde{E}$ of E to be capable of conveying information in E about *any* possible structure, it cannot be constrained to lie in a subspace of Ω. Thus, the goal of filtering is the estimation of E under a saturated model, that is, a model whose range spans Ω; such a model must have at least N functionally independent parameters (*see* **Generalized Linear Model**). This sort of filtering shares few parsimony considerations with conventional modeling. For example, because the model is saturated, there need be no parsimony criterion in model selection (e.g. parsimony is not an issue in choosing columns of X, although some covariates might be deemed redundant or uninformative for E) (*see* **Model, Choice of**). In a similar fashion, concerns about mismeasurement need not arise; for example, even without validation data, a mismeasured covariate may still provide much useful information for estimating E. Again, parallel considerations arise when including covariates in missing-data analyses [42]: a mismeasured covariate may still be useful for filling in missing values of other covariates.

"Noise removal" is a more apt term than "smoothing" for this preliminary filtering stage, for in applications like image processing and threshold detection, one wishes to sharpen rather than smooth at real edges or thresholds. An ideal method would remove as much of the noise as possible while giving back all potentially informative patterns in E ("signals" in the Y vector); that is, it would clean out the noise without removing *any* potentially important information, whether "statistically significant" or not. Again we have a shift of emphasis from parsimony criteria (e.g. "include only if significant at level α") toward more generous inclusion of model terms. This shift is needed because filtering is only an intermediate step between data description and inferential modeling (which may focus on an entirely latent parameter); it is an attempt to clean out noise before making the contextual interpretations that inference represents. The patterns in the resulting estimates $\tilde{E}$ are the information input to the inferential or structural modeling stage, and thus need to be preserved; at that later stage some patterns may be clear and others may be rejected as too imprecisely estimated to be of use.

To summarize: Much as a museum will clean and restore art for public display, the goal of filtering and imputation is to clean and restore data to facilitate interpretation by a broad audience, not to impose an interpretation (structural model) $M_E(\beta)$ on the data. This idea can be especially important when structural inference is likely to be controversial, as in the magnetic-field/leukemia example.

Filtering by Shrinkage

Fitting a saturated model (*see* **Generalized Linear Model**) by conventional "unbiased" methods, such as **least squares** or maximum likelihood (ML) would only give smoothed values $\tilde{E}$ equal to the observed Y, and so achieve nothing. Choosing methods to minimize an expected loss function of $E - \tilde{E}$, without concern for **unbiasedness** or other classical constraints, leads naturally to **shrinkage estimation** of E (as opposed to β), in which $\tilde{E}$ is "stabilized" by pulling it away from Y, toward a shrinkage point. These methods include Stein, **ridge**, **penalized**, pseudo-Bayes, semi-Bayes, **empirical-Bayes**, and **random-coefficient** estimation [4, 13, 16, 55], and their modern hierarchical-Bayes variants [6, 15, 40]. Thus, smoothing can be identified with shrinkage methods for fitting saturated models. For example, **smoothing splines** are based on penalized fitting of an overparameterized "natural" cubic spline [34]. The key questions are then: what point should Y (the conventional saturated-model estimator of E) be shrunk toward, and by how much.

Intuitive answers to these questions arise naturally from noting that all shrinkage procedures can be cast in a Bayesian format, for example, ridge and penalty parameters correspond to hyperparameters in a **prior distribution** $p(\mathrm{E})$ for E [39]. The chief difference among the procedures is that some (e.g. empirical-Bayes) estimate all hyperparameters from the data, while others (e.g. semi-Bayes) fix certain hyperparameters in advance (usually the **variance components**). The resulting shrinkage point corresponds to the estimated prior mean, and the degree of shrinkage is a function of the random error and prior variances, where the latter may be estimated or fixed.

As an example, suppose $p(\varepsilon;\upsilon)$ is **multivariate normal** (MVN) with **information matrix** (inverse-covariance) $\upsilon^{-1} = \iota_\varepsilon$, and the prior $p(\mathrm{E})$ is MVN with mean μ_E and information ι_E. With $c = (\iota_\varepsilon + \iota_\mathrm{E})^{-1}$, the posterior mean $\mu_{\mathrm{E}|Y}$ of E is $c(\iota_\varepsilon Y + \iota_\mathrm{E}\mu_\mathrm{E})$, which is the result of shrinking Y toward μ_E with the degree of shrinkage determined by the relative amount of prior information for ε and E. The posterior mean for ε is $\mu_{\varepsilon|Y} = Y - \mu_{\mathrm{E}|Y} = c\iota_\mathrm{E}(Y - \mu_\mathrm{E})$, the result of shrinking the "preposterior" residual $Y - \mu_\mathrm{E}$ toward the origin by the factor $c\iota_\mathrm{E}$; the error model $p(\varepsilon;\upsilon)$ can thus be viewed as an origin-centered prior for estimating ε, with $\mu_{\mathrm{E}|Y} = Y - \mu_{\varepsilon|Y}$ the result of subtracting the ε estimate $\mu_{\varepsilon|Y}$ from Y. Symmetrically, $\mu_{\mathrm{E}|Y} - \mu_\mathrm{E} = c\iota_\varepsilon(Y - \mu_\mathrm{E})$ is the result of shrinking $Y - \mu_\mathrm{E}$ toward the origin by the factor $c\iota_\varepsilon$.

These relations extend well beyond those in which the error and prior distributions are approximately MVN. For example, if $p(Y|\mathrm{E})$ is **multinomial** with total T and cell probability vector $\pi = \mathrm{E}/T$, and the prior $p(\pi)$ is Dirichlet with mean E_π (*see* **Multivariate Distributions, Overview**) then $\mu_{\mathrm{E}|Y} = (1 - \mathrm{c})Y + c\mathrm{E}_\pi = Y - c(Y - \mathrm{E}_\pi)$ where $c = P/(T + P)$ and P is an "effective prior sample size" [4, 17], [18, Chapter 12]. The same type of relation holds under product-multinomial-Dirichlet, product-binomial-beta, and product-Poisson-gamma models for $p(Y|\mathrm{E})p(\mathrm{E})$, with c varying across the independent subvectors (blocks) within Y.

Filtering by Hierarchical Models

Semi-Bayes (SB) or partial-Bayes procedures modify elementary Bayes procedures by estimating μ_E from Y. Usually, the prior mean μ_E is given a parametric model form $\mu_\mathrm{E} = M_\mu(\beta)$, and so estimating μ_E reduces to computing an estimate $\tilde{\beta}$ of β. This resembles conventional modeling, but now only μ_E rather than E is constrained to lie in the model manifold. Overall, we obtain a hierarchical structure $Y = M_\mu(\beta) + \delta + \varepsilon$ where the partial residual $\delta = \mathrm{E} - M_\mu(\beta)$ has a mean-zero prior density $p(\delta;\tau)$. The entire model is partial-Bayes insofar as $\mathrm{E} = M_\mu(\beta) + \delta$ is partitioned into a component $M_\mu(\beta)$ with no prior on β and a component δ with a known, mean zero proper prior [1].

Because δ and ε are aliased, the partition of the estimated preposterior residual $Y - \tilde{\mu}_\mathrm{E} = Y - M_\mu(\tilde{\beta})$ into estimates $\tilde{\varepsilon} = Y - \tilde{\mu}_{\mathrm{E}|Y}$ (which will be discarded after use in **diagnostics**) and $\tilde{\delta} = \tilde{\mu}_{\mathrm{E}|Y} - \tilde{\mu}_\mathrm{E}$, is determined by the prior $p(\delta;\tau)$. $\tilde{\delta}$ may be viewed as the result of shrinking the unconstrained δ estimate $Y - \tilde{\mu}_\mathrm{E}$ toward the origin; given $p(\varepsilon;\upsilon)$, the prior $p(\delta;\tau)$ controls the amount of shrinkage. The prior parameter vector τ may be viewed as a smoothing or tuning vector (although more often τ is a scalar, and $\lambda = 1/\tau^2$ is called the tuning parameter). Empirical-Bayes (EB) procedures estimate τ as well as β, usually under a simple model for $p(\delta;\tau)$, for example, $p(\delta;\tau)$ is often assumed MVN$(0,\tau^2 C)$ with τ unknown and C a known structured matrix (identity, exchangeable, autocorrelated, etc). Bayes empirical-Bayes (BEB) procedures introduce proper hyperpriors $p(\beta)$ and $p(\tau)$ to estimate β and τ; $p(\tau)$ may be interpreted as encoding prior information about the dispersion of δ, or about how much smoothing would be best to remove the most noise without removing too much signal.

Use of lower-dimensional prior models $M_\mu(\beta)$ and $p(\delta;\tau)$ for E resurrects questions of model and covariate choice. Perhaps by analogy with conventional modeling, much of the literature on smoothing has focused on simple $M_\mu(\beta)$, for example, most table smoothers shrink toward independence models or models reduced by a **variable-selection** algorithm [4], and many scatterplot smoothers shrink toward a line [34]. As mentioned earlier, these practices can be hazardous, insofar as any patterns falling outside the prior-model manifold will be smoothed (flattened) out, and perhaps become inapparent as a result. To avoid this problem, one can keep the prior model very large, large enough to be capable of approximating any pattern that *a priori* might be of interest and of nonnegligible probability. To cite a phrase attributed to L.J. Savage, at this stage "all models should be as

big as a house"; any data-based model reduction risks conflating inferential objectives with the more limited goal of noise reduction, which as argued above should be based on a flexible model $M_\mu(\beta)$ for μ_E.

A more specific guide for elicitation of the form of the prior model (as opposed to values of β) is that $M_\mu(\beta)$ should be rich enough to make the δ components independent given β [36]. In this sense, $M_\mu(\beta)$ should exhaust all structure in any prior beliefs or information about E, leaving only a random prior residual. This use of "random" is however distinct from the description of random error ε: δ represents an allowance for uncertainty about structure, and may contain heretofore unsuspected patterns in E (signals), whereas ε represents truly uninformative noise in Y. For data in which ε is generated by a known mechanism, such as studies conducted with known random allocation or sampling designs (and other situations admitting design-based inference), $p(\varepsilon;\upsilon)$ is known, and hence can be removed from the estimable distribution of $Y - \mu_E = \delta + \varepsilon$ to estimate $p(\delta;\tau)$.

In observational data, however, $p(\delta;\tau)$ and $p(\varepsilon;\upsilon)$ are not separately identified, and τ or some transform like $\lambda = \tau^{-2}$ is used as a tuning parameter that determines how $Y - \mu_E$ will be decomposed into the signal carrier δ and pure noise ε. When τ^2 is a prior variance, as τ goes to zero ($\lambda \to \infty$), smoothing approaches conventional fitting of the model $E = \mu_E = M_\mu(\beta)$; this extreme is complete smoothing away of everything outside the model manifold, and so treats the entire departure from $M_\mu(\beta)$ as noise. Conversely, as τ grows without bound ($\lambda \to 0$) smoothing approaches conventional fitting of the saturated model; this extreme is no smoothing at all, and so treats the entire departure from $M_\mu(\beta)$ as structural. Smoothing can thus be seen as a compromise between extremes of treating $M_\mu(\beta)$ as the only structure, and on the other hand treating all variation in Y as structural (nonrandom or systematic).

Nonlinear Models

Thus far, I have used an additive decomposition of E. Nonetheless, all the procedures extend straightforwardly to forms such as $g(E) = M_\mu(\beta) + \delta$ for some known link function (*see* **Generalized Linear Model**) g, where $g(E)$ has prior mean $M_\mu(\beta)$ and δ again has a mean-zero prior density $p(\delta;\tau)$. These extensions are just special cases of generalized-linear and nonlinear hierarchical modeling [15, 36] in which the first-stage model is saturated: $g(E)$ is the vector of first-stage coefficients, with β and δ the second-stage coefficient and residual vectors. If $M_\mu(\beta) = X\beta$ with β given no prior, β is the fixed effect in a generalized-linear mixed model with random effect δ [5].

Basic Computations

Some Considerations for Current Platforms

Most current Bayesian texts presume software for posterior sampling is available, and so emphasize BEB or more general hierarchical-Bayes approaches [15], while non-Bayesian texts focus on EB or equivalent variance-components methods [10]. Many if not most epidemiologists are however committed to one or a few major packages (SAS, SPSS, or Stata) (*see* **Software, Biostatistical**), which as of this writing lack posterior sampling; also, most prefer simple programs, at least for primary analyses. This section therefore focuses on basic numeric approximations that can be carried out with popular regression software to obtain estimated prior and posterior means $\tilde{\mu}_E$ and $\tilde{\mu}_{E|Y}$.

Smoothing with a flexible $M_\mu(\beta)$ will tend to make $Y - \tilde{\mu}_E$ close to the origin, and hence $\tilde{\delta} = \tilde{\mu}_{E|Y} - \tilde{\mu}_E$ will tend to have very small components roughly proportional to τ when τ^2 is a prior variance. While this is not a problem if τ is known, it can make τ estimators from moment-based EB procedures (used for example in SAS proc Glimmix) perform very poorly in sparse binomial data [5] (*see* **Chi-square Tests**), a case in which preliminary smoothing is most important. For example, the truncated-EB moment-estimator of τ in the MVN(0,$\tau^2 C$) model for $p(\delta;\tau)$ often collapses to 0, resulting in $\tilde{\mu}_{E|Y} = \tilde{\mu}_E$, whereas the corresponding fixed-τ (semi-Bayes) $\tilde{\mu}_{E|Y}$ behaves more reasonably under the same conditions [21, 22]. These problems with "classical" EB procedures reflect the weakness of data information for separating Y into the three model components ($\mu_E, \delta, \varepsilon$), and the consequent need for strong prior information in order to produce a contextually sensible degree of filtering.

When Y is not very informative for τ under the model but $p(\tau)$ is very informative, $p(\tau|Y)$ will differ little from $p(\tau)$ and the $\mu_{E|Y}$ obtained from a full BEB procedure using $p(\tau)$ will often be well

approximated by the $\mu_{E|Y}$ obtained from the corresponding procedure that fixes τ at the mean or median of $p(\tau)$ (which is just BEB with point prior for τ). This suggests that, for smoothing purposes, Bayes or semi-Bayes procedures can be used as convenient approximations to full hierarchical procedures. Under the model $g(E) = M_\mu(\beta) + \delta$, $\mu_{E|Y}$ has first-order asymptotic approximation $\tilde{\mu}_{E|Y} = g^{-1}[M_\mu(\tilde{\beta}) + \tilde{\delta}]$, where $(\tilde{\beta}, \tilde{\delta})$ is the posterior mode of (β, δ). One can obtain $(\tilde{\beta}, \tilde{\delta})$ by maximizing the loglikelihood for the model $g(E) = M_\mu(\beta) + \delta$ with penalty $\ln\{p(\delta;\tau)\}$, that is, $(\tilde{\beta}, \tilde{\delta})$ is the maximum penalized-likelihood (MPL) estimate of (β, δ) [40]. The negative inverse of the Hessian at $(\tilde{\beta}, \tilde{\delta})$ is then an approximate posterior covariance matrix for $(\tilde{\beta}, \tilde{\delta})$, and $\tilde{\mu}_E = g^{-1}(X\tilde{\beta})$ and $\tilde{\mu}_{E|Y} = g^{-1}(X\tilde{\beta} + \tilde{\delta})$ are approximate prior and posterior means for E.

Data-augmentation Priors

For **beta**, Dirichlet, and normal priors and some generalizations with τ fixed, one can obtain $(\tilde{\beta}, \tilde{\delta})$ and hence $\tilde{\mu}_{E|Y}$ from conventional regression packages by augmenting Y and X with "prior data" [1, 7, 25, 28, 30, 37]. To illustrate, suppose $p(Y|E)$ is product-binomial with n the vector of totals and risks $\pi_i = E_i/n_i = \text{expit}(X_i\beta + \delta_i)$, where $\text{expit}(u) \equiv 1/(1 + e^{-u})$. This is a linear-logistic regression with random effects δ_i; for Table 1, π_i and n_i are the risk of high exposure and the size of group i. A mean-zero independence conjugate prior density for δ is proportional to $\Pi_i \text{expit}(\delta_i)^{ai}\text{expit}(-\delta_i)^{ai}$, a product of logistic-beta(a_i, a_i) densities, that is, densities such that $\text{expit}(\delta_i)$ is beta(a_i, a_i); in particular, $a_i = 1$ yields the standard **logistic density**. Let a be the vector of a_i, I the N-by-N identity, $\mathbf{0}$ an N-by-J matrix of zeros, and $\updownarrow$ vertical array concatenation. An ordinary ML logistic-regression program can be tricked into entering the log of this prior as a penalty function by augmenting Y, n, and X to $Y^* = Y \updownarrow a$, $n^* = n \updownarrow 2a$, and $X^* = (X, I) \updownarrow (\mathbf{0}, I)$. The exact posterior mode $(\tilde{\beta}, \tilde{\delta})$ and an approximate posterior covariance matrix for (β, δ) will then be returned as the ML estimates; if the program will provide fitted proportions $\tilde{\pi}_i$ (as most will), the first N of the $n_i\tilde{\pi}_i$ will be $\tilde{\mu}_{E|Y}$.

This data-augmentation prior (DAP) can be generalized in a number of ways [25, 28]. To give β as well as δ a proper prior, one need only add J more augmenting rows to X^* encoding the priors for the β_j. The heaviness of the prior tails (which decreases with a_i) may be controlled while preserving the scale by rescaling the density, that is, replacing δ_i with δ_i/τ_i in the density. **Skewness** may be introduced while preserving location by using a logistic-beta(a_i, b_i) density with $a_i \neq b_i$ and recentering, that is, replacing δ_i/τ_i with $(\delta_i - m_i)/\tau_i$ where m_i is the prior mode before recentering.

Because the conjugate prior approaches normality as a_i increases, one can approximate normal$(0,\tau_i^2)$ densities for the δ_i by setting $a_i = (4/h^2\tau_i^2) - 1$ and $X^* = (X, I) \updownarrow (\mathbf{0}, hI)$, where the constant h controls closeness to normality [25]. Normality is approached as h approaches 0; estimates would not change meaningfully after h drops below a certain value, but for rounding errors produced by finite machine precision; typically, h yielding all $a_i > 100$ provides adequate numerical approximations. Prior correlations among the β and δ may be introduced by using a nondiagonal matrix in place of hI in X^*, or by using an uncorrelated reparameterization for the analysis and then transforming results back to the original parameterization.

Specification of the Prior Parameters

Incorporating Contextual Information

Specification of prior parameters from contextual (subject-matter) information can be done by back-calculating those parameters from elicited prior percentiles (e.g. [9], p. 384; 2, 24, 25, 28). Continuing the above example, suppose one is 95% certain *a priori* that the odds $\pi_i/(1-\pi_i)$ is within an R-fold range around the prior median $\exp(X_i\beta)$, which implies that $\text{logit}(\pi_i) = X_i\beta + \delta_i$ falls in an interval of width $\ln(R)$ around $X_i\beta$ with 95% prior probability. Under a normal prior for δ_i, this relation further implies that the width $2(1.96)\tau_i$ of a 95% prior interval for δ_i is $\ln(R)$, so $\tau_i = \ln(R)/3.92$ (e.g. $\tau_i = 0.764$ when $R = 20$). For other prior densities, one may use tables or numerically solve for the hyperparameter values that make the difference of the prior 97.5th and 2.5th prior percentiles for δ_i equal to $\ln(R)$ (*see* **Prior Distribution**).

Because the δ_i percentiles depend entirely on context, no purely numeric guidelines can be given for their choice. It should be borne in mind, however, that δ_i are *residual* effects after factoring out effects in the smoothing model $M_\mu(\beta)$, and so should not

be very large if the model has been chosen "maximally," that is, to capture all systematic (structural) effects expected in the context. In the present example, the only expected systematic effects on the study results are from measurement type and from power-system type, both of which will be coded into $M_\mu(\beta)$. The meaning of "not very large" is rather vague; nonetheless, in lifestyle, **environmental, and occupational epidemiology**, odds ratios between 1/3 and 3 or perhaps 1/4 and 4 may be considered not very large, corresponding to an R of 9 or perhaps 16. Still larger values are defensible as a conservative choice, however: Specifying R (and hence τ) too large relative to the true residual effects δ leads to overly wide interval estimates with supranominal coverage, but specifying R too small leads to overly narrow intervals with subnominal coverage [20, 21]; hence a value of $R = 20$ is used below. Fortunately, as will be illustrated, intervals sensitive to reasonable variation in R tend to be those so wide that no useful inference can be drawn under any reasonable choice.

An alternative approach indirectly fixes τ by instead specifying the "effective degrees of freedom" (edf) desired for the smoothing, where edf is defined as a function of the projection matrix of Y to $\tilde{E}$ implicit in the final fit [34]. The rank of X is the smallest possible value for edf and corresponds to $\tau = 0$, or simply fitting $M_\mu(\beta)$ as $M_E(\beta)$ (conventional modeling); the total (data) degrees of freedom is the largest possible value. This type of specification may be natural in the context of fitting curves or surfaces in which a somewhat qualitative prior about the complexity of the curve can be visualized in relation to polynomial curves or surfaces, but does not seem as intuitive for qualitative data.

Traditional Table Smoothing Revisited

Traditional methods of handling sparse count data can be recast as primitive versions of the data-augmentation method (*see* **EM Algorithm**) with highly constrained prior parameters [4, Chapter 12]. Adding a constant c (e.g. 1/2) to each count (Y_i and $n_i - Y_i$) in the binomial case corresponds to using independent, symmetric beta(c, c) priors for the π_i rather than the residuals expit(δ_i), which implies a prior mean $c/2c = 1/2$ and variance $(c/2c)^2/(2c + 1) = 1/(8c + 4)$ for all the π_i. This prior almost never makes sense contextually. For example, if π_i is an exposure or disease risk, it rarely will have prior mean 1/2 and (with $c = 1/2$) variance 1/8 for all π_i. In fact, most disease risks are much less than 1/2 and have fairly well-known dependencies on demographic covariates, which should be accounted for by inclusion of those covariates and a constant (or equivalent) in X. Furthermore, in pooled analyses and **meta-analyses**, risks will vary across studies.

Smoothing the Example Data by a Maximal Model

Specification and Fitting

Let L be the leukemia (case) indicator, with S the vector of 14 study indicators, D the indicator that the study used direct measurements (vs. calculated), and V the indicator of 120 volt 60 Hz power system (vs. 220 volt 50 Hz); there are no $D = 0$, $V = 1$ studies, so these indicators define only three rather than four groups of studies. The prior design matrix X will be a function of L, S, D, and V. Because S contains indicators for every study and because D and V are functions of S, once S is included, no constant and no main effect for D or V is needed.

To minimize alteration of data patterns, one should want a rich prior model $X\beta$ for the logit prior mean. The richest model is the saturated model, in which X contains 28 linearly independent columns, for example, 14 corresponding to S and 14 corresponding to the product terms in LS. To avoid zeros in the fitted values, however, certain products must be excluded (those with zero **sufficient statistics**; see [4]). On the other hand, patterns in results related to measurement D and power system V were expected. Representation of such patterns requires LD and LV products; thus, X was given columns for L, S, LD, and LV (17 columns). Any larger design matrix defined from L, S, D, V alone would introduce fitted zeros; thus, X is maximal for the smoothing objective.

The prior $p(\delta; \tau)$ was $MVN(0,\tau^2 I)$. Fitting was done by data augmentation with numerical constant $h = 0.01$, and prior standard deviation $\tau = 0.764$ $(R = 20)$; this adds 28 pseudo-observations with $a_i = (2/h\tau_i)^2 - 1 \approx 68\,489$ cases and 64 489 controls each, and 28 augmenting covariate columns equal to zero everywhere except augmenting column i, which is set to h in the ith augmenting row.

Results

Let $Z = (X, I)$ and $\theta = \beta \updownarrow \delta$. Table 1 shows the resulting smoothed counts $\text{expit}(Z_i\tilde{\theta})n_i$ (the elements of $\tilde{\mu}_{E|Y}$) for the highly exposed category; the smoothing model fit by MPL preserves the totals so those are not repeated in the table. The smoothed odds ratios are identical whether computed from the smoothed counts or from the model coefficients as $\tilde{\omega} = \exp\{(Z_{L1} - Z_{L0})\tilde{\theta}\}$, where Z_{L1} and Z_{L0} comprise the rows of Z with $L = 1$ (cases) and $L = 0$ (controls). Note the modest shrinkage of Y to $\tilde{\mu}_{E|Y}$: The largest absolute change in a count is 1.49, the changes are less than 1.00 in 12 of the 14 studies, and the mean absolute change is only 0.45. Also, to the third decimal point there is no change upon applying the ML summary point and interval estimators to the smoothed counts. Thus, as desired, the smoother has not altered the main summary from these data. In contrast, adding 1/2 to each cell inflates these estimators slightly, to 1.72 (1.31, 2.26).

Turning to patterns across studies, the three *a priori* study groups encoded in $X(D = V = 0, D = 1\&V = 0, D = V = 1)$ have estimated geometric mean odds ratios of 1.63, 2.24, 1.72 ($P = 0.88$ for deviance test of equality of these means). The smoothing procedure shrinks the odds ratios toward their group geometric means, but only the most unstable odds ratios change dramatically. Odds ratios that were previously infinite (Coghill, Dockerty) or zero (Tynes) due to zero counts, are pulled back to much more reasonable values of 2.63, 3.31, and 0.68; for comparison, adding 1/2 per cell to these studies yields odds ratios of 3.05, 6.83, and 0.21. The two other outlying odds ratios are 4.53 and 3.51 (Feychting, Savitz); the smoother shrinks these estimates to 3.30 and 2.45, whereas adding 1/2 inflates them to 4.73 and 3.68.

Estimating Variances of Smoothed Quantities

Given that an end user of a smoothed data set is likely to need variance estimates for computed quantities, there is an issue as to whether naïve computations (treating the smoothed data as if it were the observed data) are adequate. To illustrate the issues, Table 1 provides approximate 95% confidence limits $\tilde{\omega}_k\exp(\pm 1.96\tilde{\sigma}_k)$ for the smoothed odds ratios ($k = 1, \ldots, 14$), where $\tilde{\sigma}_k^2$ is the kth diagonal element of the estimated covariance matrix of the log odds ratio estimates, $(Z_{L1} - Z_{L0})\tilde{\text{Cov}}(\tilde{\theta})(Z_{L1} - Z_{L0})'$, and $\tilde{\text{Cov}}(\tilde{\theta})$ is the inverse information matrix from the fitted model. Computing the limits (incorrectly) by applying the raw-data variance formulas [50], Chapters 14 and 15 to the smoothed counts makes little difference for the larger studies and almost no difference for the summaries; for example, for the Linet study, the limits from the correct formula are 0.94, 2.43, whereas the raw-data formula naïvely applied to the smoothed counts yields 0.92, 2.47. Nonetheless, the naïve intervals tend to be excessively wide because they do not use the model information to estimate variances. This excess is small when the smoothing constraints are weak (due to the large τ and large model) relative to the data information, as in large studies. For summary measures, the excess becomes negligible because those are primarily determined by the large studies.

On the other hand, the excess can be large for small studies, for example, for the Verkasalo study, the limits from the correct formula are 0.40, 7.57, whereas the raw-data formula naïvely applied to the smoothed counts yields 0.18, 17.2. Such a difference may be of little or no practical importance, however. First, both intervals are asymptotic and so are questionable when some cell counts are very small (in Verkasalo, there is only one highly exposed case); second, when the two intervals differ greatly, both are so wide that their message is the same, that is, little of interest can be said about the odds ratio from a small study without stronger assumptions about its relation to odds ratios from other studies (this consideration is why exact methods or more refined approximations may seem academic in most epidemiologic applications).

Nonetheless, to avoid variance overestimation arising from the use of smoothed counts in place of observed counts, subsequent analyses would have to be based on $\tilde{\text{Cov}}(\tilde{\theta})$ rather than naïve variances. Analogously, in missing-data problems, valid variance estimation must account for the imputation (*see* **Missing Data in Clinical Trials**), although in those problems, the naïve formulas underestimate rather than overestimate the variances [42].

Discussion

Is Sensitivity Analysis Helpful in the Face of Arbitrary Sensitivity?

A natural question is to ask whether the smoothing results are sensitive to the choice of τ (or

equivalently, R or λ) given the smoothing model. The answer is that, like latent-structure analyses, they are arbitrarily sensitive within their mathematical bounds. In the example, when $\tau > 2$ ($R > 2500$, $\lambda < 0.25$) the smoothed counts in $\tilde{\mu}_{\mathrm{E}|Y}$ are all within 0.5 of the observed counts in Y; whereas when $\tau = 0$ ($R = 1, \lambda = \infty$) the smoothed counts in $\tilde{\mu}_{\mathrm{E}|Y}$ equal the estimated prior counts in $\tilde{\mu}_{\mathrm{E}} = M_{\mu}(\tilde{\beta})$, and deviate as much as 3.5 from the observed counts. These two count vectors bracket those obtainable under intermediate choices for τ ($2 > \tau > 0$), which include any remotely reasonable choice. Consequently, statistics that are almost the same whether computed under $\tau = \infty$ (i.e. from the observed counts) or under $\tau = 0$ (i.e. under the prior model) will be insensitive to *any* choice of τ; for example, the ML and MH summary statistics hardly change. Conversely, statistics that change much between these extremes (such as odds ratios for studies with a 0 count) are those sensitive to τ; for example, the odds ratios from Coghill and Dockerty are implausibly large (5.4 and 9.2) when $\tau = 2$, but both approach the plausible value of 1.97 as τ approaches 0.

To make contextually meaningful inferences about parameters with sensitive estimates, one must confine τ to values that are reasonable, that is, values for τ that assign the bulk of probability to plausible values for δ. This brings in subjective judgment about τ as well as δ, which could (some might say should) be summarized in a prior $p(\tau)$. One could then draw choices for τ from $p(\tau)$ or $p(\tau|Y)$ and repeat the analysis. Averaging the resulting log odds ratios over $p(\tau|Y)$, however, approximates the BEB posterior mean log odds ratios, and the BEB posterior intervals have the advantage of providing a coherent integration of uncertainties about δ, including uncertainty about τ.

Does BEB then obviate the need for sensitivity analysis of τ? One might demand a Bayesian sensitivity analysis, which varies $p(\tau)$ in the BEB analysis. Such a demand only leads back to the problem of arbitrary sensitivity, however. In the above example, any $p(\tau)$ with support above two nearly reproduces the observed counts, whereas a $p(\tau)$ concentrated near 0 reproduces the counts expected under the prior structural model. Paralleling ordinary sensitivity analysis, we will only see that stable statistics are insensitive across the range of $p(\tau)$, and that unstable statistics almost completely depend on $p(\tau)$. Again, however, both facts can be seen immediately by comparing results computed from the raw data with results computed under the model $\mathrm{E} = M_{\mu}(\beta)$ (i.e. with $\tau = 0$), or by examining raw-data confidence intervals.

One rationale for presenting analyses with different τ is that it provides inferences tailored to readers with different priors on δ. Nonetheless, such multiple presentation shifts labor and space toward interpretation of sensitive statistics, which provide a poor basis for inference. On the other hand, varying τ or $p(\tau)$ will be superfluous for inferences based only on insensitive statistics, which are a better focus of inference.

Frequentists often opt to use an "empirical" point estimate of τ (e.g. from cross-validation) rather than a prior τ or $p(\tau)$. As with use of a prior, however, this option does not display sensitivity to τ. It is also objectionable because it can result in a τ estimate that is both implausible (often seeming too small) and unstable. The subjectively and objectively poor moderate-sample performance of common τ estimators (e.g. [21]) seems a crucial problem often neglected in promotions of so-called "objective" smoothing techniques based on these estimators. One way to address this problem involves the same rationale and form as basic shrinkage estimation techniques for E and β: stabilize the τ estimator by shrinking it toward a prior mean. This solution introduces a prior $p(\tau)$, and so leads back to the BEB approach.

As a technical note, given the normal-prior specification (as in the above example), the numeric constant h is not a statistical parameter in the DAP; thus there is no philosophic issue concerning sensitivity analysis for h. One simply needs to ensure that h is small enough so that the numeric results are stable, but not so small that overflow or underflow occurs. This can be done by checking to see that reducing h leaves the results essentially unchanged, without introducing numeric warnings. One could instead treat h as a hyperparameter that determines departure from normality in the DAP (with increasingly heavy tails as h increases), in which case the debate of sensitivity analysis for h versus BEB analysis with a hyperprior $p(h)$ could again be raised.

The issues just described also arise for latent-structure analyses. For a description of a hierarchical Bayes alternative to sensitivity analysis in a structural model involving latent variables, see [27].

Smoothing and Model-robust Estimation

Another way to view smoothing is as a form of model-robust regression estimation. A large class of model-robust approaches (e.g. [60]) justify the use of a restrictive form $\mathrm{E} = M_{\mathrm{E}}(\beta)$ by appeal to theorems showing that the least-squares or ML estimator of β converges to a value β_p that would be obtained by fitting the model to the population joint distribution $p(X, Y)$, where the latter is not constrained to follow the model; they also derive standard errors without assuming the model is correct (robust or "sandwich" variance estimators). These approaches assume an absence of sampling biases and measurement errors in the data generation, and are scientifically justifiable only if the population summary afforded by $M_{\mathrm{E}}(\beta_p)$ omits no contextually important information about the true regression E.

As an example in which the latter condition can fail and appears to have failed severely at times, consider the controversy over the value of **ecologic** (group-aggregate) **studies** for inference about individual risks. With E these risks, and subdividing the population into groups indexed by k, it is well known that $\mathrm{E} = X\beta$ (risk linearity) induces a group-mean regression $\overline{\mathrm{E}}_k = \overline{X}_k\beta$, reflecting the linearity of means. This relation is often used (incorrectly) to justify unrestricted inferences from the group to individual level. If, however, the individual regression contains important nonlinearities (as for many cancer risk factors, including age, cigarette use, and asbestos exposure), the group-mean regression may not resemble the individual-level regression in form or even in direction of key associations [26, 54, 59]. This should not be surprising given that the linear risk approximation can be woefully inadequate, especially for projecting risks at extreme design points (which are often the focus of health controversies).

This sort of problem is not addressed by conventional model-robust methods, which dutifully try to estimate the best approximation *of a given form*. Smoothing instead demotes the model form from a known model $M_{\mathrm{E}}(\beta)$ for E to a model $M_{\mu}(\beta)$ for the prior mean of E, thus allowing data departures from the model (whether "significant" or not) to not only alter variance estimates (as in conventional model-robust estimation) but to also alter the fitted values $\tilde{\mathrm{E}} = \tilde{\mu}_{\mathrm{E}|Y}$.

Conclusion

Smoothing can be viewed as an exploratory technique falling between raw-data description and inference about deeper (and often latent) structures of contextual interest, as well as a modeling technique that frees one from the rigidity of typical model specifications. It can also be viewed as a data-repairing technique akin to missing-value imputation. While it cannot address fundamental data inadequacies (such as **selection bias** or poor measurement), like imputation, it can ease subsequent analyses that attempt to address those inadequacies. The hierarchical-Bayes format for smoothing (which has been around at least since the 1960s) provides useful insights for both specification of smoothing models and parameters, and eases computation of smoothed data from standard software. It would thus seem practical and timely to include such smoothing methods in basic biostatistical training, and encourage their use whenever problems arise because of unstable or zero counts.

References

[1] Bedrick, E.J., Christensen, R. & Johnson, W. (1996). A new perspective on generalized linear models (1996), *Journal of the American Statistical Association* **91**, 1450–1460.

[2] Bedrick, E.J., Christensen, R. & Johnson, W. (1997). Bayesian binomial regression: predicting survival at a trauma center, *The American Statistician* **51**, 211–218.

[3] Berk, R.A. (2004). *A Regression Analysis: A Constructive Critique*. Sage Publications, Thousand Oaks.

[4] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.

[5] Breslow, N.E. & Clayton, D. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**, 9–25.

[6] Carlin, B. & Louis, T.A. (2000). *Bayes and Empirical-Bayes Methods of Data Analysis*, 2nd Ed. Chapman & Hall, New York.

[7] Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B. & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression, *Journal of the American Statistical Association* **86**, 68–78.

[8] Coghill, R.W., Steward, J. & Philips, A. (1996). Extra low frequency electric and magnetic fields in the bedplace of children diagnosed with leukemia: a case-control study, *European Journal of Cancer Prevention* **5**, 153–158.

[9] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, New York.

[10] Cox, D.R. & Solomon, P.J. (2002). *Components of Variance*. Chapman & Hall, New York.

[11] Dockerty, J.D., Elwood, J.M., Skegg, D.C.G. & Herbison, G.P. (1998). Electromagnetic field exposures and childhood cancers in New Zealand, *Cancer Causes and Control* **9**, 299–309; Erratum (1999), **10**, 641.

[12] Eddy, D.M., Hasselblad, V. & Schachter, R. (1992). *Meta-Analysis by the Confidence Profile Method*. Academic Press, New York.

[13] Efron, B. & Morris, C.N. (1975). Data analysis using Stein's estimator and its generalization, *Journal of the American Statistical Association* **70**, 311–319.

[14] Feychting, M. & Ahlbom, A. (1993). Magnetic fields and cancer in children residing near Swedish high-voltage power lines, *American Journal of Epidemiology* **138**, 467–481.

[15] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2003). *Bayesian Data Analysis*, 2nd Ed. Chapman & Hall/CRC, New York.

[16] Good, I.J. (1965). *The Estimation of Probabilities*. MIT Press, Cambridge.

[17] Good, I.J. (1983). *Good Thinking*. University of Minnesota Press, Minneapolis.

[18] Graham, P. (2000). Bayesian inference for a generalized population attributable fraction, *Statistics in Medicine* **19**, 937–956.

[19] Greenland, S. (1990). Randomization, statistics, and causal inference, *Epidemiology* **1**, 421–429.

[20] Greenland, S. (1993a). Summarization, smoothing, and inference, *Scandinavian Journal of Social Medicine* **21**, 227–232.

[21] Greenland, S. (1993b). Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression, *Statistics in Medicine* **12**, 717–736.

[22] Greenland, S. (1997). Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analysis, *Statistics in Medicine* **16**, 515–526.

[23] Greenland, S. (1998). The sensitivity of a sensitivity analysis (invited paper), *1997 Proceedings of the Biometrics Section*. American Statistical Association, Alexandria, pp. 19–21.

[24] Greenland, S. (2000). When should epidemiologic regressions use random coefficients? *Biometrics* **56**, 915–921.

[25] Greenland, S. (2001). Putting background information about relative risks into conjugate priors, *Biometrics* **57**, 663–670.

[26] Greenland, S. (2002). A review of multilevel theory for ecologic analysis, *Statistics in Medicine* **21**, 389–395.

[27] Greenland, S. (2003a). The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia, *Journal of the American Statistical Association* **98**, 47–54.

[28] Greenland, S. (2003b). Generalized conjugate priors for Bayesian analysis of risk and survival regressions, *Biometrics* **59**, 92–99.

[29] Greenland S. (2004). Multiple-bias modeling for observational studies, *Journal of the Royal Statistical Society series A*, to appear.

[30] Greenland, S. & Christensen, R. (2001). Data augmentation for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression, *Statistics in Medicine* **20**, 2421–2428.

[31] Greenland, S., Schwartzbaum, J.A. & Finkle, W.D. (2000b). Problems from small samples and sparse data in conditional logistic regression analysis, *American Journal of Epidemiology* **151**, 531–539.

[32] Greenland, S., Sheppard, A.R., Kaune, W.T., Poole, C. & Kelsh, M.A. (2000a). A pooled analysis of magnetic fields, wire codes, and childhood leukemia, *Epidemiology* **11**, 624–663.

[33] Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall, New York.

[34] Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, New York.

[35] Kabuto, M. (2003). A study on environmental EMF and children's health: final report of a grant-in-aid for scientific research project, 1999–2001 (in Japanese). Japanese Ministry of Education, Culture, Sports, Science and Technology.

[36] Kass, R. & Steffey (1989). Approximate Bayesian inference in conditionally independent hierarchical models, *Journal of the American Statistical Association* **84**, 717–726.

[37] Landaw, E.M., Sampson, P.F. & Toporek, J.D. (1982). Advanced nonlinear regression in BMDP, *Proceedings of the Statistical Computing Section*. American Statistical Association, Washington, PP. 228–233.

[38] Lash, T.L. & Fink, A.K. (2003). Semi-automated sensitivity analysis to assess systematic errors in observational epidemiologic data, *Epidemiology* **14**, 451–458.

[39] Leamer, E.E. (1978). *Specification Searches*. Wiley, New York.

[40] Leonard, T. & Hsu, J.S.J. (1999). *Bayesian Methods*. Cambridge University Press, Cambridge.

[41] Linet, M.S., Hatch, E.E., Kleinermann, R.A., Robison, L.C., Kaune, W.T., Friedman, D.R., Severson, R.K., Haines, C.M., Hartsock, C.T., Niwa, S., Wacholder, S. & Tarone, R.E. (1997). Residential exposure to magnetic fields and acute lymphoblastic leukemia in children, *New England Journal of Medicine* **337**, 1–7.

[42] Little, R.J.A. & Rubin, D.A. (2002). *Statistical Analysis with Missing Data*, 2nd Ed. Wiley, New York.

[43] London, S.J., Thomas, D.C., Bowman, J.D., Sobel, E., Cheng, T.-C. & Peters, J.M. (1991). Exposure to residential electric and magnetic fields and risk of childhood leukemia, *American Journal of Epidemiology* **134**, 923–937.

[44] McBride, M.L., Gallagher, R.P., Theriault, H.G., Armstrong, B.G., Tamaro, S., Spinelli, J.J., Deadman, J.E.,

Fincham, S., Robson, D. & Choi, W. (1999). Power-frequency electric and magnetic fields and risk of childhood cancer, *American Journal of Epidemiology* **149**, 831–842.

[45] Michaelis, J., Schüz, J., Meinert, R., Semann, E., Grigat, J.P., Kaatsch, P., Kaletsch, U., Miesner, A., Brinkmann, K., Kalkner, W. & Karner, H. (1998). Combined risk estimates for two German population-based case-control studies on residential magnetic fields and childhood leukemia, *Epidemiology* **9**, 92–94.

[46] Olsen, J.H., Nielsen, A. & Schulgen, G. (1993). Residence near high voltage facilities and risk of cancer in children, *British Medical Journal* **307**, 891–895.

[47] Pearl, J. (2000). *Causality*. Cambridge University Press, New York.

[48] Phillips, C.V. (2003). Quantifying and reporting uncertainty from systematic errors, *Epidemiology* **14**, 459–466.

[49] Robins, J.M., Rotnitzky, A. & Scharfstein, D.O. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models, in *Statistical Models in Epidemiology*, M.E. Halloran & D.A. Berry, eds. Springer-Verlag, New York, pp. 1–92.

[50] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Ed. Lippincott, Philadelphia.

[51] Rubin, D.B. (1983). A case study of the robustness of Bayesian methods of inference, in *Scientific Inference, Data Analysis, and Robustness*, G.E.P. Box, T. Leonard & C.F. Wu, eds. Academic Press, New York, pp. 213–244.

[52] Savitz, D.A., Wachtel, H., Barnes, F.A., John, E.M. & Tvrdik, J.G. (1988). Case-control study of childhood cancer and exposure to 60-Hz magnetic fields, *American Journal of Epidemiology* **128**, 21–38.

[53] Schüz, J., Grigat, J.P., Brinkmann, K. & Michaelis, J. (2001). Residential magnetic fields as a risk factor for acute childhood leukemia: results from a German population-based case-control study, *International Journal of Cancer* **91**, 728–735.

[54] Sheppard, L. (2003). Insights on bias and information in group-level studies, *Biostatistics* **4**, 265–278.

[55] Titterington, D.M. (1985). Common structure of smoothing techniques in statistics, *International Statistical Review* **53**, 141–170.

[56] Tomenius, L. (1986). 50-Hz electromagnetic environment and the incidence of childhood tumors in Stockholm County, *Bioelectromagnetics* **7**, 191–207.

[57] Tynes, T. & Haldorsen, T. (1997). Electromagnetic fields and cancer in children residing near Norwegian high-voltage power lines, *American Journal of Epidemiology* **145**, 219–226.

[58] Verkasalo, P.K., Pukkala, E., Hongisto, M.Y., Valjus, J.E., Järvinen, P.J., Heikkilä, K.K. & Koskenvuo, M. (1993). Risk of cancer in Finnish children living close to power lines, *British Medical Journal* **307**, 895–899.

[59] Wakefield, J. (2004). Ecological inference for 2×2 tables, *Journal of the Royal Statistical Society Series A* **166**, in press.

[60] White, H. (1993). *Estimation, Inference, and Specification Analysis*. Cambridge University Press, New York.

Further Reading

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

UK Childhood Cancer Study Investigators (1999). Exposure to power-frequency magnetic fields and the risk of childhood cancer, *The Lancet* **354**, 1925–1931.

SANDER GREENLAND

Snedecor, George Waddel

Born: October 20, 1881, in Memphis, Tennessee.
Died: February 15, 1974, in Amherst, Massachusetts.

George Snedecor grew up in rural Florida and Alabama. His father, a Presbyterian minister and educator, worked training young black men for the ministry. George completed his B.S. degree in 1905 at the University of Alabama. He taught, first at the Selma Military Academy and then at Austin College, from 1905 to 1910. Then he undertook graduate work at the University of Michigan, receiving a Masters degree in physics in 1913. From 1913 to 1958 he served on the staff of Iowa State University in Ames, Iowa, beginning by teaching in the Department of Mathematics. In 1927, he became one of the two directors of the Mathematics Statistical Service in the department, a unit that provided **statistical consulting** for the campus. Iowa State organized a separate Statistical Laboratory in 1933 with Snedecor as director. He served in that position until 1947 when a policy of mandatory retirement from administrative duties required a change. The Statistical Laboratory continued, but in conjunction with a regular academic Department of Statistics founded in 1947. Snedecor continued as a staff member of the department until 1958.

The evidence indicates that George Snedecor recognized the usefulness of statistical thinking in the work of science early in his career at Iowa State. He introduced a course in statistics in 1915. He worked to bring the methods used at the Rothamsted Experimental Station to the researchers of the Iowa Agriculture and Home Economics Experiment Station. Henry A. Wallace, later Secretary of Agriculture and Vice President of the United States, shared Snedecor's views. In 1924, Snedecor assisted Wallace with a series of seminars at Iowa State on statistical methods and the use of business machines in statistical computing. This led to a publication entitled *Correlation and Machine Calculation* [1] in 1925, revised in 1931, which came to have a worldwide distribution. The Rothamsted influence came directly to Ames through summer visits by **Ronald Fisher**, first in 1931 and again in 1936. Fisher presented lectures on statistical methods to the local staff and students as well as visitors from other states and Canada.

This background, together with experience in statistical consulting in the experimental and observational sciences at Ames, led Snedecor to write a reference textbook on statistical methods [2]. The book became one of the most widely used and influential texts on the subject ever written. The first edition of *Statistical Methods Applied to Experiments in Agriculture and Biology* appeared in 1937. The eighth edition continued in print in 1996 [3]. The book has had a lifetime sales of approximately 235 000 copies. Beginning with the fifth edition, Snedecor asked **W.G. Cochran** to join in authoring the text. Cochran added a chapter on survey sampling and continued revision of the work through its sixth and seventh editions. The book had few competitors when it first appeared and became a standard reference and a graduate-level textbook on statistical methods in the agricultural research community of the US. Subsequently, it was translated into nine languages and its influence spread to other fields, serving as a model for many later textbooks on statistical methods. *Statistical Methods*, written by a man not steeped in mathematics, but endowed with a vision of the centrality of statistical thinking in science reinforced by years of consulting with active research workers, has strengthened the statistics profession and served the scientific community very well.

The world recognized the contributions Snedecor made and conferred many honors and awards including two honorary Doctor of Science degrees, one from North Carolina State University in 1956 and another from Iowa State University in 1958. Snedecor served as President of the **American Statistical Association** in 1948 and received the Samuel S. Wilks Memorial Medal in 1970. He became an Honorary Fellow of the British **Royal Statistical Society** in 1954. The building housing the statistics department on the Iowa State campus received the name Snedecor Hall in 1969.

George Snedecor directed the graduate work of **Gertrude M. Cox** and she received Iowa State's first degree in statistics. She, together with one of Snedecor's colleagues, Paul G. Homeyer, wrote a biographical and anecdotal account of the life and times of George Snedecor in 1975 [4]. The paper contains an appendix listing Snedecor's published work.

In 1959, Snedecor left Ames for work as a statistical consultant in the US Navy Electronics Laboratory in San Diego, California, which he continued

until 1963. He lived his last years in Amherst, Massachusetts with his son, James. He died in Amherst on February 15, 1974.

References

[1] Wallace, H.A. & Snedecor, G.W. (1925). *Correlation and Machine Calculation*. Iowa State College Official Publication, Vol. 23. (Revised 1931, Vol. 30.).

[2] Snedecor, G.W. (1937). *Statistical Methods Applied to Experiments in Agriculture and Biology*. Collegiate Press, Ames.

[3] Snedecor, G.W. & Cochran, W.G. (1996). *Statistical Methods*, 8th Ed. Iowa State University Press, Ames.

[4] Cox, G.M. & Homeyer, P.G. (1975). Professional and personal glimpses of George W. Snedecor, *Biometrics* **31**, 265–301.

D.F. COX

Snowball Sampling

Snowball sampling, also known as chain referral sampling, is a nonprobability method of survey sample selection that is commonly used to locate rare or difficult to find populations. Although there are several variations, this approach involves a minimum of two stages: (a) the identification of a sample of respondents with characteristic x at the zero-stage (s_0); and (b) the solicitation of referrals to other potentially eligible respondents believed to have characteristic x at snowball stages s_1 through s_k. In many applications, this referral process continues (or snowballs) until an acceptable number of eligible respondents have been located. **Statistical inferences** can be drawn from the zero-stage of a snowball sample, assuming that probability methods of selection were used. Samples drawn at s_1 through s_k, and samples that combine the zero and snowball stages are not representative, however, and cannot be used to make statistical inferences.

Goodman [8] provided the first comprehensive overview of this technique. More recent technical assessments have been made by Erickson [4], Frank [6], and TenHouten et al. [19]. Practical considerations in implementing snowball samples have been reviewed by Biernacki and Waldorf [1]. An early application of snowball sampling was reported by Menzel and Katz [15], who used this approach to study the diffusion of a medical innovation.

Snowball sampling continues to be used widely in biomedical, social, and behavioral research today. A review of recently published literature reveals the use of snowball sampling to identify numerous special populations, including homeless adolescents [3], homosexuals [2], minority community leaders [16], cancer survivors [9], drug users [12], current and former smokers [14], and women planning to use artificial insemination techniques [5]. Snowball sampling is also a commonly used method for the identification of social networks in sociometric research [6] and in qualitative studies [1]. Snowball sampling may also be used to generate **control** groups for **program evaluations** by asking program participants to identify persons similar to themselves who are not participating in the program [18].

There are several well-known disadvantages of snowball sampling. Most critical among these is the nonrandom nature of respondent selection at stages s_1 through s_k (and in many cases also at s_0) Although some researchers have attempted to deal with this problem by sampling randomly from among respondents identified at each stage (*see* **Random Sample**), persons embedded in larger social networks will nonetheless have greater probabilities, and more isolated persons will have smaller probabilities, of being referred. Selection of respondents at stages s_1 through s_k are also based on the subjective judgments of informants and may therefore be influenced by numerous considerations not easily assessed or controlled by researchers. Nominating others, particularly in studies of deviant behavior, may also raise concerns of confidentiality and discourage informant candor. Finally, this approach makes the often difficult assumption that members of the population of interest are known to one another.

In response to these limitations, several advances in snowball sampling have been proposed in recent years. Among these are variations designed to estimate the size of hidden populations [7], a "random walk" procedure for examining social networks (*see* **Stochastic Processes**) [13], a "targeted sampling" procedure [20], a "targeted personal network sampling" procedure [17], and "respondent-driven sampling" methodology [10, 11]. The latter approach, proposed by Heckathorn, may under specified circumstances be used to develop **standard errors** for population estimates constructed using this technique.

Snowball sampling is also known for several important advantages that make it an attractive approach in many situations. Perhaps, most importantly, it is a low cost and relatively efficient method for locating hard-to-find individuals. In many settings, snowball techniques can also be deployed to collect data very quickly. As such, it is an effective method for initially exploring phenomena and populations for which there are few parameters available with which to plan more formalized sample designs (*see* **Sample Surveys in the Health Sciences**).

References

[1] Biernacki, P. & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling, *Sociological Methods and Research* **10**, 141–163.

[2] Buenting, J.A. (1992). Health life-styles of lesbian and heterosexual women, *Health Care for Women International* **13**, 165–171.

[3] Clatts, M.C., Davis, W.R. & Atillasoy, A. (1995). Hitting a moving target: the use of ethnographic methods in the

development of sampling strategies for the evaluation of AIDS outreach programs for homeless youth in New York City, in *Qualitative Methods in Drug Abuse and HIV Research*, *NIDA Research Monograph* 157, E.Y Lambert, R.S. Ashery & R.H. Needle, eds. National Institute on Drug Abuse, Rockville, pp. 117–135.

[4] Erickson, B.H. (1978). Some problems of inference from chain data, in *Sociological Methodology 1979*, K.F. Schuessler, ed. Jossey-Bass, San Francisco, pp. 276–302.

[5] Etter, J.F. & Perneger, T.V. (2000). Snowball sampling by mail: Application to a survey of smokers in the general population, *International Journal of Epidemiology* **29**, 43–48.

[6] Frank, O. (1979). Estimation of population totals by use of snowball samples, in *Perspectives on Social Network Research*, P.W. Holland & S. Leinhardt, eds. Academic Press, New York, pp. 319–347.

[7] Frank, O. & Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling, *Journal of Official Statistics* **10**, 53–67.

[8] Goodman, L. (1961). Snowball sampling, *Annals of Mathematical Statistics* **32**, 245–268.

[9] Halstead, M.T. & Fernsler, J.I. (1994). Coping strategies of long-term cancer survivors, *Cancer Nursing* **17**, 94–100.

[10] Heckathorn, D.D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations, *Social Problems* **44**, 174–199.

[11] Heckathorn, D.D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations, *Social Problems* **49**, 11–34.

[12] Kaplan, C.D., Korf, D. & Sterk, C. (1987). Temporal and social contexts of heroin-using populations: An illustration of the snowball sampling technique, *The Journal of Nervous and Mental Disease* **175**, 566–574.

[13] Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities, in *The Small World*, M. Kochen, ed. Ablex, Norwood, pp. 176–210.

[14] Macaulay, L., Kitzinger, J., Green, G. & Wight, D. (1995). Unconventional conceptions and HIV, *Aids Care* **7**, 261–276.

[15] Menzel, H. & Katz, E. (1955–1956). Social relations and innovation in the medical profession: the epidemiology of a new drug, *Public Opinion Quarterly* **19**, 337–352.

[16] Michielutte, R. & Beal, P. (1990). Identification of community leadership in the development of public health education programs, *Journal of Community Health* **15**, 59–68.

[17] Spreen, M. & Zwaagstra, R. (1994). Personal network sampling, outdegree analysis and multilevel analysis: introducing the network concept in studies of hidden populations, *International Sociology* **9**, 475–491.

[18] Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

[19] TenHouten, W.D., Stern, J. & TenHouten, D. (1971). Political leadership in poor communities: applications of two sampling methodologies, in *Race, Change and Urban Society*, P. Orleans & W.R. Ellis, eds. Sage Publications, Beverly Hills, pp. 215–254.

[20] Watters, J.K. & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations, *Social Problems* **36**, 416–430.

TIMOTHY P. JOHNSON

Social Classifications

Social classifications are needed by studies that aim to describe variations in health or health care use according to socioeconomic status (*see* **Health Services Research, Overview**). Other studies, for example on the etiology of specific diseases, need social classifications to control for **confounding** or **effect modification** by socioeconomic variables. There are three core indicators of socioeconomic status: education, income, and occupation [3–5]. Each indicator is an independent predictor of health and health care use and, therefore, each is potentially relevant to studies in medicine and public health.

Education emphasizes differences among people in knowledge, skills, and attitudes. Of all socioeconomic indicators, it is the easiest to measure. The educational level of subjects can be measured as the highest level of education that has successfully been completed. If possible, this measure also takes into account technical and vocational education, and part-time study or training after leaving school. Educational levels can be grouped according to a national, hierarchical classification. Studies among the elderly should take care to distinguish between elementary and lower secondary education. Data coding can largely be avoided by using questions with set answer categories instead of open questions. Even simpler to use are questions on the number of years of education that a person has attended school full-time, or the age at leaving school.

Income level complements educational level by its emphasis on material standards of living. It is preferably measured as the net household income, if possible corrected for household size. Questions should ensure that respondents count the incomes of all household members, and include the most relevant income components such as wages and salaries, interest, pensions, and transfer payments. Income levels are measured most accurately by an extensive battery of questions but might also be approximated by one or a few general questions. In some instances, proxy measures of long-term living standards are preferable, such as the possession of durable consumption goods, house ownership, or the quality of housing. These indicators have the advantage that they are more stable over time, and do not create the problems of **nonresponse** and inaccuracy that are typical for questions on income.

Occupation is the most comprehensive socioeconomic indicator. Unfortunately, the use of occupation as a socioeconomic indicator is laborious. Basic to its measurement is the classification of subjects according to a national three-digit classification of occupational titles. This can be, supplemented with information on employment status (self-employed or in employment) and supervisory status (number of subordinates). It is important to classify economically inactive men and women (unemployed, retired, housewives, etc) according to their last occupation. Married and cohabiting women may also be classified according to the occupation of their partner. Men or women with similar occupations can be classified according to a national social class scheme that distinguishes, among others, professionals and managers, lower nonmanual employees, skilled workers, and unskilled workers [1]. An alternative is to express the socioeconomic status of each occupation by means of their score on one-dimensional status scales [2].

References

[1] Bartley, M., Carpenter, L., Dunnell, K. & Fitzpatrick, R. (1996). Measuring inequalities in health: an analysis of mortality patterns using two social classifications, *Sociology of Health and Illness* **18**, 455–475.

[2] Ganzeboom, H.B.G., Graaf, P.M. de & Treiman, D.J. (1992). A standard international socioeconomic index of occupational status, *Social Science Research* **21**, 1–56.

[3] Kunst, A.E. & Mackenbach, J.P. (1994). *Measuring Socio-Economic Inequalities in Health*. World Health Organization, Copenhagen.

[4] Liberatos, P., Link, B.G. & Kelsey, J.L. (1988). The measurement of social class in epidemiology, *Epidemiologic Reviews* **10**, 87–121.

[5] Moss, N. & Krieger, N. (1995). Measuring social inequalities in health. Report on the Conference of the National Institutes of Health, *Public Health Reports* **110**, 302–305.

ANTON E. KUNST

Social Sciences

Ideally, statistics should be a language, spoken across all the sciences, which could ease communication among disciplines, prevent duplication, and encourage research that spans disciplines. However, the types of data encountered in the social sciences are in many ways different from others. For example, it is widely acknowledged that the effects of situational variation among studies is large in the social sciences. Because of differences like this, particular concerns come to the surface in different disciplines at different times. Statistics has evolved into different dialects through the different branches of science, and statistical concerns of one group are not always discussed with others. There are four major issues of particular concern in the social sciences: **teaching statistics**, **hypothesis testing**, **categorical data analysis**, and **multilevel modeling**.

Teaching Statistics

Nobody should doubt the need for statistics and methodology training in undergraduate programmes of any science. Most agree that methodology underlies the empirical basis of any stochastic science and that statistical knowledge is a fundamental aspect of this. However, debates exist on how students should be taught. Should students be taught statistics as a set of *tools* for handling particular problems or should they be taught the underlying concepts of statistics?

While this question arises for all people concerned about students' statistical training, there is a difference between the social sciences and some other sciences with regard to the amount of prerequisite mathematics. In the social sciences we cannot assume that our students have taken, for example, calculus. In fact, we can assume that many will particularly dislike anything to do with numbers. Because of this, many textbooks present each statistical test as an unrelated **algorithm** that can passively be applied in particular situations without any conceptual understanding. Similarly, with advances in statistical computing, and the increase in "user-friendly" programs, some advocate teaching "how to run statistical tests on the computer" instead of teaching any statistical concepts. This contrasts with people who claim that the underlying concepts must be learned before they should be applied to different problems. My own view [19] is that at least some of the conceptual issues must be taught, although care must be taken not to make unrealistic demands on students' mathematical expertise.

Hypothesis Testing

Not only within psychology, but across all the social sciences, **null hypothesis** significance testing (NHST) (*see* **Hypothesis Testing**) is often used. It has been known for some time that there are conceptual and logical problems with this approach. Cohen [3] suggests that the phrase "statistical hypothesis inference testing" would yield a more appropriate acronym. The problems are particularly detrimental to the social sciences. Meehl [13] wrote that NHST is "one of the worst things that ever happened in the history of psychology". He noted [12] a difference between the physical and social sciences that he claimed was why NHST has had such detrimental effects on the social sciences. His claim was that, in the physical sciences, the null hypotheses which are put forward will often be based on a particular model that the scientists actually believe. Deviations from this substantive model allow it to be falsified in a Popperian sense (*see* **Popper, Karl R.**). In the social sciences, the null hypothesis is often a "strawman" hypothesis, like whether the relationship between social class and voting preference has changed over the years. There is no doubt that changes have taken place; it is the magnitude and direction of changes that are of substantive interest. In both physical and social sciences improved methods are increasing precision. In the physical sciences this means that more substantive models can be properly falsified. In the social sciences, as the methods become more precise, more "strawman" hypotheses are rejected. This counters the importance of falsification in theory development and validity.

Improvements have been suggested, including testing interval hypotheses rather than point hypotheses, reporting effect sizes and **confidence intervals**, considering the **power** of statistical tests, and, most importantly, testing the substantive models of interest.

Categorical Data

The work of Goodman [7, 8] and others (for example, Clogg & Shihadeh [2]) has made **loglinear modeling**

one of the most widely used techniques in the social sciences. Loglinear modeling is principally used when examining the relationships among multiple categorical – sometimes called qualitative – variables, although quantitative variables can be incorporated into the approach. In the social sciences, where assuming any sort of quantitative metric is often unjustified, the incorporation of loglinear procedures for nominal and ordinal variables into the mainstream statistics packages (*see* **Software, Biostatistical**) was welcomed.

The first extensions involved different ways of partitioning the chi-square variation (*see* **Chi-square, Partition of**) of a loglinear model. The fact that the independence model does not hold for a typical $r \times c$ **contingency table** supplies very little practical information. There are $(r-1)(c-1)$ **degrees of freedom** for the **residuals** of this model, any combination of which could account for the dependence. Researchers sought models that incorporated some of these degrees of freedom into a model that fit the data more adequately, but still were more parsimonious than the independence model. A lot of this work has been done with **square contingency tables**, where the row variable is the same as the column variable, except at a different time or different situation. An example would be comparing sons' and fathers' social class. The fact that they are related is so obvious to be uninteresting. However, how they are related, and in particular where movement off the diagonal occurs, is interesting. Goodman [8] provides a thorough discussion of the techniques for square tables.

Another strand of work in the 1990s has involved extending the loglinear model to the logmultiplicative, or the RC, model (see [2] for an introduction). This is of a class of models that attempt, in a sense, to quantify qualitative data. Assume a two-variable contingency table where the row values (R) and column values (C) are categorical. The RC model estimates new quantifications for each value so as to maximize the quantitative (linear) **correlation** between the two variables. The RC model can be extended for multiple quantifications or dimensions for each variable in a manner similar to the way in which classical **principal components analysis** operates with interval variables. Using Clogg & Shihadeh's notation, these models are called RC(M) for the M quantifications. They demonstrate the value of this approach, examining the relationship between years of schooling and occupation.

The growing realization in the value of graphing data has occurred for every type of statistics, including categorical data (*see* **Graphical Displays**). One particular approach has used RC(M) models. Often under the general title of **correspondence analysis**, these procedures (see, for example, [15] and [16] for details) produce graphical displays of **association**. There is some controversy within the social sciences as to whether these methods can replace some of the more formal model fitting and confidence interval approaches, or if they should be used to complement each other [8].

Consider the following example from Gaskell et al. [4]. They were interested in differences in age and gender for "what comes to mind when science is mentioned". They had six age categories and four categories for the field of science mentioned. When they asked a sample of about 2000 people in the UK, the model {age × gender, field} produced a residual of $\chi^2(33) = 160$. This shows that there are age and/or gender differences. Figure 1 partitions this deviation into two dimensions (see [17] for the method used here). It shows that males are more inclined to think about technology (mostly computers and engineering) and that older people are less likely to respond with "physics" (the "other" included many responses about environmental issues). Also,

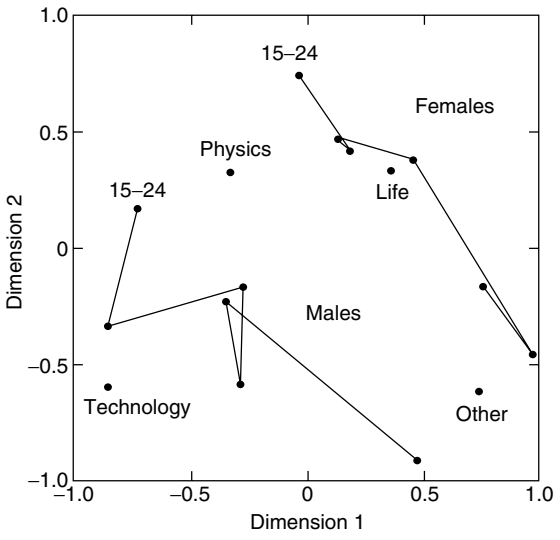

Figure 1 A two-dimensional correspondence analysis solution using a 12-category variable for gender by age. Reproduced from Wright [19], Figure 8.9, by permission of Sage Publications Ltd, 1997

there appears to be no **interaction** between age and gender on the responses {the model[age × field, age × gender, gender × field] produces a satisfactory fit [$\chi^2(15) = 20.43$, $p = 0.16$]}.

The final topic with respect to categorical analysis is its application to latent structures and **path analysis**. Different phrases are used for each of the possibilities with respect to latent (i.e. unobserved) and manifest (i.e. observed) variables being continuous or categorical. **Latent class analysis** is where latent and manifest variables are qualitative. Latent trait analysis assumes continuous latent variables but discrete manifest variables. Latent profile analysis has continuous manifest variables but discrete latent variables. Classical **factor analysis** or **LISREL** are for cases with continuous latent and manifest variables. These models have now been incorporated into path models (see [11] and [9]).

Multilevel Models

Multilevel modeling is where the data are clustered (*see* **Clustering**) or nested. For example, if you are doing research on children, you might go to many schools, visit many classrooms, and receive data from many children. An assumption in most introductory statistics textbooks is that the children are independent and identically distributed. When this is not true the **standard errors** are most often too small and therefore the estimates appear more precise than they should (e.g. [14]). Multilevel modeling accounts for this by allowing **random variables** at each level. It requires some strong assumptions about the distributions of these variables [1]. With these assumptions, some useful models about the **interactions** among groups and individuals are possible. While these models are used in biostatistics, they have become particularly popular in education and in survey research (where the geographical area or the household are often used as the higher level unit). With both of these, the structure of the hierarchy and its importance are recognized.

Goldstein [6] describes how these hierarchies occur almost everywhere you look and that ignoring the problems of independence, and more importantly not describing the model properly, will often lead to errant conclusions. These models are a particular class of **random coefficient** models. One of the reasons for the popularity of the multilevel approach is because of advantages for some estimation purposes. Kreft et al. [10] describe several specialized programs for this situation.

My own interest in multilevel models arose when analyzing the data from police eye-witness lineups in the Greater London area [21]. Many of the suspects were viewed by multiple witnesses, and hence we had witnesses nested within suspects. The response variable was categorical; witnesses could choose the suspect, a known-innocent person "picked off the street", or make no identification. When we started this work the algorithms were just being developed for multilevel modeling with categorical variables [5]. Now they are part of one of the most popular of the multilevel modeling packages, ML*n* [18].

Another aspect of multilevel modeling that is becoming of more interest is what happens when there are relatively few cases in a cluster. This happens, for example, when the hierarchy is family members nested within a household. There will be many households with only one or two people. We faced a similar situation with our witness data. Many suspects had only one or two witnesses nested within them. When we originally analyzed these data, we found extra **multinominal** variation. This is often seen as an indicator that the model and/or the hierarchical structure has been misspecified. This worried us and caused us to add various caveats to our reports on that project. More recently [20], I used **simulation** methods and found that, even when the model and structure were perfectly specified, this *sparsity* led to extra **binomial** variation (*see* **Overdispersion**). This demonstrates the importance of research on the structure of the hierarchies. While it was a big step to move from single-level analysis to multilevel analysis, the future is likely to show that not all hierarchies are the same.

References

[1] Chance, B. (1996). Hierarchical model behavior and estimation, Paper presented at the *Fourth International Social Science Methodology Conference*, July. Essex, UK.

[2] Clogg, C.C. & Shihadeh, E.S. (1994). *Statistical Models for Ordinal Variables*. Sage Publications, London.

[3] Cohen, J. (1994). The Earth is round ($p < .05$), *American Psychologist* **49**, 997–1003.

[4] Gaskell, G.D., Wright, D.B. & O'Muircheartaigh, C.A. (1993). Measuring scientific interest: the effect of knowledge questions on interest ratings, *Journal for the Public Understanding of Science* **2**, 39–57.

[5] Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data, *Biometrika* **78**, 45–51.

[6] Goldstein, H. (1995). *Multilevel Statistical Methods*, 2nd Ed. Edward Arnold, London.

[7] Goodman, L.A. (1978). *Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent-Structure Analysis*. Addison-Wesley, London.

[8] Goodman, L.A. (1991). Models, measures, and graphical displays in the analysis of contingency tables (with discussion), *Journal of the American Statistical Association* **86**, 1085–1138.

[9] Hagenaars, J.A. (1993). Loglinear Models with Latent Variables, *Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07–094*. Sage, Newbury Park.

[10] Kreft, I.G.G., de Leeuw, J. & van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, and VARCL, *American Statistician* **48**, 324–335.

[11] McCutcheon, A.L. (1987). Latent Class Models, *Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07–064*. Sage, Newbury Park.

[12] Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox, *Philosophy of Science* **34**, 103–115.

[13] Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology, *Journal of Consulting and Clinical Psychology* **46**, 806–834.

[14] Scariano, S. & Davenport, J. (1987). The effects of violations of the independence assumptions in the one way ANOVA, *American Statistician* **41**, 123–129.

[15] Van der Geer, J.P. (1993). *Multivariate Analysis of Categorical Data: Applications*. Sage Publications, London.

[16] Van der Geer, J.P. (1993). *Multivariate Analysis of Categorical Data: Theory*. Sage Publications, London.

[17] Van der Heijden, P.G.M., de Falguerolles, A. & de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis (with discussion), *Applied Statistics* **2**, 249–292.

[18] Woodhouse, G., ed. (1995). A Guide to MLn for New Users, *Multilevel Models Project*. Institute of Education, University of London.

[19] Wright, D.B. (1997). *Understanding Statistics: Introduction to Statistics for the Social Sciences*. Sage Publications, London.

[20] Wright, D.B. (1997). Extra-binomial variation in multilevel logistic models with sparse structures, *British Journal of Mathematical and Statistical Psychology* **50**, 21–29.

[21] Wright, D.B. & McDaid, A.T. (1996). Comparing system and estimator variables using data from real lineups, *Applied Cognitive Psychology* **10**, 75–84.

DANIEL B. WRIGHT

Society for Clinical Trials

The Society for Clinical Trials, established in 1978, grew out of a need for professionals working in the field of clinical trials to exchange ideas, experiences and information. The general purpose of the Society for Clinical Trials, as stated in its by-laws, is "to promote the development and exchange of information for design and conduct of clinical trials and research using similar methods" [6]. Also, it is stated in the by-laws that "the Society shall serve as a forum for discussion of philosophical, ethical, legal, and procedural issues involved in the design, organization, operation, and analysis of clinical trials and epidemiological studies using similar methods". The Society's long-term objectives are:

1. Promotion of methodological research emphasizing design, organization, operation, and analysis.
2. Promotion of the application of sound principles of design, organization, and operation through workshops and meetings sponsored by the organization. Some of these workshops and meetings may be international in character and held in countries other than the United States.
3. Promotion of communication by development, where possible, of standard terminology.
4. Promotion of better understanding to those entering the field by serving as a resource for the design and conduct of these studies.
5. Promotion of better communication through the development of standards for the analysis and reporting of results.
6. Promotion of better understanding by the general public of the importance of clinical trials for the evaluation of health care procedures [6].

The feasibility of conducting randomized clinical trials was successfully demonstrated in Great Britain during the late 1940s under the leadership of Sir Austin Bradford Hill, who also explicated the principles and methods of randomized trials in numerous writings. Between the 1950s and 1970s there was a sustained growth in randomized clinical trials, most notably in the English-speaking world. This growth, which gradually involved an increasing number of clinical specialties and professional disciplines, was accompanied by an evolution of concepts, strategies and methodologies for the design and conduct of clinical trials (*see* **Clinical Trials, Overview**). The clinicians, biostatisticians, nurses and other health professionals who had been involved in these studies from the start were gradually joined by epidemiologists, computer scientists, clinic coordinators, and ethicists, as well as by professionals from ancillary disciplines like nutrition, behavioral psychology, and management science. The evolving concepts, strategies, and methods of clinical trials entailed a broad range of topics: ethics, design, **randomization**, data collection, assurance of data quality, patient recruitment, data analysis, **data and safety monitoring**, organization and management of multicenter studies (*see* **Data Management and Coordination**), and study closeout.

To some extent, of course, there was a sharing of ideas and information during these decades. Clinical trials topics were discussed at meetings of specialty groups and articles were appearing in statistical and clinical journals, but there was no forum for the formal and informal sharing of ideas among different specialties and disciplines. As a result, technology transfer between specialties was limited.

Many of the persons who had come to devote their careers to clinical trials were working in **multicenter trials**, and these individuals provided an important stimulus to the formation of a clinical trials society. A series of annual Symposia for Coordinating Center Personnel (Cardiovascular Trials) was organized by Curtis Meinert in 1973 and continued from 1975 to 1980. As implied by the name, the content of these well-attended symposia, held in various locations in the US, was directed mainly at those concerned with the coordination of multicenter studies: biostatisticians, computer scientists, study coordinators, and data managers. The symposia were meeting an obvious need for the sharing of ideas and information and it soon became apparent that personnel and topics beyond those concerned with the coordination of multicenter studies of cardiovascular trials should be included.

In 1976, Fred Ederer, Curtis Meinert and Dale Williams – later joined by Harold Roth – proposed the idea of forming a national or international clinical trials society to the Clinical Trials Committee of the National Institutes of Health (NIH), asking for the Committee's support in the effort. The Committee expressed interest, but, lacking evidence for widespread participatory support for a society, suggested holding a conference to test that support. Accordingly, the National Conference on Clinical

Trials Methodology, sponsored by the NIH, was held in October 1977; its proceedings were published in May 1979 [4]. Attendance of that conference by more than 700 persons, far more than expected, confirmed the need for a professional society focused on clinical trials.

An *ad hoc* Board of Directors (Harold O. Conn, Thomas C. Chalmers, Fred Ederer, Robert S. Gordon, Christian R. Klimt, Paul Meier, Curtis L. Meinert, Charles Moertel, Thaddeus Prout, Harold P. Roth, and O. Dale Williams) drew up by-laws and incorporated the organization under the name Society for Clinical Trials. Under the Presidency of Harold Roth, the Society's first annual meeting was held in Philadelphia, 6–8 May 1980 [3, 5].

Coincidental with the formation of the Society for Clinical Trials was the publication of a new journal, ***Controlled Clinical Trials***, under the editorship of Curtis Meinert [1, 2]. Its objective was to meet many of the same needs – in written rather than oral communication – as the Society. The logical next step was taken: *Controlled Clinical Trials* was designated the official journal of the Society of Clinical Trials. The program and abstracts of each annual meeting are published in *Controlled Clinical Trials*.

Two joint meetings of the Society for Clinical Trials and the International Society for Clinical Biostatistics have been held: one in July 1991 and one in July 1997. A third joint meeting is scheduled for July 2003.

Additional information about the Society for Clinical Trials is available through its website [`http://www.sctweb.org`] or the Society's journal, *Controlled Clinical Trials*.

References

[1] Meinert, C.L. (1980). Why another journal?, *Controlled Clinical Trials* **1**, 1–2.

[2] Meinert, C.L. & Tonascia, S. (1998). Controlled clinical trials, in *Encyclopedia of Biostatistics*, Vol. 6, P. Armitage & T. Colton, eds. Wiley, Chichester.

[3] Roth, H.P. (1980). On the Society for Clinical Trials, *Controlled Clinical Trials* **1**, 81–82.

[4] Roth, H.P. & Gordon, R.G., Jr (1979). Proceedings of the National Conference on Clinical Trials Methodology, *Clinical Pharmacology and Therapeutics* **25**(5), part 2.

[5] Society for Clinical Trials (1980). Abstracts of the Combined Annual Scientific Sessions of the Society for Clinical Trials and the Seventh Annual Symposium for Coordinating Clinical Trials, *Controlled Clinical Trials* **1**, 167–180.

[6] Society for Clinical Trials, Inc. (1980). By-Laws, *Controlled Clinical Trials*, **1**, 83–89.

GENELL L. KNATTERUD

Software for Clinical Trials

In attempting to summarize and contrast the available clinical trials software, one immediately encounters several difficulties. First, there is not a unique set of analytical techniques for clinical trials. There are some general characteristics of most clinical trials that do result in some restriction in the statistical tests likely to be employed in the analysis. For example, recommendations for the conduct of a clinical trial include randomization of patients, sufficient sample size to have adequate statistical power of answering the proposed clinical question, and a simple, easily interpretable clinical measure as the primary outcome [13, 14] (*see* **Outcome Measures in Clinical Trials**). These guidelines usually make it less likely that methods used primarily for identifying causal relationships (such as structural equations), computationally intensive techniques used for small sample sizes, or more complex analysis of variance (ANOVA) designs will be used. However, even with some general restrictions, the wide diversity of potential clinical outcomes in the biomedical area is greater than is found in presently existing specialized software packages and often leads to utilization of the more commonly employed general statistical packages. A treatment of general biostatistical software is described in [5].

A second problem is that the available software is constantly changing. This is an advantage to the users since there is a constant source of new tools available to address their needs. However, any comparison by the reviewer contrasting advantages and limitations of various software rapidly becomes outdated. A third problem in contrasting available software is that an individual's personal preference, general background, and the group of packages with which the individual is familiar often affect the selection. Given these problems, it is not the intent of this article to attempt to provide an exhaustive list of all available software for clinical trials or present detailed evaluations of the strengths and weaknesses of various packages. Instead, we will provide an overview and inventory of some of the available software that a researcher might consider useful in conducting and analyzing a clinical trial. We have focused mostly on software that is publicly available with a proven track record. We will also discuss some of the cautions in using this software.

In identifying clinical trials software, we considered the available options from three broad categories: (1) software used in the design phase of clinical trials, (2) software used in the tracking or conduct of the day-to-day operation of the clinical trial including data collection, and (3) the analysis and reporting of the data. The last category would not only include the statistical analysis but also the presentation of the results in tabular, numerical, and graphical forms.

There has been a gradual change over time in the availability of new packages and the expanded capabilities of existing software packages. Many vendors have focused on making their packages run more efficiently and reliably in the current operating systems. The execution time of some analyses has been drastically reduced and computations that could not be performed because of software limitations may now be tractable. Many vendors have enhanced their software in its ability to import and export data files in formats that are not native to the package, thus providing for easier data exchange between packages. More vendors are now addressing the criticisms that the output from analyses, although technically correct and complete, is often too poorly formatted to be included in reports. New options and presentation formats are appearing in software to facilitate the integration of analytical output into word processed documents, for example, SAS output can be saved as MS Word and Adobe PDF files, in addition to the standard ASCII output files.

Table 1 summarizes some of the available software packages with the corresponding vendor and contact information. The computer environment (i.e. Mac, DOS, Windows) is not given since this is often version-dependent and is constantly changing. The general-purpose packages include many procedures that are seldom used in clinical trials. Conversely, given the diversity of potential outcomes in a clinical trial, it is useful to be familiar with more than one general package, since often, packages are more highly developed in some areas than in others. Detailed comparisons of the specific statistical methods available on different packages are of limited usefulness since the major packages are constantly being updated. For example, version 10 of MINITAB had procedures neither for actuarial life tables nor for logistic regression making it an unacceptable package for several important outcome variables commonly

Table 1 Summary of software which is useful in design, conduct, and analysis of clinical trials

Title	Emphasis relevant to clinical trials	Vendor
General purpose packages		
BMDP	Parametric and nonparametric survival analysis (right, left, and interval censored data; accelerated life testing; exponential, proportional hazards, Weibull, extreme value, logistic, lognormal, log-logistic regression; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; dichotomous and polychotomous logistic regression.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com
BMDP New System Professional	Functionality of BMDP with a Windows interface program. Parametric and nonparametric survival analysis (right, left, and interval censored data; accelerated life testing; exponential, proportional hazards, Weibull, extreme value, logistic, lognormal, log-logistic regression; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; dichotomous and polychotomous logistic regression; missing data analyses techniques (mean imputation, hot-deck imputation, last value carried forward, predicted mean imputation).	Statistical Solutions Stonehill Corporation Center Suite 104, 999 Broadway, Saugus, MA 01906 [781] 231-7680 [800] 262-1171 http://www.statsolusa.com
GB STAT	Parametric and nonparametric survival analysis (proportional hazards, Kaplan–Meier; log-rank procedures); repeated measures ANOVA; logistic regression.	Dynamic Microsystems, Inc. 13003 Buccaneer Road Silver Spring, MD 20904 [301] 384-2754 http://www.gbstat.com
Genstat	Parametric and nonparametric survival analysis (right, left, and interval censored data; accelerated life testing; exponential, proportional hazards, Weibull, extreme value; Kaplan–Meier; log-rank procedures); logistic regression.	NAG Ltd Wilkinson House Jordon Hill Road Oxford OX2 8DR, UK +44 1865 511245 http://www.nag.co.uk
JMP	Parametric and nonparametric survival analysis (proportional hazards model, exponential, extreme value, lognormal, Weibull; Kaplan–Meier; log-rank procedures, competing causes and recurrence analysis); ANOVA/MANOVA; logistic regression.	SAS Institute Inc. SAS Campus Drive Cary, NC 27513 [919] 677-8000 http://www.jmp.com http://www.sas.com

MINITAB	Parametric and nonparametric survival analysis (right, left, arbitrary and interval censored data; accelerated life testing; exponential, Weibull, extreme value, logistic, lognormal, log-logistic; Kaplan–Meier; log-rank procedures); probit analysis; logistic regression	Minitab, Inc. 3081 Enterprise Drive State College, PA 16801-3008 [814] 238-3280 http://www.minitab.com
NCSS	Parametric and nonparametric survival analysis (right, left, and interval censored data; accelerated life testing; exponential, proportional hazards, Weibull, extreme value, logistic, lognormal, log-logistic; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; Probit analysis; logistic regression	NCSS Statistical Software 329 North 1000 East Kaysville, UT 84037 [801] 546-0445 [800] 898-6109 http://www.ncss.com
S-PLUS	Parametric and nonparametric survival analysis (right, left, and interval censored data; accelerated life testing; exponential, proportional hazards, Weibull, lognormal, log-logistic; Kaplan–Meier; log-rank procedures; penalized survival models; smoothing splines; person-years analysis); repeated measures ANOVA; MANOVA; logistic regression	Insightful Corp. 1700 Westlake Avenue North, Suite 500 Seattle, WA 98109 [206] 283-8802 [800] 569-0123 http://www.splus.com http://www.insightful.com
SAS	Parametric and nonparametric survival analysis (right, left, and interval censored data; accelerated life testing; exponential, proportional hazards, Weibull, lognormal, log-logistic regression; Kaplan–Meier; log-rank procedures); probit analysis; repeated measures ANOVA; MANOVA; logistic regression.	SAS Institute Inc. SAS Campus Drive Cary, NC 27513 [919] 677-8000 http://www.sas.com
SPSS	Parametric and nonparametric survival analysis (proportional hazards; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; probit regression; logistic regression.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com
Stata	Parametric and nonparametric survival analysis (right and left censored data; accelerated life testing; exponential, proportional hazards, Weibull, lognormal regression; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; probit analysis; dichotomous and polychotomous logistic regression. Also has facilities for analysis of correlated data resulting from cluster designs (e.g. multiple outcomes per patient, repeated outcomes per patient, cluster randomization, and nested sampling schemes).	Stata Corporation 702 University Drive East College Station, TX 77840 [800] 782-8272 http://www.stata.com

(*continued overleaf*)

Table 1 (*continued*)

Title	Emphasis relevant to clinical trials	Vendor
STATGRAPHICS *Plus*	Parametric and nonparametric survival analysis (Weibull model; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; logistic regression.	Manugistics, Inc. 2115 East Jefferson Street Rockville, MD 20852 [800] 592-0050 http://www.statgraphics.com
STATISTICA	Parametric and nonparametric survival analysis (exponential, Gompertz, Weibull, proportional hazards, lognormal, normal; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; probit analysis; logistic regression	StatSoft 2300 East 14th Street Tulsa, OK 74104 [918] 749-1119 http://www.statsoft.com
StatView	Parametric and nonparametric survival analysis (exponential, Weibull, proportional hazards, lognormal, log-logistic; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; dichotomous and polychotomous logistic regression	SAS Institute Inc. SAS Campus Drive Cary, NC 27513 [919] 677-8000 http://www.sas.com http://www.statview.com
SYSTAT	Parametric and nonparametric survival analysis (right, left, and interval censoring; proportional hazards; exponential, accelerated exponential, Weibull, lognormal and logistic; Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; probit analysis; logistic regression	Systat Software Inc., 501 Suite 'C', Point Richmond Tech Center, Canal Boulevard, Richmond, CA 94804-2028 [800]797-7401 http://www.systat.com
True Epistat	Parametric and nonparametric survival analysis (proportional hazards; exponential, Kaplan–Meier; log-rank procedures); repeated measures ANOVA; MANOVA; logistic regression	Tracy L. Gustafson Epistat Services 280 W. Renner Road, #2112 Richardson, TX 75080 [972] 994-0904 http://www.true-epistat.com

Graphics packages

Axum	Comprehensive graphing package including: step function plots and error bars.	Math Soft 1700 Westlake Avenue North, Suite 500 Seattle, WA 98109 [206] 283-8802 [800] 569-0123 http://www.mathsoft.com
CrossGraphs	Comprehensive graphics package including: Kaplan–Meier plots and patient summary charts and graphs.	PPD, Inc. 3151 South 17th Street Wilmington, North Carolina 28412 [910] 251 0081 http://www.ppdi.com
DeltaGraph	Comprehensive graphics package including: Kaplan–Meier plots and error bars.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com
GraphExpress	Comprehensive graphing tools designed as an interface for producing graphs from SAS including: Kaplan–Meier plots and error bars.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com
SigmaPlot	Comprehensive graphics package including: Kaplan–Meier, probit, and logit plots, and error bars.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com

(*continued overleaf*)

Table 1 (*continued*)

Title	Emphasis relevant to clinical trials	Vendor
Meta-analysis		
Advanced BASIC Meta-Analysis	Meta-analysis	Brian Mullen Syracuse University Department of Psychology [315] 443-2354 http://psychweb.syr.edu/Faculty/drmullen.htm
Comprehensive Meta-Analysis	Effect size measures (standardized differences, correlation coefficients, odds ratios, relative risks, and risk differences). Fixed and random effects models.	Biostatistical Programming Associates Biostat, Inc. 14 North Dean Street Englewood, NJ 07631 [201] 692-8155 http://www.meta-analysis.com
DSTAT	Effect size measures (correlation coefficients, standardized mean difference, t-test statistics, F-test statistics, proportional data, chi-squares, P values). Fixed effects model.	Blair T. Johnson University of Connecticut Department of Psychology, U-20 406 Babbidge Road Storrs, Connecticut 06269-1020 [860] 486-2511 http://johnson.socialpsychology.org/
Epi Meta	Effect size measures (relative risks). Fixed and random effects models.	Centers for Disease Control Statistical and Epidemiology Branch Division of Prevention Research and Analytic Methods [404] 639-3806 http://www.cdc.gov/epo/dpram/epimeta/epimeta.htm
Hunter and Schmidt programs	Effect size measures (correlation coefficients, standardized mean differences) methods consistent with [8].	John E. Hunter Michigan State University Frank L. Schmidt

		University of Iowa Available in back of [8] book and available from Frank L. Schmidt, University of Iowa.
Meta-Analysis programs	Effect size measures (standardized differences correlations and differences between proportions). Fixed and random effects models.	Ralf Schwarzer, Freie Universität Berlin, Germany, University, Toronto, Canada http://www.fu-berlin.de/gesund/gesu_engl/meta_e.htm
META (Meta-analysis easy to answer).	Effect size measures (standardized differences, correlations, and difference between proportions). Fixed effects or random effects models.	David Kenny University of Connecticut http://nw3.nai.net/dakenny/meta.htm.
Meta-analyst	Effect size measures (odds ratios, relative risks, and risk ratios). Fixed and random effects models.	Joseph Lau, MD New England Medical Center http://joseph.lau@cs.nemc.org
SAS Macros	Effect size measures (probability values, standardized mean differences, correlation coefficients, odds ratios, and vote counting methods). Fixed and random effects models.	Morgan C. Wang University of Central Florida Brad J. Bushman Iowa State University http:// www.sas.com
TRUE EPISTAT	Effect size measures (standardized differences, correlation coefficients, odds ratios, relative risks, and risk differences). Fixed and random effects models.	Tracy L. Gustafson Epistat Services 280 W. Renner Road, #2112 Richardson, TX 75080 [972] 994-0904 http://www.true-epistat.com
Design, power, and sample size		
Egret-Siz	Specialty: comprehensive evaluation for Cox proportional hazards, logistic (unmatched and conditional), and Poisson regression models. Handles adjustments for ancillary variables (e.g. confounders). Testing is under the framework of likelihood ratio testing.	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617] 661-2011 http://www.cytel.com

(*continued overleaf*)

Table 1 (*continued*)

Title	Emphasis relevant to clinical trials	Vendor
N and NSURV	General package including: exponential survival (accrual and dropouts), logrank test for user-specified accrual, hazard, and dropout rates. Equivalence tests. Logistic regression.	**idv**-Data Analysis and Study Planning Wessobrunner Straβe 6 D-82131 Gauting/Germany phone: 089/850 80 01 e-mail: idvGauting@aol.com
nQuery Advisor	General package including: exponential survival (accrual and dropouts), logrank test (simulation) for user-specified accrual, hazard, and dropout rates; equivalence and bioequivalence tests - paired design, two group and crossover designs testing difference in means and ratio of means; ANOVA; logistic regression; fold-change analysis (with optional fold-change threshold for DNA microarray studies).	Statistical Solutions Stonehill Corporation Center Suite 104, 999 Broadway, Saugus, MA 01906 [781] 231-7680 [800] 262-1171 `http://www.stat-solusa.com`
PASS	General package including: exponential survival (accrual and dropouts), logrank test for user-specified accrual, hazard, and dropout rates. Group sequential analysis; ANOVA; repeated measures ANOVA; logistic regression.	NCSS Statistical Software 329 North 1000 East Kaysville, UT 84037 [801] 546-0445 [800] 898-6109 `http://www.ncss.com`
Power and Precision	General package including: exponential survival (accrual and dropouts), logrank test for user-specified accrual, hazard, and dropout rates. Equivalence tests. ANOVA; logistic regression.	Biostatistical Programming Associates Biostat, Inc. 14 North Dean Street Englewood, NJ 07631 [201] 692-8155 `http://www.power-andprecision.com`
STATISTICA Power Analysis	General package including: exponential survival; logrank test for user-specified accrual period and dropouts rates; ANOVA	StatSoft 2300 East 14th Street Tulsa, OK 74104 [918] 749-1119 `http://www.statsoft.com`
UnifyPow	General package including: logistic, log-linear, logrank tests, and Cox survival models. An SAS module/macro, which runs in base SAS and creates SAS datasets containing the results.	Ralph O'Brien Department of Biostatistics and Epidemiology Cleveland Clinic Foundation

		http://www.bio.ri.ccf.org/power.html
Report generating		
ClinPlus	A variety of SAS macro-based modules designed to provide data entry and management, statistical analysis and reports, and patient follow-up reports consistent with the procedures of the United States Food and Drug Administration (USFDA). The modules are applicable to the needs of new drug applications and adverse drug event reporting. The modules include Data Management, Coding, Remote NDA, ADE, and Report.	DZS Software Solutions, Inc. 1661 Route 22 West Bound Brook, NJ 08805 [732] 356-6961 http://www.dzs.com
Patient Profiles	Data visualization of patient demographics, medical profile, intervention, and follow-up. Includes an assortment of reports with various display options and summarizations. Interfaces with many different statistics, database management, and spreadsheet programs. Includes ASCII format and ODBC (Open DataBase Connectivity) support. Can simultaneously combine information from multiple databases with different formats.	Statistical Solutions Stonehill Corporation Center Suite 104, 999 Broadway, Saugus, MA 01906 [781] 231-7680 [800] 262-1171 http://www.stat-solusa.com
PPD Patient Profiles	Data visualization of patient demographics, medical profile, intervention, and follow-up. Includes an assortment of reports with various display options and summarizations. Interfaces with many different statistics, database management, and spreadsheet programs. Includes ASCII format support.	PPD, Inc. 3151 South 17th Street Wilmington, North Carolina 28412 [910] 251 0081 http://www.ppdi.com
Data/information entry		
Ascent Capture	Document capture application integrating batch scanning, image processing, optical character recognition (OCR), and document indexing of data and text.	Kofax Image Products 16245 Laguna Canyon Rd. Irvine, CA, 92618-3603 [949] 727-1733 http://www.kofax.com/
BMDP Data Entry	Comprehensive data entry system including double entry (two pass verification), content verification, and comprehensive edit checking. Generates ASCII data files.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com
DataFax	A fax-based data management system for data capture, forms processing and extraction, and direct data entry. Designed for management of clinical trials.	Clinical DataFax Systems Inc. 21 King Street West, Suite 305 Hamilton, Ontario, Canada L8P 4W7, [905] 522-3282 http://www.datafax.com

(*continued overleaf*)

Table 1 (*continued*)

Title	Emphasis relevant to clinical trials	Vendor
EntryPoint, FALCON	Comprehensive data entry system including double entry (two pass verification), content verification, and comprehensive edit checking.	Phoenix Software International West Century Boulevard Suite 800 Los Angeles, CA 90045 [800] 622-9292 http://www.phoenix-software.com/
FaxWare	Fax transmission and receiving software with translational capabilities to convert image documents to data files and electronic documents.	Tobit Software Limited Redwither Tower Redwither Business Park Wrexham LL13 9XT +44 1978 666900 http://www.uk.tobit.com [514] 392-9220 http://www.na.tobit.com
ImagEntry, VDE, Quantum 2000	Comprehensive data entry system including double entry (two pass verification), content verification, and comprehensive edit checking. Automated document scanning and data transfer.	Viking Software Services, Inc. 6804 South Canton Avenue, Suite 900 Tulsa, OK 74136-3419 [918] 491-6144 [800] 324-0595 http://www.viking-soft.com
Input*Accel*	Data capture, image capture, forms processing, data extraction and PDF conversion. Can handle information that arrives in the form of paper, faxes, or microfilm, and so on. User-entered digital information, such as data entered via a web site.	Captiva Software Corp. 10145 Pacific Heights Boulevard San Diego, CA 92121 [858] 320-1000 http://www.captiva-software.com
Key Entry III	Comprehensive data entry system including double entry (two pass verification), content verification, and comprehensive edit checking. Provides for monitoring and reporting operator activity and performance. Generates ASCII data files.	Scan-Optics Inc. 169 Progress Drive Manchester, CT 06040-2294 [860] 645-7878 [800] 745-6001 http://www.scanoptics.com/solutions_products.html

Rode/PC	Comprehensive data entry system including double entry (two pass verification), content verification, and comprehensive edit checking.	DPX/IDEAS, Inc. P.O. Box 7657 Menlo Park, CA 94026 [650] 233-9300 http://www.rodepc.com/
SAS/FSP	SAS module for comprehensive data entry system including ability for double entry (two pass verification), content verification, and comprehensive edit checking.	SAS Institute Inc. SAS Campus Drive Cary, NC 27513 [919] 677-8000 http://www.sas.com
SPSS Data Entry	Comprehensive data entry system including double entry (two pass verification), content verification, and comprehensive edit checking. Provides for monitoring and reporting operator activity and performance. Generates ASCII and SPSS formatted files.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com
TELEform	Data capture, image capture, forms processing, and data extraction.	Cardiff Software, Inc. 3220 Executive Ridge Drive Vista, CA 92083 [760] 936-4500 http://www.cardiff.com/
Probability calculators		
NCSS Probability Calculator	Computes cumulative distribution functions (cdfs) and probability density functions (pdfs) for a variety of continuous and discrete distributions.	NCSS Statistical Software 329 North 1000 East Kaysville, UT 84037 [801] 546-0445 [800] 898-6109 http://www.ncss.com
Pcalc	Computes cdfs and pdfs for a variety of continuous and discrete distributions. User defined variables and functions can also be programmed.	Sytse Knypstra and Arjen Merckens Rijksuniversiteit Groningen Vakgroep Econometrie Postbus 800 9700 AV Groningen The Netherlands e-mail: S.Knypstra@eco.rug.nl http://www.eco.rug.nl/medewerk/knypstra/pcalc.html

(*continued overleaf*)

Table 1 (*continued*)

Title	Emphasis relevant to clinical trials	Vendor
StaTable	Computes cdfs and pdfs for a variety of continuous and discrete distributions.	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617]661-2011 http://www.cytel.com
Utilities		
DBMS/COPY	Conversion utilities to translate system files between many different statistical, graphics, database management, spreadsheet software, and data entry programs. Includes ASCII and ODBC support. Optional integration into the SAS system.	Data Flux Corp. Cary, NC [877] 846-3589 http://www.dataflux.com/Product-Services/Products/dbms.asp
Stat/Transfer	Conversion utilities to translate system files between many different statistical, database management, and spreadsheet programs. Includes ASCII and ODBC support.	Circle Systems 1001 Fourth Ave, Ste 3200 Seattle, WA 98154 [800] 366-3794 http://www.stat-transfer.com
Specialty analysis		
ACLUSTER	Analysis and sample size estimation for cluster randomization trials for three design types: completely randomized, paired-matched and stratified. Techniques include the analysis for binary, continuous, count, and time-to-event outcomes. Provides for the generation of cluster randomization schemes. Methods consistent with [3].	Alain Pinol and Gilda Piaggio UNDP/UNFPA/WHO/World Bank Special Programme of Research, Development and Training in Human Reproduction of the World Health Organization, 1211 Geneva 27, Switzerland Fax: +41-22-7913345 acluster@who.int Update Software Ltd Summertown Pavilion Middle Way, Oxford OX2 7LG United Kingdom +44 (0)1865 513902 http://www.update-software.com/acluster/

East	Design and interim monitoring of group sequential clinical trials. Includes methods for normal, binomial, and time-to-failure data. Supports various interim monitoring error spending functions and stopping boundaries.	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617] 661-2011 http://www.cytel.com
Egret	Interactive program for a variety of regression models (logistic, conditional logistic, logistic with random effects, Poisson, proportional hazards, parametric regression for failure time data, Kaplan–Meier and traditional contingency tables).	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617] 661-2011 http://www.cytel.com
EquivTest	Equivalence and bioequivalence analysis and testing procedures – paired design, two group, and crossover designs testing difference in means and ratio of means. Includes both parametric and nonparametric procedures.	Statistical Solutions Stonehill Corporation Center Suite 104, 999 Broadway, Saugus, MA 01906 [781] 231-7680 [800] 262-1171 http://www.statsolusa.com
GLIM	Interactive statistics package design to fit generalized linear models.	NAG Ltd Wilkinson House Jordon Hill Road Oxford OX2 8DR, UK +44 1865 511245 http://www.nag.co.uk
LIMDEP	Comprehensive program for estimation and analysis of regression models, and quantitative and limited dependent variables. Includes multinomial and nested logit models, and probit analysis. Parametric and nonparametric procedures for survival analysis: Weibull, lognormal, Gompertz, log-logistic, exponential, gamma parametric, generalized F-models, Weibull and exponential models with gamma heterogeneity; arbitrary censoring and left truncation; time varying covariates; plots of survival, hazard, and integrated hazard functions; semiparametric hazard function estimation; split population survival models	Econometric Software 15 Gloria Place, Plainview, NY 11803 [516] 938-5254 http:://www.limdep.com
LogXact	Exact inference for logistic, Poisson, and polychotomous regression. Computes exact P values and confidence intervals for regression coefficients.	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617] 661-2011 http://www.cytel.com

(*continued overleaf*)

Table 1 (*continued*)

Title	Emphasis relevant to clinical trials	Vendor
Mplus	Comprehensive modeling program for binary and polychotomous logistic regression; structural equation modeling of longitudinal data; latent variable mixture modeling.	Muthén & Muthén 3463 Stoner Ave. Los Angeles, CA 90066 [310] 391-9971 http://www.statmodel.com
PEST	Design and interim monitoring of group sequential clinical trials. Includes methods for normal, binomial, and time-to-failure data. Supports various interim monitoring error-spending functions and stopping boundaries. Available as a stand-alone package or as a SAS/AF applications for Windows.	MPS Research Unit The University of Reading PO Box 240 Earley Gate, Reading RG6 6FN, UK +44 118 931-6662 http://www.reading.ac.uk/mps
PROC LogXact	SAS procedures providing the functionality of the stand-alone version of LogXact. Seamless integration with SAS.	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617] 661-2011 http://www.cytel.com
PROC StatXact	SAS procedures providing the functionality of the stand-alone version of StatXact. Seamless integration with SAS.	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617] 661-2011 http://www.cytel.com
S + SeqTrial	An S-Plus software library for designing, monitoring, and analyzing clinical trials using group sequential methods. Includes methods for normal, binomial, and time-to-failure data. Supports various interim monitoring error-spending functions and stopping boundaries.	Insightful Corp. 1700 Westlake Avenue North, Suite 500 Seattle, WA 98109 [206] 283-8802 [800] 569-0123 http://www.splus.com http://www.insightful.com
SOLAS	Analysis of missing data based on techniques developed by [19]. Provides a choice of both multiple and single imputation techniques for parametric approaches using longitudinal/repeated measures, and nonparametric single observation study designs.	Statistical Solutions Stonehill Corporation Center Suite 104, 999 Broadway, Saugus, MA 01906 [781] 231-7680

		[800] 262-1171 http://www.statsolusa.com
SPSS Missing-Value Analysis	Analytic procedures for analyzing data sets with missing data. Procedures are designed around EM algorithms.	SPSS Inc. 233 S Wacker Dr 11th Floor Chicago, IL 60606 [800] 543-2185 http://www.spss.com
StatXact	Computation of exact P values and confidence intervals for a wide variety of nonparametric statistical procedures. Exact 2 and K-sample test for censored and uncensored survival data. Includes the logrank, Wilcoxon–Gehan, and Tarone–Ware tests. Computation of power and sample for a variety of nonparametric procedures.	Cytel Software Corp. 675 Massachusetts Avenue, Cambridge, MA 02139 [617] 661-2011 http://www.cytel.com
SUDAAN	Analysis of correlated data resulting from cluster designs (e.g. multiple outcomes per patient, repeated outcomes per patient, cluster randomization, and nested sampling schemes). Includes both descriptive and regression procedures (dichotomous and polychotomous logistic regression, and proportional hazards model).	Research Triangle Institute 3040 Cornwallis Road PO Box 12194 Research Triangle Park, NC 27709 [919] 541-6602 http://www.rti.org
WesVar	Analysis of data from multistage, stratified, and unequal probability samples, resulting from cluster designs (e.g. multiple outcomes per patient, repeated outcomes per patient, cluster randomization, and nested sampling schemes). Includes descriptive, ANOVA, and regression procedures (dichotomous and polychotomous logistic regression).	WESTAT 1650 Research Boulevard Rockville, MD 20850 [301]294-2006 http://www.westat.com/wevar
Testimate	Computation of exact P values and confidence intervals for a wide variety of parametric and nonparametric statistical procedures. Exact 2- and K-sample test for censored and uncensored survival data. Includes the logrank, Peto–Wilcoxon, Wilcoxon–Gehan, and Cox–Mantel tests. Procedures for equivalence tests, and ANOVA.	**idv**-Data Analysis and Study Planning Wessobrunner Straβe 6 D-82131 Gauting/Germany phone: 089/850 80 01 e-mail: idvGauting@aol.com
XPro	Computation of exact procedures for ANOVA, MANOVA, mixed models, repeated measures, growth curves, and regression. A parametric complement to StatXact and Testimate.	X-Techniques, Inc., PO Box 58, Millburn, NJ 07041 [212] 522-4539 http://www.x-techniques.com

(*continued overleaf*)

Table 1 (*continued*)

Title	Emphasis relevant to clinical trials	Vendor
Computer program and subroutines		
Anderson Statistical Archives	A large collection of statistical programs and subroutines for study planning (Phase I studies, randomization, sample size, and power) and analysis (accelerated failure time models and survival analysis, Bayesian Phase II monitoring boundaries, dose-finding,). Archives include FORTRAN and C source code. Software is distributed in the form of program source files.	The University of Texas M.D. Anderson Cancer Center Departments of Biomathematics and Biostatistics 1515 Holcombe Boulevard Houston, TX 77030-4096 [713] 792-2600 `http://odin.mdacc.tmc.edu`
IMSL	Statistics and mathematics library for programming in C, FORTRAN and JAVA. Includes routines for parametric and nonparametric survival analyses (Kaplan–Meier estimates and proportional hazards model (exponential, linear hazard, lognormal, normal, log-logistic, logistic, log least extreme value, least extreme value, log extreme value, Extreme value, Weibull); logistic regression; ANOVA; MANOVA	Visual Numerics, Inc. 1300 W. Sam Houston Pkwy S. Suite 150 Houston, Texas 77042 [800] 364-8880 [713] 784-3131 `http://www.vni.com/index.html`
Randomization		
RANCODE RANCODE professional	Generation of randomized treatment assignments; parallel group and crossover designs; fixed block and randomly determined block sizes; multicenter stratification; output includes randomization lists, adhesive labels, envelope labels, and cards with patient/treatment assignments.	**idv**-Data Analysis and Study Planning Wessobrunner Straβe 6 D-82131 Gauting/Germany phone: 089/850 80 01 e-mail: idvGauting@aol.com
Web sites		
StatLib	A website maintaining a large archive of statistical software, subroutine, computer algorithms, datasets, and information related to statistics and biostatistics.	`http://lib.stat.cmu.edu/`
Current Index to Statistics (CIS)	Biographical index to publications in statistics and related fields. References from the 1998 database include 111 core journals and approximately 400 noncore journals.	`http://www.statindex.org`
statistics.com	Information about statistics software (commercial, shareware, and freeware; large and small packages), as well as statistical analysis, data analysis, and short courses in statistics.	`http://www.statistics.com`

employed in clinical trials. However, starting with version 12, these methods were incorporated making the package a more useful tool for clinical trials. S-PLUS is increasingly being used by many statisticians as an all-purpose package and several textbooks are available, which incorporate its use in the solution of biostatistical problems [20–22]. ACLUSTER, SUDAAN, Stata, and WesVar facilitate better handling of clustered or hierarchical data than many of the more widely used packages. These methods are required when randomization is by clinic or physician rather than by patient (*see* **Cluster Randomization**).

Most users will have one or two general packages with which they are familiar and these usually meet the majority of the needs occurring for the analysis of clinical trials data. However, in special instances, these may need to be supplemented with specific routines from other general packages or with the increasing number of programs designed to perform more specialized tasks. The choice of a specific package from several, which provide the same functionality may be based on how easily it integrates into the user's current array of software. For example, an individual using SAS may choose SUDAAN over ACLUSTER, Stata, or WesVar for clustered data analysis (provided the statistical needs are met by the packages), since the former can be run from SAS as an add-on module, whereas the others cannot. Horton [7] provides some additional insight on programs that handle nested study designs.

Time-to-event data is one of the most common endpoints (*see* **Outcome Measures in Clinical Trials**) analyzed in clinical trials. Generally, packages that support survival analysis procedures minimally provide nonparametric procedures, such as actuarial and Kaplan–Meier estimates, logrank statistics, and some form of plots. Differences usually arise in the assortment of parametric procedures supported, the available models, type of censoring, fit diagnostics, ability to handle time-dependent covariates, and plotting. Harrell and Goldstein [6] and Oster [12] provide comparisons for a number of packages supporting survival analysis.

Even though packages supporting Kaplan–Meier procedures provide survival plots, it is often advantageous to have software designed specifically for graphics. Packages such as Axum, SigmaPlot, CrossGraphs, DeltaGraph, and GraphExpress give greater flexibility and control in the look and style of the plots. Displays can be more easily tailored to include textual annotations, error bars, and the overlaying of multiple graphs. All of these packages contain routines to calculate Kaplan–Meier estimates and produce the respective plots.

Several recently developed packages have procedures relevant to the conduct of meta-analysis. Although meta-analytic procedures have been applied to a variety of study designs, they are most appropriate for combining results from clinical trials where randomization eliminates the likelihood of a consistent bias across the studies being combined [18] (*see* **Meta-analysis of Clinical Trials**).

After the design of the study and the formulation of the hypothesis, sample size and power estimates are typically calculated. Many of the integrated statistical packages such as SPSS, SAS, and S-PLUS provide capabilities for performing such calculations. However, these packages generally do not provide algorithms for estimation of power and sample size for designs involving survival, logistic regression, analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), and equivalency testing. Programs designed specifically for power and sample size estimation such as nQuery Advisor, PASS, Power and Precision, and STATISTICA Power Analysis provide for a much richer selection of designs and outcome variables. For time-to-event data, these programs provide for the incorporation of various periods of recruitment and follow-up and various rates and patterns of recruitment, attrition, and censoring. Output from these packages often includes tabulated estimates along with graphical displays. Relatively few packages, StatXact and N being notable exceptions, provide sample size and power estimation for a variety of nonparametric procedures. PASS provides procedures for estimating sample size for group sequential designs. However, packages such as East, PEST and, S + SeqTrial handle group sequential designs in a more comprehensive manner. These packages not only provide sample size and power estimation but also include a full array of tools for the subsequent analysis of the trial.

New features appearing in recent and soon to be released versions of various sample size software include procedures for estimating the power relevant to DNA microarray studies. nQuery has added procedures for the two group t-test for fold change assuming lognormal distribution (with equal or unequal sample sizes) and the two-group t-test of equal fold change with fold-change threshold (equal

or unequal sample sizes). Other enhancements found in some software include the use of Monte Carlo methods to simulate data sets when the parameters of the statistical design cannot be explicitly described in closed form.

In general, there are no database management packages specifically designed and optimized for the conduct and management of data that is associated with clinical trials (*see* **Data Management and Coordination**). Typically, users are required to select a database management system and customize that package for their particular application. The users write their own forms, queries, reports, and applications within the package itself. One guidance offered is that the selection of such database management systems should be based on the familiarity and comfort of using the package, software capabilities that match the needs of the project, compatible computer resources and knowledgeable people within the institution with respect to the support of that package.

Moving away from strictly database management packages, three products that assist in the report generation and display of patient characteristics are ClinPlus, Patient Profiles, and PPD Patient Profiles. ClinPlus runs under the SAS system and consists of a set of predefined macro modules that are customizable for the specific needs of the study. These packages provide procedures consistent with the requirements of the US Food and Drug Administration for data collection, report summarization, and statistical analysis. Patient Profiles and PPD Patient Profiles, produced by two independent companies, provide data visualization of demographics, medical profiles, and other relevant patient data that may be stored within a database. Built into the packages is an assortment of reports with various graphical and charting display options. Both, Patient Profiles and PPD Patient Profiles, provide tools for drilling down into the data so that the details of specific individuals can be reviewed. The packages provide for a variety of data input, and interface with many database management systems and spreadsheet programs. This allows greater flexibility in areas where data management is performed.

One of the fundamental tasks of any clinical trial is the actual transfer of data from the paper forms to an electronic file that resides in the computer. Software that assists in this procedure can be divided into two broad categories: (1) data entry and (2) image capturing or the abstraction of data directly from forms. There are a number of packages that provide data entry support. BMDP Data Entry, EntryPoint, FALCON, Key Entry III, Rode/PC, SAS/FSP, and SPSS Data Entry are examples of software packages that facilitate the direct entering of information from the paper forms into the computer by an operator. Depending on the packages, a variety of options are available to the user including double entry (sometimes referred to as the two-pass verification system), content verification, and comprehensive editing tools (*see* **Clinical Trials Audit and Quality Control**). Some of these packages provide facilities to monitor the activities and the performance of the operator. Typically, these packages generate a data file in ASCII (American Standard Code for Information Interchange) format. Packages such as BMDP Data Entry, SAS/FSP or SPSS Data Entry also create a system file specific to their respective analytical software packages.

The other approach for moving data from forms to an electronic file format is image capturing or data capturing. This is provided by a number of software packages, including Ascent Capture, FaxWare, Input*Accel*, and TELEform. These packages take advantage of the scanning capabilities of computer systems through either direct use of a scanner or information abstracted from faxed material. These packages provide a mechanism for designing a form that will be used to capture the data. Programs, such as DataFax, ImagEntry, VDE, and Quantum 2000, have dual capabilities for direct data entry and data capturing. Similar to the data entry systems, various procedures are available for verifying the integrity of the data, data editing, and for data updating. Furthermore, they provide varying levels of data management.

Table 1 also summarizes a variety of specialized packages. The scope of these packages is more focused and provides the user with more comprehensive tools. LogXact, StatXact, Testimate, and XPro compute exact P values and confidence intervals for a wide variety of tests, many of which are commonly used for testing hypotheses in clinical trials. This is especially useful for rare events and small sample sizes. StatXact provides for nonparametric procedures, XPro for parametric procedures, and Testimate contains both parametric and nonparametric procedures. LogXact is used for exact inference for logistic, Poisson, and polychotomous regression. SAS and SPSS also provide modules for exact testing,

which incorporate a limited subset of the statistical engines from LogXact and StatXact. Cytel Corp., offers separate products programmed as PROC procedures (called PROC StatXact and PROC LogXact), which seamlessly interface with SAS and provide the full complement of statistical procedures that are available in respective stand-alone versions of the software. EquivTest provides methods useful for the analysis of equivalence trials. Clinical trials that are designed as equivalence studies (where we are attempting to prove the hypothesis of no-treatment effect) are increasing in number and are particularly applicable when two regimens are expected to have similar effects on treatment, but different toxicities. A number of packages are available to perform analysis for missing data. SOLAS is a package that uses single and multiple imputation, while SPSS missing value analysis uses EM algorithms. Although methods for missing data have been employed mostly in observational studies, some investigators have suggested that they be employed in clinical trials [9].

At one time, the toolbox of a good researcher included an extensive set of tables containing cumulative distribution functions (cdfs) and probability density functions (pdfs) for various distributions. These were indispensable when evaluating the significance of test statistics or computing confidence intervals. Now the output from statistical software packages includes these values. Although, the need for statistical tables has diminished over time, there are occasions when such tables are necessary. A researcher may decide to do some hand computations for a small dataset, or a power/sample size computation where P values and critical regions are needed. It is no longer necessary to consult paper or computerized versions of these tables. "Probability calculators", such as Pcalc, StaTable, and NCSS Probability Calculator provide the same information. There are some major advantages in using these packages. They have greater flexibility in that they compute value of the critical region for any specified P value, or the P value for any specified critical region. This is in contrast the older style tables, where interpolation is necessary when a particular value is not found. Depending on the program and the version, these packages cover an extensive array of common and not so common distributions. Another attractive feature of these software packages is that they can be freely downloaded and used for noncommercial purposes. More detailed discussion can be found in [1].

With the use of a variety of software, there is often the necessity to move data from one package to another. This may occur when transferring data from a data entry or data management system to analytical software, or from one statistical package to another. The universal format for data transfer is ASCII text. However, often the data is stored as a "system" file specific to the software package. System files also contain ancillary information such as variable and value labels, missing data codes, transformations and computed variables. This provides for an efficient use of the data within the software but not between software. Moving data between packages is accomplished by exporting the data as an ASCII file from the first program and then importing it into the second program as ASCII. When using this approach, variable and value names and missing data codes are usually lost in the transfer. In some instances, system files from one software package may be directly read by another and this information is not lost. However, packages usually support only a few types of system formats. Two products that specialize in facilitating data transfer are DBMS/COPY and Stat/Transfer. Each of these packages have conversion utilities to directly translate system files from one package into another. They also have the ability to filter and transform data thereby creating data subsets and new variables. DBMS/COPY can be run as a stand-alone package or integrated into SAS. In addition, it supports a broad range of system files.

Although statistical procedures are being continuously developed, there is often a time lag between when these new procedures first appear in the statistical journals and when the appropriate software is developed. Although the time between the introduction of new methodology and the development of corresponding software has decreased considerably in recent years, reliance on only the analyses available in software packages may result in less than optimum statistical procedures. Increasingly, packages have programming and macro capabilities to aid researchers in the development of their own "in-house" written software. Researchers may also program in languages of their choice such as C, Pascal, and Fortran. In addition, there are higher and object-oriented languages, such as Matlab, which handle complex mathematical and statistical manipulations making the development of new software considerably easier. To assist in this process, there are published subroutines and

numerical algorithms, such as IMSL for Fortran and C, Numerical Recipes in Fortran and C [15, 16], and Anderson Statistical Archives. A useful Web site is StatLib (`http://lib.stat.cmu.edu/`), which contains numerous mathematical and statistical codes for a variety of programming languages and packages. These are many of the same algorithms found in published articles and texts. This site also contains numerous data sets appearing in the literature along with computation results that may be used for evaluation and comparison purposes. Taking advantage of such resources should lessen considerably the amount of time spent in code writing and debugging of programs.

Randomization programs to determine treatment assignment of patients in a study have been traditionally lacking. Generally, users create their own algorithms that are specific to their particular study, or borrow and modify an existing program from a previous study. Larger clinical trials research units usually have an in-house library of software with patient randomization procedures. RANCODE and RANCODE professional, developed by **idv**, are the few commercially available programs that provide randomization schemes for a variety of study designs including parallel group and crossover. Options provide for fixed and randomly determined block sizes and stratification across study centers. The program has the capabilities for generating randomization lists, adhesive labels, envelope labels, and cards with patient/treatment assignments.

More rapid development of software for useful statistical procedures increases the possibility that some of the software may have "bugs". The user must bear the primary responsibility in determining whether newly developed software is performing the statistical computations both correctly and using the method that the user believes is being applied. A useful resource in checking the accuracy of new software is the Web, where there are example files containing both data and sample programs. Often data sets from well-known studies are readily available and can be used to validate and test software. These may be found on the vendor's Web site or at StatlLib. The results of these tests can then be compared to known or published results. Another method of testing a software package is to use data from extreme cases or boundary points where the answer is known. An excellent source for this approach is Statistics Quiz available from *SYSTAT*, written by Lee Wilkinson, which provides empirical problems that can be used in evaluating statistical packages. Although the problems in Statistics Quiz are presented for the purpose of comparing programs run on microcomputers, they can be used for verification in other computer environments.

Even if the software uses a "correct" method, the specific algorithms employed by the software package may not be clearly indicated. Often, there are multiple methods of applying the same general techniques, and different decisions on approach by various packages can lead to different answers. Situations where there are often differences among packages include methods of handling tied observations and methods of handling missing data. For some statistical tests, some packages use only asymptotic procedures, while other packages use exact procedures for small sample sizes and asymptotic methods for larger samples. The sample size that determines the crossover from an asymptotic to an exact procedure may not be the same for different packages. Similarly, different packages may vary in how they address missing data. In settings where there are multiple independent variables, a software package may perform listwise deletion, whereby analysis is performed using all available data for that particular variable; or case-wise deletion where the entire case is dropped from the analysis if any of the variables are missing for a particular case. It is both useful and essential that documentation be available for the software being used. This documentation should contain references and explicit information with regard to computations being performed within the package. Different software vendors utilize differing levels of precision in the algorithms in their software. A number of articles have been written suggesting approaches for evaluating statistical software packages and assessing their reliability [4, 10, 11]. *The American Statistician*, a journal published quarterly, has a section entitled *Statistical Computing and Graphics* and often contains articles discussing issues with respect to various statistical packages.

As statistical methods become more complex and the actual use of the analytical software becomes easier, the potential for applying an incorrect procedure to a study design increases. The user needs to understand the full implication of the analyses being employed and not merely "press a button". For example, generalized additive models (GAM) are often fit by the S-Plus software package. It has recently been

demonstrated that the default convergence criteria in S-Plus (version 3.4) do not assure convergence of its iterative estimation procedure and can provide biased estimates of regression coefficients and standard errors [2]. Thus, the user needs to be savvy enough to use a more stringent convergence to assure the convergence of the iterative procedure. However, this does not necessarily reduce the potential for the underestimation of the standard error and the presence of bias in the estimate of the regression coefficients. It has been shown that this underestimation can occur if concurvity, the nonparametric analog of multicollinearity is present in the data and might lead to significance tests with inflated type 1 error (i.e. rejection of the null hypothesis when it is in fact true) [17]. This may result in erroneously declaring a statistically significant effect when none exists. Therefore, even though the software may generate technically correct results, the appropriateness for the data set may be questionable.

Software vendors are continually updating and correcting their software. Aside from version changes, vendors often issue patches or service releases that correct errors and bugs. These patches may also impart additional functionality to the software. Typically, the researcher must be proactive in his/her search to find such fixes. More recently, some vendors have made this task a little less burdensome by providing a mechanism where the user can have the software check a predefined website for the availability of new updates. Generally, this is initiated by the user and therefore should be done on a routine basis. Fortunately, some manufacturers have carried this a step further by incorporating a "live update" mechanism whereby the software will periodically check the Web for updates and automatically notify the user when fixes are available.

Vendor Web sites also provide for compilations of user notes, macros, and add-ons that extend the capability of the basic package. Links to "frequently asked questions", "knowledge base", and "technical support" found on web sites often provide additional information that may not be readily available in the manuals that are packaged with the software. Auxiliary manuals and technical reports provide a wealth of information. These web sites may also provide files of macros and programs. Instead of transcribing programs from paper textbooks and journal articles, files can be downloaded. The most current and up-to-date versions are probably those that are on the Web or obtained by contacting authors directly.

In the business world, it is common for software vendors to change names or to be taken over by other companies. When this happens, it does not necessarily imply that the software is no longer available or no longer supported. It may mean that the product is produced under another name. When this occurs, the relevant forwarding information is not readily available, and the user needs to search the web using the vast array of Internet search engines that are available at his/her disposal. For instance, BMDP is no longer available from BMDP Statistical Software, Inc. However, it is available through Statistical Solutions as BMDP New System Professional.

The developments in microcomputer technology over the past 15 years have provided the capabilities in computational speed for computer software packages that are increasingly sophisticated and complex. There are currently many packages available that are relevant for clinical trial design and analysis. Users are not restricted to just one or two standard packages. With thought and inquiry, it is now possible for individuals to find software that is tailored exactly or nearly exactly to meet most of their needs.

Acknowledgments

The authors would like to acknowledge the contribution of Hannah Rothstein for information on the meta-analytic software packages.

References

[1] Boomsma, A. & Molenaar, I.W. (1994). Four electronic tables for probability distributions, *The American Statistician* **48**(2), 153–162.

[2] Dominici, F., Daniels, M., Zeger, S.L. & Samet, J.M. (2002). Air pollution and mortality: estimation regional and national dose-response relationships, *Journal of the American Statistical Association* **97**, 100–111.

[3] Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold Publishing, London, p. 151.

[4] Francis, I., Heiberger, R.M. & Velleman, P.F. (1975). Criteria and considerations in the evaluation of statistical program packages, *The American Statistician Statistical Computing* **29**(1), 52–56.

[5] Goldstein, R. (1998). Software, biostatistical, in *Encyclopedia of Biostatistics*, Vol. 5. P. Armitage & T. Colton, eds. John Wiley & Sons, West Sussex, pp. 4180–4187.

[6] Harrell Jr., F.E. & Goldstein, R. (1997). A survey of microcomputer survival analysis software: the need for an integrated framework, *The American Statistician* **51**(4), 360–372.

[7] Horton, N.J. & Lipsitz, S.R. (1999). Review of software to fit generalized estimating equation regression models, *The American Statistician* **53**, 160–169.

[8] Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis : Correcting Error and Bias in Research Findings*. Sage Publications, Newbury Park, p. 592.

[9] Lavori, P.W., Dawson, R. & Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data, *Statistics in Medicine* **14**, 1913–1925.

[10] McCullough, B.D. (1998). Assessing the reliability of statistical software: Part I. Statistical computing software reviews, *The American Statistician* **52**(4), 358–366.

[11] McCullough, B.D. (1999). Assessing the reliability of statistical software: Part II, *The American Statistician* **53**(2), 149–159.

[12] Oster, R.A. (1998). An examination of five statistical software packages for epidemiology, *The American Statistician* **52**(3), 267–280.

[13] Piantadosi, S. (1997). *Clinical Trials: A Methodological Perspective*. John Wiley & Sons, New York.

[14] Pocock, S.J. (1984). *Clinical Trials: A Practical Approach*. John Wiley & Sons, New York.

[15] Press, W.H., Teukolsky, S.A., Vettering, W.T. & Flannery, B.P. (1992a). *Numerical Recipes in FORTRAN – The Art of Scientific Computing*, 2nd Ed. Cambridge University Press, New York.

[16] Press, W.H., Teukolsky, S.A., Vettering, W.T. & Flannery, B.P. (1992b). *Numerical Recipes in C – The Art of Scientific Computing*, 2nd Ed. Cambridge University Press, New York.

[17] Ramsay, T.O., Richard, T.B., Burnett, R.T. & Krewski, D. (2003). The effect of concurvity in generalized additive models linking mortality to ambient particular matter, *Epidemiology* **14**, 18–23.

[18] Rockette, H.E. & Redmond, C.K. (1988). Limitations and advantages of meta-analysis in clinical trials, *Recent Results in Cancer Research* **111**, 99–104.

[19] Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. John Wiley & Sons, New York, p. 258.

[20] Selvin, S. (1998). *Modern Applied Biostatistical Methods Using S-PLUS*. Oxford University Press, New York.

[21] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

[22] Venables, W.N. & Ripley, B.D. (1994). *Modern Applied Statistics with S-PLUS*, 3rd Ed. Springer, New York.

VINCENT C. ARENA & HOWARD E. ROCKETTE

Software for Genetic Epidemiology

A wide variety of software has been written to facilitate the task of managing, error-checking, and analyzing **genotype** and phenotype data for genetic studies. An exhaustive review of the software available would fill an encyclopedia of its own. A list of genetic analysis software from the Rockefeller University web site (`http://linkage.rockefeller.edu/soft/list.html`) currently includes over 150 programs. Additionally, both methodologic and software development in **genetic epidemiology** are constantly advancing and any catalog of such software will soon be rendered incomplete as new programs become available. Consequently, this review will seek to describe the types of software that are available, to discuss and characterize some of the most widely used programs, and to identify features that may differ between similar packages.

First we must define what is to be included in the category of software for genetic epidemiology. This category will be defined as including any program specifically designed for the management of genotype or pedigree data, error-checking of genotype or pedigree data, or genetic analysis using **segregation**, **linkage**, or **linkage disequilibrium** based methods. Excluded from this review are web sites such as those providing genetic **maps**, software for sequence alignment or comparison (*see* **Sequence Analysis**), software for general **variance component analysis**, and more general statistical packages, such as SAS or S-PLUS, which were not designed for genetic analysis but are often used in its service. Additionally, we will restrict our consideration to programs that are used for human genetic analysis, excluding software written for inbred lines and other study designs that are only possible in animal models. All the software discussed in this article is publicly available and, with the exception of S.A.G.E. and Cyrillic, all the programs are available free of charge for noncommercial use.

Programs for the Management and Display of Pedigree and Genotype Data

Although many researchers use standard database or spreadsheet software to store data for their genetic epidemiology studies, management software designed specifically for genetic data offers several advantages in terms of error-checking and data formatting. Errors in genotype or pedigree data may be identified through checks on Mendelian inheritance (*see* **Mendel's Laws**) and logical consistency. Data may be imported from other database systems and formatted for export to pedigree drawing or analysis programs. Commonly used software for error checking, formatting, and display of pedigree and genotype data is listed in Table 1.

Error-checking

Programs are available for detection of errors in genotype, phenotype, and pedigree data. Identification of apparent genotyping errors in family data through violations of Mendelian inheritance is performed by many data preparation and analysis programs. Checking for Mendelian consistency requires inference of the genotypes of nonsampled individuals and some error-checking programs will fill in missing **marker** genotypes when they can be unambiguously inferred. Most genetic analysis programs report errors when Mendelian inconsistencies are detected. However, only a few (e.g. ASPEX, FBAT, SimWalk2) provide explicit identifications of the likely source of the error or automatically eliminate the error by blanking suspect genotypes. PedCheck, PEDSYS, and SimWalk2 perform Mendelian consistency checks in extended pedigrees, whereas ASPEX and FBAT only accommodate nuclear families. SimWalk2 takes this one step farther and, using allele frequencies and marker maps, considers the distribution of alleles in sibships and the locations of apparent recombinations, as well as violations of Mendelian inheritance, to produce a posterior probability of error for each genotype for each individual.

The persistence of apparent genotyping errors over numerous markers may suggest pedigree errors, such as nonpaternity. Pedigree errors may also be detected by comparing empirical kinship, estimated from the observed identity-by-descent (ibd) allele sharing (*see* **Identity Coefficients**) between individuals, to the degree of relationship predicted by the assumed pedigree configuration. The program Siberror identifies probable pedigree errors in nuclear family data using genotype data at numerous markers. Similarly, RELCHECK predicts relationships between individuals (monozygotic twins, full sibs, half sibs,

Table 1 Programs for data management and pedigree drawing

Name	Mendelian error-checking	Pedigree error-checking	Data formatting	Pedigree drawing	Operating system(s)	Reference
PEDSYS	X	X	X		Mac, Unix, Windows	`http://www.sfbr.org/sfbr/public/software/pedsys/pedsys.html`
PedCheck	X				Unix	O'Connell & Weeks [36]
PedHunter		X	X		Unix	Agarwala et al. [1]
PREST		X			Unix	McPeek & Sun [33]
RELCHECK		X			Unix, Windows	Broman & Weber [4]
RELPAIR		X			Unix, Windows	Duren et al. [11], Epstein et al. [15]
Siberror		X			Unix	Ehm & Wagner [12]
Mega2			X		Unix	Mukhopadhyay et al. [34]
CoPE				X	Any	Brun-Samarcq et al. [5]
Pedigree/Draw				X	Mac	`http://www.sfbr.org/sfbr/public/software/pedraw/peddrw.html`
PEDRAW				X	Windows, X-Windows	Curtis [8]
Cyrillic			X	X	Windows	`http://www.cyrillicsoftware.com/`

or unrelated) based on their sharing at numerous genotyped markers. RELPAIR and PREST perform the same type of analysis for slightly larger pedigrees, extending to first-cousin relationships. ACT and ASPEX, although primarily **linkage analysis** programs, also perform pedigree error-checking for sibships or nuclear families.

Pedigree and phenotype errors may be identified through checks on the logical consistency of demographic data given the constraints of family relationships. For example, mothers should be female and fathers should be male. Parents should be older than their children. Dates at which individuals are examined should be later than their dates of birth. Pedigree storage programs are also equipped to manipulate and fill in pedigree structures. Analysis programs generally require that individuals who are listed as parents have their own entry in the database and some management software will create entries for missing individuals necessary to complete the pedigree structure. Additional pedigree-based manipulations include trimming of uninformative individuals. The computational burden of some types of analyses increases exponentially by pedigree size, making it most efficient to exclude individuals who are missing crucial phenotype or genotype data. When eliminating uninformative individuals, software specifically designed for the management of pedigree data is equipped to consider whether individuals with missing data are necessary to complete the pedigree structure and thus should be retained. Some of these programs will also calculate kinship coefficients based on the provided pedigree structure. PEDSYS performs all these functions. S.A.G.E. also performs many of these functions through its specialized programs. PedHunter is designed to query genealogic databases and can calculate kinships, find individuals who are related in a specific way (e.g. find all siblings of a given individual), or identify the minimum number of common ancestors to connect a specified set of individuals (e.g. trim a pedigree to affected individuals and those necessary to connect them).

Translating Between Data Formats

Genetic analysis programs require specific data formats and there is generally little overlap in the data structure required by different programs. Management software designed specifically for genetic data is often equipped to produce data files formatted for the needs of various genetic analysis programs. PEDSYS can import data from delimited field formats (e.g. comma or space delimited) and export data formatted for Pedigree/Draw, FISHER, MENDEL, CRI-MAP, PAP, SOLAR, or LINKAGE. Mega2 formats data for use in SimWalk, MENDEL, ASPEX, APM, SLINK, SIMULATE, S.A.G.E., GeneHunter, TDTMax, and SOLAR. Cyrillic can import data from programs such as MLINK, Pedigree/Draw, and CRI-MAP and output data for these programs as well as for LIPED. PedHunter outputs data formatted for the LINKAGE or Pedigree/Draw programs. SIB-PAIR, although primarily an analysis program, formats data for APM, Arlequin, ASPEX, CRI-MAP, FISHER, GAS, GDA, LINKAGE, MENDEL, PAP and S.A.G.E.

Pedigree Drawing

Drawings of pedigrees are used by both clinicians (*see* **Genetic Counseling**) and researchers to verify family relationships and to display phenotypes, genotypes, and **haplotypes**. Two of the most widely used pedigree drawing programs are Cyrillic and Pedigree/Draw. The most immediate difference between these two programs is that Cyrillic runs under Windows whereas Pedigree/Draw is Macintosh based. Both programs can accommodate large and complex pedigrees, provide a variety of marking symbols to denote phenotype and sampling status, and display genotypes or other text. Cyrillic provides some data management and formatting capabilities and includes routines for risk assessment. However, Cyrillic is a commercial program whereas the other pedigree drawing programs listed are freely available over the web. Although less widely used than Cyrillic, PEDRAW is also Windows based, performs many of the same functions, and is available without charge over the web. CoPE is designed to allow multiple researchers to access the same pedigree database and is a Java Script program that can be used on any platform through a web browser. Cyrillic can provide haplotypes and CoPE will draw haplotypes if they are specified.

Programs for Construction of Marker Maps

Although many researchers use online databases to obtain marker maps, constructing marker maps

from one's own data set can be advantageous if sufficient pedigree information is available. Methods of **multipoint linkage analysis** that are based on correlations in ibd may suffer if an ill-fitting map is specified (*see* **Genetic Map Functions**) and loci are assumed to be more or less highly correlated (i.e. closer or farther apart) than is reflected in the current data. Map construction can also be useful for error-checking purposes. For example, apparent expansion of the marker map, as compared with published maps, may highlight problematic markers or genotyping errors that cause an apparent excess of recombinations but not Mendelian inconsistencies. In some cases, adequate map data simply may not be available from other sources. Newly identified variants in candidate genes may not be placed on the online linkage maps and may be only approximately localized in physical maps.

Any of the parametric linkage packages (*see* **Linkage Analysis, Model-based**) (listed in Table 2) could be used to estimate recombination between genotyped markers and construct a map. However, several packages automate this process. CRI-MAP [29] has a variety of map construction routines that place new markers relative to a map of old markers and evaluate the **likelihoods** of alternative map orders. In its likelihood calculations, CRI-MAP considers only meioses in which the parental transmission can be unambiguously inferred, whereas standard linkage programs would weight over the possible values of missing genotypes that cannot be unambiguously inferred. This restriction leads to some loss of information but may be viewed as conservative, particularly when good estimates of allele frequency are not available. The chrompic option of CRI-MAP can also be used to examine pictorially the grandparental origin of each genotype along a chromosome for the most likely phase. MultiMap [32] further automates this process, adding markers to the map in order of locus content, which can be specified by the user but is generally measured by the marker's heterozygosity. MultiMap uses CRI-MAP for likelihood computation and thus has the same limitations with regard to the treatment of missing data. MultiMap also has a module for the construction of **radiation hybrid maps**. Both programs are available for Unix systems and as a C source code which could be compiled under a variety of operating systems. MultiMap requires a C compiler and a Lisp interpreter to run.

Software for Genetic Analysis

Numerous programs are available for genetic analyses using segregation, linkage, and linkage-disequilibrium based methods and only the most commonly used of these programs are discussed here. To obtain a representative sample of the most widely used genetic analysis software, both the proceedings of Genetic Analysis Workshop 11 [18] and the pre-conference abstract book for Genetic Analysis Workshop 12 (held in October 2000) were surveyed. Together these two volumes contain 320 papers reporting analyses of four data sets ranging from alcoholism and quantitative measures of evoked brain potentials to asthma and related risk factors to simulated diseases and quantitative traits with genome screen and single nucleotide polymorphism (SNP) data. Any genetic analysis program that was cited by a total of 5 or more papers in these two volumes is listed in Table 2.

Segregation Analysis

The programs indicated in the "segregation" column of Table 2 are those that perform classic quantitative trait or penetrance model-based segregation analyses (*see* **Segregation Analysis, Classical**; **Segregation Analysis, Complex**). PAP and S.A.G.E. can fit a variety of environmental, polygenic, and Mendelian models. SAGE also has modules for a variety of regressive models. Loki models a quantitative trait as a function of a number of diallelic quantitative trait loci (QTLs), providing posterior probability distributions for the number of QTLs, the QTL allele frequencies, and the additive genetic **heritability** of the QTLs. These programs can also be used for combined segregation and linkage analyses in which both parameters for the penetrance model or the QTLs and the location of the trait locus are estimated simultaneously. Although not marked as segregation software in the table, programs that perform variance component analyses may also be used to examine the genetic architecture of a trait. ACT, GeneHunter, and SOLAR can be used to estimate the additive and dominance components of heritability and to compare the likelihood of models with and without genetic components of variance.

Linkage Analysis

The linkage programs in Table 2 utilize a wide variety of statistical models and methods, including

Table 2 Software for genetic analysis

Name	Segregation	Linkage	Disequilibrium	Trait[a]	Methods or models	Extras	Operating system(s)	Reference
ACT		X	X	Q	Variance component, TDT	Multivariate, pedigree error-checking	Unix	Amos et al. [3]
Allegro		X		D	Penetrance model, affected pair	Haplotypes, observed map, empirical *P* values	Unix	Gudbjartsson et al. [20]
ASPEX		X	X	D	Affected pair, TDT	Mendelian and pedigree error checking	Unix	`ftp://lahmed.stanford.edu/pub/aspex/doc/usage.html`
FASTLINK		X		B	Penetrance model		DOS, Unix, VMS	Cottingham et al. [7]
FBAT			X	B	TDT	Mendelian error checking	Mac, Unix, Windows	Laird et al. [28]
GASSOC			X	D	TDT		Unix	Schaid [39]
GeneHunter		X	X	B	Penetrance model, affected pair, variance component, TDT, Haseman–Elston	Haplotypes	Unix	Kruglyak et al. [27]
GeneHunter+		X		D	Affected pair		Unix	Kong & Cox [24]
HOMOG				B		Heterogeneity testing	DOS, Unix, VMS Windows	Ott [37]
Linkage		X		B	Penetrance model		DOS, Unix, VMS	Lathrop et al. [31]

(*continued overleaf*)

(*continued*)

Name	Segregation	Linkage	Disequilibrium	Trait[a]	Methods or Models	Extras	Operating system(s)	Reference
Loki	X	X		Q	Penetrance model, Markov chain Monte Carlo	Multilocus	Unix	Heath [23]
Mapmaker/Sibs		X		B	Affected pair, Haseman–Elston		Unix	Kruglyak & Lander et al. [26]
PAP	X	X	X	B	Penetrance model, variance components, measured genotype	Multilocus, multivariate	Unix	Hasstedt [22]
S.A.G.E.	X	X	X	B	Penetrance model, affected pair, Haseman–Elston, TDT, regressive models		Unix, Windows	`http://darwin.cwru.edu/pub/sage.html`
SIB-PAIR		X	X	B	Affected pair, Haseman–Elston, TDT, measured genotype	Empirical *P* values	DOS, Unix, Windows	Duffy [10]
SimWalk2		X		D	Penetrance model, affected pair	Haplotyping, Mendelian and double recombinant error checking	Unix	Sobel & Lange [40]
SOLAR		X	X	B	Variance component, measured genotype	Multilocus, epistasis, empirical *P* values	Unix	Almasy & Blangero [2]
VITESSE		X		B	Penetrance model		DOS, Unix, VMS	O'Connell & Weeks [35]

[a]D = discrete; Q = quantitative; B = both.

model-based, affected relative pair, Haseman–Elston, and variance components (*see* **Linkage Analysis, Model-based**; **Linkage Analysis, Model-free**). Although there is a general equivalence of methods between software using the same models, there are often subtle but potentially important differences. Among programs using a Haseman–Elston test, S.A.G.E. and SIB-PAIR include both the classic [21] and modified Haseman–Elston algorithm [14], whereas GeneHunter and Mapmaker/Sibs provide only the classic Haseman–Elston [21] algorithm. Among penetrance model-based and variance component programs, differences in the handling of multipoint inference (discussed below) may lead to limitations on the size of pedigree or the number of genotyped markers that can be considered in practice.

Variance component linkage programs also differ in the types of data they use and in the ease with which various models can be parameterized. ACT and SOLAR make use of singleton individuals in addition to subjects in families to estimate trait means and standard deviations and **covariate** effects. In contrast, GeneHunter eliminates singleton individuals. Dominance components of variance can be automatically incorporated in GeneHunter but must be manually added in SOLAR through direct modification of the covariance function. However, the ability to directly modify the covariance function in SOLAR also permits advanced users to incorporate terms for **gene–environment interaction** and to construct joint tests of linkage and disequilibrium.

Affected relative pair programs differ from each other in the calculation of the test statistic and in options for correcting for the nonindependence among pairs when there are more than two affected sibs within a sibship. When some individuals are unavailable for genotyping and descent information is incomplete, the perfect data approximation of GeneHunter's nonparametric linkage (NPL) score overestimates the **variance** in its test statistic and becomes highly conservative. Allegro and GeneHunter+ use a **maximum likelihood** method that is not subject to this problem. When $n > 2$ affected sibs are available from a given family, the analytical options may include using only one pair, using the $n - 1$ independent pairs that can be constructed, or using all pairs. ASPEX also provides the option to consider only pairs where ibd sharing can be inferred unambiguously.

ASPEX, Mapmaker/Sibs, and SIB-PAIR are limited to nuclear families, whereas the other linkage programs listed can accommodate larger pedigrees. Most of the affected pair and variance component linkage programs listed estimate marker-specific and multilocus ibd allele sharing and some will output ibd matrices which may be useful for error-checking or for importation into other programs. Some of the additional features of these programs are noted in the "Extras" column of Table 2. As discussed above, some packages check for genotyping errors. A few of the programs provide haplotypes. Some permit multivariate linkage analyses of two or more traits. Some consider multilocus models that incorporate multiple loci influencing the trait with or without epistatic interaction between the loci. Several packages have some simulation capacity, permitting the estimation of empirical ***P* values**.

Linkage programs differ in their methods for multipoint inheritance inferences and these differences have implications for the type of data that can be analyzed. LINKAGE, FASTLINK, and VITESSE use the **Elston–Stewart algorithm** [13] which results in exponential increases in computing time with the number of markers analyzed. FASTLINK incorporates algorithmic improvements that make it somewhat more efficient than LINKAGE. VITESSE uses set-recoding and fuzzy inheritance methods to speed up computations, allowing it to handle more markers than LINKAGE or FASTLINK, but it is limited to pedigrees without loops descended from a single founder couple. ACT, Allegro, GeneHunter, and GeneHunter+ use the Lander–Green **Hidden Markov Model** [29] which incurs an exponential increase in computing time with the number of nonfounders in the sample, placing a practical limit on the size of the pedigrees that can be analyzed. SOLAR uses a multipoint approximation based on correlations between ibd at the individual markers [2, 16]. Computing time for this approach is linear in both the number of markers and the number of individuals, permitting analyses of larger pedigrees. However, this approximation depends on the informativeness of the genotyped markers. Thus, while it performs well for genome scanning with microsatellites, it would be suboptimal for multipoint ibd estimation given a map of SNPs. SIB-PAIR also uses this method to impute ibd between flanking markers in its Haseman–Elston linkage routine. Loki and SimWalk2 use

Markov chain Monte Carlo methods which are also approximate and linear in the number of individuals and the number of markers but do not suffer the limitations on marker informativeness imposed by the between marker correlation method. Note that SimWalk2 requires the program MENDEL [30] to compute location scores for parametric linkage analyses.

While not precisely a linkage program, HOMOG is designed to be used in conjunction with linkage software. Given the lod scores for each family in a data set, HOMOG performs heterogeneity testing to assess whether in some proportion of families the trait is unlinked to the region in question. HOMOG extension modules permit consideration of multiple locus testing, for example models in which the trait is linked to locus 1 in some families, linked to locus 2 in other families, and unlinked to either locus in another group.

Linkage Disequilibrium Based Analysis

The programs indicated in the "Disequilibrium" column of Table 2 generally fall into two broad categories: those that perform transmission disequilibrium tests (TDTs) for linkage in the presence of disequilibrium with discrete or quantitative traits, and those that model the mean of a quantitative trait as a function of genotype. A few programs perform both these functions. Some programs have additional association tests that fall into neither of these categories. GASSOC has a variety of test statistics optimized for different underlying genetic models (e.g. recessive, dominant). SIB-PAIR performs standard tests of allele frequency distribution in affected and unaffected individuals with estimation of empirical P values through gene-dropping conditional on family structure and allele frequencies.

ACT, ASPEX, FBAT, GASSOC, GeneHunter, the TDTEX module of S.A.G.E., and SIB-PAIR all perform standard TDT tests for di- or multiallelic markers using parent–child trios derived from nuclear families. ACT, ASPEX, GeneHunter, S.A.G.E., and SIB-PAIR use a variety of permutation and **Monte Carlo methods** to provide empirical P values, taking into account factors such as the use of multiple sibs within a sibship. FBAT allows **covariate** adjustments. Up to four adjacent, closely linked marker loci can be included in the TDT in GeneHunter. ACT has a macro for simultaneous consideration of multiple loci through **conditional logistic regression**. The TDT portion of ACT consists of SAS macros that require the SAS package to run.

PAP, the ASSOC module of S.A.G.E., SIB-PAIR, and SOLAR (which is based on the program FISHER) can be used for measured genotype testing in which the mean of a quantitative trait is modeled as a function of genotype. PAP, S.A.G.E., and SOLAR accommodate the inclusion of quantitative or discrete covariates in these analyses. PAP and SOLAR use maximum likelihood methods that assume a **multivariate normal distribution** for the trait values. SOLAR also has a multivariate t distribution option. S.A.G.E. uses generalized modulus **power transformations** that require less stringent assumptions about the distribution of the trait values. PAP, SOLAR, and S.A.G.E. directly account for the nonindependence among family members and model a residual familial correlation not due to the locus being tested, whereas SIB-PAIR estimates empirical P values through gene-dropping within the families.

Conclusions

There are many factors that go into deciding which of these packages to use in any given situation – the computer resources available to the project, the size and structure of family data, the type of traits to be analyzed, and the desired methods of analysis. Each of the programs discussed above has strengths and weaknesses and it is impossible to make blanket recommendations that are appropriate for all, or even most, studies. However, a number of resources exist that compare the performance and utility of some of these packages under a variety of conditions. Some of these reviews address questions of power and accuracy while others comment on ease of use and computational intensity. It should be noted that comparisons of software for genetic analysis reflect, and often cannot be separated from, comparisons of the underlying analytical methods, particularly when addressing questions of power and accuracy. Comparison of methods and software for genetic analysis is the primarily goal of the Genetic Analysis Workshop and the proceedings of this conference contain many such comparisons (e.g. [18, 19], and [41]).

Table 3 Selected publications comparing various software packages

Reference	Software compared	Context
Cervino & Hill [6]	LRAT, RCTDT, SIBASSOC, TRANSMIT	TDT with differing family structures, population stratification, nonpenetrance, nonpaternity
Davis & Weeks [9]	ASPEX, GeneHunter, Mapmaker/Sibs, S.A.G.E., and others	Linkage using affected sibpairs or sibships
Goldgar & Oniki [17]	LINKAGE, MIM	Comparison of penetrance model and ibd-based quantitative trait linkage analysis
Konigsberg et al. [25]	FISHER/MENDEL, PAP, S.A.G.E.	Quantitative trait segregation
Schaffer [38]	FASTLINK, LINKAGE	Model-based linkage with pedigree loops or ungenotyped individuals
Williams & Blangero [42]	Mapmaker/Sibs, SOLAR	Quantitative trait linkage

Other manuscripts that compare and contrast various software packages are detailed in Table 3.

There are also online sources that compare programs or are a repository of information about multiple programs. A University of Washington web site (`http://www.cs.washington.edu/homes/pmork/final_project/`) provides a fairly intensive summary of a selection of programs for quantitative trait analyses, including FBAT, GeneHunter, LOKI, and SOLAR. The Rockefeller University web site, mentioned in the introduction, is an excellent resource for finding software often with links to web sites where the programs can be downloaded. Links to downloadable versions of many programs also can be found at a European Bioinformatics Institute mirror site (`ftp://ftp.ebi.ac.uk/pub/software/linkage_and_mapping/`) or at the Weizmann Institute of Science Bioinformatics Unit (`http://bioinfo.weizmann.ac.il/repository/mapping_software.html`). Online documentation for many programs is available through `http://watson.hgen.pitt.edu/docs/` or `http://www.well.ox.ac.uk/docs/index.html`. The Computational Methods and Algorithms Group at NIH also has an extensive genetic analysis software site at `http://cmag.cit.nih.gov/Lserver.htm`.

Acknowledgment

Preparation of this manuscript was supported by US National Institutes of Health grant MH59490. I am also grateful to Ravi Duggirala and Harald Göring for their helpful comments and suggestions.

References

[1] Agarwala, R., Biesecker, L.G., Hopkins, K.A., Francomano, C.A. & Schäffer, A.A. (1998). Software for constructing and verifying pedigrees within large genealogies and an application to the old order Amish of Lancaster County, *Genome Research* **8**, 211–221.

[2] Almasy, L. & Blangero, J. (1998). Multipoint quantitative trait linkage analysis in general pedigrees, *American Journal of Human Genetics* **62**, 1198–1211.

[3] Amos, C.I., Zhu, D.K. & Boerwinkle, E. (1996). Assessing genetic linkage and association with robust components of variance approaches, *Annals of Human Genetics* **60**, 143–160.

[4] Broman, K.W. & Weber, J.L. (1998). Estimation of pairwise relationships in the presence of genotyping errors, *American Journal of Human Genetics* **63**, 1563–1564.

[5] Brun-Samarcq, L., Gallina, S., Philippi, A., Demenais, F., Vaysseix, G. & Barillot, E. (1999). CoPE: a collaborative pedigree drawing environment, *Bioinformatics* **15**, 345–346.

[6] Cervino, A.C. & Hill, A.V. (2000). Comparison of tests for association and linkage in incomplete families, *American Journal of Human Genetics* **67**, 120–132.

[7] Cottingham, R.W., Jr, Idury, R.M. & Schaffer, A.A. (1993). Faster sequential genetic linkage computations, *American Journal of Human Genetics* **53**, 252–263.

[8] Curtis, D. (1990). A program to draw pedigrees using LINKAGE or LINKSYS data files, *Annals of Human Genetics* **54**, 365–367.

[9] Davis, S. & Weeks, D.E. (1997). Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation, *American Journal of Human Genetics* **61**, 1431–1444.

[10] Duffy, D.L. (1997). Sib-pair: a program for nonparametric linkage/association analysis, *American Journal of Human Genetics* **61**, Supplement, A197.

[11] Duren, W.L., Cox, N.J., Hauser, E.R., Boehnke, M. & the FUSION Study Group (1997). Software for determining most likely relationships in relative pairs, *American Journal of Human Genetics* **61**, Supplement, A273.

[12] Ehm, M. & Wagner, M. (1998). A test statistic to detect errors in sib-pair relationship, *American Journal of Human Genetics* **62**, 181–188.

[13] Elston, R.C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Human Heredity* **21**, 523–542.

[14] Elston, R.C., Buxbaum, S., Jacobs, K.B. & Olson, J.M. (2000). Haseman and Elston revisited, *Genetic Epidemiology* **19**, 1–17.

[15] Epstein, M.P., Duren, W.L. & Boehnke, M. (2000). Improved inference of relationship for pairs of individuals, *American Journal of Human Genetics* **67**, 1219–1231.

[16] Fulker, D.W., Cherny, S.S. & Cardon, L.R. (1995). Multipoint interval mapping of quantitative trait loci using sib pairs, *American Journal of Human Genetics* **56**, 1224–1233.

[17] Goldgar, D.E. & Oniki, R.S. (1992). Comparison of a multipoint identity-by-descent method with parametric multipoint linkage analysis for mapping quantitative traits, *American Journal of Human Genetics* **50**, 598–606.

[18] Goldin, L.R., Amos, C.I., Chase, G.A., Goldstein, A.M., Jarvik, G.P., Martinez, M.M., Suarez, B.K., Weeks, D.E., Wijsman, E.M. & MacCluer, J.W. (1999). Genetic Analysis Workshop 11: analysis of genetic and environmental factors in common diseases, *Genetic Epidemiology* **17**, Supplement 1.

[19] Goldin, L.R., Bailey-Wilson, J.E., Borecki, I.B., Falk, C.T., Goldstein, A.M., Suarez, B.K. & MacCluer, J.W. (1997). Genetic Analysis Workshop 10: detection of genes for complex traits, *Genetic Epidemiology* **14**(6).

[20] Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. & Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis, *Nature Genetics* **25**, 12–13.

[21] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[22] Hasstedt, S.J. (1994). *Pedigree Analysis Package*, Revision 4.0, Department of Human Genetics, University of Utah, Salt Lake City.

[23] Heath, S.C. (1997). Markov chain segregation and linkage analysis for oligogenic models, *American Journal of Human Genetics* **61**, 748–760.

[24] Kong, A. & Cox, N.J. (1997). Allele-sharing models: LOD scores and accurate linkage tests, *American Journal of Human Genetics* **61**, 1179–1188.

[25] Konigsberg, L.W., Kammerer, C.M. & MacCluer, J.W. (1989). Segregation analysis of quantitative traits in nuclear families: comparison of three program packages, *Genetic Epidemiology* **6**, 713–726.

[26] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.

[27] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.

[28] Laird, N.M., Horvath, S. & Xu, X. (2000). Implementing a unified approach to family based tests of association, *Genetic Epidemiology* **19**, Supplement 1, S36–S42.

[29] Lander, E.S. & Green, P. (1987). Construction of multilocus genetic maps in humans, *Proceedings of the National Academy of Sciences* **84**, 2363–2367.

[30] Lange, K., Weeks, D. & Boehnke, M. (1988). Programs for Pedigree Analysis: MENDEL, FISHER, and dGENE, *Genetic Epidemiology* **5**, 471–472.

[31] Lathrop, G.M., Lalouel, J.M., Julier, C. & Ott, J. (1984). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination, *Proceedings of the National Academy of Sciences* **81**, 3443–3446.

[32] Matise, T.C., Perlin, M. & Chakravarti, A. (1994). Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map, *Nature Genetics* **6**, 384–390.

[33] McPeek, M.S. & Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data, *American Journal of Human Genetics* **66**, 1076–1094.

[34] Mukhopadhyay, N., Almasy, L., Schroeder, M., Mulvihill, W.P. & Weeks, D.E. (1999). Mega2, a data-handling program for facilitating genetic linkage and association analyses, *American Journal of Human Genetics* **65**, Supplement, A436.

[35] O'Connell, J.R. & Weeks, D.E. (1995). The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance, *Nature Genetics* **11**, 402–408.

[36] O'Connell, J.R. & Weeks, D.E. (1998). PedCheck: a program for identification of genotype incompatibilities in linkage analysis, *American Journal of Human Genetics* **63**, 259–266.

[37] Ott, J. (1986). Linkage probability and its approximate confidence interval under possible heterogeneity, *Genetic Epidemiology* **1**, Supplement, 251–257.

[38] Schaffer, A.A. (1996). Faster linkage analysis computations for pedigrees with loops or unused alleles, *Human Heredity* **46**, 226–235.

[39] Schaid, D.J. (1996). General score tests for associations of genetic markers with disease using cases and their parents, *Genetic Epidemiology* **13**, 423–449.

[40] Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics, *American Journal of Human Genetics* **58**, 1323–1337.

[41] Wijsman, E.M., Almasy, L., Amos, C.I., Borecki, I., Falk, C.T., King, T.M., Martinez, M.M., Meyers, D.,

Neuman, R., Olson, J.M., Rich, S., Spence, M.A., Thomas, D.C., Vieland, V.J., Witte, J.S. & MacCluer, J.W. (2001). Analysis of complex genetic traits: applications to asthma and simulated data, *Genetic Epidemiology* **21**, Supplement 1.

[42] Williams, J.T. & Blangero, J. (1999). Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples, *Genetic Epidemiology* **16**, 113–134.

(*See also* **Disease-marker Association**)

LAURA ALMASY

Software for Sample Survey Data, Misuse of Standard Packages

In the past 15 years, many researchers in the health sciences have become interested in performing primary and secondary analyses using data from complex **sample surveys**. These analyses are often descriptive or analytical, but they may also generate or test hypotheses within the context of a statistical model. Sample survey statisticians are aware that specialized software should be used to analyze complex sample survey data, particularly when analyses are descriptive or analytical and the survey design includes **clustering** [1, 4] (*see* **Software for Sample Survey Data**).

However, some researchers are not aware of the need to use specialized software or, if aware, prefer not to do so because of the need to learn new analytical techniques and software. A common error among data analysts is the inappropriate use of standard statistical software for sample survey data. Further, data analysts may be confused when they realize that there is a difference of opinion, even among sample survey statisticians, as to methods for analysis of sample survey data [2, 3, 4, 8], particularly when using statistical models.

This article uses sample survey data from BRFSS (Behavioral Risk Factor Surveillance System) surveys to illustrate that **biased** point estimates (*see* **Estimation**), inappropriate **standard errors** and **confidence intervals**, and misleading tests of significance (*see* **Hypothesis Testing**) can result from the incorrect use of standard statistical software packages (*see* **Software, Biostatistical**). This article is not a critique of standard statistical software but rather a critique of data analysts inappropriately choosing such software for survey data. Sample survey software has become more widely available in the past decade, giving survey data analysts both the opportunity and the responsibility to choose appropriate software for their analyses. The examples in this article of appropriate and inappropriate (but common) software choices illustrate the importance of using sample survey software for survey data.

Why Specialized Software is Needed

Standard statistical software generally assumes that the observational units have been obtained via **simple random sampling**. Thus, it does not take into account four common characteristics of sample survey data: (i) unequal probability selection of observations, (ii) clustering of observations, (iii) **stratification**, and (iv) **nonresponse** and other adjustments [6, 7]. Point estimates of population parameters are impacted by the value of the analysis weight for each observation. These weights depend upon the selection probabilities through survey design features such as stratification, oversampling and clustering and upon nonresponse adjustments. Incorrectly choosing to use standard statistical software without weighting will yield biased point estimates of population parameters. Estimated **variance** formulas for point estimates based on sample survey data are impacted by clustering, stratification, and the weights. By incorrectly choosing standard statistical software and ignoring these survey design aspects in the analysis, the estimated variance of a point estimate generally is underestimated, sometimes substantially so.

Most standard statistical software can perform weighted analyses, usually via a WEIGHT statement added to the program code. Use of standard statistical software with a weighting variable should yield the same point estimates for population parameters as sample survey software. However, the estimated variance of point estimates generally is not correct since the user is ignoring clustering and stratification. Further, the estimated variance can be substantially wrong, depending upon the particular standard software program being incorrectly used.

Description of BRFSS Surveys

The BRFSS [10] program, established by CDC (**Centers for Disease Control and Prevention**), provides state-level data to estimate the **prevalence** of **risk factors** for disease and poor health. States select a continuous **probability sample** of the adult noninstitutionalized population using some type of **random digit dialing** (RDD) telephone sampling. The Mitofsky–Waksberg RDD technique [11] was used for many years, but list-assisted or directory-based stratified RDD (*see* **Telephone Sampling**)

is commonly used now. Telephone numbers typically are stratified by density of residential telephone numbers, and the high-density stratum is oversampled. Once a residence is reached, almost all states select one adult, with equal probability, to undergo a telephone interview. Each state generally interviews between 1500 and 4500 adults per year.

BRFSS statewide surveys result in an unequal probability sample of adults because only one adult per sampled household is selected and there may be differential sampling fractions of telephone numbers in different strata. Weighting adjustments may be done for first-stage nonresponse (telephone not answered or household screening/enumeration not completed) and for second-stage nonresponse (the selected adult was not interviewed). Further, **post-stratification** of the observations to US Census data is generally done. Hence, each observation (interviewed adult) in the data set has a value for the variable FINALWT (final analysis weight). This value indicates the number of persons in the population represented by that observation. The value of FINALWT varies across observations within a state and between states, sometimes considerably so.

In addition to differential weighting, the Mitofsky–Waksberg RDD method clustered observations by telephone bank (usually defined as a group of 100 telephone numbers with identical area code, prefix, and first two digits of the suffix). Under list-assisted RDD the observations in BRFSS data sets are not clustered. Further, some states use geographic stratification in their sampling process. Although sampling details differ across states and across years, current statewide BRFSS surveys typically are weighted, stratified, and not clustered, whereas older surveys were weighted, clustered, and may be stratified.

This article uses calendar year 1993 BRFSS data on diabetes for the six states given in Table 1, yielding a total sample size of 20 049 observations over the six states. Mitofsky–Waksberg RDD was used in all six states. Presence/absence of diabetes is defined as a Yes/No answer to "Have you ever been told by a doctor that you have diabetes?"; the few observations with other than a Yes/No answer are excluded from all analyses.

Comparing Appropriate and Inappropriate Software Choices

Any sample survey software and any standard statistical software could have been chosen to illustrate these comparisons; numerical results equivalent to Tables 1–3 would have been obtained. A comprehensive and popular statistical software package was

Table 1 Sample size per state and (min, max) and sum of three types of weights per state, 1993 BRFSS Surveys

State	Sample size, (%)	FINALWT (min, max), sum, (%)	NORMWT (min, max), sum, (%)	STNORMWT (min, max), sum, (%)
California	3719	(635, 72 663)	(0.280, 32.1)	(0.104, 11.9)
	(18.6)	22 780 741	10 049	3719
		(50.1)	(50.1)	(18.6)
Florida	3087	(610, 19 131)	(0.269, 8.4)	(0.178, 5.6)
	(15.4)	10 563 183	4659	3087
		(23.2)	(23.2)	(15.4)
Maryland	4361	(70, 3876)	(0.031, 1.8)	(0.082, 4.5)
	(21.8)	3 727 710	1644	4361
		(8.2)	(8.2)	(21.8)
Minnesota	3412	(222, 4182)	(0.098, 1.8)	(0.232, 4.4)
	(17.0)	3 277 173	1446	3412
		(7.2)	(7.2)	(17.0)
Tennessee	3045	(233, 7460)	(0.103, 3.3)	(0.190, 6.1)
	(15.2)	3 747 334	1653	3045
		(8.2)	(8.2)	(15.2)
West Virginia	2425	(118, 2588)	(0.052, 1.1)	(0.211, 4.6)
	(12.1)	1 356 429	598	2425
		(3.0)	(3.0)	(12.1)
Six-state total	20 049	(70, 72 663)	(0.031, 32.1)	(0.082, 11.9)
		45 452 569	20 049	20 049

Table 2 Estimated prevalence (and standard error) of diabetes by state and analysis procedure

	Appropriate choice of sample survey software	Inappropriate choice of standard statistical software			
State	FINALWT	_ONE_	FINALWT	NORMWT	STNORMWT
California	4.48	4.88	4.48	4.48	4.48
	(0.373)	(0.354)	(0.340)	(0.340)	(0.340)
Florida	5.19	5.64	5.19	5.19	5.19
	(0.421)	(0.416)	(0.400)	(0.400)	(0.400)
Maryland	4.96	5.10	4.96	4.96	4.96
	(0.364)	(0.333)	(0.329)	(0.329)	(0.329)
Minnesota	4.23	4.37	4.23	4.23	4.23
	(0.370)	(0.350)	(0.345)	(0.345)	(0.345)
Tennessee	6.19	6.37	6.19	6.19	6.19
	(0.479)	(0.443)	(0.437)	(0.437)	(0.437)
West Virginia	6.04	6.68	6.04	6.04	6.04
	(0.520)	(0.507)	(0.484)	(0.484)	(0.484)
Six-state total	4.86	5.40	4.86	4.86	5.10
	(0.219)	(0.160)	(0.152)	(0.152)	(0.155)

Table 3 Calculated chi-square statistic and (P value) for testing independence of gender and diabetes, by state and analysis procedure

	Appropriate choice of sample survey software	Inappropriate choice of standard statistical software			
State	FINALWT	_ONE_	FINALWT	NORMWT	STNORMWT
California	1.81	0.001	13 396	5.91	2.19
	(0.178)	(0.975)		(0.015)	(0.139)
Florida	2.15	3.44	8436	3.72	2.46
	(0.143)	(0.064)		(0.054)	(0.116)
Maryland	5.48	2.96	5616	2.48	6.57
	(0.019)	(0.085)		(0.116)	(0.010)
Minnesota	0.11	0.11	104	0.05	0.11
	(0.745)	(0.743)		(0.830)	(0.742)
Tennessee	0.57	2.10	802	0.35	0.65
	(0.452)	(0.147)		(0.552)	(0.419)
West Virginia	3.09	2.74	1950	0.86	3.49
	(0.079)	(0.098)		(0.354)	(0.062)
Six-state total	6.07	8.91	28 662	12.64	13.20
	(0.014)	(0.003)		(<0.001)	(<0.001)

used to conduct two standard and common analyses: (1) estimation of a **mean** (**prevalence**) with estimated **standard error** and (2) calculation of a **chi-square test**. Results obtained from these standard analyses demonstrate the error of an inappropriate software choice. SUDAAN Version 8 [9], a specialized package for sample survey and correlated data analysis, was used to demonstrate an appropriate software choice; Taylor Series linearization [5] is used for variance estimation (*see* **Linearization Methods of Variance Estimation**).

Each state's sampling plan was described to SUDAAN in the same way; within a state no stratification was used and observations were clustered in their appropriate primary sampling unit (PSU), a telephone bank. To perform analyses for each state and for the combined states, the six-state concatenated data set was described to SUDAAN as a **stratified** (by state) **multi-stage clustered** survey. The **finite population correction factor** was not used in estimated variance calculations. For those familiar with SUDAAN and BRFSS, the PROC statement included

DESIGN = WR (with replacement sampling at stage one), the NEST statement included the state stratification variable STSTR and the clustering (telephone bank) variable PSU, and the WEIGHT statement included the variable FINALWT.

Inappropriately chosen standard analyses were conducted using four different approaches, all of which incorrectly ignored the clustering and stratification. The four approaches differ in the way the weighting variable FINALWT is handled in the standard analyses. The first standard approach ignored FINALWT and analyzed the data set unweighted; this is equivalent to using the WEIGHT statement with the variable_ONE_ (a variable whose value is 1.0 for every observation in the data set). In Table 1 (column 2) it is shown that, with this approach, the CA sample size is 3719 and it contributes 19% to the total inference population.

The second standard approach used the WEIGHT statement with the variable FINALWT. There is great variability in FINALWT; Table 1 (column 3) indicates its range as 70 to 72 663. In Table 1 it is also shown that using FINALWT implies that the CA sample contributes 50% to the total inference population, rather than only 19% in an unweighted analysis.

The third standard approach used the WEIGHT statement with the variable NORMWT, a normed weight based on FINALWT. Some data analysts claim that using normed weights with standard statistical packages yields results comparable to those from sample survey software. For observation j within state i, let finalwt (i, j) be the value of the variable FINALWT. Then, the value of NORMWT for this observation is defined as:

$$\text{normwt}(i, j) = (20049) * \frac{\text{finalwt}(i, j)}{45\ 452\ 569} \qquad (1)$$

The figure 45 452 569 is the estimated total adult population of the six states, which is the sum of the value of FINALWT over all 20 049 observations (Table 1, column 3). The variable NORMWT has values less than 1.0 and greater than 1.0, and the sum of the values of NORMWT over the entire data set is 20 049, the total sample size. In Table 1 (column 4) it is shown that, with this approach, the CA sample contributes 50% to the total inference population.

The fourth standard approach used the WEIGHT statement with the variable STNORMWT, a second normed weight calculated from FINALWT, where the norming is done within state. Hence, the sum of the values of STNORMWT over all observations within a state equals the sample size for that state (Table 1, column 5). Clearly, the sum of STNORMWT over the entire sample equals the total sample size 20 049. In Table 1 it is shown that, with this approach, the CA sample contributes 19% to the total inference population.

First, SUDAAN DESCRIPT and the four inappropriately chosen standard approaches were compared on a descriptive analysis; that is, estimation of diabetes prevalence (with estimated standard error) for the total population (six states combined) and for each state. The diabetes variable was coded as 1 or 2 for DESCRIPT and coded as 0 (no diabetes) or 100 (have diabetes) for the standard software that estimated means.

Secondly, SUDAAN CROSSTAB and the four inappropriately chosen standard approaches were compared on a chi-square test of the **null hypothesis** that gender and diabetes are statistically independent. These analyses were performed for the total population and for each state, with diabetes coded as a categorical variable (1, 2).

Results

Descriptive Analyses

Sample survey software yields correct point estimates for diabetes prevalence and for estimated standard errors (Table 2, column 2). However, the incorrect choice of unweighted standard software (column 3) causes diabetes prevalence to be overestimated by about 10% for the total population (5.40 versus 4.86%) and for half of the states. Note also that the estimated standard errors in Table 2 are smaller by incorrectly using unweighted standard software. For the entire population, the correct estimated standard error is 35% larger than the standard error estimated by the incorrect use of unweighted standard software (0.219 versus 0.160). The combination of the biased point estimate and underestimation of the standard error could result in quite misleading confidence intervals for the prevalence of diabetes.

In Table 2 (columns 4 and 5), it is shown that standard software with FINALWT or NORMWT gives identical results, with correct point estimates for diabetes prevalence. However, the incorrect choice of standard software for analysis still results in estimated standard errors that are too low. The magnitude of

underestimation of the standard error by using standard software with FINALWT or with NORMWT is somewhat worse than with unweighted standard software. Thus, standard statistical software with a WEIGHT statement using FINALWT or NORMWT yields **unbiased** point estimates of population parameters, but it yields incorrect estimated standard errors.

In Table 2 (column 6), it is shown that standard software with STNORMWT gives identical results to standard software with FINALWT or NORMWT for state-specific analyses, but yields a biased point estimate for the total population along with an underestimated standard error.

Chi-square Analyses

The chi-square analysis tests the null hypothesis that the prevalence of diabetes is the same for males and females. In Table 3 (columns 2 and 3), it is shown that the incorrect choice of unweighted standard software, compared to sample survey software, yields a higher value for the chi-square statistic for the entire population, giving a smaller ***P* value** (0.003 versus 0.014). A comparison state by state shows no consistent pattern; the *P* value for unweighted standard software is sometimes higher and sometimes lower than for sample survey software.

In Table 3 (columns 2 and 4), it is shown that the incorrect choice of standard software with FINALWT yields an unreasonably large value of the chi-square statistic for the total population and for each state. *P* values are not included in Table 3 for these very large chi-square statistics. The standard software with FINALWT considers the sample size to be the sum of the values of FINALWT (i.e. 45 452 569) as opposed to the actual sample size of 20 049. This is the reason for the very large values of the chi-square statistic.

In Table 3 (columns 2 and 5), it is shown that the incorrect choice of standard software with NORMWT yields a chi-square value for the six-state area that is twice as large (12.64 versus 6.07). However, this relationship between sample survey and standard software with NORMWT does not hold for each of the six states in Table 3. Standard software with NORMWT yields a larger chi-square statistic value for some states but a smaller value for other states. This occurs because, within each state, the standard software considers the sample size as the sum of NORMWT. Hence, the sample size for CA is artificially inflated to 10 049 from 3719, whereas the sample size for West Virginia (WV) is artificially deflated to 598 from 2425 (see Table 1). Thus, the chi-square statistic using standard software with NORMWT, compared to survey software, is much larger for CA but much smaller for WV.

In Table 3 (columns 2 and 6), it is shown that the inappropriate choice of standard software with STNORMWT yields a chi-square statistic for the total population about twice as large (13.20 versus 6.07). For each state, the chi-square statistic based on standard software with STNORMWT is about 15–20% larger than provided by survey software. Because STNORMWT is normed within a state, the sum of the weights reflects the statewide sample size. Hence, the standard software with STNORMWT shows the common pattern that standard statistical software packages generally calculate a larger value of the chi-square statistic compared to sample survey software.

Discussion

Unweighted Analyses with Standard Statistical Software

Although the empirical evidence in this article is based only on one type of survey (BRFSS), only on six states and only on 1993 data, the findings are consistent with other similar investigations [1]. Using a standard statistical package with unweighted analyses to analyze sample survey data generally will yield (i) biased point estimates of population parameters, (ii) underestimates of the standard error for point estimates, (iii) confidence intervals on population parameters that are too narrow, and (iv) tests of significance that are too likely to reject the null hypothesis because the standard errors or variability of statistics generally are underestimated.

The extent of the bias in unweighted point estimates will depend upon the particular data set and is related to the variability of the FINALWT variable. If FINALWT has little variability in the data set, then an unweighted point estimate will be close to a weighted point estimate. In the six-state BRFSS data set, the value of FINALWT ranged from 70 to 72 663 over the six states. This extreme variability in the value of FINALWT is primarily due to varying sampling fractions across the states; that is, a small variation in state sample size (2400 to 4400) but widely different statewide populations (1.4 to 22.8 million).

Another factor that contributes to the bias of point estimates of population parameters based on unweighted analyses is the relationship between the value of FINALWT and the variable being analyzed. In the data set used here, the value of FINALWT is primarily influenced by the sampling fraction in each state; you could say that certain states are "oversampled". If states were strongly related to the analysis variable (diabetes), then point estimates of diabetes prevalence from unweighted analyses could be seriously biased. In this data set, the estimated statewide prevalences of diabetes do not differ dramatically, ranging from 4 to 6%. If blacks had been oversampled within each state to a large extent, then the bias in estimated diabetes prevalence using unweighted analyses would be substantial and positive, since blacks have a higher prevalence of diabetes than do whites.

In addition to potentially biased point estimates from unweighted standard analyses, standard errors, and other measures of variability are generally underestimated because of clustering and variability in FINALWT. The intracluster **correlation** coefficients in older BRFSS data sets using Mitofsky–Waksberg RDD are generally positive but not substantial. This might be expected because most states only had about three completed interviews per PSU (telephone bank) (*see* **Telephone Sampling**) with Mitofsky–Waksberg RDD. Variability in FINALWT, and not clustering, likely is the most important factor contributing to the higher estimated variances from sample survey software using this BRFSS data set. If other sample survey data sets had been used with a higher degree of intracluster correlation, an incorrect choice of unweighted standard analyses would have produced even smaller estimates of variability, compared to sample survey software.

Weighted Analyses with Standard Statistical Software

Using standard statistical software with weighted analyses (FINALWT or NORMWT) produces unbiased point estimates of prevalence for the entire population over all six states and for any strata (states) of interest. Although not illustrated, these weighted analyses also yield unbiased point estimates of diabetes prevalence among subpopulations based on other characteristics, such as race or gender, where the subpopulations contain observations from all or some strata. Hence, either of these two weighted standard approaches is fine if only point estimates of prevalence are desired. However, weighted standard statistical software (using FINALWT or NORMWT) tends to underestimate the standard error of estimated prevalences. The degree of underestimation depends upon the size of the intracluster correlation coefficient for the variables being analyzed; the higher the intracluster correlation, the more serious the underestimation of the variability. Weighted analyses with standard statistical software (using NORMWT or FINALWT) may be a reasonable analytical approach for point estimates of population parameters under the following condition: all intracluster correlation coefficients are near zero.

The inappropriate choice of standard software with FINALWT gives substantially incorrect results for chi-square tests because the sample size is assumed to be the population size, that is, the sum of the values of FINALWT. Whether this is true in all standard statistical packages depends upon the packages' default options for weighted analyses in chi-square tests.

The inappropriate choice of standard software with NORMWT gives a larger chi-square statistic for the entire population, about twice as large. However, this procedure yields substantially incorrect chi-square statistics for state-specific analyses. The state specific analyses are wrong because the standard software assumes an incorrect sample size for the state analyses. This will occur also whenever subpopulations are analyzed using NORMWT and the variable that defines the subpopulation is related to the value of FINALWT.

The inappropriate choice of standard software with the second normed weight, STNORMWT, gives more reasonable values for the chi-square statistic for state level analyses, although the chi-square statistics were always larger than with survey software. However, if the weight STNORMWT is used for analyses over the entire population, a biased point estimate is obtained for population parameters.

Conclusions

The empirical results above illustrate that using standard statistical software with the weights FINALWT or NORMWT are the only two of the four inappropriate approaches considered that yield unbiased point estimates for population and subpopulation parameters. All four inappropriate choices for analysis using standard statistical software yield incorrect

standard errors and tests of significance, generally in the direction of underestimating variability of statistics. In particular, using normed weights with standard statistical software, a sample survey data analytic approach advocated by some data analysts, is problematic with respect to variance estimation and in some instances can give biased point estimates of population parameters. It is recommended that sample survey software be used to analyze sample survey data, especially for estimation of population parameters, descriptive analyses, and analytical analyses. Under certain circumstances, standard statistical packages can be used to provide results approximately equal to the results obtained from sample survey software. However, recognition of these circumstances and awareness of the potential pitfalls of using standard statistical packages requires detailed information about the characteristics of the survey data set (e.g. sampling plan, weighting scheme, and intracluster correlation) as well as knowledge of the particular formulas and default options used by the standard statistical software for weighted analyses. In the end, it seems easier and less time consuming to use software developed for sample survey data analysis.

Acknowledgments

This work was partially supported by CDC via the Division of Diabetes Translation and the Division's 1996 Conference Planning Committee. An invited paper based on this work was presented at the 1996 Diabetes Translation Conference, "Health Care in Transition: Diabetes as a Model for Public Health", held in Washington DC from March 31 to April 3, 1996. All statements are the sole responsibility of the author.

References

[1] Brogan, D. (2004). Sampling error estimation for survey data, in Household Surveys in Developing and Transition Countries: Design, Implementation and Analysis, S. Ibrahim Yansaneh & Graham Kalton, eds. United Nations, New York, Chapter 21.

[2] Graubard, B.I. & Korn, E.L. (1996). Modeling the sampling design in the analysis of health surveys, *Statistical Methods in Medical Research* **5**, 263–281.

[3] Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

[4] Korn, E.L. & Graubard, B.I. (1999). *Analysis of Health Surveys*. Wiley, New York.

[5] LaVange, L.M., Stearns, S.C., Lafata, J.E., Koch, G.G. & Shah, B.V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples, *Statistical Methods in Medical Research* **5**, 311–329.

[6] Levy, P.S. & Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*, 3rd Ed. Wiley, New York.

[7] Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Imprint of Brooks/Cole Publishing Company, Pacific Grove.

[8] Pfeffermann, D. (1996). The use of sampling weights for survey data analysis, *Statistical Methods in Medical Research* **5**, 239–261.

[9] Research Triangle Institute. (2001). *SUDAAN User's Manual*, Release 8.0. Research Triangle Institute, Research Triangle Park.

[10] Siegel, P.Z., Brockbill, R.M., Frazier, E.L., Mariolis, P., Sanderson, L.M. & Waller, M.N. (1991). Behavioral risk factor surveillance, 1986–1990, *Morbidity and Mortality Weekly Report CDC Surveillance Summaries* **40**, 1–23.

[11] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **73**, 40–46.

DONNA BROGAN

Software for Sample Survey Data

Introduction

A **sample survey** is a process for collecting data on a sample of observations that are selected from the population of interest using a **probability sample** design. In sample surveys, certain methods are often used to improve the precision and control the costs of survey data collection. These methods introduce a complexity to the analysis, which must be accounted for in order to produce **unbiased** estimates (*see* **Estimation**) and their associated levels of precision. This article provides a brief introduction to the impact these design complexities have on the sampling **variance**, and summarizes the characteristics and availability of software to carry out analysis on sample survey data.

Complex Sample Designs

Statistical methods for estimating population parameters and their associated variances are based on assumptions about the characteristics and underlying distribution of the observations. Statistical methods in most general-purpose statistical software tacitly assume that the data meet certain assumptions. Among these assumptions are that the observations were selected independently and that each observation had the same probability of being selected. Data collected through surveys often have sampling schemes that deviate from these assumptions. For logistical reasons, samples are often clustered geographically to reduce costs of administering the survey, and it is not unusual to sample households and then subsample families and/or persons within selected households (*see* **Multistage Sampling**). In these situations, sample members are not selected independently, nor are their responses likely to be independently distributed.

In addition, a common survey sampling practice is to oversample certain population subgroups to ensure sufficient representation in the final sample to support separate analyses. This is particularly common for certain policy-relevant subgroups, such as ethnic and racial minorities, the poor, the elderly, and the disabled. In this situation, sample members do not have equal probabilities of selection. Adjustments to sampling weights (the inverse of the probability of selection) to account for **nonresponse** as well as other weighting adjustments (such as **poststratification** to known population totals), further exacerbate the disparity in the weights among sample members.

Impact of Complex Sample Design on Sampling Variance

Because of these deviations from standard assumptions about sampling, such survey sample designs are often referred to as *complex*. While **stratification** in the sampling process can decrease the sampling variance, **cluster sampling** and unequal selection probabilities generally increase the sampling variance associated with resulting estimates. The *sampling variance* is a measure of the variation of an estimator attributable to having sampled a portion of the full population of interest. It is a measure of the variation of the *estimate* of a population parameter over repeated samples. The sampling variance becomes smaller as the sample size increases, and is zero when the full population is observed. The sampling variance differs from the *population variance*, which measures the variation among *observations* in the population, and is a constant, independent of any sampling issues.

Not accounting for the impact of the complex sample design can lead to an underestimate of the sampling variance associated with an estimate. Therefore, while standard procedures in general-purpose statistical software packages can usually produce an unbiased weighted survey estimate (*see* **Software, Biostatistical**), it is quite possible to overestimate the precision of such an estimate when using one of these standard procedures to analyze survey data.

The magnitude of this effect on the variance is commonly measured by what is known as the **design effect** [16]. The design effect is the sampling variance of an estimate, accounting for the complex sample design, divided by the sampling variance of the same estimate, assuming a sample of equal size had been selected as a simple random sample. A design effect of one indicates that the design had no impact on the variance of the estimate. A design effect of unity indicates that the design has increased the variance, and a design effect less than one indicates that the design actually decreased the variance of the estimate. The

design effect can be used to determine the *effective* sample size, simply by dividing the nominal sample size by the design effect. The effective sample size gives the number of observations that would yield an equivalent level of precision from an independent and identically distributed (i.i.d.) sample. For example, an estimate from a complex sample of size 1500 that has a design effect of 1.5 is equivalent (in terms of precision) to that same estimate from a **simple random sample** of size 1000. The benefits of the complex design in this case would be weighed against the cost of effectively losing 500 observations.

For complex designs, the exact computation of the variance of an estimate is not always possible. When estimating a total or a **mean** (when the denominator is known), the estimate is in linear form; that is, $\overline{y} = \sum_{i=1}^{n} y_i/n$ can be expressed in the form $\hat{\theta} = \sum_{i \in S} \beta_i y_i$. When an estimate is in linear form, a standard formula for the mean square error of a linear estimate can be applied to calculate the variance; however, for a weighted mean estimate, the form is no longer linear. If w_i is the weight associated with sample member i, then the weighted mean is calculated as: $\overline{y}_w = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$. The mean estimate is now a **ratio estimate** with a **random variate** in both the numerator and denominator because the sample weights depend on the units selected and differ from sample to sample.

Variance Estimation Methods

Several approaches have historically been used to compute an approximation of the true variance of an estimate when the sample deviates from i.i.d assumptions. These techniques fall into two general categories: (1) the Taylor series linearization technique (*see* **Linearization Methods of Variance Estimation**) and (2) replication techniques. Both of these were first proposed in the literature for use with survey data in the 1960s. While over the years government statistical agencies, academic departments, and private survey organizations implemented their own **algorithms** and developed their own software for carrying out these techniques, several software packages have emerged for public use, first for mainframe computer applications, and now for use on personal computers and in other computing environments. The variance estimation software available to the public uses one or the other of the two general strategies for variance estimation mentioned above. What follows is a brief description of these two types of techniques. For more detailed descriptions of these techniques, the reader is advised to consult the references given. Overviews of these techniques can be found in [1, 10, 15, 19, 25, 31].

Because estimates of interest in sample surveys are nonlinear, one approach is to *linearize* such estimates using a Taylor series expansion. This approach was first suggested for use with survey estimates in 1968 by Tepping at the US Bureau of the Census [29]. In essence, the estimate is rewritten in the form of a Taylor series expansion, and an assumption is made that all higher-order terms are of negligible size, leaving only the first-order (linear) portion of the expanded estimate. A standard formula for the **mean square error** of a linear estimate can then be applied to the linearized version to approximate the variance of the estimate. This approximation works well to the extent that the assumption regarding the higher-order terms is correct; see also [32]. Note that, with this approach to variance estimation, a separate formula for the linearized estimate must be developed for each type of statistical estimator. Most survey data analysis software includes the most widely used estimates (such as means, proportions, ratios, and regression coefficients). Binder [2] introduced a general approach that can be used to derive Taylor series approximations for a wide range of estimators, including Cox **proportional hazards** and **logistic regression** coefficients.

Replication techniques are a family of approaches that take repeated subsamples, or *replicates*, from the data, recompute the weighted survey estimate for each replicate, and then compute the variance based on the deviations of these replicate estimates from the full-sample estimate. This approach was first suggested for use with survey data in 1966 by McCarthy at Cornell University as part of his work with the National Center for Health Statistics [22]. The most commonly used replication techniques are the *balanced repeated replication* (BRR) method and the **jackknife method**. Robert Fay at the US Bureau of the Census has developed his own replication technique for this purpose as well [9] (*see* **Resampling Procedures for Sample Surveys**). Other techniques less commonly used for this purpose are **bootstrapping** [24] and the random group method [13]. All replication techniques require the computation of a set of replicate weights, which are the analysis weights recalculated for each of the

replicates selected so that each replicate appropriately represents the same population as the full sample. Such computing-intensive techniques became practical only as the computing capacity on mainframes and then personal computers increased. Unlike the Taylor series method, replication methods do not require the derivation of variance formulas for each statistical estimate because the approximation is a function of the sample, not of the estimate.

With balanced repeated replication (also known as balanced half-sampling), forming a replicate involves dividing each sampling stratum into two primary sampling units (PSUs) (*see* **Sampling in Developing Countries**), and randomly selecting one of the two PSUs in each stratum to represent the entire stratum [17, 18, 23]. The jackknife repeated replication approach involves removing one stratum at a time to create each replicate [11, 30]. Fay's method is similar to the BRR approach, except that, instead of selecting one of two PSUs in each stratum, the weights of one of the two PSUs in each stratum are multiplied by a factor k between 0 and 2 and the weights of the other PSUs are multiplied by a factor of $2 - k$ [9]. See also [7, 8] for a discussion of replication methods.

Software Packages

At the time of this writing (spring 2003), a number of packages are available to the public designed specifically for use with sample survey data. A website that is maintained by Alan Zaslavsky at Harvard University (`www.fas.harvard.edu/~stats/survey-soft/survey-soft.html`) contains a list of a dozen survey analysis variance estimation packages, along with their features, contact information, and comparative reviews. Four of these packages have been developed by government agencies (Statistics Canada, US **Centers for Disease Control** and Prevention, and two from US Bureau of the Census); three have been developed in academia (University of Essex, University of Michigan, Iowa State University); and five have been developed by private organizations (including SAS Institute, Stata, Research Triangle Institute, and Westat). Many governmental statistical agencies (including those in the United States, Canada, Sweden, Holland, and France) have developed their own sample survey software to meet their needs. This software is sometimes available from the agencies for use by others, but not marketed as such.

All but one of the packages described on the website use the Taylor series approach to variance estimation, and several of these also offer replication methods. One supports only replication methods. All are available for PC (Windows or DOS) platforms, but several offer other platforms such as Macintosh, Linux, Unix, and Sun/Solaris. Two of the packages (other than SAS [26]) require SAS in order to run. All reportedly handle complex sample designs that include stratification and cluster sampling. Some are more sophisticated and handle multistage sampling, without-replacement sampling (*see* **Sampling With and Without Replacement**), and unequal probabilities at each stage of selection. The range of estimate types available on these packages range from descriptive statistics (such as totals, means, ratios, and proportions) to **multivariate analytical** techniques (such as **linear regression** logistic regression, and proportional hazards models). The focus in this article is on several of the more commonly used packages (SUDAAN, WesVar, Stata, and SAS/STAT). The reader is encouraged to visit the web page mentioned above to find out more about CENVAR, CLUSTERS, **Epi Info** GES, IVEware, PCCARP, R Survey, and VPLX.

SUDAAN (developed by Babu Shah and others at the Research Triangle Institute (RTI)) started out as software called STDERR in 1970 for use on an IBM mainframe. Starting in 1976, the software was further developed to carry out a wider array of procedures (RATIOEST, RTIFREQS, SESUDAAN, SURREGR, and RTILOGIT), some with the support of various government statistical agencies with whom the RTI was working. SUDAAN primarily uses the Taylor series approach to variance estimation, but has added replication techniques as options. Because SUDAAN was originally developed as an SAS procedure, its syntax in batch mode is similar to that of SAS. There is a bit of a learning curve in how to specify the design parameters. It is available as a stand-alone package or as a SAS-callable procedure.

Among the four packages discussed here, SUDAAN has the greatest capabilities in terms of the types of sample designs it accommodates. It can handle stratification, cluster sampling, multistage sampling, without-replacement sampling, and unequal probabilities of selection at each stage. Specifying without-replacement sampling with unequal probabilities at the first stage (the most complicated design structure) requires that the

user compute joint-inclusion probabilities for each possible pair of PSUs within each first-stage stratum. Specifying without-replacement sampling of any kind (three types are available) requires that the user supply frame counts at each stage of selection. Being able to specify without-replacement sampling becomes a factor when the sampling fraction at the first stage is relatively high, so that the **finite population correction** factor can be incorporated.

Westat first developed its variance estimation software for the work that they were doing with the US Department of Transportation's National Accident Sampling System. The first software was developed as SAS procedures for use on IBM mainframes in the 1980s by David Morganstein and was called NASSVAR. Westat's current survey analysis software, WesVar, is a stand-alone PC package that uses replication methods to estimate variances. WesVar has the option of one of several replication techniques, including BRR, two variants of the jackknife approach, and Fay's method. To use this package, the user must supply replicate weights, or supply full-sample weights and allow the software to create the replicate weights. WesVar can handle most complex sample designs (stratification, cluster sampling, multistage sampling), but does not have the capability of handling without-replacement sampling at the first stage. It should be noted, however, that in most federal surveys, with-replacement sampling of PSUs is commonly assumed in variance estimation even if the PSUs were selected without replacement (generally a conservative assumption). WesVar also has as part of its package, the capability to **poststratify** the weights to user-specified counts to account for **sample frame** shortcomings, as does SUDAAN.

Stata (Stata Corporation, College Station, Texas) added survey data analysis capabilities to an existing general-purpose statistical package in 1995. Stata uses the Taylor series approach to variance estimation, although there are some user-written additions that incorporate Jackknife and BRR methods. Stata is very popular due to its ease of use, good user support, programmability, comprehensive set of analytical procedures, and relatively lower price-tag (compared to SUDAAN and SAS/STAT). In Stata, the Taylor series capability can be added to any user-programmable estimator. Like WesVar, Stata can incorporate most complex sample designs, but cannot handle a two-stage without-replacement design with sampling at both stages; that is, it assumes with-replacement sampling of PSUs.

SAS has added a few procedures to its SAS/STAT software over the last few years for survey data. The SURVEYSELECT procedure allows the user to select probability samples of various designs – from simple random samples to complex multistage samples with stratification, clustering, and unequal probabilities of selection. It also includes two Taylor series based procedures to estimate totals, means, and ratios (SURVEYMEANS) and regression coefficients (SURVEYREG). Future plans are for them to add similar capabilities for analyzing frequency data and logistic regression.

Issues in Selecting and Using Sample Survey Software

There are several ways to evaluate the qualities of such software packages when deciding which one to use. The user must first evaluate his or her analytical needs, such as sample design complexities, statistical procedures needed, and computing environment, and then decide among those that will be the easiest to use. None of the existing packages meet all of the recommended criteria in the following paragraphs. For the packages mentioned above, or any that use either of the two variance estimation approaches described, there is no need for the user to be concerned about whether the method used to estimate the variances is statistically sound when choosing among packages. Both the Taylor series linearization approach and replication methods were derived from well-accepted approaches previously applied to other statistical problems. Each has certain circumstances in which its approximation of the variance is better than that of the other, and certain replication techniques work better than others, depending on the sample design [3, 25]. Empirical evaluations (using national survey data such as the Current Population Survey and the National Health Interview Survey) have revealed little difference in the estimates of the variance using the different approaches [1, 11, 19].

From a practical standpoint, the software must work in the computing environment in which the user works. If the user works primarily on a mainframe or on a Macintosh, he or she will find that most of the software packages are available only for use on DOS- or Windows-based personal computers. And if

one routinely works with certain types of statistical estimates, such as those resulting from multivariate logistic regression or other nonlinear models, the user may find that many of the software packages do not have the full array of statistical procedures that are currently available in the traditional statistical software packages such as SAS and SPSS [28]. In addition, the software package should be able to import the user's data sets, whether they were created using standard statistical, database, or spreadsheet packages, as well as text (ASCII) files.

The software package should be relatively easy to use; however, there is sometimes a trade-off between this ease of use and the package's capabilities in terms of handling more complex designs or analyses. Another trade-off of being easy to use is a package's propensity for being used inappropriately. Some packages are relatively straightforward to use, with a menu-driven or Windows-type approach, enabling someone unfamiliar with the underlying assumptions to unknowingly get estimates that are inappropriate for his or her design. Other packages are less "user-friendly," and require writing lines of code to be submitted as a batch job, which generally requires the user to learn more up front about the package itself. However, when a large number of similar analyses will be run, it is often easier to run analyses in batch mode than through a menu-driven or Windows-type mode. Software packages should have an option for a batch mode of execution.

In any case, user support should be readily available through thorough and well-written documentation, helpful error messages, a complete set of "help" screens, and prompt assistance from the software provider via telephone and/or electronic mail. Support from fellow users, who may communicate with one another via listservers or other means, can also be quite helpful. It should be kept in mind that most packages were initially developed to accommodate a certain type of sample design and a certain type of user (perhaps a particular government agency and a particular survey). This tends to make the packages less user-friendly to those with different data and analysis needs. Packages should have the capacity to handle nonstandard designs, or provide guidance in the documentation as to the most appropriate design to specify in these circumstances and what the consequences are (such as an overestimate of the variance). On the other hand, for even the most sophisticated packages, it should not be cumbersome to specify a relatively simple design. The statistical package should provide technical documentation, including the formulas used for point estimates and the variance estimates.

Currently, it is commonly the case that a user creates a data file using SAS, SPSS, or some other general-purpose statistical package, and then imports the file into one of these specialized packages. Unless they are part of a general statistical package (or are "callable" procedures from such a package), the specialized packages generally do not allow for much data editing, variable construction, recoding, or sorting. Depending on the specialized package and the estimates desired, it may be necessary to then take the output from the specialized package and carry out further data manipulation in the original general-purpose (or yet another) statistical package to obtain the needed estimates. Sometimes, to carry out subgroup analyses, it is necessary to create separate files for each subgroup and go through the entire process for each subgroup. Over the last several years, variance estimation capabilities for sample survey data have been added to general-purpose statistical packages such as SAS and Stata, and SPSS is planning to add survey data analysis capabilities as well.

Several of the software packages currently available or under development are being made available free to users via the Internet, but sometimes offer less in the way of support and training. Others can run over US $1000 for a single-use license, presumably providing more comprehensive technical support and training for users and notification regarding upgrades. Training itself can be in the form of formal in-person training courses (which can be expensive), Internet-based training, or documentation that is comprehensive enough to use as a training manual. In many cases, documentation is merely a reference manual, and the user must learn how to use the package from a formal training course or by working with someone familiar with the package.

The more difficult and/or expensive a software package is to obtain, learn, and use, the less likely it would be that analysts are going to use it. Many analysts do not even realize they should use weights when deriving estimates, let alone use specialized software to estimate the variances correctly. It should be made clear to users of public use data files, through the accompanying documentation, that it is necessary to account for the design when creating estimates. If it is impossible, for confidentiality reasons, to

provide variables on these files that designate the stratum, PSU, and weight for each observation, then the agency, company, or department supplying the file should be willing and able to provide **standard errors** or design effects for certain variables on request, or should provide generalized variance curves, tables of standard errors or design effects for a wide array of variables, or at the very least provide the average design effect for certain types of variables for certain subgroups [4]. If there are no such confidentiality concerns, then the public use file should come with variables for stratum, PSU, and weight, that are clearly marked as such. Ideally, such data files would come with a set of replicate weights as well. It is unreasonable to expect secondary data users to derive a rather large set of replicate weights on their own. As mentioned above, WesVar has a procedure that can be used create replicate weights under certain circumstances. (It cannot create replicate weights that adjust for unit nonresponse or other weighting adjustments other than poststratification.)

In general, to run any of these specialized packages, one needs to specify variables on the file that correspond to sampling stratum, PSU, and analysis weight. In addition, the file needs to be sorted by stratum and then the PSU within stratum. Further, there generally have to be at least two PSUs in each stratum. (If this is not the case, then the user needs to collapse across strata.) The user should have a good understanding of the design. Was stratification employed? How many sampling stages were there? At each stage of sampling, were the units selected with or without replacement? Were the units selected with equal probability or was there disproportionate sampling (such as oversampling or sampling with probability proportionate to size)? The user should also know which variables are continuous or interval, versus categorical or ordinal (*see* **Measurement Scale**). For categorical or ordinal variables, the user should know the number of categories of each. Examples of statistical analysis using Stata and WesVar can be found in Levy and Lemeshow [21].

Several papers have been published that compare the various software packages. Because many of these packages have evolved over time, many of the criticisms and comparisons found in these papers are no longer valid [5, 6, 15].

There is an ongoing debate as to whether the sample design must be considered when deriving statistical models (as opposed to estimates of means, proportions, totals, and ratios) based on sample survey data. Analysts interested in using statistical techniques such as linear regression, logistic regression, **survival analysis** or **categorical data analysis** on survey data are divided as to whether they feel it is necessary to use specialized software. The model-based analysts argue that, as long as the model is specified correctly, they can proceed without recognizing aspects of the survey design (such as stratification, clustering, and unequal selection probabilities), and can therefore use standard statistical packages. The design-based analysts argue, to the contrary, that it is important to account for the survey design when estimating models. The debate between these two factions has been ongoing for quite a while and is not likely to be resolved soon [12, 14, 20, 27]. A compromise position adopted by some is to use standard statistical software in modeling analyses, but to incorporate into the model the variables that were used to define the strata, the PSUs and the weights.

Contact information for the providers of the specialized software mentioned are found after the references under the name of the software.

References

[1] Bean, J. (1975). Distribution and properties of variance estimators for complex multistage probability samples: an empirical distribution, in *Vital and Health Statistics*, Series 2 Number 65, National Center for Health Statistics, Rockville, MD, pp. 1–46.

[2] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review* **51**, 279–292.

[3] Brillinger, D.R. (1977). Approximate estimation of the standard errors of complex statistics based on sample surveys, *New Zealand Statistician* **11**(2), 35–41.

[4] Burt, V.L. & Cohen, S.B. (1984). A comparison of alternative variance estimation strategies for complex survey data, in *Proceedings of the American Statistical Association Survey Research Methods Section*, Philadelphia, PA.

[5] Carlson, B.L., Johnson, A.E. & Cohen, S.B. (1993). An evaluation of the use of personal computers for variance estimation with complex survey data, *Journal of Official Statistics* **9**(4), 795–814.

[6] Cohen, S.B., Xanthapoulos, J.A. & Jones, G.K. (1988). An evaluation of statistical software procedures appropriate for the regression analysis of complex survey data, *Journal of Official Statistics* **4**, 17–34.

[7] Dippo, C.S., Fay, R.E. & Morganstein, D.H. (1984). Computing variances from complex samples with replicate weights, in *Proceedings of the American Statistical*

Association Survey Research Methods Section, Philadelphia, PA.

[8] Fay, R.E. (1984). Some properties of estimates of variance based on replication methods, in *Proceedings of the American Statistical Association Survey Research Methods Section*, Philadelphia, PA.

[9] Fay, R.E. (1990). VPLX: Variance estimates for complex samples, in *Proceedings of the American Statistical Association Survey Research Methods Section*, Anaheim, CA.

[10] Flyer, P., Rust, K. & Morganstein, D. (1989). Complex survey variance estimation and contingency table analysis using replication, in *Proceedings of the American Statistical Association Survey Research Methods Section*, Washington, DC.

[11] Frankel, M.R. (1971). *Inference from Survey Samples*. Institute for Social Research, the University of Michigan, Ann Arbor.

[12] Groves, R. (1989). *Survey Errors and Survey Costs*. John Wiley, New York.

[13] Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample Survey Methods and Theory, Volume I: Methods and Applications*. Wiley, New York (Section 10.16).

[14] Hansen, M.H., Madow, W.G. & Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association* **78**(384), 776–793.

[15] Kaplan, B., Francis, I. & Sedransk, J. (1979). A comparison of methods and programs for computing variances of estimators from complex sample surveys, in *Proceedings of the American Statistical Association Survey Research Methods Section*, Washington, DC, pp. 97–100.

[16] Kish, L. (1965). *Survey Sampling*. John Wiley and Sons, New York, p. 162.

[17] Kish, L. & Frankel, M. (1968). Balanced repeated replications for analytical statistics, in *Proceedings of the American Statistical Association Social Statistics Section*, Pittsburgh, PA, pp. 2–10.

[18] Kish, L. & Frankel, M.R. (1970). Balanced repeated replications for standard errors, *Journal of the American Statistical Association* **65**(331), 1071–1094.

[19] Kish, L. & Frankel, M.R. (1974). Inference from complex samples, *Journal of the Royal Statistical Society* **B**(36), 1–37.

[20] Korn, E. & Graubard, B. (1995). Analysis of large health surveys: accounting for the sample design, *Journal of the Royal Statistical Society* **A**(158), 263–295.

[21] Levy, P.S. & Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*, 3rd Ed., Wiley and Sons, New York.

[22] McCarthy, P. (1966). Replication: an approach to the analysis of data from complex surveys, in *Vital and Health Statistics*, Series 2, Number 14, National Center for Health Statistics, Washington, DC.

[23] McCarthy, P. (1969). Pseudoreplication: Further evaluation and application of the balanced half-sample technique, in *Vital and Health Statistics*, Series 2, Number 31, National Center for Health Statistics, Washington, DC.

[24] Rao, J.N.K. & Wu, C.F.J. (1984). Bootstrap inference for sample surveys, in *Proceedings of the American Statistical Association Survey Research Methods Section*, Philadelphia, PA.

[25] Rust, K. (1985). Variance estimation for complex estimators in sample surveys, *Journal of Official Statistics* **1**(4), 381–397.

[26] SAS Institute, Inc. (1994). *SAS System for Windows*, Release 6.10 Ed., SAS, Inc., Cary, NC.

[27] Skinner, C.J., Holt, D. & Smith, T.M.F. (1989). *Analysis of Complex Surveys*. John Wiley, New York.

[28] SPSS, Inc. (1988). *SPSS/PC+ V2.0 Base Manual*, SPSS, Inc., Chicago.

[29] Tepping, B.J. (1968). Variance estimation in complex surveys, in *Proceedings of the American Statistical Association Social Statistics Section*, Pittsburgh, PA, pp. 11–18.

[30] Tukey, J.W. (1958). Bias and confidence in not-quite large samples: abstract, *Annals of Mathematical Statistics* **29**, 614.

[31] Wolter, K.M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

[32] Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association* **66**(334), 411–414.

Further Reading

Brogan, D., Flagg, E., Deming, M. & Waldman, R. (1994). Increasing the accuracy of the expanded programme on immunization's cluster survey design, *Annals of Epidemiology* **4**(4), 302–311.

Dean, A.G., Dean, J.A., Coulombier, D., Brendel, K.A., Smith, D.C., Burton, A.H., Dicker, R.C., Sullivan, K., Fagan, R.F. & Arner, T.G. (1995). *Epi Info, Version 6: A Word Processing, Database, and Statistics Program for Public Health on IBM-Compatible Microcomputers*. Centers for Disease Control and Prevention, Atlanta.

Lepkowski, J.M., Bromberg, J.A. & Landis, J.R. (1981). A program for the analysis of multivariate categorical data from complex sample surveys, in *Proceedings of the American Statistical Association Statistical Computing Section*, Detroit, MI.

Software Contact Information

SAS/STAT: `www.sas.com`
Stata: `www.stata.com`
SUDAAN: `www.rti.org/sudaan/`
WesVar: `www.westat.com/wesvar/`
Links to other survey analysis software can be found at `www.fas.harvard.edu/~stats/survey-soft/survey-soft.html`.

BARBARA LEPIDUS CARLSON

Software Reliability

The demand for complex software systems has increased more rapidly than the ability to design, implement, test, and maintain them, and the reliability of software systems has become a major concern for our modern society. In the last decade of the twentieth century, many reported system outages or machine crashes were traced down to computer software failures. Consequently, recent literature is replete with horror stories regarding software problems.

Software failures have impaired several high-visibility programs in the health industry; they have even killed people. As described in [12], the Therac-25 radiation therapy machine was hit by software errors in its sophisticated control systems and claimed several patients' lives in 1985 and 1986. In the UK, South West Thames Regional Health Authority [28] reported an incident on October 26, 1992, when the Computer Aided Dispatch system of the London Ambulance Service broke down immediately after its installation, paralyzing the capability of the world's largest ambulance service to handle the 5000 requests to carry patients in emergency situations received each day. In the aviation industry, although several airliner crashes have remained mysteries, experts pointed out that software control could be the chief suspect in some of these incidents owing to its inappropriate response to the pilots' desperate inquires during an abnormal flight condition.

To this end, software companies recognize the need for systematic approaches to measuring and assuring *software reliability*, and they devote a major share of project development resources to this. The Institute of Electrical and Electronics Engineers (IEEE) [8] defines software reliability as

> the probability of failure-free software operations for a specified period of time in a specified environment.

Software reliability engineering is the field that quantifies the operational behavior of software systems with respect to user requirements concerning reliability. It considers:

1. The definition of various metrics measuring attributes of product design, the development process, system architecture, the operational environment, and the code itself, in as far as the metrics affect the reliability.
2. The design and implementation of operational tests and field operation of the software. When software fails, the code is analyzed to identify the responsible fault, and this is corrected. (This operation may introduce a new fault.) Note the distinction between a *fault* (erroneous code) and a *failure* (the software fails to execute correctly on a specific test). A single fault may be responsible for many failures, or none if the relevant part of the code is never executed.
3. The development of models relating the metrics in point 1 to the test results. These models can be used to estimate the current reliability of the software, and to predict future performance as the software evolves.
4. Application of these technologies in specifying and guiding system architecture, development, testing, acquisition, use, and maintenance.

Reliability is an essential ingredient in customer satisfaction. In fact, ISO 9000-3 [9] specifies measurement of field failures as the only required quality metric:

> . . . at a minimum, some metrics should be used which represent reported field failures and/or defects from the customer's viewpoint. The supplier of software products should collect and act on quantitative measures of the quality of these software products.

Many of the current software reliability engineering techniques and practices are detailed in [15]; another general reference is [21].

In this article we focus on software reliability models and measurements. A software reliability model specifies the general form of the dependence of the failure process on the principal factors that affect it: fault introduction, fault removal, and the operational environment. During the test phase, the failure rate of a software system is generally decreasing owing to the discovery and correction of software faults. With careful record-keeping procedures in place, it is possible to use statistical methods to analyze the historical record with regard to failures and faults. The purposes of these analyses is two-fold: (i) to predict the additional time needed to achieve a specified reliability objective; (ii) to predict the expected reliability and faults when the testing is finished.

Implicit in this discussion is the concept of "time". For some purposes this may be calendar time, assuming that testing proceeds roughly uniformly; another

possibility is to use computer execution time, or some other measure of testing effort. Another implicit assumption is that the software system being tested remains fixed throughout (except for the removal of faults as they are found). This assumption is frequently violated.

Software reliability measurement is based on two types of models, static and dynamic reliability estimation models, used typically in earlier and later stages of development, respectively. These will be discussed in the following two sections.

Early Stage Models

One purpose of reliability models is to perform reliability prediction in an early stage of software development. This activity determines future software reliability based upon available software metrics and measures. Particularly when failure data are not available (e.g. software is in the design or coding stage), the metrics obtained from the software development process and the characteristics of the resulting product can be used to determine reliability of the software upon testing or delivery. We discuss two prediction models: the phase-based model and the Rome Laboratory model.

Phase-Based Model

Gaffney & Davis [6] proposed the phase-based model, which makes use of fault statistics obtained during the early development phases (e.g. requirement review, design, and implementation) to predict the expected fault densities during a later phase (e.g. test or operation). In order to do this, the model makes the following assumptions:

1. Code size estimates are available during the early phases of a development effort. The faults found during the requirements analysis and software design are normalized by these estimates.
2. Faults found in different phases of life cycle follow a Rayleigh density function.

Denoting the fault density (faults per 1000 lines of noncommentary source line, or KNCSL) up to the end of phase t by V_t, the model is expressed as:

$$V_t = E[1 - \exp(-Bt^2)],$$

where E is the total lifetime fault rate expressed in faults KNCSL; t is the fault discovery index ("1" means requirements analysis, "2" means software design, "3" means implementation, "4" means unit test, "5" means software integration, "6" means system test, and "7" means acceptance test); and $B = 1/2t_p^2$, where t_p is the fault discovery phase constant, the peak of the continuous Rayleigh curve fit to the discrete failure data. This is the point at which 39% of faults have been discovered. For example, $t_p = 2.64$ means the peak happens around 2/3 of the way through the design phase and the implementation phase and is closer to the latter.

As data become available B and E can be estimated. These quantities can also be used to estimate the number of remaining faults at stage t by multiplying $E\exp(-Bt^2)$ by the number of source line statements at that point. Note that since the number of data points available to fit a Rayleigh curve is very limited (at most seven of them), the prediction for the number of faults in a future phase could be very rough.

Rome Laboratory Work

The Air Force's Rome Laboratory [22] model obtains predictions of initial fault density upon testing by:

$$d_0 = A \times D \times (SA \times ST \times SQ) \times (SL \times SS \times SM \times SU \times SX \times SR),$$

where d_0 is the initial fault density and the other factors are measures of software characteristics, which can be classified into four categories:

1. *Application type* (e.g. real-time control systems, scientific, information management), denoted by A.
2. *Development environment* (characterized by development methodology and available tools), denoted by D.
3. *Requirements and design representation metrics*, including.
   - SA for anomaly management
   - ST for traceability
   - SQ for incorporation of quality review results into the software.
4. *Software implementation metrics*, including:
   - SL for language type (assembly, high-order, etc.)
   - SS for program size

SM for modularity
SU for extent of reuse
SX for complexity
SR for incorporation of standards review results into the software.

Note that A takes a baseline fault density value while all other factors are modifiers (numerical values close to 1). Rome Laboratory obtained the estimates of these values for each level based on an empirical study over 59 projects.

Once the initial fault density has been found, a prediction of the initial failure rate is made as

$$l_0 = F \times K \times d_0 \times \text{KNCSL}.$$

The number of inherent faults is $d_0 \times$ KNCLS; F is the linear execution frequency of the program, which is the average machine instruction rate divided by the number of object instructions in the program; and K is the fault expose ratio, the expected number of failures per execution per fault. Based on historical data K is between 1.4×10^{-7} and 10.6×10^{-7}.

Testing and Operational Stage Models

Software reliability estimation determines current software reliability by applying statistical inference techniques to failure data obtained during system test or during system operation. Its main purpose is to assess the current reliability. Since reliability tends to improve during the software testing and operation periods, the models are also called *reliability growth models*. Most current software reliability models fall into this category. Details of these models can be found in [15], in which a number of the best current software reliability tools that implement these models are also included. Other surveys appear in [20] and [27].

Using Reliability Models

The success of a model is often judged by how well it fits a curve $\mu(t)$ to the observed "number of faults vs. time" function. On general grounds, this may have little to do with how useful the model is in predicting future faults in the present system (a better fit can mean worse prediction), or future experience with another system, unless we can establish statistical relationships between measurable attributes of the system and estimated parameters of the fitted models (refer to the section on Rome Laboratory work for such an effort).

Different sets of assumptions can lead to equivalent models; for example the assumption that for each fault the time-to-detection is a random variable with a **Pareto distribution**, these random variables being independent, is equivalent to assuming that each fault has an **exponential** lifetime, with these lifetimes being independent, with the rates for the different faults being distributed according to a **gamma distribution** (this is Littlewood's [13] model). A single experience cannot distinguish between a model that assumes a fixed but unknown number of faults and a model that assumes this number is random. Generally, little is known about how well the various models can be distinguished.

Assumptions

Most of the published models are based on similar assumptions. These commonly include the following:

1. The system being tested remains essentially unchanged throughout testing, except for the removal of faults as they are found. Some models allow for the possibility that faults are not corrected perfectly. Miller [16] assumes that if faults are not removed as they are found, then each fault causes failures according to a stationary **Poisson process**; these processes are independent of one another and may have different rates. By specifying the rates, many of the models mentioned below can be obtained.
2. Removing a fault does not affect the chance that a different fault will be found.
3. "Time" is measured in such a way that testing effort is constant. Musa [18] reports that execution time (processor time) is the most successful way to measure time. Others prefer time measured as effort in staff hours [5].
4. The model is Markovian, i.e. at any time the future evolution of the testing process depends only on the present state (the current time, the number of faults found and remaining, and the overall parameters of the model), and not on details of the past history of the testing process. In some models a stronger property holds, namely that the future depends only on the current state and the parameters, and not on the

current time. We call this the "strong Markov" property (*see* **Markov Processes**).
5. All faults are of equal importance (contribute equally to the failure rate).
6. At the start of testing, there is some finite total number of faults, which may be fixed (known or unknown) or random; if random, their distribution may be known or of known form with unknown parameters. Alternatively, the "number of faults" is not assumed finite, so that if testing continues indefinitely, an ever-increasing number of faults will be found.
7. Between failures, the **hazard rate** follows a known functional form; this is often taken to be simply a constant.

Fixed-Shape Models

In binomial models the total number of faults is some number N; the number found by time t has a **binomial distribution** with mean $\mu(t) = NF(t)$, where $F(t)$ is the probability of a particular fault being found by time t. The number of faults found in any interval of time (including the interval (t, ∞)) is also binomial. Letting N be **Poisson** (with some mean ν) gives the related Poisson model; now the number of faults found in any interval is Poisson, and for disjoint intervals these numbers are independent. The hazard rate at time t is $NF'(t)/(1 - F(t))$. These models are Markovian but not strongly Markovian, except when F is exponential; this case was studied in [7, 10, 17, 18, 24], and [25], and for F a **Weibull distribution** in [1] and [23]; In [29] F was made a gamma distribution; and Littlewood's model [13] is equivalent to assuming F to be Pareto. In [19] the hazard rate was assumed to be an inverse linear function of time; for this 'logarithmic Poisson' model the total number of failures is infinite.

Strongly Markov Models

These can be obtained by specifying how the hazard rate of the failure process depends on the current state. Moranda [17] assumed that the hazard rate is constant between failures, and decreases geometrically at each failure. Littlewood & Verrall [14] proposed a class of models in which, after the ith fault is found, the hazard rate becomes $G_i/\xi(i)$, where $\xi(i)$ is some simple function (typically linear or quadratic) and G_i is a random variable (independent for different i) with a gamma distribution. Models of this class were studied in [11].

Reliability Growth Modeling with Covariates

So far we have discussed a number of different kinds of reliability models of varying degrees of plausibility, including phase-based models depending upon a Rayleigh curve, growth models like the Goel–Okumoto model, etc. The growth models take as input either failure time or failure count data, and fit a **stochastic process** model to reflect reliability growth. The differences between the models lie principally in assumptions made on the underlying stochastic process generating the data.

However, most existing models assume that there are no **explanatory variables** available. When the models are used to evaluate a testing process, this assumption is assuredly simplistic for all but small systems involving short development and life cycles. For large systems (e.g. greater than 100 KNCSL) there are variables, other than time, which are very relevant. For example, it is typically assumed that the number of faults (found and unfound) in a system under test remains stable during testing. This implies that the code remains frozen during testing. However, this is rarely the case for large systems since aggressive delivery cycles force the final phases of development to overlap with the initial stages of system test. Thus, the size of code, and consequently the number of faults, in a large system can vary widely during testing. If these changes in code size are not considered as a *covariate*, one is, at best, likely to have an increase in variability and a loss in predictive performance, and, at worst, a poor fitting model with unstable parameter estimates. Dalal & McIntosh [5] describe a general approach for incorporating **covariates**. They also report a case study dealing with reliability modeling during product testing when code is changing.

When to Stop Testing Software?

Dynamic reliability growth models can be used to make decisions related to when to stop testing. Software testing is a necessary but expensive process, consuming one-third to one-half the cost of a typical

development project. Testing a large software system costs thousands of dollars per day. Overzealous testing can lead to a product that is overpriced and late to market, while fixing a fault in a released system is usually an order of magnitude more expensive than fixing the fault in the testing laboratory. The question of how much to test is therefore an important economic question. We discuss an economic formulation of the "when to stop testing" issue as proposed in [2, 3]. Other formulations have also been proposed [4, 26].

Like many other reliability models, Dalal & Mallows' stochastic model assumes that there are N (unknown) faults in the software, and the times to find faults are observable and are independent, identically distributed (iid) exponential with rate m. Their economic model defines the cost of testing at time t to be $ft - cK(t)$, where $K(t)$ is the number of faults observed to time t and f is the cost of operating the testing laboratory per unit time. The constant c is the net cost of fixing a fault after rather than before release. Under somewhat more general assumptions, Dalal & Mallows [2] found the optimal stopping rule for large N. It is very nearly: stop as soon as $f(\mathrm{e}^{mt} - 1)/(mc) \geq K(t)$. Besides the economic guarantee, this rule gives a guarantee on the number of remaining faults, namely that this number has a Poisson distribution with mean $f/(mc)$. Thus, instead of determining the ratio f/c from economic considerations, we can choose it so that there are probabilistic guarantees on the number of remaining faults. Some practitioners may find that this probabilistic guarantee on the number of remaining faults is more relevant in their application; see [4] for a more detailed discussion. Finally, by using reasoning similar to that used in deriving (4.5) of [3], it can be shown that the current estimate of the additional time required for testing, Δt, is given by: $(1/m)\log cmK(t)/[f(\mathrm{e}^{mt} - 1)]$; for applications of this, see [3].

Discussions and Conclusions

Software reliability modeling and measurement have attracted a tremendous amount of attention recently in various industries concerning the quality of software. Many reliability models have been proposed, many success stories reported, several conferences and forums formed, and much project experience shared. Here we offer some caution to users regarding the application of software reliability models.

In fitting any model to a given data set, first one must bear in mind a given model's assumptions. For example, if a model assumes a fixed number of software faults will be removed within a limited period of time, but in the observed process the number of faults is not fixed (e.g. new faults are added owing to imperfect fault removal), then one should use another model which does not make this assumption.

A second model limitation and implementation issue concerns future predictions. If the software is being operated in a manner different from the way it is tested (e.g. new capabilities are being exercised that were not tested before), the failure history of the past will not reflect these changes, and poor predictions may result. Developing operational profiles, as proposed in [20], is very important if one wants to predict accurately future reliability in the user's environment.

Another issue relates to the software development environment. Most models are primarily applicable from testing onward: the software is assumed to have matured to the point that extensive changes are not being made. These models cannot have a credible performance when the software is changing and churn of software code is observed during testing. In this case the techniques described in this article should be used to handle the dynamic testing situation.

Finally, software reliability models cannot make an impact if they are not tied to software testing and operational costs to determine the optimal time to stop testing. We have described a relevant economic model in this entry.

References

[1] Crow, L.H., (1974). Reliability analysis for complex reparable systems, in *Reliability and Biometry*, F. Proshan & R.J. Serfling, eds. SIAM, Philadelphia, pp. 379–410.

[2] Dalal, S.R. & Mallows, C.L. (1988). When should one stop software testing?, *Journal of the American Statistical Association* **83**, 872–879.

[3] Dalal, S.R. & Mallows, C.L. (1990). Some graphical aids for deciding when to stop testing software, *IEEE Journal on Special Areas in Communications* **8**, 169–175. (Special issue on Software Quality & Productivity.)

[4] Dalal, S.R. & Mallows, C.L. (1992). Buying with exact confidence, *Annals of Applied Probability*, **2**, 752–765.

[5] Dalal, S.R. & McIntosh, A.M. (1994). When to stop testing for large software systems with changing code, *IEEE Transactions on Software Engineering* **20**, 318–323.

[6] Gaffney, J.D. & Davis, C.F. (1988). An approach to estimating software errors and availability. SPC-TR-88-007, version 1.0; also in *Proceedings of the Eleventh Minnowbrook Workshop on Software Reliability*.

[7] Goel, A.L. & Okumoto, K. (1979). Time-dependent error-detection rate model for software and other performance measures, *IEEE Transactions on Reliability*, **R-28**, 206–211.

[8] Institute of Electrical and Electronics Engineers (1991). *ANSI/IEEE Standard Glossary of Software Engineering Terminology*, IEEE Std. 729–1991.

[9] ISO (1991). *Quality Management and Quality Assurance Standards – Part 3: Guidelines for the Application of ISO 9001 to the Development, Supply and Maintenance of Software, ISO 9000-3* ISO, Geneva.

[10] Jelinski, Z. & Moranda, P.B. (1972). Software reliability research, in *Statistical Computer Performance Evaluation*. Academic Press, New York, pp. 465–484.

[11] Keiller, P.A., Littlewood, B., Miller, D.R. & Sofer, A. (1983). Comparison of software reliability predictions, in *Proceedings of the Thirteenth IEEE International Symposium on Fault-Tolerant Computing (FTCS-13)*. Milano, Italy, pp. 128–134.

[12] Lee, L. (1992). *The Day the Phones Stopped: How People Get Hurt when Computers Go Wrong*. Donald I. Fine, New York.

[13] Littlewood, B. (1981). Stochastic reliability growth: a model for fault-removal in computer programs and hardware designs, *IEEE Transactions on Reliability*, **R-30**, 313–320.

[14] Littlewood, B. & Verrall, V. (1973). A Bayesian reliability model with a stochastically monotone failure rate, *IEEE Transactions on Reliability*, **R-23**, 108–114.

[15] Lyu, M.R. (1996). *Handbook of Software Reliability Engineering*. McGraw-Hill, New York.

[16] Miller, D. (1986). Exponential order statistic models of software reliability growth, *IEEE Transactions on Software Engineering* **SE-12**, 12–24.

[17] Moranda, P.B. (1975). Predictions of software reliability during debugging, in *Proceedings of the Annual Reliability and Maintainability Symposium*. Washington, pp. 327–332.

[18] Musa, J.D. (1975). A theory of software reliability and its application, *IEEE Transactions on Software Engineering*, **SE-1**, 312–327.

[19] Musa, J.D. & Okumoto, K. (1984). A logarithmic Poisson execution time model for software reliability measurement, in *Proceedings of the Seventh International Conference on Software Engineering*. Orlando, pp. 230–238.

[20] Musa, J.D., Fuoco, G., Irving, N., Kropfl, D. & Juhlin, B. (1996). The operational profile, in *Handbook of Software Reliability Engineering*. McGraw-Hill, New York.

[21] Musa, J.D., Iannino, A. & Okumoto, K. (1987). *Software Reliability – Measurement, Prediction, Application*. McGraw-Hill, New York.

[22] Rome Laboratory (1987). Methodology for Software Reliability Prediction and Assessment, *Technical Report RADC-TR-87-171; Technical Report RL-TR-92-52*, 1992.

[23] Schick, G.J. & Wolverton, R.W. (1973). Assessment of software reliability, in *Proceedings of Operations Research*. Physica-Verlag, Wurzburg-Wien, pp. 395–422.

[24] Schneidewind, N.F. (1975). Analysis of error processes in computer software, *Sigplan Note* **10**, 337–346.

[25] Shooman, M.L. (1972). Probabilistic models for software reliability prediction, in *Statistical Computer Performance Evaluation*. Academic Press, New York, pp. 485–502.

[26] Singpurwalla, N.D. (1991). Determining an optimal time interval for testing and debugging software, *IEEE Transactions of Software Engineering* **17**, 313–319.

[27] Singpurwalla, N.D. & Wilson, S.P. (1994). Software reliability modeling, *International Statistical Review*, **62**, 289–317.

[28] South West Thames Regional Health Authority (1993). *Report of the Inquiry into the London Ambulance Service*.

[29] Yamada, S., Ohba, M. & Osaki, S. (1983). S-shaped reliability growth modeling for software error detection, *IEEE Transactions on Reliability* **R-32**, 475–478.

(*See also* **Algorithm**; **Reliability Study**; **Software, Biostatistical**)

S.R. DALAL, M.R. LYU & C.L. MALLOWS

Software, Biostatistical

Biostatisticians, and applied researchers using statistics, started to use statistical computer packages (by which I mean pre-written and compiled instructions to the computer for performing some form of statistical analysis) for data analysis during the 1950s. Almost immediately changes occurred in what data were analyzed and in how they were analyzed. Changes in computer hardware have brought changes in the type and quantity of software available. The advent of microcomputers in the late 1970s and early 1980s increased the rate of change and the amount of new software packages. There are currently well over 1000 statistical software packages available on a range of computer hardware platforms.

A database of citations to published reviews of statistical software is available [8]. A good review should tell potential users what the package does, how well it does it, how easy the package is to learn and to use, and how flexible the package is. Also available is information on how certain extendable packages make both vendor-written and user-written extensions available to users; information about the cost, and example contributions, have been presented in the "Editor's notes" of the Statistical Computing Section of *The American Statistician* (see, for example, [8]).

Some Historical Notes

The 1950s saw the first occurrence of statistical software, usually specialized single purpose programs that would run on one type of machine only. Some of these were written by users, but hardware vendors were the first important source (for example, SSP from IBM). The appearance of FORTRAN in the late 1950s saw the first real surge of software and the first occurrence, to my knowledge, of generally useful software not written by a hardware vendor; this was "BIMED", later called BMD, then BMDP, which was started at the University of California at Los Angeles about 1960. By the mid-1960s several other packages had appeared, including PSTAT, SPSS, and SAS in the US and Genstat from England and Australia. All of these packages still exist. A number of other packages also appeared during the 1960s (e.g. OSIRIS, Datatext), but most of these, as far as I know, are no longer available.

These packages were neither well integrated nor comprehensive in coverage by the standards of today. They often used unacceptable algorithms or were prone to coding mistakes which gave wrong, or inaccurate, answers. (For example, Longley [17], using a multicollinear data set, showed problems in a number of packages.) However, prior to the availability of packages such as these, days could be spent, using a mechanical calculator or pencil and paper, to estimate, say, one regression with two covariates on a relatively small data set.

The late 1960s and early 1970s not only saw the appearance of additional software packages, some highly specialized (e.g. just for sample-size calculations) rather than general purpose, but also witnessed the setting up of committees by statistical associations to work on evaluating and designing software: GLIM originated under the auspices of a Royal Statistical Society (RSS) Committee; the American Statistical Association (ASA) set up a Committee on Statistical Program Packages in 1973 to help in evaluating software [3, 4]. The RSS committee, now called the "GLIM Working Party" still exists, as does the software, and the RSS receives a royalty on each version of GLIM sold. The ASA, however, no longer has any such committee.

A Categorization Scheme for Biostatistical Software

The range of software currently available makes any categorization scheme somewhat problematic. The categorization presented here is limited to one dimension: the type of user to whom the vendor expects to sell (and is further limited to software aimed at professionals); a broader categorization scheme can be found in [9]. It is impossible to include all existing packages. I primarily included packages well known to me; within each category the packages are listed alphabetically. Owing to space limitations, only contact information and a brief overview of the package are given. Contact information is for the headquarters of the company; many companies have sales and support offices in other countries.

General Purpose, Useful for Biostatistics

1. Integrated packages, including
   (a) BMDP, purchase via Statistical Solutions, 8 South bank, Crosse's green, Cork, Ireland;

+353 21 4319629; SPSS Inc., 444 N. Michigan Ave., Chicago, IL 60611, USA; (312) 329-4000; its original design was aimed squarely at biostatistical goals; it is available for several computer platforms (DOS, UNIX, mainframes).

(b) Data Desk, Data Description, Inc., 840 Hanshaw road, 2nd floor, Ithaca, NY 14850; it is available on both Macintosh and Windows platforms.

(c) Genstat, Numerical Algorithms Group, Ltd, Wilkinson House, Jordan Hill Road, Oxford OX2 8DR, UK; (+44) 1865-511245; runs under Windows and several workstation operating systems, including UNIX, VMS and SunOS.

(d) GLIM, Numerical Algorithms Group, Ltd, Wilkinson House, Jordan Hill Road, Oxford OX2 8DR, UK; (+44) 1865-511245; runs under DOS and several workstation operating systems (e.g. UNIX, VMS, SunOS).

(e) JMP, SAS Institute, Inc., SAS Campus Drive, Cary, NC 27513, USA; (919) 677-8000; it is available for both the Macintosh and Windows platforms.

(f) Minitab, Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801, USA; (814) 238-3280; has been widely used in educational environments; it is available for several computer platforms (Macintosh, Windows, UNIX and mainframes).

(g) NCSS, 329 North 1000 East, Kaysville, UT 84037, USA; (801) 546-0445; runs under Windows.

(h) SAS, SAS Institute, Inc., SAS Campus Drive, Cary, NC 27513, USA; (919) 677-8000; runs under several platforms (Windows, UNIX, mainframes).

(i) SPSS, SPSS Inc., 444 N. Michigan Ave., Chicago, IL 60611, USA; (312) 329-4000; originally designed for use by social scientists; runs under several platforms (Macintosh, Windows, UNIX, mainframes).

(j) Stata, Stata Corp., 702 University Drive East, College Station, TX 77840, USA; (800) 782-8272; runs under several platforms (Macintosh, Windows, UNIX).

(k) Statistica, Statsoft, Inc., 2325 East 13th St., Tulsa, OK 74104, USA; (918) 749-1119; runs under Macintosh and Windows operating systems.

(l) Systat, Systat Software. Inc., 501 Suite "C", Point Richmond Tech Center, Canal Blvd., Richmond, CA 94804; SPSS Inc., 444 N. Michigan Ave., Chicago, IL 60611, USA; (312) 329-4000; runs under Macintosh, Windows and UNIX operating systems.

2. Packages based on programming languages; many of these, as well as at least some of the extensible packages mentioned elsewhere, can use subroutine libraries (*see* **Numerical Analysis**):

(a) Gauss, Aptech Systems, Inc., 23804 SE Kent-Kangley Road, Maple Valley, WA 98038, USA, (425) 432-7855; runs under Windows and UNIX; contains numerous statistical routines; there are also several "packages" (sets of Gauss routines) written in Gauss and relevant to biostatistical users.

(b) Matlab, The Mathworks, Inc., 3 Apple Hill Drive, Natick, MA 01760, USA; (508) 647-7000; runs under Windows, UNIX; although most early routines were aimed at engineers, there are now a sizable number of statistical routines.

(c) R, a public domain near-clone of S-Plus; this can be found on Statlib (www address: `http://lib.stat.cmu.edu/`).

(d) SC, Mole Software, 34 Greenville Road, Bloomfield, Belfast BT5 5EP, N. Ireland; (+44) (0) 1232 282654; runs under DOS.

(e) Insightful Corporation, StatSci Division of MathSoft, 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA, (800) 569-0123; based on the AT & T product "S"; S-Plus runs under Windows and UNIX; many new forms of analysis first appear as S (or S-Plus) programs (*see* **S-PLUS and S**).

(f) XLISP-STAT, available for free by anonymous ftp from umnstat.stat.umn.edu; there are versions for the Macintosh, Unix, and Microsoft Windows; there are at least four research groups that have written packages based on XLISP-STAT; an introduction to this package can be found in [22, 23]; introductions to three of the packages can be found in [21], [25], and [26].

Aimed Specifically at Biostatistical Users

1. General purpose:
   (a) EAST, CyTel Software Corp., 675 Massachusetts Avenue, Cambridge, MA 02139, USA: (617) 661-2011; for design of sequential trials; runs under DOS.
   (b) Epicure, HiroSoft International Corp., 1463 E. Republican Ave., Suite 103, Seattle, WA 98112, USA, (206) 328-5301; runs under DOS and UNIX.
   (c) EpiInfo, originated at the US Centers for Disease Control (CDC) and since then the result of collaboration between the CDC and the World Health Organization; it is available for free on the Internet (`http://www.cdc.gov/epiinfo`); can also be purchased with a printed manual of over 500 pages, from USD, Inc., 2075-A West Park Place, Stone Mountain, GA 30087, USA, (770) 469-4098; runs under DOS.
   (d) Epilog Plus, Epicenter Software, P.O. Box 90073, Pasadena, CA 91109, USA; (626) 304-9487; runs under Windows.
   (e) True Epistat, Epistat Services, 2011 Cap Rock Circle, Richardson, TX 75080, USA; (214) 680-1376; runs under DOS.
2. Special purpose:
   (a) EAST, CyTel Software Corp., 675 Massachusetts Avenue, Cambridge, MA 02139, USA; (617) 661-2011; for design of sequential trials; runs under DOS.
   (b) PEST, The MPS Research Unit, The University of Reading, Earley Gate, Reading RG6 6FN, UK; for design and analysis of sequential trials; runs under DOS.

Special Purpose Software that is Often Relevant to Biostatisticians

1. Randomization software:
   (a) RT, B.F.J. Manly, The Centre for Applications of Statistics and Mathematics, University of Otago, PO Box 56, Dunedin, New Zealand; 64-3-479-7774; randomization procedures for a number of parametric procedures, including anova, linear regression, spatial data, time series; runs on DOS.
   (b) StatXact, LogXact, CyTel Software Corp., 675 Massachusetts Avenue, Cambridge, MA 02139, USA; (617) 661-2011; StatXact includes randomization versions of a large number of nonparametric analyses; LogXact performs exact logistic regression; each runs under Windows.
   (c) Testimate, idv-Datenanalyse und Versuchsplanung, Wessobrunner Strasse 6, D-82131 Gauting/München, Germany; 089/8 50 80 01; includes randomization versions of a large number of nonparametric analyses;
2. Software for estimating sample sizes when designing studies. The following web site has information on more than two dozen such software packages: `http://www.interchg.ubc.ca/cacb/ power`. The following have specific biostatistical orientations:
   (a) EAST, CyTel Software Corp., 675 Massachusetts Avenue, Cambridge, MA 02139, USA: (617) 661-2011; for design of sequential trials; runs under DOS.
   (b) N and NSURV, idv-Datenanalyse und Versuchsplanung, Wessobrunner Strasse 6, D-82131 Gauting/München, Germany; 089/8 50 80 01; runs under DOS.
   (c) PASS, 329 North 1000 East, Kaysville, UT 84037, USA; (801) 546-0445; runs under Windows.
3. Software for correlated data, including longitudinal studies:
   (a) standard software: several of the packages included elsewhere in this list, including BMDP, LIMDEP SAS, S-Plus and Stata, include special routines for this type of analysis.
   (b) software for analyzing surveys; only one package above has adequate routines for dealing with weighted survey data: Stata; there are specialized packages, also:
      (i) SUDAAN, Research Triangle Institute, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709, USA; (919) 541-6602; runs under Windows, UNIX and mainframes;
      (ii) WESVAR, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850; (800) westat2, extension 2006.
   (c) software for hierarchical models:
      (i) HLM, Scientific Software International, 7383 N Lincoln Ave., Suite

100, Lincolnwood, IL 60712, (800) 247-6113; runs under Windows or DOS;

(ii) MLWin, Centre for Multilevel Modeling Project, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK; +44(0)207 612 6688; runs under Windows.

4. Software from other disciplines: econometric software such as LIMDEP; Econometric Software, Inc., 15 Gloria Place, Plainview, NY 11803, USA; (516) 938-5254; runs under DOS. Many of its routines are of the same type as biostatisticians use and it has some unique features, e.g. the survival analysis routines include left-truncated data and "cure" models.
5. Software for a specific form of analysis:
   (a) Survival; see [10] and [12].
   (b) Spatial; some of the above packages, especially Epilog Plus, Genstat, RT and S-Plus have some routines; there are some very specialized packages but they tend to be oriented to geostatistics and use very different jargon.
   (c) Circular; Oriana, Kovach Computing Services, 85 Nant-y-Felin, Pentraeth, Isle of Anglesey LL75 8UY, Wales, UK; (+44) (0) 1248-450414; specifically oriented to analysis of data in degrees (e.g. angular data such as might be used in a study of spinal injuries) or time (used in health services research); runs under Windows; the only other software I know of are some user-written routines in Stata.
6. Bayesian software; while there are a number of Bayesian software packages, most have never been reviewed anywhere; overviews appear in [5], [6], and [20]. Newer packages have started to appear, including
   (a) BUGS; World Wide Web address: `http://www.mrcbsu.cam.ac.uk/bugs`; versions for Windows and UNIX; "carries out Bayesian inference on complex statistical problems for which there is no exact analytic solution".
   (b) B/D: World Wide Web address: `http://fourier.dur.ac.uk:8000/stats/bd`; runs under Windows; "an interactive programming language which allows complete a priori and diagnostic analyses of Bayesian linear statistical problems".

Some Assessment Criteria

The following issues are of particular importance in assessing any statistical software package, regardless of whether it is specifically aimed at biostatistical users: the quality of the manual, the ease of learning and the ease of use of the package, and the accuracy of its computations.

Although some vendors would have purchasers believe that their package is usable without a manual, there are reasons for users to examine the manual carefully. Information in the manual should include:

1. Information on what is available (though each user must decide whether what is available is what is wanted, and, more importantly, whether it works in the way wanted and whether all the options desired are present).
2. At least one index; if it there is at least one, how good is it?
3. Examples of using the software; are the examples complete? That is to say, do the examples only display how the new commands (menu choices, etc.) work or is everything shown that one would actually need to complete an analysis?
4. Information on other sources of help, including courses, books, web sites, etc.
5. Information on how to interface this package with the operating system and/or with other types of software packages (such as word processing software).
6. Technical information relating to the algorithm used and how the vendor tested the software; there should be citations to the professional literature as well. Note that, as yet, very few vendors actually provide this (for a discussion in the context of a comparative review, see [2]).
7. A list, and explanation, of error messages; these should be clear to someone who does not have a PhD in computer science and should also be given at the same level as the statistical text.

Manuals can also be used to discover whether the package appears to be aimed at the right type of user; for this, you should examine the manual(s) with the following in mind:

1. What type of language is used in describing and explaining the package? Jargon is rampant and differs dramatically across different disciplines.
2. What level of statistical language is used (e.g. beginner or professional)?
3. What types of graphical output are available and how integrated are the graphics and the statistical routines?
4. What types of checks and diagnostic information are available to help decide whether there are problems with the results of an estimation procedure?
5. How flexible is the software with respect to:

(a) nonstandard problems, e.g. are there choices of algorithms for standard routines such as linear regression?
(b) output; can the user affect the output of the package to ensure that it is in the most usable format for that particular use?

For a discussion of some criteria useful in assessing manuals, see [1], [18], the accompanying discussion of these articles, and the rejoinders by the authors.

A criterion often mentioned is ease of learning of the package. My experience, however, has been that this is really only important for people who will be infrequent users of the software, as these people will essentially be learning the package over again each time they use it. However, for others, the cost of learning is easily overshadowed by ease of use considerations, especially since, for even the hardest-to-learn packages, it rarely takes more than a few hours to learn at least enough to obtain some output.

Ease of use is sometimes, mistakenly, listed with ease of learning as a criterion. It is however both different and much more important. It is also, generally, harder to assess since the determination of whether something is easy to use is heavily dependent on both the level of the user and what the user is trying to do, as well as on the structure of the program. Program structure affects ease of use in many ways; a simple example relates to the difference between typing a command and clicking on a menu item. How this affects a given user depends on whether the menu defaults are what is primarily wanted and how easy it is to choose different options. Of course, at the other extreme, some users want so much of what they choose to do to be dependent on the situation, that no menu-driven program could possibly be considered "easy to use". Furthermore, there are many issues that vendors have never considered and these cannot, of course, be present in a menu. Whether they are available in a command system depends on the amount of thought the vendor put into making the package flexible (a detailed example is provided in [9]). Ease of use can also be aided by the availability of books about the packages, user groups, including e-mail lists and Usenet news groups, vendor newsletters, etc. Integration and ease of recall of various parts of the numerical and graphical output, and integration of the packages to the operating system and to other software (e.g. word processors), are also important here.

Earlier, I mentioned the issue of whether the language used in the manual was appropriate to the statistical expertise of the user. A related issue has to do with the ease with which one can assess one's analytic output. This is affected by numerous factors, including the quality of the error messages, the presence of statistics that can be used to assess assumptions underlying the technique used, and the quality and integration with the statistics of the graphics.

The final criterion to be discussed here is the quality of the numerical algorithms, which affects not only the accuracy of the result, but whether the package provides an answer, and, if it does, the efficiency with which it arrives at the answer.

1. Try the examples in the manual (the vendor should supply all example data sets on the disk with the program). If the examples cannot be reproduced, then immediately contact the vendor. While this appears to be a very simple test that no vendor should ever fail, some packages do fail this test.
2. Check reviews, especially those by reputable statisticians (e.g. reviews in *The American Statistician* or in *Applied Statistics*) (unfortunately, this latter journal is dropping its review section). A good review will supply much more information than just that related to accuracy; in particular, information should be included on the level of user targeted and on the ease of learning and using the package. Furthermore, I believe that

comparative reviews are much more useful than reviews of individual packages.

3. Look in the literature for test data sets. Many "tests" are so well known that no vendor fails them anymore (this is true, for example, of the Longley [17] data). Furthermore, some tests are not relevant to the work that any particular user does. However, there are valuable benchmarks and tests in the literature (see for example, [7], [14], [15], [19], and [24]) that will help users and vendors test (a) whether the algorithms are appropriate, (b) what happens at the boundaries of either allowed data or standard language, and (c) the quality of the algorithms being used.
4. Build your own library of test data sets that are important in your own work and for which problems have previously been found. Try this library on every new package, and every upgrade received.
5. Examine the "validation" or "certification" section of the documentation, if it exists; unfortunately, most vendors do not yet provide such a section. If such a section exists, look for information regarding the algorithms used and the range and type of issues and of data used to test the software. The documentation should also discuss carefully the issue of how the vendor decided that the test result was acceptable. Also, note whether the vendor says that all tests are re-run after making any change to the software; this "regression testing" is necessary since fixing a bug in software often introduces one or more new bugs and this possibility must be checked.
6. Finally, run the analysis in at least one other software package and carefully compare the results. For this final check, the importance of having algorithm information in the manual is highlighted because for many analyses different algorithms should produce different results. This is especially true in many nonparametric analyses where the treatment of ties greatly affects the results.

Where to Go for More Information

There are no good general sources of information on what software is available. Eventually, there will probably be a source on the Internet which can be added to frequently. The number of software packages, particularly for educational uses, is growing rapidly. Some information, of course, is already available: numerous journals print reviews and a database of citations to these reviews is available [8]; numerous data sets useful for testing exist at statlib (at Carnegie Mellon University; WWW address: `http://lib.stat.cmu.edu/`) and other places on the internet. A "Statistical Software Guide" is produced approximately every two-three years. The most recent appearance of this guide, in print, was [16], but information is currently being gathered for an update report. However, none of this information is either well organized or complete. There is one commercial source of information, SciTech International, which publishes *Software for Science*; however, even their list (almost 2000 products, but including non-analysis packages such as word-processing software) is incomplete; they are especially weak, obviously, regarding shareware and freeware, which is often specialized and is often available on the Internet. Many of the vendors mentioned above have Internet World Wide Web sites; the best source for finding these in general is via a competitor: Stata, at its site (`http://www.stata.com`) maintains links to the sites of other vendors, including several suppliers of free software.

The Future – Maybe

Statistical software has been changing rapidly in recent years. The main changes, as of 1996, relate to (a) the existence of numerous specialized software packages; (b) the movement, slowly, of these specialized routines into general purpose packages; and (c) a heavy emphasis on graphical analysis, especially new types of graphics and new ways of integrating graphics into standard analysis.

While these are valuable, necessary, and will continue, there are two other changes that would be very valuable to the profession. The first relates to a better integration of what we already know about statistical assumptions with our analysis. For example, we know that the two-sample t test is somewhat affected by different variances (the amount depending on the ratio of the groups' sample sizes); many would find it helpful if, along with requested result, the software gave some information about the validity of this, and other, assumptions, for the data used. This might also help guard against the "misuse" of statistical software

by those who are not well trained in statistics. This issue has been discussed in numerous articles dating back to at least the 1970s [11]. Though noting a number of potential problems, Goodnight clearly favored this type of "offensive validity checking". Haux [13] provides a number of citations on this issue and then gives a detailed example for the Mann–Whitney test and notes that BMDP, SAS and SPSS are each unsatisfactory. Note that the software should not stop the user from doing an analysis. Rather, a user should just want be provided with some additional information without making a number of other requests to the software (e.g. a separate request for equal variances, for symmetry, for heterogeneity, etc.).

A large part of any project relates to data management and data manipulation. Much, and in many projects all, of this is done with the same statistical software used for analysis. However, no current program keeps a reversible history of what the user does to the data and many do not even keep any history (or log or journal) of what was done. The unfortunate result is that often even the analyst cannot reproduce certain results. Thus, another desirable change is the implementation of some form of reversible history of data management and data manipulation so that any particular state of the data could be recreated. Some type of coding scheme should be attached to both this history and to each analysis so that for any given output it would be clear which state of the data was used in its production. The current reliance on *ad hoc*, individual, schemes is inefficient, ineffective, and unnecessary. Version control software, as used in software development, database management and even some word-processing software should be generalizable to statistical software.

References

[1] Berk, K.N. & Francis, I.S. (1978). A review of the manuals for BMDP and SPSS (with comments and rejoinders), *Journal of the American Statistical Association* **73**, 65–70.

[2] Boomsma, A. & Molenaar, I.W. (1993). Four electronic tables for probability distributions, *American Statistician* **48**, 153–162.

[3] Francis, I. & Heiberger, R.M. (1975). The evaluation of statistical program packages – the beginning, in *Proceedings of the Computer Science and Statistics Eighth Annual Symposium on the Interface*, J.W. Frane, ed. Los Angeles, pp. 106–109.

[4] Francis, I., Heiberger, R.M. & Velleman, P.F. (1975). Criteria and considerations in the evaluation of statistical program packages, *American Statistician* **29**, 52–56.

[5] Goel, P.K. (1988). Software for Bayesian analysis: current status and additional needs, in *Bayesian Statistics*, Vol. 3, J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds. Oxford University Press, Oxford.

[6] Goel, P.K. (1988). Software for Bayesian analysis: current status and additional needs – II, in *Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface*, E.J. Wegman, D.T. Gantz & J.J. Miller, eds. American Statistical Association, Alexandria.

[7] Goldstein, R. (1987). Linear regression on IBM PC/XT/AT's, in *Computer Science and Statistics: Proceedings of the 19th Symposium on the Interface* R.M. Heiberger & M.T. Martin, eds. American Statistical Association, Alexandria, pp. 219–228.

[8] Goldstein, R. (1994). Editor's notes, *American Statistician* **48**, 254–255.

[9] Goldstein, R. (1997). Computer packages, *Encyclopedia of Statistical Sciences, Update Volume*. Wiley, New York, to appear.

[10] Goldstein, R. Anderson, J., Ash, A., Craig, B., Harrington, D. & Pagano, M. (1989). Survival analysis software on MS/PC-DOS computers, *Journal of Applied Econometrics* **4**, 393–414.

[11] Goodnight, J.H. (1975). Validity checking: how far should we go? in *Proceedings of the Computer Science and Statistics Eighth Annual Symposium on the Interface*, J.W. Frane, ed. Los Angeles, pp. 146–148.

[12] Harrell, F.E., Jr & Goldstein, R. (1997). A survey of microcomputer survival analysis software: the need for an integrated framework, *American Statistician* **51**, 360–373.

[13] Haux, R. (1983). How to detect and prevent errors in computer-supported statistical analysis: an example, *Methods of Information in Medicine* **22**, 87–92.

[14] Heiberger, R.M., Velleman, P.F. & Ypelaar, M.A. (1983). Generating test data with independently controllable features for multivariate general linear forms, *Journal of the American Statistical Association* **78**, 585–595.

[15] Knuth, D.E. (1981). *The Art of Computing*, Vol. 2: *Seminumerical Algorithms*, 2nd Ed. Addison-Wesley, Reading.

[16] Koch, A. & Haag, U. (1995). The statistical software guide '94/95, in the *Statistical Software Newsletter*, *Computational Statistics & Data Analysis* **19**, 237–261.

[17] Longley, J.W. (1967). An appraisal of least-squares programs for the electronic computer from the point of view of the user, *Journal of the American Statistical Association* **62**, 819–841.

[18] Muller, M.E. (1978). A review of the manuals for BMDP and SPSS (with comments and rejoinders), *Journal of the American Statistical Association* **73**, 71–80.

[19] Nash, J.C. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimization*, 2nd Ed. Adam Hilger, Bristol.

[20] Press, S.J. (1989). *Bayesian Statistics*. Wiley, New York.
[21] Stine, R. (1997). AXIS: an extensible graphical user interface, in *Statistical Computing Environments for Social Research*, R. Stine & J. Fox, eds. Sage, Thousand Oaks, pp. 175–192.
[22] Tierney, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York.
[23] Tierney, L. (1997). Data analysis using LISP-STAT, in *Statistical Computing Environments for Social Research*, R. Stine & J. Fox, eds. Sage, Thousand Oaks, pp. 66–88.
[24] Velleman, P.F. & Ypelaar, M.A. (1980). Constructing regressions with controlled features: a method of probing regression performance, *Journal of the American Statistical Association* **75**, 839–844.
[25] Weisberg, S. (1997). The R-code: a graphical paradigm for regression analysis, in *Statistical Computing Environments for Social Research*, R. Stine & J. Fox, eds. Sage, Thousand Oaks, pp. 193–206.
[26] Young, F.W. & Bann, C.M. (1997). ViSta: a visual statistics system, in *Statistical Computing Environments for Social Research*, R. Stine & J. Fox, eds. Sage, Thousand Oaks, pp. 207–235.

(*See also* **Software Reliability**)

RICHARD GOLDSTEIN

Software, Epidemiological

Historically, many specialized intellectual domains have had data analysis software specifically designed for that domain, including such areas as economics, geology, chemistry and epidemiology. Other domains, including sociology, astronomy and political science, have seen little of this specialization. Furthermore, in those areas, such as epidemiology, that have had numerous specialized packages, some of the packages have fallen by the wayside while others have grown and still others have appeared. In some areas, such as economics, the major specialized packages, for example, `Limdep`, have become more like standard general-purpose packages. However, this has generally not happened with the epidemiologic packages. Finally, at least one general-purpose package, `Stata`, has incorporated a number of epidemiologic analytic routines, jargon and all.

In this article I review some of the specialized epidemiologic software, whether related to the design of studies or to their analysis. I also discuss some of the advantages and disadvantages of having specialized software for a discipline. Both issues are clarified by the presence of a general-purpose analytic package that contains a set of epidemiologic analysis routines. Other general-purpose packages will generally be ignored (*see* **Software, Biostatistical**).

Why Epidemiologic Software?

To write about epidemiologic software, one must locate it, one must define it (or at least its boundaries), and one must have at least some idea regarding what distinguishes epidemiologic software from other data analysis software. While I have undoubtedly missed some small epidemiologic software packages, I hope that I have included all major packages and at least a representative selection of the smaller packages. The following, in alphabetic order, is a list of the packages I will be emphasizing in this article:

- `Cluster`, version 3.1 (free from the Centers for Disease Control (CDC)): this DOS-based package has 12 different techniques to help in disease clustering. It is fairly straightforward to use, but is certainly not elegant.
- `Egret` for Windows, version 1 (commercial; other packages from the same company include `StatXact` and `LogXact`): this package is primarily for fairly specialized modeling of epidemiologic and biomedical data. It includes **additive** and **multiplicative** versions of several models (e.g. logistic and Poisson) as well as random-effects logistic regression. It is very easy to use, much easier than when it was a DOS package. This is its first appearance as a Windows package. The same vendor sells `Egret SIZ`, a DOS package for determining sample size or power for nonlinear regressions, `StatXact`, a package for the exact analysis of tabular data, and including a power module for data to be analyzed via tables, and `LogXact`, a package for exact **logistic** or **Poisson regression**. `LogXact` also contains a Monte Carlo option for data sets for which maximum-likelihood estimates do not converge, but that are too large for exact analysis. `Egret SIZ` is the only package of these that is not easy to use. Furthermore, it requires the user to collapse continuous predictor variables into categorical variables, which can be difficult to do without biasing one's result (*see* **Bias**). The material in `Egret SIZ` is based on Self & Mauritsen [11].
- `Epicalc`, version 1.02 (free): this simple Windows package is just for analyzing epidemiologic tables. It is very easy to use when you already have summary data.
- `Epicure`, version 2.10 (commercial): this package is primarily for specialized modeling of epidemiologic and biomedical data. A wide range of models are included. This package is DOS-based and is fairly easy to use, but not as easy as, say, `Egret`.
- `EpiInfo`, version 6.04, and `EpiMap`, version 2 (free from the CDC): `EpiInfo` has extensive capabilities regarding questionnaires, data entry and data checking, but modest analytic capacity. Taking full advantage requires some programming skill. `EpiMap` has extensive mapping abilities and uses `EpiInfo` data directly. If one only wants a little from these packages, they are easy to use. However, using their full power requires some work by the user.
- `Epilog` Windows, version 1 (commercial): this package is easier to use as a Windows package than it was as a DOS package. It is very modular, which I am not a fan of, but it is also the most complete of the specialized packages.

- `EpiMeta`, version 1 (free from the CDC): relatively easy to use for its quite limited purpose of simple meta-analyses.
- `PEPI`, version 3.01 (free): this DOS package is slowly being made into a Windows package. It is comprised of a large number of separate modules, which can make it a pain to use. This pain is somewhat ameliorated in DOS by an integrated menu that accesses all modules and by a Windows help file that tells the user which module to use. The preliminary parts of the new Windows version go some way to solving the problems caused by separate modules. `PEPI` is primarily for use on summary data and has no data management facilities of its own.
- `Stata`, version 6 (commercial; general purpose): this large, general-purpose, Windows/Macintosh/UNIX package is the most complete of all considered here. It is included because it has specialized routines for epidemiologists. It is very easy to use, even at a fairly advanced level. If the user requires, the program is extensible and there is a wide-ranging user community that is actively extending this program. The Windows version was used in writing this article.
- `True Epistat`, version 5.3 (commercial): this is a DOS package and is very modular. I find the data management facilities awkward to use, though this version is an improvement over earlier versions.
- `Win Episcope`, version 2.0 and `WinEpi Ratios`, version 1.0 (free): these simple packages are easy to use for tables when you already have summary data.

Note that several major, general-purpose packages are not included here as they have no specialized epidemiologic routines, including the Biomedical Data Processing Program (BMDP), `SAS`, `SPSS`, `S-Plus` and `Systat`. The last four, at least, are under active development and are used by many epidemiologists. Each of these has many of the elements discussed below and/or listed in one of the two tables. However, because they do not use epidemiologic jargon, I have not included them in this article (*see* **Software, Biostatistical**).

As part of my preparation, I invited the producers of each of the above packages (except for `Cluster` and `EpiMeta`) to tell me why they thought there should, or should not, be specialized epidemiological software. I received responses from all except the `WinEpi` series (note: the `WinEpi` series is aimed at veterinary epidemiology). As one would expect, the providers of specialized epidemiologic software provided reasons for having such specialized software, while the people from `Stata`, a general-purpose package with some specialized epidemiologic routines included, provided reasons why one would not want specialized software. I received a total of seven responses from different software providers; since there was a lot of overlap in their responses, I just summarize the major issues raised:

1. *The output produced is the type of output used by epidemiologists*. Three vendors said this, with one adding that it was desirable to exclude the "statistical clutter" that general-purpose packages added even when they also produced the type of output desired by epidemiologists. This issue seems best exemplified by tables where many epidemiologists expect specific output (e.g. **odds ratio** (OR) and a confidence interval for this ratio). An example of the added clutter produced by some packages might be, I suppose, statistics such as lambda or Cramer's V. Note, however, some problems with this.
   (a) There are cases where one should want the additional statistics; that is, what one person might call "statistical clutter" might be desirable to other people or even to that person if the person learned about that statistic. (Lambda is an example of a proportional reduction in the error statistic; if one uses it in models, as epidemiologists do, might not one also want to use it for tables?)
   (b) One trend that has clearly increased over the last 20 years is for analysts in one area to read, and contribute to, the literature in other areas. If every discipline used its own specialized software, this would be a harder task. Certain choices by software providers already make this harder; for example, `Stata`, in its epidemiologic tables routine (`epitab`), includes McNemar's test for matched case–control data. Their output tables even use standard epidemiologic language (e.g. the rows and columns are labeled exposed and unexposed). One problem arises here because other disciplines, including various social sciences,

call **case–control studies** by other names and do not use the words "exposed" and "unexposed". A related issue, which ties this to the first point above, is that Quinn McNemar was not an epidemiologist – he was a social psychologist, and when he designed this test he was not using data that were related to health in any way. Rather, he was interested in those who changed preferences regarding presidential candidates according to consecutive public opinion polls (McNemar [6]). When general-purpose packages bend a statistic like this, they endanger their general use (note: `Stata` has another, not specifically epidemiologic, procedure that also produces McNemar's test, this time without the epidemiologic jargon). When specialized users refuse to acknowledge other uses, and other users, they risk becoming dead ends.

(c) While I agree that cluttered output is undesirable, I see no problem with software producers giving users the option to choose which output is to be shown. At the extreme, asking for a table would show only the table with no inference procedures shown unless specifically requested, and each type of inferential procedure desired would have to be specifically requested. Software that includes macros could include example macros showing, for example, the output that an epidemiologist might want, while a second macro might show the output that a psychologist would want.

2. *Ease of use, particularly via epidemiologic jargon.* At least four software providers mentioned this. Several providers specifically mentioned the desirability of this for users who were not full-time epidemiologists. I certainly agree that extreme ease of use, accompanied by standard epidemiologic jargon (to the extent it exists) makes misuse of the software less likely by these people and is therefore a good thing. Inclusion of this jargon in general-purpose statistical software is, by the same token, dangerous (unless one feels that the only possible misusers are the epidemiologists, something I strongly doubt). Related to this was the idea that single-purpose packages (i.e. software that only does one type of analysis, or that is useful with respect to a certain design issue) are generally easier to use than are general-purpose packages. Similarly, certain support issues are clearly part of this, including manuals that give fully worked out epidemiologic examples, textbooks that use a particular software package, and the presence of Usenet news groups and/or e-mail list-servers related to the package. Only the manual examples seem to argue against general-purpose packages (since all the others can, and do, easily coexist with general-purpose packages) and even this can be handled, though some ways of attempting this may work better than others.
3. *Hard-to-find statistical procedures and tests*. This was mentioned by at least five providers. An example is disease **clustering**. There is no question that in epidemiology, as in other disciplines, specialized software packages that include useful analytic routines not available in most general-purpose packages are highly desirable. Some of these might be added to an extensible general-purpose package by a user, or by the provider, but some are so unusual, and difficult to program, that a specialized package might be the only realistic alternative.
4. *Epidemiology has unique demands for data entry and for data manipulation*. One provider argued this. The examples provided (double entry and verification, changing the ordering of categories in a table) are not convincing since users from many disciplines need all of these and more.

My conclusion is that there is a place, and a need, for specialized epidemiologic software, primarily for routines that are not included, or only rarely included, in general-purpose statistical software packages. However, many of the other reasons for specialized software packages are unconvincing at best. Worse, specialized packages that offer only "simple" routines are dangerous. For example, use of tables without the ability to generalize to a model allows too much chance of misleading the researcher. Many epidemiologic tables provide exactly the same result that one would receive from a simple **logistic regression** or **Poisson regression**. The general-purpose package that provides both tables and models allows one to test assumptions and to use covariates – the specialized tool that only has the tables, no matter how perfectly epidemiologic they are, is dangerous.

Software for Designing and Implementing Studies

Historically, the primary emphasis in software has been on analyzing data. Specialized packages, outside of epidemiology, exist for sample size determination and for the assignment of subjects to groups. Some of the epidemiologic software discussed here has the capability of helping a researcher to design or implement a study and those capabilities are discussed in this section. The capabilities of nonepidemiologic specialized software and of other general-purpose software are not discussed here. Software for determination of power, or sample size, have been reviewed elsewhere (e.g. Goldstein [3], Thomas & Krebs [12]) (*see* **Software, Biostatistical**). Software for subject assignment, including matching, has unfortunately not been generally reviewed, to my knowledge.

Several of the packages included here have routines for sample size determination. For the most part, only fairly simple situations are covered. Many statisticians will use simulation for complicated studies, but most of the specialized epidemiologic software has no simulation ability. Table 1 shows what is available for those packages that have any capability for either sample size determination or for simulation. In addition, some of the developers of `Epicure` have produced a separate, and free, power program that is unique in that it can easily be used for tests of interactions. This program, "`Power`", is currently included as a module in the package "`Epitome`" and is being turned into a stand-alone Windows-based program. Note that there is little in any of the packages with respect to calculating power and/or determining sample size for any type of regression model.

None of the specialized packages includes any routines, or simulation capability, to allow for the determination of sample size for research studies involving complex samples. Furthermore, only `Egret SIZ`, a separate product requiring that it be separately purchased, has extensive regression capabilities and the requirement for categorizing any continuous predictor reduces the usefulness of this package. On the other hand, several of the specialized packages include the ability to calculate the needed sample size for simple **case–control** or for simple **cohort studies**, both of which are important in epidemiology. On the face of it, the above table is depressing – specialized packages should be, at least through simulation, capable of much more than they are. The simple studies that are currently covered are, in my opinion, mostly useful for beginners and students.

What other aspects of study design, particularly of the design of epidemiologic studies, are available? There are at least three other areas in which software can help: the design of data collection instruments (e.g. questionnaires); the assignment or allocation of study subjects (e.g. **matching** or blocked randomization); and data editing (i.e. data checking and correction). Unfortunately, in these aspects also, the various packages are fairly weak. This is not,

Table 1 Type of sample size determination

Software	Means	Proportions	Other models[a]	Simulation
`Egret SIZ/StatXact`		`StatXact`	nonlinear reg.	`Egret SIZ`
`Epicalc`[b]	Yes	Yes	cc	
`EpiInfo`		Yes	cc, cohort	
`Epilog (Power)`[c]	Yes	Yes		
`PEPI`	Yes	Yes	OLS	
`Win Episcope`	Yes	Yes	cc, cohort	
`Stata`	Yes	Yes	rm	Yes
`True Epistat`	Yes	Yes	survival	

[a]Nonlinear reg. refers to `Egret SIZ`, which will calculate sample sizes for several types of nonlinear regression: logistic, conditional logistic, Poisson, Cox proportional hazards. cc = case–control study, cohort = cohort study, rm = repeated measures study, survival = comparison of two survival curves.

[b]Mark Myatt, the provider of `Epicalc`, also provides some specialized packages for calculating sample size for simple surveys and for "LQAS triage-style surveys".

[c]Epicenter Software, the vendor of `Epilog`, also sells a program called "`Power`" specifically for sample size determination. This was reviewed in Goldstein [3] and in Thomas & Krebs [12], but was not re-reviewed for this article.

however, to say that there is nothing available or that what is available is weak – it is just that too little is available, even though what is available is generally pretty good. With respect to data collection, the EPED module in `EpiInfo` can be very helpful, particularly regarding questionnaires. When combined with the CHECK module for data checking and verification, `EpiInfo` can be very helpful to researchers. Some other packages have provided less useful routines for data checking including `Epilog` (`Proc Check`) and `Stata` (`assert`, `inspect`, etc.). Any reasonable analytic package will provide some ability to check one's data for problems through standard descriptive statistics. `Epilog`'s `Proc Check` allows one to go a little further by allowing user input to flag, in different ways, any problem records as they are found. With some programming on the part of the user, `Stata`'s relevant commands can also provide additional help.

Although many authors have pointed out problems with matching, it can still be useful in epidemiologic studies. The biggest problem, for many, is how to find the matches. Of the packages examined, only `Epilog` (`Proc Control`) provides help here: for example, caliper matching (i.e. matching within limits, such as ± 5 years of age) is fairly easy to set up and the user can select the number of matches to select per case. A related issue is the ability to set up blocked randomization; `PEPI`, `Stata` and `True Epistat` have routines for this.

One can easily imagine other tasks that would be useful. For example, in longitudinal studies, software that helped with the tracing process, either through helping to find people via hooks to internet databases, or that helped determine which of several people found was the correct one, would be very useful. `Stata`'s extensibility could be used here, though it has not been. Another possibility relates to the presence of multiple control groups – none of the packages has anything, either on the data management end or in analysis (except for relatively simple tables and hypothesis tests) related to this potentially very useful procedure.

Data Management Issues

Some data management issues (e.g. data entry and verification) were discussed above. A few others are mentioned here. However, because of the wide variety of possible issues, I cannot hope to cover everything. I believe that there is a certain minimum that each package (except the simplest special-purpose analysis packages) should contain, including:

- merging files (adding new variables to the same cases), including many-to-one merges
- appending files (adding new cases to the data set)
- recoding variables
- transformations of variables, including a wide variety of built-in mathematical, statistical and string functions; some, at least, should be usable "on the fly" when estimating models (e.g. $\log(y) = f(x_1, x_2$, etc.)
- splitting files
- generating random numbers/samples
- changing files from wide format (repeated measures as part of the same case) to long format (each measure as a separate case) and back again; this is often necessary for different kinds of longitudinal analysis
- an extensive ability to deal with character data, especially names, addresses and, at least to some extent, free text
- the ability to override any automatic decisions made by the software (e.g. one should be able to change easily the way data, or results, are displayed)
- date functions, both so that, e.g. calculation of age at any given time point is easy, and so that graphs and other output can easily be labeled in an appropriate way
- time-series, including seasonal effects
- automatic generation of appropriate indicator variables from a categorical variable, including user choice of reference group when estimating models
- easy ways to find any duplicate cases in the data
- multiple indicators for missing values (to indicate different reasons for being missing), and correct handling of missing values during transformations and the generation of new variables (e.g. when forming a new indicator variable that is coded as 1 when either $x > 5$ or $y > 3$, the new variable should be equal to 1 if x is equal to 6 even if y is missing)
- dealing with runs, or spells, or series of events (particularly those that do not fit into standard time-series formats).

None of the packages here is ideal in this respect and most of them are pretty weak with relatively short

lists of built-in abilities. Stata is the strongest in this regard, even if one excludes user-added capabilities. With the addition of user-written routines, Stata is much stronger than any of the other packages with respect to data management. Stata even has the ability to add notes to the data, or to individual variables in the data, that will be permanently saved with the data and can be edited.

Data Analysis

The heart of any of these packages is its analytic capabilities. Unfortunately, this is also the heart of one of the major disputes between some statisticians and some epidemiologists: whether to emphasize OR or **relative risk** (RR). It appears that many epidemiologists prefer measures of RR while many statisticians prefer the OR. In fact, a major strand of epidemiologic literature deals with the question of when an event is rare enough so that the OR is a good approximation of the RR! For many statisticians, the dependence of RR on the rate in the control group makes it a poor choice for any use. A cross-cutting, and more meaningful, dispute is whether to emphasize absolute or relative measures of effect. In a sense, the answer to both issues is the inclusion of multiple measures of effect, or at least the ability to obtain any effect measure the user desires. For tables, the software could offer options that the user could pick from; for models, different effects result naturally from different models. Thus, offering numerous models, especially when they are closely related, seems the obvious answer. For example, generalized linear models (GLMs) offer, through different choices of link functions, the ability to obtain different measures of effect. A simple example involves models based on the binomial distribution: using a logit link results in **logistic regression** with an OR effect; using a log link gives an RR effect, while using an identity link gives an effect measured in rate differences.

Another difference appears to be the dependence of many users on tables. However, ignoring potential **confounders**, or other relevant covariates, can be dangerous, since their inclusion could mean very different results. Again, a package that includes both tables and models should solve this problem. If the documentation of the package discusses which models appropriately build on which tables, as does Stata's, the user is in the best possible situation.

Table 2 presents a simple checklist of various relevant forms of analysis for the major packages. Those packages that only include tables (Epicalc and Win Episcope/WinEpi Ratios) are not included in the table. Also, the column for EpiInfo includes the other CDC packages (EpiMeta and Cluster). The danger of such a table, of course, is that it can be misleading for certain specialized issues and packages. There is little doubt that Epicure, for example, looks weaker than it actually is in such a table because it has a number of unique models that are not included but should be considered. In particular, it is the only included package that has models that are neither additive nor multiplicative (*see* **Additive Model**; **Multiplicative Model**). Egret also suffers in this way, though to a lesser extent. Even Stata suffers as it is the only package with a full GLM (though Epicure has aspects of GLM), which I have not included in the table.

The items in Table 2 were largely selected on the basis of three sources: Clayton & Hills [2], Oster [8], and Rothman & Greenland [9]. The table is not complete (an impossible task), but does cover most of the regularly used analytic procedures. Note that I have excluded descriptive statistics and epidemiologic tables, since all packages include these (PEPI does, however, have one interesting unique aspect to its tables: when the user wants kappa, a measure of agreement, PEPI also provides bias-adjusted and prevalence-adjusted bias-adjusted versions; see Byrt et al. [1] and Lantz & Nebenzahl [5]) (*see* **Kappa**). There is little difference in the way that epidemiologic tables are handled by these packages, too little difference to matter to most users. The same is true with respect to tabular analysis of matched case–control data: all packages offer something, most include some form of exact analysis and some form of **stratification**. I have also left out nonparametric statistics since this would require either a simple, misleading row on whether there are any, or an additional table. Instead, note that Egret, EpiInfo and PEPI have some nonparametric techniques, while Epilog, Stata and True Epistat have quite a lot of nonparametric statistical routines. All of the named packages have some exact procedures, but here I mean the more traditional nonparametric (e.g. Mann–Whitney rank sum, sign test) techniques.

The following are brief descriptions of those routines that might not be obvious from their row title.

Table 2 Data analysis routines

Analysis	Egret	Epicure	EpiInfo	Epilog	PEPI	Stata	True Epistat
Standardization	Y		Y	Y	Y	Y	Y
SMR		Y		Y	Y	Y	Y
Multidimensional tables				Y			
Complex surveys			Y			Y	
Disease clustering			Cluster	Y			
Meta-analysis			EpiMeta	Y	Y	Y	Y
Linear models	Y	Y		Y	Simple	Y	Y
Censored linear models				Y		Y	
Logistic regression	Y	Y		Y	Y	Y	Y
Additive logistic regression	Y	Y				Y	
Exact logistic	LogXact			Y			
Ordinal logistic				Y		Y	
Polytomous logistic				Y		Y	
Conditional logistic	Y	Y		Y	Y	Y	Y
Poisson regression	Y	Y		Y	Y	Y	
Cox models	Y	Y		Y		Y	Y
Parametric survival models	Y	Y		Y		Y	
Kaplan–Meier analysis	Y	Y		Y	Y	Y	Y
Extensive nonlinear models		Y		Y		Y	
Recursive partitioning (trees)				Y			
Simulations	Y	Y				Y	
Bootstrap						Y	
Multilevel models	Y			Y		Y	
Seasonality				Y	Y		
Maps			EpiMap	Y			

Note: Entries in the table, other than "Y", refer to other packages from the same provider (counting the CDC as one provider).

- *Multidimensional tables, often called loglinear analysis*: this refers to tables with three or more variables. Note that most such tables are relatively easy to estimate in packages with **Poisson regression**. Some people might expect stub-and-banner tables here, but I have not included them since only Stata has anything like such tables and its version is quite limited. "Stub-and-banner" tables can have multiple variables as the rows and/or as the columns. They are often used when summarizing the study population in one table (e.g. showing the distribution by group and by gender, showing the mean ages of each group, and the mean time since exposure).
- *Complex surveys*: this refers to analysis of data gathered via a nonsimple random sample (e.g. a clustered random sample). Note that EpiInfo can only analyze means and rates, while Stata can also estimate a number of regression models on these samples.
- *PEPI*: this provides simple, but not multiple, linear and nonparametric regression.
- *Censored linear models*: the Buckley–James model in Epilog appears to be superior to the Tobit model in Stata (Moon [7]).
- *Parametric survival models*: these include, for example, exponential or Weibull models for survival analysis. More on various forms of survival analysis and software can be found in Goldstein et al. [4] (*see* **Survival Analysis, Overview**).
- *Extensive nonlinear models*: with the recent interest in neural networks, there has been a corresponding increase in statistical models with extensive nonlinearities among the predictor variables in a model. Epilog has a neural network routine. Stata, instead, has more statistical procedures, including generalized additive models (GAMs), cubic splines and fractional polynomials. Stata also has a more traditional nonlinear least squares routine.
- *Simulations and bootstrap*: as we learn more about the weaknesses of traditional models, and as computers become more powerful, many

researchers are turning to computer-intensive methods. The row for "exact logistic regression" is one type of computer-intensive routine, a specific use of randomization of the data. Simulation and bootstrap are two other computer-intensive methods. Their implementation in Stata calls on the user to write some code for their specific use, but supplies all the front-end and back-end and housekeeping work so that actually doing a simulation to see the effect of errors in the predictor variables, for example, is very little work. The work required in Epicure is, however, considerable. Egret has especially easy-to-use Monte Carlo routines. However, the user can only use Monte Carlo where offered in the package and thus there is no flexibility. If it is offered and what it simulates is what you want (e.g. a p-value or confidence interval) then it is very easy to use; if it is not offered, then it cannot be used. Note that for small samples, either exact or simulated results will usually provide more accurate results than will use of the bootstrap.

- *Multilevel models, sometime called random coefficient or hierarchic models*: these are relatively limited in epidemiologic software as yet, but some packages have at least some capabilities (e.g. random effects logistic regression).
- *Seasonality*: the reference here is to epidemiologic uses for, for example, the detection of seasonal or secular trends in disease incidence (as compared, for example, with the use of traditional time-series techniques, which usually need more data).

Note that those packages that have extensive regression routines (Egret, Epicure, Epilog, Stata and True Epistat) also have extensive diagnostics for these regressions, including goodness-of-fit statistics and procedures. While there are minor differences in the offered routines, I believe that any of them would be sufficient for most users.

There are two other important analytic techniques that I think should be included: errors-in-variables (or misclassification) and multiple imputation. It is surprising, and disturbing, that of all the epidemiologic literature dealing with **misclassification**, only PEPI and Stata have anything of relevance, and neither is very good, though Stata's simulation abilities can be of great help. Multiple imputation (Rubin [10], Vach & Blettner [13]) is of importance when there are **missing data**, as, in my experience, there always are. Stata has an impute command, but it is for single imputation and will often give a result that is misleadingly precise – no other package has anything. I believe that these are major failures of all the included software.

None of the packages has excellent graphics, either analytic or publication. All could benefit from more effort in this area. On the other hand, the analytic graphics in most of the packages are sufficient for everyday use by a data analyst. Many of the graphic procedures would have been considered excellent as recently as 10 years ago – but now they are all behind the times. In addition, as noted above, many of these packages are still DOS-based; many of the Windows graphics adapters do not work well in DOS, further impairing the value of the graphics offered by these packages.

I have not specifically commented on importing or exporting data as a general matter (although a few comments are scattered above). In general, the presence of two specialized file-format transfer packages (DBMS/COPY and StatTransfer) makes this issue relatively unimportant. Either of these packages can be used for file import or export. DBMS/COPY can also be used to summarize data either for input into one of the summary-data-only packages (e.g. EpiCalc or PEPI) or for making stub-and-banner tables for inclusion in a report or publication. Both of these programs can import from, and export to, many other packages, including DBMS programs (e.g. ACCESS), spreadsheets (e.g. QuattroPro), statistical packages such as Stata, SAS or SPSS, and even some specialized graphics packages.

Finally, no matter how much thought a software provider puts into the output from their data management and analysis routines, not everyone will be happy. This presents another advantage of an extensible package – users, or the provider, can provide alternative ways to analyze the data or to present results. While several software providers mentioned output and reporting as one of the reasons for having specialized epidemiologic software, the examples they cited are unimpressive – in almost all cases users will want to change the presentation of results for publication purposes; in almost all cases, this will be very difficult, requiring users to type in the results anew. The only exception is Stata, where

user-written routines have added important output flexibility.

The importance of having an extensible package is, I hope, clear, even for those who do not expect ever to extend a package themselves. After all, if there are easy ways to share extensions, all can benefit from the extensions of a few. For example, `Stata`, the only extensible package here, has several ways to share additions to the package: there is a bimonthly publication, the *Stata Technical Bulletin* (STB) that is filled with new routines from both users and the vendor (and the software is available over the Internet); there is an active user community that both supplements the company's technical support and is a source of additions, primarily from users, to the package – this e-mail `listserv` can be joined via the company's web site (see below).

Documentation

Over the years, I have examined dozens, if not hundreds, of statistical software packages. The packages discussed here have particularly strong manuals – a very pleasant surprise. All the commercial packages, as well as those from the CDC, have extensive documentation, with extensive professional citations and fully worked examples of the use of the software. Furthermore, in general, the free packages, such as `PEPI`, at least have extensive help files, which are unusually good, in my opinion (note that `PEPI` also has a complete manual, as does `EpiInfo`). Several of the packages even include references to their manual in their online help, something I like and find helpful. The major weakness of most of the packages is in their indexes: only `Egret`'s is as extensive and well-organized as I like (`Epicure`'s is almost as good; its major weakness is that one has to remember to go to the index in the Release 2.0 manual to obtain coverage for all three manuals).

Conclusion

My answer to whether there should be specialized epidemiologic software is a qualified yes: there is always a place (1) for relatively limited specialized software that is primarily useful for quick calculations on tables and provides output that is almost exactly what an epidemiologist is expecting, and (2) for analytic techniques that are primarily useful for particular disciplines, and that have not, yet, made it into general-purpose packages. However, it is a mistake to think that specialized software is all that is needed: no discipline is entirely self-sufficient – there are techniques from other disciplines that are useful and important for epidemiologists. It would be prohibitively expensive for a provider of epidemiologic software to try to match `Stata`, `SAS` or `S-Plus`; it would also be a terrible waste of resources. There is also an important place, in my opinion, for both free and shareware analytic software and I am glad to find that there are several healthy free epidemiologic software packages.

In addition, there is great value to having free packages available (particularly for students and for those who are really part-time epidemiologists), especially if they are very easy to use and include unique features. Several of the free packages discussed here are very easy to use and both `EpiInfo` and `PEPI` have unique features. I hope that the commercial vendors pay attention to these, and other, free packages. However, there is a danger with these packages. Too little attention has been paid to the biasing effects of using simple statistical models, such as $2 \times K$ tables. When combined with the tendency to ignore misclassification and the effects of **missing data**, we can see, I think, one major need in the field: easy to use packages that allow users to build from tables to regression models of various kinds and that allow, and adjust, for misclassification and other types of errors in variables.

Software Sources

`Cluster: http://www.atsdr.cdc.gov/HS/cluster.html`

`Egret` for Windows: CYTEL Software Corp., 675 Massachusetts Avenue, Cambridge, MA 02139, USA; (617) 661–2011; `http://www.cytel.com` `Epicalc: http://www.myatt.demon.co.uk/index.htm`

`Epicure`: HiroSoft International Corp., 1463 E. Republican Ave., Suite 103, Seattle, WA 98112, USA; (206) 328–5301; `http://www.hirosoft.com`

EpiInfo: ftp://ftp.cdc.gov; directory for EpiInfo: /pub/software/epi/epi_info; directory for EpiMap (not discussed here): /pub/software/epi/epimap

Epilog Windows: Epicenter Software, P.O. Box 90073, Pasadena, CA 91109, USA; (626) 304–9487; http://icarus2.hsc.usc.edu/epicenter

EpiMeta: http://www.cdc.gov/epo/dpram/epimeta/epimeta.htm

PEPI: http://www.brixtonbooks.demon.co.uk/otherbks.htm#PEPI: PEPI for Windows, test version: http://www.myatt.demon.co.uk/index.htm

Stata: Stata Corp., 4905 Lakeway Drive, College Station, TX 77845, USA; (979) 696–4600, (800) 782-8272; http://www.stata.com

True Epistat: Epistat Services, 2813 Clearmeadow Drive, Mesquite, TX 75181, USA; (972) 222–3904, (800) 326-1488; epistat@attbi.com

Win Episcope and WinEpi Ratios: http://infecepi.unizar.es/pages/ratio/soft_uk.htm

References

[1] Byrt, T., Bishop, J. & Carlin, J.B. (1993). Bias, prevalence and kappa, *Journal of Clinical Epidemiology* **46**, 423–429.

[2] Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

[3] Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers, *The American Statistician* **43**, 253–260 (Correction: **44**, 264).

[4] Goldstein, R., Anderson, J., Ash, A., Craig, B., Harrington, D. & Pagano, M. (1989). Survival analysis software on MS/PC-DOS computers, *Journal of Applied Econometrics* **4**, 393–414.

[5] Lantz, C.A. & Nebenzahl, E. (1996). Behavior and interpretation of the kappa statistic: resolution of the two paradoxes, *Journal of Clinical Epidemiology* **49**, 431–434.

[6] McNemar, Q. (1947). Note on the sampling of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–157.

[7] Moon, C.-G. (1989). A Monte Carlo comparison of semiparametric tobit estimators, *Journal of Applied Econometrics* **4**, 361–382.

[8] Oster, R.A. (1998). An examination of five statistical software packages for epidemiologists, *American Statistician* **52**, 267–280.

[9] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Ed. Lippincott-Raven Publishers, Philadelphia.

[10] Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**, 473–489.

[11] Self, S.G. & Mauritsen, R.H. (1988). Power/sample size calculations for generalized linear models, *Biometrics* **44**, 79–86.

[12] Thomas, L. & Krebs, C.J. (1997). A review of statistical power analysis software, *Bulletin of the Ecological Society of America* **78**, 128–139 (or, http://sustain.forestry.ucb.ca/cacb/power).

[13] Vach, W. & Blettner, M. (1998). Missing data in epidemiologic studies, in *Encyclopedia of Biostatistics*, Vol. 4 P. Armitage & T. Colton, eds. Wiley, Chichester, pp. 2641–2654.

(*See also* **Software, Biostatistical**)

RICHARD GOLDSTEIN

Soper, Herbert Edward

Born: September 6, 1865, in London; UK.
Died: September 10, 1930.

Soper studied mathematics at Cambridge under Bertrand Russell. He then worked in electrical engineering, but studied statistics under **Karl Pearson** in 1907. During the next decade, Soper published, in ***Biometrika***, mathematical and computational studies of the distribution of the **correlation** coefficient, forerunners of Fisher's exact solution of 1915, and various other computational papers. In 1922 he published a book on *Frequency Arrays*, an individual approach to the use of probability **generating functions**. In 1923, he joined **John Brownlee** at the National Institute for Medical Research, later working there with **Major Greenwood**. In 1929, he contributed importantly to the theory of measles epidemics, formalizing mathematically the mechanism previously described by W.H. Hamer in 1906.

PETER ARMITAGE

Spatial Models for Categorical Data

Spatially referenced counts, proportions, and **rates** present particular data analytic challenges. Such statistics typically arise from a set of prespecified regions partitioning a study area. Cressie [20] refers to such an arrangement as "lattice data", where the lattice may consist of regions formed and arranged regularly (e.g. a square or hexagonal grid system), or irregularly (e.g. a set of administrative districts such as United States **census** tracts, or United Kingdom post code areas).

To begin, we may think of such data as analogous to a **contingency table**, with regions corresponding to the usual notion of "cells". Each cell contains an observed count or proportion, and we often wish to compare these values to those expected under some sort of model, perhaps based on regionally specified **covariates**. However, unlike traditional contingency tables, a spatial lattice has no margins, or underlying row by column independence hypothesis, of scientific interest. For irregular regions, there are no margins other than the observed total count, and for regular regions the margins rarely have substantive meaning (unless the orientation of the lattice happens to correspond to a particular trend).

The spatial structure of the data complicates direct application of traditional statistical methods in at least two more general ways. First, one may be reluctant to employ traditional assumptions of independence between observations (here, regional counts or proportions), instead wishing to allow for positive **correlation** between nearby observations, as would occur in the presence of an influential but unmeasured covariate with a spatial pattern or trend. Second, traditional notions of asymptotic approximation often fail in the spatial setting, where one cannot add additional regions without either expanding the study area (increasing domain asymptotics) or subdividing the current set of regions (infill asymptotics) (cf. [20, pp. 100–101]). Neither of these asymptotic approaches provides an entirely satisfactory scenario for many analyses of spatially referenced categorical data.

The lost or at least diminished applicability of these two standard tools for classical statistics impacts upon the manner in which **likelihood**-based methods of **estimation** and **inference** may be employed. Below, we briefly survey three general classes of statistical methods for modeling spatially referenced counts and proportions. The first class builds upon the family of "auto-models" defined in detail by Besag [6–8]. The second class uses **quasi-likelihood** estimation within a **generalized linear model** (GLM) framework. The third class uses mixed models wherein **random effects** induce spatial correlation. All approaches build inference from a GLM foundation based on underlying **binomial** or **Poisson** distributions for counts, resulting in **logistic** or **Poisson regression** models, respectively. Models for proportions typically consider the denominators to be fixed and known (e.g. population sizes from a census), and reduce to count-based GLMs where the denominator of the proportion becomes an offset in the model [42, p. 206].

Auto-models

Besag [7] presents a thorough development of the structure and analysis of so-called "auto-models". These are based on Markov random fields wherein, for $i = 1, \ldots, I$, the distribution of the outcome in region i, conditional on all other observations, depends only on those observations occurring in a set of neighboring regions (*see* **Markov Chain Monte Carlo**). The Hammersley–Clifford Theorem [7] specifies exactly when a set of conditional distributions for each region defines a valid joint distribution, with induced spatial correlations (*see* **Conditional Probability**).

Auto-Gaussian Models

The most flexible and frequently applied family of auto-models employs conditional Gaussian distributions yielding (under proper conditions) a valid joint multivariate Gaussian distribution for the outcome data (*see* **Multivariate Normal Distribution**). While a multivariate Gaussian joint distribution directly implies Gaussian conditional distributions, the converse is not so straightforward [3, 10]. However, as will become clear below, the conditions required for other conditional distributions to define a valid joint distribution are often more restrictive than those for Gaussian distributions. Direct applicability of

"auto-Gaussian" models to categorical data is limited, particularly for counts of rare events, where transformations and approximations are often inadequate to satisfactorily compensate for the underlying discrete, non-Gaussian variability.

Auto-logistic Models

For **binary data** associated with regions (e.g. species presence or absence), we turn instead to the "auto-logistic" model originally detailed in [6, 7]. Recent applications extend the model to allow covariates [30, 33, 34], and we present this more general formulation here. Let Y_i denote a **random variable** associated with region i and y_i its observed value, for $i = 1, \ldots, I$. For binary data, $y_i = 0$ or 1 for each region. Assume we observe a vector of covariates $\boldsymbol{x}_i$ and let N_i be a set of region i's "neighbors" (e.g. $N_i =$ the set of regions bordering region i). Then, the auto-logistic model is

$$\begin{aligned} &\text{logit}[\Pr(Y_i = 1 | \boldsymbol{x}_i, \{y_j, j \in N_i\})] \\ &\qquad = \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{j \in N_i} \gamma_{ij} y_j, \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}$ represents the vector of covariate effects and the γ_{ij} $(j \in N_i)$ denote parameters governing the impact of the neighboring observations on Y_i. The last term in equation (1) follows an autoregressive format, wherein we regress (within the link function) each observation on its neighboring observed values (*see* **ARMA and ARIMA Models**).

The autoregressive term in (1) generates an unwieldy normalizing constant, thereby hampering traditional likelihood approaches for all but very small numbers of regions. Besag [8] addresses this issue through the introduction of "**pseudo-likelihood**" estimation, in which one chooses parameter values that minimize the product of the conditional binomial probabilities in (1). While this product is not the true likelihood, it often provides a reasonable approximation. However, the addition of covariates further complicates matters. Markov chain Monte Carlo (MCMC) **algorithms** provide an alternative strategy [34] for fitting auto-logistic models. Indeed, the conditional structure of auto-models is custom-made for recursively updating algorithms such as the Gibbs sampler. Geyer [23] defines the use of MCMC to obtain **maximum likelihood** estimates (MLEs) in a variety of statistical models. Gumpertz et al. [30] illustrate pseudo-likelihood estimation and contrast it with the MCMC **maximum likelihood** approach, preferring pseudo-likelihood in their application for its computational convenience. Indeed, they note that one can obtain pseudo-likelihood estimates using maximum likelihood algorithms available from any standard logistic regression routine [49, 51, 52]. In contrast, Huffer and Wu [34] prefer an MCMC approach to obtain MLEs of model parameters, based on their **simulation** study [61] showing that MCMC estimates provide substantial improvement over pseudo-likelihood results when the amount of spatial interaction (determined by the parameters γ_{ij} above) is large. In practice, the choice of estimation technique is largely a matter of preference between computational simplicity (pseudo-likelihood) and statistical efficiency (MCMC MLE), somewhat guided by the suspected strength of any spatial correlation (*see* **Efficiency and Efficient Estimators**).

Auto-Poisson Models

Besag [7] presents an "auto-Poisson" model for counts, assigning each regional count a conditional Poisson distribution with mean dependent upon its neighboring counts. The auto-Poisson model serves as an extreme example of the care one must take in defining auto-models. Owing to the infinite support of the Poisson distribution, by the Hammersley–Clifford theorem, a valid joint distribution only exists for auto-Poisson conditional models with exclusively negative autoregressive parameters. Auto-Poisson models thus require neighboring values (counts) to be *negatively* correlated, as in resource competition models. This is quite opposite the intent of most spatial modeling, to reflect shared latent sources of variation (see also [20, pp. 427–428]).

Ferrándiz et al. [22] propose the use of an auto-model based on *truncated* Poisson counts, where the regional count of a rare disease is not allowed to exceed the number of persons at risk within the region. The truncation avoids the infinite support problem. They compare maximum pseudo-likelihood estimates to MCMC MLEs (using a Monte Carlo scoring approach) in the following model:

$$\begin{aligned} Y_i | \{y_j, j \neq i\} &\sim \text{Poisson}(\mu_i) \\ \log(\mu_i) &= \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{j \in N_i} \gamma_{ij} y_j, \end{aligned} \tag{2}$$

where $\boldsymbol{x}_i$, $\boldsymbol{\beta}$, the γ_{ij}, and N_i are as defined previously. We can extend the model to rates or proportions (i.e. models of Y_i/n_i where n_i denotes the (fixed) number of people or person-years at risk in the ith region), by including an additive term $\log(n_i)$ as an offset on the right-hand side of equation (2).

Kaiser and Cressie [35] formalize the truncation proposed in [22]. By "Winsorizing" Poisson variables through the **transformation**

$$Y_i^* = Y_i \cdot \mathcal{I}\{Y_i \leq L\} + L \cdot \mathcal{I}\{Y_i > L\}, \qquad (3)$$

for some predetermined upper limit L (where $\mathcal{I}\{\cdot\}$ denotes the indicator function), one can define automodels allowing positive spatial correlation for values $Y_i^*, i = 1, \ldots, I$ (*see* **Trimming and Winsorization**). Kaiser and Cressie [35] illustrate maximum likelihood estimation (based on standard iterative techniques such as Newton–Raphson) for such models (*see* **Optimization and Nonlinear Equations**). They note the dependence of results upon the choice of the upper limit L, and suggest use of values of L considerably larger than the largest observed value of Y_i.

Generalized Linear Models

Quasi-likelihood

The formal development of generalized linear models [48, 42] includes the class of "quasi-likelihood" estimation techniques, based on the specification of the first two **moments** of the likelihood function rather than the entire function itself. While directing most attention to independent observations, McCullagh and Nelder [42, Section 9.3] extend the ideas to dependent data.

In particular, let $\boldsymbol{X}$ be the **matrix** with rows $\boldsymbol{x}_i'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_I)'$ denote the vector of mean values for the regional counts $Y_1, \ldots, Y_I$, and $g(\boldsymbol{\mu})$ denote the link function such that $g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$. Then the quasi-likelihood function $Q(\mu_i; y_i)$ is defined by the relationship

$$\frac{\partial Q(\boldsymbol{\mu}; \boldsymbol{y})}{\partial \boldsymbol{\mu}} = \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \qquad (4)$$

where $\boldsymbol{V}$ represents a general symmetric positive definite matrix whose elements are functions of $\boldsymbol{\mu}$. Differentiating $Q(\boldsymbol{\mu}; \boldsymbol{y})$ with respect to $\boldsymbol{\beta}$, yields the quasi-likelihood score equations

$$\boldsymbol{U} = \boldsymbol{\Delta}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \qquad (5)$$

where $\boldsymbol{\Delta}$ is the matrix with elements $[\partial\mu_i/\partial\beta_j]$ and $j = 0, \ldots, p$ indexes the parameters in the GLM's linear predictor. Setting the score equations to zero and solving for $\boldsymbol{\beta}$ yields the quasi-likelihood estimates of the model parameters.

McCullagh and Nelder [42, pp. 333–335] note that the precision matrix $\boldsymbol{V}^{-1}$ must satisfy several conditions, some not easily verified in practice, to guarantee that a solution to the score equations exists. As a result, Wolfinger and O'Connell [60] and Gotway and Stroup [26], following Liang and Zeger [40, 63], limit attention to variance–**covariance matrices** written as

$$\boldsymbol{V} = \mathbf{v}_{\boldsymbol{\mu}}^{1/2}\boldsymbol{R}(\boldsymbol{\alpha})\mathbf{v}_{\boldsymbol{\mu}}^{1/2}. \qquad (6)$$

Here $\boldsymbol{R}$ denotes a matrix of correlations among the $Y_1, \ldots, Y_I$, as functions of the parameter vector $\boldsymbol{\alpha}$, and $\mathbf{v}_{\boldsymbol{\mu}}^{1/2}$ is a diagonal matrix of scale parameters, perhaps incorporating **overdispersion**. Liang and Zeger [40, 63] show that, under mild regularity conditions, even misspecified correlation matrices will generate **consistent estimators** of $\boldsymbol{\beta}$. In the spatial setting, Gotway and Stroup [26] suggest estimation of the elements of $\boldsymbol{R}$ via standard geostatistical techniques (e.g. based on the **variogram** or correlogram), then substituting the estimated matrix $\widehat{\boldsymbol{R}}$ into the score equations above, and solving for $\boldsymbol{\beta}$.

The quasi-likelihood approach outlined above provides *marginal* inference regarding covariate effects (i.e. estimates of effects averaged across the entire study population). For related analytic approaches for spatial data, see [27, 39, 44].

Generalized Linear Mixed Models (GLMMs)

Including random effects within a GLM allows covariate effects to vary between regions, and can offer different insights into data patterns than the marginal analyses above. For our purposes, the primary difference between the two approaches is how the models structure spatial correlation. **Marginal models** incorporate spatial correlation directly into the likelihood, while mixed models take an hierarchical approach, combining spatially correlated random effects with

conditionally independent observations given the random effects (*see* **Hierarchical Models**). The random effects typically follow a multivariate Gaussian joint distribution. The result is a somewhat simpler first stage (conditional independence), and a more convenient structure for modeling spatial correlations than with marginal modeling (i.e. a correlated Gaussian model rather than a correlated Bernoulli, binomial, or Poisson model).

The somewhat simpler conceptual structure still involves complicated inference, especially from the computational perspective, and various methods and approximations appear in the literature. Agresti et al. [2] provide a detailed introduction to **generalized linear mixed models**, including an excellent overview of approaches for model fitting. We outline two general classes of such approaches, the first based on modifications to a likelihood approach and the second based on MCMC implementations of Bayesian hierarchical models (*see* **Bayesian Methods**; **Bayesian Methods for Contingency Tables**).

Approximations to the Likelihood

As noted above (and detailed in [2]), specifying a GLMM involves two steps. First, conditional on the random effects, we assume the outcomes follow a distribution within the **exponential family** (e.g. Bernoulli, binomial, or Poisson distributions). Second, we specify the distribution of the random effects. In the spatial setting, we typically embed spatial correlations in the distribution of the random effects. The full likelihood function for the GLMM combines the conditionally independent first stage with the distribution of the random effects, often resulting in an intractable (or at least inconvenient) multidimensional integral. Likelihood-based approaches to fitting GLMMs involve some sort of approximation to this integral.

The type and extent of approximation varies between approaches. Some methods employ **numerical integration** and others use simulation-based **Monte Carlo** approximations (see, e.g. [14] and [59], respectively). Such approaches converge to "exact" likelihood inference as one takes finer and finer resolution, or larger and larger Monte Carlo sample sizes, respectively. The resolution of numerical methods and the Monte Carlo sample size are computational components under the control of the analyst and are not based on particular probabilistic or statistical approximations. Hence, one can make the "exact" methods as precise as computational resources allow.

Other approaches trade some amount of "exactness" for computational simplicity and stability, using additional simplifying approximations, primarily through the use of first-order Taylor series expansions of the integrand around estimates of the random effects [2, Section 4.2]. Two popular methods using this strategy are **penalized quasi-likelihood** (PQL) (Breslow and Clayton [15], Green [29]) and "pseudo-likelihood" (PL) [60]. The "penalty" in PQL adjusts the quasi-likelihood for the presence of the random effects. The term "pseudo-likelihood" derives from the use of "pseudo-data" (defined below) and differs from the "pseudo-likelihood" for auto-models introduced above, as well as from other uses of the same term (e.g. [25]), creating some potential for confusion.

The PQL and PL approaches are very similar (see [41, Section 11.4.3], and [2]), and both build from assumed normality (or near-normality) of the model residual process $\boldsymbol{y} - \boldsymbol{\mu}$, using a "working dependent variable" (Wolfinger and O'Connell's "pseudo data") defined by

$$\boldsymbol{y}_w = g(\widehat{\boldsymbol{\mu}}) + \Delta_{\widehat{\boldsymbol{\mu}}}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}), \tag{7}$$

where $\Delta_{\widehat{\boldsymbol{\mu}}}$ denotes the matrix with elements $[\partial g(\mu_i)/\partial \mu_i]$ evaluated at $\widehat{\boldsymbol{\mu}}$ (*see* **Residuals**). The working data lead to a system of score equations involving the fixed effects and the parameters defining the variance–covariance matrix of the random effects. The PQL and PL approaches update the current fixed effect given the current estimate of the covariance parameters using normal mixed model theory ([2, Section 4.2]; [50, Section 7.6]), maximizing the product of the conditional density of the data given the random effects and the density of the random effects given the current estimate of the associated covariance parameters. The update of the covariance parameters follows an assumed normal linear mixed model for the working (pseudo) data, given the current estimates of the fixed and random effects. Note that both steps involve assumptions of normality of the working (pseudo) data and linearity (at least locally) for the application of the normal linear mixed model theory (e.g. [31]) (*see* **Linear Mixed Effects Models for Longitudinal Data**).

Similar to the quasi-likelihood method outlined for generalized linear models above, Wolfinger and

O'Connell [60] focus on variance–covariance matrices of the form

$$\boldsymbol{V} = \mathbf{v}_{\boldsymbol{\mu}}^{1/2} \boldsymbol{R}(\boldsymbol{\alpha}) \mathbf{v}_{\boldsymbol{\mu}}^{1/2}. \quad (8)$$

For marginal models, $\boldsymbol{R}$ represents a spatial correlation matrix and we set the (spatial) random effects to zero. For conditional models, we set $\boldsymbol{R}$ to an identity matrix or a diagonal matrix to incorporate overdispersion, and incorporate spatial correlation through the distribution of the random effects.

PQL and PL provide two popular approaches to fitting GLMMs in general, and spatial GLMMs in particular, due to their relative simplicity, ease of implementation, and applicability to large data sets in comparison with the "exact" maximum likelihood approaches outlined earlier. (Recall that the latter are "exact" in the sense that their precision depends only on computational limitations, rather than on distributional or linearity assumptions, or on asymptotic or other approximations, that can introduce additional biases if unjustified.) McCulloch [43] provides a thorough discussion of the impact of the working pseudo data assumptions and additional numerical approximations in PQL and PL, and compares these methods to other classes of maximum likelihood algorithms for GLMMs.

Bayesian Hierarchical Models

The two-stage nature of GLMMs naturally lends itself to a hierarchical Bayes interpretation [16]. Such models are common in the "disease mapping" literature, where analysts seek accurate small area estimates of rates and proportions through "borrowing strength" from neighboring regions (*see* **Mapping Disease Patterns**). Tsutakawa [55] provides an early example. He assigns a Gaussian **prior** distribution to region-specific logits of disease risk, producing local estimates that represent a compromise between individual regional estimates and the overall disease rate. Clayton and Kaldor [18] expand this basic idea, and associated empirical Bayesian inference for region-specific **standardized** mortality/morbidity ratios (SMRs), to allow spatial correlation between neighboring regions (*see* **Empirical Bayes**). To accomplish this, they consider auto-Gaussian prior distributions inducing pairwise spatial dependence between region-specific expected counts. Besag et al. [11] extend these approaches to a fully Bayesian setting using MCMC algorithms. Clayton and Bernardinelli [17], Mollié [45], Wakefield et al. [57], and Congdon [19, Chapter 7] provide thorough introductions to the fully Bayesian approach. The models in [11] and [18] are widely applied in a variety of settings, so we detail these below.

Typical disease mapping data contain observed (Y_i) and expected (often age-standardized) disease counts (E_i) for each region $i, i = 1, \ldots, I$. The maximum likelihood estimate of the SMR for region i is Y_i/E_i. The first stage of the model presumes conditionally independent Poisson distributions for each regional count, parameterized by log-**Relative Risks** (θ_i) associated with each region i, that is

$$Y_i|\theta_i \overset{\text{ind}}{\sim} \text{Poisson}(E_i \exp(\theta_i)), i = 1, \ldots, I. \quad (9)$$

While the disease mapping models involve SMRs, the setting may be thought of more generally as a GLM with an offset (here, E_i) for each Y_i, where θ_i represents a linear function $\boldsymbol{x}_i'\boldsymbol{\beta}$ of region-specific covariates.

To induce spatial correlation at the second stage of the model, Clayton and Kaldor [18] propose the addition of region-specific random intercept terms defined by an auto-Gaussian prior distribution, resulting in regional SMR estimates compromising between the local MLE (Y_i/E_i) and the SMRs observed in neighboring regions. Besag et al. [11] expand this approach to include, separately, influence of both the overall disease rate in the entire study area, and the special influences of neighboring disease rates. In this case, one replaces θ_i by a linear combination of covariate effects and random effects, that is,

$$\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i + v_i, \quad (10)$$

where u_i and v_i denote random effects (intercepts) measuring spatial similarity and excess heterogeneity (i.e. overdispersion, extra-Poisson variation), respectively. The basic structure resembles that of the auto-models above, but incorporates spatial pattern into random rather than fixed effects.

One typically assumes independence between $\mathbf{u}$ and $\mathbf{v}$, and models excess heterogeneity through a set of **exchangeable** priors for the v_i, for example, by $v_i \overset{\text{ind}}{\sim} N(0, 1/\tau)$, $i = 1, \ldots, I$. To model spatial similarity in residuals, [18] and [11] assign an auto-Gaussian model defining a *conditional autoregressive* structure for the set of u_i. Specifically, we define

the prior distribution of each u_i conditional on the u_j, $j \neq i$, as

$$u_i | u_{j \neq i} \sim N\left(\frac{\sum_{j \neq i} w_{ij} u_j}{\sum_{j \neq i} w_{ij}}, \frac{1}{\lambda \left(\sum_{j \neq i} w_{ij} \right)} \right), \quad i = 1, \ldots, I. \tag{11}$$

The w_{ij} denote weights defining the extents to which regions i and j are neighbors (by convention all $w_{ii} = 0$), and λ denotes a hyperparameter controlling how similar u_i is to its neighboring u_j, $j \neq i$. Typical applications consider adjacency-based binary weights where $w_{ij} = 1$ if regions i and j are adjacent and $w_{ij} = 0$ otherwise, although other options appear in the literature (e.g. [12]).

As noted above, Besag [7] shows that the collection of conditional distributions uniquely defines a corresponding multivariate normal joint distribution. However, the choice of binary adjacency weights leads to a joint distribution with singular precision matrix, so that the spatial similarity implied by the conditional distributions (in this case, an *intrinsic autoregression*) does not translate directly into a model of spatial correlation [10, 38]. Also, such multivariate priors are improper by virtue of the singularity, since they only define contrasts between pairs of the u_i. However, the inclusion of any informative data (through the likelihood function) results in a proper posterior (see [9, 11]). Finally, the constraint $\sum_{i=1}^{I} u_i = 0$ is often imposed in order to allow **identifiability** of an intercept in $\boldsymbol{x}_i' \boldsymbol{\beta}$. Detailed discussions of conditional autoregressive structures are provided by [7, 10, 11] and [20, pp. 407–408, 410–423].

Specification of the hierarchical model is completed by defining (vague) priors for the covariate effects $\boldsymbol{\beta}$, and proper hyperprior distributions for the hyperparameters τ and λ. In practice, conjugate inverse **gamma distributions** are popular for the latter. Ghosh et al. [24] and Sun et al. [54] discuss restrictions on parameters for these hyperpriors to ensure posterior propriety.

Incorporation of two random intercepts u_i and v_i for each region overparameterizes the model, so the likelihood only identifies the regional sums $(u_i + v_i)$. The prior distributions allow posterior identifiability ([16], p. 308), however. A related research question involves determination of a "fair" allocation of prior variability between τ and λ to balance prior emphasis on the roles of global and local rates. This issue is complicated by the marginal nature of τ and the conditional nature of λ [4, 12, 21].

Inference proceeds via MCMC algorithms, which provide the analyst with sample-based posterior distributions for each model parameter. These in turn allow posterior inferences for SMRs, counts, proportions, and rates.

Extensions

Several authors provide spatiotemporal extensions to the approaches outlined above. Typically, either space or time plays a primary role, leading analysts to investigate temporally evolving spatial structures or spatially correlated time series [5, 32, 36, 37, 53, 58, 62]. The distinction primarily involves the structure of the data, with the former arising in space-rich, time-poor data sets and the latter in space-poor, time-rich data.

Another area of current development involves the analysis of data from incompatible spatial scales, for example, observed counts from census tracts and covariate values from counties. Such "misaligned" data complicate matters, particularly for auto-models and the hierarchical models outlined above, since such models are defined for a particular set of regions and do not readily scale up or down to different regions [1, 13, 46, 47]. Gotway and Young [28] provide a thorough review of methods from a wide variety of disciplines that attempt to address this issue.

In conclusion, the models above provide inference for spatially correlated counts, proportions, and rates. Each approach contains its own set(s) of assumptions, and most involve iterative computational techniques. Hence, analysts must be aware of the assumptions and computational requirements involved when applying such methods.

Finally, while spatial modeling provides inference for data violating some key assumptions of traditional statistical inference, spatial techniques are no panacea for the problems produced by inaccurate or **missing** observations, or the inherent major limitations of ecological inferences. Wakefield [56] provides details putting spatial modeling in a broader perspective, and illustrating several key inferential challenges in

ecologic analyses of observational data that spatial modeling does not address (*see* **Ecologic Fallacy**; **Ecologic Study**).

References

[1] Agarwal, D.K., Gelfand, A.E. & Silander, J.A. (2002). Investigating tropical deforestation using two-stage spatially misaligned regression models, *Journal of Agricultural, Biological, and Environmental Statistics* **7**, 420–439.

[2] Agresti, A., Booth, J.G., Hobert, J.P. & Coffo, B. (2000). Random-effects modeling of categorical response data, *Sociological Methodology* **30**, 27–80.

[3] Arnold, B.C., Castillo, E. & Sarabia, J.M. (1999). *Conditional Specification of Statistical Models*. Springer, New York.

[4] Bernardinelli, L., Clayton, D. & Montomoli, C. (1995a). Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine* **14**, 2411–2431.

[5] Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. & Songini, M. (1995b). Bayesian analysis of space-time variation in risk, *Statistics in Medicine* **14**, 2433–2443.

[6] Besag, J. (1972). Nearest-neighbor systems and the autologistic model for binary data, *Journal of the Royal Statistical Society, Series B* **34**, 75–83.

[7] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

[8] Besag, J. (1975). Statistical analysis of non-lattice data, *The Statistician* **24**, 179–195.

[9] Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science* **10**, 3–66.

[10] Besag, J. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions, *Biometrika* **82**, 733–746.

[11] Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

[12] Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A. & Conlon, E.M. (1999). Bayesian models for spatially correlated disease and exposure data, in *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith,eds. Oxford University Press,Oxford, pp. 131–156.

[13] Best, N.G., Ickstadt, K. & Wolpert, R.L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions, *Journal of the American Statistical Association* **95**, 1076–1088.

[14] Booth, J.G. & Hobart, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society, Series B* **61**, 265–285.

[15] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**, 9–25.

[16] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

[17] Clayton, D.G. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern,eds. Oxford University Press, London, pp. 205–220.

[18] Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risk for use in disease mapping, *Biometrics* **43**, 671–681.

[19] Congdon, P. (2003). *Applied Bayesian Modelling*. John Wiley & Sons, Chichester.

[20] Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition*. John Wiley & Sons, New York.

[21] Eberly, L.E. & Carlin, B.P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models, *Statistics in Medicine* **19**, 2279–2294.

[22] Ferrándiz, J., López, A., Llopis, A., Morales, M. & Tejerizo, J.L. (1995). Spatial interaction between neighboring counties: cancer mortality data in Valencia (Spain), *Biometrics* **51**, 665–678.

[23] Geyer, C.J. (1991). Markov chain Monte Carlo maximum likelihood, in *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation of North America, Fairfax Station, Virginia, pp. 156–163.

[24] Ghosh, M., Natarajan, K., Waller, L.A. & Kim, D. (1999). Hierarchical GLMs for the analysis of spatial data: an application to disease mapping, *Journal of Statistical Planning and Inference* **75**, 305–318.

[25] Gong, G. & Samaniego, F.J. (1981). Pseudo maximum likelihood estimation: Theory and applications, *Annals of Statistics* **9**, 861–869.

[26] Gotway, C.A. & Stroup, W.W. (1997). A generalized linear model approach to spatial data analysis and prediction, *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 157–178.

[27] Gotway, C.A. & Wolfinger, R.D. (2003). Spatial prediction of counts and rates, *Statistics in Medicine* **22**, 1415–1432.

[28] Gotway, C.A. & Young, L.J. (2002). Combining incompatible spatial data, *Journal of the American Statistical Association* **97**, 632–648.

[29] Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models, *International Statistical Review* **55**, 245–259.

[30] Gumpertz, M.L., Graham, J.M. & Ristaino, J.B. (1997). Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variables on disease presence, *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 131–156.

[31] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related

problems, *Journal of the American Statistical Association* **72**, 320–340.

[32] Heisterkamp, S.H., Doornbos, G. & Nagelkerke, N.J.D. (2000). Assessing the impact of environmental pollution sources using space-time models, *Statistics in Medicine* **19**, 2569–2578.

[33] Hoeting, J.A., Lecaster, M. & Bowden, D. (2000). An improved model for spatially correlated binary responses, *Journal of Agricultural, Biological, and Environmental Statistics* **5**, 102–114.

[34] Huffer, F.W. & Wu, H. (1998). Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species, *Biometrics* **54**, 509–524.

[35] Kaiser, M.S. & Cressie, N. (1997). Modeling Poisson variables with positive spatial dependence, *Statistics and Probability Letters* **35**, 423–432.

[36] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in Medicine* **19**, 2555–2567.

[37] Knorr-Held, L. & Besag, J. (1998). Modelling risk from a disease in time and space, *Statistics in Medicine* **17**, 2045–2060.

[38] Künsch, H.R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice, *Biometrika* **74**, 517–524.

[39] Leroux, B.G. (2000). Modelling spatial disease rates using maximum likelihood, *Statistics in Medicine* **19**, 2321–2332.

[40] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

[41] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute, Inc, Cary.

[42] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[43] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92**, 162–170.

[44] McShane, L.M., Albert, P.S. & Palmatier, M.A. (1997). A latent process regression model for spatially correlated count data, *Biometrics* **53**, 698–706.

[45] Mollié, A. (1996). Bayesian mapping of disease, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds. Chapman & Hall/CRC, Boca Raton, pp. 360–379.

[46] Mugglin, A.S. & Carlin, B.P. (1998). Hierarchical modeling in geographic information systems: population interpolation over incompatible zones, *Journal of Agricultural, Biological, and Environmental Statistics* **3**, 111–130.

[47] Mugglin, A.S., Carlin, B.P. & Gelfand, A.E. (2000). Fully model-based approaches for spatially misaligned data, *Journal of the American Statistical Association* **95**, 877–887.

[48] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society* **135**, 370–384.

[49] Preisler, H.K. (1993). Modelling spatial patterns of trees attacked by bark-beetles, *Applied Statistics* **42**, 501–514.

[50] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. John Wiley & Sons, New York.

[51] Strauss, D. (1992). The many faces of logistic regression, *American Statistician* **46**, 321–326.

[52] Strauss, D. & Ikeda, M. (1990). Pseudo-likelihood estimation for social networks, *Journal of the American Statistical Association* **85**, 204–212.

[53] Sun, D., Tsutakawa, R.K., Kim, H. & He, Z. (2000). Spatio-temporal interaction with disease mapping, *Statistics in Medicine* **19**, 2015–2035.

[54] Sun, D., Tsutakawa, R.K. & Speckman, P.L. (1999). Posterior distribution of hierarchical models using CAR(1) distributions, *Biometrika* **86**, 341–350.

[55] Tsutakawa, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates, *Journal of the American Statistical Association* **83**, 37–42.

[56] Wakefield, J. (2003). Sensitivity analyses for ecological regression, *Biometrics* **59**, 9–17.

[57] Wakefield, J., Best, N.G. & Waller, L.A. (2000). Bayesian approaches to disease mapping, in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best & D.J. Briggs, eds. Oxford University Press. Oxford, pp. 106–127.

[58] Waller, L.A., Carlin, B.P., Xia, H. & Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates, *Journal of the American Statistical Association* **92**, 607–617.

[59] Wei, G.C.G. & Tanner, M.A. (1990). Random effects in ordinal regression models, *Computational Statistics and Data Analysis* **22**, 537–557.

[60] Wolfinger, R.D. & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach, *Journal of Statistical Computing and Simulation* **48**, 233–243.

[61] Wu, H. & Huffer, F.W. (1997). Modeling the distribution of plant species using the autologistic regression model, *Environmental and Ecological Statistics* **4**, 49–64.

[62] Xia, H. & Carlin, B.P. (1998). Spatio-temporal models with errors and covariates: mapping Ohio lung cancer mortality, *Statistics in Medicine* **17**, 2025–2043.

[63] Zeger, S.L. & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**, 121–130.

(*See also* **Geographic Epidemiology**; **Geographic Patterns of Disease**).

LANCE A. WALLER

Spearman Rank Correlation

In a study of the relationship between two variables, the use of measures of **correlation** assumes that neither is functionally dependent upon the other. So, for example, we might ask whether body weight is related to height in 35-year-old men; or whether examination scores in music theory are related to examination scores in mathematics for first-year college students; or whether the blood levels of two steroid hormones are related in 20-year-old women. A quantitative measure of the strength of the correlation is a correlation coefficient, which expresses how closely a change in the magnitude of one of the variables is accompanied by a change in the magnitude of the other variable. This is also referred to as a measure of association or of correspondence (*see* **Association, Measures of**).

If the distributions underlying the two variables are far from **bivariate normal**, or if the data are ordinal (e.g. we know relative magnitudes–such as man A is taller than man C but shorter than man D–but we do not know their actual heights) (*see* **Ordered Categorical Data**), then **nonparametric** correlation techniques should be employed to test hypotheses about the relationship between variables or to set **confidence limits** around the correlation coefficients. Nonparametric correlation also is less sensitive to **outliers** than is its parametric analog. The underlying assumptions for nonparametric correlation are that the n pairs of ratio, interval, or ordinal data (*see* **Measurement Scale**) constitute a **random sample** and that the two members of each of the n pairs of data are measurements taken on the same subject.

Among the correlation coefficients proposed by Charles Spearman [19, 20] is a commonly used nonparametric correlation measure that Maurice Kendall formally associated with Spearman's name a quarter of a century later [14], and that is one of the oldest statistics based on ranks. The Spearman rank coefficient computed for a sample of data is typically designated as r_S.

If each of the n measurements of one of the variables is denoted as X_i (i.e. $X_1, X_2, \ldots, X_n$), then $\mathcal{R}(X_i)$ may represent the **rank** of X_i, where each rank is an integer, from 1 through n, indicating relative magnitude. The measurements may be ranked from high to low (e.g. rank 1 indicates the tallest person, rank 2 the next tallest, and so on, with rank n the shortest) or from low to high (rank 1 denotes the shortest and rank n the tallest). Similarly, each of the n measurements of the second variable may be denoted as Y_i (i.e. $Y_1, Y_2, \ldots, Y_n$), and $\mathcal{R}(Y_i)$ would denote the rank of Y_i, where the sequence of ranking (either high to low or low to high) is the same as for $\mathcal{R}(X_i)$. This is shown in Table 1.

Table 1

Person, i	1	2	3	4	5
Height (m), Y_i	1.59	1.66	1.82	1.73	1.91
Weight (kg), X_i	75.8	77.2	89.3	72.2	81.5
Rank of height, $\mathcal{R}(X_i)$	1	2	4	3	5
Rank of weight, $\mathcal{R}(Y_i)$	2	3	5	1	4
$d_i = \mathcal{R}(X_i) - \mathcal{R}(Y_i)$	−1	−1	−1	2	1
d_i^2	1	1	1	4	1

An $r_S = 0$ ("no correlation") indicates that the magnitudes of the ranks of one variable are independent of the magnitudes of the ranks of the second variable. A positive value of r_S ("positive correlation") indicates that the $\mathcal{R}(X_i)$s tend to increase as the $\mathcal{R}(Y_i)$s increase; a negative r_S ("negative correlation") indicates that the $\mathcal{R}(X_i)$s tend to decrease as the $\mathcal{R}(Y_i)$s increase.

If the sequence of ranks were identical for the two variables, we would say that there was a perfect positive correlation, and $r_S = 1.0$. This would occur, for example, if five pairs of data had these ranks:

$$\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{array} \quad \text{or these:} \quad \begin{array}{ccccc} 1 & 2 & 4 & 3 & 5 \\ 1 & 2 & 4 & 3 & 5 \end{array}$$

A perfect negative correlation (where $r_S = -1.0$) would be one in which the magnitudes of the ranks for one variable vary inversely with the sizes of the ranks of the second; for example,

$$\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 \end{array}$$

Computing the Coefficient

The widely used parametric correlation coefficient, known as the Pearson product–moment correlation

coefficient (*see* **Correlation**), is defined as

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2\right]^{1/2}}, \quad (1)$$

and commonly computed as

$$r = \left(\sum XY - \frac{\sum X \sum Y}{n}\right) \Bigg/ \left\{\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]\left[\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right]\right\}^{1/2}, \quad (2)$$

where $\overline{X}$ and $\overline{Y}$ are the means of the X_is and the Y_is, respectively, and where the summations ($\sum$) are each over all n data.

The Spearman **rank correlation** coefficient, r_S, may be obtained by subjecting the ranks, instead of the raw measurements, to the above calculations. For the example above, and substituting $\mathcal{R}(X_i)$ for X_i and $\mathcal{R}(Y_i)$ for Y_i:

$$\sum X_i = 1 + 2 + 4 + 3 + 5 = 15,$$

$$\sum (X_i)^2 = 1^2 + 2^2 + 4^2 + 3^2 + 5^2 = 55,$$

$$\sum Y_i = 2 + 3 + 5 + 1 + 4 = 15,$$

$$\sum (Y_i)^2 = 2^2 + 3^2 + 5^2 + 1^2 + 4^2 = 55,$$

and

$$\sum (X_i Y_i) = (1)(2) + (2)(3) + (4)(5) + (3)(1) + (5)(4) = 51.$$

Then,

$$\overline{X} = 15/5 = 3 \quad \text{and} \quad \overline{Y} = 15/5 = 3,$$

and

$$\sum (X_i - \overline{X})^2 = (1-3)^2 + (2-3)^2 + (4-3)^2 + (3-3)^2 + (5-3)^2 = 10,$$

$$\sum (Y_i - \overline{Y})^2 = (2-3)^2 + (1-3)^2 + (5-3)^2 + (3-3)^2 + (4-3)^2 = 10,$$

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = (1-3)(2-3) + (2-3) \times (3-3) + (4-3)(5-3) + (3-3)(1-3) + (5-3) \times (4-3) = 6.$$

Eq. (1) yields

$$r_S = \frac{6}{[(10)(10)]^{1/2}} = \frac{6}{10} = 0.60,$$

while (2) yields

$$r_S = \frac{51 - (15 \times 15)/5}{\left[\left(55 - (15)^2/5\right)\left(55 - (15)^2/5\right)\right]^{1/2}} = \frac{6}{[(10)(10)]^{1/2}} = 0.60.$$

As the sum of integers 1 through n (i.e. the sum of all n ranks) is $n(n+1)/2$, (2) employed for Spearman rank correlation may be written as

$$r_S = \frac{\sum \mathcal{R}(X_i)\mathcal{R}(Y_i) - n(n+1)^2/4}{\left\{\left\{\left[\sum \mathcal{R}(X_i)^2 - n(n+1)^2/4\right] \times \left[\sum \mathcal{R}(Y_i)^2 - n(n+1)^2/4\right]\right\}^{1/2}\right\}}. \quad (3)$$

Also, as the sum of the squares of all n ranks is $n(n+1)(2n+1)/6$, (2) using ranks can be reduced to

$$r_S = \frac{12\left[\sum \mathcal{R}(X_i)\mathcal{R}(Y_i) - n(n+1)^2/4\right]}{n^3 - n} \quad (4)$$

or

$$r_S = \frac{12\sum \mathcal{R}(X_i)\mathcal{R}(Y_i)}{n^3 - n} - \frac{3(n+1)}{n-1}. \quad (5)$$

Alternatively, the difference, d_i, for each pair of ranks may be obtained, and the following equation used:

$$r_S = 1 - \frac{6\sum d_i^2}{n^3 - n}, \quad (6)$$

which, for the above example, is

$$r_S = 1 - \frac{6(1+1+1+4+1)}{5^3 - 5} = 1 - \frac{48}{120} = 1 - 0.40 = 0.60.$$

Eq. (6) is most commonly encountered in textbooks, but (1) is very convenient on a computer.

Instead of the differences between pairs of ranks, one may use the sums of the ranks for each pair [15, p. 227; [20]]:

$$r_S = \frac{6\sum S_i^2}{n^3 - n} - \frac{7n + 5}{n - 1}, \tag{7}$$

where

$$S_i = \mathcal{R}(Y_i) + \mathcal{R}(X_i). \tag{8}$$

It can be shown [13] that in bivariate normal populations the Pearson correlation coefficient, ρ, is

$$\rho = 2 \sin\left(\frac{\pi}{6}\rho_S\right). \tag{9}$$

Tied Ranks

If two or more data have the same value, then they are said to be "tied", and each of their ranks may be set equal to the mean of the ranks of the positions they occupy in the ordered data set. For example, in the data set 70, 74, 74, 78, and 79 kg, data 2 and 3 are tied; the mean of 2 and 3 is 2.5, so the ranks of the five data are 1, 2.5, 2.5, 4, and 5. In the data set 1.6, 1.7, 1.9, 1.9, and 1.9 m, data 3, 4, and 5 are tied; the mean of 3, 4, and 5 is 4, so the ranks of the five data are 1, 2, 4, 4, 4.

Then these ranks would be subjected to (1), or, equivalently, the following calculation [9, p. 366] would be used as an alternative to (4):

$$r_S = \left(12\left[\sum \mathcal{R}(X_i)\mathcal{R}(Y_i) - n(n+1)^2/4\right]\right) \Big/ \left\{\left[(n^3 - n) - 12\sum t_x\right] \times \left[(n^3 - n) - 12\sum t_Y\right]\right\}^{1/2} \tag{10}$$

and the following [12, p. 38; [20]] is an alternative to (6):

$$r_S = \left((n^3 - n)/6 - \sum d_i^2 - \sum t_X - \sum t_Y\right) \Big/ \left\{\left[(n^3 - n)/6 - 2\sum t_X\right] \times \left[(n^3 - n)/6 - 2\sum t_Y\right]\right\}^{1/2}, \tag{11}$$

where

$$\sum t_X = \frac{\sum (t_i^3 - t_i)}{12}, \tag{12}$$

where t_i is the number of tied values of X in a group of ties (two in the paragraph above) and the summation is over all groups of tied Xs, and

$$\sum t_Y = \frac{\sum (t_i^3 - t_i)}{12}, \tag{13}$$

where t_i is the number of tied values of Y in a group of ties (three in the paragraph above) and the summation is over all groups of tied Ys. Similarly, the following [21] is an alternative to (7):

$$r_S = \frac{\left\{\begin{array}{c}\sum S_i^2 - [(n^3 - n)/6][(7n + 5)/(n - 1)] \\ -\sum t_X - \sum t_Y\end{array}\right\}}{\left\{\begin{array}{c}\left\{\left[(n^3 - n)/6 - 2\sum t_X\right]\right. \\ \left.\times\left[(n^3 - n)/6 - 2\sum t_Y\right]\right\}^{1/2}\end{array}\right\}}, \tag{14}$$

If $\sum t_X$ and $\sum t_Y$ are both zero, then (10) is equivalent to (4), (11) equals (6), and (14) is equivalent to (7). The results from these equations for tied and nontied data are noticeably different only if there are many ties.

Testing Hypotheses

The r_S calculated from a sample of data is an estimate of ρ_S, the Spearman rank correlation coefficient that would be obtained from the entire population of data from which that sample came; ρ_S is sometimes called "Spearman's rho".

A common desire in rank correlation analysis is to test the **null hypothesis** that there is no correlation in the population between the paired ranks, i.e. we wish to test the two-tailed hypotheses $H_0 : \rho_S = 0$ vs. $H_a : \rho_S \neq 0$ (*see* **Hypothesis Testing**). There are many tables of critical values of r_S, and if r_S is greater than the relevant critical value, then H_0 is rejected. The use of $\sum d_i^2$, instead of r_S, as the test statistic for rank-correlation testing is sometimes called the "Hotelling–Pabst test" [10]. $\sum d_i^2$ is small when r_S is large, and H_0 is rejected if $\sum d_i^2$ is *less than* the critical value. Published tables offer critical values for various sample sizes, n, and levels of significance,

α. The most extensive of such tables for r_S are those of Zar [22, Appendix, pp. 115–116] and, with slight improvements, of Ramsey [17]. If there are tied data, critical values are only approximate. It should be noted that computer software packages may use approximations that are not as accurate as published tables.

One-tailed hypotheses may also be considered. For $H_0 : \rho_S \leq 0$ vs. $H_a : \rho_S > 0, H_0$ is rejected if r_S is positive and greater than the critical value for $\alpha/2$. For $H_0 : \rho_S \geq 0$ vs. $H_a : \rho_S < 0, H_0$ is rejected if r_S is negative and its absolute value is greater than the critical value for $\alpha/2$. If n is larger than that in these large tables, then one may compute

$$t = \frac{r_S}{s}, \tag{15}$$

where s, the standard error of r_S, is

$$s = \left(\frac{1 - r^2}{n - 2}\right)^{1/2}, \tag{16}$$

for which two- and one-tailed critical values of t (**Student's *t* distribution**), for $df = n - 2$, are readily found. Equivalently, one may employ

$$F = \frac{1 + |r_S|}{1 - |r_S|} \tag{17}$$

[2], referring to two- or one-tailed critical values of the ***F* distribution** for numerator and denominator $df = n - 2$. Using t or F is valid even with tied data, and is preferable in any case to employing the normal approximation,

$$Z = r_S(n - 1)^{1/2}. \tag{18}$$

The Fisher Transformation

If n is at least moderately large, the Spearman correlation coefficient may be subjected to the Fisher z transformation by

$$z = 0.5 \ln \frac{1 + r_S}{1 - r_S}, \tag{19}$$

and there are tables, e.g. [22, Appendix pp. 110–111], available to obviate the need to perform this computation. With this transformed value, one may test null hypotheses that ρ_S equals some value other than zero; i.e. $H_0 : \rho = \rho_0$ vs. $H_a : \rho = \rho_0$, where $\rho_0 \neq 0$. This is done via

$$Z = \frac{z - \zeta_0}{\sigma_z}, \tag{20}$$

where z is the transform of r_S; ζ_0 is the transform of the hypothesized coefficient, ρ_0; the standard error of z is approximated by

$$\sigma_z = \left(\frac{1.060}{n - 3}\right)^{1/2} \tag{21}$$

[7, 8], and Z is a normal deviate. In this fashion both two-tailed and one-tailed hypotheses may be tested.

Confidence Limits

The z transformation also allows the setting of approximate $1 - \alpha$ confidence limits for ρ_S. The confidence limits for the z transformation are

$$z \pm Z_\alpha \sigma_z, \tag{22}$$

where $Z_\alpha = t_\alpha(\infty)$. Then, the lower confidence limit of the transformation, $L_1 = z - Z_\alpha \sigma_\alpha$, is converted to the lower confidence limit of ρ_S by

$$\frac{\exp(2L_1) - 1}{\exp(2L_1) + 1}, \tag{23}$$

and the upper confidence limit of the transformation, $L_2 = z + Z_\alpha \sigma_\alpha$, is converted to the upper confidence limit of ρ_S by substituting L_2 for L_1 in (23) above. Published tables, e.g. [22, Appendix, pp. 112–114], execute these conversions.

Power of Testing

For data that meet the normality assumptions of parametric correlation analysis, use of the Spearman method has a relative **efficiency** of $9/\pi^2 = 0.912$ compared with the parametric procedure for testing hypotheses about the population correlation coefficient [10]. For other data distributions, the Spearman procedure may perform even better. The power of hypothesis tests for ρ_S, and the determination of the minimum sample size needed to achieve a desired power, may be approximated by an adaptation of the procedures of Cohen [4, p. 546], as shown by Zar [22, pp. 379–380, 392].

Other Rank Correlation Measures

The Kendall rank correlation coefficient [11, 12] is the other commonly encountered rank correlation measure (*see* **Rank Correlation**). It is often referred to as Kendall's tau, with the population parameter designated as τ and the sample estimate of τ denoted as $\hat{\tau}, t, T$, or (unfortunately) τ. Whereas $\hat{\tau}$ is an unbiased estimate of τ, r_S is a biased estimate of ρ_S, with $E(r_S) = [3\tau + (n-2)\rho_S]/(n+1)$ [6], but this **bias** disappears rapidly as n increases.

The two rank-correlation procedures have different underlying premises and influences (e.g. r_S is more affected by larger d_is), so they do not necessarily yield identical coefficients, $\hat{\tau}$ and r_S; indeed, data sets may have the same $\hat{\tau}$s yet different r_Ss. However, there is a very strong correlation between the two coefficients, and they each may range between -1.0 and 1.0. Daniels [5] found the relationship

$$-(n-2) \le 3n\hat{\tau} - 2(n+1)r_S \le (n-2),$$

which, for large n, is

$$-1 \le 3\hat{\tau} - 2r_S \le 1.$$

A better relationship was proved by Durbin & Stuart [6] to be

$$\frac{3n\hat{\tau} - (n-2)}{2(n+1)} \le r_S \le 1 - \frac{(1-\hat{\tau})}{2(n+1)} \times [(n-1)(1-\hat{\tau}) + 4].$$

Whether r_S or $\hat{\tau}$ is preferable depends upon the criteria employed to make the judgment; Chow et al. [3] judged r_S to be the preferable estimator.

Spearman's [19, 20] introduction of correlation between ranks was accompanied by a correlation measure of which he was fond, the "Spearman footrule", based upon $\sum |\mathcal{R}(X_i) - \mathcal{R}(Y_i)|$ instead of $\sum[\mathcal{R}(X_i) - \mathcal{R}(Y_i)]^2$. This measure is less useful than r_S in statistical analysis and is no longer encountered.

If one's interest is predominantly in the correlation among the largest (or smallest) members in the two populations, then the weighted rank correlation concept [18, 16; see also [22], pp. 392–395] might usefully be employed.

The Spearman rank correlation coefficient, r_S, is related to the Kendall coefficient of concordance, W, when there are two sets of ranks, as

$$W = \frac{(r_S + 1)}{2}. \tag{24}$$

Basler [1] discusses a relationship between $\sum d_i^2$ and the **chi-square test** statistic in a fourfold **contingency table** with ordinal marginal categories (*see* **Two-by-Two Table**).

References

[1] Basler, H. (1988). Equivalence between tie-corrected Spearman test and a chi-square test in a fourfold contingency table, *Metrika* **35**, 203–209.

[2] Cacoullos, T. (1965). A relation between the t and F distributions, *Journal of the American Statistical Association* **60**, 528–531.

[3] Chow, B., Miller, J.E. & Dickinson, P.E. (1974). Extensions of Monte Carlo comparison of some properties of two rank correlation coefficients in a small sample, *Journal of Statistical Computation and Simulation* **3**, 189–195.

[4] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. Lawrence Earlbaum, Hillsdale.

[5] Daniels, H.E. (1950). Rank correlation and population models, *Journal of the Royal Statistical Society, Series B* **12**, 171–181.

[6] Durbin, J. & Stuart. A. (1951). Inversions and rank correlation coefficients, *Journal of the Royal Statistical Society, Series B* **13**, 303–309.

[7] Fieller, E.C., Hartly, H.O. & Pearson. E.S. (1957). Tests for rank correlation coefficients, *Biometrika* **44**, 470–481.

[8] Fieller, E.C., Hartly, H.O. & Pearson. E.S. (1961). Tests for rank correlation coefficients. II, *Biometrika* **48**, 29–40.

[9] Gibbons, J.D. & Chakraborti. S. (1992). *Nonparametric Statistical Inference*, 3rd Ed. Marcel Dekker, New York.

[10] Hotelling, H. & Pabst. M.R. (1936). Rank correlation and tests of significance involving no assumption of normality, *Annals of Mathematical Statistics* **7**, 29–43.

[11] Kendall, M.G. (1938). A new measure of rank correlation, *Biometrika* **30**, 81–93.

[12] Kendall, M.G. (1962). *Rank Correlation Methods*, 3rd Ed. Charles Griffin, London.

[13] Kruskal, W.H. (1958). Ordinal measures of association, *Journal of the American Statistical Association* **53**, 814–861.

[14] Lovie, A.D. (1995). Who discovered Spearman's rank correlation?, *British Journal of Mathematical and Statistical Psychology* **48**, 255–269.

[15] Meddis, R. (1984). *Statistics Using Ranks: A Unified Approach*. Basil Blackwell, Oxford.

[16] Quade, D. & Salama. I. (1992). A survey of weighted rank correlation, in *Order Statistics and Nonparametrics: Theory and Applications*, P.K. Sen & I. Salama, eds. Elsevier, New York, pp. 213–224.

[17] Ramsey, P.H. (1988). Critical values for Spearman's rank order correlation, *Journal of Educational Statistics* **14**, 245–253.

[18] Salama, I. & Quade. D. (1982). A nonparametric comparison of two multiple regressions by means of a weighted measure of correlation, *Communications in Statistics – Theory and Methods* **11**, 1185–1195.

[19] Spearman, C. (1904). The proof and measurement of correlation between two things, *American Journal of Psychology* **15**, 72–101.

[20] Spearman, C. (1906). "Footrule" for measuring correlation, *British Journal of Psychology* **2**, 89–108.

[21] Thomas, G.E. (1989). A note on correcting for ties with Spearman's ρ, *Journal of Statistical Computation and Simulation* **31**, 37–40.

[22] Zar, J.H. (1996). *Biostatistical Analysis*, 3rd Ed. Prentice-Hall, Upper Saddle River.

JERROLD H. ZAR

Specificity

For diagnostic or **screening** tests, the specificity is the probability that an individual without the disease will receive a correct, negative test result. A synonym is the true negative rate, this being the proportion of negative results assigned among the denominator of true noncases of disease. In the table for the entry on **sensitivity**, the specificity is $d/(b+d)$. The **false positive** rate $b/(b+d)$ is the complement of specificity; in other words, it is the probability that an individual without disease will get a positive test result.

A test with high specificity is useful clinically for ruling in potential disease; a positive result from such a test implies a relatively high chance of having the disease. Typically, if a test is designed to have high specificity, its **false negative** rate $c/(a+c)$ will also be high and the test sensitivity $a/(a+c)$ will be correspondingly low.

Achievement of high specificity is important when the implied costs of giving false positive test results to noncases are high relative to the costs of incorrectly assigning negative test results to individuals with the disease (the so-called false-negative results). For instance, in population screening programs for rare diseases such as cancer, high specificity is desirable to avoid large numbers of false positive test results that would require clinical follow-up to determine their true, nondisease status.

A second meaning of specificity refers to the capability of a measuring device (e.g. in the clinical chemistry laboratory) to detect a particular target substance in a sample of material, as opposed to giving a false positive reading with other substances.

In **multivariate analysis**, particularly in **factor analysis**, specificity refers to the proportion of total variation that is associated with a factor.

(*See also* **Clinical Epidemiology**; **Diagnostic Tests, Evaluation of**; **Diagnostic Tests, Likelihood Ratio**; **Diagnostic Tests, Multiple**; **Gold Standard Test**; **Receiver Operating Characteristic (ROC) Curves**)

STEPHEN D. WALTER

Spectral Analysis

Often when examining a **time series** $\{X(t)\}$ we are interested in cyclic effects (*see* **Circadian Variation**); that is, effects, $g(t)$, which repeat themselves at regular intervals. Sometimes such cyclic or periodic effects are fairly clear to the eye (see Figure 1), but this may not always be so. The study of this kind of cyclic behavior gives rise to harmonic analysis and spectral analysis. The spectral analysis of time series, with its concern with cyclic effects, can give quite different insights to time domain methods (*see* **ARMA and ARIMA Models**).

We start with the basic ideas of periodic functions. Formally, we say that a function $g(t)$ is periodic if it repeats itself at a fixed interval, so

$$g(t) = g(t \pm s) = g(t \pm 2s) = \cdots = g(t \pm ks) = \cdots.$$

The smallest (nonzero) s value is called the *period* of the function. The frequency f of oscillations is the number of repeats, i.e. periods or cycles per unit time. The frequency $f = 1/s$ is measured in cycles per unit time. Cycles per second, the most common measure in engineering, are called hertz.

Because the study of cycles will involve time, we will naturally have to take into account the interval Δt between observations, often known as the *sampling interval*, since we will assume a series observed at discrete time points. In theory we can have continuous records, often known as *analog signals*, but we shall assume a digital signal. The traces given in Figures 1 and 2 are examples of two quite different series; Figure 2 is sampled at 100 times per second, while the other, in Figure 1, is sampled annually.

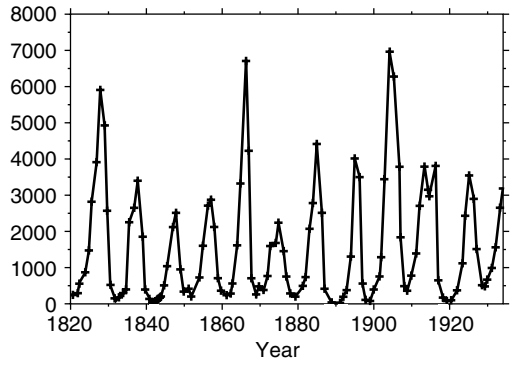

Figure 1 Annual number of lynx trapped, 1821–1934, MacKenzie River

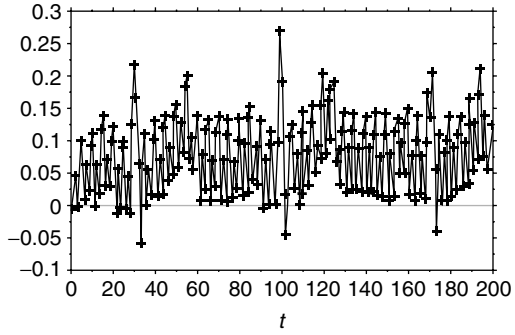

Figure 2 ECG trace, 2 s at 100 Hz

Harmonic Analysis

An obvious idea is to model a time series $X_1, X_2, X_3, \ldots, X_N$, the result of observations at times $\Delta t, 2\Delta t, 3\Delta t, \ldots, N\Delta t$ using regression techniques. If our interest is in cyclic effects, then we can think of trigonometric terms like $\cos(2\pi kt/N\Delta t)$ and $\sin(2\pi kt/N\Delta t)$ as providing the explanatory variables in a multiple regression. Since we may not know the periodic frequencies we could postulate a model of the form

$$X_t = \sum_{k=0}^{[N/2]} \left\{ a_k \cos\left(\frac{2\pi kt}{N\Delta t}\right) + b_k \sin\left(\frac{2\pi kt}{N\Delta t}\right) \right\}, \tag{1}$$

where the constant term becomes the coefficient at zero frequency. Notice that the highest frequency we may observe, the **Nyquist frequency**, is $1/2\Delta t$.

We can imagine performing some kind of stepwise regression procedure to find the nonzero coefficients and, consequently, the real frequencies. An early example of this is [6]. If we do the calculations, then we find that the component of the regression sum of squares explained by $\cos(2\pi kt/N\Delta t)$ is just $\hat{a}_k^2$, the square of the estimated coefficient of the cosine term. In the same way, $\hat{b}_k^2$ is the contribution from the $\sin(2\pi kt/N\Delta t)$ term; see [8] for details. Since these sums of squares give the importance of the corresponding frequency all we have to do is to plot

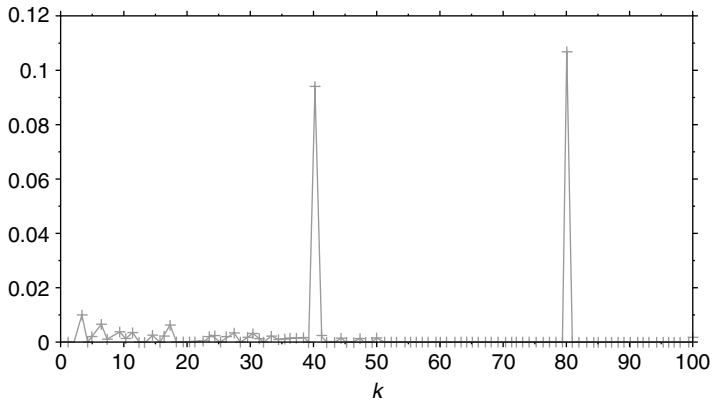

Figure 3 Periodogram of ECG series

the *periodogram*, $P(2\pi k/N\Delta t) = \hat{a}_k^2 + \hat{b}_k^2$, against $2\pi k/N\Delta t$ or, more simply, k. If the plot has a peak at k, then this implies that the corresponding frequency may well correspond to a periodic effect. Thus a spike at frequency $2\pi k/N\Delta t$ implies an interesting cyclic effect with frequency $k/N\Delta t$ (or period $N\Delta t/k$).

In Figure 3 the periodogram of an ECG trace, we see peaks in the periodogram at $k = 40$ and $k = 80$ indicating possible cyclic effects, one with period $200/40 = 5$ hundredths of a second or 0.05 s and another of 0.025 s.

Using the periodogram becomes even more attractive when we find, after some algebra, that the coefficients a_k and b_k have particularly simple forms. If we write our model in the usual complex form

$$X_t = \sum_{k=0}^{N/2} \left[a_k \cos\left(\frac{2\pi kt}{N\Delta t}\right) + b_k \sin\left(\frac{2\pi kt}{N\Delta t}\right) \right] = \sum_{k=0}^{N/2} \left\{ c_k \exp\left[i\left(\frac{2\pi kt}{N\Delta t}\right)\right]\right\}, \quad (2)$$

where $c_k = a_k + ib_k$ so that the periodogram is $|c_k|^2 = a_k^2 + b_k^2$ and in addition

$$c_k = \left(\frac{2}{N}\right)^{1/2} \sum_{t=1}^{N} X_t \exp\left(-\frac{2\pi kt}{N\Delta t}\right), \quad k = 0, 1, \ldots, \left[\frac{N}{2}\right]. \quad (3)$$

These complex trigonometric sums are widely used in engineering and signal processing where they are known as discrete Fourier transforms, or DFTs. These are popular because the coefficients c_k or, equivalently, the a_k and b_k give a unique description of the series; that is, if we know the $c_0, c_1, \ldots, c_N$ then we can reconstruct the original series since one can show that

$$c_k = \left(\frac{2}{N}\right)^{1/2} \sum_{t=1}^{N} X_t \exp\left(-\frac{2\pi kt}{N\Delta t}\right). \quad (4)$$

Many spectral calculations are based on the DFT of a series and because of the importance of DFTs, a fast and efficient method of computing them, known as the **Fast Fourier transform (FFT)**, has been developed.

We have seen that the coefficients c_k express the contribution of periodic components at differing frequencies to the observed series. This can be useful in modifying a series. If we consider (4), we could modify a series by setting the coefficient c_k to zero for frequencies of some chosen value f_0 thus eliminating higher frequency oscillations. After using (4) to recover the (modified) X_t using the modified c_ks, the resulting series with just the lower frequency contributions will be smoother. To demonstrate the effect of the just lower frequencies we set all the c_ks for the ECG data to zero after the 41st. The resulting plot in Figure 4 shows the effect of this operation. For a comprehensive account of DFTs see [9] or [16].

Angular Frequencies and Notation

We have used conventional frequencies f and have ended up with some quite complex formulas. To

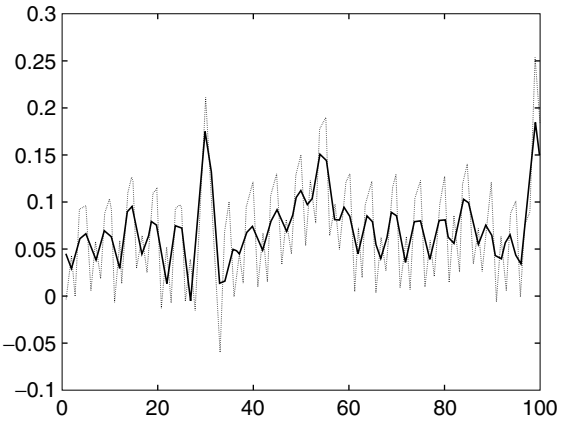

Figure 4 First 100 values of smoothed ECG trace

make matters more confusing, mathematical texts use angular frequencies, ω, measured in radians per unit time. This approach has the advantage of simplifying the formulas and is widely used in the (nonengineering) literature, so giving a more uniform notation. We shall use angular frequencies from now on and we will also simplify matters by assuming that sampling interval Δt is one. This means that we use the sampling times as the basic time measurement. To convert from f measured in cycles per unit time to angular frequencies ω is simple, we use $\omega = 2\pi f$, while in the other direction $f = \omega/2\pi$.

The Concept of the Power Spectrum

The clear and rather useful connections between the observed series and sums of trigonometric terms in the harmonic analysis above hint at a more intimate connection. Suppose we modify the expression in (2),

$$X_t = \sum_{k=0}^{p} \{a_k \cos \omega_k + b_k \sin \omega_k\}, \qquad (5)$$

and think of the $\{a_k\}$ and $\{b_k\}$ as zero mean sequences of independent random variables having variances σ_j^2. We can regard a model of this form as an attempt to explain the signal X_t in terms of contributions at the "angular frequencies" $\omega_0, \omega_1, \omega_2, \omega_3, \ldots, \omega_p$. A close analogy is that of a musical instrument; a note played on an instrument is the sum of harmonic vibrations at different frequencies.

We can rewrite (5) as

$$X_t = \sum_{k=0}^{p} \{a_k \cos \omega_k + b_k \sin \omega_k\} = \sum_{k=-p}^{p} z_k \exp(i\omega_k) \qquad (6)$$

for some p, and in angular terms. Here the z's are complex for $0 < |j| < p$, while at the end point $z_p = a_p$. The frequencies ω_{-k} are to be taken as $-\omega_k$, and if we take the obvious step of writing z_{-k} as the complex conjugate of z_k, then we have a nice equation and the novel idea of a "negative frequency".

If we use this complex formulation, then we can show that $\text{var}(X_t) = \sum_{j=0}^{p} \sigma_j^2$, so the variance of the process is made up of contributions from the individual frequencies. Suppose now we imagine that the number of frequency points in our model becomes very large and becomes a continuous range. Then the contributions from individual frequencies will tend to zero and we have a smooth function which describes the distribution of the variability over frequency. This is an exact analogy to the cumulative distribution function of a continuous random variable. We think of the contribution to the total variance made in a *frequency band* rather than at a specific frequency, thus the band $[\omega, \omega + \delta\omega]$ contributes $h(\omega)\delta\omega$ to the total variation. Notice we have assumed that there is no dominant frequency contributing a finite amount of power and the contribution *from any point frequency* is zero. This function h is called the power *spectrum*, and describes the amount of power contributed to the variance of the series in a narrow band.

The reader should keep in mind that the *h need not be a smooth and continuous function*. If a frequency ω contributes a finite amount of power, then the spectrum will have a singularity or peak at that frequency ω; indeed, a major use of the spectrum is to locate the peak and hence find to the frequencies which give finite power.

Properties of the Spectrum

For any second-order *stationary time series* we can define a power spectrum $h(\omega)$ defined on $-\pi \leq \omega \leq \pi$ and

1. The power spectrum $h(\omega)$ defines the amount of "power" or the contribution to the total variance made by frequencies in the band $[\omega, \omega + \delta\omega]$.

2. Harmonic components with finite power produce spikes or delta functions in $h(\omega)$.
3. The spectrum is symmetric (for real series), i.e. $h(\omega) = h(-\omega)$.
4. In the case where the power spectral $h(\omega)$ has no spikes we can show that the spectrum $h(\omega)$ and the autocovariance at lag k, $\gamma(k)$, are related by

$$h(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) \cos k\omega, \quad -\pi \leq \omega \leq \pi, \tag{7}$$

$$\gamma(k) = \int_{-\pi}^{\pi} h(\omega) \exp(-ik\omega)\, \mathrm{d}\omega, \quad k = 0, 1, 2, \ldots. \tag{8}$$

For real processes we can simplify a little, because $\gamma(k) = \gamma(-k)$, to give

$$h(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) \cos k\omega, \quad -\pi \leq \omega \leq \pi. \tag{9}$$

So if we have the autocovariances we can (with difficulty) calculate the spectrum, and vice versa. The implication is that all the information in the autocovariances is also contained in the spectrum, and vice versa. Some examples of spectra are:

1. A white noise process (*see* **Noise and White Noise**). Now $h(\omega) = \sum_{k=-\infty}^{\infty} \gamma(k) \cos k\omega$ for $-\pi \leq \omega \leq \pi$ and the autocovariances are zero apart from the first which is σ^2 so $h(\omega) = \sigma^2/2\pi$, a flat spectrum.
2. A first-order autoregressive (AR) model $X_t = \alpha X_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}$ is white noise (*see* **ARMA and ARIMA Models**). Using the autocovariances and some algebra we get

$$h(\omega) = \frac{\sigma^2}{2\pi(1 + \alpha^2 - 2\alpha \cos \omega)}.$$

3. A moving average (MA) process, say $X_t = \varepsilon_t - \beta\varepsilon_{t-1}$

$$h(\omega) = \frac{\sigma^2(1 + \beta^2 - 2\beta \cos \omega)}{2\pi}.$$

The flat white noise spectrum shows that the contributions to the variance are equally distributed across the entire frequency range, while for the AR model, here with parameter 0.5, the contributions are greater from the low frequency and hence long period end. This would imply that there are rather weak short period effects and in consequence a smoother series than white noise. The MA spectrum, again with a parameter of 0.5, shows the opposite effect; the power is concentrated towards the high frequency end of the frequency range and in consequence we expect a strong high frequency, i.e. short period effects giving an irregular appearance to a realization generated by such a model. (See Figure 5.)

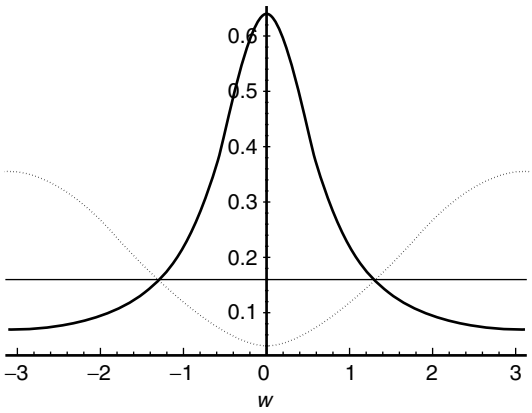

Figure 5 Three power spectra

The Spectral Representation

Looking back at (6) we have written the series as a sum of complex random variables of the form $X_t = \sum_{k=-p}^{p} z_k \exp(\omega_k t)$. One might ask: What happens to this as the number of individual frequencies becomes infinite? We can handle this by defining a stochastic process $Z(\omega)$ which is in some sense the accumulation of the z_js up to frequency ω. We take the limiting form as $X_t = \int_{-\pi}^{\pi} \exp(i\omega t)\, \mathrm{d}Z(\omega)$, where the process $Z(\omega)$ satisfies $\mathrm{E}[\mathrm{d}Z(\omega)\, \mathrm{d}Z(\phi)] = 0$ when $\omega \neq \phi$ and $\mathrm{E}[|\mathrm{d}Z(\omega)|^2] = h(\omega)\, \mathrm{d}\omega$.

This representation, known as the *spectral representation*, is mainly of technical interest for those who study the theory of spectra.

Linear Filters

Suppose we have a series $\{X_t\}$ which passes into a "black box", which produces $\{Y_t\}$ as output, rather as in the schematic in Figure 6. We could regard the

Figure 6 Black box

effect of the box as an operation on the input giving output, say $Y_t = \mathcal{L}X_t$. This would model many common situations; for example, the impact of a bump on the road is modified by the suspension system to give an output to the car occupants. The model is so general that it is difficult to handle, so we make two restrictions: (i) that the relationship is *linear*, and (ii) that the relationship is *invariant* over time. While it is not obvious, these restrictions mean, in effect, that for any t, Y_t is a weighted linear combination of past and future values of the input, namely

$$Y_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}, \quad \text{with} \quad \sum_{j=-\infty}^{\infty} a_j^2 < \infty. \qquad (10)$$

For technical details, see, for example, [11].

We call the relationship in (10) a *linear filter* and much of time series analysis is concerned with the study of such filters. Their virtue is that the relationship between the input and output *spectra* is simple. If the input series $\{X_t\}$ has a power spectrum $h_x(\omega)$ and the output $\{Y_t\}$ a corresponding spectrum $h_y(\omega)$, then they are related by

$$h_y(\omega) = \left| \sum_{j=-\infty}^{\infty} a_j \exp(-i\omega j) \right|^2 h_x(\omega). \qquad (11)$$

If we write $h_y(\omega) = |\Gamma(\omega)|^2 h_x(\omega)$, where $\Gamma(\omega) = \sum_{j=-\infty}^{\infty} a_j \exp(-ij\omega)$, then the function $\Gamma(\omega)$ is called the *transfer function* or the *frequency response function*, while $|\Gamma(\omega)|$ is often called the *amplitude gain*. The squared value, $|\Gamma(\omega)|^2$, is known as the *gain* or the *power transfer function* of the filter. The argument $\arg\{\Gamma(\omega)\}$ is the *phase gain* or just the *phase*. There is rather a rich variety of nomenclature since filters are widely used in many fields, especially in engineering.

We can see the value of this result in a simple case. Suppose we apply a **moving average** to a series, say a five-point moving average:

$$Y_t = \frac{X_{t-2} + X_{t-1} + X_t + X_{t+1} + X_{t+2}}{5}$$

$$= \frac{1}{5} \sum_{j=-2}^{2} X_{t-j}.$$

We know that this will remove a cycle in the data of period 5. Now we can investigate its properties using the transfer function. We can work out the transfer function as follows:

$$\begin{aligned} 5\Gamma(\omega) &= \exp(-2i\omega) + \exp(-i\omega) + 1 \\ &\quad + \exp(i\omega) + \exp(2i\omega) \\ &= 1 + 2\cos\omega + 2\cos 2\omega, \end{aligned}$$

and the squared gain is plotted in Figure 7. If this filter is applied to a series with a cycle of period 5, i.e. frequency $\omega = 2\pi/5$, then the resulting output spectrum, given by $h_y(\omega) = |\Gamma(\omega)|^2 h_x(\omega)$, will have a zero at this frequency and in consequence the output series will not contain this cyclic effect. Thus, as we expect, the moving average is a filter that removes cycles of period 5. In addition we see from Figure 7 that low-frequency (long-period) effects are not diminished, while high-frequency terms are reduced by the filter, so the filter will also smooth the series and will have only a small effect on long-term, low-frequency, components.

The filter above describes an action in the time domain in frequency terms, but in some circumstances it is natural to work in the other direction. Thus we might decide that the ECG trace could be distorted by the effects of the mains frequency (60 Hz) in our recording instrument. To eliminate this effect we would try to find a filter with a gain which is zero in a band around the 60 Hz mark. To find a

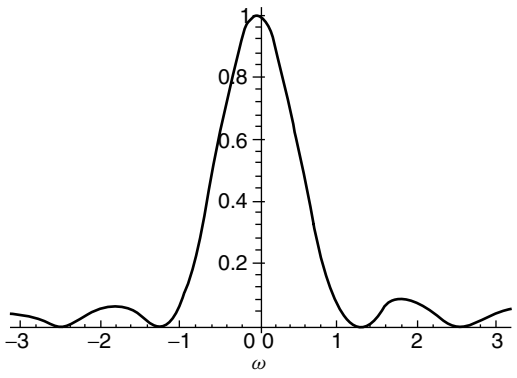

Figure 7 Squared gain for a five-point moving average

filter with a gain specified in this way requires some work (see [4]).

We can also make the black box serve our own purposes. Suppose our black box is a measuring instrument with the output signal being our experimental result. Suppose, further, that the instrumentation has an effect which distorts the observed signal. We can eliminate this effect by finding the gain due to the instrument. To do this we input a signal with a known spectrum, say a flat white noise spectrum, $f_{\mathrm{w}}(\omega)$, and find the output spectrum, $f_{\mathrm{o}}(\omega)$. The gain due to the system is then just the ratio of these two spectra, giving us the response of the instrument. This is rather like seismic surveying where an explosion provides an input signal which passes though rock while the reflected noise gives an output. The aim is to deduce the properties of the black box, the rock though which the signal has passed.

The design of filters is of considerable importance and much attention has been paid to the problems involved. See [4, 5], and [13].

Estimation of the Power Spectrum

We have discussed the spectrum and its uses but have avoided any statistics. We now consider estimation of the power spectrum. The obvious approach is to replace the autocorrelations in definition (9) by their estimates, $r(s)$, giving

$$\hat{h}(\omega) = \frac{1}{2\pi} \sum_{s=-m}^{m} r(s) \cos(s\omega), \tag{12}$$

where m is some suitable number of autocorrelations. Depending on our choice of covariance estimates we find that this is, approximately, a multiple of the periodogram.

Unfortunately, the periodogram is a poor estimate of the power spectrum. It is not a consistent estimator and because its values at adjacent frequencies are independent it is an erratic fluctuating function. Since the periodogram fluctuates wildly, one possibility is to smooth the function to make it more tractable. This nonparametric or windowed approach (*see* **Window Estimate**) uses as an estimate

$$\hat{h}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega - \theta) I_N(\theta)\, \mathrm{d}\theta. \tag{13}$$

The window $W(\theta)$ is a suitably chosen function, with $\int_{-\pi}^{\pi} W(\theta)\ \mathrm{d}\theta = 1$ and

$$I_N(\omega) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} X_t \exp(-it\omega) \right|^2$$

is the (modified) periodogram. We choose the functions $W(\theta)$ which are concentrated around zero and which decay to zero as $|\omega|$ becomes large. In fact, since the periodogram is calculated at discrete frequency points we should really have a sum $(1/N) \sum_{j=i}^{N} W(\omega - \omega_j) I_N(\omega_j)$, the idea being to average the periodogram ordinates near the frequency of interest, ω. It is rather convenient, however, to use the integral form as a notation, and this appears in much of the literature.

The problem is to select the window function $W(\theta)$ to ensure a reasonable estimate. For a sharply peaked function we can approximate crudely as follows:

$$\mathrm{E}[\hat{h}(\omega)] \approx h(\omega) \int_{-\pi}^{\pi} W(\theta)\ \mathrm{d}\theta, \tag{14}$$

$$\mathrm{var}[\hat{h}(\omega)] \approx 2\frac{h^2(\omega)\pi}{N} \int_{-\pi}^{\pi} W^2(\theta)\ \mathrm{d}\theta, \tag{15}$$

$$\mathrm{cov}[\hat{h}(\omega)\hat{h}(\phi)] \approx \frac{2\pi}{N} \int_{-\pi}^{\pi} (\omega - \theta) \times W(\phi - \theta) h(\theta)^2\ \mathrm{d}\theta. \tag{16}$$

A fairly sharply peaked window is generally a good idea, but if the window has subsidiary peaks, so-called "side lobes", then the estimate at a particular frequency ω may be contaminated by effects at other frequencies. The resulting distortion is called "leakage". To get some feel for the parameters we require we must define the peakedness or bandwidth of the window function. A simple definition of the bandwidth B_{w} is the width of a rectangular window having the same maximum height as $W(\omega)$ and the same area in the frequency of interest. Thus,

$$B_{\mathrm{w}} = \frac{1}{W(0)} \int_{-\pi}^{\pi} W(\theta)\ \mathrm{d}\theta = \frac{1}{W(0)}.$$

Three common windows are given in Table 1. These have differing bandwidths, etc. as can be seen from Table 2. The shape of the window is

Table 1 $W(\theta)$ functions

Unit	Bartlett	Parzen
$\frac{1}{2\pi}\left\{\frac{\sin\left(M+\frac{1}{2}\right)\theta}{\sin(\theta/2)}\right\}$	$\frac{1}{2\pi M}\left\{\frac{\sin(M\theta/2)}{\sin(\theta/2)}\right\}^2$	$\frac{3}{8\pi M^3}\left\{\frac{\sin(M\omega/4)}{\frac{1}{2}\sin(\omega/4)}\right\}^4$

Table 2

Window	Bandwidth	$\frac{\text{Variance}}{h^2(\omega)}$	Equivalent degrees of freedom (EDF)
Unit	$2(\pi/M)$	$2.00(M/N)$	N/M
Bartlett	$2(\pi/M)$	$2M/3N$	$3(N/M)$
Parzen	$8\pi/3M$	$0.54(M/N)$	$3.7(N/M)$

important if we are to have a *well-resolved* estimate, i.e. one that does not change very much over the bandwidth of the window. There have been many arguments over the choice of window; happily in almost all cases there is very little difference.

The choice of M is problematic. Obviously, we would like to make M as large as possible so as to decrease the bandwidth. If we do so, then the variance of the estimate at any frequency must increase. Thus, we need to find some compromise value for M. The usual pragmatic approach is to try values of M between $N/3$ and $N/5$. As M increases, the estimate becomes smoother and we choose a value that seems "smooth enough".

If there is a peak in the spectrum which is of interest and has a bandwidth, say B_f, which we can specify, we can choose the bandwidth $W(\theta)$, B_w, to fit this criterion. We require $B_w < B_f$ and take as a reasonable choice $B_w = \frac{1}{2}B_f$. Without a minimum bandwidth, B_f, any choice of the parameter M is somewhat arbitrary. For a fixed length of series the requirements of bandwidth and variance are contradictory. If one decreases the other increases and we need to come to some sensible compromise.

Lag Windows

We can look at these estimates in a rather different light if we assume that

$$W(\theta) = \frac{1}{2\pi}\sum_{-\infty}^{\infty}\lambda_j \exp(-i\theta j).$$

In this case our spectral estimate can be written

$$h(\omega) = \frac{1}{2\pi}\sum_{j=-M}^{M}\lambda_j r(j)\exp(-i\theta j) \qquad (17)$$

for some $M < N$. As we see, this is a weighted sum of estimated autocorrelations, $r(j)$, the weight sequence $\{\lambda_j\}$ being known as the *lag window*. We can in consequence think of the smoothed periodogram as a weighted sum of covariances, the trick being to choose a suitable sequence $\{\lambda_j\}$ or window $W(\theta)$.

An illustration of the effects of smoothing by changing the parameter in the spectral window is given in Figure 8. Here we look at a fox series of 93 annual observations [9], taken from Hudson Bay records. We use a Parzen window, and M values of $0.1N$, $0.2N$, and $0.3N$ are 9, 18, and 28.

As the truncation point decreases we have, as expected, a smoother estimate. The $0.1N$ seems rather too smooth so we concentrate on the $0.2N$ value. We are being rather arbitrary but having no background information it is not possible to set up any bandwidth arguments to select the appropriate smoothness. We have an apparent peak at frequency 0.26 cycles per year (1.62 radians per year) corresponding to a period of around four years and a subsidiary one at frequency 0.39.

Sampling Properties of the Smoothed Spectral Estimate

Given that the estimates we have considered are weighted sums of periodograms and the periodograms are independent χ^2 variables, we would expect to be able to approximate the distribution of our spectral estimates by a χ^2 distribution. In fact:

1. The spectral estimate $\hat{h}(\omega)$ has a distribution which is approximately χ^2 with ν degrees of

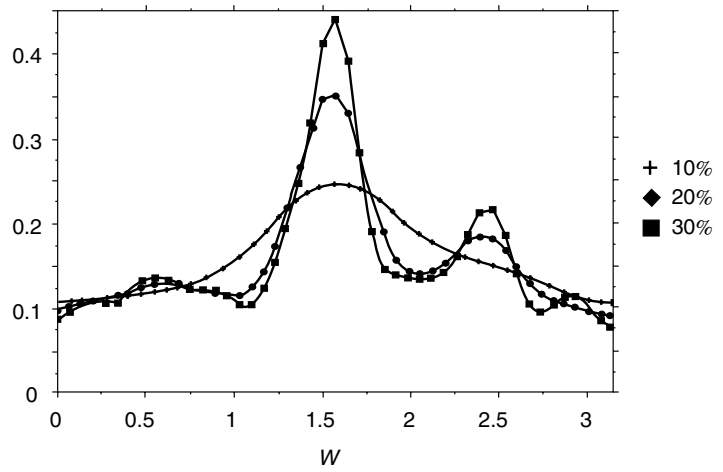

Figure 8 Estimates of spectrum of the fox series with three different truncation points

freedom. The *equivalent degrees of freedom* (EDF), ν, is defined as

$$\nu = 2\frac{\{\mathrm{E}[\hat{h}(\omega)]\}^2}{\mathrm{var}(\hat{h}(\omega))}.$$

These are given in Table 2 for three windows.

2. Spectral estimates at least one bandwidth apart will be assumed to be independent.

Given this approximate distribution it is easy to produce a confidence interval for the $\hat{h}(\omega)$. If we are given α, if a and b are the $\alpha/2$ quantiles of the χ^2 distribution with ν degrees of freedom say, $\Pr(\chi^2 \leq a) = \Pr(\chi^2 \geq b) = \alpha/2$, then the $(1-\alpha)100\%$ confidence interval is $[\nu\hat{h}(\omega)/b, \nu\hat{h}(\omega)/a]$. This gives a *pointwise* estimate rather than a confidence interval over a frequency band. For most cases this will suffice. One can find a band over all frequencies and in this case the reader is referred to [4, Chapter 6].

It is rather more satisfactory to consider $\log h(\omega)$ since the corresponding confidence intervals $[\log\hat{h}(\omega) + \log(\nu/b), \log\hat{h}(\omega) + \log(\nu/a)]$ have uniform width and are much easier to handle.

Tapering and Prewhitening

Two common techniques used for cutting down the bias in periodogram-based estimates are prewhitening and tapering. Prewhitening is a simple concept, the aim being to flatten the spectrum by filtering *before* estimating the spectrum. This is sensible because the most difficult spectra to estimate are those with sharp peaks and large ranges. In practice it is rather more complex since to design a filter to perform this "whitening" we need to know the form of the function we wish to estimate. Nevertheless it can be a valuable option and it is common for an approximate autoregressive model to be fitted to the data and for this to be used as a prewhitening filter.

Tapering is another bias reduction technique and involves adjustment of the data by multiplying by a sequence of constants $\{a_t\}$. The resulting values, $Y_t = a_t X_t$, are then used for the spectral analysis. If we regard the time series as extending into the infinite future, then our finite sample $X_1, \ldots, X_N$ is a tapered version of the infinite series with $h_t = 0$ for t exceeding N and $a_t = 1$ for $t \leq N$. The periodogram is then

$$\frac{1}{2\pi N}\left|\sum_{t=1}^{N} Y_t \exp(-i\omega t)\right|^2$$

$$= \frac{1}{2\pi N}\left|\sum_{t=1}^{N} a_t X_t \exp(-i\omega t)\right|^2$$

and we can show that the smoothed spectral estimate based on the tapered data has an effective smoothing window of the form $\int_{-\pi}^{\pi} D(\theta)W(\omega-\theta)\,\mathrm{d}\theta$, where $D(\omega) = |A(\omega)|^2 / \int_{-\pi}^{\pi} |A(\omega)|^2\,\mathrm{d}\omega$ and $A(\omega)$ is the Fourier transform of the taper sequence. The "finite

taper" above has a rectangular shape whose sharp corner gives a pronounced "ringing". The taper should smooth this corner and reduce the bias in the spectral estimate.

An important and interesting extension of this idea, called *multi-tapering*, was proposed by Thompson [17]. The data are smoothed with a sequence of "orthogonal tapers", i.e. a taper which give rise to uncorrelated periodogram estimates. The resulting periodogram estimates can then be averaged. The method appears to be very successful in reducing bias, especially at low frequencies. For a good account see Percival & Walden [13] and Waldren [18].

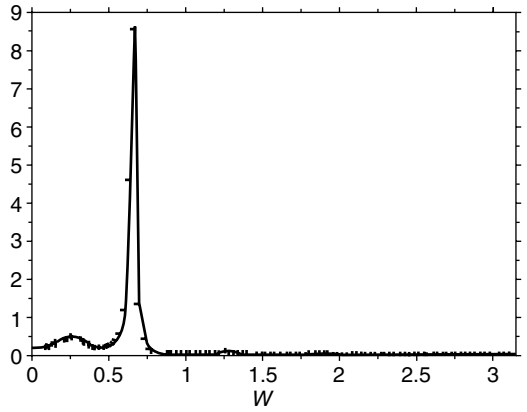

Figure 9 AR spectrum log lynx series

Parametric Spectral Estimates

The most popular parametric estimate for the spectrum are the so-called "autoregressive estimates". These are obtained by fitting an autoregressive model $\Psi(B)X_t = a_t$ and using as a spectral estimate $h(\omega) = \sigma^2/2\pi|\Psi[\exp(-i\omega)]|^2$ with the model estimates of σ and the autoregressive coefficients. A criterion is of course required to select the order (see [10]) of the autoregression and **Akaike's** information criterion (AIC) is often used. Parzen [12] suggested an alternative CAT criterion, while Akaike [1] suggested selecting the order that minimizes the final prediction error (FPE). There is little useful theory to help one choose a criterion (see [10]) while experience shows that the procedures tend to select AR orders in the range $N/3$ to $N/2$ for reasonable results.

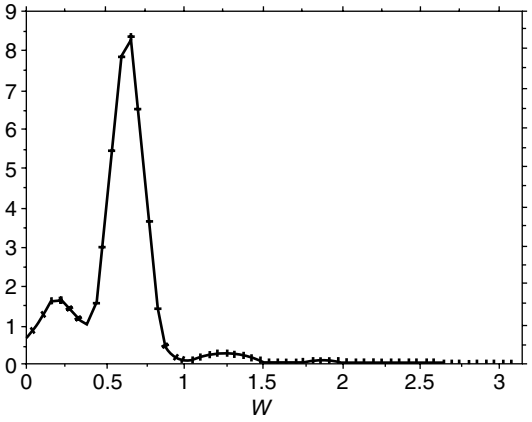

Figure 10 Nonparametric estimate of log lynx spectrum

A relatively simple approach is to use the **Yule–Walker equations** and solve to obtain estimates for the model

$$X_t + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \varphi_3 X_{t-3} + \cdots + \varphi_p X_{t-p} = \varepsilon_t$$

and then to estimate σ^2 using the following equation:

$$\gamma(0) + \varphi_1\gamma(1) + \varphi_2\gamma(2) + \varphi_3\gamma(3) + \cdots + \varphi_p\gamma(p) = \sigma^2.$$

The log of the lynx series was chosen as an example (see Figure 9) and the spectrum is given in Figure 10. The least squares estimates gave a model of the form

$$\begin{aligned} X_t &- 1.139X_{t-1} + 0.508X_{t-2} - 0.213X_{t-3} \\ &+ 0.270X_{t-4} - 0.113X_{t-5} + 0.124X_{t-6} \\ &- 0.068X_{t-7} + 0.040X_{t-8} - 0.134X_{t-9} \\ &- 0.185X_{t-10} + 0.311X_{t-11} = \varepsilon_t, \end{aligned}$$

with residual variance 0.226. The resulting spectrum is then

$$\begin{aligned} h(\omega) = \frac{0.226}{2\pi}|&1 - 1.139\exp(-i\omega) \\ &+ 0.508\exp(-i2\omega) - 0.213\exp(-i3\omega) \\ &+ \cdots + 0.311\exp(-i11\omega)|^{-2}. \end{aligned}$$

The contrast with the nonparametric estimate is characteristic of the method, see Figures 9 and 10.

While least squares is commonly used to estimate the parameters, a variant known as "maximum entropy" spectral estimation due to Burg [7] is popular: see also [2] and [3] for further details.

Multiple Series

We can extend our spectral techniques to more than one series and, as in the time domain, we get a rich and interesting theory. The drawback is that the complexity increases.

For simplicity we concentrate on the bivariate case, that is with a pair of stationary series X_t, Y_t. This is often written as a vector of the form $\mathbf{X}_t = (X_t, Y_t)$. We assume a zero mean and covariance matrix:

$$\mathbf{C}(k) = \begin{bmatrix} \mathrm{E}[X_t X_{t+k}] & \mathrm{E}[X_t Y_{t+k}] \\ \mathrm{E}[Y_t X_{t+k}] & \mathrm{E}[Y_t Y_{t+k}] \end{bmatrix} = \begin{bmatrix} \gamma_{xx}(k) & \gamma_{xy}(k) \\ \gamma_{yx}(k) & \gamma_{yy}(k) \end{bmatrix}.$$

We also define the *spectral density matrix* $\mathbf{H}(\omega)$ as

$$\frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \mathbf{C}(k) \exp(-ik\omega) = \begin{bmatrix} h_{xx}(\omega) & h_{xy}(\omega) \\ h_{yx}(\omega) & h_{xy}(\omega) \end{bmatrix}.$$

Here $\gamma_{xx}(k)$ and $\gamma_{yy}(k)$ are the autocovariances of each series while $\gamma_{xy}(k)$ and $\gamma_{yx}(k)$ are called the *cross-covariances*. In the same way, $h_{xx}(\omega)$ and $h_{yy}(\omega)$ are the *univariate* or *autospectra*, while $h_{xy}(\omega)$ is the *cross-spectrum*.

From the definition

$$h_{xy}(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_{xy}(\omega) \exp(-ik\omega)$$

$$= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_{yx}(-\omega) \exp(-ik\omega) = \overline{h_{yx}(\omega)}$$

since $\gamma_{xy}(k) = \mathrm{E}[X_t,\ Y_{t+k}] = \mathrm{E}[Y_{r-k}, X_r] = \gamma_{yx}(-k)$, so $h_{xy}(\omega)$ and $h_{yx}(\omega)$ are complex conjugates. Because we are dealing with a complex valued quantity, $h_{xy}(\omega)$, it is best to work in one of the traditional representations of complex numbers, either

1. $h_{xy}(\omega) = c_{xy}(\omega) - iq_{xy}(\omega)$, where c_{xy} is known as the *co-spectrum* and q_{xy} as the *quadrature spectrum*, or
2. The alternative polar form $h_{xy}(\omega) = \alpha_{xy}(\omega) \exp[i\phi_{xy}(\omega)]$, where $\alpha_{xy}(\omega)$ is the *amplitude spectrum* and $\phi_{xy}(\omega)$ the *phase spectrum*.

Most people find it useful to work with *coherency spectrum*, or *coherency*,

$$c(\omega) = \frac{|h_{xy}(\omega)|}{[h_{xx}(\omega) h_{yy}(\omega)]^{1/2}}$$

and the *gain*

$$G_{xy}(\omega) = \left| \frac{h_{xy}(\omega)}{h_{xx}(\omega)} \right|.$$

We use the cross-spectrum in its various guises, usually the coherency or its modulus and gain, to understand the relationship between series.

The modulus of the coherence measures the strength of the relationships between corresponding frequency components of the two series in almost exactly the same way as a correlation coefficient. The gain is the analog of the regression of the frequency ω component of the first series on the second. The lead or lag of this relationship is measured by the slope of the phase.

We can show that the coherence is unchanged under linear transformations. If $\mathbf{Z}_t$ is a filtered version of $\mathbf{X}_t$, say

$$\mathbf{Z}_t = \begin{pmatrix} a_{11}(B) & a_{12}(B) \\ a_{21}(B) & a_{22}(B) \end{pmatrix} \mathbf{X}_t,$$

then the coherency does not involve any of the filter functions $a_{ij}(z)$.

For a rather simpler case, suppose $X_t = \beta Y_{t-d} + \varepsilon_t$. Then $\phi_{xy}(\omega) = -\omega d$. This illustrates an important point: when there is a time delay the phase spectrum is a linear function of frequency with the slope representing the size of the delay. If we go further and assume that the ε_t process is uncorrelated with the Y_t series, then

$$\gamma_{yx}(k) = \mathrm{E}[Y_t X_{t+k}] = \mathrm{E}[Y_t\{\beta Y_{t+k} + \varepsilon_{t+k}\}]$$

$$= \beta \gamma_{yy}(k) \quad \text{for } k \neq 0,$$

so $h_{xy}(\omega) = \beta h_{yy}(\omega)$, while $h_{xx}(\omega) = \beta^2 h_{yy}(\omega) + \sigma^2/2\pi$, and hence the coherency is $1/\{1 + \sigma^2/(2\pi\beta)\}^{1/2}$. As we might expect, this decreases as the variance of the added noise increases.

One important application to linear relationships with extra noise is a slight extension of the above.

Suppose we have $Y_t = \sum_{S-\infty}^{\infty} g_s X_{t-s} + \eta_t$. Then we have $h_{yx} = G_{yx}(\omega)h_{xx}(\omega)$, where G_{yx} is the gain defined above. In addition, if X_t and the noise series are uncorrelated, then

$$h_{yy}(\omega) = |G_{yx}(\omega)|^2 h_{xx}(\omega) + h_{\eta\eta}(\omega).$$

Now $\phi_{yx}(\omega) = \tan^{-1}\{h_{yx}(\omega)\}$, while the gain $|h_{yx}(\omega)|$ is $|G_{yx}(\omega)|h_{xx}(\omega)$. Thus the transfer function, *complete with the gain and the phase information*, can be computed from the spectral matrix.

As one might expect, there are rather more problems involved in estimating the spectral matrix than in estimating the individual spectra. Another problem, the alignment of the series, arises because the cross-covariance is not necessarily an even function and hence its maximum need not occur at the zero lag. For details of a suitable procedure, see [14].

Evolutionary Spectra

We have only considered stationary series when dealing with spectra. The extension of the idea to types of nonstationary processes is possible, but raises some interesting problems. Suppose we have a nonstationary process and we consider a representation of the form

$$X_t = \int_{-\pi}^{\pi} \exp(i\omega t) A_t(\omega)\, \mathrm{d}Z(\omega),$$

where for each value of ω the sequence of functions $\{A_t(\omega)\}$ has a (generalized) Fourier transform. The evolutionary spectral density function $h_t(\omega)$ is defined as

$$h_t(\omega)\, \mathrm{d}\omega = |A_t(\omega)|^2 \mathrm{E}[|\, \mathrm{d}Z(\omega)|^2], \quad -\pi < \omega < \pi.$$

The evolutionary spectrum gives the decomposition of total power in the neighborhood of time point t.

To reduce the possibilities for the representations of the series, Priestley suggests that $A_t(\omega)$ should be a slowly changing function of time with a Fourier transform which is concentrated about zero. This leads to the concept of a semistationary series, which is one for which such a representation exists.

The estimation of evolutionary spectra is a two-stage process. First the series is filtered using a filter

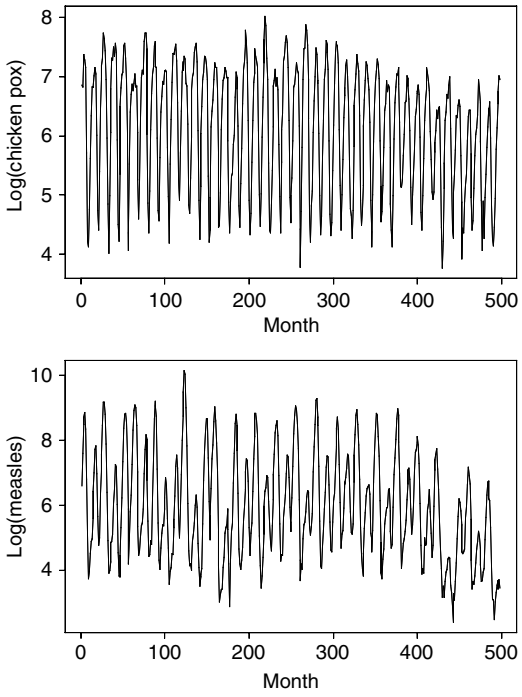

Figure 11 Log series

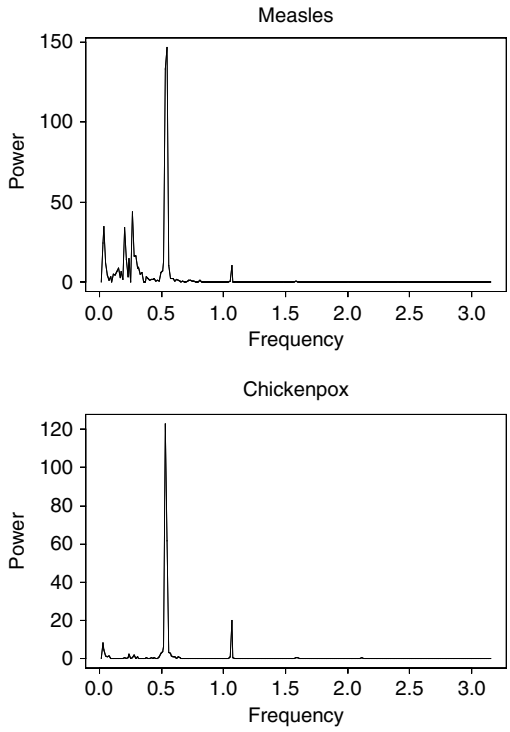

Figure 12 Periodograms

centered on a frequency ω_0, then the output of the filter is averaged in the neighborhood of time t to give the appropriate estimate. A good exposition is given by Priestley [15] who has been responsible for much of the development.

We use the techniques we considered above on two series, the reported monthly cases of measles and the reported monthly cases of chickenpox (1931–1972) in New York City. In fact, both of the observations are fairly skewed and we shall work with the log of both series.

There is a fairly obvious annual cycle in both series as can be seen from series plots and the periodograms; see Figures 11 and 12.

If we filter out a 12-month cycle in each series both spectra have spikes indicating power at period 240 months.

This seems quite curious, we almost include cross-correlations with the sunspot series! What we have done is to look at the cross spectrum between the two series; see Figure 13. As we expect at high frequencies (short periods) the high coherence indicates that the two series are very similar, they are both like noise.

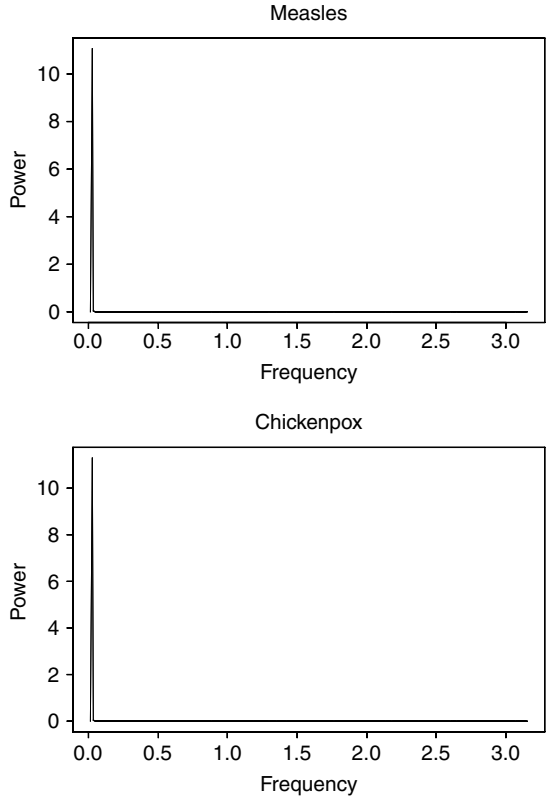

Figure 13 Spectra of smoothed series

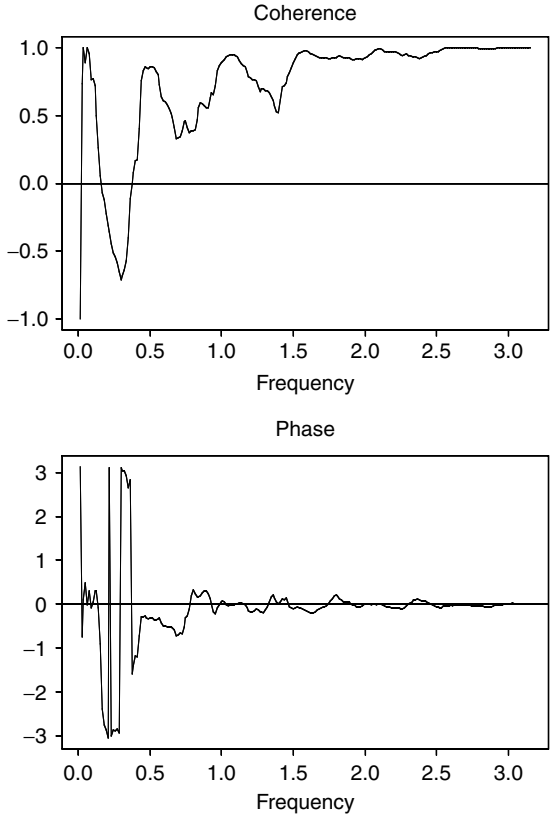

Figure 14 Coherency and phase

However, there is an interesting long period effect in the coherence with a corresponding phase change; see Figure 14. This probably reflects the fact that the series are very similar in the long term, this is the common 10-year effect while in the short term they both look random.

References

[1] Akaike, H. (1969). Fitting autoregressive models for prediction, *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

[2] Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.

[3] Beamish, N. & Priestley, M.B. (1981). A study of autoregressive estimation, *Applied Statistics* **30**, 41–58.

[4] Bloomfield, P. (1976). *The Fourier Analysis of Time Series*. Wiley, New York.
[5] Brockwell, P.J. & Davies, R.A. (1990). *Time Series: Theory and Methods*. Springer-Verlag, New York.
[6] Buijs Ballot, C.H. (1847). *Les Changements Periodiques de Temperature*. Utrecht.
[7] Burg, J.P. (1968). Maximum entropy spectral analysis, reprinted in *Modern Spectral Analysis*, D.G. Childers, ed. IEEE Press, New York.
[8] Diggle, P. (1990). *Times Series: A Biostatistical Introduction*. Oxford University Press, Oxford.
[9] Janacek, G. & Swift, A. (1993). *Time Series: Forecasting, Simulation, Applications*. Ellis Horwood, Chichester.
[10] Jones, R. (1974). Autoregressive order determination, *Geophysics* **41**, 771–773.
[11] Koopmans, L.H. (1974). *The Spectral Analysis of Time Series*. Academic Press, New York.
[12] Parzen, E. (1974). Some recent advances in time series modeling, *IEEE Transactions on Automatic Control* **19**, 723–730.
[13] Percival, D.B. & Walden, A.T. (1993). *Spectral Analysis for Physical Applications*. Cambridge University Press, Cambridge.
[14] Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Academic Press, New York.
[15] Priestley, M.B. (1988). *Non-linear and Non-stationary Time Series Analysis*. Academic Press, New York.
[16] Proakis, J.G. & Manolakis, G.M. (1988). *Introduction to Digital Signal Processing*. Macmillan, New York.
[17] Thompson, D. (1982). Spectrum estimation and harmonic analysis, *Proceedings of the IEEE* **70**, 1055–1096.
[18] Waldren, A.T. (2000)A unified view of multitaper multivariate estimation,. *Biometrika*, **84**, 767–788.

(*See also* **Coherence Between Time Series**; **Multiple Time Series**; **Nonlinear Time Series Analysis**)

G.J. JANACEK

Sphericity Test

The equal-density contours of a p-variate normal distribution (*see* **Multivariate Normal Distribution**) with **covariance matrix** $\mathbf{\Sigma}$ reduce to hyperspheres in the p-dimensional real space $\mathbb{R}^p$ with the mean vector $\boldsymbol{\mu}$ as the center, when $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$. The most common test of the sphericity hypothesis $\mathrm{H}_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}_p$ against $\mathrm{H}_1 : \mathbf{\Sigma} \neq \sigma^2\mathbf{I}_p$, where $\boldsymbol{\mu}$ and σ^2 are unknown, is the **likelihood ratio test**, based on a random sample of size N from the p-variate normal distribution $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, which rejects H_0 when

$$\lambda \equiv \frac{|\mathbf{S}|^{N/2}}{(\mathrm{tr}(\mathbf{S})/p)^{Np/2}}$$

is too small, where $\mathbf{S}$ is the sample covariance matrix. Here $\mathrm{tr}(\mathbf{S})$ means the trace of the matrix $\mathbf{S}$. This test was proposed by Mauchly in 1940 [11].

For $p = 2$, the distribution of $W \equiv \lambda^{2/N}$ is **beta** $((n-1)/2, 1)$, where $n = N - 1$; see Mauchly [11] and Anderson [1]. Steffens [16] has suggested a **Student's *t* test** for this case. The exact null distributions of W for $p = 3$, 4, and 6 are obtained by Consul [3]; see also Mathai & Rathie [10], and John [6]. The exact null distribution of W in a series form has been derived by Nagarsenker & Pillai [13] for general p. The asymptotic expansion of the distribution of $n\rho \log W$ under H_0, where $\rho = 1 - (2p^2 + p + 2)/6pn$, is given in Anderson [1]; the first term of this expansion is the **chi-square distribution** with $p(p+1)/2 - 1$ **degrees of freedom** (df).

The nonnull distribution of W is obtained by Girshick [4] in the two-root case, and by Pillai & Nagarsenker [14] and Khatri & Srivastava [7] in the general case. The asymptotic distribution of W under H_1 is given by Sugiura [17] and Gleser [5].

Kiefer & Schwarz [8] have shown that the above test is Bayes and admissible (*see* **Decision Theory**). Sugiura & Nagao [19] have proved its **unbiasedness**. A special monotonicity property of its **power** function is obtained by Carter & Srivastava [2].

The locally best invariant test of H_0 against H_1 rejects H_0 if $T = \mathrm{tr}(\mathbf{S}^2)/(\mathrm{tr}\ \mathbf{S})^2$ is too large; see Sugiura [18] and John [6]. The asymptotic expansion of the null distribution of T is given by Nagao [12]. Sugiura [18] has obtained the asymptotic distribution of T under H_1.

The **union–intersection principle** of Roy leads to a test that rejects H_0 if $(l_1 + l_p)^2/4l_1l_p$ is too large, where l_1 and l_p are the extreme roots of $\mathbf{S}$; see Srivastava & Khatri [15]. Percentage points for the null distribution of l_1/l_p are given by Krishnaiah & Schuurmann [9] for some values of the parameters.

The sphericity test discussed above can be used to test that the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$ has the following form:

$$\sigma_{ij} = \begin{cases} \sigma^2, & \text{for } i = j, \\ \rho\sigma^2, & \text{for } i \neq j. \end{cases}$$

Such a problem arises in a repeated measure analysis (*see* **Longitudinal Data Analysis, Overview**). To use the sphericity test, note that the above structure of $\mathbf{\Sigma}$ is equivalent to the following: $\mathbf{C}'\mathbf{\Sigma}\mathbf{C} = (1-\rho)\sigma^2\mathbf{I}_{p-1}$, where $\mathbf{C}$ is a $p \times (p-1)$ matrix, the columns of which form an orthonormal basis of the linear space orthogonal to the linear space spanned by the unit vector $\mathbf{1}$ (*see* **Orthogonality**).

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, New York.

[2] Carter, E.M. & Srivastava, M.S. (1977). Monotonicity of the power functions of the modified likelihood ratio criterion for the homogeneity of variances and of the sphericity test, *Journal of Multivariate Analysis* **7**, 229–233.

[3] Consul, P.C. (1967). On the exact distribution of likelihood ratio criteria for different hypotheses, in *Multivariate Analysis*, II, P.R. Krishnaiah, ed. Academic Press, New York.

[4] Girshick, M.A. (1941). The distribution of the ellipticity statistic L_e when the hypothesis is false, *Terrestrial Magnetism and Atmospheric Electricity* **46**, 455–457.

[5] Gleser, L.J. (1966). A note on the sphericity test, *Annals of Mathematical Statistics* **37**, 464–467. Correction: **39** (1968) 684.

[6] John, S. (1972). The distribution of a statistic used for testing sphericity of normal distribution, *Biometrika* **59**, 169–173.

[7] Khatri, C.E. & Srivastava, M.S. (1971). On exact non-null distribution of likelihood-ratio criteria for sphericity test and equality of two covariance matrices, *Sankhyā, Series A* **33**, 201–206.

[8] Kiefer, J. & Schwarz, R. (1965). Admissible Bayes character of $T^2 =$, $R^2 =$, and other fully invariant tests for classical multivariate normal problems, *Annals of Mathematical Statistics* **36**, 747–770.

[9] Krishnaiah, P.R. & Schuurmann, F.J. (1974). On the evaluation of some distributions that arise in simultaneous tests for the equality of the latent roots of the

covariance matrix, *Journal of Multivariate Analysis* **4**, 265–282.

[10] Mathai, A.M. & Rathie, P.N. (1970). The exact distribution for the sphericity test, *Journal of Statistical Research* **4**, 140–159.

[11] Mauchly, J.W. (1940). Significance test for sphericity of a normal n-variate distribution, *Annals of Mathematical Statistics* **11**, 204–209.

[12] Nagao, H. (1973). On some test criteria for covariance matrix, *Annals of Statistics* **1**, 700–709.

[13] Nagarsenker, B.N. & Pillai, K.C.S. (1973). The distribution of the sphericity test criterion, *Journal of Multivariate Analysis* **3**, 226–235.

[14] Pillai, K.C.S. & Nagarsenker, B.A. (1972). On the distributions of a class of statistics in multivariate analysis, *Journal of Multivariate Analysis* **2**, 96–114.

[15] Srivastava, M.S. & Khatri, C.G. (1979). *An Introduction to Multivariate Statistics*. North-Holland, New York.

[16] Steffens, F.E. (1974). A bivariate test for sphericity, *South African Statistical Journal* **8**, 59–68.

[17] Sugiura, N. (1969). Asymptotic expansions of the distributions of the likelihood ratio criteria for covariance matrices, *Annals of Mathematical Statistics* **40**, 2051–2063.

[18] Sugiura, N. (1972). Locally best invariant test for sphericity and the limiting distribution, *Annals of Mathematical Statistics* **43**, 1312–1316.

[19] Sugiura, N. & Nagao, H. (1968). Unbiasedness of some test criteria for the equality of one or two covariance matrices, *Annals of Mathematical Statistics* **39**, 1686–1692.

(*See also* **Multivariate Analysis of Variance**; **Multivariate Analysis, Overview**)

SOMESH DAS GUPTA

Spiegelman, Mortimer

Born: December 10, 1901, in Brooklyn, New York.
Died: March 25, 1969, in New York.

Mortimer Spiegelman was an important contributor to biostatistics, particularly in the areas of **demography** and public health. His major contribution to the field of public health and epidemiology came toward the end of his career when he conceived of, coordinated, edited, and carried to a successful conclusion the publication of a series of monographs sponsored by the **American Public Health Association** (APHA) and published by the Harvard University Press. Each monograph pertained to a specific set of diseases in which the 1960 **Census** was used in a standard way as the denominator for **rates** of disease. In his role as editor of this series he used his considerable powers of persuasion with the authors of the monographs to ensure comparability among them and to make certain that the work on each was completed. Sixteen monographs resulted from this effort, covering a wide range of topics as evidenced by the following titles: *Accidents and Homicides; Infectious Diseases; Trends and Variations in Fertility in the U.S.; Infant, Perinatal, Maternal, and Childhood Mortality; The Epidemiology of Oral Health; Tuberculosis; Syphilis and Other Venereal Diseases; Cardiovascular Diseases in the U.S.; The Frequency of Rheumatic Diseases; Digestive Diseases; Mental Disorders and Suicide; Cancer in the U.S.; The Epidemiology of Neurological and Sense Organ Diseases; Mortality and Morbidity in the U.S.; Differential Mortality in the U.S.*

In 1970, the Mortimer Spiegelman Gold Medal Award was established by his family and has been presented annually by the Statistics Section of the APHA to a young statistician (under 40 years of age) who has made important contributions to the field of health statistics. This has been, from the beginning, a coveted award and the list of awardees is most impressive. Many of the awardees are now heads of departments of biostatistics, deans or associate deans of schools of public health, and heads of large statistical agencies. A list of awardees from 1970 to 2001 follows:

1970 Edward Perrin
1971 P.A. Lachenbruch
1972 Manning Feinleib
1973 Joseph Fleiss
1974 Gary Koch
1975 Jane Menken
1976 A.A. Afifi
1977 David Hoel
1978 Ross Prentice
1979 Mitchell Gail
1980 Norman Breslow
1981 Robert F. Woolson
1982 Joel Kleinman
1983 J. Richard Landis
1984 Stephen Lagakos
1985 John Crowley
1986 Anastasios Tsiatis
1987 L.J. Wei
1988 Thomas Fleming
1989 Colin Begg
1990 Kung-Yee Liang
1991 Scott Zeger
1992 Ronald Brookmeyer
1993 Martin Tanner
1994 Louise Ryan
1995 Christopher Portier
1996 Jeremy Taylor
1997 Margaret Pepe
1998 Peter Bacchetti
1999 Danyu Lin
2000 Bradley Carllin
2001 Daniel Weeks

Mr Spiegelman was a native of Brooklyn, New York, and received a masters of engineering degree from the Polytechnic Institute of Brooklyn in 1923 and a masters of business administration degree from Harvard University in 1925. He spent 40 years on the staff of the Metropolitan Life Insurance Company where he published many articles and volumes that attained national and international recognition. He coauthored with Dublin and Lotka *The Money Value of Man* and *Length of Life*, both of which have been standard reference volumes. Although his employment was in an organization that was concerned primarily with **actuarial** science, his interests were much broader. He published two editions of *Introduction to Demography*, which has been a standard text in demography. The second edition, in particular, is oriented toward the general demographer and student of public health statistics rather than toward the actuary. He did extensive work on

life tables including what he referred to as "segmented generation" mortality. This approach allows one to follow the mortality experience of a given age group over successive 10-year periods as an alternative to analyzing trends in the current mortality. His development of the APHA monograph series further illustrates the breadth of his interests. Mr Spiegelman was a Fellow of the Society of Actuaries, Fellow of the **American Statistical Association**, and Fellow of the American Public Health Association. Each year, upon the presentation of the Mortimer Spiegelman Award, he is remembered again for his extraordinary contributions to public health statistics.

EARL POLLACK

Spline Function

Suppose we have n observations (t_j, y_j), $j = 1, 2, \ldots, n$, where y_j and t_j are related by the model

$$y_j = \mu(t_j) + e_j, \tag{1}$$

where $\{e_j\}$ are zero mean uncorrelated random variables with common variance σ^2. Here $\mu(t)$ is usually known as the **regression** function. If $\mu(t) = \sum_{j=1}^{p} \beta_j x_j(t)$, where $x_1(t), x_2(t), \ldots, x_p(t)$ are known functions, then the estimation of $(\beta_1, \beta_2, \ldots, \beta_p)$ given $(y_1, y_2, \ldots, y_n)$ is a classic **multiple regression** problem. This is known as a parametric regression problem. In the **nonparametric regression** context one assumes that $\mu(t)$ belongs to some infinite dimensional collection of functions. For example, μ may be differentiable with square integrable second derivatives. One can estimate $\mu(t)$ by considering the estimators of the form

$$\mu_\lambda(t) = \sum_{j=1}^{n} K(t, t_j, \lambda) y_j, \tag{2}$$

where $K(t, t_j; \lambda)$, $j = 1, 2, \ldots, n$, is a collection of weight functions. The weights are derived from a single function $K(\cdot)$ that is independent of the design. These are called *kernel estimators*. A detailed discussion of these estimators can be found in the book by Eubank [1, chapter 4]. An alternative approach is the spline approach.

Smoothing splines are related to **polynomial regression**. Consider model (1), and assume $a \le t_1 \le \cdots \le t_n \le b$, and $\mu(t)$ belongs to $W_2^{(m)}[a, b]$. Hence $W_2^{(m)}[a, b]$ is the set of all functions on $[a, b]$, where the jth derivative, $\mu^j(t)$, $j = 0, 1, 2, \ldots, m-1$, is absolutely continuous, and $\mu^m(t) \in L_2[a, b]$. One can expand $\mu(t)$ in the form

$$\mu(t) = \sum_{j=0}^{m-1} \theta_j t^j + \mathrm{Rem}(t), \tag{3}$$

where

$$\mathrm{Rem}(t) = [(m-1)!]^{-1} \int_a^b \mu^{(m)}(x)(t-x)_+^{m-1}\, \mathrm{d}x, \tag{4}$$

$(x)_+ = \max\{0, x\}$. If $\mathrm{Rem}(t)$ can be neglected, then estimating $\mu(t)$ reduces to the estimation of the coefficients $\{\theta_j\}$ of the polynomial $\sum_{j=0}^{m-1} \theta_j t^j$. One can show

$$\sup_{t \in [a,b]} |\mathrm{Rem}(t)| \le c[J_m(\mu)]^{1/2},$$

where c depends on m but not on μ. Here $J_m(\mu) = \int_a^b [\mu^{(m)}(t)]^2\, \mathrm{d}t$. We can now find an estimate μ by minimizing $(1/n)\sum_{j=1}^{n} [y_j - f(t_j)]^2$ subject to the condition that $J_m(\mu) \le \rho$ for some $\rho \ge 0$. Here $f \in W_2^{(m)}\{a, b\}$. This is essentially equivalent to estimating μ by minimizing

$$\frac{1}{n}\sum_{j=1}^{n} [y_j - f(t_j)]^2 + \lambda \int_a^b [f^{(m)}(t)]^2\, \mathrm{d}t \tag{5}$$

over $f \in W_2^{(m)}[a, b]$. This is also called the *roughness penalty approach* (see [2]) and $\lambda > 0$, is called the *smoothness parameter*. The function f that minimizes (5) is unique and is a natural spline, and it is clear that splines are related to polynomials. Splines are defined as piecewise polynomials subject to a maximum number of continuity constraints.

A spline of order r with knots of $\zeta_1, \zeta_2, \ldots, \zeta_k$ is defined to be any function

$$g(t) = \sum_{i=0}^{r-1} \theta_i t^i + \sum_{i=1}^{k} \delta_i (t - \zeta_i)_+^{r-1}. \tag{6}$$

This definition of a spline is equivalent to the following specifications:

1. $g(t)$ is a piecewise polynomial of order r on any subinterval $[\zeta_i, \zeta_{i+1}]$;
2. $g(t)$ has $r-2$ continuous derivatives; and
3. $g(t)$ has an $(r-1)$th derivative that is a step function with jumps at $\zeta_1, \zeta_2, \ldots, \zeta_k$.

If we impose further restrictions on $g(t)$ we would arrive at natural splines. A natural spline of order $r = 2m$ with $k = n$ knots at $(\zeta_1, \zeta_2, \ldots, \zeta_n)$ is defined as a polynomial $g(t)$ which satisfies conditions 1–3 and further satisfies the extra condition, namely:

4. $g(t)$ is a polynomial of order m outside $[t_1, t_n]$.

We can define $S^r(t_1, t_2, \ldots, t_k)$ as the set of all functions of the form (6) with knots $(\zeta_1, \zeta_2, \ldots, \zeta_k)$. We note that $(1, t, t^2, \ldots, t^{n-1}, (t-\zeta_1)_+^{n-1}, \ldots (t-\zeta_k)_+^{r-1})$ form the basis of this space. We can denote the collection of all natural splines with knots at

$(t_1, t_2, \ldots, t_n)$ by $S^{2m}(t_1, t_2, \ldots, t_n)$ and this is a subspace of $S^{2m}(t_1, t_2, \ldots, t_n)$.

Let $(x_1(t), x_2(t), \ldots, x_n(t))$ be the basis of $S_1^{2m}(t_1, t_2, \ldots, t_n)$. Then one can write

$$x_j(t) = \sum_{i=0}^{m-1} \theta_{ij} t^i + \sum_{i=1}^{n} \delta_{ij} (t - t_i)_+^{2m-1}, \qquad (7)$$

and the minimizer f that minimizes (5) is of the form [1, p. 205]

$$f_\lambda(t) = \sum_{j=1}^{n} \beta_{\lambda j} x_j(t), \qquad (8)$$

where $\beta_\lambda' = (\beta_{\lambda 1}, \beta_{\lambda 2}, \ldots, \beta_{\lambda n})$ and is the solution of

$$(\mathbf{x}'\mathbf{x} + n\lambda \boldsymbol{\Omega})\boldsymbol{\beta}_\lambda = \mathbf{x}'\mathbf{y}, \qquad (9)$$

$\Omega = (\int_a^b x_i^{(m)}(t)\, x_j^{(m)}(t)\, \mathrm{d}t;\ i, j = 1, 2, \ldots, n)$ and $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$. We note that the above solution depends on the smoothness parameter λ, and one can estimate λ by the method of cross validation. We refer to the books of Green & Silverman [2] and Eubank [1] for actual numerical evaluations, and also to several papers by Wahba referred to in the those books.

One interesting application of the above approach is in the context of **time series** [3, 4]. Suppose we have a zero mean second order discrete parameter Gaussian time series $(x_1, x_2, \ldots, x_n)$. Let $R(s) = \mathrm{cov}(x_t, x_{t+s})$ and let

$$f(w) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} R(s) \exp(isw),$$

and let $g(w) = \ln f(w)$. Consider the problem of estimation of $g(w)$. It is well known that the periodogram $I(w)$, where

$$I(w) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} x_t \exp(itw) \right|^2$$

provides an **unbiased** estimator of $f(w)$, but it is not a **consistent** estimator (*see* **Spectral Analysis**). Hence one can estimate $f(w)$ by smoothing $I(w)$ using a set of suitably weighted functions. Alternatively, as done earlier, one can use a spline approach [3, 4].

Let us estimate $g(w)$ at the frequencies $w_j = 2\pi j/2n$, $j = -(n-1), \ldots, 0, 1, 2, \ldots, n$. To a good approximation, we can write $I(j) = I(w_j) = f(w_j)U_j$, where U_j, $j = 1, 2, \ldots, n-1$, are independent, identically distributed as chi-square random variables. Let $y_j = \ln I_j + C_j$, $C_j = C$ be the Euler constant for $j = 1, 2, \ldots, n-1$, and $C_0 = C_n = (\ln 2 + C)/\pi$. Then

$$y_j = g(w_j) + \varepsilon_j, \qquad (10)$$

where $\varepsilon_j = \ln U_j + c_j$. The model (10) is similar to (1), and estimating the logarithm of the spectral density is similar to the estimation of $\mu(t_j)$ in (1). We refer to [3] and [4] for further details.

References

[1] Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel-Dekker, New York.

[2] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models – A Roughness Penalty Approach*. Chapman & Hall, London.

[3] Wahba, G. (1980). Automatic smoothing of the log periodogram, *Journal of the American Statistical Association* **75**, 122–132.

[4] Wahba, G. & Wold, S. (1975). Periodic splines for spectral density estimation. The use of cross validation for determining the degree of smoothness, *Communications in Statistics* **4**, 125–141.

T. SUBBA RAO

Spline Smoothing

Introduction

Assume we have observations (x_i, y_i), $i = 1, \ldots, n$, on a (one-dimensional) regressor variable x and a response variable y, that follow the model $y_i = r(x_i) + \varepsilon_i$, where the ε_i's are i.i.d. errors and $r(\cdot)$ is an unknown **regression** function. This problem is called a **nonparametric regression** problem if we allow the regression function $r(\cdot)$ to belong to some infinite dimensional collection of functions. For example, we may assume that $r(\cdot)$ is differentiable or differentiable with a square integrable derivative, and so on.

Over the past few decades, several methods for estimating $r(\cdot)$ have been proposed and studied intensively. Among these methods are kernel smoothers, **orthogonal** series estimators, **wavelet** smoothers, and spline smoothers. This entry provides a brief description of the latter by discussing regression splines, smoothing splines, and penalized splines.

Most (if not all) regression smoothing methods mentioned in the previous paragraph can be viewed as trying to estimate the response variable via

$$\hat{y}_i = \hat{a}_1 f_1(x) + \hat{a}_2 f_2(x) + \cdots + \hat{a}_k f_k(x), \quad i = 1, \ldots, n, \tag{1}$$

where the $\hat{y}_i$'s are the fitted values, the $f_i(\cdot)$'s are some basis functions (that depend on the smoothing method used) and the $\hat{a}_i$'s are estimated coefficients. The way the $\hat{a}_i$'s are estimated also differs from method to method. Spline smoothing methods typically use either ordinary **least squares** or **ridge regression** to estimate the coefficients; as basis functions $f_i(\cdot)$, they use (piecewise) polynomials. Note that if the coefficients $\hat{a}_i$'s were calculated by ordinary least squares, then k has the interpretation as being the **degrees of freedom** of the fit; otherwise, it would be necessary to resort to approximate degrees of freedom for the fit.

An example of a spline smoother is given in Figure 1 using the follicle data set from [1]. This data set contains information on the number of ovarian follicles counted from sectioned ovaries of women of various ages. Here, age was used as the regressor variable, and the number of ovarian follicles, on a log scale, as the response variable. The figure shows the data with a smoothing spline superimposed; details on how this was done using the **R** software [32] is given at the end of this entry. From this figure, it is clear that there is a nonlinear relationship between these two variables.

Preliminaries

To fix notation, a **spline function** $s(\cdot)$ of order $p + 1$ with knots at $\tau_1 < \tau_2 < \cdots < \tau_m$ is defined to be any function such that (see also [36])

1. is piecewise polynomial of order $p + 1$ on any subinterval $[\tau_i, \tau_{i+1}]$;
2. has $p - 1$ continuous derivatives; and
3. whose pth derivative is a step function with (possible) jumps at $\tau_1, \ldots, \tau_m$.

An important subset of spline functions, called *natural spline functions*, is defined by one further restriction. If $p + 1 = 2q$ is even, then $s(\cdot)$ is a natural spline if, in addition to conditions 1 to 3, it also fulfills the following condition:

4. $s(\cdot)$ is a polynomial of order q outside of $[\tau_1, \tau_m]$.

For illustrative and theoretical purposes, it is often convenient to parameterize spline functions using the

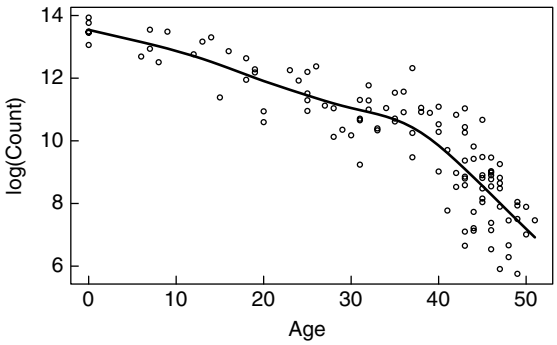

Figure 1 The picture shows a spline smooth fitted to the data on the number of ovarian follicle for women of various ages. The age of the women is depicted on the horizontal axis and the logarithm of the number of ovarian follicles is shown on the vertical axis

truncated power basis:

$$s(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{i=1}^{m} \gamma_i (x - \tau_i)_+^p, \tag{2}$$

where $(x - \tau_i)_+ = \max(0, x - \tau_i)$, and $(x - \tau_i)_+^p = \{\max(0, x - \tau_i)\}^p$.

Any spline function of order $p + 1$ with knots $\tau_1, \ldots, \tau_m$ can be expressed in the form (2). This shows that the space of spline functions of order $p + 1$ with knots at $\tau_1 < \tau_2 < \cdots < \tau_m$ is a $m + p + 1$-dimensional space. Hence, since condition 4 imposes $p + 1 = 2q$ constraints, the space of natural spline functions of order $p + 1$ with m given knots is an m-dimensional space.

For practical purposes, it is usually preferable to use other basis functions that have better numerical properties, for example, B-splines (see, among others, [4] and [7]) or the Demmler–Reinsch basis functions (see, among others, [10]). Basis function such as B-splines also provide an easier parameterization of natural splines. To represent natural splines using (2), one would have to impose some awkward constraints on the parameters $\beta_0, \ldots, \beta_p, \gamma_1, \ldots, \gamma_m$.

The next sections discuss some of the main approaches to spline smoothing; more details can be found in the books of Eubank [10], Gu [15], Wahba [48] and some chapters of Schimek [42]. Discussion on spline smoothing techniques can also be found, among others, in [13, 17, 18, 35] and [41], although these books concentrate more on statistical applications of spline smoothing.

Regression Splines

Regression spline smoothing is most conveniently discussed using representation (2). After choosing p (typically $p = 3$) and the locations for the knots, the problem of estimating $r(\cdot)$ reduces to a **multiple linear regression** problem with design matrix,

$$X = \begin{pmatrix} 1 & x_1 & \ldots & x_1^p & (x_1 - \tau_1)_+^p & \ldots & (x_1 - \tau_m)_+^p \\ 1 & x_2 & \ldots & x_2^p & (x_2 - \tau_1)_+^p & \ldots & (x_2 - \tau_m)_+^p \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^p & (x_n - \tau_1)_+^p & \ldots & (x_n - \tau_m)_+^p \end{pmatrix}. \tag{3}$$

Together with the vector $Y = (y_1, \ldots, y_n)'$ of responses, we can calculate estimates $(\hat{\beta}', \hat{\gamma}') = (\hat{\beta}_0, \ldots, \hat{\beta}_p, \hat{\gamma}_1, \ldots, \hat{\gamma}_m)$ using least squares regression:

$$\underset{\beta,\gamma}{\text{minimize}} \left(Y - X \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \right)' \left(Y - X \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \right).$$

To estimate $r(\cdot)$ at an arbitrary point x_0, we just evaluate

$$\hat{r}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \cdots + \hat{\beta}_p x_0^p + \sum_{i=1}^{m} \hat{\gamma}_i (x_0 - \tau_i)_+^p. \tag{4}$$

The advantage of this approach to spline smoothing is that the parameter estimates are least squares estimates and, hence, to study the regression estimate $\hat{r}(\cdot)$, one has the power of multiple linear regression theory available.

Not surprisingly, the biggest difficulty with this approach is the choice of the number of knots and their placement. Choose too many knots and the regression estimate $\hat{r}(\cdot)$ may show spurious features and is too wiggly, that is, the estimate shows too much variability. With too few knots the regression estimate is too restricted and may not be able to detect some important features of the underlying regression function $r(\cdot)$, that is, the estimate has too much bias. For some data sets, empirical evidence indicates that regression spline estimates that differ with respect to the placement of knots but not their number may vary markedly and give different impressions about the underlying regression function.

Thus, much research has focused on how to select the number of knots and where to place these knots. Essentially, two approaches to this problem exist. The first approach tries to choose the knots $\tau_1, \ldots, \tau_m$ from among the observed regressor variables $x_1, \ldots, x_n$. This can be done by traditional **variable selection** tools as proposed by Smith [44] (see also [46]), by **Bayesian** variable selection approaches [6, 43], or by other recently proposed variable selection methods such as the LASSO [28]. The second approach allows the knots $\tau_1, \ldots, \tau_m$ to freely vary within the range of the observed regressor variables; see [24, 25, 31] and the references therein.

Simulation studies that compare regression splines with other smoothing methods and compare some of the different knot selection schemes are reported, among others, in [2] and [49].

Smoothing Splines

Smoothing splines try to estimate $r(\cdot)$ by minimizing the residual sum of squares over a certain space of functions, typically, for technical reasons, a Sobolev space, that is, a space of functions with derivatives up to order q and the integral of the squared qth derivative being finite. Unrestricted minimization, however, would lead to a nonunique estimate, as any function in that space that interpolates the y_i's would be a solution. To avoid this nonuniqueness, a penalty is imposed on the roughness of the regression estimate $\hat{r}(\cdot)$ with the natural roughness measure being $\int\{r^{(q)}(u)\}^2\,\mathrm{d}u$, where $r^{(q)}$ is the qth derivative (*see* **Penalized Maximum Likelihood**).

Thus, smoothing splines are the solution to the following optimization problem:

$$\underset{r(\cdot)}{\text{minimize}} \sum_{i=1}^{n}(y_i - r(x_i))^2 + \lambda \int \{r^{(q)}(u)\}^2\,\mathrm{d}u, \quad (5)$$

where $\lambda > 0$ is given. Here, λ controls the influence of the penalty term and hence the smoothness of the solution of (5).

It can be shown that the solution to (5) is a natural spline of order $p + 1 = 2q$ with n knots and the set of knots equals the set of observed regressor variables. (Here, we assume for simplicity that the x_i's are distinct.) Hence, problem (5) reduces to a finite dimensional minimization problem. Specifically, if $b_1(\cdot), \ldots, b_n(\cdot)$ denotes a basis for the natural splines of order $2q$ with knots at $x_1, \ldots, x_n$ (e.g. the B-splines basis), then we can estimate $r(\cdot)$ at an arbitrary point x_0 as

$$\hat{r}(x_0) = \sum_{i=1}^{n} \hat{\beta}_i b_i(x_0), \quad (6)$$

where in this case, the parameter estimates $\hat{\beta}_1, \ldots, \hat{\beta}_n$ are obtained from

$$\underset{\beta}{\text{minimize}}\,(Y - X\beta)'(Y - X\beta) + \lambda\beta' K\beta$$

with

$$\boldsymbol{X} = \begin{pmatrix} b_1(x_1) & \ldots & b_n(x_1) \\ b_1(x_2) & \ldots & b_n(x_2) \\ \vdots & \vdots & \vdots \\ b_1(x_n) & \ldots & b_n(x_n) \end{pmatrix}$$

and

$$K = \left(\int b_i^{(q)}(u) b_j^{(q)}(u)\,\mathrm{d}u \right)_{i,j=1}^{n}. \quad (7)$$

Thus, we see that smoothing splines are essentially ridge regression estimators. The theoretical and statistical properties of smoothing splines are discussed in depth in [10, 11, 48], and the references given therein.

In (5), the smoothing parameter λ controls the balance between the residual sum of squares and the roughness penalty. If λ is small, the residual sum of squares term dominates (5) and the regression estimate will be rough and wiggly, that is, very variable, and will nearly interpolate the y_i's. However, for a large λ, the penalty term will dominate (5) and the regression estimate tends, in the limit $\lambda \to \infty$, toward a polynomial of order q.

Hence, to calculate a smoothing spline for some given data, one has to choose q and λ. A popular choice is $q = 2$ leading to cubic smoothing splines. Once q and λ are chosen, fast and numerically stable algorithms exist to calculate the solution of (5); see [21, 22, 37, 38]. The case $q = 2$ is also discussed in [13].

From a practical point of view, the choice of λ is much more crucial since this parameter controls the smoothness of the estimate. A popular way to choose λ using the data is via generalized **cross-validation**, originally proposed in [3]. A recent discussion on smoothing-parameter selection for smoothing splines is given in [47], and a simulation study comparing different methods for selecting the smoothing parameter can be found in [23].

Penalized Splines

While regression splines use only a few knots (whose placement is important), smoothing splines use a large number of knots (typically n) and achieve smoothness of the regression estimate by imposing a roughness penalty. A middle way, now popularized under the name "penalized splines", between these two extremes was suggested in [29, 30]; but see also [9] and [40].

These approaches use a moderate number of knots, more than a regression spline would use but less than smoothing splines. With such a number of knots, a regression spline estimate would be too rough and, hence, these approaches also incorporate a roughness

penalty. Typically, this roughness penalty is discrete and not continuous as in the case of smoothing splines. For example, [9] uses a B-spline basis with equidistant knots and penalizes the sum of squared higher-order finite differences of the coefficients of adjacent B-splines.

Here, we describe in more detail the approach given in [40], which is best explained using the truncated power basis (2). The knots are selected from the observed x_i's and details about this are given below. The design matrix X is the same as in the section "Regression Splines" and the response vector is Y. In addition, we require a smoothing parameter λ whose value [40] is allowed to vary from knot to knot to achieve spatial adaptiveness. Here, for the sake of simplicity, we describe a simpler version. Ruppert and Carroll [40] suggest the use of estimates $(\hat{\beta}', \hat{\gamma}') = (\hat{\beta}_0, \ldots, \hat{\beta}_p, \hat{\gamma}_1, \ldots, \hat{\gamma}_m)$ that solve the equation

$$\underset{\beta,\gamma}{\text{minimize}} \left(Y - X\begin{pmatrix}\beta\\ \gamma\end{pmatrix}\right)' \left(Y - X\begin{pmatrix}\beta\\ \gamma\end{pmatrix}\right) + \lambda\gamma'\gamma.$$

To estimate $r(\cdot)$ at an arbitrary point x_0, we just evaluate

$$\hat{r}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \cdots + \hat{\beta}_p x_0^p + \sum_{i=1}^{m} \hat{\gamma}_i (x_0 - \tau_i)_+^p. \tag{8}$$

This shows that, like smoothing splines, penalized splines are essentially ridge regression estimators. To solve this ridge regression problem in a fast and numerical stable manner, the algorithms described in [39] and [50] can be used. The theoretical and statistical properties of penalized splines are discussed in depth in the references given above. It can be shown that the approaches in [9] and [40] are identical if the x_i's are equidistant.

A suggestion is to choose the number of knots m to be in the range 5 to 40 and to set τ_i to the $i/(m+1)$th sample **quantile** of the unique x_i's; see [40]. Another suggestion is to use $m = \min(n/4, 35)$ knots and to choose τ_i as the $(i+1)/(m+2)$th sample quantile of the unique x_i's. The question on how to choose the number of knots is further explored in [39]. However, it seems that, as long as m is large enough, the number and placement of the knots are not as important as the choice of the smoothing parameter λ.

As for smoothing splines, λ is usually chosen by optimizing some criterion such as, say, generalized cross-validation. However, it seems that there is an intrinsic relation between mixed models and smoothing procedures, as first noted in [45]. For penalized splines, this relationship is thoroughly explored in [41]. This approach relates the smoothing parameter λ to the variance of certain random effects in a mixed model and, thus, leads to a new way for automatically choosing λ, which seems promising. Details of this approach are discussed in [41] and the references given therein.

Extensions and Generalizations

We note further extensions and generalizations of spline functions. The ease with which spline functions can be extended and adapted to more complicated settings is one of the reasons why they are popular for regression smoothing.

Pseudo splines, proposed in [16] and which have not been discussed here, are related to penalized splines.

Splines can be used in multivariate smoothing problems where one would want to estimate a regression function of several parameters. Here, essentially two approaches exist – tensor splines, and thin-plate splines. Details can be found, among others, in [12, 14, 27].

Splines can be used as building blocks for more complicated models, for example, **(generalized) additive models** [17] or **semiparametric** models [41]. They can also be used to extend well-known models such as **generalized linear models**; see [13]. However, these more complicated models typically use iterative methods to calculate the final estimates and care has to be taken with the choice of convergence criteria. Problems that can arise if these convergence criteria are not stringent enough are discussed in [8] which compares three different approaches that use spline smoothing, in an ongoing study that has major implications for public health decision making.

Finally, it is also straightforward to incorporate qualitative constraints into the regression estimate, such as monotonicity (see, among others, [19] or [33]). Constraints can be incorporated in essentially two ways. Either one restricts the function space over which (5) is minimized to a suitable function space that contains only functions that have the desired

property (see, among others, [5] and [26]), or one can modify the penalty term in (5) such that the penalty term enforces the desired property onto the regression estimate (see [20] and [34]).

Notes

The figure shown in the introductory section was produced using R [32] by the following code:

```
library(sm)
provide.data(follicle)
plot(Age, log(Count))
follicle.spl <- smooth.spline(Age,
                              log(Count))
lines(follicle.spl)
```

To run this code snippet, the R installation must include the additional package `sm`.

References

[1] Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.

[2] Breiman, L. & Peters, S. (1992). Comparing automatic smoothers (a public service enterprise), *International Statistical Review* **60**(3), 271–290.

[3] Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik* **31**, 377–403.

[4] DeBoor, C. (1978). *A Practical Guide to Splines*, Volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin/Heidelberg.

[5] Delecroix, M. & Thomas-Agnan, C. (2000). Spline and kernel regression under shape restrictions, in *Smoothing and Regression. Approaches, Computation and Application*. M.G. Schimek, ed. John Wiley & Sons, New York, pp. 109–133.

[6] Denison, D.G.T., Mallick, B.K. & Smith, A.F.M. (1998). Automatic Bayesian curve fitting, *Journal of the Royal Statistical Society, Series B* **60**(2), 333–350.

[7] Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Monographs on Numerical Analysis. Oxford University Press, Oxford.

[8] Dominici, F., McDermott, A., Zeger, S.L. & Samet, J.M. (2002). On the use of generalized additive models in time series of air pollution and health, *American Journal of Epidemiology* **156**(3), 193–203.

[9] Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion), *Statistical Science* **11**(2), 89–121.

[10] Eubank, R.L. (1999). *Smoothing Splines and Nonparametric Regression*. 2nd Ed. Marcel Dekker, New York and Basel.

[11] Eubank, R.L. (2000). Spline regression, in *Smoothing and Regression. Approaches, Computation and Application*, M.G. Schimek, ed. John Wiley & Sons, New York, pp. 1–18.

[12] Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**(1), 1–141.

[13] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*, Volume 58 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

[14] Gu, C. (2000). Multivariate spline regression, in *Smoothing and Regression. Approaches, Computation and Application*, M.G. Schimek, ed. John Wiley & Sons, New York, pp. 329–355.

[15] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag, New York.

[16] Hastie, T.J. (1996). Pseudosplines, *Journal of the Royal Statistical Society, Series B* **58**(2), 379–396.

[17] Hastie T.J. & Tibshirani R.J. (1990). *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

[18] Hastie, T.J., Tibshirani, R.J. & Friedman, J.H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

[19] He, X. & Shi, P. (1998). Monotone B-spline smoothing, *Journal of the American Statistical Association* **93**(442), 643–650.

[20] Heckman, N. & Ramsay, J.O. (2000). Penalized regression with model-based penalties, *Canadian Journal of Statistics* **28**, 241–258.

[21] Hutchinson, M.F. & de Hoog, F.R. (1985). Smoothing noisy data with spline functions, *Numerische Mathematik* **47**, 99–106.

[22] Hutchinson, M.F. & de Hoog, F.R. (1987). An efficient method for calculating smoothing splines using orthogonal transformations, *Numerische Mathematik* **50**, 311–319.

[23] Lee, T.C.M. (2003). Smoothing parameter selection for smoothing splines: a simulation study, *Computational Statistics & Data Analysis* **42**(1–2), 139–148.

[24] Lindstrom, M.J. (1999). Penalized estimation of free-knot splines, *Journal of Computational and Graphical Statistics* **8**(2), 333–352.

[25] Lindstrom, M.J. (2002). Bayesian estimation of free-knot splines using reversible jumps, *Computational Statistics & Data Analysis* **41**(2), 255–269.

[26] Mammen, E., Marron, J.S., Turlach, B.A. & Wand, M.P. (2001). A general projection framework for constrained smoothing, *Statistical Science* **16**(3), 232–248.

[27] Nychka, D.W. (2000). Spatial-process estimates as smoothers, in *Smoothing and Regression. Approaches, Computation and Application*, M.G. Schimek, ed. John Wiley & Sons, New York, pp. 393–424.

[28] Osborne, M.R., Presnell, B. & Turlach, B.A. (1998). Knot selection for regression splines via the LASSO, in *Dimension Reduction, Computational Complexity, and Information*, Volume 30 of *Computing Science and Statistics*, pp. 44–49, S. Weisberg, ed. Interface Foundation of North America, Inc., Fairfax Station pp. 22039–27460.

[29] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion), *Statistical Science* **1**(4), 502–527.

[30] O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators, *SIAM Journal on Scientific and Statistical Computing* **9**(2), 363–379.

[31] Pittman, J. (2002). Adaptive splines and genetic algorithms, *Journal of Computational and Graphical Statistics* **11**(3), 615–638.

[32] R Development Core Team. (2004). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3. `http://www.R-project.org`

[33] Ramsay, J.O. (1988). Monotone regression splines in action (with discussion), *Statistical Science* **3**(4), 425–461.

[34] Ramsay, J.O. (1998). Estimating smooth monotone functions, *Journal of the Royal Statistical Society, Series B* **60**(2), 365–375.

[35] Ramsay, J.O. & Silverman, B.W. (1997). *Functional Data Analysis*. Springer-Verlag, New York.

[36] Rao, T.S. (1998). Spline function, in *Encyclopedia of Biostatistics*, Vol. 6, P. Armitage & T. Colton, eds. John Wiley & Sons, New York, pp. 4210–4212.

[37] Reinsch, C.R. (1967). Smoothing by spline functions, *Numerische Mathematik* **10**, 177–183.

[38] Reinsch, C.R. (1971). Smoothing by spline functions. II, *Numerische Mathematik* **16**, 451–454.

[39] Ruppert, D. (2002). Selecting the number of knots for penalized splines, *Journal of Computational and Graphical Statistics*. **11**(4), 735–757.

[40] Ruppert, D. & Carroll, R.J. (2000). Spatially-adaptive penalties for spline fitting, *Australian & New Zealand Journal of Statistics* **42**(2), 205–223.

[41] Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical And Probabilistic Mathematics, Cambridge University Press, Cambridge.

[42] Schimek, M.G. ed. (2000). *Smoothing and Regression. Approaches, Computation and Application*. John Wiley & Sons, New York.

[43] Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics* **75**, 317–344.

[44] Smith, P.L. (1982). Curve Fitting And Modeling With Splines Using Statistical Variable Selection Techniques, Report NASA 166 034, NASA, Langley Research Center, Hampton.

[45] Speed, T. (1991). Comments on "that blup is a good thing: The estimation of random effects", *Statistical Science* **6**(1), 42–44.

[46] Stone, C.J., Hansen, M.H., Kooperberg, C. & Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion), *Annals of Statistics* **25**(4), 1371–1470.

[47] van der Linde, A. (2000). Variance estimation and smoothing-parameter selection for spline regression, in *Smoothing and Regression. Approaches, Computation and Application*, M.G. Schimek, ed. John Wiley & Sons, New York, pp. 19–41.

[48] Wahba, G. (1990). *Spline Functions for Observational Data*, Volume 59 of *CBMS-NSF Regional Conference series*. SIAM, Philadelphia.

[49] Wand, M.P. (2000). A comparison of regression spline smoothing procedures, *Computational Statistics* **15**(4), 443–462.

[50] Wood, S.N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society, Series B* **62**(2), 413–428.

(*See also* **Multivariate Adaptive Splines for Analyzing Longitudinal Data**)

BERWIN A. TURLACH

Split Plot Designs

The term "split plot" derives from agricultural experiments in which investigators test one treatment, such as a method of irrigation, on large plots of land referred to as whole plots. These whole plots then often can be separated into smaller subplots or split plots to test a second treatment, such as a type of fertilization, as part of the initial or primary experiment [11]. Assignment of treatments requires two stages. First, treatment A, with a levels, is assigned randomly (*see* **Randomization**) to the whole plots (there are ar whole plots or r replicates). Then, treatment B, with b levels, is assigned randomly to the split plots or subplots, so that every whole plot has a subplot receiving each of the b levels of treatment B. These investigations possess a "nested design", since the treatment applied to the subplots is "nested" within the treatment applied to whole plots. The essential feature within the agricultural context is the two-stage randomization: whole plots are assigned randomly to treatment and each whole plot receives all of the subplot treatments. This basic design is seen to have applications in areas of research, such as engineering and epidemiology, that extend beyond the agricultural framework. In these areas, some researchers have employed the term "split unit", instead of "split plot", to describe the study design [1].

We now present four examples to demonstrate the wide applicability of the split plot design. In each case, we measure some quantity (such as yield in bushels per acre) which we denote as Y_{ijk} meaning the ith replication for the kth subplot of the jth whole or main plot.

1. Three fields are selected from each of five farms to test the effects of herbicides and insecticides on yield. Each of the three fields is subdivided into four plots. Three concentrations of herbicides are assigned randomly to the three fields, and four different insecticides are assigned randomly to each of the four subplots within each field [8]. Y_{ijk} represents the yield per acre. This example demonstrates the classical split plot design as it developed within the agricultural framework [3]. The whole plots are represented by the three fields from each of the five farms. The split plots are represented by the four divisions within each field. Treatment A ($a = 3$) is the application of the three herbicide concentrations, and treatment B ($b = 4$) is the application of the four insecticides. There are five replicates ($r = 5$) represented by the five selected farms.
2. To measure the activated life of batteries, an engineer randomly assigns 18 batteries to be examined at three different temperatures. At each temperature, the battery is tested using four electrolytes [10]. Y_{ijk} represents the life of the battery under specified testing conditions. For this application, the whole plot corresponds to the particular battery, and the split plots to the four different electrolytes used in each battery tested. Treatment A becomes the temperature for testing ($a = 3$) and treatment B becomes the electrolyte applied ($b = 4$). There are six replicates ($r = 6$), representing the number of batteries tested at each temperature–electrolyte combination. Here we identify a split plot as a repeat test on the same experimental unit.
3. To study the effects of population density or crowding on the **prevalence** of upper respiratory tract infection, six families are chosen randomly from each of three neighborhoods (18 families in total) that are classified as overcrowded, crowded, or uncrowded [1]. Each family consists of a mother, father, and three children, designated 1, 2, and 3 by descending age. Y_{ijk} is the number of positive swabs for pneumococcus. For this application the whole plot is represented by the family, seen as a unit and the split plot by the family member. Treatment A ($a = 3$) becomes the level of crowding and treatment B ($b = 5$), the status within the family (mother, father, or child 1, 2, or 3). There are six replicates ($r = 6$), representing the six families from each of the different neighborhoods. Here, we replace random treatment assignment with sampling strata (*see* **Stratification**).
4. To determine sex differences in levels of insulin growth factor-one (IGF-1), an investigator selects one group of males and one group of females. Blood samples are drawn from each participant, and each sample split into two aliquots. Each of the two aliquots is assigned randomly to different laboratories, and each laboratory supplies an IGF-1 determination. Y_{ijk} is the determination of IGF-1. For this application, the whole

Table 1 IGF-1 determinations at two laboratories, by sex

Replicate	1	2	3	4	5	6	7	8	9	10
Males										
Laboratory 1	152	119	119	113	131	117	134	136	112	118
Laboratory 2	149	117	122	119	131	109	140	142	126	126
Females										
Laboratory 1	163	169	150	173	163	158	145	151	173	190
Laboratory 2	156	163	151	169	170	146	155	148	164	189

plot is represented by the blood sample, and the split plot is represented by the aliquot. Treatment A ($a = 2$) becomes gender, a stratified sampling, and treatment B ($b = 2$), the laboratory. Ten replicates ($r = 10$) are conducted for each laboratory–sex combination. The determinations presented in Table 1 will be used to demonstrate numerical calculations. (These are simulated data based on an actual experiment.)

Split Plot Analysis

We now examine data collected from a split plot design format using a split plot **analysis of variance**. The analysis owes its derivation to the split plot model, which we represent by the following equation:

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \beta_k + (\rho\alpha)_{ij} + (\alpha\beta)_{jk} + e_{ijk}, \quad (1)$$

where μ is a constant depicting the "grand mean", and ρ_i, α_j and so on, are parameters representing "effects". For example, α_j represents the effect of treatment A at level j, and β_k represents the effect of treatment B at level k. The effects may be **fixed**, **random**, or a mixture of both [12]. We assume that the ρ_i and $(\rho\alpha)_{ij}$ effects are random. Some researchers refer to this as the fixed-effects model [12], while others label it the random block-effects model [9].

We make the following assumptions:

1. The ρ_i are independent $N(0, \sigma_\rho^2)$ (*see* **Normal Distribution**).
2. The $(\rho\alpha)_{ij}$ are distributed $N(0, \sigma_{\rho\alpha}^2)$, and $\sum_{j=1}^{b}(\rho\alpha)_{ij} = 0$ for every i.
3. The e_{ijk} are independent $N(0, \sigma^2)$.
4. The $\rho_i, (\rho\alpha)_{ij}$, and e_{ijk} are mutually independent.

The additional fixed effects in the model are such that:

1. The α_j are constants with $\sum_{j=1}^{a}\alpha_j = 0$.
2. The β_k are constants with $\sum_{k=1}^{b}\beta_k = 0$.
3. The $(\alpha\beta)_{jk}$ are constant with $\sum_{j=1}^{a}(\alpha\beta)_{jk} = 0$ for every k, and $\sum_{k=1}^{b}(\alpha\beta)_{jk} = 0$ for every j.

With these assumptions, $E(Y_{ijk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$. All observations have the same **variance** $(\text{var}(Y_{ijk}) = \sigma_\rho^2 + \sigma_{\rho\alpha}^2 + \sigma^2)$. Observations within a whole plot have a constant **correlation**, $(\sigma^2 + \sigma_{\rho\alpha}^2)/(\sigma^2 + \sigma_\rho^2 + \sigma_{\rho\alpha}^2)$.

The general analysis of variance for the split plot design is presented in Table 2, and in Table 3 is presented the specific analysis of variance for the data in Table 1. The whole plots comprise $ar = 20$ units, with 19 degrees of freedom (df) for between whole-plot comparisons. These 19 df among whole plots can be partitioned further into $r - 1 = 9$ df for replicates; $a - 1 = 1$ df for treatment A; and $(a - 1)(r - 1) = 10$ df for error in whole plot comparisons. Within each whole plot, there are $(b - 1) = 1$ df associated with variation within the whole plot, giving a total of $ar(b - 1) = 20$ df for comparisons within whole plots. These 20 df are partitioned further into $b - 1 = 1$ df for treatment B, $(a - 1)(b - 1) = 1$ df for the **interaction** between treatments A and B, and $a(b - 1)(r - 1) = 18$ df for error in whole-plot comparisons.

There are three hypotheses of interest concerning the effects of treatments when using split plot designs: (i) no effect of treatment A; (ii) no effect of treatment B; and (iii) no interaction effect. When these three **null hypotheses** are true, we derive the F statistics presented in Table 2. The correctness of these tests can be seen intuitively by examining the expected mean squares provided in the table. For example, when there are no main-plot effects, all of the α_j

Table 2 Analysis of variance for a split plot design

Source	SS	df	MS	E(MS)	F
Between whole plots (whole plot comparisons)					
A	$br\sum_{j=1}^{a}(\overline{Y}_{.j.}-\overline{Y}_{...})^2$	$a-1$	MSA	$\sigma^2+b\sigma^2_{\rho\alpha}+\frac{rb}{a-1}\sum_{j=1}^{a}\alpha_j^2$	MSA/MSE (A)
Replicates	$ab\sum_{i=1}^{r}(\overline{Y}_{i..}-\overline{Y}_{...})^2$	$r-1$	MSR		
Error (A)	$b\sum_{i=1}^{r}\sum_{j=1}^{a}\left([\overline{Y}_{ij.}-\overline{Y}_{i..}]-[\overline{Y}_{.j.}-\overline{Y}_{...}]\right)^2$	$(a-1)(r-1)$	MSE (A)	$\sigma^2+b\sigma^2_{\rho\alpha}$	
Within whole plots (subplot comparisons)					
B	$ar\sum_{k=1}^{b}(\overline{Y}_{..k}-\overline{Y}_{...})^2$	$b-1$	MSB	$\sigma^2+\frac{ra}{b-1}\sum_{k=1}^{b}\beta_k^2$	MSB/MSE (B)
AB interaction	$r\sum_{j=1}^{a}\sum_{k=1}^{b}\left([\overline{Y}_{.jk}-\overline{Y}_{.j.}]-[\overline{Y}_{..k}-\overline{Y}_{...}]\right)^2$	$(a-1)(b-1)$	MSAB	$\sigma^2+\frac{r}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{k=1}^{b}(\alpha\beta)_{jk}^2$	MSAB/MSE (B)
Error (B)	$\sum_{i=1}^{r}\sum_{j=1}^{a}\sum_{k=1}^{b}\left([\overline{Y}_{ijk}-\overline{Y}_{.jk}]-[\overline{Y}_{ij.}-\overline{Y}_{...}]\right)^2$	$a(b-1)(r-1)$	MSE (B)	σ^2	
Total	$\sum_{i=1}^{r}\sum_{j=1}^{a}\sum_{k=1}^{b}(Y_{ijk}-\overline{Y}_{...})^2$	$abr-1$			

Table 3 Analysis of variance for data in Table 1

Source	SS	df	MS	F	P
Gender comparisons					
Gender	12 744.9	1	12 744.9	31.58	0.0003
Replicates	1 954.9	9			
Error (gender)	3 632.1	9	403.6		
Laboratory comparisons					
Lab	0.9	1	0.9	0.04	0.8420
Interaction	72.9	1	72.9	3.31	0.0854
Error (laboratory)	396.2	18	22.0		
Total	18 801.9	39			

are 0, so the MSA and MSE(A) are independent with the same expected value, $(\sigma^2 + b\sigma^2_{\rho\alpha})$. Table 3 documents a sizable gender effect ($F = 31.58$, $P = 0.0003$), but no effect due to laboratory assignment ($F = 0.04$, $P = 0.842$) or to laboratory by sex interaction ($F = 3.31$, $P = 0.0854$).

Table 2 demonstrates an important feature of the split plot design: comparisons can be made more precisely within whole plots, including comparisons of interaction effects of treatments, than between whole plots. When comparing the average IGF-1 for males with that for females in example 4, we compare the mean among males to the mean among females. Using common notation, we state the comparison as follows: $\overline{Y}_{..1} - \overline{Y}_{..2}$. The variance for the difference uses the estimator of $\sigma^2 + b\sigma^2_{\rho\alpha}$. To compare the two laboratories. we use the difference in average IGF-1 determinations, $\overline{Y}_{..1} - \overline{Y}_{..2}$. The variance of this difference uses the estimator of σ^2, which is clearly smaller than $\sigma^2 + b\sigma^2_{\rho\alpha}$. These examples demonstrate that the estimations of within-plot differences are more precise than estimations between plots. We show this theoretically by making the simple assumption that all observations have variance σ^2, and that each pair of observations within a whole plot has the correlation ρ [3]. This feature of the split plot design leads some investigators to advocate its use when the estimation of one effect takes precedence over others [10, 12].

Extensions and Related Analyses

There are many situations for which the term "split plot" is genuinely descriptive of the analyses concerned, as is the case with our examples provided above. The data fit nicely into the classical split plot design model. The design and analytic method as employed in these examples, however, have limited application. It may be advantageous to extend the use of split plot design to perform four additional types of analyses: (i) to achieve additional levels of nesting; (ii) to adjust for **covariates** using an **analysis of covariance**; (iii) to extend distributional assumptions beyond normal distributions; and (iv) to generalize assumptions concerning the correlation structure of data.

In our examples, we assume that each whole plot is split into subplots and that the subplot receiving treatment B is nested within the whole plot receiving treatment A. We can extend this concept to design studies in which each subplot is further divided into c sub-subplots. We then nest treatment C (at c levels) within treatment B. The resulting design is referred to as a split-split plot design [6]. Theoretically, the division of split plots into smaller and smaller subplots can continue indefinitely.

When conducting **observational studies**, we may find it necessary to adjust for a covariate by measuring its effects as part of the split plot design. This is accomplished by including additional terms for covariate effects in the model specified in (1) [6, 7].

To derive the F statistics in Table 2, we assume data from a normal distribution. We can however, employ (1) with the assumptions of the **generalized linear model**. Cologne et al. [4], for example, adapt a split plot analysis that assumes **Poisson** observations to a study of micronucleus frequencies and radiation sensitivity.

Finally, the requirement that the correlation between observations within a whole plot remains constant is often unreasonable. The split plot analysis of variance may be used to examine **longitudinal data** or repeated measures. For these analyses we consider measurements taken at different times to be the split plot observations. With this definition we may consider **crossover trials** to be a special type of split plot design [2]. The assumptions that we employ for (1) lead to the requirement that the data within one whole unit be equally correlated. When the split plot represents time, however, it is more realistic to assume that the correlation between measurements that are closely timed are higher than those between

measurements further apart in time. Extension of the correlation structure can be accomplished using the **generalized estimating equation** approach of Zeger and co-workers [5, 13].

References

[1] Armitage, P. (1971). *Statistical Methods in Medical Research*. Blackwell Scientific, Oxford.

[2] Castellana, J. & Patel, H. (1983). Analysis of two-period crossover designs in a multicenter clinical trial, *Biometrics* **41**, 969–977.

[3] Cochran, W. & Cox, G. (1950). *Experimental Designs*. Wiley, New York.

[4] Cologne, J., Carter, R., Fujita, S. & Ran, S. (1993). Application of generalized estimating equations to a study of *in vitro* radiation sensitivity, *Biometrics* **49**, 927–934.

[5] Diggle, P., Liang, K. & Zeger, S. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

[6] Federer, W. & Meredith, M. (1992). Covariance analysis for split-plot and split-block designs, *American Statistician* **46**, 155–162.

[7] Monzelun, C. & Blouin, D. (1988). A general nested split-plot analysis of covariance, *Journal of the American Statistical Association* **83**, 818–823.

[8] Monzelun, C., Blouin, D. & Malone, L. (1984). Contrasting split plot and repeated measures experiments and analyses, *American Statistician* **38**, 21–27.

[9] Neter, J., Wasserman, W. & Kutner, M. (1990). *Applied Linear Statistical Models*, 3rd Ed. Richard D. Irwin, Boston.

[10] Ostle, B. (1963). *Statistics in Research*, 2nd Ed. Iowa State University Press, Ames.

[11] Snedecor, G. & Cochran, W. (1989). *Statistical Methods*, 8th Ed. Iowa State University Press, Ames.

[12] Steel, R. & Torrie, J. (1960). *Principles and Procedures of Statistics*. McGraw-Hill, New York.

[13] Zeger, S., Liang, K. & Albert, P. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

DAN MCGEE

S-PLUS and S

S is an interactive programming language (*see* **Computer Languages and Programs**) developed at Bell Labs, Murray Hill, NJ, for data analysis and graphics (*see* **Graphical Displays**). S-PLUS is the commercially available software implementation of S developed and marketed by the Insightful Corporation, Seattle. In 1999, the originator of S, John Chambers, was presented with the ACM Software Systems Award for his work on the S system. The name "S" is usually reserved for the language and "S-PLUS" for the interactive environment and software tools based upon it.

The original purpose of S was to provide a vehicle for technology development, testing, and transfer within the data analysis and graphics community of Bell Labs. The S-PLUS implementation has made its facilities accessible to a general data analysis community, but it remains a fully configured developmental tool for professional data analysts.

On Unix or Linux, an S-PLUS session typically uses a command-line interface with printed output displayed in the session window and graphical output on one or more graphics windows. Since S-PLUS version 4 (released March 1997) of Windows, there has been a graphical user interface available as an alternative, with an object explorer tool, a **spreadsheet**-like data browser and menu-driven commands for many standard tasks (but only on the Windows platform).

In 1999, a revision and extension of the S language was issued by Chambers and is described officially in his book [4]. This new language version forms the basis of S-PLUS on all software platforms from Release 6 for Unix, Windows, and Linux.

The main facilities provided within the environment are

1. an interpreter for an object-oriented C-like programming language, S,
2. support for a wide variety of static and dynamic, color graphics facilities,
3. software for a wide variety of basic computations, data analysis techniques, and statistical procedures, and
4. dynamic loading of routines written in C or Fortran, and hence open access to other software.

An unusual feature of S-PLUS is that the objects it creates are permanent and available in later S-PLUS sessions until the user removes them.

Data Analysis Software Available

Since S is a complete programming language, a user may, at least in principle, write a function to do any calculation. In practice, however, user-written functions in S are typically very short and for special purposes. Most standard operations are already part of the software available within the system and most large-scale computations are done using compiled code written in C or Fortran. The categories of software available in the S-PLUS implementation include:

1. **Linear regression** models, including multistratum **analysis of variance**, mixed effects and **multivariate multiple regression** models and a suite of local, smooth, or **robust regression** techniques;
2. **Generalized linear** (including some **generalized linear mixed models**) and **generalized additive models**;
3. **Nonlinear regression** models, including **random effects**, and general nonlinear **optimization**;
4. Regular and irregular **time series** with time-domain and frequency-domain analyses;
5. Parametric and nonparametric survival analysis (*see* **Parametric Models in Survival Analysis**; **Survival Analysis, Overview**; **Survival Analysis, Software**);
6. Classification and regression trees (*see* **Tree-structured Statistical Methods**), clustering algorithms (*see* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**) and multivariate scaling (*see* **Multidimensional Scaling**).

In addition to standard built-in software, a large collection of add-on libraries exists, which offer many other, often new techniques. These are largely written by users and contributed to the user community.

In addition to libraries, S-PLUS offers "modules", which are extensive additions to the suite of functions available, which can be purchased under a separate license. These include software for financial

modeling, **wavelets**, **spatial** analysis, **environmental** statistics, **experimental design**, **sequential analysis**, and large-scale optimization.

Some Literature

What is now called "old S" was described in [1] and [2]. Release 3 of S is described in [3] and the current revision, Release 4 of S, is described in [4]. Both references are still relevant. The basic modeling software was later added and described in [5]. Third party books include [6–10].

References

[1] Becker, R.A. & Chambers, J.M. (1984). *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth.

[2] Becker, R.A. & Chambers, J.M. (1985). *Extending the S System*. Wadsworth.

[3] Becker, R.A., Chambers, J.M. & Wilks, A.R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics*. Wadsworth.

[4] Chambers, J.M. (1998). *Programming with Data: A guide to the S Language*. Springer-Verlag, New York.

[5] Chambers, J.M. & Hastie, T.J. ed. (1991). *Statistical Models in S*. Wadsworth.

[6] Krause, A. & Olson, M. (2002). *The Basics of S and S-PLUS*, 3rd Ed. Springer-Verlag, New York.

[7] Pinheiro, J.C. & Bates, D.M. (2000). *Non-Linear Mixed Effects Models in S and S-PLUS*. Springer-Verlag, New York.

[8] Venables, W.N. & Ripley, B.D. (2000). *S Programming*. Springer-Verlag, New York.

[9] Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*, 4th Ed. Springer-Verlag, New York.

[10] Zivot, E. & Wang, J. (2002). *Modelling Financial Time Series with S-PLUS*. Springer-Verlag, New York.

(*See also* **Software, Biostatistical**)

W.N. VENABLES

Sports Medicine

The statistical methods used in sports medicine differ little from those in general clinical medicine. There is a variation in the degree of sophistication of research methods applied in different subsections of sport and exercise medicine, which reflects the development of this particular field. The earliest major research studies in sport and exercise medicine were the epidemiologic studies of the benefits of exercise by Morris [11] and Paffenbarger et al. [15]. These studies were the first to demonstrate the possible cardiovascular benefits of physical activity on mortality, but were subject to many possible **confounding** effects. They were the forerunners of many epidemiologic investigations including studies by Powell et al. [16] and Berlin & Colditz [2], who were able to confirm that regular exercise was associated with a reduction in the incidence of coronary heart disease. We now have further evidence from large observational **cohort studies** of both men and women for up to 25 years demonstrating that low physical fitness is an important precursor of mortality [3]. Scientists continue to explore the relationship between physical activity and cardiovascular risk factors and, in general, are able to provide evidence linking specific risk factors with physical activity and fitness using **cross-sectional** population studies [9]. Additional evidence of a relationship may be explored using intervention studies (*see* **Clinical Trials, Overview**), and short-term **case–control studies** [19].

Sports injuries are usually first reported as case studies or a group of injuries collected into **case series**. Observations may then be explored further in a case–control study [6]. It is difficult to establish the incidence of sports injuries in the population because of problems in defining the **denominator**. Most evidence of population sports injury is recorded in **surveys** of attendance at a hospital casualty [1] or sports injury clinic [10]. A better method of recording incidence and exploring associated factors is the prospective cohort study. The highest quality methodology is of course the **randomized** controlled trial. Unlike most clinical investigations in which it is relatively easy to undertake a randomized controlled trial and blind participants (*see* **Blinding or Masking**), it is more difficult to apply this method to investigations in sport and exercise medicine where the intervention may involve some form of activity or physical device. Recruitment is often open, or by advertisement, and few people wish to be allocated to a **control** group. Blinding is usually impossible, and with any exercise program there is always the possibility of contamination. When the randomized controlled trial is applied successfully it leads to valuable high-quality research with clinical implications [8].

Another important consideration in sport and exercise sciences is how to identify the effects of risk factors (e.g. physical inactivity, dietary composition, and smoking) on health-related fitness variables (e.g. VO_2 max, grip strength, leg power, and arterial blood pressure) in the presence of confounding effects (e.g. differences in age or body weight). One obvious solution is to use techniques such as the **analysis of covariance**. However, the **frequency distributions** of many health-related fitness variables are known to be positively skewed (*see* **Skewness**) with heteroscedastic errors (*see* **Scedasticity**) [7, 12–14], and thus deviate considerably from the **normal distribution** with constant error **variance**. In addition, the relationship between many such variables and age, for example blood pressure [14], is certainly not linear. To accommodate these characteristics and associations, a **multiplicative model** with allometric body size components has been proposed [13, 14] that can explain the known proportional relationship with body size, the nonlinear age factor, and the heteroscedastic and positively skewed errors.

Prediction of athletic performance appears to be another constant source of fascination for many statisticians. When analyzing running times recorded in the Olympic Games between the years 1900 and 1976, Chatterjee & Chatterjee [5] generated a fierce debate as to the inaccuracies in their data [17] and the lack of evidence for their claimed asymptote [20]. A recent paper by Blest [4] continues to explore the relationship between running times and distance, by fitting separate power function models to the running times taken from the Olympic Games between the years 1912 and 1992. Based on these fitted exponents, the author tries to predict the limits of future performance using a variety of nonlinear curve fitting models. An alternative approach by Royston [18] uses fractional polynomials, to model changes in running *speed* over distance rather than running *time*. An examination of the **residuals** from both models favors the latter approach, which appears more successful in removing systematic residual effects from the running performance data.

References

[1] Bedford, P. & MacAuley, D. (1984). Attendances at a casualty department for sports related injury, *British Journal of Sports Medicine* **18**, 166–171.

[2] Berlin, J.A. & Colditz, G.A. (1990). A meta-analysis of physical activity in the prevention of coronary heart disease, *American Journal of Epidemiology* **132**, 612–628.

[3] Blair, S.M., Kampert, J.B., Kohl, 3rd, H.W., Barlow, C.E., Macera, C.A., Paffenbarger, R.S. Jr & Gibbons, L.W. (1996). Influences of cardio-respiratory fitness and other precursors on cardiovascular disease and all causes of mortality in men and women, *Journal of the American Medical Association* **276**, 205–210.

[4] Blest, D.C. (1996). Lower bounds for athletic performance, *Statistician* **45**, 243–253.

[5] Chatterjee, S. & Chatterjee, S. (1982). New lamps for old: an exploratory analysis of running times in Olympic Games, *Applied Statistics* **31**, 14–22.

[6] Cooper, C.J., Noakes, T.D., Dunne, T., Lambert, M.I. & Rochford, K. (1996). A high prevalence of abnormal personality traits in users of anabolic–androgenic steroids, *British Journal of Sports Medicine* **30**, 246–250.

[7] Dawber, T.R. (1980). *The Framingham Study*. Harvard University Press, Cambridge, Mass.

[8] Heinonen, A., Kannus, P., Sievanen, O., Pasanen, M., Rinne, M., Uusi-Rasi, K. & Vuori, I. (1996). Randomised controlled trial of effect of high-impact exercise of selected risk factors for osteoporotic fractures, *Lancet* **348**, 1343–1347.

[9] MacAuley, D., McCrum, E.E., Stott, G., Evans, A.E., Duly, E., Trinick, T., Sweeney, K. & Boreham, C.A. (1996). Physical activity, lipids, apolipoproteins, and Lp(a) in the Northern Ireland Health and Activity Survey, *Medicine and Science in Sports and Exercise* **28**, 720–763.

[10] Maffuli, N., Bundoc, R.C., Chan, K.M. & Cheng, J.C.Y. (1996). Paediatric sports injuries in Hong Kong: a seven year survey, *British Journal of Sports Medicine* **30**, 218–221.

[11] Morris, J.N., Hagan, A., Patterson, D.C. & Gardner, M.J. (1953). Incidence and prediction of ischaemic heart disease in London busmen, *Lancet* **ii**, 553–559.

[12] Nevill, A.M. & Holder, R.L. (1994). Modelling maximum oxygen uptake: a case study in non-linear regression formulation and comparison, *Applied Statistics* **43**, 653–666.

[13] Nevill, A.M. & Holder, R.L. (1995). Scaling, normalizing and "per ratio" standards: an allometric modeling approach, *Journal of Applied Physiology* **79**, 1027–1031.

[14] Nevill, A.M., Holder, R.L., Fentem, P.H., Rayson, M., Marshall, T., Cooke, C.B. & Tuxworth, W. (1997). Modelling the associations of BMI, physical activity and diet with arterial blood pressure: some results from the Allied Dunbar national fitness survey, *Annals of Human Biology* **24**, 229–247.

[15] Paffenbarger, R.S., Laughlin, M.E., Gima, A.S. & Black, R.A. (1970). Work activity of longshoremen as related to death from coronary heart disease and stroke, *New England Journal of Medicine* **282**, 1109–1114.

[16] Powell, K.E., Thompson, P.D., Caspersen, C.J. & Kendrick, J.S. (1987). Physical activity and the incidence of coronary heart disease, *Annual Review of Public Health* **8**, 253–287.

[17] Reid, D.D. & Sandland, R.L. (1983). New lamps for old? (letter), *Applied Statistics* **32**, 86–87.

[18] Royston, P. (1998). Modelling running speed in athletic track events, *Statistician* to appear.

[19] Woolf-May, K., Bird, S. & Owen, A. (1997). Effects of an 18 week walking programme on cardiac function in previously sedentary or relatively inactive adults, *British Sports Medicine* **31**, 48–53.

[20] Wootton, R. & Royston, J.P. (1983). New lamps for old (letter), *Applied Statistics* **32**, 88–89.

ALAN M. NEVILL & DOMHNALL MACAULEY

Spreadsheet

The term *spreadsheet* is derived from the sheet of paper employed by an accountant to set out financial calculations, often so large that it had to be spread out on a table. The first electronic spreadsheet, VisiCalc, was developed in 1979 by Bob Frankston and Dan Bricklin at Harvard. Originally designed as a business problem-solving tool, the remarkable versatility of the spreadsheet package has made it a popular working environment among computer users in many professions. The most commonly used spreadsheet packages include Excel, Lotus 1-2-3, Quattro Pro, and SuperCalc.

In medicine and the health sciences spreadsheets have proved useful in a wide variety of applications. For example, in resource management spreadsheets are used for scheduling [5], budgeting [10], **forecasting** [2], quality monitoring [9], "what-if" scenarios [7], and cost–benefit analysis [6]. Spreadsheet models [4] of complex systems and processes can be used to simulate behavior under varying conditions, allowing **optimization** [8] and **sensitivity analysis** [3] to be conducted. On a more mundane level the spreadsheet interface is easy to use by nontechnical personnel for data-entry and routine computation – for example, dosage calculations [1].

A basic spreadsheet comprises a rectangular array of cells with, say, columns referenced by letters A, B, C ... and rows by number $1, 2, 3 \ldots$. Thus the cell F3 would be in the third row of the sixth column. Cells may contain text, numbers, or formulas, or may be empty. Figure 1 shows the (formatted) results of a randomized block experiment (*see* **Randomized Complete Block Designs**). Cell A3 contains the text "Block 1", cell B3 contains the number 12.3, while cell F3 contains the formula = AVERAGE(B3:E3) which is calculated to be $(12.3 + 11.2 + 16.4 + 13.2)/4 = 13.275$. Formulas can be copied to other cells if a similar calculation is required. If the formula in F3 is copied into cell F4 it is automatically modified to = AVERAGE(B4:E4), it being assumed that any reference to row 3 should become row 4. In some situations this automatic modification needs to be overridden. To calculate the **residuals** $\{y_{ij} - y_{i\cdot} - y_{\cdot j} + y_{\cdot\cdot}\}$ in Figure 2 corresponding to data $\{y_{ij}\}$ in Figure 1, the required formula for cell B10 would be $= B3 - F3 - B6 + F6$. However, were we to copy this into B11, say, it would become $= B4 - F4 - B7 + F7$ – totally incorrect. Instead a \$ symbol can be used to force a row or column reference to remain unchanged during copying. Thus if the formula in B10 is written as $= B3 - \$F3 - B\$6 + \$F\$6$ it may safely be copied throughout the range B10:E12. Spreadsheets can be made more "transparent" to other users by attaching names to key cells or ranges. For example, in Figure 1, if the range B3:E3 were to be named "Block_1" then the formula in F3 could be entered as = AVERAGE(Block_1) – much more readable.

Formulas are dynamically linked to the cells from which they are calculated. Hence if any of the cells B3, C3, D3, or E3 is changed, the average in F3 is automatically recalculated. Automatic re-execution of formulas when data or model parameters are changed distinguishes the spreadsheet from a standard statistical package such as SAS (*see* **Software, Biostatistical**) where the program would need to be rerun. The spreadsheet thereby provides an interactive computing environment. Suppose, for example, that the investigator discovers that the response for Treatment II in Block 2 was actually missing and should not have been recorded as zero. The least squares estimate of the missing value can be found by trial

	A	B	C	D	E	F
1						
2	Response	Treat I	Treat II	Treat III	Treat IV	Mean
3	Block 1	12.3	11.2	16.4	13.2	13.275
4	Block 2	18.4	0	19.5	18.4	14.075
5	Block 3	12.8	11.9	14.4	12.7	12.95
6	Mean	14.5	7.7	16.766667	14.766667	13.433333
7						

Figure 1 Data from a randomized block experiment

8						
9	Residual	Treat I	Treat II	Treat III	Treat IV	Mean
10	Block 1	−2.0417	3.6583	−0.2083	−1.4083	0.0000
11	Block 2	3.2583	−8.3417	2.0917	2.9917	0.0000
12	Block 3	−1.2167	4.6833	−1.8833	−1.5833	0.0000
13	Mean	0.0000	0.0000	0.0000	0.0000	0.0000
14						

Figure 2 Residuals from analysis of randomized block experiment (continued from Figure 1)

and error, adjusting the value in C4 until the residual in C11 is seen to be zero. An alternative solution would be to employ a circular formula $= C4 - C11$ in C4, pressing the recalculation key to obtain successive iterates. (In fact, after 17 recalculations the missing value is estimated to be 16.6833.)

AVERAGE() is an example of an inbuilt function which calculates the arithmetic mean. Most spreadsheet packages provide functions for calculating simple descriptive statistics and linear regression coefficients. Microsoft Excel is notable for its wide range of additional statistical functions. These include all the major probability distribution functions (and their inverses) used to calculate P values (or critical values) in significance tests. There have been unfortunate errors in the algorithms employed by some spreadsheet packages. For example, in Excel 5.0 it was possible to obtain negative r^2 values when fitting a regression through the origin. Such errors cast doubt on the suitability of spreadsheets for serious statistical analysis.

An impressive feature of some recent spreadsheet packages is the facility to produce high-quality charts. These may either be stored on separate sheets in the same "workbook" as their data, or pasted onto the spreadsheet itself. Charts can be customized using different styles, fonts, colors, etc. Figure 3 shows a typical example. It is important not to rely on the default settings of the various options, otherwise "chart-junk" may result – visually beautiful but totally meaningless! Like a formula, a chart is dynamically linked to the data which it portrays. Indeed, in Excel the linkage can be selected to be two-way, so that when a data point on the chart is moved (by dragging with the mouse) the corresponding cell contents change accordingly.

The statistical facilities available within a spreadsheet package can usually be extended by writing subprograms (or macros) in the spreadsheet's own

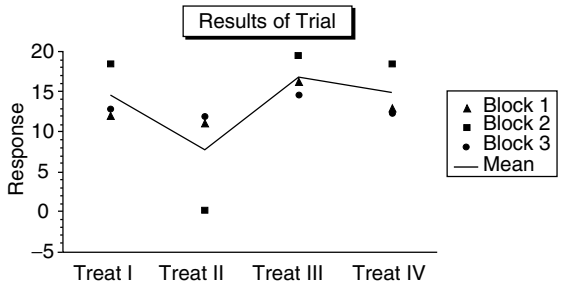

Figure 3 Spreadsheet chart of data from randomized block experiment

language – for example, Visual Basic in Excel 5. A suite of macros may even be supplied as optional "add-ins" with the software. In theory this allows even the most advanced statistical techniques to be implemented in spreadsheet form. In practice there is a danger that, unless care is taken with the algorithm and double-precision arithmetic used when appropriate, rounding errors will accumulate and numerical accuracy will suffer. Moreover the macro-output will not be dynamically linked to its input, and the spreadsheet will have lost its transparency. The proliferation of "add-ins" in spreadsheet packages, coupled with spreadsheet-like data entry facilities in statistical packages (*see* **Software, Biostatistical**), makes it likely that the distinction between the two types of package will gradually disappear. The major statistical packages such as SAS, SPSS, and Minitab already provide interface facilities with the popular spreadsheet packages, so that users can enjoy the best of both worlds.

References

[1] Balog, J.P., Sibata, C.H., Podgorsak, M.B. & Shin, K.H. (1995). The use of customized spreadsheets in

radiation therapy, *Physics in Medicine and Biology* **40**, 1057–1066.

[2] Corley, M.C. & Satterwhite, B.E. (1993). Forecasting ambulatory clinic workload to facilitate budgeting, *Nursing Economics* **11**, 77–81.

[3] Eisinger, D.S., Simmons, R.A., Lammering, M. & Sotiros, R. (1991). *Regulatory Toxicology and Pharmacology* **14**, 245–260.

[4] Kokol, P. (1990). Structured spreadsheet modelling in medical decision making and research, *Journal of Medical Systems* **14**, 107–117.

[5] Kooijman, C.J. & Klaasseen-Leil, C.C. (1995). Extraction, preparation and presentation of patient classification-data for the benefit of management overviews, *Medinfo* **8**, 1382–1385.

[6] Masobe, P., Lee, T. & Price, M. (1995). Isoniazid prophylactic therapy for tuberculosis in HIV-seropositive patients: a least-cost analysis, *South African Medical Journal* **85**, 75–81.

[7] Menzies, F.D. (1992). A microcomputer model for predicting output from beef suckler herds, *Veterinary Record* **130**, 9–12.

[8] Milsum, J.H. (1989). Determining optimal screening policies using decision trees and spreadsheets, *Computers in Biology and Medicine* **19**, 231–243.

[9] Phillips, M.S., Williams, D.B. & May, J.R. (1994). Using pharmacist clinical intervention data for quality improvement of medication use and physician assessment, *Joint Commission Journal on Quality Improvement* **20**, 569–576.

[10] Wight, J., Olliver, A. & Payne, N. (1996). Spreadsheet based computer model to predict demand for end-stage renal failure treatment, set contracts and monitor in year performance, *Nephrology, Dialysis, Transplantation* **11**, 1286–1291.

(*See also* **Algorithm**; **Graphical Displays**)

N. HUNT

Square Contingency Table

An $R \times C$ **contingency** table – that is, contingency table with R rows and C columns – is frequently seen in biomedical studies. When $R = C$, the contingency table is often referred to as a *square table*. Square tables usually arise in repeated measures experiments, in which a subject has his or her outcome variable observed repeatedly. Examples of repeated measures studies include longitudinal studies (observing the same subject over time); rater agreement studies (two investigators rate each subject in the study); and paired data studies (data obtained from a husband and wife, father and son, or case and control) (*see* **Matched Pairs With Categorical Data**; **Case–Control Study**).

First, we introduce some notation. Suppose that two discrete random variables, Y_{i1} and Y_{i2}, are observed on each of n independent subjects, where Y_{i1} can take on values $1, \ldots, R$, and Y_{i2} can take on values $1, \ldots, C$. Let the probabilities of the multinomial joint distribution of Y_{i1} and Y_{i2} be denoted by

$$p_{jk} = \Pr(Y_{i1} = j, Y_{i2} = k),$$

for $j = 1, \ldots, R$ and $k = 1, \ldots, C$. Since all the probabilities must sum to 1, there are $(RC - 1)$ nonredundant multinomial cell probabilities. We let $\mathbf{p}$ denote the $(RC - 1) \times 1$ probability vector of p_{jk}'s; for simplicity, we delete p_{RC}, and we take $R = C$ below. The marginal probabilities of the contingency table are $p_{j+} = \Pr(Y_{i1} = j)$ and $p_{+k} = \Pr(Y_{i2} = k)$, where the plus signs denote summing over the subscript they replace. We let n_{jk} denote the number of subjects with response level j on Y_{i1} and level k on Y_{i2}, and let $n = n_{++}$ be the total number of subjects in the study.

Models for square tables fall into three general classes – marginal models, **loglinear models**, and conditional models. Marginal models describe the similarity between the row and column marginal distributions. With loglinear models, we model the cell probabilities of the table. For example, the probabilities may be symmetric about the main diagonal, and a loglinear model can be used to describe this relationship among the cell probabilities. Conditional models model the conditional probabilities of the column variable given the row variable and are popularly used in longitudinal studies, letting the row variable represent a response at time 1 and the column variable represent the response at time 2. Often, in square tables, we are not always interested in modeling the data, but are interested in determining the association or agreement between the row and column variables.

Marginal Models

In a crossover study, a subject is given one treatment, has a washout period, and is then given a new treatment. Suppose that the response to each treatment is success, partial success, or failure. We are often interested in whether the marginal distribution of the outcome is the same for both treatments, often called *marginal homogeneity*. Under marginal homogeneity, the sum of the cell probabilities (i.e. the marginal probability) in the jth row of the contingency table equals the sum of the cell probabilities in the jth column; that is, $p_{j+} = p_{+j}$ for $j = 1, \ldots, R$. Lipsitz et al. [10] show that the maximum likelihood estimates for the p_{jk}'s under marginal homogeneity can be obtained via a linear model of the form

$$\mathbf{p} = \mathbf{X}\boldsymbol{\beta},$$

for the appropriate "design" matrix $\mathbf{X}$, which we now describe for $R = 3$, but which generalizes to any R. For $R = 3$, we can write the vector $\mathbf{p}$ in terms of the four cell probabilities, $\{p_{11}, p_{12}, p_{21}, p_{22}\}$ and the four marginal probabilities, $\{p_{1+}, p_{2+}, p_{+1}, p_{+2}\}$, as follows:

$$\begin{bmatrix} p_{11} \\ p_{12} \\ p_{13} \\ p_{21} \\ p_{22} \\ p_{23} \\ p_{31} \\ p_{32} \end{bmatrix} = \begin{bmatrix} p_{11} \\ p_{12} \\ p_{1+} - p_{11} - p_{12} \\ p_{21} \\ p_{22} \\ p_{2+} - p_{21} - p_{22} \\ p_{+1} - p_{11} - p_{21} \\ p_{+2} - p_{12} - p_{22} \end{bmatrix}.$$

Under marginal homogeneity, the row and column marginal probabilities are equal – that is, $p_{j+} = p_{+j} = p_j$ – which leads to the following linear model for the cell probabilities:

$$\begin{bmatrix} p_{11} \\ p_{12} \\ p_{13} \\ p_{21} \\ p_{22} \\ p_{23} \\ p_{31} \\ p_{32} \end{bmatrix} = \begin{bmatrix} p_{11} \\ p_{12} \\ p_1 - p_{11} - p_{12} \\ p_{21} \\ p_{22} \\ p_2 - p_{21} - p_{22} \\ p_1 - p_{11} - p_{21} \\ p_2 - p_{12} - p_{22} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 1 \\ -1 & 0 & -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \\ p_1 \\ p_2 \end{bmatrix}$$

$$= \mathbf{X}\boldsymbol{\beta}. \tag{1}$$

The **maximum likelihood estimates** (MLEs) can be obtained in any generalized linear modeling program, such as SAS *Proc GENMOD* [14], without any additional programming or iteration loops. In these programs, the outcome is the count n_{jk}, the covariates are the appropriate row of $\mathbf{X}$ in (1) without an intercept, and we specify a linear link with Poisson errors (we actually must include n_{33} as a value of the outcome, as described in [10]). Firth [7] also describes a method for obtaining the linear model for the p_{jk}s using Latin squares. Another method found in the literature for obtaining the MLE under homogeneity uses Lagrange multipliers [11].

After the estimates of the cell probabilities under marginal homogeneity have been obtained, a **likelihood ratio** statistic can be used to test whether marginal homogeneity holds. This statistic is approximately chi-square with $R-1$ degrees of freedom under the null. Alternately, a Wald statistic [3] (*see* **Likelihood**) can be used. It does not require the use of iterative techniques, since only the estimates of the cell probabilities for the saturated model are needed. The maximum likelihood estimate of p_{jk} for the saturated model (with R^2-1 nonredundant probabilities) is

$$\hat{p}_{jk} = \frac{n_{jk}}{n}, \tag{2}$$

and the MLEs of the marginal probabilities are $\hat{p}_{j+} = n_{j+}/n$ and $\hat{p}_{+j} = n_{+j}/n$. Suppose that we let the vector

$$\mathbf{U} = [\hat{p}_{1+} - \hat{p}_{+1}, \ldots, \hat{p}_{R-1,+} - \hat{p}_{+,R-1}]'$$

contain the first $R-1$ differences in the marginal probabilities (the Rth difference is redundant). Under the null of marginal homogeneity, $\mathrm{E}(\mathbf{U}) = \mathbf{0}$. If we let $\hat{\mathbf{V}}$ be the estimated covariance matrix of the vector $\mathbf{U}$ under the alternative (see [3]), then the Wald test statistic for homogeneity is the quadratic form

$$X^2 = \mathbf{U}'\hat{\mathbf{V}}^{-1}\mathbf{U}, \tag{3}$$

which will be approximately chi-square with $R-1$ degrees of freedom under the null. If, instead of estimating the variance under the alternative in (3), we let $\hat{\mathbf{V}}$ be the estimated covariance of $\mathbf{U}$ under the null of homogeneity, and if there are only two rows and columns in the table ($R = C = 2$), then (3) reduces to

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}. \tag{4}$$

The test statistic in (4) is popularly known as *McNemar's test* for equality of correlated proportions [13].

In the crossover study discussed earlier, the outcomes "success", "partial success", and "failure" are examples of **ordered categorical data**. Suppose that we want to test for marginal homogeneity, taking this ordering into account. We can put an ordinal model on the margins, such as the **proportional-odds model** [12], and test whether the parameters of the proportional-odds models in the two margins are the same. The test statistics are similar to those given for the nominal case above. The Wald test for marginal homogeneity for proportional-odds marginal models can be obtained in SAS *Proc CATMOD* [15] using methods derived in [9]. The likelihood ratio statistic can be obtained in a generalized linear models program, but will require some additional programming. More details on marginal modeling can be found in [2, Chapter 9] and in the article on marginal models.

Loglinear Models

The models just discussed involved testing whether the marginal probabilities of the table are equal. Now, we discuss models which describe the relationship among the cell probabilities of the $R \times R$ table. One

such model for the cell probabilities is a symmetry model, in which

$$p_{jk} = p_{kj}, \quad \text{for } j \neq k. \tag{5}$$

Note that symmetry implies marginal homogeneity since, under (5),

$$p_{j+} = \sum_{k=1}^{R} p_{jk} = \sum_{k=1}^{R} p_{kj} = p_{+j}.$$

In fact, when $R = C = 2$, marginal homogeneity and symmetry are identical, and both imply that $p_{12} = p_{21}$.

For modeling purposes, symmetry is most easily written as a loglinear model. Loglinear models for contingency tables are usually written in terms of the expected cell counts

$$m_{jk} = \mathrm{E}(n_{jk}) = np_{jk}.$$

In terms of the expected cell counts, symmetry is written as

$$m_{jk} = m_{kj}, \quad \text{for } j \neq k.$$

Then, the symmetry loglinear model is

$$\log m_{jk} = \mu + \beta_j + \beta_k + \beta_{jk}, \tag{6}$$

where $\beta_{jk} = \beta_{kj}$, with the appropriate identifiability constraints on the parameters, such as $\beta_R = 0$ and $\beta_{Rj} = \beta_{jR} = 0$ for $j = 1, \ldots, R$. Since $\beta_{jk} = \beta_{kj}$, it is easy to show that $m_{jk} = m_{kj}$, and symmetry holds for (6).

The MLEs for the expected cell counts under symmetry are

$$\hat{m}_{jj} = n_{jj} \qquad \text{and} \qquad \hat{m}_{jk} = \frac{n_{jk} + n_{kj}}{2}.$$

The estimated cell probabilities are just $\hat{p}_{jk} = \hat{m}_{jk}/n$. Intuitively, since symmetry does not concern the diagonal cells, the estimated expected diagonal counts are just the observed diagonal counts. Also, an estimated off-diagonal count is just the average of the observed counts in cells jk and kj. Under the symmetry model, we only need to estimate $R(R-1)/2$ off-diagonal probabilities on one side of the diagonal, so we have placed $R(R-1)/2$ constraints on the cell probabilities under the null, and the likelihood ratio test statistic, score test statistic (*see* **Likelihood**), or Wald test statistic for symmetry are approximately chi-square with $R(R-1)/2$ degrees of freedom. The score test statistic (equivalent to Pearson's chi-square for this problem) is the simplest [1]:

$$X^2 = \sum_{j<k} \frac{(n_{jk} - n_{kj})^2}{n_{jk} + n_{kj}}. \tag{7}$$

Recall that, when $R = C = 2$, marginal homogeneity and symmetry are identical, so that (7) is also a test statistic for marginal homogeneity when $R = 2$, and is identical to the McNemar test statistic discussed earlier.

The symmetry loglinear model puts many constraints on the probabilities. A loglinear model that has fewer constraints (and more parameters) does not force the row and column main effects in (6) to be equal,

$$\log m_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}, \tag{8}$$

where $(\alpha\beta)_{jk} = (\alpha\beta)_{kj}$. The loglinear model in (8) is called *quasi-symmetry* [5]. If symmetry holds, then $\alpha_j = \beta_j$. The **Bradley–Terry model** [4] for **paired comparisons** is a special case of this quasi-symmetry model.

Using the identifiability constraints $\alpha_R = 0$, $\beta_R = 0$, and $(\alpha\beta)_{Rj} = (\alpha\beta)_{jR} = 0$ for $j = 1, \ldots, R$, the interaction parameter $(\alpha\beta)_{jk}$ is the log-odds ratio for the jth and Rth rows and kth and Rth columns: that is,

$$(\alpha\beta)_{jk} = \log\left(\frac{p_{jk}p_{RR}}{p_{jR}p_{Rk}}\right).$$

Then, in the quasi-symmetry model, since $(\alpha\beta)_{jk} = (\alpha\beta)_{kj}$, the log-odds ratio for the jth and Rth rows and kth and Rth columns equals the log-odds ratio for the kth and Rth rows and jth and Rth columns. In particular, these log-odds ratios are symmetric.

One last loglinear model that we discuss is a *quasi-independence* model. If the row and column variables are independent, then

$$p_{jk} = p_{j+}p_{+k},$$

or, equivalently, in terms of the loglinear model,

$$\log m_{jk} = \mu + \alpha_j + \beta_k. \tag{9}$$

In repeated measures studies, most of the agreement is often on the diagonal, so that the diagonal

elements are longer than expected under independence, with independence holding in the off-diagonal elements. A simple loglinear model that expresses such a relationship is the quasi-independence loglinear model,

$$\log m_{jk} = \mu + \alpha_j + \beta_k + \gamma_j I(j = k), \qquad (10)$$

where $I(j = k)$ equals 1 for diagonal elements $j = k$ and 0 for off-diagonal elements. For off-diagonal elements, (10) is identical to (9), and, if $\gamma_j > 0$, then the diagonal elements are longer than expected under independence.

When the rows and columns are ordered, these loglinear models can be extended by assigning **scores** to the rows and columns (see, for example, [8]). For more details on loglinear models, see [1].

Conditional Models

Often in longitudinal studies, we are interested in transitions or change in states, such as the response at time 2 given the response at time 1. In general, in crossover designs, polling studies, or employment studies (mover–stayer studies), investigators are interested in these conditional probabilities for the column variable given the row variable. When the row variable represents a response at time 1, and column variable represents the response at time 2, we model the probability of response k at time 2 given response j at time 1:

$$p_{k|j} = \Pr[Y_{i2} = k | Y_{i1} = j], \quad j = 1, \ldots, R. \quad (11)$$

Suppose that, in an arthritis clinical trial for the effectiveness of a single treatment, the subjects are observed once before treatment, then are given the new treatment, and then observed again. Suppose that the possible outcomes at the two times are "no pain", "mild pain", and "severe pain". Then, we are interested in the changes in pain, as modeled by the conditional probabilities of pain status at time 2 given the pain status at time 1. If there are two levels ($R = 2$), then we can apply logistic regression to the model given in (11), treating Y_{i1} as a covariate. If ($R > 2$) and the levels are not ordered, multinomial logistic regression can be used (*see* **Polytomous Data**); if the levels are ordered (as in this example), an ordinal logistic model such as the proportional-odds model [12] can be used.

Measures of Association and Agreement

For general ($R \times C$) tables, we are often interested in the **association** between the row and column variables (*see* **Association, Measures of**). For a (2×2) table, the odds ratio is a popular measure of association. In an ($R \times R$) square table, the log-odds ratios, as described above, can be used to measure association. For the row and column variables to be associated, you should be able to predict one from another. For example, if the odds ratio is 0, the row and column variables in a (2×2) are perfectly negatively associated. In repeated measures studies, and, in particular, inter-rater reliability studies, in which two investigators rate each subject in the study, we are interested in how well the row and column variables (the two ratings on each subject) agree. When two ratings agree, most of the observations in the contingency table will be on the diagonal (most ratings are similar). Two variables can be highly associated, but agreement between the two could be very low. Suppose that both diagonal elements in a (2×2) table are zero: then the raters completely disagree, and agreement (by any measure) is very low. However, the odds ratio is 0, and, as discussed above, the two variables are perfectly negatively associated. The kappa coefficient [6] is a popular measure of agreement corrected for chance agreement (*see* the articles on **Kappa** and **Agreement, Measurement of**).

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

[3] Bhapkar, V.P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data, *Journal of the American Statistical Association* **61**, 228–235.

[4] Bradley, R.A. & Terry, M.E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons, *Biometrika* **39**, 324–345.

[5] Caussinus, H. (1965). Contribution a l'analyse statistique des tableaux de correlation, *Annales de la Faculté des Sciences Université de Toulouse* **29**, 77–182.

[6] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[7] Firth, D. (1989). Marginal homogeneity and the superposition of Latin squares, *Biometrika* **76**, 179–182.

[8] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association* **74**, 537–552.

[9] Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. & Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data, *Biometrics* **33**, 133–158.

[10] Lipsitz, S.R., Laird, N.M. & Harrington, D.P. (1990). Finding the design matrix for the marginal homogeneity model, *Biometrika* **77**, 353–358.

[11] Madansky, A. (1963). Tests of homogeneity for correlated samples, *Journal of the American Statistical Association* **58**, 97–119.

[12] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[13] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–157.

[14] SAS Institute Inc. (1993). SAS/STAT Software: The GENMOD Procedure, Release 6.09, *SAS Technical Report P-243* SAS Institute Inc., Cary.

[15] SAS Institute Inc. (1993). *SAS/STAT User's Guide Volume 1: The CATMOD Procedure* Version 6, 4th Ed. SAS Institute Inc., Cary, pp. 405–517.

(*See also* **Crossover Designs**; **Quasi-independence**; **Quasi-symmetry**)

STUART R. LIPSITZ

Staggered Entry

A biostatistical survival study is said to have *staggered entry* if the study subjects are entered into the study at times which are related to their own disease history (e.g. immediately following diagnosis at a particular hospital for a specified disease, if other criteria for entry are met), but which are unpredictable from the point of view of the study. The most common method of dealing with such data, especially in the case of analysis at a single chronological time, is via the **life-table** method one ignores the chronological times of entry and treats as response variable the time from entry until event (primary endpoint, e.g. survival time, or **censoring**) together with an indicator of whether the primary endpoint is observed [27]. Implicit in this simplest method is the assumption that the probabilistic mechanism of failure changes over time only through the time since entry into the study. However, the life-table method does not (without modification) allow survival studies with staggered entry to be monitored repeatedly or sequentially over calendar time (*see* **Data and Safety Monitoring**). Occasionally, for reasons of biomedical interpretability and statistical simplicity, the investigator will choose to analyze the data using age or time from onset of some exposure or disease condition as primary time-scale [19], and here also explicit consideration of the staggered exposure times at entry is necessary. In any case, biological age or calendar time of entry can be included as **covariates** or stratifying variables in the statistical analysis (*see* **Stratification**). A crucial assumption in analyzing data with staggered entry is that the prospective time until the primary study endpoint for individual patients can be considered *stochastically* independent of the calendar time of entry, at least conditionally given some other time-scale value such as biological age at entry. In other words, one must assume that there is no (unmodeled) tendency for the patients with worse prognosis to enter the study either systematically earlier or systematically later than the patients with better prognosis.

Mathematically, the special features of a study with staggered entry arise from the presence of at least two time-scales relevant to survival, which bear an unpredictable relationship to one another for individual study subjects. For demographic or epidemiologic population studies – with **cross-sectional**, **cohort**, or mixed designs – the relation between the calendar and individual age time-scales has classically been presented in the form of a **Lexis diagram**, which plots an individual's lifetime as a vector from study entry to failure (or censoring), using calendar time as x-axis and age or study time as y-axis. See Figure 1 for an illustrative Lexis diagram concerning 15 patients from a recent clinical trial, where times are given in months. Keiding [16, 17] gives a historical perspective on Lexis's work, with many references. However, features of *multiple time-scales* exist also in many other types of failure-time data not ordinarily regarded as coming from studies with staggered entry. Other examples of such time-scale pairs include: time from diagnosis and chronological time; time since diagnosis and time since onset of some related risk factor or exposure; and chronological time and cumulative exposure. In studies of the reliability of manufactured components, times to failure collected chronologically are often also studied in terms of the *operational time* [3, 32], which can for example be the cumulative time under significant loading, or the CPU time in a computer setting.

The random data observed in a (singly right-censored) staggered entry study are generally of the form $(T, \delta, (\mathbf{Z}(t), 0 \leq t \leq T))$, where T denotes the time for a single study subject from a specified time

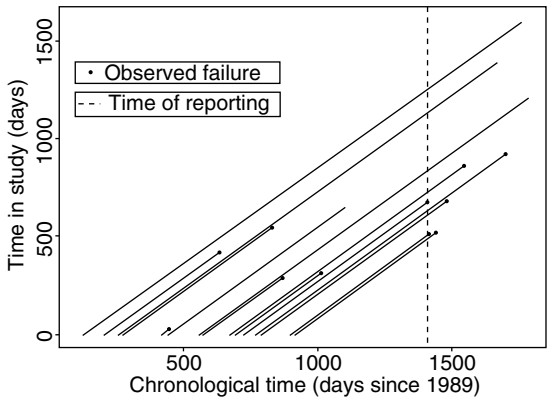

Figure 1 Lexis diagram drawn for 15 selected patients from a clinical trial. Entry time in days from the beginning of 1990 for each study subject is displayed as the initial x-coordinate for a 45° segment. At the upper-right endpoint of the segment, a dot is plotted if an observed failure occurred; otherwise, that subject's failure time was right-censored due either to loss to follow-up or to final reporting (at the time, equal to 1400, marked by the vertical dashed line) of the study results

origin (*study entry*) until the earlier of the primary endpoint and loss to follow-up; δ is the indicator variable for the event that the primary endpoint precedes loss to follow-up; and $\mathbf{Z}(t)$ denotes a vector process of *covariates*, some of which may be independent of time, observable at time $t \leq T$ after entry. There is often at least one covariate-component $Z_k(t)$ affecting survival which measures a cumulative time variable for the study subject from some time origin α until t time units following study entry. In the standard staggered entry survival study, or epidemiologic study accounting for time from onset of some condition or risk factor which never disappears, $Z_k(t) = A_k + t$ is completely determined by the time from origin α to study entry; but in studies accounting for the cumulative measurement of intermittent exposure or another *operational time*, $Z_k(t)$ may also exhibit random variation for $t > 0$. Although there are large literatures on *general* or *multiple censoring* (see [1, Chapter 3]), and on responses more complex than a single survival endpoint, we restrict attention for the rest of this article to the case described above, of right-censored survival data.

Many of the most important themes in statistical survival analysis – both theoretical and practical – are closely connected to the analysis of survival studies with staggered entry. These include: semiparametric modeling and inference for failure hazards, with a parsimonious choice of time-scale and covariates; analysis of left **truncated** and randomly right-censored data; biased sampling frames which can arise in cross-sectional designs, where some strata of the population are oversampled due to time-dependent and random entry criteria; and sequential monitoring of survival studies. In the next several paragraphs, we briefly describe the most important papers and results connecting these topics with staggered entry survival data.

Failure Models with Modified Time-Scale or Time-Dependent Covariates

Suppose that the observable data $(T, \Delta, E, \mathbf{Z})$ on each study subject consist of an event time T (time from entry until death or censoring), an observed failure indicator Δ, a biological age E at entry, and a vector of possibly time-dependent covariates $\mathbf{Z}$. The conditional failure hazard intensity $\lambda_{\mathbf{Z},E}(t|\mathbf{z}, u)$ at study-time t and entry age u, given $E = u$, $\mathbf{Z} = \mathbf{Z}(t) = \mathbf{z}$, is a general three-variable function, which one could attempt to model further by treating the age-at-entry variable as a continuous time-independent covariate or the age-at-event variable $E + T = t + u$ as a time-dependent covariate. In many clinical studies, where the age-at-entry variable E is found not to be significantly predictive of survival, one can simply drop the dependence of $\lambda_{\mathbf{Z},E}$ on E and restrict attention to the possible relationship between survival measured on the time-since-entry time-scale T and the (usually time-independent) covariates $\mathbf{Z}$.

At least in the absence of other covariates $\mathbf{Z}$, the purpose of the Lexis diagram is to help understand the dependence of λ on the two time variables T, $T + E$. A general nonparametric estimator for the corresponding two-time-variable cumulative hazard function $\Lambda(t, s) = \int_0^t \int_0^s \lambda_0(x, y)\, \mathrm{d}y\, \mathrm{d}x$ is given by McKeague & Utikal [23], who also address under minimal assumptions the basic statistical problem of testing for the presence of the covariate-adjusted relationship between a specified covariate such as *treatment* or *exposure* and the waiting time until the primary endpoint. The methodology of Cox [11], later elaborated by many authors (see [1]), applies directly when $\lambda_{\mathbf{Z},E}(t|\mathbf{z}, u)$ depends either on t alone for both values of $\mathbf{z}$ or on $t + u$ alone, in which cases we call respectively T or $T + E$ the *primary time-scale*. Another common approach [6] is to analyze data taking into account a *cohort effect* via a Cox model stratified upon (5- or 10-year groupings of) age at entry. Alternatively, Oakes [25] discusses the possibility of combining the time variables T and $T + E$ into a single time variable with respect to which survival data could be analyzed by conventional methods. The choice between the age at event, $T + E$, or time since entry, T, as the primary time-scale can be important if, for example, the covariate enters a Cox [11] proportional hazards regression model through a variable such as $\ln(E + T)$ or $(E + T)^b$ for known b. Korn et al. [19] discuss the choice of primary time-scale with special reference to examples where $T + E$ is the better choice. Here "better" means implicitly that the resulting model is simpler or more parsimonious, without the need for a time-dependent covariate involving the secondary time variable. However, it should be noted that the correctness of a "simplest" model with noticeable entry-time-dependent or cohort effects implies that any model without such effects must be misspecified.

Left Truncation and Right Censoring

Suppose that subjects of a survival study are drawn from a population who have some condition or risk factor (either the disease under investigation or some other factor), but for whom the time of onset of the condition is not known. (This can occur even in the commonest situation, where study subjects are recruited at the time of diagnosis, within population strata defined by presence or absence of some risk factor(s) at baseline.) Then the survival data are *left-truncated* in the sense that no information is collected on individuals without the condition who might otherwise have been included in the study (*see* **Delayed Entry**). It is well known that left-truncated and right-censored survival data can be analyzed via (a slight modification of) the life-table method [18]. Intuitively, this works because the life-table method infers prospective rates of failure, in terms of time since entry, from the fraction of each *risk group* of individuals under observation at specified time since entry, t, who survive to later times t'.

Random right-censoring is a characteristic feature of survival studies with staggered entry. Clinical studies are often designed to have a fixed calendar duration. If subjects were entered in a cross-sectional cohort, then those subjects without observed primary endpoints have *censoring times* from entry until analysis of the data which are identical to the (nonrandom) duration of the study. By contrast, if entry into the study is staggered, then censoring times are the waiting times from entry until the end of the study. Since recruitment into clinical studies is often assumed to be **uniformly distributed** over a specified period of calendar time, the censoring times would also be uniformly distributed. In this context, the assumption that time of entry is unrelated to survival time is equivalent to the assumption of independence between survival and right-censoring times, without which any estimation of survival-time distributions is problematic [37].

Biased Sampling for Cross-Sectional Data

Cross-sectional recruitment into survival studies at first sight seems to be the polar opposite of staggered entry. Here the subjects are entered at a single calendar time, but their biological ages, times of onset of risk factors and exposures, and times of diagnosis are still unpredictable with respect to one another. Therefore, the chronological and biological time-scales are still staggered with respect to one another, and the same considerations of modeling and analysis and multiple time-scales discussed above come into play. The papers of Keiding [16], Weldon & Potvin [41], Wang [40], and Yang & He [42] discuss various aspects of inference of disease **prevalence** and **incubation periods** as well as distributions for waiting times until medically interesting study endpoints. These papers also explicitly recognize that cross-sectional and some other sampling frames can be *biased* in the sense that study subjects are oversampled (as compared with the general patient population) in some population strata and underrepresented in others. For example, in cross-sectional studies where data on coronary risk factors are collected initially, subjects with long durations between onset of the risk factor and the study endpoint will be overrepresented. Slud & Kopylev [36] study the effect of such biased sampling in inducing dependence between death and censoring times. All of the papers mentioned in this paragraph warn that when covariate risk indicators are left-censored, as naturally happens when data are sampled cross-sectionally, straightforward regression analyses in terms of these covariates can lead to biases.

Sequential and Group-Sequential Analysis

We have described above several ways in which staggered entry affects the modeling of survival data, but staggering plays a direct role in the two-sample inference of treatment effectiveness primarily in repeated or **sequential** statistical survival analyses (*see* **Data and Safety Monitoring**). The mathematically most sophisticated work related to staggered entry has been done under this heading.

Although the early parametric sequential method of Breslow & Haug [5] did allow for random staggered entry over an accrual period, most effort has gone toward extending **nonparametric** statistics, repeated significance tests and (group) sequential procedures to the staggered setting where the calender time parameter used for repeated testing is not the same as the underlying time on test. Early work on the *group-sequential* approach to repeated significance tests in clinical trials [2, 26, 28]

advocated repeated monitoring (with possible early termination) in chronological time. Initially, the necessary asymptotic distribution theory with respect to survival-time statistics such as the **logrank** was available only for trials with *progressive censoring*, in which subjects were accrued simultaneously in a cohort or in which interim analyses would follow the termination of accrual [9, 20]. For the realistic case of trials with staggered entry, Jones & Whitehead [15] proposed – for both the logrank and Gehan-modified Wilcoxon statistics – that the asymptotic distribution theory for statistics repeatedly calculated in chronological time would be given by treating the statistics plotted against the cumulative statistical (Fisher) **information** as a **Brownian motion** process. For the logrank, this was completely substantiated: Tsiatis [39] proved it for the case of finitely many calendar times; then Sellke & Siegmund [30], with cumulative (estimated) statistical information replacing calendar time, and Slud [34], using actual calendar time, showed that the two-time-parameter normalized logrank statistic converges weakly to a continuous Gaussian process with two-dimensional time. For the modified Wilcoxon statistic in a staggered entry setting, Slud & Wei [38] showed that the asymptotic theory differed from that proposed by Jones & Whitehead, and that the repeatedly calculated statistic has dependent increments. By contrast, Slud [34, Corollary 2.4 and Proposition 2.5] (in which the functions $q(u)$ and $L(u)$ should additionally be assumed nonrandom) showed that weighted logrank statistics have uncorrelated increments (which are asymptotically jointly Gaussian, and therefore independent) if the weighting function does not depend upon calendar time.

The asymptotic distribution theory for statistics of clinical trials with staggered entry has required an understanding of the behavior of these statistics as two-time-parameter **stochastic processes** when calculated using all data available up to time on test, s, and chronological time, t. Majumdar & Sen [22] early on had the idea (further developed by Sen [31] and Sinha & Sen [33]) of using a two-time-parameter process, with time on test and number of subjects entered as the two time-scales. They provided the relevant asymptotic distribution theory, based on an assumed target sample size, for the maximum of the logrank statistic calculated at all times and numbers accrued. The more fruitful direction has been to develop the theory of clinical trial statistics as two-time-parameter processes with respect to the time on test and chronological time-scales. Sellke & Siegmund [30] and Slud [34] obtained their results on such processes via one-time-parameter martingale theory. Slud [35] represented (weighted) logrank statistics as stochastic integrals with respect to compensated two-time-parameter counting processes, using results from the theory of two-parameter (strong) martingales (see the next section) to show the weak convergence of a two-time-parameter **Kaplan–Meier estimator**. Gu & Lai [12] gave a general treatment using empirical process theory (in the case of independent and identically distributed (iid) data records for study subjects) of two-parameter weighted logrank and Kaplan–Meier statistics. This work could be used to design sequential and group-sequential clinical trial monitoring schemes based on Kaplan–Meier estimators. The empirical process approach has culminated in recent works by Gu & Ying [13] and Bilias et al. [4] on two-time-parameter analysis of Cox model estimators and **partial likelihood** score statistics under Cox-type semiparametric models with staggered entry, which could be implemented in real clinical trials if model-based group-sequential analyses were of interest.

Mathematical Tools in the Underlying Theory

It remains to describe briefly the main mathematical tools and techniques which have been used to establish the theoretical results surveyed above. To fix ideas, consider the case of iid staggered entry survival data $(E_i, T_i, \Delta_i, Z_i)$ as above, where the **binary** variables Z_i are randomized treatment-group indicators. Define

$$N_i(s,t) = I_{[E_i+T_i\le t, T_i\le s]}\cdot\Delta_i,$$
$$Y_i(s,t) = I_{[E_i+T_i\ge t, T_i\ge s]}.$$

Just as one defines a compensator for counting processes in a single time-scale, one defines a compensator for the two-time-scale process N, namely

$$A_i(s,t) = H(\min(s, C_i, t-E_i)),$$

where H is the cumulative failure hazard, assumed identical for all study subjects. Then the compensated processes $M_i(s,t) = N_i(s,t) - A_i(s,t)$ is a *two-parameter (strong) martingale* in the sense that for all t, $M_i(\cdot, t)$ is a martingale with respect to the filtration $\mathcal{F}_s^{\text{study}}$ generated by all random variables observable up to study time s, and for all fixed s, $M_i(s, \cdot)$ is a martingale with respect to the filtration $\mathcal{F}_t^{\text{chron}}$ generated by all random variables observable up to chronological time t. The two-parameter strong martingale property is a very special one, and holds in the iid staggered entry case essentially because, for each subject, after entry the two time-scales are deterministically related to one another.

Since $M(\cdot, t)$ is the usual compensated failure-counting process based upon data available up to chronological time t, the usual weighted logrank and Kaplan–Meier statistics have well-known representations as stochastic integrals with respect to

$$M(s,t) = \sum_{i=1}^{n} M_i(s,t),$$

$$M^{(1)}(s,t) = \sum_{i=1}^{n} Z_i M_i(s,t).$$

Many of the inequalities and **limit theorems** for one-parameter martingales have immediate extensions to the case of strong two-parameter martingales. The (Doob submartingale) maximal inequality has a counterpart proved by Cairoli [7]; the basic definitions and properties of stochastic integrals are due to Cairoli & Walsh [8]; the Burkholder inequalities have generalizations due to Métraux [24], Ledoux [21], and Chevalier [10]. These inequalities lead [35] to a two-parameter functional **central limit theorem** of the type of McLeish–Rebolledo–Helland [14], which Slud [35] used to obtain a two-time-parameter weak **convergence** theorem for the Kaplan–Meier statistic.

The papers of Gu & Lai cited above, beginning with [12], took a completely different and extremely productive approach to the theory of two-time-parameter weak convergence. They relied on deviation estimates and inequalities from empirical process theory [29]. These, together with functional central limit theorems for processes (derived from) $M(s,t)$, $M^{(1)}(s,t)$ above, carry over naturally to the two-time-parameter setting.

References

[1] Andersen, P., Borgan, Ø., Gill, R. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Armitage, P. (1975). *Sequential Medical Trials*. Wiley, New York.

[3] Barlow, R. & Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing: Probability Models*. To Begin With, Silver Spring.

[4] Bilias, Y., Gu, M.-G. & Ying, Z. (1995). Towards a general asymptotic theory for Cox model with staggered entry. *Preprint*.

[5] Breslow, N. & Haug, C. (1972). Sequential comparison of exponential survival curves, *Journal of the American Statistical Association* **67**, 691–697.

[6] Breslow, N., Lubin, J., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[7] Cairoli, R. (1969). Une inégalité pour martingales à indices multiples et ses applications, in *Séminaire de Probabilité IV, Strasbourg*, P.-A. Meyer, ed. Lecture Notes in Mathematics, Vol. 124. Springer-Verlag, New York, pp. 1–27.

[8] Cairoli, R. & Walsh, J. (1975). Stochastic integrals in the plane, *Acta Mathematica* **134**, 111–183.

[9] Chatterjee, S. & Sen, P. (1973). Nonparametric testing under progressive censorship, *Calcutta Statistical Association Bulletin* **22**, 13–50.

[10] Chevalier, L. (1982). Martingales continues à deux paramètres, *Bulletin des Sciences Mathématiques, Series 2* **106**, 19–62.

[11] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **33**, 187–220.

[12] Gu, M.-G. & Lai, T.-L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials, *Annals of Statistics* **19**, 1403–1433.

[13] Gu, M.-G. & Ying, Z. (1992). Group sequential methods for survival data using partial likelihood score processes with covariate adjustment, *Mathematical Sciences Research Institute Report*. Berkeley.

[14] Helland, I. (1982). Central limit theorems for martingales with discrete or continuous time, *Scandinavian Journal of Statistics* **9**, 79–94.

[15] Jones, D. & Whitehead, J. (1979). Sequential forms of the logrank and modified Wilcoxon tests for censored data, *Biometrika* **66**, 105–113; amendment: *Biometrika* **68**, (1979). 576.

[16] Keiding, N. (1990). Statistical inference in the Lexis diagram, *Philosophical Transactions of the Royal Society of London, Series A* **332**, 487–509.

[17] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[18] Keiding, N. & Gill, R. (1990). Random truncation models and Markov processes, *Annals of Statistics* **18**, 582–602.

[19] Korn, E., Graubard, B. & Midthune, D. (1995). Time-to-event analysis of longitudinal followup of a survey: choice of the time-scale. *Preprint*.

[20] Koziol, J. & Petkau, J. (1978). Sequential testing of the equality of two survival distributions using the modified Savage statistic, *Biometrika* **65**, 615–623.

[21] Ledoux, M. (1981). Inégalités de Burkholder pour martingales indexées par $N \times N$, in *Processus Aléatoires à Deux Indices*, H. Korezlioglu, G. Mazziotto & J. Szpirglas, eds. Lecture Notes in Mathematics, Vol. 863. Springer-Verlag, New York, pp. 122–127.

[22] Majumdar, H. & Sen, P.K. (1978). Nonparametric testing for simple regression under progressive censoring with staggering entry and random withdrawal. *Communications in Statistics – Theory and Methods, Part A* 349–371.

[23] McKeague, I. & Utikal, K. (1990). Inference for a nonlinear counting process regression model, *Annals of Statistics* **18**, 1172–1187.

[24] Métraux, C. (1976). Quelques inégalités pour martingales à paramètres bidimensionnel, in *Séminaire de Probabilité XII, Strasbourg*. C. Dellacherie, P.-A. Meyer & M. Weil, eds. Lecture Notes in Mathematics, Vol. 649. Springer-Verlag, New York, pp. 170–179.

[25] Oakes, D. (1995). Multiple time scales in survival analysis, *Lifetime Data Analysis* **1**, 7–18.

[26] O'Brien, P. & Fleming, T. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.

[27] Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., MacPherson, K., Peto, J. & Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient, II, *British Journal of Cancer* **35**, 1–39.

[28] Pocock, S. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika* **64**, 191–199.

[29] Pollard, D. (1990). *Empirical Processes: Theory and Applications. Regional Conference Series in Probability and Statistics* Vol. 2. Institute of Mathematical Statistics, Hayward.

[30] Sellke, T. & Siegmund, D. (1983). Sequential analysis of the proportional hazards model, *Biometrika* **70**, 315–326.

[31] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.

[32] Singpurwalla, N. (1995). Survival in dynamic environments, *Statistical Science* **10**, 86–103.

[33] Sinha, A. & Sen, P.K. (1982). Tests based on empirical perocesses for progressive censoring schemes with staggering entry and random withdrawal, *Sankhyā, Series B* **44**, 1–18.

[34] Slud, E. (1984). Sequential linear rank tests for two-sample censored survival data, *Annals of Statistics* **12**, 551–571.

[35] Slud, E. (1985). Two-parameter martingales in sequential survival analysis, in *ISI 45th Session Contributed Papers*, Vol. 1. International Statistical Institute, Amsterdam, pp. 231–232.

[36] Slud, E. & Kopylev, L. (1994). Dependent competing risks with time-dependent covariates, in *Lifetime Data: Models in Reliability and Survival Analysis*, N. Jewell, A. Kimber, M.-L. Lee & G. Whitmore, eds. Kluwer, Dordrecht, pp. 323–330.

[37] Slud, E. & Rubinstein, L. (1983). Dependent censoring and summary survival curves, *Biometrika* **70**, 643–649.

[38] Slud, E. & Wei, L.-J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic, *Journal of the American Statistical Association* **77**, 862–868.

[39] Tsiatis, A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis, *Journal of the American Statistical Association* **77**, 855–861.

[40] Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data, *Journal of the American Statistical Association* **86**, 130–143.

[41] Weldon, K. & Potvin, D. (1991). Nonparametric recovery of duration distributions from cross-sectional sample surveys, *Communications in Statistics – Theory and Methods* **20**, 3943–3973.

[42] Yang, G. & He, S. (1994). Estimating lifetime distributions under different sampling plans, in *Statistical Decision Theory and Related Topics V*, S. Gupta & J. Berger, eds. Springer-Verlag, New York, pp. 73–85.

E.V. SLUD

Standard Deviation

The standard deviation of a random variable X is a measure of the variable's spread or dispersion around its mean. The **mean** is the average or expected value of X, and is a measure of the center of its distribution. The mean can also be viewed as a "typical" value of X. By definition, however, the outcome of a **random variable** is unpredictable and will vary from one trial to the next; the standard deviation describes the amount of variation that can be expected around the average value. If the mean of X is represented by $\mathrm{E}(X)$, then its **variance** is defined by

$$\begin{aligned}\mathrm{var}(X) &= \mathrm{E}[X - \mathrm{E}(X)]^2 \\ &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2,\end{aligned}$$

provided that $\mathrm{E}(X)$ exists. The standard deviation is the positive square root of the variance,

$$\mathrm{sd}(X) = [\mathrm{var}(X)]^{1/2}.$$

The mean of a random variable X is often denoted by μ, its variance by σ^2, and the standard deviation by σ.

From the preceding definition, the standard deviation of X is a sort of average of the deviation of X from its mean. If X is a measurable quantity such as length or temperature, then, unlike the variance, the units of measurement for the standard deviation are the same as the units for X. If X is measured in meters, for example, then $\mathrm{sd}(X)$ is measured in meters as well. In general, a large standard deviation indicates that the outcomes of X are widely distributed around its mean, while a small standard deviation means that the outcomes are more homogeneous and cluster tightly around the center.

To calculate the standard deviation of a random variable, it is necessary to know the probability distribution of X. If X is a discrete random variable with mean $\mathrm{E}(X) = \mu$, then

$$\begin{aligned}\mathrm{var}(X) &= \sum_{i=1}^{k} (x_i - \mu)^2 \Pr(X = x_i) \\ &= \left[\sum_{i=1}^{k} x_i^2 \Pr(X = x_i)\right] - \mu^2,\end{aligned}$$

where $x_1, x_2, \ldots x_k$ are all outcomes of X such that $\Pr(X = x_i) > 0$. If X is a continuous random variable with probability density function $f(x)$ and mean $\mathrm{E}(X) = \mu$, then

$$\begin{aligned}\mathrm{var}(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,\mathrm{d}x \\ &= \left[\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x\right] - \mu^2.\end{aligned}$$

In each case, $\mathrm{sd}(X)$ is the square root of the variance.

A linear transformation of the random variable X affects the standard deviation in a straightforward manner. If a and b are constants and if the random variable $Y = aX + b$, then

$$\mathrm{sd}(Y) = |a|\mathrm{sd}(X).$$

There is no variability in a constant.

In practice, the standard deviation of a distribution can be estimated using the information contained in a sample of observations drawn from that distribution. If $x_1, x_2, \ldots x_n$ is a **random sample** of size n selected from a population with mean μ and standard deviation σ, then the sample standard deviation is represented by s and is defined by

$$\begin{aligned}s &= \left[\sum_{i=1}^{n} (x_i - \overline{x})^2 \Big/ (n-1)\right]^{1/2} \\ &= \left[\left(\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n\right) \Big/ (n-1)\right]^{1/2},\end{aligned}$$

where $\overline{x}$ is the sample mean

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Just as σ describes the dispersion of a distribution around its mean μ, s describes the spread of a sample of values around the sample mean $\overline{x}$. It can be thought of as a form of average of the deviations of the observations from the sample mean. While s^2 is an **unbiased** estimator of σ^2 over all possible random samples of size n, meaning that $\mathrm{E}(s^2) = \sigma^2$, s is *not* an unbiased estimator of σ.

Together, the sample mean $\overline{x}$ and standard deviation s are very useful for summarizing a set of measurements. The mean indicates where the observations are centered; the standard deviation quantifies the amount of dispersion around the center. More explicitly, Chebyshev's inequality states

that for any k which is greater than or equal to 1, at least $[1 - (1/k)^2]$ of the measurements lie within k standard deviations of the mean. Given $k = 2$, for example, at least $[1 - (1/2)^2] = 3/4$ of the values lie in the interval $\overline{x} \pm 2s$. Equivalently, it can be said that this interval encompasses at least 75% of the observations in the group. Similarly, the interval $\overline{x} \pm 3s$ contains at least 89% of the measurements. These statements are true no matter what the values of $\overline{x}$ and s, and regardless of the shape of the distribution from which the sample was drawn.

The summary provided by the sample mean and standard deviation can often be made more precise when the shape of the distribution of values is, in fact, known. If the data are symmetric and unimodal, for instance, then approximately 95% of the observations lie in the interval $\overline{x} \pm 2s$, and almost all of the values are contained in $\overline{x} \pm 3s$.

The sample mean and standard deviation can also be used to compare the variability of data sets representing different quantities or different types of measurements. The *coefficient of variation* is defined as

$$CV = 100\% \times \frac{s}{\overline{x}},$$

and is a measure of relative variability. Because s and $\overline{x}$ share the same units of measurement, these units cancel out and leave CV a dimensionless number. Since it is independent of measurement units, the coefficient of variation can be used to compare the amount of dispersion for any two sets of values. A larger coefficient implies that there is more variability among the measurements.

(*See also* **Mean Deviation**; **Moments**)

K. GAUVREAU

Standard Error

The term standard error (se) is used to refer to the **standard deviation** (sd) of an estimator or sample statistic. As such, it is a measure of the estimator's variability around its expected value (**expectation**) or **mean**.

More specifically, standard error often refers to the standard deviation of the sample mean $\overline{x}$. If $x_1, x_2, \ldots x_n$ is a **random sample** of size n selected from a population with mean μ and standard deviation σ, then the sample mean is defined by

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The sample mean can be used to estimate the true population mean μ. This statistic is itself a **random variable**; its value is unpredictable and will vary from one sample to the next. This variability or dispersion can be characterized by the standard deviation of the estimator $\overline{x}$, where

$$\begin{aligned} \mathrm{sd}(\overline{x}) &= \mathrm{se}(\overline{x}) \\ &= \frac{\sigma}{\sqrt{n}}. \end{aligned}$$

To illustrate this idea, suppose that we were to select repeated samples of size n from the underlying population with mean μ and standard deviation σ and calculate the mean of each one. We would end up with a set of sample means $\overline{x}_1, \overline{x}_2, \overline{x}_3$, etc. Each of these means can be treated as a unique observation; their collective distribution is called a **sampling distribution**. While σ measures the standard deviation of the original population and tells us how much variability to expect among the individual observations, $\mathrm{se}(\overline{x}) = \sigma/\sqrt{n}$ measures the standard deviation of the sampling distribution and tells us how much variability to expect among the means. Although the standard error is related to the population standard deviation σ, there is less variability among the sample means than there is among the individual observations. Even if a particular sample contains one or two extreme values, it is likely that these values would be offset by the other measurements in the group. As n increases, the amount of sampling variation – and thus the dispersion among the means – decreases.

The standard error of the sample mean $\overline{x}$ can be estimated by substituting the sample standard deviation s for σ, where

$$s = \left[\left(\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \right) \right]^{1/2}.$$

Therefore,

$$\widehat{\mathrm{se}}(\overline{x}) = \frac{s}{\sqrt{n}}.$$

The estimated standard error plays an important role in statistical inference on a population mean; it is vital to the construction of **confidence intervals** and the performance of **hypothesis tests**.

K. GAUVREAU

Standard Gamble Technique

In recent years the concept of **utility** has been introduced to both clinical decision making (i.e. decisions regarding the best course of action for a patient; *see* **Decision Analysis in Diagnosis and Treatment Choice**) and **program evaluation** (i.e. the best way of using available health care resources or economic evaluation of health care programs). The Standard Gamble (SG) technique is a classic method of measuring an individual's utility (i.e. preferences) under uncertainty. The measurement of individuals' preferences under uncertainty is important because decisions about health interventions at both the individual and the community levels (stemming from the unique nature of health as a nontransferable goal) are made under uncertainty [2]. It is used to measure von Neumann–Morgenstern (vNM) utility functions [20] over life-years and health states, preference weights to be used in the QALY (quality-adjusted life-years) calculations (*see* **Quality of Life and Health Status**) and the healthy years equivalent (HYEs).

The SG Technique

The SG technique is a lottery-based technique where the respondent is asked to indicate a state of indifference when comparing two lotteries or a lottery to a sure thing. The SG question is a general question. However, to be able to interpret the answers we need an underlying theory [20]. The most common one is the vNM utility theory. The SG in the context of a vNM type individual will be the focus of this article. For a review of this method and its application in general see [4], and for its use in health care see [19] or [6]. For simplicity, but without loss of generality, most of this article deals with the case of estimating a preference value for a chronic health state. In this situation an individual has a particular number of remaining life years in a given constant health state.

We use the following notations and definitions. Let Q and T denote two attributes of the outcome of concern ($Q =$ the health state of the individual, $T =$ remaining life years). Let FH represent the state of full health and D death ($D \leq Q \leq FH$). Let $U(Q, T)$ be a vNM utility function that describes the utility of being in a given health state, Q, starting now, for a period of T years, followed by death, as viewed now by the individual. For the case of chronic health state the SG technique is applied as follows (see also Figure 1):

> The subject is offered two alternatives. Alternative 1 is a treatment with two possible outcomes: either the patient is returned to normal health and lives for an additional T years (probability p), or the patient dies immediately (probability $1 - p$). Alternative 2 has the certain outcome of chronic state I (Q using the above notation) for life (T years). Probability p is varied until the respondent is indifferent between the two alternatives at which point the required preference value for state I (Q using the above notation) is simply p^* [19, p. 20].

Props and visual aids are recommended for use to help respondents understand the task.

Using the notation defined above it can be shown that $U(Q, T)$, the preference value of living T years in health state Q, is equal to p^*. More specifically, using the vNM utility theory (also known as expected utility theory) at the indifference point, the following relation holds: $U(Q, T) = p^*U(FH, T) + (1 - p^*)U(D, T)$, where p^* is the indifference probability. Denoting $U(FH, T) = 1.0$ and $U(D, T) = 0.0$, we have $U(Q, T) = p^*$.

An important question is whether one can measure the utility of a health state for a unit of time, say one year [i.e. $U(Q, 1)$]. Gafni [6] deals with this question extensively. In brief, even though in theory this is an option, it seems that in practice it will not work. For example, we can ask individuals to imagine living one year, followed by death, in each of the outcome alternatives. This is likely to be seen as threatening by participants, especially when they are evaluating less severe health states and are not very old. A common problem in many studies is that they do not incorporate any explicit statement about

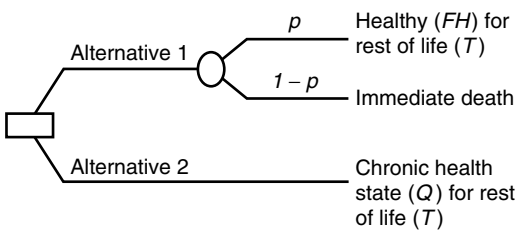

Figure 1 The Standard Gamble method for eliciting utilities for a chronic health state, Q, for rest of life T

the time period spent in the health state and what will happen after this period is over [6]. If this is not done one does not know what the respondent assumed about the length of time spent in the health state and the future states when responding to the SG question. If individuals assume different durations and future states, then their responses will embody an additional component of variation that cannot be tested out.

Interpreting the Scores

The health of an individual is unlike many other outcomes studied in **decision theory** or economics. For example, the health of an individual has a time aspect inextricably bound to it. Thus, one cannot measure only the preference value attributed by individuals to different health states while ignoring the time spent in this health state. In this sense, health is a two-dimensional phenomenon. In particular, Gafni & Torrance [5] argue that the utility (or preference value) for additional time in a given health state – measured using the SG method – depends on a quantity effect (related to the duration in the state), a time effect (reflecting time preference), and gambling effect (reflecting risk attitude). Recently, Gafni [7] added another factor – sequence effect (the order of bad and good events). Thus, observed values of $U(Q, T)$ already embody the individual's risk attitude, time preference pattern, attitudes toward additional quantities of life, and the sequences of events.

As explained, the SG technique provides an individual's preference score for living in a given health state for a given period of time, that is, $U(Q, T)$. Yet many researchers interpret this value as a "timeless" one. They define a general utility scale where a score of 1.0 represents a normal or "healthy" state and 0.0 represents the health state dead. Interpreting the values measured as "timeless" allows the development of tables in which different health states are organized in declining order from healthy (1.0) to death (0.0) (and even states worse than death), regardless of the time spent in each health state and assuming that the preference value attributed to a health state is independent of the sequence of other health states in the individual's lifetime health profile.

Another way to interpret these values is to assume, for example, that a "constant proportional trade-off" exists between quality of life (Q) and time (T), which has been shown to be a prerequisite for using the SG scores as weights in QALYs calculations (see [17]). Under this assumption, the proportion of remaining life that one would trade for a specific quality improvement is independent of the amount of remaining life. For the case of a lifetime health profile (i.e. the case where the individual can experience different health states during his or her lifetime) an additional assumption is required – that in the person's preferences, qualities of life at different times are strongly separable [3, 16]. In other words, the person's preferences about the quality of his or her life in any particular point in time are independent of the qualities of his or her life in other years.

It is important to emphasize that the constant proportional trade-off assumption and the strong separability assumption are additional assumptions to those underlying the theoretical foundation of expected utility theory (or vNM utility theory). In other words, an individual can be an expected utility maximizer without following these particular assumptions. Indeed, many who have invoked these assumptions admit that they are very restrictive and unlikely to represent individuals' behavior. Furthermore, these assumptions have no normative appeal (i.e. they do not reflect the discipline view of the world regarding how an individual should behave) nor are they supported by empirical evidence [3, 9]. Hence in treating preferences for health states as "timeless" we are thus at risk of misrepresenting the actual preferences of people (see, for example, [6]).

Aggregation of SG Scores

Until now we have concentrated on the measurement of an individual's preferences. For program evaluation (i.e. cost–utility analysis) we need a social perspective (*see* **Health Economics**). The current practice is that in cost–utility analysis "the aggregation across subjects is achieved by measuring all individual utilities on the common 0–1 dead–healthy scale and taking the arithmetic mean" [19, p. 17]. This simple (nonweighted) **mean** of individuals' responses is used to derive a "social valuation" of the weights (per unit of time) to be used in QALYs calculations. The question of aggregation of individual utilities and its validity has been

addressed by many authors (for example, [1, 11, 12, 14], and [18]). In this section I do not deal with the question of whether or not it is valid to aggregate individual utilities but assume that it can be done (albeit under a very restrictive set of assumptions).

In this section I question the meaning of aggregating individual preference scores measured using different time horizons, based on the simple arithmetic mean. As explained, each preference score represents the utility of living a period of time in a given health state [i.e. $U(Q, T)$]. For the chronic health state (which we use as an example), the time span is usually defined as the individual's expected life span (*see* **Life Expectancy**). In most health programs individuals who participate have different expected life spans. Thus, the simple aggregation rule implies subscribing to the equity criterion that the reference state of full health for the rest of the individual's life should be treated as of equal value for all individuals. However, as shown by Gafni & Birch [8], the equity criteria used to justify this method of aggregation differ from the one indicated above.

Whether a particular equity criterion is appropriate or not is a subjective issue. However, equity considerations (i.e. the relative values or weights attributed to different individuals or groups) are an intrinsic part of any evaluation. It is thus important that consideration is given to the question of whether the procedure measuring the outcome is consistent with the stated equity criteria. Gafni & Birch [8] deal with this issue extensively and derive adjustment **algorithms** based on the axioms of vNM utility theory, taking into account the different equity criteria adopted in the literature. It is important to emphasize that these adjustment algorithms add to the complexity of the measurement task.

Alternative Uses for the SG Technique

For those who do not want to subscribe to the strong assumptions of the QALY model, two alternatives exist. The first one is to use the SG technique to measure vNM utility scores directly for different potential lifetime health profiles. This can be done by asking one SG question for each lifetime health profile that we want to measure, i.e. using a holistic approach. These values can be used as measures of outcome at the endpoints of a decision tree. We can then calculate the expected utility of each treatment option and choose the one with the highest expected value. The disadvantage of this option is that it creates a communication problem. Expected utility is a theoretical notion that has no direct empirical meaning. In other words, the unit of outcome (i.e. util) has no intuitively appealing meaning to many users (e.g. clinicians or administrators). For example, it will be difficult for them to understand the meaning of cost per util.

Following the need to improve communication (i.e. to preserve the intuitively appealing meaning of the QALY measure) in a way which is consistent with the concept of utility, Mehrez & Gafni [15] suggested an alternative measure – HYE (healthy years equivalent). This measure, based on the theoretical foundation of utility theory, stems directly from the individual's utility function and thus does not require that the individual subscribes to the additional assumptions of the QALY model (or even to the underlying assumptions of the vNM utility model). It only requires that the individual's preferences should be measured under conditions of uncertainty. It combines outcomes of both quality of life (morbidity) and survival (mortality) and thus can serve as a common unit of measure for all programs, allowing comparisons across programs. It preserves the intuitively appealing meaning of the QALY by using years of life in full health as the unit of measurement.

For the case of a decision tree and a vNM type individual, the following two-stage, lottery-based procedure can be used [7, 10, 13]. In stage I, we first use the SG method to measure the utility of all potential lifetime health profiles (i.e. a holistic approach), and secondly calculate the expected utility for each treatment option. In stage II, we "convert" the expected utility of each treatment option to HYE again using the SG method but with a different type of question. In the second SG question the subject is offered two alternatives. Alternative 1 is the same lottery offered in the previous SG question with one change. The probabilities for the two outcomes are now known (i.e. $p = EU$ and $1 - p = 1 - EU$, where $EU =$ expected utility from the treatment). Alternative 2 is living H years in full health. H is varied until the individual is indifferent between the two alternatives. The indifference point H^* defines the certainty equivalent number of years in full health (HYE) that produces utility (i.e. preference value)

equal to the expected utility of the treatment. Note that this procedure illustrates the communication role of the HYE measure. By comparing the expected utility of each treatment option one can determine which treatment is preferable. However, the outcome is presented in units that are difficult to understand (i.e. utils). By adding the second stage we are able to "translate" the results to more meaningful units and thus achieve the goal of communication. For adjustment algorithms that take into account different equity criteria for the case of HYE see Gafni & Birch [8].

Debate in the literature has centered mainly around the theoretical properties of QALYs and HYEs. It seems that the theoretical superiority of the HYE has been established (e.g. in [13]); however, doubts have been raised about the feasibility of measuring HYEs. A recent paper [10] deals with this issue. In brief, the authors discuss the feasibility of measurement using the algorithm described above and the respondent burden in terms of the number and complexity of questions posed. They conclude that HYE will generally involve greater measurement burden than QALYs, but this need not always be restrictive. When the additional measurement burden on subjects is restrictive, they show how the task can be simplified and the measurement burden shared between respondents. Although the number of profiles to be assessed will make HYE infeasible for complex (i.e. very large) decision trees, analysts must view this study design issue in the broader context of the trade-off between precision of the model vs. **bias** when valuing outcome.

References

[1] Arrow, K.J. (1963). *Social Choice and Individual Values*, 2nd Ed. Yale University Press, New Haven.

[2] Ben Zion, U. & Gafni, A. (1983). Evaluation of public investment in health care: Is the risk irrelevant?, *Journal of Health Economics* **2**, 161–165.

[3] Broome, J. (1993). QALYs, *Journal of Public Economics* **50**, 149–167.

[4] Farquhar, P.H. (1984). Utility assessment methods, *Management Science* **30**, 1283–1300.

[5] Gafni, A. & Torrance, G.W. (1984). Risk attitude and time preference in health, *Management Science* **30**, 440–451.

[6] Gafni, A. (1994). The standard gamble method: what is being measured and how it is interpreted, *Health Services Research* **29**, 207–224.

[7] Gafni, A. (1995). Time in health: can we measure individuals' pure time preference?, *Medical Decision Making* **15**, 31–37.

[8] Gafni, A. & Birch, S. (1991). Equity considerations in utility-based measures of health outcomes in economic appraisals: an adjustment algorithm, *Journal of Health Economics* **10**, 329–342.

[9] Gafni, A. & Birch, S. (1995). Preferences for outcomes in economic evaluation: An economic approach to addressing economic problems, *Social Science and Medicine* **40**, 767–776.

[10] Gafni, A., Birch, S. & O'Brien, B. (1995). Healthy Years Equivalents (HYEs) and Decision Trees: a Two Stage, Lottery Based Algorithm, *CHEPA Working Paper Series No. 95-4*. McMaster University, Hamilton.

[11] Harsanyi, J.C. (1955). Cardinal welfare, individualistic ethics and interpersonal comparisons of utility, *Journal of Political Economy* **63**, 309–321.

[12] Harsanyi, J.C. (1975). Nonlinear social welfare economics, *Theory and Decision* **6**, 311–322.

[13] Johannesson, M. (1995). The ranking properties of healthy years equivalents and quality adjusted life years under certainty and uncertainty, *International Journal of Technology Assessment in Health Care* **11**, 40–48.

[14] Kalai, E. & Schmeidler, D. (1977). Aggregation procedure for cardinal preferences: a formulation and proof of Samuelson's impossibility conjecture, *Econometrica* **45**, 1431–1438.

[15] Mehrez, A. & Gafni, A. (1989). Quality-adjusted life-years, utility theory and healthy-years equivalents, *Medical Decision Making* **9**, 142–149.

[16] Mehrez, A. & Gafni, A. (1991). Healthy years equivalents: how to measure them using the standard gamble approach, *Medical Decision Making* **11**, 140–146.

[17] Pliskin, J.S., Shepard, D.S. & Weinstein, M.C. (1980). Utility functions for life-years and health status, *Operations Research* **28**, 206–224.

[18] Sen, A. (1995). Rationality and social choice, *American Economic Review* **85**, 1–24.

[19] Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal: a review, *Journal of Health Economics* **5**, 1–30.

[20] Von Neumann, J. & Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Wiley, New York.

(*See also* **Time Trade-off Technique**)

AMIRAM GAFNI

Standard Normal Deviate

If a **random variable** X follows a **normal distribution**, then the standard normal deviate measures the distance between a particular outcome of X and the **mean** of X in units of the **standard deviation**.

Suppose that X is a normally distributed random variable with arbitrary mean μ and standard deviation σ. While μ represents the center of the distribution of X, σ specifies the amount of dispersion or spread around the mean. Together, these two parameters completely define the shape of the normal curve.

We often wish to evaluate probabilities associated with various outcomes of the random variable X. For instance, if c is a constant, we might wish to know $\Pr(X \leq c)$. One way to determine this is to look up the probability using a table of areas calculated for this specific normal curve. Since a given normal distribution can have an infinite number of values for its mean and standard deviation, however, it is impossible to tabulate the probabilities associated with each and every curve. Instead, only a single curve is tabulated – the special case for which $\mu = 0$ and $\sigma = 1$. This curve is known as the standard normal distribution.

To use the table of the standard normal distribution to look up probabilities associated with a generic normal random variable X which has mean μ and standard deviation σ, we must rescale X to have mean 0 and standard deviation 1. The rescaled random variable is defined by

$$Z = \frac{(X - \mu)}{\sigma};$$

we take the random variable X, subtract its mean, and divide by its standard deviation. Instead of evaluating $\Pr(X \leq c)$ directly, therefore, we would evaluate

$$\begin{aligned} \Pr(X \leq c) &= \Pr\left(\frac{X - \mu}{\sigma} \leq \frac{c - \mu}{\sigma}\right) \\ &= \Pr\left(Z \leq \frac{c - \mu}{\sigma}\right). \end{aligned}$$

An outcome of this rescaled random variable – $(c - \mu)/\sigma$ in the example above – is known as a standard normal deviate or a z-score. Unlike the original outcome c, it does not have any units of measurement. It tells us how far the value c lies from the mean μ, measured in standard deviations. If $Z = -2$, for instance, then c must be two standard deviations below the mean.

K. GAUVREAU

Standardization Methods

Standardization methods are used to adjust for the effects of age and sex, and possibly other factors, in the comparison of disease **rates** between two or more populations. In what follows, adjustment for age will be described, but all the methods can be extended to adjust for other factors, such as sex.

Standardization methods have a long history, and rank among the earliest statistical tools developed. Keiding [21] has traced their origins to eighteenth century actuarial mathematicians (*see* **Actuarial Methods**), though they were reinvented a century later by Neison and **Farr**. Neison was a famous statistician of his day, writing regularly in the *Journal of the Statistical Society* on a wide variety of subjects. Farr was a government official who worked as the "compiler of abstracts" in the Office of the Registrar General for England and Wales from 1839 to 1880. These two eminent men recognized that the comparisons of crude death rates (*see* **Vital Statistics, Overview**) were not sufficient for examining mortality patterns over time (*see* **Morbidity and Mortality, Changing Patterns in the Twentieth Century**), or between geographic areas (*see* **Geographic Patterns of Disease**; **Mortality, International Comparisons**). They also showed that the **average age at death** was not an appropriate index for assessing differences in mortality [25].

In 1841, Farr published age-specific death rates and compared them to rates for the previous three years to show how the pattern of mortality had changed (Registrar General 1841; see [37]). Examination of age-specific rates (usually stratified by sex as well) is widely considered to be the most comprehensive way of comparing disease rates across populations. However, when many populations and types of disease are to be studied, the number of individual rates requiring scrutiny, rapidly becomes awkwardly large. A further summarization of the data is therefore required.

Farr introduced the idea of an external standard population, against which other populations could be compared (Registrar General, 1853; see [37]). His standard was the so-called "healthy counties" in England and Wales. He calculated a set of standard death rates for these counties against which those for other counties could be compared. He then took each of the age-specific rates in the "healthy counties" and multiplied them by the numbers of people of comparable age in the county of interest. In this way he derived an **expected number of deaths** in each age group.

This was not an entirely new method, as Neison had performed similar calculations on rates from two areas of London to prove that the method of comparing average ages at death was flawed [25]. Farr, however, went on to sum the age-specific expected deaths to give the total number of deaths in each county that would be expected if the mortality was the same as in the "healthy counties". The expected number could be compared with the observed number to assess how each county's mortality differed from that in the standard (*see* **Excess Mortality**). Multiplying the ratio of observed to expected deaths by the crude rate in the standard population provided a standardized rate for each county (Registrar General 1857; see [37]). This method is now known as *indirect standardization* and it has remained in widespread use to this day. Since then, other methods have been suggested, but indirect standardization is possibly still the most popular.

Rates and Ratios

Standardized rates, such as those produced by Farr, are expressed as the number of deaths (or cases of disease) per head of population. These can be compared with crude rates in the standard population and are expressed in the same units as normally used for the presentation of rates (e.g. number of deaths per 100 000 population). Possibly more often, however, standardized ratios are quoted. These compare the disease burden (*see* **Burden of Disease**) in the population of interest with that in the standard population. A ratio of 1 therefore indicates that the populations are similar in terms of the disease in question. Often, ratios are presented as percentages by multiplying them by 100, although this convention will not be used here. Some of the methods that will be described do not provide standardized rates *per se*, but multiplying the ratio by the crude rate in the standard population is a way of obtaining an adjusted rate.

Choice of Standard Population

Most methods of standardization require a standard population against which the population of interest

Table 1 Notation

Description	Index population	Standard population
Population in age group i	n_i	N_i
Total population	$n = \sum n_i$	$N = \sum N_i$
Deaths/events in age group i[a]	d_i	D_i
Total number of deaths/events	$d = \sum d_i$	$D = \sum D_i$
Death/event rate in age group i	$r_i = d_i/n_i$	$R_i = D_i/N_i$
Crude death/event rate	$r = d/n$	$R = D/N$
Number of deaths from all causes in age group i	a_i	A_i
Proportion of all deaths due to cause of interest in age group i	$p_i = d_i/a_i$	$P_i = D_i/A_i$
Number of deaths from all causes other than the specific cause of interest in age group i	$s_i = a_i - d_i$	$S_i = A_i - D_i$
Odds of death from specific cause compared to other causes	$m_i = d_i/s_i$	$M_i = D_i/S_i$
Number of years in age group i	y_i	
Mid-point of ith age group	h_i	

[a]For proportional analyses, this is the numbers of deaths from a specific cause in age group i. For all other indices, this can refer to deaths from all causes or specific causes, or to other disease rates.

(index population) is to be compared. Usually, the choice of standard is fairly obvious. Thus, for example, in trying to summarize age-specific rates for geographic regions within a country, the national population could be used as a standard. When examining rates for a variety of countries, a world population or the population of the appropriate continent would be suitable standards. Frequently, however, the sum of the set of index populations to be examined is used as the standard.

A variety of standard populations have been used in the successive volumes of *Cancer Incidence in Five Continents* [31]. These have included estimated African and European and world populations, and a truncated world population that only includes the ages 35–64 in five year age bands. The reason behind the choice of this unusual population was to avoid the examination of rates being dominated by cancers occurring at older ages; cancers at younger ages may give more clues to etiology than those occurring later in life. The most recent volume on cancer incidence [35] has, however, used only the approximate world population.

The important point to note is that different choices of standard population can give rise to different results. Thus identifying a suitable standard is a prerequisite for applying standardization methods. All standardized measures represent a comparison with a chosen standard population.

Notation

The notation used for the formulas for standardized rates and ratios varies widely. The notation used here is given in Table 1.

Indirect and Direct Standardization

Indirect and direct standardization are the two most widely used methods for standardizing rates. Other methods have been proposed, but have not achieved the same popularity.

Indirect Standardization

The information required for use of the indirect method is as follows:

1. age-specific rates in a standard population;
2. the size of the index population in each age group; and
3. the total number of deaths (or cases of disease) in the index population.

The formula for the indirectly standardized ratio is

$$\frac{d}{\sum n_i R_i}.$$

Such ratios are widely known as Standardized Mortality Ratios (SMR) when deaths have been studied. Similar names are adopted for morbidity, such as Standardized Incidence Ratios for cancer **incidence rates**. An alternate way of considering an indirectly standardized ratio is as a ratio of the observed number of events to the number expected in the index population on the basis of standard rates; in other words,

$$\text{SMR} = \frac{d}{e}, \quad \text{where } e = \sum n_i R_i.$$

One can then obtain the indirectly standardized rate by multiplying the ratio by the crude rate in the standard population:

$$\frac{dR}{\sum n_i R_i}.$$

Direct Standardization

A challenge to the indirect method of standardization came in 1883 from within the Registrar General's Office (Registrar General, 1883; see [37]). Ogle proposed the use of what is now known as the *direct method of standardization*. The method can be considered as the opposite of indirect standardization. The type of information required on the standard population for the indirect method is required for the index population in the direct method and vice versa. Thus the information needed for calculating a directly standardized ratio is:

1. age-specific rates in the index population;
2. the size of the standard population in each age group; and
3. the total number of deaths (or cases of disease) in the standard population.

The formula for the directly standardized ratio, usually termed the Comparative Mortality Figure (CMF) when deaths are being considered, is as follows:

$$\frac{\sum N_i r_i}{D}.$$

Analogous to the formula for the SMR, the CMF can be expressed as a ratio of the expected deaths in the standard population on the basis of index rates to the total number of deaths in the standard population; in other words,

$$\text{CMF} = \frac{E}{D}, \quad \text{where } E = \sum N_i r_i.$$

Multiplying by the crude rate in the standard population gives the standardized rate as follows:

$$\frac{R \sum N_i r_i}{D} = \frac{\sum N_i r_i}{N},$$

since $D/R = N$.

Discussion of Direct and Indirect Methods

The direct method is often advocated as the ideal, because it preserves consistency between different index populations. Thus if each age-specific rate in one index population is greater than the rate for the same age group in another index population, then the standardized rate in the former should be greater than in the latter. This is not necessarily true for indirect standardization. When a large number of index populations are compared to the same standard, the consistency property is important. Often, the standardized rates will be compared between index populations to make statements about differences between their disease rates. A method that may fail to preserve consistency could give rise to misleading conclusions about the disease burdens in different populations. However, it is hard to find examples in practice in which serious problems of this nature have arisen.

Direct standardization can be useful in a situation in which disease rates in the appropriate standard population are unavailable. For example, the *Cancer Incidence in Five Continents* volumes as mentioned above [31, 35] have used approximate world, African, and European populations. Since cancer registration is patchy worldwide, world cancer incidence rates are unknown. Estimating world rates by summing the numbers of cancers and population sizes in those countries with data would under-represent the cancer burden in Africa, for example. It is much easier to derive an approximate population distribution by age for the world than to estimate world cancer rates. The actual numbers in each age group need not be world figures as long as the ratios between different age groups are approximately correct. Once standard population numbers are available, directly standardized rates can be produced from the age-specific rates from the countries of interest.

Conversely, if age-specific rates in the index population are unavailable, the direct method cannot be used. It is rare to know the total number of cases of disease but not their ages. However, in such a situation, provided that the age distribution of the population is known, then the indirect method could be used.

One might wonder why the indirect method is used so widely, and indeed this question is still a subject of debate. The argument for the indirect method is that it is more stable when studying rates based on small numbers of deaths. If the age-specific rates in the index population are zero for a number of age groups, then the directly standardized rate or ratio is poorly estimated and can have a large **standard error**. Indeed, the SMR generally has a lower standard error than the CMF, no doubt in part because it is the first approximation to the **maximum likelihood** estimate of the index under the assumption that the number of deaths follows a **Poisson distribution**. These factors have led to the indirect method being widely used, particularly in Britain, for analyses of small geographic areas or occupational groups (see, for example, Office of Population Censuses and Surveys, 1986, 1990 [32, 33]; and *see* **Geographic Patterns of Disease**; **Occupational Mortality**; **Small Area Variation Analysis**). Breslow & Day [2] have pointed out that the SMR is preferred for the analysis of **cross-sectional** data according to **birth cohort** rather than calendar period. This is because the age intervals for which age-specific rates are available tend to vary for different generations, which precludes calculation of the CMF.

SMRs are widely used in analyses of **cohort studies**. The members of the study (index) population are followed through time, and the numbers of events, such as deaths or cancers, are recorded. **Person–years-at-risk** are calculated for each age group and calendar period, which provide the n_i to be multiplied by the age- and calendar period-specific rates R_i from the standard population (usually national rates). We thus obtain an expected number of events, which is compared with the observed number in the study population. The ratio of the observed to expected numbers gives the SMR. When many different causes of death or cancer sites are to be studied, many of the age-specific rates in the study population are zero and so the direct method is rarely used. Only in very large cohort studies does the use of CMFs become feasible and, even then, only for major **causes of death**.

Other Methods

A wide variety of other methods has been proposed since 1883, when the direct method was advocated. Many have been suggested for the analysis of mortality rates and thus have the word "mortality" in their name. There is no reason, though, why other forms of disease rates should not be summarized using these methods. The formulas for the various rates and ratios are given in Table 2; some methods only provide a standardized rate and no ratio, or vice versa. The origins of these methods and the reasons for them have been reviewed by Inskip et al. [19]. Most of the methods have been suggested in an attempt to circumvent problems identified in the two main methods. Sadly, nothing can circumvent the difficulty that the only reliable way of comparing disease rates is by examining age-specific data.

Few of these methods have become widely used and so they are not discussed in detail here. One that is used regularly, however, is Day's cumulative rate [7]. This has been used in the *Cancer Incidence* volumes since its proposal. It is one of the few methods ever suggested which does not require a standard population, and therefore, perhaps should not be counted as a "standardization" procedure *per se*. The principle is simple, in that it is the sum of the age-specific rates for each year to age 74. Usually, rates are only available in age bands comprising a number of years. The cumulative rate is then obtained by multiplying each age-specific rate by the number of years it spans, before summing them. The resulting rate is an approximation to the cumulative risk of acquiring the disease from birth to age 74. This gives a useful measure of the disease burden in the population and comparisons can readily be made between two populations of interest. The method does, however, assume that there is no other cause of death to be considered, and this would argue against its use for comparing two populations with widely differing all-cause mortality rates (such as comparing rates in Africa with those in Europe). No ratio is usually derived from this method, although, intuitively, two groups can be compared by taking the ratios of their cumulative rates.

Proportional Methods

Problems arise when no reliable estimates of the population at risk are available. Routine occupational

Table 2 Formulas for standardized rates and ratios

Rate		Ratio	
Name	Formula	Name	Formula
Comparative mortality rate	$\frac{1}{2}\sum\left(\frac{n_i}{n}+\frac{N_i}{N}\right)r_i$	Comparative mortality index [39]	$\frac{\sum(n_i/n+N_i/N)r_i}{\sum(n_i/n+N_i/N)R_i}$
Equivalent average death rate [44]	$\frac{\sum y_i r_i}{\sum y_i}$	Yule's index [44]	$\frac{\sum y_i r_i}{\sum y_i R_i}$
Cumulative rate [7]	$\sum y_i r_i$	Yerushalmy's relative mortality index [43]	$\frac{\sum y_i r_i/R_i}{\sum y_i}$
		Liddell's relative mortality index [24]	$\sum(N_i r_i/N R_i)$
		Relative risk index [26]	$\sum\frac{N_i n_i r_i}{(N_i+n_i)R_i}\Big/\sum\frac{N_i n_i}{N_i+n_i}$
		Kerridge's inverse method [22]	$\sum(d_i/n R_i)$
		Fisher's Ideal Index [11]	$\left(\frac{d}{\sum R_i n_i}\times\frac{\sum r_i N_i}{D}\right)^{1/2}$

mortality analyses usually suffer from this problem. This is because there are differences in the questions asked about a person's occupation in censuses and those asked of informants of a death. Indeed, the person notifying the death may be unable to give as accurate a description of the occupation as the deceased would have provided on a census form. While SMRs have been widely used for occupational mortality analyses, their weaknesses have to be acknowledged, and they are often **biased** (*see* **Occupational Epidemiology**; **Occupational Health and Medicine**).

A different approach has been adopted in many analyses of this type. In each age group, the population size in each age group is replaced by the number of all-cause deaths. Thus the rates are replaced by the proportions of all deaths due to the cause of interest. A method analogous to that for the SMR is then used to provide a **Proportional Mortality Ratio (PMR)**:

$$\frac{d}{\sum a_i P_i}.$$

Analyses of proportional mortality (although not as ratios) have a longer history than other standardization methods. As far back as 1662, **John Graunt** [16] considered the proportion of deaths due to different causes in order to assess the importance of different diseases in leading to death. However, it was not until the twentieth century that proportional methods became popular in a variety of contexts, particularly occupational analyses. The analysis of occupational mortality in England and Wales for 1931 (Registrar General, 1938; see [38]) gave some proportions of deaths due to the cause of interest, but it was not until the comparable report for 1961 (Registrar General, 1971; see [38]) that the ratios were given. They have been used ever since in the analysis of occupational mortality for England and Wales, but it was only in the latest report that they have been used exclusively (Office of Population Censuses and Surveys, 1995 [34]). Indeed, the latest volume also describes cancer incidence data by occupation, again using proportional measures, but these are Proportional Incidence Ratios (PIR), with all types of incident cancer forming the denominator of the proportions, rather than all deaths.

When suitable populations at risk are unavailable, proportional methods have to be used. Analysis of all-cause mortality does, however, become impossible, as the ratios take the value of unity. Criticisms of the method focus on the problems of bias. If the PMR is used as a proxy for the SMR, it will be biased upward when the all-cause SMR is low (and vice versa). Kupper et al. [23] and Decouflé et al. [8] have discussed the relationship between the PMR and the SMR; the PMR is approximately equal to the ratio of the cause-specific SMR to the all-cause SMR. Kupper et al. [23] termed this ratio the relative SMR (RSMR). In the absence of standardization for any factors, the RSMR and PMR are identical. Since the aim is to standardize, this is unhelpful, but empirical studies have shown that the PMR is a useful proxy for the RSMR [40]. If one is only interested in disease rates in comparison with the standard population, this presents problems for the analysis of groups with very low or high all-cause mortality. However, changes in the distribution of disease within groups should not be ignored, and so the PMR is of value in its own right. PMRs may well lead to useful etiologic clues, particularly in occupational groups with low overall mortality. In such groups, diseases with rates comparable to those in the standard population would be missed by an SMR analysis but would be identified by an elevated PMR.

PMRs can also be biased by abnormally low or high mortality from causes other than that of interest. This problem has been examined by McDowall [29]. He pointed out that it is only the largest causes that seriously influence the PMR for other causes. This led him to suggest recalculations of the PMR, successively excluding the major causes of death from both the standard and index proportions.

The method of calculation of the PMR is similar to that for the SMR, as it employs an indirect standardization approach, albeit of proportions instead of rates. In 1983, Zeighami & Morris [45] proposed an alternative to the PMR which is analogous to a direct standardization method, the formula being

$$\frac{\sum A_i p_i}{D},$$

but this does not appear to have been widely used.

Mortality Odds Ratio

A different approach to proportional mortality analyses was proposed by Miettinen & Wang [30]. Their ratio is equivalent to the **odds ratio** used in the analysis of **case–control studies**. The "cases" are deaths from the cause of interest and the "controls" are deaths from all other causes. "Exposure" is then membership of the study group of interest (e.g. a particular occupational group, or residence in a specific geographic area), and the "unexposed" are all those not in the group of interest and form the "standard" for this method. The formula for the Mortality Odds Ratio is

$$\frac{d}{\sum s_i M_i}.$$

This index is attractive, as it can be interpreted in a similar way to a case–control study, although the choice of other deaths as controls is not necessarily ideal. It is straightforward to show that the unadjusted PMR is always more conservative than the MOR. When such methods are being used for screening large amounts of data, such as in routine occupational mortality statistics, many **false positives** are identified. The use of a more conservative index may be an advantage. With the current speed of computers, the fact that the PMR is simpler to calculate is a minor point, but may still be a consideration if large data sets are to be analyzed (*see* **Proportional Mortality Study**).

Person-Years-of-Life Lost

One concern about most standardization methods is that they give most weight to the age groups that contain the largest numbers of events. In most mortality or morbidity analyses, the elderly therefore receive most emphasis. Restriction of the age groups under study can help, and indeed Yule's method [44] (see Table 2) and Day's cumulative rate [7] require an upper age limit for their calculation. However, for certain analyses, deaths occurring at younger ages may be of greatest interest.

In the early 1950s, there was considerable discussion about this problem [10, 17, 27, 28]. Whether examining changes in mortality over time, or comparing occupational groups, it is often of interest to know whether there are differences at younger ages that are missed by analyses dominated by many events among the elderly. Haenszel [17] loosely defined **person–years of life lost** as "the total number of years lost through the failure of individuals to live some allotted life span", and pointed out that

working-years lost "refer to those falling between the productive ages between 20 and 65". He went on to point out that "it has long been recognized that a count of deaths alone did not give a complete picture of mortality, and measures have been sought which would make some allowance for the widely held intuitive idea that a death at age 70, for example, does not represent as great a loss to society as death at age 35".

While standardized rates of years of life lost can be calculated and used for comparison of groups, ratios are usually more readily understood. To obtain either rates or ratios, the deaths in each age group in the index and standard population are multiplied by the years of life lost in each age group.

Two different forms of years of life lost factors have been suggested. The first is simply to choose an upper age limit of interest and subtract from it the mid-age of each age group. Thus, if years to age 70 were of interest, the years-of-life lost in age group 35–39 would be 33. The formula for the standardized ratio is therefore

$$\frac{\sum d_i(70 - h_i)}{\sum n_i R_i(70 - h_i)}$$

This is equivalent to a weighted indirectly standardized ratio, and by analogy the directly standardized form is

$$\frac{\sum N_i r_i(70 - h_i)}{\sum D_i(70 - h_i)}.$$

Upper age limits other than 70 can be used and, indeed, a lower age limit such as 15 or 20 for occupational mortality can be incorporated by ignoring deaths in childhood.

The other form of weights to represent years of life lost are obtained from **life table** estimates of **life expectancy**. The number of years that a person at the mid-age of each age group can be expected to live is estimated from life tables derived from death rates for the standard population. These weights can then be used in the above formulas replacing $70 - h_i$.

Variances and Standard Errors

In estimating **standard errors**, we usually assume that the standard rates and populations are stable and their sampling errors can be ignored. This is not always true, but, rightly or wrongly, the assumption is usually made. We also assume that the populations in the index population are fixed. Therefore, the only **random variables** to consider are the age-specific rates in the index population, or, more simply, the numbers of events in each age group in the index population. The events in each age group are also assumed to be independent of each other.

It is worth noting that almost all the formulas for standardized rates and ratios can be written as a weighted sum of the age-specific rates in the index population:

$$\sum w_i r_i,$$

where the w_i are the weights.

Thus, for the directly standardized rate, the weights are

$$w_i = \frac{N_i}{N},$$

while for the indirect method they are

$$w_i = \frac{n_i R}{\sum n_i R_i}.$$

Similar formulas exist for the ratios. Most of the ratios can be written as a weighted average of the age-specific ratios in the form

$$\frac{\sum (w_i r_i / R_i)}{\sum w_i}$$

or, sometimes more conveniently, as a ratio of the weighted age-specific rates:

$$\frac{\sum w_i r_i}{\sum w_i R_i}.$$

Using the first form, the weights for the Comparative Mortality Figure (directly standardized ratio) are D_i and for Standardized Mortality Ratio (indirectly standardized ratio) are $n_i R_i$. All the other rates and ratios in Table 2 can be written in this form with differing values of w_i, the only exception being Fisher's Ideal Index [11].

The next step is therefore to consider the standard error of the r_i (or d_i). There are two approaches to this.

Use of the Binomial Distribution

Chiang [6] developed an approach based on the **binomial distribution**. He noted that rates r_i are not proportions, but derived a formula for each rate as a function of the proportions of deaths d_i in the hypothetical population from which the deaths were drawn. This requires knowing the distribution of the ages of the events within each age group. As an approximation it is reasonable to assume that all events occur in the middle of the age range. Chiang's formula for the **variance** of the r_i then becomes

$$\frac{r_i(2 - y_i r_i)}{n_i(2 + y_i r_i)}.$$

This leads to the variance of the standardized rate being

$$\sum w_i^2 \text{var}(r_i),$$

and that of the ratio being

$$\frac{\sum \left[\frac{w_i^2}{R_i^2} \text{var}(r_i) \right]}{\left(\sum w_i \right)^2}.$$

The standard errors are then the square roots of the variances.

Use of the Poisson Distribution

The alternate approach to estimating the standard errors is to assume that the numbers of events in each age group follow a **Poisson distribution**. (It is worth nothing, however, that this assumption may not be valid and that extra-Poisson variability (*see* **Overdispersion**) may need to be investigated [3]. The variance of a Poisson variable is equal to its **expectation**, for which the observed number of events is the best approximation available. The denominators of the rates (the n_i) can be absorbed into the weights and so the formulas for the variances of the rates and ratios become

$$\frac{\sum w_i^2 r_i}{n_i}$$

and

$$\frac{\sum \frac{w_i^2 r_i}{R_i^2 n_i}}{\left(\sum w_i \right)^2},$$

respectively.

Table 3 Variances of directly and indirectly standardized rates and ratios

Method	Rate	Ratio
Direct	$\sum(N_i^2 r_i / N^2 n_i)$	$\sum(N_i^2 r_i / D^2 n_i)$
Indirect	$\frac{\sum R_i^2 d_i}{\left(\sum n_i R_i\right)^2}$	$\frac{d}{\left(\sum n_i R_i\right)^2} = \frac{d}{e^2}$

Note that when SMRs are under discussion alone, the variance is usually given as O/E^2, although the use of upper case letters and O for the total number of (observed) deaths is not consistent with the notation used here for index and standard populations.

These formulas reduce to fairly simple forms for direct and indirect standardization, and these are given in Table 3.

The formulas for the standard errors of Proportional Mortality Ratios (indirectly and directly obtained) are similar to those in Table 3. The numbers in the populations are replaced by the numbers of all-cause deaths, and the rates are replaced by the proportions of deaths. However, it is worth noting that the proportions of deaths due to the cause of interest are true proportions, unlike rates, and so the binomial distribution could be used in the derivation of the standard errors.

Confidence Intervals

In deriving **confidence intervals** for rates and ratios, similar assumptions are made as for the estimation of standard errors (se). Again, we have to make assumptions about the distribution of the rates and ratios.

Confidence Intervals for Rates

Confidence intervals for rates can be derived by assuming that the rates follow a **normal distribution**. The method, therefore, is to add and subtract 1.96 times the standard error from the rate. If rates are small, this leads to problems, as negative values can occur. In such cases, it is preferable to consider the standard error of the logarithm of the rate. Using the standard approximation

$$\text{var}(\log x) = \frac{\text{var } x}{x^2},$$

the standard error of the logarithm of a standardized rate can be obtained as

$$\text{se (log(rate))} = \frac{\text{(se (rate))}}{\text{rate}}.$$

Using this, a 95% confidence for the logarithm of the rate can be obtained, and taking exponentials gives the 95% confidence interval for the rate itself.

Such estimates of confidence intervals rely on the adequacy of the normal assumption, either for the rate or its logarithm. The assumption tends to be poor when the rates are low. This is particularly a problem when we are calculating weighted sums of the age-specific rates, each of which may be small. Dobson et al. [9] have addressed this issue. They have discussed a number of alternate methods for estimating the confidence interval for a Poisson parameter, and derived an improved estimate for the confidence interval for a weighted sum of the events in each age group.

The lower point and upper point of the interval are

$$\text{standardized rate} + \left(\frac{V}{d}\right)^{1/2}(d_{\mathrm{L}} - d)$$

and

$$\text{standardized rate} + \left(\frac{V}{d}\right)^{1/2}(d_{\mathrm{U}} - d),$$

where

$$V = \sum w_i^2 r_i,$$

and d_{L} and d_{U} are the lower and upper confidence interval for the total number of observed deaths, d. Various tables of confidence intervals for Poisson variables exist, usually for a number of levels of confidence. Those given by Gardner [14] provide 90%, 95%, and 99% confidence intervals. Alternatively, Dobson et al. [9] provide a list of approximate methods for obtaining d_{L} and d_{U}.

Recently, there has been considerable research into improved methods for obtaining confidence intervals for standardized rates. As Swift [42] has noted, most of these have been computer intensive methods, and none appears to have been used routinely. Swift himself suggested an approximate **bootstrap method** which he compared with other methods using **simulation** studies. It appears that the debate on calculation of confidence intervals has not yet run its course. Since all methods are approximate, a sensible approach might be to produce a number of confidence intervals calculated in different ways to see how they vary.

Confidence Intervals for Ratios

Similar considerations apply to the confidence interval for a ratio. Ratios are decidedly nonnormal and often the logarithm is considered, with its approximate standard error being calculated as described above. The formula for standard error of the logarithm of the directly standardized ratio (CMF) is

$$\frac{\left(\sum N_i^2 r_i / n_i\right)^{1/2}}{\sum N_i r_i},$$

whereas that for the logarithm of the indirectly standardized ratio (SMR) reduces to

$$\frac{1}{\sqrt{d}}.$$

An alternate method for the calculation of confidence intervals for the SMR (and for the indirectly standardized PMR) requires us to assume that the total number of observed deaths in the index population, d, is a Poisson variable. d_{L} and d_{U}, the lower and upper points of the confidence interval for d, are first obtained from tables or approximations and then the corresponding confidence interval for the SMR is

$$\frac{d_{\mathrm{L}}}{E} - \frac{d_{\mathrm{U}}}{E}.$$

These formulas are now widely used for the calculation of confidence intervals, and it is unusual to derive confidence intervals for SMRs using the standard errors. Computationally intensive methods could also be used, but the above formula is considered appropriate for most needs.

Regression Models

Increasingly, mortality rates are being modeled using **regression** techniques. Keiding [21] discusses some of these approaches at the end of his historical review paper. **Generalized linear modeling** can be used to analyze rates, and a number of papers have explored the issues relating to such analyses [12, 13]. More recently, there has been increasing interest

in **Bayesian** approaches using **Monte Carlo methods** [1] and **generalized estimating equations**.

Recurrent Outcomes

An important new development in the standardization of rates is the consideration of recurrent outcomes. Most of the methods described above have been developed with a single outcome per person in mind, notably death. Even when cancer has been of interest, the number of people with multiple cancers is so small that the cancers have been assumed to be independent. Epidemiology has moved on from there, to deal with outcomes that can occur more than once within an individual. Examples are admission to hospital, episodes of back pain, and attacks of asthma.

Although the standardized rates can be calculated as for nonrecurrent events, the standard errors are larger because of the lack of independence of recurrent events. Glynn et al. [15] used the **negative binomial distribution** to account for departures from the assumption, inherent in the use of the Poisson distribution, that the recurrent events occur randomly. The variance of each age-specific rate in the index population is then

$$\frac{(r_i + r_i^2/k)}{n_i},$$

where k is an index of extra-Poisson variation in the rate, with smaller values of k indicating larger departures from the Poisson distribution. k has to be estimated from the data and Glynn et al.suggest use of the **method of moments** estimator.

The variance of the standardized rate is then

$$\sum (w_i^2)\mathrm{var}(r_i),$$

from which the standard error can be obtained by taking the square root. Again, to obtain confidence intervals, taking logarithms as described above is recommended.

Often one wishes to compare a number of standardized rates (for recurrent or single events). Carriere & Roos [5] have developed a simple test for the comparison of H standardized rates against a standard rate that can be compared with the **chi-square distribution** on H **degrees of freedom**. If S_h is the standardized rate in the hth index population, and R is the crude rate in the standard population, then the test statistic is

$$\frac{\sum (S_h - R)^2}{\mathrm{var}(S_h)},$$

where the summation is over the H populations of interest. If the standard rate is simply the overall rate obtained from the combined index populations, then the test statistic should be compared with the chi-square distribution on $H - 1$ degrees of freedom.

More complex approaches are recommended when the data are available for each individual followed up over time. In this way, the events occurring for each person are known. The analysis of such outcomes requires approaches used for the analysis of **longitudinal data**. Methods mentioned above such as generalized estimating equations, Bayesian approaches using **Markov chain Monte Carlo methods**, and other methods for **multilevel modeling** could be applied. These are, however, computationally intensive and analysis of large data sets can be very time-consuming.

Computation

Despite the long history of standardization methods, few standard statistical packages allow for their use. Many users write their own procedures and link their data to the standard rates. STATA [41] is one package that now incorporates procedures to perform direct and indirect standardization, and Immonen-Räihä et al. [18] published a macro for use in SAS (*see* **Software, Biostatistical**). Spreadsheets are probably the most common computational method used to derive standardized rates/ratios; the columns of the spreadsheet hold the events/population numbers by age for the index and standard populations respectively, and the appropriate calculations performed.

For modeling approaches (*see* **Model, Choice of**) to the analysis of rates, many packages are available and will not be outlined here.

Discussion of Methods

Over the years, many methods of standardization have been proposed. No single method has emerged on top and a variety are in use. Direct and indirect standardization are undoubtedly the most popular, but other methods such as PMRs have to be employed in

certain circumstances. A recent report of occupational mortality in Italy [20] gave a table summarizing occupational mortality analyses from many countries worldwide in recent years. The method used for obtaining standardized ratios in each analysis was listed, and four different methods had been employed. None used a direct method, whereas this is commonly used in cancer studies (see, for example, Parkin et al. [35]) and in many geographic analyses (see, for example, Pickle et al. [36]).

Estimating standard errors and deriving confidence intervals are not straightforward, and many methods are in widespread use. The final verdict has not yet been reached as to which methods are best, and the debate is likely to continue for many years.

As a final note, and to return to where we began, we must be aware that in any standardization procedure we lose something. Much of the debate about which methods to use is due to the fact that no standardized measure can replace the analysis of the age-specific rates themselves. We should understand that summaries can be distorted by patterns in particular age groups. Before one employs any standardization, one should scrutinize the individual age groups. Burack et al. [4] have argued forcibly for examining the age-specific rates, but it has to be acknowledged that in large-scale studies of routine data, even examining the standardized rates or ratios for each subset of the population (such as each occupational group) is an unwieldy task. Scrutiny of the age-specific rates for every group would be impossible. However, perhaps we should take the advice of Burack et al., at least in part, and before commenting on any particular standardized rate or ratio as being particularly high or low we should be more prepared to examine the original age-specific data.

References

[1] Bean, J.A., Wiltse, C.G. & Woolson, R.E. (1987). Small sample behaviour of hypothesis tests related to indirect standardized rates: a Monte Carlo study, *Statistics in Medicine* **6**, 61–70.

[2] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, **II**, *The Design and Analysis of Cohort Studies*. International Association for Research in Cancer, Lyon.

[3] Brillinger, D.R. (1986). The natural variability of vital rates and associated statistics, *Biometrics* **42**, 693–734.

[4] Burack, T.S., Burack, W.R. & Knowlden, N.F. (1983). Cancer II: distortions in standardized rates, *Journal of Occupational Medicine* **25**, 737–744.

[5] Carriere, K.C. & Roos, L.L. (1994). Comparing standardized rates of events, *American Journal of Epidemiology* **140**, 472–482.

[6] Chiang, C.L.C. (1961). Standard error of the age-adjusted death rate, *US Department of Health Education and Welfare, Vital Statistics Special Reports*, Vol. 47, 271–285.

[7] Day, N.E. (1976). Cumulative rate and cumulative risk, in *Cancer Incidence in Five Continents*, Vol. III. J. Waterhouse, C. Muir, P. Correa & J. Powell, eds. International Association for Research in Cancer, Lyon.

[8] Decouflé, P., Thomas, T.L. & Pickle, L.W. (1980). Comparison of the proportionate mortality ratio and standardized mortality ratio risk measures, *American Journal of Epidemiology* **111**, 263–269.

[9] Dobson, A.J., Kuulasmaa, K., Eberle, E. & Scherer, J. (1991). Confidence intervals for weighted sums of Poisson parameters, *Statistics in Medicine* **10**, 457–462.

[10] Doughty, J.H. (1951). Mortality in terms of lost years of life, *Canadian Journal of Public Health* **42**, 134–141.

[11] Fisher, I. (1927). *The Making of Index Numbers*. Houghton Mifflin, Boston.

[12] Frome, E.L. (1983). The analysis of rates using Poisson regression models, *Biometrics* **39**, 665–674.

[13] Gail, M. (1978). The analysis of heterogeneity for indirect standardized mortality ratios, *Journal of the Royal Statistical Society, Series A* **141**, 224–234.

[14] Gardner, M.J. (1989). Tables for the calculation of confidence intervals, in *Statistics with Confidence*, M.J. Gardner & D.G. Altman, eds. BMJ, London, pp. 116–118.

[15] Glynn, R.J., Stukel, T.A., Sharp, S.M., Bubolz, T.S., Freeman, J.L. & Fisher, E.S. (1993). Estimating the variance of standardized rates of recurrent events, with application to hospitalizations among the elderly in New England, *American Journal of Epidemiology* **137**, 776–786.

[16] Graunt, J. (1662). *Natural and Political Observations made upon the Bills of Mortality*. London: re-published by the Johns Hopkins Press, Baltimore (1939).

[17] Haenszel, W. (1950). A standardized rate for mortality defined in units of lost years of life, *American Journal of Public Health* **40**, 17–26.

[18] Immonen-Räihä, P., Hätönen, S., Torppa, J. & Toivanen, A. (1994). A statistical analysis system macro for age-standardized incidence rates, *Computer Methods and Programs in Biomedicine* **44**, 79–83.

[19] Inskip, H., Beral, V. & Fraser, P. (1983). Methods for age-adjustment of rates, *Statistics in Medicine* **2**, 455–466.

[20] Instituto Superiore per la Prevenzione e la Sicurezza del Lavoro (1995). *Mortalità per Professioni in Italia negli Anni '80*. Collana Quaderni ISPESL, Rome.

[21] Keiding, N. (1987). The method of expected number of deaths 1786–1886–1986, *International Statistical Review* **55**, 1–20.

[22] Kerridge, D. (1958). A new method of standardizing death rates, *British Journal of Preventive and Social Medicine* **12**, 154–155.

[23] Kupper, L.L., McMichael, A.J., Symons, M.J. & Most, B.M. (1978). On the utility of proportional mortality analysis, *Journal of Chronic Diseases* **31**, 15–22.

[24] Liddell, F.D.K. (1960). The measurement of occupational mortality, *British Journal of Industrial Medicine* **17**, 228–233.

[25] Lilienfeld, D.E. (1978). "The greening of epidemiology": sanitary physicians and the London epidemiological society (1830–1870), *Bulletin of the History of Medicine* **52**, 503–528.

[26] Lilienfeld, D.E. & Pyne, D.A. (1979). On indices of mortality: deficiencies, validity and alternatives, *Journal of Chronic Diseases* **32**, 463–468.

[27] Logan, W.P.D. & Benjamin, B. (1953). Loss of expected years of life – a perspective view of changes between 1848–72 and 1952, *Monthly Bulletin of the Ministry of Health and the Public Health Laboratory Service* **12**, 244–252.

[28] Martin, W.J. (1951). Life table mortality as a measure of hygiene, *The Medical Officer* **86**, 151–153.

[29] McDowall, M. (1983). Adjusting proportional mortality ratios for the influence of extraneous causes of death, *Statistics in Medicine* **2**, 467–475.

[30] Miettinen, O. & Wang, J.D. (1981). An alternative to the proportionate mortality ratio, *American Journal of Epidemiology* **114**, 144–148.

[31] Muir, C., Waterhouse, J., Mack, T., Powell, J. & Whelan, S., eds. (1987). *Cancer Incidence in Five Continents*, Vol. V. International Association for Research in Cancer, Lyon.

[32] Office of Population Censuses and Surveys (1986). *Occupational Mortality 1979–80, 1982–83. Decennial Supplement*. HMSO, London.

[33] Office of Population Censuses and Surveys (1990). *Mortality and Geography*. HMSO, London.

[34] Office of Population Censuses and Surveys and Health and Safety Executive, (1995). *Occupational Health, Decennial Supplement*. HMSO, London.

[35] Parkin, D.M., Muir, C.S., Whelan, S.L., Gao, Y.T., Ferlay, J. & Powell, J., eds (1992). *Cancer Incidence in Five Continents*, Vol. VI. International Association for Research in Cancer, Lyon.

[36] Pickle, L.W., Mason, T.J., Howard, N., Hoover, R. & Fraumeni, J.F. (1987). *Atlas of U.S. Cancer Mortality among Whites: 1950–1980*. US Department of Health and Human Services (DHSS Publication no. (NIH) 87–2900), Washington, D.C.

[37] Registrar General (1841, 1853, 1857, 1883). *Annual Report of the Registrar General for England and Wales*. HMSO, London.

[38] Registrar General (1938, 1971). *Registrar General's Decennial Supplement on Occupational Mortality, 1931, 1961*. HMSO, London.

[39] Registrar General (1941). *Registrar General's Statistical Review of England and Wales*. HMSO, London.

[40] Roman, E., Beral, V., Inskip, H., McDowall, M. & Adelstein, A. (1984). A comparison of standardized and proportional mortality ratios, *Statistics in Medicine* **3**, 7–14.

[41] StataCorp. (2001). *Stata Statistical Software: Release 7.0*. Stata Corporation, College Station, TX.

[42] Swift, M.B. (1995). Simple confidence intervals for standardized rates based on the approximate bootstrap method, *Statistics in Medicine* **14**, 1875–1888.

[43] Yerushalmy, J. (1951). A mortality index for use in place of the age-adjusted death rate, *American Journal of Public Health* **41**, 907–922.

[44] Yule, G.U. (1934). On some points relating to vital statistics, more especially statistics of occupational mortality, *Journal of the Royal Statistical Society* **97**, 1–84.

[45] Zeighami, E.A. & Morris, M.D. (1983). The measurement and interpretation of proportionate mortality, *American Journal of Epidemiology* **117**, 90–97.

Bibliography

A list of further reading is given below. This consists of articles and books which have not been referenced above but may be of interest to those who want further information. They include some references which date back many years, but which have contributed to the development of current views on standardized methods.

Benjamin, B. (1968). *Health and Vital Statistics*. George Allen & Unwin, London.

Berry, G. (1983). The analysis of mortality by the subject-years method, *Biometrics* **39**, 173–184.

Breslow, N.E. & Day, N.E. (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data, *Journal of Chronic Diseases* **28**, 289–303.

Fox, A.J. & Adelstein, A.M. (1978). Occupational mortality: work or way of life?, *Journal of Epidemiology and Community Health* **32**, 73–78.

Gaffey, W.R. (1976). A critique of the standardized mortality ratio, *Journal of Occupational Medicine* **18**, 157–160.

Hanley, J. & Liddell, D. (1985). Fitting relationships between exposure and standardized mortality ratios, *Journal of Occupational Medicine* **27**, 555–560.

Hickey, R.J., Clelland, R.C. & Clelland, A.B. (1980). Epidemiological studies of chronic disease: maladjustment of observed mortality rates, *American Journal of Public Health* **70**, 142–150.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective, *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

Kilpatrick, S.J. (1962). Occupational mortality indices, *Population Studies* **16**, 175–187.

Kleinman, J.C. (1977). Age-adjusted mortality indexes for small areas: applications to health planning, *American Journal of Public Health* **67**, 834–840.

Liddell, F.D.K. (1979). Excess PYLL for occupational mortality comparisons, *International Journal of Epidemiology* **8**, 185–186.

Liddell, F.D.K. (1984). Simple exact analysis of the standardised mortality ratio, *Journal of Epidemiology and Community Health* **38**, 85–88.

McDowall, M. (1983). William Farr and the study of occupational mortality, *Population Trends* **31**, 12–19.

McMichael, A.J. (1976). Standardized mortality ratios and the "Healthy Worker Effect": scratching beneath the surface, *Journal of Occupational Medicine* **18**, 165–168.

Miettinen, O.S. (1972). Standardization of risk ratios, *American Journal of Hygiene* **6**, 383–388.

Milham, S. (1975). Methods in occupational mortality studies, *Journal of Occupational Medicine* **17**, 581–585.

Milham, S. (1985). Improving occupational standardized proportionate mortality ratio analysis by social class stratification, *American Journal of Epidemiology* **121**, 472–475.

Morris, J.A. & Gardner, M.J. (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates, *British Medical Journal* **296**, 1313–1316.

Osborn, J. (1975). A multiplicative model for the analysis of vital statistics rates, *Applied Statistics* **24**, 75–84.

Redmond, C. & Breslin, P.P. (1975). Comparison of methods for assessing occupational hazards, *Journal of Occupational Medicine* **17**, 313–317.

Rockette, H.E. & Arena, V.C. (1987). Evaluation of the proportionate mortality index in the presence of multiple comparisons, *Statistics in Medicine* **6**, 71–77.

Romeder, J.M. & McWhinnie, J.R. (1977). Potential years of life lost between ages 1 and 70: an indicator of premature mortality for health planning, *International Journal of Epidemiology*, **6**, 143–151.

Stukel, T.A., Glynn, R.J., Fisher, E.S., Sharp, S.M., Lu-Yao, G. & Wennberg, J.E. (1994). Standardized rates of recurrent outcomes, *Statistics in Medicine* **13**, 1781–1791.

Tsai, S.P., Hardy, R.J. & Wen, C.P. (1992). The standardized mortality ratio and life expectancy, *American Journal of Epidemiology* **135**, 824–831.

Tsai, S.P. & Wen, C.P. (1986). A review of methodological issues of the standardized mortality ratio (SMR) in occupational cohort studies, *International Journal of Epidemiology* **15**, 8–21.

Wong, O. (1977). Further criticisms on epidemiological methodology in occupational studies, *Journal of Occupational Medicine* **19**, 220–222.

Wong, O. & Decouflé, P. (1982). Methodological issues involving the standardized mortality ratio and proportionate mortality ratio in occupational studies, *Journal of Occupational Medicine* **24**, 299–304.

HAZEL INSKIP

Standardized Coefficients

A **linear regression** model relates a response variable to a linear predictor

$$\alpha + \beta_1 X_1 + \cdots + \beta_k X_k,$$

where the Xs correspond to **explanatory variables** and the βs are termed the regression coefficients. The regression coefficient β_i represents the change in the linear predictor corresponding to a one unit shift in the variable X_i, if all other variables remain the same. Thus direct comparison of the size of regression coefficients is usually meaningless, since their interpretation depends on the specific explanatory variable.

Standardized regression coefficients represent an attempt to circumvent this problem. If all the explanatory variables are standardized to have a mean of zero and a variance of one, then the coefficients corresponding to these standardized variables are termed standardized regression coefficients. Because unit shifts in these transformed variables are more comparable, then so are the corresponding regression coefficients. Note that the mean standardization is not essential, since it is reflected only in α.

VERN T. FAREWELL

Stationarity

The concept of stationarity of a time series is concerned with the series being in statistical equilibrium; the equilibrium behavior may exhibit a rich variety of features but the features themselves must not be evolving over time. This lack of evolvement is at the heart of stationarity. From practical and visual points of view the behavior of the series, in stretches of sufficient length, must appear statistically identical. A qualification about sufficient length is needed since apparently nonstationary features over the short term are not precluded from a stationary series; they must, however, be nonunique and be a repeatable feature over the long term. This aspect makes stationarity a rather slippery concept to verify in practice. Most analysts would only regard a series as stationary if there were some clear evidence of repeatability in its features. Evidence against stationarity would include cyclic behavior (*see* **Seasonal Time Series**) and the values trending over time; in the latter case, the local level or average of the series would indicate a clear pattern over the length of the series. The nonstationarity might be not in a level, but in the variability about a level; extremes, say, might become more or less frequent as time passes, and, more generally, the distribution of fluctuations of the values could be changing. A more subtle type of behavior ruling out stationarity would be a change in dependency, such as autocorrelation. Other simpler features, such as a sudden and sustained change in level, are absent from stationary series but would present little difficulty in identification. There are no particularly biostatistical aspects to stationarity, or to much of time series analysis in general, most existing methods being readily applicable. One text orientated towards biostatistics is Diggle [4]. Classical texts include Box & Jenkins [1], Priestley [5], Brillinger [2], and Brockwell & Davis [3].

Technical Definitions

Time series are usually measured either at discrete evenly-spaced time-points or continuously over time. This makes a considerable difference to their specification or modeling as **stochastic processes**, but little difference as far as definitions of stationarity are concerned. In definitions, time series are represented by random variables $X(t_1), X(t_2), \ldots, X(t_n)$ at a series of times $(t_1, t_2, \ldots, t_n), n \geq 1$, and the invariance of their joint distribution to a common translation in time is the key aspect of stationarity. Thus, the requirement of the strongest form of stationarity, called *strict* or *full stationarity*, is that the joint distribution of $[X(t_1), X(t_2), \ldots, X(t_n)]$ should be identical to that of $[X(t_1 + t), X(t_2 + t), \ldots, X(t_n + t)]$ for all integers n and all allowable $t, -\infty < t < \infty$. This form of stationarity is often unnecessarily rigorous and would be impossible to investigate practically in its entirety. It is, however, satisfied by independent and identically distributed random variables, although such a time series structure is usually no more than a useful null hypothesis model. Simpler forms of stationarity are employed in practice, beginning with stationarity in mean, which requires that $\mathrm{E}[X(t)]$ does not depend on t. A generalization would be to *marginal stationarity*, in which the marginal distribution of $X(t)$ does not depend on t. The most used form of stationarity, *second-order* or *weak stationarity,* requires that the moments up to the second-order, $\mathrm{E}[X(t)]$, $\mathrm{var}[X(t)]$ and $\mathrm{cov}[X(t_i + t), X(t_j + t)]$, $1 \leq i, j \leq n$, do not depend on translation time.

These definitions, and the main discussion here, are limited to univariate time series; with several time series considered simultaneously, there will be straightforward generalizations involving vector time series quantities.

Stationarity of Linear Time Series Models

In discrete time, $t = 0, \pm 1, \pm 2, \ldots$, a linear time series model is usually given in the form:

$$X(t) = c_0\varepsilon(t) + c_1\varepsilon(t-1) + c_2\varepsilon(t-2) + \cdots,$$

where $\varepsilon(t), t = 0, \pm 1, \pm 2, \ldots$, is an independent and identically distributed sequence of random variables, and $c_0, c_1, c_2, \ldots$ is a sequence of constants. For the model to be strictly stationary, the sequence of constants must be such that $c(x) = \sum_{i=0}^{\infty} c_i x^i$ converges for $|x| \leq 1$. A practically useful class of linear models, having a finite number of parameters, are autoregressive and moving-average processes (*see* **ARMA and ARIMA Models**) of the form

$$X(t) = a_1 X(t-1) + a_2 X(t-2) + \cdots + a_p X(t-p) + \varepsilon(t) + b_1\varepsilon(t-1) + \cdots + b_q\varepsilon(t-q),$$

where $p, q \geq 0$ and $a_1, a_2, \ldots, a_p$ and $b_1, b_2, \ldots, b_q$ are constants. The stationarity condition just mentioned is now equivalent to requiring that the roots of the pth-order polynomial equation,

$$1 - a_1 x - a_2 x^2 - \cdots - a_p x^p = 0,$$

lie outside the unit circle in the complex plane. This type of exposition was first simply set out by Box & Jenkins [1].

Statistical Analysis of Stationarity

Most series will not be stationary, and yet the statistical repeatability implied by stationarity is central to much statistical analysis of time series data. This is because the identification and elimination of nonstationary features, such as trend and seasonality, as often required, is supposed to leave a stationary series. For trend in mean, differencing of successive data values, producing the series $[X(t) - X(t-1)]$ is suggested, with iterative differencing when necessary; subtraction of a smoothed version of the series from itself is another method. Transformation by logarithms has been suggested to eliminate trend in seasonality.

References

[1] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

[2] Brillinger, D.R. (1975). *Time Series: Data Analysis and Theory*. Holt, Reinhart & Winston, New York.

[3] Brockwell, P. & Davis, R. (1987). *Time Series: Theory and Methods*. Springer-Verlag, New York.

[4] Diggle, P.J. (1990). *Time Series Analysis: A Biostatistical Introduction*. Clarendon Press, Oxford.

[5] Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

A.J. LAWRANCE

Statistical Consulting

Statistical consulting, the provision of statistical advice and/or services to those who request it, applies statistical methodology to problems in other disciplines. Consultants assist with design and conduct of the study, including **randomization** of subjects, data collection, and data analysis. They help to report the results of the study and to ensure that conclusions reached are supported by the data. The consultation may range from a five-minute chat in a hallway, involving only advice about some aspect of the study, to a many years' collaboration on a project. Although the terms *consulting* and *collaboration* are often used interchangeably, a collaboration implies more responsibility and involvement, both intellectual and time, by the statistician. In a collaborative relationship, a statistician is a full-fledged member of the team of investigators conducting the study, has more authority, receives credit for contributions made, and coauthors the research paper reporting the project. This is a relationship most conducive to statistical contribution. To connote a broad range of services, some statisticians now refer to the *practice* of statistics, meaning the communication of statistical information across disciplinary boundaries by persons who have training in statistics and related quantitative fields.

Biostatistical consulting is the application of statistical expertise in the biological or health sciences. Within the arenas of medicine, dentistry, and public health, biostatisticians work with physicians, basic scientists, dentists, nurses, pharmacists, epidemiologists and other health professionals. A biostatistician may be a faculty member in a school of public health or a professor in a quantitative sciences department in a medical or dental school or at a medical research center [4]. In this capacity, they teach graduate courses in biostatistics while working collaboratively on research grants, jointly with medical colleagues. In addition, the biostatistician might perform analyses for reports, manuscripts, and presentations for medical clients. In many universities, consulting biostatisticians belong to a statistical consulting unit [14], often within a biostatistics department that offers statistical and computing services. Some universities and schools of public health have statistical or clinical trials centers in which biostatisticians have a primary role [1]. Some biostatisticians work in cancer centers or other disease-specific research centers that may be part of a larger network [25]. Such centers are usually in a university setting; others may be independent entities. State and federal governmental agencies in the US, such as the **Food and Drug Administration** (FDA), the **National Institutes of Health** (NIH), the **National Center for Health Statistics** (NCHS), and the **Centers for Disease Control** (CDC), employ many biostatisticians in a variety of capacities. Pharmaceutical companies usually have a biostatistics unit that works with a team of research investigators in designing and analyzing **clinical trials** [10] (*see* **Statisticians in the Pharmaceutical Industry (PSI)**). Other biostatisticians might be employed in a contract research organization (CRO) (*see* **Proprietary Biostatistical Firms**) that provides statistical and other services to pharmaceutical or biotechnology companies. Finally, some biostatisticians work independently as consultants and contract individually with clients or sponsors.

Role of a Biostatistical Consultant

The role that a biostatistical consultant plays is determined by the type of organization in which they work [9, 20]. The several roles described below demonstrate the variety of circumstances in which biostatisticians are engaged.

Biostatistician in a Medical School

Kerry is the senior of four faculty biostatisticians employed in a clinical trials center at a medical school in a major university in the south east of the US. Having worked closely with physicians for more than 20 years to build and develop the center, he has had years of experience, primarily in cardiovascular clinical trials (*see* **Cardiology and Cardiovascular Disease**). Working with physicians and other medical investigators within a collaborative setting, he offers advice and provides statistical input for the design and analysis of trials of cardiovascular disease. When the trials are being designed, the medical school cardiologists meet with Kerry and investigators from pharmaceutical companies. In designing the trials, Kerry considers both statistical and clinical factors. He seeks to understand the clinical issues in these studies; he provides sample size calculations (*see* **Sample Size Determination for Clinical**

Trials), input on study design, advice on interim analyses, and **randomization** schedules (*see* **Randomized Treatment Assignment**). He usually writes the statistical portions of the study protocol (*see* **Clinical Trials Protocols**). As the trial progresses, he works with junior statisticians to monitor the trial (*see* **Data and Safety Monitoring**) and supervises the analysis. When the trial has been completed, many manuscripts reporting various aspects of the study are produced, for which Kerry specifies the appropriate statistical methods and supervises the analysis by junior biostatisticians. In addition, he is the principal investigator of three clinical trials.

Kerry's greatest strengths as a consultant are his ability to listen intently and to explain statistical design and methodology principles to physicians clearly (*see* **Teaching Statistics to Physicians**). In addition, he often conducts research conferences for a mixed audience of physicians, statisticians, trial coordinators, and other team members in the organization. His ability to speak systematically and to present statistical ideas simply make him sought after as a lecturer. He also teaches a class of physicians and other health professionals who are enrolled in a biostatistics training program to obtain a master's degree. In all aspects of Kerry's varied role, he communicates effectively with physicians, health professionals, and other statisticians, in a group setting or individually.

Biostatistician in a Pharmaceutical Company

Ellen is a doctoral-level biostatistician in the Clinical Biostatistics Department of a major pharmaceutical company on the East Coast of the US. As the lead statistician for all of the clinical trials of ophthalmic preparations that are in development or are currently manufactured by the company, she collaborates with clinical researchers to plan and design clinical trials for protocols to study new chemotherapeutic treatments. She writes the statistical sections, including plans for the designs and the analyses of the studies that will be part of a New Drug Application (NDA) (*see* **Drug Approval and Regulation**). She analyzes clinical trial data; she writes the results sections of clinical study reports and statistical technical sections of regulatory submissions; she coauthors manuscripts submitted to medical journals. Ellen attends and often presents the statistical aspects of the clinical trial at meetings with regulatory agencies and responds to queries, verbal or written, concerning the statistical aspects of the study. She serves as the statistical liaison for studies subcontracted to CROs and she organizes work for a team of masters and doctoral statisticians and statistical programmers. As the lead statistician for ophthalmic preparations, she compiles periodic reports to inform the upper management of the activities of her department concerning the status of projects, resource requirements, and important technical and regulatory issues. As the lead statistician, she is an integral member of the project team for ophthalmic preparations, a cross functional team made up of people in various areas of the company, including Basic Research, Biostatistics, Clinical Pharmacology, Clinical Research, Data Coordination, Drug Metabolism, Epidemiology, Marketing, Manufacturing, Project Planning, Regulatory Affairs, and Safety Assessment. The project team meets regularly to coordinate activities concerning the approval and promotion of ophthalmic products, to communicate new information to its members, and to make recommendations to the senior management of the company about key decisions. On the ophthalmic project team, she represents her department, participates in discussions that require statistical input and thinking, submits a time schedule of work to be accomplished, reports on the progress of projects, and presents clinical trial results to the project team.

Biostatistician in a Dental School

Stuart, a member of the biostatistics faculty of a dental school on the west coast of the United States, acts primarily as a collaborative researcher but also works on relevant methodological problems in biostatistics, teaches research methods courses, and mentors master's degree students. In most situations, he is a consultant, which always has a teaching component. He consults with clinical faculty members (dentists and hygienists), postgraduate residents, basic scientists, including basic laboratory researchers, oral epidemiologists, and public health dentists. He works on many studies including clinical trials of caries risk management, community intervention trials for spit tobacco cessation, **observational studies** of periodontal disease and temporomandibular disorders, and **case–control studies** of early childhood caries. Stuart also helps to write grant proposals, refines study designs, performs **power** analyses and **sample size determinations**, develops analysis plans, coordinates data management activities (*see* **Data**

Management and Coordination), and conducts data analyses.

Dental studies present particular statistical structures, most notably clustered observations (e.g. jaws, teeth, tooth surfaces, and periodontal probing sites) (*see* **Clustering**) within people. In addition, the variability of dental disease, both for sites within and among individuals, and lack of consistency over time (areas can heal and later relapse) create challenges for epidemiologic studies and sample surveys designed to estimate, for example, incidence and **prevalence**. Appropriate modifications are needed in sample size calculations and analyses. Issues of reliability (*see* **Agreement, Measurement of**) and validity often arise since many responses are measured on categorical scales. In all these activities, Stuart consults with colleagues, asks probing questions about the subject matter, refines study research questions, explains possible design and analysis options, and resolves the problems with clients in a partnership to develop statistically appropriate solutions based on goals and available resources.

Biostatistician in a Governmental Agency

Debby leads a biostatistical group in one of the Institutes at the National Institutes of Health in the US. Because her Institute recognizes the value of statistics in the design, analysis, and interpretation of medical studies, the group of biostatisticians finds itself in constant demand. Debby sees her consulting role divided into three very different functions – intramural consulting, participation in discussions on issues related to funding, and collaboration with non-government investigators. For the intramural scientists within the Institute, she serves as an *ad hoc* statistical consultant, with all the excitement and frustration that role entails. The investigators, often postdoctoral fellows, look to statisticians as the purveyors of small ***P*** **values**, those precious tickets to publication. Debby's first contact with an investigator may be a frantic call demanding immediate statistical input for an abstract that is complete "except for a few numbers". She recognizes that these encounters afford an opportunity for teaching young researchers about the value of statistical collaboration. Moreover, a fruitful short-term statistical fix may lead to long-term collaboration with an investigator who has become convinced that statistical thinking enhances scientific research.

In her second consultative role, Debby's activities have broad influence on public health. The NIH, as one of the primary sponsors of clinical studies, establishes priorities for funding of types and fields of studies. In her capacity as an Institute biostatistician, Debby helps to formulate the statistical aspects of research programs. Her medical colleagues at the Institute have sophisticated understanding of statistical ideas, and the biostatisticians understand the medical problems. Thus, the joint medical and statistical collaboration allows rigorous formulation of the scientific basis of scopes of work for contracts or broad areas to fund for grants. This aspect of Debby's consultation, although perhaps her most anonymous activity, is the one she finds most rewarding.

Debby's third role has one foot in government and one in academe. Her Institute funds many long-term clinical trials and epidemiologic studies and she, along with other members of the Institute, serves on committees for these studies. As a committee member, she acts just like everyone else. In some capacities she plays a role distinct from the academic researchers. As a government employee, she sits as an observer on many **data and safety monitoring boards**. Thus she brings to data monitoring a wealth of experience. Investigators, both within NIH and outside, call upon her to consult on forming data monitoring boards, planning interim analyses, and helping coordinating centers prepare statistical reports.

Debby's consulting spans a wide range of activities. Like any effective biostatistical consultant, she must understand the studies with which she is involved; she must be aware of new biostatistical methodology, and she must be poised to develop new techniques. What distinguishes her role from that of other biostatistical consultants is the perception that when she expresses an opinion, she is speaking for the government.

Biostatistician in a Department of Biostatistics and Director of a Consulting Unit

Gary, a professor in a department of biostatistics at a major university in the south east of the US, teaches two **categorical data** courses, mentors numerous doctoral dissertations, and directs a biostatistical consulting laboratory staffed by students in the department (*see* **Teaching Medical Statistics to Statisticians**). The role of the laboratory, which provides statistical consultation primarily

to investigators from the university's Schools of Medicine, Dentistry, and Public Health, is to provide funding and training for students in academic biostatistics programs and to encourage practice-oriented statistical research, particularly for master's degree papers and doctoral dissertations – and, of course, the investigators benefit by having the statistical aspects of their research problems solved. In addition to clients within the university, Gary also arranges cooperative agreements between the laboratory and sponsors in the biopharmaceutic industry.

Gary has created a structure for consultation that fosters learning and teaches team work. The doctoral students have a direct liaison with him, delegate projects to master's and undergraduate students, and mentor them. The doctoral students are responsible for meeting deadlines and writing and reviewing statistical reports. The master's students work with the doctoral students for large projects and manage databases (*see* **Database Systems**), write programs and reports, and learn to handle smaller projects directly with clients. They also mentor the undergraduates. The undergraduates perform support services including data entry, word processing, and basic programming. The benefits for students of this arrangement include mentoring by Gary, reinforcement of course work, practice in communication skills and report writing, publications, and thesis topics.

A wide variety of projects are available to students in the consulting laboratory. One example is a longitudinal periodontal study for the university's Department of Dental Ecology. Patients were examined at baseline and at three other time points. Attachment loss was measured at multiple sites on each tooth and for multiple teeth per person. There were also numerous other site-level, tooth-level, and person-level variables. The primary question was whether attachment loss during one period in time was associated with higher risk for attachment loss at a subsequent period. Analysis of these data involved using special software (*see* **Software, Biostatistical**) to adjust for the clustering of observations within each person. Three journal publications and two master's papers resulted from this investigation.

Another example is a study of Alzheimer's disease for the university's Department of Family Medicine. Data were collected from a sample of specialized dementia care units and traditional care units in five states. A **case–control** matched pair design was used, in which 307 residents from 31 specialized units and 318 patients from 32 traditional units were randomly selected. Data were obtained from questionnaires, medical record review, and direct observation. The key question was whether the use of physical restraints or pharmacologic restraints was different based on whether the individual was in a traditional or specialized care unit. **Logistic regression**, which accounted for clustering of individuals within the same unit, was used to adjust for other factors in modeling use of physical or pharmacologic restraints. A manuscript was published in a prominent medical journal.

Gray has many strengths as a consultant, including his ability to listen discerningly as clients explain their investigations. He has a great ability, based on years of experience, to assess problems and determine the appropriate statistical methodology for the situation. As he involves students in many facets of the work, he allows them to observe his interactions with clients and the development of solutions to research problems both during and after the consultation. He also explains his ideas to the students, and why he is using certain statistical techniques. Contact and coaching with real research problems in this way is excellent training for the students. Gary is a caring mentor and is highly regarded by students, faculty, and clients for both his insightful use of statistics and his ability to teach these skills to others.

Characteristics of a Successful Biostatistical Consultant

A successful biostatistical consultant possesses many skills [13, 16, 18]. The consultant must have a solid technical background in biostatistics, good knowledge of the subject matter area, excellent computing skills, and the ability to apply – or develop if necessary – statistical methods innovatively in a variety of settings. A successful biostatistical consultant should be aware of his or her limitations and know when to ask for assistance from a statistical or subject matter colleague or to learn about the appropriate statistical methodology in the literature. Other important characteristics are: enthusiasm in participating as a member of a team of research investigators; the capability of formulating problems in statistical terms; a

proclivity for problem-solving; and excellent communication skills, both oral and written.

Being a good consultant requires skills beyond ability in mathematics and knowledge of the theory and methods of statistics. Any investigation requires high-quality data, so the successful consulting biostatistician should become familiar with data collection procedures and methods for assuring the quality of the data. He or she should be genuinely interested in the subject matter area in order to become aware of issues that may have important statistical implications. Many clients fear mathematical and statistical ideas, but it is crucial that the client understand the statistical aspects that the consultant is discussing. Thus, a successful biostatistical consultant should be a good teacher who is willing to explain the statistical aspects of design and analysis in terms that a client can understand.

Working effectively on several research projects simultaneously means that the consultant should manage time efficiently and meet appropriate deadlines. Sometimes, the consultant should aim to achieve an acceptable solution to a problem when an ideal solution would require too much time, effort, and expense. Having good interpersonal skills is essential, since many personalities are encountered; sometimes the biostatistical consultant functions as an expert, while at other times he or she is a strategist or confidant. A willingness to admit mistakes and learn from failures helps a consultant in working on subsequent projects with clients.

Challenges of Consulting

Consultations are complex interactions in which the biostatistician may use many skills simultaneously. The successful biostatistical consultant often listens to the client's problem and constructs a statistical formulation. If the client brings data to be analyzed, the consultant should assess the statistical design of the project, explore the assumptions of statistical tests, discuss limitations of the analyses, and suggest appropriate solutions. After arriving at a statistical solution, the consultant should write a coherent report that is understandable to the client. Delivery of raw computer output or an uninterpreted set of tables does not serve either medicine or statistics. The consultant biostatistician should be aware that the nontechnical aspects of consulting may be at least as important as the technical aspects [2, 27]. Projects often have deadlines and budgetary constraints. Some clients apply pressures to achieve a favorable outcome, but the biostatistician must deal ethically and provide an honest report of the study. The best clients are those who understand that an honest report of the project is in the highest interest of science.

Excellence in communication is another essential skill for the consulting biostatistician. Adequate technical knowledge is simply not sufficient if the biostatistician is poor at communicating ideas to clients. Often, biostatisticians have to learn to listen and to ask questions that gradually lead to a revelation of the client's problem. This is especially important when the client is unfamiliar with statistics. Being able to discuss problems in the language of the client is a definite aid to communication. Being knowledgeable in the client's discipline, and being supportive and willing to educate the client, are attitudes that are most helpful in establishing good client–biostatistician relationships. Therefore, while a basic technical knowledge is necessary and presumed for consulting, the nontechnical aspects of consulting are often the most challenging.

Consultation may have a negative side, for the biostatistician may not receive sufficient recognition by the client and by the institution or organization in which the consultant is employed. In universities, promotions and tenure are based primarily on research, followed by teaching and service or practice-oriented research. Consulting, which usually falls into the category of supportive practice-oriented research, often is not valued greatly. Extremely active consultants who are consulting with a variety of disciplines simultaneously may not publish enough methodological research in peer-reviewed statistical journals [26]. The biostatistician spends a lot of time learning the discipline of the investigators and communicating with them regularly, particularly in large, long-term projects. The lack of recognition of consulting activities by academic departments when discussions are held about promotions and tenure, coupled with the lack of time for methodological research, compounds the frustration of an academic consultant. In addition, some clients believe that payment for services is sufficient recognition for a consultant's efforts. When the biostatistician has devoted substantial time and energy to a project, the client's failure to acknowledge the intellectual contribution,

in the form of authorship, also frustrates the consultant. Even when consultants are authors, rarely are they first authors. Since first authorship is given greater emphasis during reviews for promotion, consulting contributions tend to be undervalued. Some frustrations may be eliminated when there is a discussion with the client prior to the initiation of the project concerning authorship of papers and charges for statistical services. Some academic institutions realize that consulting biostatisticians contribute substantively to publications and subject matter areas, and consider this during reviews for promotions. Many contributions to statistical methodology have arisen through consulting projects. Also, as more attention is given to utilization of appropriate statistical methodology in research papers (e.g. the *Journal of the National Cancer Institute* has 20 consulting statistical editors), perhaps academic institutions will award more credit when a biostatistician is the coauthor of a medical paper.

Challenges may be different in other settings. In an organization in which consultation constitutes the primary role of the biostatistician, it may be difficult to satisfy the requests of several clients simultaneously in a manner consistent with the priorities of the organization. Often, the ideal analysis of a set of data may require more time and money than has been allocated to the project. The challenge to the biostatistician is to set priorities firmly and complete projects in a timely manner, subject to budgetary constraints.

Whatever the setting, biostatistical consultants must stay current with the statistical literature while maintaining their other responsibilities. In addition, they must maintain their computing skills in the face of evolving technology and a plethora of software packages. It is challenging to stay abreast of these developments in the field while simultaneously meeting work demands.

Special Challenges of Consulting with Physicians

A biostatistician working with physicians faces many challenges. Some problems are universal issues that occur with clients from all fields, and others seem to derive from differences in mind-set, approach, or inherent differences in the nature of physicians and statisticians. Physicians are trained to produce quick responses when presented with a set of patient characteristics; indeed, the patient usually expects an immediate assessment. In some clinical specialties, these decisions are made in a life-threatening situation. Physicians may weigh the results of a number of tests or factors as they make the decision. If most of the evidence leads to a certain conclusion, they often do not quibble over slight vagaries. In the clinical setting, physicians need and want the best answer in the shortest length of time. On the other hand, biostatisticians tend to be meticulous. They examine the data, check for discrepancies and **outliers**, test assumptions, and often approach questions from several angles. They like to do a thorough job. When physicians bring their clinical mind-set to the statistical consultation, they sometimes complain that biostatisticians are overly conscientious. And biostatisticians frequently complain that physicians want conclusions before the relevant data have been fully analyzed and interpreted! Each party should recognize the inherent differences in style and modify their approaches somewhat, moving toward a more "central" position. "Differences" seen in this light allow better understanding between physicians and biostatisticians.

Another complication that often surfaces when biostatisticians work with physicians is the psyche of the physician. Physicians are used to being "in charge". When coming to a biostatistician for assistance, they can feel either a lack of control because of a lack of familiarity with biostatistics or a need to dominate the biostatistician about what "statistics" should be done. Sometimes, physicians who do not understand the statistical arguments adopt a passive–aggressive stance, accepting with question the results presented by the biostatistician. This situation is usually stressful for the biostatistician and physician alike. Often this can be alleviated by a frank characterization of the situation and a statement by the biostatistician of their commitment to work with the physician, perhaps suggesting an ongoing collaboration. This, combined with some respectful coaching of the physician on statistical terms and practice by the consultant, can mollify the situation.

Many physicians and biostatisticians develop extremely productive collaborations that continue for years. This occurs most often when the physician has come to value the impact that the biostatistician has had on their work in terms of efficiency and accuracy of methodology, leading to an increase in the number of grants or papers accepted for publication. And

that usually occurs when the biostatistician has taken the time to become educated in the discipline of the physician and has fostered the relationship. In these long-term relationships, biostatisticians often have concentrated their consulting efforts in a particular area of medicine and become "experts" in the analyses that are used most often in that area.

Biostatistical Consultants as Ambassadors

The relationships that a biostatistician establishes with people in other fields can have an important impact on how those collaborators view statistics as a discipline. Often, the only contact that investigators have with statistics is through their consultant. By seeking out opportunities to demonstrate the usefulness and the power of statistics, the consultant can enhance the image of statisticians both locally and on a broader scale [3]. Locally, they can give a talk, a short course, or a workshop to a client's department or unit. Making use of actual data from designed studies in the client's field is an effective method to illustrate statistical principles. In a consulting practice, the biostatistician can also reflect interest in the client's projects by visiting the client's laboratory or office to learn about the conduct of the research project. Suggestions for improving the consulting process can come by soliciting feedback from the client directly. Feedback questionnaires administered anonymously can also elicit useful suggestions.

A biostatistician can enhance the image of statistics both nationally and internationally. For example, many biostatisticians participate in the review process for scientific manuscripts and research proposals [6]. They can serve on a journal's editorial board and develop standards for statistical review [12, 17] (*see* **Statistical Review for Medical Journals**). Numerous research journals have supported commentaries or tutorial articles on statistical issues relevant to the journal's readers [7]. All of these activities have a far-reaching impact on how clients view statisticians.

Another emissary role of a biostatistician is service on advisory committees for state or federal agencies. In particular, biostatisticians serving the Food and Drug Administration (FDA) or the National Institutes of Health (NIH) and other governmental agencies not only represent the profession, but also have an opportunity to shape decisions and make policy that may have a major impact in the health and medical sciences. Besides the statistical input, they can provide for the process of decision-making, and they can also raise awareness of the need for statistical thinking about research projects at the highest levels of these organizations.

Training of Biostatistical Consultants

Most practicing biostatisticians have graduate degrees, often from a school of public health. Undergraduate degrees are often in mathematics, statistics, or biostatistics. However, many biostatisticians have their first degree in a field such as biology, psychology, or pharmacy. Graduate training includes the usual theoretical and applied courses for a statistician as well as courses in epidemiology and other public health fields. In the past, most of the students' practical consulting experience obtained in graduate programs came from apprenticeship participation on projects with faculty advisors or from working in consulting centers, usually under faculty direction. However, the past 10–15 years has seen a growing interest in the design and development of courses in consulting to teach the specific skills needed; some of these courses have been implemented in departments of biostatistics. Although most people agree with the list of skills needed by statistical and biostatistical consultants, and many articles in the statistical literature attest to them [5, 11, 13, 15, 21], only a few people have actually designed courses to model these skills and to cultivate them in students. Most biostatisticians have not been simultaneously trained in psychological or communication skills. Even if a biostatistician does naturally possess the traits necessary for consulting, he or she may not know how to transmit these characteristics to others systematically. In addition, a university may not reward spending time in such an endeavor.

A variety of approaches have been taken to educate students about consulting. They usually include some analysis of "real" data, which most consider the main skill needed for consultation, and report writing. Other courses focus mainly on the psychological and communication aspects of consulting [19, 24]. More and more, though, the courses take a broad approach and incorporate the skills needed to carry out an entire project from start to finish. These courses offer exercises and discussions to develop and improve skills in communication, organization, analysis, and presentation of oral and written results [8,

22, 23]. Often using videotaped consultations, students are shown how to elicit information from and deliver technical information to clients as well as how to manage a consulting session. Students are often videotaped in mock or actual consultations. Carrying out a full-fledged project during the class provides experience with time management, organization and documentation of project materials, discussion of appropriate analysis approaches, writing of statistical reports, and opportunities to discuss budgets and the billing of clients. Courses such as this are usually prerequisites or adjuncts to the actual consulting experience.

Many students find additional ways to gain practical experience during their graduate programs. They either work part-time or participate in internship programs. This experience gives both parties a chance to assess potential future relationships. However, on-the-job training is perhaps the primary method for learning consulting. Being aware of what works and what does not work, and changing one's procedures accordingly, is the ultimate strategy for success in consulting. Add years of experience and you will have a seasoned consultant.

Incentives for Biostatistical Consulting

Consultation has many rewards for a biostatistician. The skills required may match those of the individual, whereas a career spent entirely in teaching and research may be less rewarding. A biostatistical consultant will almost surely design an experiment and analyze data, whereas many academic biostatisticians may have no real experience in these areas. In addition, many problems encountered in consulting relationships are challenging. Since most new statistical methodology arises from realistic problems, being a consulting biostatistician is a way to learn about new projects requiring advances in statistical methodology. Also, there is the excitement of participating as a collaborating member of a team on a research project that is addressing an important scientific or health related question. Another incentive is that financial benefits accruing to the biostatistical consultant are often greater than for those participating only in academic work. Additional benefits derive from being coauthors on medical publications, making presentations at scientific meetings, and contributing generally to the betterment of society.

Acknowledgment

We acknowledge and appreciate the contributions, critical review, and helpful suggestions for this article made by Janet Wittes, Stuart Gansky, Gary Koch, Kerry Lee, Ellen Snyder, and Gail Tudor.

References

[1] Arndt, A. & Woolson, R.F. (1991). Establishing a biostatistical core unit in a clinical research center, *American Statistician* **45**, 22–27.

[2] Boen, J.R. & Zahn, D.A. (1982). *The Human Side of Consulting*. Lifetime Learning, Belmont.

[3] Boroto, D.R. & Zahn, D.A. (1989). Promoting statistics: on becoming valued and utilized, *American Statistician* **43**, 71–72.

[4] Carter, R.L., Scheaffer, R.L. & Marks, R.G. (1987). The role of consulting units in statistics departments, *American Statistician* **40**, 260–264.

[5] DeMets, D.L., Anbar, D., Fairweather, W., Louis, T.A. & O'Neill, R.G. (1994). Training the next generation of biostatisticians, *American Statistician* **48**, 280–284.

[6] Derr, J.A. (1993). Biostatistics cores: improving the chances for funding, *American Statistician* **47**, 99–102.

[7] Derr, J.A. (1995). Statistics in nutrition, part 8: a review of good statistical practices, *Journal of Renal Nutrition* **5**, 208–209.

[8] Derr, J.A. & Rosenberger, J.L. (1992). A multi-objective course in statistical consulting, in *American Statistical Association 1992 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 269–272.

[9] Derr, J.A. & Stinnett, S.S. (1994). The interesting life of a practicing statistician, *Stats. The Magazine for Students of Statistics* **11**, 7–11.

[10] Ederer, F. (1979). The statistician's role in developing a protocol for a clinical trial, *American Statistician* **33**, 116–119.

[11] Feigl, P. (1980). The training of statisticians for clinical trials, *Biometrics* **36**, 677–678.

[12] Gardner, M.G. & Bond, J. (1993). An exploratory study of statistical assessment of papers published in the *British Medical Journal*, *Journal of the American Medical Association* **263**, 1355–1357.

[13] Gehan, E.A. (1980). The training of statisticians for cooperative clinical trials: a working statistician's viewpoint, *Biometrics* **36**, 699–706.

[14] Gibbons, J.D. & Freund, R.J. (1980). Organizations for statistical consulting at colleges and universities, *American Statistician* **34**, 140–145.

[15] Hammond, D. (1980). The training of clinical trials statisticians: a clinician's view, *Biometrics* **36**, 679–685.

[16] Hunter, W.G. (1981). The practice of statistics: the real world is an idea whose time has come, *American Statistician* **35**, 72–76.

[17] International Committee of Medical Journal Editors (1993). Uniform requirements for manuscripts submitted to biomedical journals, *Journal of the American Medical Association* **269**, 2282–2286.

[18] Kirk, R.E. (1991). Statistical consulting in a university: dealing with people and other challenges, *American Statistician* **45**, 28–34.

[19] McCulloch, C.E., Boroto, D.R., Meeter, D., Polland, R. & Zahn, D.A. (1985). An expanded approach to educating statistical consultants, *American Statistician* **39**, 159–167.

[20] Niland, J.C., Odom-Maryon, T.L., Lee, J. & Tilley, B.C. (1995). A survey of biostatistical consulting units throughout North America, *American Statistician* **49**, 183–189.

[21] Peterson, A.V. & Fisher, L.D. (1980). Teaching the principles of clinical trials design and management, *Biometrics* **36**, 687–697.

[22] Stinnett, S.S. (1990). Training statistical consultants using modules and mirrors, in *American Statistical Association 1990 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 194–199.

[23] Stinnett, S.S. (1991). Quality improvement procedures in statistical consulting education, in *American Statistical Association 1991 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 147–152.

[24] Stinnett, S.S. (1993). Are videotaping and psychology worth the effort for statistical consultants? in *American Statistical Association 1993 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 148–151.

[25] Williford, W.O., Kroll, W.F., Bingham, S.F., Collins, J.F. & Weiss, D.G. (1995). The multicenter clinical trials coordinating center statistician: "More than a consultant", *American Statistician* **49**, 221–225.

[26] Wilson, W.J. (1992). Statistical consulting is scholarship, *American Statistician* **46**, 295–298.

[27] Zahn, D.A. & Isenberg, D.J. (1983). Nonstatistical aspects of statistical consulting, *American Statistician* **37**, 297–302.

SANDRA S. STINNETT, JANICE A. DERR & EDMUND A. GEHAN

Statistical Dependence and Independence

Statistical dependence is a type of relation between any two features of units under study. These units may, for instance, be individuals, objects, or various aspects of the environment. Deterministic dependence and statistical independence can be regarded as the two opposite extreme types of relation, but also as being qualitatively distinct from the possible other forms of relation. If deterministic dependence and independence are excluded, then the remaining intermediate types of statistical dependence involve both features as proper variables such that there are differences in the distributions of one variable for at least some of the levels of the other.

If proper variables are statistically independent, then the distribution of one of them is the same no matter at which fixed levels the other variable is considered and observations for such variables will lead correspondingly to nearly equal frequency distributions. If there is deterministic dependence, then the levels of one of the variables vary in an exactly determined way with changing levels of the other. In other words, under independence, knowledge about one feature remains unaffected by information provided about the other, while under deterministic dependence it follows with certainty which level of one variable occurs as soon as the level of the other variable is known.

The definition of these opposite extreme types of relation is symmetrical between the two features involved, but in its intermediate forms, statistical dependence may or may not be considered in a symmetric way, depending on the substance matter context. A symmetrical type of dependence will be appropriate if the variables involved are considered to be on an equal footing, such as symptoms of a disease, or as length, height and depth of produced objects, or as personality characteristics of individuals. By contrast, an asymmetrical form of dependence is of main interest if, instead, one of the variables is considered as a possible **response** to the other, such as weight to caloric intake, or as depression to anxiety. The terms symmetric **association** and directed association are often used to capture this distinction.

Given observations on independent units, statistical dependence shows in a number of different ways depending on several aspects. Important are, in particular, the types of variable involved, the conditions under which the relation is recorded, and the type of association measures used to summarize the data. These issues are addressed next, in turn.

Relations Depending on Types of Variable

One important distinction for variables is whether they are qualitative or quantitative. Quantitative variables have levels that are numerical values with a substantive meaning, such as kilograms, as ranks, or as sumscores of questionnaires. Qualitative variables have, instead, categories as possible levels. With a nominal scale the categories are just of a qualitatively similar kind such as **blood groups**; numbers possibly assigned to them play the role of codes; that is, of mere labels. In the case in which levels of a qualitative variable can be ranked, the scale becomes ordinal. This information may sometimes be exploited to improve formal analysis (*see* **Measurement Scale**).

First, data summaries appropriate to detect the form of **pairwise dependence** change with the types of variable involved. They are, typically, **contingency tables** for qualitative or discretized quantitative variables, scatter plots for quantitative variables and frequency distributions (or at least selected characteristics of the distributions) of the quantitative variable displayed within each category of the qualitative variable (*see* **Graphical Displays**).

Accordingly, a great variety of more formal techniques is available. In the case of symmetric associations examples are **loglinear models** for qualitative variables, covariance selection for quantitative variables (*see* **Variable Selection**), and mixed **interaction** models for both qualitative and quantitative variables. In the case of directed associations examples are logistic [2] and probit [6] regression for discrete responses (*see* **Quantal Response Models**), **linear regression** for quantitative responses, and combinations of these for mixed joint responses. In any case it is essential to check systematically [4] for more complex dependencies involving several variables or, possibly, nonlinear relations among quantitative variables.

Relations Depending on the Conditioning Set

Every statistical dependence among observed variables is a conditional relation, since there is always some conditioning, at least implicitly on time and location of the study. A more explicit form of conditioning may result by design or by statistical analysis involving several recorded variables. In that case the distinction between conditional and marginal dependence and conditional and marginal independence becomes relevant. Both may convey different information. A marginal dependence of a response on a potential explanatory variable may, for instance, be completely explainable in terms of a corresponding conditional independence statement given an intermediate variable, which itself is strongly related to both.

One example from the German labor market in 1986 is shown here with the following 2^3 contingency table, adapted from job placement statistics [1]. The response is successful job placement, A, the intermediate variable is field of study, B, and the potential explanatory variable is gender of the applicant, C. If the marginal dependence of job placement on gender is considered, i.e. the overall association of pair (A, C), shown on the right-hand side of Table 1, it appears as if there were discrimination against women, since females have a much lower chance than men of obtaining a job.

This dependence can, however, be explained in the following way: home economics was a preferred field of qualification for women, while mechanical engineering was strongly preferred by men. At the same time there were many more successful job placements for mechanical engineers than for home economists, simply because many more job openings were available for the former. Within each of the two fields of qualification there was the same percentage of successful job placements for both, women and men. In other words, A is conditionally independent of C given B (*see* **Simpson's Paradox**).

This conditional independence, together with the strong marginal associations for pairs (A, B) and (B, C) both having variable B in common, imply the observed dependence for (A, C); that is, this dependence is generated by the intermediate variable B. The data are also an example of a simple **Markov chain** [8] and, more generally, of a graphical Markov model, a general framework (see [4, 5], and [7]) within which sequences of response, intermediate and **explanatory** variables, both types of variables, qualitative and quantitative, distinct levels of conditioning and interactive as well as nonlinear relations, may be modeled explicitly.

Judgment of Relations as Dependent on Measures of Association

In many contexts it is possible to summarize dependencies concisely with a few carefully chosen measures of association (*see* **Association, Measures of**). One example for a quantitative response and equally spaced levels of a quantitative explanatory variable is the set of coefficients of a **polynomial regression**. If, for instance, the dependence can be well captured by an orthogonal polynomial in three coefficients, then the dependence is additively decomposed into an overall mean, a linear, and a quadratic effect. A direct extension is, conceptually though not technically, the decomposition of a time dependence into a general level, a linear trend, and seasonal effects.

Some measures of association arise as parameters in multivariate distributions. In such distributions, it is typical that discrete random variables model

Table 1 Overall dependence in spite of conditional independence

	B, field of qualification				Overall; that is, summed over B	
	Home economics		Mechanical engineering			
	C, gender		C, gender		C, gender	
A, successful job placement	Female	Male	Female	Male	Female	Male
Yes	15 (3.61%)	2 (3.64%)	4 (20.0%)	95 (21.1%)	19 (4.4%)	97 (19.2%)
No	400	53	16	355	416	408
Sum	415	55	20	450	435	505

qualitative features and continuous random variables model quantitative features. For symmetric associations one prominent example is the **exponential family** called the conditional Gaussian (CG) distribution, in which the continuous variables have a joint Gaussian distribution for each level combination of the discrete variables.

In the bivariate versions of the CG distribution, the canonical association parameters are log **odds ratios** for two discrete variables, multiples of the simple **correlation** coefficient for two continuous variables, and a weighted difference in means for the mixed case. In higher dimensions these association parameters are generalized in such a way that null values of all terms involving a particular pair of variables imply conditional independence of the pair given all remaining variables: the measures of association are then conditional log odds ratios, multiples of partial correlation coefficients, and weighted differences of means, corrected for effects of the remaining variables.

The obvious danger in using measures of association which are part of a well studied joint distribution is that the true distribution of the features under study may be quite different. For instance, if the judgment of dependencies among quantitative variables were based only on simple and partial correlation coefficients, then substantial misjudgments of the actual relations might result. If the simple correlation is zero, then strong nonlinear relations of a particular type may still be present, but at least, if the simple correlation is nonzero, the variable pair will always be marginally dependent. The situation is much worse with partial correlations.

Every partial correlation coefficient is a simple correlation coefficient for **residuals** obtained after linear regression on some common set of further variables. As for the simple correlation, there may be strong nonlinear conditional associations even if a partial correlation coefficient is zero. However, the reverse may happen as well; that is, the partial correlation coefficient may be high in spite of conditional independence. This is best illustrated with an example.

Let Z, U, and V be mutually independent variables, each having a standardized Gaussian distribution; that is, in particular, each having mean zero and variance one. Define Y and X as follows:

$$Y = (Z^2 - 1) + U, \qquad X = (Z^2 - 1) + Z + V.$$

Then Y is conditionally independent of X given Z, written as $Y \perp\!\!\!\perp X|Z$, because given Z only U and V are variable, and they are independent by assumption. But the simple correlation between the residuals from linear regression is 2/3; that is, the partial correlation coefficient $\rho_{xy.z}$ is sizeable.

To see this, note that linear – instead of the appropriate nonlinear – regression of Y on Z and of X on Z would give as conditional means

$$\mathrm{E}_{\mathrm{linear}}(Y|Z) = 0, \qquad \mathrm{E}_{\mathrm{linear}}(X|Z) = Z,$$

and hence as residuals from these linear regressions

$$R_{Y.Z} = (Z^2 - 1) + U, \qquad R_{X.Z} = (Z^2 - 1) + V.$$

Since the square of a standardized Gaussian variable has a **chi-square distribution** on one **degree of freedom**, the variable Z^2 has mean 1 and variance 2 and the residuals both have zero means. Furthermore, both residuals have variance 3 and their covariance is

$$\mathrm{cov}(R_{X.Z}, R_{Y.Z}) = \mathrm{var}(Z^2 - 1) = 2,$$

so that $\rho_{xy.z} = \mathrm{cov}(R_{X.Z}, R_{Y.Z})\{\mathrm{var}(R_{X.Z})\mathrm{var}(R_{Y.Z})\}^{-1/2} = 2/3$ even though the corresponding conditional independence statement $Y \perp\!\!\!\perp X|Z$ holds. Of course, if for corresponding observations systematic checks for nonlinearities and interactions were used [3], then it would certainly be detected that nonlinear associations are present and hence it would be noticed that correcting for only linear relations of Y on Z and of X on Z is inadequate.

An alternative to assuming that a set of variables has a particular distribution is to define the joint distribution only implicitly via a sequence of recursive conditional distributions. This is typical for graphical Markov models corresponding to so-called chain graphs. In that case, conditional dependencies of potential explanatory variables are modeled separately for each response in accordance with available substance matter knowledge [4, 9]; nonlinear relations and interactions among continuous variables may be part of the model. In addition, for a given model it may often be deduced which independencies and associations are implied under other conditioning sets than those specified with the given model [10].

Another important additional advantage of such conditional modeling is that issues such as **censoring**, measurement error (*see* **Errors in Variables**), missing values, time dependencies, and effects of hidden random variables may in principle be directly

integrated into the modeling process. To date, however, the actual implementation might for some combinations still require substantial further theoretic and technical developments.

References

[1] Bundesanstalt für Arbeit (1986). *Amtliche Nachrichten* **5**, 846–847.

[2] Cox, D.R. (1958). The regression analysis of binary sequences (with discussion), *Journal of the Royal Statistical Society, Series B* **20**, 215–242.

[3] Cox, D.R. & Wermuth, N. (1994). Tests of linearity, multivariate normality and the adequacy of linear scores, *Applied Statistics* **43**, 347–355.

[4] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies – Models, Analysis and Interpretation*. Chapman & Hall, London.

[5] Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer-Verlag, New York.

[6] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, Cambridge.

[7] Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press, Oxford.

[8] Markov, A.A. (1912). *Wahrscheinlichkeitsrechnung* (German translation of 2nd Russian Ed. 1908). Teubner, Leipzig.

[9] Wermuth, N. (1997). Graphical Markov models, in *Encyclopedia of Statistical Sciences*, S. Kotz, C. Read & D. Banks, eds. Wiley, New York, to appear.

[10] Wermuth, N. & Cox, D.R. (1998). On association models defined over independence graphs, *Bernoulli*, to appear.

(*See also* **Pairwise Independence**)

NANNY WERMUTH & D.R. COX

Statistical Forensics

When genetic evidence is used for individual identification, there are generally competing explanations for the observations. A typical forensic situation arises when biological material at the scene of a crime is typed, found to have some profile A, and the circumstances of the crime suggest that the material was left by the perpetrator P. A person S suspected of having committed the crime is also typed, and is found to have the same profile. The evidence E is that the two profiles are of type A.

The competing explanations are:

H_p: the crime sample is from S
H_d: the crime sample is not from S

and the relative merits of these two explanations are compared by means of a **likelihood ratio**. This compares the probability of the evidence under the two explanations:

$$L = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}. \tag{1}$$

Values of L greater than 1 favor the explanation H_p over H_d. If there are prior odds $\Pr(H_p)/\Pr(H_d)$ on S being the contributor, then the posterior odds $\Pr(H_p|E)/\Pr(H_d|E)$ follow from **Bayes' Theorem** as

$$\text{posterior odds} = L \times \text{prior odds}.$$

One of the most common errors in interpreting genetic evidence is to confuse the posterior odds with the likelihood ratio. This transposition of the conditional is more commonly made by prosecutors, giving rise to the term "prosecutor's fallacy". It is generally the case that $\Pr(E|H_p) = 1$, and the value of $\Pr(E|H_d)$ might be 10^{-6}. The likelihood ratio is then one million, but the posterior odds depend on the prior odds. They are not a million to one on guilt. Although odds on guilt is very much the kind of information desired by courts, it cannot be found from genetic evidence alone.

Conditional Probabilities

Eq. (1) can be modified by the rules of **conditional probability**. If S_A and P_A mean that S and P, respectively, have genetic profile A, then

$$\begin{aligned} L &= \frac{\Pr(S_A, P_A|H_p)}{\Pr(S_A, P_A|H_d)} \\ &= \frac{\Pr(P_A|S_A, H_p)\ \Pr(S_A|H_p)}{\Pr(P_A|S_A, H_d)\ \Pr(S_A|H_d)}. \end{aligned}$$

It may generally be assumed that the profile type of S does not depend on either explanation of the matching profiles, so $\Pr(S_A|H_p) = \Pr(S_A|H_d)$, and that a match is certain under H_p, so $\Pr(P_A|S_A, H_p) = 1$, and then

$$L = \frac{1}{\Pr(P_A|S_A, H_d)}.$$

The focus on conditional probabilities greatly simplifies the interpretation of matching profiles. The question is clearly seen to be "What is the probability that the perpetrator of the crime is of type A given that S is of type A, when these two people are not the same?" The smaller this probability, the stronger the evidence against S. By emphasizing that L depends on the probability of an event, comparisons between L and the size of the population are avoided. There is no inconsistency between an L of one million and a population size of one thousand. One has nothing to do with the other.

In the special case that profile probabilities of different people S and P are independent, the likelihood ratio reduces to the reciprocal of the profile probability ("profile frequency")

$$L = \frac{1}{\Pr(P_A)}. \tag{2}$$

This equation will not hold if S and P are related, or if they both belong to the same subpopulation. In one case the two people are related by virtue of being in the same family, and in the second they are related in an evolutionary sense. Although the second dependence cannot be zero for two humans (see **Inbreeding**), it is usually negligible.

The Product Rule

If dependencies between profile probabilities can be ignored, (2) shows that what is needed is the probability with which an unknown (or untyped) person has a specific profile. These profiles are the joint genotypes at several loci. If a_{li} is allele i for locus l,

and has frequency p_{li}, the probability P_A of m-locus **genotype** $A = \prod_{l=1}^{m} a_{li}a_{li'}$ is

$$P_A = \prod_l 2^{h_l} p_{li} p_{li'},$$

where $h_l = 0$ if the profile is homozygous at the lth locus, $a_{li} = a_{li'}$, and $h_l = 1$ if it is **heterozygous**, $a_{li} \neq a_{li'}$. This result holds only if independence of all alleles can be assumed (*see* **Hardy–Weinberg Equilibrium**; **Linkage Disequilibrium**).

Independence of alleles among loci is usually a reasonable assumption for unlinked loci, but dependence within loci may exist. For a population with departures from Hardy–Weinberg equilibrium characterized by inbreeding F_l at locus l, the profile probability would need to be modified to

$$P_A = \prod_l \left\{(1 - h_l)[p_{li}^2 + F_l p_{li}(1 - p_{li})] + h_l 2 p_{li} p_{li'}(1 - F_l)\right\}.$$

This increase for homozygotes and decrease for heterozygotes is appropriate for departures from Hardy–Weinberg equilibrium due to inbreeding. A more likely cause of Hardy–Weinberg departures for human populations, however, is admixture (*see* **Admixture in Human Populations**). For genes with multiple alleles, admixture increases the frequency of homozygotes in the total population over the squared total allele frequency, but the frequency of any particular heterozygote may be increased or decreased over twice the product of total allele frequencies. Such doubts as to the proper adjustment for one-locus frequencies could be avoided by using observed genotype frequencies at each locus, instead of the product rule, and then multiplying genotype frequencies over loci. This procedure has difficulties for highly variable loci, when specific genotypes may not be present in population samples. A better way of avoiding doubts is to return to the conditional probabilities instead of approximating them by unconditional profile probabilities.

Relatives

A plausible defense for a suspect whose genetic profile matches that found in a crime scene sample is that he may be related to the true perpetrator, so that the match has little probative value. Unless the suspect's relatives are typed, this claim must be met by calculating the conditional probability of two relatives having the same profile. A full treatment of this issue, allowing either relative to be inbred, requires the full set of four-allele **identity coefficients**. For noninbred relatives, however, calculations are fairly simple, and involve considering whether or not the relatives share alleles. Results for some common cases are given in Table 1 [2].

Table 1

Genotype A	Relationship	$\Pr(P_A \mid S_A)$
$a_i a_j$ $(i \neq j)$	Full brothers	$(1 + p_i + p_j + 2p_i p_j)/4$
	Father and son	$(p_i + p_j)/2$
	Half brothers	$(p_i + p_j + 4p_i p_j)/4$
	Uncle and nephew	$(p_i + p_j + 4p_i p_j)/4$
	First cousins	$(p_i + p_j + 12p_i p_j)/8$
	Unrelated	$2p_i p_j$
$a_i a_i$	Full brothers	$(1 + p_i)^2/4$
	Father and son	p_i
	Half brothers	$p_i(1 + p_i)/2$
	Uncle and nephew	$p_i(1 + p_i)/2$
	First cousins	$p_i(1 + 3p_i)/4$
	Unrelated	p_i^2

Population Structure

Another plausible defense for a suspect found to have a profile matching that in a crime scene sample is that he and the true perpetrator both belong to a particular subpopulation in which the profile is more common than in the population at large, but that probability calculations have been performed with data taken from the whole population.

To address this question completely the conditional probability would need to be estimated for that subpopulation. This will not be feasible in general, and not even possible if the subpopulation is not well-defined. However, there is a framework for addressing the issue. The article on **Inbreeding** discusses conditional allelic probabilities, and gives the result for allele a_i,

$$\Pr(a_i \mid a_i) = p_i + \theta(1 - p_i). \tag{3}$$

This is the probability that two people in the same subpopulation share allele a_i, with the answer being

given as an average over all subpopulations. The allele frequency p_i applies to the total population, and may be estimated from a population-wide sample. The coancestry (or kinship) coefficient θ cannot be estimated directly without data from these subpopulations, but may be given an appropriate value from other studies or from some value considered to be a plausible upper bound.

For matching profiles, an expression equivalent to (3), but for genotypes, is needed. Such expressions require relationships among four alleles, two per person, but adequate approximations derived from evolutionary-equilibrium theory have been proposed [1]. They are:

$$\Pr(a_i a_i | a_i a_i) = \frac{[p_i + \theta(2 - p_i)][p_i + \theta(3 - p_i)]}{(1 + \theta)(1 + 2\theta)} \geq p_i^2$$

and

$$\Pr(a_i a_j | a_i a_j) = \frac{2[p_i + \theta(1 - p_i)][p_j + \theta(1 - p_j)]}{(1 + \theta)(1 + 2\theta)}.$$

Provided $\theta \geq 0$, the conditional probabilities always exceed the total probabilities for homozygotes and usually for heterozygotes. The same approach is therefore valid for all genotypes, and it is necessary only to select an appropriate value of θ. It is unlikely that θ would be as great as 0.05 for human populations (it is 0.0625 for first cousins), and so these conditional probabilities are very close to the Hardy–Weinberg probabilities when allele frequencies are 0.1 or higher.

Mixtures

Some crime-scene samples contain genetic material from more than one person. This is the case for vaginal swabs from a rape victim, for example. Unless the typing procedure indicates that some elements of the mixture necessarily come from the same person, interpretation of mixed samples proceeds very much as for single stains. Suppose the crime sample has a set $\{e\}$ of alleles for a gene. Explanation H_p is that some specified people contributed to the sample, and that p unknown people must have contributed alleles $\{u\}$ of that set, but did not have any alleles not in $\{e\}$ between them. The probability of this event is written as $P_p(u|e)$. Similarly, $P_d(v|e)$ is the probability of the event specified by explanation H_d: a number d of unknown people must have contributed alleles $\{v\}$ but did not carry any alleles not in $\{e\}$ between them. The likelihood ratio is

$$L = \frac{P_p(u|e)}{P_d(v|e)}.$$

As an example, consider the case where a woman has been raped by two men. A vaginal swab reveals alleles *abcde* for a gene. The victim is of type *ab*, and the type of a single suspect is *cd*. The prosecution explanation is that the victim, the suspect, and one unknown man were the contributors to the sample. The defense explanation may be that the victim and two unknown men were the contributors. The likelihood ratio is

$$L = \frac{P_1(e|abcde)}{P_2(cde|abcde)}.$$

Assuming allelic independence, and the absence of effects of relatives, inbreeding or population structure,

$$P_x(u|e) = T_0^{2x} - \sum_i T_i^{2x} + \sum_i \sum_{j \neq i} T_{ij}^{2x} - \cdots,$$

where T_0 is the sum of frequencies of all the alleles in $\{e\}$, T_i is T_0 minus the frequency of the ith allele in $\{u\}$, T_{ij} is T_i minus the frequency of the jth allele in $\{u\}$, and so on [3].

Conclusion

This discussion has focused on the probabilities of coincidental matches of genetic profiles from two people. It has ignored nongenetic issues, such as the possibility that a match has been declared falsely, whether by error or fraud. The discussion may be of temporary relevance because of the increasingly discriminatory power of genetic profiles. The technology has advanced to the point where it may no longer be reasonable to believe that any two people, identical twins excepted, could have the same profile. At that point, the term "genetic fingerprint" would be appropriate.

References

[1] Balding, R.A. & Nichols, R.A. (1994). DNA profile match frequency calculations: how to allow for population

stratification, database selection and single bands, *Forensic Science International* **64**, 125–140.

[2] Weir, B.S. (1996). *Genetic Data Analysis*, Vol. II. Sinauer, Sunderland.

[3] Weir, B.S., Triggs, C.M., Starling, L., Stowell, L.I., Walsh, K.A.J. & Buckleton, J.S. (1997). Interpreting DNA mixtures, *Journal of Forensic Science* **42**, 113–122.

B.S. WEIR

Statistical Map

A map can be defined as "a collection of spatially defined objects" [13, 14].

As such, a map is simply a display of the spatial properties of an object set. This usually implies a two-dimensional display of the Cartesian or polar coordinate locations of objects and also their attributes, e.g. a street map displays the locations of streets and houses on these streets (if the *resolution* of the map is high enough). In addition, the houses may have attributes that relate to the population of each household. Hence a variety of maps could be constructed even from this simple example. We could have a simple street map, a more detailed house map, and a map of household attributes at the highest resolution. The display of such varied information in a graphical form has been the concern of *cartography* for a considerable time [13]. Many of the concerns of those within statistics about the representation of data in graphical forms have also been explored within geography for mapped displays. The psychological/visual perceptual implications of chosen mapping methods has been studied extensively [14, Chapters 3– 6], and these issues also apply to the construction of maps of statistical information. Walter [22] has examined visualization issues related to medical mapping. The stages of map construction can each be associated with some form of processing of spatial information and hence can be of concern to anyone wishing to use such methods of presentation.

The main stages are:

1. choice of scale
2. choice of symbolization or representational processing
3. further processing required to construct a suitable map.

In stage 1, a suitable scale for the map must be chosen. Any choice of scale, however, inevitably leads to a process of *averaging* of spatial information from higher levels of resolution. For instance, a street map of a city will usually be represented as sets of linear features depicting street locations, but if a larger country scale was to be used, within which the city was but a small part, then the city streets could be represented by a dot. Hence in this case, the scale change has resulted in averaging of the spatial information. Stage 2 is also represented in the street map example. At the detailed scale, linear features represent the streets, while at the country scale the whole city is represented by a dot. This represents a change in symbolic representation as well as scale change. This can both have a visual perceptual effect for the map user and represent an averaging of spatial information. Stage 3, that of further processing, can occur when information on the spatial structure of the objects and/or attributes is not available in the form required by the representational system. For example, often one needs to compute a map representation from a set of sampling points that are predefined, whereas we need to have measurements at the intersections of a fixed grid which do not correspond to the sampling points. This arises in many statistical mapping problems and leads to the use of *interpolation* or smoothing of data. Another example of further processing is the use of **transformations** of the mapped data to represent some feature of the spatial structure. Map projections [14, Chapter 2] are a classic example of transformation. Schulman et al. [17] give an example of using projection and transformation in a medical statistical application.

Hence, in two of the three stages of map production, some form of statistical processing of the spatial information usually occurs. This applies in most forms of mapping exercise and hence it can be claimed that map construction is, to a large extent, a statistical processing task. Figures 1 and 2 display the transition between street and city level representations.

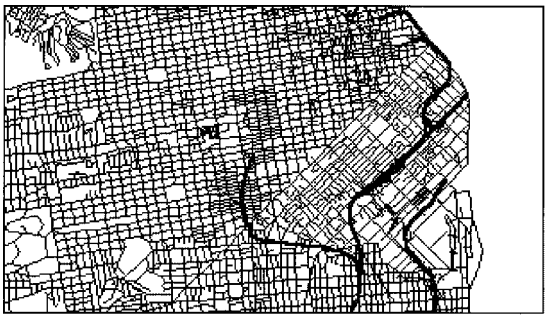

Figure 1 The streets of San Francisco: street level map. Map made with Mapinfo Professional®. © 1997 Mapinfo Corporation

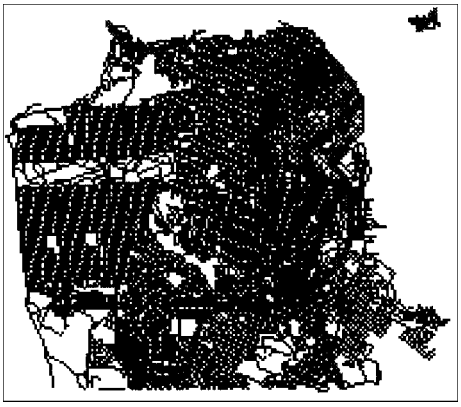

Figure 2 The streets of San Francisco: city level zoom scale change. Map made with Mapinfo Professional®. © 1997 Mapinfo Corporation

Statistical Maps and Mapping

The three stages of map production discussed above map easily onto the data types that are often the basic ingredients for mapped representation. Within the subject of spatial statistics a spectrum of spatial information and data formats is found. This spectrum ranges from the locations of points or objects (point and object processes; *see* **Point Processes**) to the measurements made on random variates at specific spatial locations (random fields). In the former case, the subject area of stochastic geometry concerns the probabilistic modeling of the locations of objects [19]. In the latter case, the subjects of geostatistics and image processing (*see* **Image Analysis and Tomography**) deal with observations made on random fields [2]. Image processing characteristically studies random fields observed on a grid mesh of regular sampling points (pixels), and its task is usually restricted to the processing of the pixel data to obtain the underlying "ground truth" or noise-free image. Hence, this form of processing is not closely akin to mapping as there is usually no need for interpolation or scale averaging. However, the subject of geostatistics does involve smoothing and interpolation and can involve the estimation of areas or blocks of information that are averages of underlying sampling point data. In addition, the analysis of object processes often involves the averaging and scale change from locational data to localized intensity data, i.e. the locations of objects are converted into a continuous surface describing the local density/intensity of objects. Both of these data types lead to scale change and interpolation/smoothing operations that are integral to the mapping process.

As image processing can be considered a special case of geostatistics, for brevity we will consider here the construction of maps and map interpretation and properties for object processes and for geostatistical data only. A review of map construction issues for disease atlases can be found in [3] (*see* **Mapping Disease Patterns**).

Object Process Mapping

An object process map is a presentation of the spatial locations of objects, usually in two dimensions. Define $\mathbf{x}_i, i = 1, \ldots, n$, to be the locations of the objects within a spatial window T. The area of T is denoted $|T|$. Usually objects are mapped at a specified point (the associated point), which can be uniquely identified for each object. For example, a process of circles could have the circle centers as associated points. Hence to construct a map of such a process it suffices to plot the locations of such points and then to construct circles with given radii. For this example, the locations of the circles could follow a **stochastic process** and the circle radii could be the realization of a **random variable**. A simpler example of this idea is the **point process**, which simply has a point location as its observation unit and the realization of point locations are the objects. For example, the address locations of cases of a disease form a point process and a map of all addresses of disease within T would be a mapping of the process. Figure 3 depicts a case address map for respiratory cancer in a small Scottish town for the period 1966–1976.

Often it is important to transform an object map by converting the object locations into a continuous surface representation of the objects. This kind of transformation can be achieved by computing the local density of objects. **Density estimation** [18] can be used to provide such local densities and the resulting density surface can be mapped over the study window. Usually such a surface is displayed as a contour plot or, in three dimensions, as a surface perspective view. The contour plot is often preferred, as some spatial information is hidden in perspective views. To demonstrate how scale and symbolization affect such mapping, the contour plot of a density estimate of the case event data in Figure 1 has been drawn for two different

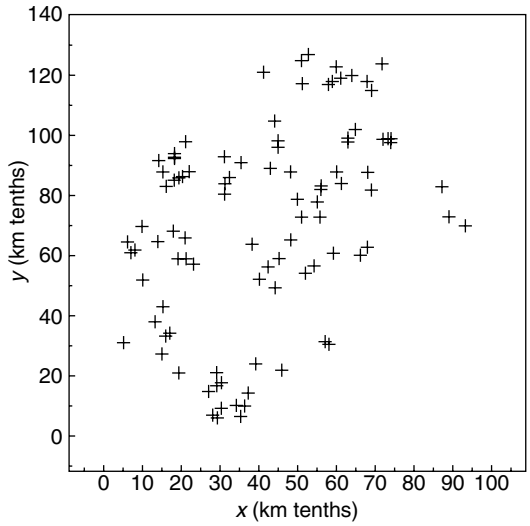

Figure 3 Arbroath: central Scotland: object map of cases of respiratory cancer within a fixed time period. Map made with Mapinfo Professional®. © 1997 Mapinfo Corporation

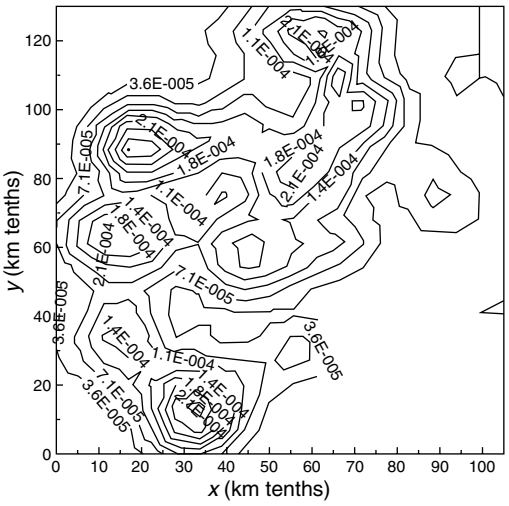

Figure 4 Contour map of case events in Figure 3: 10 contour heights. Map made with Mapinfo Professional®. © 1997 Mapinfo Corporation

contour densities (10 and five heights) in Figures 4 and 5. Note that the arbitrary choice of fewer contours effectively produces a smoother surface and can change the perception of the object map. In addition, the derivation of these contour maps has proceeded through a number of stages that may affect the final

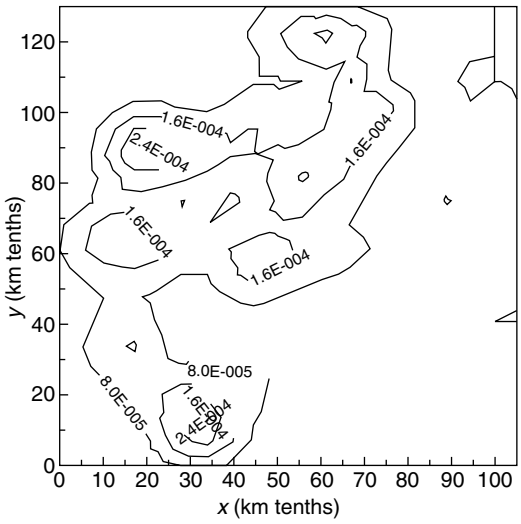

Figure 5 Contour map of case events in Figure 3: five contour heights. Map made with Mapinfo Professional®. © 1997 Mapinfo Corporation

visualization. First, the process of density estimation involves the production of estimates in a grid mesh (interpolation) and the choice of a smoothing constant (bandwidth) that controls the smoothness of the gridded data. Then a graphic package has constructed contours using a further interpolation/smoothing step.

Geostatistical Mapping

Geostatistical data differ from the above in that a network of sites is usually used to sample or measure some spatially distributed variate. For example, the early geostatistical work related to estimation of geological structures in mining applications where concentrations of particular minerals were sampled at fixed locations [2, 21]. Within biostatistics, many examples can be found where data are sampled at spatial sites. One common example is the mapping of disease rates located at the centroids of small geographic areas (*see* **Geographic Patterns of Disease**), such as census tracts. While the rate represents an average over the whole region, the approximation of allocating the rate to a centroid is often made. In this case the rate (e.g. a standardized mortality ratio (SMR) [7]; *see* **Standardization Methods**) is regarded as being associated with a fixed spatial location. Another application arises when a

spatially distributed covariate must be interpolated to a set of locations within an ecologic study. An example of this would be the use of interpolated pollution measurements as a covariate in a study of the distribution of respiratory disease morbidity (e.g. asthma). In principle, the basic mapping considerations apply in this case also: for visualization, the data can be displayed as an object map with each sample site becoming the location of an object representing the measurement at that site. For example, a circle of radius equal to the measurement could depict the distribution. Other display forms are available, such as needle plots where vertical lines of length scaled to represent the measurement are drawn at the sites [15]. Often a surface interpolated from the measured data is to be constructed. This surface also requires an interpolation or smoothing step to provide a gridded data set, which can be subsequently contoured. Such interpolation can be achieved by a wide range of smoothing techniques. The method of *Kriging* was developed within geostatistics to provide such processing. This method is not directly applicable to data that have a positivity constraint (e.g. SMRs or counts), but can be modified [10, 12].

Other notable forms of smoothing available for such data are: **nonparametric regression** or *kernel smoothing* [6], and *thin plate* **splines** [5, Chapter 7]. A wide variety of mathematical interpolation methods are available also, e.g. finite element methods [8].

Statistical Accuracy

Any step of map production which requires statistical **estimation** will have associated with it a measure of the reliability or variability of that estimation. Hence any map of estimated values (such as interpolated or smoothed data) should have a variance estimate available at the estimation points. The variance estimate can also be represented as a surface, or a pointwise **confidence interval** for the estimated surface can be produced. The visualization of such surfaces can cause some problems as there are no simple clear methods of displaying multiple surfaces without losing spatial information. If areas of the estimated surface that exceed limits of variability are of interest, then it may be possible to construct a **Monte Carlo *P*-value** surface [4] (*see* **Markov Chain Monte Carlo**).

Edge Effects

In most mapping exercises where statistical data are to be represented, edge effects are present and may require to be accommodated in the analysis. When spatial data are spatially **autocorrelated** then observations made within a study window will relate to unobserved data outside the window. This is a form of spatial **censoring**. Even when data are not autocorrelated, the method used to estimate the smoothed surface representation of the data will have greater variability at the edges. This is because such smoothing operators use neighboring data observations to compute estimates and at edges these neighborhoods are censored. Also, if only data *within* the window are used to estimate edge values, then a bias will appear in this edge estimation. The use of *guard areas* or *data augmentation* at the edges of maps may be useful [16, 20]. An example of a disease mapping application where edge effects may have a significant impact is given in [1].

Aggregation

Finally, it is important to consider the interconnection between some mapping concepts and the related statistical issue of aggregation. The effect of aggregation of data into spatially larger areas has a variety of effects on the subsequent interpretation. First, aggregation is a scale change; by accumulating observations into larger spatial units the scale of analysis is changed. In addition, aggregation acts as a smoothing operation. That is, by accumulation of data detailed variation in the data will be lost and will not be retrievable. A classic example of this is the arbitrary regionalization of case events into **census** tracts in medical small area studies. In that case the detailed spatial variation of cases is lost within the census tract count (for discussion see [9] or [11]). This type of averaging of spatial effects is inherent in scale changes, and it is important that any spatial structural effects observed in data at one scale are scale labeled, i.e. the scale at which the effect is found is permanently associated with the effect. For example, clustering of disease data in space may occur on a case event map, but when aggregated into census tract counts this effect may disappear.

References

[1] Clayton, D. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds. Oxford University Press, Oxford, pp. 205–220.

[2] Cressie, N.A.C. (1991). *Statistics for Spatial Data*. Wiley, New York.

[3] Esteve, J., Benhamou, E. & Raymond, L. (1994). *Descriptive Epidemiology*, Number 128 in Statistical Methods in Cancer Research, Vol. IV. International Association for Research in Cancer, Lyon.

[4] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

[5] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalised Linear Models*. Chapman & Hall, London.

[6] Härdle, W. (1991). *Smoothing Techniques: With Implementation in S*. Springer-Verlag, New York.

[7] Inskip, H., Beral, V., Fraser, P. & Haskey, P. (1983). Methods for age-adjustment of rates, *Statistics in Medicine* **2**, 483–493.

[8] Lancaster, P. & Salkauskas, K. (1986). *Curve and Sur-Face Fitting: An Introduction*. Academic Press, London.

[9] Lawson, A.B. (1993). On the analysis of mortality events around a prespecified fixed point, *Journal of the Royal Statistical Society, Series A* **156**, 363–377.

[10] Lawson, A.B. (1994). On using spatial gaussian priors to model heterogeneity in environmental epidemiology, *Statistician* **43**, 69–76.

[11] Lawson, A.B. & Waller, L. (1996). A review of point pattern methods for spatial modelling of events around sources of pollution, *Environmetrics* **7**, 471–488.

[12] Lawson, A.B., Biggeri, A. & Lagazio, C. (1996). Modelling heterogeneity in discrete spatial data models via map and mcmc methods, in *Proceedings of the Eleventh International Workshop on Statistical Modelling*, A. Forcina, G. Marchetti, R. Hatzinger, & G. Galmacci, eds. Graphos, Citta di Castello, pp. 240–250.

[13] MacEachren, A.M. (1995). *How Maps Work: Representation, Visualisation, and Design*. Guildford Press, New York.

[14] Monmonier, M. (1996). *How to Lie with Maps*, 2nd Ed. University of Chicago Press, London.

[15] Ripley, B.D. (1981). *Spatial Statistics*. Wiley, New York.

[16] Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

[17] Schulman, J., Selvin, S. & Merrill, D.W. (1988). Density equalised map projections: a method for analysing clusters around a fixed point, *Statistics in Medicine* **7**, 491–505.

[18] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[19] Stoyan, D., Kendall, W.S. & Mecke, J. (1987). *Stochastic Geometry and Its Applications*. Akademie-Verlag, Berlin.

[20] Tanner, M. (1991). *Tools for Statistical Inference*, Springer Series in Statistics. Springer-Verlag, New York.

[21] Wackernagel, H. (1995). *Multivariate Geostatistics*. Springer-Verlag, New York.

[22] Walter, S.D. (1993). Visual and statistical assessment of spatial clustering in mapped data, *Statistics in Medicine* **12**, 1275–1291.

(*See also* **Geographic Epidemiology**; **Graphical Displays**)

A.B. LAWSON

Statistical Methods in Medical Research

Statistical Methods in Medical Research is an international review journal that was first published in 1992. Its aim is to provide medical statisticians and others with up-to-date reviews of those areas of statistics that are most important in medical and health investigations. The journal is published six times a year, with four of the issues consisting of four or five commissioned review papers on a particular topic, and the other two papers submitted and refereed in the usual way. Each issue also contains an editorial either by one of the editors (B.S. Everitt and T. Holford) or by a guest editor, most often chosen from the 25 or so members of the editorial board, and a number of book reviews.

Some recent issues of the journal have dealt with

1. Multi-state models
2. Nonparametric longitudinal data analysis
3. Ethics, Statistics and Statisticians

The content of the issue on Ethics was as follows:

1. On the ethical aspects of the testimony of statisticians in court, E.A. Gehan.
2. The ethics of consulting for the tobacco industry, D.R. Rubin.
3. Ethics, data-dependent designs, and the strategy of clinical trials: time to start learning as we go? C.R. Palmer.
4. Ethical considerations concerning treatment allocation in drug development trials, S. Senn.
5. Placebos that harm: sham surgery controls in clinical trials, A.J. London and J.B. Kadane.
6. Ethical issues in oncology biostatistics, P.F. Thall.

The journal is published by Edward Arnold, London.

BRIAN S. EVERITT

Statistical Review for Medical Journals, Guidelines for Authors

In 1978, a panel of statisticians, practicing physicians, and medical editors convened at a statistical conference and concluded that medical articles often contained incomplete reporting of statistical results and methods, and that this problem was associated with the use of faulty statistical designs and analyses in published medical studies [11]. The panelists recognized the need for complete statistical reporting, to make medical studies both interpretable and convincing, and the role of medical journals in enforcing good statistical practice and reporting. The panel proposed the development of standards governing the format and content of reports of statistical methods and results in medical articles.

In 1980, Mosteller et al. [10] reviewed the original reports of 147 cancer trials cited in a comprehensive clinical anthology, and they found that important statistical issues, in particular the statistical methods used, were reported only 25% of the time, with statistical **power** and intended sample size (*see* **Sample Size Determination**) almost never reported. DerSimonian et al. [5] reviewed the reports of a chronological sample of 67 **clinical trials** appearing in four leading British and American medical journals, in 1979 and 1980, and found that patient eligibility (*see* **Eligibility and Exclusion Criteria**), method of **randomization**, loss to follow-up, statistical methods, and statistical power were reported ambiguously, or not at all, on average 44% of the time. Other surveys focused on individual issues. Schulz et al. [13] found that ambiguous reporting of the method of concealment of the treatment allocation (*see* **Blinding or Masking**) was associated with larger estimates of treatment effect, suggesting that inadequate methods might lurk behind inadequate reporting. Pocock et al. [12] found poor reporting of prioritization of **multiple endpoints** and statistical comparisons, of whether subgroup analyses or strategies for multiple testing over time were pre-defined (*see* **Multiple Comparisons**), and of whether sample size and interim stopping rules (*see* **Sequential Analysis**) were defined prior to the start of the study (*see* **Treatment-covariate Interaction**). They indicated that, without such provisions, multiple testing leads to serious problems of interpretation, since the nominal significance levels no longer reflect the true significance levels. For repeated interim monitoring, in particular, this was dramatically demonstrated by the simulations of Green & Fleming [7]. Moher et al. [9], in a comprehensive review of the negative trials published in three leading British and American medical journals in 1975, 1980, 1985, and 1990, found that the majority failed to report an intended sample size. A related finding was that 64% of the trials failed to have at least 80% power to detect a 50% relative improvement. Finally, George [6], in a survey of 98 medical journals (with 83 respondents), found only 16% to have a policy guaranteeing statistical review.

There have been a series of efforts by statisticians and clinicians to address the above issues. Altman et al. [1] published a detailed and comprehensive set of statistical guidelines for medical journals. Bailar & Mosteller [3], expanding on the concise guidelines adopted by the International Committee of Medical Journal Editors [8], produced a similarly comprehensive set. Zelen [17] and Simon & Wittes [14] focused their detailed guidelines on the particular issues of clinical trial reporting, as did Baar & Tannock [2] in the annotations of their whimsical twin examples of a well executed trial report and a poorly executed one. Recently, a working group of statisticians and clinicians [4], made up of representatives of two previous such groups [15, 16], issued the "CONSORT Statement", a proposal for the structured reporting of clinical trials, following a detailed list of statistical guidelines.

The following is a melding of the above sets of guidelines, structured according to and closely following the CONSORT proposal, but expanded to include additional stipulations made by the other authors. It is not meant to be either definitive or all-inclusive. Each statement is referenced to allow the reader to refer to the original sources for further explanation.

Structured Statistical Guidelines for Medical Journals

Introduction

State prospectively defined hypotheses and planned subgroup or **covariate** analyses [1, 4].

Methods: Protocol Design

Describe the planned study population, together with inclusion/exclusion criteria [1–4, 14, 17].

Give the primary and secondary **outcome measure(s)** and the minimum important difference(s), and indicate how the target sample size was projected [2, 4, 14].

Describe the rationale and methods for statistical analyses, giving the main comparative analyses and whether they were completed on an **intention-to-treat** basis, with enough detail to permit replication [1, 3, 4, 14]. Give assumptions concerning the distribution of the variables which underly the statistical methods used [1]. For randomized studies, an intent-to-treat comparison should be included for major endpoints [14].

Give the prospectively defined stopping rules [4, 14, 17].

For observational studies, explain the design, describing the selection of **controls** and the **matching** procedures, whether the study is **case–control**, **cross-sectional**, or **cohort**, and what was the participation rate [1].

Methods: Treatment Assignment

Give the unit of randomization (e.g. individual, cluster, or geographic) (*see* **Unit of Analysis**) [3, 4].

Describe the method used to generate the allocation schedule (*see* **Randomization**; **Randomized Treatment Assignment**) [1, 3, 4, 17].

Describe the method of allocation concealment and the timing of assignment [4, 17].

Give the number of eligible patients not entered or not randomized, and the reasons [2, 4, 17].

Methods: Treatment Blinding

Describe the mechanism of treatment blinding, if used, and the evidence for successful blinding of subjects and investigators, as appropriate [1, 3, 4].

Methods: Quality Control

Briefly describe the methods used to ensure that the data are complete and accurate, that all patients entered on study are reported, and that the assessment of major endpoints is reliable [14, 17]. The study should not have an inevaluability rate for major endpoints in excess of 15% [14] (*see* **Missing Data in Clinical Trials**).

Results: Follow-up Schedule and Loss to Follow-up

For each randomized group, give the timing of follow-up and the number of patients withdrawn or lost to follow-up [3, 4, 14, 17]. Not more than 15% of eligible patients should be lost to follow-up [14].

Results: Analysis

State the estimated effect of treatment on primary and secondary outcome measures, including the point estimate (*see* **Estimation**) and **confidence interval** [1, 3, 4, 14]. Give precise ***P* values**, but not to more than two or three decimal places [1, 3].

Claims of therapeutic efficacy should be based upon comparisons with a control group, except in special circumstances, such as when each patient is his own control [14]. Where historical controls are used, patient characteristics should be compared in detail with those of the experimental group, and potential sources of **bias** should be discussed [14, 17]. Comparison of survival between responders and nonresponders can not be used to establish efficacy [2, 14].

Significance tests not relating to pre-specified hypotheses must be considered exploratory [1] (*see* **Hypothesis Testing**). Claims of subset-specific treatment differences must be documented to be based on more than the random results of multiple-subset analyses [14].

Present summary data and appropriate descriptive and inferential statistics (*see* **Inference**) in sufficient detail to permit alternative analyses and replication [1, 14, 17].

Cite statistical software packages used [1, 3] (*see* **Software, Biostatistical**).

Do not use technical statistical terms, such as significance and **correlation**, in a general fashion [1, 3].

Describe **prognostic factors**, by treatment group, and any attempt to adjust for them [1, 4, 17].

Describe protocol deviations (*see* **Clinical Trials Protocols**), including the number of randomized patients subsequently found ineligible or not treated as assigned, together with the reasons [1, 4, 14]. In general, observations that appear to be inconsistent with the main body of data should not be excluded unless there are additional reasons to doubt

their credibility, and any such exclusion should be reported [1].

Comment

State specific interpretation of study findings, including sources of **bias** and imprecision [4].

State general interpretation of the data in light of the totality of the available evidence [4].

References

[1] Altman, D.G., Gore, S.M., Gardner, M.J., Pocock, S.J. (1983). Statistical guidelines for contributors to medical journals, *British Medical Journal* **286**, 1489–1493.

[2] Baar, J. & Tannock, I. (1989). Analyzing the same data in two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials, *Journal of Clinical Oncology* **7**, 969–978.

[3] Bailar, J.C. & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals, *Annals of Internal Medicine* **108**, 266–273.

[4] Begg, C., Cho, M., Easwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. & Stroup, D.F. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement, *Journal of the American Medical Association* **276**, 637–639.

[5] DerSimonian, R., Charette, L.J., McPeek, B. & Mosteller, F. (1982). Reporting on methods in clinical trials, *New England Journal of Medicine* **306**, 1332–1337.

[6] George, S.L. (1985). Statistics in medical journals: a survey of current policies and proposals for editors, *Medical and Pediatric Oncology* **13**, 109–112.

[7] Green, S.J. & Fleming, T.R. (1988). Guidelines for the reporting of clinical trials, *Seminars in Oncology* **15**, 455–461.

[8] International Committee of Medical Journal Editors (1988). Uniform requirements for manuscripts submitted to biomedical journals, *Annals of Internal Medicine* **108**, 258–265.

[9] Moher, D., Dulberg, C.S. & Wells, G.A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials, *Journal of the American Medical Association* **272**, 122–124.

[10] Mosteller, F., Gilbert, J.P. & McPeek, B. (1980). Reporting standards and research strategies for controlled trials: agenda for the editor, *Controlled Clinical Trials* **1**, 37–80.

[11] O'Fallon, J.R., Dubey, S.D., Salsburg, D.S., Edmonson, J.H., Soffer, A. & Colton, T. (1978). Should there be statistical guidelines for medical research papers?, *Biometrics* **34**, 687–695.

[12] Pocock, S.J., Hughes, M.D. & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials: a survey of three medical journals, *New England Journal of Medicine* **317**, 426–432.

[13] Schulz, K.F., Chalmers, I., Hayes, R.J. & Altman, D.G. (1995). Empirical evidence of basis: dimensions of methodological quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association* **273**, 408–412.

[14] Simon, R. & Wittes, R.E. (1985). Methodologic guidelines for reports of clinical trials, *Cancer Treatment Reports* **69**, 1–3.

[15] Standards of Reporting Trials Group (1994). A proposal for structured reporting of randomized controlled trials, *Journal of the American Medical Association* **272**, 1926–1931.

[16] Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature (1994). Call for comments on a proposal to improve reporting of clinical trials in the biomedical literature, *Annals of Internal Medicine* **121**, 894–895.

[17] Zelen, M. (1983). Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors, *Journal of Clinical Oncology* **1**, 164–169.

LAWRENCE V. RUBINSTEIN

Statistical Review for Medical Journals, Journal's Perspective

Recently I met a woman doctor who had just read the medical journal I edit – the *British Medical Journal* (*BMJ*) – for the first time in 25 years. "It doesn't have any medicine in it any more. It's full of statistics," she complained. To begin an article in a biostatistical encyclopedia with a study that has a sample of one may seem ingenuous in the extreme, but such Rip van Winkle characters are rare: most doctors who stopped reading medical journals 20 years ago have never started again. And I am fascinated that her first observation should be the predominance of statistics. It could have been the appearance of molecular biology or references to computers or the increase in sociological, economic, and political material. Perhaps the biggest change in medical journals in the past 25 years has been the dramatic increase in the amount of statistical material they contain.

Medical journals of 25 years ago did, of course, contain some statistics. **Austin Bradford Hill** published a ground-breaking series of articles on statistics in the *Lancet* in 1937, [14]. Hugh Clegg, the editor of the *BMJ* from 1947 to 1965, was a personal friend of Bradford Hill and was persuaded by him of the importance of statistics for medicine and medical journals. The *BMJ* carried one of the first randomized controlled trials (*see* **Clinical Trials, Overview**) in 1948 (*see* **Medical Research Council Streptomycin Trial**), and in 1976 the journal published a book on elementary statistics written not by a statistician but by the deputy editor, Dougal Swinscow [18]. The book, *Statistics at Square One*, has remained in print ever since and sold almost 100 000 copies – three times as many copies as any other book published by the *BMJ* Publishing Group. Many other journals have also published educational articles on statistics, and the *BMJ* has published further books.

But a survey of all papers published during 1978–79 in the *New England Journal of Medicine (NEJM)*, the world's leading general medical journal, showed that 58% included no statistical method or descriptive statistics only [8]. Three-quarters of the original papers did, however, include statistical methods. By 1990 this proportion had increased to 89%, and there was a marked increase in the use of more advanced methods of analysis [3]. Generally, medical journals have moved away from the case reports and accounts of series of patients (*see* **Case Series, Case Reports**) that were once their staple fare to experimental studies, particularly randomized controlled trials, and epidemiologic studies that use increasingly complex statistical methods. Medical journals have, however, moved at different rates. The leading general medical journals – the *Annals of Internal Medicine, BMJ, Journal of the American Medical Association (JAMA), Lancet*, and *NEJM* – increased the proportion of papers that depended on statistical analysis much more rapidly than specialist medical journals.

The appearance of statistical methods in medical journals was quickly followed by a realization that the methods were commonly misused. A great many studies have now shown that basic statistical errors – in design, analysis, and presentation – are common in medical journals [3]. A particular problem has been many studies too small to reach a confident conclusion of the absence of an effect. Many of the statistical errors found in medical papers were so serious that the conclusions of the papers were not supported by the evidence they contained (*see* **Statistical Review for Medical Journals**).

The Story of Statisticians Moving to the Heart of the *BMJ*

The appearance of studies on poor statistics in medical journals led editors, most of whom had little or no training in statistics, to recognize that they needed help. From the late 1970s medical journals began to recruit statistical advisers, although there are still many medical journals that have no regular statistical advice. I arrived at the *BMJ* in May 1979, and I have lived through the incorporation of medical statisticians into the heart of the editorial process. I want to describe the path taken by the *BMJ* – partly because I know it so well, and partly because it is, I know, very similar to the path taken by most medical journals and still being followed by many.

We began in the 1970s by recruiting a statistical adviser, the late **Martin Gardner**. He assembled a small group of statisticians who agreed to review *BMJ* papers. We thought it essential from the beginning to work with statisticians who had practical experience of medical research. We were having to

learn each other's way of thinking, and we worried that the gulf between medical editors and statisticians with no knowledge of medical research would be unbridgeable. In the early days we made the mistake of thinking that statistics was a much more exact science than clinical research and that we had to go along with exactly what the statisticians advised. Eventually we learnt that there was room for negotiation over what was acceptable. We, as editors, always did – and still do – feel vulnerable when caught in the middle of statisticians arguing with each other.

In the early days we sent our statistical advisers only those papers where we were worried about the statistics. This usually meant that we sent papers that had complex statistics, but we gradually learnt that some of the most egregious errors occurred in papers that used only simple statistical tests. There might be errors in design, sampling, and data collection. We sent only some papers because the statisticians were seen as a limited resource. But slowly over 5–10 years we moved to the point that all research papers with any statistical content at all, which is virtually all of them, were sent for a statistical opinion before publication. Statisticians are, however, usually involved towards the end of our peer review process. We reject 85% of papers (a figure that is roughly the same for the main general journals), and about 60% are rejected without ever having had a statistical opinion. Letters – which in the case of the *BMJ* are all in response to material already published in the journal – are still published without a routine statistical opinion, although we ask for one if we are worried.

A very important step for us was when we began to involve statisticians in our "hanging committee" – the committee of two editors and two outside doctors that takes the final decision on whether to publish research papers. This committee meets every week, and from the mid-1980s we began to include a statistician on some occasions and from the early 1990s on every occasion. The beauty of this arrangement is that it means that the editors, doctors, and statistician can discuss all aspects of a paper together – recognizing the inevitable trade-offs between statistical purity, what can actually be done in clinical research, and what matters to doctors treating patients. My experience is that being able to discuss a paper with a statistician is much preferable to simply having a written report. This is particularly because peer review is in my mind more about improving the papers we do publish than simply deciding which to publish. Another advantage of being able to discuss papers is that it is highly educational for those attending: we learn from each other in a way that is not possible if everything is done on paper. After moving to a system of having a statistician present at every meeting, none of the editorial team could imagine moving back to a system where they were not present.

Statistical Policies

One of the first jobs of our statistical advisers was to develop published advice to authors on statistics and to produce checklists that could be used when assessing papers. They also advised us on statistical policy. In 1983 we published "for debate" comprehensive guidelines on statistical aspects of manuscripts [5]; these were soon recommended in our Instructions to Authors (*see* **Statistical Review for Medical Journals, Guidelines for Authors**). In 1986 we published checklists used by statistical referees [12] and began to require **confidence intervals** whenever appropriate (see [11]). From the early 1990s we have not published controlled trials in which patients are allocated to different interventions in any way other than randomly (unless an acceptable argument is given on why random allocation was not possible; *see* **Randomized Treatment Assignment**) [2]. Currently we are moving towards policies on publishing **absolute risks** as well as **relative risks** and including the "**number needed to treat**" in appropriate studies. A debate is beginning on including much more **Bayesian** statistics in medical journals [10]. And medical journals have also become increasingly interested in **meta-analyses** and publish more and more [7].

Standards Across Journals

The result of these moves is that statisticians have become central to the peer review process of the *BMJ*. The same is true of other leading journals, but not of all journals. George [13] and Altman [1] have produced recommendations on the statistical aspects of medical journals:

1. All papers should be reviewed by a statistician prior to publication (perhaps only after a favorable subject-matter review).

2. Journals should recruit statistical reviewers.
3. The statistical reviewers should at least be offered the option to see the revised manuscript.
4. Journals should publish their policy on statistical review.
5. Journals should adopt written standards or guidelines for statistical reporting (usually previously published guidelines).

Altman observed some years ago that few journals meet more than two of these five recommendations [3], but my impression is that more and more are moving towards fulfilling all five. I sit on the editorial boards of the 25 specialist journals of the *BMJ* Publishing Group, and most are beginning to follow all five recommendations.

Another important move is that statisticians who advise different journals are coming together to produce advice that will be useful to all medical researchers and journals. A good example is the CONSORT proposals on how to present randomized controlled trials, [6]. These have already been adopted by more than 20 journals and will certainly be accepted by many more. Statisticians have also persuaded the International Committee of Medical Journal Editors to produce general guidelines on the use of statistics in medical journals [15].

All this activity should be raising the quality of statistics in medical journals, but there is so far little hard evidence that this is the case. I think that we urgently need to study whether the standard of statistics has improved.

Statisticians: A Fundamental Influence on Medicine

The influence of statisticians on medical journals – and medicine – has, I believe, been fundamental, and we are only just becoming aware of their full impact. Medicine is currently experiencing an intense debate over the need to move towards **evidence-based medicine** [17], and statisticians have been key in this debate [4]. Evidence-based medicine is a movement that encourages doctors to base their practice on the best evidence available. All doctors know that much medical practice is based on opinion and experience rather than on scientifically sound evidence, but they also recognize that good scientific evidence is not available on whether much of medical practice is effective. Those who are encouraging the move to evidence-based medicine are trying to work out what doctors know and what they do not know; promote research into what is not known; extract from medical journals the small proportion of studies that are scientifically sound and disseminate them; produce evidence-based guides to practice; and interest practitioners in critically appraising the evidence presented to them. All of this activity is based on statistical ideas of what constitutes good evidence. This movement will, many predict, transform medical practice.

Another important area where statisticians have had a great influence is in identifying and managing the problem of scientific fraud. Editors of medical journals – and other scientific journals – are learning that some of the papers submitted to them are fraudulent [16]. Researchers may have invented data, stolen them, or manipulated them in dishonest ways. Peer review does not easily detect fraud, but statisticians are able to help. Some statisticians, for instance, have developed tests that will help identify "the fingerprint" of invented data [9].

Statisticians have also raised consciousness of more minor scientific dishonesty. Since doctors have had access to computer programs to help them analyze data they have learnt that they can "torture the data until they confess". For instance, if doctors do not find a significant difference between placebo and a treatment in a randomized controlled trial they may keep doing subgroup analyses until they find a significant difference. Statisticians have taught us that we must declare the hypotheses we are going to test in advance, and the *post hoc* analyses are suspect. There are many other similar concerns.

My conclusion has to be that statisticians have become central to medical journals and to medical practice. There is still much room to raise statistical standards in medical journals, but statisticians have done more than simply improve statistical design, analysis, and presentation in medical journals. They have played an important part in what is looking increasingly like a paradigm shift in medicine – from opinion to evidence-based medicine.

References

[1] Altman, D.G. (1982). Statistics in medical journals, *Statistics in Medicine* **1**, 57–71.

[2] Altman, D.G. (1991). Randomization, *British Medical Journal* **302**, 1481–1482.

[3] Altman, D.G. (1991). Statistics in medical journals: development in the 1980s, *Statistics in Medicine* **10**, 1897–1913.

[4] Altman, D.G. (1994). The scandal of poor medical research, *British Medical Journal* **308**, 283–284.

[5] Altman, D.G., Gore, S.M., Gardner, M.J. & Pocock, S.J. (1983). Statistical guidelines for contributors to medical journals, *British Medical Journals* **286**, 1489–1493.

[6] Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I. et al. (1996). Improving the quality of reporting of medical studies: the CONSORT statement, *Journal of the American Medical Association* **276**, 637–639.

[7] Chalmers, I. & Altman, D.G. (1995). *Systematic Reviews*. *BMJ* Publishing Group, London.

[8] Emerson, J.D. & Colditz, G.A. (1983). Use of statistical analysis in the New England Journal of Medicine, *New England Journal of Medicine* **309**, 707–713.

[9] Evans, S.J.W. (1996). Statistical aspects of the detection of fraud, in *Fraud and Misconduct in Medical Research*, 2nd Ed. S. Lock & F. Wells, eds. BMJ Publishing Group, London 226–239.

[10] Freedman, L. (1996). Bayesian statistical method, *British Medical Journal* **313**, 569–570.

[11] Gardner, M.J. & Altman, D.G. (1989). *Statistics with Confidence*. BMJ Publishing Group, London.

[12] Gardner, M.J., Machin, D. & Campbell, M.J. (1986). Use of check lists in assessing the statistical content of medical studies, *British Medical Journal* **292**, 810–812.

[13] George, S.L. (1985). Statistics in medical journals: a survey of current policies and proposals for editors, *Medical and Pediatric Oncology* **13**, 109–112.

[14] Hill, A.B. (1937). Principles of medical statistics. I: The aim of statistical method, *Lancet* **i**, 41–43.

[15] International Committee of Medical Journal Editors (1997). Uniform requirements for manuscripts submitted to biomedical journals, *Journal of the American Medical Association* **277**, 927–934.

[16] Lock, S. & Wells, F. (1993). *Fraud and Misconduct in Medical Research*. BMJ Publishing Group, London.

[17] Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M. & Haynes, R.B. (1996). Evidence based medicine: what it is and what it isn't, *British Medical Journal* **312**, 71–72.

[18] Swinscow, D. (1976). *Statistics at Square One*. London; *BMJ* Publishing Group, London.

RICHARD SMITH

Statistical Review for Medical Journals

The statistical content of medical and epidemiologic journals has undergone a radical transformation over the past 25 years. As recently as the 1960s, articles in the major medical [13] and epidemiology [14] journals employed only the most rudimentary statistical methods. Most papers reported only **means** and **standard deviations**, while an author would occasionally use simple **linear regression**. By the end of the 1980s, however, the statistical content of articles published in the leading journals had changed dramatically. Emerson & Colditz [6] found that, of 115 Original Articles published in the *New England Journal of Medicine* in 1989, 33% employed one or more advanced statistical methods. An equally dramatic shift occurred in the design and size of studies [14]. While most studies published in the 1960s were laboratory investigations or small observational studies, often uncontrolled, the majority of today's investigations of human populations are of two major types, controlled epidemiologic studies with explicit and carefully crafted designs (*see* **Analytic Epidemiology**) and randomized **clinical trials**.

In response to this shift in the statistical and epidemiologic character of submissions, many clinical and epidemiologic journals have increased the intensity of their statistical reviews. For more than a decade, *The New England Journal of Medicine* has retained a team of statistical consultants who attend weekly editorial meetings and review every potential publication with statistical or epidemiologic content. *The Journal of the National Cancer Institute* has a total of 20 statistical editors! All articles with statistical content receive statistical review. Other leading medical journals have made similar arrangements.

The move toward increased emphasis on statistical review has been stimulated in part by periodic controversies about the design or analysis of studies that subsequently play an important role in drug approval (*see* **Drug Approval and Regulation**) (for one example, see Groothuis et al. [7], McIntosh [11], or Ellenberg et al. [5]) or national biomedical policy (for one example, see Berkel et al. [1] and Bryant & Brasher [2]). Although no review process can be completely successful in identifying weaknesses in the design, conduct, or analysis of scientific studies, the increasing intensity of statistical review has improved the quality and clarity of presentation of statistical methods employed in articles published in major medical journals. This article discusses the role of the statistical reviewer and reflects on the boundaries and limitations of statistical review.

The nature and complexity of a statistical review depends to a great degree on the design of the study under review. Review of randomized clinical trials is often straightforward because the methodologic paradigm is both well established and achievable. Review of epidemiologic studies can be more challenging for statisticians, because such reviews frequently require a deeper understanding of the biological question and its implications for design and analysis. We discuss some of the issues in each type of review in subsequent paragraphs, then conclude by mentioning some study designs and methods of analysis that are growing in importance and which present new challenges to the reviewer.

Randomized Clinical Trials

Statistical review of randomized clinical trials is often straightforward, at least when the responsibilities of a statistical reviewer are interpreted narrowly, because the standards for design and analysis of clinical trials are so clear. DerSimonian et al. [3] identified 11 aspects of the design and analysis of clinical trials that they considered important to assessing the quality of a study. They were:

1. **eligibility criteria**;
2. admission decision preceding treatment assignment;
3. random allocation to treatment;
4. the method of **randomization**;
5. patients' blindness to treatment (*see* **Blinding or Masking**);
6. blinded assessment of outcome;
7. information about treatment complications;
8. loss to follow-up data;
9. quality of statistical analysis;
10. complete description of statistical methods;
11. **power**.

Criteria for selection of patients bear primarily on the generalizability of the study. Consistent application of those criteria, however, including evidence

that all patients seen during a specified time period were screened for eligibility, insures that the study sample reflects the stated criteria, thus providing a firm basis for generalization (*see* **Validity and Generalizability in Epidemiologic Studies**).

Procedures for randomizing patients are now well standardized (*see* **Randomized Treatment Assignment**). The sequence of treatment assignments should be derived from computer-based randomization, and no aspect of the assignment procedure should enable investigators to anticipate assignments before they are revealed. These assignments should be revealed only after the decision to enroll a patient has been made. Problems associated with randomization usually result from failures of implementation rather than failures of design. There have been a number of well-publicized examples of breaches of procedure by study personnel who had access to envelopes or other insufficiently secured information about treatment assignments.

Randomization is not always ethical or feasible. In studies of, for example, survival after lung or heart transplantation, the high level of risk and ethos surrounding care of the transplant candidate have prevented consideration of randomized trials. In those settings, the reviewer must judge whether unblinding of patients or study personnel could have introduced **bias** into the comparison of treatment groups. This assessment depends in part on the degree to which the endpoints are objective (*see* **Outcome Measures in Clinical Trials**).

In the ideal trial, patients, caregivers, and evaluators are blind to treatment assignment. Knowledge of treatment assignments should be concealed from all three groups to the maximal extent feasible and ethically defensible (*see* **Ethics of Randomized Trials**). Again, in the ideal design, every patient should be followed throughout the trial and measured on every scheduled occasion. This ideal is never met in clinical research. Losses to follow-up are an unavoidable part of clinical trials involving extended follow-up. The statistical reviewer should be aware that loss to follow-up threatens the integrity of a randomized trial, and that no statistical method for analysis of incomplete data can protect against bias if losses to follow-up are informative [10]. Thus, the reviewer must judge whether (i) the investigators have achieved the best possible follow-up rate in the particular study setting, and (ii) whether the extent and nature of **missing data** have the potential to produce bias in treatment group comparisons comparable in magnitude to the treatment effects under investigation.

The statistical methods used by authors of reports of clinical trials are usually sound in their fundamentals, both because the design suggests the appropriate comparisons and because trialists tend to be experienced investigators. It is common, however, for statistical reviewers to find problems of emphasis or selection. Investigators may, for example, emphasize results in a subgroup (*see* **Treatment-covariate Interaction**) or for an endpoint that was not identified a priori as the primary hypothesis without appropriate statistical adjustment in the analysis. Simon [12] discuss statistical methods for analysis of subsets in clinical trials.

Given the importance of the distinction between a priori and a posteriori hypotheses in the interpretation of randomized trials, as well as the frequent difficulty of evaluating statistical methods from the brief synopses provided in typical submissions to medical journals, it would be helpful for statistical reviewers to receive copies of the statistical methods sections of study protocols, yet this is rarely done on a routine basis. When the study employed a less commonly used design, such as repeated measures, the reviewer must be more vigilant about the validity of the analytic approach.

Serious deficiencies of design or analysis of clinical trials are not always apparent from the manuscript submitted for review. Review of primary data has sometimes uncovered problems or discrepancies in study results that would not be detectable from a careful review of the scientific article. The risk of hidden deficiencies is unavoidable, because it will not be possible for reviewers to examine the details of every study submitted for publication. The statistical reviewer can address this concern to some degree by asking probing questions of the authors that require more detailed discussion of study procedures.

One challenging aspect of statistical review for this writer has been the judgment about whether a small, preliminary clinical trial is worthy of publication. This judgment can depend not only on the size of a trial and the strength of its evidence, but also on the degree to which the therapeutic strategy is innovative, and upon the urgency of the public health problem it addresses. When one has difficulty

making that judgment in an area in which one is not fully acquainted with the state of the field, it can be helpful to acknowledge the issue and provide clear information to the editor about the strength of the study.

Epidemiologic Studies

Epidemiologic studies are intrinsically more challenging than randomized trials for most statistical reviewers, because interpretation of an epidemiologic study often requires a high level of understanding of the scientific issues. One helpful concept is the assertion that epidemiologic studies should be conducted according to scientific principles governing the design and conduct of randomized trials in every respect save the use of randomization for assignment of treatments. Thus, epidemiologic studies should be based on a written scientific protocol, hypotheses should be stated a priori, subjects should be completely enumerated, losses to follow-up should be minimized, and so on. Several individuals and groups have prepared guidelines for the design, conduct, or evaluation of epidemiologic studies.

Even when this principle is honored, however, **observational studies** have pitfalls. Both patient selection and control of potential **confounders** are important issues in most epidemiologic studies. Statistical reviewers for medical journals must be conversant with the criteria for evaluation of such designs. Most contemporary epidemiologic studies employ either the **cohort** or **case–control** design, and thus include concurrent **controls**. The evaluation of such studies, especially case–control studies, requires special expertise and understanding of the issues involved in selection of controls, measurement of confounders, and the implications of **matching** (*see* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**).

The parallels in data analysis between epidemiologic studies and randomized trials are much closer than those in design. Statistical reviewers should be knowledgeable about statistical methods for the analysis of cohort and case–control studies, including the effects of matching (*see* **Matched Analysis**) and **stratification** on the analysis. In practice, epidemiologic studies are less likely than randomized trials to follow a written protocol that specifies the primary and secondary study hypotheses. Thus, the reviewer should attend to whether the endpoints and hypotheses emphasized in the report are the natural primary questions that motivated the study.

New Challenges to Statistical Reviewers

The use of multivariate methods in biomedical research has developed sporadically (*see* **Multivariate Analysis, Overview**). **Logistic regression** analysis emerged as an important analytic tool in the 1960s, multivariate survival analysis was popularized in the 1970s, and methods for the analysis of longitudinal and clustered data (*see* **Cluster Analysis, Variables**) became increasingly important in the 1980s. In part, this reflected the maturation of research on chronic diseases. While early studies focused on survival and severe morbidity, more recent studies have focused on changes in **quality of life** over time and in the development of risk factors for chronic disease. Both of these questions require study designs that employ repeated measures. Unfortunately, there is at present no easy introduction to the principles of **longitudinal data analysis**. The landmark book by Diggle et al. [4] is comprehensive but challenging. A recent SAS publication on mixed effects models [9] provides some helpful illustrations of the use of mixed linear and **nonlinear** models for analysis of repeated measures.

Perhaps the 1990s will be remembered for the growing importance of investigations based on genetic and molecular data. Some studies can be classified as examples of either **molecular epidemiology** or **population genetics**. In a molecular epidemiology study, a classifiable genetic trait of an individual is used as either a risk factor or an **effect-modifier** for a disease outcome. Although such studies require sophisticated molecular methods for patient characterization, the principles of study design and data analysis are identical to those encountered in epidemiologic studies using other types of risk factors. Thus, they present no fundamentally new challenges to the statistical reviewer.

The same cannot be said of studies involving **linkage analysis**, **segregation analysis**, and investigation of **polygenic inheritance** of disease. Such studies require new study designs [8] and new methods for quantifying relationships at the molecular level. Statisticians wishing to be broadly knowledgeable about statistical issues in biomedical research should master these ideas.

References

[1] Berkel, H., Birdsell, D.C. & Jenkins, H. (1992). Breast augmentation: a risk factor for breast cancer?, *New England Journal of Medicine* **326**, 1649.

[2] Bryant, H. & Brasher, P. (1995). Breast implants and breast cancer: reanalysis of a linkage study, *New England Journal of Medicine* **332**, 1535–1539.

[3] DerSimonian, R., Charette, J., McPeek, B., et al. (1992). Guidelines for statistical reporting in articles of medical journals, in *Medical Uses of Statistics*, 2nd Ed., J.C. Bailar & F. Mosteller, eds. NEJM Books, Waltham, pp. 333–347.

[4] Diggle, P.J., Liang, K.-Y. & Zeger, S. (1995). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[5] Ellenberg, S.S., Epstein, J.S., Fratantoni, J.C., et al. (1994). Trial of RSV immune globulin in infants and young children: the FDA's view, *New England Journal of Medicine* **331**, 203.

[6] Emerson, J.D. & Colditz, G.A. (1992). Use of statistical analysis in *The New England Journal of Medicine*, in *Medical Uses of Statistics*, 2nd Ed. J.C. Bailar & F. Mosteller, eds. NEJM Books, Waltham, pp. 45–60.

[7] Groothuis, J.R., Simoes, E.A.F., Levin, M.J., et al. (1993). Prophylactic administration of respiratory syncytial virus immune globulin to high-risk infants and young children, *New England Journal of Medicine* **329**, 1524–1529.

[8] Lander, E. & Schork, N. (1994). Genetic dissection in complex traits, *Science* **265**, 2037–2048.

[9] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Linear Models*. SAS Institute, Cary.

[10] Little, R. & Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[11] McIntosh, K. (1993). Respiratory syncytial virus: successful immunoprophylaxis at last, *New England Journal of Medicine* **329**, 1572.

[12] Simon, R. (1988). in *Recent Results in Cancer Research*, M. Baum, R. Kay & H. Scheurlein, eds. Springer-Verlag, Heidelberg.

[13] Ware, J.H. (1982). Comparison of medical and surgical management of coronary artery disease: methodologic issues, *Circulation* **65**, Supplement II, 32–36.

[14] Ware, J.H. (1990). The role of epidemiology in risk assessment: a statistician's perspective, *Chance* **3**, 41–47.

JAMES H. WARE

Statisticians in the Pharmaceutical Industry (PSI)

PSI (Statisticians in the Pharmaceutical Industry Limited) is a nonprofit organization, which converted to a limited company early in 2003. Though primarily UK-based, currently around 20% of PSI members reside outside of the United Kingdom, and further members from all over the world are welcomed. PSI is open to all people interested in the application of statistics in the **pharmaceutical industry**. Its major objectives are:

1. To promote professional standards in the application of statistics in matters pertinent to the pharmaceutical industry
2. To provide a forum for regular discussion on statistics and matters relating to the practice of statistics in the pharmaceutical industry
3. To influence regulatory direction and scientific methodologies that are applied to drug development
4. To contribute to the development of statistics as a profession.

PSI was founded in 1977 with around 50 members. The organization has grown rapidly, with the development of the role of statistics in the highly regulated pharmaceutical industry (*see* **Drug Approval and Regulation**) and now has over 1000 members. The majority of members are statisticians and statistical programmers working within sponsor pharmaceutical and biotechnology companies or Contract Research Organizations (*see* **Proprietary Biostatistical Firms**), in all areas of the drug development process, including research, **preclinical** and clinical development (*see* **Clinical Trials, Overview**), production, **quality control**, marketing, and market research. Further members include independent statisticians and statistical programmers, academic statisticians, teachers, students and nonstatisticians from within the industry.

PSI arranges regular scientific meetings, with a view to creating an environment where members have the opportunity to exchange scientific information. A three-day conference and a number of one-day scientific meetings are held each year. In addition, special interest groups (SIGs) communicate regularly on topics of particular interest (e.g. the Statistical Computing SIG often runs half-day meetings on topics of current interest). From time to time, working parties are established to investigate particular statistical issues, and members are actively encouraged to participate. One such group is the Clinical Research Computer Systems Validation Working Party (jointly sponsored by PSI and the Association of Clinical Data Management, ACDM), which has recently published the second edition of the "Computerised Systems Validation in Clinical Research" Guideline. The guideline can be purchased via the ACDM website (see `http://www.acdm.org.uk`).

PSI runs a program of training courses, with the aim of bringing members up-to-date in a particular area without having to commit to a lengthy period away from the office or to extensive follow-up reading. An "Introduction to Industry" course is run each year for new entrants to the pharmaceutical industry, as well as three or four short courses on statistical topics, each usually lasting two to three days (*see* **Teaching Medical Statistics to Statisticians**).

Contact is maintained with academic statistical groups and professional organizations (such as the **Royal Statistical Society**, (RSS)), as well as other pharmaceutical bodies (such as the Pharmaceutical Research and Manufacturers of America, PhRMA), thereby promoting the public image of PSI, both within and outside the pharmaceutical industry. Together with other European statistical organizations, PSI is a member of the **European Federation of Statisticians in the Pharmaceutical Industry (EFSPI)**, in which capacity PSI monitors and responds to appropriate regulatory statistical issues within the pharmaceutical industry. PSI organizes occasional workshops to discuss important regulatory issues, and invites an expert statistical group to provide comments on draft regulatory documents. Feedback from members is coordinated in conjunction with the Association of the British Pharmaceutical Industry (ABPI) and EFSPI to ensure appropriate input to regulatory authorities and industry organizations.

In 2002, PSI launched the statistical journal, *Pharmaceutical Statistics*, in conjunction with the publishers John Wiley and Sons (`http://www3.interscience.wiley.com/cgi-bin/jhome/93012805`). This international journal, sponsored by PSI, is issued electronically on a quarterly basis to

all PSI members as part of their annual PSI subscription. The journal aims to disseminate information and practical examples of the use of statistics in all stages of drug development, from discovery to production. Also, in 2002, PSI redesigned its website (`http://www.psiweb.org`) to include a resource center, membership discussion forums, and job/service advertisements in addition to electronic journal access. PSI continues to publish a quarterly newsletter, *SPIN*, to keep its members informed about PSI activities and other relevant events.

PSI promotes careers within the industry by publishing careers material and arranging a program of talks for universities. More recently, PSI has collaborated with the RSS and other organizations in attempts to nurture statistical interest in schools. A Grants Fund is available to subsidize student attendance at relevant PSI scientific meetings.

PSI employs a professional executive secretary in order to ensure efficient administration. For further information please contact: PSI Executive Office, Resources for Business, Association House, South Park Road, Macclesfield, Cheshire SK11 6SH, UK. Tel: +44 (0) 1625 267882; Fax: +44 (0) 1625 267879; e-mail: admin@psiweb.org.

DAVID MORGAN & KERRY GORDON

Statistics in Medicine

Statistics in Medicine is among the leading journals of medical statistics and epidemiology. It publishes papers on the practical applications of statistics and other quantitative methods to medicine and its allied sciences. It embraces all aspects of the collection, analysis, presentation, and interpretation of medical data, including areas such as **clinical trials**, diagnostic studies, quality control, laboratory experiments, epidemiology, and **health services research**.

The journal emphasizes the relevance of statistical techniques and aims to communicate statistical and quantitative ideas in a medical context. Examples of applications of statistics to specific projects, articles explaining new statistical methods, and reviews of general topics are published.

The ultimate goal of *Statistics in Medicine* is to enhance communication between statisticians, clinicians, and medical researchers with the common purpose of advancing knowledge and understanding of quantitative aspects of medicine. It is intended that both the readers and authors of the journal include statisticians, clinicians, epidemiologists, health researchers, mathematicians, and computer scientists interested in medicine.

Statistics in Medicine was launched as a quarterly journal in 1982, under the editorship of T. Colton (Boston), and L. Freedman and A. Johnson (Cambridge); the first volume included 40 papers in 380 pages. It progressed to six issues in 1986 with 64 papers in 678 pages; to eight in 1987, and to 12 in 1988, when L. Freedman emigrated to Bethesda, and the original editors were joined by D. Machin (Cambridge). Further expansion to 16 issues occurred in 1992, and, ultimately, to 24 in 1993; in 2003, the journal published 83 papers in 3914 pages. L. Freedman retired as an editor in 1993, an occasion marked by an account of the foundation and early development of the journal (**13**(1), 1–2 (1994)). He was succeeded by R. D'Agostino (Boston) Deputy editors R. Glynn (Boston) and J. Greenhouse (Pittsburgh) were appointed in the US at the beginning of 1995, and C. Palmer and S. Stenning (Cambridge) in the UK in September 1996. Tony Johnson stepped down in June 1999. The journal has an editorial board of over 70 eminent statisticians and epidemiologists in many centers of excellence throughout the world.

In addition to papers that are peer reviewed by the editorial board and other advisory referees, the journal welcomes letters commenting on published papers, as well as those embracing more general issues. It also includes reviews of books and reports of general interest to medical statisticians under the guidance of the book reviews editor [P. Macaskill, Sydney) The position of deputy editor was abolished at the end of 2001 when four editors-in-chief (D'Agostino, Greenhouse, Machin and Campbell) were appointed. Machin retired at the end of 2002, to be succeeded by J. Matthews (New castle). In 1995, the journal launched two further topical features: first, a series of expository articles on the application of specific biostatistical techniques, including those introduced comparatively recently, under the title of Biostatistical Tutorials, with R. D'Agostino (Boston) as editor; and, secondly, a series of occasional articles under the title Statistics in the Medical Literature formerly the editorship of D. Altman (Oxford) and now S. Evans (London) drawing attention to specific publications dealing with some aspect of biostatistics or epidemiology, and published in the biomedical literature.

Apart from the regular papers submitted to the journal, a major feature is the publication of special issues, many under the directorship of invited guest editors. Four of these have honored eminent colleagues whose work has been of fundamental importance in the foundation and development of medical statistics. **Sir Austin Bradford Hill** (**1**(4), October 1982), P. Armitage (**9**(6), June 1990), D. Newell (**14**(2), January 1995) and **Sam Greenhouse** (22(21) November 2003). Several more have been devoted to workshops on statistical methodology organized by the **National Institutes of Health**, **Centers for Disease Control**, Johns Hopkins University, United Kingdom Coordinating Committee on Cancer Research, and others. It also publishes the proceeding of the annual conference of the International Society for Clinical Biostatistics.

In 1991, a decade of publication was celebrated by a special anniversary issue which included overviews of advances over the previous ten years in clinical trials, epidemiology, diagnosis and **quality of life**, and a competition for younger medical statisticians in the US (**10**(12), December 1991).

The journal is published by John Wiley in Chichester, UK, and operates from one editorial office at Boston University School of Public Health, 80

East Concord Street, Boston, Massachusetts 02118, USA (editorial assistant, S. Thompson), since January 2004 the journal runs under an electronic system `www.sim-wiley.manuscriptcentral.com` from which submission details are available. Most papers are peer reviewed, usually by two expert referees.

THEODORE COLTON, ANTHONY L. JOHNSON &
DAVID MACHIN

Statistics, Overview

This article is an inevitably personal view of the general position of statistical science, as seen in the mid-1990s, a period of rapid development. Even within medical statistics the range of applications is great, and in statistics more broadly the variety is even more extreme, making sweeping statements about the relative importance of, for example, different techniques and different approaches difficult to substantiate in any generality.

The term *statistical science* is sometimes used for statistical theory and its applications to the natural and social sciences and to science-based technology and this roughly corresponds to the scope of the present article.

The interlinked pillars of statistics as a field of study are

1. The mathematics of probability.
2. The general principles for the design, analysis and interpretation of investigations. Formal principles of statistical **inference** are a part, but only a part, of this.

It may be tempting to add a third pillar, which gave the subject its name, namely the collection and study of economic and social statistics for government, so-called official statistics, and the closely related issues concerning large enterprises. While this aspect of the subject has indeed developed rather separately over many years, at some level the general principles seem unlikely to be different from those involved with applications to science and science-based technology. For medical statisticians official statistics connected with health have always been important (*see* **Vital Statistics, Overview**), and if the term biostatistics is interpreted more widely to include, for example, agricultural statistics, then there are other links with official statistics.

Probability

The first part of the article thus concerns the mathematics of probability; that is, issues concerned with the meaning and philosophy of probability are excluded.

Historically and, often but not always, in introductory teaching, probability starts from combinatorial problems, i.e. from the counting of the proportion of "favorable" cases in the enumeration of a set of possibilities assumed equally likely a priori. Modern **probability theory** has blossomed from that into a rich chapter of modern mathematics with links to other areas of pure mathematics. Some of the modern developments are, at the moment, fairly remote from statistical applications although study at a relatively advanced level is required:

1. To derive and underpin various statistical methods. Instances where mathematically elaborate methods have been deployed include the use of martingale theory in connection with survival data [1] (*see* **Counting Process Methods in Survival Analysis**) and more generally the rigorous derivation of limiting results in semiparametric inference (*see* **Semiparametric Regression**).
2. To derive special stochastic models for phenomena, usually systems developing in time (*see* **Stochastic Processes**; **Epidemic Models, Stochastic**).

We deal here with the second of these.

Stochastic models supply an important route for developing mathematical models of systems involving a nontrivial random element, both for the insight that the models themselves can supply and as a basis for introducing a substantive base into the interpretation of empirical data.

Among the fields in which such work has a solid history combined with much current activity are

1. Epidemic theory, again with a long history, with developments up to 1970 summarized by Bailey [4], and with major recent developments [26] stemming largely but not entirely from the study of **AIDS** [10, 2]. For an application to **BSE** (bovine spongiform encephalopathy), see [3].
2. The study of congestion and more broadly in **operational research**, dating back to the work of A.K. Erlang at the Copenhagen Telephone Company and congestion theory being stimulated nowadays by the study of complex networks.
3. Finance theory.
4. Genetics, for example, in particular, phylogenetics and **genetic epidemiology**.
5. A number of other areas of **mathematical biology**, such as competition processes, including predator–prey models.
6. Geomorphology and hydrology.

7. Statistical physics, with a very long history of the use of probabilistic ideas, often in a way that seems idiosyncratic from the viewpoint of other developments in probability theory. More recently, however, there has been rather more contact between statistical physics and the mainstream of work in stochastic processes. Two fundamental problems in physics, i.e. the foundations of quantum theory and the nature of the process generating turbulence, both seem likely to have some stochastic element to them.

In the present context, it is helpful to draw a rough distinction between four types of probability model:

1. Purely empirical models.
2. "Toy" models.
3. Intermediate models.
4. Quasi-realistic models.

Models for rainfall provide a convenient illustration. A purely empirical model for, say, daily rainfall at a single site [41] might specify the binary sequence of (no rain, rain) as an m-dependent **Markov chain** with seasonally varying transition matrix and the amount of rain, conditionally on its being nonzero, as having a **lognormal distribution** or **gamma distribution** with seasonally varying parameters. This might provide a valuable and accurate representation of the frequency properties of the rainfall process, but there would be no direct link with the underlying physical process or corresponding interpretation of the parameters.

A "toy" model is one in which a highly idealized representation is used to explore the particular circumstances under which a phenomenon of interest could be generated from simple starting assumptions. Examples are the use of idealized cascade models (clusters of clusters of ...) to show some conditions under which scaling, i.e. self-similarity, of the rainfall spatial field can occur [21], and simple models showing conditions for the explosion or extinction of epidemics, or the extinction of species by competition. Elaborate fitting to empirical data is often inappropriate.

An intermediate model is one in which some aspects of a complex physical or other process are represented with the objective of obtaining a form that can be fitted to empirical data in such a way that the resulting parameter estimates do have a link with the underlying generating process. An example with rainfall is the use of models in which there is a **Poisson process** of storms. Each storm consists of a random number of rain cells, displaced from the storm origin, each cell being of random duration and depth, the total rainfall depth consisting of the sum of all contributing cells. The notion of a rain cell has a physical interpretation and the resulting process can produce a reasonable fit to the rather complicated time series of say five-minute rainfalls, in which within periods of rain there is a large highly non-Gaussian distribution of intensity interspersed with short periods of zero rainfall. The models [36] also have the major advantage that they can be generalized to spatial–temporal form [14].

A corresponding quasi-realistic model would be a global circulation model in which the nonlinear partial differential equations representing the physical processes involved are solved numerically [30]. Similar models involving complex processes are used in studying global warming [24] and many other types of system, physical, biological, or economic. The models are frequently, although not inevitably, deterministic rather than having an explicit stochastic element.

A few general issues in this broad area of work are as follows.

1. When is the introduction of a stochastic element into a model likely to be crucial, i.e. when are deterministic models broadly adequate?
2. The relationships between a deterministic model and a roughly corresponding stochastic model [44, 25] are important in settling the kind of formulation suitable. For models consisting of linearly superimposable components the deterministic model gives the corresponding stochastic mean, but even then the mean may, for small systems, give a poor idea of the behavior of sample paths. For nonlinear systems, such as epidemic models, the deterministic model gives an approximation to the stochastic mean valid in large systems (*see* **Epidemic Models, Deterministic**).
3. "Toy" models can be highly enlightening. (The term "toy" should not be taken pejoratively!) How can they best be used in combination with more realistic and elaborate models? One route is in the interpretation of results from a complex **simulation** model by examining the ratio of relevant response variables as simulated

to those predicted by a "toy" model. This ratio, or correction factor, may be expected to vary much more slowly with relevant parameters than the response itself [12].

4. Issues arise over the fitting of intermediate models in that **likelihood** functions may be hard to compute and, not always relevant and **estimating functions** constructed by equating observed and fitted features may have a strong element of arbitrariness.

The study of deterministic models, of stochastic models and of the analysis of empirical data has often been undertaken by separate groups of investigators. While the reasons for this may be clear the separation is to be regretted.

Design, Analysis, and Interpretation

Preliminaries

The remainder of the article is concerned with the design, analysis, and interpretation of investigations. In a broad sense the same principles apply to experiments, **observational studies**, and the secondary analysis of data collected for some not very specific research purpose, for example a large family expenditure survey or a cancer registry. We assume that the starting point is a question or issue or research hypothesis of interest, although sometimes preliminary analysis may be needed to clarify the issue involved. In all types of study the key initial questions of design are

1. What individuals should be studied?
2. What properties should be measured and what do the measurements really mean?
3. What **contrasts**, including **interactions**, should be examined?

The broad requirements are

1. The avoidance of **systematic error**.
2. The control of **random error**.
3. The exploitation of the factorial structure of contrasts (*see* **Factorial Experiments**).
4. The formulation of special objectives.

So far as the last point is concerned we can for the most part regard the purpose to be the **estimation** of relevant parameters and corresponding **standard errors**, but a specific decision or prediction objective may alter the whole focus of the study. For example, the design of a plant-breeding programme for varietal selection would involve quite different considerations from that for the comparison of a small number of specific varieties. In the former, emphasis is to be placed on the properties of the small number of varieties ultimately chosen for intensive investigation rather than on specific internal comparisons among a small number of varieties.

The differing relative emphasis on the above requirements, especially the first three, explains why the literature on design appears so different in the **clinical trial** context from that in, say, the chemical engineering field. In the former, but not the latter, avoidance of systematic error is of key importance. In the design and analysis of observational studies too, attempts to eliminate systematic error are often of central importance.

It is disappointing that awareness of some of the basic principles of design has not percolated more widely into the laboratory sciences. Even in physics, where investigations of great subtlety are common, the widely held view that refinement of laboratory technique is always to be preferred to statistical technique as a base for error control is probably much less valid than it used to be.

The reason for neglect of the statistical aspects of the design of investigations may partly be that the theory of statistical design is quite widely identified with the use of complex designs. These have their place, but they are often not appropriate, key issues more commonly being simple techniques for methodical **bias** elimination and error control.

Measurement Issues

The techniques of analysis connected with **variance components** were developed in the 1930s and 1940s tied to balanced data and continuous roughly **normally distributed** data. The restriction to balanced data was removed in pioneering work by C.R. Henderson in connection with animal breeding and synthesized most satisfactorily in the residual (or **restricted) maximum likelihood** (REML) method of Patterson and Thompson [32]. Systematic extensions to **Poisson**, ordinal, and **binomial** data are the focus of current work [29].

Such techniques provide the basis for the design and analysis of interpersonal and interlaboratory

studies of measuring techniques, common in some fields of study, especially connected with the physical sciences, but regrettably less common in a medical context.

They also are in principle appropriate in instrument development. The psychometric notions of reproducibility, and of face, criterion, and concept validity and of comparison with a **gold standard** are not normally presented in terms of components of variance but would probably be better done so (*see* **Psychometrics, Overview**). Systematic techniques for the more detailed analysis of instruments consisting of many relatively similar items are needed. This is especially relevant in connection with **quality of life,** i.e. health status [13].

Classical work on error assessment in experiments and surveys emphasized the often **multilevel** or **time series** structure of error. Treating error variables as independent and identically distributed typically leads to an underestimate of the standard error of contrasts of primary interest. A more empirical adjustment for such **overdispersion** may be via the direct estimation of correction factors to apply to standard errors [6, 31, 19].

Methods of Analysis

Some methods of analysis do not depend on an explicit probability model and recent developments, especially in computer graphics, are of interest both for exploratory work and also in presenting the conclusions of more elaborate analyses; indeed, as in very elaborate analyses the connection between the data and the conclusions may get rather remote, the need for insightful methods of presentation increases.

Nevertheless the rest of this section concentrates on methods that depend at least in part on an explicit probabilistic base.

Some requirements for a probabilistic model are as follows, although not all are relevant in every application.

1. The model should establish a link with underlying substantive knowledge or theory.
2. The model should allow comparisons with previous related studies of the topic.
3. The model should be consistent with or suggest a possible process that might have generated the data.
4. Parameters defining primary features of the system should have individually clear substantive interpretations.
5. The error structure should be represented sufficiently realistically that meaningful measures of precision are obtained for the primary comparisons.
6. The fit to data should be adequate.

We comment here on only some of these points.

The first three items are related to the general issue of preferring what we previously called intermediate models to purely empirical models. Such a preference was indicated in much of the applied work of J. **Neyman**; it can have the disadvantage of making the analysis of fairly simple sets of data overcomplicated and there is a difficult broad strategical issue to be faced in each application concerning the weight to be placed on substantive vs. purely empirical models.

In fields with a quantitative theoretical base this will typically provide a key to a suitable model. In the **social sciences** and in some areas of biology there is often the problem of incorporating background knowledge that is essentially qualitative. Here the ideas of chain graphs, expressing directional relationships between variables and of substantive research hypotheses [43] expressing some conditional independencies and some strong dependencies provide a route to insertion of such knowledge. The graph theory ideas are a development from Sewell Wright's **path analysis**. For accounts with a strong statistical focus, see [18, 15] and [16], and for a more theoretical account see [28]. Spiegelhalter et al. [39] discuss applications to probabilistic expert systems.

The need to connect to previous work is in superficial conflict with the Fisherian notion that investigations provide their own estimate of error. However, the need to relate the primary conclusions in different studies is clear; this includes the examination of consistency of the conclusions. There is a broad connection with overviews, or so-called **meta-analysis**, of much current interest in medical research. The statistical principles were set out by Yates and Cochran [45] and developed further by Cochran [11]. The most challenging issues there, however, concern the choice of material for synthesis.

Models suggesting or consistent with a data-generating process provide some possible link with a causal interpretation; see the further discussion

below. While the term *causal* has a number of interpretations, a very cautious usage tends to be favored in statistical discussions, in particular that strong evidence for causality can only come from the synthesis of different kinds of data. Analysis that points towards a potentially causal interpretation can, however, be valuable; this is one reason for the importance of chain graph representations (*see* **Causation**).

The preference for primary parameters that have individual interpretations links to the previous point and to the Fisherian notion that in **analysis of variance** "treatment" sums of squares should if possible be split into single **degrees of freedom**.

Detailed Techniques

Specific developments in methods of analysis are described throughout this *Encyclopedia*. Among some predominant themes in current research are the following:

1. Nonlinearity is a widely occurring theme, as in so many areas of modern mathematical science. **Nonlinear time series** models provide one important example;
2. **Nonparametric regression** and **density estimation** have been intensively studied in their theoretical aspects. The main value in applications is likely to be in the preliminary stages of analysis.
3. Semiparametric methods in which the primary aspects of the model are represented by parameters, and such issues as distributional form are left nonparametric, raise very interesting theoretical issues. It is, of course, for consideration in each case whether the greater complication and loss of transparency involved as compared with fully parametric formulations is really justified.
4. **Markov chain Monte Carlo** methods provide a powerful general tool for the fitting of relatively complex models.
5. Computer **simulation** methods, **cross-validation**, and the **bootstrap**, provide fairly general methods of assessing precision without elaborate theoretical analysis, although some underlying assumed simple structure is needed (*see* **Computer-intensive Methods**).
6. Methods for addressing data imperfections, such as missing data, including selective **nonresponse**, and for the analysis of nonstandard sampling schemes, are important in many fields.
7. Higher-order asymptotic theory [5] aims to provide a basis for choosing between procedures equivalent to the first order of asymptotic theory and of providing more refined distributional approximations (*see* **Large-sample Theory**).

Interpretation

Design, analysis, and interpretation might suggest a sequence in which narrowly statistical considerations stop at analysis, for example ending with the estimation of relevant parameters, whereas interpretation involves essentially subject-matter considerations not specifically statistical. While, of course, there is some truth to this distinction, the emphasis on model formulation, in particular some of the criteria listed under Methods of Analysis, and the considerations of the section on Probability, are aimed, in line with current thinking, to break down that barrier.

For example, there has been increasing discussion in statistical circles of the conditions under which conclusions can be said to be causal; see, for example, [23, 38] and [16], Section 8.7. The discussion is prompted in part by developments in the computer science and philosophical literature in which a weaker definition of causality tends to be employed [33, 34, 40], much less cautious than the traditional statistical and epidemiological view summarized in **Hill's criteria** [22] (*see* **Causation**).

Closely connected with this are the issues of generalizability and specificity and the importance of absence of interaction. For example, in the light of a well-conducted randomized clinical trial showing evidence of the superiority of treatment A over treatment B, what is the basis for hoping that the conclusions generalize to a new population of patients and what is the basis for thinking that A rather than B will be beneficial for a specific new patient? (*See* **Validity and Generalizability in Epidemiologic Studies**.)

Decision Analysis

Wald [42], in effect continuing in the Neyman–Pearson tradition, proposed that all statistical problems could be formulated as a choice between possible decisions (*see* **Decision Theory**). In Wald's formulation a **utility** (or **loss**) function was assumed known but **prior distributions** enter only as technical devices to produce a complete class of decision functions.

The qualitative ideas that one should consider the objectives of the study, the possible actions that might be taken on the basis of the results and their potential consequences are clearly important. In fields such as sampling inspection and control theory, or where specific point forecasting is involved, a full specification of utilities and prior distribution will yield the preferred solution, but for most purposes summarization of evidence via the estimation of relevant parameters seems a more suitable objective. Thus formulations of clinical trials as decision making procedures, while giving valuable qualitative insights, have not been widely accepted as a realistic model of how trials are used.

Formal Theories of Inference

Much discussion in the more theoretical literature has for many years focused on the formal theory of statistical inference, in particular on the meaning of probability when used to assess the uncertainty in conclusions. The issues are important partly in setting the broad approach to specific problems and partly in detail in developing particular methods of analysis and interpretation.

There are many different approaches but, leaving aside a pure decision-theoretical approach, they can be broadly classified as

1. Pure likelihood [17].
2. Fisherian, putting emphasis on likelihood, **sufficiency**, **conditionality**, **ancillarity**.
3. Neyman–Pearson, reaching many of the same conclusions as 2 but emphasizing operational criteria such as **power**.
4. A Bayesian approach based on standardized impersonal priors [27], now often called reference priors [8, 7] (*see* **Bayesian Methods**).
5. An emphasis on personalistic (or **subjective**) **probability** leading to a wholly Bayesian analysis and rejecting the above approaches as incoherent, or at best as approximations to something else [9].

This is a controversial area on which it seems improbable (in any sense!) that unanimity will be reached. There are, however, some signs of a fairly broad agreement perhaps along the following eclectic lines:

1. Many of the issues addressed in this article, and of direct concern in applied statistical work, do not depend critically on the choice of approach to formal inference.
2. Likelihood, or some adaptation thereof, is of key importance but typically needs calibration into posterior intervals, **confidence intervals** or whatever.
3. Probability, as representing idealized properties of the real world, has to be distinguished from probability as measuring a state of an individual's knowledge.
4. Problems with many similar parameters are usually best formulated in **empirical Bayes** form, powerful numerical methods now being available for their solution.
5. Reference priors in a small number of dimensions usually produce answers with good properties also from the confidence interval viewpoint.
6. It is necessary to have some notion that one's methods of analysis have good properties, or at least are not systematically misleading, when hypothetically they are used repeatedly.
7. To the extent that the previous notion is formalized, some element of conditioning is needed, although overconditioning must be avoided, for example to escape the C.R. Rao paradox of sampling theory [35].
8. The Bayesian formalism provides a valuable representation of the merging of "prior" knowledge with new knowledge from data under analysis, although it does not deal adequately with the possibility of conflict between the two sources.
9. The Bayesian axioms of coherent personalistic probabilities are a valuable guide to opinion formation by individuals, but are not compelling as a basis for public discussion, partly because they put weight on internal consistency rather than on consistency with the real world.
10. While the importance of formal statistical significance is commonly overstated, some such notion, with a **null hypothesis** and **alternatives** (usually not formalized probabilistically), is needed partly to formalize an escape route from an initial unsatisfactory formalization.

Statistics and Public Affairs

The primary emphasis in this article has been on statistics in science and science-based technology. The organization of government statistics varies between countries with somewhat differing emphasis placed on the one hand on the provision of information and advice for direct use by government, somewhat akin to the provision of such information in a business context, and on the other hand on the provision of information to society at large and to social science research workers in particular. For both purposes independence is crucial; in the provision of advice one would hope for government statisticians to be the firm voice of reason in the face of political and economic dogmatism. In the second role, collaboration between government statisticians and social science research workers, including statisticians, is important.

One important area of social policy concerns **risk assessment** and management, i.e. especially as concerns extremely small or maybe even nonexistent risks. This has been discussed extensively by engineers, toxicologists, epidemiologists, psychologists, social anthropologists, sociologists, economists, and political scientists [37], but surprisingly little has appeared in the statistical literature. The role of judgmental probabilities in such situations is central.

The importance of appreciation by the general public of central principles of the interpretation of evidence shows itself in many aspects of material appearing in newspapers and presented on radio or television. For sample surveys, issues like the sample size, the sampling scheme, the response rate, and limits of error, are probably reported rather more often nowadays than in the past. Sensible interpretation of so-called league tables of the performance of schools, hospitals, and the like [20] depends crucially on a critical attitude to empirical data. The reporting, sometimes rather sensationally, of the results of often badly designed small medical studies is of particular concern.

Conclusion

The years 1925–1960 can be regarded as a golden era of statistical thought. For example, in terms of issues of formal inference, the period embraces most of the work of R.A. **Fisher**, of **Neyman** and E.S. **Pearson**, and of **Wald**, the objectivist Bayesian contributions of Jeffreys and the personalistic approach of F.P. Ramsey, **de Finetti**, and **Savage**. Aspects of the design of experiments and sample surveys were developed to a high pitch of elaboration; many of the key ideas of **time series** analysis and **multivariate analysis** were formulated. Statistical quality control and randomized clinical trials were firmly established.

While further important developments took place between 1960 and 1985, these years may best be seen as primarily a period of consolidation. At the beginning of that time most statisticians had access to an electronic computer but obtaining useful results could be a lengthy chore. By the end of the period all the "standard" methods, and more, were fairly readily available to a wide spectrum of users.

Encouraging features of the last 10 years or so are that while a massive educational job remains, the appreciation of statistical ideas is more widely spread among research workers in many disciplines, as shown by the relative sophistication of statistical ideas in subject-matter journals, and by an increase in the amount of substantial collaborative work involving statisticians, in contrast to short-term "consulting" on very specific and often minor details.

Viewed over a rather longer period, there has been a massive growth in the subject, as indicated by the amount of work published per year, by the introduction of new journals, by the number of people employed and by the career prospects for new graduates.

There is currently no shortage of interesting new ideas and challenging problems, many stemming from the relatively large sets of data now so common. For individual research workers freedom to follow one's own judgment of topics likely to be important and to which one is equipped to contribute is needed and is under threat from the short-term policies of many of the sources of financial support. Nevertheless, if statisticians as a group become increasingly involved in important issues in science, technology, and public affairs, if imaginative new ideas can be encouraged, and if fragmentation of the subject can be avoided, then the prospects for an important new period of major development are strong.

Acknowledgment

Support from the Leverhulme Foundation is gratefully acknowledged.

References

[1] Andersen, P.K., Borgan, Ø, Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[2] Anderson, R.M., Cox, D.R. & Hillier, H., eds (1989). Epidemiological and statistical aspects of the AIDS epidemic, *Philosophical Transactions of the Royal Society of London, Series B* **325**, 39–187.

[3] Anderson, R.M., Donnelly, C.A., Ferguson, N.M., Woolhouse, M.E.J., Watt, C.J., Udy, H.J., MaWhinney, S., Dunstan, S.P., Southwood, T.R.E., Wilesmith, J.W., Ryan, J.B.M., Hoinville, L.J., Hillerton, J.E., Austin, A.R. & Wells, G.A.H. (1996). Transmission dynamics and epidemiology of BSE in British cattle, *Nature* **382**, 779–786.

[4] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Disease*. 2nd Ed. Griffin, London.

[5] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[6] Bartlett, M.S. (1937). Some examples of statistical methods of research in agriculture and applied biology (with discussion), *Journal of the Royal Statistical Society* **2**, Supplement, 248–252.

[7] Berger, J.O. & Bernardo, J.M. (1992). On the development of the reference prior method (with discussion), in *Bayesian Statistics*, Vol. 4, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.M.F. Smith, eds. Oxford University Press, Oxford, pp. 35–60.

[8] Bernardo, J.M. (1979). Reference posterior distribution for Bayesian inference (with discussion), *Journal of the Royal Statistical Society, Series B* **41**, 113–147.

[9] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.

[10] Brookmeyer, R. & Gail, M.H. (1994). *AIDS Epidemiology*. Oxford University Press, New York.

[11] Cochran, W.G. (1954). The combination of estimates from different experiments, *Biometrics* **10**, 101–129.

[12] Cox, D.R. & Davison, A.C. (1994). Some comments on the teaching of stochastic processes to engineers, *International Journal of Continuing Engineering Education* **4**, 24–30.

[13] Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Spiegelhalter, D.J. & Jones, D.R. (1992). Quality-of-life assessment: can we keep it simple? (with discussion), *Journal of the Royal Statistical Society, Series A* **155**, 353–393.

[14] Cox, D.R. & Isham, V. (1994). Stochastic models of precipitation, in *Statistics for the Environment*, Vol. 2, V.D. Barnett & K.F. Turkman, eds. Wiley, Chichester, pp. 3–19.

[15] Cox, D.R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion), *Statistical Science* **8**, 204–283.

[16] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies*. Chapman & Hall, London.

[17] Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, Cambridge.

[18] Edwards, D. (1995). *Introduction to Graphical Modelling*. Oxford University Press, Oxford.

[19] Fitzmaurice, G., Heath, A. & Cox, D.R. (1997). Detecting overdispersion in large-scale surveys: application to a study of education and social class in Britain, *Journal of Applied Statistics* **46**, 415–432.

[20] Goldstein, H. & Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion), *Journal of the Royal Statistical Society, Series A* **159**, 385–443.

[21] Gupta, V.K. & Waymire, E. (1993). A statistical analysis of mesoscale rainfall as a random cascade, *Journal of Applied Meteorology* **32**, 251–267.

[22] Hill, A.Bradford (1965). The environment and disease: association or causation, *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[23] Holland, P.W. (1986). Statistics and causal inference (with discussion), *Journal of the American Statistical Association* **81**, 945–970.

[24] Houghton, J. (1991). The predictability of weather and climate, *Proceedings of the Royal Society of London, Series A* **337**, 521–572.

[25] Isham, V. (1991). Assessing the variability of stochastic epidemics, *Mathematical Biosciences* **107**, 161–186.

[26] Isham, V. & Medley, G.M., eds (1996). *Models for Infectious Human Diseases*. Cambridge University Press, Cambridge.

[27] Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford.

[28] Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press, Oxford.

[29] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series B* **58**, 619–678.

[30] Mason, J. (1986). Numerical weather prediction, *Proceedings of the Royal Society of London, Series A* **407**, 51–60.

[31] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. 2nd Ed. Chapman & Hall, London.

[32] Patterson, H.D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**, 545–554.

[33] Pearl, J. (1988). *Probabilistic Reasoning in Expert Systems*. Morgan Kaufman, San Mateo.

[34] Pearl, J. (1995). Causal diagrams for empirical research (with discussion), *Biometrika* **82**, 669–710.

[35] Rao, C.R. (1971). Some aspects of statistical inference in problems of sampling from finite populations (with discussion), in *Foundations of Statistical Inference*, V.P. Godambe & D.A. Sprott, eds. Holt, Rinehart & Winston, Toronto, pp. 177–202.

[36] Rodriguez-Iturbe, I., Cox, D.R. & Isham, V. (1987). Some models for rainfall based on stochastic point processes, *Proceedings of the Royal Society of London, Series A* **410**, 269–288.

[37] Royal Society (1992). *Risk: Analysis, Perception and Management*. Royal Society, London.

[38] Rubin, D.B. (1974). Estimating causal effect of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.

[39] Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L. & Cowell, R.G. (1993). Bayesian analysis in expert systems (with discussion), *Statistical Science* **8**, 219–283.

[40] Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction and Search*. Springer-Verlag, New York.

[41] Stern, R. & Coe, R. (1984). A model fitting analysis of rainfall data (with discussion), *Journal of the Royal Statistical Society, Series A* **147**, 1–34.

[42] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

[43] Wermuth, N. & Lauritzen, S.L. (1989). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion), *Journal of the Royal Statistical Society, Series B* **52**, 21–72.

[44] Whittle, P. (1957). On the use of the normal approximation in the treatment of stochastic processes, *Journal of the Royal Statistical Society, Series B* **19**, 268–281.

[45] Yates, F. & Cochran, W.G. (1938). The analysis of groups of experiments, *Journal of Agricultural Science* **28**, 556–580.

(*See also* **Biostatistics, Overview**; **Experimental Design**)

D.R. COX

StatXact

StatXact is a specialized software package for the exact analysis of small-sample **categorical** and **nonparametric** data with special emphasis on data in the form of **contingency** tables. The term "small-sample" applies equally to datasets with only a few observations, to large but unbalanced datasets, or to contingency tables with zeros and small cell-counts in some of the cells but large cell-counts in other cells (*see* **Structural and Sampling Zeros**). In these settings, StatXact produces exact ***P* values** and exact **confidence intervals** instead of relying on possibly unreliable **large-sample theory** for its inferences. The inference is based on generating permutation distributions of the appropriate test statistics in a conditional reference set (*see* **Randomization Tests**). For a discussion of the theory underlying exact inference, references to numerical **algorithms** that perform the computations, and several examples involving the analysis of biomedical data by StatXact, (*see* **Exact Inference for Categorical Data**). Different reviews of StatXact were published by Lynch, Landis and Localio [2], Wass [5], and Oster [4].

The current version, StatXact-6, offers exact P values for one, two, and K-sample problems, **2 × 2**, $2 \times c$, and $r \times c$ contingency tables, and measures of **association**. The data may be either unstratified or **stratified**. Both independent and blocked samples are accommodated. StatXact-6 computes the exact confidence intervals of odds ratio in case of 2×2 and $2 \times c$ contingency tables and exact confidence interval of median shift in ordered $2 \times c$ contingency tables. StatXact-6 has inference procedures that cater explicitly to **binomial** data, **nominal** categorical data, **ordered categorical data**, ordered correlated categorical data, continuous complete data, and continuous right-**censored** data. StatXact-5 also offers analysis of data that follow **Poisson distributions**. In case the computation of exact P value becomes infeasible due to lack of time and memory, StatXact produces exact P values with at least two decimal digits accuracy using efficient Monte Carlo simulation strategies. Changing the number of Monte Carlo simulations can change the accuracy.

StatXact-6 also computes the exact unconditional confidence interval for a difference or ratio of two independent as well as related binomial proportions and computes exact P values for tests of **equivalence** and noninferiority of two binomial proportions (*see* **Proportions, Inferences, and Comparisons**).

In addition to all the tests mentioned above, StatXact-6 also provides exact **power** and **sample size** calculations for different tests on 2×2 and ordered $2 \times c$ tables.

StatXact-6 runs on Microsoft Windows NT/2000/XP as a stand-alone product. In addition, a special version, StatXact PROCs for SAS users, is available as external SAS procedures for both the Microsoft Windows and Unix operating systems.

LogXact is a companion product to StatXact featuring exact inference for **binary data** in the presence of **covariates**. An underlying **logistic regression** model is assumed. Both exact and asymptotic inferences are provided. The current version of LogXact-5 uses powerful Monte Carlo procedures that enable fast exact inference for much larger data sets. LogXact handles matched case–control data under general M:N matching, by conditional likelihood inference. Asymptotic inference is based on maximizing the unconditional likelihood function for unstratified data and on maximizing the conditional likelihood function for stratified data *see* **Maximum Likelihood**; **Logistic Regression, Conditional**). Exact inference is based on generating the conditional distributions of the **sufficient statistics** for the coefficients of interest, **nuisance parameters** being eliminated by fixing their respective sufficient statistics at the observed values. For a detailed discussion of the theory underlying exact logistic regression, references to numerical algorithms that perform the computations, and several examples involving the analysis of biomedical data by LogXact, refer to [3].

LogXact also provides exact and asymptotic inference for **Poisson regression**. Reviews of LogXact were published by Lemeshow [1] and Oster [4].

LogXact runs on Microsoft Windows NT/2000/XP as a stand-alone product. In addition, a special version, PROC-LogXact for SAS users, is available as external SAS procedures for Microsoft Windows.

References

[1] Lemeshow, S. (1994). LogXact-Turbo: Logistic regression software featuring exact methods, *Epidemiology* **5**(2), 259–260.

[2] Lynch, J.C., Landis, J.R. & Localio, A.R. (1991). StatXact, *The American Statistician* **45**(2), 151–154.

[3] Mehta, C.R. & Patel, N.R. (1995). Exact logistic regression: Theory and applications, *Statistics in Medicine* **14**, 2143–2160.
[4] Wass, J.A. (2000). StatXact 4 for Windows, *Biotech Software & Internet Report* **1**(1), 17–23.
[5] Oster, R.A. (2002). An examination of statistical software packages for categorical data analysis using exact methods, *The American Statistician* **56**(3), 235–246.

(*See also* **Software, Biostatistical**)

CYRUS R. MEHTA, NITIN R. PATEL,
PRALAY SENCHAUDHURI &
CHRISTOPHER D. CORCORAN

Stereology

Stereology is the science of inference about three-dimensional structures based on two-dimensional sections or one-dimensional probes. The name was coined at a meeting of scientists from various disciplines in 1961. This was the first recognition that scattered results in geometrical probability could constitute a coherent field of study.

The subject has important applications in mineralogy, petrology, and metallurgy. In medical and biological research, microscope slides are thin slices through specimens of material. They can give information about the volume of different types of tissue in the specimen, about the area of membranes, and about the length of capillaries. An example of a linear probe is given by a microelectrode penetrating nervous tissue. It allows inference about the size and distribution of nerve cells and their cell membranes.

Inference is also possible about geometrical and topological properties of structures in the specimen. The curvatures of one- and two-dimensional objects may be of interest. Connectivity between structures is often important, and some progress is possible in investigating these features.

Early Results

The first stereological result to be published was the "Delesse Principle", in 1849 [17]. It states that the proportion of *area* occupied by a particular substance or "phase" in a random section of a specimen is an **unbiased** estimate of the proportion of *volume* of that phase in the specimen.

"Buffon's needle" gives the mean number of intersections between two systems of lines in a plane. The result dates from 1733, but was not published until 1777 [2]. Barbier [1], in 1860, gave the three-dimensional extension of the problem. It makes it possible to estimate the length of linear systems from counts of their intersections with a section.

The theory of geometrical probability was first seriously investigated by Crofton. In 1869, his first paper on the subject appeared [7], and his Encyclopedia Britannica article in 1885 [8] summarizes the theory.

In 1925, Wicksell [46] published a paper on the "corpuscle problem", discussing how to estimate the size distribution of spherical particles from their circular intersections with a plane. In 1926 [47], he investigated the much more difficult problem of ellipsoidal objects. These were the first statistical papers in the field of stereology, and they were followed by a number of related studies; see [23].

Sampling

If the aim of a study is to estimate the volume fractions of different phases in an object, the sampling problem is straightforward. Parallel section may be selected by any of the standard sampling procedures; see [4]. Usually, systematic sampling is preferred. Estimates of the volume ratio can then be made from each section, and combined, either by weighting according to the area of the section of the object, or by subsampling proportionally to this area.

Other stereological formulae, however, depend on the orientation of the section. For them to be valid, the section must be random in direction, as well as in position. Isotropic sampling of biological specimens for microscope slides is virtually impossible. In petrology, it may be an option if sufficiently large specimens are available, and isotropic linear sampling may be feasible. Otherwise, no statistical analysis is possible except on slides from different specimens unless it is assumed that the specimen is isotropic – generally highly implausibly for biological material.

For discussions of the sampling problem, see [13, 14, 28, 29], and [30]. The two basic types of sample that give unbiased estimates are *isotropic uniform random* and *area weighted*.

For discussions of the errors that can arise from misuse of statistics in the stereological analysis of biological material, see [34] and [43].

Fundamental Formulae of Stereology

The basic results of stereology can be expressed in terms of ratios. In the original three-dimensional specimen, interest may center on the volume of a particular phase, the area of an interface or membrane, the length of some linear feature, or the number of particles; these can all be expressed *per unit volume*. They are estimated from measurements of area, or length, or from counts on sections or linear probes, expressed per unit area, or per unit length.

A standard notation is used for these formulae. The subscripts V, A, and L refer to the divisor in the specimen, in sections, and in probes, respectively. The letters V, A, L, and N refer to volume, area, length, and count. Thus, A_V means the area of some feature per unit volume in the original specimen.

One other concept is required; $\overline{H}$ is the *mean caliper length* of a set of particles. This is defined as the mean projected length of the particles on a line normal to the section. If the sections are isotropic, or if the particles are spherical or randomly orientated, the value of $\overline{H}$ is taken to be the same for all sections.

Table 1 shows the fundamental formulae. The first line shows the Delesse principle, relating the volume ratio, the area ratio, and the length ratio. These results do not depend on isotropic sampling or structures. The second line shows estimates of area per unit volume; note that the relationship between the two estimates corresponds to the ordinary two-dimensional Buffon's needle. The third line gives the estimate, from a section, of the length per unit volume of some linear feature of the specimen. The fourth line is a possible way of estimating the number of particles or cells; it depends on a knowledge, or a separate estimate, of mean caliper length. This traditional estimate is not necessarily the best; for a discussion of it and other approaches, see [12].

Curvature

Given an interface I between two phases, there are various measures of average curvature of the surface; see [31]. At the center of an element $\mathrm{d}S$, the *principal curvatures* κ_1, κ_2 are defined as the reciprocals of the minimum and maximum radii of curvature (which are orthogonal). The radii and the curvatures are signed; they are usually taken to be positive for elements that are convex toward the bounding phase. The *integral of mean curvature* is defined as

$$K = \int_I \frac{1}{2}(\kappa_1 + \kappa_2)\,\mathrm{d}S \tag{1}$$

If the surface I is imbedded in a volume V, a possible measure of average curvature is $K_V = K/V$. DeHoff [15] (see also [6]) introduced a method of estimating K_V for a phase boundary from sections. This is based on the *area tangent count*. Given an area A, a section of an isotropic uniform specimen with a phase bounded by I, choose a fixed direction. As a line with this direction is moved across the area, it is sometimes tangential to the boundary. The tangent count is divided into two parts. T_+ is the number of times the line is tangential to a convex element of the interface, T_- the corresponding count for concave elements (see Figure 1). Now define

$$T_{\mathrm{Anet}} = \frac{T_+ - T_-}{A}. \tag{2}$$

DeHoff has shown [16] that

$$\hat{K}_V = \pi T_{\mathrm{Anet}} \tag{3}$$

is a consistent measure of the integral mean curvature per unit volume.

Table 1 Fundamental formulae of stereology

Specimen	Section	Probe
V_V	A_A (Delesse)	L_L
A_V	$\frac{4}{\pi}L_A$	$2N_L$ (Buffon)
L_V	$2N_A$ (Barbier)	
N_V	$\frac{1}{\overline{H}}N_A$	

The Spherical Particle Problem

The problem first studied by Wicksell [46] is that of estimating the distribution of the radii of spherical particles from the observed distribution of the radii of the circular sections in a random section. Suppose the sphere radii have density function $f(r)$, with **moments** μ'_k, and the radii of the sections have density $g(x)$, with moments ν'_k. The density of the radii of spheres cut by a random plane has the form

$$\frac{rf(r)}{\int_0^{R_m} rf(r)\,\mathrm{d}r} = \frac{rf(r)}{\mu'_1} \tag{4}$$

where R_m is the maximum value of r. The conditional density of x is given by

$$g(x|r = R) = \frac{x}{R(R^2 - x^2)^{1/2}} \tag{5}$$

and finally

$$g(x) = \frac{x}{\mu'_1}\int_x^{R_m} \frac{f(r)}{(r^2 - x^2)^{1/2}}\,\mathrm{d}r. \tag{6}$$

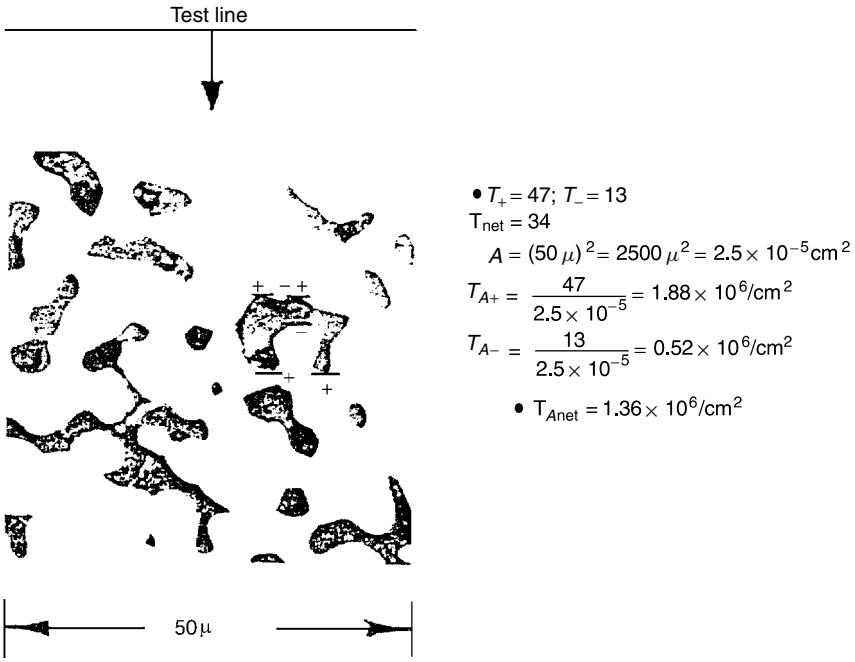

Figure 1 Procedure for measuring the net area tangent count (From DeHoff [15])

This is known as Wicksell's integral equation. The values of x are observed, and the problem is the estimation of $f(r)$. There are two difficulties with the solution. First, small values of r are underrepresented, and in theory, there might be large numbers of particles so small that they almost never appear in sections. This is not a real difficulty in biological applications, as extremely small cells or similar structures are not viable. Secondly, if the spheres are all virtually the same size, the solution may not give a density, as the **variance** of x may be lower than that arising from a homogeneous population of spheres.

Wicksell [46] himself suggested the following two approaches:

1. The moments of r and x are related by the equations

$$\mu'_{k+1} = \frac{1}{h_k}\mu'_1\nu'_k \quad k \geq -1, \tag{7}$$

where

$$h_k = \frac{1}{2}B\left(\frac{k+2}{2}, \frac{1}{2}\right). \tag{8}$$

These equations are valid for $k = 0$ and $k = -1$, with $\nu'_0 = 1$ and $1/\nu'_{-1}$ the harmonic **mean** of the section radii, so that $\mu'_1 = (\pi/2)/\nu'_{-1}$. It is then possible to estimate the moments of the sphere radii, and approximate the density by fitting a member of some family of curves, such as the Pearson family (*see* **Pearson Distributions**). This method is not satisfactory because of the problem of very small values of x. These may often be missed, and their influence on the harmonic mean is large.

2. Wicksell [46] also suggested a **nonparametric** approach, based on an approximate numerical solution of the integral equation. Suppose the observed data – the x values – are grouped in a histogram (*see* **Frequency Distribution**), and suppose R_m, the maximum radius of the spheres, is known or assumed. Then it is easy to calculate the probability that a value of r in a short range gives rise to a value of x in each cell of the histogram. This gives a set of linear equations relating the probabilities of the cells in a histogram of r values to the observed frequencies in the histogram of xs. Solving these equations, described as "unfolding" the observed histogram, gives an estimated histogram for r.

This method works well, provided the histogram intervals are well chosen and the underlying distributions are well behaved. There is no guarantee that the estimated probabilities will be positive; a gap in the distribution of r may give small negative estimates. A number of authors have used this numerical procedure modified in various ways; see [44]. For details of the numerical procedure, see [10].

Particles of Other Shapes

Wicksell [47] also studied the problem of estimating the distribution of the properties of a population of general ellipsoidal particles from a section showing elliptical sections. This is much more difficult; there are obviously problems of **identifiability**. The general problem is that of inference about the joint distribution of three variables (the three axes of the ellipsoids) from the joint distribution of two. Further assumptions are needed. These depend on the material examined, and the solutions are mathematically quite difficult. Cruz-Orive [9, 11] has solved the problem for spheroids. He shows how to estimate the distribution of measurements on particles assumed to be oblate spheroids, or prolate spheroids, and shows that there is no way, from the distribution of the properties of elliptical sections, of discriminating between the two cases.

Other particle shapes have been investigated. Nicholson [32] gives a general mathematical discussion, and illustrates, in particular, the possible sections of cylindrical particles. Sections of polyhedra are important in crystallography, but probably not in biological applications. Coleman [5] deals with sections of two-phase particles, such as cells with nuclei.

Effects of Section Thickness

Sections are usually thin slices. If a specimen has two phases, one opaque and one translucent, and a slide is examined by transmitted light, the finite thickness of the slice can cause bias in standard stereological estimates. The interface is not, in general, normal to the section, and the area ratio of the opaque phase is an overestimate of the volume ratio. This bias was first recognized by Holmes [21], and is known as the Holmes effect.

The appropriate correction in the isotropic case was given by Cahn and Nutting [3]. For a section of thickness t, the adjusted estimate is given by

$$\hat{V}_V = A_A - \frac{1}{4} A_V t, \tag{9}$$

where A_V is the volume fraction of the interface. This, of course, is not observable, and must be estimated from

$$\hat{A}_V = 4\pi L_A, \tag{10}$$

where L_A is the area fraction of the interface in the section. This gives, finally,

$$\hat{V}_V = \frac{1}{\pi} L_A t. \tag{11}$$

This is not an unbiased estimate, since the Holmes effect also implies that L_A is reduced in sections of finite thickness, but for thin sections the effect is small. If the assumption of isotropy is false, the adjustment can be seriously misleading; see [45] for a discussion of the anisotropic case, with application to the structure of trabecular bone.

Finite section thickness also affects the estimation of particle size distributions. For spherical particles, Wicksell's integral (6) becomes

$$g(x) = \frac{t f(r)}{\mu_1' + t} + \frac{x}{\mu_1 + t'} \int_x^{R_m} \frac{f(r)}{(r^2 - x^2)^{1/2}} \, \mathrm{d}r. \tag{12}$$

The solution to this equation, with the further modification of a truncation point – a minimum observable section radius – is discussed by Coleman [6]; see also [40] and [22].

Tessellations

A tessellation is a subdivision of p dimensions into closed subsets, divided by edges of dimension $p - 1$. In stereology, we are concerned with tessellations of $\mathcal{R}^3$, and tessellations in $\mathcal{R}^2$ produced by sectioning them.

The Dirichlet tessellation of a **point process** subdivides the space into those sets nearest to each point of the process. This gives convex polyhedra in $\mathcal{R}^3$, or convex polygons in $\mathcal{R}^2$. (These are sometimes known

as Voronoi polygons or polyhedra). The particular case when the underlying point process is Poisson (*see* **Poisson Processes**) has been widely studied under the name of the *cell model*. "Seed" are randomly scattered, and grow at the same rate until the circles or spheres meet, and eventually fill the complete space. The model has been used in studies of crystal growth, and in various ecological applications, such as ground cover by plants, and territories of animals.

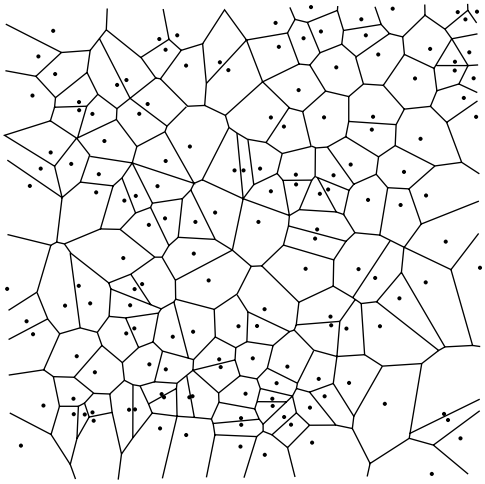

Figure 2 A two-dimensional cell model (the Dirichlet tessellation of a Poisson process). The points show the seeds. From Okabe et al. [33]

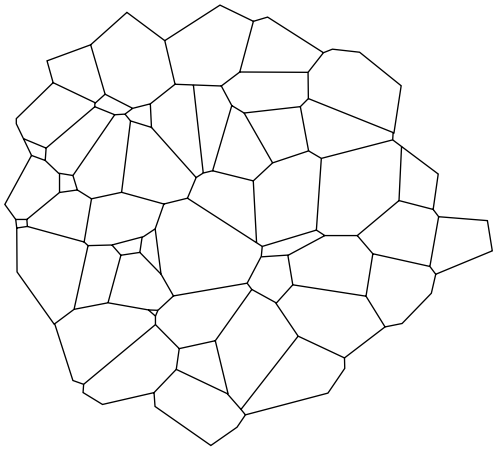

Figure 3 A two-dimensional section of a three-dimensional cell model. From Lorz [24]

The two-dimensional cell model consists of convex polygons, with an average of six edges. (This applies to all Dirichlet tessellations of random processes; it is simply a consequence of the fact that three or more boundaries meet in a point with zero probability). A two-dimensional section of the three-dimensional cell model has similar properties. In fact, Miles [29] has shown that it is a Dirichlet tessellation of a point process, but not of a Poisson process of constant intensity. The main visual difference is that the section has more cells much smaller than the average. Figures 2 and 3 show an example of a two-dimensional cell model, and of a section of a three-dimensional cell model. Table 2 lists some of the more important properties of the models. Notice that the number of vertices and the perimeter, in terms of μ, both show higher variance for the section than for the two-dimensional model.

Another model proposed for crystal growth, the Johnson–Mehl model, has seeds generated by a Poisson process in space *and time*, which then grow at a constant rate. The resulting tessellation has edges that are quadratic curves or surfaces. The cells are not convex, but have the property that all lines through the seed cut the surface twice only. Sections of this model are difficult to interpret, since a single cell may

Table 2 Some properties of the cell model in $\mathcal{R}^2$ and $\mathcal{R}^3$, and of two-dimensional sections of the $\mathcal{R}^3$ model. The intensity of the Poisson process of seeds is represented by λ, and the intensity of centroids in the section by μ. Entries with an asterisk were obtained by **Monte Carlo** simulation

Cell model in $\mathcal{R}^2$			
Vertices/cell	$E(N)$	6	
	$E(N^2)$	37.781	
Perimeter/cell	$E(P)$	$4\lambda^{1/2}$	
	$E(P^2)$	$16.945\lambda^{-1}$	
Cell model in $\mathcal{R}^3$			
Vertices/cell	$E(M)$	27.071	
Edges/cell	$E(E)$	40.606	
Faces/cell	$E(F)$	15.535	
Cell model in $\mathcal{R}^3$; 2-D section			
Intensity of centroids		$1.458\lambda^{2/3}$	μ
Edges/cell	$E(N)$	6	
	$E(N^2)$	38.827*	
Perimeter/cell	$E(P)$	$3.136\lambda^{-1/3}$	$3.79\mu^{-1/2}$
	$E(P^2)$	$11.308\lambda^{-2/3}$	$16.49\mu^{-1}$*

be represented by more than one area in the section. Properties have been studied mainly by simulation.

Early work on these models [19, 26] established the main properties, and Miles [27, 29] derived further results. Many variations are possible, by changing the underlying process, the definition of distance, or the rate of growth. For details, see [33].

Mathematical Morphology

Mathematical morphology was developed for the study of properties of three-dimensional structures from sections, particularly in the context of petrology and the porosity of rocks; see [25]. Properties of interest (porosity, the strength of bones, the healthiness of lungs) may not depend in a simple way on the standard stereological measurements, such as the area ratio. More complicated measurements on the sections might be used; an alternative is to apply transformations to the image, and investigate the relationships between the properties and measurements on the transformed images.

An image is represented by a pixel map on a regular lattice. The work of Matheron [25] and Serra [38] is based on a hexagonal lattice, but computer graphics is more often referred to a square lattice. Each pixel has an associated value z. For a black and white image, z is binary. The original theory was concerned with binary images; later work extends it to grey-level images.

Consider a binary image in which black objects ($z = 1$) appear on a white background ($z = 0$). There are now two basic operations in the calculus:
Erosion consists in peeling a layer, one or more pixels deep, off each object.
Dilatation adds a layer of one or more pixels to each object.

These operations may be combined:
Opening is erosion followed by dilatation.
Closure is dilatation followed by erosion.

Notice that changing the coding, so that white objects ($z = 1$) lie on a dark background ($z = 0$) simply interchanges erosion and dilatation, and opening and closure.

Opening has no effect on large, regular objects; dilatation simply restores the layer removed by erosion. Small and thin objects, and small irregularities on the surface of larger objects, are removed by erosion and not restored by dilatation. Opening thus gives a "smoothed" image, with a lower area ratio than the original. Closure operates in the same way on background irregularities, and gives an image with higher area ratio. Note, however, that edge effects must be considered separately.

A more formal treatment, allowing generalizations and extensions to the simple operations, was given by Serra [38, 39]. Denote by X the lattice with elements $\boldsymbol{x}$, by A the object defined by $Z(\boldsymbol{x}) = 1$, and suppose B is a *structuring element*. B is typically an approximation to a disk, defined on the lattice. On a triangular lattice, B may be hexagonal, the smallest structuring element consisting of a central point and six surrounding points. On a square lattice, it is less easy to approximate circular disks with small pixel patterns; the smallest suitable structuring elements are a point with its four nearest neighbors, or a nine-point square. The structuring element has an origin, the central point in these simple examples.

Table 3, from [37], shows the basic operations defined in terms of the structuring element B.

Minkowski addition gives a score of 1 whenever a point of A scores 1, or when a point not in A is covered by B with its origin on a point of A.

The reflection of B in the origin is $\check{B}$; for the simple structures described above, $\check{B} = B$.

Dilatation and erosion are defined in terms of the structuring element; different structuring elements give different interpretations to the idea of adding or peeling off a layer.

Table 3 Basic operations of Serra's set calculus

Minkowski addition	$A \oplus B = \{\boldsymbol{x} + \boldsymbol{y} \mid \boldsymbol{x} \in A, \boldsymbol{y} \in B\} = \cup_{\boldsymbol{y} \in B}(A + \boldsymbol{y})$
Reflection	$\check{B} = \{-\boldsymbol{x} \mid \boldsymbol{x} \in B\}$
Subtraction	$A \ominus B = (A^c \oplus B)^c = \{\boldsymbol{x} \mid (\boldsymbol{x} - B) \subset A\}$
Dilatation of A by B	$A^B = A \oplus \check{B} = \{\boldsymbol{x} \mid (B + \boldsymbol{x}) \cap A\}$
Erosion of A by B	$A_B = A \ominus \check{B} = \{\boldsymbol{x} \mid (B + \boldsymbol{x}) \subset A\}$
Opening of A by B	$A\omega B = (A \ominus \check{B}) \oplus B = \cup\{B + \boldsymbol{x} \mid (B + \boldsymbol{x}) \subset A\}$
Closure of A by B	$AfB = (A \oplus \check{B}) \ominus B = \cap\{B + \boldsymbol{x} \mid (B + \boldsymbol{x}) \cap A \neq \emptyset\}$

Opening and closure are erosion followed by dilatation, and dilatation followed by erosion respectively, as discussed above.

The extension of the set calculus to grey-level images is straightforward. An image in $\mathcal{R}^2$ can be represented by a three-dimensional plot in which the axis normal to the plane of the image indicates the grey level. This plot is known as the *umbra*. A structuring element B is defined as before in the plane, and its origin is moved to each pixel of the image. *Erosion* then reduces the grey level at the origin to the minimum level in any point of the image covered by B, and *dilatation* raises it to the maximum. Opening and closure are defined as before. Both effect a smoothing of the image, the former by flattening the peaks and the latter by filling in the troughs in the umbra.

Figure 4 illustrates, in section, the opening and closure of a grey-level function.

This section covers only the most basic functions of mathematical morphology. For further details, see [38] and [39]. Many of the extensions are of more importance in the field of **pattern recognition** than in the analysis of two-dimensional sections or projections.

Computation and Software

Devices for taking measurements on two-dimensional images have been available since the 1960s. The *Quantimet* was the earliest practical instrument; it could scan black and white images, record the areas of the two phases, and, using the principle of Buffon's needle, measure the length of an irregular curve. It was first described in 1963, and became commercially available in 1967. For details about the early developments, see [43], Chapter 7. For a more recent account of image processing software, see [36] (*see* **Image Analysis and Tomography**).

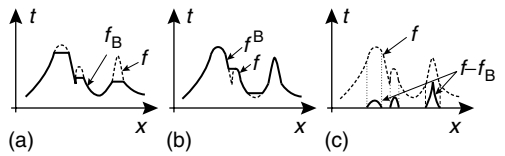

Figure 4 Grey-level morphology. (a) and (b) show respectively the opening and closure of a function by a compact convex set. (c) shows the difference between the function and its opening – the peaks smoothed off by opening. (From Serra [38])

Software for image processing is available, and is being developed and elaborated all the time. SCILAIM is a package developed by a group in Amsterdam; see [20]. The KHOROS system contains a large number of algorithms for image processing. The TargetJr software was initiated in the General Electric Corporate Research and Development department. It was originally designed primarily for photointerpretation and X-ray image analysis. Image Understanding Environment (IUE) is a project started in the United States in 1989 and still under development there and in Europe.

Bibliography

A number of books have been written on stereology; unfortunately, many are out of print or difficult to obtain. They include [6, 10, 18, 35, 41, 42, 44]. The Buffon Bicentenary meeting was influential in bringing practical stereologists and statisticians together. The Proceedings volume has been cited repeatedly, for example, [28].

There have been regular congresses on stereology, and the proceedings, usually special publications not readily available, contain many valuable papers. International Congresses for Stereology have been held at four-yearly intervals since 1963. European Congresses for Stereology have been held, also at four-yearly intervals, since 1973. International Conferences on Stereology and Stochastic Geometry were held in 1981, 1983, 1985, and 1987. Papers presented at these meetings are often published in the *Journal of Microscopy*, which regularly devotes one or more special issues to stereological topics. This journal is the main English language journal publishing papers on stereology.

Image analysis and mathematical morphology is an ever-growing subject. Serra [38] gives the basic calculus developed by Matheron and Serra. Serra [39] is more specialized. A third volume containing algorithms for image analysis has not yet appeared in English.

References

[1] Barbier, E. (1860). Note sur le problème de l'aiguille et le jeu de joint couvert, *Journal des Mathématiques pures et appliquées; par Joseph Liouville* **5**, 273–286.

[2] Buffon, G.L.L. (1777). Essai d'arithmétique morale. Supplément à l'Histoire Naturelle **4**, Paris.

[3] Cahn, J.W. & Nutting, J. (1959). Transmission quantitative metallography, *Transactions of the Metallurgical Society of AIME* **215**, 526–528.

[4] Cochrane, W.G. (1977). *Sampling Techniques*, 3rd Ed., Wiley, New York.

[5] Coleman, R. (1978). The stereological analysis of two-phase particles, in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, R.E. Miles & J. Serra, eds. Springer-Verlag, Berlin, pp. 37–48.

[6] Coleman, R. (1979). *An Introduction to Mathematical Stereology*, Memoirs No. 3. Department of Theoretical Statistics, University of Aarhus, Aarhus.

[7] Crofton, M.W. (1869). On the theory of local probability, applied to straight lines drawn at random in a plane, the method used being also extended to the proof of certain new theorems in the integral calculus, *Philosophical Transactions of the Royal Society of London* **158**, 181–189.

[8] Crofton, M.W. (1885). Probability, in *Encyclopaedia Britannica*, 9th Ed., Vol. 19, Black, Edinburgh, pp. 768–788.

[9] Cruz-Orive, L.-M. (1976). Particle size-shape distributions: the general spheroid problem, *Journal of Microscopy* **107**, 235–253.

[10] Cruz-Orive, L.-M. (1977a). *Lecture Notes, Stereology Course*. Department of Anatomy, University of Bern, Bern.

[11] Cruz-Orive, L.-M. (1977b). Particle size-shape distributions: the general spheroid problem II. Stochastic model and practical guide, *Journal of Microscopy* **112**, 153–167.

[12] Cruz-Orive, L.-M. (1980). On the estimation of particle number, *Mikroskopie (Wien)* **37**,(Suppl), 79–85.

[13] Cruz-Orive, L.-M., Gehr, P., Müller, A. & Weibel, E.R. (1980). Sampling designs for stereology, *Mikroskopie (Wien)* **37**,(Suppl), 149–155.

[14] Cruz-Orive, L.-M. & Weibel, E.R. (1981). Sampling designs for stereology, *Journal of Microscopy* **122**, 235–257.

[15] DeHoff, R.T. (1967). The quantitative estimation of mean surface curvature, *Transactions of the Metallurgical Society of AIME* **239**, 617–621.

[16] DeHoff, R.T. (1978). Stereological uses of the area tangent count, in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, R.E. Miles & J. Serra, eds. Springer-Verlag, Berlin, pp. 99–113.

[17] Delesse, M.A. (1847). Procédé mécanique pour déterminer la composition des roches, *Comptes Rendus des séances de l'Académie des Sciences, Paris* **25**, 544.

[18] Elias, H.-G. & Hyde, D.M. (1985). *Guide to Practical Stereology*. Karger, Basel.

[19] Gilbert, E.N. (1962). Random subdivisions of space into crystals, *Annals of Mathematical Statistics* **33**, 958–972.

[20] Groen, F., Verbeek, P. & de Vries, R. SCILAIM. An Image Analysis Software Package Developed in the University of Amsterdam and Delft University of Technology.

[21] Holmes, A.H. (1927). *Petrographic Methods and Calculations*. Murby, London.

[22] Keiding, N., Jensen, S.T. & Ranek, L. (1972). Maximum likelihood estimation of the size distribution of liver cell nucleii from the observed distribution in a plane section, *Biometrics* **28**, 813–829.

[23] Kendall, M.G. & Moran, P.A.P. (1963). *Geometrical Probability*. Griffin, London.

[24] Lorz, U. (1991). In *Geometrical Problems of Image Processing*, Research and Information Series, Volume 4,. U. Eckhardt, A. Hubler, W. Nagel & G. Werner, eds. Akademie-Verlag, Berlin, pp. 171–178.

[25] Matheron, G. (1967). *Eléments pour une Théorie des Milieux Poreux*. Masson, Paris.

[26] Meijering, J.L. (1953). Interface area, edge length and number of vertices in crystal aggregates with random nucleation, *Phillips Research Reports* **8**, 270–290.

[27] Miles, R.E. (1972). The random division of space, *Advances in Applied Probability* **4**,(Suppl), 243–266.

[28] Miles, R.E. (1978a). The importance of proper model specification in stereology, in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, R.E. Miles & J. Serra, eds. Springer-Verlag, Berlin, pp. 115–136.

[29] Miles, R.E. (1978b). The sampling, by quadrats, of planar aggregates, *Journal of Microscopy* **113**, 257–267.

[30] Miles, R.E. & Davy, P. (1976). Precise and general conditions for the validity of a comprehensive set of stereological fundamental formulae, *Journal of Microscopy* **107**, 211–226.

[31] Miles, R.E. & Serra, J. (1978). En matière d'introduction . . ., in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, R.E. Miles & J. Serra, eds. Springer-Verlag, Berlin, pp. 3–28.

[32] Nicholson, W.L. (1970). Estimation of linear properties of particle size, *Biometrika* **57**, 273–297.

[33] Okabe, A., Boots, B. & Sugihara, K. (1992). *Spatial Tessellations; Concepts and Applications of Voronoi Diagrams*. Wiley, Chichester.

[34] Reith, A. (1978). The non-statistical nature of biological structure and its implications in sampling for stereology of liver tissue, in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, R.E. Miles & J. Serra, eds. Springer-Verlag, Berlin, pp. 181–184.

[35] Reith, A. & Mayhew, T.M. eds. (1988). *Stereology and Morphometry in Electron Microscopy*. Hemisphere, New York.

[36] Rigaut, J.P. (1988). Analyzing electron microscopic images by computer: a guided tour, in *Stereology and Morphometry in Electron Microscopy*, A. Reith & T.M. Mayhew, eds. Hemisphere, New York, pp. 161–191.

[37] Ripley, B.D. (1981). *Spatial Statistics*. Wiley, New York.

[38] Serra, J. (1982a). *Image Analysis and Mathematical Morphology*, Vol. 1. Academic Press, New York.

[39] Serra, J. (1982b). *Image Analysis and Mathematical Morphology*, Vol. 2. Academic Press, New York.

[40] Tallis, G.M. (1970). Estimating the distribution of spherical and elliptical bodies in conglomerates from plane sections, *Biometrics* **26**, 87–103.

[41] Underwood, E.E. (1970). *Quantitative Stereology*. Addison-Wesley, Reading.

[42] Weibel, E.R. (1978). The non-statistical nature of biological structure and its implications on sampling for stereology, in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, R.E. Miles & J. Serra, eds. Springer-Verlag, Berlin, pp. 171–179.

[43] Weibel, E.R. (1979). *Stereological Methods*, Vol. 1. Academic Press, New York.

[44] Weibel, E.R. (1980). *Stereological Methods*, Vol. 2. Academic Press, New York.

[45] Whitehouse, W.J. (1976). Errors in area measurement in thick sections, with special reference to trabecular bone, *Journal of Microscopy* **107**, 183–187.

[46] Wicksell, S.D. (1925). The corpuscle problem. A mathematical study of a biometrical problem, *Biometrika* **17**, 84–99.

[47] Wicksell, S.D. (1926). The corpuscle problem. Second memoir. Case of ellipsoidal corpuscles, *Biometrika* **18**, 151–172.

F.H.C. MARRIOTT

Stimulus–Response Studies

Introduction

Exposure of individuals to environmental stress, such as heat or noise, can be damaging to health. A psychologist wishing to investigate the effect of heat on concentration, for example, might give a number of human subjects a recognition task, which proceeds as follows:

At each of a number of trials, the subjects are either presented with a signal or not, where the signal could, for example, be a light flashing on a screen. In cool conditions, each subject records his/her impression of whether the signal was present or not, and the experiment is then repeated in heat. A "hit" corresponds to a positive response to a signal, while a "false alarm" results from a positive response when no signal is present. For each subject, the overall responses to the experiment may be summarized by the proportions of hits and of false alarms recorded, for each of the two conditions.

A simple probability model for the behavior of the subject in such a stimulus–response situation is to suppose that the stimulus induces a latent random variable, X, from one of two distributions (*see* **Latent Class Analysis**). If the signal was present, then X is supposed to be a **random variable** with cumulative distribution function $F_P(x)$, and if the signal was absent, then X is supposed to be a random variable with the cumulative distribution function $F_A(x)$. The subject's response is then governed by the value of the latent variable; the subject's response is positive (the signal was present) if and only if $X > c$, where c is some cut-off value determined by the subject's natural level of performance. In this case, we can see that the probabilities of a hit and of a false alarm are given as

$$\Pr(\text{Hit}) = 1 - F_P(c); \quad \Pr(\text{False alarm}) = 1 - F_A(c). \tag{1}$$

It is usual to assume that $F_P(x) = F(x - d')$, and $F_A(x) = F(x)$, for some cumulative distribution $F()$, which is the case illustrated in Figure 1. Typically, a **normal** or **logistic** form is adopted for $F()$. The value of c varies between subjects, and may also be manipulated artificially by means of suitable reward schemes. Changing c, but without varying d', produces different points on the **receiver operating characteristic (ROC) curve** illustrated in Figure 2.

The transformation of hit and false alarm rates into a single measure d', to represent the separation of the signal from noise (when no signal was present) is very attractive, and the tables of Freeman [8], providing estimates of d' from empirical proportions of hits and false alarms, have proved very popular. Thus, for example, the effect of heat on performance may be examined by comparing two ROC curves, or equivalently, the two corresponding estimates of d'. A latent **bivariate normal** model was proposed by Metz et al. [16], for the case when the same subjects are used in both conditions, and De Long et al. [5] use the theory of generalized ***U*-statistics** to provide a **nonparametric** approach for comparing correlated ROC curves.

Designed originally for engineering problems, ROC curve methodology now has wide application, not just in psychology, but also in areas such as medicine and ecology; see [10]. An introduction is given by McNicol [15], and reviews are provided in [1, 12, 21, 23].

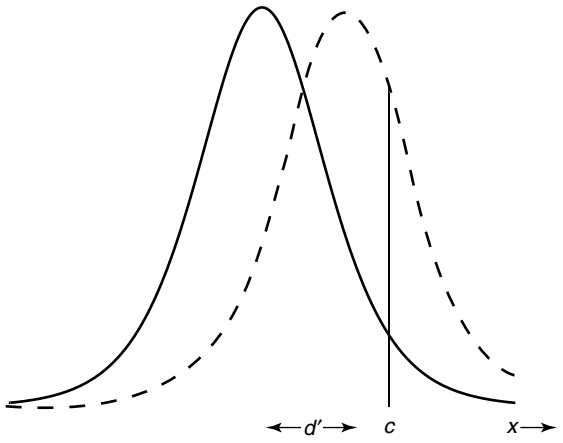

Figure 1 The basic model for the Yes–No signal-detection experiment. A subject responds "Yes" if and only if the latent variable $X > c$. The latent variable has cumulative distribution function $F(x)$ when only noise is presented (probability density function shown by ——) and cumulative distribution function $F(x - d')$ when signal is present (probability density function shown by - - - -)

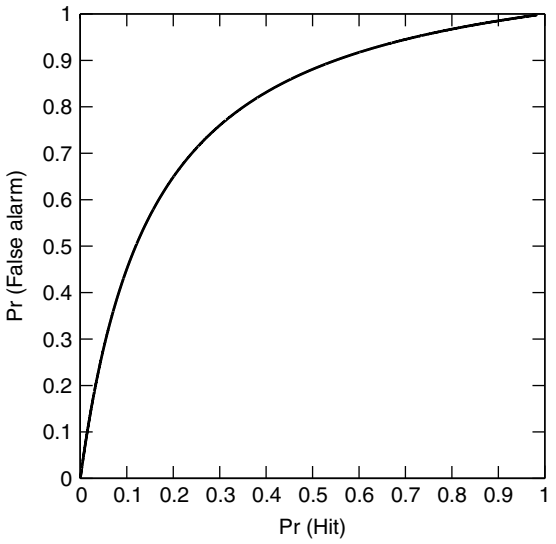

Figure 2 The ROC curve that results from varying c in Figure 1

Extensions

The simple experimental paradigm and model described above may be varied in various ways. In a forced-choice experiment [15, 8], a subject is told that on one of two trials a signal is present, and the task is to decide which of the trials contained the signal. A rating-method experiment is a natural extension of the basic stimulus–response experiment, allowing a subject to express a degree of uncertainty, rather than simply stating "Yes" or "No". The model for this experiment simply extends that of Figure 1, by increasing the number of cut-offs. For example, if there are two cut-offs, c, and c_2, then if $c_1 \leq X < c_2$, the response "maybe" would result. **Maximum likelihood** model-fitting for the rating-method experiment is described in [11]. For the case of two or more cut-offs, the two underlying distributions are not required to have equal variances, though typically, the same distributional form would be assumed. The two cut-off example is then saturated, and explicit maximum likelihood estimates result for the full set of parameters (including the cut-offs).

This model is generally useful for **contingency tables** with **ordered categories** ([14]) and also for **quantal response models** with intermediate categories of response, such as when embryos, say, are classified as deformed, as well as dead or alive; see [17]. Usually, the cut-off values are regarded as **nuisance parameters**, but in some cases, they are also of interest; see [4]. If the underlying distribution is logistic, we have a **proportional-odds** model, while if it is **extreme value**, we obtain a **proportional hazards** model; see also [18].

Applications

Applications to diagnostic medicine (*see* **Diagnostic Tests, Evaluation of**) use the terms "**specificity**", for 1− Pr(False alarm), and "**sensitivity**", for Pr(Hit). The latent variable of engineering and psychology applications is *explicit* when it plays the role of a disease indicator, which may be measured. Murtaugh [19] investigated ROC curve methodology for the case when several such markers are measured on each subject. Such data may alternatively be measured through space, rather than time, and this typically occurs in ecological monitoring studies.

The area, A, under the ROC curve is used as a measure of accuracy of a diagnostic test, for example, in radiology [1], and of discrimination in **discriminant analysis**. The Gini coefficient is defined as 2(A-1/2). A shrinkage correction for both A and the ROC curve is given in [3]; see also [7]. The nonparametric area estimator of [13] is equivalent to a ***Wilcoxon test*** for comparing diseased and normal subjects; however, this assumes exact diagnosis. When diagnosis is imperfect, it is necessary to examine the effect of verification bias. This was done by Gray, Begg, and Greeves [9] in the case of a maximum likelihood estimator of area; Zhou [22], obtained a nonparametric maximum likelihood estimate after formulating the problem in the framework of **missing data**, and making the missing-at-random assumption.

If X and Y denote the ranks of two observations, one from each of two independent populations, it is frequently important to evaluate $\Pr(X < Y)$ and $\Pr(X < Y) - \Pr(Y < X)$; see [6, 20]. Brownie [2] showed how ROC analysis may be used in this context.

References

[1] Begg, C.B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's, *Statistics in Medicine* **10**, 1887–1895.

[2] Brownie, C. (1988). Estimating $\Pr(X < Y)$ in categorized data using "ROC" analysis, *Biometrics* **44**, 615–621.

[3] Copas, J.B. & Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression, *Biometrika* **89**(2), 315–331.

[4] Craig, A. (1979). Discrimination, temperature and time of day, *Human Factors* **21**, 61–68.

[5] De Long, E.R., De Long, D.M. & Clark-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, *Biometrics* **44**, 837–846.

[6] Edwardes, M.D. & de, B. (1995). A confidence interval for $\Pr(X < Y) - \Pr(X > Y)$ estimated from simple cluster samples, *Biometrics* **51**, 571–578.

[7] Eguchi, S. & Copas, J. (2002). A class of logistic-type discriminant functions, *Biometrika* **89**(1), 1–22.

[8] Freeman, P.R. (1973). *Table of d' and β*. Cambridge University Press, Cambridge.

[9] Gray, R., Begg, C.B. & Greeves, R.A. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias, *Medical Decision Making* **4**, 151–164.

[10] Green, D.M. & Swets, J.A. (1988). *Signal Detection Theory and Psychophysics*. Wiley, Chichester.

[11] Grey, D.R. & Morgan, B.J.T. (1972). Some aspects of ROC curve-fitting: normal and logistic models, *Journal of Mathematical Psychology* **9**, 128–139.

[12] Hanley, J.A. (1989). Receiver operating characteristic methodology: the state of the art, *CRC Critical Reviews in Diagnostic Imaging* **29**, 307–335.

[13] Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under operating characteristics curve, *Radiology* **143**, 29–36.

[14] McCullagh, P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society, B* **42**, 109–142.

[15] McNicol, D. (1972). *A Primer of Signal Detection Theory*. George Allen & Unwin Ltd., Sydney.

[16] Metz, D.E., Wang, P.-L. & Kronman, H.B. (1984). A new approach for testing the significance of differences between ROC curves for correlated data, in *Information Processing in Medical Imaging*, F., Decornick ed. Nijhoff, The Hague.

[17] Morgan, B.J.T. (1992). *Analysis of quantal response data*. Chapman & Hall, London.

[18] Morgan, B.J.T. (1976). The uniform distribution in signal detection theory, *British Journal of Mathematical & Statistical Psychology* **29**, 81–88.

[19] Murtaugh, P.A. (1995). ROC curves with multiple marker measurements, *Biometrics* **51**, 1514–1522.

[20] Simonoff, J.S., Hochberg, Y. & Reiser, B. (1986). Alternative estimation procedures for $\Pr(X < Y)$ in categorized data, *Biometrics* **42**, 895–907.

[21] Swets, J.A. (1988). Measuring the accuracy of diagnostic systems, *Science* **240**, 1285–1293.

[22] Zhou, X.H. (1996). A nonparametric maximum-likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias, *Biometrics* **52**, 299–305.

[23] Zweig, M.H. & Campbell, G. (1993). Receiver operating characteristic (ROC) plots – a fundamental evaluation tool in clinical medicine, *Clinical Chemistry* **39**, 561–577.

(*See also* **Median Effective Dose**)

BYRON J.T. MORGAN

Stochastic Approximation

In the mathematical sciences, iterative methods to find roots of equations have a long and honored history dating back, of course, to Sir Isaac Newton (*see* **Optimization and Nonlinear Equations**). Stochastic approximation, which is concerned with the same problem when these roots are observable in the presence of statistical variation, has been available for less than half a century. The seminal paper by Robbins & Monro [26] appeared in 1951. Applications of stochastic approximation have been primarily in the areas of adaptive control, statistical **simulation** and sequential estimation. Biomedical areas that have used stochastic approximation include quantal response, evolution and survival analysis.

Sampson [29] provides an excellent overview of stochastic approximation from a statistical viewpoint.

Theoretical Development and Overview

Suppose an experimenter is concerned with obtaining the (unique) root of $M(x)$; that is, the value θ such that $M(\theta) = m_0$. However, $M(x)$, a real-valued, differentiable function of x, is not directly observable. What one does observe is a **random variable** $Y(x)$ with cumulative distribution function (cdf) $F(y|x)$, where the expected value of $Y(x)$ is $M(x)$. That is,

$$\mathrm{E}(Y|x) = \int_{-\infty}^{\infty} y(x)\,\mathrm{d}F(y|x) = M(x) < \infty, \quad (1)$$

for all x. The principal idea is to determine the value of θ by successive approximation.

To this end, suppose $\{X_n\}$, $n \geq 1$ is a sequence of random variables wherein X_1 is selected arbitrarily and, upon its selection, the succeeding values are defined by

$$X_{n+1} = X_n - a_n[Y(X_n) - m_0], \quad (2)$$

where $\{na_n\}$ is a sequence of bounded, positive constants. Robbins & Munro [26] proved that X_n converges to θ, stochastically, $(X_n \xrightarrow{\mathrm{p}} \theta)$ as $n \to \infty$ provided the following conditions hold:

1. there exists a θ, such that $M(\theta) = m_0$,
2. $Y(x)$ is uniformly bounded for all x, almost surely,
3. var $X_1 < \infty$,

and

4. $M(x)$ is nondecreasing and $M'(\theta) > 0$.

Since the original development by Robbins & Munro [26], there has been a myriad of refinements, improvements and significant advances in stochastic approximation. While not an all-inclusive list, the following synopsis provides some of the more important results that have accrued since 1952.

1. Wolfowitz [35] shows that condition 2 can be replaced with the weaker conditions that $M(x)$ is uniformly bounded for all x as is $\mathrm{E}[Y - M(x)]^2$. He also gives conditions under which condition 4 can be relaxed.
2. Kiefer & Wolfowitz [18], using ideas of both Robbins & Munro [26] and Wolfowitz [35], obtain a stochastic approximation procedure for finding the maximum of a regression function.
3. In 1954, Blum [1] is able to weaken, further, conditions that were considered in [18] and [35]. He also solves a similar problem when $M(x)$ is the median rather than the mean of the cdf $F(y|x)$. Further, Kallianpur [16] obtains an estimate of the order of magnitude of the nth iteration x_n in terms of $\mathrm{E}(x_n - \theta)^2$. Finally, Blum [2] considers multivariate stochastic approximation procedures.
4. Subsequently, Dvoretzky [7] develops a more general stochastic approximation scheme that includes both the Robbins–Munro and Kiefer–Wolfowitz procedures as special cases. Blum [1] also obtains a generalized version of the Robbins–Monro process that is related to the Dvoretzky procedure.
5. Issues, such as stopping rules, constrained optimization, and asymptotic normality, have been considered by several authors. Further information, including relevant references, can be obtained in [29].

Applications

Stochastic approximation has been used, to some extent, in the biomedical sciences. In particular, **neural network** research has been the focus of many of these applications.

Table 1 provides a selection of some of the biomedical studies that have used stochastic

Table 1 Application of the stochastic approximation method in biomedical sciences

Study area	Application	References
Population biology	Pharmacokinetics	[21]
	Genetic mutation rates	[17]
	Genetics	[24]
	Population ecologies	[31]
	Evolution of reproduction efforts	[25]
Epidemiology	Disease monitoring	[19]
	HIV	[32] and [33]
	Illness–death	[4]
	Survival analysis	[30]
	Maximum likelihood estimation	[5]
Forestry	Regression estimation	[20]
	Forest fire protection	[9] and [10]
Fisheries	Management	[3] and [27]
Environmental	Water management–pollution	[15]
Biology	Electrophysiology	[6]
	Myocardial contraction	[12]
	50% dosing and other drug dosing	[28]
Psychology	Memory, learning	[14]
	Learning, pattern recognition	[23]
	Psychophysical measurement–threshold determination	[34]
Neural networks	Numerous studies	[8, 11, 13], and [22]

approximation. Table 1 is, of course, not an all-inclusive list. However, it does indicate that stochastic approximation has been a useful technique, especially in neural network research.

References

[1] Blum, J.R. (1954). Approximation methods which converge with probability one, *Annals of Mathematical Statistics* **25**, 382–386.

[2] Blum, J.R. (1954). Multidimensional stochastic approximation methods, *Annals of Mathematical Statistics* **25**, 737–744.

[3] Burr, R.L. (1988). Inferring the distribution of the parameters of the Von Bertalaniffy growth model from length movements, *Canadian Journal of Fisheries and Aquatic Science* **45**, 1779–1788.

[4] Chiang, Y.-K., Hardy, R.J., Hawkins, C.M. & Kapadia, A.S. (1989). An illness–death process with time-dependent covariates, *Biometrics* **45**, 669–682.

[5] Choi, Y.J. & Severo, N.C. (1988). An approximation for the maximum likelihood estimator of the infection rate in the simple stochastic epidemic, *Biometrika* **75**, 392–394.

[6] Colquhoun, D. & Hawkes, A.G. (1990). Stochastic properties of ion channel openings and bursts in a membrane patch that contains two channels: evidence concerning the number of channels present when a record containing only single openings is observed, *Proceedings of the Royal Society of London, Series B: Biological Sciences* **240**, 453–477.

[7] Dvoretzky, A. (1956). On stochastic approximation, in *Proceedings of the Third Berkeley Symposium on the Mathematics of Statistical Probability*, Vol. 1. University of California Press, Berkeley, pp. 39–55.

[8] Elanayar, V.T.S. & Shin, Y.C. (1994). Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems. *IEEE Transactions on Neural Networks* **5**, 594.

[9] Fried, J.S. & Gilless, J.K. (1988). Stochastic representation of fire occurrence in a wildland fire protection planning model for California. *Forest Science* **34**, 948–959.

[10] Fried, J.S. & Gilless, J.K. (1989). Expert opinion estimation of fireline production rates, *Forest Science* **35**, 870–877.

[11] Gopal, S. & Fischer, M.M. (1996). Learning in single hidden-layer feedforward network models – backpropagation in a spatial interaction modeling context, *Geographical Analysis*, **28**, 38–55.

[12] Goussard, Y., Krenz, W.C., Stark, L. & Dernoment, G. (1991). Practical identification of functional expensions of nonlinear systems submitted to non-Gaussian inputs, *Annals of Biomedical Engineering* **19**, 401–427.

[13] Gusev, S.V. & Krasulina, T.P. (1995). An algorithm for stochastic approximation with a preassigned probability

of not exceeding a required threshold, *Journal of Computer Systems, Sciences International* **33**, 39–41.

[14] Heath, R.A. & Fulham, R. (1988). An adaptive filter model for recognition memory, *British Journal of Mathematical and Statistical Psychology* **41**, 119–144.

[15] Hochman, E., Zilberman, D. & Just, R. (1977). Internalization in a stochastic pollution model, *Water Resources Research* **13**, 877–881.

[16] Kallianpur, G. (1954). A note on the Robbins–Monro stochastic approximation method, *Annals of Mathematical Statistics* **25**, 386–388.

[17] Kepler, T.B. & Perelson, A.S. (1995). Modeling and optimization of populations subject to time-dependent mutation, *Proceedings of the National Academy of Sciences* **92**, 8219–8223.

[18] Kiefer, J. & Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function, *Annals of Mathematical Statistics* **23**, 462–466.

[19] Maryak, J.L., Spall, J.C. & Silberman, G.L. (1995). Uncertainties for recursive estimators in nonlinear state-space models with applications to epidemiology, *Automatica* **31**, 1889–1892.

[20] Meng, C.H., Tang, S.Z. & Burk, T.E. (1990). A stochastic restrictions regression model approach to volume equation estimation, *Forest Science* **36**, 54–65.

[21] Mentre, F., Mallet, A. & Steimer, J. (1988). Hyperparameter estimation using stochastic approximation with application to population pharmacokinetics, *Biometrics* **44**, 673–684.

[22] Najim, K. & Chtourou, M. (1994). Neural networks synthesis based on stochastic approximation algorithm, *International Journal of Systems Science* **25**, 1219.

[23] Pathak-Pal, A. & Pal, S.K. (1987). Learning with mislabeled training samples using stochastic approximation, *IEEE Transaction on Systems, Man and Cybernetics* **17**, 1072–1077.

[24] Pollak, E. (1987). On the theory of partially inbreeding finite populations. I. Partial selfing, *Genetics* **117**, 353–360.

[25] Real, L.A. & Ellner, S. (1992). Life history evolution in stochastic environments: a graphical mean–variance approach, *Ecology* **73**, 1227–1236.

[26] Robbins, H. & Monro, S. (1951). A stochastic approximation method, *Annals of Mathematical Statistics* **22**, 400–407.

[27] Ruppert, D., Reish, R.L. & Deriso, R.B. (1984). Optimization using stochastic approximation and Monte Carlo simulation (with application to harvesting of Atlantic menhaden), *Biometrics* **40**, 535–545.

[28] Rybak, E.I., Lisunkin, I.I. & Kalinin, O.M. (1966). Determination of 50 per cent and other doses by the stochastic approximation method (Russian), *Farmakologiia i Toksikologiia* **29**, 368–370.

[29] Sampson, A.R. (1988). Stochastic approximation, in *Encyclopedia of Statistical Sciences*, Vol. 8., S. Kott & N.L. Johnson, eds. Wiley, New York, pp. 784–789.

[30] Satten, G.A. (1996). Rank-based inference in the proportional hazards model for interval censored data, *Biometrika* **83**, 355–370.

[31] Sosa Burgos, L.M. (1991). Ecology of two coexisting populations of lagomorphs in the Mojave Desert, *Lepus californicus and Sylvilagus audubonii. Dissertation*, University of California, Los Angeles, DAI, 52-07B, 3359.

[32] Tan, W.Y., Lee, S.R. & Tang, S.C. (1995). Characterization of HIV infection and seroconversion by a stochastic model of the HIV epidemic, *Mathematical Biosciences* **126**, 81–123.

[33] Tan, W.Y. & Tang, S.C. (1993). A stochastic model of the HIV epidemic involving both sexual contact and IV drug use, *Mathematical and Computer Modeling* **17**, 31–57.

[34] Treutwein, B. (1995). Adaptive psychophysical procedures, *Vision Research* **35**, 2503–2522.

[35] Wolfowitz, J. (1952). On the stochastic approximation method of Robbins and Monro, *Annals of Mathematical Statistics* **23**, 457–461.

(*See also* **Median Effective Dose**; **Up-and-Down Method**)

ALAN J. GROSS & DAVID C. MCLEAN, JR

Stochastic Limit and Order Relations

The O and o notation for nonstochastic functions and sequences (*see* **Orders of Magnitude**) can be generalized to stochastic **random variables**, leading to the O_p and o_p notation. The stochastic order relations were derived originally by Mann & Wald [2], and discussed by Chernoff [1] and Pratt [3].

Assume that $\{X_n\}$ is a sequence of random variables on the extended real line $[-\infty, \infty]$, with probability distribution $\{P_n\}$. Let a_n denote a sequence of points on $[-\infty, \infty]$.

Definition 1. We write $X_n = o_p(a_n)$ if, for every $\eta > 0$,

$$P_n\left\{\left|\frac{X_n}{a_n}\right| \leq \eta\right\} \to 1 \text{ as } n \to \infty.$$

That is, the sequence $\{|X_n/a_n|\}$ approaches zero in probability. Note that $X_n = o_p(1)$ is also written: $X_n \xrightarrow{\text{p}} 0$.

Definition 1 is equivalent to Definition 1′ as follows.

Definition 1′. $X_n = o_p(a_n)$ if, for every positive ε and η, there exists an N such that

$$P_n\left\{\left|\frac{X_n}{a_n}\right| \leq \eta\right\} \geq 1 - \varepsilon, \quad \text{for } n > N.$$

Definition 1′ has a suitable analog for the O_p notation as follows.

Definition 2. We write $X_n = O_p(a_n)$ if, for every positive ε, there exist N and $\eta > 0$ such that

$$P_n\left\{\left|\frac{X_n}{a_n}\right| \leq \eta\right\} \geq 1 - \varepsilon, \quad \text{for } n > N.$$

That is, the sequence $\{X_n/a_n\}$ is *bounded in probability*.

The following theorems, derived by Pratt [3], are useful in derivations of large sample theory and rate of convergence for approximations.

Theorem. If $Y_n - Z_n = o_p(1)$, $Z_n = O_p(1)$, and g is a continuous function, then $g(Y_n) - g(Z_n) = o_p(1)$.

Theorem. If $Y_n - Y = o_p(1)$ and $Z_n - Z = o_p(1)$, and $f(y, z)$ is jointly continuous, then $f(Y_n, Z_n) - f(Y, Z) = o_p(1)$.

Example: Normal Approximation for the Binomial Distribution

Assume that $X \sim \text{Bin}(n, \pi)$ is a **binomial** random variable with index n and probability π. Then all cumulants of X (*see* **Characteristic Function**) are $O(n)$ and

$$\frac{X - n\pi}{[n\pi(1-\pi)]^{1/2}} \sim \text{N}(0, 1) + O_p(n^{-1/2}).$$

References

[1] Chernoff, H. (1956). Large sample theory: parametric case, *Annals of Mathematical Statistics* **27**, 1–22.

[2] Mann, H.B. & Wald, A. (1943). On stochastic limit and order relationships, *Annals of Mathematical Statistics* **14**, 217–226.

[3] Pratt, J. (1959). On a general concept of "In probability", *Annals of Mathematical Statistics* **30**, 549–558.

(*See also* **Central Limit Theory**; **Convergence in Distribution and in Probability**; **Large-sample Theory**)

MEI-LING TING LEE

Stochastic Processes

Time, life, and risks are three basic elements in empirical processes studied in biostatistical research. Risks of birth, risks of illness, risks of death, and other risks continuously act on human beings with varying degrees of intensity and varying degrees of frequency. Recent advances in stochastic processes have made it possible to study systematically these risks in the human population from a probabilistic point of view. Many have contributed to the theoretical development of stochastic processes to the high level of sophistication they enjoy today. Among them we mention Markov, **Kolmogorov** [22, 23], Feller [17], Doob [13, 14], and Chung [12]. The purpose of this article is to review various processes which have applications in biostatistical and epidemiologic research. Solutions are given for most problems described here, but theoretical justifications may not always be adequate. For the reader who will not take a formula or a theorem at face value without a formal proof, further references are given.

The order of presentation in this article is from discrete processes to continuous processes. The major form of discrete process is the **Markov chain**. In continuous processes, we proceed from a general birth process through the birth–death processes to the finite **Markov process**. A few special topics of interest are discussed in the sections between these on the birth processes and the birth–death processes. We end this brief introduction with a definition:

A stochastic process $\{X(t); t \in [0, \infty)\}$ *is a family of random variables describing an empirical process, whose development is governed by probability laws.* The parameter t, which is often interpreted as time, is real-valued, but it may be either discrete or continuous. The **random variable** $X(t)$ may be real- or complex-valued, or it may take the form of a vector. In diffusion processes, for example, both t and $X(t)$ are continuous variables, whereas in Markov chains, t and $X(t)$ take on discrete values. In the processes of **population growth**, time t is a continuous parameter, but the random variable $X(t)$, the population size at time t, has a discrete set of positive integers.

In some stochastic processes the one-dimensional time parameter is replaced by a multidimensional parameter, such as the coordinates of a point in multidimensional space. These are called *random fields*. A simple example would be a process describing the random fluctuations in some variable observed at points on a two-dimensional surface. The term may be applied also to more complex generalizations (see [1]).

Random Walk

The one-dimensional random walk is an extension of Bernoulli trials (*see* **Binary Data**). It is closely related to **branching processes** and to gambler's ruin in probability theory (see, for example, [27] and [17]), to **Brownian motion and diffusion processes**, to **sequential analysis** [28], and to sequential clinical trials [2] (*see* **Data and Safety Monitoring**). It is presented here as a prelude to the discrete-time Markov chains.

Position of Particle

In a one-dimensional random walk of a particle starting from the origin, the position of the particle is designated by $\pm 1, \pm 2, \ldots$. Let X_i be the outcome of the ith move, with $\Pr\{X_i = +1\} = p$ and $\Pr\{X_i = -1\} = q$, where $p + q = 1$. The expectation of X_i is $\mathrm{E}[X_i] = p - q$, and the variance of X_i is $\mathrm{var}(X_i) = 4pq$. The probability **generating function** of X_i is

$$g_i(s) = ps + qs^{-1}.$$

Let $Z_n = X_1 + \cdots + X_n$ be the position of the particle after n moves. We need the probability $\Pr\{Z_n = k\}$. The probability generating function of Z_n is

$$G_{Z_n}(s) = [ps + qs^{-1}]^n = s^{-n}[ps^2 + q]^n$$
$$= \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} s^{2i-n}.$$

Substituting $k = 2i - n$ so that $i = (n + k)/2$, we find

$$\Pr\{Z_n = k\} = \binom{n}{\frac{n+k}{2}} p^{(n+k)/2} q^{(n-k)/2},$$

with expectation $\mathrm{E}(Z_n) = n(p - q)$ and variance $\mathrm{var}(Z_n) = 4npq$. The values that Z_n assumes may be odd numbers or even numbers depending on whether n is odd or even. When n is an odd number, the probability that Z_n will assume even values

is zero. When n is an even number, the probability that Z_n will take on odd values is zero. For Z_n to assume a value k, the particle must take $(n+k)/2$ steps to the right and $(n-k)/2$ steps to the left so that $[(n+k)/2-(n-k)/2]=k$ is the net displacement to the right of the origin. Of course, k can be either positive or negative. When $Z_n=0$, the particle returns to the origin. When n is odd, the probability of the particle returning to the origin is zero. When n is even,

$$\Pr\{Z_n=0\}=\binom{n}{\frac{n}{2}}p^{n/2}q^{n/2}.$$

We now consider two extensions of the simple random walk.

Limiting Case of a Diffusion Process

The notion involved in the random walk is clear and the mathematics is simple. It is therefore quite helpful to use the random walk to explain rather more subtle concepts of Brownian movement and diffusion processes. When both the size of step and the time needed to make a move are infinitesimal, the density function of the total displacement up to time t is continuous and satisfies a partial differential equation, known as the Fokker–Planck diffusion equation. While a direct approach to establishing the differential equation from the viewpoint of diffusion processes is somewhat difficult, as a limiting case of the random walk the differential equation becomes logical (*see* **Brownian Motion and Diffusion Processes**).

Two-Dimensional Random Walk

The one-dimensional random walk presented in this section can be extended to a random walk in a two-dimensional plane or in a high-dimensional space. In a two-dimensional random walk, let X_{ij} be the outcome of the jth move in the ith coordinate, with $\Pr\{X_{ij}=+1\}=p_i$ and $\Pr\{X_{ij}=-1\}=q_i$, where $p_i+q_i=1$, for $i=1,2$ and $j=1,2,\ldots$. Let $Z_{in}=\sum_j^n X_{ij}$ be the total displacement on the ith axis after n moves. The probability $\Pr\{Z_{in}=k_i\}$ is computed separately for each coordinate i. The total displacements in the two coordinates are represented by the vector $\mathbf{Z}_n=(Z_{1n},Z_{2n})$. Therefore, the whole extension process is quite simple, and can carry over to a random walk in any high-dimensional space.

The components Z_{1n} and Z_{2n}, however, should not be treated as independent random variables in a bivariate vector. They are merely the coordinates of the particle after n moves. Furthermore, the particle will not reach everywhere in the two-dimensional plane. It will reach only those positions where the sum, in absolute values, of Z_{1n} and Z_{2n} is zero or a multiple of 2. For example, the particle could be in a position with coordinates (3, 5), or (−3, 5), but not in a position with coordinates (4, 5), or (3, 6).

Combining the above two extensions, we see that a continuous-time random walk in a three-dimensional space will be a close description of Brownian movement, since that describes phenomena in three-dimensional space.

Gambler's Ruin

Two players A and B, with their initial a and b, play a series of games. Their respective probabilities of winning a game are p and q. We seek to determine the probability, R_a, that the ruin of player A will eventually occur and the probability, R_b, that the ruin of player B will eventually occur, if the game is to be played until one of them becomes bankrupt. In the language of the random walk, these are the probabilities that a particle will be absorbed at the barriers $x=0$ and $x=a+b$, respectively.

To determine the probability of A's ruin, R_a, we first establish, and then solve, a system of difference equations. Consider more generally R_x, the probability of A's ruin when his capital is x, and establish a difference equation as the result of the first subsequent move. Gambler A may win the game with a probability p and then his probability of ruin will be R_{x+1}, or he may lose the game with a probability q and then his ruin probability will be R_{x-1}. It follows that

$$R_x=pR_{x+1}+qR_{x-1},\quad x=1,2,\ldots,a+b-1,$$

which has the general solution

$$R_x=c+d\left(\frac{q}{p}\right)^x.$$

The constants c and d are to be determined by the boundary conditions. If $x=0$, then A's ruin is certain; if $x=a+b$, then A's ruin is impossible.

Therefore the boundary conditions are $R_0 = 1$ and $R_{a+b} = 0$. Using these conditions, we find

$$R_x = \frac{(q/p)^x - (q/p)^{a+b}}{1 - (q/p)^{a+b}}.$$

When $x = a$, the probability of A's ruin is

$$R_a = \frac{(q/p)^a - (q/p)^{a+b}}{1 - (q/p)^{a+b}}.$$

Similarly, the probability of B's ruin is

$$R_b = \frac{1 - (q/p)^a}{1 - (q/p)^{a+b}}.$$

Since $R_a + R_b = 1$, it is certain that either A or B will eventually lose all his initial capital. An infinite series of games should not be expected.

When $p = q = 1/2$, we use L'Hôpital's rule to obtain the probabilities

$$R_a = \frac{b}{a+b} \quad \text{and} \quad R_b = \frac{a}{a+b}.$$

Expected Gain

R_a and R_b are the probabilities of A's gain of $-a$ and $+b$, respectively. Hence, his expected gain is:

$$\mathrm{E}(G) = -aR_a + bR_b.$$

When $a = b$,

$$\mathrm{E}[G] = a\frac{1 - (q/p)^a}{1 + (q/p)^a}.$$

In this case $\mathrm{E}[G] > 0$ if $q < p$, and $\mathrm{E}[G] < 0$ if $p < q$. When $p = q = 1/2$, $\mathrm{E}[G] = 0$.

The probabilities R_a and R_b depend not only on the probability of losing or winning a single game, but also on the amount of a player's initial capital. Suppose B is so enormously rich that $b = \infty$. Taking the limit in the formula of R_a as $b \to \infty$ yields the probability of A's eventual ruin:

$$R_a = \begin{cases} 1, & \text{if} \quad p < q, \\ 1, & \text{if} \quad p = q, \\ \left(\dfrac{q}{p}\right)^a, & \text{if} \quad p > q. \end{cases}$$

The second case above is interesting. Even if the game is fair ($p = q$), A's eventual ruin is certain simply because B is much richer than he is.

Expected Number of Games (Duration of Play)

Suppose that player A has an initial capital of x; the series of games ends as soon as either A loses all his capital or his capital becomes $a + b$. Let D_x be the expected number of games to be played. It is easy to show that D_x satisfies the following difference equation:

$$D_x = pD_{x+1} + qD_{x-1} + 1, \quad x = 1, 2, \ldots, a + b - 1.$$

The series terminates when $x = 0$ or $x = a + b$. Therefore the boundary conditions are $D_0 = 0$ and $D_{a+b} = 0$. The above difference equation is almost the same as the equation for R_x, except for the addition of unity. This addition makes the equation nonhomogeneous. Clearly, $D_x = x/(q - p)$ is a solution. The general solution of the equation is the sum:

$$D_x = \frac{x}{q-p} + \left[c + d\left(\frac{q}{p}\right)^x\right].$$

Using the boundary conditions $D_0 = 0$ and $D_{a+b} = 0$ to determine the constants c and d, we find the solution

$$D_x = \frac{x}{q-p} - \frac{a+b}{q-p}\left[\frac{1 - (q/p)^x}{1 - (q/p)^{a+b}}\right].$$

The expected number of games depends on the values of p and q. For given $a + b$ and x, D_x increases as p (and q) approaches 1/2. When $p = q = 1/2$, the expected number of games is $D_x = x(a + b - x)$. The expected number takes a maximum value of $D_x = [(a + b)/2]^2$ when $x = (a + b)/2$. If $a = \$100$ and $b = \$100$, and A and B play a series of fair games at a stake of \$1 per game, they will, on average, be expected to play 10 000 games before either one of them loses his entire capital.

Markov Chains

Most statistical and probability theory has been developed for cases where the random variables involved are independent. The classical **central limit theorem** and the **law of large numbers** are prominent examples. In many practical situations, however, the random variables involved are neither independent nor identically distributed. Such phenomena are especially prevalent when the observations are made in sequence.

For example, in sampling without replacement from a dichotomous population consisting of "successes" and "failures", the probability of choosing a "success" is a function of the previous elements sampled. In the random walk discussed in the previous section, the location of a particle after a given move depends on the previous moves. In the Markov chain, we study dependence of a particular kind. When random variables are observed in sequence, the distribution of a random variable is dependent on only the immediately preceding observed random variable and not on those that came before it.

The theory of Markov chains, or discrete-time Markov processes, is named after A.A. Markov, who in 1907 introduced the concept of chains with a discrete parameter and finite number of states. Kolmogorov [23] extended the theory for the denumerable case; Doob [13] and Paul Levy in 1951 introduced continuous-parameter chains. While many others have contributed to the advancement of Markov theory, W. Feller and K.L. Chung are among those responsible for the present status in probability theory that the Markov chain enjoys. Chung [12] gave a comprehensive theoretical treatment of the subject, and Feller [17] gave a most lucid account of both theoretical and practical aspects of Markov chains.

The purpose of this section is to introduce the Markov chain from a practical point of view. Included are the essentials necessary for an understanding and appreciation of the topic. A good reference is [17]; see also [10]. For further discussion, see the article on **Markov Chains**.

Definition 1. A sequence of random variables $\{X_\alpha, \alpha = 0, 1, \ldots\}$ is called a Markov chain if, for every collection of integers, $\alpha_0 < \alpha_1 < \cdots < \alpha_n < \beta$, the conditional distributions of X_β satisfy the relation:

$$\Pr\{X_\beta = i_\beta | X_{\alpha_0}, \ldots, X_{\alpha_n}\} = \Pr\{X_\beta = i_\beta | X_{\alpha_n}\}, \quad \text{for all } i_\beta. \tag{1}$$

Thus, given a knowledge of the present state (X_{α_n}), the outcome in the future $(X_\beta = i_\beta)$ is no longer dependent upon the past $(X_{\alpha_0}, \ldots, X_{\alpha_{n-1}})$.

Absolute Probabilities and Transition Probabilities

We denote for each X_α the absolute probability by

$$\Pr\{X_\alpha = i_\alpha\} = a_{i_\alpha}, \tag{2}$$

and for every pair of random variables X_α and $X_\beta, \alpha < \beta$, the transition (conditional) probability by

$$\Pr\{X_\beta = i_\beta | X_\alpha = i_\alpha\} = P_{i_\alpha, i_\beta}, \tag{3}$$

with the conditions that

$$\sum_{i_\alpha} \Pr\{X_\alpha = i_\alpha\} = \sum_{i_\alpha} a_{i_\alpha} = 1 \quad \text{and}$$

$$\sum_{i_\beta} P_{i_\alpha, i_\beta} = 1.$$

Therefore, the joint probabilities of $X_\alpha, X_\beta, X_\gamma$, for $\alpha < \beta < \gamma$, are given by

$$\Pr\{X_\alpha = i_\alpha, X_\beta = i_\beta, X_\gamma = i_\gamma\} = a_{i_\alpha} P_{i_\alpha, i_\beta} P_{i_\beta, i_\gamma}.$$

Generally, for any collection of integers $\alpha < \beta < \cdots < \delta < \varepsilon$ the joint probabilities are

$$\Pr\{X_\alpha = i_\alpha, X_\beta = i_\beta, \ldots, X_\delta = i_\delta, X_\varepsilon = i_\varepsilon\} = a_{i_\alpha} P_{i_\alpha, i_\beta} \ldots P_{i_\delta, i_\varepsilon}. \tag{4}$$

An important feature of the Markov chain, and indeed of stochastic processes in general, is that the random variables are observed in sequence and the order of the sequence, such as the one in (4), should not be disturbed. In a Markov chain describing a stochastic process, the totality of possible values of random variables X_α constitute the state space of the system. The event associated with the absolute probability in (2) is that the system is in state i_α at time α (or the αth step). The conditional probability in (3) describes a transition from state i_α at α to state i_β at β. A Markov chain with state space being the set of all the nonnegative integers is completely determined by the initial absolute probability distribution

$$\Pr\{X_0 = i_0\} = a_{i_0}, \quad i_0 = 1, 2, \ldots$$

and the transition probabilities

$$\Pr\{X_{\alpha+1} = i_{\alpha+1} | X_\alpha = i_\alpha\} = P_{i_\alpha, i_{\alpha+1}}, \quad i_\alpha, i_{\alpha+1} = 1, 2, \ldots \text{ for } \alpha = 0, 1, \ldots.$$

Example 1. Life Table Urns

Balls are drawn with replacement from an infinite sequence of urns numbered $0, 1, \ldots$. In the αth urn, there is a proportion p_α of white balls and a proportion q_α of black balls with $0 < p_\alpha < 1$ and $p_\alpha + q_\alpha = 1$. Beginning with the 0th urn, $X_0 = i_0$ balls are drawn of which $X_1 = i_1$ are white; a total of i_1 balls is drawn from the first urn of which $X_2 = i_2$ are white; i_2 balls are then drawn from the second urn of which $X_3 = i_3$ balls are white, and so on. In general, the number $X_{\alpha+1} = i_{\alpha+1}$ of white balls drawn from the αth urn is the number of balls to be drawn from the $(\alpha + 1)$th urn. The experiment terminates as soon as the number of white balls drawn from an urn is zero. Clearly, X_1, the number of white balls drawn from the 0th urn, has a **binomial distribution**:

$$\Pr\{X_1 = i_1 | i_0\} = \binom{i_0}{i_1} p_0^{i_1} q_0^{i_0 - i_1}, \quad i_1 = 0, \ldots, i_0.$$

The number of white balls drawn (X_2) from the first urn depends only on the number of drawings (X_1) from that urn but not on i_0. Therefore, given $X_0 = i_0$, $X_1 = i_1$, the probability of X_2 is

$$\begin{aligned} \Pr\{X_2 = i_2 | i_0, i_1\} &= \Pr\{X_2 = i_2 | i_1\} \\ &= p_{i_1, i_2} = \binom{i_1}{i_2} p_1^{i_2} q_1^{i_1 - i_2}, \\ & \quad 0 \le i_2 \le i_1 \le i_0. \end{aligned}$$

Generally,

$$\begin{aligned} &\Pr\{X_\beta = i_\beta | i_0, \ldots, i_\alpha\} = \Pr\{X_\beta = i_\beta | i_\alpha\} = P_{i_\alpha, i_\beta}, \\ &\quad i_\alpha > 0; i_\beta = 0, \ldots, i_\alpha. \end{aligned}$$

This urn model was devised to describe the **life table** where X_0 was the size of the original cohort with which a life table starts. The number X_α is the number of people of exact age α, and $X_{\alpha+1}$ is the number surviving to the end of the αth age interval.

Time-Homogeneous Markov Chains

A Markov chain is homogeneous with respect to time if the transition probabilities

$$\Pr\{X_{\alpha+1} = j | X_\alpha = i\} = p_{ij} \tag{5}$$

are independent of α. We shall be studying mainly time-homogeneous Markov chains in this section. A chain is a finite chain if there are a finite number of states, an infinite chain if there are an infinite number of states. In any case, the transition probabilities p_{ij} can be arranged in the form of a matrix

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ . & . & . & \cdots \\ . & . & . & \\ . & . & . & \cdots \end{pmatrix} \tag{6}$$

if the state space contains $1, 2, \ldots$, or

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ . & . & . & \cdots \\ . & . & . & \cdots \end{pmatrix}$$

if the state space contains nonnegative integers. These matrices are known as *stochastic matrices* with transition probabilities p_{ij} as their elements. The subscripts of each probability are the states associated with a transition from i to j. Given $X_\alpha = i$,

$$\sum_j \Pr\{X_{\alpha+1} = j | X_\alpha = i\} = \sum_j p_{ij} = 1,$$

so that each row sum in a stochastic matrix is unity.

When a system has only two states, denoted by (1,0), they may represent occurrence and nonoccurrence of an event E. In this case, p_{01} is the passage probability to the occurrence of E, and p_{11} is the recurrence probability of event E. The process becomes a **renewal**, or recurrence, process.

Example 2. Gambler's Ruin

The possible states of the system, which range from 0 to $a + b$, represent the amount of money that player A may possess during the course of the game. For $0 < i < a + b$, $p_{i,i+1} = p$ and $p_{i,i-1} = q$; and the game ends at 0 or $a + b$. The $(a + b + 1) \times (a + b + 1)$ transition probability matrix is

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ q & 0 & p & 0 & \ldots & 0 & 0 & 0 \\ 0 & q & 0 & p & \ldots & 0 & 0 & 0 \\ . & . & . & . & \ldots & . & . & . \\ . & . & . & . & \ldots & . & . & . \\ . & . & . & . & \ldots & . & . & . \\ 0 & 0 & 0 & 0 & \ldots & q & 0 & p \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{pmatrix}.$$

Example 3. The Ehrenfest Model of Diffusion

This model has $s+1$ states: $0, 1, 2, \ldots, s$. Transitions are possible only for one step to the right or one step to the left, with the respective probabilities $p_{j,j+1} = 1 - j/s$ and $p_{j,j-1} = j/s$, for $j = 0, 1, \ldots, s$. The transition probability matrix is:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 & 0 \\ \frac{1}{s} & 0 & \frac{1-1}{s} & 0 & \ldots & 0 & 0 \\ 0 & \frac{2}{s} & 0 & \frac{1-2}{s} & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \ldots & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & 0 & \frac{1}{s} \\ 0 & 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix}.$$

This model originally was described by P. & T. Ehrenfest [15] as a conceptual urn experiment where s molecules are distributed in two containers A and B. At each trial a molecule is chosen at random and moved from its container to the other. The state of the system is the number of molecules in A. Feller [17] interpreted the Ehrenfest experiment as a diffusion process with a central force in the sense that the transition probability $p_{j,j+1}$ is greater than or less than 1/2 depending on whether j is less than or greater than 1/2.

Example 4. The Bernoulli–Laplace Model of Diffusion

This model was proposed by D. Bernoulli in 1769 and analyzed by Laplace in 1812 as a probabilistic analog for the flow of two incompressible liquids between two urns. There are $2s$ particles in total, of which s are white and s are black. The system is in state k if there are k white particles in the first urn. At each trial one particle is taken from each urn and they are interchanged. The transition probabilities are:

$$p_{j,j-1} = \left(\frac{j}{s}\right)^2, \qquad p_{j,j} = \frac{2j(s-j)}{s^2},$$

$$p_{j,j+1} = \left(\frac{1-j}{s}\right)^2,$$

so that $p_{j,j-1} + p_{j,j} + p_{j,j+1} = 1$, for $j = 0, 1, \ldots, s$. The corresponding transition probability matrix $\mathbf{P}$ is $(s+1) \times (s+1)$. With the exception of the 0th row and the sth row, the matrix $\mathbf{P}$ has the probabilities $p_{j,j}$ on the diagonal line, $p_{j,j+1}$ on the upper diagonal line, and $p_{j,j-1}$ on the lower diagonal line, and zeros elsewhere. The 0th row is $(0, 1, 0, \ldots, 0)$ and the sth row is $(0, 0, \ldots, 0, 1, 0)$. For details, see [17].

High-Order Transition Probabilities $p_{ij}(n)$

The transition probability p_{ij} defined in (5) is associated with a transition taking place in one step, from $X_\alpha = i$ to $X_{\alpha+1} = j$. When a transition from i to j takes place in n steps, we have an n-step transition probability:

$$\Pr\{X_{\alpha+n} = j | X_\alpha = i\} = p_{ij}(n).$$

The matrix $\mathbf{P}(n)$, with $p_{ij}(n)$ as its elements, is related to $\mathbf{P}(1)$ by

$$\mathbf{P}(n) = [\mathbf{P}(1)]^n.$$

Classification of States

Transition from one state to another is not always possible, depending upon the type of states. The state j is said to be *reachable* from state i if there exists some positive integer n such that the probability $p_{ij}(n) > 0$, and we write $i \to j$. For $n = 0$, we define $p_{ii}(0) = 1$ and $p_{ij}(0) = 0$ for $j \neq i$. If state j is reachable from state i and state i is reachable from state j, the two states are said to be *communicative*, and we write $i \leftrightarrow j$. If state k is reachable from state j and state j is reachable from state i, then state k is reachable from state i. It is clear that the communication relation has the following properties:

1. Reflexivity: $i \leftrightarrow i$, as $p_{ii}(0) = 1$.
2. Symmetry: if $i \leftrightarrow j$, then $j \leftrightarrow i$.
3. Transitivity: if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$.

Therefore, the communication relation is an equivalence relation.

For every two states i and j, $p_{ij}(n)$ is the probability that, starting from state i, the system will enter

state j at the nth step, regardless of the number of entrances into j prior to n. Now we introduce the probability that state j is reached for the first time at the nth step, or the *first passage probability*:

$$f_{ij}(n) = \Pr\{X_n = j \quad \text{and} \quad X_m \neq j; \\ m = 1, \ldots, n-1 | X_0 = i\}.$$

The two types of probability are related as follows:

$$p_{ij}(n) = \sum_{\ell=1}^{n} f_{ij}(\ell) p_{jj}(n-\ell) \quad \text{and}$$

$$f_{ij}(n) = p_{ij}(n) - \sum_{\ell=1}^{n-1} f_{ij}(\ell) p_{jj}(n-\ell).$$

The sum

$$\sum_{n=1}^{\infty} f_{ij}(n) = f_{ij}$$

is the probability that, starting from state i, a system will enter j eventually. If $f_{ij} = 1$, the sequence $\{f_{ij}(n)\}$ is the probability distribution of the first passage time to state j. In this case, the expectation

$$\mu_{ij} = \sum_{n=1}^{\infty} n f_{ij}(n)$$

is the mean passage time from state i to state j.

When $j = i$, we have the first return (recurrence) probability at the nth step,

$$f_{ii}(n) = \Pr\{X_n = i \quad \text{and} \quad X_m \neq i; m = 1, \ldots, \\ n-1 | X_0 = i\}, \quad n = 1, 2, \ldots.$$

The probabilities $p_{ii}(n)$ and $f_{ii}(n)$ are related by the following two formulas:

$$p_{ii}(n) = \sum_{\ell=1}^{n} f_{ii}(\ell) p_{ii}(n-\ell) \quad \text{and}$$

$$f_{ii}(n) = p_{ii}(n) - \sum_{\ell=1}^{n-1} f_{ii}(\ell) p_{ii}(n-\ell).$$

The sum

$$\sum_{n=1}^{\infty} f_{ii}(n) = f_{ii}$$

is the probability of eventual return to the original state i.

Now we introduce various types of state in terms of the passage probabilities and the return probabilities.

1. *Transient state*. A state i is called a transient state if $f_{ii} < 1$. In this case there is a positive probability $1 - f_{ii}$ that, starting from state i, a system will not return to state i in a finite number of steps.
2. *Recurrent state*. A state i is called a recurrent state if $f_{ii} = 1$. In this case, the sequence $\{f_{ii}(n)\}$ represents the probability distribution of the first return (recurrence) time. The expectation
$$\mu_{ii} = \sum_{n=1}^{\infty} n f_{ii}(n)$$
is the mean recurrence time for state i.
3. *Recurrent null state and recurrent nonnull state*. A recurrent state i is called a null state if the expectation $\mu_{ii} = \infty$, and a nonnull state if $\mu_{ii} < \infty$.
4. *Periodic state and aperiodic state*. A state i is periodic with period $t > 1$ if $p_{ii}(n) = 0$ except for $n = t, 2t, \ldots$, where t is the largest integer with this property. In the gambler's ruin problem, the event that a player will break even has a period of $t = 2$: his winnings can only be zero at the nth game for $n = 2, 4, \ldots$. The number of white balls drawn from an urn in the life table example is aperiodic.
5. *Ergodic state*. An aperiodic recurrent state with a finite recurrence time is an ergodic state.
6. *Absorbing state*. A state i is an absorbing state if and only if $f_{ii}(1) = 1$. Clearly, if i is an absorbing state, $f_{ii} = 1, \mu_{ii} = 1$.

The types of state may be defined also in terms of the probabilities $p_{ii}(n)$ and $p_{ij}(n)$. The following theorems in effect further express the relations between $f_{ii}(n)$ and $p_{ii}(n)$ by way of these definitions.

Theorem 1. State j can be reached from state i if and only if $f_{ij} > 0$; states i and j communicate if and only if $f_{ij} f_{ji} > 0$.

Theorem 2. State i is transient if and only if $\sum_{n=0}^{\infty} p_{ii}(n) < \infty$, and is recurrent if the infinite sum diverges.

Theorem 3. If state j is transient, then $\sum_{n=1}^{\infty} p_{ij}(n) < \infty$, and in this case $\lim_{n\to\infty} p_{ij}(n) = 0$.

Closed Sets and Irreducible Markov Chains

Closed Set. A set C of states is closed if for every i in C

$$\sum_{j\in C} p_{ij} = 1,$$

where the summation is taken over all states j belonging to the set C. Generally, for every $n \geq 1$

$$\sum_{j\in C} p_{ij}(n) = 1.$$

Therefore, any closed subset of a system can be studied independently of all other states.

The totality of all states that can be reached from a given state i form a closed set. A closed set may contain states which do not communicate. A closed set of communicating states is a *class*. If C is a class, then for every pair i and j in C, there exists a positive integer n for which $p_{ij}(n) > 0$. An absorbing state is considered a class.

Irreducible Chain. A Markov chain is called an irreducible chain if there exists no closed subset other than the set of all states.

Theorem 4. In an irreducible Markov chain every state can be reached from every other state.

Theorem 5. The states in a class are of the same type; they are either all transient or all recurrent null or all ergodic.

Corollary. The states in an irreducible Markov chain are of the same type.

Theorem 6. All states of a class have the same period.

Corollary. All states in an irreducible Markov chain have the same period.

Ergodic Chain. An irreducible Markov chain with ergodic states is called an ergodic chain.

The above concepts are illustrated by an example.

Example 5

The stochastic matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{3}{8} & \frac{1}{6} & \frac{11}{24} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{3}{8} & \frac{1}{2} & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{2}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

describes transitions in a system of nine states numbered from 1 to 9. We divide these states into four subsets: $C_1 = \{1\}$; $C_2 = \{2, 3\}$, $C_3 = \{4, 5, 6\}$; and $C_4 = \{7, 8, 9\}$. The set C_1 consists of a single absorbing state 1; it is a closed set and a class. In set C_2 states 2 and 3 communicate and both have a period $t = 2$; C_2 is closed and is a class. C_3 is also a closed set but not a class since neither states 5 nor state 6 can be reached from state 4. State 4 is a class and is a proper closed subset of C_3; the subset $\{5, 6\}$ is not a class because it is not closed. Finally, the set C_4 is closed and is a class where states 7, 8, and 9 communicate.

It is clear, then, that a Markov chain corresponding to subset C_2 is an irreducible chain with period $t = 2$, while a chain corresponding to subset C_3 is not an irreducible chain since C_3 contains a proper closed subset; neither is a chain corresponding to subset $\{5, 6\}$ is irreducible chain since the subset is not closed. A Markov chain corresponding to set C_4 obviously is irreducible.

To summarize, we have considered examples of the following types:

1. irreducible periodic chains – C_2;
2. chains which are not irreducible, because
   (i) the corresponding state set contains a proper closed subset – C_3, or
   (ii) the state set is not closed – $\{5, 6\}$;
3. irreducible and aperiodic chains – C_4.

The matrix $\mathbf{P}$ can be decomposed into four submatrices corresponding to these four subsets:

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 & 0 & 0 & 0 \\ 0 & \mathbf{P}_2 & 0 & 0 \\ 0 & 0 & \mathbf{P}_3 & 0 \\ 0 & 0 & 0 & \mathbf{P}_4 \end{pmatrix}, \tag{7}$$

where the zeros are matrices of various dimensions. The zeros in the first row, for example, (from left to right) are 1×2, 1×3, and 1×3 matrices (or row vectors), respectively. The nonzero submatrices are:

$$\mathbf{P}_1 = [1], \quad \mathbf{P}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$\mathbf{P}_3 = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{8} & \frac{1}{6} & \frac{11}{24} \\ \frac{3}{8} & \frac{1}{2} & \frac{1}{8} \end{pmatrix},$$

$$\mathbf{P}_4 = \begin{pmatrix} 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{2}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

each corresponds to a sub-Markov chain.

The matrix on the right-hand side of (7) having submatrices on the diagonal line and zeros elsewhere is known as a *quasi-diagonal matrix*. Direct computations show that the square of the matrix $\mathbf{P}$ is also a quasi-diagonal matrix with the squares of the submatrices on the diagonal line. In general, the nth power of a quasi-diagonal matrix $\mathbf{P}$ is also a quasi-diagonal matrix with the nth power of the submatrices on the diagonal line. In the present example,

$$\mathbf{P}^n = \begin{pmatrix} \mathbf{P}_1^n & 0 & 0 & 0 \\ 0 & \mathbf{P}_2^n & 0 & 0 \\ 0 & 0 & \mathbf{P}_3^n & 0 \\ 0 & 0 & 0 & \mathbf{P}_4^n \end{pmatrix}. \tag{8}$$

Relation (8) reveals interesting properties of transitions of a Markov chain. First, the states in different classes do not communicate: $p_{ik} = 0$ whenever i and k belong to two different classes. Second, for every i and j belonging to the same class C_α, the transition probabilities $p_{ij}(n)$ are computed from the corresponding submatrix only and are independent of the other matrices. This is true even if C_α is not a class (such as C_3) and the corresponding Markov chain is reducible. Therefore, a Markov chain may be studied in terms of individual subchains, each corresponding to a closed set of states.

Asymptotic Behavior of Transition Probabilities $p_{ij}(n)$

We have seen in the preceding section that the transient probabilities $p_{ij}(n)$ behave differently for different types of state. The following theorem describes the limiting probability of $p_{ij}(n)$ as $n \to \infty$.

Theorem 7. If state i is either transient or recurrent null, then $\lim_{n\to\infty} p_{ii}(n) = 0$. If state i is recurrent with period t, then $\lim_{n\to\infty} p_{ii}(n) = t/\mu_{ii}$, where μ_{ii} is the mean recurrent time for state i. If state i is ergodic, then $\lim_{n\to\infty} p_{ii}(n) = 1/\mu_{ii}$.

Theorem 8. If state j is either transient or recurrent null, then for all i $\lim_{n\to\infty} p_{ij}(n) = 0$; if state j is ergodic, then for all i $\lim_{n\to\infty} p_{ij}(n) = 1/\mu_{jj}$.

Stationary Distribution

Definition 2. A probability distribution $\{\pi_i\}$ is called *stationary* for a given Markov chain if it satisfies the relation

$$\pi_j = \sum_i \pi_i p_{ij}.$$

For a stationary distribution, we have, for any integer n,

$$\pi_j = \sum_i \pi_i p_{ij}(n).$$

Theorem 9. If all the states in an irreducible Markov chain are ergodic, then the limits

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j$$

exist and are independent of the initial state i. Furthermore, $\pi_j > 0$,

$$\sum_j \pi_j = 1, \tag{9}$$

and the limiting distribution $\{\pi_j\}$ is stationary so that

$$\sum_i \pi_i p_{ij} = \pi_j. \tag{10}$$

Conversely, if a stationary distribution of an irreducible Markov chain exists and satisfies (9) and (10) then all the states of the Markov chain are ergodic and the stationary distribution is the limiting distribution of the chain.

It is easy to deduce from Theorems 8 and 9 that for each state j in an ergodic Markov chain,

$$\mu_{jj} = \frac{1}{\pi_j}. \tag{11}$$

Formulas for High-Order Transition Probabilities $p_{ij}(n)$

For a finite Markov chain having states $1, 2, \ldots, s$, and a stochastic matrix

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \cdots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{pmatrix}, \tag{6a}$$

with

$$\sum_{j=1}^{s} p_{ij} = 1, \quad i = 1, \ldots, s,$$

we introduce a characteristic matrix

$$\mathbf{A}(\lambda) = (\lambda \mathbf{I} - \mathbf{P}) = \begin{pmatrix} \lambda - p_{11} & -p_{12} & \ldots & -p_{1s} \\ -p_{21} & \lambda - p_{22} & \ldots & -p_{2s} \\ \vdots & \vdots & \ldots & \vdots \\ -p_{s1} & -p_{s2} & \ldots & \lambda - p_{ss} \end{pmatrix}, \tag{12}$$

where $\mathbf{I}$ is an $s \times s$ unit matrix. The determinant $|\mathbf{A}(\lambda)|$ is an s-degree polynomial in λ. The characteristic equation

$$|\mathbf{A}(\lambda)| = 0 \quad \text{or} \quad |\lambda \mathbf{I} - \mathbf{P}| = 0 \tag{13}$$

has s roots: $\lambda_1, \lambda_2, \ldots, \lambda_s$. These are known as the characteristic roots, or **eigenvalues**, of the matrix $\mathbf{P}$. The magnitude of these roots is given in the following theorem.

Theorem 10. The eigenvalues of the stochastic matrix $\mathbf{P}$ are not greater than unity in absolute value and one of the eigenvalue is $\lambda = 1$.

The cofactor of the (i, j)th element of the matrix $\mathbf{A}(\lambda)$, $A_{ij}(\lambda)$, is an $(s-1) \times (s-1)$ determinant after the ith row and the jth column are deleted from $\mathbf{A}(\lambda)$ and multiplied by $(-1)^{i+j}$.

When $\lambda = 1$, the cofactors of the elements in the same row of $\mathbf{A}(1)$ are equal:

$$A_{ij}(1) = A_{ii}(1), \quad i, j = 1, 2, \ldots, s.$$

The adjoint matrix of $\mathbf{A}(\lambda)$ is a matrix whose elements are the cofactors of $\mathbf{A}(\lambda)$, with the indices transposed: the element in the ith row and the jth column is $A_{ji}(\lambda)$, for $i, j = 1, 2, \ldots, s$.

Theorem 11. If the stochastic matrix $\mathbf{P}$ in (6) of a Markov chain has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$, then the nth-order transition probabilities are given by

$$p_{ij}(n) = \sum_{\ell=1}^{s} A_{ji}(\lambda_\ell) \lambda_\ell^n \frac{1}{\prod_{\substack{m=1 \\ m \neq \ell}}^{s} (\lambda_\ell - \lambda_m)}$$
$$i, j = 1, \ldots, s; n = 1, 2, \ldots. \tag{14}$$

Furthermore, the right-hand side of formula (14) is a real function of p_{ij} even if (13) has complex roots.

For the derivation of formula (14) and those in Theorem 11, see [10].

Limiting Probability Distribution

We have shown in Theorem 9 that the limiting probability distribution $\{\pi_j\}$ of an irreducible ergodic Markov chain exists and is stationary. We now provide explicit formulas for the limiting probabilities.

Theorem 12. Let $\mathbf{P}$ defined in (6) be the stochastic matrix of a finite ergodic Markov chain. The limiting probabilities

$$\lim_{n \to \infty} p_{ij}(n) = \pi_j, \quad i, j = 1, \ldots, s,$$

are given by

$$\pi_j = \frac{A_{jj}(1)}{\sum_{k=1}^{s} A_{kk}(1)}, \quad j = 1, \ldots s,$$

and the mean recurrence time by

$$\mu_{jj} = \frac{\sum_{k=1}^{s} A_{kk}(1)}{A_{jj}(1)}, \quad j = 1, \ldots, s,$$

where $A_{jj}(1)$ is the (j, j)th cofactor of the matrix $\mathbf{A}(1) = \mathbf{I} - \mathbf{P}$.

Theorem 12 also confirms the relation between the mean recurrence time μ_{jj} and the limiting probability π_j in formula (11).

Example 6

In a two-state Markov chain with the stochastic matrix:

$$\mathbf{P} = \begin{bmatrix} 1 - p_1 & p_1 \\ p_2 & 1 - p_2 \end{bmatrix},$$

where $0 < p_1 < 1$ and $0 < p_2 < 1$, the equation $|\lambda\mathbf{I} - \mathbf{P}| = 0$ admits two eigenvalues: $\lambda_1 = 1$ and $\lambda_2 = 1 - p_1 - p_2$. The differences between the eigenvalues are $\lambda_1 - \lambda_2 = p_1 + p_2$ and $\lambda_2 - \lambda_1 = -(p_1 + p_2)$, and the corresponding characteristic matrices are

$$\mathbf{A}(\lambda_1) = \begin{bmatrix} p_1 & -p_1 \\ -p_2 & p_2 \end{bmatrix},$$

$$\mathbf{A}(\lambda_2) = \begin{bmatrix} -p_2 & -p_1 \\ -p_2 & -p_1 \end{bmatrix}.$$

Substituting these values in formula (14) yields the final formula:

$$\mathbf{P}(n) = \begin{bmatrix} p_{11}(n) & p_{12}(n) \\ p_{21}(n) & p_{22}(n) \end{bmatrix} = \frac{1}{p_1 + p_2} \times \begin{bmatrix} p_2 + p_1(1 - p_1 - p_2)^n & p_1 - p_1(1 - p_1 - p_2)^n \\ p_2 - p_2(1 - p_1 - p_2)^n & p_1 + p_2(1 - p_1 - p_2)^n \end{bmatrix}.$$

Since $\lambda_1 = 1$, we find from $\mathbf{A}(\lambda_1) = \mathbf{A}(1)$ the limiting probabilities:

$$\pi_1 = \frac{p_2}{p_1 + p_2} \quad \text{and} \quad \pi_2 = \frac{p_1}{p_1 + p_2},$$

which can be obtained from $\mathbf{P}(n)$ as $n \to \infty$. Therefore the mean recurrence times for the two states are

$$\mu_{11} = \frac{p_1 + p_2}{p_2} \quad \text{and} \quad \mu_{22} = \frac{p_1 + p_2}{p_1}.$$

Example 7. An Application in Genetics

To describe the heredity process in a given locus (*see* **Mendel's Laws**) in terms of Markov chain, the number of A genes in the population is denoted by p and the number of a genes by q, with $p + q = 1$. The possible **genotypes**, AA, Aa and aa, are identified with the numbers 1, 2, and 3, respectively. The transition probability p_{ij} denotes the probability that an offspring will have genotype j given that a specific parent has a genotype i. Using a mother–son pair as an example, and when $i = 1$ and $j = 2$,

$$p_{12} = \Pr\{\text{the son will have genotype Aa}|\text{the mother has genotype AA}\} = q.$$

For the son to have the genotype Aa, he must inherit one A gene from his mother with probability one and acquire one a gene from the male population, through his father, with probability q. And this gene selection is the only possibility for the child to have the genotype Aa. Therefore $p_{12} = q$. Using a similar calculation, we find the other transition probabilities, and the following one-step transition probability matrix:

$$\mathbf{P}(1) = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} p & q & 0 \\ \left(\frac{1}{2}\right)p & \frac{1}{2} & \left(\frac{1}{2}\right)q \\ 0 & p & q \end{bmatrix}$$

which is well known in **population genetics**. The matrix $\mathbf{P}(1)$ also shows that the corresponding Markov chain is an irreducible ergodic chain.

To find the n-step transition probabilities $p_{ij}(n)$, we first formulate a characteristic matrix:

$$\mathbf{A}(\lambda) = [\lambda\mathbf{I} - \mathbf{P}] = \begin{bmatrix} \lambda - p & -q & 0 \\ -\left(\frac{1}{2}\right)p & \frac{\lambda - 1}{2} & -\left(\frac{1}{2}\right)q \\ 0 & -p & \lambda - q \end{bmatrix}$$

From $|\mathbf{A}(\lambda)| = 0$, we find the eigenvalues: $\lambda_1 = 1, \lambda_2 = 1/2, \lambda_3 = 0$, so that $(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) =$

$1/2$, $(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3) = -(1/2)^2$ and $(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2) = 1/2$.

For each eigenvalue, we formulate the corresponding characteristic matrix, compute the cofactors, and find the adjoint matrix. For $\lambda_1 = 1$,

$$\mathbf{A}(\lambda_1) = \mathbf{A}(1) = \begin{bmatrix} 1-p & -q & 0 \\ -\left(\frac{1}{2}\right)p & \frac{1-1}{2} & -\left(\frac{1}{2}\right)q \\ 0 & -p & 1-q \end{bmatrix},$$

and the adjoint matrix:

$$\|A_{ji}(1)\| = \begin{bmatrix} \left(\frac{1}{2}\right)p^2 & pq & \left(\frac{1}{2}\right)q^2 \\ \left(\frac{1}{2}\right)p^2 & pq & \left(\frac{1}{2}\right)q^2 \\ \left(\frac{1}{2}\right)p^2 & pq & \left(\frac{1}{2}\right)q^2 \end{bmatrix},$$

For $\lambda_2 = \dfrac{1}{2}$,

$$\mathbf{A}(\lambda_2) = \mathbf{A}\left(\frac{1}{2}\right) = \begin{bmatrix} \left(\frac{1}{2}\right)-p & -q & 0 \\ -\left(\frac{1}{2}\right)p & 0 & -\left(\frac{1}{2}\right)q \\ 0 & -p & \left(\frac{1}{2}\right)-q \end{bmatrix},$$

and the adjoint matrix:

$$\|A_{ji}\left(\frac{1}{2}\right)\| = \begin{bmatrix} -\left(\frac{1}{2}\right)pq & \left(\frac{1}{2}\right)q(p-q) & \left(\frac{1}{2}\right)q^2 \\ -\left(\frac{1}{2}\right)^2 p(p-q) & -\left(\frac{1}{2}\right)^2 (p-q) & -\left(\frac{1}{2}\right)^2 q(p-q) \\ \left(\frac{1}{2}\right)p^2 & -\left(\frac{1}{2}\right)p(p-q) & -\left(\frac{1}{2}\right)pq \end{bmatrix}.$$

In this problem, $\lambda_3 = 0$, so there is no contribution to $p_{ij}(n)$.

Now substituting the eigenvalues, the cofactors, and the products of the differences of eigenvalues in formula (14) for $p_{ij}(n)$, we find

$$\mathbf{P}(n) = \begin{bmatrix} p^2 + \frac{pq}{2}^{n-1} & 2pq + q(q-p)/2^{n-1} & q^2 - q^2/2^{n-1} \\ p^2 + p(q-p)/2^n & 2pq + (p-q)^2/2^n & q^2 + q(p-q)/2^n \\ p^2 - p^2/2^{n-1} & 2pq + p(p-q)/2^{n-1} & q^2 + pq/2^{n-1} \end{bmatrix}.$$

This formula has appeared in [17] and [24] from different approaches.

As $n \to \infty$, the limiting transition probability matrix is:

$$\lim_{n\to\infty} \mathbf{P}(n) = \begin{bmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{bmatrix}$$

and the limiting probability distribution

$$(\pi_1 \quad \pi_2 \quad \pi_3) = (p^2 \quad 2pq \quad q^2),$$

is stationary, since

$$(p^2\ 2pq\ q^2)\begin{bmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{bmatrix} = (p^2\ 2pq\ q^2).$$

The mean recurrence times are $\mu_{11} = 1/p^2$, $\mu_{22} = 1/2_{pq}$, and $\mu_{33} = 1/q^2$.

A General Birth Process

In this general birth process, the time scale t is continuous. For every $t \in [0, \infty)$, the random variable $X(t)$ is defined as the number of births occurring at or before time t, so $X(t)$ takes on discrete values. The term "the number of births" is used here quite loosely. It refers to the number of "events" occurring up to t, whatever the "event" may happen to be in a particular study. The "event" may be an accident, an incoming telephone call, a new species in a genus, a new infected case during an epidemic, or even a death. If a birth process is regarded as an "increasing" process and a death process as a "decreasing" process, the general birth process described here applies to both increasing processes and decreasing processes.

As the number of events occurring increases with time, the distribution of $X(t)$ is a increasing function of time. The purpose in this section is to review the general formula for the transition probability from which a desired transition probability for a particular process can be derived.

Let

$$p_{i,k}(0,t) = \Pr[X(t) = k | X(0) = i\}$$

be the transition probability from $X(0) = i$ at time $t = 0$ to $X(t) = k$ at time t, for $k = i, i+1, \ldots$. Given $X(t) = k$, let the probability of having a birth during $(t, t+\Delta)$ be $\lambda_k \beta(t)\Delta + o(\Delta)$. The product $\lambda_k \beta(t)$ is the birth intensity function, which plays a major role in determining the process. Here $\beta(t)$ is an integrable function of t such that the integral

$$\int_0^t \beta(\tau)\,\mathrm{d}\tau \to \infty \quad \text{as} \quad t \to \infty,$$

and λ_j is an arbitrary function of j subject to the condition that $\lambda_i \neq \lambda_j$, for $i \neq j$. Under these conditions, the transition probabilities satisfy the following differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,i}(0,t) = -\lambda_i \beta(t) p_{i,i}(0,t) \tag{15a}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,k}(0,t) = -\lambda_k \beta(t) p_{i,k}(0,t) + \lambda_{k-1}\beta(t) p_{i,k-1}(0,t), \tag{15b}$$

for $k = i, i+1, \ldots$. Eq. (15a) has the solution

$$p_{i,i}(0,t) = \exp\left[-\lambda_i \int_0^t \beta(\tau)\,\mathrm{d}\tau\right]. \tag{16}$$

Using (16) and solving (15b) successively beginning with $k = i+1$ yields the general formula:

$$p_{i,k}(0,t) = (-1)^{k-i}\lambda_i \ldots \lambda_{k-1} \times \left[\sum_{j=1}^{k} \frac{\exp\left[-\lambda_j \int_0^t \beta(\tau)\,\mathrm{d}\tau\right]}{\prod_{\substack{\ell=i \\ \ell\neq j}}^{k} (\lambda_j - \lambda_\ell)}\right], \tag{17}$$

for $k = i, i+1, \ldots$. An inductive proof of (17) is given in [10].

Now we use (17) to deduce the formulas of the transition probability $p_{i,k}(0,t)$ in some well-known processes.

Yule Processes

The intensity functions are:

$$\lambda_j = j\lambda \quad \text{and} \quad \beta(t) = 1$$

(*see* **Yule Process**). Substituting these values in formula (17) yields:

$$p_{i,k}(0,t) = (-1)^{k-i}[i\lambda]\ldots[(k-1)\lambda] \times \left[\sum_{j=1}^{k} \left(\frac{\exp(-j\lambda t)}{\prod_{\substack{\ell=i \\ \ell\neq j}}^{k} (j\lambda - \ell\lambda)}\right)\right], \tag{18}$$

where

$$[i\lambda]\ldots[(k-1)\lambda] = \lambda^{k-i}\binom{k-1}{k-i}(k-i)!$$

and

$$\prod_{\substack{\ell=i \\ \ell\neq j}}^{k} (j\lambda - \ell\lambda) = \lambda^{k-i}(-1)^{k-j}\binom{k-i}{j-i}^{-1}(k-i)!.$$

Therefore (18) can be rewritten as

$$\begin{aligned} p_{i,k}(0,t) &= \binom{k-1}{k-i}\exp(-i\lambda t) \times \left[\sum_{j=i}^{k}\binom{k-i}{j-i}\exp(-\lambda t)^{j-i}\right] \\ &= \binom{k-1}{k-i}\exp(-i\lambda t)[1-\exp(-\lambda t)]^{k-i}, \end{aligned}$$

for $k = i, i+1, \ldots$, a familiar formula in the Yule process.

If the intensity function $\beta(t)$ remains a function of t, then we have the time-dependent Yule process:

$$p_{i,k}(0,t) = \binom{k-1}{k-i}\left\{\exp\left[-\lambda\int_0^t \beta(\tau)\,d\tau\right]\right\}^i \times\left\{1-\exp\left[-\lambda\int_0^t \beta(\tau)\,d\tau\right]\right\}^{k-i}$$

for $k = i, i+1, \ldots$.

Pólya Process

The **Pólya process** usually records the events occurring during the interval $(0, t]$, so the initial value at $t = 0$ is $X(0) = 0$. The intensity function is:

$$\lambda_k = (1+\lambda k) \quad \text{and} \quad \beta(t) = (1+\lambda t)^{-1}.$$

Using these intensity functions and with reference to formula (17) for $p_{0,k}(0,t)$, we evaluate

$$\exp\left[-(1+j\lambda)\int_0^t (1+\lambda\tau)^{-1}\,d\tau\right] = (1+\lambda t)^{-(1/\lambda+j)},$$

$$(1+\lambda)(1+2\lambda)\ldots[1+(k-1)\lambda] = \binom{\frac{1}{\lambda}+k-1}{k}\lambda^k k!,$$

and

$$\prod_{\substack{\ell=0\\ \ell\neq j}}^{k}[(1+j\lambda)-(1+\ell\lambda)] = \lambda^k\binom{k}{j}^{-1}k!(-1)^{k-j},$$

and substitute these formulas in (17). As a result,

$$p_{0,k}(0,t) = \binom{\frac{l}{\lambda}+k-1}{k}\sum_{j=0}^{k}\binom{k}{j} \times(-1)^j(1+\lambda t)^{-(1/\lambda+j)} = \binom{\frac{l}{\lambda}+k-1}{k}(1+\lambda t)^{-1/\lambda}\left(\frac{\lambda t}{1+\lambda t}\right)^k$$

for $k = 0, 1, \ldots$, which is the formula of the transition probability in the Pólya process.

Birth Process with Immigration

Another plausible function for λ_j is linear:

$$\lambda_j = j\lambda+\eta \quad \text{and} \quad \beta(t) = 1.$$

Here the linear term $j\lambda$ corresponds to birth and the constant term η corresponds to immigration (*see* **Migration Processes**). With reference to (17), we compute

$$[i\lambda+\eta][(i+1)\lambda+\eta]\ldots[(k-1)\lambda+\eta] = \lambda^{k-i}\binom{k+\left(\frac{\eta}{\lambda}\right)-1}{k-i}(k-i)!$$

and

$$\prod_{\substack{\ell=i\\ \ell\neq j}}^{k}[(j\lambda+\eta)-(\ell\lambda+\eta)] = \lambda^{k-1}(-1)^{k-j}\binom{k-i}{k-j}^{-1}(k-i)!$$

As a result,

$$p_{i,k}(0,t) = \binom{k+\left(\frac{\eta}{\lambda}\right)-1}{k-i}[\exp(-\lambda t)]^{i+\left(\frac{\eta}{\lambda}\right)} \times[1-\exp(-\lambda t)]^{k-i}, \tag{19}$$

for $k = i, i+1, \ldots$. If $\beta(t)$ remains a function of time t, the transition probability $p_{i,k}(0,t)$ will be the same as in formula (19) but with t being replaced by the integral $\int_0^t \beta(\tau)\,d\tau$.

From (19) we see that the difference $X(t) - i$ has a **negative binomial** distribution with parameters $i + \eta/\lambda$ and $\exp(-\lambda t)$. (*see* **Accident Proneness**).

Death Process

Let i individuals alive at $t = 0$ be subject to the same force of mortality, or **hazard**, $\mu(t)$, and let the random variable $X(t)$ be the number of deaths occurring during the interval $(0, t]$, with $X(0) = 0$. The transition probability is defined by

$$p_{0,k}(0,t) = \Pr\{X(t) = k | X(0) = 0\}.$$

Under the assumption of independence of mortality, the intensity functions are

$$\lambda_j = i - j \quad \text{and} \quad \beta(t) = \mu(t).$$

Substituting these functions in (17), we find

$$\exp\left[-(i-j)\int_0^t \mu(\tau)\,d\tau\right] = \left\{\exp\left[-\int_0^t \mu(\tau)\,d\tau\right]\right\}^{i-k} \times \left\{\exp\left[-\int_0^t \mu(\tau)\,d\tau\right]\right\}^{k-j},$$

$$(i-0)(i-1)\dots(i-k+1) = \binom{i}{k} k!$$

and

$$\prod_{\substack{\ell=0\\ \ell=j}}^{k} [(i-j)-(i-\ell)] = (-1)^j \binom{k}{j}^{-1} k!.$$

Therefore,

$$p_{0,k}(0,t) = \binom{i}{k}\left\{\exp\left[-\int_0^t \mu(\tau)\,d\tau\right]\right\}^{i-k} \times \left\{1-\exp\left[-\int_0^t \mu(\tau)\,d\tau\right]\right\}^{k}$$

for $k = 0, 1, \dots, i$, which is the binomial distribution, with the exponential function being the survival function (*see* **Survival Distributions and Their Characteristics**).

A Divergent Process

This process was mentioned in [17]; in it $\lambda_j = j^2$ and $\beta(t)$ is a function of t. The probability derived from (17) is

$$p_{i,k}(0,t) = \binom{k-1}{k-i}^2 \sum_{j=i}^{k} (-1)^{j-i} \times \binom{k-i}{k-j}^2 \binom{2j-1}{j-1}^{-1} \binom{k+j}{k-j}^{-1} \times \exp\left[-j^2\lambda \int_0^t \beta(\tau)\,d\tau\right],$$

$k = i, i+1, \dots$. According to this model, the rate of population growth is proportional to the square of the population size j within each time element $(t, t+\Delta)$, and for every given t there is a positive probability,

$$1 - \sum_{k=i} p_{i,k}(0,t) > 0,$$

that the population will become infinitely large. This model may not be an accurate description of human population growth, but it may be applicable to the growth of microorganisms or to nuclear fission.

Poisson Process

While formula (17) holds for cases where $\lambda_i \neq \lambda_j$, for $i \neq j$, the transition probability in the **Poisson process** may be derived from the general birth process as a limiting case as $\lambda_j \to \lambda$. That is,

$$\lim_{\lambda_j \to \lambda} (-1)^k \lambda_0 \dots \lambda_{k-1} \sum_{j=0}^{k} \frac{\exp(-\lambda_j t)}{\prod_{\substack{\ell=0\\ \ell\neq j}}^{k} (\lambda_j - \lambda_\ell)} = \frac{\exp(-\lambda t)(\lambda t)^k}{k!}.$$

A description of the limiting process may be found in [10, pp. 255–256].

An Equality in Stochastic Processes

Transition probabilities $p_{i,k}(0,t)$ are the basic elements in stochastic processes. Explicit formulas for the probabilities are needed not only for an appreciation of stochastic processes as an analytic tool, but also for a better understanding of the problems on hand through analyses of the data. Generally, efforts are made to obtain explicit solutions, but these are not always successful. The equality presented in this section is a general property of (increasing) stochastic processes, but it is useful for derivation of explicit formulas when other methods have failed.

Let $X(t)$ be the number of "births" up to time t, with birth intensity function $\lambda_j\beta(\tau)$ and transition probabilities

$$p_{i,k}(0,t) = \Pr\{X(t) = k | X(0) = i\}, \qquad (20)$$

as defined in the general birth process. Let j be an arbitrary but *fixed* integer between i and k and

consider the transition $j \to j+1$. Instead of a simple transition from i to k, there is now a sequence of transitions: $i \to j \to j+1 \to k$. The transition $j \to j+1$ must take place somewhere between 0 and t; let it take place in $(\tau, \tau + d\tau)$. The probability for the sequence $i \to j \to j+1 \to k$ is

$$p_{i,j}(0,\tau)\lambda_j\beta(\tau)\,d\tau p_{j+1,k}(\tau,t).$$

For different values of τ, the corresponding sequences of transitions $i \to j \to j+1 \to k$ are mutually exclusive; the integral

$$\int_0^t p_{i,j}(0,\tau)\lambda_j\beta(\tau)p_{j+1,k}(\tau,t)\,d\tau \qquad (21)$$

is the probability that the sequence $i \to j \to j+1 \to k$ will occur during the interval $(0,t)$. Since the process must admit state j and the transition $j \to j+1$ before arriving at the value k at time t, the integral in (21) is equal to the expression in (20). That is,

$$p_{i,k}(0,t) = \int_0^t p_{i,j}(0,\tau)\lambda_j\beta(\tau)p_{j+1,k}(\tau,t)\,d\tau. \qquad (22)$$

Formula (22) was proposed in [8]; a formal proof using the Riemann integral is given in [10]. It is easy to verify that the Yule processes and others in the section on the general birth process all satisfy (22).

Example 8. The Poisson process with $i = 0$, $\lambda_j\beta(\tau) = \lambda$, has transition probability

$$p_{0,k}(0,t) = \frac{\exp(-\lambda t)(\lambda t)^k}{k!},$$

which is the left-hand side of (22). The right-hand side is

$$\begin{aligned}
&\int_0^t \left[\frac{\exp(-\lambda\tau)(\lambda\tau)^j}{j!}\right]\lambda\,d\tau \\
&\quad\times\left[\frac{\exp\{-\lambda(t-\tau)\}[\lambda(t-\tau)]^{k-j-1}}{(k-j-1)!}\right] \\
&\quad= \frac{\exp(-\lambda t)(\lambda t)^k}{j!(k-j-1)!}\int_0^1 \theta^j(1-\theta)^{k-j-1}\,d\theta \\
&\quad= p_{0,k}(0,t),
\end{aligned}$$

where $\theta = \tau/t$, and the beta function in the last integral is equal to $j!(k-j-1)!/k!$, justifying the last equality.

Equality (22) can be extended to any number of intermediate values between $X(0)$ and $X(t)$. Suppose there are two intermediate fixed values j and k, such that $i < j < k < l$; the transition probability

$$\Pr\{X(t) = l | X(0) = i\} = p_{i,l}(0,t)$$

satisfies the equality

$$\begin{aligned}
p_{i,l}(0,t) = \int_0^t \int_0^{\tau_2} &p_{ij}(0,\tau_1)\lambda_j(\tau_1)p_{j+1,k}(\tau_1,\tau_2) \\
&\times \lambda_k(\tau_2)p_{k+1,l}(\tau_2,t)\,d\tau_1\,d\tau_2.
\end{aligned}$$

An Application: Simple Stochastic Epidemic

In a simple stochastic model, a population consists of two categories of individuals: susceptibles and infectives (*see* **Epidemic Models, Stochastic**). There are no removals, no deaths, no immunes, and no recoveries from infection. Suppose that at the initial time $t = 0$, there are N susceptibles and 1 infective. Let $X(t)$ be the number of infectives at time t, so that there are $N + 1 - X(t)$ susceptibles. The primary purpose is to derive an explicit formula for the transition probability

$$p_{1,n}(0,t) = \Pr\{X(t) = n | X(0) = 1\}. \qquad (20a)$$

Under the assumption of homogeneous mixing of infectives and susceptibles, the intensity functions in the general birth process are:

$$\lambda_j = j(N+1-j) = a_j$$

and

$$\int_0^t \beta(\tau)\,d\tau = \theta(t),$$

where $\beta(t)$ is known as the *infection rate* and is a function of t, the "age" of an epidemic, and $\theta(t)$ tends to infinity as $t \to \infty$.

For $n < (N+1)/2$, $a_j \neq a_i$, for $i \neq j$, formula (17) in the general birth process applies. This means

that the solution is

$$p_{1,n}(0,t) = (-1)^{n-1} a_1 \dots a_{n-1} \times \left[\sum_{i=1}^{n} \left(\frac{\exp[-a_i \theta(t)]}{\prod_{\substack{\alpha=1 \\ \alpha \neq i}}^{n} (a_i - a_\alpha)} \right) \right], \tag{17a}$$

for $n = 1, 2, \dots, [(N+1)/2]$.

Formula (17) no longer applies for $n > (N+1)/2$, since in this case, $a_1, a_2, \dots, a_n$ are not all distinct. In particular,

$$a_j = j(N+1-j) = a_{N+1-j}.$$

However, an explicit formula for the probability $p_{1,n}(0,t)$ can be obtained by applying formula (22):

$$p_{1,n}(0,t) = \int_0^t p_{1,k}(0,\tau) a_k \beta(\tau) p_{k+1,n}(\tau,t)\, \mathrm{d}\tau. \tag{22a}$$

The integer k must be so chosen that the a_js in the probability $p_{i,k}(0,\tau)$ are distinct and the a_js in $p_{k+1,n}(\tau,t)$ are also distinct. When N is even, $k = N/2$; when N is odd, $k = (N+1)/2$. With these values of k, we apply (17) to the two probabilities $p_{1,k}(0,\tau)$ and $p_{k+1,n}(\tau,t)$ to obtain

$$p_{1,k}(0,\tau) = (-1)^{k-1} a_1 \dots a_{k-1} \times \left[\sum_{i=1}^{k} \left(\frac{\exp[-a_i \theta(\tau)]}{\prod_{\substack{\alpha=1 \\ \alpha \neq i}}^{k} (a_i - a_\alpha)} \right) \right] \tag{17b}$$

and

$$p_{k+1,n}(\tau,t) = (-1)^{n-k-1} a_{k+1} \dots a_{n-1} \times \left[\sum_{j=k+1}^{n} \left(\frac{\exp\{-a_j[\theta(t) - \theta(\tau)]\}}{\prod_{\substack{\beta=k+1 \\ \beta \neq j}}^{n} (a_j - a_\beta)} \right) \right]. \tag{17c}$$

Substituting (17b) and (17c) in (22a) and simplifying the resulting expression, we find.

$$p_{1,n}(0,t) = (-1)^{n-1} a_1 \dots a_{n-1} \times \left[-\sum_{j=k+1}^{n} \left(\frac{\theta(t) \exp[-a_j \theta(t)]}{\prod_{\substack{\beta=1 \\ a_\beta \neq a_j}}^{n} (a_j - a_\beta)} \right) + \sum_{\substack{i=1 \\ a_i \neq a_j}}^{k} \sum_{j=k+1}^{n} \left(\frac{\exp[-a_i \theta(t)] - \exp[-a_j \theta(t)]}{(a_i - a_j) \prod_{\substack{\alpha=1 \\ \alpha \neq i}}^{k} (a_i - a_\alpha) \prod_{\substack{\beta=k+1 \\ \beta \neq j}}^{n} (a_j - a_\beta)} \right) \right],$$

for $n = k+1, k+2, \dots, N+1$, where $k = N/2$ when N is even and $k = (N+1)/2$ when N is odd.

Infection Time and the Duration of an Epidemic

The length of time elapsed up to the occurrence of the nth infection is a continuous random variable taking nonnegative real values. Let it be denoted by T_n, for $1 \le n \le N+1$, with $T_1 = 0$. The duration of an epidemic is T_{N+1}, the length of time elapsed up to the infection of the last number of the population. The purpose of this section is to derive explicit formulas for the density function $f_n(t)$, the distribution $F_n(t)$, and the expectation and variance of T_n.

The density function $f_n(t)$ has a close relationship with the probability $p_{1,n}(t)$. By definition, $f_n(t)\,\mathrm{d}t$ is the probability that the random variable T_n will take values in $(t, t+\mathrm{d}t)$. This means that at time t there are $n-1$ infectives, and the nth infection takes place in the interval $(t, t+\mathrm{d}t)$. Therefore,

$$f_n(t)\,\mathrm{d}t = p_{1,n-1}(0,t) a_{n-1} \beta(t)\,\mathrm{d}t,$$

and the distribution function

$$F_n(t) = \int_0^t p_{1,n-1}(0,\tau) a_{n-1} \beta(\tau)\,\mathrm{d}\tau, \quad \text{for } n = 2, \dots, N+1.$$

Using the formulas for the probabilities in the preceding section, we can write down explicit functions for $f_n(t)$ and $F_n(t)$ for each n. For example, for

$n \le (N+1)/2$, we use formula (17a) to obtain the density function

$$f_n(t)\,\mathrm{d}t = (-1)^{n-2}a_1 \ldots a_{n-1} \times \left[\sum_{i=1}^{n-1} \left(\frac{\beta(t)\exp[-a_i\theta(t)]}{\prod_{\substack{\alpha=1 \\ \alpha\neq i}}^{n-1}(a_i - a_\alpha)} \right) \right] \mathrm{d}t$$

and the distribution function

$$F_n(t) = (-1)^{n-2}a_1 \ldots a_{n-1} \times \sum_{i=1}^{n-1} \frac{1-\exp[-a_i\theta(t)]}{\prod_{\substack{\alpha=1 \\ \alpha\neq i}}^{n-1}(a_i - a_\alpha)a_i}.$$

As t approaches infinity,

$$\lim_{t\to\infty}\theta(t) = \lim_{t\to\infty}\int_0^t \beta(\tau)\,\mathrm{d}\tau = \infty,$$

$$F_n(\infty) = (-1)^{n-2}a_1 \ldots a_{n-1} \times \sum_{i=1}^{n-1} \frac{1}{\prod_{\substack{\alpha=1 \\ \alpha\neq i}}^{n-1}(a_i - a_\alpha)a_i} = 1.$$

A proof of $F_n(\infty) = 1$, for $1 < n \le N+1$, is given in [10].

The expectation and variance of T_n can be computed directly from the definitions

$$\mathrm{E}(T_n) = \int_0^\infty t f_n(t)\,\mathrm{d}t$$

and

$$\mathrm{var}(T_n) = \int_0^\infty [t - \mathrm{E}(T_n)]^2 f_n(t)\,\mathrm{d}t.$$

Explicit formulas depend on the function $\beta(t)$. However, when the infection rate is independent of time, with $\beta(t) = \beta$, there is an alternative approach which is simpler.

The length of time elapsed until the occurrence of the nth infection may be divided into two periods: a period of length T_{n-1} up to the occurrence of the $(n-1)$th infection; and a period of length t_n between the occurrences of the $(n-1)$th and nth infections. The sum of the two periods is equal to the entire length of time:

$$T_n = T_{n-1} + t_n, \tag{23}$$

where T_{n-1} and t_n are independently distributed non-negative random variables, with respective density functions

$$f_{n-1}(t) = p_{1,n-2}(0,t)a_{n-2}\beta$$

and

$$g_n(t) = p_{n-1,n-1}(0,t)a_{n-1}\beta. \tag{24}$$

According to (23), the distribution of T_n is the convolution of the distributions of T_{n-1} and t_n, and the density functions satisfy the equation

$$f_n(t) = \int_0^t f_{n-1}(\tau)g_n(t-\tau)\,\mathrm{d}\tau.$$

Now, the density function in (24) is exponential,

$$g_n(t) = a_{n-1}\beta\exp\{-a_{n-1}\beta t)\},$$

with expectation and variance

$$\mathrm{E}(t_n) = \frac{1}{a_{n-1}\beta} \quad \text{and} \quad \sigma_{t_n}^2 = \frac{1}{a_{n-1}^2\beta^2}.$$

However, (23) can be extended so that

$$T_n = t_2 + t_3 + \ldots + t_n,$$

where $t_2, t_3, \ldots, t_n$ are independently distributed **exponential** random variables. It follows that the expectation and the variance of T_n are, respectively,

$$\mathrm{E}(T_n) = \frac{1}{\beta}\sum_{i=1}^{n-1}\frac{1}{a_i}$$

and

$$\sigma_{T_n}^2 = \frac{1}{\beta^2}\sum_{i=1}^{n-1}\frac{1}{a_i^2}, \quad n = 2, \ldots, N+1.$$

This epidemic model was originally formulated by Kermack and McKendrick [21, 26]. It has since been extensively studied; see, for example, [3], [4], [5] and [20]. An important feature of this model is that the coefficients a_n are quadratic functions of n, and the differential equations of the transition

probabilities are of second order and cannot be easily solved. The Laplace transform does not provide an explicit formula for the transition probability either. The present method, relying on the general birth process formula (17) and the equality (22), is based on [29].

Death Processes and Survival Distributions

In the death process deduced from the general birth process in the preceding section, the time interval $[0, t)$ was fixed and the number of deaths occurring in $[0, t)$ was a random variable. In this section, the survival time of an individual is treated as a random variable and the purpose is to derive its distribution.

Let a random variable T be the survival time, or lifetime, of an individual with mortality intensity function, or force of mortality, $\mu(t)$, so that, the sum $\mu(t)\Delta + o(\Delta)$ is the probability that an individual alive at time t will die in the time interval $(t, t + \Delta)$. The distribution function of T at t, $F_T(t) = \Pr\{T \leq t\}$, is the probability that an individual will die at or before time t. The complement $1 - F_T(t)$ is the survival function of T at t.

Consider now the distribution function of T at time $t + \Delta$, $F_T(t + \Delta) = \Pr\{T \leq t + \Delta\}$. For an individual to die before $t + \Delta$, either he dies before t, or else he must survive to t and die during the interval $(t, t + \Delta)$. Symbolically,

$$F_T(t + \Delta) = F_T(t) + [1 - F_T(t)][\mu(t)\Delta + o(\Delta)],$$

which leads to the differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} F_T(t) = [1 - F_T(t)]\mu(t)$$

or

$$\frac{\mathrm{d}}{\mathrm{d}t} \ln[1 - F_T(t)] = -\mu(t). \tag{25}$$

Clearly at $t = 0$, $F_T(0) = 0$. The solution of the differential equation (25) is:

$$1 - F_T(t) = \exp\left[-\int_0^t \mu(\tau)\,\mathrm{d}\tau\right]$$

and

$$F_T(t) = 1 - \exp\left[-\int_0^t \mu(\tau)\,\mathrm{d}\tau\right]. \tag{26}$$

The survival function and the distribution function in (26) are the same as those in the death process in the section on the general birth process. The density function of T is

$$f(t) = \mu(t) \exp\left[-\int_0^t \mu(\tau)\,\mathrm{d}\tau\right]. \tag{27}$$

Formulas (26) and (27) both depend on $\mu(t)$, and are the basic functions in a survival analysis. One can derive formulas for a particular survival distribution by making assumptions about the mortality intensity function $\mu(t)$. The following are a few examples.

Gompertz Distribution

In a celebrated paper on the law of human mortality, Benjamin Gompertz [18] attributed death to either one of two causes: chance or deterioration of the power to withstand destruction (*see* **Aging Models**). In deriving his law of mortality, however, Gompertz considered only deterioration, and assumed that a person's power to resist death decreases at a rate proportional to the power itself. Since the force of mortality $\mu(t)$ is a measure of a person's susceptibility to death, Gompertz used the reciprocal $1/\mu(t)$ as a measure of a person's resistance to death and thus arrived at the formula:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\mu(t)}\right) = -h\left(\frac{1}{\mu(t)}\right)$$

or

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu(t) = -h\mu(t),$$

from which he found the force of mortality $\mu(t) = Bc^t$. The distribution function and the density function are, respectively,

$$F_T(t) = 1 - \exp\left\{\frac{-B(c^t - 1)}{\ln c}\right\}.$$

and

$$f_T(t) = Bc^t \exp\left\{\frac{-B(c^t - 1)}{\ln c}\right\}.$$

Makeham Distribution

Makeham [25] suggested the modification $\mu(t) = A + Bc^t$ to restore the missing component "chance" in the Gompertz formula (*see* **Aging Models**). The corresponding distribution function is

$$F_T(t) = 1 - \exp\left\{-\left[At + \frac{B(c^t - 1)}{\ln c}\right]\right\}$$

and the density function is

$$f_T(t) = (A + Bc^t) \exp\left\{-\left[At + \frac{B(c^t - 1)}{\ln c}\right]\right\}.$$

Weibull Distribution

When the force of mortality is assumed a power function of t, $\mu(t) = ct^{c-1}$, the distribution function and the density function are:

$$F_T(t) = 1 - \exp\{-t^c\}$$

and

$$f_T(t) = ct^{c-1} \exp(-t^c).$$

This distribution, proposed by Weibull in 1939 for studies of the lifespan of materials, is used frequently in survival analysis (*see* **Weibull Distribution**).

Exponential Distribution

If $\mu(t) = \mu$ is a constant, then the distribution function and the density function are:

$$F_T(t) = 1 - \mathrm{e}^{-\mu t}$$

and

$$f_T(t) = \mu \mathrm{e}^{-\mu t},$$

which were used for illustration of the theory of life testing [16] (*see* **Exponential Distribution**).

A Staging Process and Stages of Disease

Development of many chronic conditions is characterized by stages. Generally, diseases advance with time from a mild stage through intermediate and severe stages to death. The process often is irreversible, but a patient may die while being in any of the stages. In the natural progression of cancer, for example, there are stages of the disease determined by the size of tumor and metastasis of cancer. **AIDS**, too, can be classified by stages.

Birth order and child spacing are another example of a staging process. Here the process begins when the couple decides to start a family; stages are defined by the parities of the woman, from parity zero (no children) to parity one (one child), to parity two (two children), and so on. The process is clearly irreversible, and it terminates when the couple decides to stop reproducing. We can find staging phenomena in many other areas, such as metamorphosis in biology, foraging processes in wildlife, and cascade processes in nuclear physics. Since this process was originally proposed in [9] for statistical studies of chronic illnesses, we shall use chronic diseases as an example for illustration.

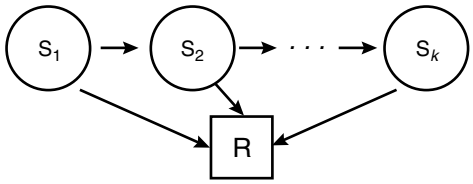

Figure 1 The stages of a disease

Denote the stages of a disease by $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_k$, and the death state by R. We can describe the disease process schematically as in Figure 1. The arrows indicate the directions in which that transitions take place. From each stage S_i, for $i = 1, 2, \ldots, k-1$, the disease process may enter the next stage S_{i+1}, or enter the death state R. From the final stage S_k, the process enters the death state R. We shall derive the distribution function of the survival time T of an individual who is in stage S_1 at the initial time $t = 0$.

The following identities are needed in deriving the formulas for the distribution function, expectation and variance of the survival time T. For proofs of these identities, see [9] and [10].

Lemma 1. For distinct numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\sum_{i=1}^{n} \frac{1}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (\lambda_i - \lambda_j)} = 0, \tag{A}$$

$$\sum_{i=1}^{n} \frac{1}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (\lambda_i - \lambda_j)\lambda_i} = (-1)^{n-1} \frac{1}{\prod_{i=1}^{n} \lambda_i}, \tag{B}$$

$$\sum_{i=1}^{n} \frac{1}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (\lambda_i - \lambda_j)\lambda_i^2} = (-1)^{n-1} \frac{1}{\prod_{i=1}^{n} \lambda_i} \left(\sum_{i=1}^{n} \frac{1}{\lambda_i}\right), \tag{C}$$

and

$$\sum_{i=1}^{n} \frac{1}{\prod_{\substack{j=1 \\ j\neq i}}^{n} (\lambda_i - \lambda_j)\lambda_i^3} = (-1)^{n-1} \frac{1}{\prod_{i=1}^{n} \lambda_i} \times \left(\sum_{j\geq i}^{n} \sum_{i=1}^{n} \frac{1}{\lambda_i \lambda_j} \right). \quad \text{(D)}$$

Intensity Functions

For an individual in stage S_i a time $\tau, 0 \leq \tau < \infty$, let

$$\nu_{i,i+1}\beta(\tau)\,d\tau = \Pr\{\text{the individual will enter stage } S_{i+1} \text{ in } (\tau, \tau + d\tau)\},$$

and

$$\mu_i \beta(\tau)\,d\tau = \Pr\{\text{the individual will enter R in } (\tau, \tau + d\tau)\}$$

and let

$$\nu_{ii}\beta(\tau) = -[\nu_{i,i+1}\beta(\tau) + \mu_i\beta(\tau)], \quad i = 1, \ldots, k-1.$$

For an individual in the final stage S_k at time τ, we let

$$\mu_k \beta(\tau)\,d\tau = \Pr\{\text{the individual will enter R in } (\tau, \tau + d\tau)\}$$

and

$$\nu_{kk}\beta(\tau) = -\mu_k\beta(\tau).$$

The function $\beta(t)$ is such that

$$\theta(t) = \int_0^t \beta(\tau)\,d\tau$$

tends to infinity as $t \to \infty$.

While diseases develop continuously, the time of transition from one stage to the next follows a definite order. Suppose the transition from stage S_i to stage S_{i+1} takes place during the time interval $(\tau_i, \tau_i + d\tau_i)$, for $i = 1, 2, \ldots, k-1$, and that $0 < \tau_1 < \tau_2 < \ldots < \tau_{k-1}$.

Density Function of Survival Time T

When death occurs during the interval $(t, t + dt)$, the individual must be in one of the states $S_1, S_2, \ldots, S_k$ at time t. By definition, $f_T(t)\,dt$ is the probability that the individual who is in stage S_1 time $t = 0$ will die in $(t, t + dt)$. Since he may enter the death state R from any one of the k states $S_1, S_2, \ldots, S_k$, the product $f_T(t)\,dt$ is the sum of k terms:

$$f_T(t)\,dt = f_1(t)\,dt + f_2(t)\,dt + f_3(t)\,dt + \cdots + f_k(t)\,dt. \quad (28)$$

Each $f_j(t)\,dt$ in (28) corresponds to the sequence of transitions. $S_1 \to S_2 \to \ldots \to S_j \to R$. The first term $f_1(t)\,dt$, for example, is the probability of transition $S_1 \to R$ occurring in $(t, t + dt)$,

$$f_1(t)\,dt = \exp\{\nu_{11}\theta(t)\}\mu_1\beta(t)\,dt. \quad (29)$$

The function $f_2(t)\,dt$ represents the sequence of transitions $S_1 \to S_2 \to R$. For the transition $S_1 \to S_2$ to take place during a particular interval $(\tau_1, \tau_1 + d\tau_1)$, the probability of the sequence $S_1 \to S_2 \to R$ is

$$\exp\left\{\nu_{11}\int_0^{\tau_1} \beta(\tau)\,d\tau\right\} \nu_{12}\beta(\tau_1)\,d\tau_1 \times \exp\left\{\nu_{22}\int_{\tau_1}^{t} \beta(\tau)\,d\tau\right\} \mu_2\beta(t)\,dt. \quad (30)$$

Integrating (30) from $\tau_1 = 0$ to $\tau_1 = t$ yields

$$f_2(t)\,dt = \nu_{12}\mu_2\beta(t)\left[\frac{1}{\nu_{11} - \nu_{22}} \exp\{\nu_{11}\theta(t)\} + \frac{1}{\nu_{22} - \nu_{11}} \exp\{\nu_{22}\theta(t)\}\right] dt. \quad (31)$$

Generally, for the sequence $S_1 \to S_2 \to \cdots \to S_j \to R$,

$$f_j(t)\,dt = \nu_{12}\ldots\nu_{j-1,j}\mu_j\beta(t) \sum_{i=1}^{j} \frac{1}{\prod_{\substack{\ell=1 \\ \ell\neq i}}^{j} (\nu_{ii} - \nu_{\ell\ell})} \times \exp\{\nu_{ii}\theta(t)\}\,dt, \quad j = 2, \ldots, k. \quad (32)$$

Substituting (29), (31) and (32) in (28) gives the density function of the survival time T:

$$f_T(t) = \exp\{\nu_{11}\theta(t)\}\mu_1\beta(t) + \sum_{j=2}^{k}\left\{\nu_{12}\ldots\nu_{j-1,j}\mu_j\beta(t) \times \sum_{i=1}^{j}\left(\exp[\nu_{ii}\theta(t)] \Big/ \prod_{\substack{\ell=1\\ \ell\neq i}}^{j}(\nu_{ii}-\nu_{\ell\ell})\right)\right\}. \quad (33)$$

Expectation and Variance of Survival Time T

When $\beta(t) = 1$, $\theta(t) = t$. In this case the density function of T is

$$f_T(t) = \mu_1\exp\{\nu_{11}t\} + \sum_{j=2}^{k}\left\{\nu_{12}\ldots\nu_{j-1,j}\mu_j \times \sum_{i=1}^{j}\left(\exp\{\nu_{ii}t\} \Big/ \prod_{\substack{\ell=1\\ \ell\neq i}}^{j}(\nu_{ii}-\nu_{\ell\ell})\right)\right\}, \quad (34)$$

and the distribution function is

$$F_T(t) = \frac{\mu_1}{\nu_{11}}(\exp\{\nu_{11}t\}-1) + \sum_{j=2}^{k}\left\{\nu_{12}\ldots\nu_{j-1,j}\mu_j \times \sum_{i=1}^{j}\left((\exp\{\nu_{ii}t\}-1) \Big/ \prod_{\substack{\ell=1\\ \ell\neq i}}^{j}(\nu_{ii}-\nu_{\ell\ell})\nu_{ii}\right)\right\}.$$

As $t \to \infty$, the distribution function $F_T(\infty) = 1$, which can be proven using (B) in Lemma 1. Therefore, the distribution is proper.

Using the density function in (34) we find the formula for the expectation $\mathrm{E}(T)$, and then using formula (C) in the lemma we simplify the formula to the following:

$$\mathrm{E}[T] = \frac{1}{-\nu_{11}} + \frac{\nu_{12}}{-\nu_{11}}\left(\frac{1}{-\nu_{22}}\right) + \frac{\nu_{12}}{-\nu_{11}}\frac{\nu_{23}}{-\nu_{22}}\left(\frac{1}{-\nu_{33}}\right) + \cdots + \prod_{i=1}^{k-1}\frac{\nu_{i,i+1}}{-\nu_{ii}}\left(\frac{1}{-\nu_{kk}}\right). \quad (35)$$

Each factor

$$\frac{\nu_{i,i+1}}{-\nu_{ii}} = \frac{\nu_{i,i+1}}{\nu_{i,i+1}+\mu_i}.$$

in (35) is the conditional probability of transition $\mathrm{S}_i \to \mathrm{S}_{i+1}$ given that a transition out of S_i takes place, while the factor $(-1/\nu_{i+1,i+1})$ is the expected duration of stay in stage S_{i+1}. In other words, the expectation $\mathrm{E}(T)$ in (35) is the sum of expected durations of stay in $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_k$, as it should be.

The variance of the survival time T is:

$$\mathrm{var}(T) = 2\left[\left(\frac{\mu_1}{-\nu_{11}}\right)\frac{1}{\nu_{11}^2} + \frac{\nu_{12}\mu_2}{\nu_{11}\nu_{22}}\sum_{j\geq i}^{2}\sum_{i=1}^{2}\frac{1}{\nu_{ii}\nu_{jj}} + \cdots + (-1)^k\left(\prod_{i=1}^{k-1}\frac{\nu_{i,i+1}}{\nu_{ii}}\right) \times \frac{\mu_k}{\nu_{kk}}\sum_{j\geq i}^{k}\sum_{i=1}^{k}\frac{1}{\nu_{ii}\nu_{jj}}\right] - [\mathrm{E}(T)]^2.$$

The staging process described in this section is a special case of the *illness – death process*, in which transitions between states of illness may be reversible. Another particular case is the **Fix–Neyman process**.

Birth–Death Processes

The stochastic processes presented so far in this article have been either increasing processes like the Yule processes, or decreasing processes like the death processes. We now consider processes that allow a population to grow as well as to decline, which are more relevant to the biological and the human populations in which both births and

deaths occur. The concept of birth–death processes existed before the theory of stochastic processes. Kendall [19] was among the first to seek a general solution of the processes. Methods of solving the differential equations have since been revised and refined. In this section, explicit solutions for two cases are presented.

Again, we let $X(t)$ be the population size at time t, for $0 \le t < \infty$, and the transition probability

$$p_{i,k}(0,t) = \Pr\{X(t) = k | X(0) = i\}, \quad \text{for } k = 0, 1, \ldots.$$

Given $X(t) = k$, let $\lambda_k(t)$ and $\mu_k(t)$ be the birth intensity and death intensity functions, respectively. The transition probabilities satisfy the following differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,0}(0,t) = -[\lambda_0(t) + \mu_0(t)] p_{i,0}(0,t) + \mu_1(t) p_{i,1}(0,t), \tag{36a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,k}(0,t) = -[\lambda_k(t) + \mu_k(t)] p_{i,k}(0,t) + \lambda_{k-1}(t) p_{i,k-1}(t) + \mu_{k+1}(t) p_{i,k+1}(t). \tag{36b}$$

The system of differential equations in (36) and the initial conditions,

$$p_{i,i}(0,0) = 1 \quad \text{and} \quad p_{i,k}(0,0) = 0 \text{ for } k \ne i,$$

completely determine the probability distribution $\{p_{i,k}(0,t)\}$. Since the differential equations are dependent on the intensity functions $\lambda_k(t)$ and $\mu_k(t)$, two specific forms of the intensity functions are assumed in the following discussion.

Linear Growth

Suppose both $\lambda_k(t)$ and $\mu_k(t)$ are independent of time but proportional to k,

$$\lambda_k(t) = k\lambda \quad \text{and} \quad \mu_k(t) = k\mu,$$

where λ and μ are constant. In this case, the differential equations (36) become:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,0}(0,t) = \mu p_{i,1}(0,t) \tag{37a}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,k}(0,t) = -k(\lambda+\mu) p_{i,k}(0,t) + (k-1)\lambda p_{i,k-1}(0,t) + (k+1)\mu p_{i,k+1}(0,t), \quad k = 1, 2, \ldots. \tag{37b}$$

Each of differential equations in (37b) has three unknown probabilities, $p_{i,k-1}(0,t)$, $p_{i,k}(0,t)$, and $p_{i,k+1}(0,t)$, and cannot be solved with the methods used in the general birth process. We resort to the method of probability generating functions (pgfs). The pgf of $X(t)$ is defined by

$$G_X(s;t) = \sum_{k=0}^{\infty} s^k p_{i,k}(0,t),$$

which is a polynomial in s, the coefficient of s^k being the transition probability $p_{i,k}(0,t)$, for $k = 0, 1, \ldots$. Therefore, when an explicit formula for the pgf is derived, one can obtain the desired transition probability by identifying the coefficient of the corresponding s^k.

In addition to the probabilities, the pgf also generates the expectation and the variance of $X(t)$.

Using the differential equations (37) we find that the generating function satisfies the partial differential equation,

$$\frac{\partial}{\partial t} G_X(s;t) + (1-s)(\lambda s - \mu)\frac{\partial}{\partial s} G_X(s;t) = 0,$$

with the initial condition at $t = 0$ given by $G_{X(0)}(s;0) = s^i$. The solution of the partial differential equation is:

$$G_X(s;t) = \left\{ \frac{\alpha(t) + [1 - \alpha(t) - \beta(t)]s}{1 - \beta(t)s} \right\}^i, \tag{38}$$

where

$$\alpha(t) = \mu \frac{1 - \exp\{(\lambda-\mu)t\}}{\mu - \lambda \exp\{(\lambda-\mu)t\}} \text{ and } \beta(t) = \frac{\lambda}{\mu}\alpha(t).$$

Now it is a simple matter of expanding the pgf as a polynomial in s, and identifying the coefficient of s^k to obtain the formula for the probability $p_{i,k}\ (0,t)$.

The numerator on the right-hand side of (38) is a binomial function:

$$\{\alpha(t) + [1 - \alpha(t) - \beta(t)]s\}^i = \sum_{j=0}^{i} \binom{i}{j} [\alpha(t)]^{i-j} [1 - \alpha(t) - \beta(t)]^j s^j.$$

For the denominator, since clearly $|\beta(t)s| < 1$,

$$\{1 - \beta(t)s\}^{-i} = \sum_{j=0}^{\infty} \binom{-i}{j} (-1)^j [\beta(t)]^j s^j = \sum_{j=0}^{\infty} \binom{i+j-1}{j} [\beta(t)]^j s^j.$$

Hence the probability

$$p_{i,k}(0,t) = \sum_{j=0}^{\min[i,k]} \binom{i}{j} \binom{i+k-j-1}{k-j} [\alpha(t)]^{i-j} \times [\beta(t)]^{k-j} [1 - \alpha(t) - \beta(t)]^j, \qquad (39)$$

for $k = 1, 2, \ldots$. For $k = 0$,

$$p_{i,0}(0,t) = [\alpha(t)]^i$$

is the probability that the population will become extinct at time t.

When $i = 1$, formula (38) is the pgf of the population size at time t when the initial population is $X(0) = 1$. This means that the random variable $X(t)$ discussed above is the sum of i independent and identically distributed random variables, each having the pgf defined in (38) with $i = 1$. In other words, the i populations reproduce and perish independent of each other, but their growth is subject to the same probability law $\{p_{1,k}(0,t)\}$.

Differentiating the pgf in (38) with respect to s yields the expectation of $X(t)$,

$$\mathrm{E}[X(t)] = i \exp\{(\lambda - \mu)t\},$$

and the variance of $X(t)$,

$$\sigma^2_{X(t)} = i \left(\frac{\lambda + \mu}{\lambda - \mu}\right) \exp\{(\lambda - \mu)t\} \times [\exp\{(\lambda - \mu)t\} - 1].$$

Two terms had been suggested for the birth–death processes, depending upon the relative values of the birth parameter λ and the death parameter μ. If $\lambda > \mu$, then the birth–death process is called supercritical; if $\lambda < \mu$, then the process is subcritical.

When $\lambda = \mu$, the probabilities are

$$p_{i,k}(0,t) = \sum_{j=0}^{\min[i,k]} \binom{i}{j} \binom{i+k-j-1}{k-j} (\lambda t)^{i+k-2j} \times (1 - \lambda t)^j (1 + \lambda t)^{-i-k+j}, \quad k > 1,$$

and

$$p_{i,0}(0,t) = \left\{\frac{\lambda t}{1 + \lambda t}\right\}^i.$$

The expectation and the variance of $X(t)$ are

$$\mathrm{E}[X(t)] = i \quad \text{and} \quad \sigma^2_{X(t)} = 2i\lambda t.$$

Thus, when the birth rate λ is equal to the death rate μ, the population size has a constant expectation but an increasing variance with time t.

As t approaches infinity, the limiting behavior of the birth–death process depends on the relative values of the birth rate λ and the death rate μ. The following are the asymptotic values of the probability generating function.

$$\lim_{t\to\infty} G_{X(t)}(s,t) = \begin{cases} 1, & \text{if} \quad \lambda \le \mu, \\ \left(\frac{\mu}{\lambda}\right)^i, & \text{if} \quad \lambda > \mu. \end{cases} \qquad (40)$$

Now when $s = 0$, $G_X(0;t) = p_{i,0}(0,t)$ is the probability that the population will become extinct at time t. According to (40), if the birth rate λ is smaller than or equal to the death rate μ, then the probability of extinction tends to unity as $t \to \infty$, and the population is certain to die out eventually. On the other hand, if the birth rate λ is greater than the death rate μ the probability of ultimate extinction is $(\mu/\lambda)^i$. Furthermore, since the limiting pgf is a constant, the population will either die out with probability $(\mu/\lambda)^i$, or increase without bound with probability $1 - (\mu/\lambda)^i$; no intermediate course is possible.

The relative values of λ and μ also influence the asymptotic values of the expectation and the variance of $X(t)$. As t approaches infinity,

$$\lim_{t\to\infty} \mathrm{E}[X(t)] = \begin{cases} 0, & \text{if} \quad \lambda < \mu, \\ i, & \text{if} \quad \lambda = \mu, \\ \infty, & \text{if} \quad \lambda > \mu, \end{cases}$$

and

$$\lim_{t\to\infty} \text{var}[X(t)] = \begin{cases} 0, & \text{if } \lambda < \mu, \\ \infty, & \text{if } \lambda \geq \mu. \end{cases}$$

When $\lambda = \mu$, we have an interesting case in which the probability of extinction tends to unity, yet the expected population size tends to i. These seemingly contradictory facts may be intuitively explained by the large value of the variance. Although most populations will eventually become extinct, a few will attain huge sizes, so that the average size will be i.

Time-Dependent Birth–Death Process

We may generalize the birth–death process presented above by letting both the birth intensity function $\lambda(t)$ and the death intensity function $\mu(t)$ be functions of time t. In this case the differential equations for the transition probability $p_{i,k}(0,t)$ assume the same general form as those in formulas (37a) and (37b) with the functions $\lambda(t)$ and $\mu(t)$ replacing the constants λ and μ, respectively.

Again we derive the formulas for the transition probabilities $p_{i,k}(0,t)$ by way of the probability generating function, Which turns out to be

$$G_X(s;t) = \left\{ \frac{\alpha(t) + [1 - \alpha(t) - \beta(t)]s}{1 - \beta(t)s} \right\}^i, \quad (38a)$$

where

$$\alpha(t) = 1 - \frac{1}{\exp[\gamma(t)] + \displaystyle\int_0^t \lambda(\tau)\exp[\gamma(\tau)]\,d\tau},$$

$$\beta(t) = 1 - \exp[\gamma(t)][1 - \alpha(t)],$$

$$\gamma(t) = -\int_0^t [\lambda(\tau) - \mu(\tau)]\,d\tau.$$

Except for the definitions of $\alpha(t)$ and $\beta(t)$, the pgf is of the same general form as in (38). Consequently, the formulas for the transition probabilities $p_{i,k}(0,t)$ assume the same form as those in (39).

The expectation of $X(t)$ is

$$\text{E}[X(t)] = i \exp\left\{ \int_0^t [\lambda(\tau) - \mu(\tau)]\,d\tau \right\}.$$

Thus $\text{E}[X(t)] \to \infty$ if the integral diverges, $\text{E}[X(t)] \to 0$ if the integral approaches minus infinity, and $\text{E}[X(t)] \to i$ if $\lambda(\tau) = \mu(\tau)$, whatever the value of τ, $0 \leq \tau < \infty$.

The probability of population extinction can be obtained directly from the pgf by setting $s = 0$,

$$p_{i,0}(0,t) = \left\{ \frac{\displaystyle\int_0^t \mu(\tau)\exp[\gamma(\tau)]\,d\tau}{1 + \displaystyle\int_0^t \mu(\tau)\exp[\gamma(\tau)]\,d\tau} \right\}^i,$$

which approaches unity if and only if the integral

$$\int_0^t \mu(\tau)\exp\left\{ -\int_0^\tau [\lambda(\xi) - \mu(\xi)]\,d\xi \right\} d\tau$$

diverges as $t \to \infty$. Obviously, the divergence occurs if and only if $\mu(\tau) > \lambda(\tau)$ for every $\tau > 0$. This conclusion is consistent with that reached in the consideration of $\text{E}[X(t)] \to 0$.

Finite Markov Processes

In a finite Markov process, a system has a finite or denumerable number of states: $1, 2, \ldots$. The state of a system at time t is identified by the value of a discrete random variable $X(t)$. "The system is in state j at time t" is the same as "$X(t) = j$". In the birth–death processes, for example, the state of the system at time t was the population size at t, and was identified by the value $X(t) = k$. But there was a rather restrictive assumption that, within a small time interval $(t, t + \Delta t)$, the population size may increase, or decrease, by only one. We now remove this restriction and allow a system to move from any state in the system to any other state in the system. And we assume that the set of states is closed and contains no proper closed subset in the set but itself. For a time interval (τ, t), for $\tau < t$, $\tau, t \in [0, \infty)$, we let

$$p_{ij}(\tau, t) = \Pr\{X(t) = j | X(\tau) = i\}, \quad \text{for } i, j = 1, 2, \ldots \quad (41)$$

be the transition (conditional) probabilities, with sum $\sum_j p_{ij}(0,t) = 1$, for every i. Formula (41) shows the stochastic dependence of $X(t)$ on $X(\tau)$. Two important forms of dependence are defined below:

Definition 3. A discrete-valued stochastic process $\{X(t) : t \in [0, \infty)\}$ is a Markov process if,

for any $t_0 < t_1 < \ldots t_i \ldots < t_j$ and any integers $k_0, k_1, \ldots, k_i, \ldots, k_j$,

$$\Pr\{X(t_j) = k_j | X(t_0) = k_0, X(t_1) = k_1, \ldots, X(t_i) = k_i\} = \Pr\{X(t_j) = k_j | X(t_i) = k_i\}. \quad (42)$$

Thus, in a Markov process, given $X(t_i)$ (present), the conditional probability of $X(t_j)$ (future) is independent of $X(t_0), \ldots, X(t_{i-1})$ (past).

Definition 4. A Markov process $\{X(t); t \in [0, \infty)\}$ is homogeneous with respect to time, or time-homogeneous, if the transition probability in (41) depends only the difference $t - \tau$ and not on τ or t separately. In such a case we may write

$$\Pr\{X(t) = j | X(\tau) = i\} = P_{ij}(0, t - \tau\}, \quad i, j = 1, 2, \ldots. \quad (43)$$

The simple Poisson process is an example of a time-homogeneous process.

Chapman–Kolmogorov Equations

Let ξ be a fixed point in the interval (τ, t), so that $\tau < \xi < t$, and let $X(\tau)$, $X(\xi)$, and $X(t)$ be the corresponding random variables. According to the assumption in (42),

$$\Pr\{X(t) = k | X(\tau) = i \text{ and } X(\xi) = j\} = \Pr\{X(t) = k | X(\xi) = j\} = P_{jk}(\xi, t)$$

and

$$\Pr\{X(\xi) = j \text{ and } X(t) = k | X(\tau) = i\} = P_{ij}(\tau, \xi) p_{jk}(\xi, t). \quad (44)$$

Formula (44) is the probability of a passage from $X(\tau) = i$ to $X(t) = k$ by way of a particular state j at time ξ. At time ξ, the system must be in one of the states $[1, 2, \ldots]$, and the set of the states is closed, $Pr\{X(\xi) = 1 \text{ or } X(\xi) = 2 \text{ or } \ldots\} = 1$. Therefore,

$$p_{ik}(\tau, t) = \sum_j p_{ij}(\tau, \xi) p_{jk}(\xi, t) \quad i, j, k = 1, 2, \ldots, \tau < \xi < t, \quad (45)$$

which is known as the *Chapman–Kolmogorov equation*. In the case of a time-homogeneous process, (45), may be replaced by

$$p_{ik}(0, \tau + t) = \sum_j p_{ij}(0, \tau) p_{jk}(0, t), \quad i, j, k = 1, 2, \ldots. \quad (46)$$

Kolmogorov Differential Equations

Kolmogorov [22] derived two systems of differential equations for the transition probabilities $p_{ij}(\tau, t)$: the *forward differential equations*, where the differentiation of $p_{ij}(\tau, t)$ is taken with respect to t; and the *backward differential equations*, where the differentiation is taken with respect to τ. When the transition probabilities $p_{ij}(\tau, t)$ satisfy certain regularity conditions, both systems may be derived from the Chapman–Kolmogorov equation (45). Following Kolmogorov, Chung [12], Doob [14], Feller [17], and others have discussed in detail theoretical aspects of these differential equations. Feller, for example, has shown that, if $\sum_j p_{ij}(\tau, t) = 1$, then there always exists a unique solution $p_{ij}(\tau, t)$ that satisfies both the forward and the backward differential equations. In this article we shall present explicit solutions for the individual transition probabilities $p_{ij}(0, t)$ for the time-homogeneous Markov processes. Reference may be made to [7] and [11].

Let the transition intensity functions be defined as follows:

$$\nu_{ij} \Delta + o(\Delta) = \Pr\{X(t + \Delta) = j | X(t) = i\}, \quad \text{for } j \neq i; i, j = 1, 2, \ldots, s,$$

and

$$\nu_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{s} \nu_{ij}, \quad i = 1, 2, \ldots, s.$$

Denote the intensity function matrix $||\nu_{ij}||$ by $\mathbf{V}$.

In the time-homogeneous Markov processes, the forward Kolmogorov differential equations are

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{ik}(0, t) = \sum_{j=1}^{s} p_{ij}(0, t) \nu_{jk}, \quad (47)$$

and the backward differential equations are

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{ik}(0, t) = \sum_{j=1}^{s} \nu_{ij} p_{jk}(0, t), \quad (48)$$

with common initial condition

$$p_{ik}(0,0) = \delta_{ik}; \tag{49}$$

δ_{ik} is the Kronecker delta, i.e. $\delta_{ik} = 1$ if $k = i$ and $\delta_{ik} = 0$ if $k \neq i$.

Formulas for Transition Probabilities $p_{ij}(0, t)$

Derivation of the formulas for the transition probabilities requires the following result:

Lemma 2. For distinct numbers $\rho_1, \rho_2, \ldots, \rho_s$,

$$\sum_{i=1}^{s} \frac{\rho_i^r}{\prod_{\substack{j=1 \\ j\neq i}}^{s} (\rho_i - \rho_j)} = \begin{cases} 0, & \text{for } 0 \leq r < s-1, \quad (50) \\ 1, & \text{for } r = s-1. \quad (51) \end{cases}$$

See [6] for proof of (50) and (51).

Formulas (47) are first-order ordinary differential equations with constant coefficients; the solution should be exponential functions of t, $p_{ij}(0, t) = c_{ij} \exp\{\rho t\}$. Substituting the suggested solution in (47) and canceling out the nonvanishing exponential functions yields a system of $s \times s$ simultaneous homogeneous equations in c_{ij}. In order for the system to have nontrivial solution for c_{ij}, the matrix of the coefficients must be zero. That is, $\mathbf{A}'(\rho) = |\mathbf{I}\rho - \mathbf{V}'| = 0$, with roots $\rho_1, \rho_2, \ldots \rho_s$. This means that these roots are the only values of ρ for which the suggested solution is a valid solution for $p_{ij}(0, t)$.

For each root, say $\rho = \rho_\ell$, there is be a system of $s \times s$ simultaneous homogeneous equations for the unknown $c_{ij\ell}$. Since the simultaneous equations are homogeneous, $c_{ij\ell}$ are proportional to the cofactors of the matrix $\mathbf{A}'(\rho_\ell)$, or $c_{ij\ell} = k_{i\ell} A'_{ij}(\rho_\ell)$, for each ℓ. When the roots $\rho_1, \rho_2, \ldots, \rho_s$ are distinct, the general solution of the differential equation (47) is the following sum:

$$p_{ij}(0, t) = \sum_{\ell=1}^{s} k_{i\ell} A'_{ij}(\ell) e^{\rho_\ell t}, \; i, j = 1, \ldots, s. \tag{52}$$

At the initial time $t = 0$, $p_{ii}(0, 0) = 1$ and $p_{ij}(0, 0) = 0$. These initial conditions impose restrictions on $k_{i\ell}$ in (52). Expanding each cofactor $A'_{ij}(\ell)$ as a polynomial in ρ_ℓ, and using the identities (50) and (51) in the lemma, we find

$$k_{i\ell} = \left[\prod_{\substack{m=1 \\ m\neq\ell}}^{s} (\rho_\ell - \rho_m) \right]^{-1}.$$

Substituting these values of k_i in (52) yields the solution:

$$p_{ij}(0, t) = \sum_{\ell=1}^{s} \frac{A'_{ij}(\rho_\ell) \exp\{\rho_\ell t\}}{\prod_{\substack{m=1 \\ m\neq\ell}}^{s} (\rho_\ell - \rho_m)}, \quad i, j = 1, \ldots, s. \tag{53}$$

The limiting probabilities can be obtained from (53), as t approaches infinity:

$$\lim_{t\to\infty} p_{ij}(0, t) = \frac{V_{jj}}{\sum_{\ell=1}^{s} V_{\ell\ell}}, \quad i, j = 1, \ldots, s, \tag{54}$$

where V_{jj} are the principal minors of matrix $\mathbf{V}$.

The proof of this result is as follows: Since $\rho_k < 0$ for $k = 2, \ldots, s$, $\exp\{\rho_k t\}$ tends to zero as $t \to \infty$. From (53),

$$\lim_{t\to\infty} p_{ij}(0, t) = \frac{A'_{ij}(\rho_1)}{\prod_{m=2}^{s} (\rho_1 - \rho_m)}. \tag{55}$$

Here $\mathbf{A}'(\rho_1) = \rho_1 \mathbf{I} - \mathbf{V}'$ and $\rho_1 = 0$; therefore,

$$A'_{ij}(\rho_1) = (-1)^{s-1} V'_{ij} = (-1)^{s-1} V_{ji} = (-1)^{s-1} V_{jj}, \tag{56}$$

since $V_{ji} = V_{jj}$. For the denominator in (55), we write

$$(-1)^{s-1} \rho_2 \ldots \rho_s = (-1)^{s-1} [V_{11} + \ldots + V_{ss}]. \tag{57}$$

Substituting (56) and (57) in (55) yields (54).

Solution of Backward Differential Equations

Formula (53) for the transition probability also satisfies the backward Kolmogorov differential equations (48). Substituting (53) in (48) yields the equation

$$\sum_{\ell=1}^{s} \frac{A'_{ik}(\rho_\ell)}{\prod_{\substack{m=1\\m\neq l}}^{s}(\rho_\ell - \rho_m)} \rho_\ell \exp\{\rho_\ell t\}$$

$$= \sum_{j=1}^{s} \nu_{ij} \sum_{\ell=1}^{s} \frac{A'_{jk}(\rho_\ell)\exp\{\rho_\ell t\}}{\prod_{\substack{m=1\\m\neq \ell}}^{s}(\rho_\ell - \rho_m)}, \qquad (58)$$

which holds if, for each ℓ,

$$\rho_\ell A'_{ik}(\rho_\ell) = \sum_{j=1}^{s} \nu_{ij} A'_{jk}(\rho_\ell)$$

or

$$\rho_\ell A_{ki}(\rho_\ell) - \sum_{j=1}^{s} \nu_{ij} A_{kj}(\rho_\ell) = 0. \qquad (59)$$

The left-hand side of (59) is an expansion of the determinant $|\mathbf{A}(\rho_\ell)|$ using the kth row cofactors and the ith row elements. It is equal to zero for $i \neq k$, and is equal to the determinant $|\mathbf{A}(\rho_\ell)| = 0$ for $i = k$, since ρ_ℓ is a root of the equation $|\mathbf{A}(\rho)| = 0$. Therefore, equation (59) is true. This implies (58), which means that (48) is satisfied.

A Time-Dependent Markov Process

When the transition intensity is a product of two functions, one being a function of the states involved in the transition and the other a function of the time at which the transition takes place, we have a time-dependent Markov process. Specifically,

$$\Pr\{X(t+\Delta) = j | X(t) = i\} = \nu_{ij}\beta(t)\Delta + o(\Delta).$$

The corresponding differential equations are

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{ik}(0,t) = \sum_{j=1}^{s} p_{ij}(0,t)\nu_{jk}\beta(t),$$

$$i, k = 1, 2, \ldots, s,$$

and the formulas for the transition probabilities are

$$p_{ij}(0,t) = \sum_{\ell=1}^{s} \frac{A'_{ij}(\rho_\ell)\exp[\rho_\ell \int_0^t \beta(\tau)\,\mathrm{d}\tau]}{\prod_{\substack{m=1\\m\neq\ell}}^{s}(\rho_\ell - \rho_m)}$$

for $i, j = 1, 2, \ldots, s$, where $A'_{ij}(\rho_\ell)$ and ρ_ℓ have the same meaning as those in (53).

References

[1] Adler, R.J. (1986). Random fields, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 508–512.

[2] Armitage, P. (1975). *Sequential Medical Trials*, 2nd Ed. Blackwell Science, Oxford.

[3] Bailey, N.T.J. (1963). The simple stochastic epidemic: a complete solution in terms of known functions, *Biometrika* **50**, 235–240.

[4] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases*. Griffin, London.

[5] Bartlett, M.S. (1956). Deterministic and stochastic models of recurrent epidemics, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, J. Neyman, ed. University of California Press, Berkeley, pp. 81–109.

[6] Chiang, C.L. (1964). A stochastic model of competing risks of illness and competing risks of death, in *Stochastic Models in Medicine and Biology*, J. Gurland, ed. University of Wisconsin Press, Madison, pp. 323–351.

[7] Chiang, C.L. (1973). A solution of Kolmogorov differential equations – a preliminary report, *Bulletin of the International Statistical Institute* **45**, 264–270.

[8] Chiang, C.L. (1974). An equality in stochastic processes and its applications, in *Progress in Statistics*, Vol. I, J. Gani, K. Sarkadi & I. Vincze, eds. North-Holland, Amsterdam, pp. 145–151.

[9] Chiang, C.L. (1979). Survival and stages of disease, *Mathematical Biosciences* **43**, 159–171.

[10] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Krieger, New York.

[11] Chiang, C.L. & Raman, S. (1973). On a solution of Kolmogorov differential equations, in *Proceedings of the Fourth Conference on Probability Theory*. Editura Academici, Bucharest. pp. 129–136.

[12] Chung, K.L. (1960). *Markov Chains with Stationary Transition Probabilities*. Springer-Verlag, Berlin.

[13] Doob, J. (1942). Topics in the theory of Markov chains, *Transactions of the American Mathematical Society* **52**, 37–64.

[14] Doob, J. (1953). *Stochastic Processes*. Wiley, New York.

[15] Ehrenfest, P. & Ehrenfest, T. (1907). Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem, *Physikalische Zeitschrift* **8**, 311–413.

[16] Epstein, B. & Sobel, M. (1953). Life testing, *Journal of the American Statistical Association* **48**, 486–502.

[17] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd Ed. Wiley, New York.

[18] Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality; and on the new mode of determining the value of life contingencies, *Philosophical Transactions of the Royal Society* **115**, 513–525.

[19] Kendall, D.G. (1948). On the generalized birth-and-death process, *Annals of Mathematical Statistics* **19**, 1–15.

[20] Kendall, D.G. (1957). La propagation d'une épidémie ou d'un bruit dans une population limitée. *Publications de l'Institut de Statistique de l'Université de Paris* **6**, 307–311.

[21] Kermack, W.O. & McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics I, *Proceedings of the Royal Society, Series A* **115**, 700–721.

[22] Kolmogorov, A.M. (1931). Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung, *Mathematische Annalen*, **104**, 415–458.

[23] Kolmogorov, A.M. (1937). Markov chains with a countable number of possible states [in Russian], *Bulletin de l'Université d'État à Moscou, Section A* **1**, 1–15.

[24] Li, C.C. (1968). *Population Genetics*. University of Chicago Press, Chicago.

[25] Makeham, W.M. (1860). On the law of mortality and the construction of annuity tables, *Journal of the Institute of Actuaries* **8**.

[26] McKendrick, A.G. (1926). Application of mathematics to medical problems, *Proceedings of the Edinburgh Mathematical Society* **44**.

[27] Uspansky, J.V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York.

[28] Wald, A. (1947). *Sequential Analysis*, Wiley, New York.

[29] Yang, G. & Chiang, C.L. (1971). A time dependent simple stochastic epidemic, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, J. Neyman, ed. University of California Press, Berkeley, pp. 147–158.

(*See also* **Epidemic Models, Spatial**; **Galton–Watson Process**; **Hidden Markov Models**; **Point Processes**; **Probability Theory**; **Queuing Processes**; **Semi-Markov Processes**)

CHIN LONG CHIANG

Stocks, Percy

Born: November 5, 1899.
Died: December 18, 1974.

After qualification in medicine and public health, Stocks joined the public health service in Bristol, before moving to **Karl Pearson's** department at University College London (UCL) in 1921, where, in 1926, he became Reader in Medical Statistics. He became Chief Medical Statistician at the General Register Office (GRO) from 1933 until retirement in 1950. At UCL and at the GRO he conducted epidemiologic studies, and at the GRO he played a leading part in revising the **International Classification of Diseases**. After retirement, he was a research fellow in cancer until 1957, and published studies on cancer epidemiology. In an obituary notice in the ***Journal of the Royal Statistical Society***, **A. Bradford Hill** wrote "Percy Stocks will be remembered, nationally and internationally, as one of the great contributors to the development and use of medical statistics".

PETER ARMITAGE

Stratification

Stratification refers in epidemiology to a design that improves the efficiency of analytical procedures to control for **confounding** by causing **controls** to have the same distribution over strata defined by levels of potential **confounders** as cases in a **case–control study** or as the exposed cohort in a **cohort study** (*see* **Matching**; **Frequency Matching**). Stratification (or stratified analysis) also refers to the analytical strategy that controls for confounding by estimating the **association** between exposure and disease status within strata defined by categorized levels of potential confounders and then combining stratum-specific results to obtain an overall estimate of exposure effect (*see* **Mantel–Haenszel Methods**; **Matched Analysis**).

In the context of survey sampling, stratification is an efficient design that usually allocates larger samples to strata of the population within which the estimate has a large **variance** (*see* **Stratified Sampling**).

MITCHELL H. GAIL

Stratified Sampling, Allocation in

Allocation describes the size of the sample to be selected in each stratum. The fact that a fixed sample size can be selected from each stratum is due to the two defining characteristics of a stratified design: (i) all members of the population can be partitioned into strata and (ii) samples of these members can be selected independently among strata (*see* **Stratified Sampling**).

Whether the total sample allocation is constrained by fixed sampling costs or whether an estimate with a specific **variance** is needed at any cost, a judicious allocation of sample size to strata can result in significant survey gains. If costs are fixed, then a good sample allocation can produce estimates with a smaller variance than other estimates sampled at equal costs. If a target variance is specified for an estimate, then a good allocation will produce estimates with variance equal to costlier alternatives. Gains in allocation are greatest when either the stratum variances or differential stratum sampling costs vary widely. To achieve a gain, only knowledge of the relative values of stratum variances is needed. Rough guesses of stratum variances and costs, possibly using related data and surveys, may even be enough to achieve large gains. Lastly, although the allocation of sample will influence the precision of an estimate and can reduce the cost of a sample, there is no wrong allocation in the sense that **unbiased** estimates of linear population parameters can always be made from any allocation in which each stratum is sampled.

When estimates of the population **mean** or population total are needed, proportional allocation or the all-encompassing optimal allocation is usually used. When individual strata estimates or estimates of strata differences are needed, equal allocation is most commonly used. Note that the population mean is defined as $\overline{Y} = \sum_{j=1}^{L} N_j \overline{Y}_j / N_\cdot$, where L is the number of strata, N_j is the number of elements in stratum j, $\overline{Y}_j = \sum_{k=1}^{N_j} Y_{jk}/N_j$ is the stratum mean and $N_\cdot = \sum_{j=1}^{L} N_j$ is the total population size. The (unbiased) estimator of the population mean, considered here, is $\overline{y} = \sum_{j=1}^{L} N_j \overline{y}_j / N_\cdot$, where $\overline{y}_j = \sum_{k\in s}^{N_j} Y_{jk}/n_j$ is the stratum mean based on a sample of size n_j. The population total is defined as $Y = \sum_{j=1}^{L} N_j \overline{Y}_j$ with an estimator: $\hat{y} = \sum_{j=1}^{L} N_j \overline{y}_j$.

As its name implies, proportional allocation involves allocating the total sample size proportionally to strata size. Given that a total sample of size n will be selected and that the number of individuals in stratum j is N_j, the proportional allocation of sample to stratum j is

$$n_j = n\frac{N_j}{N}.$$

Even though proportional allocation is most appropriate when both the variability of individuals within each stratum and sampling costs are constant across strata, it is often employed when little is known about the strata variances or collection costs. This is because it is believed that a good stratification, in producing homogeneous strata, may also produce strata having similar variances.

An optimal allocation is the sample allocation that produces an estimate with the smallest variance among all sample allocations under consideration. Note that for any allocation:

$$\mathrm{var}\overline{y} = \sum_{j=1}^{L} \left(\frac{N_j}{N}\right)^2 \left(1 - \frac{n_j}{N_j}\right) \frac{S_j^2}{n_j},$$

where

$$S_j^2 = \sum_{k=1}^{N_j} \frac{(Y_{jk} - \overline{Y}_j)^2}{(N_j - 1)} \quad \text{and} \quad \mathrm{var}\hat{y} = N^2 \mathrm{var}\overline{y}.$$

If there are L strata and the variance within stratum j is S_j^2, then the optimal allocation to stratum j, when the total sample is of size n is

$$n_j = n\frac{N_j S_j}{\sum_{k=1}^{L} N_k S_k}.$$

This type of allocation is also referred to as Neyman allocation, after **J. Neyman**, who showed in 1934 that it is optimal among all fixed-sized samples [6].

Costs may vary widely when the strata utilize different sampling methods. For example, one stratum may contain households with a telephone and another stratum households without a telephone. Here the cost of sampling a household with a telephone is generally cheaper than a nontelephone household since it involves making a telephone call instead of a face-to-face interview. If the cost per element in stratum j is c_j, and only allocations of cost, c,

are allowed, i.e. $c = \sum_{j=1}^{L} n_j c_j$, then the optimal allocation to stratum j is

$$n_j = c \frac{N_j S_j / \sqrt{c_j}}{\sum_{k=1}^{L} N_k S_k / \sqrt{c_k}}.$$

This allocation finds the sample $n_1, \ldots, n_L$ that minimizes $\mathrm{var}\overline{y}$ (or, equivalently $\mathrm{var}\hat{y}$), subject to the cost constraint $c = \sum_{j=1}^{L} n_j c_j$. An estimator based on this allocation will have a variance smaller than one based on any other allocation of equal cost.

A number of facts can be observed by viewing the sample allocation formula listed above. Given equal strata costs, the optimum allocation selects relatively more units from strata with large variances and, given equal strata variances, the optimal allocation selects relatively more units from the cheaper strata. Comparing the optimal allocation with proportional allocation, it can be seen that proportional allocation is, itself, optimal when the within-strata variance is proportional to the cost. (The most common case is when both collection costs and variances are constant across strata.) Neyman allocation can be seen to be optimal for a fixed cost design when strata collection costs are constant.

Instead of a fixed cost design, an allocation may be needed to obtain an estimate with fixed variance. Given the same linear cost structure, the optimal allocation with fixed variance takes the same form as the optimal fixed cost allocation except that the cost c is unknown and must be solved. The minimal cost to achieve a fixed variance V is

$$c = \frac{\left(\sum_{k=1}^{L} N_k S_k \sqrt{c_k}\right)^2}{N^2 V - \sum_{k=1}^{L} (N_k S_k)^2}.$$

Detailed derivations of all the above allocation formulas are in most sampling books (see, for example, Cochran [1] or Sarndal et al. [7]).

Equal allocation is used if estimates for individual strata or strata differences are needed. Generally, this will not be a good allocation for estimating the population mean or total unless sampling costs, stratum variances, and stratum sizes are all constant across strata (in this case equal, proportional, and optimal allocations are identical). Allocation for strata means (or differences of means) is an example of multipurpose design since the objective, now, is to lower more than one variance. More on multipurpose allocation will be given below.

The allocations described above apply to **unbiased**, linear estimates (estimates consisting of weighted averages of sampled values) with a linear cost (when applicable). In reality, estimates may be nonlinear (e.g. **ratio or regression estimators**) as may be the costs. In practice, allocations for nonlinear estimators and nonlinear costs usually are determined by first replacing estimates and costs by their approximate, linearized versions.

Using the basic formulas given above, a few minor problems can arise. The allocations may not be integral valued. Also, the allocation to a stratum can sometimes be larger than the stratum population size, or it may be less than one. The nonintegral problem can be bypassed by using integer programming techniques that will only find optimal integral solutions. However, rounding the allocations to their nearest integer will give nearly optimal results. Allocations that are out of bounds can be avoided by using **linear programming** techniques. This approach is not commonly employed because it does not provide an easy-to-use formula and the following quick fix is usually adequate. When an allocation in stratum j is above N_j, just select N_j from that strata and reallocate the remaining $n - N_j$ to the remaining $L - 1$ strata using as before. This procedure can also be applied for sample sizes smaller than one. That is, a sample size of one is allocated to a stratum with allocation less than one. The stratum is then removed from the variance formula and the remaining sample is allocated to the remaining strata. Note that since unbiased estimates of variance require a sample size of at least two, only allocations that include at least two sample units per stratum are usually considered. A more detailed explanation of correcting for allocations out of bounds can be found in [1, p. 104] or [7, Section 12.7].

Owing to high costs per sampling unit, most surveys collect many items. In such a situation, an allocation is needed to provide precise estimates for every item. An allocation that provides the minimal variance for estimates of all items is clearly impossible, unless each item exhibits the same relative strata variances. Since only one allocation can actually be fielded, no allocation will be optimal for all items.

There are, however, several compromise methods used in practice. For example, if it is important to measure one item with as much precision as possible, then the allocation should be based solely on that item. Proportional allocation is also used in this context because, if subpopulations are not of interest, strata variances may be relatively constant for most items. If a fixed precision is required for each estimate, then the solution is clear: determine the minimal cost sample size in which all minimum precisions are obtained. Numerical solutions requiring linear programing methods are available to do this (see, for example, [4]).

Fixed cost multipurpose design places a further constraint on allocation. The concept of an admissible allocation can be used to determine a subset of acceptable allocations. An admissible allocation, with respect to multiple estimators, is one which cannot be uniformly improved on. For example, denote a possible allocation by $\underline{n}$. Given R estimators, denote the variance of the rth estimator using allocation $\underline{n}$ by $\text{var}(\overline{y}^{(r)}|\underline{n})$. An allocation $\underline{n}_0$ is admissible if there is no allocation, $\underline{n}$, such that $\text{var}(\overline{y}^{(1)}|\underline{n}) \leq \text{var}(\overline{y}^{(1)}\,|\,\underline{n}_0), \ldots, \text{var}(\overline{y}^{(R)}\,|\,\underline{n}) \leq \text{var}(\overline{y}^{(R)}\,|\,\underline{n}_0)$ and $\text{var}(\overline{y}^{(r)}|\underline{n}) < \text{var}(\overline{y}^{(r)}|\underline{n}_0)$ for at least one r, $1 \leq r \leq R$. Folks & Antle [3] give the solution for all admissible strata allocations based on R responses. When both individual strata means and an overall population mean are needed for one item, Bankier [2] advocates power allocations. This method provides a subset of admissible allocations ranging between proportional allocation (good for populations mean) and equal allocation (good for strata means). Graphs of the tradeoffs in precision can be viewed to determine an acceptable compromise.

If the measurements are all on the same scale, an allocation that minimizes the maximum variance (**minimax**) can be useful. For example, suppose that an allocation is needed to estimate each stratum mean. The minimax allocation that will produce estimates of equal precision is the solution to

$$\left(1 - \frac{n_1}{N_1}\right)\frac{S_1^2}{n_1} = \cdots = \left(1 - \frac{n_j}{N_j}\right)\frac{S_j^2}{n_j} = \cdots = \left(1 - \frac{n_L}{N_L}\right)\frac{S_L^2}{n_L}.$$

This solution can be determined by a numerical search procedure. Note that if all N_i are large, then an approximate minimax allocation is the solution to

$$\frac{S_1^2}{n_1} = \cdots = \frac{S_j^2}{n_j} = \cdots = \frac{S_L^2}{n_L}.$$

For a fixed sample size, the minimax solution is seen to be

$$n_j = n\frac{S_j^2}{\sum_{k=1}^{L} S_k^2},$$

giving equal allocation when all variances are equal.

Minimax allocations are very useful for all types of multiple objective surveys, as was shown by Malec [5], who employs the concept of admissible designs to find minimax designs.

References

[1] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[2] Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas, *American Statistician* **42**, 174–177.

[3] Folks, J.L. & Antle, C.E. (1965). Optimum allocation of sampling units when there are R responses of interest, *Journal of the American Statistical Association* **60**, 225–233.

[4] Huddleston, H.F., Claypool, P.L. & Hocking, R.R. (1970). Optimal sample allocation to strata using convex programming, *Applied Statistics* **19**, 273–278.

[5] Malec, D. (1995). Optimal multiple objective sample design using an admissibility criterion, *Journal of Statistical Planning and Inference* **28**, 229–240.

[6] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society* **97**, 558–625.

[7] Sarndal, C.E., Swensson, B. & Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

D. MALEC

Stratified Sampling

Stratified sampling is a **probability sampling** method that is frequently implemented in sample surveys. In general, a population's elements (or in practice the **sampling frame** members) are divided into distinct groups or strata based upon the similarity of selected characteristics important to the survey. Typically, each stratum is independently sampled using a method for which an **unbiased** estimator of stratum total or stratum mean can be computed. These estimated stratum totals may then be added to obtain an estimator for the population total. Similarly, a stratum-weighted average of the estimated stratum means may be computed to estimate the overall population **mean** (*see* **Estimation**). **Stratification** is used to increase the efficiency of a sample design with respect to cost and estimator precision.

In this article we discuss the theoretical foundations of stratified sampling. For simplicity, we focus on the most elementary sampling structure, *stratified random sampling*. Here, the population elements are sampled by **simple random sampling**. "Real-life" surveys tend to use stratified multistage cluster sampling (*see* **Cluster Sampling**; **Multistage Sampling**), but most of the elementary foundations presented here can be extended to these more complicated structures.

Two simple examples of stratified populations are as follows:

1. A population of physicians is stratified by state of practice and specialty (e.g. cardiology or neurology). One such stratum is New York cardiologists.
2. A population of hospital discharged patients in the US is stratified by region, and size of hospital classified by bed size. A typical stratum could be hospitals in the South with 500 or more beds.

After we establish some basic groundwork on stratification, some actual surveys will be discussed.

Basic Foundations for Stratification

The planners of a sample survey usually start by having a set of analytic survey objectives, including domains of study and precision requirements for target estimators. Additionally, costs and administrative constraints are an integral part of the survey design process. Cochran [1, Section 5.1] discusses four principal reasons why planners consider a stratified design:

1. The population contains subpopulations or subdomains which are of primary interest to the survey planners. When distinct estimates of known precision are needed for these selected subdomains, it is advisable to treat each subpopulation as a "population" in its own right. Sample sizes within a designated subpopulation stratum may be increased to meet target precision levels.
2. It may be administratively convenient to stratify. An agency conducting a survey may have field offices which may be used to stratify a population. Each field office may be responsible for the survey administration of its part.
3. Distinct groups within a population may differ to such a degree that different sampling procedures are required. Also, since the population of study may be partitioned by multiple frames with different operational characteristics, a stratified sample may be the only workable option.
4. Stratification may improve the precision of sample estimators of the entire population. It may be possible to divide a heterogeneous population into subpopulations, each of which is internally homogeneous. Sampling variability within each stratum should be much smaller than sampling variability over the population as a whole. Precise estimates of the means of each stratum can be obtained by targeted sample sizes, and then the estimates combined to form an estimate for the entire population. With an appropriate allocation of sample, this estimator will be more precise than one created from the same size sample, but without stratification.

The gaining of precision through stratification, as outlined in reason 4 above, can be justified theoretically for many commonly used sampling methods. For simplicity, the mathematical theory presented here will focus upon simple random sample methods and linear estimators; for example, totals and means having a known base.

Suppose that a population of N elements has already been divided into L strata of known sizes N_h, $h = 1, 2, \ldots, L$, and that stratum h contains population elements Y_{hi}, $i = 1, 2, \ldots, N_h$. The true

stratum population means and **variances** are defined as $\overline{Y}_h = \sum_{i=1}^{N_h} Y_{ih}/N_h$ and $S_h^2 = \sum_{i=1}^{N_h}(Y_{hi} - \overline{Y}_h)^2/(N_h - 1)$, respectively, and the true total population mean and variance may be expressed as $\overline{Y} = \sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi}/N = \sum_{h=1}^{L}(N_h/N)\overline{Y}_h$, and $S^2 = \sum_{h=1}^{L}\sum_{i=1}^{N_h}(Y_{ih} - \overline{Y})^2/(N-1)$, respectively. Now, if a simple random sample (SRS) of size n is taken from the entire population, the typical estimator of $\overline{Y}$ is the sample mean, $\overline{y}_{\text{srs}} = \sum_{j=1}^{n} y_j/n$. For stratified random sampling over L strata, a random sample of size n_h is taken independently within each group. The sample stratum mean, $\overline{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$, is calculated and weighted by relative stratum size to define the stratified estimator, $\overline{y}_{\text{str}} = \sum_{h=1}^{L}(N_h/N)\overline{y}_h$. Both estimators are unbiased for the true population mean – that is, $\text{E}(\overline{y}_{\text{srs}}) = \text{E}(\overline{y}_{\text{str}}) = \overline{Y}$ – and they have respective variances:

$$\text{var}(\overline{y}_{\text{srs}}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \quad \text{and}$$

$$\text{var}(\overline{y}_{\text{str}}) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right).$$

Note that for an estimator of population total, one just multiplies the above mean estimators by N and the variances by N^2.

The impact of stratification on survey costs and estimator precision depends upon how the sample of n units is allocated to the strata. If the unit cost of sampling varies by some naturally defined strata, few survey planners would consider an unstratified sample. Most surveys are conducted under the constraint of a fixed budget, and the survey planners must have some ability to control sample sizes by stratum costs. For example, the National Hospital Discharge Survey (NHDS), a survey of hospital procedures performed throughout the year, stratifies hospitals into those the records of which have been automated and are served by an abstract service and into those the records of which must be processed manually. An "automated-record" hospital costs considerably less to sample and process than a "manual-record" hospital. This topic of optimal sample allocations with respect to cost and precision requirements is discussed in greater detail in Sukhatme et al.[7, Chapter 4] (*see* **Stratified Sampling, Allocation in**).

For surveys in which sample unit costs are identical over strata, both stratified and nonstratified sampling can readily be compared for various allocations of a fixed total sample size n. First, under some mild conditions it can be shown that the optimal allocation of the total sample size n into L independent samples, n_h, is defined by

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h},$$

with S_h the stratum population **standard deviation**. This is referred to as the Neyman allocation. This is optimal in the sense that $\text{var}(\overline{y}_{\text{Neyman}}) \le \text{var}(\overline{y}_{\text{str}})$ for any other stratified sample allocation. Intuitively, the largest strata are the most important in estimating the population mean, and more sample should be allocated to the large strata. Also, the strata with the largest dispersions require more sample for precise estimation. The proportionality factor $N_h S_h$ can be can be thought of as a combined measure of these two intuitive concepts. In practice, a true Neyman allocation is rarely implemented, because the true S_h is usually unknown. Instead, survey planners use a variable that has a close relationship with the variable of interest. For example, in the NHDS discussed earlier, hospitals are stratified in part by geographic location, type of hospital, and bed size. Bed size is positively correlated with the number of discharges that a hospital produces and could be used to define a Neyman allocation.

Another commonly used stratified sampling method is proportional sampling, where $n_h = n(N_h/N)$. In practice, this method is easy to implement provided that the true sizes, N_h, of the strata are known. If the S_h do not deviate much by strata, then proportional sampling will be close to the Neyman optimal. If the strata sizes N_h are large, then the following relationships among the variances for Neyman, proportional, and unstratified SRS sampling can be established:

$$\text{var}(\overline{y}_{\text{srs}}) = \text{var}(\overline{y}_{\text{prop}}) + \frac{(1 - n/N)}{n} \times \sum_{i=1}^{L}\left(\frac{N_h}{N}\right)(\overline{Y}_h - \overline{Y})^2$$

with

$$\text{var}(\overline{y}_{\text{prop}}) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)S_h^2$$

and

$$\mathrm{var}(\overline{y}_{\mathrm{prop}}) = \mathrm{var}(\overline{y}_{\mathrm{Neyman}}) + \frac{1}{n}\sum_{i=1}^{L}\left(\frac{N_h}{N}\right)(\overline{S}_h - \overline{S})^2.$$

From the above two equations, one sees that the proportional allocation will provide a sampling variance at least as small as an unstratified simple random sample. The greatest sampling efficiencies occur when the true stratum means vary to a large degree. Neyman allocation will reduce the variance by a factor proportional to the variance of the stratum standard deviation. It should be noted that for a fixed proportion allocation on the same strata, both $\mathrm{var}(\overline{y}_{\mathrm{prop}}) \leq \mathrm{var}(\overline{y}_{\mathrm{srs}})$ and $\mathrm{var}(\overline{x}_{\mathrm{prop}}) \leq \mathrm{var}(\overline{x}_{\mathrm{srs}})$ for different characteristics y and x. This is an attractive property of stratified proportional sampling. For a Neyman allocation determined by a variable y, but also used for an other unrelated survey variable x, it is possible that $\mathrm{var}(\overline{x}_{\mathrm{Neyman}[y]}) \geq \mathrm{var}(\overline{x}_{\mathrm{srs}})$. Such a phenomenon may occur if $S_h(x)$ and $S_h(y)$ are negatively correlated. For example, in a population stratified by income level, a Neyman allocation targeted to estimate occupational work-loss days would probably be quite inefficient for estimating health insurance coverage for the unemployed. Thus while proportional sampling may not be optimal, it would be a safe strategy to use when several different variables are to be estimated using the same sample.

While stratified proportional sampling reduces the variance relative to unstratified SRS, Kish [3, Section 3.4] points out that, in practice, the relative gains may be only small or moderate. This is because survey planners do not have population variables available to define a highly efficient stratification. A special case of interest is the impact of stratification on estimating a population proportion, p. Here, the population variance is $p(1-p)$, and the stratum variance becomes $S_h^2 = p_h(1-p_h)$, with p_h the stratum proportion. The nature of this variance makes it somewhat insensitive to stratification if the resulting strata p_hs are in the central range of 0.20–0.80. However, for stratified cluster sampling, typical in major surveys, the efficiency gains for proportional sampling are greater than the element sampling discussed here.

If the strata themselves are of interest and comparisons are to be made among strata, then the individual stratum estimates, $\overline{y}_h$, and not the aggregate, $\overline{y}_{\mathrm{str}}$, are most important in meeting precision requirements. Equal allocation of sample to strata, $n_h = n/L$, could be used. In such cases, especially when the strata sizes, N_h, vary greatly in size, $\mathrm{var}(\overline{y}_{\mathrm{str}})$ may exceed $\mathrm{var}(\overline{y}_{\mathrm{srs}})$. For several computationally detailed examples of the stratified random sampling method, the reader is referred to Levy & Lemeshow [4].

The above discussion has assumed that a mean or total for an entire population is the target of estimation. Also important is the subdomain or subpopulation estimator. Frequently, subdomains dictate the main precision requirements of a survey. The estimation of the mean of a population subdomain is a special case of ratio estimation (*see* **Ratio and Regression Estimates**). Here, the **target population** parameter is

$$\overline{Y}_{\mathrm{D}} = \frac{\sum_{i=1}^{N} Y_i \delta_i}{\sum_{i=1}^{N} \delta_i},$$

where $\delta_i = 1$ if unit i is in subdomain D and 0 if not. A combined ratio estimator is $\overline{y}_{\mathrm{D,strc}} = (\overline{yd})_{\mathrm{str}}/\overline{d}_{\mathrm{str}}$, where the $(\overline{yd})_{\mathrm{str}}$ and $\overline{d}_{\mathrm{str}}$ are the stratified estimators of the numerator and denominator of $\overline{Y}_{\mathrm{D}}$. For this ratio estimator, the denominator may be random, and the **linearization method** may be used to derive approximate variance formulas.

Real Examples of Stratified Sampling

The following examples are taken from the **National Center for Health Statistics (NCHS)** family of surveys. They should be typical of the large-scale surveys conducted by government agencies to produce official statistics. The somewhat involved design structures of the NCHS surveys, as with most large-scale surveys, require some oversimplification to conceptualize the fundamentals. Furthermore, the design structures of NCHS surveys tend to change over time as objectives change. Most of the examples below should be considered as core design structures which will be somewhat modified for any specified year. The reader should refer to the NCHS references to get a more thorough description of any specific survey.

Example 1: National Maternal and Infant Health Survey (NMIHS)

One component of the NMIHS targets mothers who have recently given live birth or experienced fetal or infant death. Sample mothers are contacted by mail or telephone whenever possible. With this mode of data collection, certificates of live births, certificates of infant death, and official reports of fetal death are used as sampling frames. These frames are stratified and sampled as follows. Since birth and death records are collected and processed within a registration area, typically a state, these areas define an imposed administrative stratification. Mothers of low-**birthweight** infants and of black infants are specific targets of investigation. To meet these two study objectives, the live-birth certificates are partitioned by black and nonblack status and then cross classified by the infant's birthweight to form race–birthweight strata within each state. Simple random samples are taken within each stratum, but to meet precision goals for low-birthweight and/or black infants, these particular strata are oversampled, i.e. $n_B/n > N_B/N$, where B is any oversampled stratum.

Example 2: National Ambulatory Medical Care Survey (NAMCS)

For NAMCS, the target population consists of all patient visits to physicians engaged in office-based practice in a given year. The mode of data collection for this survey involves an experienced interviewer conducting face-to-face interviews with the sampled physicians. To keep costs manageable, a stratified multistage cluster sample is used. The primary sample units (PSUs) are a sample of "representative" counties (or equivalent territorial divisions) from a stratification of all counties in the US. Counties are stratified within four regions of the US by criteria involving the similarity of US Bureau of Census classifications (e.g. metropolitan status) and Decennial Census statistics available at the county level (e.g. county income or minority populations). PSUs are selected by **sampling with probability proportionate to size**. The size used for the NAMCS' PSUs has been the person population size, not the number of physicians. For cost considerations, the same sampled counties have served as sampled PSUs for other surveys. Since the sizes of the person population and physician population in a county are highly correlated, little efficiency is lost. Next, a sampling frame of licensed physicians taken from recent professional directories is created. This frame associates physicians with the counties and, furthermore, the physicians' specialty (e.g. family practice or neurology) is available on the frame. Ambulatory care should vary somewhat by the physicians' specialty, so within each sampled PSU the physicians are then substratified by specialty. Physicians are selected within specialty substrata by SRS, in such a way that over the first two sampling stages, any physician in the US has the same probability of being in the sample. This is called a self-weighting sample and may be expressed as follows.

If a physician b lives in PSU a then

$$\begin{aligned}\Pr(b \text{ is sampled}) &= \Pr(b \text{ sampled given PSU } a \text{ is sampled}) \Pr(\text{PSU } a \text{ is sampled}) \\ &= \left(\frac{n_{ah}}{N_{ah}}\right)\left(\frac{M_a}{M_s}\right) = \frac{1}{SI},\end{aligned}$$

where M_s is the size of the stratum containing PSU a, M_a is the size of county a, N_{ah} is the size of the specialty substratum h in PSU a, SI is a sampling interval determined to meet design objectives, and $n_{ah} = (N_{ah}/M_a)(M_s/SI)$ defines the sample size of physicians to be taken from substratum h in county a. Finally, a SRS sample of about 30 patient visits is sampled from each physician's practice. The size at this level is determined so not to burden the physicians.

Example 3: National Health Interview Survey (NHIS)

The NHIS, a major health survey sampling about 50 000 households and about 120 000 persons, is conducted annually over the US noninstitutionalized population. It targets numerous health variables on age–race/ethnic–sex domains for specified precision levels. The mode of data collection involves a face-to-face interview which requires a stratified multistage geographic cluster sample to be cost effective. The target population is not directly available as a sampling frame, but is covered for the most part by two distinct frames. An area sample frame consisting of small geographic areas with dwelling units covered by the most recent US Decennial Census includes most of the target population. To

keep the coverage current, a frame consisting of places where new residential housing has been constructed since the last Decennial Census is also created. The former frame contains Decennial Census social–economic–demographic information, but the latter frame contains only location. For the survey design implemented from 1995 to 2004, a survey objective is to make the NHIS more conducive to possible future dual frame surveys at the state level. This objective results in the population primary sampling units (PSUs), counties (or equivalents), being stratified within each state by metropolitan status and poverty status. Large states have more strata than small states. For national statistics, this stratification is not as efficient as one that would allow stratum geographic boundaries to cross state borders. Within each PSU, smaller geographic units, called segments, are classified into 20 race–ethnicity substrata by the density of black and Hispanic populations covered by the area frame, and into a substratum containing the new construction frame.

At the first stage of sampling, either one or two PSUs are selected from each stratum, with probability proportional to population size. The very large population PSUs – for example, New York City and Los Angeles – are self-representing; that is, they are in the sample with certainty. At the second stage, a SRS sample of segments is taken from each race–ethnic substratum within a sampled PSU. At the third stage, all black or Hispanic households are sampled within each segment, while a SRS of the complement households is taken. The second- and third-stage SRS sample sizes are defined in such a way as to obtain precise estimates for minority populations. The end result is that black and Hispanic samples are a much greater proportion of the sample than are their population proportions. This allows for precise estimation on small minority age–sex subdomains.

Example 4: National Health and Nutrition Examination Survey (NHANES)

The NHANES family of surveys assess the health of the noninstitutionalized population. They are much more in-depth than the NHIS in that they target about 30 000 sample subjects to be given a complete physical examination. This examination is conducted in a specially designed mobile examination center, and it takes 6 years of data collection. Operational constraints and costs dictate that only about 89 geographic sites in the US can be visited, each for a period of about one month to collect data. NHANES planners have the objectives of measuring hundreds of health related variables for 52 different age–sex–race/ethnic subdomains of the US population, with an emphasis on black and Mexican-American domains. A comparison of domain considerations dictates that each domain should have a sample size of 560 or more to meet the following precision requirements:

1. If D is a targeted domain, then $\mathrm{se}(\hat{p}_\mathrm{D})/\mathrm{E}(\hat{p}_\mathrm{D}) \leq 0.30$ whenever $\mathrm{E}(\hat{p}_\mathrm{D}) = 0.10$ (relative standard error requirement).
2. If D_1 and D_2 are two distinct targeted domains, then a test of hypotheses (*see* **Hypothesis Testing**) $\mathrm{H}_0 : \mathrm{E}(\hat{p}_{\mathrm{D}_1} - \hat{p}_{\mathrm{D}_2}) = 0.0$ vs. $\mathrm{H}_1 : |\mathrm{E}(\hat{p}_{\mathrm{D}_1} - \hat{p}_{\mathrm{D}_2})| \geq 0.10$ should have **power** at least 0.90 for a size 0.05 test.

A county is the primary sampling unit, just as in the NHIS. While age and sex define domains of interest, and one might expect health variables to differ on these domains, age and sex make poor stratification variables when aggregated at the county level. This is because counties vary little in their proportional age–sex composition. On the other hand, race–ethnic aggregates make good county stratifiers, since they vary greatly by county. NHANES used Decennial Census information on race/ethnicity along with income and metropolitan status to stratify the counties into 47 strata. Unlike the NHIS, the measure of size of a county is not the total population, but a composite measure that placed more weight on a county's black and Mexican-American components. One or two counties are sampled from each stratum, and then multiple stage sampling is used to obtain a sample of persons. The precision objectives for minority populations leads to the black, Mexican-American, and the white/other populations to have approximate sample proportions of 0.30, 0.30, and 0.40, respectively, while the population proportions are about 0.13, 0.07, and 0.80, respectively.

Defining Strata

The discussion so far has considered sampling methods while treating the strata as already given, but the construction of the strata themselves is also important. Some basic issues involve the selection

of stratification variables, the boundaries between strata, and the number of strata to use (see Särndal et al.[6, Chapter 12]). Most optimality results cannot be implemented in practice, due to limited population information, conflicting design objectives, cost considerations, and administrative restrictions. Frequently, survey planners consider several stratified sampling options with respect to cost and precision to determine a design that will perform well over a wide spectrum of target variables; the selected design may not be optimal for any given variable.

Some stratification boundary defining rules have been studied under theoretic conditions. The best known is the *cum* $\sqrt{f}$ *rule*. This rule states that if y is a continuous variable, and $f(y)$ is the density function of y with support $[A_L, A_U]$ then an approximate optimal stratification of the population into H units with boundaries $A_L = a_0 < a_1 < a_2 < \ldots < a_{H-1} < a_H = A_{\mathrm{U}}$, each to be sampled by the Neyman allocation, would result from creating the H strata in such a way that

$$\int_{a_{\mathrm{H}-1}}^{a_{\mathrm{H}}} [f_Y(y)]^{1/2}\,\mathrm{d}y = \frac{1}{H}\int_{A_{\mathrm{L}}}^{A_{\mathrm{U}}} [f_Y(y)]^{1/2}\,\mathrm{d}y.$$

That is, each stratum accounts for $1/H$ of the total integral of $\sqrt{f}$. In practice, a variable related to y would be used. Using a histogram approach, Cochran [1, Section 5.A.7] provides a computational example of this method.

For element sampling, it is suggested by Kish [3, Section 3.6I] and Cochran [1, Section 5.A.8] to keep the number of strata modest in size. This suggestion is based in part on some simple theoretic structures. Let y be a **linear regression** on x, with ρ the **correlation** between y and x in the unstratified population. If the population is partitioned into L strata defined by the variable x, using the optimal strata boundaries along with sample sizes of n/L in each stratum, then

$$\frac{\mathrm{var}(\overline{y}_{\mathrm{str}})}{\mathrm{var}(\overline{y}_{\mathrm{srs}})} \geq \left[\frac{\rho^2}{L^2} + (1-\rho^2)\right].$$

As L becomes large, the lower bound tends to $(1-\rho^2)$, and a point of little return in variance reduction can be established. For $\rho < 0.95$, little reduction occurs for more than six strata. This argument would assume that, overall, estimators rather than individual stratum estimators are important.

Stratification after Sampling

Frequently, variables well-suited for partitioning the population into strata do not exist before sampling. In this case, an estimation technique, called **poststratification** can be used. Here, the sampled data are stratified after sampling, and then an estimator for population mean or total is created as if the sample had come from a presample stratification. More precisely, in the case that a simple random sample of size n is taken from an unstratified population, the sample is first poststratified into H strata, with sample stratum means $\overline{y}_g$, for $g = 1, 2, \ldots, H$, and then a poststratified mean estimator is defined: $\overline{y}_{\mathrm{pstr}} = \sum_{g=1}^{H}(N_g/N)\overline{y}_g$, where N_g/N are the population totals of the poststratification classes (see [8, Section 11.6]).

While the functional forms of the stratified and poststratified mean appear identical, there are some important distinctions as to implementation and statistical properties. First, selected poststratification classes must have known population sizes (or independently known accurate estimates of size). For example, in large-scale surveys, age–race–sex classes are frequently used for poststratification, since the US Bureau of the Census produces very accurate national tabulations, which it updates quarterly. However, a poststratification on a health status variable would be difficult, since the true class totals could not be obtained. Second, the sample sizes, n_g, observed within the poststratification cells are themselves random variables. If these sample sizes are reasonably large, perhaps having $n_g > 20$ in each class, then this method is almost as precise as the proportional stratified sampling discussed earlier. This method can be extended to more complicated sampling schemes.

Other Stratification Issues

In many large-scale surveys, the process of creating multiple levels of stratification and simple random samples can become quite involved. Instead, a process called implicit stratification along with **systematic sampling** is used [5]. For example, for an official survey of a metropolitan area, the US Bureau of the Census may provide a large sampling frame which partitions the area into identified blocks along with each block's urban status (central city or not), percentage minority population, and **median** income.

This frame may be sort-ordered by these variables. This example is a three-layer implicit stratification. Next, a systematic sample, say 1 of every 50 blocks or sampling interval of 50, can be taken from the ordered frame to obtain the sample. This method is very easy to implement and has frequently been used to obtain the within PSU samples of many NCHS surveys. The expected systematic sample sizes are proportional to the total size within any sort level. For variance estimation purposes, coarse levels containing large samples are often treated as stratified proportional samples. This facilitates analysis using conventional software (*see* **Software for Sample Survey Data**).

Well-designed surveys plan a method to compute unbiased estimators of variance for the basic total and mean estimators. In large-scale multistage cluster samples, such as NAMCS or NHIS, some strata may have only one sampled cluster. In such cases, no unbiased estimator exists. For such cases, collapsed strata may be created for variance estimation (see [2, Section 8.6]). Original strata may be collapsed by combining strata with similar stratum characteristics. The sampled clusters within a collapsed stratum are treated as having been sampled with replacement (*see* **Sampling With and Without Replacement**).

References

[1] Cochran, W.G. (1976). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[2] Foreman, E.K. (1991). *Survey Sampling Principles*. Marcel Dekker, New York.

[3] Kish, L. (1965). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[4] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

[5] Murthy, M.N. & Rao, T.J. (1988). Systematic sampling with illustrative examples, in *Handbook of Statistics*, Vol. 6, P.R. Krishnaiah, & C.R. Rao, eds. North-Holland/Elsevier, New York, pp. 147–185.

[6] Särndal, C.E., Swensson, B. & Wretnman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[7] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. & Asok, C. (1984). *Sampling Theory of Surveys with Applications*, 3rd Ed. Iowa State University, Ames.

[8] Thompson, S.K. (1992). *Sampling*. Wiley, New York.

The following publications discuss survey methods for some of the NCHS surveys. A comprehensive NCHS publication list can be obtained at The Centers for Disease Control and Prevention website: `http://www.cdc.gov/nchswww/products/products.htm`.

Further Reading

Bryant, E. & Shimizu, I. (1988). Sample Design, Sampling Variance, and Estimation Procedures for the National Ambulatory Medical Care Survey, National Center for Health Statistics, *Vital Health Statistics* Series 2, No. 108.

Kovar, M.G. (1989). Data Systems of the National Center for Health Statistics, National Center for Health Statistics, *Vital Health Statistics* Series 1, No. 23.

Massey, J.T., Moore, T.F., Parsons, V.L. & Tadros, W. (1989). Design and Estimation for the National Health Interview Survey, 1985–94, National Center for Health Statistics, *Vital Health Statistics* Series 2, No. 110.

NHANES III (1996). Reference Manuals and Reports, issued October 1996 (CD-ROM), *GPO 017-022-01358-4*, US Government Printing Office, Washington.

Schoendorf, K.C., Parker, J.D., Batkhan, L.Z. and Kiely, J.L. (1993). Comparability of the Birth Certificate and 1988 Maternal and Infant Health, National Center for Health Statistics, *Vital Health Statistics* Series 2, No. 116.

Wichura, M., Hinkelman, K. & Thisted, R., eds (n.d). *Current Index to Statistics*. Extended Database, American Statistical Association, Alexandria. This database can be used to search the literature for topics about stratification.

V. PARSONS

Stroke

World Health Organization criteria define stroke in humans as a sudden onset of signs of focal or global disturbance of cerebral function lasting more than 24 hours unless interrupted by surgery or death, with no apparent nonvascular cause [39]. For example, stroke can be characterized by the sudden loss of the ability to use the right hand and arm, the loss of the capability to produce and comprehend speech, or the loss of ability to see to one side. The causes of stroke vary. The most common is the sudden occlusion of a brain artery by a blood clot (ischemic stroke). Another cause is bleeding from a brain artery into the brain itself (intracerebral hemorrhage) or into the fluid-filled space around or within the brain (subarachnoid hemorrhage). Stroke is a major cause of morbidity and mortality in the world. In the US and other industrialized countries stroke is the third most common cause of death [2]. Although mortality has been in an accelerated decline from 1965 to 1985, studies suggest the decline is primarily attributable to a decrease in **case–fatality** rates rather than a decrease in incidence [4, 40].

Of the 500 000 strokes that occur each year, only 100 000–150 000 are fatal [2]. For most patients who survive a stroke, there is wide variability in impairment of physical and neurological function. Symptoms are highly dependent on time from stroke onset. During the course of disease, symptoms usually worsen, then gradually improve, albeit usually not completely. The disability can be a personal disaster for the patient. Loss of independence or a prolonged period of nursing home care is a common outcome [30].

Historical Development

Until 1995 no treatment was available for stroke. Studies focused on describing the epidemiology of stroke and on **clinical trials** of interventions for primary or secondary stroke prevention. Studies of mechanisms and etiology initially were restricted to animal models. With the advent of safer methods of measuring cerebral blood flow and the development of noninvasive imaging techniques such as nuclear magnetic resonance imaging and positron-emission tomography, humans could be studied (*see* **Image Analysis and Tomography**). The intense pain that accompanies a myocardial infarction, bringing patients quickly to the hospital, is rarely present in patients with stroke. Patients attribute symptoms to "flu" or symptoms go unnoticed because observers believe the patient was "sleeping". Also, in previous years, when there was no available treatment, patients with strokes were a lower priority for transport by emergency medical systems. Thus, few patients have been available for study early in their disease. With the advent of a treatment for stroke that must be given early in the course of the stroke [33], and more education of the public about the signs and symptoms of stroke, there may be more patients available for study early in the course of their stroke. However, the introduction of effective therapeutic strategies presents the medical community with new challenges. Efficient detection of treatment effects among the diverse outcomes for patients with stroke is often difficult. Given the traditional focus on risk factors and secondary prevention where the outcome is presence or absence of stroke, and given the types of symptoms that characterize stroke, standard tools for evaluating different degrees of disability due to stroke have developed slowly.

Types of Study

Risk factors for stroke have been identified both retrospectively, using **case–control** methodology, and prospectively. Methods of preventing stroke, both initially in high-risk individuals and secondarily after an initial stroke, and approaches to stroke treatment have been studied using standard clinical trials methodology.

Landmark Studies

Prospective epidemiologic studies of stroke risk factors include the **Framingham Study** and Honolulu Heart Study that assessed stroke as an adjunct to cardiovascular disease, and studies in Japan, Sweden, Denmark, Norway, Finland, the Netherlands, and Australia. Whisnant et al. reference and discuss these epidemiologic studies in their report on a case–control study of stroke risk factors in Olmsted County, Minnesota [38]. Sacco [30] also provides a summary of some case–control studies of stroke risk

factors. Primary risk factors include age, gender, race/ethnicity, heredity, hypertension, diabetes mellitus, ischemic heart disease, transient ischemic attack, atrial fibrillation, mitral valve disease (other than prolapse), and current smoking (although not in Japan).

Landmark studies of stroke prevention included studies of both surgical and nonsurgical treatments. A major surgical trial was the Extracranial/Intracranial (EC/IC) Bypass Trial [10], notable because surgery was found to be no better than best medical care. Although the results of this trial generated great debate, the results were accepted by the medical community and the procedure is now rarely used in stroke prevention. Three other landmark primary stroke prevention trials of surgical procedures were the North American Symptomatic Carotid Endarterectomy Trial (NASCET) [25], European Carotid Surgery Trial (ECST) [12], and Asymptomatic Carotid Atherosclerosis Study (ACAS) [3]. In contrast to the EC/IC trial, these trials demonstrated the effectiveness of carotid endarterectomy in reducing the incidence of subsequent strokes for those with symptomatic and asymptomatic severe carotid artery stenosis (60–70%). Dyken [9] provides a summary of these and other landmark primary and secondary prevention studies through 1992, including the Physicians' Health Study [31]. The latter study, while a landmark for cardiovascular disease prevention, gave equivocal data on the usefulness of aspirin in preventing stroke, suggesting heart and brain are different in their response to treatment. A **meta-analysis of randomized trials** of cholesterol reduction provided another example of the potential difference between heart and brain. Based on their meta-analysis, Aitkins et al. concluded that lowering serum cholesterol did not reduce stroke morbidity or mortality in middle-aged men [1].

In the early 1990s several emergent stroke treatment trials were initiated. In 1995, the National Institute of Neurological Disorders and Stroke (NINDS) t-PA Stroke Treatment Trial investigators reported a beneficial effect for patients with acute ischemic strokes treated with tissue plasminogen activator (t-PA) within 180 minutes of stroke onset [33]. The genetically engineered drug, t-PA, was previously used to break up clots in patients with heart attacks. The success of t-PA opened the field of stroke to a plethora of new treatment trials aimed at salvaging brain function after an ischemic stroke. At the same time, new surgical procedures showed some promise in treating hemorrhagic stroke [22] and Nimodipine, a calcium channel blocker, was shown to be effective in treating subarachnoid hemorrhage [27].

Statistical Concepts, Problems, and Techniques

Outcome Measures

Standard statistical approaches for **prevention trials** and risk factor studies can be applied in stroke. For modeling recurrent stroke, Foulkes suggests that a linear **hazard rate** function may fit better than the more commonly used **proportional hazard** function or other parametric approaches [13] (*see* **Parametric Models in Survival Analysis**). For studies estimating incidence and mortality, the **biases** noted in many hospital or clinic-based studies pertain [11, 14, 32]. Stroke treatment trials present more of a challenge to statisticians. By contrast to cardiovascular trials, mortality is seldom considered the sole outcome for stroke treatment trials, since, as noted above, most patients survive. Of the outcome measures available in the late 1990s, there was no single accepted measurement for quantifying stroke-related disability or the course of stroke recovery. Many trials use the Barthel Index [21], a measure of functional status. Other common measures include the Rankin Scale [29] and the Glasgow Outcome Scale [17], more general summary measures than the Barthel Index, and outcomes such as the **National Institutes of Health** (NIH) Stroke Scale [6], the Canadian Neurological Scale [7], and the European Stroke Scale [16] that measure neurological function. The outcome measures are highly **correlated** but each scale or index gives slightly different information about the patient's disability [8, 24].

Quantification of infarct volume or brain function with methods such as computerized tomography or measures of cerebral blood flow has only duplicated with technology the immense complexity of the clinical assessment of neurological conditions. Measurements of infarct volume are intuitively simple, but in practice tediously difficult and, to date, of limited value compared with the relative ease of determining more clinically relevant functional outcomes.

Clinical interpretation of the numeric value of the scale score also presents challenges. For example, a

10-point difference in the **means** between two treatment groups for a scale such as the Barthel Index or NIH Stroke Scale is difficult for both patients and clinicians to interpret. Using scores as a continuum of numeric values requires that a value be placed on death, and scales such as the NIH Stroke Scale or Barthel Index do not provide such a value in their scoring system. Hallstrom describes one approach for stroke trials, based on developing a consensus among clinicians as to the score for death [15]. For scales measuring neurologic or physical disability, the overall summary score may have limited meaning as the scale is composed of multiple components, any one of which can represent a serious disability. To avoid interpretation of a continuous measure, some stroke trials require a specified degree of patient-specific improvement from baseline. In other trials, investigators dichotomize scales into favorable or unfavorable outcomes or into "cure" vs. "no cure", considering these outcomes more clinically meaningful [5, 33]. Categories or approaches to dichotomization should be chosen before starting the trial and they should represent clinically meaningful categories for the intervention being tested. A treatment such as t-PA that is expected to have severe side effects such as hemorrhage and potentially large benefits might be categorized differently than an intervention expected to have few expected side effects and moderate benefits. Dichotomization or categorization of scores allows deaths to be assigned to the category representing the worst outcome.

Methods of Analysis

The choice of the method of analysis is also an issue. A J-shaped or U-shaped distribution is typical for many outcome measures in stroke trials. Lesaffre et al. [20] give an example for the Barthel Index. Where there are equal sample sizes, skewness alone does not affect validity of the t-test (*see* **Student's *t* Statistics**) [23]. However, sample size calculations for the t-test and the **Wilcoxon–Mann–Whitney** nonparametric test suggest that a **binary** categorization of the data and analysis by **chi-square test** or other approaches such as **logistic regression** can reduce the required sample size (*see* **Sample Size Determination for Clinical Trials**) [35]. Logistic regression with multiple **ordered categories**, particularly when death is considered a category, should be considered cautiously in analyzing stroke-related data, as it is possible that underlying assumptions may be violated.

Global Tests

A global test provides a single test statistic for comparing groups that combines correlated information on multiple outcomes per individual. Global tests for continuous and **binary data** are described by O'Brien [26], Pocock et al. [28], Lefkopoulou et al. [18] and Legler et al. [19]. Global tests have application to stroke trials since, as noted previously, no single outcome measures all degrees of disability after stroke. Of interest is whether there is a preponderance of evidence that patients in the intervention group experience a better outcome than patients in the **control** group using multiple measures of disability. For dichotomized outcomes, we can calculate a Wald statistic (*see* **Likelihood**) using **generalized estimating equations (GEE)** with a log link to take correlations among multiple outcomes for a single individual into account. We can obtain an **odds ratio** and its **confidence limits** in addition to a ***P* value**. The odds ratio can be useful to clinicians in their interpretation of stroke trial data. The GEE approach also allows for inclusion of **covariates**, important for stroke trials where baseline covariates must be taken into account (*see* **Baseline Adjustment in Longitudinal Studies**). Tilley et al. [34] describe the use of global testing in the NINDS t-PA Stroke Trial [33].

The global approach differs from **Hotelling's T^2**, a statistic that tests whether there is a treatment association with multiple outcomes, despite the direction of the association across the outcomes, a question of little interest in stroke trials. The global approach to the analysis of stroke trial data also differs from the composite outcome often used for cardiovascular disease. In cardiovascular disease, we may consider a patient a treatment failure if the patient has a new myocardial infarction or dies or has severe angina or perhaps severe congestive heart failure requiring hospitalization. This is considered a composite outcome. For stroke trials we could construct a composite outcome by defining a category for failure on each scale of interest (Barthel, NIH Stroke Scale, Rankin, etc.). However, considering a patient a failure because of a failure on only one of multiple outcomes may set too stringent a criterion. Conversely, requiring a patient to have a favorable result on every outcome to be a success might be too

stringent. The **power** of a composite test as compared to the global test depends on the correlation among outcome measures, but is generally less than the power of the global test [18]. The power of the global test is greater than or equal to the power of any test of a single outcome included in the global test and power is greater than adjusting for multiple outcomes using a **Bonferroni** approach [28]. Thus the use of global tests in stroke trials can potentially reduce sample sizes, if the investigator is willing to have less power to detect differences in individual outcomes.

Anticipated Developments and Unresolved Problems

The primary unresolved problem in the study of stroke remains the definition of an outcome (*see* **Outcome Measures in Clinical Trials**). Given that t-PA has been shown to be effective for only a brief interval after stroke onset and has a risk of hemorrhage, many other treatments for stroke are now under investigation. These trials, even those that are not successful, will provide more information on stroke outcome measures and may lead to the development of new measures. There is also the potential for further development of existing outcome measures. Researchers using the SF-36 measure, which assesses health-related **quality of life** [37], have quantified scales that incorporate several dimensions of an outcome by calculating scores for each dimension measured by the scale rather than an overall score. As yet, such an approach to analyzing the Barthel Index or NIH Stroke Scale as a set of scores for different dimensions of disability has not been thoroughly evaluated. Also, as technology improves, as magnetic resonance imaging can be done more quickly and becomes more widely available, and as statistical methodology in this field develops, this technology may become more useful as an outcome measure or in identifying subgroups of stroke patients who could benefit from different types of therapies.

A potential analytic development relates to the measure of association in global testing. The use of the odds ratio as a measure of association in global testing has only partially solved the problem of clinical interpretation. **Relative risk** as a measure of association is sometimes preferred for clinical trials. Work is currently in progress to expand the work of Wacholder [36] to allow computation of relative risks rather than odds ratios when using global tests.

Another anticipated development will be the use of **Phase II trial** methodology in stroke. The Phase II approach in stroke is as yet untested and may identify potentially beneficial treatments more quickly for Phase III testing. Most stroke trials consider short-term outcomes (3–6 months after stroke) but have been using Phase III approaches to the design of pilot trials, adding to the time and cost of stroke research.

Because of the advent of a successful stroke treatment and the upsurge of interest in studying stroke, stroke will be a fertile area for biostatistical research.

References

[1] Aitkins, D., Psaty, B.M., Koepsell, T.D., Longstreth, W.T.J. & Larson, E.B. (1993). Cholesterol reduction and the risk for stroke in men: a meta-analysis of randomized controlled trials, *Annals of Internal Medicine* **119**, 136–145.

[2] American Heart Association (1992). *Stroke Facts*. American Heart Association, Dallas.

[3] Executive Committee for the Asymptomatic Carotid Atherosclerosis Study Group (1995). Endarterectomy for asymptomatic carotid artery stenosis, *Journal of the American Medical Association* **273**, 1421–1428.

[4] Barker, W.H. & Mullooly, J.P. (1997). Stroke in a defined elderly population, 1967–1985, a less lethal and disabling but no less common disease, *Stroke* **28**, 284–290.

[5] Brott, T., Haley, E.C., Levy, D.E., Barsan, W., Broderick, J., Sheppard, G.L., Spilker, J., Kongable, G., Massey, S., Reed, R. & Marler, J.R. (1992). Urgent therapy for stroke: Part I. Pilot study of tissue plasminogen activator administered within 90 minutes, *Stroke* **23**, 632–640.

[6] Brott, T., Adams, H.P., Olinger, C.P., Marler, J.R., Barson, W.G., Biller, J., Spilker, J., Holleran, R., Eberle, R., Hertzberg, V., Rorick, M., Moomaw, C.J. & Walker, M. (1993). Measurements of acute cerebral infarction: a clinical examination scale, *Stroke* **20**, 864–870.

[7] Cote, R. & Hachinski, V. (1996). The Canadian Neurological Scale: a preliminary study in acute stroke, *Stroke* **17**, 731–737.

[8] Deehan, R., Horn, J., Limburg, M., VanderMeulen, J. & Bossuyt, P. (1993). A comparison of five stroke scales with measures of disability, handicap and quality of life, *Stroke* **24**, 1178–1181.

[9] Dyken, M.L. (1993). Overview of trends in management ad prognosis of stroke, *Annals of Epidemiology* **31**, 535–540.

[10] EC/IC Bypass Study Group (1985). Failure of extra cranial-intracranial arterial bypass to reduce the risk of ischemic stroke: results of an international randomized trial, *New England Journal of Medicine* **313**, 1191–1200.

[11] Ellenberg, J.H. (1994). Observational data bases in neurological disorders: selection bias and generalization of results, *Neuroepidemiology* **13**, 268–274.

[12] European Carotid Surgery Trials' Collaborative Group (1991). MRC European Surgery Trial: interim results for symptomatic patients with severe (70–99%) or with mild (0–29%) carotid stenosis, *Lancet* **337**, 1235–1243.

[13] Foulkes, M.A., Sacco, R.L., Mohr, J.P., Hier, D.B., Price, T.R. & Wolf, P.A. (1994). Parametric modeling of stroke recurrence, *Neuroepidemiology* **13**, 19–27.

[14] Giroud, M., Lemesle, M., Quantin, C., Vourch, M., Becker, F., Milan, C., Brunet-Lecomte, P. & Dumas, R. (1997). A hospital-based and a population-based stroke registry yield different results: the experience in Dijon, France, *Neuroepidemiology* **16**, 15–21.

[15] Hallstrom, A.P., Litwin, P.E. & Weaver, W.D. (1992). A method of assigning scores to the components of a composite outcome: an example from the MITI trial, *Controlled Clinical Trials* **13**, 148–155.

[16] Hantson, L., De Weerdt, W., De Keyser, J., Diener, H.C., Franke, C., Palm, R., Van Orschoven, M., Schoonderwalt, H., De Klippel, N., Herroelen, L. & Feys, H. (1994). The European Stroke Scale, *Stroke* **11**, 2215–2219.

[17] Jennett, B. & Bond, M. (1975). Assessment of outcome after severe brain damage: a practical scale, *Lancet* **i**, 480–484.

[18] Lefkopoulou, M., Moore, D. & Ryan, L. (1989). The analysis of multiple correlated binary outcomes: application to rodent teratology experiments, *Journal of the American Statistical Association* **84**, 810–815.

[19] Legler, J.M., Lefkopoulou, M. & Ryan, L.M. (1995). Efficiency and power of tests for multiple binary outcomes, *Journal of the American Statistical Association* **90**, 680–693.

[20] Lesaffre, E., Scheys, I., Frohlich, J. & Blujmki, E. (1993). Calculation of power and sample size with bounded outcome scores, *Statistics in Medicine* **12**, 1063–1078.

[21] Mahoney, F.I. & Barthel, D.W. (1965). Functional evaluation: the Barthel Index, *Maryland State Medical Journal* **14**, 61–65.

[22] Mayberg, M.R., Batjer, H.H., Dacey, R., Diringer, M., Haley, C., Heros, R.C., Sternau, L.L., Torner, J., Adams, H.P., Feinberg, W. & Thies, W. (1994). Guidelines for the Management of Aneurysmal Subarachnoid Hemorrhage. A statement for health care professionals from a special writing group of the Stroke Council, American Heart Association, *Circulation*, **90**, 2592–2604.

[23] Miller, R.G., Jr (1986). *Beyond ANOVA: Basics of Applied Statistics*. Wiley, New York.

[24] Muir, K.W., Grosset, D.G. & Lees, K.R. (1994). Interconversion of stroke scales: implications for therapeutic trials, *Stroke* **25**, 1366–1370.

[25] NASCET Collaborators (1991). Beneficial effect of carotid endarterectomy in symptomatic patients with high grade carotid stenosis, *New England Journal of Medicine* **325**, 445–453.

[26] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics* **40**, 1079–1087.

[27] Pickard, J.D., Murray, G.D., Illingworth, R., Shaw, M.D.M., Teasdale, G.M., Foy, P.M., Humphrey, P.R.D., Lang, D.A., Nelson, R., Richards, P., Sinar, J., Bailey, S. & Skene, A. (1989). Effect of oral nimodipine on cerebral infarction and outcome after subarachnoid hemorrhage: British aneurysm nimodipine trial, *British Medical Journal* **298**, 636–642.

[28] Pocock, S.J., Geller, N.L. & Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics* **43**, 487–498.

[29] Rankin, J. (1957). Cerebral vascular accidents in patients over the age of 60: II. Prognosis, *Scottish Medical Journal* **2**, 200–215.

[30] Sacco, R.L. (1995). Risk factors and outcomes for ischemic stroke, *Neurology* **45**, S10–S14.

[31] Steering Committee of the Physician's Health Study Research Group (1989). Final report on the aspirin component of the ongoing physicians' health study, *New England Journal of Medicine* **321**, 129–135.

[32] Stegmayr, B. & Asplund, K. (1992). Measuring stroke in the population: quality of routine statistics in comparison with a population-based registry, *Neuroepidemiology* **11**, 204–213.

[33] The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group (1995). A controlled trial of recombinant tissue plasminogen activator administered within three hours for acute stroke, *New England Journal Medicine* **333**, 1581–1587.

[34] Tilley, B.C., Marler, J., Geller, N.L., Lu, M., Legler, J., Brott, T., Lyden, P. & Grotta, J. (1996). Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial, *Stroke* **27**, 2136–2142.

[35] Tilley, B.C. & Divine, G.W. (1985). Analytic approaches to stroke clinical trials: examples from the NINDS t-PA Stroke Trial, in *Ischemic Stroke: Stroke Clinical Trials, Issues in Design, Conduct, and Analysis*, J. Grotta, L.P. Miller, A.M. Buchan & C. Sussman, eds. International Business Communications, Southborough, Mass.

[36] Wacholder, S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences, *American Journal of Epidemiology* **123**, 174–184.

[37] Ware, J.E. & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection, *Medical Care* **30**, 473–483.

[38] Whisnant, J.P., Wiebers, D.O., O'Fallon, W.M., Sicks, J.D. & Frye, R.L. (1996). A population-based model of risk factors for ischemic stroke: Rochester, Minnesota, *Neurology*, **47**, 1420–1428.

[39] WHO Task Force on Stroke and Other Cerebrovascular Disorders (1989). Recommendations on stroke prevention, diagnosis, and therapy, *Stroke* **20**, 1407–1431.

[40] Wolf, P.A., Agostino, R.B., O'Neal, A., Sytkowski, P., Kase, C.S., Belanger, A.J. & Kannel, W. (1992). Secular trends in stroke incidence and mortality: the Framingham Study, *Stroke* **23**, 1551–1555.

BARBARA C. TILLEY & JOHN MARLAR

Structural and Sampling Zeros

Structural and sampling zeros are cells with zero frequency in **contingency table** models formed by the cross-classification of categorical variables. Structural zeros arise because of the impossibility of observing a given combination of categorical factors comprising a contingency table. That is, structural zeros occur with probability equal to one. In contrast, sampling zeros occur with probability less than one, which is a function of table size, sample size, and the patterns of association between the factors pertaining to the table.

The focus on structural and sampling zeros relates closely to **loglinear models** and associated lack-of-fit statistics. The ensuing results extend in a straightforward manner with minor modifications to **logistic regression** models involving categorical covariates, as these models are derived from loglinear models [9]. We do not address logistic models with continuous covariates here, since the corresponding contingency tables are typically too complex to consider in terms of sampling zeros.

Structural Zeros

Many examples of structural zeros occur in biomedical studies. For example, in the cross classification of sex with cancer type, structural zeros occur for the following combinations: females and testicular or prostate cancer, and males and ovarian or uterine cancer.

The ease of accommodating structural zeros depends upon the method of analysis. Commonly used data analysis packages such as SAS (*see* **Software, Biostatistical**), generally offer two different ways of analyzing contingency table data: (1) chi-square and associated statistics (e.g. exact tests and measures of association) based upon cross-classification tables (e.g. PROC FREQ in SAS); and (2) model-based analyses (e.g. PROC LOGISTIC in SAS; and the `glm` function in S-PLUS). With the first approach, it is difficult if not impossible to accommodate structural zeros. In contrast, the modeling approaches provide much flexibility in this regard. If the data analyst wants to impose a structural zero, then he or she may take one of at least three approaches: (i) omit the corresponding cases from the analysis; (ii) assign weights of zero to these cases and one to the remaining cases; or (iii) fit unique parameters to the cases corresponding to the structural zero. Under all of these approaches, the data analyst should determine any **collinearity** among covariates resulting from omitting structural zero cases. One may deal with such collinearities as in any regression context; for example by removing one or more covariates from the model. Such adjustments allow the degrees of freedom for a particular model to be computed in a conventional way by packages such as SAS (e.g. PROC GENMOD and PROC LOGISTIC). These adjustments for structural zeros assume that **maximum likelihood** or weighted **least squares** methods are employed to fit the models of interest. Under another estimation approach, **iterative proportional fitting**, one sets initial table frequencies for structural zeros to zero at the start of the algorithm, such that final estimates for the corresponding cells are zero. Bishop et al. [4] address this issue, in addition to discussing different patterns of structural zeros and corresponding degrees of freedom computation and closed form parameter estimation.

Sampling Zeros

Factors influencing the likelihood of sampling zeros include the number and type of factors that constitute a contingency table, and/or the size and nature of the sample of observations contributing information to the table. For example, cross-classifying 100 patients by sex, age intervals, and a wide range of cancers is more likely to yield an observed zero count for a particular cell, such as breast cancer in young men, than cross-classifying 100 subjects according to sex and a binary cancer classification (e.g. lung cancer vs. no lung cancer). The size of the three-dimensional table, the rare occurrence of breast cancer in young men, and the limited sample size increase the probability of a sampling zero relative to the analogous probability in the two-dimensional table for lung cancer and gender. Increasing the size of the sample by a factor of ten would reduce the probability of sampling zeros in each case.

Accommodating sampling zeros in the analysis of contingency tables has been a topic of debate [2, 3, 5, 6, 9, 10]. The focus of this dialog has been on the computation of degrees of freedom for loglinear

models fitted to contingency tables under sampling zeros and the consequences for related chi-square statistics for lack of fit. The discussion by Haslett [9] of the different approaches for handling sampling zeros serves as the basis for the ensuing review.

The debate over sampling zeros has centered on the first of two ways in which sampling zeros impact degrees of freedom calculations: (1) by causing zeros in marginal tables; and (2) by causing pathologic sampling zeros [4], or what Haslett [9] calls "parity-based inestimability". The latter zeros will not be discussed further, as algorithms for detecting them do not appear to be available, and they impact only the highest-order interactions of saturated loglinear models, i.e. those models that yield expected frequencies equaling observed table frequencies.

A marginal table is formed by collapsing a multidimensional table down to a table representing the cross classification of fewer factors. This is achieved by summing over the categories or levels of the factors left out of the marginal table. For example, starting with the cross classification of three factors, sex, age, and cancer, one may form a marginal table for sex and cancer by summing the frequencies across all age categories for each sex–cancer combination.

Zero cells in marginal tables have two ramifications: (i) causing estimated cell frequencies to be zero, that is, fitted zero frequencies; and (ii) leading to the lack of existence of parameter estimates under a loglinear model, that is, to inestimable parameters [7] (*see* **Identifiability**). The two methods that account for sampling zeros differ in how they treat these two consequences of sampling zeros in the computation of degrees of freedom. The first method (method 1) subtracts the number of parameters in a given model, adjusted for inestimable parameters, from the number of zero cells in the table, adjusted for the number of fitted zero frequencies [2, 4, 5, 7]. The second method (method 2) adjusts only for inestimable parameters, ignoring fitted zero frequencies [9, 10]. Specifically, under this approach one subtracts the number of parameters of a given model, adjusted for inestimable parameters, from the adjusted number of parameters in the corresponding saturated model. Finally, a third point of view [3] advocates ignoring sampling zeros altogether in computing degrees of freedom, arguing that degrees of freedom should depend upon whether expected frequencies are zero (i.e. structural zeros) rather than whether estimated or observed frequencies are zero. This view is based upon the fact that parameter estimates always exist under certain models other than loglinear models. The focus of the debate on sampling zeros has been on the first two methods.

The distinction between methods 1 and 2 for adjusting degrees of freedom – that is, adjusting or not adjusting for fitted zero frequencies – relates to the method of fitting loglinear models. Typically, estimation procedures used for loglinear models are unconditional, in that they do not condition on the frequency counts in marginal tables, as opposed to conditional estimation approaches; for example, exact estimation procedures. Stirling [10] and Haslett [9] maintain that method 1 is appropriate for conditional estimation but not for unconditional estimation. That is, they assert that adjusting for fitted zero frequencies is necessary only if one is conditioning on marginal counts. Exact conditional tests based upon the **hypergeometric distribution** are not impacted by sampling zeros, as asymptotic distributions based upon degrees of freedom are not an issue for these tests.

Methods 1 and 2 may be extended to logistic regression models fitted conditionally (*see* **Logistic Regression, Conditional**) or unconditionally. Aston & Wilson [2] present method 1 calculations for a logistic model fitted conditionally to an example data set. Haslett [9] discusses the extension of method 2 to logistic models fitted unconditionally, while contrasting the two-degree-of-freedom methods in the context of conditional and unconditional likelihood estimation.

Parameter inestimability under loglinear models, which is a major factor in the debate about calculating degrees of freedom, is attributable to sampling zeros in marginal tables [1, 8]. To circumvent the problem of inestimable parameters due to marginal sampling zeros, some statistical packages (e.g. SAS PROC CATMOD and PROC FREQ) allow the user to add a small constant (e.g. 0.5) to every cell in a contingency table with sampling zeros. In general, this results in forcing estimates of association parameters closer to their null values (e.g. log **odds ratio** estimates are closer to zero). Agresti [1] suggests trying different constants that are very small (e.g. 10^{-8}) to assess the resulting sensitivity of parameter estimates and lack-of-fit statistics.

In addition to impacting the existence of parameter estimates and degrees of freedom of chi-square lack-of-fit statistics, sampling zeros have other effects. By diminishing expected cell frequencies, sampling zeros may affect the adequacy of chi-square approximations to the sampling distributions of lack-of-fit statistics. Agresti [1] provides a review of the literature on such effects of sparse tables in general. The bias of parameter estimates under the loglinear model may also be affected by sampling zeros, as this bias is a function of the size of the marginal table cells corresponding to the parameters of interest [8].

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Aston, C.E. & Wilson, S.R. (1984). Comment on M.B. Brown and C. Fuchs, "On maximum likelihood estimation in sparse contingency tables", *Computational Statistics and Data Analysis* **2**, 71–77.

[3] Baker, R.J., Clarke, M.R.B. & Lane, P.W. (1985). Zero entries in contingency tables, *Computational Statistics and Data Analysis* **3**, 33–45.

[4] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

[5] Brown, M.B. & Fuchs, C. (1983). On maximum likelihood estimation in sparse contingency tables, *Computational Statistics and Data Analysis* **1**, 3–15.

[6] Brown, M.B. & Fuchs, C. (1984). Rejoinder on Comment by C.E. Aston and S.R. Wilson, *Computational Statistics and Data Analysis* **2**, 79–80.

[7] Fienberg, S.E. (1977). *The Analysis of Categorical Data*. MIT Press, Cambridge, Mass.

[8] Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

[9] Haslett, S. (1990). Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables, *Computational Statistics and Data Analysis* **9**, 179–195.

[10] Stirling, W.G. (1986). A note on degrees of freedom in sparse contingency tables, *Computational Statistics and Data Analysis* **4**, 67–70.

(*See also* **Chi-square Tests**; **Square Contingency Table**; **Quasi-independence**)

THOMAS R. TEN HAVE

Structural Equation Models

Structural equation models refer to general statistical procedures for multiequation systems that include continuous latent variables ("factors" or "unmeasured variables"), multiple indicators of concepts, errors of measurement, errors in equations, and observed variables that are continuous, ordinal, dichotomous (**binary**), or **censored** (*see* **Measurement Scale**). One way to view these models is as an interrelated system of **regression** equations where some of the variables have multiple measures, and where measurement error is taken into account when estimating relationships (*see* **Errors in Variables**). From another perspective, these are **factor analysis** models in which some factor loadings are restricted to zero or other constants, and the researcher allows factors to affect each other, directly and indirectly. The most general form of the structural equation model encompasses **analysis of variance** (ANOVA), **analysis of covariance** (ANOCOVA), **multiple linear regression**, **multivariate multiple regression**, seemingly unrelated regressions, recursive and nonrecursive simultaneous equations, **path analysis**, **confirmatory factor analysis**, classical test theory (*see* **Psychometrics, Overview**), dichotomous and ordinal probit, tobit, and a variety of other procedures (*see* **Quantal Response Models**) as special cases. Occasionally, the term "structural equation model" refers to the simultaneous equation models of classical econometrics. Increasingly, though, it has come to refer to its more general form. *Covariance structure models, LISREL models, analysis of moment structures, and structural equations with latent or unobserved variables* are largely interchangeable terms for structural equation models.

Model and Notation

The structural equation models are represented in a variety of notations, but researchers most commonly use the one derived from Jöreskog [19, 20], Keesling [23], and Wiley [31]. It is called the **LISREL** notation, named after Jöreskog & Sörbom's [22] **software** package. The model has two primary components, a latent variable and a measurement model. The latent variable model is

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{1}$$

where $\boldsymbol{\eta}$ is an $m \times 1$ vector of latent endogenous variables, $\boldsymbol{\xi}$ is an $n \times 1$ vector of latent exogenous variables, $\boldsymbol{\alpha}$ is an $m \times 1$ vector of intercept terms, $\mathbf{B}$ is an $m \times m$ matrix of coefficients that give the influence of the $\boldsymbol{\eta}$s on each other, $\boldsymbol{\Gamma}$ is an $m \times n$ matrix of coefficients for the effect of the $\boldsymbol{\xi}$ on $\boldsymbol{\eta}$, and $\boldsymbol{\zeta}$ is the $m \times 1$ vector of disturbances that contains the unexplained parts of the $\boldsymbol{\eta}$s. The term "endogenous" refers to variables that are influenced by other variables in the model. "Exogenous" describes variables that are determined outside of the system of equations.

The model assumes that $\mathrm{E}(\boldsymbol{\zeta}) = \mathbf{0}$, that $\mathrm{cov}(\boldsymbol{\xi}, \boldsymbol{\zeta}') = \mathbf{0}$, and that $(\mathbf{I} - \mathbf{B})$ is nonsingular. The **covariance matrix** of the latent exogenous variables is represented by an $n \times n$ matrix, $\boldsymbol{\Phi}$, and the $m \times m$ matrix $\boldsymbol{\Psi}$ is the covariance matrix of the equation disturbances, $\boldsymbol{\zeta}$. Implicit in (1) is a subscript to index the observations. Since the same model holds for all cases, the subscript is omitted to simplify the notation.

The traditional LISREL notation has two equations for the measurement model:

$$\mathbf{y}^* = \boldsymbol{\nu}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{2}$$

$$\mathbf{x}^* = \boldsymbol{\nu}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{3}$$

where $\mathbf{y}^*$ is the $p \times 1$ vector of indicators of the latent variables in $\boldsymbol{\eta}$, $\boldsymbol{\nu}_y$ is the $p \times 1$ vector of intercept terms, $\boldsymbol{\Lambda}_y$ is the $p \times m$ **factor loading matrix** of coefficients giving the linear effect of $\boldsymbol{\eta}$ on $\mathbf{y}^*$, and $\boldsymbol{\varepsilon}$ is the $p \times 1$ vector of measurement errors or disturbances. The model assumes that $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and that $\mathrm{cov}(\boldsymbol{\eta}, \boldsymbol{\varepsilon}') = \mathbf{0}$. The covariance matrix for $\boldsymbol{\varepsilon}$ is the $p \times p$ matrix, $\boldsymbol{\Theta}_\varepsilon$. Analogous definitions and assumptions hold for (3), with $\boldsymbol{\Theta}_\delta$ the $q \times q$ covariance matrix for errors in $\mathbf{x}^*$. In addition, we assume that $\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$, and $\boldsymbol{\zeta}$ are mutually uncorrelated (*see* **Correlation**). Here too the observation index is omitted, but is implicit. The disturbance or error term for each equation in the latent variable or measurement model typically has different variances for different equations but, for a single equation, the assumption is that the disturbance's variance is homoscedastic (*see* **Scedasticity**) and uncorrelated across observations.

Eqs. (1)– (3) make up the classical form of the LISREL model, in which it is assumed that

the latent and observed variables are continuous variables. More recently, Jöreskog & Sörbom [21], Muthén [26], and others have generalized the model by allowing **categorical** or censored observed variables. In this case, some of the variables in $\mathbf{y}^*$ and $\mathbf{x}^*$ are "latent indicators" that are only observable through categorical or censored observed variables. Here the model requires an additional set of equations to link the observed variables to their underlying continuous counterparts:

$$\mathbf{y} = f(\mathbf{y}^*, \boldsymbol{\tau}_y), \tag{4}$$

$$\mathbf{x} = f(\mathbf{x}^*, \boldsymbol{\tau}_x), \tag{5}$$

where $\mathbf{y}$ and $\mathbf{x}$ are the vectors of observed variables, some or all of which can be categorical or censored, and $\boldsymbol{\tau}_y$ and $\boldsymbol{\tau}_x$ are vectors that contain threshold parameters that determine the values taken by $\mathbf{y}$ and $\mathbf{x}$, respectively. For instance, suppose that y_1 is a four-category ordinal variable. In this case, we would have

$$y_1 = \begin{cases} 1, & \text{if } y_1^* \le \tau_{01}, \\ 2, & \text{if } \tau_{01} < y_1^* \le \tau_{11}, \\ 3, & \text{if } \tau_{11} < y_1^* \le \tau_{21}, \\ 4, & \text{if } \tau_{21} < y_1^*, \end{cases} \tag{6}$$

where τ_{01}, τ_{11}, and τ_{21} are the three thresholds that determine whether the ordinal y_1 variable falls in the 1, 2, 3, or 4 category. Alternately, if the y_4 variable is censored from below, we would have

$$y_4 = \begin{cases} 0, & \text{if } y_4^* \le 0, \\ y_4^*, & \text{if } 0 < y_4^*. \end{cases} \tag{7}$$

The single threshold point is zero and when y_4^* is above zero, y_4^* and the observed y_4 are the same. If the observed variables are continuous, we have no need for (4) and (5). But when we have noncontinuous observed variables, (4) and (5) are nonlinear, deterministic equations that relate the observed variables to their underlying continuous indicator.

Many of the more familiar statistical models are derivable from this general model. Table 1 illustrates how restrictions on the general model can lead to more familiar techniques. If, for instance, we assume a scalar, continuous dependent **response variable**, no measurement error in the dependent or **explanatory variables**, and only **dummy** explanatory variables, we are led to the restrictions shown in the first row that leads to **analysis of variance** (ANOVA). Keeping the same restrictions, except allowing continuous or dummy explanatory variables, leads to multiple regression. Probit regression has the same constraints

Table 1 Common statistical models as special cases of structural equation models (SEMs)

Statistical model	$\boldsymbol{\nu}_y$	$\boldsymbol{\Lambda}_y$	$\boldsymbol{\Theta}_\varepsilon$	$\boldsymbol{\nu}_x$	$\boldsymbol{\Lambda}_x$	$\boldsymbol{\Theta}_\delta$	$\mathbf{y}$
ANOVA	0	1	0	0	$\mathbf{I}$	0	$= y^*$, scalar
Multiple regression	0	1	0	0	$\mathbf{I}$	0	$= y^*$, scalar
Probit regression	0	1	0	0	$\mathbf{I}$	0	$1, 2, \ldots, k$
Tobit regression	0	1	0	0	$\mathbf{I}$	0	$= 0$ if $y^* \le 0$ $= y^*$ if $y^* > 0$
Classical econometrics	0	$\mathbf{I}$	0	0	$\mathbf{I}$	0	$= \mathbf{y}^*$
Classical factor analysis (deviation scores)	–	–	–	–	√	Diagonal	–
Confirmatory factor analysis	–	–	–	–	√	√	–

Statistical model	$\boldsymbol{\tau}_y$	$\mathbf{x}$	$\boldsymbol{\tau}_x$	$\boldsymbol{\alpha}$	$\mathbf{B}$	$\boldsymbol{\Gamma}$	$\boldsymbol{\Psi}$
ANOVA	–	Dummy variable	–	Scalar	0	√	Scalar
Multiple regression	–	Dummy/continuous	–	Scalar	0	√	Scalar
Probit regression	$(k-1)$	Dummy/continuous	–	Scalar	0	√	Scalar
Tobit regression	$= 0$	Dummy/continuous	–	Scalar	0	√	Scalar
Classical econometrics	–	Dummy/continuous	–	√	√	√	√
Classical factor analysis (deviation scores)	C –	Continuous	–	–	–	–	–
Confirmatory factor analysis	–	Continuous	–	–	–	–	–

√, Present in model; – , absent from model.

as multiple regression, except that we have a dichotomous or ordinal dependent variable (*see* **Ordered Categorical Data**). Classical econometrics is a special case of the model that assumes perfect measurement and the absence of multiple indicators. From this perspective, structural equation models have less restrictive assumptions than many better known procedures. In addition, we can estimate many models that are not treated by the traditional procedures.

Steps in Modeling

An analysis that uses structural equation models has several components to it. These concern (i) model specification, (ii) the implied moment matrix, (iii) identification, (iv) estimation, (v) model–data fit assessment, and (vi) respecification. These are examined in the next six subsections.

Model Specification

The first step is to specify the hypothesized relations between all latent and observed variables. In other words, the researcher needs to describe the specific form that all the matrices in Table 1 take for the specific example of interest. Typically, not all of the matrices are required, so that the task is simplified. However, model specification requires the substantive expertise of the analyst to be able to formulate a set of restrictions that defines the model. A person who has little or no knowledge about the substantive area will not fare well with structural equation models (*see* **Model, Choice of**).

A path diagram for a hypothetical example, to illustrate model specification, is shown in Figure 1. A more detailed description of path diagrams is given in the article on **path analysis**, but in brief it provides a pictorial representation of the multiequation model that a researcher specifies. The ovals or circles enclose the latent variables, boxes signify observed variables, and the disturbances and error terms are not enclosed. The single-headed straight arrows indicate a linear impact of the variable at the base of the arrow on the variable at the head of the arrow. Curved two-headed arrows show linear covariances (correlations) between variables that are not explained within the model and they signify the covariances between exogenous variables or between disturbances/errors. To simplify it, the diagram does not include the

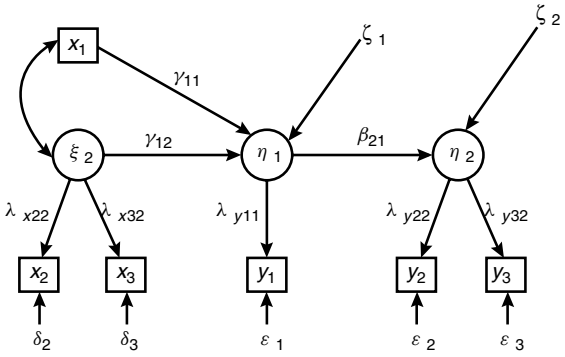

Figure 1 An hypothetical example of a structural equation model with three latent variables and six observed variables

regression constants that enter the equations for each endogenous variable.

An alternative to the path diagram is to represent the model specification using (1)–(3). In this specific example, (1) is

$$\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix} + \begin{bmatrix}0 & 0\\ \beta_{21} & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\gamma_{11} & \gamma_{12}\\ 0 & 0\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix} + \begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix}, \quad (8)$$

(2) is

$$\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} = \begin{bmatrix}0\\ \nu_{x2}\\ \nu_{x3}\end{bmatrix} + \begin{bmatrix}1 & 0\\ 0 & \lambda_{x22}\\ 0 & \lambda_{x32}\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix} + \begin{bmatrix}0\\ \delta_2\\ \delta_3\end{bmatrix}, \quad (9)$$

and (3) is

$$\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix} = \begin{bmatrix}\nu_{y_1}\\ \nu_{y_2}\\ \nu_{y_3}\end{bmatrix} + \begin{bmatrix}\lambda_{y11} & 0\\ 0 & \lambda_{y22}\\ 0 & \lambda_{y32}\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{bmatrix}. \quad (10)$$

The observed variables are continuous, so that (4) and (5) are $\mathbf{x} = \mathbf{x}^*$ and $\mathbf{y} = \mathbf{y}^*$. These relations are substituted into the above three equations. This also means that we will not need any threshold parameters ($\boldsymbol{\tau}_x, \boldsymbol{\tau}_y$) in the model. In addition, (9) shows that x_1 is perfectly measured (i.e., $x_1 = \xi_1$) and this explains the use of x_1 in place of ξ_1 in the path diagram in Figure 1. The covariance matrices of the exogenous variables and disturbances/errors are not represented in path diagrams. In the example, these

matrices are

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix},$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix}, \tag{11}$$

$$\boldsymbol{\Theta}_\delta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Theta_{\delta 22} & 0 \\ 0 & 0 & \Theta_{\delta 33} \end{bmatrix},$$

$$\boldsymbol{\Theta}_\varepsilon = \begin{bmatrix} \Theta_{\varepsilon 11} & 0 & 0 \\ 0 & \Theta_{\varepsilon 22} & 0 \\ 0 & 0 & \Theta_{\varepsilon 33} \end{bmatrix}. \tag{12}$$

The zero in the (1,1) position of $\boldsymbol{\Theta}_\delta$ follows since x_1 contains no measurement error. If we had correlated errors of measurement specified, then some of the off-diagonal elements of $\boldsymbol{\Theta}_\delta$ or $\boldsymbol{\Theta}_\varepsilon$ would contain free parameters rather than zeros.

Implied Moment Matrix

Once a model is specified, it implies that the first and second **moments** (means, variances, and covariances) of the observed variables are functions of the model parameters. Few have examined the higher-order moments of the observed variables and their relation to the model parameters. Most structural equation models focus on the implied covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The general expression for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ comes from the cov($\mathbf{z}^*$, $\mathbf{z}^{*\prime}$), where $\mathbf{z}^{*\prime}$ is [$\mathbf{y}^{*\prime}\mathbf{x}^{*\prime}$]. The linkage to the model parameters comes from substitution of (3) in for $\mathbf{x}^*$ and the reduced form equation in for $\mathbf{y}^*$,

$$\mathbf{y}^* = \boldsymbol{\nu}_y + \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}) + \boldsymbol{\varepsilon}. \tag{13}$$

The reduced form of an equation results by solving the right-hand side of the equation, so that it contains only exogenous variables, disturbances, errors, and coefficient matrices. After these substitutions and taking the cov($\mathbf{z}^*$, $\mathbf{z}^{*\prime}$), the implied covariance matrix is

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{C}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})\mathbf{C}' + \boldsymbol{\Theta}_\varepsilon & \mathbf{C}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Gamma}\mathbf{C}' & \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta \end{bmatrix}, \tag{14}$$

where $\mathbf{C} = \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}$ and $\boldsymbol{\theta}$ is the $t \times 1$ vector that contains all of the model parameters to be estimated in a given model. The upper left quadrant of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the implied covariance matrix for $\mathbf{y}^*$ (called $\boldsymbol{\Sigma}_{\mathbf{y}^*\mathbf{y}^*}(\boldsymbol{\theta})$), the lower right quadrant is the implied covariance matrix for $\mathbf{x}^*$ ($\boldsymbol{\Sigma}_{\mathbf{x}^*\mathbf{x}^*}(\boldsymbol{\theta})$), and the off-diagonal quadrants are the implied covariance matrices for $\mathbf{y}^*$ with $\mathbf{x}^*$($\boldsymbol{\Sigma}_{\mathbf{x}^*\mathbf{y}^*}(\boldsymbol{\theta})$).

The implied mean vector, $\boldsymbol{\mu}(\boldsymbol{\theta})$, is

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\nu}_y + \mathbf{C}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa}) \\ \boldsymbol{\nu}_x + \boldsymbol{\Lambda}_x\boldsymbol{\kappa} \end{bmatrix}, \tag{15}$$

where $\boldsymbol{\kappa}$ equals the mean vector of $\boldsymbol{\xi}$ [E($\boldsymbol{\xi}$) = $\boldsymbol{\kappa}$]. These general expressions for the first and second implied moments apply to any specific model. For the model in Figure 1, the implied covariance matrix for $\mathbf{y}^*$ is

$$\boldsymbol{\Sigma}_{\mathbf{y}^*\mathbf{y}^*}(\boldsymbol{\theta}) = \begin{bmatrix} \lambda_{y11}^2\text{var}(\eta_1) + \Theta_{\varepsilon 11} & & \\ \lambda_{y11}\lambda_{y22} \times \beta_{21}\text{var}(\eta_1) & \lambda_{y22}^2\text{var}(\eta_2) + \Theta_{\varepsilon 22} & \\ \lambda_{y11}\lambda_{y32} \times \beta_{21}\text{var}(\eta_1) & \lambda_{y22}\lambda_{y32} \times \text{var}(\eta_2) & \lambda_{y32}^2\text{var}(\eta_2) + \Theta_{\varepsilon 33} \end{bmatrix}, \tag{16}$$

with

$$\boldsymbol{\Sigma}_{\mathbf{x}^*\mathbf{y}^*}(\boldsymbol{\theta}) = \begin{bmatrix} \lambda_{y11} \times \text{cov}(\xi_1, \eta_1) & \lambda_{y22}\beta_{21} \times \text{cov}(\xi_1, \eta_1) & \lambda_{y32}\beta_{21} \times \text{cov}(\xi_1, \eta_1) \\ \lambda_{x22}\lambda_{y11} \times \text{cov}(\xi_2, \eta_1) & \lambda_{x22}\lambda_{y22}\beta_{21} \times \text{cov}(\xi_2, \eta_1) & \lambda_{x22}\lambda_{y32}\beta_{21} \times \text{cov}(\xi_2, \eta_1) \\ \lambda_{x32}\lambda_{y11} \times \text{cov}(\xi_2, \eta_1) & \lambda_{x32}\lambda_{y22}\beta_{21} \times \text{cov}(\xi_2, \eta_1) & \lambda_{x32}\lambda_{y32}\beta_{21} \times \text{cov}(\xi_2, \eta_1) \end{bmatrix}, \tag{17}$$

and

$$\boldsymbol{\Sigma}_{\mathbf{x}^*\mathbf{x}^*}(\boldsymbol{\theta}) = \begin{bmatrix} \phi_{11} & & \\ \lambda_{x22}\phi_{21} & \lambda_{x22}^2\phi_{22} + \Theta_{\delta 22} & \\ \lambda_{x32}\phi_{21} & \lambda_{x32}\lambda_{x22}\phi_{22} & \lambda_{x22}^2\phi_{22} + \Theta_{\delta 33} \end{bmatrix}, \tag{18}$$

where

$$\text{var}(\eta_1) = \gamma_{11}^2\phi_{11} + 2(\gamma_{11}\gamma_{12}\phi_{12}) + \gamma_{12}^2\phi_{22} + \psi_{11},$$

$$\begin{aligned} \text{var}(\eta_2) &= \beta_{21}^2 \text{var}(\eta_1) + \psi_{22}, \\ \text{cov}(\eta_1, \xi_1) &= \gamma_{11}\phi_{11} + \gamma_{12}\phi_{12}, \\ \text{cov}(\eta_1, \xi_2) &= \gamma_{11}\phi_{12} + \gamma_{12}\phi_{22}. \end{aligned} \tag{19}$$

With these implied moment matrices we have a one-to-one relation between a mean, variance, or covariance of the observed variables and a function of the parameters in the model. For instance, in the model the variance of x_2 equals $\lambda_{x22}^2\phi_{22} + \Theta_{\delta 22}$. These connections are critical to the issues of model identification, estimation, and fit assessment.

Model Identification

Model identification concerns the question whether it is possible to determine uniquely the parameters of a model from the means, variances, and covariances of the observed variables (*see* **Identifiability**). The last section gave the relation between these moments and the model parameters. Identification concerns whether it is possible to uniquely solve for the model parameters in terms of the moments of the observed variables using these equations. To illustrate this point, consider a simple example with a single observed variable, x_1, that equals $\xi_1 + \delta_1$. Here $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ has a single element, $\phi_{11} + \Theta_{\delta 11}$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$ is κ_1, the mean of ξ_1. The only second moment of the observed variable is the population variance of x_1, and the single first moment element is the population mean of x_1, μ_{x1}. The mean of x_1 identifies κ_1, but the single variance for x_1 is insufficient to identify the two parameters, ϕ_{11} and $\Theta_{\delta 11}$. For any given value of the variance of x_1, an infinite set of values of ϕ_{11} and $\Theta_{\delta 11}$ would satisfy the equation for the implied variance. The model is underidentified. More generally, if in a model $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$ are any two sets of values for $\boldsymbol{\theta}$ such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}_a) = \boldsymbol{\Sigma}(\boldsymbol{\theta}_b)$ and $\boldsymbol{\mu}(\boldsymbol{\theta}_a) = \boldsymbol{\mu}(\boldsymbol{\theta}_b)$, then $\boldsymbol{\theta}_a = \boldsymbol{\theta}_b$ must be true if the model is identified.

A necessary but not sufficient condition for identifying a model is that the researcher must assign a scale to each latent variable that is measured with error. One way to do this is to choose an indicator for each latent variable and set the coefficient or factor loading for the indicator to one. The intercept for the same observed variable should be set to zero. With this scaling, the latent variable has a metric that is similar to that of the observed variable. In the model in Figure 1, for instance, we could set λ_{x11}, λ_{y11}, and λ_{y22} to 1, and set ν_{x11}, ν_{y11}, and ν_{y22} to zero to assign scales to ξ_1, η_1, and η_2. An alternate method to scale the latent variable is to set the variance of the latent variable to one and its mean to zero. This latter option is less desirable when analyzing panel data or when testing whether models are the same across different groups.

Establishing model identification in the general structural equation model can be difficult. Algebraic manipulation of the implied moment equations can sometimes establish that each model parameter has a unique solution in terms of the means, variances, or covariances of the observed variables. In complicated models, this becomes less feasible. In special cases, such as the classical econometric model or confirmatory factor analysis, there are rules of identification that are helpful or that researchers can combine to establish model identification (see, for example, Fisher [13]; Bollen [6, pp. 88–104, 238–254, 326–333]). Also widely used are empirical checks on model identification that are based on whether the information matrix of the model parameters from a **maximum likelihood** solution is nonsingular. Singularity suggests that the model is underidentified. In most cases the empirical tests of identification work well, but it is possible for them to fail (see, for example, Bollen [6, pp. 246–251]).

Estimation

The earliest developments of structural equation models assumed that $\mathbf{y}$ and $\mathbf{x}$ were continuous and multinormally distributed. The maximum likelihood estimator under this assumption is

$$\begin{aligned} F_{\text{ML}} = {} & \ln|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) \\ & + (\bar{\mathbf{z}} - \boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})(\bar{\mathbf{z}} - \boldsymbol{\mu}(\boldsymbol{\theta})) \\ & - \ln(|\mathbf{S}|) - (p+q), \end{aligned}$$

where $\mathbf{S}$ is the sample covariance matrix of the observed variables and $\bar{\mathbf{z}}$ is the vector of sample means of the observed variables. Numerical minimization procedures find the $\hat{\boldsymbol{\theta}}$ that minimizes F_{ML}. The $\hat{\boldsymbol{\theta}}$ has the usual maximum likelihood estimator properties of being asymptotically **unbiased**, asymptotically **efficient**, **consistent**, asymptotically normal, and an asymptotic covariance matrix that is the inverse of the **information matrix** of $\hat{\boldsymbol{\theta}}$ (*see* **Large-sample Theory**).

Fortunately, the $\hat{\boldsymbol{\theta}}$ from F_{ML} retains many of its desirable properties under some conditions when $\mathbf{y}$

and $\mathbf{x}$ are not from multinormal distributions. For instance, if $\mathbf{x} = \boldsymbol{\xi}$ – that is, $\mathbf{x}$ is exogenous – then the usual properties hold assuming that the disturbances, $\boldsymbol{\zeta}$ and $\boldsymbol{\varepsilon}$, are from multinormal distributions. Even if $\mathbf{x}$ does not equal $\boldsymbol{\xi}$ and $\mathbf{y}$ and $\mathbf{x}$ are nonnormal, $\hat{\boldsymbol{\theta}}$ remains a consistent estimator. Corrections to the asymptotic standard errors from the usual maximum likelihood procedures also are available (see, for example, [9] and [28]). Furthermore, there are **robustness** conditions under which the usual maximum likelihood asymptotic **standard errors** and significance tests hold for observed variables from nonnormal distributions (see, for example, Sattora [27]).

Another class of estimators are explicitly designed to take account of nonnormality rather than relying on robustness conditions or correcting standard errors. For instance, Browne [9] proposed an asymptotically distribution free estimator (also called the weighted **least squares**) that applies to observed variables from distributions with finite eighth-order moments. Although the estimator appears to work well in moderately large samples with models that do not involve many parameters, the performance of this estimator has been disappointing in large models. An **instrumental variable** estimator, **two-stage least squares**, is a limited information estimator that also does not require observed variables from multinormal distributions. Hägglund [16] developed this estimator for factor analysis models with uncorrelated errors of measurement. Recent work proposed a two-stage least squares estimator for all the coefficients of both the measurement model and the latent variable model with or without correlated errors of measurement [7]. The two-stage least squares estimator has known asymptotic properties including standard errors that allow significance tests without assuming multivariate normality of the observed variables. The finite sample properties of the estimator are not well studied in latent variable models.

The estimation of parameters is more complicated when some of the endogenous observed variables are categorical. This would be the case if the indicators of a latent variable are ordinal, censored, or dichotomous variables, or in other situations in which the "dependent" variable of a relationship is noncontinuous. Analysts have proposed several approaches to incorporate such variables into a structural equation model, but they all share a similar strategy. It is assumed that all observed endogenous noncontinuous variables have underlying continuous variables that correspond to them. So, for example, we assume that underlying our five-point ordinal scale on self-reported health is a continuous variable of perceived health. The first step is the estimation of the correlation (covariance) matrix of the continuous variables that underlie the noncontinuous observed variables. The next step takes this matrix and analyzes it with the arbitrary distribution function (weighted least squares) estimator. Thus the main difference when endogenous categorical variables are part of the analysis is that we take the extra step of estimating what the correlation (covariance) matrix would look like if these variables were measured on continuous scales.

Model–Data Fit

Once the researcher estimates a model, attention turns to assessing its **goodness of fit**. Model fit assessments have two parts: (1) overall fit and (2) component fit. Overall fit refers to summary measures of how well the model as a whole corresponds to the data. The most widely used measure of overall fit is a test statistic (*see* **Hypothesis Testing**) that asymptotically approaches a **chi-square distribution** when the population covariance matrix equals the implied covariance matrix; that is, the null hypothesis is $\mathrm{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. In the case of F_{ML} described above, the test statistic is $T = (N-1)F_{ML}$, evaluated at the final parameter estimates. The **degrees of freedom** equal $\frac{1}{2}(p+q)(p+q+3) - t$, where p and q are the number of y and x variables and t is the number of unrestricted parameters estimated. The first term gives the number of nonredundant elements in the covariance matrix of the observed variables and the number of sample means. If the distributional assumptions of the test are satisfied, a significant value of the test statistic suggests that the model is misspecified (*see* **Misspecification**). In large samples the **power** of the significance tests is sometimes so great that even trivial departures lead to rejection of H_0. In small samples, the power of the test might be too weak to detect problems.

In response to these difficulties a variety of other overall fit measures have arisen. The **residual** matrices, $\mathbf{S} - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ and $\bar{\mathbf{z}} - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$, are two simple measures. These values show the departures of

the observed and the predicted covariance matrices and mean vectors of the observed variables. Standardization of the residuals that take account of the scaling of the observed variables or the standard errors of the residuals are sometimes employed. Numerous other overall fit measures appear in the literature. Most are normed to range approximately from 0 to 1, where the value of 1 represents an ideal fit. The fit indices are the subject of debate in the literature on structural equation models (e.g. Bollen & Long [8]).

A second type of fit assessment occurs for the components of the model rather than the overall fit of the model. These components of fit are ones that are familiar to researchers using regression techniques. Researchers examine such things as the signs and the significance of coefficients, variances, and covariances, and the R-squares for equations. They also check for "improper" solutions such as negative variances or correlations greater than one.

Respecification

It is not unusual to find that an initial model specification provides an inadequate match to the data. A common reaction is to attempt to improve the model. Once the researcher enters this more **exploratory** mode of analysis, the usual significance tests cannot be interpreted in the usual way. It is then important to seek to replicate the final model on an independent data set. The substantive expert of the research is the most valuable source for possible modifications of the initial model. It is not unusual for the analysts to have considered several plausible relationships that were excluded from the initial model. These modifications are natural ones to consider, if the initial model fit is poor.

Empirical methods that can help in respecification also are available. The residual covariance matrix and mean vector described above show poorly fit parts of the data. But care must be taken in using such residuals [10]. Other aids are Lagrangian multiplier (and Wald) test statistics, that estimate the decrease (increase) in the chi-square test statistic for the freeing up (restricting) one or more parameters at a time [2]. Too great a reliance on these empirical methods can lead to problems (e.g. MacCallum [25]), but when used in conjunction with substantive expertise, they can prove helpful.

Historical Origins

We can trace the ancestry of contemporary structural equation models to several sources: Sewall Wright's (1918 [32], 1921 [33], and 1925 [34]) **path analysis**, the factor analysis tradition in psychometrics, simultaneous equation work in econometrics, and the 1960s and early 1970s synthesis of these areas in sociometrics. Although contemporary structural equation models are distinct in many ways, Wright's path analysis is probably the closest relative. Path analysis begins with a model specified prior to estimation. It provides a method of testing the consistency of a model to the data and a method to trace the influences of variables through a system of equations. The path diagram invented by Wright in 1921 [33] is a pictorial representation of the model. It provides a simple way to represent the complex relations between a large number of latent or observed variables. These diagrams are standard in structural equation models. Wright also used these diagrams to distinguish the direct, indirect, and total effects of one variable on another. The direct effects are the influences of one variable on another that do not pass through any other variable. The indirect effect is an impact that is through at least one other variable, while the total effect is the sum of the direct and indirect effects of one variable on another. This decomposition of effects is still part of structural equation models, although researchers have elaborated the definitions to include reciprocal relations and the presence of latent variables and have debated their "causal" meaning (*see* **Causal Direction, Determination**).

Another lasting influence of Wright's path analysis is the practice of writing the variances and covariances between variables as functions of the model parameters (e.g. coefficients, variances, and covariances of exogenous variables and disturbances). Wright used these relations to explore issues of model identification and the estimation of the parameters in a path model. Through examples, he demonstrated how path analysis could incorporate latent variables (factors), reciprocal relations, and recursive relations into statistical models. See the entry on **path analysis** for further details.

Wright's [32] first application of path analysis, appearing in 1918, was a factor analysis of bone size measurements. Unknown to Wright, Spearman [29] had proposed factor analysis over a decade earlier to analyze whether a general intelligence factor

underlied individuals' performances on tests. Spearman's work launched the beginning of the factor analysis tradition in psychometrics. Factor analysts soon moved from single to multiple-factor solutions and developed various methods of "rotating factors" to improve interpretability (*see* **Rotation of Axes**). Psychometricians became the most experienced group in the analysis of latent variables measured with multiple indicators. Some applied factor analysis to test prior hypotheses about the dimensionality of measures. However, most researchers used factor analysis as a data reduction tool, in which the number of factors and the pattern of influences of the latent variables on the observed variables were determined by the statistical procedures rather than by being specified in advance.

A related but separate development in psychometrics was classical test theory [24]. It shared with factor analysis a concern with latent or true score variables, but each observed variable was a function of a true score and error rather than being possibly influenced by multiple factors. In addition, factor analysis conceives of each variable as having a specific variance that is distinct from the factors and separate from the pure random error. The true score from classical test theory would include specific variance as part of the true score, not a part of the residual term. Classical test theory developed distinct definitions and approaches to the **reliability** and **validity** of measures. And these concepts of reliability and validity still hold influence in contemporary structural equation modeling, although such models lead to far more general relations between variables than those included in classical test theory by allowing correlated errors of measurement and multiple latent variables to influence observed variables.

The contribution of econometrics to structural equation modeling comes largely from its work on *simultaneous equations*. These models focused on observed rather than latent random variables. They dealt with issues of identification and estimation of a system of equations [15]. Econometricians proposed general rules of identification for simultaneous equations [13] that systematized the study of this issue and greatly influenced the contemporary perspective on identification in the more general structural equation models. Similarly, econometricians' work on limited information (e.g. Theil [30] and Basmann [1]) and full information estimators [18] led to the more sophisticated estimators that are commonly applied to structural equation models.

In the 1960s and early 1970s, sociometrics set the stage for the cross fertilization of path analysis, factor analysis, and econometric models that eventually merged into the contemporary form of structural equation models. Blalock [4], for instance, demonstrated the power of path analysis and partial correlations in examining a researcher's model of hypothesized relationships. Duncan's [11] didactic paper on path analysis in 1966 had a tremendous impact on the spread of path analysis in sociology as well as in psychology and other disciplines. Duncan et al. [12] illustrated the synthesis of latent variable and simultaneous equation models using path analysis in a classic 1968 study of peer influence. In 1969, Heise [17] shed new light on the use of panel data to explore reliability and stability in the measurement of variables. A classic 1971 edited volume by Blalock [5] illustrates the early merging of these techniques and the diffusion of statistical approaches from one field to another.

Although separated by only a couple of years, a 1973 edited volume by Goldberger & Duncan [14] revealed the more sophisticated approach to structural equation models that now dominates the field. Included in the volume is the highly influential paper by Karl Jöreskog, where he presented an early version of the LISREL model. The papers marked a more general approach to model specification, the implied covariance matrix, identification, estimation, and testing that is typical of current research.

Structural equation models have diffused through most of the social sciences and have begun to appear in the biostatistics and public health literature. Numerous software packages to estimate structural equation models are available with LISREL [22] and EQS [3] being the two most widely used ones. Publications using structural equation models are common in sociology, marketing, psychology, and education, and the technical literature continues to grow. *Structural Equation Modeling* is a journal devoted to the technique, but other statistical journals also publish work in this area. SEMNET is a listserver devoted to structural equation models and to date has over 1400 subscribers. Structural equation models remains an active area of research and applications.

References

[1] Basmann, R. (1957). A generalized classical method of linear estimation of coefficients in a structural equation, *Econometrica* **25**, 77–83.

[2] Bentler, P.M. (1989). *EQS Structural Equations Program Manual*. BMDP Statistical Software, Los Angeles.

[3] Bentler, P.M. (1992). *EQS Structural Equations Program Manual*. BMDP Statistical Software, Los Angeles.

[4] Blalock, H.M. (1964). *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill.

[5] Blalock, H.M., ed. (1971). *Causal Models in the Social Sciences*. Aldine-Atherton, Chicago.

[6] Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

[7] Bollen, K.A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations, *Psychometrika* **61**, 109–121.

[8] Bollen, K.A. & Long, J.S., eds (1993). *Testing Structural Equation Models*. Sage, Newbury Park.

[9] Browne, M.W. (1984). Asymptotic distribution free methods in analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

[10] Costner, H.L. & Schoenberg R. (1973). Diagnosing indicator ills in multiple indicator models, in *Structural Equation Models in the Social Sciences*, A.S. Goldberger & O.D. Duncan, eds. Seminar Press, New York, pp. 167–199.

[11] Duncan, O.D. (1966). Path analysis: sociological examples, *American Journal of Sociology* **72**, 1–16.

[12] Duncan, O.D., Haller, A.O. & Portes, A. (1968). Peer influences on aspirations: a reinterpretation, *American Journal of Sociology* **74**, 119–137.

[13] Fisher, F.M. (1966). *The Identification Problem in Economics*. McGraw-Hill, New York.

[14] Goldberger, A.S. & Duncan, O.D., eds (1973). *Structural Equation Models in the Social Sciences*. Academic Press, New York.

[15] Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations, *Econometrica* **11**, 1–12.

[16] Hägglund, G. (1982). Factor analysis by instrumental variables, *Psychometrika* **47**, 209–222.

[17] Heise, D.R. (1969). Separating reliability and stability in test-retest correlation, *American Sociological Review* **34**, 93–101.

[18] Hood, W.C. & Koopmans, T.C., eds (1953). *Studies in Econometric Method. Cowles Commission Monograph* No. 14. Wiley, New York.

[19] Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system, in *Structural Equation Models in the Social Sciences*, A.S. Goldberger & O.D. Duncan, eds. Academic Press, New York, pp. 85–112.

[20] Jöreskog, K.G. (1977). Structural equation models in the social sciences: specification estimation, and testing, in *Applications of Statistics*, P.R. Krishnaiah, ed. North-Holland, Amsterdam, pp. 265–287.

[21] Jöreskog, K.G. & Sörbom, D. (1981). LISREL V*: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. National Educational Resources, Chicago.

[22] Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8.* Scientific Software Inc., Mooresville.

[23] Keesling, J.W. (1972). Maximum Likelihood Approaches to Causal Analysis. *Unpublished doctoral dissertation*, University of Chicago.

[24] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, Massachusetts.

[25] MacCallum, R.C. (1986). Specification searches in covariance structure modeling, *Psychological Bulletin* **100**, 107–120.

[26] Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika* **49**, 115–132.

[27] Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments, *Quality & Quantity* **24**, 367–386.

[28] Satorra, A. & Bentler, P.M. (1994). Corrections to test statistic and standard errors in covariance structure analysis, in *Analysis of Latent Variables in Developmental Research*, A. Von Eye & C.C. Clogg, eds. Sage, Newbury Park, pp. 399–419.

[29] Spearman, C. (1904). General intelligence, objectively determined and measured, *American Journal of Psychology* **15**, 201–293.

[30] Theil, H. (1971). *Principles of Econometrics*. Wiley, New York.

[31] Wiley, D.E. (1973). The identification problem for structural equation models with unmeasured variables, in *Structural Equation Models in the Social Sciences*, A.S. Goldberger & O.D. Duncan, eds. Academic Press, New York, pp. 69–83.

[32] Wright, S. (1918). On the nature of size factors, *Genetics* **3**, 367–374.

[33] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.

[34] Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics* **5**, 161–215.

KENNETH A. BOLLEN

Structural Nested Failure Time Models

Structural nested failure time models (SNFTMs) are causal models for the effect of a time-dependent treatment or exposure on a survival time outcome in the presence of **time-dependent** confounding covariates [1, 10, 11, 15–17, 19, 20, 23, 31, 32] (*see* **Confounding**). The simplest SNFTMs map a subject's observed failure time T, observed treatment and confounder history, and an unknown parameter ψ_0 into the time U at which the subject would have failed if, possibly contrary to fact, treatment had been withheld. The causal parameter ψ_0 is identified if, as in a sequentially randomized experiment, the treatment at time t is randomly assigned (i.e. ignorable) conditional on past treatment and confounder history. The method of g-estimation provides computationally convenient and **robust** semiparametric estimators of ψ_0 when ψ_0 is identified [16, 26].

The usual approach to the estimation of the effect of a time-varying treatment on survival has been to model the **hazard** of failure at t as a function of past treatment history using a time-dependent **proportional hazards model**. In the next section, we show that the usual approach may be biased, whether or not one further adjusts for past confounder history in the analysis, when (i) there exists a time-dependent risk factor for, or predictor of, the event of interest that also predicts subsequent treatment, and (ii) past treatment history predicts subsequent risk factor level. The following two examples demonstrate conditions (i) and (ii) will be true in studies in which there is treatment by "indication" and/or a time-dependent covariate that is simultaneously a confounder and an intermediate variable on the causal pathway from treatment to failure.

The drug AZT, used in the treatment of **AIDS**, is a direct red blood cell toxin that is often withheld in anemic subjects, since the toxic effects of AZT can worsen the anemia. Furthermore, anemic patients are at increased risk of death. Thus in a study of the effect of AZT on survival of patients with AIDS, the time-dependent covariate anemia is both a risk factor for death and a predictor of subsequent treatment with AZT. Furthermore, as a red blood cell toxin, past AZT treatment is a risk factor for the development of anemia. In **occupational mortality** studies, unhealthy workers who terminate employment early are at an increased risk of death compared to other workers and receive no further exposure to the chemical agent under study. Therefore, the time-dependent covariate, employment status at time t, is an independent risk factor for death and a predictor of future exposure to the study agent. In addition, previous exposure to the study agent may lead to early termination of employment if the agent causes a disabling illness. Epidemiologists refer to covariates such as anemia or employment status in the above examples as *time-dependent confounders*.

This article is organized as follows. We first describe the fundamental assumption of no unmeasured confounders that, if true, allows us to test for and estimate causal effects from longitudinal data. In the next section we describe a valid α-level test, the g-test, of the **null hypothesis** of no causal effect of treatment on survival. We then describe the potential for bias and lack of robustness of alternative testing procedures. In the section "Deterministic Structural Nested Failure Time Models" we introduce the simplest SNFTMs – the deterministic SNFTMs. In the section "g-Estimation of ψ_0" we show that g-estimation of deterministic SNFTMs provides a unified approach to estimation of and testing for the effect of a time-dependent treatment. In the section "Sensitivity Analysis", we describe how the consequences of violations of our assumption of no unmeasured confounders can be explored through a sensitivity analysis. In these sections, we assume censoring is absent. In the following section we extend our methods to allow for censoring by end of follow-up, loss to follow-up, and competing risks. In the section "Inference Based on Instantaneous-Rate RPSNFTMs", we show that it is difficult to incorporate a priori biological knowledge as restrictions on the functional form of our deterministic SNFTMs. However, it is straightforward to incorporate biological knowledge if we adopt a more general class of causal models, the instantaneous-rate rank-preserving structural nested failure time models (RPSNFTMs), which model the effect of a final instantaneous blip of treatment on survival. The parameters of an instantaneous-rate RPSNFTM can be consistently estimated using g-estimation. Instantaneous-rate RPSNFTMs allow the magnitude of the treatment effect to depend on the measured factors but not on unmeasured factors. This restriction is often biologically implausible. We therefore

introduce, in the following section, the class of instantaneous-rate SNFTMs which allow the magnitude of the treatment effect to depend on both measured and unmeasured factors, and include the instantaneous-rate RPSNFTMs as a special case. It is of scientific and public health interest to estimate the survival curves that would be expected under various treatment regimes in order to determine the optimal regime with which to treat future patients. Hence, we consider estimation of regime-specific survival curves. Finally, we briefly describe an alternative nonnested structural failure time model that may sometimes have advantages in survival curve estimation.

Structural nested models for repeated measure and other nonfailure time outcomes (*see* **Longitudinal Data Analysis, Overview**) are considered in [15, 17, 18, 20], and [21]; they are not considered in this article.

Causal Inference from Observational Data

The Data

For pedagogic purposes we consider a study of the effect of AZT treatment on the survival of AIDS patients. Let T_i be a continuous variable recording the survival time for the ith study subject, $i = 1, \ldots, n$, with time measured from study enrollment. Let $A_i(t)$ record subject i's AZT dosage rate at t and $\mathbf{L}_i(t)$ record the value at t of a vector of various time-dependent and time-independent covariates such as CD4 lymphocyte count, presence of anemia, and gender. For any time-dependent random variable $Z_i(t)$, let $\overline{Z}_i(t^-) = \{Z_i(u); 0 \leq u < t\}$ be the history of the Z-process up to but not including time t and let $\overline{Z}_i(t)$ be the history of the process through t. Note that $Z_i(t)$ is defined only for $t \leq T_i$. For the time being, we assume there is no censoring. In the absence of censoring, the observable variables are then $\{T_i, \overline{A}_i(T_i), \overline{\mathbf{L}}_i(T_i)\}$, which we assume are independent and identically distributed, and henceforth suppress the i subscript denoting subject. Following [3, 13, 14], and [29], we shall also assume there exists a latent (possibly counterfactual) "baseline" failure time random variable U representing a subject's survival time had, possibly contrary to fact, AZT always been withheld.

The Fundamental Assumption of No Unmeasured Confounders

Our fundamental assumption of no unmeasured confounders is

$$U \perp\!\!\!\perp A(t) | \overline{\mathbf{L}}(t^-), \overline{A}(t^-), T \geq t, \qquad (1)$$

where $A \sqcup B|C$ means A is independent of B given C [4]. We will also refer to assumption (1) as the assumption that treatment $A(t)$ is sequentially ignorable or randomized given the past. Assumption (1) states that, conditional on AZT history and the history of all recorded covariates prior to t, increments in AZT dosage rate at t are independent of the baseline failure time random variable U. This assumption will be true if all risk factors for, i.e. predictors of, the baseline failure time U that are used by patients and physicians to determine the dosage of AZT at t are recorded in $\overline{\mathbf{L}}(t^-)$ and $\overline{A}(t^-)$. For example, since physicians tend to withhold AZT from anemic subjects, and in untreated subjects anemia is a predictor of survival, assumption (1) would be false if $\overline{\mathbf{L}}(t^-)$ does not contain anemia history. It is the primary goal of the epidemiologists conducting an observational study to collect data on a sufficient number of covariates to ensure that our assumption (1) will be at least approximately true.

Assumption (1) is the fundamental condition that will allow us to draw causal inferences from observational data (*see* **Causation**). It is precisely because (1) cannot be guaranteed to hold in an observational study and is not empirically testable that it is so very hazardous to draw causal inferences from observational data. Note that if, as in a sequentially randomized trial, at each time t, the dose of AZT was chosen at random by the flip of a coin, then (1) would be true even if the probability that the coin landed heads depended on past covariate and AZT history. It is because physical randomization guarantees (1) that most people accept that valid causal inferences can be obtained from a randomized trial (*see* **Randomization**). See [5, 13], and [29] for further discussion. In a later section we describe how the consequences of violations of (1) can be explored through sensitivity analysis.

For convenience, until a later section we shall assume that the treatment $A(t)$ received at time t is dichotomous, i.e. $A(t) = 1$ if on treatment at t and zero otherwise. Robins [16] and Robins

et al. [26] consider nondichotomous treatments. For $A(t)$ dichotomous, assumption (1) can be written

$$\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), U) = \lambda(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-)), \quad (2)$$

where, if $A(t)$ is an instantaneous-rate process, $\lambda_A(t|\overline{A}(t^-), \cdot) = \lim_{\delta t \to 0} \Pr[A(t+\delta t) \neq A(t^-)| \overline{A}(t^-), T \geq t, \cdot]/\delta t$ is the hazard of the treatment process jumping in the infinitesimal interval $[t, t+\delta t)$ given $\overline{A}(t^-)$ and $\cdot$. However, $\lambda_A(t|\overline{A}(t^-), \cdot) = \Pr\{A(t) \neq A(t^-)|\overline{A}(t^-), T \geq t, \cdot\}$ is a discrete hazard if $A(t)$ can only jump at nonrandom discrete times $t_1, t_2, \ldots$, as would be the case if the $A(t_k)$ recorded whether a subject was on AZT at weekly clinic visits. Owing to measure theoretic subtleties, (2) but not (1) is a mathematically precise statement of our assumption of no unmeasured confounders.

A g-Test of the Causal Null Hypothesis

The sharp causal null hypothesis of no treatment effect on survival is that each subject's observed and baseline lifetimes are the same. That is,

$$U = T \text{ w.p.1}, \quad (3a)$$

where w.p.1 stands for with probability 1. Given our assumption (2), the restriction on the distribution of the observables implied by (3a) is that the hazard of treatment jumps at t does not depend on the survival time T given past treatment and covariate history. That is,

$$\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), T) = \lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-)). \quad (3b)$$

Hence, if the A process is a instantaneous-rate process, then we can test (ii) by specifying a time-dependent **Cox** (proportional hazards) **regression model**

$$\lambda_0(t) \exp[\boldsymbol{\alpha}' \mathbf{W}(t)] \quad (4)$$

for $\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-))$, where $\mathbf{W}(t)$ is a known vector-valued function of $(\overline{A}(t^-), \overline{\mathbf{L}}(t^-))$, $\boldsymbol{\alpha}$ is an unknown parameter vector, and $\lambda_0(t)$ is an unspecified baseline hazard function. If the A process jumps only at fixed discrete times, we interpret (4) as a model for the **odds** $\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-))/\{1 - \lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-))\}$. If model (4) is correctly specified, an asymptotic α-level Cox **partial likelihood** score, Wald (*see* **Likelihood**), or **likelihood ratio test** of the hypothesis $\theta = 0$ in the extended model that adds a term θT to $\boldsymbol{\alpha}' \mathbf{W}(t)$ in (4) is an asymptotically α-level test (*see* **Level of a Test**) of the sharp null hypothesis (3a) under the assumption (2) of no unmeasured confounders. Robins [16] refers to such a test as a g-test. Note a g-test first models the hazard of the treatment process as a function of the survival time T and past treatment and covariate history, and then tests whether the coefficient θ of T is significant. In fact, we obtain an α-level test of (3b) by testing $\theta = 0$ in the extended model that adds the term $\theta Q(t)$ to $\boldsymbol{\alpha}' \mathbf{W}(t)$ in (4), where $Q(t) = q(t, \overline{A}(t^-), \overline{\mathbf{L}}(t^-), T)$ is a function chosen by the data analyst. The choice of $Q(t)$ effects the power but not the level of the g-test. The g-test is a generalization to time-dependent treatments and confounders of Rosenbaum's [27, 28] test for the effect of a single time-independent treatment.

Bias of Standard Methods

To understand why standard approaches that use Cox regression to model the hazard of failure as a function of past treatment history are biased whether or not one adjusts for past confounder history, we consider a group of AIDS patients who are alive at 10 months, dichotomized into those who developed anemia by 8 months and those who remain free of anemia at 8 months. A Cox regression analysis that estimates the rate ratio at 10 months attributable to AZT exposure in the interval 8–10 months without adjusting for or stratifying on anemia status can make AZT appear falsely beneficial, since anemic subjects are at a higher risk of dying at 10 months and are less likely to receive AZT therapy in months 8–10. That is, anemia status at 8 months is a confounder for the causal effect of AZT treatment received in the interval from 8 to 10 months.

However, because AZT causes anemia, even if both AZT and anemia have no causal effect on the survival of any subject, the above Cox regression analysis may continue to suggest falsely that AZT is beneficial even when we adjust for anemia at 8 months. To see why, for simplicity, suppose now that 300 subjects receive AZT by 4 months, and 300 subjects never receive AZT. In both groups of 300, suppose that, regardless of AZT treatment or treatment for anemia, 100 individuals are poor-prognosis subjects who are destined to die at 10 months, 100 are moderate-prognosis subjects destined to die at 20 months, and 100 are good-prognosis subjects destined

Table 1 A Hypothetical study

Anemia by 8 months				No anemia			
	Time to death (months)				Time to death (months)		
	10	20	30		10	20	30
No AZT	100[a]	0	0	No AZT	0	100[b]	100[c]
AZT by 4 months	100[a]	100[b]	0	AZT by 4 months	0	0	100[c]

[a]Poor-prognosis patients.
[b]Moderate-prognosis subjects.
[c]Good-prognosis subjects.

to die at 30 months. Suppose AZT causes anemia in moderate-prognosis patients. Specifically, all moderate-prognosis patients would develop anemia at 8 months if given AZT, whereas none would develop anemia without AZT. All poor-prognosis and no good-prognosis patients develop anemia regardless of AZT therapy. Under these assumptions, the data would be as shown in Table 1. Inspecting Table 1, we observe that, within the stratum defined by the presence of anemia at 8 months, the mortality rate at 10 months is less in those who received AZT than in those who did not. Similarly, in the stratum defined by the absence of anemia, the mortality rate at 20 months is less in those who received AZT than in those who did not. Thus, a Cox analysis that adjusts for (or stratifies on) past anemia history would falsely suggest that AZT has a beneficial effect on survival. This bias is attributable to the fact that in Table 1 AZT by 4 months is a risk factor for subsequent anemia, and that anemia is a noncausal risk factor for death, since the death rate at 10 months is greater in those with anemia than in those without anemia among subjects without AZT, and the death rate at both 10 and 20 months is greater in those with anemia than in those without anemia among subjects with AZT. Note, by construction of our example, anemia is not a causal risk factor for death; rather, it is a proxy for the unmeasured prognosis variable. Furthermore, anemia is not an intermediate variable on the causal pathway from AZT treatment to death since, by construction, there is no such causal pathway (*see* **Path Analysis**).

It follows that we must control for the confounder "anemia status at month 8" to estimate the causal effect of AZT in the interval (8,10). However, we must not control for the variable "anemia status at month 8" to estimate the causal effect of AZT therapy in the interval (0,8) on survival. If, however, we summarize AZT history over the interval (0,10) in terms of cumulative dosage, average dose intensity, or the time since the initiation of AZT therapy, these requirements cannot be met, since we lose the ability to separate out AZT in the interval (0,8) from AZT in the interval (8,10). However, the g-test of the previous subsection is specifically designed to control for confounding by variables affected by earlier treatment by never lumping treatment received at different times. Specifically, the g-test checks, at each time t, for association between treatment $A(t)$ received at t and the failure time T after adjusting for confounder and treatment history before t, but without adjusting for the "post-treatment" variables "covariate and treatment history subsequent to t". It is essential to the validity of the g-test that treatment history $\overline{A}(t^-)$ before t be adjusted for as a potential confounding factor for the effect of the treatment $A(t)$ received at t.

Formally, results in this section reflect the fact that the null hypothesis (3b) does not imply either that

$$\lambda_T(t|\overline{A}(t^-)) = \lambda_T(t) \tag{5}$$

or that

$$\lambda_T(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-)) = \lambda_T(t|\overline{\mathbf{L}}(t^-)), \tag{6}$$

where $\lambda_T(t|\cdot)$ is the hazard of failure at t given $\cdot$. However, Robins [12] proves that (3b) does imply (5) if either of the following are true:

$$\lambda_T(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-)) = \lambda_T(t|\overline{A}(t^-)) \tag{7}$$

or

$$\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-)) = \lambda_A[t|\overline{A}(t^-)]. \tag{8}$$

If (7) holds, we say $\mathbf{L}(t)$ is not an independent predictor of failure. If (8) holds, we say $\mathbf{L}(t)$ is not an independent predictor of subsequent treatment. If

either (7) or (8) holds, we say that the $\mathbf{L}(t)$ process is not a confounder for the effect of $A(t)$ on survival. In that case, we can test the sharp null hypothesis (3a) under assumption (2) by ignoring data on $\mathbf{L}(t)$ and testing (5) using a time-dependent Cox model for failure.

Computational Complexity and Nonrobustness of Tests Based on the g-Computation Algorithm

In this subsection, to avoid the need for product integral notation, we assume the covariate history $\overline{\mathbf{L}}(T)$ can jump only at fixed times $k = 0, 1, 2, \ldots$. This is no practical limitation, since, for example, we could take the time interval between the jump times k and $k+1$ to be 1 second. The g-computation algorithm formula $r(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ of Robins [13, 14] for the effect of a treatment regime $\overline{a} = \{a(u); 0 \le u < \infty\}$ on survival to time t conditional on $\overline{\mathbf{L}}(m^-) = \overline{\mathbf{l}}(m^-)$ and $\overline{A}(m^-) = \overline{a}(m^-)$ is

$$r(t, \overline{a}, \overline{\mathbf{l}}(m^-)) = \int \cdots \int \exp\left\{-\int_0^t \lambda_T(u|\overline{\mathbf{l}}(u^-), \overline{a}(u^-))\right\} \times \prod_{k=m}^{\mathrm{int(t)}} \mathrm{d}F[\mathbf{l}(k)|\overline{\mathbf{l}}(k^-), \overline{a}(k^-)], \qquad (9)$$

where $\mathrm{int}(t)$ is the greatest integer less than t, and $\lambda_T(u|\cdot)$ is the conditional hazard of failure at u given $\cdot$. Robins [13, 14] showed that under a sequential randomization assumption, $r(t, \overline{a}, \overline{\mathbf{l}}(0^-))$ is the probability of survival to t had, contrary to fact, all subjects followed treatment regime $\overline{a}$ until failure. Researchers studying causal models based on directed acyclic graphs [12, 30] have recently rediscovered that g-computation algorithm formula. In addition, Arjas & Eerola [2] and Klein et al. [7] have also considered estimation of causal effects using this formula.

The g-null theorem of Robins [13] states that (3b) is true if and only if, for all $(t, \overline{a}, \overline{\mathbf{l}}(m^-))$, $r(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ depends on $\overline{a}$ only through $\overline{a}(m^-)$. It follows that one can, in principle, test the null hypothesis (3a) under the assumption (2) of no unmeasured confounders by: fitting a Cox proportional hazards model for $\lambda_T(u|\overline{a}(u), \overline{\mathbf{l}}(u^-))$ depending on parameters θ and a parametric or semiparametric model for $f(\mathbf{l}(k)|\overline{\mathbf{l}}(k^-), \overline{a}(k^-))$ depending on parameters η, using the fitted model to construct an estimator $\hat{r}(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ of $r(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ for various choices of $t, \overline{a}, \overline{\mathbf{l}}(m^-)$ by evaluating the right-hand side of (9); deriving estimates of the standard errors of the $\hat{r}(t, \overline{a}, \overline{\mathbf{l}}(m^-))$; and finally constructing a test of the hypothesis that $r(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ only depends on $\overline{a}$ through $\overline{a}(m^-)$ using the estimates $\hat{r}(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ and their estimated standard errors.

The difficulty with this procedure is twofold. First, it is computationally extremely demanding since (i) the integral on the right-hand side of (9) cannot in general be evaluated analytically, and a **Monte Carlo** approximation must be used; and (ii) it is difficult to compute **delta-method** estimators for the standard error of $\hat{r}(t, \overline{a}, \overline{\mathbf{l}}(m^-))$, and **bootstrap** standard errors may be computationally too demanding because of (i). Secondly, there will in general be no simple function ψ of the parameters (θ, η) that takes a fixed value (say zero) if and only if $r(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ only depends on $\overline{a}(m^-)$. As discussed in [19], this fact implies that the tests based on the $\hat{r}(t, \overline{a}, \overline{\mathbf{l}}(m^-))$ will be exquisitely sensitive to inevitable model misspecification. In summary, we suggest the g-test described earlier be used to test the null hypothesis (3b).

Deterministic Structural Nested Failure Time Models

The g-test of the hypothesis $\theta = 0$ in the extension of model (4) is an asymptotic α-level test of the sharp null hypothesis (3a) if model (4) is correctly specified and if the assumption (2) of no unmeasured confounders is true. However, we also wish to estimate the size of the treatment effect when the causal null is false. To do so, we introduce g-estimation of structural nested failure time models which will provide a unified approach to estimation of and testing for the effect of a time-dependent treatment.

The simplest SNFTM is a deterministic transformation model which assumes the counterfactual failure time U is a known function $h(T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi_0)$ of the observed data $(T, \overline{A}(T), \overline{\mathbf{L}}(T))$ and an unknown parameter ψ_0; that is

$$U = H(\psi_0), \qquad (10a)$$

where

$$H(\psi) \equiv h(T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi). \qquad (10b)$$

A specific example of a deterministic SNFTM is the strong version of the **accelerated failure-time model** of Cox & Oakes [3] which assumes

$$h(T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi) = \int_0^T \exp\{\psi A(t)\}\,\mathrm{d}t. \quad (11)$$

Any deterministic SNFTM (11) satisfies the following:

$$\text{if } \overline{A}(T) \equiv 0, \text{ then } U = T; \quad (12a)$$

$$T \equiv U \text{ w.p.1 if and only if } \psi_0 = 0. \quad (12b)$$

Statement (12a) is a natural consistency assumption stating that if a subject is, in fact, untreated, then the observed failure time T equals the failure time U when treatment is withheld. Statement (12b) implies the null hypothesis $\psi_0 = 0$ corresponds to the causal null hypothesis (3a) that treatment has no effect. To understand the implications of (11) when $\psi_0 \neq 0$, consider a subject who is continuously treated. Then, by (10) and (11), $U = \mathrm{e}^{\psi_0} T$ so $T = \mathrm{e}^{-\psi_0} U$. That is, a subject's untreated survival time U is expanded or contracted by the factor $\mathrm{e}^{-\psi_0}$ by constant treatment. Hence, if $\psi_0 > 0$, treatment is harmful and lessens survival; if $\psi_0 < 0$, treatment is beneficial and increases survival. Robins et al. [26] referred to deterministic SNFTMs as rank-preserving structural failure time models.

g-Estimation of ψ_0

We now describe how to obtain consistent asymptotically normal point and **interval estimates** of the parameter ψ_0 consistent with the g-tests of introduced above in the sense that 95% **confidence intervals** for ψ_0 will fail to include zero if and only if the corresponding 0.05 level g-test rejects. As a simple example, consider the deterministic SNFTM (11) with ψ one-dimensional. We estimate ψ by a "grid search". First, we note that for each value of ψ, $H(\psi)$ can be computed by (11) from the observed data. Hence, under the reasonable biological assumption that $|\psi_0| < 3$, separately, for each of the 61 values of ψ in the set $\{-3, -2.9, \ldots, 0, \ldots, 2.9, 3\}$, we perform a Cox partial likelihood score test (g-test) of the hypothesis $\theta = 0$ in the extended model that adds a term $\theta Q(t, \psi)$ to $\boldsymbol{\alpha}'\mathbf{W}(t)$ in (4) with $Q(t, \psi) = q(t, \overline{A}(t^-), \overline{\mathbf{L}}(t^-), H(\psi))$ a function chosen by the data analyst. A valid 95% large-sample confidence interval for ψ_0 is the set of ψ for which the score test fails to reject at the 0.05 level provided our no-confounding assumption (2), our Cox model (4), and our deterministic SNFTM (10)–(11) are correct. Furthermore, the g-estimate $\hat{\psi}$ is a consistent asymptotically normal estimator of ψ_0, where $\hat{\psi}$ is defined to be the value of ψ for which the partial likelihood score test of $\theta = 0$ is precisely zero. The parameter ψ is treated as a fixed constant when calculating the score test. The choice of the function $q()$ affects the length but not the coverage rate of the interval. The optimal choice of the function $q^*()$ is given in [16]. The method of g-estimation can be extended to estimate the parameter, say, $\psi = (\psi_1, \psi_2)'$ of a multiparameter deterministic SNFTM such as

$$H(\boldsymbol{\psi}) = \int_0^T \exp\{\psi_1 A(t) + \psi_2 L^*(t) A(t)\}\,\mathrm{d}t. \quad (13)$$

In model (13), $L^*(t)$ represents a known function of the covariate history $\overline{\mathbf{L}}(t^-)$. If the true value ψ_{20} of ψ_2 is nonzero, then there is a treatment–covariate interaction in the sense that the magnitude of the effect of the time-dependent treatment $A(t)$ depends on a subject's time-dependent covariate history $\overline{\mathbf{L}}(t^-)$ through the function $L^*(t)$. A g-estimate of the parameter vector $\boldsymbol{\psi}$ of (13) is obtained by choosing $Q(t, \boldsymbol{\psi})$ to be a known vector-valued function of dim $\boldsymbol{\psi}$ chosen by the data analyst and $\boldsymbol{\theta}$ to be a (dim $\boldsymbol{\psi}$)-valued parameter with dim $\boldsymbol{\psi}$ the dimension of the vector $\boldsymbol{\psi}$.

Estimation with Instrumental Variables

Suppose $A(t) = (A_1(t), A_2(t))$, with $A_1(t)$ recording a physician's prescribed treatment and $A_2(t)$ recording the actual treatment at time t. One then might suppose that

$$A_1(t) \coprod U | \overline{\mathbf{L}}(t^-), \overline{A}(t^-), T > t \quad (14)$$

is true but (1) is false if one believed that a predictor of both U and actual treatment $A_2(t)$ had not been included in $\overline{\mathbf{L}}(t^-)$. Under assumption (14), g-estimation of the parameter ψ_0 of the deterministic SNFTM (10) can proceed as before, except we view model (4) as a model for the cause-specific hazard $\lambda_{A_1}(t | \overline{A}(t^-), \overline{\mathbf{L}}(t^-))$ for jumps in the $A_1(t)$ process, thus ignoring jumps in the actual treatment $A_2(t)$ process in our estimation procedure. In this setting,

$A_1(t)$ is often referred to as an **instrumental variable** process, especially when prescribed treatment $A_1(t)$ has no direct causal effect on survival except through the actual treatment process $A_2(t)$. A familiar example of an instrumental variable is when $A_1(0)$ is the randomization indicator for assignment to treatment arm in a randomized **clinical trial** in which there is possibly nonrandom noncompliance and $A_2(t)$ is the actual treatment dose. (In such a setting, $A_1(t)$ can be defined to be zero by convention for times $t \neq 0$.) Then the g-estimation method described above is the method for adjusting for nonrandom noncompliance in randomized clinical trials described in [17]. For alternative **rank** estimation procedures, see [11] and [25].

Sensitivity Analysis

In observational studies, our fundamental assumption (2) of no unmeasured confounders cannot be empirically tested from the data. Hence, it is important to conduct sensitivity analyses to determine how point and interval estimates for ψ_0 would change under increasingly severe violations of (2). Let η be a sensitivity parameter that we will vary (but not estimate) in our sensitivity analysis and consider the model

$$\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), U) = \lambda_0(t) \exp[\boldsymbol{\alpha}'\mathbf{W}(t) + \eta U]. \tag{15}$$

When $\eta = 0$, both the assumption (2) of no unmeasured confounders and our Cox model (4) are true. As $|\eta|$ increasingly deviates from zero, (2) is increasingly violated. Our goal, in a sensitivity analysis, is to obtain valid point and interval estimates for the causal parameter ψ_0 of our deterministic SNFTM under the assumption that (15) is correctly specified, with α an unknown parameter to be estimated but with η known. Specifically, a 95% confidence interval for ψ_0 under these assumptions is obtained as the set of ψ for which the score test of the hypothesis $\theta = 0$ fails to reject at the 0.05 level in model

$$\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), H(\psi)) = \lambda_0(t) \exp[\boldsymbol{\alpha}'\mathbf{W}(t) + \eta H(\psi) + \theta Q(t, \psi)], \tag{16}$$

when η and ψ are treated as fixed and known when maximizing the partial likelihood over α. The data analyst should then display point and interval estimates for ψ for a moderately large number of choices for the sensitivity parameter η.

Censoring

In this section we extend our results to allow for right-censoring. We handle censoring by administrative end of follow-up differently from censoring by **competing risks** or by loss to follow-up (*see* **Bias from Loss to Follow-up**).

Censoring by End of Follow-Up

We assume that there is a fixed known calendar date at which the follow-up of all subjects will end. We then define the potential censoring time C for a subject to be the difference between this end-of-follow-up date and the date at which the subject entered follow-up. Hence, the potential censoring time C is known for all subjects, even those who fail before the end-of-follow-up date. Because the potential censoring time C is known at start of follow-up ($t = 0$), we can and do regard C as a time-independent "pretreatment" covariate that is contained in $\overline{\mathbf{L}}(t^-)$ for each time $t \geq 0$. If, as we assume in this section, the only cause of censoring is by end of follow-up, the data available for data analysis for each subject are $\{X = \min(T, C), \overline{A}(X), \overline{\mathbf{L}}(X)\}$.

Since $H(\psi)$ can only be computed for uncensored individuals, it might seem natural when calculating g-estimates of ψ_0 to replace the now partially unobservable $H(\psi)$ by the new random variable $X^*(\psi)$ obtained by replacing T by X in (10b). Unfortunately, this approach fails since, if $\psi_0 \neq 0$, then $X^*(\psi_0)$ is not independent of $A(t)$ given $(\overline{A}(t^-), \overline{\mathbf{L}}(t^-), X \geq t)$ even under the assumption (2) of no unmeasured confounders. Thus, an alternative approach is necessary. The key to our approach is to define new variables $(X(t, \psi), \Delta(t, \psi))$ that (i) in contrast to both T and $H(\psi)$, but like $X^*(\psi)$, are observed for all subjects, including those censored, and (ii) like $H(\psi_0)$, but unlike $X^*(\psi_0)$, satisfy, for $t < C$,

$$\lambda_A[t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), X(t, \psi_0), \Delta(t, \psi_0)] = \lambda_A[t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-)] \tag{17}$$

under assumption (2) and model (10). We can then estimate ψ_0 by g-estimation as before, except with

$Q(t, \psi)$ now a function $q(t, \overline{\mathbf{L}}(t^-), \overline{A}(t^-), X(t, \psi), \Delta(t, \psi))$. Below, we define $X(t, \psi)$ and $\Delta(t, \psi)$ only for the deterministic SNFTM (11). Robins [17, Appendix 4] gives the appropriate definitions for arbitrary SNFTMs. Let

$$X(t, \psi) = \min\{H(\psi), C(t, \psi)\},$$
$$\Delta(t, \psi) = I\{X(t, \psi) < C(t, \psi)\},$$

where $C(t, \psi) \equiv C - t + \int_0^t \exp\{\psi A(t)\}\, dt$ if $\psi \geq 0$ and $C(t, \psi) = \int_0^t \exp\{\psi A(t)\}\, dt + (C - t)e^{\psi}$ if $\psi < 0$. Eq. (17) is satisfied since $X(t, \psi_0)$ and $\Delta(t, \psi_0)$ are only functions of $\overline{A}(t^-), H(\psi_0)$ and C. Also $X(t, \psi)$ and $\Delta(t, \psi)$ are observables, since one can calculate that $X(t, \psi)$ is the minimum of the two observables $X^*(\psi)$ and $C(t, \psi)$. When $\Delta(t, \psi) = 0$, we say an individual is ψ-censored. Note that when $\psi \neq 0$, some failures will be ψ-censored. In practice, if **efficiency** is not of overriding concern, then it is convenient to use a very simple function $Q(t, \psi)$ that produces reasonably efficient estimators of ψ_0. The indicator function $\Delta(t, \psi)$ has been found often to satisfy this criterion in a number of examples [32]. The efficient choice of $Q(t, \psi)$ is given in [17, Appendix 4]. The simple sensitivity analysis methodology of the previous section can be extended to the censored data setting by replacing ηU by $\eta X(\psi)$ in (15); however, better methodology should be developed.

Censoring by Competing Risks

In this section, we assume that in addition to censoring by end of follow-up C, there is additional censoring by loss to follow-up and/or competing risks. Let Q be the minimum of time to loss to follow-up or to a competing risk event. For ease of exposition, we no longer distinguish censoring by loss to follow-up from censoring by competing risks and simply refer to Q as time to censoring by competing risks. The data available are $X^* = \min(T, C, Q) = \min(X, Q), \tau = I(X^* \neq Q), \overline{A}(X^*), \overline{\mathbf{L}}(X^*)$ so that $\tau = 1$ if and only if a subject was either observed to fail or to reach end of follow-up without suffering a competing risk. To adjust for censoring by competing risks, we assume that we have recorded data on a sufficient number of potential confounding factors in $\overline{\mathbf{L}}(t^-)$ so that there are no unmeasured confounders for censoring due to competing risk. That is,

$$\lambda_Q[t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), X > t, X] = \lambda_Q[t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), X > t], \quad (18)$$

in which case we shall also say that censoring by Q is ignorable given the past. Here, $\lambda_Q(t|\cdot)$ is the hazard for the random variable Q given $\cdot$.

Given the ignorable censoring assumption (18), our next task is to estimate the probability $K(X)$ of a subject surviving to $X = \min(T, C)$ without suffering a competing risk, which will be used as an inverse weight in the weighted g-estimation procedure described below. To do so, we fit the Cox proportional hazard model

$$\lambda_{0Q}(t) \exp\{\boldsymbol{\alpha}^{*'} \mathbf{W}^*(t)\} \quad (19)$$

for the hazard $\lambda_Q(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-), X > t)$, where $\mathbf{W}^*(t)$ is a known vector-valued function of $\overline{A}(t^-)$ and $\overline{\mathbf{L}}(t^-), \boldsymbol{\alpha}^*$ is the vector of unknown parameters, and $\lambda_{0Q}(t)$ is an unspecified baseline hazard. We then estimate $K(X)$ by multiplying together the estimated conditional probabilities of not suffering a competing risk before X using the time-dependent Cox model version of the **Kaplan–Meier estimator** [6]. Specifically, at each time Q_j where any subject j suffered a competing risk, we compute the Cox baseline hazard estimator

$$\hat{\lambda}_Q(Q_j) = 1 \Big/ \sum_{i=1}^{n} \{\exp[\hat{\boldsymbol{\alpha}}^{*'} \mathbf{W}_i^*(Q_j)] I(X_i^* \geq Q_j)\}$$

of $\lambda_{0Q}(Q_j)$. We then estimate a subject's $K(X)$ by the Cox model version of the Kaplan–Meier estimator

$$\hat{K}(X) = \prod_{\{j; Q_j \leq X, \tau_j = 0\}} \{1 - \hat{\lambda}_Q(Q_j) \exp[\hat{\boldsymbol{\alpha}}^{*'} \mathbf{W}^*(Q_j)], \}$$

which is the product, over the competing risk times $Q_j < X$, of the subject's estimated conditional probabilities of not suffering a competing risk. Note that a subject's estimated probability $\hat{K}(X)$ depends on his/her treatment and covariate history through the covariate $\mathbf{W}^*(t)$.

Having estimated $\hat{K}(X)$ for each subject with $X = \min(T, C)$ observed, we then estimate ψ_0 by replacing, in our g-estimation procedure of the previous subsection, the function $Q(t, \psi)$ by $Q^*(t, \psi) \equiv$

$Q(t, \psi)/\hat{K}(X)$ for each person who did not suffer a competing risk ($\tau = 1$) and by $Q^*(t, \psi) = 0$ for each person who did ($\tau = 0$). For example, if we use the simple function $Q(t, \psi) = \Delta(t, \psi)$, then $Q^*(t, \psi) = \tau\Delta(t, \psi)/\hat{K}(X)$.

We now give an intuitive explanation of why the g-estimate $\hat{\psi}$ obtained by this method should be **consistent** for ψ_0. Given the correctness of our Cox model (19) and of our assumption of ignorable censoring by competing risks (18), the following will be true: for each person with X observed ($\tau = 1$) and an estimated cumulative probability of, say, $\hat{K}(X) = 0.25$ of avoiding censoring by competing risks, there would, on average, have been three other persons (i.e. **ghosts**) who were censored by competing risks before $X(\tau = 0)$, and who would have had a similar value of X and a similar covariate and treatment history up to X, had censoring by competing risks been prevented. We therefore assign this person with $\tau = 1$ and $\hat{K}(X) = 0.25$ a weight of 4 in the g-estimation procedure by multiplying her covariate $\Delta(t, \psi)$ by the factor 4; she needs to count not only for herself but also for the three other similar subjects for whom X could not be observed due to censoring by competing risks and thus had $Q^*(t, \psi)$ set to zero.

This argument can be formalized to prove that the "competing risk" g-estimator $\hat{\psi}$ is a CAN estimator of ψ_0. However, the previous method of obtaining confidence intervals and P values is no longer valid because the contributions to the Cox partial likelihood score of the extended Cox model (4) for the treatment process are no longer uncorrelated for two distinct reasons. First, $K(X)$ and, therefore, the time-dependent covariate $\tau Q(t, \psi)/K(X)$ at time t depend on a subject's treatment and covariate history beyond t, disrupting the "martingale" structure of the Cox partial likelihood score. Secondly, the probability $K(X)$ of avoiding a competing risk is replaced by the estimate $\hat{K}(X)$ which depends on all the data. However, if we fit the extended model (4) using a Cox proportional hazards program that computes the so-called "robust variance" [8], the resulting g-intervals and tests are guaranteed to be conservative, i.e. in large samples, nominal 95% confidence intervals are guaranteed to cover ψ_0 at least 95% of the time and 0.05 level g-tests are guaranteed to reject the null hypothesis $\psi_0 = 0$ when true no more than 5% of the time. If the conservative "robust variance" g-intervals are too long to distinguish important substantive alternatives, narrower intervals that cover ψ_0 95% of the time in large samples can be obtained using the formulas provided in Appendix 1.

Remarks. Often it is reasonable to assume that censoring by end of follow-up C is also ignorable. To incorporate this assumption, we redefine Q to be the minimum of time to loss to follow-up, competing risks, and end to follow-up and replace C by a constant c^* which is slightly less than the maximal follow-up time $\max\{C_i; i = 1, \ldots, n\}$ so that $K(c^*)$ is bounded away from zero w.p.1. Then g-estimation and testing can proceed as above.

Estimation of Direct Effects

Suppose now we wish to estimate the direct effect of AZT $A(t)$ on survival when another treatment, say aerosolized pentamidine (AP), is not taken. If a reasonably large fraction of the study population, say at least 30%, were untreated with AP until failure or censoring, a quite robust approach is to regard a subject as censored at the first time the subject is on AP therapy; redefine Q to be the minimum of time to censoring by competing risks, time to loss to follow-up, and time to being on AP therapy; and estimate ψ_0 using the methods of the previous subsection.

If only a small fraction of the study population avoided AP therapy, then one can redefine $A(t)$ to be the joint treatment AP and AZT taken at time t and specify a deterministic SNFTM (10) that has separate parameters for the AZT effect and for the AP effect as described in [16, Section A2.12]. An alternative, and preferred approach is to specify a direct effect structural nested model as described in [21]. Discussion of these latter models is beyond the scope of this article.

Inference Based on Instantaneous-Rate RPSNFTMs

Difficulties Incorporating a Priori Biological Knowledge with Deterministic SNFTMs

In our simple deterministic SNFTM (11), the scientific meaning of the parameter ψ_0 was relatively straightforward; $\exp(-\psi_0)$ was the factor by which continuous treatment extended life. More specifically, since $\partial T/\partial H(\psi_0) = \exp\{-\psi_0 A(T)\}$, Cox &

Oakes [3] suggest interpreting $\exp\{-\psi_0 A(t)\}$ as the relative rate at which real time is being used up compared to baseline time at real time t. Thus, if an individual has U years of baseline time to be used up if treatment is withheld, then the actual time T at which the U years of baseline time will have been used is determined by (11). One might hope that the physical interpretation of $\partial T/\partial H(\psi_0)$ as the relative rate at which real time is being used up compared to baseline time would serve in more complex settings to allow us to easily incorporate prior biological knowledge as specific functional form restrictions on our deterministic SNFTM models. However, the following example suggests that this is not the case.

Suppose, based on a priori biological understanding, it is known that any treatment received at time t would have no effect on survival unless the subject is destined to fail within the next 5 weeks without additional treatment. An example would be a setting in which (i) failure is death from an infectious disease, (ii) if death occurs, it always occurs within 5 weeks from the time of initial unrecorded subclinical infection, and (iii) $A(t)$ is a preventive antibiotic treatment at t which is of no benefit unless the study subject is already infected by t. The challenge then is how to incorporate such biological knowledge into the functional form of a deterministic SNFTM (10). We will see that it is difficult to succeed at this challenge if we try to incorporate the biological knowledge directly into a deterministic SNFTM. However, incorporating such biological knowledge is straightforward in a more general class of causal models, the instantaneous-rate (locally) RPSNFTMs which model the effect of a final instantaneous blip of treatment on survival. Since each instantaneous-rate RPSNFTM mathematically entails a unique deterministic SNFTM (10) as a solution to a particular differential equation, it follows that by solving this differential equation, we can determine the restrictions on the functional form of our deterministic SNFTM (10) implied by the restriction that only treatment received within 5 weeks of failure can affect failure.

Instantaneous-Rate RPSNFTMs

To describe instantaneous-rate RPSNFTMs, we shall require a number of additional definitions.

Definition. A treatment regime or plan $\overline{a} \equiv a(\cdot) \equiv \{a(t); 0 \le t < \infty\}$ is a continuous from the right with left-hand limits (cadlag) function on $[0, \infty)$ that is everywhere continuously differentiable except possibly on a countable set of discontinuity points, only finitely many of which are contained in any bounded interval.

Remark. If (i) $\overline{A}(T)$ is generated by a marked point process with hazard $\lambda_A(t|\overline{A}(t^-), \overline{\mathbf{L}}(t^-))$, and (ii) we define $A(t) \equiv 0$ if $t > T$, then, with probability 1, sample paths of the **stochastic process** $A(t)$ are treatment regimes with discontinuity set the fixed or random jump times for the process depending on whether the process can jump only at fixed discrete times or in continuous time.

We assume that $\overline{\mathbf{L}}(T)$ as well as $\overline{A}(T)$ have cadlag sample paths w.p.1. We define the set *Dis* to be the possibly random set of times at which $\overline{\mathbf{L}}(T)$ or $\overline{A}(T)$ are discontinuous.

Definition. Given $\overline{a}$, let $U_{\overline{a}}$ be the (possibly) counterfactual survival time that would be observed if subjects followed treatment regime $\overline{a}$ until failure.

Note that the baseline failure time U is $U_{\overline{a}}$ for the function $\overline{a}$ that is everywhere zero.

Definition. Given $\overline{a} \equiv a(\cdot)$, let $(\overline{a}(t), 0)$ be the regime that agrees with $\overline{a}$ for $u \le t$ and is zero for $u > t$. $U_{\overline{a}(t),0}$ is the survival time had regime $\overline{a}$ been followed through t and treatment withheld after t.

We assume $U_{\overline{a}}$ obeys the following natural consistency assumptions that essentially assert that the future cannot affect the past.

Consistency Assumption A. Given $\overline{a}$ and $t > u$, the following are equivalent: $U_{\overline{a}(u),0} > u, U_{\overline{a}} > u$, $U_{\overline{a}(t),0} > u$.

Consistency Assumption B. Given $\overline{a}$ and $\overline{a}^*$ such that $\overline{a}(u) = \overline{a}^*(u)$, $U_{\overline{a}(u),0} = U_{\overline{a}^*(u),0}$.

The following consistency assumption links the counterfactual variables $U_{\overline{a}}$ to the observable variables $(T, \overline{A}(T))$.

Consistency Assumption C.

$$T = U_{\overline{A}(T),0} \text{ w.p.1.} \tag{20}$$

The instantaneous-rate RPSNFTM studied in this section requires the assumption of local rank preservation. In the following definition, parts (i) and

(ii) are the substantive parts. Part (iii) contains technical assumptions used later.

Definition of Local Rank Preservation. There is local rank preservation w.p.1 if:

(i) $U_{\overline{a}(t),0}$ is continuous in t w.p.1 and

$$\lim_{\Delta t \downarrow 0}\{U_{\overline{A}(t+\Delta t),0} - U_{\overline{A}(t),0}\}/\Delta t = D(U_{\overline{A}(t),0}, t) \text{ whenever } U_{\overline{A}(t),0} > t, \quad (21)$$

where

$$D(u,t) \equiv d\{u, t, \overline{\mathbf{L}}(t), \overline{A}(t)\}$$

(ii) $d(u, t, \overline{\mathbf{L}}(t), \overline{A}(t)) = 0$ if $a(t) = 0$;

(iii) for $t \notin Dis$ on which $\overline{A}(T)$ or $\overline{\mathbf{L}}(T)$ is discontinuous, $D(u,t)$ is bounded and its partial derivatives with respect to u and t are bounded and uniformly continuous.

Eq. (21) states that if $U_{\overline{A}(t),0} > t$, then, for infinitesimal positive Δt,

$$U_{\overline{A}(t+\Delta t),0} - U_{\overline{A}(t),0} = D(U_{\overline{A}(t),0}, t)\Delta t. \quad (22)$$

Now recall that by $\overline{A}(t)$ cadlag w.p.1, $A(t)$ is constant in $[t, t+\Delta t)$ for Δt sufficiently small. The left-hand side of (22) is the additive increment in survival time attributable to a final blip of treatment $A(t)\Delta t$ in the interval $[t, t+\Delta t)$ administered at dose rate $A(t)$. Thus (22) states that the additional increment $D(U_{\overline{A}(t),0}, t)\Delta t$ is deterministic function of $\overline{\mathbf{L}}(t)$, $\overline{A}(t)$ and $U_{\overline{A}(t),0}$. Hence we will refer to $D(u,t)$ as the instantaneous blip function.

To help understand the meaning of the instantaneous blip function $D(u,t)$, consider the infectious disease example presented earlier; it follows from (22) that the restriction on the instantaneous blip function $D(u,t)$ implied by the biological knowledge that the treatment received $A(t)$ at time t is only harmful or beneficial to those destined to fail by $t+5$ if they receive no further treatment (i.e. to those with $U_{\overline{A}(t),0} - t < 5$) is that

$$D(u,t) = 0, \quad \text{if } u - t > 5. \quad (23)$$

We now define a instantaneous-rate RPSNFT model under the assumption of local rank preservation.

Definition. If the assumption of local rank preservation holds, then we say the data follow an instantaneous-rate RPSNFT $D(u,t,\psi)$ if $D(u,t) = D(u,t,\psi_0)$, where ψ_0 is an unknown parameter and $D(u,t,\psi) \equiv \mathrm{d}(u,t,\overline{\mathbf{L}}(t),\overline{A}(t),\psi)$ is a known continuously differentiable function of ψ satisfying (i) $D(u,t,0) = 0$, (ii) $D(u,t,\psi) = 0$ if $A(t) = 0$, and (iii) for each fixed value of ψ, assumption (iii) in the definition of local rank preservation holds.

We will now show that any instantaneous-rate RPSNFTM implies a unique deterministic SNFTM (10). We first show that the instantaneous-rate RPSNFTM

$$D(u,t,\psi) = 1 - \exp\{\psi A(t)\} \quad (24)$$

implies the deterministic SNFTM (11). Model (24) states that the effect of a final instantaneous brief bit of treatment $A(t)\Delta t$ is to add or subtract $[1 - \exp\{\psi_0 A(t)\}]\Delta t$ to a subject's lifetime, so that $\psi_0 = 0$ implies no effect of treatment on survival. The regularity conditions in assumption (iii) of the definition of local rank preservation and Theorem 2.3 of [9, Section 6] on the existence and uniqueness of solutions to differential equations guarantee that w.p.1 there exists a unique continuous solution $U(t) \equiv U_{\overline{A}(t),0}$ to the differential equation

$$\frac{\mathrm{d}U(t)}{\mathrm{d}t} = D\{U(t), t\} \quad (25)$$

satisfying consistency assumption C that $U(T) \equiv T$.

We now solve the differential equations (25) corresponding to model (24). Integrating $\mathrm{d}U(t)/\mathrm{d}t = 1 - \exp\{\psi_0 A(t)\}$, we obtain $U(t) = t - \int_0^t \exp\{\psi_0 A(u)\}\,\mathrm{d}u + c$. Imposing the initial conditions $U(T) = T$ of consistency assumption C, we obtain $c = \int_0^T \exp\{\psi_0 A(u)\}\,\mathrm{d}u$ so $U(t) = t + \int_t^T \exp\{\psi_0 A(u)\}\,\mathrm{d}u$. Hence $U \equiv U(0) = \int_0^T \exp\{\psi_0 A(u)\}\,\mathrm{d}u$, reproducing model (10)–(11) as promised. It is interesting to note that the additive effect $\{1 - \exp[\psi_0 A(t)]\}\Delta t$ of the treatment $A(t)\Delta t$ implies, by model (11), a multiplicative effect of constant unit treatment, i.e. $U_{\overline{a}\equiv 1} = \exp(-\psi_0)U$ where $\overline{a} \equiv 1$ is the regime that always gives unit treatment. The reason for this is a "compound interest effect" of continuous treatment: any additional increment of survival time due to treatment received at t is itself later subjected to treatment, adding a further increment to survival time, etc. Summing the resulting "infinite series" produces the multiplicative effect on survival time of constant treatment.

More generally, for any instantaneous-rate RPSNFTM $D(u,t,\psi)$, there exists a unique solution

$H(t, \psi)$ to the differential equation

$$\frac{\partial H(t, \psi)}{\partial t} = D(H(t, \psi), t, \psi) \qquad (26)$$

satisfying the initial condition $H(T, \psi) = T$. Note $H(t, \psi)$ is a function $h(t, T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi)$ of ψ and the data. If the RPSNFTM is correctly specified with true value ψ_0, then we have by the uniqueness of the solutions to (25) and (26) that $H(t, \psi_0) = U(t) \equiv U_{\overline{A}(t),0}$. In particular, abbreviating $H(0, \psi_0)$ to $H(\psi_0)$ and $h(0, T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi_0)$ to $h(T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi_0)$, we obtain the unique deterministic SNFTM $U = H(\psi_0)$ of (10).

It follows that a CAN g-estimator $\hat{\psi}$ of the parameter ψ_0 of the instantaneous-rate RPSNFTM can be obtained by g-estimation under the assumptions described earlier (including the assumption (1) of no unmeasured confounders).

Consider next the instantaneous-rate RPSNFTM

$$D(u, t, \psi) = I(u - t < 5)\{1 - \exp\{\psi A(t)\}\}, \quad (27)$$

which satisfies the assumption (23) that treatment at t only affects those destined to fail by $t + 5$ in the absence of further treatment. Integrating (25) with $D(u, t) = D(u, t, \psi_0)$ and imposing the initial condition $U(T) = T$, we obtain $U(t) = t - \int_t^T \exp\{\psi_0 A(u)\}\, du$ for $U(t) - t \leq 5$. For $U(t) - t > 5$, $U(t)$ solves $U(t) = \{U(t) - 5\} + \int_{U(t)-5}^{T} \exp\{\psi_0 A(u)\}\, du$, i.e. $\int_{U(t)-5}^{T} \exp\{\psi_0 A(u)\}\, du = 5$. It follows that $U \equiv U(0) = \int_0^T \exp\{\psi_0 A(u)\}\, du$ for $U < 5$ and U satisfies $\int_{U-5}^{T} \exp\{\psi_0 A(u)\}\, du = 5$ when $U > 5$. This implies that $\partial U/\partial T = \exp\{\psi_0 A(T)\}$ when $U < 5$ and $\partial U/\partial T = \exp[\psi_0\{A(T) - A(U - 5)\}]$ when $U \geq 5$. It follows that when $A(u)$ varies with time u, we do not have a closed-form expression for U or for $\partial U/\partial T$. Hence, for the corresponding deterministic SNFTM (10), we will not have a closed-form expression for the function $H(\psi) = h(T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi)$ or its derivative $\partial H(\psi)/\partial T$, although $H(\psi)$ is easily evaluated by numerical means. This "nonobvious" form of $\partial H(\psi_0)/\partial T = \partial U/\partial T$ justifies the remarks of the previous subsection.

Instantaneous-Rate SNFTMs

Biological Implausibility of Local Rank Preservation

Consider two subjects, say i and j, who have identical survival times and covariate and treatment histories $(T, \overline{A}(T), \overline{\mathbf{L}}(t))$. It follows from the uniqueness of the solution to the differential equation (25) that, under the assumption of local rank preservation, the two subjects would have identical survival times U if treatment had been withheld. This assumption is biologically implausible. To see why, again consider the infectious disease example of the previous section. Suppose $A(t)$ is the dose of treatment taken at t, treatment has a beneficial biological affect, subject i and j are both infected at time t, subject i fails to absorb his/her dose due to gastrointestinal difficulties, while subject j successfully absorbs his/her dose. Then we would expect $U_i = T_i = T_j > U_j$ since subject j but not subject i experiences the benefit of treatment. Dependence of the magnitude of the treatment effect on unmeasured factors such as bioabsorption and genetic endowment is the rule. In this section, we describe the general class of instantaneous-rate SNFTMs which allow the magnitude of the treatment effect to depend on unmeasured factors. Specifically, the class of instantaneous-rate SNFTMs does not require that U be a deterministic function of $\{T, \overline{A}(T), \overline{\mathbf{L}}(T)\}$, and contains the instantaneous-rate RPSNFTMs as a subclass. Furthermore, the parameter ψ of a instantaneous-rate SNFTM can be consistently estimated using the g-estimation procedures described previously.

We now define a new function that will allow us to relax the assumption of local rank preservation. Given continuously distributed failure time variates T_1 and T_2 with survivor functions $S_1(u)$ and $S_2(u)$, recall that the quantile-quantile function $v(u) = S_1^{-1}\{S_2(u)\}$ is the unique function $v(u)$ such that $v(T_2)$ has the same distribution $S_1(u)$ as T_1. We now let $U_{\overline{A}(t),0}$ and $U_{\overline{A}(t+h),0}$ play the roles of T_1 and T_2 where, by convention, $A(u) \equiv 0$ if $u > T$. Specifically, let $\mathcal{V}(u, t, h) \equiv v(u, t, h, \overline{\mathbf{L}}(t), \overline{A}(t))$ be the unique function such that $U_{\overline{A}(t+h),0}$ and $\mathcal{V}(U_{A(t),0}, t, h)$ have the same conditional distribution given $\overline{\mathbf{L}}(t)$, $\overline{A}(t)$, $T > t$. That is,

$$\Pr[U_{\overline{A}(t+h),0} > \mathcal{V}(u, t, h) | \overline{\mathbf{L}}(t), \overline{A}(t), T > t]$$
$$= \Pr[U_{\overline{A}(t),0} > u | \overline{\mathbf{L}}(t), \overline{A}(t), T > t].$$

Note $\mathcal{V}(u, t, 0) = u$. We now make a smoothness (differentiability) assumption.

Assumption (*). We assume that (i) $D(u, t) \equiv \lim_{h \downarrow 0}\{\mathcal{V}(u, t, h) - \mathcal{V}(u, t, 0)\}/h$ exists and is bounded for all (u, t) w.p.1 where the function $D(u, t) \equiv d(u, t, \overline{L}(t), \overline{A}(t))$ satisfies assumption (iii) in the definition of local rank preservation, and (ii), for $t > x$,

$$\Pr[U_{\overline{A}(t),0} > u | \overline{\mathbf{L}}(x), \overline{A}(x)] \text{ is continuous in } t. \quad (28)$$

We have reused the notation $D(u, t)$ in Assumption (*) because, under local rank preservation, $D(u, t)$ as just defined is equal to $D(u, t)$ as defined previously. Even without local rank preservation under Assumption (*), we can regard $D(u, t)\Delta t$ as the effect of a last blip of observed treatment $A(t)$ at t sustained for an instantaneous time Δt on quantiles of $U_{\overline{A}(t),0}$. That is, for infinitesimal positive Δt, if, conditional on $\overline{\mathbf{L}}(t)$, $\overline{A}(t), u$ is, say, the zth quantile of $U_{\overline{A}(t),0}$, then $u + D(u, t)\Delta t$ is the zth quantile of $U_{\overline{A}(t+\Delta t),0}$. As before, $D(u, t)$ may be discontinuous for $t \in$ Dis.

Remark. It is important to note that we no longer assume $U_{\overline{A}(t),0}$ is continuous in t. This is scientifically important since $U_{\overline{A}(t),0}$ will be discontinuous at t if $A(\cdot)$ is exposure to cigarette smoke and a single molecule of benzpyrene inhaled at time t initiates lung cancer. However, (28) remains reasonable since the probability of lung cancer being initiated in $[t, t + \Delta t)$ is small. However, (28) and thus Assumption (*) would be inappropriate if $A(t)$ recorded whether a subject received a mammogram or any other truly "point-source" exposure at time t, where $A(t)$ is a point source if exposure of $\{t; A(t) \neq 0\}$ is a finite set w.p.1. Models for the effect of point-source exposures, such as mammography, will not be further discussed in this chapter; the SNFTMs discussed in [16, Appendix 2] are appropriate.

It then follows from Theorem 2.3 of [9, Section 6] that there exists a unique continuous solution $\mathcal{H}(t) \equiv h(t, T, \overline{\mathbf{L}}(T), \overline{A}(T))$ to the differential equation $\mathrm{d}\mathcal{H}(t)/\mathrm{d}t = D(\mathcal{H}(t), t)$ satisfying $\mathcal{H}(T) = T$.

Under local rank preservation, we have seen that this unique solution $\mathcal{H}(t)$ is precisely $U(t) \equiv U_{\overline{A}(t),0}$. This will not be true in the absence of local rank preservation, since $U(t)$ will no longer satisfy (25). However, our main result is the following theorem which states that $\mathcal{H}(t)$ and $U(t)$ continue to have the same conditional distributions.

Theorem 1. $\mathcal{H}(t)$ and $U_{\overline{A}(t^-),0}$ have the same conditional distribution given $(\overline{\mathbf{L}}(t), \overline{A}(t), T > t)$. In particular, $\mathcal{H} \equiv \mathcal{H}(0)$ and U have the same marginal distributions.

As yet, Theorem 1 has only been proved in the special case where the jump times of the $\overline{A}(t)$ and $\overline{\mathbf{L}}(t)$ processes are fixed rather than random [22], although it is almost certain that Theorem 1 holds in general. The limitation to nonrandom jump times for the measured $\mathbf{L}$ and A processes is no limitation in practice, since we can suppose them to have been measured, say, every second rather than continuously, and then the theorem is true.

We say the data follow a instantaneous-rate SNDM if there is a function $D(u, t, \psi)$ satisfying $D(u, t) = D(u, t, \psi_0)$ with $D(u, t, \psi)$ satisfying the conditions previously described under the definition of a instantaneous-rate RPSNFTMs.

Again letting $H(t, \psi) \equiv h(t, T, \overline{A}(T), \overline{\mathbf{L}}(T), \psi)$ be the solution to the differential equation (26) and setting $H(\psi) \equiv H(0, \psi)$, it immediately follows by uniqueness that $\mathcal{H} = H(\psi_0)$. Hence, a CAN estimator $\hat{\psi}$ of the causal parameter ψ_0 of a instantaneous-rate SNFTM can be obtained by g-estimation if, as we assume, (2) holds with $U_{\overline{A}(t^-),0}$ replacing U. .

Estimating the Distribution of $U_{\overline{a}}$

Given an instantaneous-rate SNFTM $D(u, t, \psi)$, we have shown how to obtain a CAN estimator $\hat{\psi}$ of ψ_0 under the assumptions given above. However, we often wish to estimate the survival curves $S_{U_{\overline{a}}}(t)$ of $U_{\overline{a}}$ for various treatment regimes $\overline{a}$. Suppose censoring is absent. Then $\hat{S}_U(t) = n^{-1}\sum_i I\{H_i(\hat{\psi}) > t\}$ is a CAN estimator of $S_U(t)$. The main tool we shall use to estimate other $S_{U_{\overline{a}}}(t)$ is the blip-up function $B(u, t) \equiv b(u, t, \overline{\mathbf{L}}(t), \overline{A}(t))$ defined to be the unique continuous solution to $\mathrm{d}B(t)/\mathrm{d}t = D(B(t), t)$ through $(0, u)$.

Example

For model (24) with $D(u, t) = 1 - \exp\{\psi_0 A(t)\}$, we obtain upon integrating that $B(t) = t - \int_0^t \exp\{\psi_0$

$A(u)\}\,du + c$. By the initial condition $B(0) = u$, we obtain $c = u$, so

$$B(u,t) \equiv B(t) = u + t - \int_0^t \exp\{\psi_0 A(u)\}\,du. \tag{29}$$

In general, $B(u,t)$ is related to the blip-down function $\mathcal{H}(t)$ by $B(\mathcal{H},t) = \mathcal{H}(t)$ where $\mathcal{H} \equiv \mathcal{H}(0)$. For any covariate and treatment histories $\bar{\mathbf{l}}$ and $\bar{a}$ defined on $[0,\infty)$, define $b^*(u,\bar{\mathbf{l}},\bar{a})$ to be the solution t^* to $t^* = b(u,t^*,\bar{\mathbf{l}}(t^*),\bar{a}(t^*))$ if one exists and $b^*(u,\bar{\mathbf{l}},\bar{a}) \equiv \infty$ otherwise. Note $b^*(\mathcal{H},\overline{\mathbf{L}},\overline{A}) = T$, since $T = B(\mathcal{H},T)$, where $\mathbf{L}(u) \equiv A(u) \equiv 0$ if $u > T$.

Example

With $B(u,t)$ given by (29), $b^*(u,\bar{\mathbf{l}},\bar{a}) = b^*(u,\bar{a})$ is the unique solution to $u = \int_0^{b^*(u,\bar{a})} \exp\{\psi_0 a(u)\}\,du$.

If (i), as in model (24), $d(u,t,\bar{\mathbf{l}}(t),\bar{a}(t)) \equiv d(u,t,\bar{a}(t))$ does not depend on $\bar{\mathbf{l}}(t)$, i.e. there is no treatment–covariate interaction, so $b^*(u,\bar{\mathbf{l}},\bar{a}) = b^*(u,\bar{a})$, and (ii) there are no unmeasured confounders for each $U_{\bar{a}}$, i.e.

$$U_{\bar{a}} \perp\!\!\!\perp A(t) | \overline{\mathbf{L}}(t^-), \overline{A}(t^-), T > t, \tag{30}$$

then

$$S_{U_{\bar{a}}}(t) = \Pr\{b^*(U,\bar{a}) > t\}. \tag{31}$$

By Theorem 1, U can be replaced by $\mathcal{H}$ in (31). Robins [17, Appendix 1] discusses conditions weaker than (30) which imply (31). It now follows that given a CAN g-estimator $\hat{\psi}$ of ψ_0, $n^{-1}\sum_i I\{b^*(H_i(\hat{\psi}),\bar{a},\hat{\psi}) > t\}$ is a CAN estimator of $S_{U_{\bar{a}}}(t)$ where $b^*(u,\bar{a},\psi)$ is $b^*(u,\bar{a})$ under $\psi = \psi_0$.

Models for Cure

The fact that $b^*(u,\bar{a})$ can be infinite reflects the possibility of "cure" (*see* **Cure Models**). As an example, suppose U represents the time from diagnosis to death from pancreatic cancer in the absence of treatment. Suppose untreated pancreatic cancer is uniformly fatal so that U is finite w.p.1. Suppose, however, that $\Pr\{b^*(U,\bar{a}) = \infty\} = p \neq 0$. Then a fraction p of the population will be cured under treatment regime $\bar{a}$.

Consider the multiplicative blip model

$$D(u,t,\psi) = (u-t)\psi A(t). \tag{32}$$

By the formula for solutions to linear first-order differential equations [9, Chapter 6],

$$b(u,t,\bar{a}(t)) = \exp\left[\int_0^t \psi_0 a(x)\,dx\right]\left\{u - \int_0^t x\psi_0 a(x) \times \exp\left[-\int_0^x \psi_0 a(v)\,dv\right]dx\right\},$$

which simplifies, when $\bar{a}$ is the constant-dose regime a^*, to $\exp(\psi_0 a^* t)[u - (\psi_0 a^*)^{-1}] + t + (\psi_0 a^*)^{-1}$. Hence, $b^*(u,\bar{a}) = \ln\{-(\psi_0 a^*)^{-1}/[u - (\psi_0 a^*)^{-1}]\}/\{\psi_0 a^*\}$ if $u < (\psi_0 a^*)^{-1}$ and $b^*(u,\bar{a}) = \infty$ if $u > (\psi_0 a^*)^{-1}$. Thus, the probability of cure under $\bar{a}$ is the probability that U exceeds $1/\{\psi_0 a^*\}$. The intuition behind this result is that, according to model (32), $d(u,t,\bar{a}(t),\psi_0)$ exceeds 1 at $t = 0$ if and only if $u\psi_0 a^* > 1$. If $d(u,t,\bar{a}(t),\psi_0) > 1$, then a blip of treatment a^* at t sustained for duration Δt adds more than Δt years to a subject's survival time. If, as in our model when $u > (\psi_0 a^*)^{-1}$, the instantaneous blip function exceeds 1 for each time t, then the subject is cured.

If the failure time variable is death from all causes, we would want $b^*(u,\bar{\mathbf{l}},\bar{a})$ to be finite for all $u,\bar{\mathbf{l}}$, and $\bar{a}$ which is guaranteed by having $d(u,t,\bar{\mathbf{l}}(t),\bar{a}(t)) < 1-\sigma, \sigma > 0$, for all $u,t,\bar{a}(t),\bar{\mathbf{l}}(t)$. A natural parameterization of an instantaneous-rate SNFTM that essentially accomplishes this is $D(u,t;\psi) = 1 - \exp\{r(u,t,\overline{\mathbf{L}}(t),\overline{A}(t),\psi)\}$ for some function $r(\cdot)$ as in models (24) and (27).

Covariate–Treatment Interaction

If $d(u,t,\bar{\mathbf{l}}(t),\bar{a}(t))$ depends on $\bar{\mathbf{l}}(t)$, we can obtain independent draws from the distribution of $U_{\bar{a}}$ under assumption (30) when the covariate process $\mathbf{L}(t)$ only jumps at non random times, say 0,1,2, ... as follows.

Step 1. Draw U from its marginal distribution.
Step 2. Draw $\mathbf{L}(0)$ from $f\{\mathbf{l}(0)|U = u\}$.
Step 3. Set $m = 1$.
Step 4. If $b[U, m-1, \overline{\mathbf{L}}(m-1), \bar{a}(m-1)] \leq m$, set $U_{\bar{a}}$ to $b^*(U,\overline{\mathbf{L}},\bar{a})$ where $\mathbf{L}(t) = 0$ for $t > m$ and agrees with the drawn $\overline{\mathbf{L}}(m^-)$ up to time m. Otherwise, draw $\mathbf{L}(m)$ from $f[\mathbf{L}(m)|(\overline{\mathbf{L}}(m^-),\bar{a}(m^-),U,T > m]$,

increment m by 1, and return to the start of this step.

To carry out this algorithm in practice, we first obtain a g-estimate $\hat{\psi}$ of the parameter ψ_0 of an instantaneous-rate SNFTM; draw U from the empirical distribution $\hat{S}_U(t) = n^{-1}\sum_i I(H_i(\hat{\psi}) > t)$; replace the functions $b(\cdot)$ and $b^*(\cdot)$ by $b(\cdot, \hat{\psi})$ and $b^*(\cdot, \hat{\psi})$; and estimate the density $f[\bar{\mathbf{l}}(m)|\bar{\mathbf{l}}(m^-), \overline{a}(m^-), U, T > m]$ by specifying a parametric model $f[\mathbf{l}(m)|\bar{\mathbf{l}}(m^-), \overline{a}(m^-), U, T > m; \eta]$ and evaluating it at $\hat{\eta}$ which maximizes

$$\prod_{i=1}^{n}\prod_{m=0}^{\text{int}(T_i)} f[\mathbf{L}_i(m)|\overline{\mathbf{L}}_i(m^-), \overline{A}_i(m^-), H_i(\hat{\psi}), T_i > m; \eta].$$

A heuristic explanation of the above algorithm is as follows. If a stimulated subject with baseline time U manages to survive to time m under regime $\overline{a}$, we randomly draw $\mathbf{L}(m)$ and then use the blip-up function $b(u, t, \bar{\mathbf{l}}(t), \overline{a}(t))$ to determine whether the subject has survived to time $m+1$ or whether the subject has died at a time $U_{\overline{a}}$, determined by the function $b^*(u, \bar{\mathbf{l}}, \overline{a})$ in the interval $(m, m+1]$. This explanation is heuristic in that it implicitly but unnecessarily assumes local rank preservation. Robins et al.[26, Appendix 2] discusses how to generalize the results of this section to allow for censoring. A drawback of SNFTMs is that in the presence of a covariate–treatment interaction $S_{U_{\overline{a}}}(t)$ cannot be calculated without modeling the law of $\mathbf{L}(m)$ given $\{\overline{\mathbf{L}}(m^-), \overline{A}(m^-), U, T > m\}$. Estimation of the nonnested structural Cox proportional hazard models for $U_{\overline{a}}$ described in Appendix 2 can obviate this problem.

Appendix 1: Calculation of the Variance of the g-Test Statistic Numerator

We shall require some notation. Given a stochastic process $G(\cdot)$, let $U_A\{G(\cdot)\} = \int_0^X \mathrm{d}M_A(u)\{G(u) - \mathrm{E}[\exp[\boldsymbol{\alpha}'\mathbf{W}(u)]G(u)]/\mathrm{E}[\exp[\boldsymbol{\alpha}'\mathbf{W}(u)]\}$ where $\boldsymbol{\alpha}'\mathbf{W}(u)$ is from model (4), $\mathrm{d}M_A(u) = \mathrm{d}N_A(u) - \lambda_0(u)\exp[\boldsymbol{\alpha}'\mathbf{W}(u)]\,\mathrm{d}u$ and $N_A(u)$ counts the number of jumps in the $A(u)$ process through time u.

Now define $U_1 \equiv U_A\{\tau Q(\cdot, \psi_0)/K(X)\}$ and $\boldsymbol{U_2 \equiv U_A\{\mathbf{W}(\cdot)\}}$. Then define $V_1 \equiv \mathrm{E}[\{U_1 - \mathrm{E}[U_1U_2']\,\mathrm{E}[U_2^{\otimes 2}]^{-1}U_2\}^{\otimes 2}]$. V_1 is the "robust variance" of the g-test numerator (i.e. Cox partial likelihood score test numerator) of the hypothesis that $\theta = 0$ in the extended Cox model (4) when the true $K(X)$ is used.

Now define $V \equiv V_1 - V_{\text{corr}}$, where V_{corr} is the correction to the variance V_1 required when we replace $K(X)$ by its estimator $\hat{K}(X)$. Specifically, $V_{\text{corr}} = V_2 + V_3$, where $V_2 = \mathrm{E}[\int_0^\infty \mathrm{d}N_Q(u)\{\mathcal{L}^Q(u, J(u))\}^{\otimes 2}]$, $N_Q(u) = I[Q \le u, \tau = 0]$, $J(u) \equiv \int_0^X \mathrm{d}M_A(t)\,Q(t, \psi_0)/K(u)$ and, for any $G(u)$, $\mathcal{L}^Q\{u, G(u)\} \equiv \mathrm{E}[K(u)\ \{K(X)\}^{-1}\ \tau\ G(u)\ I(X^* > u)\exp[\boldsymbol{\alpha}^{*'}\mathbf{W}^*(u)]/\mathrm{E}[I(X^* > u)\exp[\boldsymbol{\alpha}^{*'}\mathbf{W}^*(u)]$ with $\boldsymbol{\alpha}^{*'}\mathbf{W}^*(u)$ from model (19).

$V_3 = V_{31}\{V_{32}\}^{-1}V_{31}'$, $V_{31} = \mathrm{E}[\int \mathrm{d}N_Q(u)\{\mathcal{L}^Q\{u, J(u)\mathbf{W}^*(u)\} - \mathcal{L}^Q(u, J(u))\mathcal{L}^Q(u, \mathbf{W}^*(u))\}]$ and V_{32} is the expected partial information matrix for $\boldsymbol{\alpha}^*$ from Cox model (19). Since V_{corr} is nonnegative definite, the robust variance V_1 is greater than or equal to the true variance V in the nonnegative definite sense. A consistent estimator $\hat{V}$ of V is obtained by substitution into the above formulas according to the following six steps.

1. Replace any expectation by a sample average over the n study subjects.
2. Replace $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ by their partial maximum likelihood estimates.
3. Replace ψ_0 by the value of ψ being tested in the g-test.
4. Replace $K(\cdot)$ by $\hat{K}(\cdot)$.
5. Replace $\mathrm{d}M_A(u)$ by $\mathrm{d}N_A(u) - \mathrm{d}\hat{\Lambda}_0(u)\exp[\hat{\boldsymbol{\alpha}}'\mathbf{W}(u)]$, where $\hat{\Lambda}_0(u)$ is the Cox cumulative hazard estimate from model (4).
6. Estimate V_{32} by the observed partial information matrix from the fit of model (19).

Now let $\hat{U}_1(\psi)$ be U_1 with the substitutions described in steps 1–6 above. The g-statistic numerator is precisely $\sum_{i=1}^n \hat{U}_{1i}(\psi)$. We then have the following theorem on which our inferences are based.

Theorem. Given that (2) and (18) and that models (4), (10), and (19) are correct, then, when $\psi = \psi_0$, $n^{-1/2}\sum_i \hat{U}_{1i}(\psi)$ is asymptotically normal with mean zero and asymptotic variance V that can be consistently estimated by $\hat{V}$.

The proof is analogous to that given in [17, pp. 284–285] using the methods developed in [24].

Appendix 2

A (nonnested) structural Cox proportional hazard model specifies

$$\lambda_{U_{\bar{a}}}(t) = \lambda(t)\exp[r\{t, \overline{a}(t^-), \beta_0\}] \qquad (33)$$

where $r(\cdot)$ is a known function and $\lambda(t)$ an unspecified baseline hazard. For simplicity, assume $A(t)$ is dichotomous. Then a consistent asymptotically normal estimator $\hat{\beta}$ of β_0 under assumptions (18) and (30) and models (4), (19), and (33) is the solution to the weighted Cox score equation for T

$$0 = \sum_i \int_0^\infty \left\{\frac{dN_{Ti}(u)}{\hat{\Omega}_i(u)}\right\} \times \left\{P_i(u,\beta) - \hat{\mathrm{E}}[P(u,\beta),\beta]/\hat{\mathrm{E}}[\mathbf{1}(u),\beta]\right\},$$

where (i) $N_T(u) = I[X^* \le u, X^* = T]$; (ii) $P(u,\beta) = p\left[u, \overline{A}(u^-), \beta\right]$ is a vector function of dimβ chosen by the analyst such as $\partial r\left(u, \overline{A}(u^-), \beta\right)/\partial\beta$; (iii) $\mathbf{1}(u) = 1$ is the identity; (iv) $\hat{\mathrm{E}}[J(u,\beta),\beta] = \sum_i I(X_i^* > u)\exp\left[r\left\{u, \overline{A}(u^-), \beta\right\}\right] J_i(u,\beta)/\hat{\Omega}_i(u)$; (v) $\hat{\Omega}_i(u) = \hat{K}_i(u)\hat{K}_{Ai}(u)$, $\hat{K}(u)$ as defined in the text; (vi) $\hat{K}_A(u) = \exp\left[-\int_0^u \hat{\lambda}_A\left[t|\overline{\mathbf{L}}(t^-), \overline{A}(t^-)\right]\mathrm{d}t\right] \prod_{\{t;t<\mu \text{ and } A(t)\neq A(t^-)\}} \hat{\lambda}_A\left(t|\overline{\mathbf{L}}(t^-), \overline{A}(t^-)\right)$, where $\hat{\lambda}_A\left[t|\overline{\mathbf{L}}(t^-), \overline{A}(t^-)\right] = \hat{\lambda}_0(u)\exp\left[\hat{\boldsymbol{\alpha}}'\mathbf{W}(u)\right]$ and $\hat{\lambda}_0(u)$ is now a kernel smoothed version of the Cox estimate of $\lambda_0(t)$ of model (4) as in [31]. $\hat{\beta}$ combined with the estimate $\hat{\lambda}(t) = \sum_i dN_{Ti}(t)/\left\{\hat{\Omega}_i(t)\hat{\mathrm{E}}\left[\mathbf{1}(t),\hat{\beta}\right]\right\}$ of $\lambda(t)$ produces an estimate of $\lambda_{U_{\bar{a}}}(t)$ and thus of $S_{U_{\bar{a}}}(t)$. In the above, we have assumed that Q is the minimum of time to loss to follow-up, competing risk, and end to follow-up as discussed in the remark in the section "Censoring by Competing Risks" above. Note that $\hat{K}_A(u)$ is a consistent estimate of probability that a subject would have his observed history $\overline{A}(u)$ through time u.

References

[1] Arjas, E. (1989). Survival models and martingale dynamics (with discussion), *Scandinavian Journal of Statistics* **15**, 177–225.

[2] Arjas, E. & Eerola, M. (1993). On predictive causality in longitudinal studies, *Journal of Statistical Planning and Inference* **34**, 361–384.

[3] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[4] Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion), *Journal of the Royal Statistical Society, Series B* **41**, 1–31.

[5] Holland, P.W. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **84**, 1074–1078.

[6] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[7] Klein, J.P., Keiding, N. & Copelan, E. (1993). Plotting summary predictions in multi-state survival models: probabilities of relapse in death and remission for bone marrow transplantation patients, *Statistics in Medicine* **12**, 2315–2332.

[8] Lin, D.Y. & Wei, L.J. (1989). The robust inference for the Cox proportional hazard model, *Journal of the American Statistical Association* **84**, 1074–1078.

[9] Loomis, B. & Sternberg, S. (1968). *Advanced Calculus*. Addison Wesley, Reading.

[10] Mark, S.D. & Robins, J.M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model, *Statistics in Medicine* **12**, 1605–1628.

[11] Mark, S.D. & Robins, J.M. (1993). A method for the analysis of randomized trials with compliance information: an application to the multiple risk factor intervention trial, *Controlled Clinical Trials* **14**, 79–97.

[12] Pearl, J. (1995). Causal diagrams for empirical research, *Biometrika* **82**, 669–690.

[13] Robins, J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect, *Mathematical Modelling* **7**, 1393–1512.

[14] Robins, J.M. (1987). Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect", *Computers and Mathematics with Applications* **14**, 923–945.

[15] Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in *Health Service Research Methodology: A Focus on AIDS*, L. Sechrest, H. Freeman & A. Mulley, eds. NCHSR, US Public Health Service, Washington, pp. 113–159.

[16] Robins, J.M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors, *Biometrika* **79**, 321–334.

[17] Robins, J.M. (1993). Analytic methods for estimating HIV treatment and cofactor effects, in *Methodological Issues of AIDS Mental Health Research*, D.G. Ostrow & R. Kessler, eds. Plenum Publishing, New York, pp. 213–290.

[18] Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models, *Communications in Statistics* **23**, 2379–2412.

[19] Robins, J.M. (1995). Estimating the causal effect of a time-varying treatment on survival using structural nested failure time models, *Statistica Neederlandica*, to appear.

[20] Robins, J.M. (1997). Causal inference from complex longitudinal data, in *Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics 120, M. Berkane, ed. Springer-Verlag, New York. pp. 69–117.

[21] Robins, J.M. (1997). Estimation and testing of direct effects by reparameterizing directed acyclic graphs, in *Causation and Computation*, G. Cooper & C. Glymour, eds. MIT/AAAI Press, Cambridge, Mass., to appear.

[22] Robins, J.M. (1998). Correction for non-compliance in bioequivalence trials, *Statistics in Medicine* **17**, 269–302.

[23] Robins, J.M. & Greenland, S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial, *Journal of the American Statistical Association* **89**, 737–749.

[24] Robins, J.M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, in *AIDS Epidemiology – Methodological Issues*, N. Jewell, K. Dietz & V. Farewell, eds. Birkhäuser, Boston, pp. 297–331.

[25] Robins, J.M. & Tsiatis, A. (1991). Correcting for non-compliance in randomized trials using rank-preserving structural failure time models, *Communications in Statistics* **20**, 2609–2631.

[26] Robins, J.M., Blevins, D., Ritter, G. & Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients, *Epidemiology* **3**, 319–336.

[27] Rosenbaum, P.R. (1984). Conditional permutation tests and the propensity score in observational studies, *Journal of the American Statistical Association* **79**, 565–574.

[28] Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been adversely affected by treatment, *Journal of the Royal Statistical Society, Series A*, **147**, 656–666.

[29] Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization, *Annals of Statistics* **6**, 34–58.

[30] Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

[31] Van der Laan, M.J. & Hubbard, A. (1997). Estimation with interval-censored data and covariates, *Lifetime Data Analysis* **3**, 77–91.

[32] Witteman, J.C.M., D'Agostino, R.B., Stijnen, T., Kannel, W.B., Cobb, J.C., de Ridder, M.A.J., Hofman, A. & Robins, J.M. (1997). G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Study, *American Journal of Epidemiology* **148**, 390–401.

JAMES M. ROBINS

Structural Time Series Models

Structural time series models are set up in terms of components, such as trends, seasonals, and cycles, which have a direct interpretation. They can then be used not only for **forecasting** but also for providing a description of the main features of the series. Structural time series models have been used to tackle a wide range of problems; for example, the monograph by Harvey [8] includes applications in economics, meteorology, criminology, energy, and association football. Other illustrations are in the books by Durbin and Koopman [4], Jones [17] and Kitagawa, and Geisch [18].

The next section sets out the principal univariate structural time series models and describes their properties. The following section explains how to bring **explanatory variables** into a single equation model, thereby bringing together regression and time series techniques. Intervention analysis (*see* **Intervention Analysis in Time Series**) is treated as a special case. In a multivariate model several series are modeled jointly. Structural time series models extend naturally to this situation and are described in the section on multivariate models. Of particular interest is the fact that such models provide a framework within which to handle control groups.

The section on data irregularities explains how structural time series models can be adapted to handle problems such as **outliers**, missing observations and data not recorded at equally spaced intervals of time. In the final section we look at methods for handling non-Gaussian data such as counts, consisting of small integers, or qualitative variables, recorded as zeros or ones.

Univariate Structural Time Series Models

The basic idea of structural time series models is that they are set up as regression models in which the explanatory variables are functions of time, but with coefficients which change over time. Thus within a regression framework a simple trend would be modeled in terms of a constant and time with a random disturbance added on. That is,

$$y_t = \alpha + \beta t + \varepsilon_t, \quad t = 1, \ldots, T. \tag{1}$$

The model is straightforwardly estimated by ordinary least squares, but suffers from the disadvantage that the trend is deterministic. This is too restrictive in general and the necessary flexibility is introduced by letting the coefficients α and β evolve over time as **stochastic processes**. In this way the trend can adapt to underlying changes. The current, or filtered, estimate of the trend is obtained by putting the model in state space form and applying the Kalman filter. Related algorithms are used for making predictions and for smoothing, which means computing the best estimate of the trend at all points in the sample using the full set of observations. The extent to which the parameters are allowed to change is governed by *hyperparameters*. These can be estimated by maximum likelihood but, again, the key to this is the state space form and the Kalman filter. All these methods and algorithms are described in detail in Harvey [8], Durbin and Koopman [4], and Koopman. The *STAMP* package of Koopman et al. [21] carries out all the calculations and is set up so as to leave the user free to concentrate on choosing a suitable model. If desired, the weights implicitly assigned to observations in computing a trend, or indeed any component, may be obtained by using the algorithm set out in [20].

The model selection methodology for structural models is somewhat different to that used for autoregressive integrated moving average (ARIMA) models (*see* **ARMA and ARIMA Models**) in that there is less emphasis on looking at the correlograms of various transformations of the series in order to get an initial specification; see Box & Jenkins [3]. Instead the emphasis is on formulating the model in terms of components suggested by a knowledge of the application or an inspection of the graph. Once a model has been estimated, the same type of diagnostics tests as are used for ARIMA models can be performed on the residuals. In particular, the Box–Ljung statistic can be computed, with the number of relative hyperparameters subtracted from the number of residual autocorrelations to allow for the loss in degrees of freedom. Standard tests for nonnormality and heteroscedasticity can also be carried out, as can tests of predictive performance in a postsample period. The structural framework also allows tests for outliers and structural breaks to be performed; these are based on what Harvey & Koopman [11] call "auxiliary residuals". Plots of residuals can be augmented by graphs of the smoothed components. These can often be very informative since they enable the model builder to

check whether the movements in the components correspond to what might be expected on the basis of prior knowledge. Finally, it is shown by Andrews [1], among others, that the forecasting performance of structural time series models is as good as, and sometimes even better than, the performance of ARIMA models.

A Bayesian approach to structural time series modeling is described in West & Harrison [28]. Recent developments in Bayesian techniques, such as Markov chain Monte Carlo, are discussed in [4] and [6].

The principal univariate structural time series models are set out below. In the last subsection the formal relationship between structural and ARIMA models is discussed.

Local Level

The simplest structural time series model addresses a situation in which the underlying level of the series changes over time. This level is modeled by a random walk, on top of which is superimposed a random, or white noise (*see* **Noise and White Noise**) disturbance. The specification is therefore

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}\ (0, \sigma_\varepsilon^2), \quad t = 1, \ldots, T, \\ \mu_t &= \mu_{t-1} + \eta_t, \quad \eta_t \sim \text{NID}\ (0, \sigma_\eta^2), \end{aligned} \tag{2}$$

where NID denotes normally and independently distributed, and the two disturbances are mutually uncorrelated. An important practical feature of this model is that the estimator of the level, based on currently available information, is given by an exponentially weighted moving average (EWMA) of past observations, that is

$$\begin{aligned} \tilde{\mu}_T = \lambda[&y_T + (1-\lambda)y_{T-1} + (1-\lambda)^2 y_{T-2} \\ &+ (1-\lambda)^3 y_{T-3} + \cdots], \end{aligned} \tag{3}$$

where the "smoothing constant", λ, is a function of the *signal-to-noise ratio*, $q = \sigma_\eta^2/\sigma_\varepsilon^2$. Forecasts of future observations, however many steps ahead, are given by exactly the same expression. This was established by Muth [24]. For a pure random walk, q is infinite and $\lambda = 1$, leading to a forecast equal to the last observation. As q moves towards 0, λ also goes to zero and the forecast becomes the sample mean. That this happens is not apparent from the formula above, which is actually an approximation to the exact forecast function produced by the Kalman filter [8, Chapter 3]. The important point about the Kalman filter when applied to a nonstationary model like (2) is that it is initialized with what is called a diffuse *prior*. There has been considerable progress in the statistics literature recently on developing stable algorithms for this situation [19]. The estimation of q, the main hyperparameter in the model, can be carried out by maximum likelihood, which basically entails minimizing the sum of squares of the one-step-ahead prediction errors throughout the sample.

Within a medical context, μ_t might be thought of as the underlying state of a patient, while ε_t represents measurement error. A doctor wishing to monitor the state of the patient's health needs to have as good an estimate as possible of μ_t given the information which is currently available. If the model is correct, the Kalman filter delivers such an estimate, together with its root mean square error (RMSE). Hence a confidence interval can be constructed around the estimate, and the doctor can use this as a guide as to whether it is reasonable to suppose that the level of whatever is being monitored is acceptable see [12].

Local Linear Trend

The local linear trend model replaces the deterministic trend in (1) by a stochastic trend. The exact formulation is

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim \text{NID}\ (0, \sigma_\varepsilon^2), \\ & & t &= 1, \ldots, T, \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, & \eta_t &\sim \text{NID}\ (0, \sigma_\eta^2), \\ \beta_t &= \beta_{t-1} + \zeta_t, & \zeta_t &\sim \text{NID}\ (0, \sigma_\zeta^2), \end{aligned} \tag{4}$$

with the level and slope disturbances, η_t and ζ_t mutually uncorrelated and uncorrelated with ε_t. The extent to which the level, μ_t, and slope, β_t, change over time is governed by the relative hyperparameters, $q_\eta = \sigma_\eta^2/\sigma_\varepsilon^2$ and $q_\zeta = \sigma_\zeta^2/\sigma_\varepsilon^2$. The forecast function is a straight line starting from the estimates of the level and slope at the end of the sample. In the limiting case when both relative hyperparameters are zero, the deterministic trend model is obtained with $\alpha = \mu_0$. Other special cases of interest arise when $q_\zeta = 0$, in which case the trend is a random walk plus drift and $q_\eta = 0$, in which case the smoothed trend is related to a cubic spline (*see* **Spline Function**).

Stochastic Seasonality

Many series recorded quarterly or monthly are subject to seasonal variation. What are effectively seasonal effects also occur when observations are recorded within the day. Just as more flexibility needs to be given to a trend by allowing it to be stochastic, so a seasonal component needs to be allowed to change over time. Although the case for a stochastic seasonal component is arguably less compelling than the case for a stochastic trend, there are many reasons why changes in the seasonal pattern may take place.

If the seasonal component is deterministic, it should have the property that it sums to zero over the previous year; this ensures that it cannot be confounded with the trend. Adding a disturbance term to the sum of seasonal effects over the past year allows the seasonal pattern to evolve over time. This is a "dummy variable" form of stochastic seasonality,

$$\gamma_t = \gamma_{t-1} + \cdots + \gamma_{t-s+1} + \omega_t,$$
$$\omega_t \sim \text{NID}\ (0, \sigma_\omega^2). \tag{5}$$

An alternative way of capturing a deterministic seasonal pattern is by a set of sine and cosine functions. Allowing these to be stochastic leads to the "trigonometric form" of stochastic seasonality [8, Chapter 2]. It is better than the dummy variable stochastic seasonal model because it allows the seasonal pattern to evolve more smoothly; it can be shown that the sum of the seasonals over the past year follows a MA $(s-2)$ rather than white noise.

Seasonal effects are typically combined with trend and irregular components, usually after taking logarithms. This leads to the basic structural model (BSM),

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \quad t = 1, \ldots, T, \tag{6}$$

where the stochastic trend component, μ_t, and the irregular component are defined as in the local linear trend model above; see, for example, [18]. This model has been widely applied and can be used as the basis for seasonal adjustment of time series.

Cycle

A deterministic cycle can be expressed as a sine–cosine wave, that is

$$\psi_t = \alpha \cos \lambda t + \beta \sin \lambda t, \quad t = 1, \ldots, T. \tag{7}$$

In the previous section it was pointed out that a seasonal pattern could be modeled by a set of stochastic cycles defined at the seasonal frequencies (*see* **Seasonal Time Series**). A somewhat different situation arises when we wish to model a cycle, which may be stochastic, and, unlike the seasonal cycles, may be stationary. The statistical specification of such a cycle, ψ_t, is as follows:

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos \lambda_c & \sin \lambda_c \\ -\sin \lambda_c & \cos \lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}, \quad t = 1, \ldots, T, \tag{8}$$

where λ_c is the frequency, in radians, in the range $0 \leq \lambda_c \leq \pi$, κ_t and κ_t^* are two mutually uncorrelated white noise disturbances with zero means and common variance σ_κ^2, and ρ is a *damping factor*, such that $0 < \rho \leq 1$. Note that the *period* is $2\pi/\lambda_c$. For some purposes it is useful to take the variance of ψ_t, rather than the variance of κ_t, as a hyperparameter. Then since $\sigma_\kappa^2 = (1-\rho^2)\sigma_\psi^2$, a deterministic, but stationary, cycle is obtained when $\rho = 1$. The **autocorrelation function** (ACF) of ψ_t is

$$\rho(\tau) = \rho^\tau \cos \lambda \tau, \quad \tau = 0, 1, 2, \ldots. \tag{9}$$

This is a cycle which damps down to zero as τ goes to infinity, except when $\rho = 1$. The spectrum has a peak around λ_c, denoting irregular, or pseudocyclical, behavior (*see* **Spectral Analysis**). The peak becomes sharper as ρ approaches one, and in the limiting case when ρ equals one it manifests itself as a jump in the spectral distribution function. A test that the cycle is deterministic is given in Harvey and Streibel [14].

Cyclical components of this type have proved useful in economics for modeling the business cycle; an extended class of cycles is presented in [15]. Cycles can be combined with other components, such as trend and seasonal, as well as with other cycles or perhaps autoregressive processes.

Reduced Form ARIMA Models

All the structural time series models described in the previous section are linear and hence there is a corresponding ARIMA model which gives identical predictions. Since the ARIMA model contains only a single disturbance, it is called the "reduced form".

The specification of the reduced form can be found very easily for the simpler models. Thus for the local

level, (2), taking first differences yields

$$\Delta y_t = \eta_t + \varepsilon_t - \varepsilon_{t-1}, \quad t = 2, \ldots, T. \tag{10}$$

The ACF has the same form as that of a first-order moving average process, and so the local level has an ARIMA (0, 1, 1) reduced form,

$$\Delta y_t = \xi_t + \theta \xi_{t-1}, \quad t = 2, \ldots, T. \tag{11}$$

Equating the first-order autocorrelations in the two models gives the following relationship between the structural and reduced form parameters:

$$\theta = \frac{[(q^2 + 4q)^{1/2} - 2 - q]}{2}. \tag{12}$$

Note that θ is constrained to be negative. In more complicated models, the constraints tend to be stronger.

The reduced form of the local linear trend model is ARIMA (0, 2, 2), while that of the BSM, given in (6), is shown by Maravall [22] to be quite close to that of the "airline" model of Box–Jenkins analysis, namely the seasonal ARIMA of order $(0, 1, 1) \times (0, 1, 1)_s$.

Explanatory Variables and Interventions

Observable explanatory variables may be added to the right-hand side of a structural time series model. Some or all of these variables may be under the control of the researcher, for example they may represent the levels of drugs administered.

To keep the discussion simple, we will assume that the only component, apart from the irregular, is a random walk as in (2), and that there is a single explanatory variable, x_t. Then

$$y_t = \mu_t + \delta x_t + \varepsilon_t, \quad t = 1, \ldots, T. \tag{13}$$

The rationale for such a model is that the explanatory variable does not capture all the underlying movements in the level of the series. If it did, q would be zero, μ_t would be constant, and we would be left with a classical regression model. At the other extreme, if σ_ε^2 were zero, the model could be treated as a regression in first differences. More generally,

$$\Delta y_t = \delta \Delta x_t + (\eta_t + \varepsilon_t - \varepsilon_{t-1}), \quad t = 1, \ldots, T, \tag{14}$$

and it should be clear from the discussion surrounding (11) that the disturbance in curly brackets is equivalent to a first-order moving average process. Thus the model in (13) is equivalent to a transfer function model in which the stochastic part has an ARIMA (0, 1, 1) specification. An example of such a specification can be found in [3, pp. 409–412] and in [16], where it is used to model the effect of pollution on respiratory diseases in children. The interpretation of the structural formulation is more direct and the route by which it is obtained is simpler.

The model can be generalized so as to allow for a lag in the response of y to a change in x. There is an enormous literature, particularly in econometrics, on how to put restrictions on the lag structure so as to obtain a more stable lag structure.

Intervention variables are dummy variables which are used to take account of outlying observations and structural breaks. These data irregularities are usually thought of as arising from a specific event, for example, a specific hot day in the case of outliers or a change in policy or treatment in the case of a structural break. Consider the local level model of (2) and suppose that there is a shift in the level due to an intervention at a known time $t = \tau$. The model is then

$$y_t = \mu_t + \lambda w_t + \varepsilon_t, \qquad \varepsilon_t \sim \text{NID}\ (0, \sigma_\varepsilon^2), \quad t = 1, \ldots, T, \tag{15}$$

where w_t takes the value zero for $t < \tau$ and is one thereafter. Other components and explanatory variables can be added to the right-hand side. An example of the use of intervention analysis in structural time series models can be found in [10] on the effect of the 1983 seat belt law in Great Britain.

Multivariate Models

In a multivariate time series model there are observations on a number of individuals or groups over a period of time, and the model tries to capture the correlations and interrelationships between these series (*see* **Multiple Time Series**).

As before, each series may depend on various time series components, both nonstationary and stationary, and observable explanatory variables, including intervention effects. For example, with N series,

$$y_{it} = \mu_{it} + \psi_{it} + x_{it}'\delta_i + \lambda_i w_{it} + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{16}$$

This may represent a situation in which an experimental group is subject to some treatment, which varies over time and is measured by x_{it}, and a control group which is not subject to the treatment or has it at a different level. An intervention variable would be used for a treatment which is either on or off. The role of the time series components is to pick up changes which occur over time independently of the treatments. It may sometimes be reasonable to suppose that these effects are common to both groups.

The above framework can also be used for a situation where there are observations on a number of individuals over time. These individuals may be given different treatments or they may be subdivided into experimental and control groups. In economics, such cross sections of time series are known as "panel data", though in medicine, the term "**longitudinal data**" is more common [17].

The first subsection below generalizes the univariate structural time series models to multivariate series and explores their properties. It is then shown how such models can be used in connection with intervention analysis and control groups. The last subsection looks at longitudinal data.

Seemingly Unrelated Time Series Equations

In the seemingly unrelated time series equations (SUTSE) model, each series in an $N \times 1$ vector of observations, $\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'$, is modeled as in a univariate structural time series model, but the disturbances in each of the components may be correlated across series. There are no dynamic interactions between the series in a SUTSE model. Such features can be introduced, for example by including a vector of components which follows a stationary vector autoregression (VAR). Explanatory variables can also be included in the models.

To understand the implications of SUTSE models we focus on the special case of a multivariate local level model,

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \text{NID}\ (\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon), \\ &\qquad\qquad t = 1, \ldots, T, \\ \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \text{NID}\ (\mathbf{0}, \boldsymbol{\Sigma}_\eta), \end{aligned} \tag{17}$$

where $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\eta$ are the $(N \times N)$ covariance matrices, and $\boldsymbol{\eta}_t$ and $\boldsymbol{\varepsilon}_t$ are multivariate normal disturbances which are mutually uncorrelated in all time periods. The long-run connections between the series are captured by the covariances in the off-diagonal elements in $\boldsymbol{\Sigma}_\eta$, while the short-run correlations are in $\boldsymbol{\Sigma}_\varepsilon$. These long-run and short-run correlations may be completely different. In the special case when they are the same, the system is said to be homogeneous. Because $\boldsymbol{\Sigma}_\eta = q\boldsymbol{\Sigma}_\varepsilon$, where q is a positive scalar, the statistical treatment is simplified considerably [8, Chapter 8].

Common Factors

The form of a common factor model is similar to that of a SUTSE model except that some or all of the components are driven by disturbance vectors with covariance matrices which are less than full rank. Such models may be formulated in terms of common factors.

Consider the bivariate local level model,

$$\begin{aligned} y_{1t} &= \mu_t + \varepsilon_{1t}, \quad t = 1, \ldots, T, \\ y_{2t} &= \theta\mu_t + \overline{\mu} + \varepsilon_{2t}, \end{aligned} \tag{18}$$

where μ_t is a univariate random walk. The long-run components in the two series depend on the same underlying source, with the level in the second series being a linear function of the level in the first. If $\theta = 1$, the underlying levels remain a constant distance apart, while if $\overline{\mu} = 0$ they are identical. An important feature of this model is that the common nonstationary level can be removed by a certain linear combination of the two series. This property is known as cointegration in the econometrics literature [5]. In the present case it leads to the following relationship between y_{1t} and y_{2t}:

$$y_{2t} = \alpha y_{1t} + \overline{\mu} + \varepsilon_t, \quad t = 1, \ldots, T, \tag{19}$$

where $\alpha = \theta$ and $\varepsilon_t = \varepsilon_{2t} - \alpha\varepsilon_{1t}$.

When there are more than two series, there may be more than one common factor for each component, and these factors are not unique. A factor rotation may sometimes give components with an interesting interpretation.

Control Groups and Intervention Analysis

Suppose we wish to assess the effect of an intervention on a series. This can be done by a univariate model as described in the section on explanatory

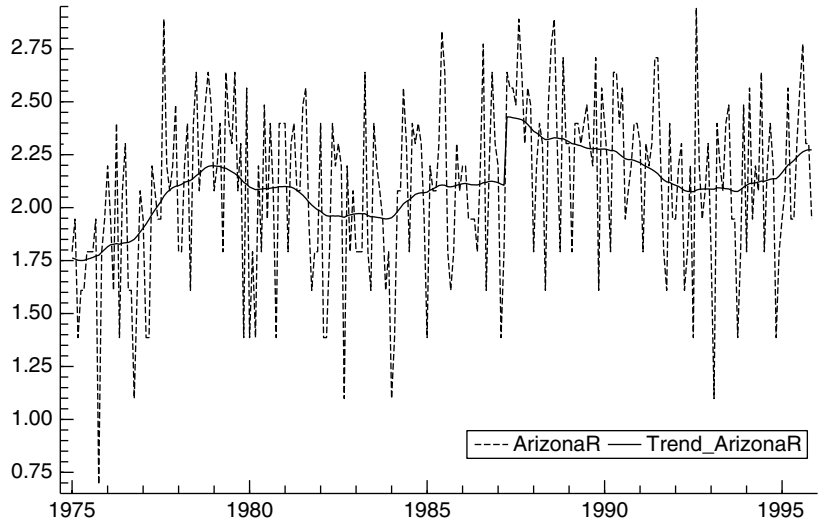

Figure 1 Monthly crashes on rural interstate highways in Arizona.

variables. However, if observations are available on another series correlated with the series of interest but unaffected by the intervention, it is possible to construct a bivariate model in which this second series acts as a control group.

Consider the following model:

$$y_{1t} = \mu_{1t} + \varepsilon_{1t}, \quad t = 1, \ldots, T,$$
$$y_{2t} = \mu_{2t} + \lambda w_t + \varepsilon_{2t}, \tag{20}$$

where the first series contains the observations on the control group. The higher the correlation between the two trends, the bigger the gain is likely to be. The most extreme case is when the two series are driven by a common level, so that they are cointegrated. The model is as in (18) except that the second equation becomes (21):

$$y_{2t} = \theta\mu_t + \overline{\mu} + \lambda w_t + \varepsilon_{2t}, \quad t = 1, \ldots, T, \tag{21}$$

and so (19) becomes

$$y_{2t} = \alpha y_{1t} + \overline{\mu} + \lambda w_t + \varepsilon_t, \quad t = 1, \ldots, T. \tag{22}$$

In contrast to (15), there is no stochastic level in (22). As a result the intervention effect can be estimated consistently and large gains can be expected when there are only a small number of observations after the intervention. For example, Harvey [9] shows that the standard error of the estimate of the parameter measuring the effect of the British seat belt law is more than halved when a control group is used. Another example, taken from [2, p. 22–4], concerns the relaxation of the speed limit on certain rural highways in the US in 1987. Figure 1 shows the random walk trend extracted from a model fitted to monthly data on crashes in rural Arizona, with the intervention in April of that year. A seasonal component was also included. The t-statistic on the intervention was 2.21, but a joint model with the series on crashes on urban highways, where the speed limit was not changed, increased the t-statistic to 2.41. The gain arose from the correlation of 0.81 in the level disturbances.

Longitudinal Data

When there are observations on a large number of individuals over time, the unrestricted SUTSE model becomes unmanageable. However, in such situations it is quite reasonable to suppose that, for each component, the correlation between the disturbances in any two individuals is the same. This restriction may be introduced by assuming that there is an effect common to all individuals and that the individual specific effects are mutually independent. Thus, in the local level model, with observable explanatory variables

$$y_{it} = \mu_t + \mu_{it}^* + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_t + \varepsilon_{it}^*,$$
$$i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{23}$$

where μ_t is the common level driven by a disturbance with variance σ_η^2, μ_{it}^*, $i = 1, \ldots, N$, are the individual specific levels each driven by a disturbance with variance $\sigma_{\eta*}^2$, and ε_t and ε_{it}^* are the common and specific irregular disturbances with variances σ_ε^2 and $\sigma_{\varepsilon*}^2$, respectively; see [12], and [23].

Data Irregularities

It is not unusual for time series to suffer from data irregularities such as missing and irregularly spaced observations. A discrete time model can handle missing observations by using the state space form and letting the Kalman filter skip over the missing observations. Irregularly spaced observations pose a more fundamental problem and the solution is to set up a model in continuous time and then use the state space form to construct the relevant discrete time model; see [13] and [17]. For a local level this is very easy. Let δ_t be the time interval between the observation at time t and the observation at time $t-1$. The level is assumed to evolve as a Wiener process, or **Brownian motion**, and if the observations are equally spaced, that is $\delta_t = 1$, the corresponding discrete time model is just (2). With unequally spaced observations the only difference is that $\text{var}(\eta_t) = \delta_t \sigma_\eta^2$, where σ_η^2 is a parameter defined with respect to the continuous time white noise process driving the Wiener process. Handling this model within the state space framework presents no problem since, although the system is not time invariant, the way in which it varies with time depends on the δ_t's and so is known.

Outliers may also arise in time series. These are observations which appear to be inconsistent with a model which is appropriate for most of the data. The simplest kind of outlier is one which is identified as an incorrect measurement, in which case it may be handled by an intervention variable or treated as a missing observation; the second solution is preferable if there are a large number of such cases. A less extreme case arises if some observations are known to be more reliable than others. For example, suppose that the tth observation is constructed as the average of the values of n_t units. In this case it may be appropriate to let the variance of the measurement error, ε_t, be proportional to $1/n_t$. As with irregularly spaced observations, there is no problem in handling such a situation with the state space form. An example involving survey data is in Pfeffermann [25]. Finally, outliers might be handled by allowing the disturbance in the measurement equation to have a heavy-tailed distribution. This makes the model more robust. References to the techniques used for dealing with heavy-tailed distributions are given in the next section, but a specific example can be found in Dubin and Koopman K [4, p. 233–5].

Non-Gaussian Observations

Suppose it is felt that a **Poisson distribution** is suitable for a set of count data. In the **generalized linear model** framework, the mean is assumed to depend on a linear combination of explanatory variables which determine the mean of the distribution via an exponential link function. A deterministic time series model can be constructed by letting the explanatory variables be functions of time. However, for all the reasons given earlier this may not be satisfactory. The functions of time may therefore be given stochastic coefficients as in the models described earlier. However, because of the nonnormality of the observations, the Kalman filter is no longer appropriate. There is no simple analytic solution and a number of ways of designing suitable filtering and estimation procedures, based on simulation techniques such as importance sampling, have been suggested; see, for example, [4], and [26]. An alternative approach is to design transition equations such that the density of the mean, conditional on the information in the previous time period, is conjugate to the observation density so that there exists an exact analytic solution. This turns out to be possible for a number of observation distributions in the simplest situation when the level of the series is assumed to change, that is the analog of the Gaussian local level model of (2); see [7, 10], and [27].

The above techniques may also be applied to other distributions used for count data and to the **binomial distribution** as used for qualitative data.

References

[1] Andrews, R.C. (1994). Forecasting performance of structural time series models, *Journal of Business and Economic Statistics* **12**, 129–133.

[2] Balkin, S. & Ord, J.K. (2001). Assessing the impact of speed-limit increases on fatal interstate crashes (with discussion), *Journal of Transportation and Statistics*, **4**, 1–26.

[3] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

[4] Durbin, J. & Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

[5] Engle, R.F. & Granger, C.W.J. (1987). Co-integration and error correction: representation, estimation and testing, *Econometrica* **55**, 251–276.

[6] Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models, *Journal of Time Series Analysis* **15**, 183–202.

[7] Grunwald, G.K., Guttorp, P. & Raftery, A.E. (1993). Prediction rules for exponentially family state space models, *Journal of the Royal Statistical Society, Series B* **149**, 187–227.

[8] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

[9] Harvey, A.C. (1996). Intervention analysis with control groups, *International Statistical Review* **64**, 313–328.

[10] Harvey, A.C. & Fernandes, C. (1989). Time series models for count data or qualitative observations, *Journal of Business and Economic Statistics* **7**, 407–417.

[11] Harvey, A.C. & Koopman, S.J. (1992). Diagnostic checking of unobserved components time series models, *Journal of Business and Economic Statistics* **10**, 377–389.

[12] Harvey, A.C. & Koopman, S.J. (1996). Structural time series models in medicine, *Statistical Methods in Medical Research* **5**, 23–49.

[13] Harvey, A.C. & Stock, J. (1994). Estimation, smoothing, interpolation and distribution for structural time-series models in continuous-time, in *Models, Methods and Applications of Econometrics*, P.C.B. Phillips, ed. Blackwell, Oxford, pp. 55–70.

[14] Harvey, A.C. & Streibel, M. (1996). Tests for deterministic versus indeterministic cycles, *Journal of Time Series Analysis* **19**, 505–529.

[15] Harvey, A.C. & Trimbur, T. (2003). General model-based filters for extracting cycles and trends in economic time series, *Review of Economics and Statistics*, to appear.

[16] Helfenstein, U., Ackermann-Liebrich, U., Braun-Fahrlander, C. & Wanner, H.U. (1991). Air-pollution and diseases of the respiratory tracts in pre-school children: a transfer function model, *Environmental Monitoring and Assessment* **17**, 147–156.

[17] Jones, R.H. (1993). *Longitudinal Data with Serial Correlation: A State-space Approach*. Chapman & Hall, London.

[18] Kitagawa, G. & Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*. Springer-Verlag, Berlin.

[19] Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for nonstationary time series models, *Journal of the American Statistical Association* **92**, 1630–1638.

[20] Koopman, S.J. & Harvey, A.C. (2003). Computing observation weights for signal extraction and filtering, *Journal of Economic Dynamics and Control* **27**, 1317–1333.

[21] Koopman, S.J., Harvey, A.C., Doornik, J.A. & Shephard, N. (1995). *STAMP 5.0 Structural Time Series Analyser, Modeller and Predictor*. Chapman & Hall, London.

[22] Maravall, A. (1985). On structural time series models and the characterization of components, *Journal of Business and Economic Statistics* **3**, 350–355.

[23] Marshall, P. (1992). Estimating time-dependent means in dynamic models for cross-sections of time series, *Empirical Economics* **17**, 25–33.

[24] Muth, J.F. (1960). Optimal properties of exponentially weighted forecasts, *Journal of the American Statistical Association* **55**, 299–305.

[25] Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys, *Journal of Business and Economic Statistics* **9**, 163–75.

[26] Shephard, N. (1994). Partial non-Gaussian state space, *Biometrika* **81**, 115–131.

[27] Smith, J.Q. (1979). A generalization of the Bayesian steady forecasting model, *Journal of the Royal Statistical Society, Series B* **41**, 375–387.

[28] West, M. & Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.

(*See also* **Time Series Regression**)

A. HARVEY & S.J. KOOPMAN

Studentization

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and standard deviation σ, then $Z = \sqrt{n}(\overline{X} - \mu)/\sigma$ follows the standard normal distribution. Here, $\overline{X}$ is the mean of the sample. If, in Z, σ is replaced by the sample standard deviation $S = [\sum(X - \overline{X})^2/(n - 1)]^{1/2}$, then the random variable $T = \sqrt{n}(\overline{X} - \mu)/S$ follows a **Student's *t* distribution** with $n - 1$ df. The process of obtaining T is called *Studentization*, initiated by William Sealy **Gosset** (1876–1937, with the pseudonym "Student"). Since then, studentization has been extended to **studentized range**, studentized extreme deviate (*see* **Extreme Values**), studentized **residuals**, and others.

AUSTIN F.S. LEE

Studentized Range

Let $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$, $j = 1, \ldots, n, i = 1, \ldots, k$, be independent observations in a (balanced) one-way layout with k treatments. Let $\overline{Y}_1, \ldots, \overline{Y}_k$ be the sample averages and let S^2 be the usual independent and unbiased estimator of σ^2 based on $\nu = k(n-1)$ df. Let W be the **range** of the $\overline{Y}_i$ and consider the *studentized range* Q,

$$Q = \frac{W}{S/\sqrt{n}}.$$

According to David [2]: "The beginning of interest in the studentized range is attributed by **Egon Pearson** [15] to a letter he received from 'Student' (**W.S. Gosset**) in 1932. Referring to the comparison of *selected* differences in variety means. . .". The test statistic for $\mathrm{H}_{ij} : \mu_i = \mu_j$ against a two-sided alternative is

$$|T_{ij}| = \frac{|\overline{Y}_i - \overline{Y}_j|}{S(2/n)^{1/2}},$$

but in view of a potential *selection effect*, Pearson and Gosset considered contrasting each $|T_{ij}|$ with upper α quantiles of

$$\max_{1 \le i < j \le k} |T_{ij}| = \sqrt{2}Q.$$

That constituted the basis for two different **multiple comparison** procedures (MCPs).

The Newman–Keuls (NK) Method

Newman [13] provided approximate tables of upper percentage points of Q using Pearson's [14] method for approximating the distribution of the range of a normal sample. Keuls [10] proposed a stepwise coherent MCP for testing all *subset homogeneity hypotheses* based on Newman's tables, and the method became known as "the NK method". The NK method rejects a given subset-homogeneity hypothesis (of means in a one-way layout, say) if, and only if, all tests of subset-homogeneity hypotheses which imply the given one end in rejection. Accordingly the NK method is a *stepdown* procedure starting with the overall homogeneity hypothesis, continuing testing all its subsets of size $k-1$ if, and only if, rejected, etc. The test of any given subset homogeneity hypothesis is an α-level studentized range test with means pertaining to the given subset, and a common pooled S^2 from all observation in the experiment. Hochberg & Tamhane [9] discuss various properties of the original NK procedure and some of its modifications that were proposed in order to turn it into a *familywise error-rate* (FWE) controlling MCP as well as other stepwise MCPs (e.g. Duncan's [4] method) which followed the NK method.

Tukey's *T*-Method

This is a **simultaneous confidence** estimation and testing MCP for all pairwise comparisons in a balanced one-way layout. With probability $1 - \alpha$ simultaneously for all $i \neq j$,

$$\mu_i - \mu_j \in \left[\overline{Y}_i - \overline{Y}_j \pm S Q_{k,\nu}^{(\alpha)}/\sqrt{n}\right],$$

where $Q_{k,\nu}^{(\alpha)}$ is the $1 - \alpha$ percentile of Q. Tukey [20] extended his method to all contrasts and to all linear combinations of the μ_i. The latter is based on percentiles of the (Studentized) augmented range defined as the maximum between the studentized maximum modulus $\sqrt{2}\max_{1 \le i \le k}\{|\overline{Y}_i - \mu_i|/(S/\sqrt{n})\}$ and Q. Tukey [20] also proposed other extensions of his original method. One of his suggestions involved the use of the critical points of the studentized range also for (all pairwise) comparisons in unbalanced designs. His suggestion was essentially repeated by Kramer [11], and has become known as the "Tukey–Kramer (TK) procedure". Tukey's original conjecture on the conservative nature of such a procedure in unbalanced **analysis of variance** (ANOVA) was proved by Hayter [7]. These and other extensions and comparisons between Tukey's method (as well as other single-step MCPs which stemmed from it) and the NK method (as well as other stepwise MCPs) are discussed in Hochberg & Tamhane [9].

The Internally Studentized Range

David [1] indicated that the term *studentized range* was also used in a different sense. "Let $X_1, \ldots, X_n$ be a random sample from a normal $\mathrm{N}(\mu, \sigma^2)$ population and let $X_{(1)} \le \cdots \le X_{(n)}$ be the corresponding

order statistics. Then $W = X_{(n)} - X_{(1)}$ is the sample range... Let $S^2 = \sum(X_i - \overline{X})^2/(n-1)$... we call... $Q = W/S$ the internally studentized range" (ISR). He refers to the first similar ratio discussed above as "the *externally* studentized range" (ESR). In the case of the ESR there is independence between W and S. In contrast, in the case of the ISR, W and S are associated. David [2] discusses the history of the ISR, including applications, approximations, and theoretical derivations of its distribution. He indicates that Snedecor [17] discussed several applications of the ISR "even before any appropriate theory was developed". Snedecor [17] considered applications of the ISR under the heading of "short cuts". He wrote: "... we introduce a topic of great utility in the common sense understanding of statistics...". He references earlier work by Tippett [19], who apparently "has given a mathematical statement of the problem, providing an extensive table..." but indicates that, "In Tippett's table the sample range is divided by the population standard deviation." Snedecor [17] studies the relation between W and S empirically in normal samples of different sizes, provides a table for the expected ratio Q, and discusses various applications, some of which are indicated in David [2].

David [2] indicates that the ISR was proposed as a test of normality [by David, H.A., Hartley, H.O. & Pearson, E.S., *Biometrika* **41** (1954) 482–493] and was found ("In empirical sampling studies") to possess "particularly good properties against symmetric, especially short-tailed (e.g. uniform) distributions but seems to have virtually no power with respect to asymmetry" (*see* **Normality, Tests of**). David [2] also discusses some distribution-free bounds on Q due to Thompson [18]. Interesting results on the power of the ESR as a test of the global null hypothesis in one-way layouts are discussed in David et al. [3].

Exact Distributions and Tables

According to David [2], "... definitive tables" for the cumulative distribution function (CDF) of the ESR are given by Harter [5] based on Hartley [6]. He also indicated that for larger values of ν one can use Pearson & Hartley's [16] percentage points of Q and that "A computing algorithm for upper tail CDF and percentage points of Q_ν is given by Lund & Lund [12]."

David [1] discusses how the distribution of the ISR was facilitated by the independence of Q and S in normal samples and indicates that this result can also be established based on Fisher, R.A., *Proceedings of the Royal Society of London, Series A* **130** (1931) 16–28 and Geary R.C. (1933), *Biometrika* **25**, 184–186. He also describes various numerical methods for approximating its percentage points.

Related Topics

Hayter [8] discusses a one-sided studentized range statistic with potential use in problems with **ordered alternatives** (*see* **Isotonic Inference**). David [2] discusses bivariate studentized range statistics and indicates potential applications.

References

[1] David, H.A. (1970, 1981). *Order Statistics*. Wiley, New York.

[2] David, H.A. (1988). Studentized range, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 39–42.

[3] David, H.A., Lachenbruch, P.A. & Brandis, H.P. (1972). The power function of range and studentized range tests in normal samples, *Biometrika* **59**, 161–168.

[4] Duncan, D.B. (1955). Multiple range and multiple F-tests, *Biometrics* **11**, 1–42.

[5] Harter, H.L. (1970). *Order Statistics and Their Uses in Testing and Estimation*, Vols. 1, 2. US Government Printing Office, Washington.

[6] Hartley, H.O. (1942). The range in random samples, *Biometrika* **32**, 334–348.

[7] Hayter, A.J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative, *Annals of Statistics* **12**, 61–75.

[8] Hayter, A.J. (1990). A one-sided studentized range test for testing against a single ordered alternative, *Journal of the American Statistical Association* **85**, 778–785.

[9] Hochberg, Y. & Tamhane, A. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[10] Keuls, M. (1952). The use of the "Studentized range" in connection with an analysis of variance, *Euphytica* **1**, 112–122.

[11] Kramer, C.Y. (1956). Extension of multiple range test to group means with unequal numbers of replications, *Biometrics* **12**, 307–310.

[12] Lund, R.E. & Lund, J.R. (1983). Probabilities and upper quantiles for the studentized range, *Applied Statistics* **32**, 204–210.

[13] Newman, D. (1939). The distribution of the range in samples from a normal population, expressed in

terms of an independent estimate of standard deviation, *Biometrika* **31**, 20–30.

[14] Pearson, E.S. (1932). The percentage limits for the distribution of range in samples from a normal population ($n \leq 100$), *Biometrika* **24**, 404–417.

[15] Pearson, E.S. (1938). "Student" as statistician, *Biometrika* **30**, 210–250.

[16] Pearson, E.S. & Hartley, H.O. (1954, 1970). *Biometrika Tables for Statisticians*, Vol. 1. Cambridge University Press, London.

[17] Snedecor, G.W. (1937). *Statistical Methods*. Collegiate Press, Ames.

[18] Thomson, G.W. (1955). Bounds for the ratio of range to standard deviation, *Biometrika* **42**, 268–269.

[19] Tippett, L.H.C. (1925). On the extreme individuals and the range of samples taken from a normal population, *Biometrika* **17**, 364–387.

[20] Tukey, J.W. (1953). The problem of multiple comparisons. Unpublished memorandum.

(*See also* **Simultaneous Inference**)

YOSEF HOCHBERG

Student's *t* Distribution

The probability density function (pdf) of the Student's t distribution is given by

$$f(x) = \frac{\Gamma[(\nu+1)/2]}{(\pi\nu)^{1/2}\Gamma(\nu/2)}(1+x^2/\nu)^{-(\nu+1)/2},$$

where $-\infty < x < \infty$ and $\nu > 0$. The parameter ν is called the **degrees of freedom** (df) of the distribution. Suppose that Z is a standard normal random variable (*see* **Standard Normal Deviate**), Y is a **chi-square** random variable with ν df, and Z and Y are statistically independent (*see* **Random Variable**). Then the random variable T obtained by

$$T = \frac{Z}{(Y/\nu)^{1/2}}$$

has a t distribution with ν df.

The t distribution is symmetric around 0. If X is a random variable that follows the t distribution, then $\mathrm{E}(X) = 0$, $\mathrm{var}(X) = \nu/(\nu-2)$, **skewness** $= 0$, and **kurtosis** $= 6/(\nu-4)$. Furthermore, as ν increases, the t distribution approximates the standard normal distribution. The percentage points of t distributions have been tabulated and can be found in many statistical texts. The t distribution plays an important role in **confidence intervals** and **hypotheses tests**.

(*See also* **Student's *t* Statistics**)

AUSTIN F.S. LEE

Student's *t* Statistics

A Student's t statistic may arise from many **hypothesis testing** problems. In general the Student's t statistic utilizes a **Student's *t* distribution** to determine the critical value (*see* **Critical Region**) for rejecting or not rejecting the **null hypothesis**. The most common Student's t statistics are given below.

1. *One-sample t statistic*. Suppose that $X_1, X_2, \ldots, X_n$ is a **random sample** from a **normal** population with mean μ and variance σ^2. To test the null hypothesis that $\mu = \mu_0$, one constructs the t statistic by $T = \sqrt{n}(\overline{X} - \mu_0)/s$, where $\overline{X}$ is the mean of the sample and $s^2 = \Sigma(X_i - \overline{X})^2/(n-1)$, the sample variance. For a given level of significance (*see* **Level of a Test**), the one-sample t-test compares T with the critical value found from the Student's t distribution with $n-1$ **degrees of freedom** (df).
2. *Two-sample t statistic*. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ_x and variance σ_x^2, and $Y_1, Y_2, \ldots, Y_m$ is another random sample from a normal population with mean μ_y and variance σ_y^2. Assume also that X_S and Y_S are independent samples and that $\sigma_x^2 = \sigma_y^2 = \sigma^2$. In testing the null hypothesis that $\mu_x = \mu_y$, the t statistic is calculated from

$$T = \frac{\overline{X} - \overline{Y}}{s_p \left(\frac{1}{n} + \frac{1}{m}\right)^{1/2}},$$

   where $\overline{X}$ and $\overline{Y}$ are the sample means of X_S and Y_S, respectively, and

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}.$$

   Here, s_p^2 is a weighted average of the sample variances s_x^2 and s_y^2 of X_S and Y_S, respectively, and called the pooled estimate for σ^2. For a given level of significance, the two-sample t test compares T with the critical value found from the Student's t distribution with $n+m-2$ df. However, if the assumption of homogeneity in variance ($\sigma_x^2 = \sigma_y^2$) is violated, the actual probability of type I error (*see* **Hypothesis Testing**) may deviate largely from the given level of significance. In the statistical literature, this is referred to as the **Behrens–Fisher problem**. In this case, another test (for example Welch's approximate τ; *see* **Aspin–Welch Test**) is recommended.
3. *The t statistic for testing significance of the regression coefficient*. In **linear regression** analysis, the t statistic $T = b/\mathrm{se}(b)$ is computed for testing significance (i.e. being statistically different from zero) of the regression coefficient β, using its **least squares** estimate b.
4. *The t statistic for testing significance of the* **correlation** *coefficient*. In correlation analysis, one computes the t statistic $T = (n-2)^{1/2} r/(1-r^2)^{1/2}$ and compares it with the critical value from the Student's t distribution with $n-2$ df. Here, r is Pearson's product–moment correlation coefficient.

Different types of Student's t statistic arise from other hypothesis testing problems.

AUSTIN F.S. LEE

Study Population

The term *study population* is often used to refer to the population from which observations are drawn; that is, the sampled population (*see* **Target Population**). In other writings, it has been used to refer to the study sample [1].

Reference

[1] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1984). *Epidemiologic Research: Principles and Quantitative Methods*. Van Nostrand, New York.

(*See also* **Validity and Generalizability in Epidemiologic Studies**)

SANDER GREENLAND

Subjective Probability

Subjective probability provides a language for organizing and expressing uncertainty, to articulate expert knowledge, and/or to serve as a meaningful framework for **Bayesian** statistical **inference**. Subjective probabilities refer directly to the strength of a person's beliefs regarding the propositions in question. The currently dominant frequentist view of probability, by contrast, claims that probabilities exist in the real world as physical properties, potential long-run frequencies. Under the frequentist notion, probabilities lack meaning as uncertainties, and their numerical values are typically unknown (although they can be subject to inference from statistical data).

Bruno de Finetti [1, 2] gave subjective probabilities an operational definition as personal prices for lottery tickets, in which $\$P(A)$ is interpreted as the person's price for a ticket that pays \$1 if A. To avoid being a potential sure loser, a person must choose his prices to have coherent values, i.e. values for which no set of hypothetical transactions exists at these prices that would combine to guarantee the person will suffer a net loss for every possible outcome. Such coherence is equivalent to numerical agreement with at least one probability measure, satisfying a finitely additive version of the usual probability axioms.

De Finetti's Fundamental Theorem of Probability enables a person to assert probability values sequentially for an open-ended sequence of propositions. Coherence is preserved at each step by restriction to an interval of available values whose endpoints are obtained by **linear programming**. Expert opinion can be articulated and quantified in probability form by this method (or by further restricting the choices to satisfy some parameterized subjective-probability model).

An equivalent alternative theory of subjective probabilities satisfying the mathematical axioms (together with subjective utilities) proceeds by axiomatizing, first, the person's preferences for lottery tickets and corresponding rewards. This approach by **L.J. Savage** [5] was inspired by ideas in Ramsey [4], and von Neumann & Morgenstern [3]. To emphasize the dependence of degree of belief on the believer, Savage used the term, "personal probability".

Conditional subjective probability was defined by de Finetti as the person's price, now, in a transaction that will be voided (with reimbursement) if the conditioning proposition turns out not to be true: $\$P(A|B)$ is the person's price for a ticket that pays 1 if A and reimburses $\$P(A|B)$ unless B. With transactions of this new type included, the requirement of coherence can be shown to be equivalent to the probability axioms *plus* the usual relationship between conditional and unconditional probability: $P(A|B) = P(A \text{ and } B)/P(B)$, if $P(B) > 0$. Thus, **conditional probability** has a natural interpretation as what the person thinks, now, that his/her opinion should be if and when B is learned. Unfortunately, there is no completely satisfactory coherence-type justification in the literature for the requirement that *after* B becomes known, the person's new opinion conforms to such previous conditional probabilities. There is no normative theory of temporal coherence.

Bayesian statistics, however, is based on the notion that inference should proceed by probability conditioning. Bayesian posterior probabilities, "posterior" to statistical data B, are conditional probabilities $P(A|B)$, which can be computed from the **"prior" probability** $P(A)$ and the statistical sampling probabilities of the data B, $P(B|A)$, $P(B|\text{Not}A)$, by use of **Bayes' theorem:** $P(A|B) = P(B|A)P(A)/[P(B|A)P(A) + P(B|\text{Not}A)P(\text{Not}A)]$. So, subjective probabilities can be of practical use as Bayesian prior or posterior probabilities, with the prior probabilities elicited from a subject-matter expert and the corresponding posterior probabilities obtained by Bayes' theorem.

Frequentist sampling probability can, itself, be viewed as a special limiting case of conditional subjective probability, namely probability expressing opinion of further future data conditional on a long sequence of observed data. According to de Finetti's Representation Theorem for **Exchangeability**, an infinite exchangeable random sequence is representable as a probability mixture of independent, identically distributed (iid) random sequences. The iid random sequences are mathematically the same as a statistical sampling model with unknown physical probabilities, and the mixing probabilities are interpretable as a corresponding prior distribution of the unknown probabilities. Hence, a person who has exchangeable uncertainties concerning an infinite sequence of observable experimental outcomes behaves as if he/she believed there to be a real-world

iid statistical model with unknown physical probabilities. Thus, frequentist probability can be subsumed within the theory of coherent subjective probability. (For a mathematically elementary treatment of the representation, see [6].)

References

[1] de Finetti, B. (1937). La prevision, ses lois logiques, ses sources subjectives, *Annales de L'Institute Henri Poincare* **7**, 1–68. H. Kyburg, translator, Foresight, its logical laws, its subjective sources, in *Studies in Subjective Probability*, H. Kyburg & H. Smokler, eds. Wiley, New York; 2nd Ed., Krieger, New York, 1980.

[2] de Finetti, B. (1974, 1975). *Theory of Probability*, 2 volumes, A.F.M. Smith & A. Machi, translators. Wiley, New York.

[3] Heath, D.L. & Sudderth, W.D. (1975). de Finetti's theorem on exchangeable variables, *American Statistician* **30**, 188–190.

[4] Ramsey, F. (1926). Truth and probability, in *Studies in Subjective Probability*, H. Kyburg & H. Smokler, eds. Wiley, New York; 2nd Ed., Krieger, New York, 1980.

[5] Savage, L.J. (1954). *The Foundations of Statistics*. Wiley, New York; 2nd Ed., Dover, New York, 1972.

[6] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economics Behavior*, 3rd Ed. Princeton University Press, Princeton, 1953.

(*See also* **Foundations of Probability**)

JAMES M. DICKEY

Sufficiency

Sir R.A. Fisher introduced the concept of sufficiency in [3] and claimed in [4] that the "sufficient statistic" is equivalent, for all subsequent purposes of **estimation**, to the original data. More generally, the *sufficiency principle* states that any **inference** should be based on the **sufficient statistic**. This principle is widely accepted by statisticians.

Two related principles are the **likelihood** principle (*see* **Foundations of Probability**) and the **conditionality principle**. The likelihood function $L_X(\theta)$ is defined as $L_X(\theta) = P_\theta(X)$ [where $P_\theta(X)$ is the density function of the data X] and is regarded as a function of the parameter θ for final X. The *likelihood principle* states that a statistical procedure should depend only on $L_X(\theta)$. It further then states that if two different experiments with the same parameter result in proportional likelihoods for the data observed, then identical conclusions should be drawn in the two experiments.

In particular, the likelihood principle implies that conclusions should be based only on the data observed, and not on any data which might have been observed. Hence such considerations as **unbiasedness**, size (*see* **Level of a Test**), power, risk (*see* **Decision Theory**), and so on, which involve averaging over the sample space, may violate the likelihood principle. **Maximum likelihood** estimates and **Bayesian methods** based on posterior distributions, on the other hand, do satisfy the likelihood principle.

The *conditionality principle* states that if the experiment actually performed is chosen randomly and independently of θ from a collection of possible experiments, then the statistical conclusions should not depend on any experiment not performed.

A famous result of Birnbaum [2] shows that, for discrete distributions, the sufficiency principle plus the conditionality principle are equivalent to the likelihood principle. Berger & Wolpert [1] extend this result to general distributions and give an extensive discussion on these principles and many related ones. They, and many others, argue that the likelihood principle ought to be followed, and that the Bayesian approach is the most reasonable way of guaranteeing its implementation. Others, including eventually Birnbaum himself, have argued that the likelihood principle is not universally valid.

In practice, we commonly have less than complete faith in our statistical models and perform residual and other analyses intended to assess model reasonableness. Even if our model appears consistent with the data, it is almost certainly the case that "nearby" models exist which fit the data essentially as well and which may well have a different set of minimal sufficient statistics. The development of **robust** methods indicates the utility of extending a given model (for example, a normal location model) to a broader model (for example, ε-contamination models), where the relatively simple sufficient statistics of the original model (such as the sample mean) are no longer sufficient for the expanded model. In this case sufficiency may provide no reduction at all for the expanded model, and the sufficiency principle may provide little in the way of guidance.

A very large collection of commonly used statistical procedures, however, owe at least a part of their development and desirable properties to the sufficiency principle and the optimality properties of sufficient statistics.

References

[1] Berger, J.O. & Wolpert, R.I. (1984). *The Likelihood Principle*. Lecture Notes No. 6. Institute of Mathematics and Statistics, Hayward.

[2] Birnbaum, A. (1964). On the foundations of statistical inference binary experiments, *Annals of Mathematical Statistics* **32**, 414–435.

[3] Fisher, R.A. (1922). On the mathematical foundation of theoretical statistics, *Philosophical Transactions of the Royal Society (London), Series A* **222**, 309–368.

[4] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

WILLIAM E. STRAWDERMAN

Sufficient Statistic

A sufficient statistic is a statistic (i.e. a function of the data) such that the **conditional** distribution of the data given the sufficient statistic does not depend on any unknown parameters. If I know the value of a sufficient statistic, then I can, with the use of an auxiliary randomization using this conditional distribution, reproduce "data" which has the same (unconditional) distribution as the original data. In this sense the sufficient statistic contains all the **information** in the original data. The notion of **sufficiency** was introduced by R.A. Fisher in [5], and has played an important role in the development of statistical theory and practice.

A simple but concrete example should help to fix ideas. Suppose that we wish to estimate the probability θ that a particular medication will cure migraine headaches. We select two migraine patients randomly and treat them with the medication. Let $X_i = 1$ if the ith patient is cured and 0 if not, and let $T = X_1 + X_2$. If $T = 0$, then the data (X_1, X_2) is (0,0) with probability 1. Similarly, if $T = 2$, then $(X_1, X_2) = (1, 1)$. If $T = 1$, then the data is either (1,0) or (0,1) each with probability $\frac{1}{2}$ regardless of the value of θ. Hence T is a sufficient statistic.

If T is known, then we may reconstruct a set of "data" (X_1, X_2) with a distribution equal to the original data (X_1, X_2) as follows. If $T = 0$, then $(X_1^*, X_2^*) = (0, 0)$. If $T = 1$, then toss a fair coin and set $(X_1^*, X_2^*) = (1, 0)$ if the coin comes up heads and $(X_1^*, X_2^*) = (0, 1)$ if tails. If $T = 2$, then $(X_1^*, X_2^*) = (1, 1)$. The distribution of (X_1^*, X_2^*) is thus the same as that of (X_1, X_2).

Note that if $\delta(X_1, X_2)$ is any estimator of θ (*see* **Estimation**), then the "estimator" $\delta(X_1^*, X_2^*)$ has the same distribution as $\delta(X_1, X_2)$ regardless of the value of θ and hence the same bias, variance, **mean square error**, and so on. Since $\delta(X_1^*, X_2^*)$ is based only on knowledge of T (and an auxiliary toss of a fair coin), we see that knowledge of T "is equivalent, for all subsequent purposes of estimation, to the original data from which it was derived", as Fisher claims in [6].

It should be reasonably clear from the previous example that, in general, knowledge of a sufficient statistic allows the construction of a (random) set of data X^* with a distribution equivalent to the original data X. Hence given any estimator $\delta(X)$ one may construct a (randomized) procedure $\delta(X^*)$ with the same behavior (bias, variance, etc.) as $\delta(X)$. In addition, in many problems, a sufficient statistic may provide a dramatic reduction in the complexity of the data. For example, if in the above example we treated n patients instead of 2, then $T = \sum_{i=1}^{n} X_i$ (= total number of people cured) is a sufficient statistic. T is one-dimensional and takes on the values $0, 1, \ldots, n$. By contrast, the original data $(X_1, \ldots, X_n)$ is an n-dimensional vector and may take on any of 2^n different values. Hence the sufficient statistic is simpler in structure (dimension n vs. 2^n) and in the number of values it takes on (n vs. 2^n).

In what follows I give a somewhat more formal definition of sufficiency, discuss how to find sufficient statistics, give other results indicating that reduction by sufficiency results in no loss of information, and discuss some related notions and results.

Finding Sufficient Statistics

A statistical model consists of data X which takes values in the sample space $\mathcal{X}$ and has a probability distribution $P_\theta(x)$ depending on an unknown parameter θ which takes values in the parameter space Θ. Unless explicitly stated otherwise, we assume that both $\mathcal{X}$ and Θ are contained in Euclidean vector spaces of possibly different dimension. A statistic is a (measurable) function, $T(X)$ possibly vector valued, defined on $\mathcal{X}$, which does not depend on θ. A sufficient statistic is a statistic such that the conditional distribution of X given T does not depend on θ (with probability one).

A rigorous discussion of sufficiency requires measure theory because the required conditional distributions are not necessarily uniquely defined for all values of T. For a rigorous measure-theoretic treatment, see Halmos & Savage [8], Lehmann [10], Bahadur [1], or Hurzurbazar [9]. We assume throughout that X is either discrete or continuous with density $p_\theta(x)$.

In some cases, such as in the example in the introductory section, it is relatively easy to find the conditional distribution of X given a statistic T. In general, however, it is quite difficult. A basic tool for determining sufficiency is the factorization theorem due originally to Fisher [5] and made more rigorous by Neyman [15] and Halmos & Savage [8].

Theorem 1 (factorization theorem). A statistic T is sufficient if and only if $p_\theta(x) = k(x)g(T(x), \theta)$ for some functions g and h.

It follows immediately that if $S(X)$ is an invertible function of a sufficient statistic $T(X)$ then $S(X)$ is also sufficient.

Example 1

Let $X_1, X_2, \ldots, X_n$ be independent, $X_i \sim \mathrm{N}(\mu, \sigma^2)$. Then $X = (X_1, \ldots, X_n)$ and $\theta = (\mu, \sigma^2)$. If $T = (T_1, T_2) = \left(\sum X, \sum X^2\right)$, then T is sufficient, since

$$p_\theta(x) = (2\pi)^{-n/2}\sigma^{-n} \exp\left(\frac{T_1\mu}{\sigma^2} - \frac{T_2}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right).$$

Furthermore, $\overline{X} = T_1/n$ and

$$S^2 = \frac{\sum(X_i - \overline{X})^2}{n-1} = \frac{T_2 - T_1^2/n}{n-1};$$

the sample mean and variance also form a sufficient statistic since $(\overline{X}, S^2)$ is an invertible function of (T_1, T_2).

Minimal Sufficiency

In any model the data X itself clearly is a sufficient statistic. It is of interest to find the sufficient statistic which in some sense has the simplest structure. One way in which to formalize this idea is to define a *minimal sufficient statistic* as a sufficient statistic that is a function of every other sufficient statistic.

One way in which to find a minimal sufficient statistic, due to Lehmann & Scheffé [12], is the following. Define an equivalence relation on the sample space $\mathcal{X}$ by $x_1 \equiv x_2$ if $p_\theta(x_1) = h(x_1, x_2)p_\theta(x_2)$ for all θ, where h is not zero and does not depend on θ. Let $C_x = \{y \in \mathcal{X} : y \equiv x\}$ be the set of ys equivalent to x. Then C_x defines a partition of the sample space. Any statistic $T(x)$ which takes on different values on different sets of the partition but is constant on each such set is a minimal sufficient statistic.

In fact, any (sufficient) statistic $S(x)$ defines a partition where a generic member is $C_x = \{y : S(y) = S(x)\}$. The partition defined in the previous paragraph is unique and is called the minimal sufficient partition. The partition corresponding to any sufficient statistic is a refinement of the minimal sufficient partition in the sense that any set in the minimal sufficient partition is the union of (at least one) sets in the given partition.

Furthermore, two sufficient statistics are invertible functions of one another if and only if they generate the same partition. It follows that if T is minimal sufficient and S is an invertible function of T, then S is also minimal sufficient. See Lindgren [13] for an extended discussion.

Perhaps the most useful example of the Lehmann–Scheffé construction of a minimal sufficient statistic is the following

Theorem 2 (exponential family). Let $p_\theta(x) = c(\theta)h(x)\exp[\sum_1^k \eta_i(\theta)T_i(x)]$. If $\eta_1(\theta), \ldots, \eta_k(\theta)$ are linearly independent, then $[T_1(x), \ldots, T_k(x)]$ is minimal sufficient.

Example 2

For the setup of Example 1, it is easily seen that $T = \left(\sum X_i, \sum X_i^2\right)$ is minimal sufficient as is $(\overline{X}, S^2)$. Here $\theta = (\mu, \sigma^2)$, $\eta_1(\theta) = \mu/\sigma^2$, and $\eta_2(\theta) = -1/(2\sigma^2)$.

Many of the families of distributions of classical statistics, including the **binomial**, **Poisson**, and **gamma**, provide examples of so-called exponential families of distributions to which the above theorem applies. In these examples the dimension of the sufficient statistic T is equal to k, the number of components of $[\eta_1(\theta), \ldots, \eta_k(\theta)]$. Often, as in Example 2, this is also the dimension of θ, but not always. For example, if in Example 2 $\sigma^2 = \mu^2$, then the dimension of $\theta = \mu$ is 1 but $\eta_1(\theta) = 1/\mu$ and $\eta_2(\theta) = -1/2\mu^2$ are still linearly independent and hence $(\overline{X}, S^2)$ remain minimal sufficient.

Occasionally, minimal sufficiency provides very little reduction of the original data. For example if $X = (X_1, \ldots, X_n)$, where the X_i are independent and identically distributed with a **Cauchy** or double exponential distribution with unknown location parameter, then the **order statistics** $(X_{(1)}, \ldots, X_{(n)})$ are minimal sufficient.

Ancillary Statistics

The degree of reduction achieved by minimal sufficient statistics varies greatly from problem to problem. As we have just mentioned, the order statistics are minimal sufficient when the data is from a Cauchy

population with an unknown location. By contrast, if the population is normally distributed with an unknown location, it follows from Theorem 2 that the sample mean is minimal sufficient.

In the former case, the minimal sufficient statistic contains ancillary information in the sense that, for example, the distribution of the sample **range** (the maximum minus the minimum) is a function of the minimal sufficient statistic, but contains no information about the location θ. A statistic $V(X)$ the distribution of which does not depend on θ is said to be an **ancillary statistic**. Reduction by sufficiency seems most successful when the minimal sufficient statistic contains no ancillary information. This is closely related to the notion of a complete sufficient statistic, which we now discuss.

Complete Sufficient Statistics

$T(X)$ is a *complete sufficient statistic* if $\mathrm{E}_\theta f(T) = 0$ for all θ implies that $f(T) = 0$ for all T (actually, for almost all T). The following result of Basu (see for example, Lehmann [10]) implies that a complete sufficient statistic carries no ancillary information.

Theorem 3 (Basu). If T is a complete sufficient statistic, then any ancillary statistic V is independent of T.

It follows that if $f(T)$ is any function of a complete sufficient statistic T, then $f(T)$ is independent of T, and by completeness must be constant. Hence no nontrivial function of a complete sufficient statistic can be ancillary.

A complete sufficient statistic is minimal complete. A result which gives a complete sufficient statistic for many classical statistical models is the following.

Theorem 4 (completeness for exponential families). If in the exponential family setup of Theorem 2, the set $N = \{\eta = [\eta_1(\theta), \eta_2(\theta), \ldots, \eta_k(\theta)] : \theta \in \Theta\}$ contains an open set, then T is a complete sufficient statistic.

Hence $\overline{X}$ is a complete sufficient statistic for sampling from a normal population with unknown location (and known scale). Similarly, $(\overline{X}, S^2)$ is a complete sufficient statistic if both the location and scale are unknown.

If, however, we sample from a normal population with unknown mean, μ, and variance equal to μ^2, then $(\overline{X}, S^2)$ is minimal sufficient (by Theorem 2) but not complete since $\mathrm{E}(n\overline{X}^2/(n+1) - S^2) = 0$ for all μ. Here $[\eta_1(\theta)\eta_2(\theta)] = (1/\mu, -1/(2\mu^2))$ and N is a curve in two dimensions, which does not contain a two-dimensional open set.

It is not necessary that N contain an open set in order for T to be minimal sufficient. For example, Messig & Strawderman [14] show that a class of commonly used **quantal response models** are exponential family models with a complete sufficient statistic, but that the space N does not contain an open set.

Some Uses of Sufficiency

There are a variety of results in different contexts which allow us to find a statistical procedure based on a sufficient statistic which is as good as, or better than, a given procedure. We discuss some of these results in this section. We suppose throughout that we are given data X with distribution $p_\theta(x)$ and that $T(X)$ is a sufficient statistic.

Estimation

Suppose that we wish to estimate $\tau(\theta)$, a possibly vector valued function of θ, and that $\delta(x)$ is any estimator. Define $\delta^*(T) = \mathrm{E}[\delta(x)|T]$. Since T is sufficient, $\delta^*(T)$ does not depend on θ and hence is itself an estimator of $\tau(\theta)$. By elementary properties of conditional expectation $\mathrm{E}_\theta\delta(x) = \mathrm{E}_\theta\delta^*(T)$, and hence δ^* is unbiased for $\tau(\theta)$ if δ is. Furthermore, the variance of δ^* is at least as small as the variance of δ and is, in fact, strictly smaller for every θ if $\delta(x)$ is not itself a function of T (and the variance of δ is finite).

If, in addition, T is a complete sufficient statistic, δ^* is the unique estimator of minimum variance among all estimators with expectation equal to that of $\delta(x)$ for all θ. In particular, if $\delta(x)$ is unbiased for $\tau(\theta)$, then δ^* is the unique minimum variance estimator provided that an unbiased estimator with finite variance exists (*see* **Minimum Variance Unbiased (MVU) Estimator**; **Unbiasedness**).

The above results can be considerably broadened along the following lines. A loss function $L(\theta, a)$ measures the "loss" (*see* **Loss Function**) when we

estimate $\tau(\theta)$ by a and θ is the true value. The risk function of an estimator $\delta(x)$ is defined by $R(\theta, \delta) = \mathrm{E}_\theta L[\theta, \delta(x)]$. For example, a common loss is so-called squared error loss, $L(\theta) = [\tau(\theta) - a]^2$ [if $\tau(\theta)$ is one-dimensional]. In this case, the risk function $R(\theta, \delta) = \mathrm{E}_\theta[\delta(x) - \tau(\theta)]^2$ is mean square error. We may generalize the above results to the following theorems.

Theorem 5 (Rao–Blackwell). If $L(\theta, a)$ is a (strictly) convex function in a, then $R(\theta, \delta^*) \leq R(\theta, \delta)$ with (strict) inequality for all θ provided that δ is not a function of T and $R(\theta, \delta)$ is finite [16].

Theorem 6 (Lehmann–Scheffé). Let $L(\theta, a)$ be convex in a and let T be a complete sufficient statistic. If $\delta(x)$ is an unbiased estimator of $\tau(\theta)$ with finite risk, then δ^* uniformly minimizes the risk among all unbiased estimators.

Hypothesis Testing

Consider testing the hypothesis $\mathrm{H}_0 : \theta \in \Theta_0$ against $\mathrm{H}_a : \theta \in \Theta_0^c$. Let $\phi(x)$ be any *critical function* (or *test function*) defining the probability of inclusion in the **critical region** for an observation x, and let $\phi^*(T) = \mathrm{E}[\phi(x)|T]$ (*see* **Hypothesis Testing**). Then ϕ^* is also a critical function. Furthermore, $\mathrm{E}_\theta\phi = \mathrm{E}_\theta\phi^*$ and hence ϕ and ϕ^* have the same **power** function. If ϕ is unbiased, so is ϕ^*.

Completeness is particularly helpful in obtaining uniformly **most powerful** unbiased (UMPU) tests of and one- and two-sided hypotheses concerning a one-dimensional parameter in the presence of **nuisance parameters** (*see* **Alternative Hypothesis**).

Suppose, for example, that the statistical model is a $k + 1$ parameter ($\theta = (\mu, \gamma_1, \gamma_2, \ldots, \gamma_k)$) exponential family with density $p_\theta(x) = c(\mu, \gamma_1, \gamma_2, \ldots, \gamma_k)$ $h(x) \exp[\mu U(x) + \sum \gamma_i T_i(X)]$, and we wish to test a hypothesis about μ. To be specific, suppose that we wish to test $\mathrm{H}_0 : \mu \leq \mu_0$ against $\mathrm{H}_\mathrm{a} : \mu > \mu_0$. Here $(\gamma_1, \ldots, \gamma_k)$ are the nuisance parameters. The boundary between the null (H_0) and alternative (H_a) hypotheses is $\Theta_B = \{\theta = (\mu, \gamma_1, \ldots, \gamma_k) : \mu = \mu_0\}$. With θ restricted to Θ_B, X has a k parameter exponential family with complete sufficient statistic $T(X) = [T_1(X), T_2(X), \ldots, T_k(X)]$. Furthermore, the distribution of $U(X)$ conditional on $T(X)$ is a one-parameter exponential family with parameter μ. The uniformly most powerful test of H_0 against H_a based on this conditional distribution is easily found to be a one-sided test $\phi^*(U, T)$ based on U (conditional on T) with the property that $\mathrm{E}_{\mu_0}[\phi(U, T)|T] = \alpha$.

However, any unbiased size α test$^*\phi(x)$ must be such that $\mathrm{E}_\theta\phi(x) = \alpha$ for all $\theta \in \Theta_B$. Completeness of T for $\theta \in \Theta_B$ implies that $\mathrm{E}_{\mu_0}[\phi(x)|T] = \alpha$ for all T, or that ϕ has what is called *Neyman structure*. But as the above one-sided test is uniformly most powerful among all such tests, it is uniformly most powerful among all unbiased size α tests and is therefore UMPU. Similar arguments work for two-sided hypotheses about μ in this model. For a more detailed development, see Lehmann [11] and Ferguson [4].

It is important to remark that the one-dimensional parameter μ about which we wish to test a hypothesis may be any one-dimensional affine function of the so-called natural parameters. In this way UMPU tests may be found, for example, for the equality of two binomial or two Poisson parameters.

Likelihood Methods and Information

The function $g(T(X), \theta)$ in the factorization theorem (Theorem 1) may be taken to be the density of T. Thus, for example, the **maximum likelihood** estimator of θ will depend on X only through $T(X)$ and coincides with the maximum likelihood estimator obtained if only $T(X)$ is observed. Furthermore, the **likelihood ratio test** depends only on the sufficient statistic T and is the same as that for which only T is observed. It also follows from the factorization theorem that the Fisher information in a sufficient statistic T is equal to the Fisher information in the original sample X.

Bayesian Methods

Bayesian methods in statistics rely on treating the parameter θ as a random variable that has a **prior distribution** $\pi(\theta)$. The statistical model $p_\theta(x)$ is treated as the model of the distribution of x conditional on the value of the random variable θ. Statistical procedures are based on the posterior distribution, the conditional distribution of θ given the data X which is proportional to $p_\theta(x)\pi(\theta)$, the constant of proportionality depending on x.

If $T(X)$ is a sufficient statistic it follows from the factorization theorem that the conditional distribution of θ given x is the same as the conditional distribution of θ given $T(X)$. Hence Bayesian statistical methods will depend on X through the sufficient statistic T (in regular models). See, however, Blackwell & Ramanoorthi [2] for an example of a nonregular model in which the Bayesian notion of sufficiency does not coincide with the classical one.

Invariance

The important role of sufficiency in statistics is due to the fact that it allows a reduction in the complexity of the sample space without a loss of information. Invariance is another method of reduction which is often useful. Briefly, a statistical model is invariant under a group of invertible transformations, $g \in G$, of the sample space on to itself if, for each $g \in G$ and $\theta \in \Theta$, the distribution of gX is $P_{\theta'}$ for some $\theta' \in \Theta$. In many problems reduction by both sufficiency and invariance is helpful. In some of these problems, it makes a difference which method of reduction is applied first. As reduction by invariance typically involves some loss of information, Ferguson [4] suggests reducing first by sufficiency. For details about the relationship between sufficiency and invariance, see Hall et al. [7].

It often happens that a best invariant procedure (if one exists) may be calculated as a Bayes procedure relative to a particular invariant prior measure. It follows from the discussion in the previous subsection that in this case the best invariant procedure will be a function of the sufficient statistic T.

Concluding Comments

Virtually every textbook on mathematical statistics discusses sufficient statistics. Lindgren [13] is a particularly nice nonmeasure theory treatment, especially of minimal sufficiency and minimal sufficient partitions: see also Casella & Berger [3]. Ferguson [4] has a nonmeasure theory discussion of sufficiency in the context of **decision theory**. Lehmann [11] presents a nice measure theory treatment of the basic properties of sufficiency and remains the basic reference to hypothesis testing.

References

[1] Bahadur, R.R. (1954). Sufficiency and statistical decision functions, *Annals of Mathematical Statistics* **25**, 423–462.

[2] Blackwell, D. & Ramanoorthi, R.V. (1982). A Bayes but not classically sufficient statistic, *Annals of Mathematical Statistics* **10**, 1025–1026.

[3] Casella, G. & Berger, R. (1990). *Statistical Inference*. Wadsworth and Brooks/Cole, Belmont.

[4] Ferguson, T.S. (1967). *Mathematical Statistics, a Decision Theoretic Approach*. Academic Press, New York.

[5] Fisher, R.A. (1922). On the mathematical foundation of theoretical statistics, *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

[6] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

[7] Hall, W.J., Wijsman, R.A. & Ghosh, J.K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis, *Annals of Mathematical Statistics* **36**, 575–614.

[8] Halmos, P.R. & Savage, L.J. (1949). Applications of the Radon-Nikodým theorem to the theory of sufficient statistics, *Annals of Mathematical Statistics* **20**, 225–241.

[9] Hurzurbazar, U.S. (1976). *Sufficient Statistics*. Marcel Dekker, New York.

[10] Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

[11] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd Ed. Wiley, New York.

[12] Lehmann, E.L. & Scheffé, H. (1950, 1955). Completeness, similar regions and unbiased estimation, *Sankhyā* **10**, 305–340; **15**, 219–236.

[13] Lindgren, B.W. (1976). *Statistical Theory*, 3rd Ed. Macmillan, New York.

[14] Messig, M.A. & Strawderman, W.E. (1993). Minimal sufficiency and completeness for dichotomous quantal response models, *Annals of Statistics* **21**, 2149–2157.

[15] Neyman, J. (1935). Sur un teorema concernte le cosidette statistiche sufficienti, *Giorn. Ist. Ital. Att.* **6**, 320–334.

[16] Rao, C.R. (1945). Information and accuracy attainable in estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society* **37**, 81–91.

(*See also* **Inference**; **Likelihood**; **Sufficiency**)

WILLIAM E. STRAWDERMAN

Summary Measures Analysis of Longitudinal Data

The method of summary measures is one of the most important and straightforward methods for the analysis of longitudinal data. If the measurements and observation times on the ith individual in a study are written as the vector $\mathbf{x}_i$, then a scalar-valued function f is chosen so that $s_i = f(\mathbf{x}_i)$ summarizes some essential feature of the response over time for that individual; the s_i are known as summary measures and some specific examples are given in the next section. Further analysis proceeds by applying standard univariate methods to the summary measures. The approach has been in use for many years [1, 3–6] and has been referred to by various names, including *response feature analysis* and *profile analysis*.

Choosing Summary Measures

The key step is the definition of the summary measure f. There are few restrictions on the type of summary that can be used, the main one being that the summary should make sense both in terms of the study and in the broader scientific context in which the study takes place. It is also advantageous if the summary can be specified before the data are collected, although this is not always possible. Some commonly used summaries include:

1. The rate at which an outcome changes (e.g. a growth rate), for which a suitable f may be the slope of a regression line (Figure 1(a)).
2. The general level of the response, which could be measured by the mean or median of all responses (Figure 1(b)); in many studies, especially in pharmacology, the area under the response curve is used.
3. Summaries defined in terms of the time axis, such as the time to the peak response or the time that a drug concentration stays above a therapeutic level: these can have particular clinical relevance (Figure 1(c)).

Although many summary measures will be closely related to one of the above types, any form of f that gives a valid numerical representation of a scientifically important aspect of the response is allowed. Indeed, there may be circumstances when the observation at a single time point is a suitable and comprehensible summary. However, such analyses must be distinguished from the flawed approach that analyzes separately all time points (*see* **Time-by-time Analysis of Longitudinal Data**).

Advantages of the Method

Three advantages of the method are: (i) the analysis avoids the need to consider the correlation structure of the whole response $\mathbf{x}_i$, and the different s_i can reasonably be assumed to be independent, so the statistical basis of the method is sounder than that of some methods in widespread use (*see* **Analysis of Variance for Longitudinal Data**); (ii) the final analysis of the s_i uses simple methods that are well known; and (iii) the analysis is readily interpreted because it is based on summaries that have been chosen for their relevance to the study.

Furthermore, determining what is a relevant summary measure forces the investigator to think carefully *and quantitatively* about the questions the study addresses; if this process occurs before data are collected, then it can lead to improvements in the design of the study, such as the proposed times at which outcomes will be recorded. It should also be noted that many summary measures will be correlated with the observed value at time zero, so if this is a legitimate covariate (as it would often be in, for example, a **clinical trial**), then using it in the analysis can lead to substantial gains in precision.

More than One Summary Measure

In many studies, more than one aspect of the response is of interest and, provided some care is taken, more than one summary can be defined in the analysis of a single study. Indeed, if a second summary measure, $t_i = g(\mathbf{x}_i)$, has been identified, then not only can the s_i and t_i be analyzed separately, but the relation between the summaries can be considered; often this will amount to no more than producing a scatterplot, but more formal methods might be adopted.

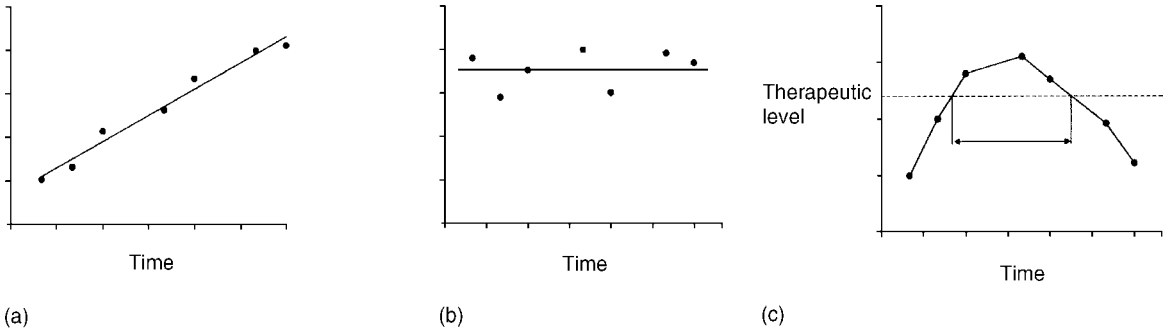

Figure 1 Examples of some common types of summary measure: (a) slope of regression line; (b) mean response; (c) time above a therapeutic level

Care must be taken to ensure that different summary measures address distinct features of the response. As with many aspects of this method, scientific judgment is as important as statistical expertise in deciding what constitutes "distinct features". The statistician should always be alert to the possibility that two apparently distinct summaries may measure the same thing. For example, in a study in which tumors of approximately the same size are transplanted into laboratory animals, the slope of the regression line of tumor size against time over the first 10 days may represent essentially the same information as the size of the tumor at the tenth day.

In practice, there are likely to be few studies in which the use of more than two or three summary measures will be helpful.

Missing Data and Irregular Time Points

In many instances, the definition of a summary measure is sufficiently flexible that missing values can easily be accommodated. Suppose measurements are anticipated on six occasions and the summary is their mean; then, if one of the measurements is missing, the summary is taken to be the mean of the remaining five observations. In practice, if the proportion of missing data points is small, then this type of device is likely to be satisfactory but there are two problems.

The first is that if the s_i are based on $\mathbf{x}_i$ that have widely differing structures then, even within apparently homogeneous groups, the s_i will not share a common distribution, contrary to the assumptions of most of the methods of analysis that are likely to be applied to the summaries. For example, means based on different numbers of observations, regression slopes based on differently located observations and maxima of sets of different sizes, will not share common distributions. Even in complete data sets, observations may be taken at different times on different individuals and this presents the same problems. The importance of this needs to be judged in each application, and will depend particularly on the relative sizes of sampling error and between-individual variation [2].

The second problem, potentially more serious, is the effect of *why* the observations are missing. If interest focuses on the maximum drug concentration and some adverse effect of high concentrations stops this being observed, for example by preventing attendance at the clinic, then it would be seriously misleading simply to use the maximum of the observations obtained. Of course, such problems occur throughout statistics but, because missing values present few computational difficulties for this technique, the analyst needs to be especially aware that this confers no special immunity from the potential dangers of missing data.

Disadvantages

The main problem with the method is that it might not be suitable for certain applications. In some instances, it may simply be impossible to identify a suitable summary measure. If there is interest in how changes over time in the response (for example, postoperative pain relief) relate to another variable (for example, plasma concentration of analgesic), then the method of summary measures has little to offer.

References

[1] Healy, M.J.R. (1981). Some problems of repeated measurements, in *Perspective in Medical Statistics*, J.F. Bithell & R. Coppi, eds. Academic Press, London, pp. 155–171.

[2] Matthews, J.N.S. (1993). A refinement to the analysis of serial data using summary measures, *Statistics in Medicine* **12**, 27–37.

[3] Matthews, J.N.S., Altman, D.G., Campbell, M.J. & Royston, P. (1990). Analysis of serial measurements in medical research, *British Medical Journal* **300**, 230–235.

[4] Oldham, P.D. (1962). A note on the analysis of repeated measurements of the same subjects, *Journal of Chronic Diseases* **15**, 969–977.

[5] Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis, *Biometrika* **30**, 16–28.

[6] Yates, F. (1982). Regression models for repeated measurements, *Biometrics* **38**, 850–853.

(*See also* **Diggle–Kenward Model for Dropouts**; **Nonignorable Dropout in Longitudinal Studies**)

JOHN N.S. MATTHEWS

Superpopulation Models in Survey Sampling

A finite population is any collection of distinct entities (*units*) such as people, businesses, medical files, hospitals, or schools. Finite population sampling, sometimes called survey sampling, is concerned with selecting subsets (*samples*) of the units, observing features of the units, and then using the observations to make **inferences** about the entire population. For example, the population is "all short-stay hospitals in the United States," a sample of short-stay hospitals is selected, and for each hospital in the sample the number of patients discharged during a particular calendar year is measured. The goal might be to estimate the total number of discharges in the entire US (*see* **Estimation**).

Invariably, we bring a host of preconceptions and, possibly, hard relevant information (*auxiliary data*), to the study of the population. For example, we may know the number of beds in each of the hospitals in the population, and may strongly suspect that the number of discharges is related to number of beds.

A *superpopulation model* of a particular population is a probability model characterizing the population, formalizing our conceptions and knowledge of the population. Quantities of interest in the population are posited to be realizations of **random variables** with a particular joint probability distribution. For example, in the case of the hospital population, we might suppose

$$Y_i = \beta x_i + x_i^{1/2}\varepsilon_i, \quad i = 1, 2, \ldots, N, \qquad (1)$$

where, for the ith of the N hospitals constituting the population, Y_i is the to-be-realized number of patient discharges per year, x_i is the known number of beds, β is an unknown constant of proportionality, and ε_i is a **random error** with **mean** zero and constant **variance** σ^2, and is assumed to be independent across units. The inclusion of the $x_i^{1/2}$ factor in the error term is meant to capture the idea of greater variation in the larger hospitals.

Such models are invaluable aids in planning sample selection, in constructing estimators of quantities of interest, and in assaying the precision of estimates.

Surveys can be undertaken with *two* distinct sorts of estimation in mind [5, Chapter 7]. In the first sort, referred to most commonly as *analytic* or *inferential*, the aim is to understand the process underlying relations between variables in the population. In this context, the superpopulation model is understood to be an attempt to characterize the way the population (and possibly others like it) has come to be. In any case, the goal is to estimate the parameter(s) of the superpopulation model. Thus, by *definition*, a superpopulation model is necessary for analytic estimation. We might, in the hospital example, be interested in the law of proportionality between beds and discharges, i.e. in estimating the unknown parameter β. A characteristic of analytic inference is that even if the entire population were available, there would be uncertainty in the estimates of the model parameters.

The second, and in fact, more common, goal, is *descriptive*: the estimation of functions, like total, mean, or **quantiles**, of the population quantities themselves. Thus, in the hospital example, we might wish to estimate $T = \sum_{i=1}^{N} y_i$, the total number of discharges from the hospitals. Such quantities are sometimes referred to as "descriptive parameters". In distinction to the analytic case, if the whole population were known, there would remain no uncertainty about the value of the descriptive parameter.

Suppose, in the hospital example, that a sample s of size n is taken from the population P and yields y_i, $i \in s$. The plot in Figure 1 of sample discharges (y) against corresponding sample beds (x), appears to be in keeping with the superpopulation model (1). Then part of T is known, namely the realization of the total for sample units, $T_s = \sum_s Y_i$. The total on the remainder $r = P - s$, namely $T_r = \sum_r Y_i$, is unrealized and unknown, but the sample ys in combination

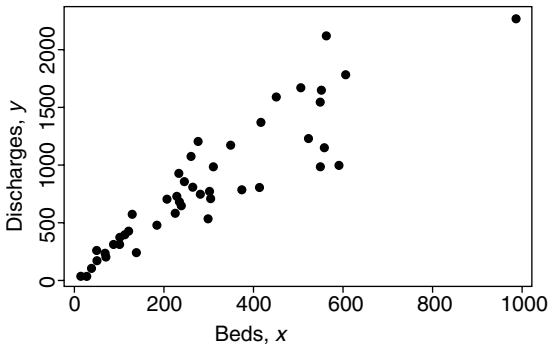

Figure 1 Number of patients discharged and number of beds in 50 short-stay hospitals

with the already known corresponding xs can be used to estimate the slope β. We might, for example, use the best linear **unbiased** estimator $\hat{\beta} = \sum_s Y_i / \sum_s x_i$, not for the sake of getting β itself (as in analysis), but as a means to estimating, or more precisely, predicting, $T_r = \sum_r Y_i$. The predictor $\hat{T}_r = \hat{\beta} \sum_r x_i$ can readily be shown to be *unbiased* for T_r, in the sense that $\mathrm{E}_M(\hat{T}_r - T_r) = \beta \sum_r x_i - \beta \sum_r x_i = 0$, where the **expectation** is with respect to the superpopulation model (1). Basically, one fills in the unknown ys by their predicted values under (1) based on the sample data, and uses these as surrogates to predict the unknown total T_r. Overall, one estimates the population total T by $\hat{T}_R = T_s + \hat{T}_r = \hat{\beta} \sum_P x_i$, which is well known as the *ratio estimator* (*see* **Ratio and Regression Estimates**) (note that if $s = P$, then $\hat{T} = T$). Then $\hat{T}$ will be the best linear unbiased predictor (BLUP) of T, under (1) (see Theorem 1 below).

(It should be noted in the above example, that no strictures were made about how to choose the sample s, other than that (implicitly) it not be in such a way as to make the *sample* pairs (x_i, Y_i) deviate from the model (1). In particular, no assumption was made about using a **random sampling** plan.)

This idea – of using a *superpopulation* model such as (1) to *predict* unknown characteristics of the *population* – has become known as the prediction, or model-based, approach, to survey sampling. The question arises: Can *every* population be characterized by a corresponding superpopulation model?

The presence of auxiliary data, such as number of beds, is *not* needed in order to posit a reasonable model. If a population is homogeneous (and that is all one knows), then a reasonable model is

$$Y_i = \mu + \varepsilon_i, \quad i = 1, 2, \ldots, N, \tag{2}$$

where, for the ith of the N units constituting the population, Y_i is the to-be-realized variate of interest, μ is an unknown constant, and ε_i, the random error, is assumed independent across units and has mean 0 and variance σ^2. It can be shown that $\hat{T} = N\overline{Y}_s$, where $\overline{Y}_s$ is the mean of sample Ys, is the BLUP under (2); see Theorem 1 below.

This particular estimator corresponds exactly to the classic *expansion estimator* for a total, that is unbiased with respect to the probability distribution of samples generated by a **simple random sample** (SRS) sampling scheme. It might be argued that this randomization distribution (of the sampler's activity in choosing the sample) is the adequate basis for choosing the expansion estimator and analyzing its properties. The method of inference that relies only on the random sampling plan for its probability distribution is known as *design-based* inference. To see that design-based inference is inadequate, consider the following thought experiment.

Suppose that one had an utterly heterogeneous population of items. For example, suppose that item 1 is a raisin, item 2 a 1957 Chevrolet, ..., item N a bit of dust. Suppose that the nature of any particular unit is unknown prior to sampling, and that one seeks the mean weight in grams of the items in the population. In this case, knowledge of any number of measurements of sampled items conveys no information regarding any of the nonsampled items. This will be the case whether or not the sample is constructed using SRS or any other probabilistic (or nonprobabilistic) scheme. Thus, we can make no inference from sample mean to population mean (or from any sample quantity to any population quantity) in this extreme sort of population. This is a reflection of the fact that neither the model (2), nor any other model, fits the population and can serve as an inferential bridge from sample to population.

(The software-ready reader might try the following simulation experiment intended to mimic to a degree the above thought experiment: generate a population of size $N = 1000$ having y-values α^i with $\alpha = 1.1$, $i = 1, 2, \ldots, N$; this population has mean $\overline{y} \approx (2.72)(10^{39})$. Suppose that the mode of generation of the population, and, in particular, the sequence number i of each unit and the constant α are unknown to you the sampler. Taking SRS samples of size $n = 100$, for each of the, say 1000, samples, calculate the sample mean $\overline{y}_s$, the SRS sample variance $v = n^{-1}(1 - n/N)(n-1)^{-1} \sum_s (y_i - \overline{y}_s)^2$, and the t-statistic $(\overline{y}_s - \overline{y})/v^{1/2}$ (*see* **Student's *t* Statistics**). Consider the distribution of the t-statistic. For adequate inference, the t statistic should lie between -2 and 2 in about 95% of runs.)

There has been much debate about the merits of model-based versus design-based inference. A key distinction between the two is that, given a sample, the model-based approach seeks to describe properties of estimators in the particular sample selected, in contrast to the design-based approach which calculates statistical properties by averaging across all the

samples that could have been selected using a specified sampling plan. Interesting additional reading can be found in [1, 9, 12, 14, 17, 23, 26].

The Need for Robustness to Model Failure

We suppose then, that, for inference from sample to population to be possible, there must be at least an implicit model of the population that the sampler can make explicit. But in making the model explicit, there are bound to be simplifications and even distortions. For example, in the model (1) adopted for the hospital population above, it was assumed implicitly that there was no curvature characterizing the relation between x and Y, that the variance went up as x rather than, say, x^2. Thus, if one is to make serious use of a superpopulation model in sampling inference, it is necessary to take into account the likelihood of deviations from the model, which may or may not be detectable from the sample data. It is necessary to "robustify" inferences against model failure (*see* **Robustness**).

One way to do this is to examine the effects, especially the **bias**, that arise if one model – the *working model* – is used as a basis for an estimator, and *another* model is correct. This idea goes back to the early stages of modern sampling: in his introduction of what effectively was a linear superpopulation model, Cochran [3] raised the question of the effect of having a quadratic model. The most serious work in this regard has been carried out by Royall and his colleagues [4, 15, 19, 20], who introduced the notion of *balanced samples* as a means of protecting against model failure of certain estimators.

To illustrate, let (1) be the working model, and suppose that if we had the population as a whole, we would recognize that the expected value of Y_i obeys

$$\mathrm{E}_M(Y_i) = \alpha + \beta x_i + \gamma x_i^2, \quad i = 1, 2, \ldots, N. \quad (3)$$

Then it is readily shown that the ratio estimator $\hat{T}_R = \hat{\beta} \sum_P x_i$, based on the working model (1) has a bias, given by

$$\mathrm{E}_M(\hat{T} - T) = N\alpha \left(\frac{\overline{x}}{\overline{x}_s} - 1 \right) + N\gamma \left(\frac{\overline{x}}{\overline{x}_s} \overline{x}_s^{(2)} - \overline{x}^{(2)} \right), \quad (4)$$

where $\overline{x}^{(2)} = N^{-1} \sum_P x_i^2$ and $\overline{x}_s^{(2)} = n^{-1} \sum_s x_i^2$. If the sample is selected to have the moments $\overline{x}_s$ and $\overline{x}_s^2$ equal to the corresponding population moments, then the bias (4) is *zero*. Such a sample is called *balanced*. Similarly, by balancing on moments up to the Jth order, the ratio estimator is protected against bias, if the underlying model is a polynomial of order J. Since smooth functions can be approximated by polynomials of high order, the implication is that, if the underlying model is unspecified but smooth, balanced sampling makes the ratio estimator bias-robust against a very wide range of underlying conditions.

This idea of balance can be generalized to a wide range of estimators (see [8, 15] and Theorem 2 below). Thus, for example, the best linear unbiased estimator $\hat{T}$ based on the superpopulation model

$$Y_i = \beta x_i + \gamma x_i^2 + x_i \varepsilon_i, \quad i = 1, 2, \ldots, N, \quad (5)$$

where ε_i is mean zero with constant variance, independent across units, will be bias-robust against high-order polynomials, if $\overline{x}_s^{(j-1)} = \overline{x}^{(j)}/\overline{x}$, for $j = 0, 2, \ldots, J$ – a sort of weighted balance with weights equal to x_i^{-1}; it will be more efficient than the ratio estimator if, as in the model (5), the variance of Y is proportional to x.

Balance is not a panacea; for example, it does not help if the goal is to estimate a distribution function $F(t) = N^{-1} \sum_P I(Y_i \le t)$ [where $I(A) = 1$, if A holds, and is *zero* otherwise]. In such cases, manipulation of the form of the estimator is necessary to protect against model **misspecification** [2, 7, 8].

Variance Estimation

Reduction or elimination of bias is not an end in itself. However, if the bias is low relative to the square root of the variance, then sound inference, in the form of valid **confidence intervals** or **hypothesis tests**, is achievable, if we can form **consistent** variance estimators. A major concern in model-based sampling is the construction of variance estimates that are unbiased for $v \equiv \mathrm{var}_M(\hat{T} - T)$, despite deviations of the working model, including its **variance component**, from the correct model.

The basic idea underlying a variety of variance estimators is as follows. Under the working model, we can write the estimation error, $\hat{T} - T$, as $\sum_s a_i Y_i - \sum_r Y_j$, where the as depend on the known values in the working model, so that, under a model with independent errors, $v = \sum_s a_i^2 \sigma_i^2 + \sum_r \sigma_j^2$, with

$\sigma_i^2 = \text{var}_M(Y_i)$. For example, in the case of the ratio estimator, $a_i = \sum_r x_i / \sum_s x_i$. Then the first term in v, i.e. the sum over the *sample* points, dominates if, as is typical, $N - n \gg n$, and the key idea is to replace the σ_i^2 by (usually slightly adjusted) squares of the residuals $r_i^2 = (Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2$, where $\mathbf{x}_i$ is a vector of auxiliaries and $\hat{\boldsymbol{\beta}}$ is an estimated parameter vector. These squared residuals are nearly unbiased for σ_i^2, even when the variance component of the working model is misstated. The second, relatively minor component, can be estimated by $[(N-n)/n]\sum_s r_i^2$ or in more sophisticated ways. The net result is a variance estimator $\hat{v} = \sum_s a_i^2 r_i^2 +$ (an estimate of $\sum_r \sigma_j^2$) that, if the working model of **expectation** is correct, is consistent for v as n, N grow to infinity and n/N goes to zero. In the case in which the working model is incorrect, but the sample is balanced, v will be conservative. The reader is referred to the literature for explicit alternatives to the unadjusted residuals, and for further details; see especially [4, 13, 16–18, 25].

The impact of model-based sampling has been very large. Although present theorists and practitioners are relatively few, many new ideas and considerations, for example, the development of "model-assisted" survey sampling, are traceable, in large measure to model-based inspiration; cf. the remarks of Särndal et al. [21, p. 535].

In the remainder of this article, we outline some theory for the **general linear model**, give a theorem on weighted balance, sketch results for the case in which errors in the model are correlated, useful for **cluster sampling**, and indicate other work in model-based sampling, not covered here.

Prediction Theory under the General Linear Model

The estimation problem can be formulated for a general linear model and the best linear unbiased predictor derived under that model. The finite population consists of N units, each of which has a value of a target variable y associated with it. In the prediction approach, the population vector $\mathbf{y} = (y_1, \ldots, y_N)'$ is treated as the realization of a random vector $\mathbf{Y} = (Y_1, \ldots, Y_N)'$. A common goal is to estimate a linear combination of the Ys, such as the total of all the Ys or their mean. A linear combination of the Ys is defined to be $\boldsymbol{\gamma}'\mathbf{Y}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)'$ is an N-vector of constants. If, for example, each $\gamma_i = 1$, then the prediction target is the total; if $\gamma_i = 1/N$, the target is the mean. Because the target Ys are modeled as random variables, any linear combination $\boldsymbol{\gamma}'\mathbf{Y}$ is a random sum, and our problem is one of prediction. The population vector of Ys, given a particular sample s, can be reordered so that the first n elements are those in the sample and partitioned as $\mathbf{Y} = (\mathbf{Y}_s', \mathbf{Y}_r')'$, where the subvectors $\mathbf{Y}_s$ and $\mathbf{Y}_r$ are $n \times 1$ and $(N-n) \times 1$. To proceed we need to define the terms "linear estimator" and "estimation error".

Definition 1. A *linear estimator* of $\theta = \boldsymbol{\gamma}'\mathbf{Y}$ is defined as $\hat{\theta} = \mathbf{g}_s'\mathbf{Y}_s$, where $\mathbf{g}_s = (g_1, \ldots, g_n)'$ is an n-vector of coefficients.

Definition 2. The *estimation error* of an estimator $\mathbf{g}_s'\mathbf{Y}_s$ is $\hat{\theta} - \theta = \mathbf{g}_s'\mathbf{Y}_s - \boldsymbol{\gamma}'\mathbf{Y}$.

We study this prediction problem under the general linear model:

$$\text{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \qquad \text{var}(\mathbf{Y}) = \mathbf{V}, \tag{6}$$

where $\mathbf{X}$ is an $N \times p$ matrix of auxiliaries, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\mathbf{V}$ is a positive definite **covariance matrix**. To compute some estimators generated from Theorem 1 below, it is necessary that all auxiliary values be known for each unit in the population, although for others summary population statistics like means are sufficient. If the population elements are rearranged so that the first n elements of $\mathbf{Y}$ are those in the sample, and the first n rows of $\mathbf{X}$ are for units in the sample, then $\mathbf{X}$ and $\mathbf{V}$ can be expressed as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix},$$

where $\mathbf{X}_s$ is $n \times p$, $\mathbf{X}_r$ is $(N-n) \times p$, $\mathbf{V}_{ss}$ is $n \times n$, $\mathbf{V}_{rr}$ is $(N-n) \times (N-n)$, $\mathbf{V}_{sr}$ is $n \times (N-n)$, and $\mathbf{V}_{rs} = \mathbf{V}_{sr}'$.

The auxiliaries being known for each unit in the population, implies that a **sampling frame** has been constructed that lists every unit that is in the survey universe. In a universe of hospitals, auxiliaries could include, in addition to the aforementioned number of beds, the number of patients admitted during a previous time period, or the type of hospital – general medical and surgical, psychiatric, rehabilitation, and so on. There are many practical situations, particularly in surveys of households, where a complete list

of every population unit is not available. Such cases may require **multistage sampling**, which is addressed in the section on clustered populations.

Given a model, we can also define unbiasedness and variance as they apply in the context of **prediction** theory.

Definition 3. The estimator $\hat{\theta}$ is *prediction unbiased* for θ under a model M if $\mathrm{E}_M(\hat{\theta} - \theta) = 0$.

Definition 4. The *error variance* or (equivalently *prediction variance*) of $\hat{\theta}$ under a model M is $\mathrm{E}_M(\hat{\theta} - \theta)^2$.

The general theorem [11], giving the BLUP of $\hat{\theta}$ under model (6) is an extension of a standard result in prediction theory [27]:

Theorem 1. Among linear, prediction unbiased estimators $\hat{\theta}$ of θ, the error variance is minimized by

$$\hat{\theta}_{\text{opt}} = \boldsymbol{\gamma}_s'\mathbf{Y}_s + \boldsymbol{\gamma}_r'\left[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})\right], \tag{7}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{Y}_s$. The error variance of $\hat{\theta}$ is for $\mathbf{A}_s = \mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s$

$$\begin{aligned}\mathrm{var}_M(\hat{\theta} - \theta) = {} & \boldsymbol{\gamma}_r'(\mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr})\boldsymbol{\gamma}_r \\ & + \boldsymbol{\gamma}_r'(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s)\mathbf{A}_s^{-1} \\ & \times (\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s)'\boldsymbol{\gamma}_r. \end{aligned} \tag{8}$$

A feature of the BLUP, as noted earlier, is that it equals the weighted sum for the sample units, $\boldsymbol{\gamma}_s'\mathbf{Y}_s$ plus a predictor of the weighted sum for the nonsample units, $\boldsymbol{\gamma}_r'[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})]$. When the sample and nonsample units are uncorrelated, $\mathbf{V}_{rs} = \mathbf{0}$, the BLUP simplifies to $\hat{\theta}_{\text{opt}} = \boldsymbol{\gamma}_s'\mathbf{Y}_s + \boldsymbol{\gamma}_r'\mathbf{X}_r\hat{\boldsymbol{\beta}}$. The assumption that $\mathbf{V}_{rs} = \mathbf{0}$ will often be reasonable in situations in which single-stage sampling is appropriate, such as institution or establishment sampling.

To appreciate the formulation of the problem as one of prediction, rather than estimation, it is instructive to look at the results for the optimum $\hat{\theta}$ if we minimize its variance, $\mathrm{var}_M(\hat{\theta}) = \mathbf{g}_s'\mathbf{V}_{ss}\mathbf{g}_s$ instead of the error variance $\mathrm{var}_M(\hat{\theta} - \theta)$. In that case, the **minimum variance estimator** is $\hat{\theta}^* = \boldsymbol{\gamma}'\mathbf{X}\hat{\boldsymbol{\beta}}$. In other words, the value for each unit in the population is estimated as its expected value from the estimated **regression** model. Contrast this to $\hat{\theta}_{\text{opt}}$, where the sum for the sample units, $\boldsymbol{\gamma}_s'$ is used directly, and the sum for the nonsample units is predicted by the estimated regression mean, $\boldsymbol{\gamma}_r'\mathbf{X}_r\hat{\boldsymbol{\beta}}$ plus an adjustment based on sample **residuals**, $\boldsymbol{\gamma}_r'\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})$.

Many commonly used estimators can be derived by applying Theorem 1 to particular models. In the examples below, the estimation target is the finite population total $T = \sum_{i=1}^{N} y_i$, implying that $\boldsymbol{\gamma} = \mathbf{1}_N$ is a vector of N ones. The model that leads to the ratio estimator, for example, is (1). The estimator itself is $\hat{T}_R = \hat{\beta}\sum_P x_i$ and its error variance under the model is

$$\mathrm{var}_M(\hat{T}_R - T) = \frac{N^2}{n}(1 - f)\frac{\overline{x}_r\overline{x}}{\overline{x}_s}\sigma^2, \tag{9}$$

where $\overline{x}_r$ is the mean of x for the nonsample units, $\overline{x}$ is the population mean, and $f = n/N$.

The linear regression estimator comes from the model $Y_i = \alpha + \beta x_i + \varepsilon_i$, with the ε_is being independent with mean 0 and variance σ^2. The BLUP is $\hat{T}_{LR} = N[\overline{Y}_s + b(\overline{x} - \overline{x}_s)]$, where $b = \sum_s(Y_i - \overline{Y}_s)(x_i - \overline{x}_s)/\sum_s(x_i - \overline{x}_s)^2$. The error variance is $\mathrm{var}_M(\hat{T}_{LR} - T) = N^2(1 - f)\sigma^2[1 + (\overline{x}_s - \overline{x})^2/\{(1 - f)c_s\}]$, where $c_s = \sum_s(x_i - \overline{x}_s)^2/n$.

Another common estimator is the stratified expansion estimator. A set of strata is a collection of mutually exclusive groups that covers the entire population (*see* **Stratified Sampling**). Strata might be regions of a country, for example. Suppose that h denotes a stratum and that the model is $Y_{hi} = \mu_h + \varepsilon_{hi}$, with the ε_{hi}s being independent with mean 0 and variance σ_h^2. The BLUP is $\hat{T}_{st} = \sum_h N_h\overline{Y}_{hs}$, where N_h is the number of population units in stratum h, $\overline{Y}_{hs} = \sum_{s_h} Y_{hi}/n_h$, s_h is the set of sample units in stratum h, and n_h is the number of sample units in the stratum. The error variance is $\mathrm{var}_M(\hat{T}_{st} - T) = \sum_h N_h^2(1 - f_h)\sigma_h^2/n_h$, where $f_h = n_h/N_h$.

A final example is the mean-of-ratios estimator, which flows from the model $Y_i = \beta x_i + x_i\varepsilon_i$. The BLUP is $\hat{T} = \sum_s Y_i + \hat{\beta}\sum_r x_i$, where $\hat{\beta} = \sum_s Y_i/(nx_i)$. The error variance is $\mathrm{var}_M(\hat{T} - T) = \sigma^2[(N - n)^2\overline{x}_r^2/n + \sum_r x_i^2]$. When the sampling fraction f is small, the BLUP is approximated by the mean-of-ratios estimator $\hat{T}_{MR} = N\overline{x}\sum_s Y_i/(nx_i)$.

Per Unit Weights

When constructing a database from a sample survey, it is often operationally convenient to have a "weight"

associated with each unit in the sample that is used to calculate linear estimates. The weights are intended to be applicable to possibly *several* variables of interest y.

Now for a single y variable, the $n \times 1$ optimal vector of coefficients in a linear estimator, implied by Theorem 1, is

$$\mathbf{g}_s = \mathbf{V}_{ss}^{-1}[\mathbf{V}_{sr} - \mathbf{X}_s\mathbf{A}_s^{-1}(\mathbf{X}_r' - \mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr})]\boldsymbol{\gamma}_r + \boldsymbol{\gamma}_s.$$

Unit i would be assigned a weight equal to the ith component of the vector $\mathbf{g}_s$. The optimal weight depends, through the covariance structure, on the particular y variable being considered and on the way in which the population is split between the sample and nonsample units. Some examples follow.

For the expansion estimator $g_i = N/n$, for all $i \in s$. The ratio estimator has $g_i = N\overline{x}/(n\overline{x}_s)$. The linear regression estimator has $g_i = N[n^{-1} + (\overline{x} - \overline{x}_s)(x_i - \overline{x}_s)/\sum_{j\in s}(x_j - \overline{x}_s)^2]$. The weight for the stratified expansion estimator is $g_i = N_h/n_h$, for $i \in s_h$, and for the mean of ratios estimator is $g_i = N\overline{x}/(nx_i)$.

Notice that the coefficients a_i, in terms of which it is convenient to express the variance $v = \sum_s a_i^2\sigma_i^2 + \sum_r \sigma_j^2$ (discussed above for models with independent errors), are given by $a_i = g_i - 1$.

Although common survey practice is to use the same weight to make an estimate for different y variables, this would appear to be reasonable only when the ys follow the same general form of model. If one variable follows the expansion estimator model while another follows the regression estimator model, using the same weight for each is not generally sensible. However, in the case in which the sample is *balanced*, in the sense given in the next section, estimators of many forms can in effect be subsumed under one form, so that per unit weights are well-grounded.

It is worth noting that all of the examples that we have considered share a certain common structure. Let $\mathbf{1}_N$ and $\mathbf{1}_s$ be vectors of N and n ones. Suppose that $\mathbf{V}$ is diagonal and that the ith diagonal element can be expressed as $v_{ii} = \sigma^2 f(\mathbf{x}_i)$, with $f(\mathbf{x}_i) = \sum_{j=1}^{p} c_i x_{ij}$ a known function and $\mathbf{x}_i$ the vector of auxiliaries for unit i. In matrix terms, suppose that $\mathbf{V}\mathbf{1}_N = \mathbf{X}\mathbf{c}$ for a $p \times 1$ vector $\mathbf{c}$. The BLUP becomes $\hat{T} = \mathbf{1}_N'\mathbf{X}\hat{\boldsymbol{\beta}}$ (see Lemma 1 below). Note that this form of $\hat{T}$ is the same as would be obtained under the general linear model if we minimized $\mathrm{var}_M(\hat{T})$ rather than $\mathrm{var}_M(\hat{T} - T)$. Even if the variance condition $\mathbf{V}\mathbf{1}_N = \mathbf{X}\mathbf{c}$ does not hold, $\hat{T} = \mathbf{1}_N'\mathbf{X}\hat{\boldsymbol{\beta}}$ is still prediction unbiased under (6). The weight for the ith sample unit is then $g_i = N\overline{\mathbf{x}}[\sum_s \mathbf{x}_i\mathbf{x}_i'/f(\mathbf{x}_i)]^{-1}\mathbf{x}_i/f(\mathbf{x}_i)$, where $\overline{\mathbf{x}} = (\overline{x}_1, \ldots, \overline{x}_p)'$ is the vector of population means of the auxiliaries.

Weighted Balance and Robustness

Models that satisfy the variance condition $\mathbf{V}\mathbf{1}_N = \mathbf{X}\mathbf{c}$ play a key role in robustness and optimality. Let $M(\mathbf{X} : \mathbf{V})$ refer to the general linear model (not necessarily polynomial) with matrix $\mathbf{X}$ of auxiliary variables, and covariance matrix $\mathbf{V}$ given by (6). we first note the following.

Lemma 1 [15]. If $\mathbf{V}\mathbf{1}_N = \mathbf{X}\mathbf{c}$ for some vector $\mathbf{c}$, then the BLUP and its error variance are

$$\hat{T}(\mathbf{X} : \mathbf{V}) = \mathbf{1}_N'\mathbf{X}\hat{\boldsymbol{\beta}},$$
$$\mathrm{var}_M[\hat{T}(\mathbf{X} : \mathbf{V}) - T] = (\mathbf{1}_N'\mathbf{X}\mathbf{A}_s^{-1}\mathbf{X}'\mathbf{1}_N - \mathbf{1}_N'\mathbf{V}\mathbf{1}_N)\sigma^2. \tag{10}$$

Definition 5. The collection of samples that satisfy

$$\frac{1}{n}\mathbf{1}_s'\mathbf{W}_s^{-1/2}\mathbf{X}_s = \frac{\mathbf{1}_N'\mathbf{X}}{\mathbf{1}_N'\mathbf{W}^{1/2}\mathbf{1}_N} \tag{11}$$

will be denoted by $B(\mathbf{X} : \mathbf{W})$ and said to *balanced with respect to the weights root*$(\mathbf{W})$ or *root*$(\mathbf{W})$ *balanced*. Here $\mathbf{W}$ is an $N \times N$ matrix and $\mathbf{W}_s$ is the $n \times n$ submatrix for the sample units.

When $\mathbf{W} = \mathbf{I}$, $B(\mathbf{X} : \mathbf{I})$ is the set of samples that are balanced on the columns of $\mathbf{X}$, i.e. $\mathbf{1}_s'\mathbf{X}_s/n = \mathbf{1}_N'\mathbf{X}_N/N$. If the model for y is a polynomial in x, then $B(\mathbf{X} : \mathbf{I})$ is the set of samples satisfying $\overline{x}_s^{(j)} = \overline{x}^{(j)}$, the balance conditions introduced earlier.

Theorem 2 [15]. Under $M(\mathbf{X} : \mathbf{V})$, with $\mathbf{V}$ diagonal, if both $\mathbf{V}\mathbf{1}_N = \mathbf{X}\mathbf{c}$ and $\mathbf{V}^{1/2}\mathbf{1}_N = \mathbf{X}\mathbf{d}$, for some vectors $\mathbf{c}$, and $\mathbf{d}$, then

$$\mathrm{var}_M[\hat{T}(\mathbf{X} : \mathbf{V}) - T] \geq [n^{-1}(\mathbf{1}_N'\mathbf{V}^{1/2}\mathbf{1}_N)^2 - \mathbf{1}_N'\mathbf{V}\mathbf{1}_N]\sigma^2.$$

The bound is achieved if and only if $s \in B(\mathbf{X} : \mathbf{V})$, in which case

$$\hat{T}(\mathbf{X} : \mathbf{V}) = n^{-1}(\mathbf{1}_N'\mathbf{V}^{1/2}\mathbf{1}_N)(\mathbf{1}_s'\mathbf{V}_{ss}^{-1/2}\mathbf{Y}_s).$$

For a given variance structure $\mathbf{V}$, the theorem states that under mild conditions on the $\mathbf{X}$ matrix, the estimator $\hat{T}(\mathbf{X} : \mathbf{V})$ will have two important properties under root($\mathbf{V}$) balance: (i) it will be bias robust, since the best linear unbiased estimator for any *other* auxiliary matrix (satisfying the mild conditions) will equal $\hat{T}(\mathbf{X} : \mathbf{V})$; and (ii) it will be most efficient under $M(\mathbf{X} : \mathbf{V})$, since the variance is the smallest possible across samples.

By way of illustration, consider the case in which the variance is proportional to x. It has already been seen that the common ratio estimator, under standard balance, will be bias robust. Now consider the model $Y_i = \gamma_{1/2} x_i^{1/2} + \gamma x_i + x_i^{1/2}\varepsilon_i$. Call the best linear unbiased estimator based on this model $\hat{T}(\gamma_{1/2} x^{1/2} + \gamma x : x)$. The lower bound on its variance is

$$\left[\frac{\left(N\overline{x}^{(1/2)}\right)^2}{n} - \sum_{i=1}^{N} x_i\right]\sigma^2, \tag{12}$$

achieved in any sample balanced in the sense that $\overline{x}_s^{(1/2)} = \sum_{i=1}^{N} x_i / \sum_{i=1}^{N} x_i^{1/2}$. Bias protection against general polynomial models is obtained by balancing on additional powers:

$$\sum_s x_i^{j-1/2}/n = \sum_{i=1}^{N} x_i^j \Bigg/ \sum_{i=1}^{N} x_i^{1/2} \quad \text{for } j = 0, 1, 2, \ldots, J,$$

that is, under *root*(x) balance (of order J). In addition, by Theorem 2, $\hat{T}(\gamma_{1/2} x^{1/2} + \gamma x : x)$ has the minimum variance in every sample under the working model $Y = \gamma_{1/2} x_i^{1/2} + \gamma x_i + x_i^{1/2}\varepsilon_i$.

Note that (12) is less than the variance (9) of the ratio estimator under the balance condition $\overline{x}_s = \overline{x}$. The ratio estimator arises on the basis of a model with variance proportional to x, but is *not* bias robust under root(x) balance and does not yield the minimal variance, in contrast to $\hat{T}(\gamma_{1/2} x^{1/2} + \gamma x : x)$. When root($x$) balanced sampling is feasible, there can be no justification for *ever* using the ratio estimator, except when one is absolutely sure of a simple through the origin model.

A word on practical methods of achieving balanced samples is in order. Consider **sampling with probability proportionate to size** (*pps*), for which there are a variety of modes of implementation. If the variable according to which *pps* is carried out is $v_i^{1/2} = [\mathrm{var}_M(Y_i)]^{1/2}$, so that (for fixed sample size n) the inclusion probabilities are $\pi_i = n v_i^{1/2} / \sum_{i=1}^{N} v_i^{1/2}$, then the sample will be balanced *in (design) expectation*:

$$\mathrm{E}_{pps}\left(\frac{\sum_s v_i^{-1/2} x_{ji}}{n}\right) = \frac{\sum_{i=1}^{N} x_{ji}}{\sum_{i=1}^{N} v_i^{1/2}},$$

for $\mathbf{X} = (x_{ji})$.

This says that an appropriate *pps* scheme *aims* at balance. One reasonable approach to getting a balanced sample is to generate some (say 100) *pps* samples, and choose one among them that comes closest to meeting the criterion of balance. Herson [10] illustrated this type of restricted **randomization** when selecting simple random samples. There may be some trade-off between one column of $\mathbf{X}$ and another; in some cases, one may wish to go beyond the original set of samples.

It is to be noted that balance will not always be achievable. For example, in the case of root(x) balance, the expression (12) for the variance at balance implies that $n \leq [(\overline{x}^{(1/2)})^2/\overline{x}]N$, so that for large enough n balance is impossible. It can be shown that if the sample inclusion weights $\pi_i = n v_i^{1/2} / \sum_{i=1}^{N} v_i^{1/2}$ are less than *one* (so that the corresponding *pps* sampling is possible), then n is small enough for balance. Thus a general strategy for getting balance is first to weed out the units from the population for which π_i exceeds *one*, putting these into a separate certainty stratum, and balance on the remainder of the population.

Large Sample Normality

For any linear estimator $\hat{\theta} = \mathbf{g}_s' \mathbf{Y}_s$ of $\theta = \boldsymbol{\gamma}' \mathbf{Y}$, whether best linear unbiased or not, the estimation error $\hat{\theta} - \theta$ is a linear function of the elements of $\mathbf{Y}$. Thus, under a model for which $\hat{\theta}$ is prediction unbiased and the Ys are independent, the standardized error $(\hat{\theta} - \theta)/[\mathrm{var}_M(\hat{\theta} - \theta)]^{1/2}$ has an asymptotic standard **normal distribution** under some reasonable conditions [16]. When a **consistent** variance estimator $\hat{v}$ (see above) is substituted in the denominator of the standardized error, $(\hat{\theta} - \theta)/\sqrt{\hat{v}}$ will also be approximately normally distributed with mean 0 and variance 1 in large samples. An interval

of the form $\hat{\theta} \pm z_{\alpha/2}\sqrt{\hat{v}}$ will then have a confidence level of $(1-\alpha)\%$ asymptotically, with $z_{\alpha/2}$ being the $\alpha/2$-quantile from the standard normal distribution (*see* **Confidence Intervals and Sets**).

Clustered Populations

Many naturally occurring populations exhibit **clustering** in which units that are, in some sense, near each other have similar characteristics. Households in the same neighborhood may tend to have similar incomes, education levels of the heads of household, and amounts of expenditures on food and clothing. Business establishments in the same industry and geographic area will pay similar wages to a given occupation because of competition. This similarity among "nearby" units can express itself statistically as a **correlation** between the target variables for different units.

In clustered populations, the methods of data collection may also differ from the methods used in other populations. In a household survey, for example, a complete list of households to use for sampling is usually not available, especially if the population is large. In the US, for instance, there are nearly 100 million households. The households may be geographically dispersed so that fieldwork can be more economically done when sample units are clustered together to limit travel costs. A practical, and widely used, technique is to select the sample in stages, using, at each stage, sampling units for which a complete list is available. In the household example, geographic areas may be selected at the first stage. At the second stage, each first stage sample unit may be further subdivided and a sample of the subdivisions selected. A list of the households in each sample subdivision is then compiled and data collected from each. In a business population, establishments may be selected at the first stage, a list of occupations compiled in each sample establishment, and a sample of occupations then drawn from each list. Although occupations are the units ultimately sampled, a complete list of occupations for each establishment in the universe is unlikely to be available while a list of establishments often is. Selecting establishments at the first stage is also sensible because survey costs may depend on the number of sample establishments more than the number of sample occupations. Cooperation must be elicited at the establishment level; and the more establishments in the sample, the more the survey will cost.

An Intracluster Correlation Model for a Clustered Population

The population of units is divided into N clusters. Cluster i contains M_i units with the total number of units in the population being $M = \sum_{i=1}^{N} M_i$. We suppose that the clusters sizes M_i, and hence the population size M, are all known. Associated with unit j in cluster i is a random variable Y_{ij} the finite population total of which is $T = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}$. One simple working model is

$$\begin{aligned} \mathrm{E}_M(Y_{ij}) &= \mu, \\ \mathrm{cov}_M(Y_{ij}, Y_{i'j'}) &= \begin{cases} \sigma_i^2, & i = i',\ j = j', \\ \sigma_i^2 \rho_i, & i = i',\ j \neq j', \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (13)$$

The model posits that units all have a common mean μ. Within cluster i, units have a common variance σ_i^2, which can be different from one cluster to another. Units in the same cluster also have a common correlation ρ_i. This type of model can also be combined with **stratification** to describe some populations better.

Elements are selected by a two-stage sampling scheme. First, a sample s of n clusters is chosen from the N. Denote the set of nonsample clusters by r. Then, from the M_i elements in sample cluster i, a sample s_i of size m_i is selected. The total number of units in the sample is $m = \sum_s m_i$. The population total is then naturally represented in three parts – the total for the observed elements, the total for unobserved elements in sample clusters, and the total for nonsample clusters:

$$T = \sum_{i \in s} \sum_{j \in s_i} Y_{ij} + \sum_{i \in s} \sum_{j \notin s_i} Y_{ij} + \sum_{i \notin s} \sum_{j=1}^{M_i} Y_{ij}. \quad (14)$$

The optimal estimator of T under model (13) is

$$\hat{T}_{\mathrm{BLU}} = \sum_{s} \sum_{s_i} Y_{ij} + \sum_{s} (M_i - m_i)\big[w_i \overline{Y}_{si} + (1 - w_i)\hat{\mu}\big] + \sum_{r} M_i \hat{\mu}, \quad (15)$$

where $w_i = m_i\rho_i/[1+(m_i-1)\rho_i]$, $\overline{Y}_{si} = \sum_{s_i} Y_{ij}/m_i$, and $\hat{\mu}$ is a weighted average of the sample means, $\hat{\mu} = \sum_s u_i \overline{Y}_{si}$, with weights

$$u_i = \frac{m_i \Big/ \left\{\sigma_i^2\,[1+(m_i-1)\rho_i]\right\}}{\sum\limits_s m_i \Big/ \left\{\sigma_i^2\,[1+(m_i-1)\rho_i]\right\}}.$$

Note that the estimator of the nonsample total for each sample cluster in the second term of $\hat{T}_{\mathrm{BLU}}$ is a kind of **composite estimator**.

Because the parameters σ_i^2 and ρ_i must be known or estimated in order to compute $\hat{T}_{\mathrm{BLU}}$, the practical use of this estimator is limited. For that reason, estimators of the form $\hat{T} = (N/n)\sum_{i\in s}\lambda_i\hat{T}_i$ are often used where $\hat{T}_i = M_i\overline{Y}_{si}$ and λ_i is a constant. Note that these will be unbiased under (13) if $(N/n)\sum_{i\in s}\lambda_i M_i = M$. The estimator $\hat{T}_p = (M/n)\sum_s \overline{Y}_{si}$ is in this class, for example (it is also the **Horvitz–Thompson estimator** under a plan in which clusters are selected with probabilities proportional to the cluster sizes M_i and an equal probability sample is selected within each cluster.)

Estimators can also be used for clustered populations that make use of a variety of auxiliary data, in cases in which $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ is appropriate. In many situations in which cluster models and multistage sampling are used, summary auxiliary data on a population may be available even though individual data for all units in the population may not. In a human population, for example, census counts of the number of persons by age, ethnic group, and sex may exist from a recent population census. If those variables are also related to the targets of a survey, regression estimation may be quite useful. The incidence and **prevalence** of some health conditions may depend on demographic characteristics, such as age, ethnic group, and sex, and on geographic place of residence, for instance. When $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, the predictors $\hat{T}_1 = \mathbf{1}_N'\mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{T}_2 = \sum_s Y_i + \mathbf{1}_{N-n}'\mathbf{X}_r\hat{\boldsymbol{\beta}}$, with $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{Y}_s$, are both prediction unbiased. In addition, the vectors $\mathbf{1}_N'\mathbf{X}$ and $\mathbf{1}_{N-n}'\mathbf{X}_r$ are the totals of the auxiliaries for the full population and the nonsample, respectively. If $\mathbf{Y}$ is set equal to $\mathbf{X}$, then $\hat{T}_1$ and $\hat{T}_2$ reproduce these totals – a feature sometimes known as **calibration** in finite population sampling [6] Although generally suboptimal, since they do not account for the covariance structure $\mathrm{var}(\mathbf{Y}) = \mathbf{V}$, $\hat{T}_1$ and $\hat{T}_2$ are practical choices because they do incorporate important auxiliary information and because $\mathbf{V}$ may be unknown and difficult to estimate.

Other Topics

Beyond what has been discussed here, superpopulation models have found many other areas of application in survey sampling. Much of biostatistical data on human populations is qualitative; for example, whether or not a person has a health condition, whether or not a certain medication is used, whether the average daily intake of calories is above a given level, and so on. Natural models for such variables include the **logistic** and other choices for **binary** responses that are often used in analysis of biostatistical data. These models can also be applied when estimating finite population totals [24]. Estimation of cumulative distribution functions and quantiles such as the **median** and first or third quartiles, may also be of interest in studies in which simple descriptive statistics alone do not suffice [2, 7, 8]. More complex, **multivariate analyses** of survey data also call for the use of superpopulation models [22]. A general reference on prediction-based survey sampling is [26].

References

[1] Basu, D. (1971). An essay on the logical foundations of survey sampling, in *Foundations of Statistical Inference*, V.P. Godambe, & D.A. Sprott, eds. Holt, Rinehart, & Winston, Toronto, pp. 203–242.

[2] Chambers, R.L. & Dunstan, R. (1986). Estimating distribution functions from survey data, *Biometrika* **73**, 597–604.

[3] Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes, *Journal of the American Statistical Association* **37**, 199–212.

[4] Cumberland, W.G. & Royall, R.M. (1981). Prediction models and unequal probability sampling, *Journal of the Royal Statistical Society, Series B* **43**, 353–367.

[5] Deming, W.E. (1952). *Some Theory of Sampling*. Wiley, New York.

[6] Deville, J.C. & Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association* **87**, 376–382.

[7] Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function, *Australian Journal of Statistics* **35**, 29–41.

[8] Dorfman, A.H. & Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression, *Annals of Statistics* **21**, 1452–1475.

[9] Hansen, M.H., Madow, W.G. & Tepping, B.J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys (with discussion), *Journal of the American Statistical Association* **78**, 776–793.

[10] Herson, J. (1976). An investigation of relative efficiency of least-squares prediction to conventional probability sampling plans, *Journal of the American Statistical Association* **71**, 700–703.

[11] Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling, *Journal of the American Statistical Association* **71**, 657–664.

[12] Royall, R.M. (1976). Current advances in sampling theory: implications for human observational studies, *American Journal of Epidemiology* **104**, 463–473.

[13] Royall, R.M. (1986). The prediction approach to robust variance estimation in two-stage cluster sampling, *Journal of the American Statistical Association* **81**, 119–123.

[14] Royall, R.M. (1988). The prediction approach to sampling theory, in *Handbook of Statistics*, Vol. 6, P.R. Krishaniah & C.R. Rao, eds. Elsevier, Amsterdam, pp. 399–413.

[15] Royall, R.M. (1992). Robustness and optimal design under prediction models for finite populations, *Survey Methodology* **18**, 179–185.

[16] Royall, R.M. & Cumberland, W.G. (1978). Variance estimation in finite population sampling, *Journal of the American Statistical Association* **73**, 351–358.

[17] Royall, R.M. & Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance (with discussion), *Journal of the American Statistical Association* **76**, 66–77.

[18] Royall, R.M. & Cumberland, W.G. (1981). The finite population linear regression estimator and estimators of its variance – an empirical study, *Journal of the American Statistical Association* **76**, 924–930.

[19] Royall, R.M. & Cumberland, W.G. (1988). Does simple random sampling provide adequate balance?, *Journal of the Royal Statistical Society, Series B* **50**, 118–124.

[20] Royall, R.M. & Herson, J. (1973). Robust estimation in finite populations I, *Journal of the American Statistical Association* **68**, 880–889.

[21] Särndal, C.E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[22] Skinner, C.J., Holt, D. & Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Wiley, New York.

[23] Smith, T.M.F. (1976). The foundations of survey sampling: a review (with discussion), *Journal of the Royal Statistical Society, Series A* **139**, 183–195.

[24] Valliant, R. (1985). Nonlinear prediction theory and the estimation of proportions in a finite population, *Journal of the American Statistical Association* **80**, 631–641.

[25] Valliant, R. (1987). Conditional properties of some estimators in stratified sampling, *Journal of the American Statistical Association* **82**, 509–519.

[26] Valliant, R., Dorfman, A.H., & Royall, R.M. (2000). *Finite Population Sampling and Inference: A prediction Approach*, Wiley, New York.

[27] Whittle, P. (1963). *Prediction and Regulation by Linear Least Squares*. The English Universities Press, London.

ALAN H. DORFMAN & RICHARD VALLIANT

Support Vector Machines

Introduction

Over the past 10 years, kernel methods such as Support Vector Machines and Gaussian Processes have become a staple for modern statistical estimation and machine learning. The groundwork for this field was laid in the second half of the twentieth century by Vapnik and Chervonenkis (geometrical formulation of an optimal separating hyperplane, capacity measures for margin classifiers), Mangasarian (linear separation by a convex function class), Aronszajn (Reproducing Kernel Hilbert Spaces), Aizerman, Braverman, and Rozonoér (nonlinearity via kernel feature spaces), Arsenin and Tikhonov (regularization and ill-posed problems), and Wahba (regularization in Reproducing Kernel Hilbert Spaces).

However, it took until the early 1990s when positive definite kernels became a popular and viable means of estimation. Firstly, this slow uptake was due to the lack of sufficiently powerful hardware, since kernel methods require the computation of the so-called kernel matrix, which requires quadratic storage in the number of data points (a computer of at least a few megabytes of memory is required to deal with 1000+ points). Secondly, many of the previously mentioned techniques lay dormant or existed independently and only recently the (in hindsight obvious) connections were made to turn this into a practical estimation tool. Nowadays, a variety of good reference books exist and anyone serious about dealing with kernel methods is recommended to consult one of the following works for further information [5, 12, 8, 15]. Below, we will summarize the main ideas of kernel method and support vector machines, building on the summary given in [13].

Learning from Data

One of the fundamental problems of learning theory (*see* **Machine Learning**) is the following: Suppose we are given two classes of objects. We are then faced with a new object, and we have to assign it to one of the two classes. This problem, referred to as *(binary)* **pattern recognition**, can be formalized as follows: we are given empirical data of m pairs

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}, \qquad (1)$$

and we want to estimate a *decision function* $f: \mathcal{X} \to \{\pm 1\}$. Here, $\mathcal{X}$ is some nonempty set from which the *patterns* x_i are taken, usually referred to as the *domain*; the y_i are called *labels* or *targets*. A good decision function will have the property that it *generalizes* to unseen data points, achieving a small value of the *risk*

$$R[f] = \int \frac{1}{2}|f(x) - y| \; \mathrm{d}P(x, y). \qquad (2)$$

(*see* **Decision Theory**). In other words, on average over an unknown distribution P that is assumed to generate both training and test data, we would like to have a small error. Here, the error is measured by means of the *zero–one loss function* $c(x, y, f(x)) := 1/2|f(x) - y|$. The loss is 0 if (x, y) is classified correctly, and 1 otherwise.

It should be emphasized that so far, the patterns could be just about anything, and we have made no assumptions on $\mathcal{X}$ other than it being a set endowed with a probability measure P (note that the labels y may, but need not, depend on x in a deterministic fashion). Moreover, (2) does not tell us how to *find* a function with a small risk. In fact, it does not even tell us how to *evaluate* the risk of a given function, since the probability measure P is assumed to be unknown.

We therefore introduce an additional type of structure, pertaining to what we are actually given – the training data. Loosely speaking, to generalize, we want to choose a fitted value $f(x)$ such that $(x, f(x))$ is in some sense similar to the training examples (1), for example, that $|y - f(x)|$ is small. To this end, we need notions of *similarity* in $\mathcal{X}$ and in $\{\pm 1\}$. Characterizing the similarity of the outputs $\{\pm 1\}$ is easy: in binary classification, only two situations can occur: two labels can either be identical or different. The choice of the similarity measure for the inputs, that is, x, on the other hand, is a deep question that lies at the core of the problem of machine learning.

One of the advantages of kernel methods is that the learning **algorithms** developed are quite independent of the choice of the similarity measure (*see* **Similarity, Dissimilarity, and Distance Measure**). This allows us to adapt the latter to the specific problems at hand without the need to reformulate the learning algorithm itself.

Kernels

Let us consider a symmetric similarity measure of the form

$$k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \text{ where } (x, x') \mapsto k(x, x'),$$

that is, a function that, given two patterns x and x', returns a real number characterizing their similarity. The function k is often called a *kernel*.

Kernels as Similarity Measures

General similarity measures of this form are rather difficult to study. Let us therefore start from a particularly simple case, the *dot product* $\langle \mathbf{x}, \mathbf{x}' \rangle$, and generalize it subsequently.

The geometric interpretation of the canonical dot product is that it computes the cosine of the angle between the vectors $\mathbf{x}$ and $\mathbf{x}'$, provided they are normalized to length 1. Moreover, it allows computation of the *length* (or *norm*) of a vector $\mathbf{x}$ as

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \tag{3}$$

Being able to compute dot products amounts to being able to carry out all geometric constructions that can be formulated in terms of angles, lengths and distances. However, this is not really sufficiently general to deal with many interesting problems.

- First, we have deliberately not made the assumption that the patterns actually exist in a dot product space (they could be any kind of object). We therefore first need to represent the patterns as vectors in some dot product space $\mathcal{H}$, called the *feature space* using a map

$$\Phi: \mathcal{X} \to \mathcal{H} \text{ where } x \mapsto \mathbf{x} := \Phi(x). \tag{4}$$

  Note that we use a boldface $\mathbf{x}$ to denote the vectorial representation of x in the feature space.
- Second, even if the original patterns lie in a dot product space, we may still want to consider more general similarity measures obtained by applying the map (4).

Embedding the data into $\mathcal{H}$ via Φ has two main benefits. First, it allows us to deal with the patterns geometrically, and thus lets us study learning algorithms using linear algebra and analytic geometry. Second, it lets us define a similarity measure from the dot product in $\mathcal{H}$,

$$k(x, x') := \langle \mathbf{x}, \mathbf{x}' \rangle = \langle \Phi(x), \Phi(x') \rangle. \tag{5}$$

The freedom to choose the mapping Φ enables us to design a large variety of similarity measures and learning algorithms.

Examples of Kernels

So far, we have used the kernel notation as an abstract similarity measure. We now give some concrete examples of kernels, mainly for the case where the inputs x_i are already taken from a dot product space. The role of the kernel then is to implicitly change the representation of the data into another (usually higher dimensional) feature space. One of the most common kernels used is the polynomial one,

$$k(x, x') = \langle x, x' \rangle^d, \text{ where } d \in \mathbb{N}. \tag{6}$$

It corresponds to a feature space spanned by *all* products of order d of input variables, that is, all products of the form $[x]_{i_1} \cdots [x]_{i_d}$. Hence, the dimension of this space is $O(N^d)$, but since we are using the kernel to evaluate dot products, this does not affect us. Another popular choice is the Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\,\sigma^2}\right), \tag{7}$$

with a suitable width $\sigma > 0$.

Examples of more sophisticated kernels, defined not on dot product spaces but on discrete objects such as strings, are the string matching kernels proposed in [16] and [7].

In general, there are several ways of deciding whether a given function k qualifies as a valid kernel. One way is to appeal to *Mercer's theorem*. This classical result of functional analysis states that the kernel of a positive definite integral operator can be diagonalized in terms of an **eigenvector** expansion with nonnegative **eigenvalues**. From the expansion, the feature map Φ can explicitly be constructed. Another approach exploits the fact that k is the kernel of a *Reproducing Kernel Hilbert Space*; see [12] for references and details.

Support Vector Classifiers

Statistical Learning Theory shows that it is imperative to restrict the set of functions from which f is chosen to one that has a *capacity* suitable for the amount of available training data. It provides *bounds* on the test error, depending on both the empirical risk and the capacity of the function class. The minimization of these bounds leads to the principle of *structural risk minimization* [15].

Support Vector Machines (SVM) can be considered an approximate implementation of this principle, by trying to minimize a combination of the *training error* (or *empirical risk*),

$$R_{\mathrm{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|f(x_i) - y_i|, \tag{8}$$

and a capacity term derived for the class of hyperplanes in a dot product space $\mathcal{H}$ [15],

$$\langle \mathbf{w}, \mathbf{x}\rangle + b = 0 \text{ where } \mathbf{w} \in \mathcal{H}, b \in \mathbb{R}, \tag{9}$$

corresponding to decision functions

$$f(\mathbf{x}) = \mathrm{sgn}\,(\langle \mathbf{w}, \mathbf{x}\rangle + b). \tag{10}$$

Hard Margin Solution

Consider first problems, which are linearly separable. There exists a unique *optimal hyperplane* [15], distinguished by the maximum margin of separation between any training point and the hyperplane. It is the solution of

$$\underset{\mathbf{w}\in\mathcal{H}, b\in\mathbb{R}}{\text{maximize}}\ \min\{\|\mathbf{x} - \mathbf{x}_i\|\ |\mathbf{x} \in \mathcal{H}, \langle \mathbf{w}, \mathbf{x}\rangle + b = 0, i = 1, \ldots, m\}. \tag{11}$$

Moreover, the capacity of the class of separating hyperplanes can be shown to decrease with increasing margin. The latter is the basis of the *statistical* justification of the approach; in addition, it is *computationally* attractive, since we will show below that it can be constructed by solving a quadratic programming problem for which efficient algorithms exist.

As one can see from the example given in Figure 1, in order to construct the optimal hyperplane, we need to solve

$$\underset{\mathbf{w}\in\mathcal{H}, b\in\mathbb{R}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 \text{ for all } i = 1, \ldots, m. \tag{12}$$

Note that the constraints ensure that $f(\mathbf{x}_i)$ will be $+1$ for $y_i = +1$, and -1 for $y_i = -1$. (One might argue that for this to be the case, we do not actually need the constraint "≥ 1". However, without it, it would not be meaningful to minimize the length of $\mathbf{w}$: to see this, imagine we wrote ">0" instead of "≥ 1." Now assume that the solution is $(\mathbf{w}, b)$. Let us rescale this solution by multiplication with some $0 < \lambda < 1$.

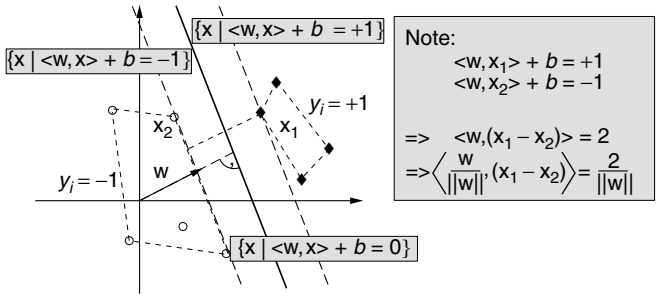

Figure 1 A binary classification toy problem: separate balls from diamonds. The *optimal hyperplane* is shown as a solid line. The problem being separable, there exists a weight vector $\mathbf{w}$ and a threshold b such that $y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) > 0$ $(i = 1, \ldots, m)$. Rescaling $\mathbf{w}$ and b such that the point(s) closest to the hyperplane satisfy $|\langle \mathbf{w}, \mathbf{x}_i\rangle + b| = 1$, we obtain a *canonical* form $(\mathbf{w}, b)$ of the hyperplane, satisfying $y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1$. Note that in this case, the *margin* (the distance of the closest point to the hyperplane) equals $1/\|\mathbf{w}\|$. This can be seen by considering two points $\mathbf{x}_1, \mathbf{x}_2$ on opposite sides of the margin, that is, $\langle \mathbf{w}, \mathbf{x}_1\rangle + b = 1$, $\langle \mathbf{w}, \mathbf{x}_2\rangle + b = -1$, and projecting them onto the hyperplane normal vector $\mathbf{w}/\|\mathbf{w}\|$ (from [12])

Since $\lambda > 0$, the constraints are still satisfied. Since $\lambda < 1$, however, the length of $\mathbf{w}$ has decreased. Hence $(\mathbf{w}, b)$ cannot be the minimizer of (12).)

The constrained **optimization** problem (12) is dealt with by introducing *Lagrange multipliers* $\alpha_i \geq 0$ ($\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_m)$) and a *Lagrangian*

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - 1\right). \tag{13}$$

L has a *saddle point* in $\mathbf{w}, b$, and $\boldsymbol{\alpha}$ at the optimal solution of the primal optimization problem. This means that it should be minimized with respect to the *primal variables* $\mathbf{w}$ and b and maximized with respect to the *dual variables* α_i. Furthermore, the product between constraints and Lagrange multipliers in L vanish at optimality, that is,

$$\alpha_i(y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - 1) = 0 \text{ for all } i = 1, \ldots, m. \tag{14}$$

To minimize w.r.t. the primal variables, we require

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = -\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{15}$$

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \tag{16}$$

The solution thus has an expansion (16) in terms of a subset of the training patterns, namely those patterns with nonzero α_i, called *Support Vectors (SVs)*. Often, only few of the training examples actually end up being SVs.

By the *Karush–Kuhn–Tucker conditions* (14) known from optimization theory, the SVs lie on the margin (cf. Figure 1) – this can be exploited to compute b once the α_i have been found. All remaining training examples $(\mathbf{x}_j, y_j)$ are irrelevant: their constraint $y_j(\langle \mathbf{w}, \mathbf{x}_j \rangle + b) \geq 1$ could just as well be left out. In other words, the hyperplane is completely determined by the patterns closest to it.

By substituting (15) and (16) into the Lagrangian (13), one eliminates the primal variables $\mathbf{w}$ and b, arriving at the so-called *dual optimization problem*, which is the problem usually solved in practice:

$$\begin{aligned} \underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\text{maximize}} \; & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K_{ij} \\ \text{subject to } & \alpha_i \geq 0 \text{ for all } i = 1, \ldots, m \\ \text{and } & \sum_{i=1}^{m} \alpha_i y_i = 0. \end{aligned} \tag{17}$$

where $K_{ij} := \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Using (16), the decision function (10) can thus be written as

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b\right), \tag{18}$$

where b is computed via (14); for details, see [5, 8, 12, 15].

The Kernel Trick

We now have all the tools to describe SVMs. Everything above was formulated in a dot product space, which we think of as the feature space $\mathcal{H}$ (see (4)). To express the formulae in terms of the input patterns in $\mathcal{X}$, we employ (5) and replace $\langle \mathbf{x}, \mathbf{x}' \rangle$ by $k(\mathbf{x}, \mathbf{x}')$ wherever it occurs. This substitution, which is sometimes referred to as the *kernel trick*, was used by Boser et al.. [3] to develop nonlinear SVMs. Now f can be rewritten as

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b\right). \tag{19}$$

Furthermore, in the quadratic program (17) the definition of K_{ij} becomes $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Figure 2 shows a toy example.

Soft Margin Solution

In practice, a separating hyperplane may not exist, for example, if a high noise level causes a large overlap of the classes. To accommodate this case, one introduces slack variables $\xi_i \geq 0$ for all $i = 1, \ldots, m$ in order to relax the constraints of (12) to

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \text{ for all } i = 1, \ldots, m. \tag{20}$$

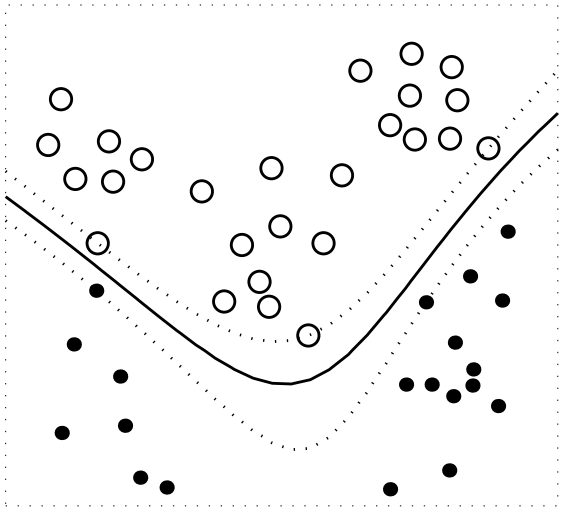

Figure 2 Example of an SV classifier found using a radial basis function kernel $k(x, x') = \exp(-\|x - x'\|^2)$. Circles and points are two classes of training examples; the middle line is the decision surface; the outer lines precisely meet the constraint of (12). Note that the SVs found by the algorithm (sitting on the dotted constraint lines) are not centers of clusters, but examples which are critical for the given classification task (from [13])

A classifier that generalizes well is then found by controlling both the classifier capacity (via $\|\mathbf{w}\|$) and the sum of the slacks $\sum_i \xi_i$. The latter can be shown to provide an upper bound on the number of training errors.

One possible realization of such a *soft margin* classifier is obtained by minimizing the objective function

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \tag{21}$$

subject to the constraints on ξ_i and (20), where the constant $C > 0$ determines the trade off between margin maximization and training error minimization. This again leads to the problem of maximizing (17), subject to modified constraints where the only difference from the separable case is an upper bound C on the Lagrange multipliers α_i.

Another realization uses the more natural ν-parameterization. In it, the parameter C is replaced by a parameter $\nu \in (0, 1]$, which can be shown to provide lower and upper bounds for the fraction of examples that will be SVs, and those that will have nonzero slack variables, respectively.

Its dual can be shown to consist in maximizing the quadratic part of (17), subject to

$$0 \le \alpha_i \le \frac{1}{\nu m}, \quad \sum_i \alpha_i y_i = 0, \text{ and } \sum_i \alpha_i = 1. \tag{22}$$

Discussion

Extensions

The applicability of the "kernel trick" extends significantly beyond the classification setting and in recent years a large number of kernel algorithms have been proposed to solve as diverse tasks as the estimation of the support (or, more generally, quantiles) of a distribution, of a **regression** function, or of a nonlinear manifold. Below, we give a brief overview of the most popular methods:

- Regression: Just as **classification** can be formulated as a quadratic optimization problem, so can regression. Here, the maximum margin condition is replaced by the requirement of finding the *flattest* function, which performs a regression within ε deviation from the observations.
- Principal Component Analysis: It can be extended to nonlinear settings by replacing PCA in input space by a feature space representation. The final algorithm consists of solving an eigenvector problem for the kernel matrix.
  Similar modifications can be carried out to obtain nonlinear versions of projection pursuit, for example, via *sparse kernel feature analysis*.
- Independent Component Analysis: Recently, an algorithm was suggested in [2] to find independent components via a modification of **canonical correlation** analysis. This is currently an active topic of research and it is likely to lead to novel criteria for factorizing distributions.
- Quantiles of a Distribution: In this problem, one attempts to find sets such that the probability of data occurring outside this set is controlled. This is done by ensuring that the set contains a certain fraction of the training data while at the same time keeping the set "simple" (where simplicity is determined by an SVM-style regularization term). This can be done

also for high-dimensional problems, and one can show that it can be cast as a classification problem with only one class. Kernel extensions exist.

- Estimation of Manifolds: Here one aims at finding smooth manifolds which approximate a dataset (i.e. manifolds for which the error incurred by projection of the data onto the manifold is small). Again, one can find optimization problems similar to the SV optimization problem (i.e. a regularization term plus a misprediction cost) and generate a kernel expansion.

These and many more kernel methods plus the corresponding references can be found in [12].

Implementations

An initial weakness of SVMs was that the size of the quadratic programming problem scaled with the number of SVs. This was due to the fact that in (17), the quadratic part contained at least all SVs – the common practice was to extract the SVs by going through the training data in chunks while regularly testing for the possibility that patterns initially not identified as SVs become SVs at a later stage. This procedure is referred to as *chunking*; note that without chunking, the size of the **matrix** in the quadratic part of the objective function would be $m \times m$, where m is the number of all training examples.

What happens if we have a high-noise problem? In this case, many of the slack variables ξ_i become nonzero, and all the corresponding examples become SVs. For this case, decomposition algorithms were proposed on the basis of the observation that not only can we leave out the non-SV examples (the x_i with $\alpha_i = 0$) from the current chunk, but also some of the SVs, especially those that hit the upper boundary ($\alpha_i = C$). The chunks are usually dealt with using quadratic optimizers. Several public domain SV packages and optimizers are listed on `http://www.kernel-machines.org`.

Empirical Results and Applications

Modern SVM implementations made it possible to train on some rather large problems. Success stories include the 60 000 example MNIST digit recognition benchmark (with record results), as well as problems in text categorization and **bioinformatics**, where two main areas of application are worth mentioning:

Firstly, there are classification and gene selection problems in DNA microarray analysis (*see* **Bioinformatics in Functional Genomics**). Given the high dimensionality of the data to begin with, the use of kernels is not advisable in this case. Instead, a linear classifier with a suitable penalty on the expansion coefficients favoring sparse expansions is found; see [4, 6] for further details and references. Finding suitable variable selection criteria is an active area of research (see e.g. [1], which points out substantial problems with the approach taken in [6], mainly due to improper testing). Reference [11] contains further empirical results on SVM performance.

Secondly, sequence analysis can often be cast into the form of a classification problem, requiring the design of custom tailored kernels for this purpose. Such research has led to excellent results (see [9, 7, 17, 16, 14, 10] and the references therein for further details).

Conclusion

During the last few years, SVMs and other kernel methods have rapidly advanced into the standard toolkit of techniques for machine learning and high-dimensional data analysis. This was probably due to a number of advantages compared to **neural networks**, such as the absence of spurious local minima in the optimization procedure, the fact that there are only few parameters to tune, enabling fast deployment in applications, the modularity in the design, where various kernels can be combined with a number of different learning algorithms, and the excellent performance on high-dimensional data.

References

[1] Ambroise, C. & McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data, *Proceedings of the National Academy of Sciences of the United States of America* **99**(10), 6562–6566.

[2] Bach, F.R. & Jordan, M.I. (2002). Kernel independent component analysis, *Journal of Machine Learning Research* **3**, 1–48.

[3] Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers, in *Proceedings of the Annual Conference on Computational*

Learning Theory, D. Haussler, ed. ACM Press, Pittsburgh, pp. 144–152, July 1992.

[4] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C., Furey, T.S., Ares, M. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data using support vector machines, *Proceedings of the National Academy of Sciences of the United States of America* **97**(1), 262–267.

[5] Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

[6] Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines, *Machine Learning* **46**, 389–422.

[7] Haussler, D. (1999). *Convolutional Kernels on Discrete Structures*, Technical Report UCSC-CRL-99-10, Computer Science Department, UC Santa Cruz.

[8] Herbrich, R. (2002). *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Cambridge, MA.

[9] Jaakkola, T.S. & Haussler, D. (1999). Exploiting generative models in discriminative classifiers, in *Advances in Neural Information Processing Systems 11*, M.S. Kearns, S.A. Solla & D.A. Cohn, eds. MIT Press, Cambridge, pp. 487–493.

[10] Leslie, C., Eskin, E. & Noble, W.S. (2002). The spectrum kernel: A string kernel for SVM protein classification, in *Proceedings of the Pacific Symposium on Biocomputing*, Lihue, Hawaii, pp. 564–575.

[11] Meyer, D., Leisch, F. & Hornik, K. (2003). The support vector machine under test, *Neurocomputing* **55**, 169–186.

[12] Schölkopf, B. & Smola, A.J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

[13] Schölkopf, B. & Smola, A.J. (2003). Support vector machines, in *The Handbook of Brain Theory and Neural Networks*, 2nd Ed., M.A. Arbib, ed. MIT Press, Cambridge, MA, pp. 1119–1125.

[14] Tsuda, K., Kin, T. & Asai, K. (2002). Marginalized kernels for biological sequences, *Bioinformatics* **18**,(Suppl. 2), S268–S275.

[15] Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York.

[16] Watkins, C. (2000). Dynamic alignment kernels, in *Advances in Large Margin Classifiers*, A.J. Smola, P.L. Bartlett, B. Schölkopf & D. Schuurmans, eds. MIT Press, Cambridge, pp. 39–50.

[17] Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T. & Müller, K.-R. (2000). Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics* **16**(9), 799–807.

BERNHARD SCHÖLKOPF & ALEX SMOLA

Support Vector Machines

BERNHARD SCHÖLKOPF & ALEX SMOLA

Volume 8, pp. 5328–5335

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Surgery

Surgeons have long been on the receiving end of criticism that they have a rather casual approach to the evaluation of new interventions. Pocock [18] discusses the problem from an objective perspective in the context of a discussion of the key role of **randomization**, but Horton [14] puts forward a more emotive argument in an editorial entitled "Surgical research or comic opera: questions but few answers". However, one approaches the issue, there is overwhelming evidence that few surgical procedures have been evaluated using randomized controlled trials [19] (*see* **Clinical Trials, Overview**), and the trials which have been conducted are often woefully inadequate methodologically [12, 17].

The history of surgery is littered with examples of procedures which were described and then widely adopted, only to be dropped when subjected to evaluation. The most recent example of a surgical "bandwagon" is laparoscopic surgery, and laparoscopic cholecystectomy in particular. This example is discussed in detail below.

What is unfair in much of the criticism of the scarcity of landmark trials in surgery is the implication that surgeons are unaware of the problem. Controlled trials in surgery are inherently more complex than drug trials, and numerous articles have been written by surgeons discussing the obstacles to performing trials in surgery. It has also been argued that the randomized clinical trial is not the only tool which can provide useful information on the efficacy of a surgical procedure, and surgical audit in particular is discussed later in this article.

The Necessity of Clinical Trials

Hierarchy of Evidence

Virtually all statisticians, and the large majority of clinicians, would agree that the randomized controlled trial is the **"gold standard"** for the evaluation of medical (and surgical!) interventions. However, it should be recognized that trials are not beyond criticism. In rather emotive terms, an editorial in the *Lancet* [1], which was discussing a proposed trial where one treatment arm involved hysterectomy, asked how we could convince a woman "that it is necessary to sacrifice her womb on the altar of science". The same article argues that trials are unnecessary in situations where "it stands to reason that the new procedure is indeed less risky". One could accept that argument if only there were not so many well-documented examples of procedures which were "obvious" advances until they were actually evaluated. One thinks, for example, of "gastric freezing", internal mammary artery ligation (see below), or extracranial–intracranial arterial anastomosis. This last procedure was widely practiced as a measure to reduce the risk of stroke, until a landmark trial demonstrated that the procedure was ineffective and possibly even harmful [8].

It is worth rehearsing the different sources of evidence on the efficacy of procedures, which were ranked by Chalmers [3] in terms of their credibility:

1. *Clinical impressions*. These should have no place in any scientific evaluation, but they do seem to be a basis for much of what is practiced in medicine.
2. *Case reports and uncontrolled case series*. Around 50% of the articles in leading surgical journals consist of **case series** [14], but their role in evaluation is severely limited. They are subject to **selection bias** and many other potential biases, including, even for surgery, a placebo effect. Their main role should be in refining procedures, and as an aid in screening potentially useful interventions prior to more formal evaluation.
3. *Case series with nonrandomized controls*. Such studies have more potential to provide useful information, but they are still prone to many sources of **bias**. For example, an experimental procedure might be introduced first with "low risk" patients, where the prognosis is atypically favorable. Alternatively, a new procedure, such as laparoscopic cholecystectomy, can be adopted so enthusiastically as to distort the indications for the procedure, so that on average a much fitter group of patients are undergoing the procedure. Comparing outcomes with historical controls then becomes a self-fulfilling prophesy because, of course, results improve when a more fit patient population is treated (*see* **Bias from Historical Controls**).
4. *Randomized controlled trial*. This is the only methodology which can control for the effects of potential biases in selection and evaluation. Any

new pharmacological agent needs to be tested to this level of rigor before it can be approved for marketing, but as yet there are no mechanisms which require "licensing" of surgical procedures.

Case Study – Internal Mammary Artery Ligation

The history of this procedure gives enormous insight into the role of trials in surgery, and also serves to highlight many of the associated problems. A very detailed history can be found in a chapter by Barsamian [2]. It was first proposed in the late nineteenth century that a heart which was receiving an inadequate blood supply via the coronary arteries could have the blood supply augmented through collateral circulation. Based on a number of animal studies and one anecdotal case report, a number of surgeons in the mid 1950s adopted an operation in which the internal mammary artery was ligated (tied off). This was used to treat patients with angina, on the principle that the operation ought to increase the blood supply to the heart. Many patients so treated experienced marked symptomatic improvement, and the operation grew in popularity. However, further animal studies failed to demonstrate any measurable effects on the blood supply, and in an uncontrolled study where 24 patients with angina were told in advance that the operation was experimental and with no physiological basis, it was reported that their response to the treatment was far less impressive. In all but four patients the anginal symptoms returned to their preoperative levels after a brief period of improvement. This suggested strongly that the response to the operation was determined by the patient's expectations, and this was reinforced by a report of two patients whose symptoms improved after an untied ligature was placed around the internal mammary arteries. When the ligatures were subsequently tied, the patients experienced no further improvement in their symptoms.

The issue was finally settled when almost simultaneously two controlled trials were reported. The studies were small (18 and 17 subjects, respectively), but they were randomized and well controlled. Indeed, the control procedure was a sham operation. All patients had their internal mammary arteries exposed, and then according to a random allocation the arteries were ligated or not, and the wound was closed. The results of both studies showed a clear improvement in symptoms in the control group and the ligated group, with no statistically significant difference between the response rates. In spite of the very small numbers, these two controlled trials led very quickly to the operation being abandoned.

This story has a number of profound implications. First, the studies showed unequivocally that surgery can have a placebo effect. This in turn throws severe doubts onto any procedure which is evaluated solely on the basis of an uncontrolled case series. Secondly, there are major ethical difficulties in surgical trials (*see* **Ethics of Randomized Trials**). There does not appear to be any other published clinical trial which has used sham operations as a control, and indeed by current standards of research ethics it is difficult to see that such a study would be regarded as ethical. Thirdly, it shows how quickly a new procedure can become popularized, based on a plausible argument for a mechanism and the most slender of anecdotal evidence.

Obstacles to Clinical Trials in Surgery

As previously mentioned, several authors have described various obstacles to performing clinical trials in surgery. Not all of these reasons are necessarily valid, but their perception as barriers has certainly dissuaded many surgeons from embarking on a clinical trial.

Blinding

The rationale for the controlled clinical trial is that it gives a mechanism which can control bias. The two key concepts which combine to achieve this are randomization and **blinding**. It is therefore a clear obstacle to surgical trials that rather obviously the operating surgeon cannot be blind to the procedure being undertaken! However, it is still possible to achieve the benefits of blinding provided (i) the allocation procedure is blinded in that the individual explaining the study to the patient, assessing the eligibility of the patient, and obtaining informed consent is unaware of the treatment which will be assigned should the patient be enrolled, and (ii) the individual assessing the response to treatment is also unaware of which treatment was allocated. There is no excuse for failing to blind the allocation procedure, and that, incidentally, is the main flaw in allocation schemes

based on apparently "random" mechanisms such as the final digit of a patient's hospital number, or the parity of the date of enrollment (*see* **Randomized Treatment Assignment**).

Blinding of the assessment of response to treatment will not always be straightforward, but ingenuity can help. Majeed et al. [16] describe a randomized comparison of a laparoscopic procedure with an open procedure, where the patients' "wounds" were dressed in an identical fashion, complete with masking bloodstains, irrespective of which procedure was performed. The impact of any loss of blinding on the assessment can also be minimized by the use of objective **outcome measures**, but even here care must be exercised. For example, in a trial of the management of severe head injuries, one might take death within seven days of injury as an end point. Death is a relatively objective end point, but in this context the precise time of death is essentially a function of when it is decided to begin the procedure for assessing brain stem death. It is conceivable that knowledge of the allocated treatment might induce a bias where deaths were delayed until after the 7-day threshold.

Consent

It can easily be seen that, in general, patients will often be unwilling to consent to a trial which involves a random choice between two surgical procedures, or between a surgical procedure and medical management. In a drug trial the effects will usually be reversible, and if a patient receives a drug which results in an undesirable side-effect, then the drug can be discontinued. Similarly, if the patient with a chronic condition is allocated to the drug which is subsequently found inferior, then at the end of the trial they can be switched to the preferred treatment. However, surgical procedures tend to be more permanent, and patients are very likely to have a preference for one or other treatment. For example, in a trial of breast-conserving surgery vs. mastectomy for early breast cancer, it is clear that some women would have a strong preference for mastectomy, to be sure that "all of the cancer is removed". Similarly, other women would have a strong preference for the less mutilating surgery because of concerns for their body image. Both opinions are perfectly valid, and it would be totally unethical to pressure a woman with a strong preference for either treatment to enter the trial. Thus, recruitment rates are likely to be slower than for trials of medical treatments.

It has been argued that the very procedure of seeking consent and implicitly or explicitly admitting uncertainty over what is the most appropriate treatment can undermine the patient's confidence in the surgeon. However, this is an argument which should be countered, since the patient–doctor relationship is totally dependent on each being truthful with the other.

One helpful approach to the problem of conducting a trial in the face of strong preferences for one or other treatment is to perform a patient preference trial, where patients with a clear preference for one type of treatment receive their preferred procedure, and patients with no preference are randomized. This results in a randomized trial together with a large body of supporting evidence which can help to place the randomized trial in context. A further solution which has been put forward is the randomized consent trial.

Surgical Skill

This is the area where surgical trials differ most obviously from drug trials. In general, a drug does not require any great skill to administer, and it is simple to standardize the administration to eliminate any potential effect of the physician. By contrast, surgeons will vary in their levels of skill, experience, and enthusiasm, and inevitably the effect of surgery will be **confounded** with the effect of the surgeon.

One aspect of this issue is the "learning curve", where one could argue that to give a fair assessment of an operation one should only recruit surgeons who are thoroughly familiar with the procedure. This needs to be balanced by an appreciation of how quickly an operation can become "standard practice" without any formal evaluation. If one waits until there are several surgeons fully trained and experienced in a new procedure, then the procedure could easily be so firmly established as to make a trial impractical, or even for a trial to be considered by some to be unethical. Chalmers [4] has argued that, to avoid this bandwagon effect, one should randomize from the very first patient.

A related problem is the question of how to compare two procedures where the surgeons participating in the trial are skilled in one procedure but inexperienced in the other. An extreme form of this

problem is when comparing procedures which would be performed by different specialists, say a conventional surgical procedure vs. a procedure performed by an interventional radiologist. Here, the only feasible design is to randomize the patients to the different specialities, meaning that the comparison of the procedures is completely confounded with the comparison of the individuals. This needs to be considered very carefully when discussing the generalizability of the results. For example, if one is comparing an "average" group of general surgeons with an atypical group of highly skilled and innovative radiologists, then the treatment comparison might be biased in favor of radiology and not generalize when the procedure becomes part of routine radiologic practice.

Technical Points

In addition to the "high-level" obstacles described above, there are several other technical difficulties which can arise in surgical trials. Many of these relate to the choice of an appropriate end point. For routine minor procedures the rate of serious complications will generally be low, but this means that without a very large sample size it would be possible to fail to detect a substantial and clinically relevant increase in the number of serious complications. This sample size issue is very relevant to the discussion of laparoscopic cholecystectomy which follows in the next section.

Another difficulty is in deciding when to assess outcome. In many procedures in surgical oncology the most relevant outcome measure is long-term survival, so there is a risk that a procedure might be obsolete even before it is fully evaluated. A more difficult question arises when the treatments compared attempt to strike different balances between short-term and long-term outcome. Typically, one might be comparing a major operation which has a high rate of early complications, but with the prospect of a long-term cure, against a more conservative procedure which might control local symptoms but compromise long-term prognosis. If such a study is analyzed too soon, then there will be a strong bias in favor of the treatment with the better short-term outcome, which might mask an important long-term benefit.

An extreme example of this problem is the assessment of a prophylactic operation such as carotid endarterectomy, where one operates on an individual thought at high risk of suffering a **stroke** in an attempt to reduce this risk. There is significant morbidity associated with the operation, and so the question is whether it is appropriate to accept the immediate risks of surgery in the hope of avoiding a stroke at some later date.

The example of carotid endarterectomy raises a further important issue, in that the problem is not to evaluate the procedure per se, but to identify patients where the procedure is indicated. There is good evidence that patients with severe stenosis (narrowing) of their internal carotid artery (70%–99% stenosis) do tend to benefit from the procedure, and those with mild stenosis (0%–29%) are more likely to be harmed than to benefit. The key statistical question is therefore not whether carotid endarterectomy is an effective treatment, but rather the problem is to estimate the threshold for percent stenosis above which the procedure is indicated. A refinement of this question would be to individualize this threshold to take account of each patient's constellation of risk factors. Preliminary results of a study of patients with moderate stenoses have been reported [9], but their analysis fails to address directly the question of estimating the threshold which determines whether the operation is indicated.

Case Study – Laparoscopic Cholecystectomy

Cholecystectomy, or removal of the gall bladder, is a very common surgical procedure. In the US, for example, approximately 500 000 such operations are performed annually. In the early 1980s a number of surgeons began experimenting with performing this operation laparoscopically rather than through a conventional incision, and by the early 1990s the laparoscopic approach had become so popular that several audits reported that over 80% of cholecystectomies were performed laparoscopically. The perceived advantages included less postoperative pain, a smaller scar, a shorter hospital stay, and a quicker return to normal activities. Indeed, the laparoscopic procedure is seen as such an advance over conventional surgery that the indications for cholecystectomy have been relaxed, and an increase of over 20% has been observed in the number of cholecystectomies performed.

In spite of the huge number of procedures performed worldwide, there is a remarkable dearth of scientifically secure data to evaluate the new approach. Cuschieri [6], one of the leading

proponents of laparoscopic surgery, has called the explosive growth in uptake of the new procedure "the greatest unaudited free-for-all in the history of surgery".

A recent review [7] identified 841 articles specifically on laparoscopic cholecystectomy published between January 1987 and October 1994. Of these 841 articles, only 15 were randomized trials, and of these, only three contained at least 50 patients in each group. Between 1994 and 1996 there have been over 1500 publications in English on laparoscopic surgery in general, but only 11 report randomized trials large enough to allow useful comparisons [15]. This is quite extraordinary given the number of procedures which are performed annually, and given that much of the debate over laparoscopic cholecystectomy centers around the risk of bile duct injury. This is a serious complication, but with an incidence of 1% or less. Clearly, to make a useful comparative statement about such a low complication rate would require a study of several thousand patients. However, it must be stressed that, for such a common procedure, which is generally performed to relieve symptoms in otherwise fit individuals, a complication rate of 1% is clinically important. Moreover, given the large number of procedures being undertaken, it would be eminently feasible to perform a trial based on several thousand individuals.

It is only now, ten years after the enthusiastic and widespread adoption of laparoscopic cholecystectomy, that a more balanced picture of the costs and benefits is beginning to emerge (*see* **Health Economics**). An influential randomized trial by Majeed et al. [16] used blinded assessment techniques to assess rates of recovery and length of hospital stay, and observed no difference between laparoscopic cholecystectomy and small-incision cholecystectomy. A number of audits have suggested that laparoscopic cholecystectomy is associated with approximately a twofold increase in bile duct injury and in other complications requiring readmission to hospital. Johnson [15] concludes his discussion of laparoscopic surgery by saying

> Laparoscopic surgery is not easier, quicker, cheaper, or safer; nor does it avoid general anesthetic. It may lead to a shorter initial hospital stay but readmissions for complications and other procedures have to be added. [...] Laparoscopic surgery [...] must be classified as an expensive luxury rather than a surgical revolution.

Surgical Audit

As discussed above, the randomized clinical trial should be regarded as the "gold standard" for the evaluation of surgical interventions. However, as also mentioned above, one of the most difficult variables to handle is the skill of the individual surgeon. In what has been an enormously influential paper, Fielding et al. [10] reported data from the Large-Bowel Cancer Project which showed dramatic differences between surgeons in the rate of breakdown of anastomoses (i.e. surgical joins of the bowel). In the summary of that paper it is stated

> The data in the current study show that the surgeon who has clinical responsibility for the care of the patient is probably the most important single factor influencing anastomotic integrity. Such a statement about surgical technique may be thought controversial, but there is a sixfold range of results (about 5–30%) that cannot be accounted for by any obvious differences in patient population.

If such results are indeed representative of surgery in general, then we could be straining at gnats when conducting trials looking for subtle differences between two variants of the same procedure, when the surgical outcome is largely determined by the responsible surgeon.

In certain unusual situations it might be feasible to conduct a randomized trial which compared different surgeons undertaking the same procedure, but the assessment of individual performance, which is termed "surgical audit", is generally undertaken less formally. There is a very large literature on surgical audit, and a review by Hayes & Murray [13] gives many of the key references. Surgical audit has many parallels with the equally controversial topic of school league tables (*see* **Quality of Care**), and a review of the statistical issues underlying both medical and school league tables is given by Goldstein & Spiegelhalter [11]. This paper brings out the idea of **multilevel modeling**, which is very relevant in a context where the data can have a clear hierarchical structure – with, say, data on individual patients, under the care of an individual consultant, working as part of a surgical team, within a hospital, within an administrative region (*see* **Hierarchical Models**).

The history of surgical audit goes back to at least the 1850s, when **Florence Nightingale** was working during the Crimean War. Throughout the first half of

the twentieth century surgical audit was developed in the US, very much from the point of view of regulation, and without notable success. An excellent review of these early historical developments is given by Wilkin & McColl [20]. There is also a long history of surgical audit in the UK, but always with more of an emphasis on education than on the "big stick". The Royal College of Surgeons of England now hosts a Surgical Epidemiology and Audit Unit, and provides a Comparative Audit Service.

The fundamental problem with all clinical audits is that one does not compare like with like. Differences in **case mix** confound direct comparisons of outcome, but this does not prevent such misleading data being sensationalized in the media. Many attempts have been made to build statistical models which can adjust for case mix, and hence give a better measure of "added value". The paper by Hayes & Murray [13] describes some of the better known scoring systems, and highlights some of their limitations. The POSSUM Score [5] is one scoring system which has been developed specifically for general surgery. It can be used either to identify patients whose outcome was unexpected, so that the case can be reviewed in detail, or else to compare the expected outcomes in a series of patients with the outcomes which were actually observed. POSSUM requires a large volume of data to be recorded, but it is claimed that the system can work in practice. One problem with the system is that the model is not well calibrated, and, in particular, the predicted risk of death cannot be below 1.08%, no matter how fit the patient or how trivial the operative procedure. Perhaps more seriously, some of the factors which are included in the model, such as blood loss, may actually reflect surgical competence. Thus one is assessing competence, adjusting for competence, which rather defeats the purpose!

No system for case mix adjustment is ever going to be perfect, but equally it seems likely that the demand for "league tables" will continue to grow. Certainly within the UK there is a growing requirement for purchasers of health care to be able to measure quality, and so further research into how best to adjust performance indicators for case mix needs to be a priority. Equally there is a pressing need to educate the consumers of "league tables" of their limitations, even after correction for case mix.

References

[1] Anonymous (1986). Positive discrimination for surgery?, *Lancet* **336**, 151.

[2] Barsamian, E.M. (1977). The rise and fall of internal mammary artery ligation in the treatment of angina pectoris and the lessons learned, in *Costs, Risks and Benefits of Surgery*, J.P. Bunker, B.A. Barnes & F. Mosteller, eds. Oxford University Press, New York.

[3] Chalmers, I. (1990). Evaluating the effects of care during pregnancy and childbirth, in *Effective Care in Pregnancy and Childbirth*, I. Chalmers, M. Enkin & M.J.N.C. Keirse, eds. Oxford University Press, Oxford.

[4] Chalmers, T.C. (1975). Randomization of the first patient, *Medical Clinics of North America* **59**, 1035–1038.

[5] Copeland, G.P., Jones, D. & Walters, M. (1991). POSSUM: a scoring system for surgical audit, *British Journal of Surgery* **78**, 356–360.

[6] Cuschieri, A. (1995). Whither minimal access surgery: tribulations and expectations, *American Journal of Surgery* **169**, 9–19.

[7] Downs, S.H., Black, M.A. & Devlin, H.B., Royston, C.M.S. & Russell, R.C.G. (1996). Systematic review of the effectiveness and safety of laparoscopic cholecystectomy, *Annals of the Royal College of Surgeons of England* **78**, 241–323.

[8] EC/IC Bypass Study Group (1985). Failure of extracranial-intracranial arterial bypass to reduce the risk of ischaemic stroke. Results of an international randomised trial, *New England Journal of Medicine* **313**, 1191–1200.

[9] European Carotid Surgery Trialists' Collaborative Group (1996). Endarterectomy for moderate symptomatic carotid stenosis: interim results from the MRC European Carotid Surgery Trial, *Lancet* **347**, 1591–1593.

[10] Fielding, L.P., Stewart-Brown, S., Blesovsky, L. & Kearney, G. (1980). Anastomotic integrity after operations for large-bowel cancer: a multicentre study, *British Medical Journal* **288**, 411–414.

[11] Goldstein, H. & Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with Discussion), *Journal of the Royal Statistical Society, Series A* **159**, 385–443.

[12] Hall, J.C., Mills, B., Nguyen, H. & Hall, J.L. (1996). Methodologic standards in surgical trials, *Surgery* **119**, 466–472.

[13] Hayes, C. & Murray, G.D. (1995). Case mix adjustment in comparative audit, *Journal of Evaluation in Clinical Practice* **1**, 105–111.

[14] Horton, R. (1996). Surgical research or comic opera: questions but few answers, *Lancet* **347**, 984–985.

[15] Johnson, A. (1997). Laparoscopic surgery, *Lancet* **349**, 631–635.

[16] Majeed, A.W., Troy, G., Nicholl, J.P., Smythe, A., Reed, M.W.R., Stoddard, C.J., Peacock, J. &

Johnson, A.G. (1996). Randomised, prospective, single-blind comparison of laparoscopic versus small-incision cholecystectomy, *Lancet* **347**, 989–994.

[17] Murray, G.D. (1988). The task of a statistical referee, *British Journal of Surgery* **75**, 664–667.

[18] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.

[19] Russell, I. (1995). Evaluating new surgical procedures, *British Medical Journal* **311**, 1243–1244.

[20] Wilkin, A. & McColl, I. (1987). Surgical audit: the clinician's view, *Theoretical Surgery* **1**, 195–206.

GORDON D. MURRAY

Surrogate Endpoints

The selection of the primary "outcome measures" or "endpoints" is a very important step in the design of **clinical trials** (*see* **Outcome Measures in Clinical Trials**). Typically, the primary goal of the clinical trial is to assess definitively a treatment's effect on these endpoints. Two major criteria should guide their selection. The endpoints should (i) be sensitive to treatment effects and (ii) be clinically relevant. Adequate attention is usually given to ensuring that the first criterion is satisfied. Unfortunately, ensuring that the endpoints also satisfy the criterion of clinical relevance is often improperly addressed. We focus on this second criterion and the corresponding controversial issues arising when surrogate endpoints are used as study outcomes.

The nature of clinical relevance depends on the stage of clinical experimentation. In **Phase II trials**, which provide a screening evaluation of treatment effect, the primary objective usually is to assess a treatment's *biological activity*. Relevant endpoints in such a trial in cancer patients might be measures of tumor shrinkage; in HIV-infected persons, measures of viral load or immune function; and in patients with cardiovascular disease, blood pressure or lipid levels. In contrast, in Phase III clinical trials, where the intent is to define the role of a therapy in standard clinical practice, the primary objective should be to assess the treatment's *clinical efficacy* through outcome measures that unequivocally reflect tangible benefit to the patient. In the treatment of patients with life-threatening diseases, such clinical efficacy measures include improvement in the duration of survival or in the **quality of life** (QOL).

Often, there is a sense of urgency in the evaluation of promising new interventions for patients having life-threatening diseases. When survival is the primary endpoint, clinical trials frequently require large sample sizes and very lengthy intervals of follow-up. The subjective nature of QOL outcome measures presents additional difficulties through the need to identify validated and widely accepted QOL instruments. To reduce the trial cost, size, and duration and to avoid complexities of QOL assessments, considerable attention has been given, in the design of definitive Phase III trials, to identifying surrogate or replacement endpoints for the true clinical efficacy endpoint. As defined by Temple [21],

> a surrogate endpoint of a clinical trial is a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives. Changes induced by a therapy on a surrogate endpoint are expected to reflect changes in a clinically meaningful endpoint.

Measures of biological activity have been chosen frequently as surrogates because usually they are readily available early in a clinical trial and because often they are strongly correlated with clinical efficacy.

Unfortunately, treatment effects on the clinical efficacy endpoints may not be predicted reliably by the observed effects on surrogate endpoints, even when natural history data reveal that these surrogates are strongly correlated with the clinical efficacy outcomes. As indicated by Fleming & DeMets [9], there are several possible explanations for this failure.

> Even though a surrogate endpoint may be a correlate of disease progression, it might not involve the same pathophysiologic process that results in the clinical outcome. Even when it does, it is likely there are disease pathways causally related to the clinical outcome and yet unrelated to the surrogate endpoint. Of the disease pathways affecting the true clinical outcome, the intervention may only affect (i) the pathway mediated through the surrogate endpoint or (ii) the pathway(s) independent of the surrogate endpoint. Most importantly, the intervention might also affect the true clinical outcome by unintended mechanisms of action independent of the disease process. The intervention's effects mediated through intended mechanisms could be substantially offset by an array of mechanisms that are unintended, unanticipated and unrecognized.

The example of lipid-lowering agents clearly illustrates the existence and impact of these unintended mechanisms. In a comprehensive overview of 50 randomized trials (*see* **Meta-analysis of Clinical Trials**) of cholesterol-lowering agents by Gordon [12], an average reduction in cholesterol of 10% was achieved along with an intended 9% reduction in coronary heart disease (CHD) mortality. However, overall mortality was unchanged, due to an unintended 24% increase in non-CHD mortality.

Illustrations

Research across a broad array of clinical settings confirms that many powerful correlates of clinical

efficacy outcomes have been poor surrogates for true clinical efficacy [8, 9]. Anti-arrhythmic drugs effectively suppress ventricular arrhythmias after myocardial infarction (MI), yet lead to more than three-fold increases in death rate. Drugs improving cardiac output as treatment for congestive heart failure have increased mortality. The rate of vessel reperfusion has not predicted adequately the effect of thrombolytic therapies on mortality. Cholesterol-lowering interventions, such as diet, fibrates, hormones, resins, and lovastatin, have not lowered mortality rates. Antihypertensive calcium channel blockers reduce blood pressure, but now appear to increase the risk of myocardial infarction. Calcium antagonists reduce the risk of developing new angiographic lesions of atherosclerosis in patients with MI, yet increase the death rate. Sodium fluoride increases bone mineral density in postmenopausal women with osteoporosis, yet substantially increases the risk of bone fractures. In patients with retinitis pigmentosa, vitamin A provides a favorable slowing of decline on electroretinograms, yet has no effect on any direct measure of visual function. In addition to these **false positive** leads, surrogates can provide **false negative** leads as well. Gamma interferon fails to have a measurable effect on superoxide production and bacterial killing in children with chronic granulomatous disease, yet substantially reduces the rate of serious life-threatening infection.

Validation of Surrogates

Proper validation of a surrogate endpoint is a difficult task. Insights about validity can be provided by empiric evidence from an array of clinical trials documenting treatment effects on both surrogate and clinical efficacy endpoints, as well as by a thorough biological understanding of causal pathways in the disease process and of mechanisms of treatment effect (*see* **Causation**).

Prentice [19] provides a definition of a valid surrogate, and gives two sufficient conditions that jointly ensure this validity, thereby providing guidance for how one might approach using empiric evidence to assess validation. By his definition, a surrogate is valid if "a test of the null hypothesis of no relationship (of the surrogate endpoint) to the treatment groups must also be a valid test of the corresponding null hypotheses based on the true endpoint". Prentice's first condition to ensure this validity is the "correlate" requirement, i.e. a valid surrogate endpoint must be **correlated** with the true clinical endpoint. This condition usually holds since, in practice, potential surrogates are often selected by identifying measures that are strongly correlated with clinical efficacy endpoints. Prentice's very restrictive second condition requires the surrogate to capture fully the treatment's "net effect" on the clinical endpoint, where the net effect is the aggregate effect accounting for all mechanisms of action. The restrictiveness of this condition provides important insight into why correlates are rarely valid surrogates. In applications, extensive analyses have been performed to assess surrogacy of CD4 cell count, using data from several large clinical trials evaluating nucleoside analogs in HIV/AIDS patients. While these analyses show consistently that CD4 cell count is a correlate of the "progression to symptomatic AIDS or death" endpoint, thereby satisfying Prentice's first condition, CD4 has not been established as a valid surrogate endpoint, since the second condition of Prentice consistently fails to hold [3, 6, 14, 15, 22].

The validity of Prentice's restrictive second condition, requiring a surrogate to capture fully the net effect of an intervention on the clinical efficacy endpoint, has been explored by Freedman et al. [11] in an epidemiologic setting. Their methods involve estimating the proportion of the net treatment effect apparently captured by the marker, allowing assessment of the strength of evidence about whether this proportion is near unity. It should be recognized, however, that while particular interest often is in evaluating the effect of treatment on the disease process pathway(s) causally inducing the clinical events, it is not possible to determine the proportion, p, of that effect that is accounted for by effects on a surrogate endpoint [20]. To demonstrate how this nonidentifiability arises, DeGruttola et al. [7] consider a simple example in which the clinical efficacy endpoint is death and where, on the control regimen, the death rate induced by the causal pathways of the disease process is μ_h, while the death rate due to other causes is μ_o. They suppose further that the experimental intervention alters the death rate induced by the causal pathways of the disease process by the multiplicative factor r, to $r\mu_h$, but increases the death rate due to other causes (including those influenced by unintended mechanisms of the drug) by the multiplicative factor k, to $k\mu_o$. If it is assumed that

the treatment-induced change in the surrogate endpoint would only influence the death rate induced by the causal pathways of the disease process, then this change would alter the overall death rate by a multiplicative factor

$$r_{so} = \frac{(\mu_h + \mu_o) - p(1-r)\mu_h}{\mu_h + \mu_o}. \quad (1)$$

This parameter r_{so} can be measured using data on **control** patients from the study itself or natural history databases that allow modeling the association of death with surrogate endpoints. One can also measure the observed overall "net effect" of the intervention on death rate,

$$r_o = \frac{r\mu_h + k\mu_o}{\mu_h + \mu_o} \quad (2)$$

and, using (1) and (2), can compute the observed portion of the net effect accounted for by the treatment-induced change in the surrogate endpoint,

$$p_o = \frac{1 - r_{so}}{1 - r_o} = \frac{p(1-r)\mu_h}{(1-r)\mu_h + (1-k)\mu_o}. \quad (3)$$

By (3), $p_o = p$ if either $\mu_o = 0$ (i.e. death can only be caused by the causal pathways of the disease process) or $k = 1$ (i.e. the intervention has no effect on the other causes of death). However, in the more common setting where $\mu_o > 0$ and $k > 1$, even when $p \gg 1$, the observed proportion p_o approaches unity as k approaches $[1 + (\mu_h/\mu_o)(1-r)(1-p)]$. Thus, surrogate endpoints, which capture only a small fraction of the change in the death rate induced by treatment effects on the causal pathways of the disease process, may appear to capture an observed portion, p_o, near unity, simply due to unanticipated and unrecognized harmful effects of the intervention on the other causes of death.

To formulate estimators of p_o in epidemiologic data, Freedman et al. [11] used linear **logistic regression** models, while Choi et al. [3], O'Brien et al. [17] and DeGruttola et al. [7] used **proportional hazards** models to conduct similar analyses in the setting of **censored** failure time data. Specifically, these three sets of authors assume that the failure rate at time t in treatment group Z is

$$\lambda(t|Z) = \lambda_o(t)\exp(\beta Z), \quad (4)$$

where $Z = 0$ for control and $Z = 1$ for experimental treatment, β is an unknown constant, and $\lambda_o(t)$ is an arbitrary positive function. From (2) and (4), the "net treatment effect" is $r_o = e^{\beta}$. In turn, incorporating the effect of the surrogate $X(t)$ on the failure rate at time t, DeGruttola et al. [7] assume the model

$$\lambda[t|Z, X(t)] = \tilde{\lambda}_o(t)\exp(\beta_a Z)\exp[\alpha X(t)], \quad (5)$$

where β_a and α are unknown constants, and $\tilde{\lambda}_o$ is an arbitrary positive function that might differ from λ_o. Strictly speaking, models (4) and (5) cannot hold simultaneously; however, they may hold approximately when either α or $\int \tilde{\lambda}_o(t)$ is small. We will assume that the effects of model **misspecification** are negligible (see Lin et al. [16] for a rigorous discussion of this issue). By (4) and (5),

$$r_{so} = \exp(\beta - \beta_a).$$

Thus,

$$p_o = \frac{1 - \exp(\beta - \beta_a)}{1 - e^{\beta}}.$$

Freedman et al. [11] approximate p_o by

$$p_o^* = 1 - \frac{\beta_a}{\beta}.$$

The two quantities, p_o and p_o^*, are equivalent when $\beta_a = \beta$ or $\beta_a = 0$, and differ only slightly for intermediate values. Of course, as shown above, the quantities are equal to p only in very special cases. Note that while p, the proportion of the intended effect on causal pathways of the disease process captured by the surrogate, is always a proportion in the mathematical sense of lying in the interval [0, 1], p_o need not lie in the interval [0, 1]. When β_a and β differ in sign (a situation that arises when the surrogate captures all of the benefit so that only the harmful effect is reflected in β_a), p_o exceeds 1; when $\beta_a > \beta > 0$, p_o is negative.

The problems of interpretation of p_o and p_o^* are compounded by the high variability of their estimators [11]. Let $\hat{\beta}$ and $\hat{\beta}_a$ denote the estimates of β and β_a, obtained by the usual method of maximum **partial likelihood**. Then p_o^* is estimated by

$$\hat{p}_o^* = 1 - \frac{\hat{\beta}_a}{\hat{\beta}}.$$

Lin et al. [16] showed that, for large samples, $\hat{p}_{\mathrm{o}}^{*}$ is approximately normal with **mean** p_{o}^{*} and with **variance**

$$\sigma^2 = \frac{V_\beta}{\beta^2}\left\{\frac{V_{\beta_{\mathrm{a}}}}{V_\beta} + (1-p_{\mathrm{o}}^{*})^2 - 2(1-p_{\mathrm{o}}^{*})\frac{V_{\beta\beta_{\mathrm{a}}}}{V_\beta}\right\}, \tag{6}$$

where V_β and V_{β_a} are the variances of $\hat{\beta}$ and $\hat{\beta}_{\mathrm{a}}$, and $V_{\beta\beta_a}$ is their covariance.

Formula (6) indicates that the factors that determine the variance of $\hat{p}_{\mathrm{o}}^{*}$ include the coefficient of variation for β (i.e. the inverse of the unadjusted treatment effect relative to its **standard error**), the value of p_{o}^{*} itself, and the values of $V_{\beta_{\mathrm{a}}}$ and $V_{\beta\beta_{\mathrm{a}}}$ relative to V_β. For illustration, suppose α is small and the correlation between treatment and marker is low. Then $V_\beta \approx V_{\beta_{\mathrm{a}}} \approx V_{\beta\beta_{\mathrm{a}}}$, in which case

$$\sigma \approx |p_{\mathrm{o}}^{*}|\frac{\mathrm{se}(\hat{\beta})}{|\beta|}. \tag{7}$$

Suppose that we have a large unadjusted treatment effect which is four times its standard error, i.e. $\beta/\mathrm{se}(\hat{\beta}) = 4$. Then (7) implies that the mean width of the 95% **confidence interval** for p_{o}^{*} is equal to p_{o}^{*} itself. In practice, (7) tends to underestimate the true variability of $\hat{p}_{\mathrm{o}}^{*}$ because $V_{\beta_{\mathrm{a}}}$ is generally larger than V_β. The estimate $\hat{\beta}_{\mathrm{a}}$ becomes increasingly unstable as the correlation between treatment and marker increases. (An extreme scenario occurs when, in a placebo-controlled trial of a treatment, all treated and no untreated patients have a marker response.) Thus, an unadjusted treatment effect that is greater than four times its standard error is a necessary, though insufficient, condition for precise estimation of p_{o}^{*}. Similar observations are made by Freedman et al. [11].

Clinical studies with treatment effects that are many times their standard errors are unusual, because studies with large treatment effects tend to be stopped early, and because most studies do not compare treatments with greatly different degrees of efficacy. Thus, meta-analyses that combine evidence across studies usually would be required for statistical evaluation of the reliability of surrogate endpoints.

There are a number of ways to make use of data collected across studies. The first is simply to estimate p_{o}^{*} (and its associated variance) corresponding to a given surrogate for each individual study, and examine the consistency of these estimates. This may be especially useful when there have been a variety of treatments under study, with differing mechanisms of action and toxicity profiles. More formally, one could treat the true values of the p_{o}^{*} for each study as latent variables, and estimate their underlying distribution (or features of the distribution) across studies. In settings where p_{o}^{*} appears to be highly variable across studies, it might be of interest to assess whether such factors as class of drug or population under study explain this variability. While such efforts are not free from the problems of **identifiability** described earlier, values of p_{o}^{*} that are consistently near 1 for studies investigating different classes of treatments may provide more persuasive evidence about the validity of a surrogate than do results from individual studies. An alternative approach to using data across studies, proposed by Daniels & Hughes [5], uses **Bayesian methods** to construct prediction intervals for the true difference in clinical outcome associated with a given estimated treatment effect on the potential surrogate.

A factor that further complicates analyses of surrogacy, especially analyses across studies, is that marker values are generally not measured continuously or without error. **Measurement error** and the fact that marker values are available only at certain times – times that are often influenced by the disease under study – can result in **bias** in the estimation of α, and hence of β and p_{o}^{*}. Tsiatis et al. [22] explored methods for correcting for bias resulting from issues related to measurement.

Auxiliary Variables

Rather than serving as surrogates to replace clinical efficacy endpoints, response variables, such as the measures of biological activity discussed earlier, can be used to strengthen clinical efficacy analyses. Such variables, S, are then called *auxiliary*. Suppose one's interest is in the effect of treatment on time to a clinical endpoint, T. Suppose, furthermore, that the auxiliary information, S, is readily observed, whereas T is **censored** in a substantial fraction of the patients because they have relatively late clinical endpoints. If S and T are strongly correlated, one can expect that S will provide useful additional information about the timing of the clinical endpoint for those patients in which T is censored.

Three approaches have been proposed for using auxiliary variables, and are referred to as "variance

reduction", "augmented score" and "estimated likelihood". The variance reduction method, explored by Kosorok & Fleming [13], is applicable when S is a time-to-event endpoint and when the treatment relationship with S is described by a statistic X with zero mean, such that $\text{cor}(X, Y) \equiv \rho$ is positive, where Y is a standard statistic used to assess the effect of treatment on T. The statistic $Y - \rho X$ proposed by Kosorok & Fleming makes use of auxiliary information to provide a variance-reduced alternative to using Y.

The "augmented score" and "estimated likelihood" methods were explored by Fleming et al. [10]. Both approaches assume the proportional hazards model of (4) for the relationship between the covariate vector Z and the **hazard** function for the clinical outcome T. Denote the **cumulative hazard** for λ_o by Λ_0. Assume T_i and U_i are independent latent failure and censoring variables for the ith patient ($i = 1, \ldots, n$), and denote $X_i = \min\{T_i, U_i\}$ and $\delta_i = I_{\{X_i = T_i\}}$, where $I_{\{A\}}$ denotes an indicator for A.

To motivate the "augmented score" approach, recall that in the **semiparametric regression** setting where λ_o is unspecified, the Cox [4] maximum *partial likelihood* estimate of β is obtained by solving the score estimating equation:

$$\sum_{i=1}^{n} Z_i \hat{M}_i(X_i|\beta) = 0, \tag{8}$$

where, for any $t \geq 0$,

$$\hat{M}_i(t|\beta) = I_{\{T_i \leq t\}} - \exp(\beta' Z_i)\hat{\Lambda}_0(t \wedge T_i)$$

is the martingale residual (*see* **Counting Process Methods in Survival Analysis**) evaluated at β, and where $\hat{\Lambda}_0$ is the semiparametric Breslow [1] estimator of Λ_0 evaluated at β.

Censorship reduces the information available in (8) that is used for the estimation of β. Specifically, $\hat{M}_i(t|\beta)$ is only known over $t \in [0, X_i]$ rather than over $t \in [0, T_i]$ and, in (8), less information is available to formulate $\hat{\Lambda}_0$. Fortunately, the surrogate information, S_i, does allow recovery of some of this lost information. Suppose τ denotes some arbitrary large time. To recover some information over $(X_i, \tau]$ for a censored case (i.e. with $\delta_i = 0$), consider

$$e_{\hat{M}_i}(\beta) \equiv \text{E}[\hat{M}_i(\tau|\beta) - \hat{M}_i(X_i|\beta)|X_i, \delta_i = 0, S_i],$$

which essentially is the conditional expectation of the lost information over $(X_i, \tau]$, given available information on case i to X_i. Fleming et al. [10] formulate an estimator $\hat{e}_{\hat{M}_i}(\beta)$ in the special case in which S_i is a censored time-to-event endpoint, and propose estimation of β based on solving the "augmented score equation":

$$\sum_{i-1}^{n} Z_i \hat{M}_i(X_i|\beta) + \sum_{i=1}^{n} (1 - \delta_i) I_{\{X_i < \tau\}} Z_i \hat{e}_{\hat{M}_i}(\beta) = 0.$$

In the "estimated likelihood" approach, following Pepe's [18] semiparametric approach in which λ_o temporarily is assumed to be known and that involves **nonparametric** estimation of $P(S|T, Z)$ to obtain greater **robustness**, the corresponding estimated **likelihood** is

$$\hat{L}(\beta) = \prod_{\delta_i=1} P_\beta(T_i|Z_i) \prod_{\delta_i=0} P_\beta(T > X_i|Z_i) \times \prod_{\delta_i=0} \hat{P}_\beta(S_i|T > X_i, Z_i), \tag{9}$$

where S_i can be an arbitrary right-censored vector-valued process providing auxiliary information. The first two terms on the right-hand side of (9) represent the usual likelihood when the auxiliary information, S, is not taken into account. Under (4), these two terms reduce to the usual Cox partial likelihood when λ_o is considered to be unspecified and, in turn, is estimated by the piecewise linear approach presented in Breslow [2]. Turning to the third term in the estimated likelihood in (9), the amount of improvement provided by the estimated likelihood relative to the usual partial likelihood depends on the degree of dependence of $P_\beta(S_i|t, Z_i)$ on t.

Improvements in efficiency with these approaches using auxiliary information are likely to be small unless S and T are highly correlated and unless there is one pool of patients having longer-term follow-up and another pool of patients with auxiliary information but with relatively short-term follow-up on the clinical endpoint. In spite of these limitations, approaches using auxiliary information are of interest since they avoid the substantial risks for false positive or false negative conclusions that arise when surrogate endpoints are used to replace measures of clinical efficacy.

Conclusions

It would be rare to be able to establish rigorously the validity of a surrogate endpoint. False positive and false negative error rates in definitive trials evaluating intervention effects on clinical outcomes are required to be very low, typically in the range of 2.5% to 10%. Hence, to be a valid replacement endpoint, a surrogate must provide a very high level of accuracy in predicting the intervention's effect on the true clinical endpoint. Predictions having an accuracy of approximately 50%, such as was provided by the CD4 surrogate in the HIV setting (see Fleming [8]), are as uninformative as random tosses of a coin. The statistical methods for validation discussed in this article usually require meta-analyses since the sample sizes needed are much larger than those necessary for the typical phase III evaluation of interventions (*see* **Sample Size Determination for Clinical Trials**). Proper validation of surrogates also requires in-depth understanding of the causal pathways of the disease process, as well as the intervention's intended and unintended mechanisms of action. Such in-depth insights are rarely achievable.

Surrogate endpoints should be used in screening for promising new therapies through the evaluation of biological activity in preliminary Phase II trials. Results of such studies can guide decisions about whether the intervention is sufficiently promising to justify the conduct of large-scale and longer-term clinical trials. In these definitive Phase III trials, while information on surrogate endpoints can provide valuable additional insights about the intervention's mechanisms of action, the primary goal should be to obtain direct evidence about the intervention's effect on safety and clinical outcomes.

References

[1] Breslow, N.E. (1972). Contribution to the discussion on the paper by D.R. Cox, Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 216–217.

[2] Breslow, N.E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–99.

[3] Choi, S., Lagakos, S.W., Schooley, R.T. & Volberding, P.A. (1993). CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine, *Annals of Internal Medicine* **118**, 674–680.

[4] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[5] Daniels, M.J. & Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers, *Statistics in Medicine* **16**, 1965–1982.

[6] DeGruttola, V., Wulfsohn, M., Fischl, M.A. & Tsiatis, A.A. (1993). Modeling the relationship between survival and CD4+ lymphocytes in patients with AIDS and AIDS-related complex, *Journal of Acquired Immune Deficiency Syndrome* **6**, 359–365.

[7] DeGruttola, V., Fleming, T.R., Lin, D.Y. & Coombs, R. (1996). Validating surrogate markers: are we being naive? *Journal of Infectious Disease* **175**, 237–246.

[8] Fleming, T.R. (1994). Surrogate markers in AIDS and cancer trials, *Statistics in Medicine* **13**, 1423–1435.

[9] Fleming, T.R. & DeMets, D.L. (1996). Surrogate end points in clinical trials: are we being misled?, *Annals of Internal Medicine* **125**, 605–613.

[10] Fleming, T.R., Prentice, R.L., Pepe, M.S. & Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research, *Statistics in Medicine* **13**, 955–968.

[11] Freedman, L.S., Graubard, B.I. & Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases, *Statistics in Medicine* **11**, 167–178.

[12] Gordon, D.J. (1994),. in *Contemporary Issues in Cholesterol Lowering: Clinical and Population Aspects*, B.M. Rifleind, ed. Marcel Dekker, New York.

[13] Kosorok, M.R. & Fleming, T.R. (1993). Using surrogate failure time data to increase cost effectiveness in clinical trials, *Biometrika* **80**, 823–833.

[14] Lagakos, S.W. & Hoth, D.F. (1992). Surrogate markers in AIDS: where are we? *Annals of Internal Medicine* **116**, 599–601.

[15] Lin, D.Y., Fischl, M.A. & Schoenfeld, D.A. (1993). Evaluating the role of CD4−lymphocyte counts as surrogate endpoints in HIV clinical trials, *Statistics in Medicine* **12**, 835–842.

[16] Lin, D.Y., Fleming, T.R. & DeGruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker, *Statistics in Medicine* **16**, 1515–1527.

[17] O'Brien, W., Hartigan, P.M., Martin, D., Esinhart, J., Hill, A., Benoit, S., Rubin, M., Simberkoff, M.S., Hamilton, J.D. & the Veterans Affairs Cooperative Study Group on AIDS (1996). Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS, *New England Journal of Medicine* **334**, 426–431.

[18] Pepe, M.S. (1992). Inference using surrogate outcome data and a validation sample, *Biometrika* **79**, 355–365.

[19] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.

[20] Schatzkin, A., Freedman, L.S., Schiffman, M.H. & Dawsey, S.M. (1990). Validation of intermediate end points in cancer research, *Journal of the National Cancer Institute* **82**, 1746–1752.

[21] Temple, R.J. (1995). A regulatory authority's opinion about surrogate endpoints, in *Clinical Measurement in*

Drug Evaluation, W.S. Nimmo & G.T. Tucker, eds. Wiley, New York.

[22] Tsiatis, A.A., DeGruttola, V. & Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS, *Journal of the American Statistical Association* **90**, 27–37.

THOMAS R. FLEMING, VICTOR DE GRUTTOLA & DAVID L. DEMETS

Surveillance of Diseases

Modern public health surveillance of disease has been defined by Langmuir [14] as "the continued watchfulness over the distribution and trends of incidence through the systematic collection, consolidation and evaluation of morbidity and mortality reports and other relevant data". It is now usual to add to this definition the final link of applying these data to prevention and control [26]. But this public health activity is not new. One of the earliest examples of population surveillance was that developed in the City of London in the sixteenth and seventeenth centuries to detect plague, so that the City Fathers could decide when to close theaters and limit the assembly of crowds, and the Royal Court could leave for the countryside [31]. Data on plague deaths were collected by parish clerks, summated each week and reported in the "Bills of Mortality". The system neatly illustrates the steps in surveillance, which are the systematic collection of data, analyses to produce statistics, interpretation to provide information–which is then reported fast enough so that action can be taken–followed by continuing surveillance to evaluate the success of the action.

The concept of surveillance is simple, but in practice there is a tendency for surveillance systems to drift from their original objectives and too easily lose their focus on public health action–"Reporting does not equal surveillance". It is therefore important that surveillance as a dynamic public health activity is distinguished from managing registries (*see* **Disease Registers**) and other health information systems such as registrations of births and deaths (*see* **Vital Statistics, Overview**), though these may be useful data sources for surveillance. It is also important to recognize that public health surveillance differs from epidemiologic research in a number of important ways [24] (Table 1). The need for ongoing reporting (which distinguishes surveillance from occasional **surveys**) to provide information for action requires that surveillance systems are simple in construction, place minimal demands on data providers, and report accurate, readily understood, and timely information. Systems have often degenerated because data requirements have not been agreed with data providers and have become overburdened with secondary objectives. Consequent failure to report in a timely way had led to the loss of credibility with data providers. Recent successful infectious disease surveillance systems have used electronic reporting to minimize the burden on data providers and provide high-quality, rapid reporting [11, 29].

Guidelines on the evaluation of surveillance systems have been proposed by the **Centers for Disease Control (CDC)** [13], although few national systems appear to have been audited as recommended. The criteria include:

1. a description of the *public health importance* of the health event, including incidence and **prevalence**, severity of disease as measured by mortality rates and **case fatality** rates, and preventability;
2. a description of *the system*, including the objectives, the population under surveillance, case definitions, a flowchart of data collection, details of data transfer, data analyses, and dissemination of information;
3. a measure of the *usefulness* of the surveillance system, including decisions and actions taken as a result of the information generated;
4. evaluation of *key attributes* of the system, including simplicity, flexibility, acceptability, **sensitivity, positive predictive value**, representativeness, and timeliness;
5. the *cost* of the system.

Surveillance Systems

Up until the 1960s, public health surveillance activities were developed mainly for infectious disease control (*see* **Communicable Diseases**). Since then surveillance has been applied to many diseases, including congenital malformations (*see* **Teratology**), injuries, occupational illness (*see* **Occupational Epidemiology**), and adverse drug reactions (*see* **Postmarketing Surveillance of New Drugs and Assessment of Risk**), principally through the work of Langmuir at CDC [25]. Similar approaches have been taken to the surveillance of uptake of vaccines, and the surveillance of hazards, such as chemical accidents and the surveillance of behavioral risk factors [9]. Surveillance systems for chronic diseases have been less well developed [26]. Specific objectives of surveillance include:

1. early detection of changes in disease or risk factor prevalence and incidence to trigger rapid investigation and control;

Table 1 Distinctions between public health surveillance and epidemiologic research (adapted from Thacker & Berkelman [24])

	Surveillance	Epidemiologic research
Main purpose	Problem detection Problem description Trigger either investigation or intervention Suggest hypotheses	Hypothesis testing Problem description
Data collection		
Frequency	Ongoing	Time-limited
Methods	Normally routine systems study	Specially tailored for
Volume of data	Minimal	Considerable
Completeness of data	Often incomplete	Usually complete
Data analyses	Usually simple and descriptive	Often complex
Dissemination of information	Timely, regular, targeted to public health agencies	Not timely, sporadic, targeted to academics and clinical audience

2. measuring trends in disease, hazards, microbial agents, and risk factors to set priorities for interventions, and to evaluate disease-control programs;
3. to describe the basic epidemiology and natural history of disease in order to develop hypotheses about **causation**, which can be tested by separate research studies (*see* **Descriptive Epidemiology**).

Data Collection

Surveillance data may be sought actively or acquired passively by making use of routinely generated data such as death registrations (*see* **Death Certification**) or hospital admissions. A common weakness of surveillance systems is the lack of agreed case definitions. This applies to most laboratory reporting and notifiable diseases in the United Kingdom. In the United States, CDC have published surveillance case definitions for infectious diseases [30].

Statistical Analysis

Usually, the routine analysis of surveillance data is simply the presentation of **incidence rates** by time, place, and person, using graphs, histograms, and maps. However, more sophisticated methods are increasingly being used [23]. Particular statistical issues include the use of **time series** analysis to model epidemics (*see* **Epidemic Models, Stochastic**), the early recognition of unusual events in routine data against a variable baseline rate [4], **small area analysis** of **clustering**, adjustment for delays and incompleteness of reporting, the use of surveillance data to predict the course of epidemics (e.g. **AIDS**) (*see* **Projections: AIDS, Cancer, Smoking**).

Reporting

Timely reporting to those responsible for public health action is an essential part of a *bona fide* surveillance system. Timeliness is defined by the objectives of the surveillance. For infectious disease, timely reporting may need to be measured in hours, a target which can now be achieved globally through the Internet [7]. For chronic diseases, annual and quarterly reporting may be sufficient. Typically, surveillance reports appear either as specifically produced publications (e.g. *Communicable Disease Report, Office of National Statistics Monitor*, *Morbidity and Mortality Weekly Report*, and *Weekly Epidemiological Record*, or as electronic bulletins, such as EPINET [19]).

Infectious Diseases Surveillance

The best recent example of the power of surveillance is the case of AIDS. Following the first reports

Table 2 Statutorily notifiable diseases in England and Wales

Under the Public Health (Control of Disease) Act 1984	
Cholera	Relapsing fever
Food poisoning	Smallpox
Plague	Typhus
Under the Public Health (Infectious Diseases) Regulations 1988	
Acute encephalitis	Ophthalmia neonatorum
Acute poliomyelitis	Paratyphoid fever
Anthrax	Rabies
Diphtheria	Rubella
Dysentery (amoebic and bacillary)	Scarlet fever
Leprosy	Tetanus
Leptospirosis	Tuberculosis
Malaria	Typhoid fever
Measles	Viral hemorrhagic fever
Meningitis	Viral hepatitis
Meningococcal septicemia (without meningitis)	Whooping cough
Mumps	Yellow fever

Notes
"Viral hemorrhagic fever" means Argentine hemorrhagic fever (Junin), Bolivian hemorrhagic fever (Machupo), Chikungunya fever, Congo/Crimean hemorrhagic fever, Dengue fever, Ebola virus disease, hemorrhagic fever with renal syndrome (Hantaan), Kyasanur forest disease, Lassa fever, Marburg disease, Omsk hemorrhagic fever, and Rift valley disease.
There are minor differences in notifiable diseases in Scotland and Northern Ireland. Some diseases are notifiable locally; for example, psittacosis in Cambridge.
AIDS is *not* statutorily notifiable, but clinicians report cases voluntarily, in strict confidence, to the directors of the CDSC in England and Wales and of the SCIEH in Scotland. Advice about reporting is available from these centers and from genitourinary medicine physicians.

of a new clinical disease, surveillance based on a complex case definition quickly established the risk groups of AIDS and thereby the probable routes of transmission, so enabling preventive advice to be promulgated, even before the HIV virus was discovered. Subsequent surveillance using clinical reports of AIDS and laboratory reporting of HIV infection has been important in confirming the risk groups, reassuring the population about the absence of risk from casual contact and identifying localities of high incidence so that services can be targeted. Mathematical modeling using surveillance data has enabled prediction of the epidemic and has identified key transmission factors (e.g. number of sexual partners) in the maintenance of the disease [20].

Statutory Notification

In England and Wales, mandatory notification of infectious disease was introduced nationally in 1899. The current list of diseases is shown in Table 2. Notifications are made by registered medical practitioners. In England and Wales, weekly summaries of these data are now published in the Public Health Laboratory Service (PHLS) *Communicable Disease Report (CDR)*. The data are later corrected and published quarterly and annually by the **Office for National Statistics (ONS)**. The chief advantages of these data are that they are available quickly, and they relate to defined populations so that rates by age and sex can be calculated. The defects of the data are lack of case definitions and variable under-notification. Interestingly, the fee to medical practitioners to notify did not improve notification rates [18].

Laboratory Reporting of Microbiological Data

The PHLS developed laboratory reporting in the 1940s and 1950s [8]. Data are analyzed within a week of receipt by the Communicable Disease Surveillance Center to produce tables and line lists which are

used in compiling narrative reports for publication in the CDR. The recent introduction of CoSurv [11] has substantially replaced manual with electronic reporting. The main benefits of laboratory reports are that they are highly specific since they are based on laboratory-diagnosed infections and the fine typing of the infecting organisms [27], they often include clinical and epidemiologic details, and they allow for free-text comment. The reporting system is flexible, and unusual or new infections can be reported, even though they were not included in the original reporting instructions. However, the reports are limited to infections for which there is a suitable laboratory test.

General Practice Reporting of Clinical Data

The Royal College of General Practitioners (RCGP) set up a reporting system in 1966 based on first consultations to a limited number of volunteer practices [6]. In 1996 there were 367 participating general practitioners in 93 practices, serving a population of about 70 000 people; similar systems exist in Wales, Scotland, and the European countries [22]. They act primarily as early warning systems, particularly for influenza epidemics, providing data rapidly, and they have the advantages that the data are related to defined practice populations; they are useful for some common diseases which are not notifiable and for which laboratory tests are not usually performed, such as chicken pox.

Serological Surveillance

In 1990 in the UK, a serologic study to measure the spread of human immunodeficiency virus (HIV) infection in the population was begun; it has continued since and become a routine surveillance system [21]. Samples from sera collected for clinical purposes are unlinked from personal identifiers but remain linked to epidemiologic information (*see* **Record Linkage**); sera remaining unused are then tested for HIV infection.

Surveillance of Vaccine Preventable Diseases, Vaccine Uptake, and Vaccine Reactions

A comprehensive system of surveillance of vaccine-preventable diseases has been developed using the notification system, laboratory reporting, and regular serologic surveys of antibody levels in stored sera taken for other purposes [1]. Vaccine uptake is followed by the COVER [2] system in the UK, which uses the national child health system in which all children in the UK are registered by a health authority. Successive cohorts of children born within three-month periods are identified and their vaccination status at predefined target dates determined. Quarterly reports are published by the Communicable Disease Surveillance Centre and the comparative uptake rates are known by Districts within a few months. Health authorities have used the data to study and remedy reasons for low uptake. Surveillance of vaccine safety has used the Yellow Card Scheme, but record linkage of district health authority child health records and computerized hospital admissions records is a promising new method of postmarketing surveillance of vaccine safety [5]. Vaccine efficacy can also be the subject of surveillance if population rates of disease and the proportion of cases vaccinated (PCV) and the proportion of the population vaccinated (PPV) can be routinely measured [3]. Vaccine efficacy is calculated by the following expression:

$$1 - \frac{\text{PCV}}{1 - \text{PCV}} \times \frac{1 - \text{PPV}}{\text{PPV}}$$

(*see* **Vaccine Studies**).

Injury Surveillance

Particularly in Australia [10], hospital-based Emergency Room data have been used successfully to follow trends in injuries and identify etiologic risk factors. In the UK, injury surveillance currently relies on mortality data and hospital admissions, which will miss most common injuries such as fractures. Locally developed population-based schemes have illustrated the potential benefit of using Emergency Room databases [16]. In the US, systems have been developed to monitor spinal cord, firearm, and sports injuries [24].

Surveillance of Birth Defects

Following the thalidomide disaster, registries of congenital malformations were set up in several countries

for the early detection of malformations in order to investigate causes. However, incompleteness and inaccuracy of reports has reduced their potential effectiveness. Monitoring etiologically linked groups of malformations rather than single defects has been recommended [12]. In Europe, the European Registration of Congenital Anomalies (EUROCAT) concerted action project of the European Union collates standardized data from national and regional registries.

Occupational Illness and Injury Surveillance

In the United Kingdom, occupational illness and injuries are reportable by law under RIDDOR (the Reporting of Injuries Diseases and Dangerous Occurrences Regulations Act), which came into force on 1 April 1986. The Health and Safety Executive are the responsible agency who will investigate incidents and develop guidelines and regulations for prevention.

Pharmacosurveillance

In many countries voluntary reporting systems for adverse reactions to drugs and vaccines have been developed. In the UK, the Yellow Card Scheme is run by the Committee on Safety of Medicines. In the US, the **Food and Drug Administration** collects data from physicians and reports findings in the *FDA Drug Bulletin*. The **World Health Organization** has set up an international registry linked to national centers (*see* **Pharmacoepidemiology, Overview**).

Chronic Disease Surveillance

The use of mortality data for surveillance was the basis of the pioneering work of **William Farr**, who, as Compiler of Abstracts at the General Register Office from 1839 to 1879, used vital statistics to alert government and the public to health problems [15]. He developed a classification of diseases that eventually led to the **International Classification of Diseases**. The routine, timely analysis and reporting of cause and age-specific death rates continued today by the Office of National Statistics can legitimately be considered as chronic disease surveillance, as is illustrated by the London smog epidemic in 1952 [17]. Publication of the death registration totals for the week ending December 13 in London identified considerable **excess mortality**, leading to a government enquiry and eventually to the Clean Air Act. Another example is the identification of excess deaths during heat waves in the US, which has resulted in development of advice for prevention. In Russia and Eastern Europe, a sudden increase in death rates in men has been observed since 1991 [28], which highlights the utility of monitoring crude death rates in identifying chronic disease problems, and emphasizes the need for chronic disease surveillance.

References

[1] Begg, N.T. & Miller, E. (1990). Role of epidemiology in vaccine policy, *Vaccine*, **8**, 180–189.

[2] Begg, N.T., Gill, O.N. & White, J.M. (1989). COVER (Cover of Vaccination Evaluated Rapidly): description of the England and Wales schemes, *Public Health* **103**, 81–89.

[3] Farrington, C.P. (1993). Estimation of vaccine effectiveness using the screening method, *International Journal of Epidemiology* **22**, 742–746.

[4] Farrington, C.P. & Beale, A.D. (1993). Computer-aided detection of temporal clusters of organisms reported to the Communicable Disease Surveillance Centre, *Communicable Disease Report* **3**, R78–R82.

[5] Farrington, P., Pugh, S., Colville, A., Flower, A., Nash, J., Morgan-Capner, P., Rush, M. & Miller, E. (1995). A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines, *Lancet* **345**, 567–569.

[6] Fleming, D.M., & Crombie, D.L. (1985). The incidence of common infectious diseases: the weekly returns service of the Royal College of General Practitioners, *Health Trends* **17**, 13–16.

[7] Giesecke, J. (1995). The fine web of surveillance, *Lancet* **346**, 196.

[8] Grant, A.D. & Eke, B. (1993). Application of information technology to the laboratory reporting of communicable disease in England and Wales, *Communicable Disease Report* **3**, R75–R78.

[9] Haperin, W. & Baker, E.L. Jr (1992). *Public Health Surveillance*. Van Nostrand Reinhold, New York.

[10] Harrison, J. & Tyson, D. (1993). National injuries surveillance in Australia, *Acta Paediatrica Japonica* **35**, 171–178.

[11] Henry, R. & Palmer, S. (1996). Evaluation of a public health electronic surveillance network, *Health Trends* **28**, 22–25.

[12] Khoury, M.J., Botto, L., Mastioiacovo, P., Skjaerven, R., Castilla, E. & Erickson, J.D. (1994). Monitoring for multiple congenital anomalies: an international perspective, *Epidemiologic Reviews* **16**, 335–350.

[13] Klauke, D.N., Buehler, J.W., Thacker, S.B., Gibson Parrish, R., Trowbridge, F.L., Berkelman, R.L. & the Surveillance Coordination Group (1988). Guidelines for evaluating surveillance systems, *Morbidity and Mortality Weekly Report* **37**, (SS-5), 1–8.

[14] Langmuir, A.D. (1963). The surveillance of communicable diseases of national importance, *New England Journal of Medicine* **268**, 182–192.

[15] Langmuir, A.D. (1976). William Farr: founder of modern concepts of surveillance, *International Journal of Epidemiology* **5**, 13–18.

[16] Lyons, R.A., Lo, S.V., Heaven, M. & Littlepage, B.N.C. (1995). Injury surveillance in children-usefulness of a centralised database of accident and emergency attendances, *Injury Prevention* **1**, 173–176.

[17] Macfarlane, A. (1906). Daily mortality and environment in English conurbations. Air pollution, low temperature, and influenza in Greater London, *British Journal of Preventive and Social Medicine* **31**, 54–61.

[18] McCormick, A. (1987). Notification of infectious diseases: the effect of increasing the fee paid, *Health Trends* **19**, 7–8.

[19] Palmer, S.R. & Henry, R. (1992). Epinet in Wales: PHLS Cadwyn Cymru: development of a public health information system, *PHLS Microbiology Digest* **9**, 107–109.

[20] Report (1996). The incidence and prevalence of AIDS and prevalence of other severe HIV disease in England and Wales from 1995 to 1999: projections using data to the end of 1994, *Communicable Disease Report* **6**, R1–R24.

[21] Report (1996). *Unlinked Anonymous HIV Prevalence Monitoring Programme, England and Wales*. Department of Health, London.

[22] Salmon, R.L. & Bartlett, C.L.R. (1995). European surveillance systems, *Review of Medical Microbiology* **6**, 267–276.

[23] Stroup, D.F. (1994). Special analytic issues, in *Principles and Practice of Public Health Surveillance*, S.M. Tentich & R.E. Churchill, eds. Oxford University Press, Oxford, pp. 136–149.

[24] Thacker, S.B. & Berkelman, R.L. (1988). Public health surveillance in the United States, *Epidemiologic Reviews* **10**, 164–190.

[25] Thacker, S.B. & Gregg, M.B. (1996). Implementing the concepts of William Farr: the contributions of Alexander D. Langmuir to public health surveillance and communications, *American Journal of Epidemiology* **144**, (supplement), S23–S28.

[26] Thacker, S.B. & Stroup, D.F. (1994). Future directions for comprehensive public health surveillance and health information systems in the United States 1994, *American Journal of Epidemiology* **140**, 383–397.

[27] Threlfall, E., Frost, J., Ward, L. & Rowe, B. (1996). Increasing spectrum of resistance in multiresistant *Salmonella typhirmurium*, *Lancet* **347**, 1053–1054.

[28] Tillinghast, S.J. & Tchernjavskii, V.E. (1996). Building health promotion into health care reform in Russia, *Journal of Public Health Medicine* **18**, 472–473.

[29] Valleron, A.J. & Garnerin, P. (1993). Computerised surveillance of communicable diseases in France, *CDR Review* **3**, R82–R87.

[30] Wharton, M., Chorba, T.L., Vogt, R.L., Morse, D.L. & Buehler, J.W. (1990). Casexsxs definitions for public health surveillance, *Morbidity and Mortality Weekly Report* **39**, RR13–RR43.

[31] Wilson, F.P. (1927). *The Plague in Shakespeare's London*. Clarendon Press, London.

Bibliography

Berkelman, R.L., Stromp, D.F. & Buehler, J.W. (1997). *Public Health Surveillance*, 3rd Ed. Oxford Textbook of Public Health, Oxford University Press, Oxford.

Detels, R., Holland, W.W., McEwan, J. & Omenn, G.S. eds. (1997). *The Methods of Public Health*, Vol. 2. Oxford University Press, New York.

Eylenbosch, W.J. & Noah, N.D. (1988). *Surveillance in Health and Disease*. Oxford University Press, Oxford.

Stroup, D.F., Wharton, M., Kafadar, K. & Dean, A.G. (1993). Evaluation of a method for detecting aberrations in public health surveillance data, *American Journal of Epidemiology* **137**, 373–380.

White, J.M., Fairley, C.K., Owen, D., Matthews, R.C. & Miller, E. (1996). The effect of an accelerated immunization schedule on pertussis in England and Wales, *Communicable Disease Report* **6**, R86–R91.

S. PALMER

Surveys, Health and Morbidity

One of the methods to assess features of health and morbidity in a population is to conduct a specific survey by means of which information on a **target population** is obtained by measuring a representative sample of that population. Besides measuring health, surveys can also be used to investigate related variables such as living conditions, housing demands, and participation in the labor force. A health survey includes measures of health characteristics, health-related behavior, and a variety of demographic and socioeconomic characteristics. If the target population for such a survey is all persons living in a certain country, the survey is usually referred to as a national health survey. If survey data are collected through face-to-face interviews, the survey is commonly referred to as a "health interview survey". Survey methodology can be used to assess the health of many different target populations, depending on the purpose of the survey. Examples of target populations are: population living in a country; population living in a certain area (e.g. a town); population of a specific age group (e.g. the elderly); population registered in a health service register (e.g. a register of a general practitioner); a specific occupational group (e.g. nurses), or the population belonging to an ethnic group. A health survey can also be limited to one single key subject such as pain, health expenditure, or dental health.

To assess information on the health and morbidity situation of such target populations, it is usually not practicable to measure every subject of that population. In most cases this is also not necessary – **probability samples** are usually sufficient to make reliable estimates for the target population. To gather information on the health and morbidity of the subjects in the sample, structured questionnaires and/or specific physical examinations are used. Most health surveys are carried out using only structured questionnaires administered by personal interview, telephone interview, or post. Surveys that consist of physical examinations (e.g. functional assessment of lungs) and/or laboratory measurements (e.g. of blood or urine) are usually called health examination surveys. It may be desirable to combine physical examinations with personal interview data in one survey because the collected data can be complementary, although the logistics of such a combined data collection are complex.

The health survey method is very popular, and applications are found in many countries as national health surveys, pain surveys, dental surveys, health expenditure surveys, and as health surveys for specific groups such as the elderly and ethnic minorities.

Survey data are used in national, regional, and local health statistics and information systems (*see* **Administrative Databases**; **Health Services Organization in the US**) and form an important base for planning, monitoring, and evaluating public health actions. One additional benefit of health surveys is that they can be used to explore the interrelationships between health, health-related behavior (*see* **Health Care Utilization and Behavior, Models of**), use of services (*see* **Health Care Utilization Data**), and social, economic, and demographic variables.

In comparison with the other sources of health and morbidity information and in particular in comparison with health service registers, the advantages of health surveys are:

1. extensive data on both health and morbidity and use of health services can be assessed
2. data on sociodemographic and other background variables, lifestyle, and many other possible determinants are assessed for the same individuals as the health and morbidity characteristics
3. subjective data such as perceived health and knowledge of health services, coping strategies, and opinions can be assessed
4. data can be collected on many subgroups in the population, including those not having contact with health services
5. they are relatively cheap and quick.

Disadvantages of health surveys compared with health service registrations include:

1. failure to contact everybody in the sample, and, therefore, the possible introduction of **bias**
2. questionable reliability for some topics assessed by self-reports
3. detailed medical information usually cannot be collected.

The design of a survey requires many decisions regarding the information that is needed, the instruments to be used (*see* **Questionnaire Design**),

sampling methods and other methodologic procedures. Survey design decisions should be directed by the desired quality of the data that are to be collected. The quality of the data is determined by the quality of the responses and the coverage of the target population. To establish an adequate insight into the health status of a population and its determinants and consequences, every detail of design of the survey should be considered in relation to the desired data quality. After designing a health survey, the implementation phase can start, and this also requires constant check-ups and management to warrant good quality. Protocols have to be carried out in detail, and adjusted when necessary. After data are collected, the coding of responses, building of data files, analysis, and reporting of the survey findings must be carried out.

Two important aspects in the design and implementation of a health survey are the kinds of information that are needed about the target population and the methodologic characteristics of the survey. Other important considerations include analysis and reporting of the data and the required survey management.

To illustrate these aspects the national health population survey will be used. A national health population survey is a general information source for those involved in health policy analysis and development. Health surveys are essential sources of information that cannot be collected routinely through registers (*see* **Disease Registers**), health records, or other available sources. National health and morbidity data are needed to provide a better foundation for decisions on priorities for public health policy action and for the effective allocation of resources.

A number of countries have experience with national health population surveys. However, the findings are usually applicable only in the country and for the population studied; international comparisons of this type of data pose problems due to differences in the methods and instruments used. Limited international comparability also may limit the use of data at the national level, since comparisons with other countries may be useful in providing insight into cultural, environmental, and economic factors associated with health problems. These limitations can be overcome not only by using comparable survey methods but also by using comparable methods of analysis and presentation. Efforts on harmonization of methods and instruments of national health interview surveys have been described in a joint **World Health Organization**/Statistics Netherlands publication, on which this article has partly been based [4]. Although the design and implementation of specific health surveys other than the general national health survey can have additional problems not covered in this article, most of the methodologic aspects described here should be considered by every health survey professional.

Kinds of Information Assessed by Health Surveys

The kinds of information assessed by health surveys are determined by their purpose. The classic health population survey includes questions on health and morbidity characteristics, use of health services (*see* **Health Services Research, Overview**), lifestyle characteristics, and sociodemographic characteristics. An overview of the most important characteristics can be found in Table 1. Health surveys can be extended with many other relevant subjects such as psychosocial factors, environmental characteristics (e.g. noise pollution), drugs use, accidents, and sexual behavior.

Detailed specification of the desired information can be seen as the first phase in designing a survey. What information is needed? At what level of detail is information needed? The choice of the content of the survey depends on many factors, including resources, actual health problems, and the need for specific information. Priorities have to be set in this phase because the length of the questionnaire must be limited to reduce both survey costs and respondent burden. After deciding on the content of the survey, specific instruments have to be chosen. In this context, the term "instrument" refers to a set of questions (or one question) which measure the characteristic of interest. The selection of instruments and the construction and wording of the questions is very important because this is the basis for the quality of the data. Ideally, instruments should:

1. Be as short as possible, i.e. the respondent burden should be kept as low as possible.
2. Collect information on characteristics that are not too rare in the target population (characteristics are only justified to be measured in a probability sample of the population if they are relevant for a sufficiently large proportion of that population. If not, other methods are preferable, such as

Table 1 Relevant characteristics measured in health surveys

Health and morbidity characteristics
Perceived health
Diseases
Disability, impairments, handicaps
Health complaints
Dental health
Mental health
Anthropometric characteristics
Height
Weight
Birthweight
Health services
General practitioner
Medical specialist
Hospital admission
Physiotherapist (and other paramedic professions)
Preventive services (e.g. participation in screening services such as cervical smear,
Mammography, and influenza vaccination)
Dentist
Maternity care (including information on pregnancy and delivery)
District nursing
General social work
Use of contraceptives
Medicine use
Alternative/complementary medicine (e.g. acupuncture and homeopathy)
Lifestyle
Smoking
Alcohol consumption
Physical activity
Food consumption
Breastfeeding
Psychosocial factors
Personality characteristics
Coping strategies
Social support
Sociodemographic characteristics
Age
Sex
Place of residence
Social class
Education
Work situation
Income
Economic position
Housing
Living arrangements
Health insurance

oversampling of subsets of the target population known to have a higher probability of possessing the characteristic, or previously identified by their response on a screening question in a larger scale survey).

3. Be simple to administer, and provide data that are easy to process (unnecessarily complex procedures should be avoided; otherwise errors may be introduced). Most of the time questions with fixed answer categories, so-called "closed questions", are preferable because "open-ended questions" require extensive efforts in the coding phase. The fixed responses should cover every possible answer to avoid the exclusion of important responses
4. Be reliable: the reliability refers to the reproducibility of the results.
5. Be valid: validity refers to the issue of whether the instrument really assesses the information that is meant to be assessed.

Adequate selection of instruments requires extensive knowledge about available instruments in each specific area. This knowledge is often obtained by consultation with experts and by extensive literature review. When available, adoption of standard questionnaire batteries should be considered. Construction and content of such standard instruments usually have been tested extensively. Additional advantages include the increased possibility of comparing the results with other data – an important consideration in the framework of international comparisons. Most countries have standard routines for questions on sociodemographic variables. Less standardization is found for the other domains in the likely content of a health survey.

Often, survey researchers use adaptations of sets of questions from existing instruments. Arguments for adapting the original instruments include their being too long or not entirely relevant for the specific setting. Modification of existing instruments and scales requires, as in the case of designing one's own questions, extensive pretesting and assessment of reliability and validity. Most of the time many revisions are necessary before the questionnaire is ready.

Careful attention has to be paid to the precise wording of the questions and the response categories. They should be simple to understand and should have a consistent meaning for all respondents. Use of introductory texts to the questions to prepare respondents for the questions following should be

chosen carefully because the context of the questions also affects response.

Additional requirements for the choice of instruments and the development of survey questionnaires are that they should not be biased by age or sex of the respondent, or by differences in culture, language, and socioeconomic status.

For most characteristics to be measured in health surveys, a large variety of instruments is available. Especially with respect to self-reported health, sometimes referred to as **"quality of life"** measures, an enormous literature has emerged. In general, the instruments that are selected for the health survey should be in correspondence with the latest developments in the field. Several handbooks with summaries of available instruments and evaluation of their quality (i.e. validity and reliability) are available [10]. However, there are often no internationally agreed standard instruments available. For national surveys that will be used in international comparisons this is a significant problem because the comparability becomes limited or impossible. Several efforts have been undertaken to reach international harmonization of methods and instruments in national health surveys.

A joint action of the World Health Organization Regional Office for Europe and Statistics Netherlands has led to the first step in harmonization of survey instruments on general health characteristics for the countries of Europe. These efforts were guided by the wish to have internationally comparable data for WHO's "Health for All" indicators [17]. A number of recommendations for common instruments which measure these indicators have been brought forward [4]. Some of these recommended instruments are presented here to give some examples of the possible content of a health survey. These illustrative examples of instruments deal with the first four domains of the content of a health survey as shown in Table 1: health and morbidity characteristics, anthropometric characteristics (*see* **Anthropometry**; **Growth and Development**), the use of health services, and lifestyle characteristics.

Health and Morbidity

The health and morbidity characteristics are part of the core characteristics measured in health surveys. The most important indicators are: subjective health assessment, disability, chronic conditions, and mental health.

Table 2 Perceived health

How is your health in general?
• Very good
• Good
• Fair
• Bad
• Very bad

Source: [4].

Subjective or *self-perceived health* is a principal health characteristic assessed by health surveys, an example question being shown in Table 2.

Even in its simplest form (see example) and despite its very general, seemingly subjective, character, perceived health is strongly associated with a number of health problems and the use of health services. It is also a strong predictor of survival in elderly people.

Many more extensive instruments measuring self-perceived health, or health-related quality of life measures, are available. These instruments usually distinguish several domains of interest, such as well-being, quality of life, life satisfaction, pain, functional disability, and handicap. Examples of such health instruments are: Nottingham Health Profile [8], the Sickness Impact Profile [3], and the MOS 36-item short-form health survey (SF36) [14]. Sometimes such instruments are referred to as generic health instruments to emphasize the distinction between them and the many disease-specific health instruments available.

These general health measures usually also have items to measure several kinds of disabilities, impairments, or handicaps. It is important to distinguish between temporary disability and long-term disability. *Temporary disability* refers to temporary restriction in an individual's level of functioning. Information on temporary disability is usually obtained by questions about days of restricted activity and bed-days. Measurement of the period of time, together with some notion of the severity of disability, can provide information on the time lost to ill-health in the society (see Table 3).

Because of changes in public health in relation to chronic diseases and the aging of the population, *long-term disability* has become an important concern to public health. It refers to long-term limitations in major daily activities. Especially for the

Table 3 Temporary disability

Think about the two weeks ending yesterday. Have you cut down on any of the things you usually do about the house, at work or in your free time because of illness of injury?
• Yes [ask questions (a) and (b)]
• No
(a) How many days was this in all during these two weeks, including Saturdays and Sundays? (01–14)
(b) On how many of these days were you in bed for all or most of the day? (00–14)

Source: [4].

elderly, in which morbidity is often characterized by multiple pathology, long-term disability is useful as an overall indicator of health problems associated with disease. Since the 1960s a large number of instruments have been developed for the assessment of long-term disability. In 1980 the International Classification of Impairments, Disabilities and Handicaps (ICIDH) [15] was introduced. This classification is a basic conceptual scheme, that has been used as a guide for the further development of instruments. Important domains of disability are: locomotion, self-care, continence, hearing, and vision. Self-care disabilities include problems with dressing, washing, feeding, and using the toilet. Instruments differ in their number of disability domains, the levels of severity, and the specification of details. In addition it can be important to distinguish between capability and performance: is it necessary to assess whether persons cannot or do not carry out an activity. Survey data on disabilities are also required to measure the *disability-free* **life expectancy** of a population [12].

Health surveys are also useful in providing data on the **prevalence** of diseases such as cancer, cardiovascular diseases, rheumatic disorders, respiratory disorders, and mental health disorders. These, often chronic, diseases are important because they are often accompanied by pain, suffering, inconvenience, and loss of physical capacity. They put pressure on health services and society in general, especially in the industrialized countries. Commonly used instruments for the assessment of the incidence and the prevalence of physical chronic conditions are still lacking. Most countries have their own list of diseases that are considered suitable for self-reporting. The available instruments show great variety in methodology, e.g. differences in the nature of the diseases, in the number of the diseases, and in the definition of severity and in the wording of the questions. Conditions like hypertension, asthma, bronchitis, thyroid disorders, diabetes, chronic skin condition, chronic heart disease, chronic cystitis, chronic dental problems, chronic back problems, arthritis, and stroke can be part of a general list of diseases measured in health surveys. For some specific disorders, such as cardiovascular diseases, internationally used survey instruments are available [13]. The wording of the questions must be based on the respondent's ability to understand the described condition. In some cases, the disease name or a popular synonym in the specific language is sufficient; in others, additional questions or lists of symptoms are necessary. For each condition measured in terms of diagnosis, respondents should be asked whether a health professional has made the diagnosis. Several methodological aspects on the survey assessment of diseases have been investigated. These include **reliability** and **validation** studies and the effects of construction of the instruments. For instance, using an "open" question such as "Do you suffer or did you last year suffer from a disease, and if as what disease(s)?" gives different results than asking the same question but showing a list of all diseases from which one can choose ("one-by-one method"). In addition, it has been shown that many diseases can be adequately assessed in a health survey. However, many methodologic problems are still to be tackled to reach internationally agreed survey instruments for many chronic diseases.

In designing or selecting an instrument to measure morbidity one should bear in mind that it is preferable to be able to summarize the results on disease prevalences into broad ICD-10 [16] categories (*see* **International Classification of Diseases (ICD)**).

In selecting diseases and the instruments to be used for population survey assessment, it is also important to distinguish between chronic mental conditions and chronic physical conditions. Of the more than 120 *mental* diagnoses that are now distinguished [2, 16], a selection of relevant disorders has to be made for inclusion in the health survey. On the basis of prevalence, severity, and duration, disorders like dementia, mental retardation, anxiety disorders, schizophrenia, and affective disorders are important for establishing public health policies. Because the knowledge of respondents about these diseases is

generally poor, instruments have to be based on a symptom approach. Such an approach is very time-consuming, as a large number of symptoms have to be checked to diagnose a mental disease. To reduce time and expense, a two-stage procedure should be considered: some screening questions followed by an extensive interview procedure. Apart from general mental health measures, such as the GHQ-12 [7], assessment of mental disorders can only be done in specialized surveys.

Anthropometric Characteristics

Data on *weight* and *height* are commonly assessed by health surveys, for instance to analyze problems related to obesity. Obesity is usually defined by the body mass index (BMI) or Quetelet's index – weight (in kilograms) divided by the square of height (in meters). Most respondents are able to state weight in kilograms and height in centimeters. Self-reported weight and height show small but systematic errors: height tends to be overstated and weight under reported. If precise information on the characteristics is necessary, actual weight and height instruments are needed.

Birthweight is a commonly used indicator for the nutritional and health status of the newborn. It is considered to be an important determinant of the survival of the infant and its ability to develop normally. A low birthweight is commonly defined as one less than 2500 g. Administrative records such as birth registration or maternity records are usually the main source for this information, but if linking of these external records to the survey questionnaires cannot be done, a question on birthweight should be included in the survey (see Table 4).

Table 4 Birthweight

1. Is the child twin or triplet?
   - Yes (multiple birth)
   - No
2. Was the child born before it was due?
   - Yes
   - No (go to question 4)
3. Was it less than one month before it was due or more than that?
   - less than one month
   - more than one month
4. How much did the child weight at birth? (record in grams)

Source: [4].

Health Services

Health surveys can be used to measure the level of use of health services. For health services for which good registers are available, it is sometimes seen that the health survey data on service use does not completely represent the total service use: on the basis of health surveys an underestimation of health services is sometimes seen. However, for many services there are no registers available, so the only tool to provide insight into service use is the health survey. In addition, the health survey allows the study of the relationships among the use of different health services and the relationships of measures of health, lifestyle, and sociodemographic characteristics to health services use. Another important advantage of a health survey is having data on the characteristics of nonusers of health services; for instance, to be able to investigate equality of access to health services. Also high users of health care can be of specific interest because it should be known to what extent the health care dependency is justified.

Beside the measurement of use of services and reasons for use, it can also be relevant to assess the knowledge of persons on specific health services or the availability of services – for instance, preventive services (*see* **Preventive Medicine**).

The actual measures to assess the different characteristics of health service use can be very different. It is, for instance, very important to evaluate the reference period (which should not be too long in order to prevent **recall bias**) and the definition of "use of" (or consultation). An example of questions on quantitative information regarding the consultation of a general practitioner (GP) is shown here as an illustration. With these questions the percentage of persons consulting in 1 year and the number of consultations in 1 year can be derived. Such questions can be followed by questions on reasons for consultation, where or how the consultation took place, whether there was a referral to a specialist, a hospital, or otherwise. In countries where there is no clear distinction between GPs and specialists, other questions should be used (e.g. using the term "medical doctor" instead of "GP"; see Table 5).

Table 5 Consultation of general practitioner

Introduction: "The following questions concern contacts with your GP. They relate to visits during surgery hours and house-calls, but also to telephone calls for other reasons than to make an appointment."

. . . . . .

1. How often have you consulted your GP during the past two weeks ending yesterday, so since . . . . . (date)?
   - . . . . . times (go to question 3)
   - no one single time (go to question 2)
2. When did you last consult your GP?
   - on . . . . . (date) or (if date not known) . . . . . weeks/months/years ago
   - never

   If last consultation <2 months ago go to question 3.
3. How often have you consulted your GP during the past 2 months, so since . . . . . (date)?
   - . . . . . times

Source: [4].

Lifestyle

To assess data on lifestyle and to study the relationship between lifestyle characteristics and health and morbidity, the health survey is an essential tool. For instance, studying the health effects of smoking, food consumption, and physical activity has expanded to specific scientific areas requiring specialized knowledge on measurements, analyses, and interpretation of these topics. For both food consumption [18] (*see* **Nutritional Exposure Measures**) and physical activity [1, 5], many instruments are available and evaluated in numerous review articles and handbooks. However, it has been difficult to agree on standard instruments internationally.

Smoking is seen as the major cause of lung cancer, ischemic heart disease, chronic bronchitis, and emphysema (*see* **Smoking and Health**). Health surveys are an important source of data on smoking behavior. They can supply information on proportions of daily smokers, occasional smokers, ex-smokers, and those who have never smoked. Other relevant parameters include number of cigarettes and other tobacco products used per day, total numbers of years smoking, whether a person has reduced smoking, how long ago a person stopped smoking, attempts to stop smoking, and opinions on the harmfulness of tobacco. The possibilities on measuring passive smoking, for instance by asking questions on exposure to tobacco smoke at work and at home, should also be considered (see Table 6).

Table 6 Illustration of a minimal set of health survey questions on smoking

1. Do you smoke?
   - Yes, daily
   - Yes, occasionally (go to question 3)
   - No (go to question 4)
2. How many cigarettes do you usually smoke on average each day?
   - Does not smoke cigarettes
   - Fewer than 20
   - 20 or more (heavy smoker)
3. Compared to two years ago would you say you now have reduced smoking?
   - Yes (end)
   - No (end)
4. Have you ever smoked?
   - Yes, daily
   - Yes, occasionally
   - No (end)
5. How long ago did you stop smoking?
   - Less than two years ago
   - Two years ago or more

Source: [4].

Questionnaire Construction

After selection of the necessary information and the choice of instruments, the health survey questionnaire has to be constructed.

The questions and instruments should be *ordered* in a logical way, determined by psychological and behavioral knowledge. In addition, the transitions from one set of questions to the other should be

clear, to indicate the content of the questions in a friendly but impersonal way. The actual introductory texts used to prepare and inform the respondent on the questions following should be carefully worded and standardized.

It is desirable to ask the respondents only those questions in the interview schedule that are *relevant* for them. For example, a man should not be asked about diseases that only affect women, and vice versa, and a college student should not be asked what profession he or she has. Besides the use of completely different questionnaires for different groups of respondents, special *routings* or questionnaire skip patterns are often used. Such routings are also necessary when in some sections of the questionnaire a procedure involving two or more stages is used. Complex routing of questions makes high demands on the design and layout of the questionnaire and increases the burden on the interviewer or, in the case of self-administered questionnaires, the respondent. The use of **computer-assisted interviews** for data collection in the field of health surveys has resolved many problems with complex routing, although it requires a heavy input of skilled resources in the preparatory phase.

During face-to-face interviews, *showcards*, that include the set of answers to a question, can be used. The showcard should be given to the respondent the moment the interviewer asks the specific question. This procedure can be used for two reasons:

1. to inform the respondent about the response possibilities, and
2. to encourage the respondent to give the correct rather than a socially desirable answer. This is facilitated when answers are precoded on the card, and the respondent needs only to reply with the appropriate code.

The extent to which answers have to rely on the *memory* of the respondent should be minimized, because recall from memory can be a source of bias. The magnitude of recall bias depends on the length of the recall period and the saliency of the events to be recalled.

Questionnaire length is least restricted in the face-to-face interview because people accept long interviews better than long self-administered questionnaires. Nevertheless, the length of a questionnaire is not unlimited because respondent burden should be as low as possible, i.e. to prevent reduction in motivation and tiredness. Interviews of 60–90 min are not uncommon. For telephone interviews (*see* **Telephone Sampling**) the duration is usually no longer than between 30 and 50 min, but is preferably shorter.

Mail questionnaires should not be longer than 12 pages; in general, most range between four and 12 pages. Longer questionnaires can be considered when the respondents are highly motivated. In general, respondents have to be rewarded for their participation; they have to be "pampered" and continuously motivated for their participation in a health survey. For self-administered questionnaires, the layout of the questions should be as attractive as possible. Easy to read type fonts, different colors, frames, "flowcharts", etc. should be used.

Health Examination Survey

The kind of information assessed by a health survey can be extended by information gathered by physical examinations. These examinations can be carried out by trained interviewers. Sometimes trained nurses, physicians, or laboratory personnel are necessary. Health examination surveys may include physical examinations (e.g. anthropometric measurements), functional assessment (e.g. ability tests and spirometry), laboratory tests on blood or urine samples (e.g. cholesterol level in blood), and the application of specific imaging techniques such as X-ray and MRI. Adding such measurements to a health survey requires extra effort in survey design, conduct, and analysis, and there will be increased respondent burden and extended duration of the survey.

Some examinations can be carried out by trained interviewers visiting respondents at home. Examples of such examinations are weight and height measures, blood pressure tests, and mobility tests. Other examinations require specific instruments that are not suitable for routine assessment at home, so specific arrangements have to be made. Respondents can be invited to visit a laboratory or clinic in order to be able to carry out the examinations. An alternative is the use of a mobile laboratory which can visit persons at home; the geographic spread of potential respondents should not be too large, otherwise such an approach would not be cost-effective.

Although the costs are very high, the general advantages of a health examination survey include the acquisition of objective medical data from

a **population-based study** sample, which can be important to assess population norms, and the possibilities to validate questionnaire-assessed data. An example of a study that combines a health interview and a health examination survey is the US National Health and Nutrition Examination survey [11].

Methodologic Considerations

There are many ways to design a health survey. The most important components of survey methodology are highlighted here. Methodologic aspects are broadly divided into those concerning the population and in those concerning data collection. All aspects of the survey process that may affect response quantity and quality should be taken into account. However, most decisions about survey design will depend on the resources available. Therefore an adequate balance between costs and quality will have to be pursued.

Survey Population

The definition of the *target population* depends on the purpose of the study and the practical possibilities of the available **sampling frames**. For a national health survey the sample should represent the general population. To get a representative sample for the target population concerned it is necessary to have a data file with all information of the subjects of the population, including information on how to locate them. The sample can be drawn by several methods, for example, by using:

1. address or postal files
2. electoral registers
3. population registers
4. telephone registers.

If available (and accessible as public records for sampling purposes), population registers should be used to obtain a representative sample. Data on name, age, sex, and address of all persons legally living in a country or a subarea are known within a population registry. An additional advantage of sample frames that include data on age and sex is the opportunity for **stratified sampling** – for example, **stratification** by age group and sex. Population registers seem to be nearly complete in some countries, but alterations such as changes of address may only be entered slowly, resulting in registers that are not fully up-to-date. The main limitation of the electoral register is that as a rule only those people are listed who are of legal voting age and who meet other voting requirements. Alternate sampling frames, e.g. address and telephone registers, have the disadvantage that not every person has the same chance of being selected. For those files the *sample unit* is the household living at the address or telephone number selected (*see* **Unit of Analysis**). When everybody in the household is interviewed, a maximum of, for example, four persons is often used to reduce "household" burden (*see* **Cluster Sampling**). Sometimes one person is interviewed on all health characteristics of the persons belonging to the household. If that is the case, the one who knows most about the health of the family is questioned. If a random choice of one person in the household is needed, then the selection can be decided by the one who is the first to have their birthday. If one person per household is invited to participate, those living in single households have a greater chance of being selected than those living in a household with more members. When there are unequal selection probabilities, the final responses have to be corrected by weighting procedures. Another disadvantage of address and telephone samples is the inclusion of inappropriate units, such as businesses, and the exclusion of certain subjects, such as the homeless. Those living in institutions such as nursing homes, homes for the handicapped, and prisons are also more difficult to reach due to sample frame limitations.

The actual sampling can have very different manifestations. A probability sample is to be preferred because precision of the sample estimates can be calculated. The strategy of generating a sample may involve several stages (*see* **Multistage Sampling**). For instance, when using address files for the sample and interviewers visiting subjects at home, first a sample of municipalities can be drawn before sampling the addresses. Such a strategy reduces the travel time of the interviewer.

The decision about *sample size* depends on many factors, including the required level of detail of the results; the finer the detail, the greater the sample size needed to provide estimates with acceptable confidence limits. This approach of calculating the necessary sample size (*see* **Sample Size Determination**) on the basis of sampling errors has two limitations.

First, most surveys are designed to make numerous estimates, and the necessary precision will vary. Secondly, in health surveys the sampling error is not the only or main source of error in a survey estimate. The decision will in most cases not depend on estimates for the total population but should depend on the estimates for the smallest subgroups of importance. The sample also depends on the expected percentage of individuals in the sample for which no data will be collected – the nonrespondents (*see* **Nonresponse**).

Nonresponse

There are several reasons why subjects eligible for the health survey are not measured (*see* **Missing Data in Epidemiologic Studies**). Groups of nonparticipating subjects include:

1. Subjects who cannot be reached – those who cannot be located because they have moved or have died. This number of nonresponders due to sample-frame defects should be determined to be able to calculate the *net response*.
2. Subjects who do not react, the not-at-homes or all those from whom no reaction is received. Efforts to reduce this group of nonresponse include increasing numbers of attempts to contact them by visits, reminders, and (multiple) telephone calls.
3. Subjects who refuse to participate. This is a difficult group because most of the time there is hardly any information on reasons for nonparticipation and so there is no possibility of tackling the question of **selection bias**. However, several approaches can be explored to get information on them, including the use of a small questionnaire with information or a preprinted refusal card with some additional questions on reasons for refusal and some additional information, e.g. perceived health and smoking habits.
4. Subjects who are unable to participate because they are too ill, do not speak the language of the interviewer, or whose reading and writing skills exclude them from filling in self-administered questionnaires. If specific subgroups are important for inclusion in the health survey, specific efforts are necessary to get information on such subjects, e.g. using proxy questioning.

Increasing *motivation for participation* is one of the most difficult and important aspects of a health survey. Traditionally, response figures for health surveys are high because health and morbidity belong to the most important aspects of life. Unfortunately, like other surveys, health surveys are experiencing reduced response rates in many countries.

General guidelines on increasing motivation for participation include (see, for example, [9]):

1. the topic should be of interest to potential respondents
2. the **confidentiality** of survey responses should be clear
3. advance contacts should be made
4. specific incentives should be used, e.g. gifts, money, or reporting back on the health status of the respondent
5. repeated visits, reminders, or telephone calls must be employed when appropriate (*see* **Callbacks and Mail-backs in Sample Surveys**)
6. use of trained interviewers, attractive questionnaires, etc.

Besides these points, the approach to the potential respondent is essential to their motivation. For interview surveys, an informative advance letter should be sent including information on the purpose of the study, why it is important, that confidentiality is assured, that it is approved by official institutions, how the person was sampled, and why it is important that they will participate. The advance notification should consist of an official letter, preferably signed by hand, to include the most important information, and an attractive and informative flyer about the research project and the participating institutions.

In spite of costly efforts to reduce nonresponse, not all eligible sample subjects will participate. When presenting results, the quality of the response group should be indicated by

1. response rate
2. estimates of nonresponse bias
3. corrections for nonresponse
4. describing to what extent the response population is representative of the target population.

In reporting on survey data, one of the key numbers readers look for is the *response rate*. In its most general form the response rate is the number of respondents divided by the number of people sampled, usually expressed as a percentage. If the denominator is reduced by those who could not be

reached for legitimate reasons, the response rate is usually higher. This *net response* figure is usually reported.

Although a low nonresponse is desirable it should always be considered whether it is likely or not that bias is introduced by the nonresponse, and what the possible size of that bias is. It is even possible that a survey with a 90% response rate is more biased than a survey with a 70% response rate. It depends on the selectivity of the response group with regard to the subject(s) of interest. "One usually does not know how biased nonresponse is, but it is seldom a good assumption that nonresponse is unbiased" [6].

Insight into nonresponse bias is only possible when there is information on the nonrespondents. In general a comparison of the response group and the nonresponse group on the basis of the sampling frame information is carried out. This information includes residence, and sometimes also age and sex. To correct for differences in these variables *weighting strategies* can be applied. In some cases, a *nonresponse study* has to be considered to measure nonresponse bias. This can be carried out by trying to contact (a random sample of) the nonresponders, using extreme extra efforts.

The main strategy to adjust for nonresponse is weighting adjustment. The response group is weighted in such a manner that the resulting response group has the same distribution of general sociodemographic characteristics as the sample or the target population. Evaluation of the *representativeness* of a response group depends on both the quality of the sample frame and the quality of the response group. One specific form of nonresponse is the item *nonresponse*, which refers to the part(s) of the questionnaires that are not answered or filled in by every respondent. This problem can be minimized by using the personal interview technique, but it is very common in self-administered questionnaires. Numerous missing data usually give big problems during analysis. Some methods of imputation are available and should be considered to reduce missing data (*see* **Missing Data**) [9].

Data Collection

An essential decision in the design of a health survey is the choice of the method of data collection, which could be either face-to-face interviews, telephone interviews, self-administered questionnaires, or a combination. Interviews can be conducted during visits to respondents at home, during the visits of respondents to a specific laboratory, or using the telephone. Self-administered questionnaires can be sent by post or can be an addition to an interview. The choice depends on the purpose of the study, sample frame, research topic, characteristics of the sample, and available resources.

Face-to-Face Interviews and Self-Completion

Personal interviews or face-to-face interviews carried out by trained interviewers who visit respondents at home and who ask questions and assess the answers by means of a structured questionnaire are generally considered to be the preferred mode of data collection (*see* **Interviewing Techniques**). Advantages include good response rates, and the questionnaires are usually filled in more completely than with other methods.

Furthermore, people will tolerate relatively long interviews much better than very long self-administered questionnaires, and the interviewers can have direct control of the quality of the response. For some topics, however, it may be useful to introduce some type of self-completion by the informant. Those measures are preferred when the questions refer to sensitive subjects, such as alcohol, drugs, contraception, sexual behavior, and/or when it is difficult to ensure privacy for an interview. One method is to introduce a self-administered questionnaire during the interview and to allow the informant time to complete it before carrying on with the interview. Another is to leave the questionnaire behind after the interview and collect it later.

Owing to some specific disadvantages of face-to-face interviews (*see* **Interviewer Bias**), the methodology of alternatives has gotten a lot of attention. These disadvantages include frequent long periods of fieldwork, high rates of "not at homes" because more persons in one household work and/or spend more leisure time outside the home, and high rates of persons who have moved. In addition, the safety of interviewers cannot be guaranteed in some areas. Furthermore, the alternatives – postal questionnaires and telephone interviews – are relatively cheap.

Postal and Telephone Surveys

Postal and telephone surveys are cheap alternatives for face-to-face interviews. Telephone surveys have

the advantages of use of interviewers, e.g. reduction of item nonresponse and control of data quality, but are limited to those persons who have a telephone, and there are measurement constraints: no visual aids or complex response categories can be used. For postal surveys, such visual aids and complex response categories can be used, and the response is the most anonymous of all data collection methods, which is preferred for sensitive topics. However, there is less control of data quality, and good reading and writing skills of the respondents are required.

Interviewers

The interviewer has the important task of tracking down the persons to be interviewed, taking care of their participation, and controlling the data quality. Considerable attention has to be paid to selection, training, and supervision.

Proxy Informants

In cases where a person acts as an informant for others, the term proxy informant is used. This is a useful approach if information on a household is needed and one person can provide all the answers. Proxy informants are usually essential for children and persons with mental or sensory disabilities, and also for the reduction of nonresponse where people are difficult to contact – for instance, married men and young single adults. In all these cases a spouse or significant other – partner or parent, or other family member (brother or sister) – can provide the necessary information. For some indicators the use of a proxy informant is not totally appropriate, such as for questions on feelings, opinions, and knowledge. Information on most other health survey assessed characteristics, such as obvious health problems and use of health services, can usually be provided adequately by proxies. Opinions on the role of proxy interviewing in relation to features such as alcohol intake and smoking differ substantially. On the one hand, it can be argued that proxy questioning may lead to increased, and possibly truer, estimates of consumption than self-reporting in areas where these habits are considered socially undesirable. On the other hand, proxies will not always know fully such habits of another person. Depending on the purpose of the survey and the cultural characteristics of the population to be investigated, the use of proxies can be considered, especially when it helps in reducing nonresponse. The philosophy of proxy use differs considerably by country. While in the UK proxy interviews are only accepted as a last resort, in France and the Netherlands they are used as a standard practice.

Specific Groups

The use of general methods can result in underrepresentation of some specific groups in the population. This can be due to the fact that they are more difficult to reach or because methods and instruments should be adapted to the specific abilities of specific groups. Problems can be caused by specific lifestyles or living situations, by language or cultural differences, and/or by their age. Groups that always have to be considered because general methodology can lead to an underrepresentation of them are the institutional population, children, elderly, homeless persons, and **ethnic groups**.

Because a lot of national surveys are based on address samples, it is usually the case that all nonprivate establishments are excluded. This means that people living in nursing homes, hospitals, prisons, hostels, and other places, such as some type of student and nurses' accommodation, the so-called *institutional population*, are excluded from the sample. Although they only make up a small proportion of the population, from a health point of view these people may be very different from people in private households, particularly regarding health problems such as dementia and long-term disability. To get national data, for instance, it can be recommended that the size of this part of the population has to be assessed adequately, and depending on the specific purpose of the study specific subsurveys should be considered. For *children* the most important source of information is the parents. Usually, separate questionnaires are necessary. If it is desirable to question children themselves, for instance with respect to topics like smoking and drinking, specific attention should be paid to a confidential environment and children should not be encouraged to exaggerate. Parental consent to interview children may be required or advisable.

Many countries have significant minorities in the population who may not be fluent in the main language. The use of questionnaires in different languages, interviewers with adequate knowledge of

alternative languages, or interpreters should be considered when the expected effect of excluding these groups is not marginal. There are some countries in which questionnaires in two main languages are essential, e.g. Canada and Belgium. In general there is some loss of standardization in questions when different languages are used – it is not always possible to find words or phrases with precisely the same meaning. In addition, the importance of health and health-related problems can differ between cultures, which may also hamper the comparability of data.

Collection Period

Many topics related to health, such as use of health services, accidents, and temporary disability are influenced by the period(s) in the year in which the fieldwork takes place. To rule out such seasonal effects, data collection should be spread evenly throughout the year, also including spreading evenly across weekdays. When analysis of trends is considered more important than average figures per year, an alternative is to collect data in the same period each year.

Frequency of Survey

Although a one-shot health survey can provide relevant data, some regularity – continuous or repeated – in assessment is usually relevant, especially to monitor changes in health status and health-related factors and to monitor effects of public health decisions.

If some regularity is needed, the costs and benefits of a continuous or regularly repeated survey, for instance every 3 years, should be carefully considered. For many health aspects, large differences are not likely to be observed from year to year. However, for any large-scale survey, the design and start-up costs will be great because large numbers of interviewers and other staff have to be recruited and trained. For this reason continuous surveys may be more cost-effective than repeated surveys. One option would be to have a continuous survey, with core questions asked each year and with a rotating element containing other items at regular intervals in turn. Additional advantages of such procedures are that new health topics can be added when necessary, and that the rotating elements can also include other national survey areas.

Data Processing and Presentation

After collection of the data they must be transformed into an appropriate form for analyses using computers. Relevant phases include deciding on the format of the data, data entry, data cleaning, weighting adjustment, and analysis of data to present the results (*see* **Data Management and Coordination**).

The format of the data in the computer refers to the way the data are organized: what kind of codes are necessary for the responses and what kind of variables should be used for what kind of questions? Information on each question is usually represented in at least one variable. A question with multiple responses can be represented by many variables.

After deciding how the data need to be organized, they should be entered into the computer. Because usually many respondents are questioned and a lot of information per respondent is assessed, the number of codes to be entered is very large. When data have to be entered by hand, specific software to assist the data entry should be used. When interviews are carried out using (laptop) computers, the data collection phase and data entry phase are the same. An alternative for data entry when written questionnaires are used is a method of computer optical scanning of the responses. These scanning procedures make specific demands for questionnaire layout. Data entry by hand can result in many errors, and with the use of computer-assisted personal interviewing such errors can be minimized.

The phase of data cleaning consists of checking the data file on inconsistencies and removing traceable errors. Three types of errors are usually distinguished: a range error (e.g. an age of 234), a consistency error (e.g. a person 6 years old and married), and a routing or skip pattern error.

Weighting adjustments may be needed to correct for unequal selection probabilities in the sample and to correct for nonresponse bias. This implies that not every respondent contributes equally to the results; the contribution depends on the specific subgroup to which the respondent belongs (most of the time based on characteristics such as age group, sex, household composition, marital status, and region).

To present the results adequately, different forms of data analysis are necessary. Many strategies are possible to make adequate tabulations of results and to analyze associations between the features assessed.

Survey Management and Logistics

The number of interviewers needed for conducting the fieldwork of a health survey largely depends on the content, the sample size, and the time period available. The usual face-to-face interview survey contains questions that can be handled by lay interviewers who have no medical skills. But when the survey includes physical examinations, these should often be left to appropriate staff (nurses, doctors, and laboratory personnel).

The number of full-time and/or part-time interviewers needed to perform an interview survey within a given period of time can be calculated when (i) the estimated duration of one interview (usually no longer than 30–45 min, to be tested in the pilot phase), (ii) the number of working days available for the fieldwork, and (iii) the net sample size are known. It should be kept in mind that interviewers may need a lot of time for traveling and other activities (phone calls, etc.) for making contacts with the respondents; time lost due to nonresponse should also be taken into account as far as possible.

For face-to-face interviews the administrative tools used to be "paper and pencil", but in recent years many of the larger survey organizations make use of laptop computers for conducting the fieldwork. Several software packages have been developed for managing the questionnaires in the laptops. The main advantages of using laptops in the field are that the actual entry of the data is done without extra effort by the interviewers (instead of afterwards at a central location by special staff) and above all that the number of administrative errors is reduced because the data can be checked and corrected interactively.

The number of scientific and administrative staff needed for an interview survey largely depends on the "history" of the survey, content and length, and the time available for preparations, analysis, and publication. A completely new survey needs more time in preparation than one that is more or less a repetition of an earlier effort. The number of organizations, ministries, and other institutes that have a say in the survey may also be an important factor of influence on the time needed for preparation and other activities. When a fairly large part of the questionnaire is not "standard" or "routine", the (scientific) staff time needed for developing and testing instruments for measuring a new topic is at least in the range of a few man-months. Once the questionnaire is ready, the scientific staff should also be involved in training the interviewers for the survey at hand (taking for granted that the interviewers have sufficient general experience).

Scientific staff should be available during the pilot phase of the fieldwork for answering the questions of interviewers and if necessary for adaptations to the questionnaire. During the fieldwork period, staff should be available for data entry and/or data checking and cleaning; preparations for the tabulations and further analyses may also be made during this period. In other words, preparations for analysis and publication should not wait until the fieldwork is completed, and staff for building the data file should be active from the very beginning. For timely reporting it is essential that a complete file becomes available shortly after conclusion of the fieldwork. Specialists in sampling and weighting – if not present in the scientific staff for the survey – should be hired at the time they are needed for the respective activities. Nowadays the technical equipment for the analysis of survey data, even if these consist of hundreds of variables and thousands of records, is no longer a problem: the storage capacity and speed of relatively cheap desktop computers is sufficient for timely and advanced analysis by means of software like SAS and/or SPSS (*see* **Software, Biostatistical**). But experienced scientific staff are needed to analyze the data and publish the results in a way that is practical and understandable for the users. In practice, the procedure may be that first the bare essentials are analyzed and published, followed by more profound analysis and publication at a later stage. Staff are also needed for making a well-documented micro data file – with sufficient protection of the privacy of respondents – for users who are willing and capable to make their own specific analysis of the data.

Finally, considerable workload may result from first and later publications of the results: questions coming from the users of the data need to be answered; additional requests for tabulation and analysis call for extra efforts that – even if additional costs are covered – need to be met by experienced staff. Although exact estimates cannot be given (all depend on the content, length, sample size, and "environment" of the survey), it is clear that any substantial health survey needs to be run by a more or less permanent staff, consisting of scientific, technical, and administrative members.

Acknowledgment

The authors wish to thank G. Van den Berg and G.T.P Bonte (Statistics Netherlands) for their comments. The views expressed in this article are those of the authors and do not necessarily reflect the policies of the National Institute of Public Health and the Environment and/or Statistics Netherlands.

References

[1] Ainsworth, B.E., Montoye, H.J. & Leon, A.S. (1994). Methods of assessing physical activity during leisure and work, in *Physical Activity, Fitness, and Health*, C. Bouchard, R.J. Shephard & T. Stephens, eds. Human Kinetic Publishers, Box 5076, Champaign, pp. 146–159.

[2] American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders*, 3rd Ed., revised (DSM-III-R). Washington.

[3] Bergner, M., Bobbitt, R.A., Carter, W.B. & Gilson, B.S. (1981). The sickness impact profile: development and final revision of a health status measure, *Medical Care* **19**, 787–805.

[4] Bruin, A.de, Picavet, H.S.J. & Nossikov, A., eds (1996). Health Interview Surveys. Towards International Harmonization of Methods and Instruments, *WHO Regional Publications, European Series No. 58*. WHO Regional Office for Europe, Copenhagen.

[5] Caspersen, C.J. (1989). Physical activity epidemiology: concepts, methods and applications to exercise science, *Exercise and Sport Sciences Reviews* **17**, 423–473.

[6] Fowler, F.J. (1993). *Survey Research Methods*, 2nd Ed. Sage, Newbury Park.

[7] Goldberg, D. & Williams, P. (1988). *A User's Guide to the General Health Questionnaire*. NFER-Nelson Publishing, Windsor.

[8] Hunt, S.M., McEwen, J. & McKenna, S.P. (1986). *Measuring Health Status*. Croom Helm, London.

[9] Kessler, R.C., Little, R.J.A. & Groves, R.M. (1995). Advances in strategies for minimizing and adjusting for survey nonresponse, *Epidemiologic Reviews* **17**, 192–204.

[10] Mc Dowell, I. & Newell, C. (1987). *Measuring Health, a Guide to Rating Scales and Questionnaires*. Oxford University Press, New York.

[11] National Center for Health Statistics (1994). Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–1994, *Vital and Health Statistics, Series 1, No. 32*. US Government Printing Office, Washington.

[12] Robine, J.M. (1989). Estimating disability-free life expectancy (DFLE) in the western countries in the last decade: how can this new indicator be used?, *World Health Statistics Quarterly* **42**, 141–150.

[13] Rose, G.A. & Blackburn, H. (1968). *Cardiovascular Survey Methods*. WHO, Geneva.

[14] Ware, J.E. & Sherbourne, C.D. (1992). The MOS 36-item short form health survey (SF-36), *Medical Care* **30**, 473–483.

[15] WHO (1980). *International Classification of Impairments, Disabilities and Handicaps: a Manual of Classification Relating to the Consequences of Disease*. WHO, Geneva.

[16] WHO (1992). *International Statistical Classification of Diseases and Related Health Problems*, 10th revision. WHO, Geneva.

[17] WHO (1993). Health for All Targets. The Health Policy for Europe, *European Health for All Series, No. 4*. WHO Regional Office of Europe, Copenhagen.

[18] Willett, W. (1990). Nutritional epidemiology in *Monographs in Epidemiology and Biostatistics*, Vol. 15. Oxford University Press, New York.

Bibliography

Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

Cartwright, A. (1983). *Health Surveys in Practice and in Potential: a Critical Review of their Scope and Methods*. King Edward's Hospital Fund, London.

Dean, K., ed. (1993). *Population Health Research-Linking Theory and Methods*. Sage, London.

Dillman, D.A. (1978). *Mail and Telephone Surveys, the Total Design Method*. Wiley, New York.

Fink, A., ed. (1995). *The Survey Kit* (Series of nine volumes). Sage, London.

(*See also* **Bias, Overview**; **Call-backs and Mail-backs in Sample Surveys**; **Data Quality in Vital and Health Statistics**)

H.S.J. PICAVET & A. DE BRUIN

Survival Analysis, Overview

Survival analysis is the study of the distribution of life times, that is, the times from an initiating event (birth, start of treatment, employment in a given job) to some terminal event (death, relapse, disability pension). A distinguishing feature of survival data is the inevitable presence of incomplete observations, particularly when the terminal event for some individuals is not observed; instead, it is only known that this event is at least later than a given point in time: *right censoring* (*see* **Censored Data**).

The aims of this entry are to provide a brief historical sketch of the long development of survival analysis and to survey what we have found to be central issues in the current methodology of survival analysis. Necessarily, this entry is rich in cross-references to other entries that treat specific subjects in more detail. However, we have not attempted to include cross-references to *all* specific entries within survival analysis.

History

The Prehistory of Survival Analysis in Demography and Actuarial Science

Survival analysis is one of the oldest statistical disciplines with roots in **demography** and **actuarial** science in the seventeenth century; see [49, Chapter 2]; [51] for general accounts of the history of **vital statistics** and [22] for specific accounts of the work before 1750.

The basic **life-table** methodology in modern terminology amounts to the estimation of a survival function (one minus distribution function) from life times with **delayed entry** (or left **truncation**; see below) and right **censoring**. This was known before 1700, and explicit **parametric models** at least since the linear approximation of de Moivre [39] (see e.g. [22, p. 517]), later examples being due to Lambert [33, p. 483]:

$$\left(1-\frac{x}{96}\right)^2 - 0.6176\left(\exp\left(-\frac{x}{31.682}\right) - \exp\left(-\frac{x}{2.43114}\right)\right) \quad (1)$$

and the influential nineteenth-century proposals by Gompertz [19] and Makeham [37], who modeled the **hazard** function as bc^x and $a + bc^x$, respectively.

Motivated by the controversy over smallpox inoculation, D. Bernoulli [5] laid the foundation of the theory of **competing risks**; see [44] for a historical account. The calculation of **expected number of deaths** (how many deaths would there have been in a study population if a given standard set of death rates applied) also dates back to the eighteenth century; see [29] and the article on **Historical Controls in Survival Analysis**.

Among the important methodological advances in the nineteenth century was, in addition to the parametric survival analysis models mentioned above, the graphical simultaneous handling of calendar time and age in the **Lexis Diagram** [35, cf. 30].

Two very important themes of modern survival analysis may be traced to early twentieth century actuarial mathematics:

Multistate modeling in the particular case of disability insurance [41] and nonparametric estimation in continuous time of the survival function in the competing risk problem under delayed entry and right censoring [13].

At this time, survival analysis was not an integrated component of theoretical statistics. A characteristic scepticism about "the value of life-tables in statistical research" was voiced by Greenwood [20] in the *Journal of the Royal Statistical Society*, and Westergaard's [50] guest appearance in *Biometrika* on "Modern problems in vital statistics" had no reference to sampling variability. This despite the fact that these two authors were actually statistical pioneers in survival analysis: Westergaard [48] by deriving what we would call the standard error of the standardized mortality ratio (rederived by Yule [52]; see [29]) (*see* **Standardization Methods**); and Greenwood [21] with his famous expression for "the 'errors of sampling' of the survivorship tables", (see below).

The "Actuarial" life table and the Kaplan–Meier Estimator

In the mid-twentieth century, these well-established demographic and actuarial techniques were presented to the medical–statistical community in influential surveys such as those by Berkson and Gage [4] and Cutler and Ederer [13]. In this approach, time

is grouped into discrete units (e.g. one-year intervals), and the chain of survival frequencies from one interval to the next are multiplied together to form an estimate of the survival probability across several time periods. The difficulty is in the development of the necessary approximations due to the discrete grouping of the intrinsically continuous time and the possibly somewhat oblique observation fields in cohort studies and more complicated demographic situations. The penetrating study by Kaplan and Meier [28] (*see* **Kaplan–Meier Estimator**), the fascinating genesis of which was chronicled by Breslow [8], in principle, eliminated the need for these approximations in the common situation in medical statistics where all survival and censoring times are known precisely. Kaplan and Meier's tool (which they traced back to Böhmer [7]) was to shrink the observation intervals to include at most one observation per interval. Though overlooked by many later authors, Kaplan and Meier also formalized the age-old handling of **delayed entry** (actually also covered by Böhmer) through the necessary adjustment for the **risk set**, the set of individuals alive and under observation at a particular value of the relevant time variable.

Among the variations on the actuarial model, we will mention two.

Harris et al. [23] anticipated much recent work in, for example, **AIDS** survival studies in their generalization of the usual life-table estimator to the situation in which the death and censoring times are known only in large, irregular intervals (*see* **Grouped Survival Times**).

Ederer et al. [(14)] developed a "relative survival rate... as the ratio of the observed survival rate in a group of patients to the survival rate expected in a group similar to the patients ..." thereby connecting to the long tradition of comparing observed with expected; see, for example, [29] and the article on **Historical Controls in Survival Analysis**.

Parametric Survival Models

Parametric survival models were well-established in actuarial science and demography, but have never dominated medical uses of survival analysis. However, in the 1950s and 1960s important contributions to the statistical theory of survival analysis were based on simple parametric models. One example is the **maximum likelihood** approach by Boag [6] to a **cure model** assuming eternal life with probability c and **lognormally distributed** survival times otherwise. The **exponential distribution** was assumed by Littell [36], when he compared the "actuarial" and the maximum likelihood approach to the "T-year survival rate", by Armitage [3] in his comparative study of two-sample tests for **clinical trials** with **staggered entry**, and by Feigl and Zelen [16] in their model for (uncensored) lifetimes whose expectations were allowed to depend linearly on covariates, generalized to censored data by Zippin and Armitage [53].

Cox [11] revolutionized survival analysis by his **semiparametric regression** model for the hazard, depending arbitrarily ("nonparametrically") on time and parametrically on covariates (*see* **Cox Regression Model**). For details on the genesis of Cox's paper, see [42, 43].

Multistate Models

Traditional actuarial and demographical ways of modeling several life events simultaneously may be formalized within the probabilistic area of finite-state **Markov processes** in continuous time. An important and influential documentation of this was by Fix and Neyman [18], who studied recovery, relapse, and death (and censoring) in what is now commonly termed an illness–death model allowing for competing risks (*see* **Fix–Neyman Process**). Chiang [9], for example, in his 1968 monograph, extensively documented the relevant stochastic models (*see* **Stochastic Processes**), and Sverdrup [46], in an important paper, gave a systematic statistical study. These models have constant transition intensities, although subdivision of time into intervals allows grouped-time methodology of the actuarial life-table type, as carefully documented by Hoem [24].

Survival Analysis Concepts

The ideal basic independent nonnegative **random variables** $X_i, i = 1, \ldots, n$ are not always observed directly. For some individuals i, the available piece of information is a *right-censoring* time U_i, that is, a period elapsed in which the event of interest has not occurred (e.g. a patient has survived until U_i). Thus, a generic survival data sample includes $((\widetilde{X}_i, D_i), i = 1, \ldots, n)$ where $\widetilde{X}_i$ is the smaller of X_i and U_i and D_i is the indicator, $I(X_i \leq U_i)$, of not being censored.

Mathematically, the distribution of X_i may be described by the *survival function*

$$S_i(t) = \Pr(X_i > t). \tag{2}$$

If the **hazard function**

$$\alpha_i(t) = \lim_{\Delta t \to 0} \frac{\Pr(X_i \leq t + \Delta t \mid X_i > t)}{\Delta t} \tag{3}$$

exists, then

$$S_i(t) = \exp(-A_i(t)), \tag{4}$$

where

$$A_i(t) = \int_0^t \alpha_i(u)\, \mathrm{d}u \tag{5}$$

is the integrated hazard over $[0, t)$. If, more generally, the distribution of the X_i has discrete components, then $S_i(t)$ is given by the **product-integral** of the cumulative hazard measure. Owing to the dynamical nature of survival data, a characterization of the distribution via the hazard function is often convenient. (Note that $\alpha_i(t)\Delta t$ when $\Delta t > 0$ is *small* is approximately the conditional probability of i "dying" just after time t given "survival" till time t.) Also, $\alpha_i(t)$ is the basic quantity in the **counting process** approach to survival analysis (see e.g. [2], and the article on **Survival Distributions and Their Characteristics**).

Nonparametric Estimation and Testing

The simplest situation encountered in survival analysis is the nonparametric estimation of a survival distribution function based on a right-censored sample of observation times $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$, where the true survival times $X_i, i = 1, \ldots, n$, are assumed to be independent and identically distributed with common survival distribution function $S(t)$, whereas as few assumptions as possible are usually made about the right-censoring times U_i except for the assumption of *independent censoring* (*see* **Censored Data**). The concept of independent censoring has the interpretation that the fact that an individual, i, is alive *and uncensored* at time t, say, should not provide more information on the survival time for that individual than $X_i > t$, that is, the right-censoring mechanism should not remove individuals from the study who are at a particularly high or a particularly low risk of dying. Under these assumptions, $S(t)$ is estimated by the *Kaplan–Meier estimator* [28]. This is given by

$$\widehat{S(t)} = \prod_{\widetilde{X}_i \leq t} \left[1 - \frac{D_i}{Y(\widetilde{X}_i)}\right], \tag{6}$$

where $Y(t) = \sum I(\widetilde{X}_i \geq t)$ is the number of individuals *at risk* just before time t. The Kaplan–Meier estimator is a **nonparametric maximum likelihood** estimator and, in large samples, $\widehat{S(t)}$ is approximately normally distributed with mean $S(t)$ and a variance that may be estimated by Greenwood's formula:

$$\widehat{\sigma^2(t)} = [\widehat{S(t)}]^2 \sum_{\widetilde{X}_i \leq t} \frac{D_i}{Y(\widetilde{X}_i)[Y(\widetilde{X}_i) - 1]}. \tag{7}$$

From this result, pointwise **confidence intervals** for $S(t)$ are easily constructed and, since one can also show weak **convergence** of the entire Kaplan–Meier curve $\{\sqrt{(n)}[\widehat{S(t)} - S(t)]; 0 \leq t \leq \tau\}, \tau \leq \infty$ to a mean zero Gaussian process (*see* **Brownian Motion and Diffusion Processes**), simultaneous confidence bands for $S(t)$ on $[0, \tau]$ can also be set up.

As an alternative to estimating the survival distribution function $S(t)$, the *cumulative hazard function* $A(t) = -\log S(t)$ may be studied. Thus, $A(t)$ may be estimated by the **Nelson–Aalen Estimator**

$$\widehat{A(t)} = \sum_{\widetilde{X}_i \leq t} \frac{D_i}{Y(\widetilde{X}_i)}. \tag{8}$$

The relation between the estimators $\widehat{S(t)}$ and $\widehat{A(t)}$ is given by the *product-integral* from which it follows that their large-sample properties are equivalent. Though the Kaplan–Meier estimator has the advantage that a survival *probability* is easier to interpret than a cumulative hazard function, the Nelson–Aalen estimator is easier to generalize to multistate situations beyond the survival data context. We shall return to this below. To give a nonparametric estimate of the hazard function $\alpha(t)$ itself requires some smoothing techniques to be applied (see **Smoothing Hazard Rates**).

Right censoring is not the only kind of data-incompleteness to be dealt with in survival analysis; in particular, *left truncation* (or **delayed entry**) where individuals may not all be followed from time 0 but maybe from a later entry time V_i conditionally on having survived until V_i, occurs frequently in, for example, epidemiological applications. Dealing with

left truncation only requires a redefinition of the *risk set* from the set $\{i: \widetilde{X}_i \geq t\}$ of individuals still alive and uncensored at time t to the set $\{i: V_i < t \leq \widetilde{X}_i\}$ of individuals with entry time $V_i < t$ and who are still alive and uncensored. With $Y(t)$ still denoting the size of the risk set at time t both (6), (7), and (8) are applicable though one should be aware of the fact that estimates of $S(t)$ and $A(t)$ may be ill-determined for small values of t due to the left truncation (*see* **Truncated Survival Times**).

When the survival time distributions in a number, k, of homogeneous groups have been estimated nonparametrically, it is often of interest to test the hypothesis H_0 of identical hazards in all groups. Thus, on the basis of censored survival data $((\widetilde{X}_{hi}, D_{hi}),\ i = 1, \ldots, n_h)$ for group $h, h = 1, \ldots, k$, the Nelson–Aalen estimates $\widehat{A_h(t)}$ have been computed, and based on the combined sample of size $n = \sum_h n_h$ with data $((\widetilde{X}_i, D_i), i = 1, \ldots, n)$, an estimate of the common cumulative hazard function $A(t)$ under H_0 may be obtained by a Nelson–Aalen estimator $\widehat{A(t)}$. As a general statistic for testing H_0, one may then use a k-vector of sums of weighted differences between the increments of $\widehat{A_h(t)}$ and $\widehat{A(t)}$:

$$Z_h = \sum_{i=1}^{n} K_h(\widetilde{X}_i)[\Delta\widehat{A_h}(\widetilde{X}_i) - \Delta\hat{A}(\widetilde{X}_i)]. \quad (9)$$

Here, $\Delta\widehat{A_h(t)} = 0$ if t is not among the observed survival times in the hth sample and $K_h(t)$ is 0 whenever $Y_h(t) = 0$, in fact all weight functions used in practice have the form $K_h(t) = Y_h(t)K(t)$. With this structure for the weight function, the covariance between Z_h and Z_j given by (9) is estimated by

$$\sigma_{hj} = \sum_{i=1}^{n} K^2(\widetilde{X}_i)\frac{Y_h(\widetilde{X}_i)}{Y(\widetilde{X}_i)}\left[\delta_{hj} - \frac{Y_j(\widetilde{X}_i)}{Y(\widetilde{X}_i)}\right]D_i, \quad (10)$$

and, letting $\mathbf{Z}$ be the k-vector $(Z_1, \ldots, Z_k)'$ and $\boldsymbol{\Sigma}$ the k by k matrix $(\sigma_{hj}, h, j = 1, \ldots, k)$ the test statistic $X^2 = \mathbf{Z}'\boldsymbol{\Sigma}^{-}\mathbf{Z}$ is asymptotically **chi-squared distributed** under H_0 with $k - 1$ **degrees of freedom** if all n_h tend to infinity at the same rate. Here, $\boldsymbol{\Sigma}^{-}$ is a generalized inverse for $\boldsymbol{\Sigma}$ (*see* **Matrix Algebra**).

Special choices for $K(t)$ correspond to test statistics with different properties for particular alternatives to H_0 (*see* **Linear Rank Tests in Survival Analysis**). An important such test statistic is the **logrank** test obtained for $K(t) = I(Y(t) > 0)$. For this test, which has particularly good power for **proportional hazards** alternatives, Z_h given by (9) reduces to $Z_h = O_h - E_h$ with O_h the total number of observed failures in group h and $E_h = \sum D_i Y_h(\widetilde{X}_i)/Y(\widetilde{X}_i)$ an "expected" number of failures in group h. For the two-sample case ($k = 2$), one may of course use the square root of X^2 as an asymptotically normal test statistic for the **null hypothesis**. For the case where the k groups are *ordered*, and where a *score* x_h (with $x_1 \leq \ldots \leq x_k$) is attached to group h, a *test for trend* is given by $T^2 = (\mathbf{x}'\mathbf{Z})^2/\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}$ with $\mathbf{x} = (x_1, \ldots, x_k)'$ and it is asymptotically chi-squared with 1 df.

The above **linear rank tests** have low **power** against certain important classes of alternatives such as "crossing hazards". Just as for uncensored data, this has motivated the development of test statistics of the **Kolmogorov–Smirnov and Cramèr–von Mises** types, based on maximal deviation or integrated squared deviation between estimated hazards, cumulative hazards or survival functions.

Parametric Inference

The nonparametric methods outlined in the previous section have become the standard approach to the analysis of simple homogeneous survival data without covariate information. However, parametric survival time distributions are sometimes used for inference, and we shall here give a brief review. Assume again that the true survival times $X_1, \ldots, X_n$ are independent and identically distributed with survival distribution function $S(t; \boldsymbol{\theta})$ and hazard function $\alpha(t; \boldsymbol{\theta})$ but that only a right-censored sample $(\widetilde{X}_i, D_i), i = 1, \ldots, n$, is observed. Under independent censoring, the likelihood function for the parameter $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}(\alpha(\widetilde{X}_i; \boldsymbol{\theta}))^{D_i} S(\widetilde{X}_i; \boldsymbol{\theta}). \quad (11)$$

The function (11) may be analyzed using standard **large-sample theory**. Thus, standard tests, that is, Wald-, score-, and **likelihood ratio tests** are used as inferential tools (*see* **Chi-square Tests**). Two frequently used parametric survival models are the **Weibull distribution** with hazard function

$\alpha\rho(\alpha t)^{\rho-1}$, and the piecewise exponential distribution with $\alpha(t, \boldsymbol{\theta}) = \alpha_j$ for $t \in I_j$ with $I_j = [t_{j-1}, t_j)$, $0 = t_0 < t_j < \cdots < t_m = \infty$. Both of these distributions contain the very simplest model, the **exponential distribution** with a constant hazard function as null cases (*see* **Parametric Models in Survival Analysis**).

Comparison with Expected Survival

As a special case of the nonparametric tests discussed above, a *one-sample* situation may be studied. This may be relevant if one wants to compare the observed survival in the sample with the *expected survival* based on a standard life table. Thus, assume that a hazard function $\alpha^*(t)$ is given and that the hypothesis $H_0 : \alpha = \alpha^*$ is to be tested. One test statistic for H_0 is the one-sample *logrank test* $(O - E^*)/(E^*)^{1/2}$ where E^*, the "expected" number of deaths is given by $E^* = \sum[A^*(\widetilde{X}_i) - A^*(V_i)]$ (with A^* the cumulative hazard corresponding to α^*). In this case, $\widehat{\theta} = O/E^*$, the *standardized mortality ratio*, is the maximum likelihood estimate for the parameter θ in the model $\alpha(t) = \theta\alpha^*(t)$. Thus, the standardized mortality ratio arises from a **multiplicative model** involving the known population hazard $\alpha^*(t)$. Another classical tool for comparing with expected survival, the so-called *expected survival function*, arises from an *additive or excess* hazard model (*see* **Excess Mortality**; **Expected Number of Deaths**; **Historical Controls in Survival Analysis**).

The Cox Regression Model

In many applications of survival analysis, the interest focuses on how *covariates* may affect the outcome; in clinical trials, adjustment of treatment effects for effects of other **explanatory variables** may be crucial if the randomized groups are unbalanced with respect to important **prognostic factors**, and in epidemiological **cohort studies**, reliable effects of exposure may be obtained only if some adjustment is made for **confounding** variables. In these situations, a *regression model* is useful and the most important model for survival data is the **Cox** [11] *proportional hazards regression model*. In its simplest form, it states the hazard function for an individual, i, with covariates $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})'$ to be

$$\alpha_i(t; \boldsymbol{Z}_i) = \alpha_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i), \tag{12}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown regression coefficients and $\alpha_0(t)$, the *baseline hazard*, is the hazard function for individuals with all covariates equal to 0. Thus, the baseline hazard describes the common shape of the survival time distributions for all individuals while the *relative risk* function $\exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)$ gives the level of each individual's hazard. The interpretation of the parameter, β_j for a dichotomous $Z_{ij} \in \{0, 1\}$ is that $\exp(\beta_j)$ is the **relative risk** for individuals with $Z_{ij} = 1$ compared to those with $Z_{ij} = 0$ all other covariates being the same for the two individuals. Similar interpretations hold for parameters corresponding to covariates taking more than two values.

The model is **semiparametric** in the sense that the relative risk part is modeled parametrically while the baseline hazard is left unspecified. This semiparametric nature of the model led to a number of inference problems, which was discussed in the literature in the years following the publication of Cox's article in 1972. However, these problems were all resolved and estimation proceeds as follows. The regression coefficients $\boldsymbol{\beta}$ are estimated by maximizing the **Cox partial likelihood**

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)} \right]^{D_i}, \tag{13}$$

where $R_i = \{j : \widetilde{X}_j \geq \widetilde{X}_i\}$, the **risk set** at time $\widetilde{X}_i$, is the set of individuals still alive and uncensored at that time. Furthermore, the cumulative baseline hazard $A_0(t)$ is estimated by the *Breslow estimator*

$$\widehat{A_0(t)} = \sum_{\widetilde{X}_i \leq t} \frac{D_i}{\sum_{j \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{Z}_j)}, \tag{14}$$

which is the Nelson–Aalen estimator one would use if $\boldsymbol{\beta}$ were known and equal to the maximum partial likelihood estimate $\widehat{\boldsymbol{\beta}}$. The estimators based on (13) and (14) also have a nonparametric maximum likelihood interpretation. In large samples, $\widehat{\boldsymbol{\beta}}$ is approximately normally distributed with the proper mean and with a **covariance**, which is estimated by the **information matrix** based on (13). This means that approximate confidence intervals for the relative risk parameters of interest can be calculated and that the usual large-sample test statistics based on (13) are available. Also, the asymptotic distribution of the Breslow estimator is normal; however, this estimate

is most often used as a tool for estimating *survival probabilities* for individuals with given covariates, $\mathbf{Z}_0$. Such an estimate may be obtained by the product integral $\widehat{S(t;\mathbf{Z}_0)}$ of $\exp(\widehat{\boldsymbol{\beta}}'\mathbf{Z}_0)\widehat{A_0(t)}$. The *joint* asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and the Breslow estimator then yields an approximate normal distribution for $\widehat{S(t;\mathbf{Z}_0)}$ in large samples.

A number of useful extensions of this simple Cox model are available. Thus, in some cases, the covariates are **time-dependent**, for example, a covariate might indicate whether or not a given event had occurred by time t, or a time-dependent covariate might consist of repeated recordings of some measurement likely to affect the prognosis. In such cases, the regression coefficients $\boldsymbol{\beta}$ are estimated replacing $\exp(\boldsymbol{\beta}'\mathbf{Z}_j)$ in (13) by $\exp[\boldsymbol{\beta}'\mathbf{Z}_j(\widetilde{X}_i)]$.

Also, a simple extension of the Breslow estimator (14) applies in this case. However, the survival function can, in general, no longer be estimated in a simple way because of the extra randomness arising from the covariates, which is not modeled in the Cox model. This has the consequence that the estimates are more difficult to interpret when the model contains time-dependent covariates. To estimate the survival function in such cases, a joint model for the hazard and the time-dependent covariate is needed (*see* **Joint Modeling of Longitudinal and Event Time Data**).

Another extension of (12) is the *stratified* Cox model where individuals are grouped into a number, k of strata each of which has a separate baseline hazard (*see* **Stratification**). This model has important applications for checking the assumptions of (12). The model assumption of proportional hazards may also be *tested* in a number of ways, the simplest possibility being to add interaction terms of the form $Z_{ij}f(t)$ between Z_{ij} and time where $f(t)$ is some specified function. Also, various forms of *residuals* as for normal linear models may be used for **model checking** in (12) (*see* **Goodness of Fit in Survival Analysis**; **Residuals for Survival Analysis**). In (12), it is finally assumed that a quantitative covariate affects the hazard *log-linearly*. This assumption may also be checked in several ways and alternative models with other relative risk functions $r(\boldsymbol{\beta}'\mathbf{Z}_i)$ may be used. Special care is needed when covariates are measured with error (*see* **Measurement Error in Survival Analysis**).

Other Regression Models for Survival Data

Though the semiparametric Cox model is the regression model for survival data that is applied most frequently, other regression models, for example, *parametric* regression models also play important roles in practice. Examples include models with a multiplicative structure, that is, models like (12) but with a parametric specification, $\alpha_0(t) = \alpha_0(t;\boldsymbol{\theta})$, of the baseline hazard, and **accelerated failure-time models**.

A multiplicative model with important epidemiological applications is the **Poisson regression** model with a piecewise constant baseline hazard. In large data sets with categorical covariates, this model has the advantage that a sufficiency reduction to the number of failures and the amount of person-time at risk in each *cell* defined by the covariates and the division of time into intervals is possible. This is in contrast to the Cox regression model (12) where each individual data record is needed to compute (13). The substantial computing time required to maximize (13) in large samples has also led to modifications of this estimation procedure. Thus, in *nested case–control studies* the risk set R_i in the Cox partial likelihood is replaced by a random sample $\widetilde{R}_i$ of R_i (*see* **Case–Control Study, Nested**).

In the accelerated failure-time model, the focus is not on the hazard function but on the survival time itself much like in classical linear models. Thus, this model is given by $\log X_i = \alpha + \boldsymbol{\beta}'\mathbf{Z}_i + \varepsilon_i$, where the error terms are assumed to be independent and identically distributed with expectation 0. Examples include **normally distributed** ($\varepsilon_i, i = 1, \ldots, n$), and error terms with a **logistic** or an **extreme value** distribution, the latter giving rise to a regression model with **Weibull** distributed life times.

Finally, we shall mention some nonparametric hazard models. In Aalen's additive model, $\alpha_i(t) = \beta_0(t) + \boldsymbol{\beta}(t)'\mathbf{Z}_i(t)$ (*see* **Aalen's Additive Regression Model**), the regression functions $\beta_0(t), \ldots, \beta_p(t)$ are left completely unspecified and estimated nonparametrically much like the Nelson–Aalen estimator discussed above. This model provides an attractive alternative to the other regression models discussed in this section. There also exist more general and flexible models containing both this model and the Cox regression model as special cases (*see* **Additive–Multiplicative Intensity Models**).

Multistate Models

Models for survival data may be considered a special case of a *multistate model*, namely, a model with a transient state *alive* (0) and an absorbing state *dead* (1) and where the hazard rate is the force of transition from state 0 to state 1. Multistate models may conveniently be studied in the mathematical framework of *counting processes* with a notation that actually simplifies the notation of the previous sections and, furthermore, unifies the description of survival data and that of more general models like the competing risks model and the illness–death model to be discussed below. We first introduce the counting processes relevant for the study of censored survival data [1] Define, for $i = 1, \ldots, n$, the stochastic processes

$$N_i(t) = I(\widetilde{X}_i \leq t, D_i = 1) \tag{15}$$

and

$$Y_i(t) = I(\widetilde{X}_i \geq t). \tag{16}$$

Then (15) is a counting process counting 1 at time $\widetilde{X}_i$ if individual i is observed to die; otherwise $N_i(t) = 0$ throughout. The process (16) indicates whether i is still at risk just before time t. Models for the survival data are then introduced via the *intensity process*, $\lambda_i(t) = \alpha_i(t)Y_i(t)$ for $N_i(t)$, where $\alpha_i(t)$, as before, denotes the hazard function for the distribution of X_i. Letting $N = N_1 + \cdots + N_n$ and $Y = Y_1 + \cdots + Y_n$ the Nelson–Aalen estimator (8) is given by the stochastic integral

$$\widehat{A(t)} = \int_0^t \frac{J(u)}{Y(u)}\, \mathrm{d}N(u), \tag{17}$$

where $J(t) = I(Y(t) > 0)$. In this simple multistate model, the *transition probability* $P_{00}(0, t)$, that is, the conditional probability of being in state 0 by time t given state 0 at time 0 is simply the survival probability $S(t)$, which, as described above, may be estimated using the Kaplan–Meier estimator, which is the product-integral of (17). In fact, all the models and methods for survival data discussed above, which are based on the hazard function have immediate generalizations to models based on counting processes. Thus, both the nonparametric tests and the Cox regression model may be applied for counting process (multistate) models (*see* **Counting Process Methods in Survival Analysis**).

One important extension of the two-state model for survival data is the *competing risks* model with one transient alive state 0 and a number, k, of absorbing states corresponding to death from cause h, $h = 1, \ldots, k$. In this model, the basic parameters are the cause-specific hazard functions $\alpha_h(t)$, $h = 1, \ldots, k$, and the observations for individual i will consist of $(\widetilde{X}_i, D_{hi})$, $h = 1, \ldots, k$, where $D_{hi} = 1$ if individual i is observed to die from cause h, and $D_{hi} = 0$ otherwise. On the basis of these data, k counting processes for each i can be defined by $N_{hi}(t) = I(\widetilde{X}_i \leq t, D_{hi} = 1)$ and letting $N_h = N_{h1} + \cdots + N_{hn}$, the integrated cause-specific hazard $A_h(t)$ is estimated by the Nelson–Aalen estimator replacing N by N_h in (17). A useful synthesis of the cause-specific hazards is provided by the transition probabilities $P_{0h}(0, t)$ of being dead from cause h by time t. This is frequently called the *cumulative incidence function* for cause h and is given by

$$P_{0h}(s, t) = \int_s^t S(u)\alpha_h(u)\, \mathrm{d}u, \tag{18}$$

and hence it may be estimated by (18) by inserting the Kaplan–Meier estimate for $S(u)$ and the Nelson–Aalen estimate for the integrated cause-specific hazard. In fact, this **Aalen–Johansen estimator** of the matrix of transition probabilities is exactly the product-integral of the cause-specific hazards.

Another important multistate model is the *illness–death* or *disability* model with two transient states, say *healthy* (0) and *diseased* (1) and one absorbing state *dead* (2). If transitions both from 0 to 1 and from 1 to 0 are possible, the disease is *recurrent*, otherwise it is *chronic*. On the basis of such observed transitions between the three states, it is possible to define counting processes for individual i as $N_{hji}(t) =$ number of observed $h \to j$ transitions in the time interval $[0, t]$ for individual i and, furthermore, we may let $Y_{hi}(t) = I$ (i is in state h at time $t-$). With these definitions, we may set up and analyze models for the transition intensities $\alpha_{hji}(t)$ from state h to state j including nonparametric comparisons and Cox-type regression models. Furthermore, transition probabilities $P_{hj}(s, t)$ may be estimated by product-integration of the intensities.

Other Kinds of Incomplete Observation

A salient feature of survival data is *right censoring*, which has been referred to throughout in the present

overview. However, several other kinds of incomplete observation are important in survival analysis.

Often, particularly when the time variable of interest is age, individuals enter study after time 0. This is called *delayed entry* and may be handled by *left truncation* (conditioning) or *left filtering* ("viewing the observations through a filter"). There are also situations when only events (such as AIDS cases) that occur *before* a certain time are included (*right truncation*) (*see* **Truncated Survival Times**). The phenomenon of *left censoring*, though theoretically possible, is more rarely relevant in survival analysis.

When the event times are only known to lie in an interval, one may use the *grouped time* approach of classical *life tables* (*see* **Grouped Survival Times**; **Life Table**), or (if the intervals are not synchronous) techniques for **interval censoring** may be relevant.

A common framework (**coarsening at random**) was recently suggested for several of the above types of incomplete observation.

Multivariate Survival Analysis

For **multivariate survival**, the innocently looking problem of generalizing the Kaplan–Meier estimator to several dimensions has proved surprisingly intricate. A major challenge (in two dimensions) is how to efficiently use singly censored observations, where one component is observed and the other is right censored.

For regression analysis of multivariate survival times, two major approaches have been taken. One is to model the marginal distributions and use estimation techniques based on **generalized estimating equations** leaving the association structure unspecified (see **Marginal Models for Multivariate Survival Data**.) The other is to specify **random effects** models for survival data based on conditional independence (*see* **Frailty**.) An interesting combination between these two methods is provided by **copula** models in which the marginal distributions are combined via a so-called copula function thereby obtaining an explicit model for the association structure.

For the special case of repeated events, both the marginal approach and the conditional (frailty) approach have been used successfully (*see* **Repeated Events**).

Concluding Remarks

Survival analysis is a well-established discipline in statistical theory as well as in biostatistics. Most books on biostatistics contain chapters on the topic and most **software** packages include procedures for handling the basic survival techniques (*see* **Survival Analysis, Software**). Several books have appeared, among them the documentation of the actuarial and demographical know-how by Elandt–Johnson and Johnson [15]; the research monograph by Kalbfleisch and Prentice [27], the first edition of which for a decade maintained its position as main reference on the central theory; the comprehensive text by Lawless [34] covering also parametric models, and the concise text by Cox and Oakes [12], two central contributors to the recent theory. The counting process approach is covered by Fleming and Harrington [17] and by Andersen et al. [2]; see also [25]. Later, books intended primarily for the biostatistical user have appeared. These include [10, 31, 32, 38, 40]. Also, books dealing with special topics, like implementation in the **S-Plus** software [47], multivariate survival data [26], and the linear regression model [45] have appeared.

References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals. of Statistics* **6**, 701–726.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[3] Armitage, P. (1959). The comparison of survival curves, *Journal of the Royal Statistical Society A* **122**, 279–300.

[4] Berkson, J. & Gage, R.P. (1950). Calculation of survival rates for cancer, *Proceedings of the Staff Meetings of the Mayo Clinic* **25**, 270–286.

[5] Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir, *Mémoires de Mathématique et de Physique de l'Académie Royale des Sciences, Paris* Année MDCCLX, pp. 1–45 of Mémoires.

[6] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society B* **11**, 15–53.

[7] Böhmer, P.E. (1912). Theorie der unabhängigen Wahrscheinlichkeiten, *Rapports, Mémoires et Procés – verbaux du 7^e Congrès International d'Actuaires, Amsterdam* **2**, 327–343.

[8] Breslow, N.E. (1991). Introduction to Kaplan and Meier (1958). Nonparametric estimation from incomplete observations, in *Breakthroughs in Statistics II*, S. Kotz & N.L. Johnson, eds. Springer, New York, 311–318.

[9] Chiang, C.L. (1968). *Introduction to Stochastic Processes in Biostatistics*. Wiley, New York.

[10] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd Ed., Chapman and Hall, London.

[11] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society (B)* **34**, 187–220.

[12] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

[13] Cutler, S.J. & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival, *Journal of Chronic Diseases* **8**, 699–713.

[14] Ederer, F., Axtell, L.M. & Cutler, S.J. (1961). The relative survival rate: A statistical methodology, *National Cancer Institute Monographs* **6**, 101–121.

[15] Elandt-Johnson, R.C. & Johnson, N.L. (1980). *Survival Models and Data Analysis*. Wiley, New York.

[16] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[17] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[18] Fix, E. & Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients, *Human Biology* **23**, 205–241.

[19] Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, *Philosophical Transactions of the Royal Society of London, Series A* **115**, 513–580.

[20] Greenwood, M. (1922). Discussion on the value of life-tables in statistical research, *Journal of the Royal Statistical Society* **85**, 537–560.

[21] Greenwood, M. (1926). The natural duration of cancer, in *Reports on Public Health and Medical Subjects*, Vol. 33 His Majesty's Stationery Office, London, pp. 1–26.

[22] Hald, A. (1990). *A History of Probability and Statistics and Their Applications before 1750*. Wiley, New York.

[23] Harris, T.E., Meier, P. & Tukey, J.W. (1950). The timing of the distribution of events between observations, *Human Biology* **22**, 249–270.

[24] Hoem, J.M. (1976). The statistical theory of demographic rates. A review of current developments (with discussion), *Scandinavian Journal of Statistics* **3**, 169–185.

[25] Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis. Regression Modeling of Time to Event Data*. Wiley, New York.

[26] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

[27] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed., Wiley, New York.

[28] Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481, 562–563.

[29] Keiding, N. (1987). The method of expected number of deaths 1786–1886–1986, *International Statistical Review* **55**, 1–20.

[30] Keiding, N. (1990). Statistical inference in the Lexis diagram, *Philosophical Transactions of the Royal Society London A* **332**, 487–509.

[31] Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis. Techniques for Censored and Truncated Data*, 2nd Ed., Springer, New York.

[32] Kleinbaum, D.G. (1996). *Survival Analysis. A Self-Learning Text*. Springer, New York.

[33] Lambert, J.H. (1772). *Beyträge zum Gebrauche der Mathematik und deren Anwendung*, Vol. III, Verlage des Buchlages der Realschule, Berlin.

[34] Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data*, 2nd Ed., Wiley, New York.

[35] Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Trübner, Strassburg.

[36] Littell, A.S. (1952). Estimation of the T-year survival rate from follow-up studies over a limited period of time, *Human Biology* **24**, 87–116.

[37] Makeham, W.M. (1860). On the law of mortality, and the construction of mortality tables, *Journal of the Institute of Actuaries* **8**, 301.

[38] Marubini, E. & Valsecchi, M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley, Chichester.

[39] de Moivre, A. (1725). *Annuities upon Lives: or, The Valuation of Annuities uon any Number of Lives; as also, of Reversions. To which is added, An Appendix concerning the Expectations of Life, and Probabilites of Survivorship*. Fayram, Motte and Pearson, London.

[40] Parmar, K.B. & Machin, D. (1995). *Survival analysis. A practical approach*. Wiley, Chichester.

[41] du Pasquier, L.G. (1913). Mathematische Theorie der Invaliditätsversicherung, *Mitteilungen der Vereinigung der Schweizerische Versicherungs-Mathematiker* **8**, 1–153.

[42] Prentice, R.L. (1991). Introduction to Cox (1972) Regression models and life-tables, in *Breakthroughs in Statistics II*, S. Kotz & N.L. Johnson eds. Springer, New York, pp. 519–526.

[43] Reid, N. (1994). A conversation with Sir David Cox, *Statistical Science* **9**, 439–455.

[44] Seal, H.L. (1977). Studies in the history of probability and statistics, XXXV. Multiple decrements or competing risks, *Biometrika* **64**, 429–439.

[45] Smith, P.J. (2002). *Analysis of Failure Time Data*. Chapman and Hall/CRC, London.

[46] Sverdrup, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health, *Skandinavisk Aktuarietidskrift* **48**, 184–211.

[47] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.
[48] Westergaard, H. (1882). *Die Lehre von der Mortalität und Morbilität*. Fischer, Jena.
[49] Westergaard, H. (1901). *Die Lehre von der Mortalität und Morbilität*, 2. Aufl., Fischer, Jena.
[50] Westergaard, H. (1925). Modern problems in vital statistics, *Biometrika* **17**, 355–364.
[51] Westergaard, H. (1932). *Contributions to the History of Statistics*. King, London.
[52] Yule, G. (1934). On some points relating to vital statistics, more especially statistics of occupational mortality (with discussion), *Journal of the Royal Statistical Society* **97**, 1–84.
[53] Zippin, C. & Armitage, P. (1966). Use of concomitant variables and incomplete survival information with estimation of an exponential survival parameter, *Biometrics* **22**, 655–672.

PER KRAGH ANDERSEN & NIELS KEIDING

Survival Analysis, Software

Most, if not all, techniques for the analysis of time-to-event data (*see* **Survival Analysis, Overview**) requires specialized **software** (*see* **Software, Biostatistical**). For example, many of the estimators used with **censored data** do not have closed-form solutions and require iterative solutions to estimating equations (*see* **Estimating Functions**).

The specialized software for survival data may be in a stand-alone package such as *SURVIVAL* [16] or may be an integrated part of a more comprehensive package such as SAS or SPSS. Whatever the package, there are a minimal number of requirements that a program should have. The package should have the ability to handle right-censored data and to properly handle tied observations (*see* **Tied Survival Times**). Better packages should have the ability to handle other types of censoring such as left and interval censoring (*see* **Censored Data**). Some packages can also handle left truncated data (*see* **Truncated Survival Times**).

The types of analysis for survival data fall into four main areas. The first area is summary or univariate statistics. Here, most packages have routines to compute the Kaplan–Meier estimator (*see* **Kaplan–Meier Estimator**) for right-censored data and make summary plots.

The second area is **hypothesis testing**. Most packages have routines to compute the **logrank test** for right-censored data. More inclusive packages have routines for weighted logrank tests.

The third area is **semiparametric regression** modeling. Most packages have at least a routine to fit the Cox proportional hazards model (*see* **Cox Regression Model**; **Proportional Hazards, Overview**) with fixed time **covariates** to right-censored data. Some packages have routines for checking model assumptions (*see* **Model Checking**) of which a check of the proportional hazards assumption is most common (*see* **Goodness of Fit in Survival Analysis**). Most packages allow for **time-dependent covariates**.

The final area is **parametric models** for right-censored data (*see* **Survival Distributions and Their Characteristics**). Most packages have routines for at least the **Weibull** model. More complete packages have routines for the **lognormal** and log-logistic models. Many of these packages allow for regression analysis for right-censored data.

Several authors have compared the adequacy of statistical packages for survival data. These include surveys by Dain et al. [6], Goldstein et al. [9], and Harrell and Goldstein [10]. We refer the interested reader to these sources for a comparison of packages for the PC.

In the remainder of this article, we compare and contrast the abilities of five of the most common statistical packages in use. These are SAS (Version 8.2), S-PLUS (version 6.1.2), SPSS (Version 11), STATA (Version 7), and BMDP (New system 2.0). We have chosen these five packages since they are comprehensive packages that include a wide range of statistical routines. They are available for a number of platforms including the PC and UNIX systems. In our interaction with other biostatisticians, we have found that these are the packages that seem to be most popular. Contact information (many packages have contact sites outside the US as well) and worldwide web site address for these packages are listed below. Newest update information, new features and modules, user's manuals, and technical support information are often available on these web sites.

1. **BMDP**, SPSS Inc., 233 S. Wacker Drive, 11th Floor, Chicago, IL 60606, USA; (312) 651–3000; http://www.statsol.ie/bmdp/bmdp.htm.
2. **SAS**, SAS Institute Inc. 100 SAS Campus Drive, Cary, NC 27513-2414, USA; (919) 677–8000; http://www.sas.com.
3. **SPSS**, SPSS Inc., 233 S. Wacker Drive, 11th Floor, Chicago, IL 60606, USA; (312) 651–3000; http://www.spss.com.
4. **S-PLUS**, Insightful Corporation Global Headquarters, 1700 Westlake Avenue, North Suite 500, Seattle, WA 98109-3044, USA; (800) 569-0123; http://www.splus.com.
5. **STATA**, STATA Corporation, 4905 Lakeway Drive, College Station, Texas 77845, USA; (800)-782-8272; http://www.stata.com.

Summary of Univariate Statistics

Each of the five packages provides the Kaplan–Meier estimator for right-censored data. The packages all present estimates of the standard errors using Greenwood's formula. Each package provides

estimates of the **median survival** function based on the Kaplan–Meier estimator and confidence intervals for the median survival based on the Brookmeyer and Crowley procedure [4]. All these interval estimates of the median are based on the naïve confidence interval for the survival function except for the interval found in STATA that uses a log–log transformed confidence interval for the survival function. Each package provides estimates of the mean survival and its standard error; however, when the last observation is censored these estimates may differ. SAS estimates the restricted mean as the area under the Kaplan–Meier curve up to the last event, while each of the other packages estimates the area under the curve up to the largest on study time. STATA allows for an alternative estimator of the mean based on completing the tail of the Kaplan–Meier estimator by an exponential curve as suggested by Brown et al. [5].

Each of the packages has additional features that are not shared by all the other packages that we now summarize.

BMDP. BMDP allows right-censored data only. It allows the user to produce plots of the estimated survival function, log survival function and cumulative **cumulative hazard** function (i.e. $-\log\{\hat{S}(t)\}$). It provides estimates of the standard error of the median. BMDP does not give confidence intervals for the survival function. BMDP also computes life-table estimates of the survival function and **hazard rate** based on grouped survival data.

SAS. SAS also only handles right-censored data. It provides Kaplan–Meier estimator in the routine PROC LIFETEST and the **Nelson–Aalen estimator** of the cumulative hazard rate can be obtained in PROC PHREG. Naïve confidence intervals are constructed for the survival function based on the Kaplan–Meier estimator. Both the Kaplan–Meier and the Fleming–Harrington [8] estimators ($\exp\{-\hat{\Lambda}(t)\}$) of the survival function, where $\Lambda_{\widehat{(t)}}$ is the Nelson–Aalen estimator of the cumulative hazard function, can be put into a SAS data set from which a wide range of plots can be made. SAS also provides classical **life-table** estimates based on **grouped data**.

S-PLUS. S-PLUS is the most flexible of the packages. It has versions of the Kaplan–Meier estimator that handle right-, left-, and interval-censored data [18, 19]. For right-censored data it allows the user to select between three variance estimators [11] and three types of confidence intervals [2] (naïve, log, and log–log transformed). It also provides the Nelson–Aalen estimator of the hazard rate and Fleming–Harrington's estimator for the survival probability. It plots the estimated survival function and log survival function. Besides making default plots, a user can easily make custom plots using saved analysis results and S-PLUS's powerful graphical capability. However, it does not provide estimates based on classical life-table results.

SPSS. SPSS handles right-censored data only. It plots the estimated survival function. It does not provide confidence intervals for the survival function. SPSS does allow the user to perform a life-table analysis on grouped event time data.

STATA. STATA handles right-censored data. It is the only one of the five packages to allow the data to be left truncated. In fact, it allows for more complicated truncation where patients may move in and out of the **risk set**. It provides log–log-transformed pointwise confidence intervals for the survival probabilities. It also allows computation of the Nelson–Aalen's estimator for the cumulative hazard function and for confidence intervals based on this statistic. It does not provide classical life-table analysis.

Hypothesis Tests

Statistical hypothesis tests for the equality of $K \geq 2$ populations using survival data are typically based on the weighted logrank test (*see* **Linear Rank Tests in Survival Analysis**). This test is based on quadratic forms constructed from the statistics

$$Z_j = \sum_{t_i} W(t_i) \left\{ \mathrm{d}_{ij} - Y_j(t_i) \frac{\mathrm{d}_{\bullet i}}{Y_\bullet(t_i)} \right\}, \qquad (1)$$

where the sum is over the event times, t_i, in the combined sample, $Y_j(t_i)$ is the number at risk, and d_{ij} the number of events at time t_i in the jth sample. Here $Y_\bullet()$ and $\mathrm{d}_{\bullet i}$ are the total number at risk and number dead in all K groups.

The most common choices for the weight function are $W(t) = 1$, which leads to the **Mantel–Haenszel**

logrank test and $W(t) = Y(t)$, which gives the Breslow–Gehan [3] version of the **Wilcoxon test**. These weights are available in all five packages. Other weight functions that can be used are summarized in Table 1. Further details on these statistics can be found in Chapter 7 of [12].

In addition to these tests STATA, BMDP, and S-PLUS all have tests of trend that can be obtained using these weight functions. Stratified tests are directly available in BMDP and STATA. In SAS, the "STRATA" command in PROC LIFETEST is used to invoke weighted logrank tests to compare the survival over subgroups, but the package does not directly compute stratified tests.

Semiparametric Regression Modeling

All five packages have extensive routines for the Cox proportional hazards model. These routines all allow for stratified regression models with both fixed and user defined time-dependent covariates. They all allow for right-censored data. At a minimum the packages return estimates of the risk coefficients, standard errors of the risk coefficients, and Wald tests of significance of the risk coefficients. Each of the packages has the ability to produce estimates and plots of the predicted survival function for models with fixed covariates. All the packages have some type of **diagnostic** plots.

The packages do differ in what they can compute in a number of ways. Table 2 summarizes the differences between the packages.

From the table we see that BMDP and SPSS do not allow delayed entry or left truncation of the data. Both S-PLUS and STATA allow for individuals to enter and leave the risk set at multiple times. SAS also allows for discontinuous time at risk by using multiple records for individuals and the counting process form of input. While one can handle categorical variables in all models, only SPSS allows the user to specify variables as categorical so that the appropriate multiple degree of freedom test is performed.

Table 1 Additional weight functions available for logrank tests

Weight	BMDP	SAS	S-PLUS	SPSS	STATA
Peto–Peto–Prentice [15] $W(t) =$ Pooled Kaplan–Meier ($S_{\mathrm{KM}}(t)$)	YES	NO	YES	NO	YES
Tarone and Ware [17] $W(t) = Y(t)^{1/2}$	YES	NO	NO	YES	YES
Fleming–Harrington [7] $S(t)^p(1 - S(t))^q$	NO	NO	For $q = 0$ only	NO	YES

Table 2 Comparison of features related to the Cox model

	BMDP	SPSS	SAS	S-PLUS	STATA
Type of Data					
Left truncation	N	N	Y	Y	Y
Discontinuous time at risk	N	N	Y	Y	Y
Categorical variables	N	Y	N	N	N
Inference properties					
Choice of likelihood for ties	N	N	Y	Y	Y
Allows robust variances	N	N	Y	Y	Y
Allows general risk function	N	N	N	Y	Y
Stepwise model building	Y	Y	Y	N	N
Frailty modeling	N	N	N	Y	N
Penalized likelihood	N	N	N	Y	N
Residuals					
Cox–Snell	Y	Y	Y	Y	Y
Martingale	N	Y	Y	Y	Y
Deviance	N	N	Y	Y	Y
Schoenfeld residuals	N	Y	Y	Y	Y
Weighted Schoenfeld residuals	N	N	Y	N	Y
Score residuals (DfBeta)	N	Y	Y	Y	N

When there are ties in the data, there are a number of choices for the **partial likelihood** in the literature (see [12]) SAS, S-PLUS, and STATA allow the user to chose between these likelihoods. S-PLUS, SAS, and STATA allow for robust versions of the variance of the Cox model including sandwich estimators that correct for correlation between estimates (see [13]). S-PLUS allows the user to use a **penalized likelihood** approach to smooth estimators. They also use the penalized likelihood to deal with random effect or frailty models.

SAS, SPSS, and BMDP all have routines for forward, backward, and stepwise model building (*see* **Variable Selection**). With the exception of SPSS these packages do not treat categorical variables as factors and the results of these automated features are suspect at times. We recommend model building by hand, which can be done quite efficiently with any of the packages. Table 2 also list which **residuals** the packages have available (*see* **Residuals for Survival Analysis**). Definitions of these residuals can be found in Chapter 11 of [12].

Parametric Modeling

Parametric models are often used to analyze survival data. The **accelerated failure-time model** has been suggested as an alternative to the Cox model when one wants to adjust for covariates. All of the packages, except SPSS, are able to perform parametric regression analysis. Parametric survival data analysis depends on the chosen distribution and residuals can be used in model diagnosis. Available distribution functions and residual types vary among the four packages. Packages can handle different types

Table 3 Capabilities in parametric analysis among various packages

	BMDP	SAS	S-PLUS	STATA
Available distributions:				
Exponential	Yes	Yes	Yes	Yes
Gompertz	No	No	No	Yes
Generalized gamma	No	Yes	No	Yes
Log-logistic	Yes	Yes	Yes	Yes
Lognormal	Yes	Yes	Yes	Yes
Rayleigh	Yes	No	Yes	No
Weibull	Yes	Yes	Yes	Yes
Fitting model with no covariate	Yes	Yes	Yes	Yes
Data type:				
Left censoring	No	Yes	Yes	No
Right censoring	Yes	Yes	Yes	Yes
Interval censoring	No	Yes	Yes	No
Left truncation	No	No	Yes	Yes
Time-dependent covariate	No	No	No	Yes
Tests significance for individual covariate:				
Wald	Yes	Yes	Yes	Yes
Likelihood-ratio	Yes	No	No	No
Score	Yes	No	No	No
Stratified regression model	No	No	Yes	Yes
Type of residuals:				
Cox–Snell	Yes	Yes	No	Yes
Deviance	No	No	Yes	Yes
Df-beta	No	No	Yes	No
Likelihood displacement	No	No	Yes	No
Martingale type	No	No	No	Yes
Standardized	Yes	No	Yes	No
Multiplicative frailty model:				
Gamma frailty	No	No	No	Yes
Inverse-Gaussian frailty	No	No	No	Yes

of censored data. Table 3 shows the capabilities of each of the four packages.

Discussion

The five packages we have discussed have the ability to compute many of the statistics a practicing biostatistician needs to analyze survival data. Several of the packages (SAS, S-PLUS, and STATA) allow the user to program additional statistical methods. These programs can also take results from the built-in routines and perform further analysis on these results. S-PLUS has an extensive collection of routines for expected survival, which can be used to compare survival data to known mortality rates.

The completeness of the packages and the ability to add additional routines make S-PLUS, SAS, or STATA the best packages we have found to analyze survival data. However, the sophistication and personal preference of the user makes the choice between the packages difficult.

We close this article with a plea to the manufacturer of this software to consider adding important missing routines for survival data. These include statistical procedures for **competing risks**. In particular, routines for the cumulative incidence function, and routines for inference for the cumulative incidence function are needed. Also missing are routines for either **Aalen's** [1] or Lin and Ying's [14] **additive hazard regression model** or for other semiparametric alternatives to the Cox model.

References

[1] Aalen, O.O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**, 907–925.

[2] Borgan, Ø. & Liestøl, K. (1990). A note on confidence intervals and bands for the survival curve based on transformations, *Scandinavian Journal of Statistics* **17**, 35–41.

[3] Breslow, N.E. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship, *Biometrika* **57**, 579–594.

[4] Brookmeyer, R. & Crowley, J.J. (1982). A confidence interval for the median survival time, *Biometrics* **38**, 29–41.

[5] Brown, J.B.W., Hollander, M. & Korwar, R.M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies, in *Reliability and Biometry: Statistical Analysis of Lifelength*, F. Proschan & R.J. Serfling, eds. SIAM, Philadelphia, pp. 327–354.

[6] Dain, B.J., Freeman, D.H. & Vredenburg, J.J. (1989). Comparison of different packages' survival test results, *Proceedings of the Statistical Computing Section*. American Statistical Association, Alexandria, pp. 315–318.

[7] Fleming, T.R. & Harrington, D.P. (1981). A class of hypothesis tests for one and two samples of censored survival data, *Communications in Statistics* **10**, 763–794.

[8] Fleming, T.R. & Harrington, D.P. (1984). Nonparametric estimation of the survival distribution in censored data, *Communications in Statistics* **13**(20), 2469–2486.

[9] Goldstein, R., Anderson, J., Ash, A. Craig, B., Harrington, D. & Pagano, M. (1989). Survival analysis software on MS/PC-DOS computers, *Journal of Applied Economics* **4**, 393–414.

[10] Harrell, F.E. & Goldstein, R. (1997). A survey of micro survival analysis software: the need for an integrated framework, *American Statistician* **51**, 360–373.

[11] Klein, J.P. (1991). Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators, *Scandinavian Journal of Statistics* **18**, 333–340.

[12] Klein, J.P. & Moeschberger M.L. (2003). *Survival Analysis: Statistical Methods for Censored and Truncated Survival Data*, 2nd Ed. Springer-Verlag, New York.

[13] Lee, E.W., Wei, L.J. & Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in *Survival Analysis: State of the Art*, J.P. Klein & P. Goel, eds. Kluwer Academic Publishers, Boston, pp. 237–248.

[14] Lin, D.Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61–71.

[15] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society* **A135**, 185–206.

[16] SURVIVAL, Salford System, 8880 Rio San Diego Dr., Suite 1045, San Diego, CA 92108, USA.

[17] Tarone, R.E. & Ware, J.H. (1977). On distribution-free tests for equality for survival distributions, *Biometrika* **64**, 156–160.

[18] Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data, *Journal of the American Statistical Association* **69**, 169–173.

[19] Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society* **B38**, 290–295.

JOHN P. KLEIN & MEI-JIE ZHANG

Survival Distributions and Their Characteristics

Many applications in biostatistics involve the modeling of lifetime data. In these applications the outcome of interest is the time, T, until some event occurs. This event may be death, the appearance of a tumor, the development of some disease, recurrence of a disease, conception, cessation of smoking, and so forth. Here T is a nonnegative **random variable** from a homogeneous population (*see* **Survival Analysis, Overview**).

In this article we examine how the distribution of T can be characterized. Four functions characterize the distribution of T: the *survival function*, which is the probability of an individual surviving beyond time t; the **hazard** *rate*, which is approximately the chance an individual of age t experiences the event in the next instant in time; the *probability density (or mass) function*, which is the approximate unconditional probability of the event occurring at time t; and the *mean residual life* at time t, which is the mean time to the event of interest, given the event has not occurred at t. If we know any one of these four functions, then the other three can be uniquely determined. These functions are introduced for continuous, discrete and mixed random variables in the following sections and the interrelationships among the four functions are discussed.

The distribution of the time to an event can also be characterized by the aging properties of the distribution of T. Aging classes are based on certain properties of one of the four basic quantities that describe the distribution of T. These classes are defined and some basic properties of these classes are discussed in the final section.

The Survival Function

The basic quantity employed to describe time-to-event phenomena is the survival function. This function, also known as the survivor function or survivorship function, is the probability an individual survives beyond time t. It is defined as

$$S(t) = \Pr(T \geq t).$$

In the context of equipment or manufactured item failures, $S(t)$ is referred to as the *reliability function*. Note that the survival function is a nonincreasing function with a value of 1 at the origin and 0 as t approaches infinity.

If T is a continuous random variable, then $S(t)$ is a continuous monotone decreasing function and the survival function is the complement of the cumulative distribution function $F(t) = \Pr(T \leq t)$. That is, $S(t) = 1 - F(t)$. The survival function is the integral of the probability density function $f(t)$. That is,

$$S(t) = \Pr(T \geq t) = \int_t^\infty f(u)\,\mathrm{d}u.$$

Thus, we have the following relationship:

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}.$$

Note that $f(t)\Delta t$ may be thought of as the "approximate" probability of the event occurring at time t and that $f(x)$ is a nonnegative function with the area under $f(x)$ being equal to one.

Example

A common distribution used in many applications is the **Weibull distribution** with probability density function $f(t) = \lambda\alpha t^{\alpha-1}\exp(-\lambda t^\alpha), \lambda > 0, \alpha > 0$. The **exponential distribution** is a special case of the Weibull distribution when $\alpha = 1$. The survival function for the Weibull distribution is $S(t) = \exp(-\lambda t^\alpha), \lambda > 0, \alpha > 0$. Survival curves with a common median of 6.93 are exhibited in Figure 1 for $\lambda = 0.26328$, $\alpha = 0.5$; $\lambda = 0.1$, $\alpha = 1$; and $\lambda = 0.00208$, $\alpha = 3$.

When T is a discrete random variable then the survival function is a nonincreasing left-continuous step function. If T can take on values $t_0 < t_1 < t_2 < \ldots$ with probability mass function (pmf) $p(t_j) = \Pr(T = t_j)$, $j = 1, 2, \ldots$, then

$$S(t) = \Pr(X \geq t) = \sum_{j:t_j \geq t} p(t_j).$$

Note that the survival function and probability mass function are related by

$$p(t_j) = S(t_j) - S(t_{j+1}).$$

Here we have defined $S(t) = \Pr(T \geq t)$ as was the case in [3] and [5]. This definition was used to make later formulas for the discrete case simpler. Other

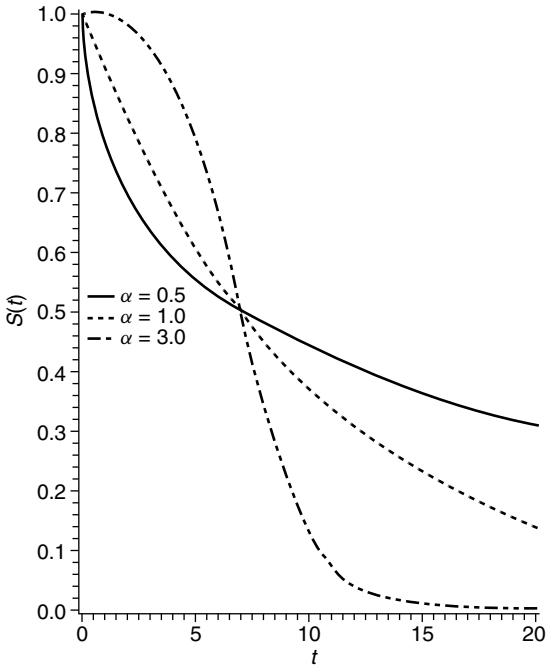

Figure 1 Comparison of Weibull survival functions

authors (see [4] and [6]) have defined $S(t) = \Pr(T > t)$, which makes the relationship $S(t) = 1 - F(t)$ hold for both the discrete and continuous case.

The Hazard Function

A basic quantity, foundational in survival analysis, is the hazard function. This function is also known as the conditional failure rate in reliability, the force of mortality in demography, the age-specific failure rate in epidemiology, the inverse of Mill's ratio in economics or simply as the hazard rate. The hazard rate is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t | T \ge t)}{\Delta t}. \quad (1)$$

The hazard rate is a nonnegative function. It tells us how quickly individuals of a given age are experiencing the event of interest. The quantity $h(t)\Delta t$ is the approximate probability that an individual who has survived to age t will experience the event in the interval $(t, t + \Delta t)$.

This function is particularly useful in determining the appropriate failure distributions utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time. There are many general shapes for the hazard rate. Some generic types of hazard rate are increasing, decreasing, constant, bathtub-shaped or hump-shaped. Models with increasing hazard rates arise when there is natural aging or wear-out. Decreasing hazard functions are much less common, but find occasional use when there is a likelihood of very early failure, as in certain types of electronic devices, or in patients experiencing certain types of transplant. Decreasing hazard rates often arise as models for heterogeneous populations where the hazard rates of members of the population are random (*see* **Frailty**). Most often a bathtub-shaped hazard is appropriate in populations followed from birth. Most population mortality data follow this type of hazard function: early in the process, deaths result primarily from infant diseases; then the death rate stabilizes; later, an increasing hazard rate sets in, due to the natural aging process. Finally, if the hazard rate is increasing early and eventually begins declining, then the hazard is termed 'hump-shaped'. This type of hazard rate is often used in modeling survival after successful surgery where there is an initial increase in risk due to infection, hemorrhaging, or other complications just after the procedure, followed by a steady decline in risk as the patient recovers.

If T is a continuous random variable, then

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \ln[S(t)].$$

A related quantity is the cumulative hazard function, $H(t)$, defined by

$$H(t) = \int_0^t h(u)\,\mathrm{d}u = -\ln[S(t)].$$

Thus for continuous lifetimes we have the following relationship:

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)\,\mathrm{d}u\right\}.$$

The Weibull distribution is flexible enough to accommodate increasing ($\alpha > 1$), decreasing ($\alpha < 1$), and constant hazard rates ($\alpha = 1$). Figure 2 plots hazard rates, $h(x) = \alpha\lambda x^{\alpha-1}$, for the same Weibull

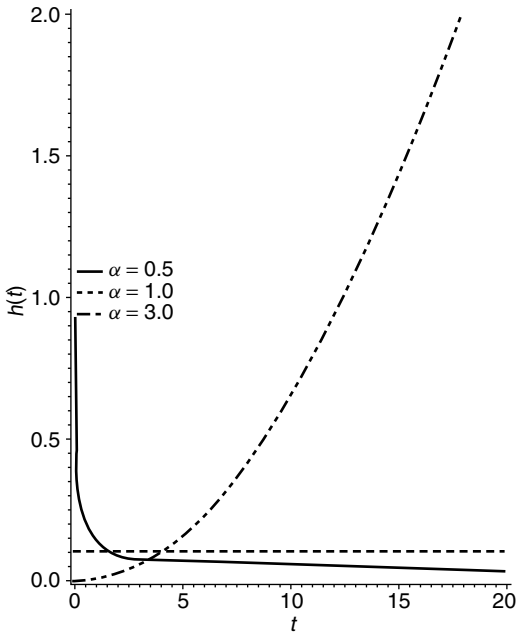

Figure 2 Comparison of Weibull hazard functions

distributions as in Figure 1. One can see that, though the three survival functions have the same basic shape, the three hazard functions are dramatically different.

When T is a discrete random variable, the hazard function is

$$h(t_j) = \Pr(T = t_j | T \geq t_j) = \frac{p(t_j)}{S(t_j)}, \quad j = 1, 2, \ldots.$$

Since $p(t_j) = S(t_j) - S(t_{j+1})$, we have

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)}, \quad j = 1, 2, \ldots,$$

so the survival function is related to the hazard function by

$$S(t) = \prod_{j:t_j<t} [1 - h(x_j)].$$

For discrete lifetimes the "cumulative hazard" function is defined by

$$H(t) = \sum_{j:t_j<t} h(t_j). \tag{2}$$

Notice that for this definition the relationship $S(t) = \exp[-H(t)]$ no longer holds true. Some authors [3] prefer to define the cumulative hazard for discrete lifetimes as

$$H(t) = \sum_{t_j<t} \ln[1 - h(t_j)]. \tag{3}$$

Note that for this definition the relationship for continuous lifetimes, $S(t) = \exp[-H(t)]$, will then be preserved for discrete lifetimes. If the $h(t_j)$ are small, (2) will be a first-order approximation to (3).

The hazard rate is a well-defined quantity for the case where T has both discrete and continuous components. In this case the hazard function defined by (1) will have a continuous part, $h_c(t)$ and a discrete part with mass h_j at time $t_1 < t_2 < \ldots$. The survival function in this case can be expressed as

$$S(t) = \exp\left\{-\int_0^t h_c(u)\,du\right\} \prod_{j:t_j<t} (1 - h_j).$$

For any survival function one can express the relationship between the hazard rate and the survival function by the using the notion of a product integral. For a function, $G()$, define the product integral of $1 - dG(u)$ over the range a to b by

$$P_a^b[1 - dG(u)] = \lim \prod_{k=1}^{r} \{1 - [G(u_k) - G(u_{k-1})]\},$$

where $a = u_1 < \cdots < u_r = b$ and the limit is taken as $r \to \infty$ and $u_k - u_{k-1} \to 0$. Here G is a function of locally bounded variation which is continuous from the right and has finite left-hand limits. If we define the cumulative hazard rate as

$$H(t) = \int_0^t h_c(u)\,du + \sum_{j:t_j<t} h_j,$$

then the survival function in the continuous, discrete or mixed case is given by

$$S(t) = P_0^t[1 - dH(u)].$$

Because of this property the product integral plays an important role in survival analytic techniques.

The Mean Residual Life Function

The fourth basic parameter of interest is the mean residual life (mrl) at time t. This parameter measures,

for individuals of age t, their expected remaining lifetime. It is defined as

$$\text{mrl}(t) = \text{E}(T - t | T \geq t).$$

It can be shown, using integration by parts or a partial summation formula, that the mean residual life is the area under the survival curve to the right of t divided by $S(t)$. Note that the mean life, $\mu = \text{mrl}(0)$, is the total area under the survival curve.

For a continuous random variable we have

$$\text{mrl}(t) = \frac{\int_t^\infty (u - t) f(t)\, \text{d}u}{S(t)} = \frac{\int_t^\infty S(u)\, \text{d}t}{S(t)}$$

and

$$\mu = \text{E}(T) = \int_0^\infty u f(u)\, \text{d}u = \int_0^\infty S(u)\, \text{d}u.$$

Also the variance of T is related to the survival function by

$$\text{var}(T) = 2 \int_0^\infty u S(u)\, \text{d}u - \left[\int_0^\infty S(u)\, \text{d}u \right]^2.$$

In some applications the median residual life, rather then the mean residual life, is of interest. To define this quantity recall that the $100p$th percentile (or pth **quantile**) of a random variable X with cumulative distribution function (survival function) $F(x)(S(x))$ is the value x_p such that

$$F(x_p) \geq p \quad \text{and} \quad S(x_p) \geq 1 - p.$$

The median lifetime is the 50th percentile, $x_{0.5}$, of the distribution of X. If X is a continuous random variable then the pth quantile is found by solving the equation $S(x_p) = 1 - p$. It follows that the median lifetime (mdrl), for a continuous random variable X, is the value $x_{0.5}$ such that

$$S(x_{0.5}) = 0.5.$$

The median residual life time of T at time t, mdrl(t), is defined as the median time to the event for an individual who has survived to time t. That is, mdrl(t) is the solution to the equation

$$\frac{S(\text{mdrl}(t))}{S(t)} = 0.5.$$

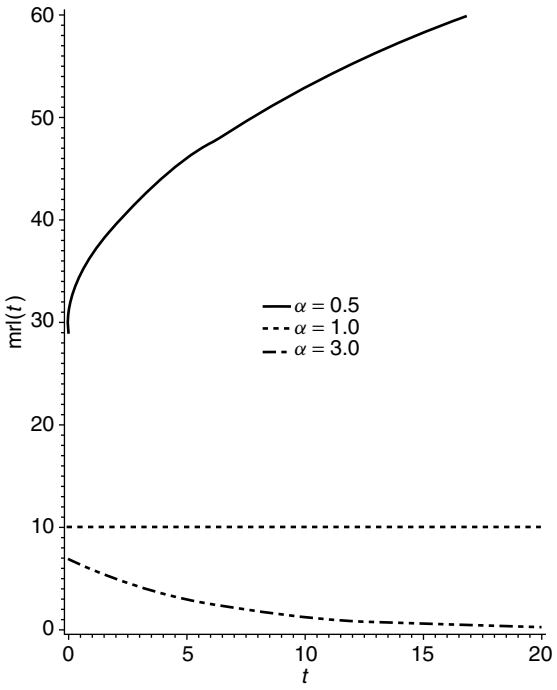

Figure 3 Comparison of Weibull mean residual life functions

The population median is simply the median residual life at time 0.

To illustrate these quantities consider the three Weibull distributions considered earlier. Figure 3 shows the mean residual life function for the Weibull models with $\alpha = 0.5$, 1.0 and 3.0. As the figure shows, the mean residual life is constant for the exponential distribution ($\alpha = 1$), decreasing for the case where $\alpha = 3$ and increasing for the case where $\alpha = 0.5$. Note that the trend in the mean residual life is reversed from the trend in the hazard rate in that when the hazard rate is increasing, reflecting aging, the mean residual life is decreasing. Figure 4 depicts the median residual life functions for the three Weibull models. The shapes of the functions are quite similar to those of the mean residual life functions.

Relationship Between Characterizations

Interrelationships between the characterizations discussed earlier, for a continuous lifetime T, may be

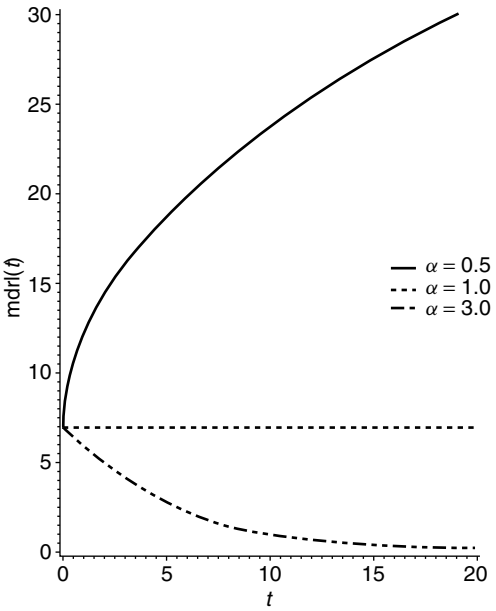

Figure 4 Comparison of Weibull median residual life functions

summarized as follows:

$$
\begin{aligned}
S(t) &= \int_t^\infty f(u)\,\mathrm{d}u \\
&= \exp\left\{-\int_0^t h(u)\,\mathrm{d}u\right\} \\
&= \exp\{-H(t)\} \\
&= \frac{\mathrm{mrl}(0)}{\mathrm{mrl}(t)}\exp\left\{-\int_0^t \frac{\mathrm{d}u}{\mathrm{mrl}(u)}\right\};
\end{aligned}
$$

$$
\begin{aligned}
f(t) &= -\frac{\mathrm{d}}{\mathrm{d}t}S(t) \\
&= h(t)S(t) \\
&= \left(\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{mrl}(t)+1\right)\left(\frac{\mathrm{mrl}(0)}{\mathrm{mrl}(t)^2}\right) \\
&\quad \times \exp\left\{-\int_0^t \frac{\mathrm{d}u}{\mathrm{mrl}(u)}\right\};
\end{aligned}
$$

$$
\begin{aligned}
h(t) &= -\frac{\mathrm{d}}{\mathrm{d}t}\ln[S(t)] \\
&= \frac{f(t)}{S(t)} \\
&= \frac{\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{mrl}(t)+1\right)}{\mathrm{mrl}(t)};
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{mrl}(t) &= \frac{\int_t^\infty S(u)\,\mathrm{d}u}{S(t)} \\
&= \frac{\int_t^\infty (u-t)f(u)\,\mathrm{d}u}{S(t)}.
\end{aligned}
$$

For a discrete random variable we have the following relationships:

$$
\begin{aligned}
S(t) &= \sum_{j:t_j \geq t} p(t_j) \\
&= \prod_{j:t_j < t} [1 - h(t_j)].
\end{aligned}
$$

If T is an integer-valued random variable with mean residual life at time k equal to $m_k, k = 0, 1, 2, \ldots,$ and m_0 is finite, then we have

$$
S(k) = \frac{1+m_0}{m_k}\prod_{j=0}^{k}\frac{m_j}{1+m_j}.
$$

Also, for any discrete survival function, we have

$$
\begin{aligned}
p(t_j) &= S(t_j) - S(t_{j+1}) \\
&= h(t_j)S(t_j), \quad j = 1, 2, \ldots; \\
h(t_j) &= \frac{p(t_j)}{S(t_j)};
\end{aligned}
$$

and

$$
\mathrm{mrl}(t) = \frac{[t_{k+1}-t]S(t_{k+1}) + \sum_{j:t_j \geq t_{k+1}} [t_{j+1}-t_j]S(t_{j+1})}{S(t)},
$$

$$
\text{for } t_k \leq t < t_{k+1}.
$$

Classes of Aging Distributions

An important characteristic of survival distribution is its aging properties (*see* **Aging Models**). There are a number of classes that have been suggested in the literature to categorize distributions based on their aging properties or their dual. The first aging class is

the class of increasing hazard rate (IHR) distributions and the dual class of decreasing hazard rate (DHR) distributions. A survival distribution is said to be in the IHR (DHR) class if and only if

$$\frac{S(t+x)}{S(t)} = S(x|t) \text{ is decreasing (increasing) in } t \text{ for all } x.$$

The definition says that T has the IHR aging property if the probability an individual of age t survives an additional period of time x is decreasing with time. If T is a continuous random variable then an equivalent definition of the IHR (DHR) class is that the hazard rate $h(t)$ is increasing (decreasing) for all t. Examples of distributions that fall in the IHR class are the Weibull distribution with $\alpha > 1$ and the gamma distribution with shape parameter greater than one.

A second, more general, aging class is the class of increasing (decreasing) hazard rate on the average, IHRA (DHRA), distributions. A distribution is said to fall in the IHRA (DHRA) class if and only if

$$-\left(\frac{1}{t}\right)\ln[S(t)] \text{ is increasing (decreasing) in } t. \quad (4)$$

The definition arises by declaring a distribution to be in the IHRA class when its cumulative hazard rate, $-\ln[S(t)]$ is increasing faster than the cumulative hazard rate of an exponential random variable, t. Since the exponential distribution reflects a model with no aging, this class is one of the distributions for which individuals are, on the average, aging. There are several equivalent definitions of a IHRA class. Since (4) implies that $S^{1/t}(t)$ is increasing in t we have that T is in the IHRA class if and only if $S(\theta t) \geq S^{\theta}(t)$. A second characterization of the IHRA class is that if T is in the IHRA class, then for any $\lambda > 0$ the quantity $S(t) - \exp(-\lambda t)$ has at most one change of sign, and if it does have a change in sign then it is from positive to negative. The class of IHRA distributions is larger than the class of IHR distributions in that every IHR distribution is an IHRA distribution but the converse is not true.

A third aging class is the class of decreasing (increasing) mean residual life, DMRL (IMRL), distributions. A distribution is said to be in the DMRL (IMRL) class if

$$\text{mrl}(t) = \frac{\int_t^{\infty} S(x)\,\mathrm{d}x}{S(t)} \text{ is decreasing (increasing) in } t.$$

This aging class, which includes all IHR models, is one where the mean remaining life of an individual of age t is becoming shorter as t increases.

A fourth aging class is the class of new better (worst) than used, NBU (NWU), distributions. Here a distribution is in the NBU (NWU) class if and only if

$$S(x+t) \leq (\geq) S(x)S(t), \quad \text{for any } x \text{ and } t.$$

An equivalent definition for the NBU class is

$$\frac{S(x+t)}{S(t)} = \Pr(T \geq x+t | T \geq t) \leq \Pr(T \geq x) = S(x).$$

From this second definition we see that T has an NBU distribution if the probability an individual of age t lives an additional x time units is smaller than the probability an individual of age 0 survives to age x. This aging class includes all the IHRA distributions.

A fifth aging class is the class of new better (worse) than used in expectation, NBUE (NWUE), distributions. A distribution is in the NBUE (NWUE) class if its mean, μ, is finite and

$$\int_t^{\infty} S(u)\,\mathrm{d}u \leq (\geq)\mu S(t), \quad \text{for all } t.$$

The NBUE class is one where the mean residual life of an individual of age t is less that the mean of an individual of age 0.

A final aging class is the class of harmonic new better (worse) than used in expectation, HNBUE (HNWUE), distributions. A distribution is said to be in the HNBUE (HNWUE) class if its mean is finite and

$$\int_t^{\infty} S(u)\,\mathrm{d}u \leq \mu \exp\left(\frac{-t}{\mu}\right).$$

An equivalent definition for the HNBUE class is

$$\left\{\frac{1}{t}\int_0^t \frac{\mathrm{d}x}{\text{mrl}(x)}\right\}^{-1} \leq \text{mrl}(0).$$

This means that for a HNBUE distribution the integral harmonic value of the residual life of an individual of age t is smaller than the same quantity for a newly born individual.

The aging classes are ordered as follows:

$$\text{IHR} \Longrightarrow \text{IHRA} \Longrightarrow \text{NBU} \Longrightarrow \text{NBUE} \Longrightarrow \text{HNBUE};$$

$$\text{IHR} \Longrightarrow \text{DMRL} \Longrightarrow \text{NBUE} \Longrightarrow \text{HNBUE};$$

$$\text{DHR} \Longrightarrow \text{DHRA} \Longrightarrow \text{NWU} \Longrightarrow \text{NWUE} \Longrightarrow \text{HNWUE};$$

$$\text{DHR} \Longrightarrow \text{IMRL} \Longrightarrow \text{NWUE} \Longrightarrow \text{HNWUE}.$$

Further discussion of these failure classes can be found in [1] and [2].

References

[1] Barlow, R.E. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, New York.

[2] Basu, A.P. & Ebrahimi, N. (1986). HNBUE and HNWUE distributions – a survey, in *Reliability and Quality Control*, A.P. Basu, ed. North-Holland, New York, pp. 33–47.

[3] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, New York.

[4] Klein, J.P. & Moeschberger, M.L. (1997). *Survival Analysis: Methods for Censored and Truncated Data*. Springer-Verlag, New York.

[5] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

[6] Lee, E.T. (1992). *Statistical Methods for Survival Data Analysis*. Wiley, New York.

(*See also* **Parametric Models in Survival Analysis**)

JOHN P. KLEIN

Synergy of Exposure Effects

Although environmental regulation of chemical exposures is typically based on laboratory studies in which animals are dosed to single agents, humans are invariably exposed to mixtures. People who drink may smoke as well. Cigarette smoke itself is an example of a mixture. Modeling and predicting the effects of combined exposures, or of exposures experienced in a particular temporal sequence, remain challenges to toxicologists and epidemiologists.

The effect associated with several exposures in combination sometimes far exceeds what would have been expected on the basis of their separate effects, a phenomenon known as "synergism" (or "synergy"). One example is the occurrence of lung cancer in relation to exposure to cigarette smoking and arsenic, where the risk in those exposed to both is high [6]. While the etiologic basis for this particular mutual enhancement of effect is not well understood, the demonstration of synergism of exposures can provide important insight into causal mechanisms and can suggest strategies for intervention. For example, mental retardation invariably ensues when a child has the metabolic disorder phenylketonuria and also consumes the amino acid phenylalanine in his or her diet; removal of dietary phenylalanine prevents the adverse effect. Negative synergism, known as "antagonism", can also arise, as when exposure to the polio virus follows exposure to the polio vaccine. Examples of more subtle forms of antagonism include scenarios where two different chemical exposures compete for the same population of receptor sites or when one interferes with the absorption or metabolism of the other.

While most would agree that epidemiologic synergism among exposures exists, defining it is problematic. Usually "synergism" is said to be present when the effect of exposure to a combination of factors exceeds the sum of the separate, factor-specific effects. We must then define what it means to "sum" effects. Such a definition would establish a model for independence, compared with which positive departures could be considered synergistic, and negative departures, antagonistic. Synergism and antagonism cannot be defined except in relation to some definition for independence of effect, except in rare scenarios where only one of the two factors has an effect when experienced without the other, and the effect of the combined exposures exceeds that of the one factor alone. The phenylketonuria example and the polio vaccine example were both of the latter, unambiguous type.

Most instances of mutual enhancement of effect are not of that simple, pure form, so a more general definition is needed. When the outcome of interest is **binary**, e.g. the occurrence of a particular disease, the exposure–response formulation often includes specification of a function to serve as a "link" between the risk of disease and the exposures. If the risk of disease, r, is first subjected to a "logit" **transformation**, by taking the logarithm of the "odds", $r/(1-r)$, then effect additivity on this inherently multiplicative scale is very different from additivity on the untransformed, "absolute", i.e. additive scale (*see* **Additive Model**). Thus, for example, if two factors each increase the risk of disease, and their combined effect can be correctly represented by a **logistic** model with no **interaction** ("product") term, then an additive formulation *would* require a positive interaction term. Conversely, the adequacy of an additive model for the combined effects would mean that a **multiplicative model** would require an interaction term. The distinction between this kind of statistical interaction, which corresponds simply to departure from additivity on some mathematically convenient scale, and biologic interaction, which corresponds to true biologic mutual enhancement of a causal mechanism, has long been appreciated by biostatisticians and epidemiologists [7, 16], who have searched for a formulation with biologic interpretability. The notion of synergism or "biologic interaction" can be defined as "the inter-dependent operation of two or more causes to produce disease" [15]. A related concept is that of independence of two factors (say, A and B) in a public health sense, which is said to occur "when the number of cases of disease that would occur in the population does not depend on the extent to which A and B occur together in the same individuals" [15].

While synergism or antagonism can involve causative factors or protective factors, the choice of null model can be different for the joint effect of protective factors [2, 23]; or for the joint effect of a protective and a causative factor. The discussion that follows will apply only to the combined effect

of causative factors, i.e. factors that each can increase the risk of the disease under study.

Rothman [14] proposes a conceptual framework for disease causality, where several "component" causes act together to form in their aggregate a "sufficient" cause, for which each component cause is necessary to its completion. (This conceptual model can be extended to continuous exposures by supposing that exceedence of a particular level of exposure is required for each sufficient cause in which the continuous exposure participates.) A single exposure may participate in (i.e. be a necessary component in) several different sufficient causes if there are several distinct pathways that involve it and can lead to the disease. If two exposures participate in the same sufficient cause, then that cause can only produce the disease when both are present, and such a pathway would imply synergism between the two exposures in a biologic sense [14] (*see* **Causation**).

For a single exposure, the difference between the incidence of the disease among those with the factor and among those without the factor can be interpreted as the **incidence rate** of completion for those sufficient causes that require that factor. By extension, the incidence rate for those with two factors, say A and B, minus those with neither is the sum of the incidence rate for completion of those sufficient causes involving A and not B and the incidence rate for completion of those sufficient causes involving B and not A, unless there exist one or more sufficient causes that require both A and B. In this way, the "causal pies" model proposed by Rothman leads naturally to the following model for independence based on incidence rates:

$$I_{AB} - I_{\overline{AB}} = I_{A\overline{B}} - I_{\overline{AB}} + I_{\overline{A}B} - I_{\overline{AB}}, \qquad (1)$$

where the overbar indicates the absence of the factor. This can be seen as additivity on the "risk difference" or absolute scale, and Rothman argues that absolute additivity is the only proper epidemiologic null model for "independent" effects. (Notice, however, that unless the lifetime risks are very small, model (1) does not imply the independence of A and B in the public health sense defined above: to the extent that A and B co-occur, the A-dependent and B-dependent pathways will compete for the same victims.) Such a model can easily be fitted to cohort data using standard statistical packages, such as SAS (the GENMOD procedure) and GLIM [19] (*see* **Software, Biostatistical**). The data are considered to provide evidence for synergism if the fit is significantly improved by inclusion of an interaction term, i.e. a nonzero γ, in the following model:

$$R[D|A,B \text{ status}] = \mu + \alpha(A \text{ present}) + \beta(B \text{ present}) + \gamma(\text{ both } A \text{ and } B \text{ present}),$$

where R denotes the incidence of the disease. For multilevel exposures, d_1 and d_2, to two exposures that now are not simply present or absent, the above zero-γ null formation generalizes to linearity in $f(d_1)$ and $g(d_2)$, for exposure-specific exposure–response functions $f(\cdot)$ and $g(\cdot)$.

It is instructive to think about what self-independence would mean for a single exposure. Exposure to 40 units of an exposure can be thought of as a combined exposure to two doses of 20 units each, or to 30 units, together with 10 units, and so on. For these separate exposures to combine independently, one can show that the exposure–response must be linear. Low-level alpha **radiation** provides an example where self-independence is plausible. Irreparable chromosomal damage at the cellular level caused by bombardment by a passing alpha particle is random, rare, and heritable, providing radiobiologists with a strong theoretical justification for a linear exposure–response.

Returning to the binary exposure scenario, when a **case–control** design is used to study the etiology of a rare disease, the incidence rates in (1) cannot be estimated. However, dividing through by the background incidence, and letting RR_{AB} denote the **relative risk** for the combined exposure, relative to the background risk, leads to the approximate relation:

$$RR_{AB} = RR_{A\overline{B}} + RR_{\overline{A}B} - 1.0, \qquad (2)$$

and thus independence can be assessed using case–control data. Wacholder & Weinberg [20] provide methods for evaluating the fit of the additive model to case–control data, under various designs.

We have thus far considered effects of two exposures on disease risk, presuming that the unexposed state is unambiguously defined, whereas Greenland & Poole [4] point out that the choice of coding of one level as "unexposed" can be arbitrary. One example is sex, where males could be considered unexposed (to

being female) or females could be considered unexposed (to being male). This raises the question as to how such factors should be incorporated in evaluations of synergistic effects. One can use algebra to show, however, that the above additivity criterion for independence is invariant under recoding of such a variable.

Under the usual understanding of the sufficient causes model, the disease invariably occurs once all components of a particular sufficient cause have been assembled, and in this sense Rothman's conceptual model is deterministic. Others have preferred to begin with a more stochastic conceptual approach and have nonetheless arrived at the same null model (1) for independence. In this way, divergent approaches to conceptualizing independence converge to a common mathematical formulation for the null model.

The stochastic approach has its roots in toxicology. Bliss [1] defines "independent joint action" between "poisons" as meaning that "the poisons or drugs act independently and have different modes of toxic action". Finney [3] provides mathematical rigor by specifying that "simple independent action" between two factors obtains if the outcomes are probabilistically independent. For one exposed to levels d_1 and d_2 of two different factors, the probability of avoiding the outcome (1 minus the risk) can be denoted $Q(d_1, d_2)$, so that the background "spontaneous" probability of nonoccurrence is $Q(0, 0)$. Probabilistic independence implies the following relationship:

$$Q(d_1, d_2) = \frac{Q(d_1, 0)Q(0, d_2)}{Q(0, 0)}. \qquad (3)$$

This reflects a scenario where the two causal mechanisms are completely separate and unrelated and where each of the exposure-dependent causal processes is independent of the background causal processes, i.e. those mechanisms that can produce the disease in the absence of either exposure.

Weinberg [23] describes a paradigm for this idealized model (3) in relation to two hunters, unaware of each other's presence, but shooting at the same ducks. To survive, a duck must stay clear of both. Under this independence scenario, the probability that the duck will survive some interval of time is the product of the probabilities that it escapes both hunters and also does not die of causes unrelated to being shot. This is the simplest probability-based notion of independence, corresponding to the situation where the exposures have completely separate biological modes of action.

This model, which can be written in **generalized linear model** form as additive in the log of the nonresponse, $\ln[Q(d_1, d_2)]$, has been applied to assessing synergism in animal experiments [21] and Korn & Liu [9] proposed a **Mantel–Haenszel**-type statistic for synergism based on follow-up with continuous failure times.

Mathematically, models (1) and (3) are equivalent. If one integrates risk over any fixed length of time, then the additive formulation (1) can be seen (replacing the factor A by the continuous d_1 and B by d_2) as equivalent to model (3), where the function $Q(d_1, d_2)$ is interpreted as the probability of survival without the disease over the specified follow-up interval for those with rates as given by (1). Conversely, the model given by (3) can be shown to imply model (1), because the negative of $\ln[Q(d_1, d_2)]$ converges in the limit, as the interval of time (hence the associated risk) becomes small, to the incidence rate associated with the combined exposure (d_1, d_2).

In the context of a rare disease, either formulation leads naturally to the case–control-based index for synergism proposed by Rothman [13]:

$$S = \frac{RR_{AB} - 1}{RR_{A\overline{B}} + RR_{\overline{A}B} - 2},$$

who also defines a synergism index for cohort data and provides **standard error** formulas for computing **confidence intervals**. Another index for synergism, resembling the usual interaction term in **analysis of variance**, provides a direct estimate for the **excess risk** associated with synergistic effects among those with both exposures: $T = R_{AB} - R_{\overline{A}B} - R_{A\overline{B}} + R_{\overline{A}\overline{B}}$ has certain advantages but can only be estimated in the context of a **cohort study** [7]. Wahrendorf et al. [21] propose a different index that can be used in cohort studies:

$$W = \frac{Q(A, B)Q(0, 0)}{A(\overline{A}B)Q(A, \overline{B})},$$

which, for a rare outcome, is approximately $\exp(-T)$, and should, in general, be one under independence. Weinberg [23] refers to W as a "health ratio" similar to a risk ratio (*see* **Relative Risk**), but interpretable

as the proportion *avoiding* the disease divided by the expected (based on independence) proportion avoiding disease, among those with both exposures. Simulations comparing W, T, and a third index, G, proposed by Korn & Liu [9], revealed that tests based on ln (W) had close-to-nominal size and relatively good **power** [12].

Statistics representing the fractional excess in the disease rate that is attributable to the synergistic effects of two factors have been developed, based on this model for independence. Hamilton [5] defines the "proportion of disease attributable to synergism" by taking the difference between the observed and expected risk among those with both divided by the overall risk and multiplying by the **prevalence** in the population of the combined exposure. Walker [22] later defines the "proportion of disease attributable to the combined effect of two factors" differently, by subtracting the expected from the observed rate of disease in those exposed to both factors and dividing by the observed rate. The difference between the indices proposed by Hamilton, and Walker, is that the Hamilton index divides by the overall population rate, while that of Walker divides by the disease rate among those with both exposures. Thus, the latter is more focused on etiology and the former on public health.

Darroch & Borkent [2] have recently revisited the question of how to assess the proportion of disease attributable to synergistic effects, within the context of a deterministic paradigm described by Hamilton [5]. They begin with the presumption that there are six types of people in the population: those who will get the disease regardless of their exposure to A and/or B; those who will *not* get the disease regardless of their exposure to A and/or B; those who will get the disease with A but will not with B or with neither; those who will get the disease with B but will not with A or with neither; those who will get the disease with A or with B but will not with neither; and those who will only get the disease in the presence of both A and B. Because only four parameters are estimable on the basis of relating the combined exposure to the observable risk of the disease, the fraction in the synergistic category, who would only get the disease with both exposures, cannot be directly estimated. Darroch & Borkent [2] propose an estimate based on maximum entropy, and illustrate its application to the problem of partitioning the cases of lung cancer among those exposed to both smoking and radiation into four parts: the fraction caused by smoking alone; the fraction caused by radiation alone; the fraction that would have developed even with neither exposure; and the fraction that developed as a result of the combined exposure. Extensions of this approach to multilevel exposures remain to be developed.

Both conceptualizations, the deterministic one championed by Rothman and Hamilton, and the probabilistic one preferred by toxicologists, are useful for clarifying our thinking about causality, but both have limitations. Suppose there are two distinct sufficient causes for the disease of interest, one requiring A and one requiring B. Koopman [8] points out that if they have a component cause in common, say C, then A and B will compete for the same pool of susceptibles, i.e. those with C, and this competition can produce apparent antagonism. This will be true even if we presume that the two sufficient causes are independent among those with C. Such shared causes may be common. Genetic factors, for example, can interact with exposures to produce disease; the existence of genetically based contributory causes, while usually unknown to the investigator, may be the rule rather than the exception.

Seen in the hunter paradigm for probabilistic independence, some ducks, depending on their size and coloring, are easier to see (hence to shoot) than others. Such variation among individuals in inherent susceptibility can produce apparent nonindependence, even when the causal processes (the two hunters) are truly functioning independently at the level of each individual at risk. Darroch & Borkent allow for the resulting inherent nonidentifiability of parameters within a deterministic conceptual framework by proposing a maximum entropy approach, while others [10, 23] compute upper and lower bounds for synergism indices that allow for covarying susceptibilities to A and to B across individuals in the population.

Whenever the epidemiologist begins with parameter relationships observed in the data and draws inferences regarding the likelihood that causal mechanisms for two different exposures are biologically linked, warning bells should sound: the same epidemiologic data can be consistent with very different underlying biologic scenarios, as discussed at length by Thompson [18]. Nevertheless, epidemiology can provide important clues regarding interdependent causal mechanisms.

A related set of issues not discussed here involves the epidemiologic identification of "initiator/promoter" relationships among pairs of exposures, where their temporal ordering can have an effect on risk [11] and [17] (*see* **Effect Modification**).

References

[1] Bliss, C.I. (1939). The toxicity of poisons applied jointly, *Annals of Applied Biology* **26**, 585–615.

[2] Darroch, J. & Borkent, M. (1994). Synergism, attributable risk and interaction for two binary exposure factors, *Biometrika* **81**, 259–270.

[3] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, New York.

[4] Greenland, S. & Poole, C. (1988). Invariants and noninvariants in the concept of interdependent effects, *Scandinavian Journal of Work Environment and Health* **14**, 125–129.

[5] Hamilton, M.A. (1979). Choosing the parameter for a 2×2 table or a $2 \times 2 \times 2$ table analysis, *American Journal of Epidemiology* **109**, 362–375.

[6] Hertz-Picciotto, I., Smith, A.H., Holtzman, D., Lipsett, M. & Alexeeff, G. (1992). Synergism between occupational arsenic exposure and smoking in the induction of lung cancer, *Epidemiology* **3**, 23–31.

[7] Hogan, M., Kupper, L., Most, B. & Haseman, J. (1978). Alternatives to Rothman's approach for assessing synergism (or antagonism) in cohort studies, *American Journal of Epidemiology* **108**, 60–67.

[8] Koopman, J.S. (1977). Causal models and sources of interaction, *American Journal of Epidemiology* **106**, 439–444.

[9] Korn, E. & Liu, P.-Y. (1983). Interactive effects of mixtures of stimuli in life table analysis, *Biometrika* **70**, 103–110.

[10] Miettinen, O.S. (1982). Causal and preventive interdependence, *Scandinavian Journal of Work Environment and Health* **8**, 159–168.

[11] Moolgavkar, S., Luebeck, E., Krewski, D. & Zielinski, J. (1993). Radon, cigarette smoke, and lung cancer: a re-analysis of the Colorado Plateau uranium miners' data, *Epidemiology* **4**, 204–217.

[12] Piegorsch, W., Weinberg, C. & Haseman, J. (1986). Testing for simple independent action between two factors for dichotomous response data, *Biometrics* **42**, 413–419.

[13] Rothman, K. (1976). The estimation of synergy or antagonism, *American Journal of Epidemiology* **103**, 506–511.

[14] Rothman, K.J. (1986). *Modern Epidemiology*. Little, Brown, & Company Boston.

[15] Rothman, K., Greenland, S. & Walker, A. (1980). Concepts of interaction, *American Journal of Epidemiology* **112**, 467–470.

[16] Saracci, R. (1980). Interaction and synergism, *American Journal of Epidemiology* **112**, 465–466.

[17] Thomas, D., Pogoda, J., Langholz, B. & Mack, W. (1994). Temporal modifiers of the radon – smoking interaction, *Health Physics* **66**, 257–262.

[18] Thompson, W. (1991). Effect modification and the limits of biological inference from epidemiologic data, *Journal of Clinical Epidemiology* **44**, 221–232.

[19] Wacholder, S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences, *American Journal of Epidemiology* **123**, 174–184.

[20] Wacholder, S. & Weinberg, C. (1994). Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling, *Biometrics* **50**, 350–357.

[21] Wahrendorf, J., Zentgraf, R. & Brown, C.C. (1981). Optimal designs for the analysis of interactive effects of two carcinogens or other toxicants, *Biometrics* **37**, 45–54.

[22] Walker, A. (1981). Proportion of disease attributable to the combined effect of two factors, *International Journal of Epidemiology* **10**, 81–85.

[23] Weinberg, C.R. (1986). Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome, *American Journal of Epidemiology* **123**, 162–173.

CLARICE R. WEINBERG

Systematic Error

Systematic error is the **bias** that results when a data-gathering process or method of analysis leads to results expected to deviate from the true quantity to be estimated. Unlike **random error**, systematic error is not ameliorated by increasing sample size, which only serves to obtain more precise **biased** estimates of the desired quantity (*see* **Random Error** for a simple example of systematic error.) Specific types of systematic errors in epidemiologic studies are discussed in several articles (*see* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**; **Bias, Overview**; **Confounding**; **Validity and Generalizability in Epidemiologic Studies**).

MITCHELL H. GAIL

Systematic Sampling Methods

Systematic sampling is a simple and convenient sampling technique that is widely used in practice (*see* **Sampling Frames**). While it is attractive in its simplicity and ease of use, care should be taken in the application of this sampling technique. Depending on the structure of the study population, systematic sampling can be at best the optimal selection method, or at worst a method that provides no more information beyond the taking of a single observation at random.

Reviews of the literature of systematic sampling, as well as some applications of systematic sampling, may be found in Bellhouse [2], Buckland [4], Iachan [13], and Murthy & Rao [17]. Monographs with significant technical detail on systematic sampling have been written by Cochran [7], Levy & Lemeshow [15], Murthy [16], and Sukhatme et al. [20].

To take a systematic sample the population must first be sequentially ordered in some way. This ordering may have structure to it, such as houses on a street or an alphabetic list of names from a directory. Alternatively, there may be no structure, such as a frame listed in random order of the sampling units. A systematic sample is chosen by selecting an initial unit using a random start in the ordered population and then by selecting subsequent units at equal intervals from the random start. Since only one unit has been randomly selected, there is no estimate of variation that is **unbiased** with respect to the sampling design.

More formally, the study population may be defined over a fixed set of units labeled $u = 1, \ldots, N$, with measurement y_u attached to the unit labeled u. A systematic sample of size n is obtained by drawing a random integer r from $1, \ldots, N$, and sampling the set of units given by $s = \{r, r + k, r + 2k, \ldots, r + (n - 1)k\}$. The term k is called the sampling interval. For any j for which $r + jk > N$, then the unit selected is $r + jk - N$. The selection method reduces to the usual notion of systematic sampling when N/n is an integer and when k is chosen as N/n. This type of systematic sampling is equivalent to **cluster sampling** with the selection of a single cluster. When N/n is not an integer, the selection method is known as circular systematic sampling. The typical choices for k in this case are the greatest integer in N/n or the integer nearest N/n. Most classical and modern textbooks on sampling discuss systematic sampling from a finite population.

Applications of systematic sampling are wide ranging and include any population which can be put into a list. Systematic sampling is also useful when a **sampling frame** or list of the ultimate sampling units is not available. Kalton [14] has given several examples of this. His examples pertain to the use of systematic sampling for the sampling of human populations when they are mobile. These examples range from exit polls on election day in which every kth person leaving a polling station is interviewed to road traffic surveys in which every kth vehicle in a particular lane of traffic is sampled.

For any finite population, **means** and **variances** of sample statistics are calculated through the first- and second-order inclusion probabilities. These are respectively π_u, the probability that unit u is included in the sample, and π_{uv}, the probability the units u and v are both included in the sample. For systematic sampling, $\pi_u = n/N$. For the special case in which N/n is an integer, the joint inclusion probability $\pi_{uv} = n/N$ when $|u - v|$ is divisible by the sampling interval k, and is 0 otherwise.

Alternately, the population may be defined on a continuum with the measurement y_u attached to u, where u is in the interval $[0, N]$. The initial unit r is chosen from the interval $[0, k]$ and, as before, the set of sampled units is given by $s = \{r, r + k, r + 2k, \ldots, r + (n - 1)k\}$. It is not necessary that k be an integer in this case. Populations on a continuum are encountered in the stereologic examination of tissues. Sampling problems here are problems in geometric probability. Detailed descriptions of the theory and application of systematic sampling to **stereology** are given in Gundersen & Jensen [9] and Cruz-Orive [8]. An application of systematic sampling in this area is given in Pache et al. [18]. They describe techniques for the estimation of lung volume and the volume of other structures inside the lung, such as bronchi and arteries. Volume is estimated from serial sections of lung specimens using a CT scan. The serial sections are obtained through systematic sampling.

In descriptive surveys, the parameter that is usually of interest is the population mean $\overline{Y}$, which may be expressed as $\sum_{u=1}^{N} y_u/N$ for the finite population, or as $\int_0^N y_u \, du/N$ for the population defined on a

continuum. For either situation the estimator of $\overline{Y}$ is given by the sample mean for the systematic sample chosen by the random start r, $\overline{y}_r = \sum_{j=0}^{n-1} y_{r+jk}/n$. The variance of this estimator, denoted by V_{sys}, is given by $\sum_{r=1}^{N}(\overline{y}_r - \overline{Y})^2/N$ in the finite population framework. When N/n is an integer and when $k = N/n$, then V_{sys} reduces to the more common expression of the variance $\sum_{r=1}^{k}(\overline{y}_r - \overline{Y})^2/k$. The equivalent expression in the population defined on a continuum is $\int_0^k (\overline{y}_r - \overline{Y})^2 \, \mathrm{d}r/k$.

Often, there is structure to a population which can be modeled mathematically. In most situations with structure present, systematic sampling can be used to advantage in the sense of reducing the variance of the estimate of $\overline{Y}$ when compared with estimates obtained from other sampling designs. The assumed structure is usually modeled by assuming that the measurement y_u is an **random variable** and by making assumptions about the first- or second-order **moments** of this random variable. There are now two sources of random variation to consider, one due to the sample selection procedure and the other to assuming that y_u is a random variable. One generally accepted measure of the variation of an estimator in this case is to average the variance obtained under the sampling design over the distribution of the random variable assumed on the measurement y_u. For systematic sampling this may be denoted by $\mathrm{E}_m V_{\text{sys}}$, where E_m is the **expectation** with respect to the model assumption on y_u. This measure was first introduced by Cochran [6].

There are several examples of populations with models assumed on the first moment:

1. $E_m(y_u) = \mu$;
2. $E_m(y_u) = \alpha + \beta u$;
3. $E_m(y_u) = \alpha + \beta u + \gamma u^2$; and
4. $E_m(y_u) = \sum_{v=-\infty}^{\infty} c_v \exp(2\pi \mathrm{i} v u/p)$ where $\pi = 3.14159\ldots$, $\mathrm{i} = (-1)^{1/2}$, c_v are Fourier coefficients and p is a constant.

For each of these examples it is assumed that the ys are uncorrelated and that the second central moment is σ^2, a constant. Models 1, 2, and 3 describe constant, linear and quadratic trends in the measurement with respect to the label number. Model 4 is a model of periodic variation with period p. Under model 1, systematic sampling has the same efficiency as **simple random sampling** without replacement (*see* **Sampling With and Without Replacement**) or any other design with constant probability of inclusion for each of the sample units. Under models 2 and 3, systematic sampling is more efficient than simple random sampling but is less efficient than **stratification** (*see* **Stratified Sampling**). The efficiency of systematic sampling can be markedly improved under models 2 and 3 by changing the estimator. In particular, the new estimator is the sample mean plus a weight times the difference between the measurements with the largest and smallest labels. The resulting estimator, known as Yate's end corrections estimator, eliminates both α and β as components of the variance (*see* **Variance Components**). In the special case of $N = nk$, the end corrections estimator is given by $\overline{y}_r + (2r - k - 1)(y_r - y_{r+(n-1)k})/[2(n - 1)k]$. The estimator under circular systematic sampling is described in Cochran [7]. Under the model for periodic variation, given as model 4, the best and worst of situations can occur. When the sampling interval k coincides with the period p, then systematic sampling is equivalent to taking a simple random sample of one unit. When the sampling interval is the half period, or $k = p/2$, then the only component of variance is σ^2.

The most common model assumption for the second central moment is **autocorrelation**. The first moment is assumed to be as in model 1, with the second-order central moments given by $\mathrm{E}_m(y_{u+t} - \mu)(y_u - \mu) = \sigma^2 \rho(t)$ with $\rho(0) = 1$, where $\rho(t)$ is the autocorrelation function at lag t. When $\rho(t)$ is concave and decreasing in t, systematic sampling is more efficient than any other sampling scheme which has constant probability of inclusion for each of the sample units. Concave decreasing autocorrelation functions are obtained when the ys are modeled by an autoregressive process of any finite order such that the roots of the **characteristic** equations are real.

A simplifying assumption that is often made for systematic sampling is that the units are in random order. Then the sample that is obtained is treated as if it were from a simple random sample. More formally, the random ordering corresponds to the assumption that the set of finite population measurements $y_1, \ldots, y_N$ are a random permutation of a set of fixed numbers $z_1, \ldots, z_N$. Under this random permutation model, expressed in a linear model framework originally by Rao [19] (*see* **General Linear Model**), systematic sampling is equivalent in efficiency to any sampling design with constant inclusion probability.

As noted previously, since a single random start was used to obtain a systematic sample, there is no unbiased estimator of the variance of the sample mean. Here unbiasedness is meant to be with respect to the sampling design or method of sampling. In view of this, there are at least three ways to obtain an estimate of variability based on the single systematic sample:

1. Make a simplifying assumption. In particular assume random ordering of the units so that the sample obtained is equivalent to a simple random sample.
2. Assume a trend in the data and base the variance estimate on squared differences between successive observations in the sample.
3. Use the data to estimate the model parameters in $\mathrm{E}_m V_{\mathrm{sys}}$, and use this as the estimate of V_{sys}.

In the first case, the unbiased variance estimator from simple random sampling $(N-n)s^2/(Nn)$ may be used, where $s^2 = \sum_{j=0}^{n-1}(y_{r+jk} - \overline{y}_r)^2/(n-1)$ is the variance within the obtained systematic sample. If the random ordering assumption is correct then this estimator is unbiased for V_{sys}, where un-biasedness is observed with respect to the random permutation model. This estimator generally works well when there is no trend in the population or when the trend is weak. When a linear trend is present as in model 2, the variance is a function of β^2, where β is the slope in the linear trend function. The variance estimator suggested in the second case will also be a function of β^2. The estimator is then of the same form as in the first case, with s^2 replaced by $\sum_{j=0}^{n-2}(y_{r+jk+k} - y_{r+jk})^2/2(n-1)$. This variance estimator, which tends to work well under various types of trends, and other similar variance estimators, are described in Wolter [22]. If the third strategy is followed, it is necessary that the model assumption be correct. The variance estimator will work well under the correct model assumption but may not be **robust** to departures from the model.

If a design unbiased estimate of variance is of primary concern, then a change to the sampling design must be considered. One way to achieve an unbiased estimator while retaining some of the convenience of systematic sampling is take more than one random start, or repeated systematic samples. When $N = nk$ this is equivalent to cluster sampling with more than one cluster selected. Under autocorrelation and trend models, this method is less efficient than systematic sampling of equivalent sample size with a single random start. What is gained in unbiased variance estimation is lost in efficiency and in some convenience. If the units are in random order, then they are equally efficient so that convenience is the only issue. A second approach is to augment the single random start systematic sample with a small simple random sample. This retains most of the convenience of systematic sampling. However, the unbiased estimate of variance can be negative.

In many large-scale surveys with stratification and two or more stages of sampling (*see* **Multistage Sampling**) the primary sampling units, or the units at the first stage of sampling, are often chosen by **sampling with probability proportionate to the size** of the primary. There are several sampling methods which yield the inclusion probability for a unit proportional to some size variable. These are reviewed in Brewer & Hanif [3]. Among these methods, probability proportional to size (pps) systematic sampling is simple and convenient to execute. An application of pps systematic sampling to a large-scale survey is given in Chambless [5]. He describes a survey taken in and near Augsberg, Germany, as part of the World Health Organization's MONICA program (Monitoring Trends in Cardiovascular Diseases). The purpose of this international program was to study the relationship between risk factor levels, estimated by sample surveys, and the **incidence rates** for coronary heart disease, estimated from population registers. The design reported by Chambless for sampling outside of Augsberg was a pps systematic sample of administrative areas, where the size variable was population size, and then a stratified sample of individuals within the chosen areas. The stratification was done on age and sex. This design yields an approximate equal probability of selection for each individual in the population.

As in the equal probability case, the population units can be or can be put in random order. Randomized pps systematic sampling, as well as the nonrandomized version, has been used extensively for primary sample unit selection in the 1970s and 1980s for Canada's national monthly survey of employment and unemployment, the Canadian Labour Force Survey. Probability proportional to size systematic sampling was used in favor of other pps sampling methods because of the flexibility of its use in

sampling on successive occasions and because of the ease with which the sample can be expanded.

To describe the pps systematic sampling scheme, denote the size variable for unit u by x_u. Any pps sampling design will yield $\pi_u \propto x_u$ or $\pi_u = nx_u/X$, where $X = \sum_{u=1}^{N} x_u$ is the population total of the size variables. A pps systematic sample is chosen in the following way. Form the cumulative totals $T_u = \sum_{i=1}^{u} x_i$ for $u = 1, \ldots, N$. Draw a random integer r from $1, \ldots, X$, and obtain the set of integers given by $s = \{r, r + k, r + 2k, \ldots, r + (n - 1)k\}$. The sampling interval k is the integer nearest X/n. If for any $j, r + jk > X$, then the integer selected is $r + jk - X$. Unit u is selected for the sample if $T_{u-1} < r + jk \leq T_u$. The selection method reduces to systematic sampling with equal probability when $x_u = 1$ for all u. When X/n is an integer then, similar to the equal probability case, the selection procedure is equivalent to drawing the random start r from $1, \ldots, k$. Randomized pps systematic sampling is obtained when the N population units are in placed in random order prior to sample selection. As in the equal probability sampling case if the simplifying assumption of random ordering of the population units is made, or if random ordering is imposed, then relatively simple and valid variance estimates can be obtained.

The estimate of $\overline{Y}$ is given by the **Horvitz–Thompson estimator**, $\sum_{u \in s} y_u/(N\pi_u)$. For large population sizes and small sampling fractions (less than 5%), the variance of the estimate may be approximated by applying the appropriate formula for pps sampling with replacement. For other situations, the Yates–Grundy form of the estimate of variance may be used. This involves knowledge of the joint inclusion probability π_{uv}. For randomized pps systematic sampling Hartley & Rao [10] have provided, to order N^{-4}, an approximation to this inclusion probability. This approximation can be accurate for populations as small as $N = 10$. Hidiroglou & Gray [11] have provided Fortran code for the exact calculation of π_{uv} which is useful for small N.

Systematic sampling techniques have also been developed for sampling units with a spatial or two-dimensional ordering. Bellhouse [2] and Thompson [21] provide reviews of the theory of spatial sampling with some applications. Early applications of these techniques were in geography; a review is given in Holmes [12]. More recent applications have been in ecology and in other areas: for example, Bellhouse [1] applied two-dimensional systematic sampling techniques to the excavation of middens at Iroquoian Indian archeological sites.

References

[1] Bellhouse, D.R. (1980). Sampling studies in archaeology, *Archaeometry* **22**, 123–132.

[2] Bellhouse, D.R. (1988). Systematic sampling, in *Sampling, Handbook of Statistics*, Vol. 6, P.R. Krishnaiah & C.R. Rao, ed. Elsevier, Amsterdam, pp. 125–145.

[3] Brewer, K.R.W. & Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer-Verlag, New York.

[4] Buckland, W.R. (1951). A review of the literature of systematic sampling, *Journal of the Royal Statistical Society, Series B* **13**, 208–215.

[5] Chambless, L.E. (1988). On the use of two-stage cluster samples in epidemiological population studies, *Biometrical Journal* **30**, 313–328.

[6] Cochran, W.G. (1946). Relative accuracy of systematic and random samples for a certain class of population, *Annals of Mathematical Statistics* **17**, 164–177.

[7] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[8] Cruz-Orive, L.M. (1989). On the precision of systematic sampling: a review of Matheron's transitive methods, *Journal of Microscopy* **153**, 315–333.

[9] Gundersen, H.J.G. & Jensen, E.D. (1987). The efficiency of systematic sampling in stereology and its prediction, *Journal of Microscopy* **147**, 229–263.

[10] Hartley, H.O. & Rao, J.N.K. (1962). Systematic sampling with unequal probabilities and without replacement, *Annals of Mathematical Statistics* **33**, 350–374.

[11] Hidiroglou, M.A. & Gray, G.B. (1980). Construction of joint probability of selection for systematic pps sampling, *Applied Statistics* **29**, 107–112.

[12] Holmes, J.H. (1970). The theory of plane sampling and its application to geographic research, *Economic Geography* **46**, 379–392.

[13] Iachan, R. (1982). Systematic sampling: a critical review, *International Statistical Review* **50**, 293–303.

[14] Kalton, G. (1991). Sampling flows of mobile human populations, *Survey Methodology* **17**, 183–194.

[15] Levy, P.S. & Lemeshow, S. (1991). *Sampling of Populations: Methods and Applications*. Wiley, New York.

[16] Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

[17] Murthy, M.N. & Rao, T.J. (1988). Systematic sampling, in *Sampling, Handbook of Statistics*, Vol. 6, P.R. Krishnaiah & C.R. Rao, eds. Elsevier, Amsterdam, pp. 147–185.

[18] Pache, J.-C., Roberts, N., Vock, P., Zimmermann, A. & Cruz-Orive, L.M. (1993). Vertical LM sectioning and parallel CT scanning designs for stereology: applications to the human lung, *Journal of Microscopy* **170**, 9–24.

[19] Rao, J.N.K. (1978). On the foundations of survey sampling, in *A Survey of Statistical Design and Linear Models*, J.N. Srivastava, ed. North-Holland, Amsterdam, pp. 489–505.

[20] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. & Ashok, C. (1984). *Sampling Theory of Surveys with Applications*, 3rd Ed. Iowa State University Press, Ames.

[21] Thompson, S.K. (1992). *Sampling*. Wiley, New York.

[22] Wolter, K.M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

(*See also* **Cluster Sampling**; **Superpopulation Models in Survey Sampling**)

D.R. BELLHOUSE

Target Population

The concept of a *target population* is an informal one, sometimes defined as "the population about which information is wanted" [1] or the "totality of elements which are under discussion and about which information is desired" [4]. Often, the word "population" refers to this concept; see, for example, Kendall & Stuart [3] or Freedman et al. [2]. The word "target" emphasizes, however, that this population is not necessarily the same as the one that we end up sampling. The latter population is sometimes called the *sampled population* [1, 4] or (in epidemiology) the *source population* [6]. Ideally, in **descriptive epidemiologic** studies, the two populations would be identical, but practical concerns usually lead to large discrepancies. For example, when a poll of the entire US population is desired but, for cost reasons, only a telephone survey of four cities is done, the sample population contains only persons who have a telephone and live in those cities, and is much smaller than the target (US) population.

In studies of causal effects (*see* **Causation**) it may be helpful or even essential for the sampled population to extend beyond or even exclude the target population. Consider a target population comprising five persons who were exposed to high asbestos levels during a job assignment, one of whom later developed mesothelioma (a very rare form of cancer). The question of whether this high rate of the disease was caused by the asbestos could not be approached by sampling the target. Only by comparison to a much larger reference experience (namely, the extensive prior data on the rate of mesothelioma in workers exposed and not exposed to asbestos) can we make any meaningful **inference** about the target. In settings such as this example, in which inferences about structural relations in the target are inferred from observations on other populations, it has been proposed to refer to the latter as *evidentiary populations* [5]. Such evidentiary populations are typically larger and sometimes more accurately measured than the target, although they may not be comparable to the target population in all important respects (*see* **Confounding**).

Issues of inferences from the sampled population to the target are sometimes classified as problems of generalizability or external validity. These issues are distinct from the issues that arise in making inferences about the sampled population from a sample (*see* **Validity and Generalizability in Epidemiologic Studies**).

References

[1] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. Wiley, New York.

[2] Freedman, D., Pisani, R. & Purves, R. (1978). *Statistics*. Norton, New York.

[3] Kendall, M. & Stuart, A. (1977). *The Advanced Theory of Statistics*, 4th Ed., Vol. 1: *Distribution Theory*. Macmillan, New York.

[4] Mood, A.M., Graybill, F.A. & Boes, D.C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

[5] Poole, C. (1987). Evidence, targets, and proportional attribution (letter), *American Journal of Epidemiology* **125**, 1095–1096.

[6] Rothman, K.J. & Greenland, S. (1997). *Modern Epidemiology*, 2nd Ed. Lippincott, Philadelphia.

SANDER GREENLAND

Teaching Medical Statistics to Statisticians

A degree in statistics does not usually prepare a graduate adequately for a career in biostatistics. Further training is needed in those aspects of statistics applicable to the health sciences, both in postgraduate degrees, and in continuing education courses for the statistician already working in a particular field (e.g. pharmaceuticals, public health, clinical medicine). In this account of teaching biostatistics, we do not intend to provide outline course contents. Rather, we bring forward a number of considerations that could help in course planning. The course planner then necessarily takes into account the assumed background knowledge of the student group, the time, physical and personnel resources available, and reasonable expectations of what the student will have acquired by the end of the course. The wide-ranging scope of this Encyclopedia makes it clear that no statistician can expect to be expert in every aspect of biostatistics. Those who plan and teach such courses must of necessity be selective.

Collaboration

Etymologically, the word biostatistics is a hybrid, its mixed parentage combining bio- (life; hence, care for health) with the statistician's art and sciences. Accordingly, instruction in biostatistics could well begin – and continue throughout – with an emphasis on the necessity of collaboration between the statistician and the health professional. Thus, it is important to recognize that even concepts which appear trivial (in the mathematical sense) to the statistician may be not only unfamiliar to, but often alien to the thought processes of, the health scientist, causing great difficulty in communication. Both teacher and pupil will benefit from an emphasis on the advantages of simplicity – in language, avoiding or explaining any jargon; and in the design, analysis and interpretation of studies. The student can be helped to develop consultancy skills by attending and contributing to consultations with health professionals who come to the teacher with real problems (see, for example, [2]) (*see* **Statistical Consulting**). Difficulties can also arise when an inappropriate statistical method has been used in a key publication in the client's area, which then spawns imitations. A common example is the use of repeated significance tests (*see* **Sequential Analysis**) at a series of time points in a **longitudinal study**, when one or more summary measures could well be more informative.

Other ways of improving understanding between the professions include organizing a "Meet the Clients" course, in which a number of established clinicians from various medical specialties, other carers, hospital laboratory staff, health care administrators, epidemiologists, pharmaceutical scientists, etc., present their own accounts of how biostatistics has impinged on their work. Discussions following such presentations can be both lively and rewarding to all concerned. Interdisciplinary understanding can also be enhanced by study of carefully selected publications from the health science literature, concentrating on the statistical methods that have been used or misused in the study design and analysis, and on the validity of the conclusions and the summary. The aims of these exercises are not only to develop constructive critical skills, but also to help to prepare the student for joint authorship with future collaborators. As medical teaching and practice is increasingly turning towards such critical appraisal of the literature to assist clinical decision making, statisticians should be aware of the important contribution of their discipline (see, for example, [1]) (*see* **Teaching Statistics to Medical Students**; **Teaching Statistics to Physicians**).

Examples from the literature can also be used to illustrate common pitfalls in the use of statistics – comparing the noncomparable (inappropriate choice of control subjects); not properly defining the denominator, or even ignoring it altogether, in the calculation of rates (*see* **Denominator Difficulties**); failing to adjust for **confounders** such as age and sex, by **standardization methods** or otherwise. Another pitfall common in health research is failure to recognize the long-established phenomenon of **regression to the mean**, which many find counterintuitive at first sight. If a patient is identified because her blood pressure reading is high, it seems strange that a subsequent reading is expected to be lower, even if no treatment has intervened. The same phenomenon will often explain why a treatment appears to produce better results in subjects with more severe disease, even when it is actually uniformly effective for all degrees of severity. These two apparent anomalies can be inspirational in

attracting bright young statisticians to this field of application – "there's more to biostatistics than meets the eye!".

How is Biostatistics Different?

What differentiates biostatistics from other applications of statistical methods? It is essentially the human context. We note three aspects: ethical issues, the failure of patients to behave as ideal "experimental units" and, in common with other biological rather than mechanical subjects, more difficulty in establishing **causation**.

Ethics

Human ethical implications arise both in research and in the implementation of health programs based on the interpretation of statistics. It is sometimes argued that ethical considerations are solely the responsibility of the health professional, and statisticians should be dispassionate and value-free in their contributions. In some instances the statistician is, however, very well placed to stand aside from the enthusiasms of a health researcher in the laboratory, the clinic or in the organization of health systems, and to see more clearly the ethical implications. These ethical issues are often important in: the design of a study (choice of subjects, sample size, likely effects of treatments, the concept of **randomization** of humans), the conduct of the study (in addition to the normal clinician–patient relationship, in most studies every patient is a volunteer and deserves to be treated as such), and the implications for the health of others of the results of the study, whether to individuals or to communities as part of a more general health program (*see* **Ethics of Randomized Trials**; **Medical Ethics and Statistics**).

Human Fallibility

Patients do not always behave as perfect experimental units. Some do not take their treatment as specified, often for very good reasons. This may occur during an experimental trial or subsequently after a successful trial when the treatment is made more generally available. Doctors may not be consistent, or even logical, when defining or making a diagnosis. Sources of variability, such as between or within observers or analytical methods, can often be identified, measured, and taken into account. Human biological variability may affect the precision of a diagnostic test (*see* **Diagnostic Tests, Evaluation of**). The dramatic effect of targeting **screening** to an at-risk population in which the **prevalence** of the disease is substantially greater than in the general community can be demonstrated as an application of **Bayes' Theorem**. The resulting increase in the **positive predictive value** of a screening test with a given **sensitivity** and **specificity** has important practical and financial implications. Cultural or educational background may affect a patient's response to a question. A biostatistician must be aware of these problems, and use personal or shared experience to take them into account in modeling the collection, analysis, and interpretation of the data.

Causation

Students will be aware that **association** does not necessarily imply causation, but the health professional expects the biostatistician to go further than that bald, and not particularly helpful, conclusion of an association. A number of accounts have been given on the topic; perhaps the most accessible is that given by **Bradford Hill**, one of the founders of twentieth century biostatistics. Drawing on, amongst other things, his literally vital contributions to establishing cigarette smoking as a cause of lung cancer (*see* **Smoking and Health**), he suggested a wide-ranging approach to establishing causation from **observational studies** on chronic diseases. Considerations include the strength of the association, consistency between studies in varied circumstances, specificity of the association, temporal consistency, **dose–response**, biological plausibility, coherence, experimental evidence, and analogy (*see* **Hill's Criteria for Causality**). The length of this list makes it clear that the biostatistician has responsibilities other than the purely mathematical. Health programs will seldom be mounted solely on the basis of statistical associations. A credible presentation of a causative link between a modifiable risk factor and a disease is necessary, and the statistician has important contributions to make to that presentation. Training should prepare the statistician for that role.

Study Types

At this stage, the epidemiologic approach to causality and the strength of evidence from various types of study may be of value in the course. In ascending order of strength of evidence for a causative link between a putative risk factor and a disease are: the **ecologic study**, the **cross-sectional study**, the **case–control study**, the **cohort study**, the randomized controlled trial (RCT) (*see* **Clinical Trials, Overview**), and the **meta-analysis of RCTs**. While the usual statistical methods of assessing and comparing **means** or proportions often conclude studies of all these types, other statistical questions arise in connection with each. Course designers may choose to structure courses on the study types, expanding on the statistical methods arising in each, or, more traditionally, to structure the course on statistical methods, giving the related study types as illustrations of their application. Whichever approach is used, examples of life-saving interventions that have been demonstrated by the various types of study design will be helpful.

We illustrate the former approach because it is a little unusual. It could be a more stimulating approach for the students, who will already be familiar with at least the theoretical aspects of **logistic regression** and perhaps **survival analysis**.

Observational Studies

While *ecologic studies* may be cited as evidence of coherence (countries or districts with higher per capita sales of cigarettes have higher death rates from disease X), they are more useful in a teaching context to illustrate the risks of **confounding** and the **ecologic fallacy**.

The *cross-sectional study* requires much attention to **questionnaire design**, use of focus groups, pilot studies, and methods of achieving adequate response rates, as well as the intricacies of sampling procedures and analyses which take account of the sampling details. No opportunity should be lost to draw attention to routinely collected statistics, national or regional (*see* **Vital Statistics, Overview**). It is all too easy for a study to produce a poor imitation on a small scale of a routine analysis, ignoring the accumulated experience of the official statisticians in data collection, coding, validation, analysis, interpretation, and limitations (*see* **Bias in Observational Studies**).

In a *case–control study* the main epidemiologic difficulty is to define the appropriate populations from which the **controls** should be selected. For example, selecting controls from hospital inpatients can lead to considerable bias. Other biases may arise in the diagnosis and coding of the disease outcome, or in the measurement of exposure to possible risk factors, which relies on recall or historical records collected for other purposes. The statistician needs to be trained to identify such biases before undertaking any analysis (*see* **Bias in Case–Control Studies**).

Case–control studies provide an ideal opportunity to introduce the concept of **odds ratios** and their interpretation as an approximation to **relative risk** for a rare disease. An example in which there is heterogeneity of the odds ratio across strata of another variable can be used to illustrate statistical **interaction**, or "**effect modification**" as it is more helpfully termed by epidemiologists. Furthermore, **logistic regression** is a way of obtaining an estimate of the approximate **relative risk** of the disease for those exposed to a particular risk factor, after adjusting for **confounding** variables. For the situation where the controls have been individually matched to the cases, **conditional logistic regression** can be demonstrated.

A *cohort study* may be affected by many of the same problems as case–control studies. In addition, attention should be drawn to the healthy worker effect in occupational cohorts (*see* **Bias in Cohort Studies**).

Randomized Trials

The *randomized controlled trial* has a central role in the evaluation of interventions in both clinical medicine and public health. Biostatisticians should have a thorough understanding of such trials, whether or not they are going to be directly involved in conducting trials. Historically, statisticians have contributed greatly to the development of valid trials, and it is good for a professional to know something of past achievements. Perhaps more importantly, the RCT introduces many apparently nonstatistical ideas that are essential to producing valid statistical analyses and interpretations. For example, **randomization** of subjects to treatments (*see* **Randomized Treatment Assignment**) raises important ethical questions including patient consent and what is meant by "informed" consent. Before randomization, however, a strict protocol (*see* **Clinical Trials Protocols**) is required to specify the population to be studied and to

whom the results will be applicable (*see* **Target Population**): diagnostic details, **eligibility and exclusion criteria**.

Careful **sample size** calculations are also important for both ethical and economic reasons. Simple randomization can be improved by techniques such as **blocking** and **stratification**, which should be taken into account in the analysis to increase **power**. If practical considerations lead to subjects being randomized in clusters (e.g. families, school classes, villages, etc.), failing to incorporate this in the analysis can produce seriously misleading conclusions (*see* **Group-randomization Designs**). Methods of **blinding** help to avoid the patient and the professional assessors of health status knowing which treatment each patient is receiving – human fallibility could introduce **biases**. Fallibility may also cause the patient not to receive the complete course of treatment as specified: analysis should almost always follow the "**intention-to-treat**" principle, which is sometimes not intuitive, particularly to the surgeon whose allocated patient dies before reaching the operating table.

Because RCTs often involve follow-up of subjects over time, **survival analysis** is often the most appropriate technique for evaluating the results. Many good medical data sets are available in the literature for demonstrating the calculation of an actuarial **life table** for grouped data or a **Kaplan–Meier** life table for individual survival data, discussing the underlying assumptions of these, illustrating the use of the **logrank test** to compare treatments, **Cox regression** to adjust for **confounders**, assessment of the **proportional hazards** assumption, and ways of dealing with violations of that assumption.

Meta-analysis is a powerful statistical method for combining the results of several RCTs. A single trial demonstrating the efficacy of a therapy is unlikely to lead to its adoption worldwide. Meta-analysis raises issues of heterogeneity and **fixed effects** vs. **random effects** models, which merit attention in this and other applications.

General Considerations

A mathematical statistics course may well omit methods relevant to biostatistics, from the very simple to the highly sophisticated. Students may be unfamiliar with, for example, the simplified formula for calculating the test statistic for a **2 × 2 table**, or with **McNemar's test** for paired proportions, which is merely an application of the **binomial distribution** with probability 1/2. There are also many practical uses of the **Poisson distribution**, ranging from the simple calculation of tail probabilities in the study of disease outbreaks, to modeling disease frequencies using **Poisson regression**. If students are given real data sets to analyze, they can learn some of the pragmatic aspects of biostatistical modeling, such as how the choice of model variables depends not only on statistical considerations, but also on biological plausibility, face validity, the "cost" of measuring a variable, confounding, and whether a risk factor is modifiable. The extent to which model assumptions may be violated can also be discussed. Students will benefit from becoming familiar with at least one of the major statistical computing packages (*see* **Software, Biostatistical**). In matching the assessment of a biostatistics course to its objectives, one needs to consider whether computers can be used, either in an open-book examination in a classroom equipped with computers, or in a take-home exam for which students have computer access.

It is clear from the above that it is appropriate to use a variety of teaching methods – lectures, seminars, computing practicals, tutorials, small-group exercises – to interest, stimulate, and extend the students. As they are mature students, the principles of adult teaching and learning should be employed. In particular, lectures are best used sparingly to motivate and expound concepts, rather than to give detail that can be better absorbed by reading, either before the lecture or afterwards.

Whatever is covered in a biostatistics course, it cannot be exhaustive, so it is important to teach and encourage students to continue their own education. They should be made aware of the resources available on the **Internet** for communicating with colleagues, finding relevant articles on unfamiliar topics, and keeping abreast of recent developments in the literature. The importance of maintaining contact by attending local meetings and conferences could also be emphasized.

We have deliberately avoided being prescriptive and giving an exhaustive list of topics to be included. Rather, we have tried to give the flavor of the process leading to decisions about course content. Those decisions will inevitably, and rightly, reflect the experience and research interests of the teachers.

References

[1] Elwood, J.M. (1988). *Causal Relationships in Medicine: A Practical System for Critical Appraisal*. Oxford University Press, Oxford.

[2] Hand, D.J. & Everitt, B.S., eds (1987). *The Statistical Consultant in Action*. Cambridge University Press, Cambridge.

DAVID NEWELL & JUDY SIMPSON

Teaching Statistics to Medical Students

It is difficult to pinpoint just when statistics was introduced into the medical curriculum, what medical school was the first to include lectures on statistics, and who was the first statistics lecturer to medical students. Clearly, the incorporation of statistics into the medical curriculum parallels publication of the early textbooks in medical statistics, with that by **Austin Bradford Hill** in 1937 being the first [13]. In the US, Colton [4] presented anecdotal material to indicate that at Johns Hopkins University, where the first US school of public health was established, there was clearly, by 1948, a course in biostatistics for medical students taught by **Margaret ("Maggie") Merrell**. Obviously, the initiation of the course dates several years earlier. Nevertheless, it is of interest to note that, in 1948, medical students at Johns Hopkins, a group that was almost exclusively male, received the required instruction in biostatistics from a female instructor who, moreover, was a professionally trained statistician and a nonphysician.

Perhaps the earliest published material on teaching statistics to medical students is an editorial written by Bradford Hill in 1947 for the *British Medical Journal* with the intriguing title, "Statistics in the Medical Curriculum?" [14]. Note the question mark in the title for which Hill comments:

> ... I should replace my querying title, "Statistics in the Medical Curriculum?", by "*What* Statistics in the Medical Curriculum?" (though I am well aware that some clinical teachers will prefer to read it as "*What!* Statistics in the Medical Curriculum?").

Hill enunciates what has become a common theme underlying the rationale for teaching statistics to medical students, namely,

> ... the medical worker, and his readers, must be at least on speaking terms with the elements of statistical reasoning and methods of analysis, and thus be able themselves to weigh numerical evidence justly in the balance.

His conclusion, which portends much discussion and debate that has appeared in the literature in the subsequent 50 years or so, is:

> Arithmetic guided by logic has been given as a fairly accurate definition of simple statistical methods, and it is that kind of teaching that would, it is my belief, be of real benefit to the medical student. It should introduce him not only to ... general viewpoints ... to be borne in mind in considering statistical evidence, but also teach him the appropriate methods of handling and presenting data, and familiarize him with the statistical concepts of variability, the ideas lying behind elementary tests of significance, simple means of measuring and interpreting associations, and so on ....

Thus, in addition to having written the first text in medical statistics, Bradford Hill is apparently among the first to state in print the rationale for teaching statistics to medical students as well as to propose the general content for such instruction.

In the US, an early publication on teaching statistics to medical students is the report in 1953 of a Committee of the Statistics Section of the **American Public Health Association** charged with consideration of the training in statistics needed by medical students, as well as that needed by public health students, including those who intended to specialize in medical and public health statistics [6]. With regard to training of medical students, the report included the findings of a questionnaire survey in 1952 sent to 90 medical schools in the US and Canada for which 82 replies (91%) were received. Of those replying, 82% indicated that instruction in biostatistics was a component of their undergraduate medical curriculum, with 46% stating that the course given was primarily one designated as biostatistics. (In the remaining schools that indicated they taught the subject, biostatistics was taught as a part of another course such as preventive medicine or scattered among several other courses such as pharmacology and physiology.) The predominant pattern was to teach biostatistics in the first two years as part of the basic science curriculum and to require the course of all medical students. Thus, by the mid-1950s, biostatistics instruction to medical students was well ensconced in the US and Canada and a clear pattern of the format and nature of such instruction was evident.

It is of interest to note the Committee's synthesis of areas of agreement among the Committee members, the respondents to the survey, and other medical school faculties whose opinions were solicited. The report states the "general agreement on a number of issues", as follows:

> It is desirable that all medical students receive statistical instruction. The important issues cannot

> well be discussed in a course of less than some 25–30 hours. The basic principles of the experimental method and of the statistical approach must be stressed, but there must also be an abundance of practical applications. The committee believes that the teaching should be done by a professional statistician (M.D., Ph.D. or other) well oriented with medical problems. People whose primary interest and concentration is in another field usually do not do the most efficient job of teaching statistics.

Subsequent to this report, findings of several other surveys of US and Canadian medical schools have been published – namely, surveys conducted in 1957 [16], in 1969–1970 and 1973–1974 [3], in 1986 [8], and most recently in 1993 (Looney, Grady & Steiner, personal communication). Most of the "issues of agreement" cited above in the 1952 survey prevail throughout the various surveys and remain areas of agreement today, except, perhaps, for the proposed 25–30 hours of class time for biostatistics. With the global trends of drastic reductions in formal lecture time, this amount of class time for biostatistics is most unrealistic in contemporary medical curricula in North America.

Another impetus in the US for the incorporation of biostatistics into the undergraduate medical curriculum has been the inclusion of questions in biostatistics on medical licensure examinations – namely, on the nationwide examinations administered by the US National Board of Medical Examiners that most US medical schools require their students to take. Traditionally, the bulk of the statistical questions appear on the Step 2 (formerly Part II) examination, the portion of the licensure examination usually taken at completion of the fourth year of medical school. At the time of the initial survey in 1952 of biostatistics instruction in US and Canadian medical schools, the biostatistics questions constituted 15% of the Preventive Medicine and Public Health component (one of six components) of the Part II examination. Undoubtedly, the inclusion of statistics questions on US medical licensing examinations offers some explanation for the proliferation of biostatistics instruction in US medical schools by the mid-1950s.

Parallel to the expansion of biostatistics instruction in medical schools in the 1950s and 1960s, additional textbooks in biostatistics began to appear and became the assigned texts for these courses. Those more commonly employed as course texts, in addition to Hill's pioneering text, were those by **Donald Mainland** published in 1952 [19], Huldah Bancroft published in 1957 [2], and Olive Jean Dunn published in 1964 [10]. In fact, by 1966, Hill's text appeared in its eighth edition.

In the US in 1969, a key event occurred – namely, the beginnings of the Subsection on Teaching of Statistics in the Health Sciences of the **American Statistical Association** (ASA). In early 1969 Anita Bahn sent a letter to several statisticians whom she knew and who had responsibility for teaching statistics to medical students. Her letter opened with the following paragraph:

> As a new teacher of biostatistics in a medical school, I have been faced with a number of problems in attempting to develop a meaningful learning experience. The rapidly changing horizon – new goals of medical schools and ways of doing things, core curricula, integrated National Board Examinations and other aspects – impinges on our teaching. It is challenging to keep one step ahead of the game.

She concluded by asking that those who shared her concerns meet informally that summer at the annual ASA meeting. The informal meeting was most successful, with keen interest indicated by virtually all who attended. Among the key next steps was the establishment of a Newsletter for which one of us (TC) served as the first editor. In the first Newsletter issued in Fall 1969, the rationale for this organization was stated as follows:

> Why is such a group needed and what are its objectives? Foremost is the need for communication among these educators to exchange information on how best to deal with the problem of student motivation, limited class time, integrated subjects, core curricula, changing medical school goals. There is much that we can learn from each other. One respondent stated that he never could understand why there has not been, heretofore, a mechanism for communication in such a sorely needed area.

Shortly after this meeting in 1969, the ASA was petitioned to form a Subsection (of the Section on Statistical Education) and thus this organization had its birth. Although the primary force behind its formation was statistics instruction to medical students, the target was more broadly defined as students in the Health Sciences, as reflected in the title. The Subsection's Newsletter continued and at each subsequent annual ASA meeting an invited papers session has appeared on the program where statistics instructors of students in the health sciences

could share their concerns, describe innovative course offerings and teaching techniques, and commiserate on their difficulties in reaching students as well as in preserving their courses intact under continued pressure from Curriculum Committees to reduce class time. In the 1990s the Subsection designation was removed and full Section status in ASA was established. Instruction of medical students remains as the key focus of the Section.

When the Subsection started, one of the major concerns of its members was the nature of the statistics questions on the National Board Examinations. Although statistics was indeed a required component of the examination, there was no professionally trained statistician responsible for determining the statistics questions that would appear on the examinations. On the preclinical examination (currently Step 1 and previously Part I), it was behavioral scientists who composed the statistics questions, and on the clinical examination (currently Step 2 and previously Part II) it was public health and preventive medicine specialists. Many of the questions that appeared on these examinations did not reflect adequately what the statistics instructors were teaching in their courses and, in several instances, the technical statistical basis of examination questions was incorrect. The Subsection lobbied successfully to place a professionally trained biostatistician, one of the founding members of the Subsection, on the National Board Test Committee responsible for creating and choosing the examination questions. By the 1970s, it was a biostatistician who composed and shared in the choice of statistics questions that would appear on the examination.

Another major effort in the early days of the Subsection was the formation of a Subcommittee to consider development of a core curriculum in biostatistics. The Committee did issue a report and its proposed core curriculum was published in 1975 [5].

In the UK, the embodiment of statistics into the medical curriculum proceeded at a considerably slower pace. Lowe [17] describes an inquiry he conducted in 1962 among the 27 undergraduate medical schools in the UK. He reports that there was some form of statistics instruction in 19 schools (70%), but that at most schools such instruction was voluntary (no examination was required) and in one school such instruction consisted of but a single hour's lecture. Despite Bradford Hill's editorial in 1947, it was not until the late 1960s that there was "official" recognition in the UK for inclusion of statistics in the medical curriculum, although as Lowe has indicated there were a substantial number of medical schools that did offer such instruction. This recognition occurred in 1968 with what has come to be known as the Todd Report [22]. With regard to statistics, this report by the Royal Commission on Medical Education states:

> Some knowledge of the principles of the statistical approach is now necessary so that doctors can make some judgement for themselves of the validity of the claims for medical advances made in journals and other communications. Instruction in statistics is a necessary part of the process of producing a graduate who can apply a scientific outlook to his future experience.

An interesting feature of the Todd Report is its Appendix, which reports the results of a survey of some 5000 students on their reactions to 18 specified courses in their curricula with regard to interest, usefulness, and difficulty. For statistics, which was offered in only some of the curricula among the schools surveyed, the report states that it

> . . . was considered dull (eighteenth and last in rank of interest), useless (seventeenth in the rank of helpfulness) and very difficult (first in degree of difficulty). This was true for all schools with the exception of Oxford where it was ranked of medium interest (ninth), of medium usefulness (ninth), and of great difficulty (first in degree of difficulty).

Consequently, by the 1970s statistics was well embodied into the medical curricula in the UK.

Over the past 25 years or so, the topic of teaching statistics to medical students has been of much concern to biostatisticians and other medical school faculties. In addition to the annual meetings of the Section on Teaching Statistics in the Health Sciences of ASA and their occasional published proceedings (the latest, at this writing, is [1]), there have been a number of symposia and workshops devoted to the topic. These include the following: a 1962 international symposium on Teaching of Statistics to Undergraduate Medical Students in Europe, organized by the **World Health Organization (WHO)**; a UK conference in 1971 organized by the Medical Section of the **Royal Statistical Society** and the Society for Social Medicine [15]; a 1978 Inter-regional Conference on Teaching Statistics

to Medical Undergraduates organized by WHO, the International Epidemiological Association (IEA), and the Government of Pakistan and that resulted in a publication entitled *The Successful Teaching of Statistics to Every Medical Student* [21]; and several conferences and workshops in the UK, a more recent one being that held in 1989 whose proceedings have been published in ***Statistics in Medicine*** [9]. The interested reader will find additional published commentaries and thoughts on this subject from the references in the above-cited publications. It is of interest to note as well that WHO and IEA had also sponsored publication in 1978 of a handbook entitled *Health Statistics: A Manual for Teachers of Medical Students* [18].

Where do we now stand with regard to teaching statistics to medical students, the contents of the courses taught, and the settings of these courses? At this writing, the most recent available information is the survey in 1993 of 125 US medical schools undertaken by Looney, Grady & Steiner (personal communication) and modeled after the previous US surveys. There were 100 responses (80%) among the 125 medical schools surveyed. Of those responding, 83% stated that their school offered such a course with 74% of those responding stating that it was a required course. Among 93% of schools that offered such a course, it occurred during the first 2 years (preclinical) and the median course length was 20 hours. Among responders who offered a course, 91% indicated provision of course notes to students, and 77% indicated use of a textbook. The three textbooks most frequently cited were those by Dawson-Saunders & Trapp [7], Morton, et al. [20], and Fletcher et al. [12]. Among the responders, 73% indicated that there was some integration of the biostatistics course with other subject matter in the curriculum. Among only 9% of the responders was there a sole instructor for the course who was a physician; in all other instances the instructor for the course had a Ph.D. or master's degree or was a nonphysician who shared with a physician major responsibility for the course. One trend noted in comparison with previous surveys is an increase in the proportion of biostatistics course directors who are full-time faculty members at their respective schools. This reflects the growing trend for US medical schools to recognize their need for and to recruit biostatisticians as full-time faculty members.

Surprisingly, with the computer revolution over the past decade, computers did not play an important role in the biostatistics courses. Only 27% of the respondents indicated use of computer-based tutorials and only 19% of respondents said they included instruction on the use of computers.

As with several of the previous surveys, the 1993 survey included questions on the topics covered in the various courses. Table 1 indicates responses to the topics covered, classified according to descriptive statistics and probability, inference, and epidemiology and clinical research (columns) and whether 75% or more of responses indicated coverage of the topic (top panel), 50%–74% indicated coverage (middle panel), or less than 50% indicated coverage (bottom panel). The survey investigators, in comparison with findings from the previous 1970 survey [3], note a shift in emphasis from more standard statistical topics to those considered more epidemiologic. There has likewise been a substantial increase compared with 1970 in coverage of **power** analysis. Undoubtedly, much of the shift in course content is commensurate with trends in the nature of the papers published in leading general medical journals such as *New England Journal of Medicine, Journal of the American Medical Association, Lancet*, and *British Medical Journal*, where there has been a substantial increase in publication of rather statistically sophisticated randomized **clinical trials** and epidemiologic **cohort studies** and **case–control studies**. Correspondingly, the issue of statistical power has received much more attention currently in the contemporary medical literature in research articles, letters to the editor, and editorials than it received at the times of the previous surveys.

It is of interest to compare the empirical results in Table 1 of what is taught in medical statistics courses in US medical schools with the topics designated as being covered currently in the Step 2 US Medical Licensing Examination [11]. Table 2 indicates the topics in Applied Biostatistics and Clinical Epidemiology that the 1997 examination covered. It is not surprising that there is considerable concordance between the topics listed in Tables 1 and 2.

Another item of interest in the 1993 survey is a solicitation of the respondent's perception of how the course was received by the students. This was compared with analogous data reported from Colton's survey in 1970 [3]. The 1993 survey investigators note that, compared with 1970, a considerably greater proportion of respondents perceived that the students had a very favorable or favorable opinion of

Table 1 Topics covered in biostatistics courses; 1993 survey of US and Canadian medical schools

Descriptive statistics and probability	Inference	Epidemiology and clinical research
Covered in 75% or more of the courses		
• Interpretation of tables and graphs • Frequency distribution • Descriptive statistics • Central tendency • Variability • Normal distribution • Probability	• Hypothesis testing • *P* values • Interpretation of confidence limits • Interpretation of the role of chance • *t* tests • Chi-square tests • Correlation	• Rates • Incidence and prevalence • Descriptive studies • Cross-sectional studies • Study design characteristics • Cohort studies • Case–control studies • Randomized clinical trials • Diagnostic tests
Covered in 50%–74% of the courses		
• Scales of measurement • Measurement issues	• Linear regression • Power analysis	• Adjusted rates
Covered in less than 50% of the courses		
• Construction of tables and graphs • Binomial distribution	• Analysis of variance • Multiple comparisons • Wilcoxon–Mann–Whitney test	• Stratified analysis

Table 2 Items in applied biostatistics and clinical epidemiology covered on the Step 2 US Medical Licensing Examination, 1997

1. Applications of concepts of measurement in medical practice (e.g. central tendency; variability, probability, and distribution; scales of measurement; disease frequency; case fatality, survival rate; relative risk, odds ratio, standardized morality rate; risk differences, attributable risk: sensitivity, specificity; positive and negative values; decision analysis).
2. Interpretation of the medical literature: study design (e.g. clinical trials, community intervention trials; cohort, case–control, cross-sectional case series; community surveys: subject eligibility and sampling; randomization, self-selection, systematic assignment; outcome assessment; validity; advantages and disadvantages of different designs; sample size).
3. Interpretation of the medical literature: statistical inference (e.g. hypothesis generation, hypothesis testing, and test statistics; statistical significance and type I error: statistical power and type II error; confidence intervals).

the course; correspondingly, a lesser proportion of respondents felt that the students' opinion of the course was neutral. Although one cannot exclude entirely the possibility of a shift over the past 20 years or so in the optimism among those who teach statistics to medical students, there is some indication that contemporary medical students are more favorably inclined to the subject and that they more readily perceive their need to understand these principles, regardless of whether or not they intend a career in medical research. Surely, one would like to think that the dismal survey results cited previously in the Appendix to the Todd Report in 1968 no longer apply to medical students in the late 1990s, both in the US and worldwide.

Finally, on a more personal note, one of us (TC) has had many years' experience in teaching statistics to medical students at several schools in the US and Canada. Sometimes it has been successful and other times it has been as disastrous as had been cited earlier in this article from the Todd Report [22]. Some view medical students as perhaps one of the toughest audiences one might have, particularly for a course in statistics. Colton [4] tried to articulate the difficulties in teaching statistics to medical students, at least those in North America, particularly in their preclinical years when such courses are almost always taught. We characterize the handicaps that the statistics instructor faces with medical students as their being the following: young, immature

and arrogant; highly competitive with one another; poorly motivated in *all* disciplines of public health; swamped by heavy demands of other courses; without any grasp or appreciation of the realities of clinical research; of considerably heterogeneous quantitative backgrounds; and a difficult group for whom to find good role models of practicing physicians as instructors. (We note that the latter has changed over time and there now is a more prevalent group of practicing physicians who are both appreciative of and knowledgeable in statistics.)

On the other hand, there are indeed some advantages to having medical students as one's audience. We characterize these advantages as their being the following: highly selected for academic performance; finely honed in knowing how to perform well in a course and to meet the instructors' demands; and that if the instructor reaches only a handful of students among a large class and can convey to them an appreciation for and enthusiasm with statistics, then the rewards and gratification are considerable.

References

[1] American Statistical Association (1998). *Statistical Education, Statistical Consulting, and Teaching Statistics in the Health Sciences Proceedings (PR7)*. American Statistical Association, Alexandria.

[2] Bancroft, H. (1957). *Introduction to Biostatistics* Hoeber-Harper, New York.

[3] Colton, T. (1975). An inventory of biostatistics teaching in American and Canadian medical schools, *Journal of Medical Education* **50**, 596–604.

[4] Colton, T. (1989). Discussion of "Teaching biostatistics – past, present future", in *Proceedings of the American Statistical Association: Sesquicentennial Invited Papers Session*. American Statistical Association, Alexandria, pp. 345–349.

[5] Committee of the American Statistical Association Subsection of Teachers of Statistics in the Health Sciences (1975). Report, *Clinical Pharmacology and Therapeutics* **18**, 127–131.

[6] Committee on Training of Medical Health Statisticians (1953). Training in medical and public health statistics, *American Journal of Public Health Year Book 1952–1953, Part 2* **43**, 129–134.

[7] Dawson-Saunders, B. & Trapp, R.G. (1994). *Basic and Clinical Biostatistics*, 2nd Ed. Appleton & Lange, Norwalk.

[8] Dawson-Saunders, B., Azen, S., Greenberg, R.S. & Reed, A.H. (1987). The instruction of biostatistics in medical schools, *American Statistician* **41**, 263–266.

[9] Day, S.J., Hutton, J.L. & Gardner, M.J. (1990). Workshop on teaching statistics to medical undergraduates, Bristol, UK, September (1989), *Statistics in Medicine* **9**, 1011–1078.

[10] Dunn, O.J. (1964). *Basic Statistics: A Primer for Biomedical Sciences*. Wiley, New York.

[11] Federation of State Medical Boards of the US, Inc. & National Board of Medical Examiners (1997). *Step 2: General Instructions, Content Description and Sample Items*. National Board of Medical Examiners, Philadelphia, p. 14.

[12] Fletcher, R.H., Fletcher, S.W. & Wagner, E.H. (1996). *Clinical Epidemiology: The Essentials*, 3rd Ed. Williams & Wilkins, Baltimore.

[13] Hill, A.B. (1937). *Principles of Medical Statistics*. The Lancet, London.

[14] Hill, A.B. (1947). Statistics in the medical curriculum?, *British Medical Journal.* **ii**, 366–368.

[15] Hill, I.D. (1971). Report on a medical statistical conference, *Journal of the Royal Statistical Society, Series C* **20**, 319–321.

[16] Hopkins, C.E. (1958). Biostatistics instruction in medical schools, *Journal of Medical Education* **33**, 370–372.

[17] Lowe, C.R. (1963). On the teaching of statistics to medical students, *Lancet* **1**, 985–987.

[18] Lowe, S.R. & Lwanga, S.K., eds (1978). *Health Statistics: A Manual for Teachers of Medical Students*. Oxford University Press, Oxford.

[19] Mainland, D. (1952). *Elementary Medical Statistics*. W.B. Saunders, Philadelphia.

[20] Morton, R.F., Hebel, J.R. & McCarter, R.J. (1996). *A Study Guide to Epidemiology and Biostatistics*, 4th Ed. Aspen, Gaithersburg.

[21] Qureishi, B.A. (1979). *The Successful Teaching of Statistics to Every Medical Student*. World Health Organization, Geneva.

[22] Royal Commission on Medical Education, 1965–68 (1968). Report. HMSO (Command 3569), London.

THEODORE COLTON & STEPHEN W. LOONEY

Teaching Statistics to Physicians

There should be little disagreement over the proposition that a practicing physician requires a sound grasp of statistical concepts, and it is not surprising that medical statisticians have emphasized this [2, 4, 6]. It is, however, particularly encouraging that professional medical bodies such as the UK General Medical Council argue that the attributes of a practicing clinician should include [8]:

1. reasoning and judgement in the application of knowledge to the analysis and interpretation of data, in defining the nature of the problem, and in planning and implementing a strategy to resolve it, and
2. understanding of the contribution of research methods, and interpretation and application of others' research in the doctor's own specialty.

The very essence of clinical medicine is decision making under uncertainty. When a patient first presents, the physician needs to adopt a strategy that drives an investigation plan. This involves a sequence of decisions. Using any available prior information on the patient, together with the presenting symptoms and a knowledge of the epidemiology and etiology of disease, the doctor must decide what investigations are needed to confirm a working diagnosis, or to exclude an implausible but sinister diagnosis. Once the diagnosis is clarified, the attention turns to prognosis and treatment. Is the condition self-limiting, requiring at most symptomatic treatment, or is the prognosis so grim that only palliative treatment is appropriate? Is the condition such that it is appropriate to intervene in an attempt to improve the prognosis, and, if so, what evidence is there to support the use of different treatments for this individual? The likely side effects and cost implications need to be considered, and if treatment is initiated, whether further monitoring would be required. Would there be any obvious early signs that the treatment is not having the desired effect, and, if so, what modification to the treatment would be appropriate? On the other hand, if the problem does resolve, are there likely to be any long-term sequelae which mean that the patient should be monitored closely in future? At every step along this process, the physician needs to make decisions against a background of uncertainty, and such decisions should be informed by an understanding of statistical principles.

Much of what has been written about teaching statistics to physicians has focused on research, and in particular it has highlighted the many statistical failings that can be found in the medical literature [3, 10, 13]. This is hardly surprising, as most medical statisticians work with a highly selected group of academic clinicians who are actively involved in conducting research. However, I would argue that it is even more important that a practicing clinician should, for example, have a sound grasp of how the results of an investigation should update the prior probability of a particular diagnosis. This does not necessarily mean that every clinical decision should be based on a back-of-an-envelope application of **Bayes' Theorem**, but in selecting a test and interpreting its results a physician should at least be aware of the distinction between a highly **sensitive screening** test and a highly **specific** investigation which might precede a final decision to undertake a major operation (*see* **Decision Analysis in Diagnosis and Treatment Choice**).

What Statistics should be Taught?

In line with the suggestions of Leinster [9], I agree that we should be targeting three identifiable groups of physicians with quite different requirements in terms of statistical education. The first and by far the largest group consists of those practicing clinicians who are never likely to work in a research environment. The second group consists of the doctors who at some stage in their clinical training will take time out to work on research as part of their career development. Finally, there is the relatively small group of doctors who will remain within the academic environment and be active researchers throughout their careers. I shall refer to these three groups as "practitioners", "casual researchers", and "professional researchers", respectively.

Practitioners

These doctors require a core knowledge of statistics very much in line with the attributes set out by the General Medical Council and quoted above. Thus, these individuals should be aware of the

basic principles of problem formulation and strategies for problem solving, particularly in the context of diagnosis and **prognosis**. In addition, practitioners need to be sufficiently familiar with the "**scientific method**" to be able to use the research results of others to inform and guide their own clinical practice.

It is implicit in this that doctors should keep up to date with the literature in their specialty, by reading research reports and/or systematic reviews. Thus, any teaching should include an introduction to critical appraisal and the principles of **evidence-based medicine**.

Casual Researchers

The rather pejorative label that I have assigned to this group is not without just cause, as much that is wrong with medical research has its roots in the system which makes a curriculum vitae embellished with a number of token publications a prerequisite for career development. This argument is very well developed in an editorial by Altman entitled "The scandal of poor medical research" [1]. Altman argues that it is the system which should be changed. While heartily endorsing this view, I think that until such a change is achieved we have to accept that a very substantial volume of research is being undertaken by clinicians who are ill-equipped as researchers, and we must seek strategies for damage limitation.

For this group the priority should be to teach the principles of study design, and in particular the requirement to have a well defined question. Almost by definition, a well designed study that addresses a well defined question will generate data that can be presented and analyzed using a very limited set of statistical tools, and so this group does not need to be exposed to a large volume of methodological detail. However, they do need to be taught when it is appropriate to seek advice from a more experienced researcher, including perhaps a professional statistician.

Professional Researchers

One could argue that this group, which is small in number but highly influential, should be the priority for our teaching efforts [11]. Not only are these individuals active in conducting their own research, but they are responsible for training and supervising junior research workers. They are the individuals who edit journals and review submitted manuscripts (*see* **Statistical Review for Medical Journals**), and they also review grant applications. They therefore have great influence on what research is undertaken, and on how results are reported. These individuals not only require a sound grasp of research methodology, but also need to be aware of their own limitations, and should recognize the benefits of working closely with an experienced statistician from the outset of a research project.

When should Statistics be Taught?

Practitioners

The only practical mechanism for reaching all practitioners is to include statistics in the undergraduate medical curriculum (*see* **Teaching Statistics to Medical Students**). Dixon [6] has proposed a syllabus for such a core course based on current practice in UK medical schools, but this tends to emphasize techniques rather than "deep" understanding.

The situation with regard to teaching statistics in the UK medical schools is very well documented. Since 1980, there has been an annual meeting of the individuals responsible for this teaching, and a summary of each school's undergraduate and postgraduate teaching is compiled each year [5]. In what was an admittedly rather cynical exaggeration, Peters [12] described the situation with undergraduate teaching of medical statistics as "all get it, while few of them want it". At least part of the problem is that, in the majority of medical schools, the statistical teaching comes very early in the curriculum, while the students are concentrating on pre-clinical subjects and before they have been exposed to any clinical problems that would motivate the use of statistical methods. Current moves within the UK to change undergraduate medical courses to become more integrated, with clinical problems being discussed from the very outset, might help to make the relevance of our discipline more obvious to the students. Dunn & Everitt have recently published an introductory textbook [7] which presents medical statistics from this angle, and which should be a useful backup to teaching in a more integrated curriculum.

Researchers

By the time that clinicians are actively involved in research there is generally no problem with motivation, although often the cry for statistical help comes too late, when a poorly designed study is beyond salvage. These individuals should have the opportunity of attending a comprehensive course in research methodology at the outset of their research. As mentioned above, this would emphasize study design, and the importance of the formulation of the research question. A healthy dose of skepticism and the acquisition of skills in critical appraisal nurtured at this stage will be invaluable to their research, whether or not the individuals eventually become career researchers.

How should Statistics be Taught?

Undergraduate

Given the size of a typical undergraduate medical class, the usual practice is for much of the teaching of medical statistics to be based around formal lectures. The paper by Dixon [6] describes a typical undergraduate curriculum, but Appleton [4] suggests that a series of well chosen case studies might be more effective in demonstrating key concepts such as variability and bias. An alternate approach that is used in the University of Edinburgh Medical School is to use a self-tutoring work book. The work book is supplemented by a small number of formal lectures, and by small group tutorials [5].

With the introduction of more integrated medical courses, there is potential to make the relevance of statistics far more obvious. For example, if a basic practical class in physiology was based around the measurement of blood pressure and heart rate before and after exercise, then there would be great scope to incorporate a statistical component looking at issues such as observer variability (*see* **Observer Reliability and Agreement**), **bias**, graphical representation of data (*see* **Graphical Displays**), and **correlation** and **association**. Moreover, if experienced statisticians were involved more actively in teaching which was integrated with other medical disciplines, it would help to ensure that bad practices were not being promoted by teaching staff who were inexpert in statistics.

An issue to consider is the role of computers in teaching medical statistics. There are two aspects to this; namely, the use of computers as a tool for data handling and statistical analysis, and the use of computers to deliver teaching materials. I would see the former as a low priority for undergraduate courses, but computer-assisted learning might well have a role in teaching even very large classes. One commercial package, for example, is *Statistics for the Terrified* (Version 3, Radcliffe Medical Press Ltd, Abingdon, UK), which could be a complement to a series of lectures.

Postgraduate

Teaching at a postgraduate level can be much more flexible, as the numbers of students are more manageable. One format that I have used extensively is an introductory course with seven or eight two-hour sessions, each session consisting of a lecture followed by a practical exercise that reinforces the lecture material. Topics covered include statistics in the medical literature, looking at data, testing (*see* **Hypothesis Testing**) and **estimation**, study design, and case studies based on data provided by the students. The practical sessions have been computer-based, using the statistical package Minitab (*see* **Software, Biostatistical**). Such an approach is not without its problems, as it risks confusing the students with both the statistical concepts and the computing! However, an almost universal finding in our course assessment questionnaires is that the students find the interplay between the concepts and their practical outworking very helpful, reinforcing their understanding. It is also unrealistic to pretend that medical researchers do not have access to computers and statistical software, and so it is surely important to teach good practice such as using graphical presentations to explore the assumptions underlying a formal statistical analysis.

At this level, statistical education should be an ongoing activity rather than simply an initiation into the world of research. An introductory course can be followed up by more specialist courses that might be discipline-based (e.g. **clinical trials** in cardiology) or topic-based (e.g. systematic reviews; *see* **Meta-analysis of Clinical Trials**). As a further part of this continuing education, there is a healthy trend for medical journals to publish series of expository articles on statistical topics. The *British Journal of Cancer*, the

British Medical Journal, and the *Journal of the American Medical Association*, for example, all have a good track record in publishing statistical articles that are accessible to a general medical readership (*see* **Medical Journals, Statistical Articles in**). This not only has a direct educational effect, but such endorsement by leading journals also helps to highlight the key role of statistics in medical research.

Finally, I would also view my role as a statistical consultant as an educational opportunity (*see* **Statistical Consulting**). Whether seeing a client in a one-off advisory role, or working long-term as a full member of a multidisciplinary research group, there is ample opportunity to reinforce statistical principles.

Summary

It is clear that a knowledge of statistical techniques and the "scientific method" is crucial for both practicing clinicians and research workers. Yet there is ample evidence that as a profession we have frequently failed in our efforts to educate our clinical colleagues. Undergraduate students are confused and alienated, and published medical research continues to demonstrate a widespread and fundamental lack of understanding of key statistical principles. I have considered the situation in the UK, but the position is broadly similar in many other countries.

If we are to succeed in resolving some of these problems, then I believe that we need to focus on three areas. First, we should aim to equip all clinicians with the statistical skills necessary for day-to-day decision making, recognizing that most clinicians will only ever be consumers of medical research. Secondly, we should recognize that much research is being conducted by inexperienced research workers whose motivation is at least partly career development. This group needs to be taught primarily the principles of study design, and helped with the presentation and interpretation of their data. Thirdly, we should be committed to developing long-term working relationships with senior clinical colleagues, so that their statistical education can be extended and reinforced in the context of their own work.

References

[1] Altman, D.G. (1994). The scandal of poor medical research, *British Medical Journal* **308**, 283–284.

[2] Altman, D.G. & Bland, J.M. (1991). Improving doctors' understanding of statistics (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 223–267.

[3] Andersen, B. (1990). *Methodological Errors in Medical Research: an Incomplete Catalogue*. Blackwell, Oxford.

[4] Appleton, D.R. (1990). What statistics should we teach medical undergraduates and graduates?, *Statistics in Medicine* **9**, 1013–1021.

[5] Appleton, D., ed. (1996). *The Teaching of Medical Statistics at Undergraduate and Postgraduate Levels in the United Kingdom*. University of Newcastle, Newcastle-upon-Type.

[6] Dixon, R.A. (1994). Medical statistics: content and objectives of a core course for medical students, *Medical Education* **28**, 59–67.

[7] Dunn, G. & Everitt, B. (1995). *Clinical Biostatistics: an Introduction to Evidence-based Medicine*. Edward Arnold, London.

[8] General Medical Council (1993). *Tomorrow's Doctors: Recommendations on Undergraduate Medical Education*. General Medical Council, London.

[9] Leinster, S.J. (1991). Contribution to the Discussion of D.G. Altman, and J.M. Bland, *Journal of the Royal Statistical Society, Series A* **154**, 253–254.

[10] Murray, G.D. (1988). The task of a statistical referee, *British Journal of Surgery* **75**, 664–667.

[11] Murray, G.D. (1990). How we should approach the future?, *Statistics in Medicine* **9**, 1063–1068.

[12] Peters, T.J. (1990). Comment on D.R. Appleton, *Statistics in Medicine* **9**, 1023–1027.

[13] Pocock, S.J., Hughes, M.D. & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials: a survey of three medical journals, *New England Journal of Medicine* **317**, 426–432.

GORDON D. MURRAY

Telephone Sampling

The use of the telephone for sample survey data collection requires the selection of samples of telephone numbers to identify sampling units for interviewing. The sampling techniques employed to make these selections are those used for many other problems where samples must be selected. However, several unique features of the **sampling frames**, the sets of materials useful for sample selection, have stimulated the development of sample designs specific to telephone surveys.

The available frames vary from one country to the next as telephone system characteristics vary. Since our own experience has been limited to telephone sampling frames available in the US, the discussion here concerns frames and sampling methods to select telephone households in the US. Frames and sampling methods in other countries will have similar features to those described here, although specific aspects of the frames and methods may require modification to improve the efficiency or other properties of survey operations.

Telephone sampling methods have largely developed and been applied in the context of household surveys. The methods can and have been adapted to other populations such as establishments. The present discussion is restricted, though, to sampling methods for household populations.

The presentation is divided into four major sections. Background on frames and basic telephone sampling methods are described in the next section. The following section addresses specific telephone sample designs, while the subsequent section presents estimators for the principal telephone sampling methods. The final section is a comparison of designs based on cost, **variance**, implementation, and **bias** considerations.

Frames and Basic Telephone Sampling Methods

The Telephone Household Population

Often the population of interest is broader than telephone households, seeking to include all households regardless of whether they have a telephone. To reduce the costs of data collection through the use of the telephone, investigators decide to compromise on the **target population**, defining a survey population of telephone households when in fact they seek to make inferences about all households. The disjuncture between target and survey population for many telephone surveys raises several important issues concerning noncoverage of households without telephones and the appropriateness of **inference** to a population other than the survey population.

Noncoverage of households without telephones can introduce **bias** into sample estimates. The bias depends both on the proportion of households that are not covered and the differences between telephone and nontelephone households on the characteristics of interest. Approximately 5% of US households do not have a telephone, and the percentage of persons who live in such households is even smaller. While the overall rate of noncoverage is small, and may be reassuring to some investigators, noncoverage varies substantially with a number of characteristics that may be related to variables being measured in a survey. For example, nontelephone households tend to have younger and more mobile populations and to be located in rural areas of the South of the US and in central cities. Noncoverage rates can rise to 15% or higher for some subpopulations, a level that is considered unacceptable to those who need to produce estimates for many small subgroups of the US population from the survey data.

The characteristics of nontelephone households have been examined in several reviews (see, for example, [15]). Nontelephone households tend to have higher rates of unemployment, have higher rates of smokers, and experience higher rates of crime victimization. Telephone surveys could produce biased estimates of employment, health, or social characteristics. Adjustments to telephone survey data to attempt to compensate for noncoverage may reduce the bias. The use of **poststratification**, or population control adjustments, for this purpose are discussed later.

Telephone Systems

Telephone systems vary from country to country, but there are features of the systems that are similar despite the variation. Telephone numbers are grouped into geographical areas. For example, in the US, telephone numbers consist of three parts: a three-digit area code, a three-digit prefix (or central office code), and a four-digit suffix. The area code and prefix

are established as part of an international system that extends across the US, Canada, Mexico, and the Caribbean. These numbers are not, of course, assigned at random across the entire geographic area covered by the phone system. Area codes are assigned to specific geographic regions that in the US do not, for the most part, cross state boundaries but otherwise do not correspond to political boundaries. Thus, there is a one-to-one correspondence between an area code and a geographic area. For instance, area code 313 is assigned to a region of southeast Michigan including a sizable portion of Detroit.

Prefixes are repeated across area codes, and within area codes are not generally geographically defined. However, prefixes are grouped into geographic areas called exchanges which are defined for the purposes of providing public service and maintaining the phone service. For example, the Ann Arbor exchange within the 313 area code is a geographic area roughly approximating the city of Ann Arbor and surrounding areas and is assigned more than 20 different prefixes. Households and businesses requesting a telephone service within the geographic area defined as the Ann Arbor exchange must be assigned a telephone number whose prefix is one of 20 serviced by the exchange. No other exchange within the same area code can use the prefixes assigned to the Ann Arbor exchange. There is little further geographic differentiation within most exchanges with respect to prefixes. Some exchanges with large numbers of prefixes will be divided into wire centers responsible for a subdivision of the area covered by the exchange and containing a subset of the prefixes assigned to the entire exchange.

The majority of the exchanges in the US are assigned only a single prefix. Exchanges have been until the recent past areas designated by public service commissions within which companies were able to obtain exclusive rights to provide a phone service. Service requirements are such that the land area covered by an exchange is limited, yet population density for a given exchange can vary enormously. Thus, some exchanges have very few customers, and enough numbers are available in a single prefix to assign to all customers. Other exchanges have large numbers of customers, and multiple prefixes are assigned to the exchange.

Suffixes are grouped in sets of 10 000. They are typically assigned by local telephone service personnel based on existing assignment patterns. There does not appear to be any particular system by which new customer requests for services are assigned suffixes within a prefix.

However, patterns of the assignment of suffixes within prefixes do emerge when the entire system is examined. In exchanges with multiple prefixes and larger numbers of customers, prefixes and suffixes are assigned haphazardly, depending on the availability of unassigned numbers within a prefix at the time of assignment. But in exchanges with a single prefix and a small number of customers, suffixes have been assigned in groupings to reduce the cost of assignment and to make telephone assignment easier. Older electromechanical switching equipment allowed smaller companies to assign all numbers in a single "1000-bank" of consecutive numbers all beginning with the same first digit of the four digit suffix. A company would only have to purchase a single bank of 1000 switches for its customers, thereby reducing costs. Telephone numbers in the more numerous single-prefix exchanges are thus effectively clustered, often at the 1000-bank level, as well as at the 100-bank level. Several telephone sampling methods described subsequently take advantage of this clustered assignment of numbers to improve the efficiency of identifying telephone numbers assigned to residential units.

Sampling Frames

There are four types of frame problems that arise in telephone sampling: listings on the frame that are not elements of the population (referred to as *blanks*); elements in the population for which there is no corresponding listing (*noncoverage*); listings on the frame which yield multiple elements in the population (*clustering*); and elements in the population which have two or more listings on the frame (*duplicates*). Each of these deficiencies can lead to bias in survey estimates or inefficiency in survey operations. Sampling statisticians develop selection procedures which try to reduce or eliminate bias due to these deficiencies. They also have been instrumental in finding selection procedures which reduce the inefficiencies associated with some of these deficiencies.

Three principal frames are used for telephone sampling: telephone numbers, directories, and commercial lists. The frame of telephone numbers can be created through a combination of a list of area

code and prefix combinations with randomly generated suffixes. The area code prefix combinations can be obtained for local studies from examination of local telephone directories, which are fairly up to date at the prefix level. For surveys covering larger areas than a local community, area code prefix combinations can be obtained in the US from Bell Core Research, Inc. The BCR frame is updated monthly and contains all area code and prefix combinations for the US, as well as for Canada, Mexico, and the Caribbean. The area codes and prefixes must be subset to the US to reduce the amount of screening of generated numbers for US telephone household surveys.

While the BCR frame affords virtually complete coverage of telephone households, it suffers from a substantial number of blank listings. Less than 25% of the generated numbers are assigned to residential units. It is operationally inefficient to use a simple **random digit dialing** scheme of area code, prefix, and randomly generated suffix. Other methods have been developed to take advantage of the inherent clustering of residential numbers in 1000 banks that increase the proportion of generated numbers assigned to residential units to more than 60%.

The BCR frame with randomly generated suffixes also has the disadvantage of duplicate listings. Households with more than one telephone number used for residential purposes are represented on the frame multiple times. **Probability sampling** methods require that the number of telephone numbers in a household be acquired and used to develop a compensatory weight for estimation.

Directories have been widely used as a frame for local studies. They are inexpensive to acquire for a local area, and simple list sampling methods can be used to select samples quickly, although not necessarily easily. Directory frames are difficult to assemble for wider geographic areas, with more than 5000 directories published across the US each year. Their popularity as a sampling frame is due to cost and convenience and to the lower proportion of blank listings in the directory compared to the telephone number frame: approximately 10%–15% of listings in a residential directory in the US are no longer residential.

On the other hand, directory frames suffer from noncoverage of the telephone household population due to unlisted numbers and changes in the telephone status of households. The percentage of telephone households which do not appear in directories exceeds 35% in the US, varying from low percentages (10%) in suburban and rural areas to more than 60% in some urban locations in the West such as Los Angeles. Survey designers generally do not ignore these high proportions of unlisted or out-of-date listings, and choose to use random digit dialing methods or other schemes that afford higher coverage.

Further, directories have higher levels of duplicate listings because subscribers can purchase additional listings. For example, a married couple at the same address with different surnames may choose for a small fee to appear in the directory under both names. The duplicate listing increases the chance of their telephone household being selected, which must, from a probability sampling point of view, be compensated through a weight for the household. Thus, a telephone directory has duplicate listings both because of multiple telephone numbers per household and multiple listings of the same telephone number in the directory.

Commercial firms now assemble electronic files in the US based on directories collected from across the country. Directory entries (name, address, and phone number) are either keyed or added to the file when an electronic format is available. Lists based on directories are supplemented by lists of automobile registrations obtained from approximately 30 states that release such data publicly. The combined file is subjected to processing to assign a zip code to each entry for the purposes of mailing. Several firms have taken advantage of the availability of telephone numbers in such files to create national directories and to draw and sell samples of telephone numbers from them. The commercial frames suffer from a small proportion of blank listings as well as the failure to cover unlisted numbers as well as duplicate listings described for directory frames.

Each of the three frames described here have generated different sampling methods that attempt to take advantage of strengths of the frame and reduce the impact of weaknesses in the frame. The methods are often classified as one of three types: simple list frame sampling methods suitable for directories; random digit dialing methods based on the telephone number frame; and list-assisted methods based on directories or commercial lists generating samples that include unlisted numbers as well. We do not discuss the sampling methods as applied to directories

here, but do examine the random digit dialing and list-assisted methods in the next section.

Telephone Sample Designs

All of the sampling designs discussed in this section assume the availability of a frame of telephone numbers which includes all possible telephone numbers in the target population. Very limited auxiliary information is available for the telephone numbers in the BCR frame: exchange name, geographic coordinates for the center of the exchange, and time zone. Importantly, though, the BCR frame does include new area code–prefix combinations approximately 3 months before they are added to the telephone system. Thus, in principle, the BCR-generated frame will provide complete coverage of the telephone household population but with little auxiliary information for the purposes of **stratification** or other design efficiencies.

The primary problem with commercial list frames is incomplete coverage of the telephone household population. Conversely, the primary problem with the BCR-generated frame is the inclusion of many telephone numbers which are not assigned to a household. The development of telephone sampling designs has been motivated almost entirely by a desire to develop an efficient methodology for sampling from the BCR frame.

Sample Designs Using the BCR Frame

The sample designs in this section document statistical sampling methodology for sampling residential households using only the BCR frame. Since approximately 95% of all residential households (and not only telephone households) can be linked to this frame, the researcher must give careful consideration to the question of how well the telephone population represents the target population with respect to the variables of primary interest.

Simple Random Digit Dialing (RDD). The simplest and most direct approach to utilizing the BCR frame is to select telephone numbers from the frame randomly, call the selected numbers and conduct the requisite interview for each number that is found to be connected to an in-scope household. Numbers are selected and called until the desired sample size, say n, of in-scope households is attained. As noted earlier, only about 20%–25% of the sample telephone numbers will be assigned to households, so the number of calls required, say n', will be considerably larger than n. The expected number of required calls is n/p, where p is the proportion of telephone numbers assigned to residential households. Thus, in order to account for the ineligible listing, the sample of telephone numbers from the BCR frame must be four to five times as large as the desired sample of n telephone households.

In general, the determination of the status of a telephone number is a costly matter, especially for telephone numbers not assigned to households. Frequently, a number must be dialed several times in order to determine its status. Since procedures must be specified for each type of dialing outcome, the use of the BCR list (or any list with a high proportion of spurious listing) will greatly increase the administrative and operational costs of telephone survey operations. This general subject is discussed in detail by Lepkowski [9]. For the purpose of constructing a simple cost model, let c_0 be the cost of determining the status of a number not assigned to an in-scope household, c_1 the cost of determining the status of a number assigned to an in-scope household, and c_2 the cost of conducting the survey interview. The total cost of the survey is then given by $C = n(c_1 + c_2) + (n' - n)c_0$ and the expected total cost of a simple random digit dialing survey is given by $\mathrm{E}(C) = n[(c_1 + c_2) + c_0(1 - p)/p]$. Obviously, for p in the neighborhood of 0.20–0.25, the component of expected cost due to unproductive calls, i.e. $nc_0(1 - p)/p$, will be a substantial proportion of total expected cost. The telephone designs described in the following sections were all motivated by a desire to reduce the proportion of cost due to unproductive calls.

The Mitofsky–Waksberg Design. The two-stage random digit dialing design proposed by Mitofsky [10] and more fully developed by Waksberg [17] has been so widely employed in telephone surveys that it has become nearly synonymous with RDD telephone surveys. The method capitalizes on the clustering of telephone numbers assigned to residential households within banks of consecutive telephone numbers. As noted above, only about 20%–25% of the numbers in the BCR frame are assigned to households; however, among banks of 100 consecutive numbers with at least one number assigned to a household,

over 60% of the numbers are assigned to residential households. Clearly, if the 100-banks with one or more residential numbers could be identified and if sampling were restricted to those banks, then the proportion of unproductive calls could be substantially reduced.

The Mitofsky–Waksberg technique starts by grouping the numbers in the BCR frame into 100-banks by using the area code, the three-digit prefix, and first two digits of the suffix to specify each bank. In the first stage 100-banks are selected at random, with replacement (*see* **Sampling With and Without Replacement**), and a telephone number within the bank is selected at random and dialed. If the selected number is found to be eligible, then the bank is retained for second-stage sampling. The process is continued until a specified sample of m 100-banks is attained. Within each retained 100-bank, telephone numbers are selected at random, without replacement, until a total of k eligible numbers (including the original number used to retain the 100-bank) have been identified.

Thus, the Mitofsky–Waksberg technique utilizes a two-stage design where 100-banks are selected – with probability proportional to number of eligible telephone numbers (*see* **Sampling With Probability Proportional to Size**) – in the first stage and a fixed-size sample of eligible households is selected in the second stage. Thus, the sample of $n = mk$ eligible households is selected with equal (but unknown) probability. The efficiency of the Mitofsky–Waksberg technique derives from the fact that the eligible telephone numbers are concentrated in a relatively small proportion of the 100-banks. Letting t be the proportion of 100-banks with no eligible numbers, then the total expected number of calls is $n[1 - t(k-1)/k]/p$ and the expected total cost is

$$\mathrm{E}(C) = n\left\{(c_1 + c_2) + \frac{c_0[1 - p - t(k-1)/k]}{p}\right\}.$$

Clearly both the expected number of calls and the expected cost decrease as k increases. Nationally, t is in the neighborhood of 0.65, so even modest values of k can lead to substantial cost savings.

Although the Mitofsky–Waksberg technique offers an elegant method to improve telephone survey efficiency, there are practical problems. The most obvious is that some 100-banks may have fewer than the requisite k eligible households, in which case all numbers in the bank will, of necessity, be called. Even then, compensatory weighting will be required. Another problem is that it is not always possible to determine accurately the eligibility status of a selected number. In the first stage this may lead to the incorrect inclusion or exclusion of 100-banks. In the second stage, some numbers may still be unresolved at the end of the survey period, so that fewer than k eligible households are identified for the bank. Another more subtle, problem is intrabank **correlation**, which is discussed in more detail in a later section.

The Potthoff Design. The design suggested by Potthoff [11] is similar to the Mitofsky–Waksberg design, except that eligibility is extended to a broader, larger class of telephone number which he termed *auspicious* numbers. Typically the auspicious numbers include not only the residential household numbers, but also ring-without-answer numbers and other results for which the residential status is unknown. This broader definition reduces the amount of screening needed for the first stage and the amount of replacement required at the second stage. Another innovative development by Potthoff [11, 12] specifies that $c \geq 2$ numbers be selected per bank in the first stage. Sampling in the second stage depends on the number of auspicious numbers observed in the first stage, but this is not discussed in detail here.

The Potthoff sampling design yields an equal probability sample of eligible numbers. Replacement is required for only a small number of selected prefix areas and it reduces ambiguities about the status of numbers dialed at the first stage. Also, as c increases, the chances of obtaining a bank that will be exhausted in the second stage are reduced.

Implementation of the Potthoff design requires knowledge about the proportion of auspicious numbers that are actually eligible numbers in order to determine the appropriate sample size. The administrative structure is more complex and the training requirements are increased for this procedure relative to the Mitofsky–Waksberg.

Sample Designs Utilizing Published Residential Telephone Numbers

As discussed in the first section, lists of published residential telephone numbers for the entire US are available from several vendors. Since 85%–90% of the telephone numbers on these lists are connected

to residential households, a straightforward random or systematic selection of numbers from such a list would be much more efficient than the designs used for sampling the BCR list. Unfortunately, the typical directory-based list only includes about 70% of the residential telephone households. Comparisons of telephone households with and without published numbers indicates that substantial bias may result if households without published numbers are omitted from the sampling frame [2]. The designs discussed in this section attempt to capitalize on the efficiency inherent in directory-based sampling while extending the coverage of the design to include the entire residential telephone population.

Designs Based on Plus Digit Dialing. Plus digit dialing is a directory-assisted procedure in which a sample of telephone numbers is selected from the directory and an integer is added to the suffix of the selected number. For instance, in plus-one dialing the integer "one" is added to the suffix of each number selected from the directory. The resulting sample of telephone numbers generally includes both listed and unlisted numbers; in addition, it yields a higher proportion of productive numbers than does the simple RDD design. Unfortunately this procedure has a number of theoretical problems. In general, the numbers in the target population have unequal and unknown probabilities of selection. In fact, some of the unlisted numbers may have a zero probability of selection unless the unlisted numbers are evenly mixed among the listed numbers. Such a mixing phenomena is, at best, difficult to verify. Generalizations of this design in which the last d digits (two or more) are replaced by a randomly generated d digit number have been suggested.

A closely related design, based on half-open intervals of telephone numbers, was suggested by Frankel & Frankel [5]. In numeric-order directories a cluster is defined to consist of a listed telephone number together with all numbers up to, but not including, the next listed number. A sample of clusters is selected from the directory by simply selecting a **simple random sample** of telephone numbers from the directory. This method achieves known, nonzero, probabilities of selection for all telephone households; however, the potentially large variation in cluster size can introduce formidable operational problems. Furthermore, this method is subject to estimation difficulties as cluster size and sample are both random variables. This basic design can be modified for use with alphabetical-order directories, but in this case the theoretical and operational problems are compounded by reporting error problems.

A Design Based on Two-Stage Sampling. A two-stage sampling design, utilizing a directory list, was proposed by Sudman [14]. This procedure, which was originally suggested by Stock [13], uses 1000-banks of telephone numbers (which are identified by the first six digits of the telephone number) as the first-stage sampling unit. The selection of 1000-banks is similar to the first-stage selection in the Mitofsky–Waksberg method except that the directory of listed numbers is used to select the first-stage sample. Thus, the probability of selection in the first stage is proportional to the number of listed numbers in the 1000-bank. In the second stage, numbers are selected until a predetermined fixed number of listed numbers are selected, and interviews are attempted for households with both listed and unlisted numbers. It should be noted that unlisted numbers in 1000-banks with no listed number have zero probability of selection, but in most cases this is not a serious problem. Of more concern is the fact that the determination of listing status often depends on a respondent report which can be in error; however, use of a directory in reverse telephone number order can eliminate this source of error.

Unlike the Mitofsky–Waksberg method, the Sudman procedure will produce unequal-size clusters of sample telephone households, although the variation in cluster size is usually not very large. Also, the potential for exhausted clusters exists, but with 1000 numbers (instead of 100 numbers, as in the Mitofsky–Waksberg method) this is of minor concern.

Designs Using Both the BCR Frame and Published Telephone Numbers

It should be noted that the designs discussed earlier require only the BCR frame, while those just discussed require only a published list of residential telephone numbers. The designs discussed in this Section require both. The basic idea behind these designs is to unite directly the desirable coverage properties of the BCR frame with the relatively high sampling efficiency of a frame of listed telephone numbers.

Dual Frame Designs. An RDD sample of n_B telephone households is selected from the BCR frame and simultaneously a sample of n_D telephone households is selected from the directory list frame. Letting n'_B and n'_D be the respective number of calls required to achieve the desired sample sizes, the cost of the dual frame design is given by

$$C = (n_B + n_D)(c_1 + c_2) + c_0(n'_B + n'_D - n_B - n_D).$$

The expected cost of a dual-frame survey is given by

$$E(C) = n\left\{c_1 + c_2 + c_0\left[\frac{\lambda(1-p_B)}{p_B} + \frac{(1-\lambda)(1-p_D)}{p_D}\right]\right\},$$

where $n = n_B + n_D$ is the total sample size, $\lambda = n_B/n$ is the proportion of the total sample allocated to the BCR frame, p_B is the proportion of telephone numbers in the BCR frame assigned to residential households, and p_D is the proportion of telephone numbers in the directory frame assigned to residential households. As p_B is in the neighborhood of 0.20–0.25 and p_D is usually in the neighborhood of 0.80–0.85, the expected cost (for a fixed total sample size n) will decrease as λ decreases.

There are several possible ways to combine the data from the two frames for estimation. In general, dual-frame estimators are more complicated than the estimators for the previously discussed designs. Groves & Lepkowski [6] provide a detailed discussion of the issue of dual-frame estimation and the problem of sample allocation to the two frames so as to attain the minimum cost for a specified **variance**.

To implement dual-frame methodology, the directory status (i.e. listed or unlisted) of each residential household from the BCR sample must be known. To avoid using potentially unreliable respondent reports regarding their listing status, numbers selected from the BCR frame can be matched to the directory list at the time of sample selection. If the directory frame contains addresses for the listed numbers, then it is possible to send advance letters for the purpose of improving response rates. In general, the dual-frame design requires a sophisticated administrative operation; also, costs may be increased by the need to match the BCD sample to the directory frame and by the use of a more complicated estimator. The benefits of a higher response rate should more than offset the costs of advance letters.

Directory-Based Stratification. For this design the directory list is used for the purpose of stratifying the BCR frame so as to improve sampling efficiency. In a typical application, the directory list is used to identify all 100-banks in the BCR frame with one or more directory listed telephone numbers. The BCR frame is then partitioned into two strata; one stratum contains all telephone numbers in 100-banks with one or more listed numbers and the other stratum contains all other numbers. The first stratum is often referred to as the high-density stratum, while the second is referred to as the residual stratum. Simple RDD samples are then selected from each stratum, with a much larger sample selected from the high-density stratum. The basic strategy behind this design is the same as for the Mitofsky–Waksberg method, i.e. telephone numbers for residential households tend to be highly clustered within 100-banks with listed numbers, so if banks containing such telephone numbers can be identified and sampled at a higher rate, then sampling efficiency can be greatly improved. Casady & Lepkowski [3] found that at the national level the proportion of the BCR frame assigned to the high-density stratum would be approximately 0.38 but it would contain about 95% of the numbers assigned to residential households. Thus, the proportion of numbers in the high-density stratum assigned to households is approximately 0.55, while the proportion of numbers assigned to households in the residual stratum is only about 0.02.

Assume that an RDD sample of n_1 telephone households is selected from the high-density stratum and that a sample of n_2 telephone households is selected from the residual stratum. Then, the cost for the stratified design is given by

$$C = (n_1 + n_2)(c_1 + c_2) + c_0(n'_1 + n'_2 - n_1 - n_2),$$

where n'_1 and n'_2 are the respective numbers of calls required to achieve the desired sample sizes. The expected cost of the stratified sample is

$$E(C) = n\left\{c_1 + c_2 + c_0\left[\frac{\gamma(1-p_1)}{p_1} + \frac{(1-\gamma)(1-p_2)}{p_2}\right]\right\},$$

where n is the total sample size, $\gamma = n_1/n$ is the proportion of the total sample allocated to the high-density stratum, p_1 is the proportion of telephone

numbers in the high-density stratum assigned to residential households, and p_2 is the proportion of telephone numbers in the residual stratum assigned to residential households. As p_1 is in the neighborhood of 0.55 and p_2 is usually in the neighborhood of 0.02, the expected cost (for a fixed total sample size n) will decrease as γ increases. The allocation of the sample to the strata to minimize the cost for a fixed variance (or minimize the variance for a fixed cost) is discussed in detail in [3].

The probability of selection is known, positive, and equal within a stratum, so the estimation of population totals is straightforward. The estimation of a population **mean** at the stratum level is also straightforward, but the estimation of the overall population mean requires that the total residential telephone population be estimated and then a ratio estimator (*see* **Ratio and Regression Estimates**) be used to estimate the population mean. A more detailed discussion of estimated means and variances is given later.

Under the relatively simple cost model given above, this design compares favorably with the Mitofsky–Waksberg design. In practice, directory-based stratification with simple RDD sampling within stratum has proven to have an advantage with respect to implementation and administration. There are two costs associated with this design that are not included in the simple model: the cost of the commercial list itself and the cost of stratifying the BCR frame. The cost of the commercial list will vary with vendor and with time, but for any large-scale, continuing survey operation this should be a relatively minor cost component. Both the costs cited above are fixed costs and can be amortized over multiple studies to reduce greatly the impact on any single study.

Directory-Based Truncation. This approach is really a special case of the preceding one in that no sample is allocated to the residual stratum, i.e. the BCR frame is truncated by removing the residual stratum. The greatly increased hit rate, together with the other advantages of the directory-based stratification design, make this an extremely attractive approach. The obvious disadvantage is that not all of the target population is accessible when the frame is truncated. In the example given above, approximately 5% of the telephone population will not be covered by the truncated frame. However, experience has indicated [1] that for many variables the out-of-scope population is very similar to the target population, so that very little bias results from truncation. As previously noted, approximately 5%–7% of the household population is not included in the telephone population, and any additional bias due to truncation of the BCR frame is probably minimal.

Estimation

The probability features of these designs must be taken into account in the computation of estimates from the samples. The basic principles of such estimation are described briefly here for means (and by implication, for proportions) and their sampling variances. In addition, poststratification, or population control adjustment, is in some cases applied to telephone survey data to attempt to adjust the telephone household sample to the distribution of all households.

Estimating Means

For the simple RDD design, let $\overline{Y}_{\mathrm{RDD}}$ be the simple mean of the n observations of the household variable y. Similarly, let $\overline{Y}_{\mathrm{MW}}$ be the simple mean of the mk observations under the Mitofsky–Waksberg design. Both $\overline{Y}_{\mathrm{RDD}}$ and $\overline{Y}_{\mathrm{MW}}$ are design-unbiased for the population mean μ; furthermore, $\mathrm{var}(\overline{Y}_{\mathrm{RDD}}) = \sigma^2/n$ and $\mathrm{var}(\overline{Y}_{\mathrm{MW}}) \cong (\sigma^2/mk)[1 + \rho(k-1)]$, where σ^2 is the population variance and ρ is the intra-100-bank correlation for the variable y.

The estimation of the population mean for the directory-based stratified designs is somewhat more complicated. Sampling within stratum is RDD, so $\overline{Y}_h$ (the simple mean of the n_h observations from the hth stratum) is unbiased for the stratum population mean μ_h. It follows that $Y'_{\mathrm{t}} = \sum_{h=1}^{H} N_h(n_h/n'_h)\overline{Y}_h$ is approximately unbiased for the population aggregate of the y values for telephone households and $N'_{\mathrm{t}} = \sum_{h=1}^{H} N_h(n_h/n'_h)$ is approximately unbiased for the total number of telephone households, say N_{t}. Thus, the ratio estimator, $\overline{Y}_{\mathrm{Strat}} = Y'_{\mathrm{t}}/N'_{\mathrm{t}}$, is approximately unbiased for the population mean and

$$\mathrm{var}(\overline{Y}_{\mathrm{Strat}}) \cong \sum_{h=1}^{H} \frac{z_h^2\sigma_h^2[1 + (1-p_h)\lambda_h]}{n_h},$$

where p_h is the proportion of telephone numbers in the hth stratum assigned to residential households, z_h

is the proportion of the telephone household population included in the hth stratum and $\lambda_h = (\mu_h - \mu)^2/\sigma_h^2$.

Several other statistical issues should be kept in mind when utilizing telephone designs:

1. In general, ratio estimators are required for estimating subclass means, in which case the relatively simple variance expressions above are not applicable.
2. The designs above yield samples of households, not persons. If persons are selected within households then additional weighting and more complex estimators are required.
3. To have unbiased estimators, the weights of households with multiple telephones must be adjusted to account for their higher probability of selection.
4. The estimators above are based on the use of random digit dialing to achieve fixed sample size. This requires that the status of all numbers selected be determined, which, in turn, requires careful record keeping and close supervision. Because fixed sample sizes are required for each retained 100-bank, the Mitofsky–Waksberg method is more complex and thus the need for tight control is even more important.

Estimating Sampling Variance

For the purpose of estimating $\text{var}(\overline{Y}_{\text{RDD}})$, we let Y_i be the value of the variable y for the ith household selected. An unbiased estimator for $\text{var}(\overline{Y}_{\text{RDD}})$ is given by $\widehat{\text{var}}(\overline{Y}_{\text{RDD}}) = \hat{\sigma}^2/n$, where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_{\text{RDD}})^2}{n-1}.$$

For the Mitofsky–Waksberg sampling we let Y_{ij} be the value of the variable y for the jth selected household in the ith retained 100-bank. An unbiased estimator for $\text{var}(\overline{Y}_{\text{MW}})$ is given by

$$\widehat{\text{var}}(\overline{Y}_{\text{MW}}) = \frac{1}{m}\frac{\sum_{i=1}^{m}(\overline{Y}_i - \overline{Y}_{\text{MW}})^2}{m-1},$$

where

$$\overline{Y}_i = \frac{\sum_{j=1}^{k} Y_{ij}}{k}.$$

For the stratified design we let Y_{hi} be the value of the variable y for the ith household selected in the hth stratum. Applying the **linearization** technique to the ratio estimator $\overline{Y}_{\text{Strat}}$ yields the variance estimator

$$\widehat{\text{var}}(\overline{Y}_{\text{Strat}}) = \sum_{h=1}^{H} \frac{\hat{z}_h^2\hat{\sigma}_h^2[1 + (1-\hat{p}_h)\hat{\lambda}_h]}{n_h},$$

where

$$\hat{p}_h = \frac{n_h}{n_h'},$$

$$\hat{z_h} = \frac{N_h\hat{p}_h}{N_t'},$$

$$\hat{\sigma}_h^2 = \frac{\sum_{i=1}^{n_h}(Y_{hi} - \overline{Y}_h)^2}{n_h - 1},$$

and

$$\hat{\lambda}_h = \frac{(\overline{Y}_h - \overline{Y}_{\text{Strat}})^2}{\hat{\sigma}_h^2}.$$

Although results are not given in detail, the linearization technique can also be used to derive estimators for the variance of the ratio estimators required for subclass means.

Poststratification

In traditional sampling theory, poststratification arises when the variables to be used to create strata are not available at the time of selection. That is, one may be interested in partitioning the population into G poststrata using variables collected during the survey. As under proportionately allocated stratified sampling, improvements in precision are possible with suitable modification to variance estimation. Poststratification requires that for each sample element the poststrata be known and that poststratum weights, say W_g, are available for each poststratum. The poststratum weights must come from an outside source such as a **census**, census projections, or administrative records. For example, poststrata based on age and

gender may be created for the respondents if suitable population counts or proportions W_g can be found for age and gender groups in the population. In telephone sampling, poststratification often adjusts not to the population residing in telephone households, but rather to the population residing in all households. This form of poststratification is applied to obtain estimates that have, in a certain sense, been adjusted to the distribution of the population in all households and not just telephone households.

In summary, poststratification is applied as follows:

1. Sort the sample into G poststrata based on some observed characteristic(s).
2. Obtain sample weights W_g for the population, typically from an outside source such as a larger survey, census or census projection data, or administrative records.
3. Compute the means $\overline{Y}_g$ for the characteristic of interest separately for each poststratum and compute the overall mean $\overline{Y}_{\text{ps}} = \sum_{g=1}^{G} W_g \overline{Y}_g$.
4. For variance estimators, use

$$\widehat{\text{var}}(\overline{Y}_{\text{ps}}) \cong \frac{1}{n}\left[\sum_{g=1}^{G} W_g S_g^2 + \sum_{g=1}^{G} W_g(1-W_g)\frac{S_g^2}{N_g}\right]$$

or, alternatively,

$$\widehat{\text{var}}(\overline{Y}_{\text{ps}}) \cong \frac{1}{n}\sum_{g=1}^{G} W_g S_g^2\left[1+\frac{1-W_g}{N_g}\right],$$

where S_g^2 is an estimator for the within-poststratum element variance, and N_g is the population size for poststratum g. The form of the estimators S_g^2 will depend on the sample design.

In almost all practical situations the poststratified estimate $\overline{Y}_{\text{ps}}$ will have smaller variances than the estimated mean without the poststratification.

Poststratification is also often referred to as population-control adjustment. Generally, poststratified weights are applied at the element level, and weighted estimates computed using the poststratified weights are "adjusted" to the outside distribution represented by the W_g. In the case of the RDD design the effects of this adjustment can be seen more clearly if we reexpress the poststratified estimate of the mean as follows. Let r denote the number of respondents in the sample and r_g denote the number of respondents in the gth poststratum. In addition, let Y_{gi} denote the value of characteristic Y for the ith respondent in the gth poststratum. Then the poststratified mean can be written in terms of element weights w_{gi} as follows:

$$\begin{aligned}
\overline{Y}_{\text{ps}} &= \sum_{g=1}^{G} W_g \overline{Y}_g = \frac{\sum_{g=1}^{G} N_g \overline{Y}_g}{N} \\
&= \frac{\sum_{g=1}^{G}\left(\frac{r}{N}\right)\left(\frac{N_g}{r_g}\right)\sum_{i=1}^{r_g} Y_{gi}}{\sum_{g=1}^{G}\left(\frac{r}{N}\right)N_g} \\
&= \frac{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi} Y_{gi}}{\sum_{g=1}^{G}\frac{N_g/N}{1/r}} \\
&= \frac{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi} Y_{gi}}{\sum_{g=1}^{G}\sum_{i=1}^{r_g}\left(\frac{1}{r_g}\right)\frac{N_g/N}{1/r}} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi} Y_{gi}}{\sum_{g=1}^{G}\sum_{i=1}^{r_g} w_{gi}}.
\end{aligned}$$

That is, the weight w_{gi} is the ratio of the proportion in the population in the gth poststratum to the proportion in the sample in the gth stratum: $w_{gi} = (N_g/N)/(r_g/r)$. Thus, poststratification of a telephone household sample of respondents to a distribution based on all households provides a simultaneous adjustment for nonresponse and noncoverage of the households without telephones.

There are several features of poststratification for telephone samples that are important to observe. Typically, the W_g are census or other related data for all households, not just telephone households. Secondly, while the poststratification adjustment may be viewed as an adjustment for both **nonresponse** and noncoverage (*see* **Nonsampling Errors**) it is

often applied in practice after some form of nonresponse compensation through weighting. Thirdly, it may not be possible to obtain population weights W_g across a full cross-classification of characteristics for the population, but marginal distributions may be available. Raking ratio adjustment procedures can be used to generate a complete distribution of the cross-classification based on the marginal distributions. For example, population weights may be available for age and education, but not their cross-classification. Raking ratio estimation can be used to generate the cross-classification based on a "main effects" model for age and education. The raked cross-classification weights for the population are then applied to the respondent distribution to generate element-level weights as indicated above.

Comparison of Designs

Cost–Variance Tradeoffs

The cost function, together with the variance of the estimator of the population mean given before, can be used to determine the size of the within-100-bank sample size k that will minimize that expected cost for a fixed variance (or minimize the variance for a fixed cost) for the Mitofsky–Waksberg design. An explicit expression for the optimal value of k can be found in [17]. Similarly, the cost function, together with the variance of the estimator of the population mean given before, can be used to determine the sample allocation to the strata that will minimize the expected cost for a fixed variance (or minimize the variance for a fixed cost) for directory-based stratification. Explicit expressions for sample allocation can be found in [3].

Using generally accepted values of cost factors and population parameters for the simple cost models and the variance expressions cited above, Casady & Lepkowski concluded that both the Mitofsky–Waksberg design and directory-based stratification offer considerable improvement over the simple RDD design. They also concluded that on the basis of the simple cost model alone there was little difference in efficiency between the two approaches; however, if the possibility of additional bias could be tolerated, then the truncated design was by far the most efficient.

Implementation Considerations for Telephone Samples

There are a host of features of the telephone system that affect the implementation of the designs described in the preceding section. We discuss here several of the more important ones briefly.

The identification of the residential status of each telephone number generated in RDD or list-assisted samples is not always an easy process. Numbers that are answered must be checked for residential use, and those used for mixed residential and business purposes must be suitably classified (usually any residential use is sufficient to classify a number as residential). Some numbers are readily identified as nonresidential because they are not in service, and a recording clearly indicates that status. Many numbers that are not in service are not connected to a recording to indicate their status, but are connected to a "ringing machine". Thus, interviewers screening telephone numbers to determine residential status cannot distinguish residential numbers where no one is at home from numbers not currently in service.

This latter problem of numbers that repeatedly ring without answer is an important consideration in the implementation of some designs. It is difficult to manage the ring-without-answer numbers in two-stage RDD designs that require the replacement of nonresidential numbers, particularly in time-limited survey data collection periods. Many survey organizations treat ring-without-answer numbers that have been called at varying times of day and days of the week as nonresidential. If the nonresidential classification is made late in the study period, then the replacement number has a relatively short period during which it can be called. Replacements often do not get the same variation in time of day and day of week calling that can be applied to original numbers. Thus, many survey organizations now prefer sampling procedures that give them a fixed sample of telephone numbers rather than one that may generate new telephone numbers late in the survey period.

At the end of the study period, ring-without-answer numbers that have been called repeatedly must be classified as residential or not in order to close out the study. If a number has been called at a variety of times and days, then it may be arbitrarily classified as nonresidential. It thus does

not count against the response rate for the survey because it has been classified as nonsample. On the other hand, ring-without-answer numbers that have not been called enough times are typically classified as residential and nonresponding, leading to a conservative calculation of the response rate.

To overcome these difficulties, and to reduce the costs of screening telephone numbers for residential status, automated screening systems have been developed to identify at a minimum telephone numbers that are connected to recordings indicating whether they are in service. The typical recording is preceded by a "tri-tone" without any ringing of a telephone number. Proprietary hardware and software has been developed which dials telephone numbers and detects the tri-tone recording. Numbers with a tri-tone recording are dropped from further sampling. Numbers without the tri-tone will often have a "ring splash" in which the telephone will ring momentarily while the hardware disconnects the call.

Surveys that are statewide or national in scope have geographic boundaries for the population that correspond to area code boundaries. Sample numbers generated within the sample area codes will be assigned to residences within the target geographic area. Many surveys target geographically defined populations whose boundaries do not match area code and exchange boundaries. In these cases, one may redefine the population, limiting it to that residing in specified exchanges, or one may select a sample from a set of exchanges that covers the entire geographic area but includes areas outside the target. Telephone numbers must then be screened not only for residential status but also for location of residence, based on respondent self-reporting. The classification of ring-without-answer numbers is even more problematic in these screening surveys.

Identification of duplicates in each of the frames also typically involves a respondent self-report. Responding households are asked if they have more than one telephone number assigned to the household, and, if so, the number of such numbers assigned. This self-reported number of telephone numbers through which a household may be reached is subsequently used to generate a weight for estimation. Many survey organizations also check for wrong connections and operator misdialing. Misdialed numbers are discarded, as are wrong connections, to avoid further complications in the weighting process for duplicate listings of a household.

Social science and health surveys (*see* **Surveys, Health and Morbidity**) also frequently select a single eligible person in a household for more interviewing. For example, on a survey involving marital satisfaction, a single adult will be selected to avoid contamination of responses among adults who converse about the content of the survey between interviews. Respondent selection must be done at an early stage in the interview. The procedure for objective respondent selection described by Kish [7] has been widely used in telephone surveys for this purpose, but it leads to an undesirable consequence – increased nonresponse rates. Households are reluctant to participate in a survey when the first questions are designed to obtain a roster of eligible persons living in the household. Alternative methods include a procedure described by Troldahl & Carter [16] and the nearest-birthday method (see [8] for a description). These latter procedures have been shown to be biased, but they continue to be used because they are easy to apply and avoid concerns about increased nonresponse rates.

Finally, answering machines and cellular telephones are posing increasing problems for telephone sampling operations. Answering machines do allow, for the most part, ready identification of residential units. Messages can be left asking that the household call a toll-free number, and calling of households with answering machines can be scheduled at a variety of times of day and days of week to try to reach the household at a time when a person will answer the phone. Cellular telephones pose a different problem. Are such telephone numbers residential or business? Further, the subscriber incurs a charge when they receive such calls. Cellular telephone numbers may be mixed in with other numbers with the same prefix, making identification difficult. Yet they are more readily answered than telephones at a residence. In addition, a well-trained interviewer can make arrangements to call a household at another number, thus reducing the cost to the telephone subscriber.

Bias

A critical issue in the use of the truncated frame is the magnitude of the bias introduced by dropping the low-density stratum. Various studies have shown that an average of less than 5% of the US household population are in the low-density stratum

[1, 4]. Thus it is likely that the additional coverage bias will not be substantial for many characteristics of the total population. Connor & Heeringa [4] show that the coverage bias associated with the truncated frame is negligible for economic attitude measurements. Brick et al. [1] show that the coverage bias for the sociodemographic measure is also small, although for some characteristics and for some subgroups of the population, the additional coverage bias may be large enough to be of concern. Generally, though, the empirical investigations have confirmed the speculation that the additional coverage bias associate with the truncated frame can be safely ignored.

Choice Among Alternative Designs

As indicated previously, the choice among alternative designs is largely based on a consideration of cost and error properties of each design. Typically, three basic cost factors are considered: the cost of generating the sample of telephone numbers, the cost of screening the sample, and the "convenience" of working with the sampling procedure in implementation (a cost consideration that is often difficult to quantify). On the error side, there are two principal concerns: coverage of the telephone household population and sampling variance.

If we examine three main competitors on these characteristics, we can see why organizations are today making particular choices among alternative designs. For example, it is inexpensive to generate telephone numbers in the Mitofsky–Waksberg two-stage RDD sample design. Screening is efficient in the second stage since nearly 65% of the telephone numbers are residential. The Mitofsky–Waksberg design presents a number of difficulties in implementation, including replacement of nonresidential numbers and exhausted clusters. These can be substantial inconveniences for some survey operations, and alternative methods that avoid these problems have substantial attraction. On the error side, the Mitofsky–Waksberg design does provide complete coverage of the telephone household population. Sampling variances are larger than for element sample designs because of the cluster sample selection and well-known increases in variance due to within-cluster homogeneity among sample elements. That is, design effects for Mitofsky–Waksberg samples are greater than one.

The stratified design has a somewhat different set of characteristics. The sample-generation costs can be high. The listed stratum sample can be purchased from a commercial sampling firm at a reasonably low cost per sample number, but the unlisted stratum sample requires further stratification of numbers and two-stage RDD samples drawn from each unlisted stratum. Screening costs are also higher than for the Mitofsky–Waksberg design since approximately 50% of the sample telephone numbers in the listed stratum are residential, and an even lower percentage are residential in the unlisted stratum. Given that different sampling methods are used across strata, sample selection is less convenient for the stratified design than for the Mitofsky–Waksberg design. However, the stratified design does eliminate the need to replace numbers in the listed stratum, and there will be no exhausted clusters in that stratum either. In terms of error, the stratified design does cover the entire telephone household population. Sampling variances will be smaller for the stratified design than for the Mitofsky–Waksberg design because it is element sampling, and some improvements in precision due to stratification can be expected.

The truncated design has the disadvantage relative to the Mitofsky–Waksberg and stratified designs of noncoverage of telephone households in 100-banks with no listed numbers. The level of noncoverage is low, and empirical investigations have shown that the difference for many characteristics between the covered and noncovered populations is small. Samples drawn using the truncated design are inexpensive when obtained from commercial sampling firms. The screening costs of the truncated design are intermediate to those of the Mitofsky–Waksberg and the stratified designs since approximately 50% of the generated telephone numbers will be residential. The truncated design is the most convenient among the three designs considered here since no replacement numbers are needed, and the sample is drawn only from the listed stratum; no two-stage sampling is needed for the unlisted stratum. The sampling variances of estimates should be the smallest for the truncated design since it is a stratified element sample with no cluster sampling.

The sampling practitioner is faced with a choice between designs which provide complete coverage but a number of inconveniences in selection and a design with less complete coverage but a number

of conveniences in selection. Given the empirical evidence on the size of the bias due to the non-coverage of telephone households in 100-banks without listed numbers, current practice favors the latter truncated design. That is, practitioners are choosing truncated sampling methods for telephone surveys based on a classic, although informal, cost–error tradeoff.

References

[1] Brick, J.M., Waksberg, J., Kulp, D. & Starer, A. (1995). Bias in list assisted telephone samples, *Public Opinion Quarterly* **59**, 218–235.

[2] Brunner, J.A. & Brunner, G.A. (1971). Are voluntarily unlisted telephone subscribers really different?, *Journal of Marketing Research* **8**, 121–124.

[3] Casady, R.J. & Lepkowski, J.M. (1993). Stratified telephone sampling designs, *Survey Methodology* **19**, 103–113.

[4] Connor, J. & Heeringa, S. (1992). Evaluation of Two Cost-Efficient RDD Designs. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St Petersburg, May 18, 1992.

[5] Frankel, M.R. & Frankel, L. (1977). Some recent developments in sample survey design, *Journal of Marketing Research* **14**, 280–293.

[6] Groves, R.M. & Lepkowski, J.M. (1985). Dual frame mixed mode survey designs, *Journal of Official Statistics* **1**, 263–286.

[7] Kish, L. (1965). *Survey Sampling*. Wiley, New York.

[8] Lavrakas, P.J. (1987). *Telephone Survey Methods: Sampling, Selection, and Supervision*. Sage, Newbury Park.

[9] Lepkowski, J.M. (1988). Telephone Sampling Methods in the United States, in *Telephone Survey Methodology*, R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, II & J. Waksberg, eds. Wiley, New York, pp. 73–98.

[10] Mitofsky, W. (1970). Sampling of Telephone Households, *CBS News Memorandum*, Unpublished.

[11] Potthoff, R.F. (1987). Some generalizations of the Mitofsky-Waksberg techniques for random digit dialing, *Journal of the American Statistical Association* **82**, 409–418.

[12] Potthoff, R.F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing: some added topics, in *American Statistical Association 1987 Proceedings of Survey Research Methods Section*. American Statistical Association, Alexandria, pp. 615–620.

[13] Stock, J.S. (1962). How to improve samples based on telephone listings, *Journal of Marketing Research* **2**, 50–51.

[14] Sudman, S. (1973). The uses of telephone directories for survey sampling, *Journal of Marketing Research* **10**, 204–207.

[15] Thornberry, O.T. & Massey, J.T. (1988). Trends in U.S. telephone coverage across time and subgroups, in *Telephone Survey Methodology*, R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, II & J. Waksberg, eds. Wiley, New York, pp. 25–50.

[16] Troldahl, V.C. & Carter, R.E., Jr (1964). Random selection of respondents within households in phone surveys, *Journal of Marketing Research* **1**, 71–76.

[17] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **19**, 103–113.

ROBERT J. CASADY & JAMES M. LEPKOWSKI

Teratology

Aristotle used *terata* to mean monsters, which he interpreted as the result of forces upsetting **reproduction** from its normal natural development. Thus teratology (derived from Greek *terata*, meaning monster/prodigy, + logy) became the term used in medicine and biology for the study of monstrosities and abnormal forms in man, animals, or plants as described by Warkany [17].

This article is restricted to the occurrence of these abnormal births in man. Literature on birth defects in laboratory animals exposed to pharmaceutical substances in the search for new treatments is described by Shepard [14] and Schardein [13].

The first use of the term, teratology, is generally attributed to Isidore Geoffroy Saint-Hilaire of Paris, son of Etienne, a distinguished teratologist. Teratology encompasses what are now known to be genetically inherited conditions, developmental conditions, e.g. Down syndrome, which are not inherited by any Mendelian (*see* **Mendel's Laws**) or other mechanisms, and the common birth defects or congenital malformations whose etiology is largely unknown. This very heterogeneous collection of defects affecting the newborn and children comes within the health services remit of the medical or clinical geneticist.

Historical Development

Historically, abnormal births were viewed with a mixture of curiosity and superstition and were generally regarded as a portent of ill luck. The Chaldeans used their occurrence to predict the future, as documented on clay tablets in the Royal Library of Nineveh in the reign of Ashurbanipal, King of Assyria in 700 BC. Likewise, the Romans used monstrous births for divination and also developed the concept of maternal impression whereby mental modification of expectant mothers was thought to influence their offspring. The Spartans passed a law requiring pregnant women to look at statues of Castor and Pollux so that their babies might be born perfect and strong. An ancient Egyptian anencephalic mummy was found in a sepulchre used for animals, monkeys, and sacred ibises, in the catacombs of Hermopolis. In Paris it was unwrapped by Etienne Saint-Hilaire and his assistants who found it not to be a monkey but an eighth month of gestation human fetus with anencephalus. The specimen joined the collection of the King of Prussia in Berlin. During bombing of that city in World War II the museum was hit, the collection destroyed, and this specimen disappeared.

From the time of the Renaissance onwards collections of descriptions of abnormal births were published including a text by the famous French surgeon, Ambroise Paré, an eight-volume treatise in Italian by Taruffi, German texts by Förster and Ahfeld, and Ballantyne's works in English together with the journal *Teratologia* which he set up and edited. By the beginning of this century a large number of birth defects were known, statistical methods of measuring the degree of likeness between relatives by **correlation** (co-relation) analysis had been discovered by **Galton** and the study of Mendelian and other types of inheritance was under way. From the 1950s onwards studies were made of early abortions, human cells and chromosomes, and birth defect registers (*see* **Disease Registers**) and monitoring systems (*see* **Surveillance of Diseases**) were set up.

Types of Study

Teratological investigations in humans can be classified as **descriptive epidemiological** studies, **analytic epidemiological** studies, and case reports (*see* **Case Series, Case Reports**). Descriptive epidemiological studies describe the frequency (usually the **prevalence** at birth) of congenital anomalies in a particular community and how this frequency varies by geographic area, year and month of birth, or with characteristics of person such as socioeconomic status, maternal age, and parity. This type of study usually is based on information available from existing sources such as birth certificates, **death certificates**, or hospital records.

Broadly classified, analytic epidemiologic studies include case–control and cohort studies, and clinical trials. In **case–control studies** in the field of teratology, a group of index cases (births/fetuses with anomalies, other adverse reproductive outcomes such as miscarriages) is identified and information is sought about prior exposures, often during a reference period such as periconception. A **control** series is identified on whom similar information is obtained. The purpose of the control series is to

provide information on the distribution of exposure in the population at risk of the adverse reproductive outcome under consideration. Information on the distribution of exposure is compared between cases and controls, and a measure of association, the **odds ratio**, may be calculated which closely approximates the **relative risk**. Important issues in evaluating the validity of data from case–control studies include the appropriateness of the control group or control groups chosen, and the similarity in the degree of accuracy of the data on exposures for cases and controls.

In a **cohort study** as applied to teratology, a group of women is classified in terms of exposure status at some defined point in time prior to the occurrence of the outcome of interest and are followed up to determine reproductive outcome. Exposure status may refer to the fact of exposure vs. nonexposure, e.g. work with video display units during pregnancy or with the husband or partner working in a specific occupation during pregnancy, or it may relate to degree of exposure, for example the reported intake of specific dietary factors per week. The frequency of adverse reproductive outcomes is compared between the groups, and on this basis the relative risk associated with the exposure may be calculated. Fewer cohort studies than case–control studies have been carried out because, for comparable statistical **power**, substantially greater numbers of subjects need to be studied. Another potentially important problem of cohort studies in this context is that ascertainment may be influenced by knowledge of the exposure status of the mother of the infant being examined. Studies of familial aggregation may be classified as a special type of cohort study in which the "exposure" is having an affected relative. Special techniques of analysis have been applied in some of these studies, such as **segregation analysis** and **linkage analysis**.

Another special type of cohort study is the **clinical trial**. In trials, assignment to exposure is determined by the investigator. The randomized control trial is the definitive method of evaluating interventions. If the trial is of adequate size, then the **randomization** ensures comparability of the group assigned to receive the intervention and the control group for potentially **confounding** factors, both known and unknown. A further refinement is to make both the woman receiving the intervention and those responsible for her care and that of the child unaware of whether or not she has received the intervention. This minimizes the chance that the women will change their behavior in a way that is related to the intervention and also the possible effect of knowledge of exposure status on the ascertainment of reproductive outcome (*see* **Blinding or Masking**). The only randomized trials carried out to date in the field of teratology have related to vitamin supplementation during the periconceptional period. **Nonrandomized trials** have been carried out regarding the possible preventive effect of multivitamin supplementation against recurrent NTDs and orofacial clefts.

In teratology, the number of analytic epidemiologic studies carried out is small compared with the fields of cancer or heart disease in adults.

Classification and Frequency

Classification and how this relates to morphology is fundamental to any understanding of teratology. However, this subject is complex. Unlike plant taxonomy, no Linnaeus has appeared to embrace the whole field of diverse birth defects and classify them according to any universally agreed order. The modern-day equivalents of the collections of the last century are McKusick's catalog [10] and the *Birth Defects Encyclopedia* [2]. The former is mainly concerned with genetically inherited conditions although it does include a large number of birth defects and associated syndromes. Other useful reference sources are those made by the March of Dimes Birth Defects Foundation in the US and the International Clearinghouse for Birth Defects Monitoring Systems [8].

The usual definition of an **incidence rate** is the number of occurrences of a disease that manifest in a unit of time in a known population of individuals at risk. This definition is not easy to apply to congenital anomalies, as it implies a process occurring regularly over the period of time chosen. In embryonic development, the frequency of an event in one week is unlikely to be the same as in the next week. It is more natural to think of prevalence, that is the proportion of living embryos with the condition, at a point in time or at the time of an event, such as birth. Prevalence is more simple because the denominator changes rapidly as a result of pregnancy loss, and because many congenital anomalies represent not so much something that occurs, but something that does not occur, such as fusion of the palatal shelves (cleft palate) or neural tube closure (NTDs). If it were possible to obtain complete information, then the

progress over time of a cohort of embryos could be followed, documenting depletion as a result of miscarriage and the events relating to organogenesis in the survivors.

The **cumulative incidence rate** is the total frequency of an event in a cohort of at-risk subjects up to a relevant time. There are theoretical and practical problems of estimating both the numerator and the denominator of this incidence rate. Theoretical problems concern the manifestation of anomalies throughout development. There is evidence that the mammalian embryo is capable of regeneration and of repairing itself. If these findings were applicable to man, then an apparently normal child may have shown defects at an early stage of intrauterine development, leading to underestimation of cumulative incidence. The practical problems of estimating incidence rates include pregnancy termination, undetected abortion, and problems of ascertainment. Antenatal diagnosis with selective termination of pregnancy has had a substantial impact on the numbers of births with certain types of congenital anomalies diagnosed at the time of delivery. Only the cases from pregnancies which went to term are eligible to be notified to the national birth defects monitoring scheme. Failure to include fetuses from terminated pregnancies may greatly distort the epidemiologic information, so it is now accepted practice to include fetuses with anomalies detected by antenatal diagnosis with subsequent termination of pregnancy in the estimation of the "prevalence at birth" of congenital anomalies.

Landmark Studies

Frequency and Types of Birth Defects

One of the first studies to address this lack of information about the frequency and types of birth defects throughout the world was that by the **World Health Organization** (WHO) in 1958 and later discussed at an informal meeting at Ann Arbor, Michigan, in April 1959. Stevenson and colleagues conducted a WHO supported prospective study of births in 24 centers in 16 countries and described the occurrence and types of birth defects found in stillborn and liveborn infants [16]. Outcomes relating to 421 781 pregnancies were traced based on 416 695 single births, 5022 sets of twins, 63 sets of triplets, and one set of quadruplets. A 400 page book containing basic tables for each center was published. The group recognized biases in the data, particularly in the centers which recorded hospital births only. This research demonstrated the large impact of NTDs on fetal wastage and a **correlation** between these defects and dizygous twinning (*see* **Zygosity Determination**). Consanguinity between parents increased stillbirth rates and early **infant mortality** rates, these being highest where parents are most closely related.

Vitamin A Deficiency

Why do birth defects occur? For the majority and wide spectrum of abnormalities observed, environmental factors have been identified in only a few instances. Hale observed a Duroc–Jersey sow, who received a ration deficient in vitamin A, at the Texas Agricultural Experimental Station. She gave birth on March 29, 1932, to 11 pigs, all of which were born without eyeballs [7]. Ten were alive at birth, one lived 4 days, one lived 3 hours, while all the others died within 5 minutes after birth. He postulated that the abnormalities were caused by vitamin A deficiency. The study marked the beginning of experimental teratology and led to many future studies using animal models for the experimental investigation of birth defects. The pharmaceutical industry built on this work extensively to investigate nutritional imbalances, drugs, chemicals, irradiation, and many other postulated teratogenic substances. Some 60 years later the Chief Medical Officer of the UK issued a warning that women who were pregnant or might become pregnant must not take excessive quantities of vitamin A. While this message relates to overprovision rather than underprovision, the origin of this work goes back to Hale's key paper.

German Measles

In the first 6 months of 1941 in Sydney, Australia, there were an unusual number of cases of congenital cataract. Gregg, an ophthalmologist, personally saw 13 cases [6]. He noted that as well as these babies being born with bilateral cataract, they were of small size, ill-nourished, and difficult to feed. Many had congenital heart defects. There were a few with monocular cataract which in two-thirds of cases was associated with microphthalmia. Close questioning of the mothers revealed that they had German measles

during early pregnancy. There had been a widespread and severe epidemic in Australia in 1940. Altogether 78 children with congenital cataract were ascertained and of these 68 were associated with a definite history of maternal rubella infection.

Thalidomide

At a meeting of the German Paediatric Society at Kassel in October, 1960, Kosenow and Pfeiffer presented photographs and X-rays of two infants with aplasia of the extremities and various other defects. Within a few months another doctor, Wiedemann, reported similar cases. At the Paediatric Society meeting on November 18, 1961, Lenz suggested that the drug, thalidomide, might be the cause, as it appeared in 17 out of the 20 maternal records that he had investigated [9]. The manufacturers withdrew thalidomide (trade names Contergan, Distaval, Softenon) and all other preparations containing this substance from the market on November 25, 1961. However, the damage was done. Some 129 cases were studied by Lenz at the University of Hamburg. Another 203 cases were reported to him by letter from the rest of West Germany and also isolated cases from Belgium, Brazil, England, Egypt, Israel, Sweden, Switzerland, and a few cases from the US. The much more strict enforcement of food and drugs legislation in the US (*see* **Drug Approval and Regulation**) ensured that many fewer births were affected in North America than in Europe.

Thalidomide is a derivative of glutamic acid, discovered in Germany, and first marketed in 1956. It was well tolerated, considered safe, and used as an analgesic, sedative, and hypnotic agent. By the time of the papers by Lenz & Knapp in 1962 [9], at least 2000 children with drug-induced abnormalities had been born in West Germany. Worldwide it is thought at least 8000 children had been affected by thalidomide.

Identification of the Relationship between Diet and Neural Tube Defects (NTDs) as a Paradigm of Teratological Investigation

Several features of the descriptive epidemiology of NTDs led to a dietary hypothesis for their etiology. Details are contained in [5]. There was an increased prevalence at birth of NTDs in the offspring of women of lower socioeconomic status compared with the offspring of other women. In the British Isles, and some other areas, the highest rates of anencephalus were amongst babies conceived in the spring and early summer, possibly linked to a lack of fresh vegetables in the winter. Body stores of certain nutrients are low in the spring. Improved all-year availability of various nutrients might in part explain the recent changes in seasonal pattern. It was known from the 1950s that therapeutic abortions could be induced by giving a folic acid antagonist 4-aminopteroylglutamic acid taken orally. This was interpreted to indicate that folic acid deficiency could induce abortion and possibly malformations.

A Case–Control Study

In western Australia, Bower & Stanley, using notifications to the local malformation registry in the period 1982–84 ascertained 77 infants with NTDs [1]. They were compared with two control groups each matched by date of last menstrual period: a group of 77 infants with other malformations registered in the same way, and a group of 154 normal infants. A telephone interview was conducted to obtain details of demographic and other factors. A three-part questionnaire was mailed to each mother, comprising a section on food frequencies (*see* **Nutritional Exposure Measures**) during the period from 9 months before to 9 months after the last menstrual period, a section on illness and drugs taken for nausea in pregnancy, cooking methods, changes in diet and other factors over the same period, and a 24-hour dietary record to be completed on a specified day after receipt. These data were used to assess the daily dietary intake of folate, including that provided by supplementation, and intakes of a number of other nutrients. Participation rates were very high: 93% for mothers of cases, 88% for mothers of infants with other malformations, and 84% for mothers of normal infants. A statistically significant association of reduced NTD risks with increased reported intake of total folate was observed.

A Cohort Study

Milunsky et al. [12] reported a cohort study based on information collected on women undergoing prenatal testing by maternal serum α-fetoprotein screening or amniocentesis in Massachusetts. Information on diet in the first 8 weeks of pregnancy was obtained using

a food frequency questionnaire, and on vitamin supplementation in the 3 months before and 3 months after conception, was collected by telephone interview. A total of 22 715 study subjects was available for analysis, of whom 49 had an infant with a neural tube defect. There was a statistically significant protective effect of taking multivitamins at least once a week before conception and during the first trimester, with a relative risk of 0.36 (95% **confidence interval** 0.15–0.83). The effect was almost entirely restricted to preparations containing folic acid. The relative risk associated with folic acid supplementation in the first 6 weeks of pregnancy was 0.29 (95% confidence interval 0.15–0.55). Amongst women who did not use supplements containing folic acid, the relative risk of NTDs associated with a dietary intake of more than 100 micrograms daily was 0.42 (95% confidence interval 0.16–1.15).

A Nonrandomized Clinical Trial

Smithells and colleagues carried out a multicenter nonrandomized prospective trial of periconceptional multivitamin supplementation in the prevention of NTDs [15]. They selected mothers who already had had one affected offspring. The oral tablets taken (Pregnavite Forte-F, made by Bencard) consisted of a mixture of folic acid, numerous vitamins, and a mineral supplement containing iron, calcium, and phosphorus. There were three groups. The fully supplemented group of 185 mothers, who took one tablet three times a day for at least 28 days prior to conception until the date of the second missed period, produced 178 infants or fetuses, of whom one (0.6%) had an NTD. This compared with 13 births with an NTD (5.0%) out of 260 infants or fetuses of unsupplemented mothers; a significant difference (relative risk 0.12, $p < 0.01$). There was also a third group of partially supplemented mothers defined as those conceiving within 28 days of beginning supplementation or commencing supplementation after conception but known to have missed tablets for more than one day. These results produced immediate and widespread interest.

An early and sustained criticism was why a randomized design had not been used. Smithells et al. eventually stated that they originally intended to use a double-blind randomized design (*see* **Blinding or Masking**; **Clinical Trials, Overview**) but that this protocol was rejected by three separate hospital research ethics committees. Owing to the absence of **randomization**, women therefore selected themselves into the supplemented and nonsupplemented groups (*see* **Selection Bias**). There were other criticisms. There was no true placebo group; which preparation might be effective, multivitamin or folic acid, was not clear.

A Randomized Clinical Trial

The debate for and against the need and ethical justification for a larger and valid randomized trial of folic acid supplementation continued (*see* **Ethics of Randomized Trials**). By September 1982 the UK Government had approved funding for 3 years and asked the Medical Research Council to carry out such a study.

The MRC vitamin study [11] was a randomized double-blind **prevention trial** with a **factorial design** conducted at 33 centers in seven countries to determine whether supplementation with folic acid or a mixture of seven other vitamins (A, D, B1, B2, B6, C, and nicotinamide) around the time of conception can prevent NTDs (anencephalus, spina bifida, encephalocele). Some 1817 women who had at least one previous affected pregnancy with an NTD were allocated at random to one of four groups (see Table 1).

Some 1195 had a completed pregnancy in which the status of the fetus or infant was ascertained with respect to having or not having an NTD. A known abnormality occurred in 27: six in the folic acid groups and 21 in the other two groups. This is a 72% protective effect (relative risk 0.28, 95% confidence interval 0.12–0.71). No significant protection effect was found with respect to the

Table 1 The 2×2 factorial design used in the UK Medical Research Council randomized trial of folic acid and other dietary supplementation in high-risk pregnancies

Group		Folic acid	
		Yes	No
Multivitamins	No	A	C
	Yes	B	D

Note: all women had a previous offspring with NTD and all received minerals supplement.

other vitamins (relative risk 0.80, 95% confidence interval 0.32–1.72). Of the centers, 17 were in the UK, seven were in Hungary, three in Australia, three in Canada, and one each in Israel, Moscow, USSR, and Lyon, France. The randomization was carried out by the Clinical Trials Service Unit at Oxford. Women took one capsule a day from the date of randomization until 12 weeks of pregnancy, estimated from the first day of the last menstrual period. The folic acid capsules contained 4 mg of the substance. The multivitamin groups contained various amounts of the various vitamins. The control substance consisting of minerals contained dried ferrous sulfate and calcium phosphate. There was an independent data monitoring committee (*see* **Data Monitoring Committees**) which reviewed progress every 6 months. The trial was double-blind with neither patients nor their medical attendants knowing which regime had been allocated.

There was a 2×2 factorial design:

A Folic acid and minerals.
B Folic acid, minerals and other vitamins.
C Minerals only.
D Minerals and other vitamins.

In this design the effect of folic acid is determined by comparing groups A and B together vs. groups C and D together. The effect of multivitamins is determined by comparing groups B and D together vs. groups A and C together. If it is thought that there might be a synergistic action between folic acid and multivitamins then only group B would benefit. This factorial design with 2000 participants would give a statistical power of 80% to detect a reduction in recurrence from 4% to 2% using a one-sided significance level of 0.05 (*see* **Alternative Hypothesis**). This degree of difference is similar to that observed by Smithells et al., who reported a reduction of recurrence risk from 4.7% to 0.7%. About one woman was being recruited each working day and it was thought that the trial would continue until around 1993. However, in April 1991 the data monitoring committee recommended that the trial be stopped. The results of the trial had been kept under review and assessed by **sequential analysis** for which the cumulative difference between the number of NTDs occurring in the folic acid and nonfolic acid groups was plotted against the total number of NTDs occurring in the study.

By April 12, 1991, results showed that this difference had passed the preset lower boundary in the sequential analysis. At this point 1817 women had been randomized and findings were known on 1195 informative pregnancies. The findings were very clear and showed a significant beneficial effect specific to folic acid supplementation. This important trial was based on measuring the reduction in recurrence risk. Roughly, only 5% of mothers who produce offspring with NTDs have previously had at least one affected birth. The great majority, about 95% of mothers who produce NTD births have only one such occurrence.

This key question of whether a periconceptional vitamin supplementation with folic acid or multivitamins would prevent the first occurrence of NTDs in a similar fashion was studied by Czeizel & Dudás in Hungary and reported in 1992 [3]. They randomized women in two groups. Some 2104 women received the folic acid and vitamin supplement and 2052 women received a trace element supplement which contained copper, manganese, zinc, and vitamin C. Birth defects were significantly more prevalent in the group receiving the trace element supplement compared with the folic acid and vitamin supplement group (22.9 per 1000 and 13.3 per 1000 births, respectively; $P = 0.02$).

The mechanism whereby folic acid supplementation in the periconceptional period reduced the occurrence and recurrence risk of NTDs has not been determined. It is clear that supplementation does not act to correct a simple nutritional deficiency, because most pregnant women carrying an affected fetus have levels of folate above the deficient range. Therefore, the possibility that an abnormality in folate metabolism is responsible for a large proportion of NTDs is being investigated. Functional variants in some of the enzymes involved in folate metabolism can be identified by demonstrating differences in the genes encoding them. Three studies have suggested that homozygosity (*see* **Heterozygosity**) for the V677T mutation in the gene coding for 5,10-methylenetetrahydrofolate reductase (MTHFR), which results in a thermolabile variant of the MTHFR enzyme, is a risk factor for spina bifida.

Particular Statistical Concepts, Problems, and Techniques

It is useful to compare progress in teratology with research in another field such as cancer. In the latter,

descriptive studies have led to case–control studies of each of the common cancers, possible treatments have been identified and tested using clinical trials. In teratology these developments have not happened. There is the difficulty of direct observation, and the major problem of assessing exposure. However, compared with cancer, the interval between exposure and outcome in a fetus or newborn is relatively short and new cohorts of births are formed by the natural process of conception, gestation, and birth.

Another key problem relates to the frequency of birth defects. For example a case–control study would need at least 750 cases and a similar number of controls to detect an exposure factor which changed the incidence of any of the most common birth defects, e.g. NTDs, by a factor of 50% (*see* **Sample Size Determination**). In turn this would require a population of around 750 000 births corresponding to a total population of at least 50 million persons. This corresponds to an entire year of births in France or England and Wales and at least 2 years of births in Australia or Canada. In the MRC trial of folic acid we have noted that this required collaboration between 33 centers drawn from 14 countries. It follows that there is a great need for international collaboration.

Fundamental problems exist relating to classification (*see* **Classification, Overview**). The debate is summarized by the phrase "lumpers and splitters". Birth defects rarely, if ever, occur as an all or nothing phenomenon. Taxonomy is crucial. Further difficulties arise due to the observation of more than one defect occurring in the same individual and the large number of complex syndromes that have now been described. A systematic study of each of the major common birth defects using rigorous case–control studies has not yet been completed. Once the selection of subjects for study has been made there may be errors and **bias** in observations (*see* **Measurement Error in Epidemiologic Studies**). One of the most difficult to assess is logical and not statistical, and relates to the definition of affected individuals. Errors of measurement may occur. Bias is common, most frequently due to incomplete ascertainment.

Solutions to These Problems

Classification problems may be clarified by new and better observations and techniques. Detailed biochemical studies and possible genetic defects have been identified using DNA methods (*see* **DNA Sequences**) in the case of NTDs. The human genome project will undoubtedly provide a great deal of knowledge relating to teratology. Progress could be made by extending existing malformation registers. These should contain more accurate observations, preferably verified by experienced clinicians, more complete ascertainment of birth defects in a defined community, and in due course larger data sets. Analysis specifically of cases with malformations of more than one system may have greater statistical power to detect associations with putative teratogens if the exposure of interest is primarily associated with a pattern of multiple defects of unknown cause rather than with isolated defects.

Monitoring

Several surveillance systems for births with malformations were established after the thalidomide tragedy of 1961. Tests for statistically significant increases in observed numbers of cases as compared with baseline expected numbers are made on a monthly basis in England and Wales and Norway, on a 1, 2, 3, 6 and 12 monthly basis in Atlanta, and on a quarterly basis in the Birth Defects Monitoring Program in the US. Many of the surveillance schemes participate in the International Clearinghouse for Birth Defects Monitoring Systems. Most of these are based solely on malformations recorded in the neonatal period. There are also systems in which malformations are recorded irrespective of the age at detection. In Europe, a number of this type are included in the EUROCAT Project, a concerted action project on registration of congenital abnormalities and twins (*see* **Twin Registers**) in the European Community.

The statistical methods used have been of the following types: (i) **graphical display** of frequencies; (ii) chi-square linear trend analysis (*see* **Trend Test for Counts and Proportions**); (iii) comparison of observed numbers of cases of specific types of congenital anomalies with the numbers expected according to the **Poisson distribution**; (iv) "self-reinforcing" techniques, notably the cumulative sum ("cusum") technique and the sets method; and (v) scan analysis (*see* **Scan Statistics for Disease Surveillance**). Some comparisons of these techniques have been carried out. In applications with small denominator populations, the Poisson technique was

found to be inefficient compared with the sets method and the cusum technique. In applications with larger denominator populations, the cusum technique is somewhat more efficient than the Poisson and sets techniques. In **simulation** analysis for four specific groups of malformations, the cusum technique showed greater **sensitivity**, **specificity**, and accuracy than the sets method. One obvious limitation of these techniques, which applies to certain types of malformations only, is the difficulty of obtaining data on terminations of pregnancy in which fetuses with certain types of anomalies have been detected by antenatal diagnosis. The size of the population surveyed is a crucial issue in view of the fact that most teratogens identified to date have had a low prevalence of exposure, and the background rate of specific anomalies is low.

Clustering in Time and Space

The recognition and investigation of clusters of congenital anomalies, and particularly the assessment of whether the occurrence of **clustering** or a particular cluster is purely a chance phenomenon, leads to considerable methodological and practical difficulties. The available studies can be classified into one of three groups based on the motivation for the investigation: (i) studies carried out to test if clustering exists within a predefined population of births, in the absence of specific spontaneous reporting of clusters or a specific hypothesis concerning an environmental agent; (ii) studies carried out when the existence of the cluster has been suspected and spontaneously reported, usually either by clinicians or local inhabitants; and (iii) studies initiated because of concern about specific environmental exposures. In the 1970s, recognition of the problems of investigating spatial clustering led to investigation being focused more on the identification of time space clustering using the methods of Ederer et al. [4]. All of these approaches have been used in investigations of neural tube defects; no clear-cut evidence of clustering was found in any study.

Anticipated Developments and Unresolved Problems

Progress in teratology, as in other fields, requires accurate and perceptive observation which in turn relates to the power and sophistication of techniques and instruments available at the time. It is likely that this will be revolutionized by the new work in genome research, the identification of chromosomal differences between mothers of affected and nonaffected offspring, and in due course by the detection of how these differences are manifest via biochemical pathways or other means. The definition and identification of clusters of birth defects is unresolved and continues to be of concern. Such outbreaks are notified from time to time usually accompanied by requests for investigation. These are difficult, expensive and time-consuming and the majority so far have rarely led to finding any new cause.

The etiology of most types of congenital anomalies remains unknown. For many types of anomaly there is an elevated recurrence risk, probably due to a combination of genetic and environmental factors (*see* **Gene-environment Interaction**). Advances in molecular genetics have made it possible to **genotype** large numbers of individuals by polymerase chain reaction (PCR) techniques. This greatly simplifies the investigation of genotype–environment interaction. For example, research has identified an increased risk of orofacial clefts among individuals with the uncommon allele for transforming growth factor alpha (TGFα).

Progress is being made at finding reasons why some rare defects occur, but this has had a small effect to date on the overall burden of the common teratological defects (*see* **Burden of Disease**), which remain a major public health problem.

References

[1] Bower, C. & Stanley, F.J. (1989). Dietary folate as a risk factor for neural-tube defects: evidence from a case-control study in Western Australia, *Medical Journal of Australia* **150**, 613–619.

[2] Buyse, M.L., ed. in chief. (1991). *Birth Defects Encyclopedia*. The Centre for Birth Defects Information Services, Inc., Dover, Mass., and Blackwell Scientific, Oxford. (A reference source describing known defects in detail.)

[3] Czeizel, A.E. & Dudás, I. (1992). Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation, *New England Journal of Medicine* **327**, 1832–1835.

[4] Ederer, H.B., Myers, M.H. & Mantel, N. (1964). A statistical problem in time and space: do leukaemia cases come in clusters?, *Biometrics* **20**, 626–638.

[5] Elwood, J.M., Little, J. & Elwood, J.H. (1992). *Epidemiology and Control of Neural Tube Defects*. Oxford University Press, Oxford. (A monograph detailing research and methodological issues relating to these malformations.)

[6] Gregg, N.M. (1941). Congenital cataract following German measles in the mother, *Transactions of the Ophthalmological Society of Australia* **3**, 35–46.

[7] Hale, F. (1933). Pigs born without eye balls, *Journal of Heredity* **24**, 105–106.

[8] International Clearinghouse for Birth Defects Monitoring Systems (1991). *Congenital Malformations Worldwide*. Elsevier, Amsterdam. (A description of the work of this nongovernmental organization of the World Health Organization and the participating centers who joined during 1974–88.)

[9] Lenz, W. & Knapp, K. (1962). Foetal malformations due to Thalidomide, *German Medical Monthly (English language edition of the Deutsche Medizinische Wochenschrift)* **7**, 253–358.

[10] McKusick, V.A., with the assistance of Francomano, C.A., Antonarakis, S.E. & Pearson P.L. (1994). *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*, 11th Ed. 2 Vols. Johns Hopkins University Press, Baltimore. (A comprehensive computerized listing of all reports in the literature relating to known and possible genetic diseases and syndromes, including many congenital malformations. Each brief entry has an identification number, description, author, and literature citation. The catalog is continuously updated and republished periodically.)

[11] Medical Research Council (MRC) Vitamin Study Research Group (1991). Prevention of neural tube defects: Results of the Medical Research Council vitamin study, *Lancet* **338**, 131–137.

[12] Milunsky, A., Jick, S.S., Bruell, C.L., MacLaughlin, D.S., Rothman, K.J. & Willett, W. (1989). Multivitamin/folic acid supplementation in early pregnancy reduces the prevalence of neural tube defects, *Journal of the American Medical Association* **262**, 2847–2852.

[13] Schardein, J.L. (1993). *Chemically Induced Birth Defects*, 2nd Ed. Marcel Dekker, New York. (A catalog relating to these teratogenic agents.)

[14] Shepard, T.H. (1995). *Catalog of Teratogenic Agents*, 8th Ed. Johns Hopkins University Press, Baltimore. (A useful computer-maintained reference file relating to this aspect of birth defect causation.)

[15] Smithells, R.W., Sheppard, S., Schorah, C.J., Seller, M.J., Nevin, N.C., Harris, R., Read, A.P. & Fielding, D.W. (1980). Possible prevention of neural-tube defects by periconceptional vitamin supplementation, *Lancet* **i**, 339–340.

[16] Stevenson, A.C., Johnston, H.A., Stewart, M.I.P. & Golding, D.R. (1966). Congenital malformations: a report of a study of series of consecutive births in 24 centres, *Bulletin of the World Health Organization* **34**, Supplement, 9–127.

[17] Warkany, J. (1971). *Congenital Malformations. Notes and Comments*. Year Book Medical Publishers, Chicago. (A large wide-ranging text based on a lifetime experience of this pediatrician and teratologist based at Cincinnati.)

J.H. ELWOOD & J. LITTLE

Textbooks in Clinical Trials

Introduction

As the number of drug, biologic, and device **clinical trials** increases at an ever-expanding rate, the number of texts discussing clinical trials seems to grow just as quickly. And just as there are a wide variety of clinical trials and aspects of clinical trials (e.g. protocol, design, monitoring, data management, regulatory, statistical analysis, and reporting results), so are there numerous texts covering these different types and aspects. Further, within each clinical trial aspect (e.g. statistical analysis), there are still a wide variety of texts. For example, some statistical texts may give an overview on how to apply standard statistical techniques to clinical trials data, while others may focus on specific topics such as analyzing **crossover** trials, **Bayesian methods** or **generalized estimating equation** (GEE) techniques. Overall, in order for anyone working with or interested in clinical trials to choose a reference text best suited to one's needs, one must first decide what one's needs are in this field. Is one new to the clinical trials field and thus interested in getting a general overview of clinical trials? Is one interested in regulatory aspects? (*see* **Drug Approval and Regulation**; Data management) Is one an entry-level **statistician in a pharmaceutical company** and wants to learn how standard statistical techniques are used to analyze clinical trials data? Is one a veteran of statistical analysis of clinical trials, but wants to learn more about a specific specialized topic?

In addition to the large amount of choices an individual has in choosing appropriate clinical trials texts, another consequence of the clinical trials boom is that more and more academic institutions are offering courses or entire programs in clinical trials. The variation in these courses mimics the variation in the types and aspects of clinical trials. There is not necessarily one course that could be considered a standard clinical trials course. As with texts, some courses may focus on introduction, conduct, ethics, components, and/or design of a clinical trial; other courses may focus on analysis. Within the different types of courses, there is not necessarily a standard syllabus that could be developed. For example, analysis courses have the following (not necessarily exhaustive) list of topics from which to choose, all of which could not be covered in a standard one semester course: Statisticians role in data management; **randomization**; **power** and **sample size** calculations; reporting results; NDA/PLA/PMA writing; application of the more common statistical techniques (**ANOVA**, **ANCOVA**, **Chi-Square**, **Mantel–Haenszel** Statistics, **Logistic Regression**, **Survival Analysis**, Repeated Measures (*see* **Longitudinal Data Analysis, Overview**)) to clinical trials; handling treatment-by-center and **treatment-by-covariate interactions**; **confounding**; the application of more complex yet now widely used statistical techniques of mixed models, GEE techniques, and **missing data** techniques (especially for dropouts); interim analyses (*see* **Data and Safety Monitoring**), group sequential methods, and **data safety monitoring boards**; crossover trials; noninferiority trials (*see* **Equivalence Trials**); analysis of safety data; **intent-to-treat** versus per-protocol analysis; **multiplicity** (multiple endpoints; multiple treatment groups; multiple looks at the data); and Bayesian methods.

Clearly, as it would not be possible for a course to cover these topics in one or even two semesters, the organizers and instructors of clinical trials courses have the not-too-easy task of first choosing the appropriate topics to cover in one course or in a set of courses, followed by the task of choosing the appropriate text to use as a reference or as a textbook. Given the wide variety of texts and courses, an instructor may have a difficult time finding a book that would be a one-to-one correspondence with the topics he or she desires to cover in a given course.

In general, whether you are a student, clinical trials professional, or an instructor, choosing references or course textbooks can be overwhelming, even after deciding upon a specific learning objective. In this article, we hope to offer guidance on appropriate textbooks or references for those interested in (a) an introduction or overview of the design and components of clinical trials; and (b) statistical analysis of clinical trial data. What follows is not a critical review of texts, since the number of texts is too numerous for such handling here. However, if this author knows of a review that does exist in journals, it is noted. Rather, what follows is a brief description of various texts discussing clinical trials and the audience for which the text is meant. Also presented is this author's judgment on whether each text might be more useful as a textbook or a reference for a clinical trial statistics course or for a clinical

trial introduction/overview/design course. Note that, again, this article focuses on various texts giving an introduction or overview of clinical trials, and on those texts focusing on analysis. The numerous other texts focusing on issues such as regulatory aspects or data management, or focusing on clinical trials in certain indications (e.g. **AIDS**, cancer) are not included here. Also, given the ever-increasing advances in the conduct and analysis of clinical trials, the concentration of this article is on clinical trials texts that have been developed or updated within approximately the last 10 years (however, a discussion of some of the books generally considered as "classics" by the statistical and clinical trial community that do not necessarily satisfy the time criterion is also given).

Discussion of Clinical Trials Texts

Table 1 gives an outline of several available clinical trials text focusing on (a) an introduction or general overviews of clinical trials, or (b) the general statistical analysis. Any text not included here must *not* be considered a reflection on the quality of the omitted text, but is rather due to the fact that the book was published after to the finalization of this report (i.e. through beginning of 2004), is not focused on design or general statistics in clinical trials in this author's opinion, was not published within approximately the last 10 years, or, despite efforts, is an oversight of the author. The table is sorted by the year of publication, with the more recent texts being listed first. Included in this table for each text is the title, author, year of publication and publisher. Also included for each text is an assessment of the level of the text (either graduate level or undergraduate level), highlights of the text's contents, this author's assessment of whether the text is more useful as a reference or textbook, the type of course where the text is most useful (categorized into two classes: introduction/overview/design, or statistics), and references for a detailed review of text, if a review is available. Note that for virtually all texts listed below, the intended audience includes both statistical and clinical researchers, where "clinical researcher" includes any nonstatistician involved in clinical trials, such as investigators and clinical trial monitors.

Note that any text listed in the table as a possibility for being used as a textbook in a clinical trials course does not necessarily imply the text has practice problems and exercises. Rather, it indicates that, in this author's (and only this author's) opinion, the text covers a relatively wide range of topics in detail, yielding sufficient quantity and quality material on which a course could be based. Unless otherwise noted, texts indicated as possible use in a clinical trial statistics course are most useful if the student has already had a one-year elementary statistics or biostatistics course. Texts indicated as possible use in as introduction/overview/ design of clinical trials often do not necessarily require such a prerequisite (nor do they usually require a medical background).

In addition to the texts listed in the table, there are numerous texts that focus on a specific statistical aspect of clinical trials (e.g. sample size; crossover designs; group sequential methods; sample size). Many of these texts are excellent references, or could be considered as textbooks for courses designed in such a specific topic, and are listed below:

Sample Size Calculation
Sample Size Calculation in Clinical Research by Chow S-C, Shao, J, Wang H (2003).
Publisher: Marcel-Dekker (This is a complete overview of sample size calculation methods for a wide variety of clinical trial scenarios; includes a discussion of sample size re-estimation in interim analyses).

Randomization
Randomization in Clinical Trials: Theory and Practice by Rosenberger WF, Lachin JM (2002).
Publisher: Wiley.

Safety
Drug Safety Evaluation by Gad SC (2002).
Publisher: Wiley.
Drug Safety Assessment in Clinical Trials by G.S. Gilbert (1993).
Publisher: Marcel-Dekker.
Review in: Biometrics (1994); 50:1231–1232;
JRSS-A (1994), 157:503.

Cross-over Trials
Design and Analysis of Cross-Over Trials, Second Edition by Kenward MG and Jones B (2003).
Publisher: Chapman & Hall/CRF.
Review in: Statistics in Medicine (1990); 9:1007;
Biometrics (1991); 47:787;
JASA (1991); 86:232.

(*this list is continued after Table 1*)

Table 1 Summary Review of Texts in Clinical Trials

Title	Author(s)	Year of publication and publisher	Level/ Audience	Concentration/ Content highlights	Possible course use – type of course/textbook or reference
Statistical Aspects of the Design and Analysis of Clinical Trials (Revised Edition) Review of first edition in: Transactions of Royal Society of Tropical Medicine and Hygiene, 2001	Everitt, B.S. and Pickles, A.	2004, World Scientific Publishing Co., Inc.	Graduate	*Statistics Concentration:* Introduction to clinical trials; randomization; sample size; reporting; monitoring (patient adherence; dropouts; interim analysis); standard statistical and analytic concepts (general linear model, survival analysis); advanced concepts (analyzing normal and nonnormal longitudinal data; Bayesian analysis; meta-analysis)	Statistics – Textbook
Advances in Clinical Trial Biostatistics	Geller, N.L.	2003, Marcel Dekker	Graduate	*Statistics Concentration:* Bayesian methods for Phase I cancer trials; equivalence trials; multiple endpoints; adaptive two-stage trials; missing data; subgroups and interaction	Statistics – Reference
Encyclopedia of Biopharmaceutical Statistics – Second Edition Review of first edition in: The Statistician (2001). 50:97	Chow, S-C. – editor	2003, Marcel Dekker	Graduate	*Statistics concentration:* Expert authors write on over 75 topics	Statistics – Reference

(*continued overleaf*)

Table 1 (*continued*)

Title	Author(s)	Year of publication and publisher	Level/ Audience	Concentration/ Content highlights	Possible course use – type of course/textbook or reference
Design and Analysis of Clinical Trials: Concept and Methodologies – Second Edition	Chow, S-C. and Lui, J.P.	2003, Wiley	Graduate	*Statistics concentration:* Basic statistical considerations; randomization; design; analysis of continuous and categorical data; censored data and interim analysis; sample size; efficacy and safety assessments and evaluation; protocol preparation; data management; noninferiority designs	Statistics – Textbook
Pharmaceutical Statistics – Practical and Clinical Applications, Fourth Edition	Bolton, S. and Bon, C.	2003, Marcel Dekker	Undergraduate/ Graduate	*Statistics concentration:* Graphics; probability; inference; sample size; linear regression; ANOVA; Factorial designs; transformations and outliers; experimental design; quality control; nonparametric	Introductory statistics textbook. Includes disk with programs. Requires no previous knowledge of statistics
Statistics Applied to Clinical Trials – Second Edition; Self-Assessment Book also available	Cleophas, T.J., Zwinderman, A.H. and Cleophas, T.F.	2002, (2nd Edition) Kluwer; 2003 (Self-Assessment); Kluwer	Graduate	*Statistics concentration:* Overview of analysis of efficacy and safety data; equivalence; sample size; interim analyses; confounding and interaction; meta-analysis; crossover studies; logistic regression; genetic analysis	Statistics – Textbook
Statistics in Drug Research Review in: Journal of Mathematical Psychology (2002) 46:791	Chow, S-C. and Shao, J.	2002, Marcel Dekker	Graduate	*Statistics Concentration*: Preclinical (assay validation, stability, bioequivalence); noninferiority; statistical methodologies for medical imaging; randomization; blinding; missing data; meta-analysis; nonparametrics	Statistics – Reference

Common Statistical Methods for Clinical Research with SAS Examples – Second Edition	Walker, G.A.	2002, BBU Press	Graduate	*Statistics Concentration:* Statistical inference; examples of standard clinical trial analyses in SAS; interim analyses; crossover trials; multiple comparisons	Statistics – Textbook
Clinical Trials	Duley, L. and Farrell, B. – editors	2001, BMJ Books	Graduate	*Overview of clinical trials (Compiled from a series of papers given at Oxford University):* Randomization; large simple trials; usefulness of controlled clinical trials; data monitoring committees; Bayesian and ethics; carrying out trials	Overview – Reference
Applied Statistics in the Pharmaceutical Industry With Case Studies Using S-Plus Reviewed in Technometrics: 2002; 44:410	Millard, S.P. and Kraus, A. – editors	2001, Springer-Verlag	Graduate	*Statistics concentration:* Experiences of over 30 statisticians in all stages of clinical trials ranging from preclinical to production; examples with S-Plus (no prior knowledge required)	Statistics – Reference
Statistical Methods for Clinical Trials Review in: Psychiatric Services (2001); 52:1540	Norleans, M.X.	2000, Marcel Dekker	Graduate	*Statistics Concentration:* Graphical analysis; analysis of variance; meta-analysis; analysis of dichotomous outcomes; survival analysis; generalized linear models; correlated data (GEE; mixed model)	Statistics – Textbook

(*continued overleaf*)

Table 1 (*continued*)

Title	Author(s)	Year of publication and publisher	Level/ Audience	Concentration/ Content highlights	Possible course use – type of course/textbook or reference
An Introduction to Randomized Controlled Clinical Trials	Matthews, J.N.S.	2000, Hodder & Stoughton Educational	Undergraduate/ Graduate	*General overview of clinical trials:* Definition of randomized control trial; sample size; randomization; blinding; analysis; data monitoring; multiplicity; protocols; various study designs	Introduction/ Overview/Design – Textbook
A Guide to Clinical Drug Research – Second Edition	Edited by: Cohen, A. and Posner, J.	2000, Kluwer	Undergraduate/ Graduate	*General overview of clinical trials:* Investigator's brochure; planning a clinical study; protocol writing; study design; discussion of data collection and analysis; ethics; study conduct; good clinical practice	Introduction/ Overview/Design – Reference
Basic Statistics and Pharmaceutical Applications Reviews in: Biometrics (2000); American Journal of Pharmaceutical Education	De Muth, J.	1999, Marcel Dekker	Undergraduate/ Graduate	*Statistics concentration geared towards non-statisticians:* Probability; sampling; normal distribution; confidence intervals; hypothesis testing; *t*-tests; analysis of variance; correlation and regression; chi-square test; nonparametrics; bioequivalence	Statistics – Textbook; requires no previous knowledge of statistics

Design and Analysis of Clinical Experiments	Fleiss, J.L.	1999, Wiley	Graduate	*Statistical concentration:* Reissue of the classic text discussing parallel groups; prognostic variables; analysis of covariance; repeated measures; crossover designs	Statistics – Reference
Randomised Controlled Trials	Jadad, A.R. – editor	1998, BMJ Books	Undergraduate/ Graduate	*General overview of clinical trials:* Basic concepts and types of randomized controlled trials; assessing quality of randomized controlled trials; reporting results; meta-analyses; decision making; evidence-based health care	Introduction/ Overview/Design – Reference
Fundamentals of Clinical Trials – Third edition	Friedman, L.M., Furberg, C.D. and DeMets, D.L.	1998, Springer-Verlag	Graduate	*General overview of clinical trials*: Introduction to clinical trials; study design; randomization; blindness; sample size; baseline assessment; recruitment; data collection and quality control; adverse event; quality of life; adherence; general and specific issues in data analysis; reporting results	Introduction/ Overview/Design – Reference
Clinical Trials: A Methodologic Perspective	Piantadosi, S.	1997, Wiley	Undergraduate/ Graduate	*Statistics concentration:* Ethics; bias and random error; objectives and endpoints; sample size; randomization; stopping the trial; estimating clinical effects; reporting; factorial designs; crossover designs; meta-analyses; fraud and misconduct	Statistics – Reference

(*continued overleaf*)

Table 1 (*continued*)

Title	Author(s)	Year of publication and publisher	Level/ Audience	Concentration/ Content highlights	Possible course use – type of course/textbook or reference
Statistical Issues in Drug Development	Senn, S.	1997, Wiley	Undergraduate/ Graduate	*Statistics concentration geared towards non-statisticians:* Treatment allocation; covariates; measuring treatment effects; subgroup analysis; multiplicity, ITT, sample size; multicenter trials; equivalence; meta-analysis; crossover trials; sequential trials; safety data	Statistics – Reference; requires no previous knowledge of statistics
Randomized Controlled Clinical Trials – Second edition Review of first edition in: Controlled Clinical Trials. 1985; 6	Bulpitt, C.J.	1996, Kluwer	Undergraduate/ Graduate	*General overview of clinical trials:* History; Ethics; objectives; designs; sample size; bias; protocol; conduct; stopping rules; analysis; adverse events; risk-benefit	Introduction/ Overview/Design – Reference
Recent Advances in Clinical Trial Design and Analysis Review in: Annals of Oncology. 1996: 7; Clinical Oncology. 1997; 9:4	Thall, P.F. – editor	1995, Kluwer	Graduate	*Statistics concentration*: Alpha-spending; AIDS trials; failure time data; Tree-based prognostic models; Bayesian methods; exact analysis of contingency tables; sample size; quality of life; phase II designs	Advanced statistics – Reference
Statistics in Medical Research Review in: Controlled Clinical Trials (1996). 17:176–177; JASA (1996). 91:430–431	Gehan, E.A. and Lemak, N.A.	1994, Kluwer	Undergraduate/ Graduate	*Historical reference*: Birth of statistics; the start of collaboration between medical researchers and statisticians; designs and analysis of important historical studies; statistics role in clinical trials	Introduction/ Overview/Design – Reference

Statistics in the Pharmaceutical Industry – Second Edition	Buncher, C.R. and Tsay, J.Y. – editors	1993, Marcel Dekker	Undergraduate/ Graduate	*General overview*: Historical Review of Pharmaceuticals; FDA; Designs; Patient selection; cancer, antiepileptic, analgesic, AIDS trials; bioavailability; crossover trials; interim analysis; CROs, reporting results; Multiplicity; data QA	Introduction/ Overview/Design – Reference
Statistical Methodology in the Pharmaceutical Sciences Review in: JASA (1991); 86:250 Stat Med (1990); 9:1382–1383	Berry, D.A. – editor	1989, Marcel Dekker	Graduate	*General statistical overview*: Designs; Bioavailability; Repeated Measures; Dose–Response; population models; regression; multicenter trials; crossover; dropouts; group sequential methods; survival analysis; robust data analysis; categorical data analysis; adverse events; Bayesian meta-analysis	First year statistics course – Reference
Statistical Issues in Drug Research and Development Review in: Appl. Stat. (1991). 40:185–186; JASA (1991). 86:249–250	Peace, K.E. – editor	1989, Marcel Dekker	Graduate	*General statistical overview:* Adverse events; crossover; optimization via combination drug trials; equivalence; intention-to-treat.	Introduction/ Overview/Design – Reference

(*continued overleaf*)

Table 1 (*continued*)

Title	Author(s)	Year of publication and publisher	Level/ Audience	Concentration/ Content highlights	Possible course use – type of course/textbook or reference
Biopharmaceutical Statistics for Drug Development Review in: JASA (1989). 84: 629; Stat Med. (1990). 9:1224–1225	Peace, K.E. – editor	1987, Marcel Dekker	Graduate	*Statistics concentration*: Regulatory aspects; clinical development; bioavailability and bioequivalence; continuous and categorical efficacy data; cancer trials; group sequential stopping rules; safety	Statistics – References
Clinical Trials: A Practical Approach 1984 edition reviewed in: Biometrics (1984). 40:1211–1212 JRSS-A (1985), 148: 62–63; JASA (1987). 82: 360–361	Pocock, S.J.	1987, Wiley	Undergraduate/ Graduate	*Classical general overview*: Design, analysis, and interpretation of clinical trials: rational, history, justification of clinical trials; randomization; blinding; ethics; crossover trials; sample size; monitoring; data management; protocol deviation; basic principles of analysis	Introduction/ Overview/ Design – Textbook
Clinical Trials – Design Conduct and Analysis Review in: JASA (1988). 83:923–924; Stat. Med. (1988). 7: 545	Meinert, C.L.	1986, Oxford	Undergraduate/ Graduate	*General overview with analysis:* Introduction; Design; carrying out of clinical trials; analysis and interpretation; managing clinical trials; reporting results	Introduction/ Overview/ Design – Reference

Cross-over Trials in Clinical Research by Stephen Senn (2002).
Publisher: Wiley.
Reviews of first edition in:
JRSS-A (1993); 156:512–513;
Biometrics (1994); 50:586;
Statistics in Medicine (1994); 13:298–300.
Cross-over Trials edited by Hothorn L (1996).
Publisher: Fischer, Stuttgart.
Cross-over Experiments: Design, Analysis, and Application by Ratkowsky DA, Evans M.A. and Alldredge JR (1993).
Publisher: Marcel Dekker.
Review in: Biometrics (1994); 50:586;
JASA (1994); 89:356–357
Statistics in Medicine (1994); 13:28–300;
JRSS-A (1995); 158:200.

Bayesian Methods
Bayesian Methods and Ethics in a Clinical Trial Design, edited by Kadane JB(1996).
Publisher: Wiley.
Review in: JASA (1997. 92:384–385.

Sequential Methods/Interim Analyses
Data Monitoring Committees in Clinical Trials: A Practical Perspective by Susan S. Ellenberg, Thomas R. Fleming, David L. DeMets (2002).
Publisher: Wiley.
Group Sequential Methods with Applications to Clinical Trials by Jennison C and Turnbull BW (1999).
Publisher: Chapman & Hall/CRC, London.
The Design and Analysis of Sequential Clinical Trials, Second Edition by Whitehead J (1997).
Publisher: Ellis Horwood.

Quality of Life
Quality of Life Assessment in Clinical Trials: Methods & Practice by Staquet MJ (1998).
Publisher: Oxford University Press.
Quality of Life Assessments in Clinical Trials by Spilker B (1990).
Publisher: Lippincott-William & Willkins.

Meta-Analysis
Meta-Analysis of Controlled Clinical Trials by Whitehead A (2002).
Publisher: Wiley.

Discussion

Whether one wants to learn more about statistical issues in clinical trials, is a clinical trials professional, or is planning or teaching a course in the design or analysis of clinical trials, the number and type of texts from which to choose is tremendous. The number of books will only increase in the future, as theory and technology further develop. Before choosing the text (or texts), the reader needs to first determine his or her objectives (not necessarily an easy task itself). After that determination, it is hoped that the above discussion and the Table 1 provide a useful guidance to those wishing to find an appropriate reference or textbook.

JOSEPH M. MASSARO

Thiele, Thorvald Nicolai

Born: 24 December, 1838, Copenhagen, Denmark.
Died: 26 September, 1910, Copenhagen, Denmark.

T.N. Thiele trained as an astronomer at the University of Copenhagen and held the position of Professor of Astronomy there from 1875 to his retirement in 1907. He worked out the **actuarial** basis for the new life insurance company Hafnia which was founded in 1872 and had Thiele as its mathematical director until 1901. Along with these activities, Thiele made deeply original contributions to mathematical statistics, described by Hald [2–4], Lauritzen [5–7] and Edwards [1].

Thiele contributed to the theory of **skew** distributions, particularly in his study of what is often called the Gram–Charlier type A distribution, and he invented cumulants (*see* **Characteristic Function**), calling them semi-invariants, and developed a theory for them, with the proposal to use empirical cumulants as estimators. He contributed importantly to linear models (*see* **General Linear Model**), using what was then a modern algebraic definition of the canonical form of the linear hypothesis and emphasizing the role of the **residuals** in **model checking**. The special cases of one- and two-way **analysis of variance** were considered separately.

In **time series**, Thiele studied a fascinating model consisting of a sum of a **regression** component, a **Brownian motion** and a white noise, as we should call it now. Along the way, he gave an attractive geometric construction and explanation of what we now call the **Kalman filter**. Lauritzen [5, 6] applied this model to the temporal development of hormone levels during pregnancy.

In estimating a **binomial** probability, Thiele made explicit use of the idea of **likelihood** using the specific Danish term "Rimelighed", which has a similar relation to "Sandsynlighed" (probability) as that of the corresponding English terms. Finally, Thiele contributed to the theory of **grouped** observations.

For a more detailed account, see [7].

References

[1] Edwards, A.W.F. (2001). Estimating a binomial parameter using the likelihood function. Comments on Thiele (1889), in *Annotated Readings in the History of Statistics*, H.A. David & A.W.F. Edwards, eds. Springer-Verlag, New York 129–135.

[2] Hald, A. (1981). T.N. Thiele's contributions to statistics, *International Statistical Review* **49**, 1–20.

[3] Hald, A. (2000). The early history of the cumulants and the Gram-Charlier series, *International Statistical Review* **68**, 137–153.

[4] Hald, A. (2001). On the history of the correction for grouping, *Scandinavian Journal of Statistics* **28**, 417–428.

[5] Lauritzen, S.L. (1976). Appendix to Winkel et al.: Method for monitoring plasma progesterone concentrations in pregnancy, *Clinical Chemistry* **22**, 427–428.

[6] Lauritzen, S.L. (1981). Time series analysis in 1880: a discussion of contributions made by T.N. Thiele, *International Statistical Review* **49**, 319–331.

[7] Lauritzen, S.L. (2002). *Thiele: Pioneer in Statistics*. Oxford University Press, New York.

NIELS KEIDING

Tied Survival Times

Tied survival, or failure, times frequently occur in survival studies. Although theoretically a lifetime is a continuous variable, in practice it is often measured to a degree of fineness due to measurement limitations on the way failure times are recorded, and the expense of more accurate measurements may outweigh the value of added information. If the number of ties are substantial, discrete failure time models may need to be considered. Therefore, discrete failure time methods or grouped data techniques such as life tables should be used. However, if there are only a few ties, then the regular procedures in handling continuous data may be used with some adjustment for tied observations. In the literature, adjustment for ties has been proposed and studied for various statistical procedures in survival analysis (see [1, 6, 7, 8, 9], and [10]). Here, we only discuss adjustment for ties for some common statistical procedures.

Consider the method of handling ties in the **Kaplan–Meier** or product-limit (PL) estimator of the survival function. If only one individual fails (no ties are present) at time t, then the factor for the single death in the PL estimator is $[1 - 1/Y(t)]$, where $Y(t)$ counts the number of individuals at risk at time $t-$. For tied uncensored observations, suppose d failures occur at time t. Split the times of the d failures infinitesimally so that the factor for the d failures in the PL estimator is

$$\left[1 - \frac{1}{Y(t)}\right]\left[1 - \frac{1}{Y(t)-1}\right]\cdots \times \left[1 - \frac{1}{Y(t)-d+1}\right] = 1 - \frac{d}{Y(t)}.$$

If censored and uncensored observations are tied at time t, then consider the uncensored individuals as having failed just before the censored observations.

In the k-sample test, the weighted **logrank test** statistic is

$$Z_h(t) = \int_0^t K(s)\, \mathrm{d}N_h(s) - \int_0^t K(s)\frac{Y_h(s)}{Y_\cdot(s)}\, \mathrm{d}N_\cdot(s)$$

for $h = 1, 2, \ldots, k-1$, where K is the weight function, $N_h(s)$ and $Y_h(s)$ are the number of failures during time period $[0, s]$ and number of individuals at risk prior to time s for the hth sample, respectively, and $N_\cdot = \sum_h N_h$, $Y_\cdot = \sum_h Y_h$. The covariance of $[Z_h(t), Z_j(t)]$ may be estimated consistently by

$$\hat{\sigma}_{hj} = \int_0^t K^2(s)\frac{Y_h(s)}{Y_\cdot(s)}\left[\delta_{hj} - \frac{Y_j(s)}{Y_\cdot(s)}\right]\mathrm{d}N_\cdot(s),$$

where δ_{hj} is a Kronecker delta, i.e. $\delta_{hl} = 1$ if $h = l$, and 0 otherwise. In the presence of tied observations, the covariance of $[Z_h(t), Z_j(t)]$ needs to be adjusted to

$$\hat{\hat{\sigma}}_{hj} = \int_0^t K^2(s)\frac{Y_h(s)}{Y_\cdot(s)}\left[\delta_{hj} - \frac{Y_j(s)}{Y_\cdot(s)}\right] \times \frac{Y_\cdot(s) - \Delta N_\cdot(s)}{Y_\cdot(s) - 1}\,\mathrm{d}N_\cdot(s).$$

Clearly, when there are no tied observations, $\hat{\sigma}_{hj}$ and $\hat{\hat{\sigma}}_{hj}$ coincide.

Cox's **partial likelihood** has been commonly used to estimate the coefficients, $\boldsymbol{\beta}$, in **Cox's (proportional hazards) regression model**. Let $t_1 < t_2 < \cdots < t_k$ be the k ordered event times. Let the set $\mathcal{D}_i$ consist of the d_i individuals who failed at time t_i and $\mathcal{R}_i$ be the **risk set** prior to t_i. Denote $\mathbf{s}_i = \sum_{l\in\mathcal{D}_i}\mathbf{z}_l$, where $\mathbf{z}_l$ is the covariate vector for individual l. If there are ties among event times, then the following adjusted partial likelihoods have been proposed:

1. Breslow [2] suggests a partial likelihood of

$$L_1(\boldsymbol{\beta}) = \prod_{i=1}^{k}\frac{\exp(\boldsymbol{\beta}'\mathbf{s}_i)}{\left[\sum_{l\in\mathcal{R}_i}\exp(\boldsymbol{\beta}'\mathbf{z}_l)\right]^{d_i}}.$$

2. Efron [5] proposed an alternative partial likelihood of

$$L_2(\boldsymbol{\beta}) = \prod_{i=1}^{k}\left(\exp(\boldsymbol{\beta}'\mathbf{s}_i) \Big/ \prod_{j=1}^{d_i}\left[\sum_{l\in\mathcal{R}_i}\exp(\boldsymbol{\beta}'\mathbf{z}_l) - \frac{j-1}{d_i}\sum_{l\in\mathcal{D}_i}\exp(\boldsymbol{\beta}'\mathbf{z}_l)\right]\right).$$

3. The third partial likelihood due to Cox [3] is based on a discrete-time hazard rate model. The

discrete logistic likelihood is

$$L_3(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{s_i})}{\sum_{\mathbf{q}\in\mathcal{Q}_\mathbf{i}} \exp(\boldsymbol{\beta}'\mathbf{s_q^*})},$$

where $\mathcal{Q}_i$ is the set of all subsets of d_i individuals who could be selected from the risk set $\mathcal{R}_i$ and $\mathbf{s}_q^* = \sum_{j=1}^{d_i} \mathbf{z}_{q_j}$ (*see* **Logistic Regression**).

4. The fourth alternative partial likelihood is (see [4])

$$L_4(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left(\int_0^\infty \prod_{j=1}^{d_i} \left\{ 1 - \exp\left[-\left(\exp(\boldsymbol{\beta}'\mathbf{z_j}) \Big/ \sum_{l\varepsilon\mathcal{R}_i^*} \exp(\boldsymbol{\beta}'\mathbf{z}_l) \right) t \right] \right\} \times \exp(-t)\, \mathrm{d}t \right),$$

where $\mathcal{R}_i^* = \mathcal{R}_i \backslash \mathcal{D}_i$ is the set of individuals whose event or censored times exceed t_i or whose censored times are equal to t_i. It is often called exact likelihood.

Note that, when the number of ties is small, Breslow's and Efron's likelihoods are quite close. Of course, if no ties occur at the event times, all four likelihood functions reduce to the regular Cox partial likelihood.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Breslow, N.E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–99.

[3] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[4] DeLong, D.M., Guirguis, G.H. & So, Y.C. (1994). Efficient computation of subset selection probabilities with application to Cox regression, *Biometrika* **81**, 607–611.

[5] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association* **72**, 557–565.

[6] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[7] Klein, J.P. & Moeschberger, M.L. (1997). *Survival Analysis: Applied Methods and Examples*. Springer-Verlag, New York.

[8] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

[9] Miller, R.G. (1981). *Survival Analysis*. Wiley, New York.

[10] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series A* **135**, 185–206.

MEI-JIE ZHANG

Time Lag Effect

The term *time lag effect* refers to the delay between the time of an intervention or exposure onset, such as the date on which a person begins smoking cigarettes, and the subsequent development of a health outcome, such as the diagnosis of lung cancer. A variety of such time lag effects are described in the article, **Latent Period**. To design and to analyze **prevention trials** efficiently, one must account for the sometimes considerable time lag between the onset of intervention and subsequent beneficial health effects.

MITCHELL H. GAIL

Time Origin, Choice of

In **survival analysis** and more general event history analysis there is often more than one relevant origin for the time to the event(s) under consideration. This problem is sometimes discussed under the somewhat unsatisfactory name "choice of time scale" – it is the *origin*, not the scale (which could be hours, years, decades, etc.), which is under debate.

Here, I briefly mention some subject matter issues as well as some technical statistical points regarding the choice of time origin. To the former I append some brief remarks on the rare occasions where alternative time *scales*, literally speaking, are relevant.

Subject-Matter Issues in Choosing Time Origin

A fundamental discussion in many epidemiologic studies of incidence and mortality, as well as the derived concepts of **prevalence** (*see* **Incidence–Prevalence Relationships**), is the **age–period–cohort** problem (also briefly mentioned in the article on the **Lexis Diagram**). In its basic version, this issue regards the decomposition of the effect on certain rates (**incidence**, mortality, lethality) of calendar time t, age a, and "cohort", i.e. time of birth $t - a$. It has been well known for many years in sociology as well as epidemiology that there are only two identifiable linear effects here (see the above entries for references; *see* **Bayesian Methods**; **Identifiability**).

Regarding *clinical* contexts, **hazards** of relapse, death, or other endpoints, often depend both on *age* and certain *duration* variables, such as time since disease onset, since primary treatment, or since onset of remission (*see* **Duration Dependence**).

In *clinical trials with staggered entry*, the substantively interesting time origin is usually duration on trial, but since the trial takes place in "real" (calendar) time, this time variable can rarely be ignored in design and analysis (*see* **Interim Analysis of Censored Data**; **Staggered Entry**).

In the examples given so far, only the time origin was under debate, and the fact that the several "time scales" all run parallel and equally fast means that apparently multivariate time representations such as the Lexis diagram may fool the uninitiated observer (see [4] and [2, Chapters 6, 31] for further discussion and the technical consequences for statistical modeling). However, in certain clinical contexts the *cumulative time on* (intermittent) *treatment* may be of central substantive interest, sometimes as a proxy for the *cumulative dose*, which may itself in some cases be treated as a time variable. Such general situations are covered in the brief discussion by Farewell & Cox [3], although their particular example concerns the choice of time origin. The considerable body of literature on **multivariate survival analysis** seems to be surprisingly modest in its discussion of classes of practical situations where the several time variables are not constrained to move in parallel [i.e. in the direction given by the vector $(1, 1, \ldots, 1)$].

Technical Statistical Issues in Choosing the Time Origin

The **semiparametric** idea in the **Cox regression model** is to fix attention on one time variable (that is, one definition of time origin) which is modeled "nonparametrically" in the *underlying* intensity, relegating other time variables to the parametric part of the model as **time-dependent covariates**. It is then useful (though surprisingly sparsely discussed in the literature, except for [2, Chapter 31]) to put some detailed thought into choosing the "underlying" time variable, usually as that for which (i) the variation in the **hazard** is unknown or is expected to be dramatic and (ii) a parametric description is less important. Other time variables yielding less dramatic variation in the hazard or requiring parametric description for interpretation purposes can more usefully be modeled in the parametric part. In many clinical studies it will be wise to choose *duration since entry on study* as the underlying variable and *age* as a time-dependent covariate. (Note that using the techniques of **delayed entry**, it is perfectly possible to choose freely.)

In the piecewise constant intensity models (often termed "**Poisson regression**", *see* **Grouped Survival Times**) all time "scales" enter symmetrically and the above complications disappear.

Arjas (*see* **Real Time Approach in Survival Analysis**) has advocated that one should always use parametric models with "real", i.e. calendar, time as time variables, and let the statistical modeling handle other time origins of interest.

A different set of problems concerns the technical analysis of **semi-Markov processes** (see [1, Examples X.1.7-8] for a survey). Here the basic problem is that, if *duration in a state* is taken as the basic time variable, then some desirable martingale properties are lost, necessitating more complicated mathematical derivation of asymptotic statistical properties.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

[3] Farewell, V.T. & Cox, D.R. (1979). A note on multiple time scales in life testing, *Applied Statistics* **28**, 73–75.

[4] Keiding, N. (1990). Statistical inference in the Lexis diagram, *Philosophical Transactions of the Royal Society of London, Series A* **332**, 487–509.

NIELS KEIDING

Time Series Regression

Time series regression is concerned with the situation in which the dependent and independent variables are measured over time. Examples might include mortality from sudden infant death syndrome (SIDS) and environmental temperature [2], or hospital admissions and air pollution [7]. An alternative is **transfer function modeling** [1] but the advantage of regression is that the method is flexible and the interpretation familiar and straightforward.

The potential for **confounding** in time series regression is very high – many variables either simply increase or decrease over time, and so will be correlated over time [8] (*see* **Correlation**). In addition many epidemiological variables are seasonal, and this variation would be present even if the factors were not causally related. It is important that seasonality and trends are properly accounted for (*see* **Seasonal Time Series**). Simply because the outcome variable is seasonal, it is impossible to ascribe causality because of seasonality of the predictor variable. For example, SIDS is higher in winter than in summer, but this does not imply that temperature is a causal factor; there are many other factors that might affect the result, such as reduced daylight, or the presence of viruses. However, if an unexpectedly cold winter is associated with an increase in SIDS, or very cold days are consistently followed after a short time by rises in the daily SIDS rate, then causality may possibly be inferred.

Often when confounding factors are correctly accounted for, the **serial correlation** of the residuals disappears; they appear serially correlated because of the association with a time-dependent predictor variable, and so conditional on this variable the residuals are independent. This is particularly likely for mortality data where, except in epidemics, the individual deaths are unrelated. Thus, one can often use conventional regression methods followed by a check for the serial correlation of the residuals and need only proceed further if there is clear evidence of a lack of independence. For further details of parametric and semiparametric approaches to modeling confounders, see [7].

Effect of Correlated Residuals on Least Squares Estimates

If the inclusion of known or potential confounders fails to remove the serial correlation of the residuals, then it is known that ordinary least squares does not provide valid estimates of the standard errors of the parameters

For a continuous outcome, suppose the model is

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + v_t, \quad t = 1, \ldots, n, \tag{1}$$

where $v_t = \varepsilon_t - \alpha v_{t-1}$, y_t is the dependent variable measured at time t, $\mathbf{x}_t$ is a vector of the predictor variables, $\boldsymbol{\beta}$ is a vector of regression coefficients and the ε_t are assumed independent normally distributed variables with mean zero and variance σ^2. Thus, we assume that v_t is generated by an AR(1) process with parameter α (*see* **ARMA and ARIMA Models**). If we also assume that $\mathbf{x}_t$ is generated by an AR(1) process with parameter γ, then it is possible to show [4] that using ordinary least squares to estimate $\boldsymbol{\beta}$, the ratio of the estimated variance to the true variance, is approximately $(1 - \alpha\gamma)/(1 + \alpha\gamma)$. Since, in general, $\mathbf{x}_t$ and v_t are likely to be positively correlated, the effect of ignoring serial correlation is to provide artificially low estimates of the standard error of the regression coefficients and thus to imply significance more often than the significance level would suggest, under the null hypothesis of no association.

Estimation using Correlated Residuals

Given the above model, and assuming α is known, a method of generalized least squares, known as the *Cochrane–Orcutt procedure*, can be employed [3].

Write $y_t^* = y_t - \alpha y_{t-1}$ and $\mathbf{x}_t^* = x_t - \alpha \mathbf{x}_{t-1}$. We can then obtain an estimate of $\boldsymbol{\beta}$ using ordinary least squares on y_t^* and $\mathbf{x}_t^*$. However, since α will not usually be known, it can be estimated from the ordinary least squares residuals e_t by

$$a = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}. \tag{2}$$

This leads to an iterative procedure in which we can construct a new set of transformed variables and thus a new set of regression estimates and so on until convergence. The iterative Cochrane–Orcutt procedure can be interpreted as a stepwise algorithm for computing maximum likelihood estimators of α and $\boldsymbol{\beta}$, where the initial observation y_1 is regarded as fixed [4]. If the residuals are assumed to be normally distributed, then full maximum likelihood methods are available, which estimate α and $\boldsymbol{\beta}$ simultaneously and this can be generalized to higher order autoregressive models and fitted using, say, PROC AUTOREG in the computer package SAS [6] (*see* **Software, Biostatistical**). However, caution is advised in using this method when the autocorrelations are high, and it is worth making the point that an autoregressive error model "should not be used as a nostrum for models that simply do not fit" [6, p. 192].

Regression with Counts

Many epidemiological series consist of counts, and require **Poisson regression** rather than ordinary linear regression. Zeger [9] described a method similar to the Cochrane–Orcutt method to allow for serial correlation, and using **generalized estimating equations** to estimate the parameters.

Essentially we assume a model of the form $\mathrm{E}(Y_t) = \mu_t$, where $\mu_t = \exp(\eta_t)$ and $\eta_t = \boldsymbol{\beta}'\mathbf{x}_t$ (a **loglinear model**). A latent process $\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}' = (\varepsilon_1, \ldots, \varepsilon_n)$, is assumed to generate the autocorrelation. However, conditional on $\boldsymbol{\varepsilon}$, we suppose that Y_t is a sequence of *independent* counts such that $\mathrm{E}(Y_t|\boldsymbol{\varepsilon}) = \mathrm{var}(Y_t|\boldsymbol{\varepsilon}) = \varepsilon_t\mu_t$. This type of model is likely to be found in practice since the reason for the counts being serially correlated is their mutual dependence on another, possibly unmeasured, variable. The covariance matrix is assumed to be of the form $\sigma^2\mathbf{D}^{1/2}\mathbf{R}(\alpha)\mathbf{D}^{1/2}$, where $\mathbf{D} = \mathrm{diag}(\mu_t + \sigma^2\mu_t^2)$, $\mathbf{R}(\alpha)$ is an autocorrelation matrix generated by an autoregressive model, and σ^2 is the variance of the latent process. The order of $\mathbf{R}(\alpha)$ is determined from the data and may be greater than one. Details of implementation are given in [2] and [8]. Further details of this type of model and others is given in [5].

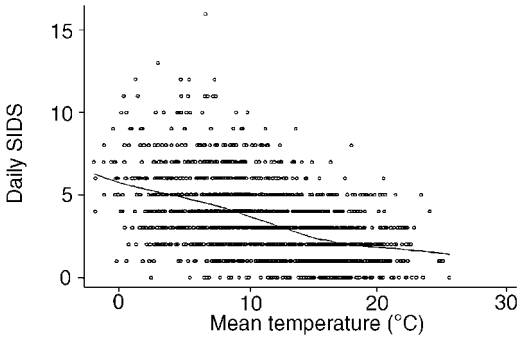

Figure 1 Daily sudden infant deaths (SIDS) in England and Wales from 1979 to 1983, with the mean temperature in London averaged over the period from two to five days earlier, with the lowess smoother plot

Example

Campbell [2] analyzed the dependence of daily deaths from SIDS in England and Wales from 1979 to 1983 on mean daily environmental temperature measured in London. A plot of the daily deaths against the temperature averaged over the period two to five days prior to the event is given in Figure 1, together with a lowess smoother plot with a bandwidth of 0.5. It can be seen that there appears to be a uniform decline in deaths with increasing temperature.

The predictor variables in the model were trend, annual and six month cycle sine and cosine terms, a weekend dummy variable, and the temperature averaged over the period two to five days prior to the event. The coefficient associated with mean temperature was -0.041 (se 0.005). We interpret this as saying that a $1\,^{\circ}\mathrm{C}$ drop in temperature is associated with a rise in SIDS by about 4%. Further investigations demonstrated that the relationship was approximately linear.

Often in practice, after controlling for seasonality and trend, the magnitude of the serial correlation is low, and estimates of the regression parameters will be little changed by incorporating serial correlation [6]. Thus, using Poisson regression on the SIDS data, having fitted annual and six-month sine and cosine terms and trend, the coefficient associated with mean temperature is -0.043 (se 0.005), which is very close to the Zeger estimate.

If the mean value for the counts is high, then methods assuming normality can be used and for this the standard software referred to earlier is available.

References

[1] Box, G.E.P. & Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden Day, San Francisco.

[2] Campbell, M.J. (1994). Time series regression for counts: an investigation into the relationship between Sudden infant death syndrome and environmental temperature, *Journal of the Royal Statistical Society, Series A* **157**, 191–208.

[3] Cochrane, D. & Orcutt, G.H. (1949). Application of least squares regression to relationships containing autocorrelated error terms, *Journal of the American Statistical Association* **44**, 32–61.

[4] Harvey, A. (1990). *The Econometric Analysis of Time Series*, 2nd Ed. Philip Alan, London.

[5] MacDonald, I.L. & Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall, London.

[6] SAS Institute Inc. (1984). *SAS/ETS User's Guide, Version 5*, Cary, NC.

[7] Schwartz, J., Spix, C., Touloumi, G., Bacharova, L., Barumamdzadeh, T., Le Terte, A., Piekarksi, T., Ponce de Leon, A., Poaska, A., Saez, M. & Schouten, J.P. (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions, *Journal of Epidemiology and Community Health* **50**, (Suppl 1), S3–S11.

[8] Yule, G.U. (1926). Why do we sometimes get nonsense-correlations between time series? A study in sampling and the nature of time series, *Journal of the Royal Statistical Society* **89**, 187–227.

[9] Zeger, S.L. (1988). A regression model for time series of counts, *Biometrika* **75**, 621–629.

(*See also* **Structural Time Series Models**)

MICHAEL J. CAMPBELL

Time Series Similarity Measures

Introduction

Time series is the simplest form of temporal data. A time series is a sequence of real numbers collected regularly in time, where each number represents a value. Time series data come up in a variety of domains, including stock market analysis, environmental data, telecommunications data, medical data, and financial data. Therefore, time series account for a large fraction of the data stored in commercial **databases**.

One interesting problem with time series data is finding whether different time series display similar behavior. More formally, we need to define a distance (or equivalently a similarity) function between two time series (*see* **Similarity, Dissimilarity, and Distance Measure**). Such a distance function can be used to cluster a set of time series in order to discover general patterns, or to classify a new time series. Different notions of *distance* between time series have been proposed in **data mining** research. The problem is hard because the similarity model should allow for imprecise matches, be efficient to compute, and allow the design of efficient indexing structures that can be used to find the most similar time series to a query [6].

Time Series Similarity

Generally, time series similarity models can be described in the following framework:

Time Series Similarity Functions

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ be two time series with n values. Let F_1 and F_2 be functions $F_i : R^n \to R^k$, for $k \leq n$, and D be a function $D : R^k \times R^k \to [0, 1]$. Then the distance of the two time series is a function $D(F_1(X), F_2(Y))$. The design of the functions F_1, F_2, D determine the characteristics of the distance function.

Setting the Function D

The simplest form for D is the p-norm distance between two k-dimensional vectors:

$$D(X, Y) = \left(\sum_{1 \leq i \leq k} |x_i - y_i|^p \right)^{\frac{1}{p}}. \qquad (1)$$

Usually, p is either two (Euclidean distance) or one (Manhattan distance). Such a function is easy to compute, allows efficient approximation, and also allows the design of efficient indexing techniques. The main disadvantage is that it does not allow shifting of the time series in time.

Figure 1 shows how to compute the Euclidean distance of two time series (in this case, the time series have been derived from electrocardiograms). The ith value of one sequence is matched to the i-th value of the other sequence. This can result in a large error if the two sequences are out of phase.

Reference [3] introduces the technique of *dynamic time warping* to time series similarity. Dynamic time warping is an extensively used technique in speech recognition, and allows acceleration–deceleration of signals along the time dimension. We are allowed to extend each sequence by repeating elements, thus creating X', Y'. The dynamic time warping distance, $DTW(X, Y)$, is defined as:

$$DTW(X, Y) = \min_{\forall X', Y' \text{ such that } |X'| = |Y'|} L_1(X', Y') \qquad (2)$$

Figure 2 demonstrates the use of the dynamic time warping technique. All values of each time series are matched with at least one value of the other time series. However, we are allowed to extend each time series by repeating some values.

A straightforward quadratic algorithm uses a bottom-up **dynamic programming** approach, where the smaller subproblems $DTW((X_1, \ldots, X_i), (Y_1, \ldots, Y_j))$ are solved first, which are then used to solve the larger subproblems, until finally $DTW(X, Y)$ is computed.

A different technique is to find the longest common subsequence ($LCSS$) of X and Y, and set $D(X, Y) = n - LCSS(X, Y)$ [4]. The $LCSS$ shows how well the two sequences can match one another if we are allowed to stretch them but we cannot rearrange the sequence of values. It can also be computed by a bottom-up dynamic programming algorithm:

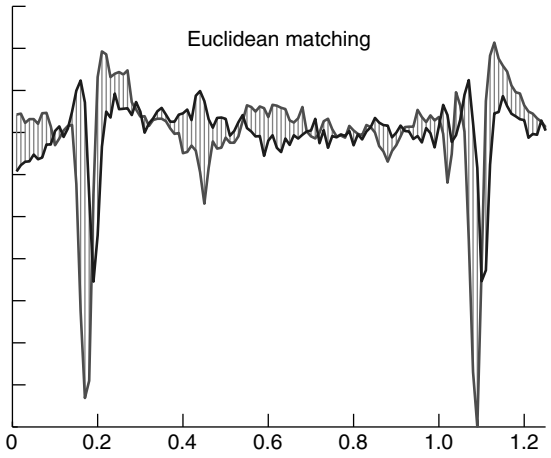

Figure 1 Using the Euclidean distance measure: the i-th value of time series X is matched to the ith value of the time series Y

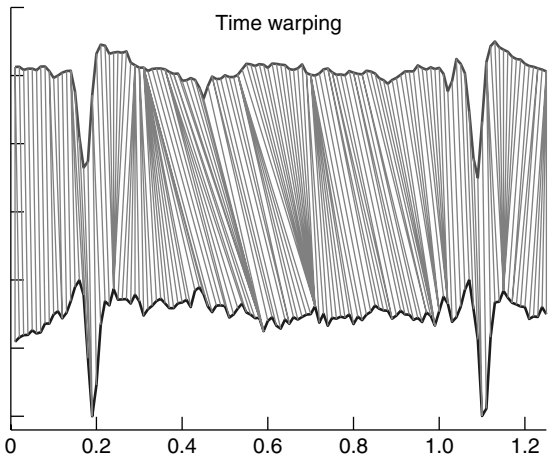

Figure 2 Using the dynamic time warping distance measure: a given value in each of the two time series can be matched to one or more consecutive values in the other time series. In the figure, the two time series have been separated vertically to make the matching clear

Given sequences $(x_1, \ldots, x_i)$, $(y_1, \ldots, y_j)$,
if $x_i = y_j$, then $LCSS((x_1, \ldots x_i), (y_1, \ldots, y_j)) =$
$\quad 1 + LCSS((x_1, \ldots, x_{i-1}), (y_1, \ldots, y_{j-1}))$
else $LCSS((x_1, \ldots x_i), (y_1, \ldots, y_j)) =$
$\quad \max(LCSS((x_1, \ldots, x_i), (y_1, \ldots, y_{j-1})),$
$\qquad LCSS((x_1, \ldots, x_{i-1}), (y_1, \ldots, y_j)))$.

Since the values are real numbers, we typically allow approximate matching. Figure 3 gives an example. The *LCSS* model allows shifting of the time series in time. One disadvantage of the *LCSS* model is that the triangle inequality does not hold, and therefore it is not formally a metric.

Setting F_1, F_2

There are many diverse proposals for what the functions F_1 and F_2 can be. They can be broken down into three main categories. However, techniques from different categories can be composed.

The first category is normalization functions: Let $\mu(X)$ and $\sigma(X)$ be the mean and variance of sequence X. The sequence X is replaced by the normalized sequence X', where

$$x'_i = \frac{x_i - \mu(X)}{\sigma(X)}. \tag{3}$$

Other similar transformations include moving averages for smoothening the time series [9, 15].

In the second case, the functions F_1 and F_2 are a specific transformation that is applied to the time series. A time series is represented as a point in n-dimensional space, and the transformation maps it to a point in a k-dimensional space ($k \leq n$). Such dimensionality-reduction techniques generally approximate the Euclidean distance of the original time series in the new space. The

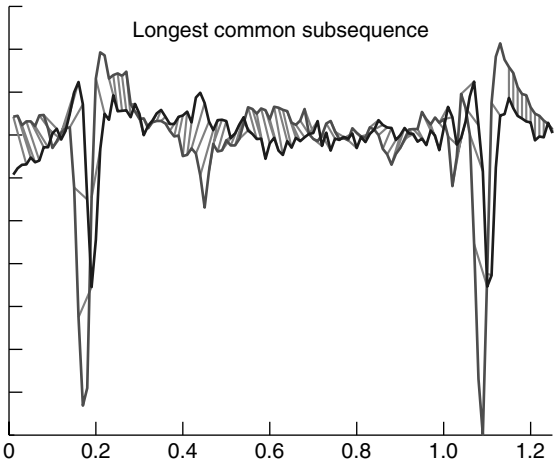

Figure 3 Using the longest common subsequence distance measure: Only similar values in the two time series are matched. Dissimilar values in one or both time series are dropped. The fraction of matched values determines the similarity of the time series

advantage of using a dimensionality reduction technique is that the distance computation is faster, and indexing structures are generally more efficient when used in lower dimensionality objects. Dimensionality reduction techniques that have been proposed include: the singular value decomposition (SVD) [12] (*see* **Matrix Computations**), the Fourier transform [1] (*see* **Fast Fourier Transform (FFT)**), the **Wavelet** decomposition [5], random projection techniques [10], FastMap [7], and Linear partitioning [13]. These techniques have specific strengths and weaknesses, making some of them better suited for specific applications and settings.

Another alternative is to define a family of functions $\mathcal{F}$, such that $F_1, F_2 \in \mathcal{F}$. The objective is to find those F_1, F_2 in $\mathcal{F}$ that minimize the distance. The distance between two time series X, Y is then

$$\text{argmin}_{F_1, F_2 \in \mathcal{F}} D(F_1(X), F_2(Y)).$$

The family of functions $\mathcal{F}$ can be global scaling, local scaling, global scaling, and different baselines.

This technique has generally been used with the *LCSS* notion of similarity. In [2], the authors develop a similarity measure that uses *LCSS*-like similarity with local scaling functions.

A simpler approach is to try and incorporate a single global scaling function with the *LCSS* similarity measure. In [4], an *LCSS*-like similarity measure is described that derives a global scaling and translation function that is independent of outliers in the data. The basic idea is that two sequences X and Y are similar if there exist constants a and b and long common subsequences X' and Y' such that Y' is approximately equal to $aX' + b$. The scale + translation linear function (i.e. the constants a and b) is derived from the subsequences, and not from the original sequences. Thus, outliers cannot taint the scale + translation function.

Probabilistic and Generative Methods

A different class of approaches to time series similarity is the class of *probabilistic and generative similarity measures*. Such measures have been studied in [8, 14]. Given a sequence X, the basic idea is to construct a probabilistic generative model M_X, that is, a probability distribution on waveforms. Once a model M_X has been constructed for a sequence X, we can compute similarity as follows. Given a new sequence pattern Y, similarity is measured by computing $p(Y|M_X)$, that is, the likelihood that M_X generates Y.

An alternate approach was undertaken by [11], who describe a general similarity framework involving a transformation rules language. Each rule in the transformation language takes an input sequence and produces an output sequence, at a cost that is associated with the rule. The similarity of sequence X to sequence Y is the minimum cost of transforming X to Y by applying a sequence of such rules. The actual rules language is application specific.

References

[1] Agrawal, R., Faloutsos, C. & Swami, A. (1993). Efficient similarity search in sequence databases, in *International Conference on Foundations of Data Organization (FODO)*, Chicago, IL.

[2] Agrawal, R., Lin, K.-I., Sawhney, H.S. & Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in time-series databases, in *Proceedings of the 21st International Conference on Very Large Databases (VLDB-95)*, Zurich, Switzerland.

[3] Berndt, D.J. & Clifford, J. (1994). Using dynamic time warping to find patterns in time series, in *KDD Workshop*, Seattle, Washington.

[4] Bollobas, B., Das, G., Gunopulos, D. & Mannila, H. (2001). Time-series similarity problems and well-separated geometric sets, *Nordic Journal of Computing*, **8**(4) 409–423.

[5] Chan, K. & Fu, W. (1999). Efficient time series matching by wavelets, in *Proceedings of the IEEE International Conference on Data Engineering*, Sydney, Australia.

[6] Faloutsos, Christos (1996). *Searching Multimedia Databases by Content*. Kluwer Academic Publishers, Boston.

[7] Faloutsos, C. & Lin, K.-I. (1995). FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets, in *ACM SIGMOD Conference*, San Jose, CA, 163–174.

[8] Ge, X. & Smyth, P. (2000). Deformable Markov model templates for time-series pattern matching, in *Proceedings of the ACM SIGKDD*, Boston, MA.

[9] Goldin, D.Q. & Kanellakis, P.C. (1995). On similarity queries for time-series data: constraint specification and implementation. *Proceedings of the 1st International Conference on Principles and Practice of Constraint Programming, Cassis, France.*

[10] Indyk, P. & Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality, in *Proceedings of the STOC*, Dallas, TX.

[11] Jagadish, H.V., Mendelzon A.O. & Milo, T. (1995). Similarity-based queries, in *Symposium on Principles of Database Systems (PODS)*, San Jose, CA.

[12] Jolliffe, I.T. (1989). *Principal Component Analysis*. Springer-Verlag, New York.

[13] Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases, *Journal of Knowledge and Information Systems.* **3**(3), 263–286.

[14] Keogh E.J. & Smyth., P. (1997). A probabilistic approach to fast pattern matching in time series databases, in *Proceedings of the KDD*, Newport Beach, CA, 24–30.

[15] Rafiei, D. & Mendelzon, A. (1997). Similarity-based queries for time-series data, in *Proceedings of 1997 ACM SIGMOD International Conference on Management of Data*, Tuscon, AZ.

(*See also* **Cluster Analysis of Subjects, Hierarchical Methods**; **Cluster Analysis of Subjects, Nonhierarchical Methods**; **Principal Components Analysis**)

DIMITRIOS GUNOPULOS

Time Series

A time series consists of values of a variable recorded, usually at regular intervals, over a long period of time. Such data arise frequently in medical investigations; for example, weekly admissions into a hospital, monthly mortality rates for a particular disease and daily concentrations of a pollutant. The observations in such a series are usually denoted by $x_1, x_2, \ldots, x_n$, where n is the length of the series. Such data usually require special methods for their analysis because of the presence of **serial correlation**. In a series of hourly blood pressure readings, for example, a "high" reading at 1 p.m. is likely to have a certain inertia and to remain relatively high at 2 p.m. Neighboring observations in a time series will frequently be positively correlated, with this correlation declining as the time interval between observations increases. The existence of a possibly complex pattern of dependency between the observations in a time series implies that methods of analysis that assume that the observations are independent will not be appropriate.

Preliminary Analysis of Time Series

Many time series can be considered to be a mixture of the following four components:

1. A trend or long-term movement.
2. Fluctuations about the trend of greater or lesser regularity.
3. A deterministic cycle; for example, a pronounced **seasonal** component.
4. A residual, irregular or random effect.

Part of the analysis of a time series might aim to provide a description of regular or systematic variation; for example, by identifying periodic effects or cycles (*see* **Circadian Variation**). Additionally, it might be hoped to develop models for the series that allow inferences about the mechanisms generating the series and also open the possibility of making predictions about the future value of the series, i.e. **forecasting**. An essential *first* step, however, when considering *any* time series is simply to plot the observations against time. Figure 1, for example, shows a plot of the number of deaths per quarter from ischemic heart disease for males in the UK, from 1967 to 1991. Figure 2 shows the reported cases of chicken pox in New York in the years 1928–1972.

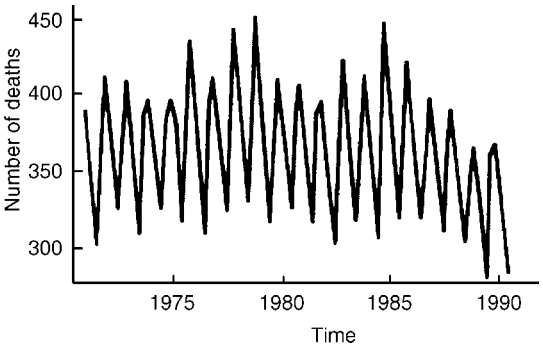

Figure 1 Number of deaths per quarter from ischemic heart disease for males in the UK – 1967–1991

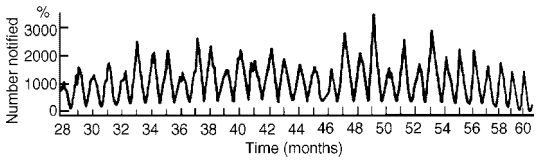

Figure 2 Reported cases of chicken pox in New York – 1928–1972

Simple plots such as those shown in Figures 1 and 2 are often valuable in highlighting qualitative features of a time series; for example, a trend, seasonality, or **outliers**, although such patterns are frequently obscured by "noise", making them less easy to detect without some formal analysis.

The simplest hypothesis that might be entertained about a time series is that it is random, i.e. a *white noise series* (*see* **Noise and White Noise**). A number of tests of randomness are available and are described in [2] (*see* **Cox's Test of Randomness**). In general, however, such tests are not applied to the original series, since, in most instances, this will clearly be nonrandom, but to the residuals arising after fitting some model or other to the series.

Stationarity

Stationarity signifies that the probability structure of a time series does not change with time. In particular, a stationary series has a constant mean and variance and a covariance structure that depends only on the difference between two time points. Many time series

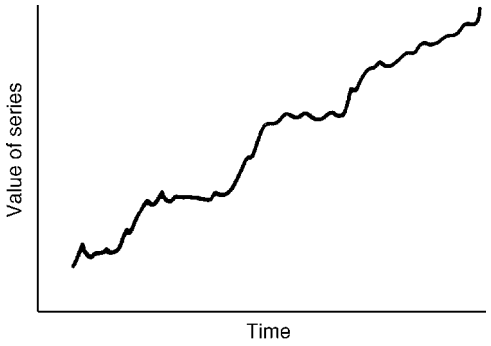

Figure 3 Time series with a trend

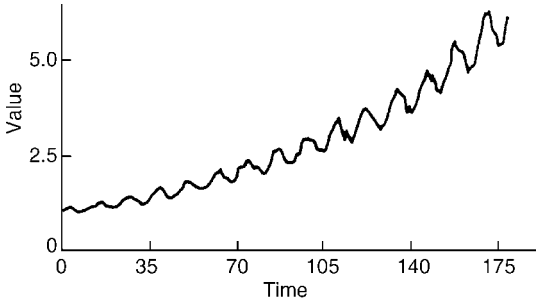

Figure 4 Time series in which variance is changing

encountered in medical studies are *not* stationary for a variety of reasons, the most common of which is the presence of a trend – see, for example, Figure 3. This feature of a series needs to be described and estimated in some way so that it can be removed before the series is analyzed further. One possibility is the use of some form of **moving average**. Another is to transform to a series of first differences,

$$z_t = x_t - x_{t-1}. \tag{1}$$

In some circumstances it might be more appropriate to take some simple **transformation** of the original series before attempting to remove any trends. If, for example, it is found that the variance is related to the mean, then a variance-stabilizing transformation would be needed. Such a series is shown in Figure 4.

Analyzing Time Series

Modern methods for the analysis for time series can be divided roughly into two classes – *frequency domain methods* and *time domain methods*.

Frequency Domain Methods

The primary aim of frequency domain methods for the analysis of time series is to identify oscillations of major importance, in the sense of explaining a large proportion of the variance in an observed series. Methods in this class are derived from the early ideas of Fourier analysis in which a series of observations is represented as a superposition of independently varying cosine and sine curves (*see* **Fast Fourier Transform (FFT)**). A typical sine wave, for example, has the form

$$x_t = A\sin(2\pi f t + \phi), \tag{2}$$

where the constant A is called the *amplitude*, f the *frequency*, and ϕ the *phase*. The curve is *periodic* with a period, $T = 1/f$. This simply means that the plot of x_t against t is the same at $t + 1/f, t + 2/f, \ldots$, etc. as at t. An example of a sine wave is shown in Figure 5.

An early tool for the analysis of time series data that used the idea of Fourier decomposition was the *Schuster periodogram* [4]. The time series observations are expressed as a sum of cosine curves of the form:

$$x_t = A_0 + \sum_{k=1}^{(n-1)/2} A_k \cos(2\pi f_k t + \phi_k). \tag{3}$$

The amplitude, A_k, and phases, ϕ_k, can be calculated from the data, with A_k indicating the importance of oscillations of period $1/f_k$ in the observed series. The periodogram is simply a plot of nA_k^2 against k. If the original series contains a well-defined cyclic component, then the periodogram can be expected to have a sharp peak at the appropriate value of k. In practice, however, such a peak is often masked because the great variability in the A_k values makes

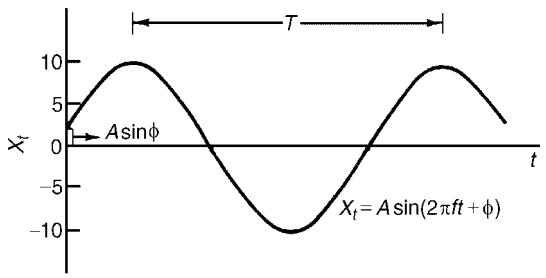

Figure 5 Sine wave

the plot extremely irregular. In other cases, apparent peaks may appear, even in the absence of genuine cycles, because one or more local maxima will seem substantially larger than neighboring values.

Most current frequency domain analyses are less concerned with the discovery of exact periodicities and more concerned with assessing how the variance of a series is distributed amongst oscillations of different frequencies by estimating the *spectrum* of the series (*see* **Spectral Analysis**). Such an approach has been most widely applied in the medical field in the analysis of EEG signals – see [1].

Time Domain Methods

The techniques used for the analysis of time series in the time domain are based on direct modeling of the lagged relationships between a series and its past. An important initial step in the search for an appropriate model for a time series is an examination of the dependence structure of the series via the *correlogram* (*see* **Autocorrelation Function**), which is a plot of the lagged correlations of a series against lag size.

The models most commonly used for time series data are those known as ARMA – autoregressive moving-average models or ARIMA – autoregressive integrated moving-average models (*see* **ARMA and ARIMA Models**). The parameters in such models can be estimated by likelihood methods. Examples of situations in which such models are of importance are:

1. Epidemiologists are often faced with assessing the relationship between a target or output series, such as the daily number of patients coming to a clinic, and explanatory or input series, such as the daily concentration of a pollutant.
2. Questions about changes in time series are frequently of great importance in medicine. An investigator might, for example, be interested in assessing how the pattern of morbidity in a population changes after an environmental accident, or in measuring the effectiveness of a campaign to make teenagers aware of the dangers of AIDS.
3. Accurate forecasts of the future values of a time series may be of great value in many areas of medicine. Public health organizations, for example, need to know what frequencies of diseases might be expected in coming years in order to plan how to allocate often limited resources.

Summary

The analysis of time series is a large and complex area. The main techniques are spectral analysis and ARMA or ARIMA models. All the main software packages have facilities for implementing both forms of analysis (*see* **Software, Biostatistical**). The **S-PLUS** package has facilities for more extensive and exotic analyses. A further software package, STAMP, developed by Koopman et al. [3], is useful for the analysis of time series using regression models in which the explanatory variables are functions of time, but with coefficients that change over time (*see* **Structural Time Series Models**).

References

[1] Gasser, T. & Molinari, L. (1996). The analysis of the EEG, *Statistical Methods in Medical Research* **5**, 67–100.
[2] Kendall, M.G. & Ord, J.K. (1990). *Time Series*, 3rd Ed. Edward Arnold, London.
[3] Koopman, S.J., Harvey, A.C., Doornik, J.A. & Shephard, N. (1995). *STAMP 5.0 Structural Time Series Analyser, Modeller and Predictor*. Chapman & Hall, London.
[4] Schuster, A. (1898). On the investigation of hidden periodicities, *Terrestrial Magnetism* **3**, 13.

(*See also* **Coherence Between Time Series**; **Multiple Time Series**; **Nonlinear Time Series Analysis**)

BRIAN S. EVERITT

Time to Pregnancy

The time from initiating attempts to become pregnant until conception occurs (*time-to-pregnancy* or TTP) is gaining importance as a measure of natural fecundity [9]. This article highlights some special features in modeling and design for this special application of **survival analysis** methods.

Models for Time-to-pregnancy (TTP) Data

Statistical models for time-to-pregnancy (TTP) data belong to the general area of survival analysis. Often, the focus is on menstrual cycle as the time unit, which makes time intrinsically discrete and has motivated most authors' use of **discrete-time survival models**. However, the actual measurement is often not made in cycles but rather in months, either as a surrogate for menstrual cycles or because months are considered to be of interest in themselves, and then there would seem to be no harm in using continuous time survival models.

The statistical models need to be able to accommodate known heterogeneity between couples in the form of **covariates**, and one will often also want to incorporate residual random heterogeneity.

Let $t = 1, 2, \ldots$ be the number of menstrual cycles since "initiation", that is, since attempts at getting pregnant started, and let x_t be a vector of covariates at time t. The task is to model the discrete **hazard rate**

$$\lambda(t|x_t) = P(T = t|T \geq t, x_t), \tag{1}$$

which is the probability of becoming pregnant at cycle t, given that this did not happen before t, and given the covariates.

An early influential model by Weinberg and Gladen [10] assumed

$$\log(\lambda(t|x_t)) = x_t'\beta, \tag{2}$$

where log is natural logarithm and

$$x_t'\beta = x_{1t}\beta_1 + \cdots + x_{kt}\beta_k. \tag{3}$$

The model has the disadvantage that when β varies across $(-\infty, \infty)$, $\lambda(t|x_t)$ is not restricted to the range [0,1] of a probability. This problem was avoided by Scheike and Jensen [6] who postulated

$$\log(-\log(1 - \lambda(t|x_t))) = x_t'\beta \tag{4}$$

in line with current practice in discrete-time survival models [2]. This model may be interpreted as a grouped-time version of the **Cox regression model**.

Weinberg and Gladen incorporated *unobserved heterogeneity* in a model with no covariates by assuming that the hazard for each given couple was constant over time: $\lambda(t) = r$, and that r follows a **beta distribution** across the population. The resulting marginal hazard in the population is

$$\lambda(t) = \frac{1}{\alpha + \mu(t-1)} \tag{5}$$

with parameters α and μ given by the beta distribution. Weinberg and Gladen extended this model to accommodate covariates by postulating

$$\lambda(t) = \frac{1}{\alpha + \mu(t-1)} + x_t'\beta, \tag{6}$$

although the interpretation of a mixture across the population is then lost; indeed, the parameter β has no interpretation at the individual level, only marginally for the population (*see* **Marginal Models**).

In contrast, Scheike and Jensen [6] extended their model to incorporate a random effect R_i for couple i by assuming

$$\log(-\log(\lambda(t|R_i, x_{it}))) = R_i + x_{it}'\beta. \tag{7}$$

In this model, the individual interpretation of β is conserved. Scheike et al. [7] discussed generalizations of this model to allow several times to pregnancy per couple and connected to the current discussion of **frailty** models and **multivariate survival analysis**, including an important discussion of the interpretations of conditional versus marginal parameterizations. Ecochard and Clayton [1] generalized to a three-parameter distribution, while assuming constant baseline, and also allowed several pregnancies per couple.

Sampling Designs

The two most common and obvious designs are a **cohort** (follow-up) **study** where couples are followed forward in time from when they start attempting to become pregnant, or a **retrospective study** of pregnant women where couples are interviewed about when they started their attempt to become pregnant. A variation of the cohort study is the *historically*

prospective design where a general sample (usually of women) from the population is asked to recall their reproductive history. Below, we discuss these designs as well as the possibilities of using **cross-sectional** samples, referring to Weinberg and Wilcox [11] for a broader epidemiological discussion.

Prospective Sampling

In principle, the cohort approach leads to standard right-**censored** survival data, where the couples who have not conceived at the end of follow-up are counted as *right-censored*. Note that couples recruited into a prospective study at a known time t after initiation will have to be counted with **delayed entry** (left **truncation**) at t. In practice, prospective studies are not very common, usually rather small, and often marred by considerable self-selection problems (*see* **Selection Bias**). A particular difficulty with assessing the effect of calendar time is whether to score it at initiation, at conception (which creates difficulties at least for the censored couples), or as current calendar time along the way. The historically prospective study suffers from recall bias and also mixes experience over a long calendar time period.

Retrospective Sampling

Large TTP surveys are often retrospective, data being gathered from pregnant women. There are obvious weaknesses with these, primarily the biased sampling based on fecundity, particularly, the nonpresence of the sterile or nonfecund couples, but also under representation of the subfecund. Juul et al. [4] demonstrated how a true age-decreasing fecundity in a heterogeneous population can be made to look age-increasing by naive analysis of a retrospective sample.

However, even beyond these unavoidable difficulties, the correct analysis of retrospective TTP data is more intricate than often realized, particularly when the focus is on revealing the dramatic trends in initiation intensity, which must be behind the observed secular trends in birth rates. As an example, in a common design, the data are gathered from interviews in a fixed time window. It is then clear that if calendar time is related to initiation, long TTPs will be over represented in the early phase, short TTPs in the late phase, with intricate patterns of left and right truncations [6]. As pointed out by Jensen et al. [3], dramatic artificial temporal trends in fecundity may be generated by disregarding the effects of these truncations. These phenomena were defined away by a tacit (hardly tenable) assumption of stationarity in the classical work of Weinberg and Gladen [10], as made explicit in Weinberg et al. [8, p. 679]. Incorporation of several TTPs per couple in a retrospective design is possible through careful **likelihood** constructions; see [7] for details.

Current Durations in a Cross-sectional Sample

A simple procedure would be to ask a cross-sectional sample of a population (or subpopulation) of women whether they are currently attempting to get pregnant and if so, for how long have they attempted to do so. With this design, it would seem reasonably realistic to minimize selection bias: there is no a priori exclusion of sterile couples as in retrospective sampling and minimal self-selection in contrast to most prospective studies. This design was briefly mentioned by Weinberg and Gladen [10] and studied in some detail by Keiding et al. [5].

It may be useful to summarize the distributions involved in these three main sampling designs, as follows.

For each attempt at becoming pregnant, let T be the time to pregnancy, U the time to discontinuation without pregnancy (for reasons such as death of the woman, disappearance of partner, couple gives up trying; in some cases, the start of fertility treatment should perhaps be included), and V the time to discontinuation of follow-up since the start of the attempt. We are interested in the distribution of T. In a *prospective* design, the problem reduces to standard survival analysis with T as the time to endpoint and $\min(U, V) = U \wedge V$ the time to censoring. In the *retrospective* design (based on pregnant women), we have a complete sample from the **conditional** distribution of $T|T < U$. (Note that this situation is different from right truncation of T by U, which corresponds to observing the conditional distribution of $(T, U)|T < U$).

In the *current-duration* design, let $X = T \wedge U$ be the waiting time until termination for whatever reason, successful or not, with probability

density $f(x)$, survival function $S(x) = \int_x^\infty f(a)\,\mathrm{d}a$, and expectation $\mu_X = \int_0^\infty xf(x)\,\mathrm{d}x = \int_0^\infty S(x)\,\mathrm{d}x$, which we shall assume finite. Cross-sectional sampling takes place at some fixed time t_0, and assume that initiations happen according to a **Poisson process** in calendar time t with intensity $\beta(t)$. In the time-homogeneous situation, $\beta(t) = \beta$, which should suffice in most situations where only short calendar intervals are considered for each "cross-section", the observed experienced waiting time at t_0 ("current duration"), $Y = X \wedge V = T \wedge U \wedge V$ will be distributed as a backward recurrence time in a **renewal process** in equilibrium with renewal distribution $f(x)$, that is, the *density* of Y is

$$g(y) = \frac{S(y)}{\mu_X}. \tag{8}$$

Note in particular that $0 < g(0) < \infty$. Thus, Y has a *decreasing density* proportional to the *survival function* of X.

References

[1] Ecochard, R. & Clayton, D.G. (2000). Multivariate parametric random effect regression models for fecundability studies, *Biometrics* **56**, 1023–1109.

[2] Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalised Linear Models*, 2nd Ed. Springer-Verlag, New York.

[3] Jensen, T.K., Keiding, N., Scheike, T., Slama, R. & Spira, A. (2000). Declining human fertility? *Fertility and Sterility* **73**, 421–422.

[4] Juul, S., Keiding, N. & Tvede, M. (2000). Retrospectively sampled time-to-pregnancy data may make age-decreasing fecundity look increasing, *Epidemiology* **11**, 717–719.

[5] Keiding, N., Kvist, K., Hartvig, H., Tvede, M. & Juul, S. (2002). Estimating time to pregnancy from current durations in a cross-sectional sample, *Biostatistics* **3**, 565–578.

[6] Scheike, T.H. & Jensen, T.K. (1997). A discrete survival model with random effects: an application to time to pregnancy, *Biometrics* **53**, 349–360.

[7] Scheike, T.H., Petersen, J.H. & Martinussen, T. (1999). Retrospective ascertainment of recurrent events: An application to time to pregnancy, *Journal of the American Statistical Association* **94**, 713–725.

[8] Weinberg, C.S., Baird, D.D. & Wilcox, A.J. (1994). Sources of bias in studies of time to pregnancy, *Statistics in Medicine* **13**, 671–681.

[9] Weinberg, C.S. & Dunson, D.B. (2000). Some issues in assessing human fertility, *Journal of the American Statistical Association* **95**, 300–303.

[10] Weinberg, C.S. & Gladen, B.C. (1986). The beta-geometric distribution applied to comparative fecundability studies, *Biometrics* **42**, 547–560.

[11] Weinberg, C.R. & Wilcox, A.J. (1998). Reproductive Epidemiology, in *Modern Epidemiology*, 2nd Ed., K.J. Rothman & S. Greenland, Lippincott Williams & Wilkins, Philadelphia, PA, Chapter 29, pp. 585–608.

(*See also* **Reproduction**)

NIELS KEIDING

Time Trade-off Technique

Decisions on medical treatments and the setting of health programs involve both technical and value judgments. An important one, for example, is evaluating tradeoffs between **quality of life** and length of life or between different domains of quality of life. In recent years the concept of **utility** has been introduced into medical decision making (*see* **Decision Analysis in Diagnosis and Treatment Choice**) to help estimate the preferences that individuals attach to the consequences of various courses of action. This information is useful for decisions at the patient level (i.e. clinical decision making) and the program level (i.e. the best way of using available health care resources or economic evaluation of health care programs; *see* **Program Evaluation**).

A commonly used measure of outcome for such analyses is QALY (i.e. quality-adjusted life-years). The QALY definition and calculation can be found elsewhere (e.g. [4]) (*see* **Quality of Life and Health Status**). In brief, the number of years spent at a given health status is multiplied by the corresponding weight (typically between zero (death) and one (full health)). The adjusted years are summed and discounted to reflect societal time preference (i.e. the rate that future benefits, and costs, should be adjusted to reflect their present value to society). This rate represents societal willingness to exchange present for future consumption. The number generated represents the equivalent in quality-adjusted life-years of a potential lifetime health profile (i.e. years of life in different health states). The time tradeoff (TTO) technique is one of the methods used to generate the preference weights to be used in the QALY calculations.

The Time Tradeoff Technique

The time tradeoff (TTO) technique was first suggested by Torrance et al. [19] as a substitute for the **standard gamble (SG) technique**, which is seen as the classical method of measuring cardinal preferences [18]. The need for a substitute for the SG technique stems from the empirical observation that subjects sometimes find it difficult to relate to probabilities. Compared with the SG technique, the TTO technique has the advantage of being simpler to use. This technique has been used extensively in many empirical studies to estimate individuals' value preferences for different health states. In this section we describe the TTO as suggested by Torrance et al. [19].

The TTO technique involves a **paired comparison** in which the subject chooses between two alternatives (a more detailed description of the TTO technique can be found in [4]). For simplicity we use the chronic health state (i.e. the case of an individual being in the same health state for the rest of their life) as an example. In this case the individual is presented with two options: alternative 1 (see Figure 1) is to have the health state under consideration (denoted as i) for time t (the remaining **life expectancy** of the individual) followed by death; alternative 2 is a shorter period of time (denoted as $x, x < t$) in full health. The period of time x is varied until the subject is indifferent between the two alternatives. The required preference value for state i (denoted as h_i) is then calculated to be $h_i = x/t$. Props and visual aids are recommended for use to help respondents understand the task.

An example for the use of this procedure is as follows. Suppose one wants to measure the preference value for the health state "severe pain". The first step is to construct scenarios describing the health states of "severe pain" and "full health". These scenarios will be used in the second step. In the second step, individuals are approached and presented with two options: alternative 1 is to have the health state "severe pain" for time t (say 20 years, which is the remaining life expectancy of that individual); alternative 2 is a shorter period of time (denoted as x) in full health. The period of time in full health is varied until the subject is indifferent between the two alternatives. Say $x = 5$ years at the point of indifference. Then, using the equation described above, the required preference value for the health state "severe pain" is then calculated to be $5/20 = 0.25$.

An Indifference Curve Interpretation for the TTO

Unlike the SG method, the TTO technique, suggested by Torrance et al. [19], was not related in a general

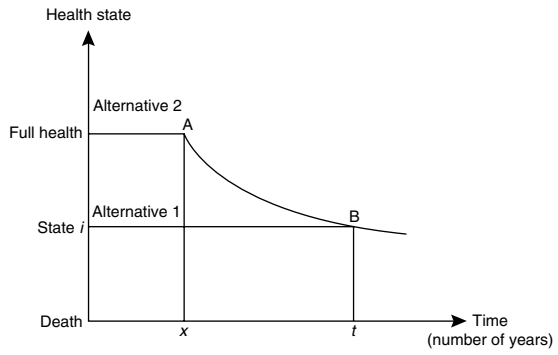

Figure 1 Time tradeoff for a chronic health state

way to any existing behavioral theory. Relating a measure to a behavioral theory is important because it helps us interpret what we measure as well as understand and test the assumptions which underlie the measurement procedure. A general theoretical interpretation for the TTO technique was provided later [10]. If the measure is to be used in the context of an economic evaluation it is also important to make sure that it is consistent with the principles of the discipline. In economics, one starts from choosing a welfare theory as the basis for the analysis. The choice of the underlying welfare theory determines, amongst other things, the types of outcome measure that can be used in the analysis [5] (*see* **Health Economics**).

Mehrez & Gafni [10] argue that the TTO technique is a method which enables us to identify different points on an individual's indifference curve in his evaluation space. An evaluation space is defined as the set of all potential alternatives (or outcomes). An indifference curve is the locus of all points (alternatives) in the individual's evaluation space among which he is indifferent. In other words, following classical utility theory, individuals are assumed to have a mechanism (that can be described as a mathematical function) which associates a real number to each alternative in the evaluation space. This function describes the individual's preferences. The indifference curve is defined as the locus of all points which have the same numerical score.

Following the above, the TTO can be seen as a two-stage procedure. First, a comparison of two alternatives in the individual's evaluation space is performed. Using the concept of indifference curves, this stage can be seen as follows: given one point in the individual's evaluation space (point B in Figure 1; the health state under consideration for period t), we search for another point (point A in Figure 1, full health for period x) which lies on the same indifference curve. In the second stage we attribute a preference score to the health state under consideration using the method described earlier. As shown in [10], in the second step additional assumptions are required about the functional form of the individual's preference function in order to calculate a preference value for the chronic health state. These assumptions, however, are not required for the first step.

To illustrate this important point, let V denote the individual's preference function. By definition, because the individual has indicated indifference between alternatives 1 and 2 (points A and B in Figure 1), then $V_2 = V_2$, i.e. the "numbers" generated by the individual's preference function will be the same for both alternatives. However, the preference value for health state i cannot be calculated without knowing the exact functional form of the preference function. Assume, for example, that the individual has a preference function of the type $V = h_i T$, where h_i and T are the two attributes of concern (h_i = preference value for health state i, $0 \leq h_i \preceq 1$, 1 = full health, 0 = death, and T = years). Using alternatives 1 and 2 in Figure 1, the score of alternative 1 is $V_1 = h_i t$ and of alternative 2 is $V_2 = 1.0x$. With indifference between the alternatives, $h_i = x/t$ (which is the formula currently recommended to calculate h_i [4]). However, if the individual has a different preference function, say $V = h_i T^a (a \not\approx 1)$, it would be incorrect to calculate the value of h_i as being equal to x/t. It is important to note that there is no empirical evidence to support the assumption that all (or most) individuals have a preference function of the functional form of $V = h_i T$. Neither are there appealing normative arguments (e.g. reflecting the discipline view of the world regarding how an individual should behave) why we should impose such preference patterns on individuals.

TTO: Some Empirical Observations

The TTO technique was offered as an empirical substitute to the SG method, and hence it is important to compare the two. From a theoretical perspective it is clear that the two techniques measure different

phenomena [10]. The two techniques have also been compared from an empirical perspective. Torrance [17] and Wolfson et al. [20] reported high **correlations** between health status ratings with SG and TTO. Read et al. [13] found a reasonable correlation between the SG and the TTO rating. However, they also found large systematic differences in the ratings obtained by the two methods. They concluded that the SG and TTO techniques produce different scale values for outcomes of clinical problems, which coincides with the claim made by Mehrez & Gafni [10] on theoretical grounds. In more recent studies, Patrick et al. [12] reported much lower correlations, and Hornberger et al. [7] report even lower correlation. Using other statistical methods, the studies by Stiggelbout et al. [16] and Nease et al. [11] report large discrepancies between TTO and SG scores.

With respect to the added assumptions about the functional form of the preference function, which are needed in the second step, it is important to ask whether they have any empirical support or normative appeal. The answer is No. An important assumption which is made is of a constant proportional tradeoff between years of life and health states. This implies that the weight which is used in the QALY calculation is constant. In other words, the value attached to any given health state is independent of the time spent in this health state. Two recent studies, Stiggelbout et al. [15] and Dolan et al. [3], found that this assumption was violated when tested. Another assumption is that the person has a zero time preference (i.e. s/he is indifferent between benefits (and costs) occurring at present or in the future). This assumption does not have normative appeal [2] or empirical support [6, 9].

Other Ways of Using the TTO Technique

For those who do not want to subscribe to the strong assumptions which are required at the second step, the TTO can still be used as a useful measure of outcome. If we just want to convert years in ill health to years in full health or lifetime health profiles (i.e. allowing individuals to be in different health states over their lifetime) to their equivalent in years in full health, then this can be done using the first step of the TTO method. As shown by Mehrez & Gafni [10], in the first step of the procedure no assumption is made (or necessary) about the functional form of the individual preference function. Hence, this step can be used to measure what Mehrez & Gafni call HYE (healthy years equivalent), i.e. the number of years in full health which is equivalent (preference-wise) to a given lifetime health profile. This measure combines both outcomes of quantity and quality of life and, like the QALY, can serve as a common unit of measurement for all programs, thus allowing comparisons across programs. It also maintains the intuitively appealing meaning of the QALY measure. However, using the TTO base HYE as a measure of outcome at the endpoints of decision trees requires additional assumptions [10].

An alternative version of the TTO technique has been suggested by [14]. This version is identical to the first step of the TTO technique and does not proceed to calculate a preference value for the health state in question. This method is not widely used in empirical studies reported in the literature. Furthermore, the authors do not provide any rationale for their suggestion.

Finally, the major limitation of the TTO technique is that it measures individuals' preferences under conditions of certainty [10]. Because decisions about health interventions at both the individual and the community levels are made under uncertainty [1], we need a measure of outcome that capture individuals' preferences under conditions of uncertainty. Recently, Johannesson [8] suggested a modified TTO question that will enable us to measure the HYE of a risky health profile. Following Johannesson [8], "... the risky health profile to be assessed is framed as a probability distribution and is equated to the certainty equivalent number of healthy years" (p. 47). This approach has not yet been tried empirically. Johannesson questions the practical feasibility of this approach and states that "it is unclear whether that type of information can be processed in a meaningful way" (p. 47). However, as Johannesson acknowledges, this is an empirical question and thus researchers should not be discouraged from trying this method in practice.

References

[1] Ben Zion, U. & Gafni, A. (1983). Evaluation of public investment in health care: is the risk irrelevant? *Journal of Health Economics* **2**, 161–165.

[2] Broome, J. (1993). QALYs, *Journal of Public Economics* **50**, 149–167.

[3] Dolan, P., Gudex, C., Kind, P. & Williams, A. (1996). The time trade-off method: results from a general population study, *Health Economics* **5**, 141–154.

[4] Drummond, M.F., Stoddart, G.L. & Torrance, G.W. (1987). *Methods for the Economic Evaluation of Health Care Programs*. Oxford Medical Publication, Oxford.

[5] Gafni, A. (1996). Proper preference based outcome measures in economic evaluations of pharmaceutical interventions, *Medical Care* **34**, DS48–DS58.

[6] Gafni, A. & Torrance, G.W. (1984). Risk attitude and time preference in health, *Management Science* **30**, 440–451.

[7] Hornberger, J.C., Redelmeier, D.A. & Petersen, J. (1992). Variability among methods to assess patients' well-being and consequent effect on a cost–effectiveness analysis, *Journal of Clinical Epidemiology* **45**, 505–512.

[8] Johannesson, M. (1995). The ranking properties of healthy-years equivalents and quality-adjusted life-years under certainty and uncertainty, *International Journal of Technology Assessment in Health Care* **11**, 40–48.

[9] Krahn, M. & Gafni, A. (1993). Discounting in the economic evaluation of health care interventions, *Medical Care* **31**, 403–418.

[10] Mehrez, A. & Gafni, A. (1990). Evaluating health related quality of life: an indifference curve interpretation for the time trade-off technique, *Social Science and Medicine* **31**, 1281–1283.

[11] Nease, R.F., Kneeland, T., O'Connor, G.T., Summer, W., Lumpkins, C., Shaw, L., Pryer, D. & Sox, H.C. (1995). Variation in patient utilities for outcomes of the management of chronic stable angina: implications for clinical practice guidelines, *Journal of the American Medical Association* **273**, 1185–1190.

[12] Patrick, D.L., Starks, H.E., Cain, K.D., Uhlmann, R.F. & Pearlman, R.A. (1995). Measuring preferences for health states worse than death, *Medical Decision Making* **14**, 9–18.

[13] Read, J.L., Quinn, R.J., Berwick, D.M., Fienberg, H.V. & Weinstein, M.C. (1984). Preferences for health outcomes: comparison of assessment methods, *Medical Decision Making* **4**, 315–329.

[14] Sox, H.C., Blatt, H.A., Higgins, M.C. & Marton, K.I. (1988). *Medical Decision Making*. Butterworth, Boston.

[15] Stiggelbout, A.H., Kiebert, G.H., Kievit, J., Leer, J.W.H., Habbema, J.D.F. & deHaes, J.C.J.M. (1995). The "utility" of the time trade-off method in cancer patients: feasibility and proportional trade-off, *Journal of Clinical Epidemiology* **48**, 1207–1214.

[16] Stiggelbout, A.M., Kiebert, G.H., Kievit, J., Leer, J.W.H., Stoter, G. & deHaes, J.C.J.M. (1994). Utility assessment in cancer patients: adjustment of time trade-off scores for the utility of life years and comparison with standard gamble scores, *Medical Decision Making* **14**, 82–90.

[17] Torrance, G.W. (1976). Social preferences for health states: an empirical evaluation of three measurement techniques, *Socio-Economic Planning Science* **10**, 128–136.

[18] Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal: a review, *Journal of Health Economics* **5**, 1–30.

[19] Torrance, G.W., Thomas, W.H. & Sackett, D.L. (1972). A utility maximization model for evaluation of health care programs, *Health Services Research* **7**, 118–133.

[20] Torrance, G.W., Wolfson, A., Sinclair, A., Bombardier, C. & McGeer, A. (1982). Preference measure for functional status in stroke patients: inter-rater and inter-technique comparisons, in *Values and Long Term Care*, R.L. Kane & R.A. Kane, eds. Lexington Books, Lexington.

(*See also* **Risk Assessment in Clinical Decision Making**)

AMIRAM GAFNI

Time-by-time Analysis of Longitudinal Data

Graphs such as Figure 1, and the associated analysis, are frequently found in reports of studies in which two groups are observed over time. The main features of the figure are:

1. Profiles based on the mean response at each time are shown.
2. Bars indicating some multiple of a standard error are displayed at each time.
3. The results of testing the equality of the group means at each time are indicated, usually by symbols, such as NS, *, ** for, respectively, $P > 0.05$, $P < 0.05$, $P < 0.01$.

Similar graphs and analogous analyses can be presented when more than two groups are involved. This kind of display, and the analysis it reports, has several serious drawbacks.

One problem is that the structure of the data is ignored: at no stage does this analysis use the information that indicates which observations are from the same individual. Consequently, the standard errors are based on between-subject variation, which for most purposes will be the wrong variance component. It is most unlikely that an analysis that ignores such an important feature of the data will be correct.

Another group of problems concerns the hypothesis tests. Since several tests are performed, the problems of multiple testing arise (*see* **Simultaneous Inference**); moreover, any attempt at interpretation will be complicated further because these tests, being based on the same individuals, are dependent. It is also highly questionable whether the hypotheses being tested; namely, the equality of the group means at each time, are of any interest. To use this collection of hypotheses to assess the equality of two or more curves is unnatural and unhelpful.

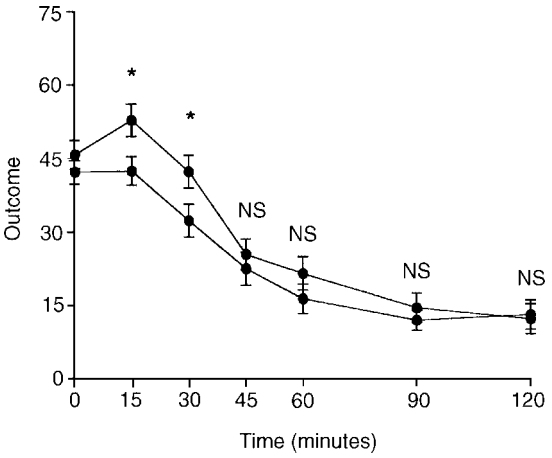

Figure 1 Typical presentation of time-by-time analysis of two groups observed over time

The profiles of the response in each group are summarized by the graph of means. There are applications where the graph of means is misleading, insofar as it may bear little resemblance to the profile of any individual. Whether it is reasonable in a particular application to summarize the response by the profile of means is, to some extent, a matter of judgment for the analyst. In making this judgment it is important that separate plots of the profile of each individual should have been studied.

(*See also* **Analysis of Variance for Longitudinal Data**; **Longitudinal Data Analysis, Overview**; **Summary Measures Analysis of Longitudinal Data**)

JOHN N.S. MATTHEWS

Time-dependent Covariate

Introduction

Regression models for survival data are frequently specified via the **hazard** function for the distribution of the survival time X (*see* **Survival Distributions and Their Characteristics**). Thus, the model specifies the conditional probability

$$\alpha(t \mid \mathbf{Z}) \approx \frac{\mathrm{P}(X \leq t + \mathrm{d}t \mid X > t, \mathbf{Z})}{\mathrm{d}t} \qquad (1)$$

of failing just after time t given survival till time t and given the **covariates**, $\mathbf{Z}$. Examples include the **Cox** proportional hazards **regression model** and Aalen's nonparametric **additive hazards model**.

In many cases, some of the covariates change over time.

1. In a medical follow-up study where t refers to time since start of treatment, the current age of the patients, and the current calendar time period are, obviously, time-dependent.
2. In an industrial life-testing experiment, the stress at which the components are tested may be designed to vary over the time period of the experiment.
3. In medical follow-up studies, events may happen to the patients under study in an unpredictable way, which may alter their **prognosis**.

An example of the latter kind is provided by the classical Stanford Heart **Transplantation** Study [17] where the hazard of patients waiting for a transplant (hopefully) changes once a transplantation is carried out.

The specification (1) lends itself to include time-dependent covariates by letting the hazard function at time t

$$\alpha(t \mid \mathbf{Z}(\cdot)) \approx \frac{\mathrm{P}(X \leq t + \mathrm{d}t \mid X > t, (\mathbf{Z}(u), 0 \leq u \leq t))}{\mathrm{d}t} \qquad (2)$$

depend on the entire covariate history over the time interval from 0 to t.

4. An example is a study of the effect of blood pressure on mortality where the blood pressure was recorded continuously and where at time $t(> \Delta t)$ the hazard was modeled to depend on the current blood pressure $Z(t)$ and the change in blood pressure $Z(t) - Z(t - \Delta t)$ over the preceding Δt hours.

Further examples of this type were discussed in [3, 24].

An important complication implied by the added generality of (2) over (1) has to do with the relation between the hazard function $\alpha(\cdot)$ and survival probabilities like

$$S(u \mid t) = \mathrm{P}(X \geq u \mid X \geq t, (\mathbf{Z}(s), 0 \leq s \leq t)), \quad u > t. \qquad (3)$$

In the fixed covariate model (1) this is simply given by

$$S(u \mid t) = \exp\left(-\int_t^u \alpha(s \mid \mathbf{Z})\mathrm{d}s\right), \qquad (4)$$

whereas for the model (2), $S(u \mid t)$ will depend on the stochastic structure of $(\mathbf{Z}(s), t \leq s \leq u)$. Kalbfleisch and Prentice [23, Section 6.3] introduced a classification of (time-dependent) covariates in survival analysis into *external* and *internal* ones. This classification is, in fact, identical to that used in econometrics where one distinguishes between *exogenous* and endogenous variables. One way of thinking of this classification is that external covariates are those for which it makes sense as in (3), at time t, to condition on the path of $\mathbf{Z}(\cdot)$ over the prediction interval from t to u, and internal covariates are those for which it does not make sense. Thus, external covariates include *time-fixed* covariates (constant from t to u), *defined* covariates whose path is fixed from t to u (e.g. current calendar time period in example 1. above), or *ancillary* covariates whose development over (t, u) is not influenced by the failure history of the individual (e.g. the stress level in the life-testing experiment of example 2. above or, as exemplified in [23, p. 197], the level of air pollution as a risk factor for the occurrence of respiratory diseases). In these examples, the survival probability is given by

$$S(u \mid t) = \mathrm{P}(X \geq u \mid X \geq t, (\mathbf{Z}(s); 0 \leq s \leq u)) = \exp\left(-\int_t^u \alpha(s \mid \mathbf{Z}(s))\mathrm{d}s\right). \qquad (5)$$

A mathematical condition on the joint distribution of X and $(\mathbf{Z}(t), t \geq 0)$ for (5) to hold was given in [36].

This may be taken as a formal definition of external covariates.

For *internal* covariates, however, as exemplified in examples 3 and 4 above, one cannot condition on $\mathbf{Z}(s)$ for $s \in (t, u)$ since the mere existence of $\mathbf{Z}(s)$ will imply that $X > s$. In this case, the survival probability is given by

$$S(u \mid t) = \mathrm{E} \exp\left(- \int_t^u \alpha(s \mid \mathbf{Z}(s)) \mathrm{d}s \right), \qquad (6)$$

where the expectation is taken with respect to the conditional distributions of $\mathbf{Z}(s)$ given $X \geq s$ and $\mathbf{Z}(v), 0 \leq v < s$ for $t \leq s < u$. The interpretation is that the (marginal) survival probability at u given the past up to t is the average over the possible paths among survivors for $\mathbf{Z}(s), t \leq s < u$.

For internal covariates, a joint model for X and $\mathbf{Z}(\cdot)$ is, therefore, in general, needed in order to calculate the survival probabilities in (3). Examples with discrete time-dependent covariates modeled as **Markov processes** were given in [4; 6, Section VII.2; 8] and examples with continuous time-dependent covariates in [21, 22, 33, 34, 37] (*see* **Joint Modeling of Longitudinal and Event Time Data**). An alternative approach for estimation of the marginal survival probabilities was discussed in [29].

Furthermore, effects of time-dependent covariates estimated in a model like (2) may only be interpreted as effects on the hazard function for given covariate histories and not as effects on the probability of survival. As a consequence, treatment effects from randomized **clinical trials** adjusted for the effect of a time-dependent covariate $Z(t)$ should be interpreted with great caution since $Z(t)$ may serve as an intermediate variable that may predict survival and whose development over time may be influenced by treatment in such a way that the treatment effect on the hazard function may be masked. (In an epidemiological setting, $Z(t)$ would *not* be a **confounder** variable in this case and should, therefore, *not* be adjusted for when evaluating the treatment effect.)

The Cox Regression Model

The concept of time-dependent covariates was introduced by Cox [14] in the fundamental paper on the **proportional hazards** model where

$$\alpha(t \mid \mathbf{Z}) = \alpha_0(t) \exp(\beta' \mathbf{Z}). \qquad (7)$$

Here, β is a vector of unknown regression coefficients and $\alpha_0(t)$, the baseline hazard, is the hazard function for individuals with $\mathbf{Z} = \mathbf{0}$. Cox's original use of time-dependent covariates was to test the basic assumption in (7) by adding a *defined*, and hence exogenous, time-dependent covariate, for example, $Z_1 \cdot t$, to the time-fixed covariates already in the model and examining whether the corresponding regression coefficient is zero. This is a powerful way of testing for proportional hazards and an important application of time-dependent covariates. Only later were models studied where the effect of certain internal time-dependent covariates was of scientific interest. Early examples are the analyses of the Stanford Heart Transplantation Data [17] and the study [20] of multiple infections. A review was given in [5].

Estimation of β is based on the **Cox partial likelihood**

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' \mathbf{Z}_i(T_i))}{\sum_{j \in R(T_i)} \exp(\beta' \mathbf{Z}_j(T_i))} \right)^{D_i}, \qquad (8)$$

[15] where $T_1, \ldots, T_n$ are independent times of observation (T_i being the failure time, X_i, if $D_i = 1$ and a right-**censoring** time if $D_i = 0$) and $R(t) = \{i : T_i \geq t\}$ is the **risk set** at time t. Furthermore, the integrated baseline hazard $A_0(t) = \int_0^t \alpha_0(u) \mathrm{d}u$ is estimated by the Breslow estimator [12]

$$\widehat{A_0(t)} = \sum_{T_i \leq t} \frac{D_i}{\sum_{j \in R(T_i)} \exp(\beta' \mathbf{Z}_j(T_i))}. \qquad (9)$$

The large sample properties of $(\widehat{\beta}, \widehat{A_0(\cdot)})$ were derived in [7]. (Note that the integrated hazard $A_0(t)$ and the estimator (9) are both well-defined in the presence of time-dependent covariates, even though the transformation $S_0(t) = \exp(-A_0(t))$ may not be interpreted as a survival probability; see the Introduction above.

It is seen that in order to compute (8) and (9), the covariate values $\mathbf{Z}_j(T_i)$ for individuals j at risk at time T_i are needed. If, for example, $\mathbf{Z}_j(t)$ is observed as repeated measurements over time for individual j, then the value $\mathbf{Z}_j(T_i)$ may not be measured. If $\mathbf{Z}_j(T_i)$ is (naively) replaced by the latest observation available for individual j before time T_i, then the

estimated coefficients based on (8) tend to be biased [10, 31]. Alternative methods use interpolation on the basis of smoothing of the observed series of repeated measurements of $\mathbf{Z}(\cdot)$ [18, 27, 32] or joint models for X and $\mathbf{Z}$ [21, 22, 34, 35] (*see* **Joint Modeling of Longitudinal and Event Time Data**). A conditional score approach was introduced in [31] while some simpler methods were discussed in [10, 19].

To compute the denominators in (8) and (9) for models with *time-fixed* covariates, a very simple **algorithm** is available. Going backwards in time from the final risk set $R(+\infty) = \emptyset$, one simply adds $\exp(\beta'\mathbf{Z}_j)$ when passing T_j (and if the survival times are left-**truncated**, one subtracts $\exp(\beta'\mathbf{Z}_j)$ when passing the entry time V_j). In models with time-dependent covariates, however, one has to recalculate each sum from scratch and this increases the computing time considerably. This, among other things has led to suggestions to replace the risk set $R(T_i)$ in these sums by some subset of the risk set usually including the individual, i, failing at that time [11, 26, 30]. Such *nested case–control designs* may save computing time and (other resources) without losing much efficiency (*see* **Case–Control Study, Nested**).

Other Regression Models

Aalen [1, 2] introduced and studied a nonparametric regression model with time-dependent effects of covariates and which can also readily take time-dependent covariates into account:

$$\alpha(t) = \beta_0(t) + \beta(t)'\mathbf{Z}(t). \tag{10}$$

Nonparametric estimation of the integrated regression functions $B_j(t) = \int_0^t \beta_j(u)\mathrm{d}u$ are fairly straightforward generalizations of the **Nelson–Aalen estimator**. A simpler version of (10) where $\boldsymbol{\beta}(t)$ is constant was discussed in [25].

The **accelerated failure-time model** for time-fixed covariates is usually defined by

$$\log X = -\beta'\mathbf{Z} + \varepsilon, \tag{11}$$

where ε is an error term and is, thus, not given by its hazard function. If U is a **random variable** distributed as $\exp(\varepsilon)$, that is, as the lifetime of an individual with $\mathbf{Z} = \mathbf{0}$, then, in distribution, $X = \exp(-\beta'\mathbf{Z})U$, or

$$U = \int_0^X \exp(\beta'\mathbf{Z})\mathrm{d}t. \tag{12}$$

On the basis of this, one may study a generalization of (11) allowing for time-dependent covariates by assuming (in the case of no censoring) $U_1, \ldots, U_n$, where

$$U_i = \int_0^{X_i} \exp(\beta'\mathbf{Z}_i(t))\mathrm{d}t, \tag{13}$$

to be i.i.d. A parametric model of this kind was studied in [16, Section 5.2] and a **semiparametric** model with the distribution of U completely unspecified and with estimation based on **linear rank tests** in [28].

Several Timevariables

In the models studied so far, the hazard function has been assumed to depend on a single timevariable but in many examples, more than one time origin may be of relevance (*see* **Time Origin, Choice of**). Thus, one may wish to consider time on study and age, age and calendar time, time on study and time since a given event, and so on.

When using a Cox regression model in such cases, it is necessary to consider one timevariable as the "basic" timescale and to model the effect of other timescales using time-dependent covariates. As an example, we may consider the three-state illness-death model with states 0: healthy, 1: diseased, and 2: dead (*see* **Fix–Neyman Process**). If the three transition intensities $\alpha_{01}(\cdot), \alpha_{02}(\cdot)$, and $\alpha_{12}(\cdot)$ all depend only on a given time, t, the modeled process is **Markov**, but if $\alpha_{12}(\cdot)$ also depends on the duration $d = t - T_{01}$ of time spent in state 1 the modeled process is **semi-Markov** and this may be modeled by including the duration d or some function of it as a time-dependent covariate. A model of this type may also be used for testing the Markov hypothesis in the obvious way. Examples were provided in [6, Section X.1].

In a **Poisson regression** model with a piecewise constant hazard function [13], several timevariables may be modeled simultaneously in a simple way. Dependence on both age and duration may be modeled by assuming the hazard function to be constant in "cells" given by a partition of age and duration,

thus treating the two timescales in "parallel" without considering one of them as basic. Criteria for choosing between Cox and Poisson regression models were discussed in [9, 13].

References

[1] Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes, *Springer Lecture Notes in Statistics* **2**, 1–25.

[2] Aalen, O.O. (1989). A linear regression model for the analysis of life times, *Statistics in Medicine* **8**, 907–925.

[3] Altman, D.G. & Stavola, B.L. (1994). Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates, *Statistics in Medicine* **13**, 301–341.

[4] Andersen, P.K. (1986). Time-dependent covariates and Markov processes, in *Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice, eds. John Wiley & Sons, New York, pp. 82–103.

[5] Andersen, P.K. (1992). Repeated assessment of risk factors in survival analysis, *Statistical Methods in Medical Research* **1**, 75–93.

[6] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[7] Andersen, P.K. & Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.

[8] Andersen, P.K., Hansen, L.S. & Keiding, N. (1991). Non- and semi-parametric estimation of transition probabilities from censored observations of a non-homogeneous Markov process, *Scandinavian Journal of Statistics* **18**, 153–167.

[9] Andersen, P.K. & Keiding, N. (2002). Multi-state models for event history analysis, *Statistical Methods in Medical Research* **11**, 91–115.

[10] Andersen, P.K. & Liestøl, K. (2003). Attenuation caused by infrequently updated covariates in survival analysis, *Biostatistics* **4**, 633–649.

[11] Borgan, O., Goldstein, L. & Langholz, B. (1995). Methods for analysis of sampled cohort data in the Cox proportional hazards model, *Annals of Statistics* **23**, 1749–1778.

[12] Breslow, N.E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–99.

[13] Clayton, D.G. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

[14] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[15] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[16] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[17] Crowley, J.J. & Hu, M. (1977). Covariance analysis of heart transplant survival data, *Journal of the American Statistical Association* **72**, 27–36.

[18] Dafni, U.G. & Tsiatis, A.A. (1998). Evaluating surrogate markers of clinical outcome measured with error, *Biometrics* **54**, 1445–1462.

[19] DeBruijne, M.H.J., LeCessie, S., Kluin-Nelemans, H.C. & van Houwelingen, H.C. (2001). On the use of Cox regression in the presence of an irregularly observed time-dependent covariate, *Statistics in Medicine* **20**, 3817–3829.

[20] Farewell, V.T. (1979). An application of Cox's proportional hazard model to multiple infection data, *Applied Statistics* **28**, 136–143.

[21] Faucett, C.L. & Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach, *Statistics in Medicine* **15**, 1663–1685.

[22] Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data, *Biostatistics* **1**, 465–480.

[23] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. Wiley, New York.

[24] Liestøl, K. & Andersen, P.K. (2002). Updating of covariates and choice of time origin in survival analysis: problems with vaguely defined disease states, *Statistics in Medicine* **21**, 3701–3714.

[25] Lin, D.Y. & Ying, Z. (1994). Semi-parametric analysis of the additive risk model, *Biometrika* **81**, 61–72.

[26] Oakes, D. (1981). Survival times: aspects of partial likelihood (with discussion), *International Statistical Review* **49**, 235–264.

[27] Raboud, J., Reid, N., Coates, R.A. & Farewell, V.T. (1993). Estimating risks of progressing to AIDS when covariates are measured with error, *Journal of the Royal Statistical Society, Series A* **156**, 393–406.

[28] Robins, J. & Tsiatis, A.A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates, *Biometrika* **79**, 311–319.

[29] Satten, G.A., Datta, S. & Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates, *Statistics and Probability Letters* **54**, 397–403.

[30] Thomas, D.C. (1977). Addendum to: methods of cohort analysis: appraisal by application to asbestos mining, By F. D. K. Liddell, J. C. McDonald and D. C. Thomas, *Journal of the Royal Statistical Society, Series A* **140**, 469–491.

[31] Tsiatis, A.A. & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error, *Biometrika* **88**, 447–458.

[32] Tsiatis, A.A., DeGruttola, V. & Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival data and CD4 counts in patients with AIDS, *Journal of the American Statistical Association* **90**, 27–37.

[33] Woodbury, M.A. & Manton, K.G. (1977). A random walk model for human mortality and aging, *Theoretical Population Biology* **11**, 37–48.

[34] Wulfsohn, M.S. & Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error, *Biometrics* **53**, 330–339.

[35] Xu, J. & Zeger, S.L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events, *Applied Statistics* **50**, 375–387.

[36] Yashin, A. & Arjas, E. (1988). A note on random intensities and conditional survival functions, *Journal of Applied Probability* **25**, 630–635.

[37] Yashin, A.I., Manton, K.G. & Stallard, E. (1986). Dependent competing risks: a stochastic process model, *Journal of Mathematical Biology* **24**, 119–140.

(*See also* **Survival Analysis, Overview**)

PER KRAGH ANDERSEN

Time-varying Treatment Effect

Comparing the effects of fixed, time-invariant treatments on a single outcome is basic to biostatistics. In this article we are concerned with the extension to the time-varying situation, i.e. the way that treatments that vary in time within an individual affect outcomes (effects of time-varying treatments). For example, consider the Lipid Research Clinics Coronary Prevention Trial [8], a placebo-controlled, double-blind, randomized trial (*see* **Clinical Trials, Overview**) of the efficacy of cholestyramine for reducing heart disease by lowering the level of cholesterol. The usual approach treats this as a comparison of fixed strategic decisions (treat with cholestyramine or placebo; *see* **Blinding or Masking**), reflected in the **randomization**, and statistical analysis follows the principle of **"intention-to-treat"**. This approach stays close to the experimental basis for **inference** about causal effects of treatment (randomization) and therefore subsumes "noncompliance" or other changes in treatment after randomization under the heading of the "pragmatic" effects of the strategies: "try to give cholestyramine" vs. "try to give placebo". This is not always completely satisfactory to investigators, who may want to know more about the effects of levels of compliance with the active treatment. This involves temporal variation in the treatment, and thus can be seen as a realistic generalization of the causal model that motivates the usual design and analysis of the RCT (*see* **Causation**; **Compliance Assessment in Clinical Trials**; **Pharmacoepidemiology, Adverse and Beneficial Effects**).

For example, an investigator might want to know what would have happened if all patients had been 100% compliant with medication assignment for the entire duration of the study. Efron & Feldman [2] (with commentary) discuss the problem of estimation of the **dose–response** relationship from such a clinical trial with uncontrolled compliance. Another investigator might try to estimate the effects on current cholesterol levels of varying periods of treatment with cholestyramine, or varying total cumulative dose. These are examples of the effects of time-varying treatments. Once the analysis departs from its basis in the randomization, many difficulties arise (see below).

To define these effects more precisely, we introduce some notation. Consider a population of patients U, a set of alternative treatments R, and a collection of possible treatment sequences specifying treatments up to time $s < S$ (an arbitrary upper time horizon):

$$Z(s) = \{z(t)\}_{0 \le t \le s < S}, \quad z(t) \in R. \tag{1}$$

To fix ideas, let U be the population of patients with depression and R be the set of possible doses of antidepressant drug that could be prescribed for some interval. Then, at each time $0 \le t < S$, $z(t)$ specifies the treatment to be given during the intervals $(t, t+1]$, and $Z(s)$ is one particular sequence of treatments specified up to time s. Treatment sequences without arguments define specific treatments for the entire span of time up to S (Z, for example). To define causal effects it is necessary to make a distinction between the set of possible treatment sequences and the one that is actually realized for a given patient. The notation is sufficiently general for realistic purposes, since treatment decisions are effectively discrete time processes, possibly with very short time intervals in some critical care situations.

Let the sequence of outcomes that would be realized up to time s if patient u received treatment sequence Z be denoted by

$$Y_Z(s, u) = \{y_{t,Z}(u)\}_{0 \le t \le s \le S}. \tag{2}$$

For example, each $y_{t,Z}(u)$ could be a vector of symptom scores, **quality of life** scores, and side-effect scores, that would have been observed at time t. If it is necessary to distinguish them from the outcomes, then **covariate** histories $X(s)$ can be specified as well. This generalization to time-varying treatments is implicit in the "potential outcomes" framework due to Rubin [13–15], who explicitly treats such important subtleties as the "stable unit treatment value assumption" (that the response by one patient to a treatment does not depend on the treatment assignments of the other patients). Robins [10, 12] and Lavori et al. [6, 7] exploit these ideas in the context of longitudinal treatments.

We suppose that the treatment decision $z(t)$ specifies the treatment received by the patient for the interval of time starting just after time t, taking effect *after* the measurement of the outcome y_t. Considerations

of causality then make it plausible to assume that the outcomes up to and including time s depend on the given treatment sequence only through its values up to time $s-1$. Thus, if treatment history Z_1 agrees with Z_2 up to and including time $s-1$, then $Y_{Z_1}(s,u) = Y_{Z_2}(s,u)$ and notation such as $Y_{Z(s)}(k,u)$ is well defined for $k \leq s+1$.

Then, following Rubin, define individual causal comparisons of the effects of pairs of treatments Z_1, Z_2 as functionals of $Y_{Z_1}(S,u), Y_{Z_2}(S,u)$. For example, suppose that $y(t)$ is a score measuring severity of depressive symptoms at time t, and Z_1, Z_2 are, respectively, the treatment sequences $\{1,1,1,\ldots,1\}$ and $\{1,1,1,\ldots,1,0,\ldots,0\}$, where there are S 1s in the first sequence, denoting constant treatment with antidepressant, and $s_0 < S$ 1s in the latter sequence. Then we can define one particular individual causal effect (for patient u) of continuous treatment vs. dropping treatment after time s_0 as the difference in time-averaged scores $(1/S)\sum_t y_{t,Z}(u)$ under the particular antidepressant treatment schedules $Z = Z_1$ or $Z = Z_2$.

The problem of inference from observed data arises (just as it does in the fixed-treatment situation) because it is only possible to observe $Y_Z(S,u)$ if Z is the actual realized treatment sequence in patient u. Holland [4] refers to this as the "central problem" of causal effects inference. The statistical approach attempts to estimate some kind of average causal effect (over U) by making assumptions about the nature of the actual assignment of patients to treatment sequences. For example, at time s, the clinician considers the patient's current and past state of depression, level of side-effects of the medication, social functioning, etc. observed under the treatment sequence defined up to the previous assignment (at $s-1$), and then assigns the dosage $z(s)$ that will be taken for the next interval. What do we need to assume about this process to estimate causal effects from the observed outcomes?

Ignorability of Treatment

In the fixed-treatment problem, Rubin [13–15] has defined the requirements for using observed responses to realized treatment conditions to make valid estimates and inferences about average causal effects. These are the "ignorability" conditions, which (for fixed treatments) are guaranteed by randomization of treatments. The crux of the matter is that the treatment actually received should be independent of the potential responses to the treatments being compared (more generally, conditionally independent given measured pretreatment covariates). Generalization of the ignorability conditions to the time-varying case raises a new issue, concerning the intermediate outcomes that may be used as "covariates" to determine subsequent treatment decisions. One omnibus ignorability condition is the following:

$$\begin{aligned}
&\text{for all times } s \text{ and all outcomes } Y(s) \text{ realized up to } s,\\
&\text{for all treatment sequences } Z_0(s-1) \text{ defined up to } s-1,\\
&\text{for all possible treatments } z_1(s), z_2(s), \text{ received at } s,\\
&\text{and for all treatment sequences } Z \text{ agreeing with } Z_0(s-1) \text{ up to } s-1,\\
&\text{for all } t \geq s, \Pr[Y_Z(t,u) | Y_{Z_0(s-1)}(s), Z_1(s)]\\
&\quad = \Pr[Y_Z(t,u) | Y_{Z_0(s-1)}(s), Z_2(s)].
\end{aligned} \tag{3}$$

In words, given any history of treatments and outcomes observed by time s, subgroups defined by the choice of treatment in the interval $(s, s+1]$ have the same distribution of potential future outcomes on any treatment strategy that agrees with the observed treatment history up to $s-1$ and extends it arbitrarily into the future. This would be satisfied if patients were subclassified on treatment and outcome history at s and then randomly assigned to treatment during $(s, s+1]$. Here we sweep all covariates into Y, to avoid unnecessary notation. The idea is that each alternative assignment of treatments during $(s, s+1]$ produces "comparable" patient groups, in the sense of having the same distribution of potential responses to treatment, across all possible treatments determined by arbitrarily extending the observed treatment assignment past s. This assumption is unobservable, and can only be *known* to be true under random assignment, possibly stratified by observables (*see* **Stratification**).

Treatment Strategies and Decision Rules

Fixed treatments can also have effects that apparently vary with time (time-varying effects of treatment).

This is not the focus of the current article, but it is useful to contrast it with the concept of effects of time-varying treatment, since they can easily be confused. Consider the comparison of survival of patients with heart disease under two alternative strategies: (i) immediate surgical treatment vs. (ii) medical treatment at first and then surgical treatment if the patient worsens. The **hazard ratio** may favor medical treatment at first, because of surgical complications, but then the balance may shift over time in favor of surgery. It is commonplace that "all survival curves cross". In our notation, there is only one occasion of treatment decision, defining $z(0)$.

The natural longitudinal extension of the fixed treatment is the fixed strategy, or decision rule. For example, Gelenberg et al. [3] studied the effects on relapse rates in patients in remission from manic depression of two treatment strategies: adjust the dose of lithium carbonate to achieve either a "standard" serum lithium level or a "low" level. The actual prescribed doses of lithium varied in response to the patient's current serum levels, to steer within the target ranges, but the treatment decision was fixed at the outset, and is thus not time-varying in our sense.

In principle, any set of longitudinal treatment strategies can be specified in advance, patients randomized to all of them, and then the outcomes measured and compared. But this breaks down quickly in practice, due to the complexity of possible decision rules. An important exception, when treatments are assigned without regard to previous outcomes, is the standard **crossover design**.

Crossover studies involve explicit use of the variation in treatments within an individual over time to reduce the interindividual variation in response. In the standard crossover study, treatment sequence assignments are randomized, and thus ignorable (barring missing data, dropouts, etc.) The aim is to compare the effects of fixed treatments, and the effects of the longitudinal component of the variation in treatment (carryover, period, and treatment by period **interaction**) are threats to the validity of the analysis rather than explicit targets of estimation. The usual analysis of the two-period two-treatment crossover study tests for the presence of treatment by period interaction effects, and then, given no rejection of the null, goes on to analyze the data as if the order of treatments were immaterial [5]. Thus, the time-varying part of the treatment effect is suppressed. More elaborate and **powerful** designs, with multiple crossovers, essentially suppress higher-order interactions to make better use of the within-individual variation in treatment, but the target is still an effect estimate in the context of a Markovian hypothesis about the irrelevance of some or all of the past history of treatment (*see* **Markov Processes**).

Observational Studies in Medicine

In medicine, simple designs with fixed treatments dominate the experimental literature, partly because of the extreme difficulty in controlling complex patterns of treatments over time in patients whose clinical state demands constant attention and commands instant revision of the individual's treatment protocol when side-effects or worsening exceed tolerance. Represent the clinical decision-making process (*see* **Decision Analysis in Diagnosis and Treatment Choice**) as an iteration of the following single step: measure current outcomes and other covariates, recall past outcomes and treatments, and, on the basis of these facts, determine the next treatment choice by a particular fixed decision rule. No controlled trial can compare many complex versions of the decision rule applied at each step, with randomization among the alternatives. Therefore, time-varying treatment effects appear in nonexperimental studies (*see* **Observational Study**), and in the secondary analysis of data from simple randomized studies when investigators try to define the effects of noncompliance, uncontrolled adjuvant treatments, and other naturalistic complications of controlled experiments.

The study of the effects of time-varying naturalistic treatment is formally identical to the study of the effects of environmental exposures (*see* **Environmental Epidemiology**) that vary in time, and this links the growing literatures in both medical and epidemiologic fields [10]. Since epidemiologists are usually concerned with exposures that are not the result of planned experiments, issues of **confounding** with **time-dependent covariates** are central, including variables that are in the causal pathway from exposure to outcome (intervening variables). An important difference between the epidemiologic and medical treatment contexts is that few epidemiologic exposures are "selected" in order to produce a therapeutic effect. This has

enormous consequences for the likelihood that there are unmeasured **confounders**.

Adjustment for Confounders

There is little controversy over the principles governing adjustment for covariates that are measured prior to treatment (or exposure). In contrast, adjustment for time-varying confounders in the longitudinal setting can yield causally incorrect interpretations (see below). Careful statement of the goals and assumptions is critical to avoiding pitfalls of inference. Mark & Robins give a detailed discussion of this point [9].

For example, in analyzing the effect of dropping antidepressant treatment in patients with major depression in remission, one may suspect that the decision to drop treatment may be influenced by the patient's current state of symptoms, which may vary enough to convince the patient (or clinician) to continue treatment but not enough to qualify as a full relapse. However, adjusting for the patient's current level of symptoms without simultaneously stratifying on the history of treatment can distort the estimation of causal effects on full relapse if the current level of symptoms predicts future relapse and also future treatment, and past treatment predicts current symptoms.

Because of the importance of qualitative changes in health state, such as death, onset of disease, and recurrence of acute illness, survival and other "time-to-event" studies predominate in medicine and epidemiology. Here we concentrate on examples of time-varying treatments in a "survival" setting (including circumstances where the event is not fatal). The survival context offers a substantial statistical simplification: the total probability of the individual trajectory can be factored into a sequence of conditional survival probabilities, with no need to consider the past history of events. This can be generalized by counting process techniques (*see* **Counting Process Methods in Survival Analysis**) to cover situations with reentrance into the risk set [1].

Suppose that measurements of survival and covariates are available, and decisions about a dichotomous treatment are taken, at discrete times 0 (baseline), 1, 2,..., S. Specializing the notation from the first section above to a dichotomous treatment situation, and separating the notation for the survival outcome and possible confounders, let

$$z(t) = \begin{cases} 1, & \text{if patient is treated during } (t, t+1], \\ 0, & \text{otherwise;} \end{cases} \tag{4}$$

$$Z(s) = \{z(t)\}_{0 \le t \le s < S},$$

so that $Z(s)$ defines a particular treatment history. Suppose that each individual u has a set of survival times – one for each possible treatment history – so that $T_{u,Z}$ is the survival time for individual u that would be observed if that individual followed treatment history Z. Note that this is a summary of the sequence of dichotomous indicators of survival at each time: $Y_Z(S, u) = \{y_{t,Z}(u)\}_{0<t\le S}$. (Assume for simplicity that S is larger than any survival time.) Then the obvious individual causal effect of treatment Z_1 vs. Z_2 is just $T_{u,Z_1} - T_{u,Z_2}$, and population average causal effects are functionals of the corresponding survival distributions over U:

$$S_{Z_j}(t) = \Pr(T_{Z_j} > t), \quad j = 1, 2. \tag{5}$$

In the depression maintenance example used above, the "event" could be "relapse into depression", and the treatment dichotomy could be defined by whether the patient continued to receive antidepressant medications during the interval. Some obvious comparisons include "always treat" vs. "never treat" or "treat up to k and then stop". Most possible sequences would be irrelevant or uninteresting; we usually are interested in sequences that are realized naturalistically by many patients.

To introduce possible confounders, let $x(t)$ and $X(t)$ be, respectively, the value of the covariate (vector) measured at time t and the entire history of the covariate up to and including time t. Robins [10] generalizes Rubin's "ignorable treatment assignment" in the following way, stated informally: for each treatment regimen defined up to but not including time s, patients who have survived that regimen and are observed to have the same covariate history up to and including s fall into two groups defined by the actual treatment assignment at time s (treated or not). The assignment is ignorable if the two groups are "comparable" in the sense that their survival distribution from s on would be the same under every treatment history that extends the regimen observed up to but not including s. That is, not only are the groups similar in the unobservable probability of

surviving through $s + 1$, on both the treatment they received and the other option, but also with respect to survival on any future version of the treatment. In symbols,

$$\begin{aligned}&\Pr[T_Z > k | Z_0(s-1), X(s), z(s) = 1] \\ &\quad = \Pr[T_Z > k | Z_0(s-1), X(s), z(s) = 0]\end{aligned} \tag{6}$$

for all Z agreeing with $Z_0(s-1)$, up to $s - 1 < k$.

Robins [10] shows that, under this set of assumptions, causal effects can be calculated by summing (over all possible covariate histories) the joint survival and covariate probabilities of each covariate history, under the two treatment histories being compared. Of course, one must have observed these histories.

Mark & Robins [9] point out that, even with these assumptions and a well-specified time-varying **Cox regression model** for survival given treatment and confounders, the model-based test of the treatment **null hypothesis** does not necessarily test the causal null hypothesis that for each individual the survival times are identical across treatment histories. The example they give involves a hypothetically impotent treatment which has a **correlation** with the discrepancy between the measured confounder and the actual prognostic (latent) variable that it purports to measure. Robins et al. [12] describe an example involving prophylaxis for the intervening variable "onset of pneumocystis carinii infection" in prevention of death from **AIDS**. Lavori et al. [6] investigate a more restricted set of treatment alternatives, defined by the time of cessation of antidepressant therapy, and a "prompt" treatment effect, and find that under these circumstances the time-varying Cox regression model gives valid causal inferences. Robins has also noted that prompt effects can be estimated with the Cox model. Thus, the causal interpretation of the "effect" estimates from the time-varying Cox model depends on more than just the correct specification of the model for the observed data. This is a lively area that should see much activity in the coming years.

Structural Models for Treatment Effect

In the sociological literature, **structural equation models** (*see* **Path Analysis**) are routinely used to model panel data (*see* **Panel Study**) with multiple independent and dependent variables. Such models can be used to analyze time-varying treatment effects, although the assumptions necessary for identification of the parameters include ignorability assumptions that are hidden from view by the structural formulation. Holland [4] contrasts the typical structural equation setup for causal inference with the experiment-based setup described above, which is most familiar to biostatisticians (in its fixed treatment form), and which has come to be known as "Rubin's causal model".

Closely related to the structural equation models used by the sociologists are similar methods used by economists. In particular, **instrumental variables** approaches are often used in econometrics to analyze the structural effects of decisions, inputs, or other variables, which often vary with time. Issues of identification are paramount.

Robins & Tsiatis [11] have proposed an approach to the estimation of causal effects on survival that relies on a structural parametric model for the longitudinal causal effects (an **accelerated failure-time model**). In this model, the strong parametric assumptions make it possible to infer the time to failure of an individual under every possible assignment to treatment, given the observed time to failure under the actual treatment received, and the unknown value of the coefficients of the model. Thus, the parameters have direct causal interpretations. Of course, the method relies on the correct specification of the parametric model for unobservable effects, which is stronger than the usual assumption that the observables are correctly modeled. It will be useful to have several such methods available, to test the sensitivity of conclusions to the specific structural assumptions. This area is currently a focus of intense effort.

References

[1] Aalen, O.O., Borgan, O., Keiding, N. & Thormann, J. (1980). Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring, *Scandinavian Journal of Statistics* **7**, 161–171.

[2] Efron, B. & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials, *Journal of the American Statistical Association* **86**, 9–17.

[3] Gelenberg, A.J., Kane, J.M., Keller, M.B., Lavori, P.W., Rosenbaum, J.F., Cole, K. & Lavelle, J. (1989). Comparison of standard versus low serum levels of lithium for maintenance treatment of bipolar disorder, *New England Journal of Medicine* **321**, 1489–1493.

[4] Holland, P.W. (1981). Statistics and causal inference, *Journal of the American Statistical Association* **81**, 945–963.

[5] Jones, B. & Kenward, M.G. (1989). *Design and Analysis of Cross-over Trials*. Chapman & Hall, London.

[6] Lavori, P.W., Dawson, R. & Mueller, T.I. (1994). Causal estimation of time-varying treatment effects in observational studies: application to depressive disorder, *Statistics in Medicine* **13**, 1089–1100.

[7] Lavori, P.W., Keller, M.B., Sheftner, W., Fawcett, J., Mueller, T.I. & Coryell, W. (1994). Recurrence after recovery in unipolar MDD: An observational follow-up study of clinical predictors and somatic treatment as a mediating factor, *International Journal of Methods in Psychiatric Research* **4**, 211–229.

[8] Lipid Research Clinic Program (1984). The Lipid Research Clinic Primary Prevention Trial Results, Parts I and II, *Journal of the American Medical Association* **251**, 351–374.

[9] Mark, S.D. & Robins, J.M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model, *Statistics in Medicine* **12**, 1605–1628.

[10] Robins, J.M. (1989). The control of confounding by intermediate variables, *Statistics in Medicine* **8**, 679–701.

[11] Robins, J.M. & Tsiatis, A.A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models, *Communications in Statistics – Theory and Methods* **20**, 2609–2631.

[12] Robins, J.M., Blevins, D., Ritter, G. & Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients, *Epidemiology* **3**, 319–336.

[13] Rubin, D. (1974). Estimating causal effects of treatment in randomized and non-randomized studies, *Journal of Educational Psychology* **66**, 688–701.

[14] Rubin D. (1978). Bayesian inference for causal effects: the role of randomization, *Annals of Statistics* **6**, 34–58.

[15] Rubin, D. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism, *Biometrics* **47**, 1213–1234.

(*See also* **Predictive Modeling of Prognosis**)

PHILIP W. LAVORI & REE DAWSON

Tolerance Interval

Tolerance intervals are statistical intervals that *contain* (or *cover*) at least a proportion β of a population, either on average, or else with a stated *confidence*, γ. Tolerance intervals (and in the multivariate case **tolerance regions**) summarize uncertainty about values of a **random variable**, usually a future observation. This should be contrasted with **confidence intervals** and regions, which provide confidence statements about uncertainty in unknown constants (parameters). Tolerance intervals are sometimes referred to as tolerance bounds or tolerance limits. Prediction intervals and confidence intervals on **quantiles** can be regarded as special cases of tolerance limits.

The statistical theory of tolerance intervals, although conceived to address problems in manufacturing [31, 38], has numerous applications to problems in biostatistics. Some examples include clinical chemistry [9, 30] (*see* **Normal Clinical Values, Reference Intervals for**) and **bioequivalence** [3]. Tolerance interval theory parallels that of confidence intervals in most respects. There is a theory of nonparametric tolerance intervals, a well-developed normal theory, a less complete theory for other parametric families, as well as tolerance limits for linear models. We review some of these areas in this article, with an emphasis on normal theory – which, because of its flexibility, is most important in applications.

Definitions

Let $Y_1, \ldots, Y_n$ denote a sample from a probability distribution with distribution function F. A (β, γ) two-sided β-content tolerance interval is a statistical interval $[L(Y_1, \ldots, Y_n), U(Y_1, \ldots, Y_n)]$ for which

$$\Pr[F(U) - F(L) \geq \beta] \geq \gamma. \tag{1}$$

The constants β and γ are referred to as the *content* (or *coverage*) and the *confidence*, respectively. A two-sided β-expectation interval satisfies

$$\mathrm{E}[F(U) - F(L)] \geq \beta;$$

that is, it has an expected content of at least β. As was first noted by Paulson [26], β-expectation tolerance limits are equivalent to prediction intervals for a future observation.

We define one-sided intervals in the obvious way. A (β, γ) lower tolerance limit is a statistic $L(Y_1, \ldots, Y_n)$ such that

$$\Pr[F(L) \leq 1 - \beta] \geq \gamma.$$

That is, at least a proportion β of the distribution of F exceeds L with confidence at least γ. A (β, γ) lower tolerance limit is thus a lower confidence limit on the $1 - \beta$ quantile of F, with a confidence coefficient of at least γ. Upper tolerance limits are defined similarly.

General References and Bibliographies

The only book-length treatment of statistical tolerance intervals in English is a monograph by Guttman [4], which is particularly useful for its exposition of nonparametric tolerance region theory, and for its discussion of normal and **Bayesian** theory for simple **random samples**. Tolerance intervals are also discussed in some detail in books by Aitchison & Dunsmore [1, Chapters 5 and 6] and, more recently, by Hahn & Meeker [6]. The encyclopedia article by Guttman [5] is also useful for nonparametric, normal, and Bayesian theory for simple random samples, and the article by Noether [20] discusses nonparametric intervals.

The most extensive bibliography of the literature up to 1989 has been published by Jilek, in two parts [10, 11]. Review articles by Patel on tolerance intervals [24] and prediction intervals [25] are particularly useful for their coverage of univariate theory for various parametric families.

An Example: Serum Glucose Measurements

In the remainder of this article we discuss tolerance limits for simple random samples, and briefly review tolerance limits for linear models. We use the following numerical example to illustrate various statistical methods.

At the National Bureau of Standards, measurements were made using isotope dilution/mass spectrometry (ID/MS) of the concentration of glucose in frozen bovine serum. Glucose concentration in frozen serum tends to decrease with time, so these measurements were made on several days over a period of

Table 1 ID/MS concentration of glucose in frozen bovine serum (mg/l)

Serum age (days)				
1	8	15	140	290
1980	1979	1988	1957	1940
1981	1982	1987	1951	1923
1986	1978	1982	1948	1945
1975	1978	1982	1953	1938

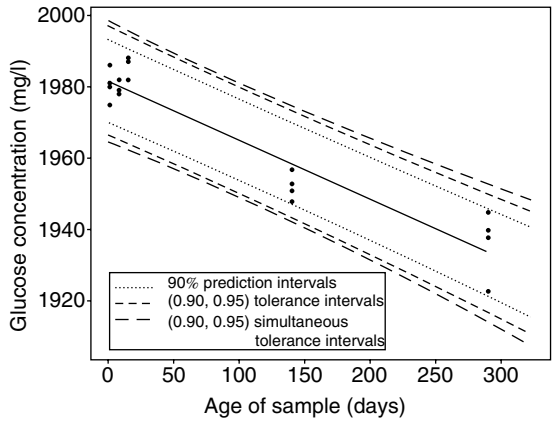

Figure 1 Glucose in bovine serum

months (see Table 1 and [28, p. 14]). A simple plot (e.g. Figure 1 below) of these data suggests that measurements on days 1, 8, and 15 can be pooled, and this is supported by formal tests. However, there is a roughly linear decrease in concentration for the last two sets of measurements.

Simple Random Samples

We begin by assuming that $Y_1, \ldots, Y_n$ are independent and identically distributed (iid). We consider primarily nonparametric tolerance intervals, and tolerance intervals under normality, illustrated using the data in Table 1.

Nonparametric

In one of the earliest papers on tolerance intervals [38], Wilks discusses nonparametric limits (*see* **Nonparametric Methods**). He notes that if $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ are the **order statistics** of an iid sample from a continuous distribution F, then, for $1 \leq i < j \leq n$, the $\Pr[F(Y_{(j)}) - F(Y_{(i)}) \geq \beta]$ does not depend on F, and can be calculated. This follows from the well-known fact that $F(Y_{(i)})$ and $F(Y_{(j)})$ have **beta distributions**. To determine a (β, γ) tolerance interval, one can select i and j for which

$$\int_0^{1-\beta} \left\{1 - \text{beta}\left[\frac{\beta}{(1-x)}; j-i, n-j+1\right]\right\} \times \text{beta}'(x; i, n-i+1)\, \mathrm{d}x \geq \gamma, \qquad (2)$$

where beta $(\cdot, \lambda_1, \lambda_2)$ denotes the beta distribution with parameters λ_1 and λ_2, and beta$'$ $(\cdot, \lambda_1, \lambda_2)$ denotes the corresponding density. The above integral, derived by conditioning on $F(Y_{(i)})$, is straightforward to evaluate numerically. Some tables (e.g. [6, pp. 318–324]) and graphs (e.g. [19] and [27, p. 123]) are available. Voluminous tables are avoided through the use of the concept of *statistically equivalent blocks* [27, 33], an idea which is central to the theory of multivariate nonparametric tolerance regions. In this univariate case it can be shown that the random variables $C_i = F(Y_{(i)}) - F(Y_{(i-1)})$, where $Y_{(0)} \equiv -\infty$ and $Y_{(n+1)} \equiv \infty$, have the same distribution as the differences of successive order statistics from a sample of size n from a **uniform distribution**. Also, the sum, S_t, of any t of the C_i has a beta $(t, n+1-t)$ distribution. If q denotes the smallest t for which $\Pr(S_t \geq \beta) \geq \gamma$, then $(Y_{(i)}, Y_{(i+q)}), i+q \leq n+1$, is a (β, γ) nonparametric tolerance interval. Since $\mathrm{E}(C_i) = 1/(n+1)$, the expected content of this tolerance limit is simply $q/(n+1)$.

If we pool the data in Table 1 for the first three days, we have $n = 12$, $Y_{(1)} = 1975$, and $Y_{(12)} = 1988$. Substituting $i = 1$, $j = 12$, and $\beta = 0.6613$ into (2) gives $\gamma \doteq 0.9500$; so [1975, 1988] is a (0.6613, 0.95) nonparametric tolerance interval, with expected content $11/13 = 0.846$. Note that for $n = 12$ and $\gamma = 0.95$, nonparametric tolerance intervals of this form do not exist for $\beta > 0.6613$. One-sided nonparametric tolerance limits involving linear combinations of order statistics have been proposed [8, 34]. These limits are valid for a smaller class of distributions than the Wilks limits, but they do not have sample-size limitations.

Normal Distribution

Let $Y_1, \ldots, Y_n$ be an iid sample from a **normal distribution**, and denote the usual sample mean and

standard deviation by $\overline{y}$ and s. We can determine one- and two-sided (β, γ) tolerance limits of the form $\overline{y} - ks$, $\overline{y} + ks$, or $\overline{y} \pm ks$, where k is an appropriate constant. Extensive tables of tolerance limit factors, k, are available; the best source of such tables is Odeh & Owen [21]. For one-sided tolerance limits, the tolerance limit factors are easily shown (e.g. [21, pp. 269–270]) to be proportional to quantiles of a noncentral **Student's *t* distribution**. For two-sided tolerance limits, Wald & Wolfowitz [36] show that $k \approx ru$, where $u = [(n-1)/\chi^2_{n-1,1-\gamma}]^{1/2}$ and r is the solution to

$$\frac{1}{(2\pi)^{1/2}} \int_{r-1/\sqrt{n}}^{r+1/\sqrt{n}} \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t = \beta. \qquad (3)$$

This Wald–Wolfowitz approximation is usually very good, and was used almost exclusively before the advent of modern computing capabilities. Today, exact tolerance limit factors are easy to obtain numerically, and extensive tables are available (e.g. [21]).

Normal tolerance limits for which, with confidence γ, $(1-\beta)/2$ of the population is both less than the lower limit and greater than the upper limit are discussed by Owen [23], and corresponding tolerance limit factors have been tabulated [21, pp. 115–145]. From definition (1), one can see that this symmetry condition is not required for a tolerance limit.

Returning to the example dataset (days 1, 8, and 15), we have $n = 12$, $\overline{y} = 1981.5$ and $s = 3.9196$. To construct, for example, a (0.90, 0.95) two-sided β-content normal tolerance interval, use $k = 2.670$ (e.g. from [21, Table 3.4.1, p. 98]), and calculate the desired limit (1971.0, 1992.0). For a (0.90, 0.95) tolerance limit which controls the probability in both tails, we would use $k = 2.978$ [21, Table 4.4.1, p. 128]. For a β-expectation (i.e. prediction) interval with $\beta = 0.9$, it is easy to show that $k = t_{(1+\beta)/2}(n-1)[n+1)/n]^{1/2} = 1.869$ is appropriate.

Other Distributions

Tolerance interval methods for simple random samples are available for many distributions, continuous and discrete. Patel [24] provides a review. For a **log-normal** model, one need only transform the data by taking logarithms, calculate normal-theory tolerance limits, and exponentiate the result. For the **Weibull distribution** (and, taking logarithms, the **extreme value** distribution), Thoman et al. [32] obtain one-sided tolerance limits using quantiles of pivotal random variables obtained by simulation; a good source for the necessary tables is [2]. Lawless [13] demonstrates that if one conditions on **ancillary statistics**, then one-sided Weibull tolerance limits can be determined numerically, without the need for simulation or tables. One-sided tolerance limits for the log **gamma distribution** are discussed by Jones et al. [12] in an article that emphasizes regression models. For some discussion of discrete distributions, see [24] and [25], and the references cited therein.

Linear Regression

The concept of a tolerance limit can also be applied to **linear regression** models, and the literature in this area is extensive. Let $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$, where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unknown constant parameters, and $\mathbf{X}$ is a known (for convenience, full rank) $n \times p$ matrix of covariates. The **least squares** estimator $\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is normally distributed with expectation $\boldsymbol{\alpha}$. The residual mean square $S = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\alpha}})'(\mathbf{Y} - \hat{\boldsymbol{\alpha}})/(n-p)$ is σ^2 times a χ^2_{n-p} **(chi-square distributed)** random variable divided by its **degrees of freedom**. For any $p \times 1$ vector $\mathbf{w}$, $\mathbf{w}'\hat{\boldsymbol{\alpha}}$ is normally distributed, and an independent estimate of its variance is proportional to a χ^2 random variable. As a consequence, the theory developed for normal tolerance and prediction intervals for a simple random sample can be applied directly to **fixed-effects** regression problems. This was apparently first demonstrated by Wallis [37], who employed what amounts to the approximation of Wald & Wolfowitz [36]. Although the Wald–Wolfowitz approximation is easy to use, and usually quite accurate, exact computations are straightforward using modern computers.

Similarly, it is reasonable to extend the notion of a prediction interval to a regression setting. Prediction intervals for a single future observation are, of course, classical results in most elementary regression analysis textbooks. Various formulations of multiple prediction problems, along with corresponding solutions, are reviewed by Hahn [7], Hahn & Meeker [6], and Patel [25].

Random-Effects and Mixed-Effects Models

Consider a large (effectively, infinite) population of individuals who respond differently to a treatment,

with each individual having his/her own mean response. Assume that the response for each individual is normally distributed with a common variance σ^2, and that means for individuals are themselves distributed $N(\mu, \sigma_b^2)$. On the basis of, say, J measurements on each of I individuals, one might be concerned with estimating an interval that contains at least a proportion β of the population of potential measurements from a randomly selected member of the population, with confidence γ. Or one might want a prediction interval for a future measurement made on a randomly selected person. This is an example of a **random effects** tolerance limit. Mee & Owen [17] provide conservative methods for one-sided β-content tolerance limits, and Mee [16] treats two-sided tolerance and prediction intervals. Vangel [35] presents an approach for approximate one-sided tolerance intervals for (possibly unbalanced) mixed models having two **components of variance**.

Simultaneous Tolerance Intervals

We now return to the usual regression setup; that is, the data have the form $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is a known $n \times p$ matrix of covariates, $\boldsymbol{\alpha}$ is an unknown vector of fixed coefficients, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Simultaneous prediction intervals constructed for m future $\mathbf{x}$s, when m is large, can often be so wide as to be unusable. For such situations, one would typically use tolerance intervals, or simultaneous tolerance intervals.

A regression tolerance interval covers, for any fixed vector of covariates $\mathbf{x}$, at least a proportion β of future responses $y(\mathbf{x})$, with confidence γ. A *simultaneous* regression tolerance interval is a band about a regression surface such that if $\mathbf{x}_1^*, \mathbf{x}_2^*, \mathbf{x}_3^*, \ldots$ denote arbitrary covariate vectors, and if $y_1^*, y_2^*, y_3^*, \ldots$ denote the corresponding future ys, then at least a proportion β of these future responses will be contained within a tolerance interval, with confidence γ. One can think of the regression sample as being a "training sample" used to estimate a surface, which provides a bound on arbitrarily many future responses. The simultaneous tolerance interval statement means that for at least a proportion γ of repeatedly estimated regression surfaces, at least $100\beta\%$ of all future y^*s will fall within the simultaneous tolerance intervals. Various approximate simultaneous tolerance limit procedures, using different approximations, have been proposed; among these are methods of Lieberman & Miller [14], Limam & Thomas [15], Scheffé [29], and Mee et al. [18].

Regression tolerance intervals, individual and simultaneous, are important in multiple-use **calibration**. In this scenario, a regression curve, once estimated, is to be used many times to obtain estimates and confidence intervals for $\mathbf{x}^*$s based on future y^*s. For a discussion of the use of tolerance intervals in calibration, see the review by Osborne [22]; Lieberman & Miller [14] present an immunoassay example.

An Example Using Serum Glucose Data

In Figure 1, a straight line is fit, and three two-sided tolerance bands are displayed, for the data in Table 1. The innermost band consists of individual 90% prediction intervals (that is, β-expectation tolerance limits) for a single observation, and the outermost bands provide (0.90, 0.95) tolerance intervals: individual and simultaneous. The simultaneous intervals in this figure were calculated using the method of Mee et al. [18].

References

[1] Aitchison, J. & Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.

[2] Bain, L.J. (1978). *Statistical Analysis of Reliability and Life-Testing Models*. Marcel Dekker, New York.

[3] Brown, E.B., Iyer, H.K. & Wang, C.-M. (1997). Tolerance intervals for assessing individual bioequivalence, *Statistics in Medicine* **16**, 803–820.

[4] Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian*. Hafner, Darien.

[5] Guttman, I. (1988). Tolerance regions, statistical, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 272–287.

[6] Hahn, G.J. & Meeker, W.Q. (1991). *Statistical Intervals: A Guide for Practitioners*. Wiley, New York.

[7] Hahn, G.J. & Nelson, W. (1973). A survey of prediction intervals and their applications, *Journal of Quality Technology* **5**, 178–188.

[8] Hanson, D.L. & Koopmans, L.H. (1964). Tolerance limits for the class of distributions with increasing hazard rate, *Annals of Mathematical Statistics* **35**, 1561–1570.

[9] Holst, E. & Christensen, J.M. (1992). Intervals for the description of the biological level of a trace element in a reference population, *Statistician* **41**, 233–242.

[10] Jilek, M. (1981). A bibliography of statistical tolerance regions, *Mathematische Operationsforschung und Statistics Series Statistics* **12**, 441–456.

[11] Jilek, M. & Ackermann, H. (1989). A bibliography of statistical tolerance regions, II, *Statistics* **20**, 165–172.

[12] Jones, R.A., Scholz, F.W., Ossiander, M. & Shorack, G.R. (1985). Tolerance bounds in log-gamma regression models, *Technometrics* **27**, 109–118.

[13] Lawless, J.F. (1975). Construction of tolerance bounds for the extreme-value and Weibull distributions, *Technometrics* **17**, 255–261.

[14] Lieberman, G.J. & Miller, R.G., Jr (1963). Simultaneous tolerance intervals in regression, *Biometrika* **50**, 155–168.

[15] Limam, M.M.T. & Thomas, D.R. (1988). Simultaneous tolerance intervals for the linear regression model, *Journal of the American Statistical Association* **83**, 801–804.

[16] Mee, R.W. (1984). β-expectation and β-content tolerance limits for balanced one-way ANOVA random model, *Technometrics* **26**, 251–254.

[17] Mee, R.W. & Owen, D.B. (1983). Improved factors for one-sided tolerance limits for balanced one-way ANOVA random model, *Journal of the American Statistical Association* **78**, 901–905.

[18] Mee, R.W., Eberhardt, K.R. & Reeve, C. (1991). Calibration and simultaneous tolerance intervals for regression, *Technometrics* **33**, 211–219.

[19] Murphy, R.B. (1948). Non-parametric tolerance limits, *Annals of Mathematical Statistics* **19**, 581–589.

[20] Noether, G. (1985). Nonparametric tolerance intervals, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 331–332.

[21] Odeh, R.E. & Owen, D.B. (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. Marcel Dekker, New York.

[22] Osborne, C. (1991). Statistical calibration: a review, *International Statistical Review* **59**, 309–336.

[23] Owen, D.B. (1964). Control of percentage in both tails of the normal distribution, *Technometrics* **6**, 377–387.

[24] Patel, J.K. (1986). Tolerance limits: a review, *Communications in Statistics – Theory and Methods* **15**, 2719–2762.

[25] Patel, J.K. (1988). Prediction intervals: a review, *Communications in Statistics – Theory and Methods* **18**, 2393–2465.

[26] Paulson, E. (1943). A note on tolerance limits, *Annals of Mathematical Statistics* **14**, 90–93.

[27] Pratt, J.W. & Gibbons, J.D. (1981). *Concepts of Nonparametric Theory*. Springer-Verlag, New York, pp. 118–130.

[28] Schaffer, R., Mandel, J., Sun, T. & Hertz, H.S. (1982). *Evaluation by an ID/MS Method of the AACC Reference Method for Serum Glucose*, NBS Special Publication 260–80. National Bureau of Standards, Gaithersburg.

[29] Scheffé, H. (1973). A statistical theory of calibration, *Annals of Statistics* **1**, 1–53.

[30] Selberg, H.E. (1983). The theory of reference values. Part 5. Statistical treatment of collected reference values. Reference limits, *Journal of Clinical Chemistry and Clinical Biochemistry* **21**, 749–760.

[31] Shewhart, W.A. (1931). *Economic Control of Quality of Manufactured Products*. Van Nostrand, New York.

[32] Thoman, D.R., Bain, L.J. & Antle, C.E. (1970). Maximum likelihood estimation, exact confidence intervals for reliability, and tolerance limits in the Weibull distribution, *Technometrics* **12**, 363–371.

[33] Tukey, J.W. (1947). Nonparametric estimation II. Statistically equivalent blocks and tolerance regions–the continuous case, *Annals of Mathematical Statistics* **18**, 529–539.

[34] Vangel, M.G. (1994). One-sided nonparametric tolerance limits, *Communications in Statistics – Theory and Methods* **23**, 1137–1154.

[35] Vangel, M.G. (1996). Design allowables from regression models using data from several batches, in *Composite Materials: Testing and Design*, R.B. Deo & C.R. Saff, eds. ASTM STP 1274, American Society for Testing and Materials, Philadelphia, pp. 358–370.

[36] Wald, A. & Wolfowitz, J. (1946). Tolerance limits for a normal distribution, *Annals of Mathematical Statistics* **17**, 208–215.

[37] Wallis, W.A. (1951). Tolerance intervals for linear regression, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 43–52.

[38] Wilks, S.S. (1941). Determination of sample sizes for setting tolerance limits, *Annals of Mathematical Statistics* **12**, 91–96.

MARK G. VANGEL

Tolerance Region

Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ denote iid p-dimensional vectors with distribution function F. A (β, γ) β-content tolerance region is a set $R(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$ such that

$$\Pr\left(\int_R \mathrm{d}F \geq \beta\right) \geq \gamma.$$

The constants β and γ are referred to as the *content* (or *coverage*) and *confidence* of the tolerance region R, respectively. A β-expectation tolerance region (or *prediction region*) is a region R for which $\mathrm{E}[\Pr(Y \in R)] \geq \beta$. Univariate tolerance regions, often called **Tolerance Intervals**, are discussed in a separate article, which also contains references to bibliographies and reviews on multivariate theory.

The statistical theory on this topic consists mostly of nonparametric and normal theory. We provide an introduction to the main ideas of nonparametric theory below (*see* **Nonparametric Methods**). Some references for **multivariate normal** methods include the monograph by Guttman [3], and articles by Chew [1] and Hall & Sheldon [4].

Nonparametric Regions

Multivariate tolerance regions are determined using *statistically equivalent blocks*, a concept due to Tukey [6], which is introduced in the article on **Tolerance Intervals**, but which we can better appreciate in the multivariate case.

Let $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ be the order statistics from an iid univariate sample with continuous distribution F. The intervals $C_i = F(Y_{(i)}) - F(Y_{(i-1)})$ (with $Y_{(0)} \equiv -\infty$ and $Y_{(n+1)} \equiv \infty$) are an example of statistically equivalent blocks of **exchangeable** random variables, each of which has the same **beta distribution** as the difference in consecutive uniform **order statistics**. The expectation of each of these random variables is $1/(n+1)$, and the sum of any t of them has the same distribution as the tth order statistic of a uniformly distributed sample of size n. As a consequence, we can easily construct β-content and β-expectation tolerance intervals; this is discussed in some detail in **Tolerance Interval**.

Wald [7] generalizes this idea to multiple dimensions, providing essentially the following **algorithm** for a multivariate tolerance region. First, construct a one-dimensional tolerance interval by ordering the data according to one of the coordinates, keeping only those data values that correspond to blocks chosen to provide a tolerance interval in that coordinate. Repeat this with each coordinate in turn. The region that results is a multivariate tolerance region. Tukey [6] extends Wald's work greatly, introducing the concept of *ordering functions*. Fraser [2] provides further results; Pratt & Gibbons [5] give an elementary introduction.

References

[1] Chew, W. (1966). Confidence, prediction and tolerance regions for the multivariate normal distribution, *Journal of the American Statistical Association* **61**, 605–617.

[2] Fraser, D.A.S. (1953). Nonparametric tolerance regions, *Annals of Mathematical Statistics* **24**, 44–55.

[3] Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian*. Hafner, Darien.

[4] Hall, I.J. & Sheldon, D.D. (1979). Improved bivariate normal tolerance regions with some applications, *Journal of Quality Technology* **11**, 13–19.

[5] Pratt, J.W. & Gibbons, J.D. (1981). *Concepts of Nonparametric Theory*. Springer-Verlag, New York, pp. 118–130.

[6] Tukey, J.W. (1947). Nonparametric estimation II. Statistically equivalent blocks and tolerance regions – the continuous case, *Annals of Mathematical Statistics* **18**, 529–539.

[7] Wald, A. (1943). An extension of Wilks' method for setting tolerance limits, *Annals of Mathematical Statistics* **14**, 45–55.

MARK G. VANGEL

Total Time on Test

Total time on test (TTT) statistics [6] and plots [3, 5] have been discussed quite frequently in the reliability literature on failure-time data as tools for assessing that a **hazard** (or intensity) is constant. The ideas have been phrased in terms of the multiplicative intensity model for counting processes by Aalen & Hoem [1] and Gill [7], and surveyed by Andersen et al. [2]. In this article we introduce this methodology (in counting process notation) with emphasis on biostatistical applications, and present an example of testing constant hazard of death in a follow-up study of liver cirrhosis patients.

We consider a univariate counting process $N(t)$ on an interval $[0, \tau]$ satisfying the **multiplicative** intensity model, i.e. it has an intensity process of the form

$$\lambda(t) = \alpha(t)Y(t),$$

where we want to test the hypothesis that there exists a θ such that

$$\mathrm{H}_0 : \alpha(t) = \theta$$

for all t. Let

$$R(t) = \int_0^t Y(s)\,\mathrm{d}s;$$

then under H_0 the counting process $N(t)$ has $\theta R(t)$ as its compensator and the usual decomposition yields

$$N(t) = \theta R(t) + M(t),$$

with M a (local square integrable) martingale. Therefore, we have

$$\mathrm{E}[R(t)] = \mathrm{E}\left[\frac{N(t)}{\theta}\right]$$

(provided that the expectations exist), and it follows that a plot of $R(t)$ against $N(t)$ should give approximately a straight line with slope θ^{-1}. This plot may be transformed to the unit square by choosing a (possibly random) time T and plotting $R(t)/R(T)$ against $N(t)/N(T)$. This is the TTT plot, the name being due to the fact that $R(t)$ measures the "exposure" or "total time on test" when $Y(t)$ is the size of the risk set. This is, for instance, the case when $N(t) = \sum_i I(X_i \le t, D_i = 1)$ counts the number of failures in $[0, t]$ among independent, identically distributed (iid) (possibly right-**censored**) survival times $X_1, \ldots, X_n$ with D_i being the failure indicator. In this case $Y(t) = \sum_i I(X_i \ge t)$ equals the number at risk at time $t-$. The plot should approximate a straight line with unit slope under H_0, which, for survival data, corresponds to the exponential distribution.

We plot $R(t)/R(T)$ against $N(t)/N(T)$ for all $t \in [0, T]$ and connect the vertical lines one gets in this manner by horizontal lines.

The TTT plot is especially well-suited for situations where the alternative to the hypothesis H_0 of special interest is that $\alpha(t)$ is monotone. Since we have

$$\mathrm{E}[N(t)] = \mathrm{E}\left[\int_0^t \alpha(s)Y(s)\,\mathrm{d}s\right] = \int_0^t \alpha(s)\,\mathrm{dE}[R(s)],$$

it is seen that $\mathrm{dE}[R(t)] = \mathrm{dE}[N(t)]/\alpha(t)$. Therefore, the TTT plot will tend to be concave for an increasing $\alpha(t)$ and convex when $\alpha(t)$ is decreasing.

Closely related to the TTT plot is the cumulative total time on test statistic, defined as $N(T)$ times the area under the TTT plot. According to the just-mentioned properties of the TTT plot, this statistic for testing H_0 tends to take on large values when $\alpha(t)$ is increasing and small values when it is decreasing. One may also use a **Kolmogorov–Smirnov**-type test, i.e. reject the hypothesis when the maximum distance between the TTT plot and the diagonal line $y = x$ is large.

To study formally the properties of the TTT plot, we follow Aalen & Hoem [1] and note that in the new (random) time scale measured by "the total time on test" $R(t)$, N is transformed into a counting process N^* given by

$$N^*(u) = N[R^{-1}(u)],$$

on the (random) interval $[0, R(\tau)]$. Here, $R^{-1}(u) = \inf\{t : R(t) \ge u\}$. This counting process has intensity process $\lambda^*(u) = \alpha[R^{-1}(u)]$, and under H_0 it is a counting process with a constant intensity process θ on $[0, R(\tau)]$, i.e. a randomly stopped **Poisson process** with constant intensity.

As shown by Barlow et al. [4] and Barlow & Campo [3] for certain censoring patterns, and by Gill [7] in greater generality, the asymptotic distribution of the signed area between the TTT plot and the

diagonal line $y = x$, times $[N(T)]^{1/2}$, is the same as that of $\int_0^1 W^0(x)\,dx$, W^0 being the standard Brownian bridge. Therefore, the normalized cumulative total time on test statistic

$$[N(T)]^{1/2}\left[\frac{1}{N(T)}\sum_{i=1}^{N(T)}\frac{R(T_i)}{R(T)} - \frac{1}{2}\right]$$

(with $T_1 < T_2 < \cdots$ being the jump times of N) is asymptotically normally distributed with mean zero and variance 1/12 under the hypothesis. Furthermore, the Kolmogorov–Smirnov-type test for hypothesis H_0 is to reject at the level α when

$$[N(T)]^{1/2}\sup_{0\le t\le T}\left|\frac{R(t)}{R(T)} - \frac{N(t)}{N(T)}\right| > e_\alpha,$$

where e_α is the upper α fractile in the distribution of $\sup_{0\le x\le 1}|W^0(x)|$.

Example

In a **clinical trial** of prednisone vs. placebo treatment of liver cirrhosis, patients entered during 1962–1969 and were followed until September 1974. We consider the survival experience of the 237 placebo patients which could be reevaluated histologically [8]. Figure 1 shows the TTT plot with

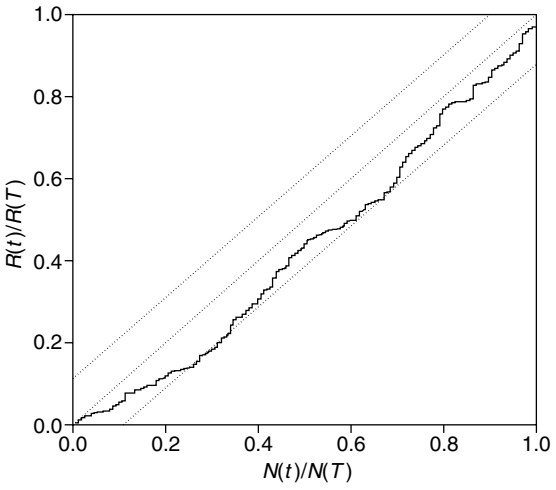

Figure 1 Total time on test plot for mortality of liver cirrhosis patients, including the identity line and boundaries corresponding to the Kolmogorov–Smirnov test

$T = 4892$ days and $N(T) = 150$ observed deaths. The plot is generally *convex*, corresponding to a generally *decreasing* hazard rate. The plot crosses the boundaries $\pm e_\alpha[N(T)]^{1/2}$ around the identity line, where $e_{0.05} = 1.36$ [9, Table 9], indicating significant departure from the hypothesis of exponentiality as judged by a Kolmogorov–Smirnov test at the 5% level. Indeed, the maximal deviation is 0.1214, yielding a Kolmogorov–Smirnov test statistic of $(150)^{1/2} \times 0.1214 = 1.486$, corresponding to $P = 0.024$. The cumulative TTT statistic takes the values -0.8234, corresponding to an approximately normal deviate of $(12)^{1/2} \times (-0.8234) = -2.85$ or a (two-sided) $P = 0.004$.

References

[1] Aalen, O.O. & Hoem, J.M. (1978). Random time changes for multivariate counting processes, *Scandinavian Actuarial Journal*, 81–101.

[2] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[3] Barlow, R.E. & Campo, R. (1975). Total time on test processes and application to failure data analysis, in *Reliability and Fault Tree Analysis*, R.E. Barlow, J. Fussell & N.D. Singpurwalla, eds. SIAM, Philadelphia, pp. 451–481.

[4] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

[5] Bergman, B. (1985). On reliability theory and its applications (with discussion), *Scandinavian Journal of Statistics* **12**, 1–41.

[6] Epstein, B. & Sobel, M. (1953). Life testing, *Journal of the American Statistical Association* **48**, 486–502.

[7] Gill, R.D. (1986). The total time on test plot and the cumulative total time on test statistic for a counting process, *Annals of Statistics* **14**, 1234–1239.

[8] Schlichting, P., Christensen, E., Andersen, P.K., Fauerholdt, L., Juhl, E., Poulsen, H. & Tygstrup, N., for The Copenhagen Study Group for Liver Diseases (1983). Identification of prognostic factors in cirrhosis using Cox's regression model, *Hepatology* **3**, 889–895.

[9] Schumacher, M. (1984). Two-sample tests of Cramér–von Mises and Kolmogorov–Smirnov type for randomly censored data, *International Statistical Review* **52**, 263–281.

PER KRAGH ANDERSEN & NIELS KEIDING

Transfer Function Models

Time series data arise in a wide range of subject areas; biology and medicine (see, for example, [4]), environment, engineering, economics, etc. To analyze and model such data a number of different tools are available. The simplest is the univariate stochastic model, which associates the value of a variable at time point t to some function of its values at previous time points, i.e. $Y_t = f(Y_{t-1}, Y_{t-2}, \ldots; \boldsymbol{\nu}, \{N_t\})$, where $\boldsymbol{\nu}$ is a vector of parameters and $\{N_t\}$ a noise process. Often the series $\{Y_t\}$ is thought to be related to a second series $\{X_t\}$. A simple example of a *transfer function model* relating the two time series would be

$$Y_t = \nu_0 X_t + \nu_1 X_{t-1} + \cdots + N_t, \tag{1}$$

where $\{X_t\}$ is referred to as the "input" series and $\{Y_t\}$ the "output" series. Transfer function models have proved very useful in the assessment of relationships, if any, between time series measurements, and for prediction and **forecasting** [2, 9]. For example, relating "the daily number of patients attending a clinic for respiratory disease" (output) to "daily pollen count", "daily level of air pollutant", "daily humidity value" (inputs); "monthly employment statistics" related to "inflation rate", "level of investment", etc., in previous months; or, "daily demand for electricity" related to "daily mean temperature" and "mean night temperature".

Assuming there to be no noise (*see* **Noise and White Noise**) in the system, we can write (1) as

$$Y_t = \nu(B) X_t \tag{2}$$

where B is the usual **backward shift operator** $B^h X_t = X_{t-h}$, the operator $\nu(B) = \sum_{i=0}^{\infty} \nu_i B^i$ a polynomial in B known as the *transfer function* of this *linear filter*, and the set of *weights*, $\nu_0, \nu_1, \ldots$ the *impulse response function* of the system.

Note that if $y_t = \nabla Y_t = Y_t - Y_{t-1}$ and $x_t = \nabla X_t = X_t - X_{t-1}$ denote incremental changes in Y and X, then we have

$$y_t = \nu(B) x_t \tag{3}$$

the *same* transfer function model.

Eq. (2) contains a large, possibly infinite, number of parameters and in many systems there may be a time lag, b, in Y_t for a change in X_t. In this case, we can reparameterize the system as

$$Y_t = \delta_r^{-1}(B)\omega_s(B) X_{t-b}, \tag{4}$$

where $\delta_r(B) = (1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r)$ and $\omega_s(B) = (\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s)$. Thus, if $\Omega(B) = \omega_s(B) B^b$, the transfer function of this noise-free model is $\nu(B) = \delta_r^{-1}(B)\Omega(B)$, the ratio of two finite-order polynomials, and in this case $\nu_k = 0$ for $k = 0, \ldots, b-1$.

For many practical applications values of r and s that are both ≤ 2 provide suitable models. Box & Jenkins [2] give details of the properties of transfer function models for all combinations of $r = 0, 1, 2$ and $s = 0, 1, 2$.

If multiple input series are available, then the basic model (4) can be generalized to

$$Y_t = \sum_{j=1}^{J} \delta_j^{-1}(B)\omega_j(B) X_{j,t-b_j}, \tag{5}$$

where each input X_{jt} has a transfer function representation the same as for the single-input case, with time lag b_j. The same property for *differencing* as that stated in (3) also applies in this multiple-input case.

For the potential user of transfer function models, the most important task is to identify (*fit*) the model that best describes the process under investigation. This is frequently a mixture of art and science! There are, however, some basic steps to be followed.

Model Identification

An output series $\{Y_t\}$ will almost always be subject to some kind of error. Let us consider the process of identification of a combined transfer function–noise model of the form

$$y_t = \delta_r^{-1}(B)\omega_s(B) X_{t-b} + N_t, \tag{6}$$

where $\{N_t\}$ is a noise process assumed to be generated by an ARIMA (*see* **ARMA and ARIMA Models**) process *statistically independent* of the input process $\{X_t\}$. If $n_t = \nabla^d N_t$, where $\nabla = 1 - B$ denotes the (backward) difference, is a stationary process, then (6) can be rewritten as

$$y_t = \delta_r^{-1}(B)\omega_s(B) x_{t-b} + \phi_p^{-1}(B)\theta_q(B) a_t, \tag{7}$$

where $\{a_t\}$ is a *white noise* process.

In the same way that estimated autocorrelation and partial autocorrelation functions are used to identify univariate ARIMA models, so the estimated *cross correlation function* is used in the identification of transfer function models.

The cross correlation function between two *stationary* series $\{U_t\}$ and $\{V_t\}$ at lag k is defined as

$$\rho_{U,V}(k) = \frac{\gamma_{U,V}(k)}{\sqrt{\sigma_U^2 \sigma_V^2}}, \quad k = 0, \pm 1, \pm 2, \ldots,$$

where $\gamma_{U,V}(k) = \mathrm{E}\{[U_t - \mathrm{E}(U_t)][V_{t+k} - \mathrm{E}(V)]\}$, and σ_U^2 and σ_V^2 are the variances of the input and output series, respectively.

If data $(u_t, v_t; t = 1, \ldots, m)$ is available, then an estimate of the cross correlation function is given by

$$\hat{\rho}_{U,V}(k) = \frac{\hat{\gamma}_{U,V}(k)}{\sqrt{\hat{\sigma}_U^2 \hat{\sigma}_V^2}},$$

where

$$\hat{\gamma}_{U,V}(k) = \sum_{t=1}^{m-k} \frac{(u_t - \overline{u})(v_{t+k} - \overline{v})}{m},$$

$$\hat{\sigma}_U^2 = \sum_{t=1}^{m} \frac{(u_t - \overline{u})^2}{m}, \quad \text{and}$$

$$\hat{\sigma}_V^2 = \sum_{t=1}^{m} \frac{(v_t - \overline{v})^2}{m}$$

are estimates of the covariance and variances, respectively. If the two time series, U and V, are not cross correlated and one is *white noise*, then the standard error for $\hat{\rho}_{U,V}(k)$ is $1/\sqrt{m}$. This result is used to test the statistical significance of cross correlations at each lag k.

Step 1. Identify *univariate* ARIMA models for both the input and output series. The ARIMA model for the input series converts the correlated series $\{X_t\}$ into an approximately independent series $\{\alpha_t\}$. Suppose that this fitted ARIMA model is

$$\alpha_t = \hat{\phi}_X(B)\hat{\theta}_X^{-1}(B)X_t \tag{8}$$

The identical ARIMA model is then applied to the output $\{Y_t\}$ to produce a new series

$$\beta_t = \hat{\phi}_X(B)\hat{\theta}_X^{-1}(B)Y_t; \tag{9}$$

this process is known as *prewhitening* [2].

The transfer function model then becomes

$$\beta_t = \nu(B)\alpha_t + \varepsilon_t, \tag{10}$$

where $\varepsilon_t = \hat{\phi}_X(B)\hat{\theta}_X^{-1}(B)N_t$ is the transformed noise process.

Step 2. Furthermore, it can be shown that $\nu_k = \rho_{\alpha\beta}(k)\sigma_\beta/\sigma_\alpha$. Although statistically inefficient, a reasonable initial estimate for ν_k is $\hat{\nu}_k = \hat{\rho}_{\alpha\beta}(k)\hat{\sigma}_\beta/\hat{\sigma}_\alpha$. Because $\{\alpha_t\}$ is an approximately independent white noise series, the result stated above can be used to identify which cross correlations are significantly different than zero. This step provides an initial estimate of the impulse response function $\{\hat{\nu}_k\}$.

Step 3. The next step is to use $\{\hat{\nu}_k\}$ to "estimate" b, r, and s in (4). The general approach is initially to consider values of $r, s \leq 2$, and to use the results that $\{\nu_k\}$ comprise

1. b zero values $\nu_0, \nu_1, \ldots, \nu_{b-1}$;
2. a further $s - r + 1$ values $\nu_b, \nu_{b+1}, \ldots, \nu_{b+s-r}$, with no fixed pattern;
3. values ν_j, $j \geq b + s - r + 1$, follow the pattern dictated by an rth order difference equation which has starting values $\nu_{b+s}, \ldots, \nu_{b+s-r+1}$. Starting values for ν_j, for $j < b$, will be zero.

Examining the set of values $\{\hat{\nu}_k\}$ in the light of these results, it is possible to obtain preliminary choices for r, s, and b.

Step 4. The identity $\nu(B) = \delta_r^{-1}(B)\Omega(B)$ can be established, and by comparing coefficients of powers of B, preliminary estimates for $\delta' = (\delta_1, \ldots, \delta_r)$ and $\omega' = (\omega_0, \ldots, \omega_s)$ can be obtained.

Step 5. Approximations to the maximum likelihood estimates of parameters δ', ω', ϕ', and θ' are obtained by minimizing the conditional sum of squares function

$$S_0(\delta', \omega', \phi', \theta') = \sum_{t=1}^{m} a_t^2(\delta', \omega', \phi', \theta' | b, x_0, y_0, a_0), \tag{11}$$

where $a_t = \hat{\theta}_X^{-1}(B)\hat{\phi}_X(B)\hat{n}_t$, $\hat{n}_t = y_t - \hat{y}_t$, $\hat{y}_t = \hat{\delta}_r^{-1}(B)\hat{\omega}_s(B)\hat{x}_{t-b}$, and b, x_0, y_0, and a_0 are starting values. This step involves the use of an algorithm to minimize the nonlinear function S_0.

Further details, for each of these steps, can be found in [2] and [9].

Model Checking

The adequacy of any fitted model must be examined. *Residual analysis* is used to do this. Suppose that the identification process gives residuals $\{\hat{a}_t(\hat{\delta}', \hat{\omega}', \hat{\phi}', \hat{\theta}'); t = 1, \ldots, m\}$. The **autocorrelation function** of these residuals, $\hat{\rho}_{\hat{a},\hat{a}}(k) = \gamma_{\hat{a},\hat{a}}(k)/\sigma^2_{\hat{a}}$, should be that of white noise. If this is not the case, then the identified model is not correct.

The inadequacy of either the transfer function or noise process can be checked by examining the cross correlation function, $\hat{\rho}_{\alpha,\hat{a}}(k)$, between the prewhitened input $\{\alpha_t\}$ and the residuals $\{\hat{a}_t(\hat{\delta}', \hat{\omega}', \hat{\phi}', \hat{\theta}')\}$.

If $\hat{\rho}_{\hat{a},\hat{a}}(k)$ exhibits structure – e.g. significant correlations – and $\hat{\rho}_{\alpha,\hat{a}}(k)$ does not, then the noise model alone is incorrect, whilst if both exhibit structure, then the transfer function and noise model are incorrect. Box & Jenkins [2] provide a number of additional diagnostic checks, and statistical tests, that can be carried out to assess the adequacy of various parts of the fitted model.

The application of these models is therefore a cycle of *identification* → *fitting* → *checking* → *re-identification*. A number of software packages are available to assist this process, including BMDP™ [1] and SAS™ [10] (*see* **Software, Biostatistical**).

Examples

Example (A)

Helfenstein [5] (see also [6]) gives an example of the use of a transfer function model to examine the relationship between environmental time series (daily concentrations of SO_2, NO_2, and other factors) and the incidence of respiratory disease in young children (daily number of respiratory symptoms). Measurements were made over a period of approximately one year. Autoregressive models AR(1) were identified for both input series $\ln(SO_2)$ and NO_2, and an autoregressive integrated moving average model ARIMA(0,1,1) identified for the symptoms output series. Transfer function models were identified using the prewhitened cross correlation function for three cases: $\ln(SO_2)$ and $\ln(NO_2)$ as univariate input series, and then both of these as a two-input series.

The paper also gives details of a special case of a transfer function model known as an *intervention model* (*see* **Intervention Analysis in Time Series**). Intervention models incorporate sudden and unusual events into the identification process. In this particular example the sudden event was a chemical spillage.

Example (B)

Crabtree et al. [3] present several examples of the analysis of biomedical time series data. In one of these a transfer function model is identified to describe the relationship between exercise (miles walked per day) and fasting blood glucose concentration (mg/dl). The exercise input series was identified as an autoregressive process with lags at 5 and 8. After prewhitening, the cross correlation function was calculated and revealed a significant negative peak at lag 1. This indicated that exercise significantly reduced blood glucose the following day.

Some other examples of the use of transfer function models in biostatistics can be found in [7, 8], and [11].

References

[1] *BMDP*, SPSS Inc., Illinois, USA.

[2] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, California.

[3] Crabtree, B.F., Ray, S.C., Schmidt, P.M., O'Connor, P.J. & Schmidt, D.D. (1990). The individual over time: time series applications in health care research, *Journal of Clinical Epidemiology* **43**, 241–260.

[4] Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

[5] Helfenstein, U. (1996). Box–Jenkins modelling in medical research, *Statistical Methods in Medical Research* **5**, 3–22.

[6] Helfenstein, U., Ackermann-Liebrich, U., Braun-Fahrländer Ch. & Wanner, H.U. (1991). Air pollution and diseases of the respiratory tracts in pre-school children: a transfer function model, *Journal of Environmental Monitoring and Assessment* **17**, 147–156.

[7] Katzoff, M. (1989). The application of time series forecasting methods to an estimation problem using provisional mortality statistics, *Statistics in Medicine* **8**, 335–341.

[8] Martinez-Schnell, B. & Zaidi, A. (1989). Time series analysis of injuries, *Statistics in Medicine* **8**, 1497–1508.

[9] Montgomery, D.C. & Weatherby, G. (1980). Modelling and forecasting time series using transfer function and intervention methods, *AIIE Transaction* **12**, 289–307.

[10] SAS (SAS Institute Inc., North Carolina, USA).

[11] Zaidi, A.A., Schnell, D.J. & Reynolds, G.H. (1989). Time series analysis of Syphilis surveillance data, *Statistics in Medicine* **8**, 353–362.

CLIVE J. LAWRENCE

Transformations

Statistical analyses make use of transformations which change some quantity into a function of the quantity in a variety of ways. Common transformations are the taking of logarithms, the calculation of powers (*see* **Power Transformations**), and exponentiation but there are many other possibilities. The primary purpose of transformation is frequently to permit the application of standard statistical methodology in a situation in which some features required for the methodology are not present.

In a statistical model, transformations can be applied to a **response variable**, an **explanatory variable**, or to a parameter of the model. The first two situations can be illustrated in the context of **linear regression** which relates a response variable, Y, to an explanatory variable, X. The **normal distribution** assumptions necessary may not be satisfied for the response Y, but may be more reasonably supposed for some transformation of Y such as its logarithm or square root. Furthermore, the linear dependence of the response on X, represented by the linear predictor $\alpha + \beta X$, may be extended to represent a nonlinear relationship between Y and X by adding powers of X to the model. For example, consideration of the quadratic function $\alpha + \beta X + \gamma X^2$ is often used to provide a first test of the assumption of a linear relationship (*see* **Polynomial Regression**). There also may be features of the application which make models which involve nonlinear functions of the explanatory variables particularly appealing (*see* **Nonlinear Regression**). Similar considerations apply in the use of other regression models.

The transformation of a parameter may be required because the natural parameter of interest is not the most convenient for statistical analysis. For example, in epidemiologic studies interest often focuses on **odds ratios**, but inferences are most conveniently undertaken for the logarithm of odds ratios. In this case, the parameter of interest is transformed for the purposes of inference and, most sensibly, the resulting inferences are transformed back to be summarized in terms of the original parameter.

Another illustration of this arises in **generalized linear models** which are based on "linking" a function of the mean of a response to a set of explanatory variables. In the special case of normal theory regression, the mean itself is used but, for example, with **binary data**, the mean of the response is equal to the probability of "success", say, and **logistic regression** links the logarithm of the odds, Pr(success)/Pr(failure), to a linear predictor involving the explanatory variables.

VERN T. FAREWELL

Transfusion Medicine

Transfusion medicine is the branch of medicine concerned with all aspects of the use of blood products and components – from donor selection and care, through testing (blood grouping and microbiology), component preparation, and the indications for their clinical use. The discipline also encompasses pregnancy testing for blood group incompatibilities between mother and baby. New areas in transfusion medicine now overlap with hematology to include tissue banking, stem cells, and immunotherapy. Useful background is provided in references [1, 4, 8, 9].

History of Transfusion

The science of transfusion medicine had a false start for about 2500 years with the mistaken belief that blood letting was actually beneficial for sick patients. Eventually, in the seventeenth century, William Harvey (1578–1657) discovered that blood circulates around the body, and this was the start of experimentation with blood transfusion. In 1666, the English anatomist Richard Lower (1631–1691) found that the pressure difference between an artery and vein would force blood from donor to recipient, and the first successful animal transfusion was achieved, using two dogs as donor and recipient. The first, albeit unsuccessful, human to human transfusion occurred in London in 1818. Although blood transfusion was eventually found to be effective in some recipients, others experienced a severe, sometimes fatal, reaction. This led to the discovery in 1900, by the Austrian biologist Karl Landsteiner (1868–1943), that the blood of one donor was sometimes incompatible with the blood of another and the A, B, and O, blood groups were identified; these are determined by the presence or absence of the antigens A and B in the erythrocytes, and the agglutinating antibodies anti-A and anti-B in the plasma. A fourth group, AB, was discovered two years later. Other blood subgroups were later identified, one of the most important being RhD (Rhesus factor), named after its discovery in Rhesus monkeys, which frequently had fatal implications for foetuses and neonates. These are now rare due to a highly successful prevention programme involving administration of anti-D to RhD negative mothers at delivery to prevent sensitization.

During the Great War, it was discovered that the addition of sodium citrate would stop blood clotting outside the body and prolong its life when refrigerated. The theory of blood transfusion was eventually put into practice during World War II, when thousands of transfusions were needed to save soldiers' lives and the evidence supporting its use was unequivocal. This marked the birth of the specialty of transfusion medicine, which needed to make huge advances to keep up with demand. It was found that albumin and plasma could be used instead of whole blood, and that plasma could be pooled and fractionated to isolate clotting factors, such as factor VIII, for the treatment of hemophilia. Unfortunately, this later led to HIV infection in thousands of sufferers worldwide before HIV screening was introduced (*see* **AIDS and HIV**). Blood clotting is a mechanism to seal the circulatory system after injury. The wound is initially blocked with platelets, some of which rupture to release chemicals that combine with proteins and enzymes in the plasma to form a tough fibrous clot. These clotting factors play an essential role in this mechanism.

Blood Transfusion Today

Whole blood is collected from suitable donors, screened for viral infection, and typed for blood group. As it is inefficient to transfuse whole blood, donations are separated into the three major components of red cells, platelets, and plasma. Leucocytes are discarded to minimize the risk of variant Creutzfeldt-Jakob disease (vCJD) transmission. Individual blood components, such as platelets or plasma, can also be collected by apheresis, where the selected individual components are separated during donation and the remaining components are returned to the donor. The components are stored at blood banks under appropriate conditions until they are either released for transfusion, or expire.

Transfusion of blood products to patients depends on the clinical indication. The evidence base for the efficacy of transfusion lies mainly in clinical experience and laboratory-based research rather than through clinical trials. The key purpose of red cell transfusion is to improve the delivery of oxygen to the tissues, while platelets and plasma are used to

treat and prevent life-threatening bleeding by clotting the blood. Plasma can be fractionated to produce several components, such as clotting factor concentrates, to treat congenital coagulation deficiencies, like hemophilia; immunoglobulins to treat acute infections and to treat immunological disorders; and albumin. This last product was used to treat hemorrhagic shock, although the evidence for this is now doubtful; it is now used for patients with low albumin levels. Transfusions are mostly given without removing blood from the recipient, but exchange transfusions can be performed where the patient's own blood is removed and replaced with donor blood. This is done to counteract the effects of severe jaundice in neonates, or to reduce the proportion of distorted cells in sickle disease.

Millions of blood units are transfused each year, an estimated 23 million in the United States of America, 6 million in Africa, and 2.6 million in the United Kingdom.

The safety of blood is of paramount importance, and safely procedures are in place at every step from donation to transfusion to help prevent transfusion of incompatible blood and infection. There are limits for the shelf life of each component; this depends on the viability of cells and potential for bacterial contamination after prolonged periods. As well as screening for infectious agents, components and recipient blood samples are screened for antibodies that could cause serious allergic reactions in recipients. Numerous safety checks are performed by the nurse at the bedside prior to transfusion to ensure a patient receives correctly matched blood. Despite these safety measures, blood transfusion still carries some risk, and should only be used when absolutely necessary. Several alternatives to transfusion can be considered, such as erythropoietin to improve patient's own red cell production, volume expanders, and red cell substitutes. Techniques to retransfuse the patient's own blood are also employed to minimize the risks of donor transfusion. In the United Kingdom, the Serious Hazards of Transfusion (SHOT) organization exists to audit serious transfusion complications and publishes an annual report on its findings to guide blood safety policies.

Articles on transfusion medicine are published in general medical journals, hematology journals such as *Blood*, and the specialist journals, *Transfusion, Transfusion Medicine,* and *Vox Sanguinis.*

Statistics and Transfusion Medicine

Statistics has had limited use in transfusion medicine, but one of its important practical applications is in quality control, and in studies of transfused versus untransfused patients, to look for associations between perioperative transfusion and adverse clinical outcomes such as risk of cancer recurrence [10]. Parametric models are also employed to determine time to restoration of blood cell concentration, and to select the optimum transfusion time and triggers for transfusion. Models have also been used to estimate induction time of transfusion associated HIV infection and in monitoring survival of transfusion recipients.

Clinical Studies

Blood banks store the products derived from volunteer donations, maintaining meticulous records of vital information, including identity of donors, blood-screening procedures, processing, and dates. They also keep very precise records of the issue of each pack of blood product, including its storage, despatch from blood bank, destination, and the patient who will receive it. Thus, while much information is routinely available about the use of packs of drug products, even within individual hospitals over comparatively short periods, there is a paucity of information about the recipients. Such information is now of vital importance when implementing hemovigilance, a surveillance system for monitoring transfusion safety, through notification of unexpected or adverse events linked to transfusion. This problem was addressed in a landmark study in France by taking a random sample of blood recipients using a complex multistage sampling process [6]. At the first stage, the 39 regional blood centers in France were stratified into small, medium, and large, annual (1997) total of blood units distributed, and five or six centers were randomly selected within each stratum. The selected centers provided lists of the public and private hospitals they each supplied, and the hospitals themselves were then stratified into two groups according to small or medium versus large number of units used. The second stage sampling randomly selected hospitals within these two strata. Within each hospital, all patients who received a transfusion during a specified calendar period of 7 or 14 days were

noted, and details of the first transfusion recorded were analyzed. The recipients were characterized by age, medical history, and ICD-10 diagnostic category, and provided, for the first time, detailed information about transfusions at the population level.

The first population-based survey of the use of red cells in the United Kingdom, was reported in a study confined to north England in 2002 [11]. It was based on transfusion information covering two 14-day periods in 1999 and 2000, and reported broad indications for red cell transfusion, as well as distribution of use among surgical procedures. In addition, it provided age-specific use of red cells, and projected regional demand in 2008. Such surveys will become increasingly important during the next decade, and will require extension to longitudinal designs to assess the consequences of multiple transfusions, and to study survival.

Reducing the amount of unnecessary blood transfused is a key aspect of improving blood safety and can be implemented by the augmentation of current guidelines to rationalize the indications for transfusion. One method is to reduce the trigger for transfusion to lower thresholds. In a Canadian trial, Hébert et al. randomized 838 critical care patients between 1994 and 1997 to either a restrictive red cell transfusion strategy, where they were transfused if their hemoglobin fell to less than 7 g/dL, or a liberal strategy, where transfusion was administered if the hemoglobin fell to less than 10 g/dL [5]. They found no evidence that the restrictive policy was worse in terms of 30-day mortality, with 19% deaths in the restrictive group versus 23% in the liberal arm. Carson et al. carried out a meta-analysis of nine red cell trigger trials in critical care, cardiac, and orthopedic surgery and found that where a lower trigger was used, the probability of receiving a transfusion was reduced without affecting mortality except in patients with serious cardiac disease [2].

The same logic has been applied to other blood components. A trial carried out by Rebulla et al. in Italy randomized 276 patients between 1994 and 1996 with acute myeloid leukemia to either receive platelet transfusions when their platelet count fell below 10×10^9/L, or when it fell below 20×10^9/L [7]. They found no difference in the risk of major bleeding, which was 22% amongst those assigned the lower threshold compared to 20% in the liberal group, with no evidence of a difference in mortality. Of all the blood components, and despite its widespread use, least is known about the efficacy of fresh frozen plasma, useful in treating some congenital deficiencies of clotting factors, and thrombotic thrombocytopenic purpura; its ability to provide benefit in major bleeding remains unconfirmed. A meta-analysis of trials in the use of fresh frozen plasma (FFP) in cardiac surgery found no evidence that its use resulted in reduced blood loss compared to patients receiving no FFP or placebo infusions [3]. Trials in transfusion medicine present more practical issues than standard drug trials, for example, the composition of a component can differ between countries, and timing of randomization also presents difficulties, as a transfusion may not ultimately be needed in all eligible patients. It is also almost always impossible to achieve blinding due to the essential identity checks on the blood bag at the time of transfusion. Clinicians may also be reluctant to randomize patients to receive transfusion or no transfusion, in some clinical situations in which there is limited evidence of benefit; dependent on their perception of risks and benefits involved, some may decide that it is unethical to transfuse, whereas others could take the opposite view.

Limitations of Current Evidence

The clear-cut benefits of transfusing blood in emergency situations, and of routine transfusions in diseases such as leukemia, together with understanding of the severity of outcomes resulting from transfusion of the wrong blood, have led to transfusion medicine becoming a major speciality of medicine. However, the evidence base for current practice is still very limited; there have been very few randomized clinical trials, many of which are underpowered, and poorly designed and conducted.

While there is no question that transfusion can rectify blood loss and imbalances caused by some haematological diseases, or that the biochemical and immunological properties of blood have been proven, it remains to be shown whether the transfusion of blood or blood components is always effective in the treatment of certain diseases and conditions, and whether the benefits of transfusion outweigh its risks.

References

[1] Marcella, C. *ABC of Transfusion*, 3rd Ed. BMJ Books, London, 1998.

[2] Carson, J.L., Hill, S., Carless, P., Hébert, P. & Henry, D. (2002). Transfusion triggers: a systematic review of the literature, *Transfusion Medicine Reviews* **16**, 187–199.

[3] Casbard, A.C., Williamson, L.M., Murphy, M.F., Rege, K. & Johnson, T. (2004). The role of prophylactic fresh frozen plasma in decreasing blood loss and correcting coagulopathy in cardiac surgery. A systematic review, *Anaesthesiology* **59**, 550–558.

[4] McClelland, D. *Handbook of Transfusion Medicine*, 3rd Ed. The Stationary Office, London, 2001.

[5] Hébert, P.C., Wells, G., Blajchman, M.A., Marshall, J., Martin, C., Pagliarello, G., Tweeddale, M., Schweitzer, I. & Yetisir, E. (1999). A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care, *New England Journal of Medicine* **340**, 409–417.

[6] Mathoulin-Pélissier, S., Salmi, L.R., Verret, C. & Demoures, B., for the RECEPT investigators. (2000). Blood transfusion in a random sample of hospitals in France, *Transfusion* **40**, 1140–1146.

[7] Rebulla, P., Finazzi, G., Marangoni, F., Avvisati, G., Gugliotta, L., Tognoni, G., Barbui, T., Mandelli, F. & Sirchia, G. (1997). The threshold for prophylactic platelet transfusions in adults with acute myeloid leukemia, *New England Journal of Medicine* **337**, 1870–1875.

[8] Regan, F. & Taylor, C. (2002). Blood transfusion medicine, *British Medical Journal* **325**, 143–147.

[9] Starr, D. (2000). *Blood: An epic history of medicine and commerce*. Warner Books, London.

[10] Vamvakas, E.C. & Blajchman, M.A. (2001). Deleterious clinical effects of transfusion-associated immunomodulation: fact or fiction? *Blood* **97**, 1180–1195.

[11] Wells, A.W., Mounter, P.J., Chapman, C.E., Stainsby, D. & Wallis, J.P. (2002). Where does blood go? Prospective observational study of red cell transfusion in north England, *British Medical Journal* **325**, 803–804.

ANGELA CASBARD

Transition Models for Longitudinal Data

Suppose that the sequence of random variables $Y_1, Y_2, \ldots, Y_T$ represents the observations from a subject in a longitudinal trial or study. In a transition model, the distribution of each variable is considered *conditionally* on previous outcomes in the sequence. That is, the model represents the behavior of *changes* from the previously established position. More formally, the model is expressed in terms of the **conditional** functions:

$$f(Y_t|y_1, \ldots, y_{t-1}), \quad t = 2, \ldots, T.$$

In practice, a model may use only the most recent history of the process. For example, in a first-order model:

$$f(Y_t|y_1, \ldots, y_{t-1}) = f(Y_t|y_{t-1}), \quad t = 2, \ldots, T.$$

A general discussion of transition regression models can be found in [1].

Such a construction should be contrasted with a *marginal* regression model in which the behavior of Y_t is considered after averaging over the possible outcomes at all other times. Unlike the marginal model, the transition model has no representation in terms of cross-sectional data. Generally, the parameters in analogous regression structures embedded in the two types of model will differ in their interpretation and the choice of the appropriate type should be based on the inferences that are required from the analysis.

Due to the conditional form of the transition model, the associated likelihood is typically easy to construct using the chain rule for probabilities. As a consequence, the models are often comparatively straightforward to use in practice. Applications with binary data, for which the transition models represent **Markov chains**, are widespread. See, for example, [2] and [3]. In contrast, the marginal distribution of Y_t is typically a very complicated function of the parameters of the defining transition model. This may have served to inhibit the theoretical development of the models, which is, at present, less extensive.

References

[1] Diggle, P.D., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

[2] Korn, E.L. & Whittmore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution, *Biometrics* **35**, 795–802.

[3] Zeger, S.L., Liang, K.-Y. & Self, S.G. (1985). The analysis of binary longitudinal data with time-independent covariates, *Biometrika* **72**, 31–38.

(*See also* **Multivariate Methods for Binary Longitudinal Data**)

M.G. KENWARD

Transplantation

The first successful human corneal transplant was around 1900 by Zirm, corneal xenografting having been tried as early as 1837 but without notable success. From the early 1950s, Calne envisaged kidney transplantation as practicable therapy; and surgical, immunological, and immunosuppressive advances made it so by the mid-1980s. By 1970, Barnard had pioneered heart transplantation; and liver transplantation was also under way. Xenografting from transgenic pigs is the challenge of the next decades, together with improvements in unrelated donor bone marrow transplantation and therapeutic exploitation of stem cell banking. Interventional ventilation has not been proceeded with, and there has been more emphasis on realizing national potentials for living related kidney transplantation together with ethical and matching safeguards for unrelated living renal transplantation.

By 1990, transplantation had achieved one-year graft survival rates of 80% or more for most solid organs, and has done so through surgical innovation, advances in immunosuppression, beneficial and favourable matching of kidney donor to recipient, better preservation solutions, and by studying center variation in donor rates as well as in transplant outcome. In the 1990s, shortage of cadaveric donor organs has been a limiting factor that use of split livers or of domino heart transplants from cystic fibrosis heart-lung block recipients has mitigated only very partially. Epidemiological studies monitor malignancies secondary to immunosuppression. **Quality of life** as well as length of life (*see* **Life Expectancy**) is improved by transplantation.

Statistical science has underpinned most of this progress. Well-conducted randomized controlled trials (*see* **Clinical Trials, Overview**) of new immunosuppression therapies and preservation fluids have been published [26]; there has been occasional but critical early stopping of trials, because of overimmunosuppression, on the basis of **surrogate endpoints** of rejection episodes and major infections [16]. Proposals for the design and analysis of randomized trials with recurrent events [3] have had application in kidney transplantation. Two small trials in bone marrow transplantation were used to illustrate a new statistical measure to aid in the interpretation of published trials [1].

Beneficial matching [9, 12]; that is, the rules by which cadaveric donor kidneys have been exchanged in the UK, had a statistical basis and has persisted for 10 years up to 1997 when extended, also on statistical grounds, to favorable matching. Similar work on matching and matchability (see below) has been done independently by Mickey and colleagues [14, 24]. **Validation studies** have featured, whether in independent data sets (matching effects in distinct epochs of follow-up [11, 29, 31]) or by **meta-analysis** (DR mismatching in corneal transplantation [19]). Matchability score, dependent upon human leukocyte antigen (HLA) phenotype and exchange rules, for patients on the kidney transplant waiting list was introduced by Gilks [8, 10, 12] to summarize a patient's chance of getting a well-matched donor kidney in two or five years, and hence to aid individual decision-making on whether to accept or reject an offered kidney.

Special studies such as Corneal Transplant Follow-up Study (CTFS) and International Marrow Unrelated Search and Transplant (I MUST) Study have been set up to establish the core data that national registries (*see* **Disease Registers**) should seek to collect because they determine either waiting times [20], tissue allocation or prognosis [21] or quality of outcome, for which visual acuity is a natural measure [30]. In the I MUST Study, minimization, as in randomized trials, was adapted to select prospectively a **control** cohort of twice as many HLA-identical sibling transplants to correspond to the unrelated donor transplants in terms of marginal frequency for age group, diagnosis, risk, and transplant center. A second aspect of the design of the I MUST Study is noteworthy: in unrelated bone marrow donor searches, the patients for whom the search procedure finds an unrelated HLA-identical donor are effectively selected by "genetic randomization", which has broader epidemiological application than in studies of transplantation, for example, to understanding environmental determinants of disease [5]. A **time-dependent covariate** indicator (or several to account fully for nonproportionality of hazards (*see* **Proportional Hazards, Overview**) post transplant) can be switched on at that time and, by following all patients for whom an unrelated donor search was initiated, the effect of unrelated HLA-identical bone marrow transplantation against alternative management can be estimated in an **unbiased** manner. Effective **randomization** makes

the proposed analysis even more powerful than the use of a time-dependent indicator to switch patients from "awaiting cardiac transplantation" to "recipient status" [23], leading to appropriate analyses of the cost-effectiveness of heart transplantation (*see* **Health Economics**).

Cardiothoracic transplantation has posed other important statistical problems, including analysis of repeated biopsies after cardiac transplantation [28], informative **censoring** of quality-of-life measurements [4], and individualization of cyclosporine dose by monitoring the variability of cyclosporine blood levels and also the patient's kidney and liver function [2]. Kalman filter techniques [22], applied to weight-adjusted reciprocal creatinine for detection of kidney rejection episodes, were pioneering but did not become routine, perhaps because they were developed before cyclosporine. Sharples [13, 27] used a Gibbs sampling approach (*see* **Markov Chain Monte Carlo**) to modeling the longer-term risk of developing coronary occlusive disease after heart transplantation and thereby showed that there were particularly high transition intensities from mild to severe disease and from severe disease to death; thus, once mild disease developed, a patient's deterioration was rapid and research should focus on reducing progression from mild to severe disease.

Renal graft failure rates have been published in the UK on a center-anonymized basis since the early 1970s. This tradition in transplantation was in contradistinction to center-identified "name and shame" publication of performance data in the public services that took hold in UK in the late 1990s. Dissemination strategies were among the issues reported in October 2003 by a Royal Statistical Society Working Party on Performance Monitoring in the Public Services (see `www.rss.org.uk` for "Performance Indicators: Good, Bad, and Ugly"). Center variation has reduced considerably in the post-cyclosporine era [7] and further analysis by the confidence ranking methods developed by Goldstein & Spiegelhalter [15] would allow comparison of centers over calendar time, with or without adjustment for **case mix**, but taking account of center **covariates** such as whether a department of transplant immunology or transfusion medicine was responsible for tissue typing and cross-matching. In renal transplantation where centers' policies, let alone practice, on acceptance of older or asystolic donors, adherence to favorable matching, and retransplantation of older or diabetic or highly sensitized recipients may differ greatly, there is merit in no adjustment for case mix on the basis that the case mix is effectively center-determined. Ohlssen [25] used center variation in mortality after transplantation as one of three running examples that motivated a predominantly Bayesian analytical framework, easily programmable in WINBUGS, for the identification of unusual performance in a limited number of, versus many, centers.

Donor statistics are as important in transplantation as understanding the determinants of graft outcome. Confidential audit of all deaths in intensive care units in England and Wales in 1989–1990 [17, 18] showed that the second reason, after relatives' refusal, for missed suitable organs differed for the different organs – e.g. failure to ask in the case of kidneys but nonprocurement of offered suitable livers. That confidential audit also showed that even if all potential kidney donors in intensive care units became actual donors the need for cadaveric kidneys would not be met. Since then, the problem of nonprocurement of donor livers has been solved by designation of new centers but the shortage of donor kidneys has been exacerbated by the successful introduction of rear seat-belt legislation which saves lives. UK's most recent donor audit, begun in 2003, additionally records ethnicity because special allocation measures have had to be introduced to achieve better equity for blood group B patients on the kidney transplant waiting list, and differential relatives' consent rate by ethnicity needs to be investigated. Worryingly, preliminary results suggest that, overall, relatives' consent rate in 2003 had reduced markedly compared to 1990. Reduced altruism may be an adverse consequence of an organ-retention scandal emanating from the Royal Liverpool Children's NHS Trust, which has led to a revised Human Tissue Bill [6].

References

[1] Begg, C.B. (1985). A measure to aid in the interpretation of published clinical trials, *Statistics in Medicine* **4**, 1–10.

[2] Best, N.G., Trull, A.K., Tan, K.K., Spiegelhalter, D.J., Cary, N. & Wallwork, J. (1996). Pharmacodynamics of cyclosporine in heart and heart-lung transplant recipients I: blood cyclosporine concentrations and other risk factors for cardiac allograft rejection, *Transplantation* **62**, 1429–1435.

[3] Cook, R.J. (1995). The design and analysis of randomized trials with recurrent events, *Statistics in Medicine* **14**, 2081–2098.

[4] Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Jones, D.R. & Spiegelhalter, D.J. (1992). Quality of life assessment: can we keep it simple? (with discussion), *Journal of the Royal Statistical Society, Series A* **155**, 353–393.

[5] Davey Smith, G. & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? (30th Thomas Francis Jr. Memorial Lecture), *International Journal of Epidemiology* **32**, 1–22.

[6] Furness, P. & Sullivan, R. (2004). The human tissue bill. Criminal sanctions linked to opaque legislation threaten research, *British Medical Journal* **328**, 533–534.

[7] Gilks, W.R. (1987). Some applications of hierarchical models in kidney transplantation, *Statistician* **36**, 127–136.

[8] Gilks, W.R. (1991). Tissue matching and matchability in kidney transplantation, *Applied Statistics* **40**, 317–336.

[9] Gilks, W.R., Bradley, B.A., Gore, S.M. & Klouda, P.T. (1987). Substantial benefits of tissue matching in renal transplantation, *Transplantation* **43**, 669–674.

[10] Gilks, W.R., Gore, S.M. & Bradley, B.A. (1988). Matchability in kidney transplantation, *Tissue Antigens* **32**, 121–129.

[11] Gilks, W.R., Gore, S.M. & Bradley, B.A. (1990). Renal transplant rejection – transient immunodominance of HLA mismatches, *Transplantation* **50**, 141–146.

[12] Gilks, W.R., Gore, S.M. & Bradley, B.A. (1991). Predicting match grade and waiting time to kidney transplantation, *Transplantation* **51**, 618–624.

[13] Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. & Kirby, A.J. (1993). Modelling complexity: applications of Gibbs sampling in medicine, *Journal of the Royal Statistical Society, Series B* **55**, 39–52.

[14] Gjertson, D.W., Terasaki, P.I., Takemoto, S. & Mickey, M.R. (1991). National allocation of cadaveric kidneys by HLA matching. Projected effect on outcome and costs, *New England Journal of Medicine* **324**, 1032–1036.

[15] Goldstein, H. & Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion), *Journal of the Royal Statistical Society, Series A* **159**, 385–444.

[16] Gore, S.M. (1995). Statistical thinking and when to stop a clinical trial, in *Logic in Medicine*, 2nd Ed., C.I. Phillips, ed. British Medical Journal, London, pp. 116–132.

[17] Gore, S.M., Taylor, R.M.R. & Wallwork, J. (1991). Availability of transplantable organs from brain stem dead donors in intensive care units, *British Medical Journal* **302**, 149–153.

[18] Gore, S.M., Cable, D.J. & Holland, A.J. (1992). Organ donation from intensive care units in England and Wales: two year confidential audit of deaths in intensive care, *British Medical Journal* **304**, 349–355.

[19] Gore, S.M., Vail, A., Bradley, B.A., Rogers, C.A., Easty, D.L. & Armitage, W.J. (1995). HLA-DR matching in corneal transplantation: systematic review of published evidence, *Transplantation* **60**, 1033–1039.

[20] Howard, M.R., Gore, S.M., Hows, J.M., Downie, T.R. & Bradley, B.A. (1995). A prospective study of factors determining the outcome of unrelated marrow donor searches: report from the International Marrow Unrelated Search and Transplant Study Working Group on behalf of collaborating centers, *Bone Marrow Transplant* **15**, 499–503.

[21] Hows, J., Bradley, B.A., Gore, S., Downie, T., Howard, M. & Gluckman, E. (1993). The International Marrow Unrelated Search and Transplant (I MUST) Study, *Bone Marrow Transplant* **12**, 371–380.

[22] Knapp, M.S., Smith, A.F.M., Trimble, I.M., Pownall, R. & Gordon, K. (1983). Mathematical and statistical aids to evaluate data from renal patients, *Kidney International* **24**, 474–486.

[23] Mantel, N. & Byar, D.P. (1974). Evaluation of response-time data involving transient states: an illustration using heart transplant data, *Journal of the American Statistical Association* **69**, 81–86.

[24] Mickey, M.R. (1985). HLA matching in transplants from cadaver donors, in *Clinical Kidney Transplants*, P.I. Terasaki, ed. UCLA Tissue Typing Laboratory, Los Angeles, pp. 45–56.

[25] Ohlssen, D.I. (2004). Methodological issues in the use of random effects modles for comparisons of health care providers. University of Cambridge, PhD Thesis.

[26] Ploeg, R.J., van Bockel, J.H., Langendijk, P.T., Groenewegen, M., van der Woude, F.J., Persijn, G.G., Thorogood, J. & Hermans, J. (1992). Effect of preservation solution on results of cadaveric kidney transplantation. The European Multicentre Study Group, *Lancet* **340**, 129–137.

[27] Sharples, L.D. (1993). Use of Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation, *Statistics in Medicine* **12**, 1155–1170.

[28] Spiegelhalter, D.J. & Stovin, P.G.I. (1983). An analysis of repeated biopsies following cardiac transplantation, *Statistics in Medicine* **2**, 33–40.

[29] Thorogood, J., Persijn, G.G., Schreuder, G.M., d'Amaro, J., Zantvoort, F.A., van Houwelingen, J.C. & van Rood, J.J. (1991). The effect of HLA matching on kidney graft survival in separate post-transplantation intervals, *Transplantation* **50**, 146–150.

[30] Vail, A., Gore, S.M., Bradley, B.A., Easty, D.L. & Rogers, C.A. on Behalf of Corneal Transplant Follow-up Study Collaborators (1994). Corneal graft survival and visual outcome, *Ophthalmology* **101**, 120–127.

[31] Van Houwelingen, H.C. & Thorogood, J. (1995). Construction, validation and updating of a prognostic model for kidney graft survival, *Statistics in Medicine* **14**, 1999–2008.

Further Reading

Bird, S.M. (2004). Recipients of blood or blood products "at vCJD risk". We need to define their rights and responsibilities and those of others, *British Medical Journal* **328**, 118–119.

Chadeau-Hyam, M., Tard, A., Bird, S., le Guennec, S., Bemrah, N., Volatier, J.-L. & Alperovitch, A. (2003). Estimation of the exposure of the French population to the BSE agent: comparison of the 1980–95 consumption of beef products containing mechanically recovered meat in France and the UK, by birth cohort and gender, *Statistical Methods in Medical Research* **12**, 247–260.

Cooper, J.D. & Bird, S.M. (2003). Predicting incidence of variant Creutzfeldt-Jakob disease from UK dietary exposure to bovine spongiform encephalopathy for the 1940 to 1969 and post-1969 birth cohorts, *International Journal of Epidemiology* **32**, 784–791.

Cousens, S., Everington, D., Ward, H.J.T., Huillard, J., Will, R.G. & Smith, P.G. (2003). The geographical distribution of variant Creutzfeldt-Jakob disease cases in the UK: what can we learn from it? *Statistical Methods in Medical Research* **12**, 235–246.

Gravenor, M.B., Stallard, N., Curnow, R. & McLean, A.R. (2003). Repeated challenge with prion disease: The risk of infection and impact on incubation period, *PNAS* **100**(19), 10960–10965 (see also `www.pnas.org/cgi/doi/10.1073/pnas.1833677100`).

Houston, F., Foster, J.D., Chong, A., Hunter, N. & Bostock, C.J. (2000). Transmission of BSE by blood transfusion in sheep, *Lancet* **356**, 999–1000.

Huillard d'Aignauz, J.N., Cousens, S.N., Maccario, J., Costagliola, D., Alpers, M.P., Smith, P.G., Alperovitch, A. (2000). The incubation period of kuru, *Epidemiology* **13**(4), 402–408.

Hunter, N., Foster, J. & Chong, A et al. (2002). Transmission of prion diseases by blood transfusion. *Journal of General Virology* **83**, 2897–2905.

Llewelyn, C.A., Hewitt, P.E., Knight, R.S.G., Amar, K., Cousens, S., Mackenzie, J. & Will, R.G. (2004). Possible transmission of variant Creutzfeldt-Jakob disease by blood transfusion, *Lancet* **363**, 417–421.

Scientific Steering Committee. (2000). *Opinion on Geographical Risk of Bovine Spongiform Encephalopathy (GBR)*. European Commission, Health and Consumer Protection Directorate-General, Brussels, July 2000.

Scientific Steering Committee. (2003). *Opinion and Report on BSE in Great Britain's cattle born after 31 July 1996 [BARBs]*. European Commission, Health and Consumer Protection Directorate-General, Brussels, March 2003.

Valleron, A.J., Boelle, P.Y., Will, R. & Cresbon, J.Y. (2001). Estimation of epidemic size and incubation time based on age characteristics of vCJD in the United Kingdom, *Science* **294**, 1726–1728.

S.M. GORE

Travel Medicine

The subject of travel medicine is aimed at preventing illnesses associated with travel or travelers. The speciality provides advice and protection for travelers, immigrants, and refugees, and covers issues relating environmental, climatic, physical, and infectious agents during or after a journey. In the nineteenth century, colonial expansion led to the development of the speciality of tropical medicine in order to maintain the health of colonists working in the tropics. As tropical medicine changed with world politics, it became clear that preventing disease dissemination and transmission of cholera, smallpox, and yellow fever required international control and regulation. The **World Health Organization** enacted statutes requiring travelers to show certificates of immunization from registered centers for smallpox, yellow fever, and cholera. On the eradication of smallpox, with control of yellow fever, and with the recognition that cholera was not controlled by immunization, vaccination centers focused on other common travel health problems. Research focused on the epidemiology of diseases imported by travelers, particularly diseases preventable by vaccines or prophylactic drugs. Steffen [6] quantified health problems that afflicted Swiss travelers, and noted that travelers' diarrhea was the most frequent cause of illness. Surveillance studies in travelers revealed that vaccine preventable illnesses such as polio, typhoid, tetanus, and hepatitis B occurred very rarely. Patterns of illness were not always similar across nations; for example, hepatitis A in UK travelers was significantly lower than the incidence of hepatitis A in Swiss travelers [2, 6]. For problems where preventive measures existed, investigators established their effectiveness and safety.

Malaria is important, as it poses a threat to most tropical travelers, and therefore its prevention has attracted much research. Using surveillance reports, Phillips-Howard [5] described the pattern of malaria imported into the UK and factors associated with a high risk of infection. Ethnic travelers visiting West Africa to visit friends and relatives were at particularly high risk. It was also noted that region of travel and the reason for travel and use of a chemoprophylactic regimen had a bearing on the risk of developing malaria. Large-scale questionnaire-based surveillance studies of returning and returned travelers allowed estimates of protective efficacy of various malaria chemoprophylaxis regimens to be made. The largest study, of 145 500 European travelers returning from Mombassa in Kenya, undertaken by Steffen et al. [7], reported comparative prophylaxis effectiveness of mefloquine 91% (95% **confidence interval** 85–94) to chloroquine/proguanil 72% (95% CI 56–82). The study did not report CI between the drug regimens. The authors identified at least six different malaria regimens used by tourists visiting Kenya.

The study was also used to estimate adverse events associated with chemoprophylaxis use. Problems associated with this and other studies of adverse events are in the definitions of an adverse event and in using these definitions across studies. In this cohort, serious was defined as life threatening, disabling, or fatal, and occurred in 1 in 10 000 travelers using mefloquine. In a more recent study of adverse events associated with malaria prophylaxis, Barrett et al. [1] describe severe adverse reactions as those associated with hospital admission and define a disabling reaction as that which interferes with normal daily activity. In relation to mefloquine use, the incidence of severe reactions was 0.5%, and disabling side-effects were reported to occur in 0.7% of users.

Much confusion now exists on the true rate of adverse events and how they should be defined. The size and cost of such projects has restricted subsequent confirmatory or supportive studies. The study relied on adverse events reported on return and on follow-up. This design meant that individuals who developed reactions that prevented them from traveling (most travelers start prophylaxis a week or two before traveling) would not be recognized, a design fault that may have significantly **biased** the true estimates of adverse events. Evaluations of interventions on avoidance of travelers' diarrhea, looking at the impact of advice on food and water hygiene and the benefits of chemoprophylaxis against a problem that affected up to 50% of travelers, have been undertaken by many clinicians. Advice to change travelers' eating and drinking had no impact on diarrhea incidence [4], but antibacterial chemoprophylaxis used in a double blind (*see* **Blinding or Masking**) controlled **clinical trial** in students visiting Mexico had a significant impact. The method of **randomization** was unclear, but two groups received a daily antibacterial agent and the third received a placebo. One third of the placebo group suffered

diarrhea in a 14 day period, while only one of the subjects receiving trimethoprim–sulphamethoxazole had diarrhea [3]. More recent research methodologies have focused on the risk to benefits of interventions (*see* **Data Monitoring Committees**), especially where drugs/vaccines have appreciable toxicity and the risk of infection is variable. As many vaccines and drugs are costly, the cost–benefit [2] of preventive measures is also under investigation (*see* **Health Economics**). Many of the studies examining interventions use standard statistical methods, and the study design is occasionally difficult, as monitoring events in traveling subjects requires innovative techniques.

References

[1] Barrett, P.D., Emmins, P.D., Clarke, P.D. & Bradley, D.J. (1996). Comparison of adverse events associated with use of mefloquine and combination of chloroquine and proguanil as antimalarial prophylaxis: postal and telephone survey of travellers, *British Medical Journal* **313**, 525–528.

[2] Behrens, R.H. & Roberts, J.A. (1994). Is travel prophylaxis worth while? Economic appraisal of prophylactic measures against malaria, hepatitis A, and typhoid in travellers, *British Medical Journal* **309**, 918–922.

[3] DuPont, H.L., Galindo, E., Evans, D.G., Cabada, F.J., Sullivan, P. & Evans, D.J. (1996). Prevention of travelers' diarrhoea with trimethoprim–sulfamethoxazole and trimethoprim alone, *Gastroenterology* **84**, 75–80.

[4] Kozicki, M., Steffen, R. & Schar, M. (1985). "Boil it, cook it, peel it or forget it": does this rule prevent travellers' diarrhoea? *International Journal of Epidemiology* **14**, 167–172.

[5] Phillips-Howard, P.A., Bradley, D.J., Blaze, M. & Hurn, M. (1988). Malaria in Britain: 1977–86, *British Medical Journal* **296**, 245–248.

[6] Steffen, R. (1991). Travel medicine – prevention based on epidemiological data, *Transactions of the Royal Society of Tropical Medicine and Hygiene* **85**, 156–162.

[7] Steffen, R., Fuchs, E., Schildknecht, J., Funk, M., Schlagenhauf, P., Phillips-Howard, P.A., Nevill, C. & Sturchler, D. (1993). Mefloquine compared with other malaria chemoprophylactic regimens in tourists visiting East Africa, *Lancet* **341**, 1299–1303.

R.H. BEHRENS

Treatment Delay

Patients showing symptoms characteristic of a life-threatening disease are always advised to report to a physician without delay. Early reporting and subsequent diagnosis permit early treatment intervention. Prognosis is likely to be improved by early diagnosis and treatment, although the benefit may be small or zero if the disease has already progressed irrevocably at the time of diagnosis, or if no effective treatment is available.

The assertions made in the above paragraph cannot be confirmed by randomized trials (*see* **Clinical Trials, Overview**) because it would be both unethical and impracticable to require patients to incur unnecessary delay in reporting symptoms or receiving treatment. However, it seems clear on general grounds that delays in diagnosis or treatment cannot be beneficial except in the unlikely event that the treatment to be offered is actually harmful.

Paradoxically, many data sets appear to show the opposite [3]. Table 1 describes a historically interesting series of 950 cases of breast cancer operated on between 1889 and 1931 [2] (reported in [3]), of whom 420 had died by 1932. The mean duration of survival after operation is related to the preoperative delay. There is little effect of delays less than three years, but four-year delay appears to be associated with longer postoperative survival. The effect is more remarkable when survival is measured from onset of symptoms, since patients in the last group in Table 1 survived some five years longer, after onset of symptoms, than those in the first group.

Other series reported in [3] show higher five-year postoperative survival rates in patients with very short delays or very long delays, with poorer results for the intermediate lengths of delay.

Table 1 Preoperative delay and mean duration of survival after operation, in a series of patients with breast cancer

Preoperative delay (months)	Number of patients	Mean duration of survival after operation (years)
0–	66	3.87
3–	67	2.86
6–	50	3.07
9–	66	2.57
12–	74	3.54
24–	20	3.65
36–48	17	4.91

From [2], quoted in [3, p. 163].

Table 2 Delay in admission to hospital and case–fatality rate, in patients with clinical tetanus

Time from first symptom to admission (hours)	Deaths/total (% fatality rate)
1–9	53/89 (60)
10–18	90/150 (60)
19–36	120/302 (40)
37–72	69/388 (18)
73–144	31/262 (12)
145–	5/90 (6)

Based on [1, Table 3].

It would be easy, but wrong, to conclude from such evidence that patients benefit by deliberate delay in seeking diagnosis and treatment. Long delay will tend to occur with relatively slow progression of disease (for instance, with a slowly growing tumor), and patients with this type of disease will tend to survive longer after onset of symptoms and longer after initiation of treatment. The malignancy of the disease, with the consequent rapidity of progression of symptoms, acts as a **confounder**, being associated negatively with both pretreatment delay and survival. A slight reversal of the positive **correlation** between delay and survival may occur for the group of patients with the very shortest delays, because this group may include some patients whose prognosis is improved by early access to effective treatment.

A similar phenomenon was noted in a study [1] of prognosis in patients with clinical tetanus. Table 2 shows the relation between the time from first symptoms to admission to hospital, and the case–fatality rate. Patients with a delay in admission of less than 10 hours had a fatality rate of 60%, whereas those with a delay longer than 145 hours had a fatality rate of 6%. There is great variation in the rapidity of development of symptoms in clinical tetanus, and mildly affected patients may experience delay in admission but nevertheless have a favorable prognosis.

References

[1] Armitage, P. & Clifford, R. (1978). Prognosis in tetanus: use of data from therapeutic trials, *Journal of Infectious Diseases* **138**, 1–8.

[2] Lewis, D. & Rienhoff, W.F., Jr (1932). A study of the results of operations for the cure of cancer of the breast, *Annals of Surgery* **95**, 336–400.
[3] Sutherland, R. (1960). *Cancer: The Significance of Delay*. Butterworth, London.

PETER ARMITAGE

Treatment-covariate Interaction

A *treatment–covariate interaction* (TCI) is said to exist when the effect of a treatment varies according to the value of a specified **covariate**, the latter being a function of one or more patient characteristics measured at baseline. The existence of such an **interaction** is important for the clinical practice of medicine, because it implies that the optimal choice of treatment differs for different patients. For example, knowledge of aspects of the patient's history and clinical presentation may permit a better treatment selection than would be possible without that knowledge.

The usual objective of a controlled **clinical trial** is to study the effects of a particular treatment given to patients of a particular type. The main conclusion from the trial is usually assumed to relate to any persons who meet the trial's eligibility criteria (*see* **Eligibility and Exclusion Criteria**). For instance, Fischl et al. [4] reported the results of a trial of two drug regimens for delaying disease progression in patients with "advanced" HIV disease, where the investigators defined "advanced" operationally as either symptomatic disease with a CD4 cell count of not more than 300 cells/mm^3 or asymptomatic disease with a CD4 cell count of not more than 200 cells/mm^3.

Two natural questions arise. First, do the results of this trial apply equally to all of the types of persons represented in the study? Because individuals differ in an unlimited number of ways, there is unfortunately no certain answer to the question. However, as we describe below, one can address a more limited question when suitable data are in hand.

Secondly, one might also inquire whether the results of this trial apply to some types of persons who were excluded from entering. For example, do the results apply to individuals with "limited" HIV disease? This second question has to do with study generalizability and involves a somewhat different set of issues.

Interaction Modeling

To approach answering the first question in a limited manner, suppose that the data from the trial include variables indicating not only the treatment and the clinical outcome, but also patient characteristics measured at baseline, such as gender, age, overall health status (such as the Karnofsky performance status score), stage or extent of disease, and other factors that may be prognostic.

A TCI may be defined, estimated, and tested for significance within the framework of commonly used **regression** models. To keep the exposition relatively simple, suppose that the aim of treatment is to affect the average value of outcome which is **normally distributed**, namely, $Y \sim \mathrm{N}(\mu, \sigma^2)$. We choose a model that expresses the influence of the **binary** variable for treatment (x_1, say) and baseline covariates ($x_i, i = 2, \ldots, p$) on outcome through the relationship $\mu = \mathbf{x}'\boldsymbol{\beta}$.

Suppose that the covariate for age is x_2. To introduce an interaction between treatment and age into the model, define the new variable $x_{p+1} = x_1 x_2$, and append it to the list of variables $\mathbf{x}$ in the model $\mu = \mathbf{x}'\boldsymbol{\beta}$. **Mean** outcomes for individuals with covariate values $x_2, \ldots, x_{p+1}$ are $\beta_1 x_{11} + \beta_2 x_2 + \cdots + \beta_p x_p + \beta_{p+1} x_{11} x_2$ and $\beta_1 x_{12} + \beta_2 x_2 + \cdots + \beta_p x_p + \beta_{p+1} x_{12} x_2$, with x_{11} and x_{12} designating the two treatment variable values. The difference between these means, or treatment effects, is $\beta_1(x_{11} - x_{12}) + \beta_{p+1}(x_{11} - x_{12})x_2$, which depends on x_2 unless $\beta_{p+1} = 0$. If x_2 is a continuous covariate, β_{p+1} is the amount by which the treatment effect changes per unit change in x_2. As a special case, if x_2 is binary, taking values 0 and 1, or 1 and 2, say, β_{p+1} is the amount by which the treatment effect changes when the covariate changes from one level to the other. Then, to assess the statistical significance of a particular TCI, one can test the null hypothesis H_0 : $\beta_{p+1} = 0$, using standard methods (*see* **Hypothesis Testing**).

Because the data often include many covariates, one can test many TCIs. One set of choices involves the definitions of covariates to evaluate for characteristics with more than two possible categories, such as age or prior therapy. Another issue is whether to consider higher-order interactions. Note especially for continuous covariates that interactions that appear to be substantial may largely disappear under a **transformation** of the covariate (e.g. from a linear to a logarithmic scale).

Making these choices can leave the analyst with a large multiple testing problem (*see* **Multiplicity in Clinical Trials**), arousing suspicion that any large

observed interaction is spurious. Many writers (e.g. Buyse [1]) recommend limiting the analysis to a small number of TCIs, preferably those identified as plausible before the trial.

Qualitative Interactions

Qualitative interactions are said to exist "when the direction of the true treatment differences varies among subsets of patients" [6, p. 361]. Differences in the magnitudes, when all are in the same direction, are considered to be unremarkable. A change in direction, however, implies a change in recommended treatment. Interactions that are not qualitative are called *quantitative*. Quantitative interactions are model-dependent, because it is sometimes possible to remove them by a monotone transformation of the covariate. Qualitative interactions are model independent [2]. A focus on qualitative interactions will thus: (i) concentrate on cases involving different treatment preferences for different types of patients; (ii) reduce the interactions claimed that are really artifacts of the model chosen; and (iii) limit exposure to multiplicity effects.

Gail & Simon [6] derived a **likelihood ratio test** of the **null hypothesis** that there is no qualitative interaction when the covariate is categoric. Starting with statistically independent estimates $D_i, i = 1, \ldots, I$, of the treatment effect differences in each of I nonoverlapping categories, or subsets, and their **variances** σ_1^2, one calculates $\Sigma(D_i^2/\sigma_i^2)I(D_i > 0)$ and $\Sigma(D_i^2/\sigma_i^2)I(D_i < 0)$, where $I(S)$ is the indicator function for the set S. If both of these quantities exceed a critical value from Table 1 of Gail & Simon [6], which depends on both the nominal type I error rate (*see* **Hypothesis Testing**) and the number of categories being examined, then the null hypothesis is rejected.

Most workers in the field believe that qualitative interactions are quite uncommon. Indeed, it seems implausible that a treatment beneficial in one subgroup (e.g. young patients) would actually be harmful in another (e.g. the old), even though it would not be at all surprising if the average benefit of the treatment varied in magnitude among subgroups. Byar [2] did observe a qualitative interaction concerning the use of diethylstilbesterol (DES) for the treatment of prostate cancer. Patients with advanced stage disease who received DES had lower mortality from prostate cancer than their counterparts not receiving DES, and this was the predominant cause of death in this group. Patients with early stage disease who received DES also had lower prostate cancer mortality but higher mortality from cardiovascular causes than similar patients who did not receive DES. In this early stage disease group, prostate cancer did not predominate as the cause of death, and the increased mortality from cardiovascular diseases became the overriding factor. It would be wrong, Byar concluded, never to look beyond overall results.

Bayesian Approach

A Bayesian approach (*see* **Bayesian Methods**) can be useful in some situations. Starting with estimated regression coefficients at least approximately normally distributed, Dixon & Simon [3] proposed using **exchangeable** normal **priors** centered at zero for the regression coefficients corresponding to TCIs, and vague priors for the other regression coefficients. This model leads to a **shrinking** of estimates of subset-specific treatment differences toward the estimated overall treatment difference. The precision of the estimates, however, need not be reduced on account of multiplicity. Dixon & Simon [3] illustrated this feature of a Bayesian approach using data from a clinical trial of chemotherapy for colorectal cancer.

Subgroup Analysis

As indicated earlier, when describing the parametric model for studying TCIs, the real interest is in detecting subgroups of patients for whom the optimal treatment differs from the overall patient population. An alternative to studying interactions in a parametric model, one that has been used very often in the clinical research literature, is simply to compare treatments in each of a number of patient subsets or subgroups, and to highlight those subgroups in which the treatment difference attains conventional statistical significance.

Uncritical presentation of numerous subgroup-specific tests of a significant treatment effect is certainly to be discouraged. For one thing, clinical trials accrue sufficient participants to provide adequate precision for estimating quantities of primary interest, usually overall treatment effects. Confining attention to subgroups almost always results in

estimates of inadequate precision. Furthermore, the chances of making a type I error increase rapidly with the number of subgroups examined. Provided that the parametric model above holds, there is a one-to-one correspondence between nonzero regression coefficients for interaction terms and a collection of subgroup-specific treatment effects that vary from one subgroup to another. Testing hypotheses about interactions, however, is a much more efficient way to detect these variations than to compare treatments within many different subgroups.

In addition, there is usually good reason for regarding the overall treatment groups as comparable in a randomized trial. Subgroups may not enjoy the same degree of balance in patient characteristics, leading to apparent treatment differences, due to the two treatment groups within a subgroup having markedly different prognoses.

Finally, there are multiple ways to create subgroups. For example, one may examine four mutually exclusive subgroups determined by the cross classification of two binary covariates, as well as two pairs of subgroups determined by the two covariates considered marginally. The choice of subgroups to examine is thus almost boundless, and selection is inevitably somewhat arbitrary.

Gender and Minority Subgroups

In the past few years, assessment of TCIs has assumed new importance because of its relationship to questions of generalizability of findings from clinical trials and access to clinical trials by traditionally under-represented groups. If one assumes no TCIs, one is free to design a study of usual size, or one may even reduce the sample size by favoring volunteers at highest risk of the unfavorable outcome being observed. In the extreme, this leads to omitting whole classes of individuals; for example, women. The problem is that this leaves no possibility of later checking for the existence of TCIs using data from the trial. It therefore seems scientifically sensible to provide reasonable access to clinical trials of new treatments by all segments of the general population.

Unfortunately, a trial just large enough to evaluate an overall treatment effect reliably will almost inevitably lack precision for evaluating differential treatment effects between different population subgroups. **Meta-analyses** of similar trials may carry sufficient **power**, however, especially if one is interested mainly in qualitative interactions. For a thorough exposition of these issues, see Freedman et al. [5].

References

[1] Buyse, M.E. (1989). Analysis of clinical trial outcomes: some comments on subgroup analyses, *Controlled Clinical Trials* **10**, 187S–194S.

[2] Byar, D.P. (1985). Assessing apparent treatment-covariate interactions in randomized clinical trials, *Statistics in Medicine* **4**, 255–263.

[3] Dixon, D.O. & Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial, *Statistics in Medicine* **11**, 13–22.

[4] Fischl, M.A., Stanley, K., Collier, A.C., Arduino, J.M., Stein, D.S., Feinberg, J.E., Allan, J.D., Goldsmith, J.C., Powderly, W.B. and the NIAID AIDS Clinical Trials Group (1995). Combination and monotherapy with Zidovudine and Zalcitabine in patients with advanced HIV disease, *Annals of Internal Medicine* **122**, 24–32.

[5] Freedman, L.S., Simon, R., Foulkes, M.A., Friedman, L., Geller, N.L. & Mowery, R. (1995). Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993 – the perspective of NIH clinical trialists, *Controlled Clinical Trials* **16**, 277–285.

[6] Gail, M. & Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets, *Biometrics* **41**, 361–372.

DENNIS O. DIXON

Trees, Probabilistic Functional

Decision Trees

A decision tree (*see* **Computer-aided Diagnosis**) uses a divide-and-conquer strategy. It attacks a complex problem by dividing it into simpler problems and recursively applying the same strategy to the subproblems. The solutions of subproblems can be combined in a form of a tree to yield a solution of the complex problem. The power of this approach comes from the ability to split the instance space into subspaces and each subspace is fitted with different models. This *recursive partitioning* (*see* **Tree-structured Statistical Methods**) idea is behind well-known decision tree–based **algorithms**, such as CART [4] and C4.5 [17]. More recently several statistical packages, *S-PLUS*, *Statistica*, *SAS*, and *SPSS* (*see* **Software, Biostatistical**) [14] have incorporated functions that implement decision trees for classification and regression problems.

Formally, a decision tree is a direct acyclic graph in which each node is either a *decision node* with two or more successors or a *leaf node*. A *leaf node* is labeled with a *class*. A *decision node* has some *condition* based on attribute values. The hypothesis space of these algorithms is within the *disjunctive normal form* (DNF) formalism. Classifiers generated by those systems encode a DNF for each class. For each DNF, the conditions along a branch represent conjuncts and the individual branches can be seen as disjuncts. Each branch forms a rule with a conditional part and a conclusion. The conditional part is a conjunction of conditions. Conditions are tests that involve a particular attribute, operator (e.g. $=$, $\geq$, etc.) and a value from the domain of that attribute. These kinds of tests correspond, in the input space, to a hyperplane that is orthogonal to the axes of the tested attribute and parallel to all other axis. The regions produced by these classifiers are all hyperrectangles. Each leaf corresponds to a region. The regions are mutually exclusive and exhaustive (i.e. cover all the instance space). It is known that the problem of building a minimal decision tree (in terms of number of nodes), consistent with a set of data, is an *NP hard* problem [18]. Usually, algorithms exploit heuristics that locally perform a one-step lookahead search. Once a decision is taken, it is never reconsidered. This hill-climbing search without backtracking is sensitive to the usual risks of converging to locally optimal solutions that are not globally optimal. However, this strategy allows building decision trees in time *linear* to the number of examples.

The standard algorithm to build univariate trees consists of two phases. In the first phase, a large tree is constructed. In the second phase, this tree is pruned back. The algorithm to grow the tree follows the standard divide-and-conquer approach. The most relevant aspects are the splitting rule, the termination criterion, and the leaf assignment criterion. With respect to the last criterion, the usual rule consists of assignment of a constant to a leaf node. Considering only the examples that fall at this node, the constant is usually the constant that minimizes the **loss function**: the mode of y values in the case of classification problems or the mean of the y values in the regression setting. With respect to the splitting rule, we distinguish between nominal attributes and continuous ones. In the former, the number of partitions is equal to the number of values of the attribute; in the latter, a binary partition is obtained. To estimate the merit of the partition obtained by a given attribute, several heuristics have been used [4, 17]. A nice review appears in [13]. In any case, the attribute that maximizes the criterion is chosen as test attribute at this node. The pruning phase consists of traversing the tree in a depth-first order. At each non–leaf node two measures should be estimated. An estimate of the error of the subtree below this node, which is computed as a weighted sum of the estimated error for each leaf of the subtree, and the estimated error of the nonleaf node if it was pruned to a leaf. If the latter is lower than the former, the entire subtree is replaced to a leaf. All of these aspects have several and important variants (e.g. [4, 17]). Nevertheless, all decision nodes contain conditions based on the values of one attribute, and leaf nodes predict a constant.

Multivariate Trees

One of the most appealing extensions to the basic decision tree algorithm is the use of combinations of attributes in decision tree learning. One of the earliest works is CART [4] where a linear multivariate split is found using a hill-climbing search. An extension to

avoid local minima has been proposed in [15]. Other methods include the use of gradient descent to construct attribute combinations. In [9], in each decision node, a multilayer perceptron is trained with back-propagation leading to a binary split. In [5], in each decision node, a linear machine is trained leading to a k-way splitting. Methods using **linear programming** have been used, for example, in [1, 2]. All these methods are search intensive; they are prone to overfitting and local minima. The most successful methods seems to be those based on **discriminant analysis** that have been used in [6, 7, 11, 12]. A related research line explores the use of functional tree leaves. Functional tree leaves is almost the rule [16, 20, 8], in **regression** problems. In the classification setting, few works explore this idea [22, 10, 19, 8].

Probabilistic Functional Trees

Multivariate trees can be seen as a combination of multiple models. This idea has been fully explored in functional trees. Functional trees combine a standard univariate tree with a discriminant function by means of constructive induction. At each decision node, a discriminant function is built using the examples that fall at that node. Each of these examples is extended with new attributes computed as the probability that the example belongs to a class. The merit of each new attribute is evaluated, in competition with the original attributes, using the meritfunction of the univariate tree. If one of the new attributes is chosen by the merit function, this corresponds to a multivariate split. Functional trees use two types of decision nodes: those based on a test of one of the original attributes, and those based on the values of the discriminant function.

Once a tree has been constructed, it is pruned back. The basic mechanism consists of replacing a decision node by a leaf based on an estimation of the error. Functional trees consider two types of leaves. Those that predict a constant and those that make a prediction using the discriminant function stored at the node before pruning. The pruning algorithm produces two different types of leaves: *ordinary leaves* that predict a constant, and *discriminant leaves* that predict the value of the discriminant function learned (in the growing phase) at this node.

Conclusions

Functional trees extend and generalize multivariate trees. Functional trees generate hybrid models that combine a univariate tree with a discriminant function using a kind of local stacked generalization [23]. The components of the hybrid algorithm use different representation languages and search strategies. While the tree uses a divide-and-conquer method, a discriminant function performs a global minimization approach. While the former performs feature selection, the latter uses all (or almost all) the attributes to build a model. From the point of view of the bias-variance decomposition of the error [3], a decision tree is known to have low bias but high variance, while discriminant functions are known to have low variance but high bias. This is the desirable behavior for components of hybrid models. An extensive experimental study [8] has shown that functional trees are competitive algorithms both in terms of accuracy and learning times. An analysis of the bias-variance decomposition of the error shows that the use of multivariate decision nodes is a bias reduction process, while the use of multivariate leaves is a variance reduction process.

It is interesting to note that there are standard algorithms for topological transformations on decision trees: trees to decision rules [17], multivariate trees to multilayer networks. [21] This is an interesting aspect because it points to a common representation for different generalization languages.

References

[1] Bennet, K. (1992). Decision tree construction via linear programming, in Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, 97–101.

[2] Bennett, K. & Mangasarian, O. (1994). Multicategory discrimination via linear programming, *Optimization: Methods and Software* **3**, 27–39.

[3] Breiman, L. (1998). Arcing classifiers, *The Annals of Statistics* **26**(3), 801–849.

[4] Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group, USA.

[5] Brodley, C.E. & Utgoff, P.E. (1995). Multivariate decision trees, *Machine Learning* **19**, 45–77.

[6] Gama, J. (1997). Probabilistic linear tree, in *Machine Learning, Proceedings of the 14th International Conference*, D. Fisher, ed. Morgan Kaufmann, San Francisco, CA, pp. 134–142.

[7] Gama, J. (1999). Discriminant trees, in *Machine Learning, Proceedings of the 16th International Conference*,

I. Bratko & S. Dzeroski, eds. Morgan Kaufmann, San Francisco, CA, pp. 134–142.

[8] Gama, J. (2002). An analysis of functional trees, in *Machine Learning, Proceedings of the 19th International Conference*, C. Sammut, ed. Morgan Kaufmann, San Francisco, CA, pp. 155–162.

[9] Gelfand, S., Ravishankar, C. & Delp, E. (1991). An iterative growing and pruning algorithm for classification tree design, *Transactions on Pattern Analysis and Machine Intelligence* **13**(2), 163–174.

[10] Kohavi, R. (1996). Scaling up the accuracy of naive Bayes classifiers: a decision tree hybrid, in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, E. Simoudis, J. Wei Han, & U. Fayyad, eds. AAAI Press, USA, pp. 202–207.

[11] Loh, W. & Shih, Y. (1997). Split selection methods for classification trees, *Statistica Sinica* **7**, 815–840.

[12] Loh, W. & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association* **83**, 715–728.

[13] Martin, J.K. (1997). An exact probability metric for decision tree splitting and stopping, *Machine Learning* **28**, 257–291.

[14] Mattison, R. (1998). *AnswerTree algorithm User's guide* SPSS Inc., USA.

[15] Murthy, S., Kasif, S. & Salzberg, S. (1994). A system for induction of oblique decision trees, *Journal of Artificial Intelligence Research* **2**, 1–32.

[16] Quinlan, R. (1992). Learning with continuous classes, in *5th Australian Joint Conference on Artificial Intelligence*, A. Adams & L. Sterling, eds. World Scientific, pp. 343–348.

[17] Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

[18] Rivest, R.L. (1987). Learning decision lists, *Machine Learning* **2**, 229–246.

[19] Seewald, A.K. & Fürnkranz, J. (2001). An evaluation of grading classifiers, in *Advances in Intelligent Data Analysis – IDA01*, LNCS 2189 F. Hoffmann, D. Hand, N. Adams & G. Guimaraes, eds. Springer Verlag, Berlin, pp. 115–124.

[20] Torgo, L. (2000). Inductive Learning of Tree-based Regression Models. PhD thesis, University of Porto.

[21] Towell G.C., Shavlik JW. Extracting refined rules from knowledge-based neural networks. *Machine Learning* **13**, 71–101.

[22] Utgoff, P. (1988). Perceptron trees – a case study in hybrid concept representation, in *Proceedings of the Seventh National Conference on Artificial Intelligence*, AAAI Press, St Paul MN, pp. 601–606.

[23] Wolpert, D. (1992). Stacked generalization. *Neural Networks*, **5**, 241–260.

JOÃO GAMA

Tree-structured Statistical Methods

This article is about binary tree-structured methods for biostatistics. The techniques, sometimes called "recursive partitioning," can facilitate the automation of diagnoses and prognoses in clinical contexts. A brief account of the larger topic of rules for clinical prediction provides context for the more specific discussion that follows. Wasson et al. [42] give a broad view of such rules.

Classifying patients is a central element of the physician's work. Typically, the physician asks questions like these. "Is this patient with chest pain suffering a heart attack, or does he simply have a strained muscle? What is the best diagnostic test for this patient with chest pain? During the next year, is this survivor of a heart attack likely enough to die that I should do a costly test that might detect a life-threatening, correctable problem?" Answering questions helps in crafting good care, but it also is essential for matching health care resources to the patients who need them the most.

Until recent decades physicians had no choice but to answer these questions in a subjective, intuitive, idiosyncratic manner. Skill probably depended on clinical experience, but equally experienced physicians varied in their levels of skill. Physicians seldom wrote down the clinical findings that they used to estimate probabilities or their rules for combining findings. As a result, there was, and is, interphysician variability and intraphysician variability in estimating probabilities and in making prognoses. Data-based rules are now available for these purposes. They enable the physician to interpret a patient's findings in quantitative terms by reference to a large number of patients with similar findings and a known diagnosis or clinical outcome. The goal of using a clinical prediction rule is to use clinical findings to place the patient in a subgroup whose disease prevalence, outcome rate, or survival rate is known and to use that placement to infer some aspect of the patient's course.

Clinical prediction rules are, as we have indicated, empirical. Their basis is a cohort of patients with known clinical findings and a known diagnosis or outcome. This cohort, the training set (also called the learning sample), is the set of patients from whom the key clinical predictors and the rule for combining them are discovered. Applying the rule to a separate cohort, the test set, can provide the subgroup-specific disease prevalences or outcome rates that are the basis for estimating probability or prognosis in an individual patient.

There are several steps involved in developing a clinical prediction rule.

Assemble the cohort. The first step is to decide upon the criteria for including the patient in the cohort that will form the training set. The investigator must ask, "What is the problem to be solved? Estimating a probability of coronary artery disease in patients with chest pain? Estimating the one-year death rate in heart attack survivors?" The answers to these questions determine the clinical criteria for admitting a patient to the cohort. One must also pay attention to the generalizability of the findings when defining the cohort. Will the clinical prediction rule apply to all hospitals or clinics? If so, the study must enroll patients from a variety of care settings. Will the findings apply to all patients with the cohort-defining problem (e.g. chest pain) or just those of a certain age, gender, or socioeconomic standing?

The second step in assembling the cohort is to decide on the size of the training set, though circumstances frequently limit the choice. A large cohort maximizes the chance that the clinical prediction rule will be optimal for other populations of patients. One empirical rule is to include five patients in the smallest outcome category for every clinical predictor in the rule.

Decide upon the outcome measure. One must answer this question, "What is the outcome to be predicted?" Then, one must state the criteria for deciding if the outcome has occurred, being sure that it is possible to collect the outcome criteria on all patients. Ideally, the measure is an unequivocal feature of the outcome, such as the result on a reliable indicator that disease is present (the "**gold standard test**") or death from the disease. The outcome should be useful to a clinician, such as an intermediate point on the path to a decision.

Decide which predictors to obtain. The predictors should be pertinent to the clinical problem. Obtaining the information should be feasible. Precise instructions on how to obtain the information are important to the clinician who wants to classify a patient accurately. The list of predictors should include all the clinical findings that could be pertinent, so that the study does not overlook an important predictor.

Collect the data and determine the outcome on a series of patients. It is fundamental here to avoid bias in collecting the data and deciding upon outcome. If the outcome is a clinical diagnosis, then it is all too easy to define it from the predictors for a patient for whom the outcome is not obvious. This can lead to excessive optimism regarding the worth of the predictors at hand. Thus, it is important that, for purposes of the study, predictors and outcome are distinct. Moreover, a source of bias is avoided if the person who assigns the diagnosis is ignorant of any findings that comprise or inform the determination of predictors.

Identify the predictors and the rule for combining them. Most of what remains of the article is a detailed description of this step.

Determine the **misclassification** rate of the rule. The most important principle is to measure the misclassification rate in a new cohort of patient (the test set). If it is not possible to follow this principle, there are several **cross-validatory** techniques by which to estimate misclassification rates on new populations by using the training set patients. By far the best approach is to enroll a new cohort of patients, preferably by a new research group in a new clinical setting. The article by Wasson et al. [42] describes the measurement of the misclassification rates. See, especially, [2] and [14].

The tree-structured statistical techniques we have mentioned are by now widely used in biostatistical inference, e.g. [1, 3, 4, 11, 17, 18, 24, 45–51], and [52]. While there are many approaches, all have in common the successive partitioning of a "feature space" of predictors into subsets. The partitioning is done on the basis of a *learning sample*, and then, if one is fortunate, it is validated by a *test sample*. In this article, it is always the case that a nonterminal node of a tree has only two daughter nodes; thus the trees are *binary trees*. Each node of the trees corresponds uniquely to a subset of the feature space and thus to a unique set of constraints on the predictors of outcome. A decision rule or summary statistic or value of a regression, depending on the application, is the same within the region determined by the terminal node. Learning sample observations have (predictor, outcome) pairs. The hope is to partition so that regions are simple enough to be understandable in terms of the subject matter, yet homogeneous as to outcome. Prediction is made to future data for which predictors are known but outcomes are not. These techniques have been applied with success to classification, regression, survival analysis, and clustering. One popular way to form trees from data is that of *Classification and Regression Trees* (CART^{TM}). Depending on the nature of the response, the techniques may be referred to as classification trees (discrete response), regression trees (continuous response), survival trees (censored positive response), or tree-structured vector quantization (when the predictors and response are the same, univariate, or vectorial but there are constraints on the complexity of the prediction). In the literature, all of them are termed tree-based (or tree-structured) methods or recursive partitioning techniques. Those of the many ideas they have in common are well described by Breiman et al. [2], and Zhang and Singer [48], which has much historical background [8, 15, 16, 19, 32].

In what follows, we first present an early tree-based analysis that will serve to illustrate many aspects of tree-based methodology, not least the simplicity of the ultimate answers. Secondly, we lay out the key and common ground for all of the tree-based methods. Then, we fill in details to distinguish the major types of tree-based methods. In light of the recent surge of the development and use of survival trees, we devote particular attention to this area. Finally, we discuss some common tips, tricks, and traps one encounters in applying the tree-based methods.

An Example

One early notable application of binary classification trees was for the purpose of diagnosing patients who enter hospital emergency rooms with chief complaints of acute chest pain [18]. See also Goldman et al. [17]. Starting with about 100 initial variables that were thought to be predictors of a heart attack, Goldman and colleagues went through a preliminary screening of the predictors and selected 40 of them for further consideration. Their goal was to construct a classification rule that can guide physicians in emergency rooms to decide in a timely manner (i.e. before levels of fundamental enzymes are known) whether a patient has suffered or is suffering a myocardial infarction (MI, or heart attack). Although a definitive diagnosis of heart attack is typically done by

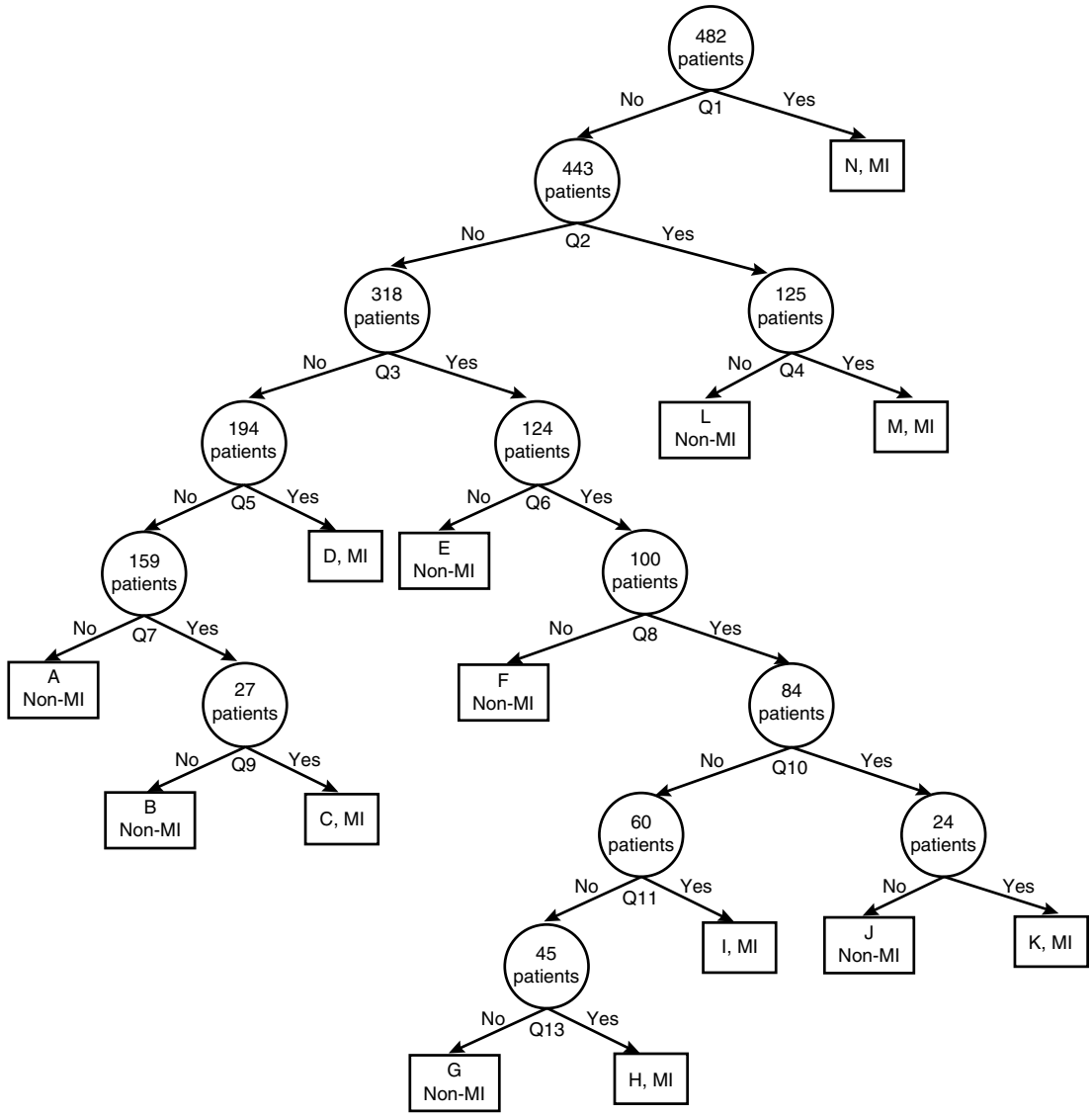

Figure 1 Classification tree for diagnosing heart attack. Table 1 provides the questions (Q1–Q13) used in this tree. This figure is based on Figure 1 of Goldman et al. [18]

testing the levels of these enzymes which tend to be released by damaged heart muscle, the importance of the computerized decision rule is that it is based on clinical measurements that are available almost immediately when a patient is admitted. By answering a maximum of 13 questions (Figure 1), any patient can be classified as having a high or low risk for heart attack.

Outline of the Tree-Based Methods

The Data and the Objective

Suppose that we have observed p covariates $\mathbf{x}$ and a response y for n individuals. For the ith individual, the measurements are

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \quad \text{and} \quad y_i, i = 1, \ldots, n.$$

The objective is to model the probability distribution of $\Pr(y|\mathbf{x})$ or some functional of this conditional distribution. Here, $\mathbf{x}$ can be an array of mixed categorical (nominal or ordinal) and continuous variables. Some components may have missing values. It is the nature of y that mandates the choice of methodology. In most applications, the outcome, y, is either a continuous (with or without censoring) or categorical variable. Recently, the tree-based methods have been developed to allow for vectorial y [16, 36, 44]. Here, our discussion focuses on **binary** and **censored** continuous y, for (i) these are the situations where the tree-based methods are applied for the most part in medicine, and (ii) **logistic regression** and **linear discriminant function analysis** (binary case), and the **Cox regression model** for **proportional hazard** modeling (in the case of **survival analysis**) are standard approaches to analyzing such outcomes; it is worthwhile to understand the strengths and limitations of both the more classical and the tree-based methods.

Basics of the Tree-Based Technique

Look again at the tree in Figure 1. This tree has eight layers of nodes. In general, the number of layers varies from case to case. The first layer is always the unique root node, namely, the circle on the top. There are 13 internal (the circle) and 14 terminal (the box) nodes that are scattered among the various layers. The root and the internal nodes are connected to two nodes in the next layer that are called left and right daughter nodes, but terminal nodes do not have "children". Moreover, the tree is not necessarily "balanced" in that not all nodes in the same layer have daughter nodes. The thrust of the tree-based technique is to answer these questions:

1. What are the contents of the nodes and how do we split a node?
2. How do we declare a node terminal?
3. What inferences do we make for the various terminal nodes?
4. What have we learned about our data and the possibly complex relationships among the predictors and outcome as a result of studying the tree?

The subsection below, "Splitting a Node", addresses the first item, and it is followed by a subsection on Terminal Nodes. There, we discuss how to determine terminal nodes. The last question is best answered on a case-by-case basis.

Splitting a Node

The root node contains the learning sample. The learning sample summarizes the information from past experience and allows us to learn the underlying data structure. In Goldman et al. [18], it contains 482 patients. The terminal nodes correspond, as was indicated, to disjoint subgroups of this learning sample. The union of two subgroups in the daughter nodes comprises the subgroup of their parent node. For example, the root node in Figure 1 has 482 patients who are divided into two subgroups: one with 443 patients and the other with 39. So, a node is merely a subgroup of the learning sample.

A critical step of the tree-based technique is to determine the split from one parent node to the two daughter nodes. Since splitting the root node is identical in terms of criterion to that for other nodes, it suffices to explain how to split the root node. Thus, we consider how the 482 patients in the study of Goldman et al. [18] might be divided into two subgroups.

First, the division of the root node is described and implemented by means of a predictor. The purpose of splitting is to generate two offspring whose union is preferred to the root node in some sense. As was mentioned earlier, there were 40 selected potential predictors of a heart attack, denoted by $\mathbf{x}$, that entered into the tree-based analysis. If x_j is an ordered covariate such as age, two subgroups result from the question of the form "Is $x_j > c$?" Here the cutoff point c is in the range of the observed values of x_j. The ith subject goes to the right or left node according to whether or not $x_{ij} > c$. Q2, 8, 9, 12, and 13 in Figure 1 and Table 1 are precisely this type of question. On the other hand, many medical studies involve nominal covariates. For example, the body sites of pain in the present example include the chest, shoulder, and neck. We can send a patient to the left or right node by asking questions such as "is the pain in the neck only?" and "is the pain in the neck and shoulder?" Given the number of covariates (here, it is 40) and the number of possible cutoff points for every covariate, there are many possibilities to split the root node into two nodes. Therefore, we must be specific in what we mean by a desirable split.

Table 1 Questions used in Figure 1

Label	Question
Q1	Does the emergency room EKG show ST-segment elevation or a Q wave that is suggestive of infarction and is not known to be old?
Q2	Did the present pain or episodes of recurrent pain begin 42 or more hours ago?
Q3	Is the pain primarily in the chest but radiating to the shoulder, neck, or arms?
Q4	Does the emergency room EKG show ST-segment elevation or a Q wave that is suggestive of ischemia or strain and not known to be old?
Q5	Is the present pain (a) similar to but somehow worse than prior pain diagnosed as angina or (b) the same as pain previously diagnosed as an MI?
Q6	Does local pressure reproduce the pain?
Q7	Has the chest pain associated with diaphoresis?
Q8	Is the patient 40 years or older?
Q9	Is the patient 70 years or older?
Q10	Was this pain diagnosed as angina (and not an MI) the last time the patient had it?
Q11	Is the pain primarily in the chest but radiating to the left shoulder?
Q12	Did the present pain or episodes of recurrent pain begin 10 or more hours ago?
Q13	Is the patient 50 years or older?

These questions are taken from Figure 1 of Goldman et al. [18]

Table 2

		Non-MI	MI	
Left node (t_L)	$x_j \le c$	n_{11}	n_{12}	$n_{1\cdot}$
Right node (t_R)	$x_j > c$	n_{21}	n_{22}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	

If we take age as a tentative splitting covariate and consider its cutoff at 40, as a result of the question "Is x_j(age) $> c$ (40)?", then we have Table 2. What would be desirable in this case? Obviously, we would want to choose a split so that the distributions of y in the daughter nodes are homogeneous. To reflect this idea in the table above, a desirable left (right) node t_L (t_R) should have the property that either n_{11} (n_{21}) is much greater than n_{12}(n_{22}) or vice versa. In other words, we force most of the MI cases to either the left or right node. In a perfect situation where $n_{11} = n_{22} = 0$, the two nodes are pure (or completely homogeneous) because each of them contains only one value of the outcome. In contrast, their parent node includes a mixture of n_{11} non-MI and n_{22} MI patients. This is what we mean by "more desirable" here. Mathematically, one frequently used measure of node homogeneity is defined through the entropy function as follows:

$$h(t_L) = \frac{n_{11}}{n_{1\cdot}} \log\left(\frac{n_{11}}{n_{1\cdot}}\right) + \frac{n_{12}}{n_{1\cdot}} \log\left(\frac{n_{12}}{n_{1\cdot}}\right). \quad (1)$$

Then, we select a split that maximizes the weighted node homogeneity:

$$\frac{n_{\cdot 1}}{n} h(t_L) + \frac{n_{\cdot 2}}{n} h(t_R). \quad (2)$$

It is also interesting to view the criterion (1) from other points of view. Thus, suppose that y in node t_L has a **binomial distribution** with a frequency of θ so that

$$\Pr(y = 1|t_L) = \theta.$$

Then, the log **likelihood** function from the $n_{1\cdot}$ observation in node t_L is

$$n_{11} \log(\theta) + n_{12} \log(1 - \theta).$$

The maximum of this log likelihood function is proportional to (1). Not surprisingly, many criteria of "more desirable" are couched as maxima of certain likelihood functions. See the section "Use of Likelihood Functions" below.

Terminal Nodes

After the node-splitting procedure described above is applied to the root node, the resulting daughter nodes can also be split in the same way, followed by the granddaughter nodes, and so on. This splitting process always terminates because the number of study subjects is finite. For example, the number of possible splits for the data of Goldman et al. [18]

cannot exceed 481. Of course, we can force the process to stop at any point. In usual practice, we end up with a large tree, which is generally too large to be useful. In order that we end up with a useful tree, a rigorous rule for pruning some overfitting nodes is required. Goldman and colleagues did not know a priori the 14 terminal nodes in Figure 1. Initially, these terminal nodes had offspring.

For the purpose of illustration, take a part of Figure 1 as displayed in Figure 2, and use it as if it is an entire initial tree. The question is: "can we prune away some of the nodes?" If we can answer this question in a general way, then we will know how to prune any tree. To this end, we introduce a measure of the quality of a tree. Recall that the objective of the tree-based method is to extract homogeneous subgroups of the study sample. Whether we have achieved it depends on whether the terminal nodes are indeed sufficiently homogeneous. Hence, the quality of a tree, denoted by T, is really the quality of its terminal nodes, and we have

$$R(T) = \sum_{t \in \tilde{T}} p(t) r(t), \tag{3}$$

where $\tilde{T}$ is the set of terminal nodes of tree T, $r(t)$ summarizes the quality of node t, and $p(t)$ is the proportion of subjects falling into node t. For binary outcomes, $r(t)$ is usually taken to be the within-node misclassification cost.

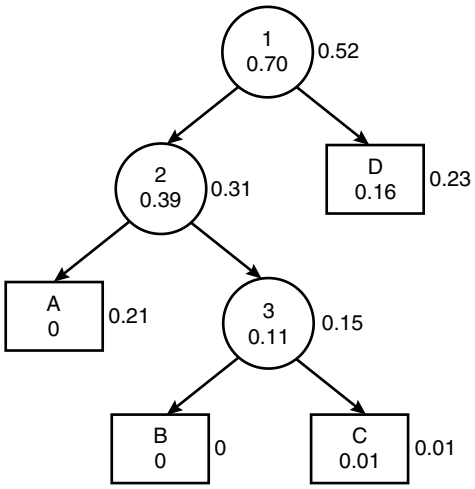

Figure 2 An illustrative tree for pruning. The root node (labeled 1) here comes from the node above the split, Q5, in Figure 1. It contains 194 subjects. Inside each node is the node label and the misclassification cost. The misclassification cost from a test sample is given outside the node

The size of a tree is another important aspect, which here is the fundamental measure of its complexity. Note that the total number of nodes in a tree, T, is $2|\tilde{T}| - 1$, where $|\tilde{T}|$ is the number of the terminal nodes of T. Hence, the complexity of T can be defined directly as $|\tilde{T}|$. Usually, a unit cost, called a complexity parameter, is assigned to each terminal node, and the sum of all costs becomes the penalty for the tree complexity. Therefore, the final quality measure of a tree is the following cost-complexity:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|, \tag{4}$$

where $\alpha(> 0)$ is the complexity parameter.

For a given complexity parameter and an initial tree such as the one in Figure 2, there is a unique smallest subtree of the initial tree that minimizes the cost-complexity measure (4). Importantly, if $\alpha_1 > \alpha_2$, then the optimally pruned subtree corresponding to α_1 turns out to be a subtree of the one corresponding to α_2. So, as we increase the complexity parameter, we have a sequence of nested optimally pruned subtrees. This sequence has to have finite length, and the last one is the root node. That the successive optimally pruned subtrees are nested can entail important savings in computation [2].

Here is how pruning works for the tree in Figure 2. Before we start, we must specify a misclassification cost that reflects the severity of the mistake that results when an MI patient is classified to non-MI or vice versa. Let $C(i|j)$ be the misclassification costs that a class j patient is classified as a class i patient. Here, there are two classes of patients: 0 for non-MI and 1 for MI patients. For medical reasons, it is natural to choose $C(0|1) > C(1|0)$ because the consequence is potentially more severe when an MI patient is wrongly diagnosed than when a non-MI patient is. As did the authors, we take $C(1|0) = 1$ and $C(0|1) = 15$, which means that a false positive diagnosis costs as much as 15 false negative ones. Table 3 gives the misclassification costs for all nodes and their designated classes. The third and fourth columns list the misclassification costs as a result of classifying the node as MI and non-MI, respectively. The minimum of these two types of cost determines the final class membership (column 5) of a node. The expected node cost, $r(t)$,

Table 3 Misclassification costs

Node label	Node size	Misclassification costs[a]		Designated class	$r(t)$[b]	Weighted cost[c]
		MI	non-MI			
1	194	185	$9 \times 15 = 135$	Non-MI	0.70	0.70
2	169	154	$5 \times 15 = 75$	Non-MI	0.44	0.39
3	37	22	$5 \times 15 = 75$	MI	0.59	0.11
A	132	132	0	Non-MI	0	0
B	20	20	0	Non-MI	0	0
C	17	2	$5 \times 15 = 75$	MI	0.12	0.01
D	46	31	$4 \times 15 = 60$	MI	0.67	0.16

[a]Number of misclassified subjects multiplied by the cost unit. [b]Misclassification cost divided by the node sample size. [c] $r(t)$ multiplied by $p(t)$, the proportion of subjects in the node.

Table 4 Nested sequence of subtrees

Subtree	Range of α	Nodes in the subtree	Cost complexity
T_0	0–0.1	1, 2, 3, A, B, C, D	$0.01 + 0.16 + 4 \times 0.005 = 0.19$
T_1	0.1–0.215	1, 2, 3, A, D	$0.11 + 0.16 + 3 \times 0.19 = 0.74$
T_2	0.215+	1	$0.70 + 0.28 = 0.98$

is the minimum of the two costs divided by the node size; for instance, $r(D) = \min(31, 60)/46 = 31/46 = 0.67$. The final within-node cost (the last column) is obtained by weighting $r(t)$ by $p(t)$. For example, within node 3, $r(3) = 0.59$ and $p(3) = 37/194 = 0.19$. Hence, the final cost equals $0.59 \times 0.19 = 0.11$.

From Table 3 we calculate the misclassification cost for the tree in Figure 2 as follows. Note that it has four terminal nodes, labeled A, B, C, and D. As is defined in (3), the tree misclassification cost is the sum of the weighted misclassification costs of its terminal nodes. Based on the last column of Table 2, the weighted costs for terminal nodes A through D are, respectively, 0, 0, 0.01, and 0.16. Thus, the tree misclassification cost is

$$0 + 0 + 0.01 + 0.16 = 0.17.$$

The complexity of this tree is 4 because it has four terminal nodes. If we choose a complexity parameter, $\alpha = 0.005$, it follows from (4) that the present tree cost complexity equals $0.17 + 4 \times 0.005 = 0.19$. Table 4 provides three ranges of the complexity parameter that correspond to three nested subtrees. The cost complexities are also given in this table when a complexity parameter is chosen in the range. The thresholds of the range are determined by these considerations. We prune off some terminal nodes only if the tree cost complexity is improved after the pruning. This decision obviously depends on the choice of the complexity parameter, α. For instance, if $\alpha = 0$, then the initial tree, T_0, has a smaller cost complexity than any of its subtrees. Therefore, we cannot prune off any terminal nodes with $\alpha = 0$. What is the smallest α such that some of the terminal nodes can be removed? It turns out to be

$$\min_{t \notin \tilde{T}_0} \frac{r(t)p(t) - R[T(t)]}{|\tilde{T}(t)| - 1},$$

where $T(t)$ is a subtree rooted at node t and the minimization is over all internal nodes of T_0 [2]. Now, T_0 has three internal nodes 1, 2, and 3, and 0.1 is the minimum of the corresponding three numbers: $(0.7 - 0.17)/(4 - 1) = 0.18$, $(0.39 - 0.01)/(3 - 1) = 0.19$, and $(0.11 - 0.1)/(2 - 1) = 0.1$. When $\alpha = 0.1$ is applied, we can prune off terminal nodes B and C without loss of cost complexity, leading to tree T_1 in Table 4. Next, we can ask the same question: what is the smallest α such that some of the terminal nodes of T_1 can be removed? This tree has two internal nodes labeled 1 and 2. It is easy to see that the smallest α equals 0.215 and it leads to the single node tree, T_2. In general, we repeat the same process until we reach the single node tree.

The next step is to select a subtree from the nested sequence. A special aspect of the study of Goldman et al. was that the tree was used to classify patients at another hospital [18]. These additional data constitute a validation data set, also called a test sample. The misclassification costs for the three nested trees were, respectively, $R^{ts}(T_0) = 0.45$, $R^{ts}(T_1) = 0.59$, and $R^{ts}(T_2) = 0.52$. Because $R^{ts}(T_0)$ is the smallest, T_0 is the best choice, implying that we cannot prune any nodes. When an independent test sample is not available, a cross-validation procedure is usually recommended. We refer to Breiman et al. [2] for details; see [28] for a different approach.

Survival Trees

In this Section, we explain how to use the ideas expressed above to analyze censored survival data. Censored survival data arise from many medical studies; see, for example, [1, 3], and [4] for some typical examples. We face the same basic issues. One is to define a splitting criterion by which to divide a node into two, and the other is to choose a "right-sized" tree for subsequent use. Many criteria have been proposed in the literature, but they differ primarily in the way of declaring what daughter nodes are desirable. Segal [37] and Intrator & Kooperberg [23] are two important and helpful reviews. See also LeBlanc & Crowley [27] and Crowley et al. [9].

Gordon and Olshen's Rule

One early proposal was made by Gordon & Olshen [22]. The idea is this: when a node is divided into two, we can compute the **Kaplan–Meier** curves (see, for example, [31]) separately for each. A desirable split can be characterized as one that results in two very different survival functions in the daughter nodes. They used the so-called L^p Wasserstein metrics, $d_p(\cdot, \cdot)$, as the measure of discrepancy between the two survival functions. Specifically, for $p = 1$, the Wasserstein distance, $d_1(S_L, S_R)$, between two Kaplan–Meier curves, S_L and S_R, is the shaded area in Figure 3.

An optimal split is chosen to maximize the distance, $d_1(S_L, S_R)$. Here, S_L and S_R are, respectively, the Kaplan–Meier curves for the left and right daughter nodes. Replacing the quantity (2) with $d_1(S_L, S_R)$ we can produce an initial tree as described above in the section on Splitting a Node.

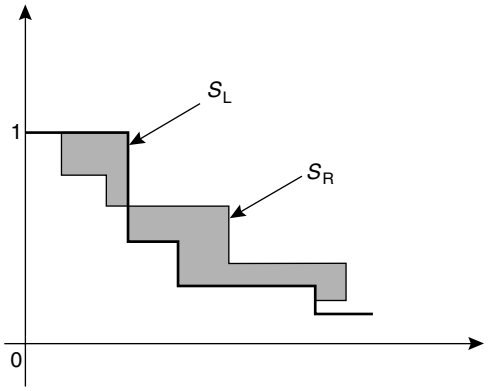

Figure 3 The L^1 Wasserstein distance between two Kaplan–Meier curves. Note that one curve (S_L) is darker than the other (S_R)

To prune an initial survival tree, T, Gordon & Olshen [22] suggested a tree cost complexity as follows. Consider a terminal node, $t \in \tilde{T}$. First, estimate the Kaplan–Meier curve S_t. Secondly, find the closest δ_t to S_t in terms of $d_1(S_t, \delta_t)$; here δ_t must be chosen from piecewise constant survival functions that have at most one point of discontinuity. That is, δ_t has at most two constant pieces. Then, define the within-node cost, $R(t)$, as $d_1(S_t, \delta_t)$. This can be viewed as the deviation of survival times about their median. Finally, applying the same formula (4), we have the tree cost complexity. Obviously, the same principle applies as we use different Wasserstein metrics. It should be noted, however, that when censoring depends on the covariates, the L^p Wasserstein metrics tend to produce splits (due to structure in the censoring) when in fact there is no dependence of survival upon covariates [9].

Use of the Logrank Test

In survival analysis, the **logrank test** is a popular approach for testing the significance of differences between the survival times of two groups. Motivated by this fact, Ciampi et al. [7] and Segal [35] suggested selecting a split that results in the largest logrank test statistic, which is defined as follows. A partition gives a sequence of 2×2 tables at times when failures occurred (Table 5).

Table 5

		Dead	Alive	
Left node (t_L)	$x_j \le c$	a_i		m_{i1}
Right node (t_R)	$x_j > c$			
		n_{i1}		n_i

The logrank test statistic is

$$LR = \frac{\sum_{i=1}^{k}(a_i - E_i)}{\left(\sum_{i=1}^{k} V_i\right)^{1/2}},$$

where k is the number of distinct failure times,

$$E_i = \frac{m_{i1}n_{i1}}{n_i},$$

and

$$V_i = \left[\frac{m_{i1}(n_i - m_{i1})n_{i1}}{n_i(n_i - 1)}\right]\left(1 - \frac{n_{i1}}{n_i}\right).$$

The logrank test (or any similar two-sample test) is a measure of between-node difference. However, with this approach a measure of cost for each node is not readily available for use in pruning. Segal [35] also recommended a practical bottom-up procedure. The basic idea is this. For each internal node (including the root node) of an initial tree, we assign it a value that equals the maximum of the logrank statistics over all splits starting from the internal node of interest. Then, we plot the values for all internal nodes in increasing order and decide a threshold from the graph. If an internal node corresponds to a smaller value than the threshold, then we prune all of its offspring.

LeBlanc & Crowley [26] introduced the notion of "goodness-of-split" complexity as a substitute for cost complexity in pruning the tree. Let $G(t)$ be the value of the logrank test at node t. Then the split-complexity measure is

$$G(T) = \sum_{t \notin \tilde{T}} G(t) - \alpha(|\tilde{T}| - 1).$$

Note that the summation above is over the set of internal (nonterminal) nodes and $|\tilde{T}| - 1$ is the number of internal nodes. The negative sign is a reflection of the fact that G is to be maximized, whereas the cost complexity R is minimized. LeBlanc & Crowley [26] recommend choosing α between 2 and 4 (when the logrank test is expressed in the χ^2 form) and using bootstrap techniques to deflate the value of G. An alternative pruning method based on permutation of ***P* values** for the logrank test is described in LeBlanc & Crowley [27].

In some medical situations such as in cancer, the goal of a tree-based analysis is to arrive at a few (perhaps three or four) groups that define the "stages" of disease. Treatment strategies or randomization algorithms within a **clinical trial** can then be designed with these prognostic groups or stages in mind. Even an optimally pruned tree may have many terminal nodes, so nodes with similar survival need be combined in a final staging system. Ciampi et al. [7] termed this process "amalgamation", and suggested combining terminal nodes based on comparisons using the logrank statistic. LeBlanc & Crowley [26] define an ordered categorical variable (based, for example, on median survival) describing the terminal nodes, and subject that single variable to a recursive partitioning scheme to amalgamate the nodes. Less formal techniques are described in LeBlanc & Crowley [27].

Use of Likelihood Functions

Several likelihood-based splitting and pruning criteria have been proposed. Davis & Anderson [12] assume that the survival function within any given node is an exponential function with a constant hazard. The splitting criterion of LeBlanc & Crowley [25] and Ciampi et al. [6] are both based on the assumption that the hazard functions in two daughter nodes are proportional, but unknown. The difference between their two approaches is whether the full or **partial likelihood** function in the Cox proportional hazards model should be used. For the same logic, these authors defined various tree cost complexities using the likelihood ratio statistic by comparing the survival times in a parent node with those in its daughter nodes. A related method due to Therneau et al. [41] makes use of what are termed martingale residuals from the Cox model as the input to a cost-complexity scheme using least squares as the cost (*see* **Residuals for Survival Analysis**).

A Straightforward Extension

Zhang [43] examined a straightforward tree-based approach to censored survival data. Note that we observe a binary death indicator and the (failure or censored) time. If we treat these two outcomes separately, then we can compute the within-node impurity, i_δ, of the death indicator and the within-node quadratic loss function, i_y, of the time as already defined by Breiman et al. [2]. Then, the within-node impurity for both the death indicator and the time is a weighted combination, $w_\delta i_\delta + w_y i_y$. Some choices of weights w_δ and w_y have been recommended by Zhang [43].

Several applications to real data have indicated that this approach and the use of the logrank test produce very similar tree structures. Perhaps surprisingly, a preliminary simulation suggests that this simple extension outperforms the more sophisticated ones in discovering the underlying structures of data. More extensive simulations are still warranted to study the performance of the various splitting criteria.

Which is Better?

This is still an open question, and perhaps it has no clear answer. Obviously, there is no shortage of splitting criteria for survival analysis. There is, however, very little evidence to suggest which approach is best under what circumstances. Some limited simulations comparing several of the methods have been reported in the literature [9, 10, 43]. Our recommendation is to construct survival trees using a number of approaches. Experts are likely to see, on their own subject matter grounds, which tree makes better sense than others.

Software

The best tested software is the commercial CART program as distributed by Salford Systems, San Diego. It has various versions for Windows, DOS, Macintosh, and Unix systems. A tree function is also available in **S-PLUS** [40]. Free software for survival trees is available, but it is less organized and tested. Four of the splitting criteria introduced above are implemented together in the C language and are available upon request to `heping.zhang@yale.edu`. Specific programs are also available by sending e-mail to various sites, such as `dstein@scott.cts.com` (Salford Systems), `mark@biostat.ucsf.edu` [35], `rdaids@sdac.harvard.edu` [12], and `mikel@fhcrc.swog.org` [26].

Other extensions of the tree-structured method have also been developed to analyze **longitudinal** data and clustered binary responses (*see* **Correlated Binary Data**). Manuscripts and programs are available upon request to `mark@biostat.ucsf.edu` for continuous longitudinal data [36] and `heping.zhang@yale.edu` for multiple correlated binary responses [44]. Also see Dr. Zhang's website (http://peace.med.yale.edu) for additional information.

Discussion

The application of tree-structured methods to many areas of research is growing (see, for example, [1, 3–5, 11, 24, 46–51], and [52]). Nevertheless, logistic regression for binary data and Cox proportional hazard models for censored survival data still dominate applications. The main advantage of tree-based methods is their ability to produce intuitive and appealing tree structures without requiring the users to specify and select conventional models. This advantage is more obvious when the classical, parametric models are not appropriate (see [23] for interesting examples). Several authors have compared the tree-structured methods with other methods [29, 38, 39]. Related programs are available upon request from `wjl@mit.edu` in addition to the sites given above. The computational complexity was an issue, but is no longer. To date, the application of the tree-based methods has been mostly for **exploratory** and secondary analyses. Recently, Zhang & Bracken [46] have demonstrated the use of tree-based methods as an intermediate step in **hypothesis testing**. Tree stability is another important concern. The tree is not a parameter, and it is not necessarily stable to small perturbations in the data. However, the resulting decision rules tend to be. Bayes theory may shed some light on this problem (*see* **Bayesian Methods**). Much work remains to strengthen the basis for statistical inference in this area. The theoretical properties of the tree-structured methods are largely unexplored, but exceptions include [13, 19–21, 26, 30, 33], and [34].

References

[1] Bacchetti, P. & Segal, M.R. (1995). Survival trees with time-dependent covariates: application to estimating changes in the incubation period of aids, *Lifetime Data Analysis* **1**, 35–47.

[2] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont. (Since 1993 this book has been published by Chapman & Hall, New York.)

[3] Carmelli, D., Zhang, H.P. & Swan, G.E. (1997). Obesity and 33 years of coronary heart disease and cancer mortality in the western collaborative group study, *Epidemiology*, **8**, 378–383.

[4] Carmelli, D., Halpern, J., Swan, G.E., Dame, A., McElroy, M., Gelb, A.B. & Rosenman, R.H. (1991). 27-year mortality in the western collaborative group study: construction of risk groups by recursive partitioning, *Journal of Clinical Epidemiology* **44**, 1341–1351.

[5] Chou, P.A., Lookabaugh, T. & Gray, R.M. (1989). Optimal pruning with applications to tree-structured source coding and modeling, *IEEE Transactions on Information Theory* **35**, 299–315.

[6] Ciampi, A., Hogg, S., McKinney, S. & Thiffault, J. (1988). A computer program for recursive partition and amalgamation for censored survival data, *Computer Methods and Programs in Biomedicine* **26**, 239–256.

[7] Ciampi, A., Thiffault, J., Nakache, J.-P. & Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates, *Computational Statistics and Data Analysis* **4**, 185–204.

[8] Cosman, P.C., Gray, R.M. & Olshen, R.A. (1994). Vector quantization: clustering and classification trees, *Proceedings of the IEEE* **82**, 919–932.

[9] Crowley, J., LeBlanc, M., Gentleman, R. & Salmon S. (1995). Exploratory methods in survival analysis, in *IMS Lecture Notes – Monograph Series 27*, H.L. Koul & J.V. Deshpande, eds. IMS, Hayward, pp. 55–77.

[10] Crowley, J., LeBlanc, M., Jacobson, J. & Salmon S. (1997). Some exploratory methods for survival data, in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D.Y. Lin & T.R. Fleming, eds. Springer-Verlag, New York, pp. 199–229.

[11] Curran, W.J., Jr., Scott, C.B., Horton, J., et al. (1993). Recursive partitioning analysis of prognostic factors in three radiation therapy oncology group malignant glioma trials, *Journal of the National Cancer Institute* **85**, 704–710.

[12] Davis, R. & Anderson, J. (1989). Exponential survival trees, *Statistics in Medicine* **8**, 947–962.

[13] Donoho, D.L. (1997). CART and best-ortho-basis: a connection, *Annals of Statistics* **25**, 1870–1911. The manuscript is available from the Web site: `http://playfair.stanford.edu/donoho/bob.ps.Z`.

[14] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[15] Friedman, J.H. (1977). A recursive partitioning decision rule for nonparametric classification, *IEEE Transactions on Computers* **C-26**, 404–407.

[16] Gersho, A. & Gray, R.M. (1992). *Vector Quantization and Signal Compression*. Kluwer, Boston.

[17] Goldman, L., Cook, F., Johnson, P., Brand, D., Rouan, G. & Lee, T. (1996). Prediction of the need for intensive care in patients who come to emergency departments with acute chest pain, *New England Journal of Medicine* **334**, 1498–1504.

[18] Goldman, L., Weinberg, M., Olshen, R.A., Cook, F., Sargent, R., Lamas, G.A., Dennis, C., Wilson, C., Deckelbaum, L., Fineberg, H. & Stiratelli, R. (1982). A computer protocol to predict myocardial infarction in emergency department patients with chest pain, *New England Journal of Medicine* **307**, 588–597.

[19] Gordon, L. & Olshen, R.A. (1978). Asymptotically efficient solutions to the classification problem, *Annals of Statistics* **6**, 515–533.

[20] Gordon, L. & Olshen, R.A. (1980). Consistent nonparametric regression from recursive partitioning schemes, *Journal of Multivariate Analysis* **10**, 611–627.

[21] Gordon, L. & Olshen, R.A. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes, *Journal of Multivariate Analysis* **15**, 147–163.

[22] Gordon, L. & Olshen, R.A. (1985). Tree-structured survival analysis, *Cancer Treatment Reports* **69**, 1065–1069.

[23] Intrator, O. & Kooperberg, C. (1995). Trees and splines in survival analysis, *Statistical Methods in Medical Research* **4**, 237–262.

[24] Kwak, L.W., Halpern, J., Olshen, R.A. & Horning, S.J. (1990). Prognostic significance of actual dose intensity in diffuse large-cell lymphoma: results of a tree-structured survival analysis, *Journal of Clinical Oncology* **8**, 963–977.

[25] LeBlanc, M. & Crowley, J. (1992). Relative risk trees for censored survival data, *Biometrics* **48**, 411–425.

[26] LeBlanc, M. & Crowley, J. (1993). Survival trees by goodness-of-split, *Journal of the American Statistical Association* **88**, 457–467.

[27] LeBlanc, M. & Crowley, J. (1995). A review of tree-based prognostic models, in *Recent Advances in Clinical Trial Design and Analysis*, P.F. Thall, ed. Kluwer, New York, pp. 113–124.

[28] Loh, W.Y. & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association* **83**, 715–725.

[29] Long, W.L., Griffith, J.L., Selker, H.P. & D'Agostino, R.B. (1993). A comparison of logistic regression to decision tree induction in a medical domain, *Computers and Biomedical Research* **26**, 74–97.

[30] Lugosi, G. & Nobel, A.B. (1996). Consistency of data-driven histogram methods for density estimation and classification, *Annals of Statistics* **24**, 687–706.

[31] Miller, R.G. (1981). *Survival Analysis*. Wiley, New York.

[32] Morgan, J.N. & Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association* **58**, 415–434.

[33] Nobel, A.B. (1996). Histogram regression estimation using data-dependent partitions, *Annals of Statistics* **24**, 1084–1105.

[34] Nobel, A.B. & Olshen, R.A. (1996). Termination and continuity of greedy growing for tree structured vector quantizers, *IEEE Transactions on Information Theory* **42**, 191–206.

[35] Segal, M.R. (1988). Regression trees for censored data, *Biometrics* **44**, 35–48.

[36] Segal, M.R. (1992). Tree-structured methods for longitudinal data, *Journal of the American Statistical Association* **87**, 407–418.

[37] Segal, M.R. (1995). Extending the elements of tree-structured regression, *Statistical Methods in Medical Research* **4**, 219–236.

[38] Segal, M.R. & Bloch, D.A. (1989). A comparison of estimated proportional hazards models and regression trees, *Statistics in Medicine* **8**, 539–550.

[39] Selker, H.P., Griffith, J.L., Patil, S., Long, W.L. & D'Agostino, R.B. (1995). A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients, *Journal of Investigative Medicine* **43**, 468–476.

[40] StatSci (1993). *S-PLUS: Guide to Statistical and Mathematical Analysis*. MathSoft, Inc., Seattle.

[41] Therneau, T.M., Grambsch, P.M. & Fleming, T.R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**, 147–160.

[42] Wasson, J.H., Sox, H.C., Neff, R.K. & Goldman, L. (1985). Clinical prediction rules: applications and methodologic standards, *New England Journal of Medicine* **313**, 793–799.

[43] Zhang, H.P. (1995). Splitting criteria in survival trees, *Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling*. Springer-Verlag, Innsbruck, pp. 305–314.

[44] Zhang, H.P. (1998). Classification trees for multiple binary responses, *Journal of the American Statistical Association* **93**, 180–193.

[45] Zhang, H.P. & Bracken, M.B. (1995). Tree-based risk factor analysis of preterm delivery and small-for-gestational-age birth, *American Journal of Epidemiology* **141**, 70–78.

[46] Zhang, H.P. & Bracken, M.B. (1996). Tree-based, two-stage risk factor analysis for spontaneous abortion, *American Journal of Epidemiology* **144**, 989–996.

[47] Zhang, H.P., Holford, T. & Bracken, M.B. (1996). A tree-based method of analysis for prospective studies, *Statistics in Medicine* **15**, 37–49.

[48] Zhang, H.P. & Singer, B. (1999). *Recursive Partitioning in the Health Sciences*. Springer-Verlag, New York.

[49] Zhang, H.P., Tsai, C.-P., Yu, C.-Y. & Bonney, G. (2001). Tree-based linkage and association analyses of asthma, *Genetic Epidemiology* **21**, S317–S322.

[50] Zhang, H.P., Yu, C.Y. & Singer, B. (2003). Cell and tumor classification using gene expression data: Construction of forests, *Proceedings of the National Academy of Sciences USA* **100**, 4168–4172.

[51] Zhang, H.P., Yu, C.-Y, Singer, B. & Xiong, M. (2001). Recursive partitioning for tumor classification with gene expression microarray data, *Proceedings of the National Academy of Sciences USA* **98**, 6730–6735.

[52] Zhu, H.T., Yu, C.Y. & Zhang, H.P. (2003). Tree-based Disease Classification Using Protein Data, *Proteomics* **3**, 1673–1677.

(*See also* **Decision Analysis in Diagnosis and Treatment Choice**; **Multivariate Analysis, Overview**)

HEPING ZHANG, JOHN CROWLEY,
HAROLD C. SOX, JR & RICHARD A. OLSHEN

Trend Test for Counts and Proportions

In the comparison of counts or proportions across various populations, it is often important to consider the intrinsic ordering of the populations with regard to some particular characteristic. For example, one may be interested in assessing whether the proportion of women reporting insomnia increases with age group, or whether the number of car accidents is increasing over calendar periods. Such a comparison can be accomplished through the use of a trend test. Trend tests arise naturally within a wide variety of biostatistical applications, such as animal **bioassays**, epidemiologic studies, and evaluations of environmental exposures, in which demonstration of a **dose–response** relationship may be important. The characteristic of the population may be measured on a continuous scale, such as an assigned treatment level, or on an ordinal scale (*see* **Ordered Categorical Data**), such as age group or initial severity of a health condition.

In considering a trend test for counts arising from independent populations, suppose that Y_i is a random variable representing the count of interest and x_i is the quantitative (continuous or ordinal) **covariate** for the ith population. In addition, let w_i be a known design variable for the ith population; this often relates to the sample or population size so that Y_i/w_i represents a "rate" of a certain event. The general data framework for a trend test is shown in Table 1.

We assume that the expected count can be related to the covariate through a continuous function f, as follows:

$$\mathrm{E}[Y_i] = w_i f(x_i).$$

Under the general null hypothesis, there is no difference in expected counts due to differences in x_i, so that the null hypothesis can be stated as:

$$\mathrm{H}_0\colon f(x_i) = f(x_j), \quad \text{for all } i, j.$$

Note that, because the sample or population sizes (or other relevant known weights) w_i may differ across populations, the expected counts themselves may not necessarily be equal even under the **null hypothesis**.

The general **alternative** to this null hypothesis is that $f(x_i) \neq f(x_j)$ for $i \neq j$. However, the trend test considers a narrower alternative, which reflects either an increasing or decreasing ordered alternative. For example,

$$\mathrm{H}_{a1}\colon f(x_i) < f(x_j), \quad \text{for } x_i < x_j$$

reflects an increasing trend alternative, while

$$\mathrm{H}_{a2}\colon f(x_i) > f(x_j), \quad \text{for } x_i < x_j$$

reflects a decreasing (or reverse) trend alternative. When either alternative is true, the trend tests tend to have more power than tests of the general alternative.

Trend tests are typically developed for the situation in which independent random samples are selected from each of the $i = 1, \ldots, k$ populations. However, they can often be applied to data collected under other types of data sampling, such as **cross-sectional studies**, where the sample sizes in each of the k groups become known only after the study is completed. More detail on data sampling plans is provided in Fleiss [14].

The trend test can be developed by considering the function $f(x)$ to be restricted to a linear function of x,

$$f(x) = \alpha + \beta x, \tag{1}$$

or a monotone (increasing or decreasing) continuous function of the above; that is

$$f(x) = g(\alpha + \beta x). \tag{2}$$

Examples of functions commonly used in this context include the normal cdf, **logistic**, arcsine, **extreme value**, and one-hit models. Several of these models, especially the normal and logistic, have been justified in toxicological applications as arising from tolerance distributions, where x represents the dosage (or log dosage) of exposure to a particular chemical [12, 23, 32] (*see* **Logistic Regression**; **Quantal Response Models**). Similarly, the one-hit model $g(x) = 1 - \exp[-(\alpha + \beta x)]$ has been used extensively in cancer

Table 1 Data framework for trend test

Population	Population covariate	Weight	Observed count	Expected count
1	x_1	w_1	y_1	$w_1 f(x_1)$
2	x_2	w_2	y_2	$w_2 f(x_2)$
⋮	⋮	⋮	⋮	⋮
i	x_i	w_i	y_i	$w_i f(x_i)$
⋮	⋮	⋮	⋮	⋮
k	x_k	w_k	y_k	$w_k f(x_k)$

risk assessment (*see* **Dose–Response Models in Risk Analysis**). The inverse of $g(x)$ is referred to as the link function, and allows $g^{-1}[f(x)]$ to be modeled as the linear function $(\alpha + \beta x)$. For example, the link functions for the normal, logistic, and extreme value models are the probit, logit, and complementary log–log links, respectively. Use of such link functions has facilitated the development of trend tests using **generalized linear models**, **quasi-likelihood** methods, and **generalized estimating equations** [32, 41].

After choosing the appropriate model (*see* **Model, Choice of**), a test for trend can be constructed as a test of $H_0 : \beta = 0$, with the alternative of an increasing trend $H_{a1} : \beta > 0$ or decreasing trend $H_{a2} : \beta < 0$. The specific form of the trend test depends on the distribution of the random variables Y_i. Three cases will be considered here:

1. $Y_1, \ldots, Y_k$ are independent **binomial** random variables (see next section);
2. $Y_1, \ldots, Y_k$ follow a **multinomial distribution** (see the section "Trend Tests for Multinomial Counts" below);
3. $Y_1, \ldots, Y_k$ are independent **Poisson** random variables (see the section "Trend Tests for Poisson Counts" below).

The trend test for proportions follows the same form as that for independent binomial random variables, and will be discussed in the next section. For all three cases, it can be shown that the **sufficient statistic** for the trend test is $\sum x_i Y_i$. In the special case in which the covariates x_i are equally spaced or represent ordinal categories, the sufficient statistic for the trend test can be simplified to $\sum_i Y_i$.

Trend Tests for Binomial Counts and Proportions

When $Y_1, \ldots, Y_k$ represent independent binomial random variables, then the design variable w_i is equal to the sample size for the ith population, $w_i = n_i$, and the function of interest is

$$f(x_i) = p_i = g(\alpha + \beta x_i),$$

such that

$$\mathrm{E}[Y_i] = n_i p_i = n_i g(\alpha + \beta x_i).$$

In this situation, the null hypothesis of interest is

$$H_0: p_1 = p_2 = \cdots = p_k \tag{3}$$

and the alternatives of increasing trend and decreasing trend can be written, respectively, as

$$H_{a1}: p_1 < p_2 < \cdots < p_k,$$
$$H_{a2}: p_1 > p_2 > \cdots > p_k.$$

To test these ordered alternatives, the **likelihood** must be specified in terms of the model chosen for p_i. For example, when g is the identity function as in (1), the likelihood becomes

$$\prod_{i=1}^{k} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$
$$= \prod_{i=1}^{k} \binom{n_i}{y_i} (\alpha + \beta x_i)^{y_i} [1 - (\alpha + \beta x_i)]^{n_i - y_i}.$$

In general, one obtains the **maximum likelihood** estimator $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\beta})$ by solving the following score equations:

$$\mathbf{u}(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{k} \begin{bmatrix} 1 \\ x_i \end{bmatrix} [y_i - n_i \hat{p}_i] = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where $\hat{p}_i = g(\hat{\alpha} + \hat{\beta} x_i)$. For the particular case in which g is the identify function, the maximum likelihood estimator for β is

$$\hat{\beta} = \frac{\sum_{i=1}^{k} x_i (y_i - n_i \tilde{p})}{\sum_{i=1}^{k} n_i (x_i - \overline{x})^2},$$

where $\tilde{p} = \sum y_i / \sum n_i$ and $\overline{x} = \sum x_i n_i / \sum n_i$. Under other models, such as the **logistic regression** model,

$$p_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}, \tag{4}$$

there may be no closed form solution for the maximum likelihood estimators, but iterative techniques such as the Newton–Raphson or Fisher scoring algorithm can be used to identify the MLEs (*see* **Optimization and Nonlinear Equations**).

When p_i can be written as $g(\alpha + \beta x_i)$, then a score test (*see* **Likelihood**) of the null hypothesis $H_0: \beta = 0$ can be constructed as

$$Z^2_{\text{linear}} = \mathbf{u}(\alpha_0, \beta_0)' \mathbf{I}^{-1}(\alpha_0, \beta_0) \mathbf{u}(\alpha_0, \beta_0),$$

where $\beta_0 = 0$, and $\mathbf{I}^{-1}(\alpha_0, \beta_0)$ is the inverse of the **information matrix** evaluated at the null hypothesis. The information matrix can be shown to be:

$$\begin{aligned}
\mathbf{I}(\alpha, \beta) &= \begin{bmatrix} \mathbf{I}_{\alpha^2} & \mathbf{I}_{\alpha\beta} \\ \mathbf{I}_{\alpha\beta} & \mathbf{I}_{\beta^2} \end{bmatrix} \\
&= -\begin{bmatrix} \dfrac{\partial^2 \log L(\alpha, \beta)}{\partial \alpha^2} & \dfrac{\partial^2 \log L(\alpha, \beta)}{\partial \alpha \partial \beta} \\ \dfrac{\partial^2 \log L(\alpha, \beta)}{\partial \beta \partial \alpha} & \dfrac{\partial^2 \log L(\alpha, \beta)}{\partial \beta^2} \end{bmatrix} \\
&= \sum_{i=1}^{k} n_i p_i (1 - p_i) \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} \\
&= \sum_{i=1}^{k} \text{var}(Y_i) \mathbf{x}_i \mathbf{x}_i',
\end{aligned}$$

where $\mathbf{x}_i' = [1 \quad x_i]$. After taking the inverse, we have

$$\begin{aligned}
\mathbf{I}^{-1}(\beta) &= (\mathbf{I}_{\beta^2} - \mathbf{I}_{\alpha\beta} \mathbf{I}_{\alpha^2}^{-1} \mathbf{I}_{\alpha\beta})^{-1} \\
&= p(1 - p) \left[\sum_{i=1}^{k} n_i x_i^2 - \frac{\left(\sum_{i=1}^{k} n_i x_i \right)^2}{\sum_{i=1}^{k} n_i} \right] \\
&= p(1 - p) \sum_{i=1}^{k} n_i (x_i - \overline{x})^2,
\end{aligned}$$

so that the score test can be written as

$$\begin{aligned}
Z^2_{\text{linear}} &= \frac{\mathbf{u}(\beta)^2}{\mathbf{I}^{-1}(\beta)} \\
&= \frac{\left[\sum_{i=1}^{k} x_i (y_i - n_i \tilde{p}) \right]^2}{\tilde{p}(1 - \tilde{p}) \sum_{i=1}^{k} n_i (x_i - \overline{x})^2}. \qquad (5)
\end{aligned}$$

There are many other algebraically equivalent forms of the score test, some of which facilitate numerical computations, including

$$\begin{aligned}
&Z^2_{\text{linear}} \\
&= \frac{\left[\sum_{i=1}^{k} n_i (x_i - \overline{x})(\hat{p}_i - \tilde{p}) \right]^2}{\tilde{p}(1 - \tilde{p}) \sum_{i=1}^{k} n_i (x_i - \overline{x})^2} \\
&= \frac{\left[\sum_{i=1}^{k} n_i (x_i - \overline{x}) \hat{p}_i \right]^2}{\tilde{p}(1 - \tilde{p}) \sum_{i=1}^{k} n_i (x_i - \overline{x})^2} \\
&= \frac{\left[\sum_{i=1}^{k} (x_i - \overline{x})(y_i - n_i \tilde{p}) \right]^2}{\tilde{p}(1 - \tilde{p}) \sum_{i=1}^{k} n_i (x_i - \overline{x})^2} \\
&= \frac{\left[\sum_{i=1}^{k} x_i - \overline{x} \sum_{i=1}^{k} y_i \right]^2}{\tilde{p}(1 - \tilde{p}) \left\{ \sum_{i=1}^{k} n_i x_i^2 - \left[\left(\sum n_i x_i \right)^2 \Big/ \sum n_i \right] \right\}}
\end{aligned}$$

or, in matrix form, as

$$Z^2_{\text{linear}} = \mathbf{x}'(\mathbf{Y} - \mathbf{E})[\mathbf{x}' \mathbf{V} \mathbf{x}]^{-1}$$

where $\mathbf{x} = [(x_1 - \overline{x}), \ldots, (x_k - \overline{x})]'$, $\mathbf{Y} = [y_1, \ldots, y_k]'$, $\mathbf{E} = [n_1 \tilde{p}, \ldots, n_k \tilde{p}]'$, and $\mathbf{V}$ is the diagonal matrix with elements $n_i \tilde{p}(1 - \tilde{p})$ on the diagonal.

For the special case in which g is the identity function, it is also possible to express the Wald test (*see* **Likelihood**) in closed form. When the asymptotic variance is evaluated under the null (i.e. $p_i = p$ for all i), this test statistic is equivalent to the score test, as follows:

$$\begin{aligned}
Z^2_{\text{Wald}, H_0} &= \frac{(\hat{\beta}_0)^2}{\widehat{\text{var}}(\hat{\beta})} \\
&= \frac{\left[\sum_{i=1}^{k} x_i (y_i - n_i \tilde{p}) \right]^2}{\tilde{p}(1 - \tilde{p}) \sum_{i=1}^{k} n_i (x_i - \overline{x})^2}.
\end{aligned}$$

The test statistic Z^2_{linear} is widely known as the Cochran–Armitage trend test, based on the work of Armitage [2] and Cochran [8]. Asymptotically, it follows a **chi-square distribution** with one **degree of freedom**. Alternately, its square root Z_{linear} follows a **normal distribution**. A test of $H_{a1}: \beta > 0$ is constructed by rejecting the null hypothesis of no trend if $Z_{\text{linear}} > z_{(1-\alpha)}$, where $z_{(1-\alpha)}$ is the α-level upper critical value of the normal distribution. Similarly, a test of H_{a2} is constructed by rejecting H_0 for $Z_{\text{linear}} < z_\alpha$. When the values of x_i represent integer **scores** or rankings, then a test of monotone trend can be analogously constructed by replacing x_i with i in (5) to form Z^2_{monotone}. Cox [10] demonstrated that these tests are uniformly **most powerful** for their respective alternatives under the logistic model given by (4). Although Agresti [1] notes that they are not uniformly most powerful (UMP) for any departure from independence, Tarone & Gart [38] showed that both the monotone and linear trend tests are asymptotically locally optimal $C(\alpha)$ tests, provided that g is a twice differentiable monotone function of x. Other extensions of the Cochran–Armitage trend test and associated **power** considerations have been described by Chapman & Nam [7], Gross [21], Tarone & Gart [38], and Wood [44].

Although the maximum likelihood estimators may not have closed form expressions for many choices of the link function g, it is also possible to construct both Wald tests and **likelihood ratio tests** of the null hypothesis of no trend using the appropriate numerical estimates; these test statistics are provided along with the score test in most statistical **software** packages. The Wald test is calculated as

$$Z^2_{\text{Wald},H_a} = \frac{(\hat{\beta})^2}{\widehat{\text{var}}(\hat{\beta})},$$

where the asymptotic variance of $\hat{\beta}$ is estimated under the trend alternative as

$$\widehat{\text{var}}(\hat{\beta}) = \frac{\sum_{i=1}^{k} n_i \hat{p}_i (1-\hat{p}_i)(x_i - \overline{x})^2}{\left[\sum_{i=1}^{k} n_i (x_i - \overline{x})^2\right]^2}$$

and $\hat{p}_i = g(\hat{\alpha} + \hat{\beta} x_i)$ is the estimate of p_i under the alternative. Similarly, a likelihood ratio test statistic for trend can be calculated as

$$\begin{aligned} D^2_{\text{linear}} &= 2\{\log[L(\hat{\alpha}, \hat{\beta})|H_a] - \log[L(\tilde{\alpha}, 0)|H_0]\} \\ &= 2\sum_{i=1}^{k}\left[y_i \log\left(\frac{\hat{p}_i}{\tilde{p}}\right) + (n_i - y_i) \right. \\ &\quad \left. \times \log\left(\frac{1-\hat{p}_i}{1-\tilde{p}}\right)\right], \end{aligned}$$

where $\tilde{\alpha}$ is the MLE for α under the null, and $\tilde{p}_i = g(\tilde{\alpha})$. Like the Cochran–Armitage trend test Z^2_{linear}, the above two test statistics are approximately χ^2_1 in large samples under the null hypothesis of no trend. When some cells of the **contingency table** are sparse or sample sizes are otherwise not adequate for asymptotic approximations, exact tests for trend can be based on the multiple **hypergeometric distribution** [1, 10, 39] (*see* **Exact Inference for Categorical Data**). To accomplish this, one calculates Z^2_{linear} for all tables with the same fixed row and column margins, and then sums the probabilities from the multiple hypergeometric distribution associated with all tables having values of Z^2_{linear} equal to or exceeding that of the observed table.

The **goodness of fit** of the linear model specified by the trend hypothesis can be assessed by comparing the chi-square statistic for testing the hypothesis of independence, that is, the general alternative to the null hypothesis given by (3), to the trend statistic. For example, the difference between the score tests of independence (i.e. the Pearson **chi-square test**) and trend (i.e. the Cochran–Armitage trend test) can be calculated as

$$\chi^2_{\text{gof}} = \chi^2_{\text{ind}} - Z^2_{\text{linear}},$$

where Pearson's chi-square test is given by

$$\chi^2_{\text{ind}} = \frac{\sum_{i=1}^{k} n_i (\hat{p}_i - \tilde{p})^2}{\tilde{p}(1-\tilde{p})}.$$

Similarly, one can construct an analogous goodness of fit test from the corresponding likelihood ratio test statistics:

$$D^2_{\text{gof}} = D^2_{\text{ind}} - D^2_{\text{linear}},$$

where

$$D^2_{\text{ind}} = \sum_{i=1}^{k} y_i \log\left(\frac{y_i}{n_i \tilde{p}}\right).$$

Under the linear trend model, the goodness-of-fit test statistics χ^2_{gof} and D^2_{gof} both follow a chi-square distribution with $k-2$ df.

It can be shown that there is a simple relationship between the Cochran–Armitage Trend test and the Pearson **correlation** coefficient. If the jth individual from sample i is defined to have the binary random variable y_{ij} and covariate x_{ij}, then

$$Z_{\text{linear}} = \frac{\sum_{i=1}^{k} x_i(y_i - n_i\tilde{p})}{\left[\tilde{p}(1-\tilde{p})\sum_{i=1}^{k} n_i(x_i-\overline{x})^2\right]^{1/2}} = N^{1/2}\text{corr}(x_{ij}, y_{ij}),$$

where $\overline{y} = \tilde{p}$ is the sample mean of the y_{ij}s, $N = \sum n_i$, and $\text{corr}(x_{ij}, y_{ij})$ is the Pearson correlation between the individual responses y_{ij} and individual scores x_{ij}. Agresti [1, p. 284] derived a similar formulation, but with $N-1$ in place of N in the above equation. Mantel [30] proposed an extension of this correlation-based statistic for stratified data.

In cases in which the observed $\hat{p}_i$s reflect slight departures from the hypothesized ordering under the trend alternative (e.g. $\hat{p}_i > \hat{p}_j$ for some $i < j$ under H_{a1}), the **isotonic regression** approach of Barlow et al. [3] can be applied. The hypothesis of no trend can be tested by

$$Z^2_{\text{iso}} = \frac{\sum_{i=1}^{k} n_i(\hat{p}^*_i - \tilde{p})^2}{\tilde{p}(1-\tilde{p})},$$

where $\hat{p}^*_i$ is the estimate of p_i under the appropriate order restriction (H_{a1} or H_{a2}). Collings et al. [9] demonstrated that Z^2_{iso} and Z^2_{linear} have the same asymptotic power for rejecting the null in favor of the trend alternative.

Much discussion has surrounded the issue of the appropriate scores x_i to use in constructing the trend test. For example, in a bioassay study with laboratory animals randomly assigned to one of four doses (say 1, 10, 100, and 1000 mg) of a chemical carcinogen, one may question whether to use the actual doses or the dose levels on the $\log_{10}$ scale (i.e. 0, 1, 2, and 3). In general, any set of scores will give a valid test under the null hypothesis and will protect the type I error rate (*see* **Level of a Test**). However, the most powerful scores to assign are those of the true model $p_i = g(\alpha + \beta x_i)$, which may be unknown in practice. Gross [21] has shown that if model (4) holds and the true scores are $x_1, \ldots, x_k$, but the incorrect scores $z_1, \ldots, z_k$ are used in the trend test, then the asymptotic relative efficiency of the Cochran–Armitage trend test is equal to the squared Pearson correlation between the scores, $[\text{corr}(z_i, x_i)]^2$.

Often, the explanatory variable has an underlying continuous scale, but data have been either collected or summarized into groups which specify a range of the continuous variable. For example, a variable such as age may be summarized as 20–29 years, 30–39 years, 40–49 years, and ≥50 years. In such situations, the midpoint of each interval is often chosen as x_i. When the highest or lowest category is specified only as above or below a certain cutoff, as in the previous example, one possible choice is to use the median level of x_i among the individuals in that group (if such data is available) (*see* **Categorizing Continuous Variables**). An alternate selection of scores for ordinal covariates can be provided by the approach of ridit analysis, as described by Bross [6], Fleiss [14], and Mantel [31]. Further discussion of the choice of scores is given by Agresti [1], Armitage [2], Cochran [8], Graubard & Korn [20], Snedecor & Cochran [35], and Yates [45] (*see* **Scores**).

Analysis of animal bioassay data was one of the original applications and motivations for development of trend tests (*see* **Tumor Incidence Experiments**). In a typical bioassay, animals are randomized to various exposure or "dose" levels of a drug, chemical, or other stimulus, and the proportion exhibiting the response of interest is observed. An example of bioassay data and associated trend tests is shown in Table 2. In this study, female mice were administered one of three doses of the chemical 1, 2-dichloroethane, and the proportion with lung tumors was observed. The Cochran–Armitage trend test yields $Z^2_{\text{linear}} = 10.64$ (df = 1, $P = 0.001$). A goodness of fit test for the linear model is calculated as $\chi^2_{\text{gof}} = 11.09 - 10.64 = 0.45$ (df = 1, $P = 0.650$), where $\chi^2_{\text{ind}} = 11.09$ is the Pearson chi-square test for independence; the nonsignificance of χ^2_{gof} suggests that the linear model is appropriate. A logistic regression model fit to this data yields likelihood ratio and Wald tests for trend of $D^2_{\text{linear}} = 11.51$ (df = 1, $P = 0.001$) and $Z^2_{\text{Wald},H_a} = 9.54$ (df = 1, $P = 0.002$), respectively. Based on these results,

Table 2 Binomial counts for lung tumors in female mice exposed to 1,2-dichloroethane

Dose (mg/kg)	Number exposed	Number with tumor	Percentage with tumor(%)
0	40	2	5
1	50	7	14
2	48	15	31

Test statistics

Test statistic	Label	Chi-square statistic	df	P value
Cochran–Armitage trend test	Z^2_{linear}	10.64	1	0.001
LR test of trend	D^2_{linear}	11.51	1	0.001
Wald test of trend	$Z^2_{\text{Wald,H}_a}$	9.54	1	0.002

we can conclude that the data provide evidence of an increasing trend in tumor rates with higher 1, 2-dichloroethane doses.

There is a wealth of literature on the application of trend tests to bioassay data. Bliss [4, 5] and Finney [12, 13] provided some of the original treatises on the subject of logit and probit models for analyzing trend in bioassays. More recent texts include those by Gart et al. [18], Govindarajulu [19], Hubert [23], and Morgan [32]. Adjustments to trend tests for bioassay data have been proposed by Gart et al. [17] to account for differential survival, and by Ibrahim & Ryan [24], Ryan [34], and Tarone [36] to incorporate information on historical controls into trend tests.

Extensions of the trend test have been suggested for stratified analysis, multiple outcomes, clustered data, and missing data. Using an example of combining information from animal bioassays conducted in several different sex-species groups, Tarone & Gart [38] described how the trend test for a general $(2 \times k)$ contingency table can be extended to incorporate **stratification** factors. Trend tests are also applied routinely to data from developmental toxicity bioassays, in which pregnant dams are exposed but primary interest lies in assessment of the effects of exposure on development and growth of their offspring. Since there are typically multiple outcomes of interest, trend tests in this setting must take into account both the clustering induced by litter effects and the correlation between multiple outcomes measured on the same embryo (*see* **Correlated Binary Data**). A test for global trend was proposed by Lefkopoulou & Ryan [29]; this test evaluates the overall effect of exposure on multiple correlated outcomes which may also be clustered. Williams & Ryan [42] adapted this test to account for missing data, and examined the efficiency of such global trend tests under various patterns of missing data. Because of the complexity of maximum likelihood analysis in such settings, use of quasi-likelihood methods and generalized estimating equations (GEEs) has become very popular in evaluating trends in proportions in the context of clustered binary data [32, 33, 43] (*see* **Quasi-likelihood**). An example of fitting three different types of models – logit, probit, and extreme value – to the probability of fetal malformation resulting from exposure to ethylene glycol is shown in Figure 1. These models were fit using both standard estimation methods and GEEs. While the parameter estimates for α and β were actually fairly similar, the standard errors for the standard models (which ignored clustering) were underestimated and led to inflated trend test statistics; for example, the trend Z statistic for the logit model was 14.6 based on the standard method and 8.4 using the GEE method.

Trend Tests for Multinomial Counts

In the case in which $(Y_1, \ldots, Y_k)$ represents a vector of counts following a multinomial distribution with $\sum_{i=1}^{k} Y_i = N$ and corresponding probabilities $p_1, \ldots, p_k$, we again assume that the probabilities p_i are linked to ordinal or continuous covariates x_i which are monotonically increasing or decreasing. Lee [26] described a method for testing the null hypothesis that the probability of falling into any

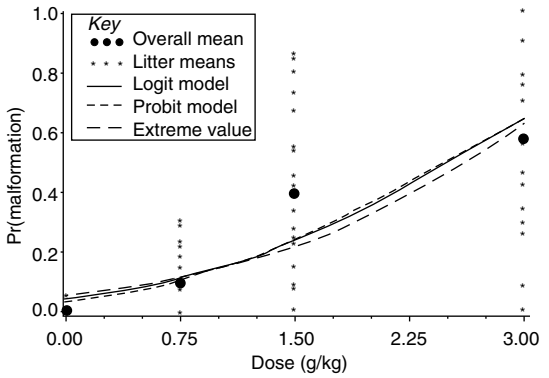

Figure 1 An illustration of logit, probit, and extreme value models fitted to the EG data on malformation

level is equal, that is,

$$H_0: p_1 = p_2 = \cdots = p_k = 1/k$$

vs. the increasing trend alternative

$$H_a: p_1 \leq p_2 \leq \cdots \leq p_k.$$

In fact, a slightly narrower alternative than H_a is considered by Lee as H_a^*: $p_{i+1} \geq \lambda_i p_i$, where $\lambda_i \geq 1$, $i = 1, \ldots, k-1$, with strict inequality for at least one λ_i. The trend statistic was then developed by choosing the test which maximized the minimum power over the parameter space $\Omega(\lambda)$, i.e. a **minimax** test.

This approach was later generalized, again by Lee [27], to one that allows incorporation of weights for each p_i. The test for trend is specified in terms of the related probabilities π_i which are subject to weights w_i, such that

$$p_i = f(w_i, \pi_i),$$

where $\sum w_i = 1$. Consider, for example, a cross sectional sample of N subjects summarized in a $(2 \times k)$ contingency table, with the first row representing the response of interest and the columns indicating the levels of the covariate $(x_1, \ldots, x_k)$. Since each of the N subjects can fall into only one of the $2k$ cells, the vector $(Y_{11}, Y_{21}, \ldots, Y_{1k}, Y_{2k})$ follows a multinomial distribution. In addition, conditional on $\sum_{i=1}^{k} Y_{1i}$, the vector of counts $(Y_{11}, \ldots, Y_{1k})$ also follows a multinomial distribution with probabilities $(p_1, \ldots, p_k)$. The weights w_i in this case are the proportion of the total sample N falling into the ith covariate level, that is $w_i = (Y_{1i} + Y_{2i})/N$, and the functional relationship is $p_i = w_i \pi_i$, where $\pi_i = \delta_i / \sum_{i=1}^{k} w_i \delta_i$ and δ_i is the conditional probability of response given the subject has covariate level x_i. The values of π_i can be related to the covariates via the function g, so that $\pi_i = g(x_i)/N$. The null hypothesis and alternative hypotheses are then stated as

$$H_0: \pi_1 = \pi_2 = \cdots = \pi_k$$

vs.

$$H_{a1}: \pi_1 < \pi_2 < \cdots < \pi_k$$

or

$$H_{a2}: \pi_1 > \pi_2 > \cdots > \pi_k,$$

with $\sum_{i=1}^{k} w_i \pi_i = 1$. Lee showed that the linear trend test for multinomial counts is:

$$Z_{\text{linear}} = \frac{\sum_{i=1}^{k} x_i (y_i - N w_i)}{\left\{ N \left[\sum_{i=1}^{k} w_i (x_i - \overline{x})^2 \right] \right\}^{1/2}},$$

where $\overline{x} = \sum w_i x_i / \sum w_i$. The monotone trend test Z_{monotone} can be constructed by replacing x_i with i in Z_{linear}. These test statistics follow an asymptotic standard normal distribution, and are both minimax tests. Like the tests derived in the binomial count setting, they are asymptotically efficient. Lee recommends use of the monotone trend test statistic, since it is more generally applicable regardless of the values of λ_i or the functional form of g. Tiwari & Sen [40] derived a rank statistic which incorporates historical control data for testing for a trend in multinomial proportions.

An example is shown in Table 3 for a sample of 1178 HIV-infected patients enrolled in a clinical trial comparing treatments for prevention of *Mycobacterium avium* complex (MAC) disease, a serious opportunistic infection contributing towards mortality

Table 3 Multinomial counts for MAC disease by pre-entry CD4 count

	CD4 lymphocyte level (cells/mm^3)				
	<25	25–50	50–75	>75	Total
MAC disease	78	25	11	7	121
No MAC disease	473	251	150	183	1057
Total	551	276	161	190	1178

Table 4 Estimated multinomial probabilities by pre-entry CD4 count

CD4 level	Number with MAC (Y_{1i})	Estimated multinomial probability ($\hat{p}_i$)	Weight (w_i)	Estimated conditional probability (δ_i)	Estimated trend value (π_i)
<25	78	0.645	0.467	0.142	1.378
25–50	25	0.207	0.234	0.091	0.882
50–75	11	0.091	0.137	0.068	0.665
>75	7	0.058	0.161	0.037	0.359
Total	121	1.000	1.000	–	–

in this population. The 1178 patients are cross-classified as to their MAC disease status and pre-entry CD4 lymphocyte count (cells/mm^3). Conditional on the 121 total observed cases of MAC disease, the vector of counts of patients with MAC disease in each of the four CD4 categories follows a multinomial distribution. While the estimated multinomial probabilities shown in Table 4 appear to follow a decreasing trend, such a comparison neglects the fact that the overall proportions of patients in the four CD4 levels are not equal. Instead, the appropriate trend test is based on the conditional probabilities of response δ_i given CD4 level x_i. This can be formulated as a test of trend in $\pi_i = p_i/w_i$; the values of π_i are proportional to δ_i but satisfy the constraint that $\sum p_i = \sum w_i \pi_i = 1$. For the data shown in Table 3, the linear trend test based on setting x_i equal to the median values of CD4 for each level (10, 36, 63, and 93, respectively) yields $Z_{\text{linear}} = -4.210$ ($P < 0.0001$). Since the values of x_i are almost evenly spaced, this result is very similar to that of the monotone trend test $Z_{\text{monotone}} = -4.227$. Either test indicates a significant departure from the null hypothesis of equal values of π_i in favor of the reverse trend alternative H_{a2}.

Trend Tests for Poisson Counts

For the case in which $Y_1, \ldots, Y_k$ are independent Poisson random variables, we still assume that $E[Y_i] = w_i f(x_i)$ where x_i is an ordered covariate, but now $f(x_i) = \lambda_i$ is the mean of the Poisson variable Y_i. Lee [28] has noted that we may consider the weights w_i arising from one of two scenarios: either (i) Y_i may be the number of rare events during an interval of length w_i, where λ_i is the event rate per unit time; or (ii) each Y_i may be a sum of w_i independent Poisson random variables, that is $Y_i = \sum_{j=1}^{w_i} Y_{ij}$, where $Y_{i1}, \ldots, Y_{iw_i}$ are identically distributed with mean λ_i. The former situation may also be relevant when w_i is the cumulative exposure in units such as person-years. Examples of Poisson data in biostatistical applications include incidence of new AIDS or cancer cases per calendar year, number of injuries or accidents over a set time period, number of bacteria per unit volume of suspension, number of revertants in microbial mutagenesis assays, or number of tumors observed in w_i animals exposed to dose x_i in an animal bioassay.

Our interest is again in testing for an increasing or decreasing trend in the means $\lambda_i = E[Y_i]/w_i$ with increasing levels of x_i. The relationship between λ_i and x_i is specified as

$$\lambda_i = g(\alpha + \beta x_i), \tag{6}$$

where g is a twice-differentiable monotone function. Commonly used functions in the context of Poisson data include g equal to the identity function or $g(x) = \exp(x)$, with the latter implying the loglinear regression model (*see* **Poisson Regression**)

$$\log(\lambda_i) = \alpha + \beta x_i. \tag{7}$$

The special case in which g is the identity and the intercept is zero, resulting in the regression through the origin $\lambda_i = x_i \beta$, has been considered by Frome et al. [15], Gart [16], and Jorgensen [25]. The test of trend developed by Gart in this context is exact, while the other two rely on asymptotic approximations.

Armitage [2] derived a chi-square test for trend in frequencies based on the asymptotic distribution of β in the model

$$\lambda_i = \alpha + \beta x_i,$$

and Cochran demonstrated that this test was valid for Poisson data under the assumption of equal weights,

$w_i = w$ for all i. For the more general model given by (6), the likelihood can be written as

$$\mathcal{L} = C(y_1, \ldots, y_k) \prod_{i=1}^{k} \exp\{-w_i g(\alpha + \beta x_i)\} \times \{g(\alpha + \beta x_i)\}^{y_i},$$

where C is a constant not involving α or β. The maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ are found by solving the score equations $\partial \log \mathcal{L}/\partial\alpha = 0$ and $\partial \log \mathcal{L}/\partial\beta = 0$ simultaneously, but iterative numerical methods such as the Newton–Raphson approach or Fisher scoring algorithm are necessary. Frome et al. [15] described iterative approaches for obtaining estimates by the methods of maximum likelihood, weighted least squares, and minimum chi-squared criteria leading to "best asymptotically normal" **(BAN) estimates** of parameters, and noted that these estimates are computationally equivalent under certain conditions.

Since score tests only require MLEs under the null, this approach proves more practical for deriving tests of trend which have a closed form representation. Tarone [37] showed that a score test of the null hypothesis $H_0: \beta = 0$ in (6) could be derived as

$$\chi^2_{\text{Poisson}} = \frac{\left[\sum_{i=1}^{k} x_i(y_i - w_i\overline{y})\right]^2}{\overline{y}\sum_{i=1}^{k} w_i(x_i - \overline{x})^2}, \qquad (8)$$

where $\overline{y} = \sum_{i=1}^{k} Y_i / \sum_{i=1}^{k} w_i$. Asymptotically, χ^2_{Poisson} is distributed as chi-square with one degree of freedom under H_0. While this statistic is identical to that proposed by Armitage [2], Tarone [37] provided a more rigorous justification for its use by demonstrating that it was asymptotically locally optimal against any smooth monotone alternative, i.e. regardless of the choice of g, provided g was a smooth monotone function on $[0, x_k]$. In addition, Tarone noted that when $g(x) = \exp(x)$ as in (7), the chi-square test proposed in (8) is an asymptotic approximation to the UMP test of $H_0: \beta = 0$. Because the Poisson distribution is the limiting distribution of the binomial for small p_i and large n_i, the efficiency properties described for tests of trend for binomial proportions extend to the Poisson setting.

A number of computational difficulties can arise when applying the BAN method or other estimation approaches which require inversion of the information matrix, particularly if the values of $E[Y_i]$ are small [28]. When sample sizes are moderate to large, however, maximum likelihood estimates of α and β can be obtained from statistical **software** packages such as GLIM and Stata. These estimates and their associated asymptotic standard errors can be used as the basis for constructing a Wald or likelihood ratio test of trend, using methods analogous to those described for binomial data in an earlier section.

Lee [28] suggested that an alternative approach is first to condition on the sum of the Poisson random variables, $\sum_{i=1}^{k} Y_i$, and then use the methods derived for multinomial data. Since the vector of Poisson random variables $(Y_1, \ldots, Y_k)$ follows a multinomial distribution conditional on $\sum Y_i$, the tests Z_{linear} and Z_{monotone} of the previous section can be considered as conditional minimax tests when applied to tests of trend for Poisson means.

Jorgensen [25] extended the trend tests for Poisson regression to multiple variables, and both linear and nonlinear extensions of the zero-intercept model are addressed by Frome et al. [15]. Both Gart [16] and Frome et al. [15] have discussed methods for evaluating the goodness of fit of the Poisson regression model. Tarone [37] used an approach similar to that employed for binomial data [36] to incorporate historical control data into tests for trend in Poisson means; this technique is appropriate for dose–response experiments which include a control group. El-Sayyad [11] described a Bayesian method and a related Bayesian approximation to test for trends in Poisson-distributed data. Hakulinen & Dyba [22] have described use of trend models for Poisson data to predict future disease incidence, and developed associated prediction intervals for new cases of melanoma and lung, stomach, or colon cancer. In cases in which the variance of Poisson counts appears to be inflated relative to the mean, quasi-likelihood methods can be used to account for such overdispersion ([1, pp. 456–457], and *see* **Overdispersion**; **Quasi-likelihood**).

An example of Poisson data for which a trend test is of interest is presented in Table 5. This table shows the number of cases of melanoma reported between 1969 and 1971 for six age groups, along with the person-years of employment in each age group. Also

Table 5 Poisson counts of melanoma cases by age group

Age group midpoint (x_i)	Number of observed melanoma cases (Y_i)	Person-years of exposure (w_i)	Observed rate per 100 000 person-years ($y_i/w_i \times 10^5$)	Predicted rate per 100 000 person-years ($\hat{\lambda}_i \times 10^5$)
30	61	2 880 262	2.12	3.90
40	76	564 535	13.46	6.64
50	98	592 983	16.53	11.29
60	104	450 740	23.07	19.22
70	63	270 908	23.26	32.70
80	80	161 850	49.43	55.66

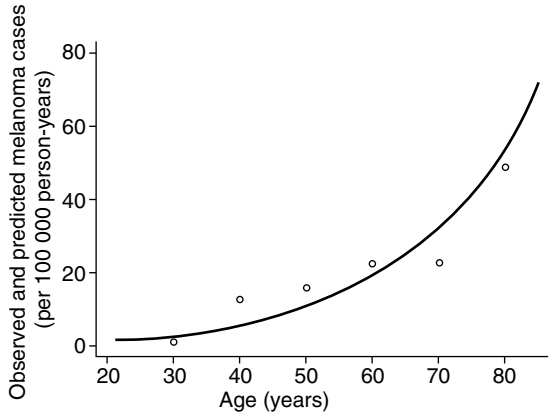

Figure 2 An illustration of Poisson regression to melanoma data

shown is both the observed rate of melanoma per 100 000 person-years, and the predicted rate based on fitting the loglinear model in (7) (see Figure 2). The MLE for β is 0.0532 with associated asymptotic standard error of 0.0025, yielding a Wald test of trend of $Z_{\text{Wald},H_a} = 21.2$. The score test of trend is calculated as 23.5, leading to a similar conclusion that the data indicate a significant increasing trend in rates of melanoma with increasing age level.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Armitage, P. (1955). Tests for linear trends in proportions and frequencies *Biometrics* **11**, 375–386.

[3] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, New York.

[4] Bliss, C.I. (1935). The calculation of the dosage–mortality curve, *Annals of Applied Biology* **22**, 134–167.

[5] Bliss, C.I. (1952). *The Statistics of Bioassay*. Academic Press, New York.

[6] Bross, I.D.J. (1958). How to use ridit analysis, *Biometrics* **14**, 18–38.

[7] Chapman, D.G. & Nam, J. (1968). Asymptotic power of chi-square tests for linear trends in proportions, *Biometrics* **24**, 315–327.

[8] Cochran, W.G. (1954). Some methods of strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.

[9] Collings, B.J., Margolin, B.M. & Oehlert, G.W. (1981). Analyses for binomial data, with application to the fluctuation test for mutagenicity, *Biometrics* **37**, 775–794.

[10] Cox, D.R. (1958). The regression analysis of binary sequences, *Journal of the Royal Statistical Society, Series B* **20**, 215–242.

[11] El-Sayyad, G.M. (1973). Bayesian and classical analysis of Poisson regression, *Journal of the Royal Statistical Society, Series B* **35**, 445–451.

[12] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, Cambridge.

[13] Finney, D.J. (1978). *Statistical Method in Biological Assays*, 3rd Ed. Griffin, London.

[14] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[15] Frome, E.L., Kutner, M.H. & Beauchamp, J.J. (1973). Regression analysis of Poisson distributed data, *Journal of the American Statistical Association* **68**, 935–940.

[16] Gart, J.J. (1964). The analysis of Poisson regression with an application in virology, *Biometrika* **51**, 517–521.

[17] Gart, J.J., Chu, K. & Tarone, R.E. (1979). Statistical issues in interpretation of chronic bioassay tests for carcinogenicity, *Journal of the National Cancer Institute* **62**, 957–974.

[18] Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research*, Volume III: *Design and Analysis of Long-term Animal Experiments*. Oxford University Press, Oxford.

[19] Govindarajulu, Z. (1988). *Statistical Techniques in Bioassay*. Karger, Basel.

[20] Graubard, B.I. & Korn, E. (1987). Choice of column scores for testing independence in ordered $2 \times k$ contingency tables, *Biometrics* **43**, 471–476.

[21] Gross, S.T. (1981). On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications, *Journal of the American Statistical Association* **76**, 935–941.

[22] Hakulinen, T. & Dyba, T. (1994). Precision of incidence predictions based on Poisson distributed observations, *Statistics in Medicine* **13**, 1513–1523.

[23] Hubert, J.J. (1992). *Bioassay*, 3rd Ed. Kendall–Hunt, Dubuque.

[24] Ibrahim, J.G. & Ryan, L.M. (1996). Use of historical controls in time-adjusted trend tests for carcinogenicity, *Biometrics* **52**, 1478–1485.

[25] Jorgensen, D.W. (1961). Multiple regression analysis of a Poisson process, *Journal of the American Statistical Association* **56**, 235–245.

[26] Lee, Y.J. (1977). Maximin tests of randomness against ordered alternatives: the multinomial distribution case, *Journal of the American Statistical Association* **72**, 673–675.

[27] Lee, Y.J. (1980). Test of trend in count data: multinomial distribution case, *Journal of the American Statistical Association* **75**, 1010–1014.

[28] Lee, Y.J. (1988). Tests for trend in count data, in *Encyclopedia of Statistical Sciences*, N.L. Johnson, & S. Kotz, eds. New York: Wiley, pp. 328–334.

[29] Lefkopoulou, M. & Ryan, L. (1993). Global tests for multiple binary outcomes, *Biometrics* **49**, 975–988.

[30] Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, *Journal of the American Statistical Association* **58**, 690–700.

[31] Mantel, N. (1979). Ridit analysis and related ranking procedures – use at your own risk, *American Journal of Epidemiology* **109**, 25–29.

[32] Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. Chapman & Hall, London.

[33] Ryan, L. (1992). Quantitative risk assessment for developmental toxicity, *Biometrics* **48**, 163–174.

[34] Ryan, L. (1993). Using historical controls in the analysis of developmental toxicity data, *Biometrics* **49**, 1126–1135.

[35] Snedecor, G.W. & Cochran, W.G. (1971). *Statistical Methods*. Iowa State University Press, Ames.

[36] Tarone, R.E. (1982). The use of historical control information in testing for a trend in proportions, *Biometrics* **38**, 215–220.

[37] Tarone, R.E. (1982). The use of historical control information in testing for a trend in Poisson means, *Biometrics* **38**, 457–462.

[38] Tarone, R.E. & Gart, J.J. (1980). On the robustness of combined tests for trends in proportions, *Journal of the American Statistical Association* **75**, 110–116.

[39] Thomas, D.G., Breslow, N. & Gart, J.J. (1977). Trend and homogeneity analysis of proportions and life table data, *Computational and Biomedical Research* **10**, 373–381.

[40] Tiwari, R.C. & Sen, P.K. (1991). Incorporating historical controls in testing for a trend in multinomial proportions, *Journal of Statistical Planning and Inference* **27**, 143–156.

[41] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss–Newton method, *Biometrika* **61**, 439–447.

[42] Williams, P. & Ryan, L. (1996). Design of multiple binary outcome studies with intentionally missing data, *Biometrics* **52**, 1498–1514.

[43] Williams, P.L. & Ryan, L.M. (1996). Dose–response models for developmental toxicology, in *Handbook of Developmental Toxicology*. R.D. Hood, ed. CRC Press, Boca Raton, Florida, pp. 635–666.

[44] Wood, C.L. (1978). Comparison of linear trends in binomial proportions, *Biometrics* **34**, 496–504.

[45] Yates, F. (1948). The analysis of contingency tables with grouping based on quantitative characters, *Biometrika* **35**, 176–181.

(*See also* **Isotonic Inference**)

PAIGE L. WILLIAMS

Trigonometric Regression

This is essentially a method of fitting a periodic regression function to data of the form $\{y_i, t_i; i = 1, 2, \ldots, n\}$, where y is a response variable and t usually denotes time. Common applications arise in modeling biological cycles that are tied to environmental cycles such as the 24 h solar day, the approximately 25 h lunar day, or the annual cycle of 12 months (seasonal variation; *see* **Circadian Variation**). In such cases the fundamental period is determined a priori. Other biological cycles, such as the menstrual period in women, have no clear link to the environment and may vary in length between individuals; so it may be of interest to estimate the length of the cycle in addition to its amplitude and phase. Furthermore, many diurnal cycles may be slightly changed in length by changes in the environment, such as exposure to continuous light (or dark). Applications where t is spatial rather than temporal arise in modeling variation with orientation on a circle (*see* **Circular Data Models**). Bliss [2, Chapter 17], gives examples and references.

Consider a model equation

$$y_i = g(t_i) + e_i, \tag{1}$$

where g is periodic with period τ, i.e. $g(t + \tau) = g(t)$ for all t, and e_i are error random variables with zero mean. In the simplest nontrivial case, $g(t)$ is a single cosine (or sine) wave with amplitude ρ, angular frequency, ω, and phase angle ϕ. This may be written in the equivalent forms

$$g(t) = \rho \cos(\omega t - \phi) \tag{2}$$

$$= \rho \sin(\omega t - \phi + \frac{\pi}{2}) \tag{3}$$

$$= \alpha \cos \omega t + \beta \sin \omega t, \tag{4}$$

where $\alpha = \rho \cos \phi$ and $\beta = \rho \sin \phi$. Expressing ρ and ϕ in terms of α and β gives

$$\rho = (\alpha^2 + \beta^2)^{1/2} \quad \text{and} \quad \phi = \tan^{-1}\left(\frac{\beta}{\alpha}\right).$$

The function $g(t)$ given by 2, 3, or 4 is periodic with period $\tau = 2\pi/\omega$ time units; ω is the angular frequency in radians per unit time, and $\omega/2\pi = 1/\tau$ is the frequency in cycles per unit time. The value of $g(t)$ cycles between ρ and $-\rho$ and reaches its first peak after $t = 0$ at time $t_\phi = \phi/\omega$. The parameters α and β are useful for fitting models; they may then be transformed to ρ and t_ϕ for interpreting results.

The model equation (1) for a constant term plus a single cosine wave becomes

$$y_i = \alpha_0 + \alpha \cos \omega t_i + \beta \sin \omega t_i + e_i. \tag{5}$$

When the period τ (and hence the frequency ω) is known, this is equivalent to an ordinary linear regression model with a constant term and explanatory variables $x_{1i} = \cos \omega t_i$ and $x_{2i} = \sin \omega t_i$. Note that, unless the phase is known, it is not sensible to include one of these terms without the other. The constant term α_0 can be regarded formally as a trigonometric function 4 with $\omega = 0$, $\alpha = \alpha_0$, and $\beta = 0$.

A cosine wave is limited in shape, but in principle a periodic function of any shape can be expressed as a linear combination of sine and cosine terms with frequencies $\omega, 2\omega, 3\omega, \ldots$, see, for example, [8, p. 11]. Here ω is called the fundamental frequency and $2\omega, 3\omega, \ldots$ are higher harmonics. Thus consider

$$g(t) = \alpha_0 + \sum_{j=1}^{m} (\alpha_j \cos j\omega t + \beta_j \sin j\omega t). \tag{6}$$

This is still periodic with period $\tau = 2\pi/\omega$, because a function that repeats itself every $2\pi/j\omega$ time units also does so every $2\pi/\omega$ time units. Again, when ω is known, and with the usual assumptions about the errors e_i, (1) is an ordinary linear regression model once we have fixed which terms to include.

In applications where the cycle varies about a nonstationary trend, the model equation may need to be generalized further – for example, by adding to 6 a parametric trend function, such as a low-order polynomial in t or a low-frequency trigonometric function to be fitted along with the cyclic function, or alternatively by removing the trend nonparametrically before fitting (6).

When the period τ (or frequency ω) is regarded as an unknown parameter, then (1) with $g(t)$ given by 6 is no longer a linear regression model. However, a natural approach is to fit 6 by linear regression for each of a range of values of ω and to choose ω to optimize the fit. If the e_i are assumed to be independent and normal, then **maximum likelihood** estimates for ω and the other parameters may be obtained by this method.

There are obvious extensions to **generalized linear models**, or types of error models other than

independent normal. For example, y_i might have a **Poisson distribution** with mean $\exp\{g(t_i)\}$, independently for each i, where $g(t)$ is given by 6. When ω is known, this is a standard **loglinear model**. For a single trigonometric term, the frequency and phase have the same meaning as before, but ρ is now the amplitude on a log scale, so that $\mathrm{E}[y(t)]$ cycles between $\mu_0 e^{\rho}$ and $\mu_0 e^{-\rho}$.

Another type of extension of the model is to nonindependent errors. Particularly when $y_1, y_2, \ldots, y_n$ is an observed **time series**, it may sometimes be more reasonable to assume that the e_i are correlated, perhaps described by a low-order **moving average** or autoregressive process (*see* **ARMA and ARIMA Models**).

When observed at a sufficient number of reasonably spaced times, different trigonometric functions are nearly orthogonal. These **orthogonality** properties are exact for the equally spaced times and frequencies discussed in the next section, which makes trigonometric functions particularly convenient for modeling periodic phenomena.

Equally Spaced Times and Fourier Frequencies

Often y is observed at n equally spaced times. We now take unit time to be the sampling interval and denote the data by $\{y_t, t; t = 1, 2, \ldots, n\}$ to avoid possible confusion with the complex number i which is sometimes used in the derivation of the orthogonality results used here and in related representations.

The shortest period that can be observed is $\tau = 2$, corresponding to $\omega = 2\pi/2 = \pi$, so without loss of generality we consider frequencies in the range $0 \le \omega \le \pi$. A trigonometric function 4 with $\omega = \pi$ may be written as

$$\alpha \cos \pi t + \beta \sin \pi t = \alpha(-1)^t, \tag{7}$$

which is represented by just one parameter, α. The absolute value of α is the amplitude and the sign of α determines the phase ($t_\phi = 0$ or 1). Thus, to fit a function with period 24 h one would need to observe it at least every 12 h; and observing it every 12 h would just allow one to fit (7).

For reasonably large n, the frequency range $0 \le \omega \le \pi$ is well spanned by the set of *Fourier frequencies*

$$\omega_j = \frac{2\pi j}{n}, \quad j = 0, 1, 2, \ldots, m, \tag{8}$$

where $m = n/2$ when n is even and $m = (n-1)/2$ when n is odd. Furthermore, the sampling interval is usually chosen so that the frequencies of interest are a subset of 8. For these frequencies it can be shown (see, for example, [1, p. 95]) that the functions

$$\cos \omega_j t, \sin \omega_j t, \cos \omega_k t, \sin \omega_k t$$

for $j \ne k$, are mutually orthogonal. That is, the sum over $t = 1, 2, \ldots, n$ of products of any pair of these is zero.

We present results for the case n even, so $m = n/2$. The formulas for n odd are the same except that there is no one-parameter term corresponding to $\omega = \pi$. The saturated model equation, including sine and cosine terms at all frequencies 8, is now

$$y_t = \alpha_0 + \sum_{j=1}^{m-1} (\alpha_j \cos \omega_j t + \beta_j \sin \omega_j t) + \alpha_m \cos \pi t + e_t. \tag{9}$$

This has n coefficients. If these are all estimated by **least squares**, then the data will be fitted exactly. In practice, a model with a subset of these terms, corresponding to the fundamental frequency and some higher harmonics for each underlying periodicity, will be of interest. Because of the orthogonality, such a model may be fitted simply by selecting the relevant terms from the saturated model, analogously to using orthogonal polynomials (*see* **Orthogonality**) in **polynomial regression**.

The least squares estimates for $j = 1, 2, \ldots, m-1$, are

$$\hat{\alpha}_j = \frac{2}{n} \sum_{t=1}^{n} y_t \cos \omega_j t, \; \hat{\beta}_j = \frac{2}{n} \sum_{t=1}^{n} y_t \sin \omega_j t, \tag{10}$$

and for $j = 0, m$, are

$$\hat{\alpha}_0 = \frac{1}{n} \sum_{t=1}^{n} y_t, \qquad \hat{\alpha}_m = \frac{1}{n} \sum_{t=1}^{n} (-1)^t y_t. \tag{11}$$

These estimates are unaffected by omitting other terms from the model. Under the usual sampling model for linear regression, where the errors e_t are uncorrelated with mean 0 and variance σ^2, all of the estimates (10) and (11) are mutually uncorrelated.

Table 1 Harmonic analysis of variance

Source frequency (cycles per n time units)	Degrees of freedom (df)	Sum of squares (ss)
0	1	$n\bar{y}^2$
1	2	$(n/2)(\hat{\alpha}_1^2 + \hat{\beta}_1^2)$
2	2	$(n/2)(\hat{\alpha}_2^2 + \hat{\beta}_2^2)$
...	...	...
$m-1$	2	$(n/2)(\hat{\alpha}_{m-1}^2 + \hat{\beta}_{m-1}^2)$
m	1	$n\hat{\alpha}_m^2$
Total	n	$\sum_{t=1}^{n} y_t^2$

Each of the estimates (10) has variance $2\sigma^2/n$, and each of 11 has variance σ^2/n.

There is a corresponding harmonic **analysis of variance**, described in Table 1. Sources of variation are the different frequencies j in cycles per n sampling intervals, or corresponding periods in numbers of sampling intervals. The frequency $j = 0$ corresponds to the constant term; this is usually omitted from the analysis of variance table, but is included here for completeness. The sums of squares are proportional to the squared amplitude of each fitted cosine wave.

The n coefficients $\frac{1}{2}\hat{\alpha}_j$, $\frac{1}{2}\hat{\beta}_j$, $\hat{\alpha}_0$, and $\hat{\alpha}_m$ given by 10 and 11 constitute the *discrete Fourier transform* of the data, and the sums of squares in Table 1 constitute the *periodogram*. These are basic statistics used in the frequency domain analysis of **time series**.

Example

As an illustrative example consider the data in Table 2. The height of the tide at a point on a tidal river in West Scotland has been measured every 62 min for $n = 24$ times.

The interval between successive high tides at the relevant time of year is known to be 12 h 24 min, so there should be exactly two tidal cycles in these data. Here it is convenient to take the time unit as 62 min. Strictly speaking, two successive tidal cycles within a day are not identical, but as the data are limited to just one lunar day we fit a single function with period 12.

Table 3 shows the analysis of variance (cf. Table 1) along with the parameter estimates for the saturated model, and hence for any submodel.

To fit a model with period 12, one would include only terms with j even. The remaining terms ($j = 1, 3, 5, 7, 9, 11$) can be pooled to provide a residual mean square of 1.53 with 12 degrees of freedom. Including all of the even terms (12 parameters) effectively fits an arbitrary periodic function with period 12. The $j = 2, 4$, and 8 mean squares are individually significantly large by an F test. The pooled mean square for $j = 6$, 10, and 12 is only 1.97 with five degrees of freedom, so the tidal variation is well explained by just the 2, 4, and 8 terms. This leads to the fitted model:

$$\begin{aligned} y_t = 10.07 &+ 7.04\cos(t - 23.6)\frac{\pi}{6} \\ &+ 2.58\cos(t - 0.81)\frac{\pi}{3} \\ &+ 1.62\cos(t - 1.01)\frac{\pi}{3} + e_t, \end{aligned}$$

with error standard deviation σ estimated (from the pooled mean square for $j = 6$, 10, 12 and j odd) to be 1.29 feet.

Figure 1 shows a time plot of the data along with each component cosine wave separately and fitted models with and without the $j = 8$ term. The higher harmonics $j = 4$ and 8 have no particular meaning on their own. The $j = 4$ term has the effect of shaping the basic wave so that it spends 8 time units dropping from maximum to minimum

Table 2 Height of tide y_t (ft) at 24 equally spaced times t

t	1	2	3	4	5	6	7	8	9	10	11	12
y_t	19.8	18.1	12.7	13.5	10.5	10.5	6.8	2.4	1.4	4.7	6.0	13.6
t	13	14	15	16	17	18	19	20	21	22	23	24
y_t	20.8	16.7	14.6	14.0	8.3	9.2	9.2	3.5	1.1	4.7	8.8	10.7

Table 3 Parameter estimates and ANOVA for the data in Table 2

Frequency (j)	Period ($24/j$)	$\hat{\alpha}_j$	$\hat{\beta}_j$	Phase ($\hat{t}_{\phi j}$)	Amplitude ($\hat{\rho}_j$)	Degrees of freedom (df)	Sum of squares (ss)
0	∞	10.067			10.067	1	2432.27
1	24	−0.001	−0.141	17.97	0.141	2	0.24
2	12	2.320	6.650	2.36	7.043	2	595.30
3	8	−0.327	−0.298	4.94	0.442	2	2.35
4	6	1.708	1.934	0.81	2.581	2	79.91
5	4.80	0.221	0.066	0.22	0.231	2	0.64
6	4	−0.517	0.317	1.65	0.606	2	4.41
7	3.43	−0.856	−0.498	2.00	0.991	2	11.78
8	3	−0.842	1.386	1.01	1.621	2	31.54
9	2.67	−0.256	−0.315	1.71	0.406	2	1.98
10	2.40	−0.653	−0.134	1.28	0.667	2	5.34
11	2.18	−0.230	−0.244	1.37	0.335	2	1.35
12	2	0.067		0.00	0.067	1	0.11

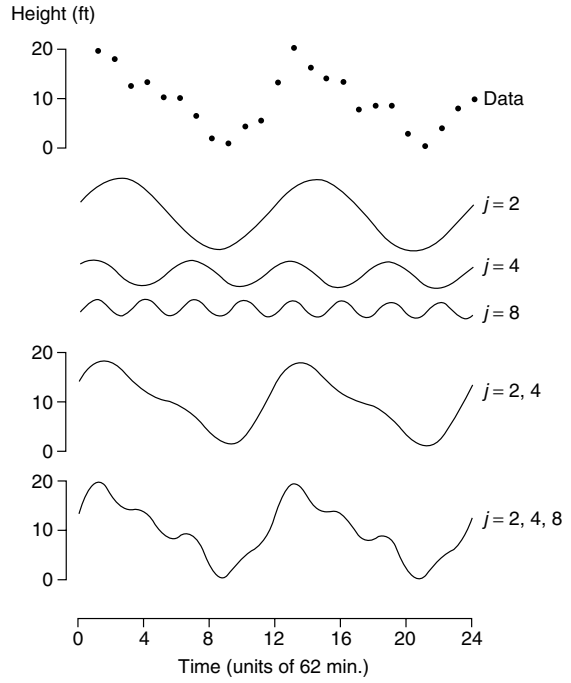

Figure 1 Plot of height of tide vs. time, individual trigonometric terms with j = 2, 4, and 8 cycles per 24 time units, and fitted regression functions with $j = 2, 4$ and $j = 2, 4, 8$

and 4 time units climbing back to the maximum. This is presumably due to the flow of the river. The $j = 8$ term puts steps into the wave to that it does not decrease monotonically. This seems physically implausible, though not impossible; for example, it might be due to local currents caused by a sandbank. Although there is some hint of this feature in both cycles of the data, one should beware of reading too much detail into the fitted curve.

Extensions and Further Material

This topic is discussed in [2, Chapter 17] and in books on time series analysis, particularly [1, Chapter 4] and [3, Chapter 2]. Related Encyclopedia articles include [6] and [7].

There is a generalization of 9, minus the error term, that is fundamental to the frequency domain theory of stationary time series, i.e. when the coefficients α_j and β_j are independent (or in some contexts merely uncorrelated) random variables with zero means and variances given by $\text{var}(\alpha_j) = \text{var}(\beta_j) = \sigma_j^2$. The second-order properties of the model are defined by the set of variances σ_j^2, which constitute the (discrete, nonnormalized) spectral density function (*see* **Spectral Analysis**). It can be shown that, when this model is extended to the continuous range of frequencies $0 \leq \omega \leq \pi$, it defines the class of *all* second-order **stationary** processes (with zero mean).

For cyclic phenomena that are not strictly periodic there are other types of model that may be more **parsimonious**. Perhaps the simplest is the second-order autoregression

$$y_t = \mu_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t,$$

where μ_0 is a constant term, e_t are independent errors with zero means and common variance σ^2, and the

coefficients α_1 and α_2 are such that $-1 < \alpha_2 < 0$ and $\alpha_1^2 + 4\alpha_2 < 0$ (see, for example, [4, Section 3.2.4]). Such a process displays pseudo-periodic behavior, oscillating with varying frequencies around ω_0 given by $\cos\omega_0 = |\alpha_1|/2\sqrt{-\alpha_2}$, i.e. a large amount of the variation in y_t comes from frequencies in the neighborhood of ω_0. Among published examples of such modeling are several analyses of data on annual trappings of the Canadian lynx, including two discussion papers: Tong [9] fitted an autoregression of order 11, while Campbell & Walker [5] fitted a superposition of a pure sine wave and a second-order autoregression.

References

[1] Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.

[2] Bliss, C.I. (1970). *Statistics in Biology*, Vol. II. McGraw-Hill, New York.

[3] Bloomfield, P. (1976). *The Fourier Analysis of Time Series: An Introduction*. Wiley, New York.

[4] Box, G.E.P. & Jenkins, G.M. (1970). *Time Series Analysis, Forecasting and Control*. Holden Day, San Francisco.

[5] Campbell, M.J. & Walker, A.M. (1977). A survey of the statistical work on the McKenzie River series of annual Canadian lynx trappings for the years 1821–1934 and a new analysis (with discussion), *Journal of the Royal Statistical Society, Series A* **140**, 411–431, 448–468.

[6] Ord, J.K. (1985). Periodogram analysis, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson, eds. Wiley, New York.

[7] Parzen, E. (1982). Cycles, in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz & N.L. Johnson, eds. Wiley, New York.

[8] Stuart, R.D. (1961). *An Introduction to Fourier Analysis*. Methuen, London.

[9] Tong, H. (1977). Some comments on the Canadian lynx data (with discussion), *Journal of the Royal Statistical Society, Series A* **140**, 432–436, 448–468.

(*See also* **Seasonal Time Series**)

R.F. GALBRAITH

Trimming and Winsorization

There are several practical problems with the population mean, μ, and its usual estimator, the sample mean, $\overline{X}$. First, μ is not **robust**. Roughly, this means that a small proportion of a distribution can dominate its value. Also, very small shifts in a distribution can result in large changes in μ. For example, μ might correspond to the 0.8 **quantile**, in which case, at least in some situations, it provides a poor reflection of the typical subject under study. The sample mean is not resistant (*see* **Robustness**), meaning that a single unusual observation can completely dominate its value. Another problem is that slight departures from normality can inflate its **standard error**, which in turn can result in relatively low **power** for **hypothesis testing**. Also, a single unusual value, or **outlier**, can inflate the estimate of the standard error. It might be hoped that these problems rarely arise in applied work, but the exact opposite seems to be true.

The contaminated normal distribution provides the classic example of how small departures from normality can inflate the standard error of $\overline{X}$. The distribution is given by

$$H(x) = (1 - \varepsilon)\Phi(x) + \varepsilon\Phi(x/k),$$

where $\Phi(x)$ is the standard **normal distribution**. That is, with probability $1 - \varepsilon$, an observation is sampled from a standard normal distribution; otherwise, sampling is from a normal distribution having standard deviation k. The standard normal and contaminated normal distributions for $\varepsilon = 0.1$ and $k = 10$ are shown in Figure 1. The distributions are similar (as measured by the Kolmogorov distance function; *see* **Kolmogorov–Smirnov Test**), yet the contaminated normal has variance 10.9 versus a variance of 1 for the standard normal.

When **testing hypotheses** or computing **confidence intervals**, more problems arise. First, standard methods for computing confidence intervals can have probability coverage substantially different from the nominal level. The probability coverage can be too low when sampling from **skewed**, light-tailed distributions, and it can be too high when distributions have heavy tails. The left panel of Figure 2 illustrates the first problem by showing the distribution of the one-sample **Student's *t* statistic**

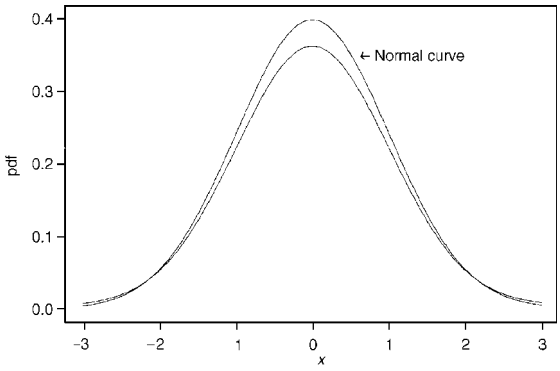

Figure 1 Normal and contaminated normal distributions

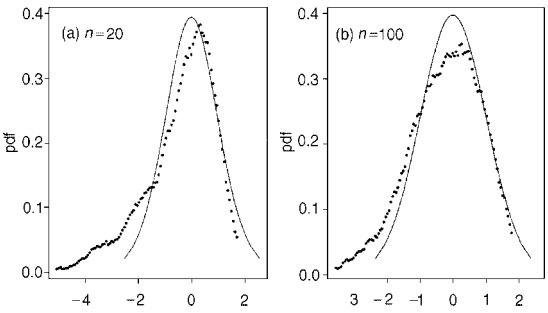

Figure 2 The probability density function of Student's t when sampling from a lognormal distribution (the solid line is the assumed distribution)

when sampling from a **lognormal distribution**. Under standard assumptions, the distribution is symmetric about zero, as shown by the solid line in Figure 2. In actuality, the left tail is too thick, the right tail is too thin, and the mean of the test statistic is approximately -0.5, not zero as is commonly assumed. The result is that power can go down as we move away from the null hypothesis, although eventually it goes up. For a lower-tail test at the 0.05 level, the actual probability of a type I error is approximately 0.15. The right panel of Figure 2 shows that the tail of the actual distribution is still too thick when $n = 100$. Just how large the sample size has to be, to ensure accurate probability coverage, is unknown. From [4], $n = 160$ is not large enough. In the comparison of two or more distributions, similar problems arise when distributions have unequal skewnesses. The second problem arises because sampling from heavy-tailed distributions inflates the sample variance, resulting in

confidence intervals that are too long, which, in turn, can mean poor power.

When attention is focused on making inferences about distributions using some measure of location, trimming and Winsorization provide one approach that has been found to be relatively effective for dealing with the problems just described. (M-estimators (*see* **Robustness**) represent another approach.) The resulting measure of location is more robust than the mean, efficiency is not overly sensitive to small changes in the tails of the distributions, and accurate confidence intervals can be computed for a wider range of situations vs. methods based on means, particularly when distributions are skewed.

Trimming deals with problems associated with the tails of a distribution by removing them. That is, it concentrates on the "middle" portion of the distribution. For example, 20% trimming means that a distribution would be trimmed at the 0.2 and 0.8 quantiles. This is not to say that observations in the tails are uninteresting or unimportant, but for certain purposes they do more harm than good. For a random sample, $X_1, \ldots, X_n$, let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the **order statistics** (the observations written in ascending order). The sample trimmed mean is

$$\overline{X}_t = \frac{X_{(g+1)} + \cdots + X_{(n-g)}}{n - 2g},$$

where $g = [\gamma n]$, the notation $[\gamma n]$ meaning that γn is rounded down to the nearest integer, and γ is the desired amount of trimming. The optimal amount of trimming varies from one situation to another. A common recommendation is $\gamma = 0.2$ (20%), because it maintains reasonably high efficiency under the normal model, and it can have substantially higher efficiency, vs. the sample mean, when distributions have heavy tails. In terms of probability coverage for confidence intervals, there are advantages to having γ close to 0.5, but a negative consequence is low **efficiency** under normality. Also, it seems that as the amount of trimming increases, problems in obtaining accurate confidence intervals diminish substantially up to about $\gamma = 0.2$.

A common misconception is that trimming is equivalent to randomly discarding $2g$ observations. Another common mistake is to apply standard methods for means after trimming. The problem is that the order statistics $X_{(g+1)}, \ldots, X_{(n-g)}$ are dependent random variables, so application of the usual estimate of the standard error of the sample mean to these $n - 2g$ values, to estimate the standard error of $\overline{X}_t$, is inappropriate. If $2g$ observations are randomly removed, the remaining observations would be independent, which differs from trimming. The result is that some of the practical advantages of trimming are not intuitive to many researchers.

A practical problem is to find an appropriate estimate of the standard error of the trimmed mean, and there are theoretical results that supply a useful solution. The resulting estimator depends in part on the Winsorized sample mean (named after the American statistician, C.P. Winsor),

$$\overline{X}_w = \frac{1}{n}\Sigma W_i,$$

where

$$W_i = \begin{cases} X_{(g+1)}, & \text{if } X_i \leq X_{(g+1)}, \\ X_i, & \text{if } X_{(g+1)} < X_i < X_{(n-g)}, \\ X_{(n-g)}, & \text{if } X_i \geq X_{(n-g)}. \end{cases} \quad (1)$$

An estimate of the squared standard error of $\overline{X}_t$ is

$$\frac{1}{n^2(1-2\gamma)^2}\sum(W_i - \overline{W})^2.$$

Another, nearly equivalent, estimate, that seems to have a slight practical advantage in certain situations, is

$$\frac{1}{nh}\sum(W_i - \overline{W})^2,$$

where h, the so-called effective sample size, is $n - 2g$, the number of observations left after trimming. Both estimates are slightly biased, but a Winsorized unbiased estimate can be obtained by replacing n^2 in the first estimate with $n(n-1)$. (See [5] for details.)

Methods for comparing trimmed means have been examined by **simulations** for many problems including one-way and two-way designs, repeated measures (*see* **Longitudinal Data Analysis, Overview**), **random effects** models and **multiple comparisons**. Extensions to **split plot** and higher-way designs are relatively straightforward. Trimming and Winsorization also play a role in **correlation** and **regression**. Details about all of these methods, with related techniques, are summarized in [5]. For relevant theoretical results, see also [1–3].

References

[1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*. Wiley, New York.
[2] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.
[3] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*. Wiley, New York.
[4] Westfall, P.H. & Young, S.S. (1993). *Resampling Based Multiple Testing*. Wiley, New York.
[5] Wilcox, R.R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego.

R. WILCOX

Truncated Survival Times

Truncation of survival data arises when observation of an experimental subject can only occur if the value of the failure time (survival time) lies within a certain interval, (l, r), where l or r may be infinite. When the failure time is outside the truncation interval, no information about the subject is observable; hence the subject is said to be "sampled from a conditional distribution". This feature distinguishes truncation from **censoring**; in the latter situation, subjects are known to have failure times greater (or less) than some fixed constant. An important example of truncation in the analysis of data on **AIDS** arose from investigation of blood transfusions contaminated with the human immunodeficiency virus (HIV) [11] published in 1986. Data from an AIDS registry that included date of onset of AIDS and retrospective determination of transfusion times were used to estimate the distribution of latency times from HIV infection to onset of clinical disease (*see* **Latent Period**). Lui et al. [11] were able to observe only latency times short enough to result in onset of AIDS before the end of the observation period, December 1985. For example, they knew about only those contaminated transfusions from June 1982 that were associated with latency periods of less than 3.5 years. These data are said to be *right-truncated* with a truncation time equal to the difference between the end of the observation period and time of transfusion. Because AIDS was not identified until 1981, latency periods of greater than 2 years were required for transfusions that took place in 1979 to be included in the AIDS registry. When failure times must exceed a certain value for the subject to be observable, such data are said to be *left-truncated*. In the transfusion case, truncation times are equal to the difference between the beginning of the observation period and the time of transfusion. Thus, some transfusion dates result in data that are only right-truncated; other dates yield data that are left- as well as right-truncated. The following sections describe estimation procedures for data that are only right-truncated, or both left- and right-truncated; for methods for data that are only left-truncated, *see* **Delayed Entry**.

Nonparametric Maximum Likelihood Methods

Methods for closed-form **nonparametric maximum likelihood estimation** are available when data are truncated only on one side; the simplest situation occurs when data are only left-truncated. Following an approach described by **Kaplan & Meier** [7] and the formalization of Keiding [8], let $l_1, \ldots, l_n$ be arbitrary left-truncation times, and X be the underlying **random variable** (latency period in our example) with continuous distribution function F. Consider $Y_1, \ldots, Y_n$ to be independent failure times, with Y_i following the conditional distribution of X given $X > l_i$. Here and throughout this section we assume there are no ties (*see* **Tied Survival Times**). We can calculate the number at risk at time x as

$$R(x) = n(l_i \leq x) - n(Y_i < x),$$

where $n(\cdot)$ refers to the number of subjects. The product-limit estimator of F is given by

$$1 - \hat{F}(x) = \Pr(X > x) = \prod_{Y_i \leq x} \left(\frac{1 - 1}{R(Y_i)} \right).$$

Left truncation is equivalent to delayed entry into the risk set. The **Nelson–Aalen estimator** may be used to estimate the integrated hazard under left truncation

$$\hat{\Lambda}(x) = \sum_{Y_i \leq x} \frac{1}{R(Y_i)}.$$

When data are right-truncated, F is **identifiable** only if the longest right-truncation time, $r_M = \sup\{r_i\}$, exceeds the longest possible time of failure. Otherwise one can identify only $G(x) = F(x)/F(r_M)$ [10]. Let z_i denote the chronologic time of transfusion, and R_* denoted the end of the observation period. The right-truncation times r_i are equal to $R_* - z_i$. Once again, we let Y_i denote the observed failure times (latency in our example), sampled from X given $X < r_i$. Nonparametric maximum likelihood estimation in the setting where only right truncation occurs can also be performed using a product-limit estimator. Once again, the presence of right censoring can easily be accommodated. Intuition into the approach is

provided by considering the problem in reverse time. Consider a "reverse time" transformation, $S = R_* - X$, which transforms right truncation into (easier to deal with) left truncation. We define a risk set at (reverse) time s

$$R(s) = n(R_* - r_i \le s) - n(R_* - Y_i < s).$$

Thus, $R(s)$ consists of subjects who have z_i, such that $z_i \le s$ (or $r_i \ge x$), and failure times such that $R_* - Y_i \ge s$ or $(Y_i \le x)$. The product-limit estimator is then

$$\Pr(S > s | S \ge 0) = \prod_{R_* - Y_i \le s} \left(\frac{1-1}{R(R_* - Y_i)} \right)$$

for $R_* - Y_M \le s \le R_*$, and 1 for $0 \le s \le R_* - Y_M$, where Y_M is the maximum of the observed failure times. Since

$$\begin{aligned} \Pr(S > s | S \ge 0) &= \Pr(X < R_* - s | X < R_*) \\ &= G(R_* - s), \end{aligned}$$

$$\hat{G}(x) = \prod_{Y_i \ge x} \left(\frac{1-1}{R(R_* - Y_i)} \right),$$

for $0 \le x \le Y_M$, and 1 for $Y_M \le x \le R_*$. In practice one can compute these estimates with software for product-limit estimates that permit left truncation (delayed entry). Subjects enter the risk set at time z_i and fail at time $R_* - Y_i$. Asymptotic properties of the estimators have been studied by Woodroofe [19], Wang et al. [18], and Keiding & Gill [9].

Example 1

Lagakos et al. [10] have considered the problem of estimating and comparing the induction-time distribution for 258 adults (group 0) and 37 children (group 1) infected by blood transfusion, diagnosed by June 30, 1986, and reported to the Center for Disease Control before January 1, 1987. The data were been condensed by grouping dates of infection and AIDS into 3-month intervals beginning April 1, 1978. Thus $z = 0$ denotes an infection occurring between April 1, 1978 and June 30, 1978, $z = 0.25$ denotes an infection occurring between July 1, 1978, and September 30, 1978, and $R_* = 8$. The data are right-truncated because only cases (z_i, Y_i) such that $z_i + Y_i \le R_*$ were observed. Since there are ties of Y_is in this example, we have to modify the methodology described above. Following [10], let $v_1 < \ldots < v_m$ denote the distinct values of $(Y_1, \ldots, Y_n)$, let $u_j = R_* - v_j$, and define $n_j = n(Y_i = v_j)$, for $j = 1, \ldots, m$. Notice that N_j in [10] is the risk set $R(R_* - v_j)$. The nonparametric maximum likelihood estimator of $G(x)$ is

$$\hat{G}(x) = \prod_{v_j \ge x} \left(\frac{1 - n_j}{N_j} \right),$$

for $0 \le x \le v_m$, and 1 for $v_m \le x \le R_*$.

Figure 1 shows the estimates $\hat{G}(x)$ for the groups of adults and children. Conditional on being less than 8 years, the induction times for children tend to occur sooner than those for adults, with estimated medians of about 3 and 5.5 years, respectively. To compare to the above nonparametric results, Lagakos et al. [10] also used a parametric likelihood method to estimate F. They fit the **Weibull** model $F(x) = 1 - \exp\{-(\theta x)^r\}$ to the adult data, giving a **likelihood** function that is extremely flat over a range of parameter values representing a wide range of induction distributions. Figure 1 also displays the maximum likelihood estimates of G corresponding to Weibull distributions with medians of 8.5 years and 210 years. Both parametric maximum likelihood

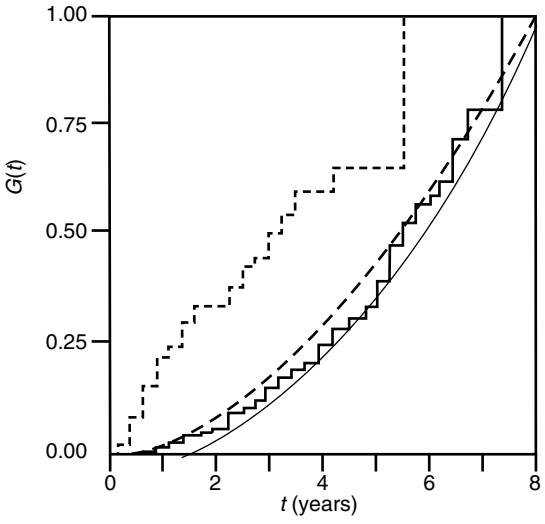

Figure 1 Nonparametric estimates of G based on data in Table 1: - - - - - children; ——— adults. Parametric estimates of G based on Weibull distributions with medians of: – – – 8.5 years; ——— 210 years. Reprinted from [10] by permission of *Biometrika*

estimates agree with the nonparametric maximum likelihood estimate $\hat{G}(x)$ reasonably well, indicating that very different $F(\cdot)$ distributions can have similar $G(\cdot)$ components. See also [5, 9], and [13].

When both left and right truncation occur, closed-form solutions are not possible. A method for obtaining the nonparametric maximum likelihood estimate (NPMLE) was proposed by Turnbull [17], that also accommodates left, right, and interval censoring (*see* **Turnbull Estimator**). Turnbull described a self-consistent approach to estimation that turns out to be a version of the **EM algorithm**. In the presence of both arbitrary censoring and truncation, however, Turnbull's self-consistent algorithm has to be modified [3]. Alioum & Commenges [1] provide a detailed correction of Turnbull's method and an extension to the regression analysis. Here, we consider only the problem of truncation, and assume that the failure times are uncensored. Let the observed times of failure be at $Y_i, i = 1, \ldots, N$. Let the probabilities of failure associated with Y_i be p_i. The likelihood can be written as

$$L(p_1, \ldots, p_N) = \prod_{i=1}^{N} \frac{p_i}{\sum \beta_{ij} p_j},$$

where β_{ij} is 1 for all Y_j that are within the truncation interval for the ith person. Each observed individual might be considered to represent others–Turnbull called them "**ghosts**"–whose events occurred outside the truncation interval. For example, in the transfusion example above, the ghosts corresponding to case i would include people infected by transfusion at time z_i but who had latencies longer than $R_* - z_i$. The expected number of such subjects who will have disease onset at time Y_j is

$$\mathrm{E}(I_{ij}) = \frac{(1 - \beta_{ij}) p_j}{\sum \beta_{ik} p_k}. \tag{1}$$

Once again, in order for F to be identifiable, r_M must exceed the longest possible failure time.

Expression (1) permits construction of an EM algorithm. The complete data log likelihood (if I_{ij} were known) is

$$\log L_{\mathrm{c}} = \sum_{ij} (\alpha_{ij} + I_{ij}) \log \ p_j,$$

where $\alpha_{ij} = 1$ if the ith subject is observed to fail at time j and 0 otherwise. Therefore the rth iteration of the E-step is evaluation of

$$\mathrm{E}^r(\log \ L_{\mathrm{c}} | p, Y) = \sum_{ij} (\alpha_{ij} + \mathrm{E}^r(I_{ij})) \log \ p_j$$

using (1); and the M-step maximizes this simple expectation to obtain

$$p_k^{r+1} = \frac{\sum_i \alpha_{ik} + \mathrm{E}^r(I_{ik})}{\sum_{ij} \alpha_{ij} + \mathrm{E}^r(I_{ij})}.$$

Regression Models

Regression models have also been developed for settings in which the effect of **covariates** on failure time is of interest. Suppose that we are interested in determining whether a covariate, **Z**, such as age at transfusion, chronologic time of transfusion, or region, affected the time from transfusion to onset of AIDS. We divide the time interval $[0, R_*]$ into units of equal length, where 0 is the time of the earliest transfusion. Let $A(j|\mathbf{z})$ denote the number of cases corresponding to covariate $\mathbf{z}$ that have latency j for $0 \le j \le T$, where T is the longest latency time that can be reliably estimated. Since the observed cases have different transfusion times and are observed in the same chronologic time interval $[0, R_*]$, the observed portion of $A(j|\mathbf{z})$, denoted $A(j|\mathbf{z}, r) (0 \le j \le r)$, depends on the truncation interval $[0, r]$.

If the $A(j|\mathbf{z})$ are regarded as multiple responses at $0 \le j \le T$, models in **categorical data analysis** can be used to model $A(j|\mathbf{z})$. A number of investigators have considered the following **multinomial** response model (also with censoring), where failure time X takes only discrete values j. The probability that failure time, X, equals j, given covariate $\mathbf{Z}$ is

$$p(j|\mathbf{Z}) = \Pr(X = j|\mathbf{Z}) = \frac{\exp(\eta_j(\mathbf{Z}))}{\sum \exp(\eta_k(\mathbf{Z}))},$$

where

$$\eta_j(\mathbf{Z}) = \begin{cases} 0, & \text{if } j = 0, \\ \alpha_j + \mathbf{Z}'\boldsymbol{\beta}_j, & \text{if } 1 \le j \le T. \end{cases}$$

An alternative is the discrete **proportional hazards model** [15]:

$$\Pr(X = j|\mathbf{Z})$$

$$= \begin{cases} (p_0, \ldots p_{j-1})^{\exp(\mathbf{Z}'\boldsymbol{\beta})}[1 - p_j^{\exp(\mathbf{Z}'\boldsymbol{\beta})}], \\ \quad \text{if } j < T, \\ (p_0, \ldots p_{T-1})^{\exp(\mathbf{Z}'\boldsymbol{\beta})}, \\ \quad \text{if} j = T, \end{cases}$$

where

$$p_j = \Pr(X > j + 1 | X > j), \quad \text{for } 0 \leq j \leq T - 1.$$

This model implies that the probability that a case is reported at lag time j conditional on being reported at time j or later is given approximately by

$$P(X = j | \mathbf{z}, X \geq j) = \exp(\alpha_j + \mathbf{z}'\boldsymbol{\beta}),$$

where $\alpha_j = \log(-\log(p_j))$.

We let $\boldsymbol{\theta} = (\alpha, \beta)$ and denote the dependence of $p(j|\mathbf{Z})$ on $\boldsymbol{\theta}$ by $p(j|\mathbf{Z}, \boldsymbol{\theta})$. We also denote the truncation interval for the ith subject as $[0, r_i]$. The log likelihood under a chosen model and an observed sample of size N, denoted $\{A(j|\mathbf{z}_i, r_i)\}_{1 \leq j \leq r_i; 1 \leq i \leq N}$, is given by

$$\log\ L(\theta) = \sum_i^N \sum_j^{r_i} A(j|\mathbf{z}_i, r_i) \times \left\{ \log[p(j|\mathbf{z}_i, \boldsymbol{\theta})] - \log \left[\sum_{j=0}^{r_i} p(j|\mathbf{z}_i, \boldsymbol{\theta}) \right] \right\}. \quad (2)$$

Note that the second term above is the result of right truncation and disappears if the data are not truncated.

Estimation of model parameters as well as the variance–**covariance matrix** can be obtained by maximizing the log likelihood above using the Newton–Raphson method (*see* **Optimization and Nonlinear Equations**). When the data are not truncated, the analytical calculations of the first- and second-order derivatives, necessary for the Newton–Raphson method, are simplified. In this case, some standard **software** packages can be utilized to obtain the estimates. For example, estimation of the parameter vector $\boldsymbol{\theta}$ for the multinomial response model can be consistently estimated by using the loglinear model

$$\log[\mu_j(\mathbf{z})] = \phi(z) + \alpha_j + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_j, \quad 0 \leq j \leq T, \quad (3)$$

with the error following a **Poisson** distribution, where $\mu_j(z) = \mathrm{E}[A(j|z)]$ and the **nuisance parameter** ϕ (**z**) is termed *incidental* [12]. So statistical packages such as GLIM, which also implement the complementary log-log model (*see* **Quantal Response Models**), are easily utilized to estimate the parameter vector $\boldsymbol{\theta}$. The simplicity of estimation in the absence of truncation suggests using the EM algorithm for estimation under either model, in the presence of truncation [14].

Some other discrete-time regression models for right-truncated data have been developed and applied in the analysis of AIDS incidence and induction-time distributions [2, 4, 6, 16]. Alioum & Commenges [1] discussed methods for fitting a continuous proportional hazards model for truncated and censored data.

Example 2

We illustrate the use of regression models for right-truncated data by applying these methods to a problem that arises in analyses of data on AIDS surveillance (*see* **Surveillance of Diseases**) in the US. A data set released by the US Centers for Disease Control in the first quarter of 1990 includes AIDS diagnosis date and reporting date in six geographic regions: Northeast, Central, West, South, Mid-Atlantic, and Other. In this setting, R_*, the end of the observation period, is the first quarter of 1990, and the observed portion of $A(j|\mathbf{Z}), 0 \leq j \leq r$, depends on the truncation interval $[0, r]$, where $r = R_* - x$ and x is the chronologic time of AIDS diagnosis. To adjust for reporting delay, we first model the covariate effects of the chronologic time of AIDS diagnosis and region by fitting the multinomial response model [14] with

$$\eta_j(Z) = \alpha_j + \text{region} \times \gamma_j + x\zeta_j + (\text{region} \times x)\psi_j, \quad 0 \leq j \leq T,$$

where x is the year of diagnosis, and "region" is a vector of five indicator variables that designate any five of the six regions. When $\psi_j = 0$, for $0 \leq j \leq T$, there is no interaction between region and x and the delays have the same trend across the regions.

The reported cases were grouped on quarterly intervals, and cases reported with a delay time of more than 12 quarters were grouped into one category, $T = 12$. The EM algorithm was used for estimation under the model, as described above. The median reporting delays for the six geographic regions obtained from the above model are plotted in Figure 2. The reporting delays appear to have

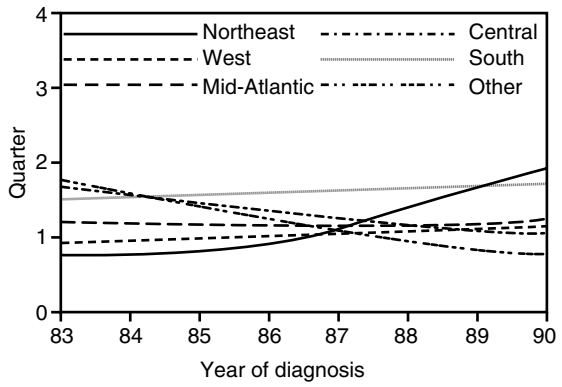

Figure 2 Reporting delays for six geographic regions. Reprinted from *Biometrics*, with permission

lengthened in the late 1980s, except for the Central region and Other, in which the delays seem to have shortened from 1983 to 1990. The South region has the longest delay in reporting. Figure 2 also shows strong interaction between the chronologic time trend and geographic regions. Such analyses allow us to distinguish between trends in reporting and in AIDS incidence; without reporting delay estimates, AIDS surveillance data are uninterpretable. The estimated AIDS incidence can be obtained by dividing the AIDS incidence at each chronologic time period by the estimated probability that an AIDS case would be reported in time for inclusion in the data base.

References

[1] Alioum, A. & Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data, *Biometrics* **52**, 512–524.

[2] Brookmeyer, R. & Liao, J. (1990). The analysis of delays in disease reporting: methods and results for acquired immunodeficiency syndrome, *American Journal of Epidemiology* **132**, 355–365.

[3] Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations, *Journal of the Royal Statistical Society, Series B* **56**, 71–74.

[4] Harris, J.E. (1990). Reporting delays and the incidence of AIDS, *Journal of the American Statistical Association* **85**, 915–924.

[5] Kalbfleisch, J.D. & Lawless, J.F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS, *Journal of the American Statistical Association* **84**, 360–372.

[6] Kalbfleisch, J.D. & Lawless, J.F. (1991). Regression models for right-truncated data with application to AIDS incubation times and reporting lags, *Statistica Sinica* **1**, 19–32.

[7] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation for incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[8] Keiding, N. (1988). Nonparametric estimation under truncation, in *Encyclopedia of Statistical Sciences* Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 357–359.

[9] Keiding, N. & Gill, R.D. (1990). Random truncation models and Markov processes, *Annals of Statistics* **18**, 582–602.

[10] Lagakos, S.W., Barraj, L.M. & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS, *Biometrika* **75**, 515–523.

[11] Lui, K.J., Lawrence, D.N., Morgan, W.M., Peterman, T.A., Haverkos, H.H. & Bregman, D.J. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome, *Proceedings of the National Academy of Science* **83**, 2913–2917.

[12] McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models*. Chapman & Hall, London.

[13] Medley, G.F., Billard, L., Cox, D.R. & Anderson, R.A. (1988). The distribution of the incubation period for the acquired immunodeficiency syndrome (AIDS), *Proceedings of the Royal Society of London, Series B* **233**, 367–377.

[14] Pagano M., Tu X.M., De Gruttola V. & Mawhinney S. (1994). Regression analysis of censored and truncated data: estimating reporting delay distributions and AIDS incidence from surveillance data, *Biometrics* **50**, 1203–1214.

[15] Prentice, R.L. & Gloeckler, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics* **34**, 290–295.

[16] Tu, X.M., Meng, X.L. & Pagano, M. (1993). The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data, *Journal of the American Statistical Association* **88**, 26–36.

[17] Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored, and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 290–295.

[18] Wang, M.C., Jewell, N.P. & Tsai, W.Y. (1986). Asymptotic properties of the product limit estimate under random truncation, *Annals of Statistics* **14**, 1597–1605.

[19] Woodroofe, M. (1985). Estimating a distribution function with truncated data, *Annals of Statistics* **13**, 163–177.

VICTOR DE GRUTTOLA & Q. LIAO

Tukey, John Wilder

Born: June 16, 1915, New Bedford, Massachusetts.
Died: July 26, 2000, New Brunswick, New Jersey.

John Wilder Tukey (JWT) was one of the most influential statisticians of the twentieth century. The combination of his unorthodox education, scientific interests, and exposure to a diverse range of applied problems facing scientists and engineers across many disciplines, enabled him to make significant contributions in many areas and to advance a new basic philosophy for how statisticians approach data. He may be best known to statisticians for founding the field of exploratory data analysis [70], for introducing the jackknife as a tool for characterizing the uncertainty in a statistic [63], and for guiding and contributing to research in robust methods [3, 8, 67]. But his name is familiar to scientists in many other fields for diverse reasons: Tukey's lemma (mathematics); fast Fourier transform, or FFT (digital computing, engineering, and medicine), for which he received the Medal of Honor from the Institute of Electronic and Electrical Engineers (1982); multiple comparisons (psychology and education); principles of sampling (social science and medicine); binomial probability paper (quality control); cloud seeding experiments (meteorology); spectrum and cepstrum estimation (geophysics); smoothing (science and engineering); and coining the word "bit" (computer science).

Tukey had tremendous vision, addressing problems with solutions whose need was recognized, sometimes only years or decades later. Two notable examples are **exploratory data analysis** [70], with its emphasis on statistical graphics (*see* **Graphical Displays**), and *Index to Statistics and Probability* [69], forerunner to the present *Current Index to Statistics* [2]; both foreshadowed needs to cope with the data explosion of the 1990s. He believed that solving the *exact* problem, even with only an *approximate* solution, was better than solving the convenient and easier (but only approximate) problem with the exact solution: "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" [68, p. 13]. Consequently, an important theme throughout his work is the validity of **inference** that does not depend heavily on assumptions (e.g. Gaussian-distributed errors). He published over 300 papers (nearly 100 of which appeared after his federally mandated retirement in 1985), graduated 55 Ph.D. students, and advised countless others, both undergraduate (e.g. David L. Donoho, William F. Eddy, Paul A. Tukey) and graduate (e.g. Frederick Mosteller, Marvin L. Minsky, Yoav Benjamini).

Tukey received numerous awards, including honorary doctorates from seven universities; Deming and Shewhart Medals (American Society for Quality); S.S. Wilks Medal (American Statistical Association); James Madison Medal (Princeton University); and the National Medal of Science from President Nixon in 1973, "for his studies in mathematical and theoretical statistics. . .and for his outstanding contributions to the applications of statistics in the physical, social, and engineering sciences". His entire career was devoted to academic, government, and public service, through his association with Princeton University (1937–2000), employment and consultant for the Bell System (1945–2000), membership in the National Academy of Sciences (1961–2000), and participation on numerous government panels and academic committees. He was a member of the President's Scientific Advisory Committee (PSAC) for Presidents Eisenhower and Kennedy, and headed PSAC working groups for Presidents Johnson and Nixon (environment, 1964–1965; chemicals and human health, 1971–1972). He was also a member of Technical Working Group 2 of the Conference on the Discontinuance of Nuclear Weapons Tests in Geneva (1959) and the United Nations Conference on the Human Environment in Stockholm (1972) for the US State Department. Retired Chairman of the Board of Bell Laboratories William O. Baker commented on his influential contributions to the labs and to science at large, saying, "We have watched at least four Presidents of the United States listen to him and heed his counsel" [38, p. 335].

Tukey's career included a variety of projects that led to important methodological developments in biometry and biostatistics:

The National Halothane Study. This study [16, 42] was conducted under the auspices of the National Research Council in response to concerns about a possible association between the use of halothane as an anesthetic in surgical operations and fatal hepatic necrosis (liver failure). The critical statistical issue was the quantitative comparison of death rates for

various anesthetics (halothane, ether, cyclopropane, nitrous oxide-barbiturate, "other"), in the presence of unavoidable **confounding** variables (e.g. type of operation; hospital where operation was performed; length of operation; and age class in 10-year intervals, gender, ethnicity, and physical status of the patient), whose effects grossly dominated the differences among the anesthetics. With millions of **covariate** combinations and only 800 000 cases, Tukey and his colleagues developed a method to standardize rates (*see* **Standardization Methods**), or adjust for the joint effects of suitably selected combinations of variables, which they called "smear-and-sweep analysis" [48]. Similar issues arose in quantifying the effect of sulfinpyrazone after myocardial infarction [5]. Tukey returned to the important issue of adjustment in the 1980s in connection with the US decennial census [25]: in the spirit of approximate solutions to right problems, he wrote, "adjustment to reduce bias cannot wait for perfection but must be considered as soon as we recognize that it will help. Incomplete adjustment to reduce bias, since that is all we can ever do, is desirable and not to be denigrated" [76, p. 127].

Multiple Comparisons. The widely circulated 1953 manuscript, "The problem of multiple comparisons," appeared in Volume VIII of *The Collected Works of John W. Tukey* [80]. It set out the concepts of "error rate per comparison/determination", "error rate per family/batch", and "error rate familywise/batchwise", as well as "error rate budgeting" [80, p. 5], which guided the research into these methods for the next 40 years. He advocated the "wholly significant difference" (WSD), for the allowance when comparing any two means in a **fixed-effects**, one-way analysis of variance (*see* **Experimental Design**) or **linear regression** model (based on the **Studentized range** statistic, and later called "honestly significant difference" (HSD); see [41, p. 92]), as opposed to the "least significant difference" (LSD), which is based on the F-statistic (*see* ***F*** **Distributions**) in a one-way analysis of variance; see also [9] for a broad overview. With increasing volumes of data (e.g. microarrays, county rates of disease mortality and incidence), Tukey returned to this area in the 1990s, in his advisory role on committees for the National Assessment of Education Progress (NAEP), which involved thousands of comparisons of measures among various states in the United States [1], and for presenting results of plant breeding experiments [7] and of animal studies [22, 82]. He acknowledged "the usefulness of – and need for – a variety of procedures reflecting the *varied strength* of different experiments" [81, p. liv].

Jackknife. In perhaps the most oft-cited abstract in the statistics literature, Tukey described, in only six sentences, the **jackknife** as a procedure for assessing the uncertainty in a statistic [63]. By leaving out one observation (or one group of observations) at a time, one can recompute the statistic, and then calculate the usual sample **standard deviation** of these "leave-out-one" statistics, to obtain an estimate of the **standard error** of the statistic computed on the entire sample (see also [43, Chapter 7]). The use of the jackknife had a huge impact on statistical practice intervals. Efron later analyzed the bias and **variance** of such jackknifed estimates of the standard error, leading him to propose the **bootstrap** [23, 24]. Fernholz, Morgenthaler, and Tukey [27] later combined jackknife samples using principles of experimental design and Hadamard matrices (*see* **Response Surface Methodology**) to derive methods of "nominating" **outliers**.

Clinical Trials. Tukey's recommendations for the design and analysis of **clinical trials** included the unequivocal value of the focused randomized clinical trial and the ethical consequences of lesser, unfocused alternatives [71], the combination of covariates in assessing treatment effects [77], problems of **multiplicity** [78], and designs whose analysis relies on "probability statements that depend on only exactly how the trial was conducted – not at all on assumptions", which he termed the "platinum standard" [79, p. 266]. Several of the ideas recommended by Brillinger, Jones, and Tukey [15], for the design of cloud-seeding experiments, are applicable to the design of clinical trials with tight constraints.

Statistical Problems of the Kinsey Report. In the appendix to their evaluation of the "Kinsey report" [17], **Cochran**, Mosteller, and Tukey discussed important principles in survey sampling (*see* **Sample Surveys in the Health Sciences**), such as sample selection, generalization from sample to population, accuracy of interview data, systems of interviewing,

methods of checking and analyzing data, and the reporting of results [17–19].

Statistical Mapping. Tukey emphasized graphical displays and used them heavily in all his work, particularly for geographical data. In "Statistical mapping: what should *not* be plotted" [74], he argued that chloropleth ("patch") maps inappropriately draw too much attention to region size and political boundaries, and that displayed rates, adjusted properly for age, should be further adjusted for other variables known to have strong effects (e.g. lung cancer and smoking, or a proxy for smoking such as urbanization; see [31]), and then smoothed [72, 74]. The *Atlas of United States Mortality* [44] implemented many of these ideas.

Tukey was famous for parallel processing; rarely did he engage in only one activity at a time. Seminar speakers recall seeing him in the audience, working on an entirely different project or even seemingly completely asleep, yet rising at the end to deliver insightful comments on the presentation [37, p. xlv] – often telling the speaker not only what he did wrong, but also how he might do it right the next time [32]. He was a large man, often seen wearing a black polo shirt, whose pocket bulged with his address book and four-color pens. He relaxed by frequenting mystery book stores and organizing birding expeditions in various places where meetings took him, all over the world. He was able "to carry out two or three times the load of ordinary men" [12]. He was also known for "making up words" – but those who worked with him knew that he proposed a new word only when he was absolutely certain that no other word (in any of the large number of languages that he knew) precisely matched his intended meaning (e.g. "hinge" or "fourth" versus "sample quartile"; "batch" versus "random sample"; [34, p. 4]).

Tukey was born in New Bedford, Massachusetts, the only child of Adah M. (Tasker) and Ralph H. Tukey, both 1898 graduates (first and second) of Bates College (Lewiston, Maine), and both educators (his father earned a Ph.D. degree in classics and chaired the Latin department at New Bedford High School). Recognizing JWT's genius at age 3, they educated him at home. He did attend high school "for one term in French and some mechanical drawing" [14, p. 26], and spent much time in the New Bedford public library, reading the *Journal of the American Chemical Society* and the *Transactions of the American Mathematical Society*. He was admitted to Brown University in 1933 on the basis of his College Board Examinations. He received the Sc.B. and Sc.M. degrees in chemistry in 1936 and 1937 and entered the Ph.D. program in chemistry at Princeton in September 1937. His concurrent interest in mathematics flourished around the stimulating environment at Fine Hall and the Institute for Advanced Study at Princeton, so he soon transferred to the Mathematics Department, passed his oral examinations in May 1938, and completed his dissertation the following spring (Tukey [52], published by Princeton University Press as *Convergence and Uniformity in Topology*, [53]). During his years as a graduate student, his friends and later collaborators included physicists Richard Feynman and Lyman Spitzer [49, 50]; mathematicians Ralph Boas [11] and Arthur Stone [51]; William Baker, later chairman of the board of Bell Laboratories; and Frederick Mosteller, a lifelong friend and frequent collaborator (of four books and 24 articles). Tukey joined on the mathematics faculty at Princeton, as Instructor (1939–1941), Assistant Professor (1941–1948), Associate Professor (1948–1950), Professor (1950–1965), and later as Professor of Statistics (1965–1985), Donner Professor of Science (1976–1985), and Professor Emeritus and Senior Research Statistician (1985–2000).

During World War II and while teaching at Princeton, Tukey held the position of Research Associate in the Fire Control Research Office (1941–1944). He worked closely with his mentor, Charles P. Winsor, for whom he later named the robust location estimate, "Winsorized mean" [30, 66], and to whom he dedicated the book *Exploratory Data Analysis* [70]. Tukey and his colleagues worked on projects related to the war effort, including stereoscopic range finders and testing of rocket powders. In 1945, he joined Bell Laboratories, first as a Member of the Technical Staff (MTS), working on anti-aircraft guided missiles (later Nike and Nike-Ajax). Samuel S. Wilks in the Mathematics Department at Princeton asked him to teach statistics, and Tukey remained part-time at both institutions for the next 40 years. His dual career at Bell Labs proceeded from MTS to Assistant Director of Research in the Department of Communications Principles (1958–1961), and then as Associate Executive Director of Research in Information Sciences (1961–1985) and consultant (1985–2000), working on problems of signal processing, communications

engineering, and information retrieval and interpretation. *The Measurement of Power Spectra from the Point of View of Communications Engineering* [10] was the leading authoritative work on spectrum analysis (*see* **Spectral Analysis**); its principles were used even more widely after Tukey introduced computer scientists to an **algorithm** that permitted fast and efficient computation of **fast Fourier transforms (FFT)**; [20].

In the 1950s and 1960s, Tukey wrote papers on concepts in mathematical statistics, **regression** and **analysis of variance**, and methodology that assume no particular distributional form (i.e. **nonparametric**); for example, population **tolerance** limits and **confidence** bands for a continuous, cumulative, distribution function [46, 47]; **means**, variances, and covariances of **order statistics** for small samples from Gaussian (*see* **Normal Distribution**) and non-Gaussian distributions [33]; a useful bound on the ratio of the variance of the *mis*-weighted mean to the variance of the *optimally* weighted mean [54]; the robustness of **Student's t** confidence intervals (Tukey [55]; see also [6]).

His numerous contributions in **Biometrics** have had important consequences for later practice and research in statistics. Tukey [56] derived the allowance for comparing all pairwise differences among several means in a one-way analysis of variance based on the studentized range distribution (distribution of the range of Gaussian random variables all having the same variance, estimated by the **mean square for error**), now called "Tukey's method of multiple comparisons." In "One degree of freedom for nonadditivity" [57], Tukey developed a test for an **interaction** term (of a specific form) between the two factors in a two-way layout with no replication. McNeil and Tukey [40] proposed diagnostic plots for estimating this interaction parameter (see also [43, Chapter 9], and [35, Chapter 3]). A short abstract published in *Biometrics* [65] contained technically advanced theory for the development of highly fractionated and saturated experimental designs (*see* **Fractional Factorial Designs**) to investigate many factors in very few runs (see also [64]), potentially applicable to the design of clinical trials today. Several other papers addressed the analysis of single and higher-order classifications [59–61] Papers on **transformations**, which he later called reexpressions (which later included combinations of two or more transformations to different segments of the data), also have had a lasting impact on statistical practice ([28, 62]; see also [34]).

Tukey's association with Winsor led him to distrust analyses that relied heavily on often unverifiable assumptions about the data. Many of his papers provide methods with minimal reliance on distributional assumptions. Cornfield and Tukey [21] concentrated on the expectation of mean squares in crossed and nested classifications, "based on a model of sufficient generality and flexibility that the necessary assumptions concern only the selection of the levels of the factors [e.g. fixed or random] and not the behavior of what is being experimented upon [e.g. the underlying distributions]" (p. 907). In "Components in regression," Tukey was "principally concerned with simple linear regression where both variates are subject to 'error"' [58, p. 34], and laid out the problems later addressed by the field now known as "**errors in variables**." In a seminal paper, "A survey of sampling from contaminated distributions," Tukey [82] showed that the asymptotic variance of the mean absolute deviation from the mean is *less* than that of the sample standard deviation if the underlying distribution is a contaminated normal, when the fraction of the contamination by a normal distribution having three times the standard deviation of the target normal, is as small as 0.0018 (less than two observations per thousand). This paper guided the development of robust methods (*see* **Robustness**) for the next two decades, including the "Princeton Robustness Study" [3], which investigated the performance of 65 estimators of location. A by-product of this research was the development of a clever **simulation** algorithm, called the Monte Carlo Swindle, described in Andrews et al. [3, § 4D, pp. 61–63]. Later, Beaton and Tukey [8, p. 151] proposed the "biweight" as a robust estimator of location; (see also [43, pp. 205–206]). Beaton and Tukey were concerned with a regression coefficient, but the biweight has been shown to be remarkably efficient and to perform extremely well in a variety of other contexts.

Tukey's extensive experience with data and association with Charles P. Winsor and Edgar Anderson led him to develop a more flexible approach to data, "exploratory data analysis", which he later described as "an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as for those we believe might be there. Except for emphasis on graphs, its

tools are secondary to its purposes" [73]. His book, *Exploratory Data Analysis*, or *EDA* [70] describes many such tools, including stem-and-leaf displays; boxplots (extended by McGill, Tukey, Larsen [39], and by Rousseeuw, Ruts, Tukey [45]); letter value displays; robust smoothing by running medians; median polish; and resistant line fitting. Subsequent books [34–36] emphasized the exploratory approach in practice. His methods were designed to extract the "fit" in the data decomposition $data = fit + residual$, with special emphasis on graphical displays. "Projection pursuit," an algorithm to identify interesting unexpected structure in higher-dimensional data [29], has a similar goal. In contrast, "confirmatory analysis" emphasized fitting specific models and testing particular hypotheses. Despite the distinction between these two approaches, Tukey nonetheless believed that, "We need both exploratory *and* confirmatory" [75].

Even after this age-mandated retirement from AT & T and Princeton in 1985, Tukey's commitment was unflagging. He continued to write, attend conferences and workshops, and discuss statistical problems until his final days. He consulted extensively for various organizations, including Xerox Palo Alto Research Center (1985–2000), which led to 10 patents on which his name appears as coinventor; Health Effects Institute, which sponsors epidemiological studies on matters related to health; Merck Laboratories (1952–2000); Educational Testing Service (1965–2000); Schering-Plough; and Pfizer. He remained generous with his time and ideas to both students and colleagues, and the research ideas that he conceived continue to be pursued today. Upon his death, a press release from Princeton University (26 July 2000) quoted Princeton Emeritus Professor of Physics John A. Wheeler as saying: "I believe that the whole country – scientifically, industrially, financially – is better off because of him and bears evidence of his influence."

Further information about his life appears in various sources: *The Collected Works of John W. Tukey* (brief biography by Frederick Mosteller); *The Practice of Data Analysis* ([14]; biography, pp. 5–8; interview conducted by L.T. Fernholz and S. Morgenthaler, pp. 26–45), [13], interviews conducted by Anscombe [4] and Fernholz and Morgenthaler [26], and special issues of *Technometrics* (August 2001), *The Annals of Statistics* (May 2002), and *Statistical Science* (August 2003).

Acknowledgments

The author expresses sincere thanks to Dr. David C. Hoaglin and Mr. F. R. Anscombe for their comments and suggestions on an earlier version of this article, which led to several important corrections and additions, and also to Dr. Barry I. Graubard and Dr. Michael Cohen, for copies of Tukey's less easily obtained articles. (The author takes responsibility, with regret, for remaining errors and omissions.)

References

Note: Letters after years cited in Tukey's publications between 1939 and 1993 correspond to his bibliography in "The Publications and Writings of John W. Tukey," *The Annals of Statistics* **30** (Volume 6), 1666–1680, 2002. Except for articles that appeared in 1948 and 1949, and discussion or comment papers, these letters are the same as those given in *The Collected Works of John W. Tukey, Volume VIII: Multiple Comparisons* (H.I. Braun, ed.), pp. xvii–xli.

[1] Almond, R.G., Lewis, C., Tukey, J.W. & Yan, D. (2000). Displays for comparing a given state to many others, *The American Statistician* **54**, 89–93.

[2] American Statistical Association. (2003). *Current Index to Statistics*, `http://www.amstat.org`.

[3] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972e). *Robust Estimates of Locations: Survey and Advances*. Princeton University Press, Princeton.

[4] Anscombe, F.J. (1988). A conversation with Frederick Mosteller and John W. Tukey, *Statistical Science* **3**, 136–144.

[5] Anturane Reinfarction Trial Research Group. (1978d). Sulfinpyrazone in the prevention of cardiac death after myocardial infarction, *The New England Journal of Medicine* **298**, 289–295.

[6] Arthur, S.P. (1979). Skew/Stretched Distributions and the t-Statistic, Unpublished Ph.D. Dissertation, Department of Statistics, Princeton University, Princeton.

[7] Basford, K.E. & Tukey, J.W. (1999). *Graphical Analysis of Multiresponse Data*. Chapman & Hall, Boca Raton.

[8] Beaton, A.E. & Tukey, J.W. (1974c). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics* **16**, 147–185.

[9] Benjamini, Y. & Braun, H.I. (2002). John W. Tukey's contributions to multiple comparisons, *The Annals of Statistics* **30**, 1576–1594.

[10] Blackman, R.B. & Tukey, J.W. (1959b). *The Measurement of Power Spectra from the Point of View of Communications Engineering*. Dover, New York. (Table of contents, preface, and glossary reprinted in: *The Collected Works of John W. Tukey*, *Volume I: Time series, 1949–1964*, D.R. Brillinger, ed. Wadsworth Advanced Books & Software, Monterey, pp. 257–277.)

[11] Boas, R.P. Jr. & Tukey, J.W. (1938a). A note on linear functionals, *Bulletin of the American Mathematical Society* **44**, 523–528. (Correction: 1940, **46**, 566.)

[12] Bode, H.W. (1947). Letter to Samuel S. Wilks.

[13] Brillinger, D.R. (2002). John Wilder Tukey (1915–2000), *Notices of the AMS*, February 2002, 193–201.

[14] Brillinger, D.R., Fernholz, L. & Morgenthaler, S. (1997). *The Practice of Data Analysis: Essays in Honor of John W. Tukey*. Princeton University Press, Princeton.

[15] Brillinger, D.R., Jones, L.V. & Tukey, J.W. (1978g). *The Management of Weather Resources II: The Role of Statistics in Weather Resources Management*. United States Government Printing Office, Washington.

[16] Bunker, J.P., Forrest, W.H., Jr., Mosteller, F. & Vandam, L.D. (eds.) (1969). *The National Halothane Study*. Report of the Subcommittee on the National Halothane Study of the Committee on Anesthesia, Division of Medical Sciences, National Academy of Sciences-National Research Council, Government Printing Office, Washington.

[17] Cochran, W.G., Mosteller, F. & Tukey, J.W. (1953b). Statistical problems of the Kinsey report, *Journal of the American Statistical Association* **48**, 673–716.

[18] Cochran, W.G., Mosteller, F. & Tukey, J.W. (1954a). Principles of sampling, *Journal of the American Statistical Association* **49**, 13–35.

[19] Cochran, W.G., Mosteller, F. & Tukey, J.W. (1954c). *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male*. American Statistical Association, Washington.

[20] Cooley, J.W. & Tukey, J.W. (1965a). An algorithm for the machine calculation of a complex Fourier series, *Mathematics of Computation*, **19**, 297–301. Reprinted in *The Collected Works of John W. Tukey*, *Volume II: Time Series, 1965–1984*, D.R. Brillinger, ed. Wadsworth, Monterey, 1985, pp. 651–658.

[21] Cornfield, J. & Tukey, J.W. (1956f). Average values of mean squares in factorials, *The Annals of Mathematical Statistics* **27**, 907–949. Reprinted in *The Collected Works of John W. Tukey*, *Volume VII: Factorial & Anova, 1949–1962*, D.R. Cox, ed. Wadsworth Advanced Books & Software, Pacific Grove, 1992, pp. 179–239.

[22] Cox, J.L., Heyse, J.F. & Tukey, J.W. (2000). Efficacy estimates from parasite count data that include zero counts, *Experimental Parasitology* **96**, 1–8.

[23] Efron, B. (1979). Bootstrap methods: another look at the Jackknife, *Annals of Statistics* **5**, 1–26.

[24] Efron, B. (1982). *The Bootstrap, the Jackknife, and Other Resampling Plans*, Volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.

[25] Ericksen, E.P., Kadane, J.B. & Tukey, J.W. (1989x). Adjusting the 1980 census of population and housing, *Journal of the American Statistical Association* **84**, 927–944.

[26] Fernholz, L.T. & Morgenthaler, S. (2000). A conversation with Elizabeth and John Tukey, *Statistical Science* **15**, 79–94.

[27] Fernholz, L.T., Morgenthaler, S. & Tukey, J.W. (2004). An outlier nomination method based on the multihalver, *Journal of Statistical Planning and Inference* 122, 125–139.

[28] Freeman, M.F. & Tukey, J.W. (1950h). Transformations related to the angular and the square root, *Annals of Mathematical Statistics*, **21**, 607–611. Reprinted in *The Collected Works of John W. Tukey*, *Volume VI: More Mathematical, 1938–1984*, C.L. Mallows, ed. Wadsworth, Monterey, 1990, pp. 149–155.

[29] Friedman, J.H. & Tukey, J.W. (1974a). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers* **C-23**, 881–890. Reprinted in: *The Collected Works of John W. Tukey*, *Volume V: Graphics, 1965–1985*, W.S. Cleveland, ed. Wadsworth Advanced Books & Software, Pacific Grove, 1988, pp. 149–170.

[30] Gans, D.J. (1988). Trimmed and winsorized means, tests for, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz, N.L. Johnson & C. Read, eds. Wiley, New York, pp. 346–348.

[31] Goodall, C.R., Kafadar, K. & Tukey, J.W. (1998). Computing and using rural versus urban measures in statistical applications, *The American Statistician* **52**, 101–111.

[32] Goodman, A. (2000). `http://stat.bell-labs.com/who/tukey`.

[33] Hastings, C., Mosteller, F., Tukey, J.W. & Winsor, C.P. (1947b). Low moments for small samples: a comparative study of order statistics, *Annals of Mathematical Statistics* **18**, 361–426.

[34] Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (1983b). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

[35] Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (1985b). *Exploring Data Tables, Trends, and Shapes*. Wiley, New York.

[36] Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (1991h). *Fundamentals of Exploratory Analysis of Variance*. Wiley, New York.

[37] Jones, L.V. (1986). Philosophy and principles of data analysis, *The Collected Works of John W. Tukey*, *Volume III Philosophy and Principles of Data Analysis, 1949–1964*. Wadsworth & Brooks/Cole, Monterey.

[38] Mallows, C.L. (2003). John Tukey at Bell Labs, *Statistical Science* **18**, 332–335.

[39] McGill, R., Tukey, J.W. & Larsen, W.A. (1978a). Variations of box plots, *The American Statistician* **32**, 12–16. Reprinted in *The Collected Works of John W. Tukey*, *Volume V: Graphics, 1965–1985*, W.S. Cleveland, ed. Wadsworth Advanced Books & Software, Pacific Grove, 1988, pp. 63–77.

[40] McNeil, D.R. & Tukey, J.W. (1975). Higher-order diagnosis of two-way tables, illustrated on two sets of demographic empirical distributions, *Biometrics* **31**, 487–510.

[41] Miller, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Ed. Springer, New York.

[42] Moses, L.E. & Mosteller, F. (1989). Safety of anesthetics, in *Statistics: A Guide to the Unknown*, 3rd Ed., J.M. Tanur, F. Mosteller, W.H. Kruskal, E.L. Lehmann, R.F. Link, R.S. Pieters & G.R. Rising, eds. Wadsworth & Brooks/Cole, Belmont, CA, pp. 15–24.

[43] Mosteller, F. & Tukey, J.W. (1977b). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading.

[44] Pickle, L.W., Mungiole, M., Jones, G.K. & White, A.A. (1996). *Atlas of United States Mortality*. National Center for Health Statistics, Hyattsville.

[45] Rousseeuw, P.J., Ruts, I. & Tukey, J.W. (1999). The bagplot: A bivariate boxplot, *The American Statistician* **53**(4), 382–387.

[46] Scheffé, H. & Tukey, J.W. (1944a). A formula for sample sizes for population tolerance limits, *Annals of Mathematical Statistics* **15**, 217.

[47] Scheffé, H. & Tukey, J.W. (1945a). Non-parametric estimation, I: Validation of order statistics, *Annals of Mathematical Statistics* **16**, 187–192.

[48] Scott, R.C. (1988). Smear and sweep, in *Encyclopedia of Statistical Sciences*, Vol. 8, S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, NY, pp. 515–517.

[49] Spitzer, L. & Tukey, J.W. (1949c). Interstellar polarization, galactic magnetic fields, and ferromagnetism, *Science* **109**, 461–462.

[50] Spitzer, L. & Tukey, J.W. (1951c). A theory of interstellar polarization, *Astrophysical Journal* **114**, 187–205.

[51] Stone, A. & Tukey, J.W. (1942b). Generalized "sandwich" theorems, *Duke Mathematical Journal* **9**, 356–259. Reprinted in *The Collected Works of John W. Tukey*, *Volume VI: More Mathematical, 1938–1984*, C.L. Mallows, ed. Wadsworth, Monterey, 1990, pp. 11–13.

[52] Tukey, J.W. (1939b). Denumerability in Topology, Ph.D. Thesis, Princeton University, Princeton, (Pam 4572).

[53] Tukey, J.W. (1940a). Convergence and uniformity in topology, *Annals of Mathematical Studies*, Number 2. Princeton University Press, Princeton.

[54] Tukey, J.W. (1948b). Approximate weights, *Annals of Mathematical Statistics* **19**, 91–92.

[55] Tukey, J.W. (1948c). Some elementary problems of importance to small sample practice, *Human Biology* **20**, 205–214.

[56] Tukey, J.W. (1949f). Comparing individual means in the analysis of variance, *Biometrics* **5**, 99–114.

[57] Tukey, J.W. (1949h). One degree of freedom for non-additivity, *Biometrics* **5**, 232–242. Reprinted in *The Collected Works of John W. Tukey*, *Volume VII: Factorial & Anova, 1949–1962*, D.R. Cox, ed. Wadsworth Advanced Books & Software, Pacific Grove, 1992, pp. 1–13.

[58] Tukey, J.W. (1951a). Components in regression, *Biometrics* **7**, 33–49.

[59] Tukey, J.W. (1956d). Variances of variance components: I. Balanced designs, *Annals of Mathematical Statistics* **27**, 722–736. Reprinted in *The Collected Works of John W. Tukey*, *Volume VII: Factorial & Anova, 1949–1962*, D.R. Cox, ed. Wadsworth Advanced Books & Software, Pacific Grove, 1992, pp. 157–178.

[60] Tukey, J.W. (1957a). Variances of variance components: II. The unbalanced single classification, *Annals of Mathematical Statistics* **28**, 43–56. Reprinted in *The Collected Works of John W. Tukey*, *Volume VII: Factorial & Anova, 1949–1962*, D.R. Cox, ed. Wadsworth Advanced Books & Software, Pacific Grove, 1992, pp. 241–260.

[61] Tukey, J.W. (1957b). Variances of variance components: III. Third moments in a balanced single classification, *Annals of Mathematical Statistics* **28**, 378–384.

[62] Tukey, J.W. (1957c). On the comparative anatomy of transformations, *Annals of Mathematical Statistics* **28**, 602–632. Reprinted in *The Collected Works of John W. Tukey*, *Volume VI: More Mathematical, 1938–1984*, C.L. Mallows, ed. Wadsworth, Monterey, 1990, pp. 167–209.

[63] Tukey, J.W. (1958g). "Bias and confidence in not-quite large samples [abstract]," *Annals of Mathematical Statistics* **29**, 614. Reprinted in *The Collected Works of John W. Tukey*, *Volume VI: More Mathematical, 1938–1984*, C.L. Mallows, ed. Wadsworth, Monterey, 1990, p. 391.

[64] Tukey, J.W. (1959d). Discussion of the papers by Messrs. Satterthwaite and Budne, *Technometrics* **1**, 166–174.

[65] Tukey, J.W. (1959g). Little pieces of mixed factorials (abstract), *Biometrics* **15**, 641–642.

[66] Tukey, J.W. (1960d). Discussion of Anscombe and Daniel papers, *Technometrics* **2**, 159–163.

[67] Tukey, J.W. (1960f). A survey of sampling from contaminated distributions, Chapter 39 in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, S.B. Churye, W. Hoeffding, W.C. Madow, H.B. Mann, eds. Stanford University Press, Stanford, pp. 448–485.

[68] Tukey, J.W. (1962a). The future of data analysis, *The Annals of Mathematical Statistics* **33**(1), 1–67. Reprinted in *The Collected Works of John W. Tukey*, *Volume IV: Philosophy and Principles of Data Analysis, 1949–1964*, L.V. Jones, ed. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, pp. 391–484.

[69] Tukey, J.W. (1973f). *Index to Statistics and Probability: Citation Index*, *Volume 2 of the Information Access Series*. The R&D Press, Los Altos.

[70] Tukey, J.W. (1977a). *Exploratory Data Analysis*. Addison-Wesley, Reading.

[71] Tukey, J.W. (1977d). Some thoughts on clinical trials, especially problems of multiplicity, *Science* **198**, 679–684. Reprinted in *Evaluation Studies Review Annual*, T.D. Cook & Associates, eds. **3**, Sage: Beverly Hills, 1978, pp. 327–332.

[72] Tukey, J.W. (1979e). Methodology, and the statistician's responsibility for BOTH accuracy AND relevance, *Journal of the American Statistical Association* **74**, 786–793.

[73] Tukey, J.W. (1979f). Discussion on 'Nonparametrics statistical data modeling' by E. Parzen, *Journal of the American Statistical Association* **74**, 121–122. Reprinted in *The Collected Works of John W. Tukey*, *Volume IV: Philosophy and Principles of Data Analysis, 1949–1964*, L.V. Jones, ed. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, pp. 805–809.

[74] Tukey, J.W. (1979g). Statistical mapping: What should not be plotted, *Proceedings of the 1976 Workshop on Automated Cartography*, DHEW Publication No. (PHS) 79–1254, 18–26. Reprinted in *The Collected Works of John W. Tukey*, *Volume V: Graphics, 1965–1985*, W.S. Cleveland, ed. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, 1988, pp. 109–121.

[75] Tukey, J.W. (1980a). We need both exploratory *and* confirmatory, *The American Statistician* **34**, 23–25. Reprinted in *The Collected Works of John W. Tukey*, *Volume IV: Philosophy and Principles of Data Analysis, 1949–1964*, L.V. Jones, ed. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, pp. 811–817.

[76] Tukey, J.W. (1985d). Comment on: Estimating the population in a census year: 1980 and Beyond (E.P. Ericksen, J.B. Kadane), *Journal of the American Statistical Association* **80**, 127–128.

[77] Tukey, J.W. (1991o). Use of many covariates in clinical trials, *International Statistical Review* **59**(2), 123–137.

[78] Tukey, J.W. (1992r). Souvenir sheets for 'Seventeen points relevant to multiplicity in clinical trials' Unpublished manuscript (distributed at the Merck-Temple conference, 13 November 1992).

[79] Tukey, J.W. (1993i). Tightening the clinical trial, *Controlled Clinical Trials*, **14**, 266–285.

[80] Tukey, J.W. (1993b). The Problem of Multiple Comparisons, 1953 unpublished manuscript, printed in *The Collected Works of John W. Tukey*, *Volume VIII: Multiple Comparisons, 1948–1983*, H.I. Braun, ed. Chapman & Hall, New York, 1993, pp. 1–300.

[81] Tukey, J.W. (1994). Foreword to the multiple comparisons volume, in *The Collected Works of John W. Tukey*, *Volume VIII: Multiple Comparisons, 1948–1983*, H.I. Braun, ed. Chapman & Hall, New York, 1993, pp. liii–liv.

[82] Tukey, J.W., Ciminera, J.L. & Heyse, J.F. (1985p). Testing the statistical certainty of a response to increasing doses of a drug, *Biometrics* **41**, 295–301.

Further Reading

Dolby, J.L. & Tukey, J.W. (1973e). *The Statistics CumIndex*. The R&D Press, Los Altos.

Hoaglin, D.C. (1983). Letter values: a set of selected order statistics, Chapter 2 in *Understanding Robust and Exploratory Data Analysis*, D.C. Hoaglin, F. Mosteller & J.W. Tukey, eds. Wiley, New York.

Mason, T.J., McKay, F.W., Hoover, R., Blot, W.J. & Fraumeni, J.F., Jr. (1975). *Atlas of Cancer Mortality for U.S. Counties: 1950–1969*, DHEW Publication No. (NIH)75–780, U.S. Government Printing Office, Washington.

Quenouille, M.H. (1956). Notes on bias in estimation, *Biometrika* **43**, 353–360.

Ross, I.C. & Tukey, J.W. (1973g). *Index to Statistics and Probability: Locations and Authors*, *Volume 5 of the Information Access Series*. The R&D Press, Los Altos.

Tukey, J.W. (1959a). A quick, compact two-sample test to Duckworth's specifications, *Technometrics* **1**, 31–48.

Tukey, J.W. (1962c). Keeping research in contact with literature: citation indices and beyond, *Journal of Chemical Documentations* **2**, 34–37.

Tukey, J.W. (1963f). A citation index for statistics and probability *Bulletin of the International Statistical Institute* **40**, 747–756.

Tukey, J.W. (1963g). A tagging system for journal articles and other citable items: a status report, Annual Report for 1963 under National Science Foundation Grant NSF-GN-297 (from the Office of Science Information Services).

KAREN KAFADAR

Tumor Growth

Since the control of tumor growth is the purpose of cancer treatment, it is natural to ask: What are the characteristics of that growth which can be measured in a quantitative fashion for prognostic, diagnostic or explanatory purposes? Loosely, there are three different types of measurements that might be made to describe the growth of a tumor. The simplest measure of total growth is a series of sequential size measurements to produce an overall growth curve. At the cellular level, measurements can be made of proliferative markers related to the growth of the tumor and, possibly, to the mechanisms controlling growth. Finally, at the molecular level, measurements are made of genes and gene products controlling the proliferation of a tumor. Each of these measurements has limitations, but can be of great interest.

Tumor Growth Curves

Of the different types of measurements, that of the total tumor growth is undoubtedly the most important, but often unavailable for analysis in the clinical setting. Indeed, if the future growth of a tumor could be predicted then it might be possible to develop individualized treatments for patients depending on the specific properties of their tumors. In practice, comparing the growth patterns of tumors induced in animals following treatment is a standard method for screening potential anti-cancer agents prior to clinical use and for the study of carcinogenesis (*see* **Animal Screening Systems**; **Tumor Incidence Experiments**).

There are several reasons why total tumor growth measurements may not be possible. Normally, unless the patient chooses otherwise, treatment begins as soon as the tumor is discovered. While retrospective studies of earlier diagnostic procedures, e.g. mammograms, can sometimes be done with detection of the tumor at an earlier size [9], it is more likely that only bounds on the rate of growth can be found. Even in animal studies, the growth rate of tumors can usually be followed only for short periods of time because of ethical considerations or early death of the animal. For example, in murine tumors, commonly used in **radiation** biology studies, the tumors seldom can be measured for much more than four volume doublings so that a comparison of growth rates is usually determined among murine experimental tumors on the basis of growth between 4.0 mm mean geometric diameter and 12.5 mm [12]. (Methods do exist for obtaining estimates at smaller sizes, but these are less precise.) Inferential estimates on tumor growth rates can be made also on the basis of size at detection [2] and as done in some carcinogenesis studies [15].

While the actual form of the total growth curve is unknown, it is generally accepted that tumors, unlike cells in tissue culture, do not increase exponentially. (A detailed study of different models has been made in model systems by Marusic et al. [8].) The key question then is to understand and predict the changing nature of tumor growth. Generally, it is agreed [10, 11] that there are three cellular determinants of the population growth. First is the **cell cycle** time T_C, or the minimal time required for a cell to grow, duplicate its DNA, and divide. Second is the fraction of cells in a tumor actively dividing, known as the growth fraction, GF. Third is the rate of death or cell loss from the tumor. Since the tumor doubling time of an individual tumor is not accessible, a variety of methods have been devised to identify **surrogate** markers of proliferation in the population reflecting the overall growth of the tumor on the basis of measurements made at a single time, such as at biopsy or surgery.

Proliferation Markers

At the cellular level, the methods of analytical cytometry [3] are employed to identify subpopulations of tumor cells having specific markers associated with growth such as DNA content. Since cells go through a regular progression of DNA contents or phases, known as G_1 (post-mitotic and the state of the majority of cells in tissues), S (actively synthesizing DNA), G_2 (pre-mitotic), and M (mitosis), it is possible to measure the fractions of cells in a population in any given state of the division cycle. If we assume that the population is in a steady state where the fraction of cells in each phase is constant, from **branching process** theory [6] we assert that the phase fractions are proportional to the time spent in each phase. Thus, one

can, for example, compare the fractions of cells in the S-phase in two tumors and argue that the tumor with the greater fraction synthesizing DNA will be the faster growing. Unfortunately, these methods only give the relative fractions in each phase and two tumors with widely differing cycle times might appear identical [13]. There is an extensive literature on the validity of such markers as **prognostic factors** [3, 4].

An alternative approach to obtaining dynamic markers of proliferation, including specific times for each tumor, is based on the fact that certain halogenated thymidine analogs, e.g. bromodeoxyuridine, which label cells exclusively during DNA synthesis, can be detected by monoclonal antibodies. It is thus possible to identify a cell's bivariate analysis of DNA content and label. The duration of the S-phase, T_S, can then be estimated from the change in DNA content of the labeled cells [1]. Using the phase fractions and the known T_S, an estimate of the doubling time, known as the potential doubling time or T_{pot}, can be computed [14]. T_{pot}, introduced by Steel [11], is related to T_C by $T_{pot} = \ln(1 + GF)T_C / \ln 2$ and is considered to be the doubling time that would occur without the presence of cell loss. Usually the GF cannot be measured so that T_C cannot be determined, but T_{pot} may provide a useful dynamic estimate of the state of a tumor. As such T_{pot} has formed the basis for a series of prospective trials for assigning patients to altered radiotherapy treatments [12].

Other proliferation markers, such as Ki67 and PCNA, are commonly studied and new ones [5] are continually being developed. The advantage of these markers is that they can be determined in individual cells so that population averages may be determined, but in many cases, such as DNA content, these markers are removed from the molecular machinery controlling cell proliferation and may not be closely associated with long-term tumor growth.

Molecular Markers

Most recently, at the molecular level, a series of methods are being employed to determine the growth status on the basis of the occurrence of aberrant gene products. In this methodology, specific genes or gene products known to control cell regulation are measured in cells obtained from tumors. Unfortunately, the tools are largely qualitative, leaving one with a series of statements such as p53 is overexpressed or BCL is upregulated in a tumor without it being possible to describe the fraction of cells actively involved in this process. The advantage of this methodology is that it provides a systematic approach for gene therapy to suppress specific genes and possibly control the tumor growth. For a current overview of these approaches, see [7].

References

[1] Begg, A.C., McNally, N.J. et al. (1985). A method to measure the duration of DNA synthesis and potential doubling time from a single sample, *Cytometry* **6**, 620–626.

[2] Brown, B.W., Atkinson, E.N. et al. (1984). Estimation of tumor growth rate from distribution of tumor size at detection, *Journal of the National Cancer Institute* **72**, 31–38.

[3] Eudey, T.L. (1996). Statistical considerations in DNA flow cytometry, *Statistical Science* **11**, 320–334.

[4] Hedley, D.W., Shankey, T.B. et al. (1993). DNA Cytometry Consensus Conference, *Cytometry* **14**, 471. [Other papers in this issue of *Cytometry* are of great interest.]

[5] Hideyuki, K. & Steinbach, G. (1996). Histone H3 messenger RNA *in situ* hybridization correlates with *in vivo* bromodeoxyuridine labeling of S-phase cells in rat colonic epithelium, *Cancer Research* **6**, 434–437.

[6] Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley, Chichester.

[7] Kasten, M.B. (1997). Molecular biology of cancer: the cell cycle, in *Cancer: Principles and Practice of Oncology*, 5th Ed., Vol. 1, V.T. DeVita, S. Hellman & S.A. Rosenberg, eds. Lippincott-Raven, Philadelphia, pp. 121–134.

[8] Marusic, M., Bajzer, Z. et al. (1994). Analysis of growth of multicellular spheroids by mathematical models, *Cell Proliferation* **27**, 73–94.

[9] Spratt, J.A., von Fournier, D. et al. (1992). Mammographic assessment of human breast cancer growth and duration, *Cancer* **71**, 2013–2019.

[10] Steel, G.G. (1993). The growth rate of tumours, in *Basic Clinical Radiobiology*, G.G. Steel, ed. Edward Arnold, London, pp. 8–13.

[11] Steel, G.G. (1968). Cell loss from experimental tumours, *Cell and Tissue Kinetics* **1**, 193–207.

[12] Terry, N.H.A. (1996). Predictive assays for radiotherapy: the role of tumor proliferation (T_{pot}) measurements, *Onkologie* **19**, 322–327.

[13] Watson, J.V. (1992). *Flow Cytometry Data Analysis: Basic Concepts and Statistics*. Cambridge University Press, New York.

[14] White, R.A., Terry, N. et al. (1990). Improved method for computing potential doubling time from flow cytometric data, *Cytometry* **11**, 314–317.
[15] Yakolev, A.Y. & Tsodikov, A.D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

R. ALLEN WHITE

Tumor Incidence Experiments

The primary purpose of a typical tumorigenesis study is to evaluate the rate at which new tumors develop. Thus, the focal point in the analysis of such an experiment should be the tumor **incidence rate**. Consequently, the ideal statistical analysis treats tumor onset as the endpoint of interest, estimates the tumor incidence rates, and formally compares treatment groups with respect to these incidence rates.

In practice, several factors complicate what otherwise might be a straightforward failure-time (or survival) analysis. Among the complexities are: no uncensored onset times (*see* **Censored Data**), differential survival, and tumor lethality. Some of the proposed solutions, which introduce complications of their own, include: focusing on a different endpoint (*see* **Response Variable**), assuming a particular level of tumor lethality, assigning individual **causes of death**, requiring some animals to be sacrificed, specifying parametric models, and imposing functional restrictions. This article summarizes the advantages and disadvantages associated with these approaches, and gives a short description of each method.

Background

Tumor Incidence

Tumor incidence refers to the rate of tumor onset during a given time period, where onset is the earliest stage of tumor development at which the lesion could be detected microscopically. The time interval over which new tumors are accrued can vary in length from infinitesimal to an entire lifetime. If time is treated as continuous, the incidence rate is the **hazard rate** for tumor onset, which is an instantaneous (conditional) failure rate. Often, the timescale is partitioned into intervals, and a discrete incidence rate can be defined for each interval. Here, the incidence rate is the probability that tumor onset occurs in a particular interval, conditional on being alive and tumor-free upon entering that interval. If time is completely ignored, by treating the entire study as one large time interval, the incidence rate reduces to the lifetime probability of developing a tumor.

Study Design

Generally, studies involve both sexes of two rodent species, usually one strain of mice and one strain of rats. Exposures often begin when animals are 6–8 weeks of age and typically continue for two years, which corresponds to late middle age in these rodents, at which point all live animals are killed and necropsied. Some two-year studies incorporate interim sacrifices, which call for randomly selected subsets of animals to be killed and examined at intermediate times (*see* **Serial-sacrifice Experiments**). For each sex/species combination, the study usually includes one control group and several (e.g. three) exposed groups. For a single type of exposure, the treatment groups differ, in theory, only with regard to the dose level of exposure. After being stratified on weight, typically 50 animals are randomly assigned to each group, with perhaps ten more animals per group for each interim sacrifice planned.

Conditions and Assumptions

We focus on a single sex/species combination and we restrict attention to tumors at a single site. All tumors are assumed to be irreversible and, except for illustrating the ideal analysis, all tumors are assumed to be occult (unobservable in live animals). When multiple tumors occur at the site of interest, we do not adjust for multiplicity; we simply classify animals as tumor-free or tumor-bearing. In these situations, tumor onset refers to the onset of the first tumor. The time variables are measured from a common origin, such as birth, weaning, or study initiation. The term "sacrifice" refers to the intentional killing of a randomly selected animal, which acts as a random censoring mechanism with respect to the failure times, where failure might represent either tumor onset or death.

Stochastic Model

The simplest way to view the problem is in terms of a **competing risks** framework, in which each animal is subject to two competing risks: tumor onset or death without the tumor. Each animal is initially tumor-free, but eventually either develops the tumor or dies without the tumor. Thus, we can imagine a three-state stochastic model with one initial state (alive and tumor-free), one transient state (alive and tumor-bearing), and one absorbing state (dead)

(*see* **Fix–Neyman Process**). If followed long enough, each animal would travel from the initial state to the absorbing state, either directly or through the transient state. We observe which of the two paths is taken from birth to death, as well as the sojourn time for the overall journey, but not the sojourn times for the transitions to and from the intermediate state.

Notation

Let X denote the time to the first event, either tumor onset or tumor-free death, and let T denote the time to (natural) death. Sacrifices randomly censor both X and T. Define an indicator $Y(t)$ (*see* **Dummy Variables**) that is 1 if the tumor is present at time t and 0 otherwise. Let Y indicate whether the tumor is present or absent at death, either from natural causes or sacrifice.

Suppose that X and T are continuous variables. The transition intensities associated with the two competing risks can be expressed as event-specific hazard functions:

$$\lambda(t) = \lim_{\varepsilon \to 0} \Pr(t \le X < t + \varepsilon, Y(X) = 1 | X \ge t)/\varepsilon, \tag{1}$$

$$\beta(t) = \lim_{\varepsilon \to 0} \Pr(t \le X < t + \varepsilon, Y(X) = 0 | X \ge t)/\varepsilon, \tag{2}$$

where $\lambda(t)$ represents the tumor incidence rate and $\beta(t)$ is called the tumor-free death rate. Once an animal develops a tumor, say at time x, the risk of death at time t is

$$\alpha(t|x) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon | T \ge t, Y(t) = 1, X = x)/\varepsilon, \tag{3}$$

where $t \ge x > 0$. Some analyses involve a simplified version of this conditional death rate:

$$\alpha(t) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon | T \ge t, Y(t) = 1)/\varepsilon, \tag{4}$$

which can be regarded as an average rate of death among all animals having the tumor.

Define the following "pseudo" survivorship functions:

$$S_\lambda(t) = \exp\left[-\int_0^t \lambda(u)\,du\right],$$

$$S_\beta(t) = \exp\left[-\int_0^t \beta(u)\,du\right], \tag{5}$$

$$S_\alpha(t|x) = \exp\left[-\int_x^t \alpha(u|x)\,du\right],$$

which are used purely for notational convenience and have no particular interpretation.

Observed Data

The basic information that we observe for each animal is the time of death and an indicator of whether a tumor was found at the organ site of interest. In particular, we observe $\{T = t, Y(t) = y\}$ if an animal dies of natural causes, and $\{T > t, Y(t) = y\}$ if an animal is sacrificed ($y = 0, 1$). Note that among animals that die of natural causes, $Y(t)$ generally is not observable for $t < T$, except in special situations such as when tumors are palpable or visible in live animals. Occasionally, we have access to additional information, such as the tumor's role in causing death, but these extra data are not routinely available.

Suppose that there are $K \ge 1$ exposed groups and a single control group. Let N_k denote the number of animals randomized to the kth group ($k = 0, 1, \ldots, K$). Let $d_0 < d_1 < \cdots < d_K$ denote the ordered dose levels, where the control value is assumed to be $d_0 = 0$. Often, transforms of the dose levels are used, such as logarithms or equally spaced scores, in which case d_k represents the transformed value. Let J be the number of distinct natural death times across all groups, with $t_1 < t_2 < \cdots < t_J$ denoting their ordered values.

Likelihood Contributions

Sacrifice times randomly censor tumor onset times and natural death times. Thus, an animal sacrificed at time t and found to be tumor-free contributes only the information that the time to the first event exceeds $t : \Pr(X > t)$. In terms of the underlying transition intensities, the **likelihood** contribution from such an animal can be expressed as

$$S_\lambda(t) S_\beta(t). \tag{6}$$

Similarly, a tumor-free animal that dies of natural causes at time t contributes

$$\beta(t) S_\lambda(t) S_\beta(t). \tag{7}$$

The likelihood contributions from animals having an occult tumor are more complicated, as the components must be integrated over all possible (unobserved) tumor onset times. Thus, an animal sacrificed at time t and found to have an occult tumor contributes

$$\int_0^t \lambda(x)S_\lambda(x)S_\beta(x)S_\alpha(t|x)\,\mathrm{d}x, \tag{8}$$

whereas the corresponding contribution for a natural death with an occult tumor is

$$\int_0^t \lambda(x)S_\lambda(x)S_\beta(x)S_\alpha(t|x)\alpha(t|x)\,\mathrm{d}x. \tag{9}$$

The overall likelihood is the product of contributions such as those in (6)–(9).

If tumors were observable, contributions for tumor-bearing animals would simply be the integrands in (8) and (9), and maximization of the likelihood would be much easier. For occult tumors, the likelihood often is simplified by making assumptions about tumor lethality, cause of death, parametric models, or functional restrictions. With enough sacrifices, these additional assumptions are unnecessary and the likelihood can be maximized nonparametrically by treating certain failure times as discrete.

Goals

The main objective of a carcinogenicity study is to compare treatment groups with respect to tumor incidence and to provide summaries of the incidence rates within these groups. Thus, the analysis should focus on estimating the incidence rate in the kth group, say $\lambda_k(t)$, as well as testing the **null hypothesis** of equal incidence rates across groups

$$\mathrm{H}_\lambda : \lambda_0(t) = \lambda_1(t) = \cdots = \lambda_k(t), \tag{10}$$

against certain **alternative hypotheses** of interest. The usual choices are the general alternative that not all incidence rates are identical (i.e. group heterogeneity) and the specific alternative that incidence rates increase linearly with dose (i.e. a positive linear trend). Often, pairwise comparisons of each exposed group with the control group are performed. These comparisons can be viewed as special cases of testing H_λ against either of the above alternatives when there are only two groups ($K = 1$).

Lifetime Incidence Rates

The simplest incidence analysis focuses on lifetime incidence rates, comparing the group-specific proportions of animals that develop the tumor during the experiment. The lifetime incidence rate, say θ, can be expressed in terms of the age-specific incidence rate, $\lambda(t)$, and the tumor-free death rate, $\beta(t)$:

$$\begin{aligned}\theta &= \Pr(X \le T^*, Y(X) = 1)\\ &= \int_0^{T^*} \lambda(x)S_\lambda(x)S_\beta(x)\,\mathrm{d}x,\end{aligned} \tag{11}$$

where T^* is the time at which the study ends.

No Survival Adjustment

The standard nonparametric estimate of the lifetime incidence rate in the kth group, denoted by θ_k, is the overall group-specific proportion of tumor-bearing animals:

$$\hat{\theta}_k = \sum y_{ik}/N_k, \tag{12}$$

where y_{ik} is the observed value of Y for the ith animal in the kth group and the summation is over all i from 1 to N_k. In general, the null hypothesis of equal lifetime incidence rates across groups neither implies, nor is implied by, the hypothesis of equal age-specific incidence rates, H_λ. Usually, pairwise comparisons are based on **Fisher's exact test**; an overall assessment of group heterogeneity is based on an omnibus **chi-square test**; and dose-related changes are judged on the basis of the linear trend test of Cochran [11] and Armitage [3] (*see* **Trend Test for Counts and Proportions**). See Haseman [26] and Gart et al. [25] for a review of these tests and for some worked examples based on data from real carcinogenicity studies.

The primary advantage of this approach, in addition to using well known statistical methods, is that it does not rely on the usual simplifying assumptions. The analysis of lifetime tumor incidence rates does not make tumor lethality assumptions; occult tumors pose no problems; and there is no need for sacrifice data, cause-of-death information, parametric models, or functional restrictions. The major drawback is that the lack of any time adjustment can lead to biased inferences when mortality rates differ across groups.

Tests regarding lifetime incidence rates assume that all animals in the same group have the same

lifetime risk of developing a tumor, which clearly is violated if some die earlier than others. Even so, if mortality patterns are similar across groups, then these unadjusted tests are valid [24], though possibly less powerful than survival-adjusted tests [48] (*see* **Power**). Misleading results can be obtained, however, when mortality rates differ across groups. For example, if exposure causes animals to die early, then unadjusted tests can miss true carcinogens if treated animals die before having a chance to develop many tumors. In fact, tumorigenesis analyses should focus on $\lambda(t)$ rather than θ, and tests comparing lifetime incidence rates will tend to reject H_λ less often than desired when exposure is toxic because θ is a decreasing function of β. Again, true carcinogens can be missed when no survival adjustments are incorporated, and thus all analyses discussed subsequently make some adjustment for survival.

Adjusting for Survival

One simple way to adjust for survival when analyzing lifetime incidence rates is to modify the number of animals at risk by reducing the denominator of each rate. Rather than giving equal weight to all animals, less weight can be assigned to those dying early without a tumor. For example, Gart et al. [24] assign a weight of zero to any animal dying without the tumor at a time before the first death with the tumor, while assigning all other animals a weight of one. Extending this idea, Bailer & Portier [5] assign a weight of one to animals dying with the tumor and otherwise assign a weight proportional to a fixed power of the time on study, where the choice of the time exponent depends on the assumed shape of the tumor onset distribution. Bieler & Williams [7] propose a variance correction for the Bailer–Portier procedure. Often, the survival adjustment made by these methods is helpful; but, generally, an analysis focusing on age-specific incidence rates is preferable to one focusing on lifetime incidence rates, as inferences about the latter are not necessarily well correlated with inferences about the former.

Age-Specific Incidence Rates

There are various ways to adjust for survival, but no one method is appropriate in all situations. Several factors complicate the adjustment, and the proper analysis depends on what additional data can be obtained or which extra assumptions are plausible. For example, unless tumors are observable in live animals, all onset times are censored. Tumor lethality worsens the problem; if animals with the tumor die at a different rate than those without the tumor, then deaths from natural causes do not randomly censor tumor onset times. Thus, without direct observations on the onset times, survival adjustments usually are accomplished at the expense of additional data or assumptions.

Observable Tumors

Ideally, an incidence analysis should focus on the tumor onset times directly, which is only possible for observable tumors, such as visible skin tumors or palpable mammary tumors. When event times are observable, most analyses use standard nonparametric survival methods (see Kalbfleisch & Prentice [28]), which typically treat event times as discrete random variables. If X is discrete, then the tumor incidence rate at time t_j is

$$\lambda_j = \Pr(X = t_j, Y(X) = 1 | X \geq t_j). \tag{13}$$

Within the kth group, let O_{jk} be the (observed) number of animals developing a tumor at time t_j and let R_{jk} be the number of animals at risk of developing a tumor at time t_j:

$$\begin{aligned} O_{jk} &= \#\{X_{ik} = t_j, Y(X_{ik}) = 1\} \quad \text{and} \\ R_{jk} &= \#\{X_{ik} \geq t_j\}, \end{aligned} \tag{14}$$

where $\#\{e\}$ is the number of animals experiencing event e and X_{ik} is the value of X for the ith animal in the kth group. Under H_λ, the expected counts and variance–covariance terms associated with O_{jk} are

$$\begin{aligned} E_{jk} &= R_{jk}\left[\frac{O_{j+}}{R_{j+}}\right] \quad \text{and} \\ V_{jkm} &= \left[\frac{O_{j+}(R_{j+} - O_{j+})}{(R_{j+} - 1)}\right]\left[\frac{R_{jk}}{R_{j+}}\right] \\ &\quad \times \left[A_{km} - \frac{R_{jm}}{R_{j+}}\right], \end{aligned} \tag{15}$$

where the plus sign indicates the summation over all $K + 1$ groups and A_{km} is an indicator that equals 1 if $k = m$ and 0 otherwise. For each k (and m), create

group-specific summaries for the observed counts, expected counts, and variance–covariance terms:

$$O_k = \sum O_{jk}, \quad E_k = \sum E_{jk}, \quad V_{km} = \sum V_{jkm}, \tag{16}$$

where the summations are over all J times. Let $\mathbf{D}' = (d_1, \ldots, d_K)$ be the vector of dose levels, let $\mathbf{O}' = (O_1, \ldots, O_K)$ be the vector of observed counts, let $\mathbf{E}' = (E_1, \ldots, E_K)$ be the vector of expected counts, and let $\mathbf{V}$ be the matrix of variance terms $V_{km} (k = 1, \ldots, K; m = 1, \ldots, K)$. Note that these arrays do not include the summary terms from the control group ($k = 0$).

The nonparametric estimate of the incidence rate at time t_j in group k is

$$\hat{\lambda}_{jk} = O_{jk}/R_{jk}. \tag{17}$$

The usual test for heterogeneity of groups is based on the statistic:

$$\chi^2_{\mathrm{H}} = (\mathbf{O} - \mathbf{E})'\mathbf{V}^{-1}(\mathbf{O} - \mathbf{E}), \tag{18}$$

which follows asymptotically the **chi-square distribution** on K **degrees of freedom** (df) under H_λ. Often, the test for a dose-related trend in incidence rates is based on the **logrank** statistic

$$\chi^2_{\mathrm{T}} = [\mathbf{D}'(\mathbf{O} - \mathbf{E})]^2/(\mathbf{D}'\mathbf{V}\mathbf{D}), \tag{19}$$

which is also distributed asymptotically as chi-square under H_λ, but on a single df. Finally, a test for departures from linearity can be based on the statistic

$$\chi^2_{\mathrm{D}} = \chi^2_{\mathrm{H}} - \chi^2_{\mathrm{T}}, \tag{20}$$

which is distributed asymptotically as chi-square on $K - 1$ df under the null hypothesis that tumor incidence rates increase linearly with dose. For details on these tests and other related analyses, see Tarone [50], Tarone & Ware [51], Kalbfleisch & Prentice [28], and Gart et al. [25].

The only disadvantage of this approach is that most tumor types are unobservable in live animals. Thus, in the remaining sections we discuss methods that require additional data or assumptions to overcome the complications associated with occult tumors.

Rapidly Lethal Tumors

We assume that lethality refers to an intrinsic property of the tumor type and does not change on an individual animal basis, as opposed to the concept of cause of death, which allows each tumor's effect on death to vary across animals. The presence of a nonlethal tumor does not alter the risk of death, whereas a lethal tumor increases the risk of death. Conceivably, a tumor can be protective if it lessens the risk of death, although such instances are rare and are not considered here. The degree of tumor lethality is often characterized by how much a tumor's presence hastens death.

Suppose that a tumor type is rapidly lethal, which means that post-onset survival is short and thus time to death with the tumor is a good **surrogate** for time to tumor onset. If this rapid lethality assumption is accurate, then the analysis can focus on time to death with the tumor, which is an observable event. Therefore, as with the observable tumors discussed previously, ordinary **life table** analyses based on the estimates in (17) and the tests in (18)–(20) are reasonable when tumors are rapidly lethal.

The main problem with this approach is that few tumor types are instantly lethal, and life table analyses perform worse as tumor lethality decreases, especially as the mortality patterns across groups become more disparate. Tests comparing the rates of death with the tumor can yield biased conclusions about tumor incidence rates when the post-onset survival times are relatively long. The bias is in the opposite direction of the bias in the unadjusted analysis, and Gart et al. [24] characterize this effect as an over-adjustment for survival. For example, if exposure is toxic and causes treated animals to die sooner than controls, then nonlethal tumors will be discovered earlier in the exposed groups and a life table analysis might falsely conclude that exposure is carcinogenic even when the tumor incidence rates are identical across groups. Thus, a life table test applied to data on tumors that are not rapidly lethal rejects too often [34].

Nonlethal Tumors

Suppose that the tumor type of interest is strictly nonlethal, so that tumor presence has no effect on the risk of death. In this situation, the incidence rate is a one-to-one function of the **prevalence** rate, which in turn equals the response rate. Therefore, a standard

prevalence analysis of tumor response rates, which are directly observable, allows us to make inferences about the incidence rates of nonlethal tumors.

The tumor prevalence rate, say $\pi(t)$, is the expected proportion of animals having a tumor among those *alive* at time t, which can be expressed as

$$\pi(t) = \Pr(Y(t) = 1 | T > t). \tag{21}$$

In general, $\pi(t)$ is a function of the incidence rate, $\lambda(t)$, and the death rates, $\beta(t)$ and $\alpha(t|x)$:

$$[1 - \pi(t)]^{-1} = 1 + \int_0^t \lambda(x) \exp\left\{ \int_x^t [\lambda(u) + \beta(u) - \alpha(t|u)]\, \mathrm{d}u \right\} \mathrm{d}x. \tag{22}$$

When the tumor type is nonlethal, the conditional death rates for tumor-free and tumor-bearing animals are equal, $\alpha(t|x) \equiv \beta(t)$, and thus $\pi(t)$ reduces to a function of $\lambda(t)$ only:

$$\pi(t) = 1 - S_\lambda(t). \tag{23}$$

Similarly, the tumor response rate, say $p(t)$, is the expected proportion of animals having a tumor among those *dying* at time t, which can be expressed as

$$p(t) = \Pr(Y(t) = 1 | T = t). \tag{24}$$

In general, $p(t)$ is a slightly different function of the three underlying transition rates:

$$[1 - p(t)]^{-1} = 1 + \beta(t)^{-1} \int_0^t \lambda(x) \exp\left\{ \int_x^t [\lambda(u) + \beta(u) - \alpha(t|u)]\, \mathrm{d}u \right\} \alpha(t|x)\, \mathrm{d}x, \tag{25}$$

but when tumors are nonlethal, $p(t)$ also reduces to $1 - S_\lambda(t)$, and thus equals $\pi(t)$.

For stabilization purposes, a nonparametric analysis of the response rates usually groups the data into time intervals. Within the jth interval, say I_j, the response rate is

$$p_j = \Pr(Y(T) = 1 | T \in I_j). \tag{26}$$

The nonparametric estimate of the jth tumor response rate in the kth group is simply

$$\hat{p}_{jk} = \#\{T_{ik} \in I_j, Y(T_{ik}) = 1\} / \#\{T_{ik} \in I_j\}. \tag{27}$$

Depending on the intervals, these estimates can fluctuate greatly. Of course, there is no consensus on the best way in which to choose intervals. Some analyses specify fixed intervals, while others use data-dependent intervals that contain equal numbers of deaths. As the response rates are nondecreasing when tumors are nonlethal and irreversible, Hoel & Walburg [27] apply the pool-adjacent-violators algorithm [4] to estimate $p(t)$ under a monotonicity constraint. The resulting **isotonic regression** estimate of $p(t)$ is constant over data-dependent intervals, with step heights of the form given in (27).

Hoel & Walburg [27] propose a survival-adjusted prevalence analysis that uses intervals to stratify on death times, compares observed and expected counts within time intervals, and applies **Mantel–Haenszel methods** [39] to combine results over the intervals. Regardless of how time intervals are determined, the test statistics still are given by (18)–(20), except that now for group k the observed count (O_{jk}) is the number of deaths in I_j with a tumor and the number at risk (R_{jk}) is the total number of deaths in I_j.

One potential problem is that differential mortality can yield time intervals within which all (most) of the deaths come from the same group and thus these intervals make no (little) contribution to the test statistics. With this in mind, Dinse & Lagakos [22] model tumor response as a **logistic regression** on age and dose. Likelihood methods can be used to estimate and to compare prevalence rates. This regression approach avoids the arbitrariness of choosing time intervals at the expense of parameterizing the time term. Despite its parametric nature, however, the logistic analysis is fairly robust. For example, linear age and dose terms produce a test with operating characteristics that match or exceed those of the interval-based tests [14].

Similar to the problems that arise when a life table analysis is applied to nonlethal tumors, the use of a prevalence analysis when the tumors are lethal also can produce misleading results, except that the bias is in the opposite direction. Once again, suppose that exposure is toxic and causes the treated animals to die sooner than the controls. As the tumor response rate, $p(t)$, is a decreasing function of the tumor-free death rate, $\beta(t)$, a prevalence test oriented toward detecting an increase in $p(t)$ will not reject as often as a test appropriate for assessing H_λ should reject, unless

tumors are nonlethal. Therefore, a prevalence test applied to data on lethal tumors will not reject often enough [34], which could result in a true carcinogen being missed.

Fixed Intermediate Lethality

In general, it is unlikely that death always follows tumor onset immediately or that death is completely unaffected by tumor onset. More realistically, all tumors of a given type might have a fixed but intermediate level of lethality, which could even vary with age. Formally, lethality usually is defined as some function of the conditional death rates, $\beta(t)$ and $\alpha(t|x)$. If this lethality function is known, then $\lambda(t)$ is identifiable, and thus appropriate incidence estimators and tests can be derived [40].

Typically, tumor lethality is unknown. Often, both life table and prevalence methods are applied in hope that the two procedures will give similar results [26]. While this approach has some intuitive appeal, the two tests do not always give the same results. Lagakos & Louis [35] show that the significance levels from life table and prevalence analyses need not bracket the P value from a test based on an intermediate lethality. Alternatively, Lagakos & Louis suggest a sensitivity analysis, which evaluates the data for various assumed lethalities and illustrates the range of possible inferences. A more definitive solution, however, requires additional data or assumptions.

Cause of Death

Rather than assuming that lethality is an intrinsic property of the tumor type itself, suppose that an individual context of observation [43, 44] can be specified for each tumor discovered, which characterizes that tumor's effect on the risk of death. Within this framework, tumors that do not alter longevity and are observed merely as the result of a death from an unrelated cause are classified as *incidental*. Conversely, tumors that affect mortality either by directly causing death, or by indirectly increasing the risk of death from other causes, are classified as *fatal*. This information on context of observation commonly is referred to as data on cause of death.

The availability and reliability of cause-of-death data are subject to debate. Many pathologists will not make these assessments and several investigations have shown that cause-of-death data, when available, can be inaccurate [33, 36]. Nevertheless, suppose that accurate cause-of-death data are provided, at least for a subset of the animals. Note that any tumor found in a randomly sacrificed animal is observed in an incidental context.

Information on cause of death allows us to identify the tumor onset distribution and perform a survival-adjusted incidence analysis without lethality assumptions, sacrifices, parametric models, or functional restrictions. Let C denote cause of death. We extend the three-state model to four states by replacing the single absorbing state with two absorbing states: death from the tumor ($C = 1$) and death from other causes ($C = 0$). Deaths without the tumor must be due to other causes; thus, the tumor-free death rate is

$$\beta(t) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon, C = 0 | T \ge t, Y(t) = 0)/\varepsilon. \tag{28}$$

The death rates with fatal and incidental tumors among the tumor-bearing animals are

$$\gamma(t|x) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon, C = 1 | T \ge t, Y(t) = 1, X = x)/\varepsilon, \tag{29}$$

$$\delta(t|x) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon, C = 0 | T \ge t, Y(t) = 1, X = x)/\varepsilon. \tag{30}$$

Note that the death rate for tumor-bearing animals is the sum of two cause-specific death rates, $\alpha(t|x) = \gamma(t|x) + \delta(t|x)$. Define a marginal rate of death from the tumor among all of the tumor-bearing animals:

$$\gamma(t) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon, C = 1 | T \ge t, Y(t) = 1)/\varepsilon. \tag{31}$$

Finally, as an incidental tumor does not affect mortality, the associated death rate is the same as the death rate in the absence of the tumor:

$$\delta(t|x) \equiv \beta(t), \quad \text{for all } t \ge x > 0. \tag{32}$$

Kodell & Nelson [31] propose a parametric estimator for the incidence rate by assuming **Weibull** models for all of the transition rates. Later, within a nonparametric framework, Kodell et al. [32] derive an estimator for the tumor onset distribution under the assumption that tumor prevalence is a nondecreasing

function of age. Dinse & Lagakos [21] and Turnbull & Mitchell [53] generalize this analysis to allow nonmonotonic prevalences (see also [16]).

Peto [43] and Peto et al. [44] describe a testing procedure that simultaneously compares groups with respect to tumor prevalence and tumor mortality. This approach applies a prevalence analysis to the subset of animals dying from other causes, applies a life table analysis to all animals, and combines these two components to obtain an overall assessment of group differences. The life table component treats deaths from the tumor as uncensored events and deaths from other causes as censored events. Regression extensions of this combined analysis have been proposed [22, 23], based on a logistic model for the prevalence part and a **proportional hazards** model [12] for the life table part (*see* **Cox Regression Model**). Lagakos & Louis [35] derive the Peto test as a **partial likelihood** score test under linked **proportional-odds** and proportional hazards models for the prevalence and mortality rates, respectively (see also [9]).

Pathologists are reluctant to label every tumor as definitely incidental or definitely fatal, and classification errors can produce biases [34, 47]. Peto et al. [44] suggest adding categories for probably fatal and probably incidental. In practice, however, the analysis often combines categories to form a new dichotomy and then proceeds as usual. Lagakos [34] considers a single intermediate category to allow an unknown cause of death and proposes two strategies: (i) relabel the unknowns as incidentals and fatals according to the proportions observed in those categories; or (ii) relabel all unknowns as incidentals in the lifetable component and as fatals in the prevalence component. Alternatively, analyses can formally account for uncertain contexts of observation either by having a pathologist assign a probability to each category [47] or by estimating these probabilities from the data [2, 15, 16, 30].

There are several difficulties with cause-of-death analyses. Most studies do not provide information on cause of death, and even when these data are available, they are often unreliable. Furthermore, even if data on cause of death are available and accurate, most tests are oriented toward prevalence and mortality rather than incidence, and H_λ does not always correspond to $\pi_0(t) = \pi_1(t) = \cdots = \pi_K(t)$ and $\gamma_0(t) = \gamma_1(t) = \cdots = \gamma_K(t)$. In fact, the prevalence portion of a Peto type test focuses on $p(t)$ rather than on $\pi(t)$, and thus its validity depends on whether tumor patterns in animals that die from other causes are representative of those in live animals, an assumption that is not always true [1, 36]. Finally, given that representativeness holds, McKnight & Wahrendorf [41] discuss situations in which these cause-of-death tests should provide appropriate inferences about tumor incidence.

Random Sacrifices

The intentional killing and examination of a random sample of healthy animals gives a cross-sectional view of the tumorigenic process. The proportion of animals having the tumor among those sacrificed at time t provides an unbiased estimate of $\pi(t)$. Therefore, together with observations on time to death and tumor response, sacrifice data permit an analysis of tumor incidence. The advantage of having sacrifice data is that lethality assumptions, cause-of-death data, parametric models, and functional restrictions are not necessary. The survival adjustment improves with the number of sacrifice times and the number of animals killed at each sacrifice time. The main disadvantages of multiple sacrifices are the extra expense and complexity associated with a large experiment.

McKnight & Crowley [40] state that without knowledge of how tumor incidence affects the risk of death, sacrifice data are needed to identify $\lambda(t)$. They express the tumor incidence rate in terms of the prevalence rate, its derivative, and two death rates:

$$\lambda(t) = [\pi'(t) + g(t) - \pi(t)h(t)]/[1 - \pi(t)], \quad (33)$$

where $\pi'(t)$ is the derivative of $\pi(t)$, $g(t)$ is the rate of death with the tumor

$$g(t) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon, Y(T) = 1 | T \ge t)/\varepsilon, \quad (34)$$

and $h(t)$ is the overall death rate:

$$h(t) = \lim_{\varepsilon \to 0} \Pr(t \le T < t + \varepsilon | T \ge t)/\varepsilon. \quad (35)$$

Lacking further structure, however, the prevalence rate is estimable only at the sacrifice times and thus the "resolution" of a nonparametric incidence analysis is limited by the number of sacrifice times. For example, a nonparametric analysis is generally based on intervals with endpoints at the sacrifice times. The amount of age adjustment increases with the number of sacrifice times and most experiments have few, if

any, in addition to the terminal sacrifice. Even carcinogenicity studies with one or two interim sacrifices would yield very coarse nonparametric estimates of the derivative of the prevalence rate, and thus poor estimates of the incidence rate.

McKnight & Crowley [40] describe the advantages of multiple sacrifice times, discuss identifiability, and give nonparametric estimators and tests for the incidence rates. Dewanji & Kalbfleisch [13] propose an iterative maximum likelihood analysis, while Malani & Van Ryzin [38] suggest a closed form solution that coincides with the maximum likelihood analysis when the data are well behaved, but otherwise can produce negative incidence estimates. Williams & Portier [54, 55] derive similar explicit nonparametric estimators, with and without restricting the incidence rates to be positive.

Many other authors have dealt with survival/sacrifice experiments, but most have assumed parametric models or have not focused on tumor incidence rates. Turnbull & Mitchell [52] consider the simultaneous analysis of several diseases, but they use **loglinear** and logistic models to investigate tumor prevalence and lethality rates (see also [6] and [42]). On the basis of a piecewise constant model for the transition rates, Borgan et al. [8] show that multiple sacrifices can greatly increase efficiency. In rare cases, such as in the ED_{01} study [10], an experiment is large enough and has a sufficient number of sacrifices to support a reasonable nonparametric analysis of tumor incidence rates, but this is clearly the exception rather than the rule. In practice, an analysis should not rely on routinely having numerous sacrifice times.

Parametric Models

Kalbfleisch et al. [29] discuss a fully parametric setting and suggest that a maximum likelihood analysis is complicated, but feasible, if the assumed models are identifiable. Borgan et al. [8] describe a special case in which the underlying intensities are piecewise constant. By reformulating in terms of more easily estimable quantities, Dinse [17] uses simple, flexible models to produce reasonable estimates of the incidence rate and a measure of tumor lethality. Without additional data, though, such as from multiple sacrifices, the modeling assumptions in a parametric analysis are untestable. Conclusions based on one model can differ greatly from those based on another. Thus, one must take care to select a biologically sensible model, or at least a general model that is fairly robust to misspecification.

Several recent analyses combine the benefits of parametric models and sacrifice data. That is, much information on tumor incidence can be gained through relatively few sacrifices as a result of the increased structure provided by a parametric model. For example, Portier [45] proposes a semiparametric analysis that assumes a **Weibull distribution** for the tumor onset times, which gives incidence rates of the form

$$\lambda(t) = bct^{c-1}, \tag{36}$$

but imposes no parametric model on the conditional death rates and requires only two sacrifice times (see also [16] and [46]).

Functional Restrictions

Rather than assuming particular distributions for the underlying random variables or modeling the components of the onset/death process, constraints can be placed on the way in which the components relate to each other. For example, Dinse [18] proposes a constant risk difference model, which assumes that the difference between the death rates for animals with and without the tumor in the kth group is constant over time:

$$\alpha_k(t|x) - \beta_k(t) \equiv \Delta_k. \tag{37}$$

Dinse [18] also considers a constant risk ratio model, which assumes that the ratio of these death rates is constant over time:

$$\alpha_k(t|x)/\beta_k(t) \equiv \rho_k. \tag{38}$$

Although multiple sacrifice times are useful, only one is needed to fit these models.

Under either restriction, the tumor incidence rates and tumor-free death rates can vary with each death time t_j; which, together with Δ_k or ρ_k, can yield as many as $2J + 1$ unknowns per group in this nonparametric setting. The number of unknowns can be reduced by modeling either $\lambda_k(t)$ or $\beta_k(t)$ as a function of time t. In fact, any of the quantities $\lambda_k(t)$, $\beta_k(t)$, Δ_k, or ρ_k can be modeled as functions of dose and other covariates. Lindsey & Ryan [37] describe a constant risk ratio model with piecewise constant tumor incidence rates and tumor-free death

rates. Ryan & Orav [49] propose a constant risk ratio model and suggest incorporating covariates such as tumor size and histology, although they focus on tumor prevalence rather than tumor incidence.

This approach enjoys several advantages, as it adjusts for differential mortality, allows for occult tumors, avoids lethality assumptions, and does not require information on cause of death. Only one sacrifice time is necessary, in theory, and certain functional constraints seem less restrictive than many parametric models. One downside is that the validity of any functional constraint can not be tested without additional information, such as data from multiple sacrifices. Thus, the appropriateness of such methods in general must be judged on the basis of simulations and large data sets, such as the ED_{01} study [10]. The estimates and the operating characteristics of tests derived under the constant risk difference model appear promising [18–20].

Discussion

Our understanding of the tumor onset/death process has evolved over time, as has our appreciation of which endpoints are important, what data must be collected, and how studies should be designed. If our goal is to analyze tumorigenesis, then we must focus on the time to tumor onset, which is characterized by the tumor incidence rate. In the vast majority of organ sites, however, tumors simply are not observable in live animals and thus the experiment provides no direct observations on the primary endpoint of interest. Consequently, we must either focus on secondary endpoints, collect additional data, design studies differently, or make unverifiable assumptions.

If our goal is to estimate and to compare tumor incidence rates, an analysis oriented toward some other endpoint is unacceptable, unless a clear equivalence exists between tumor incidence and the other endpoint. The common alternatives, such as time-specific tumor prevalence, time to death with tumor, and time to death from tumor, suffer various shortcomings. Except in special situations, an analysis based on one of these other endpoints can produce biases with respect to inferences about tumor incidence. The best strategy, in general, is to focus on tumor incidence if at all possible.

In a perfect world, with no budgetary limits, the ideal study would have many interim sacrifices. Given enough sacrifice data, the analysis could focus on tumor incidence without worrying about tumor lethality, cause of death, parametric models, or functional restrictions. Unfortunately, such studies are extremely rare, and thus analyses that rely on an abundance of sacrifices are not routinely applicable, regardless of how satisfying they are conceptually. In general, cost and other practical considerations demand an approach that can be applied with few sacrifice times.

Assumptions about tumor lethality provide the simplest survival-adjusted analysis of tumor incidence rates. In this case, however, simplicity comes at the expense of general applicability. Most researchers would agree that there are few organ sites for which all tumors are strictly nonlethal or all tumors are instantly lethal. Consequently, analyses based on either of these extreme lethality assumptions produce biased inferences with regard to tumor incidence when that lethality assumption is false. Techniques that rely on cause-of-death determinations, which are not commonly available, experience similar problems due to misclassification errors.

The last resort seems to be the use of parametric models or functional constraints. Without sacrifice data, or other extraordinary information such as reliable cause-of-death data, parametric assumptions typically are untestable. Methods that rely on unverifiable assumptions must be used with caution. The increased structure of a parametric model, however, provides many benefits and thus such models should be considered carefully.

In conclusion, perhaps the most promising approach is based on a combination of some sacrifice data and some type of formal structure. Although multiple sacrifice times are rare, most studies are terminated after a fixed period of time, at which point all of the remaining animals are killed. In many cases, the data from this terminal sacrifice can provide enough information about the overall tumorigenesis puzzle to allow an analysis that makes limited parametric assumptions or imposes some functional restrictions. The terminal sacrifice data permit the incidence function to be identified even when only a subset of the underlying transition intensities is modeled parametrically or constrained in some way. For example, assuming a constant difference between the death rates for animals with and without the tumor

is sufficient to provide an otherwise nonparametric analysis of the tumor incidence rates. This approach and other similar methods appear to be the most fruitful avenues for future research.

References

[1] Archer, L. & Ryan, L. (1989). On the role of cause-of-death data in the analysis of rodent tumorigenicity experiments, *Applied Statistics* **38**, 81–93.

[2] Archer, L. & Ryan, L. (1989). Accounting for misclassification in the cause-of-death test for carcinogenicity, *Journal of the American Statistical Association* **84**, 787–791.

[3] Armitage, P. (1955). Tests for linear trends in proportions and frequencies, *Biometrics* **11**, 375–386.

[4] Ayer, M., Brunk, H., Ewing, G., Reid, W. & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information, *Annals of Mathematical Statistics* **26**, 641–647.

[5] Bailer, A. & Portier, C. (1988). Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples, *Biometrics* **44**, 417–431.

[6] Berlin, B., Brodsky, J. & Clifford, P. (1979). Testing disease dependence in survival experiments with serial sacrifice, *Journal of the American Statistical Association* **74**, 5–14.

[7] Bieler, G. & Williams, R. (1993). Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity, *Biometrics* **49**, 793–801.

[8] Borgan, Ø., Liestøl, K. & Ebbesen, P. (1984). Efficiencies of experimental designs for an illness–death model, *Biometrics* **40**, 627–638.

[9] Burnett, R., Krewski, D. & Bleuer, S. (1989). Efficiency robust score tests for rodent tumorigenicity experiments, *Biometrika* **76**, 317–324.

[10] Cairns, T. (1980). The ED_{01} study: introduction, objectives, and experimental design, *Journal of Environmental Pathology and Toxicology* **3**, 1–7.

[11] Cochran, W. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.

[12] Cox, D. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[13] Dewanji, A. & Kalbfleisch, J. (1986). Non-parametric methods for survival/sacrifice experiments, *Biometrics* **42**, 325–341.

[14] Dinse, G. (1985). Testing for a trend in tumor prevalence rates: I. Nonlethal tumors, *Biometrics* **41**, 751–770.

[15] Dinse, G. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data, *Journal of the American Statistical Association* **81**, 328–336.

[16] Dinse, G. (1988a). Estimating tumor incidence rates in animal carcinogenicity experiments, *Biometrics* **44**, 405–415.

[17] Dinse, G. (1988b). Simple parametric analysis of animal tumorigenicity data, *Journal of the American Statistical Association* **83**, 638–649.

[18] Dinse, G. (1991). Constant risk differences in the analysis of animal tumorigenicity data, *Biometrics* **47**, 681–700.

[19] Dinse, G. (1993). Evaluating constraints that allow survival-adjusted incidence analyses in single-sacrifice studies, *Biometrics* **49**, 399–407.

[20] Dinse, G. (1994). A comparison of tumor incidence analyses applicable in single-sacrifice animal experiments, *Statistics in Medicine* **13**, 689–708.

[21] Dinse, G. & Lagakos, S. (1982). Nonparametric estimation of lifetime and disease onset distributions from incomplete observations, *Biometrics* **38**, 921–932.

[22] Dinse, G. & Lagakos, S. (1983). Regression analysis of tumour prevalence data, *Applied Statistics* **32**, 236–248; corrigenda **33**, (1984) 79–80.

[23] Finkelstein, D. & Ryan, L. (1987). Estimating carcinogenic potency from a rodent tumorigenicity experiment, *Applied Statistics* **36**, 121–133.

[24] Gart, J., Chu, K. & Tarone, R. (1979). Statistical issues in interpretation of chronic bioassay tests for carcinogenicity, *Journal of the National Cancer Institute* **62**, 957–974.

[25] Gart, J., Krewski, D., Lee, P., Tarone, R. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research*, Vol. III: *The Design and Analysis of Long-term Animal Experiments*. IARC Scientific Publications No. 79. International Agency for Research on Cancer, Lyon.

[26] Haseman, J. (1984). Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies, *Environmental Health Perspectives* **58**, 385–392.

[27] Hoel, D. & Walburg, H. (1972). Statistical analysis of survival experiments, *Journal of the National Cancer Institute* **49**, 361–372.

[28] Kalbfleisch, J. & Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[29] Kalbfleisch, J., Krewski, D. & Van Ryzin, J. (1983). Dose–response models for time-to-response toxicity data, *Canadian Journal of Statistics* **11**, 25–49.

[30] Kodell, R. & Chen, J. (1987). Handling cause of death in equivocal cases using the EM algorithm, *Communications in Statistics – Theory and Methods* **16**, 2565–2585.

[31] Kodell, R. & Nelson, C. (1980). An illness–death model for the study of the carcinogenic process using survival/sacrifice data, *Biometrics* **36**, 267–277.

[32] Kodell, R., Shaw, G. & Johnson, A. (1982). Nonparametric joint estimators for disease resistance and survival functions in survival/sacrifice experiments, *Biometrics* **38**, 43–58.

[33] Kodell, R., Farmer, J., Gaylor, D. & Cameron, A. (1982). Influence of cause-of-death assignment on time-to-death analyses in animal carcinogenesis studies, *Journal of the National Cancer Institute* **69**, 659–664.

[34] Lagakos, S. (1982). An evaluation of some two-sample tests used to analyze animal carcinogenicity experiments, *Utilitas Mathematica* **21B**, 239–260.

[35] Lagakos, S. & Louis, T. (1988). Use of tumor lethality to interpret tumorigenicity experiments lacking cause-of-death data, *Applied Statistics* **37**, 169–179.

[36] Lagakos, S. & Ryan, L. (1985). On the representativeness assumption in prevalence tests of carcinogenicity, *Applied Statistics* **34**, 54–62.

[37] Lindsey, J. & Ryan, L. (1993). A three-state multiplicative model for rodent tumorigenicity experiments, *Applied Statistics* **42**, 283–300.

[38] Malani, H. & Van Ryzin, J. (1988). Comparison of two treatments in animal carcinogenicity experiments, *Journal of the American Statistical Association* **83**, 1171–1177.

[39] Mantel, N. & Haenszel, W. (1959). Statistical aspects of analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[40] McKnight, B. & Crowley, J. (1984). Tests for differences in tumor incidence based on animal carcinogenesis experiments, *Journal of the American Statistical Association* **79**, 639–648.

[41] McKnight, B. & Wahrendorf, J. (1992). Tumor incidence rate alternatives and the cause-of-death test for carcinogenicity, *Biometrika* **79**, 131–138.

[42] Mitchell, T. & Turnbull, B. (1979). Log-linear models in the analysis of disease prevalence data from survival/sacrifice experiments, *Biometrics* **35**, 221–234.

[43] Peto, R. (1974). Guidelines on the analysis of tumor rates and death rates in experimental animals (editorial), *British Journal of Cancer* **29**, 101–105.

[44] Peto, R., Pike, M., Day, N., Gray, R., Lee, P., Parish, S., Peto, J., Richards, S. & Wahrendorf, J. (1980). Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments, in *Long-term and Short-term Screening Assays for Carcinogens: a Critical Appraisal*. IARC Monographs, Annex to Supplement 2. International Agency for Research on Cancer, Lyon, pp. 311–426.

[45] Portier, C. (1986). Estimating the tumor onset distribution in animal carcinogenesis experiments, *Biometrika* **73**, 371–378.

[46] Portier, C. & Dinse, G. (1987). Semiparametric analysis of tumor incidence rates in survival/sacrifice experiments, *Biometrics* **43**, 107–114.

[47] Racine-Poon, A. & Hoel, D. (1984). Nonparametric estimation of the survival function when cause of death is uncertain, *Biometrics* **40**, 1151–1158.

[48] Ryan, L. (1985). Efficiency of age-adjusted tests in animal carcinogenicity experiments, *Biometrics* **41**, 525–531.

[49] Ryan, L. & Orav, E. (1988). On the use of covariates for rodent bioassay and screening experiments, *Biometrika* **75**, 631–637.

[50] Tarone, R. (1975). Tests for trend in life table analysis, *Biometrika* **62**, 679–682.

[51] Tarone, R. & Ware, J. (1977). On distribution-free tests for equality of survival distributions, *Biometrika* **64**, 156–160.

[52] Turnbull, B. & Mitchell, T. (1978). Exploratory analysis of disease prevalence data from survival/sacrifice experiments, *Biometrics* **34**, 555–570.

[53] Turnbull, B. & Mitchell, T. (1984). Nonparametric estimation of the distribution of time to onset for specific diseases in survival/sacrifice experiments, *Biometrics* **40**, 41–50.

[54] Williams, P. & Portier, C. (1992a). Analytic expressions for maximum likelihood estimators in a nonparametric model of tumor incidence and death, *Communications in Statistics – Theory and Methods* **21**, 711–732.

[55] Williams, P. & Portier, C. (1992b). Explicit solutions for constrained maximum likelihood estimators in survival/sacrifice experiments, *Biometrika* **79**, 717–729.

GREGG E. DINSE

Tumor Modeling

Introduction

There are various stages in the development of a particular cancer into a life-threatening disease. In the first of these, one or more mutations occur within an individual cell. The uncontrolled division of that cell leads to the growth of a (small) avascular tumor, which may then remain dormant unless and until it acquires its own blood system by the process of angiogenesis (*see* **Tumor Growth**). The resulting vascular tumor is relatively well-supplied by nutrients and may grow to a much larger size. Finally, malignant tumors are able to invade the surrounding tissue, leading to metastatic spread, with secondary tumors arising elsewhere in the host. Further mutations within the population of tumor cells, including drug resistance, may underpin these later stages of development.

Tumorigenesis, the process by which normal cells transform into cancer cells, is associated with the progressive loss of function of a range of regulatory genes, including repair genes that correct mutations and DNA damage before cell division and tumor suppressor genes that signal for cell-cycle arrest or induce programmed cell death (apoptosis) if substantial genetic damage is detected.

Once a solid tumor has developed and been detected, it may be categorized in many ways, such as: according to the tissue and cell-type of origin, the fractal dimension of its periphery [24] and whether it is benign or malignant.

A further classification is based on whether it possesses a blood supply: vascular tumors have a blood supply, whereas avascular ones do not. Diffusion controls the delivery of nutrients (e.g. oxygen and glucose) to, and the removal of waste products from, avascular tumors [13, 43]. The diameter to which avascular tumors grow is thus limited to several millimeters and they are relatively harmless. By contrast, vascular tumors are life-threatening for two reasons. Firstly, being connected to the host's blood supply, they have access to an almost limitless supply of nutrients. The consequent rapid growth of such tumors may impair the function of neighboring vital organs.

Second, tumor fragments that enter the vasculature may be transported to other parts of the body where they may establish secondary tumors (metastases) that further jeopardize the host. The switch from avascular to vascular growth is effected by angiogenesis [12]. During this process, the tumor cells secrete a range of diffusible chemicals (e.g. vascular endothelial growth factor and tumor necrosis factor-β), which are known collectively as angiogenic factors. The angiogenic factors stimulate neighboring blood vessels to proliferate and migrate towards the tumor, eventually furnishing it with a circulating blood supply so that vascular growth may commence.

Invasion of the surrounding tissue is another key feature of solid tumors: contact with the tissue stimulates the production of enzymes such as matrix metallo-proteases, which digest the tissue. This creates spaces into which the tumor cells may then migrate [40].

At a cellular level, the sequence of events that are needed to establish a well-developed, vascularized tumor may be associated with genetic mutations. For example, mutations in the tumor suppressor gene p53 have been linked with tumor angiogenesis and cell immortality, the latter being a hallmark of many tumors [37, 39].

Tumorigenesis

As stated above, disruption of a number of regulatory genes is necessary for tumorigenesis (the initiation of tumor growth), the number and type of genes required remaining open questions [19]. The earliest models of tumorigenesis were **stochastic** and developed to investigate the age-specific incidence rates for certain adult cancers [3]. These **multistage carcinogenesis models** were extended to distinguish between inherited and spontaneous cancers [22, 29] and adapted to investigate the effects of apoptosis [46] and variable mutation rates [47]. More recently, Plotkin and Nowak [36] developed a stochastic model that investigates the extent to which loss of DNA repair genes and tumor suppressor genes contribute to tumorigenesis.

The corresponding literature for deterministic models is less well developed. Coldman and Goldie develop **compartmental models** in which reversible mutations may occur [10]. Thompson and Royds [45] study competition between subpopulations of tumor cells that differ in the status of the tumor suppressor gene p53 (the gene functions normally in

one population and gives a survival advantage under low oxygen to the second population). Norris [33] has extended these concepts to allow for the effects of spatial variation within avascular tumors.

Avascular Growth

During avascular growth, nutrients that enter the tumor are consumed by live, proliferating cells as they diffuse towards the tumor center. As the tumor grows, the amount of nutrient reaching the center declines until there is insufficient to sustain viable cells. There ensues the formation of a central core of dead (necrotic) cellular material whose size increases as the tumor continues to grow. Thus, a well-developed avascular tumor comprises an outer rim of nutrient-rich, proliferating cells and a central core of nutrient-starved, necrotic debris. These regions may be separated by a layer of oxygen-poor (hypoxic) cells, which are quiescent (viable but nonproliferating).

Many mathematical models of avascular tumor growth that reproduce the phenomena described above have been developed and shown to exhibit good qualitative and quantitative agreement with experimental data. Probabilistic models that focus on individual cells and their interactions with neighboring cells use concepts ranging from **Markov chain** processes [17], through cellular automata [11] to stochastic energy minimization techniques [42]. By contrast, deterministic models tend to focus on cell populations or continua and are formulated as systems of ordinary differential equations [30], spatiotemporal partial differential equations [6, 15, 16, 30, 50] or age-structured partial differential equations [16].

Angiogenesis

Deterministic models of angiogenesis, the process by which an avascular tumor acquires a blood supply from the host tissue, have successfully reproduced many macroscopic features of the developing vascular network. These include: the acceleration of the vascular network and the increase in the number of capillary tips as the vascular network approaches the tumor [7] and regression of the vasculature in response to angiogenic inhibitors such as angiostatin [27].

However, such deterministic models are unable to provide details of microscopic features, such as vessel lengths and distances between buds or anastomoses, that can be obtained using stochastic models [41] and hybrid deterministic-stochastic models in which certain processes (e.g. nutrient diffusion and its consumption by the tumor cells) are viewed deterministically while others (e.g. cell proliferation, death and migration) are treated probabilistically [2].

Vascular Growth

Once a tumor has acquired its own blood supply by angiogenesis (see [28], for example, for an approach to modeling the onset of vascularization), it is able to grow to a large (ultimately life-threatening) size. There has been relatively little modeling work on this crucial stage of tumor development, in part reflecting the complexities that arise; these include the following: the interactions between angiogenesis and the growing tumor (e.g. [32, 34]); the balance between nutrient delivery by the vasculature and nutrient consumption by the tumor (e.g. [1]); the complexities of the immature vascular networks through which the nutrients are delivered, in terms both of their highly tortuous geometry and of the interactions between blood flow and interstitial fluid pressures within the tumor (see [31, 38, 52], for instance); and the complex microenvironments (as described in [5, 23], for example), involving both spatial and temporal variations in oxygen tension and pH in particular, which arise because of nonuniformities in nutrient delivery and waste product removal [35] by the vasculature. These effects have important implications for the effectiveness of different, blood-borne treatment protocols [21].

Invasion

One of the earliest approaches to the modeling of tumor invasion and metastatic spread, pioneered by Greenspan [15], involves linear stability studies of existing models for the growth of radially symmetric tumors; the growth of fingers due to the instability of the symmetric state provides a mechanism for a tumor to invade the surrounding tissue. Such studies, which typically involve an analysis of the interactions between nutrient-limited growth and the physical properties of a growing tumor, remain an active

area of research (see [8], for instance). However, numerous additional complexities have also now been incorporated, including: the role of degradative proteases, for example, in leading to a distinction between noninvasive (benign) tumors, which are often confined by a capsule comprised of connective tissue, and malignant ones (see, for example, [20, 25, 26]); the role of pH (e.g. [51]); interactions with the immune system (surveyed in [4]); interactions with the underlying tissue matrix (e.g. [49]); interactions between primary and secondary tumors (e.g. [14]); and the implications of "diffuse" invasion for imaging and therapy (see [44]). Because cellular mutation may enhance a tumor's invasiveness and because a single cell can lead to metastatic spread, stochastic effects have a crucial role to play in invasion; [48] provides a discussion of the role of deterministic and stochastic models.

Conclusions

The modeling of tumor growth continues to present enormous challenges. A vast range of scales is involved, from subcellular through cellular and tumor to patient. Numerous phenomena are thus of interest, from stochastic effects relating, in particular, to the mutation or survival of single cells to mechanical ones, which may not only influence the interactions of a growing tumor with the surrounding tissue [9, 18] but may also create difficulties in, say, delivering drugs against an adverse pressure gradient [52]. Multiple (and age-structured) populations of cell types are present and three-dimensional effects can be crucial, leading to severe computational challenges. The value of the insights, which experimentally validated modeling can provide is nevertheless, increasingly widely recognized.

References

[1] Adam, J.A. & Noren, R.D. (1993). Equilibrium model of a vascularized spherical carcinoma with central necrosis – some properties of the solution, *Journal of Mathematical Biology* **31**, 735–745.

[2] Anderson, A.R.A. & Chaplain, M.A.J. (1998). Continuous and discrete mathematical models of tumor-induced angiogenesis, *Bulletin of Mathematical Biology* **60**, 857–899.

[3] Armitage, P. & Doll, R. (1954). The age distribution of cancer and multistage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

[4] Bellomo, N. & Preziosi, L. (2002). Modelling and mathematical problems related to tumour evolution and its interaction with the immune system, *Mathematical and Computer Modelling* **32**, 413–452.

[5] Breward, C.J.W., Byrne, H.M. & Lewis, C.E. (2001). Modelling the interactions between tumour cells and a blood vessel in a microenvironment within a vascular tumour, *European Journal of Applied Mathematics* **12**, 529–556.

[6] Breward, C.J.W., Byrne, H.M. & Lewis, C.E. (2002). The role of cell-cell interactions in a two-phase of solid tumor growth, *Journal of Mathematical Biology* **45**, 125–152.

[7] Byrne, H.M. & Chaplain, M.A.J. (1995). Mathematical models for tumour angiogenesis: numerical simulations and nonlinear wave solutions, *Bulletin of Mathematical Biology* **57**, 461–486.

[8] Chaplain, M.A.J. & Sleeman, B.D. (1993). Modelling the growth of solid tumours and incorporating a method for their classification using nonlinear elasticity theory, *Journal of Mathematical Biology* **31**, 431–473.

[9] Chen, C.Y., Byrne, H.M. & King, J.R. (2001). The influence of growth-induced stress from the surrounding medium on the development of multicell spheroids, *Journal of Mathematical Biology* **43**, 191–220.

[10] Coldman, A.J. & Goldie, J.H. (1983). A model for the resistance of tumor-cells to cancer chemotherapeutic-agents, *Mathematical Biosciences* **65**, 291–307.

[11] Duchting, W. (1996). Cancer: A challenge for control theory and computer modelling, *European Journal of Cancer* **32**, 1283–1292.

[12] Folkman, J. (1974). Tumour angiogenesis, *Advances in Cancer Research* **19**, 331–358.

[13] Folkman, J. & Hochberg, M. (1973). Self-regulation of growth in three-dimensions, *The Journal of Experimental Medicine* **138**, 745–753.

[14] Gatenby, R.A., Gawlinski, E.T., Tangen, C.M., Flanigan, R.C. & Crawford, E.D. (2002). The possible role of postoperative azotemia in enhanced survival of patients with metastatic renal cancer after cytoreductive nephrectomy, *Cancer Research* **62**, 5218–5222.

[15] Greenspan, H.P. (1976). On the growth and stability of cell cultures and solid tumours, *Journal of Theoretical Biology* **56**, 229–242.

[16] Gyllenberg, M. & Webb, G. (1990). A nonlinear structured population model of tumour growth with quiescence, *Journal of Mathematical Biology* **28**, 671–684.

[17] Hanin, L.G., Rachev, S.T., Tsodikov, A.D. & Yakovlev, A.Y. (1997). A stochastic model of carcinogenesis and tumour size at detection, *Advances in Applied Probability* **29**, 607–628.

[18] Helmlinger, G., Netti, P.A., Lichtenbeld, H.C., Melder, R.J. & Jain, R.K. (1997). Solid stress inhibits the growth of multicellular tumor spheroids, *Nature Biotechnology* **15**, 778–783.

[19] Ilyas, M., Straub, J., Tomlinson, I.P. & Bodmer, W.F. (1999). Genetic pathways in colorectal and other cancers, *European Journal of Cancer* **35**, 1986–2002.

[20] Jackson, T.L. & Byrne, H.M. (2002). A mechanical model of tumour encapsulation and transcapsular spread, *Mathematical Biosciences* **180**, 307–328.

[21] Jackson, T.L., Lubkin, S.R., Siemers, N.O., Kerr, D.E., Senter, P.D. & Murray, J.D. (1999). Mathematical and experimental analysis of localisation of anti-tumour antibody-enzyme conjugates, *British Journal of Cancer* **80**, 1747–1753.

[22] Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma, *Proceedings of the National Academy of Sciences of the United States of America* **68**, 820–823.

[23] Kraus, M. & Wolf, B. (1998). Physicochemcial microenvironment as key regulator for tumour microevolution, invasion and immune response: targets for endcytotechnological approaches in cancer-treatment, *Endocytosis Cell Research* **12**, 133–156.

[24] Landini, G. & Rippin, J.W. (1996). How important is tumour shape? *The Journal of Pathology* **179**, 210–217.

[25] Landman, K.A. & Pettet, G.J. (1998). Modelling the action of proteinase and inhibitor in tissue invasion, *Mathematical Biosciences* **154**, 23–57.

[26] Larreta-Garde, V. & Berry, H. (2002). Modelling extracellular matrix degradation balance with proteinase/ transglutamine cycle, *Journal of Theoretical Biology* **217**, 105–124.

[27] Levine, H.A., Pamuk, S., Sleeman, B.D. & Nilsen-Hamilton, M. (2001). Mathematical modelling of capillary formation and development in tumour angiogenesis: penetration into the stroma, *Bulletin of Mathematical Biology* **63**, 801–864.

[28] Maggelakis, S.A. (1996). The effects of tumour angiogenesis factor (TAF) and tumour inhibitor factors (TIFs) on tumour vascularization: A mathematical model, *Mathematical and Computer Modelling* **23**, 121–133.

[29] Mao, J.H., Lindsay, K.A., Balmain, A. & Wheldon, T.E. (1998). Stochastic modelling of tumorigenesis in p53 deficient mice, *British Journal of Cancer* **77**, 243–252.

[30] Marusic, M., Bajzer, Z., Vukpavlovic, S. & Freyer, J.P. (1994). Tumour-growth in-vivo and as multicellular spheroids compared by mathematical models, *Bulletin of Mathematical Biology* **56**, 617–631.

[31] McDougall, S.R., Anderson, A.R.A., Chaplain, M.A.J. & Sherratt, J.A. (2002). Mathematical modelling of flow through vascular networks: implications for tumour-induced angiogenesis and chemotherapy strategies, *Bulletin of Mathematical Biology* **64**, 673–702.

[32] Michelson, S. & Leith, J.T. (1996). Host response in tumour growth and progression, *Invasion & Metastasis* **16**, 234–246.

[33] Norris, E.S. (2002). *Modelling the Growth of Avascular Tumours and their Response to Chemotherapy*. PhD Thesis, University of Nottingham.

[34] Orme, M.E. & Chaplain, M.A.J. (1996). A mathematical model of vascular tumour growth and invasion, *Mathematical and Computer Modelling* **23**, 43–60.

[35] Patel, A.A., Gawlinski, E.T., Lemieux, S.K. & Gatenby, R.A. (2001). A cellular automaton model of early tumour growth and invasion: the effects of native tissue vascularity and increased anaerobic tumour metabolism, *Journal of Theoretical Biology* **213**, 315–331.

[36] Plotkin, J.B. & Nowak, M.A. (2002). The different effects of apoptosis and DNA repair on tumorigenesis, *Journal of Theoretical Biology* **214**, 453–467.

[37] Royds, J.A., Dower, S.K., Qwarstrom, E.E. & Lewis, C.E. (1998). Responses of tumour cells to hypoxia: role of p53 and NFkB, **51**, 55–61. *Journal of Clinical Pathology: Molecular pathology*.

[38] Secomb, T.W., Hsu, R., Dewhirst, M.W., Klizman, B. & Gross, J.F. (1993). Analysis of oxygen transport to tumour tissue by microvascular networks, *International Journal of Radiation Oncology, Biology, Physics* **25**, 481–489.

[39] Sinik, Z.T., Albibay, T., Ataoglu, O., Biri, H., Sozen, S., Deniz, N., Karaoglan, U. & Bozkirli, I. (1997). Nuclear p53 overexpression in bladder, prostate and renal carcinomas, *International Journal of Urology* **4**, 546–551.

[40] Stetler-Stevenson, W.G., Aznavoorian, S. & Liotta, L.A. (1993). Tumour cell interactions with the extracellular matrix during invasion and metastasis, *Annual Review of Cell Biology* **9**, 541–573.

[41] Stokes, C.L. & Lauffenburger, D.A. (1991). Analysis of the roles of microvessel endothelial cell random motility and chemotaxis in angiogenesis, *Journal of Theoretical Biology* **152**, 377–403.

[42] Stott, E.L., Britton, N.F., Glazier, J.A. & Zajac, M. (1999). Stochastic simulation of benign avascular tumour growth using the Potts model, *Mathematical and Computer Modelling* **30**, 183–198.

[43] Sutherland, R.M. & Durand, R.E. (1984). Growth and cellular characteristics of multicell spheroids, *Recent Results in Cancer Research* **95**, 24–49.

[44] Swanson, K.R., Alvord, E.C. & Murray, J.D. (2000). A quantitative model for differential motility of gliomas in grey and white matter, *Cell Proliferation* **33**, 317–329.

[45] Thompson, K.E. & Royds, J.A. (1999). Hypoxia and reoxygenation: A pressure for mutant p53 cell selection and tumour progression, *Bulletin of Mathematical Biology* **61**, 759–778.

[46] Tomlinson, I., Novelli, P.M. & Bodmer, W.F. (1995a). Failure of programmed cell-death and differentiation as causes of tumours – some simple mathematical-models, *Proceedings of the National Academy of Sciences of the United States of America* **92**, 11130–11134.

[47] Tomlinson, I., Novelli, P.M. & Bodmer, W.F. (1995b). The mutation rate and cancer, *Proceedings of the National Academy of Sciences of the United States of America* **93**, 14800–14803.

[48] Tracqui, P. (1995). From passive diffusion to active cellular migration in mathematical models of tumour invasion, *Acta Biotheoretica* **43**, 443–464.

[49] Turner, S. & Sherratt, J.A. (2002). Intercellular adhesion and cancer invasion: A discrete simulation using the extended Potts model, *Journal of Theoretical Biology* **216**, 85–100.

[50] Ward, J.P. & King, J.R. (1997). Mathematical modelling of avascular-tumour growth, *IMA Journal of Mathematics Applied in Medicine and Biology* **14**, 39–69.

[51] Webb, S.D., Sherratt, J.A. & Fish, R.G. (1999). Alterations in proteolytic activity at low pH and its association with invasion: A theoretical model, *Clinical and Experimental Metastasis* **17**, 397–407.

[52] Zlotecki, R.A., Baxter, L.T., Boucher, Y. & Jain, R.K. (1995). Pharmacological modification of tumour blood flow and interstitial fluid pressure in a human tumour xenograft – network analysis and mechanistic interpretation, *Microvascular Research* **50**, 429–443.

(*See also* **Mathematical Biology, Overview**)

H.M. BYRNE & J.R. KING

Turnbull Estimator

The Turnbull estimator [27] is a **nonparametric maximum likelihood estimator** (NPMLE) of the distribution function F of a real-valued **random variable** X based on N independent, arbitrarily interval-censored and/or truncated observations $(X_1, X_2, \ldots, X_N)$ of X. The observation X_i is said to be truncated by a set B_i if X_i is drawn from the conditional distribution $F(x : B_i) = \Pr(X \leq x | X \in B_i)$, and X_i is said to be **interval-censored** if X_i is only known to lie in an interval $[L_i, R_i]$.

Interval censoring occurs naturally when X_i represents the time to an event of interest and there is intermittent monitoring for this event. For example, in **AIDS** studies interval-censored data arise in connection with the time of infection with human immunodeficiency virus (HIV) in individuals exposed to the virus. Since only periodic assessment of HIV status is feasible, the time to infection with HIV will be known only to lie in an interval specified by the last negative and the first positive assessment, or it will be right-censored if no positive assessment was made by the time of last examination.

Truncation occurs if X_i is drawn from the population in which observations with values outside truncating set B_i have been removed (*see* **Truncated Survival Times**). For example, if X is a survival time associated with a chronic disease, and individuals suffering from the disease are not followed from the time of diagnosis, but instead are recruited into a study at some later times, then X_i will be included in the sample only if $X_i > V_i$, where V_i is the time from the diagnosis to the time of recruitment (*see* **Delayed Entry**). Thus, X_i is drawn from the population of observations with the same time of diagnosis but from which observations with survival times shorter than V_i were removed. The observation X_i is then said to be left-truncated by $B_i = (V_i, \infty)$.

The observations can be both truncated and interval-censored as in [2, Example I.3.11]. In this example, the data set includes interval-censored and left-truncated observations on the time from the diagnosis of diabetes to the onset of severe complications associated with diabetes.

The sample of truncated and interval-censored observations is then represented by N pairs $(A_1, B_1), (A_2, B_2), \ldots, (A_N, B_N)$, where X_i is truncated by B_i and furthermore is censored by $A_i = [L_i, R_i] \subseteq B_i$. In particular, X_i is right (left) censored if $R_i = +\infty$ ($L_i = -\infty$) and is known exactly if $L_i = R_i$. Interval, right, and left truncation are defined in a similar manner. The truncating sets, B_i, and censoring sets, A_i, can be either fixed or random. Turnbull's method of estimation is also applicable when A_i is a union of disjoint closed intervals.

We note that grouped data are a special case of interval-censored data in which, for each i, the censoring interval $[L_i, R_i]$ is a member of a known, fixed partition of the range of X.

Interval-censored and/or truncated data arise in a wide range of research areas including AIDS studies (see e.g. [4, 5, 7], and [25]) and cancer research (see for example, [12] and [24]). Examples of studies which involve application of the Turnbull estimator to medical data are [21, 22], and [24].

Derivation of the Turnbull Estimator

We sketch the derivation of $\hat{F}$, the NPMLE of F. Assuming that either (A_i, B_i) are fixed, or that they were generated by a random mechanism independent of X_i, the **likelihood** function is proportional to

$$L(F) = \prod_{i=1}^{N} \frac{[F(R_i+) - F(L_i-)]}{P_F(B_i)}, \qquad (1)$$

where $P_F(B_i) = \Pr(X \in B_i)$.

The estimator is derived in two steps. In the first step it is shown that $\hat{F}$ increases on only a finite number of disjoint intervals. This characterization of the support of $\hat{F}$ is used in the second step to compute the estimator. For the case of nontruncated, interval-censored data Peto [23] proposed a Newton–Raphson algorithm (*see* **Optimization and Nonlinear Equations**) for computation of $\hat{F}$. Turnbull [27] developed a simple and intuitively appealing self-consistency algorithm for obtaining $\hat{F}$.

The Support of the Turnbull Estimator

When data are not truncated, i.e. $P_F(B_i) = 1, (1 \leq i \leq N)$, the intervals on which $\hat{F}$ may increase are derived as follows. Let $L = (L_i, 1 \leq i \leq N)$ and $R = (R_i, 1 \leq i \leq N)$, be the sets of left and right endpoints of censoring intervals. The form of (1) indicates that $L(F)$ will be maximized when the values of $F(x)$ are as large as possible for $x \in R$

and as small as possible for $x \in L$ subject to the constraint that F is the distribution function. Accordingly, let $C = \bigcup_{j=1}^{m}[q_j, p_j]$, where $q_1 \le p_1 \le q_2 \le p_2 < \ldots < q_m \le p_m$, be a union of disjoint, closed intervals whose left and right endpoints lie in the sets L and R, respectively, and which contain no other members of L or R. Then on examination of $L(F)$ it can be seen that the support of $\hat{F}$ is contained in C, and that for fixed values of $F(p_j+)$ and $F(q_j-)$, $L(F)$ is independent of the behavior of F within each interval $[q_j, p_j]$. Thus, $\hat{F}$ is flat outside C and is only unique up to the class of distributions with the same values of $\hat{F}(p_j+) - \hat{F}(q_j-)$, $1 \le j \le m$.

When data are also truncated the described construction is in general not valid. The necessary modification of Turnbull's construction of the support of $\hat{F}$ in the presence of truncation involves endpoints of truncation intervals and is discussed by Frydman [14]. In the following we assume that, if data are truncated, the set C is constructed as in [14].

Let $s_j = F(p_j+) - F(q_j-)$, for $1 \le j \le m$. The foregoing discussion shows that the problem of finding the NPMLE of F reduces to one of maximizing

$$L(\mathbf{s}) = \prod_{i=1}^{N}\left(\sum_{j=1}^{m}\alpha_{ij}s_j \Big/ \sum_{j=1}^{m}\beta_{ij}s_j\right) \qquad (2)$$

with respect to $\mathbf{s} = (s_1, \ldots, s_m)$, subject to $\sum_{j=1}^{m} s_j = 1$ and $s_j \ge 0$, $1 \le j \le m$, where $\alpha_{ij} = 1$ if $[q_j, p_j] \in A_i$, 0 otherwise, $\beta_{ij} = 1$ if $[q_j, p_j] \in B_i$, 0 otherwise.

The Self-Consistency Algorithm

Turnbull proposed an **algorithm** for the maximization of $L(\mathbf{s})$ based on the idea of self-consistency. This idea was introduced by Efron [11] and is described, for example, by Cox & Oakes [6]. We briefly describe the algorithm. For $1 \le i \le N$ and $1 \le j \le m$, let $I_{ij} = 1$ if $X_i \in [q_j, p_j]$ and 0 otherwise. Also let J_{ij} be the number of individuals corresponding to the observation X_i who were never observed because their X-values are in the complement of B_i and which have X-values in $[q_j, p_j]$. Turnbull termed these X_is **"ghosts"**. Because of censoring I_{ij} is in general not known but its expectation conditional on the observed data computed under a given value of $\mathbf{s}$ is equal to

$$\mathrm{E}(I_{ij}|\text{data}, \mathbf{s}) = \alpha_{ij}s_j \Big/ \sum_{k=1}^{m}\alpha_{ik}s_k \equiv \mu_{ij}(\mathbf{s}). \qquad (3)$$

The expectation of J_{ij} conditional on the observed data, under $\mathbf{s}$, is given by

$$\mathrm{E}(J_{ij}|\text{data}, \mathbf{s}) = (1 - \beta_{ij})s_j \Big/ \sum_{k=1}^{m}\beta_{ik}s_k = \nu_{ij}(\mathbf{s}). \qquad (4)$$

A self-consistent estimate of $\mathbf{s}$ is defined as a solution to the following system of equations

$$s_j = \sum_{i=1}^{N}[\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})] \Big/ \sum_{i=1}^{N}\sum_{j=1}^{m}[\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})], \quad 1 \le j \le m. \qquad (5)$$

It can be demonstrated that (5) are simply loglikelihood equations for the maximization of $L(\mathbf{s})$ with respect to $\mathbf{s}$. This shows that $\hat{\mathbf{s}}$, the **maximum likelihood estimator** (MLE) of $\mathbf{s}$, is a self-consistent estimator of $\mathbf{s}$. The algorithm for finding the MLE of $\mathbf{s}$ starts with the initial value $\mathbf{s}^0 > 0$, and the improved estimate $\mathbf{s}^1$ is obtained from the right-hand side of (5) evaluated at $\mathbf{s}^0$. One iterates in this fashion until convergence is achieved. We note that the described algorithm is an example of the EM algorithm [9]. The complete data likelihood function is based on the fact that the random variables $\sum_{i=1}^{N}(I_{ij} + J_{ij})$, $1 \le j \le m$, have a **multinomial** distribution with cell probabilities $(s_1, \ldots, s_m)$. Combining the E-step, where the expectations are computed as in (3) and (4), with the M-step gives self-consistency equations in (5). The convergence of the algorithm was demonstrated by Turnbull [27], but it is also assured by the theory of the EM algorithm for an **exponential family** of distributions of which the multinomial distribution is a member.

The Turnbull estimator of F is given by

$$\hat{F}(x) = \begin{cases} 0, & \text{if} \quad x < q_1, \\ \hat{s}_1 + \hat{s}_2 + \cdots + \hat{s}_j, & \text{if} \quad p_j < x < q_{j+1}, \\ & \quad 1 \le j \le m-1, \\ 1, & \text{if} \quad x > p_m, \end{cases}$$

and is undefined for $x \in [q_j, p_j]$, $1 \le j \le m$, so that the way an increase $\hat{s}_j$ occurs over an interval $[q_j, p_j]$ is arbitrary. Thus, when plotted, $\hat{F}$ is a step function with gaps which occur over the intervals in C.

Special Cases

For some special cases of interval-censored data an NPMLE of F has an explicit representation. If all observations are exact or right-censored only, then $\hat{F}$ is the well known **Kaplan–Meier estimator** [17]. Current status data arise when all observations are either right- or left-censored; see the survey by Diamond & McDonald [10]. In this case, $\hat{F}$ also has an explicit representation [3, 16]; see also [18] for an exposition.

Asymptotic Properties

Groeneboom & Wellner [16], using an elegant approach based on **isotonic regression** theory, demonstrated **consistency** and developed asymptotic distribution theory (*see* **Large-sample Theory**) for an NPMLE of F derived from current status data. They also demonstrated consistency and presented partial asymptotic distribution results for an NPMLE of F obtained from "case 2" interval-censored data. The "case 2" interval censoring is the same as arbitrary interval censoring except that exact observations can never be observed.

The asymptotic distribution result obtained by Groeneboom & Wellner [16] for an NPMLE of F derived from current status data is nonstandard and involves $\sqrt[3]{n}$ norming. This shows that there will never be a completely general asymptotic result for the Turnbull estimator of F derived from arbitrarily interval-censored data.

However, assuming that exact observations occur with some positive probability, Li et al. [20] presented an EM algorithm for obtaining a self-consistent estimator of F defined on $(0, \infty)$. The obtained estimator is a smoothed version of the Turnbull estimator. The authors state that consistency and asymptotic normality of the smoothed estimator are established in [28].

Generalizations

Turnbull's method has been generalized by a number of authors to the estimation of multistate survival models in which times of transitions to states may be interval-censored and/or truncated (*see* **Multivariate Survival Analysis**). Many generalizations were motivated by the problems encountered in AIDS studies (unknown time of infection with HIV, left truncation of the **incubation period** of AIDS, and delays in the reporting of AIDS cases), where the typical framework is an irreversible three-state disease model (state 1, not infected; state 2, infected; state 3, clinical AIDS). For example, De Gruttola & Lagakos [7] simultaneously estimated the distribution of the time of infection with HIV, and the distribution of the incubation time (time between infection and the appearance of clinical symptoms of AIDS; *see* **Latent Period**), assuming independence between these times, in the case where the time of infection is interval-censored and there is no truncation. Assuming that a three-state disease model is a time-nonhomogeneous Markov process, Frydman [13] proposed a self-consistent algorithm for estimating the cumulative transition intensities for data of the same form as in [7]. Other generalizations are given by De Gruttola et al. [8], Sun [25], and Frydman [15]. The generalizations by Kim et al. [19], Tu et al. [26], and Alioum & Commenges [1] incorporate covariates.

References

[1] Alioum, A. & Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data, *Biometrics* **52**, 512–524.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1991). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[3] Ayer, M., Brunk, H.D., Ewing, G.M., Reid W.T. & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information, *Annals of Mathematical Statistics* **26**, 641–647.

[4] Bacchetti, P. & Jewell, N.P. (1991). Nonparametric estimation of the incubation distribution of AIDS based on a prevalent cohort with unknown infection times, *Biometrics* **47**, 947–960.

[5] Becker, N. & Melbye, M. (1991). Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity, *Australian Journal of Statistics* **33**, 125–133.

[6] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, New York.

[7] De Gruttola, V. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics* **45**, 1–11.

[8] De Gruttola, V., Tu, X.M. & Pagano, M. (1992). Pediatric AIDS in New York City: estimating the distributions of infection, latency and reporting delay and projecting future incidence, *Journal of the American Statistical Association* **87**, 633–640.

[9] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

[10] Diamond, I.D. & McDonald, J.W. (1991). The analysis of current status data, in *Demographic Applications of Event History Analysis*, T.J. Trussell, R. Hankinson & J. Tilton, eds. Oxford University Press, Oxford.

[11] Efron, B. (1967). The two sample problem with censored data, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. LeCam & J. Neyman, eds. University of California Press, Berkeley, pp. 831–853.

[12] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data, *Biometrics* **42**, 845–854.

[13] Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS, *Journal of the Royal Statistical Society, Series B* **54**, 853–866.

[14] Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations, *Journal of the Royal Statistical Society, Series B* **56**, 71–74.

[15] Frydman, H. (1995). Nonparametric estimation of a Markov "illness - death" process from interval-censored observations, with application to diabetes survival data, *Biometrika* **82**, 773–789.

[16] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser Verlag, Basel.

[17] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[18] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.

[19] Kim, M.Y., De Gruttola, V.G. & Lagakos, S.W. (1993). Analyzing doubly censored data with covariates, with application to AIDS, *Biometrics* **49**, 13–22.

[20] Li, L., Watkins, T. & Yu, Q. (1996). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data, *Scandinavian Journal of Statistics* **24**, 531–542.

[21] Paneth, N., Pinto-Martin, J., Gardiner, J.C., Wallenstein, S., Katsikiotis, V., Hegyi, T., Hiatt, M.I. & Susser, M. (1993). Incidence and timing of germinal matrix/intraventricular hemorrhage in low birthweight infants, *American Journal of Epidemiology* **137**, 1167–1176.

[22] Peckham, C.S. (1991). Children born to women with HIV-1 infection: natural history and risk of transmission. The European Collaborative Study, *Lancet* **337**, 253–260.

[23] Peto, R. (1973). Experimental survival curves for interval-censored data, *Applied Statistics* **22**, 86–91.

[24] Rücker, G. & Messerer, D. (1988). Remission duration: an example of interval censored observations, *Statistics in Medicine* **7**, 1139–1145.

[25] Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies, *Biometrics* **51**, 1096–1104.

[26] Tu, X.M., Meng, X-.L. & Pagano, M. (1993). The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data, *Journal of the American Statistical Association* **88**, 26–36.

[27] Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 290–295.

[28] Yu, Q., Li, L. & Wong, G.Y.C. (1996). On consistency of the self-consistent estimator of survival functions with interval censored data, *Scandinavian Journal of Statistics*, to appear.

H. FRYDMAN

Twin Analysis

The first use of twin resemblance as a means of resolving alternative hypotheses about the causes of human differences appears to have been in 426 AD by Augustine of Hippo in Book V of the *City of God*. Augustine argued that since twins, who were highly correlated in their times of birth, nevertheless had such discrepant life histories, there was little empirical support for planetary influence on human destiny. For Augustine's purpose, it was sufficient that at least some twin pairs showed markedly different life histories. In the nineteenth century, **Sir Francis Galton** suggested the etiologic basis of the distinction between identical (monozygotic or "MZ") and fraternal (dizygotic or "DZ") twins [15] (*see* **Heterozygosity**). He viewed twins as a unique natural experiment which allowed the resolution of the effects of heredity and environment on human development. These effects had been **confounded** in his studies of hereditary human traits in nuclear families and derived pedigrees [16]. Galton solicited letters from twins describing some of their experiences and characteristics. Based on these anecdotal data, Galton likened the developmental trajectories of identical twins to the path of two sticks dropped simultaneously into a stream. Although the eddies of the stream meant that the sticks would probably alternate in relative position as they flowed downstream, the overwhelming pressure of the current, corresponding to hereditary influences on human development, ensured that, on average, the sticks moved at essentially the same rate.

Galton's data, and his approach to data analysis, would receive little recognition today. In the century since Galton's pioneering studies, the twin study, though still subject to many inherent difficulties, has played a critical part in establishing the prima facie case for a significant role of genetic factors in a wide range of human traits and disorders. This article outlines some of the principal contours of past and current thought on the analysis of twin data.

Basic Analytic Method for Continuous Measures

Although the study of twins separated at birth (*see* **Adoption Studies**) has obvious appeal [48], practical issues of obtaining large and representative samples have dictated that the majority of twin studies employ twins reared together. Furthermore, although there are reports of other kinds of twins [3], we focus on the two most frequent classes: monozygotic (MZ) and dizygotic (DZ) twins, who most often have been reared together since birth.

Currently, there is no single perfect method of twin data analysis. The approach taken will depend somewhat on the kinds of data being analyzed and the genetic and environmental issues being explored. It is convenient, however, to begin with the approach laid down by Jinks & Fulker [24], as this is based on the familiar **analysis of variance**. Twin pairs are regarded as **random samples** from a population of pairs, and hence as a random sample of the genetic and environmental factors creating variation in the population of interest. Twins within a pair are assumed to sample at random the **genes** segregating within a family and the environmental differences operating within families. Birth order notwithstanding, there is no reason a priori to order the twins within a pair, so the data for each kind of twin pair may be summarized (Table 1) by a nested analysis of variance recognizing sources of variation within and between pairs or by the derived intraclass **correlation** coefficient. The strength of the twin method stems from the fact that the **components of variance** within and between pairs involve different proportions of the genetic and environmental contributions in MZ vs. DZ pairs.

Under the simplest model, which assumes no **genotype × environment interaction** (G × E) and no genotype–environment covariance (CGE), the contributions of genes and environment to the components of variance are shown in Table 2, where G_1 and G_2 are, respectively, the components of variance due to genetic differences within and between sibships, and E_1 and E_2 are, respectively, the components due to environmental differences within and between sibships.

Table 1 Expectations under a nested analysis of variance of n families of size m, in terms of between (σ_b^2) and within (σ_w^2) components of variance

Source	df	Expected MS
Between families	$n-1$	$\sigma_w^2 + m\sigma_b^2$
Within families	$n(m-1)$	σ_w^2

Table 2 Variance components for MZ and DZ twin pairs, reared together or apart, under a simple genetic model. Subscripts 1 and 2, respectively, denote within- and between-family components of genetic (G) and environmental (E) variance; $G = G_1 + G_2$ and $E = E_1 + E_2$. These components contribute to within (σ_w^2), between (σ_b^2), and total (σ_t^2) variance components in the analysis of variance

Component	Reared together		Reared apart	
	MZ twins	DZ twins	MZ twins	DZ twins
σ_w^2	E_1	$G_1 + E_1$	$E_1 + E_2$	$G_1 + E_1 + E_2$
σ_b^2	$G + E_2$	$G_2 + E_2$	G	G_2
σ_t^2	$G + E$	$G + E$	$G + E$	$G + E$

Other authors sometimes use different notations, especially for the between-families environmental component (also known as "common" environment, "CE", "shared" environment, or "family" environment). The total variance is $G + E$ for all relatives under the model. The expectations of the intraclass correlations $t = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$ for MZ and DZ twins reared together are

$$t_{MZT} = G_1^* + G_2^* + E_2^*,$$
$$t_{DZT} = G_2^* + E_2^*,$$

where the asterisks denote the components of variance expressed as proportions of the total.

The proportion of the total variance attributable to genetic differences (the "broad **heritability**") is $G_1^* + G_2^*$. Early twin researchers proposed a number of tests for the importance of genetic effects in twin studies and ratios intended to summarize the relative importance of genetic factors. Holzinger [23] proposed using the ratio $H = (r_{MZ} - r_{DZ})/(1 - r_{DZ}) = G_1/(G_1 + G_2)$. Similarly, Vandenberg [51] proposed a statistic $F = 1/(1 - H) = (G_1 + E_1)/E_1$ that measures the relative contribution of genes and environment to differences within families. Significance levels and **confidence intervals** could be obtained from the sampling distribution of the intraclass correlation coefficient and variance ratio as appropriate. However, under the model above, it is clear that neither of these ratios corresponds to the broad (or narrow) heritability understood by geneticists.

Genetic Meaning of Statistical Parameters

Two components of variance G_1 and G_2 are merely convenient ways of partitioning the genetic variance in twin pairs into those effects that arise as a result of **segregation** (G_1) and those that arise because of the sampling of parents (G_2). They do not correspond directly to the additive, dominant, or epistatic effects of the genes contributing to variation. Considering only the additive (V_A) and dominant (V_D) components of genetic variance, it may be shown (e.g. 32) that $G_1 = \frac{1}{2}V_A + \frac{3}{4}V_D$ and $G_2 = \frac{1}{2}V_A + \frac{1}{4}V_D$ when mating is random. Thus, in the absence of dominance, $G_1 = G_2 = \frac{1}{2}V_A$.

Discontinuous Traits

Univariate analysis of discontinuous data, specifically concordance rates and **relative risks**, is described in the article on **Twin Concordance**. **Multivariate analysis** of both continuous and discontinuous traits is described below.

Structural Equation Model for Twin Resemblance

Although the biometric genetic model allows for the specification of additive and nonadditive genetic components, it does not deal well with the effects of **assortative mating** and the various forms of nongenetic inheritance. The approach of **path analysis**, though not ideal for the specification of nonlinear effects, provides a convenient way of representing family resemblance in the presence of assortative mating and cultural inheritance [12, 49]. The basic path model for the similarity of MZ and DZ twins reared together is shown in Figure 1 (cf. [39]). The model allows for the additive and dominance effects of genes (A and D), and for the common (shared) family environment (C) and the unique (within-family) environment (E). The basic model assumes **polygenic** autosomal inheritance, random mating, and the additivity and independence of genetic and environmental effects. In the simplest model, it is assumed that genes and environment have the same effects in males and females.

Model Fitting

If MZ and DZ twins are random samples of the genetic and environmental effects in the population,

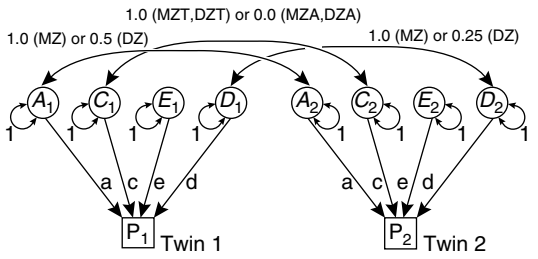

Figure 1 Univariate model for data from monozygotic (MZ) or dizygotic (DZ) twins reared together (T) or apart (A). Additive (A_i) and dominant (D_i) genetic, and common (C_i) and specific (E_i) environment latent variables cause the phenotypes P_i in a linear additive structural equation model

then there should be no difference between the means of MZ and DZ twins, nor any difference between their total phenotypic variances. Some types of social interaction, such as sibling contrast effects [5, 9], can produce zygosity differences in mean or variance or both, as can nonrandom sampling. The model of Figure 1 yields the following predicted values for the **covariance matrices** of phenotypic values in pairs of MZ and DZ twins:

$$\boldsymbol{\Sigma} = \begin{bmatrix} a^2 + c^2 + e^2 + d^2 & \alpha_i a^2 + \beta_i c^2 + \delta_i d^2 \\ \alpha_i a^2 + \beta_i c^2 + \delta_i d^2 & a^2 + c^2 + e^2 + d^2 \end{bmatrix}, \tag{1}$$

where a, c, e, and d are path coefficients for additive genetic, common environment, random environment, and genetic dominance effects, respectively; $\alpha_i = 1.0$ for MZ and 0.5 for DZ twins; $\delta_i = 1.0$ for MZ and 0.25 for DZ; and $\beta_i = 1.0$ for twins reared together and 0.0 for twins reared apart.

The parameters of the **structural equation model** for the pattern of MZ and DZ covariance may be estimated by several approaches, including **maximum likelihood** and weighted **least squares**. Parameters a, d, c, and e cannot be estimated simultaneously from the "classical twin study", which consists of data on MZ and DZ twins living together. Dominance effects tend to reduce the DZ correlation relative to that of MZ pairs, whereas the common environment tends to increase the DZ correlation relative to that of MZs. If d and c are both zero, then the DZ correlation is predicted to be exactly half that of MZs as long as mating is random. Typically, a set of reduced models (omitting either c or d) is fitted to the covariances to explore the major contributions of genes and environment to family resemblance. Where appropriate, **likelihood ratio tests** of **alternative hypotheses** may be conducted (e.g. that c or d is zero) and confidence intervals obtained for the parameter estimates [44]. These operations can be conducted efficiently, with many programs currently available for the structural equation modeling of covariance matrices. Recently, Mx [38] has been widely used in twin and adoption studies. It has the advantage of being able to fit models to raw data, which enables appropriate treatment of data missing completely at random or missing at random [30] (*see* **Missing Data**) as well as the detection of **outliers**.

Multiple Variables and Developmental Change

The univariate model for the analysis of twin data directly extends to the multivariate case. Multivariate path analysis [52] may be used to derive predicted covariances between relatives measured on several variables. In Figure 1, every latent and observed variable is replaced by a vector of variables. Thus the path coefficients are replaced by matrices of path coefficients, which are organized such that column variables cause row variables. The predicted covariances among twins are then:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{AA}' + \mathbf{CC}' + \mathbf{EE}' + \mathbf{DD}' & \alpha_i \mathbf{AA}' + \beta_i \mathbf{CC}' + \delta_i \mathbf{DD}' \\ \alpha_i \mathbf{AA}' + \beta_i \mathbf{CC}' + \delta_i \mathbf{DD}' & \mathbf{AA}' + \mathbf{CC}' + \mathbf{EE}' + \mathbf{DD}' \end{bmatrix}.$$

The form of the matrices **A**, **C**, **E**, and **D** dictates the form of the multivariate model for each component. One simple form is the Cholesky decomposition, in which all matrices are square and lower triangular. This decomposition provides a **robust** way to estimate the genetic and environmental sources of variation in, and covariation between, multiple measures. Covariance due to additive genetic sources is simply $\mathbf{AA}'$, and this matrix may be standardized to obtain additive genetic correlations between traits. The genetic variances and covariances may be expressed as proportions of the total (phenotypic) covariances, to obtain a multivariate analog of the narrow heritability coefficient, yielding information on the relative importance of genetic vs. environmental factors to covariation between traits. Some caution is required because it is possible for some sources of covariance to be negative while others are

positive. Such estimates might arise, for example, if the phenotypic covariance between traits is zero, but the cross-twin cross-trait covariance is positive. Thus the multivariate twin study has a remarkable potential to identify relationships between traits that are uncorrelated within individuals.

The Cholesky decomposition is not a theory-based model – it merely factorizes the covariance matrices **A**, **D**, **C**, and **E**. Other forms of the path coefficient matrix can be used to test alternative (usually simpler) models. A natural example from confirmatory **factor analysis** is to postulate that a single latent factor is responsible for all genetic (or environmental) covariation between traits, along with residual factors specific to each trait. Models of this type are called *biometric factor* or *independent pathway* models. Matrix **A** would be specified as a partitioned matrix **F**:**D**, where **F** is an $m \times 1$ vector of paths from the latent factor and **D** is an $m \times m$ diagonal matrix of residual factors. The form of the environmental factor matrices might be similarly constructed, but there is no obligation to keep the factor structure the same for the different variance components. If specific environmental variation consists of **measurement error**, a diagonal form for **E** would be sufficient.

Another natural branch of multivariate model is the *phenotypic factor*, or *common pathway*, model [26, 33], in which the genetic and environmental factors combine to form a latent factor which subsequently causes variation and covariation between the phenotypes. In addition, there may be trait-specific genetic and environmental factors. This common pathway model predicts that the cross-trait variation and covariation due to the general genetic and environmental factors is proportionate for all traits. Usually, this model has fewer parameters than the biometric factor model and does not fit as well by the χ^2 criterion. However, the model may prove to be a simpler, more **parsimonious**, account of the data, as judged by other indices of fit such as **Akaike's information criterion** [1].

Understanding of the sources of variance and covariance between traits gleaned from multivariate analyses can greatly enhance our knowledge of the role of risk factors in the interplay between genotype, environment, and phenotypic outcome. In studies of complex disorders, it is not good practice to assume that a risk factor is purely environmental, nor that it is a cause rather than a consequence of liability to a disorder. Multivariate genetically informative studies are particularly useful for the identification of risk factors and for the quantification of their effects.

Longitudinal Genetically Informative Studies

Just as the measurement of twins on many variables at a single occasion can provide information about the proportion of covariance between traits that is due to genetic vs. environmental factors, so can multiple measures made on the same twins at several points in time yield information about the sources of longitudinal stability and change. One starting point for analysis is the use of the Cholesky decomposition, which has a good conceptual basis if the variables $P_1, \ldots, P_t$ are ordered chronologically from occasion 1 to t. The first Cholesky factor F_1 causes variation at all occasions. The second causes variation at all occasions *except the first*, and may be conceived of as sources of variation not present at the first occasion, i.e. new variance or "innovations" at the second occasion. All subsequent Cholesky factors $F_{i=3,\ldots,t}$ represent innovations which occur at time i of measurement. In the context of twin studies, this distinction between persistent and innovation variance can be made separately for genetic and environmental components. Either genetic or environmental factors or both may account for patterns of stability and change over time, and the longitudinal twin study allows us to test alternative hypotheses about the origins of individual differences in change over time.

In 1986, Eaves et al. [10] described a number of alternative models of genetic and environmental stability and change. Because data are collected at several points in time, and because **causation** by definition operates from earlier to later events, it is natural to think of earlier variables causing later ones, and not vice versa. Such models are known as **Markov chain**, or Simplex, and have been used in psychological science for at least half a century [18]. In the context of a twin study, each of the components of variation may have a Simplex form, so that, for example, environmental variance at each occasion is partitioned into that due to previous occasions, and that due to innovation since the previous measurement. Application of the genetic Simplex to twin data is described in [39].

A novel component to the Eaves et al. [10] approach was the development of methodology to deal with irregular spacing of intervals between study.

Though not widely used, these methods are valuable for taking into account differences in age between subjects, which is particularly useful when the twin study is expanded to include other relatives.

Application to Discontinuous Data

In the univariate case, data analysis may proceed by maximum likelihood analysis of the **contingency tables** of twin 1 against twin 2 (*see* **Twin Concordance**). Most analyses assume that there is an underlying **normally distributed** liability continuum on which there are one or more abrupt thresholds that subdivide the population into ordinal classes. In principle, it is possible to analyze multivariate data this way, but the likelihood computations require the computation of **multivariate normal** integrals over twice (for twins) as many dimensions as there are variables in the analysis. Beyond-bivariate analysis of twin data (requiring four-dimensional integration) is numerically tedious at this time, even with advanced integration methods [17].

As an alternative to direct maximum likelihood analysis of multivariate ordinal data, it is possible to fit models to matrices of polychoric and polyserial correlations computed using software such as PRELIS [25], together with a weight matrix based on fourth-order **moments** [4]. These methods are practical only when the sample sizes are very large relative to the number of variables being analyzed; with small sample sizes the departure of the fit statistics from chi-square can be substantial.

Other Models

There are many approaches to the analysis of twin data, only some of which involve more elaborate structural equation models. To some extent, any multivariate statistical method can be applied to twin data, often yielding tests of salient genetic and environmental hypotheses. However, it is rare that a statistical method developed for unrelated subjects can be directly applied to data collected from twins. Methodologic development is usually required to specify that the subjects are related and that while certain parameters may be expected to be equal, others may need reparameterization to reflect the a priori knowledge from Mendelian (*see* **Mendel's Laws**) or Fisherian theory of the degree of genetic relatedness.

Continuous Indices of the Environment

A common misconception is that variation for something that seems environmental – like socioeconomic status – is purely environmental. Often, empirical inquiry with a twin or adoption study reveals substantial genetic variability for such traits. Therefore, it is wise to collect genetically informative data on all variables, wherever possible, before exploiting one or other variable as an environmental index. If a suitable environmental index does exist, it is possible (in principle) to identify genetic variance in an otherwise confounded design [36]. Such situations are relatively rare, so the multivariate twin or adoption study – which does not require the existence of an environmental index and is therefore of ubiquitous utility – is more widely used.

$G \times E$ and $G \times$ Sex Interactions

A common question is whether the set of relevant genetic and environmental factors, and the sizes of their effects, remain the same under different conditions [32]. One example would be examination of exposure to stressful experiences (though this may not be entirely environmental). Another example is male vs. female sex. In both cases multigroup structural equation modeling is a viable statistical approach. Twin pairs may be subclassified as concordant for exposure, discordant, or concordant for nonexposure. The basic procedure is to fit models in which parameters are allowed to differ between the subsamples, and to compare them, by likelihood ratio tests, with models in which the parameters are constrained to be equal across all groups. This comparison tests whether the magnitude of the parameters is the same under different environmental conditions. It is possible to carry out this test separately for one or more of the genetic or environmental sources of variation, to test whether, for example, differences in genetic effects alone are sufficient to account for differences in phenotypic variability. The subtler question of whether the same genetic and environmental factors are operating in the different groups may be addressed by developing a more elaborate model, which involves genetic and environmental factors that only operate on those individuals exposed to the environment (see Figure 2). In the special case of sex interactions, the twin study is limited to assessing sex-specific effects for either additive or dominance

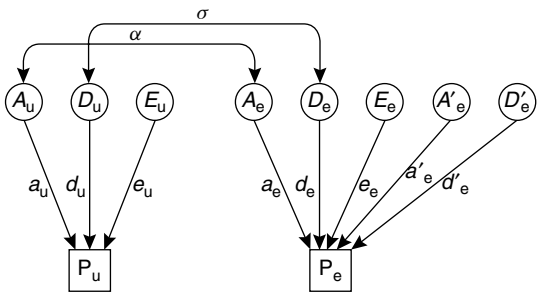

Figure 2 Path diagram of genotype × environment interaction in twins discordant for environmental exposure. For MZ pairs, $\alpha = 1.0$ and $\beta = 1.0$; for DZ pairs, $\alpha = 0.5$ and $\delta = 0.25$. The subscripts u and e identify variables and parameters for unexposed and exposed twins, respectively

(or common environment) genetic effects but not both (a'_e or d'_e in Figure 2). This limitation is because the identical twins are always of the same sex.

While the method may in principle be extended to test simultaneously the effects of several environmental variables, the number of subgroups increases quadratically, requiring an even greater increase in sample size to maintain sufficient numbers in the less frequent subgroups. To some extent this problem – which also occurs when more than a **binary** subclassification of twin pairs is used – can be mitigated by the analysis of the raw data instead of summary statistics [38].

Age Effects

Age can affect parameter estimates in a twin study in many different ways. Perhaps the simplest way, and one that is easiest to correct, is where there is a linear relationship between age and the phenotype. Assuming that the age distribution is equivalent in the MZ and DZ pairs, that the members of a twin pair are measured at the same age, and that twin pairs span a variety of ages, the linear effects of age will inflate the estimate of c^2. Neale & Martin [42] showed how the linear effects of age could be added to the structural equation model for twin data to control for, and estimate, the effects of age. Another approach is to regress out the effects of age on the phenotype prior to structural equation modeling. This latter technique is useful when nonlinear effects of age are of concern. Failure to correct for age in extended twin designs (twins plus their parents, children, or other relatives) where age difference is unequal for different types of relationship (such as parent–offspring vs. twin) can give rise to **biased** parameter estimates.

Similar to genotype × environment interaction, the impact of genes and the environment may vary between ages. In addition, different genetic and environmental factors may be operating at different ages. Such age × genotype or age × environment interactions are best resolved with longitudinal, genetically informative studies (see above).

A further recent development in genetic studies is the use of growth curve modeling [43]. Based on dynamical systems theory, specific functional forms for asymptotic growth or decay are specified. It is then possible, in a single-step analysis, to estimate genetic and environmental components to variation in initial level, rate of growth, and final asymptote components of variation in growth. Such models have great economy for predicting means, variation, and covariation of relatives across many points in time.

Censoring

A common statistical problem concerns the nonrandom observation of test scores. For example, examination scores from a sample of students may be available only for those who passed. Such a sample is termed **censored**. Uncorrected analysis of data from twin pairs in which both members passed the examination would yield biased estimates of genetic and environmental components of variation for the population. Typically, such censoring biases correlations towards zero, but in a nonlinear fashion [46], so the impact on additive genetic and common environment estimates depends on their true population values.

Correction for censoring requires knowledge of the proportion of the sample that has been censored, and, in the case of twins, some assumption about the underlying distribution so that the correlated **ascertainment** can be controlled. Assuming that a variable is normally distributed in the population, the likelihood for a pair of censored observations is

$$\frac{\phi(x_i)}{A}, \tag{2}$$

where $\phi(x_i)$ is the **bivariate normal** pdf

$$|2\pi\,\mathbf{\Sigma}|^{-n/2} \exp\left[-\tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)\right],$$

in which $\mathbf{\Sigma}$ is the population covariance matrix, $\boldsymbol{\mu}_i$ is the (column) vector of population means of x_1 and x_2,

and $|\mathbf{\Sigma}|$ and $\mathbf{\Sigma}^{-1}$ denote the determinant and inverse of the matrix $\mathbf{\Sigma}$, respectively. The divisor A in (2) is the ascertainment correction, and is the probability that a pair will both have passed the examination. If the passing grade is t, the correction term is

$$A = \int_t^\infty \int_t^\infty \phi(x_i)\, \mathrm{d}x_2\, \mathrm{d}x_1.$$

In a similar fashion, data from the pairs discordant for passing the examination may be analyzed jointly. These data increase the precision of the estimates of the means and variances of the examination scores, but add little to the precision of the estimated covariances between the twins, and hence add little to the precision of estimates of heritability or common environmental variance.

Nonrandom Sampling

Similar to the analysis of data from censored samples, other forms of nonrandom sampling are sometimes used to increase statistical **power**. For example, in "four corners" sampling, pairs are selected because both members have scores either above a cutoff point t or below $-t$. This design will, after correction, yield estimates of genetic and environmental effects that are more precise (i.e. have smaller confidence intervals) than those from a random sample of equal size. By itself, this observation is of little practical use because scores on the whole sample are required in order to effect the nonrandom sampling, in which case the data from the whole sample should be used. However, the method gains value when subjects are being screened for a more expensive measurement protocol such as genotyping (see below).

Another common source of nonrandom samples in twin data occurs when all patients at a hospital are asked if they are a twin. In this case a key concept is the *probability of ascertainment*, usually denoted as π. Maximum likelihood based methods for estimating π and twin concordances are available [2]. Especially for rare disorders, nonrandom sampling, such as that through hospital records, can yield large gains in statistical power [45]. However, nonrandom sampling may suffer from nonrepresentativeness, if, for example, hospital cases do not constitute a random sample of cases in the population.

Co-Morbidity and Causation

An interesting feature of the classical twin study is its ability, under certain conditions, to test causal models of the relationship between traits or disorders [8, 22, 39] (*see* **Causation**). The principle can be seen in a simplified example, in which trait A correlates between twins, and trait B does not. However, trait A and trait B correlate within individuals. If trait A causes trait B, then we would predict a cross-correlation between trait A in twin 1 and trait B in twin 2. On the other hand, if B causes trait A, then the cross-twin, cross-trait correlation would be predicted to be zero. This simple example extends to the more general case; identification of the model requires that the twin correlation (either MZ or DZ) for trait A differs from the twin correlation for trait B. A limitation of this method is that, to avoid incorrect inferences about causation, measurement error variance should be approximately equal for the two traits.

Simple causation is only one of a variety of possible sources of covariation between traits, or of *co-morbidity* between disorders. We discussed above several multivariate models of resemblance based on genetic and environmental correlations between traits. In addition to these, there are a number of other models of co-morbidity that are clinically relevant [28, 41]. In particular, there is the idea that succumbing to a disorder in and of itself causes an increase in liability to a second disorder. This is quite distinct from simple correlation or causation between the liabilities to two disorders, and has different implications for etiology and treatment.

Other models of co-morbidity are based on hypotheses that two disorders are either a single-liability dimension, two correlated dimensions, or three independent dimensions, where excess co-morbid cases are a third, independent disorder [41]. All the models described in this section can be assessed with a maximum-likelihood or minimum χ^2 goodness-of-fit function. When cell frequencies are low, minimum χ^2 often performs better than maximum likelihood.

Dimensionality

Biostatisticians and psychometricians frequently rely on the joint distribution of the items on a test to

discern the dimensionality of a trait. Item response theory, **latent class analysis**, and **multidimensional scaling** are all examples of devices to assess dimensionality with multivariate data. Twins offer a unique perspective on dimensionality, gleaned from the pattern of covariance between twins measured *on a single scale*. A simple example of this is for cigarette smoking, where individuals may be current smokers, ex-smokers, or nonsmokers [21]. The question is whether these categories represent an ordinal scale based on a single underlying continuum of liability or whether there are different processes involved in initiation vs. persistence of tobacco use. Data from twins may be summarized as contingency tables, and used to test against frequencies predicted under a bivariate normal model with two thresholds and a polychoric correlation coefficient. This single-liability dimension model may be compared with two main alternatives. First is a model of two independent normally distributed liability dimensions in which one dimension discriminates between those who initiate and those who do not, and the second dimension discriminates between those who quit smoking and those who persist. Second is a combined model, which is the same as the two-dimensional model except that those intermediate on the initiation dimension invariably become ex-smokers. Empirically, the combined model has the greatest support, and the single-liability dimension model is usually statistically rejected, indicating that smoking status cannot be regarded as a unitary trait.

Latent Class Analysis

It is possible to extend latent class analysis [34] for use with twin data by constraining the pattern of joint class frequencies in pairs of twins [11] to accord with various models of twin resemblance. A simple form of patterning allows members of a twin pair to correlate for class membership. The usual reparameterization to estimate variation due to additive genetic, shared environmental, and random environmental components can be applied to class membership frequencies, assuming some ordering of classes (polychoric correlation) or nonparametric association such as Cramer's C [6]. Of greater interest from a biometric genetic standpoint is the patterning of classes according to **gene frequencies** under a major locus model (see Table 3). Optionally, these cell frequencies may be multiplied by **penetrance** parameters, which estimate the probability that each genotype gives rise to the three latent classes. Not all penetrance parameters can be estimated in this model.

Table 3 Pairwise class membership frequencies for MZ and DZ twins under a single biallelic major locus model, with gene frequencies p and $q = 1 - p$

		MZ twin 1			DZ twin 1		
		1	2	3	1	2	3
Twin 2	1	p^2	0	0	p^4	$2p^3q$	q^2p^2
	2	0	$2pq$	0	$2p^3q$	$4p^2q^2$	$2pq^3$
	3	0	0	q^2	q^2p^2	$2pq^3$	q^4

Extending the Twin Design

The classical twin study should be regarded only as a good starting point for **genetic epidemiologic** investigations. It is clearly limited in a number of ways, several of which can be overcome by extending the twin design to include other relatives. As noted above, data from either MZ or DZ twins reared apart enable the joint estimation of the effects of both dominance genetic factors and the common environment, along with the effects of additive genes and specific environment. However, data from adoptees are themselves subject to a number of critical assumptions (*see* **Adoption Studies**), so it is worthwhile to consider extended twin designs.

A simple and practical extension is to assess twins and their parents [13, 47]. Parents yield two new types of data: marital and parent–offspring covariances. The marital covariance provides information about the presence of **assortative mating**. Usually, assortment is assumed to be phenotypic [12]. The consequences for the twin study are an increase in total phenotypic variance, the same increase in MZ covariance (hence no increase in correlation) and the same increase for DZ twins. Functionally, with the MZ and DZ covariances increasing by equal amounts, assortative mating will be estimated as part of the shared environment variance in the uncorrected classical twin study. It should be noted that appropriate modeling of twin–parent data will yield estimates of heritability that include a portion due to the effects of assortment.

Parent–offspring covariances can be used to identify several alternative parameters. When environmental transmission between generations is assumed

to be absent, the parent–offspring covariance may be used to identify genetic dominance, as parents and their offspring do not share genetic dominance deviations (*see* **Genotype**). It seems naive to the behavioral scientist to suppose that human parents, who expend much effort in rearing their offspring, do not influence the environment of their children. A genetically informative design, such as twins and their parents or a full adoption study, is required to discriminate between genetic and environmental transmission from parent to child. Path models for mixed environmental and genetic transmission in the twin–family design appear to have been described first by Fulker [13]. Of these models, those that involve an effect of the parents' phenotypes on their offspring ($P \rightarrow$ E transmission) generate *genotype–environment covariance* because the genes and the environment have a common origin (parental genes affect both parental phenotype, and hence children's environment as well as children's genotypes). Genotype–environment covariance is typically assumed to be at equilibrium over generations, yielding a nonlinear constraint among the parameters of the model.

Much parent–child environmental transmission is modeled in a univariate fashion, e.g. a parent's body mass is assumed to form part of the environment for their children's body mass. Transmission might, however, operate indirectly via other variables, e.g. a parent's dietary preference forms part of the environment for their children's body mass. Multivariate models which overcome this limitation have been described [47]. Twin–parent data are still limited by the assumption that heritability is the same in both generations. While this may be defensible for adult populations where the age of the children from some families overlaps that of the parents in others, it is much more questionable in studies of juvenile twins and their parents.

Beyond twins and parents, the next most obvious extension is twins, their spouses, and their children. Interestingly, twins and their spouses – and in-law data in general – allow the resolution of several alternative models of assortative mating [20]. Of particular note is the distinction between phenotypic homogamy – a matching of the phenotypes of spouses – from social homogamy, where there may be some social stratification or other environmental basis for spouse similarity. These models have quite different consequences for genetic and environmental sources of variation, so incorrect modeling of marital resemblance can give rise to biased estimates, especially when the mate resemblance is high.

A great advantage of twins and their children is that both generations contain a genetically informative design – the classical twin study in the parental generation, and siblings, cousins, and half-siblings related through MZ twins in the offspring generation. Thus a test of equality of heritability over generations is possible by a likelihood ratio test. If heritability varies continuously over the lifespan, and different genetic and environmental factors operate at different ages, then longitudinal data would be needed to quantify these changes.

Recently, models for twins, their parents, spouses, siblings, and children have been devised and applied [50]. Initially, these models were fitted to z-transformed correlations using a weighted least-squares procedure, but more recently Mx has been used to fit the models to the raw data, enabling statistically appropriate maximum likelihood analysis of pedigrees of irregular size and structure.

Measured Genotypes in Studies of Twins

Assessments of the genotype, such as **blood groups**, have long been used to help distinguish MZ from DZ twins [35]. More recently, greater precision has been obtained through **polymorphic** markers, against which questionnaire methods have proved approximately 95% accurate. Beyond **zygosity determination**, DZ twins can be regarded as a special type of sibling, where age and intrauterine effects have been matched. As such, they are suitable for studies of genetic **linkage** and association.

Genetic marker studies fall into three broad categories: studies of candidate loci, studies of linkage, and studies of association. A candidate locus is a specific region of the genome that is thought to cause phenotypic variation. DZ twins, who may share zero, one, or two genes identical by descent (IBD) at this locus, can yield information about the locus' effects, but MZ twins, who share two alleles IBD at every locus (somatic mutations notwithstanding), are uninformative about the effects of the candidate locus [37].

Studies of genetic linkage use data from highly polymorphic genetic markers approximately evenly

spaced along the genome. Two branches of linkage studies concern binary traits – especially putative genetic disorders – and quantitative traits. Again, DZ twins are informative for linkage, whereas MZ twins provide information to resolve background genetic from common environment effects. Several approaches to the analysis of linkage data have been described, from the **regression** methods of Haseman & Elston [19] to recent multipoint maximum-likelihood methods [29]. All rely on estimating the probabilities, $p(0)$, $p(1)$, and $p(2)$, that a sib-pair shares zero, one, or two alleles IBD, and the association of these probabilities with the phenotypic resemblance. These probabilities may be estimated by a variety of methods; those of Kruglyak & Lander [29] are currently popular. There are two main ways of using the probabilities: separately to weight the likelihood under different models, or jointly using the summary statistic $\hat{\pi} = 0.5p(1) + p(2)$. In addition, there are two main ways to utilize univariate data from sib-pairs or DZ twins: signed intrapair differences or raw data. Intrapair differences may be convenient when selected samples are used, but they carry a cost of loss of information and hence statistical power. Similarly, use of $\hat{\pi}$ can simplify data analysis but lose statistical power, although – in some cases at least – less so than the use of intrapair differences [14].

Structural equation modeling can clarify the use of genetic marker data to detect the effects of quantitative trait loci, and provides a straightforward extension to the multivariate case. Figure 3 shows path diagrams for the alternative approaches with $\hat{\pi}$ and weighted models. The likelihood maximized with the $\hat{\pi}$ approach is based on the multivariate normal pdf, but the predicted population covariance matrix $\mathbf{\Sigma}$ changes for each sib-pair, according to the $p(i)$ values. The log likelihood of the weighted model is given by

$$L_{\mathrm{M}} = \log \sum_{i=0}^{2} p_i L_i,$$

where L_i is the likelihood under the model for siblings that share i alleles IBD, and p_i is the probability based on the marker data that the siblings share i alleles IBD. Likelihood for mixtures of

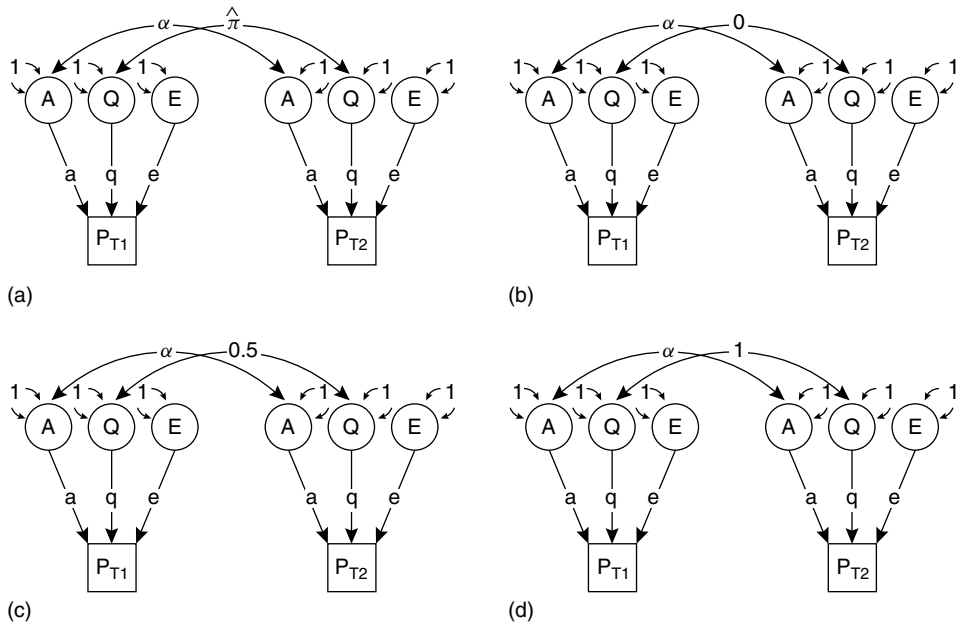

Figure 3 Models for quantitative trait locus effects (Q) in data from monozygotic (MZ) or dizygotic (DZ) twins reared together. Additive polygenic background (A_i) and specific (E_i) environment latent variables also cause the phenotypes (P_i) in a linear additive structural equation model. Figure (a) shows the use of $\hat{\pi}$ as an estimate of the correlation between QTL effects of siblings; models (b) through (d) are used jointly to compute the likelihood from weighted mixture distributions

normal distributions can be evaluated in the structural equation modeling program Mx [38].

Estimates of the path coefficient in Figure 3 from **Q** to the phenotype give an indication of the impact of the putative QTL on the phenotype, and likelihood-based confidence intervals may be obtained to judge their statistical significance. A formal likelihood ratio test may be used for this purpose by fixing the parameter q to zero and comparing the **goodness of fit** under the two models. An alternative test is available through the LOD score, being the log of the odds under the model where the sibling IBD 0, 1, and 2 probabilities are fixed at 0.25, 0.5, and 0.25, respectively, against the model where the probabilities are set to their estimated values. The log likelihood has well-known distributional properties for the multivariate case (where **Q** affects multiple traits) and is therefore a natural choice. It should be noted that there remain problems with the method in that the likelihood ratio does not asymptote exactly to χ^2, so alternative methods using **bootstrapping** may be preferred.

Problems of the Twin Method

Several assumptions of the twin method are open to question. Most of these may be tested empirically either within the twin study itself, by reference to population statistics, or through expansion of the twin study to include other relatives.

Representativeness

A general problem with any **observational** or **experimental study** is that the sample collected may not be representative of the population, and therefore any conclusions drawn would not necessarily generalize to the population. Of course, studies that use volunteer samples always run the risk of being unrepresentative of nonvolunteers. In twins, some information concerning the degree of volunteer bias can be obtained by comparison of pairs discordant for study participation with concordant participating pairs [40]. This method requires nonzero correlation between twins for volunteering, which is commonly observed.

Obstetric Complications

A particular concern with twin data is that their uterine development is unusual, and that this has lasting effects on development. However, these effects, if they exist, are hard to detect because the means and variances of twins are comparable with population norms in many different domains, from personality and psychopathology to body mass index and cardiovascular function. There is evidence that twins have lower than average verbal abilities, but this seems to be associated with the social aspect of having a same-aged companion rather than biological complications [35]. One possibility that has not yet been tested thoroughly is the effects of chorionicity (MZ twins may occupy either one or two amniotic sacs, depending on the time in development at which the zygote splits). Systematic ascertainment of twins during pregnancy – such as currently occurs in Belgium – may help address this question.

Sibling Interaction

As mentioned above, twins may have a direct influence on one another during development. If they do, we would predict differences in total variance of twins compared to nontwins [9]. Even within the twin study, if the MZ and DZ correlations differ prior to interaction, then we would expect differences in their total variance. Structural equation modeling of twin data implicitly tests few these effects, for substantial sibling interaction will cause the model to fit badly. Few cases of sibling effects have been observed in empirical studies. Few parental ratings of activity in their children interaction seems to exist, but multivariate analyses and those of behavioral observations suggest that this is a parental contrast (rater bias) effect rather than genuine interaction.

Equal Environments

The "equal environments" assumption requires that MZ and DZ twins share environmental experiences to the same extent. It is often found that MZ twins are dressed more similarly and are treated more similarly than are DZ twins. However, the important factor here is whether or not these similarities lead to appreciable variation in measured phenotypes. Usually this is tested by examining the correlation of absolute intrapair differences with the measure of treatment similarity (contact between twins is amenable to similar methods). More recently, treatment effects have been incorporated into structural equation models [27]. In general, the assumption appears valid for

psychopathology and other behavioral traits [27, 31]. Furthermore, to the extent that MZ twins elicit more similar treatment by others than do DZ twins, variation is correctly estimated as, and is ascribed to, genetic factors, albeit that they are acting through the environment. Effects of this type are sometimes referred to as those of the *extended phenotype* [7].

No Genetic Dominance

The **confounding** of genetic dominance with the common environment is a difficult problem which is best overcome by the addition of further types of family relatives, as described above. Failure of the assumption without test will cause an underestimation of the effects of the common environment; however, twice as much genetic dominance as common environment would be needed to mask it completely.

No Assortative Mating

Tests for assortative mating are best carried out by assessing a sample of parents; as noted above, the parents of twins make an excellent sample for this purpose.

Linearity

Almost all models considered to date assume a linear effect of genotype on phenotype. Nonlinear effects are entirely plausible, but are difficult to test in data from human subjects. In practice, most nonlinear effects carry a substantial linear component, and it is good scientific practice (Occam's razor) to keep the models simple unless further complications are warranted.

The Future of the Twin Method

The rise of molecular genetics led some researchers to infer that the twin study is dead. As may be seen from this article, we believe such reports to be a great exaggeration. While the proportion of funds and research effort devoted to molecular genetics has clearly grown to dwarf that spent on twin studies, the latter continues to grow in absolute terms. Indeed, the effort spent on twin studies seems to be growing faster than traditional epidemiologic areas. Twin studies have the advantage that linkage studies which include MZ and DZ twins can assess the proportion of QTL and non-QTL genetic variance. In addition, there are large **databases** of twin data that have been gathered on a variety of phenotypes worldwide. These are valuable resources from which samples may be specially selected (for extreme phenotypic values) to increase the power of linkage studies.

Finally, it should be said that the best way to study the environment is to use a genetically informative design. Without controlling for genotype – which the twin study allows us to do statistically – putative environmental factors almost always may have a biological component. Thus the twin study will likely play a significant role in the assessment of environmental as well as genetic causes of individual differences and disease.

References

[1] Akaike, H. (1987). Factor analysis and AIC, *Psychometrika*, **52**, 317–332.

[2] Allen, G. & Hrubec, Z. (1979). Twin concordance: A move general model, *Acta Geneticae et Medicae Gemellologiae* **28**, 3–13.

[3] Boklage, C.E. (1987). Twinning, nonrighthandedness and fusion malformations: evidence for heritable causal elements held in common, *American Journal of Medical Genetics* **28**, 67–84.

[4] Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

[5] Carey, G. (1986). A general multivariate approach to linear modeling in human genetics, *American Journal of Human Genetics* **39**, 775–786.

[6] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[7] Dawkins, R. (1982). *The Extended Phenotype: The Gene as the Unit of Selection*. Oxford University Press, Oxford.

[8] Duffy, D.L. & Martin, N.G. (1994). Inferring the direction of causation in cross-sectional twin data: theoretical and empirical considerations. *Genetic Epidemiology* **11**, 483–502.

[9] Eaves, L.J. (1976). A model for sibling effects in man, *Heredity* **36**, 205–214.

[10] Eaves, L.J., Long, J. & Heath, A.C. (1986). A theory of developmental change in quantitative phenotypes applied to cognitive development, *Behavior Genetics* **16**, 143–162.

[11] Eaves, L., Silberg, J., Hewitt, J., Meyer, J., Rutter, M., Simonoff, E., Neale, M. & Pickles, A. (1993). Genes, personality and psychopathology: a latent class analysis of symptoms of attention-deficit hyperactivity disorder

in twins, in *Nature, Nurture and Psychology*, R. Plomin & G.E. McClearn, eds. American Psychological Association, Washington.
[12] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.
[13] Fulker, D.W. (1982). Extensions of the classical twin method, in *Human Genetics, Part A: The Unfolding Genome*. Alan R. Liss, New York, pp. 395–406.
[14] Fulker, D.W. & Cherny, S.S. (1996). An improved multipoint sib-pair analysis of quantitative traits, *Behavior Genetics* **26**, 527–532.
[15] Galton, F. (1865). Hereditary talent and character, *MacMillan's Magazine* **12**, 157–166, 318–327.
[16] Galton, F. (1869). *Hereditary Genius: An Inquiry into its Laws and Consequences*. Macmillan, London.
[17] Genz, A. (1992). Statistics applications of subregion adaptive multiple numerical integration, in *Numerical Integration*, T.O. Espelid & A. Genz, eds. Kluwer, Dordrecht, pp. 267–280.
[18] Guttman, L. (1954). A new approach to factor analysis: The radex, in *Mathematical Thinking in the Social Sciences*, P.F. Lazarsfeld, ed. Free Press, Glencoe, pp. 258–349.
[19] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a locus, *Behavior Genetics* **2**, 3–19.
[20] Heath, A.C. & Eaves, L.J. (1985). Resolving the effects of phenotype and social background on mate selection, *Behavior Genetics* **15**, 15–30.
[21] Heath, A.C. & Martin, N.G. (1993). Genetic models for the natural history of smoking: evidence for a genetic influence on smoking persistence, *Addictive Behaviors* **18**, 19–34.
[22] Heath, A.C., Kessler, R.C., Neale, M.C., Hewitt, J.K., Eaves, L.J. & Kendler, K.S. (1993). Testing hypotheses about direction-of-causation using cross-sectional family data, *Behavior Genetics* **23**, 29–50.
[23] Holzinger, K.J. (1929). The relative effect of nature and nurture influences on twin differences, *Journal of Educational Psychology* **20**, 245–248.
[24] Jinks, J.L. & Fulker, D.W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior, *Psychological Bulletin* **73**, 311–349.
[25] Jöreskog, K.G. & Sörbom, D. (1993). *New Features in PRELIS 2*. Scientific Software International, Chicago.
[26] Kendler, K.S., Heath, A.C., Martin, N.G. & Eaves, L.J. (1987). Symptoms of anxiety and symptoms of depression: same genes, different environments? *Archives of General Psychiatry* **44**, 451–457.
[27] Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C. & Eaves, L.J. (1993). A test of the equal-environment assumption in twin studies of psychiatric illness, *Behavior Genetics* **23**, 21–27.
[28] Klein, D.N. & Riso, L.P. (1994). Psychiatric disorders: problems of boundaries and comorbidity, in *Basic Issues in Psychopathology*, C.G. Costello. The Guilford Press, New York, pp. 19–66.
[29] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.
[30] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
[31] Loehlin, J.C. & Nichols, R.C. (1976). *Heredity, Environment, and Personality*. University of Texas Press, Austin.
[32] Mather, K. & Jinks, J.L. (1982). *Biometrical Genetics: The Study of Continuous Variation*, 3rd Ed. Chapman & Hall, London.
[33] McArdle, J.J. & Goldsmith, H.H. (1990). Alternative common-factor models for multivariate biometric analyses, *Behavior Genetics* **20**, 569–608.
[34] McCutcheon, A.L. (1987). *Latent Class Analysis*. Sage, Newbury Park.
[35] Mittler, P. (1971). *The Study of Twins*. Penguin, Harmondsworth.
[36] Morton, N.E. (1982). *Outline of Genetic Epidemiology*. Karger, New York.
[37] Nance, W.E. & Neale, M.C. (1989). Partitioned twin analysis: a power study, *Behavior Genetics* **19**, 143–150.
[38] Neale, M.C. (1997). *Mx: Statistical Modeling*, 4th Ed. Box 980126 MCV, Richmond VA 23298.
[39] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer, Boston.
[40] Neale, M.C. & Eaves, L.J. (1993). Estimating and controlling for the effects of volunteer bias with pairs of relatives, *Behavior Genetics* **23**, 271–277.
[41] Neale, M.C. & Kendler, K.S. (1995). Models of comorbidity for multifactorial disorders, *American Journal of Human Genetics* **57**, 935–953.
[42] Neale, M.C. & Martin, N.G. (1989). The effects of age, sex and genotype on subjective expressions of drunkenness following a challenge dose of alcohol, *Behavior Genetics* **19**, 63–78.
[43] Neale, M.C. & McArdle, J.J. (1997). Structured latent growth curves for twin data: Mx models, *Behavior Genetics*, to appear.
[44] Neale, M.C. & Miller, M.M. (1997). The use of likelihood-based confidence intervals in genetic models, *Behavior Genetics* **27**, 113–120.
[45] Neale, M.C., Eaves, L.J. & Kendler, K.S. (1994). The power of the classical twin study to resolve variation in threshold traits, *Behavior Genetics* **24**, 239–258.
[46] Neale, M.C., Eaves, L.J., Kendler, K.S. & Hewitt, J.K. (1989). Bias in correlations from selected samples of relatives: the effects of soft selection, *Behavior Genetics* **19**, 163–169.
[47] Neale, M.C., Walters, E.E., Eaves, L.J., Maes, H.H. & Kendler, K.S. (1994). Multivariate genetic analysis of twin-family data on fears: Mx models, *Behavior Genetics* **24**, 119–139.
[48] Plomin, R. & DeFries, J.C. (1990). *Behavioral Genetics: A Primer*, 2nd Ed. W.H. Freeman, Oxford.

[49] Rice, J., Cloninger, C.R. & Reich, T. (1978). Multifactorial inheritance with cultural transmission and assortative mating. I. Description and basic properties of the unitary models, *American Journal of Human Genetics* **30**, 618–643.

[50] Truett, K.R., Eaves, L.J., Heath, A.C., Hewitt, J.K., Meyer, J.M., Silberg, J., Neale, M.C., Martin, N.G., Walters, E.E. & Kendler, K.S. (1992). A model system for analysis of family resemblance in extended kinships of twins, *Behavior Genetics* **24**, 35–49.

[51] Vandenberg, S.G. (1966). Contributions of twin research to psychology, *Psychological Bulletin* **66**, 327–352.

[52] Vogler, G.P. (1985). Multivariate path analysis of familial resemblance, *Genetic Epidemiology* **2**, 35–53.

M.C. NEALE

Twin Concordance

The comparison of similarity, or concordance, for a **binary** trait between monozygotic (MZ) and dizygotic (DZ) twin pairs can be used to test the null hypothesis that genetic factors do not influence the variance of that trait. Under the classic twin model it is assumed that nongenetic factors relevant to the trait variance are shared to the same extent within MZ pairs as they are within DZ pairs, so that greater concordance within MZ pairs is evidence against the null nongenetic hypothesis and in favor of a genetic alternative.

For binary traits there are several measures of association within twin pairs, such as the Pearson **correlation** (which is related to the classic test for independence in a **two-by-two table**), the tetrachoric correlation, the **odds ratio**, and several probabilities conditional on the trait status of one or both twins. The latter conditional probabilities have been referred to traditionally in the twin literature as "concordance rates".

Some methods of estimation allow adjustment for measured factors that might influence trait **prevalence**. This is important. For example, if the expression of the trait depends on age, methods that do not take this into account will give higher estimates of concordance. This is because twins within a pair are perfectly matched for age, whereas twins from different pairs may differ in age. In the classic twin method, failure to adjust for such putative factors would be misleading if those factors were not independent of zygosity. Crude measures that use only the numbers of twins in the cells of the 2×2 table implicitly assume that the MZ and DZ pairs are comparable for such factors.

Statistical Framework and Definitions

Let Y_1 and Y_2 be random variables representing the binary trait status of twin 1 and twin 2, respectively, within the same pair. Often the trait represents disease status, and this situation will be used for illustrative purposes. Let $Y_j = 1$ if twin j is affected, otherwise let $Y_j = 0$, for $j = 1, 2$. For simplicity, let the probability that a twin is affected be the same for both members of a pair, i.e. let $P = \Pr(Y_j = 1)$ be independent of j. Define the joint probabilities as: $P_{11} = \Pr(Y_1 = 1, Y_2 = 1)$, $P_{10} = \Pr(Y_1 = 1, Y_2 = 0) = \Pr(Y_1 = 0, Y_2 = 1) = P_{01}$, and $P_{00} = \Pr(Y_1 = 0 \text{ and } Y_2 = 0)$. Note that

$$P_{11} + P_{10} + P_{01} + P_{00} = 1 \quad \text{and}$$
$$P = P_{11} + P_{10} = P_{11} + P_{01}. \tag{1}$$

The *casewise concordance*, P_c, the probability that one member of a pair is affected given that the other twin is affected, is

$$\begin{aligned} P_c &= \Pr(Y_2 = 1 | Y_1 = 1) \\ &= \Pr(Y_1 = 1 | Y_2 = 1) \\ &= \frac{P_{11}}{P}. \end{aligned} \tag{2}$$

The *pairwise concordance*, P_p, the probability that both members of a pair are affected given that at least one member is affected, is

$$\begin{aligned} P_p &= \Pr(Y_1 = 1 \text{ and } Y_2 = 1 | Y_1 = 1 \text{ or } Y_2 = 1) \\ &= \frac{P_{11}}{P_{11} + 2P_{10}}. \end{aligned} \tag{3}$$

Therefore, casewise and pairwise concordance are two different measures, and the decision of which to work with depends on the question(s) being asked. For example, if as in **genetic counseling** one wants to know the probability that a twin will become affected, given that the other twin is affected, then casewise concordance is appropriate. However, if interest is in predicting the pair disease status when all one knows is that at least one twin is affected, then pairwise concordance is relevant.

The *Pearson correlation* is

$$\begin{aligned} \rho &= \frac{\operatorname{cov}(Y_1, Y_2)}{[\operatorname{var}(Y_1)\operatorname{var}(Y_2)]^{1/2}} \\ &= \frac{P_{11} - P^2}{P(1 - P)} \\ &= \frac{P_c - P}{1 - P}. \end{aligned} \tag{4}$$

Note that

$$P_{jk} = \Pr(Y_1 = j) \Pr(Y_2 = k) + \delta_{jk} \rho D, \tag{5}$$

where $D = P(1 - P)$ and $\delta_{jk} = 1$ if $j = k$, else -1.

The **odds ratio** is

$$\psi = \frac{P_{11} P_{00}}{P_{10} P_{01}}. \tag{6}$$

It can be shown from (1), (5), and (6) that

$$\rho^2 - \rho\left(2 + \frac{1}{K}\right) + 1 = 0, \tag{7}$$

where $K = (\psi - 1)P(1 - P)$.

The *tetrachoric correlation* is the Pearson correlation of a presumed underlying normally distributed "liability" score, which, when dichotomized at appropriate cut-points, gives expected proportions in the four cells that best approximate the observed numbers in the 2×2 table (*see* **Genetic Liability Model**).

Estimation

If all twin pairs in a defined population are sampled at random, estimation of the above measures of association is straightforward. If pairs are studied only because (at least) one twin is affected, the correlations and odds ratio cannot be estimated. The concordances may be estimable, but to do so the process whereby the pairs were "ascertained" needs to be known. If there is complete **ascertainment**, the concordance rate estimator one would naturally use when all pairs are randomly sampled is appropriate; it is an unbiased estimator of the respective conditional probability. When there is incomplete ascertainment, however, this natural estimator is no longer unbiased.

Under Random Population Sampling

Let n_{ij} be the number of twin pairs with $Y_1 = i$ and $Y_2 = j$, $n_{\,\mathrm{d}} = n_{01} + n_{10}$ be the number of pairs discordant for disease, and n be the total number of pairs observed.

The **maximum likelihood** estimates of P and P_{c} are, respectively,

$$\hat{P} = \frac{2n_{11} + n_{\,\mathrm{d}}}{2n} \tag{8}$$

and

$$\hat{P}_{\mathrm{c}} = \frac{2n_{11}}{2n_{11} + n_{\,\mathrm{d}}}. \tag{9}$$

The asymptotic variances of these estimates are

$$\frac{\hat{P}(1 - \hat{P})}{n} - \frac{n_{\,\mathrm{d}}}{4n^2}$$

and

$$\hat{P}_{\mathrm{c}}^2(1 - \hat{P}_{\mathrm{c}})\left[\frac{1}{n_{11}} - \frac{1}{n_{\,\mathrm{d}}}\right],$$

(see [1, Appendix C], where in addition an expression for the asymptotic variance of the maximum likelihood estimate of ρ is given).

Smith [2] presents methods for interpreting the casewise concordance (referred to there as the "proband concordance rate"; see below) in terms of an estimate of the tetrachoric correlation, and a standard error.

The square of the maximum likelihood estimator of ρ is X^2/n, where $X^2 = (n_{11}n_{00} - n_{10}n_{01})^2 n \div [(n_{00} + n_{01})(n_{00} + n_{10})(n_{01} + n_{11})(n_{10} + n_{11})]$ is Pearson's χ^2 statistic (*see* **Chi-square Tests**) for the 2×2 table of disease in pairs with $n_{10} = n_{01} = 1/2n_{\,\mathrm{d}}$. This provides a simple approximate χ_1^2 test of the hypothesis $\rho = 0$.

Sampling of Affected Pairs Only

Suppose that pairs are observed only if one or both twins are affected. Under complete ascertainment, the maximum likelihood estimate of P_{c} is

$$\hat{P}_{\mathrm{c}} = \frac{2n_{11}}{2n_{11} + n_{\,\mathrm{d}}}, \tag{10}$$

and for large samples the standard error is approximately $[\hat{P}_{\mathrm{c}}(1 - \hat{P}_{\mathrm{c}})(2 - \hat{P}_{\mathrm{c}})/[2n_{11} + n_{\,\mathrm{d}})]^{1/2}$.

The maximum likelihood estimator of P_{p} is

$$\hat{P}_{\mathrm{p}} = \frac{n_{11}}{n_{11} + n_{\mathrm{d}}}, \tag{11}$$

with standard error $[\hat{P}_{\mathrm{p}}(1 - \hat{P}_{\mathrm{p}})/(n_{11} + n_{\mathrm{d}})]^{1/2} = [n_{11}n_{\mathrm{d}}/(n_{11} + n_{\mathrm{d}})^3]^{1/2}$.

Under incomplete ascertainment, for the n_{11} pairs concordant for disease a distinction must be made between those $n_{11\mathrm{D}}$ *doubly* ascertained pairs in which both twins were found to be affected in the original sampling from the population, and those $n_{11\mathrm{S}}$ *singly* ascertained pairs in which only one member was found to be affected in the original sampling, and the other was subsequently found to be affected only on further examination, for whatever reason. Therefore $n_{11} = n_{11\mathrm{D}} + n_{11\mathrm{S}}$. The ascertainment probability, π, is the probability that an affected twin will be identified from the original sampling from the population. Its maximum likelihood estimator is

$$\hat{\pi} = \frac{2n_{11\mathrm{D}}}{2n_{11\mathrm{D}} + n_{11\mathrm{S}}}. \tag{12}$$

The *probandwise estimator* is

$$\hat{P}_{\mathrm{pr}} = \frac{2n_{11\mathrm{D}} + n_{11\mathrm{S}}}{2n_{11\mathrm{D}} + n_{11\mathrm{S}} + n_{\mathrm{d}}}. \tag{13}$$

Therefore, whereas under complete ascertainment $\hat{P}_{\mathrm{c}}$ is the maximum likelihood estimator of the casewise concordance, P_{c}, under *incomplete* sampling the maximum likelihood estimator of P_{c} is $\hat{P}_{\mathrm{pr}}$. $\hat{P}_{\mathrm{c}}$ given in (10) becomes $(2n_{11\mathrm{D}} + 2n_{11\mathrm{S}})/(2n_{11\mathrm{D}} + 2n_{11\mathrm{S}} + n_{\mathrm{d}})$, which is not equal to $\hat{P}_{\mathrm{pr}}$. This shows that under incomplete ascertainment $\hat{P}_{\mathrm{c}}$ is a biased estimator of P_{c}. If π is close to 1, however, the bias is small.

The asymptotic variances of $\hat{\pi}$ and $\hat{P}_{\mathrm{pr}}$ are

$$\hat{\pi}^2(1-\hat{\pi})^2\left[\left(\frac{1}{n_{11\mathrm{D}}}\right) + \left(\frac{1}{n_{11\mathrm{S}}}\right)\right]$$

and

$$\hat{P}_{\mathrm{pr}}(1-\hat{P}_{\mathrm{pr}})^2\left\{1 + \left[(4n_{11\mathrm{D}} + n_{11\mathrm{S}})n_{\mathrm{d}} / (2n_{11\mathrm{D}} + n_{11\mathrm{S}})^2\right]\right\}/n_{\mathrm{d}}.$$

Unfortunately, the difference between the concordance in the population (in statistical parlance a *parameter*) and the concordance *estimator* based on the observed numbers of pairs sampled, has usually been obfuscated in the twin literature through use of the same notation for both entities.

In summary, therefore:

1. Under random or complete ascertainment:
   (a) the casewise estimator is unbiased for the casewise concordance, and
   (b) the pairwise estimator is unbiased for the pairwise concordance.
2. Under incomplete ascertainment:
   (a) the casewise estimator is biased for the casewise concordance, and
   (b) the pairwise estimator is biased for the pairwise concordance, but
   (c) the probandwise estimator is unbiased for the casewise concordance.

Taking into Account Main Effects of Measured Variables on the Trait Mean

As mentioned at the start, the trait mean may depend on measured factors, and inference about the binary trait covariance (concordance) may differ depending on whether or not these are taken into account. The maximum likelihood methods can be extended to allow the parameter P to depend on these factors, introducing a new set of parameters. By using a numerical maximization routine all parameters can be straightforwardly estimated, allowing the twin concordance to be estimated while adjusting the trait prevalence for covariates. Furthermore, the twin concordances can also be estimated as functions of measured factors, such as zygosity, age, sex, cohabitation status, years living apart; see [1]. This allows for a more critical appraisal of the null and alternate genetic hypotheses.

References

[1] Hannah, M.C., Hopper, J.L. & Mathews, J.D. (1983). Twin concordance for a binary trait. I. Statistical models illustrated with data on drinking status, *Acta Geneticae Medicae et Gemellologiae* **32**, 127–137.

[2] Smith, C. (1974). Concordance in twins: methods and interpretation, *American Journal of Human Genetics* **26**, 454–466.

(*See also* **Twin Analysis**)

JOHN L. HOPPER

Twin Registers

Twins offer unusual and interesting opportunities for biomedical research [7, 19]. In the classic (and widely used) twin study design the degree of similarity (concordance) of identical (presumed monozygotic) and like-sex nonidentical (presumed dizygotic) twin pairs are compared [7] (*see* **Zygosity Determination**). This approach indicates an upper limit to the relative importance of genetic factors, in the context of the prevailing lifestyle and environment, although it has long been recognized that the method cannot rigorously prove genetic causation [14]. However, potential uses of twins extend well beyond the classic twin method to studies in molecular genetics, **environmental epidemiology**, developmental biology and behavioral science. In combination, such research can address the influence of both genetic and environmental factors, and their interaction [3, 8, 17].

Dizygotic twins who are concordant or highly discordant for disease or a quantitative trait may be studied for **genetic markers** as a special case of the *sib-pair technique* [5, 15]. Monozygotic pairs who are discordant for a particular condition, however rare, offer powerful evidence for nongenetic determinants of disease [4]. Experimental studies based on monozygotic twin pairs are particularly informative for outcomes where a powerful **gene–environment interaction** is suspected [2]. Twins differ from singletons in both intrauterine environment and upbringing. Comparison of adult characteristics of twins and singletons may point to long-term effects of these unusual early experiences [1]. Certain experiences which are unique to twins, such as the occurrence of death or disease in a co-twin, may also deserve special study.

The rarity of twins in general population samples poses a problem for research studies focusing on dichotomous disease outcomes, rather than continuously measured physiological, psychological or other variables. Twin series based on volunteers or respondents to mass media appeals may be **biased** in important respects [7, 14, 19], in particular towards an excess of females and identical pairs concordant for behavioral and disease characteristics [10]. To overcome these problems, national twin registers have been compiled in Sweden [13], Norway [11], Finland [9], and Denmark [6, 12]. No national twin register exists in the UK, although the UK shares with Scandinavian countries the basic ingredients required to create one: a near-complete population register and identity numbers allocated at birth or in other ways which are informative with regard to twin status [18].

Work with the Scandinavian registers has shown that, among adults and older children, zygosity can be determined reliably in most cases by postal questionnaire, enquiring about the visual identity of co-twins [16]. Twins who are visually identical are almost always monozygotic, although a few monozygotic pairs are sufficiently dissimilar to be considered visually nonidentical. This simple method of estimating zygosity can be validated by more objective methods, such as tissue typing or DNA "fingerprinting" (*see* **DNA Sequences**), techniques which may be required to ascertain zygosity reliably in infants and younger children. These more expensive and invasive techniques have yet to be applied systematically to national twin registers, although a selective application, to pairs in whom the questionnaire information is equivocal, may be considered.

References

[1] Alin-Akerman, B. & Fischbein, S. (1991). Twins: are they at risk? A longitudinal study of twins and non-twins from birth to 18 years of age, *Acta Geneticae Medicae et Gemellologiae* **40**, 29–40.

[2] Bouchard, C., Perusse, L. & Leblanc, C. (1990). Using monozygotic twins in experimental research to test for the presence of a genotype-environment interaction effect, *Acta Geneticae Medicae et Gemellologiae* **39**, 85–89.

[3] Bryan, E.M. (1992). The role of twins in epidemiologic studies, *Paediatric and Perinatal Epidemiology* **6**, 460–464.

[4] Bundey, S. (1991). Uses and limitations of twin studies, *Journal of Neurology* **238**, 360–364.

[5] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[6] Hauge, M., Harvald, B., Fischer, M., Gotlieb-Jensen, K., Juel-Nielsen, N., Raebild, I, Shapiro, R. & Videbach, T. (1968). The Danish twin register, *Acta Geneticae Medicae et Gemellologiae* **2**, 315–331.

[7] Hrubec, Z. & Robinette, C.D. (1984). The study of human twins in medical research, *New England Journal of Medicine* **310**, 435–441.

[8] Kaprio, J., Koskenvuo, M. & Rose, R.J. (1990). Population-based twin registries: illustrative applications in genetic epidemiology and behavioural genetics from

the Finnish Twin Cohort Study, *Acta Geneticae Medicae et Gemellologiae* **39**, 427–439.

[9] Kaprio, J., Sarna, S., Koskenvuo, M. & Rantasalo, I. (1978). The Finnish twin registry: formulation and compilation, questionnaire study, zygosity determination procedures and research program, *Progress in Clinical and Biological Research* **24B**, 179–184.

[10] Kendler, K.S. & Holm, N.V. (1985). Differential enrollment in twin registries: its effect on prevalence and concordance rates and estimates of genetic parameters, *Acta Geneticae Medicae et Gemellologiae* **34**, 125–140.

[11] Kringlen, E. (1978). Norwegian twin registers, in *Twin Research. Proceedings of the Second International Congress of Twin Studies. Part B. Biology and Epidemiology*, W.E. Nance, ed. Alan R. Liss, New York, pp. 189–195.

[12] Kyvik, K.O., Green, A. & Beck-Nielsen, H. (1995). The new Danish twin register: Establishment and analysis of twinning rates, *International Journal of Epidemiology* **24**, 589–596.

[13] Medlund, P., Cederlöf, R., Floderus-Myrhed, B., Friberg, L. & Sörensen, S. (1976). A new Swedish twin registry, *Acta Medica Scandinavica* **600**, *Supplement*, 1–104.

[14] Price, B. (1950). Primary biases in twin studies: a review of prenatal and natal difference-producing factors in monozygotic pairs, *American Journal of Human Genetics* **2**, 293–352.

[15] Risch, N. & Zhang, H. (1995). Extreme discordant sib pairs for mapping, quantitative trait loci in humans, *Science* **268**, 1584–1589.

[16] Sarna, S., Kaprio, J., Sistonen, P. & Koskenvuo, M. (1978). Diagnosis of twin zygosity by mailed questionnaire, *Human Heredity* **28**, 241–254.

[17] Segal, N.L. (1993). Twin, sibling and adoption methods. Tests of evolutionary hypotheses, *American Psychologist* **48**, 943–956.

[18] Strachan, D.P. & Burnett, A.C. (1997). Systematic identification of twins by computerized searches of National Health Service patient registers in the UK, *Journal of Epidemiology and Community Health* **51**, 96–100.

[19] World Health Organization (1996). The use of twins in epidemiological studies. Report of the WHO meeting of investigators on methodology of twin studies, *World Health Organization Chronicle* **20**, 121–128.

(*See also* **Twin Analysis**; **Twin Concordance**)

DAVID PETER STRACHAN

Two-by-Two Table

A two-by-two table is a 2×2 array of frequencies, obtained by classifying items according to two dichotomous characteristics. We denote by n_{ij} the frequency in row i and column j $(i, j = 1, 2)$. The marginal row totals are $n_{i\cdot} = n_{i1} + n_{i2}$ $(i = 1, 2)$, the marginal column totals are $n_{\cdot j} = n_{1j} + n_{2j}$ $(j = 1, 2)$ and the total frequency is $n = n_{11} + n_{12} + n_{21} + n_{22}$

n_{11}	n_{12}	$n_{1\cdot}$
n_{21}	n_{22}	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	n

Two-by-two tables arise very often in biomedical studies. The following are three examples:

Example 1 Consider a clinical trial (*see* **Clinical Trials, Overview**) to compare two treatments. If the outcome variable has only two categories, say "recovered" and "not recovered", then the results can be displayed in a 2×2 table with rows corresponding to the treatments and columns to the outcomes. For example, n_{11} is the number of patients who received the first treatment and have recovered.

Example 2 Consider a study to compare men and women with respect to cigarette smoking. Samples of $n_{1\cdot}$ men and $n_{2\cdot}$ women are drawn from the target population, and each person is classified as "smoker" or "nonsmoker". The results can be summarized in a 2×2 table where the rows represent males and females, and the columns represent the two outcome categories.

Example 3 To investigate the association between smoking and a certain lung disease, a sample of n persons is drawn from the target population and each sampled individual is classified with respect to both smoking status (smoker or nonsmoker) and disease status (diseased or nondiseased).

It is important to note a difference between the first two examples and the last one. Examples 1 and 2 represent comparative studies, where the row totals $n_{i\cdot}$, which are the number of patients in each treatment group in Example 1, and the number of persons from each gender in Example 2, have been determined before the beginning of the study. However, Example 3 represents a cross-sectional study, where only the total sample size, n, is fixed in advance, while the marginal totals $n_{i\cdot}$ and $n_{\cdot j}$ are random. We refer to these two situations as case 1 (row totals are fixed, column totals are random) and case 2 (both row and column totals are random), respectively. While most analytical methods for 2×2 tables are the same in both cases, it is important to understand the differences between the two cases for the sake of formulating the questions regarding a specific table and for the interpretation of the results of the statistical analysis. In case 1, the main interest is to compare the two rows with regard to the probability of a specific outcome (i.e. the probability of being in a specific column of the table). Denote by p_i the probability of the first column for an individual in row i $(i = 1, 2)$. Then the main interest is the comparison of p_1 and p_2. In case 2, one is mainly interested in the association between the two characteristics. The natural null hypothesis in this case is that of independence, i.e. that the probability p_{ij} of falling into row i and column j equals the product of the marginal probabilities $p_{i\cdot}$ of row i and the marginal probability $p_{\cdot j}$ of column j. An equivalent formulation of this hypothesis is $p_{11}p_{22} = p_{12}p_{21}$.

Measures of Association and Agreement

Numerous measures of the strength of the association (*see* **Association, Measures of**) between the row and column classifications in a 2×2 table have been proposed over the years. In case 1, one may be interested in the difference $p_1 - p_2$ or in the ratio p_1/p_2. When the first category of the outcome variable represents a disease, then these measures are known as the *risk difference* and the *risk ratio*, respectively (*see* **Relative Risk**). To estimate these measures, one simply replaces each p_i by its natural estimator $n_{i1}/n_{i\cdot}$. In case 2, the most important measure of association in medical applications is the **odds ratio** $p_{11}p_{22}/p_{12}p_{21}$, which can be estimated by the observed odds ratio, $n_{11}n_{22}/n_{12}n_{21}$. The odds ratio is also a useful measure in case 1, where it is defined as $p_1(1 - p_2)/[p_2(1 - p_1)]$. For example, in case-control studies the odds ratio is used as an approximation to the risk ratio, which cannot be measured directly in this kind of study.

A measure of a different type is the **kappa** *coefficient of agreement*. It is used to assess the

amount of agreement between two raters who assign the same item to one of two (or more) categories. This coefficient, which is often used in reliability studies (*see* **Reliability Study**), is defined as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where $p_0 = p_{11} + p_{22}$ is the probability that both raters classify an item to the same category, and $p_e = p_{1\cdot}p_{\cdot 1} + p_{2\cdot}p_{\cdot 2}$ is the probability of agreement expected by chance.

Tests of Hypotheses

The main hypothesis of interest in case 1 is that the probability of an item being classified into the first column is the same in both rows, i.e. H_0: $p_1 = p_2$. The most commonly used test for this hypothesis is based on the statistic comparing the observed proportions in the two groups:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{[\hat{p}(1-\hat{p})(1/n_{1\cdot} + 1/n_{2\cdot})]^{1/2}},$$

where $\hat{p}_i = n_{i1}/n_{i\cdot}$ is the estimated probability of the first column in row i, and $\hat{p} = n_{\cdot 1}/n$ is the estimated pooled probability of the first column under H_0. The denominator in the expression for Z is the standard error of the numerator, hence the distribution of Z under H_0 is approximately standard normal as long as the sample sizes are not too small. A one-sided test against the alternative H_1: $p_1 > p_2$ rejects H_0 when $Z \geq z_{1-\alpha}$, where $z_{1-\alpha}$ is the upper 100α percentile of the standard normal distribution. A two-sided test against the alternative H_1: $p_1 \neq p_2$ rejects H_0 when $|Z| \geq z_{1-\alpha/2}$.

From a computational viewpoint, it is more convenient to use *Pearson's χ^2 statistic* (*see* **Chi-square Tests**), defined as:

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}.$$

The two-sided test rejects H_0 when χ^2 exceeds the upper 100α percentile of the χ^2 distribution with one degree of freedom. Since $\chi^2 = Z^2$, the two tests are equivalent when the alternative is two-sided. In the one-sided case, it is more natural to use the Z test.

In case 2, the main hypothesis of interest is that of independence of the row and column classification variables, i.e. H_0: $p_{ij} = p_{i\cdot}p_{\cdot j}$. It can be shown that Pearson's χ^2 test is appropriate in case 2 as well.

The adequacy of the use of the χ^2 distribution as an approximation to the exact sampling distribution of Pearson's statistic has been investigated by several statisticians. A conservative rule of thumb requires all the expected frequencies under H_0, $e_{ij} = n_{i\cdot}n_{\cdot j}/n$, to exceed 5. A more liberal rule allows one of the expected frequencies to be as small as 1, as long as the other three are at least 5. Other statisticians suggested modifications to Pearson's statistic in attempts to improve the χ^2 approximation for small samples. The most frequently used modification is **Yates's continuity correction**. The Yates-corrected χ^2 statistic is defined as

$$\chi^2_C = \frac{n\left(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2}\right)^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}.$$

This correction is obtained by subtracting $\frac{1}{2}$ from the two frequencies n_{ij} which are larger than the corresponding expected frequencies e_{ij}, and adding $\frac{1}{2}$ to the other two frequencies (which are smaller than the expected frequencies under H_0). Since χ^2_C is always smaller than χ^2, the test based on Yates's correction is more conservative.

An alternative approach to the small sample problem is the use of an *exact test* (*see* **Exact Inference for Categorical Data**). One should note that the term "exact" does not mean that the test has a probability of exactly α of rejecting H_0 when this hypothesis is true. All it means is that the test is based on the exact sampling distribution (under H_0) of a particular test statistic. An exact test may be conditional or unconditional. The difference between these two types of exact tests will be explained in the context of case 1. Under the null hypothesis, the distribution of the observations n_{ij}, and therefore the distribution of every test statistic, depends on the unknown common value of p_1 and p_2, denoted by p. Let T be a test statistic, and suppose that H_0 is rejected for large values of T. Then one must determine a critical value, say $t_{1-\alpha}$, such that $\Pr_{H_0}(T \geq t_{1-\alpha}) \equiv \alpha$. However, this last probability depends on the unknown value of p; hence, it is usually impossible to determine a single critical value which will guarantee a type I error probability of exactly α for every p. One way to overcome this problem is to use the *conditional* distribution of T when the marginal column totals, $n_{\cdot j}$, are fixed at their observed values. This

distribution is independent of p and hence one can find a critical value such that $\Pr_{H_0}(T \geq t_{1-\alpha}|n_{\cdot j}) \equiv \alpha$. Alternatively, one can calculate the conditional P value as $\Pr_{H_0}(T \geq t_0|n_{\cdot j})$, where t_0 is the value of T for the observed table, and then reject H_0 when this conditional P value is less than the nominal α.

The conditional distribution of the cell frequencies when the marginal totals are held fixed is hypergeometric (*see* **Hypergeometric Distribution**):

$$\Pr(n_{11}, n_{12}, n_{21}, n_{22}|n_{\cdot 1}, n_{\cdot 2}) = \frac{\binom{n_{\cdot 1}}{n_{11}}\binom{n_{\cdot 2}}{n_{12}}}{\binom{n}{n_{1\cdot}}}.$$

Therefore, the conditional P value can be calculated as the sum of the hypergeometric probabilities of all the possible 2×2 tables (with the same marginal totals as the observed table) for which the value of T is greater than or equal to the observed value, t_0. The most popular conditional exact test for 2×2 tables is **Fisher's exact test**, where the test statistic T is Pearson's statistic χ^2. It is easy to see that if another test statistic, T' is a monotonic function of T in the conditional sample space, then the conditional exact tests based on T and T' are equivalent. For the one-sided test, n_{11} is a monotonic function of χ^2; hence, the one-sided Fisher test can be based on n_{11}. In the two-sided case, Fisher's test can be based on the absolute difference between the observed and the expected frequencies, $|n_{11} - e_{11}|$, which is a monotonic function of χ^2. (The absolute difference is the same for all four cells.)

To derive the *unconditional* distribution of a test statistic under H_0 one uses the fact that when the column totals are considered random, then n_{11} and n_{21} are independent binomial variables (*see* **Binomial Distribution**) with a common probability p of success. Thus:

$$\Pr(n_{11}, n_{12}, n_{21}, n_{22}; p) = \binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}} p^{n_{11}+n_{21}} (1-p)^{n_{12}+n_{22}}.$$

If T is, as before, a test statistic and H_0 is rejected for large values of T, then the type I error probability associated with a critical value c is

$$\alpha(c, p) = \sum_{T \geq c} \Pr(n_{11}, n_{12}, n_{21}, n_{22}; p).$$

The summation is over all the possible 2×2 tables (with the fixed values of the row totals and any values of the column totals) for which the test would reject H_0. This critical value depends on the unknown value of the nuisance parameter p. One way to eliminate p is by averaging according to some prior distribution. Alternatively, one can substitute the estimate $\hat{p}$ for p. However, in order to stay within the traditional framework of the theory of testing hypotheses, it seems more natural to require that the size of the test, i.e. the maximum (over p) of the type I error probability, will never exceed the nominal significance level α. Thus, the critical value for the unconditional exact test based on T is the smallest value of c for which

$$\max_{0 \leq p \leq 1} \alpha(c, p) \leq \alpha.$$

Unconditional tests of this type have been discussed and compared by McDonald et al. [10], Suissa & Shuster [11], and Haber [6].

The development of computation technology enabled statisticians to investigate the properties of asymptotic and exact tests for 2×2 tables. Actual type I error probabilities and powers were calculated either by complete enumeration of all possible outcomes or via simulations. The results of these computations lead to a controversy regarding the use of conditional exact tests. The opponents of the conditional exact tests argue that these tests are very conservative and, as a result, their power is much lower as compared with unconditional exact tests. The proponents of the conditional approach have raised several arguments, four of which are presented here:

1. The correct sample space should be the conditional one, since once a particular 2×2 table has been observed, the only other 2×2 tables of interest are those with column totals equal to the observed values of $n_{\cdot 1}$ and $n_{\cdot 2}$. The conditional exact tests are not too conservative when their *conditional* type I error and power are calculated.
2. The column totals do not provide any relevant information regarding the equality of p_1 and p_2, and hence conditioning on these marginal totals should not result in any loss of information.
3. Fisher's exact test, when supplemented by a randomization process to ensure that the size

of the test will be equal to the nominal level of significance (α), is uniformly most powerful among the unbiased tests (UMPU) (*see* **Most Powerful Test**), and therefore the conditional approach provides the "optimal" test.

4. Fisher's exact test seems conservative when one uses a fixed nominal α. However, for most practical purposes one is interested in the P value as a measure of the degree to which the data support H_0, rather than as a tool for deciding whether to "accept" or "reject" the null hypothesis. When tests are compared with regard to their attained P values, Fisher's test is no longer conservative or powerless.

The first argument is a philosophical matter; there is no "correct" or "incorrect" sample space. The second argument seems unconvincing, as the lack of information provided by the column totals is not a sufficient reason to consider them as fixed. The third argument has two major flaws: (i) in practice, nobody uses a random device to decide whether to accept or reject H_0 when the test statistic equals the critical value; (ii) the optimality property in this case is meaningless, since none of the tests used in practice is unbiased. (An unbiased test is one for which the probability of rejecting H_0 is always larger when $p_1 \neq p_2$ than when $p_1 = p_2$.) In fact, there are situations when an unconditional biased test has a larger power than the UMPU test [6]. The fourth argument is incorrect. Calculations have shown [7] that when tests are compared with respect to their attained P values, Fisher's test is still conservative and has a low relative efficiency compared to other tests.

Despite all this, there are two pieces of evidence that suggest that conditioning on the column totals per se is not the cause for conservatism of the conditional tests: (i) the conservatism (and the resulting lack of power) almost completely disappears when larger tables (e.g. 2×3 tables) are considered; (ii) a simple modification of Fisher's test has adequate power even though it is still a conditional test [5]. For an exact test which rejects H_0 for large values of a statistic T, this modified test defines the P value as

$$\Pr(T > t_0) + \tfrac{1}{2}\Pr(T = t_0),$$

where the probabilities are calculated in the conditional sample space. It rejects H_0 when this P value is less than α. This procedure is known as the *mid P value test*. Thus, it seems that the main reason for the conservatism of Fisher's exact test is the very small number of points in the conditional sample space for a 2×2 table, rather than any loss of information resulting from conditioning. For example, when $n_{1\cdot} = n_{2\cdot} = 10$, and a table with $n_{\cdot 1} = 5$ and $n_{\cdot 2} = 15$ is observed, then the conditional sample space has only six points, compared with 121 points in the unconditional sample space. Therefore, one can expect a large difference between $\Pr(T > t_0)$ and $\Pr(T \geq t_0)$, both of which could be used as measures of the amount of evidence for or against H_0.

In summary, is seems that the mid P value test should be acceptable to both proponents and opponents of the conditional approach. It is based on the conditional distribution of the test statistic, but its attained type I error probability is usually close to the nominal significance level and its power is similar to that of unconditional tests.

Sample Size Determination

A common question asked by investigators when designing a study is "How large should our sample be?" Determination of the sample sizes (*see* **Sample Size Determination**) will be discussed here in the context of a comparative study, i.e. case 1. The row totals $n_{1\cdot}$ and $n_{2\cdot}$ represent the sizes of the samples drawn from the two groups compared. It is assumed that Pearson's χ^2 statistic is used to test the null hypothesis $p_1 = p_2$ at a given significance level α, and that the desired power of the two-sided test is $1 - \beta$ when the actual absolute difference between p_1 and p_2 is δ. If nothing is known about the probabilities p_1 and p_2, then the total sample size is minimized with $n_{1\cdot} = n_{2\cdot}$ [9]. However, if the relative magnitude of p_1 and p_2 is known, then it is advantageous to take a larger sample from the group whose p is closer to zero or one [1]. The following discussion is limited to the case of equal sample sizes.

Let m denote the common value of $n_{1\cdot}$ and $n_{2\cdot}$. Suppose first that one has some rough estimates of p_1 and p_2 and let $\overline{p} = (p_1 + p_2)/2$. Then a first approximation to the required sample sizes is [3]

$$m' = \left\{ z_{1-\alpha/2}[2\overline{p}(1-\overline{p})]^{1/2} + z_{1-\beta}[p_1(1-p_1) + p_2(1-p_2)]^{1/2} \right\}^2 \Big/ \delta^2.$$

It turns out that the actual power attained with this value of m is less than the specified value of

$1 - \beta$. Casagrande et al. [2] suggested to adjust m' as follows:

$$m = \frac{m'}{4}\left[1 + \left(1 + \frac{4}{m'\delta}\right)^{1/2}\right]^2.$$

This value of m provides a close approximation to the desired power. Fleiss [3] gives tables of the values of m derived by this method for various combinations of α, $1 - \beta$, p_1, and p_2.

If nothing is known about the actual values of p_1 and p_2, then one can obtain a conservative estimate of the required sample size for each group, m, by replacing each of the three products of the form $p(1 - p)$ in the expression for m' by $\frac{1}{4}$[since $p(1 - p) \leq \frac{1}{4}$ for all $0 \leq p \leq 1$].

There are several other methods for determining the required sample sizes (see [4] for a recent review). A computer program that calculates the sample sizes based on the arcsine transformation is included in the NCSS-PASS package [8]. Software for calculating sample sizes for exact tests can be found in StatXact (see www.cytel.com).

References

[1] Brittain, E. & Schlesselman, J.J. (1982). Optimal allocation for the comparison of proportions, *Biometrics* **38**, 1003–1009.

[2] Casagrande, J.T., Pike, M.C. & Smith, P.G. (1978). An improved approximate formula for calculating sample sizes for comparing two binomial distributions, *Biometrics* **34**, 483–486.

[3] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.

[4] Gordon, I. (1994). Sample sizes for two independent proportions: a review, *Australian Journal of Statistics* **36**, 199–209.

[5] Haber, M. (1986). A modified exact test for 2×2 contingency tables, *Biometrical Journal* **28**, 455–463.

[6] Haber, M. (1987). A comparison of some conditional and unconditional exact tests for 2×2 contingency tables, *Communications in Statistics – Simulation and Computation* **16**, 999–1013.

[7] Haber, M. (1992). On the expected significance probability and Bahadur efficiencies of tests for comparing two binomial proportions, *Journal of Statistical Computation and Simulation* **43**, 243–251.

[8] Hintze, J.L. (1991). *NCSS Power Analysis and Sample Size*. Hintze, Kaysville.

[9] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd Ed. Wiley, New York.

[10] McDonald, L.L., Davis, B.M. & Milliken, G.A. (1977). A nonrandomized unconditional test for comparing two proportions in 2×2 contingency tables, *Technometrics* **19**, 145–156.

[11] Suissa, S. & Shuster, J.J. (1985). Exact unconditional sample sizes for 2×2 binomial trials, *Journal of the Royal Statistical Society, Series A* **148**, 317–327.

(*See also* **Binary Data**; **Contingency Table**; **Mantel–Haenszel Methods**; **McNemar Test**)

MICHAEL HABER

Two-mutation Carcinogenesis Model

Stochastic models of carcinogenesis were first proposed in the 1950s [1, 2, 34] to explain the observation that the age-specific incidence curves of many human carcinomas increase roughly with a power of age (*see* **Multistage Carcinogenesis Models**). In the four decades since the introduction of these models our understanding of the processes underlying malignant transformation has increased considerably. Yet the basic assumption – that a malignant tumor arises from a single cell that has sustained a small number of critical insults to its genetic apparatus – on which these early models were predicated remains valid today. From the perspective of carcinogenesis modeling, perhaps the most important insight has been the realization that, in addition to heritable changes to the genome (mutations), the kinetics of cell division and apoptosis (programmed cell death) play an important role in carcinogenesis (*see* **Cell Cycle Models**).

In 1971, on the basis of epidemiologic observations, Knudson proposed a model for retinoblastoma, a rare tumor of the retina in children [16]. In contrast to earlier models of carcinogenesis, Knudson's model took explicit account of cell proliferation kinetics. Subsequent work in molecular biology leaves little doubt that the salient features of Knudson's model for retinoblastoma are correct [3]. Knudson's so-called recessive oncogenesis model has been generalized and applied to analyses of both epidemiologic [14, 18, 25, 31] and experimental data [17, 22, 30]. In its modern garb, this model is often referred to as the two-mutation clonal expansion model. The main goal of this article is a brief discussion of this model.

Although there is much that we do not understand about the biological events underlying carcinogenesis, there are some things we do know. Disruption of normal cell proliferation and differentiation are the *sine qua non* of the malignant state. Conversely, as indicated above, there is accumulating evidence that the kinetics of cell proliferation and differentiation in normal and premalignant cells are important in the carcinogenic process. Increases in cell division rates may lead to increases in the rates of critical mutational events, and an increase in cell division without a compensatory increase in differentiation or death leads to an increase in the size of target cell populations, leading to increased probability of malignant transformation. It is these important aspects of the process of carcinogenesis that the two-mutation clonal expansion model attempts to capture.

The two-mutation clonal expansion model has been used for analyses of time-to-tumor data in epidemiologic and experimental studies (*see* **Tumor Incidence Experiments**), and for analyses of data on the number and size distribution of intermediate lesions on the pathway to malignancy in initiation-promotion experiments. The requisite mathematical and statistical development required for these applications is briefly described.

The Model

The version of the two-mutation model discussed here has been widely used for data analysis. Similar models were considered by Armitage & Doll [2], Neyman & Scott [33] and Kendall [15]. A detailed description of the model can be found in Moolgavkar & Luebeck [26]. The development here follows the development in that paper and uses the same notation. The fundamental biological assumptions are: (i) In any tissue there is a pool of cells susceptible to malignant transformation. This pool is generally identified with the stem cell pool in the tissue of interest and may change in size during life. (ii) Malignant tumors are clonal, i.e. they arise from a single progenitor cell that has become malignantly transformed. (iii) Malignant transformation of a susceptible cell is the result of two specific, rate-limiting, hereditary (at the level of the cell) and irreversible events. For a discussion of the biological interpretation of the two events, see Moolgavkar & Luebeck [26] and Moolgavkar & Knudson [25]. The model also provides a natural framework for the interpretation of initiation and promotion, as discussed in [25] and [26]. Succinctly, the first rate-limiting event is identified with initiation, the second rate-limiting event with progression, and the clonal expansion of initiated cells with promotion.

The following assumptions are used in the mathematical development. Let $X(s)$ represent the number of normal susceptible cells in the tissue of interest at time (age) t, and suppose that intermediate cells, i.e. cells that have sustained the first rate-limiting event on the pathway to malignancy, arise from normal

cells as a nonhomogeneous **Poisson process** with intensity $\nu(s)X(s)$, where $\nu(s)$ is the first event rate. Note that although ν and X are not separately **identifiable**, it is preferable to model the two separately because information on one or the other may be available from independent sources. In the time interval $(s, s+\Delta s)$, an intermediate cell divides into two intermediate cells with probability $\alpha(s)\Delta s + o(\Delta s)$; it dies or differentiates with probability $\beta(s)\Delta(s) + o(\Delta s)$ (note that death and differentiation are equivalent events for carcinogenesis because both events remove the cell from the pool of susceptible cells); it divides into one intermediate cell and one cell that has sustained the second event (malignant cell) with probability $\mu(s)\Delta s + o(\Delta s)$; the probability of more than one event is $o(\Delta s)$. In many applications the parameters are assumed to be constant or piecewise constant. In particular, this implies that the distribution of waiting times to cell division and cell death is assumed to be **exponential**.

Some comments on these mathematical assumptions are in order. The cell kinetics of intermediate cells are modeled in primitive fashion. There are entire tomes on the mathematical modeling of the cell cycle and it is clear that cells do not divide or die with exponential waiting times. Nevertheless, in the context of carcinogenesis modeling these simplifications appear to be entirely appropriate as a first approximation. Once a malignant cell is generated it is assumed to give rise to a detectable tumor after a suitable lag time. This assumption is clearly false and there is clearly a time-to-detection distribution. Furthermore, malignant cells undoubtedly execute a birth–death process and as a consequence become extinct with nonzero probability.

The Hazard Function

Let $Y(t)$ and $Z(t)$, represent the number of intermediate and malignant cells, respectively, at time t and let

$$\Psi(y, z; t) = \sum_{j,k} P_{j,k}(t) y^j z^k$$

be the probability **generating function**, with

$$P_{j,k}(t) = \Pr[Y(t) = j,\ Z(t) = k | Y(0) = 0, Z(0) = 0].$$

Then $(Y(t), Z(t))$ is a **Markov Process**, and Ψ satisfies the Kolmogorov forward differential equation

$$\begin{aligned}\Psi'(y, z; t) = \frac{\partial \Psi(y, z; t)}{\partial t} &= (y-1)\nu(t)X(t)\Psi(y, z; t) \\ &+ \{\mu(t)yz + \alpha(t)y^2 + \beta(t) \\ &- [\alpha(t) + \beta(t) + \mu(t)]y\}\frac{\partial \Psi}{\partial y} \qquad (1)\end{aligned}$$

with initial condition $\Psi(y, z; 0) = 1$ [26] (*see* **Stochastic Processes**). $\Psi(1, 0; t)$ is the survival function for this model, and the **hazard** (incidence) function is given by

$$h(t) = -\frac{\Psi'(1, 0; t)}{\Psi(1, 0; t)}. \qquad (2)$$

It follows immediately from the Kolmogorov equation that

$$\Psi'(1, 0; t) = -\mu(t)\frac{\partial \Psi}{\partial y}(1, 0; t)$$

and thus

$$h(t) = \mu(t)\mathrm{E}[Y(t) | Z(t) = 0] \qquad (3)$$

where E denotes the expectation and we have used the relationship

$$\mathrm{E}[Y(t) | Z(t) = 0] = \frac{\partial \Psi}{\partial y}\frac{1, 0; t}{\Psi(1, 0; t)}.$$

If the probability of tumor is small enough, then $\mathrm{E}[Y(t)] \approx \mathrm{E}[Y(t) | Z(t) = 0]$ and $h(t) \approx \mu(t)\mathrm{E}[Y(t)]$. The differential equation, derived from the Kolmogorov equation, for $\mathrm{E}[Y(t)]$ can be readily solved to yield

$$\begin{aligned}h(t) \approx \mu(t) \int_0^t \Bigg\{ &\nu(s)X(s) \\ &\times \exp \int_s^t [\alpha(u) - \beta(u)]\, \mathrm{d}u \Bigg\}\, \mathrm{d}s. \qquad (4)\end{aligned}$$

This approximate solution sometimes has been used for the analysis of epidemiologic data. It is reasonably accurate when tumors are rare, as in epidemiologic data. However, even with epidemiologic data, this approximation could yield misleading results (see below).

The exact hazard function can be obtained by solving the characteristic equations associated with

the Kolmogorov equation. The survival function is given by

$$\Psi(1,0;t) = \exp \int_0^t [y(u,t) - 1]\nu(u)X(u)\,\mathrm{d}u, \quad (5)$$

where, for each $t > 0$, $y(u,t)$ satisfies the Riccati equation

$$\frac{\mathrm{d}y}{\mathrm{d}u} = -\{\mu(u)yz + \alpha(u)y^2 + \beta(u) - [\alpha(u) + \beta(u) + \mu(u)]y\},$$

with $y(t,t) = 1$.

The hazard function is then given by

$$h(t) = -\frac{\Psi'(1,0;t)}{\Psi(1,0,t)} = -\int_0^t \nu(u)X(u)y_t(u,t)\,\mathrm{d}u, \quad (6)$$

where y_t denotes the derivative of y with respect to t.

Suppose now that $0 = t_0 < t_1 < \cdots < t_k = t$, and suppose that the parameters α, β, and μ are piecewise constant, i.e. on (t_{i-1}, t_i) the parameters are α_i, β_i, and μ_i. Suppose, furthermore, that A_i and B_i are the two roots of the polynomial $\alpha_i x^2 - [\alpha_i + \beta_i + \mu_i]x + \beta_i$. It can be easily shown that $0 < A_i < 1 < B_i$. Then, for $u \in (t_{i-1}, t_i)$, $y(u,t)$ can be defined inductively by

$$y(u,t) = \frac{B_i - A_i \dfrac{y(t_i,t) - B_i}{y(t_i,t) - A_i} \exp[\alpha_i(A_i - B_i)(u - t_i)]}{1 - \dfrac{y(t_i,t) - B_i}{y(t_i,t) - A_i} \exp[\alpha_i(A_i - B_i)(u - t_i)]}, \quad (7)$$

with $y(t_k,t) = 1$. The derivative $y_t(u,t)$ is now straightforward, albeit cumbersome, to compute by repeated use of the chain rule. The equations for $\Psi(1,0;t)$ and $h(t)$ can be integrated using the values of $y(u,t)$ and $y_t(u,t)$ computed above. If $\nu(u)$ is piecewise constant too and if, as is often the case, $X(u)$ is taken to be constant, then, in principle, these equations can be integrated in closed form.

Sometimes, the time-scale of interest is not the age of the animal, or time since start of treatment, but the age of *individual* intermediate clones. Then, $(Y(t), Z(t))$ is not Markovian, and the Kolmogorov differential equation does not exist. A second approach, described in [29] can then be used.

Properties of the hazard function are discussed in detail in [11] and [29]. See also [36]. A brief summary is given here. For a general class of multistage models, the hazard function for the kth malignant transformation (given that $k-1$ malignant transformations have already occurred) is given by an expression that is analogous to (3) above:

$$h_k(t) = \mu_n(t)\mathrm{E}[Y_{n-1}(t)|Z(t) = k-1],$$

where $Y_{n-1}(t)$ represents the number of cells in the penultimate stage on the pathway to malignancy and $\mu_n(t)$ is the last mutation rate, i.e. the rate of transition from the penultimate stage into the malignant stage [6]. Then $h_1(t)$ is just the usual hazard function. It can be quite easily shown that, for any k, $h_k(t) > h_{k-1}(t)$. This inequality implies, for example, that the hazard function for a second malignancy is higher than that for the first. This phenomenon has been observed and attributed to biological changes (decrease in resistance) associated with the first malignancy. While this may well be true, it is worth remembering that an increased hazard for a second malignant tumor is a logical consequence of a multistage process.

The approximate expression $h(t) = \mu_n(t)\mathrm{E}[Y_{n-1}(t)]$ has often been used. The Armitage–Doll approximation, for example, retains only the first nonzero term in the Taylor series expansion of this approximate expression [23]. While replacing the conditional expectation in the exact hazard function with the unconditional expectation would appear, at first glance, to lead to a good approximation if the probability of tumor is low, the qualitative behavior of the approximate hazard function is quite different. For example, consider exposure to an environmental carcinogen that increases the hazard by affecting one or more of the parameters of the model. When exposure stops, the exact hazard function eventually approaches the background hazard, whereas the approximate hazard function never returns to background levels [11, 18, 29]. This has important implications for interpretation of epidemiologic data. For example, the relative risk of lung cancer among smokers who quit declines with time since smoking cessation (*see* **Smoking and Health**). It is not

generally appreciated that this phenomenon, which is often attributed to repair, is predicted by multistage carcinogenesis.

For **likelihood** construction and applications to analyses of time-to-tumor data, the reader is referred to the original papers [22, 28, 32]. Identifiability of parameters of the two-mutation clonal expansion model is discussed in publications [8–11]. Likelihoods for case–control studies are discussed in [13, 24]. The impact of covariate measurement errors on parameter estimates is discussed in [12]. Extensions of the model to incorporate more than two mutations can be found in [18, 20, 27, 36, 37].

Number and Size Distribution of Intermediate Lesions

In the two-mutation clonal expansion model, clones of cells in the intermediate compartment can be identified with premalignant lesions that arise in initiation–promotion experiments, which are typically done with the mouse skin or the rodent liver as the target organs. In the rodent liver the lesions of interest are microscopic foci that exhibit typical enzyme alterations characterized by specific stains. The number and sizes of these altered foci (in two-dimensional sections) are generally reported as functions of the doses of agents of interest and time since beginning of exposure. The requisite mathematical quantities for analyses of such lesions have been developed [7, 19], and used for analyses of substantial data sets (e.g. [21, 26], and [32]). Since the observations are made in two-dimensional sections of three-dimensional objects (the foci) stereologic considerations play an important role in the analyses (see, for example, [5] and [35] and the references above). In the mouse skin system, the intermediate lesions are papillomas. These lesions are directly observable, so that the animal does not have to be sacrificed for the data to be collected. Typically, several observations are made on the same animal over the course of the experiment. This leads to correlated longitudinal observations. The reader is referred to the articles by Kopp-Schneider & Portier [17] and Dewanji et al. [4] for more details.

References

[1] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

[2] Armitage, P. & Doll, R. (1957). The two-stage theory of carcinogenesis in relation to the age distribution of human cancers, *British Journal of Cancer* **11**, 161–169.

[3] Cavanee, W.K., Dryja, T.P., Phillips, R.A., Benedict, W.F. & Godbout, R. (1983). Expression of recessive alleles by chromosomal mechanisms in retinoblastoma, *Nature* **305**, 779–784.

[4] Dewanji, A., Goddard, M., Krewski, D. & Moolgavkar, S.H. (1999). Two stage model for carcinogenesis: number and size distributions of premalignant clones in longitudinal studies, *Mathematical Biosciences* **155**, 1–12.

[5] Dewanji, A., Luebeck, E.G. & Moolgavkar, S.H. (1996). A biologically based model for the analysis of premalignant foci of arbitrary shape, *Mathematical Biosciences* **135**, 55–68.

[6] Dewanji, A., Moolgavkar, S.H. & Luebeck, E.G. (1991). Two-mutation model for carcinogenesis: joint analysis of premalignant and malignant lesions, *Mathematical Biosciences* **104**, 97–109.

[7] Dewanji, A., Venzon, D.J. & Moolgavkar, S.H. (1989). A stochastic two-stage model for cancer risk assessment II. The number and size of premalignant clones, *Risk Analysis* **9**, 179–187.

[8] Hanin, L.G. & Yakovlev, A.Yu. (1996). A nonidentifiability aspect of the two-stage model of carcinogenesis, *Risk Analysis* **16**, 711–715.

[9] Hazelton, W.D., Luebeck, E.G., Heidenreich, W.F. & Moolgavkar, S.H. (2001). Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette, and pipe smoke exposures using the biologically-based two-stage clonal expansion model. *Radiation Research* **156**, 78–94.

[10] Heidenreich, W.F. (1996). On the parameters of the clonal expansion model, *Radiation and Environmental Biophysics* **35**, 127–129.

[11] Heidenreich, W.F., Luebeck, E.G. & Moolgavkar, S.H. (1997). Some properties of the hazard function of the two-mutation clonal expansion model, *Risk Analysis* **17**, 391–399.

[12] Heidenreich, W.F., Luebeck, E.G. & Moolgavkar, S.H. (2004). Effects of exposure uncertainties in the TSCE model and application to the Colorado miners data, *Radiation Research* **161**, 72–81.

[13] Heidenreich, W.F., Wellmann, J., Jacob, P. & Wichmann, H.E. (2002). Mechanistic modelling in large case–control studies of lung cancer from smoking, *Statistics in medicine* **21**, 3055–3070.

[14] Kai, M., Luebeck, E.G. & Moolgavkar, S.H. (1997). Analysis of solid cancer incidence among atomic bomb survivors using a two-stage model of carcinogenesis, *Radiation Research* **148**, 348–358.

[15] Kendall, D.G. (1960). Birth-and-death processes, and the theory of carcinogenesis, *Biometrika* **47**, 13–21.

[16] Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma, *Proceedings of the National Academy of Sciences* **68**, 820–823.

[17] Kopp-Schneider, A. & Portier, C.J. (1992). Birth and death/differentiation rates of papillomas in mouse skin, *Carcinogenesis* **13**, 973–978.

[18] Little, M.P. (1995). Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon and Knudson, and of the multistage model of Armitage and Doll, *Biometrics* **51**, 1278–1291.

[19] Luebeck, E.G. & Moolgavkar, S.H. (1991). Stochastic analysis of intermediate lesions in carcinogenesis experiments, *Risk Analysis* **11**, 149–157.

[20] Luebeck, E.G., Moolgavkar, S.H. (2002). Multistage carcinogenesis and the incidence of colorectal cancer, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15095–15100.

[21] Luebeck, E.G., Moolgavkar, S.H., Buchmann, A. & Schwarz, M. (1991). Effects of polychlorinated biphenyls in rat liver: quantitative analysis of enzyme altered foci, *Toxicology and Applied Pharmacology* **111**, 469–484.

[22] Luebeck, E.G., Curtis, S.B., Cross, F.T. & Moolgavkar, S.H. (1996). Two-stage model of radon-induced malignant lung tumors in rats: effects of cell killing, *Radiation Research* **145**, 163–173.

[23] Moolgavkar, S.H. (1991). Stochastic models of carcinogenesis, in *Handbook of Statistics*, Vol. 8, C.R. Rao & R. Chakraborty, eds. Elsevier, Amsterdam, pp. 373–393.

[24] Moolgavkar, S.H. (1995). When and how to combine results from multiple epidemiological studies in risk assessment, in *The Proper Role of Epidemiology in Regulatory Risk Assessment*, John Graham, ed. Elsevier, New York, 77–90.

[25] Moolgavkar, S.H. & Knudson, A.G. (1981). Mutation and cancer: a model for human carcinogenesis, *Journal of the National Cancer Institute* **66**, 1037–1052.

[26] Moolgavkar, S.H. & Luebeck, E.G. (1990). Two event model for carcinogenesis: biological, mathematical and statistical considerations, *Risk Analysis* **10**, 323–341.

[27] Moolgavkar, S.H. & Luebeck, E.G. (1992). Multistage carcinogenesis: population-based model for colon cancer, *Journal of the National Cancer Institute* **84**, 610–618.

[28] Moolgavkar, S.H., Cross, F.T., Luebeck, G. & Dagle, G.E. (1990). A two-mutation model for radon-induced lung tumors in rats, *Radiation Research* **121**, 28–37.

[29] Moolgavkar, S.H., Dewanji, A. & Venzon, D.J. (1988). A stochastic two-stage model for cancer risk assessment. I: The hazard function and the probability of tumor, *Risk Analysis* **8**, 383–392.

[30] Moolgavkar, S.H., Luebeck, E.G., Buchmann, A. & Bock, K.W. (1996). Quantitative analysis of enzyme-altered liver foci in rats initiated with diethylnitrosamine and promoted with 2,3,7,8-tetrachlorodibenzo-p-dioxin or 1,2,3,4,6,7,8-heptachlorodibenzo-p-dioxin, *Toxicology and Applied Pharmacology* **138**, 31–42.

[31] Moolgavkar, S.H., Luebeck, E.G., deGunst, M., Port, R.E. & Schwarz, M. (1990). Quantitative analysis of enzyme altered foci in rat hepatocarcinogenesis experiments, *Carcinogenesis* **11**, 1271–1278.

[32] Moolgavkar, S.H., Luebeck, E.G., Krewski, D. & Zielinski, J.M. (1993). Radon, cigarette smoke and lung cancer: a re-analysis of the Colorado Plateau uranium miners' data, *Epidemiology* **4**, 204–217.

[33] Neyman, J. & Scott, E. (1967). Statistical aspects of the problem of carcinogenesis, in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. LeCam & J. Neyman, eds. University of California Press, Berkeley, pp. 745–776.

[34] Nordling, C.O. (1953). A new theory of the cancer inducing mechanism, *British Journal of Cancer* **7**, 68–72.

[35] Nychka, D., Wahba, G., Goldfarb, S. & Pugh, T. (1984). Cross validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two-dimensional cross-sections, *Journal of the American Statistical Association* **79**, 832–846.

[36] Tan, W.Y. (1991). *Stochastic Models of Carcinogenesis*. Marcel Dekker, New York.

[37] Tan, W.Y. (2002). *Stochastic Models with Applications to Genetics, Cancers, AIDS and other Biomedical Systems*. World Scientific Publishing. New Jersey.

(*See also* **Serial-sacrifice Experiments**; **Tumor Growth**)

SURESH H. MOOLGAVKAR

Two-phase Sampling

Historical Background

Suitably enough, the theory of two-phase sampling was created by Jerzy Neyman [8] in response to a problem posed at a conference on sampling human populations in April, 1937. Neyman [8] introduces his solution "by describing the problem in much the same form as it was stated to me, without using any mathematical symbols": Simply put, a field survey is to be undertaken to determine the average value of some character of a population; for example, the amount of money families spend on food. As the collection of data requires long interviews by specially trained enumerators, the cost per family is quite high. The cost of the survey is constrained within a specified amount but the sample does not appear to yield an estimate of desired precision because of the great variability of the character. Nevertheless, the character is correlated with a second character that can be determined at a lower cost per family so that a precise estimate of the distribution of this second character is readily obtained. Hence, a more precise estimate of the original character can be found by first estimating the distribution of the second character alone from a large random sample, then dividing this sample, as in stratified sampling, into classes or strata according to the value of the second character and to draw at random from each of the strata a small sample for the costly procedure of measuring the first character.

Neyman [8] called this method *double sampling*, and this term remains in use among statisticians working in the area of quality control and assurance. Survey statisticians, however, tend instead to use the term *two-phase sampling* so that this method is distinguished from *two-stage sampling*. Two-phase sample designs differ from two-stage sample designs in that the stratification occurs after the first sample is collected (i.e. *post hoc*) in the case of two-phase designs rather than before the first sample is collect (i.e. *ante hoc*) in the case of two-stage designs. It is understandably regrettable and a source of confusion that the biometrics literature refers to a two-phase survey sampling design as a two-stage design, as noted by Whittemore [13].

In the case of **genetic epidemiology**, the interest lies primarily with the estimation of means associated with Bernoulli distributed random variables, such as disease prevalence and allele frequencies (*see* **Gene**), rather than economic variables. Nevertheless, two-phase sampling designs are applicable to quantitative phenotypes.

Formulation and Allocation

Assume a finite population size N of which n' are to be selected by a simple random sample without replacement at the first phase of the design. Suppose that there are determined to be K strata with (possibly unknown) population size N_h for the hth stratum. Suppose n'_h sample units are observed to be in the hth stratum in the first phase sample with a sample mean of $\overline{y}_h$. Let $\overline{Y}_h$ denote the population mean and S_h^2 the population variance for the hth stratum. Let s_h^2 denote the usual **unbiased** estimator of S_h^2 based on the n'_h sample units in the first phase. Let n_h denote the sample size for the hth stratum at the second phase. For convenience, let $W_h = N_h/N$, $w_h = n'_h/n'$, and $\nu_h = n_h/n'_h$ with $0 < \nu_h \leq 1$ for all h. Assuming that $\Pr(n'_h = 0) = 0$ for all h, Rao [9] showed that an unbiased estimator of the population mean $\overline{Y} = \Sigma W_h \overline{Y}_h$ is $\overline{y} = \Sigma w_h \overline{y}_h$ with **variance**

$$\operatorname{var}(\overline{y}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \sum_{h=1}^{K} \frac{W_h S_h^2}{n'} \left(\frac{1}{\nu_h} - 1\right),$$

where S^2 is the population variance. These results are available elsewhere in the literature, but Rao [9] obtained them under the assumption that the second-phase sample sizes $\{n_h\}$ for the strata are random variables, unlike Cochran [1] who assumed that they are fixed values. Rao [9] further showed that a nonnegative unbiased estimator of $\operatorname{var}(\overline{y})$ is

$$v(\overline{y}) = \frac{1}{Nn'} \left[\left(\frac{N-1}{n'-1}\right) \sum_{h=1}^{K} n'_h s_h^2 \left(\frac{1}{\nu_h} - 1\right) + \left(\frac{N-n'}{n'-1}\right) \left(\sum_{h=1}^{K} \frac{1}{\nu_h} \sum_{j=1}^{n_h} y\{h_j\}^2 - n' \overline{y}^2 \right) \right],$$

provided n' is sufficiently large so that $\Pr(n_n \geq 2) = 1$ for all h. Särndal & Swensson [10] showed that the result for $v(\overline{y})$ continues to be valid if K is a random variable or if there is random nonresponse at the second phase described by a Bernoulli distribution with a fixed but unknown probability of inclusion

within each stratum with the possibility that the probability of inclusion varies among strata.

With respect to the optimal allocation of the first-phase sample size n' and the second-sample sampling fractions $\{\nu_h\}$, the cost function is taken as

$$C = n'c' + \sum_{h=1}^{K} n_h c_h,$$

where c' is usually much smaller than c_h. Since C is a random variable, we take

$$C^* = \mathrm{E}(C) = n' \left(c' + \sum_{h=1}^{K} W_h c_h \nu_h \right).$$

From the Cauchy inequality, the optimal ν_h for the hth stratum for given C^* and $\mathrm{var}(\overline{y})$ is

$$\nu_h = S_h \left[\frac{c'}{c_h \left(S^2 - \sum W_h S_h^2 \right)} \right]^{1/2}.$$

As noted by Singh & Singh [12], it is important to realize that the upper limit on the second-phase sample size is n'_h if randomly sampled without replacement from the first-phase sample. As suggested by Rao [9] in this case for which $\nu_h > 1$, set the corresponding $\nu_h = 1$ and repeat the procedure until all the $\nu_h \leq 1$.

By Rao [9], if the strata weights $\{W_h\}$ are not known, then the subsampling fraction $\nu_h = n_h/n'_h$ varies as a function of the observed value of n'_h. Nevertheless, in this case, replace W_h by its estimate w_h.

Bayesian Approaches

Draper & Guttman [3] assumed that the observations $\{y_{hj}\}$ are normally distributed with independent improper prior distributions for the mean μ_h and variance σ_h^2 of the hth stratum given by

$$p(\mu_h)\,\mathrm{d}\mu_h \propto \mathrm{d}\mu_h, \qquad p(\sigma_h^2)\,\mathrm{d}\sigma_h^2 \propto \frac{\mathrm{d}\sigma_h^2}{\sigma_h^2}.$$

Draper & Guttman [3] further assumed that C, K, and $\{n'_h\}$ are fixed with $\Sigma n_h c_h < C$ but with the prior information concerning the means and variances of the strata available before the first phase and showed that the posterior distribution of

$$T_h = \frac{(n'_h + n_h - 1)(\mu_h - \tilde{\mu}_h)}{\tilde{\sigma}_h}$$

is student's t with $n'_h + n_h - 1$ degrees of freedom where, if $\{x_{hj}\}$ denotes the observations from the first phase,

$$\tilde{\mu}_h = \frac{n'_h \overline{x}_h + n_h \overline{y}_h}{n'_h + n_h},$$

and

$$\tilde{\sigma}_h^2 = \frac{1}{n'_h + n_h} \left[\left(\frac{n'_h n_h}{n'_h + n_h} \right) (\overline{x}_h - \overline{y}_h)^2 + (n'_h - 1)s_h^2 + (n_h - 1)t_h^2 \right]$$

with $\overline{x}_h$ and s_h^2 the usual unbiased estimators of the mean and variance of the h stratum from the first phase and $\overline{y}$ and t_h^2, respectively, from the second phase. Furthermore, Draper & Guttman [3] showed that the posterior distribution of $(n_h - 1)s_h^2/\sigma_h^2$ is $\chi^2_{n'_h - 1}$. From these results concerning the posterior distributions, the posterior exception of $\mu = \Sigma_h W_h \mu_h$ is

$$\sum_{h=1}^{K} W_h^2 \frac{(n'_h - 1)s_i^2}{(n'_h + n_h)(n'_h - 3)}.$$

Choosing the sample size n_h for the h stratum at the second phase subject to $n_h \geq 0$ for all h leads to

$$n_h = \frac{C}{c_h} q_h - n'_h,$$

where

$$q_h = \frac{\left(\frac{n_h - 1}{n_h - 3} \right)^{1/2} W_h s_h \sqrt{c_h}}{\sum_h \left(\frac{n_h - 1}{n_h - 3} \right)^{1/2} W_h s_h \sqrt{c_h}}.$$

There is the possibility that this allocation rule will lead to negative n_h for some strata. This merely indicates that the hth statum has been oversampled. Draper & Guttman [3] discussed an algorithmic adjustment to the optimal allocation rule to compensate for this.

If, on the other hand, the posterior after the first phase is used to provide the prior for the second

phase, then by Draper & Guttman [3], the optimal allocation rule becomes instead $n_h \propto Cq_h/c_h$. Compare this with the optimal allocation rule

$$C\frac{W_h S_h}{\sum_h W_h S_h}$$

of Neyman [8] assuming $\{W_h\}$ are known with n' and $\sum_h n_h$ fixed. This so-called *Neyman allocation* can also be obtained from the expression derived by Rao [9] for $\text{var}(\overline{y})$ using the Lagrange multiplier.

Draper & Guttman [3] also considered the situation when the strata weights $\{W_h\}$ are no longer known but rather follow a Dirichlet prior distribution with parameter v_h corresponding to the hth stratum. In this case, the answer is the same as the case for $\{W_h\}$ known except that the unbiased estimator

$$\tilde{w}_h = \frac{n'_h + v_h}{\sum_h (n'_h + v_h)}$$

replaces W_h everywhere. Note that Jeffrey's prior coincides with the uniform prior (all $v_h = 1$).

For a **multivariate normal** extension to this approach, see Draper & Guttman [4]. Although the approach of Draper & Guttman [3, 4] is suitable for quantitative phenotypes with a normally distributed **likelihood**, it is not suitable for estimation of disease prevalence or allele frequencies for which a solution is given by Zacks [15] assuming a hypergeometric likelihood and a discrete uniform prior distribution for the number of successes out of the number of trials for the hth stratum. For a heterogeneous situation in which the prevalence varies among strata, see the optimal allocation rules of Newbold [7] which assumes the invariant Jeffreys' prior distribution

$$p(P_h) \propto P_h^{-1/2}(1 - P_h)^{-1/2}$$

and a binomial likelihood for the parallel cases comparable to those of Draper & Guttman [3].

Prevalence Estimation and Practical Considerations

For a Bayesian solution to the problem of estimation of prevalence in a two-phase sampling design using a **Markov chain Monte Carlo** method for a Dirichlet conjugate prior distribution for **sensitivity**, **specificity**, and prevalence jointly with a beta posterior distribution, see Erkanli et al. [5].

On the other hand, the results of Neyman [8] and Rao [9] do not assume a distributional form for the likelihood and thus apply to the problem of estimation of disease prevalence. While these results do assume a finite population, the hypothetical case of an infinite population is easily derived as a limiting case.

As discussed by Deming [2], a two-phase design is not necessarily more efficient than a one-phase design, nor is Neyman allocation necessarily more efficient than *proportional allocation*: $n_h \propto w_h$ for all h.

Calculations in Deming [2], suggest that, as a rule-of-thumb, it is only when the ratio of interview cost per sampling unit at the second phase compared with screening cost per sampling unit at the first phase exceeds $6:1$ that two-phase sampling will be more advantageous. Note that the ratio is likely to be high when the screening and stratification is done on the basis of records, typically on the order of $40:1$ or $100:1$ according to Deming [2].

A sample design using Neyman allocation that incorporates an estimate of the proportion of false negatives that is wide of the mark may well yield an estimate of the prevalence with greater variance than the estimate by proportional sampling. Deming [2] also noted that it is easy "to fall into the trap in the planning stages by putting unwarranted credence into an advance estimate" of the proportion of false positives when in fact a large sample or a long history of usage of the exact plan of screening is required. The example of the heavy workload encountered by a psychiatrist interviewing 30 subjects for a pilot study is cited despite the fact that the estimate of a small proportion of false positives in such a small sample is subject to a wide standard error. Whereas, a fairly large preliminary sample will often reveal problems that one would not otherwise foresee, for example, a set of admission records intended to contain individuals only aged 21 to 60 but actually including admissions of age 20 and under.

For a discussion of two-phase sampling designs in the context of prevalence for a rare disease for which all those screened positive in the first phase must ethically be included in the second phase, see Shrout & Newman [11].

For a discussion concerning the estimation of disease prevalence with nonresponse at the second phase, see Särndal & Swensson [10], and Gao et al.

[6] for a representation incorporating a logistic model for nonresponse that is not completely random.

A **maximum likelihood** approach for the multinomial distribution with discussion of options involving the **EM algorithm** and the **bootstrap method** are given in Zhou et al. [16] together with a **likelihood ratio test** for the null hypothesis of completely random nonresponse.

Multistage Sampling in Genetic Epidemiology

One of the greatest challenges to successfully concluding a **disease–marker association** study is heterogeneity in the distribution of alleles among races, ethnic and regional groups (*see* **Bias in Case–Control Studies**). For example, cystic fibrosis (CF) can be caused by many different mutations. The most common mutation in the North American non-Ashkenazi population is $\Delta 508$. The proportion of CF genes that are $\Delta 508$ varies widely among different countries, within a country, and among different ethnic and racial groups. But in the case of CF, these observations were noted after the gene was successfully cloned.

A case–control study using a two-phase design is discussed by Whittemore & Halpern [14] in which men were asked whether they were diagnosed with prostate cancer and whether they had a first degree male relative with prostate cancer at the first phase. The subjects are stratified according to diagnosis and family history in preparation for second-phase sampling. The parameters of interest in this study were prostate cancer hazard rates in carriers and noncarriers and the probability that an arbitrary allele contains a deleterious **mutation**. The study budget could accommodate a second-phase sample of size 570 from the first-phase sample of size 1500. Calculations showed that Neyman allocation of 570 sampling units resulted in little loss of efficiency compared with a complete one-phase sample of 3000 with respect to the variances of the three parameter estimators.

With respect to the theoretical discussion in Whittemore & Halpern [14], the use of the Horvitz–Thompson estimating equation for multiphase sampling is treated in greater detail in Whittemore [13] where it is noted that although it can yield estimates less efficient than the maximum likelihood estimates, substantial efficiency loss appears to occur chiefly when multiphase sampling is unnecessary.

References

[1] Cochran, W.G. (1963). *Sampling Techniques*, 2nd Ed. Wiley, New York.

[2] Deming, W.E. (1977). An essay on screening, or on a two-phase sampling, applied to surveys of a community, *International Statistical Review* **45**, 29–37.

[3] Draper, N.R. & Guttman, I. (1968). Some Bayesian stratified two-phase sampling results, *Biometrika* **55**, 131–139.

[4] Draper, N.R. & Guttman, I. (1968). Bayesian stratified two-phase sampling results: *k* characteristics, *Biometrika* **55**, 587–589.

[5] Erkanli, A., Soyer, R. & Stangl, D. (1997). Bayesian inference in two-phase prevalence studies, *Statistics in Medicine* **16**, 1121–1133.

[6] Gao, S., Hui, S.L., Hall, K.S. & Hendrie, H.C. (2000). Estimating disease prevalence from two-phase surveys with non-response at the second phase, *Statistics in Medicine* **19**, 2101–2114.

[7] Newbold, P. (1971). Optimum allocation in stratified two-phase sampling for proportions, *Biometrika* **58**, 587–589.

[8] Neyman, J. (1938). Contributions to the theory of sampling human populations, *Journal of the American Statistical Association* **33**, 101–116.

[9] Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys, *Biometrika* **60**, 125–133.

[10] Särndal, C.-E. & Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse, *International Statistical Review* **55**, 279–294.

[11] Shrout, P.E. & Newman, S.C. (1989). Design of two-phase prevalence surveys of rare disorders, *Biometrics* **45**, 549–555.

[12] Singh, B.D. & Singh, D. (1965). Some remarks on double sampling for stratification, *Biometrika* **52**, 587–590.

[13] Whittemore, A.S. (1997). Multistage sampling designs and estimating equations, *Journal of the Royal Statistical Society, Series B* **59**, 589–602.

[14] Whittemore, A.S. & Halpern, J. (1997). Multi-stage sampling in genetic epidemiology, *Statistics in Medicine* **16**, 153–167.

[15] Zacks, S. (1970). Bayesian design of single and double stratified sampling for estimating proportion in finite population, *Technometrics* **12**, 119–130.

[16] Zhou, X.-H., Castelluccio, P., Hui, S.L. & Rodenberg, C.A. (1999). Comparing two prevalence rates in a two-phase design study, *Statistics in Medicine* **18**, 1171–1182.

(*See also* **Prevalence**)

K.J. KEEN

Two-stage Least Squares Regression

This procedure is widely used in econometrics where ordinary **least squares** regression (OLS) gives inconsistent estimators (*see* **Consistent Estimator**) for systems of simultaneous equations if some **explanatory variables** are correlated with errors. Two-stage least squares (2SLS) was proposed by Theil [5] to overcome this difficulty. A description for extensive systems is given by Amemiya [1] but the principle may be illustrated by a two-equation econometrics model. This is discussed more fully by Johnston [3].

If y_t represents consumption expenditure, x_t income and z_t nonconsumption expenditure in period t, where z_t (often called an exogenous or instrumental variable) is determined by external influences such as taxation, interest rates, savings, or mortgage contracts, it may be reasonable to model income and expenditure by the equations

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{1}$$

$$x_t = y_t + z_t. \tag{2}$$

In more general cases, (2) may contain coefficients and an error term.

We assume that the ε_t for different t are independent and that for all t, $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{var}(\varepsilon_t) = \sigma^2$, and that z_t is independent of ε_t. Then, using (1) to eliminate y_t from (2) gives

$$x_t = \frac{\beta_0}{1-\beta_1} + \frac{z_t}{1-\beta_1} + \frac{\varepsilon_t}{1-\beta_1},$$

whence

$$\mathrm{E}(x_t) = \frac{\beta_0}{1-\beta_1} + \frac{z_t}{1-\beta_1}$$

and

$$\mathrm{E}\{\varepsilon_t[x_t - \mathrm{E}(x_t)]\} = \frac{\sigma^2}{1-\beta_1},$$

implying that ε_t and x_t are correlated. The OLS estimate of β_1, obtained from (1) is easily shown to be inconsistent.

In 2SLS, the first stage is to form the OLS estimators, say p_0 and p_1, of intercept and slope, for the regression of x_t on z_t and use this fitted regression to give estimators $\hat{x}_t$ of x_t. In the second stage, these estimators are inserted in (1) in place of x_t, and OLS is applied to the amended equation to estimate β_1. In practice, the two stages may be telescoped and a direct estimate of β_1 obtained as $b = s_{yz}/s_{xz}$, where s_{yz} and s_{xz} denote the usual sums of products of deviations from the mean. Estimators with this structure also occur in procedures described by Barnett [2] which involve instrumental variables in models that allow for errors in explanatory variables (*see* **Errors in Variables**).

Applications outside econometrics are not common, but Permutt & Hebel [4] applied 2SLS using a slightly different model to results of a clinical trial designed to estimate the effect of smoking by pregnant women upon birth weight of offspring. The smokers were randomly allocated to a control and treatment group, and in the latter group intervention measures to discourage smoking were used. Although the original analysis indicated higher birth weights among the treatment group, it did not give an accurate measure of birth weight gain associated with giving up smoking because, despite the intervention, not all women in the treatment group gave up smoking. The authors proposed and justified an alternative model:

$$B = \beta_{01} + \beta_{11} S + \varepsilon_1,$$

$$S = \beta_{02} + \beta_{12} I + \varepsilon_2,$$

where S is the number of cigarettes smoked per day during the eighth month of pregnancy, B the birth weight and I an indicator variable for control or treatment group. Although B does not occur explicitly in the second equation, I, which is related to B, does. A 2SLS analysis was recommended on the basis of evidence that ε_1 and ε_2 were correlated for some individuals because heavy smokers with the same intervention status may tend systematically to have lighter (or perhaps heavier) offspring for reasons other than smoking. This introduces correlation between S and ε_1. Details of the analysis using 2SLS are given in [4].

References

[1] Amemiya, T. (1988). Two-stage least squares, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York.

[2] Barnett, V.D. (1969). Simultaneous pairwise linear structural relationships, *Biometrics* **25**, 129–142.

[3] Johnston, J. (1972). *Econometric Methods*, 2nd Ed. McGraw-Hill, New York.

[4] Permutt, T. & Hebel, J.R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight, *Biometrics* **45**, 619–622.

[5] Theil, H. (1957). Specification errors and the estimation of economic relationships, *Review of the International Statistical Institute* **25**, 41–51.

(*See also* **Structural Equation Models**)

P. SPRENT

Type-specific Covariates in Survival Analysis

The presentation will be in terms of the **Cox regression model** [3] for survival data, though the concept of type-specific covariates is equally relevant for other regression models.

The Cox regression model in its simplest form states that the **hazard** function $\alpha(t)$ for an individual with covariates $\mathbf{Z}$ has the form

$$\alpha(t) = \alpha_0(t) \exp(\boldsymbol{\beta}'\mathbf{Z}), \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of unknown regression coefficients and the baseline hazard, $\alpha_0(t)$, is an unknown and unspecified hazard function for individuals with covariates $\mathbf{Z} = 0$.

In some cases, however, individuals may experience several types ($h = 1, \ldots, k$) of events, the intensities of which may all be of scientific interest. Examples include the **competing risks** model, where $h = 1, \ldots, k$ refers to different causes of failure, and the Markov illness–death model for a chronic disease, with states $0 =$ healthy, $1 =$ diseased, and $2 =$ dead, and where the types of events may be $h = 01$ (occurrence of disease), $h = 02$ (death without disease), and $h = 12$ (death while diseased) (*see* **Stochastic Processes**). In such cases, an immediate extension of the model (1) would be

$$\alpha_h(t) = \alpha_{h0}(t) \exp(\boldsymbol{\beta}'_h\mathbf{Z}) \tag{2}$$

for events of type h. If, however, some regression coefficients are the same for different types of events, then it is often more convenient to define *type-specific covariates* $\mathbf{Z}_h$, $h = 1, \ldots, k$, and formulate the model as

$$\alpha_h(t) = \alpha_{h0}(t) \exp(\boldsymbol{\beta}'\mathbf{Z}_h) \tag{3}$$

(see, for example, Andersen et al. [1, Section VII.1]).

To illustrate this, consider an example of a competing risks model with two causes of death (A and B) and with covariates sex (S), age (A), and treatment (T). Suppose that S has different effects on causes A and B, A has the same effect on causes A and B, while T only affects A. Then the model formulated as (2) would be

$$\alpha_{\mathrm{A}}(t) = \alpha_{\mathrm{A}0}(t) \exp(\beta_{\mathrm{A}1}S + \beta_{\mathrm{A}2}A + \beta_{\mathrm{A}3}T)$$

and

$$\alpha_{\mathrm{B}}(t) = \alpha_{\mathrm{B}0}(t) \exp(\beta_{\mathrm{B}1}S + \beta_{\mathrm{B}2}A),$$

where $\beta_{\mathrm{A}2} = \beta_{\mathrm{B}2}$. To write the same model in the form (3), type-specific covariates for A and B are defined as follows: for cause A, let

$$Z_{\mathrm{A}1} = S, \qquad Z_{\mathrm{A}2} = 0, \qquad Z_{\mathrm{A}3} = A, \qquad Z_{\mathrm{A}4} = T,$$

and for cause B, let

$$Z_{\mathrm{B}1} = 0, \qquad Z_{\mathrm{B}2} = S, \qquad Z_{\mathrm{B}3} = A, \qquad Z_{\mathrm{B}4} = 0.$$

Then the covariate vector $\boldsymbol{\beta}$ in (3) is

$$\boldsymbol{\beta}' = (\beta_{\mathrm{A}1}, \beta_{\mathrm{B}1}, \beta_{\mathrm{A}2} = \beta_{\mathrm{B}2}, \beta_{\mathrm{A}3}).$$

To test the hypothesis $\beta_{\mathrm{A}1} = \beta_{\mathrm{B}1}$ of identical effects of S for causes A and B, one may replace the model above with one where both $(Z_{\mathrm{A}1}, Z_{\mathrm{A}2})$ and $(Z_{\mathrm{B}1}, Z_{\mathrm{B}2})$ are replaced by S. It is seen that, to formulate the model using type-specific covariates, one has to define, for all types $h = 1, \ldots, k$, a p-dimensional covariate vector where p is the total number of regression coefficients to be estimated.

Type-specific covariates provide a flexible means for analyzing multistate survival regression models. Since model (3) is formally identical to a stratified Cox regression model, estimates may be obtained using standard software for this model. Thus, to analyze the model exemplified above based on competing risks survival data, we create a new input file by duplicating each patient's data in the following way: in version A, we include the survival/censoring time, the indicator for failure of type A, and the type-specific covariate vector $\mathbf{Z}_{\mathrm{A}}$, and set the stratum variable to A; in version B, we include the survival/censoring time, the indicator for failure of type B, and the type-specific covariate vector $\mathbf{Z}_{\mathrm{B}}$, and set the stratum variable to B. An example of this is described in detail by Andersen & Keiding [2].

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[2] Anderson, P.K. & Keiding, N (2002). Multi-state models for event history analysis, *Statistical Methods in Medical Research* **11**, 91–115.

[3] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

(*See also* **Survival Analysis, Overview**)

PER KRAGH ANDERSEN

Unbiasedness

The term *unbiasedness* is used in somewhat different senses in the theories of **estimation** and **hypothesis testing**, with a unified interpretation in **decision theory**.

Unbiased Estimation

There is usually more than one way to estimate a parameter. Applying alternate estimation procedures to the same data set, it is possible to obtain distinct estimates of the same quantity. Making an informed choice among such estimates may be difficult or impossible if the choice is to be made simply on the basis of the estimates themselves, particularly if little is known about the quantity being estimated (the *estimand*). The choice must be governed by the typical performance of the estimators that produced the estimates.

It is helpful to imagine a large collection of hypothetical data sets, replicates obtained under essentially similar conditions to those which generated the data at hand. In application to these, a single estimator produces a collection of replicate estimates of the same estimand; summaries of these hypothetical replicates provide means for evaluating the performance of the estimator.

In particular, it is desirable that the average of the values produced by an estimator be close to the value of the estimand. The difference between the average value of an estimator and the estimand is called *bias*; an estimator is said to be unbiased if its bias is zero.

The existence of an unbiased estimator depends, in varying degrees, on the parameter being estimated and on the nature of the data. We illustrate this in the following examples, supposing that $X_1, X_2, \ldots, X_n$ is a sample of random variables with common distribution function $F(x) = \Pr(X \leq x)$.

Example 1

If F has a finite mean μ, then the sample mean $\overline{X} = (1/n)\sum_i X_i$ is unbiased for μ, without making any assumptions on the parametric form of F, or even whether the Xs are independent.

Example 2

If F has a finite variance σ^2, then the sample variance $S^2 = \sum(X_i - \overline{X})^2/(n-1)$ is unbiased for σ^2 provided that the Xs are independent. However, if the Xs are not independent, S^2 is likely to be biased. For example, if $\mathrm{corr}(X_i, X_j) \equiv \rho$ for all $i \neq j$, the expected value of S^2 is $\mathrm{E}(S^2) = (1-\rho)\sigma^2$.

Example 3

If F is an **exponential distribution**, then $T_p = -\overline{X}\ln(1-p)$ is unbiased for the pth **quantile** of F. The unbiasedness of this estimator is a consequence of the relation between means and quantiles of exponential random variables; T_p is likely to be biased if F is not an exponential distribution.

The criterion of unbiasedness is not without inadequacies, some of which are discussed in the next set of examples (see also [6] and [9]).

Example 4

Unbiasedness is not maintained by transformation. $\mathrm{E}[g(X)]$ is rarely the same as $g[\mathrm{E}(X)]$; this is true if $g(\cdot)$ is linear, but seldom otherwise. Consequently, if $\hat{\theta}$ is unbiased for θ, it is not to be anticipated that $g(\hat{\theta})$ is unbiased for $g(\theta)$: the opposite is usually the case. For example, the unbiasedness of S^2 for σ^2 in Example 2 guarantees that S is biased for σ; $0 < \mathrm{var}(S) = \mathrm{E}(S^2) - [\mathrm{E}(S)]^2 = \sigma^2 - [\mathrm{E}(S)]^2$ implies that $\mathrm{E}(S) < \sigma$.

Example 5

Unbiased estimators may not exist. If X is a **binomial** random variable with index N and success parameter p, unbiased estimators of $\psi(p)$ exist if and only if ψ is a polynomial of degree $\leq N$. Thus, for example, there does not exist an unbiased estimator of the expected number of trials until the next success; that is, $\psi(p) = 1/p$.

In cases such as this, a reasonable biased estimator can be constructed by considering a surrogate parameter for which an unbiased estimator does exist. Setting $\theta = 1/p$, it is easily demonstrated that $\theta = \theta^* - B$, where

$$\theta^* = \sum_{i=0}^{N} (1-p)^i$$

and

$$B = \frac{-(1-p)^{N+1}}{p}.$$

Since θ^* is a polynomial of degree $\leq N$ in p, an unbiased estimator of θ^* exists; namely, $\hat{\theta}^* = (N+1)/(X+1)$. Considered as an estimator of θ, $\hat{\theta}^*$ has bias $B < 0$, which is readily seen to be a negligible fraction of the estimand, for large N. It can be shown that $\hat{\theta}^*$ is a consistent estimator of θ.

The bias of an estimator can sometimes be reduced through the use of **bootstrap** or **jackknife** techniques [2, 3, 8].

Example 6

When unbiased estimators exist, they may be obviously unreasonable. Let X be a **Poisson** random variable with mean λ. We consider the problem of estimating $\theta_a = \exp(a\lambda)$, for some known value $a \neq -1$. The unique unbiased estimator of θ_a is $T_a(X) = (a+1)^X$.

For negative values of a, $0 < \theta_a < 1$. However, $a < 0$ implies that $\Pr[0 < T_a(X) < 1] < 1$; there is a chance that the estimator will not fall within the range of the estimand. In fact, for $a \leq -2$, $\Pr[0 < T_a(X) < 1] = 0$, so the estimator never falls within the range of the estimand. A similar problem can occur in estimating **variance components**: unbiased estimators of variance sometimes produce negative estimates.

An alternative estimator of θ_a is $S_a(X) = \exp(aX)$. It can be shown that if $a < 0$, the average squared distance from θ_a to $S_a(X)$ is smaller than that from θ_a to $T_a(X)$. Thus, even though $S_a(X)$ is biased, it has a smaller **mean square error** (MSE) than $T_a(X)$.

Example 7

More than one unbiased estimator may exist. If $X_1, X_2, \ldots, X_n$ are independent and **uniformly** distributed on $(0, \theta)$, then the following estimators are all unbiased for θ : $\hat{\theta}_1 = (n+1)X_{\min}$, $\hat{\theta}_2 = 2X_1$, $\hat{\theta}_3 = 2\overline{X}$, $\hat{\theta}_4 = X_{\min} + X_{\max}$, and $\hat{\theta}_5 = [(n+1)/n]X_{\max}$.

When more than one unbiased estimator exists, selection can be made on the basis of the variance of the estimators. The variances of estimators $\hat{\theta}_1, \ldots, \hat{\theta}_5$ can be shown to be in the proportions

$$n^2 : \frac{n(n+2)}{3} : \frac{n+2}{3} : \frac{2n}{n+1} : 1 ;$$

on this basis, the estimator $\hat{\theta}_5$ is clearly the best. In fact, it can be shown that $\hat{\theta}_5$ is a *uniformly minimum variance unbiased estimator* (UMVUE): no unbiased estimator exists having smaller variance.

Under regularity conditions, a lower bound for the variance of an unbiased estimator is available; an estimator with variance equal to this lower bound is therefore a UMVUE. This "**Cramér–Rao** lower bound" [1] is the reciprocal of the Fisher **information** number $\mathcal{I}(\theta)$: in words, $\mathcal{I}(\theta)$ is the expectation of the squared derivative with respect to θ of the log likelihood. It is possible, however, that a UMVUE does not attain this bound. Construction of UMVUEs is facilitated by the Lehmann–Scheffé theorem, which states that the conditional expectation of an unbiased estimator given a complete **sufficient statistic** is a UMVUE [1].

The existence of a UMVUE does not rule out consideration of other estimators. It is easily demonstrated that if T is unbiased for θ and $\mathrm{var}(T) = a\theta^2$, then $S = T/(1+a)$ has smaller MSE than does T. This result can be used to show that the UMVUE of a normal variance, S^2, has larger MSE than does $[(n-1)/(n+1)]S^2$: in Example 7, the UMVUE $\hat{\theta}_5$ has larger MSE than does $[(n+2)/(n+1)]X_{\max}$.

Other considerations may weigh against the choice of a UMVUE. In Example 7, the UMVUE $\hat{\theta}_5$ has a **skewed** distribution so that, for any sample size $n > 5$, there is a better than 60% chance that the UMVUE exceeds θ. Thus, $\hat{\theta}_4$, while more variable than $\hat{\theta}_5$, may be deemed superior on the basis of the symmetry of its distribution. Since $\Pr(\hat{\theta}_4 \leq \theta) = 0.5$, $\hat{\theta}_4$ is said to be median unbiased.

Unbiased Hypothesis Testing and Risk Unbiasedness

Let $\mathcal{F}_0$ and $\mathcal{F}_1$ denote disjoint collections of distribution functions. A test of $\mathrm{H}_0 : F \in \mathcal{F}_0$ is said to be unbiased against $\mathrm{H}_1 : F \in \mathcal{F}_1$ if, for any $F_0 \in \mathcal{F}_0$ and $F_1 \in \mathcal{F}_1$, $\Pr(\text{Reject } \mathrm{H}_0 | F_0) \leq \Pr(\text{Reject } \mathrm{H}_0 | F_1)$; one is at least as likely to reject the null hypothesis when the alternative is true, as when the null is true. This criterion of unbiasedness is of importance in the theory of uniformly most powerful tests [7].

Unbiased estimation, unbiased hypothesis testing, and other optimal procedures for testing and estimation are unified by the general decision-theoretic concept of *risk unbiasedness* [4, 5] (*see* **Decision**

Theory). Given data $\boldsymbol{X}$ and a **loss function** $L(\theta, d)$, a decision rule $\delta(\boldsymbol{X})$ is said to be risk unbiased if $\mathrm{E}_\theta[L(\theta^*, \delta(\boldsymbol{X}))] \geq \mathrm{E}_\theta[L(\theta, \delta(\boldsymbol{X}))]$ for all θ, θ^*. For example, choosing a squared-error loss function, $L(\theta, d) = [\psi(\theta) - d]^2$, risk unbiasedness of $\delta(\boldsymbol{X})$ is equivalent to unbiasedness of $\delta(\boldsymbol{X})$ as an estimator of $\psi(\theta)$.

References

[1] Bickel, P.J. & Doksum, K.A. (1977). *Mathematical Statistics*. Holden-Day, San Francisco.

[2] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

[3] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

[4] Karlin, S. & Rinott, Y. (1983). Unbiasedness in the sense of Lehmann in n-action decision problems, in *Festschrift for Erich Lehmann*, P.J. Bickel, K.A. Doksum & J.L. Hodges, eds. Wadsworth, Belmont.

[5] Lehmann, E.L. (1957). A theory of some multiple decision problems, I, *Annals of Mathematical Statistics* **28**, 1–25.

[6] Lehmann, E.L. (1981). An interpretation of completeness and Basu's theorem, *Journal of the American Statistical Association* **76**, 335–340.

[7] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd Ed. Wiley, New York.

[8] Quenouille, M.H. (1956). Notes on bias in estimation, *Biometrika* **43**, 353–360.

[9] Romano, J.P. & Siegel, A.F. (1985). *Counterexamples in Probability and Statistics*. Wadsworth and Brooks/Cole, Monterey.

W.A. LINK

Uniform Distribution

We can distinguish between the *discrete* and *continuous* uniform (or *rectangular*) distributions.

Discrete Uniform Distribution

Let X be a **random variable** taking integer values from m_1 to m_2 with the probability function

$$\Pr(X = j) = \frac{1}{m_2 - m_1 + 1}, \quad m_1 \le j \le m_2.$$

There are many examples where a uniform distribution is appropriate. For example, observing a number from tossing a die, or drawing a random digit from 0 to 9, are two common examples.

Mean and Variance

$$\mathrm{E}(X) = \frac{m_1 + m_2}{2},$$

$$\mathrm{var}(X) = \frac{(m_2 - m_1 + 1)^2 - 1}{12}.$$

Continuous Uniform Distribution

Let X be a random variable taking real values from a to b with probability density function (pdf)

$$f(x) = \frac{1}{b - a}, \quad a \le j \le b.$$

It is common to denote $X \sim \mathrm{U}(a, b)$.

Properties

Mean and variance:

$$\mathrm{E}(X) = \frac{a + b}{2},$$

$$\mathrm{var}(X) = \frac{(b - a)^2}{12}.$$

Linear transformation:
For any constants $c > 0$ and d,

$$Y = cX + d \sim \mathrm{U}(ac + d, bc + d).$$

In particular,

$$Y = \frac{X - a}{b - a} \sim \mathrm{U}(0, 1).$$

More Properties of U*(0, 1)*

1. *Probability integral transformation*. Let $F(x)$ be the cdf of a continuous random variable X, then $Y = F(X) \sim \mathrm{U}(0, 1)$.

2. *Inverse cdf transformation*. If the $F(x)$ in point 1 is strictly increasing, then for $Y \sim \mathrm{U}(0, 1)$, $F^{-1}(Y)$ is a random variable with cdf $F(x)$. This result enables us to simulate observations of a continuous random variable that has as its cdf $F(x)$. For example,
   (i) $-\ln Y$ has an **exponential distribution** with mean 1;
   (ii) $\ln[Y/(1 - Y)]$ has a **logistic distribution**;
   (iii) $Y^{1/n}$ has a **beta(n, 1) distribution**.
3. We can also use $Y \sim \mathrm{U}(0, 1)$ to **simulate** observations of a discrete random variable. In particular,
$$X = m_1 + [(m_2 - m_1 + 1)Y]$$
will have a discrete uniform distribution from m_1 to m_2. Here $[x]$ is the largest integer $\le x$.

For additional properties and references related to the uniform distribution, see [2, 3], and [4].

Characterization

For two independent random variables, U and V, Deng & George [1] gave some characterizations of a function g such that $g(U, V) \sim \mathrm{U}(0, 1)$ is independent of V if and only if $U \sim \mathrm{U}(0, 1)$. They also studied several classes of function $g(u, v)$ which yield a U(0, 1), if U and/or V follows a uniform distribution. Examples of such functions are

1. $g_1(u, v) = \min[u/v, (1 - u)/(1 - v)]$,
2. $g_2(u, v) = u + v \bmod 1$,
3. $g_3(u, v) = \min(u, v)/\max(u, v)$,
4. $g_4(u, v) = \log u/(\log u + \log v)$.

References

[1] Deng, L.Y. & George, E.O. (1992). Some characterizations of the uniform distribution with applications to

random number generation, *Annals of the Institute of Statistical Mathematics* **44**, 379–385.

[2] Johnson, N.L. & Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Wiley, New York.

[3] Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics: Continuous Distributions*. Wiley, New York.

[4] Read, C.B. (1988). Uniform (or rectangular) distributions, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 411–414.

LIH-YUAN DENG

Uniform Random Numbers

The use of empirical studies based on computer-generated **(pseudo)-random numbers** has become a common practice in the development of statistical methods, particularly when the analytical study of a statistical procedure becomes intractable. Often there are several generating methods that can be used to produce a random number sequence for a given distribution. Most often these methods are based on the generation of independent variates from the **uniform distribution**, U(0, 1). Usually, pseudorandom integers from 0 to m are generated and then they are transformed into [0, 1] by a scale of $1/m$. Several common uniform number generators are presented here.

Linear Congruential Generator (LCG)

The congruential method, proposed by Lehmer [19], is the most commonly used pseudorandom number generator. A sequence of random numbers is obtained by setting

$$X_i = (BX_{i-1} + A) \bmod m, \quad i \geq 1,$$

where X_i, B, A, and m are nonnegative integers. The quality of the generator is determined by the choice of the increment A, multiplier B, initial seed X_0, and modulus m.

Case 1

If $A = 0$, then it is called a *multiplicative linear congruential generator* (MLCG), in which case it becomes

$$X_i = BX_{i-1} \bmod m, \quad i \geq 1.$$

The maximum period of the sequence $\{X_0, X_1, X_2, \ldots\}$ generated depends on the choice of the modulus m [14, p. 20]:

1. $m = 2^t$, $t > 3$. The maximum period attainable is 2^{t-2}.
2. $m = p > 3$, a large prime number. The maximum period attainable is $p - 1$.
3. Any composite modulus. See Knuth [14, p. 21] for a formula of the maximum period attainable. However, a composite modulus (other than 2^t) is rarely used for a MLCG.

Case 2

If A is not zero, it is possible to achieve the full period m [14, p. 16]. However, according to Marsaglia [24], the "effective period" cannot be greater than the period of the corresponding MLCG. The value of A does not have any effect on the structure of the random number sequence. For example, different values of A give the same results in the spectral test [14, p. 91] and the lattice test [4, p. 28].

In the early days, a popular choice of m was 2^{w-1} (w is the machine word-size) for efficiency considerations. For example, the IBM generator RANDU used $m = 2^{31}$, $B = 65539$, and $A = 0$. However, sequences generated using $m = 2^t$ have a serious drawback: the lower-order bits of X_i are not very random. The Lth least significant bit of the X_i has period equal to $\max(1, 2^{L-2})$. The lowest-order bit is always 1 (odd); the second-lowest-order bit has an order of 1 (if $B = 8k - 3$) or 2 (if $B = 8k + 3$).

Recently, $m = 2^{31} - 1$, a well-known Mersenne prime number [14, p. 390], has become the most popular modulus. It is also the maximum positive number representable in a 32-bit computer. The multiplier $B = 16807$ was suggested by Lewis et al. [20]. It is also used in the scientific library from IMSL. Payne et al. [28] suggested the use of $B = 630360016$, which is commonly used in the SIMSCRIPT II simulation programming language.

Marsaglia [23] was the first to show that successive overlapping sequences of k random numbers fall on at most $(k!m)^{1/k}$ planes, where m is the modulus chosen. This shortcoming may yield grossly wrong results for certain applications, such as in the **Monte Carlo** multiple-integration method.

Shift Register Generator (SRG)

Tausworthe [29] suggested combining L-bits of a binary sequence generated by the linear recurrence relation

$$X_i = (\alpha_1 X_{i-1} + \cdots + \alpha_k X_{i-k}) \bmod 2, \quad i \geq k,$$

for any initial nonzero binary vector $(X_0, \ldots, X_{k-1})$, where $\alpha_j = 0$ or 1 for $1 \leq j \leq k - 1$ and $\alpha_k = 1$.

The necessary and sufficient condition for achieving the maximum period $2^k - 1$ is that the polynomial $f(x)$,

$$f(x) = x^k - \alpha_1 x^{k-1} - \cdots - \alpha_k,$$

is a primitive polynomial over the finite field consisting of $\{0, 1\}$ and $\gcd(L, 2^k - 1) = 1$. For computing efficiency, we usually consider only the primitive trinomial

$$f(x) = x^k + x^h + 1, \quad 1 \le h \le k - 1.$$

Watson [30] tabulated primitive polynomials with degree ≤ 100. A more comprehensive listing of primitive trinomials with modulus 2 was given in [34] and [35].

The advantage of the Tausworthe generator is that if we choose parameter values carefully, then the generated sequences are guaranteed to have a nice property of equidistribution over some multidimensional space. The disadvantage is that it is not efficient because it requires L to be big enough to resemble a continuous random variable over [0, 1]. Lewis & Payne [21] proposed a more efficient method known as the generalized feedback shift register (GFSR) in which numbers are formed by phase-shifted elements of the binary sequence generated by a primitive trinomial. One of the shortcomings of the GFSR is that no theoretical equidistribution over multidimensional space can be proved for all GFSRs. Fushimi & Tezuka [9] gave a necessary and sufficient condition for the k distribution of the GFSR.

Although the Tausworthe shift register generator has a k-space equidistribution property, its empirical performance gives poor results (see, for example, [25]).

Multiple Recursive Generator (MRG)

The MRG is a natural extension of the SRG. It is generated from a degree k primitive polynomial [14, pp. 28–29]

$$f(x) = x^k - \alpha_1 x^{k-1} - \cdots - \alpha_k,$$

with period $p^k - 1$, by

$$X_i = (\alpha_1 X_{i-1} + \cdots + \alpha_k X_{i-k}) \bmod p, \quad i \le k,$$

for any initial nonzero vector $(X_0, \ldots, X_{k-1})$, where p is a large prime number. A polynomial of degree k is said to be a "primitive polynomial modulo p" if this polynomial has a root that is a primitive element of the field with p^k elements. Knuth [14], Zierler [33], Golomb [10], and Lidl & Niederreiter [22] proved and summarized several very important properties about the primitive polynomial and the MRG. The main difference is that the classical SRG uses the modulus $p = 2$, large L (decimation), and large degree, k, whereas the MRG uses small k, $L = 1$, and very large modulus p. Clearly, when $k = 1$, the MRG is reduced to the LCG. Therefore, the MRG includes both the LCG and the SRG as special cases.

Knuth [14, p. 29, conditions (i)–(iii)] gave a search algorithm for finding primitive polynomials. L'Ecuyer & Blouin [17] implemented Knuth's algorithm into a computer program using the generalized spectral tests [14] to find the MRG. The main drawback of Knuth's algorithm is that it involves polynomial modulus arithmetic, and it needs some additional programming work. Deng et al. [7] proposed an efficient search algorithm using only usual arithmetic and no polynomial arithmetic was required.

Example Modulus $m = p = 2^{31} - 1$

A listing of two MRGs and their periods for $k \le 3$ is given in Tables 1–3. For additional listings with other prime modulus and/or a larger k, see [7].

When $k = 1$, period $= 2^{32} - 2 = 2,147,483,646$. As mentioned before, the MRG is reduced to the LCG when $k = 1$. When $k = 2$, period $= 4,611,686,014$,

Table 1 $k = 1$

No.	α_1
1	16 807
2	630 360 016

Table 2 $k = 2$

No.	α_1	α_2
1	7 732	19 398
2	1 644 975 444	1 454 071 610

Table 3 $k = 3$

No.	α_1	α_2	α_3
1	34 482	41 200	20 226
2	1 090 176 785	2 064 992 429	1 835 451 531

132, 420, 608. When $k = 3$, period = 9, 903, 520, 300, 447, 984, 150, 353, 281, 022.

Matrix Congruential Generator (MCG)

The matrix generator, considered by Franklin [8], Grothe [11], and Niederreiter [26], is defined by

$$\mathbf{X}_i = \mathbf{B}\mathbf{X}_{i-1} \bmod p, \quad i \geq 1,$$

where the $\mathbf{X}_i$s are k-dimensional vectors, $\mathbf{B}$ is a $k \times k$ matrix, and p is usually chosen as a large prime number. The maximum period of the MCG is $p^k - 1$. A brief review of the matrix generator is given by L'Ecuyer [16]. Niederreiter [27] derived the "discrepancy" of the sequence generated by the matrix generator. The procedure proposed in Grothe [11] of finding the matrix multiplier $\mathbf{B}$ with the maximum period depends on the availability of the primitive polynomial of degree k. Deng et al. [7] also proposed an efficient algorithm for searching the MRG and MCG.

Example: Modulus $m = p = 2^{31} - 1$

A sample listing of the matrix $\mathbf{B}$ for two MCGs and their periods for $k \leq 3$ is given below. When $k = 1$, the MCG is again reduced to the LCG. When $k = 2$, period = 4 611 686 014 132 420 608.

$$\begin{pmatrix} 17\,943 & 43\,665 \\ 9\,283 & 33\,768 \end{pmatrix},$$

$$\begin{pmatrix} 1\,238\,321\,839 & 1\,336\,529\,607 \\ 995\,629\,446 & 359\,802\,910 \end{pmatrix}.$$

When $k = 3$, period = 9 903 520 300 447 984 150 353 281 022.

$$\begin{pmatrix} 1\,853 & 8\,475 & 32\,652 \\ 25\,307 & 36\,979 & 23\,868 \\ 39\,567 & 34\,683 & 41\,419 \end{pmatrix},$$

$$\begin{pmatrix} 1\,347\,553\,617 & 1\,521\,896\,970 & 1\,565\,253\,766 \\ 1\,454\,071\,610 & 1\,644\,975\,444 & 1\,616\,800\,968 \\ 519\,234\,310 & 1\,463\,044\,428 & 2\,045\,396\,196 \end{pmatrix}.$$

Combination Generator

Consider the following n multiplicative linear congruential generators (MLCGs), proposed by Lehmer [19]:

$$X_{j,i+1} = B_j X_{j,i} \bmod m_j, \quad i \geq 0,$$
$$j = 1, 2, 3, \ldots, n,$$

where $X_{j,0}$ (initial seed), B_j (multiplier) are positive integers and m_j (modulus) are different prime numbers. Wichmann & Hill [31] suggested adding three MLCGs and take the fractional part:

$$U_{W,i} = \sum_{j=1}^{3} \frac{X_{j,i}}{m_j} \bmod 1.$$

Through a simple example, they claimed that this procedure "ironed out" the imperfections in the component variates.

Example

Listed in Table 4 are the three MLCGs used in Wichmann & Hill [31]. The period of this generator is 1cm, $(m_1 - 1, m_2 - 1, m_3 - 1) \approx 6.95 \times 10^{12}$. Zeisel [32] observed that a linear combination of several MLCGs with different moduli is equivalent to another MLCG with a large multiplier ($B = 16\,555\,425\,264\,690$) and a large modulus ($m = 27\,817\,185\,604\,309$).

L'Ecuyer [15] considered a variation of Wichmann & Hill's method:

$$U_{L,i} = \sum_{j=1}^{n} \frac{\delta_j X_{j,i}}{m_1} \bmod 1,$$

where $\delta_j = (-1)^{j-1}$. He proved that if generators are independent of each other and if one of the generators is uniformly distributed, then the combined generator will also be uniformly distributed. L'Ecuyer & Tezuka [18] studied the structural properties for these two classes of the combined random number generators (RNGs) and extended the observation by Zeisel [32].

Table 4

j	B_j	m_j
1	170	30 323
2	171	30 269
3	172	30 307

Empirical and Statistical Justifications

The technique of combining several random number generators to obtain a "more random" generator has been suggested by many authors. Wichmann & Hill [31], Marsaglia [25], Collings [3], L'Ecuyer [15], and Anderson [1] all suggested the use of the combination generator. In fact, as pointed out in [7] the, MRG (and MCG) can also be considered as a combination generator. Marsaglia [25] concluded that the combination generator seems to be the best, according to his empirical study of several popular generators. Several other authors also performed empirical studies about the combination generator. Collings [3], L'Ecuyer [15], and Anderson [1] also found good empirical performance of the combination generators.

Some theoretical support for the combination generator is given in the literature. See, for example, Horton [12], Horton & Smith [13], Brown & Solomon [2], Marsaglia [25], and Deng & George [5]. However, they all made an unrealistic assumption that the individual generators are independent of each other. Deng et al. [6] proved the (asymptotic) uniformity and independence of the combined generator without assuming independence between the generators. Deng et al. [7] also gave some intuitive explanations on the excellent performance of the combination generators.

References

[1] Anderson, S.L. (1990). Random number generators on vector super-computers and other advanced architectures, *SIAM Review* **32**, 221–251.

[2] Brown, M. & Solomon, H. (1979). On combining pseudorandom number generators, *Annals of Statistics* **3**, 691–695.

[3] Collings, B.J. (1987). Compound random number generators, *Journal of the American Statistical Association* **82**, 525–527.

[4] Dagpunar, J. (1988). *Principles of Random Variate Generation* Oxford University Press, New York.

[5] Deng, L.Y. & George, E.O. (1990). Generation of uniform variates from several nearly uniformly distributed variables, *Communications in Statistics – Simulation and Computation* **19**, 145–154.

[6] Deng, L.Y., George, E.O. & Chu Y.C. (1991). On improving pseudo-random number generators, in *Proceedings of the 1991 Winter Simulation Conference*, B.L. Nelson, W.D. Kelton & G.M. Clark, eds. WSC, Phoenix, Arizona, pp. 1035–1042.

[7] Deng, L.Y., Rousseau, C. & Yuan, Y. (1992). Generalized Lehmer–Tausworthe random number generators, in *Proceedings of the Thirtieth Annual ACM Southeast Regional Conference*. Raleigh, North Carolina, April 8–10, pp. 108–115.

[8] Franklin, J.N. (1964). Equidistribution of matrix-power residues modulo one, *Mathematics of Computation* **18**, 560–568.

[9] Fushimi, M. & Tezuka, S. (1983). The k-distribution of generalized feedback shift register pseudorandom number, *Communications of the Association of Computing Machinery* **26**, 516–523.

[10] Golomb, S.W. (1967). *Shift Register Sequence*. Holden-Day, San Francisco.

[11] Grothe, H. (1987). Matrix generators for pseudo-random vector generation, *Statistics Papers* **28**, 233–238.

[12] Horton, H.B. (1948). A method for obtaining random numbers, *Annals of Mathematical Statistics* **19**, 81–85.

[13] Horton, H.B. & Smith III, R.T. (1949). A direct method for producing random digits in any number system, *Annals of Mathematical Statistics* **20**, 82–90.

[14] Knuth, D.E. (1981). *The Art of Computer Programming*, Vol 2. *Seminumerical Algorithms*, 2nd Ed. Addison-Wesley, Reading.

[15] L'Ecuyer, P. (1988). Efficient and portable combined random number generators, *Communications of the Association of Computing Machinery* **31**, 742–748, 774.

[16] L'Ecuyer, P. (1990). Random numbers for simulation, *Communications of the Association of Computing Machinery* **33**, 85–97.

[17] L'Ecuyer, P. & Blouin, F. (1988). Linear congruential generators of order $k > 1$, in *Proceedings of the 1988 Winter Simulation Conference*, IEEE Press, New York: pp. 432–439.

[18] L'Ecuyer, P. & Tezuka, S. (1991). Structural properties for two classes of combined random number generators, *Mathematics of Computation* **57**, 735–746.

[19] Lehmer, D.H. (1951). Mathematical methods in large-scale computing units, in *Proceedings of the Second Symposium on Large Scale Digital Computing Machinery*. Harvard University Press, Cambridge, Mass., pp. 141–146.

[20] Lewis, P.A.W., Goodman, O.S. & Miller, J.W. (1969). A pseudo-random number generator for the System 360, *IBM Systems Journal* **8**, 136–146.

[21] Lewis, T.G. & Payne, W.H. (1973). Generalized feedback shift register pseudo-random number algorithms, *Journal of the Association of Computing Machinery* **20**, 456–468.

[22] Lidl, R. & Niederreiter, H. (1986). *Introduction to Finite Fields and Their Applications*. Cambridge University Press, Cambridge.

[23] Marsaglia, G. (1968). Random numbers fall mainly in planes, *Proceedings of the National Academy of Sciences* **61**, 25–28.

[24] Marsaglia, G. (1972). The structure of linear congruential sequences, in *Applications of Number Theory to*

Numerical Analysis, S.K. Zaremba, ed. Academic Press, New York, pp. 249–287.

[25] Marsaglia, G. (1985). A current view of random number generators in *Proceedings of the Sixteenth Symposium on the Interface*, L. Billard, ed. Elsevier, Amsterdam, pp. 3–10.

[26] Niederreiter, H. (1986). A pseudorandom vector generator based on finite field arithmetic, *Mathematica Japonica* **31**, 759–774.

[27] Niederreiter, H. (1990). Statistical independence properties of pseudorandom vectors produced by matrix generators, *Journal of Computational and Applied Mathematics* **31**, 139–151.

[28] Payne, W.H., Rabung, J.R. & Bogyo, T. (1969). Coding the Lehmer pseudo number generator, *Communications of the Association of Computing Machinery* **12**, 85–86.

[29] Tausworthe, R.C. (1965). Random numbers generated by linear recurrence modulo two, *Mathematics of Computation* **19**, 201–209.

[30] Watson, E.J. (1962). Primitive polynomials (Mod 2), *Mathematics of Computation* **16**, 368–369.

[31] Wichmann, B.A. & Hill, I.D. (1982). An efficient and portable pseudo-random number generator, *Applied Statistics* **31**, 188–190.

[32] Zeisel, H. (1986). A remark on algorithm AS 183, *Applied Statistics* **35**, 89.

[33] Zierler, N. (1959). Linear recurring sequences, *Journal of the Society for Industrial and Applied Mathematics* **7**, 31–48.

[34] Zierler, N. & Brillhart, J. (1968). On primitive trinomials (Mod 2), I, *Information Control* **13**, 541–554.

[35] Zierler, N. & Brillhart, J. (1969). On primitive trinomials (Mod 2), II, *Information Control* **14**, 566–569.

(*See also* **Simulation**)

LIH-YUAN DENG

Unimodality

Let X be an absolutely continuous **random variable** with density function $f(x)$ and corresponding distribution function $F(x)$. We say that X is unimodal if $f(x)$ is either decreasing, increasing, or, more typically, increasing to the **mode** m and then decreasing (in other words, $F(x)$ is convex if $x < m$ and concave if $x > m$). For example, **normal** and **gamma** random variables are unimodal.

Often we are interested in some characteristic θ of X (e.g. **median**, 95th percentile; *see* **Quantiles**). If we know $f(x)$, then θ is easily calculated. If we know that $f(x)$ belongs to some parametric family, then θ is a function of the parameters of the family. If we have no information on $f(x)$, then little can be said about θ.

The question we entertain is: If we know that $f(x)$ belongs to the class of unimodal density functions, what can be said about θ? Two results are described:

1. *Chebyshev's inequality*. If we know that $f(x)$ is unimodal, what can we say about $\Pr(|X - \mu| > x)$?
2. *Interpolation*. If we know that $f(x)$ is unimodal and the values of $F(x)$ at selected points; $F(a_i) = p_i$ for $i = 1, \ldots, n+1$, what can be said about $F(b)$ or $F^{-1}(p)$ for some b and p of interest?

Chebyshev's inequality states that $\Pr(|X - \mu| \geq x) \leq \sigma^2/x^2$. Camp [2] and Meidell [9] showed that, if X is unimodal with mode at 0, then, for any $x > 0$,

$$\Pr(|X| \geq x) \leq \left(\frac{r}{r+1}\right)^r \frac{\mathrm{E}(|X|^r)}{x^r} \quad \text{for any } r > 0. \tag{1}$$

The result for $r = 2$ dates back to Gauss [6]. If X is unimodal and symmetric about μ (i.e. has a mode at μ), then (1) provides the following modification to Chebyshev's inequality:

$$\Pr(|X - \mu| \geq x) \leq \frac{4}{9}\frac{\sigma^2}{x^2}. \tag{2}$$

For example, if $x = 2\sigma$, then Chebyshev's bound is 1/4, while the bound in (2) is 1/9. This bound is achieved by mixing a point mass at $x = \mu$ with a **uniform** density function. In [1] this bound is further refined if, in addition to unimodality and symmetry, smoothness conditions in the form of bounds on $|f'(x)|$ are assumed. Specifically, if $|f'(x)|$ does not exceed the maximum of $|f'(x)|$ for a normal random variable, then the bound is further reduced from 1/9 to 0.067. Similar results for restrictive classes of distributions also appear in [3, 10].

The interpolation result appears in [8]. We present the result for a nonnegative random variable with decreasing density; the result for unimodal densities, which appears in [8], is more complicated and in the same spirit.

Assume that the p of interest is in the jth interval (i.e. $p_j \leq p \leq p_{j+1}$). The bound on the pth quantile is determined by the line connecting (a_i, p_i) to (a_{i+1}, p_{i+1}) with slope $s_i = (p_{i+1} - p_i)/(a_{i+1} - a_i)$. This line crosses the line $y = p$ at $t_{i,p} = (p - p_i)/s_i + a_i$:

$$F^{-1}(p) \geq LB \equiv \begin{cases} \max(0, t_{2,p}), & \text{if } j = 1, \\ \max(t_{j-1,p}, t_{j+1,p}), & \text{if } 2 < j < n, \\ t_{n-1,p}, & \text{if } j = n, \end{cases} \tag{3a}$$

and

$$F^{-1}(p) \leq UB \equiv t_{j,p}. \tag{3b}$$

A reasonable estimate for $F^{-1}(p)$ is $(LB+UB)/2$. The results for the **exponential** random variable with $\lambda = 1$, where $a_i = 0.5i$, $i = 1, \ldots, 8$, are presented in Table 1.

Two other questions have been raised regarding unimodal densities. The first is: When is the sum of two unimodal densities unimodal? If the modes of the two random variables are not the same, then the sum need not be unimodal. For example, the sum of two normal random variables with different means (i.e. modes) is not unimodal if the means are sufficiently different. Even if the modes are the same then it is also necessary (see the example in [5]) and sufficient [11] that the random variables are symmetric.

The other question that has received attention is: What is the relationship between the mode (m), **mean** (μ), and the median (ν) for unimodal densities? Groeneveld & Meeden [7] showed that if $F_{Y_1}(y) \leq F_{Y_2}(y)$, where $Y_1 = (X - \nu)^+$ and $Y_2 = (\nu - X)^+$, then $m \leq \nu \leq \mu$ (*see* **Skewness**).

The reader is referred to Dharmadhikari & Joagdev [4] for a more mathematical treatment of the

Table 1 Interpolation of exponential quantiles

p	$\theta = F^{-1}(p)$	LB	UB	$\hat{\theta} = (LB + UB)/2$	$(\hat{\theta} - \theta)/\theta$
0.5	0.693	0.635	0.723	0.679	−0.020
0.75	1.386	1.347	1.407	1.377	−0.007
0.90	2.303	2.223	2.331	2.277	−0.011
0.95	2.996	2.995	2.997	2.996	0.0

above topics, and for extensions to multivariate unimodal densities.

References

[1] Bickel, P.J. & Krieger, A.M. (1992). Extensions of Chebychev's inequality with applications, *Probability and Mathematical Statistics* **13**, 293–310.

[2] Camp, B.H. (1922). A new generalization of Tchebycheff's statistical inequality, *Bulletin of the American Mathematical Society* **28**, 427–432.

[3] DasGupta, A. (2000). Best constants in Chebyshev's inequalities with various applications, *Metrika* **51**, 185–200.

[4] Dharmadhikari, S. & Joag-dev, K. (1988). *Unimodality, Convexity, and Applications*. Academic Press, New York.

[5] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. II. Wiley, New York.

[6] Gauss, C.F. (1821). *Theoria Combinationis Observationum*, Article 10, Gottingen.

[7] Groeneveld, R.A. & Meeden, G. (1977). The mode, median and mean inequality, *American Statistician* **31**, 120–121.

[8] Krieger, A.M. & Gastwirth, J.L. (1984). Interpolation from grouped data for unimodal densities, *Econometrica* **52**, 419–426.

[9] Meidell, B. (1922). Sur un problème du calcul des probabilités et les statistiques mathématiques, *Comptes Rendus del'Académie des Sciences, Paris* **173**, 806–808.

[10] Selke, T. & Selke, S. (1997). Chebyshev's inequalities for unimodal distributions, *American Statistician* **51**, 34–40.

[11] Wintner, A. (1938). *Asymptotic Distributions and Infinite Convolutions*. Edwards, Ann Arbor.

ABBA M. KRIEGER

Union Internationale Contre le Cancer (UICC)

The International Union Against Cancer (UICC) is devoted to all aspects of the worldwide fight against cancer (*see* **Oncology**). Its objectives are to advance scientific and medical knowledge in research, diagnosis, treatment, and prevention of cancer, and to promote all other aspects of the campaign against cancer throughout the world. Founded in 1933, the UICC is a nongovernmental, independent association of more than 270 member organizations in about 80 countries. Members are voluntary cancer leagues and societies, cancer research and/or treatment centers and, in some countries, ministries of health. The UICC has played a major role in the field of cancer biostatistics, notably by creating the TNM committee and the Controlled Clinical Trials Committee.

The TNM committee was created in 1954 with the aim of characterizing and classifying malignancies. The TNM system, devised by Pierre Denoix, was based on:

1. the tumor (T),
2. regional lymph nodes (N), and
3. distant metastases (M).

Precise clinical description and classification of malignant neoplasms by anatomical extent of disease may serve a number of related objectives, namely:

1. to aid the clinician in the planning of treatment.
2. to help the clinician to making a prognosis.
3. to facilitate the exchange of information between treatment centers.
4. to pursue studies on the natural history of cancer.

Finally, the TNM classification improved the quality of **clinical trials**; first, by precisely defining the type of patients concerned, and secondly, by enabling the comparison of treatment efficacy between groups, the stage of malignancy being controlled. The TNM staging system is still being actively developed by UICC.

The Controlled Clinical Trials Committee, created in 1966, was composed of a dozen members (clinicians, surgeons, and biostatisticians) of various nationalities. I was the first President, later succeeded by Robert Flamant. There were two main aims.

The first was to analyze the different methodologic, ethical, and practical problems encountered during clinical trials, and to offer solutions. Committee meetings were held yearly to discuss new aspects. UICC technical reports on a large number of methodologic points were prepared and distributed.

The committee's second objective was to establish a list of all randomized cancer trials throughout the world, either completed or ongoing. This information was aimed at avoiding redundancy: teams planning duplicate trials could either withdraw or join forces. In 1968 the International Information Office was set up under the direction of Robert Flamant; five compilations of clinical trials were published. The International Information Office only ceased its activity when the US National Cancer Institute (*see* **National Institutes of Health (NIH)**) created the International Data Bank (ICRDB), a body with the same vocation.

Independently of these actions, the UICC has conducted and is still conducting a program on epidemiology and prevention. Current studies include in particular Chernobyl Disaster Follow-up, Evaluation of Primary Prevention of Cancer, Familial Cancer and Prevention, and Nutrition, Diet, and Cancer.

In these and other fields of cancer research UICC has made an important contribution.

D. SCHWARTZ

Union–Intersection Principle

Multivariate analysis refers to the branch of statistics in which we attempt to analyze multiple response (or dependent) variables simultaneously. This is in contrast to the simpler situation of univariate analysis in which we focus on only one response variable. Not surprisingly, multivariate analysis is more complex than univariate analysis and this is especially true for **hypothesis testing**. Roy [8] developed the *union–intersection principle* as a tool for solving statistical inference problems in multivariate analysis.

We illustrate the applicability of the union–intersection principle with an example. Suppose we have an experimental situation in which we measure pretreatment and posttreatment values of p response variables on each of n subjects, and we are interested in determining whether the treatment has an effect. For convenience, we let $\boldsymbol{\mu} = [\mu_1 \ldots \mu_p]'$ denote the $p \times 1$ vector of population means for the post- minus pretreatment responses. A treatment effect is apparent if $\boldsymbol{\mu}$ is different from the null vector. Therefore, we form our hypothesis testing problem as

$$\mathrm{H}_0 : \boldsymbol{\mu} = \mathbf{0} \text{ vs. } \mathrm{H}_1 : \boldsymbol{\mu} \neq \mathbf{0} \tag{1}$$

We let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote the $p \times 1$ vectors of responses for the subjects $1, \ldots, n$, respectively, and we let $\overline{\mathbf{X}} = (1/n) \sum_{i=1}^{n} \mathbf{X}_i$ denote the $p \times 1$ vector of sample means. We attempt to determine from $\overline{\mathbf{X}}$ whether there is enough evidence to reject H_0 in favor of H_1. If our situation consisted of only one response variable ($p = 1$) that followed a **normal distribution**, then we could apply the **paired *t* test**,

$$t = \sqrt{n} \frac{\overline{X}}{s}, \tag{2}$$

where s is the sample standard deviation. The statistic t in (2) follows a t_{n-1} distribution (**Student's *t* distribution** on $n - 1$ **degrees of freedom**).

However, for the multivariate testing problem in (1) we want to test the **null hypothesis** that simultaneously p response means are different from zero. To apply the union–intersection principle, we let $\mathbf{b} = [b_1 \ldots b_p]$ be any nonnull $p \times 1$ vector. Then $\boldsymbol{\mu} = \mathbf{0}$ if and only if $\mathbf{b}'\boldsymbol{\mu} (= \sum_{i=1}^{p} b_i \mu_i) = 0$ for every nonnull vector $\mathbf{b}$. Analogously, $\boldsymbol{\mu} \neq \mathbf{0}$ if and only if $\mathbf{b}'\boldsymbol{\mu} \neq 0$ for at least one nonnull vector $\mathbf{b}$. This suggests that we consider the following hypothesis testing problem for each nonnull vector $\mathbf{b}$:

$$\mathrm{H}_0(\mathbf{b}) : \mathbf{b}'\boldsymbol{\mu} = 0 \text{ vs. } \mathrm{H}_1(\mathbf{b}) : \mathbf{b}'\boldsymbol{\mu} \neq 0. \tag{3}$$

Over all nonnull vectors $\mathbf{b}$, the null hypothesis described in (1) is equivalent to the intersection of all the null hypotheses described in (3), and the alternative hypothesis in (1) is equivalent to the union of all the **alternative hypotheses** described in (3). In other words,

$$\mathrm{H}_0 \equiv \bigcap_{\mathbf{b}} \mathrm{H}_0(\mathbf{b}) \quad \text{and} \quad \mathrm{H}_1 \equiv \bigcup_{\mathbf{b}} \mathrm{H}_1(\mathbf{b}). \tag{4}$$

This decomposition of the null and alternative hypotheses leads to the chosen nomenclature of the union–intersection principle.

For a particular nonnull vector $\mathbf{b}$, we let $T(\mathbf{b})$ be a statistic for the hypothesis testing problem in (3) and we let $R(\mathbf{b})$ denote its critical region. If we observe $T(\mathbf{b})$ within $R(\mathbf{b})$, then we reject $\mathrm{H}_0(\mathbf{b})$ in favor of $\mathrm{H}_1(\mathbf{b})$. In general, we strive to select a statistic $T(\mathbf{b})$ that has optimal properties, such as uniformly **most powerful**, **unbiased**, etc. If large values of $T(\mathbf{b})$ lead to the rejection of $\mathrm{H}_0(\mathbf{b})$, then H_0 is rejected if any $T(\mathbf{b})$ is in the **critical region** $R = \cup R(\mathbf{b})$. However, this is equivalent to rejecting H_0 for large values of

$$T = \sup_{\mathbf{b}} \{T(\mathbf{b})\}, \tag{5}$$

which we label as the *union–intersection statistic*.

With respect to our particular example, under the assumption that the post- minus pretreatment responses follow a **multivariate normal distribution**, we select $T(\mathbf{b})$ as the paired t statistic, described in (2). For convenience, we use

$$T^2(\mathbf{b}) = \frac{n\mathbf{b}'\overline{\mathbf{X}}\overline{\mathbf{X}}'\mathbf{b}}{\mathbf{b}'\mathbf{S}\mathbf{b}}, \tag{6}$$

where $\mathbf{S}$ is the sample $p \times p$ variance–**covariance matrix**,

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'. \tag{7}$$

It turns out that the vector $\mathbf{b} = \mathbf{S}^{-1}\overline{\mathbf{X}}$ yields the union–intersection statistic

$$T^2 = n\mathbf{X}'\mathbf{S}^{-1}\overline{\mathbf{X}}. \tag{8}$$

The statistic in (8) is known as **Hotelling's T^2** and $(n-p)T^2/p(n-1)$ follows an ***F* distribution** with $(p, n-p)$ degrees of freedom. In the same context, if we let $T^2_{1-\alpha,p,n-p}$ denote the $100\,(1-\alpha)$ upper percentile from Hotelling's T^2 distribution, then the union–intersection principle leads to the following simultaneous probability statement for all linear combinations $\mathbf{b}'\boldsymbol{\mu}$:

$$\Pr\left\{\mathbf{b}'\overline{\mathbf{X}} - \left[\left(\frac{1}{n}\right)\mathbf{b}'\mathbf{S}\mathbf{b}\right]^{1/2} \times T_{1-\alpha,p,n-p} \le \mathbf{b}'\boldsymbol{\mu} \le \mathbf{b}'\overline{\mathbf{X}} + \left[\left(\frac{1}{n}\right)\mathbf{b}'\mathbf{S}\mathbf{b}\right]^{1/2} \times T_{1-\alpha,p,n-p}\right\} = 1-\alpha. \qquad (9)$$

The above examples with multivariate paired data are the simplest applications of the union–intersection principle. The union–intersection principle also provides test statistics and **simultaneous confidence** regions in the context of the multivariate two-sample problem, **multivariate multiple regression**, and **multivariate analysis of variance** [7]. Other applications of the union–intersection principle include testing sphericity of the variance matrix, i.e. proportionality of the variance matrix to the identity matrix [6, 11], and hypothesis testing in multivariate settings when the alternative hypothesis imposes restrictions on the model parameters [1–5, 9, 10, 12].

References

[1] Boyd, M.N. & Sen, P.K. (1984). Union–intersection rank tests for ordered alternatives in a complete block design, *Communications in Statistics – Theory and Methods* **13**, 285–303.

[2] Boyd, M.N. & Sen, P.K. (1986). Union–intersection rank tests for ordered alternatives in ANOCOVA, *Journal of the American Statistical Association* **81**, 526–532.

[3] Chinchilli, V.M. & Sen, P.K. (1981). Multivariate linear rank statistics and the union–intersection principle for hypothesis testing under restricted alternatives, *Sankhyā* **43**, 135–151.

[4] Chinchilli, V.M. & Sen, P.K. (1981). Multivariate linear rank statistics and the union–intersection principle for the orthant restriction problem, *Sankhyā* **43**, 152–171.

[5] Hauschke, D., Kieser, M., Diletti, E. & Burke, M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distributed data, *Statistics in Medicine* **18**, 93–105.

[6] Lombard, C.J. (1983). Union–intersection tests for sphericity, *South African Statistical Journal* **17**, 165–175.

[7] Morrison, D.F. (1976). *Multivariate Statistical Methods*, 2nd Ed. McGraw-Hill, New York.

[8] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics* **24**, 220–238.

[9] Sen, P.K. & Tsai, M.-T.M. (1999). Two-stage likelihood ratio and Union-Intersection tests for one-sided alternatives multivariate mean with nuisance dispersion matrix, *Journal of Multivariate Analysis* **68**, 264–282.

[10] Tsai, M.-T.M. (1993). Union–intersection score tests for some restricted alternatives in exponential families, *Journal of Multivariate Analysis* **45**, 305–323.

[11] Venables, W. (1976). Some implications of the union–intersection principle for tests of sphericity, *Journal of Multivariate Analysis* **6**, 185–190.

[12] Wada, C.Y. & Hotta, L.K. (2000). Restricted alternatives tests in a bivariate exponential model with covariates, *Communications in Statistics – Theory and Methods* **29**, 193–210.

(*See also* **Roy's Maximum Root Criteria**)

V.M. CHINCHILLI

Unit of Analysis

The *unit of analysis* in a given study is the type of item on which data values are summarized in order to draw statistical **inferences**. It is what "*n*" counts. In **hierarchical** data sets, the unit of analysis specifies a level of aggregation.

Introductory Examples

In many research situations, there is one obvious choice for the unit of analysis. In a simple two-arm **clinical trial**, for example, individual patients are assigned at random to either of two treatments, and an outcome is observed for each patient. The results are expressed in terms of how many patients received each treatment and the distribution of outcomes among the patients in each treatment group. The patient is the unit of allocation, the unit of measurement, and thus the natural unit of analysis.

Ambiguity about the unit of analysis may arise, however, in more complex study designs. Consider the following situations.

1. *Community intervention trials*. A multicommunity study to determine whether a mass-media campaign motivates smokers to quit smoking might involve allocating entire communities *en bloc* to intervention or control groups. The primary endpoint, however, may be whether a person who smoked at baseline had quit smoking by the end of the study period, which is determined at the individual-person level. Communities are the units of allocation, while individual people nested within those communities are the units of observation. Is the unit of analysis the community or the individual? A similar question may arise in **group-randomization** studies more generally.
2. *Multilevel observational studies*. A study that seeks to identify the determinants of compliance with guidelines on mammographic screening in a large managed care plan might measure characteristics of individual female enrollees, attributes and care practices of these women's physicians, and local barriers or conveniences at several clinic sites at which these physicians practice. Individual women are nested within physician practices, which are nested within clinic sites. Data are obtained at all three levels. What is the appropriate unit of analysis (*see* **Multilevel Models**)?
3. *Longitudinal studies*. A study of prognosis among **AIDS** patients may seek to model time trends in CD4 lymphocyte counts, which reflect the effects of AIDS on the immune system. If the CD4 counts are obtained in the course of routine clinical care of each patient rather than at fixed time points dictated by a common protocol, then the number and timing of CD4 count values will vary among patients, and measurement occasions can be considered as nested within (rather than crossed with) patients. Is the unit of analysis the individual CD4 count or the patient?

Study Design Features That Lead to Uncertainty About Unit of Analysis

Four features seem to characterize studies such as these, in which ambiguity arises about the unit of analysis. First, they involve nested (or hierarchical) data arrangements in which there is clustering of subunits within aggregates (*see* **Cluster Sampling**). In the three examples above, individuals are clustered within communities, patients within provider panels within clinic sites, and CD4 test measurements within patients. In general terms, the issue is whether the subunit or the aggregate should be the unit of analysis. Some studies involve two or more levels of nesting, as in the mammography example; however, the underlying analytic issue does not depend critically on the number of levels of nesting, and the discussion here will consider only the simplest, two-level case.

Secondly, at least some of the observations are made at the subunit level. Thus, subunit-level data can either be collapsed to the aggregate level or not, depending on the choice of a unit of analysis. (If only aggregate-level data were available, this decision would not arise; *see* **Ecologic Study**.)

Thirdly, aggregate-level effects on outcome are likely. This phenomenon can be viewed equivalently in two ways: (i) greater variability in outcomes among aggregates than would be expected based on within-aggregate variation; or (ii) less variability within aggregates than would be expected from total variation among subunits across all aggregates. In other words, observations on subunits within a

given aggregate tend to be **correlated**, for any of several reasons. One mechanism is *self-selection*. In the community-trial context, people can choose to reside in a certain community because they share attributes of other community residents and thus "fit in", and those shared characteristics may in turn be predictors of the outcome variable. *Contagion* is another mechanism. Attitudes, norms, and behaviors may be transmissible from person to person within a community, leading to homogeneity. *Shared exposures* constitute yet another mechanism. Patients of a certain physician are all subject to his or her individual practice style. Similarly, observations on the same patient over time are influenced by systematic differences between that patient and other patients as well as by temporal **autocorrelation**. Regardless of the mechanism, because of this kind of clustering, subunit-level observations cannot be assumed to be statistically independent across aggregates.

Fourthly, the particular aggregates studied represent, in some sense, a sample from a universe of similar aggregates to which we may wish to generalize the study findings. In **analysis of variance** parlance, the aggregate-level design factor can be considered a **random effect** rather than a **fixed effect**.

Approaches to Analysis

An Incorrect Approach: Treat Subunit-level Data as Statistically Independent

A naive and incorrect approach to analysis of data from studies fitting this description is to analyze data at the subunit level, treating the subunits as though they were statistically independent across aggregates. Failure of the data to conform to this independence assumption leads to inflated type I error rates (*see* **Hypothesis Testing**), sometimes substantially so [2, 9, 11]. Unfortunately, reviews of published medical research suggest that this incorrect practice is all too common [3, 9, 10].

Analysis Based on Aggregate-level Means

An alternative approach that avoids this pitfall is to use the aggregate, rather than the subunit, as the unit of analysis. Subunit-level data are combined to yield a **mean** value for each aggregate, and these means are then treated as elementary data points to be compared across aggregates. (If there are two or more levels of nesting, further aggregation may be necessary to reach a level of aggregation at which the independence assumption is likely to be satisfied.)

Applying this method to the community-trial smoking-cessation example, a "quit rate" for each study community would first be obtained. These quit rates can be regarded as community-specific mean values of an individual-level indicator variable that takes the value 1 for each smoker who quits smoking and 0 for each smoker who continues to smoke. The success of the community-level intervention would then be assessed by comparing the location of the distribution of quit rates between all intervention and all control communities, possibly using a t-test (*see* **Student's *t* Statistics**) or a **nonparametric** analog.

Basing the analysis on group means has long been advocated for group-randomized studies, such as community intervention trials. It follows the classical Fisherian principle to "analyze as you randomize" [4, 6]. Under the **null hypothesis, randomization** justifies an assumption of independence among the communities within a treatment group, even if no such assumption holds among individuals within each community.

But while analysis based on aggregate-level means circumvents some difficulties, it has important shortcomings. If the number of subunits varies markedly among aggregates, then estimates of aggregate-level means based on many subunits will be more precise than those based on fewer subunits. This difference in precision implies violation of the homoscedasticity assumption (*see* **Scedasticity**) behind many standard statistical tests. In principle, this problem can be circumvented by conducting a weighted analysis, weighting each mean in proportion to the inverse of its estimated **variance** (taking care to include both subunit- and aggregate-level components), but other superior methods of analysis described below obviate the need to do so.

A second major difficulty is that subunit-level **covariates** are not easily accommodated in an analysis based on aggregate-level means. Their omission from the analysis may prevent the removal of **bias** due to subunit-level **confounding** factors and may be a lost opportunity to enhance **power**. For example, if women in some physicians' practices are generally younger than those in other practices, then differences in mammographic screening practices may

Table 1 Abridged ANOVA tables

Source of variation	Symbol	df	MS	E(MS)
For the model given in (1)				
Treatment group	G_i	$g-1$	MS_G	$\sigma_P^2 + p\sigma_C^2 + pc\Sigma G_i^2$
Community	$C_{j(i)}$	$g(c-1)$	MS_C	$\sigma_P^2 + p\sigma_C^2$
Person	$P_{k(ij)}$	$gc(p-1)$	MS_P	σ_P^2
For the model given in (2)				
Treatment group	G_i	$g-1$	MS_G	$\sigma_*^2 + \sigma_C^2 + c\Sigma G_i^2$
Community	$C_{j(i)}$	$g(c-1)$	MS_C	$\sigma_*^2 + \sigma_C^2$

be partly due to patient age differences rather than to practice style. In addition, the ability to detect treatment effects can be enhanced if otherwise unexplained variability in outcomes can be reduced by including patient-level predictors of outcome. Finally, absence of subunit-level covariate data from the main analysis sacrifices the ability to consider **interaction** effects involving those factors.

Multilevel Analysis

Fortunately, several analytic methods are now available that obviate the need to choose between using subunits or aggregates as units of analysis. They accommodate data from two or more levels at once and account properly for the nonindependence of subunits within aggregates.

The *mixed-model analysis of variance* is historically the oldest of these methods, being based on classical analysis of variance (ANOVA) theory. Both subunits and aggregates are treated as random effects. In a community-trial context, a simple statistical model for a continuous outcome variable Y would be:

$$Y_{ijk} = \mu + G_i + C_{j(i)} + P_{k(ij)}, \qquad (1)$$

where Y_{ijk} is the outcome value for person k ($k = 1, \ldots, p$) within community j ($j = 1, \ldots, c$) within treatment group i ($i = 1, \ldots, g$), μ is the grand mean, G_i is the effect of being in treatment group i, $C_{j(i)}$ is the effect of being in community j within treatment group i, and, $P_{k(ij)}$ is the effect of being person k within community j within treatment group i. Basically, a particular value of Y is regarded as the sum of treatment-group, community, and individual-person effects.

Because there is only one observation per person, variation among individuals within a community is implicitly combined with **measurement error** in $P_{k(ij)}$. The model includes two random effects, one at the subunit level and one at the aggregate level: $P_{k(ij)} \sim \mathrm{N}(0, \sigma_P^2)$ and $C_{j(i)} \sim \mathrm{N}(0, \sigma_C^2)$. The one fixed effect is for treatment group: G_i. This simple model includes no covariates at the subunit (person) or aggregate (community) levels.

If the number of individuals per community, p, is assumed to be constant across communities and the number of communities per treatment group, c, is constant across treatment groups, then an abridged ANOVA table for the model given in (1) is shown in the top panel of Table 1. A test of the main study hypothesis about treatment effectiveness ($\mathrm{H}_0 : G_i = 0$ for all i) would be $F = MS_G/MS_C$, with $g-1$ and $g(c-1)$ **degrees of freedom** in the numerator and denominator, respectively, neither of which depends on p.

It is interesting and satisfying to note that, for a balanced study design and in the absence of covariates, an analysis based on community-level means yields the same test statistic for the main study hypothesis. In particular, the corresponding additive statistical model for a community-level mean would be:

$$\overline{Y}_{ij} = \mu + G_i + C_{j(i)} + e_{j(i)}, \qquad (2)$$

where $\overline{Y}_{ij}$ is the mean outcome value for community j within treatment group i, $e_{j(i)}$ is the random error of a community-specific mean, and other symbols are as defined above for the model given in (1). Here, $e_{j(i)} \sim \mathrm{N}(0, \sigma_*^2)$. Because there is only one "observation" (a mean value of Y) per community, σ_*^2 and σ_C^2 cannot be separately estimated, and the expected mean square for community is their sum. But σ_*^2 is the sampling variance of an estimated community-specific mean based on p randomly chosen individuals, while σ_P^2 is the variance of Y in the community from which those individuals were

sampled. Hence $\sigma_*^2 = \sigma_P^2/p$. With this substitution, the F statistics for a treatment-group effect become equivalent in the top and bottom panels of Table 1.

Modern statistical software (*see* **Software, Biostatistical**), such as PROC MIXED in SAS, substitute **restricted maximum likelihood** estimation for classical **least squares** estimation, thus relaxing the need for a balanced study design in order to estimate the parameters of the model given in (1) and to calculate a valid test of the null hypothesis of no treatment-group effect [8].

The flexibility and utility of multilevel modeling is illustrated in several sources describing closely related statistical approaches that accommodate covariates at both the subunit and aggregate levels to explain variation at those levels. Bryk & Raudenbush [1] describe how *hierarchical linear modeling* combines subunit- and aggregate-level linear models into a single framework, with an emphasis on continuous response variables. Hedeker et al. [5], describe *random-effects regression* methods that have been developed to accommodate clustered data with unbalanced designs. **Generalized estimating equations** [7] offer yet another way in which to create multilevel models for outcome variables with nonnormal distributions in a **general linear model** framework.

References

[1] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park.

[2] Cornfield, J. (1978). Randomization by group: a formal analysis, *American Journal of Epidemiology* **106**, 100–102.

[3] Divine, G.W., Brown, J.T. & Frazier, L.M. (1992). The unit of analysis error in studies about physicians' patient care behavior, *Journal of General Internal Medicine* **7**, 623–629.

[4] Fisher, R.A. (1949). *The Design of Experiments*, 5th Ed. Oliver & Boyd, Edinburgh.

[5] Hedeker, D., Gibbons, R.D. & Flay, B.R. (1994). Random-effects regression models for clustered data with an example from smoking prevention research, *Journal of Consulting Clinical Psychology* **62**, 757–765.

[6] Hopkins, K.D. (1982). The unit of analysis: group means versus individual observations, *American Educational Research Journal* **19**, 5–18.

[7] Liang, K.Y. & Zeger, S.L. (1993). Regression analysis for correlated data, *Annual Review of Public Health* **14**, 43–68.

[8] Murray, D.M. & Wolfinger, R.D. (1994). Analysis issues in the evaluation of community trials: progress toward solutions in SAS/STAT MIXED, *Journal of Community Psychology*, *Special Issue*, 140–154.

[9] Simpson, J.M., Klar, N. & Donner, A. (1995). Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993, *American Journal of Public Health* **85**, 1378–1383.

[10] Whiting-O'Keefe, Q.E., Henke, C. & Simborg, D.W. (1984). Choosing the correct unit of analysis in medical care experiments, *Medical Care* **22**, 1101–1114.

[11] Zucker, D.M. (1990). An analysis of variance pitfall: the fixed effects analysis in a nested design, *Education and Psychological Measurement* **50**, 731–737.

THOMAS D. KOEPSELL

Univariate Response

The term univariate response is used to refer to a single **response variable** and contrasts with a multivariate response which refers to a number of response variables. Traditionally, **multivariate analysis** considers a number of response variables jointly, and univariate analysis considers a single or univariate response variable.

Many analyses, however, examine the relationship between a response variable and one or more **explanatory variables**. A univariate response may be related to a single, or to multiple, explanatory variables, as can a multivariate response. It is increasingly common in the medical literature to use the term *univariate analysis* to refer to analyses which examine only a single explanatory variable's relationship to a response variable. Similarly, multivariate analysis is often used to refer to analyses like **multiple linear regression** which examine a number of explanatory variables jointly. The term *multifactorial* is sometimes used in the latter situation to avoid confusion, but the context usually makes the intention clear.

VERN T. FAREWELL

University Group Diabetes Program (UGDP)

The University Group Diabetes Program (UGDP) was one of the first **multicenter clinical trials** designed and implemented to evaluate treatments for a chronic disease. It was designed to test the accepted and widely used methods available in the late 1950s and early 1960s for treating adult-onset, noninsulin-dependent diabetes. The three major objectives were: (i) evaluation of the effects of different hypoglycemic treatments (diet, insulin, and oral agents) on the development of vascular complications in patients with adult-onset diabetes; (ii) study of the **natural history** of vascular disease in noninsulin-dependent diabetics; and (iii) development of methods appropriate for the design and conduct of cooperative clinical trials.

Twelve clinics, two lipid laboratories, and a Coordinating Center participated in the study. Each clinic was responsible for recruiting patients, for collecting observations on these patients according to a common study protocol (*see* **Clinical Trials Protocols**) and for providing medical care to each patient enrolled in that clinic. The Coordinating Center was the data repository for all study forms and data. The Coordinating Center staff helped to design the study, were responsible for the inventory, filing, editing, and storing of all study material and had major responsibility for the analysis of study data. The study was funded by grants from the National Institute of Arthritis, Metabolism and Digestive Diseases, US Public Health Service.

Patient recruitment was started in February 1961 and was completed in February 1966, after 1027 patients were enrolled. Follow-up examinations were scheduled through the end of August 1975. Thus patients were followed for 10.0–14.5 years; the mean follow-up for all patients was 12.25 years.

Methods

Patient Eligibility

Patients were considered for enrollment only if the diagnosis of diabetes had been established within the 12-month period preceding the date of the screening examination. In addition, the diagnosis of diabetes had to be confirmed by a glucose tolerance test performed during the screening examination. Each clinic physician was asked to use his best judgment to screen prospective study candidates for absence of life-endangering diseases so as to select patients with a good prognosis for five-year survival. The study design was described in detail in several study reports [9, 10, 11, 13, 16] (*see* **Eligibility and Exclusion Criteria**).

Candidates for the study were placed on a diet with caloric content designed to achieve or to maintain the patient's body weight within $\pm 15\%$ of his/her desirable body weight. The prescribed diet consisted of a fixed proportion of calories derived from fat, protein, and carbohydrates. Each patient was observed for four weeks on treatment with diet alone and only those patients who did not develop major diabetic symptoms, in particular ketosis, were asked to participate in the study. Only the patients who met all of the above requirements and who indicated a willingness to participate, including a willingness to accept any of the treatments under study, were enrolled and allocated to treatment. Initially, patients were asked to give verbal consent after a detailed explanation of the study had been given. Later, all patients including those already enrolled were asked to sign a consent form after the study design and methods were reviewed in detail (*see* **Ethics of Randomized Trials**).

Treatments

Patients were randomly assigned to one of the five treatments listed in Table 1 (*see* **Randomized Treatment Assignment**). There were two insulin treatment groups. The insulin variable treatment was designed to resemble as much as possible the use of insulin in clinical practice. Adjustments in the insulin dosage for patients assigned to the insulin variable treatment group during the course of the study were based on blood glucose values from an abbreviated glucose tolerance test (short GTT). The insulin dosage was to be increased by at least two units per day whenever the fasting blood glucose value obtained from this test was 110 mg per 100 ml or greater and the one-hour value from this test was 210 mg per 100 ml or greater. The fasting value for this test was based on a blood sample drawn from the patient after a

Table 1 Study treatments

Treatment	Abbreviation	Dosage
Insulin variable (U-80 Lente Iletin or other insulins)	IVAR	Amount required to maintain "normal" blood glucose; minimum dose five units per day
Insulin standard (U-80 Lente Iletin insulin)	ISTD	10, 12, 14, or 16 units per day, depending on the patient's body surface
Tolbutamide (Orinase)	TOLB	1.5 g per day
Phenformin[a] (DBI-TD)	PHEN	100 mg per day
Placebo	PLBO	Dosage schedules similar to those used for the oral agents

[a]Added to the study 18 months after patient recruitment started for the other four treatment groups.

12-hour fast. The one-hour value for this short GTT was based on a blood sample drawn 1.5 hours after the patient took his assigned study medication and one hour after ingesting a drink containing 50 g of glucose. The insulin dosage was to be decreased if hypoglycemic episodes were reported by the patient. In the second insulin group, each patient was given a prescription which was based solely on an estimate of the individual's body surface (function of height and weight). The range of dosage was 10–16 units per day.

Two oral agents were studied: tolbutamide, a member of the sulfonylurea family of compounds, and phenformin, a member of the biguanide family of compounds. Tolbutamide was selected because in 1960 clinical experience with this agent was greater than with any of the other sulfonylurea drugs. A dose of 1.5 g per day (1 g in the morning and 0.5 g in the evening) was used for all patients. The dosage of phenformin was 100 mg per day (50 mg before breakfast and 50 mg before the evening meal).

Phenformin was added to the study approximately 18 months after the study had started for the other four treatment groups. When the decision was made, six of the original seven UGDP clinics were so far along with patient recruitment that it was decided not to include phenformin as one of the study treatments in these six clinics. Five additional clinics were recruited for the study when phenformin was added. These five clinics, along with one of the original seven clinics, had allocation schedules providing for assignments to all five treatment groups. In these clinics, three times the number of assignments were made to phenformin as were made to each of the remaining treatment groups. This procedure was used to provide almost the same total number of assignments to phenformin in these six clinics as were made to each of the other four treatment groups in all 12 clinics at the end of patient recruitment. This design feature required that the evaluation of phenformin effects be based on the results for patients in the six clinics administering phenformin rather than the results from all 12 clinics.

Separate **randomization** schedules were used for each clinic, and these schedules were designed to provide balance in the number of patients assigned to the different treatment groups at specified intervals throughout the period of patient recruitment. All treatment allocations were issued by the Coordinating Center. The oral agents were administered in a double-blind fashion; that is, neither the patient nor the physician knew whether the patient was receiving an active drug or placebo.

Examination Schedule

Patients were given an extensive battery of examinations at the time of enrollment and at three-month intervals thereafter. A general clinical review as well as a detailed examination of the eyes, heart, kidney function or peripheral vascular and neurologic systems was performed in conjunction with each quarterly follow-up visit to the study clinic.

Results for Tolbutamide Therapy

The use of tolbutamide therapy was discontinued in June 1969 when it became apparent that there was

an excess cardiovascular mortality for patients in the tolbutamide treatment group compared with the mortality experience for patients in the placebo treatment group as well as compared to patients in either of the two insulin treatment groups (*see* **Excess Mortality**). As soon as possible after June 1969 patients assigned to tolbutamide or placebo tablets were recalled by the clinics to discontinue the prescription for these tablets. A closing date of October 7, 1969, was used for the evaluation of tolbutamide therapy for the first published report on findings for patients treated with tolbutamide [10].

A total of 89 deaths was reported for the four treatment groups considered in the analysis of tolbutamide findings. The number and percentage of patients who died in each of the treatment groups by cause are given in Table 2. As indicated, more patients in the tolbutamide group were observed to have died from all causes, as well as from cardiovascular causes, than in the other groups. The cumulative annual mortality rates per 100 population at risk were computed with **life table** methods, and the results are shown in Figure 1 for all causes as well as for cardiovascular causes. The final judgment regarding the principal cause of death of each deceased study patient was made by a special review team without knowledge of the treatment group to which the patient had been assigned. Their decision regarding principal cause of death was based on information provided in a detailed death report prepared at the clinic (*see* **Cause of Death, Underlying and Multiple**).

Except for blood glucose levels, few differences in evaluations made at scheduled follow-up examinations were observed among the groups treated with placebo and diet, tolbutamide and diet, or insulin and diet [10, 11, 14]. The fasting blood glucose levels in the tolbutamide treated group were lower than the levels in the placebo group and about the same as the levels in the group treated with a fixed dose of insulin. The patients treated with variable doses of insulin had consistently lower levels than the patients in the other three groups (Figure 2).

Most of the excess mortality observed in the tolbutamide treated group over that of patients treated with diet or diet plus insulin appeared to be a result of increased mortality due to myocardial infarction among the tolbutamide treated patients [14]. The mortality for all patients who had at least one myocardial infarction during the course of follow-up was 50% for patients in the tolbutamide group, and 18% for patients in the placebo group, 35% for patients in the insulin standard group, and 40% for patients in the insulin variable group.

Results for Phenformin Therapy

The observed mortality for all causes and from cardiovascular causes for patients assigned to phenformin was higher than the mortality in any of the other treatment groups about two years after the decision concerning tolbutamide. Also, there was no evidence that phenformin was more effective than any of the other treatments in preventing nonfatal

Table 2 Number of deaths by cause, October 7, 1969

Cause	PLBO, $N = 205$	TOLB, $N = 204$	ISTD, $N = 210$	IVAR, $N = 204$
Myocardial infarction	0	10	3	2
Sudden death	4	4	4	5
Other heart disease	1	5	1	2
Extracardiac vascular disease	5	7	5	3
All cardiovascular (CV)	10	26	13	12
Cancer	7	2	4	2
Other or unknown	4	2	3	4
All causes	21	30	20	18
Percent dead				
CV causes	4.9	12.7[a]	6.2	5.9
All causes	10.2	14.7	9.5	8.8

[a]Chi-square P value = 0.005 for PLBO vs. TOLB comparison.

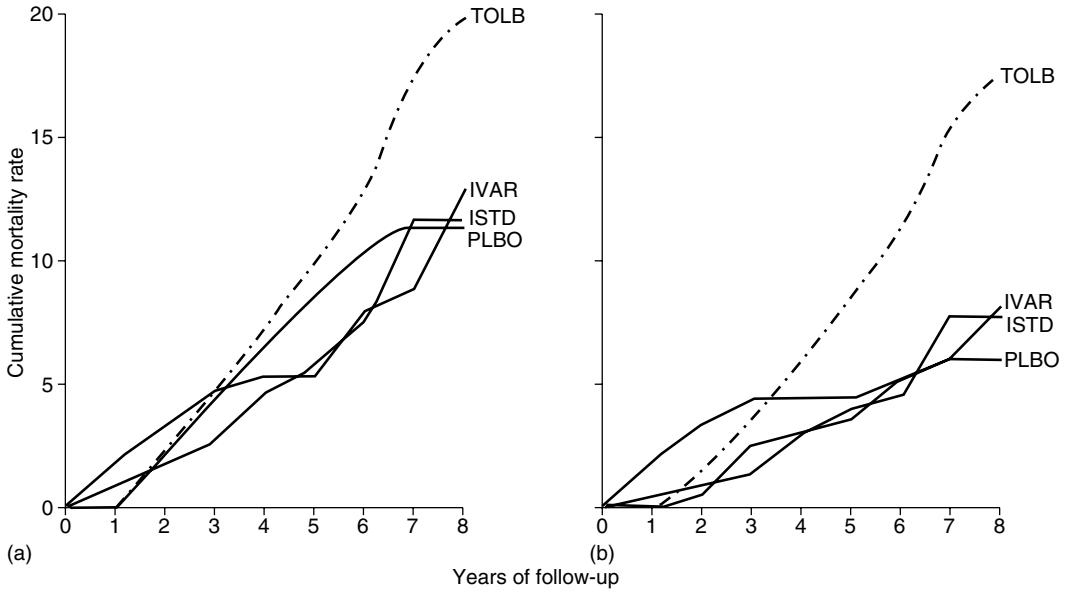

Figure 1 Cumulative mortality rates per 100 population at risk by year of follow-up, as of October 7, 1969. (a) All causes; (b) cardiovascular causes

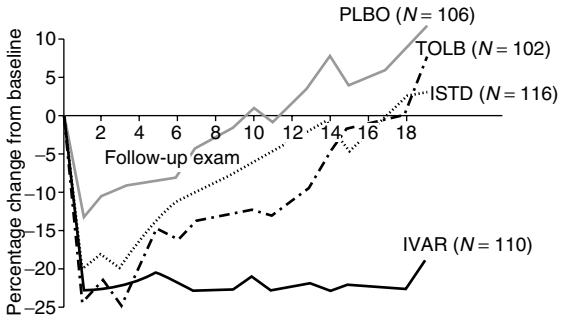

Figure 2 The percentage change in fasting blood glucose levels from baseline to each follow-up examination for the cohort of patients followed through the 19th follow-up examination, as of October 7, 1969

vascular complications. For these reasons, all patients originally assigned to phenformin and those assigned to the corresponding placebo were recalled by the UGDP clinics for a special examination in order to discontinue phenformin or placebo capsules as soon as possible after May 15, 1971, the date of the UGDP decision to discontinue this therapy in the UGDP. A preliminary report on these findings was published based on mortality findings through January 6, 1971, shortly after the decision was made to discontinue phenformin [12]. A more complete report based on data through October 7, 1971, was published later [13], and these results are summarized here.

The number and percentage of patients who died in each treatment group by cause are given in Table 3. The cumulative annual mortality rates per 100 population at risk calculated with life table methods are shown in Figure 3. The mortality results for patients in the two insulin groups were quite similar although the amount of insulin received was different, as specified by protocol. Furthermore, there was no evidence of beneficial or adverse effects for either of the insulin therapies compared to diet alone (placebo group). Therefore, the results in these three treatment groups were pooled to obtain a larger population for comparison with the patients in the phenformin treated group. In the phenformin group, one death was attributed to lactic acidosis, there were also two reported cases of nonfatal lactic acidosis. The possibility that this side effect is a consequence of phenformin therapy had been reported in the literature previously, although there was no convincing proof of causal relationship.

The changes in fasting blood glucose levels during the course of follow-up are shown in Figure 4. There was a large drop in blood glucose levels in all treatment groups after the initiation of treatment, but

Table 3 Number of deaths by cause, October 7, 1971

Cause	PLBO, $N = 64$	PHEN, $N = 204$	ISTD, $N = 68$	IVAR, $N = 65$	PL + INS, $N = 197$
Myocardial infarction	1	5	1	0	2
Sudden death	1	6	2	1	4
Other heart disease	0	8	1	0	1
Extracardiac vascular disease	1	8	2	2	5
All cardiovascular (CV)	3	27	6	3	12
Cancer	3	3	0	0	3
Other or unknown	1	4	0	1	2
All causes	7	34	6	4	17
Percent dead					
CV causes	4.7	13.2[a]	8.8	4.6	6.1
All causes	10.9	16.7[a]	8.8	6.2	8.6

[a]Chi-square P value = 0.02 for PL + INS vs. PHEN comparison.
PL + INS = PLBO + ISTD + IVAR.

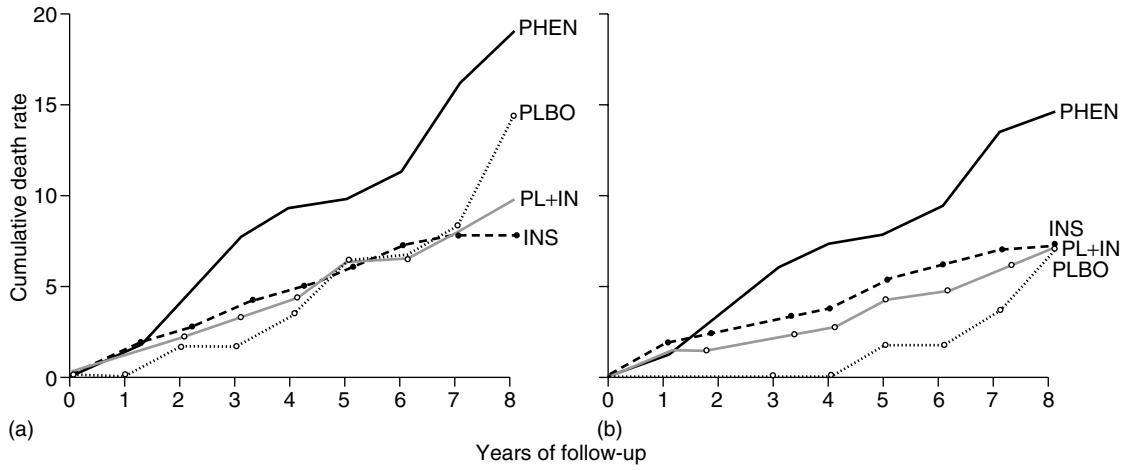

Figure 3 Cumulative mortality rates per 100 population at risk by year of follow-up, as of October 7, 1971. (a) All causes; (b) cardiovascular causes

the changes did not persist except in the insulin variable group. Patients in the phenformin-treated group had an increase in both systolic and diastolic blood pressure levels and in heart rate. These adverse effects of phenformin on blood pressure and heart rate in addition to the observed excess mortality led to the discontinuation of phenformin in the UGDP.

Results for Insulin Therapy

A preliminary report of the findings for the two insulin groups was published in 1978 [15] and a more detailed report in 1982 [16]. The number and percentage of patients who died in the two insulin groups are shown in Table 4. The cumulative mortality rates per 100 population at risk are shown in Figure 5. Almost the same number of cardiovascular deaths was observed in the three treatment groups; more cancer deaths were reported for placebo patients than for patients in either of the two insulin groups.

The changes in fasting blood glucose levels observed during the course of follow-up for each treatment group are presented in Figure 6. There was a large drop in blood glucose levels in all

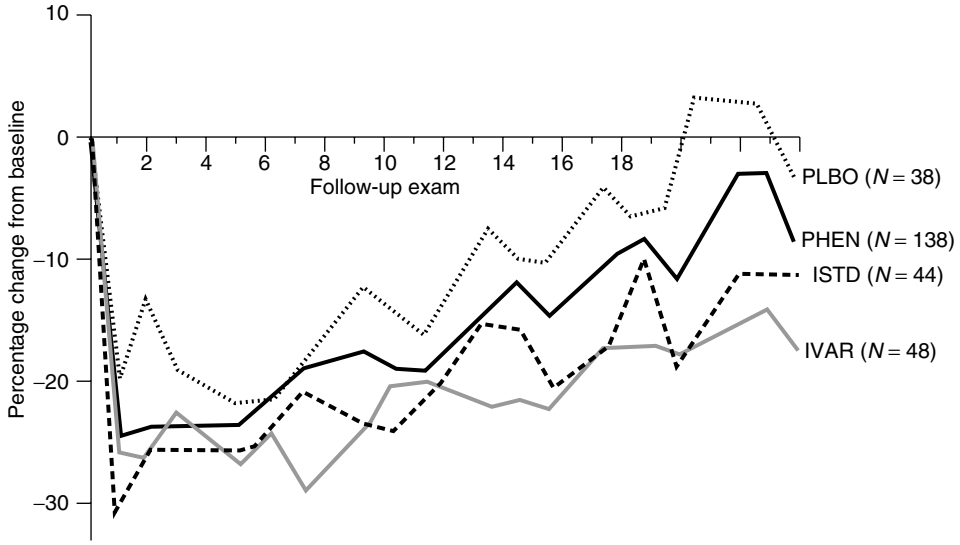

Figure 4 The percentage change in fasting blood glucose levels from baseline to each follow-up examination for the cohort of patients followed through the 23rd follow-up examination, as of October 7, 1971

Table 4 Number of deaths by cause, August 31, 1975

Cause	PLBO, $N = 205$	ISTD, $N = 210$	IVAR, $N = 204$
Myocardial infarction	1	7	4
Sudden death	12	10	12
Other heart disease	4	5	6
Extracardiac vascular disease	14	10	9
All cardiovascular (CV)	31	32	31
Cancer	17	10	8
Other or unknown	12	10	13
All causes	60	52	52
Percent dead			
CV causes	15.1	15.2	15.2
All causes	29.3	24.8	25.5

treatment groups after the initiation of treatment. However, the fasting blood glucose levels for patients in the placebo and insulin standard treatment groups showed a trend toward baseline and then exceeded baseline values. The lower fasting blood glucose levels in the insulin variable treatment group were maintained during the course of the study by increasing the mean number of units of insulin from ten units at the first quarterly follow-up examination to 47 units at the 39th quarterly follow-up examination in the cohort followed for 39 quarters.

The occurrence of microvascular complications such as diabetic retinopathy and diabetic nephropathy was remarkably low during the course of follow-up and there were no differences among the three treatment groups. The UGDP investigators pointed out

> ... the 12–14 years of observations on the course of vascular complications in patients with type II disease in the UGDP show that in spite of the gradual progression of the carbohydrate abnormality and in spite of a relatively high incidence of cardiovascular risk factors, the overall mortality and the proportion of mortality due to cardiovascular complications were not significantly greater than would be expected in a nondiabetic population of comparable age, race and sex [16].

Discussion

Meinert [5] provided a detailed chronology of the University Group Diabetes Program as well as a summary of criticisms of the UGDP and comments on these criticisms. He also provided an assessment of the impact of the UGDP on prescribing practices in the United States. Meinert concluded that the study did have some effect on treatment practices, but perhaps the most important result was that it had led physicians to reexamine the underlying rationale for treatment of noninsulin-dependent diabetics.

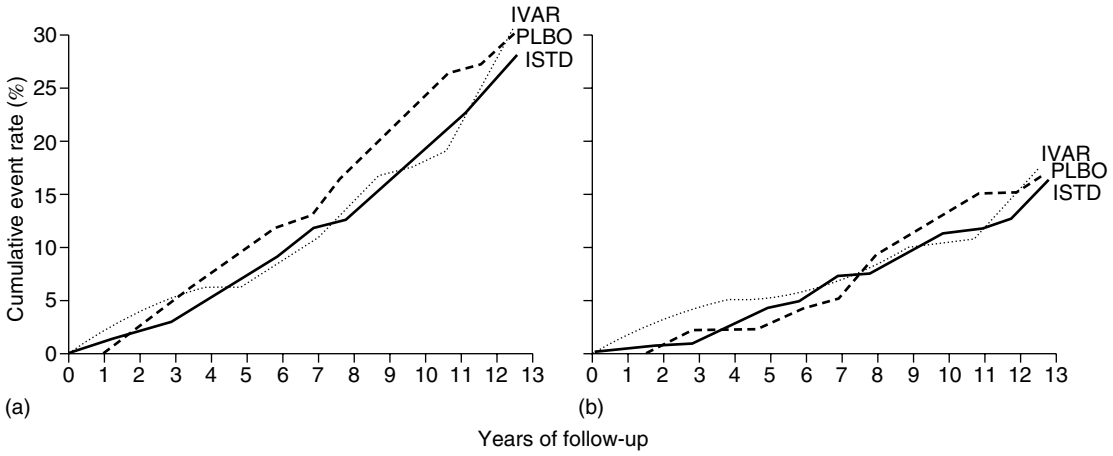

Figure 5 Cumulative mortality rates per 100 population at risk by year of follow-up, as of August 31, 1975. (a) All causes; (b) cardiovascular causes

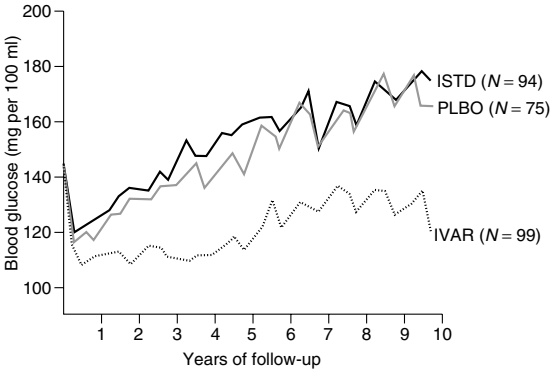

Figure 6 Mean fasting blood glucose levels at baseline and each follow-up examination for the cohort of patients followed through the 39th follow-up examination, as of August 31, 1975

The methods and procedures used in the UGDP Coordinating Center provided a framework for standards of operations for such Coordinating Centers in multicenter clinical trial. The Coordinating Center investigators (C.R. Klimt, C.L. Meinert, G.L. Knatterud, and P.L. Canner) utilized the UGDP experience to further the development of clinical trial methodology and the development of procedures for Coordinating Center operations in the implementation of many multicenter clinical trials conducted after the UGDP was initiated. Two statistical procedures, **Monte Carlo** monitoring procedures and a **likelihood** approach, were used to evaluate the effects of tolbutamide [10] and the effects of phenformin therapy [13]. The monitoring approach developed by Canner and colleagues [2, 3] used a computer **simulation** procedure to generate boundaries to evaluate a test statistic at different times during the course of a study (*see* **Data and Safety Monitoring**). The likelihood approach was developed by Cornfield [3, 4] and this **Bayesian method** was designed to evaluate treatment effects. The value generated by this approach was called "relative betting odds" (*see* **Relative Odds**) and for a drug–placebo comparison estimated the odds in favor of the **null hypothesis** of no difference relative to a specified set of alternatives. In the case of tolbutamide, these procedures were applied after the investigators realized that there was an unfavorable trend for patients treated with tolbutamide. It is perhaps worth pointing out that the UGDP did not have an independent Data Monitoring Committee (*see* **Data Monitoring Committees**) to review accumulating results for adverse or beneficial trends. The strategies used in the UGDP have been replaced by other methods, but perhaps their use in the UGDP pointed out the need for approaches to take account of interim monitoring in the assessment of treatment effects.

Few clinical trials have generated such an unusually acerbic and long lasting controversy as the UGDP. Every aspect of the design, execution, analysis and results of the trial have been subjected to extraordinary scrutiny. Two audits (Bilstad et al. [1], and Report of Committee for the Assessment of

Biometric Aspects of Controlled Trials of Hypoglycemic Agents [7]) of the UGDP were conducted. Both of these identified some shortcomings that had already been acknowledged by the UGDP investigators. The two audits concluded that any errors in data reporting or processing that did occur were infrequent and no more than might be expected in a long-term multicenter clinical trial.

The UGDP findings which precipitated this controversy were the results for tolbutamide which were first presented at the American Diabetes Association annual meeting and subsequently published in *Diabetes* [10]. The UGDP investigators review of mortality and occurrence of nonfatal events had led to formulation of the following conclusion:

> All UGDP investigators are agreed that the findings of this study indicate that the combination of diet and tolbutamide therapy is no more effective than diet alone in prolonging life. Moreover, the findings suggest that tolbutamide and diet may be less effective than diet alone or than diet and insulin at least insofar as cardiovascular mortality is concerned. For this reason, use of tolbutamide has been discontinued in the UGDP. . . . It should be noted that any conclusion reached in this study pertains only to the type of patient studied and the specific hypoglycemic agents and dosage schedules used. Extrapolation of findings to other dosage schedules of the same drug or to other chemically related hypoglycemic agents not included in this study must be made on a judgmental and nonstatistical basis.

It is remarkable that such a conservative conclusion should have generated such controversy. The UGDP investigators responded to several of the criticisms of the study in 1972 [6].

The results of phenformin were released to the medical community by a short report in the *Journal of the American Medical Association* [12] rather than by means of a presentation at a national meeting. This report did not generate the same controversy. The results for the two insulin treatment groups were initially released at the time of the presentation of the tolbutamide findings and the lack of differences among the insulin groups and the diet alone group did not change substantially from that time through the end of follow-up. The reports on insulin therapy reopened the challenges to the study. The UGDP investigators' perspective on the extended controversy was summarized in the detailed report on the insulin findings [16].

Since the last UGDP major report was published in 1982, the Diabetes Control and Complication Trial [8] has provided strong evidence that control of blood glucose by diet and insulin will delay microvascular diabetic complications. Such evidence has not yet been provided from a study of oral hypoglycemics or adult-onset noninsulin-dependent diabetics.

Acknowledgment

The tables and figures in this article reproduced by permission of the American Diabetes Association.

References

[1] Bilstad, J.M., Gurian, J.M., Lisook, A.B., Litt, B.D. & Shanahan, E.J. (1978). *The Food and Drug Administration Audit of the University Group Diabetes Project*. US Department of Health, Education, and Welfare, October 16.

[2] Canner, P.L. (1977). Monitoring treatment differences in long-term clinical trials, *Biometrics* **33**, 603–615.

[3] Canner, P.L. (1983). Monitoring of the data for evidence of adverse or beneficial treatment effects, *Controlled Clinical Trials* **4**, 467–483.

[4] Cornfield, J. (1969). The Bayesian outlook and its applications, *Biometrics* **25**, 617–657.

[5] Meinert, C.L. (1986). Impact of clinical trials on the practice of medicine, in *Clinical Trials: Design, Conduct and Analysis*. Oxford University Press, Oxford, Chapter 7, pp. 52–61.

[6] Prout, T.E., Knatterud, G.L., Meinert, C.L. & Klimt, C.R. (1972). The UGDP controversy: clinical trials versus clinical implications, *Diabetes* **21**, 1035–1040.

[7] Report of the Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents (1975). *Journal of the American Medical Association* **231**, 583–608.

[8] The Diabetes Control and Complications Trial Research Group (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus, *New England Journal of Medicine* **329**, 977–986.

[9] University Group Diabetes Program I (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes, I: Design, methods and baseline characteristics, *Diabetes* **19**, Supplement 2, 747–783.

[10] University Group Diabetes Program II (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II: Mortality results, *Diabetes* **19**, Supplement 2, 785–830.

[11] University Group Diabetes Program III (1971). Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes, III. Clinical implications of UGDP results, *Journal of the American Medical Association* **218**, 1400–1410.

[12] University Group Diabetes Program IV (1971). Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. IV. A preliminary report on phenformin results, *Journal of the American Medical Association* **217**, 777–784.

[13] University Group Diabetes Program V (1975). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. V. Evaluation of phenformin therapy, *Diabetes* **24**, Supplement 1, 65–184.

[14] University Group Diabetes Program VI (1976). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. VI. Supplementary report on nonfatal events in patients treated with tolbutamide, *Diabetes* **25**, 1129–53.

[15] University Group Diabetes Program VII (1978). Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes, VII. Mortality and selected nonfatal events with insulin treatment, *Journal of the American Medical Association* **240**, 37–42.

[16] University Group Diabetes Program VIII (1982). Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. VIII. Evaluation of insulin therapy: final report, *Diabetes* **31**, Supplement 5, 1–81.

GENELL L. KNATTERUD

Up-and-Down Method

In biological experiments, often the response is dichotomous (or **binary**); for example, death or survival of an animal exposed to a toxic substance. In the standard situation, there are k levels of exposure, where, at the ith level, $n_i + m_i$ animals are placed, of which n_i show a positive response, for instance by death. The probability of a positive response, p_i, is a function of the exposure level, Z_i. If, over the k levels of response, animals are independent within and among these levels, the **likelihood** over the entire experiment is

$$L = \prod_{i=1}^{k} \binom{n_i + m_i}{n_i} p_i^{n_i}(Z_i) q_i^{m_i}(Z_i), \qquad (1)$$

where $q_i(\cdot) = 1 - p_i(\cdot)$, $i = 1, \ldots, k$.

It should be noted, concerning (1), that Z_i, instead of being a scalar quantity such as exposure level, is more generally a vector that includes other covariates related to response. Furthermore, $p_i(Z_i)$ can be represented by a plethora of possible models (*see* **Quantal Response Models**). Three examples are probit analysis [11], **logistic regression** [19], and reliability growth [16].

Dixon & Mood [10] describe a modification of (1), for which the response of the ith animal at exposure level Z_i determines the next exposure level. More precisely, assume that the first animal receives a particular stimulus at a dose level Z_0. Should the animal respond positively, the next animal receives a smaller dose Z_{-1}, whereas a negative response implies that the next animal receives a larger dose Z_1. This process is then repeated, sequentially, with either Z_{-1} or Z_1 as the dose for the next animal, depending on whether the first animal responded positively or negatively to Z_0. (Compare this with the play-the-winner rule [36] (*see* **Adaptive and Dynamic Methods of Treatment Assignment**) and **sequential analysis** [34].)

The basic theory assumes an initial dose level Z_0 and dose increments of amount d. Hence, the ith dose level (or logarithm thereof) is

$$Z_i = Z_0 + jd, \qquad (2)$$

$i = 0, 1, \ldots$, where j is the excess of negative over positive responses at doses 0 to $i - 1$, inclusive. The likelihood is

$$L(n, m|Z_0) = K \prod_i p_i^{n_i} q_i^{m_i}, \qquad (3)$$

where $q_i = 1 - p_i$ and K is a constant, independent of the p_i and q_i. Furthermore, in probit analysis the tolerance (dose or logdose at which an animal would respond positively; *see* **Quantal Response Models**) is assumed to be normally distributed with mean and variance μ and σ^2, respectively.

Hence, if Z_i is the dosage metameter that is normally distributed,

$$q_i = \Phi\left(\frac{Z_i - \mu}{\sigma}\right), \qquad (4)$$

$i = 0, 1, \ldots$. The parameters μ and σ^2, and hence p_i and q_i, are estimated by **maximum likelihood**. Details of the procedure are found in [10]. (*See* **Median Effective Dose**.)

Example

Table 1 displays a hypothetical up-and-down experiment to estimate the LD_{50} (the dose that produces a response in 50% of all subjects under test) of a new analgesic. First, a series of concentrations of 1%, 4%, 8%, 16%, 32%, and 64% was used. Furthermore, the number of tests performed in this series is $N' = 8$. However, as Dixon [7, 8] indicates, it is convenient to reduce the nominal sample size by one less than the number of like responses at the commencement of the trial. This provides a nominal sample size of $N = 6$. There is no loss in information in estimating an appropriate dosage level since all early responses are used in the tabled estimates.

In this example, an estimate of the LD_{50} is obtained as $Z_f + \theta d$ where Z_f is the final dose that

Table 1 Example of testing an analgesic

Log dose	Results of tests[a]							
1.806								
1.505				+				
1.204			−		+		+	
0.903		−				−		−
0.602	−							
0								

[a]+ indicates a positive response, − indicates a negative response.

Table 2 Application of the up-and-down method in biomedical studies

Study area	Application	References
Anesthesiology	Determination of the minimum alveolar concentration	[24, 28], and [30]
	Anesthesia dosing	[20, 31], and [33]
Toxicity studies	Drug dosing	[3]
	Determination of the LD_{50} of toxic agents	[21] and [35]
Visual studies	Determination of visual perception	[13, 18, 29], and [32]
	Visual acuity studies	[17, 23], and [25]
	Visual testing studies	[15, 17], and [38]
	Visual threshold determination	[27]
Auditory studies	Auditory perception	[1, 6], and [22]
	Auditory facilitation	[26]
	Auditory awakening threshold	[37]
Miscellaneous	Taste testing in diabetics	[12]
	Analysis of dental adhesive stress	[2]
	Determination of defibrillation threshold in dogs	[5]
	Psychophysical pain assessment	[14]
	Pain measurement	[4]

is administered, d is the common interval between doses and θ is obtained from [9, Table 19–3], or [8, Table 1]. Thus, $\theta = 0.831$ and $Z_f = 1.153$.

Applications

The up-and-down (or "staircase") method is a very practical and useful technique that has enjoyed a wide variety of scientific applications. In particular, many biomedical applications are featured. Most of these applications indicate that the up-and-down method saves both time and money as well as providing improved subject tolerance when compared with alternate sampling methods. Table 2 indicates a selection of the type and number of biomedical studies that have used the up-and-down method. While Table 2 is not an all-inclusive list of the biomedical applications of the up-and-down method, it does indicate the wide application and usefulness of this technique.

References

[1] Aoki, K. (1991). Visual scoring in the auditory brainstem response. Second report: determination of response threshold (Japanese), *Nippon Jibiinkoka Gakkai Kaiho (Journal of the Oto-Rhino-Laryngological Society of Japan)* **94**, 705–711.

[2] Aquilino, S.A., Diaz-Arnold, A.M. & Piotrowski, T.J. (1991). Tensile fatigue limits of prosthodontic adhesives, *Journal of Dental Research* **70**, 208–210.

[3] Boissel, J.P., Durieu, L., Girard, P., Nony, P., Chauvin, F. & Haugh, M. (1995). Dose-ranging trials: guidelines for data collection and standardized descriptions, *Controlled Clinical Trials* **16**, 319–330.

[4] Chaplan, S.R., Bach, F.W., Pogrel, J.W., Chung, J.M. & Yaksh, T.L. (1994). Quantitative assessment of tactile allodynia in the rat paw, *Journal of Neuroscience Methods* **53**, 55–63.

[5] Chen, P.S., Feld, G.K., Mower, M.M. & Peters, B.B. (1991). Effects of pacing rate and timing of defibrillation shock on the relation between the defibrillation threshold and the upper limit of vulnerability in open chest dogs, *Journal of the American College of Cardiology* **18**, 1555–1563.

[6] Chiasson, C.R. & Davis, R.I. (1991). Speech recognition in noise for hearing-impaired subjects: effects of an adaptive filter hearing aid, *Journal of the American Academy of Audiology* **2**, 146–150.

[7] Dixon, W.J. (1965). The up- and-down method for small samples, *Journal of the American Statistical Association* **60**, 967–978.

[8] Dixon, W.J. (1989). Staircase method, in *Encyclopedia of Statistical Sciences*, Vol. 8. S. Kotz, N.L. Johnson & C.B. Read, eds. Wiley, New York, pp. 622–625.

[9] Dixon, W.J. & Massey, F.J. (1969). *Introduction to Statistical Analysis*, 3rd Ed. McGraw-Hill, New York, pp. 380–393.

[10] Dixon, W.J. & Mood, A.M. (1948). A method for obtaining and analyzing sensitivity data, *Journal of the American Statistical Association* **43**, 109–126.

[11] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, Cambridge.

[12] Fontvielle, A.M., Faurion, A., Helal, I., Rizkalla, S.W., Falgon, S., Letanoux, M., Tchobroutsky, G. & Slama, G. (1989). Relative sweetness of fructose compared to sucrose in healthy and diabetic subjects, *Diabetes Care* **12**, 481–486.

[13] Gardner, R.M. & Morrell, J., Jr (1991). Body-size judgments and eye movements associated with looking at body regions in obese and normal weight subjects, *Perceptual and Motor Skills* **73**, 675–682.

[14] Gracely, R.H., Lota, L., Walter, D.J. & Dubner, R. (1988). A multiple random staircase method of psychophysical pain assessment, *Pain* **32**, 55–63.

[15] Griswold, M.S. & Stark, W.S. (1992). Scotopic spectral sensitivity of phakic and aphakic observers extending into the near ultraviolet, *Vision Research* **32**, 1739–1743.

[16] Gross, A.J. & Clark, V.A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley, New York.

[17] Harris, L.R. & Lott, L.A. (1995). Sensitivity to full-field visual movement compatible with head rotation: variations among axes of rotation, *Visual Neuroscience* **12**, 743–754.

[18] Hirata, T. (1982). *Tohoku Psychologica Folia* **41**, 35–41.

[19] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[20] Inomata, S., Watanabe, S., Taguchi, M. & Okada, M. (1994). End-tidal sevoflurane concentration for tracheal intubation and minimum alveolar concentration in pediatric patients, *Anesthesiology* **80**, 93–96.

[21] Jung, H. & Choi, S.C. (1994). *Journal of Biopharmaceutical Statistics* **4**, 19–30.

[22] Kochanek, K., Grzanka, A., Dawidowicz, J., Jaskiewicz, M., Zajac, J. & Mika, U. (1992). Wplyw rodzaju bodzca dzwiekowego na oznaczanie progue slichowego metoda ABR. (The effect of brief tone envelopes on ABR and behavioral thresholds), *Otolaryngologia Polska* **46**, 296–301.

[23] Lam, A.K.C., Chau, A.S.Y., Lam, W.Y., Leung, G.Y.O. & Man, B.S.H. (1996). Effect of naturally occurring visual acuity differences between two eyes in stereoacuity, *Ophthalmic and Physiological Optics* **16**, 189–195.

[24] Licina, M.G., Schubert, A., Tobin, J.E., Nicodemus, H.F. & Spitzer, L. (1991). Intrathecal morphine does not reduce minimum alveolar concentration of halothane in humans: results of a double-blind study, *Anesthesiology* **74**, 660–663.

[25] Meese, T.S. & Georgeson, M.A. (1996). The tilt after-effect in plaids and gratings: channel codes, local signs and "patchwise" transforms, *Vision Research* **36**, 1421–1437.

[26] Miskiewicz, A., Buus, S. & Florentine, M. (1994). Auditory facilitation: procedural or sensory effect?, *Journal of the Acoustical Society of America* **96**, 1429–1434.

[27] Mitrani, L., Yakimoff, N. & Shekerdjiiski, S. (1987). Determination of the threshold of simultaneity and temporal order using the "step-ladder" method, *Acta Physiologica et Pharmacologica Bulgarica* **13**, 36–39.

[28] Murray, D.J., Mehta, M.P. & Forbes, R.B. (1991). The additive contribution of nitrous oxide to isoflurane MAC in infants and children, *Anesthesiology* **75**, 186–190.

[29] Pressey, A. & Martin, N.S. (1990). The effects of varying fins in Muller–Lyer and Holding illusions, *Psychological Research* **52**, 46–53.

[30] Rampil, I.J., Lockhart, S.H., Zwass, M.S., Peterson, N., Yasuda, N., Eger, E.I., II, Weiskopf, R.B. & Damask, M.C. (1991). Clinical characteristics of desflurane in surgical patients: minimum alveolar concentration, *Anesthesiology* **74**, 429–433.

[31] Sebel, P.S., Glass, P.S., Fletcher, J.E., Murphy, M.R., Gallagher, C. & Quill, T. (1992). Reduction of the MAC of desflurane with fentanyl, *Anesthesiology* **76**, 52–59.

[32] Sokol, S., Moskowitz, A., McCormack, G. & Augliere, R. (1988). Infant grating acuity is temporally tuned, *Vision Research* **28**, 1357–1366.

[33] Taguchi, M., Watanabe, S., Asakura, N. & Inomata, S. (1994). End-tidal sevoflurane concentrations for laryngeal mask airway insertion and for tracheal intubation in children, *Anesthesiology* **81**, 628–631.

[34] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

[35] Yam, J., Reer, P.J. & Bruce, R.D. (1991). Comparison of the up-and-down method and the fixed-dose procedure for acute oral toxicity testing, *Food and Chemical Toxicology* **29**, 259–263.

[36] Zelen, M. (1969). Play the winner rule and the controled clinical trial, *Journal of the American Statistical Association* **64**, 131–146.

[37] Zepelin, H., McDonald, C.S. & Zammit, G.K. (1984). Effects of age on auditory awakening thresholds, *Journal of Gerontology* **39**, 294–300.

[38] Zhang, L. & Sturr, J.F. (1995). Aging, background luminance, and threshold-duration functions for detection of low spatial frequency sinusoidal gratings, *Optometry and Vision Science* **72**, 198–204.

(*See also* **Stochastic Approximation**)

ALAN J. GROSS & DAVID C. MCLEAN, JR

U-Shaped Distribution

A U-shaped distribution is a **probability** distribution or **frequency distribution** shaped, more or less, like a letter U, although not necessarily symmetrical. Such a distribution has its greatest frequencies at the two extremes of the range of the variable. An example of a variable which commonly has a U-shaped distribution and is often encountered in medical research is the Barthel index; this is a **quality-of-life** measure used to assess the ability of a patient to perform daily activities. A score of zero corresponds to complete dependence on others (and, in some investigations, to the death of a patient), and a score of 100 implies that the patient can perform all usual daily activities without assistance. An example of the distribution of the Barthel index for the patients in a particular investigation is shown in Figure 1.

Variables having U-shaped distributions, such as the Barthel index, are sometimes referred to in the statistical/medical literature as *bounded scores* (see [2]); that is, scores bounded below and above, in which the bounds can and will be attained in a nontrivial proportion of the population. Such variables might be analyzed by considering patients scoring zero (or the lower bound) separately from the others. In some cases, this will correspond to analyzing the mortality rate separately from the values of the variable amongst survivors. Alternately, the two-sample **Wilcoxon–Mann–Whitney** rank sum test might be used to test for a difference between, for example, a group given an active treatment and one given a placebo. Some discussion of the appropriateness of this approach is given in [1].

The problem of calculating sample sizes in studies with bounded outcome scores having U-shaped distributions is taken up in [2].

References

[1] Hilton, J.F. (1996). The appropriateness of the Wilcoxon test in ordinal data, *Statistics in Medicine* **15**, 631–645.

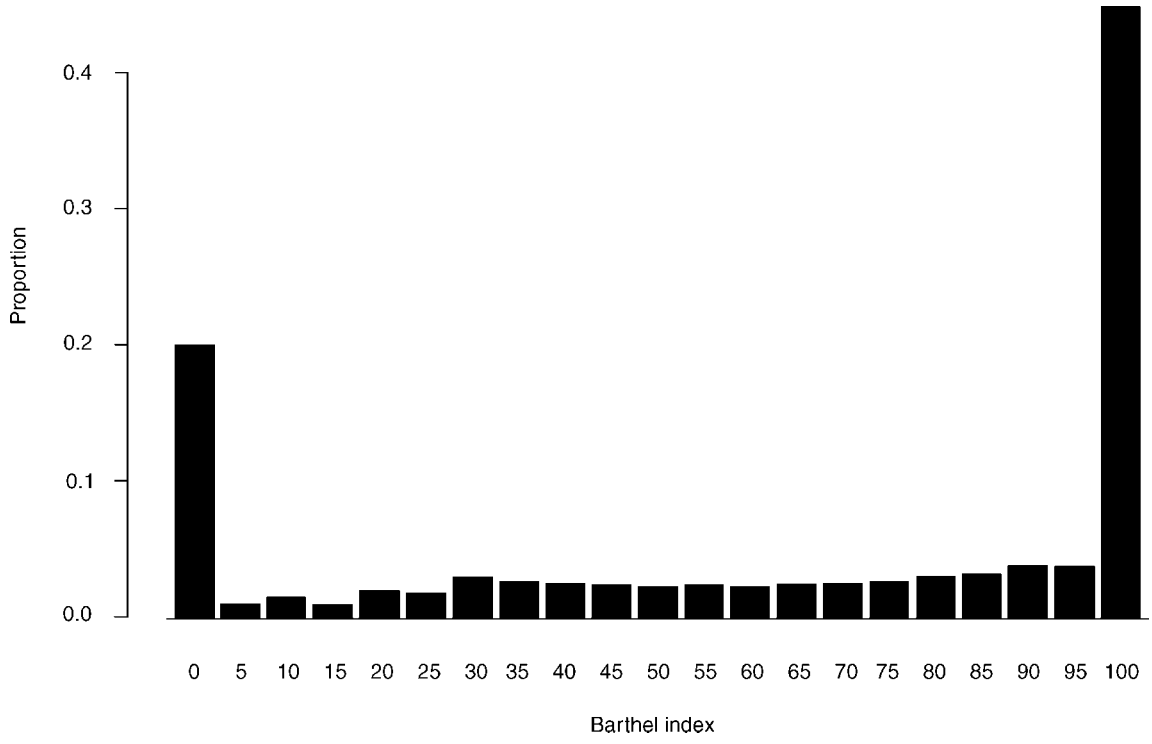

Figure 1 The distribution of the Barthel index

[2] Lesaffre, E., Scheys, I., Frölich, J. & Bluhmki, E. (1993). Calculation of power and sample size with bounded outcome scores, *Statistics in Medicine* **12**, 1063–1078.

BRIAN S. EVERITT

U-Statistics

Given a **random sample** (a sequence of independent **random variables** $X_1, \ldots, X_n$ with common distribution function F), the study of the statistical properties of the sample mean, $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, is a well-established part of probability theory. The notion of averaging over the observations has been generalized by Hoeffding [14] in the following way: given a measurable real-valued function h, symmetric in its m arguments, a *U-statistic* is obtained by averaging over the outcomes $h(X_{i_1}, \ldots, X_{i_m})$ where $(i_1, \ldots, i_m) \in C_{nm} = \{(i_1, \ldots, i_m) \in IN^m : 1 \leq i_1 < \ldots < i_m \leq n\}$, i.e.

$$U_n = \binom{n}{m}^{-1} \sum_{C_{nm}} h(X_{i_1}, \ldots, X_{i_m}).$$

Note that, because of the symmetry of h (a nonrestrictive assumption), it is sufficient to average over the *ordered* m-tuples. U_n is called a U-statistic with *kernel* h of *degree* m. We assume, of course, that $n \geq m$.

Many statistics in estimation and testing theory can be represented as U-statistics. We give two illustrations.

Example 1

Assume $0 < \sigma^2 = \mathrm{var}(X_1) < \infty$. The sample variance $S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$, the **minimum variance unbiased estimator** for σ^2, can be rewritten as

$$S_n^2 = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \frac{(X_i - X_j)^2}{2}.$$

Therefore, the sample variance is a U-statistic with kernel $h(x, y) = (x - y)^2/2$. In general, we have that the minimum variance unbiased estimator of the mth central **moment** is a U-statistic with kernel of degree m. See, for example, Hoeffding [14, p. 295] and Serfling [21, p. 176] for details.

Example 2

The Cramér–von Mises statistic (*see* **Kolmogorov–Smirnov and Cramer–Von Mises Tests in Survival Analysis**), a **goodness-of-fit** statistic to test if the unknown distribution function F equals some specified distribution function F_0, is given by

$$V_n = \int_{-\infty}^{+\infty} [F_n(x) - F_0(x)]^2 \, \mathrm{d}F_0(x),$$

with $F_n(x) = n^{-1} \sum_{i=1}^{n} \mathrm{I}\{X_i \leq x\}$, the empirical distribution function of the sample $X_1, \ldots, X_n$. With

$$h(x, y) = \int_{-\infty}^{+\infty} [\mathrm{I}\{x \leq t\} - F_0(t)][\mathrm{I}\{y \leq t\} - F_0(t)] \, \mathrm{d}F_0(t)$$

we can write $V_n = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j)$. An asymptotically equivalent statistic is the U-statistic

$$U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h(X_i, X_j).$$

See de Wet [7] for a detailed discussion.

For both examples we have that the parameter of interest is of form

$$\theta(F) = Eh(X_1, X_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x, y) \, \mathrm{d}F(x) \, \mathrm{d}F(y).$$

With h as in Example 1 we have $\theta(F) = \sigma^2$, and with h as in Example 2 the goodness-of-fit parameter is $\theta(F) = \int_{-\infty}^{+\infty} [F(x) - F_0(x)]^2 \, \mathrm{d}F_0(x)$. Under the null hypothesis $F = F_0$ we have $\theta(F_0) = 0$. If, in general, a real-valued functional θ defined on a set $\mathcal{F}$ of distribution functions can be written as the expectation with respect to $F \in \mathcal{F}$ of a properly chosen kernel h of degree m, the functional θ is called a *regular functional*. Such functionals have U-statistics as minimum variance unbiased estimators. See Lee [19, Chapter 1] for details. His book also includes a variety of further examples (Chapter 6).

Note that a naive estimator for $\theta(F)$ can be obtained by the plug-in method (replace F by F_n), i.e. use $\theta(F_n)$ as an estimator for $\theta(F)$. The resulting (biased) estimator is the von Mises statistic. The goodness-of-fit statistic, V_n, in Example 2 is a plug-in estimator. U-statistics and von Mises statistics are closely related.

A U-statistic with kernel of degree m can be written in terms of uncorrelated U-statistics of degree $1, \ldots, m$. In fact,

$$U_n - \theta(F) = \sum_{c=1}^{m} \binom{m}{c} U_{cn},$$

with

$$U_{cn} = \binom{n}{c}^{-1} S_{cn} = \binom{n}{c}^{-1} \sum_{C_{nc}} h_c(X_{i_1}, \ldots, X_{i_c}).$$

See Lee [19, Section 1.6] for the explicit expression of h_c and an excellent further discussion. This *H-decomposition* is due to Hoeffding [15]. Other important structural properties are the *forward martingale structure* of $\{S_{cn}, \mathcal{F}_n\}_{n \geq c}$, with $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ and the *reverse martingale* structure of $\{U_n, \tilde{\mathcal{F}}_n\}_{n \geq m}$, with $\tilde{\mathcal{F}}_n = \sigma(X_{1:n}, \ldots, X_{n:n}, X_{n+1}, X_{n+2}, \ldots)$ and $X_{i:n}$ the ith **order statistic** of $X_1, \ldots, X_n$ [19, Section 3.4] (*see* discussion of martingales in **Counting Process Methods in Survival Analysis**).

So far we have demonstrated that many statistics are in fact U-statistics and we have discussed some structural properties. Also highly relevant is the appearance of U-statistics as terms in stochastic approximations of smooth statistics. U-statistics are, for example, extremely useful to approximate important estimators in nonparametric **density estimation** and **nonparametric regression** theory (see, for example, [13] and [20]) and **survival analysis** (see, for example, [5]). The basic idea is that the estimator of interest can be approximated by a sum of uncorrelated U-statistics. This idea is closely related to the H-decomposition of a U-statistic (see [19, Section 4.1] and [9]) and to von Mises expansions, a generalization of the projection method (a technique discussed in more detail below). For further reading we refer to [21, Chapter 6] and [10].

A more detailed discussion would require a number of technical concepts and definitions. We therefore restrict ourselves to one illustration.

Example 3

Let $T_1, \ldots, T_n$ denote iid nonnegative survival times with a continuous distribution function F and let $C_1, \ldots, C_n$ denote iid nonnegative censoring times with a continuous distribution function G. For $i = 1, \ldots, n$, we denote $X_i = \min(T_i, C_i)$ and $\delta_i = \mathrm{I}\{T_i \leq C_i\}$. Let $\hat{F}_n(t)$ denote the product-limit or **Kaplan–Meier estimator** for $F(t)$. With $\hat{\Lambda}_n(t)$ the **Nelson–Aalen estimator** and $\Lambda(t)$ the cumulative hazard function, a U-statistic representation has been established in [5] for $\hat{\Lambda}_n(t) - \Lambda(t)$. On the basis of the relation

$$\hat{F}_n(t) - F(t) = \exp[-\Lambda(t)] \times \{1 - \exp[-(\Lambda_n(t) - \Lambda(t)]\}$$

and using Taylor expansion ideas, a U-statistic representation for the Kaplan–Meier estimator can be obtained.

Asymptotic Properties

A basic contribution to the study of the asymptotic behavior of U-statistics (*see* **Large-sample Theory**) is the following result.

Theorem 1. If $E|h(X_1, \ldots, X_m)| < \infty$, then $U_n \to \theta(F)$ almost surely (a.s.).

This theorem states that the classical strong **law of large numbers** for the sample mean generalizes to U-statistics. Different proofs are available. They rely on the martingale structure of U-statistics mentioned above. For full proofs and references to the original papers, see Lee [19, Section 3.4].

Next, we briefly discuss the asymptotic distribution theory for U-statistics. The limit distribution of a (properly standardized) U-statistic will be Gaussian if we can obtain a stochastic approximation, $\hat{U}_n$, of iid structure that is close to U_n (in the sense that U_n inherits the asymptotic distributional behavior of $\hat{U}_n$). The appropriate approximation is obtained from the projection technique, which is in fact the first term in the H-decomposition. We have

$$\hat{U}_n = \sum_{i=1}^{n} \mathrm{E}(U_n | X_i) - (n-1)\theta(F).$$

With

$$h_1(x) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} h(x, x_2, \ldots, x_m)\, \mathrm{d}F(x_2) \ldots \times \mathrm{d}F(x_m) - \theta(F)$$

we can write

$$\hat{U}_n - \theta(F) = \frac{m}{n} \sum_{i=1}^{n} h_1(X_i).$$

If $h_1 \equiv 0$, then the U-statistic is called *degenerate* or *pure*; otherwise, the U-statistic is *nondegenerate*. Pure U-statistics do not admit an iid approximation,

and as a consequence the limit distribution is not Gaussian. For nondegenerate U-statistics the following central limit result is valid.

Theorem 2. [14]. If $Eh^2(X_1, \ldots, X_m) < \infty$ and $\zeta_1 = \text{var}h_1(X_1) > 0$ (U_n is nondegenerate), then

$$\frac{\sqrt{n}[U_n - \theta(F)]}{(m\zeta_1^{1/2})} \xrightarrow{d} Z,$$

with Z a standard normal random variable.
A simple calculation shows that

$$\zeta_1 = \text{E}\{[h(X_1, X_2, \ldots, X_m) - \theta(F)] \times [h(X_1, X_{m+1}, \ldots, X_{2m-1}) - \theta(F)]\}.$$

For a pure U-statistic (the first term in the H-decomposition vanishes and $\zeta_1 = 0$) with $\zeta_2 = \text{E}\{[h(X_1, X_2, X_3, \ldots, X_m) - \theta(F)]\ [h(X_1, X_2, X_{m+1}, \ldots, X_{2m-2}) - \theta(F)]\} > 0$, the H-decomposition is

$$U_n - \theta(F) = \frac{m(m-1)}{n(n-1)} \sum_{1 \le i < j \le n} h_2(X_i, X_j) + \sum_{c=3}^{m} \binom{m}{c} U_{cn}.$$

For h_2, define the operator

$$Az(x) = \int_{-\infty}^{+\infty} h_2(x, y) z(y)\, \mathrm{d}F(y)$$

with z square integrable with respect to F. Let $\lambda_1, \lambda_2, \ldots$ denote the (not necessarily distinct) **eigenvalues** corresponding to the distinct solutions $z_1, z_2, \ldots$ of the equation $Az = \lambda z$.

Theorem 3. [12]. If $\text{E}[h^2(X_1, \ldots, X_m)] < \infty$ and $\zeta_1 = 0 < \zeta_2$, then

$$n[U_n - \theta(F)] \xrightarrow{d} \frac{m(m-1)}{2} Y,$$

with Y a random variable of the form $Y = \sum_{j=1}^{\infty} \lambda_j [\chi_j^2(1) - 1]$, where $\chi_1^2(1), \chi_2^2(1), \ldots$ are independent $\chi^2(1)$ random variables (*see* **Chi-square Distribution**; **Convergence in Distribution and in Probability**).

Example 1 (Continued)

For the sample variance an application of Theorem 2 yields (with μ_k the kth central moment): if $\mu_4 < \infty$ and $\mu_4 - \mu_2^2 > 0$, then $\sqrt{n}(S_n^2 - \mu_2)$ has a limiting normal distribution with mean zero and variance $\mu_4 - \mu_2^2$.

Example 2 (Continued)

Under the null hypothesis $F = F_0$ the Cramér–von Mises statistic is easily seen to be a pure U-statistic. Theorem 3 is applicable, the eigenvalues are $\lambda_j = (j\pi)^{-2}$. See [7] for details.

Remarks and Extensions

1. For U-statistics with a kernel of degree $m > 2$, more terms in the H-decomposition might vanish (higher order degeneracy). Asymptotic distribution theory has been established. The resulting limit distributions are characterized in terms of multiple Wiener integrals [8].
2. We reviewed some basic results for one-sample U-statistics. Extensions to multisample or generalized U-statistics are available. See the books by Lee [19], Koroljuk & Borovskich [18] and Borovskikh [4] for details. These books also deal with other variations on the theme: incomplete U-statistics, random U-statistics, weighted U-statistics, generalized L-statistics, Edgeworth expansions for U-statistics, etc.
3. **Bootstrap** theory for U-statistics is reviewed in Janssen [17]. Bickel & Freedman [3] is a basic reference.
4. A further important topic, especially for applications in nonparametric density and regression estimation, is the study of U-statistics with the kernel depending on the sample size n. Key references are Jammalamadaka & Janson [16] and Mammen [20]. We also mention the work by Frees [11] on infinite order U-statistics.
5. In Serfling [22] the study of U-processes and U-**quantiles** is initiated. Important recent contributions on U-processes and U-quantiles include Arcones & Giné [2], Stute [23], and Arcones [1]. Key words in the development of new results for U-processes are martingales and decoupling. For details we refer to the book by de la Peña & Giné [6].

References

[1] Arcones, M. (1995). The asymptotic accuracy of the bootstrap U-quantiles, *Annals of Statistics* **23**, 1802–1822.

[2] Arcones, M. & Giné, E. (1993). Limit theorems for U-processes, *Annals of Probability* **21**, 1494–1542.

[3] Bickel, P. & Freedman, D. (1981). Some asymptotic theory for the bootstrap, *Annals of Statistics* **9**, 1196–1217.

[4] Borovskikh, Yu.V. (1996). *U-Statistics in Banach Spaces*. VSP, Utrecht.

[5] Chang, M.N. & Rao, P.V. (1989). Berry-Esseen bound for the Kaplan-Meier estimator, *Communications in Statistics–Theory and Methods* **18**, 4647–4664.

[6] de la Peña, V. & Giné, E. (1998). *An Introduction to Decoupling Inequalities with Applications*. Springer-Verlag, New York.

[7] de Wet, T. (1987). Degenerate U- and V-statistics, *South African Statistical Journal* **21**, 99–129.

[8] Dynkin, E.B. & Mandelbaum, A. (1983). Symmetric statistics, Poisson point processes, and multiple Wiener integrals, *Annals of Statistics* **11**, 739–745.

[9] Efron, B. & Stein, C. (1981). The jackknife estimate of variance, *Annals of Statistics* **9**, 586–596.

[10] Fernholz, L. (1983). *Von Mises Calculus for Statistical Functionals*. Springer-Verlag, New York.

[11] Frees, E.W. (1989). Infinite order U-statistics, *Scandinavian Journal of Statistics* **16**, 29–45.

[12] Gregory, G. (1977). Large sample theory for U-statistics and tests of fit, *Annals of Statistics* **5**, 110–123.

[13] Härdle, W. & Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association* **84**, 986–995.

[14] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *Annals of Mathematical Statistics* **19**, 293–325.

[15] Hoeffding, W. (1961). The strong law of large numbers for U-statistics, *University of North Carolina Institute of Statistics Mimeo Series No. 302*.

[16] Jammalamadaka, S.R. & Janson, S. (1986). Limit theorems for a triangular scheme of U-statistics with applications to interpoint distances, *Annals of Probability* **14**, 1347–1358.

[17] Janssen, P. (1997). Bootstrapping U-statistics. *South African Statistical Journal*, to appear.

[18] Koroljuk, V.S. & Borovskich, Yu.V. (1994). *Theory of U-Statistics*. Kluwer Academic Publishers, Dordrecht.

[19] Lee, A.J. (1990). *U-Statistics*. Marcel Dekker, New York.

[20] Mammen, E. (1992). *When Does the Bootstrap Work?* Springer-Verlag, New York.

[21] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[22] Serfling, R. (1984). Generalized $L-$, M- and R-statistics, *Annals of Statistics* **12**, 76–86.

[23] Stute, W. (1994). U-statistic processes: a martingale approach, *Annals of Probability* **22**, 1725–1744.

PAUL JANSSEN

Utility in Health Studies

Utility is a technical term from economics and **decision theory** with a very precise meaning based on an underlying theory to be described later. However, in a general way, utility can be thought of as a measure of strength of preference. Utilities are applicable, and are used, in all sectors of the economy including business, defense, environment, education, and health. Applications of utility in **pharmacoepidemiology** do not differ from applications of utility to health in general. Accordingly, the utility material described in this article applies not just to pharmacotherapy but equally well to all interventions or programs designed to improve health.

In health applications the preferences generally relate to different outcomes that can be achieved, but could equally well refer to different programs or treatments. As an example of utilities for outcomes, if a particular individual prefers outcome A to outcome B, and in turn prefers outcome B to outcome C, cardinal numbers can be assigned to each outcome such that they represent the preferences of the individual on an interval scale. For example, if the numbers so assigned are $A = 12$, $B = 6$, and $C = 4$, then they would indicate that the individual prefers A to B and B to C, and that their preference difference between A and B (6 units) is three times as great as their preference difference between B and C (2 units). Note that it would not indicate that the person prefers A twice as much as B, because the preference scale is only an interval scale and not a ratio scale (*see* **Measurement Scale**).

More formally, utility can be defined by a precise set of axioms that form the foundation of expected utility theory [36, 64, 71]. The axioms represent a fundamental statement or definition of consistent and rational decision making under uncertainty. That is, they represent compelling rules that are widely seen as logical and appropriate for rational decision making when there are uncertainties regarding the outcomes. Clearly this applies to decision making in health – hence the view that utilities can play a useful role in the analysis of alternative courses of action in the field of health (*see* **Decision Analysis in Diagnosis and Treatment Choice**).

When utility is defined formally on the basis of the axioms mentioned above, it is properly referred to as "von Neumann–Morgenstern utilities", or NM utilities for short. NM utilities are measured using a technique called the **standard gamble**. The standard gamble is a direct application of one of the fundamental axioms of expected utility theory. In the standard gamble an individual expresses his or her preference by choosing between two alternatives. For example, the individual described above, who preferred A to B to C, would be asked to choose between one alternative in which outcome B would be received with certainty and a second alternative in which outcome A would be received with probability p and outcome C with probability $(1 - p)$. The probability p would then be varied until the individual was indifferent between these two choices. The indifference probability is used to calculate the NM utility that the individual has for outcome B relative to the utilities for outcomes A and C. Details of the methods are widely available [6, 24, 48, 62].

In addition to the technical and precise use of the term utility to represent NM utilities measured according to the fundamental theory, the term is also used broadly to refer simply to preferences, however measured. This is unfortunate, and causes considerable confusion in the literature. When readers come across the term utility they should first determine whether it is being used in the technical NM sense or in the broader sense of preferences.

We prefer to use the term preference to refer to the broad construct, and the term utility to refer to NM utilities. We also prefer to make a distinction between preferences that are measured with a standard gamble instrument, which contains uncertainty, and other instruments such as the **time trade-off** and rating scales that do not. The former are called utilities, while the latter are called values. However, readers should beware that many writers fail to make this distinction.

Utility as a Measure of Health-Related Quality of Life

Quality of life is an extremely broad concept that includes health, wealth, freedom, environment, political system, family, future prospects, and indeed an endless list of all-encompassing considerations. However, within this list is a subset that is generally known as "health-related quality of life". It is widely accepted that the goal of the health system is to improve both the quantity of life **(life expectancy)**

and the health-related quality of life. Accordingly, many health status instruments have been developed to measure health-related quality of life. A useful taxonomy for these instruments is that developed by Guyatt and colleagues in which the instruments are partitioned into three major sets: specific instruments, generic profiles, and utility measures [18, 28, 30].

Specific instruments include those that are disease-specific, such as the Functional Living Index–Cancer [12] or the Western Ontario–McMaster Osteoarthritis Index [5]. Specific instruments also include those designed for individuals in a particular age group, for example, the Care and Resource Evaluation Tool for the elderly [23]. Specific instruments are generally felt to be the most sensitive and responsive, but they clearly lack generalizability [29].

Generic profiles include instruments such as the Short Form 36 [72], the Sickness Impact Profile [13], and the Nottingham Health Profile [43]. These instruments produce **scores** on a number of different dimensions (profile of scores), are applicable to a wide variety of diseases and individuals, and accordingly are more generalizable but probably less responsive than specific instruments [29].

The third category, utility measures, produces a single summary index that represents health-related quality of life by a cardinal number. Like the generic profiles, these approaches are applicable to a wide variety of diseases and individuals and thus are highly generalizable. However, they are likely to be the least responsive of the three approaches [29].

A major advantage of the utility approach is that the single cardinal score for health-related quality of life can be combined with quantity of life to provide an integrated measure of health improvement that captures both the impact on quantity of life and the impact on quality of life. The usual method of combining quality and quantity of life is to calculate the quality-adjusted life years (QALY) involved, although alternative measures such as the healthy years equivalent [44] or the disability-adjusted life years (DALY) [45] have been suggested. The single measure of QALY, or its alternative, can be used in economic evaluation studies such as cost-effectiveness analysis and cost–utility analysis (*see* **Health Economics**) and can also be used as an **outcome measure in clinical trials** [27, 63].

Which measure of health-related quality of life should be used in a study? Our general answer to that question is that if health-related quality of life is an important outcome for the study, then the study should contain one instrument from each of the three types. The specific instrument is likely to be the most informative at a detailed level, particularly to clinicians interested in the impact on the disease and in specific patients. The generic profile will be useful in comparing the health impact more broadly to other types of patients and other diseases. The utility measure is necessary if one wishes to undertake economic evaluations such as cost–effectiveness or cost–utility analyses. The utility measure is also necessary if the intervention has an impact on both quantity and health-related quality of life and one wishes to have a single effectiveness measure that aggregates both of these effects.

Utility for Economic Evaluation Including Pharmacoeconomics

The primary application of utility is for use in economic evaluation. Economic evaluation is the comparative analysis of the costs and consequences of two or more alternative treatments or programs to improve health [15]. The techniques of economic evaluation include cost analysis, cost–effectiveness analysis, cost–utility analysis, and cost–benefit analysis (*see* **Cost–Benefit Analysis, Willingness to Pay**). When these approaches are applied to pharmacotherapy they have been labeled as pharmacoeconomics. However, the methods are the same whether the intervention being evaluated is a pharmaceutical product or a nondrug therapy.

When quantity of life and health-related quality of life are both important, the principal approach in economic evaluation is to use the utility approach, broadly defined. That is, utilities may be measured according to expected utility theory using the standard gamble instrument, or they may be measured more generally using other instruments such as the time trade-off or rating scales. The most common approach is to use the utility score to combine the quantity and health-related quality of life into a single outcome measure, the quality-adjusted life year (QALY) gained. This is then used as the denominator to determine the cost per QALY gained for one intervention or program as compared to another. The approach is known as cost–effectiveness analysis or cost–utility analysis, depending upon the writer.

CUA is a special case of CEA, and some researchers, including ourselves, distinguish it with its own title, while others do not [27]. Methods for calculating QALYs and for undertaking economic evaluation are widely available [15, 27, 56, 63, 64].

Utility as an Outcome Measure for Clinical Studies

Because utilities can be used to combine the impact on quantity of life and health-related quality of life, it makes an ideal outcome measure for clinical studies that affect both. Utilities have been used in this way as secondary measures of effectiveness for some time [1, 3, 8, 49, 50, 59, 61]. Recently, utility has been designated as the primary effectiveness measure for a planned Canadian randomized **multicenter trial** of lung volume reduction surgery for patients with pulmonary emphysema [46]. We expect that use of utility as a primary clinical endpoint in trials will become more common in the future.

How to Measure Utility

Utilities can be measured directly or indirectly; see [67] for a comprehensive review.

Direct Measurement of Utility

As mentioned above, and discussed elsewhere, utilities can be measured directly using a number of different instruments: the standard gamble, the time trade-off, and the rating scale with its variant, the visual analog scale. The conventional approach to using these instruments is to interview the respondent face-to-face, with a carefully trained and scripted interviewer leading the respondent through a highly structured interview, complete with visual aids and props to elicit the relevant preferences (*see* **Interviewing Techniques**). This approach is elaborate and costly, but has been included in many studies to date [11, 14, 37, 40, 41, 42, 49, 51, 54, 55, 58, 68, 69, 73]. Recently, a number of new and more efficient approaches have been developed and are being tested. These include telephone administration [35], self-administration by pencil and paper [47], and self-administration by an interactive computer program [27, 52]. In fact, a number of commercial systems are now available to provide computer interactive self-administered interviews for utility assessment. Although these are recent developments, they are likely to become the standard approach for direct measurement of utility in the future on many types of respondents.

Indirect Measurement of Utilities

An alternative approach that is being increasingly and widely used in studies is to use one of the multiattribute health status classification systems that include a utility scoring formula [31]. Three such systems are the Quality of Well-Being [30, 31], the Health Utilities Index [17, 22, 25], and the EQ-5D [39, 53]. All of these systems are similar in that they consist of attributes of health status (such as physical function, emotional function, or cognitive function) and levels on each attribute ranging from good function to bad or no function. Patients in studies, or populations under study, are classified into the system, and a scoring formula provides the utility score for the particular combination of levels across attributes. The systems differ in terms of the attributes included, the levels described for each attribute, and the type of scoring formula provided. As one example of the systems, we will describe the Health Utilities Index.

The Health Utilities Index (HUI) has developed through a series of studies over a number of years [9, 10, 19, 20, 65, 66]. The studies defined and refined the attributes and levels in the system, and the methods for determining the utility scoring formula. Currently there are two versions of the system, which are closely interrelated – the HUI2 (Table 1) and the HUI3 (Table 2). Scoring formulae are available for each [22, 69]. HUI2 and HUI3 have been widely used in clinical studies, and HUI3 has also been used widely in population health studies [7, 32, 57, 70, 74]. Thus, population norm data are available for HUI3. Currently, our recommendation is that studies incorporate both systems by using a combined questionnaire, which we have developed. The self- administration version consists of 15 questions and provides all of the information required to map the health status of the respondent into both HUI2 and HUI3. The questionnaire is available in a number of formats: self-administered, interviewer-administered face-to-face, and interviewer-administered by telephone. It is currently available in a wide range of languages, and

Table 1 Health status classification system for HUI2

Attribute	Level	Level description
Sensation	1	Ability to see, hear, and speak normally for age
	2	Requires equipment to see or hear or speak
	3	Sees, hears, or speaks with limitations even with equipment
	4	Blind, deaf, or mute
Mobility	1	Able to walk, bend, lift, jump, and run normally for age
	2	Walks, bends, lifts, jumps, or runs with some limitations but does not require help
	3	Requires mechanical equipment (such as canes, crutches, braces, or wheelchair) to walk or get around independently
	4	Requires the help of another person to walk or get around and requires mechanical equipment as well
	5	Unable to control or use arms and legs
Emotion	1	Generally happy and free from worry
	2	Occasionally fretful, angry, irritable, anxious, depressed, or suffering night terrors
	3	Often fretful, angry, irritable, anxious, depressed, or suffering night terrors
	4	Almost always fretful, angry, irritable, anxious, depressed
	5	Extremely fretful, angry, irritable, or depressed, usually requiring hospitalization or psychiatric institutional care
Cognition	1	Learns and remembers schoolwork normally for age
	2	Learns and remembers schoolwork more slowly than classmates as judged by parents and/or teachers
	3	Learns and remembers very slowly and usually requires special educational assistance
	4	Unable to learn and remember
Self-care	1	Eats, bathes, dresses, and uses the toilet normally for age
	2	Eats, bathes, dresses, or uses the toilet independently with difficulty
	3	Requires mechanical equipment to eat, bathe, dress, or use the toilet independently
	4	Requires the help of another person to eat, bathe, dress, or use the toilet
Pain	1	Free of pain and discomfort
	2	Occasional pain; discomfort relieved by nonprescription drugs or self-control activity without disruption of normal activities
	3	Frequent pain; discomfort relieved by oral medicines with occasional disruption of normal activities
	4	Frequent pain, frequent disruption of normal activities; discomfort requires prescription narcotics for relief
	5	Severe pain; pain not relieved by drugs and constantly disrupts normal activities
Fertility[a]	1	Ability to have children with a fertile spouse
	2	Difficulty in having children with a fertile spouse
	3	Unable to have children with a fertile spouse

[a]Fertility attribute can be deleted if not required. Contact developers for details. Source: Table II in [19]

further translations are under way. The questionnaire takes from 2 min (interviewer-administered telephone version) to under 10 min (self-administered written version) to administer, has been used on thousands of respondents, and is simple to complete. Proxy respondents can be used for patients who are unable to answer the questions on their own.

The HUI system is useful in clinical studies in two ways. First, the classification is useful in its own right and has been widely used as a systematic method

Table 2 Health status classification system for HUI3

Attribute	Level	Level description
Vision	1	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses
	2	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses
	3	Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses
	4	Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses
	5	Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses
	6	Unable to see at all
Hearing	1	Able to hear what is said in a group conversation with at least three other people, without a hearing aid
	2	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people
	3	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people with a hearing aid
	4	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	5	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	6	Unable to hear at all
Speech	1	Able to be understood completely when speaking with strangers or friends
	2	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know the respondent well
	3	Able to be understood partially when speaking with strangers or people who know the respondent well
	4	Unable to be understood when speaking with strangers but able to be understood partially by people who know the respondent well
	5	Unable to be understood when speaking to other people (or unable to speak at all)
Ambulation	1	Able to walk around the neighborhood without difficulty, and without walking equipment
	2	Able to walk around the neighborhood with difficulty, but does not require walking equipment or the help of another person

(*continued overleaf*)

Table 2 (*continued*)

Attribute	Level	Level description
	3	Able to walk around the neighborhood with walking equipment, but without the help of another person
	4	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighborhood
	5	Unable to walk alone, even with walking equipment; able to walk short distances with the help of another person, and requires a wheelchair to get around the neighborhood
	6	Cannot walk at all
Dexterity	1	Full use of two hands and ten fingers
	2	Limitations in the use of hands or fingers, but does not require special tools or help of another person
	3	Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person)
	4	Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools)
	5	Limitations in use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools)
	6	Limitations in use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools)
Emotion	1	Happy and interested in life
	2	Somewhat happy
	3	Somewhat unhappy
	4	Very unhappy
	5	So unhappy that life is not worthwhile
Cognition	1	Able to remember most things, think clearly and solve day to day problems
	2	Able to remember most things, but having a little difficulty when trying to think and solve day to day problems
	3	Somewhat forgetful, but able to think clearly and solve day to day problems
	4	Somewhat forgetful, and having a little difficulty when trying to think or solve day to day problems
	5	Very forgetful, and having great difficulty when trying to think or solve day to day problems
	6	Unable to remember anything at all, and unable to think or solve day to day problems
Pain	1	Free of pain and discomfort
	2	Mild to moderate pain that prevents no activities
	3	Moderate pain that prevents a few activities
	4	Moderate to severe pain that prevents some activities
	5	Severe pain that prevents most activities

Source: Table III in [19].

of describing and comparing patients and monitoring their changes over time [1–4, 21, 26, 38, 60]. In addition, the scoring formula for the HUI is based directly on NM utilities measured on a **random sample** of the community. Thus the HUI scores represent both appropriate utility weights (NM utilities) for calculating quality-adjusted life years and undertaking cost–effectiveness or cost–utility analyses, and they represent the appropriate source of these preferences, i.e. the community at large [27].

Given the ease and simplicity of using these multiattribute systems in clinical studies, we generally recommend their use, rather than the direct measurement of utilities using a standard gamble or other instrument. The latter is generally reserved for studies in academic centers that also have methodologic hypotheses about the utilities being measured. Investigators who simply wish to use utilities are generally better served by using one of the multiattribute systems.

Conclusions

Utilities are being incorporated increasingly into clinical studies. They can be used as a measure of health-related quality of life, as an outcome measure for the trial either in terms of health-related quality of life or in terms of a combined index of quality and quantity of life, and they can be used to undertake economic evaluations of interventions or programs. For most studies the simplest and most appropriate method of obtaining utilities is to use a multiattribute health status classification system that includes a utility scoring formula. In some studies the researchers may wish to measure the utilities directly using one of the preference measurement tools available.

References

[1] Barr, R.D., Feeny, D., Furlong, W., Weitzman, S. & Torrance, G.W. (1995). A preference-based approach to health-related quality of life for children with cancer, *International Journal of Pediatric Hematology/Oncology* **2**, 305–315.

[2] Barr, R., Furlong, W., Feeny, D., Horsman, J., Rosenbaum, P. & Weitzman, S. (1995). Evaluating treatments for childhood cancer, *International Journal of Technology Assessment in Health Care* **11**, 1–10.

[3] Barr, R., Pai, M., Weitzman, S., Feeny, D., Furlong, W., Rosenbaum, P. & Torrance, G. (1994). A multi-attribute approach to health status measurement and clinical management–illustrated by an application to brain tumors in childhood, *International Journal of Oncology* **4**, 639–648.

[4] Barr, R., Furlong, W., Dawson, S., Whitton, A., Strautmanis, I., Pai, M., Feeny, D. & Torrance, G. (1993). An assessment of global health status in survivors of acute lymphoblastic leukemia in childhood, *The American Journal of Pediatric Hematology/Oncology* **15**, 284–290.

[5] Bellamy, N., Buchanan, W., Goldsmith, C.H., Campbell, J. & Stitt, L. (1988). Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to anti-rheumatic drug therapy in patients with osteoarthritis of the hip or knee, *Journal of Rheumatology* **15**, 1833–1840.

[6] Bennett, K.J. & Torrance, G.W. (1996). Measuring health state preferences and utilities: rating scale, time trade-off and standard gamble techniques, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 253–265.

[7] Berthelot, J., Roberge, R. & Wolfson, M. (1993). The calculation of health-adjusted life expectancy for a Canadian province using a multi-attribute utility function: a first attempt, in *Calculation of Health Expectancies: Harmonization, Consensus and Future Perspectives*, Vol. 226, J.M. Robine, C.D. Mathers, M.R. Bone & I. Romieu, eds. John Libbey Eurotext Ltd, pp. 161–172.

[8] Bombardier, C., Ware, J., Russell, I., Larson, M., Chalmers, A. & Reid, J.L. (1986). Auranofin therapy and quality of life in patients with rheumatoid arthritis, *American Journal of Medicine* **81**, 565–578.

[9] Boyle, M., Torrance, G., Sinclair, J. & Horwood, S. (1983). Economic evaluation of neonatal intensive care of very-low-birth-weight infants, *New England Journal of Medicine* **308**, 1330–1337.

[10] Cadman, D., Goldsmith, C., Torrance, G.W., Boyle, M. & Furlong, W. (1986). *Development of a Health Status Index for Ontario Children, Final Report to Ontario Ministry of Health for Research Grant DM648 (00633)*. McMaster University, Centre for Health Economics and Policy Analysis, Hamilton.

[11] Churchill, D., Torrance, G., Taylor, D., Barnes, C., Ludwin, D., Shimizu, A. & Smith, E. (1987). Measurement of quality of life in end-stage renal disease: the time trade-off approach, *Clinical and Investigative Medicine* **10**, 14–20.

[12] Clinch, J.J. (1996). The functional living index–cancer: ten years later, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 215–225.

[13] Damiano, A.M. (1996). The sickness impact profile, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 347–354.

[14] Dolan, P., Gudex, C., Kind, P. & Williams, A. (1996). Valuing health states: a comparison of methods, *Journal of Health Economics* **15**, 209–231.

[15] Drummond, M., O'Brien, B., Stoddart, G. & Torrance, G. (1997). *Methods for the Economic Evaluation of Health Care Programmes*, 2nd Ed. Oxford University Press, Oxford.

[16] EuroQol Group. (1990). EuroQol–a new facility for the measurement of health-related quality of life, *Health Policy* **16**, 199–208.

[17] Feeny, D.H., Torrance, G.W. & Furlong, W.J. (1996). Health utilities index, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 239–252.

[18] Feeny, D.H., Torrance, G.W. & Labelle, R. (1996). Integrating economic evaluations and quality of life assessments, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 85–95.

[19] Feeny, D., Furlong, W., Boyle, M. & Torrance, G. (1995). Multi-attribute health status classification systems: health utilities index, *Pharmacoeconomics* **7**, 490–502.

[20] Feeny, D., Furlong, W., Torrance, G., Rosenbaum, P. & Weitzman, S. (1992). A comprehensive multi-attribute system for classifying the health status of survivors of childhood cancer, *Journal of Clinical Oncology* **10**, 923–928.

[21] Feeny, D., Leiper, A., Barr, R., Furlong, W., Torrance, G., Rosenbaum, P. & Weitzman, S. (1993). The comprehensive assessment of health status in survivors of childhood cancer: application to high-risk acute lymphoblastic leukaemia, *British Journal of Cancer* **67**, 1047–1052.

[22] Feeny, D., Furlong, W., Torrance, G.W., Goldsmith, C.H., Zhu, Z., Depauw, S., Denton, M. & Boyle, M. (2002). Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system, *Med Care* **40**(2), 113–128.

[23] Fretwell, M.D. (1996). Frail older patients: creating standards of care, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 809–817.

[24] Furlong, W., Feeny, D., Torrance, G., Barr, R. & Horsman, J. (1990). *Guide to Design and Development of Health-State Utility Instrumentation, Working Paper No. 90–9*. McMaster University, Centre for Health Economics and Policy Analysis, Hamilton.

[25] Furlong, W.J., Feeny, D.H., Torrance, G.W. & Barr, R.D. (2001). Health Utilities Index(HUI) system for assessing health-related quality of life in clinical studies, *Annals of Medicine* **33**(5), 375–384.

[26] Gemke, R.J.B.J., Bonsel, G.J. & van Vught, A.J. (1995). Long term survival and state of health after paediatric intensive care, *Archives of Disease in Childhood* **73**, 196–201.

[27] Gold, M.R., Siegel, J.E., Russell, L.B. & Weinstein, M.C. (1996). *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York.

[28] Guyatt, G., Feeny, D. & Patrick, D. (1993). Measuring health-related quality of life, *Annals of Internal Medicine* **118**, 622–629.

[29] Guyatt, G.H., Jaeschke, R., Feeny, D.H. & Patrick, D.L. (1996). Measurements in clinical trials: choosing the right approach, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 41–48.

[30] Guyatt, G., Veldhuyzen van Zanten, S., Feeny, D. & Patrick, D. (1989). Measuring quality of life in clinical trials: a taxonomy and review, *Canadian Medical Association Journal* **140**, 1441–1448.

[31] Hawthorne, G. & Richardson, J. (2001). Measuring the value of program outcomes: a review of multiattribute utility measures, *Expert Review of Pharmacoeconomics and Outcomes Research* **1**(2), 215–228.

[32] Hood, S., Beaudet, M. & Catlin, G. (1996). A healthy outlook, *Health Reports* **7**, 25–32.

[33] Kaplan, R. & Anderson, J. (1988). A general health policy model: Update and applications, *Health Services Research* **23**, 203–235.

[34] Kaplan, R.M. & Anderson, J.P. (1996). The general health policy model: an integrated approach, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 309–322.

[35] Katz, J., Phillips, C., Fossel, A. & Liang, M. (1994). Stability and responsiveness of Utility Measures, *Medical Care* **32**, 183–188.

[36] Keeney, R. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York.

[37] Kennedy, W., Reinharz, D., Tessier, G., Contandriopoulos, A.P., Trabut, I., Champagne, F. & Ayoub, J. (1995). Cost utility of chemotherapy and best supportive care in non-small cell lung cancer, *Pharmacoeconomics* **8**, 316–323.

[38] Kiltie, A.E. & Gattamaneni, H.R. (1995). Survival and quality of life of paediatric intracranial germ cell tumour patients treated at the Christie Hospital, 1972–1993, *Medical and Pediatric Oncology* **25**, 450–456.

[39] Kind, P. (1996). The EuroQol instrument: an index of health-related quality of life, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 191–201.

[40] Kuppermann, M., Shiboski, S., Feeny, D., Elkin, E.P. & Washington, A.E. (1997). Can preference scores for discrete states be used to derive preference scores for an entire path of events? An application to prenatal diagnosis, *Medical Decision Making* **17**, 42–55.

[41] Laupacis, A., Wong, C., Churchill, D. & The Canadian Erythropoietin Study Group (1991). The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin, *Controlled Clinical Trials* **12**, 168S–179S.

[42] Laupacis, A., Bourne, R., Rorabeck, C., Feeny, D., Wong, C., Tugwell, P., Leslie, K. & Bullas, R. (1993). The effect of elective total hip replacement on health-related quality of life, *The Journal of Bone and Joint Surgery* **75-A**, 1619–1626.

[43] McEwen, J. & McKenna, S.P. (1996). Nottingham health profile, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 281–286.

[44] Mehrez, A. & Gafni, A. (1992). Preference based outcome measures for economic evaluation of drug

interventions: quality adjusted life years (QALYs) versus healthy years equivalents (HYEs), *Pharmacoeconomics* **1**, 338–345.

[45] Miller, J.D., Malthaner, R.A. & Goldsmith, C. (1996). *The Canadian Lung Volume Reduction Surgery Project*, Medical Research Council of Canada, Operating Grant Application, March 1997. McMaster University, Department of Surgery and Father Sean O'Sullivan Research Centre, Hamilton.

[46] Murray, C.J.L. & Lopez, A.D. (1997). Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: global burden of disease study, *Lancet* **349**, 1347–1352.

[47] Nease, R.F. (1996). Do violations of the axioms of expected utility theory threaten decision analysis? *Medical Decision Making* **16**, 399–403.

[48] O'Brien, B.J., Torrance, G.W. & Moran, L.A. (1994). *A Practical Guide to Health State Preference Measurement: A Video Introduction, Working Paper No. 95–2*. McMaster University, Centre for Health Economics and Policy Analysis, Hamilton.

[49] Oldridge, N., Furlong, W., Feeny, D., Torrance, G., Guyatt, G., Crowe, J. & Jones, N. (1993). Economic evaluation of cardiac rehabilitation soon after acute myocardial infarction, *American Journal of Cardiology* **72**, 154–161.

[50] Oldridge, N., Guyatt, G., Jones, N. et al.. (1991). Effects on quality of life with comprehensive rehabilitation after acute myocardial infarction, *American Journal of Cardiology* **67**, 1084–1089.

[51] Patrick, D., Starks, H., Cain, K., Uhlmann, R. & Pearlman, R. (1994). Measuring preferences for health states worse than death, *Medical Decision Making* **14**, 9–18.

[52] Pellissier, J.M. & Hazen, G.B. (1994). Implementation of continuous risk utility assessment: the total hip replacement decision, *Socio-Economic Planning Sciences* **28**, 251–276.

[53] Rabin, R. & de Charro, F. (2001). EQ-5D: a measure of health status from the EuroQol group, *Annals of Medicine* **33**(5), 337–343.

[54] Read, J., Quinn, R., Berwick, D., Fineberg, H. & Weinstein, M. (1984). Preferences for health outcomes–Comparisons of assessment methods, *Medical Decision Making* **4**, 315–329.

[55] Revicki, D., Brown, R.E., Palmer, W., Bakish, D., Rosser, W., Anton, S. & Feeny, D. (1995). Modelling the cost effectiveness of antidepressant treatment in primary care, *Pharmacoeconomics* **8**, 524–540.

[56] Rittenhouse, B.E. (1996). Designing and conducting cost-minimization and cost-effectiveness analyses, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 1093–1103.

[57] Roberge, R., Berthelot, J.M. & Wolfson, M. (1995). The health utility index: measuring health differences in Ontario by socioeconomic status, *Health Reports* **7**, 25–32.

[58] Russell, D., Beecroft, M., Ludwin, D. & Churchill, D. (1992). The quality of life in renal transplantation–a prospective study, *Transplantation* **54**, 656–660.

[59] Saigal, S., Feeny, D., Furlong, D., Rosenbaum, P., Burrows, E. & Torrance, G. (1994). Comparison of the health-related quality of life of extremely low birthweight children and a reference group of children at age eight years, *Journal of Pediatrics* **125**, 418–425.

[60] Saigal, S., Rosenbaum, P., Stoskopf, B., Hoult, L., Furlong, W., Feeny, D., Burrows, E. & Torrance, G. (1994). Comprehensive assessment of the health status of extremely low birthweight children at eight years of age: comparison with a reference group, *Journal of Pediatrics* **125**, 411–417.

[61] Thompson, M.S., Read, J.L., Hutchings, H.C. (1988). The cost effectiveness of auranofin: results of a randomized clinical trial, *Journal of Rheumatology* **15**, 35–42.

[62] Torrance, G.W. (1986). Measurement of health-state utilities for economic appraisal: a review, *Journal of Health Economics* **5**, 1–30.

[63] Torrance, G.W. (1996). Designing and conducting cost-utility analyses, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 1105–1111.

[64] Torrance, G. & Feeny, D. (1989). Utilities and quality-adjusted life years, *International Journal of Technology Assessment in Health Care* **5**, 559–575.

[65] Torrance, G.W., Boyle, M.H. & Horwood, S.P. (1982). Application of multi-attribute utility theory to measure social preferences for health states, *Operations Research* **30**, 1043–1069.

[66] Torrance, G., Furlong, W., Feeny, D. & Boyle, M. (1995). Multi-attribute preference functions: health utilities index, *Pharmacoeconomics* **7**, 503–520.

[67] Torrance, G.W., Furlong, W. & Feeny, D. (2002). Health utility estimation, *Expert Review of Pharmacoeconomics and Outcomes Research* **2**(2), 99–108.

[68] Torrance, G.W., Paterson, M. & Harris, Jr., E.D. (1976). Social preferences for health states: An empirical evaluation of three measurement techniques, *Socio-Economic Planning Sciences* **10**, 129–136.

[69] Torrance, G.W., Feeny, D.H., Furlong, W.J., Barr, R.D., Zhang, Y. & Wang, Q. (1996). Multiattribute utility function for a comprehensive health status classification system: health utilities index mark 2, *Medical Care* **34**, 702–722.

[70] Tully, P. & Mohl, C. (1995). Older residents of health care institutions, *Health Reports* **7**, 27–30.

[71] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton University Press, Princeton.

[72] Ware, J.E. (1996). The SF-36 health survey, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed., B. Spilker, ed. Lippincott-Raven, Philadelphia, pp. 337–345.

[73] Wolfson, A.D., Sinclair, A.J., Bombardier, C. & McGeer, A. (1982). Preference measurements for

functional status in stroke patients: inter-rater and inter-technique comparisons, in *Values and Long Term Care*, R. Kane & R. Kane, eds. D.C. Heath, Lexington, pp. 191–214.

[74] Wolfson, M.C. (1996). Health-adjusted life expectancy, *Health Reports* **8**, 41–46.

(*See also* **Decision Analysis in Diagnosis and Treatment Choice**; **Risk Assessment in Clinical Decision Making**)

GEORGE W. TORRANCE

Utility

Utility theory is concerned with the quantitative representation of individual preferences for the outcomes of a decision. In biomedical research, utility theory is used to provide guidance for individual and sometimes societal decisions regarding health, and to provide a rational foundation for the choice of appropriate **experimental designs** and data analysis strategies (*see* **Decision Theory**).

The ideas underlying modern utility theory arose during the Age of Enlightenment with the work of Daniel Bernoulli [3, 5] (*see* **Bernoulli Family**). Bernoulli analyzed the behavior of rational individuals in the face of risk from a Newtonian perspective, viewing science as an operational model of the human mind. His empirical observation that prudent and thoughtful individuals do not necessarily take the actions that maximize their expected monetary return, as in the St Petersburg's paradox [3, 11], led him to investigate a formal model of individual choices based on the direct quantification of value, and to develop a prototypical utility function for wealth.

Quantification of value has been a central component of economic thought since then, and has more recently gained an important role in the theories of decision making under uncertainty, both descriptive and normative. (Descriptive theories aim at portraying the way individuals or groups make decisions, while normative theories aim at guiding decision making based on adherence to accepted fundamental principles. While descriptive theories are important in public health areas such as risk communication (*see* **Risk Assessment**) and patient counseling, this article is concerned with normative theories.)

Most utility theories consider the problem of representing an individual's preferences for the elements of a set R of possible outcomes. Examples of outcomes relevant for biomedical applications are the health states following a treatment or intervention, the consequences of marketing a drug, the change in the exposure to a toxic agent that may result from a regulatory change, the knowledge gained from a study design, and so forth. If r_1 and r_2 are two outcomes in R, then $r_1 \prec r_2$ indicates that r_2 is preferred to r_1. Formally, $\prec$ is a binary relation on R, usually taken to be asymmetric. Indifference between two outcomes (neither $r_1 \prec r_2$ nor $r_2 \prec r_1$) is indicated by $r_1 \sim r_2$.

A cardinal utility function is a real-valued score u assigned to the outcomes in R, so that

$$r_1 \prec r_2 \iff u(r_1) < u(r_2). \tag{1}$$

Not all preferences are amenable to this kind of representation, but, for example, if R is a countable set, and if both $\prec$ and $\sim$ are transitive relationships, then a u that represents these preferences can be constructed. An in-depth discussion of conditions for the existence of such cardinal utility representations of preferences is in [6].

Expected Utility Theory

One contribution of utility theory to decision making is in the possibility of characterizing preferences over complex sets of options in terms of much simpler utility specifications. In the prototypical problem of decision under uncertainty, a decision maker must choose an action whose consequences are uncertain. Each action is then described by a given probability distribution p over the set R of possible outcomes. A simple operational approach is to assign a utility score $u(r)$ to each of the outcomes, and choose the action p that maximizes the expected value of the utility of the outcome, or

$$u(p) = \sum_{r \in R} p(r)u(r). \tag{2}$$

A fundamental contribution to the foundations of this approach is the work of von Neumann & Morgenstern [21], who showed how the expected utility representation (2) can be derived from conditions on the ordinal relationships among the set of all actions P. In particular, they provided necessary and sufficient conditions for preferences over a convex set P of options to be representable by a utility function of the form (2). These conditions can be thought of as basic rationality requirements, and are taken to be the primitives, or axioms, in the von Neumann & Morgenstern theory of utility. To gain a basic understanding of these axioms, assume that R is countable and that P is the set of all probability distributions on R. The decision maker has preferences over elements of P. The axioms, in the format given in [10], are:

1. *Weak order axiom.* Both $\prec$ and $\sim$ are transitive.
2. *Archimedean axiom.* If $p_1 \prec p_2 \prec p_3$, then there are α and β in (0, 1) such that $\alpha p_1 + (1 -$

$\alpha)p_3 \prec p_2 \prec \beta p_1 + (1-\beta)p_3$. Here $\alpha p_1 + (1-\alpha)p_3$ indicates the action that leads to outcome r with probability $\alpha p_1(r) + (1-\alpha)p_3(r)$. In words, p_3 can be preferred to p_2, but not so strongly that mixing p_3 with p_1 cannot lead to a reversal of preference. So p_3 cannot be incommensurably better than p_2. Likewise, p_1 cannot be incommensurably worse than p_2.

3. *Independence axiom.* If $p_1 \prec p_2$, then, for every p_3 in P and α in (0, 1), $\alpha p_1 + (1-\alpha)p_3 \prec \alpha p_1 + (1-\alpha)p_2$. In words, the two composite lotteries should be compared solely based on the component that is different.

Axioms 1, 2, and 3 hold if and only if there is a real-valued utility u such that the preferences for the options in P can be represented as in (2). A given set of preferences identifies a utility function u only up to a linear transformation with positive slope.

The von Neumann & Morgenstern theory also provides the basis for practically assessing an individual decision maker's utilities for outcomes. A widely used approach is the so-called **"standard gamble"**, illustrated here in the case of a finite set of outcomes R. If we avoid the trivial case in which all outcomes are equally valued by the decision maker, then the weak ordering assumption permits us to identify a worst outcome r_1 and a best outcome r_2. For example, in assessing the utility of health states, "death" is often chosen as the worse outcome and "full health" as the best, although in some problems there are health outcomes that could be ranked worse than death [20]. Worst and best outcomes need not be unique. Because all utility functions that are positive linear transformations of the same utility function lead to the same preferences over P, we can arbitrarily set $u(r_1) = 0$ and $u(r_2) = 1$, leading to a convenient and interpretable utility scale.

Then a decision maker's utility for outcome r can be inferred by eliciting the value of π such that the decision maker is indifferent between the following two actions:

p_1: outcome r for certain;

p_2: outcome r_1 with probability $1-\pi$ and outcome r_2 with probability π.

The existence of a value of π reaching indifference is implied by the Archimedean and independence properties of the decision maker's preferences. It is easy to check that the expected utility of both actions p_1 and p_2 is π, and that therefore $u(r) \simeq \pi$. Alternative assessment methods used in health sciences are reviewed by [20].

Expected utility maximization proved to be a fundamental tool in guiding practical decision making under uncertainty, including clinical decision making (*see* **Decision Analysis in Diagnosis and Treatment Choice**) and cost-effectiveness analysis (*see* **Health Economics**). The literature on the extensions of this characterization is extensive. Good entry points are [13, 8], and [9]. Because of the centrality of the expected utility paradigm, the von Neumann–Morgenstern axiomatization and its derivatives have been deeply scrutinized and criticized from both descriptive and normative perspectives. Empirically, it is well documented that individuals sometimes violate the independence axiom [1, 12, 13, 19]. Normative questions have also been raised about the weak ordering assumption [17, 18].

Subjective Expected Utility Theory

A more general decision setting occurs when outcomes depend on uncertain events, whose probabilities are not given externally, as in the theory of von Neumann–Morgenstern, but must be assessed by the decision maker together with the utilities. An action can then be described as a function $a(s)$ from states $s \in S$ to outcomes $r \in R$: the states describe the alternative realizations of the uncertain events that affect the outcome of the decision. For example, consider deciding whether to be vaccinated against influenza in anticipation of the cold season. The two actions are "vaccine" and "no vaccine". A simple description of the problem could include four states, defined by the combinations of the events "adverse reaction to the vaccine" and "influenza is contracted later in the winter". To each state S there corresponds an outcome $r(s)$ that could be a description of health states, costs, and so forth, ensuing from state s.

Extending the results of von Neumann & Morgenstern, and echoing earlier groundbreaking work by Ramsey [14], **L.J. Savage** [15] developed a system of axioms for preferences over the type of actions exemplified above. These axioms hold if, and only if, there is a utility function u over outcomes and a probability distribution p over states that represent

the agent preferences by

$$u(a) \simeq \sum_{s \in S} p(s)u[r(s)]. \tag{3}$$

Savage's theory is based on seven axioms [15]. As an alternative to the rewarding but demanding reading of Savage's fundamental book, one can consult [13] or [7]. Some of the axioms are similar in spirit to those of von Neumann & Morgenstern. An important additional requirement is the so-called state independence of utilities, which in essence requires the decision maker to give the same value to identical outcomes ensuing from different states. A critical appraisal of the consequences of this assumption is in [16]. Later work on subjective expected utility theory is reviewed in [7].

In addition to having direct implications for decision making under uncertainty, Savage's theory aims at providing a rational foundation for statistical **inference**. The actions represent the results of a statistical analysis, such as rejecting a hypothesis; the states are the alternative values of the parameters of interest such as the **null** and the **alternative hypotheses**; the outcomes represent the rewards, or losses, resulting from the analysis. If an agent's preferences satisfy the axioms, then the agent's choices will be consistent with assigning a **subjective probability** distribution to the states and a utility function to the outcomes, and choosing the action that maximizes the resulting expected utility. In particular, all Bayesian optimal procedures (*see* **Bayesian Methods**) are consistent with Savage's system of rationality axioms.

References

[1] Allais, M. (1953). Le comportement de I'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine, *Econometrica* **21**, 503–546.

[2] Berger, J.O. (1985). *Bayesian Analysis and Statistical Decision Theory*. Springer, Berlin.

[3] Bernoulli, D. (1738). Specimen Theoriae Novae de mensura sortis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, Vol. **V**, 175–192; English translation as "Exposition of a new theory on the measurement of risk", *Econometrica* **22**(1954), 23–35.

[4] DeGroot, M.H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

[5] Eatwell, J., Milgate, M. & Newman, P., eds (1987). *Utility and Probability*. Norton, New York.

[6] Fishburn, P.C. (1970). *Utility Theory for Decision Making*. Wiley, New York.

[7] Fishburn, P.C. (1981). Subjective expected utility: a review of normative theories, *Theory and Decision* **13**, 139–199.

[8] Fishburn, P.C. (1982). *The Foundations of Expected Utility*. Reidel, Dordrecht.

[9] Gärdenfors, P. & Sahlin, N.-E. (1988). *Decision, Probability and Utility*. Cambridge University Press, Cambridge.

[10] Jensen, N.E. (1967). An introduction to Bernoullian utility theory: I. Utility functions, *Swedish Journal of Economics* **69**, 163–183.

[11] Jorland, G. (1987). The Saint Petersburg Paradox 1713–1937, in *The Probabilistic Revolution*, Vol. 1, L. Kruger, L.J. Daston & M. Heidelberger, eds. MIT Press, Cambridge, Mass, pp. 157–190.

[12] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York.

[13] Kreps, D.M. (1988). *Notes on the Theory of Choice*. Westview, Boulder.

[14] Ramsey, F. (1926). Truth and probability, in *The Foundations of Mathematics*. Routledge & Kegan Paul, London, 1931, pp. 156–211.

[15] Savage, L. (1954). *The Foundations of Statistics*. Wiley, New York.

[16] Schervish, M.J., Seidenfeld, T. & Kadane, J.B. (1990). State dependent utilities, *Journal of the American Statistical Association* **85**, 840–847.

[17] Seidenfeld, T. (1988). Decision theory without "independence" or without "ordering". What is the difference? (with discussion), *Economics and Philosophy* **4**, 267–290.

[18] Seidenfeld, T., Schervish, M.J. & Kadane, J.B. (1995). A representation of partially ordered preferences, *Annals of Statistics* **23**, 2168–2217.

[19] Shoemaker, P.J.K. (1982). The expected utility model: its variants, purposes, evidence and limitations, *Journal of Economic Literature* **20**, 529–563.

[20] Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal, *Journal of Health Economics* **5**, 1–30.

[21] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Wiley, New York.

GIOVANNI PARMIGIANI

Validation Study

Validation studies obtain information on **measurement errors** in exposures and other **covariates** used in epidemiologic studies by comparing the conventional exposure measurements with "**gold standard**" measurements. **Reliability studies**, unlike validation studies, provide information on the measurement error process by replicating the conventional exposure measurements. Validation studies can be applied to study a broader range of error processes than reliability studies, which are based on a special error model. Data from validation studies or reliability studies are needed to correct relative **risk** estimates for **bias** and to obtain valid **inference** in the presence of measurement error (*see* **Misclassification Error**).

To define these ideas more precisely, let $\mathbf{Y}$ be the response variable, and let $\mathbf{X}$ be the true value(s) of the covariate. In some cases, $\mathbf{X}$ can never be observed and can be thought of as a *latent* variable. In other cases, $\mathbf{X}$ is a "gold standard" method of covariate assessment which is infeasible and/or expensive to administer to large numbers of study participants. Usually, instead of observing $\mathbf{X}$, an error-prone measurement $\mathbf{W}$ is observed. Finally, there may be covariates $\mathbf{Z_1}$ upon which the model for response depends that are never misclassified or measured with error. In main study/validation study designs, the main study yields the data $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i), i = 1, \ldots, n_1$. If the validation study is *internal*, it yields the observations $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i), i = n_1 + 1, \ldots, n_1 + n_2$. If the validation study is *external*, it produces observations $(\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_{2i}), i = n_1 + 1, \ldots, n_1 + n_2$ observations. There may be covariates, denoted $\mathbf{Z}_2$, upon which the measurement error and/or misclassification model depend but of which the model for response is independent; we denote the unique elements of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ by $\mathbf{Z}$. An external validation study is a useful option only when there are a priori reasons to believe that measurement error/misclassification is **nondifferential**, i.e. that $f(\mathbf{Y}|\mathbf{W}, \mathbf{X}, \mathbf{Z}) = f(\mathbf{Y}|\mathbf{X}, \mathbf{Z})$. This definition can be rewritten as $f(\mathbf{W}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = f(\mathbf{W}|\mathbf{X}, \mathbf{Z})$, in which form the nondifferential error feature is more apparent.

Without validation or reliability data, it is possible only to perform **sensitivity analyses** under hypothesized scenarios for measurement error and/or misclassification. Without information about the nature and extent of the measurement error and/or misclassification, sensitivity analyses yield wide ranges of the parameter(s) estimates, and cannot assess the true uncertainty of the estimates. It is possible in certain instances to test hypotheses about $\mathbf{X}$, however, even in the absence of validation data. For **generalized linear models** with $\dim(\mathbf{X}) = \dim(\mathbf{W}) = 1$, the usual score test, based upon the main study data alone, will have the correct size, although its **power** will be reduced unless $\mathbf{X}$ is linearly related to $\mathbf{W}$ and $\mathbf{Z}$ [16]. The same results apply for the global **null hypothesis** about $\mathbf{X}$. Validation and/or reliability data are required for valid **estimation** and inference in nearly all other circumstances. An exception occurs when the model for $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z}_1)$ is **logistic** and the model for $\mathbf{X}$ given $(\mathbf{W}, \mathbf{Z}_2)$ is Gaussian (*see* **Normal Distribution**). In this case, when $\dim(\mathbf{X}) = \dim(\mathbf{W}) = 1$, the parameters of both models are **identifiable** from the main study alone [10], although as yet unpublished work by Spiegelman & Rosner indicates that estimates of these parameters are difficult to obtain, and when obtained, are usually very imprecise.

A reliability study can be used to estimate **variance components** in the classic, random within-person, measurement error model

$$\mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{1}$$

where $\mathbf{U}$ is a mean zero error term with variance–covariance Σ. It is only when (1) applies that replicate data, as would be obtained in a reliability study, can be used for valid estimation and inference in the presence of covariate measurement error. Eq. (1) has been applied to the assessment of measurements of blood pressure, serum hormones, and other serum biomarkers such as vitamin concentrations.

A validation study can be used for a wide range of error models. A validation study may be expensive or infeasible because it requires that the true value $\mathbf{X}$ be observable, at least in some small sample of n_2 study participants. That is, a "gold standard" technique for measuring the quantity of interest without error must be available. In most situations, it is impossible to measure exposure perfectly. If measurement error/misclassification methods are used with an "alloyed" or imperfect gold standard, $\mathbf{X}'$, the results of the analysis can be misleading. If $\mathbf{X}'$ and $\mathbf{Z}$ are uncorrelated, however, the results may be interpreted as those which would have been obtained had

the (imperfect) gold standard measurements $\mathbf{X}'$ been available for all study participants, rather than just those in the validation study. Under certain circumstances, if the errors in $\mathbf{X}'$ and $\mathbf{W}$ are uncorrelated, the regression calibration estimate will provide **unbiased** measurement error correction for the gold standard, $\mathbf{X}$ [14, 18]. In some realistic examples, the bias in regression calibration estimates is small even when the errors are moderately correlated.

In this article it is assumed, unless stated otherwise, that the validation study is sampled completely at random. That is, if V is a **random variable** which equals 1 if a participant is in the validation study and 0 otherwise, we assume $\Pr(V = 1|\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{Z}) = \pi$, independent of $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Z}$.

Two-stage designs (*see* **Case–Control Study, Two-phase**) allow V to depend upon $\mathbf{Y}$, $\mathbf{W}$ and $\mathbf{Z}$. Two-stage designs allow one to control the selection of validation study participants so as to increase statistical efficiency or reduce cost. These options require the validation study to be *internal*. Two-stage designs have several limitations. Many validation studies yield data on numerous covariates. For example, in a prospective **cohort study** yielding information on cancer and cardiovascular endpoints, it may be necessary to validate many nutrient measures simultaneously. An optimal design for one response/covariate pair may be inefficient for another. Although some authors have found that the optimal sampling probability function, π depends on $\mathbf{Y}$ ([17] and as yet unpublished work by Holcroft & Spiegelman) in cohort studies and many **nested case–control** studies, it will not be possible to identify the optimal π as a function of $\mathbf{Y}$, since covariate status is best ascertained before observing $\mathbf{Y}$.

The remainder of this article will discuss strategies for optimizing (n_1, n_2), assuming that the validation sample is taken completely at random. Solutions to this problem are mathematically complex, and software is not widely available. In a two-stage design, $(n_1 + n_2, \pi)$ must be optimized, which poses an even more difficult theoretical and computational problem. Nevertheless, further research on two-stage designs may lead to improvements on the designs presented below.

The valid use of an external validation study requires that the measurement error model $f(\mathbf{X}|\mathbf{W}, \mathbf{Z}_2)$ is the same in the external population as in the main study. This assumption is necessarily true for an internal validation study, supplied completely at random. By **Bayes' Theorem** $f(\mathbf{X}|\mathbf{W}, \mathbf{Z}_2) = f(\mathbf{W}|\mathbf{X}, \mathbf{Z}_2) f(\mathbf{X}, \mathbf{Z}_2)/f(\mathbf{W}, \mathbf{Z}_2)$. Although in many instances it may be reasonable to assume that $f(\mathbf{W}|\mathbf{X}, \mathbf{Z}_2)$ may be "transportable" from one population to the next, provided the instruments used to measure $\mathbf{X}$ and $\mathbf{W}$ are identical, it is less reasonable to assume that the unobservable marginal density $f(\mathbf{X}, \mathbf{Z}_2)$ in the main study population is the same as that in an external validation population. Thus, an internal validation study is more convincing than an external validation study.

Two recent epidemiologic textbooks devote a chapter to the design and analysis of validation and reliability studies. Armstrong et al. [1, Chapter 4] focus primarily on the design and analysis of reliability studies, but there is some consideration of validation studies as well. Willett [19] discusses validation study design for dietary intake questionnaires (*see* **Nutritional Exposure Measures**). Willett gives a simple formula for calculating the sample size of a validation study, based on the criterion of testing $\mathrm{H}_0 : \rho = \rho_0$ vs. $\mathrm{H}_\mathrm{a} : \rho = \rho_A$ with prespecified power $1 - \beta$ and nominal size α, where ρ is the **correlation** between the usual exposure method ($\mathbf{W}$) and the gold standard ($\mathbf{X}$):

$$n = 3 + \frac{(Z_\alpha + Z_\beta)^2}{|\rho_A - \rho_0|},$$

where Z_α and Z_β are the **standard normal deviates** for α and β, respectively (*see* **Sample Size Determination**). This criterion for validation study sample size is not designed to ensure adequate sample size for measurement error correction, and is strictly useful only when the estimation of the correlation between $\mathbf{X}$ and $\mathbf{W}$ is the end goal. Willett reports some otherwise unpublished data which examined the influence of validation study sample size on the precision of **odds ratio** corrected for measurement error by regression calibration, where it had been found that for "realistic conditions" $(0.5 < \rho < 0.7)$, validation studies with more than 150–200 subjects provide little additional precision.

Design of Main Study/Validation Studies

Choice of the Optimization Criterion for Efficient Study Designs

In the biomedical research setting, research proposals will usually not be approved for funding unless the

proposed study has power of 80% or more to test the central scientific hypothesis. The study design process seeks to minimize the proposed budget while assuring adequate statistical power. Because the unit cost of measuring **X** can be 100 or more times that of measuring **W**, validation studies can be expensive, and efficient main study/validation study designs are essential. Incorporating these design features is important, because it is seldom possible to collect validation data after the main study has been completed. Calculation of a point estimate, typically an odds ratio or **hazard ratio**, and the construction of **confidence intervals** around this estimate are often primary analytic goals in observational biomedical research, where measurement error and misclassification frequently arise. Greenland proposed the *discriminatory power* criterion [5] for these settings. One specifies the two sample sizes needed, respectively, to test the **null hypothesis**, $H_0 : \beta = \beta_L$ against the **alternative**, $H_a : \beta = \beta_U$, each with prescribed size and power. According to the discriminatory power criterion, the required sample size is the maximum of these two sample sizes. In **relative risk models**, because the **variance** of the estimate of the parameter of interest, $\hat{\beta}$, depends on the value of the parameter of interest, β, typically the log odds ratio or log hazard ratio, the discriminatory power criterion will usually produce optimal sample sizes larger than those produced by the traditional power criterion. An alternative criterion is to specify the expected confidence interval width at a specified relative risk, but this criterion may produce designs with unknown, possibly subnominal confidence levels for different, equally plausible values of the relative risk.

Given unit costs $r_{\mathbf{Y}}, r_{\mathrm{W}}$, and r_{X} for a single measurement of **Y**, **W**, and **X**, respectively, the optimization criterion we prefer minimizes the total study cost, C, with respect to (n_1, n_2), subject to minimum discriminatory power requirements. In a main study/external validation study design,

$$C(n_1, n_2) = (r_{\mathbf{D}} + r_{\mathbf{W}})n_1 + (r_{\mathbf{W}} + r_{\mathbf{X}})n_2$$

and maximization over (n_1, n_2) is subject to the constraints

$$1 - \Phi\left[\frac{Z_{1-\alpha/2}[V_L(n_1, n_2)]^{1/2} - \beta_U + \beta_L}{[V_U(n_1, n_2)]^{1/2}}\right] \geq \Pi,$$

and

$$\Phi\left[\frac{-Z_{1-\alpha/2}[V_U(n_1, n_2)]^{1/2} - \beta_L + \beta_U}{[V_L(n_1, n_2)]^{1/2}}\right] \geq \Pi.$$

Here $V_L(n_1, n_2)$ and $V_U(n_1, n_2)$ are the expected values of the variances of $\hat{\beta}$ evaluated at β_L and β_U, respectively, over the distribution of (**D**, **X**, **W**) for the study population to be investigated, Π is the minimal acceptable discriminatory power, and z_γ is γth quantile of the standard normal distribution. For a main study/internal validation study design, the cost function

$$C(n_1, n_2) = \min[(r_{\mathbf{D}} + r_{\mathbf{W}})n_1 + (r_{\mathbf{W}} + r_{\mathbf{X}} + r_{\mathbf{D}})n_2, (r_{\mathbf{D}} + r_{\mathbf{X}})n_2^*]$$

is used, since there may be a discontinuity point in the main study/internal validation study design optimization equations at which the fully validated design, consisting only of observations $(\mathbf{Y}_i, \mathbf{W}, \mathbf{Z}_{1i})$, $i = 1, \ldots, n_2$, is optimal.

Within this framework, one must supply the necessary design specifications, which will vary from one setting to another, and substitute the appropriate formulas for V_L and V_U to obtain the optimal values of n_1 and n_2. Even with a simple formula for $\mathrm{var}\hat{\beta}$, the solution cannot be written in closed form because the constraints are complex. Numerical solutions can be found using a nonlinear multiparameter optimization subroutine such as DNCONF in IMSL [8], with a call to an application-specific subroutine implementing the appropriate variance formula.

Although the design criteria discussed above are all functions of the variance of $\hat{\beta}$ as well as other parameters, the expected confidence interval width criterion is simply proportional to the square root of this variance. A limitation of most of the papers on validation study design is that they investigate design issues only by the criterion of expected confidence interval width.

Optimal Study Design for Misclassified Binary Exposure Variables

When **binary** exposure variables are subject to misclassification, the simplest misclassification model, $f_2(\mathbf{X}|\mathbf{W}; \theta)$, can be completely described with two parameters, $\theta = (\omega, \varphi)$, namely the **sensitivity**

$$\omega = \Pr(\mathbf{W} = 1|\mathbf{X} = 1),$$

and the **specificity**

$$\varphi = \Pr(\mathbf{W} = 0|\mathbf{X} = 0).$$

This model is less restrictive than the measurement error model (1), since the error distribution depends on **X**, except in the special case when $\omega = 1 - \varphi$. Further complexity in the misclassification model can be introduced by allowing that ω and/or φ vary with **Y**, as in the case of **recall bias** in a **case–control study**, or allowing ω and/or φ to vary with some other covariate(s), **Z**. In most instances, optimal study designs will depend on the underlying true exposure **prevalence**, $\Pr(\mathbf{X} = 1)$.

Palmgren [11] studied internal validation designs for case–control studies and with the same proportions of the sample allocated to validation for cases and controls. Palmgren found the optimal allocation proportion, $n_2/(n_1 + n_2)$, to minimize the null variance of the estimated log odds ratio, $\hat{\beta}$, subject to fixed cost. She also determined the optimal allocation proportion to minimize the variance of $\hat{\beta}$ subject to fixed cost. She did not base her designs on a classical power criterion nor on the discriminatory power criterion.

To use Palmgren's formulation, the investigator must specify the odds ratio, the sensitivity and specificity for measuring exposure, the exposure prevalences in cases and controls, and the costs of measuring **W** and **X**. When $\beta = 0$ and the sensitivity and specificity are assumed to be the same for cases and controls, the optimal design for minimizing the variance of $\hat{\beta}$ is the fully validated design, $(n_1 = 0)$, unless the square of the correlation between **W** and **X** is greater than the cost ratio $r_{\mathbf{W}}/r_{\mathbf{X}}$, in which case the optimal design is the main study only $(n_2 = 0)$. When sensitivity and specificity are allowed to depend upon case status, the optimal allocation ranges between 0 and 1, depending on the other design parameters. Palmgren also provided some results for minimizing the variance of the **maximum likelihood** estimator of β, for $\beta \neq 0$, subject to fixed cost, under the assumptions that the sensitivity and specificity are equivalent for cases and controls and for a one-to-one case–control ratio. It is shown that for small β, case–control ratios near one-to-one and in most cases when the cost ratio, $r_{\mathrm{X}}/r_{\mathrm{W}}, \geq 4$, the main study/internal validation study design is more efficient than the fully validated design. When the exposure is rare, the optimal design depends more heavily on the value of the specificity parameter, and when the exposure is common, the optimal design depends more heavily on the sensitivity parameter.

A useful contribution of [11] are equations (A1) and (A2), which give the nonnull and null formulas for the variance of the maximum likelihood estimator of β.

Greenland [6] also minimized the variance of $\hat{\beta}$ subject to fixed cost, but he used the matrix method for estimating β [2, 7, 12], rather than maximum likelihood. Greenland found that the optimal proportion allocated to the validation study increases dramatically when **differential** misclassification is assumed. He concluded that the fully validated design is optimal or near optimal in many cases and has the additional advantages of permitting standard methods of data analysis and of assuring representativeness of the validation sample. Although Greenland's recommendations are clearly appropriate for low cost ratios ($r_{\mathbf{X}}/r_{\mathbf{W}}$ ranging between 3 and 12) and high case–control ratios (e.g. where $\Pr(\mathbf{Y}) \approx 0.50$), it is not clear at what point this recommendation no longer applies. For cost ratios, $r_{\mathrm{X}}/r_{\mathrm{W}}$, above 100, or lower case–control ratios, the main study/validation study design is likely to be preferable.

Chernoff & Haitovsky [4] and Zelen & Haitovsky [20] considered optimal design in the estimation and testing settings, respectively. They admitted as the class of optimal designs a linear combination of the eight designs derived from four possibilities for case and controls: (i) main study only; (ii) validate all subjects with $\mathbf{W} = 1$; (iii) validate all subjects with $\mathbf{W} = 0$; and (iv) validation study only. Designs were optimized by minimizing the variance of the estimate of $\Pr(\mathbf{X} = 1|\mathbf{Y} = 1) - \Pr(\mathbf{X} = 1|\mathbf{Y} = 0)$, for fixed cost, and differential misclassification was assumed at known misclassification rates. They found that the optimal design was a combination of two of the eight sampling plans, one for cases and the other for controls. As long as the sensitivity, specificity, and costs of collecting the data are the same for cases and controls, the same type and sample size of design for cases and controls is optimal.

Optimal Study Designs for a Continuous Covariate Measured with Error

Buonaccorsi [3] provided optimal allocation formulas for minimizing the variance of the estimate of the

odds ratio subject to specified cost, when $\dim(\mathbf{X}) = \dim(\mathbf{W}) = 1$, $\dim(\mathbf{Z})$ is arbitrary, and $\mathbf{Y}$ is a binary outcome. These formulas apply when ($\mathbf{X}$, $\mathbf{W}$, $\mathbf{Z}$) are jointly **multivariate normal** given $\mathbf{Y}$ but where the measurement error model, $f_2(\mathbf{X}|\mathbf{W})$, may be a function of $\mathbf{Y}$ or dependent on other covariates $\mathbf{Z}_2$. Under these assumptions, the odds ratio can be validly and efficiently estimated through the normal **discriminant analysis** model instead of the **logistic regression** model. Buonaccorsi derived a closed-form expression for the optimal proportion of study subjects validated under the main study/internal validation study design, $n_2/(n_1 + n_2)$, as a function of six quantities: (i) unit costs for $\mathbf{Y}$, ($\mathbf{W}$, $\mathbf{Z}$) and $\mathbf{X}$ ($r_{\mathbf{Y}}$, $r_{\mathbf{W}}$, and $r_{\mathbf{X}}$); (ii) total cost for the study (C); (iii) value for the log odds ratio of $\mathbf{Y}$ from a unit change in $\mathbf{X}$ (β); (iv) multiple correlation between $\mathbf{X}$ and ($\mathbf{W}$ $\mathbf{Z}$) (this quantity can be taken to represent the extent of measurement error resulting from failure to observe $\mathbf{X}$); (v) the variance of $\mathbf{X}$ (σ_x^2); and (vi) the **marginal probability** of $\mathbf{Y}$ [Pr($\mathbf{Y}$)]. With the exception of the measurement error parameter and the quantities relating to cost, all of these quantities would be required for study design calculations even when $\mathbf{X}$ were perfectly measured. Buonaccorsi gave a simple formula for the variance of the measurement-error corrected estimate of β.

Spiegelman & Gray [13] also investigated optimal study design for binary **regression** in the case of a single continuous covariate measured with error but, unlike Buonaccorsi, they relied on the discriminatory power criterion. In addition to considering the main study/internal validation study design paradigm, they also considered the main study/external validation study designs. Since external validation study data will often be obtained as an afterthought at the end of a prospective study, choosing the optimal sample sizes, n_1 and n_2, is of less practical importance. However, it is instructive to compare cost, power, and sampling ratios as given by optimal internal and external validation study settings. From efficiency considerations, it was shown that an internal validation study is optimal. However, as Pr($\mathbf{Y}$) becomes small, the efficiency advantage of the main study/internal validation study design virtually disappears relative to the main study/external validation study design.

Rather than assuming that ($\mathbf{W}$, $\mathbf{X}$, $\mathbf{Z}$) are jointly normal given $\mathbf{Y}$, Spiegelman & Gray assume that $\mathbf{X}|\mathbf{W}$ is multivariate normal. Unlike Buonaccorsi, they did not consider the presence of additional perfectly measured exposure variables $\mathbf{Z}$, and they require that $\mathrm{E}(\mathbf{X}|\mathbf{W})$ is linear. Without the joint multivariate normality assumption, the normal discriminant model cannot be used to obtain an **unbiased** estimate of β. Instead, the logistic regression model must be used. This is the model upon which Spiegelman & Gray based their sample size calculations. Iterative methods must be used to find optimal main study and validation study sample sizes when $f_2(\mathbf{X}|\mathbf{W})$ is normal and $f_1(\mathbf{Y}|\mathbf{X})$, the model for the outcome conditional on the true exposure, is logistic. In order to find the optimal design in this framework, the investigator needs to identify six quantities: (i) $r_{\mathbf{Y}}, r_{\mathbf{X}}$, and $r_{\mathbf{W}}$; (ii) Pr($\mathbf{Y}$); (iii) β_{L} and β_{U}, the two values of the log odds ratio between which the study is designed to discriminate; (iv) the mean and variance of $\mathbf{W}$; (v) the parameters for the conditional mean of $\mathbf{X}$ given $\mathbf{W}$, α', and γ, where $\mathrm{E}(\mathbf{X}|\mathbf{W}) = \alpha' + \gamma\mathbf{W}$, and $\mathrm{var}(\mathbf{X}|\mathbf{W})$; and (vi) the desired confidence level, α, and the required discriminatory power, Π.

Spiegelman & Gray found that the fully validated design is optimal only when $r_{\mathbf{X}}/r_{\mathbf{W}}$ is small, Pr($\mathbf{Y} = 1$) is large, and the magnitudes of β_{U} and $\beta_{\mathbf{L}}$ are relatively far from the null but close together. They found that the optimal percent allocation to the validation study increased as the unit cost of $\mathbf{Y}$ increased. Figure 1 shows the cost of the optimal main study/internal validation study designs for sample disease frequencies [Pr($\mathbf{Y}$)] equal to 0.005. In

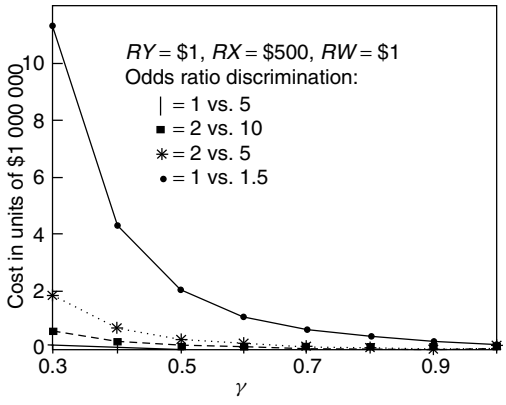

Figure 1 Plot of minimized cost to discriminate between two hypothesized odds ratios, against γ, where $\mathrm{E}(\mathbf{X}|\mathbf{W}) = \alpha' + \gamma\mathbf{W}$. When $\mathbf{X}$ and $\mathbf{W}$ are standardized, γ is corr ($\mathbf{W}$, $\mathbf{X}$)

this figure RY, RX, and RW are the unit costs for measuring **Y**, **X**, and **W**, respectively, and, when **X** and **W** are standardized, γ corresponds to the correlation between **X** and **W**. Designs are optimized to discriminate with 95% power between two hypothesized values of the odds ratios, and the scenarios considered are given in the legend. Cost increases dramatically as the distance between the two hypothesized odds ratios decreases, and as measurement error increases.

In the context of **nutritional epidemiology**, two additional papers on validation study design have appeared. Stram et al. [15] considered validation study design for minimizing the variance of the regression calibration estimate of the odds ratio, under the constraint of fixed total cost for an external validation study. They derived equations for the optimal choice of the number of subjects and the number of days per subject of diet records or diet recalls ($\mathbf{X}'$) when the food frequency questionnaire (**W**) is used to assess diet in the main study. It is assumed that the relationship between **X** and $\mathbf{X}'$ is given by the assumption of random within-person variation, following (1). They found that the optimal validation study size and number of replicates per subject depend on the ratio between the costs of the initial and subsequent 1-day diet records, and on the ratio of the variance in a single replicate of $\mathbf{X}'$ to the variance of the true underlying diet. The authors concluded that, in most settings, the optimal study design will rarely require more than five 1-day diet records per validation study participant. Kaaks et al. [9] derived a closed-form expression for the increase due to measurement error in the number of cases needed in a main study/external study design using regression calibration, as a function of the validation study sample size, the correlation between **X** and **W**, the odds ratio, and the conditional variance of **X** given **W**. They inverted this expression to optimize n_2 as a function of these other parameters, subject to a specified degree of precision in the regression calibration estimate of β per subject.

Conclusion

In main study/validation study design, the criterion used for design optimization should be carefully chosen. For both validity and efficiency considerations, internal validation studies are preferred over external ones. Particularly when the sample disease frequency is not rare and when the cost of the gold standard is not prohibitive, completely validated designs may be optimal. Owing to the lack of user-friendly software, it does not appear that explicitly optimized main study/validation study designs have been used by scientific investigators. Further work could involve making the identification of optimal designs more accessible in the field.

Acknowledgment

This work was supported by National Cancer Institute grants CA50587 and CA03416.

References

[1] Armstrong, B.K., White, E. & Saracci, R. (1992). *Principles of Exposure Measurement in Epidemiology*. Oxford University Press, Oxford, Chapter 4.

[2] Barron, B.A. (1977). The effects of misclassification on the estimation of relative risk, *Biometrics* **33**, 414–418.

[3] Buonaccorsi, J.P. (1990). Doubling sampling for exact values in the normal discriminant model with application to binary regression, *Communications in Statistics – Theory and Methods* **19**, 4569–4586.

[4] Chernoff, H. & Haitovsky, Y. (1990). Locally optimal designs for comparing two probabilities from binomial data subject to misclassification, *Biometrika* **77**, 797–806.

[5] Greenland, S. (1988). On sample size and power calculations for studies using confidence intervals, *American Journal of Epidemiology* **128**, 231–236.

[6] Greenland, S. (1988). Statistical uncertainty due to misclassification: implications for validation substudies, *Journal of Clinical Epidemiology* **41**, 1167–1174.

[7] Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification, *Statistics in Medicine* **7**, 745–757.

[8] IMSL (1987). *User's Manual: Math Library*. IMSL, Houston.

[9] Kaaks, R., Riboli, E. & van Staveren, W. (1995). Sample size requirements for calibration studies of dietary intake measurements in prospective cohort investigations, *American Journal of Epidemiology* **142**, 557–565.

[10] Küchenhoff, H. (1990). *Logit- und Probitregression mit Fehlen in den Variabeln*. Anton Hain, Frankfurt am Main.

[11] Palmgren, J. (1987). Precision of double sampling estimators for comparing two probabilities, *Biometrika* **74**, 687–694.

[12] Selen, J. (1986). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data, *Journal of the American Statistical Association*, **81**, 75–81.

[13] Spiegelman, D. & Gray, R. (1991). Cost-efficient study designs for binary response data with Gaussian covariate measurement error, *Biometrics* **47**, 851–869.

[14] Spiegelman, D., Schneeweiss, S. & McDermott, A. (1997). Measurement error correction for logistic regression models with an "alloyed gold standard", *American Journal of Epidemiology* **145**, 184–196.

[15] Stram, D.O., Longnecker, M.P., Shames, L., Kolonel, L.N., Wildens, L.R., Pike, M.C. & Henderson, B.E. (1995). Cost-efficient design of a diet validation study, *American Journal of Epidemiology* **142**, 353–362.

[16] Tosteson, T. & Tsiatis, A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates, *Biometrika* **77**, 11–20.

[17] Tosteson, T.D. & Ware, J.H. (1990). Designing a logistic regression study using surrogate measures for exposure and outcome, *Biometrika* **77**, 11–21.

[18] Wacholder, S., Armstrong, B. & Hartge, P. (1993). Validation studies using an alloyed gold standard, *American Journal of Epidemiology* **137**, 1251–1258.

[19] Willett, W.C. (1990). *Nutritional Epidemiology*. Oxford University Press, New York, pp. 115–118.

[20] Zelen, M. & Haitovsky, Y. (1991). Testing hypotheses with binary data subject to misclassification errors: analysis and experimental design, *Biometrika* **78**, 857–865.

DONNA SPIEGELMAN

Validity and Generalizability in Epidemiologic Studies

The validity of a study of human subjects is often separated into two components: the validity of the **inferences** drawn as they pertain to members of the source population (*internal validity*), and the validity of the inferences as they pertain to people outside that population (*external validity* or *generalizability*). Internal validity parallels the statistical concept of generalizing from sample to source population, while generalizability involves more informal inference beyond a source population to **target populations**.

Scientific generalization extends beyond statistical generalization of study results to the formulation of abstract concepts relating the study factors. The concepts are abstract in the sense that they are not tied to specific populations; instead they amount to the specification of a more general scientific theory. Internal validity is a prerequisite for the study to contribute usefully to this process of abstraction, but the generalization process is otherwise separate from the concerns of internal validity and the mechanics of the study design.

Validity

Internal validity implies validity of inference for the study subjects themselves. Specifically, it implies an accurate measurement apart from **random errors**. Numerous types of **biases** can detract from internal validity; for examples, see [19]. The distinction among these biases is occasionally difficult to make, but three general types can be identified: **selection bias**, **confounding**, and information bias. These categories are not always clearly demarcated; factors that appear to be responsible for a selection bias can also be viewed, under some circumstances, as confounding factors. Occasionally, certain information biases can also be construed as confounding.

Selection Bias

Most epidemiologic studies involve a comparison of two or more groups with regard to either disease or exposure frequency. Bias is a distortion of the effect that is measured. Selection biases are distortions that result from procedures used to select subjects, and from factors that influence study participation.

Self-Selection Bias. One form of such bias is *self-selection* bias. When the **Centers for Disease Control (CDC)** investigated subsequent leukemia incidence among troops who had been present at the Smoky Atomic Test in Nevada [3], 76% of the troops identified as members of that cohort (*see* **Cohort Study**) had known outcomes. Of this 76%, 82% were traced by the investigators, but the other 18% contacted the investigators on their own initiative in response to publicity about the investigation. This self-referral of subjects is ordinarily considered a threat to validity, since the reasons for self-referral may be associated with the outcome under study [6]. In the Smoky study, there were four leukemia cases among the $0.18 \times 0.76 = 15\%$ of cohort members who referred themselves and four among the $0.82 \times 0.76 = 62\%$ of cohort members traced by the investigators, for a total of eight cases among the 76% of the cohort with known outcomes. These data indicate that self-selection bias was a small but real problem in the Smoky study. If the 24% of the cohort with unknown outcomes had a leukemia incidence like that of the subjects traced by the investigators, then we should expect that only $4(24/62) = 1.5$ or about one or two cases occurred among this 24%, for a total of only nine or 10 cases in the entire cohort. If, however, we assumed that the 24% with unknown outcomes had a leukemia incidence like that of subjects with known outcomes, then we would calculate that $8(24/76) = 2.5$ or about two or three cases occurred among this 24%, for a total of 10 or 11 cases in the entire cohort.

Self-selection can also occur before subjects are identified for study. For example, it is routine to find that the mortality of active workers is less than that of the population as a whole [9, 15]. This "healthy-worker effect" presumably derives from a screening process, perhaps largely self-selection, that allows relatively healthy people to become or remain workers, whereas those who remain unemployed, retired, disabled, or otherwise out of the active worker population are as a group less healthy [23] (*see* **Occupational Epidemiology**).

Diagnostic Bias. Another type of selection bias occurring before subjects are identified for study is

diagnostic bias [19]. When the relation between oral contraceptives and venous thromboembolism was first investigated with **case–control studies of hospitalized patients**, there was concern that some of the women had been hospitalized with a diagnosis of venous thromboembolism because their physicians suspected a relation between this disease and oral contraceptives and had known about oral contraceptive use in patients who presented with suggestive symptoms [20]. A study of hospitalized patients with thromboembolism could lead to an exaggerated estimate of the effect of oral contraceptives on thromboembolism if the hospitalization and determination of the diagnosis were influenced by the history of oral contraceptive use.

Many varieties of selection bias could be described. The common element of such biases is that the relation between exposure and disease is different for those who participate and those who should be theoretically eligible for study, including those who do not participate. The result is that associations observed in the study represent a mix of forces determining participation as well as forces determining disease. It is sometimes (but not always) possible to disentangle the effects of participation determinants from those of disease determinants using analytic methods for the control of confounding.

Confounding. The term **confounding** has been used for several different concepts. Although this bias can occur in experiments, it is a considerably more important issue in nonexperimental research.

On the simplest level, confounding may be considered a confusion of effects. Specifically, the apparent effect of the exposure of interest is distorted because the effect of an extraneous factor is mistaken for or mixed with the actual exposure effect (which may be null). The distortion introduced by a confounding factor can be large, and it can lead to overestimation or underestimation of an effect depending on the direction of the **associations** that the confounding factor has with exposure and disease. Confounding can even change the apparent direction of an effect.

A more precise definition of confounding begins by considering the manner in which effects are estimated. Let us assume that we wish to estimate the degree to which exposure has changed the frequency of disease in an exposed cohort. To do so, we must estimate what the frequency of disease would have been in this cohort had exposure been absent. To accomplish this task, we observe the disease frequency in an unexposed cohort. But rarely could we take this unexposed frequency as fairly representing what the frequency would have been in the exposed cohort had exposure been absent, because the unexposed cohort would differ from the exposed cohort on many factors that affect disease frequency besides exposure. To express this problem, we say that the comparison of the exposed and unexposed is *confounded*, because the difference in disease frequency between the exposed and unexposed results from a mixture of several effects, including (but not limited to) any exposure effect.

The extraneous factors responsible for difference in disease frequency between the exposed and unexposed are called **confounders**. In addition, factors associated with these extraneous causal factors that can serve as surrogates for these factors are also commonly called confounders. The most extreme example of such a surrogate is chronologic age. Increasing age is strongly associated with *aging* – the accumulation of cell mutations and tissue damage that lead to disease – but increasing age does not itself cause such pathogenic changes, for it is just a measure of how much time has passed since birth.

Regardless of whether a confounder is a cause of the study disease or merely a surrogate for such a cause, its chief characteristic is that it would be predictive of disease frequency within the unexposed (reference) cohort – otherwise it could not explain why the unexposed cohort fails to represent properly what the exposed cohort would experience in the absence of exposure. For example, suppose all the exposed were men and all the unexposed women. If unexposed men would have the same incidence as unexposed women, then the fact that all the unexposed were women rather than men could not account for any confounding that is present.

Information Bias

Once the subjects to be compared have been identified, the information to be compared must be obtained. Bias in evaluating an effect can occur from errors in obtaining the needed information. Information bias can occur whenever there are errors in the measurement of subjects, but the consequences of the errors are different depending on whether the distribution of errors for one variable (for example,

exposure or disease) depends on the actual value of other variables. and errors in other variables.

For discrete variables, **measurement error** is usually called classification error or **misclassification**. Classification error that depends on the values of other variables is referred to as *differential misclassification* (*see* **Differential Error**). Classification error that does not depend on the values of other variables is referred to as *nondifferential misclassification* (*see* **Nondifferential Error**).

Differential Misclassification. Suppose a cohort study were undertaken to compare incidence rates of emphysema among smokers and nonsmokers. Emphysema is a disease that may go undiagnosed without unusual medical attention. If smokers, because of concern about health-related effects of smoking or as a consequence of other health effects of smoking (such as bronchitis), seek medical attention to a greater degree than nonsmokers, then emphysema might be diagnosed more frequently among smokers than among nonsmokers simply as a consequence of the greater medical attention. Unless steps were taken to ensure comparable follow-up, an information bias would result: a spurious excess of emphysema incidence would be found among smokers compared with nonsmokers that is unrelated to any biologic effect of smoking. This is an example of differential misclassification, since the underdiagnosis of emphysema, a classification error, occurs more frequently for nonsmokers than for smokers. Unlike the diagnostic bias in the studies of oral contraceptives and thromboembolism described earlier, it is not a selection bias, since it occurs among subjects already included in the study. Nevertheless, the similarities between some selection biases and differential misclassification biases are worth noting.

In case–control studies of congenital malformations, the etiologic information may be obtained at interview from mothers. The case mothers have recently given birth to a malformed baby, whereas the vast majority of control mothers have recently given birth to an apparently healthy baby. Another variety of differential misclassification, referred to as **recall bias**, can result if the mothers of malformed infants recall exposures more thoroughly than mothers of healthy infants. It is supposed that the birth of a malformed infant serves as a stimulus to a mother to recall all events that might have played some role in the unfortunate outcome. Presumably such women will remember exposures such as infectious disease, trauma, and drugs more accurately than mothers of healthy infants, who have not had a comparable stimulus. Consequently, information on such exposures will be ascertained more frequently from mothers of malformed babies, and an apparent effect, unrelated to any biologic effect, will result from this recall bias. Recall bias is a possibility in any case–control study that uses an anamnestic response, since the cases and controls by definition are people who differ with respect to their disease experience, and this difference may affect recall.

The bias that is caused by differential misclassification can either exaggerate or underestimate an effect. In each of the examples above, the misclassification serves to exaggerate the effects under study, but examples to the contrary can also be found. Because of the relatively unpredictable effects of differential misclassification, some investigators go through elaborate procedures to ensure that the misclassification will be nondifferential, such as **blinding** of exposure evaluations with respect to outcome status. Unfortunately, even in situations when blinding is accomplished or in cohort studies in which disease outcomes have not yet occurred, collapsing continuous or categorical exposure data into fewer categories can induce differential misclassification [8, 21].

Nondifferential Misclassification. Nondifferential exposure or disease misclassification occurs when the proportion of subjects misclassified on exposure does not depend on disease status, or when the proportion of subjects misclassified on disease does not depend on exposure. When the misclassification is independent of other errors, bias introduced by such nondifferential misclassification of a **binary** exposure or disease is predictable in direction, namely toward the null value [5, 11, 12, 17] (*see* **Bias Toward the Null**). Contrary to popular misconceptions, however, nondifferential exposure or disease misclassification can sometimes produce bias away from the null, especially if the errors in exposure and disease classification are correlated [4, 7, 13, 22]. If the misclassification is extreme, the misclassification can go beyond the null value and reverse direction.

When the exposure is **polytomous** (that is, has more than two categories) and there is nondifferential misclassification between two of the categories and no others, the effect estimates for those two categories will be biased toward one another [1,

22]. In particular, the effect estimate for the lower exposure category will be shifted toward that of the higher exposure category, and away from the null value. It is also possible for independent nondifferential misclassification to bias trend estimates away from the null or reverse a trend [7]. Such examples are unusual, however, because trend reversal cannot occur if the **mean** exposure measurement increases with true exposure [24].

Nondifferential Misclassification of Disease. The effects of nondifferential misclassification of disease resemble those of nondifferential misclassification of exposure. In most situations, nondifferential misclassification of a binary disease outcome will produce bias toward the null, provided that the misclassification is independent of other errors. There are, however, some useful special cases in which such misclassification produces no bias in the risk ratio (*see* **Relative Risk**); in addition, the bias in the risk difference is a simple function of the **sensitivity** and **specificity**. For a discussion, see Rothman & Greenland [18].

Pervasiveness of Nondifferential Misclassification. Since the bias from independent nondifferential misclassification of a dichotomous exposure is always in the direction of the null value, historically it has not been a great source of concern to epidemiologists, who have generally considered it more acceptable to underestimate effects than to overestimate effects. Nevertheless, such misclassification is a serious problem: the bias it introduces may account for certain discrepancies among epidemiologic studies. Many studies ascertain information in a way that guarantees substantial misclassification, and many studies use classification schemes that can mask effects in a manner identical to nondifferential misclassification.

Suppose aspirin transiently reduces risk of myocardial infarction. The word *transiently* implies a brief induction period. Any study that considered as exposure aspirin use outside of a narrow time interval before the occurrence of a myocardial infarction would be misclassifying aspirin use: there is relevant use of aspirin, and there is use of aspirin that is irrelevant because it does not allow the exposure to act causally under the causal hypothesis with its specified induction period. Many studies ask about "ever use" (use at any time during an individual's life) of drugs or other exposures. Such cumulative indices over an individual's lifetime inevitably augment possibly relevant exposure with irrelevant exposure, and can thus introduce a bias toward the null value through nondifferential misclassification.

In cohort studies in which there are disease categories with few subjects, investigators are occasionally tempted to combine outcome categories to increase the number of subjects in each analysis, thereby gaining precision. This collapsing of categories can obscure effects on more narrowly defined disease categories.

Nondifferential exposure and disease misclassification is a greater concern in interpreting studies that seem to indicate the absence of an effect. Consequently, in studies that indicate little or no effect, it is crucial for the researchers to consider the problem of nondifferential misclassification to determine to what extent a real effect might have been obscured. On the other hand, in studies that describe a strong nonzero effect, preoccupation with nondifferential exposure and disease misclassification is rarely warranted, provided that the errors are independent. Occasionally, critics of a study will argue that poor exposure data or a poor disease classification invalidate the results. This argument is incorrect, however, if the results indicate a nonzero effect and one can be sure that the classification errors produced bias towards the null, since the bias will be in the direction of underestimating the effect.

Generally speaking, it is incorrect to dismiss a study reporting an effect simply because there is substantial nondifferential misclassification of exposure, since an estimate of effect without the misclassification could be even greater, provided that the misclassification probabilities apply uniformly to all subjects. Thus, the implications of nondifferential misclassification depend heavily on whether the study is perceived as "positive" or "negative". Emphasis on measurement instead of on a qualitative description of study results lessens the likelihood for misinterpretation, but even so it is important to bear in mind the direction and likely magnitude of a bias.

Misclassification of Confounders. If a confounding variable is misclassified, the ability to control confounding in the analysis is hampered [2, 10, 14]. While independent nondifferential misclassification of exposure or disease usually biases study results in the direction of the null hypothesis, independent nondifferential misclassification of a confounding

variable will usually reduce the degree to which confounding can be controlled and thus can cause a bias in either direction, depending on the direction of the confounding. For this reason, misclassification of confounding factors can be a serious problem.

If the confounding is strong and the exposure–disease relation is weak or zero, misclassification of the confounding factor can lead to extremely misleading results. For example, a strong causal relation between smoking and bladder cancer, coupled with a strong association between smoking and coffee drinking, makes smoking a strong confounder of any possible relation between coffee drinking and bladder cancer. Since the control of confounding by smoking depends on accurate smoking information, and since some misclassification of the relevant smoking information is inevitable no matter how smoking is measured, some residual confounding is inevitable [16]. The problem of residual confounding would be even worse if the only available information on smoking were a simple dichotomy such as "ever smoked" vs. "never smoked", since the lack of detailed specification of smoking prohibits adequate control of confounding. The resulting confounding is especially troublesome because to many investigators and readers it may appear that confounding by smoking has been controlled.

Generalizability

Many epidemiologists and statisticians have taught that generalization from a study group depends on the study group being a representative subgroup of the target population, in the sense of a sample. If scientific generalization were simply a matter of statistical generalization, however, it would be limited literally to those individuals who might have been included, through sampling, as study subjects. If this notion were correct, there would be no application to humans of any results obtained from animal research. In addition, every population would require its own set of studies, and these studies would have to be repeated for every new generation.

The tendency to use "representative" study groups probably derives from early experience with **surveys** for which the inferential goal was only description of the surveyed population. Social scientists often rely on statistical inference because decisions about what is relevant for generalization are more difficult in the **social sciences**, and populations are considerably more diverse in sociologic phenomena than in biologic phenomena. In the biologic sciences, however, investigators conduct experiments using animals with characteristics selected to enhance the validity of the experimental work rather than to represent the target population. Epidemiologic study designs are usually stronger if subject selection is guided by the need to make a valid comparison, which may call for severe restriction of admissible subjects to a narrow range of characteristics, rather than by an attempt to make the subjects representative, in a sampling sense, of the potential target populations.

Ultimately, the goal of a purely scientific study is to contribute to scientific knowledge. The process of synthesizing knowledge from observations is, after centuries of examination, not yet well understood. In most sciences, however, the process involves moving from the particulars of a set of observations to the abstraction of a scientific hypothesis or theory that is more or less divorced from time and place: the abstractions apply to a broader domain of experience than that observed or sampled from. Such scientific generalization amounts to moving from time- and place-specific observations to an abstract "universal" hypothesis, such as "cigarette smoking causes lung cancer". This process is neither mechanical nor statistical, nor does it involve specific target populations (although the hypothesis may be limited to certain biological subgroups, such as a specific **genotype**). In this sense, the term external validity is a misnomer, and the term generalization must be interpreted as abstraction. Selection of study groups that are representative of larger populations in the statistical sense will generally not enhance the ability to abstract universal statements from observations, but selection of study groups for characteristics that enable a study to distinguish effectively between competing scientific hypotheses will do so.

In addition to scientific goals, some studies also have a goal of measuring effects and predicting the impact of interventions in a specific target population. In contrast to scientific inference, these pragmatic goals may depend more closely on the representativeness of study subjects with respect to the target population. For example, if a **clinical trial** is conducted using patients with a good **prognosis**, the results from the trial may not predict well the results when the new intervention is applied to patients with a poor prognosis. Thus, some effort may be needed

in the study design to ensure that enough subjects are included from each of several major subgroups of the target, such as males and females. Even in this situation, complete representativeness is not always desirable, for a more efficient study might be obtained by oversampling some subgroups and then standardizing the study estimate to the target population.

Acknowledgment

Adapted from Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Ed. Lippincott, Philadelphia, Chapter 8, with permission.

References

[1] Birkett, N.J. (1992). Effect of nondifferential misclassification of estimates of odds ratios with multiple levels of exposure, *American Journal of Epidemiology* **136**, 356–362.

[2] Brenner, H. (1993). Bias due to non-differential misclassification of polytomous confounders, *Journal of Clinical Epidemiology* **46**, 57–63.

[3] Caldwell, G.G., Kelley, D.B. & Heath, C.W. Jr (1980). Leukemia among participants in military maneuvers at a nuclear bomb test: a preliminary report, *Journal of the American Medical Association* **244**, 1575–1578.

[4] Chavance, M., Dellatolas, G. & Lellouch, J. (1992). Correlated nondifferential misclassifications of disease and exposure, *International Journal of Epidemiology* **21**, 537–546.

[5] Copeland, K.T., Checkoway, H., Holbrook, R.H. & McMichael, A.J. (1977). Bias due to misclassification in the estimate of relative risk, *American Journal of Epidemiology* **105**, 488–495.

[6] Criqui, M.H., Austin, M. & Barrett-Connor, E. (1979). The effect of non-response on risk ratios in a cardiovascular disease study, *Journal of Chronic Diseases* **32**, 633–638.

[7] Dosemeci, M., Wacholder, S. & Lubin, J. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value?, *American Journal of Epidemiology* **132**, 746–749.

[8] Flegal, K.M., Keyl, P.M. & Nieto, F.J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement, *American Journal of Epidemiology* **134**, 1233–1244.

[9] Fox, A.J. & Collier, P.F. (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry, *British Journal of Preventive and Social Medicine* **30**, 225–230.

[10] Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564–569.

[11] Gullen, W.H., Berman, J.E. & Johnson, E.A. (1968). Effects of misclassification in epidemiologic studies, *Public Health Reports* **53**, 1956–1965.

[12] Keys, A. & Kihlberg, J.K. (1963). The effect of misclassification on the estimated relative prevalence of a characteristic, *American Journal of Public Health* **53**, 1656–1665.

[13] Kristensen, P. (1992). Bias from nondifferential but dependent misclassification of exposure and outcome, *Epidemiology* **3**, 210–215.

[14] Marshall, J.R. & Hastrup, J.L. (1996). Mismeasurement and the resonance of strong confounders: uncorrelated errors, *American Journal of Epidemiology* **143**, 1069–1078.

[15] McMichael, A.J. (1976). Standardized mortality ratios and the "healthy worker effect": scratching beneath the surface, *Journal of Occupational Medicine* **18**, 165–168.

[16] Morrison, A.S., Buring, J.E., Verhoek, W.G., Aoki, K., Leck, I., Ohno, Y. & Obata, K. (1982). Coffee drinking and cancer of the lower urinary tract, *Journal of the National Cancer Institute* **68**, 91–94.

[17] Newell, D.J. (1962). Errors in interpretation of errors in epidemiology, *American Journal of Public Health* **52**, 1925–1928.

[18] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Ed. Lippincott, Philadelphia, Chapter 8.

[19] Sackett, D.L. (1979). Bias in analytic research, *Journal of Chronic Diseases* **32**, 51–63.

[20] Sartwell, P.E., Masi, A.T., Arthes, F.G., Greene, G.R. & Smith, H.E. (1969). Thromboembolism and oral contraceptives: an epidemiologic case-control study, *American Journal of Epidemiology* **90**, 365–380.

[21] Wacholder, S., Dosemeci, M. & Lubin, J.H. (1991). Blind assignment of exposure does not prevent differential misclassification, *American Journal of Epidemiology* **134**, 433–437.

[22] Walker, A.M. & Blettner, M. (1985). Comparing imperfect measures of exposure, *American Journal of Epidemiology* **121**, 783–790.

[23] Wang, J.D. & Miettinen, O.S. (1982). Occupational mortality studies: principles of validity, *Scandinavian Journal of Environmental Health* **8**, 153–158.

[24] Weinberg, C.R., Umbach, D.M. & Greenland, S. (1994). When will nondifferential misclassification preserve the direction of a trend?, *American Journal of Epidemiology* **140**, 565–571.

(*See also* **Bias in Case–Control Studies**; **Bias in Cohort Studies**; **Bias in Observational Studies**; **Bias, Overview**)

KENNETH J. ROTHMAN &
SANDER GREENLAND

Variable Selection

Automated variable selection is often used for the selection of **explanatory variables** in regression (*see* **Multiple Linear Regression**). This may be in the context of linear or nonlinear models with continuous or categorical data. More broadly, variable selection concerns the choice of variables derived from an original set of available explanatory variables, including **interaction** and nonlinear terms, and orthogonal series decompositions (principal components (*see* **Principal Components Analysis**), Fourier series, **splines**, **wavelets** (*see* **Orthogonality**). This involves substantive considerations of the application, and cannot necessarily be based on automated means. In short, it concerns the modeling process.

The most common use is for the selection of explanatory variables in least squares linear regression, often by automated means, to arrive at a single "best" model. Variables are included or excluded on the basis of:

1. Statistical tests that regression coefficients or groups of regression coefficients are zero.
2. Model choice criteria (*see* **Model, Choice of**), e.g. **Mallows' C_p statistic**, **Bayes** information criterion (BIC) (see the section "Bayes Selection" below).
3. **Cross-validation** or **bootstrap** prediction.

The first of these broadly relates to **estimation** accuracy, whereas the latter two relate to **prediction** accuracy.

Many caveats need to be attached to such automated procedures, and we discuss some of them later in this article. With k explanatory variables there are 2^k possible regression models, depending on whether or not each variable is included. When k is more than about 15 (or perhaps less) a complete search of all models is often abandoned in favor of a restricted automated search, broadly categorized as:

1. backward, forward, and stepwise selection
2. using a branch-and-bound **algorithm**
3. using a stochastic algorithm for finding a good model.

The search algorithms may also aim to provide a subset of models, all of which may be judged to be alternative good models. One may then wish to choose from within this primitive subset a model which makes physical sense or alternatively report the range of suitable models, or even provide predictions based on averaging the range of good models; see, for example, [6] and [18]. The use of stochastic algorithms is mushrooming, especially in the area of Bayesian variable selection (*see* **Bayesian Methods**), using **Markov chain Monte Carlo** (MCMC).

Before considering these aspects in more detail we illustrate some simple selection algorithms with data on the factors associated with infant birth weight, collected from Baystate Medical Centre, Massachusetts, during 1986, given by Hosmer & Lemeshow [12]. The birth weights of the 189 infants ranged from 709 g to 4990 g. It was thought that variation in birth weight might be explained by nine variables: *AG*e of mother (years), *W*eigh*T* in pounds at last menstrual period, *SM*oking status during pregnancy, number of *PR*evious premature births, *HY*pertension presence, uterine *IR*ritability, number of physician *VI*sits during first trimester, and two race dummy variables to define *BL*ack and *OT*her to cover the three categories, White, Black, and Other. Our purpose here is not to examine whether these are the most sensible variables to employ, but rather to illustrate some of the selection issues elaborated on in the following sections of this article. The regression equation for all nine variables included is

$$\begin{aligned} Birthwt = {} & 2928 - 3.57AG + 4.35WT - 352SM \\ & - 48PR - 593HY - 516IR - 14.1VI \\ & - 488BL - 355OT. \end{aligned} \quad (1)$$

Only 24% of the variation is explained, and Student's t-ratios on individual coefficients range from 3.72 on IR to 0.30 on VI (in absolute terms), with three <2: AG, PR, and VI. If one successively removes the most insignificant variable, refitting the remaining variables and stopping when all remaining variables are significantly different from zero on the basis of a nominal 5% F test, then a six-variable model results. This model is

$$\begin{aligned} Birthwt = {} & 2837 + 4.24WT - 356SM - 585HY \\ & - 526IR - 475BL - 348OT. \end{aligned} \quad (2)$$

The t ratios for the six variables (based on newly estimated residual error) range from 2.53 to 3.90 in absolute terms with 24% variation explained. In

fact, this model contains all the variables significant at the 5% level. None of the remaining variables, either singly or in combination, explains significant variation. It also happens to be the preferred model of the various search methods based on statistical testing as well as C_p model choice. In this example it hardly leads to an impressively well-fitting model, and has a sizable residual standard deviation of 646, and yet at least has coefficients which are plausible in direction. A more detailed summary of alternative models is given in Table 1, with the best two models of each size from 1 to 8 together with the full nine-variable model.

The residual standard deviation in the full model, $\hat{\sigma} = 650.3$, is used for

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2 + 2p - n} \tag{3}$$

in all p variable models fitted, $p = 1, \ldots, 9$. Here, RSS_p is the residual sum of squares for the particular p-variable model, and the estimated residual variance for this model is $s^2 = \text{RSS}_p/(n - p - 1)$.

If one were to base choice of model on percentage of variation explained (R^2), then the full nine-variable model is best. Adjusted R^2, given as

$$\text{adj. } R_p^2 = 1 - \left(1 - R_p^2\right) \frac{n}{(n - p)}, \tag{4}$$

suggests a six-variable model with WT, SM, PR, IR, BL, and OT included. Mallows' C_p leads to the same six-variable model.

The hoped for advantage of subset models rests with reduced **mean square error** loss, to be achieved by reduction of variance at the expense of minimal increase in bias. It may offer prediction or estimation robustness by avoiding those variables whose effect is nonlinear or which interact with other variables. Occasionally there is real benefit in terms of reduced cost in not needing to measure particular variables, and this may be particularly true in the medical context where some measurements may be invasive to the patient.

Automatic selection procedures are perhaps most appropriate when there is a large number of candidate variables and no clear prior knowledge of variables likely to be related to the response. These procedures are built into many statistical packages (*see* **Software, Biostatistical**). With least squares multiple regression they provide perhaps the most heavily used, and abused, statistical technique. Focusing on just one model often ignores other good models and the uncertainty, due to the process of sifting, that ought to attach to that selection. Some of the technical aspects are described below. For further discussion of strategic considerations both in the choice of measurement and **measurement scale**, and choice

Table 1 Best two models of each size for predicting birth weight from nine explanatory variables, percentage variation explained (R^2), adjusted R^2, Mallows' C_p, and residual standard deviation

					Variables								
Size	R^2	Adj. R^2	C_p	s	AG	WT	SM	PR	HY	IR	VI	BL	OT
1	8.1	7.6	32.3	701						X			
1	3.6	3.1	42.8	718			X						
2	11.2	10.3	26.8	691					X	X			
2	11.1	10.1	27.2	691			X			X			
3	14.8	13.4	20.4	679		X			X	X			
3	14.7	13.3	20.6	679			X			X			X
4	18.9	17.1	12.7	664			X			X		X	X
4	17.7	16.0	15.4	669			X		X	X			X
5	21.4	19.2	8.9	655			X		X	X		X	X
5	20.5	18.3	11.0	659		X	X			X		X	X
6	24.0	21.5	4.6	646		X	X		X	X		X	X
6	21.6	19.0	10.3	656			X	X	X	X		X	X
7	24.2	21.2	6.3	647		X	X	X	X	X		X	X
7	24.1	21.2	6.3	647	X	X	X		X	X		X	X
8	24.2	20.9	8.1	649	X	X	X	X	X	X		X	X
8	24.2	20.8	8.1	649		X	X	X	X	X	X	X	X
9	24.3	20.5	10.0	650	X	X	X	X	X	X	X	X	X

of variables, see [5]. Raab [22], in a cross-sectional study of blood lead levels on children's abilities, emphasizes the importance of study design to separate **confounders** from exposure variables and the desirability of purposeful predetermined inclusion of some covariates. Related topics are **Model, Choice of**, **Parsimony**, **Akaike's Criteria**, **Model Checking**, **Diagnostics**, and **Residuals**.

Restricted Search Methods

Forward, Backward, and Stepwise Selection

These methods are all based on significance tests as to whether to enter or delete a variable. The estimated residual variance is usually kept constant during the search, either being specified from the full model with k variables or from replicates or near replicates in the data. For stepwise selection there are two quantities, *FADD* and *FDROP*, and these may be specified by the appropriate value of the tabulated F distribution (*see* ***F*** **Distributions**) to test whether the coefficient can be assumed to be zero for the candidate variable in the context of the currently entertained model involving p of the k regressors, typically on 1 and $n-k-1$ degrees of freedom. Suppose the model currently includes explanatory variables $X_1, X_2, \ldots, X_p$. For each of these p variables the F statistic (= square of t statistic) is compared with *FDROP*. If the minimum F statistic is less than *FDROP*, then that variable is removed from the current model, $p \leftarrow p-1$, and the variable goes into the pool of variables not currently used, otherwise no variable is dropped at this stage. Each variable in this pool of unused variables is examined now in turn to see whether it should be added to the current model. If the largest F statistic from augmenting the model with a variable from the available pool is greater than *FADD*, then the variable is added to the current model and $p \leftarrow p+1$, otherwise no variable is added at this stage. Clearly, to make sense, $FADD \geq FDROP$. The process of adding or subtracting variables continues until the model no longer changes, and all variables included have F statistics greater than *FDROP* and all variables not included have F statistics for inclusion less than *FADD*. The stepwise algorithm may start from either no variables in the model and build up the model, or from all variables in the model and strip down the model. It might also start from some arbitrarily specified model with some of the main variables included, or even a randomly chosen starting set. The build-up mode is the only feasible option when there are more variables than observations and the full variable model is overparameterized, and fits perfectly, with zero residual variation.

Often in stepwise regression one chooses $FADD = FDROP$. A value around 4 corresponds to a 5% significance test ($1.96^2 = 3.84$) for infinite error degrees of freedom and consequent normality of the estimated regression coefficient. For moderate or few degrees of freedom on the residual error estimate, then the *FADD*, *FDROP* pair ought to be correspondingly larger than 4. The use of Mallows' C_p as a criterion for subset choice implies $FADD = FDROP = 2$ since the residual sum of squares for the candidate p variable model, divided by estimated residual variance, is penalized by adding $2p$, that is 2 per new variable added. Thus Mallows' C_p will tend to choose larger models than arising out of 5% significance tests. However, C_p was not designed by Mallows [19] as an automatic selection procedure, but rather as a graphical method of comparing models of differing dimensions exhaustively.

Forward selection and backward elimination may be thought of as special cases of the stepwise algorithm. Forward selection starts with no variables in the model and has $FDROP = 0$, so that variables are added one at a time until no variable not in the equation has an F statistic for entry greater than *FADD*. There is no guarantee that the variables in the final equation all have F statistics greater than *FADD*, since once entered in the sequential process they are not considered for removal at any future stage.

Backward elimination corresponds to the stepwise algorithm with $FADD = \infty -$ in practice, some chosen large number. It starts from all variables in the equation. The procedure stops when no variable in the model has an F statistic less than *FDROP*.

None of these methods guarantees that the model chosen is the best out of the 2^k possible models. Particular combinations of variables may be missed completely. The attraction of the algorithms is that they are very fast, whereas a complete enumerative search of the 2^k models becomes infeasible for k much greater than about 15.

A number of issues are raised by adoption of one of these automatic procedures, aside from whether they find the "best" model. First, the idea of finding a single best model is suspect because there may

be other near "best" models which are much more plausible and substantively interesting. Secondly, the extent of the sifting process with repeated significance tests and search for the maximum F statistic is not quantified in terms of increased uncertainty that the optimal model is best or even good. This is discussed in some detail by Miller [20, Chapter 3]. Breiman & Spector [3] evaluate backward selection techniques in terms of prediction error with special emphasis on a bootstrap and cross-validation choice of model. Bootstrap and 5-fold cross-classification choices fare particularly well.

Branch-and-Bound Algorithms

Criteria such as minimizing the residual sum of squares or prediction error have a particular monotonicity property. If $\mathcal{A}$ is a set of variable labels and $\mathcal{B}$ is a subset of $\mathcal{A}$, then

$$\mathrm{RSS}(\mathcal{A}) \leq \mathrm{RSS}(\mathcal{B}).$$

Many of the criteria used in identifying the "best" subset of variables are monotone in RSS given subsets with the same number of independent variables. These include adjusted R^2 and Mallows' C_p; exceptions are the bootstrap procedures of Breiman & Spector [3] and the leave-one-out cross-validatory statistic PRESS of Allen [2], although the latter is asymptotically equivalent to C_p [25].

The branch-and-bound algorithm, in its most satisfying form given by Furnival & Wilson [9], relies on this simple monotonicity. If a model with $p < p^*$ variables has a smaller RSS (and cannot therefore be a subset), then the model with p^* variables and *all* its submodels must be inferior to the model with the alternative set of p variables. If models are thought of as being generated by a binary tree, then the branch with the particular set of p^* variables can be cut off and all its submodels ignored. Furnival & Wilson's approach provides a clever exploitation of this with an implementation that simultaneously creates two tableaux, one being for bounds, moving through the tree in complementary directions (*see* **Tree-structured Statistical Methods**).

The procedure can be readily modified to provide, say, the five best subsets of each size. Whilst all-subsets regression is only feasible for up to around 15 variables, the branch-and-bound algorithm can extend this to 30 or more variables, although its benefits reduce with increasing correlation between the independent variables. It cannot begin to tackle data generated by modern instruments used for chemometrics in the pharmaceutical industry, where 700 explanatory absorbances/reflectances at 700 wavelengths are not unusual. See [4, Chapter 7] for a variety of graphical and algorithmic techniques for such high dimensions.

Non-Gaussian Models

The techniques described readily extend to nonnormal models, common in medical studies. The residual sum of squares for a Gaussian model considered above is equal to $-2 \times$ ln likelihood maximized over the regression parameters, aside from a scale factor of $1/\sigma^2$. In the non-Gaussian case this so-called deviance (*see* **Generalized Linear Model**, or [7]), can be used as a basis of using **likelihood ratio tests** for comparison of two models, with either an assumed asymptotic **chi-square distribution** or an ***F* distribution** when the scale factor is estimated, in direct analogy with the Gaussian case.

In moving from Gaussian to non-Gaussian models, there is usually the need for iterative methods in fitting models by **maximum likelihood** (*see* **Optimization and Nonlinear Equations**). This imposes a considerable computational overhead. Lawless & Singhal [14] use a first-order approximation to the log likelihood of the submodel to speed up computations. Nordberg [21] shows how to incorporate such approximations so that standard least squares regression computer packages may be used, together with their associated best subsets/stepwise selection routines.

Bayes Selection

Model Choice

Suppose M_1 and M_2 denote two different regression models, with p_1 and p_2 variables and nonzero regression parameters β_1 and β_2, e.g. 1 and 2 above, respectively. The *Bayes factor* for M_1 vs. M_2 is

$$B_{12} = \frac{P(Y|X, M_1)}{P(Y|X, M_2)},$$

where

$$P(Y|X, M_i) = \int P(Y|X, \beta_i, M_i) P(\beta_i|M_i)\, \mathrm{d}\beta_i$$

is the probability of the data n-vector Y averaged over the prior distribution of the regression vector β_i. One way of comparing models is through Bayes factors. In fact under a wide set of model assumptions and prior assumptions, as n becomes large, it may be shown that

$$2 \ln B_{12} \approx \Lambda - (p_1 - p_2) \ln(n).$$

Here Λ is the generalized log likelihood ratio, and the relative error of the approximation implied for B_{12} is $O(1)$; see [13]. Large values of B_{12} support model 1 compared with model 2, and if $p_1 > p_2$ then the log likelihood ratio is penalized by $\ln(n)$ times the difference in dimensions. This $\ln(n)$ factor, known as the Bayes information criterion (BIC) [23], contrasts with the value 2 (AIC) argued by Akaike [1], and corresponds to Mallows' C_p in the special case of least squares multiple regression. The adjusted R^2 of 4 involves a factor $(1 - p/n)^{-1} \doteq 1 + p/n$ and therefore penalizes by 1 rather than the 2 of Mallows' C_p. The F test 5% significance method described earlier produces a penalization factor of around 4. Both AIC and BIC penalize maximum likelihood for overfitting. For large n the BIC penalization is much greater and leads to smaller models being promoted. For n smaller than $e^2 = 7.4$, AIC tends to favor smaller models. The BIC procedure is consistent as $n \to \infty$, whereas AIC is not [24], but BIC is not asymptotically efficient as $p \to \infty$. Consistency requires at least a penalization factor that increases with n. A factor of $2c \ln \ln(n)$, $c > 1$, is necessary and sufficient in the special case of autoregression; see [11]. Consistency may be regarded as less important when the aim is prediction rather than estimation.

Bayes Averaging

Whether the focus be estimation or prediction, any reasonable loss function (apart from 1–0 loss) will lead to model averaging rather than selection of a single "best" model if applying Bayesian decision theory. One exception is where costs on observing variables are included, as in [17], leading to a minimum expected posterior loss submodel. For prediction of a future Y_f at x_f,

$$P(Y_f | x_f, D) = \sum \int P(Y_f | X, \beta_i, M_i) P(\beta_i, M_i | D) \, \mathrm{d}\beta_i,$$

where D represents the n observation data. When there are 2^k models and k is large, this involves summation over a very large number of models. Madigan & Raftery [18] propose an Occam window principle (*see* **Parsimony**) which reduces the summation to be over the subset of models that are more probable *a posteriori*. Interestingly, **ridge regression** may be viewed as a particular weighted average of all subset models; see Leamer & Chamberlain [16].

Other Prior Distributions

A class of **prior distributions** with equiprobability contours proportional to

$$\prod_i (\gamma + |\beta_i|^\delta)$$

has been shown [15] to generate a range of densities which generate modal models favoring subsets. The idea seems to have been rediscovered by Frank & Friedman [8] to lend insight into a variety of chemometric regression tools, and taken up in a particular instance by the lasso of Tibshirani [26], corresponding to a double exponential prior distribution, with $\delta = 1$, $\gamma = 0$.

An approach gaining sway and more in keeping with model averaging is based on mixture models. Each regression parameter β_i is thought to come from one of two distributions, each one centered on zero, but one having a much smaller variance than the other, and may even be zero corresponding to a spike of probability at zero. There is an indicator random variable γ_i which determines whether the parameter has large variance ($\gamma_i = 1$) or small variance ($\gamma_i = 0$). The posterior distribution of the k-vector γ gives all the information about probable subset models. With a natural conjugate prior distribution and Gaussian errors, direct computation is feasible provided 2^k is not too large, say with $k < 20$. Otherwise Markov chain Monte Carlo (MCMC) allows one to summarize the posterior distribution by simulation; see [10].

References

[1] Akaike, H. (1974). A new look at the statistical identification model, *IEEE Transactions on Automatic Control* **19**, 716–723.

[2] Allen, D.M. (1971). Mean square error of prediction as a criterion for selecting variables, *Technometrics* **13**, 469–475.

[3] Breiman, L. & Spector, P. (1992). Submodel selection and evaluation in regression. The x-random case, *International Statistical Review* **60**, 291–319.

[4] Brown, P.J. (1993). *Measurement, Regression, and Calibration*. Clarendon Press, Oxford.

[5] Cox, D.R. & Snell, E.J. (1974). The choice of variables in observational studies, *Applied Statistics* **23**, 51–59.

[6] Dempster, A.P. (1973). Alternatives to least squares in multiple regression, in *Multivariate Statistical Inference*, D.G. Kabe & R.P. Gupta, eds. American Elsevier, New York, pp. 25–40.

[7] Firth, D. (1991). Generalized linear models, in *Statistical Theory and Modelling: In Honour of Sir David Cox*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, pp. 55–82.

[8] Frank, I.E. & Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics* **35**, 109–147.

[9] Furnival, G.M. & Wilson, R.W. Jr (1974). Regressions by leaps and bounds, *Technometrics* **16**, 499–511.

[10] George, E.I. & McCulloch, R.E. (1997). Approaches for Bayesian variable selection, *Statistica Sinica* **7**, 339–373.

[11] Hannan, E.J. & Quinn, B.G. (1979). The determination of order of an autoregression, *Journal of the Royal Statistical Society, Series B* **41**, 190–195.

[12] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[13] Kass, R.E. & Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**, 773–795.

[14] Lawless, J.F. & Singhal, K. (1978). Efficient screening of nonnormal regression models, *Biometrics* **34**, 318–327.

[15] Leamer, E.E. (1978). Regression selection strategies and revealed priors, *Journal of the American Statistical Association* **73**, 580–587.

[16] Leamer, E.E. & Chamberlain, G. (1976). A Bayesian interpretation of pretesting, *Journal of the Royal Statistical Society, Series B* **38**, 85–94.

[17] Lindley, D.V. (1968). The choice of variables in multiple regression (with discussion), *Journal of the Royal Statistical Society, Series B* **30**, 31–66.

[18] Madigan, D. & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1535–1546.

[19] Mallows, C.L. (1973). Some comments on C_p, *Technometrics* **15**, 661–675.

[20] Miller, A. (1990). *Subset Selection in Regression.* Chapman & Hall, London.

[21] Nordberg, L. (1982). On variable selection in generalized linear and related regression models, *Communications in Statistics – Theory and Methods* **11**, 2427–2449.

[22] Raab, G.M. (1994). Selecting confounders from covariates, *Journal of the Royal Statistical Society, Series A* **157**, 271–283.

[23] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

[24] Shibata, R. (1976). The selection of order of an autoregressive model by Akaike information criterion, *Biometrika* **63**, 117–126.

[25] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B* **39**, 44–47.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the lasoo, *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

(*See also* **Shrinkage**)

P.J. BROWN

Variance Component Analysis

Fisher [7] introduced the concept of environmental and **genetic correlations and covariances** in 1918. He assumed that a trait could be influenced by unmeasured genetic factors transmitted from parent to offspring in accordance with Mendelian inheritance (*see* **Mendel's Laws**), and showed that under certain conditions these factors would result in *genetic components of variation* that would make a stable contribution to population **variance**. It is also possible to develop models for *environmental components of variation* to represent the effects on trait covariation of sharing nongenetic factors.

These variance component models can be fitted to data collected on sets of individuals, generally referred to as pedigrees. Each pedigree could consist of a single individual on its own, a twin pair, a nuclear family or a multigenerational kinship. Pooled pedigrees need not all be of the same size and structure; see, for example, [9]. Although some methods for fitting variance components may be applicable to pedigrees that are "regular", in that they are all of the same size and structure (such as twin pairs), more general methods are needed to analyze data from sets of "irregular" pedigrees.

Even if a variance component model appears to give a good or parsimonious fit to the data, this does not necessarily imply that the hypothesized components correctly represent the true causes. The ability to differentiate the effects of shared (usually unmeasured) **genes** from those of (usually unmeasured) shared environment is strongly dependent on the design and the ability to model correctly the putative causes of familial associations.

A General Model

Within each pedigree, let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ be a vector of possibly dependent measures on a continuous trait. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ be the vector of conditional trait means expressed as a not necessarily linear function, f, of measured covariates specified by a set of parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots)'$. Similarly, for any two individuals i and j, let the covariance $\mathrm{cov}(Y_i, Y_j) = \boldsymbol{\Omega}_{ij}(\boldsymbol{\beta})$ be a not necessarily linear function of measured **covariates** specified by a set of parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)'$. The covariance matrix can be modeled in terms of variances and **correlations**, or the covariances themselves, or it can be modeled in terms of variance components. It can also be modeled in terms of path coefficients derived from path diagrams; although this approach is often referred to as **path analysis**, it is essentially the same as variance component analysis.

A Descriptive Model for Familial Associations

The covariance between relatives can be modeled by, for example, letting $\mathrm{cov}(Y_i, Y_j) = \sigma^2$ if $i = j$, $\rho_{\mathrm{sib}}\sigma^2$ if i and j are siblings, $\rho_{\mathrm{PO}}\sigma^2$ if i and j are parent and offspring, $\rho_{\mathrm{sp}}\sigma^2$ if i and j are a spouse pair, and so on, where ρ_{sib}, ρ_{PO}, and ρ_{sp} are the correlations between sibling, parent–offspring, and spouse pairs, respectively. Note that σ^2 can itself be modeled in terms of measured characteristics of the individuals, such as their age and sex. In this case, the expression above for the covariance must be adjusted to allow two individuals, i and j, to have different variances by replacing σ^2 by $(\sigma_i^2\sigma_j^2)^{1/2}$, and allowing σ_i^2 and σ_j^2 to depend on those variables.

A Basic Variance Component Model

A basic variance component model is represented by

$$Y_i = \mu_i + C_{\mathrm{i}} + E_{\mathrm{i}}, \tag{1}$$

where C_i and E_i are independent random variables with zero mean, and variances σ_{c}^2 and σ_{e}^2 representing factors common to a group of individuals, and factors specific to the individual, respectively. That is, for $i, j = 1, \ldots, n$, $\mathrm{cov}(C_i, C_j) = c_{ij}\sigma_{\mathrm{c}}^2$, where $c_{ii} = 1$ and $-1 \le c_{ij} \le 1$, while $\mathrm{cov}(E_i, E_j) = \sigma_{\mathrm{e}}^2$ if $i = j$, and 0 otherwise. If C and E are independent, then $\mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(C_i, C_j) + \mathrm{cov}(E_i, E_j)$, and the variance of Y is $\sigma^2 = \sigma_{\mathrm{c}}^2 + \sigma_{\mathrm{e}}^2$. This model can be extended to include multiple variance components representing independent familial factors, $C^1, C^2, \ldots$, in which case $\mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(C_i^1, C_j^1) + \mathrm{cov}(C_i^2, C_j^2) + \cdots + \mathrm{cov}(E_i, E_j)$, and the total variance of Y is $\sigma^2 = \sigma_{\mathrm{c1}}^2 + \sigma_{\mathrm{c2}}^2 + \cdots + \sigma_{\mathrm{e}}^2$.

The model can be extended by letting c_{ij} take different values depending on the characteristics of the individuals i and j, such as their relationship to

one another, or whether they actually live together and for how long, and how often they see each other. The coefficients c_{ij} can also be considered as parameters, and estimated rather than fixed a priori. They can also be estimated as a function of measured variables (see below).

The descriptive model can be represented as a variance component model as follows: let C represent factors common to, for example, siblings, so that $c_{ij} = 1$ if i and j are siblings, otherwise 0. The correlation between siblings, ρ_{sib}, is therefore σ_c^2/σ^2. The variance component E encompasses measurement error, which sets an upper limit on the ratio σ_c^2/σ^2, "the amount of variation attributed to the variance component" C.

Note that in the descriptive model above, and in this simple variance component model, the causes of familial aggregation need not be specified. The correlations between relatives, and the variance components representing factors common to relatives, could be caused by genetic and/or nongenetic factors shared by the relatives.

A More General Variance Component Model

A more general model that allows interpretation of variance components is given by

$$Y_i = \mu_i + G_i + C_i + E_i, \tag{2}$$

where G, C and E are independent with zero mean, and represent genetic factors, factors common to relatives, and factors specific to an individual (including measurement error), respectively.

The variance of G, or *genetic variance* σ_g^2, can be decomposed into σ_a^2, the additive genetic variance, representing the additive effects of alleles at a locus, and σ_d^2, the dominance genetic variance, representing the nonadditive effects of alleles at a locus [7]. This applies whether there is one, several or a multitude of loci influencing the trait. It has traditionally been used to model the effects of one or more putative genetic loci that have not been measured, and is used to make inferences about the existence and magnitude of genetic etiologies, even though the genes responsible are not identified. The model can be extended to include epistasis (*see* **Genotype**), and there are a number of ways of expressing the genetic correlations and covariances in terms of **identity coefficients**. Furthermore, the components of variance themselves can be modeled as a function of measured covariates, in particular age and sex. Variations in the genetic variance with age or, for example, with geographic location, are consistent with **gene–environment interactions**.

The dominance component is difficult to detect in the presence of additive genetic factors, even with a large number of observations on relatives (e.g. twin pairs), because in most designs the correlation between the estimate of σ_a^2 and the estimate of σ_d^2 is typically close to -1 (for an example, see the section "Statistical Power" below). Although in theory a dominance effect can occur at a single locus without there being an additive effect [in which case the heterozygote must be on a different side of the mean than the homozygote(s)], a polygenic dominance component is implausible without a polygenic additive component (*see* **Polygenic Inheritance**).

Even when there is a "purely" dominant or recessive effect (in that one homozygote and the heterozygote have the same residual value about the mean that is different to the residual value of the other homozygote) there is both an additive and a dominance variance component. That is, "dominance variance" is not the same concept as "dominant inheritance"; one refers to a variance component and the other to a pattern of (expected) trait values for given genotypes. Furthermore, estimates of the dominance component are strongly confounded with those of a common sibling environment component in most designs.

As indicated above, the *environmental variances* that represent the effects of factors common to relatives, C, can be defined in a number of ways. For example, $\text{cov}(C_i, C_j) = \sigma_c^2$ if i and j live in the same household, or 0 otherwise, represents a common household effect. An effect related to length of cohabitation and to time spent living apart might take the form

$$\text{cov}(C_i, C_j) = \begin{cases} \sigma_c^2(1 - e^{-\lambda t}), & \text{if } t \le t', \\ \sigma_c^2(1 - e^{-\lambda t'})e^{-\nu(t-t')}, & \text{if } t > t', \end{cases} \tag{3}$$

where t represents time measured from when i and j begin living together possibly up to and beyond a time t', when i and j begin to live apart. The parameters λ and ν can be allowed to vary across relationships. Some theoretical justifications have been proposed [5, 14, 19].

Example 1: Blood Lead Levels

The lead content of blood was measured in 617 individuals from 80 families of two or three generations [14, 15]. After extensive descriptive modeling, which examined the shape of correlations as a function of type of relationship and ages of pairs, the following model was fitted to the log transformed blood lead levels: $Y_i = \mu_i + G_i + C_i^1 + C_i^2 + E_i$, where G, C^1, C^2 and E are independent with zero mean, and represent additive genetic factors, factors common to siblings, factors related to cohabitation, and factors specific to an individual (including measurement error), respectively, with variance components σ_a^2, σ_{c1}^2, σ_{c2}^2 and σ_e^2, respectively. For two members i and j of the same family, $\mathrm{cov}(G_i, G_j) = 2\phi_{ij}\sigma_a^2$ (i.e. it is assumed there are no dominance effects so that $\sigma_d^2 = 0$), $\mathrm{cov}(C_i^1, C_j^1) = \sigma_s^2$ if i and j are siblings, otherwise 0, and $\mathrm{cov}(C_i^2, C_j^2)$ follows (3) with t' set at 16 years.

A parsimonious model gave the following estimates (standard errors in parentheses): $\sigma_{c1}^2 = 0.008$ (0.005), $\sigma_{c2}^2 = 0.037$ (0.009), $\sigma_e^2 = 0.042$ (0.017), with $\sigma_a^2 = 0$. [When σ_a^2 was fitted it was estimated to be 0.007 (0.008), justifying its exclusion.] Referring to (3), the estimates were: $\lambda = 0.072$ for parent–offspring pairs, ∞ for sibling pairs, and 0 for spouse pairs; and $\nu = 0.140$.

The predicted correlations between parent–offspring and between sibling pairs based on this fit are depicted in Figure 1, across the possible range of cohabitation times and times spent living apart of pairs of study subjects. For sibling pairs both under 16 years the predicted correlation was $(\sigma_{c1}^2 + \sigma_{c2}^2)/\sigma^2 = (0.008 + 0.037)/0.087 = 0.52$, where $\sigma^2 = \sigma_{c1}^2 + \sigma_{c2}^2 + \sigma_e^2$, while for older sibling pairs the effect of C^1 halved every $\log 2/\nu = 5$ years towards the asymptote $\sigma_{c1}^2/\sigma^2 = 0.09$. For parent–offspring pairs the correlation was $[1 - \exp(-0.072a)]\sigma_{c1}^2/\sigma^2$ for offspring of age a, being 0.22 for $a = 10$, and peaking at 0.30 for $a = 16$, then halving for every 5 years thereafter. When sibling correlations were estimated from pairs of specific ages they were about 0.5 for siblings aged 11–18, 0.2 for siblings both aged 30 or older, and 0.1 for siblings both over the age of 50, in general accord with the predicted values displayed in Figure 1.

The factor C can also be used to represent measured genetic factors. Suppose a genetic **marker** is measured using **DNA** from individuals i and j; an additive effect could be modeled by $\mathrm{cov}(C_i, C_j) = \sigma_c^2$ if i and j share both alleles, $\frac{1}{2}\sigma_c^2$ if i and j share one allele, 0 otherwise, so that trait similarity is equated with the number of shared alleles [14, 17, 25]. By considering sharing of alleles both within and across pedigrees, in which case alleles are said to be shared "identity-by-state", the factor C represents a random association effect. If i and j are in the same pedigree, then it may be possible to determine how many of the alleles are shared identical-by-descent (ibd), or at least the probability that 0, 1 or 2 alleles are shared ibd. With c_{ij} equal to either half the observed number of alleles shared ibd, or half the expected number based on the probabilities, the factor C represents a random linkage effect (see below).

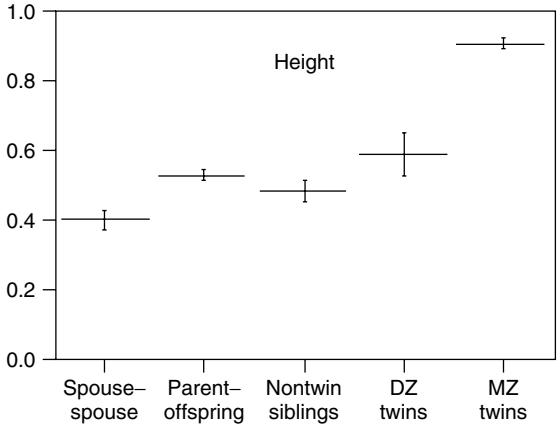

Figure 1 Predicted values of correlation in blood lead levels as a function of years, t, since cohabitation began, based on the fitted model. For sibling pairs, t is the age of the younger sibling, while for parent–offspring pairs t is the age of the offspring. It is assumed that individuals live in the same household as their parents up until the age $t' = 16$ years. Reproduced from [16] with permission from Blackwell Science

A Model for Twin and Family Data

Data from pairs of monozygous (MZ) and dizygous (DZ) twins can be used to test if there is evidence for a genetic component of variance, under the assumptions of the classic twin model (*see* **Twin Analysis**). By including data from relatives of the twins, or by supplementing the twin data with data from nontwin families, more critical assessment can be made of the underlying causes of familial aggregation. For

example, the coefficients of a common environment component can be defined by $c_{ij} = 1$ if i and j are members of the same twin pair, irrespective of zygosity (*see* **Zygosity Determination**), c_{sib} if i and j are siblings, c_{po} if i and j are a parent–offspring pair, c_{sp} if i and j are spouse of one another, etc. Unlike Fisher's model for the genetic components of variance, which are defined in terms of the identity coefficients, there is no established model for the relationship between these environmental coefficients, $c_{\text{sib}}, c_{\text{po}}, c_{\text{sp}}, \ldots$. Nevertheless, within the limitations of the design, the coefficients can be estimated [9]. This model presumes that the common environment effects are strongest within twin pairs, and the effects within other pairs of individuals are the same, or a proportion of, those effects. If, for example, the correlation within spouse pairs is more than can be explained by this parameterization of a common environment, then the model may not be realistic (e.g. there could be spouse-specific effects not shared by twins or other relatives), or there could be effects of **assortative mating**.

Example 2: Height

Height was measured in 2959 adult individuals in 783 families, including 89 MZ and 86 DZ twin pairs [9]. Figure 2 shows that the correlations for age- and sex-adjusted height. Independent of the sex of the parent or offspring, or the sexes of sibling pairs, the correlations in parent–offspring pairs were no different to the sibling correlations with all estimates in the range 0.4–0.5. The correlation in DZ pairs was 0.6 (standard error about 0.1). For MZ pairs the correlation was more than 0.9, adhered closely to the pattern anticipated under a model that attributes most familial aggregation to additive genetic factors with a small component to shared environment, and includes assortative mating (the correlation in spouse pairs was 0.4). After taking into account assortative mating [7], 55% of variance was attributed to additive genetic factors, 15% to the effects of environmental factors common to siblings, twins and parent–offspring pairs while they cohabited in the past, and the remainder to effects specific to individuals.

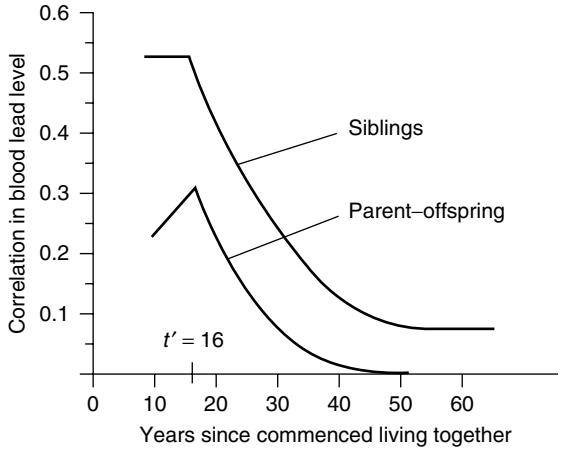

Figure 2 Correlation coefficients and their standard errors for the following pairs of family members: spouse–spouse, parent–offspring, nontwin siblings, DZ twins, MZ twins for height, Victorian Family Heart Study, 1990–1996. Reproduced from [9] by permission of Oxford University Press

Estimation, Statistical Inference and Model Fitting Under Multivariate Normality

Least squares analysis of variance (ANOVA), **maximum likelihood** and Bayesian methods have all been used to estimate variance components from samples of pedigrees [11, 12, 23, 31]. For simple balanced or "regular" designs, such as collections of twin pairs, the mean squares between and within pairs are sufficient statistics to estimate variance components. For more complex pedigree structures and unbalanced designs, ANOVA-type methods do not use the data efficiently and have unknown sampling properties, so that maximum likelihood (ML) methods are usually preferred because of the desirable asymptotic properties of estimates.

Perhaps the most flexible and practically useful approach, given the current wide availability of high-speed computers, assumes that $\mathbf{Y}$ has an n-variate normal distribution with mean $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Omega}$, and uses ML estimation to simultaneously estimate mean and (co)variance parameters [23, 31]. Within a pedigree, the log likelihood (LL) of the observed values $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is, to a constant,

$$\text{LL} = -\tfrac{1}{2}\log|\boldsymbol{\Omega}| - \tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu}), \qquad (4)$$

where $\boldsymbol{\mu}$ is a function of fixed effects parameters $\boldsymbol{\alpha}$ and the covariance matrix $\boldsymbol{\Omega}$ defined by a set of

parameters $\boldsymbol{\beta}$. For a sample of k independent pedigrees, not necessarily of the same size and structure, asymptotically **unbiased** parameter estimates can be obtained by maximizing the sum of LLs over all pedigrees. Asymptotic standard errors can be calculated from the inverse of the observed **information matrix**. The (Fisher) information matrix contains minus the second differentials of LL with respect to the parameter estimates. A choice between nested models, the selection of a parsimonious model, and hypothesis testing can be carried out using the **likelihood ratio** criterion.

Typically the size of pedigrees is of the order of 2 to 20, and small relative to the total number of pedigrees, which is often in excess of 100. In such instances the negative bias in ML estimates of variance components is small, even when within pedigree correlations are large.

The ML estimation of variance components in effect assumes that the fixed effects are known without error, which leads to biased estimates of the variance components. [In the simplest case of $Y_i = \mu_i + E_i$ and n observations, the ML estimate of σ_e^2 is $\sum(y - \overline{y})^2/n$, which is biased by a factor of $n/(n-1)$.] When the number of observations is large relative to the number of fixed effects or covariates to be estimated, this bias is small. In residual (or restricted) maximum likelihood (REML) [27], only the part of the likelihood that is independent of fixed effects is maximized, by taking into account the loss in degrees of freedom by estimating fixed effects. In balanced designs, REML estimates are identical to ANOVA estimates of variance components. For the analysis of samples from human populations, the number of covariates is usually small relative to the number of observations, so that the use of either ML or REML is likely to lead to the same statistical inference.

Transformation of Data

In order to apply the above methodology, the distribution of the residuals about the mean $\boldsymbol{\mu}$ must be approximate multivariate normal. If this condition is not satisfied, then it may be after scale **transformation** and modeling of the mean, for example using a **power transformation** [3]. Transformation may have other desirable properties, such as stabilizing the variance.

Tests of Fit and Detection of Outliers

Under a fitted model, the observed trait of an individual can be compared with its expected distribution independent of, or conditional on, all or a subset of the observed traits of other individuals in the pedigree [14, 16]. The conditional residuals can be orthogonally transformed to approximately independent univariate normal variates. These can reveal potential outliers, and such individuals may have trait values that are typical for the population but atypical given the trait values of their relatives and the patterns of within-family associations evident in the pooled data. **Goodness of fit** can be assessed by comparing the overall distribution of these residual variates with the standard normal distribution, and by examining a plot of expected vs. observed normal order statistics.

Outlying pedigrees may be identified by noting that, after replacing $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ by their estimated values, the observed quadratic form for a pedigree, $Q = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, has an approximate χ^2 distribution with n degrees of freedom. For each pedigree, $P = P(\chi_n^2 > Q)$ should have a uniform distribution on [0, 1]. An excess of small values for P can be detected by counting the number of values less than a particular cutpoint, e.g. 0.1, 0.05 or 0.01, and comparing with the binomial distribution defined by the number of pedigrees and the cutpoint [2]. The detection of outliers is important not only because of the implications for valid statistical inference, but also because of the potential biologic insights that can follow, e.g. evidence for a genetic locus with a major influence.

Robustness of Statistical Inference

It is well known that outliers can be masked when estimation is based on normal theory, under which standard tests on means are influenced by skewness while tests on variances are influenced by **kurtosis**. Departures from normal theory influence statistical inference on the genetic regressions between relatives [4]. Under non-normality, ML estimates of components of variance and correlation are robust in the sense of being not greatly biased, but are inefficient [29]. Standard errors are under- or over-estimated when there is positive or negative kurtosis, respectively [30].

One robust approach to calculating standard errors for variance components is based on the observed

covariance matrix of the score vectors [1]. Although the likelihood ratio test is not robust, the score test may be modified using a consistent estimate of the variance to allow hypotheses regarding specific components to be tested without relying directly on the assumption of multivariate normality.

Other robust procedures exist. One involves downweighting observations with large standardized residuals [18]. Another involves replacing the **multivariate normal distribution** by a multivariate t distribution [22]. In theory this approach can be extended to other non-normal distributions, including nonsymmetric ones.

Adjustment for Ascertainment

If pedigrees are sampled (ascertained) through one or more individuals (probands) with particular features, e.g. having an extreme value on the trait, then the data are no longer a random sample (*see* **Ascertainment**). Suitable adjustments must be made to avoid major biases in estimates. One approach is to maximize the LL conditional on the observed trait value of proband p, y_p, $\mathrm{LL_c} = \mathrm{LL} - \mathrm{LL}_p$, where LL is given by (4), μ_p is the expected mean, and $\mathrm{LL}_p = -\frac{1}{2}\log|\Omega| - \frac{1}{2}(y_p - \mu_p)^2/\sigma^2$. However, if it is known that probands have trait values above a given threshold, or have been selected by a specific criteria, then the conditional LL should reflect that information. This may be problematic in practice, however, because it would necessitate computation of an area under the multivariate normal density. Conditioning on the observed value of a proband can lead to biased estimates if probands are selected under specific criteria [2, 6].

Estimability

An important issue in the estimation of multiple (co)variance parameters is whether there is sufficient information to estimate each parameter. Although models containing a number of components can be specified, the components may not be identifiable from a given data set. For example, from twin data alone it is not possible to estimate uniquely σ_a^2, σ_d^2, σ_c^2 and σ_e^2, where σ_c^2 represents the variance due to common environmental effects shared by twins independent of their zygosity, under the classic twin model. Often lack of identifiability is obvious from the design, or becomes evident during model-fitting computations due to matrix singularities, or by correlations between estimates being close to -1. Problems arise, however, when researchers ignore realistic components, or model them simplistically or poorly.

Perhaps the most important issue is the confounding between the additive genetic factor and a shared environment factor. For both these factors, theoretical considerations and common sense predict that the associated correlation between relatives will decrease the weaker is their relationship to one another. Often model fits attribute this pattern of correlations in pedigree data to a genetic, rather than a shared environment, component. This could be a consequence of the genetic model predicting a more detailed pattern of correlations within a pedigree than that predicted by the usual simplistic model of the shared features of the environment (such as a dichotomous "common family" or "common household" effect, presumed to be independent of the age, sex, cohabitational status, etc. of pedigree members). The classic twin model, where any increase in the MZ pair correlation over the DZ same-sex pair correlation can only be attributed to genetic effects, is an extreme case of this bias. There are a number of other reasons why the typical modeling paradigm used historically to fit genetic and environmental components of variance to twin data tends to conclude that familial aggregation is due to genetic factors, and not to the effect of shared environment [13].

Note that if it is presumed that common environment effects do not exist, then the standard error of the genetic component(s) of variance will be greatly reduced from what they would be if it was assumed that common environment effects do exist, even though they may not be nominally statistically significant. For example, consider the classic twin model and suppose there exists a common environment effect even though it may not be highly likely to go undetected given the sample size. When fitting both an additive genetic component and a common environment component, the estimate of σ_a^2 is based on twice the difference in covariance between MZ and DZ pairs. Its standard error must be greater than the sum of the standard errors of the MZ and DZ covariance estimates. If, instead, it is assumed that there is no common environment effect, then the estimate of σ_a^2 is based on a weighted pooling of the MZ and DZ covariance estimates, and will therefore have a considerably smaller standard error.

For human traits for which it is known that environment or lifestyle factors influence the mean values, and that these factors are themselves familial (i.e. correlated in individuals within families), it would be hard to argue that there are definitely no effects of common environment on variation in trait values across the population. Furthermore, real deviations from the assumption of the effect of common environment is independent of zygosity (in that MZ pairs actually share those effects more strongly than DZ pairs) will result in overestimating the additive genetic component, and consequently underestimating the common environment component and making it more likely to be "not significant". Therefore it would seem prudent always to quote a confidence interval for the genetic component, or for the **heritability**, by reference to the fits of a variance components model that included a common environment effect, even if the estimate of that variance component were negative.

A negative estimate of a variance component designed to represent factors causing similarities between individuals may occur if, in reality, there are factors causing dissimilarities between individuals. The existence of such "competition" effects has been recognized in behavioral studies, where for example outgoing or extroverted behavior in an individual may induce introverted behavior in a close relative. Note, however, that many software packages that estimate (co)variance components using ML have inbuilt constraints that allow only positive-definite covariance matrices. With these packages one would not be able to observe a negative estimate of a variance component.

Analysis of Multivariate Traits

The likelihood framework is easily extended to multiple traits, allowing the estimation of genetic and environmental correlations and covariances between traits [20]. The main difficulty regarding multitrait analysis is computational, because many parameters are estimated and they may be strongly negatively correlated. For an analysis of k traits and I random effects per trait, a total of $l * k(k+1)/2$ parameters are estimated.

Another approach to considering the genetic and environmental links between multiple traits is to conduct a univariate analysis of a "primary" trait, and to fit the effect of one or more "secondary" traits as fixed effects on the mean of the primary trait. As each secondary trait is entered into the equation representing that mean, the residual variance will decrease. The extent to which the genetic or environmental components of variance decrease reflects the pathways through which the secondary trait(s) influence the primary trait. The approach of adjusting the mean of a trait for different factors and observing the relative reductions in variance components can also be applied to multivariate analyses.

Example 3: Bone Density and Lean Mass

Hip bone density and lean mass were measured in 56 MZ and 56 DZ female twin pairs [28]. After adjusting mean bone density for age, the correlation (standard error in parentheses) was 0.62 (0.08) in MZ pairs and 0.33 (0.11) in DZ pairs. When a variance component model was fitted the estimates were 109 for σ_a^2 and 65 for σ_e^2 (for this exercise, σ_c^2 was set at zero). After adjusting mean lean mass for age, the correlations were 0.87 (0.03) and 0.30 (0.11) for MZ and DZ pairs, respectively, and the same variance component modeling gave estimates of 17.4 and 2.5 for σ_a^2 and σ_e^2, respectively. After also adjusting mean lean mass for height, the variance component estimates became 8.6 and 2.2 for σ_a^2 and σ_e^2, respectively, indicating that about half the genetic variation in lean mass for age was explained by the association between lean mass and height.

The (cross-trait) correlation between age-adjusted hip bone density and age-adjusted lean mass in the same individual was 0.43 (0.06). The (cross-trait cross-twin) correlation between these two measures in different members of a twin pair was 0.31 (0.07) in MZ pairs and 0.09 (0.09) in DZ pairs (one-sided P value $= 0.05$). This is consistent with about 75% of the covariance between the two traits being attributable to genetic factors that influence variation of both traits.

After adjusting the means of both traits for height as well as age, the cross-trait correlation reduced to 0.26 (0.07), suggesting that about 40% of the original within-person association between the two traits was explained by their height being associated with both traits. The cross-trait cross-twin correlations, however, were no longer different between MZ and DZ pairs, becoming 0.16 (0.08) and 0.13 (0.09), respectively. This suggests that, after also adjusting for

height, genetic factors no longer explained the residual correlation between hip bone density and lean mass. Furthermore, the genetic factors that explained the original association between the two age-adjusted traits must be associated with height.

Variance Component Linkage Analysis

Fisher introduced the concept of variance, and its partitioning into causal components, partly to separate unmeasured genetic from unmeasured nongenetic sources of variance [7]. With the advent of molecular genetics, it is now possible to decompose the genetic variance further, in contributions from individual loci [quantitative trait loci (QTL)]. The above theory and estimation procedures readily lend themselves for this extension. The essence of detection of a QTL is that *observed* proportions of alleles shared ibd at genetic marker loci are used instead of (co)variances among relatives based upon the *expected* proportion of alleles shared ibd. This allows the estimation of within-family genetic variance and therefore the partitioning of genetic variance into components due to individual trait loci. One way has already been described above as the last example of a More General Variance Component Model.

Haseman & Elston [10] proposed a simple least-squares method to detect linkage between a QTL and a marker locus for collections of sibling pairs, essentially by estimating a variance component associated with a marker locus. For each pair of siblings i and j, let $Z_i = Y_i - \mu_i$ and $Z_j = Y_j - \mu_j$, and use all pairs to fit the linear regression model $(Z_i - Z_j)^2 = a + b\pi_{ij}$, where π_{ij} is the proportion of alleles shared ibd between individuals i and j, at a marker locus. For a fully informative marker, π takes the values 0, $\frac{1}{2}$ or 1. The expected value of the regression coefficient is $b = -2(1-2\theta)^2\sigma_q^2$, where θ is the recombination fraction between the marker and QTL, and σ_q^2 the variance component due to the QTL. ML variance component methods have subsequently been proposed for linkage analysis in sib pairs and in more complex pedigrees [32]. The mixed linear model [(3)] is readily extended to incorporate random QTL effects. The difficulty in estimating QTL variance for complex pedigrees is in deriving the probabilities of sharing 0, 1 or 2 alleles ibd for all pairs of members within a pedigree from multiple marker loci, when individual marker loci are not fully informative and when marker genotypes are available on a subset of the pedigree only.

Statistical Power

The statistical power of a balanced random-effect ANOVA design to estimate variance components has been addressed [24] by assuming a general random-effect model and considering the test statistic $F = MS_x/MS_y$, where MS_i is the mean squares for stratum i. Let $n(x)$ and $n(y)$ be the degrees of freedom pertaining to each stratum. Then, to obtain a power of $(1-\beta)$, a sample size is needed such that $F_{n(x),n(y),[1-\alpha]} = [E(MS_x)/E(MS_y)]F_{n(x),n(y),[\beta]}$.

As a simple example, consider a sample of n pairs under the basic variance component model $Y_i = \mu_i + C_i + E_i$ given by (1), with $c_{ij} = 1$ for all $i, j = 1, 2$. Assume that $C_i \sim \mathrm{N}(0, \sigma_c^2)$ and $E_i \sim \mathrm{N}(0, \sigma_e^2)$, and let $c^2 = \sigma_c^2/(\sigma_c^2 + \sigma_e^2)$. The test statistic $F = MS_C/MS_E$ is distributed as $F \sim [1 + 2c^2/(1-c^2)]F_{n-1,n}$, and the power of the test is $\Pr[F_{(n-1),n} > (F_{n-1,n,[1-\alpha]})/(1 + 2c^2/(1-c^2))]$.

Another approach is to consider the ML estimate of the correlation between two traits, Pearson's product moment correlation coefficient, r. When estimated from n pairs, r estimates c^2 and has an asymptotic variance of $(1-r^2)^2/(n-1)$. The sampling distribution tends to normality slowly, so that even if the two traits follow a bivariate normal distribution the distribution of r is not normal, especially if the true value correlation is close to 1 or -1 when it has substantial **skewness**. The Fisher z-transform, $z = \frac{1}{2}\log[(1+r)/(1-r)]$, has an approximate variance $1/(n-3)$ and a distribution that tends to normality rapidly as the sample size increases, irrespective of the true value of r. Provided the true values of r is between -0.4 and 0.4, there is little difference between r and z (i.e. if $r = 0.4$, $z = 0.42$), but as r increases in absolute value, the difference increases rapidly (i.e. if $r = 0.6$, $z = 0.69$, and if $r = 0.8$, $z = 1.10$). For 80% power at the 0.05 level of significance (one-sided) the true value must be 2.45 times the standard error of the test statistic. Therefore, at least for true correlations or values of c^2 in the range of -0.4 to 0.4, a sample size of about $n = (2.45/c^2)^2 + 3$ is needed (compare with Table 1).

That is, the number of pairs required is small if the correlation between members of the pairs is large, or equivalently if the variance component representing

Table 1 Number of pairs, n, needed to detect a variance component that explains the proportion c^2 or more of total variance, or a correlation between members of the pair of r, with 80% or more power at the 0.05 level of significance (one-sided)

c^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
n	617	152	66	36	22	15	10	7	5

effects common to the pair is large relative to the overall variance. When the correlation is small, which may particularly apply to effects of a QTL, the sample size to detect even a substantial variance component can become very large.

For the classic twin model, let the sample size be sufficiently large that the total variance, σ^2, is estimated with negligible error. Let A and C be the estimates of $a^2 = \sigma_a^2/\sigma^2$ and $c^2 = \sigma_c^2/\sigma^2$, respectively so that $A = 2(r_{MZ} - r_{DZ})$ and $C = 2r_{DZ} - r_{MZ}$ [and the correlation between A and C is $-(0.9)^{0.5} = -0.95$]. Given that the twin pairs are independent, however, the standard errors of A and C are $2[\mathrm{var}(r_{MZ}) + \mathrm{var}(r_{DZ})]^{0.5}$ and $[4\mathrm{var}(r_{DZ}) + \mathrm{var}(r_{MZ})]^{0.5}$, respectively. (Note that most of the variance of C comes from the variance of r_{DZ}, which can only be reduced by increasing the number of DZ pairs.) Let the correlations within both MZ and DZ pairs be < 0.4, say, and let there be the same number, n, of MZ pairs as there are DZ pairs, so their sampling variances are each approximated by $1/(n-3)$. The standard error of A is then about $2.83(n-3)^{0.5}$, and the standard error of C is about $2.24(n-3)^{0.5}$. Therefore, if c^2 is zero or small, one would need more than 4800, 1200, 500 or 300 pairs of each zygosity to detect values of $a^2 = 0.1$, 0.2, 0.3 and 0.4, respectively. Similarly, if a^2 is zero or small, one would need more than 3000, 750, 330 or 200 pairs of each zygosity to detect values of $c^2 = 0.1$, 0.2, 0.3 and 0.4, respectively. Conversely, if $n = 1000$, say, and the true situation is that $a^2 = 0.3$ and $c^2 = 0.2$, then the standard error of A will be about 0.09 and the standard error of C about 0.07, so there would be about 95% power of detecting the additive effect, and about 90% power of detecting the common environment effect. If $n = 100$, the standard errors would be 0.29 and 0.25, respectively, and the powers would be reduced to about 25% and 20%, respectively. That is, failure to detect evidence for a genetic effect, or for a common environment effect, must be interpreted carefully after consideration of sample sizes. For a further discussion of power in twin studies, see [26].

Statistical Software Packages

There are a great many statistical packages capable of being used for variance component analysis in general. For specific applications to pedigree data, the multivariate normal model can be fitted by ML using the program FISHER [21]. REML can be fitted using ASREML [8]. Mx is specifically written for structural equation modeling of pedigree data, in particular for regular designs [26]; see [`http://views.vcu.edu/mx`].

References

[1] Beaty, T.H. & Liang, K.-Y. (1987). Robust inference for variance component models in families ascertained through probands: I. Conditioning on probands' phenotype, *Genetic Epidemiology* **4**, 203–210.

[2] Boehnke, M. & Lange, K. (1984). certainment and goodness of fit of variance component models for pedigree data, in *Genetic Epidemiology of Coronary Heart Disease: Past, Present, and Future*, D.C. Rao et al., eds. Liss, New York, pp. 173–192.

[3] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* **26**, 211–250.

[4] Bulmer, M.G. (1985). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

[5] Eaves, L.J., Long, J. & Heath, A.C. (1986). A theory of developmental changes to quantitative phenotypes applied to cognitive development, *Behavior Genetics* **16**, 143–162.

[6] Ewens, W.J. (1988). Problems in statistical modelling in human genetics, *Australian Journal of Statistics* **30A**, 100–106.

[7] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[8] Gilmour, A.R., Thompson, R. & Cullis, B.R. (1995). Average information REML, an efficient algorithm for variance parameter estimation in linear mixed models, *Biometrics* **51**, 1440–1450.

[9] Harrap, S.B., Stebbing, M., Hopper, J.L., Hoang, H.N. & Giles, G.G. (2000). Familial patterns of variation for cardiovascular risk factors in adults: the Victorian Family Heart Study, *American Journal of Epidemiology* **152**, 704–715.

[10] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.

[11] Henderson, C.R. (1953). Estimation of variance and covariance components, *Biometrics* **9**, 226–252.

[12] Hopper, J.L. (1993). Variance components for statistical genetics: applications in medical research to characteristics related to human disease and health, *Statistical Methods in Medical Research* **2**, 199–223.

[13] Hopper, J.L. (1999). why "common environment effects" are so uncommon in the literature, in *Advances in Twin and Sib-Pair Analysis*, T. Spector, H. Sneider & A. MacGregor, eds. Greenwich Medical Media, London, pp. 151–165.

[14] Hopper, J.L. & Mathews, J.D. (1982). Extensions to multivariate normal models for pedigree analysis, *Annals of Human Genetics* **46**, 373–383.

[15] Hopper, J.L. & Mathews, J.D. (1983). Extensions to multivariate normal models for pedigree analysis. II. Modeling the effect of shared environment in the analysis of blood lead levels, *American Journal of Epidemiology* **117**, 344–355.

[16] Hopper, J.L. & Mathews, J.D. (1994). A multivariate normal model for pedigree and longitudinal data and the software "FISHER", *Australian Journal of Statistics* **36**, 153–176.

[17] Hopper, J.L., Tait, B.D., Propert, D.N. & Mathews, J.D. (1982). Genetic analysis of systolic blood pressure in Melbourne families, *Clinical and Experimental Pharmacology and Physiology* **9**, 247–252.

[18] Huggins, R.M. (1993). On the robust analysis of variance components models for pedigree data, *Australian Journal of Statistics* **35**, 43–57.

[19] Lange, K. (1986). Cohabitation, convergence, and environmental covariances, *American Journal of Medical Genetics* **24**, 483–491.

[20] Lange, K. & Boehnke, M. (1983). Extensions to pedigree analysis: IV. Covariance components models for multivariate traits, *American Journal of Medical Genetics* **14**, 513–524.

[21] Lange, K., Boehnke, M. & Weeks, D. (1987). *ograms for pedigree analysis*, Department of Biomathematics, UCLA, Los Angeles.

[22] Lange, K.L., Little, R.J.A. & Taylor, J.M.G. (1989). Robust statistical modeling using the *t* distribution, *Journal of the American Statistical Association* **84**, 881–896.

[23] Lange, K., Westlake, J. & Spence, M.A. (1976). Extensions to pedigree analysis. III. Variance components by the scoring method, *Annals of Human Genetics* **39**, 485–491.

[24] Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, Appendix 5.

[25] Martin, N.G., Clark, P., Ofule, A.F., Eaves, L.J., Corey, L.A. & Nance, W.E. (1987). Does the PI polymorphism alone control alpha-1-antitrypsin expression?, *American Journal of Human Genetics* **40**, 267–277.

[26] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer, London.

[27] Patterson, H.D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal, *Biometrika* **58**, 545–554.

[28] Seeman, E., Hopper, J.L., Young, N.R., Formica, C., Goss, P. & Tsalamandris, C. (1996). Do genetic factors explain associations between muscle strength, lean mass, and bone density? A twin study, *American Journal of Physiology* **270**, E320–E327.

[29] Smith, C.A.B. (1980). Estimating genetic correlations, *Annals of Human Genetics* **43**, 265–284.

[30] Tan, B.Y. (1987). Statistical properties and applications of the multivariate normal model for pedigree analysis. M.Sc. thesis, The University of Melbourne, Australia.

[31] Thompson, R. (1977). The estimation of heritability with unbalanced data. II. Data available on more than two generations, *Biometrics* **33**, 496–504.

[32] Visscher, P.M. & Hopper, J.L. (2002). Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data, *Annals of Human Genetics*, in press.

(*See also* **Polygenic Inheritance**)

JOHN L. HOPPER & PETER M. VISSCHER

Variance Components

In a **simple random sample**, one observation is made on each of a number of separate individuals and the variation is assumed to be represented by independent and identically distributed **random variables**, one for each individual. This forms the basis of **regression** and other models widely used in biostatistics. However, there are two ways in which the assumption of a single random component corresponding to each individual might fail to be adequate. In the first, the random variation may have a more complex structure arising from several identifiable sources. The variation is then considered to have multiple components, which we call *components of variance*. This is the classical field of variance components and has a long history dating from the nineteenth century. The second way in which the assumption can fail is when the parameters describing the systematic part of the variation may themselves change randomly, for example, between individuals or groups of individuals. This forms the basis of **hierarchical** or **multilevel modeling** in which the emphasis is on **computer intensive** methods for handling unbalanced or nonnormal data.

We begin this article with three examples to illustrate the key concepts and objectives involved in variance component analysis. Example 1 presents the simplest situation of the balanced one-way model. Example 2 describes a more complex model for microarray data, which involves *nesting* and *cross-classification* and helps distinguish these features. Examples 1 and 2 are classical variance component models. Example 3 outlines a linear **random effects** regression for a marker of HIV/**AIDS** disease and is an example of a multilevel model.

Example 1 *One-way balanced model*. Consider a group of patients, each of whom has a 'true' value of cholesterol say, or blood pressure, denoted by μ_j, $j = 1, \ldots, n_J$. For each patient, one measurement is made by a conditionally unbiased method; this means that for a given patient, μ_j has corresponding observation $Y_j = \mu_j + \varepsilon_j$, where the random term ε_j has mean zero and variance σ_ε^2. We call σ_ε^2 the *component of variance within patients*, which usually represents sampling or measurement error or some such.

Suppose now that the n_J patients are to be regarded as a random sample from a hypothetical infinite population of patients of true mean μ. This situation could arise, for example, in a **clinical trial** in which a homogeneous group of patients has been **randomized** to a treatment and interest centers on the efficacy of that treatment. The mean for patient j becomes a random variable, which can be written as the sum of the overall population mean, μ, and an independent random contribution from the patient, ξ_j. This gives $Y_j = \mu + \xi_j + \varepsilon_j$, where ξ_j has mean zero and variance σ_ξ^2. The latter is called the *component of variance between patients*. It follows that the variance of Y is the sum of two components, $\sigma_\xi^2 + \sigma_\varepsilon^2$, which are not separately estimable without either an external estimate of σ_ξ^2 from other studies or repeated measurements on each patient.

Suppose that several measurements are made on each patient for whom the response is assumed to remain stable. This gives observations

$$Y_j = \mu + \xi_j + \varepsilon_{js} \tag{1}$$

in which n_S, $s = 1, \ldots, n_S$, repeat observations are nested within patients. This means that observation 1 on patient i is assumed to have no special connection with observation 1 on a different patient k, and so on. The simplest situation assumes that all the random variables ξ and ε are mutually uncorrelated, but such an assumption should not be made uncritically. For example, errors would be uncorrelated if repeated samples were taken from a patient, homogenized, then split into n_S subsamples. Many such considerations relate to the design of the investigation.

This is the balanced one-way model in which there are two components of variance, between-patients and within-patients, each with zero mean. The random variables are usually, although by no means necessarily, assumed independently normally distributed. Repeat observations for a randomly chosen patient are correlated in the one-way model with *intraclass* **correlation** *coefficient* $\rho = \sigma_\xi^2/(\sigma_\xi^2 + \sigma_\varepsilon^2)$. This is a dimensionless measure and such measures are in general useful for formal inference, such as in genetics, but the variance components themselves are more informative as a basis for comparing the spread between and within patients.

Some statisticians prefer to represent variance component models via **covariance matrices** rather than random variables. The covariance matrix of the full $n_J n_S \times 1$ random vector formed by stacking the

rows of $\{Y_{js}\}$ into a single column is a block diagonal matrix of the form

$$(\tau_\xi J_{n_S} + \tau_\varepsilon I_{n_S}) \otimes I_{n_J} = \tau_\xi U_\xi + \tau_\varepsilon U_\varepsilon, \qquad (2)$$

where $\otimes$ denotes the Kronecker product [27], and I_{n_S} and J_{n_S} are the $n_S \times n_S$ identity matrix and the matrix all of whose elements are one, respectively; U_ξ, U_ε are associated matrices connected with indicator matrices defining the contribution of the component random variables to the observations (*see* **Matrix Algebra**). This formulation paves the way for a very general version with each separate component of variance identified with its own associated matrix. For interpretation and inference, however, we regard the representation in terms of component random variables as primary and this is the focus of the present article.

Example 2 *A model for cDNA microarray data* (*see* **DNA Sequences**). In cDNA microarrays, known single-stranded DNA clones are robotically spotted out and fixed onto a glass microscope slide. At the same time, two mRNA samples from the cell populations to be compared are reversed transcribed into cDNA and separately labeled with dyes, usually red (Cy5) and green (Cy3). The two labeled targets are mixed together and applied to the microarray slide. During hybridization, single strands in the target solution competitively combine with their complementary base-pair nucleotide sequences spotted on the slide. The relative intensities of red and green at a spot are extracted by image processing the scanned microarray images. The motivation for the technique is that the mRNA in the original cell sample reflects which genes are being used by the cell, and that the intensity ratio at a spot is a measure of the relative abundance of that gene in the two samples. The intensity ratios are usually adjusted for background noise on the slide, normalized to remove systematic sources of variation, transformed to log base 2 to induce approximate normality and additivity of effects, and denoted by the random variable M. For a detailed description of the biological and technical background, see [29].

In a study of osteoarthritis, n bone samples from diseased patients are compared to n bone samples taken from the same site in nondiseased control cadavers. The aim of the investigation is to identify which genes are differentially expressed in the osteoarthritis and control bone samples. In a simplified situation, the patients are assumed to be homogeneous for the **risk factors** age and sex. There is no shortage of slides so each case i is hybridized with each control j, m times. Replicates are assumed to be independent. One model for the observed log intensity ratio for gene g is

$$M_{gijk} = \mu_g + \xi_{gi}^D + \varepsilon_{gi}^D - \xi_{gj}^N - \varepsilon_{gj}^N + \varepsilon_{gijk}, \qquad (3)$$

where μ_g represents the true mean difference in expression of gene g in the two samples and all the remaining terms are independent random variables with zero means. In particular, ξ_{gi}^D and ξ_{gj}^N are crossed random effects specific to the diseased and control individuals with variances $\sigma^2_{g\xi D}$ and $\sigma^2_{g\xi N}$, respectively. The random variable ε_{gi}^D is an error term with component of variance $\sigma^2_{g\varepsilon D}$ specifically associated with the ith diseased case and believed to arise from random errors accumulating through the mRNA extraction, amplification, and labeling steps prior to hybridization; $\sigma^2_{g\varepsilon N}$ is the analogous component of error for the jth control sample. Finally, ε_{gijk} is the measurement error associated with the hybridization, scanning and image processing of patient i with control j and is assumed to have variance $\sigma^2_{g\varepsilon}$ for gene g. The k replicates across slides are nested within the disease-control classification (i, j). The variance of M_{gijk} is then

$$\mathrm{var}(M_{gijk}) = \sigma^2_{g\xi D} + \sigma^2_{g\varepsilon D} + \sigma^2_{g\xi N} + \sigma^2_{g\varepsilon N} + \sigma^2_{g\varepsilon}. \qquad (4)$$

In practice, it may not be feasible to estimate the separate components of variance in the model, not least because many sources of systematic and random variation in microarray experimentation are still not well understood. In this example, it would be adequate for determining differential expression to combine the sources of error into a single variance component term corresponding to the variability between log intensity ratios across slides for gene g. This illustrates an important general point that it is often adequate to use a model in which many sources of error are combined into a single variance term.

Microarray data analysis is receiving increasing attention from statisticians. Speed and Yang [44] are among the first researchers to critically examine the assumption of independent random variables and replication in this context.

Example 3 *A random effects regression model*. Suppose that a marker of disease progression such as log viral load or CD4 cell count in individuals infected with HIV varies roughly linearly over time in each individual. An initial analysis might be reasonably based on a linear regression with time, in which each individual j has intercept and slope parameters β_0 and β_1, that is,

$$Y_{jt} = \beta_0 + \beta_1 x_t + \varepsilon_{jt}. \quad (5)$$

(*see* **Nonlinear Mixed Effects Models for Longitudinal Data**).

However, a cohort of infected individuals would be very unlikely to have the same parameters. The next step might then be to regard the intercept and slope as responses regressed on individual characteristics, or to consider models in which the parameters themselves have random structure; that is, to model the slope for individual j as $\beta_{1j} = \beta_1 + \xi_{1j}$, where β_1 is the mean slope and ξ_{1j} is a random term, and similarly for the intercept, which we write as $\beta_{0j} = \beta_0 + \xi_{0j}$. In this model, interest focuses on the magnitudes of the random variation of individual responses about their regression line, in the variation in the intercepts and slopes, as well as on explanatory determinants of the regression parameters. The random effects themselves are often assumed to be normally distributed although it may not be possible to test the assumption, and it will nearly always be essential to allow these random terms to be correlated so that σ_ξ^2 denotes the covariance matrix of (ξ_{0j}, ξ_{1j}). These ideas generalize to nonnormal response data and to binary **logistic regression** models in particular.

There are only really two key ideas involved in these examples and in variance component problems generally. The first is the distinction between nesting and cross-classification. This is a qualitative rather than a statistical issue, and is to do with the design and logical structure of the data under study and not with any probabilistic or distributional model assumptions (*see* **Experimental Design**). The second key idea is statistical: are we going to treat the levels of factors as intrinsically interesting (i.e. as **fixed effects**) or are the factors to be regarded as random variables (i.e. as **random effects**) where interest might be in their variances? For example, in genetics, an investigator may want to partition the variability into environmental versus inherited components (*see* **Twin Analysis**).

Both dichotomies are subject-matter considerations. There are some general principles, which can be helpful in deciding whether a factor should be regarded as fixed or random. If the levels of a factor are treatments, for example, different therapies for breast cancer, they would usually be treated as fixed effects. Exceptions arise, such as in a clinical trial comparing the effects of many antibiotics.

The key to variance component analysis is to build models that represent different situations and explain levels of variability that are plausible approximations of what we actually observe. The motivation may be intrinsic interest in the variance components themselves, such as in a comparison of different measuring techniques, or on estimating the precision of the mean or other model parameters. Alternatively, the motivation may be the design of further studies via a *synthesis of variance*, which we discuss below.

History

The idea of partitioning variability can be traced at least as far back as Airy's interest in errors of measurement in astronomy [1]. The more recent systematic study of splitting variation into components dates from R.A. Fisher's introduction of the **analysis of variance**; his original motivation was to improve on the intraclass correlation. There followed periods of intense activity during the last century in biometrical genetics as described by Bulmer [6] (*see* **Polygenic Inheritance**), in the analysis of variability in industrial processes dating from the 1930s work in the cotton industry by Tippett [46] and in the wool industries by Daniels [12], and on error structures especially in randomized experimental designs in the 1950s [8]. Eisenhart made explicit the distinction between fixed and random interpretations of an analysis of variance and introduced this terminology [13].

Much of the early work dealt with balanced data. Henderson, in a long series of papers starting in the 1950s, gave noniterative methods for handling unbalanced data based on equating suitable quadratic forms to their expectation [18, 19]. This more intuitive approach has now largely been replaced by **likelihood**-based methods. Hartley and Rao [17] gave a general matrix formulation and **maximum likelihood** estimation for the unbalanced linear model. The important subsequent generalization of maximum likelihood to REML (reduced, **restricted or**

residual maximum likelihood, which we discuss in the next section) for unbalanced data was developed in detail by Patterson and Thompson [30]. Searle et al. [38] provide a very detailed and systematic account of the normal theory formulations and the associated matrix algebra for balanced and unbalanced data. Rao gives a broad account of normal theory aspects too [35]. Rao and Kleffe [34] emphasize the point **estimation** of variance components using quadratic error **loss**, and we discuss this and other methods of estimation in the next section.

Variance component problems with discrete response data have a long history going back to the Lexis urn models of dispersion associated with the **binomial** distribution; see, for instance, [20] (*see* **Overdispersion**). For an early paper on the **beta-binomial distribution**, see [40]. Greenwood and Yule [16] derived the **negative binomial distribution** as a **Poisson distribution** with an additional source of variation in connection with an analysis of accidents to London bus drivers (*see* **Accident Proneness**). Anscombe [2] compared the theoretical properties of various methods of estimation of its parameters. Cox [9] proposed simple methods for variance components in multiplicative models for Poisson variables.

The literature on multilevel modeling has been steadily growing over the past decade and is now very large. See Goldstein [15] for a thorough discussion. In addition to the references already mentioned, Snijders and Bosker [41] contains important computational work and guidance for fitting random effects and other models and Verbeke and Molenberghs [48] give an extremely thorough account of **linear mixed** models. McCulloch and Searle [28] discuss **generalized, linear, and mixed models** as do Fahrmeir and Tutz [14]. Pinheiro and Bates [32] focus on nonlinear normal theory models (*see* **Nonlinear Mixed Effects Models for Longitudinal Data**).

Variance components arise implicitly or explicitly in many problems in sampling and experimental design. Important applications include industrial processes and reliability studies, genetics, animal and plant breeding, econometrics, the design and analysis of interlaboratory standardization trials, epidemiology, psychometric testing, and education. Khuri and Sahai [23] review developments in variance components analysis to the mid-1980s and include a comprehensive bibliography, and a recent issue of *Statistical Methods in Medical Research* was devoted to variance components [42].

Estimation

The most important and often most difficult issue in variance component problems is the appropriate formulation of a model, or equivalently, the formulation of an analysis of variance table. We begin with the simplest situation described in Example 1. It is well known from the analysis of variance that for balanced systems, there are parallel **orthogonal** decompositions of the data vector, of sums of squares of the components, and of the **degrees of freedom**. The observation vector is decomposed into orthogonal components as

$$Y_{js} = \overline{Y}_{..} + (\overline{Y}_{j.} - \overline{Y}_{..}) + (Y_{js} - \overline{Y}_{j.}), \qquad (6)$$

and if we write the data as one long vector, orthogonality implies that the cross-product terms on the right-hand side vanish.

It is conventional to write out the analysis of variance table for the components and this is shown in Table 1, in which MS denotes Mean Square. Roughly, the mean square measures the sum of squares per dimension for the component. The analysis of variance formulation is entirely structural and does not involve model or distributional assumptions. For interpretation, we bring in the probability model, although we still only need the theory of a simple random sample to derive the key properties, in particular, for equating mean squares to their expected values.

For the one-way balanced arrangement, the first important property is that $E(\mathrm{MS}_\varepsilon) = \sigma_\varepsilon^2$, which only concerns how repeat observations for an individual vary around the true mean for that individual. It is also straightforward to show that $E(\mathrm{MS}_\xi) =$

Table 1 Analysis of variance table for the one-way balanced variance component model

Source	SS	df	
Mean	$\Sigma_{j,s}\overline{Y}_{..}^2$	1	MS
Between individuals	$\Sigma_{j,s}(\overline{Y}_{j.} - \overline{Y}_{..})^2$	$n_J - 1$	MS_ξ
Within individuals	$\Sigma_{j,s}(Y_{js} - \overline{Y}_{j.})^2$	$n_J(n_S - 1)$	MS_ε
Total	$\Sigma_{j,s}Y_{js}^2$	$n_J n_S$	

$n_S\sigma_\xi^2 + \sigma_\varepsilon^2$, from which we deduce σ_ξ^2 is estimated via $(\mathrm{MS}_\xi - \mathrm{MS}_\varepsilon)/n_S$.

If we are interested in the overall mean μ, for instance, to compare the means in two or more groups treated in different ways, we want $E(\overline{Y}_{..}) = \mu$ and $\mathrm{var}(\overline{Y}_{..}) = \sigma_\xi^2/n_J + \sigma_\varepsilon^2/(n_J n_S)$. Hence, a pivot for the estimation of μ is

$$\frac{\overline{Y}_{..} - \mu}{\sqrt{\mathrm{MS}_\xi/(n_J n_S)}}.$$

Assuming the pivot is approximately normally distributed, we can also obtain (approximate) confidence limits for μ.

These estimates are sometimes called the least-squares–based estimators and are **unbiased** estimates of the variance components. The overall approach can be generalized to more complex situations in which estimating equations are formed by equating suitable functions of the data (here sums of squares) to their expectations under the assumed model (*see* **Estimating Functions**). Alternative (biased) estimators are given by the method of maximum likelihood and these are discussed below.

If we make the further assumption that all the random variables are independently normally distributed, several important properties follow that also extend to general balanced cases. The most important is that the two sums of squares and the sample mean are minimal sufficient statistics implying various strong optimum properties, and in particular, that as long as the model is adequate, all we need for analysis are the sums of squares and the mean. The assumption of normality should not be made uncritically however, and some effort should be expended on investigating the sensitivity of the conclusions. We discuss ways of assessing model adequacy later.

Certain exact inferential procedures for the three unknown parameters μ, σ_ξ^2, and σ_ε^2 follow from the assumption of normality. For example, a technical refinement of the pivot for μ is that it then has the **Student *t* distribution** with $n_J - 1$ degrees of freedom. However, only certain combinations of the parameters can be tackled by these procedures, which may not be of substantive interest. For example, we can obtain exact confidence limits for the ratio of variances $\sigma_\xi^2/\sigma_\varepsilon^2$, but not for σ_ξ^2 itself, which is of interest in comparing estimates from two or more similar sets of data, subject to checks of homogeneity. The safest general procedure for doing this is the use of **profile likelihood** or one of its generalizations. There are however simpler and essentially equivalent methods. For example, if T is an approximately unbiased estimate of a positive parameter θ with effective degrees of freedom d, then $\log T$ is approximately normally distributed around mean $\log\theta$ with variance $2/d$, and further issues of analysis are in a normal theory least-squares framework. See [11] for a more thorough discussion of these less standard procedures.

Example 4 *Angiogenesis microarray data*. In a collaborative study, the author has been investigating genes involved in the growth of blood vessels, a process known as angiogenesis. The ability to stimulate new blood vessel growth is a prerequisite for the expansion of a solid tumor and future anticancer treatments are postulated to involve therapy directed to both cancer cells and the expanding vascular system. COX2 (Prostaglandin endoperoxide synthase 2) is a gene known to regulate angiogenesis and cell migration, and served as a control gene in a cDNA microarray experiment comparing mRNA samples from time three hours with time zero. The microarray consisted of a subtracted library of 10 400 clones, each duplicated on the slide. The duplicate spots were printed next to each other and are therefore spatially correlated, but we will ignore this special feature of the data. Four slides were hybridized and we assume that the hybridized slides are independent.

The observed log intensity ratios for COX2 are given in Table 2, which illustrates the data structure for the simple one-way model with two replicate observations. Note that in general, the ordering of the observations within rows is arbitrary. In the notation of Table 1, $n_J = 4$ and $n_S = 2$. The appropriate analysis is based on the pivot for the mean, μ, which under the null hypothesis of no differential expression and the assumption that the log ratios are normally distributed, is t with 3 degrees of freedom.

Table 2 Log intensity ratios for a COX2 in a cDNA microarray experiment with four slides and duplicate spots within slides

Slide	Log ratios M	
1	3.5040	3.4757
2	3.7160	3.7896
3	3.6215	3.7496
4	2.9467	2.8873

Thus, $T = 3.4613/(0.3796/2) = 18.23$ on 3 degrees of freedom. The associated P value is 0.00036 with an estimated 95% confidence interval for the true mean difference in expression (2.8572, 4.0654), indicating that COX2 is significantly upregulated at three hours.

We are ignoring here issues of multiple testing, which can be important in microarray experiments when many thousands of genes are analyzed simultaneously (*see* **Multiple Comparisons**).

Negative estimates: All variances are by definition nonnegative. However, the standard least-squares estimates of the upper variance component in the one-way balanced model are based on differences of mean squares and hence may sometimes be negative. The simplest way to deal with negative estimates arising from this and similar situations is to replace them by zero. For example, we would take $\max\{(\mathrm{MS}_\xi - \mathrm{MS}_\varepsilon)/n_S, 0\}$ as an estimate of σ_ξ^2. There are two qualifications to this recommendation. Firstly, if the mean square between individuals is substantially smaller than the mean square within individuals, this indicates that the data are inconsistent with the model and may be a warning that a systematic effect has been omitted. Alternatively, it may be a warning that important correlations between the random variables have been ignored. Secondly, in an analysis that synthesizes an estimate of σ_ξ^2 from several separate sets of data, such as in a **meta-analysis** of **case–control** studies, then negative values should be retained to avoid systematic error in the pooled estimate.

The procedures described so far extend directly to more complex models provided the data are balanced. In practice, however, data are often not balanced, either by design or as a result of various forms of missingness. The concepts involved are not affected by lack of balance, but the analytical details are. In particular, the decompositions for the balanced case no longer hold and the underlying algebra is more complicated. It is not always obvious how to find the variance estimates for more complicated models and general procedures are required. One very powerful procedure is maximum likelihood for which we find algebraically, or more commonly numerically, the combination of parameter values that maximize the likelihood.

Maximum likelihood and REML: It is well known that the maximum likelihood estimate of the variance in a simple random sample is biased, having divisor n_S rather than $n_S - 1$. In more complex models, the resulting estimates of variance may be entirely unsatisfactory especially if the number of **nuisance parameters** is large, and alternative methods of estimation need to be deployed. The most widely used method and preferred basis for the formal analysis of unbalanced normal models is REML, which maximizes the likelihood of judiciously chosen parts of the data, rather than that of all the data.

REML may be formulated as follows for the one-way analysis. We may apply an **orthogonal** transformation to each individual (or sample) to replace the n_S values by the quantity $\overline{Y}_{j.}\sqrt{n_S}$ and $n_S - 1$ variables, which are independently normally distributed with zero mean and variance σ^2. The contribution of the individual to the likelihood is thus the product of two factors, one depending on μ_j and σ^2, and the other depending only on σ^2 and involving the data only via $\sum(Y_{js} - \overline{Y}_{j.})^2$. In many problems, especially when little is known initially about μ_j, the first factor contains little or no information about σ^2. Thus, for inference about σ^2, we use only the second factor. This leads to a loglikelihood based on $n_J(n_S - 1)$ observations that are independently normally distributed with mean zero and variance σ^2. The corresponding maximum likelihood estimate then has the correct divisor, which is the degrees of freedom within individuals. The same idea can be applied to the general linear mixed model with fixed and random effects.

REML has the advantage of returning the usual least-squares estimates of the variance components for balanced data. It is a particular case of the use of **marginal likelihood** and **conditional** likelihood; see [21] for a general study of both. Barndorff-Nielsen and Cox [3] show that REML is a special case of modified profile likelihood.

Alternative methods of estimation: Powerful and efficient methods for model fitting are important. Indeed, the lack of such methods for unbalanced data held the subject of variance components back until relatively recently. A disadvantage of these developments, however, is that the relationship between the data and the conclusions can be obscure, and for complicated problems, simpler methods may be useful for conceptual clarity and interpretation.

For the unbalanced one-way arrangement, the two simplest procedures are to base the estimation of the upper-level variance component σ_ξ^2 on either the unweighted sum of squares $\sum(\overline{Y}_{j.} - \overline{Y}_{..}^{(u)})^2$, where $\overline{Y}_{j.}$ is the mean of the r_j responses for individual j

and $\overline{Y}_{..}^{(u)}$ is the unweighted average of these means, or on the usual analysis of variance sum of squares $\sum r_j(\overline{Y}_{j.} - \overline{Y}_{..}^{(r)})^2$, where $\overline{Y}_{..}^{(r)}$ is the average of the $\overline{Y}_{j.}$ weighted by the group size. The idea is to decide informally whether the upper or lower component of variance is dominant and to use the unweighted or standard analysis of variance sum of squares as a basis for examining the upper-level component of variance. The same idea can be extended to general models. These simpler approaches are related to the various methods of estimation proposed by Henderson and are described in detail in [38].

An important special case is when $T_1, \ldots, T_{n_J}$ are estimates of a parameter θ obtained from independent sets of data, each with its own internal estimate of error. For example, in combining the results of a number of case–control studies, θ could be the log **odds ratio** for treatment versus control after adjustment by maximum likelihood logistic regression for imbalance with respect to **explanatory variables**, which might be different in the different studies. Note that it is not necessary that the same model is fitted to each group of data, only that the parameter θ has the same interpretation. The estimates may vary more than would be expected on the basis of internal error and it may not be feasible to explain the extra variability as systematic. In this case, we may represent the additional variability as random, and in particular, take as a reasonable approximation $T_j = \theta + \xi_j + \varepsilon_j$, where the ξ and ε are approximately normally distributed and independent. The idea is that a simple analysis helps decide whether a component of variance σ_ξ^2 is necessary, whether there are **outlying** groups, and which of the weighted or unweighted estimates of θ are likely to have high efficiency. Similar arguments apply if θ is a vector. Cox [10] outlines the approach and Cox and Solomon [11] give details of these simpler procedures in applications to nonnormal response data and random effects logistic regression.

We mention briefly another class of methods for estimating variance components known as minimum norm quadratic unbiased estimators (MINQUE) described in detail in [34]. In this and related criteria, low moment assumptions are made about the component random variables and attention focusses on quadratic point estimates that satisfy conditions such as unbiasedness and minimum variance (*see* **Minimum Variance Unbiased (MVU) Estimator**).

Synthesis of Variance

This refers to the process of putting the variance components back together with a view to determining what the variability would be in different sampling situations, or the variance that should be attached to a nonstandard type of comparison. Calculations of this sort are particularly important in designs of systems to achieve a balance between the number of groups or individuals that need to be studied and the number of replicates within each individual.

The simplest example is to estimate the variance of a mean if n_{S_1} repeat observations are to be made on each of n_{J_1} individuals. This is $(\sigma_\varepsilon^2 + n_{S_1}\sigma_\xi^2)/(n_{J_1}n_{S_1})$, which can be estimated. We may be interested to know how much better will the precision be if we take three or four repeat observations on each individual rather than one, say. If the different individuals are very different, that is, σ_ξ^2 is large, then there is little point in replicating more within individuals. But if σ_ε^2 is large relative to σ_ξ^2, there will be an $n_{J_1}n_{S_1}$ effect and increasing the number of replicates will improve the precision of the overall mean.

The estimated synthesized variance of the mean under the new design with n_{S_1} rather than n_S repeated observations within each individual is

$$\frac{1}{n_{S_1}n_{J_1}}\left(\mathrm{MS}_\varepsilon + n_{S_1}\frac{\mathrm{MS}_\xi - \mathrm{MS}_\varepsilon}{n_S}\right),$$

where the observed mean squares and degrees of freedom are those from the original data.

Example 4 revisited: *Angiogenesis microarray data*. The estimated components of variance for the gene COX2 in Example 4 are 0.3731 for the between-slide component and 0.0131 for the within-slide component. In view of the considerations outlined above, increasing the number of replicate spots within a slide would have little impact on the precision as compared with increasing the number of slides hybridized in the experiment.

Components of Covariance and Regression

In the simplest case of a number of groups with a regression of Y, say, on X within each group, work on multilevel modeling has tended to stress the effect of the correlation and additional variation

within groups on the regression coefficient of Y on X and its standard error. But if the groups represent **bivariate** populations, there are two regression coefficients, one within groups and another that would be defined by a scatterplot of the means of X and Y. In an early exposition of **analysis of covariance**, Pearson stressed the distinction between these coefficients [31].

To formulate the issues explicitly, consider the bivariate one-way balanced model in which each observation Y_{js} is a 1×2 row vector giving an $n_J n_S \times 2$ data matrix Y. The pairs of observations are

$$Y_{js} = \mu_Y + \xi_j^Y + \varepsilon_{js}^Y,$$
$$X_{js} = \mu_X + \xi_j^X + \varepsilon_{js}^X, \tag{7}$$

for which there are four variances $\sigma^2_{Y\xi}$, $\sigma^2_{X\xi}$, $\sigma^2_{Y\varepsilon}$ and $\sigma^2_{X\varepsilon}$ as well as covariances $\mathrm{cov}(\xi_j^Y, \xi_j^X)$ and $\mathrm{cov}(\varepsilon_{js}^Y, \varepsilon_{js}^X)$. In the general case with p response variables, each observed random variable is replaced by a set of p components.

As explained above, we can view the bivariate decomposition in two different ways. If Y and X are treated on an equal footing, we have two covariance matrices for the interpretation of associations at the two different levels. Sometimes it may be helpful to estimate separately the two **correlations** $\mathrm{corr}(\xi_j^Y, \xi_j^X)$ and $\mathrm{corr}(\varepsilon_j^Y, \varepsilon_j^X)$. The second possibility is that X should be considered as explanatory to the response Y. Then there are two regression coefficients of Y on X, namely, $\beta_{\xi,YX}$ and $\beta_{\varepsilon,YX}$, regression coefficients from the between- and within-group structure, respectively.

Suppose as an illustration that on a large sample of subjects of stable health and in a narrow age range, measurements are made of blood pressure, Y, and sodium (Na) intake, X. For each subject, the observations are repeated some months later. If we ignore possible time trends, we may consider a one-way analysis. The regression coefficient $\beta_{\varepsilon,YX}$ measures the mean increase in blood pressure Y when the Na intake, X, of a particular subject varies by one unit, for example, 10 mg per day. By contrast $\beta_{\xi,YX}$ is the average difference in the mean blood pressure of two different subjects whose long-run mean Na intakes differ by 10 mg per day. The naive interpretation of $\beta_{\xi,YX}$ would imply that if subjects changed their long-run mean Na intake by 10 mg per day, then there would be a corresponding change in long-run mean blood pressure as determined by $\beta_{\xi,YX}$. The naive interpretation of $\beta_{\varepsilon,YX}$ would imply that individuals increasing their Na intake by 10 mg per day would on average have an increase in blood pressure determined by the regression coefficient.

In an observational study, both interpretations involve substantial assumptions and would be quite speculative. If individuals had been randomized to Na levels on the other hand, the interpretation of the regression coefficients would be unambiguous. In the absence of randomization, however, there may be explanatory variables, observed or unobserved, and long-run features of individuals that are themselves explanatory to both Y and X. These arguments extend to more complicated structures. The difficulties of applying aggregate-level conclusions to individuals in this way is often called **ecological**.

Empirical Bayes

When a frequency probability analysis is based on empirical data with structural assumptions, for example, that certain terms in a regression are random, we call the analysis **empirical Bayes**. No special conceptual issues to do with defining a suitable **prior** probability and so forth are involved.

Classical empirical Bayes analysis proceeds as follows. Consider the univariate one-way model again where, as before, the ξ_j and the ε_{js} are independently normally distributed with zero mean. There are three unknown parameters in the model, $\theta = (\mu, \sigma_\xi^2, \sigma_\varepsilon^2)$. Suppose θ is known and that interest is in the mean of the first group, $\mu + \xi_1$; ξ_1 is an unobserved random variable, which itself partly determines the distribution of the observations. It is therefore appealing, and can be justified formally from various points of view that information about ξ_1 is best summarized by its conditional distribution given the data. This is derived by **Bayes's theorem**. We can show [11, Chapter 3] that the required conditional distribution is normal with mean of the form of an optimally weighted mean obtained from combining the information from the data $\overline{y}_{1\cdot}$ and that from the distribution of ξ_1 around μ, denoted $\tilde{\xi}_1$. In effect, the sample mean $\overline{y}_{1\cdot}$ is shrunk towards the general mean (*see* **Shrinkage**). By the same argument, the estimate of any contrast is obtained by shrinking the sample contrast towards zero (*see* **Shrinkage Estimation**).

It can be shown that if in the originating model, the random variables are not normally distributed, then the above estimates are in a sense the best linear estimates. Viewed as a point predictor of ξ_1, the quantity $\tilde{\xi}_1$ has a property summarized in the term *best linear unbiased predictor* (BLUP; see [37]).

Nonnormal Models

There are broadly parallel developments for the Poisson and binomial distributions with one extra level of variability to those discussed above for continuous random variables. An alternative approach to analysis may be to use (approximate) weighted least-squares methods on the basis of an empirical **transform** of the response variable, for example, the square root or logarithm of Poisson variables and the empirical logistic, probit, or log−log transform of binomial variables.

There may be some loss of efficiency in these approximate procedures. But a more general discussion of variance component models for **generalized linear models** and normal theory **nonlinear regression** is difficult, primarily due to the fact that formally efficient methods of estimation involve high-dimensional integration. The general form of the full likelihood is given by

$$\int \mathrm{lik}(\theta \mid \xi; y)\, \mathrm{d}F(\xi; \tau),$$

where $F(\xi; \tau)$ is the distribution function of ξ depending on parameters τ, which are typically components of variance and their generalizations. In cases without **time series** or similar structure, ξ will consist of independent components so that $F(\xi; \tau)$ factorizes into a product component by component. The integral will factorize into subintegrals but, even so, the dimension of each may be large.

In the special case in which, given the random terms ξ, the observations have an **exponential family** distribution, we obtain a **generalized linear mixed model**. In the simplest formulation, the random effects ξ_j are independently and identically normally distributed with zero mean and q-dimensional variance matrix $D(\tau)$, where τ is a vector of unknown variance components; D is often called the *dispersion matrix*. The conditional independence of the observations within an individual or cluster allows us to write the exact marginal likelihood

$$\mathrm{lik}(\beta, \tau; y) = \prod_{j=1}^{n_J} \int \prod_{s=1}^{r_j} f(y_{js} | \xi_j; \beta) g(\xi_j; \tau)\, \mathrm{d}\xi_j, \tag{8}$$

where g is the link function for the generalized linear model.

Formal inference can be based on maximum likelihood or on **Bayesian** considerations, and there are currently three ways to approach the **numerical integration** problem. The most direct and appealing method is direct or preferably adaptive quadrature. The second, applicable when the integrals can be resolved into a sequence of one-dimensional integrals, is to use an analytical approximation, usually based on a few terms of a Laplace expansion [5, 39, 43]. Such expansions are based on the idea that integrals involving an exponential of a function are dominated by behavior of that function near its maximum. This method can sometimes yield relatively simple interpretable results. Calculation of higher terms in the expansions may be feasible, especially if aided by **computerized algebra**. Higher terms are important to give at least a partial check on the adequacy of the approximations but there is often some uncertainty about the range of applicability of the approximations.

The third method is **Markov chain Monte Carlo (MCMC)**. In the Bayesian version, a **Markov chain** is defined, which has as its equilibrium distribution, the posterior distributions of interest. The chain is then simulated a very large number of times and if the realizations appear to have converged to stationarity, the frequency distribution of realized values, excluding a run-in period, is used to estimate the posterior distributions. MCMC is a powerful and general technique but there is the possibility, in theory at least, that apparent convergence to a stationary state is illusory. Some protection can be achieved by starting the **simulations** from very different initial states.

There are also at least two other approaches to these problems. Lee and Nelder [25, 26] study a notion of h-likelihood in which, in effect, realized values of individual random variables representing portions of variability are treated like unknown parameters. This is likely to be effective when there is substantial information about each such realized value. Another mode of analysis called *penalized*

quasi-likelihood concentrates on the underlying estimating equations and their justification in a broader setting than a fully parametric one [4, 5, 24, 45] (*see* **Penalized Maximum Likelihood**; **Quasi-likelihood**). Rabe-Hesketh et al. [33] provide a valuable comparison of methods for estimation in generalized linear mixed models.

In **survival** or **event history** data, a random term for each individual with an associated variance component is often called *frailty*. The terminology arises from applications in which the randomly occurring events are failures or adverse reactions of some kind. See **Frailty** for a detailed discussion of these and related models.

Model Assessment and Prediction

Although there is a very large literature on formal and informal tests of model adequacy, little of it is directly relevant to variance component models (*see* **Model Checking**). The most important type of failure of a model stems from omitting a substantial effect, for example, treating a cross-classification as if nested. This destroys the independence assumptions underlying the discussion and is likely to be detected by anomalous behavior of the mean squares, possibly leading to substantial negative estimates of variance.

Outlying observations or individuals can influence the usual quadratic estimates of variance. For example, in the one-way arrangement, an anomalous single observation has a large effect on the estimated component of variance within individuals, but relatively little effect on the estimated component between individuals. In more complex situations, the distinction between outliers at the different levels becomes harder to detect empirically. **Robust** methods provide one way of dealing with outlying observations but do not retain key parameter properties, which are central to variance component analysis, in particular, the additivity of variance as a parameter.

Other important departures from the standard formulation include nonnormality of one or more of the component random variables, or dependence between the variability within an individual and the individual mean. Mild nonnormality of the variances within or between individuals may be of relatively minor concern, but the dependence described above may lead to inappropriate predictions or supplementary analyses. Solomon and Cox [43] suggested a formal analysis in which the nonnormal variances and dependence features are separated.

Further, special topics on model criticism and improvement include the prediction of exceedances, the analysis of **panel** data, fitting more elaborate models, transformations, and study of the distributional form of the underlying random variables. Many of these methods and ideas are discussed in detail in [11]. An important general point when assessing model adequacy is that the analysis should focus on issues that are of substantive importance. For example, discriminating between heterogeneous variances versus constant variances with differing individual means is only worth attempting if the distinction can be given a physical interpretation.

Generalizations and Further Topics

There are many additional areas of current work related to variance components including the following.

Measurement error models: The main emphasis in this article has been on the estimation of variance components as parameters of intrinsic interest. One situation where the real emphasis lies elsewhere and the components of variance are of concern because they affect this primary aspect, is the effect of measurement error in explanatory variables on regression analysis (*see* **Errors in the Measurement of Covariates**). **Measurement error models** have a long history and an extensive literature; see, for example, [7, 36] for a recent application.

Design of investigations: The objective of variance components analysis is the study of patterns of variation as they exist rather than the assessment of interventions under controlled conditions, which is the purpose of formal design of experiments. However, many of the general principles of **experimental design**, and especially those common with the principles of sampling (*see* **Sample Surveys in the Health Sciences**), apply. Khuri [22] gives a systematic review and bibliography of work on design for the estimation of variance components, and Cox and Solomon [11, Chapter 3] present some new ideas.

Finite population aspects: Occasionally, the individuals are not regarded as individuals or as sampled from an infinite population but as from an existing **finite population** of known size, or in particular, as forming the whole of the population in which

variation is to be assessed. The finite population variance component is relevant only in very special situations, and in some industrial problems in particular. The importance of distinguishing between finite and infinite populations when defining variance components was first stressed by Daniels in the context of studies of variation in industrial processes [12]. Tukey [47] extended these ideas to sampling a finite population. A formulation relevant to the industrial context is outlined in [11].

Synthesis of studies: In many fields of application, the synthesis of information from several studies is crucial. Variation between studies and interactions of such variation with the treatment effects under investigation may involve representation by components of variance. In biostatistics, the term *overview* or **meta-analysis** is often used and is an integral part of evidence-based medicine. A representation in terms of random effects would only be indicated if no direct explanation of important observed variation in treatment effect is apparently available, such as nonconstancy of the treatment effect being confined to certain contrasts. The use of variance components in meta-analysis is not without controversy.

Acknowledgments

Example 2 is based on joint work in progress with G. Glonek and N. Fazzalari. Example 4 is from collaborative work with C. Hahn, J. Gamble and A. Tsykin. I am grateful to Sir David Cox, FRS, for helpful comments.

References

[1] Airy, G.B. (1861). *On the Algebraic and Numerical Theory of Errors of Observation and the Combination of Observations*. McMillan, London.

[2] Anscombe, F.J. (1950). Sampling theory of the negative binomial and log series distributions, *Biometrika* **37**, 358–382.

[3] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[4] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of American Statistical Association* **88**, 9–25.

[5] Breslow, N.E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**, 81–91.

[6] Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

[7] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

[8] Cornfield, J. & Tukey, J.W. (1956). Average values of mean squares in factorials, *Annals of Mathematical Statistics* **27**, 907–949.

[9] Cox, D.R. (1955). Some statistical methods connected with series of events (with discussion), *Journal of the Royal Statistical Society B* **17**, 129–164.

[10] Cox, D.R. (1998). Components of variance: a miscellany, *Statistical Methods in Medical Research* **7**, 3–12.

[11] Cox, D.R. & Solomon, P.J. (2002). *Components of Variance*. Chapman & Hall/CRC, Boca Raton.

[12] Daniels, H.E. (1939). The estimation of components of variance, *Supplement of Journal of Royal Statistical Society* **6**, 186–197.

[13] Eisenhart, C. (1947). The assumptions underlying the analysis of variance, *Biometrics* **47**, 1–21.

[14] Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Iinear Models*. Springer, New York.

[15] Goldstein, H. (1995). *Multilevel Statistical Models*. Wiley, New York.

[16] Greenwood, M. & Yule, G.U. (1920). An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of Royal Statistical Society* **83**, 255–279.

[17] Hartley, H.O. & Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model, *Biometrika* **54**, 93–108.

[18] Henderson, C.R. (1953). Estimation of variance and covariance components, *Biometrics* **9**, 226–252.

[19] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics* **31**, 423–447.

[20] Johnson, N.L., Kotz, S. & Kemp, A.W. (1993). *Univariate Discrete Distributions*, 2nd Ed. John Wiley & Sons, New York.

[21] Kalbfleisch, J.D. & Sprott, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion), *Journal of Royal Statistical Society B* **32**, 175–208.

[22] Khuri, A.I. (2000). Designs for variance component estimation: past and present, *International Statistical Review* **68**, 311–322.

[23] Khuri, A.I. & Sahai, H. (1985). Variance components analysis: a selective bibliography and survey, *International Statistical Review* **53**, 279–300.

[24] Laird, N. (1978). Empirical Bayes methods for two-way contingency tables, *Journal of American Statistical Association* **65**, 581–590.

[25] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion), *Journal of Royal Statistical Society B* **58**, 619–678.

[26] Lee, Y. & Nelder, J.A. (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structural dispersions, *Biometrika* **88**, 987–1006.

[27] McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall, London.
[28] McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.
[29] Nguyen, D.V., Arpat, A.B., Wang, N. & Carroll, R.J. (2002). DNA microarray experiments: biological and technical aspects, *Biometrics* **58**, 701–717.
[30] Patterson, H.D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**, 545–554.
[31] Pearson, E.S. (1932). Discussion of paper by B.H. Wilsdon, *Supplement of the Journal of the Royal Statistical Society* **1**, 200–202.
[32] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-effects Models in S and S-PLUS*. Springer, New York.
[33] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature, *The Stata Journal* **2**, 1–21.
[34] Rao, P.S.R.S. (1997). *Variance Component Estimation*. Chapman & Hall, London.
[35] Rao, C.R. & Kleffe, J. (1988). *Estimation of Variance Components and Applications*. North-Holland, Amsterdam.
[36] Reeves, G.K., Cox, D.R., Darby, S.C. & Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models, *Statistics in Medicine* **17**, 2157–2177.
[37] Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion), *Statistical Science* **6**, 15–51.
[38] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.
[39] Shun, Z. (1997). Another look at the salamander mating data: a modified Laplace approximation approach, *Journal of American Statistical Association* **92**, 341–349.
[40] Skellam, J.G. (1948). A probability distribution derived from the binomial by regarding the probability of success as variable between sets of trials, *Journal of Royal Statistical Society B* **10**, 257–261.
[41] Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modelling*. Sage, London.
[42] Solomon, P.J. ed. (1998). Five papers on variance components in medical research, *Statistical Methods in Medical Research* **7**, 1–84.
[43] Solomon, P.J. & Cox, D.R. (1992). Nonlinear component of variance models, *Biometrika* **79**, 1–11.
[44] Speed, T.P. & Yang, Y.W. (2003). Direct and indirect hybridizations for cDNA microarray experiments, *Sankya Series A* **64**, 707–721.
[45] Stiratelli, R., Laird, N. & Ware, J. (1984). Random effects models for serial observations with binary responses, *Biometrics* **40**, 961–971.
[46] Tippett, L.H.C. (1931). *Methods of Statistics*. Matthew Norgate, London.
[47] Tukey, J.W. (1950). Some sampling simplified, *Journal of American Statistical Association* **45**, 501–519, Reprinted in *The collected works of* John W. Tukey, Vol. 7. Wadsworth, Pacific Grove.
[48] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

P.J. SOLOMON

Variance

The variance of a random variable X is a measure of the variable's spread or dispersion around its mean. The **mean** is the average or expected value of X, and is a measure of the center of its distribution. The mean can also be viewed as a "typical" value of X. By definition, however, the outcome of a **random variable** is unpredictable and will vary from one trial to the next; the variance describes the amount of variation that can be expected around the average value. If the mean of X is represented by $\mathrm{E}(X)$, then its variance is defined by

$$\begin{aligned}\operatorname{var}(X) &= \mathrm{E}[X - \mathrm{E}(X)]^2 \\ &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2,\end{aligned}$$

provided that $\mathrm{E}(X)$ exists. The mean of a random variable X is often denoted by μ and its variance by σ^2.

From the preceding definition, the variance of X is the average value of the squared deviation of X from its mean. Since a quantity which is squared cannot be negative, the variance is never less than 0. If X is a measurable quantity such as length or temperature, then the units of measurement for the variance are the square of the units for X. If X is measured in meters, for example, then the variance of X is measured in meters squared. This is a major drawback in the variance's use a measure of dispersion – it is difficult for most people to think in terms of squared units. Nevertheless, a large variance indicates that the outcomes of X are widely distributed around its mean, while a small variance means that the outcomes cluster tightly around the center. In practical applications, a measure of dispersion called the **standard deviation** is often used in place of the variance; the standard deviation is the positive square root of the variance and is denoted by σ.

To calculate the variance of a random variable, it is necessary to know the probability distribution of X. If X is a discrete random variable with mean $\mathrm{E}(X) = \mu$, then

$$\begin{aligned}\operatorname{var}(X) &= \sum_i (x_i - \mu)^2 \Pr(X = x_i) \\ &= \left[\sum_i x_i^2 \Pr(X = x_i)\right] - \mu^2,\end{aligned}$$

where $x_1, x_2, \ldots x_i \ldots$ are all outcomes of X such that $\Pr(X = x_i) > 0$. If X is a continuous random variable with probability density function $f(x)$ and mean $\mathrm{E}(X) = \mu$, then

$$\begin{aligned}\operatorname{var}(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,\mathrm{d}x \\ &= \left[\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x\right] - \mu^2.\end{aligned}$$

A linear transformation of the random variable X affects the variance in a straightforward manner. If a and b are constants and if the random variable $Y = aX + b$, then

$$\operatorname{var}(Y) = a^2 \operatorname{var}(X).$$

A constant has variance 0.

In practice, the variance of a distribution can be estimated using the information contained in a sample of observations drawn from that distribution. If $x_1, x_2, \ldots, x_n$ is a **random sample** of size n selected from a population with mean μ and variance σ^2, then the sample variance is represented by s^2 and is defined by

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1},\end{aligned}$$

where $\overline{x}$ is the sample mean

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Just as σ^2 describes the dispersion of a distribution around its mean μ, s^2 describes the spread of a sample of values around the sample mean x. It can be thought of as the average squared deviation of each observation from the sample mean. While it might seem more natural to estimate σ^2 by the true average

squared deviation, or

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n},$$

this estimator is seldom used in practice. This is because s^2 is an **unbiased** estimator of σ^2 over all possible random samples of size n – meaning that $\mathrm{E}(s^2) = \sigma^2$ – while $\hat{\sigma}^2$ is not unbiased.

(*See also* **Moments**)

K. GAUVREAU

Varimax Rotation

Varimax rotation is probably the most popular **orthogonal rotation** procedure for use with **principal components analysis** and **factor analysis**. Given a matrix $\mathbf{V}$ of dimension $p \times k$ consisting of a set of k vectors defining a set of principal components or factors, a new set of transformed variables is obtained by an orthogonal rotation of $\mathbf{V}$, namely $\mathbf{B} = \mathbf{V}\boldsymbol{\Theta}$. $\boldsymbol{\Theta}$ is a $k \times k$ matrix determined such that the coefficients of $\mathbf{B}$, a $p \times k$ matrix, will maximize the quantity

$$Q = \sum_{j=1}^{k} \left[\sum_{i=1}^{p} b_{ij}^4 - \left(\frac{1}{p}\right) \left(\sum_{i=1}^{p} b_{ij}^2 \right)^2 \right],$$

where p is the number of original variables and k is the number of retained components or factors. Varimax rotation is a special case of *orthomax* rotation with $c = 1.0$ (*see* **Factor Analysis, Overview**; **Rotation of Axes**) In this procedure, the sums of squares of $\mathbf{B}$ are maximized *columnwise* as contrasted with **quartimax rotation**, which maximizes them *rowwise*. Varimax is due to Kaiser [3, 4] and was used in the original Little Jiffy of Kaiser & Rice [5]. It is the default in some computer packages. The form above is referred to as *raw varimax* and, in this form, the effect of each variable on the rotation is a function of the amount of the variability in that variable accounted for by the retained components or factors. To put the variables on an equal footing, the individual b_{ij}^2 may be divided by the corresponding diagonal element of $\mathbf{VV}'$. The resultant form is called *normal varimax*, and is the default in some other computer packages. To deal with some pathological situations which could arise, Cureton & Mulaik [2] introduced another version called *weighted varimax*. The standard errors for the vector coefficients produced by varimax rotation were given by Archer & Jennrich [1].

For the decathlon example introduced in the articles on **Rotation of Axes** and **Factor Analysis, Overview**, $\mathbf{V}$ and $\mathbf{B}$ are repeated here in Table 1.

References

[1] Archer, C.O. & Jennrich, R.I. (1973). Standard errors for rotated factor loadings, *Psychometrika* **38**, 581–592.

[2] Cureton, E.E. & Mulaik, S.A. (1975). The weighted varimax rotation and the promax rotation, *Psychometrika* **40**, 183–195.

[3] Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

[4] Kaiser, H.F. (1959). Computer program for varimax rotation in factor analysis, *Educational and Psychological Measurement* **19**, 413–420.

[5] Kaiser, H.F. & Rice, J. (1974). Little Jiffy, Mark IV, *Educational and Psychological Measurement* **34**, 111–117.

(*See also* **Axes in Multivariate Analysis**)

J. EDWARD JACKSON

Table 1 Decathlon data: characteristic and rotated vectors

	Characteristic vectors				Varimax rotation			
	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
100 m run	0.69	0.22	−0.52	−0.21	0.88	0.14	0.16	−0.12
Long jump	0.79	0.18	−0.19	0.09	0.63	0.19	0.52	−0.01
Shotput	0.70	−0.53	0.05	−0.18	0.24	0.82	0.22	−0.15
High jump	0.67	0.13	0.14	0.40	0.24	0.15	0.75	0.08
400 m run	0.62	0.55	−0.08	−0.42	0.80	0.07	0.10	0.47
110 m hurdle	0.69	0.04	−0.16	0.35	0.40	0.15	0.64	−0.17
Discus	0.62	−0.52	0.11	−0.23	0.19	0.81	0.15	−0.08
Pole vault	0.54	0.09	0.41	0.44	−0.04	0.18	0.76	0.22
Javelin	0.43	−0.44	0.37	−0.24	−0.05	0.74	0.11	0.14
1500 m run	0.15	0.60	0.66	−0.28	0.05	−0.04	0.11	0.93

Variogram

The variogram is a description of the second-order dependence properties of a **stochastic process**. It was first proposed by Jowett [4] in the context of industrial sampling and subsequently popularized by the French geostatistical school [5], where it is a fundamental ingredient in the method of spatial prediction known as kriging. Its importation to the field of longitudinal data analysis stems from Diggle [2]. In this article we will describe the variogram in the context of longitudinal data analysis. For a counterbalancing view emphasizing spatial applications, see [1].

For a stochastic process $Y(t)$, where t denotes time, the *covariance function* is the function $\gamma(t, s) = \text{cov}\{Y(t), Y(s)\}$. If $Y(t)$ is **stationary**, the covariance between $Y(t)$ and $Y(s)$ only depends on $u = |t - s|$, in which case we write the covariance function as $\gamma(u)$. The *variogram* of $Y(t)$ is the function

$$V(u) = \tfrac{1}{2}\text{Var}\{Y(t) - Y(t-u)\}, \qquad (1)$$

if this function exists. Note that if $\text{E}[Y(t)] = \mu$, a constant for all t, then

$$V(u) = \text{E}[\tfrac{1}{2}\{Y(t) - Y(t-u)\}^2].$$

Some authors call $V(u)$ the *semivariogram*. The variogram exists for all stationary processes $Y(t)$ and for a limited class of nonstationary processes. In the stationary case, $V(u) = \gamma(0) - \gamma(u)$.

One practical advantage of the variogram over the covariance function is that estimation from observed data is more straightforward, especially when the underlying stochastic process is observed at irregular time-points. Consider a set of longitudinal data in the form $(t_{ij}, y_{ij}) : j = 1, \ldots, n_i; i = 1, \ldots, m$, in which y_{ij} denotes the jth of n_i observations on the ith of m subjects, and t_{ij} the corresponding observation times. We assume that y_{ij} is a realization of $Y(t_{ij})$ and that observations from different subjects (different values of i) are independent.

The *empirical variogram* is the set of points (u_{ijk}, v_{ijk}): $k > j$; $i = 1, \ldots m$, where $u_{ijk} = |t_{ij} - t_{ik}|$ and $v_{ijk} = \frac{1}{2}(y_{ij} - y_{ik})^2$. A scatterplot of the points (u_{ijk}, v_{ijk}) is called a *variogram cloud*. In principle, the variogram cloud can be used to suggest a parametric model for $V(u)$, and to obtain quick, if inefficient, estimates of the parameters of $V(u)$ by nonlinear ordinary least squares regression, the justification for this being that $\text{E}(v_{ijk}) = V(u_{ijk})$. The usefulness of the variogram cloud for exploratory data analysis is limited by the fact that the sampling distributions of the empirical variogram ordinates v_{ijk} are typically highly variable and highly skewed. For example, if $Y(t)$ is a Gaussian process, then $v_{ijk}/V(u_{ijk})$ has a χ^2_1 sampling distribution. Furthermore, pairs of variogram ordinates from the same subject are typically dependent.

A more useful graphical display is the *sample variogram*, defined by averaging empirical variogram ordinates v_{ijk} at common values of u_{ijk}. In most designed longitudinal studies there is a high degree of commonality amongst the observation times associated with the different subjects, and the averaging imparts a high degree of stability to the sample variogram by comparison with the empirical variogram. When each of the subjects have their own unique set of observation times, an alternative graphical display can be obtained by averaging the v_{ijk} corresponding to *approximately* equal values of u_{ijk}, or by applying a nonparametric scatterplot smoother to the points (u_{ijk}, v_{ijk}). See, for example, [3, p. 53].

Because variogram ordinates from different subjects are independent, in the stationary case an estimate of $\gamma(0) = \text{var}\{Y(t)\}$ can be obtained as the average of all quantities of the form $\frac{1}{2}(y_{ij} - y_{i'k})^2$ for all j, k and $i' > i$.

Mathematically legitimate forms for the theoretical variogram $V(u)$ are constrained by the need for $V(u)$ to correspond to a legitimate covariance structure. An algebraic condition that a variogram must satisfy is that $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j V(t_i - t_j) \leq 0$, for any positive integer n, set of times $t_i; i = 1, \ldots, n$, and real numbers $a_i, i = 1, \ldots, n$, such that $\sum_{i=1}^{n} a_i = 0$. For examples of variogram models for longitudinal data, see [3, Chapter 5]. Common features of $V(u)$ in the longitudinal setting are the following:

1. *A nonzero value at the origin*. We interpret $V(0)$ as the expectation of one half the squared difference between two independent determinations of $Y(t)$ from the same subject. If we think in terms of a model in which $Y(t)$ is observed with additive measurement errors, $V(0)$ is the variance of the measurement error. In the spatial setting, $V(0)$ is called the *nugget effect*.

2. *Increasing trend*. An increasing function $V(u)$ corresponds to a decreasing $\gamma(u)$. Most models for the serial correlation in longitudinal (or spatial) data assume that correlation decreases monotonically with increasing time-separation.
3. *Limiting behavior at large time-separation*. The difference between the estimated variance of $Y(t)$ and the limiting value of $V(u)$ as $u \to \infty$ arises because of variation between subjects, which is included in var$\{Y(t)\}$ but excluded from $V(u)$.

Often, in practice, the available data are generated by a model of the form $y_{ij} = \mu_i(t_{ij}) + z_{ij}$ where the function $\mu_i(t)$ represents the mean response for subject i at time t and z_{ij} is a realization at time t_{ij} of an underlying **stationary** random process $Z(t)$. In this case, we estimate the unobservable z_{ij} as $\hat{z}_{ij} = y_{ij} - \hat{\mu}_i(t_{ij})$ and use the $\hat{z}_{ij}$ as the basis for calculating the empirical or sample variogram. In a designed experiment with a common set of observation times for all subjects, the effects of estimating the $\mu_i(t_{ij})$ can be minimized by fitting a saturated model for the mean response. More generally, the recommended practice is to base the estimates $\hat{\mu}_i(t_{ij})$ on a deliberately over-parameterized model. For further discussion, see [3, Chapter 5].

Figure 1 shows the sample variogram derived from a set of data on the log-bodyweights of 27 cows, allocated amongst four treatment groups. Each cow was weighed at 23 unequally spaced time-points over a period of approximately two years. The variogram was computed using the residuals from a saturated treatments-by-times model for the mean response. The horizontal line denotes the estimate of the variance of the residual process, assuming that this process is stationary. The large number of plotted points is a consequence of the fact that time intervals between successive weighings were not common to all cows, which limits the opportunities for averaging the empirical variogram ordinates. The other salient features of Figure 1 are:

1. *Behavior near the origin*. Simple extrapolation suggests that $V(0)$ is close to zero, which in turn suggests, reasonably, that the measurement error in determining the log-bodyweight is negligible.
2. *Increasing trend*. Log-bodyweights at different times become less correlated as the time-separation increases; the shape of this rising trend conveys information about the shape of the underlying covariance function $\gamma(u)$; note, however, that the sampling fluctuations in $\hat{V}(u)$ are substantial, and increase with u.
3. *Difference between the estimated variance and the values of $\hat{V}(u)$ at large values of u*. Assuming that $V(u)$ has leveled out within the plotted range of u, this difference appears to be substantial, suggesting a large component of variance between cows.

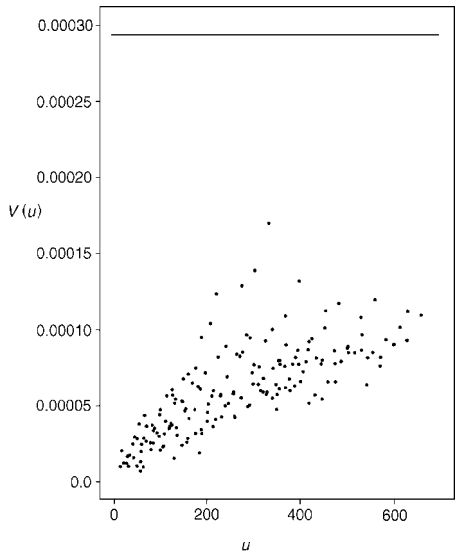

Figure 1 The sample variogram of ordinary least squares residuals from a saturated treatment-by-times model fitted to data on the log-bodyweights of cows

References

[1] Cressie, N. (1988). Variogram, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson. eds. Wiley, New York, pp. 489–491.
[2] Diggle, P.J. (1988). An approach to the analysis of repeated measurements, *Biometrics* **44**, 959–971.
[3] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.
[4] Jowett, G.H. (1952). The accuracy of systematic sampling from conveyor belts, *Applied Statistics* **1**, 50–59.
[5] Matheron, G. (1963). Principles of geostatistics, *Economic Geology* **58**, 1246–1266.

PETER J. DIGGLE

Vector Field Plot

Slope Fields

One may think of a slope field as a scatterplot (*see* **Graphical Displays**) where at each selected pair (x, y) denoting a location with respect to abscissa and ordinate axes is plotted the expected change in y for a one unit change in x as indicated by a short line segment of that slope. As an example, suppose a variable y were related to its slope with respect to time t such that

$$\frac{\Delta y(t)}{\Delta t} = \beta(y(t) - C), \tag{1}$$

where $y(t)$ is some time varying score, $\Delta y(t)/\Delta t$ is the expected change in $y(t)$ over some fixed interval of time Δt, β is some fixed coefficient, and C is some asymptotic value (i.e. *fixed point equilibrium*). From this equation, a slope field may be plotted such that on each point in a grid of pairs of values of Y and t a line segment is centered with the expected slope given that pair of values (i.e. *initial conditions*)(see Figure 1). The lengths of the line segments are all equal to one another. Three salient characteristics of the slope field shown are that (a) since each row of line segments has no variance in slope, the slopes are independent of time, (b) since each column of line segments has variance in slope, the slope is not independent of y, and (c) since the slopes are near zero when y is near 60, the equilibrium C in this plot must be near 60.

Slope fields may be plotted in which one variable represents time, or may also be plotted so as to visualize the expected slope of two variables with respect to each other. The expected slope could potentially vary with respect to the value of one, both, or neither of the two variables.

Vector Fields

Vector fields differ from slope fields in that they are directional, implying some evolution of the slope (i.e. first derivative) with respect to time. Vector fields are useful for visualizing the implications of differential equations over a range of initial conditions [5, 10]. A vector field is composed of a grid of arrows that may vary in direction and length. Direction and length

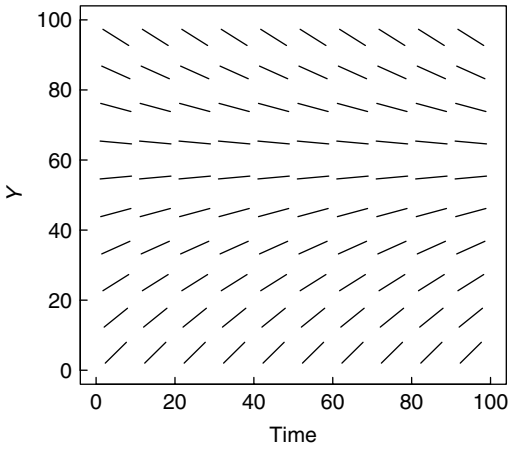

Figure 1 Slope field plotted for Equation 1

may be mapped to a variety of concepts. The most commonly used vector field display maps the vectors so that given the values of the variables in a system at time t plotted as the base of a vector, the length and the direction of the vector point to the values of those variables after some chosen interval τ has elapsed.

As an example in continuous time, the relationship between two variables X and Y might be modeled as a set of simultaneous differential equations

$$\begin{aligned} \frac{\mathrm{d}X}{\mathrm{d}t} &= \alpha_x + \beta_x X(t) + \gamma_x Y(t) \\ \frac{\mathrm{d}Y}{\mathrm{d}t} &= \alpha_y + \beta_y Y(t) + \gamma_y X(t), \end{aligned} \tag{2}$$

where the coefficients $(\alpha,\ \beta,\ \gamma)$ for each variable are used to describe the instantaneous trajectory of the bivariate processes. Or equivalently in discrete time might be formed a set of simultaneous difference equations

$$\begin{aligned} &\frac{\Delta X(t)}{\Delta t x \alpha_x} + \beta_x X(t - \Delta t) + \gamma_x Y(t - \Delta t) \\ &\frac{\Delta Y(t)}{\Delta t x \alpha_y} + \beta_y Y(t - \Delta t) + \gamma_y X(t - \Delta t), \end{aligned} \tag{3}$$

where Δt is defined by the application, and the coefficients for each variable describe the step-by-step trajectory of the bivariate processes [4, 6, 9]. Such a system is plotted in Figures 2 (a) and (b).

Vector field plots such as that shown in Figure 2b can be read to provide several forms of information.

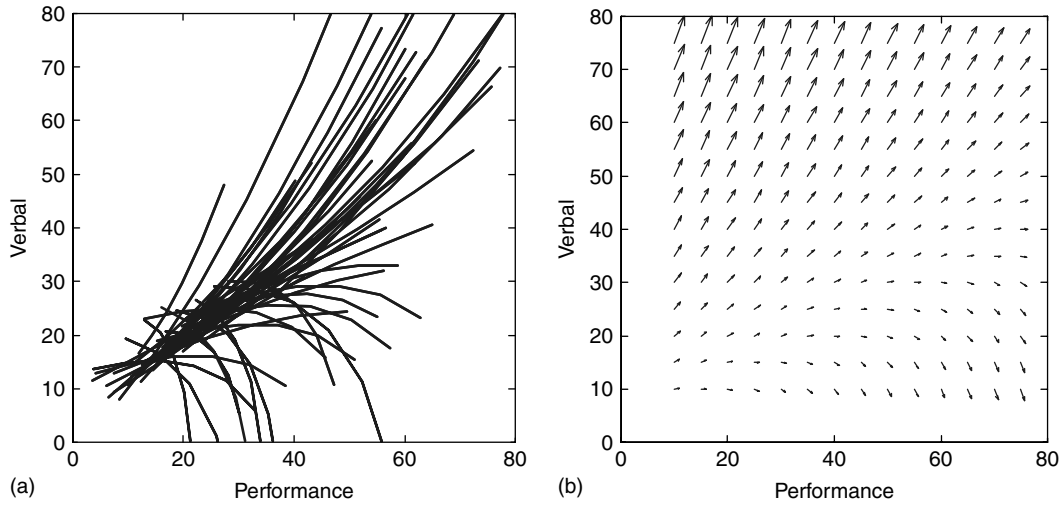

Figure 2 Trajectories and vector field generated from coefficients estimated from WISC data from $N = 204$ Children aged 6–11 [7]. (a) Hypothetical individual trajectories evolving over a lifetime from a few selected initial conditions. (b) A vector field plotting the evolution of a grid of initial conditions over a short interval of time

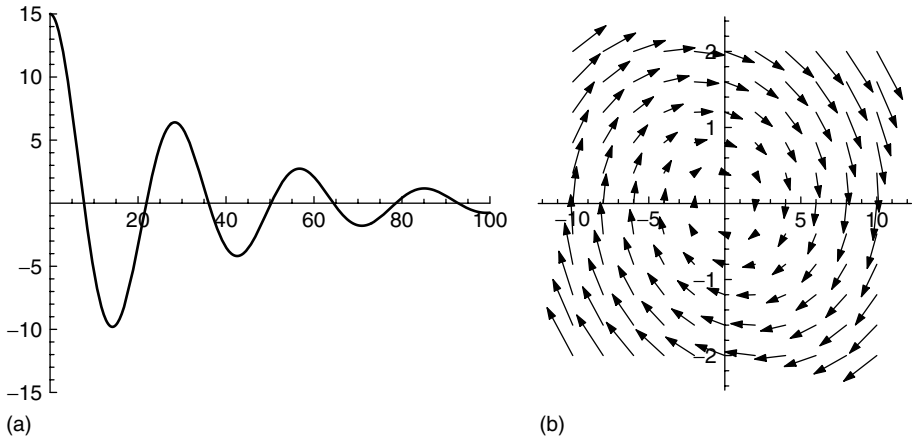

Figure 3 (a) Time series plot and (b) vector field plot of a damped linear oscillator conforming to Equation 4

In areas of the graph where vector lengths are small, the system is near an equilibrium. If vectors point away from an equilibrium, then that equilibrium is unstable. If vectors point towards an equilibrium, then that equilibrium is stable. If vectors appear to "circle" an equilibrium, then the system may oscillate under some initial conditions.

Vector fields are also used to visualize the relationship between a variable and its first and second derivatives with respect to time [1, 3]. The vector field in Figure 3 plots the expected change in x and its first derivative $\dot{x}$ over a short interval of time for the damped linear oscillator system

$$\ddot{x} = \eta x + \zeta \dot{x} \tag{4}$$

when η and ζ are both negative. This system has a stable equilibrium at $x = 0$ and oscillates about that equilibrium for a time that is dependent on its initial conditions.

Statistical Slope Fields

When performing **exploratory** analyses on **longitudinal data**, it is often useful to visualize the expected change of one variable with respect to another or in a variable with respect to time. Statistical slope fields employ **nonparametric methods** of local aggregation or smoothing to develop local estimates of the derivatives of systems and then plot these empirically derived estimates. Typically, a statistical slope field appears similar to a parametric slope field, but varies the length of the line segment as a means of displaying the relative proportion of the data near to the center of the line segment.

As an example, suppose a random sample of 100 individuals of different ages were drawn from a population whose scores were evolving according to the autoregressive system from (1) (*see* **ARMA and ARIMA Models**). Suppose each of these individuals were measured at two occasions separated by 15 years and that the measured score were the sum of a true score and **normally distributed** independent

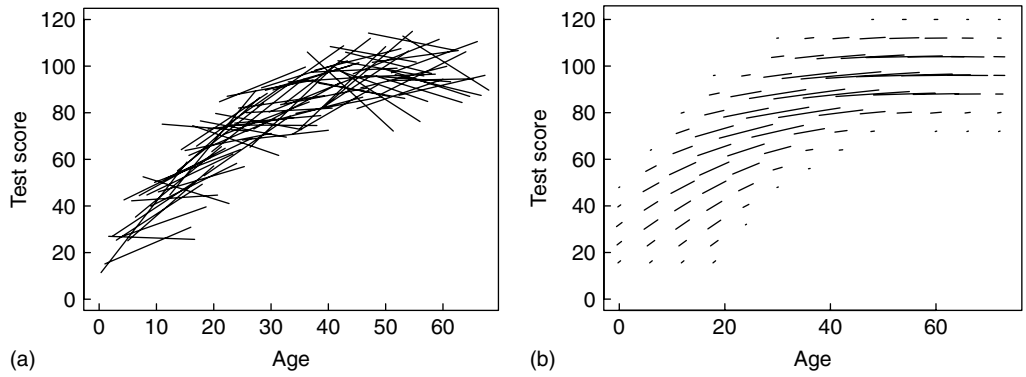

Figure 4 (a) Longitudinal time series plot and (b) statistical slope field plot of an autoregressive process conforming to Equation 1

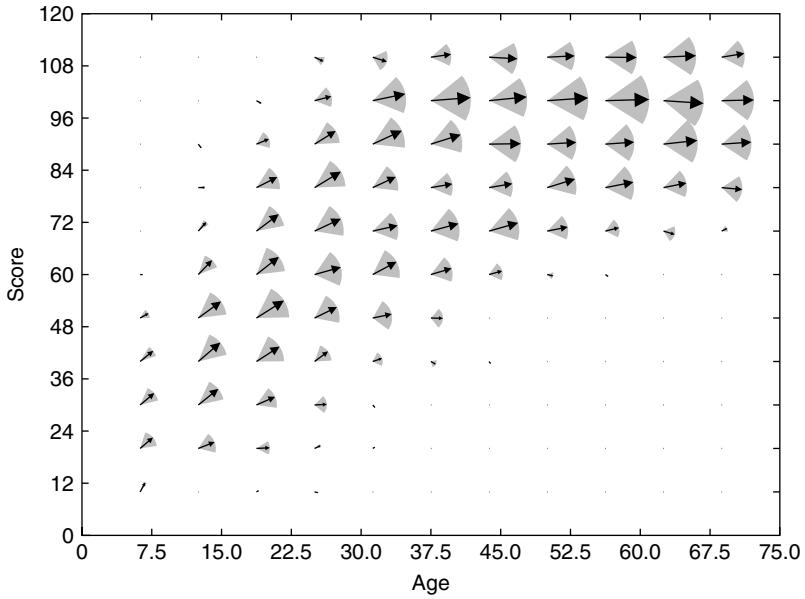

Figure 5 A statistical vector field plot of data from an autoregressive process conforming to Equation 1

error. These longitudinal scores could be plotted as in Figure 4(a). The statistical slope field of the same data was calculated using loess smoothing (*see* **Graphical Displays**) to locally estimate the derivative of the score with respect to time at the center of each point in a grid of age and score pairs. In some areas, there is no data and thus no estimate is made. Derivative estimates with larger Ns are plotted as longer line segments.

Statistical Vector Fields

A statistical vector field [2] is similar to a statistical slope field except that there is a time direction of the evolution of the system and an estimate of the variability of the slopes is made in the locality of each initial condition pair. The direction of the vector represents the expected change in the value of the variable shown on the ordinate axis with respect to a unit change in the variable shown on the abscissa. The length of the vector plots the proportion of the data in the vicinity of the initial condition pair at the base of the vector. The estimated **standard deviation** of this slope is plotted as a gray arc centered around the vector. An example statistical vector field of data conforming to (1) is plotted in Figure 5 [8, 9].

Acknowledgment

Funding for this work was provided in part by NIH Grants AG−14983 and AG−07137. Correspondence may be addressed to Steven M. Boker, Department of Psychology, The University of Notre Dame, Notre Dame Indiana 46556, USA; email sent to sboker@nd.edu. Vector field and slope field plots can be created in Matlab, Mathematica, Maple, and S plus or R. Example code to produce the figures in this entry may be obtained from `http://www.nd.edu/sboker`.

References

[1] Boker, S.M. & Graham, J. (1998). A dynamical systems analysis of adolescent substance abuse, *Multivariate Behavioral Research* **33**(4), 479−507.

[2] Boker, S.M. & McArdle, J.J. (1995). Statistical vector field analysis applied to mixed cross−sectional and longitudinal data, *Experimental Aging Research* **21**(1), 77−93.

[3] Boker, S.M. & Nesselroade, J.R. (2002). A method for modeling the intrinsic dynamics of intraindividual variability: Recovering the parameters of simulated oscillators in multi−wave panel data, *Multivariate Behavioral Research* **37**(1), 127−160.

[4] Hamagami, F., McArdle, J. & Cohen, P. (2000). A new approach to modeling bivariate dynamic relationships applied to evaluation of comorbidity among DSM−III personality disorder symptoms, in *Temperament and Personality Development Across the Life Span*, V.J. Molfese, ed. Lawrence Erlbaum Associates, Mahwah, pp. 253−280.

[5] Hubbard, J.H. & West, B.H. (1991). *Differential equations: A dynamical systems approach*. Springer-Verlag, New York.

[6] McArdle, J.J. (2000). A latent difference score approach to longitudinal dynamic structural analyses, in *Structural Equation Modeling: Present and Future*, R. Cudeck, S. du Toit & D. Sorbom, eds. Scientific Software International, Lincolnwood, pp. 342−380.

[7] McArdle, J.J., (2001). A latent difference score approach to longitudinal data analysis, in *Structural Equation Models: Present and Future*, R. Cudeck, S. du Toit & D. Sorbom, eds. Scientific Software, Chicago, pp. 341−379.

[8] McArdle, J. & Hamagami, F. (2003). Longitudinal tests of dynamic hypotheses on intellectual abilities measured over sixty years, in *Quantitative Methodology in Aging Research*, C.S. Bergeman & S.M. Boker, eds. Lawrence Erlbaum Associates, Mahwah in press.

[9] McArdle, J.J., Hamagami, F., Meredith, W. & Bradway, K.P. (2001). Modeling the dynamic hypotheses of gf-gc theory using longitudinal life−span data, *Learning and Individual Differences* **12**, 53−79.

[10] Thompson, J.M.T. & Stewart, H.B. (1986). *Nonlinear Dynamics and Chaos*. John Wiley & Sons, New York.

(*See also* **Graphical Presentation of Longitudinal Data**)

STEVEN M. BOKER & JOHN J. MCARDLE

Viral Population Growth Models

Viruses are amongst the most dangerous and devastating threats to human health. They may invade a human or animal population and spread rapidly amongst its members, sometimes causing a large number of fatalities, possibly on a recurrent basis. A well-known case is the influenza virus, a new strain sweeping the globe and causing approximately 25 million deaths in 1919. At the present time, there are several countries, particularly in Africa, with up to 35% of their populations between the ages of 15 and 50 years infected by human immunodeficiency virus (HIV) (*see* **AIDS and HIV**). Throughout the world, already over 16 million deaths have been caused by this virus. More recently, a corona virus causing SARS (severe acute respiratory syndrome) [9] has created panic and economic chaos in certain southeast Asian countries. Renewed interest in the dynamical processes involved in the spread of viruses has also arisen recently because of the threat of bioterrorist attacks, especially with such viruses as smallpox [6], which was eradicated many years ago. We will first briefly describe in a simplified fashion some of the processes that determine the outcome of a viral invasion [14].

Viral Reproduction

Viral reproduction depends on host cells. The sequence of steps after the virus has penetrated the body's initial physical barriers (skin, mucosal lining) is

(a) the virus attaches to a host cell at a receptor on the cell surface,
(b) penetration occurs,
(c) the virus sheds its protein coat and releases its nucleic acid (RNA or DNA) into the cell,
(d) transcription occurs followed by replication of the virus genetic material and the production of proteins for new coats, and
(e) virus particles are assembled and released and may infect new host cells; the original host cell may die.

The time taken for some of these steps is extremely variable. For HIV, replication is unpredictable but may occur in a few hours; with herpes virus there may be a delay of weeks or up to many years, which seems to be an evolutionary strategy [25].

Models of viral population growth may be distinguished by whether they consider the within-host population, or the total across all individuals. To determine the latter, the dynamics of transmission amongst members of the host population are needed and this is usually the domain of classical models such as **SIR** (susceptible infected, recovered). These are discussed, for example, in Bailey [2] and Hethcote [8]. Recently, we have formulated nonlinear dynamical spatial network models for determining the total viral load (see [28], and references therein). In this work, we only consider within-host dynamics.

Deterministic Models

Since the growth of a viral population depends on the ability of the virus to penetrate new host cells, the simplest growth model has the following three components for a given volume of tissue: $x(t) =$ number of uninfected cells; $y(t) =$ number of infected cells; and $v(t) =$ number of free virus particles. Then,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \lambda - \mu x - \beta v x \tag{1}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \beta v x - \alpha y \tag{2}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = cy - \gamma v - \beta v x, \tag{3}$$

where uninfected host cells are supplied at rate λ and have a per cell death rate of μ, the parameter β describes the rate at which virus infects host cells, c is the rate at which free virions are produced per infected cell, and α and γ are the "per capita" rates of attrition of infected cells and virions, respectively. Equation (3) is a slightly modified version of the standard model formulated by Herz et al. [7], which is analyzed in [18]. The modification, suggested by Tuckwell and Le Corfec [27] and others consists of the additional term $-\beta v x$ in (3) to allow for the fact that whenever a cell is attacked, a free virus must disappear. The more complete model (1)–(3) is analyzed in detail by Tuckwell and Wan [29].

The system (1)–(3) has two equilibrium points:

$$P_1 = \left(\frac{\lambda}{\mu}, 0, 0\right),$$
$$P_2 = \left(\frac{\alpha\gamma}{\beta c}, \frac{\lambda}{\alpha} - \frac{\gamma\mu}{\beta c}, \frac{c\lambda}{\alpha\gamma} - \frac{\mu}{\beta}\right). \quad (4)$$

For $\lambda > 0$, P_1 is either a saddle point or an asymptotically stable node, but for usual parameter values, the former. P_2 is usually an asymptotically stable spiral point. Note that P_2 is unphysical when $\beta c\lambda < \alpha\gamma\mu$. When P_2 is in the first octant, solutions of the system (1)–(3) approach this point in an oscillatory fashion so that eventually there remain infected cells, virions, and uninfected cells in equilibrium. In the less usual case that P_2 is unphysical, the free virus must be extinguished.

The above model does not include an "immune response". The presence of infected cells may stimulate the production of cytotoxic T-cells, which attack infected cells. If these have concentration $z(t)$, then a plausible model system is given by (1) and (3) with (2) changed by the addition of a term representing the removal of infected cells

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \beta vx - \alpha y - \rho yz \quad (2')$$

and an additional equation for the cytotoxic cells:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = kyz - \delta z, \quad (5)$$

where δ is their natural death rate. In (2′), ρ is the per capita rate of removal of infected cells per immune cell, and k is the per capita rate of production of immune cells per infected cell. Some dynamical properties of the model (1),(2′),(3) (without the correction term) and (5) are also discussed in [18].

HIV

The virus that has attracted the most dramatic attention in the previous two decades is HIV (human immunodeficiency virus, usually type 1). A distinguishing feature of this virus is that the infected cells are those of the immune system itself, being CD4+ T-cells (helper cells). After infection, there is a rapid initial rise of virion density followed usually by a similarly paced fall, the latter being originally ascribed to an immune response. Modeling indicated that the early dynamics might be reasonably explained without invoking such a response [12, 21], and this has been mainly vindicated by subsequent studies. However, fitting the data on viral loads after the primary peak in some patients seemed to require the introduction of more complex dynamics such as the inclusion of cytotoxic T lymphocytes [23]. There is a large number of models that have been posited to describe these phenomena, and they may be distinguished by whether they address only the early phases or the later phases of the disease – Perelson and Nelson [20] and Nowak and May [18] can be consulted for numerous references.

The general model of McLean et al. [12] was later used in a different context by Phillips [21] to explain the decline in viral load in HIV after the initial rise to about 5000 per mm^3. Two classes of infected cells [17] are introduced because the insertion of viral genetic material may be followed by a delay before virions emerge. Thus, $x(t)$ is the density of uninfected CD4+ T-cells, but now $u(t)$ is the density of latently infected cells and $w(t)$ is the density of infected cells producing virus. With $v(t)$, the density of free virions in the plasma we have, with an added correction term for $\mathrm{d}v/\mathrm{d}t$:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \lambda - \mu x - \beta vx \quad (6)$$
$$\frac{\mathrm{d}u}{\mathrm{d}t} = \beta pvx - (\mu + \alpha)u \quad (7)$$
$$\frac{\mathrm{d}w}{\mathrm{d}t} = \beta(1-p)vx + \alpha u - aw \quad (8)$$
$$\frac{\mathrm{d}v}{\mathrm{d}t} = cw - \gamma v - \beta vx. \quad (9)$$

Here, p is the fraction of infected cells, which become latent, α is the rate of conversion of latent to actively infected cells, and a is the death rate of actively infected cells. The term λ in (6) gives the rate at which new T-cells are created from sources in the body. Another term may be added, such as a logistic, to represent growth due to the proliferation of existing T-cells. The parameters are possibly time-varying so that, for example, if the immune response is weakening, then γ may decrease. A long-term model not dissimilar to the above and which leads to the break down of immunity (AIDS) was analyzed in [24].

In all the above dynamical models, not only for HIV but also for other viruses, the law of "mass action" is assumed to operate giving a rate of

new infections proportional to the product of viral and host cell numbers. However, there are strong grounds for sometimes questioning the validity of this assumption. For such a law to apply, there must be homogeneous mixing and the latter is unlikely if there are only one or two virions and say 10^{11} host cells. The law may, nevertheless, be accurate for a very restricted volume of tissue or fluid. It is also worth mentioning that the viral density in plasma may be low, or essentially zero, but that due to the continued existence of latently infected host cells, the viral density may increase at a later time, making the disease extremely difficult to eradicate.

Other Viruses

There has been a considerable effort to model the invasion and within-host growth of influenza virus populations, sometimes with very complex systems [3]. Although there have been many epidemiological studies of the spread of such viruses as smallpox, measles, herpes, and so on, mathematical models for the growth of their within-host viral populations have been sparse. Nowak and May [18] have described simple differential equations for the hepatitis B virus, which presently afflicts about 300 million people, and Neumann et al. [15] have analyzed the dynamics of the hepatitis C virus.

Stochastic Dynamical Models for HIV

As pointed out by Tan and Wu [26], a stochastic description of viral population growth is more realistic than a deterministic one because of the nature of the subcellular processes. Furthermore, it is expected that a stochastic approach can provide a more accurate quantitative basis for evaluating the efficacy of drug treatments in infected host populations. Nevertheless, relatively little attention has been given to stochastic dynamical viral growth models. From an analytic viewpoint, this is probably due to the complexity of the systems. The general viral growth model (1)–(3) does not have a stochastic counterpart, though one could easily be developed using approaches similar to those described below for HIV. One of the first stochastic models for HIV was a simple **branching process** [13].

In general, the following effects are stochastic:

(a) generation and fluctuations in the rate of appearance of new host cells,
(b) contacts between viruses and the host cell and random attachment,
(c) transition to active or latent infected cell,
(d) time for the emergence of new virions,
(e) number of new virions emerging from a host cell,
(f) death process for infected and uninfected host cells and virions,
(g) mutation to other viral strains, and
(h) appearance and action of immune system components, which assist in the removal of virions.

Models for a Single Viral Genotype

Here, it is assumed that all virus particles have the same genetic properties. Vector valued **Markov process** models for the growth of HIV populations have been introduced by Tuckwell and Le Corfec [27] and Tan and Wu [26]. Both models have the same four biological components. A comparison of these models was made in Kamina et al. [10], but the nature of the boundaries for the diffusions (see e.g. [4]) needs further investigation. The Tuckwell–Le Corfec model is a four-dimensional diffusion process and as it is simpler than Tan and Wu's model, it will be described first, after a consideration of a more fundamental model involving **Poisson processes**.

Stochastic Model for Early HIV Growth. In this work, emphasis is on the early period (to several weeks) after infection and not on the later progression to the acquired immune deficiency syndrome, which may follow. The stochastic properties included are (b) and (c) in the above list, so that emphasis is on the viral production process. It is assumed that the number of virions produced by each actively infected cell and the birth and death rates of the various components take fixed mean values, so variability is underestimated but with the advantage of simplification and far fewer parameters.

Letting the components be X_k, $k = 1, 2, 3, 4$, we have that at time t (days) after initial infection, for a fixed and relatively small (in order to enhance the validity of the mass action principle) volume of plasma, $X_1(t)$ is the number of uninfected CD4+ T-cells (called "activated" by Phillips [21]), $X_2(t)$ is

the number of latently infected cells, $X_3(t)$ is the number of actively infected cells, and $X_4(t)$ is the number of circulating HIV-1 virions. The attachment of virus to CD4+ T-cells (assumed to be one on one) occurs according to a Poisson process $N = \{N(t), t \geq 0\}$ with rate $\beta X_1 X_4$ so that the system of stochastic differential equations corresponding to the above deterministic model is

$$\mathrm{d}X_1(t) = (\lambda - \mu X_1(t)\,\mathrm{d}t - \mathrm{d}N(t) \tag{10}$$

$$\mathrm{d}X_2(t) = -(\mu + \alpha)X_2(t)\,\mathrm{d}t + \mathrm{d}X(t) \tag{11}$$

$$\mathrm{d}X_3(t) = (\alpha X_2(t) - aX_3(t))\,\mathrm{d}t + \mathrm{d}Y(t) \tag{12}$$

$$\mathrm{d}X_4(t) = (cX_3(t) - \gamma X_4(t))\,\mathrm{d}t - \mathrm{d}N(t). \tag{13}$$

The random variable $X(t)$ is **binomial** with parameters $N(t)$ and p and $Y(t)$ is binomial with parameters $N(t)$ and $1 - p$. That is, at time t for each of $N(t)$ virus-uninfected cell interactions, the probability of a transition to a latently infected cell is p; and the probability of a transition to an actively infected cell is $(1 - p)$. Note that N has different units in, for example, (10) and (13), but equal numerical values. In a diffusion approximation [27] (*see* **Brownian Motion and Diffusion Processes**), the evolution of the system is described by

$$\begin{aligned}\mathrm{d}X_1 &= (\lambda - \mu X_1 - \beta X_1 X_4)\,\mathrm{d}t \\ &\quad - \sqrt{(\beta X_1 X_4)}\,\mathrm{d}W \end{aligned} \tag{10A}$$

$$\begin{aligned}\mathrm{d}X_2 &= [\beta p X_1 X_4 - (\mu + \alpha)X_2]\,\mathrm{d}t \\ &\quad + \sqrt{(\beta p X_1 X_4)}\,\mathrm{d}W \end{aligned} \tag{11A}$$

$$\begin{aligned}\mathrm{d}X_3 &= [\beta(1 - p)X_1 X_4 + \alpha X_2 - aX_3]\,\mathrm{d}t \\ &\quad + \sqrt{(\beta(1 - p)X_1 X_4)}\,\mathrm{d}W \end{aligned} \tag{12A}$$

$$\begin{aligned}\mathrm{d}X_4 &= (cX_3 - \gamma X_4 - \beta' X_1 X_4)\,\mathrm{d}t \\ &\quad - \sqrt{(\beta' X_1 X_4)}\,\mathrm{d}W. \end{aligned} \tag{13A}$$

The model parameters are as defined above for the deterministic model. The prime on β indicates that this parameter has different units from those of β. Note that there is only one primary Wiener process, not four as portrayed in [10] as each component carries the same one, derived from the Poisson process N.

For the four-component diffusion process (10A)–(13A), the transition probability density function $P(\mathbf{y}, t; \mathbf{x}, s)$, $s < t$, where $\mathbf{y}$ is a 4-vector of forward variables and $\mathbf{x}$ is a 4-vector of corresponding backward variables, satisfies the following backward Kolmogorov equation [5]

$$\frac{\partial P}{\partial s} + L_{\mathbf{x}} P = 0, \tag{14}$$

where the operator $L_{\mathbf{x}}$ is defined through

$$\begin{aligned} L_{\mathbf{x}} &= \left[\lambda - \mu x_1 - \beta x_1 x_4\right]\frac{\partial}{\partial x_1} \\ &\quad + \left[\beta p x_1 x_4 - (\mu + \alpha)x_2\right]\frac{\partial}{\partial x_2} \\ &\quad + \left[\beta(1 - p)x_1 x_4 + \alpha x_2 - a x_3\right]\frac{\partial}{\partial x_3} \\ &\quad + \left[c x_3 - \gamma x_4 - \beta' x_1 x_4\right]\frac{\partial}{\partial x_4} \\ &\quad + x_1 x_4 \Bigg[\frac{1}{2}\left\{\beta\frac{\partial^2}{\partial x_1^2} + \beta p\frac{\partial^2}{\partial x_2^2}\right. \\ &\quad \left. + \beta(1 - p)\frac{\partial^2}{\partial x_3^2} + \beta'\frac{\partial^2}{\partial x_4^2}\right\} \\ &\quad + \beta\sqrt{p}\frac{\partial^2}{\partial x_1 \partial x_2} + \beta\sqrt{1 - p}\frac{\partial^2}{\partial x_1 \partial x_3} \\ &\quad - \sqrt{\beta\beta'}\frac{\partial^2}{\partial x_1 \partial x_4} + \beta\sqrt{p(1 - p)}\frac{\partial^2}{\partial x_2 \partial x_3} \\ &\quad - \sqrt{\beta\beta' p}\frac{\partial^2}{\partial x_2 \partial x_4} - \sqrt{\beta\beta'(1 - p)}\frac{\partial^2}{\partial x_3 \partial x_4}\Bigg]. \end{aligned}$$

For the diffusion model, some simulated sample paths obtained with a strong Euler scheme for X_1, X_2, and X_4 are shown in Figure 1. Good agreement with the time course and variability of the acute phase of HIV-1 infection is found. In addition, it was found useful to find the times at which the virion density attains levels corresponding to the thresholds for detection of the virus in plasma samples; a typical distribution is shown in the bottom part of Figure 1. Such results are useful in ascertaining the risks in tests of blood donations for infection by HIV [11].

It is possible to find the properties of the distribution of the time to detection by using first passage time theory for diffusion processes, which results in the following analytical framework. Let the threshold level of detection of the virus be θ/mm^3. Let A be a set in R^4 containing the initial value $\mathbf{x}$ of the process such that $x_4 \in (0, \theta)$. In particular, let $A = (x_1', x_1'') \times [0, x_2'') \times [0, x_3'') \times (0, \theta)$. Then, we consider the time to detection as the first exit time, $T_\theta(\mathbf{x})$,

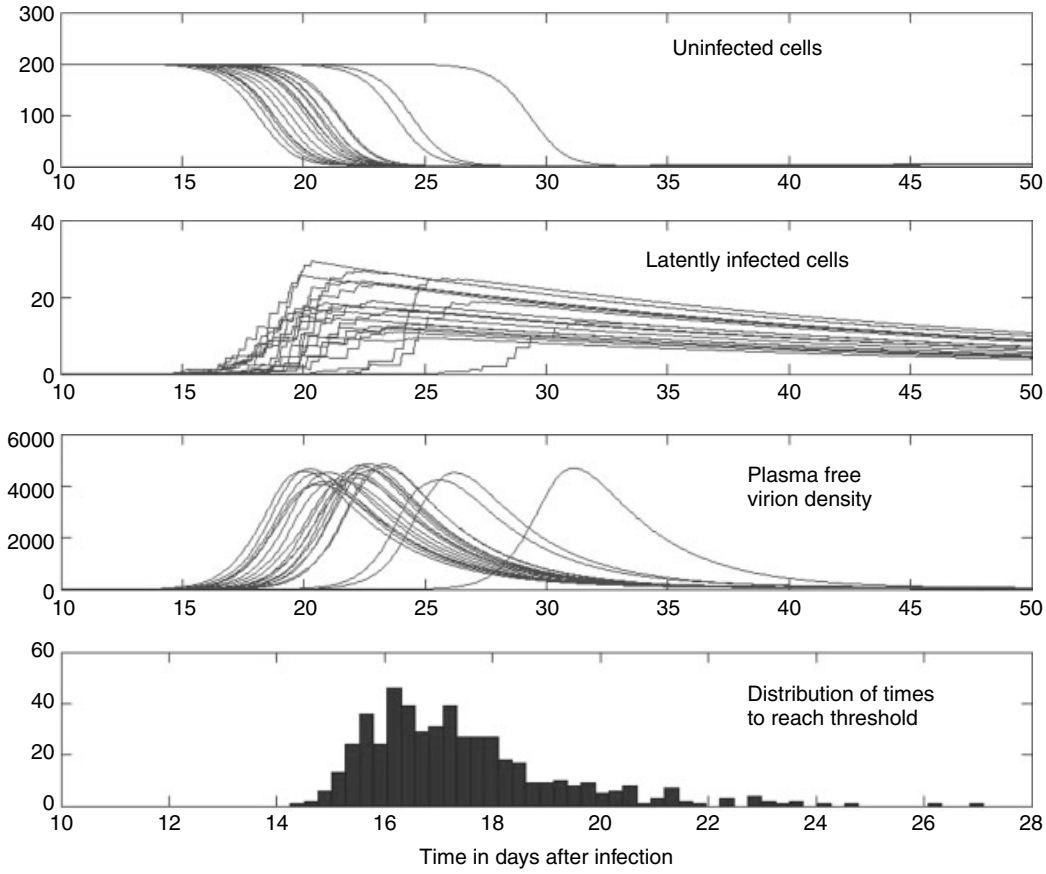

Figure 1 Numerical solutions of the stochastic differential equations (10A)–(13A), showing 20 sample paths for the components X_1, activated uninfected cells, X_2, latently infected cells, and X_4, free plasma virus. For more details, see Tuckwell and Le Corfec [28]. An estimate of the distribution of the time to reach an assumed detection threshold (100 virions/cubic millimeter) is also shown, based on 500 trials

of the process (X_1, X_2, X_3, X_4) from A. Here, x_1' is chosen small enough and both x_2'' and x_3'' are chosen large enough so that escapes through $x_1 = x_1'$ or $x_2 = x_2''$ or $x_3 = x_3''$ are extremely unlikely. Also, the actual initial value of X_1 must be less than x_1''.

The distribution function of this quantity, $F_{\boldsymbol{\theta}}(\mathbf{x}; t) = Pr\{T_\theta(\mathbf{x}) \leq t\}$, satisfies

$$\frac{\partial F_{\boldsymbol{\theta}}}{\partial t} = L_{\mathbf{x}} F_{\boldsymbol{\theta}}, \tag{15}$$

with initial condition $F_{\boldsymbol{\theta}}(\mathbf{x}; 0) = 0$, if $\mathbf{x} \in A$ and $F_{\boldsymbol{\theta}}(\mathbf{x}; 0) = 1$, if $\mathbf{x} \notin A$, with boundary condition $F_{\boldsymbol{\theta}}(\mathbf{x}; t) = 1$, $\mathbf{x} \notin A, t \geq 0$. Furthermore, the **moments** $\mu_n = E\left[T_\theta^n(\mathbf{x})\right]$, $n = 1, 2, \ldots$, satisfy the recursive system

$$L_{\mathbf{x}} \mu_n = -n\mu_{n-1}, \tag{16}$$

for $\mathbf{x} \in A$, with boundary conditions $\mu_n(\mathbf{x}) = 0, \mathbf{x} \in \partial A$. Here, $\mu_0 = 1$ is the probability of ever leaving A. There may be some escape of probability mass at zero virion level but this is expected to be insignificant compared to that associated with paths, which attain level θ, so $T_\theta(\mathbf{x})$ will be very close to the time to detection.

Approximation for Small Times. The above four-component framework can be simplified to a two-component one at early times by not distinguishing

between latently and actively infected CD4+ T-cells, as in [7], and by considering the number of uninfected CD4+ T-cells as constant. Neglecting also the interaction term in the viral dynamical equation, one obtains a simplified stochastic model for the very early (less than 15 days) period of HIV-1 dynamics [30].

Long-term Stochastic HIV Model. The mathematical model of Tan and Wu [26] has the same four components as above, but the dynamics of the uninfected cells are more complicated. These are generated by a means of a (possibly temporally nonhomogeneous) Poisson process with rate $s(t)$ (replacing the constant λ), the rate declining as free virion level increases. Furthermore, these cells are stimulated by HIV and antigen to produce new X_1 by a stochastic logistic birth process with rate $r(t) = r_0[1 - (X_1 + X_2 + X_3)/T_{\max}]$, where r_0 is a constant and $T_{\max}$ is the saturating level of host cells in all stages. The proportion of infected cells that become latent is possibly time dependent and given by $\omega(t)$ and the number of virions released by an actively infected cell is (possibly) random and given by $M(t)$. Furthermore, the three kinds of host cells and free HIV die according to simple death processes with rates μ_1, μ_2, μ_3 and μ_4, respectively (*see* **Stochastic Processes**). The stochastic equations take the form

$$\mathrm{d}X_1 = s(t)\,\mathrm{d}t + r(t)X_1\,\mathrm{d}t - X_1[\mu_1 + k_1X_4]\,\mathrm{d}t + \varepsilon_1(t)\,\mathrm{d}t \tag{17}$$

$$\mathrm{d}X_2 = \omega(t)k_1X_1X_4\,\mathrm{d}t - X_2[\mu_2 + k_2]\,\mathrm{d}t + \varepsilon_2(t)\,\mathrm{d}t \tag{18}$$

$$\mathrm{d}X_3 = [1 - \omega(t)]k_1X_1X_4\,\mathrm{d}t + k_2X_2\,\mathrm{d}t - \mu_3X_3\,\mathrm{d}t + \varepsilon_3(t)\,\mathrm{d}t \tag{19}$$

$$\mathrm{d}X_4 = M(t)\mu_3X_3\,\mathrm{d}t - k_1X_1X_4\,\mathrm{d}t - \mu_4X_4\,\mathrm{d}t + \varepsilon_4(t)\,\mathrm{d}t. \tag{20}$$

The constants k_1 and k_2 are the interaction rate between HIV and uninfected cells and the transition rate from X_2 to X_3. Note that here virions are released (only) whenever an X_3 cell "dies". The terms ε_k, $k = 1, \ldots, 4$, are "random noises". The means of the X_k's are not the same as the deterministic model. By means of simulation, Tan and Wu were able to distinguish three regimes: (a) an early infection period; (b) a transition period; and (c) a steady state period. Using parameters estimated from patient data, in both their simple and complex models, they found there was a positive probability of approaching a noninfected state ($X_3 = X_4 = 0$), even in the absence of drug treatments.

Models with Mutant Virus Strains

It is important to address the effects of the appearance of mutant strains of virus, which may lead to ineffective or less effective drug treatments [1, 16, 18, 22]. Abundo and Rossi [1] consider a multicomponent diffusion process model. In addition to distinguishing viral strains, $V_i, i = 1, \ldots, n$, each of which has its own unique growth rate r_i, it was assumed that there are associated specialized classes of CD4+ T-cells, X_i. There are also an uninfected cell population Y and a nonspecific class of activated cells, X. A virus of any strain may attack host cells. The drift terms correspond to the deterministic model of Nowak *et al.* [19]. The process of viral mutation does not appear explicitly in the model, which seems undesirable. However, in the simulated solutions of the diffusion model, explosions of viral load eventually occur, accompanied by the collapse of the immune system. For a perspective on the ramifications for strategies of drug treatment, see [22].

Statistical Models and Drug Treatments

Using simple differential equations, similar to those in the section "Deterministic Models", modified to incorporate the effects of drug treatments, patient data on CD4 cell counts and HIV loads may be used to estimate model parameters from explicit solutions. For example, using **nonlinear regression** analysis, the viral clearance rate constant, and the rate of loss of virus-producing cells were estimated for each of a group of patients [20]. However, as there is presently no suitable long-term HIV model, Wu and Zhang [32] have applied a class of semiparametric nonlinear mixed-effects models developed for longitudinal data [33] (*see* **Random Coefficient Repeated Measures Model**). At the population level, combination antiretroviral therapies (ARV) are widely used to treat HIV, but drug-resistant strains have quickly evolved and the overall impact that ARV will have on HIV epidemics remains unclear. Velasco-Hernandez et al. [31] used a mathematical model to determine the effectiveness of current therapies in reducing the severity of HIV epidemics. They claimed that even

a high-prevalence HIV epidemic could be eradicated using current ARV.

Conclusions

Mathematical models of viral dynamics are potentially very useful for understanding the progression and treatment of virally induced diseases. Simple statistical models and the incorporation of mutant viral strains can lead to optimized drug treatments for HIV, which ameliorate the disease but do not eliminate it. The stochastic modeling of viral dynamics is in its infancy, and existing models omit many important agents and properties of both virus, host cell and immune response.

References

[1] Abundo, M. & Rossi, C. (1994). A stochastic model of the impact of genetic variability of HIV on the immune system in infected patients, in *Modeling the AIDS epidemic: Planning, Policy and Prediction* E.H. Kaplan, & M.L., Brandeau, eds. Raven, New York, pp. 481–498.

[2] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Application*. Griffin, London.

[3] Bocharov, G.A. & Romanyukha, A.A. (1994). Mathematical model of antiviral immune response III. Influenza A virus infection, *Journal of Theoretical Biology* **167**, 323–360.

[4] Feller, W. (1952). The parabolic differential equations and the associated semigroups of transformations, *Annals of Mathematics* **55**, 468–519.

[5] Gihman, I.I. & Skorohod, A.V. (1972). *Stochastic Differential Equations*. Springer-Verlag, Berlin.

[6] Henderson, D.A., Inglesby, T.V., Bartlett, J.G., Ascher, M.S., Eitzen, E., Jahrling, P.B., Hauer, J., Layton, M., McDade, J., Osterholm, M.T., O'Toole, T., Parker, G., Perl, T., Russell, P.K., Tonat, K. (1999). Smallpox as a biological weapon: medical and public health management, *Journal of American Medical Association* **281**, 2127–2137.

[7] Herz, A.V.M., Bonhoeffer, S., Anderson, R.M., May, R.M. & Nowak, M.A. (1996). Viral dynamics *in vivo*: limitations on estimates of intracellular delay and virus decay, *Proceedings of the National Academy of Sciences USA* **93**, 7247–7251.

[8] Hethcote, H.W. (2000). The mathematics of infectious diseases, *SIAM Review* **42**, 599–653.

[9] Holmes, K.V. (2003). SARS-associated corona virus, *New England Journal of Medicine* **348**, 1946–1951.

[10] Kamina, A., Makuch, R.W. & Zhao, H. (2002). A stochastic modeling of early HIV-1 population dynamics, *Mathematical Biosciences* **170**, 187–198.

[11] Le Corfec, E., Le Pont, F., Tuckwell, H.C., Rougioux, C. & Costagliola, D. (1999). Direct HIV testing in blood donations: variation of the benefit with detection threshold and pool size, *Transfusion* **39**, 1141–1144.

[12] McLean, A.R., Emery, V.C., Webster, A. & Griffiths, P.D. (1991). Population dynamics of HIV within an individual after treatment with zidovudine, *AIDS* **4**, 374–378.

[13] Merrill, S. (1989). Modeling the interaction of HIV with the cells of the immune system, *Mathematical and Statistical Approaches to AIDS Epidemiology, Lecture Notes in Biomath*, Vol. 83, Springer-Verlag, New York.

[14] Nash, T. (2001). Immunity to viruses, in *Immunology*, I. Roit, J. Brostoff, & D. Male, eds. Mosby, Edinburgh, Chapter 14 235–244.

[15] Neumann, A.U., Lam, N.P. & Dahari, H. (1998). Hepatitis C viral dynamics *in vivo* and the antiviral efficiency of interferon-alpha therapy, *Science* **282**, 103–107.

[16] Nowak, M.A. (1992). Variability of HIV infections, *Journal of Theoretical Biology* **155**, 1–20.

[17] Nowak, M.A. (1999). The mathematical biology of human infections, *Conservation Ecology* **3**, 12–23.

[18] Nowak, M.A. & May, R.M. (2000). *Virus Dynamics*. Cambridge University Press, Cambridge UK.

[19] Nowak, M.A., May, R.M. & Anderson, R.M. (1990). The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease, *AIDS* **4**, 1095–1103.

[20] Perelson, A.S. & Nelson, P.W. (1999). Mathematical analysis of HIV-1 dynamics *in vivo*, *SIAM Review* **41**, 3–44.

[21] Phillips, A.N. (1996). Reduction of HIV concentration during acute infection: independence from a specific immune response, *Science* **272**, 497–499.

[22] Phillips, A.N., Youle, M., Johnson, M. & Loveday, C. (2001). Use of a stochastic model to develop understanding of the impact of different patterns of antiretroviral drug use on resistance development, *AIDS* **15**, 2211–2220.

[23] Stafford, M.A., Corey, L. & Cao, Y. (2000). Modeling plasma virus concentration during primary HIV infection, *Journal of Theoretical Biology* **203**, 285–301.

[24] Stilianakis, N.I., Dietz, K. & Schenzle, D. (1997). Analysis of a model for the pathogenesis of AIDS, *Mathematical Biosciences* **147**, 27–46.

[25] Stumpf, M.P.H., Laidlaw, Z. & Jansen, V.A.A. (2002). Herpes viruses hedge their bets, *Proceedings of the National Academy of Sciences USA* **99**, 15234–15237.

[26] Tan, W.-Y. & Wu, H. (1998). Stochastic modeling of the dynamics of $CD4^+$ T-cell infection by HIV and some Monte Carlo studies, *Mathematical Biosciences* **147**, 173–205.

[27] Tuckwell, H.C. & Le Corfec, E. (1998). A stochastic model of early HIV-1 population dynamics, *Journal of Theoretical Biology* **195**, 451–463.

[28] Tuckwell, H.C., Toubiana, L. & Vibert, J.-F. (2001). Epidemic spread and bifurcation effects in two-dimensional

network models with viral dynamics, *Physical Review E* **64**, 0419181–0419188.

[29] Tuckwell, H.C. & Wan, F.Y.M. (2000a). Nature of equilibria and effects of drug treatments in some viral population dynamical models, *IMA J Math. Appl. Biol. Med.* **17**, 311–327.

[30] Tuckwell, H.C. & Wan, F.Y.M. (2000b). First passage time to detection in stochastic population dynamical models for HIV-1, *Applied Mathematics Letters* **13**, 79–83.

[31] Velasco-Hernandez, J.X., Gershengorn, H.B. & Blower, S.M. (2002). Could widespread use of combination antiretroviral therapy eradicate HIV epidemics?, *The Lancet Infectious Diseases* **2**, 487–493.

[32] Wu, H. & Zhang, J.T. (2002a). The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models, *Statistical Medicine* **21**, 3655–3675.

[33] Wu, H. & Zhang, J.T. (2002b). Local polynomial mixed effects models for longitudinal data, *Journal of American Statistical Association* **97**, 883–897.

Further Reading

Le Corfec, E. & Tuckwell, H.C. (1998). Variability in early HIV-1 population dynamics, *AIDS* **12**, 960–962.

HENRY C. TUCKWELL

Vital Statistics, Overview

Vital statistics, as a scientific discipline, is a subdomain of **demography**, the study of the characteristics of human populations. Vital statistics comprises a number of important events in human life including birth, death, fetal death, marriage, divorce, annulment, judicial separation, adoption, legitimation, and recognition. The term "vital statistics" is also applied to individual measures of these vital events. Thus, a birth rate is an example of a vital statistic and an analysis of trends in birth rates is an example of an application in the field of vital statistics. A vital statistics system is the total process of collecting by civil registration, enumeration, or indirect estimation, information on the frequency of occurrence of vital events, selected characteristics of the events and the persons concerned, and the compilation, analysis, evaluation, and dissemination of these data in summarized statistical form. Other life events of demographic importance such as change of place of residence (migration), change of citizenship (naturalization), and change of name are not included, mainly because information on these is usually derived from other statistical systems such as population registers [1].

Systems for Collecting Vital Statistics

It is generally accepted that the preferred method for individual countries to collect vital statistics is through a civil registration system. This is recognized by the United Nations (UN) and other international organizations, as well as by the many countries that have had civil registration laws and regulations in place and in operation for many years [1, 2]. Nevertheless, a number of newly emergent and developing nations, facing the difficulties and length of time it takes to create a satisfactory civil registration system, have instituted alternative procedures to acquire statistical data to describe the levels and trends for key vital events, particularly for fertility and mortality measurements. The UN recognizes the importance of a civil registration system for each country as the preferred source of vital statistics data for the long run. However, use of an alternative data collection system is recommended as an interim measure for meeting needs for essential information where a civil registration system of acceptable quality does not yet exist. Other systems include, for example, probability area samples (*see* **Probability Sampling**), purposeful area samples, records-based **surveys**, and **record linkage**. Furthermore, the UN recommends a priority order for the types of vital statistics data to be collected. The highest need is given to data on births and deaths, followed in order by marriages, divorces, fetal deaths, annulments, judicial separations, adoptions, legitimations, and recognitions [1].

Uses of Vital Records and Vital Statistics

Vital records created through a civil registration system have two classes of use. They have value individually as legal documents for the persons named thereon; they also constitute the input, when aggregated, for the various vital statistics measures that are used to study the demographics and health of populations and population subgroups.

For the individual, a birth record is a legal document establishing name, parentage, birth data, order of birth for multiple births, legitimacy, and citizenship, nationality, or geographic place of birth. A wide variety of individual rights and civil entitlements depends on these facts, including proof of age for school entrance, motor vehicle drivers' licenses, military service and other age-related activities, establishment of eligibility for family allowances, insurance benefits, tax benefits, inheritance rights, issuance of passports, etc. The death record provides documentary proof of the facts of death needed for social security and insurance purposes such as time and place of death and the medical cause of death. Proof of death and the associated facts are also used for property inheritance rights, for remarriage rights of surviving spouses, etc. Marriage and divorce records serve to document rights to special social and economic programs and benefits for the married, including tax privileges for couples, alimony, change of nationality based on marriage, and the right to remarry. Many rights of children, their parents, and their guardians are dependent on records of adoption, legitimation, and recognition.

Individual vital records may also be used administratively as the basis for initiating maternal and child health services, including child immunization programs, or for epidemiologic investigations into disease outbreaks or assessments of causes of accidents and injuries. Another important administrative

use of individual vital records especially of death records (*see* **Death Certification**), is for the updating or clearing of files such as electoral rolls, social security files, **disease registers**, cohort follow-up studies, tax registers, etc.

In aggregated form, vital records become a collection of vital statistics, most often in the form of **means**, **medians**, and various ratios such as proportions and **rates**. Whether collected by civil registration or by other means, vital statistics serve as key demographic variables in the analysis of population size, growth and geographic distribution, especially when used in conjunction with periodic population **censuses**. When census data are used as a base, current intercensal estimates of population size can be made, and projections into the future can be prepared using estimates of future trends in fertility, natality, and mortality linked with estimates of net migration. In addition to the importance of vital statistics to the study of population size and growth trends, other national and subnational economic and social concerns such as health, welfare, education, occupation, housing, urbanization, family structure, and income are also affected by these measures. In the fields of public health and medicine, for example, levels and trends of **infant and perinatal mortality** are often used as surrogate measures of levels and trends in the overall health and well-being of nations. **Life expectancy** at birth is also frequently used to compare the overall effects of mortality and its determinants. **Cause of death** information provides a foundation upon which much research into diseases and disease prevention is based.

Differentials in mortality by sex, age, racial groups, and other variables are often the basis for the planning of health and medical intervention programs. In addition, the planning and provision of public and private housing, educational facilities, social security and private insurance plans, medical facilities, and consumer goods of all kinds are examples of activities dependent on vital statistics data. At the international level, vital statistics provide a basis for comparing important demographic, social, and economic differences and trends over time among countries or regions of the world.

Definitions of Selected Vital Events

Standard statistical definitions of vital events have been promulgated by international agencies [1, 5]. In some cases, legal definitions may differ from the international standards in varying degrees, but, in many cases, national vital statistics reports are either based on the standard statistical definitions or do not differ in principle. In cases where comparability among countries is compromised because of the use of nonstandard definitions, international agencies and others presenting national comparisons of tabular, graphical or descriptive vital statistics usually provide appropriate cautions to users. Nevertheless, users of vital statistics data need to ascertain the comparability of the data before drawing reliable conclusions about national differences. The **World Health Organization (WHO)** promulgates a number of vital statistics definitions as part of the **International Classification of Diseases (ICD)**. These definitions are incorporated in regulations adopted by the World Health Assembly and which each WHO member country has agreed to follow [4]. Nevertheless, it is still necessary to ensure that the standard definitions have been followed for a given data set. The international standard definitions for selected vital events are given below.

Live Birth. This is the complete expulsion or extraction from its mother of a product of conception, irrespective of the duration of the pregnancy, which, after such separation, breathes or shows any other evidence of life, such as beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles, whether or not the umbilical cord has been cut or the placenta is attached; each product of such a birth is considered liveborn [5].

Fetal Death. This is death prior to the complete expulsion or extraction from its mother of a product of conception, irrespective of the duration of pregnancy; the death is indicated by the fact that after such separation the fetus does not breathe or show any other evidence of life, such as beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles [5].

Maternal Death. This is the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and the site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes. Maternal deaths may be subdivided into two groups: direct obstetric

deaths which are the result of obstetric complications of the pregnant state (pregnancy, labor, and the puerperium), from interventions, omissions, incorrect treatment, or from a chain of events resulting from any of these; and indirect obstetric deaths which are the result of previously existing disease or disease that developed during pregnancy and which was not due to direct obstetric causes, but which was aggravated by physiologic effects of pregnancy [5].

Infant Death. This is the death of a liveborn infant who dies before completing its first year of life.

Neonatal Death. This is the death of a liveborn infant who dies during the first 28 completed days of life. These may be subdivided into early neonatal deaths, occurring during the first seven days of life, and late neonatal deaths, occurring after the completion of the seventh day but before the completion of 28 days [5].

Perinatal Death. This is the death of a fetus or newborn infant occurring after 22 completed weeks (154 days) of gestation (the time when fetal weight is normally about 500 g), but prior to the completion of seven days after birth [5].

Marriage. This is the act, ceremony or process by which the legal relationship of husband and wife is constituted. The legality of the union may be established by civil, religious, or other means recognized by the laws of each country [1].

Divorce. This is a final legal dissolution of a marriage which confers on the parties the right to remarriage under civil, religious, or other provisions, according to the laws of each country [1].

Definitions of Selected Vital Statistics Measures

Raw vital statistics most often are comprised of counts of how often a specified vital event has occurred, rather than on measurements of continuous variables such as height, weight, or blood pressure. The analysis of vital data depends mainly on the conversion of observed frequencies into indices, ratios, and probabilities. Counts of vital events often do have utility, but, for the majority of uses, absolute frequencies are not sufficient and it becomes necessary to compute relative numbers, including crude rates, various types of specific rates, percentages, probabilities, and other ratios.

Some of the more commonly encountered vital statistics relative numbers are defined and calculated as follows.

Crude Death Rate

The most common form of mortality measurement is the crude death rate. It is computed from the following formula [3]:

$$m_{\mathrm{cd}} = \left(\frac{D}{P}\right) k,$$

where m_{cd} is the crude death rate, D is the total number of deaths for a given area and time period, usually a calendar year, P is the size of population at risk of dying, usually taken as the estimated population at the midpoint of the calendar year, and k is a constant, usually taken as 1000.

The crude rate is so named to differentiate it from various specific and adjusted rates and represents the total or overall death rate without regard to the various component elements which combine to produce the total figure. The crude death rate is usually expressed as "the number of deaths per 1000 persons" for a specified place (country, city, state, etc.) for a given year.

Specific Death Rate

Detailed analyses of vital statistics frequently go beyond the overall risk of death in the population as a whole. Many studies deal with subsets of the population or with particular classes of deaths. Epidemiologists often focus on deaths from a particular disease or class of diseases. Actuaries and demographers are concerned with differences in mortality by sex and in different age groups within the population. Environmental and **occupational health** specialists are interested in the differential risks of dying in selected occupations, and in different geographic subdivisions such as urban and rural areas. To meet these kinds of needs, various specific death rates are calculated. Specific rates for different age groups are called *age-specific death rates*; rates for males and females

are called *sex-specific death rates*, rates for particular causes of death are called *cause-specific death rates*. Rates may be specific for combinations of characteristics. For example, age–sex–race-specific death rates are computed separately for each age group by race and sex. Specific death rates are approximations of true probabilities. That is, the denominator of the ratio is an estimate of the total number of events of a particular type that could happen, while the numerator is a count of those that did happen.

Specific death rates are computed as follows [3]:

$$m_{\mathrm{sd}} = \left(\frac{d_i}{p_i}\right) k,$$

where m_{sd} is the specific rate for any defined ith class, d_i is the number of deaths occurring in the ith class for a given area and time, p_i is the number of persons in the ith class of the population for the same area and time, and k is a constant, usually 100 000.

For cause-specific death rates, the denominator, p_i, in the above formula is replaced by P, the total population exposed to the risk of death. Therefore, a cause-specific death rate measures the risk in the total population of dying from a specified cause of death.

Infant Mortality Rate

The infant mortality rate is considered by many as one of the important indicators of the overall level of health and social well-being of a country or other geopolitical area. This is, in part, because a large proportion of deaths in the first year of life are considered to be preventable through adequate prenatal care, good nutrition for women and infants, and improved control of the environment, including injury prevention.

The infant mortality rate is computed as follows [3]:

$$m_i = \left(\frac{d_{<1}}{B}\right) k,$$

where m_i is the infant mortality rate, $d_{<1}$ is the number of deaths to liveborn infants under one year of age during a specified time period, usually one year, B is the total number of live births during the same time period, and k is a constant, usually 1000.

The infant mortality rate is a proxy for the age-specific death rate for the "under one year of age group" and is intended to be a measure of the risk of dying during the first year of life. The numerators of the infant mortality rate and the "under one year of age" age-specific death rate are the same. For a denominator, however, a reliable estimate of the size of the population under one year of age for a given time period is hard to obtain, even in a census year. As a proxy measure, the denominator may be considered to be the number of births occurring during the period. For either of these choices of denominator, there is some mismatch with the numerator in terms of a true probability number. Not all events in the numerator arise from the events in the denominator. For example, in the infant mortality rate, some of the deaths under one year of age in a given year and counted in the numerator were actually born in the previous year and are not represented in the denominator, while some of the births represented in the denominator will die before their first birthday but the deaths will occur in the next year and are not included in the numerator. However, when the birth rate is fairly stable from one year to the next, calculation of the infant mortality rate results in a ratio that closely approximates the probability of a live-born infant dying within the first year of life. When the birth rate is not stable from year to year, a more accurate mortality rate may be computed by following each live birth occurring during a one year period and measuring how many of them die before their first birthday.

Neonatal, Early Neonatal and Postneonatal Mortality Rates

The *neonatal mortality rate* is defined as follows [5]:

$$m_{\mathrm{n}} = \left(\frac{d_{<1\ \mathrm{mo}}}{B}\right) k,$$

where m_{n} is the neonatal mortality rate, $d_{<1\ \mathrm{mo}}$ is the number of deaths of infants under 1 month of age during a specified time period, B is the number of live births occurring during the same time period, and, k is a constant, usually 1000.

The neonatal mortality rate, like the infant mortality rate, is a proxy for an age-specific death rate. It approximates the risk of dying in the first month of life. The relative importance of an infant mortality rate compared with the corresponding neonatal mortality rate depends on the proportionate age distribution of the deaths under one year of age. Generally, when the infant mortality rate is low, a large proportion of infant deaths occur during the first month

of life. The neonatal mortality rate then reflects an important measure of the mortality risk for infants. Conversely, when the infant mortality rate is high, larger proportions of deaths fall into the older age groups under a year. Often it is useful to partition the deaths of infants under one year of age into two groups: those dying before one month of age, and those dying between one month and their first birthday. The former comprise the numerator, $d_{<1\ \mathrm{mo}}$, of the neonatal mortality rate, while the latter can be used to calculate the *postneonatal mortality rate*:

$$m_{\mathrm{pn}} = \left(\frac{d_{1\ \mathrm{mo}-1\ \mathrm{yr}}}{B} \right) k,$$

where m_{pn} is the postneonatal mortality rate, $d_{1\ \mathrm{mo}-1\ \mathrm{yr}}$ is the number of deaths occurring between 1 month and 1 year of age during a specified time period, B is the number of live births occurring during the same time period, and k is the same constant used in the neonatal mortality rate, usually 1000.

In similar fashion, the neonatal deaths may be partitioned into those dying within the first week of life and the remainder that survive the first seven days but die before one month of age. The risk of dying in the first week of life is measured by the *early neonatal mortality rate*, m_{en}, as follows [5]:

$$m_{\mathrm{en}} = \left(\frac{d_{<7\ \mathrm{days}}}{B} \right) k,$$

where the components of the calculation are the same as in the neonatal and postneonatal mortality rates, except that the numerator contains only those deaths to infants occurring during the first week of life.

Perinatal Mortality Rate

The perinatal period, as defined earlier, is the period of time surrounding the event of birth. It includes the time that a fetus spends in utero after it has reached 22 weeks of gestation and continues through the birth process until the end of the first week of life after birth. The perinatal mortality rate measures mortality occurring during this period. The rate, therefore, combines deaths of fetuses of specified **gestational age** with deaths of liveborn infants who die in their first week of life. The determination of whether a fetus is born dead or whether it shows any sign of life before expiring is not always clear-cut; social, economic, and cultural factors, as well as medical and biological considerations, tend to push the fetal death rate in one direction or the other in different societies, thus making comparisons of neonatal or infant mortality among countries difficult. By using the perinatal mortality rate for comparisons, this difficulty is minimized since fetuses dying just before or during the birth process as well as those born alive but dying shortly thereafter are all included in the calculation [5]:

$$m_{\mathrm{peri}} = \left[\frac{d_{\mathrm{peri}}}{F + B} \right] k,$$

where m_{peri} is the perinatal mortality rate, d_{peri} is the number of deaths of fetuses of 22 or more weeks of gestation plus deaths of liveborn infants of less than 7 days of age during a specified period, usually a calendar year, F is the number of fetal deaths of 22 or more weeks of gestation during the same period, B is the number of live births during the same period, and, k is a constant, usually 1000.

Note that, unlike the infant and neonatal mortality rates, the denominator of the perinatal mortality rate combines both the number of live births and the number of fetal deaths of 22 or more weeks of gestation. This denominator is called "total births" and better approximates the population from which the numerator could arise than would a denominator restricted to only live births. On the other hand, it is recognized that it is easier to collect reliable counts of live births than of fetal deaths, thus introducing another source of error into the calculation of the perinatal mortality rate.

Maternal Mortality Rate

The **maternal mortality** rate is calculated as follows [5]:

$$m_{\mathrm{m}} = \left[\frac{d_{\mathrm{md}} + d_{\mathrm{mi}}}{B} \right] k,$$

where m_{m} is the maternal mortality rate, d_{md} is the number of direct maternal deaths in a specified time period, usually 1 year, d_{mi} is the number of indirect maternal deaths in the same period, B is the number of live births in the same period, and k is a constant, usually 10 000 or 100 000.

A related measure, the *direct obstetric mortality ratio*, may be calculated from the above formula

but using in the numerator only the direct maternal deaths, d_{md}.

Proportionate Mortality

Proportionate mortality, sometimes known as the death ratio (*see* **Proportional Mortality Ratio (PMR)**), is defined as [3]:

$$p_d = \left(\frac{d_i}{D}\right) k,$$

where p_d is the proportionate mortality, d_i is the number of deaths in a specified class during a stated time period, D is the total number of deaths in the same time period, and k is a constant, usually 100 or 1000.

Proportionate mortality ratios may be calculated for any class of deaths, but their most common uses are for given causes or group of causes of death expressed as percentages of deaths from all causes, or for deaths at a specified age expressed as percentages of deaths at all ages.

Crude Birth Rate

The crude birth rate is the most frequently used overall measure of the reproduction of a population. Like its counterpart, the crude death rate, it is influenced by many factors and represents a proxy for more specific fertility measurements. It is calculated as follows [3]:

$$m_{cb} = \left(\frac{B}{P}\right) k,$$

where m_{cb} is the crude birth rate, B is the total number of live births for a given area and time period, P is the total population at the midpoint of the time period, and, k is a constant, usually 1000.

Comparing Vital Statistics Data

Aggregated vital statistics data, whether in tabular or graphical form, often appear as time trends for particular variables such as causes or groups of causes of death, or for age and sex groups of the population. They also appear frequently as comparisons between countries or other geographical entities for a point in time, usually a particular year. In either case, great care must be taken to ensure that the quality of the data in the groups being compared warrants making the comparisons. In registration based systems, measures or estimates of completeness of reporting of vital events should be known. In sample based systems, the representativeness of the sample and the **nonresponse** rate is important. In the comparison of data between two or more geographic places, it is important to ascertain if common definitions and procedures were used to collect, process, analyze, and present the data; in looking at time trends, it is essential to know if the definitions of the events and the procedures for classifying the data remained constant over the entire time period being studied. This latter point is particularly important when looking at trends in causes of death since the instrument for grouping diseases into categories for study, the ICD, is revised approximately every 10 years (*see* **Morbidity and Mortality, Changing Patterns in the Twentieth Century**; **Mortality, International Comparisons**). Vital statistics data are often presented in statistical compendia published by official national and international organizations that attempt to include important notes for interpretation of the data in headnotes and footnotes to tables, appendices, etc. (*see* **Data Access, National and International**). The user is cautioned to pay careful attention to such explanatory or cautionary notes.

References

[1] United Nations (1973). Principles and Recommendations for a Vital Statistics System. *Statistical Papers, Series M, No. 19, Rev. 1*. United Nations, New York.

[2] United States Bureau of the Census (1971). *The Methods and Materials of Demography*, H. Shryock, J. Siegel et al., eds. US Government Printing Office, Washington.

[3] United States Department of Health, Education, and Welfare (1965). Techniques of Vital Statistics. Reprint of Chapters I-IV, *Vital Statistics Rates in the United States, 1900–1940*. National Center for Health Statistics, Washington.

[4] World Health Organization (1967). *WHO Nomenclature Regulations*. World Health Organization, Geneva.

[5] World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems: 10th revision*, 3 vols. World Health Organization, Geneva.

ROBERT A. ISRAEL

Wald, Abraham

Born: October 31, 1902, in Cluj, Rumania.
Died: December 13, 1950.

Abraham Wald emigrated to the US in 1938 and, in the period from 1938 till his death in a plane crash in 1950, he studied and revolutionized modern statistics. His most important contributions were in introducing **decision theory** and **sequential analysis**.

As the son of an orthodox Jew, he would not attend school on Saturday, the Jewish sabbath, and was not admitted to the local gymnasium (high school). He studied by himself with the help of an older brother, Martin, an electrical engineer, and was later admitted to the University of Cluj. After graduating, he spent a year in the engineering school at Vienna, and was finally admitted to the University of Vienna in the fall of 1927. After introducing himself to Karl Menger and expressing his interest in geometry, he spent some time serving in the Rumanian army instead of at the university.

In February 1930, Wald began to attend Menger's lectures and became part of an exciting group of young mathematicians taking part in an active mathematical colloquium. He received his Ph.D. in 1931. At this time of political and economic unrest, it was impossible to get a position at the university, and Menger recommended that he become involved with applied mathematics. Menger introduced Wald to Karl Schlesinger, a well-to-do banker and economist who wished to broaden his knowledge of higher mathematics. The association between Schlesinger and Wald led to publications on the existence of meaningful solutions for systems of equations of the theory of production, and on the cost of living index, and to contact with Oskar Morgenstern. Morgenstern was then director of the Institut für Konjunkturforschung, and later became a coauthor with von Neumann of their book on the theory of games [5]. He employed Wald at his institute and in the late 1930s Menger, Morgenstern, and Wald emigrated to the US.

Before emigrating, in the summer of 1938, to become a fellow of the Cowles Commission for Research in Economics at the University of Chicago, Wald also worked on a problem of the consistency of the von Mises concept of "Kollektiv", which lies at the heart of the von Mises axiomatization of probability.

In the fall of 1938, the Cowles Commission released Wald to accept a fellowship of the Carnegie Corporation, obtained for him by **Harold Hotelling** at Columbia University. Wald spent a busy year learning modern statistics by reading and attending Hotelling's lectures. At the same time, he started writing the publications on probability and statistics that are the foundation of his fame. The following year, he also began his career as an outstanding teacher of statistics at Columbia University. One of his earliest contributions in statistics, one considered by his frequent collaborator Jacob Wolfowitz to be his most important paper, was published in 1939, introducing decision theory [6].

On July 1, 1942, the Statistical Research Group (SRG) was formed to assist in the war effort. It was sponsored by Columbia University under the directorship of W. Allen Wallis. Shortly afterwards, another group was formed, the Statistical Research Group at Princeton, sponsored by Princeton University under the directorship of Samuel S. Wilks. In response to a question on sampling inspection raised by Captain Schuyler of the US Navy, Wallis and Milton Friedman proposed the use of sequential methods in sampling inspection. Failing to find a satisfactory resolution, they approached Wald in April 1943 [10]. At first, Wald was cool to the concept that seemed to violate a basic dogma of statistical **inference**, but the next day he changed his mind and produced the Sequential Probability Ratio Test, and derived some of its properties the day after that.

In 1941, Wald married Lucille Lang (who died with him in the plane crash, leaving two children, Betty and Robert born in 1943 and 1947). Columbia University recognized Wald's talent and quickly promoted him from Assistant Professor of Economics to Associate Professor and then to full Professor in 1944.

In 1946, Hotelling was recruited by **Gertrude Cox** to go to the University of North Carolina in support of her grand plan for the Institute of Statistics combining departments at the University and at North Carolina State. His departure led Columbia University to start a Statistics Department in the Faculty of Political Sciences with Wald as chair. The department had Wolfowitz and T.W. Anderson, Jr as regular appointees and, to supplement this small faculty, the department invited visitors to give courses. Among these visitors were **J. Neyman**, J.L. Doob, R.C. Bose, M.M. Loeve, E.J.G. Pitman,

and S.N. Roy. In the next few years, H. Scheffé and H. Levene were added to the regular faculty. In this postwar period, many relatively mature students, supported by the G.I. bill, returned to school, and the Statistics Department had a substantial collection of students who later became prominent in the field.

Much of the above text has been paraphrased from the more detailed articles by J. Wolfowitz, K. Menger, and G. Tintner, which were written in memorials for the *Annals of Mathematical Statistics* [3, 4, 12, 13].

The theoretical work of **R.A. Fisher**, suggested by applied problems, did much to accelerate the process of introducing mathematical considerations into theoretical statistics. The **Neyman–Pearson** theory helped to clarify many of the issues of inference that had hitherto been implicit, but never clearly articulated. These developments opened up a world of research for mathematical statisticians, the recognition of which led to the publication of the *Annals of Statistics* in 1930 and The Institute of Mathematical Statistics in 1933. When Wald arrived in the US, the number of talented mathematicians working in statistics was still quite small, but ready to explode. Hotelling had published a list of problems, the solution of which he felt would contribute much to the advancement of statistical theory. This was an environment well suited for talented mathematicians, and especially for those who had good statistical insights. Wald had the ability to know which problems were important, how to solve those that were tractable, and how to get good approximations for those that were not tractable.

The first paper on decision theory did not get an exceptionally good reaction, but one of his collaborators, Henry B. Mann, introduced it in a reading course in statistics in the summer of 1945 at Brown University. Wald returned to that subject in 1946. This formulation of the problem of statistical inference completed the work of Neyman and **Pearson**, by introducing cost considerations into hypothesis testing as well as in **estimation** and inference in general. It clearly points out one essential source of subjectivity in inference, since different investigators may very well have different cost functions. This formulation may be regarded as that of a game of the statistician against nature, with the difference that nature cannot be regarded as an active opponent. It has been suggested that this paper may have been influenced in part by a 1928 paper by von Neumann on game theory [11].

Although there was considerable resistance against this formulation, and, in particular, against Wald's tentative proposal of the **minimax** criterion, that resistance has died down, and it is difficult to deny the essential clarity it has contributed to the understanding of inference. The book on the subject [8] was not easy to read, since it dealt with some of the mathematical difficulties in deriving general results in this field.

Sequential analysis was immediately accepted by theoreticians. The initial results applied to the test of a simple hypothesis against a simple alternative (*see* **Hypothesis Testing**). Wald proved that the sequential probability ratio test (SPRT) would terminate with probability one and that the error probabilities and expected sample size under both hypotheses could be approximated very accurately. In typical applications, he obtained effective savings of about 50% in the sample size required to achieve given error probabilities. (In theory, for problems involving nearby alternatives and small error probabilities, this saving approaches 75%.) Wald conjectured that the SPRT was optimal, but was unable to prove that until his paper with Wolfowitz in 1948 [9] established a unique optimal property. The SPRT minimizes the expected sample sizes for the two hypotheses among all tests that achieve a given pair of error probabilities or better. The proof in this paper had some measure-theoretic difficulties that were resolved in a later paper by Arrow et al. [2], which was a foundation paper for dynamic programing. But the essential idea of relating the SPRT to the solution of a Bayes problem appeared in the Wald–Wolfowitz paper. In fact, the original derivation of the SPRT was based on a Bayesian approach, but Wallis convinced Wald to omit this from his publications, since the published results at that time did not require prior probabilities. Wald often used Bayesian arguments as part of his analysis to derive frequentist results. At the time, **Bayesian methods** were generally regarded as unacceptable.

While the main results were for testing a simple hypothesis vs. a simple alternative, Wald extended the method for testing composite hypotheses by introducing the notion of an indifference region. As with many of his approximations, this device worked remarkably well for the typical problems dealt with by most experimenters. However, it failed to address

the real problem in a satisfactory fashion designed to derive optimal results. Wald published a remarkably clear and simple presentation of his main results in a book [7], with an elegant appendix with more mathematical detail, in 1947, before the derivation of the optimality result.

While sequential analysis was an important tool in military applications where each observation was extremely costly, and the need to reduce sample size important, it tended to be neglected in applied work until recently when it began to be used more often in **clinical trials**. In much engineering work a major cost of experimentation is the setup cost, and it is relatively inconvenient to sample and observe the results one at a time. In the early post-World War II days, the computational complexity, more apparent than real, inhibited its use. A foundational problem also played a role here. When a sequential test with significance level 0.05 leads to rejection, it is possible that the investigator, acting as though the random sample size had been selected in advance, will calculate the ***P* value** and find a value quite different than 0.05 in either direction. With a fixed sample size, rejection would always lead to a smaller P value. The naive investigator would have difficulty in interpreting the data. The sophisticated investigator with Bayesian leanings would reject the concept that the interpretation should depend on the rule that was used to decide when sampling should be stopped.

Although we have concentrated on Wald's most important contributions to the statistical literature, his publications covered an enormous range of problems of importance where his mathematical skills and his statistical insights were effective. These included contributions to asymptotic theory, econometrics, theory of **nonparametric methods, analysis of variance, experimental design, multivariate analysis**, and the exploitation of the theorem of Lyapounov on the range of a vector measure. The following is a very brief description of some of these papers.

In asymptotic theory, Wald characterized asymptotically optimal procedures, presented a clarifying proof of the consistency (*see* **Consistent Estimator**) of **maximum likelihood** estimates, treated problems where the number of parameters approached infinity, and described a calculus of stochastic limit and order relations (*see* **Order Statistics**). In econometrics he dealt with the identification problem and presented one of the earliest effective attacks on the estimation of a linear relationship when observations on both variables are subject to error (*see* **Regression**). Exploiting the Lyapounov theorem with Wolfowitz and Dvoretzky, together they showed that in problems involving continuous distributions, inference did not require **randomization** to achieve specified significance levels. Moreover, the **sufficiency** conditions of Neyman–Pearson theory were also necessary. At SRG, he was responsible for revolutionizing the Air Force view of vulnerability with a report that exploited his insight that the places where returning planes did not have bullet holes were the ones that needed reinforcement.

Wald collaborated with several statisticians. His most frequent collaborator was J. Wolfowitz. Several interesting papers were written with H.B. Mann and, as indicated above, Dvoretzky and Wolfowitz worked with Wald on the consequences of the Lyapounov theorem. He also wrote with C. Stein and M. Sobel. As one of his students, Chernoff developed his asymptotic approach to the Fisher–Behrens problem in his thesis.

I wish to thank Professor T.W. Anderson for information about Wald's life [1] and for helpful comments on a previous draft of this biography.

References

[1] Anderson, T.W. (1955). The Department of Mathematical Statistics, in *A History of the Faculty of Political Science of Columbia University*, R.G. Hoxie, ed. Columbia University Press, New York, pp. 250–255.

[2] Arrow, K.J., Blackwell, D. & Girshick, M.A. (1949). Bayes and minimax solutions of sequential decision problems *Econometrica* **17**, 213–243.

[3] Menger, K. (1952). The formative years of Abraham Wald and his work in geometry, *Annals of Mathematical Statistics* **23**, 14–20.

[4] Tintner, G. (1952). Abraham Wald's contributions to econometrics, *Annals of Mathematical Statistics* **23**, 21–28.

[5] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

[6] Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses, *Annals of Mathematical Statistics* **10**, 299–326.

[7] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

[8] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

[9] Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* **19**, 326–399.

[10] Wallis, W.A. (1980). The Statistical Research Group, 1942–1945, *Journal of the American Statistical Association* **75**, 320–330.

[11] Wallis, W.A. (1980). Rejoinder: The Statistical Research Group, 1942–1945, *Journal of the American Statistical Association* **75**, 334–335.

[12] Wolfowitz, J. (1952). Abraham Wald, 1902–1950, *Annals of Mathematical Statistics* **23**, 1–13.

[13] Wolfowitz, J. (1952). The publications of Abraham Wald, *Annals of Mathematical Statistics* **23**, 29–33.

H. CHERNOFF

Wald's Identity

Wald [1, 2] discovered a remarkable identity which can solve approximately, or sometimes exactly, a number of boundary problems on random walk in one dimension with discrete time steps (*see* **Stochastic Processes**). Let $X_1, X_2, \ldots$ denote a sequence of independent and identically distributed **random variables**. Assume that $M(t) = \mathrm{E}[\exp(tX_1)]$, the **moment generating function** of the sequence $\{X_n\}$, exists for all real values t in some interval. Consider the restricted random walk problem such that the walk ends as soon as the sum

$$S_n = X_1 + \cdots + X_n$$

satisfies the condition that $S_n \geq a$ or $S_n \leq -b$, for some finite constants $a > 0$ and $b > 0$. Let $N = n$ be the first integer for which $S_n \geq a$ or $S_n \leq -b$. Then Wald shows that

$$\mathrm{E}[M(t)^{-N} \exp(tS_N)] = 1 \tag{1}$$

for every real or complex value of t for which $1 \leq |M(t)| < \infty$.

Differentiating (1) once and letting $t = 0$, we get the useful Wald equation.

Wald's Equation

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables having finite expectations. Assume that the integer-valued random variable N is a *stopping time* for the sequence $X_1, X_2, \ldots$ (i.e. the event $\{N = n\}$ is independent of $X_{n+1}, X_{n+2}, \ldots$ for all $n = 1, 2, \ldots$) such that $\mathrm{E}[N] < \infty$; then

$$\mathrm{E}[S_N] = \mathrm{E}[N]\mathrm{E}[X_1]. \tag{2}$$

Differentiating (1) twice and letting $t = 0$, we get another useful equation:

$$\mathrm{E}[(S_N)^2] = \mathrm{E}[N]\mathrm{E}[X_1^2]. \tag{3}$$

References

[1] Wald, A. (1944). On cumulative sums of random variables, *Annals of Mathematical Statistics* **15**, 283–296.

[2] Wald, A. (1946). Differentiation under the integral sign in the fundamental identity in sequential analysis, *Annals of Mathematical Statistics* **17**, 493–497.

(*See also* **Sequential Analysis**)

MEI-LING TING LEE

Wavelet Analysis

Wavelet analysis is an approach to signal representation that has grown in popularity by virtue of its ability to overcome some of the limitations of Fourier series approximation (*see* **Time Series**; **Fast Fourier Transform (FFT)**). Whereas Fourier analysis provides a stationary approximation of a signal in terms of its frequency components but lacks the ability to capture the local features of a signal, wavelet analysis does not suffer from this drawback. It is thus used extensively to capture transient behavior in signals. The technique derives its name from the characteristic shape of the analytic functions that replace the sines and cosines used in Fourier analysis.

Like sines and cosines, wavelet functions also show oscillatory behavior about zero, but unlike sines and cosines, they decay to zero, a property underpinning their ability to detect local features in a signal (see Figure 1). The so-called "father" wavelets integrate to a value of 1, and represent the smooth low frequencies well. In contrast, "mother" wavelets integrate to a value of zero and represent the detail and high-frequency features well. There are many different types of wavelet functions, with properties that lend themselves to different approaches to wavelet analysis and the detection of different features of signals [4]. The first few wavelet coefficients contain information about the overall shape of the time series, whilst the higher-order coefficients describe localized trends.

Fundamentally, wavelet analysis involves decomposition of a signal by computation of its inner products, with analysis functions formed by dilation and translation of a prototypical wavelet. The value of a wavelet coefficient (magnitude of the inner product) will be maximal when the shape and position of a particular feature in the image match those of the chosen wavelet (hence the choice of different wavelets for the analysis of different types of signals).

Using Unser's notation [4], the so-called *continuous* wavelet transformation can be represented as

$$(W_\varphi f)(a,b) = (f, \varphi_{(a,b)})$$

$$\varphi_{(a,b)} = a^{-1/2}\varphi\left(\frac{x-b}{a}\right), \qquad (1)$$

where $\varphi_{(a,b)}$ is a set of analysis functions derived by applying scalings and translations to a wavelet. The scalings, a, capture the properties of the signal at different resolutions (analogous to the frequency representation in a Fourier series), and the translations, b, capture local behavior.

The continuous wavelet transform maps a function of a single independent variable (e.g. time) to a function of two independent variables (scale and position in the time series) and is thus redundant and computationally inefficient. This problem can be overcome by sampling the transform on a discrete grid in the scale/position plane, leading to the discrete wavelet transform (DWT). With a suitable choice of wavelet (e.g. the orthogonal wavelet set described by Daubechies [1]) and the so-called pyramidal decomposition scheme described by Mallat [3], a rapid, invertible DWT can easily be accomplished.

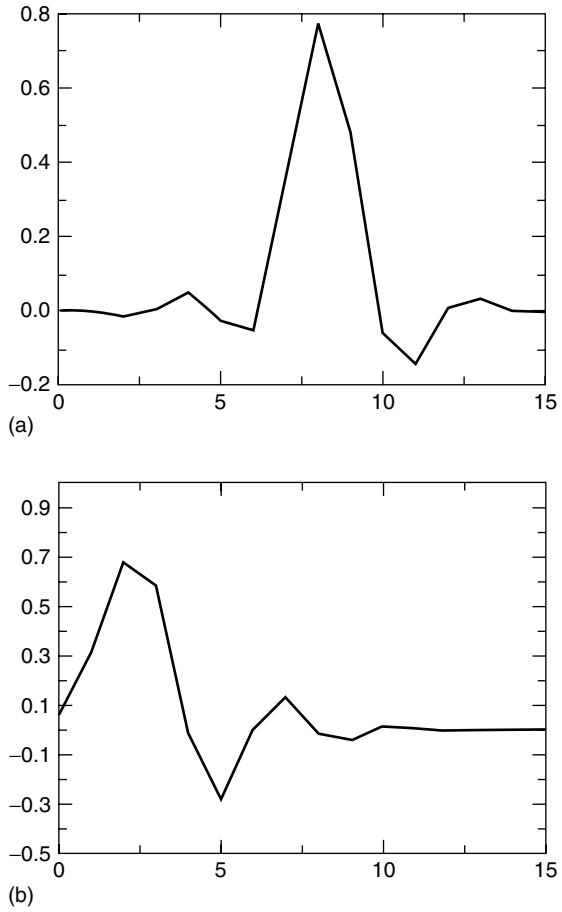

Figure 1 (a) 16 coefficient Symmlet wavelet (b) 16 coefficient Daubechies wavelet

Wavelet analysis/transformation has been extensively used in data compression, where a signal can often be represented by a number of wavelet coefficients equal to only a small fraction of the number of data points in the original data set. It has also been used for signal denoising. Donoho and his colleagues [2] have worked extensively in this area and have pioneered the technique of wavelet shrinkage. This technique involves an initial DWT, followed by computation of a threshold value for wavelet coefficients, rejection of coefficients below this threshold, and inversion of the DWT, using the modified coefficient set. Another important aspect of the DWT is its so-called decorrelating property. Thus, strong **correlations** existing in the temporal or spatial domain are often much less evident following DWT, a property that has begun to be exploited in statistical analysis.

References

[1] Daubechies, I. (1988). Orthogonal bases of compactly supported wavelets, *Communications on Pure and Applied Mathematics* **41**, 909–996.

[2] Donoho, D.L. & Johnstone, I.M. (1994). Ideal spatial adaptation via wavelet shrinkage, *Biometrika* **81**, 425–455.

[3] Mallat, S. (1989). A theory of multiresolution signal decomposition: the wavelet representation, *IEEE Transactions Pattern Analysis on Machine Intelligence* **PAM-11**(7), 674–693.

[4] Unser, M. (1996). A practical guide to the implementation of the wavelet transform, in *Wavelets in Medicine and Biology*, A. Aldroubi & M. Unser, eds. CRC press, Boca Raton, New York, London, Tokyo, pp. 37–73.

(*See also* **Spectral Analysis**)

M.J. BRAMMER

Weibull Distribution

A commonly used model for survival data is the two-parameter Weibull model with hazard rate $h(t) = \lambda\alpha t^{\alpha-1}, \alpha > 0, \lambda > 0$. This is a flexible model that has an increasing hazard rate when $\alpha > 1$, a decreasing hazard rate when $\alpha < 1$, and a constant hazard rate when $\alpha = 1$ (*see* **Exponential Distribution**). Its survival function is $S(t) = \exp(-\lambda t^{\alpha})$ and the density function $f(t) = \lambda\alpha t^{\alpha-1}\exp(-\lambda t^{\alpha})$.

The cumulative hazard rate of the Weibull is $H(t) = \lambda t^{\alpha}$, so that $\ln[H(t)] = \ln\lambda + \alpha \ln t$. This characterization can be used to assess whether the Weibull model fits data by plotting an empirical estimate of the cumulative hazard rate on a log–log scale. Such a plot should be linear if the model holds.

The rth moment of the Weibull distribution is $[\Gamma(1 + r/\alpha)]\lambda^{-r/\alpha}$. The mean and variance are $[\Gamma(1+ 1/\alpha)]\lambda^{-1/\alpha}$ and $\{\Gamma(1 + 2/\alpha) - [\Gamma(1 + 1/\alpha)]^2\}\lambda^{-2/\alpha}$, respectively, where $\Gamma(\alpha)$ is the **gamma** function. The pth **quantile** of the Weibull distribution is $x_p = \{-[\ln(1 - p)]/\lambda\}^{1/\alpha}$.

The Weibull distribution arises as a limiting distribution of the minimum of n independent random variables. If $X_1, \ldots, X_n$ are a random sample from a population with a survival function which, for t close to 0, is of the form $S(t) = 1 - \lambda t^{\alpha} + o(t^{\alpha})$, then the limiting distribution of $T_n = n^{1/\alpha}\min(X_i, \ldots, X_n)$ is Weibull with shape parameter α and scale parameter λ.

The Weibull distribution arises from a **multistage model** of carcinogenesis. In this model, cancer in a cell occurs if r different mutations in the cell occur. Suppose that θ_j is the mutation rate at the jth locus per unit time and that the probability that a mutation at the jth locus occurs in a cell prior to time t is approximately equal to $\theta_j t$, a small number. For each cell the probability that the required r mutations occur prior to time t is $(\prod_{j=1}^{r}\theta_j)t^r$. The average number of clones that develop from mutated cells up to time t is $\mu = c(\prod_{j=1}^{r}\theta_j)t^r = \lambda t^r$, where c is the number of cells at risk of mutation. If the number of clones at time t has an approximate **Poisson distribution** with mean μ, then the waiting time to first occurrence of disease follows a Weibull distribution with shape parameter r and scale parameter λ [1].

The Weibull distribution is related to the **extreme-value** distribution. If we let $Y = \ln T$, then Y has an extreme-value distribution with density function

$$\alpha\exp\left(\alpha\left[y - \left(\frac{-\ln\lambda}{\alpha}\right)\right] - \exp\left\{\alpha\left[y - \left(\frac{-\ln\lambda}{\alpha}\right)\right]\right\}\right), \quad -\infty < y < \infty.$$

If we let $\mu = -\ln\lambda/\alpha$ and $\sigma = \alpha^{-1}$, then $Y = \mu + \sigma E$, where E has the standard extreme-value distribution.

Estimation in the Weibull model based on independent, possibly right-**censored** survival times $X_1, \ldots, X_n$ is usually based on **maximum likelihood**. The estimator $(\hat{\alpha}, \hat{\lambda})$ is not given in a closed-form expression, but many of the standard software packages may be used for the estimation.

Two popular models that are used to adjust the survival function for covariates are the **proportional hazards model** (the **Cox regression model**) and the **accelerated failure-time model**. For the proportional hazards model, the conditional hazard rate of T, given a set of covariates $\mathbf{Z}$, is of the form

$$h(t|\mathbf{Z}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}),$$

while, for the accelerated failure-time model,

$$h(t|\mathbf{Z}) = h_0[t\exp(-\boldsymbol{\beta}'\mathbf{Z})]\exp(-\boldsymbol{\beta}'\mathbf{Z}),$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. The Weibull regression model is the only model which admits both an accelerated failure-time model and a proportional hazards model representation.

Reference

[1] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.

JOHN P. KLEIN, PER KRAGH ANDERSEN & NIELS KEIDING

Weighted Distributions

Traditional statistical theory and practice have been occupied largely with statistics involving **randomization** and replication (*see* **Experimental Design**). However, in biomedical and public health work, observations also fall in the nonexperimental, nonreplicated, and nonrandom categories (*see* **Observational Study**). The problems of model specification and data interpretation acquire special importance and great concern (*see* **Misspecification**). The theory of weighted distributions provides a perceptive and unifying approach for the problems of model specification and data interpretation. Weighted distributions take into account the observer–observed interface, i.e. the method of **ascertainment**, by adjusting the probabilities of actual occurrence of events to arrive at a specification of the probabilities of those events as observed and recorded. Failure to make such adjustments can lead to wrong conclusions.

The concept of weighted distributions can be traced to the study of the effect of methods of ascertainment upon estimation of frequencies by Fisher [5]. In extending the basic ideas of Fisher, Rao [16, 17] saw the need for a unifying concept, and identified various sampling situations that can be modeled by what he called weighted distributions. Within the biomedical context of cell kinetics and the early detection of disease, Zelen [23] introduced weighted distributions to represent what he broadly perceived as **length-biased** sampling introduced earlier by Cox [1]. In a series of papers with his co-workers, Patil has pursued weighted distributions in theory and practice for purposes of encountered data analysis, equilibrium population analysis subject to harvesting and predation, **meta-analysis** incorporating publication bias and heterogeneity, modeling clumping and extraneous variation, etc. See, for example, [2, 7–13, 15], and [20], and for more references, see [14].

To introduce the concept of a weighted distribution, suppose X is a nonnegative observable **random variable** (rv) with its natural probability density function (pdf) $f(x;\theta)$, where the natural parameter $\theta \in \Omega$, the parameter space. Suppose a realization x of X under $f(x;\theta)$ enters the investigator's record with probability proportional to $w(x,\beta)$, so that

$$\frac{\Pr(\text{recording}|X=y)}{\Pr(\text{recording}|X=x)} = \frac{w(y,\beta)}{w(x,\beta)}.$$

Here, the recording (weight) function $w(x,\beta)$ is a nonnegative function with parameter β representing the recording (sighting) mechanism. Clearly, the recorded x is not an observation on X, but on the rv X^w, say, having pdf

$$f^w(x;\theta,\beta) = \frac{w(x,\beta)f(x;\theta)}{\omega},$$

where ω is the normalizing factor obtained to make the total probability equal to unity by choosing $\omega = \mathrm{E}[w(X,\beta)]$. The rv X^w is called the weighted version of X, and its distribution in relation to that of X is called the weighted distribution with weight function w. Note that the weight function $w(x,\beta)$ need not lie between zero and one, and actually may exceed unity, as, for example, when $w(x,\beta) = x$, in which case, $X^* = X^w$ is called the size-biased version of X. The distribution of X^* is called the size-biased distribution with pdf

$$f^*(x;\theta) = \frac{xf(x;\theta)}{\mu},$$

where $\mu = \mathrm{E}[X]$. The pdf f^* is called the length-biased or size-biased version of f, and the corresponding observational mechanism is called length- or size-biased sampling. The concept of weighted distributions has been much used recently as a useful tool in the selection of appropriate models for observed data, especially when samples are drawn without a proper frame. In many situations, the model given above is appropriate, and the statistical problems that arise are the determination of a suitable weight function $w(x,\beta)$ and drawing inference on θ. Appropriate statistical modeling approaches help accomplish unbiased inference in spite of the biased data and, at times, even provide a more informative and economic setup (see [13]).

The following examples may help illustrate a few situations generating weighted distributions and their applications.

Example 1: Analysis of Family Data, $w(x,\beta) = w(x) = x$

Various **demographic** studies involve family size and sex ratio as important factors which have some

Table 1 Analysis of family data

Family size	1	2	3	4	5	6	7	8	9	10	11	12	13	15	Total
No. of families	1	6	6	13	12	7	14	11	12	8	6	5	2	1	104
Brothers	1	8	12	34	34	29	59	50	54	46	32	31	16	8	414
Sisters	0	4	6	18	26	13	39	38	54	34	34	29	10	7	312

bearing on the main study. This example shows how a weighted distribution arises as a result of size-biased sampling.

Consider the data in Table 1 relating to brothers and sisters in families of 104 boys admitted to a postgraduate course. Assume that in families of given size n, the probability of a family with x boys coming into the record is proportional to x. Also, suppose that the number of boys follows a **binomial distribution** with probability parameter π. Then

$$f(x;\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x},$$

$$w(x) = x, \omega = n\pi,$$

$$f^w(x;\pi) = \binom{n-1}{x-1} \pi^{x-1} (1-\pi)^{n-x},$$

$$\mathrm{E}\left[\frac{X^w}{n}\right] = \pi + \frac{1-\pi}{n} > \pi, \quad \text{and}$$

$$\mathrm{E}\left[\frac{X^w - 1}{n-1}\right] = \pi.$$

If k boys representing families of size $n_1, n_2, \ldots, n_k$ report $x_1, x_2, \ldots, x_k$ boys, an unbiased estimate of π is

$$\tilde{\pi} = \frac{\sum x_i - k}{\sum n_i - k} = \frac{414 - 104}{726 - 104} \doteq \frac{1}{2}$$

(see [14] and [19]).

Example 2: Analysis of Intervention Data, $w(x, \beta) = w(x) = x$

The expected value of the duration to the completion of a random event sampled randomly at the end of its duration turns out to be approximately equal to the expected duration to its random intervention. This can be explained using the concept of size-biased/length-biased sampling with weight function $w(x, \beta) = w(x) = x$, where x represents the duration of the random event. The applications in medical and public health sciences include: (i) **cell cycle** analysis and pulse labeling [23]; (ii) efficacy of early **screening** for disease and scheduling of examinations [23]; and (iii) cardiac **transplantation** [22]. Simon [18] uses length-biased sampling in etiologic studies for estimation of antigen frequencies to compare patients with the antigen that are more likely to be alive and included in the study than patients without the antigen.

Example 3: Analysis with Damaged Observations, $w(x, \beta) = \beta^x$

Consider a damage model where an observation $X = x$ is reduced to y by a destructive process with pdf $d(y|x)$. Then the probability that the observation $X = x$ is undamaged is $d(x|x)$, and the distribution of the undamaged observation is the weighted distribution with $w(x) = d(x|x)$. For example, under the binomial survival model, $d(x|x) = \theta^x, 0 < \theta < 1$. An investigator recording only undamaged observations will need to work with a corresponding weighted distribution.

Example 4: Modeling Clumped Sampling, Heterogeneity, and Extraneous Variation, $w(x, \beta) = w(x, \beta, \theta)$

During their examination of the problem of toxoplasmosis, Diaconis & Efron [3, 4] looked at the data sets coming from different cities with different rainfall and found that there was more dispersion in the data sets than the existing models could accommodate (*see* **Overdispersion**), and therefore they introduced a model called the double exponential family (DEF). This family enjoys the **exponential family** properties simultaneously for the mean and the dispersion parameters. It allows the data analyst to model overdispersion while carrying out the usual regression analyses for the mean as a function of the predictors. The overdispersion may be due to one or more possible causes, such as clumped sampling (*see* **Clustering**), heterogeneity, **selection bias**, etc.

Interestingly, the DEF can be seen as a weighted distribution (see [9] and [13]). The weight function turns out to have the following interesting form:

$$w(x, \beta) = w(x, \beta, \theta) = \exp(1 - \beta) I[x, \mu(\theta)],$$

where $I(x, \mu)$ is the **Kullback–Leibler** distance function between x and $\mu(\theta) = \mu = \mathrm{E}(X)$ of the usual exponential family density function f with parameter θ. Kullback–Leibler distance increases with the distance from the mean μ, thus allowing a more distant observation larger weight and accommodating extra dispersion in the data set when $1 - \beta > 0$.

Example 5: Meta-analysis Incorporating Heterogeneity and Publication Bias

Meta-analysis consists of quantitative methods for combining evidence from different studies about a particular issue. Its objective is to summarize quantitatively a research literature with respect to a particular question and to examine systematically the manner in which a collection of studies contributes to knowledge about that question.

The weight function enters the analysis to represent the publication/selection bias and the heterogeneity among different studies. It also helps model the overdispersion/underdispersion in the data caused by publication bias and the inherent heterogeneity.

The weight functions examined include a:

1. Critical value model: $w(x) = (x/x_{\mathrm{crit}})^{\beta}$ if $|x| < x_{\mathrm{crit}}$, and $= 1$, otherwise,
2. half-normal model: $w(x) = \exp[-\beta p(x)^2]$, and
3. negative exponential model: $w(x) = \exp[-\beta p(x)]$.

Here $p(x)$ is the ***P*** **value** when the test statistic takes value x, and x_{crit} stands for the critical value under the test statistic (see [6, 7], and [13]).

Example 6: Statistical Analysis Incorporating Overdispersion and Heterogeneity in Teratologic Binary Data in Developmental Toxicity Studies

The problem of overdispersion and heterogeneity in binary data arises quite naturally in developmental toxicity studies. Since a pregnant female is exposed to the chemical dose, litter becomes the primary unit. The random effect of the litter, i.e. the biological response of the mother to the chemical dose, affects the toxic responses of the fetuses. This potential random litter effect causes heterogeneity, extravariation, and also intralitter correlation between the responses of two fetuses within the litter.

To incorporate this random litter effect in the analysis of such binary data, research workers have introduced the **beta-binomial** model. The beta-binomial model, regarded as the binomial mixture model, takes care of overdispersion, heterogeneity, and also the clumping of observations within the litter. This clumping of observations occurs because of the tendency of the fetuses within the litter to "behave" alike. For this reason, the use of the double binomial family model seems in place as an alternative model in the analysis of developmental toxicity data.

The beta-binomial model provides an estimate of the index of overdispersion and also an estimate of intralitter **correlation** coefficient, while the double-binomial model provides an estimate of the overdispersion parameter.

The problem of analysis of overdispersed binary data coming from different litters at a given dose or coming from different cities with given rainfall can also be looked upon as the problem of encounter data where one is trying to combine the observational data coming from different sources. Therefore, one can look upon both the beta-binomial and the double-binomial as weighted binomial distributions where each model has its own separate weight function.

The two distributions are quite comparable in their capability for describing overdispersed binary data. It appears that the choice between them may be made on such grounds as parameter interpretation and inferential convenience (see [15] and [21].

Acknowledgment

This article was prepared with partial support from the Statistical Analysis and Computing Branch, Environmental Statistics and Information Division, Office of Policy, Planning, and Evaluation, United States Environmental Protection Agency, Washington, DC, under a Cooperative Agreement Number CR-821531. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

References

[1] Cox, D.R. (1962). *Renewal Theory*. Barnes & Noble, New York.

[2] Dennis, B. & Patil, G.P. (1984). The gamma distribution and weighted multimodal gamma distributions as models of population abundance, *Mathematical Biosciences* **68**, 187–212.

[3] Diaconis, P. & Efron, B. (1985). Testing the independence of a two-way table: new interpretations of the chi-square statistic (with discussion and rejoinder), *Annals of Statistics* **13**, 845–913.

[4] Efron, B. (1986). Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association* **81**, 709–721.

[5] Fisher, R.A. (1934). The effects of methods of ascertainment upon the estimation of frequencies, *Annals of Eugenics* **6**, 13–25.

[6] Iyengar, S. & Greenhouse, J.B. (1988). Selection models and the file drawer problem, *Statistical Science* **1**, 109–135.

[7] Laird, N., Patil, G.P. & Taillie, C. (1988). Comment on S. Iyengar and J. B. Greenhouse, "Selection models and the file drawer problem", *Statistical Science* **3**, 126–128.

[8] Patil, G.P. (1981). Studies in statistical ecology involving weighted distributions, in *Statistics Applications and New Directions: Proceedings of ISI Golden Jubilee International Conference*, J.K. Ghosh & J. Roy, eds. Statistical Publishing Society, Calcutta, pp. 478–503.

[9] Patil, G.P. (1991). Encountered data, statistical ecology, environmental statistics, and weighted distribution methods, *Environmetrics* **24**, 377–423.

[10] Patil, G.P. (1996). Statistical ecology, environmental statistics, and risk assessment, in *Advances in Biometry*, P. Armitage & H.A. David, eds. Wiley, Chichester. pp. 213–240.

[11] Patil, G.P. & Ord, J.K. (1976). On size-biased sampling and related form-invariant weighted distributions, *Sankhyā* **38**, 48–61.

[12] Patil, G.P. & Rao, C.R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families, *Biometrics* **34**, 179–184.

[13] Patil, G.P., & Taillie, C. (1989). Probing encountered data, meta analysis and weighted distribution methods, in *Statistical Data Analysis and Inference*, Y. Dodge, ed. Elsevier, Amsterdam, pp. 317–345.

[14] Patil, G.P., Rao, C.R. & Zelen, M. (1988). Weighted distributions, in *Encyclopedia of Statistical Sciences*, Vol. 9. S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 565–571.

[15] Patil, G.P., Taillie, C. & Talwalker, S. (1993). Encounter sampling and modelling in ecological and environmental studies using weighted distribution methods, in *Statistics for the Environment*, V. Barnett & K.F. Turkman, eds. Wiley, New York.

[16] Rao, C.R. (1965). On discrete distributions arising out of methods of ascertainment, in *Classical and Contagious Discrete Distributions*, G.P. Patil, ed. Pergamon Press and Statistical Publishing Society, Calcutta, pp. 320–332.

[17] Rao, C.R. (1985). Weighted distributions arising out of methods of ascertainment, in *A Celebration of Statistics*, A.C. Atkinson & S.E. Fienberg, eds. Springer-Verlag, New York, Chapter 24, pp. 543–569.

[18] Simon, R. (1980). Length biased sampling in etiologic studies, *American Journal of Epidemiology* **111**, 444–452.

[19] Stene, J. (1981). Probability distributions arising from the ascertainment and the analysis of data on human families and other groups, in *Statistical Distributions in Scientific Work*, Vol. 6: Applications in Physical, Social, and Life Sciences, C. Taillie, G.P. Patil & B. Baldessari, eds. Reidel Publishing Company, Dordrecht and Boston, pp. 233–264.

[20] Taillie, C., Patil, G.P. & Hennemuth, R.C. (1995). Modelling and analysis of recruitment distributions, *Environmental and Ecological Statistics* **2**, 315–330.

[21] Talwalker, S., Patil, G.P. & Taillie, C. (1995). Qualitative and quantitative assessment of the risk from the exposure to fetotoxic chemical compounds, *Environmental and Ecological Statistics* **2**, 71–79.

[22] Temkin, N. (1976). Interactive Information and Distributional Length Biased Survival Models, Unpublished PhD Dissertation. University of New York at Buffalo.

[23] Zelen, M. (1974). Problems in cell kinetics and the early detection of disease, in *Reliability and Biometry*, F. Proschan & R.J. Serfling, eds. SIAM, Philadelphia, pp. 701–706.

G.P. PATIL

Wigner–Ville Distribution

The Wigner–Ville distribution is a time–frequency distribution developed for the analysis of time-varying spectra [1]. Indeed, the interpretation of classic Fourier analysis (*see* **Fast Fourier Transform (FFT)**), which decomposes the power of a signal into frequency components by quantifying the power spectrum, may became problematic when the frequency content changes over time. In fact, the power spectrum cannot indicate when specific spectral components occur or how they change in intensity and frequency. Thus, when the signal is characterized by a time-varying spectrum, it is preferable to decompose the signal power by a joint function of time and frequency. The Wigner–Ville distribution $W(t, f)$ quantifies the fraction of the power in a certain frequency band during a certain time range, representing the "instantaneous" spectrum of a nonstationary signal $s(t)$. It is defined by

$$W(t, f) = \frac{1}{2} \int s^* \left(t - \frac{1}{2}\tau \right) \times s \left(t + \frac{1}{2}\tau \right) \mathrm{e}^{-j2\pi f \tau} \, \mathrm{d}\tau, \qquad (1)$$

where $s^*(t)$ is the complex conjugate of $s(t)$.

$W(t, f)$ satisfies the marginal conditions, which means that the integral of the distribution over the frequency f at a certain time t gives the instantaneous energy of the signal, $|s(t)|^2$, and the integral over the time t at each frequency f gives the energy spectrum $E(f)$ (the power spectrum is the energy spectrum per unit of time). Thus, $W(t, f)$ satisfies the intuitive idea of a time-varying spectrum of $s(t)$, and the product $W(t, f)\,\mathrm{d}t\,\mathrm{d}f$ can be interpreted as the fraction of the $s(t)$ energy at time t and frequency f in the $(\mathrm{d}t \times \mathrm{d}f)$ cell.

The formulation of the Wigner–Ville distribution for a discrete time signal $s(n)$ is

$$W(n, \omega) = \frac{1}{\pi} \sum_{k=-\infty}^{+\infty} s^*(n-k) s(n+k) \mathrm{e}^{-j2\omega k} \qquad (2)$$

A negative characteristic of the Wigner–Ville distribution is the presence of important *interference terms*. In fact, when $s(t)$ is a multicomponent signal (e.g. the sum of sinusoids occurring at different frequencies and/or in different time ranges), the $W(t, f)$ distribution may be not zero during time periods when the signal is not expected or in frequency bands where spectral components are not expected. Moreover, interference terms may also produce negative $W(t, f)$ values, which cannot be considered energy components of $s(t)$. These undesired components have no physical meanings but can be in part suppressed by smoothing $W(t, f)$. However, a smoothed $W(t, f)$ has a lower resolution in time and frequency and might not satisfy the marginal conditions. A popular smoothed $W(t, f)$ is the smoothed pseudo-Wigner–Ville, $PSW(t, f)$, proposed in [2]:

$$PSW(t, f) = \int \left| h\left(\frac{\tau}{2}\right) \right|^2 \int g(u-t) s^* \times \left(u - \frac{1}{2}\tau \right) s \left(u + \frac{1}{2}\tau \right) \mathrm{d}u \times \mathrm{e}^{-j2\pi f \tau} \, \mathrm{d}\tau \qquad (3)$$

The $h()$ window provides frequency smoothing and $g()$ suppresses interference terms. The Wigner–Ville distribution has been successfully implemented to detect changes in the structure of several biological signals, like nonstationary cardiovascular signals, ultrasonic Doppler signals, auditory neuron activity, and acoustic signals [3] (*see* **Clinical Signals**).

References

[1] Cohen, L. (1989). Time-frequency distributions–a review, *Proceedings of the IEEE* **77**, 941–981.

[2] Martin, W. & Flandrin, P. (1985). Wigner-Ville spectral analysis of nonstationary processes, *IEEE Transactions on Acoustic, Speech and Signal Processing* **33**, 1461–1470.

[3] Novak, P. & Novak, V. (1993). Time/frequency mapping of the heart rate, blood pressure and respiratory signals, *Medical and Biological Engineering and Computation* **31**, 103–110.

(*See also* **Choi–Williams Distribution**)

PAOLO CASTIGLIONI

Wilcoxon Signed-rank Test

The Wilcoxon signed-rank test, due to Wilcoxon [9], is a nonparametric test procedure (*see* **Nonparametric Methods**) used for the analysis of matched-pair data or for the one-sample problem. In the matched-pair setting it is used to test the hypothesis that the probability distribution of the first sample is equal to the probability distribution of the second sample (*see* **Hypothesis Testing**). This hypothesis can be tested from statistics calculated on the intrapair differences. The hypothesis commonly tested is that these differences come from a distribution centered at zero.

Consider the following example. A study was conducted in which nine people of varying weights were put on a particular exercise regimen to determine the program's effect on the resting heart rate of the subjects. Given that a low resting heart rate is beneficial in reducing blood pressure and increasing overall cardiovascular fitness, this exercise regimen was developed to help people lower their resting heart rate. To test the effectiveness of the regimen, the resting heart rate measurement for each subject was taken before the induction of the regimen, and at six months after beginning the regimen. Table 1 presents the data from this study.

Because this study involves before and after measurements of the same individuals, an independent sample test procedure cannot be executed. The **null hypothesis** in the Wilcoxon signed-rank test is that the set of pairwise differences have a probability distribution centered at zero. A key assumption is that the differences arise from a continuous, symmetric distribution. In the example, the null hypothesis would be that there is no resting heart rate difference before and after the exercise regimen ($H_0 : \mu_d = 0$). In this instance, μ_d represents the location parameter for the distribution of differences. One **alternative hypothesis** is that the resting heart rate before the exercise regimen is higher than the resting heart rate after the exercise regimen ($H_1 : \mu_d > 0$).

To execute the test, the absolute values of the differences, $|d_i|$, are computed. These values also are given in Table 1. After computing the absolute values, one must order them from smallest to largest disregarding any zeros (*see* **Ranks**). In the case of absolute differences being tied for the same ranks, the mean rank (mid-rank) is calculated and assigned to each tied value. The rankings for the absolute differences are also given in Table 1. Test statistics for the Wilcoxon signed-rank test are calculated by either summing the ranks assigned to the positive differences (T_+) or by summing the ranks assigned to the negative differences (T_-). If there are n differences, then the two sums are related through

$$T_- = \left\{\frac{[n(n+1)]}{2}\right\} - T_+. \tag{1}$$

In the example, the sum of the ranks of the positive differences is

$$T_+ = 5 + 3 + 9 + 7 + 4 + 6 + 8 = 42.$$

Note that $\{[n(n+1)]/2\} = 45$ so

$$T_- = 45 - 42 = 3.$$

To test the null hypothesis, a rejection region can be determined for the test statistic, T_+. This rejection region can be determined from the exact null hypothesis distribution of T_+. This null distribution is easily derived from a permutational argument, as each of the possible configuration of signs (+ or −) is equally likely under the null hypothesis. Tables of this exact null distribution are available in standard nonparametric texts such as [3, 4], or [7]. This null distribution depends only on n, hence the test procedure is nonparametric, i.e. distribution-free.

For **large samples** the standard normal distribution Z, can be used as an approximation to test hypotheses. For this situation a two-tailed rejection region for the null hypothesis based on T_+, is given:

$$Z_+ = \frac{\left\{\frac{[T_+ - n(n+1)]}{4}\right\}}{\left\{\frac{[n(n+1)(2n+1)]}{24}\right\}^{1/2}} > Z_{1-\alpha/2} \tag{2}$$

or

$$Z_- = \frac{\left\{\frac{[T_- - n(n+1)]}{4}\right\}}{\left\{\frac{[n(n+1)(2n+1)]}{24}\right\}^{1/2}} > Z_{1-\alpha/2}. \tag{3}$$

A one-tailed test is conducted in a similar fashion with the comparison made to $Z_{1-\alpha}$.

Table 1 Resting heart rate of nine people before and after initiation of an exercise regimen

Subject	Heart rate at baseline (y_i)	Heart rate at 6 months (x_i)	Difference $d_i = y_i - x_i$	Absolute value of difference $\lvert d_i \rvert$	Rank of absolute difference (sign)
1	80	72	+8	8	5(+)
2	76	70	+6	6	3(+)
3	78	82	−4	4	2(−)
4	90	76	+14	14	9(+)
5	84	86	−2	2	1(−)
6	86	76	+10	10	7(+)
7	81	74	+7	7	4(+)
8	84	75	+9	9	6(+)
9	88	76	+12	12	8(+)

Other issues regarding the Wilcoxon signed-rank test include its testing efficiency and the construction of estimators. The **asymptotic relative efficiency** (ARE) of this test relative to the **paired *t* test** is never less than 0.864 in the entire class of continuous symmetric distributions, and is 0.955 if the underlying distribution of differences is normal, see [2]. The handling of ties and zeros is discussed by Pratt [5] and Cureton [1]. Point and **confidence interval** estimators are easily derived from the test procedure, and details are described in Lehmann [4]. Lehmann [4] also describes power properties for the test procedure when shift alternatives are of interest. Extensions to **censored data** are discussed by Woolson & Lachenbruch [10] and Schemper [8]. References for other aspects of the Wilcoxon signed-rank test are given by Randles & Wolfe [7] and Randles [6].

References

[1] Cureton, E.E. (1967). The normal approximation to the signed-rank sampling distribution when zero differences are present, *Journal of the American Statistical Association* **62**, 1068–1069.

[2] Hodges, J.L., Jr & Lehmann, E.L. (1956). The efficiency of some nonparametric competitors of the *t*-test, *Annals of Mathematical Statistics* **27**, 324–335.

[3] Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods.* Wiley, New York.

[4] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

[5] Pratt, J.W. (1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures, *Journal of the American Statistical Association* **54**, 655–667.

[6] Randles, R.H. (1988). Wilcoxon signed rank test, in *Encyclopedia of Statistical Sciences,* Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 613–616.

[7] Randles, R.H. & Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics.* Wiley, New York.

[8] Schemper, M. (1984). A generalized Wilcoxon test for data defined by intervals, *Communications in Statistics – Theory and Methods* **13**, 681–684.

[9] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

[10] Woolson, R.F. & Lachenbruch, P.A. (1980). Rank tests for censored matched pairs, *Biometrika* **67**, 597–600.

(*See also* **Signed-rank Statistics**)

R.F. WOOLSON

Wilcoxon, Frank

Born: September 2, 1892, in County Cork, Ireland.
Died: November 18, 1965, in Tallahassee, Florida.

Photograph supplied by the Department of Statistics, at Florida State University

Frank Wilcoxon is best known in statistics for his fundamental work on ranking methods (*see* **Nonparametric Methods**; **Ranks**), but he was also an excellent chemist and made research contributions to physical chemistry, biochemistry, plant pathology, and entomology. Also, Frank Wilcoxon was an outstanding human being, with an enthusiasm for understanding the world and communicating his excitement to all around him.

Wilcoxon was born in Glengarriffe Castle near Cork, Ireland, of wealthy American parents. He was raised in Catskill, New York, and developed a lasting love for nature there. In his early years, he had varied experiences as a merchant sailor, pumper of gas, and tree surgeon before receiving his B.S. degree from Pennsylvania Military College in 1917. After World War I, Wilcoxon received an M.S. degree in chemistry from Rutgers University in 1921 and a Ph.D. degree in physical chemistry from Cornell University in 1924. At Cornell, Frank met Frederica Facius, an undergraduate, and they were married in 1926. Frank and Freddie were long-time attendees at the Gordon Research Conferences on Statistics and Chemistry and Chemical Engineering, and became well known and loved in the statistical community.

From 1924 to 1950, Frank Wilcoxon did research related to chemistry, first at the Boyce Thompson Institute for Plant Research in Yonkers, New York, and later at the Nichols Copper Company in Queens, Long Island and the Ravenna Ordance Plant operated by the Atlas Power Company. He worked at the American Cyanamid Company beginning in 1943 and continued until his retirement in 1957, first with the Stamford Research Laboratories as head of a group developing insecticides and fungicides and later as head of the statistics group of the Lederle Division in Pearl River, New York. Subsequent to his retirement from American Cyanamid, he served as a consultant to various organizations until 1960, when he joined the faculty of the newly formed Department of Statistics at Florida State University in Tallahassee. He remained active in research and teaching and contributed to the development of the Department until his death in 1965. Florida State University honored Wilcoxon by designating the statistics library and reading room as the Frank Wilcoxon Memorial Room.

Wilcoxon's interest in statistics began in 1925 with a study of **R.A. Fisher's** book, then newly published, *Statistical Methods for Research Workers* [6]. This study was done in a small reading group, of which W.J. Youden was a member and **C.I. Bliss** was a visitor. Wilcoxon's first publication in a statistics journal was on the usage of statistics in plant pathology [8], and in the same year he published his most significant contributions to statistics, the two-sample rank sum statistic and the one-sample signed rank statistic, both proposed in a very brief paper [9]. These statistics are well known as the Wilcoxon two-sample rank-sum test (*see* **Wilcoxon–Mann–Whitney Test**) and the **Wilcoxon signed-rank test**, and inspired much subsequent research on ranking methods, in addition to having a major impact on applied statistics, especially for applications in the **social sciences**.

In collaboration with Bradley and other colleagues, Wilcoxon extended the basic rank procedures to sequential testing situations [2, 3, 11, 12], including **screening**-type experiments. Wilcoxon was also interested in **multiple comparisons** problems, and the 1964 revision of the booklet, *Some Rapid Approximate Statistical Procedures* [10], gives multiple comparison procedures based on the rank-sum test for one- and two-way designs. This booklet was

widely circulated and played a significant role in the widespread use of nonparametric multiple comparison procedures.

In collaboration with Cuthbert Daniel, Wilcoxon devised **factorial experimental** designs that would be **robust** against certain linear and quadratic trends. From their paper, "The basic idea, due to Frank Wilcoxon, is that certain of the ordered contrasts appearing in the 2^{p-q} system are orthogonal to linear and quadratic trends." [4].

Wilcoxon made contributions to chemistry and biochemistry with about 40 publications. His publications spanned varied areas: acidimetry and alkalimetry; the mode of action of sulphur and copper fungicides; a mercury reduction method for the determination of pyrethrin I; synthesis of a number of plant growth substances; and research leading to the development of the insecticides Parathion and Malathion.

Although Wilcoxon was not an academician for most of his career, he was a teacher and student throughout life. He communicated his enthusiasm for statistics to students at Florida State University and was always available to talk about problems. He had wide interests including being an accomplished musician, a student of languages, and was fascinated by mathematical games, puzzles, combinatorics, and other pursuits. Through middle-age, he and his wife often traveled by bicycle, and he did not own an automobile until late in life. At Florida State University, he regularly rode a motorcycle to work.

Wilcoxon was a Fellow of the **American Statistical Association** and of the American Association for the Advancement of Science. He was an early chairman and leader in the development of the Gordon Research Conferences on Statistics in Chemistry and Chemical Engineering. The Chemical Division of the American Society for Quality Control annually awards a Frank Wilcoxon prize for the best papers of the year on practical applications published in *Technometrics*.

For a fuller account of his life and work, see [1] and [5], and a complete bibliography is given in [7].

References

[1] Bradley, R.A. & Hollander, M. (1978). Biography of Frank Wilcoxon, in *International Encyclopedia of Statistics*, Vol. 2, W.H. Kruskal & J.M. Tanur, eds. The Free Press (Division of Macmillan Publishing Company, Inc.), New York, pp. 1245–1250.

[2] Bradley, R.A., Martin, D.C. & Wilcoxon, F. (1965). Sequential rank tests: I. Monte Carlo studies of the two-sample procedure, *Technometrics* **7**, 463–483.

[3] Bradley, R.A., Merchant, S.D. & Wilcoxon, F. (1966). Sequential rank tests: II. Modified two-sample procedures, *Technometrics* **8**, 615–623.

[4] Daniel, C. & Wilcoxon, F. (1966). Factorial 2^{p-q} plans robust against linear and quadratic trends, *Technometrics* **8**, 259–278.

[5] Dunnett, C.W. (1966). Frank Wilcoxon, 1892–1965, *Technometrics* **8**, 195–196.

[6] Fisher, R.A. (1925). *Statistical Methods for Research Workers*, 14th Ed., 1970, revised & enlarged, Oliver & Boyd, Edinburgh; Hafner, New York.

[7] Karas, & Savage, I.R. (1967). Publications of Frank Wilcoxon (1892–1965), *Biometrics* **23**, 1–11.

[8] Wilcoxon, F. (1945). Some uses of statistics in plant pathology, *Biometrics Bulletin* **1**, 41–45.

[9] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics Bulletin* **1**, 80–83.

[10] Wilcoxon, F. (1964). *Some Rapid Approximate Statistical Procedures*. Stamford Research Laboratories. American Cyanamid Company, 1947, revised 1949, and revised, jointly with Roberta A. Wilcox, in 1964.

[11] Wilcoxon, R. & Bradley, R.A. (1964). Two sequential two-sample grouped rank tests with applications to screening experiments, *Biometrics* **20**, 892–895.

[12] Wilcoxon, F., Rhodes, L.J. & Bradley, R.A. (1963). Two sequential two-sample grouped rank tests with applications to screening experiments, *Biometrics* **19**, 58–84.

EDMUND A. GEHAN

Wilcoxon–Mann–Whitney Test

This test for comparing two samples with respect to their "general size" is based on ranking the observations – both samples combined – and then comparing the average **ranks** in the two samples. Though this idea had appeared several times in various disciplines [2], the statistical community first recognized the idea when Wilcoxon proposed it in 1945 [9]; thereafter developments followed fast, the first of which was Mann & Whitney's paper [3].

Representations of the Test

To fix ideas we introduce some notation. Until further stated, we consider continuous data, thus excluding ties. Let $X_1, X_2, \ldots, X_m$ be independent and identically distributed, with unknown cumulative distribution function F. Define

$$F_m(t) = \frac{1}{m} \sum_{i=1}^{m} u(x_i, t), \tag{1}$$

where the function $u(a, b) = 1$ if $a < b$ and zero if not.

Similarly, define $Y_1, Y_2, \ldots, Y_n$, and G,

$$G_n(t) = \frac{1}{n} \sum_{j=1}^{n} u(y_j, t), \tag{2}$$

and write $N = m + n$.

If $r_i = r(x_i)$ is the rank of x_i in the combined sample, let $R(x) = \sum_{i=1}^{m} r(x_i)$; and if $s_i = s(y_i)$ is the rank of y_j in the combined sample, let $R(y) = \sum_{j=1}^{n} s(y_j)$.

Observe that $R(x) + R(y) = N(N + 1)/2$ since each side represents the sum of the integers $1, 2, \ldots, N$.

Define

$$U(x < y) = \sum_{i=1}^{m} \sum_{j=1}^{n} u(x_i, y_j), \tag{3}$$

where again $u(x_i, y_j) = 1$ if $x_i < y_j$ and zero otherwise.

$U(x < y)$ reports how many of the mn distinct pairs comprising one x_i and one y_j have $x_i < y_j$. Mann & Whitney showed that

$$R(y) = \frac{n(n + 1)}{2} + U(x < y), \tag{4}$$

and that, hence, properties of Wilcoxon's test could be learned by studying $U(x < y)$. The relation between $R(y)$ and $U(x < y)$ also implies that one may choose to calculate whichever is more convenient with any particular data set. (In what follows we write W-M-W for Wilcoxon–Mann–Whitney.)

Because a monotone continuous **transformation** (like $x^{1/2}$ or $\log x$) does not change order relations, both $U(x < y)$ and $R(y)$ are also unaffected.

Distribution Theory When $F = G$ (i.e. H_0 Holds)

The exact distribution of $U(x < y)$ is obtained by enumeration, which is much expedited by using recursion relationships.

Under H_0 the mean and variance of $U(x < y)$ are:

$$\mathrm{E}_0[U(x < y)] = \frac{mn}{2} \tag{5}$$

and

$$\mathrm{var}_0[U(x < y)] = \frac{mn(N + 1)}{12}. \tag{6}$$

Both results are readily obtained by regarding $R(y)$ as the sum of n random observations chosen without replacement from $(1, 2, \ldots, N$; *see* **Sampling With and Without Replacement**).

Asymptotic normality (shown below) provides good approximation to the exact distribution for m and n both large ($m \geq 8, n \geq 8$, suffices at $2p = 0.05$; *see* **Large-sample Theory**).

Owen [7] tabulates distributions of both $U(x < y)$ and $R(y)$.

Distribution Theory: General Case; x and y Continuous

It is evident that

$$\begin{aligned} \mathrm{E}[u(x_i, y_j)] &= 1 \times \Pr(x < y) + 0 \times \Pr(y < x) \\ &= \Pr(x < y), \end{aligned}$$

whence

$$\mathrm{E}\left[\frac{1}{mn}U(x < y)\right] = \frac{1}{mn}\mathrm{E}\sum_{i}^{m}\sum_{j}^{n}u(x_i, y_j)$$
$$= \Pr(x < y).$$

Hereafter we write $\widehat{\Pr}(x < y)$ for $U(x < y)/mn$.

Now,

$$\widehat{\Pr}(x < y) = \frac{1}{mn}\sum_{i}^{m}\sum_{j}^{n}u(x_i, y_j)$$
$$= \frac{1}{n}\sum_{j=1}^{n}\left[\frac{1}{m}\sum_{i=1}^{m}u(x_i, y_j)\right].$$

Hence, applying (1),

$$\widehat{\Pr}(x < y) = \frac{1}{n}\sum_{j=1}^{n}F_m(y_j). \tag{7}$$

It follows that, as $m \to \infty$ [and hence $F_m(t) \to F(t)$],

$$\lim \widehat{\Pr}(x < y) = \frac{1}{n}\sum_{j=1}^{n}F(y_j). \tag{8}$$

From (8), for large m, $\widehat{\Pr}(x < y)$ is nearly an average of n independent identically distributed bounded random variables, and hence is asymptotically normally distributed as n (in addition to m) grows large (*see* **Central Limit Theory**).

From (7),

$$\widehat{\Pr}(x < y) = \frac{1}{n}\sum_{j=1}^{n}F_m(y_j) = \sum F_m \, \mathrm{d}G_n,$$

which estimates $\int F \, \mathrm{d}G$.

Examination of (8) yields a one-sample version of the Wilcoxon–Mann–Whitney test [4]. If F is known (say from census figures), then we can test whether a given set of data $(y_1, \ldots, y_n)$ comes from that distribution, against an alternative that $G(y)$ is some other distribution with $\int F \, \mathrm{d}G \neq 1/2$.

Under H_0, $G = F$ and the statistic $(1/n)\sum_{j=n}^{n}F(y_j)$ is a sum of **uniform** (0, 1) random variables; the statistic has mean 1/2, variance $1/12n$, and, if n is large (say 8 or more), its distribution is effectively normal, providing the test for H_0: $G = F$.

Some Properties of the Test

In comparison with the two-sample t test (*see* **Student's *t* Statistics**), the W-M-W enjoys a very strong property as a test against translation alternatives. First, if normality, with $\sigma_x^2 = \sigma_y^2$, governs the data, the **asymptotic relative efficiency** of the W-M-W procedure is $0.955 = 3/\pi$, which is nearly 1. Secondly, if the data come from a heavy-tailed distribution, then that efficiency rises *above* 1.0 and, for some distributions, much above 1.0. Hodges & Lehmann showed [1] that the asymptotic relative efficiency of W-M-W vs. the t test *never* falls below 0.864 for translation alternatives.

Some Practical Aspects

Where the distributions F and G are believed to differ by translation, a *confidence interval* for that translation can be constructed by a simple graphical procedure, based on W-M-W test theory [5].

The parameter $\Pr(x < y)$ and its estimate $\widehat{\Pr}(x < y)$ are sometimes readily interpretable. They are unit-free, and can serve as indicators of "effect size".

We have seen that $\widehat{\Pr}(x < y)$ is asymptotically normal, and **unbiased** for $\Pr(x < y)$ under H_0 when

$$\mathrm{var}[U(x < y)] = \frac{mn(N+1)}{12}$$

and

$$\mathrm{var}[\Pr(x < y)] = \frac{N+1}{12mn}.$$

Unfortunately, except when $F = G$, the standard error of $\widehat{\Pr}(x, y)$ is not a simple matter, though the following upper bound can be justified; $\mathrm{se}[\widehat{\Pr}(x < y)] = [p(1-p)]^{1/2}/k$, where p denotes $\Pr(x < y)$ and k denotes the smaller of m and n [7].

Tied Data (Discrete Probabilities)

When ties occur only in the xs or only the ys, they do not affect anything. However, where there are t observations, including at least one x and at least one y, sharing a common value, they are handled in the following manner. To calculate $U(x \leq y)$, count each tied pair (x_i, y_j) in the tied set as contributing 1/2 to $U(x \leq y)$.

To calculate $R(y)$, consider the t consecutive ranks that would be assigned were the tied data perturbed slightly to become distinct. The average of

those t distinct ranks is then assigned as the rank for every observation in that tied set. These two approaches are consistent – they lead to compatible values of R and U. The variance of R (and U) is somewhat reduced by the ties. Indeed, the variance appropriate for untied data is multiplied by CF (for "correction factor") as follows:

$$\mathrm{CF} = 1 - \frac{\sum(t^3 - t)}{N^3 - N},$$

where the sum runs over all the sets of x-with-y ties, and t denotes the length of such a tie.

If no x-with-y tie includes as many as half of the observations, the variance will need correction only in borderline situations, as the correction factor stays near 1.0 unless at least half the observations are in one tied set.

Example (Table 1)

$$\mathrm{CF} = 1 - \frac{\left\{\begin{array}{c}[(21^3 - 21) + (20^3 - 20)\\ +(23^3 - 23) + (24^3 - 24)]\end{array}\right\}}{88^3 - 88}$$
$$= 0.93665.$$

So the routine null standard $\sigma = [(22 \times 66 \times 89)/12]^{1/2}$ is reduced; it is multiplied by $\sqrt{0.93665} = 0.9678$.

For this table, $\widehat{\Pr}(x, y)$ is calculated thus:

$$\widehat{\Pr}(x < y) = \frac{\left\{\begin{array}{c}\{2(47) + 4(31) + 5(13) + \frac{1}{2}[2 \times 19\\ +4 \times 16 + 5 \times 18 + 11 \times 13\}\end{array}\right\}}{22 \times 66}$$
$$= 0.3103.$$

In the above calculation we have organized our work by choosing successive subgroups of x; thus, the 2 xs in "poor" form (x, y) pairs, where $x < y$, with $16 + 18 + 13 = 47$ ys, etc. The tied x, y pairs each contribute 1/2.

Table 1

	Poor	Fair	Good	Excellent		
x	2	4	5	11	22	m
y	19	16	18	13	66	n
t	21	20	23	24	88	N

The example illustrates a salient application of W-M-W methodology. It is mistaken to apply the usual **chi-square test** of significance where the categories have a relevant order, because that order, the key to the problem, is not taken into account by the χ^2 statistic. For these data $\chi_3^2 = 8.64$ ($p = 0.0345$), and the W-M-W statistic is

$$Z = \frac{0.31026 - 0.5000}{\left(\dfrac{89}{12 \times 22 \times 66}\right)^{1/2}(0.9678)} = -\frac{0.18974}{0.06917}$$
$$= -2.74 \ (2p = 0.0061).$$

A more detailed treatment of the ordered $2 \times k$ contingency table appears in [6] (*see* **Ordered Categorical Data**).

References

[1] Hodges, J.L. & Lehmann, E.L. (1956). The efficiency of some non-parametric competitors of the t-test, *Annals of Mathematical Statistics* **27**, 324–335.

[2] Kruskal, W.H. (1957). Historical notes on the Wilcoxon unpaired two-sample test, *Journal of the American Statistical Association* **52**, 356–360.

[3] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[4] Moses, L.E. (1964). One sample limits of some two sample rank tests, *Journal of the American Statistical Association* **59**, 645–651.

[5] Moses, L.E. (1965). Confidence limits from rank tests (query), *Technometrics* **7**, 257–260.

[6] Moses, L.E. (1986). *Think and Explain with Statistics*. Addison Wesley, Reading, pp. 184–187.

[7] Owen, D.B. (1962). *Handbook of Statistical Tasks*. Addison Wesley, Reading.

[8] Pratt, J.W. & Gibbons, J.D. (1981). *Concepts of Non-Parametric Theory*. Springer-Verlag, New York, pp. 264–265.

[9] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

(*See also* **Nonparametric Methods**; **Trend Test for Counts and Proportions**; **Wilcoxon-type Scale Tests**)

LINCOLN E. MOSES

Wilcoxon-type Scale Tests

Populations (distributions) are often described and compared on the basis of certain features or aspects called "parameters" (*see* **Estimation**). Two of the most widely used parameters are the location and the scale. The location parameter measures "central tendency" and represents the size of a typical observation, whereas the scale parameter indicates how variable or spread out the observations can be. Related to the idea of scale is the concept of dispersion, which indicates how close the observations are, on an average, to a central value. The **mean** and the **median** are two popular location parameters while the **standard deviation**, the mean absolute deviation (*see* **Mean Deviation**), the **range** and the interquartile range are often used to represent the scale or dispersion. If the application at hand permits a specific parametric model assumption (such as normality; *see* **Normal Distribution**) about the underlying distribution(s), suitably designed techniques can be used to make statistical **inference** regarding the location and/or the scale parameters. These methods are called parametric statistical methods of inference. On the other hand, if a complete model assumption is hard to justify (perhaps because not much is known about the populations), use of **nonparametric** or distribution-free methods is advocated. Nonparametric methods are often intuitively appealing and their implementation is quite simple. Of course, when the true distribution is indeed of a specific form, the nonparametric methods will be less **efficient** than their parametric counterparts, but the fact is that in many practical situations the form of the true distribution cannot be specified completely. Distribution-free methods for scale parameters are the focus of this article. We use the terms nonparametric and distribution-free interchangeably, since readers from different areas might be familiar with one of the terms and not the other.

One of the most popular distribution-free tests to compare the location parameters of two populations is the Wilcoxon rank-sum test. The rank-sum test statistic is linearly related to the Mann–Whitney statistic, so that the corresponding tests are equivalent (*see* **Wilcoxon–Mann–Whitney Test**). In the rank-sum form of the statistic, observations from two independent random samples drawn from two populations are combined into a single array and are **ranked** from the lowest to the highest, keeping track of whether each observation was from the first sample or the second. When ties occur (i.e. more than one of the observations have the same value), the average of the tied ranks is assigned to each of the tied observations. The rank-sum test is based on the sum of ranks of the observations that are, say, from the first sample. In the Mann–Whitney form of the statistic, each observation from the first sample is compared with each observation from the second, and, assuming there are no ties, a score of 1 or 0 is given depending on whether or not the observation from the first sample is larger or smaller than the observation from the second. When there is a tie, a score of 1/2 is assigned to that comparison. The test statistic is simply the sum of these scores summed over all such comparisons.

The idea behind the Wilcoxon–Mann–Whitney (WMW) test has been extended to the problem of testing for scale differences between two populations and such tests are referred to as Wilcoxon-type scale tests. Here observations are often first "centered" by subtracting some measure of the central tendency and the WMW tests is applied to the absolute values of the deviations. A nice feature of the Wilcoxon-type tests is that they lead to a **confidence interval** for the ratio of the scale parameters, which can be used for a test of **hypothesis** as well as for **estimation** purposes.

A review of Wilcoxon-type scale tests is given by Gibbons [6]. Discussions on various nonparametric tests for scale can be found, for example, in [8, Chapter 9] and [7, Chapter 10]. A general review of scale tests, including distribution-free tests, is given by Fligner [4], who also gives examples of situations where inference on the scale parameter is of interest.

Assumptions

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent **random samples** from continuous populations from the **"location-scale" family**, with cumulative distribution functions (cdfs) $F_1(x) = F\{(x - \theta_x)/\tau_x\}$ and $F_2(y) = F\{(y - \theta_y)/\tau_y\}$, where is some unknown continuous cdf, θ_x and θ_y are the respective location parameters, and τ_x and τ_y are the respective scale parameters. Thus, F_1 and F_2 are assumed to have the same shape, but they could have different location and/or scale parameters. To understand this better, note that we can write $X_i = \theta_x + \tau_x Z$ and

$Y_j = \theta_y + \tau_y Z$, where Z is a continuous random variable with cdf $F(\cdot)$. Therefore, X and Y can be viewed as linear functions of the random variable Z and it is clear that a change in the location parameter(s) shifts the center of the distribution(s), whereas increasing (decreasing) the scale parameter(s) causes the distribution(s) to be more (less) spread out. While the locations of two distributions are usually compared on the basis of the difference $\theta_x - \theta_y$, the scale comparison is often done in terms of the ratio $\gamma = \tau_x/\tau_y$. We are interested in distribution-free methods of inference about γ. Both tests of hypothesis and confidence intervals are considered. The latter can also be used for estimation purposes.

Tests: Locations Unknown

For ease of presentation, suppose that the location parameters are the respective medians of F_1 and F_2. Consider the situation in which the two medians are known, and we would like to test the **null hypothesis** $H_0 : \gamma = \gamma_0$, where γ_0 is some specified value. If only the equality of the scale parameters is of interest, γ_0 is set equal to 1.

Two tests are available, both based on the absolute values of the adjusted variables $S_i = (X_i - \theta_x)/\gamma_0, i = 1, 2, \ldots, m$ and $T_j = (Y_j - \theta_y), \quad j = 1, 2, \ldots, n$. Absolute values of the differences are used, since they reflect dispersion about the respective medians. Under the null hypothesis the distributions of $|S_i|$ and $|T_j|$ are identical, which implies that their medians must be the same. This leads us to proceed as in the case of the usual Wilcoxon rank-sum test for location, except that now we work with the absolute values of the deviations from the medians. Therefore, we arrange the $|S_1|, |S_2|, \ldots, |S_m|$ and $|T_1|, |T_2|, \ldots, |T_n|$, from the lowest to the highest, and assign ranks $1, 2, \ldots, m + n$ to each of the ordered values (the lowest gets rank 1, the highest gets rank $m + n$; use average ranks if there are ties), keeping track of whether an observation was an S or a T. The test statistic is W^*, the sum of the ranks of the absolute values of the Ss. The test is consistent when the $F(\cdot)$ is symmetrically distributed about 0; that is, when F_1 and F_2 are symmetrically distributed about θ_x and θ_y, respectively. The rejection regions for the test can be argued as follows. Letting F_1^* and F_2^* represent the cdfs of $|S|$ and $|T|$ respectively, it can be seen that when γ is greater (less) than γ_0, F_1^* is stochastically larger (smaller) than F_2^*. Thus, when the **alternative hypothesis** is $\gamma > (<)\gamma_0$, we should reject H_0 if W^* is large (small).

The **critical** values as well as the P values can be found from the distribution of W^* under the null hypothesis. As explained earlier, the null distribution of W^* is the same as that of the Wilcoxon rank-sum test statistic for comparing two location parameters. The latter has been shown to depend only on the sample sizes m and n and has been studied and tabulated by several authors, the most extensive tabulation being available in [16]. These tables can be used to implement the test. Note, however, that since W^* is a discrete random variable, not all commonly used levels of significance might be exactly achievable for all m and n (*see* **Level of a Test**). When the sample sizes are large, a normal approximation to the critical value or the P value can be found. This is based on the fact that the distribution of the "standardized" random variable

$$Z^* = \frac{W^* - m(m+n+1)/2}{[mn(m+n+1)/12]^{1/2}} \tag{1}$$

can be approximated by the normal distribution with mean 0 and variance 1 (the **standard normal** distribution). Using the normal approximation is convenient in practice, since the standard normal tables are widely available. The normal approximation-based rejection regions are given in Table 1. In situations in which a number of ties are present in the data, it is advisable to "correct" the null variance of W^* and use the corrected variables in the standardized statistic Z^*. The correction for ties is detailed, for example, [8, p. 300].

A continuity correction of 0.5 can improve the approximation. The **asymptotic relative efficiency** (ARE) of this test relative to the normal theory F test (*see* ***F* Distributions**) is $6/\pi^2 = 0.61$, when the underlying populations are normal.

The second test that one could use with known medians is the Sukhatme [13] test. Here the positive Ss and Ts are put in one group and the negative Ss

Table 1 Rejection regions for various alternative hypotheses

Alternative hypothesis	Rejection region
$H_a : \gamma > \gamma_0$	$Z^* > z_\alpha$
$H_a : \gamma < \gamma_0$	$Z^* < -z_\alpha$
$H_a : \gamma \neq \gamma_0$	$Z^* > z_{\alpha/2}$ or $Z^* < -z_{\alpha/2}$

and Ts are put in a second group. The first group of values are ranked and we let S^+ be the sum of the ranks of the Ss.

For the second group, absolute values are ranked and we let S^- be the sum of the ranks of the Ss. The Sukhatme test is based on $W_S = T^+ + T^-$. Actually, the original Sukhatme [13] test is based on a Mann–Whitney-type "***U*-statistic**", which is a linear function of W_S. The exact distribution of W_S is complicated, since this must be obtained conditioned on the number of positive (or negative) deviations, which itself is a random variable. For large m and n, Gibbons [6] states that normal approximation can be used with mean $m(m+n+2)/4$ and variance $mn(m+n+7)/48$, to find an approximate critical value or a P value. Sukhatme [13] showed that the test is consistent without the assumption of symmetry. When the underlying distributions are normal, the ARE of the Sukhatme test relative to the F test is 0.61, which is the same as the ARE of the W^* test relative to the F test. On the other hand, when the underlying distributions are double-**exponential**, the ARE of the Sukhatme test relative to the F test is 0.94. Laubscher & Odeh [11] showed that the normal approximation to W_S is good for m and n larger than 10, when used with a **continuity correction**. They also considered a statistic, which is a linear function of W_S and tabulated the exact critical values for $2 \le m,\ n \le 10$.

To summarize, when the populations can be assumed to be symmetrically distributed about their known medians, one should use the test based on W^*. When symmetry cannot be assumed, the test based on W_S should be used.

Tests: Locations Unknown

In many practical applications, the underlying medians θ_x and θ_y are likely to be unknown. First, assume that the two population medians are the same but the common median is unknown. Fligner & Kileen [5] suggested using the median of the combined sample of $X's$ and $Y's$, say M, to estimate the common median and applying the Wilcoxon rank-sum test, to the absolute deviations $|X_i - M|$ and $|Y_j - M|$. The resulting test, although not a linear rank test, is distribution-free and the same tables for the Wilcoxon rank-sum test, cited earlier, can be used to implement it. They showed that the test is consistent whether or not the underlying distributions are symmetric. In a small sample study with normal and double exponential distributions, the **power** of their test was found to be significantly higher than some popular linear rank tests for scale. Moreover, when the medians are unknown and unequal, their test was shown to be consistent for scale differences as long as the populations are symmetrically distributed and the sample sizes are equal.

In situations in which the population medians are both unknown and cannot be assumed to be equal, one approach would be to use the sample medians M_x and M_y to calculate $|X_i - M_x|$ and $|Y_j - M_y|$ and apply the W^* test on these absolute deviations. However, such a test might not always be distribution-free, not even when the sample sizes are large; see [4, 12] for some discussions on this issue and recommendations for tests to be used in this case.

Sukhatme's test can also be used when medians are unknown and unequal, by applying the test based on W_S discussed earlier, to $|X_i - M_x|$ and $|Y_j - M_y|$. Sukhatme [14] showed that the resulting test has the same asymptotic null distribution as that of W_S and is asymptotically distribution-free when F_1 and F_2 are symmetrically distributed about their respective medians and have bounded density functions. Fligner [4] remarked that in terms of power, these rank-like tests based on the absolute values of the estimated deviations do not compare very well with what are called *robust* (*see* **Robustness**) tests for scale. See his paper for proposed robust tests.

Confidence Intervals

If the two medians are known and the populations can be assumed to be symmetrically distributed about their respective medians, a confidence interval for γ can be constructed from the ratios $|X_i - \theta_x|/|Y_j - \theta_y|$. The procedure is similar to that of finding a confidence interval for the difference between two population medians based on the WMW test. Here, the mn values of the ratio are arranged from the lowest to the highest. The $100(1-\alpha)\%$ confidence interval is obtained simply by locating the uth smallest and the uth largest of the mn ratios. For a given confidence level, the integer u is calculated from the null distribution of a two-sample Wilcoxon rank-sum test. In fact, as described in [8], it is often convenient to think of u *as the rank of*

a left-tail critical value for the Wilcoxon rank-sum test, at a level of significance $\alpha/2$. When m and n are small, an exact table for the null distribution of the rank-sum test statistic should be used to find the confidence interval. Table J in [8] is useful for this purpose. To illustrate, suppose that $m = 4$, $n = 7$, and that a 95% confidence interval is desired. Then we have $\alpha/2 = 0.025$ and, from Table J in [8], we find that the closest we can come to the desired confidence level is $100(1 - 2 \times 0.021) = 95.8\%$ and this corresponds to a left-tail critical value of 13, which has the rank of 4. Thus, $u = 4$ and a 95.8% confidence interval for it is obtained by arranging the $mn = 28$ values of the ratio in an increasing order and locating the fourth smallest and the fourth largest (or the $28 - 4 + 1 = 25$th value from the smallest). This also highlights the fact that for small sample sizes, the desired confidence coefficient often cannot be achieved exactly. When m and n are moderately large, the normal approximation, with a continuity correction, to u can be found as

$$u = \left[\frac{mn}{2} + 0.5 - z_{\alpha/2} \left(\frac{mn(m+n+1)}{12} \right)^{1/2} \right], \tag{2}$$

where $[a]$ denotes the greatest integer in a (e.g. if $a = 3.5$, then $[a] = 3$) and $z_{\alpha/2}$ is obtained from the standard normal table, so that $\alpha/2$ is the area to the right of $z_{\alpha/2}$. For our example, $u = [4.12] = 4$, so that the normal approximation yields the same solution as the exact answer.

Corresponding to the Sukhatme test, the confidence interval for γ is obtained from the ratios $(X_i - \theta_x)/(Y_j - \theta_y)$ that are positive. It may be noted that this ratio will be positive when both the numerator and the denominator have the same sign. These positive ratio values are arranged from the lowest to the highest, and the $100(1 - \alpha)\%$ confidence interval is given by the uth smallest and the uth largest of the positive ratios. The quantity u can be found from the tables of Laubscher & Odeh [11]. For large m and n, the normal approximation to u with a continuity correction, can be found from

$$u = \left[\frac{mn}{4} + 0.5 - z_{\alpha/2} \left(\frac{mn(m+n+7)}{48} \right)^{1/2} \right]. \tag{3}$$

Generalizations

Deshpande & Kusum [2] considered a generalized version of Sukhatme's test where the location parameter is some **quantile**. When the scale parameters of more than two populations need to compared, the well-known Kruskal–Wallis test (*see* **Nonparametric Methods**) can be used after transforming the data into absolute values of the deviations from the respective locations parameters. Tsai et al. [15] compared the performance of this procedure with other multisample tests for scale. Duran [3] gave a comprehensive survey of distribution-free tests for scale.

In some situations, we are interested in comparing both the location and the scale parameters. There is a body of literature for distribution-free tests for this type of problem; see [9] for a review.

Blair & Thompson [1], following Moses [12], considered a class of distribution-free rank-like tests for scale differences when the locations are unknown. Their tests do not require the assumptions on the location parameters discussed earlier. In addition, the tests are "robust for **skewed** data, are resolving and have significant power advantages". One of their test statistics is the Wilcoxon rank-sum test applied to the absolute values of the differences $|X_i - X_j|$. They studied asymptotic properties of the tests, including a normal approximation and the asymptotic relative efficiency. The asymptotic relative efficiency was calculated for several distributions. An extension to the multisample problem was also considered.

Kössler [10], also following Moses [12], proposed a distribution-free test in the case of unequal and unknown location parameters. The two samples are randomly separated into groups of size k and the Wilcoxon test or the Savage test is applied to the ranges or the variances of the subgroups; $k = 4$ is recommended. Asymptotic power function of the proposed test is derived and **simulation** studies are carried out for small to moderate sample sizes. Power calculations suggest using the Wilcoxon test for original densities with long tails and using the Savage test for original densities with small or medium tails.

References

[1] Blair, R.C. & Thompson, G.L. (1992). A distribution-free rank-like test for scale with unequal population locations, *Communications in Statistics – Theory and Methods* **21**, 353–371.

[2] Deshpande, J.V. & Kusum, K. (1984). A test for the nonparametric two sample scale problem, *Australian Journal of Statistics* **26**, 16–24.

[3] Duran, B.S. (1976). A survey of nonparametric tests for scale, *Communications in Statistics – Theory and Methods* **5**, 287–1312.

[4] Fligner, M.A. (1988). Scale tests, in *Encyclopedia of Statistical Sciences*, Vol. 8, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 271–278.

[5] Fligner, M.A. & Kileen, T.J. (1976). Distribution-free two-sample tests for scale, *Journal of the American Statistical Association* **71**, 210–213.

[6] Gibbons, J.D. (1988). Wilcoxon-type scale tests, in *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 616–619.

[7] Gibbons, J.D. (1996). *Nonparametric Methods for Quantitative Analysis*, 3rd Ed. American Sciences Press, Columbus.

[8] Gibbons, J.D. & Chakraborti, S. (2003). *Nonparametric Statistical Inference*, 4th Ed. Marcel Dekker, New York.

[9] Goria, M.N. (1982). A survey of two-sample location-scale problem, asymptotic relative efficiencies of some rank tests, *Statistica Neerlandica* **36**, 3–13.

[10] Kössler, W. (1999). Rank tests in the two-sample scale problems with unequal and unknown locations, *Statistical Papers* **40**, 13–35.

[11] Laubscher, N.F. & Odeh, R.E. (1986). A confidence interval for the scale parameter based on Sukhatme's two-sample statistic, *Communications in Statistics – Theory and Methods* **5**, 1393–1407.

[12] Moses, L.E. (1963). Rank tests of dispersion, *Annals of Mathematical Statistics* **34**, 973–983.

[13] Sukhatme, B.V. (1957). On certain two sample non-parametric tests for variances, *Annals of Mathematical Statistics* **28**, 188–194.

[14] Sukhatme, B.V. (1958). Testing the hypothesis that two populations differ only in scale, *Annals of Mathematical Statistics* **29**, 60–78.

[15] Tsai, W.S., Duran, B.S. & Lewis, T.O. (1975). Small sample behavior of some multisample nonparametric tests for scale, *Journal of the American Statistical Association* **70**, 791–796.

[16] Wilcoxon, F., Katti, S.K. & Wilcox, R.A. (1972). *Selected Tables in Mathematical Statistics*, Vol. 1. American Mathematical Society, Providence, pp. 171–259.

SUBHA CHAKRABORTI

Window Estimate

The most common occurrence of this terminology is in the context of frequency domain estimation in **time series** analysis. Suppose that an observed time series is a realization of a stationary random process $\{Y_t\}$ with spectral density $f(\omega) = (1/2\pi)\sum_{k=-\infty}^{\infty}\gamma_k\cos(k\omega)$, where γ_k is the autocovariance function of $\{Y_t\}$. The fundamental tool for estimating $f(\omega)$ is the periodogram, $I(\omega)$, which can be shown to be estimated by the discrete Fourier transform of the sample **autocorrelation function** r_k, of the observed series $(Y_t;\ t = 1, \ldots, n)$. It is given by $c_0\{1 + 2\sum_{k=1}^{n-1} r_k\cos(k\omega)\}$, where c_0 is the sample autocovariance at lag k [8, p. 56]. There are certain unsatisfactory statistical properties possessed by $I(\omega)$, such as inconsistency [11, p. 426]. The periodogram (*see* **Spectral Analysis**) at Fourier frequencies of the form $\omega_j = 2\pi j/n,\ j = 1, \ldots, n/2$, are asymptotically independent of each other [8, p. 96], so that if we take a simple average of the estimates $I(\omega_j)$, the resultant statistic $\hat{f}(\omega_j)$ will gain in precision as n increases. More generally, *weighted* averages are used. Thus, an equivalent way of improving the estimates is provided by introducing a nonincreasing sequence of weights, λ_k called a *lag window*, and defining $\hat{f}_\lambda(\omega) = c_0\{1 + 2\sum_{k=1}^{n-1}\lambda_k r_k\cos(k\omega)\}$. The function $\hat{f}_\lambda(\omega)$ is called the *spectral window*, after Blackman & Tukey [2].

A good deal of effort has been put into investigating the statistical properties of windowed estimates and choosing the sequence λ_k. More than 11 different lag windows are reviewed by Priestley [11, Section 6.2.3]. However, using lag windows is no longer fashionable, mainly because of computational considerations: since the periodogram is essentially a discrete Fourier transform, it can be computed very quickly using a Fast Fourier transform (FFT) [7]. In the early days of this development in spectral analysis, computing power was relatively expensive, and using lag windows reduced this cost. It is now more logical to compute the periodogram first, perhaps using a FFT, and then to smooth it. A more recent development for locally adaptive windows is given by Buhlmann [4].

The idea of window estimates also arises in the context of time domain estimation in time series analysis. Let $Y_1, Y_2, \ldots, Y_n$ be n observations from a time series $\{Y_t\}$ with mean μ, theoretical autocorrelation function ρ_k, and variance σ_Y^2. Estimates of these and other summary statistics for a time series usually use all n observations, but it is sometimes useful to divide the data up into segments or *windows* as follows. Consider a subset of m observations in the data as indicated in $Y_1, Y_2, \ldots, \boxed{Y_l, \ldots, Y_{l+m-1}}, Y_{l+m}, \ldots, Y_n$. If we let $l = 1, 2, \ldots, (n - m + 1)$ then this represents a (forward) moving window of fixed length m. If we fix $l = 1$, and let $m = p + 1, p + 2, \ldots, n$ for fixed p, this represents a window of increasing size with starting length p. The estimates of statistics within the moving window can be useful to judge whether those statistics are changing through time. In general, the statistics can be plotted against window number, and a visual assessment made of whether they are changing. Formal significance tests on the window estimates can be carried out, but these are complicated by the fact that successive estimates will be highly correlated. In the case of the window increasing in size from a fixed starting length, p, the estimates are sometimes called *recursive*. The use of recursive residuals in time series regression was pioneered by Brown et al. [3] and generalized for lagged values of the dependent variable by Kramer et al. [9] and Ploberger & Kramer [10]. Another approach is to make one- or multistep predictions of data outside the window and judge whether these are consistent with the data that actually occur.

The term *window* is also used in kernel **density estimation**. The classic books on this topic are Silverman [12] and Wand & Jones [13]. Suppose that $X_1, X_2, \ldots, X_n$ is a set of continuous random variables having common density f. A nonparametric density estimator assumes no prespecified functional form for f. Such an estimator is given by $\hat{f}(x;h) = (nh)^{-1}\sum_{i=1}^{n} K\{(x - X_i)/h\}$. Here, K is a function satisfying $\int K(x)\,dx = 1$, called the kernel, and h is a positive number called the *bandwidth* or *window width*. The simplest and most common density estimate is the histogram defined by $\hat{f}(x;h) = (nh)^{-1}$ (number of X_i in the same bin as x). Each bin, or rectangle, may be regarded as a window that summarizes the behavior of the data between the bin extremities. The choice of origin and h can greatly affect the appearance of the histogram. In the general case defined above, much effort has been put into determining optimal values for h [13, Chapter 3].

So-called *kernel* and *regression* smoothers are designed locally to smooth data by slicing through it in windows of fixed width. See, for example, [1] for an introduction to the topic. Special cases include average smoothers that use the within-slice average to summarize the data in a slice, and regression smoothers that use a fixed proportion of the data in a specified neighborhood. In this context, Cleveland [5] introduced the *lo*cally *we*ighted *s*catterplot *s*moother (*lowess*) and this has become the most commonly used method in modern graphical and regression analyses. See, for example, [6, p. 31].

References

[1] Altman, N.S. (1992). An introduction to kernel and nearest neighbour nonparametric regression, *American Statistician* **46**, 175–185.

[2] Blackman, R.B. and Tukey, J.W. (1959). *The Measurement of Power Spectra*. Dover, New York.

[3] Brown, R.L., Durbin, J. & Evans, J.M. (1975). Techniques of testing the constancy of regression relationships over time, *Journal of the Royal Statistical Society, Series B* **37**, 141–192.

[4] Buhlmann, P. (1996). Locally adaptive lag-window spectral estimation, *Journal of Time Series Analysis* **17**, 247–270.

[5] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

[6] Cook, R.D. & Weisberg, S.W. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

[7] Cooley, J.W. & Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series, *Mathematics of Computation* **19**, 297–301.

[8] Diggle, P. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

[9] Kramer, W. Ploberger, W. & Alt, R. (1988). Testing for structural change in dynamic models, *Econometrica* **56**, 1355–1369.

[10] Ploberger, W. & Kramer, W. (1992). The CUSUM test with OLS residuals, *Econometrica* **60**, 271–285.

[11] Priestley, M.B. (1981). *Spectral Analysis and Time Series*, Vol. 1. Academic Press, London.

[12] Silverman, B.W. (1985). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[13] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

NEVILLE DAVIES

Wishart Distribution

Suppose that $X_1, X_2, \ldots, X_n$ is a **random sample** from a normal population with mean μ and variance σ^2. Also let $f = n - 1$, the sample mean $\overline{X} = \sum X_i/n$, and the sample variance $s^2 = \sum(X_i - \overline{X})^2/f$. It is known that fs^2/σ^2 is distributed as χ_f^2, where χ_f^2 represents a **chi-square distributed** random variable with f **degrees of freedom**. In other words, $v = fs^2 = \sum(X_i - \overline{X})^2$ is distributed as $\sigma^2\chi_f^2$. The probability density function (pdf) of v is given by

$$f(v) = \frac{1}{\Gamma(f/2)(2\sigma^2)^{f/2}} v^{(f-2)/2} \exp\left(\frac{-v}{2\sigma^2}\right) \propto \frac{v^{(f-2)/2}}{(2\sigma^2)^{f/2}} \exp\left(\frac{-v}{2\sigma^2}\right), \quad v > 0. \tag{1}$$

It is a **gamma distribution** with parameters $f/2$ and $2\sigma^2$.

Let us now consider a **bivariate normal** random variable $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and **covariance matrix** $\boldsymbol{\Sigma}$, where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is a random sample on $\mathbf{X}$. Define the sums of squares v_{11}, v_{22}, and the sum of cross-products v_{12} as follows:

$$v_{11} = (n-1)s_1^2 = \sum(X_{1j} - \overline{X}_1)^2,$$

$$v_{22} = (n-1)s_2^2 = \sum(X_{2j} - \overline{X}_2)^2,$$

$$v_{12} = v_{21} = (n-1)s_{12} = \sum(X_{1j} - \overline{X}_1)(X_{2j} - \overline{X}_2).$$

Also, let

$$\mathbf{V} = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}.$$

Then the joint distribution of v_{11}, v_{22}, and v_{12} is given by

$$f(\mathbf{V}) \equiv f(v_{11}, v_{22}, v_{12}) = c\frac{|\mathbf{V}|^{(f-3)/2}}{|\boldsymbol{\Sigma}|^{f/2}} \times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{V}\right)\right], \tag{2}$$

where both $\mathbf{V}$ and $\boldsymbol{\Sigma}$ are positive definite. Here, c is a normalizing factor which is defined later.

For the sake of comparison with (1), let us assume that $\sigma_{12} = 0$. Then (2) reduces to

$$f(v_{11}, v_{22}, v_{12}) \propto \frac{(v_{11}v_{22} - v_{12}^2)^{(f-3)/2}}{(4\sigma_1^2\sigma_2^2)^{f/2}} \times \exp\left[-\frac{1}{2}\left(\frac{v_{11}}{\sigma_1^2} + \frac{v_{22}}{\sigma_2^2}\right)\right].$$

The marginal distribution of v_{11} or v_{22} is (1).

Assume that $\mathbf{X}$ is now a p-variate **multivariate normal** random vector with mean $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$. Define

$$\mathbf{V} = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'.$$

The distribution of $\mathbf{V}$ is given by

$$f(\mathbf{V}) = c\frac{|\mathbf{V}|^{(f-p-1)/2}}{|\boldsymbol{\Sigma}|^{f/2}} \exp\left[-\frac{1}{2}\ \mathrm{trace}\ \left(\boldsymbol{\Sigma}^{-1}\mathbf{V}\right)\right], \tag{3}$$

where both $\mathbf{V}$ and $\boldsymbol{\Sigma}$ are positive definite, and $f(\mathbf{V}) = 0$ otherwise. The normalizing constant is

$$c = \left[2^{fp/2}\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\left(\frac{f+1-j}{2}\right)\right]^{-1}.$$

The distribution with the pdf (3) is called the Wishart distribution and is denoted as $W(\boldsymbol{\Sigma}, p, f)$ or $W_p(f; \boldsymbol{\Sigma})$. It has the first moment

$$\mathrm{E}(\mathbf{V}) = f\boldsymbol{\Sigma}.$$

Furthermore, if $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m$ are independently distributed as $W(\boldsymbol{\Sigma}, p, f_k)$, $k = 1, 2, \ldots, m$, then the sum $\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2 + \cdots + \mathbf{V}_m$ is distributed as $W(\boldsymbol{\Sigma}, p, \sum f_k)$. A detailed derivation of the pdf and its properties are given by Johnson & Kotz [1].

Reference

[1] Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

AUSTIN F.S. LEE

Women's Health Initiative: Statistical Aspects and Selected Early Results

Introduction

The women's health initiative (WHI) is perhaps the most ambitious population research investigation ever undertaken. The centerpiece of the WHI program is a randomized, controlled **clinical trial** (CT) to evaluate the health benefits and risks of three distinct interventions (dietary modification, postmenopausal hormone therapy, and calcium/vitamin D supplementation) among 68 132 postmenopausal women. Participating women were identified from the general population living in proximity to any one of 40 participating clinical centers throughout the United States. The WHI program also includes an **observational study** (OS) comprising 93 676 postmenopausal women recruited from the same population base as the CT. Enrollment into WHI began in 1993 and concluded in 1998. Intervention activities in the combined hormone therapy component of the CT ended early in July 2002 when evidence had accumulated that the risks exceed the benefits for combined hormone therapy. Follow-up on all participating women is planned through March 2005, giving an average follow-up duration of about 8.5 years in the CT and 7.5 years in the OS.

WHI Clinical Trial and Observational Study

The WHI CT includes three overlapping components, each a randomized controlled comparison among women who were postmenopausal and in the age range of 50 to 79 at randomization. The dietary modification (DM) component randomly assigned 48 835 (target 48 000) eligible women to either a sustained low-fat eating pattern (40%) or self-selected dietary behavior (60%), with breast cancer and colorectal cancer as designated primary outcomes and coronary heart disease as a secondary outcome (*see* **Outcome Measures in Clinical Trials**). From the outset, the nutrition goals for women assigned to the DM intervention group have been to reduce total dietary fat to 20%, and saturated fat to 7% of corresponding daily calories and, secondarily, to increase daily servings of vegetables and fruits to at least five and of grain products to at least six, and to maintain these changes throughout trial follow-up. The randomization of 40%, rather than 50%, of participating women to the DM intervention group was intended to reduce trial costs, while testing trial hypotheses with specified **power**.

The postmenopausal hormone therapy (PHT) component is composed of two parallel randomized, double-blind (*see* **Blinding or Masking**) trials among 27 347 (target 27 500) women, with coronary heart disease (CHD) as the primary outcome, with hip and other fractures as secondary outcomes, and with breast cancer as a potential adverse outcome. Of these, 10 739 (39.3% of total) had a hysterectomy prior to randomization, in which case there was a 1 : 1 randomized double-blind allocation between conjugated equine estrogen (E-alone) 0.625 mg/day or **placebo**. The remaining 16 608 (60.7%) of women, each having a uterus at baseline, were randomized 1 : 1 to the same preparation of estrogen plus continuous 2.5 mg/day of medroxyprogesterone (E + P) or placebo. These numbers compare to design goals of 12 375 for the E-alone comparison, and 15 125 for the E + P comparison, based on an assumption that 45% of women would be post hysterectomy. Over 8000 women were randomized to both the DM and PHT clinical trial components.

At their one-year anniversary from DM and/or PHT trial enrollment, all women were further screened for possible randomization in the calcium and vitamin D (CaD) component, a randomized double-blind trial of 1000 mg elemental calcium plus 400 international units of vitamin D_3 daily, versus placebo. Hip fracture is the designated primary outcome for the CaD component, with other fractures and colorectal cancer as secondary outcomes. A total of 36 282 (53.3% of CT enrollees) were randomized to the CaD component. While the WHI design estimated that about 45 000 women would enroll in the CaD trial component, protocol planning activities also included projected **sample sizes** of 35 000 and 40 000 and noted that most WHI objectives could be met with these smaller sample sizes.

The total CT sample size of 68 132 is only 60.6% of the sum of the individual sample sizes for the three CT components, providing a cost and logistics

justification for the use of a partial **factorial design** with overlapping components.

Postmenopausal women of ages 50 to 79 years, who were screened for the CT but proved ineligible or unwilling to be randomized, were offered the opportunity to enroll in the observational study (OS). The OS is intended to provide additional knowledge about risk factors for a range of diseases, including cancer, cardiovascular disease, and fractures. It has an emphasis on biological markers of disease risk, and on risk factor changes as modifiers of risk.

There was also an emphasis on the recruitment of women of racial/ethnic minority groups. Overall 18.5% of CT women and 16.7% of OS women identified themselves as other than white. These fractions allow meaningful study of disease risk factors within certain minority groups in the OS. Also, key CT subsamples are weighted heavily in favor of the inclusion of minority women in order to strengthen the study of intervention effects on specific intermediate outcomes (e.g. changes in blood lipids or micronutrients) within minority groups.

Age distribution goals were also specified for the CT as follows: 10%, ages 50 to 54 years; 20%, ages 55 to 59 years; 45%, ages 60 to 69 years; and 25%, ages 70 to 79 years. While there was substantial interest in assessing the benefits and risks of each CT intervention over the entire 50 to 79 year age range, there was also interest in having sufficient representation of younger (50–54 years) postmenopausal women for meaningful age group-specific intermediate outcome (biomarker) studies, and of older (70–79 years) women for studies of treatment effects on quality of life measures, including aspects of physical and cognitive function. Differing age and **incidence rates** within the 50 to 79 age range, and across the outcomes that were hypothesized to be affected, favorably or unfavorably, by the interventions under study, provided an additional motivation for a prescribed age-at-enrollment distribution.

The enrollment of such a large number of women, meeting designated **eligibility and exclusionary criteria** [11] for each CT component and for the OS proved to be a challenge, particularly for the PHT component of the CT, since many women who volunteered for WHI were already taking hormones and did not wish to be randomized to take hormones or placebo, while other women had already made a decision against the use of hormones.

Study Organization

In addition to the clinical centers, the study is implemented through a Clinical Coordinating Center (CCC) located in Seattle with various collaborators providing specific expertise, as described below (*see* **Multicenter Trials**). The National Heart Lung and Blood Institute (NHLBI) sponsors the program with input from the National Cancer Institute, the National Institute of Aging, the National Institute of Arthritis and Musculoskeletal and Skin Diseases, the NIH Office of research on women's health, and the NIH director's office. A steering committee, consisting of the principal investigators of the 40 CCs, CCC, and NHLBI representatives are responsible for major scientific and operational decisions. An executive committee identifies, prioritizes, and coordinates items for the steering committee discussion. Program activities are implemented through a regional organization that categorizes CCs geographically (West, Midwest, Northeast, Southeast). Principal investigators, and staff groups defined by project responsibilities (clinic manager, clinic practitioner, nutritionist, recruitment coordinator, data coordinator, outcomes coordinator) meet regularly through conference calls within regions to discuss implementation plans and issues, and regional staff group representatives also confer regularly to ensure national coordination. Nine standing advisory committees (behavior, calcium and vitamin D, design and analysis, dietary modification, hormone therapy, morbidity and mortality, observational study, publications and presentations, special populations) composed of study investigators having expertise in the major substantive areas involved in the program, provide recommendations on relevant issues as they arise. The CCC participates and provides liaison support in these various contexts. Figure 1 shows the WHI governance more generally, including NIH advisory committees. Specifically, the directors of participating NIH institutes and offices form a consortium that advises the NHLBI director concerning the WHI. A special working group of the NHLBI council also advises the NHLBI director concerning the WHI.

Principal Clinical Trial Comparisons, Power Calculations, and Safety and Data Monitoring

This section provides sample sizes by age for each CT component and for the OS, and provides power

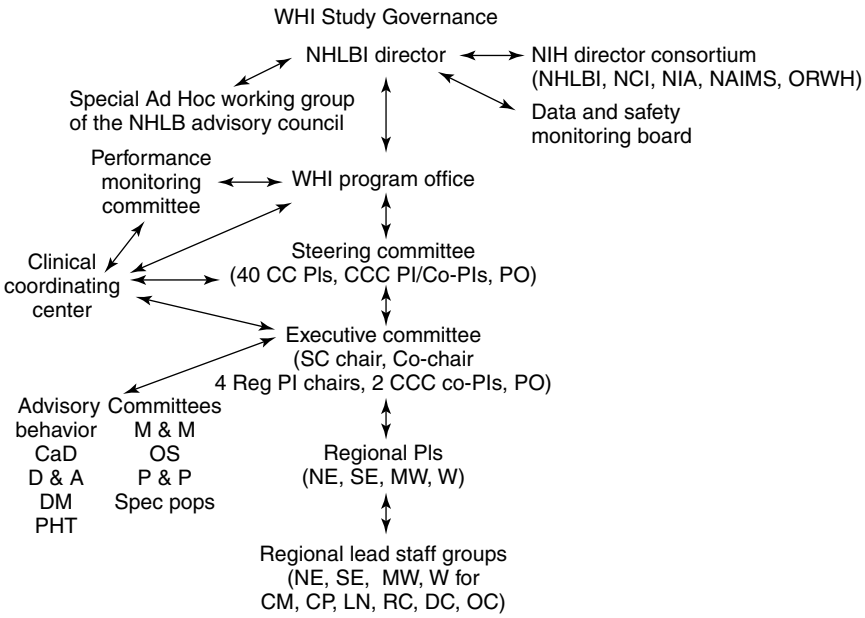

Figure 1 Organizational structure of the Women's Health Initiative. NHLBI: National Heart, Lung, and Blood Institute; NIH: National Institutes of Health; PI: Principal Investigator; SC: Steering Committee; CC: Clinical Center; CM: Clinic Manager; LN: Lead Nutritionist: CP: Clinic Practitioner; DC: Data Coordinator; OC: Outcomes Coordinator

calculations for key outcomes for each continuing CT component. Relative to the basic WHI design manuscript [11], these calculations have been updated to reflect the sample size and age distribution achieved, and to reflect the actual average follow-up duration, which will be realized by March 2005.

The target sample sizes noted above were based on consideration of the probability of rejecting the **null hypothesis** of no treatment effect (i.e. power) on the designated primary outcome under a set of design specifications concerning age-specific control group primary outcome incidence rates, intervention effects on incidence rates as a function of time from randomization, intervention adherence rates, and **competing risk** mortality rates. These assumptions have previously been listed in [11] where an extensive bibliography is cited providing the rationale for these assumptions.

The power calculations were based on weighted **logrank** statistics that accumulate the differences between the observed numbers of primary outcome events in the intervention group and the expected number of such events under the null hypothesis, across the follow-up time period. Early events that may be less likely to be affected by intervention activities are downweighted relative to later events. Specifically, the observed minus expected differences are weighted linearly from zero at randomization to a maximum value of one at a certain time from randomization and are constant(at one) thereafter. For cardiovascular disease and fracture incidence, this "certain time" was taken to be three years, whereas for cancer and mortality, it was taken to be 10 years. For coronary heart disease, incidence of the event times are grouped into three-year follow-up periods, in order to accommodate the inclusion of silent myocardial infarctions detected by routine electrocardiograms, which are to be obtained at baseline and every three years during follow-up for CT participants. A weighted **odds ratio** test statistic is then used to acknowledge this grouping.

Table 1 shows the number of enrollees, and percentages of the total, by age category for each component of the CT and the OS. Note the degree of correspondence to the target age distribution, especially in the PHT component. Such correspondence was achieved by the closure of age-specific cells as the target numbers were approached.

Table 2 shows the projected power; that is, the probability of rejecting the null hypothesis, for the

Table 1 Women's Health Initiative sample sizes (% of total) by age group (as of 4/1/00)

Age group	Dietary modification	Postmenopausal hormone therapy		Calcium and vitamin D	Observational study
		Without uterus (E-alone)	With uterus (E + P)		
50–54	6961 (14)	1396 (13)	2029 (12)	5157 (14)	12 386 (13)
55–59	11 043 (23)	1916 (18)	3492 (21)	8265 (23)	17 321 (18)
60–69	22 713 (47)	4852 (45)	7512 (45)	16 520 (46)	41 196 (44)
70–79	8118 (17)	2575 (24)	3575 (22)	6340 (17)	22 773 (24)
Total	48 835	10 739	16 608	36 282	93 676

Table 2 Statistical power for each component for the CT

Outcome	Disease probability (%) (×100)[a]		Intervention effect[b] (%)	Avg. follow-up duration (yrs)	Projected power (%)
	Control	Intervention			
Dietary modification component	–	–	–	–	–
Breast cancer	2.72	2.35	14	8.5	84
Colorectal cancer	1.39	1.12	19	8.5	87
CHD	3.78	3.27	14	8.5	84
Postmenopausal hormone therapy – E-alone	–	–	–	–	–
CHD	4.63	3.67	21	8.5	72
Hip fracture	2.86	2.25	21	8.5	55
Combined fracture[c]	11.02	8.81	20	8.5	97
Breast cancer	4.38	5.36	(22)	13.5	71
Calcium and vitamin D	–	–	–	–	–
Hip fracture	2.23	1.77	21	7.5	88
Combined fracture[c]	8.93	7.23	19	7.5	>99
Colorectal cancer	1.25	1.02	18	7.5	66

[a] Cumulative disease probability to planned termination (×100)
[b] One minus ratio of control to intervention cumulative incidence rates at study termination (×100)
[c] Includes proximal femur, distal forearm, proximal humerus, pelvis, and vertebra

key outcomes for each continuing component of the CT, taking account of the age-specific sample sizes in Table 1. Projected power is given at planned termination in early 2005, in which case the average follow-up duration will be about 8.5 years in the DM and PHT components and about 7.5 years in the CaD component. The intervention effects shown in Table 2 represent the projected effect size after accounting for assumed nonadherence and loss to competing risks. Comparison with projected power calculations at the design stage [11] indicates that a somewhat prolonged recruitment period, and the minor departures from target in sample sizes by age category had rather little effect on projected study power. The CHD and hip fracture power projections for the E-alone versus placebo comparison is somewhat reduced by a smaller than targeted sample size (10 739 versus 12 375) in this CT component. Power calculations for representative comparisons in the OS have been previously given [11].

An independent **Data and Safety Monitoring Board** (DSMB) is charged with monitoring the CT to ensure participant safety, to assess conformity to program goals, and to examine whether there is a need for early stoppage or other modification of any CT component. The DSMB is composed of senior researchers, otherwise not associated with the study, who have expertise in relevant areas of medicine, epidemiology, biostatistics, clinical trials, and ethics. The DSMB meets biannually to review study progress, including its status in the context of emerging external data. The Board provides recommendations to the NHLBI Director (see Figure 1). The DSMB reviewed and approved the protocol

(*see* **Clinical Trials Protocols**) and consent forms (*see* **Ethics of Randomized Trials**) prior to study implementation. They are apprised of any significant changes to protocol.

Throughout the period of study conduct, the DSMB reviews data on recruitment, adherence, retention, and outcomes. The DSMB is the only group given access to treatment arm comparisons outside of the necessary CCC and NHLBI staff. As such, they determine whether the existing data demonstrate either significant or unanticipated risk or unexpectedly strong benefits, in which case early trial termination, or modification, may be recommended (*see* **Benefit/Risk Assessment in Prevention Trials**). A particular complexity in this study, as often exists in prevention studies, is the need to consider effects on multiple disease processes that may differ in direction, timing, and magnitude.

In the WHI, CT monitoring for consideration of early stopping (*see* **Data and Safety Monitoring**) is based on the following principles and procedures:

- Each trial component (DM, Estrogen alone, Estrogen plus Progestin, CaD) is evaluated separately, so that a stopping decision for one will not necessarily impact the continuation of the other three.
- The evaluation of each intervention includes an assessment of the overall intervention effects on health, through the use of a global index. This global index is defined for each woman as time to first incident event. The events to be included were selected on the basis of *a priori* evidence for each intervention, and supplemented with evidence of death from other causes to capture serious unanticipated intervention effects, as shown in Table 3.
- Early stopping for benefit would be considered, if the primary endpoint comparison crossed a 0.05 level O'Brien–Fleming (OBF) boundary (*see* **Data and Safety Monitoring**), and the global index provided supportive evidence defined by crossing the 0.1 level OBF in favor of the intervention. For the DM, a **Bonferroni** correction is used to acknowledge the fact that there are two designated primary endpoints. This correction allows a stopping recommendation to be made if the boundary is crossed for either of the primary endpoints, without exceeding the designated probability (0.05) of falsely rejecting the overall null hypothesis.
- Early stopping for adverse effects uses a two-step procedure with a 0.1 level OBF boundary for primary safety endpoints, a Bonferroni corrected 0.1 level OBF boundary for all other safety endpoints, and a lower boundary of $z = -1.0$ for the global index to signify supportive evidence for overall harm.

As mentioned above, weighted logrank test statistics are used to test the difference between intervention and control event rates for each outcome. These weights were specified to yield efficient test statistics for the primary outcome under CT design assumptions. As such, these tests may not be sensitive to unexpected effects, whether adverse or beneficial, on any of the study outcomes. Consequently, the DSMB also informally examines unweighted logrank statistics, as well as weighted and unweighted tests for various intervals of time since randomization and for selected subgroups of participants (e.g. specific age groups), toward ensuring participant safety. Further detail on CT monitoring methods and their rationale is given in [6].

Table 3 Trial monitoring endpoints for the WHI clinical trial components

	PHT (E-alone and E + P)	DM	CaD
Primary endpoint	CHD	Breast cancer, Colorectal cancer	Hip fractures
Primary safety endpoint	Breast cancer	N/A	N/A
Other endpoints included in the global index	Stroke, Pulmonary embolism, Hip fractures, Colorectal cancer, Endometrial cancer, Death from other causes	CHD, Death from other causes	Colorectal cancer, Breast cancer, Other fractures, Death from other causes

CT monitoring reports, prepared on a semiannual basis throughout trial follow-up, also present data on the adherence to intervention goals, the rates of participation in follow-up and other program activities, and control group incidence rates. These data are used to update power calculations, along the lines of Table 2, to help assess conformity to overall design goals, and to alert the DSMB to emerging problems. Data on selected biomarkers and intermediate outcomes are also assembled, as such data can provide an objective assessment of the extent to which intervention goals are achieved, and can provide insights into processes that can explain intervention effects on disease outcomes.

Biomarkers and Intermediate Outcomes

Beyond testing primary and secondary hypotheses, the CT is designed to support specialized analyses to explain any treatment effects in terms of intermediate outcomes, and both the CT and OS are designed to produce new information on **risk factors** for cardiovascular disease, cancer, and other diseases. To do so, the basic WHI program supports a substantial infrastructure of archival blood product storage, which includes serum and plasma from CT and OS participants at baseline, and at selected follow-up times (one year from enrollment in the CT and three years from enrollment in the OS). In addition, baseline white blood cells (buffy coat) are stored in both the CT and OS. These blood specimens are used for specialized studies related to participant safety and CT intervention adherence, and for externally funded ancillary studies. Stored blood components collected from each CT and OS participant during screening include 7.2 mL serum (in 4×1.8 mL vials), 5.4 mL citrated plasma (in 3×1.8 mL vials), 5.4 mL EDTA plasma (in 3×1.8 mL vials), and two aliquots of buffy coat.

Intermediate outcome data collected in the CT include electrocardiograms (obtained as baseline, 3, 6, and 9 years among all CT women) to ascertain "silent" myocardial infarctions and other cardiac diagnoses, and bilateral mammograms (obtained annually for PHT women and biennially for other CT participants). In addition, all PHT women, 65 years of age and older, have cognitive function assessment, and a 25% sample have functional assessment, at baseline and follow-up. A sample of women in both the CT and OS (all those who are enrolled at any one of three specified clinical centers) have dual x-ray absorptiometry at baseline, and at follow-up years 1 (CT only), 3, 6, and 9, to measure change in bone mass in the hip, spine, and total body; these women also provide urine specimens that are stored for studies of the interventions' effects on bone metabolites.

Analyses to explain CT treatment effects, and CT/OS analyses to elucidate disease risk factors, generally take place in a **case–control** or **case–cohort** fashion, to limit the number of specialized analyte determinations. Extensive self-report questionnaire data at baseline and selected follow-up times are also available for use in these analyses, and can be used to inform the case–control sampling procedure.

Data Management and Computing Infrastructure

The size and scope of the WHI creates a large and rather complex data processing load (*see* **Data Management and Coordination**). Each clinical site has recruited at least 3000 participants creating a local data management load as large as that for many multicenter trial coordinating centers.

The data collected for WHI fall roughly into three categories: self-report, clinical measurements, and outcomes data. Self-reported information includes demographic, medical history, diet, reproductive history, family history, and psychosocial and behavioral factors. For these areas, standardized questionnaires were developed from instruments used in other studies of similar populations. Current use of medications and dietary supplements is captured directly from pill bottles that participating women bring to the clinic (*see* **Compliance Assessment in Clinical Trials**). To capture details of hormone therapy use prior to WHI enrollment, an in-person interview was conducted with each woman at baseline to determine her entire history of postmenopausal hormone use. For additional diet information in the DM trial beyond routine food frequency questionnaires, 4-day food records and 24-hour recall of diet are obtained from a subsample of women. Dietary records are completed by the participant, reviewed and documented by certified clinic staff, and a subsample is sent to the CCC for nutrient coding and analysis. The 24-hour recalls of diet are obtained by telephone contact from the coordinating center and these data were coded using the same methods as for the dietary records.

Clinical measures such as anthropometrics, blood pressure, functional status, and results from gynecologic exams are obtained by certified WHI clinic staff using standardized procedures and data collection forms and key-entered into the local study database. Limited blood specimen analyses are conducted locally and recorded. The remaining blood specimens are sent to a central blood repository where they are housed until the appropriate subsamples are identified and sent to the central laboratory for the selected analyses. Electrocardiogram and bone densitometry data are submitted electronically to respective central reading and coordination facilities.

Information on significant health outcomes is initially obtained by self-report. If the type of event is of particular interest for WHI research, additional documentation is obtained from local health care providers and this information is used by a clinic physician to classify and code the event. Additional details of outcomes definitions and methods appear elsewhere [5].

Data quality assurance mechanisms are incorporated at several levels, in addition to the overall quality assurance program described below (*see* **Clinical Trials Audit and Quality Control**). Data entry screens incorporate range and validity checks, and scanning software rejects forms containing critical errors. Routine audits of randomly selected charts document errors and provide feedback to CC and CCC staff. Additional data quality checks are used in creating analytic data sets. Multiple versions of most forms have been used, so some data items require mapping across versions.

To support the large requirement of local operations as well as central analyses and reporting, the CCC developed and implemented a standardized computing and **database management system** that serves each clinical center site and the coordinating center. This computing system can be logically divided into three major areas: computing at the clinical centers, computing at the CCC, and a private wide area network (WAN). The study-wide database uses this infrastructure to provide the appropriate data management tools to all sites.

Each clinical center is equipped with its own local area network consisting of a file server, ethernet switch, 10 to 20 workstations, two or more printers, a mark sense form reader, bar code readers, and a router. The router provides connectivity back to the CCC over the WAN. In some cases, the router also provides connectivity to the parent institution. The file server is configured with Windows NT Advance Server and runs its own instance of the study's Oracle database. The server also provides standardized office applications (Microsoft Office) and e-mail (Microsoft Exchange Web client). The workstations are Windows 98 clients.

The CCC maintains a cadre of application servers dedicated to the development, testing, and warehousing of the consolidated database, currently requiring 100 GB. The CCC also maintains several other servers dedicated to statistical analysis, administrative support for CCC staff, website and e-mail services for study-wide communication, and centralized automated backup for all study servers. The website and e-mail system dedicated to WHI staff and investigators is critical to managing the challenges of study communications with nearly 1500 WHI staff and investigators spread across 5 timezones. The website provides a kind of electronic glue for bringing together disparate groups. WHI e-mail access is available through the website either over the WAN or through the Internet.

The WHI WAN is a private network, which connects CCs to the CCC using a combination of 56 k and T1 frame-relay circuits. The WAN enables the CCC to conduct nightly backups of clinical center file servers. It also facilitates remote management and troubleshooting of clinical center equipment. In addition, it provides CCs direct access to the WHI e-mail system and website.

The WHI database management system is a distributed replicated database, implemented in Oracle 8.0 for Windows NT. Database design and table structure are identical across CCs but are populated only with data specific to that site. The average clinical center database currently requires approximately 15 GB of space. Data acquisition relies heavily on mark sense scanning, supplemented with traditional key entry and bar code reading. The database supports and enforces the study protocol through its participant eligibility confirmation, randomization, drug dispensing, and collection, visit and task planning, and outcomes processing functions. Security is provided both by password protection and by limiting access to specific data based on the identified role of the user. Local access to clinical site-specific data is supported through centrally defined reports and a flexible data extract system.

The CCC database provides the superstructure into which the CC data are consolidated routinely. Additional data are obtained from the central laboratories and specimen repository and are merged with, and checked against, the corresponding participant data. The central database serves as the source of all data reports and analyses.

Quality Assurance Program Overview

The WHI program involves a complex protocol, with an extensive set of required procedures. The CT intervention goals and the study timeline are demanding program elements. With these challenges, an organized quality assurance (QA) program was needed to identify and correct emerging problems. The QA program is an integral part of the study protocol, procedures, and database, and covers all aspects of WHI. The program seeks to balance the need to assure scientific quality of the study with available resources. The complexity, size, and fiscal responsibility of WHI necessitated establishing priorities to guide local and central QA activities.

The WHI QA priorities were developed by a task force comprising WHI investigators and staff, under the premise that aspects critical to the main scientific objectives of WHI would be of highest priority. As the centerpiece of WHI, the fundamental elements of the CT are considered of highest priority. The next highest priority is given to key elements of the OS and elements of the CT that are important for interpretive analyses. The remaining elements are given a lower priority. The implementation of these priorities is manifested in the frequency and level of detailed QA activities.

QA methods and responsibilities include activities performed at the CCs as well as activities initiated and coordinated by the CCC. The QA Program includes: extensive documentation of procedures; training and certification of staff; routine QA visits conducted by the CCC (all CCs received an initial and an annual QA Visit while subsequent visits are done approximately every other year, or more frequently as needed); and database reports for review by CCs and pertinent committees describing the completeness, timeliness, and reliability of tasks at the CCs. For example **moving average** monthly intervention adherence rates, and major task completeness rates, for each CC are used as one up-to-date indicator of CT status.

WHI has established performance goals for various important tasks that are centrally monitored. These goals were determined on the basis of design assumptions and, where available, on previously published standards of quality and safety.

The performance of each CC is reviewed on a regular basis under a performance monitoring plan. This plan is used to identify clinic-specific performance issues in a timely fashion, to reinforce good performance and to provide assistance or to institute corrective action if performance is inadequate. Much of this work is conducted under the auspices of a performance monitoring committee (PMC), comprising representatives of the CCC, CCs, and PO. The PMC follows up on persistent issues with specific CCs, and conducts site visits to facilitate the resolution of specific areas of concern. Some additional detail on the implementation of the WHI design is given in [2].

Early Results from the WHI Clinical Trial

In late May 2002, after an average follow-up of 5.2 years, the DSMB recommended the early stopping of the E + P trial component because the weighted logrank test for invasive breast cancer exceeded the OBF stopping boundary for this adverse effect and the global index supported risks exceeding benefits.

Participating women in the E + P trial were asked to stop taking their study pills on July 8, 2002 and principal trial results were published soon thereafter [13]. Women in the E + P trial continue to be followed without intervention through 2005, and a plan is under development for the additional nonintervention follow-up of PHT women from 2005 to 2007.

On the basis of data through April 2002, the E + P trial generated **hazard ratio** estimates and nominal 95% **confidence intervals** as follows: coronary heart disease 1.29 (1.02–1.63), breast cancer 1.26 (1.00–1.59), stroke 1.41 (1.07–1.85), pulmonary embolism 2.13 (1.39–3.25), colorectal cancer 0.63 (0.43–0.92), endometrial cancer 0.83 (0.47–1.47), hip fracture 0.66 (0.45–0.98), and death due to other causes 0.92 (0.74–1.14). The global index, defined as the earliest event of these just listed, had a hazard ratio estimate (nominal 95% confidence interval) of 1.15 (1.03–1.28). Absolute **excess risks** per 10 000 **person years** were estimated as seven for coronary

heart disease, eight for stroke, eight for pulmonary embolism, and eight for breast cancer, while corresponding absolute risk reductions were estimates as six for colorectal cancer and five for hip fracture. The absolute excess risk for global index events were estimated as 19 per 10 000 person years. Confidence intervals adjusted for sequential monitoring (*see* **Sequential Methods for Clinical Trials**), and for multiple testing (*see* **Multiplicity in Clinical Trials**) in accordance with the CT monitoring plan are also given in [13].

Even though these risk alterations are fairly modest, they have substantial population implications for morbidity and mortality. As a result of these findings, various professional organizations have altered their recommendations concerning combined hormone use, and labeling changes have been made or are under consideration. These results follow decades of observational studies supporting a cardioprotective benefit for hormone therapy, and the discussion following the reporting of E + P trial results has sharpened the understanding of comparative properties of trials and observational studies among scientific groups and the general population. Additional outcome events through July 7, 2002 have been adjudicated and several more specialized results papers have been published [1, 3, 4, 7, 8, 12]. An ancillary study examining the effects of E + P on dementia and cognitive function has also been published [9, 10].

Summary and Discussion

The WHI, CT, and OS was implemented in close correspondence to design specifications [11]. Departures from design assumptions concerning sample size, age distribution, and projected average trial follow-up have limited effect on the adequacy of primary outcome study power for continuing CT components, with the possible exception of the E-alone versus placebo comparison, where some power reduction for coronary heart disease arises from a smaller than targeted sample size. Substantial infrastructure for specimen storage, routine analyte determination, data management and computing, and for data and protocol quality control was also implemented. Principal results from the trial of combined hormones (E + P) have been presented following the early stopping of intervention.

Ongoing challenges in the CT and OS include retaining the active participation of study subjects over a lengthy follow-up period, ensuring the unbiased and timely ascertainment of outcome events in each CT component and in the OS and, perhaps the most challenging, ensuring an adequate adherence to intervention goals for each continuing CT intervention.

The astrogen-alone component of the WHI clinical trial stopped early on March 1, 2004 with principal results presented in the Journal of the American Medical Association 291; 1701–1712, 2004.

Acknowledgments

This work was supported by NIH contracts for the WHI. Parts of this entry are closely related to a forthcoming monograph chapter [2] by the authors and other WHI colleagues.

References

[1] Anderson, G.L., Judd, H.L., Kaunitz, A.M., Barad, D.H., Beresford, S.A., Pettinger, M., Liu, J., McNeeley, S.G. & Lopez, A.M., for the WHI Investigators. (2003). Effects of estrogen plus progestin on gynecologic cancers and associated diagnostic procedures in the Women's Health Initiative: a randomized trial, *Journal of the American Medical Association* **290**, 1739–1748.

[2] Anderson, G.L., Manson, J., Wallace, R., Lund, B., Hall, D., Davis, S., Shumaker, S., Wang, C.Y., Stein, E. & Prentice, R.L. (2003). Implementation of the Women's Health Initiative study design, *Annals of Epidemiology* **13**, 55–517.

[3] Cauley, J.A., Robbins, J., Chen, Z., Cummings, S.R., Jackson, R., LaCroix, A.Z., LeBoff, M., Lewis, C.E., McGowan, J., Neuner, J., Pettinger, M., Stefanick, M.L., Wactawski-Wende, J. & Watts, N.B. (2003). The Effects of estrogen plus progestin on the risk of fracture and bone mineral density. The Women's Health Initiative clinical trial, *Journal of the American Medical Association* **290**, 1729–1738.

[4] Chlebowski, R.T., Hendrix, S.L., Langer, R.D., Stefanick, M.L., Gass, M., Lane, D., Rodabough, R.J., Gilligan, M.A., Cyr, M.G., Thomson, C.A., Khandekar, J., Petrovitch, H. & McTiernan, A., for the WHI Investigators. (2003). Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women. The Women's Health Initiative randomized trial, *Journal of the American Medical Association* **289**, 3243–3253.

[5] Curb J.D., McTiernan, A., Heckbert, S.R., Kooperberg, C., Stanford, J., Nevitt, M., Johnson, K., Proulx-Burns, L., Pastore, L., Criqui M. & Daugherty, S. (2003). Outcomes ascertainment and adjudication methods in the Women's Health Initiative, *Annals of Epidemiology* **13**, 5122–5128.

[6] Freedman, L.S., Anderson, G.L., Kipnis, V., Prentice, R.L., Wang, C.Y., Rossouw, J.E., Wittes, J. & De Mets, D. (1996). Approaches to monitoring the results of long-term disease prevention trials: examples from the Women's Health Initiative, *Controlled Clinical Trials* **17**, 509–525.

[7] Hays, J., Ockene, J.K., Brunner, R.L., Kotchen, J.M., Manson, J.E., Patterson, R.E., Aragaki, A.K., Shumaker, S.A., Brzyski, R.G., LaCroix, A.Z., Granek, I.A. & Valanis, B.G. for the WHI Investigators. (2003). Effects of estrogen plus progestin on health-related quality of life, *New England Journal of Medicine* **348**, 1839–1854.

[8] Manson, J.E., Hsia, J., Johnson, K.C., Rossouw, J.E., Assaf, A.R., Lasser, N.L., Trevisan, M., Black, H.R., Heckbert, S.R., Detrano, R., Strickland, O.L., Wong, N.D., Crouse, J.R., Stein, E. & Cushman, M., for the WHI Investigators. (2003). Estrogen plus progestin and the risk of coronary heart disease, *New England Journal of Medicine* **349**, 523–534.

[9] Rapp, S.R., Espeland, M.A., Shumaker, S.A., Henderson, V.W., Brunner, R.L., Manson, J.E., Gass, M.L., Stefanick, M.L., Lane, D.S., Hays, J., Johnson, K.C., Coker, L.H., Daily, M. & Bowen, D., for the WHIMS Investigators. (2003). Effect of estrogen plus progestin on global cognitive function in postmenopausal women. The Women's Health Initiative memory study: a randomized controlled trial, *Journal of the American Medical Association* **289**, 2663–2672.

[10] Shumaker, S.A., Legault, C., Rapp, S.R., Thal, L., Wallace, R.B., Ockene, J.K., Hendrix, S.L., Jones, III, B.N., Assaf, A.R., Jackson, R.D., Kotchen, J.M., Wassertheil-Smoller, S. & Wactawski-Wende, J., for the WHIMS Investigators. (2003). Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women. The Women's Health Initiative memory study: a randomized controlled trial, *Journal of the American Medical Association* **289**, 2651–2662.

[11] The Women's Health Initiative Study Group. (1998). Design of the women's health initiative clinical trial and observational study, *Controlled Clinical Trials* **19**, 61–109.

[12] Wassertheil-Smoller, S., Hendrix, S.L., Limacher, M., Heiss, G., Kooperberg, C., Baird, A., Kotchen, T., Curb, J.D., Black, H., Rossouw, J.E., Aragaki, A., Safford, M., Stein, E., Laowattana, S. & Mysiw, W.J., for the WHI Investigators. (2003). Effect of estrogen plus progestin on stroke in postmenopausal women. The Women's Health Initiative: a randomized trial, *Journal of the American Medical Association* **289**, 2673–2684.

[13] Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the Women's Health Initiative randomized controlled trial, *Journal of the American Medical Association* **288**, 321–333.

ROSS L. PRENTICE & GARNET L. ANDERSON

Worcester, Jane

Born: December 5, 1910.
Died: October 8, 1989, in Falmouth, Mass.

Jane Worcester was the first female Department Chair of Biostatistics at Harvard School of Public Health. Her career in the department spanned from 1931, when she came to the department as a mathematical computing assistant after receiving her A.B. from Smith College, to 1977, when she retired as Department Chair and Professor of Biostatistics and Epidemiology. Dr Worcester liked to relate to her students the story of her first position in the department as "Computer". Dr Worcester received a Dr. P.H. from Harvard in 1947 and an honorary Sc.D. from Smith in 1968. She was one of the earliest non-Radcliffe women to become a full professor at Harvard at a time when there were no women in the Faculty of Arts and Sciences, the Law School, or the Medical School. She served as an important role model for many women in academics.

During the 46 years that Dr Worcester spent with the department, she devoted her time to research, teaching, and service both to the School and to the discipline of biostatistics. She was the center of gravity of the department for a very long time, during which both faculty and students sought her advice. Dr Worcester stimulated her students to achieve more than they thought they could accomplish and encouraged them to grow (*see* **Teaching Medical Statistics to Statisticians**). In describing Dr Worcester's influence on the students in the department, Ray Neff, Sc.D. '77, considered her "an archetypal mentor". Charles Ralph Buncher, Sc.D. '67, remembered the story of when Dr Worcester returned from the oral qualifying examination of Joseph Brain, now Professor and Chair of the Department of Environmental Health at the School. She exclaimed, "What a joy it was because I learned so much!"

Dr Worcester provided a consulting role for research in a range of areas in public health. She worked very closely with the editors and editorial board of the *New England Journal of Medicine* and taught her students how to review the medical literature (*see* **Statistical Review for Medical Journals**). She also worked with laboratory scientists and the Department of Nutrition at the School and viewed the involvement as a tool to provide students with an active, practical, learning environment and also to offer help to other scientists. Jane Menken, Ph.D., described Dr Worcester as "someone whose intense dedication to the School and scientific inquiry were obvious to all who knew her, and those who associated with her were indeed very fortunate".

NAN M. LAIRD

World Health Organization (WHO): Biostatistics and Epidemiology

The World Health Organization (WHO) was established in April 1948 as a specialized agency of the United Nations, taking over the functions of the Office International d'Hygiène Publique, the Health Organization of the League of Nations, and the Health Division of the United Nations Relief and Rehabilitation Administration (UNRRA). The practical tasks of WHO were seen as supplying technical aid in combating epidemics such as cholera, malaria, smallpox, tuberculosis, etc., and particularly as assisting less developed and financially weak countries to improve their medical and health services.

In the general area of health statistics and epidemiological **surveillance**, WHO's responsibilities included the establishment and maintenance of administrative and technical services, including epidemiologic and statistical services; the provision of information, counseling and assistance in the field of health; and the founding, with revision where necessary, of an international nomenclature of diseases, causes of death and public health practices (*see* **International Classification of Diseases (ICD)**). Member States had corresponding obligations to communicate to WHO important laws, regulations, official reports, and statistics relating to the health field, and especially to provide appropriate statistical and epidemiologic reports. For a more detailed historical account of developments over the period 1948–1988, see the excellent review by Uemura [8].

In addition to WHO's headquarters (WHO/HQ) in Geneva, there are separate Regional Offices in the six regions of Africa, the Americas, Eastern Mediterranean, Europe, South-East Asia and the Western Pacific, besides several other special offices. Like many other international organizations, WHO's staff and responsibilities are widely, but thinly, spread and increasingly insufficiently funded. Despite these difficulties, much excellent work is carried out in maintaining a high standard of public health review through the mechanism of Expert Committees, promotion of collaboration with country projects, development of advisory services and training programs, etc.

Essential statistical activities were developed and formulated by the WHO Expert Committee on Health Statistics, and in 1949 a Health Statistics Division was created in WHO/HQ. The direct use of biostatistical methods expanded only gradually until in 1967 a special effort was made to establish at HQ a Division of Research in Epidemiology and Communication Science (RECS). This Division incorporated individual units covering **communicable diseases**, noncommunicable diseases, social systems, environmental aspects, **operational research**, and computer science – following the installation of a central computer facility in 1966.

In principle, this development should have greatly increased WHO's stature and capabilities over the whole area of applied biostatistics. Unfortunately, the expected collaboration between RECS and the long-standing Units and Divisions of WHO/HQ never took place. A great deal of first class research work was carried out and duly published, but as the initial expectations were not realized, RECS was disbanded in 1972 after five years of serious biostatistical efforts.

Most of the RECS technical staff were then transferred to other areas of WHO/HQ, and the Division of Health Statistics was itself greatly strengthened in this way. In particular, one can cite the development of dynamic modeling applied to several communicable diseases including typhoid fever, cholera, tetanus, diphtheria, whooping cough, and tuberculosis. Early on there was the epidemiologic modeling of the spread and control of tuberculosis by Waaler & Piot [9] in the Tuberculosis Unit which involved some discussions with RECS. Later on it became clear that one of the most significant achievements of RECS was the construction of a malaria model closely tied to field investigations in Africa (see Dietz et al. [4]). Also to be mentioned are the books by Bailey on the mathematical theory of infectious diseases [1] and the biomathematics of malaria [2]. Many other studies, such as the MONICA project on monitoring trends and determinants in cardiovascular disease [10], carried out in a large number of countries, were also promoted and strengthened by the spread of ideas and techniques arising from the RECS period. For further discussion of these matters see Uemura [8], as already mentioned above.

On the other hand, there is the failure of WHO's communicable disease statistics, dealing largely with limited country-by-country data, to elucidate the worldwide process of spatial diffusion. Much important work of direct practical relevance has already been done by geographers and statisticians. See, for example, Cliff et al. [3] on the spread of measles; Gould [5], and Gould & Wallace [6], on the spatial diffusion of HIV and **AIDS**; as well as a whole recent issue of *Statistical Methods in Medical Research* [7] entitled "Spatial Epidemiology" (*see* **Epidemic Models, Spatial**).

Now, what is the role of biostatistics in the whole WHO program? Most technical biostatistical publications are highly mathematical, and are largely not understood, even in principle, by those who might use the methods and results in practical applications to medicine, epidemiology, and public health. Even when the research papers seem to be applied they are often using simplified real data only for illustrative purposes. The real-life problems faced by administrators and decision-makers are therefore mostly neglected. A fully effective application of biostatistics in the whole public health field requires a major collaborative effort of an operational research character.

Finally, let us consider what should be done. Well-organized multidisciplinary teams to cover the epidemiologic modeling of both communicable and noncommunicable diseases, involving relatively small groups of individuals with overlapping skills, should be established so as to include a spectrum from purely scientific expertise to those engaged in real-life decision-making. This implies operational research on a global level.

The world clearly needs all the applied technical skills it can get to control, and eradicate where possible, a very wide range of life-threatening diseases – some well known, others newly emerging.

References

[1] Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases*. Griffin, London.

[2] Bailey, N.T.J. (1982). *The Biomathematics of Malaria*. Griffin, London.

[3] Cliff, A., Haggett, P. & Smallman-Raynor, M. (1993). *Measles: an Historical Geography of a Major Human Viral Disease*. Blackwell, Oxford.

[4] Dietz, K., Molineaux, L. & Thomas, A. (1974). A malaria model tested in the African savannah, *Bulletin of the World Health Organization* **50**, 347–357.

[5] Gould, P. (1993). *The Slow Plague: a Geography of the AIDS Pandemic*. Blackwell, Oxford.

[6] Gould, P. & Wallace, R. (1994). Spatial structures and scientific paradoxes in the AIDS pandemic, *Geografiska Annalen* **76B**, 105–116.

[7] Statistical Methods in Medical Research (1995). *Spatial Epidemiology* **4**,.

[8] Uemura, K. (1988). World health situation and trend assessment from 1948–1988, *Bulletin of the World Health Organization* **66**, 679–687.

[9] Waaler, H.T. & Piot, M.A. (1969). The use of an epidemiological model for estimating the effectiveness of tuberculosis control measures, *Bulletin of the World Health Organization* **41**, 75–93.

[10] WHO (1988). The World Health Organization MONICA project (monitoring trends and determinants in cardiovascular disease): a major international collaboration, *Journal of Clinical Epidemiology* **41**, 105–114.

NORMAN T.J. BAILEY

World Health Organization (WHO): Global Health Situation

To recognize the development of epidemiologic and statistical activities and trend assessment in the World Health Organization (WHO), it is first important to review the contribution of epidemiology to world health. Epidemiology originated in response to a need to understand and control the highly infectious epidemic diseases, such as cholera, plague, smallpox, and yellow fever. It was only with time that appreciation grew of the fact that all conditions of disease and ill-health are interrelated, and that the emerging science of epidemiology provided the tools for helping to understand the major factors underlying these issues as well [1].

International health work also began with a concentration on infectious diseases (*see* **Communicable Diseases**), and then moved towards a wider concept of health as part of overall development. The roots of the World Health Organization, as an international health agency, go back to efforts in the last century, and early in this century, particularly to the Rome Agreement of 1907, which established the Office International d'Hygiène Publique, with the express purpose "to combat infectious diseases". The progressive shift of the concept of health, from the prevention of infectious diseases to viewing health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity", is reflected in the successive evolution of the Pan American Sanitary Bureau, founded in 1902, the health services arm of the League of Nations, founded in 1918, and finally WHO, founded in 1948 [1].

Epidemiology has provided the tools for a better understanding of the incidence, **prevalence, natural history**, causes (*see* **Causation**), and effects of control and other measures that are relevant to each of the communicable disease control programs of WHO. More than this, the epidemiologic sciences have enabled us also to understand noncommunicable diseases such as cancer (*see* **Oncology**), cardiovascular diseases, (*see* **Cardiology and Cardiovascular Disease**), and genetic disorders (*see* **Genetic Epidemiology**). In the area of primary prevention this understanding has allowed for intervention before the onset of disease [1].

WHO has a constitutional responsibility for the global epidemiologic **surveillance of disease**. It receives information on outbreaks of communicable disease and distributes this information throughout the world by telecommunication and its publication *Weekly Epidemiological Record*. The WHO system of international epidemiologic surveillance provides countries with access to information. This includes the countries that otherwise would not be able to communicate directly with each other. WHO is responsible for the *International Health Regulations*, the **International Classification of Diseases**, and a great many international standards, which together make international epidemiologic comparisons possible. WHO also publishes the *World Health Statistics Quarterly*, the *World Health Statistics Annual*, the *World Health Report*, and other publications of an epidemiologic nature [1].

Twenty years ago, the Thirtieth World Health Assembly decided that the main social target of governments and WHO in the coming decades should be health for all by the year 2000. One year later, in 1978, at a major international conference at Alma-Ata, primary health care was declared to be the key to attaining this goal, in the spirit of social justice. The policies and strategies for "health for all" have subsequently been defined by the World Health Assembly at an international level, in the light of its own health and socioeconomic situation. Furthermore, WHO Member States have committed themselves, with the Organization, to monitoring progress towards and evaluating the attainment of this common goal, using a basic set of global indicators in addition to those applicable within each country. For the first time in the history of international health work, an epidemiologic framework is being applied on a global scale. The implication is that the science of epidemiology must be applied for strategic health planning and evaluation, in a systematic manner, in practically all countries of the world, for national and international health development purposes [1].

The health-for-all monitoring and evaluation process is intended to establish a baseline of current health and socioeconomic conditions, against which progress towards defined targets and objectives can be measured. Periodic measurement should establish trends that will permit anticipation of future conditions, and to start planning for them in advance. The three cycles of monitoring and evaluation of the health-for-all strategy that have taken place so far

make WHO optimistic that the information obtained can be used to reorient national and international priorities and directions for health development work, on the basis of sound epidemiologic evidence, reported by countries with honesty [1].

The latest evaluation was carried out late 1996/early 1997; national reports will be consolidated into regional reports to be reviewed by WHO regional committees in September–October 1997; the global findings will be reported to WHO Governing Bodies in 1998.

At the present time epidemiology and evaluation are used substantially to support health future trend assessment through scenario planning at the global, regional, and country level. A new reorientation of health statistics is also taking place in response to the trend for broader and higher-quality data requested by users. In WHO, approaching the end of the twentieth century, epidemiology, statistics, and future trend assessment provide a substantive contribution to the formulation of the health-for-all policy and strategy for the next century.

Program Activities to the End of the Twentieth Century

The Health Situation and Trend Assessment Program is at present responsible for global health situation analysis and projection; strengthening of country health information; and partnerships and coordination of epidemiology, statistics, and trend assessment. Global epidemiologic surveillance is under the responsibility of the Emerging and other Communicable Diseases Surveillance and Control Program. Various epidemiologic and statistical activities are also carried out by technical programs at WHO headquarters, e.g. Expanded Program on Immunization; Health and Environment; Information System Management; Special Program of Research, Development, and Research Training in Human Reproduction; Special Program for Research and Training in Tropical Diseases; and also its six regional offices.

As already mentioned, one of the normative functions of WHO is monitoring the health situation and trends throughout the world, for which the compilation, use, and coordination of relevant health information and statistical activities are essential. WHO evaluates the world health status and trends every 6 years and publishes its findings. It also assesses the global health status annually and, since 1995, has published *The World Health Report*. The report gives an overview of the global situation and identifies priority areas for international health action; it also links the work of WHO to global health needs and priorities (*see* **Morbidity and Mortality, Changing Patterns in the Twentieth Century**; **Mortality, International Comparisons**).

In headquarters this Program is carried out by the Division of Health Situation and Trend Assessment, which is comprised of the Units of Health Situation Analysis and Projection and Strengthening Country Health Information, and the Director's Office responsible for partnerships and coordination in epidemiology, statistics, and trend assessment and the *International Statistical Classification of Diseases and Related Health Problems* (ICD) besides providing direction and supervision of the Program.

The six regional offices of WHO also undertake health situation and trend assessment activities. From the Program review in 1995 the focus of the regional programs is as follows.

African Region

In the African Region priority is given to generation and utilization of epidemiologic and health information, particularly through the strengthening of national health information systems (*see* **Administrative Databases**) and epidemiologic surveillance systems. The development of epidemiology practice is based on the following approaches: integrated management of epidemiologic information systems; strengthening both data management capacity and decision-making processes, particularly at local level; combined training in epidemiology and management of health programs and health systems. Efforts are made to equip all district teams with a set of essential epidemiologic capabilities. WHO established a position of epidemiologist in each Country Office in order to support this effort.

Region of the Americas

The major achievements of the Region of the Americas include coordinating the response to the cholera epidemic, supporting the evaluation and strengthening of national health surveillance systems, support for modernization of records and statistical information systems, implementation and application of

the ICD-10, Geographic–Epidemiologic Information Systems (*see* **Geographic Patterns of Disease**; **Geographic Epidemiology**), supporting national studies of social inequities that affect health status, producing regional publications of the health situation (*Health Conditions in the Americas, Strategies to Monitor Health for All (HFA), PAHO Epidemiological Bulletin*, and *Health Statistics from the Americas*), and establishing a technical information system including country profiles, databases on mortality and population, and a bibliography on epidemiology. Efforts will continue to strengthen national epidemiologic capacities, to study inequity in health, training in epidemiology, statistics, and health situation analysis and systems, and to disseminate information (*see* **Health Services Data Sources in Canada**; **Health Services Data Sources in the US**).

Eastern Mediterranean Region

The Eastern Mediterranean Region has cooperated and will continue to cooperate in the development of health information systems in Member States, by providing technical support, strengthening national capabilities through fellowships and workshops, developing a regional database and publishing guidelines and manuals to improve health information management design, implementation, and use in the decision making process.

European Region

In the European Region the main tasks are collecting and analyzing health information for periodic reports of progress towards health for all by the year 2000 (entitled "Health in Europe"), regularly updating and disseminating information from the health-for-all database to Member States (available also from `http://www.who.dk`), and supporting training in epidemiology and health information. A European Public Health Information Network (EUPHIN) is being established (together with the EC) to enable telematic reporting and exchange of data for international comparisons. As part of this initiative, efforts continue to improve international data comparability by developing and encouraging countries to use standard definitions, measurement instruments, and methods.

At the country level, the European Regional Office assists countries in developing and using national health and health service databases like the health-for-all database, as a means for making better use of available health data at national and local level. The European Regional Office also produces country "highlights" which give an overview of the health and health-related situation in a given country and compare, where possible, its position in relation to other countries in the WHO European Region (also available from `http://www.who.dk`).

The intention in all cases is to make better use of available health information, which will in itself help to improve data quality and comparability, and enable national and local agencies and institutions with health responsibilities to have easy access and benefit from the activities and product of the European Regional Office (*see* **Health Services Data Sources in Europe**).

Southeast Asia Region

In the Southeast Asia Region support to countries has focused on the strengthening of health management information systems at the central and district level, and the enhancement of mortality and morbidity statistics (*see* **Vital Statistics, Overview**). Attention will continue to focus on developing health management information systems in interested countries. Epidemiologic surveillance, the use of "health futures" methodology, and health data processing and rational use is being further enhanced.

Western Pacific Region

The Western Pacific Region is engaged in technical cooperation to strengthen epidemiologic surveillance and cholera control, support field epidemiology training, improve medical records, birth registration documentation and health information systems, and provide guidelines and manuals. Future activities include reformulation and construction of a new regional database, increased coordination with technical units, strengthening of current activities, and research on more sensitive indicators and associated analytic methods for monitoring and evaluation.

Main Activities at Global and Regional Level

The Program's main activities at the global and regional level from around 1997 to 1999 can be summarized as follows:

1. *Global health situation analysis and projection*. Updating databases on mortality, health-for-all data resulting from the third global evaluation and global health futures information; reporting on the third global evaluation of the implementation of the health-for-all strategy (to be published in the *World Health Report 1998*); improving and formulating new indicators and updating the common framework for the fourth monitoring of the implementation of the health-for-all strategy; issuing the *World Health Statistics Annual 1997 and 1998*, and the *World Health Statistics Quarterly*, volumes 51 and 52; updating the documents "Global health situation analysis and projection" and "Demographic data for health situation and projections"; and validating global health and health-related data and information. At the regional level preparing and distributing regional reports on the world health situation; maintaining and updating databases of health statistics; and improving the health-for-all strategy through the findings of the third evaluation of the implementation of the Strategy.
2. *Strengthening country health information*. Contributing to defining universal public health functions; providing guidance on assessing performance of public health functions and defining the minimum information required for their monitoring and management; preparing methodology for rapidly assessing availability and use of information for managing the essential public health functions; preparing guidance materials and supporting processes for enhancing various information functions and support systems; and undertaking a series of applications of the above methods within interested countries. At the regional level providing advisory services and support to countries; enhancing national information systems and assisting countries in applying various methodologies such as health futures and rapid evaluation.
3. *Partnerships and coordination in epidemiology, statistics, and trend assessment*. Forming partnerships and coordination in epidemiology, statistics and trend assessment with related WHO programs, regional offices and international organizations; preparing an international agreement on a taxonomic approach for medical procedures and guidelines for establishing national classifications; providing guidance on medical certification of causes of death. At the regional level improving partnerships in epidemiology, statistics, and trend assessment, and holding computer-based training courses in the use of ICD-10.

It is important to note that the Division of Health Situation and Trend Assessment is responsible for the development, maintenance, and coordination of the *International Statistical Classification of Diseases and Related Health Problems* (ICD) and other members of the "family" of disease and health-related classifications in both English and French.

The future prospect of ICD can be highlighted as follows. More than 60 Member States submit national mortality data to WHO on a routine basis. Twenty-eight of these Member States have already implemented the tenth revision of the ICD (ICD-10), which was published by WHO in 1992–93, for either mortality or morbidity coding or both. Twenty-two Member States in the Region for the Americas have received training in the use of ICD-10 but have not yet implemented it. A further nine Member States have indicated that they plan to implement it before the year 2000. It can be assumed, therefore, that by the year 2000 the vast majority of the countries currently submitting mortality data will have moved to ICD-10, though the actual data may not be available until some 2–3 years later. It is hoped that the planned publication of a simplified three-character version of ICD-10 will encourage more developing countries to use the classification. In general, it is estimated that, approaching the twenty-first century, most developed countries will have implemented ICD-10. On the other hand, in developing countries the classification will be implemented step by step in line with the speed of the strengthening of the capability of their vital and health statistical infrastructures.

At the global level the Division of Emerging and other Communicable Diseases Surveillance and Control publishes the *Weekly Epidemiological Record*; updates annually *International Travel and Health*; revises the International Health Regulations; and rapidly exchanges information through electronic media with WHO Collaborating Centers, public health administrations and the general public.

At the present time the Health Situation and Trend Assessment Program is implemented with support from various professional staff, especially

epidemiologists, statisticians, medical officers, and public health specialists.

Vision for the Use and Generation of Data in the First Quarter of the Twenty-first Century

Epidemiologic and health statistical activities in WHO are now operational through dynamic networking with various parties at the global, regional, and country level. From experience at the global and regional level and in many countries, the trend of information required by users is towards more specific and higher-quality information. Figure 1 highlights the information on health status and determinants required by most users.
The generation of data and information can usually be grouped into four main activities:

1. Collection, validation, analyses, and dissemination of data and information.
2. Support activities and resources, especially through cooperation with Member States in strengthening country health information.
3. Research and development activities.
4. Partnerships and coordination of information activities.

Figure 2 shows these activities and their interconnection. From experience this diagram is helpful in visualizing the complex issues of information activities.

On the basis of experience and future prospects the trends of the use and generation of information are briefly described in Table 1. The trends here shown take into consideration the main change in the use of data and the various activities involved in the generation of data. It is hoped that the main trends from 1975 to the end of the century are clearly highlighted in the table.

From the WHO Global Health Situation Analysis and Projection from 1950 to 2025 it is possible to foresee the future prospects of health situation and trend assessment in the early part of the twenty-first century, as summarized below.

1. The various users of information will request many types of information or data, all of which should be up-to-date and of high quality. This kind of request will not only come from the user in developed countries but also from many users in developing countries.
2. On the generation of information:
   (i) In the collection, validation, analysis and dissemination of data and information the emphasis will be on data concerning health status (*see* **Quality of Life and Health Status**) (health, **life expectancy**, mortality, disability, morbidity), health resources (including financial) (*see* **Health Care Financing**), physical and manpower resources (*see* **Health Workforce Modeling**), health services,

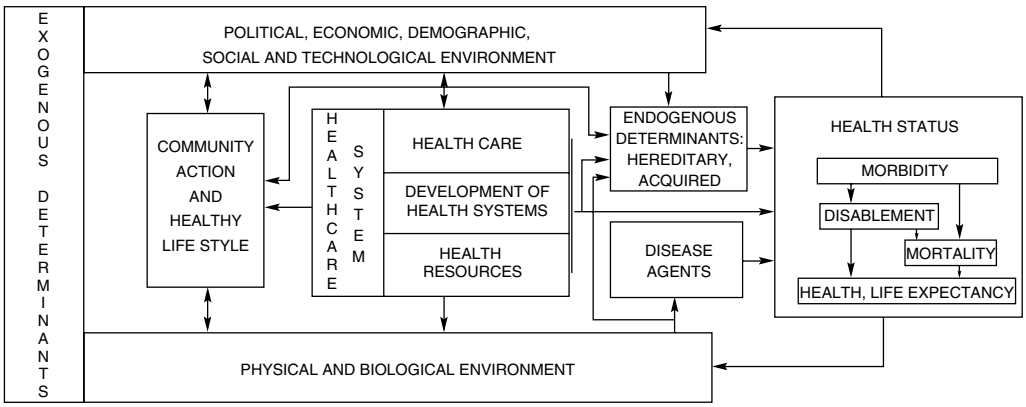

Figure 1 Health status and determinants. Adapted from *Sistem Kesehatan Nasional [National Health System]*, Annex 1. Jakarta, Ministry of Health, 1984 and *Public Health Status and Forecast*, Figure 1.3.3, The Hague, National Institute of Public Health and Environmental Protection, 1994

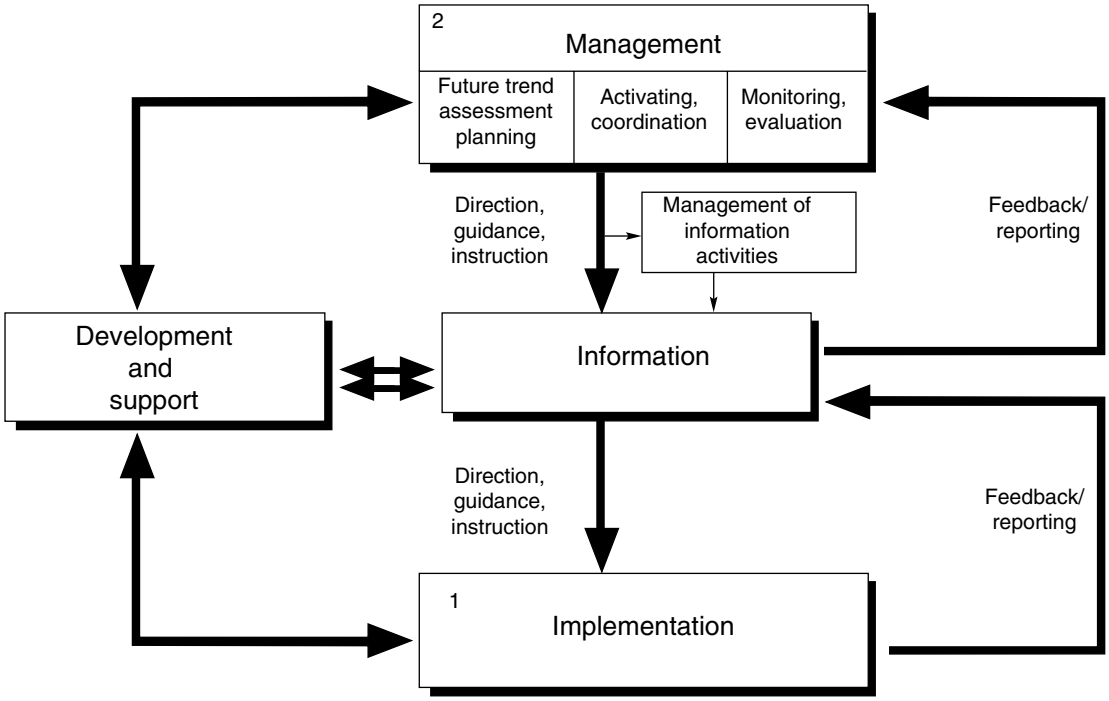

Figure 2 Relationship between information and implementation, management and development activities

Table 1 Brief review of development of main health situation and trend assessment activities at country, regional, and global level

Main activities	1975–87	1988–95	1996–2001
1. *The use of information*	For planning, management and evaluation Limited quality of data could be tolerated	Limited use of information for past and present trends Limited quality of some data could be tolerated	Broader use, including future trends, of information and data High quality of data is needed
2. *The generation of information*			
(i) Collect, validate, analyze, and disseminate data and information	Data on health status, utilization of services, resources, and demographic, social and environmental data	Emphasis on data on monitoring and evaluation of Health for All Partial support of informatics	Emphasis on data on mortality, morbidity, disability Continue to give emphasis to data on monitoring and evaluation of Health for All Full support of informatics
(ii) Research and development activities	Developments of indicators; methods of presentation and dissemination; operational research methods	Development of monitoring and evaluation	Further development of health indicators

Table 1 (*continued*)

Main activities		1975–87	1988–95	1996–2001
		Statistical support to health research	Started health futures trend assessment	Health futures trend assessment
		ICD	ICD, ICIDH	Further development of health futures methods
				Implementation of ICD
(iii)	Support activities and resources	Support given country by country	Strengthening country health information – started cooperation with Member States, especially developing countries and countries in transition, through regional offices	More on strengthening country health information through regional offices
				Support health futures trend assessment
				Support the monitoring and evaluation of HFA
				Support the enhancement of statistical capabilities
				Support the use of computers
(iv)	Partnerships and coordination of information activities	Improve the uniformity and comparability of data		Dynamic networking of health information and statistical activities
		Balancing between decentralized and centralized systems		Coordination of various information, statistical activities, and trend assessment
		Cooperation between statisticians and users	Started through dynamic networking of health information and statistical activities	

Source: Division of Health Situation and Trend Assessment, WHO.

socioeconomic environment (especially economic, sociocultural, and technological environment), and biological and physical environments (including pollution) (*see* **Environmental Epidemiology**). Full informatic support in these information activities will become a reality.

(ii) The future focus on research and development will be on health measurement of the above-mentioned data and various health classifications in a more elaborated way. Besides this, further development of future trend assessment in support of scenario planning will be needed by many countries.

(iii) The trend in support activities and resources will continue to focus on providing support to many developing countries, especially to the least developed countries in strengthening their country health information. The specific support activities and resources as mentioned in Table 1 will continue throughout the beginning of the twenty-first century.

(iv) Partnerships, coordination and provision of direction of information activities will become WHO's most important and challenging tasks in health situation and trend assessment in the future.

Reference

[1] Nakajima, H. (1991). Epidemiology and the future of world health – The Robert Cruickshank Lecture, *International Journal of Epidemiology* **20**, 589–594.

Bibliography

WHO (1992). *Global Health Situation and Projections, Estimates*, WHO/HST/92.1. WHO, Geneva (updated 1997).

WHO (1993–96). *Implementation of the Global Strategy for Health for All by the Year 2000- – Second Evaluation: Eighth Report on the World Health Situation*. Vol. 1, *Global Review* (WHO, Geneva, 1993); Vol. 2, *African Region* (WHO, Brazzaville, 1994); Vol. 3, *Region of the Americas* (WHO, Washington, 1993); Vol. 4, *South-East Asia Region* (WHO, New Delhi, 1993); Vol. 5, *European Region* (WHO, Copenhagen, 1993); Vol. 6, *Eastern Mediterranean Region* (WHO, Alexandria, 1996); Vol. 7, *Western Pacific Region* (WHO, Manila, 1993).

WHO (1994). *Ninth General Programme of Work, Covering the Period 1996–2000*. WHO, Geneva.

WHO (1995). *Programme Budget for the Financial Period 1996–1997*. WHO, Geneva.

WHO (1995). Progress towards health for all: Third monitoring report, *World Health Statistics Quarterly* **48**(3/4).

WHO (1995). *Renewing the Health-for-All Strategy: Elaboration of Policy for Equity, Solidarity and Health*, Consultation document, WHO/PAC/95.1. WHO, Geneva.

WHO (1996). *Catalogue of Health Indicators*, WHO/HST/SCI/96.8. WHO, Geneva.

WHO (1996). *Programme Budget for the Financial Period 1998–1999*. WHO, Geneva.

Weekly Epidemiological Record. WHO, Geneva.

WHO (1995). *World Health Report*. WHO, Geneva.

WHO (1996). *World Health Report*. WHO, Geneva.

World Health Statistics Annual. WHO, Geneva.

World Health Statistics Quarterly. WHO, Geneva.

H.R. HAPSARA

X-Linkage

In humans, there are 22 pairs of homologous chromosomes called *autosomes*, and an additional pair of *sex chromosomes*, denoted by X and Y. Normal females are XX and normal males are XY. The Y chromosome is small and lacks most of the loci found on the X chromosome; loci found only on the X chromosome are called *X-linked loci*. Regions of homology between the X and Y chromosomes are called the *pseudo-autosomal regions* (see below).

The phenomenon of X-linked inheritance has been recognized since ancient times in the case of hemophilia. Jewish law banned the circumcision of male offspring of females with a family history of bleeding; however, this ban did not apply to the offspring of males who were members of families with a history of bleeding. Hemophilia became infamous in the nineteenth and early twentieth centuries with the spread of hemophilia through the royal houses of Europe. Queen Victoria of Great Britain was a carrier of the disease gene, which she passed on to her son Prince Leopold as well as to numerous other members of the royal family. The most famous victim was her great-grandson Alexis Romanov who met his demise not from hemophilia but when he was executed at the age of 14 with his family at Ekaterinburg in 1918. The causes of X-linked recessive hemophilia A and hemophilia B are now known to be due to **mutations** in the factor VIII **gene**, which maps to Xq28 and in the factor IX gene mapping to Xq27.1–q27.2, respectively.

In 1911, color blindness was localized to the X chromosome [27]. This was the first trait to be mapped to the X chromosome in mammals. To date a large number of genes for X-linked traits have been mapped and isolated. They include the genes for retinitis pigmentosa 2 [26], retinitis pigmentosa 3 [25], and X-linked Alport syndrome [1]. These are a few examples of the over 1000 X-linked traits and genes described in Online Mendelian Inheritance in Man [22].

X-linked Transmission and Modes of Inheritance

For X-linked diseases, affected males pass their only X chromosome to all their female offspring, therefore 100% of their female children carry a chromosome with the disease gene. Since males can only pass their Y chromosome to their male offspring, there is no male-to-male transmission for X-linked traits. A **heterozygous** female has two copies of the X chromosome and therefore passes the X chromosome carrying the disease gene to 50% of her children, regardless of their sex. For many X-linked recessive traits (*see* **Genotype**), females are unaffected carriers of the trait, while males express the disease phenotype; a female is affected only in the rare circumstance of receiving a copy of the mutated gene from both her mother and father. As an example, X-linked color blindness has a frequency of 8% in western European males but also affects about 0.6 % of females. In X-linked dominant inheritance, both males and females who carry only one copy of the disease allele are affected.

For some X-linked recessive traits, female carriers display a milder phenotype than affected males. For example, for sensorineural deafness caused by the DFN4 gene, affected males display congenital bilateral profound sensorineural hearing impairment affecting all frequencies, while female carriers manifest a much milder phenotype of bilateral mild to moderate high frequency sensorineural hearing impairment with onset during adulthood [17]. In the case of retinitis pigmentosa 3, carrier females display tapetal-like retinal reflex (a brilliant, scintillating, golden-hued, patchy appearance most striking around the macula), but no visual defect [10].

Ascertainment of Families

From which individuals should DNA samples be collected, when ascertaining a kindred segregating an X-linked recessive trait? For the example in Figure 1 (the “a” indicates the family member should be genotyped), all females are phenotypically normal and it is only possible to determine whether they are carriers if they have an affected son. For example, it is not possible to determine the carrier status of II.4 or II.5, since neither female has a male offspring. Therefore, they do not provide **linkage** information. In addition, the affected male II.10 cannot pass the disease gene to his sons (who need not be ascertained), but his daughter III.10 has an affected son and therefore is a carrier; III.10 and her son IV.1 should be ascertained.

Even if DNA cannot be collected for an individual, in certain circumstances they should still be retained

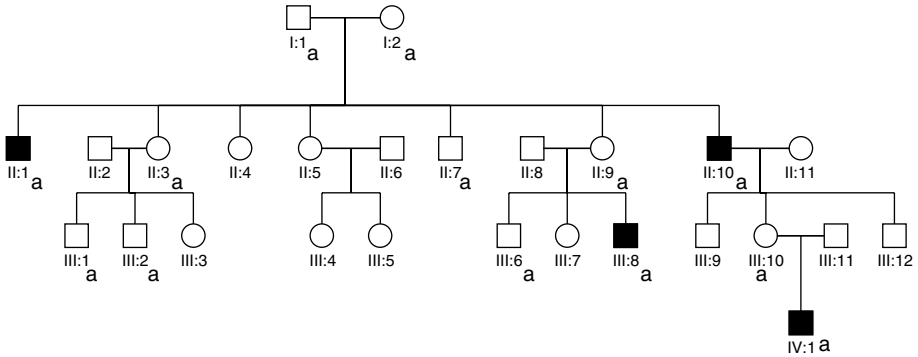

Figure 1 Pedigree 1: ascertainment scheme for an X-linked recessive trait. Individuals marked with an "a" subscript are appropriate for ascertainment for a linkage study

in the pedigree for linkage analysis. For example, individuals III.1 and III.2 should be retained in the analysis even if they lack genotypic information, since the number of unaffected sons that II.3 has changes the probability of whether or not she is a mutant allele carrier, with the probability of her being a carrier decreasing with increasing number of unaffected sons. In addition, affected male offspring III.8 and IV.1 should be retained in the analysis even if a DNA sample cannot be obtained; including them in the analysis classifies the carrier status of their mothers II.9 and III.10.

If it is possible to identify whether or not the females are carriers (or the trait is X-linked dominant and females with one copy of a mutated allele are affected), then ascertainment should be extended to include additional females whose affection or carrier status is known.

Linkage Analysis for X-linked Loci

For **complex diseases**, it may be difficult to determine whether one or more susceptibility loci are X-linked; in this case, the whole genome, including the X chromosome, should be scanned. In general, the phenomenon of X-linkage can be distinguished from other sex-related phenomena, such as **parental effects** and sex-dependent **penetrances**. **Segregation analyses** may be carried out to determine whether or not an X-linked model of inheritance best fits a set of pedigree data.

If there is strong evidence that a disease or trait segregating in a pedigree is X-linked, then genotyping of **markers** is usually restricted to the X chromosome. Markers that are located on the X chromosome can be selected from a variety of genetic maps; for example, maps created at the Marshfield Medical Center for Medical Genetics [2] and the Foundation Jean Dausset Centre d'Etude du Polymorphisme Humain (CEPH) [7]. Markers that are approximately 5 cM–10 cM apart can be genotyped initially to perform a scan of the X chromosome. Additional markers can then be genotyped to aid in establishing linkage and to fine map a locus to a region.

Model-based Linkage Analysis

Model-based methods for X-linked loci may be viewed as modifications of model-based methods for autosomal diseases (*see* **Linkage Analysis, Model-based**; **Linkage Analysis, Multipoint**); as a result, relatively little methodologic literature on analysis of X-linkage is available. A summary of early work is given by Edwards [9]. For X-linked loci, the transmission probability (the probability that an individual of a given genotype transmits a particular gamete to his or her offspring) depends on the sex of both parent and offspring. First, consider a single X-linked locus [6]. Let $k = 1, 2, \ldots, K$ denote the possible alleles at a given locus and let 0 denote a null (absent) allele. For a male parent, the genotype is in a haploid state for one of these possible alleles. Let $\tau_{k0m \to k'm}$ ($\tau_{k0m \to k'f}$) be the probability that a male parent with genotype $k0$ transmits a k' allele to a male (female) offspring. Then $\tau_{km \to k'm} = 0$ for all k, k', and $\tau_{k0m \to k'f} = 1$, if $k = k'$, and 0 otherwise. For a female parent, the transmission probabilities

are the same as in the autosomal case, regardless of offspring sex.

When two loci are analyzed together in a linkage analysis, the recombination fraction is included as a parameter in the transmission probabilities. For X-linked loci, no recombination occurs in the male parent and the transmission probabilities are the same as above, substituting the two-locus haploid genotype for the single locus haploid genotype. Only the transmission probabilities associated with the female parent include the (female) recombination fraction; these transmission probabilities are the same as in the autosomal case.

Next, consider the mode of inheritance. For X-linked traits, the penetrance functions are usually defined separately for males and females. For a two-allele locus, the possible genotypes for a female are DD, D+ or ++, where D is the disease allele and + is the wild type allele. The male has only one copy of the X chromosome; the Y chromosome does not carry the locus. The possible genotypes for a male are therefore DY or +Y. As examples of X-linked traits with full penetrance for disease genotypes and no phenocopies, Table 1 gives penetrances for an X-linked recessive trait, Table 2 for an X-linked dominant trait, and Table 3 for an X-linked recessive trait for which female carriers can be identified.

Morton [20] recommends using, as a critical value, a lod score of 2.0 in order to establish linkage of a trait locus to the X chromosome. For a scan of only the X chromosome using a dense map of markers [18], a lod score of 2.0 corresponds to a significance level of 0.024 (*see* **Genome-wide Significance**).

Model-free Linkage Analysis

For model-free methods, sometimes called nonparametric methods, the mode of inheritance of a trait is not specified a priori (*see* **Linkage Analysis, Model-free**). Like the case of model-based analysis, model-free linkage analysis of X-linked traits can be viewed as a modification of model-free linkage analysis of autosomal traits. When analyzing relative-pair data, the number of alleles shared identical-by-descent (ibd) (*see* **Identity Coefficients**) between the members of the pair is estimated, conditional on the available marker data. For an autosomal marker, siblings can share either 0 or 1 maternal alleles (each with probability $\frac{1}{2}$) and 0 or 1 paternal alleles (each with probability $\frac{1}{2}$). Therefore, for autosomal markers and under the null hypothesis of no linkage, the expected probabilities of sibs sharing 0, 1, and 2 alleles ibd are $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$, respectively.

For X-linked markers, female–female sib pairs must share exactly 1 paternal allele ibd, male–male sib pairs must share exactly 0 paternal alleles ibd (because the Y chromosome does not carry the locus of interest), and opposite-sex sib pairs must also share exactly 0 paternal alleles ibd. For alleles inherited from the mother, both same-sex and opposite-sex sib pairs can share either 0 or 1 allele ibd (each with probability $\frac{1}{2}$). Therefore, under the null hypothesis of no linkage, female–female sib pairs can share either 1 or 2 alleles ibd (each with probability $\frac{1}{2}$), male–male sib pairs can share either 0 or 1 allele ibd (each with probability $\frac{1}{2}$), and opposite-sex sib pairs can share 0 or 1 allele (each with probability $\frac{1}{2}$).

Once ibd-sharing probabilities are estimated, they can be incorporated into standard methods of model-free analysis, with the caveat that, for some methods, different parameters are estimated for the three types

Table 1 Penetrances for an X-linked recessive trait without phenocopies or reduced penetrance

	Female			Male	
Disease status	DD	D+	++	DY	+Y
Affected	1	0	0	1	0
Unaffected	0	1	1	0	1

Table 2 Penetrances for an X-linked dominant trait without phenocopies or reduced penetrance

	Female			Male	
Disease status	DD	D+	++	DY	+Y
Affected	1	1	0	1	0
Unaffected	0	0	1	0	1

Table 3 Penetrances for an X-linked recessive trait without reduced penetrance or phenocopies, but for which carriers can be identified

	Female			Male	
Disease status	DD	D+	++	DY	+Y
Affected	1	0	0	1	0
Carriers	0	1	0	–	–
Unaffected	0	0	1	0	1

of sex-specific pairs (female–female, male–male, and male–female), and the results then combined for an overall test of linkage. For example, Cordell et al. [4] gives an extension of the Risch [24] affected-relative-pair likelihood method for autosomal loci to the analysis of X-linked loci. To derive their method, they specify sex-specific genetic **variances** and covariances and use them to obtain sex- and pair-type-specific recurrence **risk ratio** parameters and their corresponding multinomial parameters.

In general, let z_{Ri} be the probability that an affected relative pair of type R (sex-specific) shares i alleles ibd, for $i = 0, 1, 2$, and let α_{Ri} be the associated prior probability. In the case of sib pairs, $\alpha_{bb2} = \alpha_{bs2} = \alpha_{ss0} = z_{bb2} = z_{bs2} = z_{ss0} = 0$ (where bb = brother–brother, bs = brother–sister, and ss = sister–sister). The remaining αs equal $\frac{1}{2}$. Within each sex-specific pair type, the zs must sum to one; therefore, there are three free parameters in the likelihood: z_{bb0}, z_{bs0}, and z_{ss1}. For each sex-specific pair type, the lod score is maximized, subject to constraints on the parameters consistent with genetic inheritance. The asymptotic distribution of the corresponding likelihood ratio statistic is a 50 : 50 mixture of a point mass at zero (χ_0^2) and χ_1^2.

The three lod scores can then be summed to give an overall lod score; the asymptotic distribution of the corresponding **likelihood ratio** statistic is a mixture of χ_0^2, χ_1^2, χ_2^2, and χ_3^2, with mixing proportions equal to the binomial probabilities 0.125, 0.375, 0.375, and 0.125, respectively. To achieve the nominal (point-wise) significance level of 0.05, a lod score exceeding 1.18 is required [21]. If the entire genome is scanned, then the Lander & Kruglyak [18] criteria for suggestive and significant linkage correspond to lod scores equaling 3.06 and 4.62, respectively, on the X chromosome [21].

The Cordell et al. [4] method is general, allowing for different recurrence risk ratio parameters for each sex-specific sib-pair type at the expense of additional degrees of freedom. For X-linked recessive inheritance, the probabilities of sharing a maternal allele are the same for each type of affected sib pair [13]. As a result, all maternal ibd sharing can be combined and a single parameter (here denoted as z_1) to be estimated. The null hypothesis H_0: $z_1 = \frac{1}{2}$ is tested against the alternative H_a: $z_1 > \frac{1}{2}$; the asymptotic distribution of the corresponding likelihood ratio statistic is distributed as a 50 : 50 mixture of χ_0^2 and χ_1^2. Nyholt [21] argues that this statistic has the same properties as the likelihood ratio statistic from a model-based analysis and that, as a result, the Morton [20] criterion applies if only the X chromosome is scanned and the standard Lander & Kruglyak [18] criteria for model-based analysis apply if the entire genome is scanned.

The Pseudo-autosomal Regions

There are two small regions on the X chromosome where recombination does occur with the Y chromosome, the distal regions on Xp [3] and on Xq [11]. These are known as the pseudo-autosomal regions; inheritance of loci and traits in these regions of the sex chromosomes mimic autosomal inheritance. As a result, when testing for linkage within the pseudo-autosomal regions, penetrance functions take the same form as for the autosomal case (i.e autosomal dominant or recessive) [23]. If the trait locus is within a pseudo-autosomal region and the marker locus/loci lies outside this region, a different set of penetrance functions must be used (see [23]).

Dupuis & Van Eerdewegh [8] present a method to carry out affected-sib-pair analysis within the pseudo-autosomal regions. This method takes into account the fact that, in the pseudo-autosomal regions, same-sex siblings will share more genetic material ibd, even when a disease locus is not present. This increased sharing will be greater in those regions closer to the sex-specific region. Likewise, opposite-sex siblings will share less genetic material ibd. If this difference is not taken into account, for either dichotomous or quantitative traits, then those samples with more same-sex siblings will have higher type I error rates, while samples with an abundance of opposite-sex siblings will have reduced power to detect linkage.

Software for Analysis of X-linked Traits

Programs that can be used to carry out model-based linkage analysis for X-linkage include LINKAGE [19], FASTLINK [5], GENEHUNTER2.1 [16], and ALLEGRO1.1 [12]. For affected-sib-pair data, programs include MAPMAKER/SIBS [15], which incorporates the methods of Cordell et al. [4], and ASPEX, an exclusion mapping program that incorporates a one-parameter model [14]. GENEHUNTER2.1 and ALLEGRO1.1 also perform non-parametric analysis of more general pedigrees (*see* **Software for Genetic Epidemiology**).

Acknowledgments

This work is supported by NIH grant DC03 594. Special thanks to Dale Nyholt and Andrew DeWan for their comments and suggestions.

References

[1] Barker, D.F., Hostikka, S.L., Zhou, J., Chow, L.T., Oliphant, A.R., Gerken, S.C. et al. (1990). Identification of mutations in the COL4A5 collagen gene in Alport syndrome, *Science* **248**, 1224–1227.

[2] Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination, *American Journal of Human Genetics* **63**, 861–869.

[3] Cooke, H.J., Brown, W.R. & Rappold, G.A. (1985). Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal, *Nature* **317**, 687–692.

[4] Cordell, H.J., Kawaguchi Y., Todd, J.A. & Farrall, M. (1995). An extension of the maximum lod score method to X-linked loci, *Annals of Human Genetics* **57**, 920–934.

[5] Cottingham, R.W. Jr, Idury, R.M. & Schaffer, A.A. (1993). Faster sequential genetic linkage computations, *American Journal of Human Genetics* **53**, 252–263.

[6] Demenais, F.M. & Elston, R.C. (1981). A general transmission probability model for pedigree data, *Human Heredity* **31**, 93–99.

[7] Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P. et al. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites, *Nature* **380**, 152–154.

[8] Dupuis, J. & Van Eerdewegh, P. (2000). Multipoint linkage analysis of the pseudoautosomal regions, using affected sibling pairs, *American Journal of Human Genetics* **67**, 462–475.

[9] Edwards, J.H. (1971). The analysis of X-linkage, *Annals of Human Genetics* **34**, 229–250.

[10] Falls, H.F. & Cotterman, C.W. (1948). Choroidoretinal degeneration: a sex-linked form in which heterozygous women exhibit a tapetal-like retinal reflex, *Archives of Ophthalmology* **40**, 685–703.

[11] Freije, D., Helms, C., Watson, M.S. & Donis-Keller, H. (1992). Identification of a second pseudoautosomal region near the Xq and Yq telomeres, *Science* **258**, 1784–1787.

[12] Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. & Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis, *Nature Genetics* **25**, 12–13.

[13] Hallmeyer, J., Hebert, J.M., Spiker, D., Lotspeich, L., McMahon, W.M., Petersen, P.B. et al. (1996). Autism and the X chromosome, *Archives of General Psychiatry* **53**, 985–989.

[14] Hinds, D.A. & Risch, N. (1996). The ASPEX package: affected sib-pair exclusion mapping, `ftp://lahmed.stanford.edu/pub/aspex`.

[15] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.

[16] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.

[17] Lalwani, A.K., Brister, J.R., Fex, J., Grundfast, K.M., Pikus, A.T., Ploplis, B. et al. (1994). A new nonsyndromic X-linked sensorineural hearing impairment linked to Xp21.2, *American Journal of Human Genetics* **55**, 685–694.

[18] Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits:guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**, 241–247.

[19] Lathrop, G.M., Lalouel, J.M., Julier, C. & Ott, J. (1984). Strategies for multilocus linkage analysis in humans, *Proceedings of the National Academy of Sciences* **81**, 3443–3446.

[20] Morton, N.E. (1955). Sequential tests for detection of linkage, *American Journal of Human Genetics* **7**, 277–318.

[21] Nyholt, D.R. (2000). All LODs are not created equal, *American Journal of Human Genetics* **67**, 282–288.

[22] Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda), 2001. World Wide Web URL: `http://www.ncbi.nlm.nih.gov/omim/`.

[23] Ott, J. (1986). Y-linkage and pseudoautosomal linkage, *American Journal of Human Genetics* **38**, 891–897.

[24] Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs, *American Journal of Human Genetics* **46**, 229–241.

[25] Roepman, R., van Duijnhoven, G., Rosenberg, T., Pinckers, A.J.L.G., Bleeker-Wagemakers, L.M. & Bergen, A.A.B. (1996). Positional cloning of the gene for X-linked retinitis pigmentosa 3: homology with the guanine-nucleotide-exchange factor RCC1, *Human Molecular Genetics* **5**, 1035–1041.

[26] Schwahn, U., Lenzner, S., Dong, J., Feil, S., Hinzmann, B., van Duijnhoven, G. et al. (1998). Positional cloning of the gene for X-linked retinitis pigmentosa, *Nature Genetics* **19**, 327–332.

[27] Wilson, E.B. (1911). The sex chromosomes, *Archiv fuer Mikroskopische Anatomie und Entwicklungsmechanik* **77**, 249–271.

SUZANNE M. LEAL

Yates's Algorithm

Yates's **algorithm** is a computationally efficient method of calculating main effects and **interactions** in a balanced **factorial experiment**. The method, developed by **Frank Yates** [13] in 1937, originally applied to designs in which all factors have two levels, but it is easily extended to include factors with three or more levels. The calculations can be performed readily by hand through a series of multiplications and additions, and are easily programmed. A reverse algorithm yields fitted values and **residuals**.

The structural characteristics that lead to the computational efficiency of Yates's algorithm form the basis of the **fast Fourier transform** [1, 5, 8, 9], and also bear a direct relationship to computational aspects of **wavelet** shrinkage [7, 11] in, for example, **nonparametric regression**.

Standard Order of Experimental Conditions

The calculations for Yates's algorithm are organized in a table, with the data arranged in what is known as *standard order* in the first column. Before illustrating the method through examples, let us establish some notation. Consider an experiment with three factors, A, B, and C, each at two levels, designated "high" and "low". Combinations of the lowercase letters a, b, and c represent experimental conditions defined by the levels of these factors. The presence of a letter indicates that the corresponding factor is at its higher level, and its absence indicates that the factor is at its lower level. Accordingly, ab denotes observations at the high levels of A and B and the low level of C. Now suppose that an additional factor, D, with three levels, is included in the design. Successive levels of this factor are written, 1, d, and d^2. Hence, the letters bd^2 stand for an observation at the low levels of A and C, the high level of B, and the highest level of D. The observation at the lowest level of all factors is denoted 1. For factors with unordered levels, the designations "low" and "high" are, of course, arbitrary.

Observations arranged in standard order start at the lowest level of all factors, and cycle through the levels of the first factor most rapidly. In the notation just introduced, this consists of writing the levels of the first factor, then writing the letters for each level of the next factor times these terms, one level at a time, and continuing on to succeeding factors in the same manner. For example, the standard order for three factors A, B, and C with two levels each is:

$$1 \quad a \quad b \quad ab \quad c \quad ac \quad bc \quad abc.$$

Standard order for a $2 \times 2 \times 3$ design with factors A, B, and D is:

$$1 \quad a \quad b \quad ab \quad d \quad ad \quad bd \quad abd \quad d^2 \quad ad^2 \quad bd^2 \quad abd^2.$$

Notice that the first four symbols are like those for a 2×2 experiment, that multiplying each of these by d produces the next four symbols, and multiplying them by d^2 produces the last four.

Yates's Algorithm for 2^k Factorial Designs

Table 1 illustrates the steps of Yates's algorithm for a 2^4 factorial design. The column labeled Treatment total contains the data, the totals over r replications for each experimental condition, arranged in standard order. Columns (1)–(4) are calculated from these totals as follows. To generate column (1), first group the data in the preceding column into pairs. Then calculate the top half of the column as the sums of these pairs and the bottom half as their differences, subtracting the second member of the pair from the first. Thus, the entries in the top half of column (1) are $24.9 = 15.8 + 9.1$ through $63.0 = 22.6 + 40.4$, and the entries in the bottom half are $-6.7 = 9.1 - 15.8$ through $17.8 = 40.4 - 22.6$. Generate columns (2) and (3) from their preceding columns in the same manner, with sums of pairs in the top half and differences in the bottom. The number of columns calculated in this way is the same as the number of factors in the design. The final such column contains the *effect totals*, or raw effects.

The *mean effects*, which are the main effects and interactions identified in the final column, are obtained by dividing the effect totals by $r2^k$ for the average effect and by $r2^{k-1}$ for all other effects. These effects are **orthogonal**, and their corresponding one-degree-of-freedom sums of squares are calculated as $[\text{effect total}]^2/2^k$. The **standard errors** of the mean effects are $(2\sigma^2/r2^{k-1})^{1/2}$. Replicates, when available, provide an estimate of the error **variance**. In the absence of replication,

Table 1 Yates's algorithm for a 2^4 design

Experimental condition	Treatment total	(1)	(2)	(3)	Effect total (4)	Divisor	Mean effect = Effect total/ divisor	Sum of squares = Effect total^2/ $r2^k$	Identification
1	15.8	24.9	58.7	124.4	324.6	$r2^k = 16$	20.29	6585.32	Average
a	9.1	33.8	65.7	200.2	30.2	$r2^{k-1} = 8$	3.78	57.00	*A*
b	14.6	27.8	88.2	−2.2	44.0	8	5.50	121.00	*B*
ab	19.2	37.9	112.0	32.4	48.0	8	6.00	144.00	*AB*
c	16.8	38.6	−2.1	19.0	30.8	8	3.85	59.29	*C*
ac	11.0	49.6	−0.1	25.0	14.4	8	1.80	12.96	*AC*
bc	16.1	49.0	10.0	22.8	4.2	8	0.53	1.10	*BC*
abc	21.8	63.0	22.4	25.2	1.4	8	0.18	0.12	*ABC*
d	19.8	−6.7	8.9	7.0	75.8	8	9.48	359.10	*D*
ad	18.8	4.6	10.1	23.8	34.6	8	4.33	74.82	*AD*
bd	19.3	−5.8	11.0	2.0	6.0	8	0.75	2.25	*BD*
abd	30.3	5.7	14.0	12.4	2.4	8	0.30	0.36	*ABD*
cd	22.2	−1.0	11.3	1.2	16.8	8	2.10	17.64	*CD*
acd	26.8	11.0	11.5	3.0	10.4	8	1.30	6.76	*ACD*
bcd	22.6	4.6	12.0	0.2	1.8	8	0.23	0.20	*BCD*
abcd	40.4	17.8	13.2	1.2	1.0	8	0.13	0.06	*ABCD*

the higher-order interactions are often taken to be zero, and their effects pooled to obtain an estimate of σ^2. A **half-normal** plot of the absolute standardized effects [|effect total|/(divisor)$^{1/2}$], preferably excluding main effects, is useful in determining which higher-order effects to pool.

Yates's Algorithm for Factors with More than Two Levels

In general, orthogonal **contrasts** among factor levels determine the steps of Yates's algorithm. For a factor with two levels, the contrasts are unique, but for a factor with three or more levels, these contrasts, which decompose the average and main effect into one-degree-of-freedom effects, can be chosen in any number of ways. If the factor levels are quantitative and evenly spaced, then **polynomial** contrasts are appropriate. The columns of $\mathbf{C}_2$ and $\mathbf{C}_3$ below contain the polynomial contrast coefficients for two and three factors, with the sums of squared coefficients in the margins. The steps just described for a 2^k design clearly correspond to use of the coefficients in the columns of $\mathbf{C}_2$:

$$\begin{array}{c} \\ \mathbf{C}_2 = \\ \\ \end{array} \begin{array}{c} \text{Average} \quad \text{Linear} \\ \left[\begin{array}{cc} 1 & -1 \\ 1 & 1 \\ 2 & 2 \end{array}\right], \end{array}$$

$$\begin{array}{c} \\ \mathbf{C}_3 = \\ \\ \end{array} \begin{array}{c} \text{Average} \quad \text{Linear} \quad \text{Quadratic} \\ \left[\begin{array}{ccc} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \\ 3 & 2 & 6 \end{array}\right] \end{array}$$

Table 2 contains the steps of Yates's algorithm for a $2 \times 3 \times 2$ design using polynomial contrasts. Because *A* and *C* each have two levels, columns (1) and (3) are calculated from the preceding columns by computing sums and differences of pairs as described above for 2^k designs. Thus, only calculation of column (2) corresponding to the three-level factor *B* remains to be explained. To obtain this column, first divide the preceding column into groups of three (the number of levels of the factor). Then apply the coefficients in the columns of $\mathbf{C}_3$ to these groups. That is, compute the first third of column (2) as the sums of each group, compute the middle third using the coefficients −1, 0, 1 for the linear contrast, and compute the last third using the coefficients 1, −2, 1 for the quadratic contrast. Thus, the first four entries in column (2) are $74.5 = 21.8 + 18.0 + 14.7$ through $53.5 = 12.1 + 15.6 + 25.8$; the next four are $12.9 = 34.7 - 21.8$ through $13.7 = 25.8 - 12.1$, and the last four are $20.5 = 21.8 - 2 \times 18.0 + 34.7$ through $6.7 = 12.1 + 2 \times 15.6 + 25.8$.

The contrasts not only define the operations on the column elements in Yates's algorithm, but also

Table 2 Yates's algorithm for a $2 \times 3 \times 2$ design using polynomial contrasts

Experimental condition	Treatment total	(1)	(2)	Effect total (3)	Divisor	Mean effect = Effect total/ divisor	Sum of squares = Effect total^2/ divisor	Identification
1	10.6	21.8	74.5	183.8	$2 \times 3 \times 2 \times r = 12r$	15.317	2815.30	Average
a	11.2	18.0	109.3	69.6	$2 \times 3 \times 2r = 12r$	5.800	403.68	A
b	7.2	34.7	16.1	33.6	$2 \times 2 \times 2r = 8r$	4.200	141.12	B_L
ab	10.8	29.9	53.5	25.0	$2 \times 2 \times 2r = 8r$	3.125	78.13	AB_L
b^2	11.4	28.8	12.9	43.4	$2 \times 6 \times 2r = 24r$	1.808	78.48	B_Q
ab^2	23.3	50.6	20.7	12.0	$2 \times 6 \times 2r = 24r$	0.500	6.00	AB_Q
c	8.9	0.6	11.3	34.8	$2 \times 3 \times 2r = 12r$	2.900	100.92	C
ac	21.0	3.6	13.7	37.4	$2 \times 3 \times 2r = 12r$	3.117	116.56	AC
bc	6.6	11.9	20.5	7.8	$2 \times 2 \times 2r = 8r$	0.975	7.61	B_LC
abc	22.2	12.1	22.9	2.4	$2 \times 2 \times 2r = 8r$	0.300	0.72	AB_LC
b^2c	12.4	15.6	5.3	2.4	$2 \times 6 \times 2r = 24r$	0.100	0.24	B_QC
ab^2c	38.2	25.8	6.7	1.4	$2 \times 6 \times 2r = 24r$	0.058	0.08	AB_QC
	Coefficients:	1st half: 1, 1 2nd half: −1, 1	1st third: 1, 1, 1 2nd third: −1, 0, 1 Last third: 1, −2, 1	1st half: 1, 1 2nd half: −1, 1				

determine the divisors used to obtain mean effects and sums of squares from effect totals. The divisors are products of the sums of squared contract coefficients, with each factor in the design contributing to the product. If a particular factor is not involved in the effect (e.g. C is not in the AB interaction), then its average contrast supplies its contribution to the divisor. Thus, the divisor for the A main effect is $2 \times 3 \times 2 = 12(A_{\text{linear}} \times B_{\text{average}} \times C_{\text{average}})$; the divisor for the AB_Q interaction is $2 \times 6 \times 2 = 24(A_{\text{linear}} \times B_{\text{quadratic}} \times C_{\text{average}})$.

The Reverse Algorithm

The reverse algorithm operating on the mean or raw effects produces the original data. If certain effects that are considered negligible or nonsignificant are set to zero, then the reverse algorithm produces the fitted values for the corresponding model. On the other hand, if significant effects are set to zero and the values for the nonsignificant effects retained, then this back-transformation produces residuals.

The reverse algorithm for 2^k designs is straightforward. Starting with raw effects, columns are again generated as sums and differences of pairs of elements in the preceding column. However, in the reverse algorithm, the bottom half of the column consists of sums of pairs and the top half of differences obtained by subtracting the second member of the pair from the first. These steps correspond to taking linear combinations with coefficients from the columns of $\mathbf{C}_2'$. Equivalently, the reverse algorithm can be implemented by applying the forward algorithm to the raw effects listed in reverse standard order, which produces the original data (fitted values, residuals) in reverse order as well [2].

When the design includes factors with more than two levels, the reverse algorithm is applied to mean effects rather than effect totals and no divisors are used. The steps for two-level factors are as described above. The columns of $\mathbf{C}_3'$ contain the coefficients for calculating each third of a column corresponding to a three-level factor, assuming polynomial contrasts have been used in the forward algorithm:

$$\mathbf{C}_2' = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \mathbf{C}_3' = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix}.$$

Table 3 The reverse algorithm for a $2 \times 3 \times 2$ design using polynomial contrasts
(a) Fitted values

Identification	Mean effect	(1)	(2)	Fitted values (3)	Experimental condition
Average	**15.32**	9.52	10.26	11.45	1
A	**5.80**	1.07	−1.19	10.55	a
B_L	**4.20**	1.81	15.60	6.12	b
AB_L	**3.13**	−0.22	5.05	11.48	ab
B_Q	**1.81**	0.97	5.90	11.65	b^2
AB_Q	0.00	0.00	−0.22	23.27	ab^2
C	**2.90**	21.12	17.50	9.07	c
AC	**3.12**	7.33	6.02	20.65	ac
B_LC	**0.97**	1.81	12.40	5.68	bc
AB_LC	0.00	6.02	0.75	23.52	abc
B_QC	0.00	0.97	30.26	13.15	b^2c
AB_QC	0.00	0.00	6.99	37.25	ab^2c

(b) Residuals

Identification	Mean effect	(1)	(2)	Residuals (3)	Experimental condition
Average	0.00	0.00	−0.50	−0.84	1
A	0.00	0.00	0.34	0.64	a
B_L	0.00	−0.50	0.50	1.08	b
AB_L	0.00	0.00	−0.14	−0.68	ab
B_Q	0.00	−0.30	1.00	−0.24	b^2
AB_Q	**0.50**	0.04	−0.08	0.04	ab^2
C	0.00	0.00	−1.00	−0.16	c
AC	0.00	0.00	−0.32	0.36	ac
B_LC	0.00	0.50	−0.50	0.92	bc
AB_LC	**0.30**	0.00	−0.26	−1.32	abc
B_QC	**0.10**	0.30	0.50	−0.76	b^2c
AB_QC	**0.06**	0.16	0.46	0.96	ab^2c
	Coefficients:	1st half: 1, −1 2nd half: 1, 1	1st third: 1, −1, 1 2nd third: 1, 0, −2 Last third: 1, 1, 1	1st half: 1, −1 2nd half: 1, 1	

Table 3 illustrates the reverse algorithm for the $2 \times 3 \times 2$ example presented in Table 2, with two-way interactions involving the quadratic effect of B and all three-way interactions set to zero a priori. Calculations of fitted values and **residuals** follow exactly the same steps, so we describe only the former. To compute column (1) corresponding to the first factor which has two levels, group the preceding column into pairs. Apply the coefficients −1, 1 from the first column of $\mathbf{C}_2'$ to obtain the differences $9.52 = 15.32 - 5.80$, and so forth in the top half of column (1). Similarly, apply the coefficients in the second column of $\mathbf{C}_2'$ to obtain the sums of each pair in the bottom half of column (1). Since the second factor has three levels, compute column (2) by first dividing the preceding column into groups of three. Then apply the coefficients in each column of $\mathbf{C}_3'$ to obtain the linear combinations that fill each third of the

column:

$$\begin{aligned} 10.26 &= 9.52 - 1.07 + 1.81 \text{ using } 1, -1, 1 \\ &\vdots \qquad\quad \vdots \\ 5.90 &= 9.52 - 2 \times 1.81 \text{ using } 1, 0, -2 \\ &\vdots \qquad\quad \vdots \\ 12.40 &= 9.52 + 1.07 + 1.81 \text{ using } 1, 1, 1 \\ &\vdots \qquad\quad \vdots \end{aligned}$$

Column (3), the last column calculated in this manner, contains the estimated effects for the fitted model with terms $A + B_L + B_Q + C + AB_L + AC + B_LC$.

Computational Efficiency: Yates's Algorithm, and the Fast Fourier and Wavelet Transforms

Yates's algorithm for a $t_1 \times t_2 \times \ldots \times t_k$ design requires $(t_1 + t_2 + \ldots + t_k)N$ multiplications and additions, in contrast to the N^2 operations in direct calculation from the design matrix, $\mathbf{X}$. The former quantity is proportional to $N \log_2 N$, with constant of proportionality that is a weighted average of $t_1/\log_2 t_1, \ldots, t_k/\log_2 t_k$ [8, 5]. The computational efficiency of Yates's algorithm derives from the fact that $\mathbf{X}'$, which is the direct (Kronecker) product of the transposes of the relevant $t \times t$ contrast matrices, can be written as the usual matrix product of sparse $N \times N$ matrices with at most $t \times N$ nonzero elements. Calculation of columns (1) through (k) in Yates's table corresponds to multiplication by these sparse matrices.

Good [8, 9] generalized Yates's algorithm as presented above and demonstrated its application to the calculation of Fourier series. Cooley & Tukey [5] developed the fast Fourier transform based on this work. McCullagh [11] noted the similarity in thresholding of wavelet coefficients and back-transformation in Yates's algorithm, and Donoho & Johnstone [7] verified that the formal manipulations of Yates's algorithm are exactly the same as computations in particular cases of wavelet transformation.

Yates's algorithm is described and illustrated in various books on the statistical analysis of designed experiments in addition to those cited, such as Box et al. [3], Cochran & Cox [4], Daniel [6], Johnson & Leone [10], and Miller & Freund [12].

References

[1] Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*. Wiley, New York.

[2] Box, G.E.P., Hunter W.G. & Hunter J.S. (1978). *Statistics for Experimenters*. Wiley, New York.

[3] Box, G.E.P., Connor, L.R., Cousins, W.R., Davies, O.L., Himsworth, F.R. & Sillitto, G.P. (1963). *The Design and Analysis of Industrial Experiments*, 2nd Ed. Oliver & Boyd, Edinburgh.

[4] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York.

[5] Cooley, J.W. & Tukey, J.W. (1965). An algorithm for the machine computation of complex Fourier series, *Mathematical Computation* **19**, 297–301.

[6] Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. Wiley, New York.

[7] Donoho, D.L. & Johnstone, I.M. (1995). Wavelet shrinkage: asymptopia?, *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

[8] Good, I.J. (1958). The interaction algorithm and practical Fourier analysis, *Journal of the Royal Statistical Society, Series B* **20**, 361–372.

[9] Good, I.J. (1960). The interaction algorithm and practical Fourier analysis: an addendum, *Journal of the Royal Statistical Society, Series B* **22**, 372–375.

[10] Johnson, N.L. & Leone, F.C. (1977). *Statistics and Experimental Design in Engineering and the Physical Sciences*, 2nd Ed. Wiley, New York.

[11] McCullagh, P. (1995). In discussion of Donoho, D.L. & Johnstone, I.M. (1995). Wavelet shrinkage: asymptopia?, *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

[12] Miller, I. & Freund, J.E. (1985). *Probability and Statistics for Engineers*, 3rd Ed. Prentice-Hall, Englewood Cliffs.

[13] Yates, F. (1937). The design and analysis of factorial experiments, *Technical Communication 35*. Imperial Bureau of Soil Science, Harpenden.

M. DRUM

Yates's Continuity Correction

For testing the null hypothesis of independence in a **2 × 2 contingency table**,

$$\begin{array}{cc|c} a & b & n_1 \\ c & d & n_2 \\ \hline m_1 & m_2 & N \end{array},$$

Yates [4] defined a continuity correction to provide a P value that approximates the P value from **Fisher's exact test** better than that based on the uncorrected **chi-square test** statistic,

$$T = \frac{N(ad - bc)^2}{m_1 m_2 n_1 n_2}. \tag{1}$$

Yates's continuity-corrected chi-square statistic is

$$T_c = \frac{N\left(|ad - bc| - \frac{1}{2}N\right)^2}{m_1 m_2 n_1 n_2}. \tag{2}$$

Conditional on the table's margins, only one cell count, say $A = a$, is random (A denotes a random variable taking value a). The discrepancy of outcome A from its expected value under H_0, $A' = A - E(A)$, where $E(A) = m_1 n_1 / N$, identifies outcomes in opposite tails of the distribution. Keeping these tails distinct, Yates's [4, 5] two-sided exact P value is twice the one-sided P value, with a maximum of one:

$$P = \begin{cases} 2\ \Pr(A' \le a'), & \text{if } a' < 0, \\ 2\ \Pr(A' \ge a'), & \text{if } a' > 0, \end{cases} \tag{3}$$

where $\Pr(A' = a') = \Pr(A = a)$ is the **hypergeometric probability** of outcome a,

$$\Pr(A = a) = \frac{\binom{n_1}{a}\binom{n_2}{m_1 - a}}{\binom{N}{m_1}}, \tag{4}$$

and $\max(0, m_1 + n_1 - N) \le a \le \min(m_1, n_1)$.

A two-sided asymptotic P value, $\overline{F}(t)$ [or $\overline{F}(t_c)$], is found by referring t (or t_c) to the **chi-square distribution** with 1 df; F is the cumulative distribution function of this distribution and $\overline{F} = 1 - F$.

The exact distribution is discrete and its tails $\Pr(A' \le a')$ and $\Pr(A' \ge a')$ can be asymmetric, whereas the approximating distribution is continuous and assumes that these tails are symmetric. The continuity-corrected chi-square P value greatly improves the approximation to the exact P value.

Example

For the 2×2 table with $a = 5$ and margins $\{m_1 = 6, n_1 = 8, N = 14\}$, we obtain $E(A) = 3.429$ and $a' = 1.571$. Table 1 lists the seven possible values of a, conditional on the margins, showing the two tails of the distribution. The two-sided exact conditional P value is $2\Pr(A' \ge 3.429) = 0.2424$. The chi-square statistics are $t = 2.941$ (uncorrected) and $t_c = 1.367$ (continuity-corrected); the asymptotic P values are $\overline{F}(t) = 0.0864$ and $\overline{F}(t_c) = 0.2423$, respectively. Table 1 illustrates general results: $\overline{F}(t)$ underestimates the exact distribution, while $\overline{F}(t_c)$ is nearly identical to it, and $\overline{F}(t_c)$ is slightly conservative when very small (but can be improved by using the Fisher–Yates reference table VIII [2]).

Table 1 Two-sided exact and asymptotic P values for all 2×2 contingency tables with margins $\{m_1 = 6, n_1 = 8, N = 14\}$

			Exact	Not corrected		Corrected	
a	a'	$\Pr(A = a)$	P	t	$\overline{F}(t)$	t_c	$\overline{F}(t_c)$
0	−3.429	0.0003	0.0006	14.00	0.0002	10.21	0.0014
1	−2.429	0.0160	0.0326	7.024	0.0080	4.430	0.0353
2	−1.429	0.1399	0.3124	2.431	0.1190	1.027	0.3109
3	−0.429	0.3730	1.0000	0.2187	0.6400	0.0061	0.9379
4	0.571	0.3497	0.9418	0.3889	0.5329	0.0061	0.9379
5	1.571	0.1119	0.2424	2.941	0.0864	1.367	0.2423
6	2.571	0.0093	0.0186	7.875	0.0050	5.110	0.0238

Discussion

Much of the continuing controversy over Yates's continuity correction is really over the exact reference distribution. It makes sense to use this correction if one believes that the conditioning used in Fisher's exact test is appropriate. Opponents argue that the reference distribution should sometimes be unconditional, depending on the underlying sampling process [3, 5].

In addition, Yates's reference P value, [3], which is based on outcomes in the observed tail only, has been misrepresented by a P value based on outcomes in both tails [$\Pr(|A'| \geq |a'|)$], and, in turn, his approximation has been miscalculated [1, 5]. Eq. (3) is increasingly recognized as correct [3, p. 369]; it is a smooth function of $|a'|$, while the miscalculated P value is not.

Yates recommends the continuity correction whenever the smallest expected value is less than 500 [2]. In practice, however, the exact P value itself is typically computed today (*see* **Exact Inference for Categorical Data**).

References

[1] Conover, W.J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables (with discussion), *Journal of the American Statistical Association* **69**, 374–382.

[2] Fisher, R.A. & Yates, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver & Boyd, Edinburgh (6th edition, 1963).

[3] Haviland, M.G. (1990). Yates' correction for continuity and the analysis of 2×2 contingency tables (with discussion), *Statistics in Medicine* **9**, 363–383.

[4] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society* **1**, Supplement, 217–235.

[5] Yates, F. (1984). Tests of significance for 2×2 contingency tables (with discussion), *Journal of the Royal Statistical Society, Series A* **147**, 426–463.

JOAN F. HILTON

Yates, Frank

Born: May 12, 1902, in Manchester, UK.
Died: June 17, 1994, in Harpenden, UK.

Reproduced by permission of the Royal Statistical Society

Frank Yates spent almost the whole of his working life as Head of the Statistics Department at Rothamsted Experimental Station, the large agricultural research station situated some 40 km northwest of London. The department had been founded by **R.A. Fisher** on his appointment to Rothamsted in 1919. He recruited Yates in 1931 and, shortly afterwards, left to take the Chair of Eugenics at University College, London. Yates was left in charge with just one other member of staff; his long career saw the department grow to a total of over 20 statisticians and he maintained its reputation as one of the world's leading centers of statistical research.

Before going to Rothamsted, Yates had been working in the Gold Coast (present-day Ghana) as mathematician to the colony's geodetic survey. Here, he obtained a thorough grounding in the theory and practice of **least squares** calculations, a body of knowledge that provided a foundation for much of his future work. At Rothamsted, his first main field of interest was **experimental design**. Fisher had laid down the basic principles of replication, **randomization, stratification**, and factorial treatment structure (*see* **Factorial Experiments**) in the 1920s and these had been adopted remarkably quickly by agricultural research workers, but some of the complexities had been imperfectly appreciated, even by members of Fisher's staff. Yates was able to put the subject onto a sound footing and to relate it properly to the **analysis of variance**, a technique at which he became a virtuoso. His work on factorial designs, including mixed factorials and **Latin square designs**, was expounded in his booklet *The Design and Analysis of Factorial Experiments* [5], which was, for several years, the only available manual on the subject. At the same time, he developed the wide range of **incomplete block designs**, including the quasi-factorial lattices (*see* **Lattice Designs**), and solved a variety of the associated combinatorial problems [8].

Throughout this period and later, the Rothamsted department was responsible for the analysis of the field experiments done on the Rothamsted farm, as well as for many other calculations. Yates brought to this work an unrivaled concern for numerical accuracy, which he had developed during his work on geodetic surveys. His ongoing interest in computational matters was to bear fruit later on.

During World War II, Yates's interests moved to different areas. He was responsible for undertaking several large exercises in what would today be called **meta-analysis**, examining and summarizing all the available evidence on responses to fertilizers and on the feeding of dairy cattle [1]. These studies – a good deal more sophisticated than most in current practice – were the basis for the wartime control of imported fertilizers and animal feedstuffs. He was also engaged in **operations research** work under S. Zuckerman. At the same time, he and the Rothamsted department became involved in the design and analysis of several large-scale sample surveys, and the theory and practice of survey work became the second of Yates's main areas of interest. His work was summarized in his book *Sampling Methods for Censuses and Surveys* [6].

Yates's interest in computation led him to experiment productively with punched-card equipment, but came to full fruition with the advent of electronic computers in the 1950s. He was, in fact, among the first to realize the enormous potential importance of these machines for the development of statistics. In 1954, Rothamsted was able to install the first British electronic computer to operate away from its designers. Within a year or two this machine (unimaginably

small-scale and primitive by today's standards) was in regular use for the analysis of experiments and surveys on a large scale, and for much other work, notably in the area of **multivariate analysis**. Yates played a leading role in the design and construction of the necessary programs, insisting from the start on a degree of discipline and user-friendliness that was most unfamiliar at the time.

Experimental design, sample survey, and computing were the three main headings under which Yates's work can be classified, but there were other notable contributions. An important and neglected article [7] discussed an economic argument for deciding upon the amount of experimentation justifiable on a given topic, showing that in an agricultural context a much larger number of experiments than usually envisaged would amply pay for themselves. It should be noted that concepts of size and **power**, the conventional determinants of sample size (*see* **Sample Size Determination**), did not enter into the case. An important paper on the analysis of **two-by-two tables** was published in 1934 [4] and followed (surely uniquely) by a substantial sequel just 50 years later [9] during his active retirement. It is ironic that **"Yates's continuity correction"**, which he may not even have originated, seems to be his best-known contribution among non-statisticians. Also worthy of mention are the several editions of Fisher and Yates's Statistical Tables [2], a collection (with its substantial introduction) that is still useful even in the computer era.

Yates's working life was spent mainly in an agricultural research setting. He is a giant figure from the heroic age of modern statistics, rivaling Student (**William Sealy Gosset**) as the greatest of applied statisticians, and much of his work has yet to be taken full advantage of by the biostatistical community.

A fuller account of Yates's career appears in [3].

References

[1] Crowther, E.M. & Yates, F. (1941). Fertilizer policy in wartime: the fertilizer requirements of arable crops, *Empire Journal of Experimental Agriculture* **9**, 77–97.

[2] Fisher, R.A. & Yates, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver & Boyd, Edinburgh. Subsequent enlarged editions in 1942, 1948, 1953, 1957, 1963.

[3] Healy, M.J.R. (1995). Frank Yates, 1902–1994 – the work of a statistician, *International Statistical Review* **63**, 271–88.

[4] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society* **1**, Supplement, 217–235.

[5] Yates, F. (1937). *The Design and Analysis of Factorial Experiments*. Imperial Bureau of Soil Science Technical Communication no. 35, Harpenden.

[6] Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. Griffin, London. Subsequent revised and expanded editions in 1953, 1960, 1981.

[7] Yates, F. (1952). Principles governing the amount of experimentation required in development work, *Nature* **170**, 138–140.

[8] Yates, F. (1972). *Experimental Design*. Griffin, London.

[9] Yates, F. (1984). Tests of significance for 2×2 contingency tables, *Journal of the Royal Statistical Society, Series A* **147**, 426–463.

M.J.R. HEALY

Youden Squares and Row–Column Designs

The technique of **blocking** is used to control variability due to extraneous sources, each source of variability constituting a blocking factor. Two sources can be controlled using a Youden square or a row–column design, the rows and columns comprising two different systems of blocking. For example, in a medical study, if each subject is to be used several times, then rows and columns may represent subjects and occasions, respectively. In an animal experiment conducted over several days, litters of animals and days may comprise the row and column blocking factors, respectively. As an illustration, the following design is a row–column design for four treatments in four rows and six columns:

1	3	1	4	4	2
2	1	2	3	1	4
4	4	3	2	2	3
3	2	4	1	3	1

Let D_1 be the block design obtained by taking rows as blocks, ignoring the columns. Similarly, D_2 denotes the design obtained by taking columns as blocks, ignoring the rows. D_1 and D_2 are called the row and column component designs, respectively. In the above design the row component design D_1 is variance balanced for estimating **paired comparisons** of treatment 1 with treatments 2, 3, and 4. The column component D_2 is a **randomized complete blocks design**. It is important to arrange the treatments in rows and columns so that the design has a high overall efficiency factor and the individual **contrasts** of interest are also estimated with high efficiencies. With this in mind, classes of row–column designs have been defined based upon the **orthogonality** of the component designs D_1 and D_2. A design in which both D_1 and D_2 are randomized complete block designs is called a **Latin square**. Therefore, Latin squares are row–column designs with the numbers of rows, columns, and treatments all equal, and all the treatments occur once in each row and in each column. If practical constraints do not make it possible to have both D_1 and D_2 as randomized complete block designs, one of them may be taken to be an **incomplete block design**. A row–column design in which D_1 is a randomized complete block design and the other component D_2 is a **balanced incomplete block design** is called a Youden square. Obviously, the roles of D_1 and D_2 are interchangeable. Incomplete Latin squares of Yates [60], obtained by deleting a row or a column from a Latin square, are examples of Youden squares. Youden [61] constructed them using symmetrical balanced incomplete block designs; the construction method was later provided by Smith & Hartley [57] for all symmetrical balanced incomplete block designs. A Youden square for seven treatments in three rows and seven columns is shown below:

1	2	3	4	5	6	7
2	3	4	5	6	7	1
4	5	6	7	1	2	3

Analysis of Designs

We first consider the analysis of a general row–column design for v treatments arranged in p rows and q columns. The analysis of a Youden square design will be provided later as particular case. It is assumed that a treatment is allocated to each of the combinations of rows and columns, which means that the row and column classifications are orthogonal. For a valid **randomization** of row–column designs, we first randomly permute the rows and then randomly permute the columns [47]. Let treatment t be assigned to the experimental unit in the ith row and jth column, and let y_{ijt} be the observed response from this unit. Then, the model for the data is assumed to be

$$y_{ijt} = \mu + \rho_i + \gamma_j + \tau_t + \in_{ijt},$$

where μ is the overall mean, ρ_i and γ_j are the effects of the ith row and jth column, respectively, τ_t is the effect of the tth treatment, and $\in_{ijt}$ are **random errors** assumed to be independently distributed with zero **mean** and a constant **variance** σ^2, $i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, q$; $t = 1, 2, \ldots, v$. Let $\mathbf{N}_1$ and $\mathbf{N}_2$ be the incidence matrices of the row and column component designs D_1 and D_2, respectively. The (i, j)th element of $\mathbf{N}_1(\mathbf{N}_2)$ equals 1 if the ith treatment occurs in the jth row (column), and is zero otherwise. Then the reduced normal equations for estimating treatment parameters are given by $\mathbf{C}\tau = \mathbf{Q}$, where $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_v)'$,

$$\mathbf{C} = \mathbf{r}^{\delta} - \frac{1}{q}\mathbf{N}_1\mathbf{N}_1' - \frac{1}{p}\mathbf{N}_2\mathbf{N}_2' + \frac{1}{n}\mathbf{r}\mathbf{r}'$$

is the **information matrix** of the design, $\mathbf{r} = r_1, r_2, \ldots, r_v)'$, r_i is the number of replications of treatment i, $\mathbf{r}^\delta$ is the diagonal matrix with elements those of $\mathbf{r}$, a prime denotes transpose,

$$\mathbf{Q} = \mathbf{T} - \frac{1}{q}\mathbf{N}_1\mathbf{T}_1 - \frac{1}{p}\mathbf{N}_2\mathbf{T}_2 + \frac{G}{n}\mathbf{1}$$

is the vector of adjusted treatment totals, $\mathbf{1}$ is a column vector of ones, $\mathbf{T}$, $\mathbf{T}_1$, $\mathbf{T}_2$ are the vectors of treatment, row and column totals, respectively, G is the grand total of observations, and $n = pq$ is the total number of observations. A solution to the normal equations is given by $\hat{\boldsymbol{\tau}} = \boldsymbol{\Omega}\mathbf{Q}$, where $\boldsymbol{\Omega}$ is a generalized inverse of $\mathbf{C}$ satisfying $\mathbf{C}\boldsymbol{\Omega}\mathbf{C} = \mathbf{C}$. It is important to note that $\boldsymbol{\tau}$ cannot be estimated uniquely, and only contrasts among treatment parameters are estimable. The adjusted treatment sum of squares is $\hat{\boldsymbol{\tau}}'\mathbf{Q} = \mathbf{Q}'\boldsymbol{\Omega}\mathbf{Q}$. The **analysis of variance** table is constructed as shown in Table 1.

The sum of squares due to error is

$$SSE = \mathbf{Y}'\mathbf{Y} - \frac{\mathbf{T}_1'\mathbf{T}_1}{q} - \frac{\mathbf{T}_2'\mathbf{T}_2}{p} - \mathbf{Q}'\boldsymbol{\Omega}\mathbf{Q} + \frac{G^2}{n}.$$

The information matrix $\mathbf{C}$ has rank less than or equal to $v - 1$. A design is said to be connected if the rank of $\mathbf{C}$ is equal to $v - 1$, otherwise it is said to be disconnected. All treatment comparisons are estimable in a connected design. We will assume that all comparisons among treatments are of interest, and thus we restrict our attention to connected designs only. Connectedness in row–column designs has been considered by Shah & Khatri [54], Raghavarao & Federer [49], Russell [50], and Sia [56]. Let $\mathbf{s}'\boldsymbol{\tau}$, $\mathbf{s}'\mathbf{1} = 0$, be a contrast of interest. The (unique) **least squares** estimate of $\mathbf{s}'\boldsymbol{\tau}$ is given by $\mathbf{s}'\hat{\boldsymbol{\tau}} = \mathbf{s}'\boldsymbol{\Omega}\mathbf{Q}$ with variance $\mathrm{var}(\mathbf{s}'\hat{\boldsymbol{\tau}}) = \mathbf{s}'\boldsymbol{\Omega}\mathbf{s}\sigma^2$. The sum of squares due to $\mathbf{s}'\boldsymbol{\tau}$ is $(\mathbf{s}'\boldsymbol{\Omega}\mathbf{Q})^2/\mathbf{s}'\boldsymbol{\Omega}\mathbf{s}$. For a Youden square, $\mathbf{C} = (\lambda v/p)(\mathbf{I} - 1/vJ)$ and $\boldsymbol{\Omega} = (p/\lambda v)\mathbf{I}$, where λ is the number of times each pair of treatments occurs in D_2. Therefore, the adjusted treatment sum of squares is $(p/\lambda v)\mathbf{Q}'\mathbf{Q}$, and $\mathbf{s}'\hat{\boldsymbol{\tau}} = (p/\lambda v)\mathbf{s}'\mathbf{Q}$ with $\mathrm{var}(\mathbf{s}'\hat{\boldsymbol{\tau}}) = (p/\lambda v)\mathbf{s}'\mathbf{s}\sigma^2$.

Table 1 Analysis of variance calculations

Source of variation	Degrees of freedom	Sum of squares
Treatments (adjusted)	$v - 1$	$\mathbf{Q}'\boldsymbol{\Omega}\mathbf{Q}$
Rows	$p - 1$	$\frac{\mathbf{T}_1'\mathbf{T}_1}{q} - \frac{G^2}{n}$
Columns	$q - 1$	$\frac{\mathbf{T}_2'\mathbf{T}_2}{p} - \frac{G^2}{n}$
Error	$(p-1)(q-1) \times (v-1)$	SSE (by subtraction)
Total (corrected)	$n - 1$	$\mathbf{Y}'\mathbf{Y} - \frac{G^2}{n}$

Classes of Designs

A design is variance-balanced if all pairwise comparisons between treatments are estimated with the same variance. For variance-balanced designs $\mathbf{C} = \alpha[\mathbf{I} - 1/v\mathbf{J}]$ and $\boldsymbol{\Omega} = 1/\alpha\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, $\mathbf{J}$ is the square matrix of 1s, and α is a constant whose value depends upon the design. Youden squares are variance-balanced since for these designs $\mathrm{var}(\hat{\tau}_i - \hat{\tau}_j) = (2p/\lambda v)\sigma^2$, $i \neq j = 1, 2, \ldots, v$. The following design given by Pearce [46] is also variance-balanced:

1	4	2	3	1	4
3	3	4	1	2	2
4	2	3	2	4	1
2	1	4	1	3	3
2	2	3	4	1	1
3	4	1	4	3	2

For this design, $\mathrm{var}(\hat{\tau}_i - \hat{\tau}_j) = (3/25)\sigma^2$, $i \neq j = 1, 2, 3, 4$. Generalized Youden designs by Kiefer [33], which have been listed by Ash [2], provide a large class of variance-balanced designs. The earliest example of a variance-balanced design in which both D_1 and D_2 are partially balanced appears to be the design of Kshirsagar [34] for $v = 9$, $p = q = 6$. In some experiments all pairwise comparisons among treatment parameters are not of equal importance. Supplemented balanced, or S, designs introduced by Hoblyn et al. [22] and Pearce [45] are useful where the main object is to compare several test or new treatments with a standard treatment which is called the control treatment. With treatment 1 coded as the control, in these designs, $\mathrm{var}(\hat{\tau}_1 - \hat{\tau}_i) = a_1\sigma^2$ and $\mathrm{var}(\hat{\tau}_i - \hat{\tau}_j) = a_2\sigma^2$, $i \neq j = 2, 3, \ldots, v$, where a_1 and a_2 are some constants. In the block design setting, reinforced designs of Das [11] obtained by augmenting each block of a balanced incomplete block design with a constant number of replications of the control treatment are of type S. In the incomplete block design setting, Bechhofer & Tamhane [5] called type S designs balanced treatment incomplete block designs.

Several methods of constructing optimal type S block designs have been provided in the literature; excellent review are by Hedayat et al. [21] and Majumdar [40]. The row–column design for four treatments in four rows and six columns which was given in the introduction is an S design with $\text{var}(\hat{\tau}_1 - \hat{\tau}_i) = (51/140)\sigma^2$, $\text{var}(\hat{\tau}_i - \hat{\tau}_j) = (48/140)\sigma^2$, $i \neq j = 2, 3, 4$. Note that there is only a small difference between the two variances, such designs being called nearly balanced. In general, $\text{var}(\hat{\tau}_1 - \hat{\tau}_i) < \text{var}(\hat{\tau}_i - \hat{\tau}_j)$, $i \neq j = 2, 3, \ldots, v$, for S designs, as is the case for the following design:

1	1	2	3	4
1	1	4	2	3
2	3	1	4	1
3	4	1	1	2
4	2	3	1	1

For this design $\text{var}(\hat{\tau}_1 - \hat{\tau}_i) = 0.3\sigma^2$, $\text{var}(\hat{\tau}_i - \hat{\tau}_j) = 0.4\sigma^2$, $i \neq j = 2, 3, 4$. Ture [59] gave a catalog of efficient type S row–column designs. Some type S row–column designs can be constructed using the methods of Gupta et al. [20], Kumari et al. [36], Majumdar [39], Majumdar and Tamhane [41] and Pearce [48]. Nair & Rao [43] defined intra- and inter-group balanced designs that are not restricted to two types of treatments; Pearce [46] referred to them as multipartite designs. In **factorial experiments** the contrasts of interest among treatment parameters represent different main effects and **interactions**. Let $\mathbf{s}_i'\boldsymbol{\tau}$ represent these single df main effects and interaction normalized contrasts, normalized such that $\mathbf{s}_i'\mathbf{s}_i = 1$, $i = 1, 2, \ldots, v - 1$. Then var $(\mathbf{s}_i'\hat{\boldsymbol{\tau}}) = a\sigma^2$ in a variance-balance design, where a is some constant. The variance-balanced design given earlier in this Section for $v = 2^2$, $p = q = 6$, is appropriate for a 2^2 experiment for which $\text{var}(\mathbf{s}_i'\hat{\boldsymbol{\tau}}) = 0.06\sigma^2$, $i = 1, 2, 3$. Practical constraints often dictate the use of a partially variance-balanced design for factorial experiments. In a partially balanced design with factorial balance (Shah [52], and Kshirsagar [35]), or balanced confounded designs of Nair & Rao [44], all the single df contrasts belonging to the same main effect or interaction are estimated with the same variance. The contrasts belonging to two different factorial effects may have unequal variances. A wide class of such designs for two-factor experiments is provided by group divisible designs. Some group divisible designs may have factorial balance even when more than two factors are involved. In a group divisible design treatments are divided into m groups, each group containing n treatments, and $v = mn$. Pairwise comparisons of any two treatments in the same group are all estimated with the same variance, say $a_1\sigma^2$, and pairwise comparisons of any two treatments belonging to two different groups are also estimated with the same variance, say $a_2\sigma^2$, with $a_1 \neq a_2$, where a_1, a_2 are some constants. With appropriate coding of the treatment labels, as mentioned earlier these designs are balanced factorially as well for two-factor experiments. A group divisible row–column design is shown below:

1	4	2	6	5	3
2	5	1	3	4	6
3	6	5	1	2	4
4	1	3	2	6	5

Several designs for two or more than two factors have been given by Suen & Chakravarty [58]. John & Lewis [29] gave a wide class of row–column designs for factorial experiments using a generalized cyclic method of construction, but their designs may not be factorially balanced. Under the general setting, Freeman [18] gave several series of partially balanced designs and Freeman [19] considered designs with unequal replications.

A design is called row-orthogonal if the row component D_1 is orthogonal to the treatments. The group-divisible design given above satisfies the condition of row orthogonality. Statistical properties of a row-orthogonal design can be evaluated from its column component D_2. Several group divisible designs listed by Clatworthy [6] can be rearranged to provide row-orthogonal row–column designs. Cyclic designs of John et al. [31] and Lamacraft & Hall [37] also provide a wide class of efficient row-orthogonal designs. A row–column design has adjusted orthogonality [15, 16] if the estimates of the row (column) parameters do not depend on whether or not column (row) parameters are included in the model. For adjusted orthogonal designs, the information matrices of D_1, D_2 and the row–column designs have the same **eigenvectors** [14]. All row-orthogonal designs have adjusted orthogonality. The α-designs of John & Eccleston [28] provide a wide class of efficient adjusted orthogonal designs. Some useful adjusted orthogonal designs have also been considered [1, 4, 12, 38, 53]. Row–column designs satisfying various optimality criteria have been considered in [3, 7–10,

17, 23–27, 30, 39, 42, 55], and [59]. General **algorithms** for constructing row–column designs using computers have been considered in [13, 32, 49], and [51].

References

[1] Anderson, D.A. & Eccleston, J.A. (1985). On the construction of a class of efficient row–column designs, *Journal of Statistical Planning and Inference* **11**, 131–134.

[2] Ash, A. (1981). Generalized Youden designs: construction and tables, *Journal of Statistical Planning and Inference* **5**, 1–25.

[3] Bagchi, S. & Shah, K.R. (1989). On the optimality of a class of row–column designs, *Journal of Statistical Planning and Inference* **23**, 397–402.

[4] Bagchi, S. & van Berkum, E.E.M. (1991). On the optimality of a class of adjusted orthogonal designs, *Journal of Statistical Planning and Inference* **28**, 61–65.

[5] Bechhofer, R.E. & Tamhane, A.C. (1981). Incomplete block designs for comparing treatments with a control: general theory, *Technometrics* **23**, 45–57.

[6] Clatworthy, W.H. (1973). Tables of two-associate-class partially balanced designs, *Applied Mathematics, Series 63*. National Bureau of Standards.

[7] Das, A. (1993). E-optimal block and row-column designs with unequal numbers of replicates, *Sankhyā, Series B* **55**, 77–90.

[8] Das, A. & Dey, A. (1990). Optimality of row-column designs, *Calcutta Statistical Association Bulletin* **39**, 63–72.

[9] Das, A. & Dey, A. (1991). Optimal variance- and efficiency-balanced designs for one- and two-way elimination of heterogeneity, *Metrika* **38**, 227–238.

[10] Das, A. & Dey, A. (1992). Universal optimality and non-optimality of some row–column designs, *Journal of Statistical Planning and Inference* **31**, 263–271.

[11] Das, M.N. (1958). On reinforced incomplete block designs, *Journal of the Indian Society of Agricultural Statistics* **10**, 73–77.

[12] Eccleston, J.A., John, J.A. & Whitaker, D. (1993). Some row–column designs with adjusted orthogonality, *Journal of Statistical Planning and Inference* **36**, 331–346.

[13] Eccleston, J.A. & Jones, B. (1980). Exchange and interchange procedures to search for optimal row-and-column designs, *Journal of the Royal Statistical Society, Series B* **42**, 372–376.

[14] Eccleston, J.A. & Kiefer, J. (1981). Relationships of optimality for individual factors of a design, *Journal of Statistical Planning and Inference* **5**, 213–219.

[15] Eccleston, J.A. & Russell, K.G. (1975). Connectedness and orthogonality in multi-factor designs, *Biometrika* **62**, 341–345.

[16] Eccleston, J.A. & Russell, K.G. (1977). Adjusted orthogonality in nonorthogonal designs, *Biometrika* **64**, 339–345.

[17] Eccleston, J.A. & Russell, K.G. (1980). (M,S)-optimal row–column designs, *Communications in Statistics – Theory and Methods* **9**, 449–452.

[18] Freeman, G.H. (1958). Families of designs for two successive experiments, *Annals of Mathematical Statistics* **29**, 1063–1078.

[19] Freeman, G.H. (1975). Row-and-column designs with two groups of treatments having different replications, *Journal of the Royal Statistical Society, Series B* **37**, 114–128.

[20] Gupta, V.K., Ramana, D.V.V. & Agarwal, S.K. (1998). Weighted *A*-optimal row-column designs for treatment-control comparisons, *Journal of Combinatorics, Information and System Sciences* **23**, 333–344.

[21] Hedayat, A.S., Jacroux, M. & Majumdar, D. (1988). Optimal designs for comparing test treatments with controls (with discussion), *Statistical Science* **3**, 462–491.

[22] Hoblyn, T.N., Pearce, S.C. & Freeman, G.H. (1954). Some considerations in the design of successive experiments in fruit plantations, *Biometrics* **10**, 503–515.

[23] Jacroux, M. (1985). Some E and MV-optimal designs for the two-way elimination of heterogeneity, *Annals of the Institute of Statistical Mathematics* **37**, 557–566.

[24] Jacroux, M. (1986). Some E-optimal row-column designs, *Sankhya, Series B* **48**, 31–39.

[25] Jacroux, M. (1987). Some E and MV-optimal row-column designs having equal numbers of rows and columns, *Metrika* **34**, 361–381.

[26] Jacroux, M. (1990). Some E-optimal row-column designs having unequally replicated treatments, *Journal of Statistical Planning and Inference* **26**, 65–81.

[27] Jarrett, R.G., Piper, F.C. & Wild, P.R. (1997). Efficient two-replicate resolvable row-column designs, *Journal of Statistical Planning and Inference* **58**, 65–77.

[28] John, J.A. & Eccleston, J.A. (1986). Row-column α-designs, *Biometrika* **73**, 301–306.

[29] John, J.A. & Lewis, S.M. (1983). Factorial experiments in generalized cyclic row-column designs, *Journal of the Royal Statistical Society, Series B* **45**, 245–251.

[30] John, J.A. & Williams, E.R. (1997). The construction of efficient two-replicate row-column designs for use in field trials, *Applied Statistics* **46**, 207–214.

[31] John, J.A., Wolock, F.W. & David, H.A. (1972). Cyclic Designs, *Applied Mathematics, Series 62*. National Bureau of Standards.

[32] Jones, B. (1979). Algorithms to search for optimal row-and-column designs, *Journal of the Royal Statistical Society, Series B* **41**, 210–216.

[33] Kiefer, J. (1975). Constructions and optimality of generalized Youden designs, in *A Survey of Statistical Design and Linear Models*, J.N. Srivastava, ed. North Holland, New York, pp. 333–353.

[34] Kshirsagar, A.M. (1957). On balancing designs in which heterogeneity is eliminated in two-directions, *Calcutta Statistical Association Bulletin* **7**, 161–166.

[35] Kshirsagar, A.M. (1966). Balanced factorial designs, *Journal of the Royal Statistical Society, Series B* **28**, 559–567.

[36] Kumari, S., Mehta, B.D. & Batra, S.D. (2000). Some supplemented row-column designs with varying replications, *Calcutta Statistical Association Bulletin* **50**, 43–48.

[37] Lamacraft, R.R. & Hall, W.B. (1982). Tables of cyclic incomplete block designs: $r = k$, *Australian Journal of Statistics* **24**, 350–360.

[38] Lewis, S.M. & Dean, A.M. (1991). On general-balance in row-column designs, *Biometrika* **78**, 595–600.

[39] Majumdar, D. (1986). Optimal designs for comparisons between two sets of treatments, *Journal of Statistical Planning and Inference* **14**, 359–372.

[40] Majumdar, D. (1996). Optimal and efficient treatment-control designs, in *Handbook of Statistics*, Vol. 13, S. Ghosh & C.R. Rao, ed. North Holland, Amsterdam, pp. 1007–1053.

[41] Majumdar, D. & Tamhane, A.C. (1996). Row-column designs for comparing treatments with a control, *Journal of Statistical Planning and Inference* **49**, 387–400.

[42] Mandeli, J.P. (1999). Construction of systematic row-column designs with treatments unconfounded with the linear row × columns and quadratic row × columns interactions, *Statistics and Probability Letters* **42**, 229–237.

[43] Naur, K.R. & Rao, C.R. (1942). Incomplete block designs for experiments involving several groups of varieties, *Science and Culture* **7**, 615–616.

[44] Nair, K.R. & Rao, C.R. (1948). Confounding in asymmetrical factorial experiments, *Journal of the Royal Statistical Society, Series B* **10**, 109–131.

[45] Pearce, S.C. (1960). Supplemented balance, *Biometrika* **47**, 263–271.

[46] Pearce, S.C. (1963). The use and classification of non-orthogonal designs (with discussion), *Journal of the Royal Statistical Society, Series A* **126**, 353–377.

[47] Pearce, S.C. (1975). Row-and-column designs, *Applied Statistics* **24**, 60–74.

[48] Pearce, S.C. (1994). Reinforced lattices, *Journal of the Royal Statistical Society, Series B* **56**, 469–476.

[49] Raghavarao, D. & Federer, W.T. (1975). On connectedness in two-way elimination of heterogeneity designs, *Annals of Statistics* **3**, 730–735.

[50] Russell, K.G. (1976). The connectedness and optimality of a class of row-column designs, *Communications in Statistics – Theory and Methods* **5**, 1479–1488.

[51] Russell, K.G., Eccleston, J.A. & Knudsen, G.J. (1981). Algorithms for the construction of (M. S)-optimal block designs and row-column designs, *Journal of Statistical Computation and Simulation* **12**, 93–105.

[52] Shah, B.V. (1958). On balancing in factorial experiments, *Annals of Mathematical Statistics* **29**, 766–779.

[53] Shah, K.R. & Eccleston, J.A. (1986). Some aspects of row–column designs, *Journal of Statistical Planning and Inference* **15**, 87–95.

[54] Shah, K.R. & Khatri, C.G. (1973). Connectedness in row–column designs, *Communications in Statistics – Theory and Methods* **2**, 571–573.

[55] Shah, K.R. & Sinha, B.K. (1993). Optimality aspects of row–column designs with non-orthogonal structure, *Journal of Statistical Planning and Inference* **36**, 331–346.

[56] Sia, L.L. (1977). Some properties of connectedness in two-way designs, *Communications in Statistics – Theory and Methods* **6**, 1165–1170.

[57] Smith, C.A.B. & Hartley, H.O. (1948). The construction of Youden squares, *Journal of the Royal Statistical Society, Series B* **10**, 262–263.

[58] Suen, C.-Y. & Chakravarty, I.M. (1985). Balanced factorial designs with two-way elimination of heterogeneity, *Biometrika* **72**, 391–402.

[59] Ture, T.E. (1994). Optimal row–column designs for multiple comparisons with a control: a complete catalog, *Technometrics* **36**, 292–299.

[60] Yates, F. (1936). Incomplete Latin squares, *Journal of Agricultural Science* **26**, 301–315.

[61] Youden, W.J. (1937). Use of incomplete block replication in estimating tobacco-mosaic virus, *Contribution of Boyce Thompson Institute* **9**, 41–48.

SUDHIR GUPTA

Yule Process

The Yule process is a birth process (*see* **Stochastic Processes**) based on the assumptions of independence and of a constant birth rate, λ. Consider a time interval (t_0, t) and let $X(t)$ be the number of individuals present at time t, with the initial number $X(0) = n_0$ at $t = t_0$. The transition **(conditional) probabilities** of $X(t)$,

$$P_{n_0,n}(t_0, t) = \Pr[X(t) = n | X(0) = n_0], \quad n = n_0, n_0 + 1, \ldots, \tag{1}$$

satisfy the differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{n_0,n_0}(t_0, t) = -n_0 \lambda P_{n_0,n_0}(t_0, t) \tag{2a}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{n_0,n}(t_0, t) = -n\lambda P_{n_0,n}(t_0, t) + (n-1)\lambda P_{n_0,n-1}(t_0, t), \tag{2b}$$

for $n = n_0 + 1, n_0 + 2, \ldots$, with the initial conditions at $t = t_0$,

$$P_{n_0,n_0}(t_0, t_0) = 1 \quad \text{and} \quad P_{n_0,n}(t_0, t_0) = 0, \quad \text{for} \quad n \neq n_0.$$

Solving the differential equations (2a) and then (2b) successively, beginning with $n = n_0 + 1$, we find:

$$P_{n_0,n}(0, t) = \binom{n-1}{n-n_0} [\exp(-\lambda t)]^{n_0} \times [1 - \exp(-\lambda t)]^{n-n_0}, \quad n = n_0, n_0 + 1, \ldots. \tag{3}$$

Now let $Y(t) = X(t) - n_0$ be the number of births occurring during the interval (t_0, t). Then

$$\Pr[Y(t) = k] = \Pr[X(t) = k + n_0 | X(0) = n_0] = \binom{k + n_0 - 1}{k} [\exp(-\lambda t)]^{n_0} \times [1 - \exp(-\lambda t)]^k, \quad k = 0, 1, \ldots, \tag{4}$$

which is a **negative binomial distribution** with parameters n_0 and $\exp(-\lambda t)$. The expectations of $Y(t)$ and $X(t)$ are

$$\mathrm{E}[Y(t)] = \mathrm{E}[X(t)] - n_0 = n_0[\exp(\lambda t) - 1],$$

and the variances are

$$\mathrm{var}[Y(t)] = \mathrm{var}[X(t)] = n_0 \exp(\lambda t)[\exp(\lambda t) - 1].$$

This type of distribution was first studied by **G.U. Yule** [2, 3]. In his mathematical theory of evolution, Yule studied the number of species in a genus, where n_0 is the number of species at initial time $t = t_0$, $Y(t)$ is the number of species produced during the time period $(0, t)$, and $X(t)$ is total number of species in a genus present at time t. Yule did not use differential equations, but derived formulas (3) and (4) by a limiting process.

A General Case

Suppose r observations $[X(t_1), X(t_2), \ldots, X(t_r)]$ are made at r points $[t_1, t_2, \ldots, t_r]$ on the time axis. Since clearly the distribution of any random variable $X(t_i)$ depends only on the values of the random variable $X(t_{i-1})$, and not on any random variable before t_{i-1}, the joint probability distribution of $[X(t_1), X(t_2), \ldots, X(t_r)]$ is a product of r conditional probabilities:

$$\begin{aligned}\Pr[X(t_1) = n_1, X(t_2) = n_2, \ldots, X(t_r) = n_r | X(t_0)] &= \prod_{i=1}^{r} \Pr[X(t_i) = n_i | X(t_{i-1}) = n_{i-1}] \\ &= \prod_{i=1}^{r} \binom{n_i - 1}{n_i - n_{i-1}} [\exp(-\lambda \tau_i)]^{n_{i-1}} \\ &\quad \times [1 - \exp(-\lambda \tau_i)]^{n_i - n_{i-1}},\end{aligned} \tag{5}$$

where $\tau_i = t_i - t_{i-1}$, for any positive integer r.

Now let $Y(t_i) = X(t_i) - n_{i-1}$ be the number of births taking place during the interval (t_{i-1}, t_i). The joint distribution of $[Y(t_1), Y(t_2), \ldots, Y(t_r)]$ is obtained directly from formula (5) with the substitution of $n_i - n_{i-1} = k_i$:

$$\Pr[Y(t_1) = k_1, Y(t_2) = k_2, \ldots, Y(t_r) = k_r] = \prod_{i=1}^{r} \binom{n_i - 1}{k_i} [\exp(-\lambda \tau_i)]^{n_{i-1}} \times [1 - \exp(-\lambda \tau_i)]^{k_i}, \tag{6}$$

for any positive integer r. Formula (6) is a chain of negative binomial distributions [1].

References

[1] Chiang, C.L. (1980). *An Introduction to Stochastic Processes and Their Applications*. Krieger, New York.

[2] Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*, 2nd Ed. Wiley, New York.

[3] Yule, G.U. (1924). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S., *Philosophical Transactions of the Royal Society of London, Series B* **213**, 21–87.

CHIN LONG CHIANG

Yule, George Udny

Born: February 18, 1871, near Haddington, Scotland.
Died: June 26, 1951, in Cambridge, UK.

Reproduced by permission of the Royal Statistical Society

George Udny Yule was educated at Winchester College from where, at the age of only 16, he transferred to University College London to study engineering. **Karl Pearson**, then Professor of Applied Mathematics, was beginning to develop his own interest in statistics, and in 1892 offered Yule a post as demonstrator. In 1896, Yule was appointed Assistant Professor of Applied Mathematics, a post that he held for three years, until he resigned it in favor of more remunerative employment.

His continuing interest in statistics, however, led to his appointment as Newmarch Lecturer in Statistics at University College in 1902, a post which he held concurrently with his other work until 1909, and which led to the publication of the book which made his name, *An Introduction to the Theory of Statistics*. The first edition appeared in 1911 [4], and during his lifetime there were 13 further editions. The 11th edition was the first to be jointly undertaken with **M.G. Kendall** [6], and by the time of the 14th and last edition of "Yule & Kendall", in 1950, Kendall's own two-volume *The Advanced Theory of Statistics* was already establishing itself. It, and its present-day descendants, still bear the marks of Yule's pioneering effort.

In 1912, a Lectureship in Statistics was established for Yule by the University of Cambridge, to be held in the Faculty of Agriculture, and this, coupled with a Fellowship at St John's College from 1922, provided him with congenial employment (save for the war years) until 1931 when he retired, by then Reader in Statistics. He kept up his College teaching until the second war, and died in Cambridge in 1951.

Yule played an important part in the affairs of the **Royal Statistical Society**, of which he was honorary secretary for 12 years and subsequently President (1924–1926). He was elected a Fellow of the Royal Society in 1922.

Yule's main contributions in the theoretical field were concerned with **regression** and **correlation**, association in **contingency tables**, Mendelian genetics (*see* **Mendel's Laws**), epidemiology, and **time series**. In the Pearsonian fields of regression and correlation he gave more prominence than his mentor to the former, perhaps easing the path toward **R.A. Fisher's** invention of the **analysis of variance**.

Yule's studies of the correlation of continuous variables led him, in 1900 [1], to study measures of association for discrete variables, in particular the cross ratio θ (**odds ratio**) in a **two-by-two table** and its transform $Q = (1-\theta)/(1+\theta)$, now known as "Yule's coefficient". This led to an altercation with Pearson, in which Pearson's capacity for acrimonious and ill-directed criticism was displayed, in marked contrast to Yule's gentler mode of expression. Even Fisher, who as a young man had also felt the sharpness of Pearson's pen, was later moved to remark "Pearson attacked Yule's work at one time much more violently than ever he did mine". In 1903 [3], Yule, making use of his understanding of partial correlation, described what was much later to be termed **Simpson's paradox**, in which the pairwise associations at two levels in a $2 \times 2 \times 2$ table can be seemingly incompatible with the marginal association. In his work on association, all of his numeric examples were drawn from biology.

In Mendelian genetics, Yule [2] was the pioneer in suggesting that the observed correlations between parent and offspring could be accounted for by

multifactorial Mendelian inheritance, as Fisher fully acknowledged in his classic treatment of the correlation between relatives in 1918.

In 1914, in collaboration with F.L. Engledow, Yule invented the method of minimum chi-square for estimating a genetic recombination fraction, and the following year, with **M. Greenwood**, he was the first to recognize that there was something wrong with Pearson's **chi-square test** of association in respect of the **degrees of freedom** used, as Fisher was later to prove. He introduced the simple birth process of stochastic theory (the "**Yule process**") in connection with evolution in 1924 [5].

F. Yates ended his Royal Society obituary notice of Yule as follows: "We may ... justly conclude that although Yule did not fully develop any completely new branches of statistical theory, he took the first steps in many directions which were later to prove fruitful lines for further progress. ... In the biological field ... his work provided a corrective to many of the errors committed by the biometric school, and served to spread the use of statistical methods amongst biologists who might otherwise have been wholly repelled by them. He can indeed rightly claim to be one of the pioneers of modern statistics."

References

[1] Yule, G.U. (1900). On the association of attributes in statistics: with illustrations from the material of the Childhood Society, &c., *Philosophical Transactions of the Royal Society of London, Series A* **194**, 257–319.

[2] Yule, G.U. (1902). Mendel's Laws and their probable relations to intra-racial heredity, *New Phytologist* **1**, 193–207, 222–238.

[3] Yule, G.U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* **2**, 121–134.

[4] Yule, G.U. (1911). *An Introduction to the Theory of Statistics*. Griffin, London.

[5] Yule, G.U. (1924). A mathematical theory of evolution, based on the conclusions of Dr J.C. Willis, F.R.S., *Philosophical Transactions of the Royal Society of London, Series B* **213**, 21–87.

[6] Yule, G.U. & Kendall, M.G. (1973). *An Introduction to the Theory of Statistics*. Griffin, London.

Bibliography

Yule, G.U. (1971). *Selected Papers of George Udny Yule*, A. Stuart & M.G. Kendall, eds. Griffin, High Wycombe.

A.W.F. EDWARDS

Yule–Walker Equations

The sequence $\{X_t : t = 0, \pm 1, \pm 2, \ldots\}$, defined by

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + e_t, \quad (1)$$

where $\phi_p \neq 0$ and $\{e_t\}$ is a sequence of uncorrelated random variables each with zero mean and variance σ^2, is called an *autoregressive time series of order p* (*see* **ARMA and ARIMA Models**). The notation AR(p) is commonly used. Define $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, where B is the backward shift operator (*see* **Backward and Forward Shift Operators**) such that $BX_t = X_{t-1}$. Then the condition for **stationarity** is that the roots of the equation $\phi(Z) = 0$ must lie outside the unit circle [1, Section 3.2]. Under this condition, X_t may be expressed as a convergent (in the mean square sense) series in terms of e_s, $s \leq t$.

An important recurrence relation for the **autocorrelation function** of a stationary autoregressive process AR(p) is found by multiplying by X_{t-k} throughout (1) to obtain

$$X_{t-k}X_t = \phi_1 X_{t-k}X_{t-1} + \phi_2 X_{t-k} X_{t-2} + \cdots + \phi_p X_{t-k}X_{t-p} + X_{t-k}e_t, \quad (2)$$

On taking expected values in (2), we obtain the difference equation

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p}, \quad k > 0, \quad (3)$$

where γ_k is the autocovariance function, defined by

$$\gamma_k = \mathrm{E}[(X_t - \mathrm{E}X_t)(X_{t-k} - \mathrm{E}X_{t-k})] = \mathrm{E}X_t X_{t-k},$$

since $\mathrm{E}X_t = \mathrm{E}X_{t-k} = 0$. Note that the expectation $\mathrm{E}(X_{t-k}e_t)$ vanishes when $k > 0$, since X_{t-k} is a linear combination of e_s up to $t - k$, which are uncorrelated with e_t. On dividing through in (3) by γ_0, it is seen that the autocorrelation function satisfies the same form of the difference equation

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}, \quad k > 0, \quad (4)$$

where the autocorrelation function ρ_k is defined by $\rho_k = \gamma_k/\gamma_0$.

If we substitute $k = 1, 2, \ldots, p$ in (4) and use $\rho_k = \rho_{-k}$, we obtain a set of linear equations for $\phi_1, \phi_2, \ldots, \phi_p$ in terms of $\rho_1, \rho_2, \ldots, \rho_p$. Then

$$\begin{array}{lcllll} \rho_1 & = & \phi_1 & + \phi_2\rho_1 & \cdots & + \phi_p\rho_{p-1} \\ \rho_2 & = & \phi_1\rho_1 & + \phi_2 & \cdots & + \phi_p\rho_{p-2} \\ \vdots & & \vdots & \vdots & \cdots & \vdots \\ \rho_p & = & \phi_1\rho_{p-1} & + \phi_2\rho_{p-2} & \cdots & + \phi_p. \end{array} \quad (5)$$

These are usually called the *Yule–Walker equations* [3, 4]. If we first estimate the autocorrelation functions by moment estimators we can then use the Yule–Walker equations to obtain estimators of $\{\phi_1, \ldots, \phi_p\}$. These estimators are called the *Yule–Walker estimators*.

For the multivariate Yule–Walker equations, see [2, Section 2.8].

References

[1] Box, G.E.P. & Jenkins, G.M. (1970). *Time Series Analysis, Forecasting and Control*. Holden Day, San Francisco.

[2] Fuller, W.A. (1996). *Introduction to Statistical Time Series*, 2nd Ed. Wiley, New York.

[3] Walker, G. (1931). On periodicity in series of related terms, *Proceedings of the Royal Society, Series A* **131**, 518–523.

[4] Yule, G.U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, *Philosophical Transactions of the Royal Society* **A226**, 267–298.

BING CHENG

Z Analysis

Z-analysis is a method for determining the **association** between two probabilistic events, A and B, on the basis of the calculation of a statistical link coefficient, $Z(A, B)$. The $Z(A, B)$ coefficient may range between -1, when A and B are mutually exclusive events, and $+1$, when A is included within B; it is equal to 0 when A and B are independent. Intermediate values between -1 and 0 quantify the partial exclusion of B by A, while values between 0 and $+1$ quantify the partial dependence of B on A.

Let us call $p(A)$ and $p(B)$ the probabilities of A and B, and $p(B|A)$ the probability of B when A occurred. $Z(A, B)$ is defined as [2]

$$Z(A, B) = \frac{[p(B|A) - p(B)]}{[1 - p(B)]} \quad \text{when } p(B|A) - p(B) \geq 0 \qquad (1)$$

$$Z(A, B) = \frac{[p(B|A) - p(B)]}{p(B)} \quad \text{when } p(B|A) - p(B) < 0. \qquad (2)$$

$Z(A, B)$ is continuous in zero because both (1) and (2) tend to 0 when $p(B|A)$ tends to $p(B)$. If A and B are mutually exclusive, $p(B|A) = 0$ and $Z(A, B) = -1$ from (2). If A is included within B, then $p(B|A) = 1$ and $Z(A, B) = 1$ from (1).

$Z(A, B)$ can be estimated by approximating the probabilities by the observed frequencies:

$$Z^{\wedge}(A, B) = \frac{[n_{AB}/n_A - n_B/N]}{[1 - n_B/N]} \quad \text{when } \frac{n_{AB}/n_A - n_B}{N} \geq 0 \qquad (3)$$

$$Z^{\wedge}(A, B) = \frac{[n_{AB}/n_A - n_B/N]}{[n_B/N]} \quad \text{when } \frac{n_{AB}/n_A - n_B}{N} < 0, \qquad (4)$$

where n_A = number of times that A occurred, n_B = number of times that B occurred, n_{AB} = number of times that A and B occurred simultaneously, and N = total number of observations. Simulations showed that $Z^{\wedge}(A, B)$ is an unbiased estimator of $Z(A, B)$ [2].

The computation of Z can be clarified by the following fictitious study investigating the association between the **smoking** habit and educational level in a sample of 48 subjects. Three educational levels are considered: lower than high school ($L1$), high school graduate ($L2$), and college graduate ($L3$). Subjects are classified as heavy smokers (h, more than 10 cigarettes smoked daily), mild smokers (m, less than 10 cigarettes), or nonsmokers (n). Data are summarized in Table 1.

The association between heavy smoking and each educational level is quantified by $Z(L3, h)$, $Z(L2, h)$, and $Z(L1, h)$. To calculate $Z(L3, h)$, we first observe that the number of heavy smokers in the $L3$ class is $\mathrm{n}_{L3,h} = 2$ (first cell of the last column) and that $(\mathrm{n}_{L3,h}/\mathrm{n}_{L3}\text{-}\mathrm{n}_h/\mathrm{N}) = (2/15\text{–}17/48)$ is lower than 0. Thus, we should use (4), obtaining

$$Z(L3, h) = \frac{(\mathrm{n}_{L3,h}/\mathrm{n}_{L3} - \mathrm{n}_h/\mathrm{N})}{(\mathrm{n}_h/\mathrm{N})} = \frac{(2/15 - 17/48)}{(17/48)} = -0.62. \qquad (5)$$

Similarly, we also obtain

$$Z(L2, h) = \frac{(\mathrm{n}_{L2,h}/\mathrm{n}_{L2} - \mathrm{n}_h/\mathrm{N})}{(\mathrm{n}_h/\mathrm{N})} = \frac{(6/17 - 17/48)}{(17/48)} = -0.003. \qquad (6)$$

Since $(\mathrm{n}_{L1,h}/\mathrm{n}_{L1} - \mathrm{n}_h/\mathrm{N}) = (9/16 - 17/48)$ is greater than 0, we calculate $Z(L1, h)$ from (3):

$$Z(L1, h) = \frac{(\mathrm{n}_{L1,h}/\mathrm{n}_{L1} - \mathrm{n}_h/\mathrm{N})}{(1 - \mathrm{n}_h/\mathrm{N})} = \frac{(9/16 - 17/48)}{(1 - 17/48)} = +0.34. \qquad (7)$$

$Z(L3, h) < 0$ and $Z(L1, h) > 0$ indicate respectively that heavy smoking is partially excluded by the highest educational level and associated with the lowest level. $Z(L2, h)$ close to 0 means that the probability

Table 1 Number of heavy (h), mild (m), or non (n) smokers in each educational level

Smoking habit	Educational level			
	$L1$	$L2$	$L3$	
h	9	6	2	$\mathrm{n}_h = 17$
m	5	5	4	$\mathrm{n}_m = 14$
n	2	6	9	$\mathrm{n}_n = 17$
	$\mathrm{n}_{L1} = 16$	$\mathrm{n}_{L2} = 17$	$\mathrm{n}_{L3} = 15$	$\mathrm{N} = 48$

Table 2 Example of Z-analysis to assess the links between systolic blood pressure (SBP) and RR-interval (RR) series; beat-by-beat data is derived from a 24-h recording in a healthy subject. First row and first column contain 10 SBP and 8 RR classes, cells show $Z^{\wedge}(A, B)$ only for couples of classes with $n_{AB} > 10$. Black, grey, and white backgrounds indicate cells of exclusion ($Z^{\wedge} \leq -0.2$), independence ($-0.2 < Z^{\wedge} < 0.2$), and bond ($Z^{\wedge} \geq 0.2$) respectively

SBP(mmHg) / *RR (ms)*	<80	80–97.5	97.5–115	115–132.5	132.5–150	150–167.5	167.5–185	185–202.5	202.5–220	>220
<500					−0.83	0.00	0.02	0.04		
500–625			−0.97	−0.77	−0.11	0.02	0.07	0.15	0.20	0.29
625–750		−0.29	−0.91	−0.56	−0.09	0.12	0.41	0.53	0.45	0.22
750–875		−0.61	−0.82	0.00	0.07	0.06	−0.02	−0.29	0.01	
875–1000	0.17	−0.05	−0.36	0.04	0.09	0.01	−0.58	−0.87		
1000–1125	0.01	0.33	0.39	0.07	−0.20	−0.43	−0.77			
1125–1250		−0.42	0.14	0.03	−0.55	−0.80	−0.80			
1250–1375			0.02	0.01	−0.84					

of finding a heavy smoker does not change if we know that the subject belongs to the $L2$ educational class.

Z-analysis can be also used to quantify the statistical link between two time series, $\{x_n\}$ and $\{y_n\}$ (*see* **Coherence Between Time Series**). In this case, the A and B events are the occurrence of x_n and y_n within specific amplitude intervals. By choosing N intervals for $\{x_n\}$ and M for $\{y_n\}$, $Z(A, B)$ is defined over a grid of $N \times M$ couples of (A, B) events. An example is shown in Table 2, where a matrix of $Z(A, B)$ estimates is obtained from a 24-h monitoring of systolic blood pressure and RR-interval data. Similar grids of $Z(A, B)$ have been used to separately quantify the links between blood pressure and heart rate time series due to direct central controls and those due to the baroreflex control of blood pressure [1].

References

[1] Cerutti, C., Ducher, M., Lantelme, P., Gustin, M.P. & Paultre, C.Z. (1995). Assessment of spontaneous baroreflex sensitivity in rats: a new method using the concept of statistical dependence, *American Journal of Physiology* **268**, R382–R388.

[2] Ducher, M., Cerutti, C., Gustin, M.P. & Paultre, C.Z. (1994). Statistical relationships between systolic blood pressure and heart rate and their functional significance in conscious rats, *Medical and Biological Engineering and Computing* **32**, 649–655.

PAOLO CASTIGLIONI

Zelen Leadership Award and Lecture

The Marvin Zelen Leadership award was instituted by the Department of Biostatistics at the Harvard School of Public Health (HSPH) in 1997 to commemorate Professor Zelen's contributions to statistical science. The award recognizes an individual in government, industry, or academia, who by virtue of his/her outstanding leadership, has greatly influenced the theory and practice of statistical science. While individual accomplishments are considered, the most distinguishing criterion is the awardee's contribution to the creation of an environment in which statistical science and its applications have flourished. The selection committee consists of two members from the Department of Biostatistics (HSPH), and the three previous awardees. Nominations for candidates for the award are solicited from all members of the profession and are due in the fall of each year. Nominators should include a letter describing the contributions of a candidate, specifically highlighting the criteria for the award, and a curriculum vita.

The award is given annually each spring at a ceremony in Boston in which the awardee delivers a lecture of their choosing. The awardees and the lecture titles are:

1997 Dr. C. Frederick Mosteller, Professor Emeritus, Harvard University, "The Importance of Clinical Trials in Education"

1998 Sir David Cox, Nuffield College, University of Oxford, "Graphical Models in Statistics: A Review"

1999 Dr. **John W. Tukey**, Professor Emeritus, Princeton University, "A Smorgasbord of Handy Techniques That Can Help In Analyzing Data"

2000 Dr. Lincoln Moses, Professor Emeritus, Stanford University, "Deciding Whether Large Clinical Trials And Meta-Analyses Agree Or Not"

2001 Professor Niels Keiding, Professor of Biostatistics, University of Copenhagen, "Event Histories And Their Analysis"

2002 Dr. Robert O'Neill, Director of the Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, "A Perspective on the Development and Future of Statistics at the FDA"

2003 Dr. Wayne A. Fuller, Emeritus Distinguished Professor in Liberal Arts and Sciences, Iowa State University, "Analytic Studies with Complex Survey Data"

2004 Professor Robert C. Elston, Case Western University, "The analysis of Case-control Data to Detect Candidates Genes"

DAVID HARRINGTON

Zero Padding

Zero-padding is a procedure that consists of extending the length of a **time series** by adding zeros. For instance, if $\{x_0, x_1, \ldots, x_{M-1}\}$ is a sequence of M data, the zero-padded sequence of length $N = M + 5$ is $\{x_0, x_1, \ldots, x_{M-1}, 0, 0, 0, 0, 0\}$.

Adding zeros to a time series before computing the Discrete Fourier Transform (DFT) results in the evaluation of a Fourier transform with additional interpolated values. In fact, if $\{y_k\}$ is obtained by padding P zeros to a time series $\{x_k\}$ of length M, the DFTs before and after appending zeros are

$$X_m = \sum_{k=0}^{M-1} x_k e^{-j2\pi mk/M} \qquad m = 0, 1, \ldots, \frac{M}{2}. \tag{1}$$

$$Y_m = \sum_{k=0}^{M+P-1} y_k e^{-j2\pi mk/(M+P)} = \sum_{k=0}^{M-1} x_k e^{-j2\pi mk/(M+P)} \qquad m = 0, 1, \ldots, \frac{M+P}{2}. \tag{2}$$

The equations differ for the exponents only, indicating that the two DFTs provide components of the same Fourier transform evaluated, however, at different frequencies. In both cases, the maximum angular frequency is π, but after zero-padding, there are $P/2$ additional components, resulting in a more densely spaced DFT. Interpolation by zero-padding does not improve the basic resolution of the spectral estimate, but results are smoother and the identification of spectral peaks may be easier (Figure 1).

Zero-padding is also used when **Fast Fourier Transform (FFT)** algorithms, which require the number of data to be a power of two, are applied on shorter time series. Zeros are appended to cause the sequence length to become a power of two, and the FFT is calculated on the zero-padded sequence.

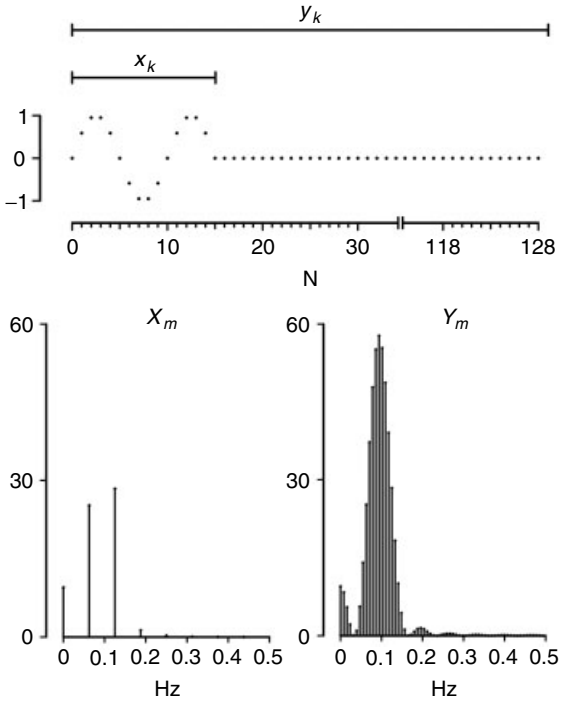

Figure 1 Effects of zero-padding. The series $\{x_k\}$ contains 16 samples of a 0.1-Hz sinusoid sampled at 1 Hz (1 cycle and half), and its DFT $\{X_m\}$ has 8 components only; $\{y_k\}$ is obtained by adding 112 zeros to $\{x_k\}$, and the DFT $\{Y_m\}$ has 64 components, 8 of which coincide with $\{X_m\}$. Although the frequency resolution of the two DFTs is the same, the 0.1-Hz spectral peak is much more easily identifiable after zero-padding

Zero-padding can be also applied to a DFT in order to interpolate in the time domain. Given a DFT $\{X_m\}_{m=0,1,\ldots,M/2}$, complex zeros are added after the component with $m = M/2$. The frequency of this component is half the sampling rate of the input series $\{x_k\}$ of length M. The maximum frequency of the DFT increases by adding $P/2$ zeros, and the padded DFT corresponds to the DFT of a new series $\{y_k\}$ sampled at a higher rate. The inverse DFT after zero-padding gives the series $\{y_k\}$ of length $M + P$, which is an interpolation of $\{x_k\}$ of length M.

PAOLO CASTIGLIONI

Zygosity Determination

Accurate determination of whether a twin pair is identical/monozygotic or fraternal/dizygotic has important implications with regard to the validity of research studies utilizing twins to assess the importance of genetic and/or environmental factors in explaining observed variation in specific traits or disease susceptibilities (*see* **Twin Analysis**). It is also important to an individual twin pair where questions of organ transplantation or risk for specific inherited diseases might arise. The approaches used in assigning zygosity include: (i) **genetic markers**, (ii) use of questionnaires to obtain self- or surrogate-reported information related to the similarity of pair members, (iii) number of fetal membranes, (iv) comparison of physical similarities of pair members, and (v) Weinberg's differential rule.

Zygosity Determination Using Genetic Markers

Zygosity determination methods using polymorphic marker systems (*see* **Polymorphism**) such as blood type (*see* **Blood Groups**), red cell enzymes, or **DNA sequence** markers have the advantage over other methods in that they provide an objective measure of **twin concordance** that is based upon qualitative genetically determined biological markers whose population frequency and modes of transmission are known. The efficiency of a given genetic marker is dependent upon the number of alleles and their frequency in the population from which the twin pair has been sampled. Since the chance of differentiating a dizygotic (DZ) twin pair from a monozygotic (MZ) pair increases with the inclusion of additional independent genetic markers, zygosity determination is usually based on typing information for a number of marker systems. Opposite-sexed pairs are assumed to be DZ on the basis of the sex difference, so that zygosity determination using genetic marker information is usually only done on like-sexed twin pairs. Any like-sexed pair that is found to be discordant for any marker is automatically classified as dizygotic. Since all MZ pairs and only a few DZ pairs will be concordant for all of the genetic markers examined, the probability statistic that is of primary interest with regard to an individual twin pair is the likelihood that that pair, if concordant, is DZ.

The probability that a pair of DZ twins will be concordant for all of n independent loci is $\prod_{i=1}^{n} \Pr_i(C|D)$, where $\Pr_i(C|D)$ is the probability of concordance for marker i given that the twin pair is dizygotic. As Lykken [13] has shown, this probability can be written in terms of the odds that a concordant twin pair is DZ, where

$$\text{odds}_{\text{DZ}|C_i} = \frac{\Pr_i(C|D)}{\Pr_i(C|M)} = \Pr_i(C|D).$$

In the situation where parental **genotype** is completely known, the final probability of dizygosity is simply the product of all of the probabilities that both members of the dizygotic pair are concordant, i.e. received the same allele from each parent, for each of the genetic makers examined. In most cases, however, marker information is available for only the twin pair, the probability that a DZ twin pair will be concordant for a given marker must be inferred from population frequencies, and all possible mating combinations capable of producing the observed phenotypes must be taken into account.

The estimate of the total probability that a random DZ twin will be concordant for a given genetic marker or markers provides a measure of the efficiency of the markers used in distinguishing MZ from DZ twins. A more detailed discussion of the estimation of the probability of mono- or dizygosity using genetic marker information has been provided in [17] and [18]. Historically, the genetic marker systems typically used in zygosity determination have included blood group type (*see* **Blood Groups**) (ABO, Rh, MNS, P, Lewis, Lutheran, Kidd, Kell, and Duffy), serum group (haptoglobin, C3, Gc, and lipoprotein), and red cell enzyme (PGM, adenylate kinase, adenosine deaminase, and acid phosphatase). Jablon et al. [11] found an average "efficiency" of 4% for five marker systems that included ABO, Rh, MN, haptoglobin, and Gm, while Magnus et al. [14] calculated it in their sample to be 0.0023 using 17 marker systems. DNA polymorphisms have come into wide use in zygosity determination as increased numbers of markers have become available and the cost associated with the molecular genetic analyses involved has decreased. In comparisons of DNA with other markers, both Derom et al. [5] and Eufinger et al. [8] have found DNA markers to be more

efficient in determining the correct zygosity than the other markers used in the past. Another important advantage of this method lies in the extremely small DNA sample needed to carry out analyses. This method is less invasive than those requiring larger samples because adequate amounts of DNA for zygosity determination can be extracted from buccal scrapings or from blood spots obtained as part of newborn screening.

Zygosity Assignment by Questionnaire

Issues of cost and efficiency generally preclude the use of genetic markers or DNA as a means of assigning twin zygosity in large-scale epidemiologic studies. As has been shown by a number of investigators, the use of self- or surrogate-reported information on the degree of similarity of twin pair members has been found to be an inexpensive and generally reliable method for zygosity determination in large-scale questionnaire surveys.

Cederlöf et al. [4] examined the accuracy of questions about the twins' similarity when growing up and the degree to which they were confused as children when both questionnaire and genotyping information were available on five independent blood group systems. On the basis of responses to two questions: "When growing up, were you and your twin 'as alike as two peas in a pod' or of ordinary family likeness only?" and "Were you and your twin mixed up as children by parents, brothers and sisters or teachers?" the diagnosis of monozygosity agreed with blood-typing results in 72 of 73 cases for MZ twins and 99 of 108 cases for DZ twins, with 19 pairs being unclassifiable on the basis of questionnaire information. Over all, it was possible to assign the correct zygosity accurately in 92% of the twin pairs examined using questionnaire information. In a similar study of twins identified from a sample of US high school juniors who took the Nation Merit Scholarship Qualifying examination, Nichols & Bilbro [15] found that zygosity could be accurately assigned on the basis of questionnaire information on physical similarity of pair members in 93% of cases; however, as was found previously, zygosity determination using this approach was more accurate for MZ than for DZ pairs. Magnus et al. [14] examined the accuracy of questionnaire information in determining zygosity in a sample of Norwegian twin pairs where zygosity had been previously established using genetic markers for 17 polymorphic systems. Twins were queried as to whether they were as alike as "two drops of water" in childhood, or were just as alike as siblings; the extent to which parents, siblings, grandparents, classmates, teachers and strangers had difficulty in telling the twins apart; the degree of similarity of eye color, hair color, hair type, height, weight, teeth, voice, muscular strength, dexterity, temperament, musicality, and language ability; whether they thought they were identical or fraternal or did not know; and why they thought they were identical or fraternal. Results of discriminant analyses of this information estimated that the zygosity would be misclassified for 2.4% of the pairs if the questionnaire responses of both pair members were used and for 3.9% of the pairs if questionnaire information was provided by only a single pair member. These results indicated that zygosity could be reasonably accurately assigned on the basis of information provided by a single twin pair member. Eisen et al. [7] obtained similar results for male veteran twin pairs, which suggested that the results obtained were independent of nationality and provided further evidence that questionnaire surveys are a reliable means of assigning twin zygosity in large epidemiologic studies.

Although a number of studies have been conducted to establish the reliability of questionnaire information obtained from adults as a method of zygosity determination, less information is available with regard to the reliability of this method for children. Bønnelykke et al. [2] examined the reliability of information on twin pair similarity obtained from mothers of twins between 6 months and $6\frac{1}{2}$ years of age. When compared to the results of zygosity determination based on genotyping information, the frequency of misclassification using information provided by the mother concerning whether the twins had ordinary or more than ordinary family likeness, the same hair and eye color or different hair and/or eye color, and whether the mother did or did not have difficulty in telling the twins apart was 4%. Five percent of pairs could not be classified using questionnaire information. Goldsmith [9] has constructed a zygosity determination questionnaire specifically for use in studies of infants and young children. However, information on the validity and reliability of this instrument in assigning correct zygosity is not available.

Zygosity Determination Based upon the Number of Fetal Membranes

Dizygotic twins arise from the independent fertilization of two eggs by two sperm. The two blastocysts then implant separately in the wall of the uterus, which results in each embryo being wrapped it its own chorionic membrane. Although these membranes can fuse as a result of the blastocysts being implanted in close proximity in the uterus, the placenta is always of the dichorionic-diamniotic type, i.e. two chorions and two amnions. Fusion between the chorionic membranes of DZ twins occurs in about 50% of cases. Monozygotic twins, on the other hand, arise from a division of the embryo that can occur either before or after the formation of the blastocyst, or, if it occurs after implantation of the blastocyst in the uterine wall, before or after the development of the amnionic membranes. If the division occurs before development of the blastocyst, placental development will occur in the same manner as seen for DZ twins and the placenta will be dichorionic-diamniotic. This occurs in approximately one third of MZ twin pairs making them indistinguishable from DZ pairs with regard to placenta type. If the division of the embryo occurs after implantation of the blastocyst in the uterus, there is only one chorion and the placenta is of the monochorionic type. MZ twin pairs arising from the division of the blastocyst after implantation but before the development of the amniotic cavity have individual amnions but only one chorionic membrane and are termed monochorionic-diamniotic. Those arising from division of the blastocyst after development of the amniotic cavity share that cavity and have a single chorionic membrane and are termed monochorionic-monoamniotic.

Since DZ twins cannot have monochorionic placentas, examination of the number of fetal membranes present in a twin placenta can be and has been successfully used for zygosity determination with regard to MZ pairs. In general, examination of fetal membranes is the only method currently available for identifying MZ twins with complete certainty [6]. An accurate assessment of fetal membranes, however, can only be made by an experienced pathologist since, in many cases, it will require microscopic examination of a cross-section of the placenta. Since fused dichorionic placentas are virtually indistinguishable to the naked eye from monochorionic placentas, information that there was only a single placenta at birth, in the absence of a pathology report, does not support the conclusion that the twin pair is identical.

Zygosity Determination Based on Comparisons of Physical Similarities

Similarity for physical characteristics such as height, weight, hair and eye color, and dermatoglyphic or fingerprint patterns have also been used in assigning twin zygosity. Allen [1] examined the diagnostic efficiency of fingerprint differences in zygosity determination and found both finger ridge count and pattern type to be useful in discriminating between MZ and DZ twin pairs. Jablon et al. [11] examined the reliability of zygosity determination on the basis of recorded height and weight, hair and eye color and fingerprints in 2805 twin pairs who served in the Armed Forces of the US during World War II. On the basis of fingerprint information alone, the average rate of zygosity misclassification was 22.6%. When fingerprint data were supplemented by information on height, weight, hair color and eye color, the average misclassification rate dropped to approximately 13%. Segal [16] found results of dermatoglyphic analyses to agree with those based on blood typing in 85% of cases and noted that, although fingerprint patterns are highly heritable, they are sensitive to environmental influences inutero which can differ between members of a twin pair depending upon the type of placental membrane structure present and should not be used as primary indices for assigning zygosity. Similar caution should be exerted when zygosity determination is based on a global impression of zygosity reached during a face to face contact with a twin pair. It should be noted that the effects of twin–twin transfusion syndrome can still be seen in many cases even into adulthood, and there can be significant differences in heights and body sizes of affected MZ twin pair members.

Weinberg's Differential Rule

In large population studies where it is only possible to obtain the sex of the twin pair, the frequency of MZ and DZ twin pairs can be estimated for that population using Weinberg's Differential rule. Under the assumptions that males account for 50% of DZ twins and the sex of both members of a

DZ pair is independently determined, this method estimates the number of MZ twins as $L - U$ and the number of DZ twins as $2U$, where L is the number of like-sexed pairs and U is the number of unlike-sexed pairs. Bulmer [3] has shown that the effects of violating the first assumption are negligible. Although James [12] has raised a number of points that suggest that Weinberg's rule is flawed, Vlietinck et al. [19] and Husby et al. [10] found the observed distributions of MZ and DZ pairs in two consecutive series of twins to be in agreement with those predicted using Weinberg's rule. This suggests that this method is valid as a rule of thumb but should not be considered as definitive for estimating the distribution of MZ and DZ twin pairs within a given population.

References

[1] Allen, G. (1968). Diagnostic efficiency of fingerprint and blood group differences in a series of twins, *Acta Geneticae Medicae et Gemellologiae* **17**, 359–374.

[2] Bønnelykke, B., Hauge, M., Holn, N., Kristoffersen, K. & Gurtler, H. (1989). Evaluation of zygosity diagnosis in twin pairs below age seven by means of a mailed questionnaire, *Acta Geneticae Medicae et Gemellologiae* **38**, 305–313.

[3] Bulmer, M.G. (1976). Is Weinberg's method valid?, *Acta Geneticae Medicae et Gemellologiae* **25**, 25–28.

[4] Cederlöf, R., Friberg L., Jonsson, E. & Kaij, L. (1961). Studies on similarity diagnosis in twins with the aid of mailed questionnaires, *Acta Genetica Statistica Medica (Basel)* **11**, 338–362.

[5] Derom, C., Bakker, E., Vlietinck, R., Derom, R., Van den Berghe, H., Thiery, M. & Person, P. (1985). Zygosity determination in newborn twins using DNA variants, *Journal of Medical Genetics* **22**, 279–282.

[6] Derom, R., Vlietinck, R.F., Derom, C., Keith, L.G. & Van den Berghe, H. (1991). Zygosity determination at birth: a plea to the obstetrician, *Journal of Perinatal Medicine* **19S1**, 234–240.

[7] Eisen, S., Neuman, R., Goldberg, J., Rice, J. & True, W. (1989). Determining zygosity in the Vietnam Era Twin Registry: an approach using questionnaires, *Clinical Genetics* **35**, 423–432.

[8] Eufinger, H., Rand, S.P. & Schutte, U. (1995). Use of single and multi-locus and polymerase chain reaction systems for zygosity determination – clinical application in twins with clefts of lip and palate, *Acta Geneticae Medicae et Gemellologiae* **44**, 25–30.

[9] Goldsmith, H.H. (1991). A zygosity questionnaire for young twins: a research note, *Behavior Genetics* **21**, 257–269.

[10] Husby, H., Holm, N.V., Grenow, A., Thomsen, S.G., Kock, K. & Gürtler, H. (1991). Zygosity, placental membranes and Weinberg's rule in a Danish consecutive twin series, *Acta Geneticae Medicae et Gemellologiae* **40**, 147–152.

[11] Jablon, S., Neel, J.V., Gershowitz, H. & Atkinson, G.F. (1967). The NAS–NRC Twin Panel: methods of construction of the panel, zygosity diagnosis, and proposed use, *American Journal of Human Genetics* **19**, 133–161.

[12] James, W.H. (1979). Is Weinberg's differential rule valid?, *Acta Geneticae Medicae et Gemellologiae* **28**, 69–71.

[13] Lykken, D.T. (1981). Blood typing and twin zygosity: a comparison of two methods, *Acta Geneticae Medicae et Gemellologiae* **30**, 293–295.

[14] Magnus, P., Berg, K. & Nance, W.E. (1983). Predicting zygosity in Norwegian twin pairs born 1915–1960, *Clinical Genetics* **24**, 103–112.

[15] Nichols, R.C. & Bilbro, W.C. Jr (1966). The diagnosis of twin zygosity, *Acta Genetica Statistica Medica (Basel)* **16**, 265–275.

[16] Segal, N.L. (1984). Zygosity testing: laboratory and the investigator's judgement, *Acta Geneticae Medicae et Gemellologiae* **33**, 515–521.

[17] Smith, S.M. & Penrose, L.S. (1955). Monozygotic and dizygotic twin diagnosis, *Annals of Human Genetics* **19**, 273–289.

[18] Sutton, H.E., Clark, P.J. & Schull, W.J. (1955). The use of multi-allelic genetic characters in the diagnosis of twin zygosity, *American Journal of Human Genetics* **7**, 180–188.

[19] Vlietinck, R., Derom, C., Derom, R., Van den Berghe, H. & Thiery, M. (1988). The validity of Weinberg's rule in the East Flanders Prospective Twin Survey (EFPTS), *Acta Geneticae Medicae et Gemellologiae* **37**, 137–141.

L. COREY